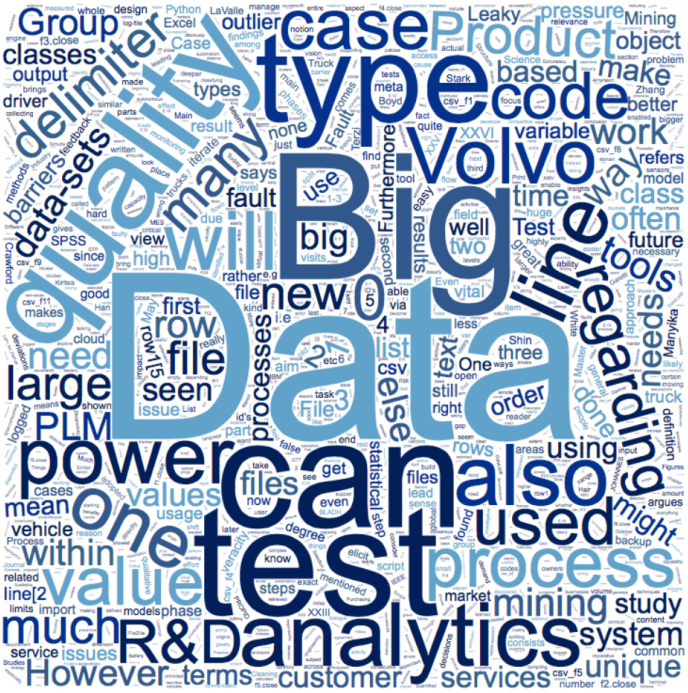
CHALMERS
UNIVERSITY OF TECHNOLOGY



Big Data and Product Lifecycle Management

Case Studies from the Automotive Industry

Master's thesis in Master Programme Product Development

JOHANNES BLADH AND KRISTOFFER HERTZMAN

Department of Industrial and Materials Science
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2017

Master's thesis 2017

Big Data and Product Lifecycle Management

Case Studies from the Automotive Industry

JOHANNES BLADH
KRISTOFFER HERTZMAN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Industrial and Materials Science
Division of Product Development
Chalmers University of Technology
Gothenburg, Sweden 2017

Big Data and Product Lifecycle Management
Case Studies from the Automotive Industry
JOHANNES BLADH AND KRISTOFFER HERTZMAN

Supervisor: Amer Catic
Examiner: Dag Bergsjö, Industrial and Materials Science

Master's Thesis 2017
Department of Industrial and Materials Science
Division of Product Development
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2017

Big Data and Product Lifecycle Management
Case Studies from the Automotive Industry
Johannes Bladh and Kristoffer Hertzman
Department of Industrial and Materials Science
Chalmers University of Technology

Abstract

Big Data is on the horizon for many industries and the automotive is no different. Therefore is the aim of this project to explore Big Data opportunities and barriers for an automotive company, focusing on knowledge-driven product development and new services. It does so by looking at it from a Product Life-cycle Management (PLM) perspective, where opportunities and barriers are examined and compared between different divisions of the company. These three divisions are R&D, Operations and Aftermarket. The project also aims to obtain a holistic view, and discuss how these divisions interconnects and what requirements this puts on Big Data solutions.

The approaches used are interviews and observations presented as case studies, where people with connection to Big Data and the automotive company gives their view on the subject. It also includes a data mining experiment where a large set of data is examined in order to get a deeper understanding of requirements and barriers for working with Big Data analytics.

The project concludes that the potential for knowledge-driven product development powered by Big Data is huge. Actual data on product and user behavior can be used to make decisions on product design, errors can be detected live and shorten feedback-loops and software updates can be done wirelessly. Barriers for this are capabilities in terms of personnel skills, data quality, automatic analytic tools and traceability.

Other Big Data opportunities in automotive are new services such as predictive maintenance in Operations and Aftermarket, leading to quicker identification of quality problems enabling better customer services and less quality issues in production.

Keywords: Big Data, PLM, Automotive, Product Development, Data Mining.

Acknowledgements

Thanks to our supervisors Dag Henrik Bergsjö at Chalmers, Amer Catic at and Johan Olofsson at Yolean for support and guidance along the way. Thanks also to all interviewees for your time and effort during this project.

Johannes Bladh and Kristoffer Hertzman, Gothenburg, May 2017

Contents

List of Figures

1
Introduction

This thesis is a part of the PROPID (2015-06912) project financed by the Swedish agency for innovation systems (Vinnova) through the BADA (Big Automotive Data Analytics) Program. In this section the background to the project is introduced along with aim, objectives, scope and thesis structure.

1.1 Background

Big Data has been a buzz word for quite a while, but companies still work on how to best use it. The automotive company and Chalmers are interested in how Big Data can be used in product development, and this project is a joint venture between the two parties. For the automotive company the potential of using Big Data is, for instance, to better predict product behavior, based on production data and customer behavior. This could be used to foresee maintenance needs, and adapt the production line based on how the vehicles on the roads performs. In the automotive industry, there is a gap between the time product and process designs are changed and the time the effects of these changes are starting to show on the market. The consequence is that companies lose out on the most important feedback, i.e. the one from their customers. The gap makes it hard to conduct cause and effect conclusions (PROPID, 2015). Big Data could be a way of bridging this gap in the future. Because of this, the automotive company have invested time and resources to this project. Another stakeholder is the truck dealership Rejmes, which provides services for the aftermarket.

From an academic perspective, the question of how data can be used from one product phase to another is one of the focus areas in Product Lifecycle Management (PLM). One example of this practise is when 3D-models from the design phase are used in manufacturing. Using data from one phase to another are in reality complex, one reason being that different software often are used during different phases. Furthermore, experts in one product phase might not always have insight on how decisions effect other later product stages. Since a Big Data solution would collect data from different parts of the organization, it makes sense to apply methods and theory from PLM because of the fact that these problems often are faced there. Before PLM was the name of the game to listen to the voice of the customer. Even if this still applies, with PLM, it is now also to listen to the product. With new technologies such as Internet of Things (IoT) products can communicate with managers in a whole new way (Stark, 2015). This paradigm shifts has led to a large amount of data over the whole lifecycle, leading in to how Big Data connects to

PLM.

Big Data have the potential to have radical impact in PLM, and will improve many areas e.g. manufacturing, R&D and supply chain (Manyika et al., 2011). It will optimize assembly processes, reduce cycle time, and elicit customer needs so the future of Big Data in PLM is promising (Li, Tao, Cheng & Zhao, 2015). Thusfar, much of the PLM research has focused on specific areas in the lifecycle, not as much has focused on the chain as a whole (Li et al., 2015).

The use of Big Data is expected to increase, and according to the vice president at Gartner (2016), three trends will drive this change in coming years:

- Data will not only reflect performance, but also drive the business operations
- Organizations will aim for a more holistic approach to data and analytics, connecting the different departments of the business. This will allow for data management and analytics from the core to the edge of the business and possibly between different lifecycles.
- Executives will use data and analytics to shape business strategy, allowing new roles for analytics professionals.

With this in mind, studying the change and development of Big Data management and analytics in such a huge business as automotive is of high interest and relevance.

1.2 Aim

The aim of this project is to take a step towards a paradigm shift in the automotive industry. A shift where product individuals can be connected to Big Data and used to predict future product behavior and with this develop better products and services. The long term hope is to achieve:

- Increased speed and accuracy by which product developing companies can create new knowledge about their own products and markets based on actual data collected from production and end-users.
- Maximize the value of each product individual by using each individuals production and usage history to identify and predict quality problems in the whole product lifecycle.

1.3 Objective

The aim is to be reached by a case study at a automotive company from a PLM perspective. This means dividing the organization into three phases: R&D, Operations, and Aftermarket. By analyzing data and talk to stakeholders in each phase the following questions will hopefully get answered:

- What are the barriers and opportunities for knowledge-driven product development across the lifecycle using Big Data analytics?
- What new services can come from Big Data analytics and what are the barriers for taking advantage of these?

These questions will be answered by gaining knowledge from interviews and study visits, so called experience based knowledge. However, it will also be done

by examining a large set of data provided by the automotive company, where the expectation is to gain insight from it without a clearly defined problem. This is called data based knowledge, and is a way to understand correlations and root-causes by analyzing data without knowing the problems before hand. The result will give insight in the potential of a feedback approach to PLM and Big Data in the automotive industry. The findings from the objectives will be presented in a report, including recommendations for further work. If positive results are achieved, it might also include recommendations regarding how to implement the findings.

1.4 Scope

Motors and batteries are in the periphery for discussion due to interest of the automotive company and Rejmes as components of importance. The scope is limited to a few specific sets of data and the findings is therefore only true for this specific data-set, and can not necessarily be used to explain product behavior connected to other data-sets or products.

1.5 Thesis Structure

The thesis is based on a theoretical framework, a qualitative part and a data mining experiment. The qualitative part includes interviews and study visits of stakeholders from three different lifecycle phases, in this report defined as R&D, Operations and Aftermarket. The experiment is an attempt to find root-causes of quality issues by applying statistical models to production data. A summary of the chapters and their content can be seen in Figure 1.1.

Introduction	Describes the background of the thesis, the aim of the project, objectives and scope.
Theoretical Framework	Previous studies regarding Big Data, PLM and automotive.
Methodology	The research approach, including the quantitative and qualitative methods used.
Results	Aggregation of the conducted interviews and visits, along with the result from the quantitative case.
Analysis	Compares the results with previous studies and identifies similarities and differences between R&D, Operations and Aftermarket in the Product Life-cycle.
Discussion	Discussion about the project and how to move forward.
Conclusion	Concluding all parts.

Figure 1.1: Thesis Structure

2
Theoretical Framework

This chapter aims at introducing the reader into relevant concepts for the project along with a review of earlier work.

2.1 Big Data

The definition of Big Data is unfortunately not an exact quantity. One definition of Big Data consists only of the notion that supercomputers are needed to manage or analyze the vast quantity of data (Boyd & Crawford, 2012, 5; Manyika et al., 2011; Chen, Chiang & Storey, 2012, 4; Zikopoulus, Eaton, deRoos, Deutsch & Lapis, 2011). The definition is intentionally abstract as the quantity of data will increase and therefore would an exact definition in terms of quantity not be relevant within a time frame (Manyika et al., 2011). What was considered to be Big Data then can now be managed by a desktop computer (Boyd & Crawford, 2012, 5). Both processing power and quantity of data increases. Where some are satisfied with that definition, some have tried to get a better feel for what Big Data entails. Big Data can be defined by the three V's: Volume, Variety and Velocity(Zikopoulus et al., 2011; McAfee & Brynjolfsson, 2012).

Volume - quite intuitively, refers to the fact that it usually involves a very large number of data-sets or several terabytes to zettabytes of data (Zikopoulus et al., 2011). This puts constraints and requirements on the hardware used for analytics.

Variety - refers to the different types of data that can be used for analysis. The data could be sensoric data, images, GPS signals, etc. (McAfee & Brynjolfsson, 2012) and it can also be of interest to cross-reference these in your analysis (Boyd & Crawford, 2012, 5), which will put requirements on pre-processing.

Velocity - an aspect that differs from classic analytics is that Big Data applications often use real-time data for decision-making and makes predictive analysis more agile (McAfee & Brynjolfsson, 2012). This therefore put requirements on infrastructure for data collection and storage.

In addition to the three V's is another two V's suggested (White, 2012, 4; Demchenko, Grosso, de Laat & Membrey, 2013). Value being represented by the fourth V, refers to the value that the data may contribute with through the predictive analysis or other activity. The fifth and final V stands for veracity. It adds the dimension of quality to the definition. Consistency, certainty and trustworthiness are attributes that can be assigned the data due to the recording process of the data (Demchenko et al., 2013). Bias, origin and infrastructure of that process needs

to be considered due to the risk of deducting false correlations. This demands high level of competence of the analyst in terms of interpretation and analysis of the data (White, 2012, 4).

2.1.1 Big Data Analytics

Beside the V definitions, there are some criticism regarding the general notion of Big Data. Boyd and Crawford (2012, 5) define Big Data as "a cultural, technological, and scholarly phenomenon that relies on the interplay of: Technology, Analysis and Mythology". Boyd and Crawford (2012, 5) Continues to explain that mythology refers to that there is a widespread belief that vast amounts of data will lead to extraordinary insights regarding ones business. Quality over quantity is still applicable to Big Data analytics (Lazer, Kennedy, King & Vespignani, 2014, 6176). The claim of objectivity and accuracy can be misleading, much because of the veracity dimension of Big Data.

Data analytics generally can contribute greatly to overall business development. Top performing organizations cross industries has adopted analytics to a higher degree than the low performers which gives a notion of the capabilities of Big Data and analytics (LaValle, Lesser, Shockley, Hopkins & Kruschwitz, 2011, 2).

2.1.2 Opportunities

Applications such as real-time micro customer segmentation, replacing human intuition with automated algorithm based decision-making, reduce rework in manufacturing, preventative maintenance and many other can be credited the wave of Big Data (Manyika et al., 2011). There are however boundaries in the organizations to adopt more data-driven business processes. There will be a need for new talent within the field of data science along with new types of leadership, new technology, new company culture and decision-making. The adoption is considered an organizational change which can be facilitated with easy to use tools and technologies along with the right leadership (LaValle et al., 2011, 2; McAfee & Brynjolfsson, 2012).

2.2 PLM

Product lifecycle management is the activity of managing a firm's products trough their lifecycles, and doing so in the most effective way (Stark, 2015). A product's lifecycle spans from the first product idea to disposal of the product, "from the cradle to the grave". PLM includes management for all company products and how they integrate with each other. At the highest level, the objective of PLM is to increase product profit and maximize product value of existing and coming products for customers and the company shareholders (Stark, 2015). Before the concept of PLM, the information management of the company was divided into departments, such as marketing, engineering, manufacturing and after-sales. The logic behind this was that persons with most knowledge in an area seemed best suited to manage because of the training and familiarity with connected activities. However, this lead to incompatibilities between departments, and poor interfaces. There was therefore

a need for connection between departments and make the company work efficiently as a whole (Stark, 2015).

The products lifecycle is often divided into three phases:

- BOL: Beginning of lifecycle
- MOL: Middle of lifecycle
- EOL: End of lifecycle

The first phase constitutes the product creation, i.e. Research & Development and the manufacturing of the product. MOL refers to the phase where the product is used by the customer and EOL refers to the grave of the product i.e. recycling or disposal of the product (Li et al., 2015).

Throughout the value-chain resides different conceptions regarding the product, its attributes and performance. Sharing the same conception throughout the value-chain creates many benefits. Advantages that can be reaped includes reducing time required in product development cycles, quality improvement, logistics optimization, better maintenance activities etc. (Terzi, Bouras, Dutta, Garetti & Kiritsis, 2010, 4). Being on the same page require knowledge and information sharing throughout.

2.2.1 ICT-systems for PLM

There has been an effort of developing ICT (Information Communication Technology) tools aiding the sharing for quite some time. One of those tools are often referred to as PDM-systems (Product Data Management) and are central for the philosophy of PLM (Stark, 2015). The system enables knowledge and information sharing at the early phases, the information then requires other systems for the other phases. For example, PDM-systems passes the sharing ability to ERP's (Enterprise Resource Planning) where the product development process and the product delivery process coincide. More enterprise applications are developed and customized to fit the lifecycle phase's needs (Terzi et al., 2010, 4).

Integrating different systems is often considered a problem, because it often involves manual transferring of files. Most commonly used types of integration between systems for ERP and PDM (Saaksvuori & Immonen, 2005):

- Transfer file integration
- Database integration

Transfer file integration is cheap but time consuming because of the need of manual work and replication over several databases. Personnel at one department exports content from the system into a standardized file format such as *csv* and sends via email or other to the other department personnel. Database integration requires heavy implementation and is thereby expensive but speed, information in one place and automation is among the advantages. It does not mean the use of one common database but rather systematic and automated information transfer between databases (Saaksvuori & Immonen, 2005).

2.2.2 Traceability

An issue that is common for large companies is that the same item is created several times, with item referring to document, cad-file etc. It is easier to create the item

than to find the existing one. Traceability is enabled by assigning attributes to items and thereby classifying them. The traceability is equally constrained by the information structure as well as the quality of the meta data of the product (Saaksvuori & Immonen, 2005). Terzi, Panetto, Morel and Garetti (2007, 3) suggest through an extensive literature study a range of user requirements throughout the lifecycle regarding traceability. Manage warranties, legal, manage recalls, product reuse or rework, remote maintenance, information on product life and usage and many other business process requires traceability. Increasing product individual data quantity and sharing product-centric data throughout the organization puts requirements on the design of the data models (Terzi et al., 2007, 3).

2.2.3 Opportunities

There is a gap between how a product is intended to be used and how it is actually used. Collecting data in the middle of the lifecycle can help build deeper understanding about the products critical attributes with regards to degradation. Shin, Jun, Kiritsis and Xirouchakis (2009, 1-3) applies a multi-linear regression model to better understand use-patterns correlation to a specific fault code in a locomotive. There is uncertainty regarding the relations between the predicting variables. Furthermore, the statistical model contains subjective weighing approximated by seasoned engineers (Shin et al., 2009, 1-3). There are however more ways to utilize collected quantified measured data regarding product usage. It presents an opportunity for extracting the voice of the customer. Schulte (2008, 4) developed a prototype for aggregating different levels of product information to build a complete picture of the voice of the customer. Through a web-based platform, customer feedback and meta data together with physical data and ontologies enables assigning the feedback to modules of the product structure directly in the PDM-system. The integration automates the interpretation and analysis of the customer feedback usually carried out by the developer (Schulte, 2008, 4).

