FinalProject_Step3_PuppalaSucharitha

2022-08-11

1. INTRODUCTION.

- In this paper I would like to discuss about the research on "COFFEE".Before getting into the data analysis I would like to give a quick introduction about "COFFEE".
- Coffee, is brewed from the roasted and ground seeds of the tropical evergreen coffee plants of African origin. Coffee is one of the three most popular beverages in the world alongside water and tea, and one of the most profitable international commodities. Though coffee is the basis for an endless array of beverages, its popularity is mainly attributed to its invigorating effect, which is produced by caffeine, an alkaloid present in coffee.
- For coffee drinkers, there's no denying the feeling of waking up to a good cup of joe, or snatching one at your favorite coffee shop while out and about. With the many options when it comes to coffee beverages, the multiple flavors and aromas, it's no wonder more and more people are becoming fans of the coffee industry.
- Coffee lovers around the world who reach for their favorite morning brew probably aren't thinking about its health benefits or risks.
- Coffee is a rich source of

1. Caffeine
2. Vitamin B2 (riboflavin)
3. Magnesium
4. Plant chemicals: polyphenols including chlorogenic acid and quinic acid,

and diterpenes including cafestol and kahweol.

- In 2022, the number of coffee drinkers in America rose to 66%. This makes coffee the most popular beverage thanks to a 14% increase since January of 2021.
- I have selected three data sets that contains the data of the coffee survey, coffee chain and the caffeine content of coffee.All the data sets collected contains the data of America.
- The data sets are collected from Kaggle.com and data.world.

2. SUMMARIZE THE PROBLEM STATEMENT YOU ADDRESSED.

- While Americans may love coffee, it's clear that other countries enjoy it as well. The United States ranks 25th when it comes to countries that consume the most coffee per capita.
- In 2022, E-Imports reports that American coffee drinkers are drinking more coffee per day. In recent years, it was reported that coffee drinkers enjoyed an average of 2 to 3 cups per day. Now, coffee drinkers are downing 3.1 cups a day on average.
- A moderate amount of coffee is generally defined as 3-5 cups a day, or on average 400 mg of caffeine, according to the Dietary Guidelines for Americans.

*It may not be a surprise to learn that the Number One state for drinking coffee is New York. New Yorkers sure love their coffee. In NYC, there seems to be a coffee shop on every corner. They not only consume more of it than any other state, but they also pay the most for a cup of cappuccino compared to other states.

- With all the above information, in this analysis I would like to know the reason for consumption of coffee,category of people who consume more coffee, the category of drink that has high amount of caffeine, which state in USA consumes more amount of coffee and has more profit.

3. HOW YOU ADDRESSED THIS PROBLEM STATEMENT?

- In this analysis the problem statement I have taken to address is ' What is the reason for consumption of coffee?,Which category of people consume more coffee?,Which drink has highest amount of caffeine?'
- To address the above problem statement have selected three data sets.

1. Coffee Survey data set - This data set contains the survey of the people who drink coffee, the reason for coffee drinking,how many cups they drink,does coffee drinking works.
2. Caffeine data set - This data set has the details of different types of drinks and their caffeine content,calories of each drink,volume of the drink.
3. Coffee chain data set - This data set contains the details of the coffee consumed in the states of America and the share of the amount of coffee they have consumed, the profits obtained by each state ,etc.

- From the three data sets collected, I have analysed the three data sets individually and grouped the analysis for the final conclusion.The following are the steps that are followed in the analysis.
- From all the data sets collected I have selected the variables that are helpful in each data set for the analysis.
- After selecting the variables, found that few additional data is required, i.e. the amount of caffeine in some herbal tea, etc. which were not available in the data sets collected. The required additional data is collected and the inserted the data in the required columns and rows.

For analyzing the data to arrive at the problem statement the below steps are followed:

- Imported the required libraries.
- Next working directory is set and the CSV file is read using the read.csv().
- The data is clean by checking for any duplicates, null values etc.
- Once the data is cleaned it is ready for the analysis.
- Using the "ggplot2" package I have plotted scatter plots, boxplots , bar graphs for different variables from the three datasets.
- I have used the linear regression model and logistic regression model and performed analysis.
- I would suggest logistic regression model best fits for the analysis of all the three data sets.

4. SUMMARIZE THE INTERESTING INSIGHTS THAT YOUR ANALYSIS PROVIDED.

- In some cases, visualizations are much better than plain numbers at conveying information. The relationships among variables, the distribution of variables, and underlying structure in data can easily be discovered using data visualization techniques.
- In my analysis of the three data sets I found data visualizations provided a unique perspective on the data sets that I have collected. I have visualized the data in different ways.

