NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

AY 2023/24 SEMESTER 2

MH3511 Data Analysis with Computer

Group Project

<Coffee, Tea or Me: ME!>

Name	Matriculation Number
Suki Ng	U2210602K
Tay Shu Shuang	U2222462C
Rachel Lim (Lin Jiahuan)	U2220056H
Chai Jie En	U2221398J
Rheanne Leong En Ting	U2222661F

***Abstract:** The beverage landscape is divided between the two: coffee and tea, each with its own nuances and preferences. In this study, we delve into the intricate world of coffee and tea sales in the United States, leveraging a comprehensive dataset spanning the years 2012-2015. Our analysis aims to uncover key insights into consumer behaviour, market dynamics, and the factors influencing sales patterns in this industry.*

Content Page

1. Introduction

Coffee or Tea? Regular or Decaffeinated? The world has been split into two.

With the rise of major franchises and artisanal cafes, we seek to delve deeper and understand the factors affecting the intricacies of each beverage. Above all, we want to enable upcoming businesses to predict optimal locations for establishing shops and to determine what products to offer. In our project, a dataset from Kaggle containing the profit margins of coffee chains around the United States from 2012-15 with additional variables such as market size, type of coffee and product type. By utilising this dataset, we aim to answer the following questions regarding coffee and tea:

1. Which is the most popular type of coffee and tea respectively?
2. Is the major market or small market more popular for coffee and tea sales respectively?
3. Does the sales of the coffee increase over the years?
4. Which variables affect sales the most?
5. Does the major market hit their target sales better than the small market?

This report will cover the data descriptions and analysis using R language. For each of our research objectives, we performed statistical analysis and drew conclusions in the most appropriate approach, together with explanations and elaborations.

2. Data Description and data cleaning

The dataset “Coffee Chain Sales Analysis” is obtained from the machine learning and data science online community Kaggle.
Before proceeding to data analysis, we first performed a preliminary data cleaning to ensure that:

- Irrelevant columns are removed: Area Code, Marketing, Target_cogs, Target_margin, Target_profit, Target_sales
- Duplicated rows are removed

We also added one more variable ‘id’ to assist us in data cleaning.

After all the preparation, 1062 observations with 16 variables are retained for analysis:

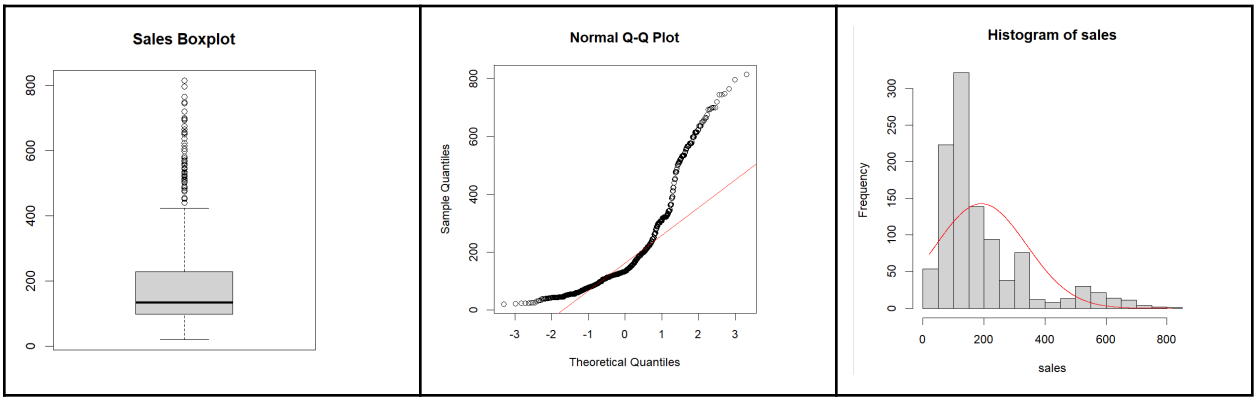
1. COGS (Cost of Goods Sold): The total cost incurred by the coffee chain in producing or purchasing the products it sells.
2. Difference between Actual and Target Profit: This attribute indicates how well the company performed in terms of profit compared to its target. It reflects the financial performance against predefined goals.
3. Date: The date of sales transactions, which allows for time-based analysis of sales trends and patterns.
4. Inventory Margin: The difference between the cost of maintaining inventory and the revenue generated from selling those inventory items. (positive is good - means got profits)
5. Margin: The profit margin, which is the percentage of profit earned from sales. It's a critical financial metric.

6. Market Size: Information about the size of the market in each area, helping to understand the potential customer base and market dynamics.
7. Market: Whether market is Central, South, West or East
8. Product_line: Whether product is made from leaves (tea) or beans (coffee)
9. Product_type: The type of product (e.g Herbal Tea, Espresso, Tea, Coffee)
10. Product: Name of product (e,g Mint, Earl Grey, Caffe latte)
11. Profit: financial gain achieved by the company after deducting the cost of goods sold (COGS) and other expenses from the revenue generated through sales.
12. Sales: represent the revenue generated from the coffee chain's products, reflecting its financial performance and customer demand.
13. State: Name of State the product origins from (e.g California, Utah, Washington)
14. Total_expenses: Total expenses on the product
15. Type: Regular or Decaffeinated
16. Id: unique identifying numbering for each tuple to aid in data cleaning

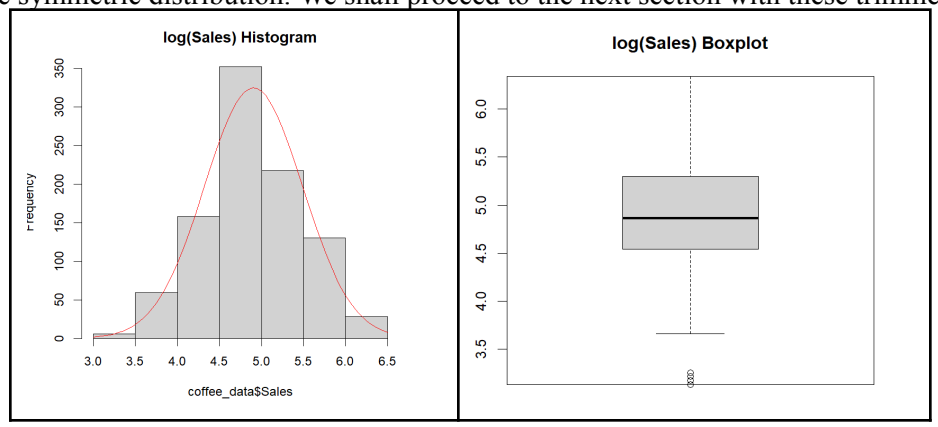
3. Summary Statistics

In this section, we analyse each variable in detail by removing outliers and performing a log base e transformation to avoid highly skewed data.

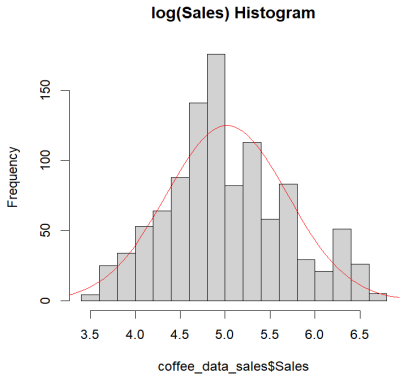
3.1 Summary statistics for the main variable of interest, Sales

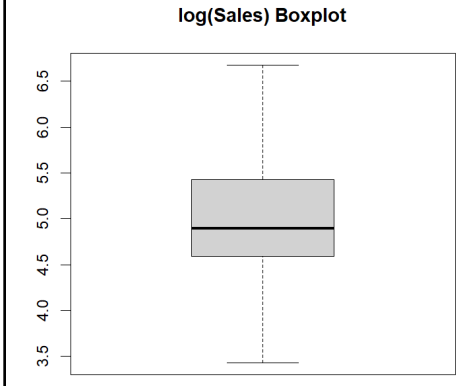


It appears that the variable *Sales* is highly left skewed, we apply a log-transformation to the variable to achieve a more symmetric distribution. We shall proceed to the next section with these trimmed datasets.



After removing outliers:



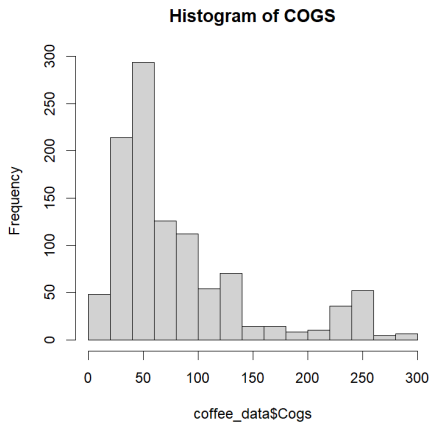


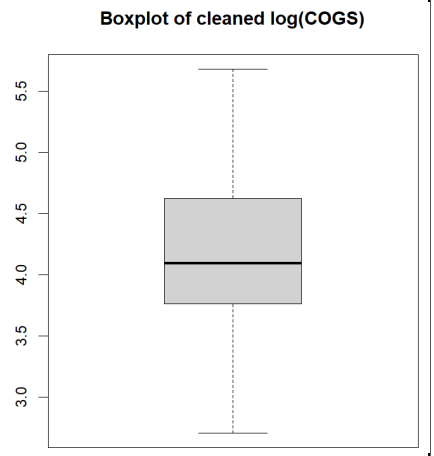
- 9 outlying values were removed using boxplot rule

```
summary(coffee_data_sales$Sales)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3.434   4.595   4.898   5.020   5.429   6.680
```

3.2 Summary statistics for the other variables

3.2.1 Cost of Goods Sold, COGS

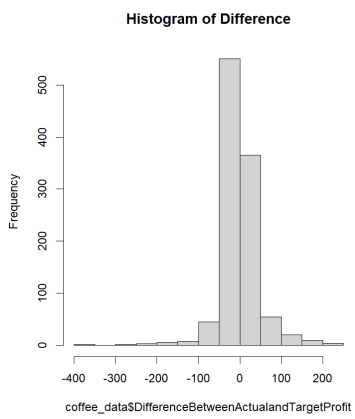


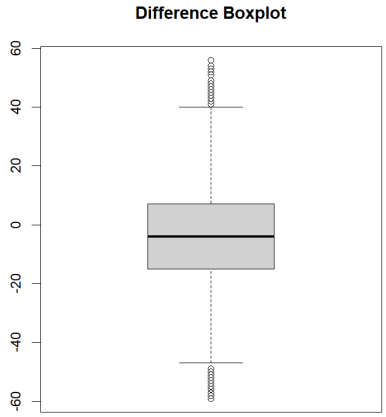


The log-transformation (base e) is applied

- 18 rows with COGS of value zero were removed from original dataframe
- 6 outlying values were removed using boxplot rule

