LEVERAGING CUSTOMER TRANSACTION AND SALES DATA FOR FUTURE MARKETING ACTIONS

AN ANALYSIS REPORT
APRIL 2023

Team Yellow
Muhurthana Subramaniam
Gurleen Kaur
Daksh Sethi
Yu Chen
Dharampreet Singh

Table of Contents

Executive Summary

This report analyzes customer transaction and sales data for an independent tire and automotive retail outlet in Kelowna. This analysis aims to identify trends and understand the features of the source data to provide insights that can aid the company in making data-driven decisions to develop effective marketing strategies.

The analysis uncovered several key insights about the company's customers, including their birth years, education levels, family size, and income levels. For example, it was revealed age of customers ranged from 20 – 70 years, and these customers belonged to distinct generational groups, namely Baby boomers, Gen X, Gen Y, and Gen Z. In terms of income, most households fell into income ranging from CAD 32,000 to CAD 70,000.

In addition, it was discovered that the Baby Boomer generation had the highest number of complaints, accounting for 50.5% of the total. The key takeaways obtained from the data were that nearly half of the customers belong to the city of Kelowna at 49.3%, winter tires were purchased the most during the span of 8 years, customers preferred in-store purchases more than online for tires and finally Baby Boomers and Gen X tends to purchase tires instore whereas Gen Y and Z prefer online purchases. All three codes performed well during the months of June and November. The most used discount code was the buy 1 set, get a second set at 15% off.

It would be advisable to maintain grip on sales for the main profitable region, which is Kelowna, and try to use blue ocean strategy for untapped markets. Furthermore, advertise the Specialized wheel rim quality in Vernon as average affordability is high based on their income level in comparison to other main neighborhoods. Increase social media promotional content that would direct users to the client’s website to increase sales. Reach Baby boomers and Gen X

through traditional media, and Gen Y and Gen Z through email promotions. Keep promoting Happy Dozen and Buy 1 set and get second set 15% off discount in the months of June and November. Furthermore, it is advisable for the client to survey Baby Boomers and Gen X to find the root cause of their complaints and address them accordingly.

In summary, the company can utilize these insights and the further analysis conducted to develop effective marketing strategies and make data-driven decisions to maximize potential opportunities and profit.

Introduction

The Tire Industry is a continuously growing market, and the Kelowna Tire Market has space for vast development though it undergoes fierce competition. A study mentioned that the Canadian Tire Market was worth USD 5.42 billion in 2021 and will reach around USD 7.68 billion by the end of 2028 (Tire and Rubber Association of Canada, 2021). Kelowna is one of the top five fastest growing (statcan, 2022) large urban in Canada; the population increased quickly, and the growth rate from 2016 to 2021 was 14%, and the rising trend will continue. Population growth leads to increasing demand for automotive parts and tires. Meanwhile, the competition is also severe.

Our client, an independent tire and automotive retail outlet based in Kelowna, has entrusted the in-house team with company data requiring support in analyzing customer transactions and sales data to evaluate future marketing actions to maximize potential opportunities. This report lays out the exploratory data analysis. An initial study will be conducted to discover patterns and understand the features via data visualization methods to present insights that will be utilized in further analysis and to make data driven decisions.

Opportunities for Growth

1. Determine the best methods to reach new and existing customers.
2. Develop future marketing initiatives to boost sales based on email promotions.
3. Target the customer based on popularity and preference of promotional campaigns.
4. Target profitable regions and untapped markets based on the purchasing behavior of customers.

Methodology

Initial analysis was performed utilizing the data set provided via using techniques such as Exploratory Data Analysis, Data Cleaning, and Imputation Techniques (See Glossary and Appendix) to obtain better understanding of the data. In addition, Secondary Data Research of the tire industry, Canadian Economy, and competitors was performed to support the findings obtained through the exploration of the data.

Findings and Insights

Exploratory Data Analysis

Customer Base

The following figures and tables below will provide prominent data analysis that will uncover insights from the source data.

Figure 01 represents a map of Okanagan with the number of customers based on their respective regions. The size of the circle depicts the aggregation of customers belonging to each region. The bigger the circle is, the higher the customer base will be. The most extensive customer base is Kelowna, with 831 customers, followed by Vernon. Penticton and West Kelowna are almost tied at a count with 207 and 205 customers, respectively. Peachland has the lowest number of customers, which amounts to 207. Based on the city profile of Kelowna, this region has the largest community in the regional district, with a population of 143,000. This helps justify why Kelowna has the most customers compared to other regions surrounding it (City of Kelowna, 2023).

Figure 01 – Regional customer base of Okanagan depending on tire sales.

Generational Cohorts

Interesting insights were acquired through Generational Cohorts, which were created based on the birth year of the customer. The customers were segmented into four categories such as Baby Boomer (Born between 1944 - 1964), Gen X (Born between 1965 - 1979), Gen Y or Millennial (Born between 1980 - 1994), and Gen Z (Born between 1995 to 2015). Figure 02 indicates that Gen Y owns the most significant number of vehicles surpassing 600, followed by Gen X and Gen Z, placing Baby Boomers in the last place.

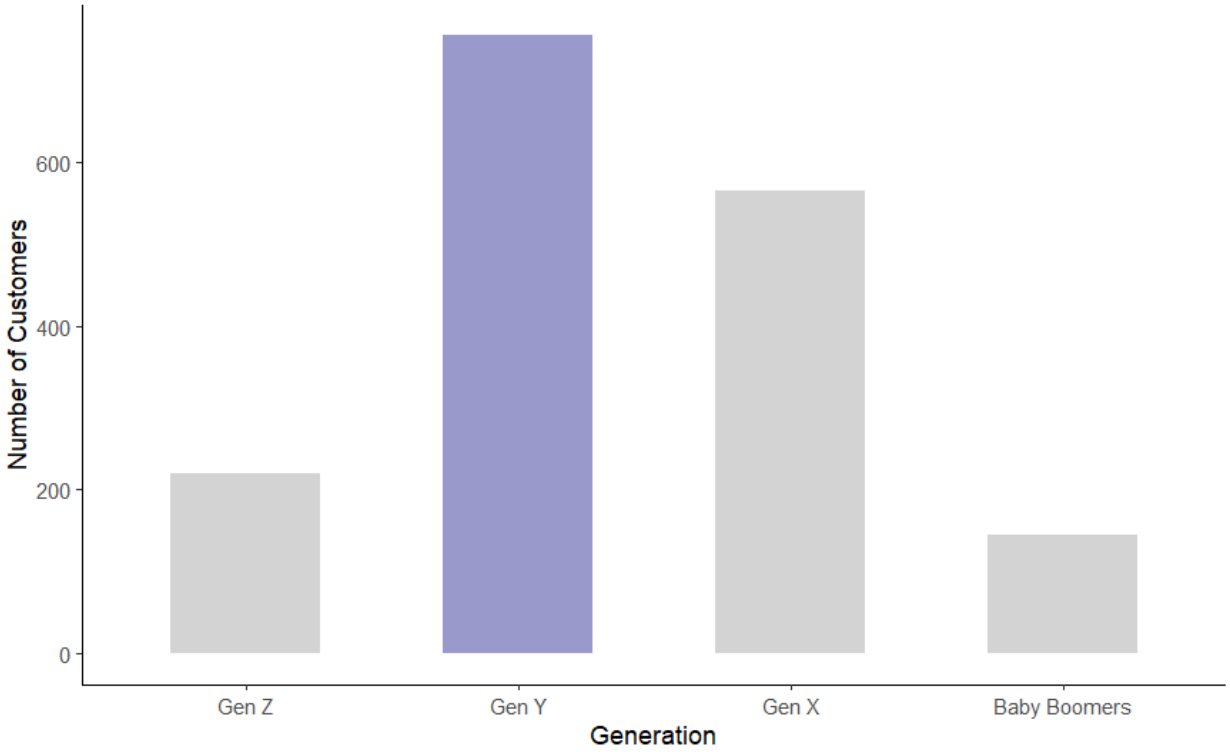


Figure 02 – Number of vehicles owned based on Generation types.

