From Bean to Cup: A Data Science Perspective on Coffee Bean Quality Analysis and Optimization

Introduction

Coffee is one of the most widely consumed beverages worldwide, with over 400 billion cups consumed annually. Coffee production is a complex process that involves multiple factors such as altitude, processing method, and variety of coffee bean. To understand the production and characteristics of coffee beans from various regions and countries, datasets containing information on coffee production have been collected and analyzed.

In this analysis, we explore the coffee dataset to gain insights into the factors that affect the quality and characteristics of coffee beans. We investigate the impact of various variables such as altitude, processing method, and variety of coffee bean on the quality of coffee beans as measured by the Total Cup Points. Additionally, we examine the relationship between the number of bags and Total Cup Points and investigate the impact of unique regions and countries on coffee production.

Through this analysis, we aim to provide valuable insights into the production and characteristics of coffee beans, which can be used to inform further research and improve the quality of coffee production. Understanding the factors that impact coffee quality and production can help coffee growers and producers make informed decisions about the cultivation and processing of coffee beans, ultimately leading to a better cup of coffee for consumers.

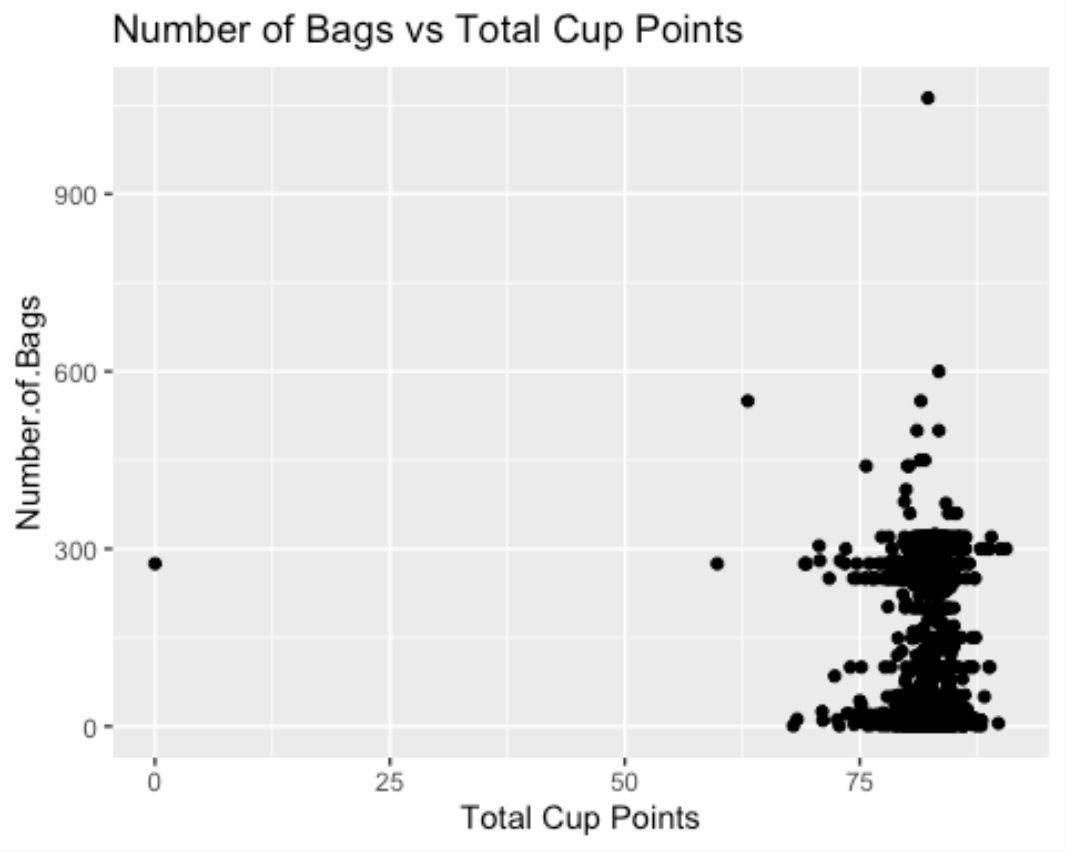
Reference Variables

This data set does not have an attribute that can be compared to find the best coffee bean or kind of coffee bean, that I understand. The closest attribute, that looks like a possibility to use is `Total.Cup.Points`. This variable at first does not make much sense as to ho it was calculated. Lets take some time to get a better understanding of this variable, so that it can be used for further.

Total.Cup.Points vs. Number.of.Bags

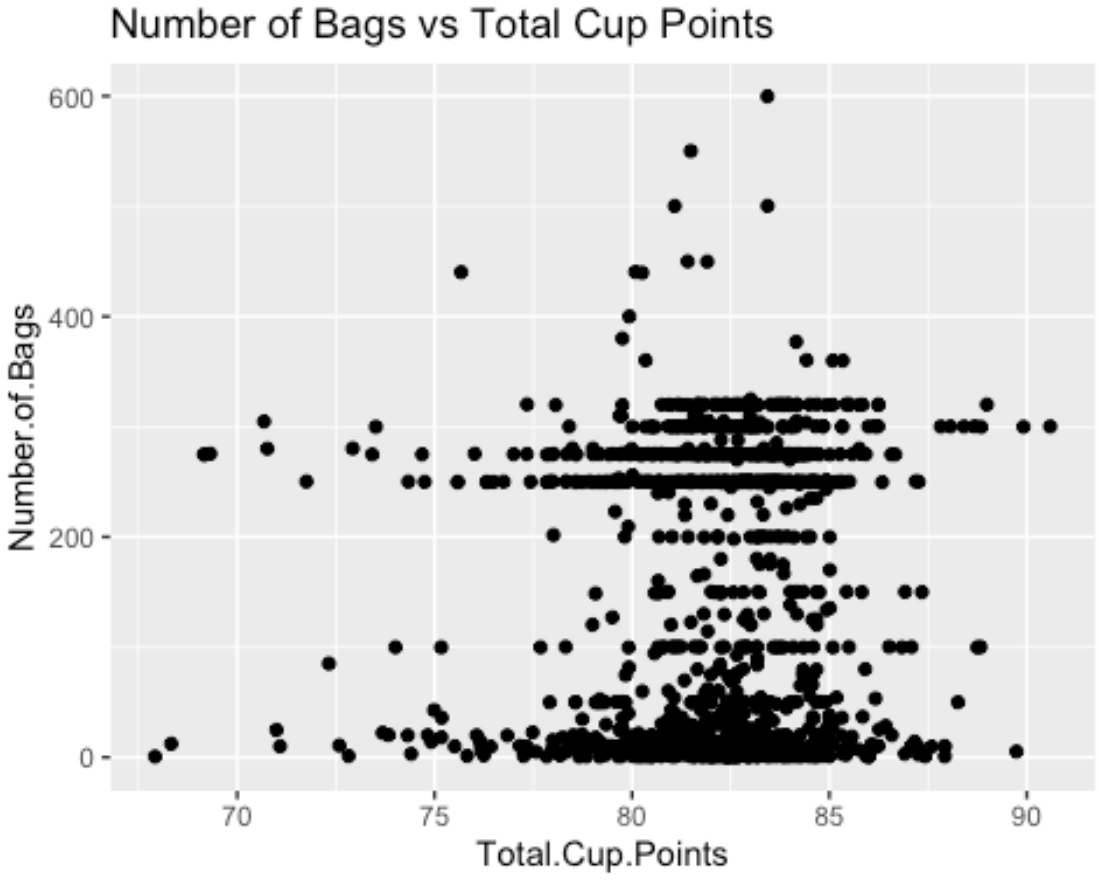
```
coffee %>%
  ggplot(mapping = aes(x = Total.Cup.Points, y = Number.of.Bags)) +
  geom_point() +
  labs(title = "Number of Bags vs Total Cup Points", x = "Total Cup
Points")
```

```
## Warning: Removed 12 rows containing missing values
(`geom_point()`).
```