2.3 Big Data in Automotive

This section describes how Big Data is applicable in the automotive industry. There are some opportunities riding the trend along with some constraints regarding the subject. The utilization of Big Data are really on the verge to change many businesses (Wozniak, Valton & Fjeld, 2015). However, PLM lags behind and remains relatively unexploited. Manufacturers do often not store data, or do not know how to use it which makes for breaks in the manufacturing chain. To reach success with Big Data, the company needs to know which data to store and how to transmit data between different departments (Li et al., 2015).

2.3.1 Opportunities

Overall, the rapid development of wireless communication has lead to great potential in improving products and services in automotive vehicles. The ability to take advantage of this and use it well can be a great competitive advantage in automotive

in the years to come (Johanson, Belenki, Jalminger, Fant & Gjertz, 2014). The product development process will benefit from streaming data from vehicles because it it will enable more informed decisions in all stages of the lifecycle. This is what refers to as knowledge-driven product development, and it is a definition that is shared in this report (Johanson et al., 2014).

The value of Big Data comes in many forms, and the opportunities are as previously stated considerable. It enables for new aftermarket services and products such as predictive and preventive maintenance services, support services, autonomous driving and active safety and various infotainment services, just to scratch the surface (Johanson et al., 2014; Luckow et al., 2015).

In addition to improve and enable product-driven product development and new aftermarket services, Big Data can generate new streams of revenue in more direct ways for automotive companies. Data can be stored, analyzed and sold to third parties (Johanson et al., 2014; Manyika et al., 2011). For example, sensors in the cockpit can track what kind of music the driver listens to and in real time suggest targeted adds connected to this music (Palade, Nicolaescu & Kifor, 2015). This is only one example, but the potential is great and this kind of information can be interesting to i.e. insurance companies, governments and e-service developers (Johanson et al., 2014).

Data recorded in the MES (Manufacturing Execution System) and in the production system presents an opportunity to develop the production processes and enhance product quality (Zhang, Xu & Wood, 2016; Luckow et al., 2015). The production data together with diagnostics data recorded later in the lifecycle can help build predictive models that can be used in production (Luckow et al., 2015). Advanced analytics, such as outlier detection, can be applied to the entire flow of data for new ways of quality control (Zhang et al., 2016).

2.3.2 Barriers

According to Johanson et al. (2014) the 5V model of Big Data applies well into an automotive context, so there is no need to redefine the term in an automotive context. Yet, there is some differences from many other Big Data application areas that needs to be highlighted. For one, the data sources is often mobile since they are in moving vehicles. This makes wireless communication networks a necessity to collect data. Secondly, there is often a emphasis in time-series, and traceability. It is vital to see how things are developing over time, and to trace root-causes problems. Furthermore, safety is a huge priority in automotive, and this also reflects the use of Big Data applications in automotive (Johanson et al., 2014; Luckow et al., 2015).

The data recorded in the automotive industry also withhold great variety in regards to different formats. With the rise of connected vehicles and autonomous vehicles are GPS-data, image and video added to the group of formats (Luckow et al., 2015). Already in the systems for manufacturing and PDM-systems are data recorded as text, which presents a challenge for analytics (Arnarsson, Malmqvist, Gustavsson & Jirstrand, 2016; Luckow et al., 2015).

The data that comes from fleets of vehicles using telematic tools and services can be of very high volume, for example, a fleet consisting of 1000 vehicles can pro-

duce 4.5 terabytes of data per day. This demands a system that is highly efficient and scalable and can be done by having reliable data and statistics, as well as customer behavior data available quick and efficiently and feel this to the development process (Johanson et al., 2014). According to Johanson et al. (2014):

"The challenge is being able to efficiently capture, collect, manage, analyze and make good use of the large volumes of data, i.e. to convert collected data into useful knowledge".

2.3.3 Data, Analytics, the Process and Previous Learnings

Big Data in automotive is a more and more occurring topic and there is much to be learned from previous projects. According to Wozniak et al. (2015) there are 6 lessons learned from building a Big Data service in automotive, this boils down to 6 lessons learned:

- Identify sources first
- Expect the unexpected correlations
- Let the users prove you wrong
- Focus on domain knowledge
- Understand data through dialogue
- Build corporate Big Data policies early

Identify sources first. Investigate what the data is before starting analyzing and identify the stakeholders connected to the data. It is important to know how the data is gathered and maintained in order to validate the quality and relevance.

Expect the unexpected correlations. Wozniak et al. (2015)'s study showed that mathematical sense not necessarily will lead to business value. This is because various data sources often are included, and that the relationship within the data can be obscure. Discussion between user might give the insight needed.

Let the users prove you wrong. Their lack of knowledge in the field sparked a discussion between users which lead to further insight. In a big organization no employee can be expected to have insight of a significant part of the data, so having different part of the company explore and correcting each other can be a starting point for exploring the data-set.

Focus on domain knowledge. Struggled to match business intelligence with domain expertise throughout the study. They found that it is important to include the ones that are supposed to use the end product. They identified work flows to adapt the product to the real environment, and stretches the importance of business-driven development.

Understand data through dialogue. In the study Wozniak et al. (2015) concludes that to optimize the potential benefits from Big Data opportunities, dialogue is vital. Big Data usage is an evolutionary process and results must be delivered step-wise in order to keep all the stakeholders in the loop.

Build corporate Big Data policies early. Much can be made easier if future databases are designed with Big Data in mind. Wozniak et al. (2015) states that whenever new business branches launches Big Data initiatives they should do so with the knowledge that data bases much likely will have to be integrated into the larger infrastructure of the company.

3
Methodology

This chapter will account for the methodologies used for the qualitative respectively the data mining experiment.

The project has used a research approach containing a qualitative one along with an experiment. The experiment entails hands-on data mining, where the data origins from the automotive company's production plant in Skövde, Sweden. The factory produces engines for all the company's business units. The qualitative part explores Big Data's role in the automotive industry through a product lifecycle perspective, where interviews have been conducted with people from R&D, Operations and Aftermarket.

3.1 Qualitative

The data collection is a combination between observation and semi-structured interviews. The interviewees were carefully selected in the organization in order to get the PLM-perspective. This means that people from R&D which represents the beginning of the lifecycle, Operations the middle of the lifecycle and Aftermarket the end of the lifecycle. In some cases, study visits has been made in order to really understand the work flows and demands connected to the capabilities and potentials of Big Data. Some emphasis was put on upstream Big Data flow. In the beginning of the interview phase the questions was more open-ended in order to really get the sense of interesting topics connected to the lifecycle phases. As the project proceeded, the questions were more specific on areas of interest in order to gain deeper knowledge. The study visits was made after first having interviews, so that the relevance of the visit was determined beforehand. The observations at these visits was vital to really understand the work environment and processes.

The interviews are distributed as following based on competence:

- R&D: 3 data collection specialists and 3 product specialists
- Operations: 2 quality and 1 production development specialist
- Aftermarket: 3 aftermarket professionals

3.2 Data Mining Experiment

The experiment was conducted with the aim to attain some practical knowledge regarding Big Data analytics and illustrating the challenges. It was furthermore testing the notion of eliciting data based knowledge.

The material and information that outlined the prerequisites of the study was the following:

- Quality issue case of 7 product individuals: Quality case 1
- Second quality case of 202 product individuals: Quality case 2
- MES (Manufacturing Execution System) database backup (an external hard-drive)
- Information model describing the database UML (Unified Modeling Language)

Quality case 1 consists of a set of product individuals that had the same exact quality issue who's root-cause has eluded the automotive company personnel and was thereby of interest. In Quality case 1, product individuals has experienced the utility phase or middle of lifecycle, i.e. they have been in the hands of the customer. Quality case 2 consisted of product individuals that quite far downstream in manufacturing has failed function testing and needed rework. The case 2 populace does not all have the exact same quality issue but they are well classified and clustered within the populace.

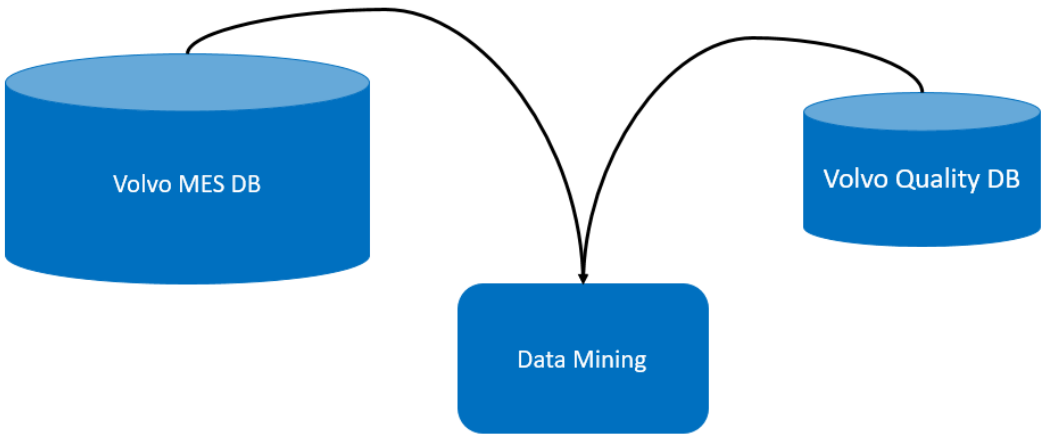


Figure 3.1: The sources of data for quality case 1 are recorded in different stages of the lifecycle of the products.

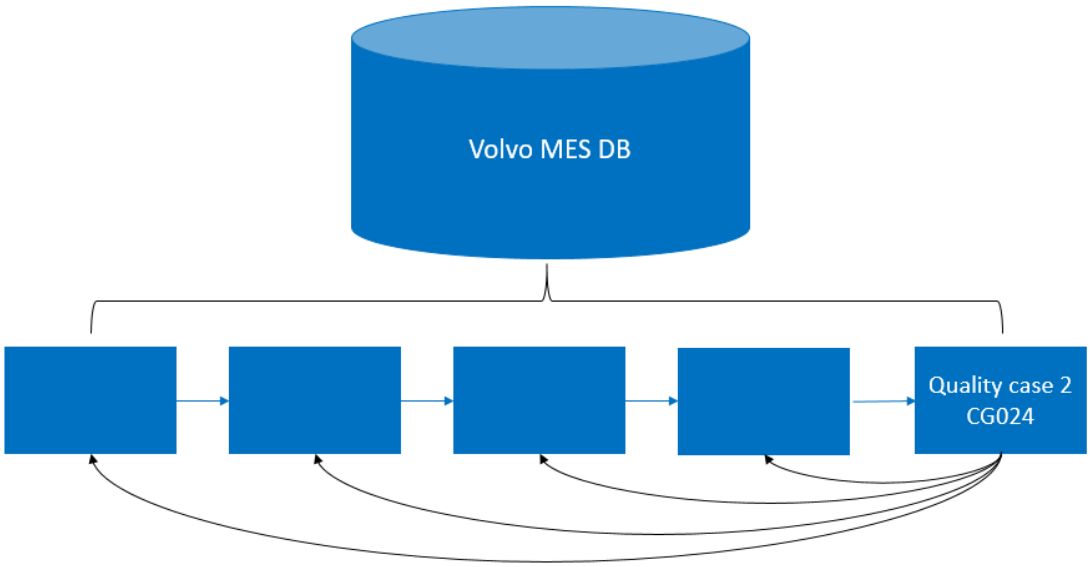


Figure 3.2: The sources of data are within the same system and quality case 2 is one of the last process steps.

The variety of data is in this experiment is delimited from text, image or video analysis. The database does not contain image or video but large amounts of the recorded data was in the form of text strings. The following sections will account for the process undertaken for the experiment along with some support from literature.

3.2.1 Data Mining Process

Data mining is a somewhat misrepresenting name to what it entails. Data mining is the extraction of knowledge through large sets of data. It is closely related to the KDD-process which is short for, Knowledge Discovery from Data (Han, Kamber & Pei, 2012; Gorunescu, 2011). The two are by some considered to be synonyms(Witten & Frank, 2005; Gorunescu, 2011). However, data mining is also viewed by some as merely the smart way of extracting knowledge using methods at the intersection of statistics, artificial intelligence, machine learning and database systems (Han et al., 2012). In this report KDD and data mining are considered synonyms, as it refers to the process of extracting knowledge. The KDD-process and thereby the data mining process is roughly outlined as the following steps:

1. Cleaning
2. Selection
3. Preparation
4. Mining
5. Pattern evaluation
6. Knowledge presentation

Cleaning refer to the task of deleting noise and inconsistencies which usually is produced to some degree at the data recording. Selecting the relevant data-sets

for the analysis is always a re-occurring task in the data mining process. Preparation is broad term for the sequence of actions needed to prepare the data for the smart knowledge extraction algorithms enclosed by the mining term. Preparation can entail further cleaning, integrating different data sources, changing formats etc. (Han et al., 2012; Fayyad, Piatetsky-Shapiro & Smyth, 1996).

Usually the need for the respective steps in the process differs from situation to situation. The variant of the process executed through this experiment can be seen in Figure 3.3.

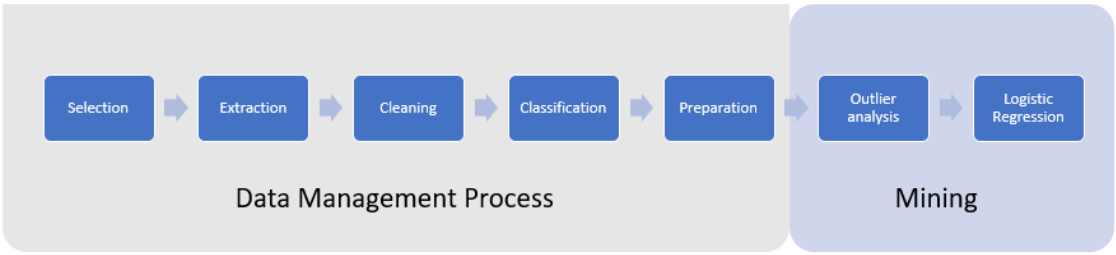


Figure 3.3: The process flow of the experiment

Like Figure 3.3 depicts, the entire process is divided into two main categories. The data management process entails all the steps required up to the actual analysis or mining steps. The mining in this project consists of two different sets of analysis, i.e. outlier analysis and logistic regression which will be described later. The steps are chronological and will be described in detail in upcoming sections.

3.2.2 Tools and Setup

The database used in the experiment is a relational database and came in the format of a large image-file and a virtual machine was needed to access the database. The system is quite old and require a specific type of hardware, an Alpha processor. The virtual machine mimics that hardware enabling the system to run. Furthermore, to navigate the database Oracle RDB, being the ICT-system, and a SQL-tool was used. The SQL-tool enabled a way to pose queries about specific engines within specific classes. It was possible to retain a sample from classes to get a clearer view of what they contained. A class is in this report referring to a collection of parameters with respective sequentially recorded data. Querying a class from a database will print all data contained in the class in the form of a table with the parameter names as headers and recorded data in each column, a row would represent one instance of recording. It is also viewed as probable that a class in the database represents a station in the production line.

To retrieve data from the virtual machine there was a need for a way to transfer files. FileZilla, a FTP-client was used to retrieve the data by setting up the virtual machine to the same network as the desktop computer.

The mining can be done in various ways to extract different knowledge, outlier analysis, cluster analysis or association/dependency analysis. Outlier analysis is a way to see whether specific data-sets differ from the total quantity. Cluster analysis is a way of determining frequency of certain events or grouping similar objects

within a data-set. The third type refers to mapping relationships between objects and determining correlation (Han et al., 2012). The analysis is to be performed with the statistical analysis tool provided by IBM: SPSS (Statistical Package for the Social Sciences) (IBM, 2016). The software uses and needs the same format as Excel, which is basically tab or comma delimited text-files. It is therefore a useful tool for the mining step of the process. SPSS is limited in capacity, by the hardware where it is run (IBM, 2016) whereas Excel is limited to 1,048,576 rows and 16,384 columns, as shown in Figure 3.4. Even though cleaning can be done with features of SPSS, there was a need for better hardware or smaller files in order to manage the process because of limited hardware. The smaller files approach was adopted by splitting the files using a developed *C++* statement script, see Appendix A. As an example, a file containing one class, extracted from the backup, consisting of approximately 1,6 million rows would be split into four files with 500,000 rows in each (approximately 100,000 in one). This enabled data management in SPSS and/or Excel. However, the data management tools was still running slow and there were files with 40 million rows resulting in 80 splits making manipulation a too large amount of work. Moreover, many of the manipulation operations needed was work intense and it was finally considered unsustainable. This became an incentive for another approach than using Excel or SPSS for data management.

The programming language *Python* is a high-level object-oriented programming language which is easy to use for rapid script and application development (van Rossum & Fred L. Drake, 2001). When dealing with data in the format of text, csv and Excel-files it provides a good tool for preparing the data. Writing Python scripts to manipulate the data was considered a good alternative and was therefore adopted for all the remaining steps of the data management process. The manipulation required only the usepackage *csv*.

3.2.3 Data Management Process

The consecutive sections will describe the data management process for the experiment.

3.2.3.1 Selection

The format needed for performing the analysis is considered to be the same for both quality cases. Data-sets directly linkable to the product individuals in each case is of interest. There is also a need for data not linked to the cases but from the same source in order to see whether the cases can display a pattern that stand out from the total quantity. The product individuals in the quality cases should thereby be traceable in the classes or data-sets:

Requirement 1

Secondly, the datasets should contain measured and quantified recorded physical data:

Requirement 2

It is deemed that that type of data is the most likely to have a correlation to the quality case populaces and also that it is a critical characteristic.