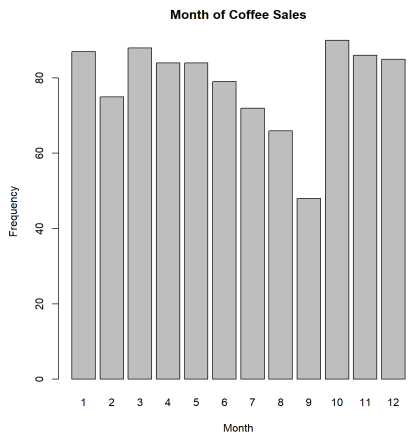
3.2.2 Difference between actual and target profit

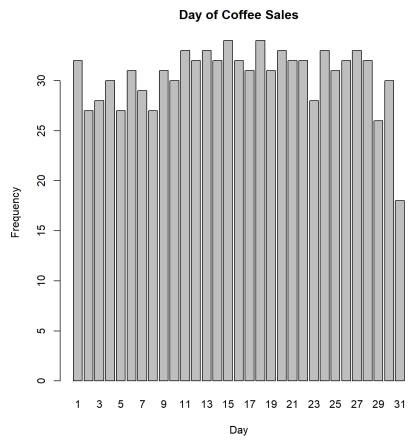




- 114 outlying values were removed using the boxplot rule

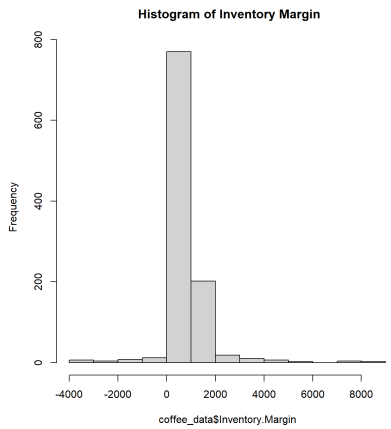
3.2.3 Date

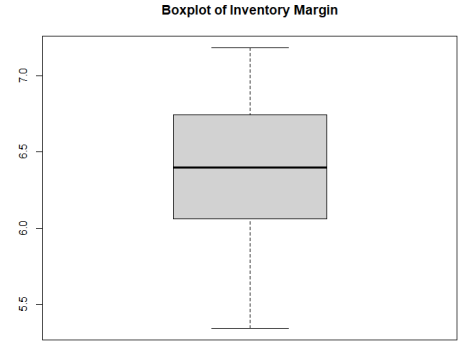




- No outlying values removed
- There is no huge difference in the Month and Day

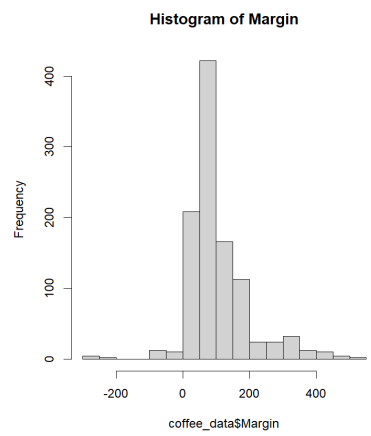
3.2.4 Inventory Margin

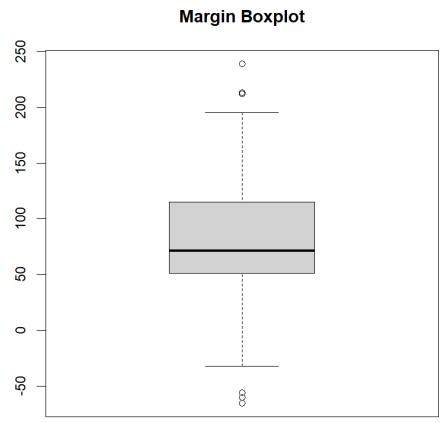




- Since there are Inventory Margin <0, we do not log the data
- 114 outliers removed
- Data is normally distributed

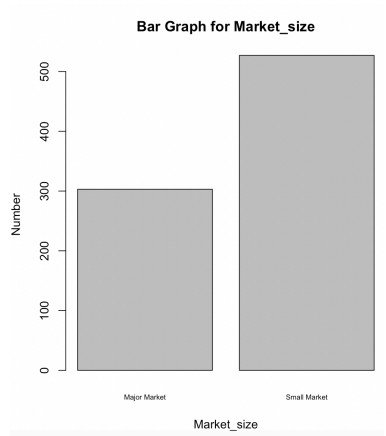
3.2.5 Margin





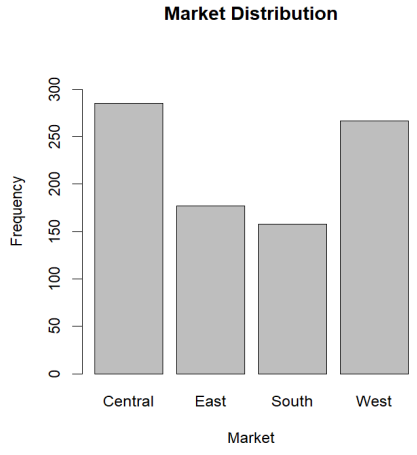
- 96 outliers removed
- Since there exist <0 Margin data, we do not log the data
- Data is normally distributed

3.2.6 Market Size



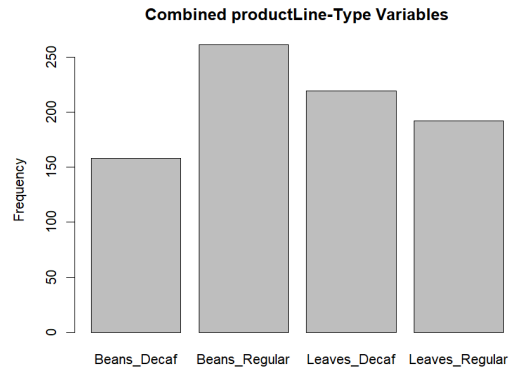
- No removal of outliers
- There is a significant difference in the frequency of major and small markets present in the data set
- More small markets than major markets

3.2.7 Market



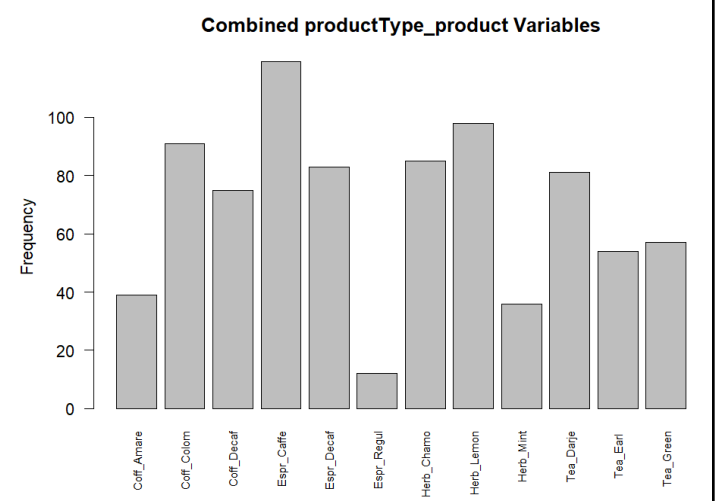
- No outlying value of market is removed
- There is a relatively greater distribution of the Central and Western market as compared to the Eastern and Southern market

3.2.8 Combined ProductLine-Type



- No outlying value of ProductLine-Type is removed
- There is no huge difference in the ProductLine-Type distribution

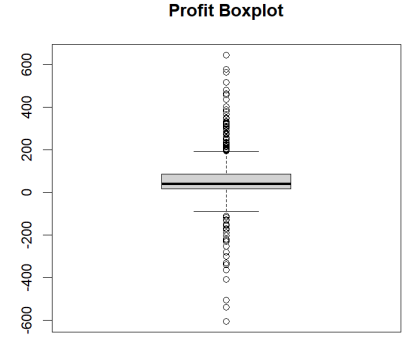
3.2.6 Combined Product_type-Product

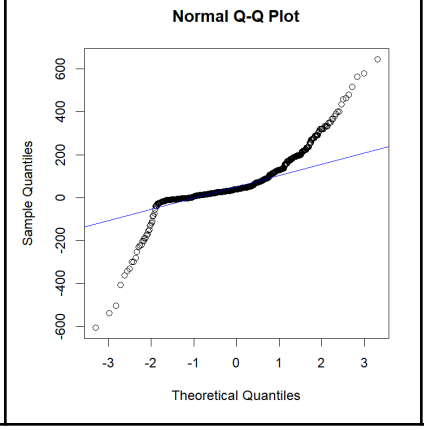


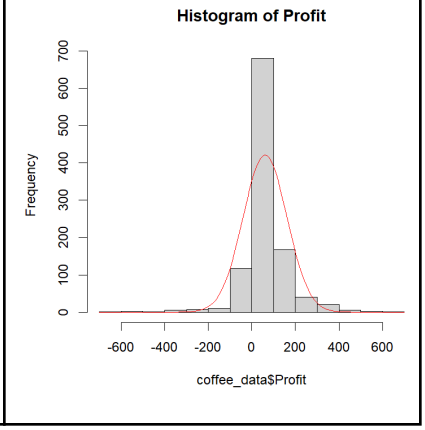
- No outlying value is removed
- There is no significant difference in the distribution of Product_type-Product.
- Caffeinated espresso has the highest frequency, and regular espresso has the lowest frequency in the dataset