Based on the graph above, it is not the representation of the data. Most of the data is all clumped in the bottom right hand corner of the graph. We can get a better representative of the graph and we will have an easier time making accurate observations. We can put a cap on the number of bags to 600, as all the data but 1 is below 600. We can also filter out any data with less than 50 `Total Cup Point" as most of the coffee beans have more than 65 total cup points.

```
coffee %>%
  filter(Number.of.Bags <= 600, Total.Cup.Points > 65) %>%
  ggplot(mapping = aes(x = Total.Cup.Points, y = Number.of.Bags)) +
  geom_point(position = "jitter") +
  labs(title = "Number of Bags vs Total Cup Points")
```



I can see that this is not a very correlated graph, nor is it a readable graph either. This does, however, confirm that that there is no relationship between `Total.Cup.Points` and `Number.of.Bags`. These variable must mean completely different things, such as

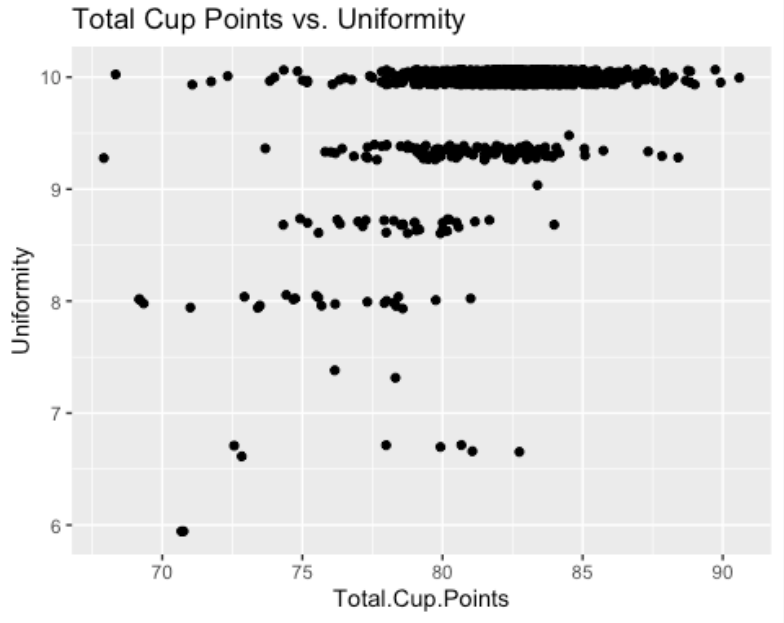
Number.of.Bags might be recording the number of bags of beans the producers could make given their circumstances before circumstances before the public influenced their decision. The Total.Cup.Points circumstances before the public influenced their decision. The Total.Cup.Points circumstances before the public influenced their decision. The Total.Cup.Points difference is between Number.of.Bags and Total.Cup.Points. Let's use Total.Cup.Points might be a variable resulting from a combination of several other attribute of the coffee bean, such as the Aroma, Body, and Flavor, just to name a few.

Now that we have an idea of what the Total.Cup.Points, and see what attribute affect this the most.

Trying to Understand Total.Cup.Points

Let's do a little more research in if there is any affect on the `Total.Cup.Points`, based on`Aroma`, `Body`, and `Flavor` of coffee beans. Before preforming these analysis I am predicting that `Total.Cup.Points` works like `Uniformity` does. Uniformity in terms of coffee beans refers to the consistency of characteristics among the beans in a batch or lot. This includes the size, shape, color, moisture content, and other physical attributes of the beans. So my guess is that the `Aroma`, `Body`, and `Flavor` and other variables like these, of coffee beans, directly affect both the `Total.Cup.Points` and `Uniformity`, and after studying each variable we will confirm our evidence and research by use a regression equation to predict both the `Total.Cup.Points` and `Uniformity`.

```
coffee %>%
  filter(Total.Cup.Points > 65) %>%
  ggplot(mapping = aes(x = Total.Cup.Points, y = Uniformity)) +
  geom_point(position = "jitter") +
  labs(title = "Total Cup Points vs. Uniformity")
```

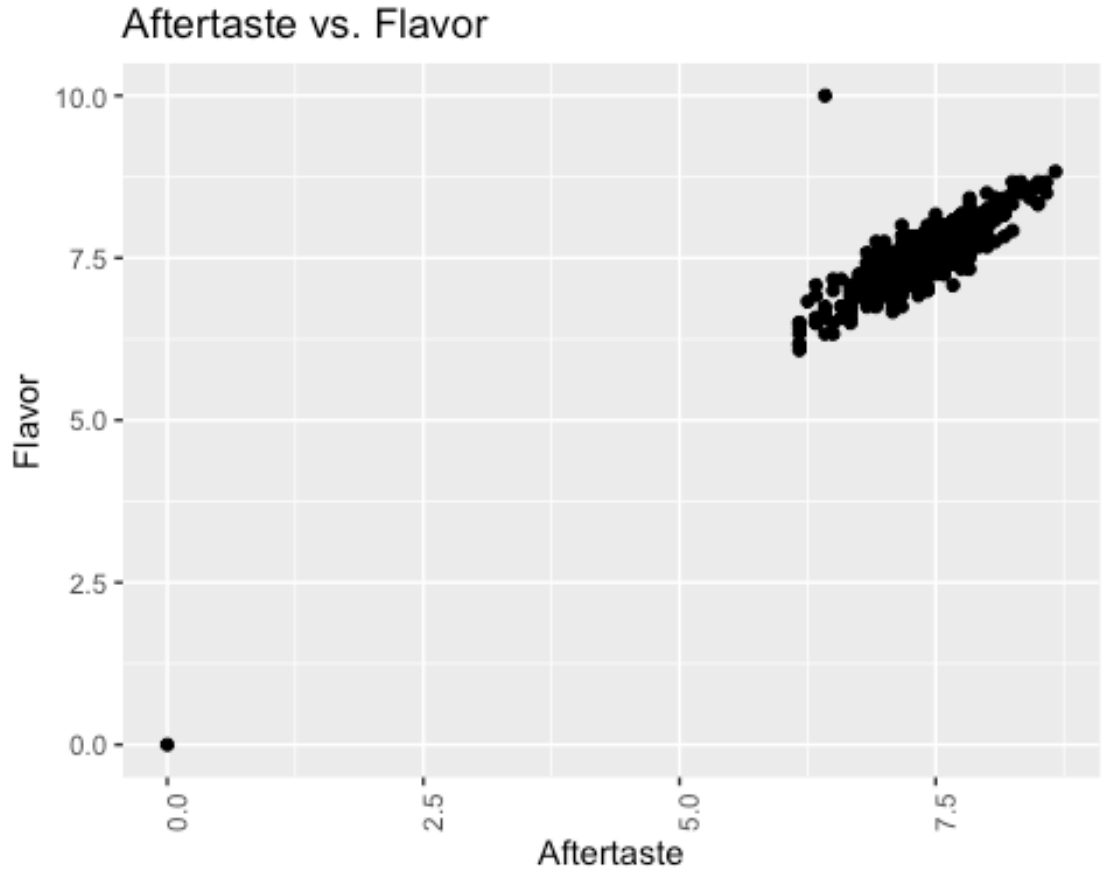


We can see that there is a somewhat of a weak positive correlation between `Total.Cup.Points` and `Uniformity`. The correlation is not strong enough to make accurate predictions probably. Looking at the data set more, there are some other variables that might have more of an effect on Uniformity, such as, `Defects`, and `Color`, to name a few. For now we are going to be using just `Total.Cup.Points`.

###Taste Let's by looking at all of the variables having to do with the taste of a coffee bean. This will help us to understand how each variable that hs to do with taste acts compared to the other variables in the "taste" category.