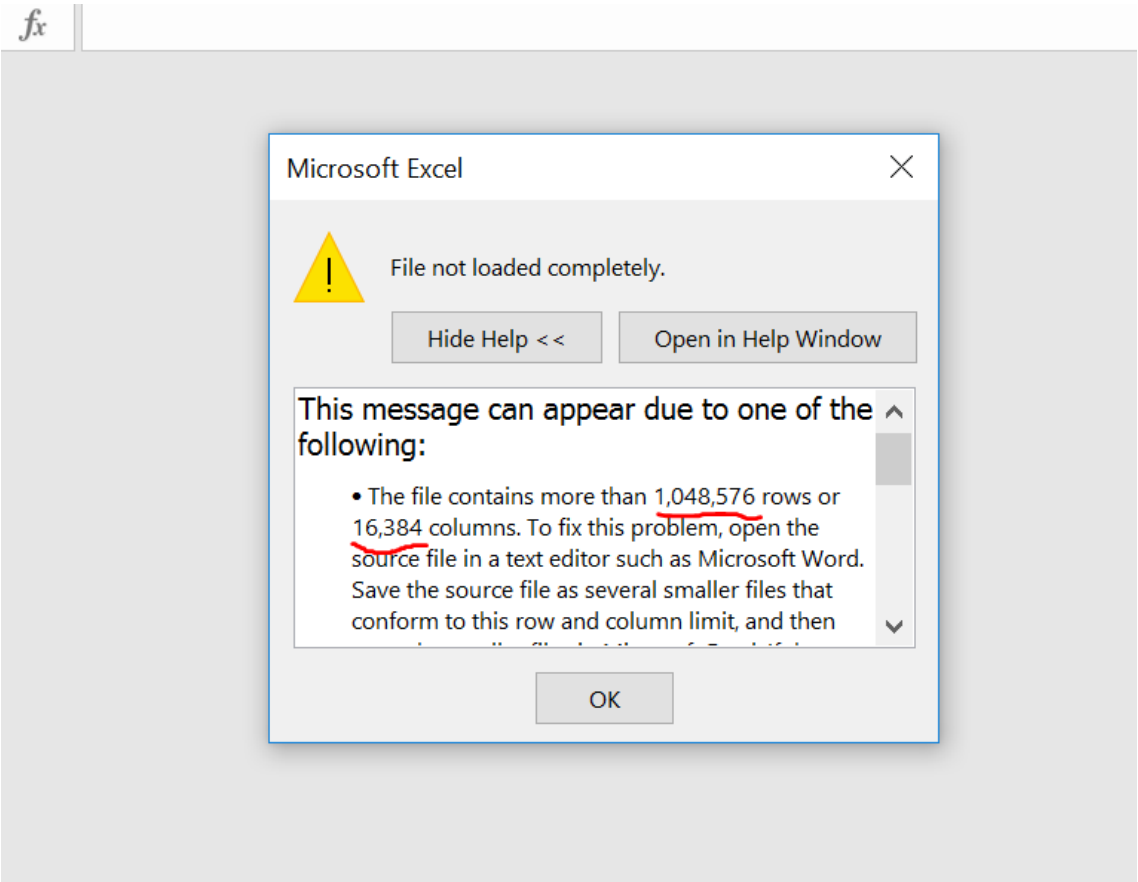


Figure 3.4: The notice Excel gives when loading too large files.

The information model over the content of the database was thoroughly studied and classes with potential regarding the requirements for the data was selected for extraction with priority. Extracting and downloading content via FileZilla was a time-consuming process and it was therefore a need for prioritization. The information model also clearly depicted the irrelevant supporting processes.

3.2.3.2 Extraction

The classes was extracted from the database by setting a log-file that recorded everything printed in the terminal of the system. By using the SQL-tool to pose queries, the contents of a class was printed to the terminal and thereby recorded in the log-file. The sequence of commands was the following:

```
set linesize 5000;
set logfile 'SQLDMP:[SQLDUMP]CG000.TXT';
SELECT * FROM CG000;
```

The log-file was dumped into a directory accessible by FileZilla. The log-file was then downloaded to the PC, in the format of a txt-file using the FTP-client.

3.2.3.3 Cleaning

All of the classes were subjected to cleaning. Extracted classes were split every row after 80 byte or characters and also contained several blank spaces making cleaning necessary. The split was done by RDB and was easily avoided by limiting each row in the SQL-tool to 5000 characters instead of 80 (set linesize 5000;). The blank spaces were eliminated using the linux tool *tr*, through *Cygwin* which enables usage on Windows operating system. These two cleaning processes resulted in csv-files with a varying number of rows depending on class.

3.2.3.4 Classification

The end goal of manipulating the format of the data is to have quantified measurement data **(Requirement 2)**, where every row is classified by the most important denominators and is unique.

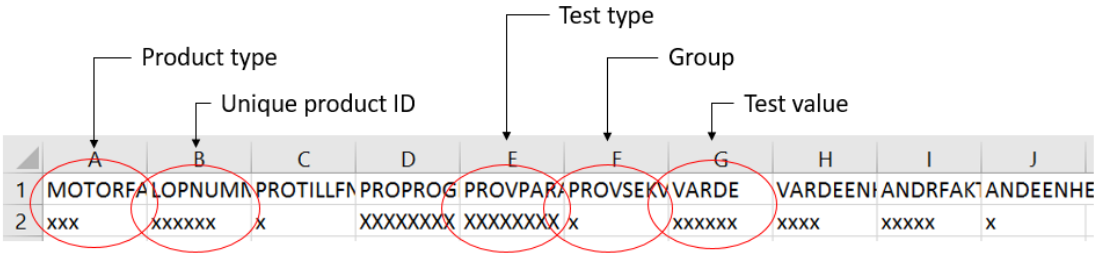


Figure 3.5: A sample of the class containing test values, CG063.

Figure 3.5 depicts a small sample of the class CG063 where the first row is a header. All of the circled columns are parameters that confine the unique row. Except *Test value* which is the nominal data that has been recorded in the manufacturing process step and is the data of interest for analysis. One product individual would undergo several different tests where all were logged in the database. There was therefore a need to classify and split into smaller data-sets using the circled parameters.

Firstly, was the class split using the *Group* parameter as Figure 3.6 shows. Each of the groups had in turn 11 different *Test types* which was the next parameter used for classification. It was printed out in a total of 5*11=55 output files for the sake of data management and keeping track of the data. Each of these files now only contained the same *Group* and *Test type* and each row was one unique product individual. The output files looked as shown in Figure 3.5 but within a file was only one and the same variable in the *Group* and *Test type* columns.

There was also a need to find the product individuals from quality case 2. A Python script indexed the rows of both files enabling comparison of the content of each row in the two files. The product ID's were compared in order to locate the product individuals from the quality case. If they were the same, the row from CG063 was printed to an output file. Thereafter was the same procedure used for classifying the output file as shown in 3.6. There was now another 55 output files only containing the same products as in the quality case.

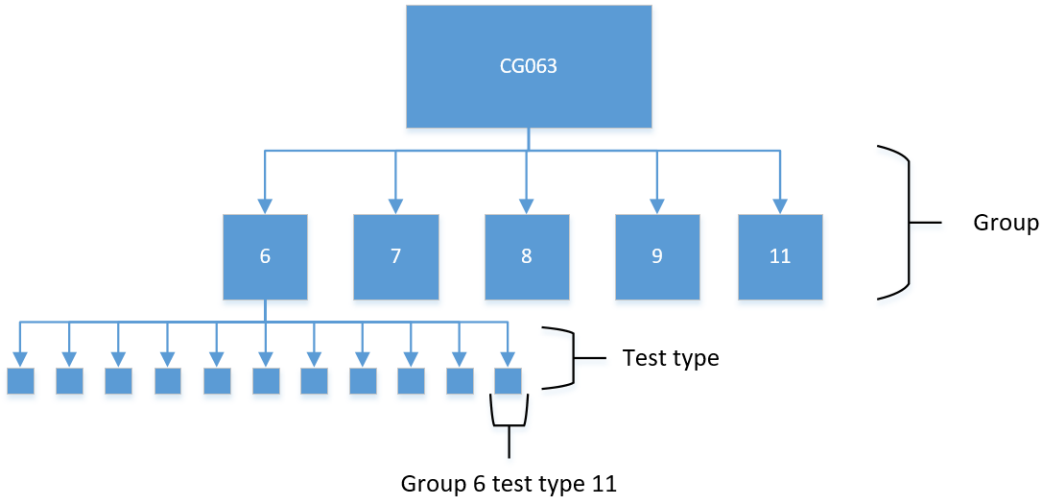


Figure 3.6: Classifying the data-sets within CG063 into 5*11 different types

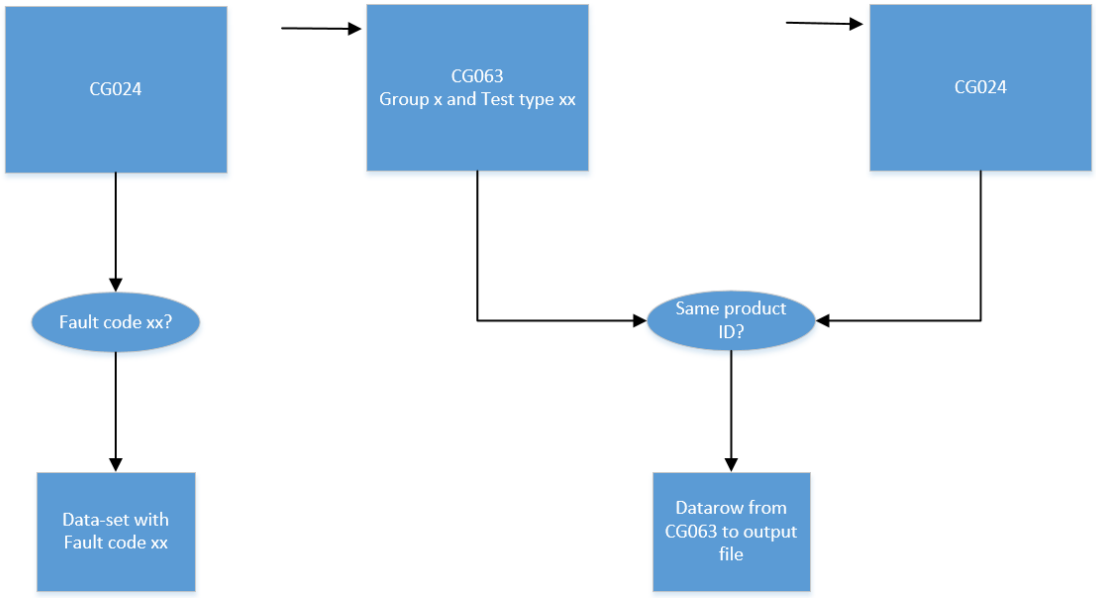


Figure 3.7: The process for compiling the data-sets for the second analysis

The logistic regression analysis was done on a different data-set than quality case 2. CG024 was a larger populace of the same type of data-set as quality case 2, but was, as described, discovered after the first analysis. This enabled a larger data-set and could thereby be classified to a higher degree without risking a too small sample size. The data-set were to be classified by *Test type*, *Product type* and *Fault code*. Product individuals with the same *Fault code* was extracted from CG024. Product individuals classified by *Test type* and *Group* was iterated over CG024 and extracted if they existed in that data-set. This procedure made certain that all the product individuals used in the analysis had passed through both CG063 and CG024. Which in turn means that all product individuals in the data-set have been assured that they do or do not have the faulty function, they've all been tested.

However, the data logged in CG024 only consisted of product individuals that has a faulty function, meaning that the entire data-set was populated by faulty individuals but not all by the specific fault of interest for the analysis. The process for compiling the data-sets are shown in Figure 3.7.

All code written for this process step can be found in Appendix B.

3.2.3.5 Preparation

Different formats were preferred for the statistical methods. The first preparation entailed aggregating the 55 files generated by the classification of the data-sets related to quality case 2. The second needed aggregation of one *Fault code* data-set with one *Test type* for each analysis.

Outlier analysis preparation

The outlier analysis required a format and alignment of the data as shown in Figure 3.8.

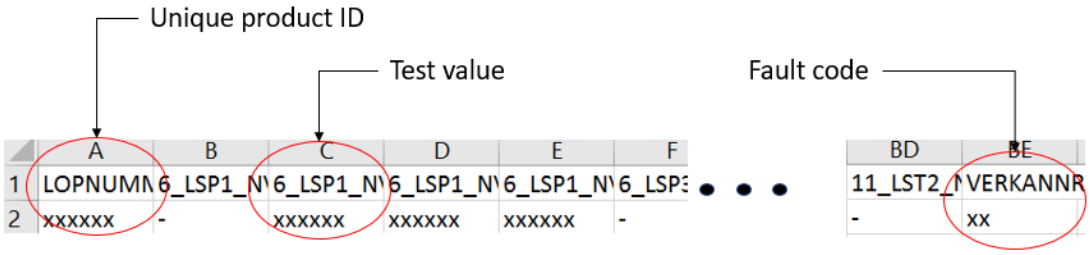


Figure 3.8: The prepared format for the outlier analysis where each row is one unique product individual from quality case 2.

The only data necessary from those files was the *Test value*. Similar to the script developed for extracting the data-sets related to the quality case, was a new script developed. It extracted and printed the test value from each file in a row assigned to a unique product ID. It did so by iterating through two files and comparing the ID's and thereby was able to find data corresponding to each ID. The output, as depicted by Figure 3.8, was a file with one row corresponding to one product. All the test values from the *Test types* that the product had gone through in 55 consecutive columns. The last column denoted what fault had been discovered on the product. Missing test values was initially denoted with a '0'.

At the first aggregation, it was realized that the data-set contained doublets. At second iteration of the aggregation, the doublets where filtered out. In addition to the small cleaning procedure was also the missing value representation of '0' changed to '-' which can be seen in Figure 3.8. SPSS interpreted the '0' as the value 0 where it was rather supposed to interpret it as a missing value.

The analysis required additional descriptive statistics that was easier retrieved by developing Python scripts than importing the data to SPSS. The mean across the whole range along with standard deviations of all the *Test values* was retrieved from each of the 55 output files. The formula for calculating the standard deviation is as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}}$$

Where n is the number of *Test values* with the index i representing the *ith* element. To comply with the realm of Python code, the standard deviation had to be calculated sequentially. The *Test values* was represented in a list object, similar to a vector, and where x had to be squared element by element using a loop in that list object. The square of the sum did not require a list object but rather just an integer and the result was printed to the terminal and manually written to the file used in SPSS.

Logistic regression preparation

The required format for the logistic regression analysis can be seen in Figure 3.9. If a product individual had failed the function test of interest and had been assigned the fault code in CG024 then that product individual received a '1' in the *Fault code* column. The *Product type* parameter was used to group rows in the analysis and thereby classifying the data-set one bit further.

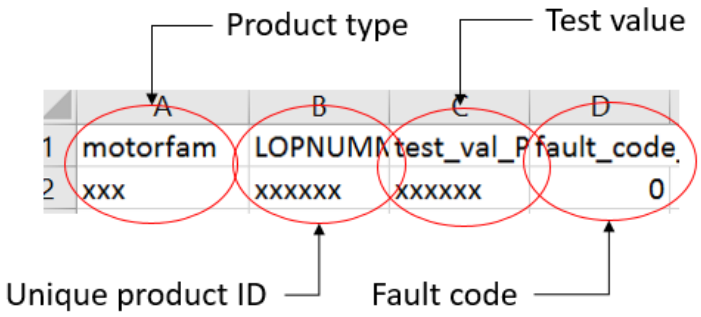


Figure 3.9: The format required for the logistic regression analysis

The preparation process for the second analysis can be seen as an extension of its classification as depicted in Figure 3.10. The outputs of the classification was iterated through and were the ID's the same then that row was consequently marked with a '1' in the *Fault code* column if not, a '0' was printed.

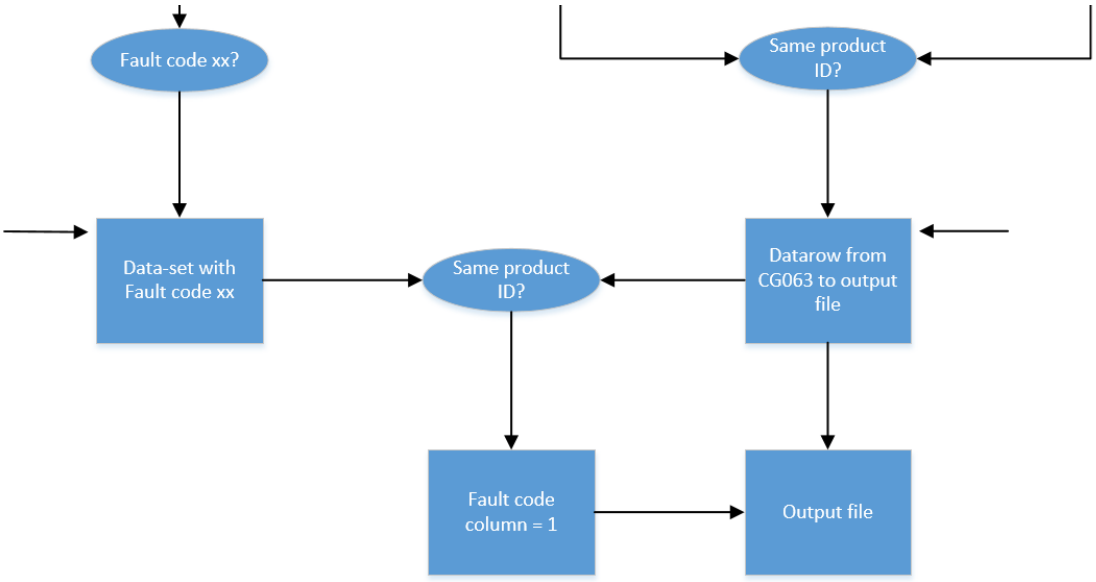


Figure 3.10: The preparation of the second analysis can be seen as an extension of its classification process

Before conducting the analysis it was realized that also these output files contained doublets. These were filtered out in a similar way as previous procedures.

