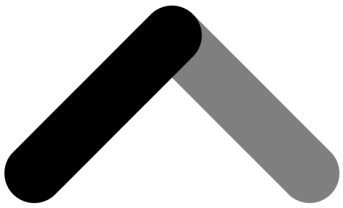


Optimisering av Datadriven Marknadsföring

CTRL DIGITAL - Data Hantering, Data Insamling och Maskininlärning för Google Analytics 4

Examensarbete inom högskoleprogrammet Datateknik

GABRIEL RASK THURIN, JONATHAN ROOTZÉN

INSTITUTIONEN FÖR DATA- OCH INFORMATIONSTEKNIK

CHALMERS TEKNISKA HÖGSKOLA

Göteborg 2025

www.chalmers.se

Examensarbete 2025

Optimisering av Datadriven Marknadsföring

CTRL DIGITAL - Data Hantering, Data Insamling och Maskininlärning för Google Analytics 4

GABRIEL RASK THURIN, JONATHAN ROOTZÉN

CHALMERS

Institutionen för Data- och informationsteknik
Chalmers Tekniska Högskola
Göteborg 2025

Optimisering av Datadriven Marknadsföring
CTRL DIGITAL - Data Hantering, Data Insamling och Maskininlärning för Google Analytics 4
GABRIEL RASK THURIN, JONATHAN ROOTZÉN

© GABRIEL RASK THURIN, JONATHAN ROOTZÉN, 2025.

Akademisk Handladare: Sakib Sistek, Chalmers, Data- och informationsteknik
Teknisk Handledare: Theodor Öberg, CTRL DIGITAL
Projekthandledare: Ted Solomon & Elin Svensson, CTRL DIGITAL
Examinator: Nicolas Smallbone, Chalmers, Data- och informationsteknik

Examensarbete 2025
Institutionen för Data- och informationsteknik
Chalmers Tekniska Högskola
SE-412 96 Göteborg
Telefon +46 31 772 1000

Omslagsbild: CTRLs logotyp.

Skriven i LaTeX
Göteborg 2025

Optimisering av Datadriven Marknadsföring
Examensarbete inom högskoleprogrammet Datateknik
Gabriel Rask Thurin, Jonathan Rootzén
Institutionen för Data
Chalmers Tekniska Högskola

Sammanfattning

Detta examensarbete bedrevs i samarbete med CTRL DIGITAL och syftade till att utveckla en analyslösning för att segmentera kunder baserat på beteendedata från Google Analytics 4 (GA4). Arbetet inleddes med att skapa en pipeline för att hämta historisk data via GA4:s API och lagra den i Google BigQuery. Datan strukturerades och berikades med information från CRM-system och externa annonseringskällor, varefter relevanta nyckeltal (KPI:er) definierades i dialog med verksamheten.
Med hjälp av en Recency, Frequency, Monetary (RFM) modell och KMeans-klustring genomfördes en prediktiv analys av kundbeteende, där tre kundsegment med varierande köpbeteende och affärsvärde identifierades. Segmenteringen baserades på parametrarna aktualitet, köpfrekvens och monetärt värde. Projektet resulterade i en strukturerad process för datahantering och kundanalys som kan återanvändas i liknande uppdrag.
Insikterna har presenterats för kundens mediebyrå och kan fungera som underlag för framtida riktade marknadsföringsinsatser.

Nyckelord: backfill, KPI, maskininlärning, Google Analytics, BigQuery, marknadsföring, datadriven, AI, prediktiv analys, CRM

Data Management and Machine Learning for Google Analytics 4
Bachelor's Thesis in Computer Science
Gabriel Rask Thurin, Jonathan Rootzén
Computer Science and Engineering
Chalmers University of Technology

Abstract

This thesis project was conducted in collaboration with CTRL DIGITAL and aimed to develop an analytics solution for segmenting customers based on behavioral data from Google Analytics 4 (GA4). The project began by building a pipeline to retrieve historical data through the GA4 API and store it in Google BigQuery. The data was structured and enriched with information from the client's CRM system and external advertising platforms. Key performance indicators (KPIs) were then defined in close collaboration with the business stakeholders.
Using a Recency, Frequency, and Monetary (RFM) model and KMeans clustering, the project conducted a predictive analysis to identify three customer segments with distinct purchasing behavior and business value. The segmentation was based on recency, frequency and monetary value. The outcome was a structured process for data handling and customer analysis, designed to be reusable in future projects and to support data-driven marketing decisions.
The insights have been presented to the client's media agency and serve as a foundation for future targeted marketing efforts.

Key-words: backfill, KPI, machine learning, Google Analytics 4, BigQuery, marketing, datadriven, AI, predictive analysis, CRM

Förord

Detta examensarbete har genomförts under vårterminen 2025 vid Chalmers tekniska högskola, i samarbete med konsultbyrån CTRL DIGITAL Och deras kund Brothers. Arbetet har gett oss möjlighet att fördjupa oss i datadriven kundanalys, med inriktning på klustringsmetoder, dataplattformar och affärsnära beslutsunderlag.

Vi vill rikta ett stort tack till vår handledare på Chalmers, Sakib Sistek, för stöd och konstruktiv återkoppling under arbetets gång. Vi vill även tacka handledarna på CTRL DIGITAL - Ted Solomon, Elin Svensson och Theodor Öberg - för insiktsfulla diskussioner, teknisk vägledning och förtroendet att genomföra ett projekt med verklig tillämpning.

Slutligen vill vi tacka alla medarbetare På CTRL DIGITAL som bidragit med sin tid, kunskap och trevliga stämning på kontoret.

Gabriel Rask Thurin, Jonathan Rootzén, Göteborg, maj 2025

Akronymer

Nedan listas i bokstavsordning de akronymer som kommer att användas i rapporten:

AI	Artificial Intelligence
AOV/AUV	Average Order/Unit Value
API	Application Programming Interface
Colab	Google Colabaratory
CR	Conversion Rate
CRM	Customer Relationship Management
CSV (file)	Comma-Separated Values File
GCP	Google Cloud Platform
GCF	Google Cloud Functions
KPI	Key Performance Indicators
ML	Machine Learning
Pseuduo ID	Pseudonym identifierare
RFM	Recency Frequency Monetary
ROAS	Return On Ad spend
SQL	Structured Query Language
WCSS	Within-Cluster Sum of Squares
RMSE	Root Mean Squared Error

Innehåll

Akronymer **x**

Figurer **xv**

Tabeller **xvii**

1 Inledning **1**
1.1 Bakgrund . . . 1
1.2 Syfte . . . 1
1.3 Mål . . . 2
1.3.1 Datainsamlings- och hanteringslösning . . . 2
1.3.2 Utveckling av en proof-of-concept för Maskininlärning . . . 2
1.3.3 Dokumentation och rekommendationer för vidareutveckling . . . 2
1.4 Avgränsningar . . . 2

2 Metod **3**
2.1 Insamling, lagring och bearbetning av historisk data . . . 3
2.2 Dataanalys och etablering av nyckeltalsstruktur (baseline) . . . 5
2.3 Tillämpning av maskininlärning för kundsegmentering . . . 5
2.4 Standardisering . . . 6
2.5 Etiska och hållbarhetsmässiga överväganden . . . 6

3 Teknisk Bakgrund **9**
3.1 Begrepp inom digital marknadsföring . . . 9
3.1.1 Key Performance Indicators . . . 9
3.2 Programspråk . . . 10
3.2.1 SQL . . . 11
3.2.2 Javascript . . . 11
3.2.3 Python . . . 11
3.3 Molntjänster och tekniska verktyg . . . 11
3.3.1 Google Analytics 4 . . . 11
3.3.2 BigQuery . . . 11
3.3.3 Google Cloud Functions . . . 12
3.3.4 Dataform . . . 12
3.4 Klustringsmetodik . . . 12
3.4.1 Recency, Frequency, Monetary-modellen . . . 12
3.4.2 KMeans-algoritmen . . . 13

3.4.3 Val av antal kluster 13

4 Genomförande **15**
4.1 Fas 1: Uppbyggnad av dataplattform 15
4.1.1 Backfill - Datainsamling och lagring 15
4.1.2 Tekniska utmaningar 15
4.1.3 Levererat system och datakvalitet 16
4.1.4 CRM- och annonsdata 16
4.2 Fas 2: Databearbetning och KPI-modellering 17
4.2.1 Tekniska utmaningar 17
4.2.2 Dataform skelett 17
4.2.3 KPI sammanställning 18
4.2.4 Annonsdata 18
4.2.5 Visualisering 18
4.3 Fas 3: Klustringanalys 18
4.4 Fas 4: Standardisering och avslut 18
4.5 Sammanfattning 19

5 Resultat **21**
5.1 Resultatsavgränsning 21
5.2 Fas 1: Datainsamling 21
5.3 Fas 2: KPI-modellering 22
5.4 Fas 3: Klustring 24
5.5 Fas 4: Standardisering och utvärdering 26
5.6 Sammanfattning av resultat 27

6 Diskussion **29**
6.1 Metodreflektion 29
6.2 Begränsningar i datakällor 30
6.3 Effekter på modellval och segmentering 31
6.4 Etiska och hållbarhetsmässiga överväganden 31
6.5 Förslag på framtida förbättringar 33

7 Slutsats **35**

Bibliography **37**

Figurer

2.1 Historisk data med viss censur p.g.a. NDA. 4
2.2 Exempel på ett GA4-event med tillhörande parametrar (modifierad) . 4

3.1 Exempel på KPI-träd 9

4.1 Schematisk översikt över datapipelinen från GA4 till BigQuery. 16

5.1 Käll- och mellanlagringstabeller i Dataform 22
5.2 Bearbetning- och slutdatatabeller i Dataform 23
5.3 Kundtyper över tid i Colab 24
5.4 källprestation 24
5.5 Visualisering av klusterprofiler för den valda modellen med tre kluster ($k = 3$). 26

Tabeller

5.1 Exempel på dataposter från backfill-processen 22
5.2 Sammanställning av tre utvalda resultat för olika k-värden vid klustring 25

1
Inledning

I takt med att digital marknadsföring blir alltmer datadriven ökar behovet av att kunna samla in, hantera och analysera stora mängder användardata på ett effektivt sätt. Google Analytics 4 (GA4) är ett verktyg som möter detta behov, då det möjliggör spårning och bättre förståelse av användarbeteenden, försäljningsstatistik och kampanjers prestationer. Som komplement finns även BigQuery, BigQuery tillåter djupare analys med Structured Query Language (SQL) och stödjer processflöden för databehandling med verktyget *Dataform*.
Trots dessa verktyg finns det utmaningar med att integrera GA4-data i BigQuery och att på ett strukturerat sätt använda informationen för att identifiera värdefulla målgruppssegment. En lösning är att kombinera datainsamling med prediktiv analys genom maskininlärning, vilket har visat goda resultat när det gäller att förbättra kundretention, identifiera högriskkunder och skapa mer riktade och individualisera kampanjer [1]. Detta visar på potentialen i att förena spårningsdata från GA4 med strukturerad analys i BigQuery, något som undersöks vidare i detta projekt.

