University of Groningen

Data Engineering for MADS

Assignment 2 Report

Authors:
Nils Depner (s4766946)
Daniel Hsu (s5680115)
Ruben Meijer (s4543831)
Erlangga Roesdyoko (s2667983)
Group 10

October 30, 2024

1. Introduction

We explored the management dilemma at [CENSORED], a Dutch online retail company, focusing on the beachwear category, which, despite an 18% year-on-year sales growth and a 9.72% conversion rate, experiences significant seasonal volatility. [CENSORED]'s management identifies this volatility as a potential growth opportunity rather than a challenge, aiming to increase the conversion rate by at least 1%. This led us to our central management question: How can [CENSORED] increase the conversion rate for its beachwear category?

Our research question further refined this objective, seeking to identify the specific factors influencing conversion rates for beachwear. We hypothesized that conversion is affected by customer demographics, behaviors, and environmental variables like weather. Hypotheses were developed to test various aspects, such as gender, income, customer age, device type, and environmental factors like temperature, sunlight, and number of covid cases.

The following hypotheses will be explored:

H1: Conversion rates of beach wear are higher when the gender of the customer is male.

H2: When household income is higher, conversion rates are also higher.

H3: Frequent shoppers are more likely to make purchases and thus have a higher conversion rate.

H4: Conversion rate decreases when there are more new daily covid cases

H5: Conversion rates of some brands are higher than others.

H6: The conversion rate of beachwear increases when hot weather appears consecutively.

H7: The conversion rate of beachwear is positively related to temperature.

H8: The conversion rate of beachwear is positively related to sunlight.

H9: Customer purchases predominantly occur during the first half of the week compared to the later.

H10: Customers using mobile devices have the highest percentage of conversion rate.

H11: There is a positive correlation between Google search trends for "bikini" and "zwembroek" with [CENSORED] beachwear purchases.

2. Managerial insights - Summary

This section provides a strategic outline for actionable insights derived from the hypotheses, aimed at increasing conversion rates of the beachwear category.

Demographic Insights

Hypotheses H1 and H2 examine the influence of gender and household income on conversion rates. Analysis shows that male customers exhibit a higher conversion rate (10.68%) than female customers (8.42%). This difference, though statistically significant, suggests that [CENSORED] could leverage gender-specific marketing strategies. Targeted campaigns aimed at male customers, especially reaching more customers as males represent a smaller segment relative to females, may further capitalize on this high conversion tendency. For female customers, which represent the majority of [CENSORED]’s customers in the beachwear segment, marketing

approaches that resonate with their preferences could help bridge the conversion gap. Income data indicates that lower and higher income groups have higher conversion rates. By tailoring marketing efforts toward lower and higher-income households, [CENSORED] can enhance its appeal to this segment, potentially boosting conversion rates. Furthermore, it could prove beneficial to target customers with a reported household income between €100,000 and €200,000, as this is the largest segment.

Shopping Behavior Insights

Hypotheses H3, H9, and H10 explore shopping frequency, device use, and weekly patterns. Frequent shoppers showed lower conversion rates in categories 1 and 3, which implies that targeting average-frequency shoppers yields higher conversions. However, category 3 shoppers yielded the most sales, indicating that tring to improve their conversion rate would be beneficial. Additionally, mid-week (especially Wednesday) sees the highest conversion rates (9.25%), while weekends show a dip. This pattern aligns with established shopping behaviors, suggesting [CENSORED] might enhance mid-week promotions to capture peak interest. In terms of device usage, desktops yield the highest conversion rate despite mobile devices having the most sessions. This disparity points to potential friction in the mobile purchasing process. Improving security features and product information on the mobile site could encourage more conversions among mobile users, who may hesitate to complete purchases on these devices.

Product Insights

Hypothesis H5 revealed substantial variance in conversion rates across different brands, with high-conversion brands such as Brand21 and Brand29 standing out. These brands not only convert well but also align with customer preferences. [CENSORED] should prioritize these high-performing brands in marketing and site placement. Additionally, for brands with high conversion but low purchase volume, increasing visibility and availability may drive more sales while maintaining conversion rates. Conversely, for low-conversion brands, addressing possible barriers like pricing or product imagery could improve performance. [CENSORED] may also consider streamlining the catalog to focus on higher-performing brands, optimizing the product lineup for better overall conversions.

External Driven Insights

Hypotheses H4, H6, H7, H8 and H11 examined the impact of COVID-19 cases, weather and Google search trends. Two years after the initial breakout, COVID-19 cases still affect the customer's purchase intention, the cases negatively affect the conversion rate. Thus, the increase of COVID-19 cases would lead to a less conversion rate. On the other hand, temperature and sunlight positively affect conversion rates; specifically, a 0.1°C rise in temperature correlates with a 0.01108% increase in conversion, while increased sunlight duration also boosts conversions. Additionally, consecutive hot days drive higher conversions but lower conversion rate, likely due to heightened traffic for beachwear. [CENSORED] could implement targeted ads

and promotions during the first consecutive hot days, aligning with the seasonal interest spike. Although Google search trends for keywords like "bikini" correlate with [CENSORED]'s beachwear purchases, the effect is modest. Nonetheless, search engine optimization could maintain brand visibility during peak search periods, supporting conversion goals.

3. Methodology & Data Collection

3.0 Conversion Rate Formula

In the Data Engineering course, the conversion rate formula is defined as sum(conversion) / length(unique(internet_session_id)) * 100%. However, since internet_session_id is **not** unique and can have multiple customer_ids associated with it, we introduced a new column called actual_session_id. This column combines internet_session_id and customer_id, providing a better representation of a session.

With this change, the conversion rate calculation is changed to sum(conversion) / length(unique(actual_session_id)) * 100%, using a summarized table containing actual_session_id and conversion. Including article_id in this table would lead to incorrect conversion rates because multiple articles can be viewed per actual_session_id by a customer, inflating the row count.

Using the new calculation method, the total average conversion rate of the dataset will be 52.155 / 60.0015 * 100%= 8.69% in comparison to the original method with 9.72%

For 2021 it will be 8.99% (22.275 / 247.795 * 100%) in comparison to 10.00%
For 2022 it will be 8.48% (29.880 / 352.220 * 100%) in comparison to 9.52%

3.1 Data sources

The main dataset has been acquired directly through [CENSORED], and includes data from January 1st 2021 up until December 31st 2022. Since these are the 2 years following the COVID-19 pandemic, which has had tremendous impacts on the e-commerce industry, data regarding covid cases from Github has been implemented in the analysis. Furthermore, weather data from KNMI and google trends data from google has been included to gain further insights on consumer behavior and trends. In total we have one internal dataset from [CENSORED] with four tables (events, customers, orders, and article) and three external dataset which are COVID-19 data, KNMI weather data, and Google trend data.

3.2 Data combination

In this research, we performed the data preparation by following these steps: data combination, data cleaning, and data formatting. Our consideration on doing data combination, in this case doing aggregation, before data cleaning is that we have defined our main research objective. Thus, we want to simplify our dataset to focus on our main research questions and subquestions and we have reduced our dataset to only include essential variables by doing aggregation.

The data aggregation started from our internal data, the [CENSORED] dataset, which includes four tables, namely article, customers, events and orders. We use internet_session_id from the events table as the key variable to aggregate with the customers table. Then, the table is aggregated with the article table by using customer_id as the key variable. The final table from the internal data consists of 16 columns: internet_session_id, customer_id, session_date, conversion, pdview, article_id, brand_name, class_2, class_3, live_year, device_category_desc, gender_code, geom_household_age, geom_household_income, geom_consumption_frequency, and geom_clothing_budget. Then, we change the hashed value in the gender_code variable to its original value. We also convert all categorical variables into numeric values.

We continue combining our internal data with external data in R. For aggregating the external data, we use session_date as the key variable to aggregate with the COVID dataset and weather dataset, and we create a new column containing weekly date information based on session_date as the key variable for aggregating the google trend data to the main dataset. Detailed description about each variable in our final table (Table 0.1) can be found in Appendix Description of the Variables..

3.3 Data cleaning

The data cleaning step is performed after we define our final dataset, thus it includes our main table from internal data and weather data, covid case data, and google trends data from our external data. Start with checking the summary for each variable in the final dataset and then continue with analyzing the distribution of each unique value for each variable related to customer identification to check whether there are outliers or missing data. We found a large amount of missing data in the columns related to customer identification, including customer_id, gender_code, geom_household data, and device_category_desc.

Before handling the missing observations, we identify which type of missing data type each variable has. For customer_id, the missing observations are due to the customer's decision whether to create an account or not. Thus, we categorize customer_id as Missing Not At Random (MNAR) because the missingness is related to the existence of the account itself. The missingness arises from a meaningful, non-random choice. On the other hand the missingness of the gender variable depends on the presence or absence of customer id. The same applies to the device category. Furthermore, the household information has only been bought for customers

who have a customer id, implying that the missingness of these variables also depend on the value of an observed variable (customer_id). Thus, all of these variables are categorized as Missing At Random (MAR). The presence and proportions of these missing values is visualized in Figure 0.1. Further treatment for those types of missing observations will be discussed in Chapter 3.4.

Before testing our hypothesis using statistical analysis, we performed data formatting to change the type of values in our categorical variables from character to numerical in our final dataset (Table 0.2).

3.4 Missing Observations and Outliers

There are two types of missing observations in our dataset. For Missing At Random (MAR), we can apply a statistical technique for handling missing data by using imputation, in this case Multiple Imputation by Chained Equations (MICE). MICE fills the missing values with multiple imputed dataset instead of a single guess, so it also takes into account the uncertainty of what the missing value would be. A predictor matrix was employed to ensure columns would not predict themselves, and that logical columns would predict certain columns. For example brand_name, class_2 and class_3 predict gender_code. And all of the geom_household variables are allowed to predict each other (Figure 0.2). The use of MICE is not without drawbacks, however, this method helps us maintain our sample size and preserve the power and robustness of statistical analyses, which we need to test our hypothesis further. On the other hand, in MNAR the likelihood of missingness is related to the missing data itself, in this case the cause of missingness is related to customer decisions that lead to the existence of the customer account. In this study we decided not to go deeper in solving the missingness in customer_id, since we do not use it for any hypothesis testing.

4. Descriptive Statistics

Some key descriptive statistics (Table 0.3) inform us about interesting distributions in the dataset. Our final table has 600,015 observations. Some interesting patterns can be noted, such as the mean of gender_code being 1.847, informing us that the gender distribution in the dataset is largely represented by females relative to males. For household income, the median income level is 6, while income levels 7 and 8, representing higher incomes, show a smaller but important segment. Frequency of purchase behavior is coded from 1 to 3, with a mean value of 2.4, indicating most households have moderate to high purchasing frequency. Sessions are primarily conducted on mobile devices, followed by desktops and tablets. The number of page views per session varies widely, with a mean of around 1.26 and a maximum of 12, indicating that some sessions involved extensive browsing.

5. Results

5.1 Key Findings & Recommendations

H1: Conversion rates of beach wear are higher when the gender of the customer is male.

To investigate whether there is a difference between conversion rates of different genders, a linear model has been used. For this linear model, a summarized table with the max conversion and grouped by actual_session_id and gender_code. Then the gender_code needs to be seen as a factor, followed by the computation of the linear model (Table 1.1, Figure 1.1). The results of the model show that male customers have a significantly higher conversion rate than female customers (10.68% opposed to 8.42%, p-value = <2e-16). However, nothing can be said about the difference in comparison to "Other" gender as this effect is insignificant (p = 0.558). The reason for this can be the small amount of data points with gender "Other" (Figure 1.2). When looking at these results, it can be concluded that indeed conversion rates of beachwear are higher when the gender of the customer is male.

Given these findings, [CENSORED] might consider tailoring marketing strategies to further leverage these gender-specific tendencies. [CENSORED] might choose to focus on reaching a larger male audience, or to use other strategies to try to increase the conversion rates of females to match those of males. However, it is crucial to consider the female demographic in beachwear marketing, as a significantly higher number of females are engaging with and purchasing products compared to their male counterparts (Figure 1.2).

H2: When household income is higher, conversion rates are also higher.

To investigate whether household income had a statistically significant impact on conversion rates, we conducted a chi-square test. The Chi Square test was significant (p = <2.2e-16), resulting in the rejection of the null hypothesis. A linear model was used, revealing that all income groups listed have significantly lower average conversion rates than the reference group (Geom_Household_Income = "Unknown"), with reductions ranging from approximately 13% to 15% in conversion rate. For example, households in income group 1 have a conversion rate that is approximately 12.97 percentage points lower than the reference group, on average (Table 2.1). It can also be seen that as income groups increase, conversion rates tend to decrease first, with the middle income group having the lowest conversion rate, and then increasing again towards the higher income groups (Figure 2.1). Furthermore, the data is heavily skewed, with category 7 having the largest number of observations (Figure 2.2).

The results indicate that lower-income households exhibit higher conversion rates than other income groups, while the middle class has the lowest. This suggests that in order to boost conversion rates, [CENSORED] should tailor product offerings and advertisements to different income groups, with a focus on lower- and higher-income segments. By also taking into consideration the size of these segments, [CENSORED] can aim to increase the number of customers in lower income segments, which have high conversion rates.

H3: Frequent shoppers are more likely to make purchases and thus have a higher conversion rate.

To investigate whether shopping frequency had a statistically significant impact on conversion rates, we conducted a chi-square test. The chi-square test was significant ($p < 2.2e-16$, Table 3.2), resulting in the rejection of the null hypothesis and indicating an association between shopping frequency and conversion rate. A linear model was then used to assess the specific impact of different shopping frequencies on conversion. The model (Table 3.1) revealed that all shopping frequency groups had significantly lower average conversion rates than the reference group (Geom_Consumption_Frequency = "Unknown"), with reductions ranging from approximately 12.67% to 16.24% in conversion rate. The category with the highest conversion rate is the customer group with an average frequency of shopping. The customers with the highest frequency of shopping have the lowest conversion rate, but also have by far the most sessions (Figure 3.2) and the most purchases (Figure 3.3).

Given these findings, [CENSORED] should focus on targeting customers in the average shopping frequency group, as these have the highest conversion rate. However, this is only the second largest consumption frequency group, and the consumption frequency group with the most sales has the lowest conversion rate. In addition to targeting customers in the average shopping frequency group, [CENSORED] should therefore try to use tailored marketing strategies to try and increase the conversion rate of the most highest consumption frequency group.

