university of groningen

Data Engineering 2024

Wehkamp Case
Final Report

Group 6

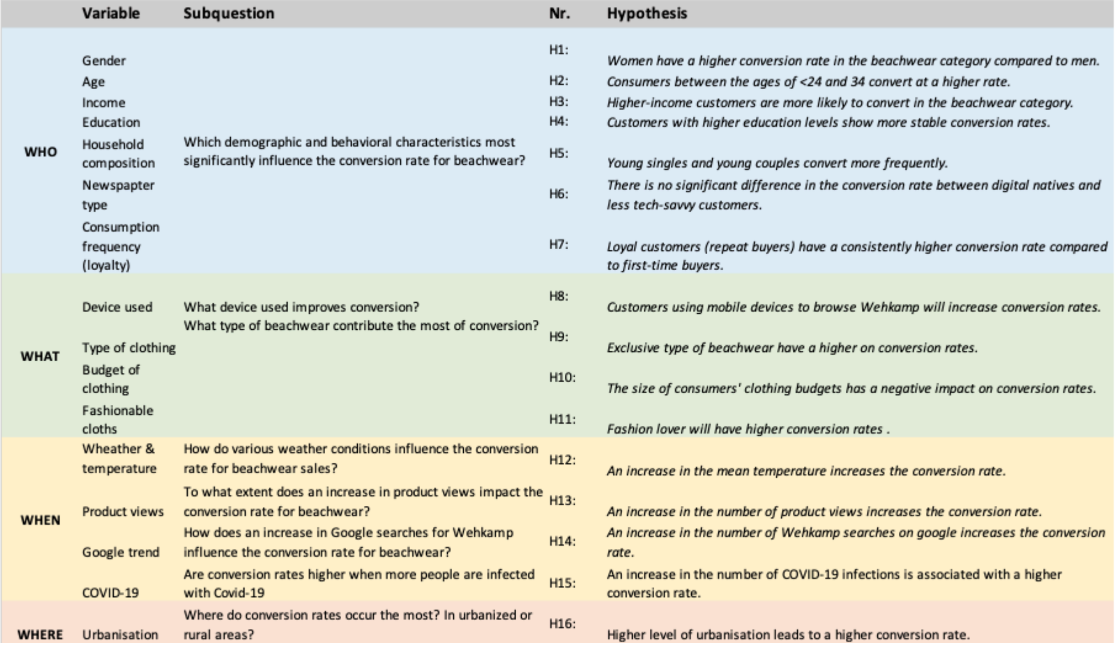
Sarah Isabella Fokken S5395453
Arjan Homsma S4538897
Jakob Schmid S6017967
Yichong Tao S5065623

1. Introduction

Wehkamp, a leading online retailer in the Netherlands, serves over 2.25 million active customers with a diverse selection of 300,000 products. Despite strong customer engagement (486,000 daily visits), recent analysis has revealed unexpected fluctuations in beachwear sales. Temperature, a previously unexamined factor, may be influencing these shifts, alongside other factors like demographics and search behavior.

The **management** is faced with a **dilemma**: How to mitigate these fluctuations, including the impact of year-on-year growth, and stabilize sales in the beachwear category. The current conversion rate stands at 9.72%, and the goal is to identify actionable strategies to increase this rate by 1%. Therefore, the core **management question** guiding this report is: *How can Wehkamp optimize the customer journey in the beachwear category to achieve this improvement?*

To answer the **management question**, a research question was developed to understand the underlying influences on conversion rates: *What independent variables affect the conversion rate, particularly in the beachwear category?* By delving into these factors, the report aims to uncover the impact of external elements like temperature, as well as internal aspects such as customer demographics and behavior patterns. To achieve this, several sub-questions as well as hypotheses were developed (see Table 1). In appendix, a more detailed version is available.



	Variable	Subquestion	Nr.	Hypothesis
WHO	Gender	Which demographic and behavioral characteristics most significantly influence the conversion rate for beachwear?	H1:	*Women have a higher conversion rate in the beachwear category compared to men.*
	Age		H2:	*Consumers between the ages of <24 and 34 convert at a higher rate.*
	Income		H3:	*Higher-income customers are more likely to convert in the beachwear category.*
	Education		H4:	*Customers with higher education levels show more stable conversion rates.*
	Household composition		H5:	*Young singles and young couples convert more frequently.*
	Newspapter type		H6:	*There is no significant difference in the conversion rate between digital natives and less tech-savvy customers.*
	Consumption frequency (loyalty)		H7:	*Loyal customers (repeat buyers) have a consistently higher conversion rate compared to first-time buyers.*
WHAT	Device used	What device used improves conversion?	H8:	*Customers using mobile devices to browse Wehkamp will increase conversion rates.*
	Type of clothing	What type of beachwear contribute the most of conversion?	H9:	*Exclusive type of beachwear have a higher on conversion rates.*
	Budget of clothing		H10:	*The size of consumers' clothing budgets has a negative impact on conversion rates.*
	Fashionable cloths		H11:	*Fashion lover will have higher conversion rates .*
WHEN	Wheather & temperature	How do various weather conditions influence the conversion rate for beachwear sales?	H12:	*An increase in the mean temperature increases the conversion rate.*
	Product views	To what extent does an increase in product views impact the conversion rate for beachwear?	H13:	*An increase in the number of product views increases the conversion rate.*
	Google trend	How does an increase in Google searches for Wehkamp influence the conversion rate for beachwear?	H14:	*An increase in the number of Wehkamp searches on google increases the conversion rate.*
	COVID-19	Are conversion rates higher when more people are infected with Covid-19	H15:	An increase in the number of COVID-19 infections is associated with a higher conversion rate.
WHERE	Urbanisation	Where do conversion rates occur the most? In urbanized or rural areas?	H16:	Higher level of urbanisation leads to a higher conversion rate.