The clustering of *Product types* did not require an additional classification script. It was easier to use the "filter"-feature in Excel to do that grouping.

All code written for this process step can be found in Appendix C.

3.2.4 Mining

The following sections will account for the mining part of the process.

3.2.4.1 Outlier Analysis

In this analysis, product individuals within quality case 2 was grouped after fault code. The mean value of each test in each group was calculated. Thereafter was the mean compared to a global mean derived from the rest of populace. The distance between the mean of the product individuals and the global corresponded to whether the product individuals with the fault code where outliers or not. It was assessed that two standard deviations from the global mean constituted an outlier, as can be seen in Figure 3.11. 2 standard deviations from the mean was chosen because 95 percent of the values will statistically be within the limits of 2 standard deviations (Joseph F. Hair, Black, Babin, Anderson & Tatham, 2006), so the other 5 percent was regarded as outliers.

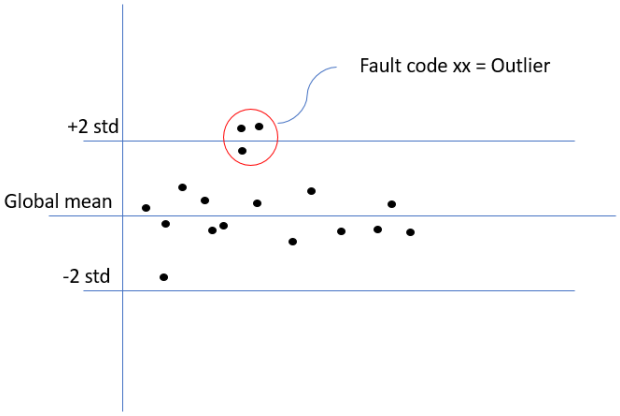


Figure 3.11: An illustration of the outlier analysis

3.2.4.2 Logistic Regression

Logistic regression is much like ordinary regression analysis but specialized to deal with some extraordinary cases. It is formulated to predict a binary categorical variable rather than a continuous metric dependant. The input is therefore one dependant with on single multivariate relationship. An example of a good application of the analysis would be the impact a variable (e.g. torque applied on a screw) has in the event of success or failure "1" or "0"(e.g. of quality test) (Joseph F. Hair et al., 2006).

The format in which the data is recorded is suited for logistic regression analysis. The *Fault code* can be expressed as a binary dependant variable, a product individual is either assigned the variable '1' and thereby has that faulty function or not and is then assigned '0'. The logistic regression analysis is considered robust and easy to interpret with these prerequisites (Joseph F. Hair et al., 2006).

4
Results

This chapter will account for some background about the main stakeholders along with the presented results of both the qualitative cases along with the main result of the experiment.

4.1 Stakeholder Information

This section starts with some background into the organizational layout and the IT-infrastructure surrounding the product individual data.

4.1.1 Case Company Information

The company is one of the world's leading manufacturers of trucks, buses, construction equipment along with marine and industrial engines. They also provide complete solutions for financial services. The company employs nearly 100.000 people worldwide, with facilities in 18 countries and sales in over 190 markets. The organizational structure consists of three divisions (Technology, Operations and Purchasing) and 10 business units, spanning from its several truck brands to financial services.

The main divisions and their function are described in further detail below:

Technology - is responsible for technology research, product design, engine development and all technology and product development connected to truck operations and supporting product development in the aftermarket. The division has 7.000 employees working in global teams.

Operations - handles all production of trucks, as well as for the other business units of different brands.

Purchasing - is covering the purchase of automotive products globally, including products for the aftermarket, for all truck brands within the company. It also includes the Indirect Products & Services (IPS) Purchasing division.

4.1.2 Rejmes Information

Rejmes is an official retailer of personal vehicles, trucks and buses (Rejmes, 2017). Rejmes also provides service, and is one stakeholder in Aftermarket in this report. With trucks becoming much more digital now, better performance for batteries is of high demand. When talking about Big Data in product development, batteries was of high interest for both Rejmes and the interviewees in R&D, and is one reason why they participated in this project. When doing interviews about knowledge-driven

product development with Big Data, it has to a large extent been done with batteries in mind. More can be seen in the sections named Aftermarket in this report.

4.2 Product Individual Data in the Lifecycle

The information in this section has been elicited from interviews and observations. Product individual data is recorded and produced throughout the lifecycle of the product.

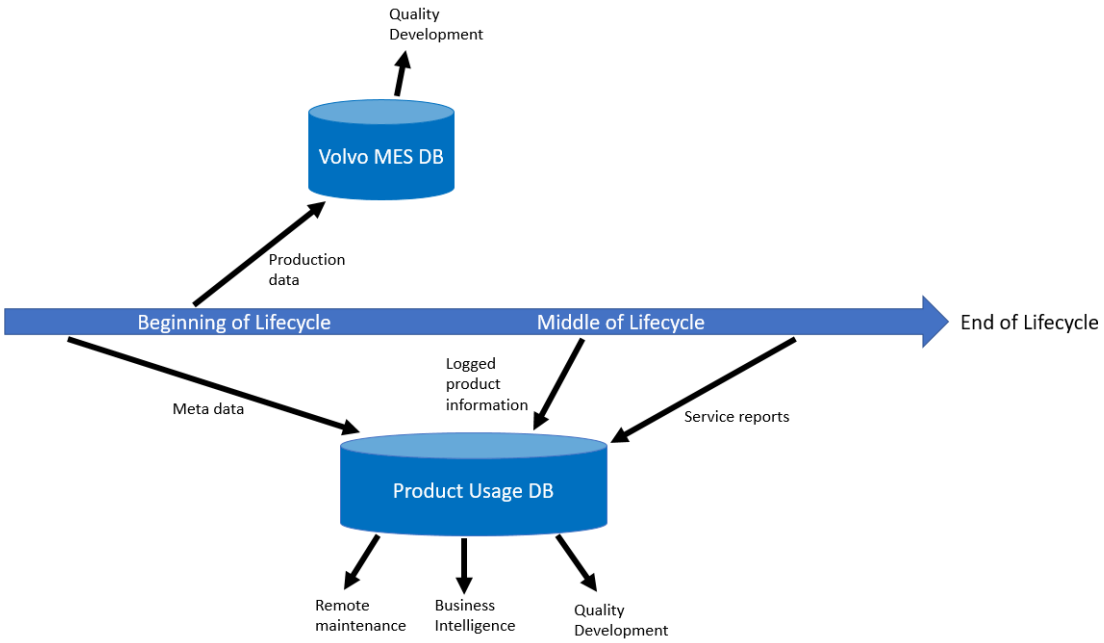


Figure 4.1: Product individual data through the lifecycle. DB is short for Database

In Figure 4.1 can the lifecycle and the related data collection be seen. The Product usage DB represents a system of data storage that are dispersed down to several applications such as preventative maintenance, business intelligence applications and statistical tools for quality engineering etc. The system is more complex than depicted in Figure 4.1, but it provides a general understanding of the data generated throughout the lifecyle regarding the product individual. Production data and its storage is detached from the rest of the data storage system. Product usage data is kept in a system that is administered by one organizational entity. Production data is administered through often old systems and some are also obviously managed by suppliers which creates organizational barriers. The old systems makes database integration harder and more cost intensive. The current integration between the entities is file transfer integration. This means that stakeholders outside the function needs to request data, upon which it will be extracted from the system in a file and then sent to the stakeholder.

Meta data in Figure 4.1 refers to parameters such as product configuration data and technical specifications of the product individual. Logged product information can be sensory measurement data retrieved from the diagnostics system. It can also be fault codes that are shown to the driver or programmed in the diagnostics system. Service reports are most commonly recorded as text with associated meta data in workshops.

Logged product information is recorded through sensors and local ECU:s (Electronic Control Unit) in the vehicle. The diagnostic system is quite extensive and not all of the data collection points are utilized. Rather, specific data-sets are requested from the vehicle, upon which the data starts being logged and then transferred to a cloud solution through telematics. Telematics refers to a box in the vehicle which receives the data from the diagnostic system and then uploads it via 4G, 3G or GSM networks to the cloud. There is a another channel to access logged product information. At the service vendor, a tool access the diagnostic system via a cable. The tool is used for root-cause and fault search as a maintenance activity. As the tool is connected, logged data is uploaded to the cloud as well. The service reports are recorded simultaneously. It consists of text strings describing maintenance activities and has meta data also connected in the same data-set.

4.3 Data Mining Experiment

The database is from the projects perspective detached from reality. It is a black box in terms of how and where the data is recorded and exactly what the data entails. This is partly to test the data based knowledge aspect of the objective of this project. It can be realized when examining the information model that large parts of the database is concerning supporting processes and thereby entirely irrelevant for the purpose. An example are classes concerning personnel data. Some classes are generated by automation and some is recorded manually by personnel. That conclusion can be drawn from the first look into the backup as some classes break pattern and become corrupt, behaviour not possible for an automated recording process. There are also many classes that has a text string as input where it is likely that the operator has entered the data.

During the time of extracting classes was a number of issues realized concerning the database:

- Many of the classes was empty or inconclusive
- Data-sets concerning the individuals in quality case 1 could not be found in any of the classes
- The names of many of the classes which was given by the information model did not exist in the database
- Some rows in some classes looked corrupt because they broke the pattern in the data

These issues was strong evidence that the backup of the database did not withhold required quality for the expected result. The last item on the list of issues is indicating that random data has been entered by personnel in the factory. The first two items indicate that the database has been flushed at several occasions or that all product individuals do not go through all processes where data is recorded. The third item indicated that the information model was inaccurate and no longer admissible as a way of selecting relevant data. The second issue meant the end of analyzing quality case 1 as no data related to those product individuals could be found. The following presented analysis is therefore regarding quality case 2 or

CG024.

Because the information model was considered inadmissible, a list of all classes contained in the database was printed out using the SQL-tool. The entire backup was now scrutinized by querying a sample from each class.

It finally resulted in one class fulfilling both **Requirement 1** and **Requirement 2** which was called CG063. Even that class did not fulfill the requirements entirely satisfactory. 102 product individuals out of 202 from *quality case 2* could be linked to data-sets in the class. The class contained different kinds of test results measured in kPa, kW etc. Each product individual was subjected to varying number of different tests. One row corresponded to one variant of a test and its result on one product individual.

While scrutinizing the database, the originating class of *quality case 2* was found, called CG024. Bigger populaces with faults could thereby be extracted increasing the quality of the logistic regression analysis. One unique row in this class contained one product individual, a fault code, a cause code and descriptions of the codes enabling a crude assessment of the data's relevance.

Quality case 2 and the aforementioned class fulfilling the requirements was used for the outlier analysis.

4.3.1 Outlier Analysis

The analysis only showed values outside two standard deviation for one fault code. However, the number of observations was way to few to draw conclusions from this. All the other values is within the margin of error (two standard deviations) and can therefore not be assumed to be the reason for the faulting engine. This test would preferably be done for all test-stations that could be a possible reason for the quality issue. When checking which fault code that had anomalies in test values it turned out that it was "specification right, no fault" which further more indicates that the result is invalid since the engine does not even seem to have quality issues.

It was also investigated whether there could be a pattern in the sequence of the nominal values in the classified groups. For example, if a group was tightly clustered in all nominal values. It was done through studying all standard deviations for all tests in the group. No significance could however be detected. The results can be seen in Appendix F.

4.3.2 Logistic Regression

For this analysis was certain parameters chosen. Information available to make the assessment regarding relevant variables is as described very scarce. In the CG024 class are however, short descriptions regarding the *Fault codes* available.

For the first trial was the description "pores" and "cracks" grouped together and constituted the dependent variable. It was thought reasonable that those kind of faults in the casting could correlate with a test of power of the engine. In this analysis was also *Product type* used to classify the data-set. Two examples of the relationships between the variables can be seen in Figures 4.2, 4.3 and all the relationships can be seen in Appendix E.

Second trial tested the description "leaky" being the dependent variable and a pressure test measured in kPa. The *Product type* was also used in this data-set. Both trials had similar results which can be seen in Appendix D.

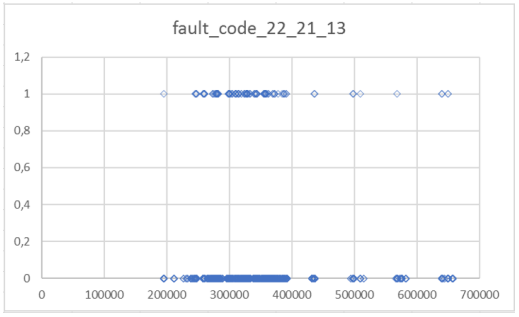


Figure 4.2: Pores/Cracks code, power test and product type 13

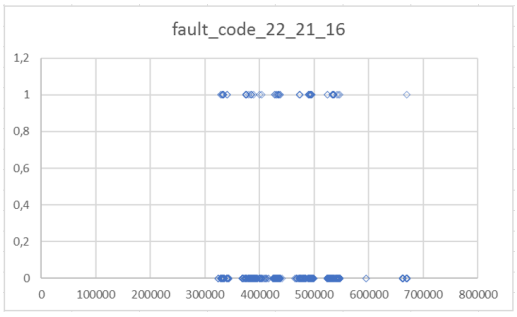


Figure 4.3: Pores/Cracks code, power test and product type 16

As can be seen in Figure 4.2 and 4.3, the *Fault code* is highly uncorrelated to the *Test value*. The Y-axis denotes the dependent variable and X-axis the test value. Similar result was achieved for all *Product types* and *Fault codes* in the trials. If there was strong correlation, a predictive model could have been created enabling a new quality assurance process.

4.4 Case I: R&D

In this section, R&D professionals with relations to the automotive company are interviewed and asked about Big Data related topics. Three of the interviewees have special knowledge about batteries, and three have special knowledge about data collection and analytics.

The main topics that have been covered is the following:

- Data
- Analytics
- Opportunities
- Barriers

4.4.1 Data

Much focus from the R&D perspective is on the data itself. Big Data poses many interesting opportunities and those will be elaborated on later, but the characteristics of such data also leads to many requirements. The data trend has lead to problems with data storage and collection argues one interviewee, and the ability to collect data is vital. R&D professionals brings up many different aspects of this. Some of the interviewees focus much on storage itself and how this could be done best. A view is that more data is collected via cloud solutions, and that the trends goes towards that. On the other hand, it is still common to collect data in batch instead of streaming it live. Examples on these occasions are when trucks are at

service vendors, or when products are collected for quality checks before truck assemblies. According to some interviewees a big challenge will be how to get clean quality data from large data-sets, and uncertainties in organizations regarding how to handle this might be a problem. One opinion often mentioned by the R&D professionals is the need to standardize the data recording in the future, just to make sure that the data is of enough quality. High veracity data is not enough, a large part of the interviewees talks about the difficulties with monitoring the data once it is collected. R&D professionals in this study says that it is important to know what data is really necessary before collecting it. By knowing this beforehand there is no need to waste storage on data that is not useful, and the analysis becomes faster and more manageable. One with very good knowledge in Big Data analytics says that it is more important to have a flexible Big Data solution that rapidly can change what data is collected, instead of just collecting everything. Furthermore the person argues that monitoring all data is very difficult. A professional at the case company says that breaking down data is necessary, and that it has to be filtered before one starts to work with it and analyze it.

4.4.2 Analytics

The second topic discussed to a larger extent is analytics. Monitoring data is hard so analysis needs to be done automatically says one with experience in Big Data analytics. "The ability to collect data, and conduct analytics that are being sent back to the customer is vital", he argues further. The old way of doing it is to send back data and then conduct the analysis. But a common view is that it is hard to monitor all the data that comes in, and that the analysis have to be done instantly and automatically. This is so important that it is considered a bigger challenge than data warehousing, especially in automotive. Better to design data collection and analysis in the consecutive steps:

- Aim
- Design collection/infrastructure
- Statistical analysis

Interviewees in this study are mentioning some key points regarding Big Data analytics. When talking about monitoring data from batteries, one interviewee raised concerns about monitoring a large fleet, just because each battery is unique in terms of previous usage. This might make it hard to draw conclusions from mean values according to the interviewee, just because it might not be valid for the battery sample. With that said, it is still common to look at large samples and compare it to individual batteries. Another brings up that with large data-sets irrelevant correlations are found in data mining, and it is important to be critical to correlations. One professional in Big Data analytics within automotive says that it is still many test vehicles they work on in terms of Big Data, and that simulations are more common than testing vehicles on the field. There is however clear that they think measuring on roads and customer behaviors poses big opportunities. In terms of tools, there needs to be filtering and sorting tools, along with easily used analytic tools in order to facilitate data driven decision-making.

4.4.3 Opportunities

Big Data in automotive has large potential and there are many opportunities. Both R&D personnel at the automotive company and at external companies working with Big Data within the automotive industry acknowledge preventive maintenance as a great opportunity. General product intelligence from customer behavior can generate value to R&D says an employee at the automotive company. Another interviewee brings up benefits for both products and services, and adds that text based analysis in automotive is an area of interest in the future. The interviewee's explain that much can be accomplished with search and text analysis in organizations. The company's core is highly applicable on the deviation reports recorded by R&D, which is engineering reports recorded during development. Much of the data recorded in manufacturing is also text-based. Insights derived from text-based data are often just intuition from the individual. That is because there is a lack of tools for analyzing text, which is therefore a great business opportunity for third parties. This was also observed during a root-cause analysis session at the automotive company.

4.4.4 Barriers

There are of course obstacles and capability issues that needs to be solved to go forward with Big Data in automotive. The interviewees says that the right competencies are imperative for analytics, and that some roles in companies today might have to be redefined in the future. Today, it is good if you know what to look for, otherwise it is not easy, but there is great potential. Sensible filters and sorting is everything, then it is required in the end, someone who understands the data and find the root cause of the problem. The task of traceability (much focused on in this project) is considered to be a very interesting challenge, and some R&D professionals in this study requests more traceability. However, one barrier for traceability is the use of many different systems. According to Big Data experts, the key is connecting many smaller systems, rather than having one enormous system. Other capability issues that might arise is the fact that many of these projects includes many stakeholders, and that prioritizing between resources is not always easy.