- With the help of the scatter plots , boxplots, and bar graphs I was able to arrive at the problem statement before computing the statistical analysis.
- I have analysed the reason for consumption of more coffee in USA with the help of scatter plot, bar graph and boxplot.For this I have used the coffee survey data set and with the output obtained by the analysis ,the maximum count for the reason for the consumption of more coffee is to relieve the Study Stress and next is to have refreshing every morning, and next comes the living habits. With the help of the data set we can also say that the students category are consuming high amount of coffee to get relieved from the stress.Below are visualizations that are done on the data sets collected.
- With the help of the caffeine data set collected I have analysed for the drink that has the highest amount of caffeine. The data set contains the different types of drinks and different varieties in each type of drink.The data set contains the data of Coffee, Energy drinks, Energy shots,Soft Drinks, Tea and Water.In my analysis I found that the coffee has different varieties that has the highest amount of caffeine.
- When considering the analysis of the coffee chain data set, I have added an extra variable that is the amount of caffeine that is present in each type of coffee drink , which when added helped me in knowing more about the varieties of the coffee based on the caffeine content.
- Some of the visualizations on the coffee survey data set, caffeine data set and the coffee chain data sets data variables are below.
- These plots helped me in visualizing the problem statement in a easily understandable way.
- I have also used the logistic regression model for the summary analysis of the data sets collected.