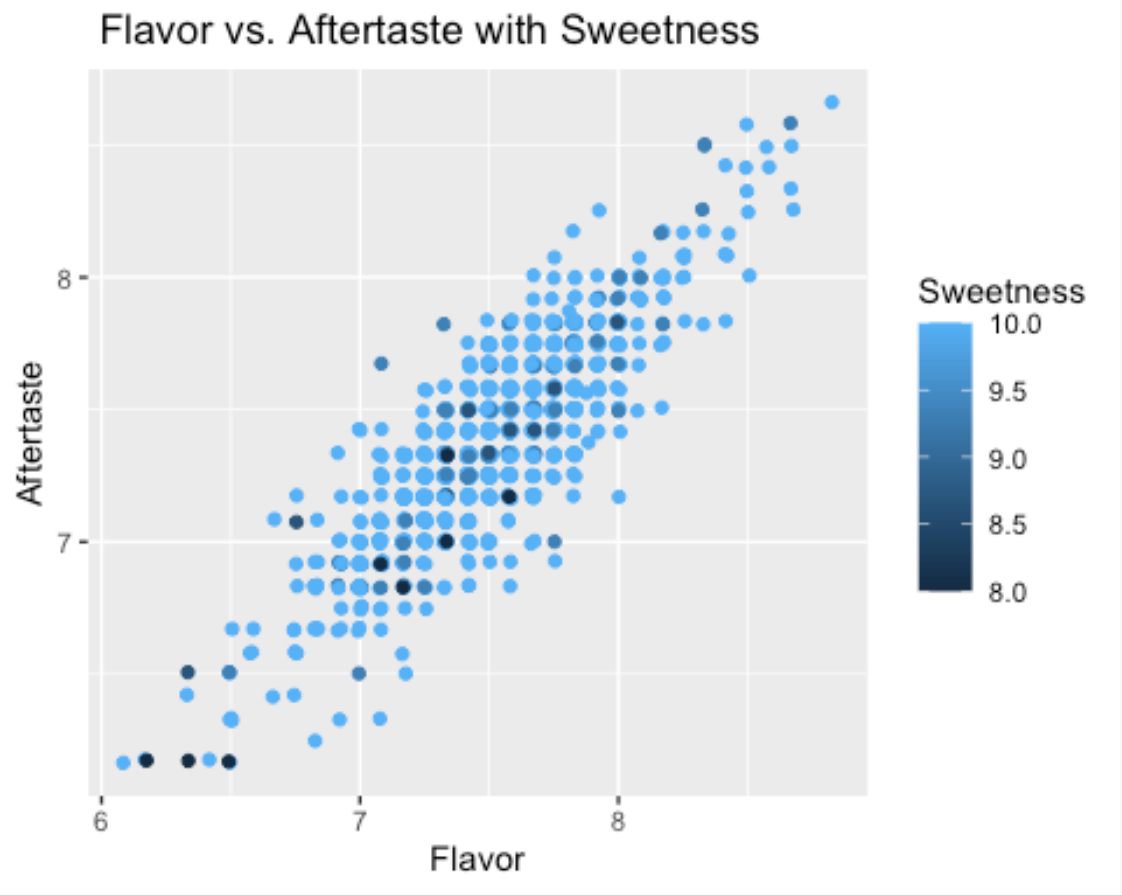
```
coffee %>%
  ggplot(mapping = aes(x= Aftertaste, y = Flavor)) +
  geom_point() +
  labs(title = "Aftertaste vs. Flavor") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.title = element_text(size = 12),
        legend.text = element_text(size = 8, margin = margin(t = 2)))
```

```
## Warning: Removed 11 rows containing missing values
(`geom_point()`).
```



This is a decent graph but most of our data is ver condensed to accommodate the outlines of the data set. By filtering these out, we can get a better graph to make conclusions. We can also flip the variables, it make more sense for `Flavor` to be first, rather than `Aftertaste`. To account for any overlapping data, lets also use `jitter`.

```
coffee %>%
  filter(Flavor > 5, Flavor < 9, Sweetness > 6.75) %>%
  ggplot(mapping = aes(x= Flavor, y = Aftertaste, color = Sweetness)) +
  geom_point(position = "jitter") +
  labs(title = " Flavor vs. Aftertaste with Sweetness”) +
theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.title = element_text(size = 12),
        legend.text = element_text(size = 8, margin = margin(t = 2))
```



The filters that have been preformed, make it easier to read the graph. Their is a better representation of variance in aftertaste, now that the range of the aftertaste has been

condensed. By also filtering out outlines, the plot is not overlapping and we have a more precise graph.

Like the last graph, we are still at the conclusion that moisture does not effect the taste very much. We can make a new conclusion that the more flavor a coffee bean has the more of an aftertaste the bean will give. This can tell us that both `Flavor` and `Aftertaste` have at least some sort of correlation between them. This makes sense, as the more flavorful something is, the longer we will taste it in our mouths. This might be giving an idea into what kind of coffee bean someone like.

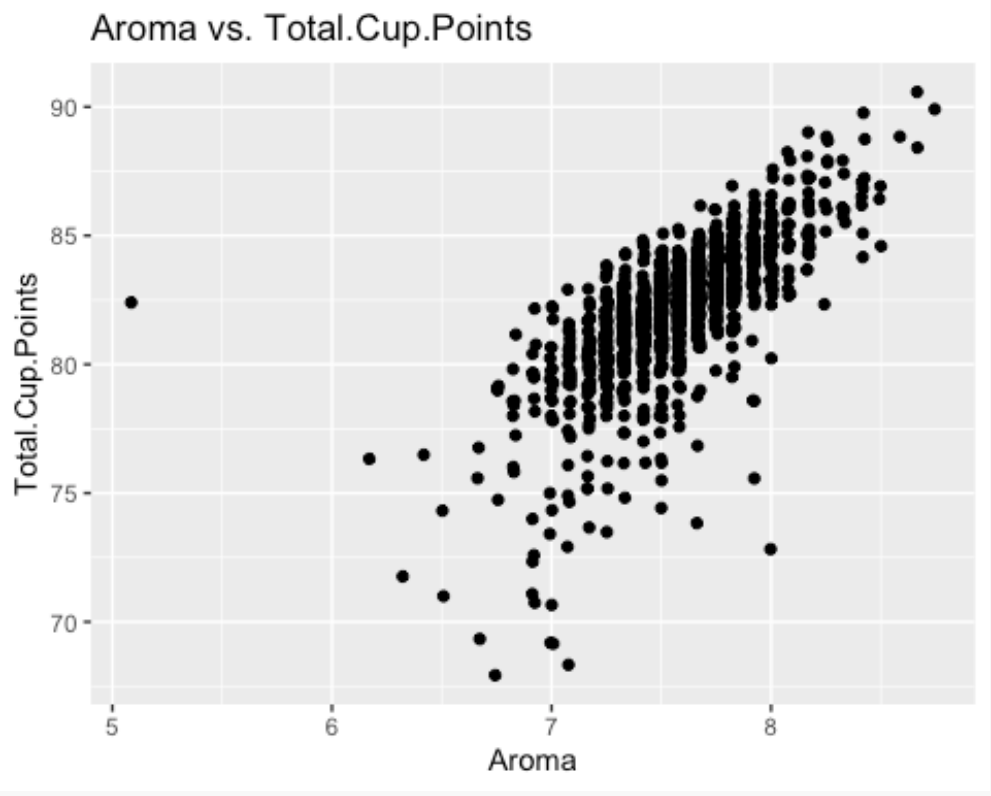
We can also see that when we color by `Sweetness`, we see that there is not much variation within it, therefore it probably is not helpful in using it to predict the `Total.Cup.Points`.

In conclusion all of the "taste" kind of variables pretty much act the same way. For the sake of the study, we are going to use `Flavor` as a the representative, of all the variable having to do with taste. Let's just hope that `Flavor` has come to play and does not make the other variables in its family mad.

