DSC520 Final Project on Consumerism

Danish Khan

2023-10-31

Introduction

The world of consumerism is a huge part of an individual's daily life, and a strong factor in driving the economy. The concept of consumerism outlines the increase of the consumption of goods is the basis for a thriving economy. In other words, the more that is purchased and consumed, the more prosperous the societal economy becomes.

Though there are many aspects to consumerism such as goods and services, I would like to explore the retail aspect of the concept. Large companies and manufacturers provide mass-produced products to customers, where they demonstrate how their improved products can create value for the customer at a desired price point. Creating successful products is only a fraction of the battle a profit-oriented business needs to be successful; the remaining challenges include understanding customer spending strategies, categorical interests, marketing campaigns, discounts, yearly sales seasons, etc. Understanding what drives a customer to the store to purchase products and maximize profits is critical for any business's individuals to understand. To be able to predict sales and forecast profit based on trends and consumer factors is core to a retailer's success.

These are some of many questions of consumerism that data science can facilitate in answering. One common application of data science in retail consumerism is forecasting product demand. Knowing the direction of the market trend can be advantageous to increase sales and guide product releases. For instance, exploring historical data will provide an idea of time frame for when sales in the year are high or low, allowing the retailer to provide mitigate the waves of profit by introducing discounts as price optimization or releasing new products. Another common application of data science in retail consumerism is consumer segmentation. Identifying which consumers are likely to purchase certain products together or how they weigh which products are needs vs wants.

Research Questions

1. Which time of the year are consumer sales at their highest?
2. Which time of the year are consumer sales at their lowest, and how is this slump mitigated?
3. Are there product categories that consumers tend to purchase simultaneously (i.e., electronics and traveling gear)?
4. Do coupons tend to help with increasing the demand of products?
5. What are the highest sold product categories that a retailer can identify for launching a new product?
6. Are there multiple driving variables that a consumer considers with their purchasing decisions?

Approach

To address this, I plan to gather relevant information regarding consumer sales data across any category of retail. Ideally the data will span across a lengthy time frame to identify seasonal pattern changes. Comparing different variables within consumer spending decisions (i.e., sales, time of year, product types, etc.) will allow me to determine which factors are critical in forecasting consumer behavior. From there I can propose how product sales can be predicted, and how the forecast may be influenced by changing any researched independent variables. Understanding relationships between variables and which factors correlate with each other will be important express through a multitude of visuals and graphs.

How Your Approach Addresses the Problem

My approach will provide insight into how retail consumerism can be influenced. I may not be able to fully address the topic because there is much more that goes into sales forecasting beyond what is readily available to me. My goal is to explore a sales forecasting model that more than just relies on simple historical data. Through exploring numerous variables such as market trends, customer product preferences, incentives, and marketing costs, I can approach the idea of a sales forecasting model through multivariate analysis. The potential resulting model will be comprised of independent variables that can each be manipulated for an enterprise to accurately forecast as necessary. Though this approach is more accurate than simple historical-based predictions due to the incorporation of a broader range of factors, I understand that statistical limitations on my part will allow me to explore a solution only partially.

Data

Source:

Kaggle – "Marketing Insights for E-Commerce Company" https://www.kaggle.com/datasets/rishikumarrajvansh/marketing-insights-for-e-commerce-company The data sets provided are based on a small excerpt of consumer transaction data across the country, from 1/1/2019 to 12/31/2019. The data has been optimized for use as a teaching tool and exploration.

Customers_Data.xlsx

In this data set, there are nearly 1500 total customers each with a unique identifier, along with male or female gender, location of residence, and the duration of residence in months. This dataset is straightforward with no visible abnormalities; however, a customer's unique identification will have to be omitted as a change since the identifier will not be used as a variable.

Variables:

CustomerID: Individual identifier Gender: Male/Female Location: Location of residence Tenure_Months: Time of residence

Discount_Coupon.csv

In this data set, there are a few helpful variables present. There are categorical variables such as month, coupon code, and product category. The only numerical variable present is discount total. Coupon code will need to be omitted since the written data is arbitrary and I'm only concerned with the percent discount value instead of the name of the coupon itself. This dataset can be used to break consumer product types into the various categories listed. I can use the information to understand during which time of the year do specific types of products go on sale, and by how much.

Variables:

Month: Month with active coupon Product_Category: Product type Coupon_Code: Coupon redemption code Discount_pct: Coupon price discount

Marketing_Spend.csv

This dataset is fairly simple; however, it contains marketing expenditure information, which will be useful in determining how marketing expenditures of products might directly influence the profit margin and number of products sold.

Variables:

Date: Time of the year Offline_Spend: Marketing spend outside of the internet Online_Spend: Marketing spend on online sources

Online_Sales.csv

This is the bulk of the data that will be used and contains nearly 53000 items of information. All previously discussed variables come together to identify how each consumer went about purchasing. Transaction days and coupons used are listed, as well as both product types and product categories. Quantity of items are also listed with an overall final checkout price. A few columns will need to be omitted, such as delivery/shipping fees and product SKU. These identifiers may not affect the model that will be proposed. There are a few oddities that will need to be sorted, such as ensuring all product categories do not contain any unique symbols nor numbers, as a few of them do. Date format will also need to be normalized throughout the data set.

Variables:

CustomerID: Individual identifier Transaction_ID: Transaction identifier Transaction_Date: Date of transaction Product_SKU: Product SKU identifier Product_Description: Product itself Product_Category: Product type Quantity: Total number of items Avg_Price: Price per item Delivery_Charges: Delivery cost Coupon_Status: Coupon usage

Tax_Amount.xlsx

This data set is minimal and only includes possible sales tax on product categories. This may or may not affect modeling, however it cannot be omitted until it has been proven that final sales tax does not affect consumer purchase decisions.

Variables:

Product_Category: Product type GST: Sales tax percentage

Required Packages

Ggplot2: for graphing Dplyr: for analyzing and transforming data frames Purrr: for filtering through data Metrics: for predictive and modeling purposes

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(purrr)
library(Metrics)
```

Plots and Tables Needed

Data frames and matrices will both be useful forms of table formats to organize and transpose data for analysis as needed. A scatter plot will enable me to compare two independent variables that are integers directly against each other. Histograms will enable me to categorize the numbers of sales per type of item purchased. A q-q plot will allow me to assess the accuracy of a line-of-fit between an independent and a dependent variable and help with assessing normal distribution. Box plots could be potentially useful to determine the mean, medium, and quantile range of sales over the range of one year.

Questions for Future Steps

In terms of data for analysis, I do not know the extent of sales platforms as the subsets of data I've chosen pale in comparison to true big data available on the subject. Learning how this can limit my modeling proposals will enable me to work around data limitations to still come up with an accurate suggestion. In terms of data analysis and modeling, one thing I do not know is how to run simulations through a multiple regression model repeatedly, while assessing best-of-fit tests to ensure a more accurate model fit over time. I imagine related concepts will be touched on in the future under machine learning.

How to Import and Clean My Data

The first step is to convert all .xlsx files into .csv formatted files to correctly import and assign to variables. Once completed I grouped all variables in each spreadsheet using the “stats” extension in excel to find any typos or misspellings. I did end up finding one major misspelling and it was in the “Online_Sales.csv” sheet, where some products were listed as “Dog Frisbee” and others were listed as “7” Dog Frisbee.” I replaced all incorrect variables with the correct product “Dog Frisbee.” I then looked through all data sheets to see if there are any obvious variables that can be eliminated. I eliminated columns “Product_SKU” and “Delivery_Charges” from “Online_Sales.csv” as these variables are arbitrarily assigned and the products already have their own descriptive unique identifiers. Once sorted, I imported the data using the read.csv() function to automatically convert the files to a data frame with observations and variables.

