UGANDA MARTYRS UNIVERSITY

FACULTY OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION SYSTEMS

ENTERPRISE BUSINESS ANALYTICS: TAKE HOME PROJECT

Name	Registration No.	Student No.
MUSINGUZI JIMMY	**2023-M132-21504**	**2300501504**

Table of Contents

List of Figures

ENTERPRISE BUSINESS ANALYTICS: TAKE HOME PROJECT

1.0 Introduction

This project used the *Coffee Quality Dataset*, collected by the Coffee Quality Institute in January 2018. The data was collected on Arabica coffee beans from across the world and professionally rated on a 0-100 scale based on factors like acidity, sweetness, fragrance, balance, etc. The dataset also contained information about coffee bean origin country, harvesting and grading date, colour, defects, processing and packaging details [2].

Analyzing this dataset is very crucial since Uganda is one of the leading exporters of Arabica Coffee which is extremely competitive on the international market. The crop is mainly grown in slightly raised areas and in Uganda it is mainly grown in areas like the Rwenzori Mountains, mountain Elgon and the west Nile region. The coffee varieties that are grown in Uganda include Robusta-screen 15, Arabica-Bugisu A, Arabica- Bugisu AA, Arabica-Bugisu PB, Arabica –Drugar and Arabica-Bugisu B. In 2023, Uganda exported 74,235 bags of Arabica valued at US$16.68 million [1].

1.1 Project Data Analysis Question

The data analytics question that this project is attempting to answer is: "**Given a dataset with a set of characteristics, what is Uganda's competitive advantage over other arabica coffee exporting countries?"**

1.2 Objective of the assignment

The main objective of the assignment is to analyze the quality of arabica coffee beans exported globally and get insights on how to be competitive as a country on the arabica coffee global market.

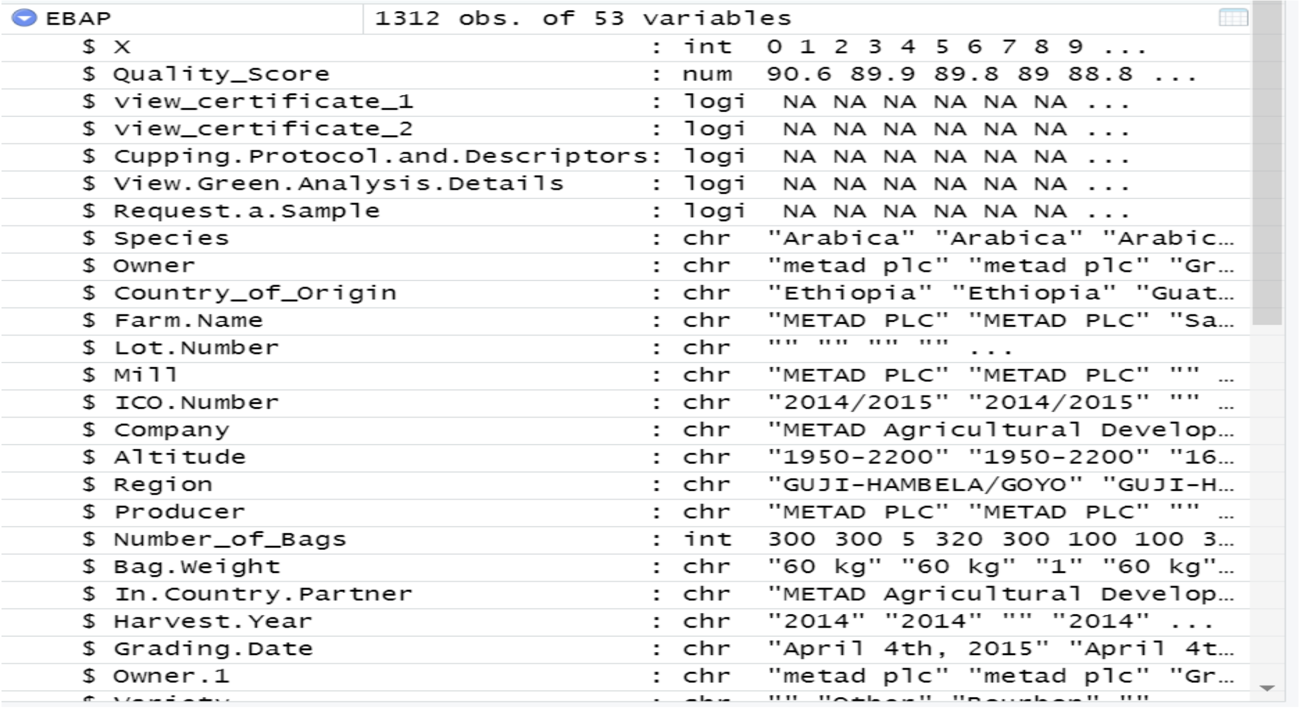
1.3 Description of the dataset

This project uses the Coffee Quality Dataset, collected by the *Coffee Quality Institute* in January 2018. The data was retrieved from tidytuesday, courtesy of James LeDoux, a Data Scientist at Buzzfeed. The data is collected on Arabica coffee beans from across the world and professionally rated on a 0-100 scale based on factors like acidity, sweetness, fragrance, balance, etc. The dataset also contains information about the country of origin of the coffee beans, harvesting and grading date, colour of the beans, defects, processing and packaging details [2]. There are 1312 observations in the dataset and 53 variables.

The dataset had a total number of 1, 0497 null values, 1,312 unique values and 53 columns. The main features from the dataset that were used in data analysis included quality score, country of origin, flavor, acidity, sweetness, aroma, moisture, bag weight, colour and defects.

2.0 Data Collection

This project uses the Coffee Quality Dataset, collected by the *Coffee Quality Institute* in January 2018. The data was retrieved from tidytuesday, courtesy of James LeDoux, a Data Scientist at Buzzfeed. The data is collected on Arabica coffee beans from across the world and professionally rated on a 0-100 scale based on factors like acidity, sweetness, fragrance, balance, etc. The dataset also contains information about the country of origin of the coffee beans, harvesting and grading date, colour of the beans, defects, processing and packaging details [2]. There are 1312 observations in the dataset and 53 variables and 37 unique countries of origin as shown below.



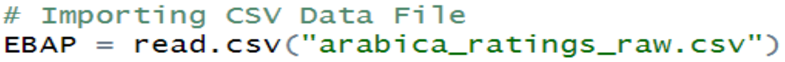
```
EBAP                       1312 obs. of 53 variables
    $ X                               : int  0 1 2 3 4 5 6 7 8 9 ...
    $ Quality_Score                   : num  90.6 89.9 89.8 89 88.8 ...
    $ view_certificate_1              : logi  NA NA NA NA NA NA ...
    $ view_certificate_2              : logi  NA NA NA NA NA NA ...
    $ Cupping.Protocol.and.Descriptors: logi  NA NA NA NA NA NA ...
    $ View.Green.Analysis.Details     : logi  NA NA NA NA NA NA ...
    $ Request.a.Sample                : logi  NA NA NA NA NA NA ...
    $ Species                         : chr  "Arabica" "Arabica" "Arabic…
    $ Owner                           : chr  "metad plc" "metad plc" "Gr…
    $ Country_of_Origin               : chr  "Ethiopia" "Ethiopia" "Guat…
    $ Farm.Name                       : chr  "METAD PLC" "METAD PLC" "Sa…
    $ Lot.Number                      : chr  "" "" "" "" ...
    $ Mill                            : chr  "METAD PLC" "METAD PLC" "" …
    $ ICO.Number                      : chr  "2014/2015" "2014/2015" "" …
    $ Company                         : chr  "METAD Agricultural Develop…
    $ Altitude                        : chr  "1950-2200" "1950-2200" "16…
    $ Region                          : chr  "GUJI-HAMBELA/GOYO" "GUJI-H…
    $ Producer                        : chr  "METAD PLC" "METAD PLC" "" …
    $ Number_of_Bags                  : int  300 300 5 320 300 100 100 3…
    $ Bag.Weight                      : chr  "60 kg" "60 kg" "1" "60 kg"…
    $ In.Country.Partner              : chr  "METAD Agricultural Develop…
    $ Harvest.Year                    : chr  "2014" "2014" "" "2014" ...
    $ Grading.Date                    : chr  "April 4th, 2015" "April 4t…
    $ Owner.1                         : chr  "metad plc" "metad plc" "Gr…
    $ Variety                         : chr  "" "Other" "Bourbon" ""
```

Figure 1: Observations and Variables

3.0 Data Cleaning

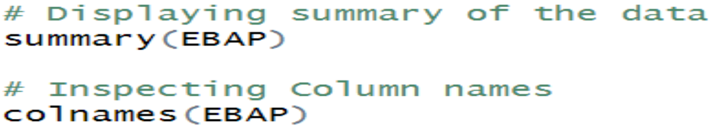
3.1 Data cleaning

- The dataset “*arabica_ratings_raw*” was imported as ***EBAP,*** a csv file using the code below. The *library(readr)* package was installed to read the csv file.