H4: Conversion rate decreases when there are more new daily covid cases

To investigate whether daily new covid cases had a statistically significant impact on conversion rates, we conducted a linear regression test. The linear regression (Table 4.1) is not very significant ($p = 0.0614$). However, the distribution (Figure 4.1) indicates one big outlier of daily covid case. We replaced the outlier with the mean of the neighbor value and redo the regression analysis and got a significant result with p value of 0.0328 (Table 4.2). Whenever there are 100k new covid cases, the conversion rate that day will decrease 1.022%. However, if we standardize new covid cases and perform a linear regression test with standardized temperature data, covid cases become insignificant ($p =0.293$). Thus, daily covid cases have significant impact on conversion rates, but compared to other factors, it's not that important.

H5: Conversion Rates of some brands are higher than others.

To determine if conversion rates vary significantly among brands, we conducted a Pearson's Chi-Square test for independence on conversion rates across 31 distinct brands. First, we assigned different names (Brand1 to Brand31) to the hashed brand names (Table 5.1). The results indicate a statistically significant difference in conversion rates across brands ($p < 2.2e-16$). Thus, we reject the null hypothesis of equal conversion rates, supporting H5 that conversion rates differ significantly among brands.

The calculated conversion rates for each brand (Table 5.2) reveal substantial variation, with Brand21 exhibiting the highest conversion rate at 17.17%, followed closely by Brand29 at 15.59%, and Brand2 at 15.15%. At the lower end, Brand31 shows no conversions (0%), and Brand25 and Brand10 have conversion rates of 3.30% and 4.38%, respectively. This range in conversion rates highlights that certain brands perform significantly better in terms of conversions than others (Figure 5.1).

However, the distribution plot of purchases per brand (Figure 5.2) shows that some brands dominate in purchase count, such as Brand `0, Brand 11, and Brand 18. On the other hand, Brand 21 and 29 have an extremely low purchase count.

The variation in brand conversion rates provides clear guidance for [CENSORED]'s strategy. High-conversion & purchase count brands should be prioritized in marketing, product placement, and promotions to capitalize on their appeal. [CENSORED] could also consider increasing the number of products shown in the website for high conversion but low purchase count brands such as Brand29, in hopes that as sales increase the conversion rates stay consistent.

For low-conversion brands it may be useful to identify barriers like pricing, imagery, or customer alignment. If these brands lack strong appeal, [CENSORED] might consider adjusting their positioning or even removing them from the product catalog.

H6: The conversion rate of beachwear increases when hot weather appears consecutively.

To investigate how hot weather affects the customer behavior, we compare conversion rate per day and conversions per day with temperature data. We first derive two new variables: hot_weather (set top 10% hottest days as 1 and the rest as 0), consequtive_hot_days (set consecutive hot_weather as 1 and the rest as 0). First, we perform a t-test of conversion rate on consecutive hot days with an insignificant result, with p-value of 0.2475 (Table 6.1). Thus, we should reject H6 and conclude that the conversion rate per day won’t significantly increase when hot weather appears consecutively.

However, when we looked into the graph of the conversion rate (Figure 6.1), we discovered that some winter days have higher conversion rates, which is not intuitively correct. From Figure 6.2, we can see the distribution of conversions have significant seasonal distribution. What’s more, the first few consecutive hot days almost perfectly matched the trend of the conversions. Thus, we think consecutive hot days might affect the conversions. The t-test of all conversions per day on consecutive hot days was proved significantly with p-value of 6.297e-08 (Table 6.2). The linear regression model shows that whenever there is a consecutive hot day, the conversions will increase 93.559 more conversions than normal days (65.550 conversions) (Table 6.3). We assume that consecutive hot days not only increases the conversions, but also increases the traffic, thus it won’t significantly affect the conversion rate.

Based on the findings, we can conclude that consecutive hot days might not increase conversion rate, but will increase the traffic and conversions of beachwear, especially at the first few consecutive hot weather days. We believe that the first few consecutive hot days will make customers think of the start of the summer and start thinking about shopping for beachwear.

According to these findings, [CENSORED] can increase their marketing strategy whenever they observe the consecutive hot days starting to happen and exploit the first shopping conversions of the start of the season.

H7: The conversion rate of beachwear is positively related to temperature.

To investigate how temperature affects the customer behavior, we compare conversion rate per day and the max temperature of the day. The linear regression shows that when the max temperature of the day increases 0.1 ℃ the conversion rate will significantly increase 0.01108%, with p-value smaller than 2e-16 (Table 7.1). The intercept of the conversion rate is 5.891% for 0℃ max temperature. We did a further analysis of how temperature affects the conversion rate if it is a consecutive hot day. From Table 7.2, we can see that when the day is a consecutive hot day, whenever the max temperature increases 0.1 ℃, the conversion rate decreases 0.01922%. This proves that a hotter consecutive hot day will have a lower conversion rate than a hotter normal day. The highest conversion rate might happen a few days before the hot days strike. By observing the distribution of conversion rate versus hot consecutive days (Figure 6.1), we could see that the conversion rates are lower when the consecutive hot days happen.

When the We performed granger test on the max temperature on conversion rate and discovered that for lagging 1 day to 4 days passed the granger test (Figure 7.1). This means the conversion rate would be affected by the max temperature of the previous 1 to 4 days.

Combining the discoveries in H6, [CENSORED] can boost advertising 1 to 4 days before the predicted hot days occur and slowly reduce advertising when the hot days strike to save money.

H8: The conversion rate of beachwear is positively related to sunlight.

To investigate how sunlight affects the customer behavior, we compare conversion rate per day and the sunshine duration of the day. The linear regression shows that when the sunshine duration of the day increases 0.1 hour the conversion rate will significantly increase 0.01215%, with p-value smaller than 5.2e-11 (Table 8.1). Thus we should accept H8 as conversion rate is positively related to sunshine duration.

To sum up the discoveries from H6, H7, H8, we perform a linear regression for sunlight duration, max temperature on consecutive hot days on conversion rate. All variables have a significant effect on conversion rate. Sunshine duration will increase 0.004045% of the conversion rate for every 0.1 hour more sunshine duration, with p-value 0.05036; max temperature will increase 0.01252% of the conversion rate for every 0.1 ℃ increasing, with p-value <2e-16; when the day is consecutive hot weather, max temperature will reduce 0.02197% of the conversion rate for every 0.1 ℃ increasing, with p-value of 0.00239. Then we conduct linear regression of standardized sunshine duration and max temperature to compare the importance between sunshine duration and max temperature. From Table 8.3, we can see that max temperature (0.0088496) is more influencing than sunshine duration (0.0017698). Thus, we can conclude that sunshine duration has the least effect on conversion rate and temperature will increase conversion rate if the weather is not too hot.

H9: The conversion rate is higher during the first half of the week compared to the later.
To examine differences in conversion rates by day of the week, we used a summarized table of actual_session_id, DayOfTheWeekNumber, and conversion, calculating conversion rate per session. After converting DayOfTheWeekNumber to a factor, a linear model revealed that Wednesday and the weekend (Saturday and Sunday) have significantly different conversion rates compared to Monday (Figures 9.1 and 9.2). Monday's conversion rate is around 8.89%; on Wednesday, it rises by 0.36% (9.25%, p = 0.00665), while on Saturday and Sunday, it drops by about 1% (7.88% and 7.87%, p-values of 1.46e-12 and 1.14e-15).
This pattern suggests higher conversion rates in the first half of the week, peaking on Wednesday, and supports Gong et al. (2019), who noted midweek online shopping peaks. Consumers typically browse early in the week and complete purchases by Wednesday, with lower weekend conversions due to offline activities and family time.

H10: Customers using mobile devices have the highest percentage of conversion rate.
To assess the significance of device type on conversion, a Chi-square test was conducted, showing a significant association (p = 2.2e-16, Figure 10.1). However, H10 was rejected as desktops had the highest conversion rate, not mobile devices (Figure 10.2). Despite this, mobile devices accounted for 79% of sessions (476k sessions, Figure 10.3). According to de Haan et al. (2018), customers feel more secure purchasing on fixed devices, leading them to browse on mobile but complete purchases on desktops.
This suggests that improving the mobile experience alone may not solve the conversion rate issue. However, adding security features and detailed product information on mobile platforms could reduce friction, helping customers complete purchases without switching devices. Enhancing these aspects on mobile may boost the overall conversion rate effectively.

H11: There is a positive correlation between Google search trends for "bikini" and "zwembroek" with [CENSORED] beachwear purchases.
To investigate the correlation between google search and [CENSORED]'s conversion rate, we used point-biserial correlation using Pearson method to test the correlation of google search trend with conversion. The test shows that there is a significant relationship between bikini and zwembroek google search and the conversion, however the relationship has a very low positive correlation (Figure 11.1). In reality, this correlation may not have strong predictive value due to its small effect size thus may not be a strong predictor of conversion on its own.
We performed another research on google search, where we do another point-biserial correlation test for [CENSORED] and its competitors' as a keywords to check whether it has an effect on conversion (Figure 11.2). The result is the same as before, statistically significant but very low impact on conversion. In the digital era, we believe the impact of digital presence would have higher correlation with customer's engagement with a company, including purchases. However, we can conclude that search engine optimization on google search alone would not produce a significant result on conversion rates.

Appendices

Description of the Variables

Table 0.1: Descriptive Statistics

No	Column Name	Description
1	internet_session_id	Identification for the session
2	customer_id	Identification of the customer
3	session_date	The date of the internet session, implied variable from internet_session_id
4	conversion	Identification of purchase (1,0), implied variable from action_type_desc
5	pdview	Implied variable from the amount of actions per session_id
6	article_id	Identification of the article
7	brand_name	Identification of the brand name
8	class_2	Second-level merchandise classification
9	class_3	Third-level merchandise classification
10	live_year	The year the article was live on platforms
11	device_category_desc	Category of the device was used during the session: desktop, mobile, tablet
12	gender_code	Identification of the gender
13	geom_household_age	Age category of the household: 0-13 0. Unknown 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

No	Column Name	Description
		10. 65 - 69 years 11. 70 - 74 years 12. 75 - 79 years 13. >= 80 years
14	geom_household_income	Income category of the household: 0-8 0. Unknown 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
15	geom_consumption_frequency	Consumption frequency of the household: 0-4 0. Unknown 1. Little 2. Average 3. Much
16	geom_clothing_budget	Budget allocated to clothing: 0. Unknown 1. Little 2. Below average 3. Average 4. Above average 5. Much
17	mean_temp	Daily mean temperature in (0.1 degrees Celsius)
18	wind_gust_speed	Daily mean wind speed (in 0.1 m/s)
19	sunshine_duration	Sunshine duration (in 0.1 hour) calculated from global radiation (-1 for <0.05 hour)
20	precipitation_total	Daily precipitation amount (in 0.1 mm) (-1 for <0.05 mm)
21	mean_cloud_cover	Mean daily cloud cover (in octants, 9 = sky invisible)
22	max_temp	Maximum temperature (in 0.1 degrees Celsius)
23	min_temp	Minimum temperature (in 0.1 degrees Celsius)
24	Cumulative_Covid_Cases	The cumulative number of covid cases up to the given date

No	Column Name	Description
25	Month	Month of the session
26	Year	Year of the session
27	DayofTheWeek	Day of the session
28	YearWeek	Combination of Year and Week of the session
29	Week	Week of the session
30	[CENSORED]	The number of google search on “[CENSORED]” in a week
31	bikini	The number of google search on “bikini” in a week
32	zwembroek	The number of google search on “zwembroek” in a week
33	Zalando	The number of google search on “Zalando” in a week
34	CA	The number of google search on “C&A” in a week
35	actual_session_id	Implied variable indicating internet_session_id per unique customer_id

Table 0.2: Transformation of categorical to numerical variables

No	Column	Character Values	Numerical Values
1	device_category_desc	mobile, desktop, tablet	1, 2, 3
2	gender_code	Male, Female, Other	1, 2, 3
3	geom_household_age	See Appendix Description of variables - No. 13	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
4	geom_household_income	See Appendix Description of variables - No. 14	0, 1, 2, 3, 4, 5, 6, 7, 8
5	geom_consumption_ frequency	See Appendix Description of variables - No. 15	0, 1, 2, 3
6	geom_clothing_budget	See Appendix Description of variables - No. 16	0, 1, 2, 3

7	brand_name	See Appendix Description of variables - No. 7	Converted using as.factor for MICE
8	class_2	See Appendix Description of variables - No. 8	Converted using as.factor for MICE
9	class_3	See Appendix Description of variables - No. 9	Converted using as.factor for MICE