```r
customers <- read.csv(
  "/Users/danishk/Documents/DSC520/Final Project/Customers_Data.csv")
coupons <- read.csv(
  "/Users/danishk/Documents/DSC520/Final Project/Discount_Coupon.csv")
marketing <- read.csv(
  "/Users/danishk/Documents/DSC520/Final Project/Marketing_Spend.csv")
sales <- read.csv(
  "/Users/danishk/Documents/DSC520/Final Project/Online_Sales.csv")
tax <- read.csv(
  "/Users/danishk/Documents/DSC520/Final Project/Tax_Amount.csv")
```

What Your Final Data Set Looks Like

```r
head(customers)
```

```
##   CustomerID Gender   Location Tenure_Months
## 1      17850      M    Chicago            12
## 2      13047      M California            43
## 3      12583      M    Chicago            33
## 4      13748      F California            30
## 5      15100      M California            49
## 6      15291      M California            32
```

```r
head(coupons)
```

```
##   Month Product_Category Coupon_Code Discount_pct
## 1   Jan          Apparel      SALE10           10
## 2   Feb          Apparel      SALE20           20
## 3   Mar          Apparel      SALE30           30
## 4   Jan         Nest-USA      ELEC10           10
## 5   Feb         Nest-USA      ELEC20           20
## 6   Mar         Nest-USA      ELEC30           30
```

```r
head(marketing)
```

```
##       Date Offline_Spend Online_Spend
## 1 1/1/2019          4500      2424.50
## 2 1/2/2019          4500      3480.36
## 3 1/3/2019          4500      1576.38
## 4 1/4/2019          4500      2928.55
## 5 1/5/2019          4500      4055.30
## 6 1/6/2019          4500      3796.85
```

```r
head(sales)
```

```
##   CustomerID Transaction_ID Transaction_Date
## 1      17850          16679           1/1/19
## 2      17850          16680           1/1/19
## 3      17850          16681           1/1/19
## 4      17850          16682           1/1/19
## 5      17850          16682           1/1/19
## 6      17850          16682           1/1/19
##                                      Product_Description Product_Category
## 1 Nest Learning Thermostat 3rd Gen-USA - Stainless Steel         Nest-USA
## 2 Nest Learning Thermostat 3rd Gen-USA - Stainless Steel         Nest-USA
## 3                  Google Laptop and Cell Phone Stickers           Office
## 4   Google Men's 100% Cotton Short Sleeve Hero Tee Black          Apparel
## 5                        Google Canvas Tote Natural/Navy             Bags
## 6                                              Sport Bag             Bags
##   Quantity Avg_Price Coupon_Status
## 1        1    153.71          Used
## 2        1    153.71          Used
## 3        1      2.05          Used
## 4        5     17.53      Not Used
## 5        1     16.50          Used
## 6       15      5.15          Used
```