3.2.7 Profit

The following plots show the overall distribution of the variable *Profit*

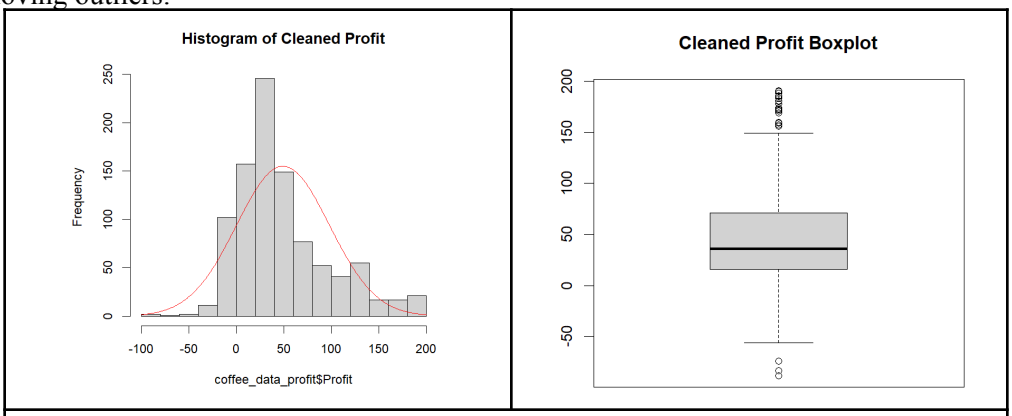






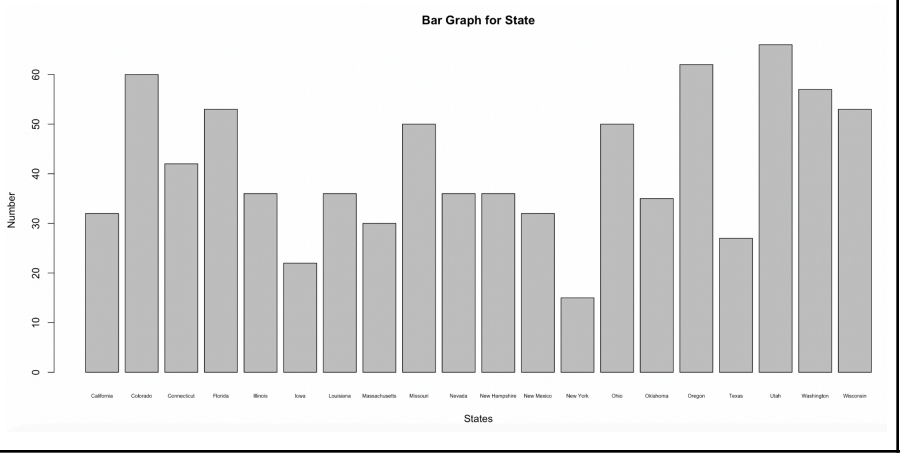
The Profit data appears to have some outlying data on both the left and right tail. Therefore, we will remove outliers using the boxplot rule to achieve a trimmed dataset with the following summary statistics.

After removing outliers:



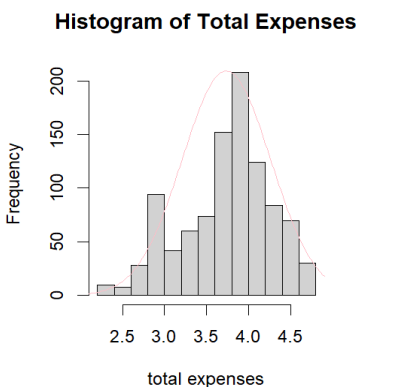

```
> summary(removed_profit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -88.00   16.00   36.00   48.93   71.00  191.00
```

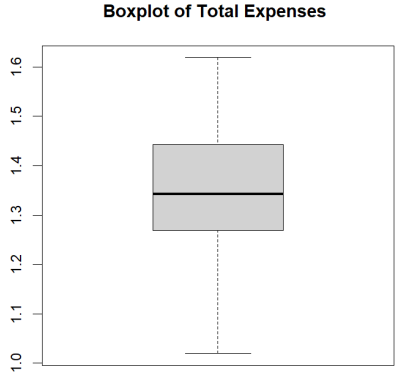
3.2.8 State



- No outlying value is removed
- There is a huge difference in the distribution
- Utah has the highest amount of coffee/tea sold
- New York has the lowest amount of coffee/tea sold

3.2.9 Total expenses





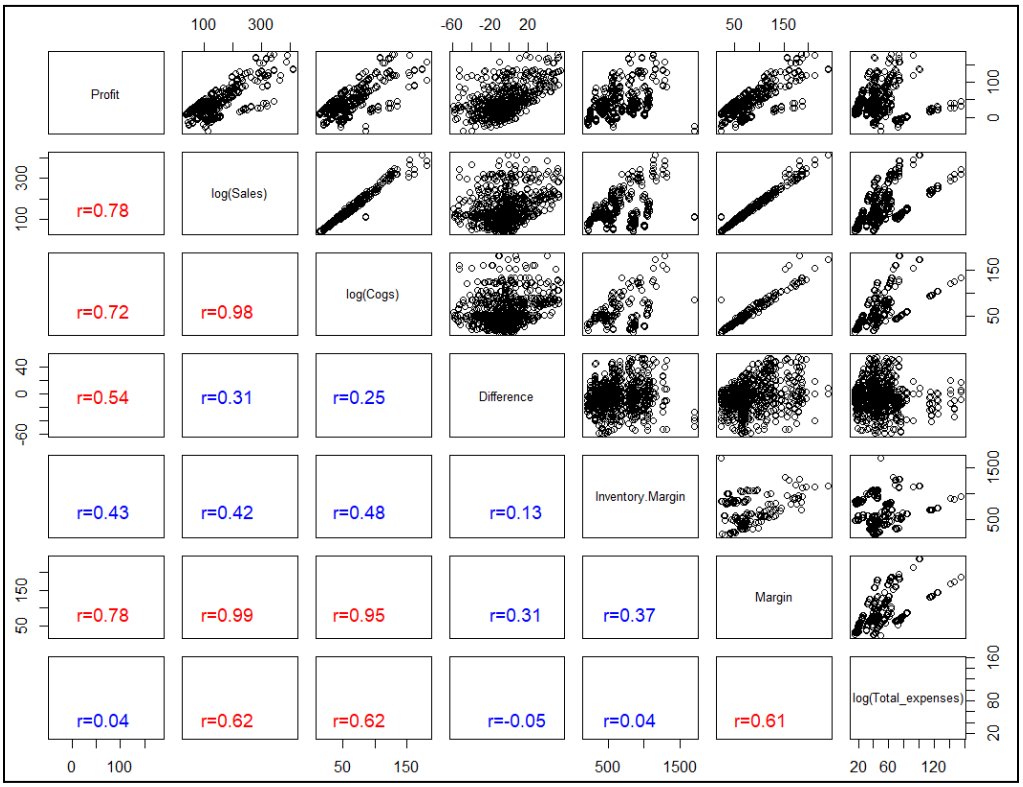
- The log-transformation (base e) is applied
- 14 outliers removed

3.3 Summary of final dataset for analysis

Based on the above analysis, the final dataset is further reduced to 830 observations and 16 variables, with the suggested transformation. Namely applying log(base e) to the Sales variable.

4. Statistical analysis

4.1 Correlations between main variable, *log(Sales)* and other continuous variables



Based on the scatter plot and correlation coefficient, it appears that *log(Sales)* is more highly correlated to *Profit, log(Cogs)*, *Margin* and *log(Total_expenses)*. *Profit* is more highly correlated to *log(Sales)*, *log(Cogs)* and *Margin* than other variables. As such, we have decided to focus on the statistical analysis of *log(Sales)* against other variables since they appear to have more correlations than *Profits* against other variables. Additionally, to answer one of our main questions on the popularity of coffee and tea respectively, sales seems to be a better gauge than profits as it is directly related to consumer demand and consumption patterns.

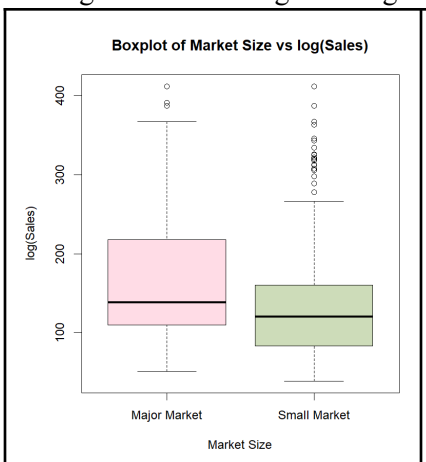
Among the variables, the interesting observations are:

- *log(Cogs)* and *Margin* are highly positively correlated (r=0.95)
- *log(Cogs)* and *log(Total_expenses)* are quite positively correlated (r=0.62)
- *Margin* and *log(Total_expenses)* are quite positively correlated (r=0.61)

4.2 Statistical tests

4.2.1 Relation between log(sales) and Market_size

In this section, we carried out f-test and t-test to determine whether log(Sales) is determined by the *Market_size*, namely the Major and Small Market, since *Market_size* is a categorical variable. Studying the mean of log(Sales) between Major and Small markets, we can identify that the log(Sales) mean of Small Markets is slightly lower than that of Major Markets, indicating a possibility of companies in the larger markets being able to generate greater sales.




```
> t.test(market_size_table$Sales ~ market_size_table$Market_size, var.equal = FALSE)

        Welch Two Sample t-test

data:  market_size_table$Sales by market_size_table$Market_size
t = 6.628, df = 553.41, p-value = 8.091e-11
alternative hypothesis: true difference in means between group Major Market and group Small Market is not equal to
0
95 percent confidence interval:
 26.31544 48.48244
sample estimates:
mean in group Major Market mean in group Small Market
                  172.5413                   135.1423

> t.test(market_size_table$Sales ~ market_size_table$Market_size, var.equal = FALSE)

        Welch Two Sample t-test

data:  market_size_table$Sales by market_size_table$Market_size
t = 6.628, df = 553.41, p-value = 8.091e-11
alternative hypothesis: true difference in means between group Major Market and group Small Market is not equal to 0
95 percent confidence interval:
 26.31544 48.48244
sample estimates:
mean in group Major Market mean in group Small Market
                  172.5413                   135.1423
```

Variance test

H_0: Variances of Major and Small markets' log(Sales) are equal;

H_1: Variances of Major and Small markets' log(Sales) are not equal;

At a significance level of 0.05, we reject the null hypothesis and conclude that the variances of the two samples are not equal, since p-value = 8.091e-11< 0.05.

