Final Project

Alex Khaykin

2024-12-17

library(tidyverse)  
library(lubridate)  
library(ggplot2)  
library(corrplot)  
library(caret)  
library(randomForest)  
library(rpart)  
library(rpart.plot)

# INTRODUCTION

This (dataset)[“<https://www.kaggle.com/datasets/jehanzaibbhatti/sales-data?resource=download>”] is of a particular interest to me because I am a lond distance outdoor cyclist. The dataset contains individual sales transactions for a single cycling enthusiast products provider. They provide various products to customers in multiple countries and sell both bicycles, accessories, and clothing. The landscape of modern business increasingly relies on data-driven decision-making, and effective analysis of sales data that can significantly enhance operational efficiency and customer satisfaction. For my project, I focus on trying to identify the critical factors influencing decline of Revenue over time. I employ both a linear regression model and a random forest model.

Revenue typically reflects the total sales from customer transactions. Revenue and sales are crucial for any company due to operational demands as well as profitability. This project will articulate the statistical relationship to explain why this companies’ revenue may have declined over time and I will try to provide actionable recommendations.

# DATA CLEANING

sales\_data <- read.csv("C:\\Users\\akhay\\OneDrive\\Documents\\DATA\_SCIENCE\\DATA\_622\\Assignments\\Final\_Project\\sales\_data.csv") %>%   
 mutate(Date=as.Date(Date))

Although this dataset requires minimal cleaning, I did choose to make a version of the dataset for modeling that removes age group due to redundancy, month because of I turned month into a number, product because too much variation, and state for the same reason. I also dropped date because I chose to use year and month as time variables.

sales\_data2 <- sales\_data %>%   
 select(-Age\_Group, -Month, -Product, -State) %>%   
 mutate(Month=month(Date, label=FALSE)) %>%   
 select(-Date)

## 

## EDA

The dataset comprises 113,036 entries with various attributes such as Country, Product, Order\_Quantity, Date, and Unit\_Cost. As well as customer demographics such as Age and Customer\_Gender. The date ranges for the date are from 2011 through 2016, and the average age of customer is 35, it is also worth mentioning that the average quantify of products ordered is 12.

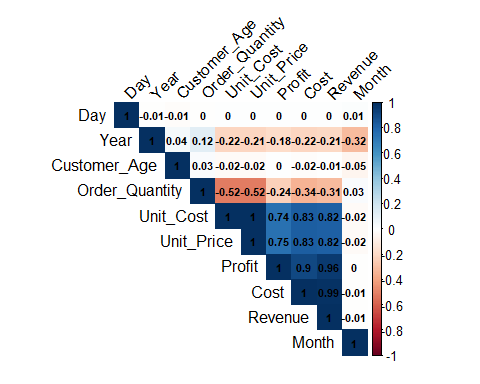
summary(sales\_data2)

## Day Year Customer\_Age Customer\_Gender   
## Min. : 1.00 Min. :2011 Min. :17.00 Length:113036   
## 1st Qu.: 8.00 1st Qu.:2013 1st Qu.:28.00 Class :character   
## Median :16.00 Median :2014 Median :35.00 Mode :character   
## Mean :15.67 Mean :2014 Mean :35.92   
## 3rd Qu.:23.00 3rd Qu.:2016 3rd Qu.:43.00   
## Max. :31.00 Max. :2016 Max. :87.00   
## Country Product\_Category Sub\_Category Order\_Quantity  
## Length:113036 Length:113036 Length:113036 Min. : 1.0   
## Class :character Class :character Class :character 1st Qu.: 2.0   
## Mode :character Mode :character Mode :character Median :10.0   
## Mean :11.9   
## 3rd Qu.:20.0   
## Max. :32.0   
## Unit\_Cost Unit\_Price Profit Cost   
## Min. : 1.0 Min. : 2.0 Min. : -30.0 Min. : 1.0   
## 1st Qu.: 2.0 1st Qu.: 5.0 1st Qu.: 29.0 1st Qu.: 28.0   
## Median : 9.0 Median : 24.0 Median : 101.0 Median : 108.0   
## Mean : 267.3 Mean : 452.9 Mean : 285.1 Mean : 469.3   
## 3rd Qu.: 42.0 3rd Qu.: 70.0 3rd Qu.: 358.0 3rd Qu.: 432.0   
## Max. :2171.0 Max. :3578.0 Max. :15096.0 Max. :42978.0   
## Revenue Month   
## Min. : 2.0 Min. : 1.000   
## 1st Qu.: 63.0 1st Qu.: 4.000   
## Median : 223.0 Median : 6.000   
## Mean : 754.4 Mean : 6.453   
## 3rd Qu.: 800.0 3rd Qu.:10.000   
## Max. :58074.0 Max. :12.000

numeric\_columns <- sales\_data2 %>%   
 select(where(is.numeric))  
correlation\_matrix <- cor(numeric\_columns, use = "complete.obs")

Creating a correlation plot and as anticipated there is a strong correlation between units’ cost, unit price, profit, cost and revenue. However, it is quite the opposite when it comes to the relationship between unit price, unit cost and order quantity. There is a slight negative correlation with revenue and year, suggesting there has been a possible decline in sales over time.