```r
head(tax)
```

```
##   Product_Category    GST
## 1         Nest-USA 10.00%
## 2           Office 10.00%
## 3          Apparel 18.00%
## 4             Bags 18.00%
## 5        Drinkware 18.00%
## 6        Lifestyle 18.00%
```

Questions for Future Steps

One thing I don't know currently is how to proceed with a data clean-up by misspellings in R instead of in excel. Summarizing all possible variables in a list of any selected data sheet without the statistical calculations that the function summary() would allow me to visualize how many possible misspellings of variables exist. This would also allow me to strictly complete all cleaning and data manipulation within R, further minimizing human-error.

Information That Is Not Self-Evident

Looking at each data set, nearly all columns represent a form of independent variable that can be defined as used. The correlation between variables is not evident as the data sets are laid out in a random order. The causation of variables is also not evident going going the data sets. These supposed impacts that variables have on each other, such as whether or not discounts drive a specific product's sale, will need to be explored as these conclusions are not obvious. Sorting through the data sets in ascending/descending order or categorically may make some observations self-evident but not absolute.

Different Ways to Look at the Data

All the different ways to look at the data set will allow for different forms of interpretation. One important way to look at the data is by identifying which variables can compared, which variables cannot, and which variables can be combined. Splitting the variables into numerical and non-numerical data is easiest, followed by understand which non-numerical data is considered nominal so as not to redundantly combine it with other such variables. The non-numerical can further be qualitatively analyzed through observations, assessing possible patterns present in which columns that represent the qualitative variable. Visually these variables can be understood through histograms or distribution graphs. Naturally, the next type of data after understand qualitative variables, is to explore and isolate which quantitative variables are present. This will allow for direct comparison, as numerical values can easily be understood visually through scatter plots, or numerically via matrix formulations. These values can contribute to simple tests such as through functions like p-value, correlation, and covariance. Juxtaposing related variables will allow a quick visual check of preliminary conclusions before any statistical exploration.

Plan to Slice and Dice the Data

I plan to combine as many aspects of the data sheets are possible. The "Online_Sales.csv" sheet is considered to be the main source of data, where the other spreadsheets act as definitions of measured variables. For example, The "Online_Sales.csv" mentions which coupons were used and who the customer was that used them. The "Customers_Data.csv" sheet contains personal and residential information of all possible customers and the "Discount_Coupon.csv" sheet contains all possible coupons and total percentage discount off pricing. Combining these definitions and variables into one master data frame would allow for direct comparisons between variables both through quantitative tests and through visualizing on graphs. I also would like to explore filtering the rows in the data by key variables under specific columns for comparison purposes. This will allow me look at other independent variables in the context of individual selected variables to identify any patterns.

How to Summarize the Data to Answer Key Questions

There are a few ways to summarize the data to help portray the pattern or story of what key factors are that drive retail consumerism. First it's important to identify the customer and sales geography. Who are the customers? Differences in genders? Are there major metropolitan locations that can be targeted for product sales or launches? This will be critical for market information since identifying an audience is a major component for targeting advertisements. Another important aspect of data summary is to understand spending patterns throughout the year, where all the quantities and average prices will need to be multiplied together and graphed to demonstrate any possible trends. A possible trend here would suggest any seasonal relevance that contributes to the identified customers' spending goals Once pricing summaries have been established, the distribution of the prices across the product ranges offered will be important in determining which kinds of products are in demand. From here possible conclusions can be made utilizing the mentioned summaries, helping to both tell a story and to suggest a model for retail consumerism predictions.

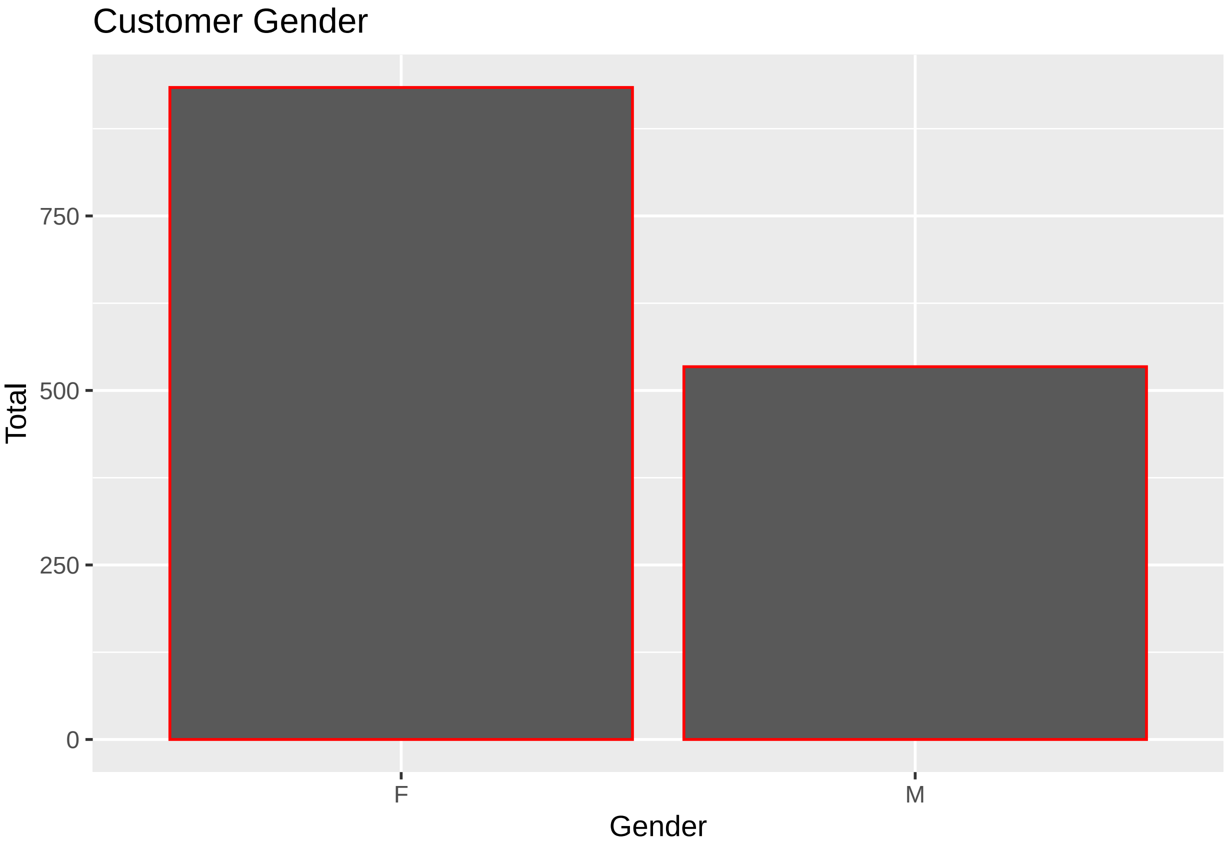
Types of Plots and Tables to Help Illustrate Findings

To identify customer demographics, I used the histograms below to visualize genders and popular cities of residence. According to the Customer Gender graph, nearly 40% more customers are female compared to males. According to the Customer Location graph, the most popular cities for consumerism are California and Chicago, followed closely by New York.

```
cust_graph1 <- ggplot(customers, aes(x=Gender)) +
  geom_histogram(binwidth=1, color="red", stat="count") +
  xlab("Gender") + ylab("Total") + ggtitle("Customer Gender")
```

```
## Warning in geom_histogram(binwidth = 1, color = "red", stat = "count"):
## Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```

```
cust_graph1
```

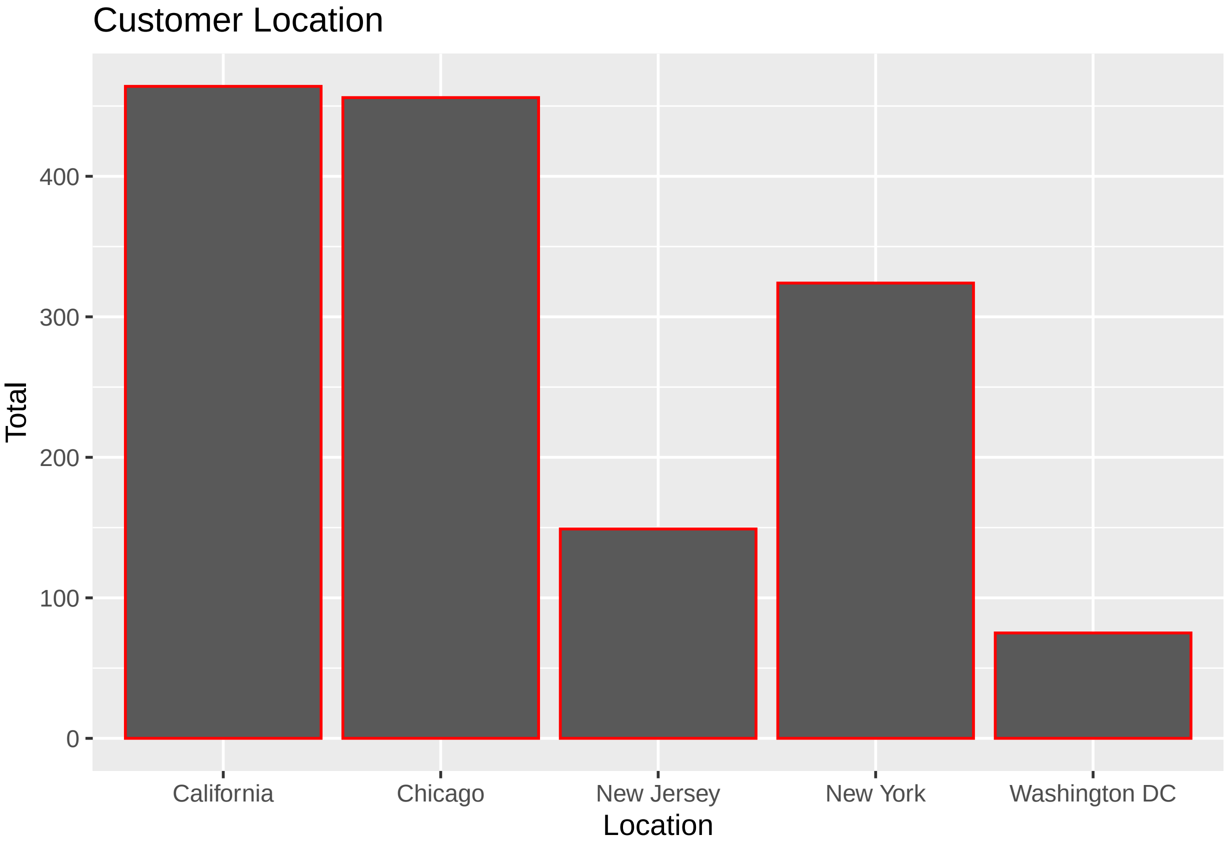


```
cust_graph2 <- ggplot(customers, aes(x=Location)) +
  geom_histogram(binwidth=1, color="red", stat="count") +
  xlab("Location") + ylab("Total") + ggtitle("Customer Location")
```

```
## Warning in geom_histogram(binwidth = 1, color = "red", stat = "count"):
```