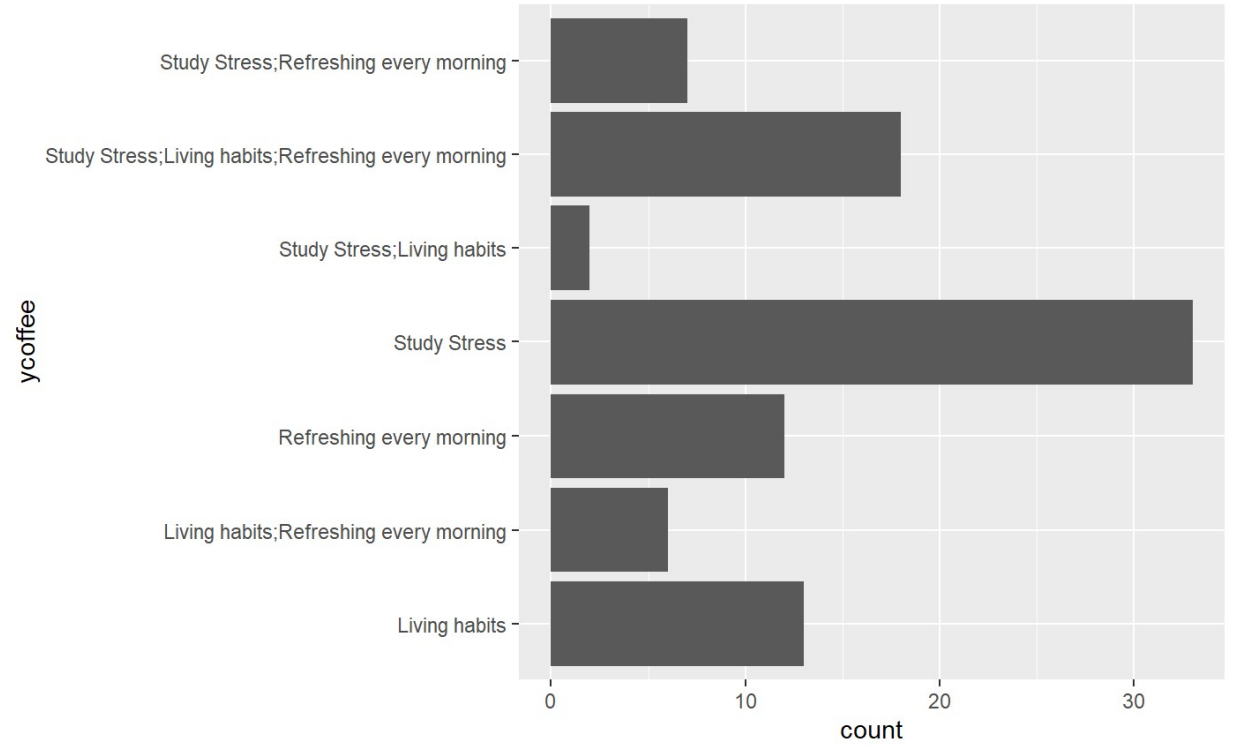

Figure 1: Bar graph for Ycoffee

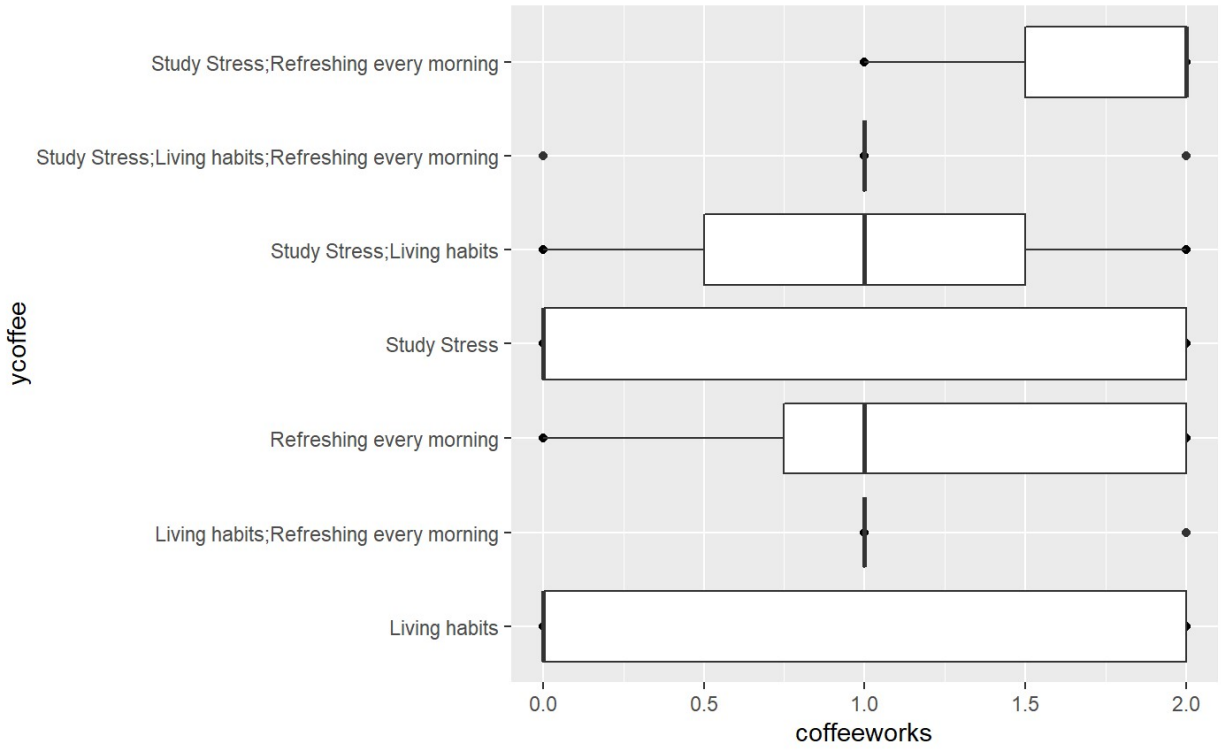

Figure 2: ycoffee vs coffee works