A key issue is the personal integrity connected to this, and this must be closely monitored. There are different perspective on this, one R&D engineer says that changed laws connected to data collection might make the work much harder. At the same time the potential danger with logging more and more behavioral data are apparent and not taken lightly by any interviewee. There are different rules depending on where the market is geographically. At the Chinese market the government demands that data are collected from the vehicles, so the laws are highly interchangeable.

4.5 Case II: Operations

This perspective comes from stakeholders that very often look at and work with quantitative data. Decision-making is both driven by data and tacit knowledge. Data in the form of text tends to introduce a small element of subjective bias which

was observed at study visits. For example, if the root-cause of a problem eludes workshop personnel then frustration might be expressed in the documenting process, or data recording because that personnel does the documenting.

4.5.1 Data

The general perception of data is "the more the better". That is concerning the amount of inputs for specific cases rather than a bigger populaces to work with. This is due the root-cause analysis before taking action regarding a quality issue from the field. This a time-consuming process where common denominators in the populace of the case is identified. It is in the root-cause identification process that the large amount of information about the products are required. Traceability thereby becomes an imperative requirement for much of the analysis. It is expressed that "a data-trail across the lifecycle would be great". There are organizational barriers to this but also plain access to data. For instance, accessing production data requires getting a hold of an administrative middle man that can assist. "Some cases require the extraction of CC's, critical characteristics that is". Critical characteristics are features of the product where the company guarantees quality for the customer. That type of data origins from manufacturing.

The biggest input to the work comes from the market, i.e. workshops. The data recorded at the workshops varies greatly in quality and quantity. Some of the data is recorded in a way that gives room for much variation. Text fields where the workshop personnel enter the fault and action taken varies in language, quantity of the description and foremost precision of the description. Standardizing the recording procedure is an obstacle from many perspectives but would minimize the investigative work needed as many hours are spent decrypting what the individual in each case tries to convey.

4.5.2 Analytics

Quality issues are mainly detected via input from the market, where different indexes are monitored. However, prioritization is still an issue and which in itself needs investigative work, specifically digging through data. There is much room for interpretation error in the data-sets, due to bias but also because circumstantial effects on specific cases. One fictional example mentioned was the higher fault frequency of a big type of battery (big capacity) where the fast conclusion would be that it is a bad product which should be excluded. However, circumstances are that the customers having that configuration are using much more of the capacity than the usual customer. Therefore should the conclusion rather be to offer customers even bigger batteries.

The data trend within production development is moving towards utilizing advanced algorithms and models for optimizing and monitor production flows. "It is of interest to analyze deviations in real time in the flow". There are barriers in terms of old systems recording data and competence within the relevant fields such as machine learning etc. A learning deducted from the experiment is that much production data is recorded in text strings entered by the operator which adds to

the competence requirement. However, that issue could be exclusive for the case.

Customized statistical analysis tools are used for the most common and standardized tasks. The tools range from products from SAS (Statistical Analysis Software) to more simple products like Qlikview and Excel. "Different data visualizations and sources require different tools", there is also a constraining factor of knowledge regarding some tools.

4.5.3 Opportunities and Barriers

The experiment conducted in this project is considered interesting by the interviewees. Training algorithms with data from later in the lifecycle and correlating with production data could enable prediction of quality issues in the field. However, there are so many parameters and dimensions to consider that it will be very difficult to realize.

An expectation on future capabilities of data analytics is what analyzing more streamed data will bring. Moving toward finding the quality issue before the customer brings it to the workshop is the main foresighted capability. "Things are happening fast, but there still a way to go and barriers along the way". Barriers refers to a notion of the technological but also legal compliance.

4.6 Case III: Aftermarket

The aftermarket has knowledge about the market and are in close contact with the automotive company's customers. Therefore their view on how Big Data can improve life for themselves and the customers are of vital importance. To spark an interest for this question, the project focuses on a product that is in the center of attention right now due the massive increase of electronics in trucks. This product are batteries, and it is a product that obviously is vital to further development of sensors and data streaming.

4.6.1 Data

Today, there are service packages available for truck owners provided by the vendors where data are streamed from vehicles and monitored to inform driver when the components are in need for reparation. The only package where the drivers data are logged and analyzed are the gold package. This however will change and include more packages, the difference between these packages being the price. When not streaming data directly from the vehicle, the data are collected manually in the workshops. When only collected at workshops, the vendors lose the continuity in data collection. Much focus when talking to Aftermarket professionals have been on batteries, and information such as oil levels, topography and State of Health for batteries are collected to determine the life span for the batteries. There is no way to monitor all vehicles due to capacity, they can only check 4-5 vehicles at the time. The service vendor has contact with the automotive company, and they can make wishes regarding what logged data they want in the vehicle. However this kind of development takes time, and they have their own focus as well. Furthermore, they

have to make it work technically as well which is not always easy. However, the company are grateful for input from the market side.

4.6.2 Analytics

Even with today's data levels, there is no way to monitor all data that comes in. This would be preferable of course, but would demand too much personnel for it to be realistic. What is needed is automated analysis done instantly aftermarket professionals says, especially since the levels of data collected are going to increase. Today there are a signal when something is wrong in a truck, but there is no information regarding what it is about. Someone have to log in and manually check (this only applies to gold packages). The value today is that the driver can call and ask what the problem is, but it does not happen automatically.

4.6.3 Opportunities

In the business there is emphasis on developing predictive maintenance, and being able to tell this to the customers before they even notice. It is where the big potential competitive advantages are, being able to tell what action the driver needs to take and in what time-span. In the end, it is all about the money, and standing still is expensive for the customers. Instead of the driver noticing something wrong, often way to late, the determining of the root-cause and the wait for the parts to arrive are costly. With predictive maintenance, it will only be one stop, the fault can be solved earlier and parts can be delivered in time without having to be in stock for longer than necessary.

4.6.4 Barriers

There are multiple barriers to successfully implement Big Data analytics in the aftermarket according to the interviewees. Better algorithms that can determine appropriate actions, as well as doing it automatically. It should be possible to log more data, and doing so in more detail. There needs to be ID's on not only each vehicle, but at all important components in order to really make sense of root-causes. Furthermore, there needs to be better and more robust sensors that does not brake, and if they do, they need to be monitored as well to be able to detect false data recording. Today, much are up to the drivers when it comes to avoid and detect faults.

Another barrier could be personal integrity, but it does not seem to be a problem as things stand. According to one interviewee truck owners signs an umbrella agreement in which the owner of the vehicle has signed the right to collect data. It is the owners responsibility to inform the driver about these agreements, and the driver seldom has a choice but have to accept since the agreement applies to many vehicle owners. This is so central in the industry now and there are no personal data collected, when they talk about ID, the interviewees do not mean ID on the driver but ID on components in the truck. There are also laws that some data needs to be collected, and the vehicle owner obviously needs to follow this.

5
Analysis

The analysis provides insights regarding the results and compares them with previous studies and findings. The analysis is structured so that the current situation is assessed. Then the vision and opportunities for the automotive case company are mentioned, both from employees and people connected to the comapny, Rejmes and from a theoretical standpoint. Lastly the requirements and barriers for going from today into the future vision will be presented. The analysis covers the interviews, study visits and observations, and the data mining experiment conducted in order to learn first hand how Big Data analytics practically can be conducted and where the pitfalls are.

5.1 Data

The data itself and collection methods are central for R&D, Operations and Aftermarket, however to different degrees. Aftermarket is interested in data, but do not want to get too specific regarding how to collect it. For them it is more about application, and how they want the data presented.

For Operations the data are important, and much emphasis is put on traceability and volume of data. It is also of importance where it is gathered and what is needed to make sense of it in terms of connecting data throughout the lifecycle. Operations at the company are often links between departments and since being a large corporation it is often hard to find existing information (Saaksvuori & Immonen, 2005). There is also value for companies to collect information in the middle of the lifecycle so Operations plays a big part in terms of linking the company together (Shin et al., 2009, 1-3). Quality of data are important for Operations, however they also want as much data as possible and then make their own analyses. This is quite different from how R&D likes to work and also different from literature where it is much emphasized that one should know what analyses to do before collecting data (Johanson et al., 2014; Wozniak et al., 2015). One reason for this might be that Operations does not have to think about how the data should be collected technically, something R&D has to do.

R&D is also the division that focuses the most on the collection itself, how to store it and analyze it. This makes sense since it is R&D's task to make all things works, whereas Operations and Aftermarket is more application inclined. R&D stress that the ability to collect data is vital and that the data trend demands more from the data divisions, a fact seen in many businesses (Manyika et al., 2011). R&D are focusing not only on the quantity, but also on the quality of the data.

Many argues that there will have to be standardize ways to collect data, and really knowing what data are collected. This is in line with the learning's from Wozniak et al. (2015)'s study, where investigation of the data is considered vital. Furthermore (Wozniak et al., 2015) says that it is important to know how the data has been collected and maintained to guarantee its quality and relevance. This view is also strengthened after conducted the data mining experiment in this project, where parameters and variables were unknown. This makes it hard to analyze results and find validity. Even if correlations is found it does not mean that they are relevant (Wozniak et al., 2015). This view is strongly agreed upon among R&D professionals and will be discussed further in Analysis.

In terms of storage, R&D argues that the development goes more towards cloud solutions, but it is clear that much data logs at particular moments, such as when the truck are in service workshops or when component quality are logged in production. Different types of data storage are expected to continue and R&D argues that the key will be to get several systems to work together. (Johanson et al., 2014) has a model where data first are stored based on geographic location, and then aggregated into a cloud. Solutions like this might be appropriate for the automotive case company as well.

5.2 Analytics

As the volume of data increases, so does the requirements on the tools and methods for analysis (McAfee & Brynjolfsson, 2012). There seem to be consensus between the interviewees regarding the need for automating some processes for data monitoring and analysis. Specifically for Operations and Aftermarket, who's daily tasks much consists of digging in to product individual data. Aftermarket see the need arise for new processes with the increase of customers connected to their remote maintenance service. Similar needs reside in Operations where the interviewees sees the investigative work as time-consuming and a bad way to utilize themselves as a resource. For Operations, much efficiency could be accomplished by standardizing the data recording process. It would require less knowledge about the circumstances under which it was recorded (Wozniak et al., 2015). Standardizing the recording would increase the veracity of the data and decrease the risk of false correlations and interpretation (Demchenko et al., 2013; White, 2012, 4).

R&D also express concern regarding the risk of deducting false correlations and interpretations. Quality over quantity is preferred and considered more important (Lazer et al., 2014, 6176). Meta data, regarding the data-sets, of interest for all interviewees are product configuration, climate where the product is used or supposed to be used and lastly, how the product is used. There is no need for bigger populaces but rather more dimensions that describes the circumstances and the life-cycle of the product. Expanding dimensions will require more effort into designing data models (Terzi et al., 2007, 3).

Facilitating data-driven decision-making in the automotive is a challenge due to the many dimensions required for obtaining high veracity data (White, 2012, 4). There is a need to fit the statistical model to the format of the data-set (Shin et al., 2009, 1-3), which was realized during the experiment and adopted for instance with

the logistic regression. Furthermore, it is considered a harder task to interpret the results and elicit value from models with many dimensions (Demchenko et al., 2013), also realized during the experiment. Some of the interviewees from R&D suggest that the right technical competence is necessary for interpreting analysis results to elicit value. Knowledge-driven product development certainly require a high level of competence of the analyst (White, 2012, 4; Wozniak et al., 2015). However, there might be value in challenging pre-existing assumptions which was tried in the experiment to a limited extent through the data based experience approach.

5.3 Opportunities

All three division sees large potential with Big Data in automotive, as have experts within the industry before them (Johanson et al., 2014; Wozniak et al., 2015). R&D mentions that the value of each product can increase enormously, since they could provide insight in how products are used and what their status are in terms of quality. This development have been identified as positive and advantageous before (Johanson et al., 2014) and the prospect of knowledge-driven product development where the actual product guides decisions are exciting among the interviewees. Bringing data into the product development process is much focused on by R&D, and the potential for shortening the feedback-loop for how the products perform is huge according to interviewees and PROPID (2015). Another opportunity is predictive maintenance, and this could bring insight and value to all three divisions. The potential for aftermarkets are described to be huge and imperative from a competitive point of view. It can lead to less stops for drivers, less components in storage and better knowledge about product and driver behaviour. In operations product quality can be predicted based on test-values in manufacturing as tested in this projects experiment, and the value and interest in this kind of tests are validated by interviewees at the case company. These kinds of services and solutions have previously been identified in Big Data and automotive (Johanson et al., 2014; Luckow et al., 2015) and there is consensus among the interviewees that this will with certainty increase. With more of the right data, the production and quality processes could be enhanced and improved (Zhang et al., 2016; Luckow et al., 2015). There will be flows of data with the launch of autonomous vehicles (Luckow et al., 2015). There are also large amounts of reports created during product development that may hold some opportunity in terms of data analytics (Arnarsson et al., 2016).

Manyika et al. (2011) talks in their Big Data report about the opportunity to sell the data itself to third parties. This has not been discussed to a large degree in this project, even tough there is potential for this kind of new revenue streams also in the automotive industry (Johanson et al., 2014). This might be because of the characteristics of the interviewees, it might be mentioned more if business developer was interviewed, and this could be interesting to look into in the future.

5.4 Barriers

The majority of the interviewees are concerned about future legislation regarding personal integrity. There is a upcoming EU-directive regarding data collection and personal integrity that will take effect in May of 2018 (Datainspektionen, 2017). However, some of the interviewees brushes of the responsibility and refer to the compliance department. But there is general consensus that security with data collection is an aspect that definitely needs to be considered. It is mentioned that the legislation requires changes from a technological perspective as well. According to an interviewee in R&D, will the legislation demand that users should be able to delete personal related data-sets. This means that all data needs to be traceable.

Shifting towards more data-driven decision-making requires new and added competence in the organization (LaValle et al., 2011, 2; McAfee & Brynjolfsson, 2012). The interviewees are aware of this, Operations interviewees suggest new tools for facilitating the change. Some from R&D suggest automation of the analytics and make it centralized and cloud-based. This presents a major challenge for the case company but is an opportunity for third parties. However, the conception still is that the developer or appropriate stakeholder should interpret the results even of advanced analytics.

5.5 Data Mining Experiment

There is definitely potential in cross-lifecycle analysis to improve product quality (Luckow et al., 2015; Zhang et al., 2016), as was aimed for with quality case 1 in the experiment. In the experiment it was also realized that traceability plays a large role for such analytics as it was the main failing component of quality case 1. However, the state of the backup harddrive and other parameters might be the cause of not finding the product individuals from quality case 1. The quality case that generated result is however not irrelevant. Operations stated that it is of interest to pursue predictive analysis and machine learning methods for improving quality, which coincides with efforts in other industries (Zhang et al., 2016). Such pursuits should however be executed with a different approach. Firstly define the aim with the analysis, then ensure the data collection and recording method and lastly perform the statistical analysis (Shin et al., 2009, 1-3).

The logistic regression analysis was considered good and easy to understand (Joseph F. Hair et al., 2006). Studying the relationships between the variables in the plots, there seems to be clusters i the X-axis in most of the results, see Appendix D and E. As the data-sets are classified after product family, perhaps the clusters are specific product types within the family. There is no meta data in the database to make that distinction even though it seems likely. If there were a way to make the distinction between specific product types, it could lead to a different result.

The outlier analysis was considered even less fruitful. In hindsight it seems likely that a nominal value has tolerance limits and is thereby already monitored. However, the analysis could point towards patterns in the data if a certain distance from the global mean was attained for a sequence of nominal values. This would

mean that it is not a specific parameter that causes a function failing but rather a synergistic effect between a range of parameters. The sample sizes in the experiment showed however no such result.

6

Discussion

The experiment conducted in this project has generated some important insights. It was found that the 5V definition of Big Data to a large degree explained the challenges and opportunities regarding Big Data analytics. As depicted in Figure 6.1, there are certain things to consider before engaging in analytics. As mentioned, the volume and variety of the data-set determines the statistical model that should be used (Shin et al., 2009, 1-3) and the aim of the analysis in terms of value and velocity decides which application e.g. machine learning, data mining.

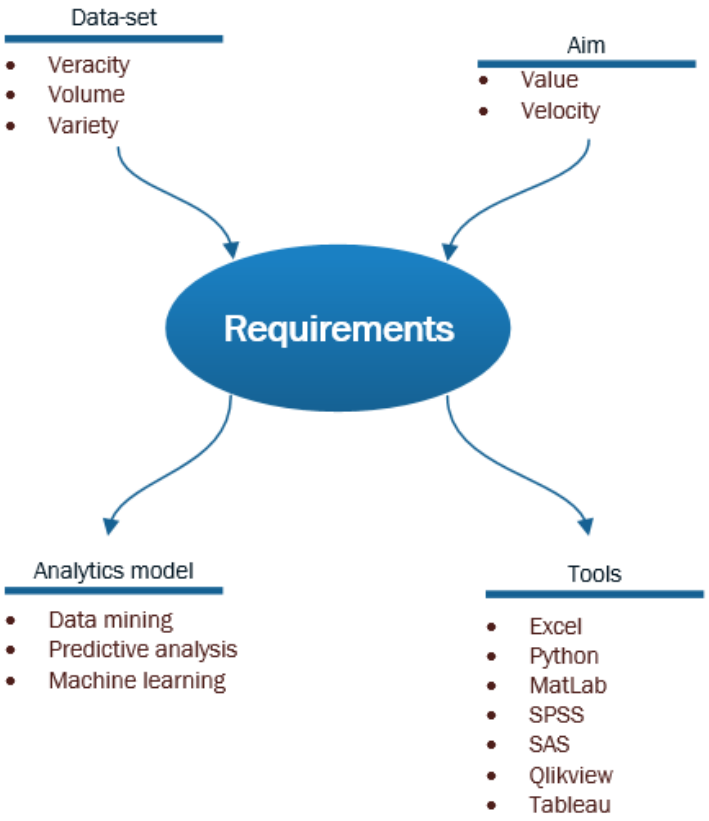


Figure 6.1: Big Data analytics requirements

It was elicited from the interviews that tools is often chosen after personal preference. The functionality regarding simpler analytics are considered to be fairly equal making it a less important aspect. However, some of the tools requires a long learning period and are required for realizing some analytics which was realized

during the experiment. Data science competence seems to be an elusive resource in the organization. To tap into the wave of opportunities coming with Big Data, that type of competence should be acquired and developed in the organization (LaValle et al., 2011, 2; McAfee & Brynjolfsson, 2012).