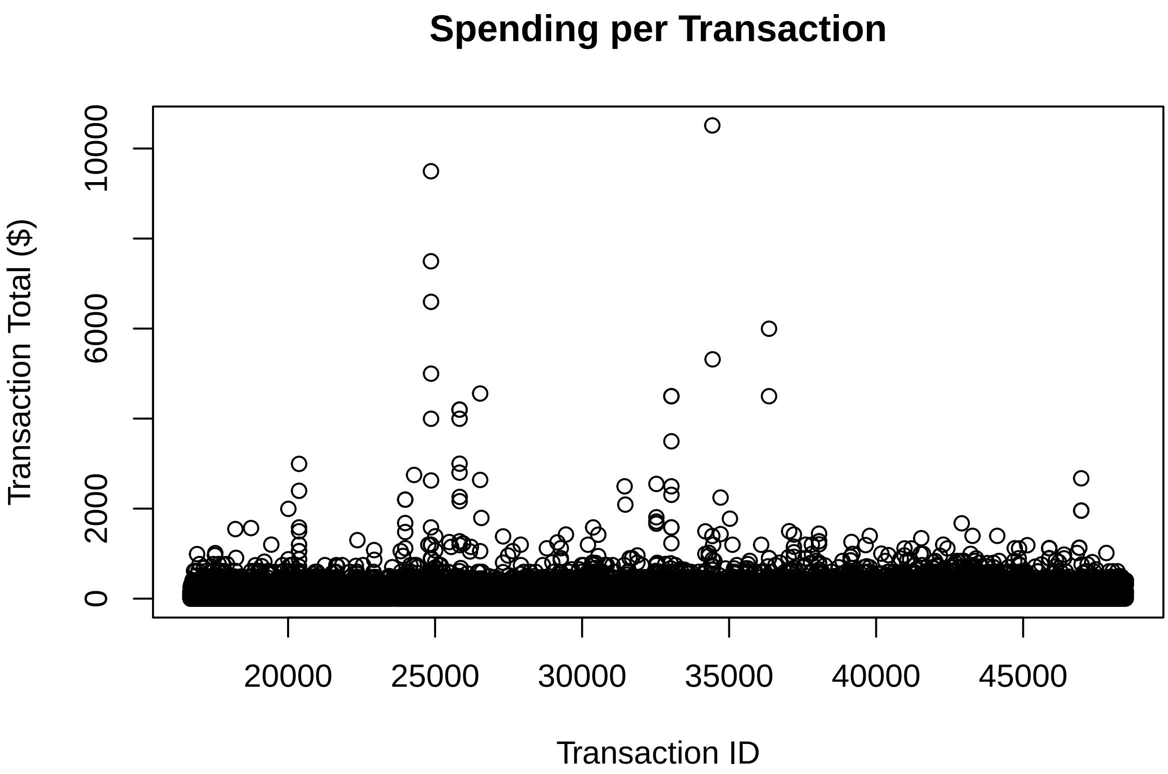
```
## Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```