Income levels

The average customer income ranges between CAD 32,000 and CAD 70,000, which accounts for 39.3%. Only 79 customers fall into the category of 200,000 or above. The lowest income bracket consists of 269 customers. These findings can be validated via Figure 03, as shown below. These have been classified based on Secondary research conducted on Income levels (Revolution, 2022)

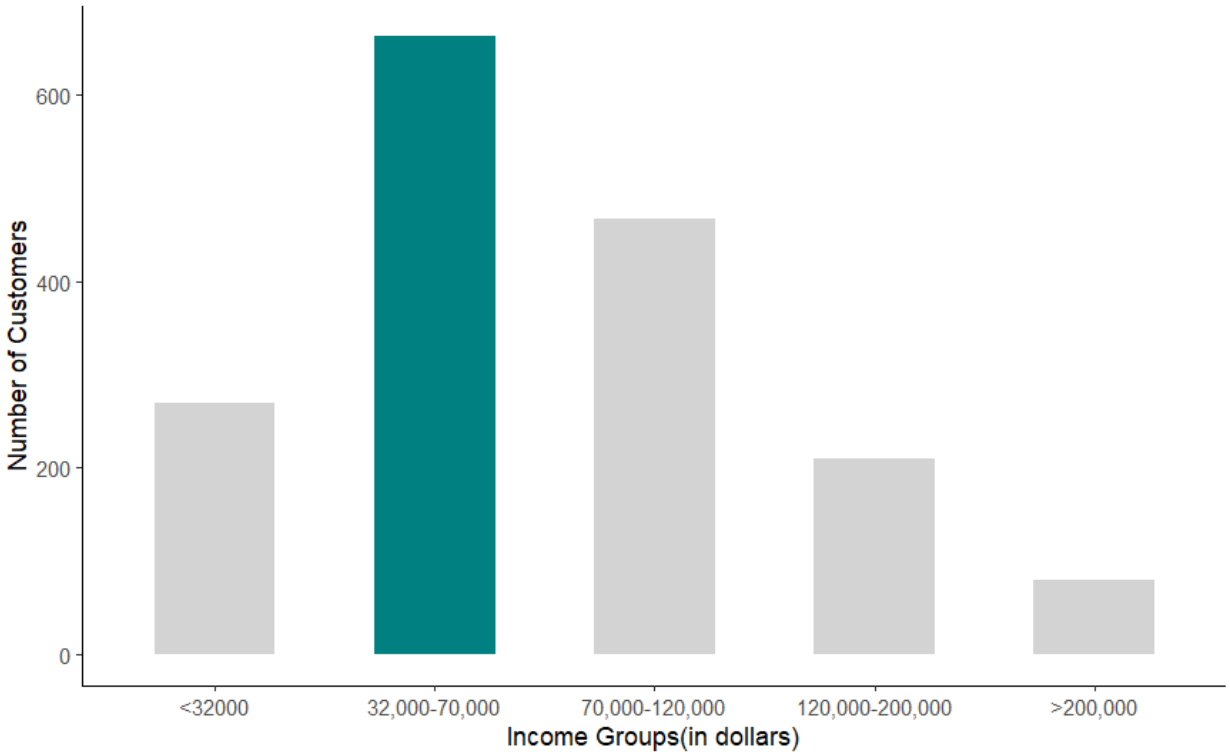


Figure 03 – Categorization of customers based on Income levels.

Tire Sales based on Seasons.

The following figure shows the number of tire sets purchased within a span of 8 years for the years 2014 to 2022 (8 months) with a projection for the years 2022, and 2023. The types of tires analyzed in this figure are all season, winter, and summer. The following insights were discovered during the analysis, and according to the data provided, Summer, winter, and All-season started in August for the years 2014, 2017, and 2018 respectively. The tire sets purchased for the summer season have been constant through the years with few fluctuations and the highest sale was observed in 2017.

In terms of winter tires, the most significant number of sales was generated in the year 2020. However, all-season tire purchases were constant for 3 years (2019 to 2021). The projected tire sales for the years 2022 and 2023 states that, winter tires will provide more revenue, followed by all-season and summer tires. This shows that consumers understand the harsh winter season and the importance of using appropriate tires for their safety and vehicle. This is supported by the Winter Tracks 2020 report, which was considered secondary data analysis; the Canadian consumer winter tire study shows that tire usage increased up to 72% in 2020 compared to 2017 (66%) and 2014 (58%).(Winter Tire Report, 2020).

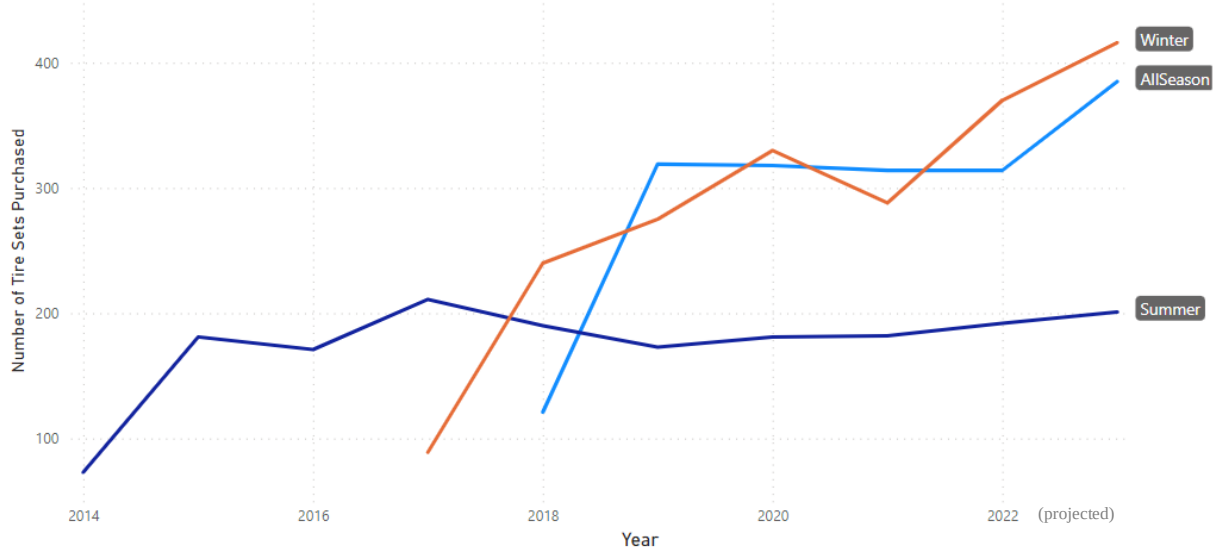


Figure 04 – Total sets of tires purchased in a span of 8 years

Sales Channels (Click and Mortar)

The figure below shows the online vs instore purchases and includes the projected sales for the years 2022 and 2023. On further analysis, it was observed that in-store purchases were almost twice the number of online purchases. There were few fluctuations throughout 8 years, and a significant drop was witnessed in 2019, which applied to both in-store and online purchases. In addition, it shows that the number of instore purchases projected for the years 2022 and 2023 surpasses online purchases like previous years.

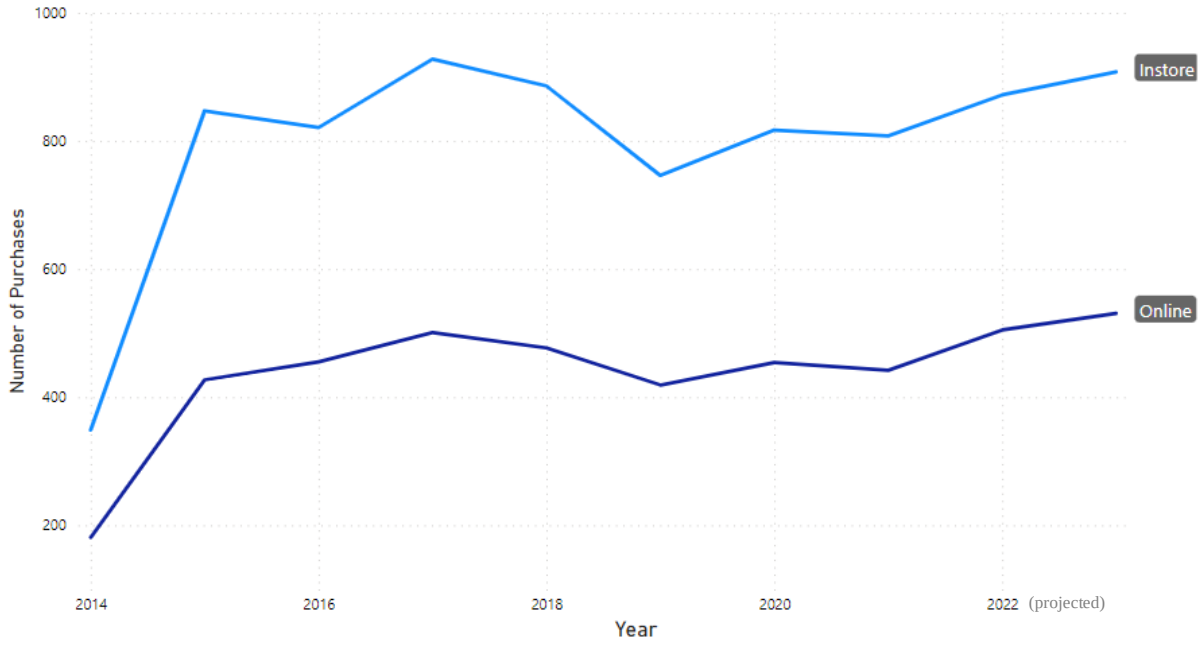


Figure 05 – Number of purchases made in-store vs. online over the years.

