Case Study for Google Data Analytics Professional Certificate

Freyja Thoroddsen Sigurdardottir

2023-10-29

Introduction

This is my Case Study for the Google Data Analytics Professional Certificate. I use RStudio and am using the packages Tidyverse, Lubridate, and Janitor.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

ASK & PREPARE

This project is based on the *Comprehensive Credit Card Transactions Dataset*, uploaded by Rajatsurana979 to Kaggle

> **Data Source:** This dataset is a compilation of publicly available credit card transaction records from various financial institutions.
>
> **Data Collection Date:** The data was collected between January 2023 and October 2023.
>
> **Data Authorship:** The dataset was curated by Rajat Surana. Credit card transaction data is contributed by various financial institutions

I saved the data to my hard drive and loaded it into RStudio.

```r
transactions <- read_csv("credit_card_transaction_flow.csv")
```

```
## Rows: 50000 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): Name, Surname, Gender, Birthdate, Date, Merchant Name, Category
## dbl (2): Customer ID, Transaction Amount
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Let's take a look at the first few rows of the dataset. Here you can see information such as the data types for each column.

```r
head(transactions)
```

```
## # A tibble: 6 x 9
##   `Customer ID` Name     Surname   Gender Birthdate  `Transaction Amount` Date
##           <dbl> <chr>    <chr>     <chr>  <chr>                     <dbl> <chr>
## 1        752858 Sean     Rodriguez F      20/10/2002                 35.5 03/04~
## 2         26381 Michelle Phelps    <NA>   24/10/1985               2553.  17/07~
## 3        305449 Jacob    Williams  M      25/10/1981                116.  20/09~
## 4        988259 Nathan   Snyder    M      26/10/1977                 11.3 11/01~
## 5        764762 Crystal  Knapp     F      02/11/1951                 62.2 13/06~
## 6        576539 Monica   Bartlett  F      20/10/2001                 99.1 24/08~
## # i 2 more variables: `Merchant Name` <chr>, Category <chr>
```

I am now going to change the column names. They will only include lowercase letters and no spaces.Then we will take a look at the new column names.

```r
transactions <- transactions %>%
  clean_names()
```

My next step is to check for missing values and duplicates.

```r
transactions <- transactions %>%  drop_na()

transactions <- transactions %>%
  distinct()
```

There were missing values in Gender, but those rows have now been removed. There were no duplicates.

I see that the columns 'birthdate' and 'date' are formatted as chr, and I want them to be dates.

```r
transactions <- transactions %>%
  mutate(
    birthdate = as.Date(birthdate, format = "%d/%m/%Y"),
    date = as.Date(date, format = "%d/%m/%Y")
  )
```

ANALYZE & SHARE

Now it's time to analyze the data. Before that, here is a list of the categories used:

