CHALMERS
UNIVERSITY OF TECHNOLOGY

Data Driven Maintenance at Gothenburg Trams

A study at Gothenburg trams

Master's thesis in in Production Engineering and Sustainable Energy Systems

SHARATH. J. BHUSHI

Department of Industrial and Materials Science
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2018

MASTER'S THESIS 2018

Data Driven Predictive Maintenance
A Case Study at Gothenburg Trams

SHARATH. J. BHUSHI

Department of Industrial and Materials Science
Division of Productions Systems
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2018

Gothenburg, Sweden 2018
Data Driven Predictive Maintenance
A Case Study at Gothenburg Trams
Master of Science Thesis [Industrial and Materials Science, IMS]
SHARATH J BHUSHI

Supervisors:
Mukund Subramaniyan, Department of Industrial and Materials Science, Chalmers
Lars Sandberg and Håkan Sjöberg, Prevas
Lalla Fondin and Krste Cvetkovski, Gothenburg Trams
Examiner: Anders Skoogh, Department of Industrial and Materials Science, Chalmers

SHARATH.J.BHUSHI, 2018
Department of Industrial and Materials Science
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Data driven predictive maintenance

SHARATH.J.BHUSHI
Department of Industrial and Materials Science
Chalmers University of Technology

Abstract

As the new era dawns in, the lines demarcating the division of the service and the secondary sector seems to be reducing. With multiple Secondary sectors (Manufacturing) shifting to IoT and Industry 4.0, the tendency to take key business decisions based on the data has become the norm. The improvements and advances in the field of Artificial Intelligence and Machine learning on the data sets ranging from medical, defence, manufacturing, finance etc. has taken the world by surprise with multiple industries investing resources to make strides in this field. The technology does seem to be mature enough to be applied to the manufacturing sectors as well and is deemed to usher in a new era of taking business decisions.

The purpose of the thesis is to achieve a smarter and effective predictive maintenance in a workshop whose key aim included was to evaluate and understand the potential for a data driven predictive maintenance at Gothenburg Trams and thereby paving a way for the future work to be carried out in the field of AI. A model called CRISP-DM was implemented as a key methodological process and was used during the entire course of project. A time tested data mining process model which is used to solve many problems in the data science field and commonly adopted across most of the Industry 4.0 compatible industries.

The project resulted in a successful descriptive analysis of the different tram models with having identified the key components that contributed to major losses for each of the tram segments. The critical component of M29 and M32 tram types had close to 80% as unscheduled maintenance. The average distance travelled by the critical component is not even 20% of the expected value! But as the data sets were not sufficient enough to provide a solid regression models, the project had to be cut short to building and identification of the key issues in a maintenance activity. The learning, the key issues and the mistakes from the project can be seen as a vital feedback to the maintenance industry whose mistakes and the drawbacks can be seen as various impediments that occur in an maintenance industry which try to be Industry 4.0 compliant. The work done can be utilised for building a good predictive maintenance program for Gothenburg Trams.

Key Words: IoT (Internet of things), Industry 4.0, AI(Artificial Intelligence), CRISP-DM, predictive maintenance, data science.

Acknowledgements

This master thesis has been written at the Department of Industrial and Materials Science at Chalmers University of Technology and in cooperation with Göteborgs Spårvägar (aka Gothenburg Trams) and PREVAS AB in the spring of 2018.

I would like to thank *Mukund Subramaniyan*, at the Department of Industrial and Materials Science at Chalmers Technical University, for his help and insight in both statistical analysis and predictive modelling.

I also want to give recognition and a thank you to Lars Sandberg from PREVAS for having brought up the thesis topic and Håkan Sjöberg for having helped me set up the working environment

I would also like to thank Lalla Fondin and Krste Cvetkovski for helping me get the business insights and providing me an indepth knowledge about the functioning of the company

A big thanks to Johan Bengtsson of Smarta Fabriker for having Identified Prevas as a partner entrusted with IoT and big data technologies.

Thanks to Parents and other well-wishers for their continued support and commitment.

Sharath Bhushi , Gothenburg, July 2018

Contents

1

Introduction

The first chapter brings to the reader the introductory aspect of the subject area. A brief background to the subject area along with overall purpose, objective and goals of the project are thereby presented in the following sections. A brief review of what predictive maintenance is and different approaches of obtaining the scheduled operations are also discussed. The delimitations are also further defined in this section in order to confine the topic thereby defining the scope of the thesis and also a brief skeletal structure of the entire thesis can be understood from this topic as well.

1.1 Background

In the contemporary era where data driven decisions have become the norm in the services sector, it's only a matter of time when the concepts are employed to the many industrial systems consisting of a production floor, a factory setting or maybe a component even. In such cases the aspects of costs, profits and safety are of topmost concern. The significance of maintenance operations in this regard has become an important tool to have a sustainable and a viable manufacturing systems. As per authors (Maletic et al., 2014) the aforementioned concerns are addressed by practicing an efficient maintenance program impacting both productivity and profitability. In order to address these issues, suitable course of action needs to be implemented one of which is the predictive maintenance which is one of the plethora of maintenance schemes that are usually employed by the industries.

In any transportation and other service oriented systems, avoiding downtime is the key to profitability. Unplanned and unscheduled maintenance are where it hurts the most for these companies. As per (Susto et al., 2012) the maintenance management can be grouped into three main categories namely: Run-to-failure (R2F), Preventive maintenance (PvM) and Predictive maintenance (PdM). The current low buffer time provided, along with the high costs associated commands for a system that's dependable and trustworthy giving rise to a need that makes the workshops more dependable.

With the advent of Industry 4.0 and the emergence of "smart factories" the aspect of maintenance is undergoing a tremendous change. The ever increasing data from multitude of sensors, the ERP (Enterprise resource planning) systems etc. is changing the way decisions are being taken when it comes to scheduling, maintenance management and other quality improvement techniques (Susto et al., 2015). As per the author (Prajapati et al., 2012), precise and accurate data from the production systems form the backbone of the predictive maintenance (PdM) ideals of which can be implemented in the transportation sector as well.

With this amount of data that ranges from 100 Terabytes/day (Thaduri et al., 2015) and above that's available, historical trends can be observed in order to analyse the point at which a component starts to deteriorate. The paper also goes on to say that these trends can then be used to study the criticality of the component and thereby plan the maintenance operations accordingly in order to prevent the unplanned and unscheduled maintenance thereby saving costs and time. The output that is obtained can then be used to build multiple ML models and thereby train them as per convenience.

As far as the research field is concerned, there aren't many research papers to discuss about the Maintenance of trams . Most of the papers deal with the maintenance of the tracks and the entire railway network in general. A study by the author (Zhang, 2012) made on the high speed train maintenance was a credible source of information concerning the research being done in this domain. It talks about the design of the EMU (Electric Multiple unit) IoT system. The paper gave a blue print on how the IoT (Internet of Things) system could be achieved by seamless integration of multiple RFID tags with which data concerning various production aspects such as trains flow, parts flow, labour jobs flow and equipments, monitor productive process and logistics of train maintenance in total life cycle could be recorded. The paper discusses about implementing the system in general and doesn't discuss about the key issues one might face while analysing the dataset. The authors (Fumeo et al., 2015) talk about the train axle bearings maintenance based on the real time big data streaming analysis. The paper talks about the different algorithmic models that are implemented in achieving the same. The paper talks about the computational requirements and about the algorithms in detail, but doesn't dwell much into the topic of how the data structure seems to be.

The most suited research paper that is close suited to the topic that the paper dealt with has to be the study conducted about the Swedish railways (Thaduri et l., 2015). The paper talks about how the railway assets act as a potential domain for the big data analytics. The paper talks about the dilemmas infrastructure managers face when dealing with hundreds of sources which has structured data like timings, speed etc., semi-structured like images and videos and unstructured like maintenance records. The data management issue deals with the multiple modules of data the railways are supposed to deal with. The multiple data sources thus make it a difficult task in order to calculate the remaining useful life of any component in the railway system. Gaps such as heterogenous sources of information, real time requirements and algorithms that are suited for lab conditions and not in the real life scenario is also discussed in the paper. However, the paper doesn't go in detail with the different data set provided and this current thesis hopefully will throw a light on how smarter decisions are got to be made in the workshop setting when it comes to maintenance which most of the earlier studies haven't spoken much about and hopefully the insights that are brought out at the end of the project provide a small basis for the maintenance industry in trams.

1.2 Purpose

Thus, the purpose of this project is to enable a smarter and effective descriptive maintenance in a workshop setting using data driven decision making. The key aim in this particular project would be to analyse the behaviour of the maintenance structure being followed at Gothenburg trams and to help them take better decisions by analysing the hidden patterns and coming up with a predictive maintenance description in the project. The smarta fabriker project, which

Prevas AB is part of collaboration between school and business and a platform for creating expertise and disseminating knowledge about industrial digitization and thereby preparing the workshops to be ready for the Industry 4.0.

1.3 Objective

The task in this thesis was to evaluate as to how predictive maintenance could be achieved for a smart workshop setting, condition monitoring in the workshop, performing a data analysis and preparing a conclusive report as to how the maintenance activities can be carried forward. To have a brief understanding of the same the following two research question were framed:

RQ1. What insights can be drawn with the given data set ?

RQ2. What are the key gaps to be addressed in the data set and what has been done in the literature about it?

The first research question will be answered in the results section and discussion regarding the same will be in the discussion section. The question can further be answered by analysing the data set provided and also with the help of researchers who are from the academic background in the college in order to understand the kind of algorithms or predictive models that were implemented. The overall aim of this particular question is to see what insights that can be drawn with the available dataset and also explore the possibility of using it in order to build a good predictive model in the future.

The second research question concerns the issues that must be addressed before building a good predictive model. In order to implement the techniques or even build a predictive models as stated, there needs to be a series of steps that must be conducted in prior in order to achieve that and preparation of good data set is the first and the foremost step to go ahead. There needs to be multiple issues that need to be addressed with the data set ranging from the availability, quality and business context. The studies regarding the same in the previous research don will also be looked upon.

1.4 Scope

The key focus of the project as was stated earlier is there are five key areas that need to be looked up to in order to achieve predictive maintenance in a smart factory, Based on how the situation might persist, the data collection, storage, analysis and visualization needs to be carried out thereby involving a gradual step. The scope though not limited to the aforementioned steps, will also be based on the type and the quality of the data that was received from the client and gradual progress across each of the steps. The final potential implementation of the PdM needs to be achieved and its implementation in the client environment too needs to be assessed. The subsequent implementation on a larger scale is by the company the author is associated with, which can use this document as a means of understanding the implementation of the project.

1.5 Description of the case

Gothenburg Trams is a part of public transportation system in the city of Gothenburg whose operations are organised by Göteborg Spårvägar (Hereby addressed as Gothenburg Trams) and is controlled by Västrafik. As per the company's homepage it has a total of 263 trams and 64 buses. The trams come in four different models, where the oldest wagons are from the mid-1960s. All the vehicles are driven with the least possible environmental impact in mind. (Gothenburg trams, 2018). The further details of the trams are:

Model	Year Built	Weight (Ton)	Power (kW)	Number of Trams (2013)
M28	1965-1967	16.8	4*44	60
M29	1969-1972	17	4*50	58
M31	1984-1992	34.5	300	80
M32	2004-Date	40.5	4*106	65

The responsibility for the overall maintenance and upkeep of the trams lies with mainly two workshops where one is situated at Svingeln and the other situated at Majorna. The schedule for maintenance (discussed more in the results section) vary across each components and across different trams as well. The expectation was that all the sensory level data would be used for the analysis, but the data relating to work order was provided which did have a lot of loopholes and empty values in it. By continuously collecting and analyzing the real time data, costly corrective maintenance which is basically taking the tram to the workshop, can be avoided. (Carnero, 2006). The overall maintenance operations are coordinated by the Maintenance department of the company but the actual planning and the maintenance methodologies depends on the location and division of the workshop. (Gothenburg trams, 2018). The pictures of the different models of the trams are given below:

Built by: ASEA AB headquartered in Västerås, Sweden

Figure 1: Tram model M28

Built by: Hägglunds and Söner headquartered in Örnskäldsvik, Sweden

Figure 2: Tram model M29

Built by: ASEA and ABB headquartered in Västerås, Sweden

Figure 3: Tram model M31

Built by: Ansaldobreda headquartered in Naples, Italy

Figure 4: Tram model M32

1.6 Delimitations

The various delimitations of the project include:

1. There was a delay in getting the data initially
2. The data considered for the descriptive analysis is from 2015-2018. The components are then selected accordingly. It can vary if an another timeline is chosen.
3. There can be discrepancies in the data set prior to the year 2016 since the database system was completely overhauled in that year
4. There were multiple views (tables) of the data provided but only 4 of them were utilised initially.
5. The shop floor process data could not be obtained since the personnel felt that data would be out of scope for the project.
6. The quality of data was not as good as expected with multiple blank values and errors.
7. Most of the research in the smart maintenance are done in Big railway systems where large amount of data is available. NO credible research done on trams and with limited data.
8. The models M28 and M29 showed very similar characteristics, and in some cases the components in one fit in the other tram type as well. So there can be similar component and breakdown characteristics for the two tram types.
9. The number of data points for the M31 and M32 trams are higher due to their capabilities to traverse long distances and are thereby used more frequently. The data is of good quality too from these two tram types.
10. A good regression model could not be unfortunately constructed as the p values obtained were not satisfactory and the r-Squared value for the model built for the model built was nowhere close to 0.75
11. The project doesn't look into the aspects of how The company would go on to solve the problem of the clients but reflects only the author's understanding and way of approaching the problem
12. . A holistic approach is followed while solving the problem of the respective clients. The data again seemed to be specific to a particular process and limited to a particular geographical region which would again with it bring in geographical differences associated.
13. The guidelines for building a predictive models were followed but due to the complexity of the data bases, a predictive model could not be built and instead a descriptive analysis of the problem has been obtained.

2

Theoretical Concepts

The current chapter serves as a medium to throw light on some of the theoretical concepts that forms the basis of the current project. Along with concepts, a case background is also provided with the clients in focus in the given project. Some key maintenance concepts, key findings in the PdM methodologies, few machine learning concepts are explained too. Though not an in-depth explanation of each of the concepts, the basic framework underlying the project has been explained in brief

2.1 Basic Terminologies

Since the dataset is limited to geographical region of Sweden, the definitions of the terms are of Swedish standard institute (2001).

Maintenance: As per the standard institute the word is defined as: "the combination of all technical, administrative and managerial actions during the life cycle of an item intended to retain it in, or restore it to, a state in which it can perform the required function"

Reliability: As per the definition of the Swedish standard institute the word reliability is defined as: the ability of an item to perform a required function under given conditions for a given time interval"

In the field of maintenance management, the maintenance of the equipments are usually classified under three major subcategories. As per authors (Susto et al., 2012), the three main orders in the increasing order of complexity and efficiency are:

Run-to-failure (R2F)—where maintenance interventions are performed only after the occurrence of failures. Easiest approach courtesy which it's more widely used than the other two. Thus making it more frequently adopted , but it is also the least effective one, as the cost of interventions and associated downtime after failure are usually high than those when compared to the ones with corrective actions that are employed usually in advance.

Preventive maintenance (PvM)—in this scenario, the maintenance operations are carried out usually based on planning and iterations. Also called as *scheduled maintenance* approach, failures are usually prevented, but a huge amount of resources is spent in terms of time and money in getting the corrective actions performed which usually ends up seen as a huge increase in operating costs and time as well.