During a comparison of in-store vs. online purchases, it could be seen that there is a difference of 3107 customers who chose in-store purchases over online since the total number of purchases made in-store was 6733, whereas the online purchases amounted to 3626.

Generational Cohorts and Source of Discount

The below figure portrays the Generation types and the source used to obtain the discount on their purchases. Gen Y and Gen Z mostly prefer online. These two generations have embraced technology and are more tech-savvy (Digital Generations, 2023). On the other hand, Baby boomers and Gen X preferred in-store more because they only interact with technology to keep in touch with peers and family. Furthermore, they favor direct communication as much as possible regarding face-to-face conversations, leading to in-store being chosen by those two generations. (Four Generations, 2023)

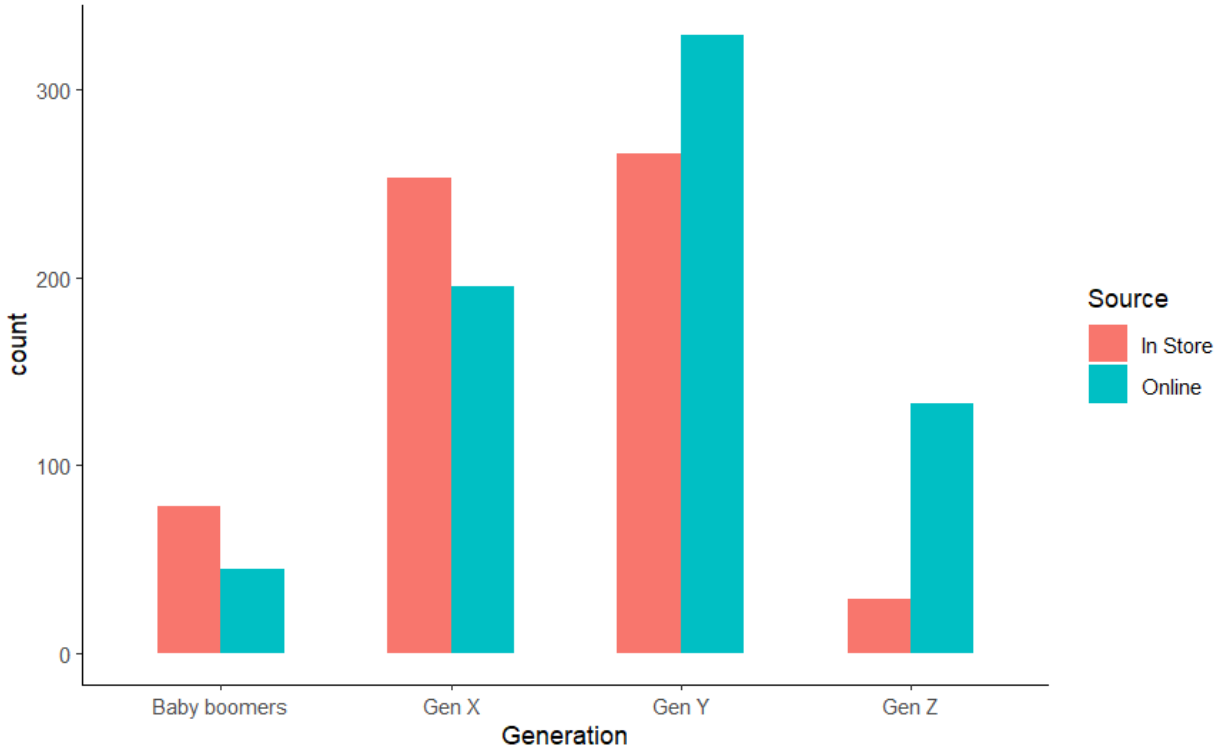


Figure 06 - Sources of discount based on generation types.

Rim Quality Preference

The figure below illustrates the greatest quality of rim that consumers preferred and purchased in the previous five years. Wheel rims are graded as follows based on secondary research: Specialized, Aluminum, and Steel. Yet, according to the data studied, aluminium (30.49%) is the most popular type of metal acquired between 2018 and 2022. Specialized and Steel came in second and third, with 28.35% and 23.84% respectively.

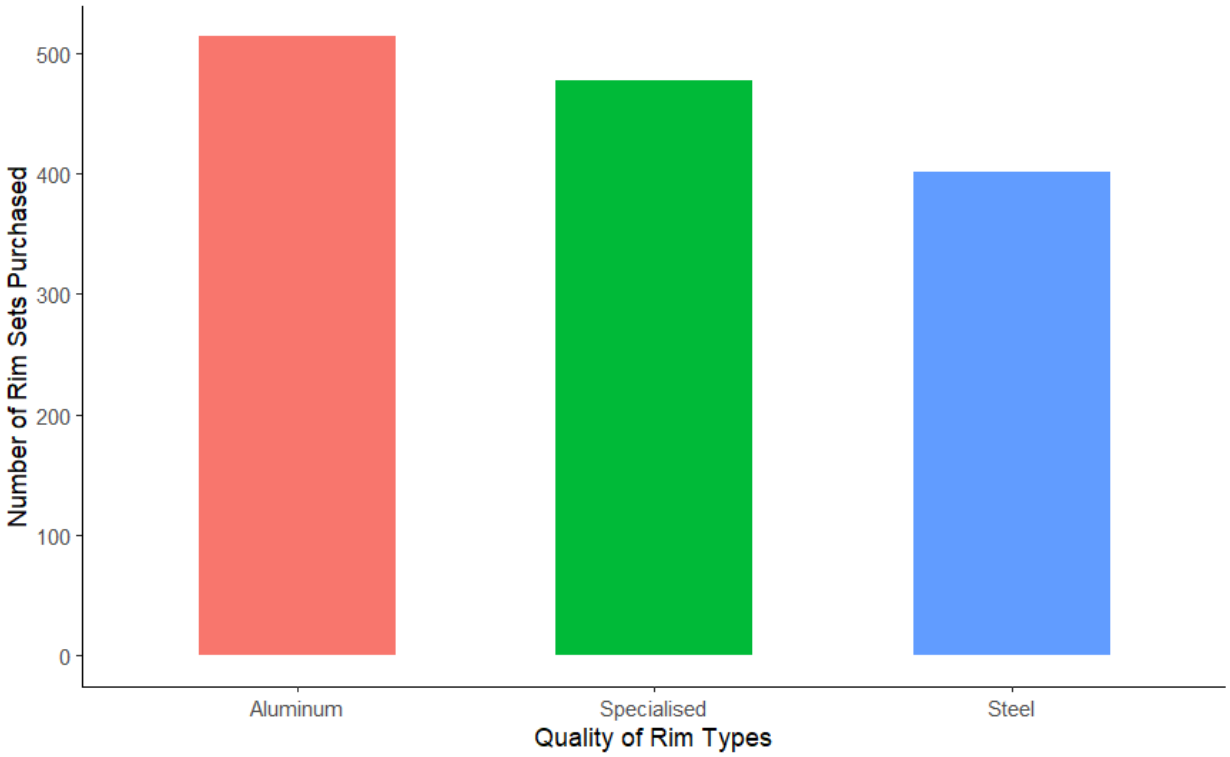


Figure 07 – The highest quality rim purchased in the last five years

Generational Cohorts and Complaints

The figure below represents a stacked bar chart regarding the complaints made by different generation cohorts regarding marketing and special offers. Majority of the complaints were lodged by baby boomers followed by Gen X. Furthermore, the least number of complaints were made by Gen X and Z.