Table 1. Sub-questions & hypothesis

This report begins with a management summary, followed by an overview of the relevant data and the methodology used for data extraction and analysis. It then presents the results of the analysis and concludes with concrete action recommendations.

2. Managerial Insights

The results of this research have several implications for managers who are interested in increasing Wehkamp's beachwear sales. To understand how to boost sales, various factors influencing the conversion rate were identified.

From the research, it was found that a number of factors significantly influence the sales of Wehkamp's beachwear:

- Demographic factors such as age, income, and education level have a substantial impact on conversions. Middle-aged customers with higher incomes and higher education levels demonstrated the highest conversion rates.
- Gender emerged as a significant factor, with women displaying higher conversion potential compared to men, suggesting that gender-specific marketing strategies may be beneficial.
- Conversion was found to be influenced by environmental factors like temperature and sunshine, highlighting the potential of weather-responsive marketing strategies.
- Sales forecasts showed an inverse relationship with COVID-19 case rates, suggesting reduced conversion during periods of higher infection rates.
- Customers who prefer "fashionable" clothing segments, have higher income levels, and live in urban areas were more likely to convert.
- Interestingly, an inverse relationship was found between conversion rates and the number of product views, implying that customers who eventually convert tend to make quick purchasing decisions rather than extensively browsing multiple products.

Wehkamp could benefit from simplifying the customer journey by minimizing the number of product views required before purchase, perhaps by providing tailored product recommendations or emphasizing popular items. By focusing on these areas, Wehkamp can likely increase conversion rates and reduce volatility in beachwear sales.

Addressing this management dilemma is crucial to ensure that Wehkamp remains competitive. By understanding these variables, Wehkamp can optimize the customer journey, reduce sales volatility, and enhance customer loyalty. The insights from this research will help management make data-driven decisions to improve conversion rates.

3. Methodology

Data Sources
For the analysis, four different datasets were used, all of which are presented in this chapter. Additionally, the procedures for preparing, cleaning, and merging the data are explained in detail.

Wehkamp data (internal)
The Wehkamp dataset includes clickstream data from Wehkamp.nl, with a sample size limited to a small subset of visitors. The dataset covers activity in the beachwear category between January 1 and July 31, 2022. In total, six data tables encompassed the Wehkamp dataset: Article, Customer, Session, Article events, Order. All the data is anonymized and denormalized.

Weather data (external)
As an external data source, we used the KNMI dateset, which contains weather information for the Netherlands. The extracted data covers the period from January 1 to December 31, 2022. For our analysis, we agreed to use the following variables: daily mean temperature (°C), daily sun hours, amount and duration of rainfall, and cloud cover.

Covid-19 data (external)
Given the significant impact of the global COVID-19 pandemic during the analysis period, we incorporated the global daily count of new COVID-19 cases into our research, utilizing data from the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The analysis focuses on data specific to mainland Netherlands, covering the period from January 1, 2021, to December 31, 2022.

Google Trends data (external)
To account for times Wehkamp is trending on google search, the weekly google trends for the search term Wehkamp was included. Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

Method of Combining Data Sources

We have done an internal data integration using SQL, mainly with the customer and event tables. We found an unusual situation where two customer IDs appeared under the same session ID with equal session durations. We attributed all operations to a single customer and aggregated them to the same session ID. We use the customer ID as the key for the merged table and the session ID, and create a new dummy variable for customer actions in order to include time and all action types in a single observation. This will help us to study the relationship between customer interactions and time-dependent variables.

We used three external datasets: weather data from KNMI, trend data from Wehkamp, and COVID data from GitHub. We obtained these CSV files from the Internet and imported them into RStudio. For the weather data, we processed the date variables using as.date(). Since COVID and trend data are not continuous, we created weekly averages as new variables. Finally, we processed the date variable using as.date().

We merged all the data using left joins. First, we extracted the internal data from SQL using the SQL Toolkit in R. Then, we used the internal data as the left table for the left join, using date as the key to join the weather, COVID, and trend data. The final dataset consisted of 536,765 records.

Method of Data Cleaning

After merging the data, we began cleaning. Our dataset included multiple types, such as integer, numeric, character, and integer64. To streamline processing, we converted character and integer64 data to integers. For non-numeric variables like device and gender, we used mutate()

and case_when() from dplyr to assign values. For ordered numeric variables like age, we used gsub() with these functions to remove non-numeric characters and convert them to integers. For integer64 variables, we called the bit64 toolkit for type conversion. To handle "other" values as missing data, we converted "other" to 0, then used mutate() and ifelse() to convert 0 to NA.