Predictive maintenance (PdM)—in this scenario as per the same author, the maintenance activities are carried out based on the working condition of the piece of the whole equipment

or the entire equipment itself. The PdM systems when successfully implemented, helps for early detection, helps to enable early corrections of equipment's, predictive tools based on the historical data, statistical approaches and other machine learning algorithms help achieve an efficient maintenance system.

As per Lenz and Barak (2013), with the advent of various statistical methods, the ones based on machine learning are the most suitable for dealing with modelling for high dimensional problems, such as the production process systems where most of the components variabilities are subjected to numerical constraints such as pressures, voltages and other product specific factors.

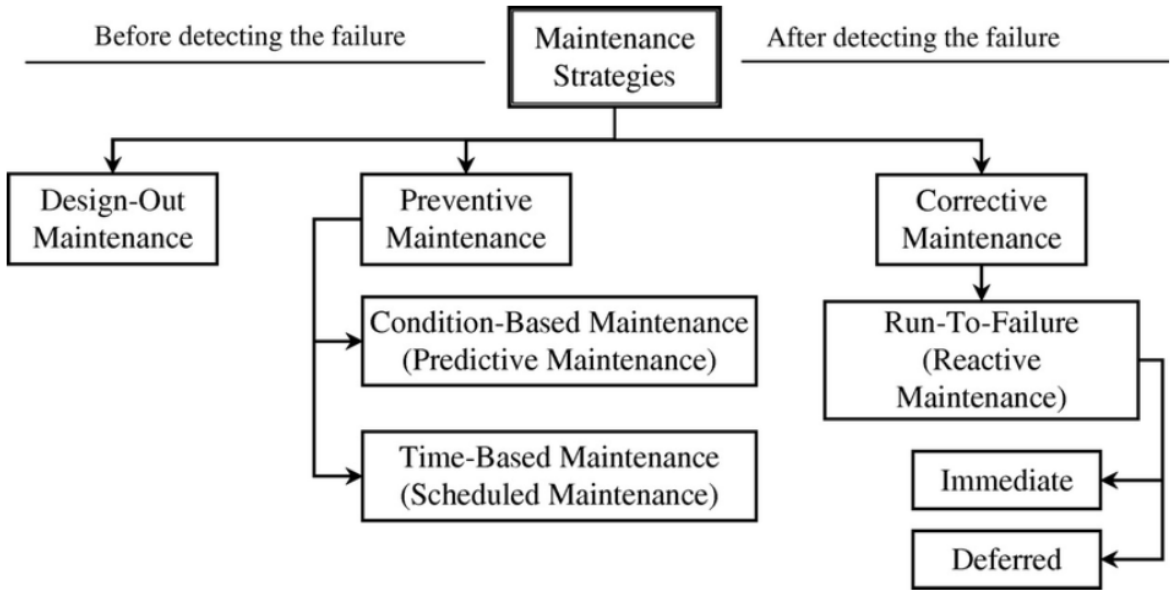


Figure 5 :Maintenance cycles as per Mostafa et al (2015). Note that the current thesis deals with 3 major parts of this maintenance cycles.

As explained in the previous paragraphs there are major two types of maintenance activities that can be carried out. One that's before the detection of the failure and the other that's usually after the detection of the failure and is the most commonly employed tactic. The scheduled maintenance activities are more resource consuming both in terms of time and money as well. For which a predictive maintenance framework needs to be based upon. Based on the diagnosis result of failure modes and severity, predictive techniques can help determine how fast the degradation is expected to progress from its current state to functional failure and offer a trade-off maintenance strategy. (Wang et al., 2017)This can be analysed by looking at the following paradigm.

2.2 Predictive maintenance Paradigm

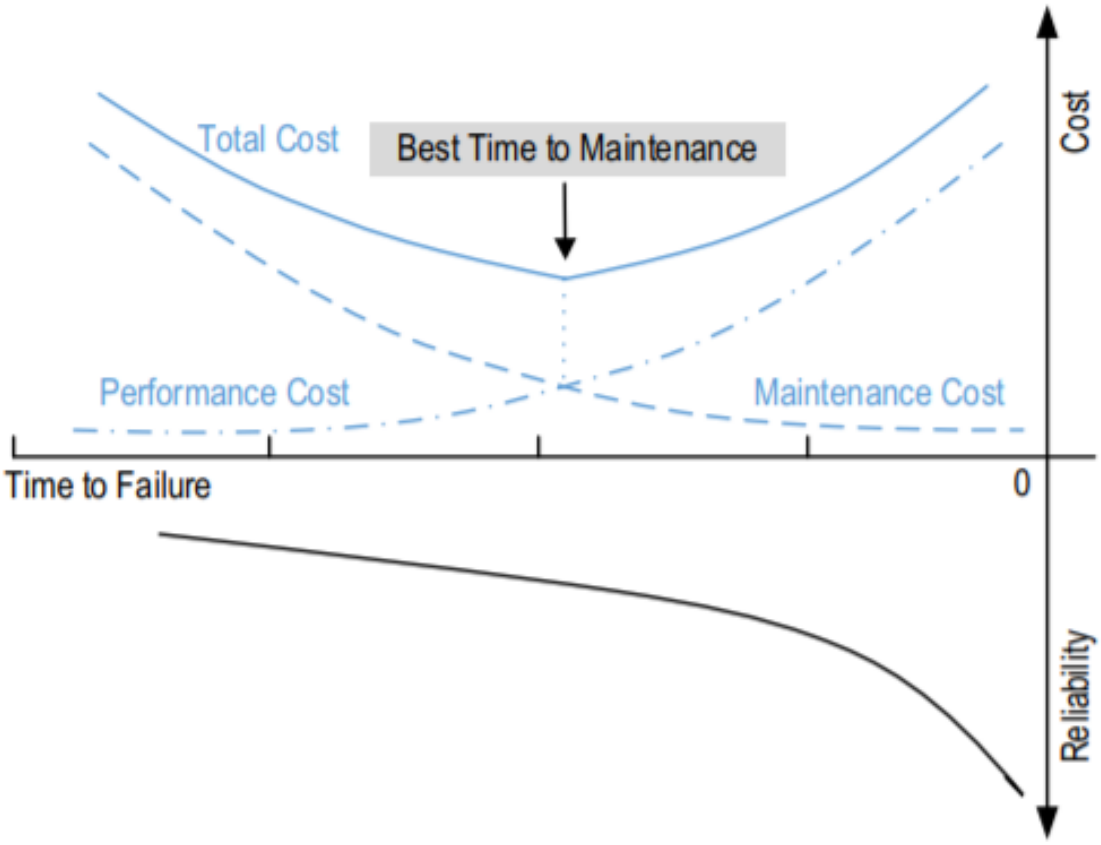


Figure 6: The relationship between failure rate, reliability and cost

The above figure shows the relationship among the cost, time to failure and the reliability of machines. As the author (Larry, 1995) has discussed in his paper, when the time to failure equals zero, the system or the component gets into a breakdown status. The reliability of the system keeps reducing *(getting more negative in the above graph)* to a minimal level as the time to failure of the system approaches to zero as seen in the graph. When compared to the graphs that are present in the second quadrant of the same graph, the performance cost of the system reduces; thereby impairing the component's productivity and the maintenance reduces too; indicating the fact that the less reliable a component has become (the more it's closer to the scrap level) the maintenance cost is minimal. The total cost, as seen by the parabolic curve , which basically includes the sum of the performance and the maintenance cost, reduces first to a minimal level and then rises all over again back to where it stared. It's here that the predictive maintenance with the capacity of precisely predicting the time to failure and reliability of the system can provide meaningful insights and information for the decision to be made on whether the maintenance has to be carried out considering the economic aspects of the project.

2.3 Maintenance strategy and decision making

As per Kobbacy & Murthy (2008), maintenance across complex systems have emerged from being regarded as a trivial issue to a matter of strategic importance. The idea of maintenance

being considered as a necessary distraction to something that can drive strategic and operational decisions. As per Campbell & Reyes-Picknell (2015) it is of high importance to complement a strategy with tactical and operational directives as well as aligning it with the organizational business strategy. Though the definition of the term strategy remains unclear, when it comes to the manufacturing sector it's usually defined as means of transforming the business priorities into maintenance driven key priorities (Salonen, 2011). The definition is further enhanced by the author (Tsang, 2002) who believes a maintenance strategy must serve as a support for strategic issues such as resource allocation, operation prioritization and other key business goals. There are multiple strategies to go ahead for selecting a maintenance strategy:

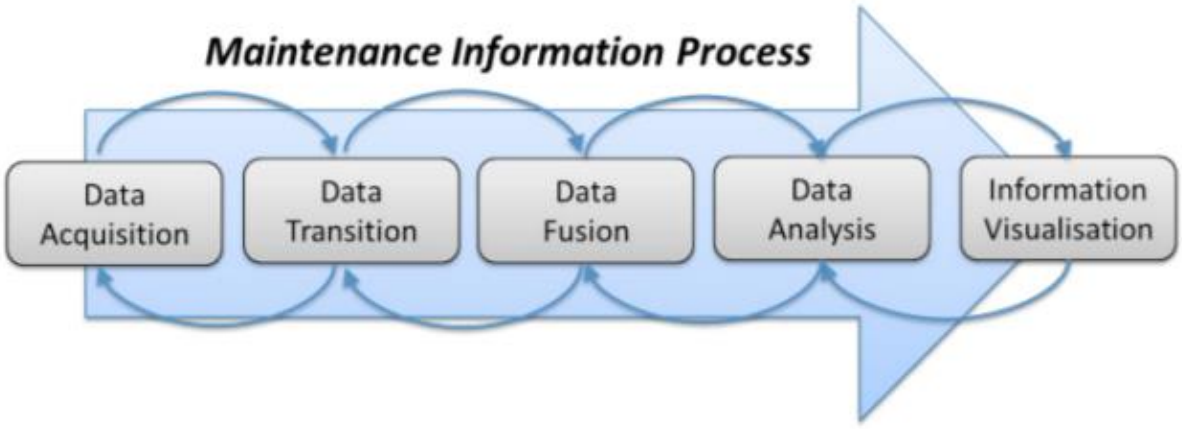


Figure 7: Knowledge discovery process (Karim et al., 2016)

The knowledge discovery process consists of the aforementioned steps as shown in the figure. The first step is obtaining the relevant data and managing its content. Data transition includes communicating the collected data. The data fusion step consists of data compilation from multiple sources. Data analysis includes analysing the given data to extract information and knowledge. The reverse cyclic process that occurs after each steps shows a rework needs to be done in case the data obtained in the current process is not satisfactory. The information obtained through the visualization is further used to support the maintenance decision (Karim et al., 2016).

2.3.1 The maintenance analytics concept

In the contemporary era where technological changes are occurring more rapidly, wherein the smart factories, Industry 4.0, IoT enables the companies and the users to have a enhanced level of information. This calls for a need of a structured approach and concept to improve information extraction and knowledge discovery. A concept is proposed by the authors to deal with the usage of analytics in methodological perspective. The concept is based on the four interconnected time-lined phases which aim to help and aid the maintenance action through analysing and understanding the data. It consists of four different aspects and shown and explained below (Karim et al., 2016).

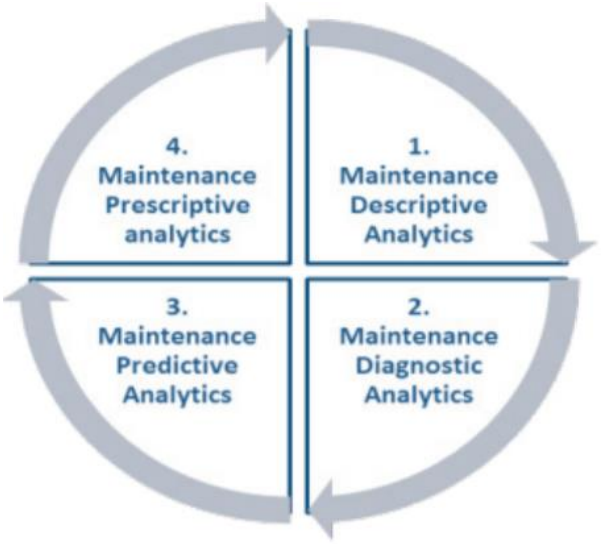


Figure 8: Phases of maintenance analytics concept (Karim et al., 2016)

- The ***maintenance descriptive analytics*** phase answers as to what events or situations have happened. Data access to system operation, condition is of utmost importance. It becomes crucial to understand the relationship between states and events and their associated time frame in a logged data. This makes the events associated with the system configuration acting on a certain time interval. Synchronization of time becomes a key part in support maintenance analytics.
- The ***maintenance diagnostics analytics*** phase answers the why aspect of the maintenance paradigm. The outcome of 'maintenance descriptive analytics' is used having a pre-requisite that the data used should be of good quality and reliability.
- The ***maintenance predictive analytics*** phase aims to answer 'what would the outcome be'. This phase needs the outcome of the previous phase. This phase is a crucial aspect of the cycle too as it incorporates certain business data such as planned operation and maintenance in order to give an output that is highly reliable and well within the confidence interval.
- The ***maintenance prescriptive analytics*** phase of the maintenance analytics concept tries to answer as 'what is supposed to be done'? The outcome of the two phases is dependent on this phase. In addition, in order to predict upcoming failure and fault there is a need to provide resource planning data and business data.

Analytics is the process of generating knowledge based on understanding of the underlying process. The above *Maintenance analytics* is a concept for big data analytics and maintenance. It focuses on key technological aspects such as service-orientation, distributed computing, modularization and usability. With the help of such cutting edge technologies the above concept can be put into practice thus making it a strong case for the decision-making process in the field of maintenance (Karim et al., 2016).

2.4 Framework for implementing PdM

As per the authors (Groba et al, 2007), the process of carrying out a Predictive Maintenance program can be divided into 5 key steps. Namely:

- Identification of significant equipment and indicators
- Measuring of the indicators
- Modelling of the indicators
- Forecasting of the indicators
- Developing effective decision-making procedures

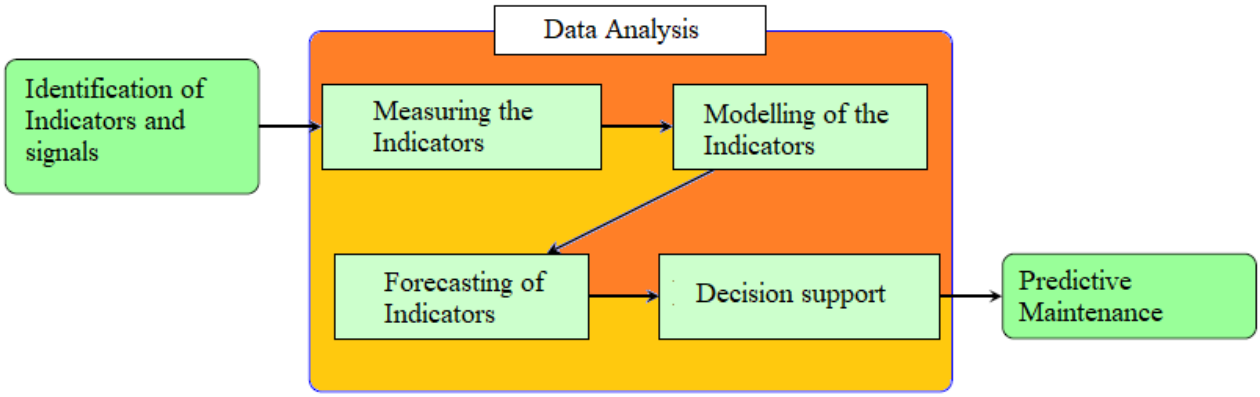


Figure 9: Flowchart for PdM implementation as per Groba et al., 2007

The above image depicting a flowchart indicating, how predictive maintenance can be carried out using the skeletal framework. Data that's acquired from various systems and sensors are measured in the first place and a certain pre-processing activity is carried out. A certain rules are then implied based on the business context and related algorithms are devised. The various results from the algorithms thereby derived are then tested and approximate forecasts are made from the historical dataset to ascertain the feasibility of the model. The decision support is then carried out in the end explaining the credibility of the model and if found accurate to a satisfactory level, the key actions are then taken in order to achieve predictive maintenance.