1.1 Bakgrund

CTRL DIGITAL är en konsultbyrå inom dataanalys och mätning med fokus på att hjälpa företag fatta välgrundade beslut i en allt mer integritetsstyrd marknadsföringsmiljö. Företaget beskriver sig som tekniskt drivna med djup kompetens inom datainsamling, molnteknologi och aktivering av data för affärstillväxt [2].
CTRL DIGITAL verkar i en tid där ökade krav på dataskydd och spårningsbegränsningar har försvårat möjligheten att bedriva effektiv och personaliserad marknadsföring. För att möta dessa utmaningar har företaget utvecklat lösningar som bygger på en modern förståelse för hur data samlas in, lagras och används. De plattformar som har centralt fokus är GA4 och BigQuery.
Examensarbetet är en del av detta sammanhang och har syftat till att bygga en teknisk lösning för strukturerad datahantering samt undersöka hur maskinlärning kan användas för att identifiera affärsrelevanta kundsegment.

1.2 Syfte

Syftet med detta examensarbete är att utveckla en lösning som integrerar GA4 API med BigQuery, vilket tillåter hämtning av både historiska och realtidsdata. Projektet kommer även undersöka tillämpningen av maskininlärning för att identifiera

värdefulla målgruppssegment och därigenom stödja utvecklingen av datadriven och differentierad marknadsföring.

1.3 Mål

Målet med detta examensarbete är att utveckla en funktionell och återanvändbar lösning för att hantera data från GA4 i BigQuery samt att undersöka möjligheten att använda maskininlärning för att identifiera värdefulla målgruppssegment. För att uppnå detta definieras följande delmål.

1.3.1 Datainsamlings- och hanteringslösning

En komplett datahanteringspipeline ska utvecklas. Lösningen ska kunna hämta, lagra och hantera historisk- och realtidsdata samt kunna transformera insamlad data till en analysfärdig struktur.

1.3.2 Utveckling av en proof-of-concept för Maskininlärning

En proof-of-concept ska tas fram där maskininlärningsmodeller appliceras på verklig data för att analysera och identifiera värdefulla målgruppssegment baserat på jämförelser via KPI-analys.

1.3.3 Dokumentation och rekommendationer för vidareutveckling

Arbetet ska resultera i en detaljerad dokumentation av lösningens arkitektur, implementation och prestanda. Rekommendationer för vidare utveckling och optimering ska formuleras, med betydande fokus på hur datahantering och maskininlärningsmodeller kan förbättras för datadriven marknadsföring.

1.4 Avgränsningar

Projektet genomförs inom ramen för examenarbetets tidsbegränsning och kan därför inte inkludera långsiktig utvärdering. Endast data från GA4 och Customer Relationship Management-data (CRM) från kundens interna datakälla används, och inga andra källór kommer att inkluderas i analysen. Lösningen utvecklas och testas även enbart inom Google Cloud Platform (GCP) och Google Colab (Colab) och kommer inte att använda några andra molntjänster.

2
Metod

För att genomföra detta examensarbete har en metodik tagits fram som innefattar fyra huvudsakliga moment, så kallade Faser.

- **Fas 1:** insamling, lagring och bearbetning av historisk data.
- **Fas 2:** dataanalys och etablering av nyckeltalsstruktur (baseline).
- **Fas 3:** tillämpning av maskininlärning för kundsegmentering.
- **Fas 4:** standardisering av den färdigställda produkten samt utvärdering av modellresultat och förslag på fortsatt arbete.

Syftet med metoden är att utforma en datadriven lösning som kan vägleda marknadsföringsinsatser utifrån faktisk användardata. Metodvalen baseras på relevanta tekniska och praktiska förutsättningar inom GCP, samt i samråd med projektägaren CTRL DIGITAL.
Valet av verktyg har inte enbart styrts av tillgång utan också av ett tydligt mål att skapa en lösning som kan återanvändas i liknande projekt. Genom att arbeta med etablerade plattformar som GA4 och BigQuery kan metodiken enklare tillämpas i andra organisationer med liknande behov. Det bidrar till att minska både kostnader och tidsåtgång, vilket ökar möjligheterna att använda redan existerande data på ett mer affärsdrivet sätt.

2.1 Insamling, lagring och bearbetning av historisk data

För att bygga en datadriven grund genomförs en process för att hämta, strukturera och analysera användardata från GA4. Arbetet inleds med att konfigurera en datainsamlingspipeline som kopplar GA4 Data API till Google BigQuery. Efter hämtning förbereds datan analys genom rensning och bearbetning i tabeller. Detta görs med hjälp av verktyg som SQL och Dataform, där projektets struktur följer en etablerad mall från Google Marketing Solutions [3].

Användning av GA4 API och rådataexport

Eftersom GA4:s automatiska export av rådata endast gäller från och med den dag exporten aktiveras, används GA4 Data API för att hämta data för tidigare perioder. API:et returnerar data i aggregerad form och omfattas av begränsningar i antalet tillåtna dimensioner, metriker och rader per förfrågan [4]. Trots dessa begränsningar används API-lösningen som ett komplement för att nå en tillräckligt omfattande historisk täckning.

API-datan är aggregerad och varje rad motsvarar en summerad händelsetyp för en viss kombination av dimensioner. Denna struktur skiljer sig från den mer detaljerade rådataexporten [4]. Det kräver särskilda anpassningar i både modellering och analys. Valet att kombinera dessa två datakällor har gjorts utifrån projektets tekniska ramar och efter dialog med CTRL DIGITAL vars insikter om spårbarhet och koppling till affärsmål har haft stor betydelse. Exempel på den aggregerade API-datan och GA4-datan visualiseras i Figur 2.1 respektive Figur 2.2. Alla potentiellt känsliga fält är anonymiserade eller censurerade i enlighet med gällande sekretessavtal.

Row	date	event_name	campaign_id	campaign_name	source_medium	transaction_id	device	ecommerce_purc...
1	En del av en dags rådata	purchase	(not set)	NDA	eClub / email	NDA	mobile	2
2		purchase	(not set)	(direct)	(direct) / (none)		desktop	1
3		purchase	(not set)	(direct)	(direct) / (none)		desktop	1
4		purchase	(not set)	(direct)	(direct) / (none)		desktop	1
5		purchase	(not set)	(direct)	(direct) / (none)		desktop	1
6		purchase	(not set)	(direct)	(direct) / (none)		desktop	1
7		purchase	(not set)	(direct)	(direct) / (none)		desktop	1
8		purchase	(not set)	(direct)	(direct) / (none)		desktop	1
9		purchase	(not set)	(direct)	(direct) / (none)		desktop	1
10		purchase	(not set)	(direct)	(direct) / (none)		desktop	1
11		purchase	(not set)	(not set)	(not set)		desktop	1
12		purchase	(not set)	(organic)	bing / organic		desktop	1
13		purchase	(not set)	(organic)	google / cpc		desktop	1
14		purchase	(not set)	(organic)	google / cpc		desktop	1
15		purchase	(not set)	(organic)	google / cpc		desktop	1
16		purchase	(not set)	(organic)	google / organic		desktop	1
17		purchase	(not set)	(organic)	google / organic		desktop	1

Figur 2.1: Historisk data med viss censur p.g.a. NDA.

Row	event_date	event_timestamp	event_name	event_params. key	event_params.value. string_value	event_params.val...
1	XXXX-XX-XX	NDA	purchase	page_referrer	NDA	null
				coupon_value	null	0
				affiliation	Online Store	null
				value	null	469
				engaged_session_event	null	1
				page_title	Brothers	null
				batch_ordering_id	null	2
				session_engaged	0	null
				transaction_id	NDA	null
				coupon_amount	null	0
				currency	SEK	null
				engagement_time_msec	null	370
				page_location	NDA	null
				ignore_referrer	true	null
				batch_page_id	null	NDA
				ga_session_number	null	1
				shipping	null	69

Figur 2.2: Exempel på ett GA4-event med tillhörande parametrar (modifierad)

2.2 Dataanalys och etablering av nyckeltalsstruktur (baseline)

När datan har strukturerats utförs en inledande analys för att definiera en grundläggande referensram för projektets nyckeltal. Fokus ligger på att etablera ett nuläge, där exempelvis ROAS analyseras för att förstå nuvarande affärsprestationer. Arbetet inkluderar en kvalitetsgranskning av datan för att identifiera eventuella felkällor eller ofullständiga mätpunkter. I denna fas beslutas vilka KPI:er som ska användas för att styra och utvärdera modellen, i nära dialog med både CTRL DIGITAL och kunden.

2.3 Tillämpning av maskininlärning för kundsegmentering

För att identifiera kundsegment baserat på användarbeteende används KMeans-klustring. Algoritmens funktion och utvärderingsmått som Elbow Method, Silhouette Score, Gap Statistic och Davies-Bouldin Index (DBI) beskrivs i Kapitel 3.

Valet av KMeans grundar sig i dess tillämpning inom modellering baserad på RFM-modellen. RFM-modellen utgörs av aktualitet mätt i tid sedan senaste köp (Recency), köpfrekvens som totalt antal köp (Frequency) och monetärt värde som total köpesumma (Monetary value).

Datan kommer då för varje kund i kundbasen ha ett värde för; hur länge sedan deras senaste köp gjordes, hur många köp de har gjort totalt och hur mycket de har köpt för totalt.

Denna metod är vanligt förekommande inom kundanalys och används för att särskilja kundgrupper utifrån dess beteendemönster och ekonomiska betydelse för verksamheten.Tidigare forskning har visat att en strukturerad segmentering av användare baserat på beteendedata kan förbättra precisionen i marknadsföringsinsatser, öka konverteringsgraden och stärka kundrelationer [5].
Arbetet inleds med att datan förbehandlas genom standardisering, så att variabler med olika skala (t.ex. monetärt värde och frekvens) vägs lika vid klustring. Därefter analyseras datan i Colab med hjälp av Python-biblioteket *scikit-learn*, där KMeans tillämpas i en förberedande analys för att bedöma datakvalitet och identifiera lämpliga värden på antal kluster k. För detta används metoderna Elbow Method, Silhouette Score och Gap Statistic.
När datan och parametrarna har verifierats kommer den slutliga klustringsmodellen att tränas i BigQuery ML. Valet att vänta med detta moment motiveras av den kostnad som BigQuery kräver per körning, vilket ställer krav på effektiv resursanvändning. Det slutliga antalet kluster kommer även att diskuteras med projektägaren CTRL DIGITAL, som även gett vägledning i fråga om affärsmässig tillämpning.

2.4 Standardisering

För att säkerställa att projektets resultat kan återanvändas för framtida kunduppdrag, kommer lösningen att standardiseras och dokumenteras på ett strukturerat sätt. Målet är att den tekniska produkten inte endast ska lösa detta kunduppdrag, utan ska designas på ett modellerbart sätt så att det kan tillämpas i liknande projekt framöver. Dokumentationen kommer även att kompletteras med praktiska guider för projektets olika faser, vilket ska möjliggöra för framtida användare att enkelt förstå, följa och anpassa processen.
I Fas 1 kommer särskild vikt att läggas vid att beskriva hur datakällor identifieras, vilka tabeller som krävs samt hur dessa ska importeras. En detaljerad instruktion kommer att redogöra för varje moment för att säkerställa reproducerbarhet.
Fas 2 kommer använda tydliga och beskrivande namn på tabeller, variabler och parametrar, i syfte att öka transparensen i datamodellen. Transformationerna kommer att kommenteras noggrant och kopplas till respektive analytiskt syfte.
I Fas 3 kommer all kod och metodik att dokumenteras, så att algoritmval, uppdelning av data, utvärderingsmetoder och resultat enkelt ska kunna förstås och återanvändas i framtida kontexter. Genom denna struktur och dokumentationsstrategi kommer ett hållbart, skalbart och återanvändbart system att säkerställas.