Method of treating missing Values

After converting data types, we checked for missing values. Most missing data were in the internal data, especially variables from the customer table, with two main types missing (308,175 and 308,976 values) and fewer missing values for gender (289,552). Missing device type data from the event table totaled 111,050. We believe these missing values are due to users opting out of cookies for privacy, classifying this data as Missing at Random (MAR). With about 60% data loss using listwise deletion, we applied Multivariate Imputation by Chained Equations (MICE) to fill in missing values based on variable relationships. This method generates multiple complete datasets, analyzes each, and combines results for reliable estimation. We then extracted the first complete dataset and continued with further data processing and analysis.

Method of removing outliers

To identify outliers, we considered the deviation of all observed variables simultaneously using Mahalanobis distance as our criterion. The Mahalanobis distance density distribution and QQ plot confirmed outliers in the data. We then created an ordered scatter plot and marked different standard deviation heights. Based on this, we set three standard deviations as our outlier threshold, identifying 17,883 outliers (3.33% of the data), which we then removed. This left us with a total of 518,882 observations.

Method of Analysis

The analysis began with defining and structuring the business challenge, followed by data collection and processing, as previously described. Next, we conducted the analysis using various statistical methods, including chi-square tests, logistic regression, and multiple regression. We then presented our findings and recommendations in this report. The final step, implementation of the proposed solutions, falls outside our scope of responsibility.

Description of Variables (General and Statistical)

Variable	**Description**	**Data Type**	**Value description**	**Descriptive Statistics**
internet_session_id	Session ID	nominal	Random individual session number	Length:518882
most_customer_id	Customer ID	nominal	Random individual customer number	Length:518882
last_session_dtime	Date of the session	date	Date of the session	Length:518882
n_product_view	Number of product views per session	discrete	In number of views	Min.: 0 Max.: 9 Median: 1 Mean: 1.395
WeekAverage	Daily new covid	discrete	In number of cases per day	Min.: 1567

	cases			Max.: 30266 Median: 3759 Mean: 5696
mean_temp	Daily mean temperature	continuous	in 0.1 degrees Celsius	Min.: -58.0 Max.: 269.0 Median: 165.0 Mean: 150.1
session_conversion	Conversion or no conversion of the session	binary	0 = no sale 1 = sale	Min.: 0 Max.: 1 Median: 0 Mean: 0.09257
clothing_budget	Budget allocated to clothing:	continuous	1. Little 2. Below average 3. Average 4. Above average 5. Much	Min.: 1 Max.: 5 Median: 3 Mean: 2.714
clothing_fashionable	How much clothes are bought in the 'fashionable' segment	continuous	1. Little 2. Below average 3. Average 4. Above average 5. Much	Min.: 1 Max.: 5 Median: 4 Mean: 2.848
consumption_frequency	Consumption frequency of the household	continuous	1. Little 2. Average 3. Much	Min.: 1 Max.: 3 Median: 3 Mean: 2.314
education	Education level	nominal	1. Primary education 2. LBO/VMBO-K/VMBO-B/MBO 1 3. MAVO/MULO/VMBO-T/VMBO-G 4. MBO (2, 3 or 4) 5. HAVO/VWO/HBS 6. HBO or WO bachelor 7. HBO or WO master/MBA/postdoc	Min.: 1 Max.: 7 Median: 6 Mean: 5.102
age	Age category	continuous	1. < 25 years 2. 25 - 29 years 3. 30 - 34 years 4. 35 - 39 years 5. 40 - 44 years 6. 45 - 49 years 7. 50 - 54 years 8. 55 - 59 years 9. 60 - 64 years 10. 65 - 69 years 11. 70 - 74 years 12. 75 - 79 years 13. >= 80 years	Min.: 1 Max.: 13 Median: 6 Mean: 6.283
household_composition	Composition of the household	nominal	1. Young singles 2. Middle-aged singles 3. Older singles 4. Families with only young children 5. Families with older children 6. Young couples without children 7. Middle-aged couples without children	Min.: 1 Max.: 8 Median: 5 Mean: 4.889

			8. Older couples without children	
Income	Income category of the household	nominal	1. < 18,000 2. 18,000 - 26,000 3. 26,000 - 35,000 4. 35,000 - 50,000 5. 50,000 - 75,000 6. 75,000 - 100,000 7. 100,000 - 200,000 8. >= 200,000	Min.: 1 Max.: 8 Median: 7 Mean: 6.039
urbanization	Degree of urbanisation	nominal	1.00 – High urbanisation 2.00 3.00 4.00 5.00 – Low urbanisation	Min.: 1 Max.: 5 Median: 3 Mean: 3.108
gender	Gender of the customer	nominal	1. Male 2. Female 3. Other	Min.: 1 Max.: 2 Median: 2
device	Type of device used.	nominal	1. mobile 2. tablet 3. desktop	Min.: 1 Max.: 3 Median: 1

Table 2. Sub-questions & hypothesis

4. Results of the Research Question

Analysis of WHO variables

H1: Gender and session conversion

The analysis reveals a statistically significant association between gender and session conversion ($p=0.0004$), indicating that conversions differ between gender groups. Specifically, women have notably higher conversions compared to men. The observed frequencies deviated from the expected values, further supporting this finding. This suggests that gender plays a significant role in influencing conversion behavior.