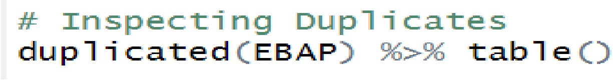
```
# Importing CSV Data File
EBAP = read.csv("arabica_ratings_raw.csv")
```

- The codes below were used to display as summary of the dataset and the column names.



```
# Displaying summary of the data
summary(EBAP)

# Inspecting Column names
colnames(EBAP)
```

- The duplicates in the dataset were then inspected using the command as shown below:



```
# Inspecting Duplicates
duplicated(EBAP) %>% table()
```

No duplicates were detected in the dataset.

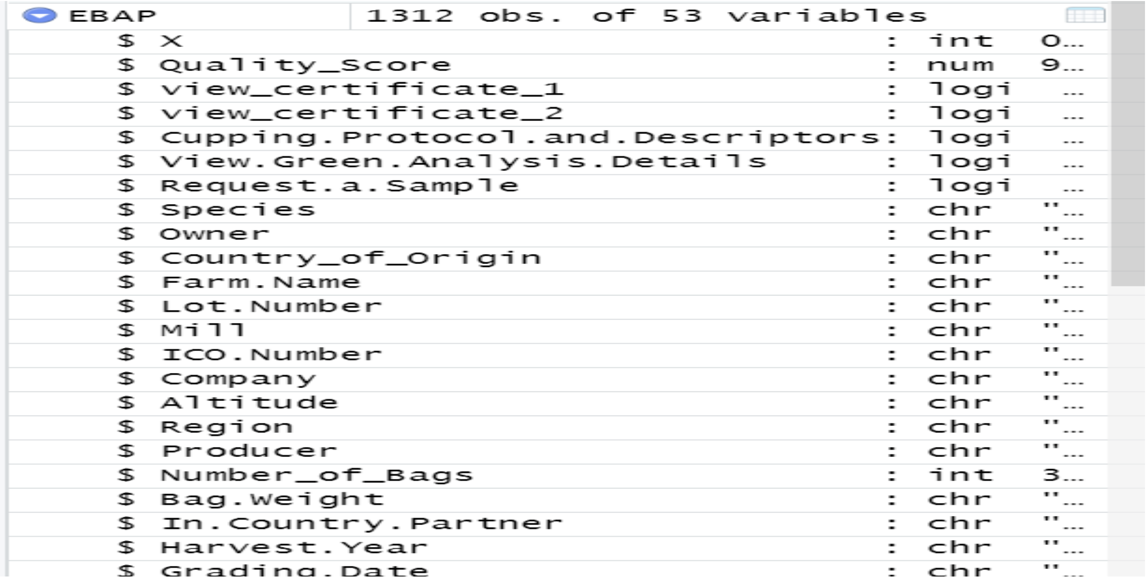
- The data set was then inspected for null and missing values and columns with any missing values were removed. The data set reduced from *1,312 obs. and 53 variables* to *1,312 obs. and 44 variables.*

```
#Inspecting location of missing values
which(is.na(EBAP))

# Count total missing values
sum(is.na(EBAP))

# Remove columns with any NA values
EBAP <- EBAP %>% select_if(~ !any(is.na(.)))
```

- The unwanted columns were eliminated using the code below.

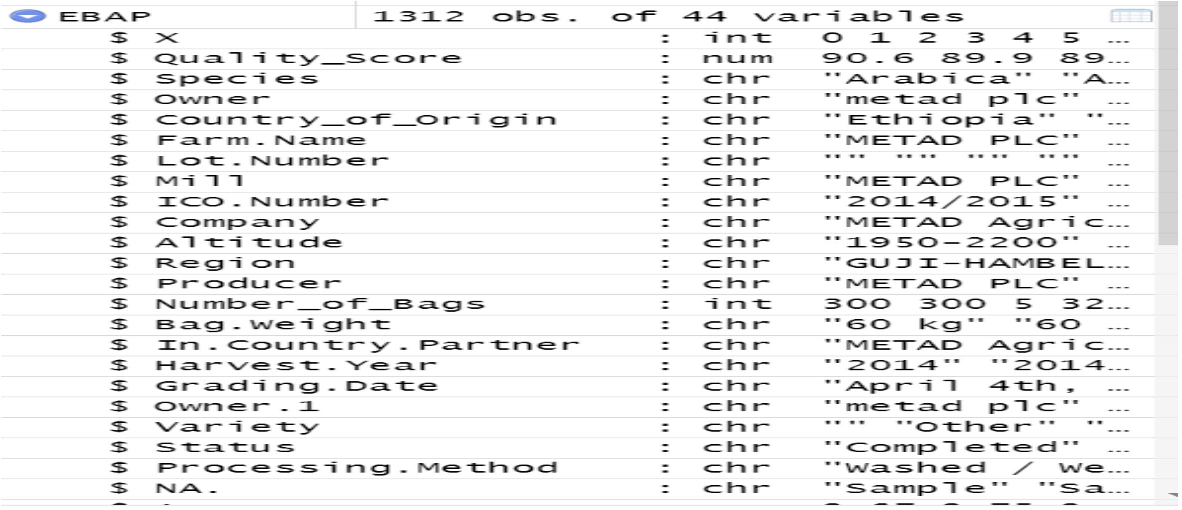
3.2 Before and After Snapshots

The figure 1 below show a summary of the arabica raw data set before data manipulation. The figure 2 show a summary of data set after removing columns and rows with missing values and nulls.



```
EBAP                  1312 obs. of 53 variables
    $ X                                    : int   0...
    $ Quality_Score                        : num   9...
    $ view_certificate_1                   : logi   ...
    $ view_certificate_2                   : logi   ...
    $ Cupping.Protocol.and.Descriptors: logi   ...
    $ View.Green.Analysis.Details          : logi   ...
    $ Request.a.Sample                     : logi   ...
    $ Species                              : chr   "...
    $ Owner                                : chr   "...
    $ Country_of_Origin                    : chr   "...
    $ Farm.Name                            : chr   "...
    $ Lot.Number                           : chr   "...
    $ Mill                                 : chr   "...
    $ ICO.Number                           : chr   "...
    $ Company                              : chr   "...
    $ Altitude                             : chr   "...
    $ Region                               : chr   "...
    $ Producer                             : chr   "...
    $ Number_of_Bags                       : int   3...
    $ Bag.Weight                           : chr   "...
    $ In.Country.Partner                   : chr   "...
    $ Harvest.Year                         : chr   "...
    $ Grading.Date                         : chr   "...
```

Figure 2: Raw Data set



```
EBAP                  1312 obs. of 44 variables
    $ X                           : int   0 1 2 3 4 5 ...
    $ Quality_Score               : num   90.6 89.9 89...
    $ Species                     : chr   "Arabica" "A...
    $ Owner                       : chr   "metad plc" ...
    $ Country_of_Origin           : chr   "Ethiopia" "...
    $ Farm.Name                   : chr   "METAD PLC" ...
    $ Lot.Number                  : chr   "" "" "" "" ...
    $ Mill                        : chr   "METAD PLC" ...
    $ ICO.Number                  : chr   "2014/2015" ...
    $ Company                     : chr   "METAD Agric...
    $ Altitude                    : chr   "1950-2200" ...
    $ Region                      : chr   "GUJI-HAMBEL...
    $ Producer                    : chr   "METAD PLC" ...
    $ Number_of_Bags              : int   300 300 5 32...
    $ Bag.Weight                  : chr   "60 kg" "60 ...
    $ In.Country.Partner          : chr   "METAD Agric...
    $ Harvest.Year                : chr   "2014" "2014...
    $ Grading.Date                : chr   "April 4th, ...
    $ Owner.1                     : chr   "metad plc" ...
    $ Variety                     : chr   "" "Other" "...
    $ Status                      : chr   "Completed" ...
    $ Processing.Method           : chr   "Washed / We...
    $ NA.                         : chr   "Sample" "Sa...
```

Figure 3: Cleaned Dataset

4.0 Data Integration

Data integration process

- The dataset was loaded using the library *tidyverse* in R.
- The data set was checked for missing values and duplicates using the code below.