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



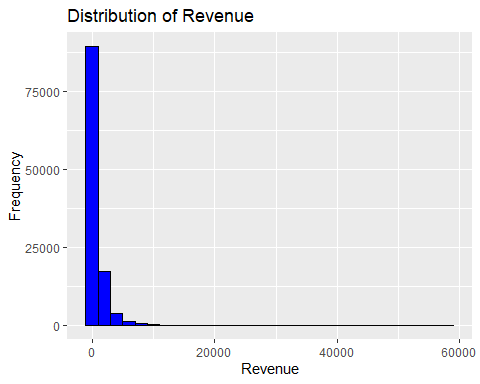
## VISUALIZATIONS

### Histogram of Fulfilment Speed

This chart shows the distribution of Revenue that most sales generate revenue under $10k, with few that generate revenue upwards of $60k. The average revenue per sale is $754.

ggplot(sales\_data2, aes(x = Revenue)) +  
 geom\_histogram(fill = "blue", color = "black") +  
 labs(title = "Distribution of Revenue", x = "Revenue", y = "Frequency")

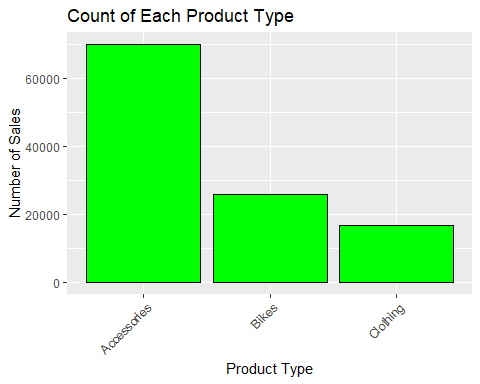
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



### Bar Plot of Item Types

This chart shows the count of each product category sold. Biking accessories have the highest number of sales, followed by bikes, and then clothing.

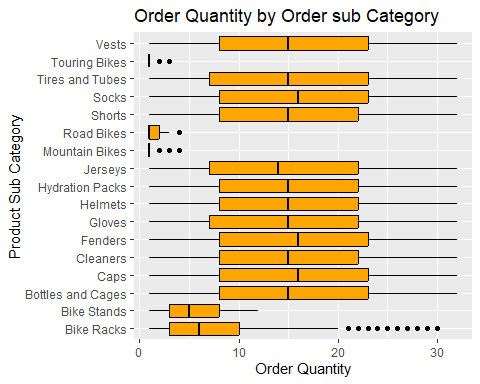
ggplot(sales\_data2, aes(x = Product\_Category)) +  
 geom\_bar(fill = "green", color = "black") +  
 labs(title = "Count of Each Product Type", x = "Product Type", y = "Number of Sales") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



### Box Plot of Unit Price by Sales Channel

This chart compares the product sub-category by order quantity. Most order quantities are similar in their distributions, with the exceptions of largest items such as bikes and bike racks.

ggplot(sales\_data, aes(x = Sub\_Category, y = Order\_Quantity)) +  
 geom\_boxplot(fill = "orange", color = "black") +  
 labs(title = "Order Quantity by Order sub Category", x = "Product Sub Category", y = "Order Quantity") +  
 coord\_flip()

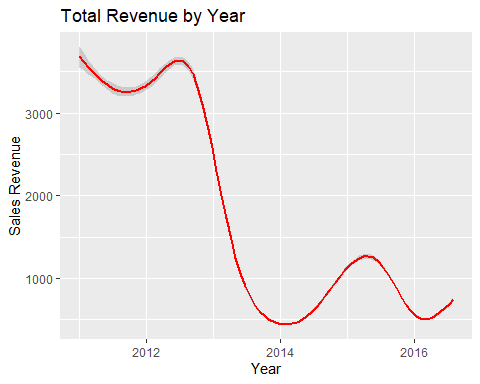


### Line Plot of sales revenue by year.

The plot confirms my initial observation that revenues looked like they have declined over time. The highest drop have was in 2013.