```r
#List of the categories
unique_categories <- unique(transactions$category)
unique_categories
```

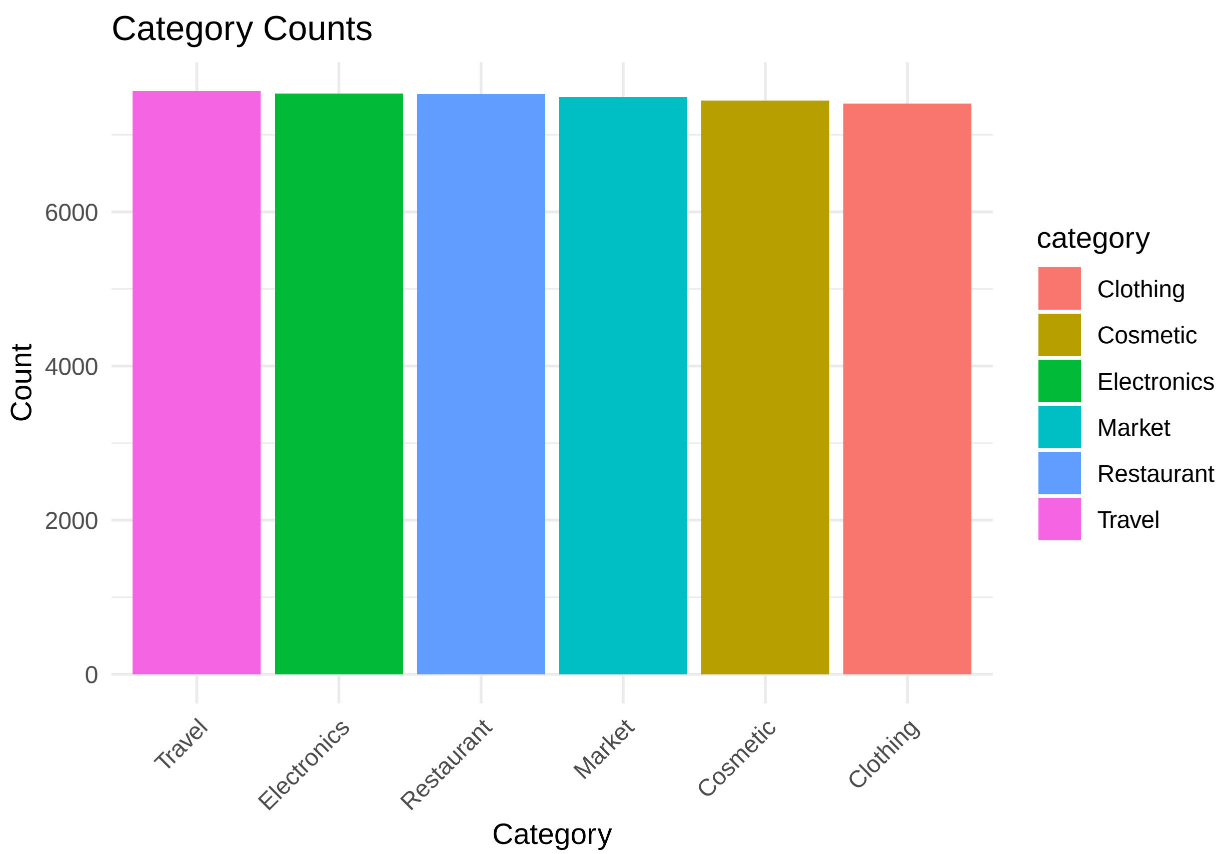
```
## [1] "Cosmetic"    "Clothing"    "Electronics" "Restaurant"  "Travel"
## [6] "Market"
```

For some basic descriptive statistics, I start by counting how many times each category appears in the dataset. This is displayed in descending order.

```r
category_counts <- transactions %>%
  group_by(category) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

category_counts
```

```
## # A tibble: 6 x 2
##   category    count
##   <chr>       <int>
## 1 Travel       7563
## 2 Electronics  7534
## 3 Restaurant   7527
## 4 Market       7488
## 5 Cosmetic     7440
## 6 Clothing     7401
```

```r
ggplot(data=category_counts, aes(x = reorder(category, -count), y = count)) +
  geom_bar(stat = "identity", aes(fill = category)) +
  ggtitle("Category Counts") +
  xlab("Category") +
  ylab("Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