Utvärdering

Efter att klustringsmodellen har tränats och tillämpats på den förberedda datan genomförs en utvärdering av resultatets kvalitet. Detta sker dels med hjälp av klustervärderingsmåttet DBI, och dels genom affärsrelaterade indikatorer som genomsnittlig intäkt per kluster och kundengagemang. Målet är att bekräfta att segmenten inte bara är tekniskt väldefinierade utan även affärsstrategiskt relevanta.
Utvärderingen sker med CTRL DIGITAL och innefattar en analys av hur väl det identifierade klustret motsvarar önskade målgruppsprofiler. Om det framkommer att nuvarande modell brister i precision eller affärsnytta föreslås en justering av parametrar eller val av alternativ algoritm. Sådana justeringar kan innefatta ny normalisering, annan viktning av RFM-variabler eller test av annan klustringsmetod.
Slutligen identifieras potentiella förbättringsområden för framtida arbete, inklusive mer individuell analys, funnel-tracking och ytterligare databerikning om fler variabler blir tillgängliga.

2.5 Etiska och hållbarhetsmässiga överväganden

I metoden kommer etiska och hållbarhetsmässiga aspekter att beaktas vid hantering av användardata från GA4, CRM-system och annonseringsplattformar. All persondata kommer att behandlas i enlighet med GDPR [6], med särskild försiktighet vid användning av pseudonyma identifierare och kund-ID:n. Ett sekretessavtal kommer att undertecknas, och inga försök till återidentifiering av individer kommer att göras.

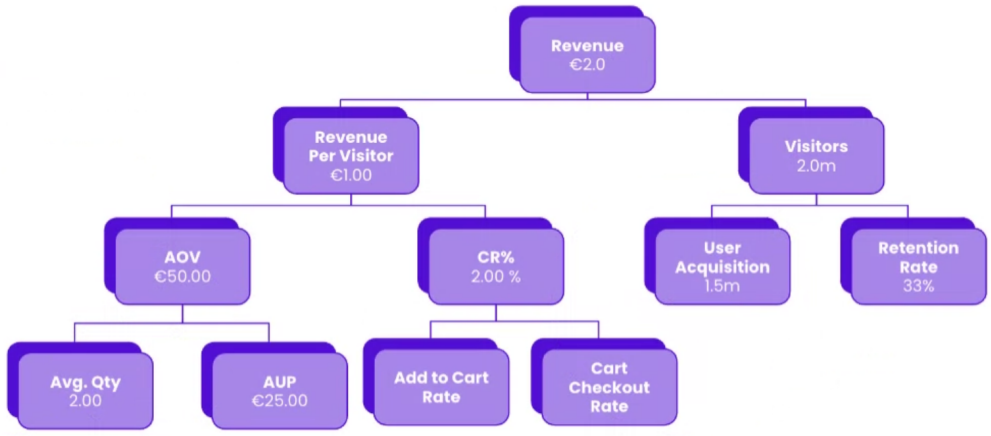
Segmenteringsmodellen (KMeans) kommer att baseras enbart på aggregerade siffror utan möjlighet till identifikation och utvecklas i dialog med handledare och kontaktperson. Hänsyn kommer att tas till potentiella risker för snedfördelning i kundbemötande, och resultatet betraktas som explorativt.
Analysen kommer att genomföras i molnbaserade miljöer med fokus på resurseffektivitet. Datapunkter kommer att avgränsas, tabeller struktureras för återanvändning, och onödig dataprocessering undvikas i syfte att minska klimatpåverkan. Etiska och hållbarhetsmässiga överväganden kommer därmed att integreras i metodval och genomförande.

3
Teknisk Bakgrund

Det genomförda projektet berör tre holistiska områden: digital marknadsföring, moltjänster samt maskininlärningsmetodik. För att ge en helhetsförståelse av projektet introduceras och förklaras relevanta begrepp och termer i detta kapitel.

3.1 Begrepp inom digital marknadsföring

Mätpunkter och Key Performance Indicators (KPI:er) utgör centrala verktyg för att övervaka, styra och optimera marknadsföringsinsatser. De möjliggör en objektivt utvärdering av hur väl olika aktiviteter uppfyller uppsatta mål. Inom ramen för detta projekt ligger fokus på ett antal nyckel-KPI:er kopplade till e-handel och användarbeteende.
För att tydliggöra struktur och samband mellan de olika KPI:erna har ett KPI-träd utformats, vilket illustreras i Figur 3.1.



Figur 3.1: Exempel på KPI-träd

3.1.1 Key Performance Indicators

KPI:er är specifika mätetal som används för att kvantifiera hur väl en verksamhet når sina uppsatta mål och utgör en grund för att strukturera och prioritera relevanta mätpunkter. KPI:er tillåter också intern jämförelse vilket kan synliggöra över- och

underpresterande dimensioner. De KPI:er som analyserats inom ramen för detta projekt är: *Total Revenue*, *Total Sessions*, *Average Order Value*, *Conversion Rate*, *Return On Ad Spend*, *Cost Of Sale* och *Cost Per Click* .

Total Revenue

Total Revenue (sammanlagd intäkt) avser den totala försäljningsintäkten och beräknas som summan av varje enskilt köp, inklusive tillämpade rabatter och avdrag.

Total Sessions

Total Sessions (totala sessioner) påvisar antalet besök.

Average Order Value

AOV (genomsnittligt ordervärde) mäter den genomsnittliga intäkten per genomfört köp. AOV beräknas som total försäljning dividerat med antal ordrar och används för att analysera kunders köpbeteende och bidra till identifiering av högvärdesegment.

CR - Conversion Rate

CR (konverteringsgrad) beskriver andelen besökare som genomför en önskad handling, exempelvis ett köp. CR är en indikator på effektiviteten i kundresan och används för att bedöma hur väl marknadsföringsinsatser leder till faktiska resultat.

Return On Ad Spend

ROAS (annonsavkastning) mäter intäkter genererade per spenderad krona i annonsering. Det är en central KPI inom digital marknadsföring som används för att utvärdera lönsamheten i kampanjer och kan kopplas till kundvärde i efterföljande analys.

Cost Of Sale

CoS (försäljningskostnad) är alla kostnader som direkt krävs för att sälja en produkt eller tjänst, till exempel inköp av varor, tillverkning, eller leverans. Det är alltså inte reklam eller löner till administration, utan bara det som behövs för att själva försäljningen ska kunna ske.

Cost Per Click

CPC (kostnad per klick) är ett mått på hur mycket det kostar varje gång någon klickar på en annons.

3.2 Programspråk

Följande programspråk används i detta projekt.

3.2.1 SQL

SQL är ett språk som används för att hämta, uppdatera och hantera data i databaser. SQL används för att t.ex. hitta specifik data i en tabell, lägga till nya rader eller filtrera fram vissa resultat. [7]

3.2.2 Javascript

JavaScript är ett programmeringsspråk för webben. Det används för att göra hemsidor interaktiva – t.ex. menyer som öppnas, knappar som reagerar eller innehåll som uppdateras utan att sidan laddas om.

3.2.3 Python

Python är ett programmeringsspråk som används till allt från dataanalys och AI till webb och automation. Det är populärt tack vare sin enkel- och smidighet.

3.3 Molntjänster och tekniska verktyg

Följande molntjänster och medföljande tjänster används i detta projekt.

3.3.1 Google Analytics 4

GA4 är Googles senaste plattform för webbanalys, där datainsamlingen är event-baserad snarare än session-baserad. GA4 möjliggör insamling och analys av data från både webbplatser och mobila applikationer, vilket skapar en mer sammanhållen bild av användarbeteende över flera kanaler [4].

GA4 Data API

För att hämta historisk data används GA4 Data API. API:et tillåter åtkomst till rapportdata baserat på angivna dimensioner och metriker. Begränsningar inkluderar maximalt nio dimensioner, tio metriker och kvotregler för antalet rader per förfrågan [4]. Dessa begränsningar har påverkat utformningen av datainsamlingsprocessen i projektet.
Vid användning av GA4 Data API returneras data i aggregerad form, där varje rad summerar förekomsten av en viss händelsetyp (t.ex. *purchase* eller *add_to_cart*) för en specifik kombination av dimensioner, som datum, kampanj, enhet och källa. Det innebär att varje eventtyp utgör sin egen rad, och att vissa fält (exempelvis *totalRevenue*) endast förekommer i samband med specifika events. Denna struktur skiljer sig från rå eventdata och påverkar både tolkning och vidare analys i projektet.

3.3.2 BigQuery

BigQuery är en fullt hanterad, serverlös dataplattform som erbjuder lagring och analys av mycket stora datamängder [8]. Plattformen är en central del av Google

Cloud och tillåter realtidsanalyser genom SQL-liknande syntax, vilket används för att strukturera och bearbeta data i projektet.

BigQuery ML

BigQuery ML utökar BigQuerys kapacitet genom att främja träning och användning av maskininlärningsmodeller direkt via SQL-querys [9]. I projektet används BigQuery MLs implementation av KMeans för att segmentera kunder baserat på historiskt beteende.

3.3.3 Google Cloud Functions

Google Cloud Functions (GCF) är en serverlös lösning från Google som tillåter exekvering av kod som svar på händelser, utan att behöva hantera servrar [10]. I detta projekt används GCF för att automatisera insamlingen av data från GA4 API och exportera denna data till BigQuery, vilket skapade en kontinuerlig och underhållsfri datapipeline.

3.3.4 Dataform

Dataform är ett verktyg för datamodellering och transformation i BigQuery [11]. Det används för att skapa och hantera vyer och modeller av insamlad data, exempelvis genom att standardisera strukturer eller aggregera datapunkter över tid. I detta projekt används Dataform för att organisera och förbereda data för analys i senare faser.
GCP har ett eget verktyg för att hantera, modellera och transformera data i BigQuery på ett organiserat sätt – Dataform.

3.4 Klustringsmetodik

Klustring är en form av icke-superviserad maskininlärning som används för att identifiera mönster och gruppera datapunkter baserat på likheter. Inom ramen för detta projekt används klustringsmetodik för att segmentera kunddata och identifiera underliggande strukturer i användarbeteenden.

3.4.1 Recency, Frequency, Monetary-modellen

Recency, Frequency, Monetary-modellen (RFM-modellen) är en etablerad metod inom marknadsanalys för att segmentera kunder utifrån köpbeteende, och har visat sig särskilt effektiv i kombination med klustringsalgoritmen KMeans [12]. Modellen baseras på tre parametrar:

- **Recency** - aktualitet, det vill säga antal dagar sedan senaste köp
- **Frequency** - köpfrekvens, hur ofta kunden har handlat
- **Monetary** - monetärt värde, alltså den totala intäkten per kund

Genom att kombinera dessa parametrar kan man identifiera kundsegment med olika affärsvärde, såsom lojala stamkunder, nya köpare eller inaktiva användare. RFM-modellen används frekvent som grund för riktade marknadsföringsinsatser och prediktiv analys [12]. I detta projekt har modellen använts för att skapa ett analysunderlag till KMeans-klustringen.