2.5 Basic Statistical concepts

The presentation of the results from a statistical analysis can be split in three categories:

- Descriptive Statistics
- Inferential statistics
- Explorative Statistics

2.5.1 Descriptive and Inferential Statistics

Descriptive statistics aims to describe various aspects of the data obtained in the study such as listings, graphics and summary statistics. Summary statistics includes mean, mode, median,

standard deviation of the dataset provided etc. In the current thesis and in most of the conclusions that are supposed to be drawn from data, inferential statistics forms a basis for conclusion regarding a specified objective addressing a given set of data of the population under the study.

Confirmatory analysis forms a pattern:

HYPOTHESIS → RESULTS → CONCLUSION

This forms the backbone of the descriptive and the inferential statistics wherein a set of hypothesis is concluded by the initial exploration of the data. If the results are in accordance with the assumed hypothesis, a viable conclusion is thereby drawn. The conclusion can either reject the hypothesis assumed or fail to reject the hypothesis due to lack of statistical inference.

2.5.2 Explorative statistics

Explorative statistics aims to find interesting results that can be used to formulate new objectives/hypothesis for further investigation in future studies.

Exploratory analysis forms a pattern:

RESULTS → HYPOTHESIS

2.5.3 Hypothesis Testing, p-values and confidence intervals

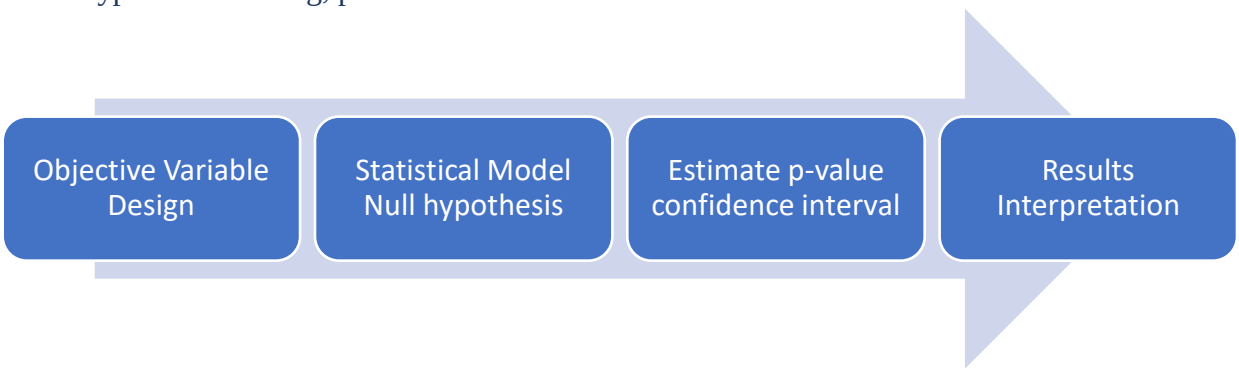


Figure 10: Basic Statistical Modelling process

Process Explanation:

1. An initial research hypothesis is proposed and the null and alternate hypothesis are thereby stated.
2. Statistical assumptions are made and tests to check if the values are normally distributed are then conducted.
3. A significance level (α) is selected, a probability below which null hypothesis will be rejected. The p- values (The smallest significance level for which the null hypothesis can be rejected) for this to be considered are 0.05 or 0.01
4. The distribution of the test statistic under the null hypothesis partitions the possible values into those for which the null hypothesis is rejected if it falls in the critical region whose probability is (α). The decision to reject or fail to reject the null hypothesis is made in this step.

Confidence value is an interval estimate calculated from the statistics of the given data that might contain the true value of un unknown population parameter which could be the median, or even the average value. (Basic Mathematical Concepts- Chalmers, n.d.)

2.5.4 Goodness of fit Statistics

A well-fitting regression model helps us to obtain the predicted values close to the observed values. For the evaluation of the regression models and to estimate their fit there are three key statistics used in regression to evaluate a modular fit. These are based on basicaclly the sum of the squares concept. Sum of squares Total (SST) and sum of squares error (SSE). SSt measures how far rthe data points are from the mean and SSE tells howfar the data points are from the predicted values. The three key statistics are:

- R-Squared and adjusted R-Squared

The difference between SST and SSE is the improvement in prediction from the regression model, compared to the mean model. Dividing that difference by SST gives R-squared. It is the proportional improvement in prediction from the regression model, compared to the mean model. It indicates the goodness of fit of the model. It ranges from 0 to 1 with 1 bing the best fit for a particular regression. An issue with R-Squared is that it can increase as the number of variables tend to increase (more the number of predictors). Sometimes when the predictors actually don't improve the fit of the model, adjusted R-Squared is then used incorporating the model's degrees of freedom. Adjusted R-Squared must be used in instances when there are more than one predictor variables.

- F-Test

The F-test determines if the relationship considered in the null hypothesis, between the response variable and predictors is statistically reliable and useful when the objective of the research is to make a good predictive model or even in an explanatory analysis.

- RMSE (Root Mean Square Error)

The RMSE is the square root of the variance of the residuals. It shows the model fit of the data showing how close the observed data points are with respect to the model's predicted values. RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit and is the most used criteria for predictive modelling. (Sweet and Grace-Martin, 1999)

3
Methodology

This current section description of the main work process by describing the basic methodology followed. This section would also deal with the strategy adopted to go ahead with the project and also describe the model thus implemented.

3.1 Research strategy adopted

In the subject research methodology, it was made quite clear that having formulating the right research question and by developing a key research strategy, the success of any research project is half achieved. (Bryman and Bell, 2015). Along with which quantitative and qualitative research, questionnaires and workshops, qualitative and quantitative data analysis and writing up the business research was part of the study as well.

The choice of research strategy varies across the projects. Since the project that I dealt already had a set of historical data being collected, opted for a case study strategy too. The interviews could have been carried out as well, but courtesy the busy schedule of the employees at the office, it unfortunately couldn't be executed. A simple **AIM model** could have been implemented too as it's known to be quite effective in analysing the business problems well (Alänge and Scheinberg, 2005), but the key stakeholders were busy and couldn't get the managerial people under one room to have a discussion about what the current problems are. A basic quantitative research had to be executed in the concerned project as there was availability of the raw unstructured data. In addition to going through multiple research papers the following research process was adopted:

1) **Plan**
a) Problem identification
b) Stakeholder identification
c) Defining time plan
2) Develop research instruments
a) Literature review
b) Archival research
3) Collect data
a) Data exploration
4) Analysing the data
a) Descriptive analysis
b) Analysis of literature review and archival research
5) Pen down the Findings
a) Results presentation
b) Report preparation

The key step which usually involves in such projects are usually the ones that are dealt with exploration of the data and understanding the various attributes of the data set that is provided. *Cross Industry standard process for data mining* known by its acronym **CRISP-DM** (Shearer, 2000) is a data mining process model which describes the most ubiquitous approaches to tackle any data science problems. As per Mariscal et al. (2010), the model was to be considered as "de facto standard for developing data mining and knowledge discovery projects." There are various reviews of the models that were carried out by multiple authors. There were quite a few data mining approaches that were linked to this model too. Most notably the SEMMA approach which stands for *Sample, Explore, Modify, Model and Assess.* (Muenchen, 2012). The SEMMA model is a sequential algorithm that has been developed by the SAS institute. As per Azevedo et al. (2008), it acts as a useful guide in implementing data mining applications with companies like SAS acclaiming it to be a methodology that is meant to carry out the core tasks of data mining.

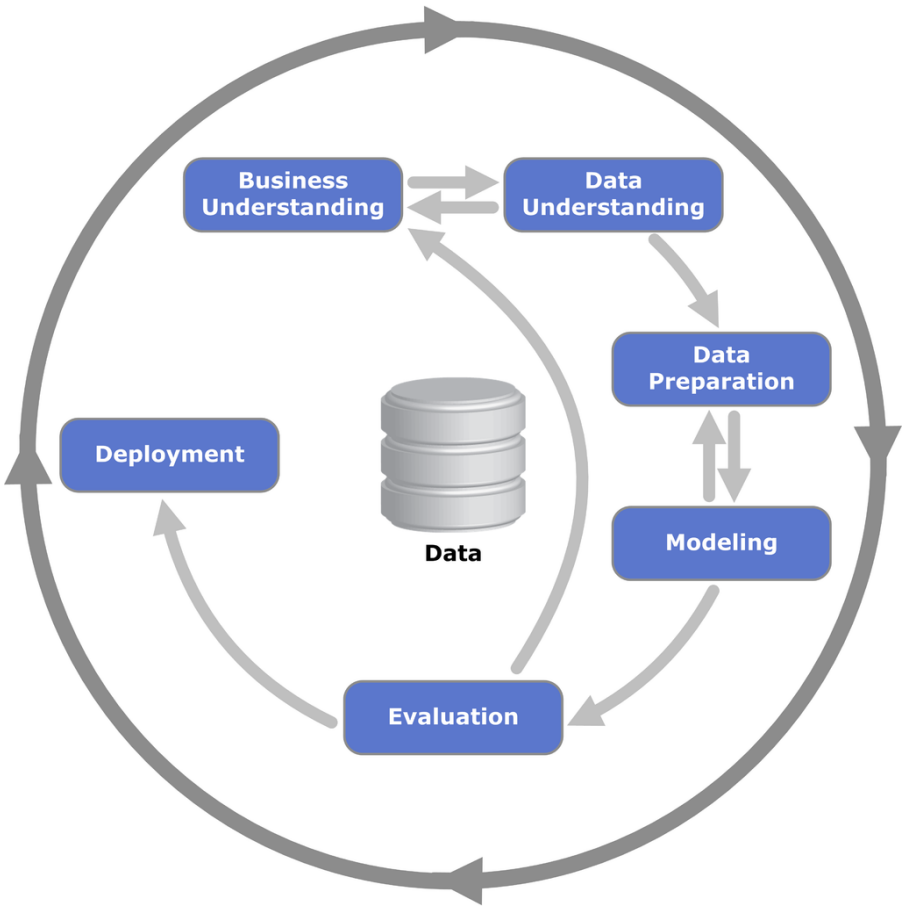


Figure 11: Process diagram showing the relationship between the different phases of CRISP-DM. A generic reference model (Jensen 2012).

3.2 CRISP-DM model phases

As per Harper and Pickett (2006) the model breaks down the entire data mining process into 6 major phases. The phase sequences are dynamic in nature and always fluctuate

depending on the requirement of any particular assignment. The multiple phases or steps contain arrows in the process diagrams which represent the vital and frequently occurring dependencies. The external cycle indicates that the process is cyclic and has to be redone and executed at all times for the gradual improvement of the model. This is basically the cyclic nature of the data mining process. The process of finding the most efficient process carries on spite of having figured out a solution in the first place. The experiences that are gathered and collected during the entire process helps to target more business oriented concepts and questions and thereby help the model to be more efficient by learning from the previous glitches from the model. (Harper and Pickett, 2006)

3.2.1 Business Understanding

Understanding the business context is the key role in solving any key business oriented problem. The projects' objectives and perspectives need to be understood first. This knowledge needs to be converted into a data mining problem definition. A preliminary plan can then be devised to achieve the objectives. This will give rise to a foundation of a decision model for which all the business decisions can be based upon. This is the key step towards data oriented decision making.

The project had many interesting questions that were meant to be dealt with. There were many key business decisions that were supposed to be made without taking into consideration the type of data and its quality. Variations among different components across different trams were successfully studied. The number of planned versus the unplanned work was analysed too, but only considering the data from the last 3 years. The third key insight which will be presented in the scatter plots below are the performance of critical components with respect to the distance covered. The success of this insight would prompt the company to either shorten the maintenance intervals of the critical components or change a particular maintenance plan. This was a key issue to know whether the components are able to handle the wear and tear for a right kilometre range.

There was a proposal to study the trends in case they existed among multiple components leading to a complete system failure. This was to determine if there was a link among multiple subsystems that would eventually lead to breakdown of the entire system or even the tram itself. Carrying out a correlation between the multiple components is possible but actually proving the hypothesis that the failure of one component leads to other needs a very high statistical significance which couldn't be obtained in this scenario. There was a study that was supposed to be conducted over the trams that were driven more than their expected kilometre range before having undergone a maintenance at the depot. The author (of this report) was supposed to attribute this to whether it was the depot that prompted this or whether it was the person in charge of maintenance. The data was quite blurry regarding this in the work order history. The same data set had an insight to cover wherein repeated failures or breakdowns of a particular component could be attributed to the last servicing that was done. There was a proposal to study the trends in different set of time ranging from time of the day to the time of the year but the semi structured data set of the work order history ensured otherwise.

3.2.2 Data Understanding

This phase usually begins with collection of the data. Various activities are then executed in order to get familiarized to the data and the working environment in general. Quality issues such as identifying the data quality and ranking them as per satisfactory levels is carried out too in this step. First insights of the data are carried out in this step. Any interesting subsets are found out to formulate key necessary hypothesis needed and also to check the validity of such hypothesis (if any) from the missing values or information.

In the data set that was provided, it did take quite a bit of time adjusting to the language since it was primarily in Swedish. As it was seen in the previous paragraphs, most of the insights that had to be covered were expected from the author to have a better data understanding. As the organization had gone major changes post 2016, it became quite clear that the data must be from the recent one. So only the data from the past three years were considered as a sample. There were a set of business rules to be followed too which included the overall scheduled maintenance for the trams after a certain kilometres. This was applicable to components too. Data understanding regarding the multiple discrete values in the data set had to be dealt with carefully since they had a symbolic meaning and were not mere random numbers.

3.2.3 Data Preparation

The key step in the model wherein the success of the project literally depends on this one key step. This step includes all the activities that help to transform the raw data from its initial to final form. This includes missing value treatment, removing the NA values and removal of values that do not fit the business context. This ensures that the data that's remaining at the end of the step is of good quality and ensures better model building capacity and better accuracy. The tasks are likely to be performed multiple times in order to make sure that the data comes out clean and decipherable at the end of the day. It does not follow any one specific order and can occur in any possible way. Various tasks that come under this step include tabulation, recording, selecting key attributes, cleaning and transformation of the existing data set.

