Analysis of Factors Effecting Coffee Quality

It has been said, "Happiness is coffee on a fall day!" However, what determines a good cup of coffee from a great cup of coffee? This R project data analysis will be an exploration of the variables that have the most effect on quality as expressed through coffee ratings. The data set being used for this analysis is from the Coffee Quality Institute as provided by the TidyTuesday project. (Reference link: https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-07-07/coffee_ratings.csv). This data was collected from the Coffee Quality Institute's review pages in January of 2018.

To begin our analysis of this data set, let us examine the variables contained within and their associated definitions.

```
# Load data set and assign to variable name coffee_ratings.
coffee_ratings <- read.csv("coffee_ratings.csv", header = TRUE, sep =",")

# Determine the dimensions and column variables of the data set.
dim(coffee_ratings)
```

```
## [1] 1339   43
```

```
names(coffee_ratings)
```

```
##  [1] "total_cup_points"      "species"               "owner"
##  [4] "country_of_origin"     "farm_name"             "lot_number"
##  [7] "mill"                  "ico_number"            "company"
## [10] "altitude"              "region"                "producer"
## [13] "number_of_bags"        "bag_weight"            "in_country_partner"
## [16] "harvest_year"          "grading_date"          "owner_1"
## [19] "variety"               "processing_method"     "aroma"
## [22] "flavor"                "aftertaste"            "acidity"
## [25] "body"                  "balance"               "uniformity"
## [28] "clean_cup"             "sweetness"             "cupper_points"
## [31] "moisture"              "category_one_defects"  "quakers"
## [34] "color"                 "category_two_defects"  "expiration"
## [37] "certification_body"    "certification_address" "certification_contact"
## [40] "unit_of_measurement"   "altitude_low_meters"   "altitude_high_meters"
## [43] "altitude_mean_meters"
```

Initial analysis finds the data set dimensions are 1339 rows by 43 columns. Upon review of the column variables, focus for this analysis will include the following column variables: total_cup_points, species, country_of_origin, variety, processing_method, aroma, flavor, aftertaste, acidity, body, balance, uniformity, clean_cup, sweetness, cupper_points, color, and altitude_mean_meters.

The column keys for each of the column variables are listed below.

total_cup_points : Total rating based on a scale from 0 to 100 points.

species: Species of the coffee bean. Either Arabica or Robusta.

country_of_origin: Country where coffee bean was grown.

processing_method: Processing method for the coffee bean.

aroma: Aroma grade based on a score from 0 to 10 points.

flavor: Flavor grade based on a score from 0 to 10 points

aftertaste: Aftertaste grade based on a score from 0 to 10 points.

acidity: Acidity grade based on a score from 0 to 10 points.

body: Body grade based on a score from 0 to 10 points.

balance: Balance grade based on a score from 0 to 10 points.

uniformity: Uniformity grade based on a score from 0 to 10 points.

clean_cup: Clean cup grade based on a score from 0 to 10 points.

sweetness: Sweetness grade based on a score from 0 to 10 points.

cupper_points: Cupper grade based on a score from 0 to 10 points.

color: Color of the coffee bean.

altitude_mean_meters: Mean altitude in meters at which the coffee bean was grown.

Note: The variables aroma, flavor, aftertaste, acidity, body, balance, uniformity, clean_cup, sweetness, and cupper_points are summed to determine the overall total_cup_points for each coffee sample. Additional information on sample grading can be found at the Coffee Quality Institute's website. (Reference link: https://database.coffeeinstitute.org/coffee/357789/grade)

Review of the data set indicates that the majority of the data being examined is clean. Total_cup_points are available for all samples with the exception of one sample, for which all zeros have been entered. This sample will be filtered out of the analysis. For some of the additional column variables there appears to be NA fields present. Analysis involving these column variables will be filtered to not include samples with NA fields, as no additional information is available to update these fields. As no identification number is available for each sample, an additional column will be added to the data set named coffee_id based on the row number filtered by total_cup_points to aid in analysis.The packages being used during this analysis are the tidyverse, ggridges, and epiDisplay packages. These packages help provide a more robust graphical representation of the data.

```
# Load packages.
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
library(ggridges)
```

```
## Warning: package 'ggridges' was built under R version 4.0.5
```

```
library(epiDisplay)
```

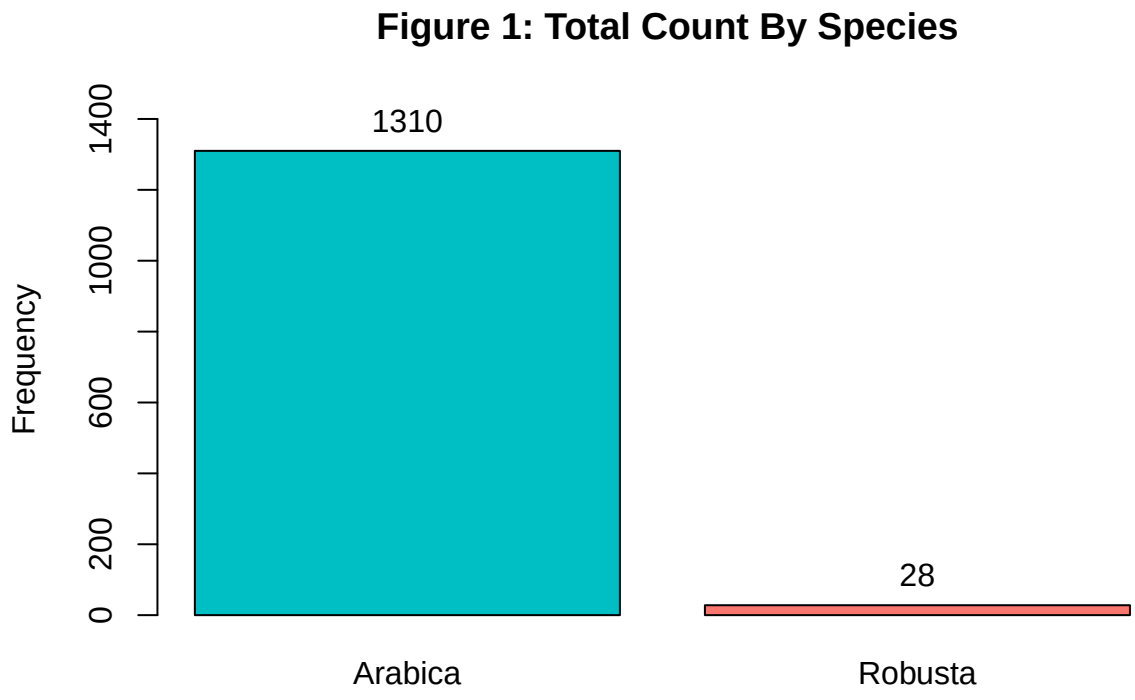
```
## Warning: package 'epiDisplay' was built under R version 4.0.5
```

```
## Warning: package 'foreign' was built under R version 4.0.3

## Warning: package 'survival' was built under R version 4.0.5

## Warning: package 'MASS' was built under R version 4.0.5

## Warning: package 'nnet' was built under R version 4.0.5
```

```
# Add coffee_id column to data set and filter any total_cup_points equal to 0.
coffee_ratings <- read.csv("coffee_ratings.csv", header = TRUE, sep =",")%>%
mutate(coffee_id = row_number())%>%
filter(total_cup_points > 0)
```

Let us begin the analysis by looking at coffee bean species. The main two species used in coffee production worldwide are Arabica and Robusta, with each species containing multiple different varieties. From Figure 1 and the summary chart below, we see the overwhelming majority of samples in this data set are the Arabica species. Given the number samples in this data set and that the Robusta species only accounts for 2.1% of the samples in this set, this would be an indication that the Arabica species more than likely has a better associated taste.