3.4.2 KMeans-algoritmen

KMeans är en populär klustringsalgoritm som syftar till att dela in data i k distinkta grupper baserat på minimering av "inomkluster-varians"(Within-cluster sum of squares, WCSS) [13]. Algoritmen tilldelar varje datapunkt till det närmaste klustercentroidet och uppdaterar sedan centroidernas position iterativt tills konvergens uppnås.

3.4.3 Val av antal kluster

Att identifiera ett lämpligt antal kluster är avgörande för att uppnå segment som är både interna homogena och externa distinkta. För att vägleda detta val tillämpades flera valideringsmetoder. Dessa metoder utvärderar modellens prestanda genom olika kriterier, som klusters kompakthet, separation och stabilitet. Nedan följer en kort beskrivning av de metoder som användes i projektet: Elbow Method, Silhouette Score, Gap Statistic och Davies-Bouldin Index.

Elbow method

Elbow-metoden innebär att WCSS plottas som en funktion av antal kluster. Det optimala k-värdet identifieras vid den punkt där ytterligare ökning av k ger avtagande förbättring i WCSS [14].

Silhouette Score

Silhouette Score mäter hur väl varje datapunkt passar i sitt tilldelade kluster jämfört med andra kluster. Ett högre genomsnittligt värde indikerera tydligare klusterstruktur [14].

Gap Statistic

Gap Statistic jämför den observerade WCSS med WCSS från slumpmässigt genererade dataset. Det k-värde där skillnaden (gapet) är störst anses ge den bästa klustringsstrukturen [14].

Davies-Bouldin Index

DBI är ett mått på klusterkvalitet som utvärderar både klustrens kompakthet och separation. Ett lägre DBI indikerar att klustren är väl separerade och individuellt kompakta. Det optimala antalet kluster motsvarar det värde på k som ger det lägsta DBI-värdet, eftersom ett lägre värde indikerar hög intern kompakthet och stor extern separation mellan klustren [14].

Root Mean Squared Error

Root Mean Squared Error (RMSE) är ett mått för avvikelsen mellan datapunkter och deras tilldelade klustercentroid. Ett lägre RMSE indikerar att datapunkterna ligger närmare sina respektive centroids, vilket tyder på mer homogena kluster. I BigQuery ML används RMSE som ett standardmått för att bedöma klustringsmodellens noggrannhet och stödjer valet av ett optimalt k-värde utifrån intern koherens [14].

4
Genomförande

Det praktiska arbetet i examensarbetet har delats in i fyra faser, där varje fas bygger vidare på resultaten från föregående. Strukturen valdes för att ge ett tydligt flöde mellan datainsamling, modellering och analys, samtidigt som arbetet kunde anpassas efter insikter som uppstod längs vägen. Kapitlet redogör för hur lösningen successivt utvecklades. Uppbyggnad av dataplattform till databerikning och klustringsanalys och vidare till förslag på fortsatt utveckling.

4.1 Fas 1: Uppbyggnad av dataplattform

Den första fasen av examensarbetet fokuserade på att etablera en stabil pipeline för insamling, lagring och transformering av historisk data från GA4 till Google BigQuery. Syftet var att skapa förutsättningar för efterföljande analys och modellering baserat på högkvalitativ och strukturerad data.

4.1.1 Backfill - Datainsamling och lagring

För att automatisera processen utvecklades en lösning i GCP. Funktionen var ansvarig för att hämta daglig data från GA4 API, inom ett angivet tidsintervall (initialt satt 2021-09-01 till 2024-10-21), samt att exportera datan till en fördefinierad tabell i BigQuery. De dimensioner och metriker som användes i rapportförfrågningarna valdes ut i nära samarbete med CTRL DIGITAL. I gemensamma workshops identifierades de datapunkter som ansågs vara mest relevanta för att underlätta analys av ett så brett spektrum av KPI:er som möjligt. Exempelvis inkluderades dimensionerna *device_category*, *campaign_id* och *transaction_id*, samt metriker relaterade till e-handeln och användarbeteende: *sessions*, *ecommerce_purchases* och *total_revenue*. Tabellen som mottog datan i BigQuery skapades manuellt efter initial testning, då möjligheten att automatiskt replikera strukturen från GA4 inte var möjlig.

4.1.2 Tekniska utmaningar

Arbetet i denna fas präglades av flera tekniska begränsningar i GA4 API. Utöver ett maximalt antal rader per förfrågan, fanns även restriktioner gällande antalet tillåtna dimensioner och metriker per rapport, vilket är begränsat till 9 respektive 10. Detta påverkade utformningen av rapportförfrågningarna och krävde noggrant urval för att kunna täcka in projektets KPI-krav. Begränsningarna hade kunnat hanteras mer flexibelt om det varit möjligt att integrera data i BigQuery baserat på

en gemensam unik nyckel, specifikt *transaction_id*. Tyvärr fanns inte denna nyckel tillgänglig i data äldre än 13 april 2023, vilket begränsade möjligheten att koppla samman data över flera tabeller i BigQuery.
För att hantera detta utformades en strategi där data hämtades en dag i taget, med fokus på ett balanserat urval av dimensioner och metriker. Denna strategi gjorde det möjligt till förbättrad felhantering, loggning och successiv validering av datans kvalitet. Begränsningarna hanterades genom kontinuerlig dialog med CTRL DIGITAL, där tekniska kompromisser identifierades och implementerades successivt.

4.1.3 Levererat system och datakvalitet

Fas 1 resulterade i en automatiserad och robust pipeline för datainsamling, som låg till grund för fortsatt analysarbete i projektet. Under denna period importerades data från 2022-04-13 till 2024-10-21, med ett överlapp till 2025-01-01 för valideringsändamål. Den mest centrala tabellen i BigQuery bestod av över en miljon rader och strukturerades platt för att tillåta flexibel hantering i efterföljande faser. Ytterligare tabeller genererades för att stödja analys per dag, vecka, månad och år.
Datans kvalitet utvärderades genom återkommande jämförelser med GA4:s gränssnitt. Testning genomfördes bland annat efter att exporten aktiverats, där innehållet i BigQuery jämfördes mot data i GA4 för att bekräfta överensstämmelse. Eftersom datan var aggregerad kunde fullständig validering inte genomföras, men de övergripande nivåerna och mönstren visade på konsekvens.

Figur 4.1: Schematisk översikt över datapipelinen från GA4 till BigQuery.

Denna pipeline låg därefter till grund för arbetet med analys, segmentering och prediktiv modellering i Fas 2.

4.1.4 CRM- och annonsdata

Projektet inkluderade flera externa datakällor utöver den primära händelsedatan från GA4. Dessa användes för att berika analysen med kundinformation och annonseringskostnader.

CRM

GA4 använder cookies och pseudonyma ID:n för att spåra användarbeteende, vilket innebär att samma kund kan förekomma flera gånger över olika enheter. Eftersom denna spårning inte nödvändigtvis speglar det faktiska antalet kunder, krävdes en mappning mot CRM-systemet. Ett internt dataset innehållande kundernas *customer_id* och deras respektive pseudo-ID:n hämtades därför från Brothers databas, och kopplades därefter till det dataset som användes i analysen.

Google Ads

Det ursprungliga försöket att använda data från GA4 tillsammans med Google Ads visade sig vara otillräckligt, då kampanjinformation inte kunde kopplas korrekt mellan dessa. Orsaken låg i hur GA4 bearbetar och exponerar annonseringsdata, vilket försvårade möjligheten att genomföra en pålitlig sammanfogning baserat på kampanjnamn eller ID.
Som alternativ valdes att hämta Google ads-data separat, frikopplat från GA4 men relaterat till Brothers annonsering. Denna lösning gav tillgång till rådata innehållande bland annat kampanjnamn, kostnader och intäkter. Därefter användes ett Python-skript för att justera intäktsvärden och beräkna nyckeltal, som CoS, med hjälp av jämförelse mot interna referensvärden.

Facebook Ads

Datainsamlingen genomfördes utan större hinder men kampanjnamn saknade enhetlig struktur. För att hantera detta utvecklades en kategoriseringsfunktion baserad på nyckelord i kampanjnamnet, vilket gjorde det möjligt att analysera kampanjtyper över tid.

4.2 Fas 2: Databearbetning och KPI-modellering

Fas 2 utspelade sig i GCP, GA4 och Colab, syftet var att sammanställa datan från Fas 1 och den kontinuerligt uppdaterande datan från GA4:s egna backfill. Datan skulle därefter bearbetas till tabeller för källor (*sources*) och mellanlagring (*staging*) tagen från fas ett, och därefter tabeller för bearbetning (*intermediate*) och slutdata (*output*). I efterhand krävdes ett efterarbete i Colab för att sammanställa och presentera datan från GCP samt inhämtad annonsdata.

4.2.1 Tekniska utmaningar

Eftersom projektet var först i sitt slag inom CTRL DIGITAL krävdes en inledande fas med explorativt arbete och gemensamma workshops för att konkretisera struktur, målbilder och arbetsfördelning. Tidsplanen var till en början öppen, då det saknades tidigare referensramar för hur lång tid respektive moment skulle ta. Lösningen behövde byggas från grunden, vilket bidrog till ett iterativt arbetssätt i Fas 2 - en datatabell kunde exempelvis omdefinieras från slutdata till bearbetningsled i takt med att nya behov och KPI:er formulerades. En särskild utmaning var att annonseringsdata inte kunde integreras i Dataform och därför hanterades separat i Colab.

4.2.2 Dataform skelett

Det första steget i Fas 2 var att skapa ett Dataform-projekt baserat på Google Marketing Solutions [3]. Därefter byggdes projektet upp stegvis: först länkades källor till källtabeller, och därefter skapades mellanlagringstabeller, för att filtrera ut irrelevanta kolumner, urnästa vissa fält samt omdöpa andra parametrar.

4.2.3 KPI sammanställning

Tabellerna togs fram iterativt under projektets gång, där analysbehoven och KPI definitionerna växte fram successivt. Exempelvis behandlades en tidig tabell med kundtyper initialt med slutdata, men efter hand insågs behovet av vidare bearbetning. Den omklassificerades därför till en bearbetningstabell, för att möjliggöra aggregering över tid och vidare segmentering i de slutgiltiga visualiseringarna.

4.2.4 Annonsdata

Då annonsdatan från Google Ads och Facebook Ads inte gick att väva in i Dataform så inhämtades, bearbetades och presenterades den i Colab.

4.2.5 Visualisering

GCP:s begränsade visualiseringsmöjligheter innebar att efterbearbetning behövde utföras i Colab. CSV-filer laddades ned från projektet och bearbetades med hjälp av Python-biblioteket *pandas*, varefter visualisering i form av linje- och stapeldiagram genomfördes med hjälp av *matplotlib*.