# Inspecting Duplicates

duplicated(EBAP) %>% table()

# Inspecting location of missing values

which(is.na(EBAP))

# Count total missing values

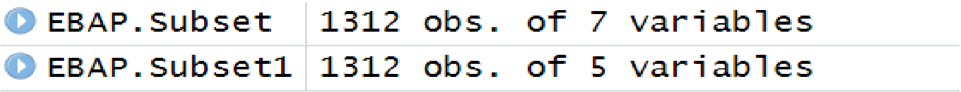
sum(is.na(EBAP))

- The data set was inspected for columns with missing values which were removed from the data set using code below.

  # Remove columns with any NA values

  EBAP <- EBAP %>% select_if(~ !any(is.na(.)))

- Two smaller datasets ***EBA.Subset*** *and* ***EBA.Subset1*** were created from the ***EBAP*** data set as shown below.



```
EBAP.Subset    1312 obs. of 7 variables
EBAP.Subset1   1312 obs. of 5 variables
```

- The codes below were used to create ***EBA.Subset*** *and* ***EBA.Subset1*** data set.

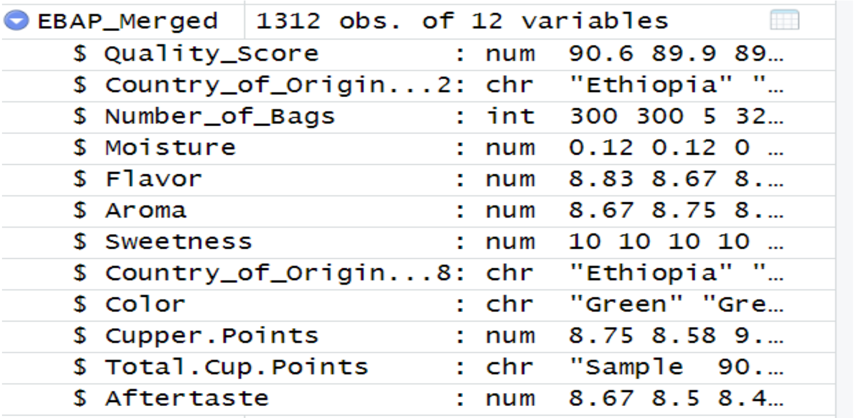
EBAP.Subset <- subset(EBAP, select = c(Quality_Score, Country_of_Origin,Number_of_Bags, Moisture,Flavor,Aroma,Sweetness))

EBAP.Subset1 <- subset(EBAP, select = c(Country_of_Origin,Color,Cupper.Points,Total.Cup.Points, Aftertaste))

- The two datasets **EBAP.Subset** and **EBAP.Subset1** were integrated.

# The two data sets EBAP.Subset and EBAP.Subset1will be merged below.

EBAP_Merged <- bind_cols(EBAP.Subset, EBAP.Subset1)



```
EBAP_Merged    1312 obs. of 12 variables
   $ Quality_Score         : num  90.6 89.9 89…
   $ Country_of_Origin...2: chr  "Ethiopia" "…
   $ Number_of_Bags        : int  300 300 5 32…
   $ Moisture              : num  0.12 0.12 0 …
   $ Flavor                : num  8.83 8.67 8.…
   $ Aroma                 : num  8.67 8.75 8.…
   $ Sweetness             : num  10 10 10 10 …
   $ Country_of_Origin...8: chr  "Ethiopia" "…
   $ Color                 : chr  "Green" "Gre…
   $ Cupper.Points         : num  8.75 8.58 9.…
   $ Total.Cup.Points      : chr  "Sample  90.…
   $ Aftertaste            : num  8.67 8.5 8.4…
```

Figure 4: Integrated Data set

5.0 Data Analysis

Exploratory data analysis (EDA)

The target variables for this exercise were Country of origin, quality score, aroma, flavor, aftertaste, acidity, sweetness, moisture, defects.

To get a sense of the data we exploratory data analysis was done in which I plotted distributions of select categorical features: Country_of_Origin and Quality Score, Country_of_Origin and Sweetness, Country_of_Origin and Aroma, Country_of_Origin and Moisture, Country_of_Origin and Bags of Coffee from the dataset.

In doing so, it was discovered that the average coffee quality rating differed between the various countries of origin: the highest was from United States & Papua New Guinea and the lowest was from Haiti & India.

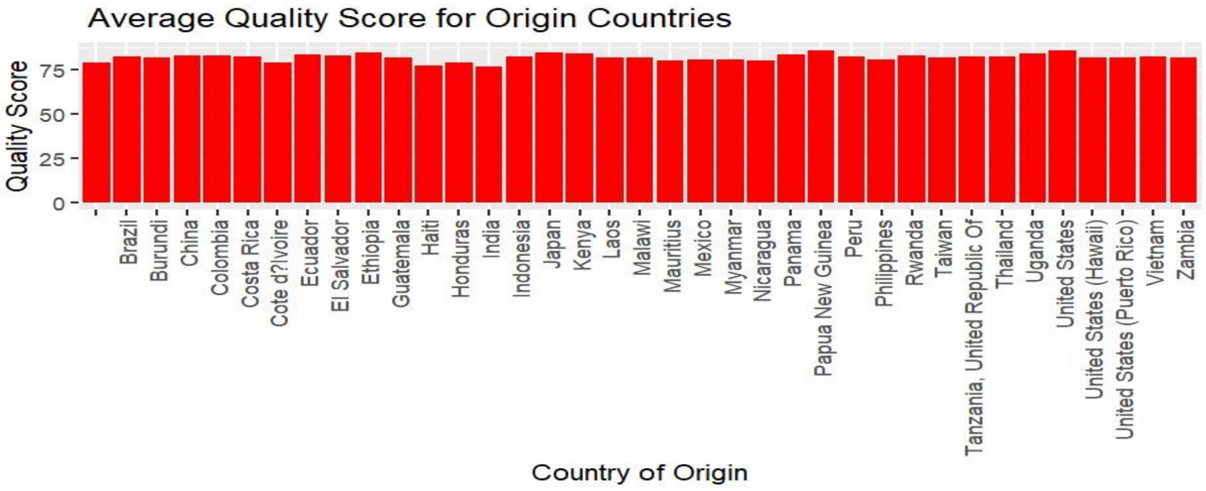


Figure 5: Average Quality Rating

Looking at the average aroma rating from the different countries of origin, Papua New Guinea scores the highest while Haiti has the lowest.

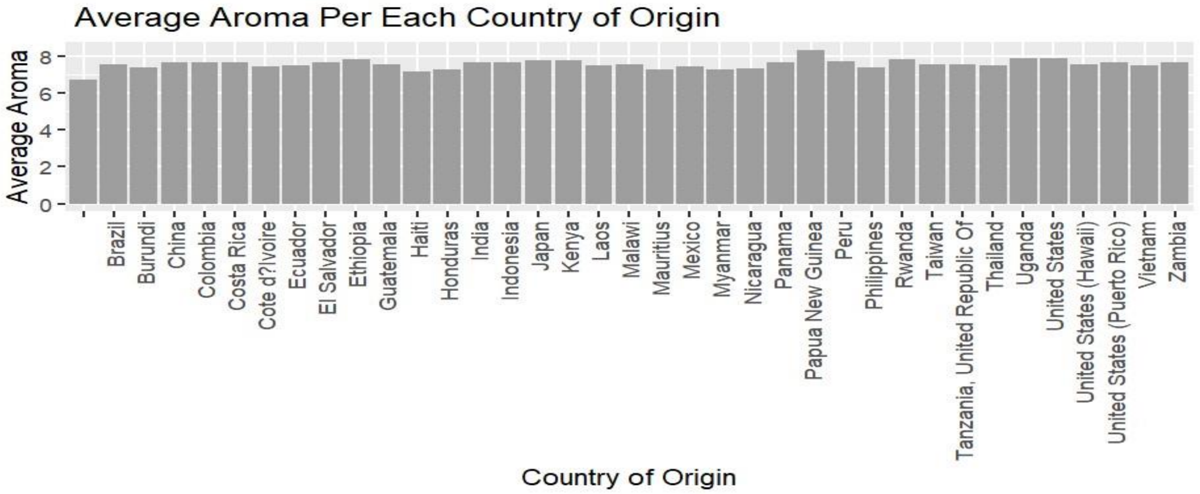


Figure 6: Coffee Aroma

There was little variation is sweetness per cup of coffee tasted from the countries of origin reserve for India which had an average below 7.5.