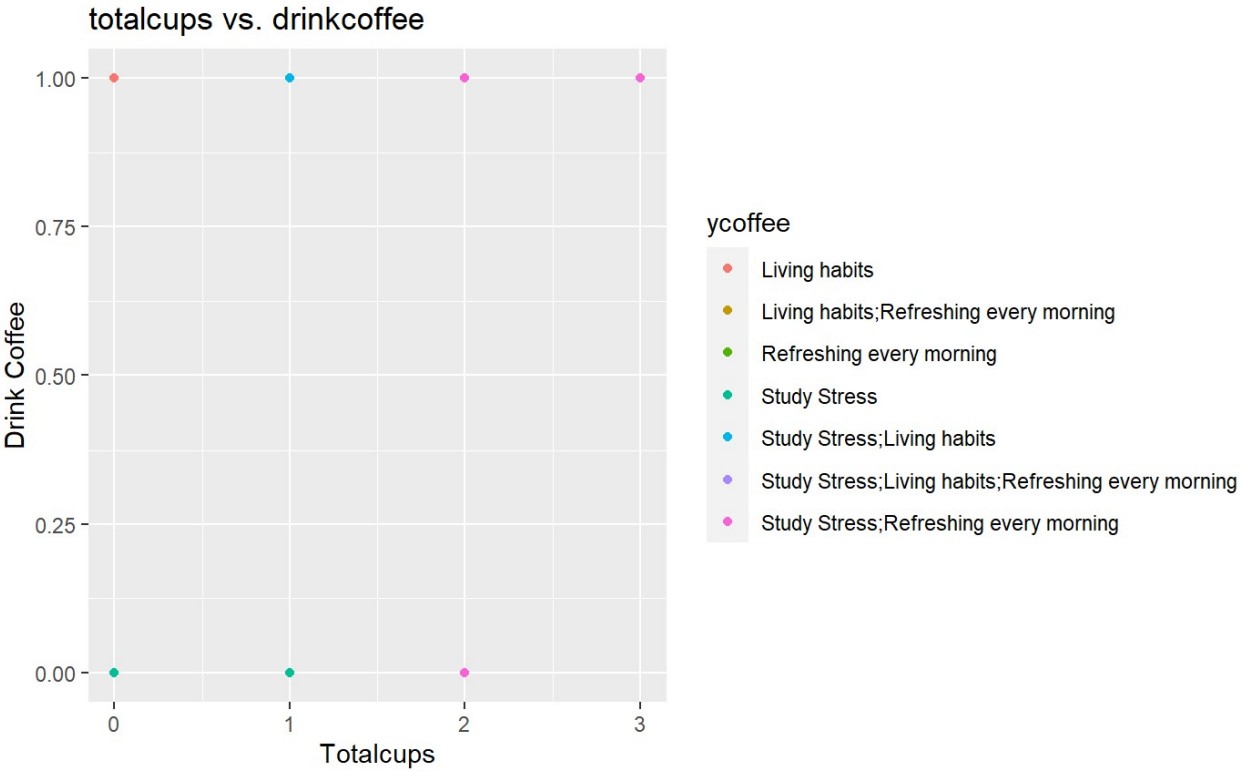

Figure 3: Scatter Plot for Total Cups,Drink coffee and Ycoffee

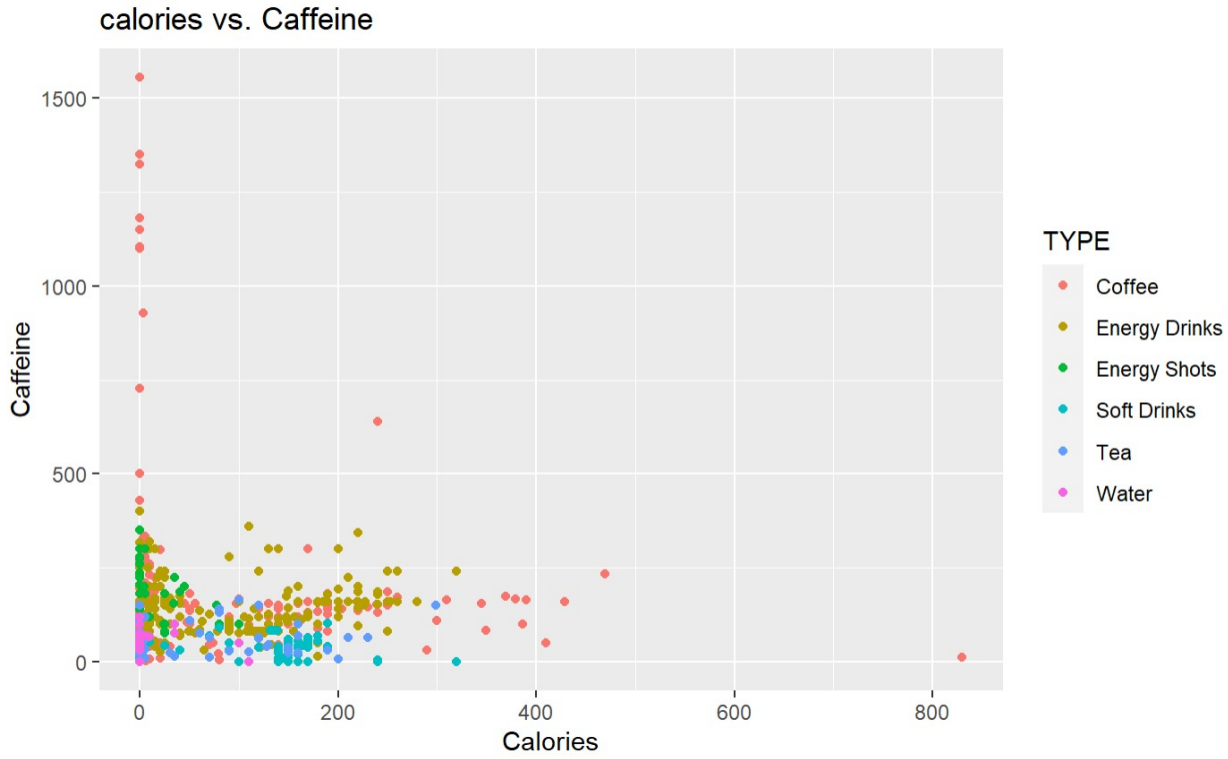