4.3 Fas 3: Klustringanalys

Arbetet i Fas 3 inleddes med målet att identifiera lämpligt antal kluster (k-värdet) för att segmentera kunddata med hjälp av KMeans-algoritmen. Tre utvärderingsmetoder användes för att identifiera det mest lämpliga värdet på k, i linje med diskussioner och riktlinjer från CTRL DIGITAL:

- Elbow Method, där det värde på k där minskningen av WCSS avtar anses vara optimalt,
- Silhouette Score, som mäter hur väl varje observation överensstämmer med sitt tilldelade kluster jämfört med andra,
- Gap Statistic, som jämför den observerade klustringen med en slumpmässig referensfördelning.

Under det inledande arbetet identifierades också vissa avvikelser i databehandlingen. Exempelvis upptäcktes att recency initialt bestod av datum i strängformat, vilket gav ett inkorrekt resultat. Detta justerades till ett numeriskt format baserat på antal dagar sedan senaste köp, vilket är mer lämpligt för klustring i RFM format. Ytterligare förbättringar gjordes även i dataförberedelsen, där variablerna recency, frequency och monetary normaliserades för att erhålla en gemensam skala och därmed förbättra klustringens kvalitet.

4.4 Fas 4: Standardisering och avslut

Fas 4 syftade till att utvärdera klustringsmodellen från Fas 3 och att inleda ett arbete med att standardisera processflödet för framtida användning. Enligt den ursprungliga metodplanen skulle modellens prestanda utvärderas mot uppsatta KPI:er och

alternativ omträning ske vid behov. På grund av projektets tidsramar prioriterades dock att färdigställa den primära modellen och presentera resultatet för verksamheten. Fullständig KPI-utvärdering och modellomträning föreslås istället som fortsatt arbete.

Migrering

Under projektets gång eftersträvades en konsekvent struktur i hanteringen av data och analyslogik, särskilt inom ramen för Dataform-projektet. I slutet av arbetet identifierades ett behov av att ytterligare förenkla implementeringen vid framtida kunduppdrag. Tillsammans med handledare på CTRL DIGITAL föreslogs därför att central kod från projektet skulle migreras till företagets interna kodbas, med syftet att skapa en återanvändbar struktur för kommande uppdrag.
Förberedelserna innebar att identifiera vilka delar av projektets logik som var generiska, exempelvis KPI-beräkningar och RFM-segmentering, och vilka som behövde anpassas beroende på kunddata. Diskussionen kring migrering bidrog även till att utvärdera projektets tekniska arkitektur och lyfte fram förbättringsområden vad gäller underhåll och skalbarhet.
Migreringen gav också möjligheter för modifiering av tidigare struktur i Fas 2 och Dataform, eftersom några av tabellerna för bearbetning (*intermediate*) var beroende av andra bearbetningstabeller. Därför implementerades en mellanliggande nivå kallad *refined*-tabeller som gav en mjukare övergång mellan tabeller för mellanlagring (*staging*) och bearbetning.

4.5 Sammanfattning

Genomförandet har resulterat i en komplett teknisk lösning som täcker datainsamling från GA4, databerikning vid externa källor samt analys och segmentering med hjälp av klustring. Arbetet har bedrivits iterativt i fyra faser och har anpassats utifrån de praktiska begränsningar och nya insikter som uppstått under projektets gång. Resultatet är en skalbar plattform för fortsatt analys och beslutstöd, vilket bedrivs vidare i efterföljande kapitel.

5
Resultat

Detta kapitel presenterar de huvudsakliga resultaten från projektets fyra faser, med fokus på hur datainsamling, bearbetning och analys resulterat i en segmenteringslösning baserad på kundbeteende. Arbetet har innefattat både teknisk uppbyggnad av en dataplattform i BigQuery och analys av nyckelparametrar kopplade till kundvärde. Resultaten visar hur data strukturerats, hur relevanta nyckeltal extraherats och hur dessa använts för att identifiera kundsegment med hjälp av KMeans-klustring.

5.1 Resultatsavgränsning

Projektet genomfördes i samarbete med ett externt företag (Brothers), och all data som använts omfattas av ett sekretessavtal (NDA). Detta innebär att ingen faktisk kunddata, enskilda nyckeltal eller detaljerade resultat från specifika kampanjer får redovisas i rapporten. Visualiseringar och analyser har därför anonymiserats eller aggregerats för att bibehålla dataskydd och konfidentialitet. Resultaten som presenteras syftar till att illustrera metodens tillämpning snarare än att avslöja företagets interna information.

5.2 Fas 1: Datainsamling

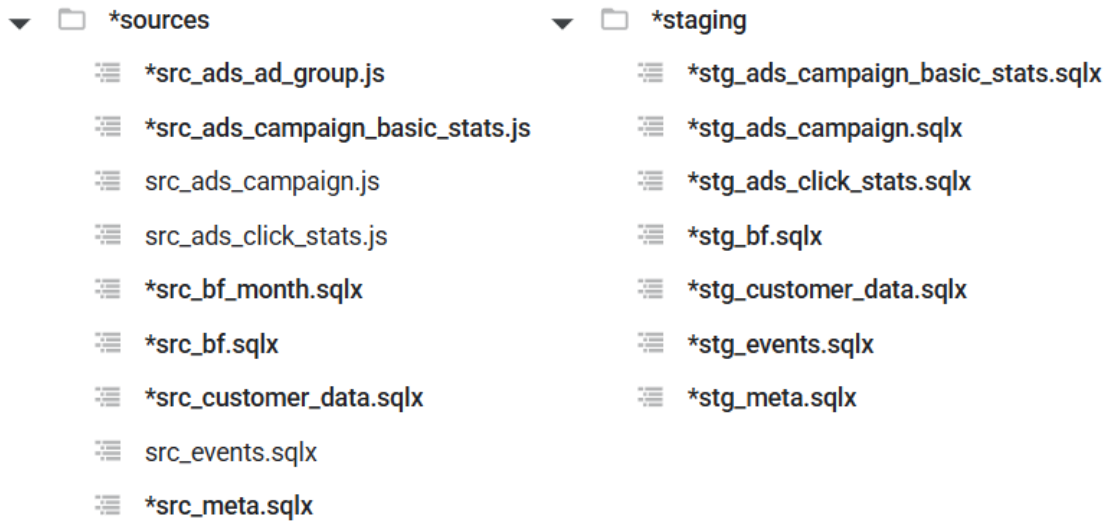
Fas 1 resulterade i en automatiserad lösning för att hämta data från GA4 och lagra den i BigQuery. *1 006 237* rader importerades via GA4 Data API, och omfattade 19 dimensioner och metriker. De mest centrala är *transaction_id*, *total_revenue*, *campaign_name* och *source_medium*.
Datan är strukturerad enligt GA4:s eventbaserade modell, där varje rad representerar en summering av ett specifikt händelseförlopp (t.ex. *purchase* eller *add_to_cart*) för en given kombination av dag, kampanj, kanal och enhet. Som ett resultat varierar fältens innehåll beroende på vilken typ av event som registrerats.
En särskilt viktig observation var att *transaction_id* saknades i data före april 2023. Eftersom denna nyckel används för att koppla händelsedata från GA4 till kunddata via *customer_id*, är den avgörande för att kunna göra analys på individnivå. Utan denna koppling hade det inte varit möjligt att identifiera återkommande kunder eller beräkna kundvärde i senare faser. Av denna anledning fokuserades den fortsatta analysen enbart på data från 2024 och framåt.
Tabell 5.1 visar två typiska exempelrader: en med ett registrerat köp, och en med enbart interaktioner i form av tillägg till varukorg.
Samtliga värden är anonymiserade eller modifierade för att skydda konfidentialitet.

Tabell 5.1: Exempel på dataposter från backfill-processen

date	source_medium	transaction_id	device	sessions	total_revenue	add_to_cart	purchase
2024-03-15	eClub / email	Exempel-ID 1	desktop	1	1 234,0	0,0	1,0
2024-03-15	(direct) / (none)	–	desktop	12	0,0	49,0	0,0

5.3 Fas 2: KPI-modellering

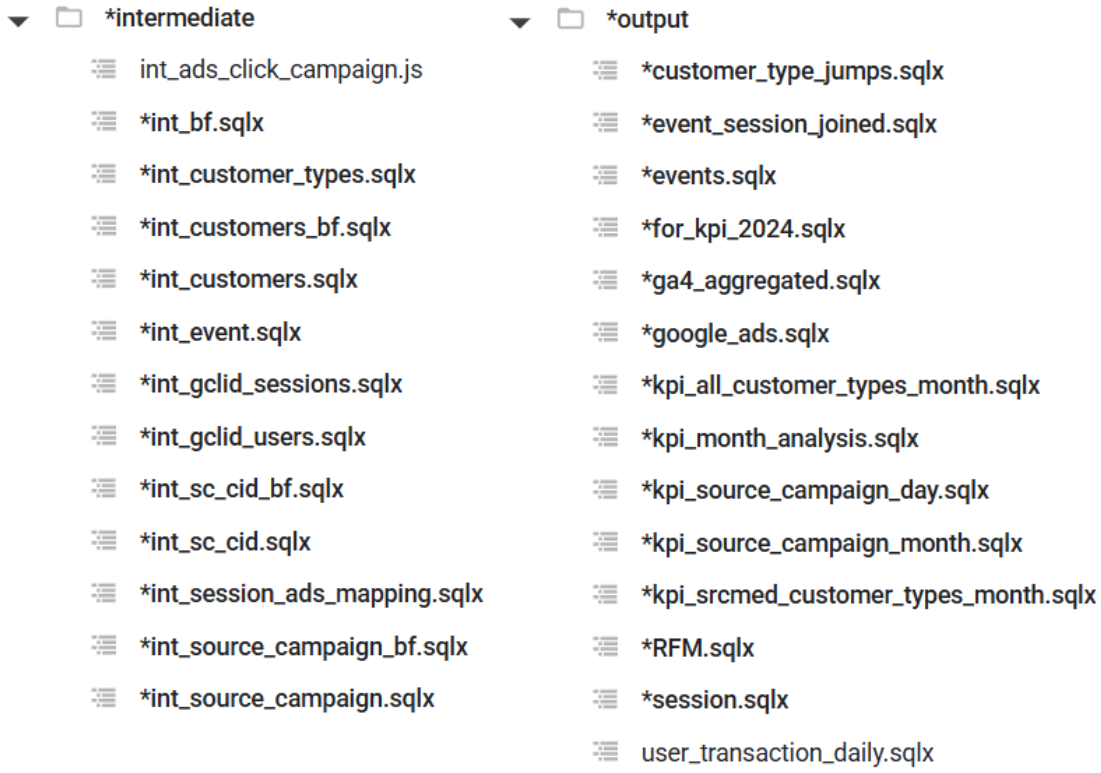
I Fas 2 utvecklades projektets Dataform efter strukturen: källtabeller, mellanlagringstabeller, bearbetningstabeller och slutdatatabeller. Tabeller för de första två stegen innehöll mycket omformatering och transformering för att underlätta sammanvävningen i efterföljande steg. Flera av de tabeller som utvecklades under denna fas kom dock inte att användas i det slutgiltiga flödet. Figur 5.1 visar en översikt över samtliga käll- och mellanlagringstabeller i Dataform samt försök att integrera annonsdata, vilka avbröts på grund av tekniska begränsningar. Tabellerna *src_bf* och *src_bf_month* representerar källtabeller för *backfill*-data och tillsammans med *src_events* (GA4 data) och *src_customer_data* (kund data) var de enda källtabellerna som kom att användas. Dessa transformerades i nästa steg till *stg_bf*, *stg_events* respektive *stg_customer_data*.