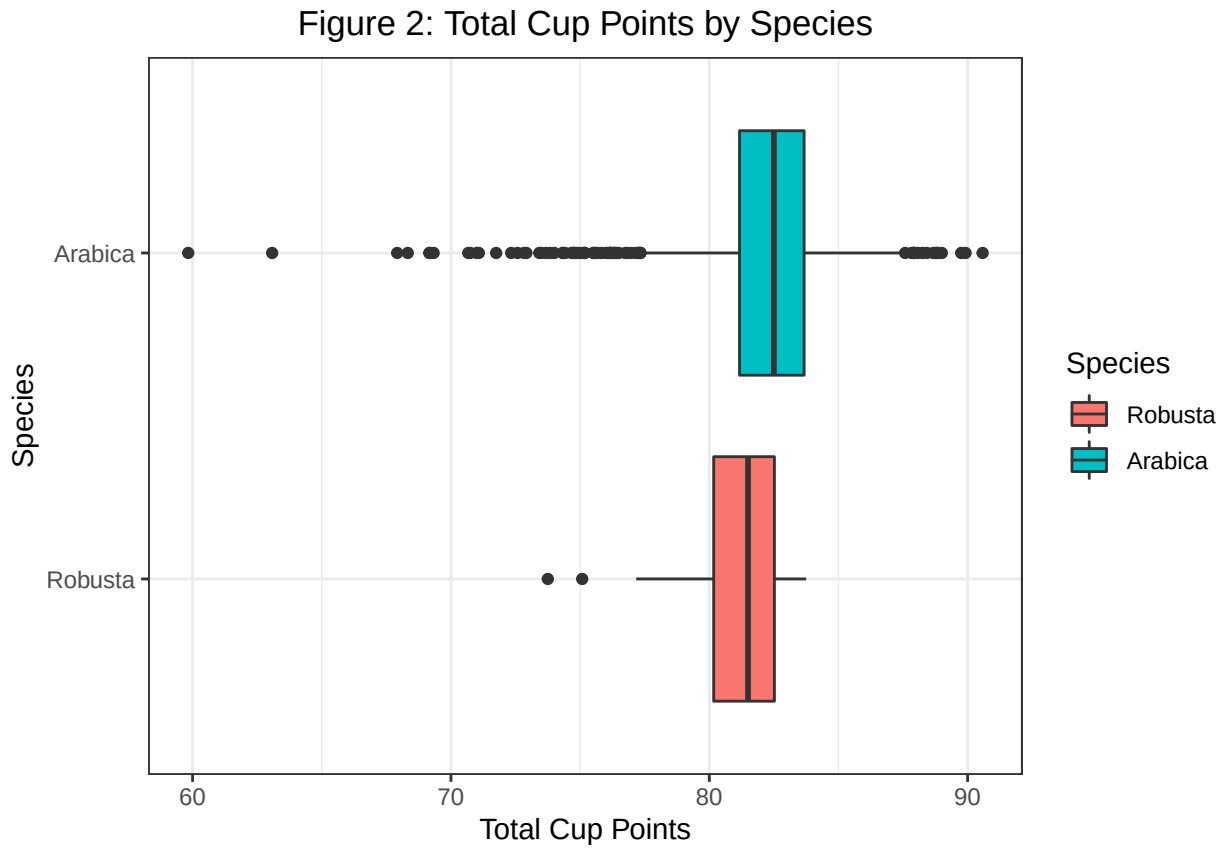
```
# Calculate total count for coffee bean species in data set.
tab1(coffee_ratings$species,sort.group = "decreasing", cum.percent = TRUE,
main= " Figure 1: Total Count By Species", col=c("#00BFC4","#F8766D"))
```



```
## coffee_ratings$species :
##         Frequency Percent Cum. percent
## Arabica      1310    97.9         97.9
## Robusta        28     2.1        100.0
##   Total      1338   100.0        100.0
```

While the sample size for the Robusta species is rather small in comparison to the Arabica species, we will compare the total cup points for each species using the boxplot below in Figure 2.

```
# Create boxplot of coffee bean species by total cup points.
# Mutated to reorder by species and total_cup_points.
coffee_ratings %>%
  mutate(species = fct_reorder(species, total_cup_points))%>%
  ggplot(aes(total_cup_points, species, fill = species)) +
  geom_boxplot()+ xlab("Total Cup Points")+ ylab("Species") +
  ggtitle("Figure 2: Total Cup Points by Species") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))+labs(fill= "Species")
```