In the data set that was provided there were a lot of issues. Since the time frame was fixed, there was a limited amount to be dealt with. The quality issues were a lot too. There were columns with multiple NA and even blank values. The business context of the data had to be considered too as well. There were instances where the tram ran close to 2000 kms a day and had to take out that data too as it didn't make business sense. There were many instances where date wasn't reported and the average run time of the component/tram had to be considered. The delimitations in doing so included the wrong assumption of the data which would then hamper the modelling aspect later. Erasing the data row made the matters worse too since there was a limited amount of quality data that was available.

3.2.4 Modelling

A phase wherein different models are made, verified and tested in order to select the best fit model. This step also includes the parameters to be calibrated to their optimal values. Typically, there are multiple techniques for the same data mining problem. The techniques

differ on basic requirements of the data. There are quite a few techniques which require data formats to be of certain structure (character, discrete data, continuous etc) . Therefore this phase involves stepping back and forth to the previous step as and when it's needed. The modelling phase will also involve the use of crucial parameters that are needed to build a particular model and dropping and combining of variables if any.

In the modelling phase, there were a lot of hindrances at the beginning. The descriptive analysis for each of the critical component across multiple trams were carried out. The work order history level data gave no significant insights on the business insights discussed earlier. There was no statistical inference on whether a particular workshop or an individual accounted for more number of maintenance as the data showed otherwise. After the descriptive analysis was carried out, a regression model for each of the components level data was planned to carry out as well, but due to low statistical inference (very high p-values) a known algorithm couldn't be used to construct a model. The overall damage models for the tram sets was carried out which had very high R squared values indicating a good fit. But since the major scope was to check the component- tram level data and not on all the components overall, it had to be discontinued.

3.2.5 Evaluation

This stage of the project involves evaluating the various models that the analyst has built. It may consist of only one model but nonetheless, testing of the same becomes a very crucial factor in the analysis. The best model needs to have high quality from the data analysis perspective. Before heading to the final model deployment, it becomes very crucial to evaluate the performance and also review the steps to construct the model. This is to ascertain that it is strictly in accordance with the business objectives that is supposed to be derived from it. A key concern that remains at this stage is to ascertain if a key business insight has not been considered and has to be accordingly adjusted. Usually by comparing the models by their error rate and by analysing the best fit for each of the model obtained, a decision on the use of data mining results need to be obtained.

The models were evaluated based on the summary function command in R and by analysing the graphs. While evaluating the models built for the overall damage of all the components across multiple set of trams, a very high R squared value was obtained. The lower RMSE value ensured that the model was well evaluated but the idea was supposed to have been implemented for the component level data. While evaluating the regression models that were built for the individual component taking multiple factors into consideration (**lm command in R)** , the results showed otherwise. The regression models thus formed had very high p-values under the 95% confidence level interval and thus were evaluated to be a loose fit models. Had the data remained in a more structural format and multiple years of data been considered, the evaluation would have been slightly better off due to high amount of data points that could be processed to evaluate the data.

3.2.6 Deployment

The project doesn't usually end after the creation of the model. If the sole purpose of the project is to increase the knowledge of the domain, the insights that are obtained throughout the process needs to be presented in such a way to the customer that it's beneficial to them in a business perspective. Based on the requirement and on the successful implementation of the above steps the deployment phase could vary from building a knowledge repository to implementing a sturdy algorithmic process to gain useful insights. In most of the cases it's observed that it's usually the customer who dictates the data mining processes and not the data scientist himself and will hence take the call for the entire deployment process too. The way analyst handles the situation and deploys the model is of utmost importance for the customer who will have to understand the kind of actions and procedures that are needed in order to actually make use of the models thus made.

Since the evaluation part didn't give any significant results and merely the descriptive part showed the current scenario, a robust model couldn't be built and therefore there wasn't deployment of any particular model to be considered. The image below shows the ideal wa of executing the CRISP-DM methodology.

3.2.7 A general Overview of the model

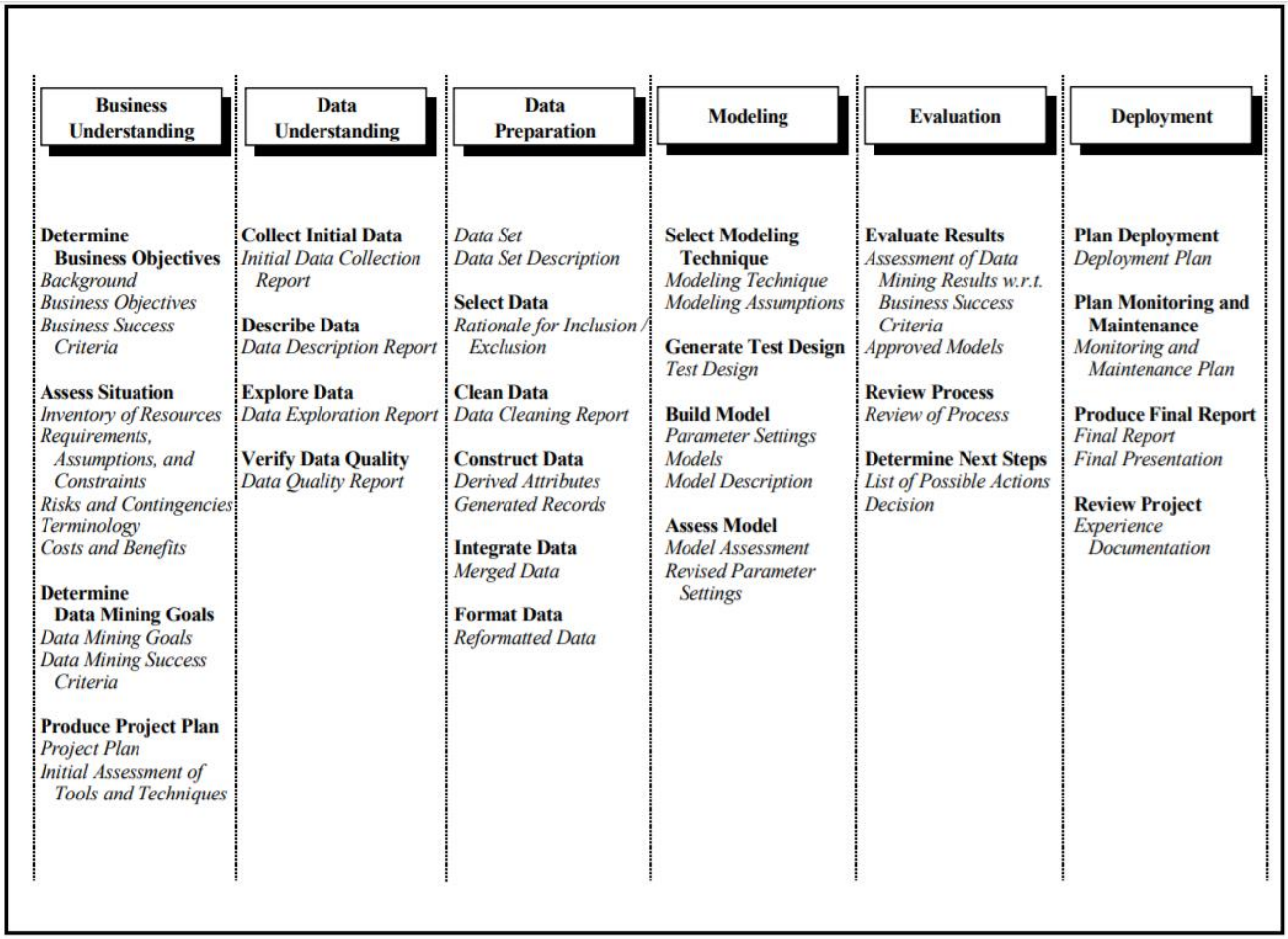


Figure 12: A general overview of the CRISP-DM model along with the desired output (Wirth and Hipp, 2000)

4

Data exploration and Results

The current section deals with the insights and the knowledge gained during the implementation of CRISP-DM methodology. The situational analysis and insights based on the data is presented. The section mainly covers the descriptive analysis and of course the parts and features that need to be improved upon. The section ends with suggestions and insights regarding the implementation of a predictive model and summary from the findings that were carried out throughout the project.

4.1 Business Understanding

In order to completely comprehend what the project was really concerned about and what issues that were supposed to be looked upon, a key set of objectives and desired results were agreed upon as discussed in the previous section without considering the feasibility of the data set. The following section discusses the some of the objectives that were supposed to be achieved and its consequent process.

4.1.1 Objective definition and situational analysis

As per the clients requirements, the project was mainly divided into the descriptive and predictive aspects respectively. The project was mainly concerned about the problems that were occurring at the component level and also at the overall maintenance work that was carried out in general. It was done so in order to analyse the downtime of the respective trams whether they were caused because of components basis or were formed because of some other reason which had to be ascertained. Though the insights that were supposed to be drawn were almost clearly well defined, the author of the report didn't go ahead to form a questionnaire model based approach to get the bigger issue of the problem.

The maintenance strategy adopted by the Gothenburg trams was similar in both the workshops. A periodic check-up after running a certain distance based on the classified tram types. There are failure based models as well wherein the components when reported as failed tend to get replaced by taking the tram to the maintenance workshop and checking for the overall health of the tram as well during that particular visit. This actually meant that there would be no specific time that would be allocated to the maintenance (apart from the kilometre based maintenance) which would actually be an hindrance to the predictive model that was being thought upon.

There were quite a set of objectives to be achieved. The first being on the component level. The variations of components and their behavioural study was to be carried out. The difference between the components and whether if some of the components had a greater chance of failure for a particular component. The study of planned maintenance and ascertaining of the performance of the component was supposed to be carried out as well. The trends (which might have been invisible) regarding the system or the component failure was decided to study as well. Subsequently, based on the kind of data that was supposed to be provided ,it was agreed upon to build a predictive model on the same. The maintenance based on the human factors such as the workshop type and the person who handled this was supposed to be studied as well. The linkage of the human factors with respect to the repeated failure was to be looked upon as well. Time and yearly based trends along with a brief study of routine was supposed to be studied as well.

As for as the tools are concerned regarding the project, Microsoft SQL server management was utilized to pull the data. Since the author has had previous experience working in R, the language was used to plot graphs and for testing various hypothesis and models that may have been suited for the predictive maintenance of the Trams. R is a free and open source language and has plethora of tools and libraries for execution of aforementioned tasks.

4.2 Data Access and Understanding

The data for the operations was stored in the SQL server. There could be two way s to access it. Access the database by a Citrix server (which was executed later) or use the dataset that was provided. Due to the extreme complications in the dataset that was provided only the descriptive part of the maintenance system could be obtained. The dataset that was provide to the author was about the component order history and of work order history of different set of trams. In order to build a robust predictive model and then employ the machine learning models on it, it's essential that the data remains of high quality and without much noise.

It used to occur quite frequently that the most desired columns of the tables used to be NA values or were either blank. Discussions were held regarding the same as to whether any specific rules and regulations could be applicable for the assumption of data in such cases, but no firm decision was made on this regard. The insights from the component or der history table were quite useful in analysing and understanding the depth of the business case scenario, but the work order history dataset which was provided played no key role in either descriptive or in predictive analysis.

The access to the data played another key role in the project. For the first few weeks the data that was shared with me which consisted of 4 major view tables having data from year 2000 till 2018. The scope was defined and it was decided to look into the data from these views but were found not enough to provide any suitable predictive model. Though with the help of citrix server the access to the database of the Gothenburg trams was later permitted, it unfortunately had no impact on the result of the project.

The two sets of data sets (component history and work order history) had multiple variables of which only a few were relevant. The data in both the tables weren't clean and data preparation had to follow. Plus with the business logic, new set of variables had to be created

as well. The two data sets had no data field that was common to them, so tracing back to the other table would turn out to be a problem. So it can be assumed that the data from the two tables were independent of each other. The variables are explained as below:

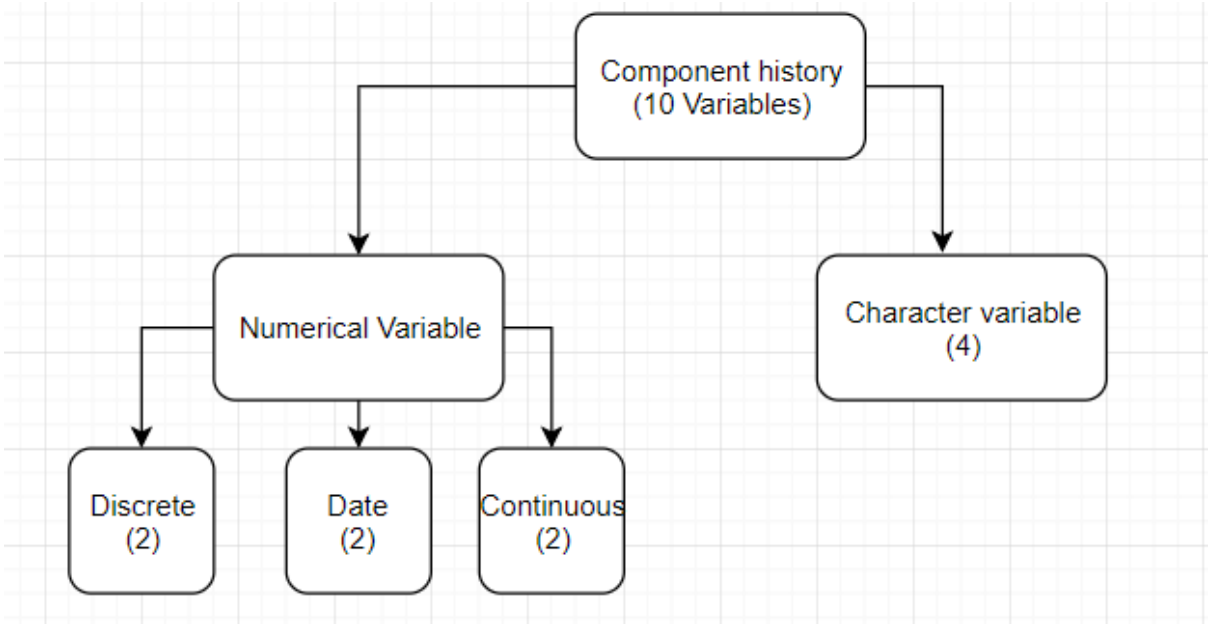


Figure 13: Component order history variables

Suhv_kopid	Suhv_kopdesc	Suhv_vagnid	Suhv_vagndesc	suhv_fromdate	suhv_todate	suhv_km	total_days	tram_type	komponent

Figure 14: Different column names of Component order history

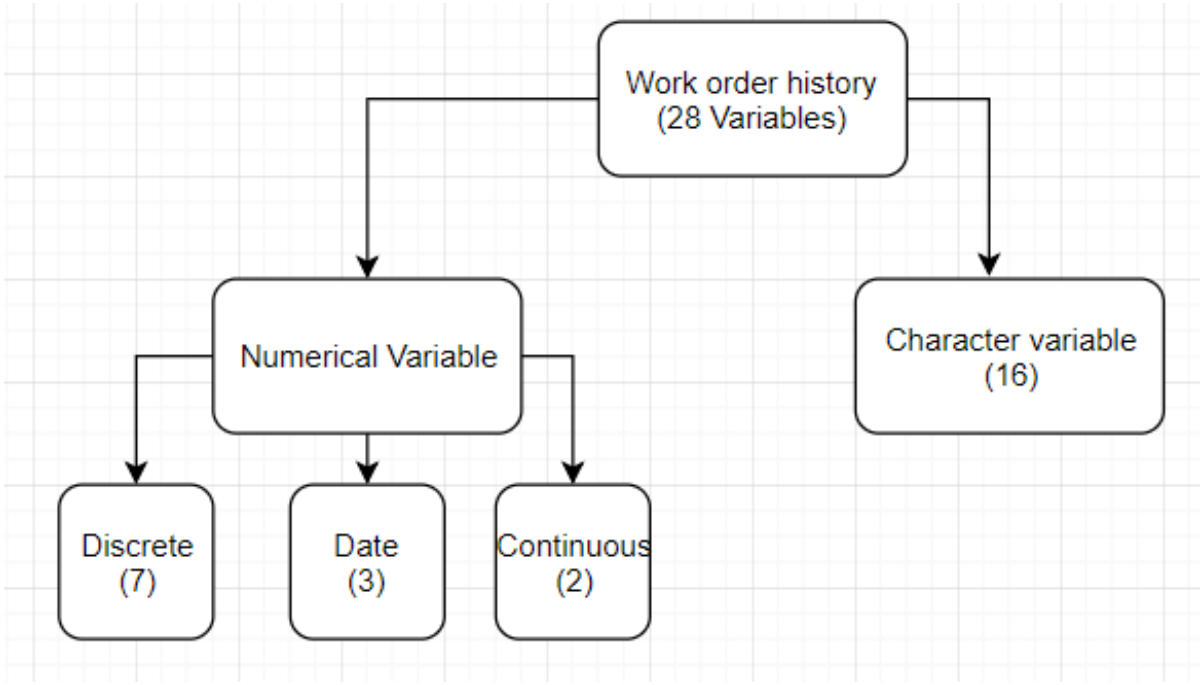


Figure 15: Work order history variables

Figure 16: Different Column names of work order history

It becomes quite evident that a good predictive models which requires the essence of continuous variables is missing in the data set provided. Nonetheless, there are plethora of problems that are based on the logistic regression (which makes use of discrete variables) and many text based algorithms that cold be written. But both of them were beyond the scope for the project and all the three set of segments had lot of NA values and blank values in them too.