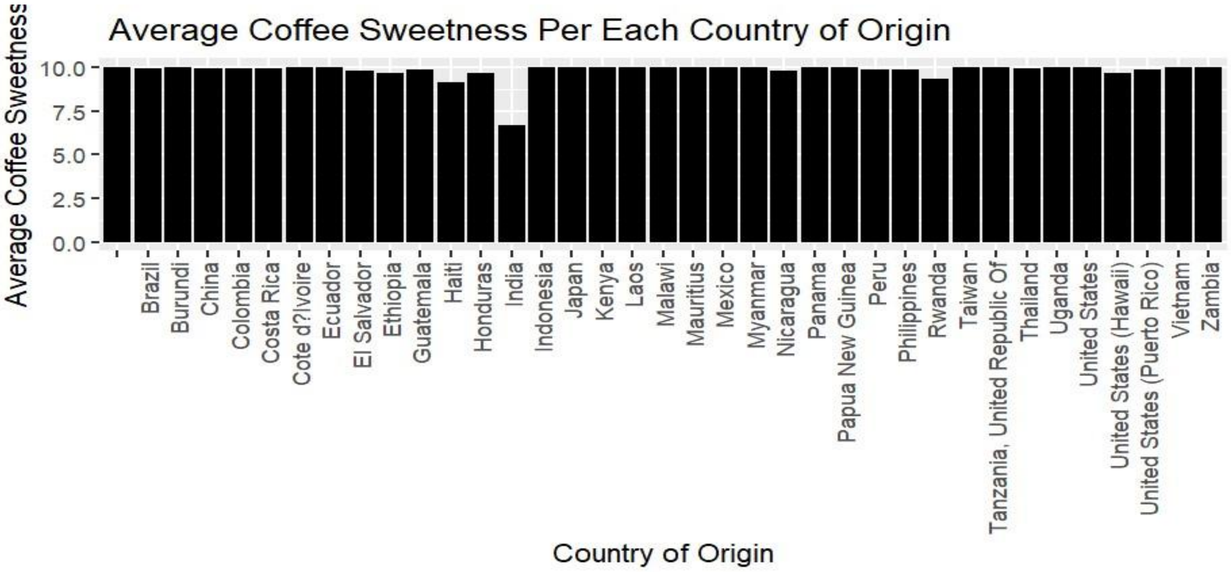


Figure 7: Average Sweetness Rating

The average moisture for coffee from countries of origin had high variations with Cote D'Ivoire registering the highest average moisture. There was no moisture detected in coffee from United States and India.

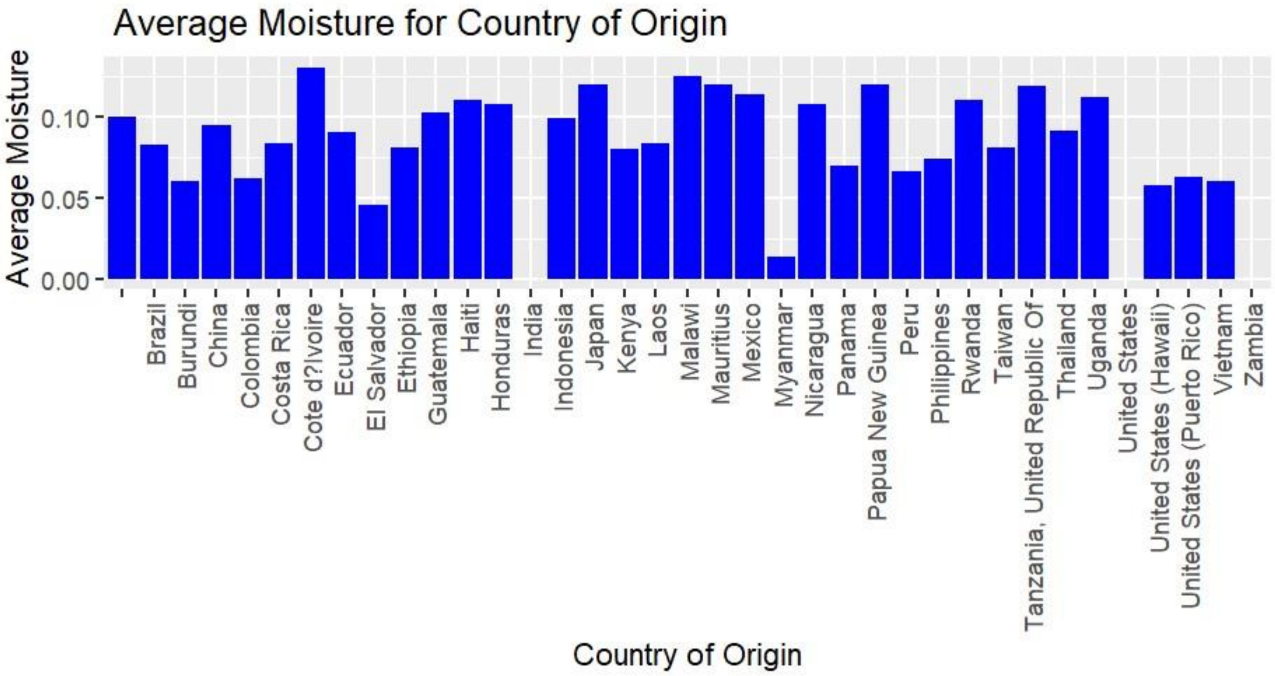


Figure 8: Average Moisture

Looking at the bags of coffee exported for the data set, Burundi, Ethiopia topped the list while the likes of Cote d'Ivoire, Ecuador, Mauritius, Myanmar and Vietnam were at the bottom of the list. However, this is not a determinant on the volumes by country of origin since the samples were randomly selected.

