1. **Dataset Background**

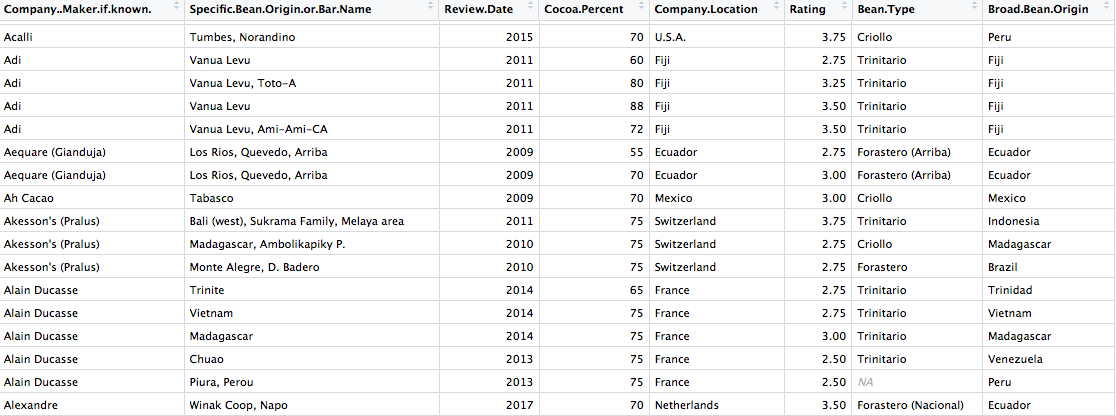
The dataset contains expert ratings of 1,795 individual chocolate bars from year 2006 to 2017, and all the data are collected from the Kaggle. The original dataset has 5 categorical data (company name, company location, origin, bean type and broad bean origin) and 4 numerical data (review date, cocoa percent, rating and ref). For this particular data analysis, the columns of broad bean origin and ref are excluded.

1. **Objective**

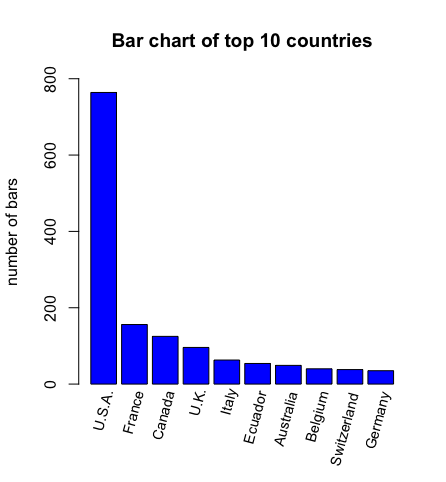
The goal of this data analysis is to get more detailed information about countries produce most number of chocolate bars; regions grow the best cocoa beans; popular bean types; and average percentage of cocoa of each bar.

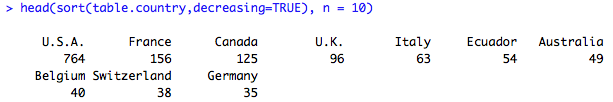
1. **Data Preparation**





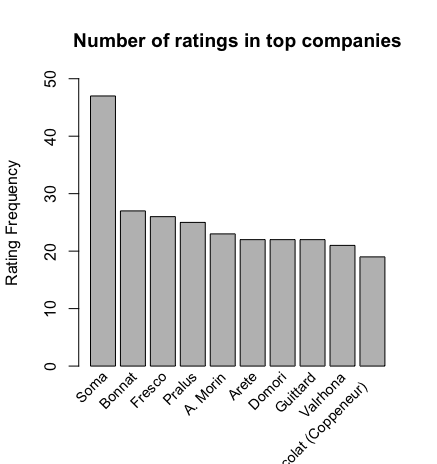
1. **Data Analysis**
2. **Categorical Data**
3. Country

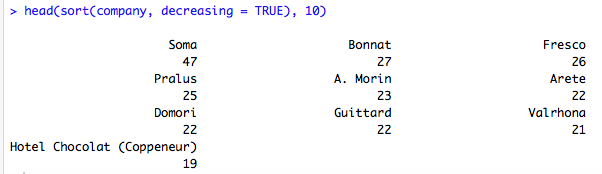




The above results show that top 10 countries which have the most numbers of chocolate vendors from the dataset, and U.S.A. has the most chocolate vendors which is about 764.