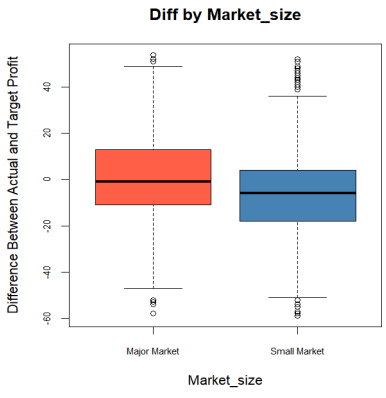
T-test

H_0: The mean of log(Sales) under Major Market is equal to that under Minor Market;

H_1: The mean of log(Sales) under Major Market is larger than that under Minor Market;

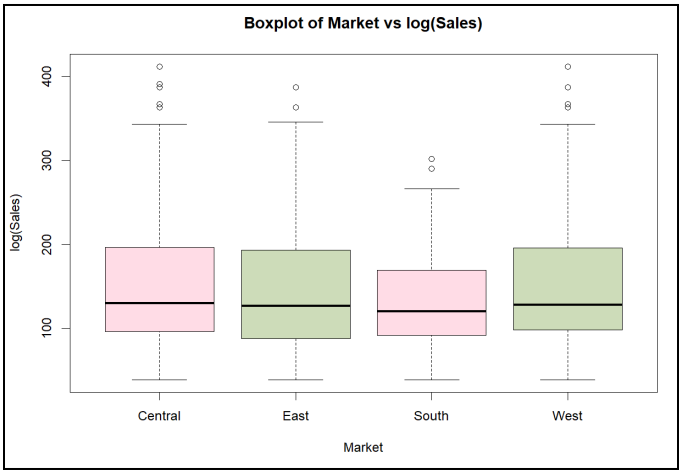
Using a sided t-test with unequal variances, the p-value is less than 2.2e-16 < 0.05. At a significance level of 0.05, we reject the null hypothesis and conclude that the mean of log(Sales) under Small Market is significantly less than that under Major Market. Therefore, we conclude that major markets have higher sales.

4.2.2 Relation between the Difference Between Actual and Target Profit against Market_size



We delved deeper to study the effects of *Market_size*, and analysed its relation to *DifferenceBetweenActualandTargetProfit.* From the box plot, we can see that Major Markets are better able to meet their target profits as compared to small markets, based on the median values of the two.

4.2.3 Relation between log(Sales) and Market



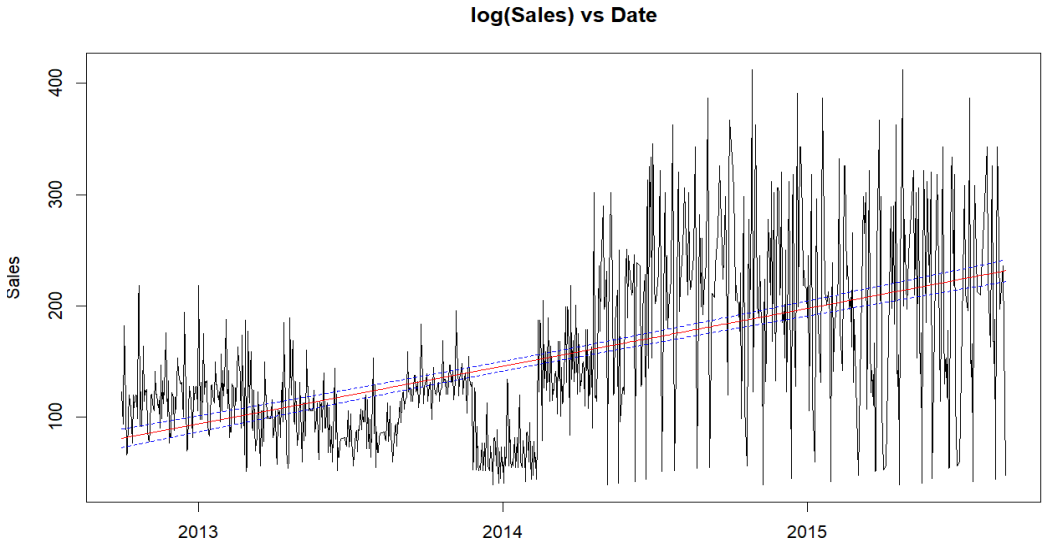
In this section, we determine whether the sales are affected by the Market. We did a boxplot of *log(Sales)* against the *Market*. Studying the means of *log(Sales)* against Central, East, South, and West markets respectively, we could identify that the sales are relatively equal for each market just based on the boxplots. Hence we will conduct ANOVA testing to confirm our hypothesis.

We test, H_0: $\mu_{Central} = \mu_{East} = \mu_{East} = \mu_{East}$ against H_1: not all μ_i are equal

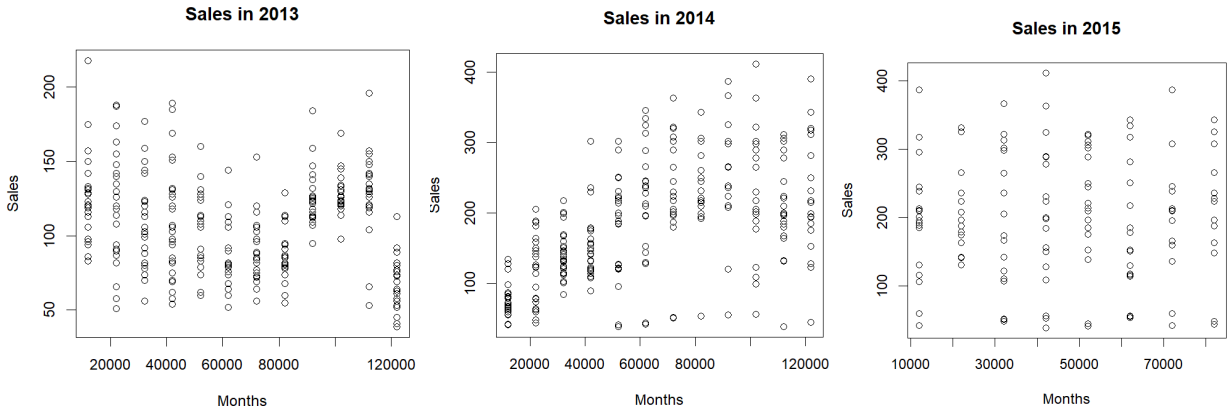
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
merged_coffee_data$Market_size	1	269088	269088	47.76	9.59e-12
Residuals	828	4664662	5634		

The p-value of the ANOVA test is 0.59e-12(<0.05), therefore we would reject H_0 at a significance level of 0.05. Thus we would conclude that the sales of the coffee is not independent of the market.

4.2.4 Time series analysis of log(Sales) vs Date



The sales graph over time shows a clear upward trend over the years but with increased volatility. The red regression line depicts the overall trend in sales, while the blue lines represent the 95% confidence interval. The narrow spread of the blue lines around the regression line signifies higher certainty in the trend's accuracy, despite the observed volatility in sales data.

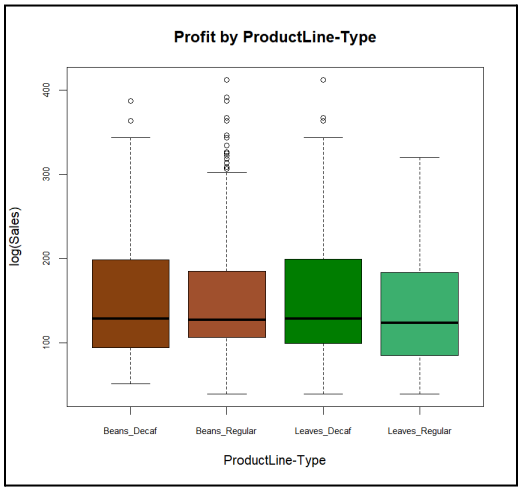


We zoom into each individual year and plot the sales per year against months, noting the following trends: There is no noticeable dependency between sales and month of the year. There is a general increase in sales over the years 2013 and 2014, with relatively constant sales throughout the year 2015. Similar to the analysis above, sales data appears to be more volatile as the years increase.

4.2.5 Boxplot for log(Sales) against ProductLine-Type

Since we aim to answer our question, is coffee (caf vs decaf) or tea (caf vs decaf) more popular? We would perform the analysis of variance (ANOVA) test to see if the productline-type of coffee affects the log(Sales) of coffee. To test,

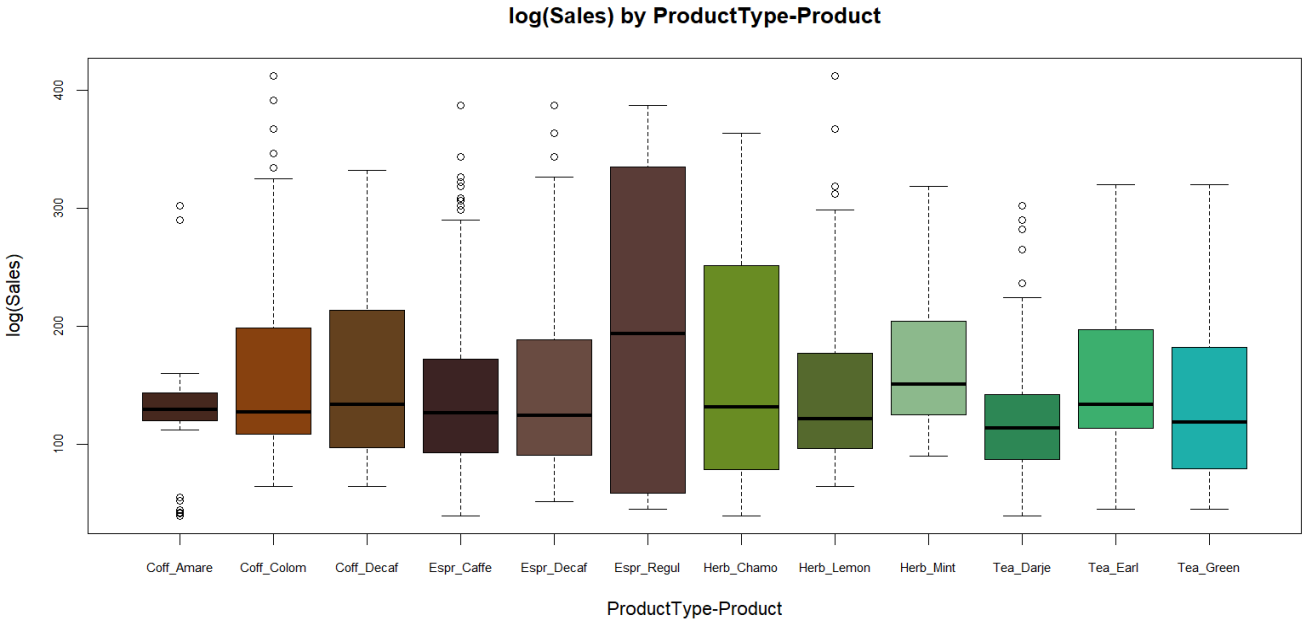
$$H_0: \mu_{beans\ decaf} = \mu_{beans\ regular} = \mu_{leaves\ decaf} = \mu_{leaves\ decaf} \text{ against } H_1: \textit{not all } \mu_i \textit{ are equal}$$



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
merged_coffee_data$productLine_type	3	52178	17393	2.943	0.0323
Residuals	826	4881571	5910		

The p-value of the ANOVA test is 0.0323(<0.05), therefore we would reject H_0 at a significance level of 0.05. Thus, we conclude that the sales of the coffee is not independent of the product line type. In the next section, we continue to investigate which specific product type is the most popular amongst consumers.

