Assignment 1

Alex Khaykin

2024-10-19

library(tidyverse)  
library(lubridate)  
library(ggplot2)  
library(corrplot)

# INTRO

The landscape of modern business increasingly relies on data-driven decision-making, and effective analysis of sales data can significantly enhance operational efficiency and customer satisfaction. For my project, I focus on understanding the relationship between two critical factors influencing Fulfilment\_Speed: Order.Priority and and other influeincing factors such as geography and the companies Total.Revenue. By employing a linear regression model, I aim to uncover patterns within my chosen dataset composed of 1,000 sales transactions, examining how these variables interact and impact overall sales outcomes.

Order priority typically reflects the urgency and importance assigned to each sale, potentially impacting the fulfillment process and customer experience. Fulfillment speed, on the other hand, denotes the time taken to process and deliver an order. In a competitive market, optimizing both order priority and fulfillment speed is essential for meeting customer expectations and driving sales growth. Through linear regression analysis, we will explore whether order priority status can predict fulfillment speed, providing valuable insights for stakeholders looking to enhance their operational strategies.

This project will articulate the statistical relationship between order priority and fulfillment speed and provide actionable recommendations for improving order management and fulfillment processes. By leveraging this analysis, businesses can make informed decisions to enhance their performance in a fast-paced market environment.

# DATA

Reading in the data downloaded from the site, and converting it to a data frame from where I will be unitizing R to run summary statistics to learn about the data and the information contained within.

sales\_1000 <- read.csv("C:\\Users\\akhay\\OneDrive\\Documents\\DATA\_SCIENCE\\DATA\_622\\Assignments\\Assignment\_1\\1000 Sales Records.csv")

## EDA

The dataset comprises 1,000 entries with various attributes such as Region, Country, Item.Type, Sales.Channel, Order.Priority, Order.Date, and Order.ID, all stored as character types. The Order.ID values range significantly, indicating diverse transactions. Numeric columns like Units.Sold, Unit.Price, Unit.Cost, Total.Revenue, Total.Cost, and Total.Profit show substantial variability. For instance, Units.Sold ranges from 13 to 9,998, with a mean of 5,054, while Total.Revenue spans from 2,043 to 6,617,210, reflecting the wide range of sales activities captured in the dataset.

### Summary

This sections give the names of the colums(variable), and the basic statistics of the dataset.There seems to be no missing data in this dataset.

summary(sales\_1000)

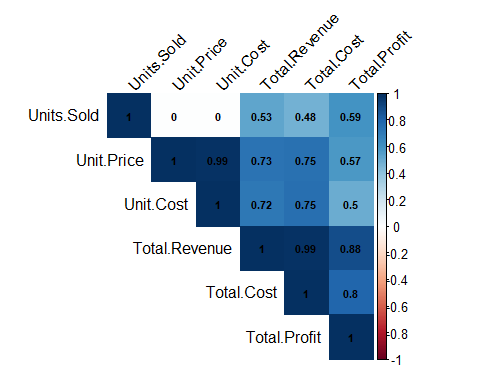
## Region Country Item.Type Sales.Channel   
## Length:1000 Length:1000 Length:1000 Length:1000   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## Order.Priority Order.Date Order.ID Ship.Date   
## Length:1000 Length:1000 Min. :102928006 Length:1000   
## Class :character Class :character 1st Qu.:328074026 Class :character   
## Mode :character Mode :character Median :556609714 Mode :character   
## Mean :549681325   
## 3rd Qu.:769694483   
## Max. :995529830   
## Units.Sold Unit.Price Unit.Cost Total.Revenue   
## Min. : 13 Min. : 9.33 Min. : 6.92 Min. : 2043   
## 1st Qu.:2420 1st Qu.: 81.73 1st Qu.: 56.67 1st Qu.: 281192   
## Median :5184 Median :154.06 Median : 97.44 Median : 754939   
## Mean :5054 Mean :262.11 Mean :184.97 Mean :1327322   
## 3rd Qu.:7537 3rd Qu.:421.89 3rd Qu.:263.33 3rd Qu.:1733503   
## Max. :9998 Max. :668.27 Max. :524.96 Max. :6617210   
## Total.Cost Total.Profit   
## Min. : 1417 Min. : 532.6   
## 1st Qu.: 164932 1st Qu.: 98376.1   
## Median : 464726 Median : 277226.0   
## Mean : 936119 Mean : 391202.6   
## 3rd Qu.:1141750 3rd Qu.: 548456.8   
## Max. :5204978 Max. :1726181.4