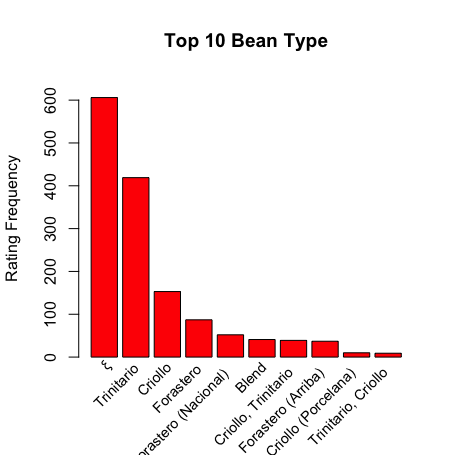
1. Company

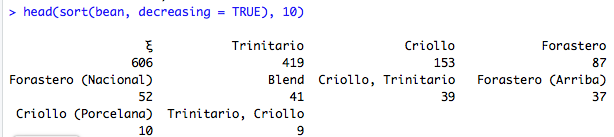




For the companies, the chart lists the most ratings of top 10 countries. Among them, Soma, which is located in Canada, is the top one.

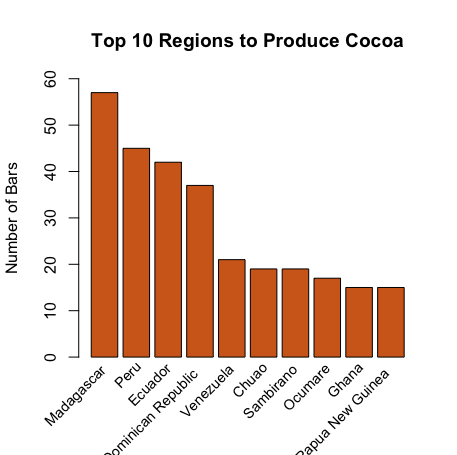
1. Bean Type

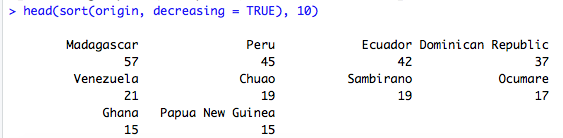




The bar plot above shows the top 10 most popular bean types based on the ratings. The top one which has 606 rating has unknown name from the dataset, and the next popular one is Trinitario which is the world’s finest cocoa hybrid.

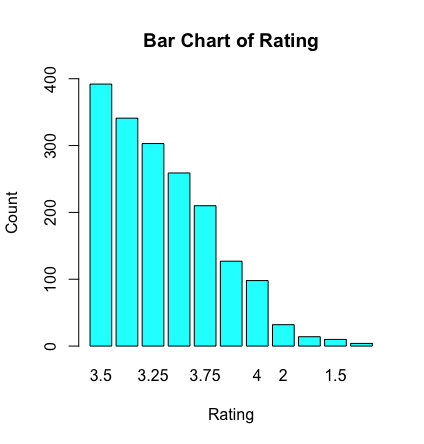
1. Origin of Bean Types

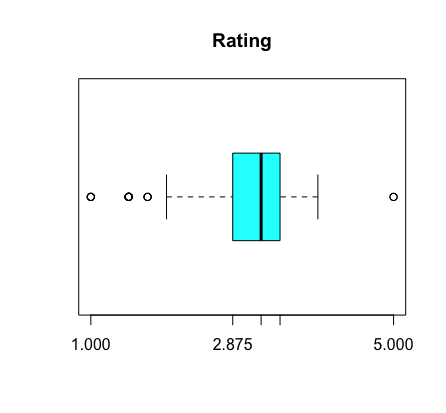


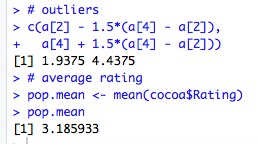


The chart lists the top 10 regions that produce the most number of cocoa beans, and Madagascar is the top 1.