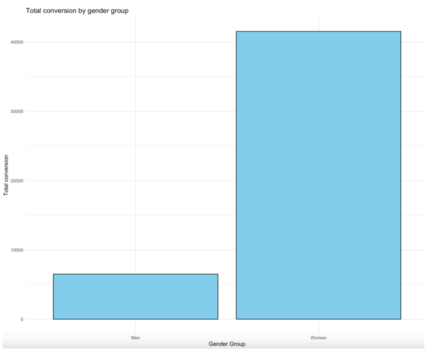


Figure 1. Gender and session conversion

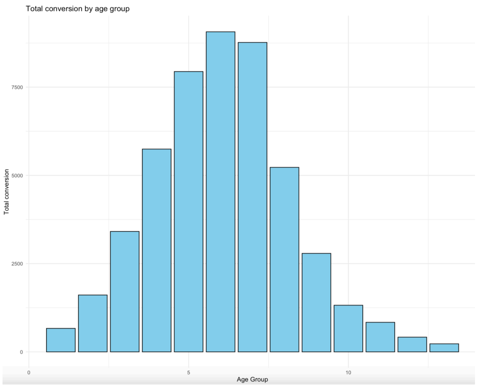


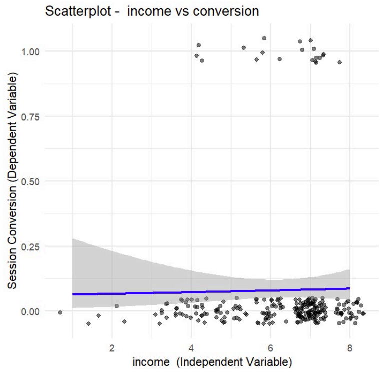
Figure 2. Age and session conversion

H2: Age and session conversion

The Chi-Square test shows a highly statistically significant association between age and session conversion (p<2.2e−16), with the age groups 35-39 years (group 4), 40-44 years (group 5), and 45-49 years (group 6) demonstrating the highest conversions. This indicates that middle-aged consumers, specifically those between 35 and 49 years, are more likely to convert compared to other age groups. Thus, these findings do not fully support the original hypothesis that "Consumers between the ages of 25 and 34 convert at a higher rate." Instead, conversion peak for consumers aged 35 to 49.

H3: Income and session conversion

The analysis reveals a strong significant association between income levels and session conversion (p<2.2e−16), indicating that conversion rates vary across different income categories. Specifically, higher-income groups, such as 100,000 to 200,000 (Group 7), showed the highest conversion rates, followed by groups 75,000 to 100,000 (Group 6)and >= 200,000 (Group 8) (see Figure 3). This supports the hypothesis that higher-income customers are more likely to convert in the beachwear category.



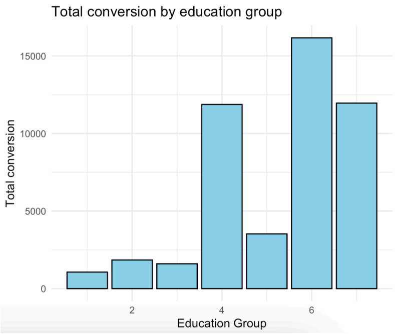
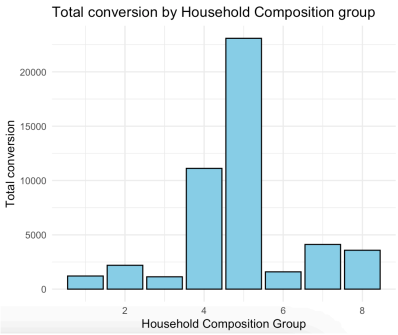


Figure 3. Income and session conversion

Figure 4: Education and session conversion

H4: Education level and session conversion

With a p-value of <2.2e−16, the analysis reveals a strong relationship between education level and session conversion. Customers with higher education levels, specifically those with bachelor's degrees (Group 6) and master's/postdoc degrees (Group 7), show both higher and more stable conversion rates compared to those with lower education levels. This supports the hypothesis that customers with higher education levels show more stable conversion rates.



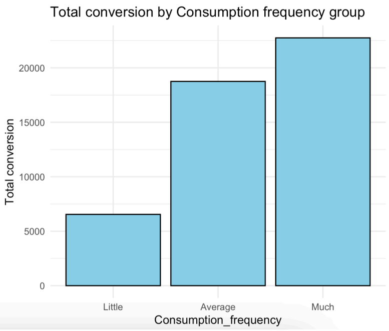


Figure 5. Household composition and session conversion

Figure 6. Household composition and session conversion

H5: Household composition and session conversion

The analysis shows that household composition significantly affects conversion rates (p-value < 2.2e-16). Families with older children (Group 5) have the highest conversions, indicating strong engagement. Families with young children (Group 4) also convert substantially, showing active engagement. In contrast, younger households, such as young singles (Group 1) and young couples without children (Group 6), have lower conversion rates than initially hypothesized, suggesting that these groups are less likely to convert.