4.3 Data preparation

The data preparation part for most of the projects accounts for the major chunk of the time. In the data set that was provided, both the datasets namely the component history and the work order history had multiple values which couldn't be utilised (NAs and white spaces). The discussion regarding the insertion of average values of kilometres and days to be substituted in the blank column was unfruitful too, therefore these data rows had to be deleted. As far as the pre-processing is concerned, there were a few business cases applied such as the tram serial number conversion to the tram types, time stamp formatting etc

4.4 Data description and investigation

The next step after cleaning the dataset was to analyse the existing data set in a descriptive format. After the missing value treatment ,outlier treatment and removal of values that are illogical the following set of frequency table was obtained from both the tables. Under this section the descriptive results of these analysis are presented.

The following table gives a brief insight of the data set of the different set of trams across the two tables. Note these are the data sets which were cleaned, i.e. fit for further processing of the dataset for modelling and analytical description.

Tram Types	Component order history	Work order history
M28	468	15609
M29	723	20788
M31	2364	62026
M32	1066	48029

Table 1: Brief summary of datasets

4.4.1 Data Investigation of work order history

The above table clearly shows that the work order history seems to have a lot of data points remaining after the data cleaning part. The fact of the matter was most of the data points were repeated. For example when a tram was taken to maintenance 3-4 data points regarding different changes done to the tram was noted down. This was apparent as the event code

registered for the same tram had duplicating values but were unique in terms of work done on the trams. The key issue with this data set was the inability to find out as to when a data was entered regarding a maintenance done on the tram, was the maintenance or repair done immediately after the previous maintenance/repair or was it due in the long run? The maintenance cycles couldn't be traced.(more in the discussion). The time difference between the reported event and the finished event was erroneous in most occasions too. Some of the trams studies in this table covered 800+Kms/day which doesn't make business sense. The correlation between the concerned person in the repair and the failure was ruled out due to lack of statistical evidence. In short a predictive model based on the set of variables provided and a cleaned data was simply insufficient and the models couldn't be statistically proven having very low p-values when considered for hypothesis testing method.

4.4.2 Data Investigation of component order history

Component order history dataset is a data set of primary concern. The dataset contained the data of each component being replaced after every maintenance and repair schedule. The correlation between the number of days travelled and the total distance could be built using this data set. Unlike the work order history, most of the data points in this table were accurate and a descriptive model could be built on this. A predictive model which could be built with the given variables had very high p value but gave no contribution to the business case that was developed. The scope of the project had to be further reduced to analysing the key components that caused the repeated failures in each tram types

4.5 Evaluation

The Evaluation section of the current chapter deals with the evaluation of the descriptive analysis carried out. It also looks at some of the predictive models that were built but couldn't be used in an effective manner due to low statistical evidence or even due to having no significant contribution to the business objectives that were defined. The results cover a few spectrum including the overall descriptive analysis, component based analysis and basic statistical description for each of the critical components studied.

The first step in the process was to analyse the data which had been cleaned and which had complied to all the business models that were discussed.

4.5.1 An overview of the data description

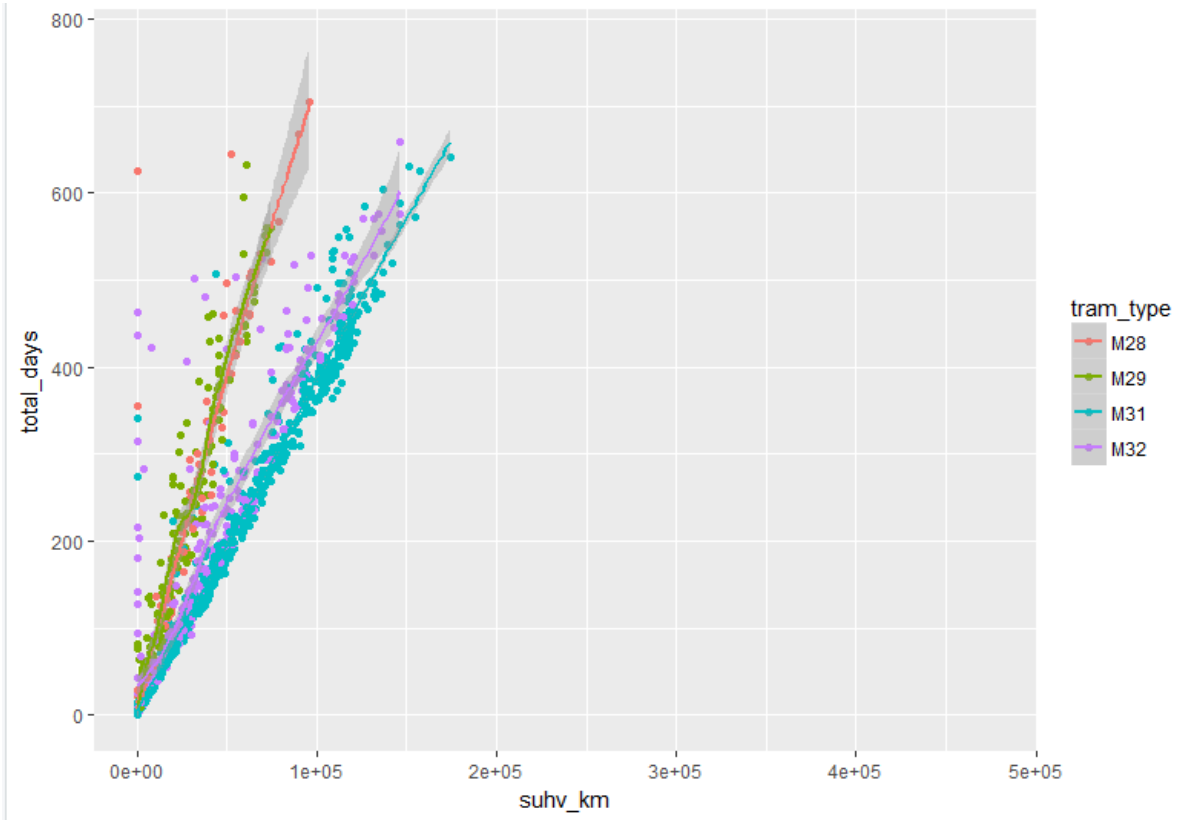


Figure 17: Days travelled vs Distance covered for all tram types

A brief exploration of the data helped fetch the above graph. All the graphs in the following section were plotted in R using the ggplot2 package. Analysing the component level data was a key study in the descriptive modelling. The plot of total days (defined as the date wherein the component was last replaced in the tram) and the total days it ran provided a great deal of insights. The pattern followed for the trams were all in a linear structure, with only the slopes of the graphs being varied. The corresponding lines for each of the tram types show the regression lines and the grey regions indicated the confidence interval of the region which is 95%CI. It can be observed from the above graph that the M31 and the M32 trams have a large data points making their regression lines more statistically accurate. The next step in the process involved analysing on the tram basis to ascertain if there was a correlation between the damage caused and the total time. There is a catch to the total days column, wherein the days can also include the durations the tram was basically lying idle and also on those occasions where it covered very less distance. The following graphs on provide an in-depth insights between the correlation of the every component being reported as damaged with respect to the total duration and total kilometres traversed. These graphs were obtained to get an overview if the damages do follow a certain trend and whether they could be explained by a mathematical equation.

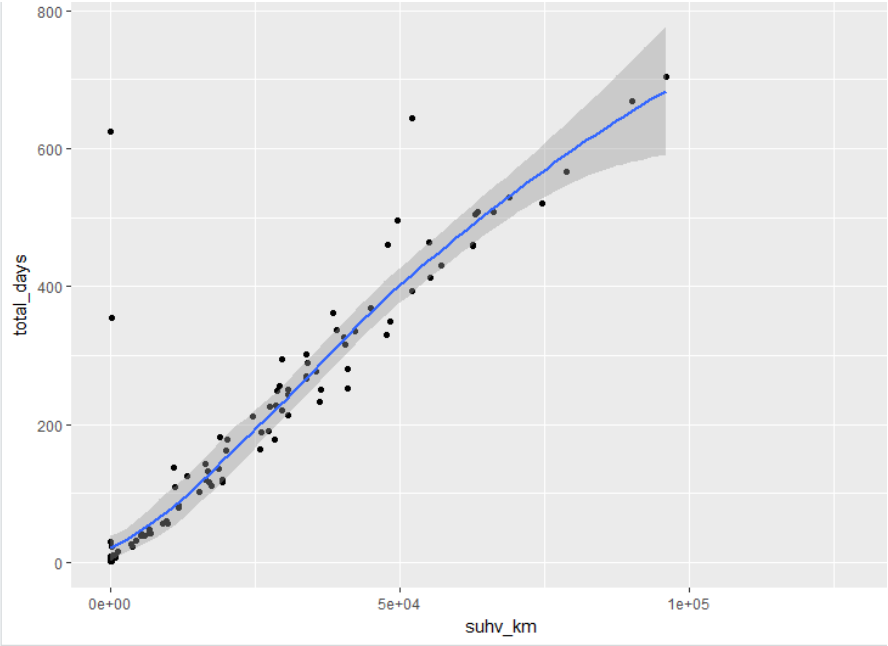


Figure 18: M28 Trams

The descriptive analysis of components considering the days between the maintenance/replacement with respect to the total distance covered in the same duration The regression equation is *y=-989+115*x.* It has a r-squared value of 0.859 which is a decent fit.

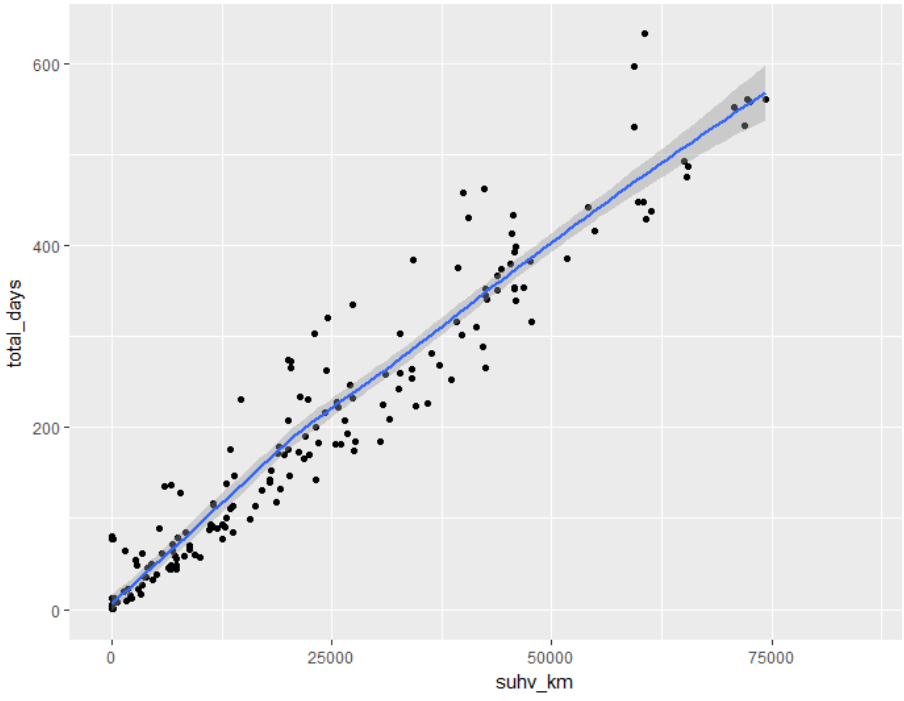


Figure 19: M29 Trams

The regression equation is *y=-244+118*x.* It has a r-squared value of 0.927 which is a good fit.

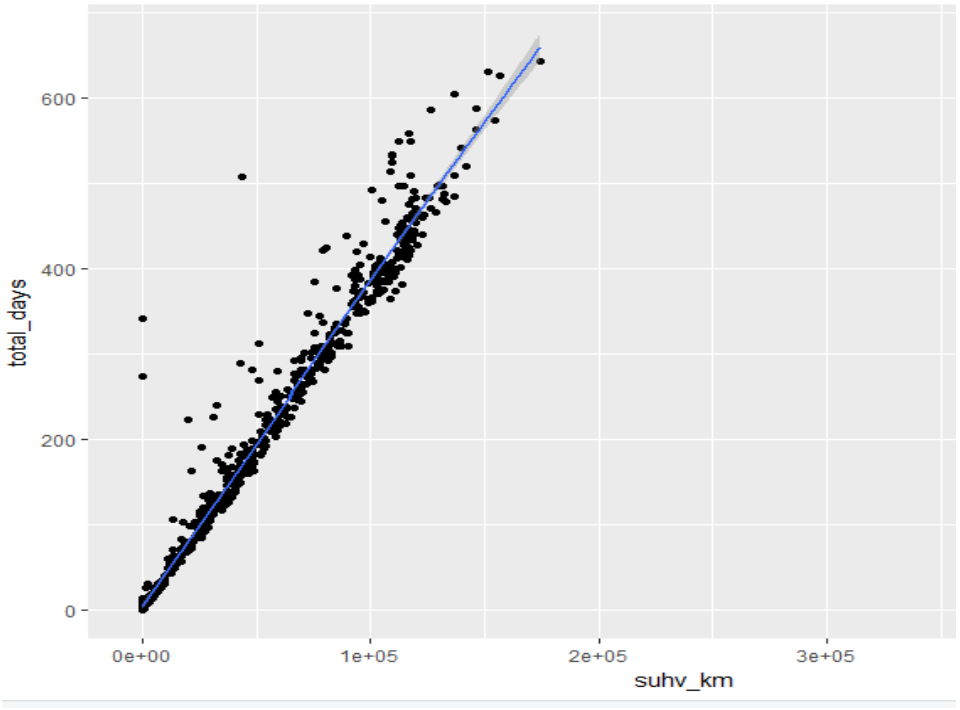


Figure 20: M31 Trams

The regression equation is *y=-754+253*x*. It has a r-squared value of 0.966 which is an excellent fit!