ggplot(sales\_data, aes(x = Date, y = Revenue)) +  
 geom\_smooth(color="red") +  
 labs(title = "Total Revenue by Year", x = "Year", y = "Sales Revenue")

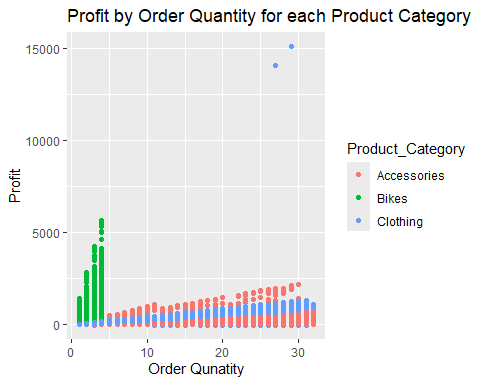
### ## `geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'



### Scatter Plot of Total Revenue vs. Total Profi

This chart shows the relationship between sales profit and order quantity, colored by the product category. This shows that the company makes most profit per sale for the single purchase items such as bikes, but sells more accessories and clothing, which generate lower profit although generating higher volume.

ggplot(sales\_data2, aes(x = Order\_Quantity, y = Profit, fill = Product\_Category, color=Product\_Category)) +  
 geom\_point() +  
 labs(title = "Profit by Order Quantity for each Product Category", y = "Profit", x = "Order Qunatity")



# MODEL

## Regression model

I first chose to run a regression model because Revenue is a continuous variable. My goal is to try to explain what factors have led to the decline of revenue over time.

model1<- lm(Revenue ~.,data=sales\_data2)

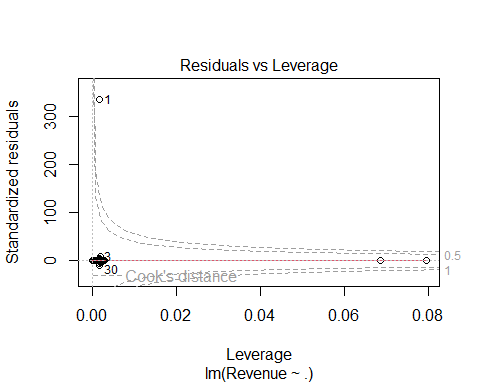
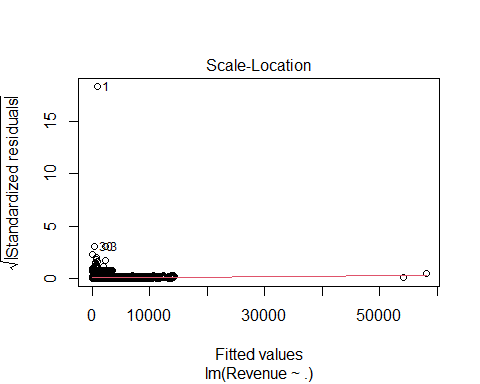
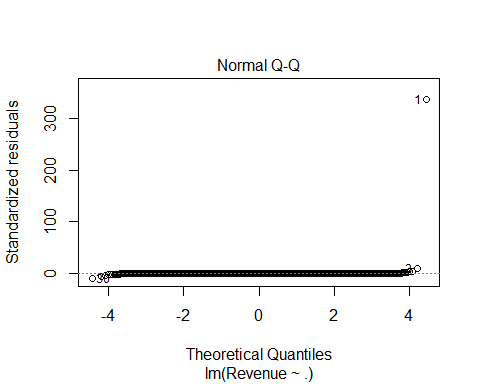
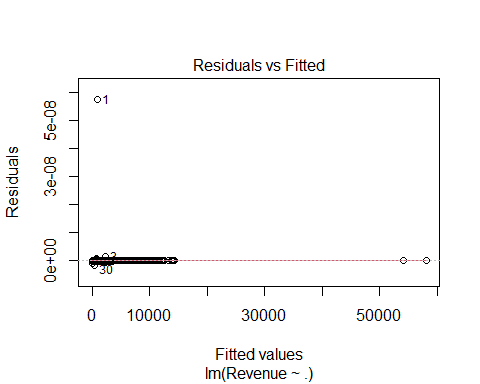
### Model Summary

summary(model1)