Figure 4: Scatter Plot for Calories,Caffeine and Type

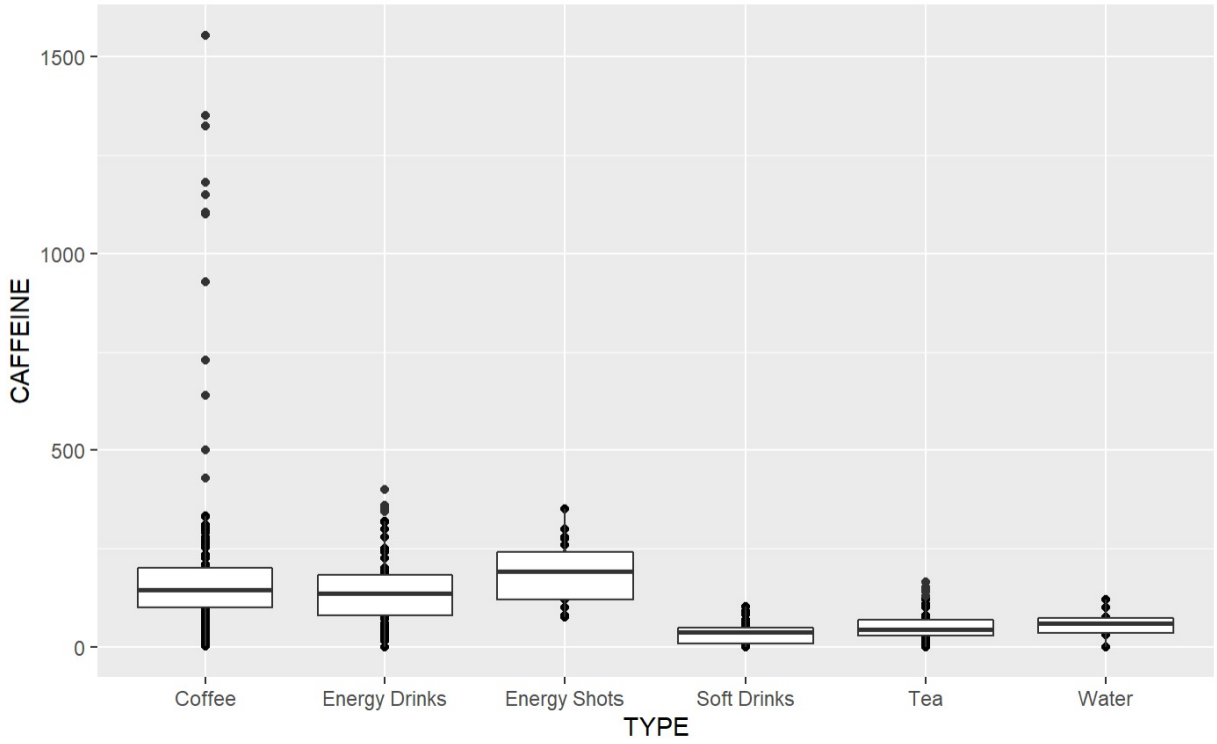

Figure 5: Box Plot for Drink vs Caffeine