Table 0.3: Descriptive Statistics

```
internet_session_id customer_id        session_date         conversion          pdview         article_id
Length:600015       Length:600015      Length:600015      Min.   :0.00000   Min.   : 0.0000   Length:600015
Class :character    Class :character   Class :character   1st Qu.:0.00000   1st Qu.: 1.0000   Class :character
Mode  :character    Mode  :character   Mode  :character   Median :0.00000   Median : 1.0000   Mode  :character
                                                          Mean   :0.08692   Mean   : 0.9645
                                                          3rd Qu.:0.00000   3rd Qu.: 1.0000
                                                          Max.   :1.00000   Max.   :12.0000

 brand_name          class_2            class_3           live_year         device_category_desc  gender_code
Length:600015      Length:600015      Length:600015      Length:600015      Min.   :1.000        Min.   :1.000
Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000        1st Qu.:2.000
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.000        Median :2.000
                                                                            Mean   :1.267        Mean   :1.847
                                                                            3rd Qu.:1.000        3rd Qu.:2.000
                                                                            Max.   :3.000        Max.   :3.000

geom_household_age geom_household_income geom_consumption_frequency geom_clothing_budget   mean_temp
Min.   : 0.000     7      :279402        Min.   :0.000              Min.   :0.000        Min.   :-58
1st Qu.: 4.000     6      :102387        1st Qu.:2.000              1st Qu.:2.000        1st Qu.:112
Median : 6.000     8      : 67849        Median :3.000              Median :3.000        Median :165
Mean   : 5.869     5      : 55959        Mean   :2.396              Mean   :3.136        Mean   :149
3rd Qu.: 7.000     4      : 43439        3rd Qu.:3.000              3rd Qu.:4.000        3rd Qu.:184
Max.   :13.000     3      : 27625        Max.   :3.000              Max.   :5.000        Max.   :269
                   (Other): 23354
wind_gust_speed sunshine_duration precipitation_total mean_cloud_cover    max_temp        min_temp
Min.   : 8.00   Min.   :  0.00    Min.   : -1.00      Min.   :0.000    Min.   :-48.0   Min.   :-109.00
1st Qu.:21.00   1st Qu.: 37.00    1st Qu.:  0.00      1st Qu.:4.000    1st Qu.:162.0   1st Qu.:  60.00
Median :28.00   Median : 75.00    Median :  0.00      Median :6.000    Median :210.0   Median : 102.00
Mean   :29.72   Mean   : 76.27    Mean   : 20.43      Mean   :5.388    Mean   :198.2   Mean   :  91.24
3rd Qu.:35.00   3rd Qu.:120.00    3rd Qu.: 12.00      3rd Qu.:7.000    3rd Qu.:239.0   3rd Qu.: 129.00
Max.   :89.00   Max.   :155.00    Max.   :428.00      Max.   :8.000    Max.   :355.0   Max.   : 188.00

Cumulative_Covid_Cases    Month                Year      DayOfTheWeek         YearWeek            Week
Min.   : 813765        Length:600015      Min.   :2021   Length:600015      Length:600015      Length:600015
1st Qu.:1766102        Class :character   1st Qu.:2021   Class :character   Class :character   Class :character
Median :7897946        Mode  :character   Median :2022   Mode  :character   Mode  :character   Mode  :character
Mean   :5350808                           Mean   :2022
3rd Qu.:8190158                           3rd Qu.:2022
Max.   :8569095                           Max.   :2022

   Wehkamp          bikini        zwembroek         Zalando            CA        actual_session_id
Min.   :27.00   Min.   : 5.00   Min.   : 1.000   Min.   :36.00   Min.   :15.00   Length:600015
1st Qu.:31.00   1st Qu.:14.00   1st Qu.: 3.000   1st Qu.:44.00   1st Qu.:21.00   Class :character
Median :35.00   Median :20.00   Median : 6.000   Median :49.00   Median :23.00   Mode  :character
Mean   :35.82   Mean   :20.91   Mean   : 5.936   Mean   :48.63   Mean   :24.53
3rd Qu.:40.00   3rd Qu.:25.00   3rd Qu.: 9.000   3rd Qu.:53.00   3rd Qu.:28.00
Max.   :61.00   Max.   :41.00   Max.   :13.000   Max.   :78.00   Max.   :58.00

  pred_base
Min.   :0.04916
1st Qu.:0.08647
Median :0.08712
Mean   :0.08692
3rd Qu.:0.08712
Max.   :0.23453
```

Figure 0.1: Missing Values

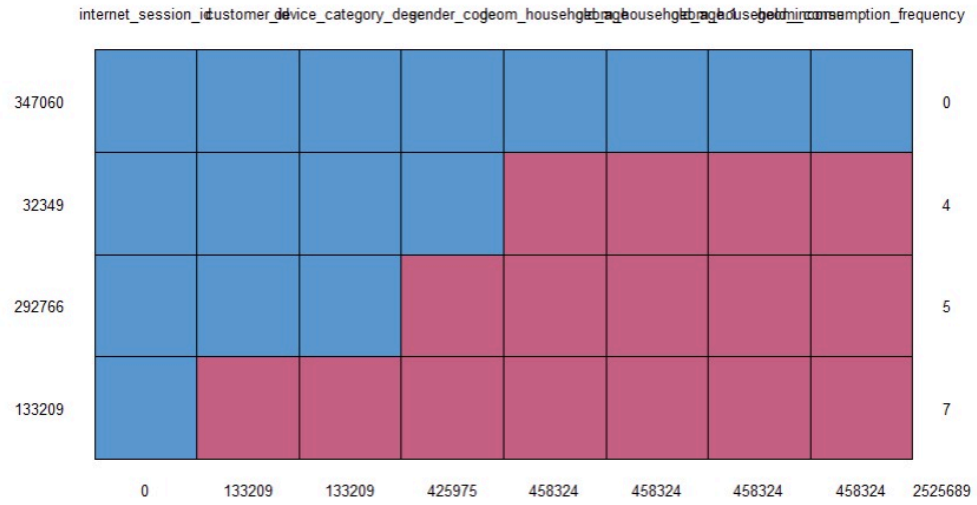


Figure 0.2: PredictorMatrix

```
> print(PredictorMatrix)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17]
 [1,]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [2,]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [3,]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [4,]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [5,]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [6,]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 [7,]    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0
 [8,]    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0
 [9,]    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0
[10,]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
[11,]    0    0    0    0    0    0    0    0    0    0    0    0    1    1    1    1    0
[12,]    0    0    0    0    0    0    0    0    0    0    1    0    1    1    1    1    0
[13,]    0    0    0    0    0    0    0    0    0    0    1    0    0    1    1    1    0
[14,]    0    0    0    0    0    0    0    0    0    0    1    0    1    0    1    1    0
[15,]    0    0    0    0    0    0    0    0    0    0    1    0    1    1    0    1    0
[16,]    0    0    0    0    0    0    0    0    0    0    1    0    1    1    1    0    0
[17,]    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

1. Tables & Figures

Table 1.1: Gender linear model

```
> actual_session_table <- CustomerConversion %>%
+   arrange(desc(conversion)) %>%
+   distinct(actual_session_id, .keep_all = TRUE)
>
> actual_session_table$gender_code<- as.factor(actual_session_table$gender_code)
> model3 <- lm(conversion ~ gender_code, data = actual_session_table)
>
> summary(model3)

Call:
lm(formula = conversion ~ gender_code, data = actual_session_table)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14286 -0.08421 -0.08421 -0.08421  0.91579

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.106836   0.001051 101.671   <2e-16 ***
gender_code2 -0.022623   0.001120 -20.200   <2e-16 ***
gender_code3  0.036022   0.061465   0.586    0.558
---
```

Figure 1.1: Visualization of gender conversion rates

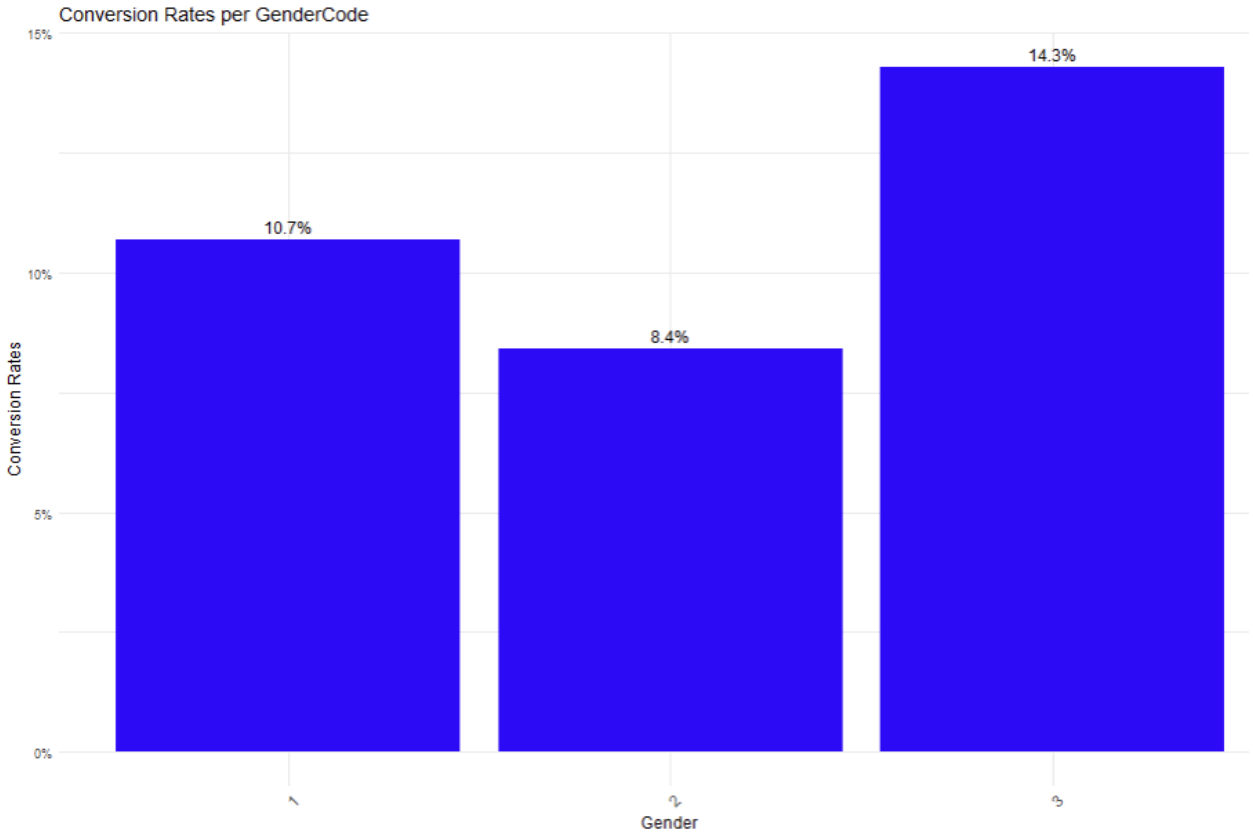


Figure 1.2: Distribution of genders

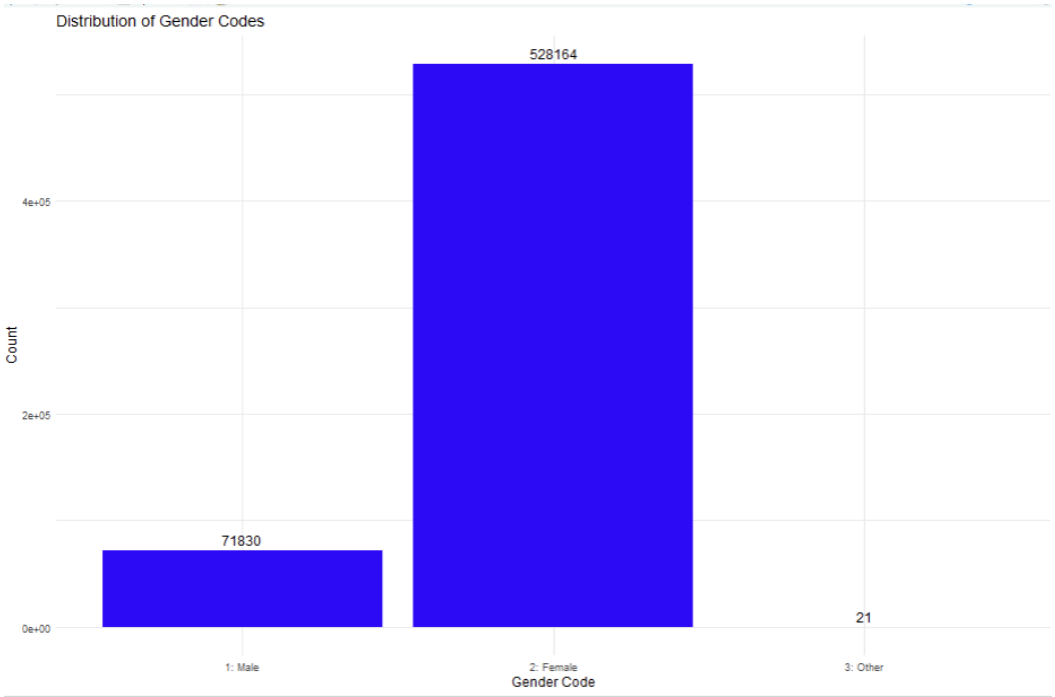


Table 2.1: Household income linear model

Distribution of Gender Codes

```
Call:
lm(formula = conversion ~ geom_household_income, data = actual_session_table)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23453 -0.08994 -0.08429 -0.08136  0.92437