To create a sub-set of all numeric data from the original dataset prior to creating the corroplot to test for overrelaction.

numeric\_columns <- sales\_1000 %>%   
 select(Units.Sold, Unit.Price, Unit.Cost, Total.Revenue, Total.Cost, Total.Profit)  
correlation\_matrix <- cor(numeric\_columns, use = "complete.obs")

Creating a corroplot. As anticipated there is a strong correlation between units sold and total revenue, as well as units price and and total revenue. However, quite opposite when it comes to the relationship between units price and units sold.

corrplot(correlation\_matrix, method = "color", type = "upper",   
 tl.col = "black", tl.srt = 45,   
 addCoef.col = "black", number.cex = 0.7)



The dataset contains 1,000 sales records, with 14 columns. However for the purpose of the question I am asking, I will be treating these as company records and dropping the following columns: Country, Order Id. Additional y I will be making a new primary predictor variable Fullfilment Speed, which is the lag between Order Date and Ship Date, after which I will drop the two date columns entirely. I am treating this as the average fulfillment speed for each of the hypothetical companies and their records contained in the dataset.

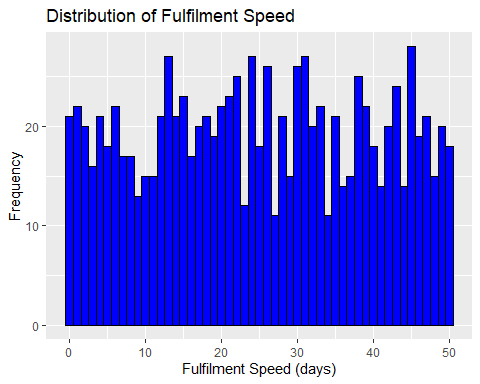
sales\_1000 <- sales\_1000 %>%   
 select(-Country, -Order.ID) %>%   
 mutate(Order.Date=as.Date(Order.Date, format="%m/%d/%Y"),   
 Ship.Date=as.Date(Ship.Date, format= "%m/%d/%Y"),  
 Fulfilment\_Speed =as.numeric(Ship.Date - Order.Date)) %>%   
 select(-Order.Date, -Ship.Date)

## VISUALIZATIONS

### 1. Histogram of Fulfilment Speed

This chart shows the distribution of fulfillment speeds. With a few exceptions, there is a approximately 20 occurrences in the dataset.

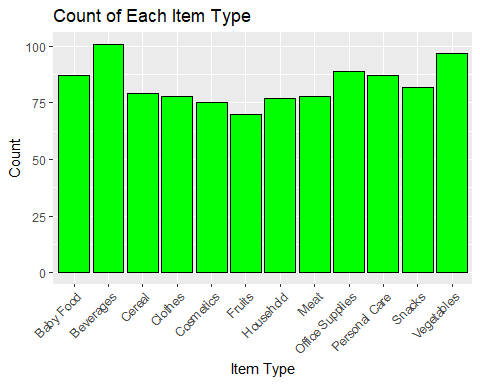
ggplot(sales\_1000, aes(x = Fulfilment\_Speed)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "Distribution of Fulfilment Speed", x = "Fulfilment Speed (days)", y = "Frequency")



### 2. Bar Plot of Item Types

This chart shows the count of each item type.