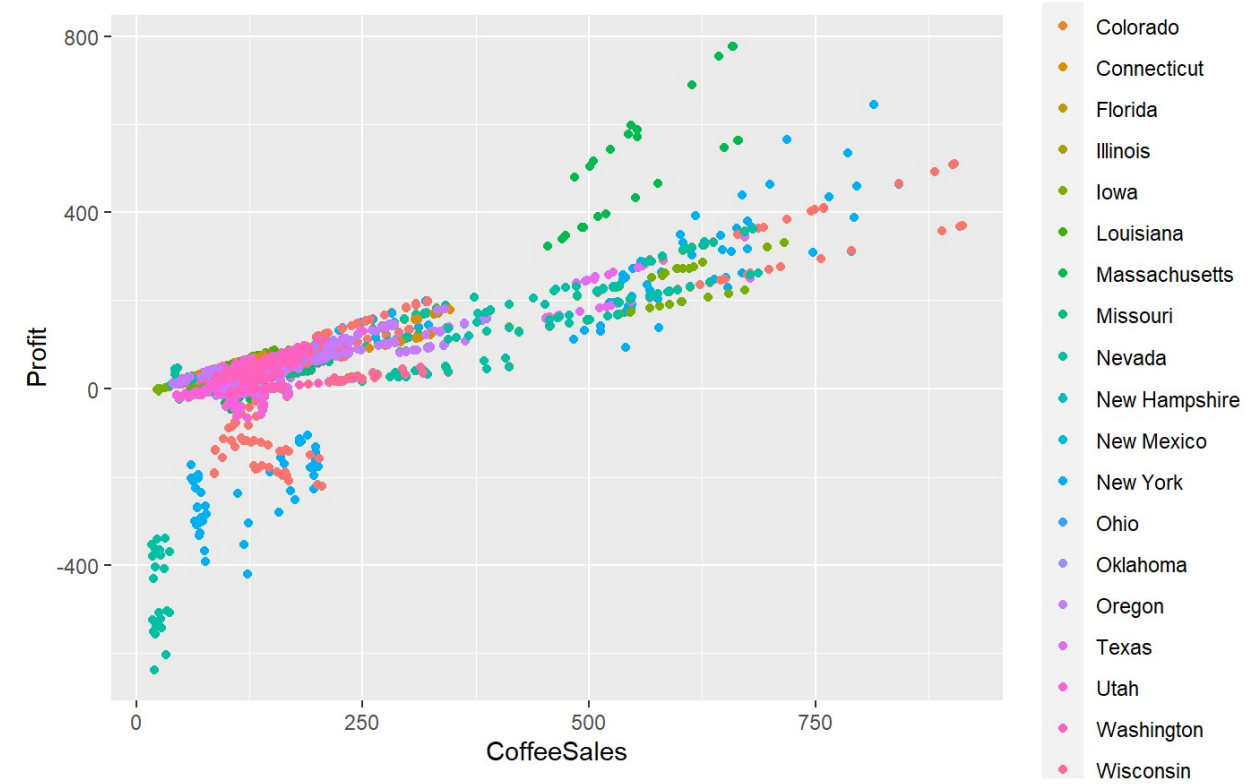

Figure 6: Scatter Plot for CoffeeSales,Profit and State

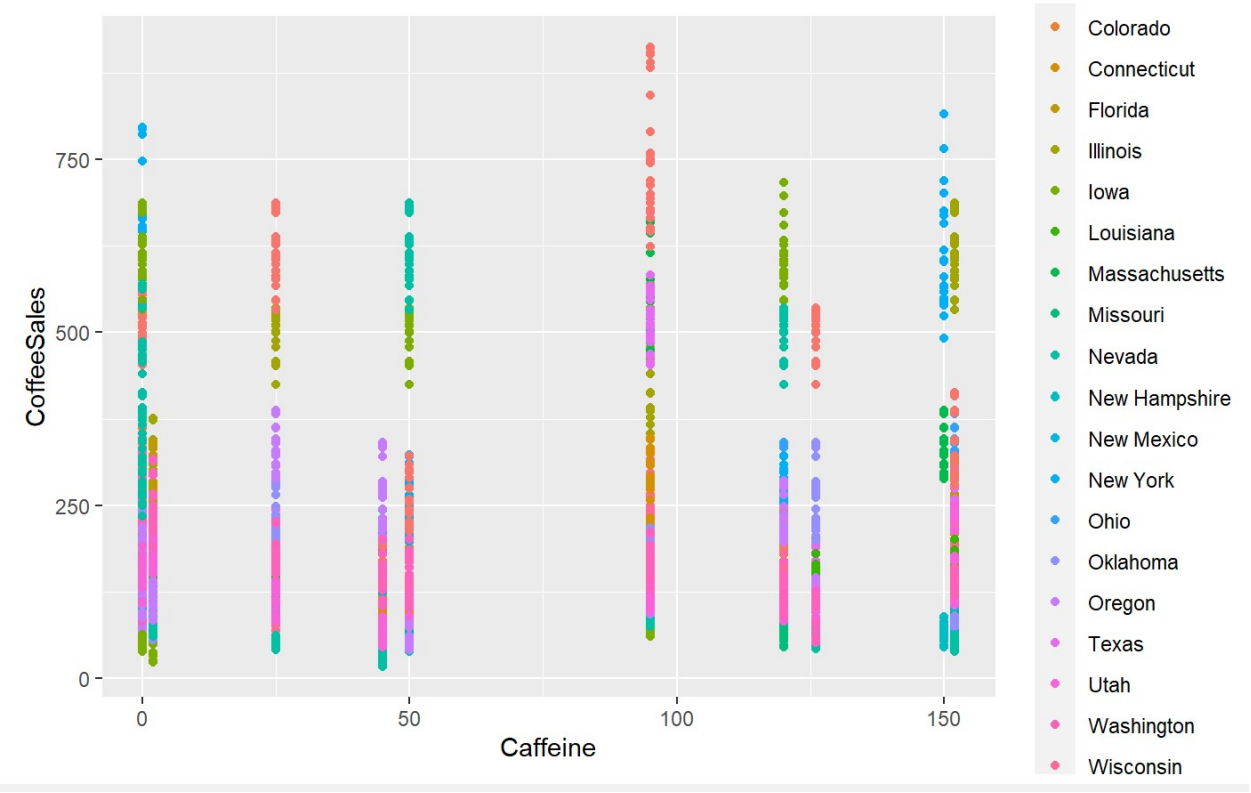