##   
## Call:  
## lm(formula = Revenue ~ ., data = sales\_data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.590e-09 -1.000e-12 0.000e+00 1.000e-12 5.759e-08   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.746e-09 8.974e-10 9.746e+00 < 2e-16 \*\*\*  
## Day 8.999e-14 5.818e-14 1.547e+00 0.12191   
## Year -4.202e-12 4.453e-13 -9.436e+00 < 2e-16 \*\*\*  
## Customer\_Age -6.790e-14 4.675e-14 -1.452e+00 0.14640   
## Customer\_GenderM 7.233e-13 1.023e-12 7.070e-01 0.47971   
## CountryCanada 2.478e-12 1.881e-12 1.317e+00 0.18771   
## CountryFrance -1.952e-12 1.984e-12 -9.840e-01 0.32512   
## CountryGermany -2.937e-12 1.991e-12 -1.475e+00 0.14011   
## CountryUnited Kingdom -4.103e-12 1.890e-12 -2.171e+00 0.02992 \*   
## CountryUnited States -3.799e-12 1.443e-12 -2.632e+00 0.00849 \*\*   
## Product\_CategoryBikes -1.090e-10 8.330e-12 -1.308e+01 < 2e-16 \*\*\*  
## Product\_CategoryClothing -1.212e-10 8.990e-12 -1.348e+01 < 2e-16 \*\*\*  
## Sub\_CategoryBike Stands -1.053e-10 1.071e-11 -9.832e+00 < 2e-16 \*\*\*  
## Sub\_CategoryBottles and Cages -1.029e-10 7.342e-12 -1.401e+01 < 2e-16 \*\*\*  
## Sub\_CategoryCaps 1.735e-11 6.256e-12 2.773e+00 0.00555 \*\*   
## Sub\_CategoryCleaners -9.607e-11 8.264e-12 -1.162e+01 < 2e-16 \*\*\*  
## Sub\_CategoryFenders -1.012e-10 7.654e-12 -1.321e+01 < 2e-16 \*\*\*  
## Sub\_CategoryGloves 2.535e-11 6.493e-12 3.905e+00 9.44e-05 \*\*\*  
## Sub\_CategoryHelmets -1.034e-10 7.296e-12 -1.417e+01 < 2e-16 \*\*\*  
## Sub\_CategoryHydration Packs -1.041e-10 8.512e-12 -1.223e+01 < 2e-16 \*\*\*  
## Sub\_CategoryJerseys 8.118e-12 6.220e-12 1.305e+00 0.19186   
## Sub\_CategoryMountain Bikes -1.171e-11 6.206e-12 -1.887e+00 0.05922 .   
## Sub\_CategoryRoad Bikes 3.969e-13 3.244e-12 1.220e-01 0.90264   
## Sub\_CategoryShorts 7.370e-12 6.863e-12 1.074e+00 0.28284   
## Sub\_CategorySocks 1.003e-11 7.613e-12 1.317e+00 0.18790   
## Sub\_CategoryTires and Tubes -1.005e-10 7.260e-12 -1.384e+01 < 2e-16 \*\*\*  
## Sub\_CategoryTouring Bikes NA NA NA NA   
## Sub\_CategoryVests NA NA NA NA   
## Order\_Quantity -4.613e-13 7.022e-14 -6.570e+00 5.07e-11 \*\*\*  
## Unit\_Cost -2.834e-14 3.810e-14 -7.440e-01 0.45696   
## Unit\_Price 1.052e-14 2.262e-14 4.650e-01 0.64178   
## Profit 1.000e+00 3.516e-15 2.844e+14 < 2e-16 \*\*\*  
## Cost 1.000e+00 1.989e-15 5.029e+14 < 2e-16 \*\*\*  
## Month 1.209e-13 1.564e-13 7.730e-01 0.43941   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.717e-10 on 113004 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 2.12e+29 on 31 and 113004 DF, p-value: < 2.2e-16

### Model Diagnostic Plots

plot(model1)



Model summary indicates a problem with fit of the model because we are warned of singularities. The model diagnostic plots confirm issues with the model. Most of the residuals are clustered near zero on the y-axis, which suggests that the fitted values are close to the observed values for most data points. Most of the points appear tightly clustered near the line, but there are extreme outliers that deviate significantly from the expected normal behavior. The model summary also confirm issues with model fit which very likely is due to the high multicollinearity detected above. Linear regression is not able to handle multicollinearity and thus I will need to use a different a different regression model that can like random forest.

## Random Forest Model

First, I partitioned the data using a 72/25 split.

ind <- createDataPartition(sales\_data2$Revenue, p=0.75, list=FALSE)  
train <- sales\_data2[ind, ]  
test <- sales\_data2[-ind, ] %>%   
 mutate\_if(is.character, as.factor)

I also set up the random forest to run with a 5-fold cross validation after which I generated predictions to assess how well the model performed with an RMSE.