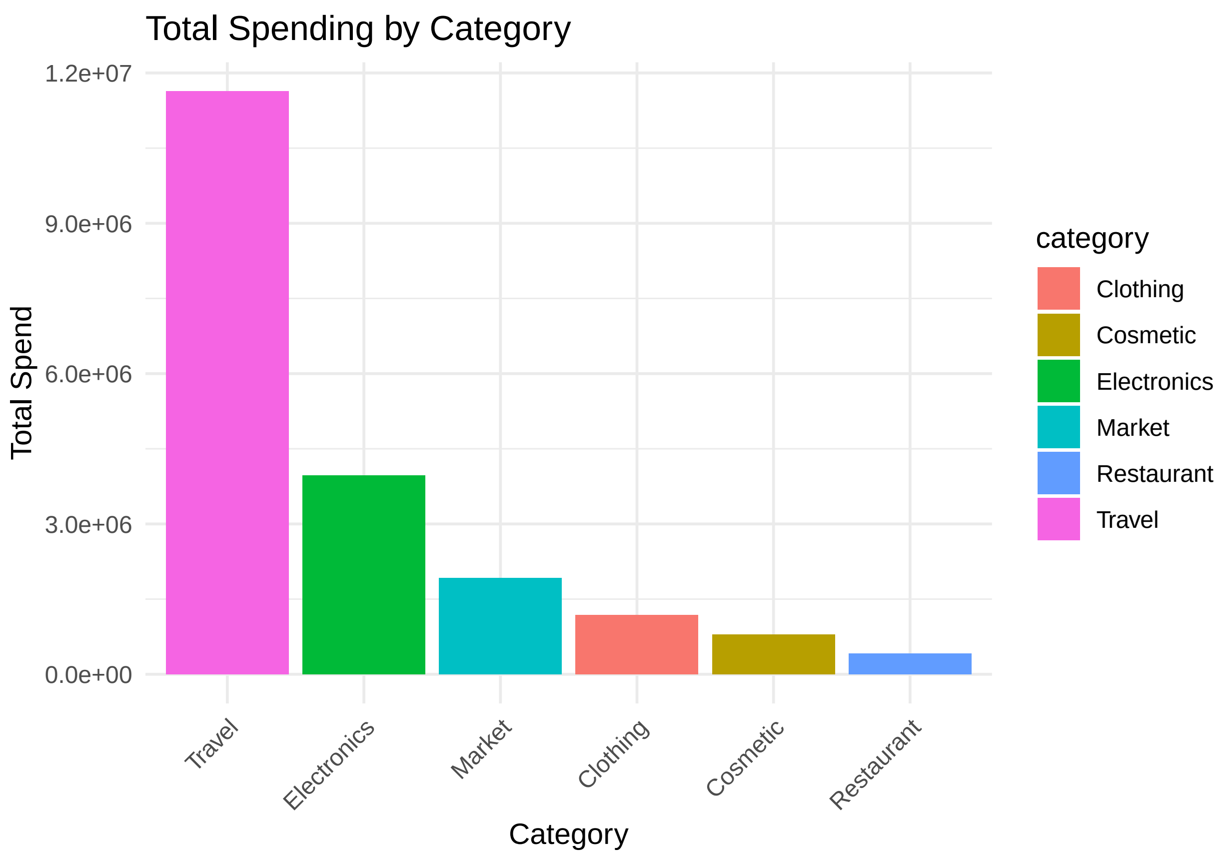


What about the total spend on each category?

```
category_spend <- transactions %>%
  group_by(category) %>%
  summarise(total_spend = sum(transaction_amount)) %>%
  arrange(desc(total_spend))

category_spend
```

```
## # A tibble: 6 x 2
##   category    total_spend
##   <chr>             <dbl>
## 1 Travel        11632036.
## 2 Electronics    3971321.
## 3 Market         1920270.
## 4 Clothing       1182933.
## 5 Cosmetic        790886.
## 6 Restaurant      416556.
```

```
ggplot(data = category_spend, aes(x = reorder(category, -total_spend), y = total_spend)) +
  geom_bar(stat = "identity", aes(fill = category)) +
  ggtitle("Total Spending by Category") +
  xlab("Category") +
  ylab("Total Spend") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