The other variable that I think play a key role in the `Total.Cup.Points`, are `Body`, `Acidity`, and `Aroma`. Along with `Flavor`, these variables I think make up a lot consists of `Total.Cup.Points`.

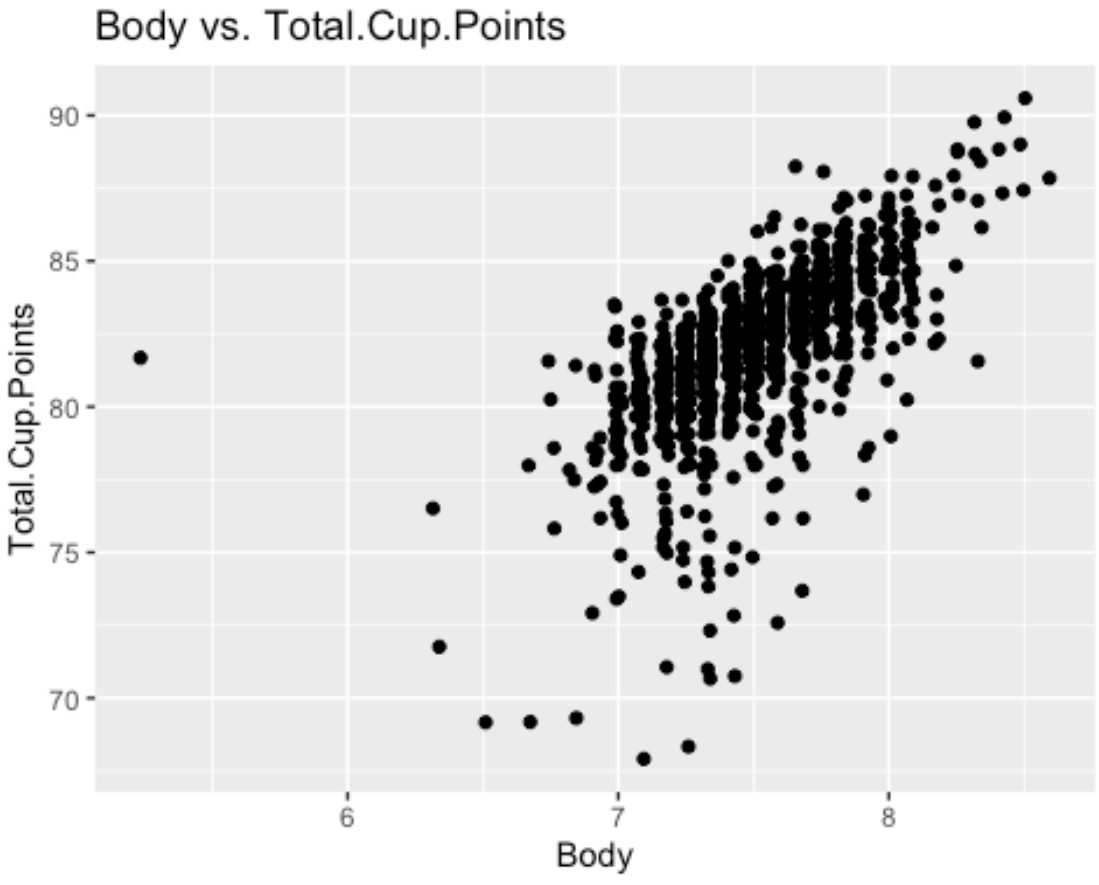
Aroma

```
coffee %>%
  filter(Total.Cup.Points > 65) %>%
  ggplot(mapping = aes(x = Aroma, y = Total.Cup.Points)) +
  geom_point(position = "jitter")+
  labs(title = "Aroma vs. Total.Cup.Points “)
```



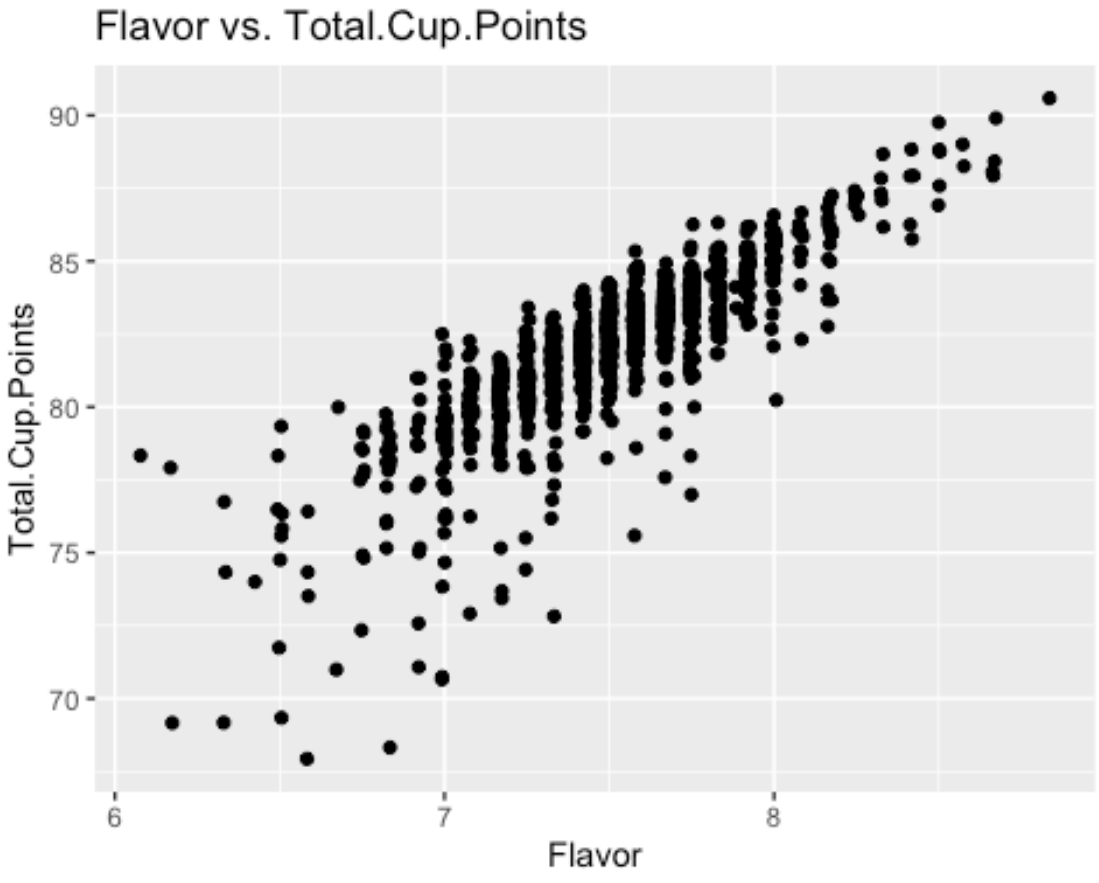
Body

```
coffee %>%
  filter(Total.Cup.Points > 65) %>%
  ggplot(mapping = aes(x = Body, y = Total.Cup.Points)) +
    geom_point(position = "jitter") +
  labs(title = "Body vs. Total.Cup.Points")
```



Flavor

```
coffee %>%
  filter(Total.Cup.Points > 65) %>%
  ggplot(mapping = aes(x = Flavor, y = Total.Cup.Points)) +
    geom_point(position = "jitter") +
  labs(title = "Flavor vs. Total.Cup.Points")
```



By graphing `Aroma`, `Body`, and `Flavor`, we can observe that there is some correlation between each variable respectively. But we are not getting the same exact graph for each variable when graphed individually with `Total.Cup.Points`. There is still a chance that all of these variable have a role in contributing to `Total.Cup.Points`, lets calculate a multiple regression line and see if we can use these three variables, in order to predict the Total Cup Points, of a coffee bean. lets start by making a smaller data set of just these attributes