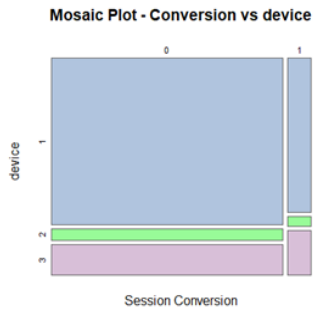
H7: Consumption frequency (Loyalty) and session conversion

The Chi-squared test indicates a strong association, with households in the "Much" consumption group showing the highest conversion rates compared to the "Average" and "Little" groups. Specifically, households in the "Much" group demonstrate a noticeably higher conversion rate, followed by those in the "Average" group, while those in the "Little" group have the lowest rate.

Analysis of WHAT variables

H8: Customers using mobile devices have a positive influence on conversion rate.

In order to investigate the relationship between device use (categorical variable) and conversion (categorical variable), a chi-square test was applied. From the test we can judge that there is a significant difference in the conversion using different devices (P<0.01). From the fig.8, we can see that the blue color has the largest surface, followed by the purple, which means that the number of people using mobile phones is the largest, followed by those using computers. Mobile phones bring the most conversions. But comparing the 0/1 blocks of the same color, many of the mobile users did not purchase, and the conversion rate is lower than computer users. Therefore, the hypothesis is rejected, mobile use doesn't positively affect the conversion rate.



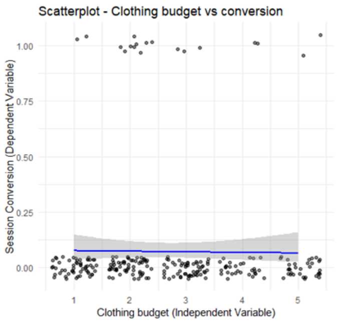


Figure 7. device and session conversion **Figure 8.** Clothing budget and session conversion

H10: Higher shopping budget has a positive influence on conversion.

In order to investigate the relationship between clothing budget (interval variable) and conversion (categorical variable), we used a logistic regression test. From the results of the test, we can judge that the effect of clothing budget on conversion is significantly positive, but this effect is extremely small (p<0.01, B=0.037).

H11: Fashion level has a positive influence on conversion.

To examine the relationship between clothing fashionable (interval variable) and conversion rate (categorical variable), we used a logistic regression test. From the results of the test, we can determine that the effect of clothing fashionable on conversion rate is significantly positive (p<0.01, B=0.24). We visualized the relationship using a scatter plot + fitted curve, which shows that the dots become denser as the fashion level increases, and the higher the fashion level the higher the conversion. The fitted curve shows an upward trend, which means that the conversion rate also increases with fashion level. The original hypothesis is true, fashion level has a positive effect on conversion rate.

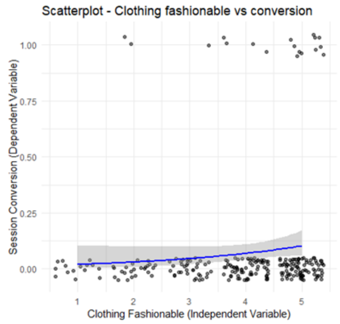


Figure 9. Clothing fashionable and session conversion

Analysis of WHEN variables

H12: Increase in the number of product views increases the conversion rate

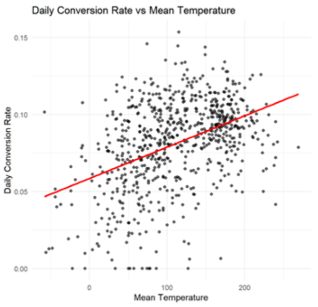
The hypothesis, that a increase in the number of product views increases the conversion rate was tested with a logistic regression analysis. The results show a negative effect of -0.055 on the log-odds of the conversion. The p-value is highly significant at <2e-16. The null hypothesis (no effect) is rejected, and we conclude a negative effect of the number of product views on conversion rate. Fewer product views correlate with a higher conversion likeliness.

H13: Increase of google searches of Wehkamp leads to higher conversion rate

The hypothesis, that a increase in the number of Wehkamp searches on google increases the conversion rate was tested with a logistic regression analysis. The results show a slight positive effect of 0008392 on the log-odds of the conversion. The p-value is with 0. 0916 significant only at a 10 % significance level. The null hypothesis (no effect) is accepted, and we conclude no significant effect of Wehkamp in google trends on conversion rate at a 0.05 threshold.

H14: Increase in mean temperature leads to higher conversion rate

The hypothesis that an increase in the mean temperature increases the conversion rate was tested with a logistic regression analysis. The results show a positive effect of 1.952e-04 on the log-odds of the conversion. The p-value is highly significant at <2e-16. The null hypothesis (no effect) is rejected, and we conclude a positive effect of mean temperature on conversion rate. Higher temperatures correlate with a higher conversion likeliness.