```
cust_graph2
```

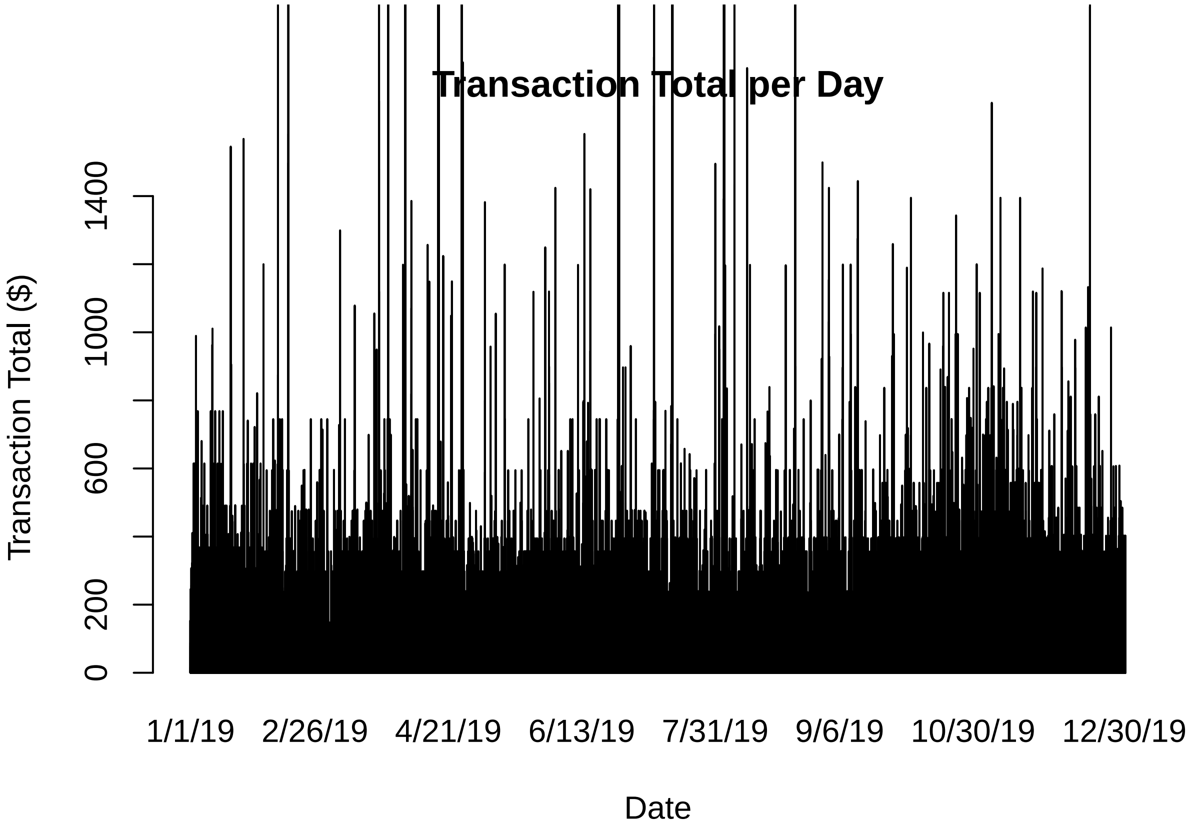


The next parameter to identify is the range of spending, and to see if there's a possible pattern visible within that range of spending. I used a basic plot combining arbitrary transaction IDs against the total money spent, and to obtain the total money spent I multiplied the columns of total items in each transaction with the column of average price of each item in each transaction. Visually I can create an upper limit of $1500 and adjust the y-range to further amplify spending patterns, keeping other exceeding purchases from tweaking the visual trend. Once the spending range was identified, I created a bar plot to visualize spending on each day of the year from January 1st all the way through December 31st. I created another bar plot that combines both online marketing spend (i.e. internet adverts) and offline marketing spend (i.e. magazines) to understand if advertising trends are similar to spending trends. After taking a look at spending and market graphs, there may be a visual trend that will need to be explored between the months of November, December, and January.

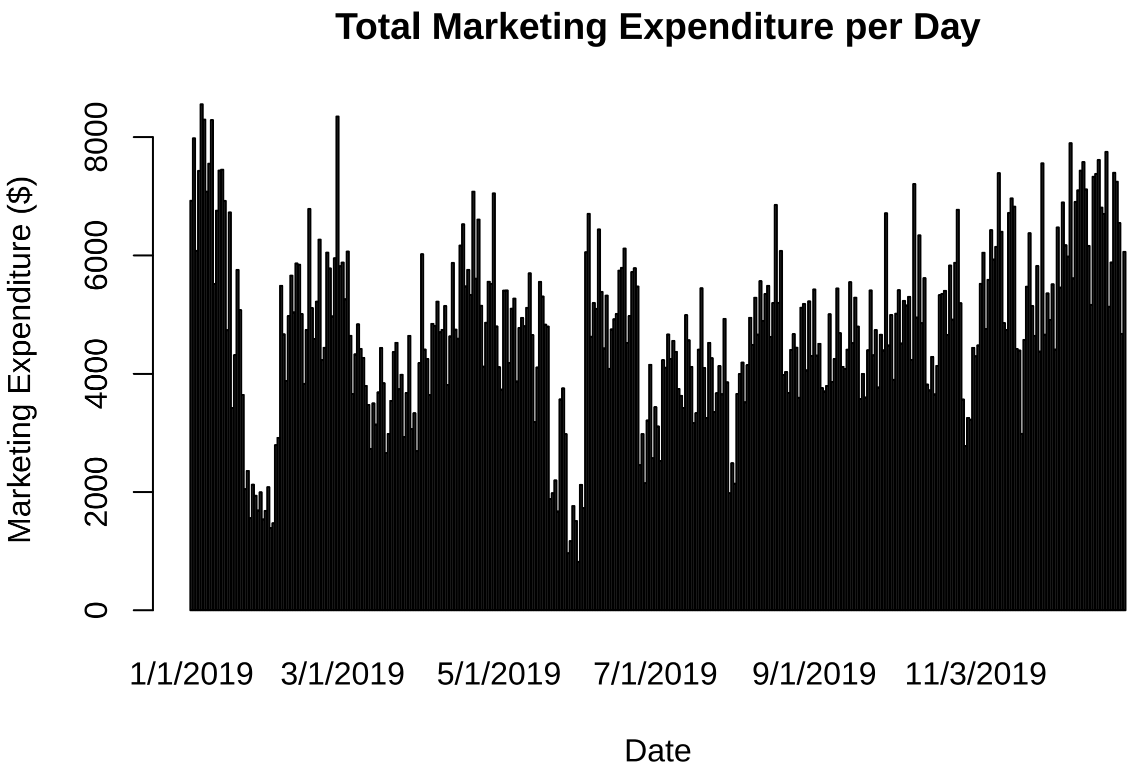
```
plot(sales$Transaction_ID, sales$Avg_Price*sales$Quantity,
     xlab="Transaction ID", ylab="Transaction Total ($)",
     main="Spending per Transaction")
```



```r
barplot(sales$Avg_Price*sales$Quantity, names.arg=sales$Transaction_Date,
        xlab="Date", ylab="Transaction Total ($)",
        main="Transaction Total per Day", ylim = c(0,1500))
```



```r
barplot(marketing$Offline_Spend+marketing$Online_Spend,
        names.arg=marketing$Date, xlab="Date", ylab="Marketing Expenditure ($)",
        main="Total Marketing Expenditure per Day")
```

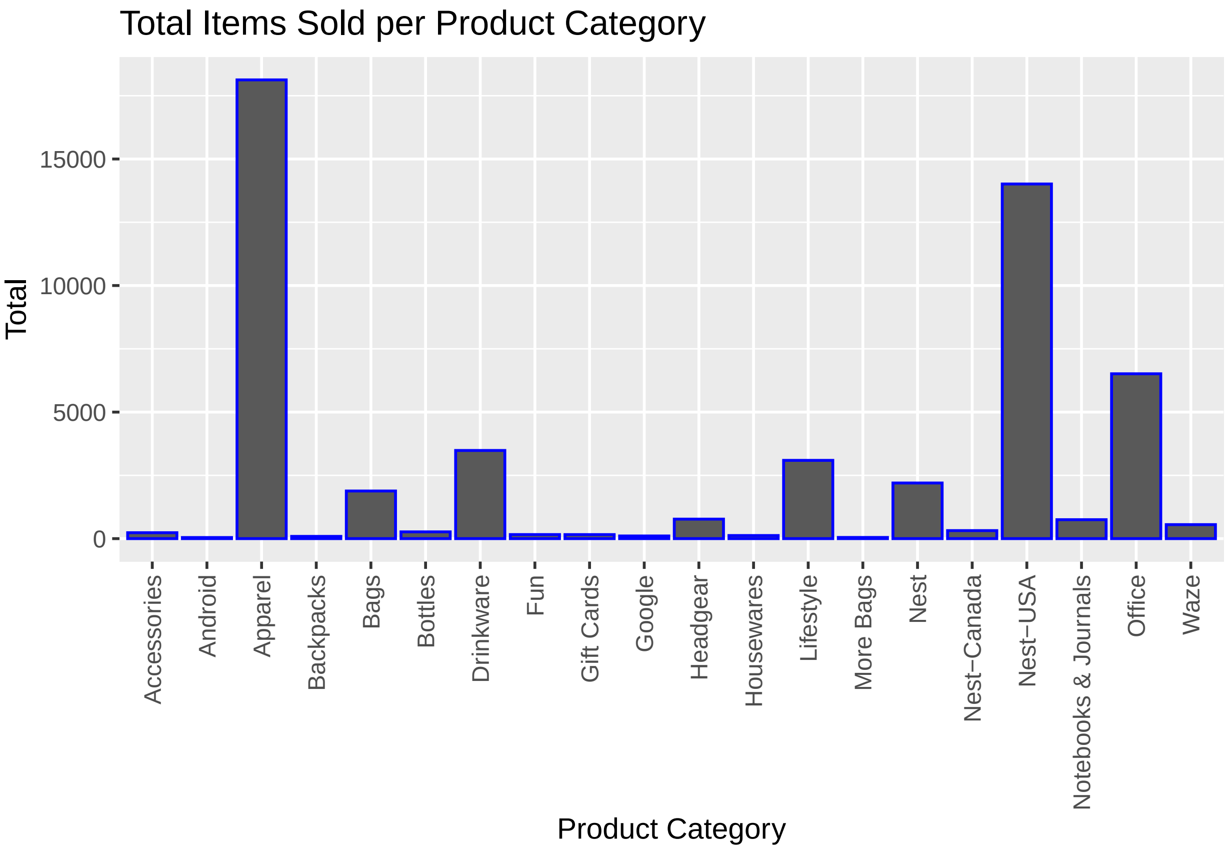


I next wanted to take a look at the sales data more closely, to understand a couple of things: what products categories are the most popular? Why are they popular and by how much? For the first step I used a histogram to categorically visualize how many items of each category was sold. Based on the graph, the category Apparel sold the most products while the category Nest-USA followed closely behind.