Building a Multiple Regression Equation

```
coffeeSmall <- coffee %>%
  dplyr::select(Aroma, Body, Flavor, Total.Cup.Points)
```

Before build a whole regression equation, lets form a correlation matrix to make sure we see similar results to what can be taken away from the graphs above.

```
coffeeSmall <- coffeeSmall %>%
  filter(!is.na(Aroma), !is.na(Total.Cup.Points))

cor(coffeeSmall)
```

```
##                      Aroma      Body    Flavor Total.Cup.Points
## Aroma            1.0000000 0.6956131 0.8135973        0.7966370
## Body             0.6956131 1.0000000 0.7612108        0.7763467
## Flavor           0.8135973 0.7612108 1.0000000        0.8776274
## Total.Cup.Points 0.7966370 0.7763467 0.8776274        1.0000000
```

We can see by the correlation graph above that `Aroma`, `Body`, and `Flavor`, all have similar correlation patterns to `Total.Cup.Points`. So this chart confirms the information gathered from looking at the graphs made above. Lets now build a regression equation to estimate the `Total.Cup.Points` of a coffee bean.

```
coffeeModel <- lm( Total.Cup.Points ~ Aroma + Body + Flavor, data =
coffee)
summary(coffeeModel)

##
## Call:
## lm(formula = Total.Cup.Points ~ Aroma + Body + Flavor, data =
coffee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6591  -0.3437   0.2386   0.7233   4.3959
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.7495     0.9444  16.676   <2e-16 ***
## Aroma         1.8034     0.1968   9.164   <2e-16 ***
## Body          2.1740     0.1858  11.700   <2e-16 ***
## Flavor        4.8393     0.2064  23.451   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 1303 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.8103, Adjusted R-squared:  0.8099
## F-statistic:  1856 on 3 and 1303 DF,  p-value: < 2.2e-16
```

Let's also look at the p-values between the different attributes and `Total.Cup.Points`, to make sure that the correlation between them all is significant.

By running the function above, and see that our , we can build a multiple regression equation. This can be seen down below.

```
Total.Cup.Points = 15.75 + (1.80 * Aroma) + (2.17 * Body) + (4.84 *
Flavor)
```

If calculations have been done correctly, we can predict the `Total.Cup.Points` of a coffee bean. Let's try it out with some different numbers.

We can also observe that aside from the coefficient, `Flavor` is multiplied to the next biggest number. This may be that `flavor` is the strongest predictor. Let's run the `lm.beta` function to either confirm or deny our results.

```
lm.beta(coffeeModel)

##
## Call:
## lm(formula = Total.Cup.Points ~ Aroma + Body + Flavor, data =
coffee)
##
## Standardized Coefficients::
## (Intercept)       Aroma         Body       Flavor
##          NA   0.1941813   0.2222581   0.5504568
```

By running `lm.beta`, we can see that my prediction was correct. Based on the data and attributes being studied, we can conclude that `Flavor` is the highest predictor when estimating `Total.Cup.Points`.

Now we are going to try to test our equation with data about the coffee bean with strong flavor and a strong aroma

```
Total.Cup.Points <-  15.75 + (1.80 * 9.2) + (2.17 * 8.7) + (4.84 *
9.9)
Total.Cup.Points

## [1] 99.105
```

When using the equation made above, we can see that when we use `9.2` for the Aroma, `8.7` for the body, and `9.9` for the Flavor of the Coffee bean, we an estimated `99.105` for the `Total.Cup.Points`. By look at the dataset, we can compare the number that we put in the equation and get similar `Total.Cup.Points`, to what is in the data set.

After all of that works we have made a pretty good idea of what `Total.Cup.Points` accounts for and now we can use this to see what is produced between the various countries, and regions of the world.

Studying different Origins of Coffee Beans

Figuring out ways to group the coffee beans

I want to look at coffee beans from different areas regions or countries. To figure how to graph the data correctly, lets figure out how many different regions and countries that are represented within the data set. Upon further reference from our textbook, we can use the `Unique()` function to figure this out.

```
unique(coffee$Country.of.Origin)
```

```
##  [1] "Ethiopia"
##  [2] "Guatemala"
##  [3] "Brazil"
##  [4] "Peru"
##  [5] "United States"
##  [6] "United States (Hawaii)"
##  [7] "Indonesia"
##  [8] "China"
##  [9] "Costa Rica"
## [10] "Mexico"
## [11] "Uganda"
## [12] "Honduras"
## [13] "Taiwan"
## [14] "Nicaragua"
## [15] "Tanzania, United Republic Of"
## [16] "Kenya"
## [17] "Thailand"
## [18] "Colombia"
## [19] "Panama"
## [20] "Papua New Guinea"
## [21] "El Salvador"
## [22] "Japan"
## [23] "Ecuador"
## [24] "United States (Puerto Rico)"
## [25] "Haiti"
## [26] "Burundi"
## [27] "Vietnam"
## [28] "Philippines"
## [29] "Rwanda"
## [30] "Malawi"
```

```
## [31] "Laos"
## [32] "Zambia"
## [33] "Myanmar"
## [34] "7.42"
## [35] "Natural / Dry"
## [36] "0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660"
## [37] "Mauritius"
## [38] NA
## [39] "7.25"
## [40] "m"
## [41] "Cote d?Ivoire"
## [42] "India"
## [43] "oriente"
```

WE can see by doing this, we get 43 different entries represented, with applying filters, to just get Countries and not numbers, we get 40 different countries represented. Lets see how many unique regions are represented in this data set as well.

When running this, we can see that there are 341 different regions. There are to many different regions to handle in order to build a easy to read graph. We are going to avoid using Region, as it might be a little to advanced to handle for this project. I can also see that there are a lot of different regions, That I have never heard of and there would not be useful for this project.

```
unique(coffee$altitude_mean_meters)
```

```
##   [1]   2075.0000   1700.0000   2000.0000          NA   1635.0000
1822.5000
##   [7]   1905.0000   1872.0000   1943.0000    609.6000   2080.0000
1500.0000
##  [13]   1450.0000   1850.0000   2019.0000   1300.0000   1320.0000
2112.0000
##  [19]   1250.0000   1950.0000   1400.0000   1200.0000   1775.0000
1800.0000
##  [25]   1941.0000     12.0000   1000.0000   1754.0000   1860.0000
1650.0000
##  [31]   1750.0000    426.7200   1600.0000   1900.0000   1524.0000
1417.3200
##  [37]   1350.0000   1680.0000   1731.2640   1770.0000   1550.0000
1325.0000
##  [43]   2560.0000   2136.0000      1.0000   1580.0000   1100.0000
1325.8800
```

```
##  [49]    800.0000   1620.0000    350.0000    925.0000    170.0000
1150.0000
##  [55]   1310.6400   1219.2000   1575.0000    900.0000    442.0000
1275.0000
##  [61]   1050.0000   1493.5200   1386.8400   1099.4136   1170.0000
1116.7872
##  [67]   1565.0000   3280.0000   1606.5000   1530.0000    250.5000
1038.5000
##  [73]    968.0000   1480.0000   1706.8800   1260.0000    890.0000
934.0000
##  [79]    950.0000   2527.0000    525.0000   1676.5000   1218.0000
1499.9208
##  [85]    750.0000    853.4400    700.0000   1625.0000   1356.6648
1128.3696
##  [91]   1799.0000    940.0000   1402.0800    439.0000   1330.0000
1813.0000
##  [97]   1227.0000   1280.1600   1880.0000   2100.0000   1261.0000
1483.0000
## [103]    175.0000   1901.0000    168.0000   1653.0000    532.0000
1089.0000
## [109]   1380.0000    905.0000    894.0000    872.0000   1473.0000
1645.0000
## [115]   1371.6000    775.0000   1059.0000  11000.0000   1248.0000
1040.0000
## [121]    157.8864    650.0000   1442.0000    165.0000   1296.0000
1240.0000
## [127]   3850.0000    200.0000   1317.0000    975.0000   1679.0000
2285.0000
## [133]   1560.0000   1396.0000    982.0000   1268.0000   1676.4000
600.0000
## [139]   1752.6000    441.0000   1180.0000   1338.0000    533.0000
1390.8024
## [145]   1144.0000     13.0000    944.0000   2500.0000   1020.0000
1877.0000
## [151]    981.0000    690.0000    973.0000   4001.0000   1425.0000
1556.0000
## [157]    774.0000    150.0000 190164.0000    850.0000   1600.2000
180.0000
## [163]   1599.0000    825.0000    125.0000    110.0000   1422.0000
763.0000
## [169]   1130.0000    680.0000    995.0000    695.0000   1456.0000
1642.0000
```