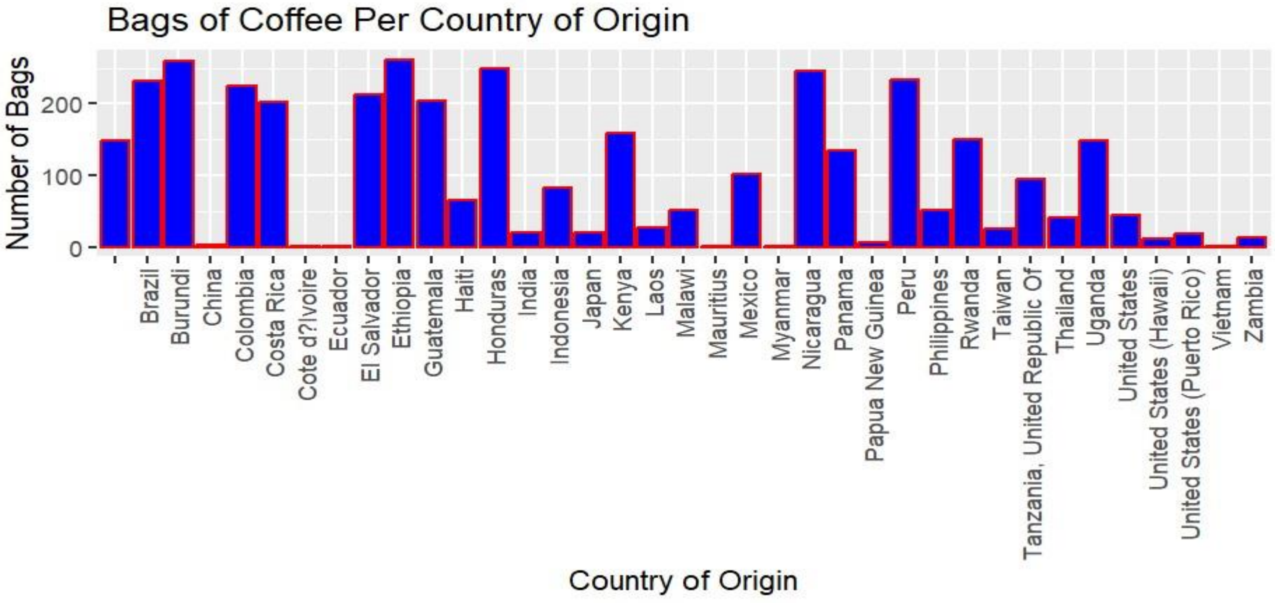


Figure 9: Bags of Coffee

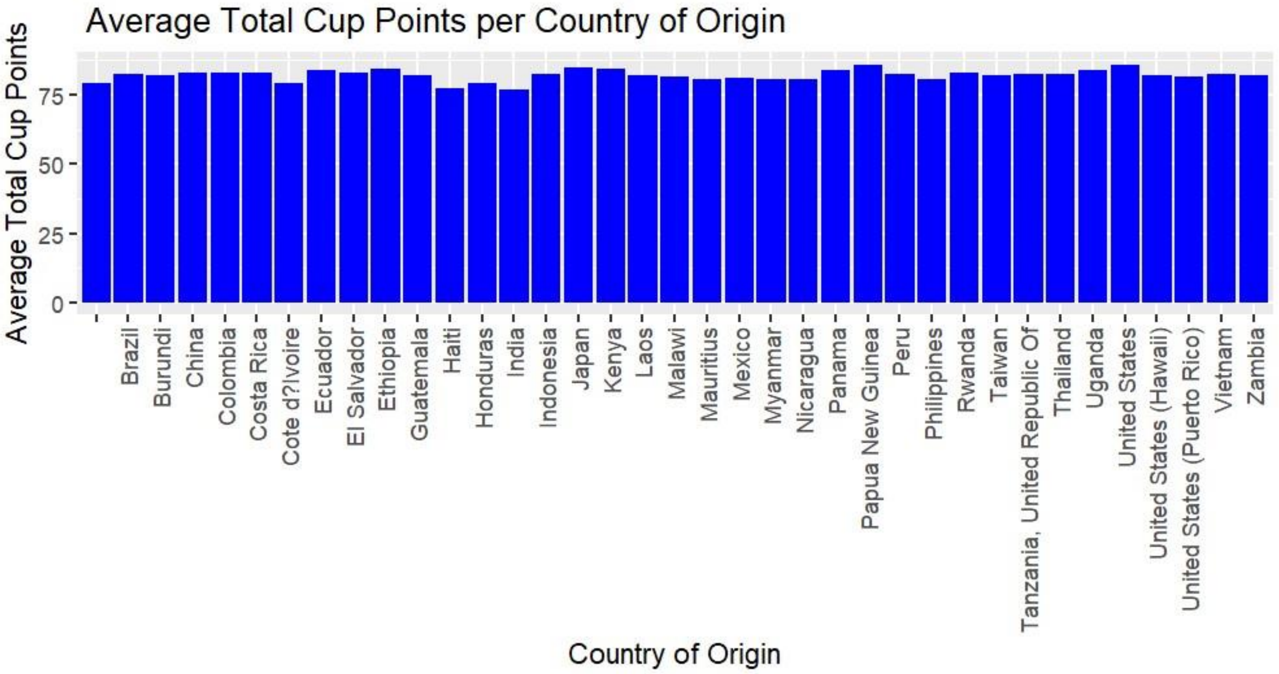


Figure 10: Average Total Cup Points

Features such as aftertaste, acidity, body, balance, uniformity, make up the sum of the total cup points variable and so were not analyzed.

6.0 Data Visualization

6.1 Insights from the Data Visualizations

Power BI was used to conduct data visualization for the data set and below were my findings.

- **Aroma:** Uganda's coffee scores 7.9 points and comes second after Papua New Guinea with 8.33 points out of 10. Aroma of a coffee is assessed by inhaling the aroma released when hot water is poured over the ground coffee. The aroma can vary from floral and fruity to nutty or chocolatey, and it provides important clues about the coffee's potential flavor profile.

- **Flavor:** Uganda's coffee scores 7.75 0ut of 10 points and comes 6th after Papua New Guinea (8.42), United States (8.14), Ethiopia (7.96), Rwanda (7.92), Kenya (7.78). Flavor defines the length of the flavor once the coffee has been swallowed. The longer the pleasant trail, the better the score.
- **Sweetness:** Most of the countries tied with a score of 10 Uganda inclusive. Sweetness is evaluated based on the coffee's natural sweetness and lack of unpleasant bitterness.

- **Total Cupper Points:** Uganda scored 84.052 out of 100 points and is the 6th after United States, Papua New Guinea, Japan, Ethiopia, and Kenya. This qualifies Uganda Arabica Coffee to be categorized as a Specialty Coffee. Cupping is a standardized tasting method used to assess the quality and characteristics of coffee beans. It involves brewing coffee samples and evaluating them in a controlled environment. This process allows cuppers (professional coffee tasters) to objectively analyze and compare various coffees based on specific criteria.

  The cupping score is essentially a numerical rating assigned to a coffee after a cupping session. It serves as a benchmark to assess the coffee's quality, flavor profile, cleanliness, and potential defects. The cupping score ranges from 0 to 100, with 100 being the highest possible score. In practice, the majority of high-quality specialty coffees usually fall within the range of 80 to 90 points.

These are the major determinants of the quality rating of Arabica Coffee.