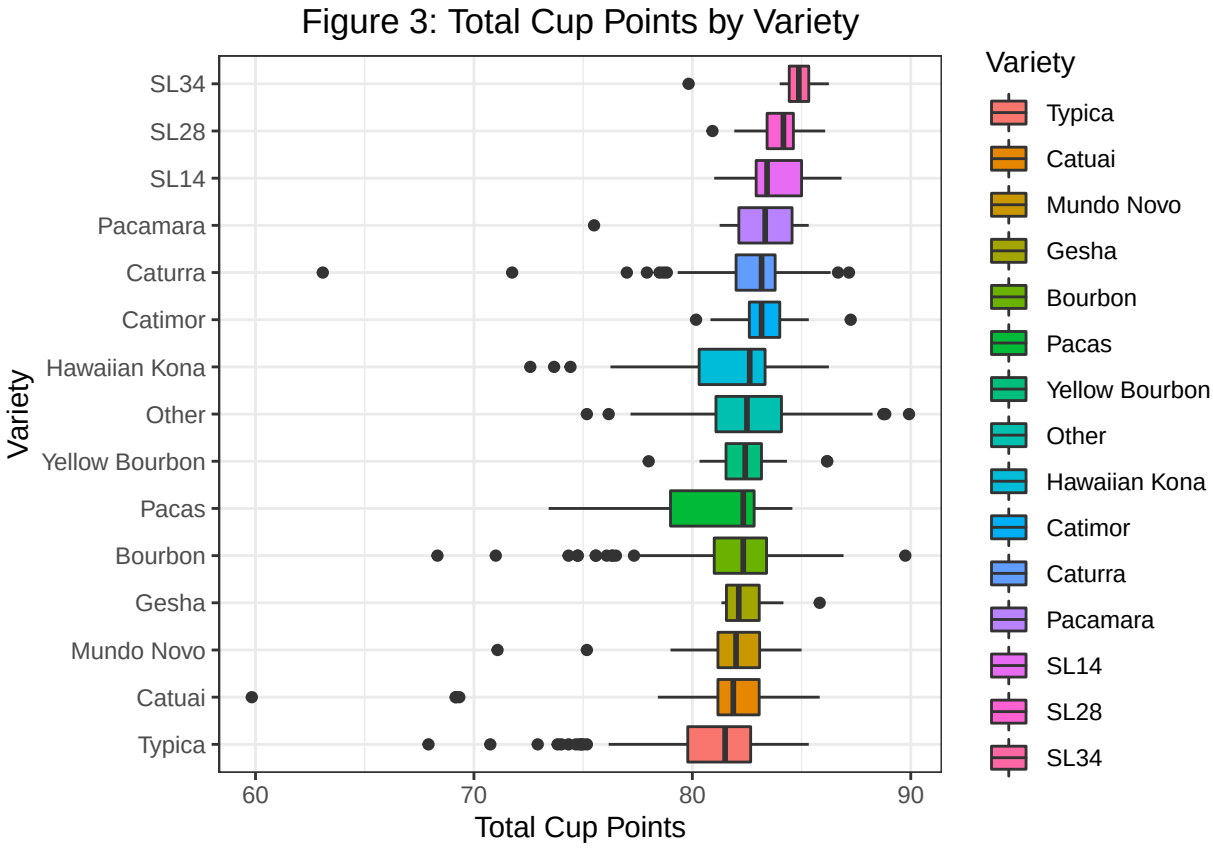


From the boxplot in Figure 2, we see that the overall median total cup score is higher for Arabica as compared to Robusta, thus adding credence to our earlier observation that Arabica most likely has a more desired taste than Robusta. However, to thoroughly investigate this comparison in greater depth, a larger sample size for Robusta should be used.

Having looked at species, let us now examine how variety impacts the overall total cup score. As there are over 30 different varieties listed, we will only explore the top 15 varieties in count. The variety column

also contains 226 entries that are NA, so these will be filtered out of the analysis. The top 15 varieties are compared below in Figure 3.

```
# Filter out NA field in variety column and lump variety to filter results for top 15
# varieties.
# Created new variable for the lumped data called coffee_variety.
coffee_variety <- coffee_ratings%>%
  filter(!is.na(variety))%>%
    mutate(variety = fct_lump(variety, 15), sort = TRUE)
# Create boxplot of coffee bean variety by total cup points for top 15 varieties.
# Mutated to reorder by variety and total_cup_points.
coffee_variety %>%
  mutate(variety = fct_reorder(variety, total_cup_points))%>%
  ggplot(aes(total_cup_points, variety, fill = variety)) +
  geom_boxplot()+ xlab("Total Cup Points")+ ylab("Variety") +
  ggtitle("Figure 3: Total Cup Points by Variety") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))+labs(fill= "Variety")
```

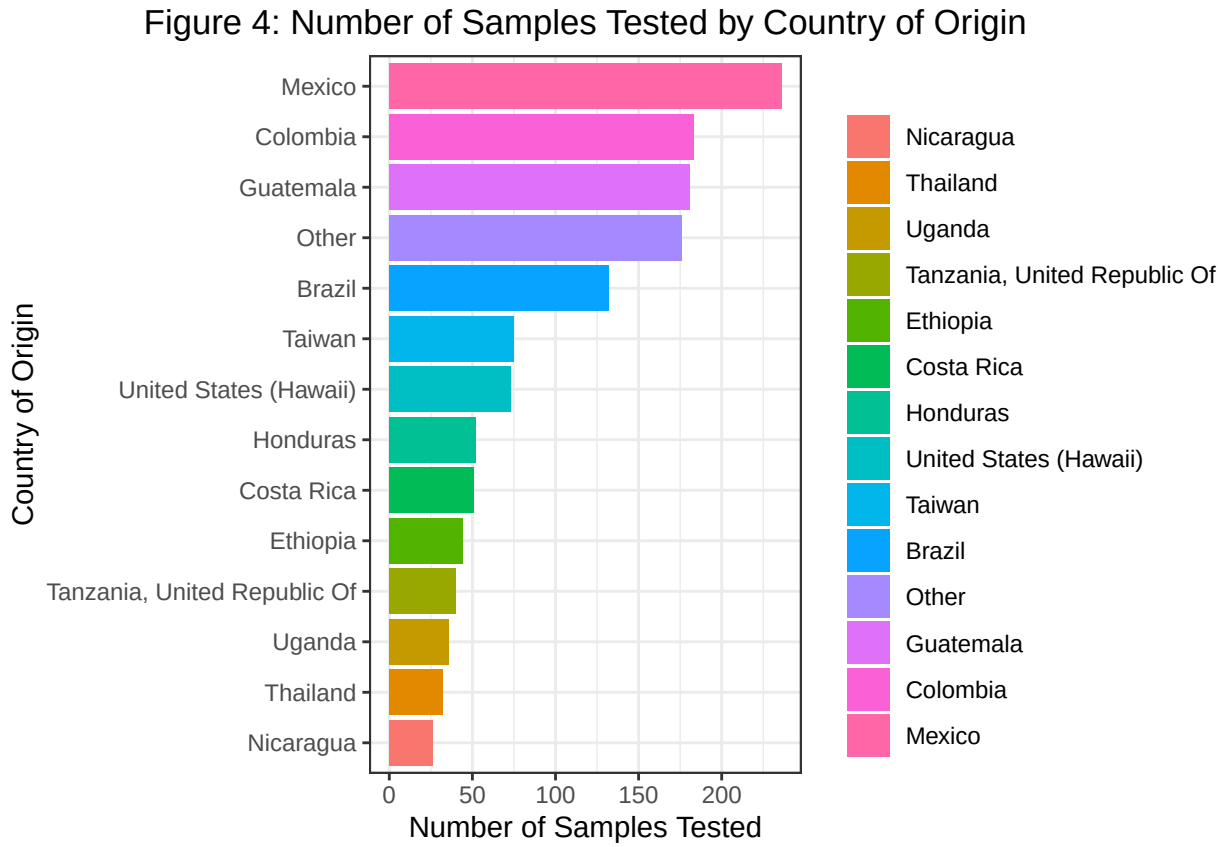


As indicated by the boxplot in Figure 3 above, the SL34 variety has the highest overall median total cup points score followed closely by the SL28 variety. The results of the plot indicate an association between coffee variety and total cup points.