Figur 5.1: Käll- och mellanlagringstabeller i Dataform

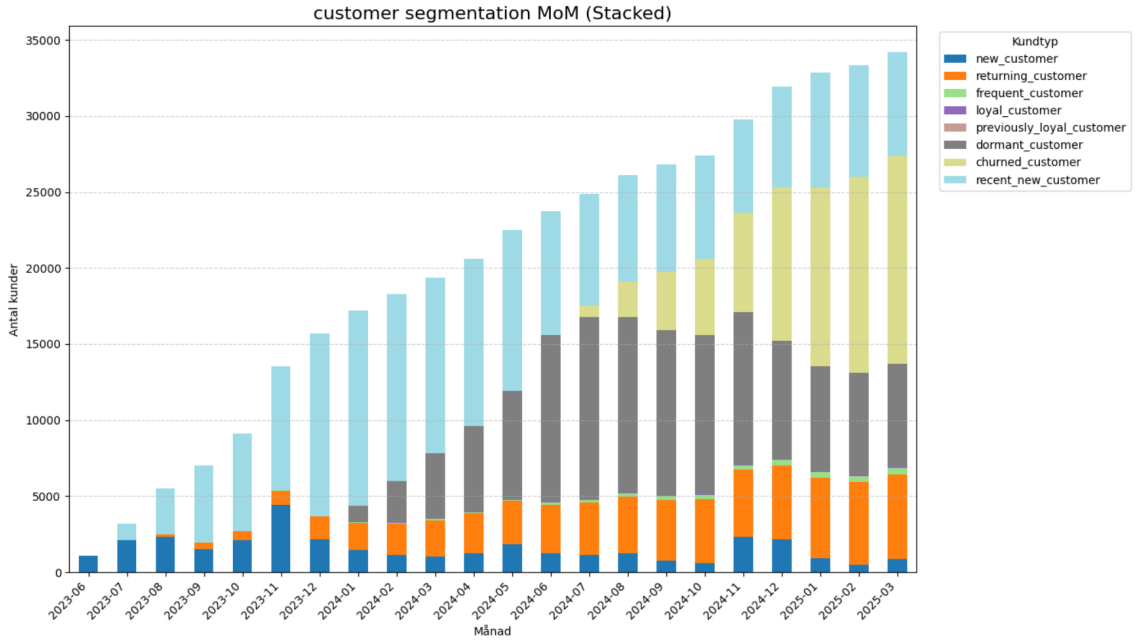
Även tabeller för bearbetning och slutdata präglades av samma utmaningar som i föregående steg, då flera av dem inte kom att användas i det slutgiltiga flödet. Bearbetningstabellerna kan i huvudsak delas in i två kategorier: tabeller för bearbetning av föregående datakällor samt tabeller för extraktion av specifika parametrar med egna definitioner. En översikt över dessa datakällor och deras kopplingar visas i Figur 5.2
Till den första kategorin hör *int_bf*, *int_event*, *int_customers* och *int_customers_bf*. De två sistnämnda sammanfogade kunddata med övriga dataset för att göra den mer lätthanterlig.
Den andra kategorin omfattar tabeller för specifika parametrar, såsom källpresta-

tioner i *int_source_campaign* och *int_source_campaign_bf*, samt aggregering av kunder utifrån trafikkällor i *int_sc_cid* och *int_sc_cid_bf*. Utöver dessa definierades olika kundtyper per individ i tabellen *int_customer_types*.



Figur 5.2: Bearbetning- och slutdatatabeller i Dataform

Bland slutdatatabellerna blir det iterativa arbetetssättet tydligt, då tabeller som *events*, *session* och *event_session_joined* främst bearbetade data och skulle inte betraktas som en slutprodukt. Tabeller som kom att användas i efterföljande analyser var: *for_kpi_2024* som sammanställde KPI:er för 2024, *kpi_source_campaign_day* och *kpi_source_campaign_month*, vilka användes för att analysera källprestationer. Vidare utgjorde *RFM* datasetet för klustringsanalysen med KMeans, medan *customer_type_jumps* och *kpi_srcmed_customer_typers_month* gav djupare analyser över alla kundtyper. Utöver dessa togs *kpi_all_customer_types_month* fram för att följa hur fördelningen av kundtyper har förändrats över tid. Denna tabell visualiserades i Colab, vilket illustreras i Figur 5.3.



Figur 5.3: Kundtyper över tid i Colab

Före visualiseringen analyserades resultatet även med hjälp av GCP:s Explorer-funktion. Ett utdrag av resultatet för *kpi_source_campaign_day* under en specifik dag visas i Figur 5.4.

Row	campaign_name	event_date	purchases	total_revenue	total_sessions	carts	reclassified_source
1	NDA skyddat	Alla kolumner på denna bild har samma datum	25	23297.0	1256	46	(direct) / (none)
2			17	14996.0	983	55	google / cpc
3			15	11709.0	1324	36	google / organic
4			11	9557.0	484	23	eClub / email
5			3	7757.0	120	5	google / cpc
6			1	3068.0	56	1	(direct) / (none)
7			3	1594.0	62	6	google / cpc
8			1	1586.0	48	0	eClub / email
9			1	1325.0	4	0	eClub / email
10			1	1169.0	136	2	google / cpc
11			1	999.0	7	1	eClub / email
12			1	667.0	74	3	google / cpc

Figur 5.4: källprestation

5.4 Fas 3: Klustring

För att identifiera ett lämpligt antal kluster genomfördes ett antal experiment med olika värden på k, från 2 upp till 11. Varje modell utvärderades med hjälp av fyra mått: medelfel (RMSE), Silhouette Score, DBI och Gap Statistic. Analysen gjordes initialt i Colab med hjälp av Python, för att minska behovet av kostsamma körningar i BigQueryML. Därefter testades utvalda konfigurationer i BigQuery för att

generera slutresultatet.
Tabell 5.2 visar tre utvalda modellkonfigurationer för $k = 3$, $k = 4$ och $k = 10$, vilka ansågs särskilt intressanta utifrån en kombination av teknisk kvalitet och affärsrelevans. Trots att modellen med $k = 10$ uppvisade goda tekniska resultat, bestod vissa av klustren av färre än 100 kunder, vilket bedömdes som svåranvänt i en praktisk kontext, eftersom aktiveringskostnaderna för mycket små kundgrupper ofta överstiger den potentiella nyttan
Slutligen valdes modellen med $k = 3$, eftersom den gav en balanserad indelning av låg-, mellan- och högvärdeskunder. Strukturen bedöms vara tillräckligt nyanserad för vidare analys, samtidigt som den är tillräckligt kompakt för att omsättas i operativa beslut. Valet har dessutom stämts av med CTRL DIGITAL, där efterfrågan fanns på just denna typ av överskådlig segmentering.

Tabell 5.2: Sammanställning av tre utvalda resultat för olika k-värden vid klustring

Antal kluster (k)	Davies–Bouldin Index	Mean Squared Distance	Kommentar
3	**0.9916**	**1.2904**	Mest balanserad och affärsrelevant
4	1.028	1.0224	Bra spridning men svag siluett
10	1.0067	0.5435	Tekniskt stark men svårtolkad

Kluster 1

Som illustreras i Figur 5.5 visar Kluster 1 tydligt lägre värden på både frekvens och monetärt värde. Kluster 1 omfattar cirka 16 874 kunder och kännetecknas av en låg köpfrekvens (1,11 köp i snitt), senaste köp genomfört för cirka 516 dagar sedan samt en relativt låg total intäkt (cirka 990 kronor per kund). Kundbeteendet indikerar en låg engagemangsnivå och en hög sannolikhet för kundbortfall. Givet gruppens egenskaper och affärsstrategins fokus på aktiva och högvärdeskunder, prioriteras inte denna grupp i den fortsatta analysen.

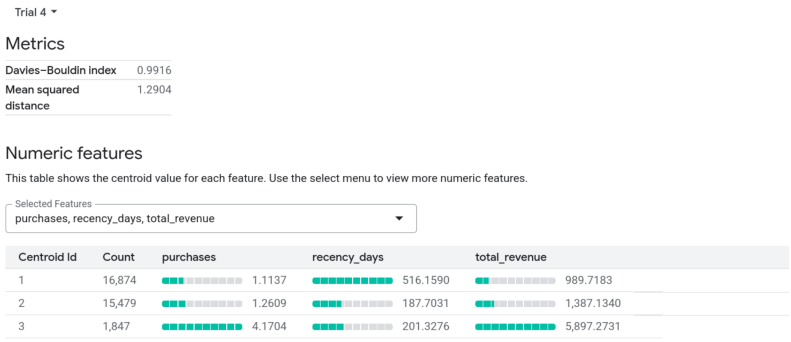
Kluster 2

Kluster 2 omfattar cirka 15 479 kunder och representerar en grupp med relativt hög köpfrekvens, omkring 1,26 köp per kund, samt ett senaste köp för i genomsnitt 188 dagar sedan. Den genomsnittliga totala intäkten per kund uppgår till cirka 1 387 kronor. Kunderna i detta kluster uppvisar ett aktivt köpbeteende och ett medelhögt livstidsvärde. Denna grupp bedöms vara strategiskt viktig för att stärka lojaliteten genom riktade insatser, exempelvis via återkommande kommunikation eller lojalitetsprogram.

Kluster 3

Figur 5.5 bekräftar att Kluster 3 har högst värden på samtliga RFM-parametrar. Kluster 3 utgör den minsta men mest värdefulla gruppen, med cirka 1 847 kunder. Kundbeteendet karaktäriseras av en hög köpfrekvens (i genomsnitt 4,17 köp per kund), ett senaste köp för cirka 201 dagar sedan, samt en betydligt högre total intäkt, omkring 5 897 kronor per kund. Dessa kunder kan betraktas som ett högvärdesegment (VIP-kunder) och utgör en väsentlig del av intäktsbasen. Det är

därför av stor vikt att fokusera på att bibehålla och vidareutveckla denna grupp, exempelvis genom personaliserade erbjudanden eller exklusiva lojalitetsprogram



Figur 5.5: Visualisering av klusterprofiler för den valda modellen med tre kluster ($k = 3$).

5.5 Fas 4: Standardisering och utvärdering

I projektets fjärde fas låg fokus på att utvärdera resultatet från klustringsmodellen och förbereda dokumentation och överföring till verksamheten. Arbetet inleddes med att analysera resultatet från de olika k-värden som testats i Fas 3, där stor vikt lades att bedöma hur väl modellen kunde användas i ett affärssammanhang. Valet av $k = 3$ validerades tillsammans med handledare på CTRL DIGITAL, och sågs som en bra avvägning mellan tolkbarhet, affärsrelevans och teknisk stabilitet.
Parallellt med detta har arbetet med att dokumentera analyslogik, datakällor och tabellflöden påbörjats. Syftet är att underlätta för verksamheten att i framtiden underhålla, vidareutveckla eller integrera modellen i operativa processer. Den slutliga migreringen av modellen till CTRL DIGITAL:s centrala plattform har dock inte färdigställts vid tidpunkten för denna rapport.
Utöver själva migreringen utgör denna fas ett resultat i form av:

- Ett validerat segmenteringsresultat med god teknisk och affärsmässig motivering
- En tydlig struktur för hur modellen kan införas i organisationens befintliga miljö
- En verifierad analys som kan ligga till grund för riktade åtgärder mot olika kundsegment

Denna fas har därmed skapat förutsättningar för att modellen kan tillämpas operativt och integreras i framtida beslutsprocesser.