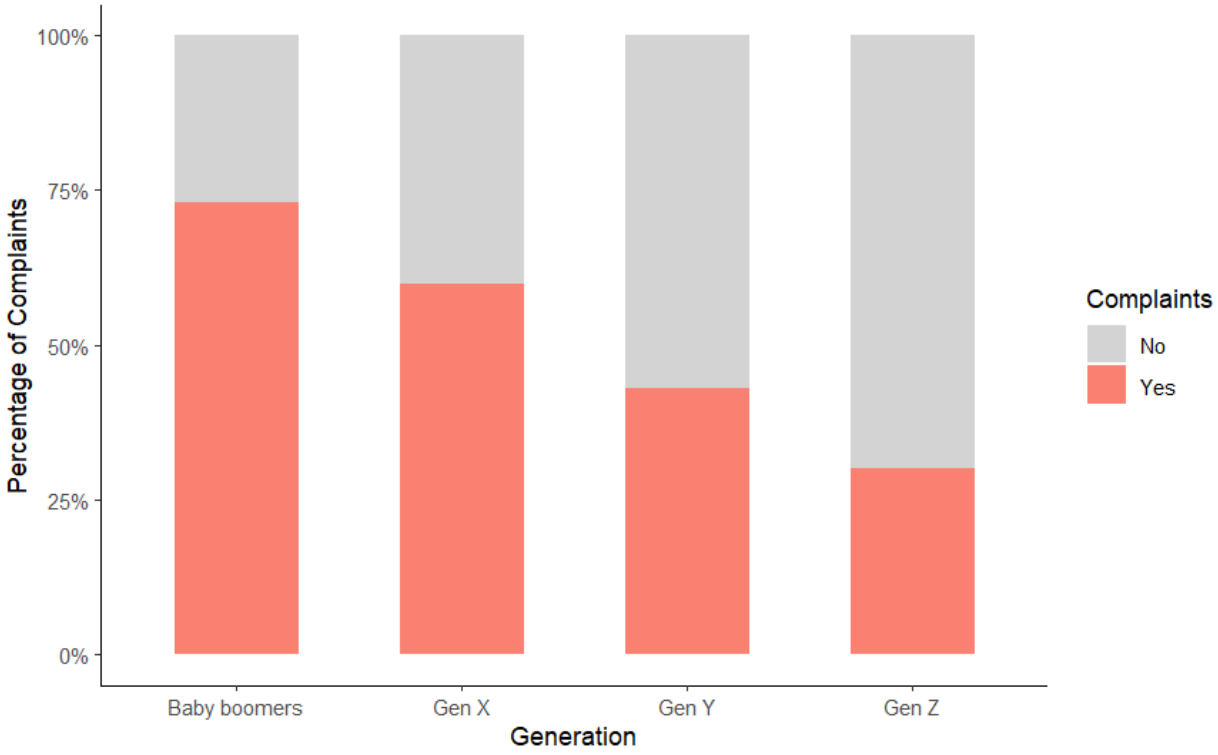


Figure 08 – Generation groups and complaints

Customer Segmentation

Customer segmentation aids in pinpointing the characteristics and behaviours of the client that are more significant, as well as aids the company in reaching underserved customers. It would also assist in deciding which client segments we ought to focus our marketing efforts on.

Based on Customer profiling, four, customer profiles were discovered and have been laid out below in terms of user personas.

Category	**Persona Robin**	**Persona Spencer**	**Persona Jasper**	**Persona Bliss**
Year	1986	1982	1984	1982
Neighborhood	Penticton	Kelowna	West Kelowna	Vernon
Driver	1	1	1	1
Primary vehicle	SUV	SUV	Van	Van
Approx of km driven	29,903	30,658	30,183	32,889
Income range	32,000-52,000	32,000-52,000	32,000-52,000	52,000 – 70,000
Rim quality	Aluminum	Aluminum	Specialized	Aluminum
Discount source and code	Online Buy 1 set, get second 15% off	Online Happy Dozen	Online Buy 1 set, get second 15% off	Online Happy Dozen
Number of instore purchases	4	3	5	3
Number of online purchases	1	2	0	2

Table 01 – Customer Segmentation

Based on the above table, it could be witnessed that, there are 4 types of customers representing the neighborhoods Kelowna, West Kelowna, Penticton, and Vernon. The annual km per year driven by all 4 types of customers is at least 30,000. Majority of the customers income ranges between 32,000 to 70,000. Preference for purchasing tires was favored towards instore rather than online for all four types of customers. The most used discount codes were Buy 1 set, get second 15% off, and 12% Happy Dozen, and majority of these codes were obtained online. Furthermore, when studying the similarities between these cities based on customer segmentation, Customers in Penticton and Kelowna prefer SUV, whereas Van is preferred by West Kelowna and Vernon.

Analysis of Coupons and Promotions

According to Figure 09, Discount code 2 (Buy 1 set, get 15% off your next invoice with us), generated the most significant tire sales. The second most frequent code used was number 3 (Happy Dozen, 12% off entire purchase), which generated 1336 sales and placed discount code 1 (Buy one set, get a second set at 15% off), in the last place with a sale of 1258 tires. Overall, all three coupons have roughly equal popularity.

Most responses received for promotional emails were in June and November all discount codes and this has been highlighted in the figure. This could be due to switching to all-season tires for June, and winter tires for November which resulted in maximum utilization of discount codes and an increase in response to promotional campaign emails.

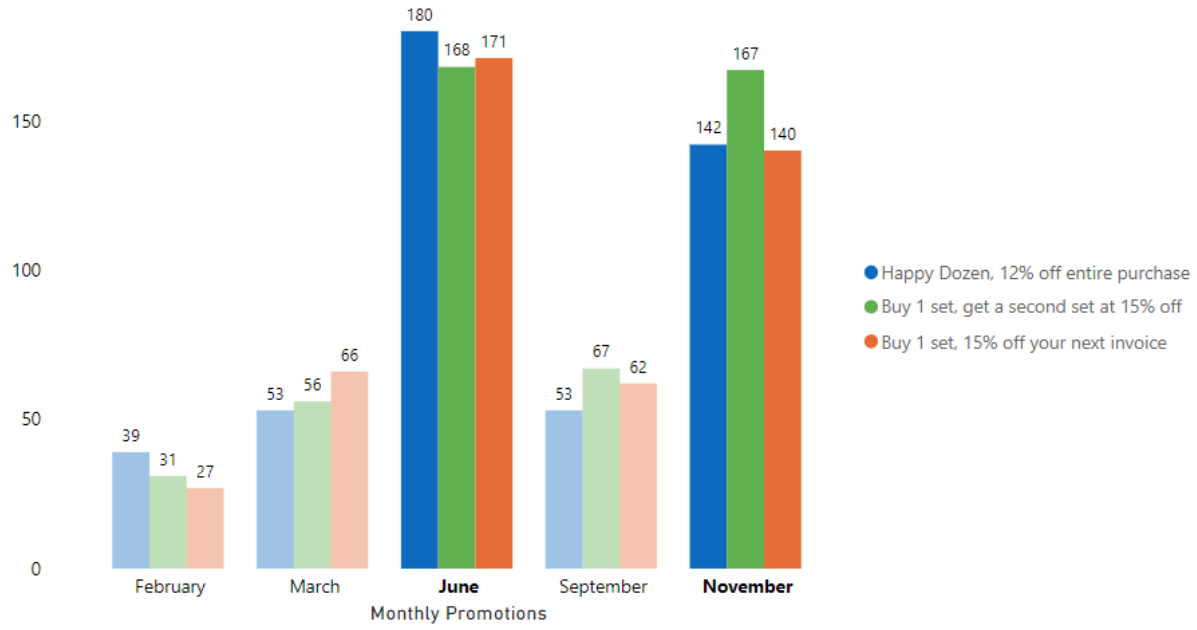


Figure 09 – Coupon alignment with specific promotions

Further Analysis

To evaluate the most profitable discount coupon and find the revenue generated via each discount coupon the in-house team has prepared a Revenue Calculator (See Appendix: External Links) for the client to be utilized. Based on the findings via calculator, it was discovered that, out of the three discount codes, buy one set, get a second set at 15% off (VK855356Y) earned the most income after analysing the data given by the revenue calculator and was the most popular among consumers. The data will be useful in influencing our future pricing and promotional activities. It might possibly enhance our income and consumer engagement by making this discount available more frequently or merging its benefits into other promotions.

The revenue calculator might be a great asset to the customer in terms of pricing strategies and revenue forecasts. The option to enter different discount codes and obtain an expected revenue calculation will allow you to make more educated pricing selections.

Secondary research was used to determine the cost and sales price of the tires, demonstrating a thorough effort to constructing this instrument.

Further testing and development of the revenue calculator can be done in the future to guarantee that it appropriately reflects the real-world revenue possibilities of various pricing methods. We also urge the team to think about adding new factors to the tool, such as changes in demand or market trends, to make it even more powerful.

Conclusions

- Nearly half of the customers belong to the city of Kelowna at 49.3%.
- Most households fell into income ranging from CAD 32,000 to CAD 70,000.
- 38% of customers had purchased an extra warranty.
- Baby boomers and Gen X are less tech savvy than Gen Y and Gen Z
- Gen Y owns the greatest number of vehicles.
- Winter tires were purchased the most during the span of 8 years.
- Customers preferred in-store purchases more than online for tires.
- Baby Boomers and Gen X tends to purchase tires instore whereas Gen Y and Z prefer online purchases.
- Most popular rim quality among customers is aluminum.
- Baby Boomers complained the most whereas Gen Z complained the least regarding marketing and special offers.
- All three discount codes seemed to perform well in the months of June and November.
- Winter tires generated the most revenue for the company.
- Buy 1 set, get a second set at 15% off was the most used discount code by the customers, based on the insights obtained via the revenue calculator.