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.234535   0.004616   50.81   <2e-16 ***
geom_household_income1 -0.129694   0.006695  -19.37   <2e-16 ***
geom_household_income2 -0.138620   0.005422  -25.57   <2e-16 ***
geom_household_income3 -0.148943   0.005014  -29.71   <2e-16 ***
geom_household_income4 -0.158902   0.004784  -33.22   <2e-16 ***
geom_household_income5 -0.153170   0.004758  -32.19   <2e-16 ***
geom_household_income6 -0.151424   0.004709  -32.16   <2e-16 ***
geom_household_income7 -0.144595   0.004647  -31.11   <2e-16 ***
geom_household_income8 -0.150245   0.004709  -31.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2814 on 600006 degrees of freedom
Multiple R-squared:  0.002007,  Adjusted R-squared:  0.001994
F-statistic: 150.8 on 8 and 600006 DF,  p-value: < 2.2e-16
```

Figure 2.1: Conversion Rates by Household

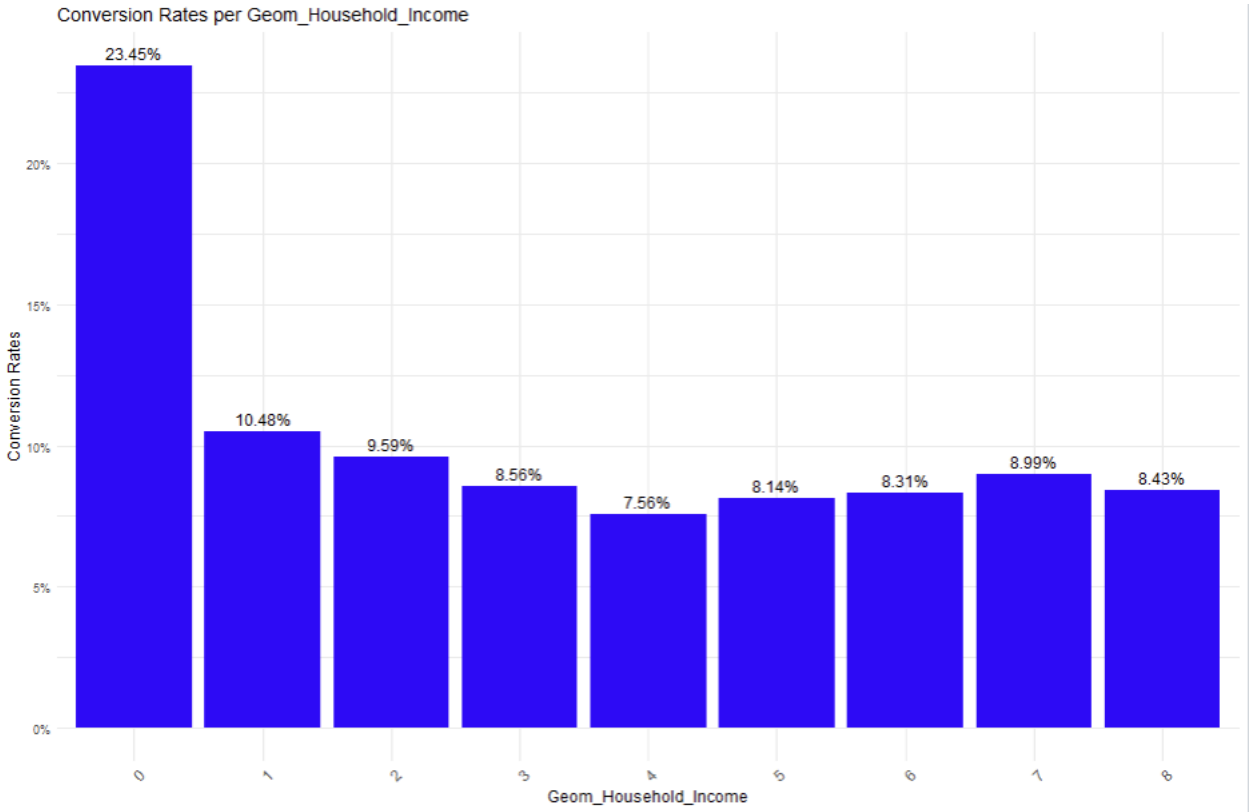


Figure 2.2: Distribution of household income categories

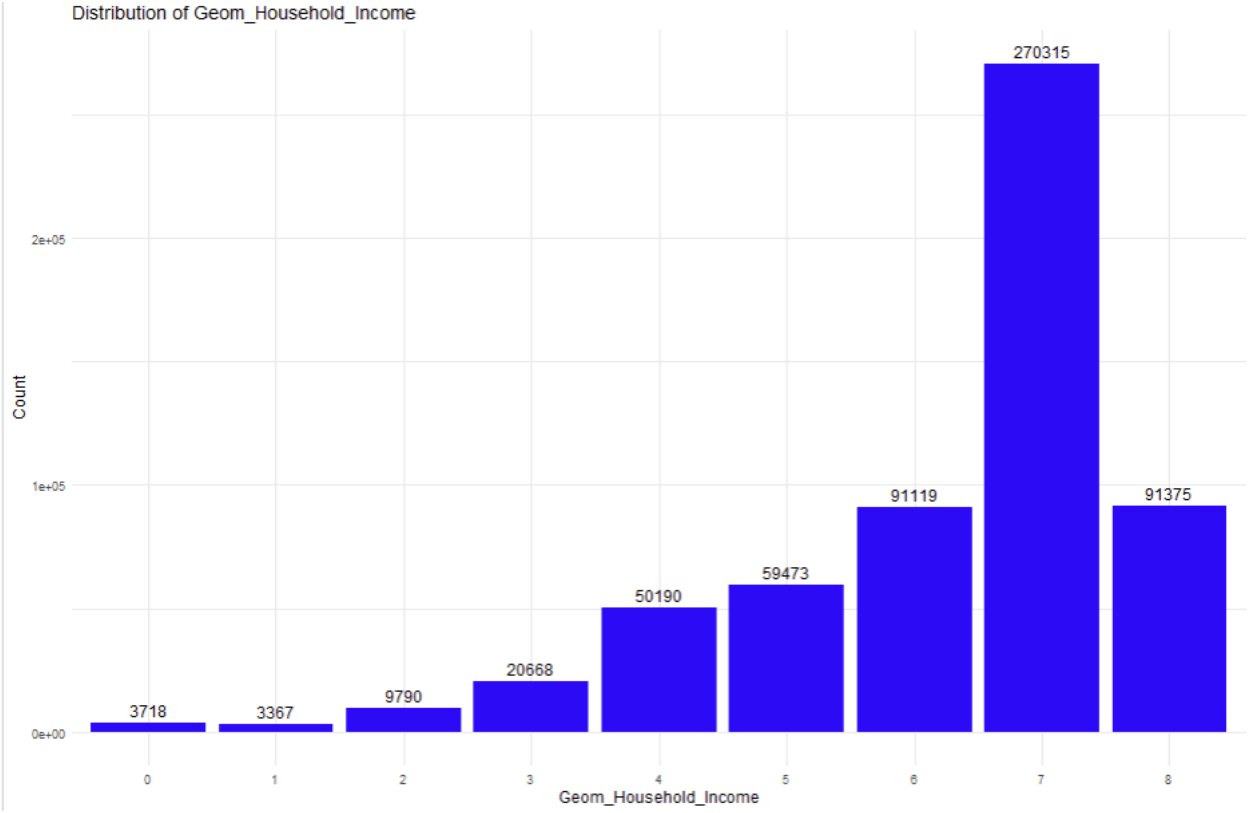


Table 3.1: Household consumption linear model

```
> summary(model_consumption)

Call:
lm(formula = conversion ~ geom_consumption_frequency, data = actual_session_table)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23453 -0.10785 -0.07212 -0.07212  0.92788

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.234535   0.004609   50.89   <2e-16 ***
geom_consumption_frequency1 -0.138832   0.004720  -29.41   <2e-16 ***
geom_consumption_frequency2 -0.126688   0.004655  -27.21   <2e-16 ***
geom_consumption_frequency3 -0.162419   0.004634  -35.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.281 on 600011 degrees of freedom
Multiple R-squared:  0.005054,  Adjusted R-squared:  0.005049
F-statistic:  1016 on 3 and 600011 DF,  p-value: < 2.2e-16
```

Table 3.2: Chi-squared test - Geom_Consumption_Frequency and Conversion

```
> chisq_consumption_conversion <- chisq.test(table(actual_session_table$geom_consumption_frequency, actual_session_table$conversion))
> print(chisq_consumption_conversion)

        Pearson's Chi-squared test

data:  table(actual_session_table$geom_consumption_frequency, actual_session_table$conversion)
X-squared = 3032.6, df = 3, p-value < 2.2e-16
```

Figure 3.1: Conversion rates across household consumption levels

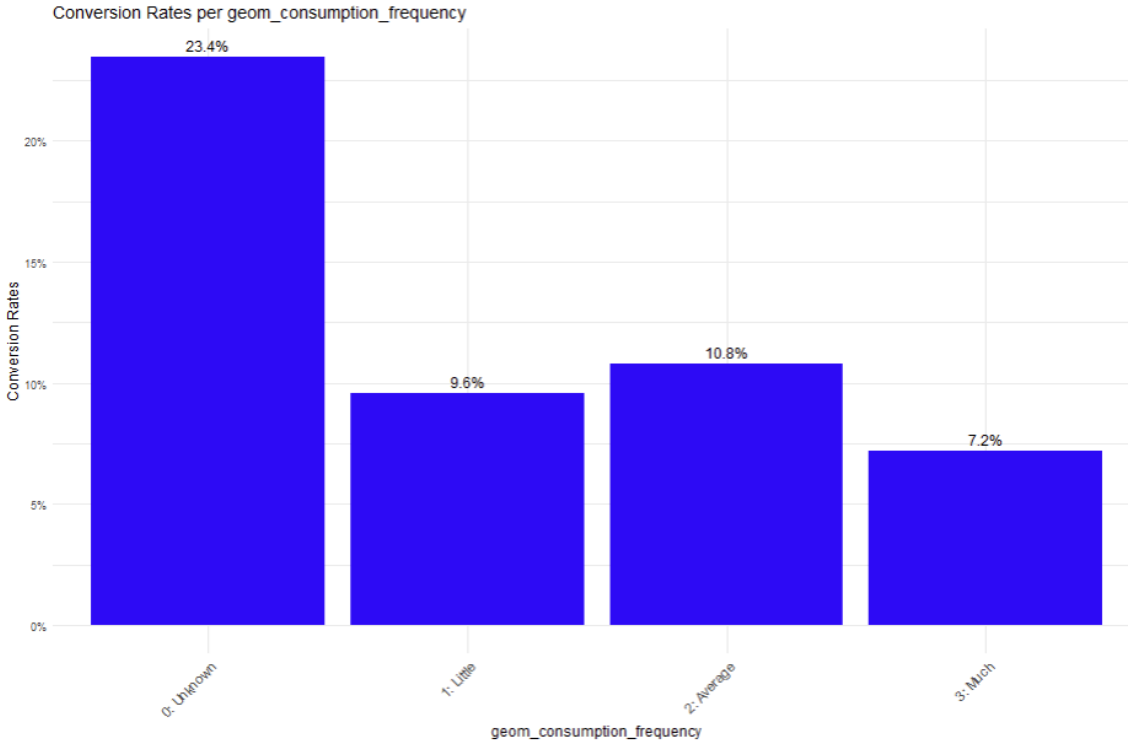


Figure 3.2: Distribution of geom_consumption_frequency

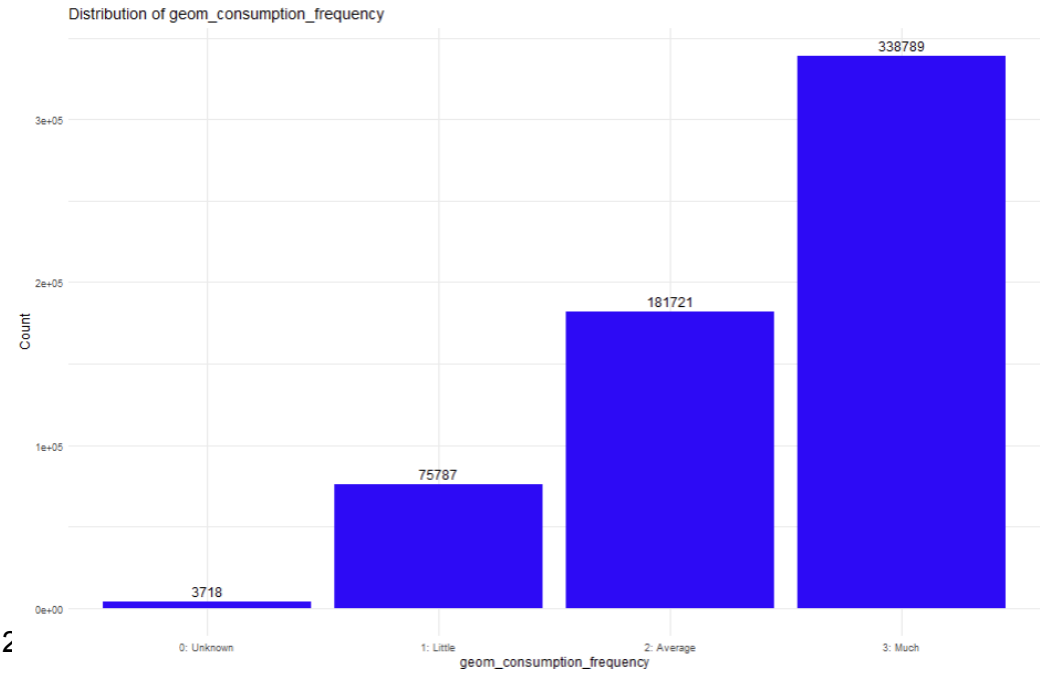


Figure 3.3: Total Conversion (Sales) per Consumption Frequency

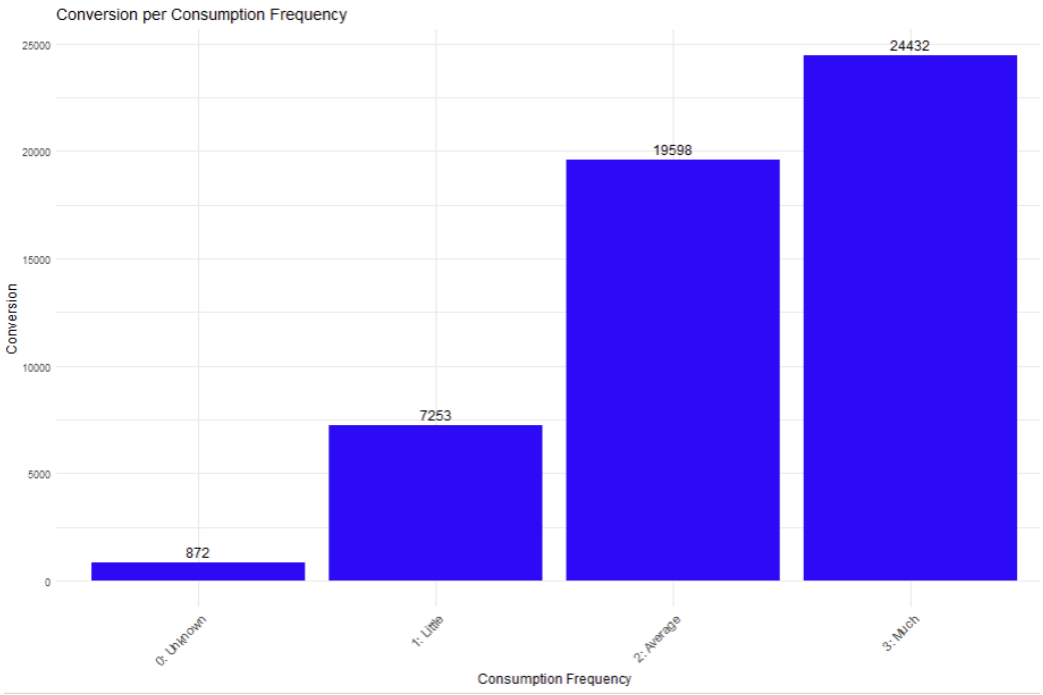


Table 4.1 : Linear regression on daily covid new cases

```
> H4 <- lm(conversion_rate ~ Covid_increase, data = Day_Covid_Merged_Table)
> summary(H4)

Call:
lm(formula = conversion_rate ~ Covid_increase, data = Day_Covid_Merged_Table)