4.2.6 Boxplot for log(Sales) against ProductType-Product



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
merged_coffee_data$productType_product	11	174675	15880	2.729	0.0018
Residuals	818	4759074	5818		

To test,

$$H_0: \mu_{\{productType-product\}} \; are\ all\ equivalent \quad \text{against } H_1: \; not\ all\ \mu_i \; are\ equal$$

The p value of this ANOVA test is 0.0018(<0.05), thus we would reject H_0 at a significance level of 0.05. From the boxplot, we can also tell that *regular espressos* have higher sales on average relative to all the other types of coffee. Thus we can conclude that there is a difference in sales among different types of coffees.

4.3 The single most important variable that is affecting Sales?

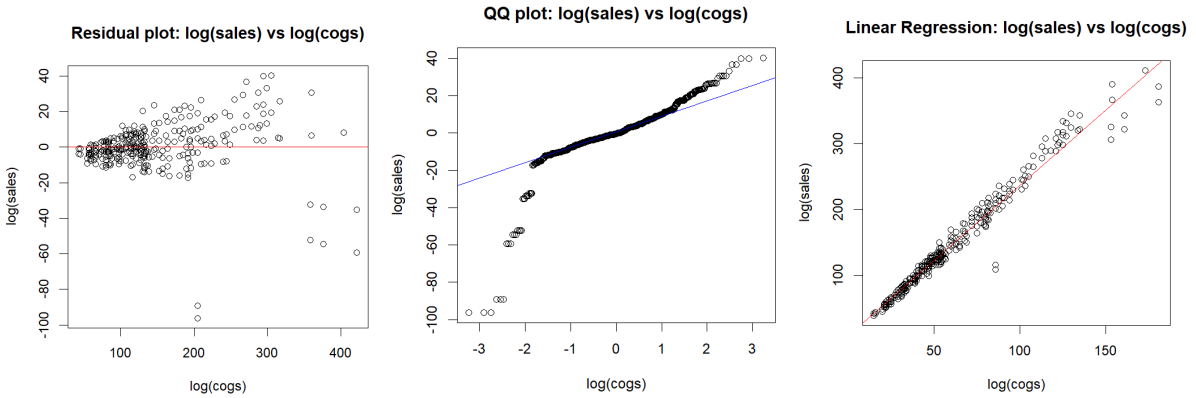
We can see that *log(Cogs)*, *Margin* and *log(total expenses)* have the highest correlation with *log(sales)*. We can perform simple linear regression analysis to determine which of the 3 performance measures can is the most suitable to model a linear function as shown below,

$$\log(Sales) = aX + b$$

where X could be any one of the variables that was mentioned above. By comparing the R-squared value, QQ plot, residual plot and linear regression plot, we can determine the single most important performance measure to model the log(Sales) with a simple linear model.

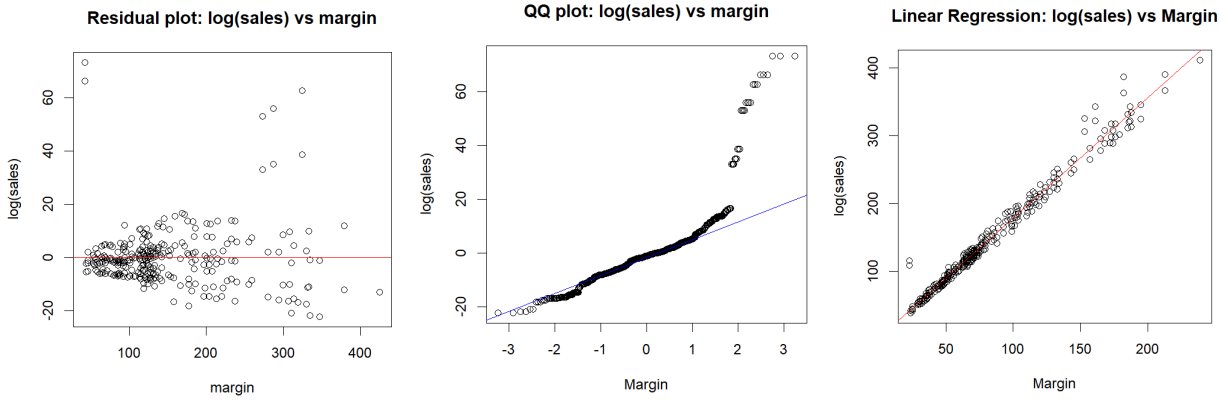
To determine the strongest independent X, the residual plot should be randomly scattered near the line x=0, the QQ norm should align with the QQ line and the R-squared value should be near to the value 1.

4.3.1 log(Cogs)



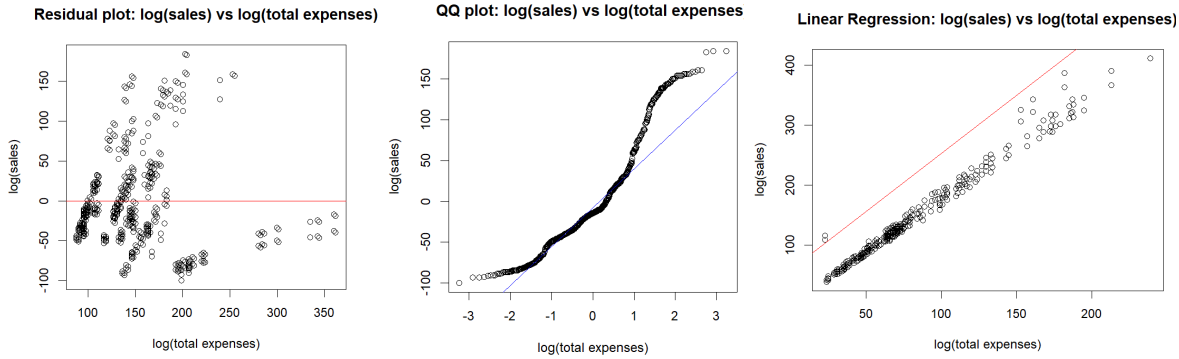
Fitted model: 2.2862X + 8.4691, p-value: 2.2e-16, R-squared value: 0.965536

4.3.2 Margin



Fitted model: 1.7713X + 1.7780, p-value: 2.2e-16, R-squared value: 0.975153

4.3.3 log(total expenses)



Fitted model: 1.9436X + 58.614, p-value: 2.2e-16, R-squared value: 0.388924

Hence, the strongest independent variable is Margin as the R-squared value is the largest and the graphs show that it is a strong variable

4.4 Multiple Linear Regression

In this section, we build a multiple linear regression model based on our 5 numerical variables. We decided on a backward elimination method in order to identify the best fitted model. From our results, it is evident that *log*(Cogs), Difference, Margin and *log*(Total_Expenses) can be used to model *log*(Sales). The fitted model is:

$$\log(\text{Sales}) = 2.28 + 1.08*\log(\text{Cogs}) + 0.13*\text{Difference} + 0.94*\text{Margin} + 0.06*\log(\text{Total\_Expenses})$$

```
Start:  AIC=2548.36
`log(Sales)` ~ `log(Cogs)` + Difference + Inventory.Margin +
    Margin + `log(Total_expenses)`

                       Df Sum of Sq    RSS    AIC
- Inventory.Margin      1         6  17635 2546.7
<none>                              17629 2548.4
- `log(Total_expenses)` 1       941  18569 2589.5
- Difference            1      5231  22860 2762.0
- `log(Cogs)`           1     76990  94619 3941.0
- Margin                1    113098 130727 4209.3

Step:  AIC=2546.66
`log(Sales)` ~ `log(Cogs)` + Difference + Margin + `log(Total_expenses)`

                       Df Sum of Sq    RSS    AIC
<none>                              17635 2546.7
- `log(Total_expenses)` 1      1009  18644 2590.8
- Difference            1      5225  22860 2760.1
- `log(Cogs)`           1    100143 117778 4120.8
- Margin                1    120485 138120 4253.0

Call:
lm(formula = `log(Sales)` ~ `log(Cogs)` + Difference + Margin +
    `log(Total_expenses)`, data = merged_coffee_data)

Coefficients:
          (Intercept)            `log(Cogs)`             Difference                 Margin  `log(Total_expenses)`
              2.28377                1.08047                0.13327                0.93883                0.05961
```

5. Conclusion and Discussion

Through the course of our statistical analysis, we delved into the world of two beloved beverages, studying the factors that affect the sales of coffee and tea. In order for franchises and artisanal cafes alike to decide on their business models and attract more customers, they will have to pay close attention to costs, setting reasonable profit targets and controlling their expenses.

We conclude that:

1. Regular espressos have higher sales on average relative to all the other types of coffee
2. The size of a market, major or minor indeed affects coffee and tea sales
3. Over time, the sales of coffee and tea increase but with greater volatility
4. The variable that affects sales the most is margin
5. Major markets hit their target sales better than small markets

Looking ahead, further research could explore additional variables that may influence sales trends, such as demographic factors, consumer preferences, and the impact of marketing strategies. Additionally, expanding the scope of analysis to encompass broader geographical regions or a more extensive time frame could provide deeper insights into long-term market dynamics and trends.