Recommendations

1. Maintain grip on sales for the main profitable region, which is Kelowna, and try to reach untapped markets namely Vernon, Penticton, and West Kelowna.
2. Advertise the Specialized wheel rim quality in Vernon as average affordability is high based on their income level in comparison to other main neighborhoods.
3. Increase online sales through social media promotional content that would direct users to the client's website.
4. Reach Baby boomers through traditional media Gen Y and Gen Z through email promotions based on their preference towards technology.
5. To reduce the complaints for Baby boomers and Gen X, it is advisable to lower the number of promotional emails sent as they prefer in-store purchases more than online.
6. It is advisable to keep promoting Happy Dozen and Buy 1 set and get second set 15% off discount in the months of June and November.

Glossary

Term	Definition
Data cleaning	The process of correcting or eliminating data that is erroneous, corrupted, poorly formatted, duplicated, or incomplete within a dataset
Data Transformation	The process of modifying the format, structure, or values of data is known as data transformation.
Data Imputation	The process of substituting replacement values for missing data

References

Antonio (Anthony) de la Mora y Madrigal. (n.d.). *Boomers, gen X, gen y, and gen Z explained*. LinkedIn. Retrieved February 6, 2023, from https://www.linkedin.com/pulse/boomers-gen-x-y-z-explained-delamora-y-madrigal-cfe-cams-cfci

City profile. City of Kelowna. (2021, June 17). Retrieved February 8, 2023, from https://www.kelowna.ca/our-community/about-kelowna/city-profile

Shibboleth authentication request. (n.d.). Retrieved February 8, 2023, from https://my-ibisworld-com.okanagan.idm.oclc.org/ca/en/industry/44132ca/industry-performance

Wilmes, M. (2022, September 14). *4 generations, 4 styles of communication*. AnswerNet. Retrieved February 8, 2023, from https://answernet.com/blog-generations-styles-communication/

Written by Chloe Pilette for NortonLifeLock. (n.d.). *Digital Generations: The technology gap between seniors, parents, and kids*. Norton. Retrieved February 8, 2023, from https://us.norton.com/blog/how-to/digital-generations

Winter Tire Report 2020. TRAC. (2023, February 2). Retrieved February 24, 2023, from https://tracanada.ca/tire-reports/winter-tire-report-2022/

What is the upper middle class income in Canada? Reviewlution. (2022, December 4). Retrieved February 25, 2023, from https://reviewlution.ca/resources/upper-middle-class-income-in-canada/

Appendix

Technical Methodology

This section will address the issues and concerns and the necessary transformation steps.

1. **Eliminating Irrelevant Variables**

   During the initial screening phase, it was observed that the Acct key, Customer ID, and Additional vehicles would not be a good fit for analysis in further stages. Hence, these variables were removed because they would contribute to reducing the dimension of the data set and the complexity of working with it.

2. Eliminating Irrelevant Observations

There were three observations, namely the 189th,236^{th,} and 335th observations of the sample data, which did not consist of their birth years or respective discount codes. Imputation of birth years could lead to potentially misleading information and could affect the results of the analysis. Therefore, those observations were removed since they only accounted for 0.17%.

3. Renaming Variable Names

The variables were renamed for simplification purposes to access the variables with ease.

4. Imputation Techniques

There are a total of 13,272 missing values which amounted to 20.1%. The following imputation techniques were utilized.

Variables to be imputed with "None."

There were a few variables, such as primary vehicle type, 2nd vehicle type, 3rd vehicle type, 4th vehicle type, highest quality rim purchased, and discount code, including blank cells. However, these cells were replaced with a category named "none." This is because certain customers do not possess a vehicle hence a vehicle type does not exist for them.

Variables to be imputed with "Zero."

There were certain blank cells in the Discount code obtained variable, which specifies the source of the Discount code obtained by the customer. This includes values such as in-store, online, zero, and blanks. These blanks were filled with the number zero. as it was observed that the customers did not have any discount code to be used for purchasing tires.

Variables to be imputed with "N."

The five variables that accounted for the response regarding promotions for February, March, June, September, and November only contained "Y" in their respective columns. The rest of the cells had no value. Hence, these were replaced with "N," assuming the customer did not respond.

5. Removal of variables with near-zero variance

Near zero variance variables contain some unique values but do not contribute to the model by offering less to no useful information (Boehmke & Greenwell, 2020). Hence, the dataset was checked for near-zero variance. Next, the variables corresponding to the criteria were

removed, such as the number of young drivers in the home, adult drivers, 3rd vehicle type, and 4th vehicle type.

6. Encoding Categorical variables as Dummy and Ordinal

Some machine learning algorithms require encoding the categorical variables as they cannot process categorical data. Hence, label encoding is utilized when a particular order exists, and dummy encoding when unordered categorical features exist (Boehmke & Greenwell, 2020).

Furthermore, categorical variables were converted into factors and went through re-leveling as they consist of pre-specified levels.

Variables that were Dummy Encoded are as follows:

- Cust neighborhood
- Primary vehicle type, 2nd vehicle type
- Responded Feb, March, June, September, and November
- Warranty purchased.
- Discount code obtained.

Variables that were Ordinal Encoded are presented in Table 1 provided below,

Ordinal Encoding	Levels Assigned
Cust Education	1 - Basic 2 - HS Diploma 3 - College

	4 - Bachelors 5 - Graduate
Household size	1- 1 person 2- 2 persons 3- 3 persons 4- 4 persons 5- 5 persons 6- 6 or more
Household income	1 - <32000 2 - 32000 - 52000 3 - 52000 - 70000 4 - 70000 - 89000 5 - 89000 - 120000 6 - 120000 - 200000 7 - >2000000
Rim quality	0 – None 1 – Steel 2 – Aluminum 3 – Specialized
Discount code	0 – None 1-VK855356y 2 -UK016353x 3 – WS014878x

Table 2 - Ordinal encoding and their respective levels

Note to be taken into consideration:

Ordinal levels were assigned to Customer Education (Levels of Study, 2019) and Rim Quality (lessch, 2023) based on Secondary Research.

The correlation plot (See Additional Graphs) measures linearity among different variables. During analysis, some variables had a correlation value greater than 0.5, such as:

- Number of adults and adult drivers (correlation of 0.72)
- Adults in home and total drivers (correlation of 0.70)
- Adult drivers and total drivers (correlation of 0.97)

Additional Graphs

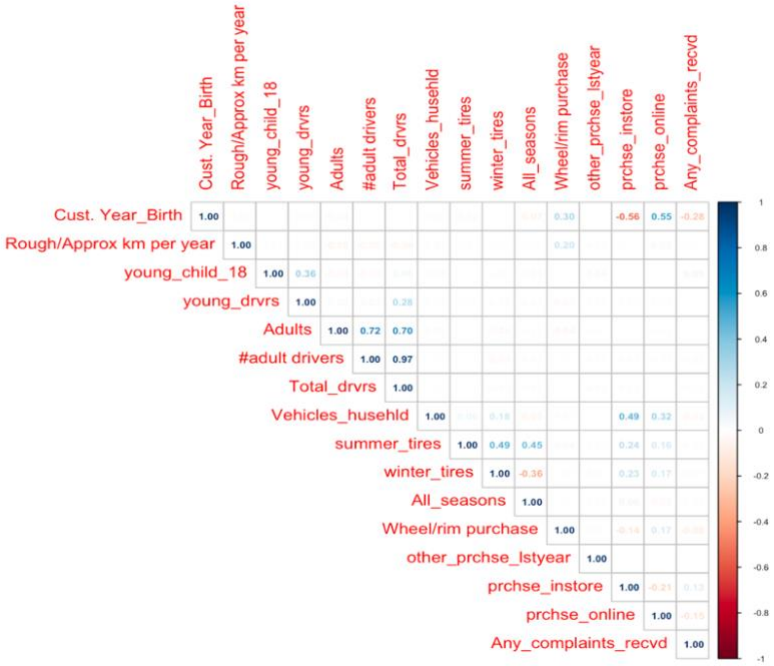


Figure 11 – Correlation plot

External links

Revenue Calculator

Code Book

Variable Name	Variable Label	Missing Data	Typical Range	Data Type	Value	Label
AcctKey	Primary Key	-	1-1689	Int		
Cust_ID	Internal ID for customer	-	-	Int		
CustYear_Birth	Year of Customer's birth	Yes	1953-2003	Int		
Cust_Neigh	Area where customer resides	-	-	Char		
Rough_km	Self-declared customer estimate of annual driving	-	0-110,000	Int		
Cust_Ed	Highest level of Education Attained	-	-	Char		
House_size	number of persons with in the household	-	1-6	Int		
Num_child	Number of drivers 16 - 18 yrs of age	-	0-4	Int		
Num_youn_driv	Customer's household primary vehicle	-	0-3	Int		
Num_adults_home	Number of adult driver in the home	-	1-6	Int		
adult_driv	Total number of drivers in the home	-	0-5	Int		
Num_driv	Total number of drivers in the home	-	0-5	Int		
House_income	Self-declared income	-	-	Char		
Num_veh	Total number of licensed, registered vehicles	-	0-4	Int		
Prim_veh	Customer's household primary vehicle	yes	-	Char		
2nd_Vehc_type		yes	-	Char		
3rd_Veh_type		yes	-	Char		
4th_Veh_type		yes	-	Char		
Add_veh	Any additional vehicles	yes	-	Char		
Sum_tires	number of sets summer tires purchased	-	0-2	Int		
Wint_tires	number of sets winter tires purchased	-	0-2	Int		
All_seas	number of sets all seasons purchased	-	0-2	Int		
Date_pur_summer		-	-	Char		
Date_pur_winter		-	-	Char		
Date_pur_allseas		-	-	Char		