Residuals:
      Min        1Q    Median        3Q       Max
-0.067634 -0.012411  0.002382  0.015491  0.066001

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.973e-02  9.109e-04  87.524   <2e-16 ***
Covid_increase -7.040e-08  3.758e-08  -1.873   0.0614 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0221 on 727 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.004804,  Adjusted R-squared:  0.003435
F-statistic: 3.509 on 1 and 727 DF,  p-value: 0.06144
```

Figure 4.1 : Conversion rate per day with Covid new cases

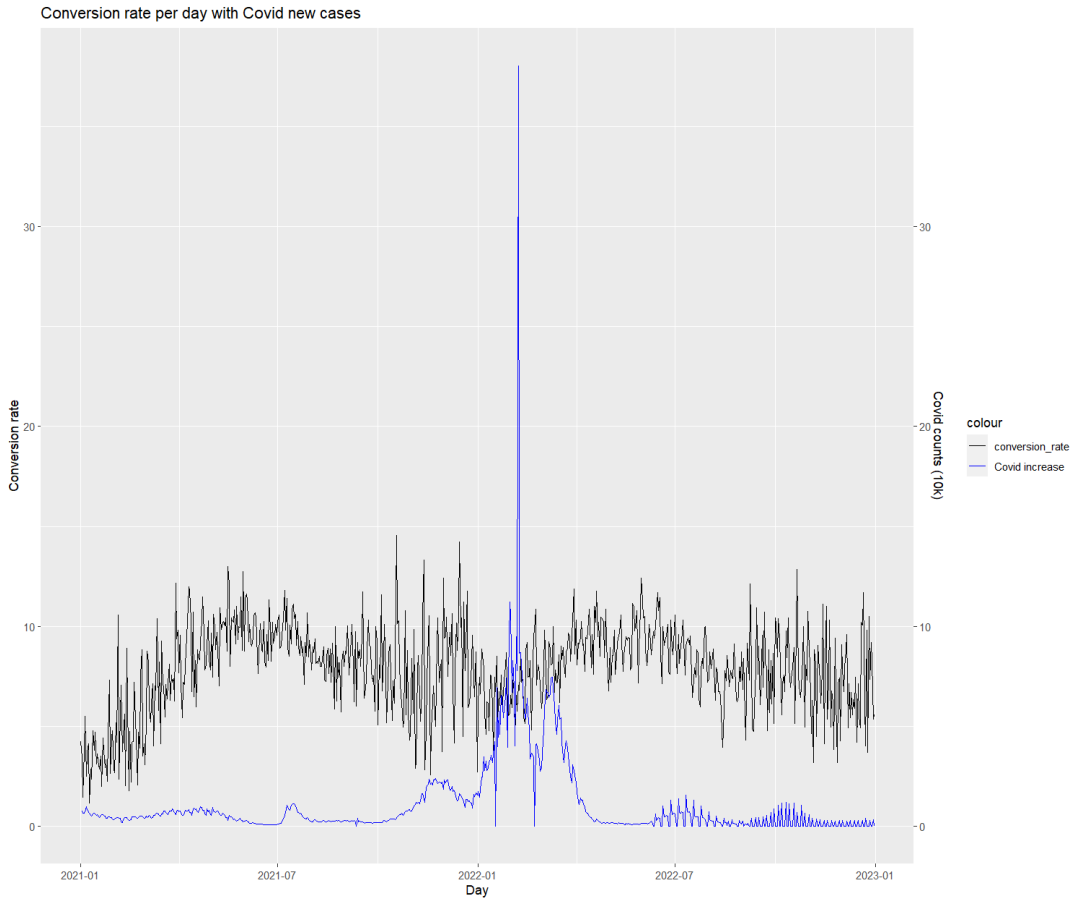


Table 4.2 : Linear regression on daily covid new cases without outlier

```
> H4_outlier <- lm(conversion_rate ~ Covid_increase_outlier, data = Day_Covid_Merged_Table)
> summary(H4_outlier)

Call:
lm(formula = conversion_rate ~ Covid_increase_outlier, data = Day_Covid_Merged_Table)

Residuals:
      Min        1Q    Median        3Q       Max
-0.067720 -0.012608  0.002309  0.015353  0.065829

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             8.002e-02  9.528e-04  83.991   <2e-16 ***
Covid_increase_outlier -1.022e-07  4.780e-08  -2.138   0.0328 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02208 on 727 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.006251,  Adjusted R-squared:  0.004884
F-statistic: 4.573 on 1 and 727 DF,  p-value: 0.03282
```

Figure 4.2 : Conversion rate per day with Covid new cases

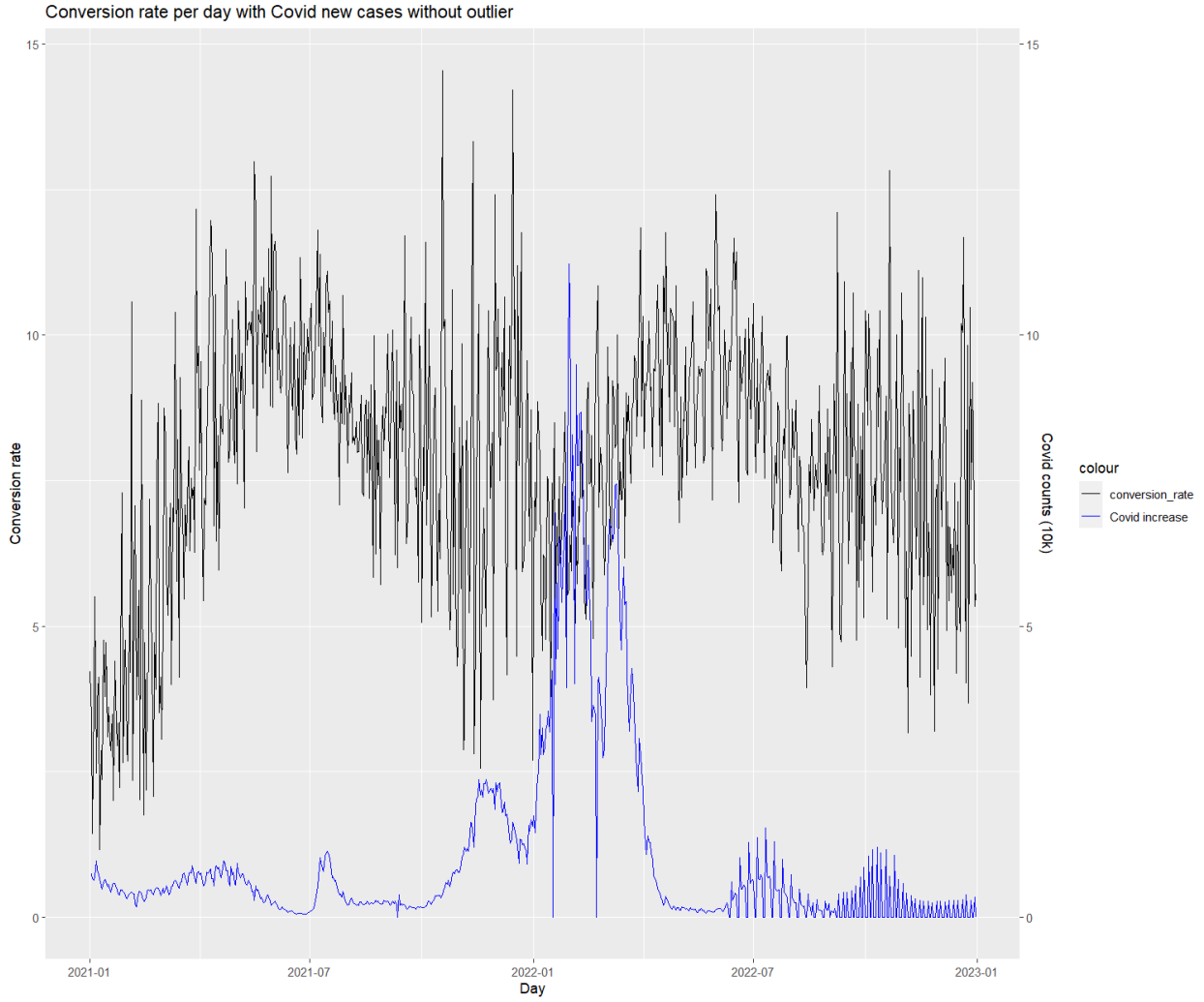


Table 4.3 : Linear regression on standardized covid new cases without outlier and standardized max temperature

```
> H4_outlier_std <- lm(conversion_rate ~ std_covid + std_max_temp, data = Day_Covid_Merged_Table)
> summary(H4_outlier_std)

Call:
lm(formula = conversion_rate ~ std_covid + std_max_temp, data = Day_Covid_Merged_Table)

Residuals:
      Min        1Q    Median        3Q       Max
-0.059095 -0.011736  0.001119  0.014555  0.067618

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0789638  0.0007680 102.815   <2e-16 ***
std_covid    0.0008542  0.0008121   1.052    0.293
std_max_temp 0.0080639  0.0008126   9.924   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02074 on 726 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.125,     Adjusted R-squared:  0.1225
F-statistic: 51.84 on 2 and 726 DF,  p-value: < 2.2e-16
```

Table 5.1 : Hashed Brand Names with matching Brand#

	Hashed_Brand_Value	Brand_Label
1	98a96fa8a4de67ce18da5e0ec1de640702cea7ae054225f70e85b74b	Brand1
2	5bcbe0a1bc3731cab57367fc24a4e31818c5c8eeb3723164b4da1cc7	Brand2
3	e57e600c55ae45a6eacacde6f4a8b6cb43b3488dfb80b8d8154fa9d0	Brand3
4	10806da77b0fa0ea52bc425a457e9d3eaf8309dc02a805d198c9d0e4	Brand4
5	6b50157fb0c5a443e84e033dfc5d48970e05e9ef716e4d40efcab51e	Brand5
6	9f58f6f50a6359f7ac02d569a5242c303067f04e8bc9dbac0421d30d	Brand6
7	b4a8f26e12e23444c09ef766b7b6abc7b4deba92f7820e9a1ee2f2b8	Brand7
8	8d294f5ea7b7a8af081dc6e42cb8272fe6eeeafbac094891302f7a30	Brand8
9	4a1b8d62da25fc57119d1b6daee92aa36da09d51c4349877c717a16a	Brand9
10	298592e67e5f682be6ffcae269605c43bfb89adb877fd65bfe18503c	Brand10
11	ad30737c31ed4d577144dc2b4d1bbf375bb8b4c7c8c651014b802646	Brand11
12	5bc8219c995b80eb7a108600784193532a7646e9ad57afd5bf26eb8b	Brand12
13	fc525cf56b917ae4330e05415ffb221f81f5356b80983318309ce89d	Brand13
14	d49bcde038b0e29f73f408dbb6196d5328c89e7296dfd1f803ab8b65	Brand14
15	b7212013d20d13bade477470fbe11b6610bd3dd713618c8c3d6147e3	Brand15
16	b5d243270eb4d1fe799b1b05c957d5ed521893a78e1af7a00671f404	Brand16
17	ab6367e28161e90479df418c0dd968a9df9b89315178edeb2a892dcb	Brand17
18	008ef3eec37727573b8fd5225b200433ab64ed28b90f7fbbfaf634e5	Brand18
19	99647b1963fad175769c0b0e235422c152099a517eb6912d591880e3	Brand19
20	c7b8762038d9770cd9eb152ddc0837c2a75ef0030bbc877dc8c3782c	Brand20
21	93b8f43a826b433e8e1b354453b89ffb73dba4a3d3d01f05634ff313	Brand21
22	0b1da5ba53e2fe4972aa50c481ed54c4172d45b24c43c4ecd1097809	Brand22
23	d6e6100c42d4b571272eb6584d07760db8e8e4ef3223fd1db18039a4	Brand23
24	a51bc9f20418bba0d7a18f2a1bbd182039bcd4907ecec991f85d8f36	Brand24
25	311d6634d9460cf55a7da3a87d85d24d49c4cd845669229ac86249ec	Brand25
26	da952362ac3df58afabd8c89b8246e6fe5705ade7c5ee7fd5ca66a56	Brand26
27	330539c7bd8c8b23f000b03ff05998f11dbdb7a6da75174d2b9a5873	Brand27
28	801d60da07c7f0a8d18c5126b071684c0f39f72d46ecb76b4c97619a	Brand28
29	85869feed3ee905db9aa4086228ad6b5f5c2d9f6b3b4ab43d6d818bb	Brand29
30	b5dd00bdc5b35365902d98fc4a0893d8790cf324c5e146718ee24d20	Brand30
31	6fd1023c26a329213f7757995cfe39275636f5a4ce680fdf1e9b481e	Brand31

Table 5.2 : Conversion Rates across brands

	brand_name	conversion_rate
1	Brand21	17.174515
2	Brand29	15.591398
3	Brand2	15.149197
4	Brand15	12.711989
5	Brand12	12.631727
6	Brand6	12.163494
7	Brand14	11.581751
8	Brand5	11.433959
9	Brand4	11.364852
10	Brand8	11.100837
11	Brand1	10.733548
12	Brand22	10.478876
13	Brand11	10.105795
14	Brand27	9.995728
15	Brand3	9.618027
16	Brand24	9.458298
17	Brand30	8.776267
18	Brand16	8.701739
19	Brand17	8.068813
20	Brand7	7.940933
21	Brand28	7.618734
22	Brand13	7.244733
23	Brand9	6.658429
24	Brand23	6.425532
25	Brand18	5.992138
26	Brand20	5.720619
27	Brand26	5.087960
28	Brand19	5.077951
29	Brand10	4.379781
30	Brand25	3.300639
31	Brand31	0.000000

Figure 5.1: Brand Conversion rates

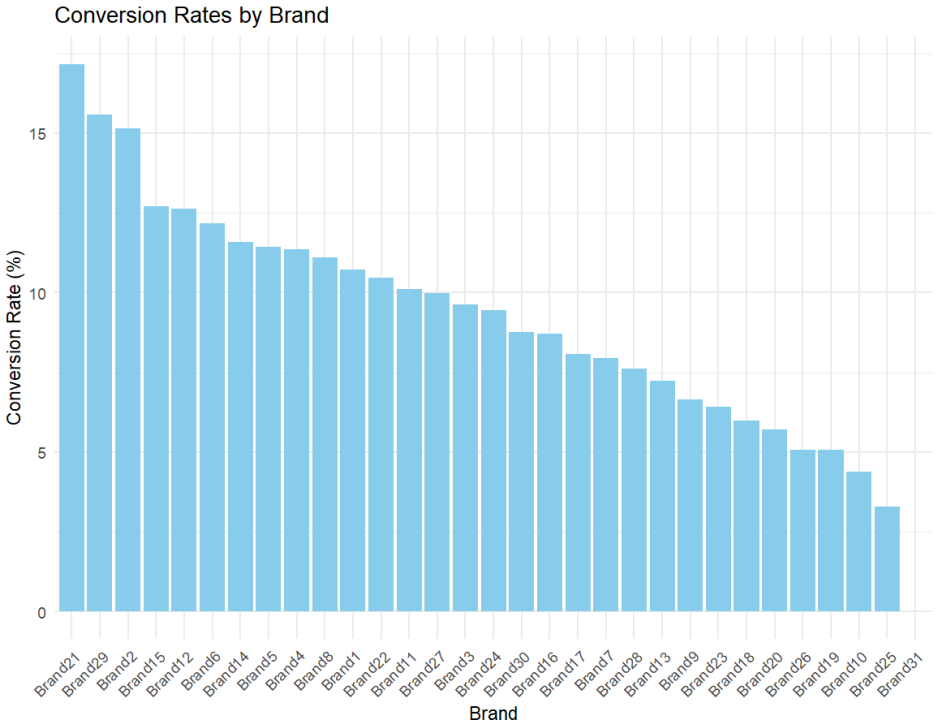


Figure 5.2: Purchase counts of brands

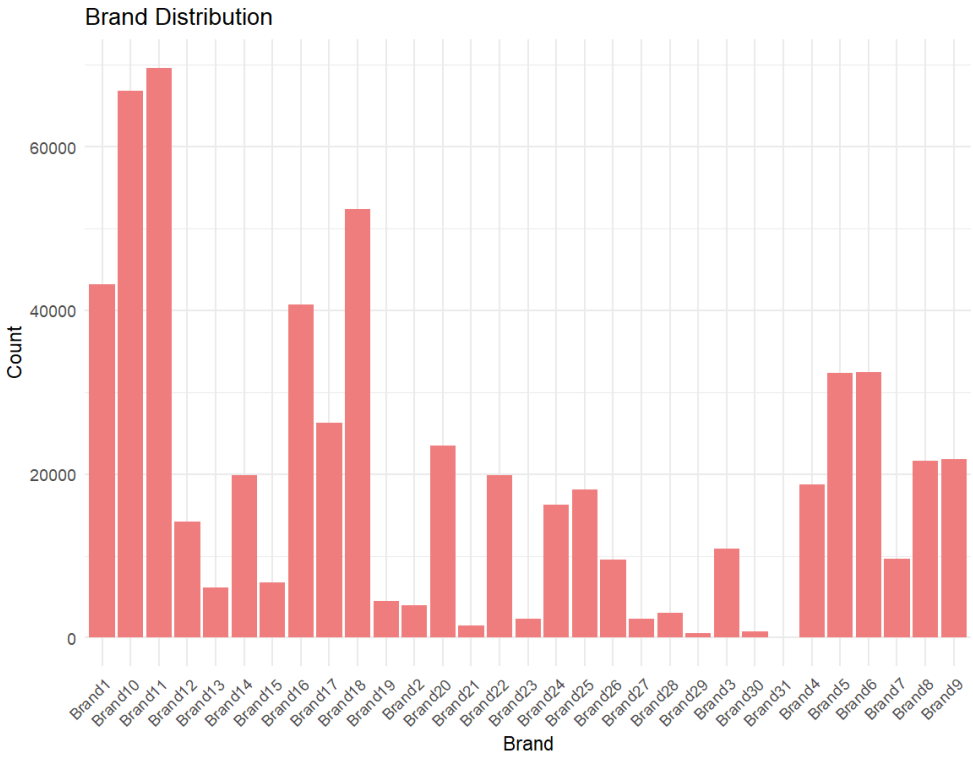


Table 6.1: Conversion rates, consecutive hot days t-test