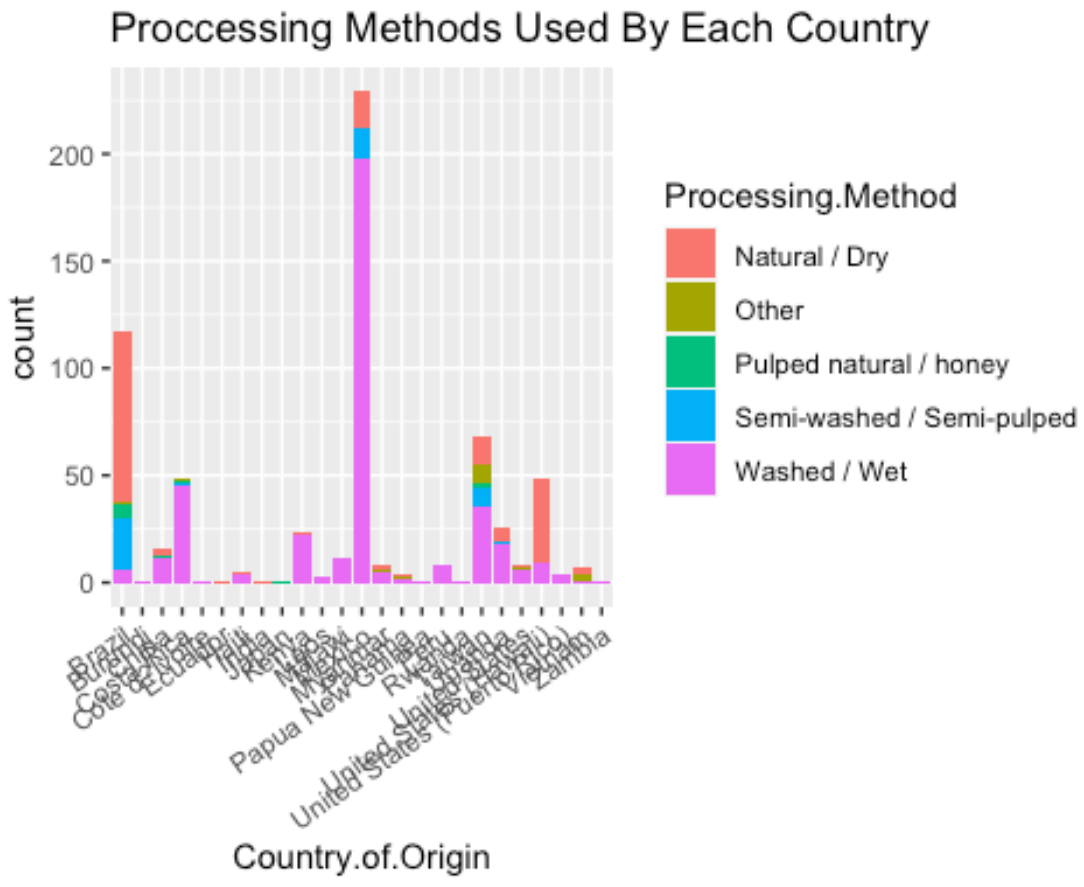
```
## [175]    758.0000   1525.0000   1066.8000    550.0000    688.0000
300.0000
## [181]   1186.0000   1210.0000    100.0000   3825.0000 110000.0000
185.0000
## [187]   1383.7920     50.0000    280.0000   3800.0000   1859.2800
914.4000
## [193]   4287.0000    808.0000   1187.5000    250.0000   3845.0000
1022.0000
## [199]   1264.0000   3500.0000   1280.0000   1140.0000
```

At first glance, I thought that this data set only had one columns of `Altitude which forced me to take a decent chunk of time in order to clean the data. After further diging through the data set, I found that the altitude has bee cleaned for me. Lets use for some further investigations.

Processing Method

Let';s now look into what `Processing.Method` is used the most.

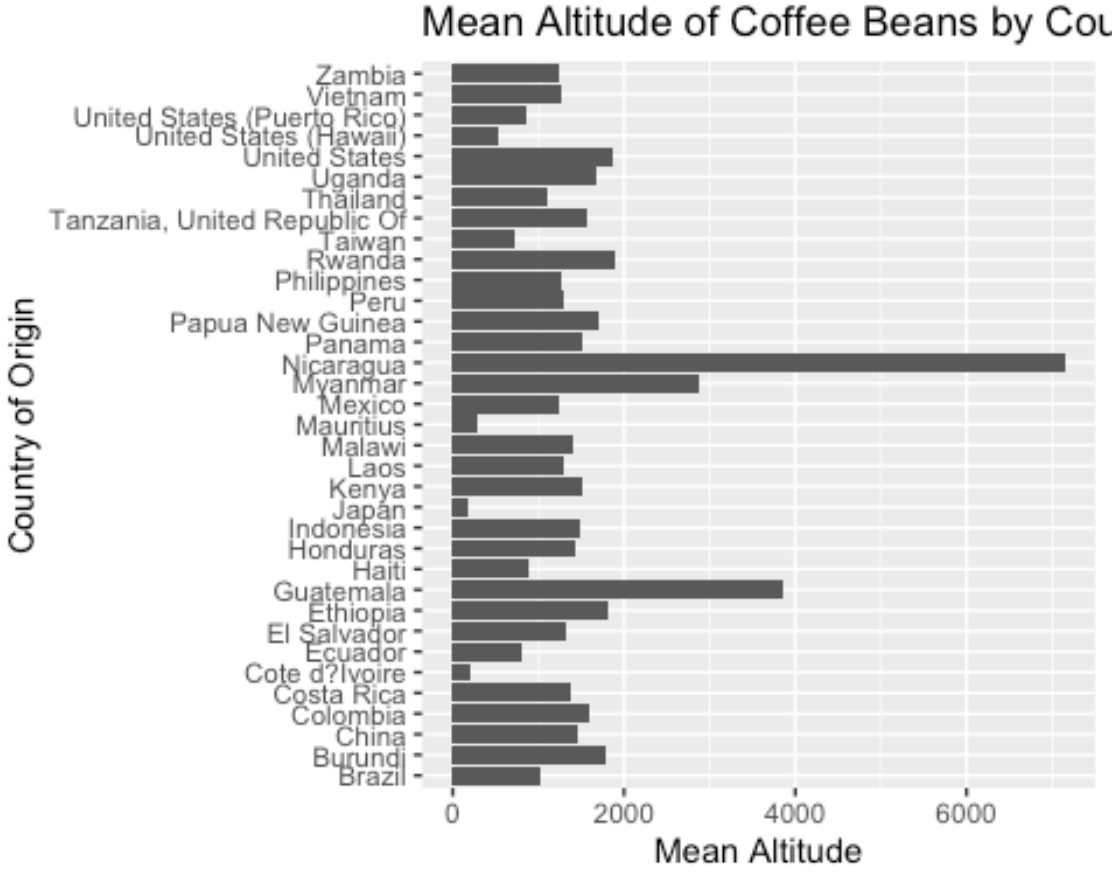
```
coffee %>%
  filter(!grepl("[0-9]", Country.of.Origin),
         !grepl("[a-zA-Z]{8,}", Country.of.Origin),
         !grepl("[0-9]", Processing.Method),
         Processing.Method != "N/A") %>%
  group_by(Country.of.Origin) %>%
  ggplot(mapping = aes(x = Country.of.Origin, fill =
Processing.Method)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 35, hjust = 1))+
  labs(title = "Processing Methods Used By Each Country")
```



WE can see by comparing the `Country.Of.Origin` to the count of how many times a `Proccessing.Method` is used, that `Washed/Wet` is the most popular method for production of coffee beans. We can see that there Mexico has the largest count, and that is probably due to the fact that Mexico is a major producer of coffee in the world. Some countries are real big. Within each country there might be different variations in altitude, lets compare the `altitude` to `Processing.Method` and see if the altitude changes any results.