If species and variety are associated with overall total cup score, does the growing region also have an association with coffee quality? To perform this analysis, let us start looking at the number of countries with samples submitted for testing in this data set. Analysis shows that there are 36 countries, or regions of a country, that have samples represented in this data set with one entry containing an NA value. For

consideration of this analysis, we will only consider countries with a sample size of 25 or greater. This data set contains 14 countries with a sample size 25 or more. Figure 4 represents the number of samples tested by country of origin for these 14 countries.

```
# Filter out NA field in country_of_origin column and lump countries to filter results
# for top 14 countries.
# Mutated to reorder by country and number of samples.
# Create Bar graph of country of origin by number of samples.
coffee_ratings %>%
  count(country = fct_lump(country_of_origin, 13), sort= TRUE) %>%
  filter((!is.na(country)))%>%
  mutate(country = fct_reorder(country, n))%>%
  ggplot(aes(n, country, fill= country)) +
  geom_col()+xlab("Number of Samples Tested")+ ylab("Country of Origin") +
  ggtitle("Figure 4: Number of Samples Tested by Country of Origin") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))+labs(fill="")
```



As presented in Figure 4, the countries with the largest number of samples tested in the data set are from Mexico, Colombia, Guatemala, and Brazil. There also appears to be a large number of samples from "Other". There is no indication from where the samples in this category were obtained. As such, very little can be ascertained for this response. To continue with the analysis regarding country of origin, we will look at the relationship between country of origin and total cup points. Reference Figure 5 below.

```
# Filter out NA field in country_of_origin column and lump countries to filter results
# for top 14 countries.
```

```
# Mutated to reorder by country and total cup points.
# Create boxplot of country of origin by total cup points.
coffee_ratings %>%
  filter(!is.na(country_of_origin))%>%
  mutate(country = fct_lump(country_of_origin, 13),
         country = fct_reorder(country, total_cup_points)) %>%
  ggplot(aes(total_cup_points, country, fill= country)) +
  geom_boxplot() + xlab("Total Cup Points")+ ylab("Country of Origin") +
  ggtitle("Figure 5: Total Cup Points by Country") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))+labs(fill= "Country")
```

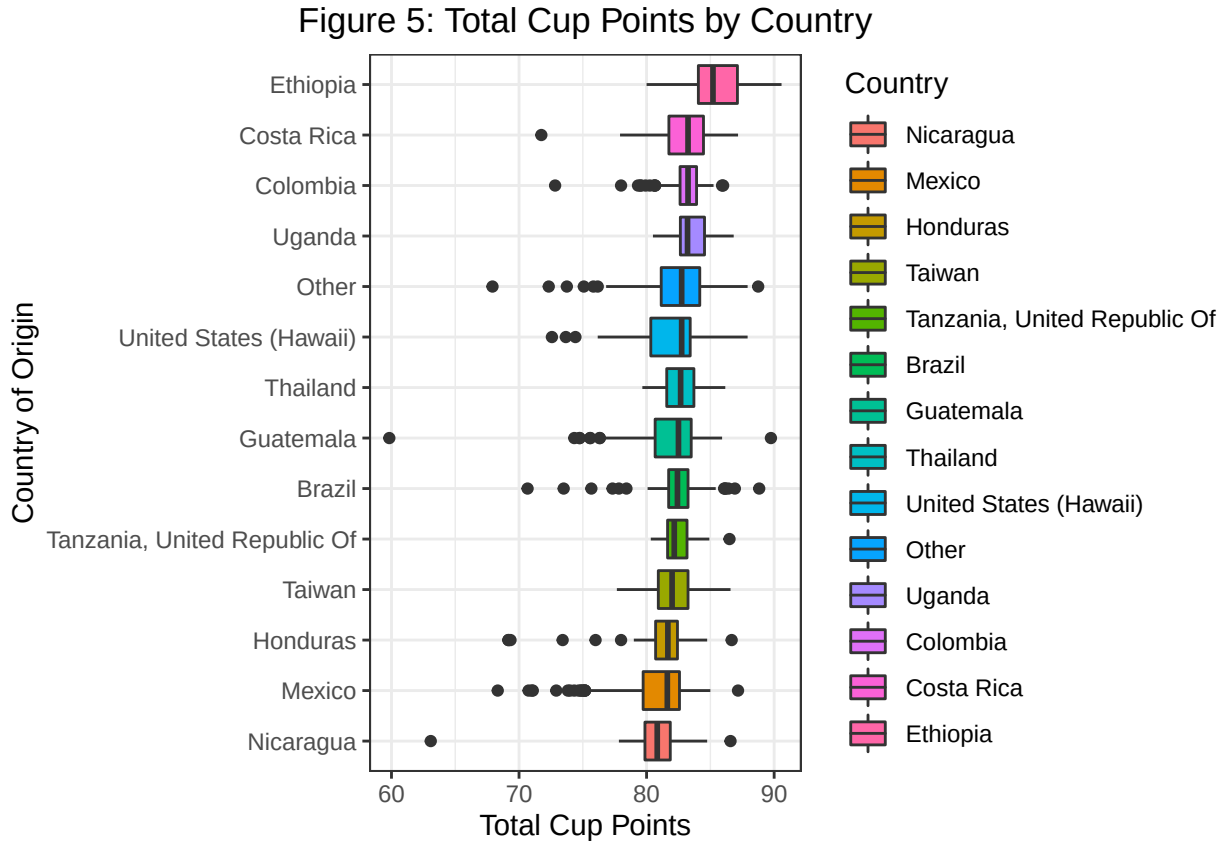
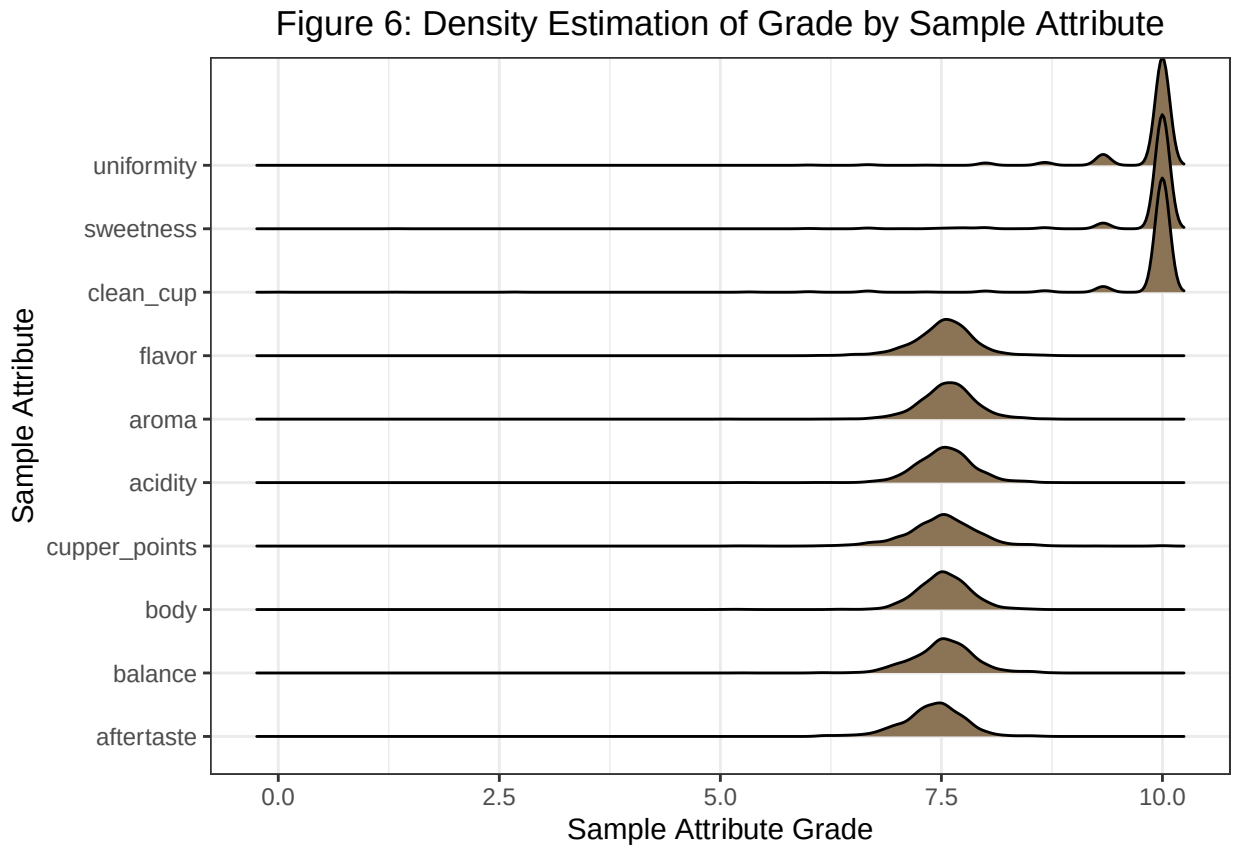