```r
cat_graph1 <- ggplot(sales, aes(x=Product_Category)) +
  geom_histogram(binwidth=1, color="blue", stat="count") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  xlab("Product Category") + ylab("Total") +
  ggtitle("Total Items Sold per Product Category")
```

```
## Warning in geom_histogram(binwidth = 1, color = "blue", stat = "count"):
## Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```

```r
cat_graph1
```



To dive into both Apparel and Nest-USA sales, I filtered out the sales data and created sub-data sets to explore what drives sales behind these items, choosing to explore coupon usage. Below are various different filters applied to section-off and filter through the data, once all shopping categories were graphed and the top two purchased product categories are used as filters.

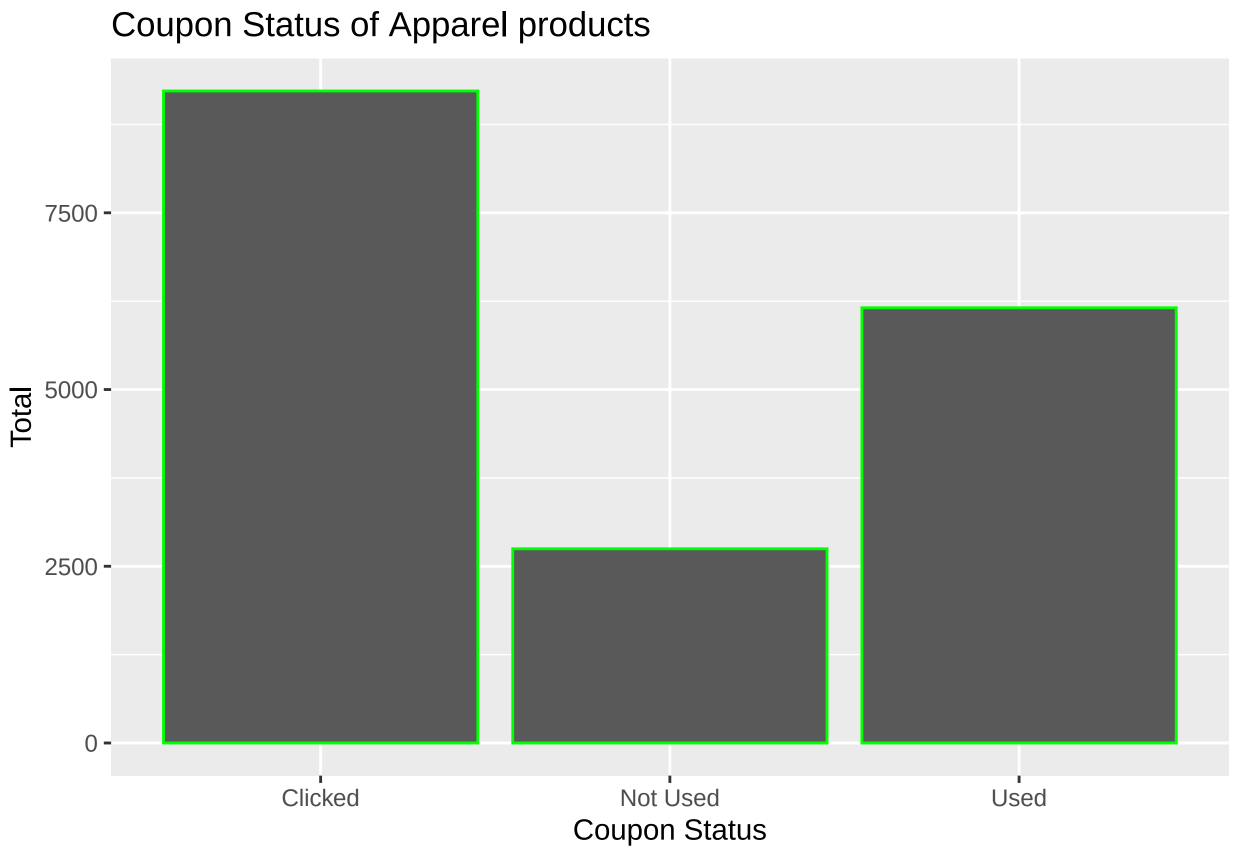
```
Apparel <- filter(
  sales, Product_Category =="Apparel")
Nest_USA <- filter(
  sales, Product_Category =="Nest-USA")
```