Migreringen

Migreringen innebar att Dataform-projektet från Fas 2 flyttades över till CTRL Digitals interna Dataform-miljö. Arbetet resulterade i ett mer dynamiskt och konfigurerbart Dataform-projekt, där hårdkodade parametrar ersattes med generiska

lösningar. Detta möjliggör enklare anpassning till olika kunders datakällor samt generering av sluttabeller.
Migreringen ledde också till en mer strukturerad och överskådlig tabellstruktur, vilket underlättar framtida modifieringar och implementationer i liknande projekt.

5.6 Sammanfattning av resultat

Resultaten visar hur projektet i fyra steg byggde upp en fungerande dataplattform och genomförde analys av kunddata med stöd av nyckeltal och klustring. I Fas 1 skapades en automatiserad pipeline för att hämta och strukturera data från GA4 I BigQuery. I Fas 2 bearbetades datan vidare i Dataform där flera KPI:er och kundparametrar togs fram för analys. Fas 3 resulterade I en segmenteringsmodell med tre kluster, baserade på kundernas köpmönster, där särskild fokus lades på att hitta grupper med hög affärspotential. I Fas 4 utvärderades modellen tillsammans med handledare, och förberedelser påbörjades inför en framtida migrering till verksamhetens centrala miljö.
Helheten visar att det är möjligt att skapa en iterativ analyslösning med affärsnära tillämpning, där både datainsamling, analys och visualisering sker i samma tekniska miljö. Resultatet utgör en grund för vidare utveckling av kundstrategier och riktade insatser baserat på datadrivna insikter.
Utöver vad som uppnåtts inom projektets ramar, bör även den framtida tillämpbarheten av resultaten uppmärksammas. På sikt kan denna segmenteringslösning användas som grund för automatiserade kampanjflöden, där kunder i olika kluster tilldelas anpassade erbjudande utifrån sina köpbeteenden. Genom att koppla samman RFM-segmenten med externa system som e-postverktyg, annonseringsplattformar eller lojalitetsprogram kan verksamheten skapa mer träffsäkra och kostnadseffektiva marknadsinsatser. Det gör det möjligt att gå från efterhandsanalys till löpande styrning av kommunikation baserat på realtidsdata. Resultaten från detta projekt kan därmed ses som ett första steg som en mer reaktiv och datadriven marknadsföringsstrategi.

6
Diskussion

Detta kapitel diskuterar arbetets genomförande, valda metoder, tillgången till data och hur dessa faktorer påverkat resultatet. Syftet är att kritiskt reflektera över projektets styrkor, begränsningar och de avvägningar som gjorts under arbetets gång. Diskussionen tar även upp hur resultaten kan tolkas i en affärskontext, samt ger förslag på framtida förbättringar.

6.1 Metodreflektion

Arbetet strukturerades i fyra faser: datainsamling, bearbetning, modellering och utvärdering. Denna struktur utformades i samråd med CTRL DIGITAL för att skapa en tydlig process och följa en logik som både stödjer teknisk utveckling och affärsanalys. Retrospektivt visade sig denna uppdelning fungera väl, särskilt då projektet präglades av osäkerhet i både datakvalitet och målbild. Fasindelningen gav utrymme att agera iterativt, vilket var avgörande när tekniska begränsningar och nya insikter uppstod.
De metodval som gjordes var i flera fall avsiktliga och styrda av projektets syfte att använda verktyg med stark förankring i branschen. CTRL DIGITAL valde att samarbeta med Brothers eftersom de redan använde GA4 som datakälla och BigQuery som analysplattform. För dessa aktörer innebar det inga större förändringar i arbetsflöde, medan projektgruppen samtidigt fick möjlighet att bygga upp kunskap i verktyg som är vanliga inom digital analys. Dessa val bedömdes även som lämpliga för att stödja praktisk tillämpning och jämförbarhet. Andra lösningar togs inte i beaktning, eftersom en avvikande teknisk miljö hade försvårat generalisering till liknande kontexter. Utöver de praktiska fördelarna fanns även ett mer långsiktigt syfte. Genom att arbeta med plattformar som redan används brett inom branschen minskar tröskeln för återanvändning i liknande sammanhang. Kodstrukturen i Dataform och klustringslogiken i BigQuery ML är uppbyggda på ett sätt som gör det enkelt att anpassa lösningen. För att använda modellen på en ny datamängd krävs i regel endast att datan förs in i ett RFM-format och att ett nytt värde på k anges för klustringen. Den övriga analyskedjan kan i stor utsträckning användas direkt. Detta gör att resultatet från projektet kan ligga till grund för fler analyser utan omfattande omarbetning.
Särskilt Fas 2 och 3 bedrevs i nära samverkan med handledare och tekniska kontaktpersoner hos CTRL DIGITAL. Redan från start fanns ett väldefinierat KPI-träd (Se Figur 3.1) som beskrev de nyckeltal som var centrala för verksamheten, exempelvis AOV, ROAS och CR. Däremot saknades en färdig struktur för hur kundsegment

skulle definieras och tolkas baserat på dessa KPI:er. Klassificeringen av kundgrupper och de tröskelvärden som skilde olika segment från varandra behövde därför utformas iterativt under projektets gång. Denna process krävde löpande dialog och justeringar i datamodellen, men bidrog samtidigt till att analysen bättre kunde förankras i affärsverkligheten.
Slutligen bör det nämnas att Fas 4, utvärdering och justering av modeller, inte kunde genomföras fullt ut inom den givna tidsramen. Detta var ett metodologiskt avsteg från den ursprungliga planen, men motiverat utifrån att det primära fokuset låg på att färdigställa en fungerande analyslöning. Istället kom Fas 4 att användas för att sammanställa, dokumentera och migrera kodbasen till CTRL DIGITALs centrala plattform. Detta bedöms ha varit ett motiverat val, eftersom det ökade systemets framtida användbarhet och underhållsbarhet.

6.2 Begränsningar i datakällor

En central utmaning i projektet har varit att arbeta med den strukturerade och aggregerade data som GA4 Data API returnerar. Den eventbaserade modellen i GA4 innebär att varje rad i datan representerar en viss händelsetyp (t.ex. *add_to_cart* eller *purchase*) summerad per kombination av dimensioner som datum, enhet, kampanj och kanal.
Detta skapar en begränsning i analysdjupet, särskilt för events som inte är köp. Eftersom *add_to_cart*, *begin_checkout* och liknande händelser inte kopplas till unika identifierare som *transaction_id*, går det inte att följa en enskild användares väg genom köpresan. Resultatet är att funnel-analyser (dvs. analys av köpresans konverteringar och avhopp) inte kan genomföras med önskad detaljnivå enbart baserat på API-datan.
Endast *purchase*-eventet returneras på transaktionsnivå, där varje rad motsvarar ett faktiskt köp med unikt *transaction_id*. Det tillåter att i efterhand koppla köp till pseudo-ID:n och därigenom segmentera på kundnivå. Däremot saknas denna möjlighet för interaktionsdata, vilket innebär att analys av exempelvis lojalitet eller beteendetrender över tid endast kan ske med stöd av externa CRM-data.
Eftersom stora delar av datan är aggregerad, begränsas den analytiska precisionen och möjligheten att göra individuella tolkningar. Varje datapost kan representera ett genomsnitt eller en summering av många användares beteenden, vilket medför en risk för att betydelsefulla variationer döljs. detta ställer särskilda krav På försiktighet vid tolkning och på metodval som kan hantera denna osäkerhet.

Kombination av rådataexport och API

Under projektets gång användes både GA4:s automatiska rådataexport till BigQuery och GA4 Data API. Rådataexporten aktiverades den 21 oktober 2024, vilket innebar att detaljerad, icke-aggregerad data fanns tillgänglig för perioden därefter. För att analysera historisk data innan detta användes därför GA4 Data API för att komplettera datan bakåt i tiden.
API-lösningen var det enda tillgängliga alternativet för att täcka historiska data-

punkter innan exporten var aktiv. Denna kombination var nödvändig för att skapa ett så heltäckande dataset som möjligt, även om den samtidigt innebar att datans struktur varierade beroende på källa. Det krävde anpassningar i både datahantering och analyslogik.

6.3 Effekter på modellval och segmentering

Begränsningarna i datans struktur påverkade inte bara analysens omfattning utan även valet av maskininlärningsmodeller. Eftersom data saknade sekventiell händelseinformation på individnivå, kunde mer avancerade tekniker som sekvensmodeller eller funnel-baserad segmentering inte tillämpas. KMeans-algoritmen valdes därför för sin förmåga att gruppera kunder utifrån summerade RFM-värden snarare än baserat på tidsberoende mönster. Metoden var väl anpassad till den tillgängliga datanivån, men andra alternativ som exempelvis DBSCAN eller hierarkisk klustring hade kunnat ge mer nyanserade segment givet en högre upplösning i datan.

Möjligheter till prediktiv användning

Trots de begränsningar som fanns i tillgången till detaljerad beteendedata visar tidigare forskning att personligt anpassad kommunikation och segmentering utifrån kundbeteende kan leda till konkreta resultat. Studien av Wu och Lin [5] visar att användning av RFM-segmentering inte bara förbättrar precisionen i marknadsföringsinsatser, utan också ökar kundens upplevda relevans. Detta stärker både lojalitet och köpbenägenhet, vilket gör segmenteringen användbar även som grund för prediktiva modeller.
En fortsatt utveckling av segmenteringen kan ske genom att successivt bredda det variabelunderlag som används i modellen. Genom att inkludera fler beteenderelaterade datapunkter från GA4, exempelvis sessionslängd, trafikkälla eller kampanjinteraktioner, kan segmenten göras mer nyanserade. Ett nästa steg är att fördjupa kopplingen till CRM-data genom att lägga till information om exempelvis köpta produktkategorier eller om kunden handlar som privatperson eller företag. Dessa tillägg skapar en rikare datagrund som på sikt kan användas för att träna prediktiva modeller. Sådana modeller kan bland annat användas för att uppskatta återköpsbenägenhet, identifiera risk för kundavhopp och bedöma hur mottaglig en viss grupp är för olika erbjudanden. Detta kan i sin tur bidra till mer träffsäkra kampanjer och bättre prioritering av resurser i kommunikationen med olika kundgrupper.

6.4 Etiska och hållbarhetsmässiga överväganden

Examensarbetet innefattade hantering av användardata från GA4, kunddata från ett CRM-system samt annonseringsdata från Google Ads och Facebook Ads. Dessa datakällor krävde särskild hänsyn till etiska aspekter i form av personlig integritet, algoritmiskt ansvar och dataskydd, samt hållbarhetsaspekter kopplade till energiåtgång och resursanvändning.