6.2 Links/Screenshots of Power BI dashboards

Data Visuals Github Link

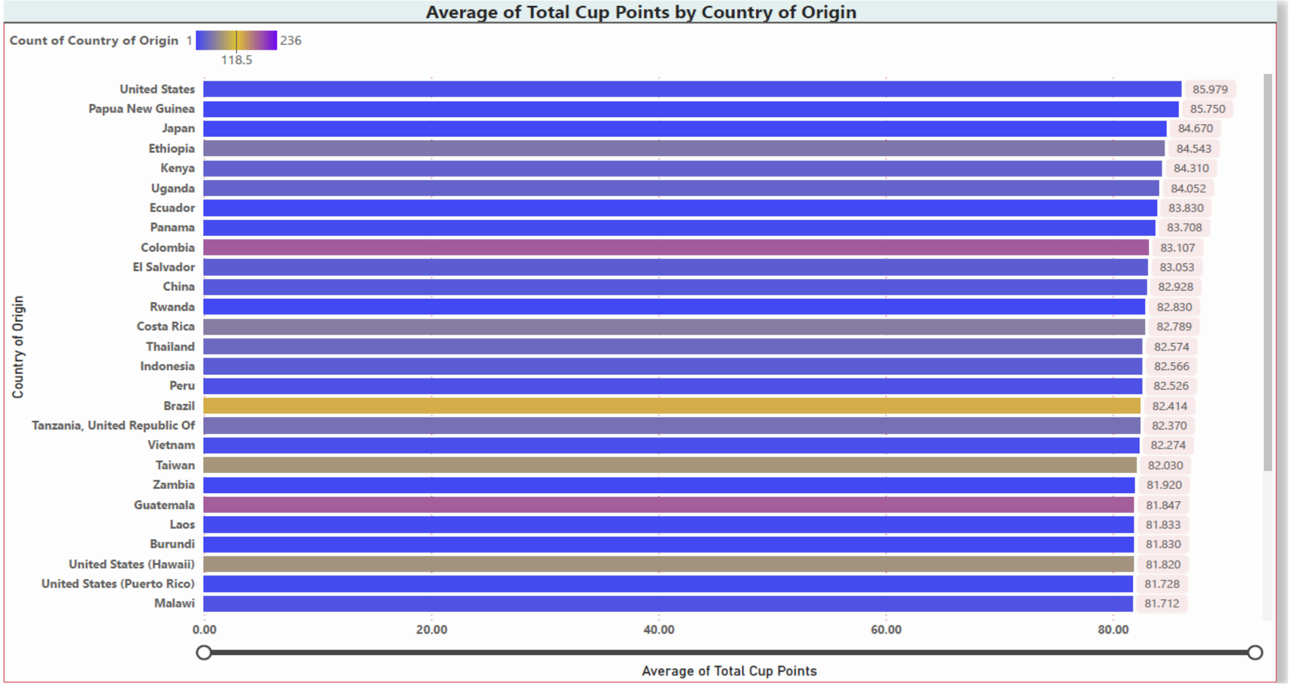


Figure 11: Average of Total Cup Points by Country of Origin

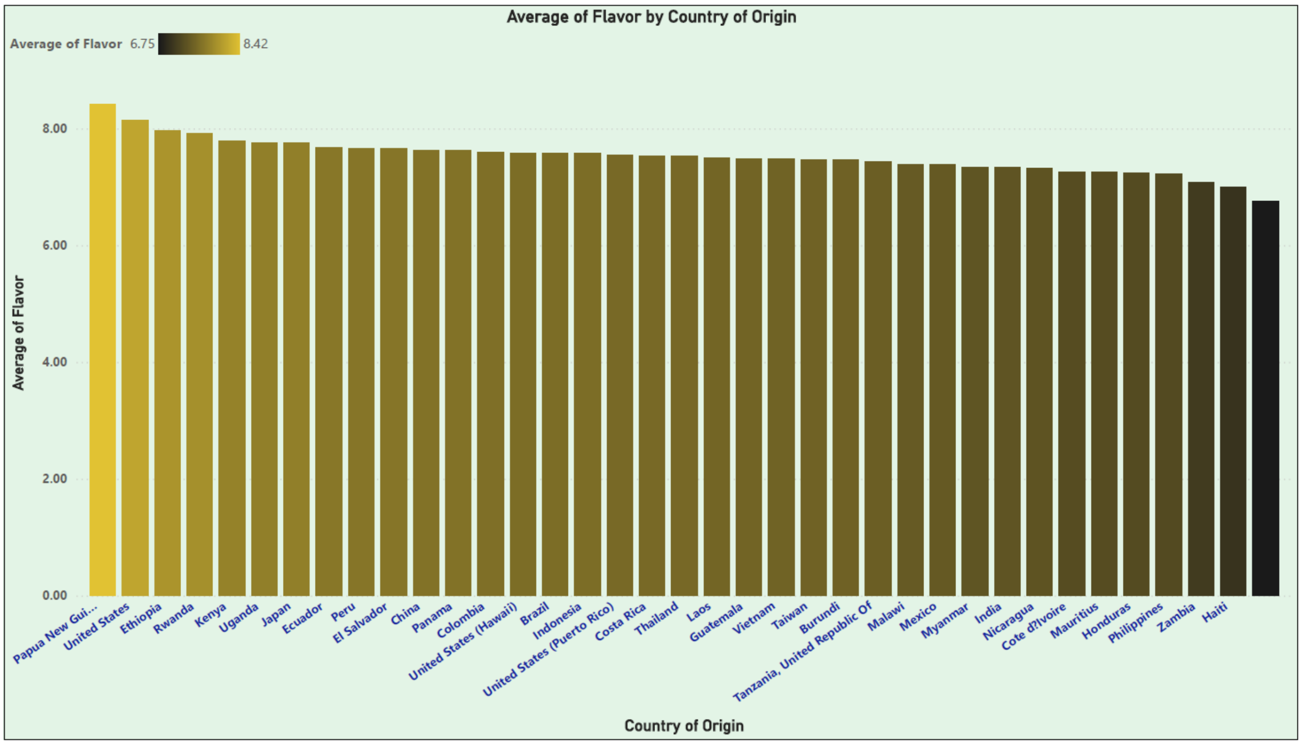


Figure 12: Flavor by Country of Origin

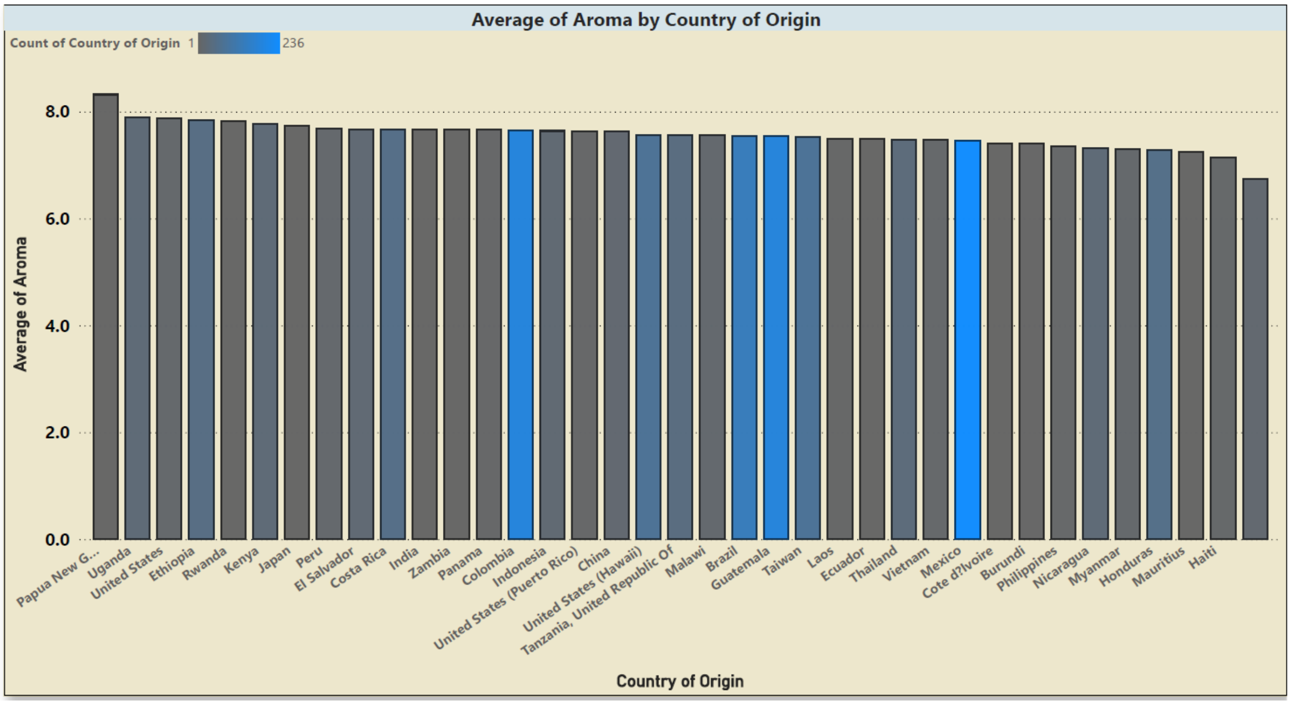


Figure 13: Average Aroma Rating

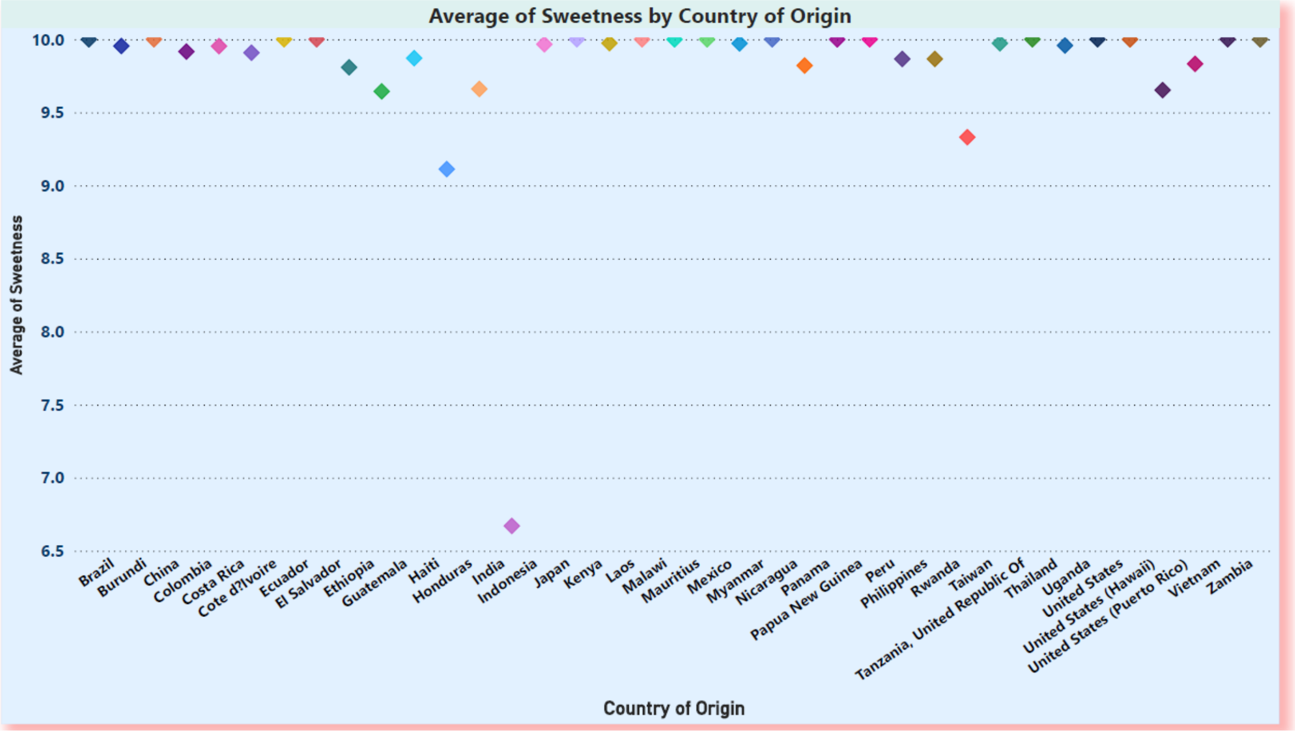


Figure 14: Average Sweetness Rating

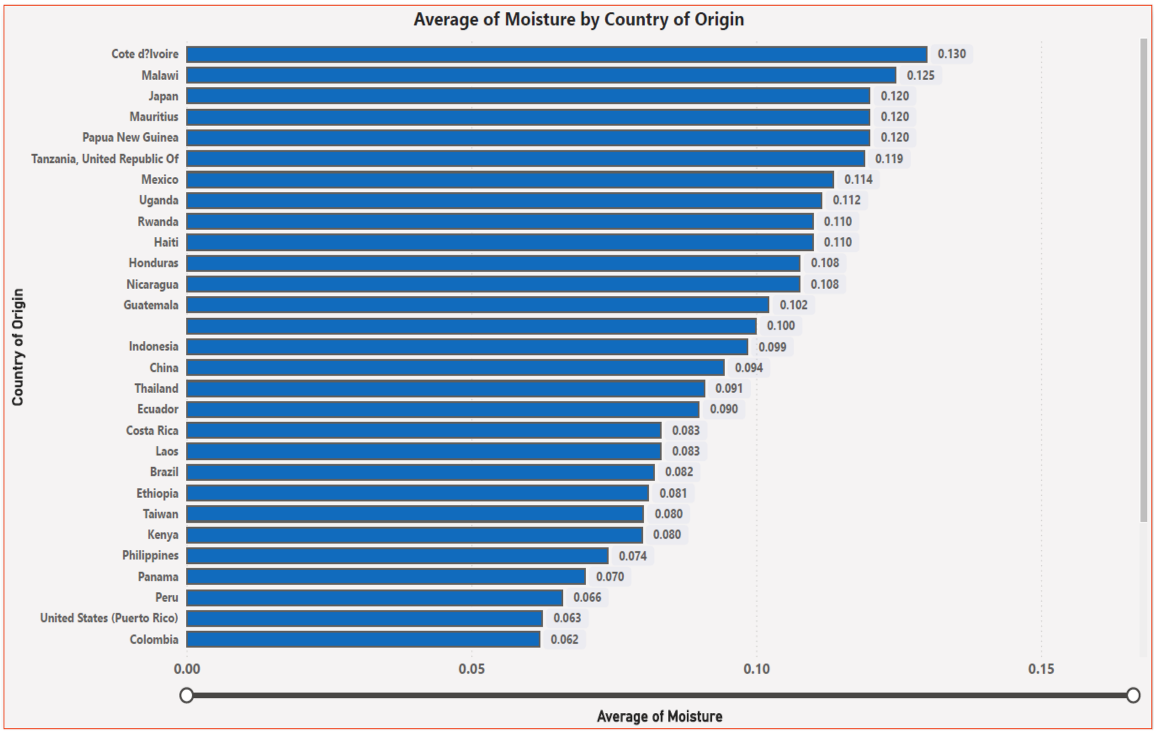


Figure 15: Moisture by Country of Origin

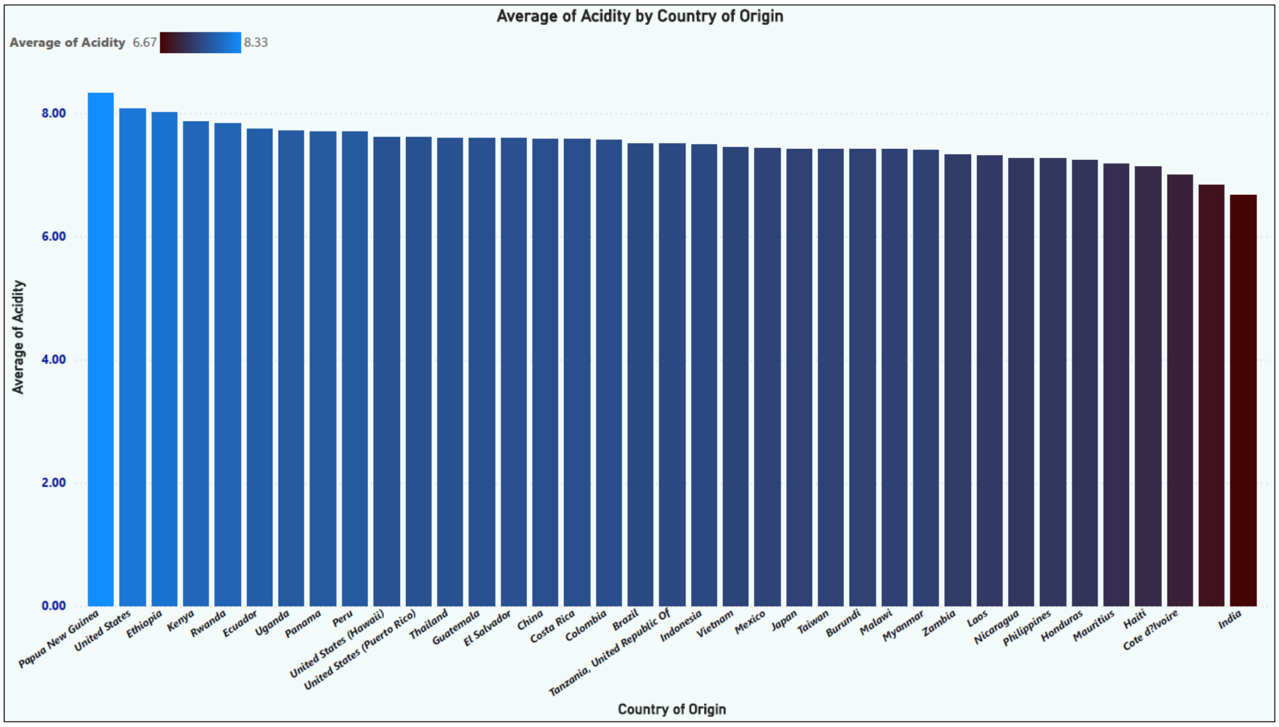


Figure 16: Acidity by Country of Origin

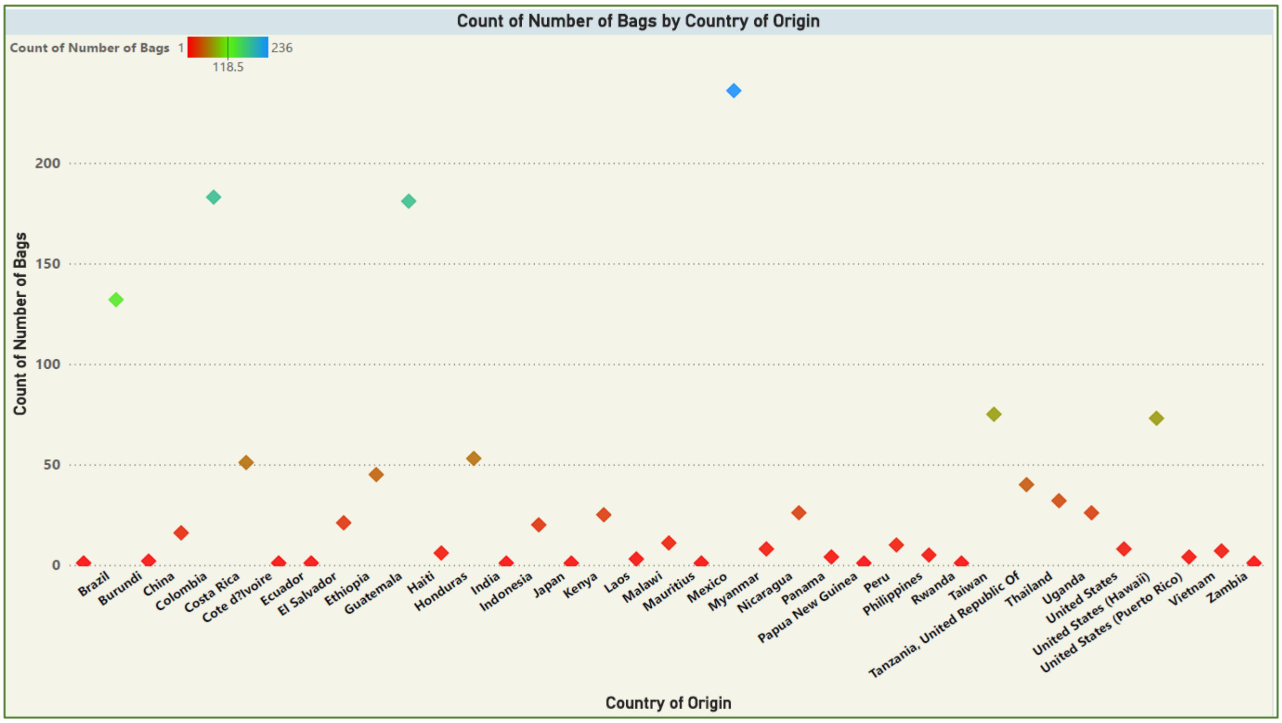


Figure 17: Sampled Bags per Country of Origin

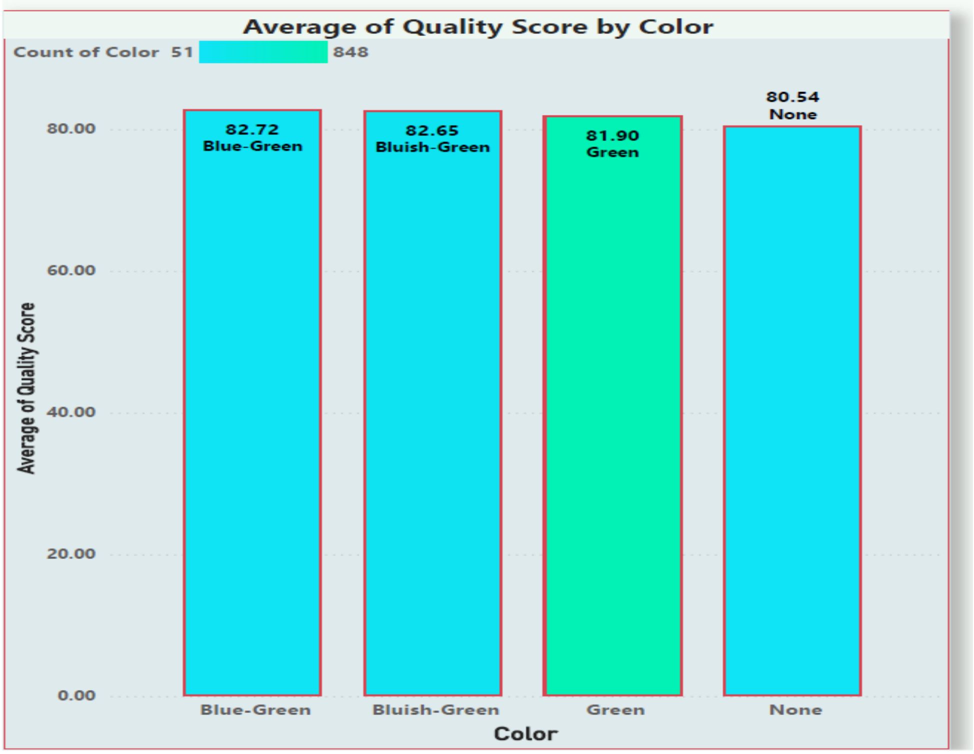


Figure 18: Quality Rating by Color of Beans

7.0 Conclusion and Recommendations

On a rate of 0 – 100, the arabica coffee from Uganda scores an average of 84.052 on the total cupper points and is in the top 10 countries that produce the best arabica coffee globally. Given this rating, Uganda has a competitive advantage on the global market, but more efforts are needed to increase on the arabica coffee volumes produced annually.

The government of Uganda and the different coffee stakeholders should invest more in the arabica coffee value chain right from training the farmers to coffee processing. I believe this will increase the arabica coffee volumes as well as the quality of the coffee produced which will enhance our bargaining power and competitiveness as a country.

References

1. https://www.almacenaplatform.com/uncategorized/ugandan-coffee-export-success-record-revenues-and-market-dynamics-in-june-2023/#
2. https://github.com/jldbc/coffee-quality-database/blob/master/README.md
3. https://stackoverflow.com/questions/2643939/remove-columns-from-dataframe-where-all-values-are-na

4. https://www.infoworld.com/article/2264570/how-to-merge-data-in-r-using-r-merge-dplyr-or-datatable.html
5. https://www.thrillist.com/drink/nation/the-world-s-best-coffee-growing-countries-ethiopia-kenya-colombia-and-more
6. https://medium.com/@yaskalidis/the-data-speak-ethiopia-has-the-best-coffee-91f88ed37e84