Altitude

```
coffee %>%
  group_by(Country.of.Origin) %>%
  summarise(mean_altitude = mean(altitude_mean_meters, na.rm = TRUE))
%>%
  filter(mean_altitude > 0) %>%
  ggplot(aes(x = mean_altitude, y = Country.of.Origin)) +
  geom_bar(stat = "identity") +
  ylab("Country of Origin") +
  xlab("Mean Altitude") +
  ggtitle("Mean Altitude of Coffee Beans by Country")
```

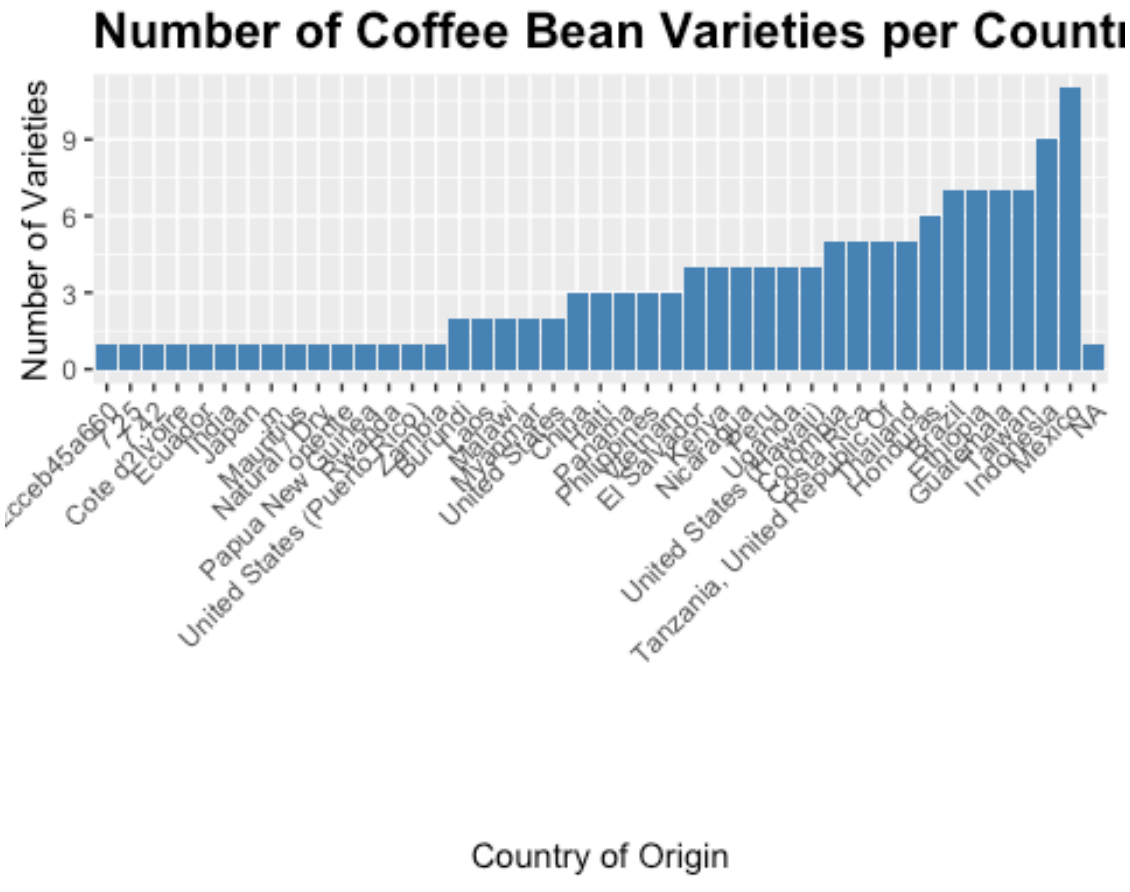


We can see by this graph, that altitude is not a big factor in producing coffee. In the graph before this one, we can determine that Mexico is one of the biggest producers of coffee in this data set, but its average altitude is similar with a lot of other countries, whom do not producer as much as Mexico does.