Furthermore should the veracity be considered to a high degree before trying to elicit value from the analysis. Especially in the automotive industry, it seems like the product is so complex that it requires many dimensions of meta data to ensure the value of the analysis. Through the interviews and observations it can be concluded that many correlations are circumstantial and that there is a need to cope with high dimension to enable value-adding analytics. The way that it was coped with in the experiment was to classify the data-set into smaller data-sets with the same defining meta data. However, if that is done to a larger degree then the value also might become insignificant and only applicable to very specific situations.

6.1 Research Impact

With increased data analytics, more and more knowledge will be acquired regarding products and processes. It has been disregarded in this project, however, knowledge management will be highly relevant to consider in the future within the same scope. As analytics are automated and centralized in the cloud, the knowledge and value elicited, needs to be presented to the relevant stakeholders. Transforming it in to actionable information is a future challenge and will require even more PLM research. Things to consider regarding knowledge management is the integration between the systems, visualization and presentation of the value elicited from the analytics and distribution of knowledge.

Large parts of all collected data are in the form of text. The interviewees seem to refrain from analyzing that type of data because of low veracity but also because there is a lack of tools for text analysis. During the experiment and the scouring of the backup of the database it was discovered that especially in production data there was much data recorded as text strings. Next step for the aim of challenging pre-existing assumptions could be to explore text analysis to a much greater degree. However, there should be some control over and insight regarding the circumstances of the recording which was lacking in this experiment.

Another next step to research cross lifecycle analytics could be to build and try a logistic regression model for vehicles in the field and production data. Fault flags from the field as the dependent variable together with carefully selected variables from production could bring deeper insight into tolerancing and manufacturing processes impact on quality. Including the learnings from this project, should the data-set be selected after similar meta data such as similar product configuration, same market and same time period to elicit value and cope with the risk of false correlations.

If such analytics prove fruitful, then it should be considered another type of integration between the MES and the rest of product individual related data storage solutions. The high demand on automating processes would be greatly facilitated if a heavier type of integration would be adopted. There are however barriers to this. Even though partnerships might be strong, there could be a need for developing

further policies regarding ownership and distribution of data.

6.2 Societal Impact

Big Data analytics is a technology concept that is certainly here to stay and to be further developed. The extensive list of potentially value-adding and disruptive applications are long. The most elusive, yet most striven for application is predicitive analytics. The application has the potential within the realm of the automotive industry to maximize the value of each resource and product leading to a more sustainable society and more lean and efficient businesses. Supply chain management and logistics is a common denominator as important function across industries that will be greatly impacted by Big Data analytics. Optimizing logistics has inherently a good effect on the environment.

Big Data analytics is also playing a large role for the new technology of autonomous vehicles. The technology will thereby have an immense positive impact on society as autonomous vehicles are thought to minimize casualties in traffic. There are also technologies being developed that enables optimization of fuel consumption using new streams of data such as GPS-locations.

Even though there are barriers and resistance regarding data collection as an intrusion of integrity, it also infuses transparency. Adapting to the change certainly needs deep consideration for laws and regulations to satisfy and protect all stakeholders. However, all data collection does not intrude on integrity and the trend will go mainstream, even though new hurdles continuously arise.

7
Conclusion

This project explores how the case company can use Big Data in product development and create opportunities by answering the questions:

- What are the barriers and opportunities for knowledge-driven product development across the lifecycle using Big Data analytics?
- What new services can come from Big Data analytics and what are the barriers for taking advantage of these?

The two main themes of knowledge-driven product development discussed in this projects is Data and Analytics. These two themes are closely linked and the methods that can be used to bring insight is directly connected to which data is available. The opportunities with knowledge-driven product development is that actual data on product and user behavior can be used to make decisions on product design. Products can then be adapted based on differences in usage depending on e.g. the market it is on. Furthermore, the feedback loop gets shorter since there is no need to wait for a component to break, signs of this can be seen live and way before an actual breakage. It is also possible to update software via cloud, something that otherwise has to be done manually. Barrier for Big Data is the data collection, which R&D emphasize. Storage is not seen as the biggest issue, there are many companies that have to store more data than an automotive company. A bigger barrier is getting the right data in terms of quality and avoiding false correlations. There will have to be personnel with the right skills, traceability in data to identify root-causes, hardware that can handle the requirements of a moving vehicle and software able to present the analysis in an easy and understandable way.

It has been considered a fact that data-driven decision-making enhances the general performance of a company. In this report are several opportunities and barriers with adopting the concept discussed. The great opportunity that many interviewees in this study includes in their vision is systems that not only presents data, but automatically presents suggestions on actions to take. In aftermarket, near the end of the lifecycle, this could mean less storage, fever stops for the driver leading to reduced costs of being on the road, and better knowledge of when and why errors occur. The same can be seen in Operations, where quality issues can be predicted based on specific test-values in production and stopped much earlier in production.

Barriers for Big Data applications are legislation's regarding personal integrity, and that data can be collected legally. Also traceablity between company divisions and capabilities within the company are barriers in a shift towards Big Data. Through the PLM perspective of this report, many different stakeholders have been interviewed. Although many have a shared vision of what Big Data can become for

the automotive company, there are some contrasting views on what horse to bet on in terms of data collection and analytic strategy. Getting all stakeholders going the same way will be a challenging and but vital task in the shift towards data-driven decision-making in the company and in the automotive industry in general.

7.1 Recommendations

The following recommendations are made for the case company based on the findings on this report. As discussed in Scope 1.4, these findings are representative only of the roles interviewed and the the data that was made available to the project.

- Know the data, where it is collected and who is responsible for it. This will make for faster tracing its origins when analyzing it.
- Put effort into automating data analytics processes to retain value more efficiently.
- Acquire data science competence in places with high strategic significance, i.e. product development and product planning and strategies departments in order to facilitate data-driven decision-making.
- Quality over quantity. Instead of collecting all data, have a flexible solution which easily can change what it collects. Makes for less storage need and faster analytics.
- Consider the integration of systems and access of data throughout the lifecycle of the product and aim for traceability.
- Conduct cluster analysis on text-based data from the MES and thereby elicit value through knowledge regarding production processes.
- Standardize data collection to verify the veracity of the data wherever possible.
- Test cross-lifecycle analytics further with DTC's (Diagnostic Trouble Code) from vehicles in the field and carefully selected parameters from production.
- Increase quality of sensors so that they can withstand the requirements of being in a moving vehicle.
- Provide tools for managing the increase in customers with remote maintenance deals.

Bibliography

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*.

van Rossum, G. & Fred L. Drake, J. (2001). *Python reference manual*. Python Labs.

Saaksvuori, A. & Immonen, A. (2005). *Product lifecycle management, second edition*. Springer.

Witten, I. H. & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Elsevier.

Joseph F. Hair, J., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham, R. L. (2006). *Multivariate data analysis, sixth edition*. Pearson Prentice Hall.

Terzi, S., Panetto, H., Morel, G. & Garetti, M. (2007). A holonic metamodel for product traceability in product lifecycle management. *International Journal of Product Lifecycle Management*, *2*.

Schulte, S. (2008). Customer centric plm: Integrating customers' feedback into product data and lifecycle processes. *International Journal of Product Lifecycle Management*, *3*.

Shin, J.-H., Jun, H.-B., Kiritsis, D. & Xirouchakis, P. (2009, January). Function performance evaluation and its application for design modification based on product usage data. *International Journal of Product Lifecycle Management*, *4*, 84–113.

Terzi, S., Bouras, A., Dutta, D., Garetti, M. & Kiritsis, D. (2010). Product lifecycle management – from its history to its new role. *International Journal of Product Lifecycle Management*, *4*.

Gorunescu, F. (2011). *Data mining: Concepts, models and techniques*. Springer.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S. & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management review*, *52*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011, May). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute, MGI*.

Zikopoulus, P. C., Eaton, C., deRoos, D., Deutsch, T. & Lapis, G. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw Hill.

Boyd, D. & Crawford, K. (2012, June). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication society*, *15*, p.662–679.

Chen, H., Chiang, R. H. L. & Storey, V. C. (2012, December). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*, p.1165–1168.

Han, J., Kamber, M. & Pei, J. (2012). *Data mining: Concepts and techniques, 3rd ed.* Elsevier.

McAfee, A. & Brynjolfsson, E. (2012, October). Big data: The management revolution. *Harvard Business Review.*

White, M. (2012). Digital workspaces: Vision and reality. *Business information review, 29.*

Demchenko, Y., Grosso, P., de Laat, C. & Membrey, P. (2013, May). Addressing big data issues in scientific data infrastructure. *IEEE.*

Johanson, M., Belenki, S., Jalminger, J., Fant, M. & Gjertz, M. (2014, October). Big automotive data: Leveraging large volumes of data for knowledge-driven product development. *IEEE*, p.736–741.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014, March). The parable of google flu: Traps in big data analysis. *American Association for the Advancement of Science, 343*, p.1203–1205.

Li, J., Tao, F., Cheng, Y. & Zhao, L. (2015, May). Big data in product lifecycle management. *Springer-Verlag London.*

Luckow, A., Kennedy, K., Manhardt, F., Djerekarov, E., Vorster, B. & Apon, A. (2015, October). Automotive big data: Applications, workloads and infrastructures. *IEEE*, p.1201–1210.

Palade, H. C., Nicolaescu, S. S. & Kifor, C. V. (2015, May). The impact of big data and knowledge management on r&d projects from automotive industry. *Springer International Publishing Switzerland.*

PROPID. (2015, December). Ansokan inom big automtive analytics (bada). *Vinnova.*

Stark, J. (2015). Product lifecycle management. *Springer International Publishing, 1.*

Wozniak, P., Valton, R. & Fjeld, M. (2015). Volvo single view of vehicle : Building a big data service from scratch in the automotive industry. *ACM.*

Arnarsson, I. Ö., Malmqvist, J., Gustavsson, E. & Jirstrand, M. (2016). Towards big-data analysis of deviation and error reports in product development projects. *Chalmers University of Technology.*

Gartner. (2016). 2017: The date that data and analytics go mainstream. *Gartner.* Retrieved February 24, 2017, from http://www.gartner.com/smarterwithgartner/2017-the-date-that-data-and-analytics-go-mainstream/

IBM. (2016). Homepage. Retrieved February 15, 2017, from http://www.ibm.com/analytics/se/sv/technology/spss/

Zhang, D., Xu, B. & Wood, J. (2016). Predict failures in production lines: A two-stage approach with clustering and supervised learning. *IEEE.*

Datainspektionen. (2017). Homepage. Retrieved May 3, 2017, from http://www.datainspektionen.se/lagar-och-regler/eus-dataskyddsreform/

Rejmes. (2017). Homepage. Retrieved May 8, 2017, from http://www.dealer.volvotrucks.se/tage-rejmes/our-depots.html

A

Appendix

The script used for splitting large data-sets.

```
# ---- C-script for splitting files ---- #

var list = new List<string>();
var fileSuffix = 0;

using (var file =
File.OpenRead(@"C:\Users\Johan\Desktop\Python\CG051\CG051_split\CG051_formatted.cs
v"))
using (var reader = new StreamReader(file))
{
    while (!reader.EndOfStream)
    {
        list.Add(reader.ReadLine());

        if (list.Count >= 500000)
        {

File.WriteAllLines(@"C:\Users\Johan\Desktop\Python\CG051\CG051_split\split" +
(++fileSuffix) + ".csv", list);
            list = new List<string>();
        }
    }
}

File.WriteAllLines(@"C:\Users\Johan\Desktop\Python\CG051\CG051_split\split" +
(++fileSuffix) + ".csv", list);
```

B

Appendix

The code written for the classification step of the Data mining process.

```
# --------- Data Extraction Level 1 -------- #

import csv                                          # import python csv-module

f1 = open('SDE_Rejcom_one_row.csv', "r")            # open quality file
f2 = open('CG063_formatted.csv', "r")               # open Big File
f3 = open('case2_CG063.csv', "w")                   # open new output file


csv_f1 = csv.reader(f1, delimiter='\t')             # create read/write objects for the files
csv_f2 = csv.reader(f2, delimiter=',')
csv_f3 = csv.writer(f3, lineterminator='\n')

# ---------- Main ---------- #

masterlist = list(csv_f2)                           # Save big file as list object

for row1 in csv_f1:                                 # iterate through quality file
   for row2 in masterlist:                          # iterate through Big file
      if row1[1] == row2[1]:                        # if indexes pointing at same id on rows
         csv_f3.writerow(row2)                      # Print Big file row to outputfile


# ---------- EOS ---------- #

f1.close()                                          # close files
f2.close()
f3.close()
```

```
# ---------- Data classification level 2 ---------- #

import csv

f1 = open('CG063_formatted.csv', "r")                    # open Big file
f2 = open('LSP1_NV_BSFC.csv', "w")                       # create new unique test files
f3 = open('LSP1_NV_FlwFuel.csv', "w")
f4 = open('LSP1_NV_PCAC.csv', "w")
f5 = open('LSP1_NV_PwrEng.csv', "w")
f6 = open('LSP3_NV_BSFC.csv', "w")
f7 = open('LSP3_NV_FlwFuel.csv', "w")
f8 = open('LSP3_NV_PCAC.csv', "w")
f9 = open('LSP3_NV_PwrEng.csv', "w")
f10 = open('LST2_NV_BSFC.csv', "w")
f11 = open('LST2_NV_FlwFuel.csv', "w")
f12 = open('LSP3_NV_TrqEng.csv', "w")

csv_f1 = csv.reader(f1, delimiter=',')                   # Create reader object for big file
csv_f2 = csv.writer(f2, lineterminator='\n')             # Create writer objects for the unque test
csv_f3 = csv.writer(f3, lineterminator='\n')             # files
csv_f4 = csv.writer(f4, lineterminator='\n')
csv_f5 = csv.writer(f5, lineterminator='\n')
csv_f6 = csv.writer(f6, lineterminator='\n')
csv_f7 = csv.writer(f7, lineterminator='\n')
csv_f8 = csv.writer(f8, lineterminator='\n')
csv_f9 = csv.writer(f9, lineterminator='\n')
csv_f10 = csv.writer(f10, lineterminator='\n')
csv_f11 = csv.writer(f11, lineterminator='\n')
csv_f12 = csv.writer(f12, lineterminator='\n')

masterlist = list(csv_f1)                                # Create list object for big file

# ------------ Main ------------ #

for row1 in masterlist:                                  # iterate over big file
   if row1[4] =='LSP1_NV_BSFC' and row1[5] =='11':       # Determine classifiers for the row
      csv_f2.writerow(row1)                              # place in unique test file
   if row1[4] =='LSP1_NV_FlwFuel' and row1[5] =='11':
      csv_f3.writerow(row1)
   if row1[4] =='LSP1_NV_PCAC' and row1[5] =='11':
      csv_f4.writerow(row1)
   if row1[4] =='LSP1_NV_PwrEng' and row1[5] =='11':
      csv_f5.writerow(row1)
   if row1[4] =='LSP3_NV_BSFC' and row1[5] =='11':
      csv_f6.writerow(row1)
   if row1[4] =='LSP1_NV_FlwFuel' and row1[5] =='11':
      csv_f7.writerow(row1)
   if row1[4] =='LSP1_NV_PCAC' and row1[5] =='11':
```

```
        csv_f8.writerow(row1)
    if row1[4] =='LSP1_NV_PwrEng' and row1[5] =='11':
        csv_f9.writerow(row1)
    if row1[4] =='LST2_NV_BSFC' and row1[5] =='11':
        csv_f10.writerow(row1)
    if row1[4] =='LST2_NV_FlwFuel' and row1[5] =='11':
        csv_f11.writerow(row1)
    if row1[4] =='LST2_NV_TrqEng' and row1[5] =='11':
        csv_f12.writerow(row1)

# ----------- EOF ----------- #

f1.close() # mste stnga loopen
f2.close()
f3.close()
f4.close()
f5.close()
f6.close()
f7.close()
f8.close()
f9.close()
f10.close()
f11.close()
f12.close()
```

```
# ---------- Data Classification level 6 ---------- #

import csv

f1 = open('SDE_Rejcom_one_row.csv', "rt")          # open quality file
f2 = open('MasterMasterClean.csv', "r")            # Master file
f3 = open('CG063_Rejcom.csv', "w")                 # create output file

csv_f1 = csv.reader(f1, delimiter='\t')            # Create reader/writer objects
csv_f2 = csv.reader(f2, delimiter=',')
csv_f3 = csv.writer(f3, lineterminator='\n')

# ---------- Main ---------- #

masterlist = list(csv_f2)                          # save Master file to list object

for row1 in csv_f1:                                # iterate over quality file
   for row2 in masterlist:                         # iterate over Master file
      if row1[0] == row2[0]:                       # comparing id's on rows
         row2.extend([row1[1]])                    # if same, add classifier from quality file to
                                                   # Master file object

for row in masterlist:                             # Print Master file object to outputfile
   csv_f3.writerow(row)

f1.close()                                         # close files
f2.close()
f3.close()
```