7. https://ikigai.coffee/what-is-a-cupping-score/

Appendix

Code snippets

# Setting working directory

```
setwd("C:/Users/jimmy.musinguzi/OneDrive - ED & F Man Holding Limited/Documents/R Programming")
```

# Importing Libraries

```
library(tidyverse)

library(tidyr)

library(readr)

library(dplyr)

library(ggplot2)

library(knitr)
```

# Importing CSV Data File

```
EBAP = read.csv("arabica_ratings_raw.csv")
```

# Displaying summary of the data

```
summary(EBAP)
```

# Inspecting Column names

```
colnames(EBAP)
```

# Inspecting Duplicates

```
duplicated(EBAP) %>% table()
```

# Inspecting location of missing values

```
which(is.na(EBAP))
```

```
# Count total missing values
sum(is.na(EBAP))
# Remove columns with any NA values
EBAP <- EBAP %>% select_if(~ !any(is.na(.)))


#Dropping Duplicates
EBAP<- distinct(EBAP)


# Using the subset() function, I will create two smaller datasets  called EBA.Subset and
EBA.Subset1
EBAP.Subset <- subset(EBAP, select = c(Quality_Score, Country_of_Origin,Number_of_Bags,
Moisture,Flavor,Aroma,Sweetness ))
EBAP.Subset1 <- subset(EBAP, select =
c(Country_of_Origin,Color,Cupper.Points,Total.Cup.Points, Aftertaste ))


# Display the new EBA.Subset created using the subset() Function
print(EBAP.Subset)
print(EBAP.Subset1)


# Inspecting Column names for EBA.Subset
colnames(EBAP.Subset)
colnames(EBAP.Subset1)


# The two data sets EBA and  EBA.Subset will be merged below.
EBAP_Merged <- bind_cols(EBAP.Subset, EBAP.Subset1)


#Display summary and column names of integrated data set
summary(EBAP_Merged)
colnames(EBAP_Merged)
```

```
# plot the data using ggplot
# Average Quality Score for Origin Countries
ggplot(EBAP_Merged, aes(x=Country_of_Origin...8, Quality_Score)) +
geom_bar(stat = "summary", fill="red")+
scale_x_discrete(guide = guide_axis(angle = 90)) +
xlab("Country of Origin")+
ylab("Quality Score")+
labs(title= " Average Quality Score for Origin Countries",)

ggplot(EBAP_Merged, aes(x=Country_of_Origin...8, Quality_Score)) +
geom_histogram(stat = "summary", fill="red")+
scale_x_discrete(guide = guide_axis(angle = 90)) +
xlab("Country of Origin")+
ylab("Quality Score")+