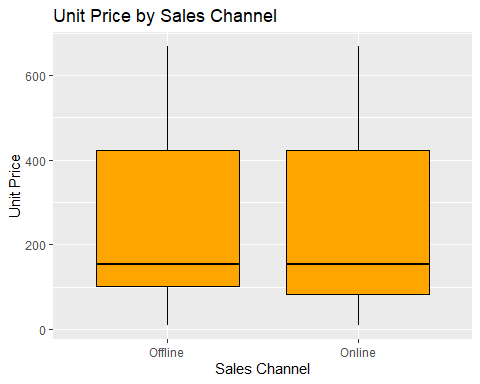
ggplot(sales\_1000, aes(x = Item.Type)) +  
 geom\_bar(fill = "green", color = "black") +  
 labs(title = "Count of Each Item Type", x = "Item Type", y = "Count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



### 3. Box Plot of Unit Price by Sales Channel

This chart compares the distribution of unit prices across different sales channels.

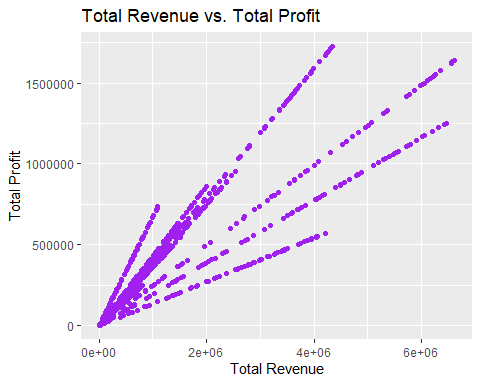
ggplot(sales\_1000, aes(x = Sales.Channel, y = Unit.Price)) +  
 geom\_boxplot(fill = "orange", color = "black") +  
 labs(title = "Unit Price by Sales Channel", x = "Sales Channel", y = "Unit Price")



### 4. Scatter Plot of Total Revenue vs. Total Profit

This chart shows the relationship between total revenue and total profit.

ggplot(sales\_1000, aes(x = Total.Revenue, y = Total.Profit)) +  
 geom\_point(color = "purple") +  
 labs(title = "Total Revenue vs. Total Profit", x = "Total Revenue", y = "Total Profit")



# MODEL

model\_sales <- lm(Fulfilment\_Speed ~ Order.Priority + Region + Total.Revenue + Sales.Channel, data = sales\_1000)

## Model Summary

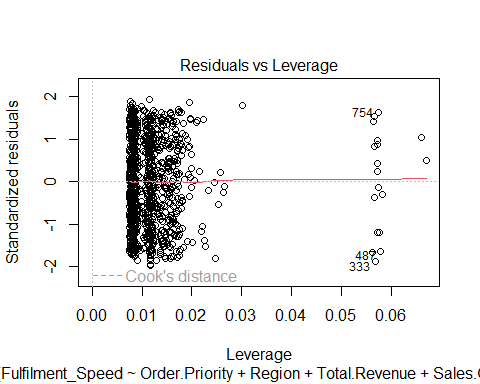
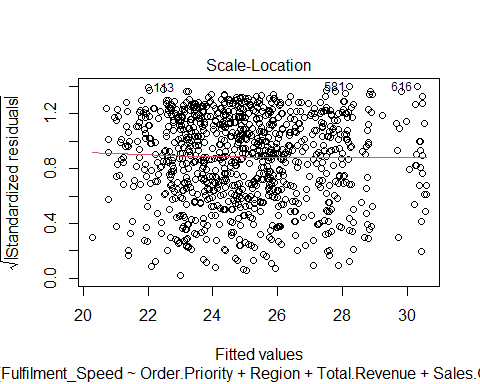
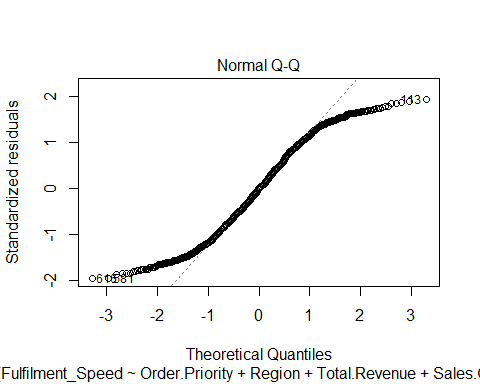
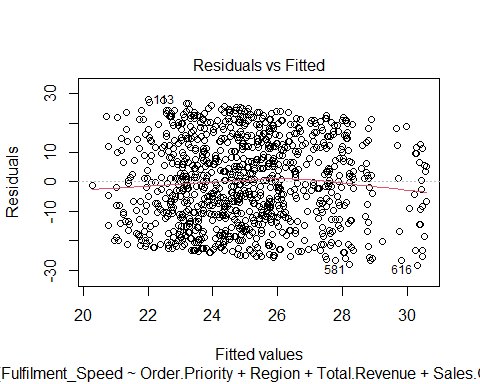
summary(model\_sales)