C

Appendix

The code written for preparing data-sets for statistical analysis.

```
# ---------- Data Aggregation Level 3 ----------- #


import csv

f1 = open('Masterlist.csv', "r")                    # Open Masterlist containing quality case
f2 = open('LSP1_NV_BSFC.csv', "r")                  # id's
f3 = open('LSP1_NV_FlwFuel.csv', "r")               # rest of the file's contain unique test
f4 = open('LSP1_NV_PCAC.csv', "r")                  # values per quality case id
f5 = open('LSP1_NV_PwrEng.csv', "r")
f6 = open('LSP3_NV_BSFC.csv', "r")
f7 = open('LSP3_NV_FlwFuel.csv', "r")
f8 = open('LSP3_NV_PCAC.csv', "r")
f9 = open('LSP3_NV_PwrEng.csv', "r")
f10 = open('LST2_NV_BSFC.csv', "r")
f11 = open('LST2_NV_FlwFuel.csv', "r")
f12 = open('LSP3_NV_TrqEng.csv', "r")
f13 = open('MasterClean.csv', "w")

csv_f1 = csv.reader(f1, delimiter=',')              # Create reader objects
csv_f2 = csv.reader(f2, delimiter=',')
csv_f3 = csv.reader(f3, delimiter=',')
csv_f4 = csv.reader(f4, delimiter=',')
csv_f5 = csv.reader(f5, delimiter=',')
csv_f6 = csv.reader(f6, delimiter=',')
csv_f7 = csv.reader(f7, delimiter=',')
csv_f8 = csv.reader(f8, delimiter=',')
csv_f9 = csv.reader(f9, delimiter=',')
csv_f10 = csv.reader(f10, delimiter=',')
csv_f11 = csv.reader(f11, delimiter=',')
csv_f12 = csv.reader(f12, delimiter=',')
csv_f13 = csv.writer(f13, lineterminator='\n')      # writer object, creater of output file

MasterList = list(csv_f1)                           # create list object containing the quality
                                                    # case id's

# -------- Duplicate removal -------- #

DataList1 = []
for line in csv_f2:                                 # iterate over LSP1_NV_BSFC test file
   if line in DataList1 or line[2] != '1':          # skip duplicate rows or test nr>1
      continue
   else:
      DataList1.append( line )                      # save unique row to list object

DataList2 = []                                      # exactly as above for the other unique
for line in csv_f3:                                 # test files
   if line in DataList2 or line[2] != '1':
```

```
        continue
    else:
        DataList2.append( line )

DataList3 = []
for line in csv_f4:
    if line in DataList3 or line[2] != '1':
        continue
    else:
        DataList3.append( line )

DataList4 = []
for line in csv_f5:
    if line in DataList4 or line[2] != '1':
        continue
    else:
        DataList4.append( line )

DataList5 = []
for line in csv_f6:
    if line in DataList5 or line[2] != '1':
        continue
    else:
        DataList5.append( line )

DataList6 = []
for line in csv_f7:
    if line in DataList6 or line[2] != '1':
        continue
    else:
        DataList6.append( line )

DataList7 = []
for line in csv_f8:
    if line in DataList7 or line[2] != '1':
        continue
    else:
        DataList7.append( line )

DataList8 = []
for line in csv_f9:
    if line in DataList8 or line[2] != '1':
        continue
    else:
        DataList8.append( line )

DataList9 = []
for line in csv_f10:
```

```
    if line in DataList9 or line[2] != '1':
        continue
    else:
        DataList9.append( line )

DataList10 = []
for line in csv_f11:
    if line in DataList10 or line[2] != '1':
        continue
    else:
        DataList10.append( line )

DataList11 = []
for line in csv_f12:
    if line in DataList11 or line[2] != '1':
        continue
    else:
        DataList11.append( line )

# --------- Main ---------- #


for row1 in MasterList:                          # iterate over list of quality case id's
    for row2 in DataList1:                       # locate quality case id in unique test
        if row1[0] == row2[1]:
            row1.extend([row2[6]])               # extend the row with matching id
    if len(row1) < 2 :                           # with the test value
        row1.extend(['0'])                       # if no test value for id extend with "0"

for row12 in MasterList:                         # same as above for all unique test files
    for row3 in DataList2:
        if row12[0] == row3[1]:
            row12.extend([row3[6]])
    if len(row12) < 3 :
        row12.extend(['0'])

for row13 in MasterList:
    for row4 in DataList3:
        if row13[0] == row4[1]:
            row13.extend([row4[6]])
    if len(row13) < 4 :
        row13.extend(['0'])

for row14 in MasterList:
    for row5 in DataList4:
        if row14[0] == row5[1]:
            row14.extend([row5[6]])
    if len(row14) < 5 :
```

```
        row14.extend(['0'])

for row15 in MasterList:
    for row6 in DataList5:
        if row15[0] == row6[1]:
            row15.extend([row6[6]])
    if len(row15) < 6 :
        row15.extend(['0'])

for row16 in MasterList:
    for row7 in DataList6:
        if row16[0] == row7[1]:
            row16.extend([row7[6]])
    if len(row16) < 7 :
        row16.extend(['0'])

for row17 in MasterList:
    for row8 in DataList7:
        if row17[0] == row8[1]:
            row17.extend([row8[6]])
    if len(row17) < 8 :
        row17.extend(['0'])

for row18 in MasterList:
    for row9 in DataList8:
        if row18[0] == row9[1]:
            row18.extend([row9[6]])
    if len(row18) < 9 :
        row18.extend(['0'])

for row19 in MasterList:
    for row10 in DataList9:
        if row19[0] == row10[1]:
            row19.extend([row10[6]])
    if len(row19) < 10 :
        row19.extend(['0'])

for row20 in MasterList:
    for row11 in DataList10:
        if row20[0] == row11[1]:
            row20.extend([row11[6]])
    if len(row20) < 11 :
        row20.extend(['0'])

for row21 in MasterList:
    for row12 in DataList11:
        if row21[0] == row12[1]:
            row21.extend([row12[6]])
```

```
    if len(row21) < 12 :
        row21.extend(['0'])

# -------- Print to file -------- #

i = 6;

csv_f13.writerow(['LOPNUMMER', str(i)+'_LSP1_NV_BSFC', str(i)+'_LSP1_NV_FlwFuel',
str(i)+'_LSP1_NV_PCAC', str(i)+'_LSP1_NV_PwrEng', str(i)+'_LSP3_NV_BSFC',
str(i)+'_LSP3_NV_FlwFuel', str(i)+'_LSP3_NV_PCAC', str(i)+'_LSP3_NV_PwrEng',
str(i)+'_LST2_NV_BSFC', str(i)+'_LST2_NV_FlwFuel', str(i)+'_LST2_NV_TrqEng'])

for line in MasterList:
    csv_f13.writerow(line)

# --------- Comment field ----------#
# Above code creates headers for the MasterClean that will contain unique quality id's with all
# respective test values. The ouput file contains 103 rows with unique quality id's with 50+ columns
# with only test values
# --------- End comment field -------- #

f1.close() # mste stnga loopen
f2.close()
f3.close()
f4.close()
f5.close()
f6.close()
f7.close()
f8.close()
f9.close()
f10.close()
f11.close()
f12.close()
f13.close()
```

```
# --------- Data Aggregation level 5 --------- #

import csv

f2 = open('LSP1_NV_BSFC.csv', "r")                    # open all unique test value files
f3 = open('LSP1_NV_FlwFuel.csv', "r").
f4 = open('LSP1_NV_PCAC.csv', "r")
f5 = open('LSP1_NV_PwrEng.csv', "r")
f6 = open('LSP3_NV_BSFC.csv', "r")
f7 = open('LSP3_NV_FlwFuel.csv', "r")
f8 = open('LSP3_NV_PCAC.csv', "r")
f9 = open('LSP3_NV_PwrEng.csv', "r")
f10 = open('LST2_NV_BSFC.csv', "r")
f11 = open('LST2_NV_FlwFuel.csv', "r")
f12 = open('LSP3_NV_TrqEng.csv', "r")
f13 = open('Mean_agg.csv', "w")                       # create an output file

csv_f2 = csv.reader(f2, delimiter=',')                # create write and reader objects for the
csv_f3 = csv.reader(f3, delimiter=',')                # files
csv_f4 = csv.reader(f4, delimiter=',')
csv_f5 = csv.reader(f5, delimiter=',')
csv_f6 = csv.reader(f6, delimiter=',')
csv_f7 = csv.reader(f7, delimiter=',')
csv_f8 = csv.reader(f8, delimiter=',')
csv_f9 = csv.reader(f9, delimiter=',')
csv_f10 = csv.reader(f10, delimiter=',')
csv_f11 = csv.reader(f11, delimiter=',')
csv_f12 = csv.reader(f12, delimiter=',')
csv_f13 = csv.writer(f13, lineterminator='\n')

# -------- Main -------- #

mean_total=[]                                         # Create list object to place mean values
 # ------- File 1 -------- #
mean_data_1=[]                                        # create list object place unique test values
for line in csv_f2:                                   # iterate over file
   mean_data_1.append(float(line[6]))                 # save test value and change datatype
if mean_data_1 is none:
   mean_total.append('0')                             # if the file is empty denote missing value
else:                                                 # with ‘0’
   mean_total.append(sum(mean_data_1)/len(mean_data_1))          # calculate mean and place
                                                                 # in mean_total object

# ------- File 2 -------- #
mean_data_2=[]
for line in csv_f3:
   mean_data_2.append(float(line[6]))
if mean_data_2 is none:
```

```
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_2)/len(mean_data_2))


# ------- File 3 -------- #
mean_data_3=[]
for line in csv_f4:
    mean_data_3.append(float(line[6]))
if mean_data_3 is none:
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_3)/len(mean_data_3))


# ------- File 4 -------- #
mean_data_4=[]
for line in csv_f5:
    mean_data_4.append(float(line[6]))
if mean_data_4 is none:
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_4)/len(mean_data_4))


# ------- File 5 -------- #
mean_data_5=[]
for line in csv_f6:
    mean_data_5.append(float(line[6]))
if mean_data_5 is none:
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_5)/len(mean_data_5))


# ------- File 6 -------- #
mean_data_6=[]
for line in csv_f7:
    mean_data_6.append(float(line[6]))
if mean_data_6 is none:
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_6)/len(mean_data_6))


# ------- File 7 -------- #
mean_data_7=[]
for line in csv_f8:
```

```
    mean_data_7.append(float(line[6]))
if mean_data_7 is none:
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_7)/len(mean_data_7))


# ------- File 8 -------- #
mean_data_8=[]
for line in csv_f9:
    mean_data_8.append(float(line[6]))
if mean_data_8 is none:
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_8)/len(mean_data_8))


# ------- File 9 -------- #
mean_data_9=[]
for line in csv_f10:
    mean_data_9.append(float(line[6]))
if mean_data_9 is none:
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_9)/len(mean_data_9))


# ------- File 10 ------- #
mean_data_10=[]
for line in csv_f11:
    mean_data_10.append(float(line[6]))
if mean_data_10 is none:
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_10)/len(mean_data_10))

# ------- File 11 ------- #
mean_data_11=[]
for line in csv_f12:
    mean_data_11.append(float(line[6]))
if mean_data_11 is none:
    mean_total.append('0')
else:
    mean_total.append(sum(mean_data_11)/len(mean_data_11))

# --------- Output File ----------- #

i = 6;
```

```
csv_f13.writerow([str(i)+'_LSP1_NV_BSFC', str(i)+'_LSP1_NV_FlwFuel', str(i)+'_LSP1_NV_PCAC',
str(i)+'_LSP1_NV_PwrEng', str(i)+'_LSP3_NV_BSFC', str(i)+'_LSP3_NV_FlwFuel',
str(i)+'_LSP3_NV_PCAC', str(i)+'_LSP3_NV_PwrEng', str(i)+'_LST2_NV_BSFC',
str(i)+'_LST2_NV_FlwFuel', str(i)+'_LST2_NV_TrqEng'])

csv_f13.writerow(mean_total)

# ------------ Comment field ----------- #
# above creates headers and the appropriate mean values are printed in a row below
# ------------ End comment field ------------ #


# ------- EOS -------#

f2.close()                                    # close the files
f3.close()
f4.close()
f5.close()
f6.close()
f7.close()
f8.close()
f9.close()
f10.close()
f11.close()
f12.close()
f13.close()
```

```
# --------- Fault code extraction ---------- #
import csv

f2 = open('CG024_formatted.csv', "r", newline='\n')
f3 = open('Orsaksnr_52.csv', "w")



csv_f2 = csv.reader(f2, delimiter=',', quotechar=' ') # csv-fil delimiter
csv_f3 = csv.writer(f3, lineterminator='\n')

for row in csv_f2:
   if row[4] == '21' or row[4] == '22':
      csv_f3.writerow(row)



f2.close()
f3.close()
```

```
# -------- Data preparation Logistic regression --------- #

import csv

p1 = open('LSP1_NV_PwrEng.csv', "r")
p2 = open('LSP3_NV_PwrEng.csv', "r")
f3 = open('Orsaksnr_22_21.csv', "r", newline='\n')
f4 = open('21_22_1_PwrEng.csv', "w")
f5 = open('21_22_3_PwrEng.csv', "w")

csv_p1 = csv.reader(p1, delimiter=',')
csv_p2 = csv.reader(p2, delimiter=',')
csv_f3 = csv.reader(f3, delimiter=',', quotechar=' ')
csv_f4 = csv.writer(f4, lineterminator='\n')
csv_f5 = csv.writer(f5, lineterminator='\n')

helplist = []

for row in csv_p1:                                  # Extract all wanted data from infile to
   helplist.append(row[1])                          # to local variable
   helplist.append(row[6])
   for line in csv_f3:
      if line[1] == row[1]:
         helplist.append('1')
   if len(helplist) < 3:
      helplist.append('0')
   csv_f4.writerow(helplist)                        # Print to outfile
   helplist.clear()

helplist.clear()                                    # clear local variable

for row in csv_p2:                                  # Same procedure for next infile
   helplist.append(row[1])
   helplist.append(row[6])
   for line in csv_f3:
      if line[1] == row[1]:
         helplist.append('1')
   if len(helplist) < 3:
      helplist.append('0')
   csv_f5.writerow(helplist)
   helplist.clear()

p1.close()
p2.close()
f3.close()
f4.close()
f5.close()
```

D
Appendix

The Logistic Regression scatter-plots with leakage as dependent variable.

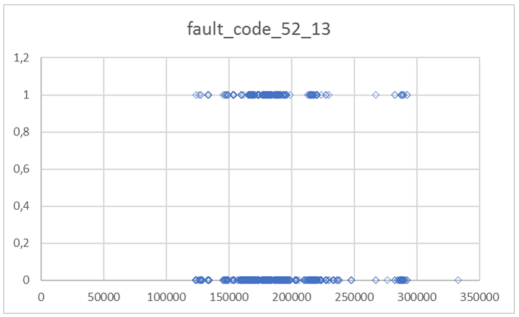


Figure D.1: Leaky code, pressure test 1 and product type 13

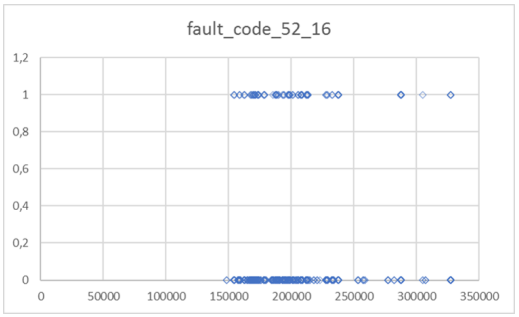


Figure D.2: Leaky code, pressure test 1 and product type 16

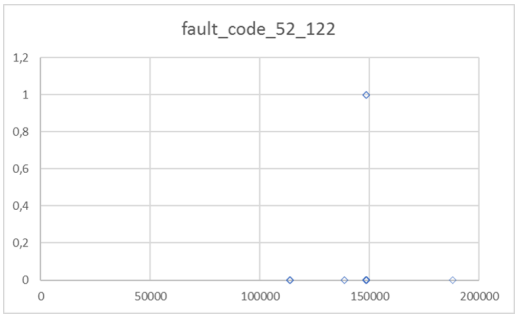


Figure D.3: Leaky code, pressure test 1 and product type 122

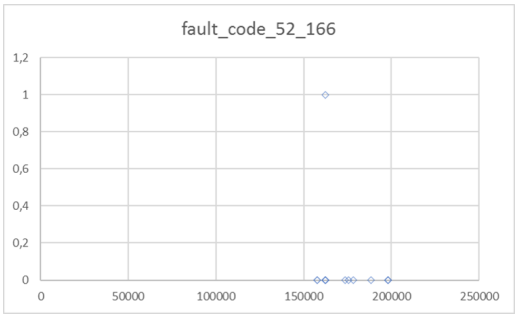


Figure D.4: Leaky code, pressure test 1 and product type 166

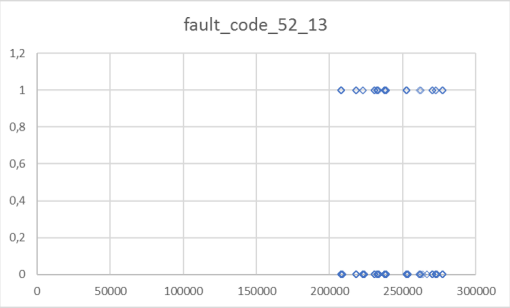


Figure D.5: Leaky code, pressure test 2 and product type 13

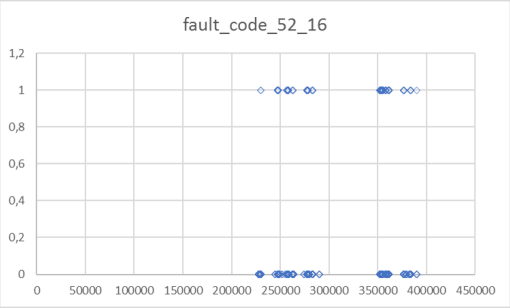


Figure D.6: Leaky code, pressure test 2 and product type 16

E

Appendix

The Logistic Regression scatter-plots with pores and cracks as dependent variable.