H19: More Covid-19 infections lead to higher Conversion rates

An ordinal logistic regression was carried out in order to analyze the relationship between conversion rates and weekly averages of Covid-19 infections in the Netherlands. The estimate shows a converse significant effect (<2e-16) on the log-odds of the conversion rates. The effect of -1.728e-05, however, is very small and therefore not considered as an important factor. We conclude that conversations are lower when more people are infected with Covid-19. However, the effect is very small and negligible.

Analysis of WHERE variables

H A higher level of urbanization leads to higher Conversion rates

An ordinal logistic regression was carried out to analyze the relationship between conversion rates and the level of urbanization. The estimate shows that there is a positive effect of 0.07 on the log-odds of the conversion rate, our P-value shows that this effect is significant (<2e-16). We reject the null hypothesis and thereby conclude that with one unit increase, the log-odds of conversion rate increase with 0.07. With 5 being the lowest level of urbanization: conversion rates occur the most in rural areas.

Final Regression Model

As a last step we build a regression model with all of the scaled variables (n_product_view, WeekAverage, mean_temp, weekly_trend, urbanisation, clothing_budget, clothing_fashionable, consumption_frequency, age, income). All the variables were standardized before the analysis in order to make the estimates comparable. The results suggest significant effects from the number of product views, new covid cases, urbanization level of the customer, their clothing budget, their fashionability, consumption frequency, age and income. All of them provide significant influence on the conversion likeliness at a significance level of 0.001 as can be seen in the table below. The estimate explains the effect strength. The more the value is distant from zero, the larger the effect. Variables with a negative estimate have a negative effect on the conversion likeliness. Therefore the more clothes from the “fashionable”-segment are bought, the more likely a conversion is (estimate of 0.246834). In close second place is income. After that the budget for clothing and consumption frequency. Number of new covid cases has a relevant negative effect on the conversion likelihood as well. The other data can be drawn from the table below.

Term	Estimate	Pr(>|z|)	Term	Estimate	Pr(>|z|)
clothing_fashionable	0.246834	< 2e-16 ***	n_product_view	-0.061617	< 2e-16 ***
income	0.228736	< 2e-16 ***	urbanisation	0.058694	< 2e-16 ***
clothing_budget	0.116479	< 2e-16 ***	age	-0.043803	1.38e-15 ***
consumption_frequency	-0.113543	< 2e-16 ***	weekly_trend	0.008032	0.118
WeekAverage	-0.105758	< 2e-16 ***	mean_temp	0.002836	0.649

5. Managerial Advice & Conclusion

Managerial Advice

So how can Wehkamp optimize the customer journey in the beachwear category to achieve the goal of increasing the conversion rate by 1 %?

Almost all of the one to one analysis of the tested sub-questions showed a significant effect. First and foremost it could be shown that the sales of swimwear are affected by weather conditions. That however might also be due to the correlation between weather and season and summer is just the more prevalent season to go swimming. Another significant external factor was the covid pandemic, thats ending might contribute positively to conversions.

Wehkamp should target those groups that have a higher conversion likeliness per shopping session. Therefore women, aged 35 to 50 with higher and high incomes (75,000-200,000 € per year), higher education (bachelor or even master degree), who already have children and live in more rural areas. They are however digital natives as can be seen by the newspaper type and consume frequently and regularly from their phone. Buying is however usually done on the computer as it might be seen as more secure. They have a high sense for fashion and style and a high budget for it. Additionally the analysis showed that the number of product views is strongly negatively contributing to the conversion likeliness. The larger final regression model showed the relevance of the single variables on conversion. Therefore Wehkamp should focus especially on fashionable, high income individuals.
To reach this target group and therefore the goal of an increase in conversion, Wehkamp can focus their marketing efforts on channels such as the online version of fashion magazines. Target the specified group on social media like instagram and pinterest with targeted advertising. Another way of reaching the target audience could be through the children. Families need products for their kids and there might be the chance to lure especially moms in need of toys or childs fashion onto the page.

Conclusion

Our analysis of Wehkamp's beachwear market revealed important variables influencing conversion rates and provided useful information that could increase sales. We gained an extensive understanding of customer behavior in this field by looking at a variety of variables, including demographics, device usage, weather, and economic issues. Customers who were middle-aged, have higher incomes, and have more education have the highest conversion rates, indicating that demographic factors like age, income, and education level have a significant impact. Women demonstrate higher conversion potential indicating the potential benefits of customized, gender-specific marketing. Targeting devoted and tech-savvy clients may be successful, according to the behavioral variables analysis, which showed that customers who preferred digital newspapers and consumed them frequently converted at a greater rate. Conversion was significantly affected by environmental factors including temperature and sunshine, highlighting the need of weather-responsive marketing tactics. On the other hand, variables that Wehkamp can track for sales forecasts, and COVID-19 case rates, showed an inverse effect with conversion. While the comprehensive regression model revealed an inverse effect between conversion and increased product views—a finding that implies customers who make faster decisions tend to have a higher chance of conversion—it also emphasized the positive effects of variables such as preference for "fashionable" clothing segments, higher income, and urbanization level. This analysis provides Wehkamp with a blueprint for optimizing the customer journey in the beachwear category. By focusing on demographic targeting, leveraging seasonal trends, and tailoring marketing strategies to frequent buyers and higher-income customers, Wehkamp can likely increase conversion rates and reduce sales volatility.