```
> t.test(conversion_rate ~ consecutive_hot_days, data=Day_Weather_Merged_Table)# insignificant

        Welch Two Sample t-test

data:  conversion_rate by consecutive_hot_days
t = -1.1688, df = 55.035, p-value = 0.2475
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.008774560  0.002309858
sample estimates:
mean in group 0 mean in group 1
     0.07872460      0.08195695
```

Figure 6.1: Conversion rate per day with temperature and consecutive hot days

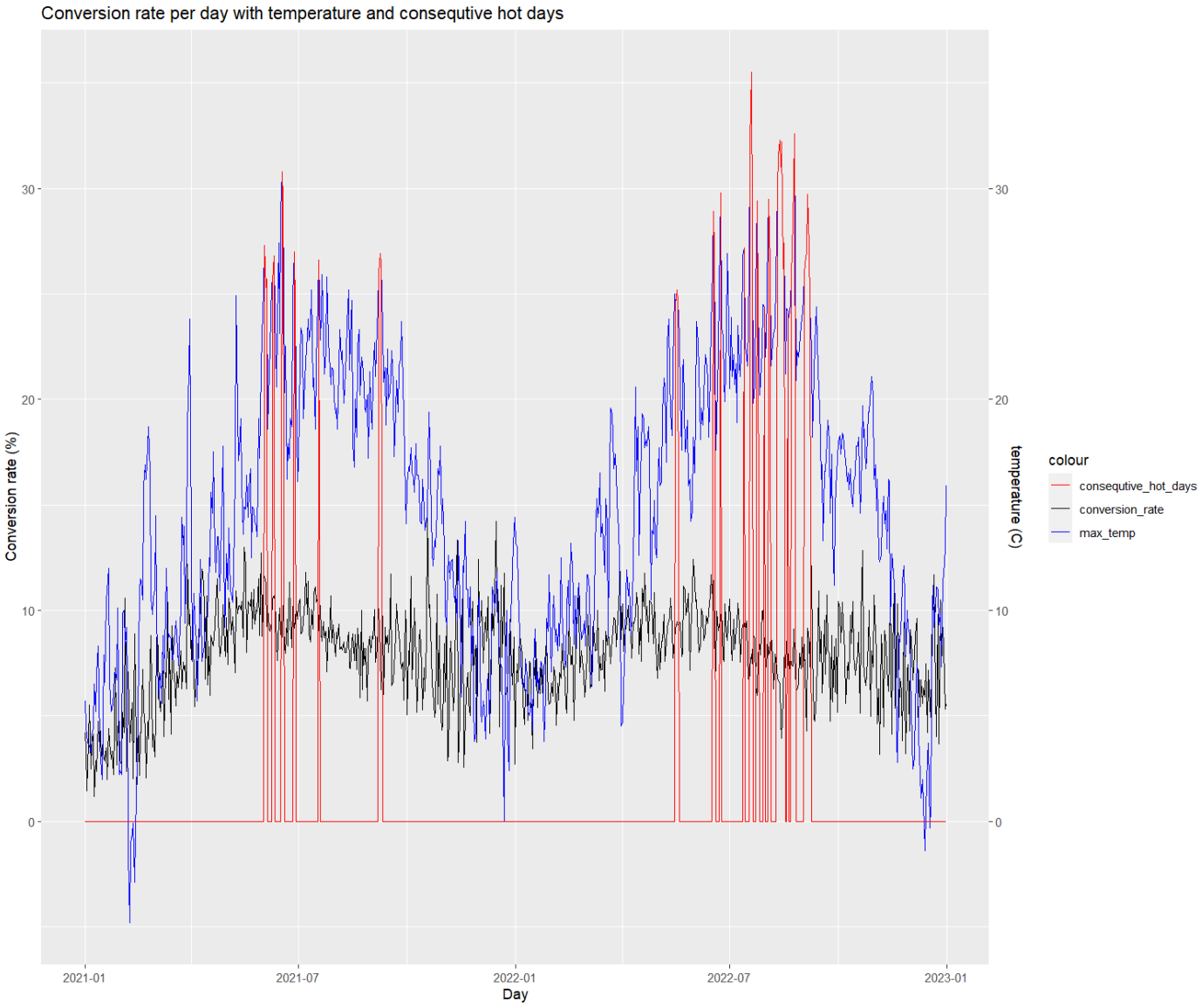


Figure 6.2: Conversions per day with temperature and consecutive hot days

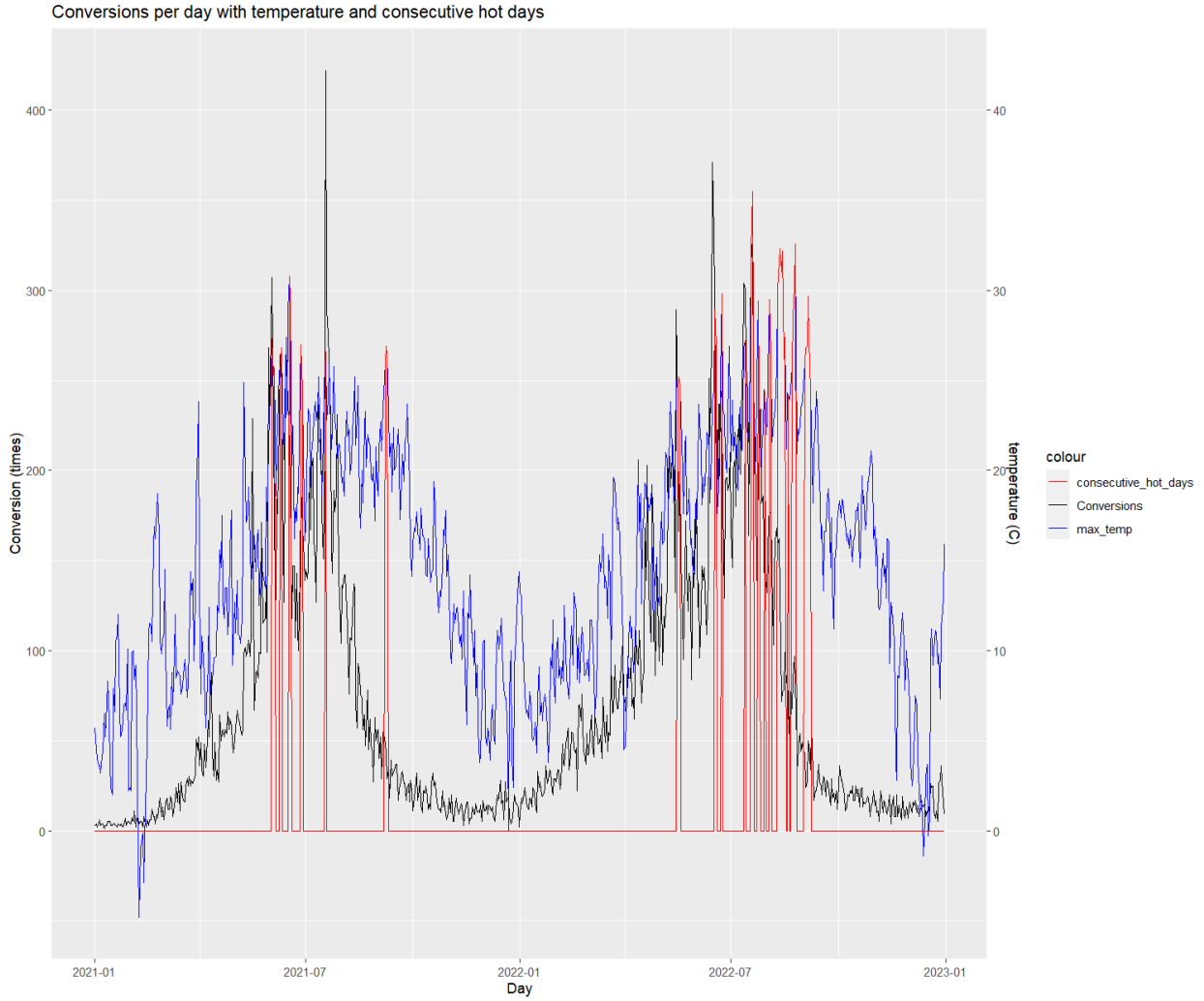


Table 6.2: Conversions, consecutive hot days t-test

```
> t.test(Conversions ~ consecutive_hot_days, data=Day_Weather_Merged_Table)

        Welch Two Sample t-test

data:  Conversions by consecutive_hot_days
t = -6.3882, df = 48.167, p-value = 6.297e-08
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -123.00345  -64.11453
sample estimates:
mean in group 0 mean in group 1
       65.54971       159.10870
```

Table 6.3: Linear regression model on consecutive hot days

```
> H6 <- lm(Conversions ~ consecutive_hot_days, data = Day_Weather_Merged_Table)
> summary(H6) # significant

Call:
lm(formula = Conversions ~ consecutive_hot_days, data = Day_Weather_Merged_Table)

Residuals:
    Min      1Q  Median      3Q     Max
-135.11  -51.55  -31.55   38.20  305.45

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          65.550      2.756  23.781   <2e-16 ***
consecutive_hot_days 93.559     10.980   8.521   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 72.09 on 728 degrees of freedom
Multiple R-squared:  0.09068,   Adjusted R-squared:  0.08943
F-statistic:  72.6 on 1 and 728 DF,  p-value: < 2.2e-16
```

Table 7.1: Linear regression model max temperature

```
> H7 <- lm(conversion_rate ~ max_temp, data = Day_Weather_Merged_Table)
> summary(H7)

Call:
lm(formula = conversion_rate ~ max_temp, data = Day_Weather_Merged_Table)

Residuals:
      Min        1Q    Median        3Q       Max
-0.059346 -0.011830  0.001037  0.014363  0.067824

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.212e-02  1.819e-03   34.16   <2e-16 ***
max_temp    1.108e-04  1.087e-05   10.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02075 on 728 degrees of freedom
Multiple R-squared:  0.125,     Adjusted R-squared:  0.1238
F-statistic:   104 on 1 and 728 DF,  p-value: < 2.2e-16
```

Table 7.2: Linear regression model of max temperature on consecutive hot days