Figure 7: Scatter Plot for Caffeine, Coffee Sales and State

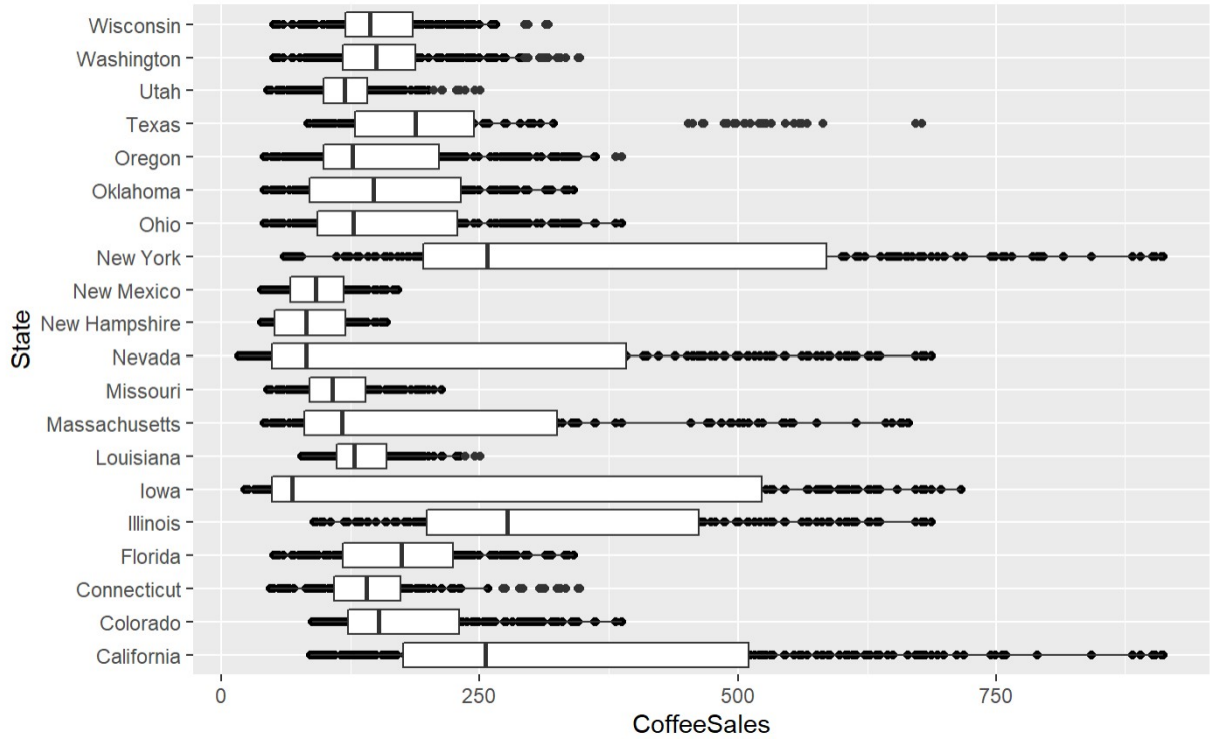

Figure 8: Boxplot for Coffeesales and State.