Integritet och dataskydd

Ingen direkt personidentifierbar information användes av projektgruppen. GA4 tillämpar så kallade pseudonyma identifierare, vilket innebär att användare spåras via unika ID:n utan att kopplas till namn eller e-postadresser. Därutöver tillhandahöll Brothers ett internt dataset innehållande kundens *customer_id*, vilket kunde kopplas till *transaction_id* och matchas mot GA4-datasetet. Eftersom *customer_id* i teorin kan länkas till andra personuppgifter inom organisationen, klassas dessa som personuppgifter enligt dataskyddsförordningen GDPR [6]. Projektgruppen undertecknade ett sekretessavtal (NDA) som reglerade hanteringen av dessa uppgifter. Även om det uttryckligen eller underförstått förbjöd försök till identifiering, valde projektgruppen att aldrig försöka koppla *customer_id* till andra attribut, och hade heller inte tillgång till några direkt identifierande uppgifter. All analys genomfördes med fokus på aggregerade mönster, och *customer_id* användes endast för segmenteringen baserat på beteendedata.
Projektet eftersträvade att arbeta i linje med principerna om ändamålsbegränsningen och dataskydd genom design. Även om datamängden var omfattande, inskränktes analysen till de delar som var nödvändiga för att uppfylla projektets syfte. Eftersom *customer_id* i vissa fall kan härledas till en individ inom organisation, omfattas dessa uppgifter även av rätten till radering enligt artikel 17 i GDPR[6], det vill säga rätten att bli glömd.
Vad gäller rättslig grund för behandlingen av uppgifterna förutsattes att data från GA4, Google Ads och CRM samlats in i enlighet med gällande regelverk, exempelvis genom giltigt cookie-samtycke eller medlemsvillkor.

Vidare reflektion kring integritet och framtidssäkring

Även om datan var pseudonymiserad hade ytterligare steg kunnat tas för att minska risken för återidentifiering, exempelvis genom hashning av ID:n vid export. Sådan metod hade gett ett förstärkt skydd samtidigt som analysmöjligheterna bevarats.
För framtida liknande projekt bör en struktur definieras redan vid datainsamlingen början. Det innebär att rådata exporteras i format som ger tillgång till detaljerad analys men samtidigt följer principer för dataminimering och ändamålsbegränsning. Ett tidigt samarbete mellan datateam och juridisk kompetens ger bättre förutsättningar att följa regelverk utan att begränsa den analytiska potentialen.

Algoritmisk segmentering och ansvar

Segmentering genomfördes med hjälp av KMeans utifrån RFM-parametrar. Syftet med analysen var att identifiera kundgrupper som kan ligga till grund för framtida riktade insatser inom kommunikation och marknadsföring. Även om modellen under projektet användes i explorativt syfte, är målet att resultaten på sikt ska tillämpas operativt i verksamheten. Det innebär att den analys som utvecklades i projektet kan komma att påverka hur olika kundtyper behandlas. Denna typ av segmentering kan bidra till förbättrad affärsanalys men kan också innebära risker. Exempelvis kan grupper med lägre köpfrekvens prioriteras ned, vilket kan skapa snedfördelning i kundbemötande och marknadsföring. Inga känsliga variabler ingick i segmente-

ring, och modellen togs fram i dialog med handledare och teknisk kontaktperson. Samtidigt är det viktigt att framtida användning sker med eftertanke, där resultat inte tolkas mekaniskt eller används utan att beakta kontext. Organisationer som använder dessa modeller har ett ansvar att utvärdera konsekvenser över tid och att sätta resultat i relation till både affärsmål och kundupplevelse.

Digital hållbarhet och resursanvändning

Analysen genomfördes i molnbaserade miljöer, huvudsakligen BigQuery. Storskaliga databeräkningar medför miljöpåverkan i form av energiförbrukning. Google har uppgett att all elförbrukning sedan 2017 matchas med förnybar energi och att samtliga datacenter ska vara koldioxidfria dygnet runt senast 2030 [15]. Denna målsättning sänker den klimatmässiga belastningen jämfört med traditionella lösningar men innebär inte att hållbarhetsfrågan är löst.
Projektet tog hänsyn till resursfrågan genom att avgränsa vilka datapunkter som ingick och genom att strukturera tabeller för återanvändning snarare än upprepning. Överdriven datainsamling och ineffektiv analyslogik undveks. Det visar att digital hållbarhet bygger på medvetna val i design och datamodellering redan från början snarare än som ett efterhandsintresse.

Reflektion kring projektets ansvar

Projektet visar att analys av beteendedata kan genomförs med hänsyn till både etiska riktlinjer och hållbarhetsperspektiv. Det förutsätter att varje beslut i datainsamling, lagring och analys vägs mot risker för integritet, ökad komplexitet och miljöpåverkan. Etik och hållbarhet bör ses som integrerade delar av den tekniska utvecklingsprocessen snarare än något som behandlas separat i slutet av ett projekt.

6.5 Förslag på framtida förbättringar

För att utöka analysens detaljnivå och prediktiva förmåga föreslås följande förbättringar:

- **Funnel-analys med rådata:** I takt med att fler veckors rådata exporteras till BigQuery kan sekventiell analys och konverteringsoptimering på sikt implementeras.
- **Utökad feature engineering:** Genom att inkludera flera beteendeparametrar (t.ex. sessionlängd eller trafikkälla) kan segmenteringen göras mer nyanserad, vilket öppnar för användning av mer avancerade modeller.
- **Omträning och validering över tid:** Med tiden, när datamängden växer och affärsmiljön förändras, bör den nuvarande modellen regelbundet utvärderas mot uppdaterade KPI:er. Detta kan inkludera omträning med justerade RFM-parametrar eller övergång till mer adaptiva metoder som hierarkisk klustring eller tidsseriebaserade analyser.

Ett ytterligare område som bör övervägas i framtida arbete är integritetsmedveten datamodellering. För att skapa analyslösningar som både är kraftfulla och ansvarsfulla behöver struktur och export av data planeras redan i ett tidigt skede. Det

innebär att definiera vilka variabler som är nödvändiga, hur dessa kan aggregeras eller pseudonymiseras och hur rådataformatet kan anpassas för att ge detaljerad analys utan att öka återidentifieringsrisken. En sådan strategi stärker både analysens hållbarhet och dess förenlighet med gällande regelverk, vilket skapar bättre förutsättningar för långsiktig användning av modellen.

7
Slutsats

Detta examensarbete har visat hur data från GA4 kan struktureras och analyseras i BigQuery för att skapa affärsrelevanta kundsegment med hjälp av maskininlärning. Arbetet har resulterat i en fungerande pipeline för datainsamling, bearbetning och analys samt en proof-of-concept-modell för kundsegmentering baserad på KMeans och RFM-parametrar.

Kombinationen av GA4:s aggregerade API-data och mer detaljerad rådataexport innebar både möjligheter och begränsningar. För att hantera detta krävdes tekniska kompromisser, särskilt i hanteringen av historisk data. Trots detta kunde segmenteringen genomföras med god precision, och modellen visade tydligt skilda kundgrupper med olika affärsrelevans.

Valet av KMeans var väl anpassat till datans struktur, och resultaten visar att en uppdelning i tre kluster ger en balanserad och tolkbar segmentering. Arbetet har också understrukit vikten av affärsnära tolkning, där tekniska metoder anpassats till verksamhetens behov och förutsättningar.

Sammantaget visar projektet på den praktiska tillämpbarheten av datadriven segmentering i en modern marknadsföringsmiljö. Det framtagna systemet erbjuder en grund för vidare utveckling, där framtida förbättringar kan inkludera sekventiell funnel-analys, mer dynamiska modeller och prediktiva analyser baserade på klustertillhörighet.

Arbetet visar även hur väl valda verktyg och en iterativ metodik kan bidra till att bygga lösningar som är både tekniskt hållbara och affärsmässigt relevanta. Det skapar en grund för vidare tillämpningar, där liknande arbetssätt kan skalas upp och användas i fler kontexter.

Litteraturförteckning

[1] S. K. Noor, M. A. U. Imran, M. B. Aziz, B. Biswas, S. Saha, and R. Hasan, "Using data-driven marketing to improve customer retention for U.S. businesses," in *Proc. 2024 Int. Conf. on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 2024, pp. 338–343, doi: 10.1109/ICICYTA64807.2024.10913232.

[2] CTRL DIGITAL, "About us," CTRL DIGITAL, 2024. [Online]. Available: `https://ctrldigital.com/about`

[3] CTRL DIGITAL, "Data pipeline project," GitHub repository, 2024. [Online]. Available: `https://github.com/google-marketing-solutions/ga4_dataform`

[4] Google, *Google Analytics Data API v1 Documentation*. Hämtad 27 april 2025 från: `https://developers.google.com/analytics/devguides/reporting/data/v1`

[5] Zhi, T., Social Media User Behavior Prediction and Personalized Digital Marketing Model Based on Deep Learning", *Journal of Electrical Engineering and Automation*, vol. 6, nr. 1, s. 10–18, 2024. DOI: `https://doi.org/10.23977/jeea.2024.060102`

[6] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," *Official Journal of the European Union*, L119, pp. 1–88, 2016. [Online]. Available: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`

[7] SQL, *SQL*. Hämtad 14 maj 2025 från: `https://en.wikipedia.org/wiki/SQL`

[8] Google Cloud, *BigQuery Documentation*. Hämtad 27 april 2025 från: `https://cloud.google.com/bigquery/docs`

[9] Google Cloud, *BigQuery ML Documentation*. Hämtad 27 april 2025 från: `https://cloud.google.com/bigquery-ml/docs`

[10] Google Cloud, *Cloud Functions Documentation*. Hämtad 27 april 2025 från: `https://cloud.google.com/functions/docs`

[11] Google Cloud, *Dataform Documentation*. Hämtad 27 april 2025 från: `https://cloud.google.com/dataform/docs`

[12] M. Majilya, G. Maji, P. Ghosh och S. Sen, *Online Retail Customer Segmentation using RFM Quantiles and Clustering Technique*. 2024 4th International Conference on Computer, Communication, Control & Information Technology (C3IT), Hooghly, Indien, 28–29 sep. 2024, s. 1–6. Tillgänglig via: `https://doi.org/10.1109/C3IT60531.2024.10829407`

[13] T. Hastie, R. Tibshirani och J. Friedman, *The Elements of Statistical Learning*. Springer, 2:a upplagan, 2009. Tillgänglig via: `https://web.stanford.edu/~hastie/ElemStatLearn/`

[14] Scikit-learn developers, *Clustering - scikit-learn documentation*. Hämtad 27 april 2025 från: `https://scikit-learn.org/stable/modules/clustering.html`

[15] Google, "Net-zero carbon" 2025. [Online]. Available: `https://sustainability.google/operating-sustainably/net-zero-carbon/`

[16] Gustaver, M. (2020) A Chalmers University of Technology Master's thesis template for LaTeX. Unpublished.

INSTITUTIONEN FÖR DATA- OCH INFORMATIONSTEKNIK
CHALMERS TEKNISKA HÖGSKOLA
Göteborg, Sverige
www.chalmers.se

CHALMERS