Figure 5 shows that the median overall total cup point score for Ethiopia is the highest of all countries listed. However, one interesting aspect noted was with Mexico and Colombia having the most samples tested, the median result for Columbia was also at the top of the list for total cup points, but Mexico fell to the near the bottom when analyzed for total cup points. There does appear to be a fair number of lower outliers for Mexico which could cause the lower median result. This analysis would indicate an association between region and coffee quality as well.

Let us turn our focus now to the ten sample attributes that make up the total cup score. The attributes of aroma, flavor, aftertaste, acidity, body, balance, uniformity, clean cup, sweetness, and cupper points are used to determine total cup score. For the initial analysis of these attributes, we will compare the density variation of the related scores for each attribute to determine if there are any noticeable differences between the various attributes. To perform this type of analysis, we will need to restructure our current data set to create one (coffee_measurement) in which the various attributes are combined into one column named measurement with subsequent scores added to another column named value.

```r
# Create data frame named coffee_measurement selecting only coffee_id, total_cup_points,
# species, variety, country_of_origin, aroma:cupper_points, altitude_mean_meters
# column variables. Then pivot_long the variables associated with sample attributes calling
# this new column for attributes "measurement" and assigning the associated score for the
# attribute to a column named "value".
coffee_measurement <- coffee_ratings %>%
  dplyr::select(coffee_id, total_cup_points, species, variety, country_of_origin,
aroma:cupper_points, altitude_mean_meters)%>%
  pivot_longer(aroma:cupper_points, names_to = "measurement", values_to = "value")
# Mutated to reorder by measurement and value.
# Create density ridge plot of grade by sample attribute.
coffee_measurement %>%
  mutate(measurement =  fct_reorder(measurement, value)) %>%
  ggplot(aes(value, measurement)) +
  geom_density_ridges(fill="burlywood4")+
  xlab("Sample Attribute Grade")+
  ylab("Sample Attribute") +
  ggtitle("Figure 6: Density Estimation of Grade by Sample Attribute") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

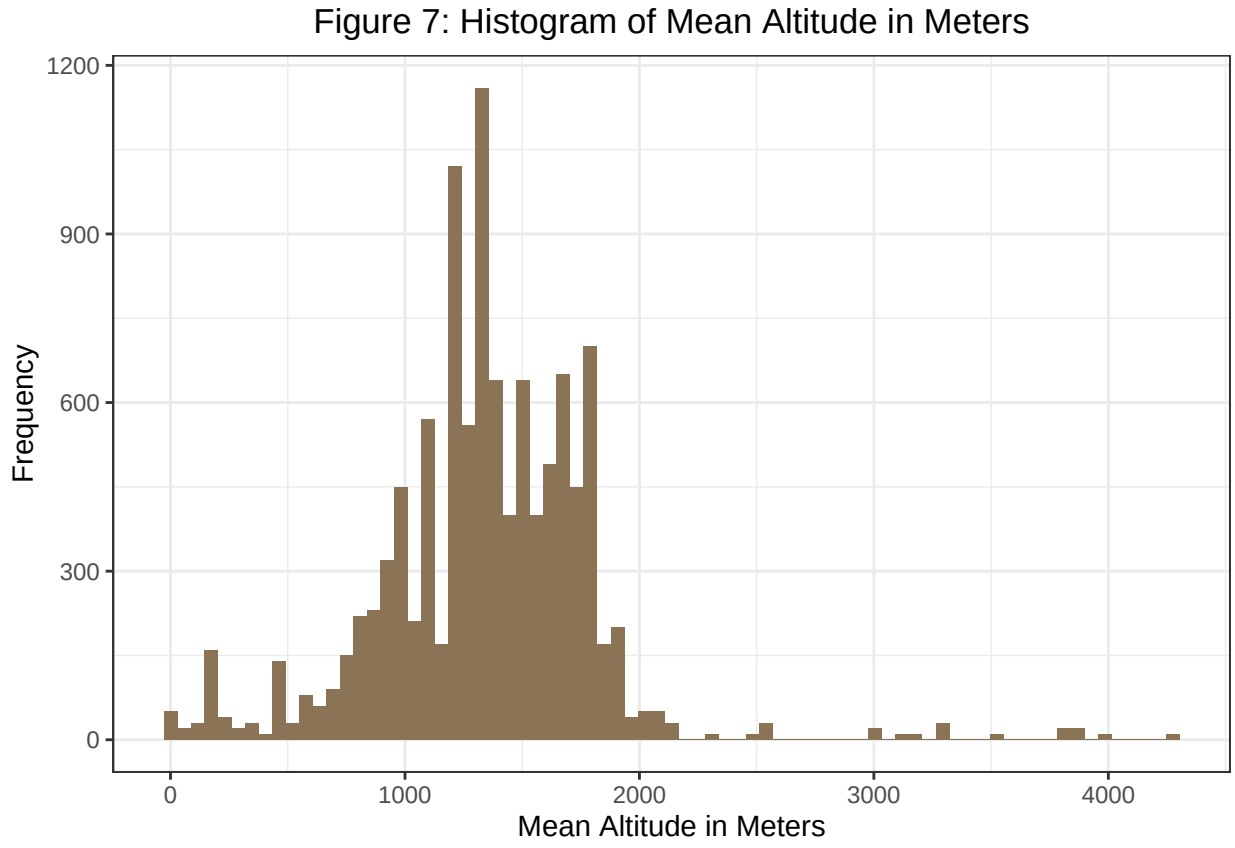


The density estimation results from Figure 6 indicate that the attributes of aroma, flavor, aftertaste, acidity, body, balance, and cupper points have a similar type of distribution with the values all around a 7.5 mean score. However, the attributes of uniformity, sweetness, and clean cup have a different distribution than the other 7 attributes. The distribution for these 3 attributes have less variability with mean scores around