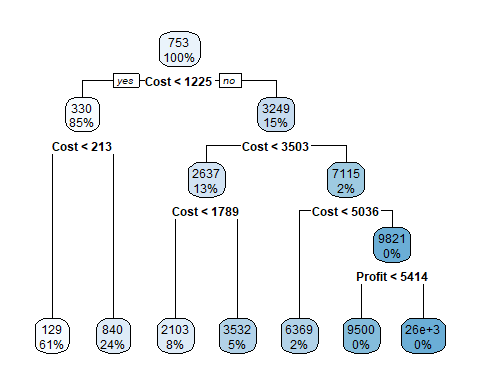
# train\_cont <- trainControl(method = "cv", number = 5)  
# rf <- train(Revenue~.,data = train, method = "rf", trControl = train\_cont, tuneLength = 3)

# rf <- randomForest(Revenue~.,data = train, importance=TRUE)

# preds <- predict(rf, test)  
# sqrt(mean((preds-test$Revenue)^2))

## Decision Tree

cart <- rpart(Revenue~.,data = train, method = "anova")  
rpart.plot(cart)



preds <- predict(cart, newdata = test)  
sqrt(mean((preds-test$Revenue)^2))

## [1] 313.8018

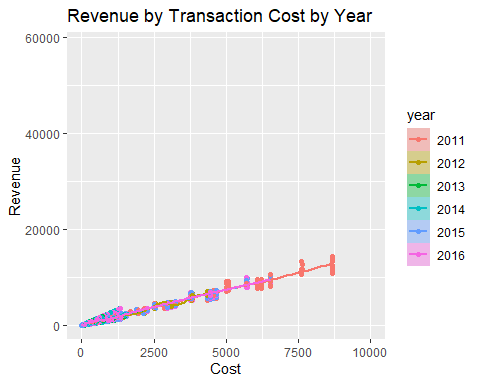
Even though the random forest model failed, a decision tree model did successfully run. This model predicts that the sole predictor of Revenue is the item cost which yields an RMSE of $356.29 per transaction. This seems like a very high error rate and suggests that the model is not a doing a great job of predictive model. However, it may help understand why revenues have declined if sales of high-cost items such as bicycles have fallen over time.

ggplot(sales\_data2, aes(x = Cost, y = Revenue, fill = as.factor(Year), color=as.factor(Year))) +  
 geom\_point() +  
 geom\_smooth() +  
 xlim(0,10000) +  
 labs(title = "Revenue by Transaction Cost by Year", y = "Revenue", x = "Cost", fill="year", color="year")

## `geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_smooth()`).

## Warning: Removed 2 rows containing missing values or values outside the scale range  
## (`geom\_point()`).



Two very large outlier were removed but this shows for all years in the dataset increasing transaction cost led to increasing revenue, however this was most notable in the earlier year of 2011.

ggplot(sales\_data2, aes(y = Revenue, x = Year, fill = Product\_Category, color=Product\_Category)) +  
 geom\_col() +  
 labs(title = "Revenue by Product Category over Time", y = "Revenue",

x = "Year")A graph of sales

Description automatically generated

This plot further confirms that over time smaller slices of revenue proportionally have come from high-cost items like bikes, because more revenue is being generated by other lower cost items like clothing and accessories than in the early years of 2011 and 2012.

# CONCLUSION

An online cycling products sales company reported sales transactions from 2011 - 2016. My initial observation showed that revenue had been declining over time which I attempted to explain with a regression, random forest, and a decision tree model. My regression model was inappropriate due to high multicollinearity, so I chose a random forest model which is known to be more robust to this situation. Unfortunately, I was unable to achieve success with running even a simple random forest model due to computational limitations. My code above (commented out) shows my attempt at this. Because an RF is an ensemble of decision trees, I chose to fit a single regression-based decision tree instead. I found that transaction cost was the sole predictor of revenue, but this yielded a high predicted error rate. This suggests an area for the company to more deeply explore. For example, the scatter plot above shows the transaction costs were the highest in 2011 and the plot below shows all revenue in 2011 and 2012 came from the sales of bicycles. Revenue declined in 2013 which corresponds with the addition of other items for sale such as accessories and clothing. Which may suggest that although these products are important for company solvency, these lower cost items cannot generate the revenue that the higher cost items did. The company may want to explore promotional sales of high-cost items to boost revenue. Additional analysis could include an SVM or NAIVE Bayes model even if RF cannot run.