Varieties of Coffee Beans

Let's now look at the number of different varieties that are produces in each country and

```
coffee %>%
  group_by(Country.of.Origin) %>%
  summarise(num_varieties = n_distinct(Variety)) %>%
  ggplot(aes(x = reorder(Country.of.Origin, num_varieties), y =
num_varieties)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Country of Origin", y = "Number of Varieties",
       title = "Number of Coffee Bean Varieties per Country of
Origin") +
  theme(plot.title = element_text(size = 16, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

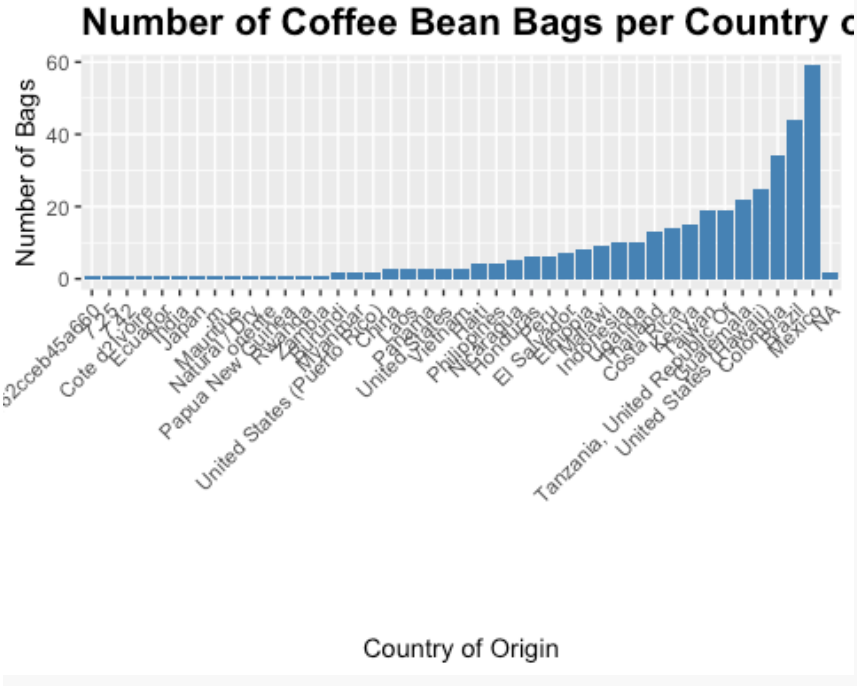


While it's true that Mexico is a significant coffee-producing country, it's important to note that the graph we generated only shows the number of distinct coffee bean varieties for each country of origin. It does not necessarily indicate that Mexico produces the most coffee or has the highest coffee production levels.

To get a better idea of the coffee production levels for each country, we would need to analyze a different variable, such as the total number of bags produced per country or the total export value of coffee per country.

However, it's still interesting to note that Mexico has a large number of distinct coffee bean varieties, which could be attributed to the country's diverse geography and climates, as well as its long history of coffee production.

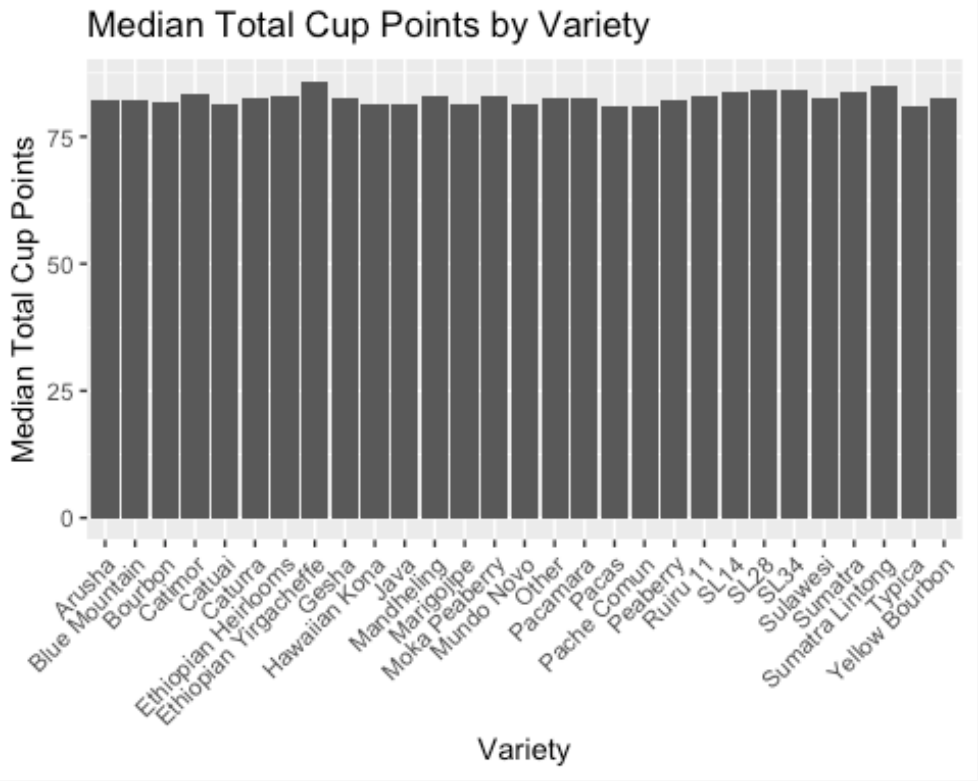
```
coffee %>%
  group_by(Country.of.Origin) %>%
  summarise(num_bags = n_distinct(Number.of.Bags)) %>%
  ggplot(aes(x = reorder(Country.of.Origin, num_bags), y = num_bags))
+
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Country of Origin", y = "Number of Bags",
       title = "Number of Coffee Bean Bags per Country of Origin") +
  theme(plot.title = element_text(size = 16, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1))
```



So it's possible that Mexico has a high number of distinct coffee bean varieties due to its large coffee production levels. However, it's important to note that the number of coffee bean varieties is not directly proportional to coffee production levels, as other factors such as climate, soil, and coffee cultivation techniques can also play a significant role in the number of distinct coffee varieties produced in a particular region or country.

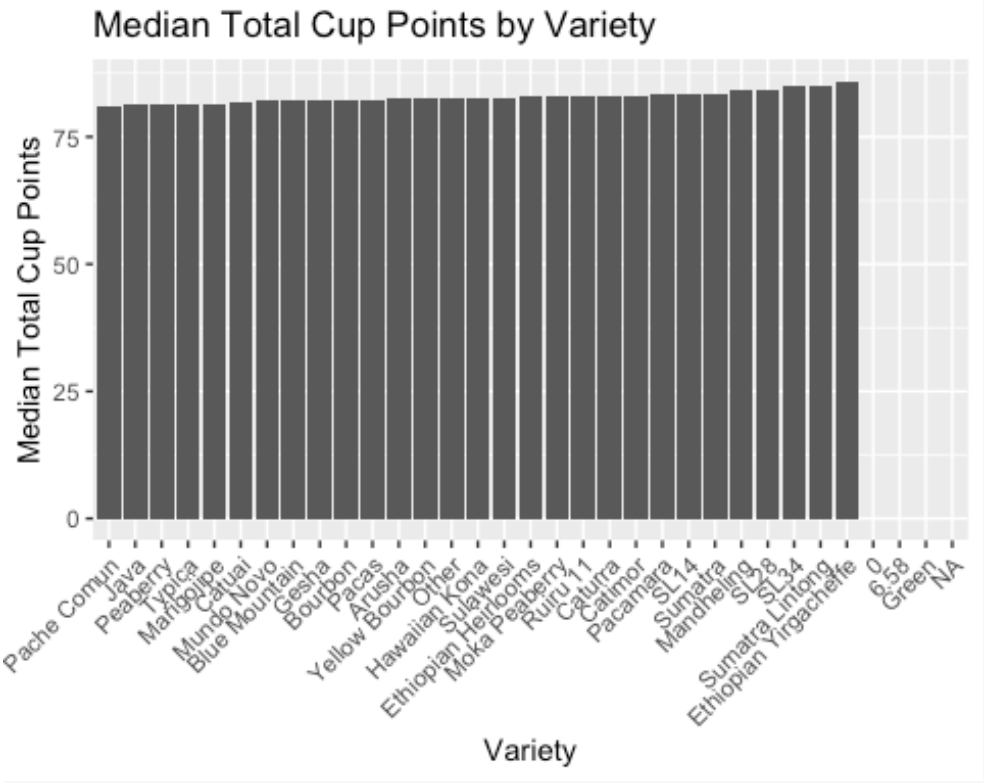
Let's Now look at what varieties have the most amount of `Total.Cup.Points`

```
coffee %>%
  group_by(Variety) %>%
  summarize(avg_cup_points = mean(Total.Cup.Points)) %>%
  filter(avg_cup_points > 10) %>%
  ggplot(aes(x = Variety, y = avg_cup_points)) +
  geom_bar(stat = "identity") +
  labs(title = "Median Total Cup Points by Variety",
       x = "Variety",
       y = "Median Total Cup Points") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)
```



It might be Unclear to see which one is has the most or least average cup points, so let's order it by least to greatest.

```
coffee %>%
  group_by(Variety) %>%
  filter(!is.numeric(Variety)) %>%
  summarize(median_cup_points = median(Total.Cup.Points)) %>%
  ggplot(aes(x = reorder(Variety, median_cup_points), y =
median_cup_points)) +
  geom_bar(stat = "identity") +
  labs(title = "Median Total Cup Points by Variety",
       x = "Variety",
       y = "Median Total Cup Points") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



We can see by this graph that the varieties all have similar average cup points. The `Ethiopian Yirgachette` has the highest average cup points. Some Could say that this is the best coffee bean.

Conclusions

Based on the analysis, it was determined that the variable `Number.of.Bags` is not related to `Total.Cup.Points`. The focus then shifted to studying the impact of the variables `Aroma`, `Flavor`, and `Body` on `Total.Cup.Points`. The correlation between each variable and `Total.Cup.Points` were graphed and confirmed with a correlation matrix. The p-values were significant, indicating a strong relationship. A multiple regression equation was formed to predict `Total.Cup.Points`, with Flavor being the strongest predictor. The number of unique countries and regions was investigated, and it was found that there were over 340 different regions, making grouping by region unfeasible for this study. We have also explored various aspects of coffee beans using the coffee data set. We found that there were 40 different countries represented and that `Washed/Wet` was the most popular processing method used for coffee production. Altitude did not seem to be a significant factor in producing coffee, as countries with similar average altitudes had vastly different coffee production levels. Additionally, we found that there were a variety of coffee bean types represented in the dataset, with similar average cup points. However, `Ethiopian Yirgachette` had the highest average cup points. Overall, this analysis provides insight into the production and characteristics of coffee beans in various regions and countries.