'Travel' seems to represent a large share of the total spending. I would like to see the percentage of total spend represented by that category.

```
percentage_travel <- category_spend %>%
  filter(category == "Travel") %>%
  select(total_spend) %>%
  sum() / sum(category_spend$total_spend) * 100

percentage_travel
```

```
## [1] 58.41134
```

58.36% of the total spend is on 'Travel'.

My next question: Is there a difference in buying behavior between quarters of the year?

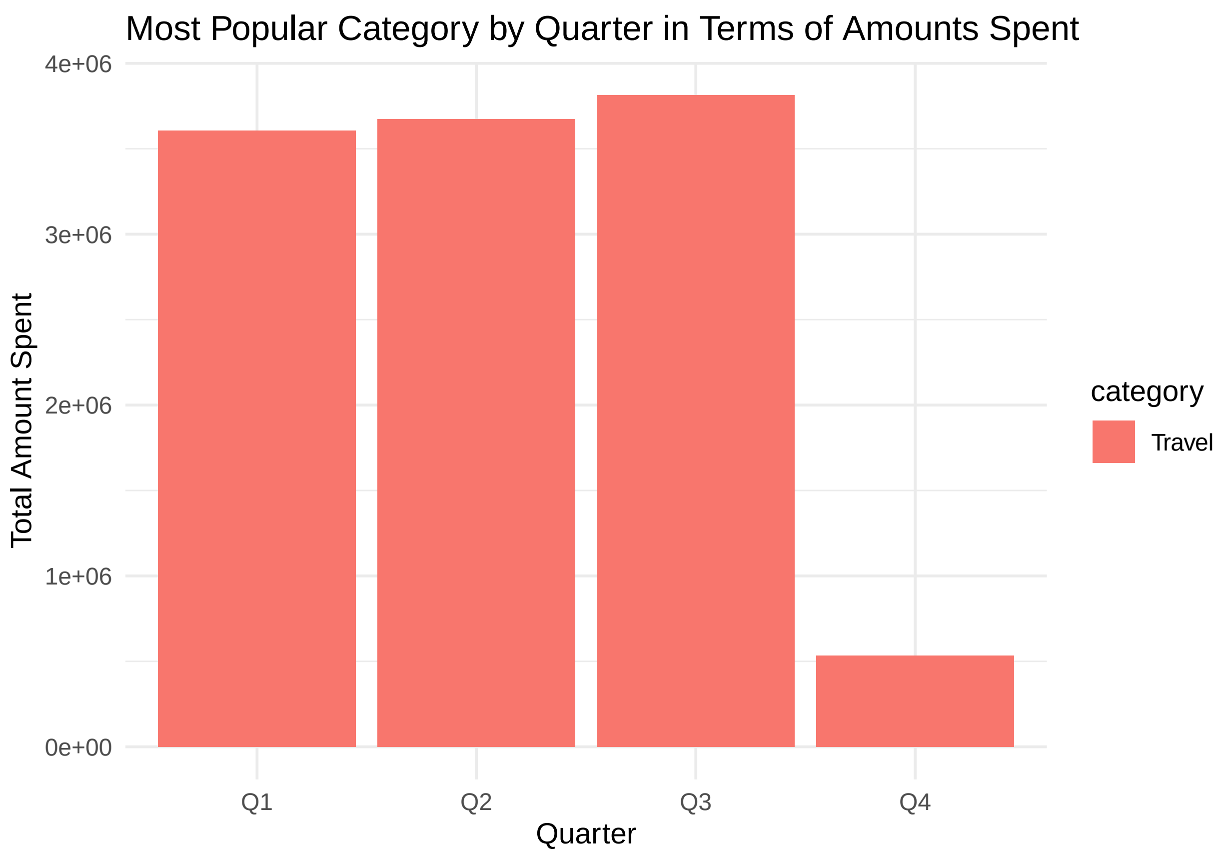
```
#Adding a 'quarter' column
transactions$quarter <- case_when(
  lubridate::month(transactions$date) %in% c(1, 2, 3) ~ "Q1",
  lubridate::month(transactions$date) %in% c(4, 5, 6) ~ "Q2",
  lubridate::month(transactions$date) %in% c(7, 8, 9) ~ "Q3",
  TRUE ~ "Q4"
)