Variable Name	Variable Label	Missing Data	Typical Range	Data Type	Value	Label
wheel_pur	Customer's household primary vehicle	-	0-3	Int		
Rim_qual	May or may not include recent purchase	-	-	Char		
warranty_pur	If extra warranty purchased on tires	-	-	Char		
other_parts	Other services that we provide	-	0-7	Int		
Num_store_purch	Includes parts, and services, other than tires	-	0-14	Int		
Num_onl_purch	Includes parts, and services, other than tires	-	0-7	Int		
Disc_obt	Method for obtaining code	yes	-	Char		
Disc_code	Specific discount code used	yes	-	Char		
Res_F	Did customer respond to promo?	yes	-	Char		
Res_M	Did customer respond to promo?	yes	-	Char		
Res_J	Did customer respond to promo?	yes	-	Char		
Res_S	Did customer respond to promo?	yes	-	Char		
Res_N	Did customer respond to promo?	yes	-	Char		
complaints	any complaint(marketing or special offers)	-	-	Char		

Code

```
library(readxl)
TiresData <-
  read_excel(
    "C:/Users/17789/Desktop/College/SEM 4/DSCI 490 Capstone project/Data set/
TiresDataFile_DSCI490_F22.xlsx"
  )

# dimensions of the data
dim(TiresData)

## [1] 1689   39

# checking near zero variance
library(caret)
nearZeroVar(TiresData)
```

```
## [1]  9 19 34 35 36 37 38

# changing variable names
colnames(TiresData)[1] <- "Key"
colnames(TiresData)[2] <- "ID"
colnames(TiresData)[3] <- "YearBirth"
colnames(TiresData)[4] <- "Neighbourhood"
colnames(TiresData)[5] <- "KmPerYear"
colnames(TiresData)[6] <- "Education"
colnames(TiresData)[7] <- "HouseholdSize"
colnames(TiresData)[8] <- "YoungChild_18"
colnames(TiresData)[9] <- "YoungDrivers"
colnames(TiresData)[10] <- "Adults"
colnames(TiresData)[11] <- "AdultsDrivers"
colnames(TiresData)[12] <- "TotalDrivers"
colnames(TiresData)[13] <- "Income"
colnames(TiresData)[14] <- "HouseVehicles"
colnames(TiresData)[15] <- "VehicleType_Primary"
colnames(TiresData)[16] <- "VehicleType_2nd"
colnames(TiresData)[17] <- "VehicleType_3rd"
colnames(TiresData)[18] <- "VehicleType_4th"
colnames(TiresData)[19] <- "AdditionalVehicle"
colnames(TiresData)[20] <- "SummerTires"
colnames(TiresData)[21] <- "WinterTires"
colnames(TiresData)[22] <- "Allseasons"
colnames(TiresData)[23] <- "Lstdate_prchse_summer"
colnames(TiresData)[24] <- "Lstdate_prchse_winter"
colnames(TiresData)[25] <- "Lstdate_prchse_allseason"
colnames(TiresData)[26] <- "WheelRimPurchase"
colnames(TiresData)[27] <- "RimQuality_highest"
colnames(TiresData)[28] <- "ExtraWarranty"
colnames(TiresData)[29] <- "Other_prchse_lstyear"
colnames(TiresData)[30] <- "Prchse_instore"
colnames(TiresData)[31] <- "Prchse_online"
colnames(TiresData)[32] <- "DiscountCode_obtained"
colnames(TiresData)[33] <- "DiscountCode"
colnames(TiresData)[34] <- "February"
colnames(TiresData)[35] <- "March"
colnames(TiresData)[36] <- "June"
colnames(TiresData)[37] <- "September"
colnames(TiresData)[38] <- "November"
colnames(TiresData)[39] <- "Complaints"

#check missing data
library(naniar)
gg_miss_var(TiresData, show_pct = TRUE) +
  theme(axis.text.y = element_text(size = 7))
```

```r
#Imputing Na's with N, 0 or none  as per variables requirements


TiresData["February"][is.na(TiresData["February"])] <- "N"
TiresData["March"][is.na(TiresData["March"])] <- "N"
TiresData["June"][is.na(TiresData["June"])] <- "N"
TiresData["September"][is.na(TiresData["September"])] <- "N"
TiresData["November"][is.na(TiresData["November"])] <- "N"

TiresData["DiscountCode_obtained"][is.na(TiresData["DiscountCode_obtained"])]
<-
  "0"
TiresData["RimQuality_highest"][is.na(TiresData["RimQuality_highest"])] <-
  "None"
TiresData["VehicleType_Primary"][is.na(TiresData["VehicleType_Primary"])] <-
  "None"
TiresData["VehicleType_2nd"][is.na(TiresData["VehicleType_2nd"])] <-
  "None"
TiresData["VehicleType_3rd"][is.na(TiresData["VehicleType_3rd"])] <-
  "None"
TiresData["VehicleType_4th"][is.na(TiresData["VehicleType_4th"])] <-
  "None"

# Deleting NA's in customer year birth row wise.
TiresData <- TiresData[-c(189, 236, 335), ]

# Removing unnecessary variables Accnt key, Cust ID and Addtnl Vehicles
TiresData <- TiresData[, -c(1, 2, 19)]

# checking NA's
sum(is.na(TiresData[, -c(14, 15, 16)]))

## [1] 0

# assigning weights to discount code
TiresData$DiscountCode[TiresData$DiscountCode == "None"] <- 0

TiresData$DiscountCode[TiresData$DiscountCode == "VK855356y"] <- 1
# buy one set, get a second set at 15% off

TiresData$DiscountCode[TiresData$DiscountCode == "UK016353x"] <- 2
# buy one set, 15% off your next invoice with us

TiresData$DiscountCode[TiresData$DiscountCode == "WS014878x"] <- 3
# Happy Dozen: 12% off entire purchase


# Assigning weights to Highest rim quality variable
TiresData$RimQuality_highest[TiresData$RimQuality_highest == "None"] <-
  0
TiresData$RimQuality_highest[TiresData$RimQuality_highest == "Steel"] <-
```

```
  1
TiresData$RimQuality_highest[TiresData$RimQuality_highest == "Aluminum"] <-
  2
TiresData$RimQuality_highest[TiresData$RimQuality_highest == "Specialised"] <
-
  3

#Assigning orders to Income levels
TiresData$Income[TiresData$Income == "< 32000"] <- 1
TiresData$Income[TiresData$Income == "32000 - 52000"] <- 2
TiresData$Income[TiresData$Income == "52000 - 70000"] <- 3
TiresData$Income[TiresData$Income == "70000 - 89000"] <- 4
TiresData$Income[TiresData$Income == "89000 - 120000"] <- 5
TiresData$Income[TiresData$Income == "120000 - 200000"] <- 6
TiresData$Income[TiresData$Income == "> 200000"] <- 7

#Assigning orders to Household Size
TiresData$HouseholdSize[TiresData$HouseholdSize == "1 person"] <- 1
TiresData$HouseholdSize[TiresData$HouseholdSize == "2 persons"] <- 2
TiresData$HouseholdSize[TiresData$HouseholdSize == "3 persons"] <- 3
TiresData$HouseholdSize[TiresData$HouseholdSize == "4 persons"] <- 4
TiresData$HouseholdSize[TiresData$HouseholdSize == "5 persons"] <- 5
TiresData$HouseholdSize[TiresData$HouseholdSize == "6 or more"] <- 6

# Assigning orders to Education level
TiresData$Education[TiresData$Education == "Basic"] <- 1
TiresData$Education[TiresData$Education == "HS dipl."] <- 2
TiresData$Education[TiresData$Education == "College"] <- 3
TiresData$Education[TiresData$Education == "Bachelors"] <- 4
TiresData$Education[TiresData$Education == "Graduate"] <- 5


# Separating out year from the last date purchase summer, winter & all season
TiresData$Lstdate_prchse_summer<-
  format(as.Date(TiresData$Lstdate_prchse_summer, format="%Y/%m/%d"),"%Y")

TiresData$Lstdate_prchse_winter<-
  format(as.Date(TiresData$Lstdate_prchse_winter, format="%Y/%m/%d"),"%Y")

  TiresData$Lstdate_prchse_allseason<-
    format(as.Date(TiresData$Lstdate_prchse_allseason, format="%Y/%m/%d"),"%Y
")

#write.csv(TiresData, file = "TiresData.csv")

# Checking Near zero variance
library(tibble)
library(recipes)
caret::nearZeroVar(TiresData, saveMetrics = TRUE) %>%
  rownames_to_column() %>%
  filter(nzv)
```

```
# Removing variables which have near zero variance

TiresData <- TiresData[, -c(7, 9, 15, 16)]

#young drivers, adult drivers, vehicle 3rd type and vehicle 4th type

# Converting categorical variables to factors
TiresData[sapply(TiresData, is.character)] <-
  lapply(TiresData[sapply(TiresData, is.character)], as.factor)

TiresData$Complaints <- as.factor(TiresData$Complaints)

# Dummy encoding for the variables which doesn't have any order
TiresData_dumy <-
  fastDummies::dummy_cols(
    TiresData,