10.0. There is also a smaller distribution of results for these 3 attributes around the score of 9.0. This would indicate that the majority of scores for these 3 attributes are normally 10.0, but a small number of samples do receive a lower score in the range around 9.0. From the results presented in Figure 6, it can be ascertained that the attributes of aroma, flavor, aftertaste, acidity, body, balance, and cupper points would have the greatest bearing on determining the total cup score.

Next, let us now look at the relationship between the coffee bean growing altitude and the desired quality of the coffee. There are some NA results and values that are illogical for this column variable, so we will filter the results to remove NA values and only include results between 2 and 5000 meters.

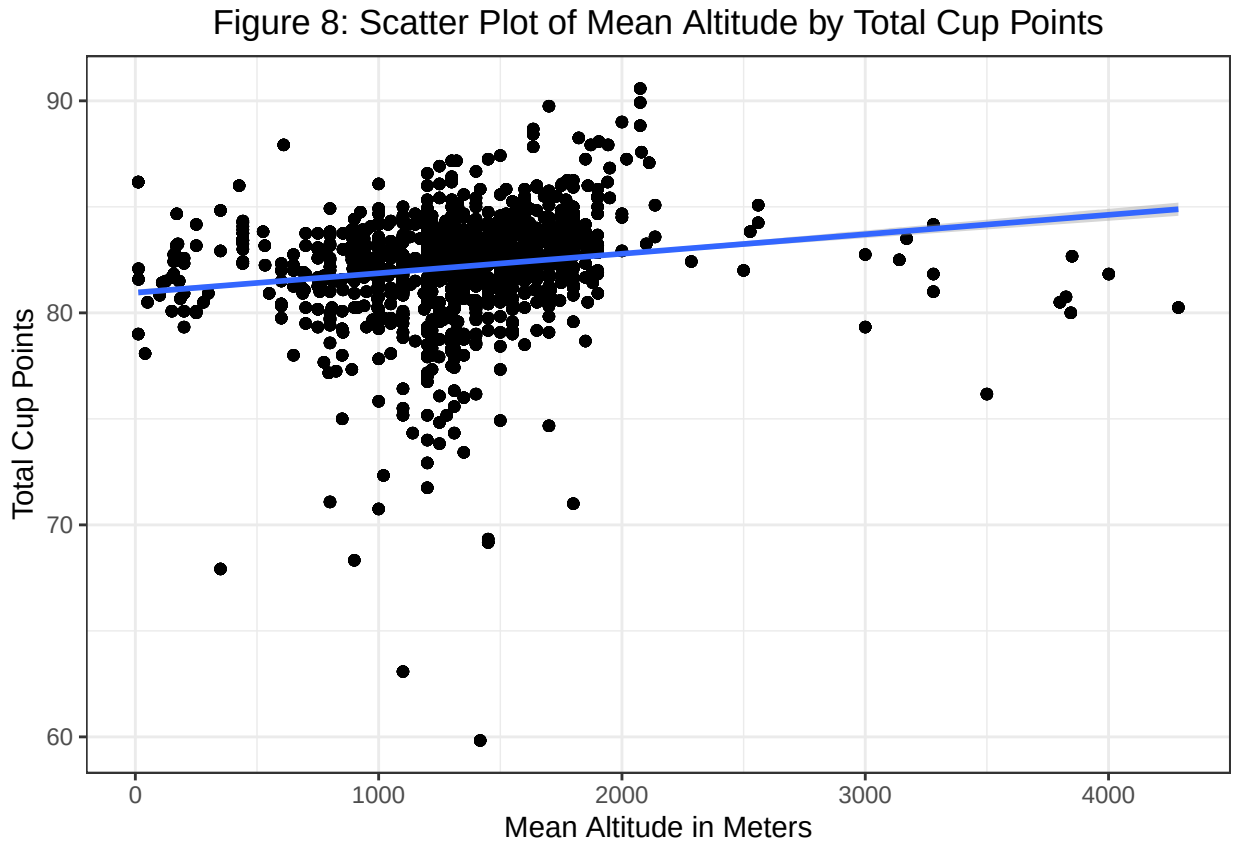
```
# Filter coffee_measurement to remove any NA entries for the column variable
# altitude_mean_meters and any altitude values 5000 meters or more.
# Create histogram of mean altitude values.
coffee_measurement %>%
  filter(!is.na(altitude_mean_meters))%>%
  filter(altitude_mean_meters <= 5000, altitude_mean_meters >=2)%>%
  ggplot(aes(altitude_mean_meters))+
  geom_histogram(fill="burlywood4", bins=75)+
  xlab("Mean Altitude in Meters")+
  ylab("Frequency") +
  ggtitle("Figure 7: Histogram of Mean Altitude in Meters") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



As demonstrated by the histogram in Figure 7, the majority of samples are between 500 to 2200 meters in mean altitude with the most frequent value occurring in the 1200 to 1300 meter range. Now let us determine

if a correlation exists between mean altitude and total cup points. To analyze this relationship, a scatter plot of these two variables will be used. Reference Figure 8 below.

```
# Filter out NA field in altitude_mean_meters column.
# Filter altitude_mean_meters column results to values between 2 and 5000 meters.
# Create scatter plot of altitude_mean_meters by total cup points.
coffee_measurement %>%
  filter(!is.na(altitude_mean_meters))%>%
  filter(altitude_mean_meters <= 5000, altitude_mean_meters >=2)%>%
  ggplot(aes(altitude_mean_meters, total_cup_points))+
  geom_point()+ geom_smooth(method = "lm", formula = y ~ x)+
  xlab("Mean Altitude in Meters")+
  ylab("Total Cup Points") +
  ggtitle("Figure 8: Scatter Plot of Mean Altitude by Total Cup Points") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



The scatter plot in Figure 8 above indicates a positive correlation between mean altitude and total cup points. From the plot, it appears values above 2200 meters may start to have opposite effect. As the majority of the data is contained within the range of 500 to 2200 meters, let us determine the correlation coefficient of this range.