labs(title= " Average Quality Score for Origin Countries",)

# Average Quality Score for Origin Countries
ggplot(EBAP_Merged, aes(x=Country_of_Origin...8, Quality_Score)) +
geom_point( color="blue") +
scale_x_discrete(guide = guide_axis(angle = 90,)) +
xlab("Country of Origin")+
ylab("Quality Score")+
labs(title= " Average Quality Score for Origin Countries",)

# Average Quality Score for Origin Countries
ggplot(EBAP_Merged, aes(x=Country_of_Origin...8, Quality_Score)) +
geom_boxplot( fill="red",color="blue") +
scale_x_discrete(guide = guide_axis(angle = 90)) +
theme_classic()+
xlab("Country of Origin")+
ylab("Quality Score")+
labs(title= " Average Quality Score for Origin Countries",)



# Number of sampled bags per country of origin
ggplot(EBAP_Merged, aes(x=Country_of_Origin...8, Number_of_Bags)) +
geom_bar(stat = "summary",fill="blue",color = "red")+
scale_x_discrete(guide = guide_axis(angle = 90)) +
xlab("Country of Origin")+
ylab("Number of Bags")+
labs(title= " Bags of Coffee Per Country of Origin",)

# Average moisture per country of origin
ggplot(EBAP_Merged, aes(x=Country_of_Origin...8,Moisture)) +
geom_bar(stat = "summary",fill="blue")+
scale_x_discrete(guide = guide_axis(angle = 90)) +
xlab("Country of Origin")+
ylab(" Average Moisture")+
labs(title= " Average Moisture for Country of Origin",)

# Average Aroma per country of origin
ggplot(EBAP_Merged, aes(x=Country_of_Origin...8,Aroma)) +
geom_bar(stat = "summary",fill="008080")+
scale_x_discrete(guide = guide_axis(angle = 90)) +
xlab("Country of Origin")+
ylab("Average Aroma")+
labs(title= " Average Aroma Per Each Country of Origin",)
```

```
# Average Sweetness per country of origin
ggplot(EBAP_Merged, aes(x=Country_of_Origin...8,Sweetness)) +
geom_bar(stat = "summary",fill="9f2b68")+
scale_x_discrete(guide = guide_axis(angle = 90)) +
xlab("Country of Origin")+
ylab("Average Coffee Sweetness")+
labs(title= " Average Coffee Sweetness Per Each Country of Origin",)

#Calculating mean, median, variance and standard deviation
mean(EBAP_Merged$Quality_Score)
median(EBAP_Merged$Quality_Score)
range(EBAP_Merged$Quality_Score)
var(EBAP_Merged$Quality_Score)
sd(EBAP_Merged$Quality_Score)
```