#Note that I am least interested in Q4, as the dataset doesn't cover the entire year.
```

```r
popular_category_by_quarter <- transactions %>%
  group_by(quarter, category) %>%
  summarise(total_amount = sum(transaction_amount), .groups = 'drop') %>%
  arrange(quarter, desc(total_amount)) %>%
  group_by(quarter) %>%
  slice_head(n=1)

popular_category_by_quarter
```

```
## # A tibble: 4 x 3
## # Groups:   quarter [4]
##   quarter category total_amount
##   <chr>   <chr>           <dbl>
## 1 Q1      Travel       3607639.
## 2 Q2      Travel       3673502.
## 3 Q3      Travel       3815578.
## 4 Q4      Travel        535316.
```

Travel is the most popular category every quarter (in terms of amounts spent). Please remember Q4 is incomplete.

```r
ggplot(data = popular_category_by_quarter, aes(x = quarter, y = total_amount, fill = category)) +
  geom_bar(stat = "identity") +
  ggtitle("Most Popular Category by Quarter in Terms of Amounts Spent") +
  xlab("Quarter") +
  ylab("Total Amount Spent") +
  theme_minimal()
```

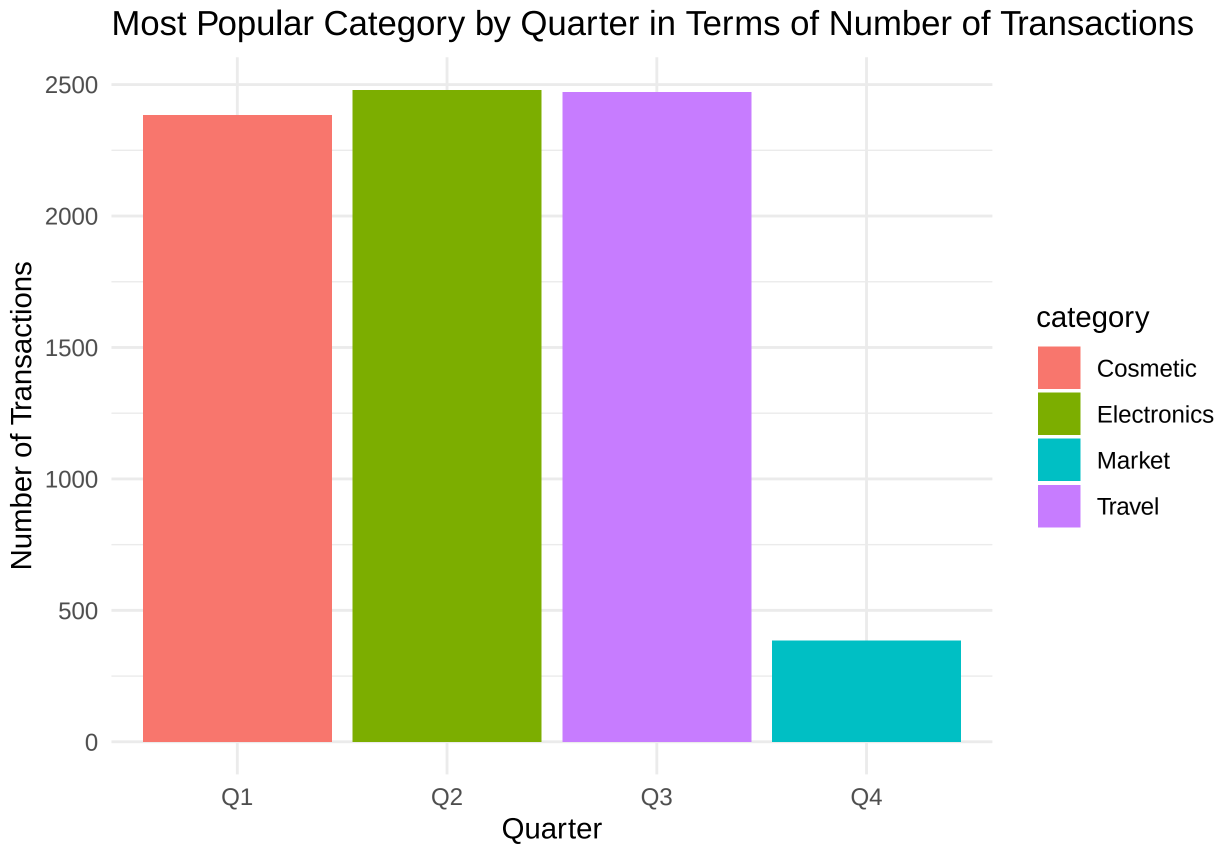


How about the most popular category every quarter in terms of number of transactions?