1. **Numerical Data**
2. Rating



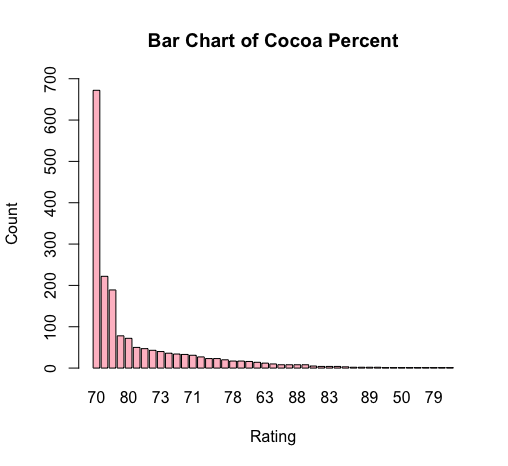




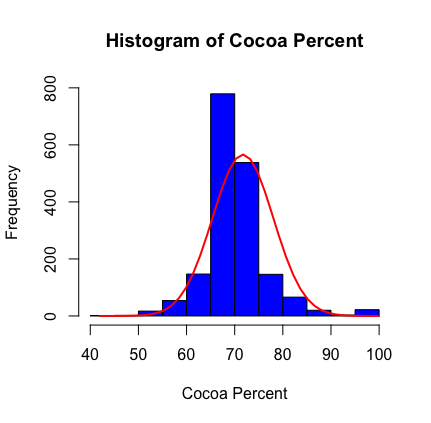
The data analysis of the numerical variable rating shows the rang of the rating is between 1 to 5, and most of the ratings are 3.5. The mean of the ratings is about 3.185933.

1. Cocoa Percent

The bar chart indicates that the range of cocoa percentage is between 50 to 89, and the most chocolate bars contain about 70% of cocoa beans.



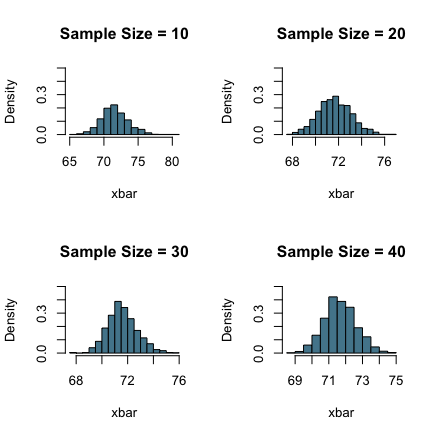
1. **Data Distribution of Percentage of Cocoa Beans**

****

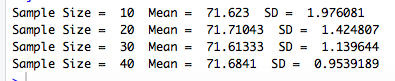
As the histogram and the curve show that the data distribution of the percentage of cocoa beans is normal distribution, and most of the values are within the range of the mean.

1. **Central Limit Theorem**

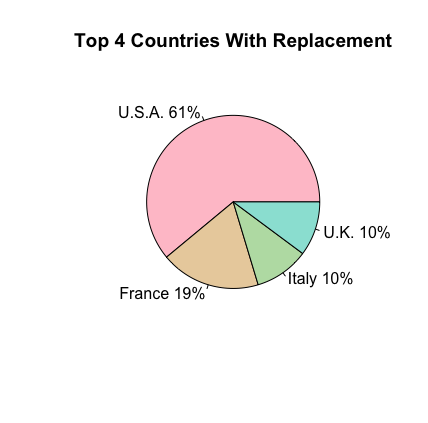
The analysis has the sample means of 1000 random samples with four different sample size which are 10, 20, 30 and 40 show a normal distribution, and the original population is the percentage of cocoa beans.