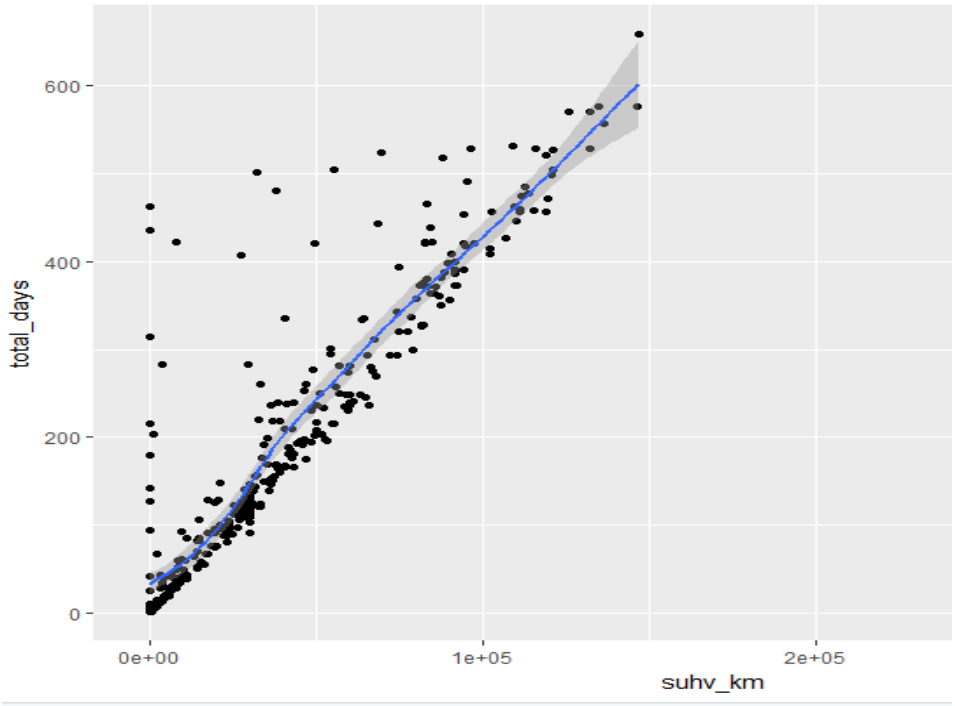


Figure 21: M32 Trams

The regression equation is *y=-189+211*x*. It has a r-squared value of 0.847 which is a good fit.

The above graphs which had a really good fit models could have been used to build a perfect predictive model. But the scope of the initial investigation was to find out if there are any hidden trends in the data set. The author informed the concerned people about the behaviour of each set of trams but it garnered no attention. Primarily being the total days vs the duration graphs gives no insight about the amount of time the tram was idle. There could be other components such as a repeated failure of a particular component that causes the breakdown to occur and thereby having to take the tram for a maintenance. The graphs although provide a good information about how each of the trams and their all components show a mathematical pattern but was of no significant business importance. Therefore the scope was narrowed down further to study on the component-tram basis level.

4.5.2 Study of components

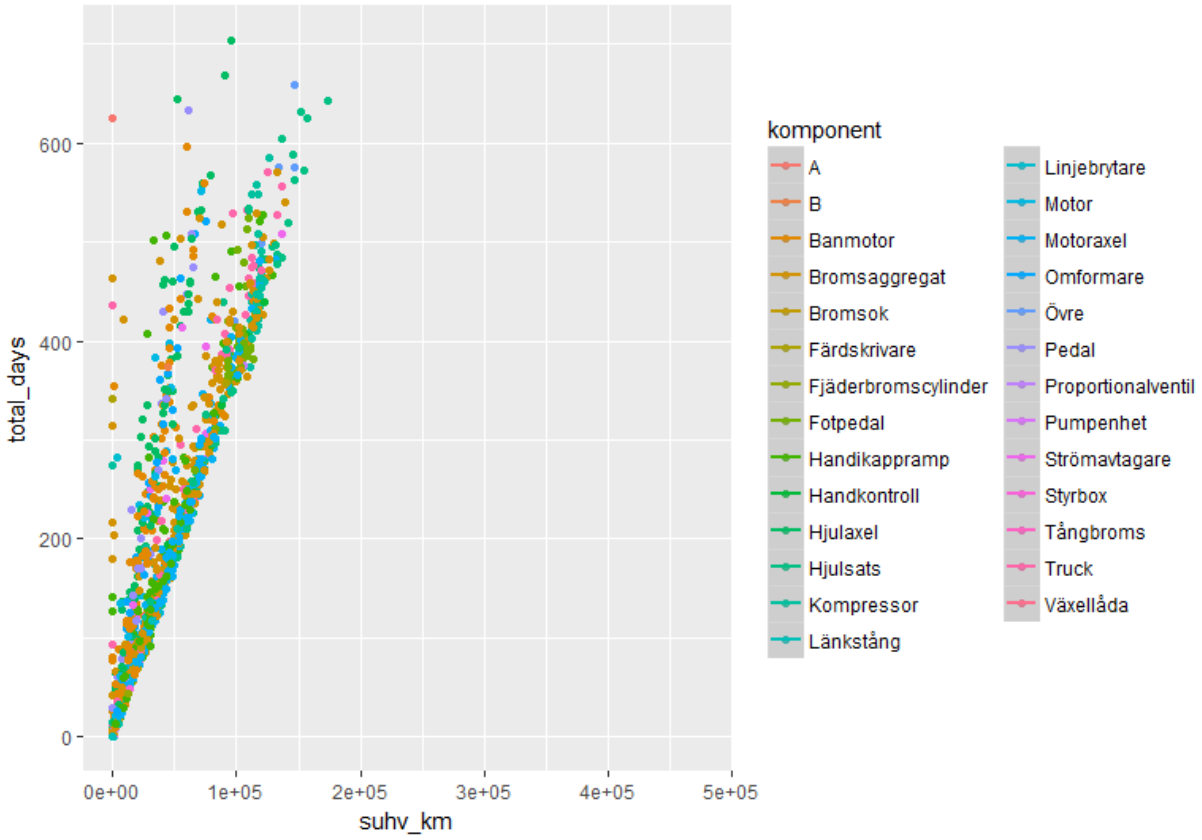


Figure 22: Analysis using the components

As is clear with the graph plot above, it can be observed that different set of components show a multiple degree of trends. Some components tend to fail at the initial stages and others later in the running stage. Majority of the components tend to get replaced at 100000 Km range which by the pareto principle account for the total 80% of the overall components. The key issue concerning the study of the components on the overall level was to ascertain if any particular type of components tend to get replaced at high frequency that is have a very high failure rate. But the descriptive analysis of all the components studied as above shows no such inference. It was again decided upon to study on the individual tram basis later looking at the above output, and the a result is shown in the next graph plotted.

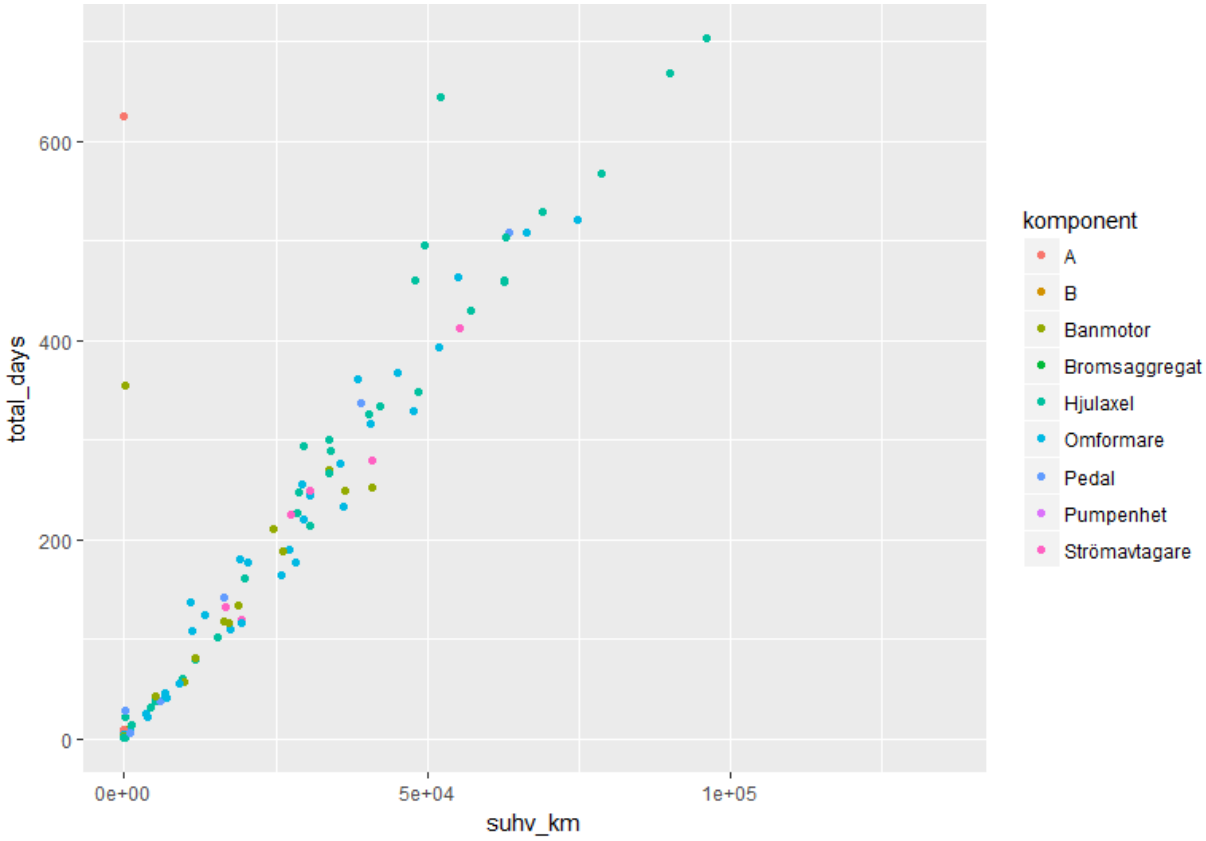
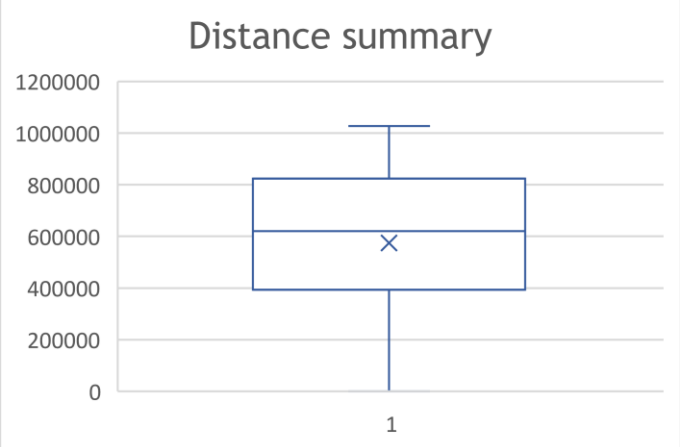


Figure 23: Scatterplot for M28 trams

Analysing on the component level basis on the individual tram series had to be the next task to analyse if there is a underlying pareto principle 20% of the components failing for 80% of the time) underlying the maintenance systems. It proved to be indeed a positive result and was in compliant to the business objectives that were defined initially at the start of this analysis. The critical components that were identified were Wheel axle, Motor, Motor Axle and Brake assembly for M28, M29, M31 and M32 respectively.

4.5.1 Critical Components on Tram classes

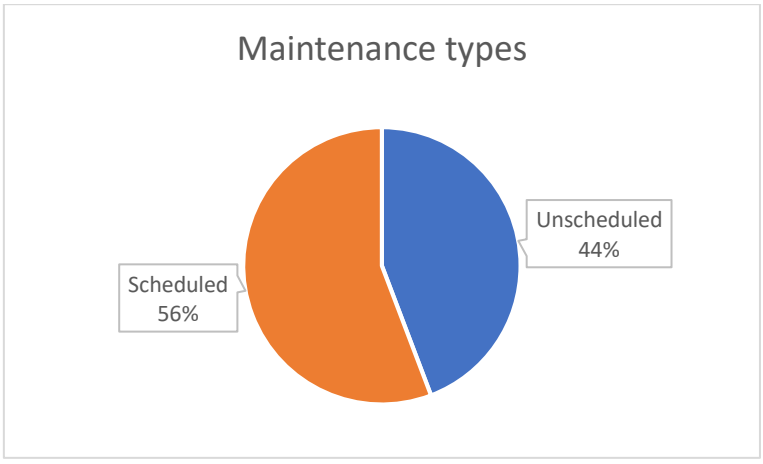
4.5.1.1 M28- Wheel Axle

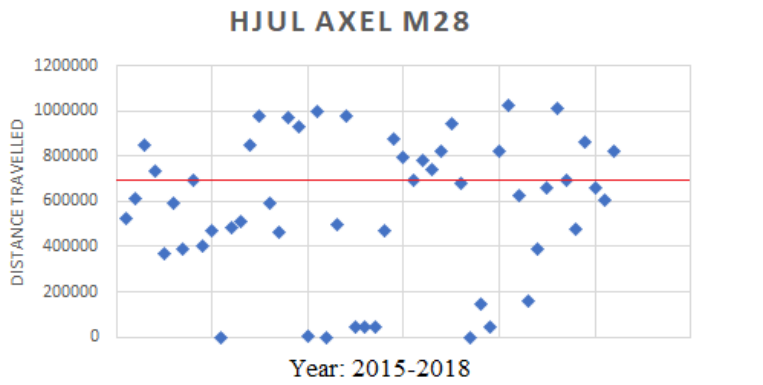


Feature	Distance (km)
Cut off	700000
Mean	574171
Median	620057

Wheel axle was found out to be the most frequently occurring component which needed a replacement in the M28 set of trams. The cut off distance is 700000 Km. In other words it's the *distance travelled since the last revision.* The cut off distance is defined as the minimum distance which a component has to run before which it can be replaced with a margin of +- 20000 Km. It can be seen in the above box plot and the chart beside that which states clearly that the average distance run by the component is not close to the 700000 Km mark. The distance is set by the Göteborgs Spårvägar

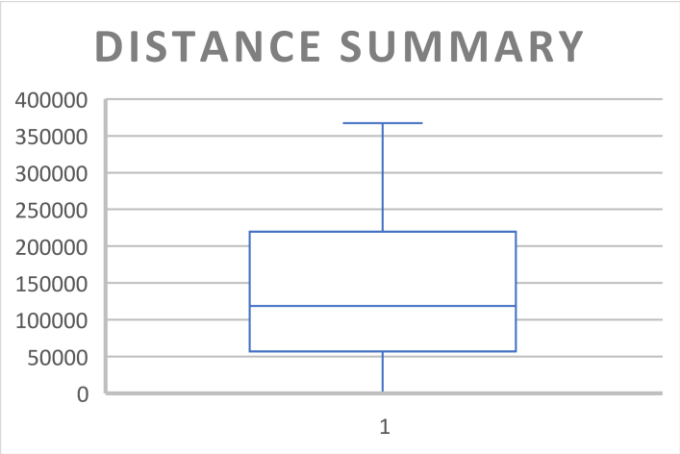
considering the total lifespan and the operating condition of the equipment. If the component has served a total distance of 700000 Km with an allowance of 20000 Km, it can then be sent to the maintenance system to be replaced. An unscheduled maintenance is observed when it doesn't fulfil the criteria and has to be taken for servicing well before the expected limit is reached. The pie chart and the scatter plot show the overall maintenance activities carried out for the given component.