6. Appendix

6.1 Data Cleaning (Numerical)

Read data and drop columns:

```
> coffee_data = read.csv("C:/Users/rhean/Documents/RHEANNE/Y2S2_MH3511/Project/Coffee_Chain_Sales .csv")
> #Remove unwanted columns
> coffee_data<-coffee_data[,-16:-19,]#remove columns 16-19 from the dataframe
> coffee_data<-coffee_data[,-9,]#remove 9th column
> coffee_data<-coffee_data[,-1,]#remove 1st column
> coffee_data<-coffee_data[!duplicated(coffee_data),]
> #add new column for id
> coffee_data$id<-seq(1,nrow(coffee_data),1)
> coffee_data_profit<-coffee_data
> coffee_data_sales<-coffee_data
```

Box Plot, QQ norm, Histogram and Cleaning Profits:

```
> #Profit column
> boxplot(coffee_data$Profit, main ="Profit Boxplot")
> qqnorm(coffee_data$Profit, main = "Normal Q-Q Plot for Profits")
> qqline(coffee_data$Profit, col="blue")
> hist(coffee_data$Profit, main = "Histogram of Profit")
> xpt<-seq(-700,700,by=10)
> n_den<-dnorm(xpt,mean(coffee_data$Profit),sd(coffee_data$Profit))
> ypt<-n_den*length(coffee_data$Profit)*100
> lines(xpt,ypt,col="red")
> #removing of outliers using boxplot rule
> coffee_data_profit<-coffee_data[coffee_data$Profit>=quantile(coffee_data$Profit,0.25)-1.5*IQR(coffee_data$Profit)&quantile(coffee_data$Profit,0.75)+1.5
*IQR(coffee_data$Profit)>= coffee_data$Profit,]
> hist(coffee_data_profit$Profit, main = "Histogram of Cleaned Profit")
> xpt<-seq(-100,200,by=1)
> n_den<-dnorm(xpt,mean(coffee_data_profit$Profit),sd(coffee_data_profit$Profit))
> ypt<-n_den*length(coffee_data_profit$Profit)*20
> lines(xpt,ypt,col="red")
> boxplot(coffee_data_profit$Profit, main ="Cleaned Profit Boxplot")
> summary(coffee_data_profit$Profit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -88.00   16.00   36.00   48.93   71.00  191.00
```

Box Plot, QQ norm, Histogram and Cleaning Sales:

```
> #Sales Column
> boxplot(coffee_data$Sales, main ="Sales Boxplot")
> qqnorm(coffee_data$Sales, main = "Normal Q-Q Plot for Sales")
> qqline(coffee_data$Sales, col="blue")
> hist(coffee_data$Sales, main = "Histogram of Sales")
> xpt<-seq(0,600,by=10)
> n_den<-dnorm(xpt,mean(coffee_data$Sales),sd(coffee_data$Sales))
> ypt<-n_den*length(coffee_data$Sales)*50
> lines(xpt,ypt,col="red")
> #apply log transformation for Sales
> coffee_data$Sales<-log(coffee_data$Sales)
> hist(coffee_data$Sales)
> xpt<-seq(3,6.5,by=0.05)
> n_den<-dnorm(xpt,mean(coffee_data$Sales),sd(coffee_data$Sales))
> ypt<-n_den*length(coffee_data$Sales)*0.5
> lines(xpt,ypt,col="red")
> #remove outliers for Sales
> coffee_data_sales<-coffee_data[coffee_data$Sales>=quantile(coffee_data$Sales,0.25)-1.5*IQR(coffee_data$Sales)&quantile(coffee_data$Sales,0.75)+1.5*IQR
(coffee_data$Sales)>= coffee_data$Sales,]
> hist(coffee_data_sales$Sales, main="log(Sales) Histogram")
> xpt<-seq(3,7,by=0.05)
> n_den<-dnorm(xpt,mean(coffee_data_sales$Sales),sd(coffee_data_sales$Sales))
> ypt<-n_den*length(coffee_data_sales$Sales)*.2
> lines(xpt,ypt,col="red")
> boxplot(coffee_data_sales$Sales, main="log(Sales) Boxplot")
> summary(coffee_data_sales$Sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.434   4.595   4.898   5.020   5.429   6.680
```

Box Plot, QQ norm, Histogram and Cleaning COGs:

```
> #Cogs Column
> boxplot(coffee_data$Cogs, main ="COGS Boxplot")
> hist(coffee_data$Cogs, main = "Histogram of COGS")
> xpt<-seq(0,300,by=10)
> n_den<-dnorm(xpt,mean(coffee_data$Cogs),sd(coffee_data$Cogs))
> ypt<-n_den*length(coffee_data$Cogs)*10
> lines(xpt,ypt,col="red")
> #remove 0 value
> coffee_data <- coffee_data[coffee_data$Cogs!=0,]
> #remove NA values
> coffee_data<-coffee_data[!is.na(coffee_data$Cogs),]
> #log transformation for COGS
> coffee_data$Cogs<-log(coffee_data$Cogs)
> hist(coffee_data$Cogs, main ="Histogram of COGs")
> xpt<-seq(2,6,by=0.05)
> n_den<-dnorm(xpt,mean(coffee_data$Cogs),sd(coffee_data$Cogs))
> ypt<-n_den*length(coffee_data$Cogs)*0.5
> lines(xpt,ypt,col="red")
> #remove outliers for COGS
> coffee_data_cogs<-coffee_data[coffee_data$Cogs>=quantile(coffee_data$Cogs,0.25)-1.5*IQR(coffee_data$Cogs)&quantile(coffee_data$Cogs,0.75)+1.5*IQR(coffe
e_data$Cogs)>= coffee_data$Cogs,]
> hist(coffee_data_cogs$Cogs, main="Histogram of cleaned log(COGs)")
> xpt<-seq(2.5,6,by=0.01)
> n_den<-dnorm(xpt,mean(coffee_data_cogs$Cogs),sd(coffee_data_cogs$Cogs))
> ypt<-n_den*length(coffee_data_cogs$Cogs)*0.1
> lines(xpt,ypt,col="red")
> boxplot(coffee_data_cogs$Cogs, main="Boxplot of cleaned log(COGS)")
> summary(coffee_data_cogs$Cogs)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.708   3.761   4.094   4.192   4.625   5.684
```

Box Plot, QQ norm, Histogram and Cleaning Inventory.Margin:

```
> #Inventory Margin
> boxplot(coffee_data$Inventory.Margin, main ="Inventory Margin Boxplot")
> hist(coffee_data$Inventory.Margin, main = "Histogram of Inventory Margin")
> xpt<-seq(-4000,6000,by=10)
> n_den<-dnorm(xpt,mean(coffee_data$Inventory.Margin),sd(coffee_data$Inventory.Margin))
> ypt<-n_den*length(coffee_data$Inventory.Margin)*1000
> lines(xpt,ypt,col="red")
> #remove outliers for inventory margin
> coffee_data_inventory<-coffee_data[coffee_data$Inventory.Margin>=quantile(coffee_data$Inventory.Margin,0.25)-1.5*IQR(coffee_data$Inventory.Margin)&
quantile(coffee_data$Inventory.Margin,0.75)+1.5*IQR(coffee_data$Inventory.Margin)>= coffee_data$Inventory.Margin,]
> hist(coffee_data_inventory$Inventory.Margin, main="Inventory Margin Histogram")
> boxplot(coffee_data_inventory$Inventory.Margin, main="Inventory Margin Boxplot")
> summary(coffee_data_inventory$Inventory.Margin)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  209.0   454.0   627.0   713.7   885.0  1744.0
```

Box Plot, QQ norm, Histogram and Cleaning Margin:

```
 #Margin Column
 boxplot(coffee_data$Margin, main ="Margin Boxplot")
 hist(coffee_data$Margin, main = "Histogram of Margin")
 xpt<-seq(-200,400,by=10)
 n_den<-dnorm(xpt,mean(coffee_data$Margin),sd(coffee_data$Margin))
 ypt<-n_den*length(coffee_data$Margin)*100
 lines(xpt,ypt,col="red")
 #remove outliers
 coffee_data_margin<-coffee_data[coffee_data$Margin>=quantile(coffee_data$Margin,0.25)-1.5*IQR(coffee_data$Margin)&quantile(coffee_data$Margin,0.75)+1.5
IQR(coffee_data$Margin)>= coffee_data$Margin,]
 hist(coffee_data_margin$Margin, main="Margin Histogram")
 boxplot(coffee_data_margin$Margin, main="Margin Boxplot")
 summary(coffee_data_margin$Margin)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-65.00   51.00   71.00   86.18  115.00  239.00
```

Box Plot, QQ norm, Histogram and Cleaning Total_Expenses:

```
#Total Expenses column
boxplot(coffee_data$Total_expenses, main ="Total Expenses Boxplot")
hist(coffee_data$Total_expenses, main = "Histogram of Total Expenses")
xpt<-seq(0,150,by=1)
n_den<-dnorm(xpt,mean(coffee_data$Total_expenses),sd(coffee_data$Total_expenses))
ypt<-n_den*length(coffee_data$Total_expenses)*10
lines(xpt,ypt,col="red")
#log transformation for total expenses
coffee_data$Total_expenses<-log(coffee_data$Total_expenses)
#remove outliers for total expenses
coffee_data_total_expenses<-coffee_data[coffee_data$Total_expenses>=quantile(coffee_data$Total_expenses,0.25)-1.5*IQR(coffee_data$Total_expenses)&quant
e(coffee_data$Total_expenses,0.75)+1.5*IQR(coffee_data$Total_expenses)>= coffee_data$Total_expenses,]
hist(coffee_data_total_expenses$Total_expenses, main="Total Expenses Histogram")
xpt<-seq(1,1.6,by=0.01)
n_den<-dnorm(xpt,mean(coffee_data_total_expenses$Total_expenses),sd(coffee_data_total_expenses$Total_expenses))
ypt<-n_den*length(coffee_data_total_expenses$Total_expenses)*0.05
lines(xpt,ypt,col="red")
boxplot(coffee_data_total_expenses$Total_expenses, main="Boxplot of Total Expenses ")
summary(coffee_data_total_expenses$Total_expenses)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.708   3.555   3.829   3.843   4.234   5.050
```