Appendix

Use of GenAI

For the analysis of the report, occasionally generative AI was used to enhance the research report. Where useful, generative AI was strategically deployed in order to ensure overall quality. Specifically, in the areas of code optimizing and spelling and grammar generative AI was used. Parts of the script that was used for the data merging, data cleaning and the analysis of the final dataset were made with the assistance of generative AI.

An example of this is where GenAI was used to merge the Google Trends data with the Wehkamp data. The Google Trends data was aggregated on a weekly average of interest frequency, the Wehkamp data is aggregated on a daily level. To solve this a piece of code was suggested by the GenAI to create a daily sequence and expand the data. This code was adjusted and implemented to fit with the rest of the program.
The prompt that was used to generate this code:

"Can you help me write an R script that takes a dataset with weekly data, converts the date column to Date type, and then expands it into a daily sequence for each week? I'd like to use ***lubridate*** *for date handling and* ***dplyr*** *for expanding the data. For each row, I want to generate a daily sequence starting from the weekly date, creating one row per day, and include the weekly trend value in each new row."*

Overall, the code provided answers that were not directly applicable to the script, even when large parts of the script were sent together with the prompt. Minor changes to the suggested code of the GenAI had to be made to successfully get the program running without errors. This, however, was not a large issue as the errors were mostly very intuitively solvable. Whenever a bug came up that wás harder to solve at first sight, prompting a step-by-step plan on solving the bug always proved to easen the problem. Conclusively: whether it is troubleshooting errors in the script or minor spelling mistakes in the text, GenAI is very helpful and speeds up the process of problem solving significantly.

Reflection on groupwork

Course name: Data Engineering for MADS
Group number: Group 6
Group members: Arjan Homsma (S4538897), Jakob Schmid (S6017967), Yichong Tao (S5065623), Sarah Isabella Fokken (S5395453)

Team goal

1. Good team collaboration
2. Good product delivery
3. Strive for the best (between 7-8)
4. Having fun while working on the assignments
5. Learn and grow together

Roles

Arjan: Communication, creative
Jakob: Eye for data, detail-oriented.
Sarah: leadership, disciplined, organized
Yichong: Communication, punctual

Agreements

Ways of working:

- Internal team deadline (maybe 2 days)
- Standardized times every week
- Be productive in the tutorials

Collaboration:

- Do weekly catch-ups (online)
- Respond to each others messages within a reasonable timeframe (few hours)
- Honesty (when meeting deadlines becomes hard)
- Respectfulness
- Listening to others opinions

Reflection

Like mentioned before it is a joint responsibility to have a successful cooperation. Giving feedback to each other and reflect on your own contribution is essential. Especially with respect to the agreements that have been made.

Our collaboration was a joint effort, relying on everyone’s responsibility to contribute to a successful outcome. Reflecting on our ways of working and our adherence to the agreements we made at the beginning, it is clear that we effectively supported each other through consistent communication and active participation.

1. To which extent did you meet the team goal? Compliment each other for their contribution.
We met our team goals quite well overall. Here, is our feedback to each point:

- Good communication: We communicated effectively and regularly, which helped us stay on track despite tight individual schedules.

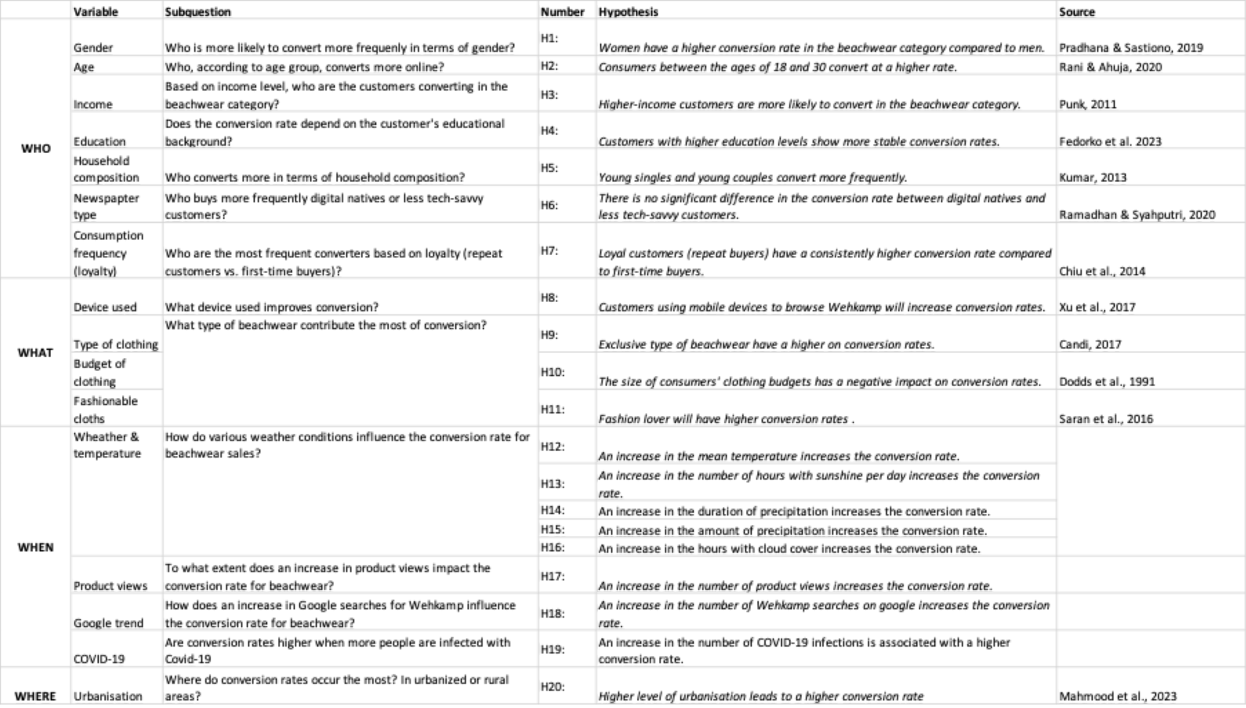
- Good Product Delivery: We delivered the best possible outcome within our scope, capabilities, and the time and effort we had available. We are proud of our work and the quality we achieved.