Using the above filters, I explore coupon usage patterns to understand if that can be a driving force behind the highest purchase product categories. Looking through the graphs, I can conclude that coupons for both Apparel and Nest-USA products were both clicked on for exploring sales and also used during the checkout. By comparison, the frequency of coupons not used was much lower than the clicked and used frequencies.

```
coup_graph1 <- ggplot(Apparel, aes(x=Coupon_Status)) +
  geom_histogram(binwidth=1, color="green", stat="count") +
  xlab("Coupon Status") + ylab("Total") +
  ggtitle("Coupon Status of Apparel products")
```

```
## Warning in geom_histogram(binwidth = 1, color = "green", stat = "count"):
## Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```

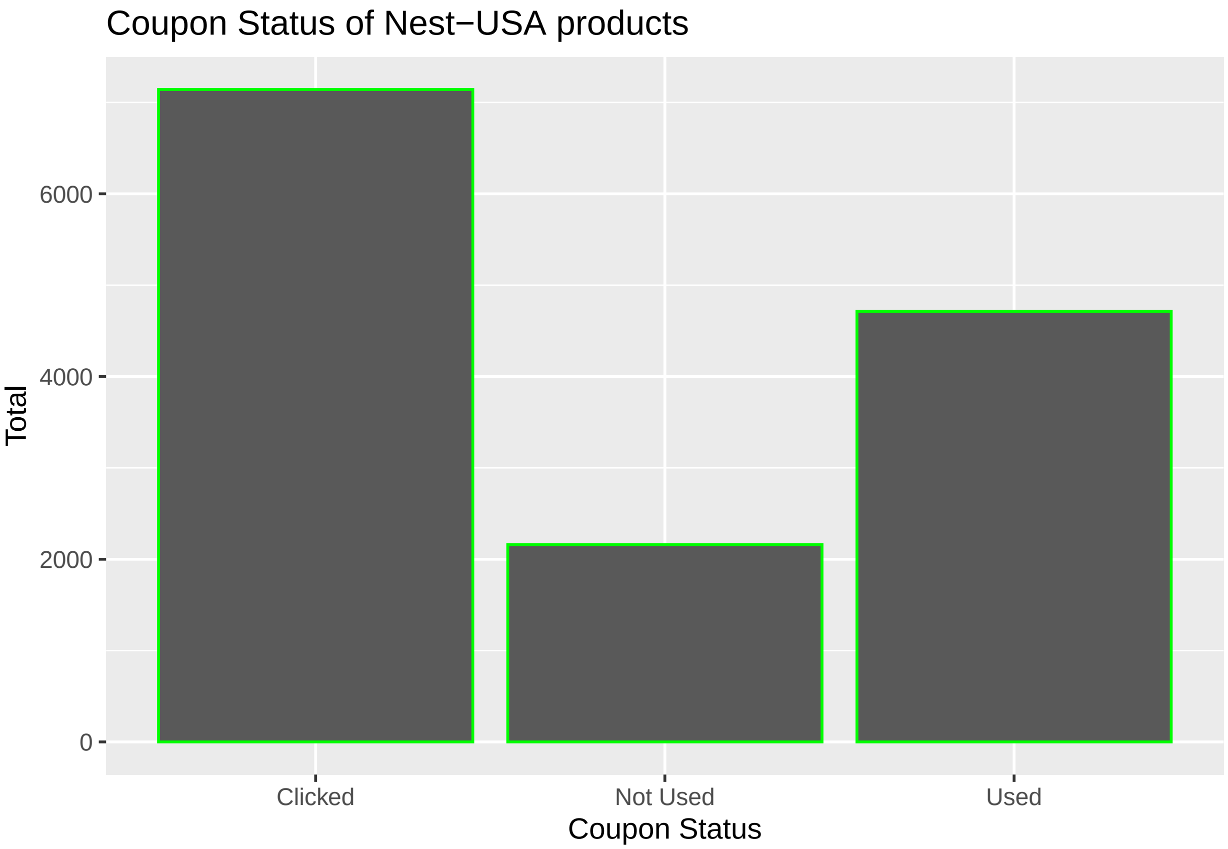
```
coup_graph1
```



```
coup_graph2 <- ggplot(Nest_USA, aes(x=Coupon_Status)) +
  geom_histogram(binwidth=1, color="green", stat="count") +
  xlab("Coupon Status") + ylab("Total") +
  ggtitle("Coupon Status of Nest-USA products")
```

```
## Warning in geom_histogram(binwidth = 1, color = "green", stat = "count"):
## Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```

```
coup_graph2
```



I further dived into Apparel and Nest-USA products to take a look at specifically which products or product types are most popular. The best way to go through this data is to visualize through purchase frequency tables for each product category in descending order, instead of graphs. For the Apparel category the top items sold are an assortment of men and women t-shirts in various colors. For the Nest-USA category the top items sold are smart thermostat and smart security camera.

```
app_table1 <- sort(table(Apparel$Product_Description), decreasing=TRUE)
head(app_table1)
```

```
##
## Google Men's 100% Cotton Short Sleeve Hero Tee Black
##                                                  595
##                                     Google Twill Cap
##                                                  546
## Google Men's 100% Cotton Short Sleeve Hero Tee White
##                                                  504
##                   Google Men's Vintage Badge Tee Black
##                                                  496
##                                       BLM Sweatshirt
##                                                  445
##           Google Men's Bike Short Sleeve Tee Charcoal
##                                                  431
```

```
nest_table2 <- sort(table(Nest_USA$Product_Description), decreasing=TRUE)
nest_table2
```

```
##
## Nest Learning Thermostat 3rd Gen-USA - Stainless Steel
##                                                    3511
##                  Nest Cam Outdoor Security Camera - USA
##                                                    3328
##                   Nest Cam Indoor Security Camera - USA
##                                                    3230
##         Nest Protect Smoke + CO White Battery Alarm-USA
##                                                    1361
##            Nest Learning Thermostat 3rd Gen-USA - White
##                                                    1089
##           Nest Protect Smoke + CO White Wired Alarm-USA
##                                                    1065
##           Nest Learning Thermostat 3rd Gen-USA - Copper
##                                                     393
##           Nest Protect Smoke + CO Black Wired Alarm-USA
##                                                      19
##         Nest Protect Smoke + CO Black Battery Alarm-USA
##                                                      17
```

Questions for Future Steps

One thing I do not know how to do right now and would like to learn, is to successfully code if/else decision parameters to expand the columns in my data frames. I would like to add more columns to concatenate my data using variables in other columns as contingencies. This would be a simplified way of testing my variables hypothesis, where I can input thresholds of coupon usage, marketing expenditures, customer identifiers, etc. and get a potential predicted output of how much an individual customer would've spent and in which category. This can be compared across all 53000 rows of data in the sales data frame.

Plan to Incorporate Machine Leanring Techniques

Currently I am unsure if there's a possibility to add machine learning concepts into understanding a potential model at this time. Ideally machine learning would come in at a later stage. For example, the first step is to use my data to propose a regression model that is predictive. From there machine learning concepts can be incorporated to self-correct the model as the data set turns into a rolling data set, forever changing as addition points of consumer data are added as the years pass by.

Narrative

Consumerism is an economic theory that consumer spending is the key to an individual's well-being and one of the greatest factors in driving a country's economic growth. The concept of consumerism outlines the increase of the consumption of goods is the basis for a thriving economy. Consumerist societies measure their success through gross domestic product (GDP). Consumerism is a prominent part of any country's economic drive and a strong component of its population's routine life, where spending effectively increases GDP. Capitalist economies depend on the consumption of products and encourage populations to purchase beyond their basic needs to promote a thriving economy. Basically, the more products that are purchased and consumed, the more prosperity is brought about.

In my research, I explored the retail aspect of consumerism, rather than offered services. Large companies and manufacturers provide mass-produced products to customers, where such entities demonstrate how improved products can create value for consumers across all needs and requirements. Therefore, it's critical to know the direction of the market trend, as this can be advantageous to increase sales and guide product releases. Within a trending market's scope, a retailer must also be able to mitigate. the waves of profit by introducing discounts as price optimization or releasing new products.

Problem Statement You Addressed

The problem is that to create successful products, a profit-oriented business first needs to understand customer spending strategies, categorical interests, marketing campaigns, discounts, sale seasons, etc. Understanding what drives a customer to the store to purchase products and maximize profits is critical for any business to understand, and to be able to predict sales and forecast profit based on trends and consumer factors is core to a retailer's continued success. Without such insights, predictive profits will fail to yield any usable information that a retailer can use when trying to understand which products should be released, when they should be realized, and all the different avenues consumers will use to approach and purchase the products. I also had formed my own set of research questions that guided my analysis towards my conclusions:

1. Which time of the year are consumer sales at their highest?
2. Which time of the year are consumer sales at their lowest, and how is this slump mitigated?
3. Are there product categories that consumers tend to purchase simultaneously (i.e., electronics and traveling gear)?
4. Do coupons tend to help with increasing the demand of products?
5. What are the highest sold product categories that a retailer can identify for launching a new product?
6. Are there multiple driving variables that a consumer considers with their purchasing decisions?