    select_columns = c(
      "Neighbourhood",
      "VehicleType_Primary",
      "VehicleType_2nd",
      "DiscountCode_obtained",
      "ExtraWarranty",
      "February",
      "March",
      "June",
      "September",
      "November"
    ),
    remove_first_dummy = TRUE,
    remove_selected_columns = TRUE
  )

a<-ggplot(TiresData)+
  geom_bar(aes(Lstdate_prchse_summer,SummerTires)
           ,stat = "identity")

b<-ggplot(TiresData,aes(Lstdate_prchse_winter,WinterTires))+
  geom_bar(stat = "identity")
c<-ggplot(TiresData,aes(Lstdate_prchse_allseason,Allseasons))+
  geom_bar(stat = "identity")

gridExtra::grid.arrange(a,b,c,nrow = 2)
```

```
library(ggplot2)
library(ggbreak)
ggplot(TiresData,aes(KmPerYear, fill = ExtraWarranty))+
  geom_bar( position = "fill",show.legend = T,)+
  ggtitle("Extra Warranty purchased based on Kms driven per year")+
  theme(panel.grid.major = element_blank(),
```

```
               panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(labels = scales::percent_format(accuracy = 1))+
  labs(x="Approximation of Kms driven per year",y="Number of Customers")+
  scale_fill_discrete(labels=c("No","Yes"))+
   scale_x_continuous(limits = c(NA, 120000), breaks = c(0,10000,20000,30000,
40000,50000,60000,70000,80000,90000,100000,110000,120000))+
  theme(axis.text.x = element_text(angle = 45,hjust=1))+
  scale_fill_brewer(palette="Set2")
```

```
 TiresData_dumy[,c(2,5,6,7,9,10,11,12,16,18,19,20)]

library(cluster)
library(NbClust)


#source:https://dpmartin42.github.io/posts/r/cluster-mixed-types

TiresData[,c(12)]

Tirecluster<-TiresData[,-c(4,6,7,12,16,17,18,27,28,29,30,31,32)]

gower_dst <- daisy(Tirecluster, metric = "gower")



summary(gower_dst)

gower_mat<-as.matrix(gower_dst)

# Output most similar pair
Tirecluster[
  which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]),
        arr.ind = TRUE)[1, ], ]

# Output most dissimilar pair
Tirecluster[
  which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]),
        arr.ind = TRUE)[1, ], ]

## calculating silhoutte width for several k
#source referencee:https://dpmartin42.github.io/posts/r/cluster-mixed-types

sil_width<-c(NA)

for (i in 2:10) {
    pam_fit<-pam(gower_dst,diss = TRUE, k = i)
    sil_width[i]<-pam_fit$silinfo$avg.width
}
```

```
plot(1:10,sil_width,xlab = "Number of clusters",
     ylab = "Silhouette Width")
lines(1:10, sil_width)
```

```
## Pick the number of cluster with the highest silhoutte width

pam_fit_new<-pam(gower_dst,diss = TRUE, k = 2)


library(dplyr)

pam_results <- Tirecluster %>%
    mutate(cluster = pam_fit_new$clustering) %>%
    group_by(cluster) %>%
  do(the_summary = summary(.))

pam_results$the_summary



Tirecluster[pam_fit_new$medoids,]

#source referencee:https://dpmartin42.github.io/posts/r/cluster-mixed-types


# Visualization
library(Rtsne)
tsne_obj <- Rtsne(gower_dst, is_distance = TRUE)

tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit_new$clustering),
         feature1 = Tirecluster$YearBirth,
         feature2 = Tirecluster$Neighbourhood,
         feature3 = Tirecluster$KmPerYear,
         feature4 = Tirecluster$Education,
         feature5 = Tirecluster$HouseholdSize,
         feature6 = Tirecluster$YoungChild_18,
         feature7 = Tirecluster$Adults,
         feature8 = Tirecluster$TotalDrivers,
         feature9 = Tirecluster$Income,
         feature10 = Tirecluster$HouseVehicles,
         feature11 = Tirecluster$VehicleType_Primary,
         feature12 = Tirecluster$VehicleType_2nd,
         feature13 = Tirecluster$SummerTires,
         feature14 = Tirecluster$WinterTires,
         feature15 = Tirecluster$Allseasons,
         feature16 = Tirecluster$WheelRimPurchase,
         feature17 = Tirecluster$RimQuality_highest,
```

```
          feature18 = Tirecluster$ExtraWarranty,
          feature19 = Tirecluster$Other_prchse_lstyear,
          feature20 = Tirecluster$RimQuality_highest,
          feature21 = Tirecluster$ExtraWarranty,
          feature22 = Tirecluster$Other_prchse_lstyear,
          feature23 = Tirecluster$Prchse_instore,
          feature24 = Tirecluster$Prchse_online,
          feature25 = Tirecluster$DiscountCode_obtained,
          feature26 = Tirecluster$DiscountCode)

ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster))
```

```
library(ggplot2)
# Highest Quality rim purchased
filter(TiresData,TiresData$RimQuality_highest!="None")%>%
ggplot(.)+
geom_bar(aes(x=RimQuality_highest,fill = "RimQuality_highest"),show.legend =
F) +  labs(x='Quality of Rim types',y="Number of Wheel/Rim sets purchased")+
  ggtitle("Highest quality of Wheel/Rim sets purchased in the last 5 years")+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(
),
        plot.title = element_text(hjust = 0.5))+
  scale_fill_brewer(palette="Set3")
```

```
# Income level of customers
ggplot(TiresData) +
geom_bar(aes(x=Income, fill = Income),show.legend = F)+
  scale_fill_brewer(palette="Set2")+
  xlab('Household Income')+
    ylab('Number of customers(in hundreds)')+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(
))+
  ggtitle("Income levels of customers")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
ggplot(TiresData)+
  geom_bar(aes(x=DiscountCode,fill = March),position = "dodge")
```

```
library(dplyr)

#generation buying behaviour
```

```
TiresData <-TiresData %>% mutate (NEW = case_when(
    TiresData$YearBirth >= 1946 & TiresData$YearBirth <= 1964 ~ "Baby boomers
",TiresData$YearBirth >= 1965 & TiresData$YearBirth <= 1979 ~ "Gen X",
     TiresData$YearBirth >= 1980 & TiresData$YearBirth <= 1994 ~ "Gen Y",
    TiresData$YearBirth >= 1995 & TiresData$YearBirth <= 2010 ~ "Gen Z"
                                      ))

unique(TiresData$NEW)
```

```
## [1] "Gen X"        "Gen Z"        "Gen Y"        "Baby boomers"
```

```
library(ggplot2)

ggplot(TiresData)+
  geom_bar(aes(x = NEW,fill = NEW),show.legend = F)+
  scale_fill_brewer(palette="Set2")+
  labs(x='Generation type', y='Number of vehicles') + theme(panel.grid.major
= element_blank(),
                panel.grid.minor = element_blank()) +  ggtitle("Number of vehi
cles based on Generations")
```

```
# str(TiresData)


ggplot(TiresData, aes(x=reorder(NEW,NEW, function(x)-length(x)))) +
  geom_bar(fill='lightblue') +  labs(x='Generation types', y='Number of custo
mers') +
  theme(panel.grid.major = element_blank(),panel.grid.minor = element_blank()
) +
  ggtitle("Generations and their origin of discount") +
  geom_bar(aes(fill = TiresData$Disc_Code_Ob)) +
  guides(fill=guide_legend(title="Source of Discount"))
```

```
ggplot(TiresData, aes(x=KmPerYear)) +
 geom_histogram( colour="black", fill="white")+
  theme(panel.grid.major = element_blank(),
                panel.grid.minor = element_blank())
```

```
TiresData %>%
  ggplot( aes(x=Lstdate_prchse_summer, y=SummerTires)) +
    geom_line() +
    geom_point()
```

```
TiresData$Lstdate_prchse_summer<-
  format(as.Date(TiresData$Lstdate_prchse_summer, format="%Y/%m/%d"),"%Y")

TiresData$Lstdate_prchse_winter<-
  format(as.Date(TiresData$Lstdate_prchse_winter, format="%Y/%m/%d"),"%Y")

  TiresData$Lstdate_prchse_allseason<-
    format(as.Date(TiresData$Lstdate_prchse_allseason, format="%Y/%m/%d"),"%Y
")

TiresData%>%
  filter(Allseasons !=0 & Lstdate_prchse_allseason == 2022)%>%
  summarise(SummerTires=sum(SummerTires))
```