- Striving for the Best: Our goal was to aim for a result between a 7 and 8. While we do not yet know our final evaluation, we are hopeful that our efforts have left a positive mark.

- Having Fun While Working on the Assignments: We all remained motivated and genuinely enjoyed working on the project. The process of analyzing data, brainstorming, and collaborating made the work enjoyable for everyone.

- Learning and Growing Together: This was an important aspect of our collaboration. We discovered each other's strengths and weaknesses and saw improvements both individually and as a team. The experience allowed us to learn from each other, grow our skills, and support one another effectively.

We also gave each other compliments when someone made a notable contribution (Yichong did an amazing job), although we acknowledge that more frequent position.



	Variable	Subquestion	Number	Hypothesis	Source
WHO	Gender	Who is more likely to convert more frequenly in terms of gender?	H1:	*Women have a higher conversion rate in the beachwear category compared to men.*	Pradhana & Sastiono, 2019
	Age	Who, according to age group, converts more online?	H2:	*Consumers between the ages of 18 and 30 convert at a higher rate.*	Rani & Ahuja, 2020
	Income	Based on income level, who are the customers converting in the beachwear category?	H3:	*Higher-income customers are more likely to convert in the beachwear category.*	Punk, 2011
	Education	Does the conversion rate depend on the customer's educational background?	H4:	*Customers with higher education levels show more stable conversion rates.*	Fedorko et al. 2023
	Household composition	Who converts more in terms of household composition?	H5:	*Young singles and young couples convert more frequently.*	Kumar, 2013
	Newspapter type	Who buys more frequently digital natives or less tech-savvy customers?	H6:	*There is no significant difference in the conversion rate between digital natives and less tech-savvy customers.*	Ramadhan & Syahputri, 2020
	Consumption frequency (loyalty)	Who are the most frequent converters based on loyalty (repeat customers vs. first-time buyers)?	H7:	*Loyal customers (repeat buyers) have a consistently higher conversion rate compared to first-time buyers.*	Chiu et al., 2014
WHAT	Device used	What device used improves conversion?	H8:	*Customers using mobile devices to browse Wehkamp will increase conversion rates.*	Xu et al., 2017
	Type of clothing	What type of beachwear contribute the most of conversion?	H9:	*Exclusive type of beachwear have a higher on conversion rates.*	Candi, 2017
	Budget of clothing		H10:	*The size of consumers' clothing budgets has a negative impact on conversion rates.*	Dodds et al., 1991
	Fashionable cloths		H11:	*Fashion lover will have higher conversion rates .*	Saran et al., 2016
WHEN	Wheather & temperature	How do various weather conditions influence the conversion rate for beachwear sales?	H12:	*An increase in the mean temperature increases the conversion rate.*	
			H13:	*An increase in the number of hours with sunshine per day increases the conversion rate.*	
			H14:	An increase in the duration of precipitation increases the conversion rate.	
			H15:	An increase in the amount of precipitation increases the conversion rate.	
			H16:	An increase in the hours with cloud cover increases the conversion rate.	
	Product views	To what extent does an increase in product views impact the conversion rate for beachwear?	H17:	*An increase in the number of product views increases the conversion rate.*	
	Google trend	How does an increase in Google searches for Wehkamp influence the conversion rate for beachwear?	H18:	*An increase in the number of Wehkamp searches on google increases the conversion rate.*	
	COVID-19	Are conversion rates higher when more people are infected with Covid-19	H19:	An increase in the number of COVID-19 infections is associated with a higher conversion rate.	
WHERE	Urbanisation	Where do conversion rates occur the most? In urbanized or rural areas?	H20:	*Higher level of urbanisation leads to a higher conversion rate*	Mahmood et al., 2023