How You Addressed the Problem Statement

To address the problem statement I identified, I created an outline of how to approach the idea of proposing a useful model. I gathered relevant data, identified key variables and contributors, and use a multitude of graphs and analysis to arrive at conclusions through identified patterns.

I gathered relevant data sets regarding consumer sales data across any category of retail. Overall the selected data sets are based on a small excerpt of consumer transaction data across the country, from 1/1/2019 to 12/31/2019 and optimized for analytical use. In this data set, there are nearly 1500 total customers with nearly 53000 line items of transaction information. The information spanned across a lengthy time frame to identify seasonal pattern changes, and the variables I compared all fell within consumer spending. Through these available categories I looked to determine which factors are critical in forecasting consumer behavior. Through exploring numerous variables such as market trends, customer product preferences, incentives, and marketing costs, I was able to approach the idea of a sales forecasting model through multivariate analysis. The potential resulting model will be comprised of independent variables that can each be manipulated for an enterprise to accurately forecast as necessary.

After importing and simplifying the data sets, I employed data frames and various visuals as useful forms analysis formats to reorganize and visualize transposed data as analysis. Understanding relationships between variables and which factors correlate with each other is important to express through a multitude of visuals and graphs. Even though my approach is more accurate than simple historical-based predictions due to the incorporation of a broader range of factors, I kept in mind statistical limitations and the possible emergence of only a partial solution.

Analysis

The correlation and causation between variables were not evident in the beginning, as the variable information was laid out in randomized order. Understanding which variables can compared, which variables cannot, and which variables can be combined is the first step in planning my analysis. The data was analyzed in the order of identifying the customer and sales geography. Followed by understand spending patterns throughout the year, where all the quantities and average prices were combined and graphed to demonstrate any possible trends. Trends suggest any seasonal relevance that contributes to identified customers' spending goals. Established pricing summaries and their distribution across the product ranges offered will also enable for targeted product releases.

initially, I used the histograms to visualize customer genders and their cities of residence. According to the Customer Gender graph, nearly 40% more customers are female compared to males. According to the Customer Location graph, the most popular cities for consumerism are California and Chicago, followed closely by New York. The next parameter I identified was the range of spending and a potential spending pattern. I used a basic plot combining arbitrary transaction IDs against the total money spent, and to obtain the total money spent I combined the columns of total items in each transaction with the column of average price of each item in each transaction. Visually I determined an upper spending cap $1500 and adjusted the y-range to observe potential spending patterns. I next created a bar plot to visualize spending on each day of the year from January 1st all the way through December 31st, and another bar plot that combines online and offline marketing trends. According to the graphs, there is an observable trend that shows the spend on marketing is higher during the months of November, December, and January. This trend also compares to an increase in consumer spending along the months of November and December. The next step was to identify popular categories. Based on the histogram created, the category Apparel sold the most products while the category Nest-USA followed closely behind. This suggests both clothing home-security products are some of the highest categories of product sold. Using the coupon data set to identify a possible explanation, I used additional histograms within each product category to further identify spending patterns. According to the graphs, in both categories consumer has clicked on and used coupons at a higher rate compared to other product categories. Finally, generated frequency tables revealed the most popular products within the identified categories. The highest sold products are men and women t-shirts, followed closely by home wi-fi items such as cameras and thermostats. Overall, there are several conclusions that I interpreted about consumer spending habits and target products as a result of the data analysis above:

1. The target customer demography is a female that lives in an affluent state such as California or New York.
2. The popular times for retail consumer spending is around the holidays from November to January of the following year.
3. On average in a single transaction during peak season, a consumer can spend up to $800.
4. Coupons are a large part of a consumer's decision on when to purchase a product and in which product category.
5. The monetary resources spent on marketing translates to increased product sales.
6. The most popular item category sold throughout the year is apparel, specifically tops such as t-shirts and sweaters.

For a model, I propose a time series regression model for predicting a future response based on the response history and relevant predictors. This assumes that the dependent variable has a linear relationship with its independent various, which I believe to be case based on my conclusions. Time series allows for the amendment of new data, as transaction data from each passing year would only make the model more accurate. The independent variables should be: time of year, advertising resources spent, target customer gender, target customer location, active sales or coupons, sales tax in region, and product category. The dependent variable should either be total money spent for a linear relationship or specific product for a categorical predictor. I am not able to successfully code a true model as there are variables made up of string values and not integers.

Implications

There are a few implications the data provides when trying to understand consumerism. These can be viewed as an ethical discussion or an opportunity to promote awareness for informed purchasing decisions. Extracting and utilizing consumer information to influence purchasing decisions imply that retail corporations have a large direct influence on the success of a capitalistic society. Usually, the influential marketing tactics tied with spending or financing incentives leads to consumers overspending beyond their means, year after year. Understanding trend and targeted marketing as observed in my research, can drive individuals to become more aware to limit their consuming habits to needs, rather than wants. Unearthing the social implications of consumerism is an entirely other subject of behavioral psychology. Where individuals may share a preoccupation with hoarding retail products that do not serve neither a need nor want, but with an aim to increase social perception. All of this can be further amplified with corporations' better understand their target audience year after year, refining their models and forecasts.

Limitations

There are quite a few limitations to the data sets, including statistical limitations and my own personal limitations. The data selected is only a fraction of the true big data available on the subject. There are unexplored platforms, such as physical and online retail outlets. Products are well are minimally represented, as consumerism reaches far beyond just apparel. I only focused on one year's worth of data in a couple of select cities, ignoring decades of global research available. For modeling, I do not know is how to run simulations through a multiple time series regression model repeatedly, while assessing best-of-fit tests to ensure a more accurate model fit over time. And some of my variables return characters rather than integers, further complicating a simple linear fit. Employing machine learning concepts into understanding the potential model would also help expand on my current limitations. For example, the first step is to use my data to propose a regression model that is predictive. From there machine learning concepts can be incorporated to self-correct the model as the data set turns into a rolling data set, constantly changing as addition points of consumer data are added. Further available data can also be appended through data mining capabilities.

Concluding Remarks

Overall, I've been able to understand the overarching theory of consumerism. And how available patterns and data can be used to feed into the theory, in turn driving the economy and result societal impact. This practical translation from theory to practice is a great example of how data science can be used to propose a solution and directly drive the theory into practice. Through unearthing spending patterns and exploring the balance between retailers and consumers, I believe both sides of the spectrum can benefit. Where retailers and business can provide an ideal shopping experience available to consumers, and consumers can actively recognize targeted experience to make better personal spending decisions.