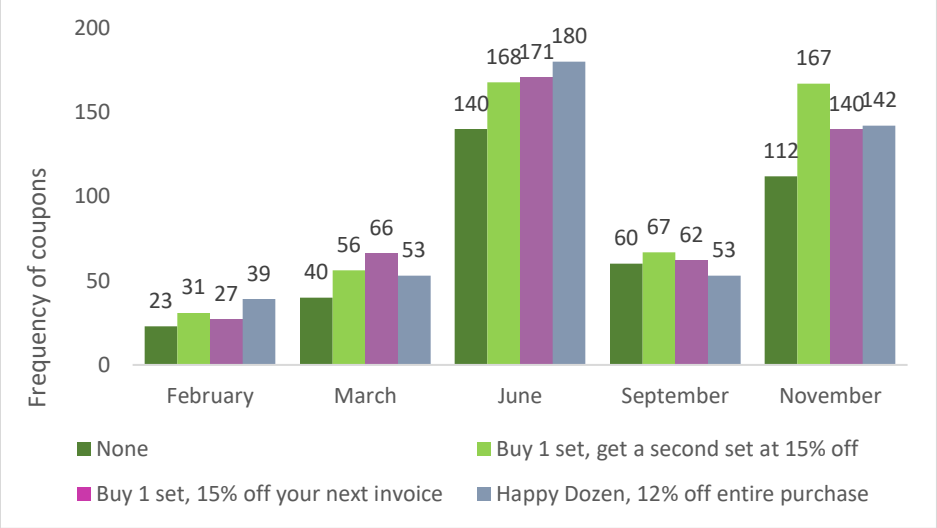
```
## # A tibble: 1 × 1
##   SummerTires
##         <dbl>
## 1           0
```

```
library(readxl)
Newdata <- read_excel("C:/Users/17789/Desktop/College/SEM 4/DSCI 490 Capstone
project/Data set/Newdata.xlsx")
Newdata
```

```
## # A tibble: 20 × 3
##    Season     Year `Tires Purchased`
##    <chr>     <dbl>             <dbl>
##  1 Summer     2014                73
##  2 Summer     2015               181
##  3 Summer     2016               171
##  4 Summer     2017               211
##  5 Summer     2018               190
##  6 Summer     2019               173
##  7 Summer     2020               181
##  8 Summer     2021               182
##  9 Summer     2022               117
## 10 Winter     2017                89
## 11 Winter     2018               240
## 12 Winter     2019               275
## 13 Winter     2020               330
## 14 Winter     2021               288
## 15 Winter     2022               186
## 16 AllSeason  2018               121
## 17 AllSeason  2019               319
## 18 AllSeason  2020               318
## 19 AllSeason  2021               314
## 20 AllSeason  2022               194
```

```
don <- Newdata %>%
  filter(Season %in% c("Summer", "Winter", "AllSeason"))
```

```
don %>%
  ggplot( aes(Year, y=`Tires Purchased`,group = Season,color = Season)) +
    geom_line()+
    geom_point()+
  ggtitle("Total Tire sets Purchased throughout the years (2014-2022)")+
  theme(panel.grid.major = element_blank(),
                panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5))+
  labs(y="Number of Tire sets purchased")
```




```
TiresData2 <- TiresData

TiresData2$Income[TiresData$Income == "< 32000"] <- 1
TiresData2$Income[TiresData$Income == "32000 - 52000"] <- 2
TiresData2$Income[TiresData$Income == "52000 - 70000"] <- 3
TiresData2$Income[TiresData$Income == "70000 - 89000"] <- 3
TiresData2$Income[TiresData$Income == "89000 - 120000"] <- 3
TiresData2$Income[TiresData$Income == "120000 - 200000"] <- 4
TiresData2$Income[TiresData$Income == "> 200000"]<-5



TiresData1 <- TiresData2 %>%
  mutate(income_band=case_when(Income == "1" ~ "Lower class", Income=="2"~ "L
ower middle class", Income=="3" ~ "Middle class",Income=="4" ~ "Upper middle
class", TRUE ~ "Upper class"))


 ggplot(data = TiresData1) +
  geom_bar(mapping=aes(x=factor(income_band,level= c('Lower class','Lower mid
dle class','Middle class','Upper middle class','Upper class')),fill = income_
band),show.legend = F)+
```

```
  ggtitle("Number of customers based on Income groups")+
   scale_fill_brewer(palette="Set2")+
  theme(panel.grid.major = element_blank(),
                panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5))+
  labs(x="Income Groups",y="Number of Customers")
```

```
table (TiresData1$income_band)

library(dplyr)


c <- data.frame(summer=TiresData$SummerTires,winter=TiresData$WinterTires,
                all=TiresData$Allseasons)
c$total <- rowSums(c)
Tires<-TiresData%>%
  mutate(Totaltires=c$total)%>%
  group_by(DiscountCode)%>%
  summarise(Totaltiresales=sum(Totaltires))




tiresemils <- mutate(TiresData,Tiresales=total)
tiresemils%>%group_by(February)%>%summarise(feb=sum(Tiresales))

tiresemils%>%group_by(March)%>%summarise(mar=sum(Tiresales))

tiresemils%>%group_by(June)%>%summarise(jun=sum(Tiresales))

tiresemils%>%group_by(September)%>%summarise(sep=sum(Tiresales))

tiresemils%>%group_by(November)%>%summarise(nov=sum(Tiresales))

filter(Tires,Tires$DiscountCode!="None")%>%
ggplot(.)+
  geom_bar(aes(x=DiscountCode,y=Totaltiresales,fill=DiscountCode),stat = "Ide
ntity",show.legend = F)+
  ggtitle("Total number of Tire sets sold based on Discount Code")+
   scale_fill_brewer(palette="Set2")+
  theme(panel.grid.major = element_blank(),
                panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5))+
  labs(x="Discount Code",y="Total Number of Tire sets sold")+
  scale_y_continuous(limits = c(NA, 1500), breaks = c(0,300,600,900,1200,1500
))
```

```
TiresData_dumy <- read.csv("TiresData_dumy.csv")

TiresData_dumy <- TiresData_dumy[,-1]
TiresData_dumy_cluster <- TiresData_dumy[ ,-c(3,4,5,6,10,11,12,13,14,15,16,18,30:40,43)]
tire_data_scaled  <- scale(TiresData_dumy_cluster)



fviz_nbclust(tire_data_scaled, kmeans, method = "silhouette")



fviz_nbclust(tire_data_scaled, kmeans, method = "wss")



km.res <- kmeans(tire_data_scaled, 7, nstart = 50)



TiresData_factor <- read.csv("TiresData_factor.csv")
tiredata <- TiresData_factor[,-c(1,5,6,7,8,13,17:19,14,15,16,20,22,23)]
cust_seg <- cbind(tiredata, cluster = km.res$cluster)
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}



categ <- cust_seg %>% group_by(cluster) %>%
  summarise(yearbirth=getmode(YearBirth), Neighbour=getmode(Neighbourhood),kilometer=me
an(KmPerYear),
            driver=getmode(TotalDrivers),income=getmode(Income),
            Houseveh=getmode(HouseVehicles),primary=getmode(VehicleType_Primary),

            rimqual=getmode(RimQuality_highest),instorepurchase=getmode(Prchse_instore),
            Prchse_online=getmode(Prchse_online), discobtained=getmode(DiscountCode_obtaine
d),discode=getmode(DiscountCode),
            Feb=getmode(February),ma=getmode(March),jun=getmode(June),sep=getmode(Septe
mber),
            nov=getmode(November),complaint=getmode(Complaints))
```