```
# Filter out NA field in altitude_mean_meters column.
# Filter altitude_mean_meters column results to values between 500 and 2200 meters.
# Calculate correlation coefficient for altitude_mean_meters by total cup points.
```

```
coffee_measurement %>%
  filter(!is.na(altitude_mean_meters))%>%
  filter(altitude_mean_meters <= 2200, altitude_mean_meters >=500)%>%
summarize(correlation = cor(altitude_mean_meters, total_cup_points))
```

```
## # A tibble: 1 x 1
##   correlation
##         <dbl>
## 1       0.274
```

As noted above, the calculated correlation coefficient over the range of 500 to 2200 meter is 0.274. This value indicates a weak positive correlation, but is close to being within a moderate correlation range. Let us now look at the correlation between mean altitude and sample attribute to determine if a relationship is present.

```
# Filter out NA field in altitude_mean_meters column.
# Filter altitude_mean_meters column results to values between 500 and 2200 meters.
# Group sample attribute in measurement column.
# Calculate correlation coefficient for altitude_mean_meters by each sample attribute score.
coffee_measurement %>%
  filter(!is.na(altitude_mean_meters))%>%
  filter(altitude_mean_meters <= 2200, altitude_mean_meters >=500)%>%
  group_by(measurement)%>%
  summarize(correlation = cor(altitude_mean_meters, value), .groups ='drop')%>%
  arrange(desc(correlation))
```

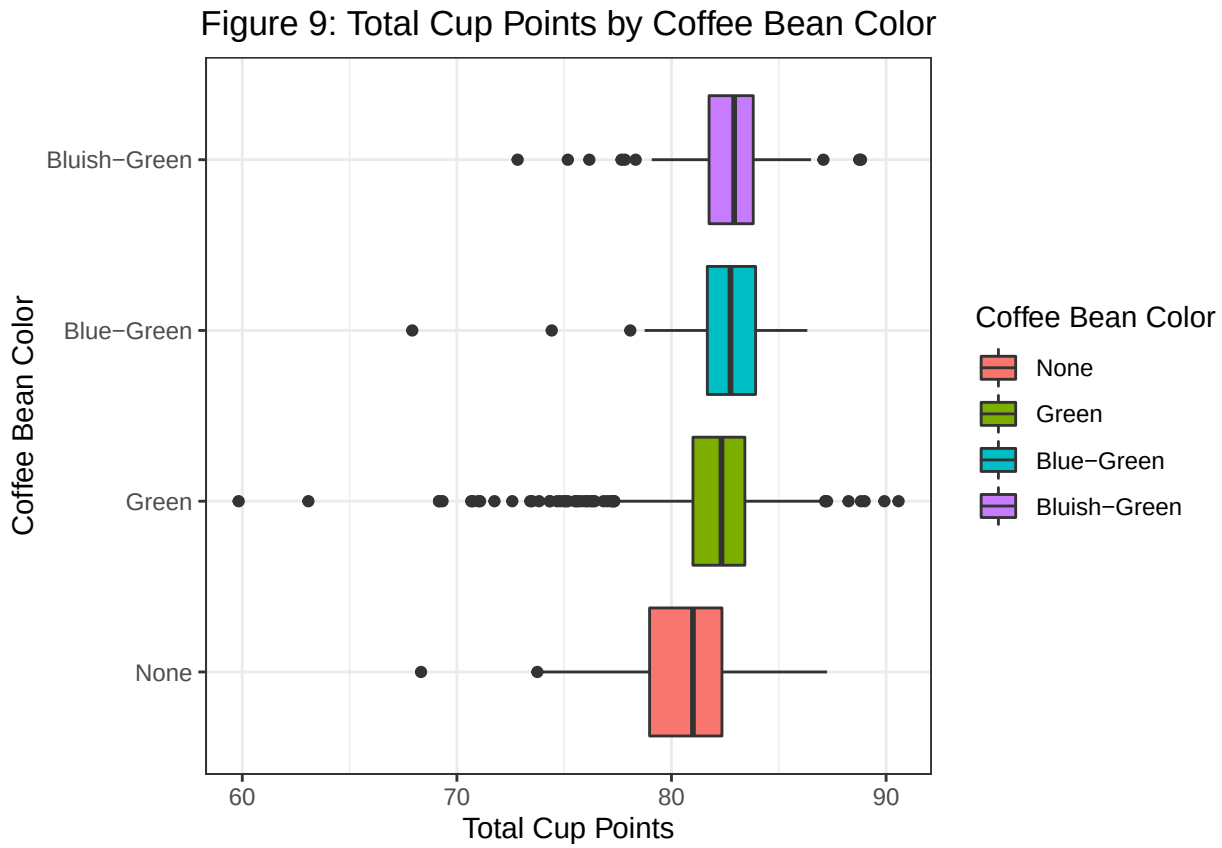
```
## # A tibble: 10 x 2
##    measurement   correlation
##    <chr>               <dbl>
##  1 balance            0.269
##  2 acidity            0.256
##  3 cupper_points      0.254
##  4 flavor             0.245
##  5 aroma              0.242
##  6 aftertaste         0.242
##  7 body               0.212
##  8 uniformity         0.0887
##  9 clean_cup          0.0884
## 10 sweetness          0.0871
```

The results above indicate a weak positive correlation between the attributes of aroma, flavor, aftertaste, acidity, body, balance, and cupper points, but little to no correlation among the attributes of uniformity, clean cup, and sweetness. This difference among the attributes had been previously noted during performance of the density estimation, giving further indication that these attributes do not have a large effect on the overall quality of the coffee.

Does the color of the coffee bean influence the total cup rate? To investigate, we will compare coffee bean color to total cup points using a boxplot of the results. The color column variable also contains some NA values, so for this analysis those samples will be filtered out. See Figure 9 below for the boxplot of this analysis.

```
# Filter out NA field in color column.
# Mutated to reorder by color and total cup points.
```