Merging of all DataFrames after cleaning each variable:

```
> merged_coffee_data<-merge(coffee_data_profit,coffee_data_sales,by="id")
> merged_coffee_data<-subset(merged_coffee_data, select = c("id","Cogs.x", "DifferenceBetweenActualandTargetProfit.x","Date.x", "Inventory.Margin.x","Mar
gin.x", "Market_size.x","Market.x", "Product_line.x", "Product_type.x", "Product.x", "Profit.x", "Sales.x", "State.x", "Total_expenses.x", "Type.x"))
> #merge with cogs
> merged_coffee_data<-merge(merged_coffee_data,coffee_data_cogs,by="id")
> merged_coffee_data<-subset(merged_coffee_data, select = c("id","Cogs.x", "DifferenceBetweenActualandTargetProfit.x","Date.x", "Inventory.Margin.x","Mar
gin.x", "Market_size.x","Market.x", "Product_line.x", "Product_type.x", "Product.x", "Profit.x", "Sales.x", "State.x", "Total_expenses.x", "Type.x"))
> #merge with difference
> merged_coffee_data<-merge(merged_coffee_data,coffee_data_difference,by="id")
> merged_coffee_data<-subset(merged_coffee_data, select = c("id","Cogs.x", "DifferenceBetweenActualandTargetProfit.x","Date.x", "Inventory.Margin.x","Mar
gin.x", "Market_size.x","Market.x", "Product_line.x", "Product_type.x", "Product.x", "Profit.x", "Sales.x", "State.x", "Total_expenses.x", "Type.x"))
> #merge with inventory margin
> merged_coffee_data<-merge(merged_coffee_data,coffee_data_inventory,by="id")
> merged_coffee_data<-subset(merged_coffee_data, select = c("id","Cogs.x", "DifferenceBetweenActualandTargetProfit.x","Date.x", "Inventory.Margin.x","Mar
gin.x", "Market_size.x","Market.x", "Product_line.x", "Product_type.x", "Product.x", "Profit.x", "Sales.x", "State.x", "Total_expenses.x", "Type.x"))
> #merge with margin
> merged_coffee_data<-merge(merged_coffee_data,coffee_data_margin,by="id")
> merged_coffee_data<-subset(merged_coffee_data, select = c("id","Cogs.x", "DifferenceBetweenActualandTargetProfit.x","Date.x", "Inventory.Margin.x","Mar
gin.x", "Market_size.x","Market.x", "Product_line.x", "Product_type.x", "Product.x", "Profit.x", "Sales.x", "State.x", "Total_expenses.x", "Type.x"))
> #merge with total expenses
> merged_coffee_data<-merge(merged_coffee_data,coffee_data_total_expenses,by="id")
> merged_coffee_data<-subset(merged_coffee_data, select = c("id","Cogs.x", "DifferenceBetweenActualandTargetProfit.x","Date.x", "Inventory.Margin.x","Mar
gin.x", "Market_size.x","Market.x", "Product_line.x", "Product_type.x", "Product.x", "Profit.x", "Sales.x", "State.x", "Total_expenses.x", "Type.x"))
> colnames(merged_coffee_data) <- c("id", "log(Cogs)","Difference","Date", "Inventory.Margin","Margin", "Market_size","Market", "Product_line", "Product_
type", "Product", "Profit", "log(Sales)", "State", "log(Total_expenses)", "Type")
```

6.2 Summary Statistics (Categorical)

6.2.1 Market Size

```
> #Market_size Column
> marketsize_counts <-table(merged_coffee_data$Market_size)
> barplot(marketsize_counts, main = "Bar Graph for Market_size", xlab = "Market_size",
ylab = "Number", cex.names = 0.6)
```

6.2.2 Market

```
> #Market Column
> market_counts <-table(merged_coffee_data$Market)
> barplot(market_counts, main = "Bar Graph for Market", xlab = "Market", ylab = "Area",
cex.names = 0.6)
```

6.2.3 Barplot for productLine-Type

```
#PRODUCTLINE-TYPE
#Combine the abbreviated words into a new variable named combined_category
# Create combined productLine-Type variable
merged_coffee_data$productLine_type <- paste(merged_coffee_data$Product_line,
                                             merged_coffee_data$Type,
                                             sep = "_")
# Count frequencies
productLine_type <- table(merged_coffee_data$productLine_type)
# Plot a bar plot with larger margins and specified colors
barplot(productLine_type,
        main = "Combined productLine-Type Variables",
        ylab = "Frequency")
```

6.2.4 Barplot for productType-Product

```
#PRODUCTTYPE-PRODUCT
# Abbreviate Product_type and Product to four letters
abbreviated_product_type <- substr(merged_coffee_data$Product_type, 1, 4)
abbreviated_product <- substr(merged_coffee_data$Product, 1, 5)
# Combine the abbreviated words into a new variable named combined_category
merged_coffee_data$productType_product <- paste(abbreviated_product_type,
                                        abbreviated_product,
                                        sep = "_")
productType_product = table(merged_coffee_data$productType_product)
# Set larger margin size
par(mar = c(5, 8, 4, 2))  # Adjust the right margin to fit longer labels
barplot(productType_product,
        main = "Combined productType_product Variables",
        ylab = "Frequency",
        las = 2,
        cex.names = 0.7)  # Rotate y-axis labels vertically
```

6.2.5 State

```
> #State Column
> state_df <- merged_coffee_data
>
> state_counts <- table(state_df$State)
> barplot(state_counts, main = "Bar Graph for State", xlab = "States", ylab = "Number",
cex.names = 0.5)
```

6.3 Statistical Analysis

Relation between Sales and ProductLine-Type:

```
> #Relation between Sales and ProductLine-Type
> #summary stats
> summary_prodLine_type <- aggregate(merged_coffee_data$Sales, by = list(merged_coffee_data$productLine_type),
FUN = summary)
> # Define custom colors for coffee and tea types
> coffee_colors <- c("#8B4513", "#A0522D")  # Brown and Sienna colors for coffee
> tea_colors <- c("#008000", "#3CB371")     # Green and Medium Sea Green colors for tea
> #visualise the data against sales
> boxplot(merged_coffee_data$Sales ~ merged_coffee_data$productLine_type,
+         main = "Profit by ProductLine-Type",
+         ylab = "log(Sales)",
+         xlab = "ProductLine-Type",
+         col = c(coffee_colors, tea_colors),
+         cex.axis = 0.7)  # Reduce font size of x-axis labels
> #ANOVA test
> anova_prodLine_type <- aov(merged_coffee_data$Sales ~ factor(merged_coffee_data$productLine_type))
> summary(anova_prodLine_type)
                                        Df  Sum Sq Mean Sq F value Pr(>F)
factor(merged_coffee_data$productLine_type)   3   52178   17393   2.943 0.0323 *
Residuals                                   826 4881571    5910
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Relation between Profit and ProductType-Product:

```
> # Relation between Profit and ProductType-Product
> # Summary stats
> summary_prodType_product <- aggregate(merged_coffee_data$Sales, by = list(merged_coffee_data$productType_prod
uct), FUN = summary)
> any(is.na(merged_coffee_data$productType_product))
[1] FALSE
> any(!is.numeric(merged_coffee_data$productType_product))
[1] TRUE
> unique(merged_coffee_data$productType_product)
 [1] "Herb_Lemon" "Herb_Mint"  "Tea_Darje"  "Tea_Green"  "Espr_Decaf" "Coff_Decaf" "Coff_Amare" "Coff_Colom"
 [9] "Espr_Caffe" "Herb_Chamo" "Tea_Earl "  "Espr_Regul"
> # Define coffee and tea colors
> coffee_colors <- c("#4A2C21", "#8B4513", "#654321", "#3E2723", "#6D4C41", "#5D4037")
> tea_colors <- c("#6B8E23", "#556B2F", "#8FBC8F", "#2E8B57", "#3CB371", "#20B2AA")
> # Get unique values in the productType_product column
> unique_values <- unique(merged_coffee_data$productType_product)
> # Create a named vector of coffee-tea palette
> coffee_tea_palette <- c(coffee_colors, tea_colors)
> names(coffee_tea_palette) <- unique_values[1:length(coffee_colors)]
> # Visualize the data against profit with coffee-tea palette
> boxplot(merged_coffee_data$Sales ~ merged_coffee_data$productType_product,
+         main = "log(Sales) by ProductType-Product",
+         ylab = "log(Sales)",
+         xlab = "ProductType-Product",
+         col = coffee_tea_palette,
+         cex.axis = 0.7)  # Reduce font size of x-axis labels
> #Anova test
> anova_prodType_product <- aov(merged_coffee_data$Sales ~ merged_coffee_data$productType_product)
> summary(anova_prodType_product)
                                     Df  Sum Sq Mean Sq F value Pr(>F)
merged_coffee_data$productType_product  11  174675   15880   2.729 0.0018 **
Residuals                              818 4759074    5818
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Barplot for DifferenceBetweenActualandTargetProfit(small and large markets) against Sales:

```
> #plot DifferenceBetweenActualandTargetProfit(small and large markets) against Sales
> # Define custom colors for small and large markets
> small_market_color <- "#FF6347"  # Tomato color for small market
> large_market_color <- "#4682B4"  # Steel Blue color for large market
> # Create a boxplot of Sales by Market_size
> boxplot(merged_coffee_data$DifferenceBetweenActualandTargetProfit ~ merged_coffee_data$Market_size,
+         main = "Diff by Market_size",
+         ylab = "Difference Between Actual and Target Profit",
+         xlab = "Market_size",
+         col = c(small_market_color, large_market_color),
+         cex.axis = 0.7)  # Reduce font size of x-axis labels
> var.test(merged_coffee_data$DifferenceBetweenActualandTargetProfit ~ merged_coffee_data$Market_size)

        F test to compare two variances

data:  merged_coffee_data$DifferenceBetweenActualandTargetProfit by merged_coffee_data$Market_size
F = 1.3143, num df = 302, denom df = 526, p-value = 0.006651
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.078579 1.610433
sample estimates:
ratio of variances
          1.314322

> t.test(merged_coffee_data$DifferenceBetweenActualandTargetProfit ~ merged_coffee_data$Market_size)

        Welch Two Sample t-test

data:  merged_coffee_data$DifferenceBetweenActualandTargetProfit by merged_coffee_data$Market_size
t = 4.6155, df = 562.24, p-value = 4.862e-06
alternative hypothesis: true difference in means between group Major Market and group Small Market is not equal
to 0
95 percent confidence interval:
  4.099551 10.173669
sample estimates:
mean in group Major Market mean in group Small Market
                 0.8481848                 -6.2884250
```