```
popular_category_by_quarter_count <- transactions %>%
  group_by(quarter, category) %>%
  summarise(count_transactions = n(), .groups = 'drop') %>%
  arrange(quarter, desc(count_transactions)) %>%
  group_by(quarter) %>%
  slice_head(n=1)

popular_category_by_quarter_count
```

```
## # A tibble: 4 x 3
## # Groups:   quarter [4]
##   quarter category    count_transactions
##   <chr>   <chr>                    <int>
## 1 Q1      Cosmetic                  2384
## 2 Q2      Electronics               2480
## 3 Q3      Travel                    2471
## 4 Q4      Market                     385
```

```
ggplot(data = popular_category_by_quarter_count, aes(x = quarter, y = count_transactions, fill = catego
  geom_bar(stat = "identity") +
  ggtitle("Most Popular Category by Quarter in Terms of Number of Transactions") +
  xlab("Quarter") +
  ylab("Number of Transactions") +
  theme_minimal()
```



Here we see different results. Please remember that Q4 is incomplete.

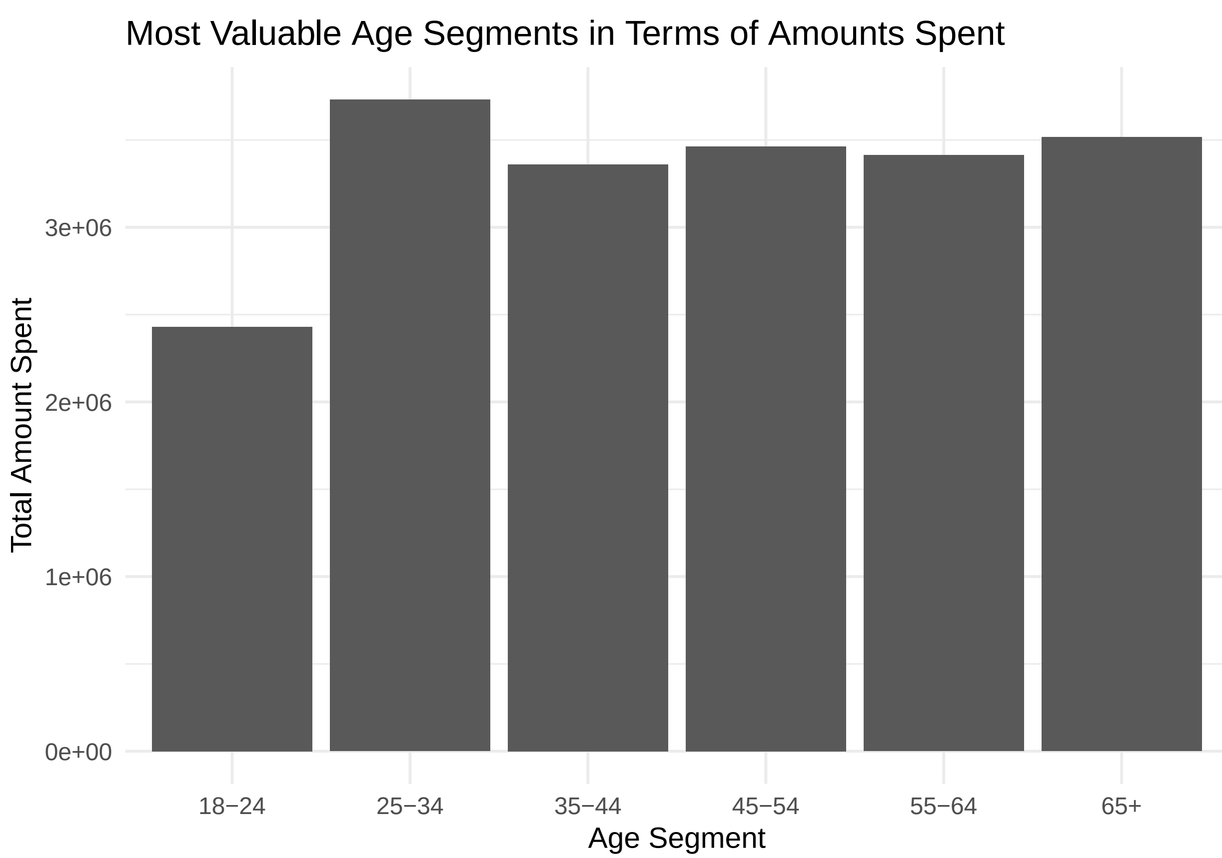
Another analysis I want to do is to segment the customer based on age, and see which segment is the most valuable in terms of amounts spent.