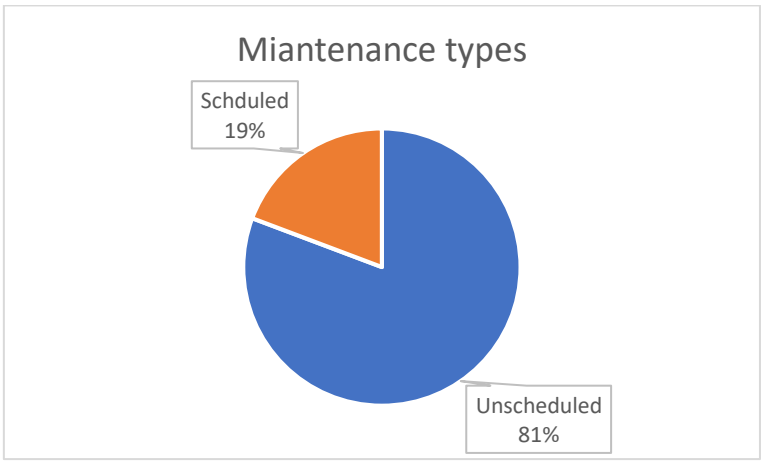
As can be seen with the pie-chart, 44% of the components are replaced in a unscheduled manner which again will lead to increase in overall costs as discussed earlier. The scatter plot on the chart above shows the instances of each replacement of the wheel axle made during the years 2015-2018. As can be seen there are a lot of points scattered well below the cut off mark of 700000 Km. The data from 2015 was included too in all the critical components as there were quite a few points in the last 2 years when looked at the component specific level. The reason for not having built a predictive model shall be explained later. The following parts will contain the same set of observations that were carried out on other critical components on different classes of trams as well.

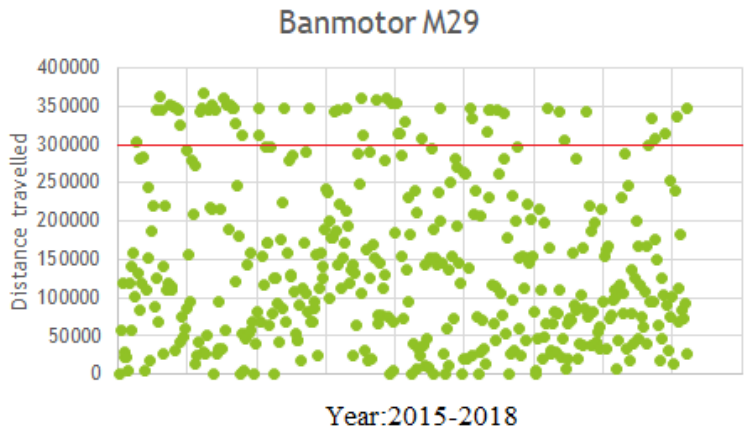
4.5.1.2 M29- Propulsion motor (Motor)



Feature	Distance (km)
Cut off	700000
Mean	146032
Median	118712

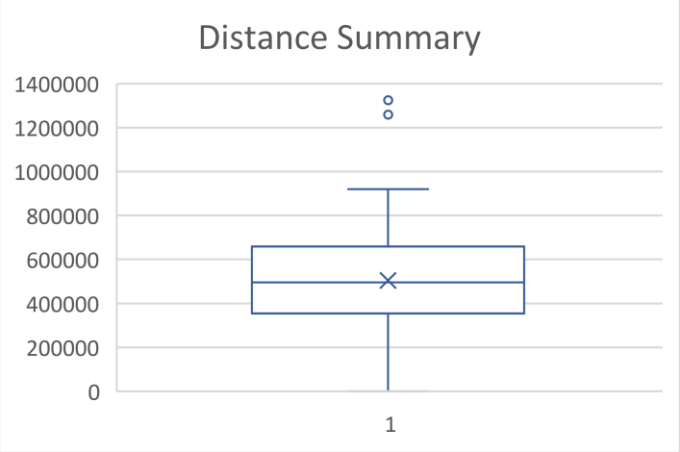
The propulsion motor of tram M29 was found out to be the component that had failures on most of the occasions. It can be observed that the cut off seems to be optimistic value for the mean and the median values so obtained. The values are nowhere close to the cut off value of 700000 Km and is a serious matter of concern. As expected there were quite a large number of replacements done for the motor in the last three years. The overall summary of the components are given below.





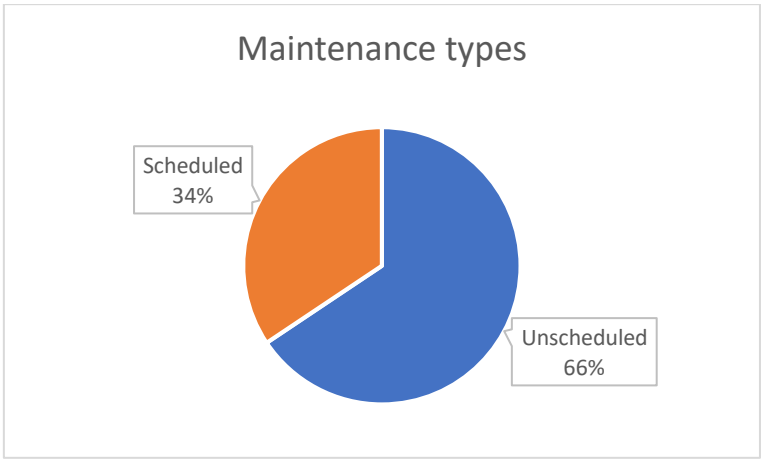
As is evident from the pie chart and the scatter plot, there needs to be an efficient predictive model that needs to be built which it reduce the overall maintenance cost for this particular component. 81% of the maintenance which is unscheduled is a huge loss of resources in terms of both time and the money.

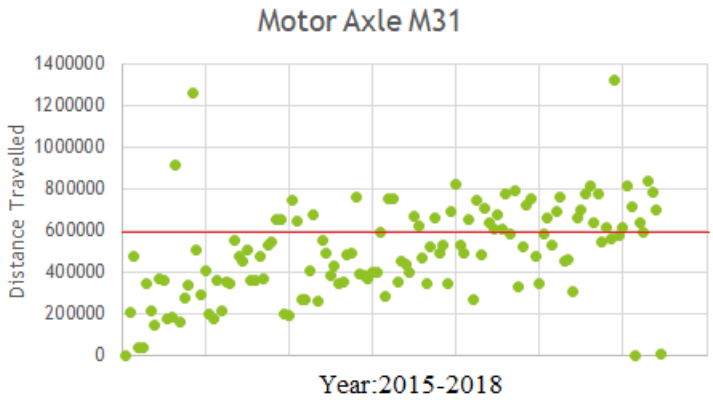
4.5.1.3 M31- Motor Axle



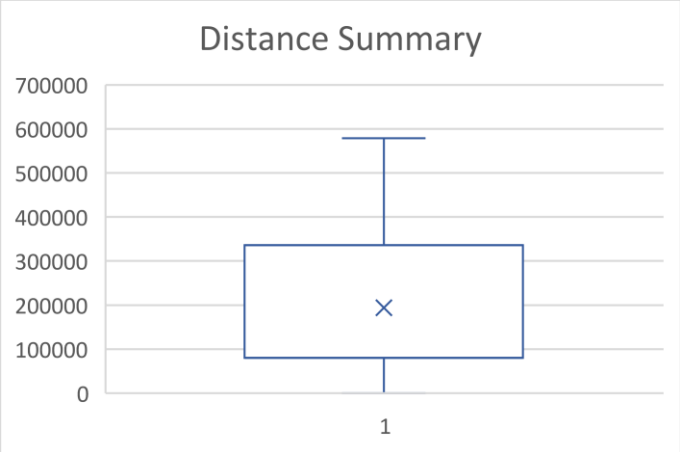
Feature	Distance (km)
Cut off	700000
Mean	503924
Median	495093

The motor axle of tram M31 was found out to be the component that had failures on most of the occasions. It can be observed that the cut off value is 700000 Km. The mean and the median values of the components do not lie in the 20000 Km buffer that has been provided for the component, thereby raising an alarm for the maintenance division. There were high number of instances for the component replacements as seen with the plots below with 66% of the maintenance being unscheduled. The scatter plot has many instance below the cut off limit too.



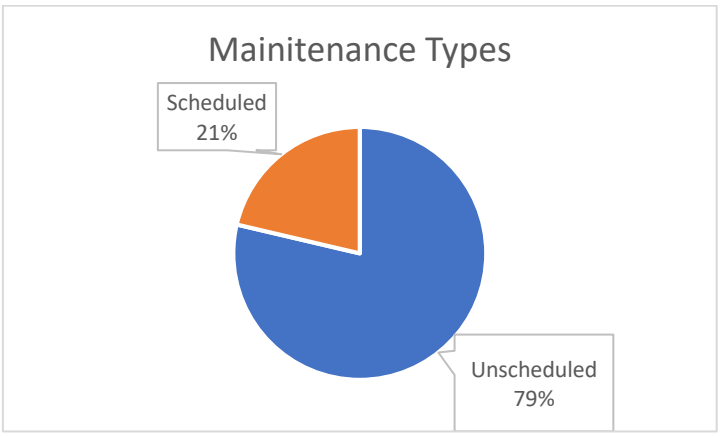


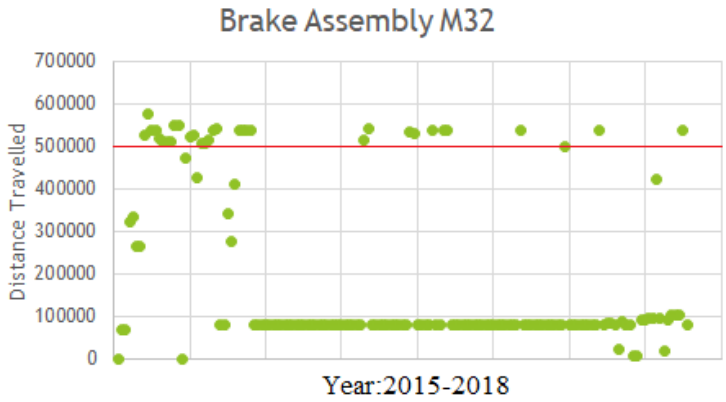
4.5.1.4 M32- Brake assembly



Feature	Distance (km)
Cut off	500000
Mean	193582
Median	80000

The Brake assembly of tram M31 was found out to be the component that had failures on most of the occasions. It can be observed that the cut off value is 500000 Km. The mean and the median values of the components do not lie in the 20000 Km buffer that has been provided for the component and can also be observed that the median value is very low which reflects in the bar plot as the median line is not plotted since it lies outside the box. This is a matter of serious concern for the maintenance department. There were high number of instances for the component replacements as seen with the plots below with 79% of the maintenance being unscheduled. The scatter plot has many instance below the cut off limit too. The number of repeated failure occurrences that occur well below the 100000 Km range is clearly shown in the scatter plot below thus justifying the box plot values.





4.6 Literature evaluation

The second research question dealt with knowing what the gaps are in the literature and what the author faced in the real world scenario. The following table gives an illustration of the same

Literature	Data	Gap
Big data solutions designed to enhance travel and business Experience (1)	When looking at the maintenance database, it looked otherwise	Customer centric solutions are executed well but not the maintenance solutions
Estimation of RUL by regression analysis over historical data (2)	Many regression models were explored but with very low p-values	The quality of data plays a key role in model building
Usage of SVM (Support Vector Machine) technique for alarm prediction in railways (1)	Can be used effectively for large data set and in case where there is a time series.	An effective technique cannot be built when data volume and quality differ.
Availability of structured, unstructured and semi-structured data in railways (1)	Availability of structured and unstructured data in smaller quantity especially after cleaning	Data quality is often skipped in the literature.
Multiple data modules carrying different set of information (1)	Limited data module namely component and work order history	Infrastructural issues in the smaller tram companies compared to the larger railway network
Multiple Data layers for IoT implementation across the railway maintenance system (3)	Unclear of the data layers found in the SQL database in the system	A known hierarchal structure helps in smoother implementation of IoT
A seamless integration between the working infrastructure (live) and the cloud databases (1)	A discontinuity among the different modules of the data.	Modernization of the infrastructure differs across multiple regions.

Unable to collate multiple data streams and take a good business decisions.	Similar scenario wherein the data is being just collected and stored	Similar in terms but quantity varied a lot in the trams project
No literature dealt with work order history	Data that deals with the work order history of the maintenance activities	A lot needs to be done in analysing the discrete data coming out from work order history

Although multiple literatures were considered, it was the three key papers that shed light on the maintenance scenario. The numbers indicated in brackets in the above paper are from the following research papers.

1) Thaduri et al., 2015
2) Fumeo et al., 2015
3) Zhang, 2012

5
Discussion

The current chapter covers both the successful and unsuccessful results of the project, attempting to explain the causes and what can be gained from the results. The chapter is divided in two sections, understanding and insights from data set, future scope and benefits gained from the project.

5.1 Understanding and Insights from Discussions, Data set and Literature review

The initial project definition and definition of the scope at the earlier stages were quite exciting since there was a lot at stake. Though the company did get a better insight and a overall view of what are the critical component and the overall damage pattern of the multiple component, a lot could have been achieved under the project. The initial discussions held with multiple stakeholders in the company had many interesting ideas and issues that had to be dealt with. Most of the descriptive part was successfully executed and some of the predictive models failed. The initial discussions that were held were all under the assumption that the company had a lot of data and it was a matter of extracting it and utilization. The initial discussions never dealt with the quality aspects of the data and was assumed that the system and implementation was smooth. The initial discussions also suggested that the company possessed large amount of structured data which was proved otherwise. The data sets, both: Component and work order history datasets were provided with a purpose for building a strong predictive models to have a robust algorithm to predict a maintenance as an when it is necessary as suggested by the initial meetings but could unfortunately be not done so. The company did posses a certain data structure as was evident with the initial meetings but data sets from the work order history proved difficult to have any solid regression models be built.

The results were mainly descriptive. The key issue for the current status was the data set. The component order history had meaningful contents but were simply not enough to build a predictive model due to lack of sufficient data, unclean data which led to very few data in the end for any algorithm to be trained. The results for building a linear and multiple regression model (built in R) was not shown because of abysmal p values for the variables obtained in building so. The p values for such models were in the range 0.2-0.35 which are extremely unsatisfactory to be even built in the first place and therefore the process was skipped. The work order history data on the other hand provided no real insights. Apart from having unclear shop-floor level data, the data regarding actual schedule, time between the scheduled maintenance and the current maintenance cycle was completely missing in the list. Added to the above concern data regarding the names of operators having carried out a maintenance on a tram, date columns that were left blank added more to the woes as the author couldn’t make

use of such data in order to build any kind of model. Had there been a time series based maintenance model a robust ARIMA model could have been built and with the help of a stable time series data a machine learning model could be built as well.