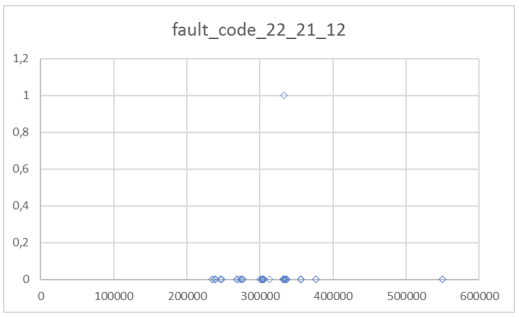


Figure E.1: Pores/Cracks, power test 1 and product type 12

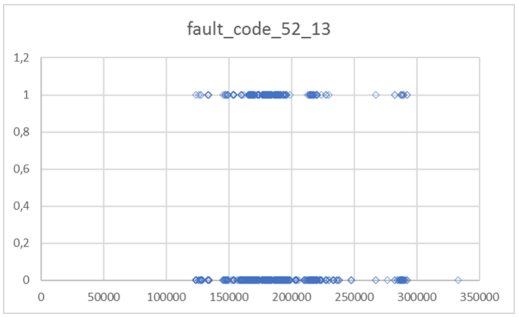


Figure E.2: Pores/Cracks, power test 1 and product type 13

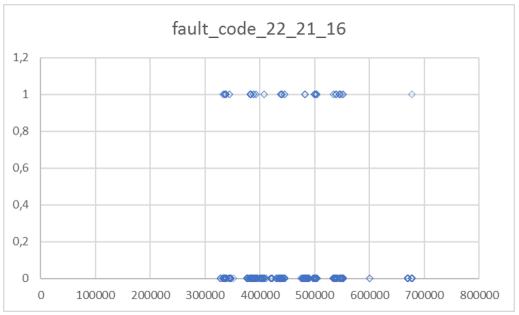


Figure E.3: Pores/Cracks, power test 1 and product type 16

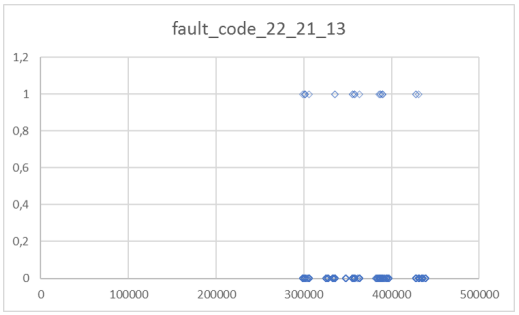


Figure E.4: Pores/Cracks, power test 2 and product type 13

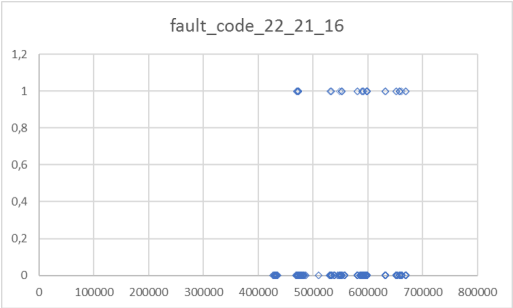


Figure E.5: Pores/Cracks, power test 2 and product type 16

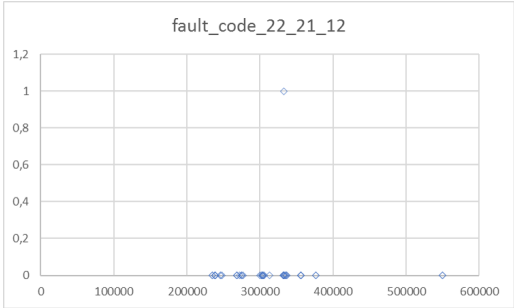


Figure E.6: Pores/Cracks, power test 3 and product type 12

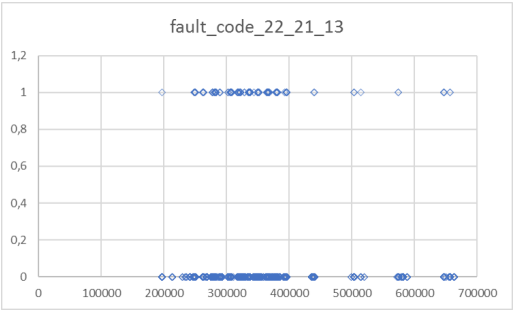


Figure E.7: Pores/Cracks, power test 3 and product type 13

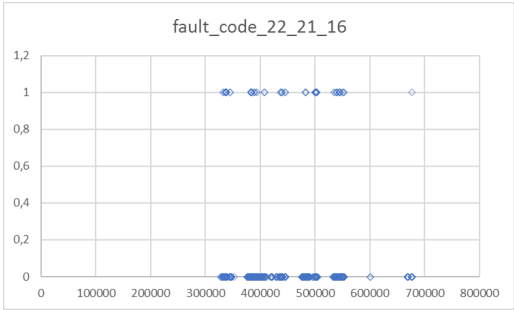


Figure E.8: Pores/Cracks, power test 3 and product type 16

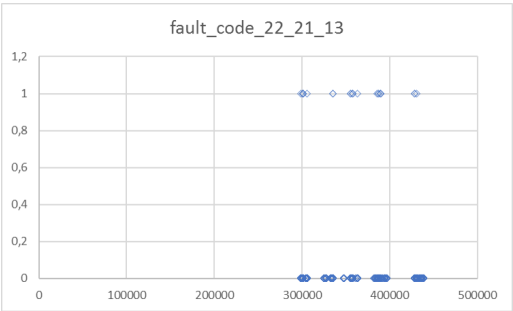


Figure E.9: Pores/Cracks, power test 4 and product type 13

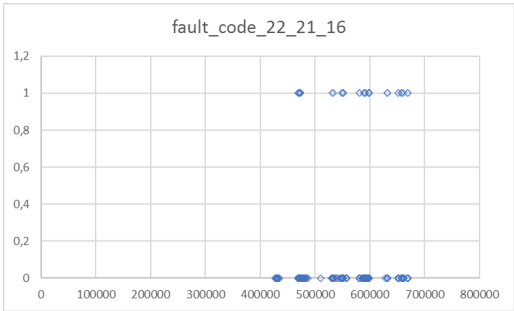


Figure E.10: Pores/Cracks, power test 4 and product type 16

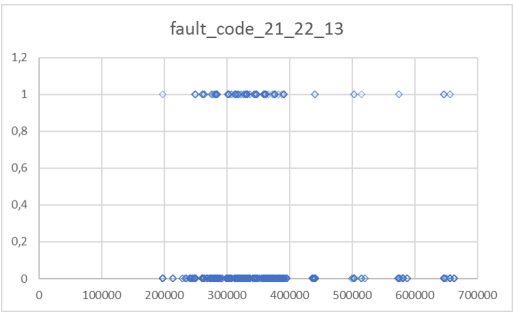


Figure E.11: Pores/Cracks, power test 5 and product type 13

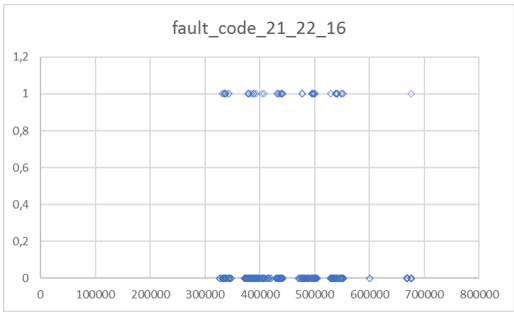


Figure E.12: Pores/Cracks, power test 5 and product type 16

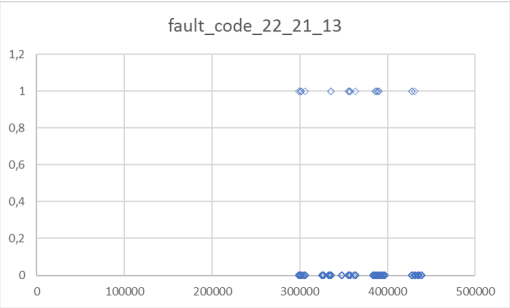


Figure E.13: Pores/Cracks, power test 6 and product type 13

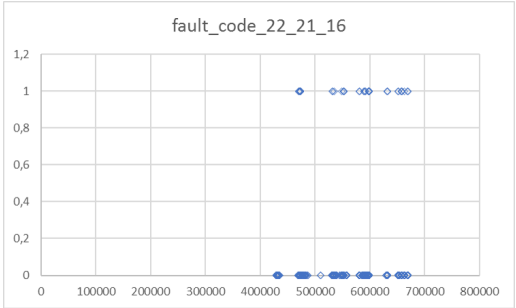


Figure E.14: Pores/Cracks, power test 6 and product type 16

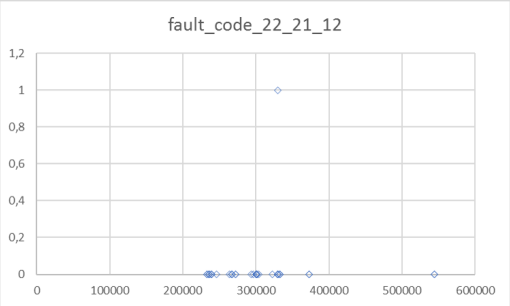


Figure E.15: Pores/Cracks, power test 7 and product type 12

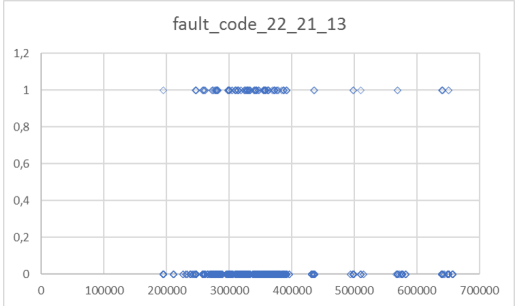


Figure E.16: Pores/Cracks, power test 7 and product type 13

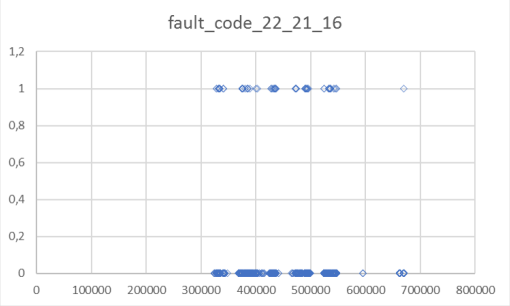


Figure E.17: Pores/Cracks, power test 7 and product type 16

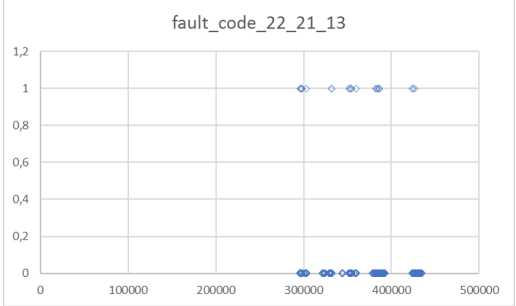


Figure E.18: Pores/Cracks, power test 8 and product type 13

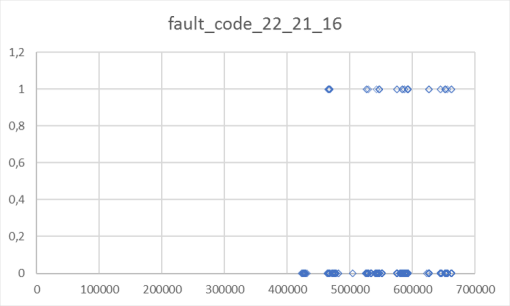


Figure E.19: Pores/Cracks, power test 8 and product type 16

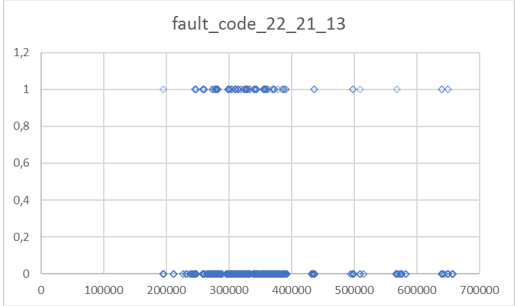


Figure E.20: Pores/Cracks, power test 9 and product type 13

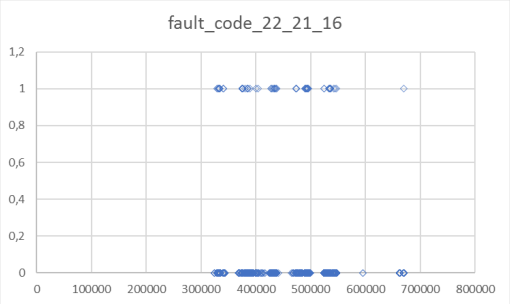


Figure E.21: Pores/Cracks, power test 9 and product type 16

F

Appendix

The result of the outlier analysis. Red cell represents a value within the limits, green cell represents an outlier and empty cell, missing test value. Furthermore are fault code groupings on the X-axis and tests on the Y-axis. The populace consists of the product individuals from quality case 2.

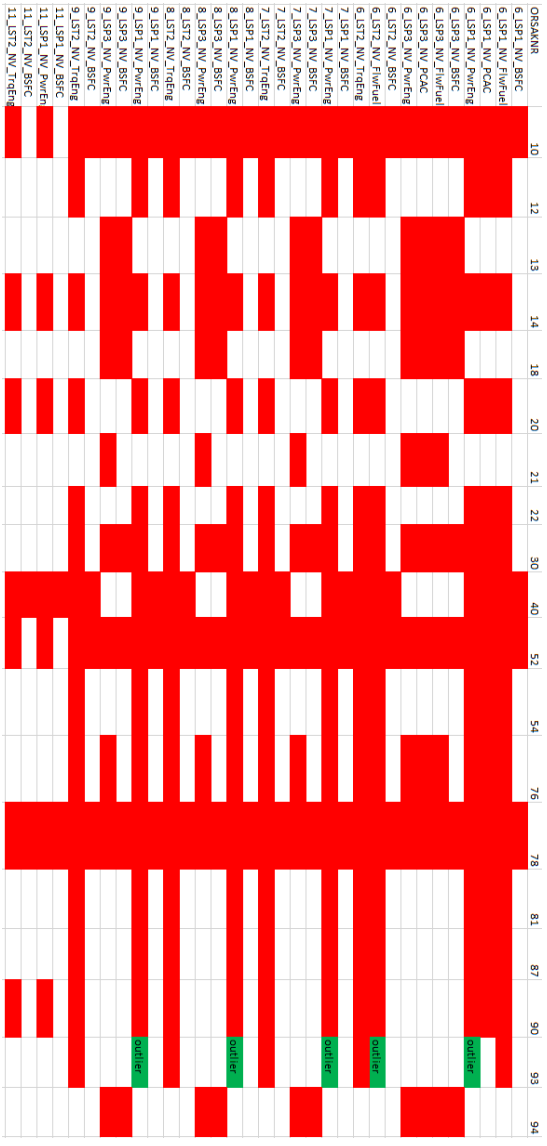


Figure F.1: Outlier analysis with 2 standard deviation limits

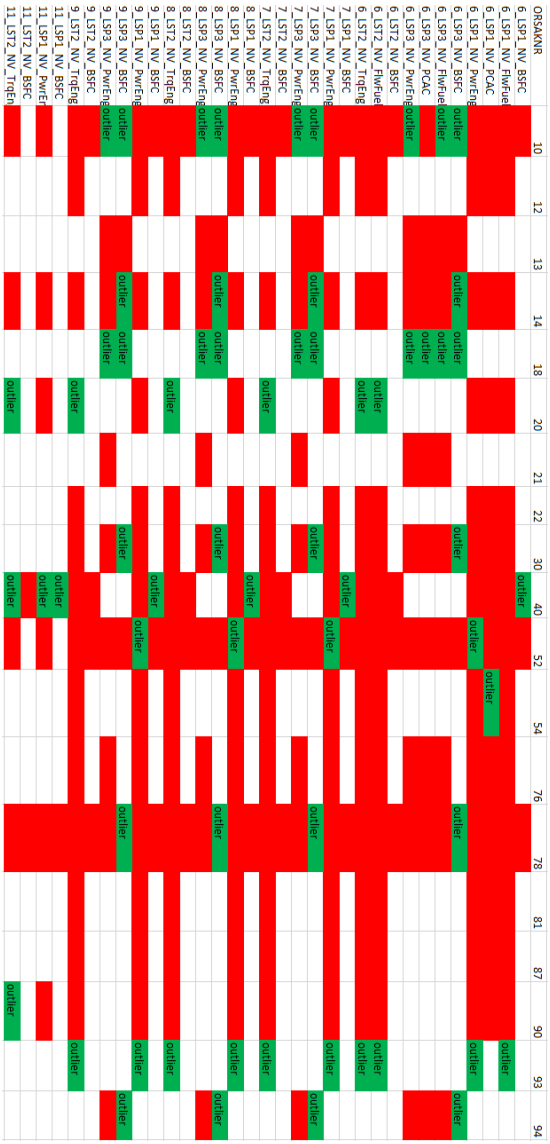


Figure F.2: Outlier analysis with 1 standard deviation limits