```
> H7_hot <- lm(conversion_rate ~ max_temp*consecutive_hot_days, data = Day_Weather_Merged_Table)
> summary(H7_hot)

Call:
lm(formula = conversion_rate ~ max_temp * consecutive_hot_days,
    data = Day_Weather_Merged_Table)

Residuals:
      Min        1Q    Median        3Q       Max
-0.060601 -0.012100  0.000202  0.014099  0.068000

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 5.891e-02  1.904e-03  30.936  < 2e-16 ***
max_temp                    1.384e-04  1.213e-05  11.405  < 2e-16 ***
consecutive_hot_days        7.646e-02  3.146e-02   2.430  0.01533 *
max_temp:consecutive_hot_days -3.306e-04  1.132e-04  -2.922  0.00359 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02041 on 726 degrees of freedom
Multiple R-squared:  0.1559,    Adjusted R-squared:  0.1524
F-statistic: 44.69 on 3 and 726 DF,  p-value: < 2.2e-16
```

Figure 7.1: Granger Max temp on Conversion rate p-value plot

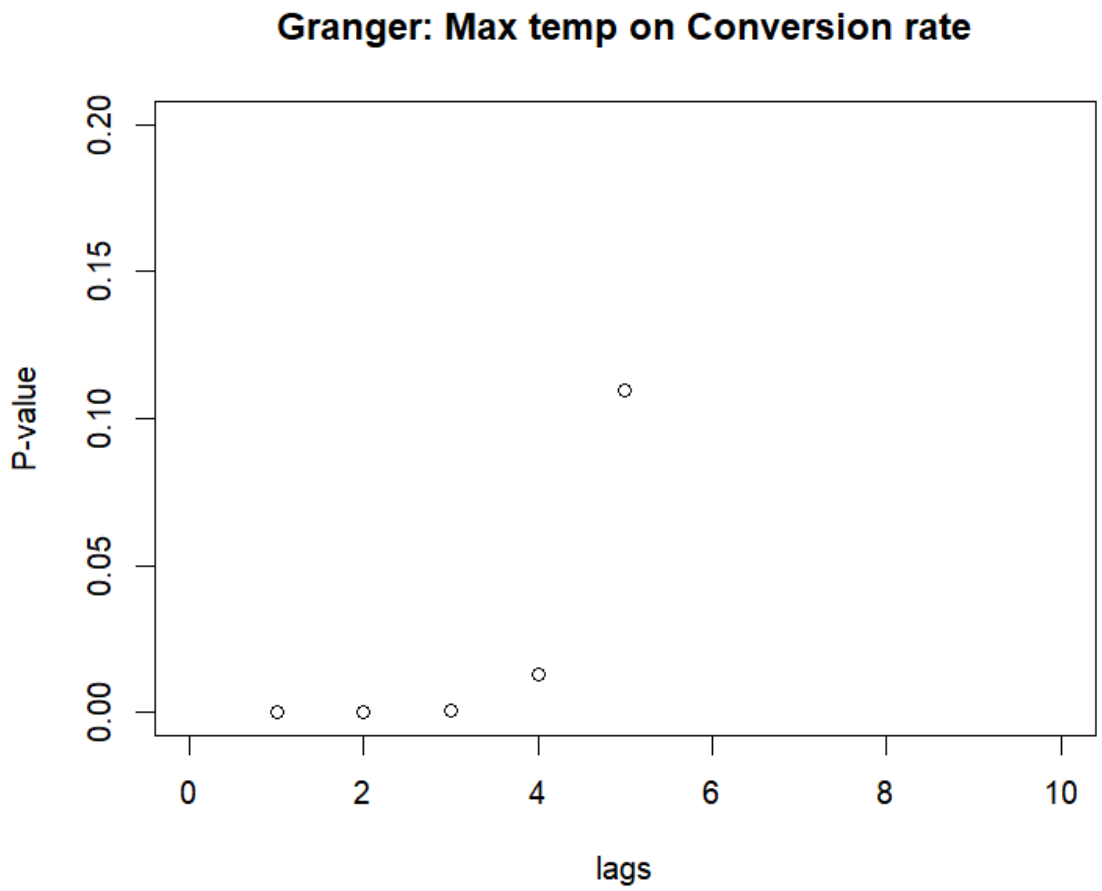


Table 8.1: Linear regression model of sunshine duration

```
> H8_rate <- lm(conversion_rate ~ sunshine_duration, data = Day_Weather_Merged_Table)
> summary(H8_rate)

Call:
lm(formula = conversion_rate ~ sunshine_duration, data = Day_Weather_Merged_Table)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06731 -0.01227  0.00142  0.01481  0.06975

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.238e-02  1.265e-03  57.220  < 2e-16 ***
sunshine_duration 1.215e-04  1.823e-05   6.666  5.2e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02153 on 728 degrees of freedom
Multiple R-squared:  0.05752,   Adjusted R-squared:  0.05623
F-statistic: 44.43 on 1 and 728 DF,  p-value: 5.197e-11
```

Figure 8.1: Conversion rate per day with Sunshine duration

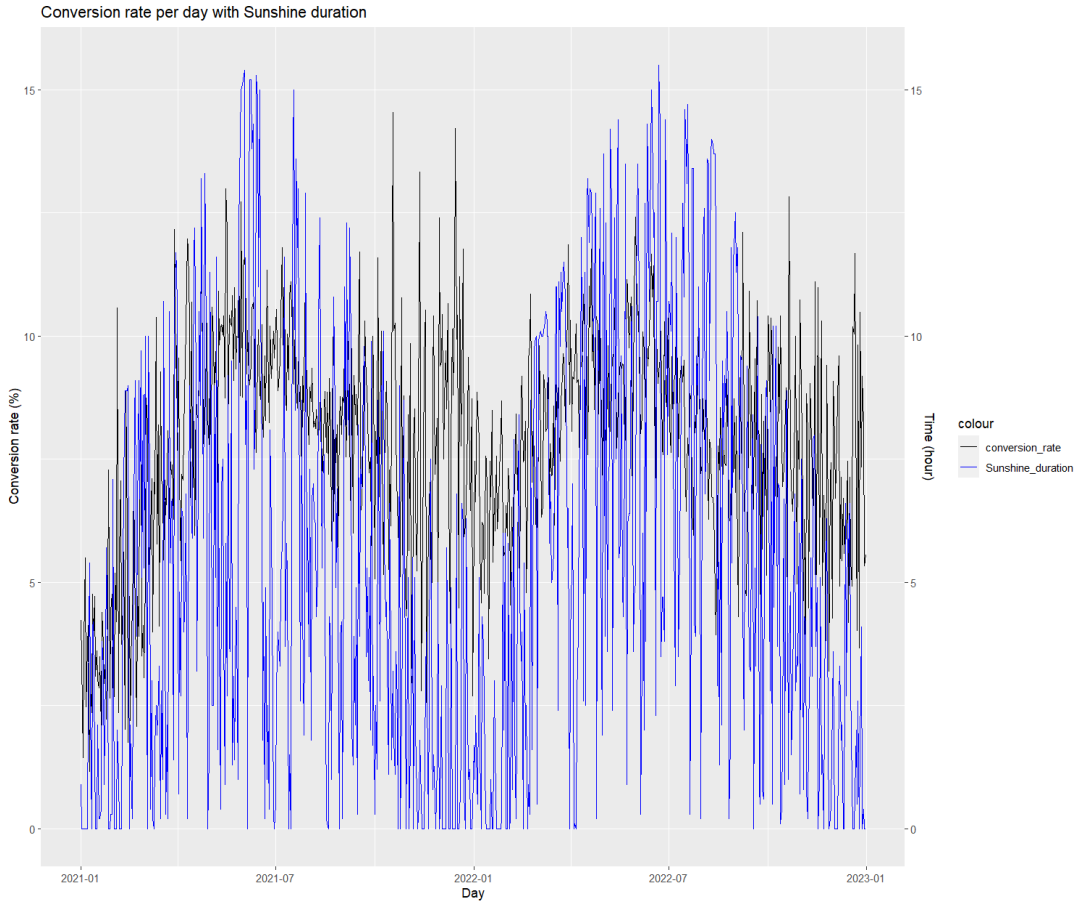


Table 8.2: Linear regression model of sunshine duration, max temperature on hot consecutive days

```
> Hweather <- lm(conversion_rate ~ sunshine_duration + max_temp*consecutive_
> summary(Hweather)

Call:
lm(formula = conversion_rate ~ sunshine_duration + max_temp * 
    consecutive_hot_days, data = Day_Weather_Merged_Table)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.059877 -0.011572  0.000087  0.013613  0.069616 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  5.875e-02  1.902e-03  30.879  < 2e-16 ***
sunshine_duration            4.045e-05  2.064e-05   1.960  0.05036 .  
max_temp                     1.252e-04  1.385e-05   9.038  < 2e-16 ***
consecutive_hot_days         8.021e-02  3.146e-02   2.549  0.01099 *  
max_temp:consecutive_hot_days -3.449e-04  1.132e-04  -3.048  0.00239 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02037 on 725 degrees of freedom
Multiple R-squared:  0.1603,    Adjusted R-squared:  0.1557 
F-statistic: 34.61 on 4 and 725 DF,  p-value: < 2.2e-16
```

Table 8.3: Linear regression model of standardized sunshine duration, standardized max temperature on hot consecutive days

```
> Hweather_std <- lm(conversion_rate ~ sunshine_duration_std + max_temp_std*co
> summary(Hweather_std)

Call:
lm(formula = conversion_rate ~ sunshine_duration_std + max_temp_std * 
    consecutive_hot_days, data = Day_Weather_Merged_Table)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.059877 -0.011572  0.000087  0.013613  0.069616 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     0.0799132  0.0007855 101.739  < 2e-16 ***
sunshine_duration_std           0.0017698  0.0009029   1.960  0.05036 .  
max_temp_std                    0.0088496  0.0009791   9.038  < 2e-16 ***
consecutive_hot_days            0.0278919  0.0145273   1.920  0.05525 .  
max_temp_std:consecutive_hot_days -0.0243845  0.0080012  -3.048  0.00239 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02037 on 725 degrees of freedom
Multiple R-squared:  0.1603,    Adjusted R-squared:  0.1557 
F-statistic: 34.61 on 4 and 725 DF,  p-value: < 2.2e-16
```

Figure 9.1: Average Conversion Rates per DayOfTheWeek

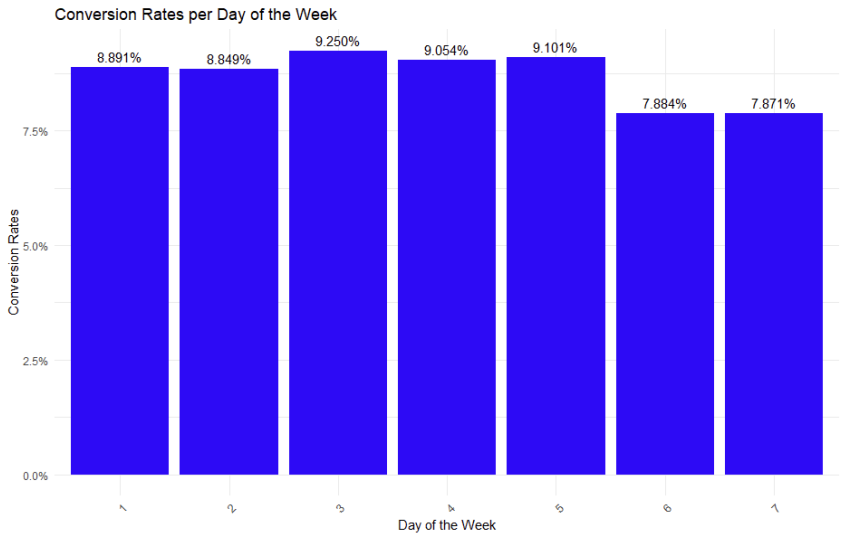


Figure 9.2: Results of the linear model for Conversion per DayOfTheWeek

```
Call:
lm(formula = Conversion ~ DayOfTheWeekNumber, data = Conversion_SDW)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09250 -0.09054 -0.08891 -0.07884  0.92129

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.0889130  0.0009328  95.317  < 2e-16 ***
DayOfTheWeekNumber2 -0.0004240  0.0013271  -0.319  0.74936
DayOfTheWeekNumber3  0.0035839  0.0013206   2.714  0.00665 **
DayOfTheWeekNumber4  0.0016233  0.0013528   1.200  0.23015
DayOfTheWeekNumber5  0.0020942  0.0014028   1.493  0.13549
DayOfTheWeekNumber6 -0.0100694  0.0014226  -7.078 1.46e-12 ***
DayOfTheWeekNumber7 -0.0102076  0.0012742  -8.011 1.14e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2817 on 600008 degrees of freedom
Multiple R-squared:  0.000363,  Adjusted R-squared:  0.000353
F-statistic: 36.31 on 6 and 600008 DF,  p-value: < 2.2e-16
```

Figure 10.1: Chisq results for significance of device on conversion

```
> conversion_device <- table(actual_session_table$device_category_desc, actual_session_table$conversion)
> chi_test_device <- chisq.test(conversion_device)
> print(chi_test_device)