Correlation and ScatterPlot Matrix for numerical variables:

```
> # Calculate the correlation matrix
> correlation_matrix <- cor(merged_coffee_data[c("Profit", "log(Sales)", "log(Cogs)", "Difference", "Inventory.Margin", "Margin", "log(Total_expense
s)")])
> # Print the correlation matrix
> print(correlation_matrix)
                      Profit log(Sales) log(Cogs)  Difference
Profit             1.00000000  0.7761674 0.7214713  0.53817580
log(Sales)         0.77616738  1.0000000 0.9826169  0.31391566
log(Cogs)          0.72147128  0.9826169 1.0000000  0.25217824
Difference         0.53817580  0.3139157 0.2521782  1.00000000
Inventory.Margin   0.43206145  0.4227573 0.4758868  0.13085245
Margin             0.78039658  0.9874984 0.9502260  0.30857574
log(Total_expenses) 0.03656469  0.6236377 0.6163241 -0.04567231
                   Inventory.Margin    Margin log(Total_expenses)
Profit                   0.43206145 0.7803966          0.03656469
log(Sales)               0.42275727 0.9874984          0.62363774
log(Cogs)                0.47588680 0.9502260          0.61632410
Difference               0.13085245 0.3085757         -0.04567231
Inventory.Margin         1.00000000 0.3733448          0.03842745
Margin                   0.37334478 1.0000000          0.61133278
log(Total_expenses)      0.03842745 0.6113328          1.00000000
> # Create scatterplot matrix with correlation coefficients below the diagonal
> pairs(merged_coffee_data[c("Profit", "log(Sales)", "log(Cogs)", "Difference", "Inventory.Margin", "Margin", "log(Total_expenses)")],
+       upper.panel = function(x, y) {
+         points(x, y)
+         cor_val <- round(cor(x, y), 2) # Calculate correlation coefficient
+         text(mean(x, na.rm = TRUE), mean(y, na.rm = TRUE), labels = NULL, col = "red")
+       },
+       lower.panel = function(x, y) {
+         cor_val <- round(cor(x, y), 2)  # Calculate correlation coefficient
+         label <- sprintf("r=%.2f", cor_val)  # Format label as "r=0.78"
+         color <- ifelse(cor_val > 0.5, "red", "blue")  # Set color based on correlation coefficient
+         text(mean(x, na.rm = TRUE), mean(y, na.rm = TRUE), labels = label, col = color, cex = 1.2, adj = c(0.3, 0.5))  # Display label with color
+       }
```

Time analysis plot for log(sales) against date:

```
# Convert Date column to Date format
merged_coffee_data$Date <- as.Date(as.character(merged_coffee_data$Date), format = '%m/%d/%Y')
# Create character column with a custom date format
merged_coffee_data$newDate <- strftime(merged_coffee_data$Date, '%d%m%Y')
# Fit a linear regression model with Sales as the dependent variable
salesdate_model <- lm(Sales ~ Date, data = merged_coffee_data)
# Plot the Sales against Date
plot(merged_coffee_data$Date, merged_coffee_data$Sales,
     type = "l",  # 'l' for line plot
     xlab = "Date", ylab = "Sales",      # labels for x-axis and y-axis
     main = "log(Sales) vs Date")
# Plot the linear regression line
lines(merged_coffee_data$Date, predict(salesdate_model), col = "red")
# Add confidence intervals for the regression line if needed
ci_salesdate <- predict(salesdate_model, interval = "confidence")
lines(merged_coffee_data$Date, ci_salesdate[, "lwr"], col = "blue", lty = 2)
lines(merged_coffee_data$Date, ci_salesdate[, "upr"], col = "blue", lty = 2)
```

```
> summary(salesdate_model)

Call:
lm(formula = Sales ~ Date, data = merged_coffee_data)

Residuals:
     Min       1Q   Median       3Q      Max
-185.768  -38.895   -4.512   33.984  223.685

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.136e+03  1.211e+02  -17.64   <2e-16 ***
Date         1.420e-01  7.522e-03   18.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64.55 on 828 degrees of freedom
Multiple R-squared:  0.3008,    Adjusted R-squared:    0.3
F-statistic: 356.2 on 1 and 828 DF,  p-value: < 2.2e-16
```

Linear Regression

```r
#sales and cogs
# Fit a linear regression model
model_sales_cogs <- lm(Sales ~ Cogs, data = merged_coffee_data)
summary(model_sales_cogs)

# Assess R-squared value
r_squared_sales_cogs <- summary(model_sales_cogs)$r.squared
cat("R-squared for log(sales) vs log(cogs)", r_squared_sales_cogs, "\n")

# Plot residuals
plot(model_sales_cogs$fitted.values, resid(model_sales_cogs),
     xlab = "log(cogs)", ylab = "log(sales)",
     main = "Residual plot: log(sales) vs log(cogs)")
# Optionally, you can add a horizontal line at y = 0 for reference
abline(h = 0, col = "red")

#plot qq
qqnorm(resid(model_sales_cogs),
       xlab = "log(cogs)", ylab = "log(sales)",
       main = "QQ plot: log(sales) vs log(cogs)")
qqline(resid(model_sales_cogs), col = "blue")  # Add a line to the Q-Q plot

# Extract coefficients
slope_sales_cogs <- coef(model_sales_cogs)["Cogs"]
intercept_sales_cogs <- coef(model_sales_cogs)["(Intercept)"]

# Plot the data points
plot(merged_coffee_data$Cogs, merged_coffee_data$Sales,
     xlab = "log(cogs)", ylab = "log(sales)",
     main = "Linear Regression: log(sales) vs log(cogs)")

# Add the regression line
abline(a = intercept_sales_cogs, b = slope_sales_cogs, col = "red")
print(intercept_sales_cogs)
print(slope_sales_cogs)

#sales and margin
# Fit a linear regression model
model_sales_margin <- lm(Sales ~ Margin, data = merged_coffee_data)
summary(model_sales_margin)

# Assess R-squared value
r_squared_sales_margin <- summary(model_sales_margin)$r.squared
cat("R-squared for log(sales) vs margin:", r_squared_sales_margin, "\n")

# Plot residuals
plot(model_sales_margin$fitted.values, resid(model_sales_margin),
     xlab = "margin", ylab = "log(sales)",
     main = "Residual plot: log(sales) vs margin")
abline(h = 0, col = "red")
```

```r
qqnorm(resid(model_sales_margin),
       xlab = "Margin", ylab = "log(sales)",
       main = "QQ plot: log(sales) vs margin")
qqline(resid(model_sales_margin), col = "blue")  # Add a line to the Q-Q plot

# Extract coefficients
slope_sales_margin <- coef(model_sales_margin)["Margin"]
intercept_sales_margin <- coef(model_sales_margin)["(Intercept)"]

# Plot the data points
plot(merged_coffee_data$Margin, merged_coffee_data$Sales,
     xlab = "Margin", ylab = "log(sales)",
     main = "Linear Regression: log(sales) vs Margin")

# Add the regression line
abline(a = intercept_sales_margin, b = slope_sales_margin, col = "red")
print(intercept_sales_margin)
print(slope_sales_margin)

#sales and log(total_expenses)
# Fit a linear regression model
model_sales_total <- lm(Sales ~ Total_expenses, data = merged_coffee_data)
summary(model_sales_total)

# Assess R-squared value
r_squared_sales_total <- summary(model_sales_total)$r.squared
cat("R-squared for log(sales) vs log(total expenses):", r_squared_sales_total, "\n")

# Plot residuals
plot(model_sales_total$fitted.values, resid(model_sales_total),
     xlab = "log(total expenses)", ylab = "log(sales)",
     main = "Residual plot: log(sales) vs log(total expenses)")
abline(h = 0, col = "red")

qqnorm(resid(model_sales_total),
       xlab = "log(total expenses)", ylab = "log(sales)",
       main = "QQ plot: log(sales) vs log(total expenses)")
qqline(resid(model_sales_total), col = "blue")  # Add a line to the Q-Q plot

# Extract coefficients
slope_sales_total <- coef(model_sales_total)["Total_expenses"]
intercept_sales_total <- coef(model_sales_total)["(Intercept)"]
print(model_sales_total)

# Plot the data points
plot(merged_coffee_data$Margin, merged_coffee_data$Sales,
     xlab = "log(total expenses)", ylab = "log(sales)",
     main = "Linear Regression: log(sales) vs log(total expenses)")

# Add the regression line
abline(a = intercept_sales_total, b = slope_sales_total, col = "red")
print(intercept_sales_total)
print(slope_sales_total)
```

Multiple Linear Regression:

```
> #Multiple Linear Regression
> model <- lm(`log(Sales)` ~ `log(Cogs)` + Difference + Inventory.Margin + Margin + `log(Total_expenses)`, data = merged_coffee_data)
> step(model, direction = "backward")
Start:  AIC=2548.36
`log(Sales)` ~ `log(Cogs)` + Difference + Inventory.Margin +
    Margin + `log(Total_expenses)`

                       Df Sum of Sq    RSS    AIC
- Inventory.Margin      1         6  17635 2546.7
<none>                              17629 2548.4
- `log(Total_expenses)` 1       941  18569 2589.5
- Difference            1      5231  22860 2762.0
- `log(Cogs)`           1     76990  94619 3941.0
- Margin                1    113098 130727 4209.3

Step:  AIC=2546.66
`log(Sales)` ~ `log(Cogs)` + Difference + Margin + `log(Total_expenses)`

                       Df Sum of Sq    RSS    AIC
<none>                              17635 2546.7
- `log(Total_expenses)` 1      1009  18644 2590.8
- Difference            1      5225  22860 2760.1
- `log(Cogs)`           1    100143 117778 4120.8
- Margin                1    120485 138120 4253.0

Call:
lm(formula = `log(Sales)` ~ `log(Cogs)` + Difference + Margin +
    `log(Total_expenses)`, data = merged_coffee_data)

Coefficients:
          (Intercept)           `log(Cogs)`            Difference                Margin  `log(Total_expenses)`
              2.28377               1.08047               0.13327               0.93883               0.05961
```

7. References

Amrutha yenikonda. 2023. Coffee Chain Sales Analysis. Kagglecom. [accessed 2024 Mar 27]. https://www.kaggle.com/datasets/amruthayenikonda/coffee-chain-sales-dataset.