```r
# Create boxplot of color by total cup points.
coffee_ratings %>%
  filter(!is.na(color))%>%
   mutate(color = fct_reorder(color, total_cup_points))%>%
  ggplot(aes(total_cup_points, color, fill=color)) +
  geom_boxplot()+xlab("Total Cup Points")+ ylab("Coffee Bean Color") +
  ggtitle("Figure 9: Total Cup Points by Coffee Bean Color") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))+labs(fill="Coffee Bean Color")
```

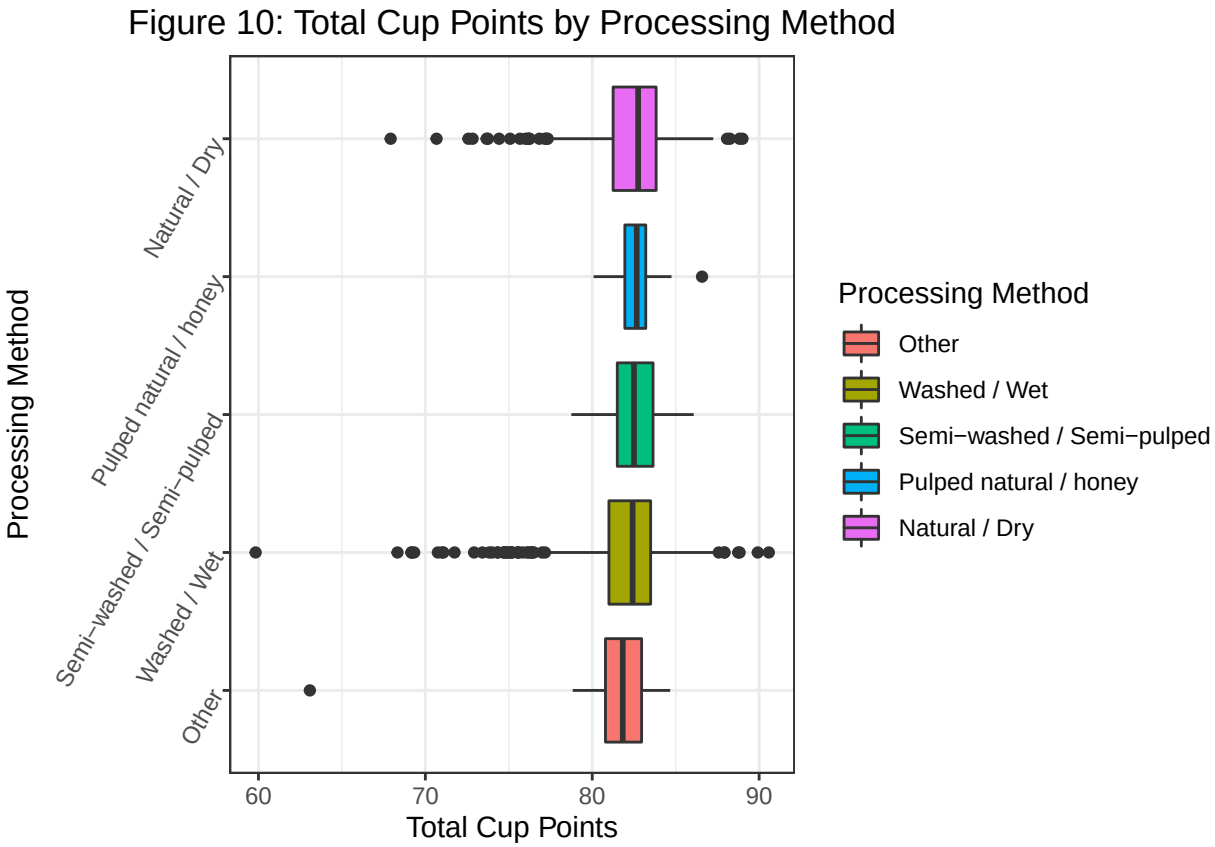


While the data set does not readily state the definition for none, it is inferred that this value represents the normal color variation for coffee beans while the other values represent the green to blueish hues present in some coffee varieties. In Figure 9, we see the boxplots values for the associated coffee bean colors as compared to total cup points. From the results of this plot, it is apparent that the coffee beans with green to blue hues have a higher median total cup store than those with none. It is also evident that bluer hues have a slightly higher median score than the green coffee beans. However, there is a lot of variation in the green coffee bean color, which could be responsible the differences seen in the plot.

The final analysis of this data set will examine the processing method for the coffee bean sample to determine if this has any effect on the total cup score. Like other column variables in this data set, there are a few NA entries that will be filtered out. Reference Figure 10 below for the boxplot comparing processing method and total cup points.

```r
# Filter out NA field in processing method column.
# Mutated to reorder by processing method and total cup points.
# Create boxplot of processing by total cup points.
```

```
coffee_ratings %>%
  filter(!is.na(processing_method))%>%
  mutate(processing = fct_reorder(processing_method, total_cup_points))%>%
  ggplot(aes(total_cup_points, processing, fill=processing)) +
  geom_boxplot()+ xlab("Total Cup Points")+ ylab("Processing Method") +
  ggtitle("Figure 10: Total Cup Points by Processing Method")+ theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))+labs(fill="Processing Method")+
  theme(axis.text.y = element_text(angle = 60, hjust = 1))
```



As seen above in Figure 10, the following processing methods were utilized for the samples in this data set: Natural/Dry, Pulped/Honey, Semi-washed/Semi-pulped, Washed/Wet, and Other. The results of the boxplot do not significantly indicate a difference between processing methods as compared to total cup points. The only method that indicates a somewhat lower median score would be the Other processing method designation when compared to the remaining listed processing methods.

Through our analysis of this coffee rating data set, we have looked at several variables contained within the set. This analysis determined that coffee bean species has an association with a higher quality coffee as determined through total cup point. It was found that the Arabica species was favored over the Robusta species of coffee beans. It was shown that different varieties of coffee beans are associated with a high median value and it was noted that the top 3 varieties were SL34, SL28, and SL14, which were all developed by Scott Agricultural Laboratories during the 1930s. When examining the coffee bean country of origin, the analysis showed that country of origin led to higher overall total cup scores. A significant difference was shown for coffee originating from Ethiopia, but other countries such as Costa Rica, Columbia, and Uganda also had higher overall median scores with the highest rated coffee from the United States being grown in Hawaii. We then looked at the sample attributes of aroma, flavor, aftertaste, acidity, body, balance, cupper

points, uniformity, sweetness, and clean cup, which when added together results in the total cup point score for each coffee sample. In this analysis, we looked at the density estimation for each attribute and found that the attributes of aroma, flavor, aftertaste, acidity, body, balance, and cupper points where similar in distribution centered around a similar mean score. However, the attributes of uniformity, sweetness, and clean cup had a different distribution and were centered around a different mean score. This indicated that aroma, flavor, aftertaste, acidity, body, balance, and cupper points had more of an influence in overall total cups points. The effects of altitude on coffee beans were also examined and this analysis found a weak positive correlation between mean altitude in meters and total cup points. This was also shown at the sample attribute level by calculating the correlations for each of the 10 sample attributes compared to mean altitude. Once again, we saw where the attributes of aroma, flavor, aftertaste, acidity, body, balance, and cupper points had weak positive correlations with mean altitude, but uniformity, sweetness, and clean cup had little to no correlation with mean altitude. When examining coffee bean color, it was found that the green to blueish green color resulted in higher median total cup point scores. Finally, we looked for any association between the coffee bean processing method and total cup point. This analysis did not indicate any major difference in determining overall total cup points.

During the analysis certain limitations were noted, the most prevalent one being the lack of Robusta coffee bean species samples in the sample set as compared to the Arabica species. Additional amounts of the Robusta species would have provided greater exploration into the difference between these species. Due to the size of the data set and the column variables present, additional comparisons could be performed. Further explorations into the different farms, owners, and harvest years could provide a more extensive analysis into the factors effecting overall coffee bean quality.

Upon completion of the analysis, the question that begs to be asked is, “So, what coffee is the best?” Ultimately, the answer to this question resides with the taste of the person drinking the coffee. However, from the results of this analysis, there will definitely be some of the noted variables found during this analysis in future coffee selections by this data analyst.