```
#I will calulate the age for each customer using their birthdate and current date.

transactions$age <- as.integer(difftime(Sys.Date(), transactions$birthdate, units = "weeks") / 52.25)

#Now there's a column for age

#Now calculate the different age segments

transactions <- transactions %>%
  mutate(
    age_segment = case_when(
      age <= 24 ~ "18-24",
      age >= 25 & age <= 34 ~ "25-34",
      age >= 35 & age <= 44 ~ "35-44",
      age >= 45 & age <= 54 ~ "45-54",
      age >= 55 & age <= 64 ~ "55-64",
      age >= 65 ~ "65+"
    )
  )

#And to find out which segment is the most valuable in terms of amounts spent:
```

```r
most_valuable_age_segment <- transactions %>%
  group_by(age_segment) %>%
  summarise(total_amount_spent = sum(transaction_amount), .groups = 'drop') %>%
  arrange(desc(total_amount_spent))

most_valuable_age_segment
```

```
## # A tibble: 6 x 2
##   age_segment total_amount_spent
##   <chr>                    <dbl>
## 1 25-34                 3730895.
## 2 65+                   3515765.
## 3 45-54                 3463041.
## 4 55-64                 3412965.
## 5 35-44                 3360915.
## 6 18-24                 2430422.
```

```r
ggplot(data = most_valuable_age_segment, aes(x = age_segment, y = total_amount_spent)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle("Most Valuable Age Segments in Terms of Amounts Spent") +
  xlab("Age Segment") +
  ylab("Total Amount Spent") +
  theme_minimal()
```



As seen here, the age segment 65+ is the most valuable in terms of amounts spent.

How about some deeper analyses, beyond descriptives?

I'm doing an ANOVA to compare means of transaction amounts across different age groups and transaction amounts

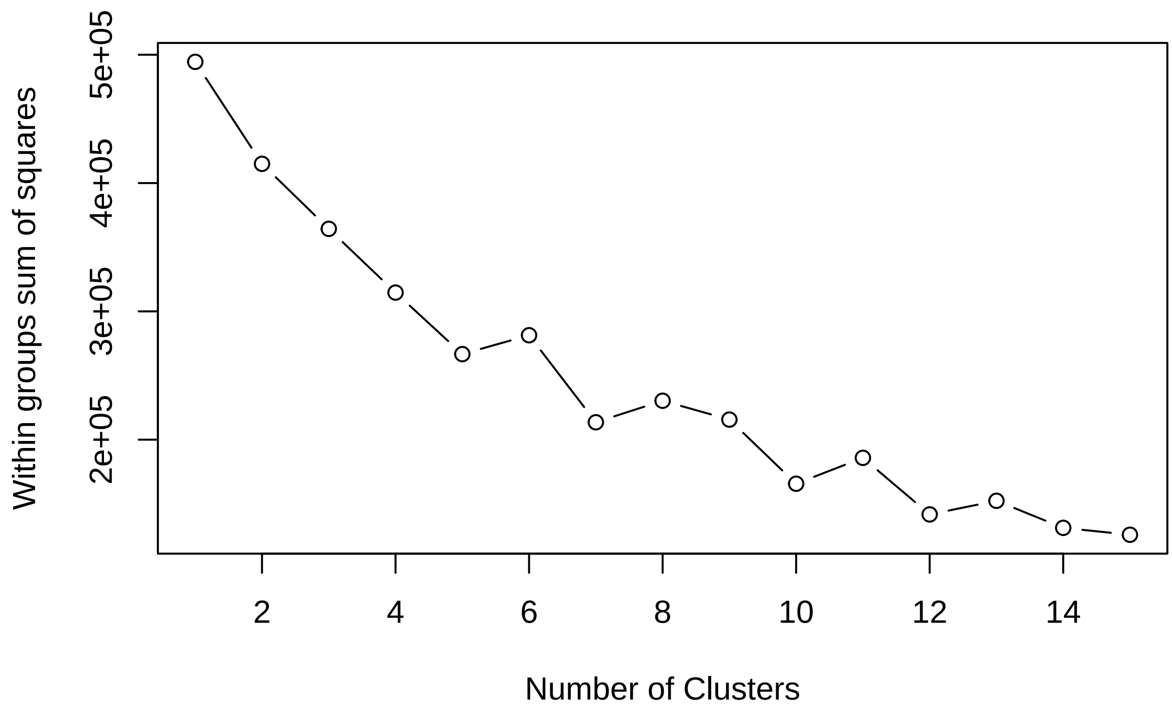
```
anova_result <- aov(transaction_amount ~ age_segment, data = transactions)
summary(anova_result)
```

```
##                Df    Sum Sq Mean Sq F value Pr(>F)
## age_segment     5 9.658e+05  193153   0.483  0.789
## Residuals   44947 1.798e+10  400062
```

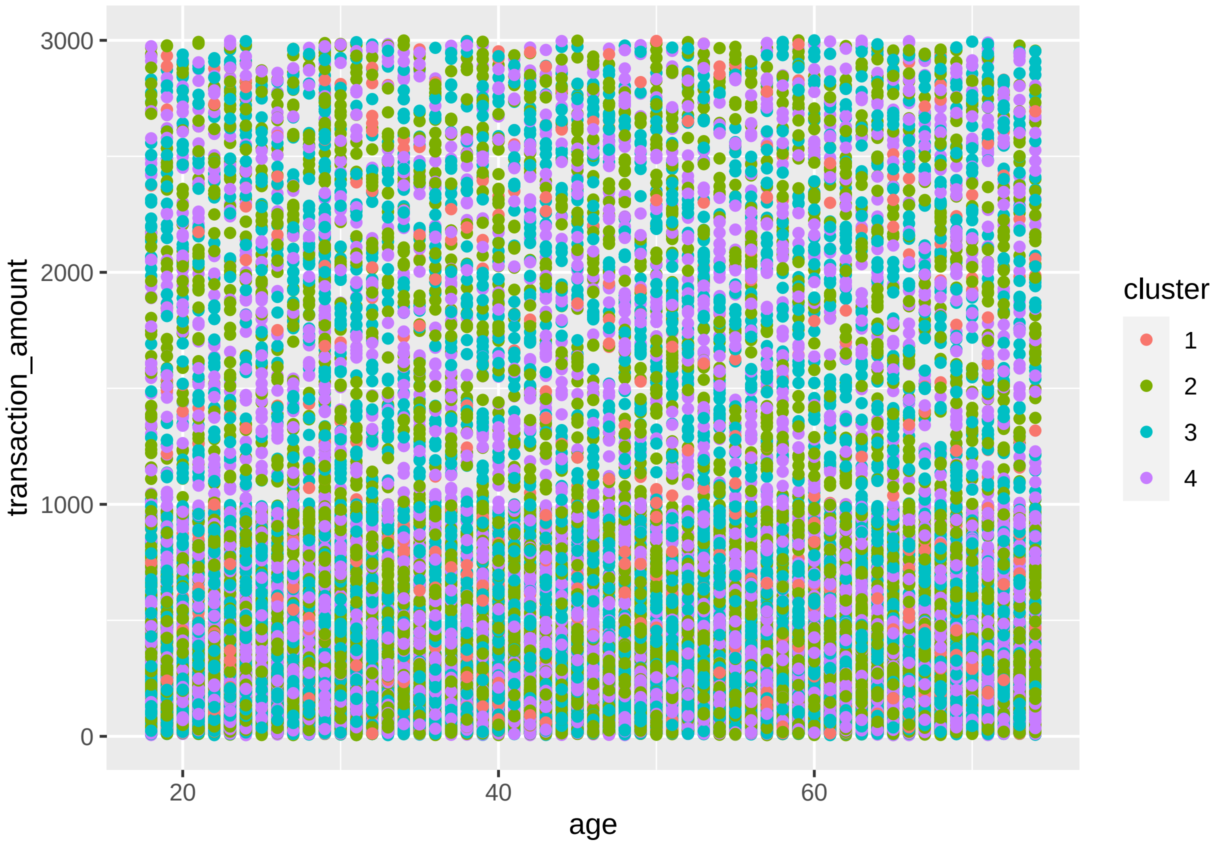
There does not seem to be a statistically significant difference in 'transaction_amount' across the different 'age_segment' groups.

#I want to try a cluster analysis.