        Pearson's Chi-squared test

data:  conversion_device
X-squared = 2078.6, df = 2, p-value < 2.2e-16
```

Figure 10.2: Conversion rates based on type of device

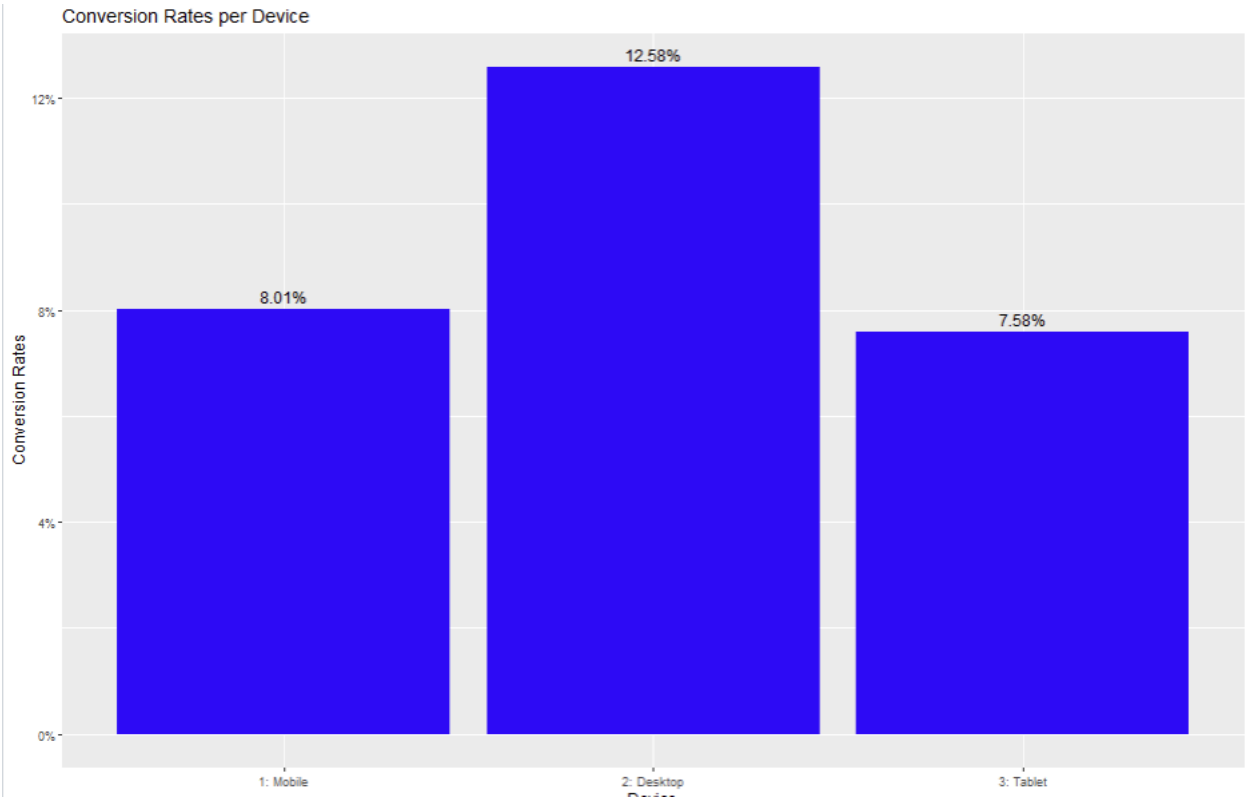


Figure 10.3: Total of Sessions based on type of the device

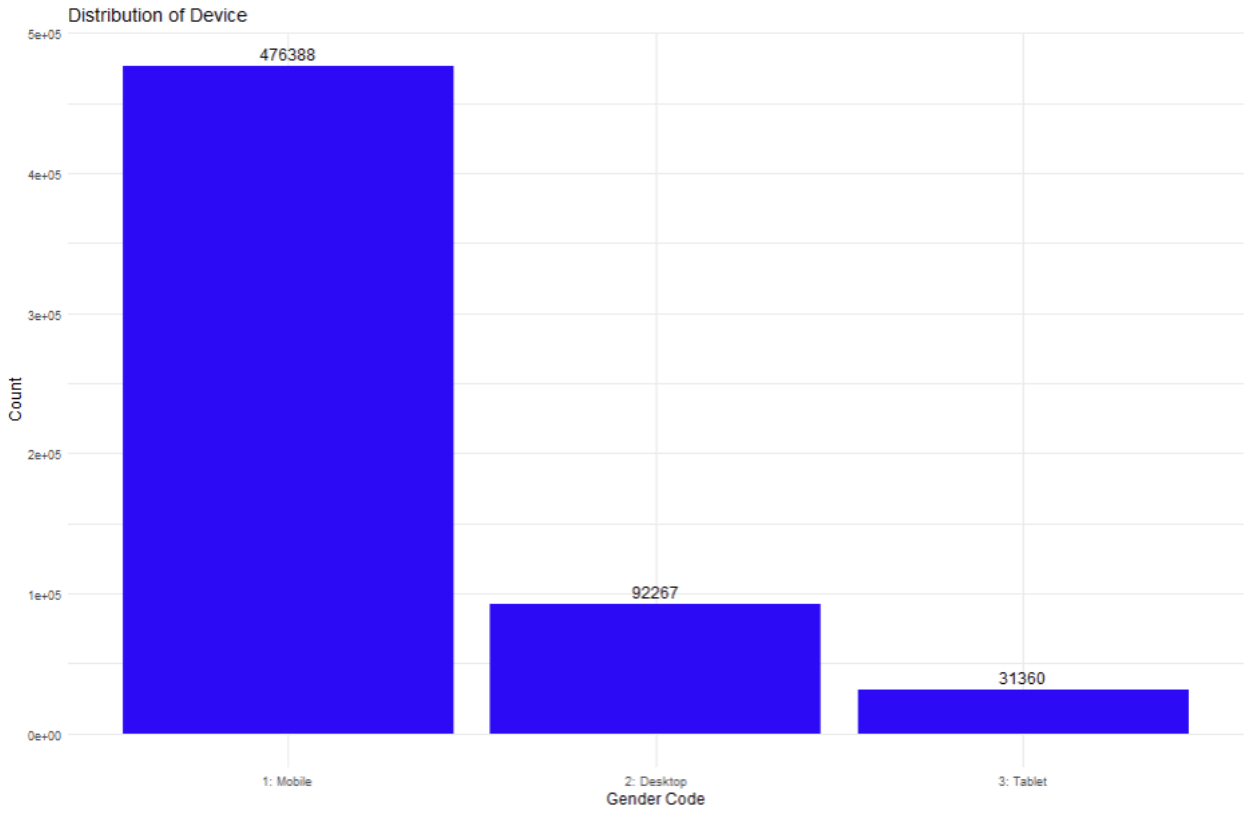


Figure 11.1: Pearson Correlation test result for google search on bikini and zwembroek with Conversion

```
> print(cor_test_result_bikini)

        Pearson's product-moment correlation

data:  actual_session_table$bikini and actual_session_table$conversion
t = 14.742, df = 600013, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01649862 0.02155733
sample estimates:
      cor
0.0190281

> print(cor_test_result_zwembroek)

        Pearson's product-moment correlation

data:  actual_session_table$zwembroek and actual_session_table$conversion
t = 11.568, df = 600013, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01240290 0.01746231
sample estimates:
      cor
0.0149327
```

Figure 11.1: Pearson Correlation test result for google search on [CENSORED], Zalando and C&A with Conversion

```
> print(cor_test_result_wehkamp)

        Pearson's product-moment correlation

data:  actual_session_table$Wehkamp and actual_session_table$conversion
t = 2.9278, df = 600013, p-value = 0.003413
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.001249499 0.006309971
sample estimates:
        cor
0.003779759

> print(cor_test_result_zalando)

        Pearson's product-moment correlation

data:  actual_session_table$Zalando and actual_session_table$conversion
t = 6.0473, df = 600013, p-value = 1.474e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.005276503 0.010336738
sample estimates:
        cor
0.007806671


> print(cor_test_result_ca)

        Pearson's product-moment correlation

data:  actual_session_table$CA and actual_session_table$conversion
t = 4.9108, df = 600013, p-value = 9.073e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.003809393 0.008869734
sample estimates:
        cor
0.006339604
```

2. Bibliography

de Haan, E., Kannan, P. K., Verhoef, P. C., & Wiesel, T. (2018). Device switching in online purchasing: Examining the strategic contingencies. Journal of Marketing, 82(5), 1–19. https://doi.org/10.1509/jm.17.0113

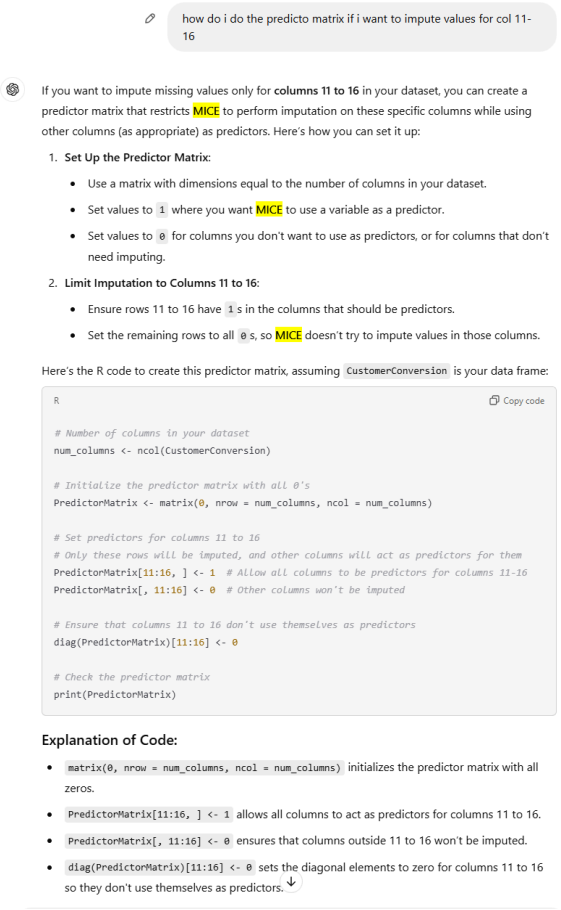
Gong, K., Peng, Y., Wang, Y., & Xu, M. (2018). Time series analysis for C2C conversion rate. *Electronic Commerce Research, 18*, 763–789.

Knmi. KNMI. (n.d.). https://www.knmi.nl/over-het-knmi/about

3. Use of GenAI

For this assignment, we use ChatGPT to get a better understanding on both theoretical and practical subjects that were given during the lecture and tutorial sessions. We used ChatGPT incrementally, doing small prompts here and there to help explain theory, help fine tune our code, and to analyze and confirm whether our approaches were correct. However, all of the ideas presented in this research paper, such as our approaches to methodology, as well as our interpretations, are our own.

Example Prompt 1: Understanding MICE test



how do i do the predicto matrix if i want to impute values for col 11-16

If you want to impute missing values only for **columns 11 to 16** in your dataset, you can create a predictor matrix that restricts MICE to perform imputation on these specific columns while using other columns (as appropriate) as predictors. Here's how you can set it up:

1. **Set Up the Predictor Matrix**:
   - Use a matrix with dimensions equal to the number of columns in your dataset.
   - Set values to `1` where you want MICE to use a variable as a predictor.
   - Set values to `0` for columns you don't want to use as predictors, or for columns that don't need imputing.
2. **Limit Imputation to Columns 11 to 16**:
   - Ensure rows 11 to 16 have `1`s in the columns that should be predictors.
   - Set the remaining rows to all `0`s, so MICE doesn't try to impute values in those columns.

Here's the R code to create this predictor matrix, assuming `CustomerConversion` is your data frame:

```r
# Number of columns in your dataset
num_columns <- ncol(CustomerConversion)

# Initialize the predictor matrix with all 0's
PredictorMatrix <- matrix(0, nrow = num_columns, ncol = num_columns)

# Set predictors for columns 11 to 16
# Only these rows will be imputed, and other columns will act as predictors for them
PredictorMatrix[11:16, ] <- 1  # Allow all columns to be predictors for columns 11-16
PredictorMatrix[, 11:16] <- 0  # Other columns won't be imputed

# Ensure that columns 11 to 16 don't use themselves as predictors
diag(PredictorMatrix)[11:16] <- 0

# Check the predictor matrix
print(PredictorMatrix)
```

Explanation of Code:

- `matrix(0, nrow = num_columns, ncol = num_columns)` initializes the predictor matrix with all zeros.
- `PredictorMatrix[11:16, ] <- 1` allows all columns to act as predictors for columns 11 to 16.
- `PredictorMatrix[, 11:16] <- 0` ensures that columns outside 11 to 16 won't be imputed.
- `diag(PredictorMatrix)[11:16] <- 0` sets the diagonal elements to zero for columns 11 to 16 so they don't use themselves as predictors.

Example Prompt 2: Get better understanding of missing values type
Prompt: "what is the definition and implication of MAR, MCAR and MNAR?"

In data analysis, missing data mechanisms are often classified into three categories: Missing at Random (MAR), Missing Completely at Random (MCAR), and Missing Not at Random (MNAR). Each describes a different pattern of missingness and has implications for how the missing data should be handled.

MCAR (Missing Completely at Random):
Definition: Data is missing completely at random if the probability of a value being missing is unrelated to any observed or unobserved variables in the dataset.
Implication: The missingness is purely random, with no systematic pattern. Therefore, analyses can be conducted without much bias if data is MCAR, and methods like listwise deletion or pairwise deletion can be valid.

MAR (Missing at Random):
Definition: Data is missing at random if the probability of a value being missing is related to some observed variables in the dataset but not to the unobserved data itself.
Implication: With MAR, the missing data pattern can be accounted for by adjusting for observed variables. For instance, if data on income is missing more often for younger people, age can be used to predict income. Techniques like multiple imputation or maximum likelihood estimation are commonly used to handle MAR data.

MNAR (Missing Not at Random):
Definition: Data is missing not at random if the probability of missing data is related to the value of the missing data itself or other unobserved factors.
Implication: MNAR is the most challenging case because the missingness depends on unobserved information, making it hard to accurately estimate the missing values. Dealing with MNAR often requires specialized modeling approaches or sensitivity analyses to account for potential biases.
Understanding these distinctions is key to selecting appropriate methods for handling missing data in analyses.

4. Group Reflection

In our team, we effectively achieved our purpose by supporting each other and maintaining constant communication via WhatsApp. Each member brought unique insights, which helped us learn from one another. This mutual learning enriched our understanding and kept us aligned with our goals.

Our agreed way of working, especially setting regular check-ins and using WhatsApp for continuous updates, was efficient and effective. It enabled quick responses to questions and kept the whole team engaged without needing formal meetings. This structure was crucial for maintaining momentum and addressing issues as they arose.

Our group uses university google drive for sheets, slides, docs type of documents; and github for code-related documents as our file sharing platform. By using these sharing platforms, we can seamlessly check our peers' work in real time. Which leads to a better team collaboration and performance.

Our biggest challenge was with aggregating the dataset, due to some issues in the dataset. We had many challenging discussions within our team, debating what the best course of action was, but every conversation was proactive and productive. We complemented each other very well in aspects of technical skills, analytical skills and abstract thinking. This created a learning environment within the group where everybody was teaching new skills to one another. We would say that we even had a fun time working on this assignment!