##   
## Call:  
## lm(formula = Fulfilment\_Speed ~ Order.Priority + Region + Total.Revenue +   
## Sales.Channel, data = sales\_1000)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -28.3071 -11.7424 0.0249 12.3163 27.9851   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.606e+01 1.607e+00 16.214 < 2e-16  
## Order.PriorityH 2.947e+00 1.322e+00 2.229 0.02605  
## Order.PriorityL 5.043e-01 1.273e+00 0.396 0.69200  
## Order.PriorityM 6.961e-01 1.298e+00 0.536 0.59197  
## RegionAustralia and Oceania -2.197e+00 2.062e+00 -1.065 0.28699  
## RegionCentral America and the Caribbean -1.367e-01 1.926e+00 -0.071 0.94346  
## RegionEurope -2.801e+00 1.533e+00 -1.827 0.06797  
## RegionMiddle East and North Africa -4.620e+00 1.764e+00 -2.620 0.00894  
## RegionNorth America -6.298e-01 3.563e+00 -0.177 0.85972  
## RegionSub-Saharan Africa -3.253e+00 1.538e+00 -2.116 0.03464  
## Total.Revenue -2.747e-07 3.097e-07 -0.887 0.37538  
## Sales.ChannelOnline 1.607e+00 9.250e-01 1.737 0.08263  
##   
## (Intercept) \*\*\*  
## Order.PriorityH \*   
## Order.PriorityL   
## Order.PriorityM   
## RegionAustralia and Oceania   
## RegionCentral America and the Caribbean   
## RegionEurope .   
## RegionMiddle East and North Africa \*\*   
## RegionNorth America   
## RegionSub-Saharan Africa \*   
## Total.Revenue   
## Sales.ChannelOnline .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.54 on 988 degrees of freedom  
## Multiple R-squared: 0.02002, Adjusted R-squared: 0.009111   
## F-statistic: 1.835 on 11 and 988 DF, p-value: 0.04436

## Model Diagnostic Plots

Model diagnostic plots show a robust model, with no violations of assumptions, although there are so possible outliers by Cooks Distance.

plot(model\_sales)



# CONCLUSION

I fitted the model to test whether fulfillment speed was affected by an orders priority, sales channel, geography, or the companies total revenue. I found that only high-priority orders have significantly faster fulfillment speed. Which suggests, that customers paying for medium priority are not receiving value for the shipping speed purchased while shopping. The region of delivery for the order was also significantly slower for the middle east, Northers Africa and Sub-Saharan Africa. Which points to that clients in these potential areas of sales need to be advised of potential slower fulfillment speeds. This is regardless of whether a customer requests and pays for a higher order priority at time of sale. Additionally, there was a marginally significant difference between sales channel with online orders being fulfilled nearly two day faster on average, after controlling for all other factors. It is interesting to note that, there is no significant affect of total revenue on fulfillment speed, which suggests that larger companies are not necessarily able to fulfill orders faster that smaller companies. This is contrary to what I would expect from companies like Amazon, which have build themselves around customer expectation of faster fulfillment speeds. This is likely because this dataset is fabricated and is not an accurate reflection of the true relationship between companies actual sales and the fulfillment speed that they are able to deliver. in Conclusion these simulated sales from 1000 orders may have been drawn from actual orders, hence the difference in fulfillment speed in geography and order priority. With the larger dataset, these relationships if true should be found again. If they they would not be found again, then it would suggest that my finding are the results of the small sample size of the smaller simulated dataset, and thus random chance and nothing more.