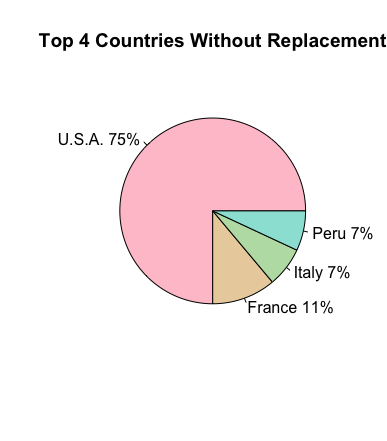
pop.mean (cocoa.percent) = 71.69833

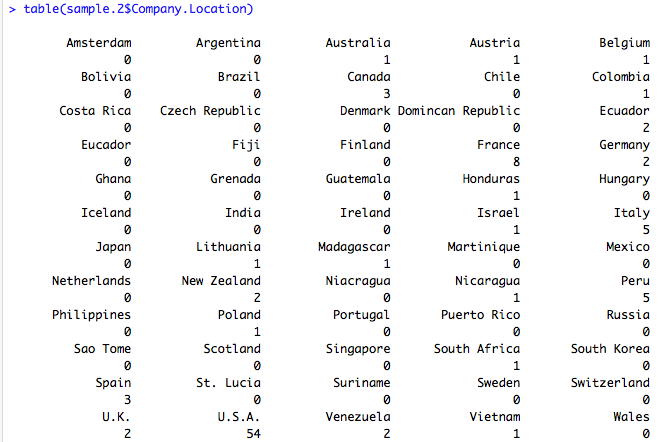


1. **Sampling Methods**
2. Simple Random Sampling





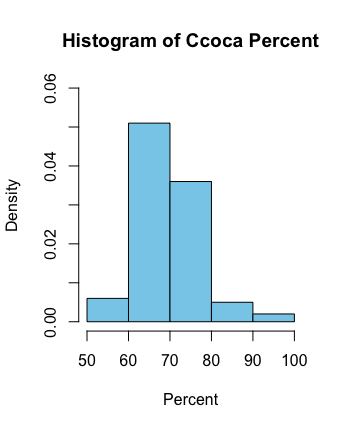


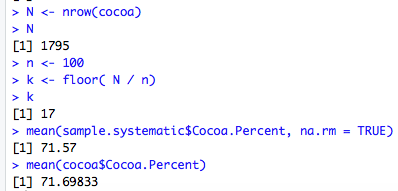


The sample size is 200, and the original data is from company location column which has 1795 data frames. Compared with the population data, the result of simple random sampling with replacement is identical; whereas, the result of simple random sampling without replacement has the fourth country is Peru instead of U.K. In this case, the method of simple random sampling with replacement is more accurate for the prediction.

1. Systematic Sampling

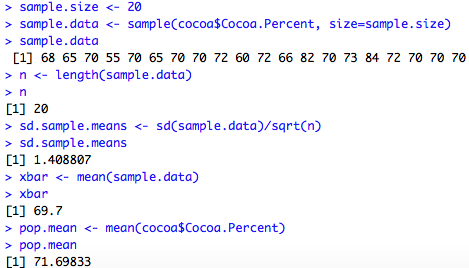
For the systematic sampling, the analysis has 17 subsets of the percentage of cocoa beans, and each subset contains 100 values of the data. According to the histogram and the result, the sample size also follows a normal distribution with a sample mean which is 71.57, close to the population mean which is 71.69833.

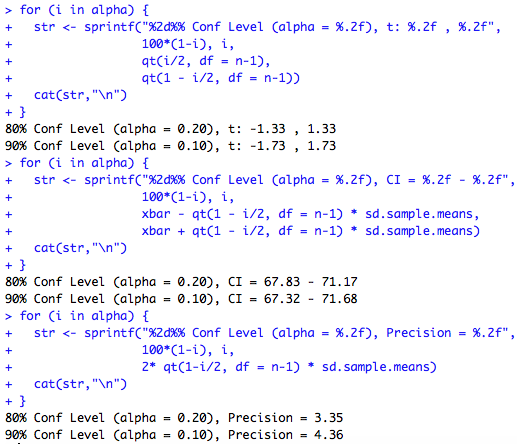




1. **Confidence Intervals of 80 and 90**

For the analysis of confidence intervals, a simple size of 20 is chosen from the percentage of cocoa beans. As the results show that the 90% confidence interval is more accurate to predict the population mean, and the interval is between 67.32 to 71.68. Also, the precision increases with the increasing of confidence levels.





References

Kalathur, S. (2017). *CS544 Module 3*. Retrieved from

<https://onlinecampus.bu.edu/bbcswebdav/pid-5845767-dt-content-rid-21345546_1/courses/18sprgmetcs544_a1/CS544_Module3.pdf>

Kalathur, S. (2017). *CS544 Module 4*. Retrieved from

<https://onlinecampus.bu.edu/bbcswebdav/pid-5886572-dt-content-rid-21702699_1/courses/18sprgmetcs544_a1/CS544_Module4.pdf>

Kalathur, S. (2017). *CS544 Module 5*. Retrieved from

<https://onlinecampus.bu.edu/bbcswebdav/pid-5911469-dt-content-rid-21899288_1/courses/18sprgmetcs544_a1/CS544_Module5.pdf>

Kalathur, S. (2017). *CS544 Module 6*. Retrieved from

<https://onlinecampus.bu.edu/bbcswebdav/pid-5917979-dt-content-rid-21982038_1/courses/18sprgmetcs544_a1/CS544_Module6.pdf>

Tatman R. (2017). Chocolate Bar Ratings. *Kaggle.* Retrieved from

<https://www.kaggle.com/rtatman/chocolate-bar-ratings>