There were multiple instances when the results that would turn up had no correlation with the overall objective that was defined. On some occasions there were the results from the missing data (with the total distance and date being provided) would give a result of trams running more that a 1000 km a day and made no real business context. Using the average values on the NA values in dates and distances column would only make the matters worse by introducing more outliers in the data set which had to be cleaned all over again. In some of the components the mean and the median distance was distant from the cut-off value that was accepted as a standard by the maintenance team. All these unexpected insights and results could only be obtained due to the CRISP methodology that was deployed. Understanding the business context, preparation and analysis of the data gave multiple insights which proved to be a very important part of the learning process of the project. It's because of the methodology that was adopted, relevant insights and suggestions regarding the next step in the project could be gathered.

The literature reviews on most occasions provided little help to tackle the issues in the current project. As pointed out in the results section there was a lot of gap as far as the literature and this project is concerned. Though the methodology followed is a time tested strategy, it doesn't account the practical implications and gaps faced while executing the project. The literatures in this field mainly talk about the execution of multiple algorithms in multiple ways given the fact that they all have a large amount of data to deal with. In places where the data is scarce and the amount of unstructured data is high, there were quite few literatures regarding the same. Some of the literatures also talk about how difficult it is to consider multiple data sources and take a good business decision, but none dwell into the topic of scarce data sources. There aren't many papers in the trams segment and most of them are catering to the smart maintenance in railways. The kind of data plays key roll too. In most of the literature the implementation of SVM algorithm is made possible thanks to a large amount of data and time series based data. The literatures that were browsed didn't dwell upon the work order history aspect of the maintenance workshops. So in short yes there are a lot of literature available for smart maintenance for railway systems having a larger data set to play with, but not in the trams segment.

5.2 Benefits and Future Scope

There are several key aspects where the company can benefit from the mistake and the learnings of the current project. The first of course is the knowledge gained by the author both in theoretical terms and the hands on experience on the data set which of course resulted in more drawbacks than the desired result itself. The project benefitted the stake holders too as they could realise their drawbacks and some basic insights that were covered during the course of thesis at the company. The project demonstrates that there's a considerable amount of work to be carried out at the maintenance department at the company. Given the way the data is stored in the database with key columns such as dates and kilometre values missing, it becomes essential to make them as a mandatory field during any data input by the concerned persons. The project must also demand for keeping a time based dataset of the components (preferably

on day basis), to have an effective time series model which can be built based on the component level data.

However, given the data set such as work order history, it becomes a arduous task to correlate the comments of the different users (Drivers and technicians) and many character level data columns such as the names provide no real consequences in building a dataset. IN order to build and execute a perfect predictive model, a lot of data needs to be generated and also cleaned. In the given project both of which were limited in quantity and as a result only descriptive analysis could be carried out. Application of predictive modelling and further implementing a machine learning level model needs plethora of data and the regression models built are supposed to have a decent accuracy of 0.5-0.7 % whose performance can be improved at a later stage by having multiple training instances. However this seems to be a herculean task given the difficulties faced during the project and the only advice at this stage would be to improve the data quality.

Apart from the data being made mandatory in certain columns, a new set of rules are needed to be brought out as well. The existing system makes it impossible to track the previous maintenance carried out and the current stage of the component condition. Having a stage based monitoring would do great deeds along with the introduction of day-level data for the critical or even the multiple individual components which would help the company track the point from which component started deteriorating and which would eventually lead to a breakdown. There needs to be a perfect coordination among multiple departments too staring from the personnel at the workshop to the team engineers working in the maintenance department.

As we are aware of the fact that the consumer loyalty and satisfaction is the key to sustenance of any business, any inconvenience caused to the passengers due to the aforementioned issues will lead to a sense of lot of dissatisfaction and mistrust which the company tries to avoid at best. This unplanned and unscheduled maintenance has key aspects of losing out on revenue in terms of losing out on revenue and the brand image value affected because of poor service provided.

6
Conclusion

It became quite evident that the effect of pareto principle could be applied in this project as well. That is 20% of the components causing 80% of the issues. Tram types M29 and M32 register close to 80% of unscheduled maintenance. The critical component of M32 tram type (brake assembly) runs at 16% of its desired capacity (cut-off distance). Descriptive models for the same were built and shown in the results section. Unfortunately with dismal p-values in the hypothesis testing and also while building a predictive modelling, the author could not build a stable predictive model. The machine learning process which had been though about at the start of the project as a key process that could be implemented by the students wishing to pursue this project could take some time to achieve. In order to train a predictive model it's usually recommended to take the 6 quarters (or 2 years) of data. But in the current project that was deemed unfit for training due to multiple issues.

Although the project did manage to get some of the business requirements that were set by the company, it couldn't unfortunately meet all of them. The study carried out acknowledged the fact that there seemed to be a strong correlation between the tram types and the type of components that are needed to be replaced. The project does justify the kind of research questions posed by the author. A lot of work is supposed to be done in the component order history database and the work order history database needs multiple changes. The insights drawn were desirable and answered a few of the business objectives placed but could not be sufficient enough for building a predictive model. The lack of availability of a predictive model at this stage of the project can only be attributed to the kind of database that exists in the system. Further changes are supposed to be brought about in order to build a platform for the machine learning take place. As far as the overall purpose was concerned, it has been met given the fact that some of it were descriptive in nature. But the quality of data unfortunately couldn't help the author to build a robust model. The first research question though broadly remains answered. The second research question has covered the aspect of the literature research undertaken in this field though most of them were as stated earlier in the railways segment where large amount of structured data was made available.

Lastly, free and opensource statistical languages like R and Python provide a plethora of libraries and packages for the machine learning platform. The descriptive models that were built in this project was done in R with the data being extracted from a SQL server. This is a great step ahead when the discussion regarding IoT and Industry 4.0 crops up. The client has got the basic infrastructure right but collating it in a better way and accommodating some changes as suggested by the author will help the company be better compatible to Industry 4.0 thereby ushering the era of Internet of things.

7

Bibliography

1. Anderson, R. T., & Neri, L. (Eds.). (2012). *Reliability-centered maintenance: management and engineering methods*. Springer Science & Business Media.
2. Andrew K.S. Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, 20(7):1483 – 1510, 2006.
3. Alsyouf, I. (2006). Measuring maintenance performance using a balanced scorecard approach. *Journal of Quality in Maintenance Engineering*, *12*(2), 133-149.
4. Ayodele, T. O. (2010). Types of machine learning algorithms. In *New advances in machine learning*. InTech.Campbell, J. D., & Reyes-Picknell, J. V. (2015). *Uptime: Strategies for excellence in maintenance management*. CRC Press.
5. Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.
6. Alänge, S., & Scheinberg, S. (2005). *Innovation Systems in Latin America: Examples from Honduras, Nicaragua and Bolivia*. External organization.
7. *Basic Statistical Concepts*, Department of Mathematics, Chalmers : http://www.math.chalmers.se/Stat/Grundutb/GU/MSA620/V08/BasicStatistical%20.pdfaccessed: Sep 1 2018
8. Bergman, B., & Klefsjö, B. (2010). *Quality from customer needs to customer satisfaction*. Studentlitteratur AB.
9. Boyce, C., & Neale, P. (2006). Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input.
10. Bryman, A., & Bell, E. (2015). *Business research methods*. Oxford University Press, USA.
11. Carnero, M. (2006). An evaluation system of the setting up of predictive maintenance programmes. *Reliability Engineering & System Safety*, *91*(8), 945-963.
12. Fumeo, E., Oneto, L., & Anguita, D. (2015). Condition based maintenance in railway transportation systems based on big data streaming analysis. *Procedia Computer Science*, *53*, 437-446.
13. Gubata, J. (2008). Just-in-time manufacturing. *Research Starters Business*, 1-8.
14. Gothenburg Trams : http://goteborgssparvagar.se/om-oss/var-flotta/ accessed: June 10 2018
15. Groba, C., Cech, S., Rosenthal, F., & Gossling, A. (2007, June). Architecture of a predictive maintenance framework. In *Computer Information Systems and Industrial Management Applications, 2007. CISIM'07. 6th International Conference on*(pp. 59-64). IEEE.

16. Harper, G., & Pickett, S. D. (2006). Methods for mining HTS data. *Drug Discovery Today*, *11*(15-16), 694-699.
17. Hashemian, H. M., & Bean, W. C. (2011). State-of-the-art predictive maintenance Techniques. *IEEE Transactions on Instrumentation and measurement*, *60*(10), 3480-3492.
18. Hjorth, U. (1980). A reliability distribution with increasing, decreasing, constant and bathtub-shaped failure rates. *Technometrics*, *22*(1), 99-107.
19. J., Alexander. (2017, March 13). A Cost-Benefit Case for Condition Monitoring. Retrieved June 21, 2018, from https://www.efficientplantmag.com/2017/03/cost-benefit-case-condition-monitoring/
20. Karim, R., Westerberg, J., Galar, D., & Kumar, U. (2016). Maintenance analytics–the new know in maintenance. *IFAC-PapersOnLine*, *49*(28), 214-219.
21. Kobbacy, K. A. H., & Murthy, D. P. (Eds.). (2008). *Complex system maintenance handbook*. Springer Science & Business Media.
22. Kumar, U., Galar, D., Parida, A., Stenström, C., & Berges, L. (2013). Maintenance performance metrics: a state-of-the-art review. *Journal of Quality in Maintenance Engineering*, *19*(3), 233-277.
23. L. Zhang, X. Li, and J. Yu. A review of fault prognostics in condition based maintenance. *Proc. of SPIE*, 6357:635–752, Nov 2006.
24. Larry, T. (1995). Machinery oil analysis: Methods, automation and benefits(pp. 1–383). Park Ridge: Society of Tribologists and Lubrication Engineers
25. Lenz, B., & Barak, B. (2013, January). Data mining and support vector regression machine learning in semiconductor manufacturing to improve virtual metrology. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 3447-3456). IEEE.
26. Lienig, J., & Bruemmer, H. (2017). *Fundamentals of Electronic Systems Design*. Springer.
27. Makeham, W. M. (1860). On the law of mortality and construction of annuity tables. *Journal of the Institute of Actuaries*, *8*(6), 301-310.
28. Maletic, D, Maletic M, Al-Najjar, B & Gomišcek, B 2014, 'The role of maintenance in improving company's competitiveness and profitability', Journal of Manufacturing Technology Management, vol 25, no. 4, pp. 441-456.
29. Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(2), 137-166.
30. Marquez, F. P. G., Lewis, R. W., Tobias, A. M., & Roberts, C. (2008). Life cycle costs for railway condition monitoring. *Transportation Research Part E: Logistics and Transportation Review*, *44*(6), 1175-1187.
31. Mobley, R. K. (2002). *An introduction to predictive maintenance*. Butterworth-Heinemann.
32. Mobley, R. K. (2013). Eliminate These Obstacles Before You Implement Predictive Maintenance. Retrieved June 21, 2018, from https://www.lce.com/Eliminate-These-Obstacles-Before-You-Implement-Predictive-Maintenance-1328.html
33. Mostafa, Sherif & Lee, Sang-Heon & Dumrak, Jantanee & Chileshe, Nicholas & Soltan, Hassan. (2015). Lean thinking for a maintenance process. Production & Manufacturing Research: An Open Access Journal. 3. 236-272.

34. Muenchen, R. A. (2012). The popularity of data analysis software. *UR L http://r4stats. com/popularity*.
35. Nakajima, S. (1988). Introduction to TPM: Total Productive Maintenance (preventative maintenance series). *Hardcover. ISBN 0-91529-923-2*.
36. Otilia Elena Dragomir, Rafael Gouriveau, Florin Dragomir, Eugenia Minca,and Noureddine Zerhouni. Review of prognostic problem in condition-based maintenance. In *Control Conference (ECC), 2009 European*, pages 1587–1592.
37. IEEE, 2009.
38. Prajapati, A., Bechtel, J., & Ganesan, S. (2012). Condition based maintenance: a survey. *Journal of Quality in Maintenance Engineering*, *18*(4), 384-400.
39. Prabhuswamy, M. S., Nagesh, P., & Ravikumar, K. P. (2013). Statistical analysis and reliability estimation of total productive maintenance. *IUP Journal of Operations Management*, *12*(1), 7.
40. Salonen, A. (2011). *Strategic maintenance development in manufacturing industry* (Doctoral dissertation, Mälardalen University).
41. Samuel, A. L. (2000). Some studies in machine learning using the game of checkers.*IBM Journal of Research and Development*, *44*(1), 206-226.
42. Schwan, C. A. (1999). Introduction to Reliability Centered Maintenance. In *Proceedings of Reliability Centered Maintenance for Substations, Transmission and Distribution Conference* (pp. 3-12). Electric Utility Consultants, Inc..
43. Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, *5*(4), 13-22.
44. Siddiqui, A. W., & Ben-Daya, M. (2009). Reliability centered maintenance. In *Handbook of maintenance management and engineering* (pp. 397-415). Springer, London.
45. Smith, A. M., & Hinchcliffe, G. R. (2003). *RCM--Gateway to world class maintenance*. Elsevier.
46. Susto, G. A., Beghi, A., & De Luca, C. (2012). A predictive maintenance system for epitaxy processes based on filtering and prediction techniques. *IEEE Transactions on Semiconductor Manufacturing*, *25*(4), 638-649.
47. Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, *11*(3), 812-820.
48. Swedish Standard Institute 2001, 'SS-EN 13306 - *Maintenance terminology'*, Swedish Standard Institute, Stockholm.
49. Sweet, S. A., & Grace-Martin, K. (1999). *Data analysis with SPSS* (Vol. 1). Boston, MA: Allyn & Bacon.
50. Thaduri, A., Galar, D., & Kumar, U. (2015). Railway assets: A potential domain for big data analytics. *Procedia Computer Science*, *53*, 457-467.
51. TPM (Total Productive Maintenance). (n.d.). Retrieved from https://www.leanproduction.com/tpm.html, Accessed: 20 June 2018
52. Tsang, A. H. (2002). Strategic dimensions of maintenance management. *Journal of Quality in Maintenance Engineering*, *8*(1), 7-39.
53. Umiliacchi, P., Lane, D., Romano, F., & SpA, A. (2011, May). Predictive maintenance of railway subsystems using an Ontology based modelling approach. In *Proceedings of 9th world conference on railway research, May* (pp. 22-26).

54. Wang, J., Zhang, L., Duan, L., & Gao, R. X. (2017). A new paradigm of cloud-based predictive maintenance for intelligent manufacturing. *Journal of Intelligent Manufacturing*, *28*(5), 1125-1137.
55. Wienclaw, R. A. (2008). Operations and business process management. *EBSCO, Research Starters*, *5*.
56. Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39).
57. Ying Peng, Ming Dong, and Ming Jian Zuo. Current status of machine prognostics in condition-based maintenance: a review. *The International Journal of Advanced Manufacturing Technology*, 50(1):297–313, 2010.
58. Zhang, W. (2012, November). Study on internet of things application for high-speed train maintenance, repair and operation (MRO). In *Proceedings of the National Conference on Information Technology and Computer Science (CITCS 2012), Lanzhou, China* (pp. 16-18).