```
# Selecting relevant features for clustering. Transaction amount, age, quarter, category
cluster_data <- transactions %>% select(transaction_amount, age, quarter, category)

# Converting categorical variables into dummy variables
cluster_data <- as.data.frame(model.matrix(~.-1, data=cluster_data))

# Standardizing the data
scaled_data <- scale(cluster_data)

# Computing total within-cluster sum of square
wss <- (nrow(scaled_data)-1) * sum(apply(scaled_data,2,var))

for (i in 2:15) wss[i] <- sum(kmeans(scaled_data, centers=i)$tot.withinss)

# Plotting the elbow graph
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")
```




```r
set.seed(123) # For reproducibility

# Applying k-means clustering
kmeans_result <- kmeans(scaled_data, centers=4)


# Adding the cluster assignments back to the original data
transactions$cluster <- as.factor(kmeans_result$cluster)

# Summarizing clusters
cluster_summary <- transactions %>%
  group_by(cluster) %>%
  summarise(
    avg_transaction = mean(transaction_amount),
    most_common_category = names(sort(table(category), decreasing = TRUE)[1]),
    avg_age = mean(age),
    .groups = 'drop'
  )

# Plotting clusters
ggplot(transactions, aes(x=age, y=transaction_amount, color=cluster)) + geom_point()
```




```r
print(cluster_summary)
```

```
## # A tibble: 4 x 4
##   cluster avg_transaction most_common_category avg_age
##   <fct>             <dbl> <chr>                  <dbl>
## 1 1                  436. Market                  45.8
## 2 2                  444. Electronics             45.7
## 3 3                  448. Travel                  45.8
## 4 4                  438. Cosmetic                45.6
```

Cluster 1: "Market regulars" **Average Transaction Amount:** $220.52 **Most Common Spending Category:** Market **Average Age:** 45.6 **Interpretation:** This cluster consists of middle-aged individuals who mostly spend at markets. They may be focused on everyday purchases like groceries.

Cluster 2: "Tech Enthusiasts" **Average Transaction Amount:** $222.14 **Most Common Spending Category:** Electronics **Average Age:** 45.6 **Interpretation:** These individuals are also middle-aged (like Cluster 1) and spend slightly more than the first cluster, but their primary interest is in electronics.

Cluster 3: "Dine-Out Lovers" **Average Transaction Amount:** $220.65 **Most Common Spending Category:** Restaurant **Average Age:** 45.4 **Interpretation:** This cluster has a similar spending average and age to the first two clusters but prefers spending their money on dining out.

Cluster 4: "High-Spending Travelers" **Average Transaction Amount:** $1,541 **Most Common Spending Category:** Travel **Average Age:** 45.5 **Interpretation:** This is the high-spending group among the clusters, focusing mainly on travel. Their average transaction is significantly higher than the others, indicating that they may be less price-sensitive when it comes to travel expenses.

Business Strategy

- The cluster “Market Regulars” might be good targets for grocery store promotions or loyalty programs.
- “Tech Enthusiasts” could be targeted with electronics promotions or with information on the launch of new tech gadgets.
- “Dine-Out Lovers” might be interested in restaurant week or other dining promotions.
- “High-Spending Travelers could be attracted through travel packages or loyalty programs that offer significant rewards for high spending.

Limitation and Further Studies

The main limitation of this dataset is the fact that it does not cover an entire year. Quarters 1-3 are complete, but Q4, which is often associated with increased sales, is incomplete. Future studies could include an entire year (or more), to get a more complete idea of spending habits across the year. These analyses are relatively straightforward but represent my current skill level. As I improve I will hopefully be able to conduct more thorough analyses.