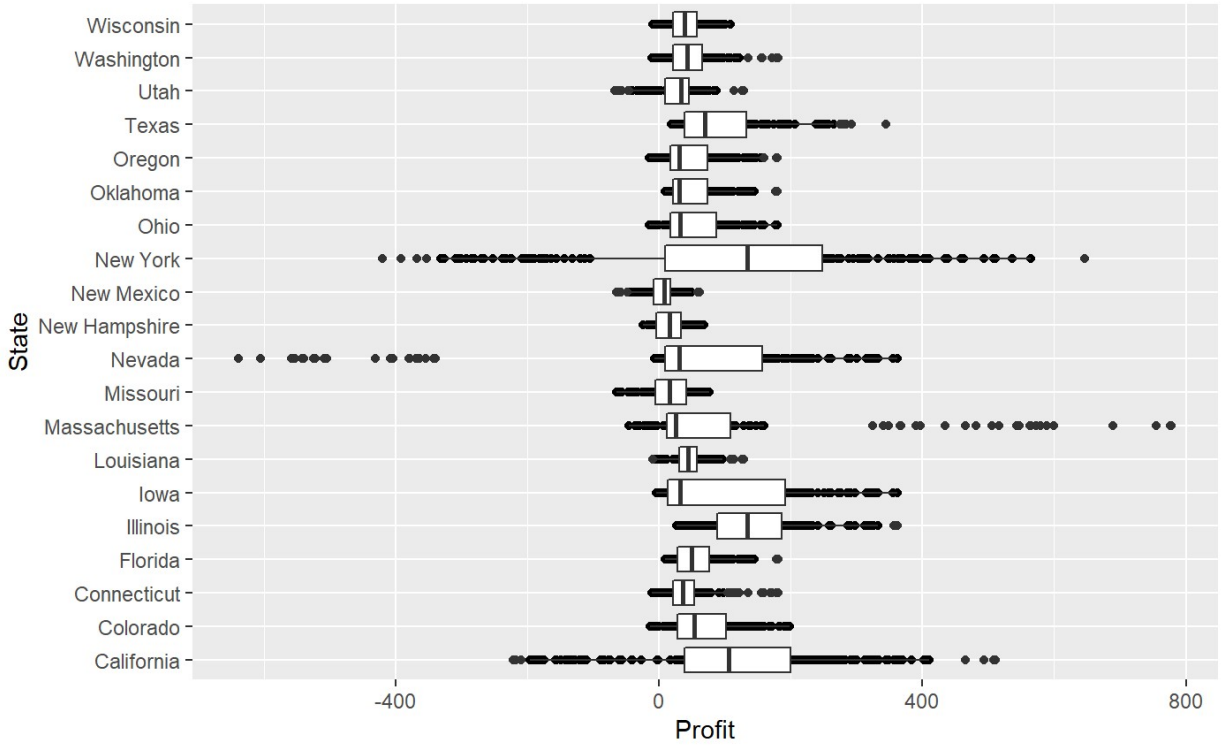

Figure 9: Boxplot for State and Profit.

5.SUMMARIZE THE IMPLICATIONS TO THE CONSUMER OF YOUR ANALYSIS.

- This is the analysis of the consumption of coffee, amount of caffeine content in different drinks, states that consumes highest amount of coffee and the amount of profit obtained by the coffee consumption.
- All the analysis is based on the data collected based on survey, actual figures, etc.
- With the help of the analysis the coffee shop brands can select the area where the business will be increased base on the target customers, and can also know in which state the coffee business has been increased from the previous years business and the amount of business that can be increased can be predicted.
- The amount of caffeine content coffee that has more sales in each state can be known, which helps in preparing the Coffee shop menu that contains different varieties of coffee depending on the caffeine values, which indirectly increases the customers in more frequent buying of coffee and thus increasing the business and earning profit.

6. LIMITATIONS OF YOUR ANALYSIS AND HOW YOU, OR SOMEONE ELSE COULD IMPROVE

- While analyzing all the three data sets in arriving at the solution for the problem statement I felt that the analysis could have been more clear if the survey details has been taken from each state separately, which would have helped in predicting the actual category of people.
- The age factor can also play an important role in the survey, which when added can be helpful in knowing the tastes and preferences of the people age wise.

- The model would have been more perfect if I have considered the age wise data and the state wide data individually to data and have analysed the data.

7. CONCLUSIONS.

- With the above analysis done on the three data sets collected we can say that the state NewYork in the USA has the highest amount of coffee sales, the study stress and the living habits are the reasons for the consumption of more coffee in America.

*The majority of regular coffee drinkers report drinking it for the flavor, not the caffeine. Coffee has different varieties that has different percentages of caffeine in them that changes the flavor of each type of coffee drink.

- Coffee is a global phenomenon. The growth and change in the coffee industry each year show that more people are showing interest in this delicious drink, and the different varieties of coffee beverages gaining popularity, to make it through their day.