11.3 Final Project Step 3

Sadie Harper

6/5/2021

Introduction

Coffee is one of the most widely consumed beverages out there. Hot, iced, sweet, or any other variation, it is enjoyed in many countries by people of all ages. Not all coffee tastes alike and the opinions of regular coffee drinkers can range quite a bit. Opinions aside, what accounts for quality coffee? Maybe it depends on where it comes from, maybe there are good years and bad years for growing, or maybe there is not all that much of a difference. I am going to take the opinions of trained coffee reviewers and coffee bean growth information to see what, if anything, truly drives the quality of this popular beverage. Whether you drink coffee three times a day, once a year, or sell it to those caffeine enthusiasts, this should interest you. Using data science to answer this question can provide those coffee drinkers and sellers key insights into what increases or decreases quality and how that might affect prices. Data science techniques can show patterns and relationships that may not be immediately known. This can all be done in a timely manner as well which is why this is an ideal way to tackle this question.

The Problem

What contributes to quality coffee? This is the overarching question that fuels this analysis. Diving deeper into this question I want to know if the quality differs between countries? Does it change over the years? Does the price change with the quality? Are production heavy years related to the quality? There are a lot of factors that might drive the quality, price, and production of this caffeinated beverage and I want to question all of them to address overall coffee quality.

Addressing the Problem

Data used

I used three datasets to address Arabica coffee quality in this analysis. Arabica coffee is one of the most commonly used coffee beans in the world. There are many different varieties and mutations of this bean and the majority of my data was focused on this type of coffee. The first dataset I have includes coffee prices in US dollars per pound per year from 1998 to 2018. This included prices for three different types of arabica coffee and robusta coffee as well. Coffee prices came from ICO Indicator Prices collected from Statista. The second dataset I used also came from Statista and has information on the production of arabica coffee from 2005 to 2020 in units of 1,000 60 kg bags. The information is from the USDA Foreign Agricultural Service. Lastly, the third dataset I used came from a GitHub Repository that scraped data from the Coffee Quality Institute review pages. This is a large dataset that contains a number of different quality metrics of different coffee varieties from all over the world.

Approach

I have began addressing the question of coffee quality by performing different statistical exercises on the coffee quality data. To start, I cleaned and transformed the data. There were a couple typos and date transformations that needed to be fixed. Then I started the exploratory data analysis, consisting of visualizations, summary statistics and other statistical tests. R has many different functions and packages to make this step easy. I began identifying important variables that would be useful for building models. The rest of this analysis will show what I have found in the data in a presentable format so that it can be easily interpreted for anyone interested in learning about what might contribute to coffee quality.

Methods

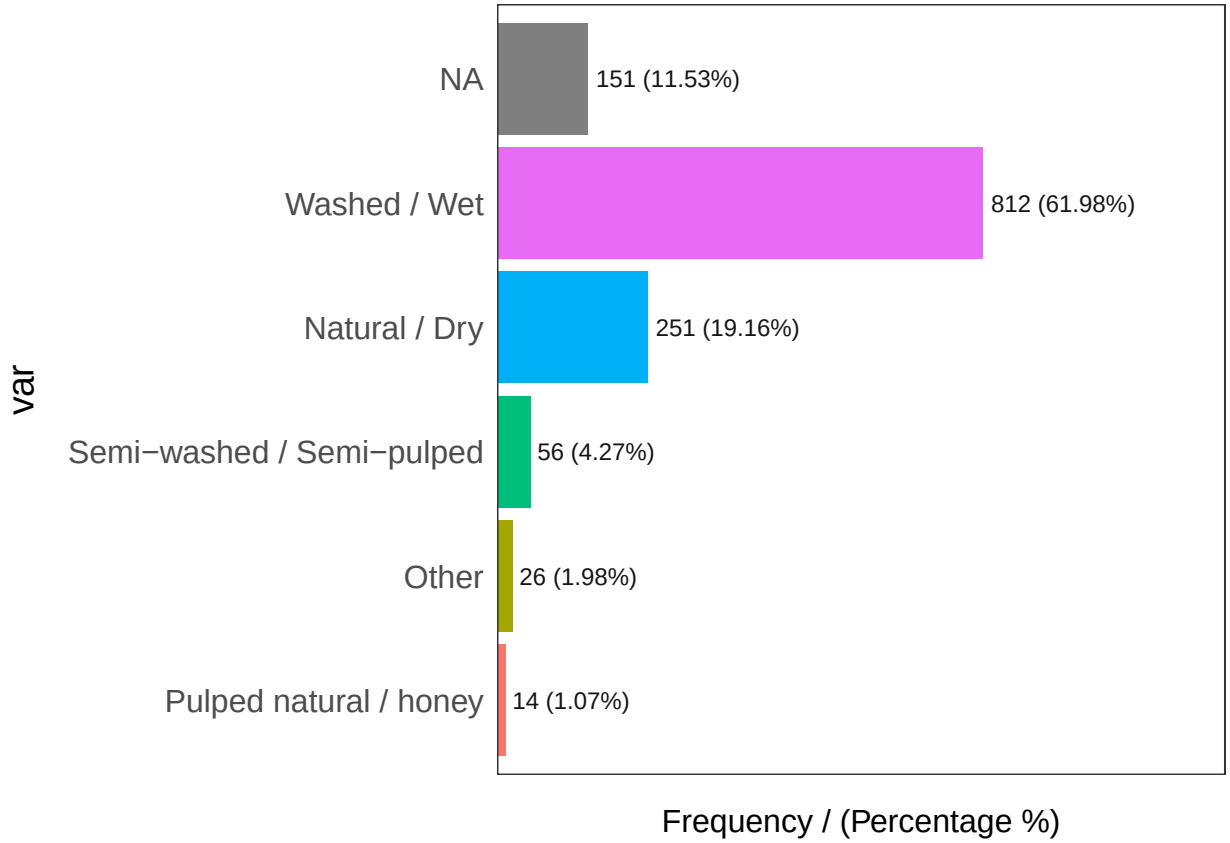
I am using different methods to learn about my data. Once I have evaluated the different metrics and variations I can decide what model will best fit the data. I have many methods shown below that check for a number of things. The frequency of values to check for variation, histograms to understand the distribution, line charts and scatter plots to look for trends and relationships, and bar graphs that show differences between categorical variables in the dataset.

Analysis

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Frequency

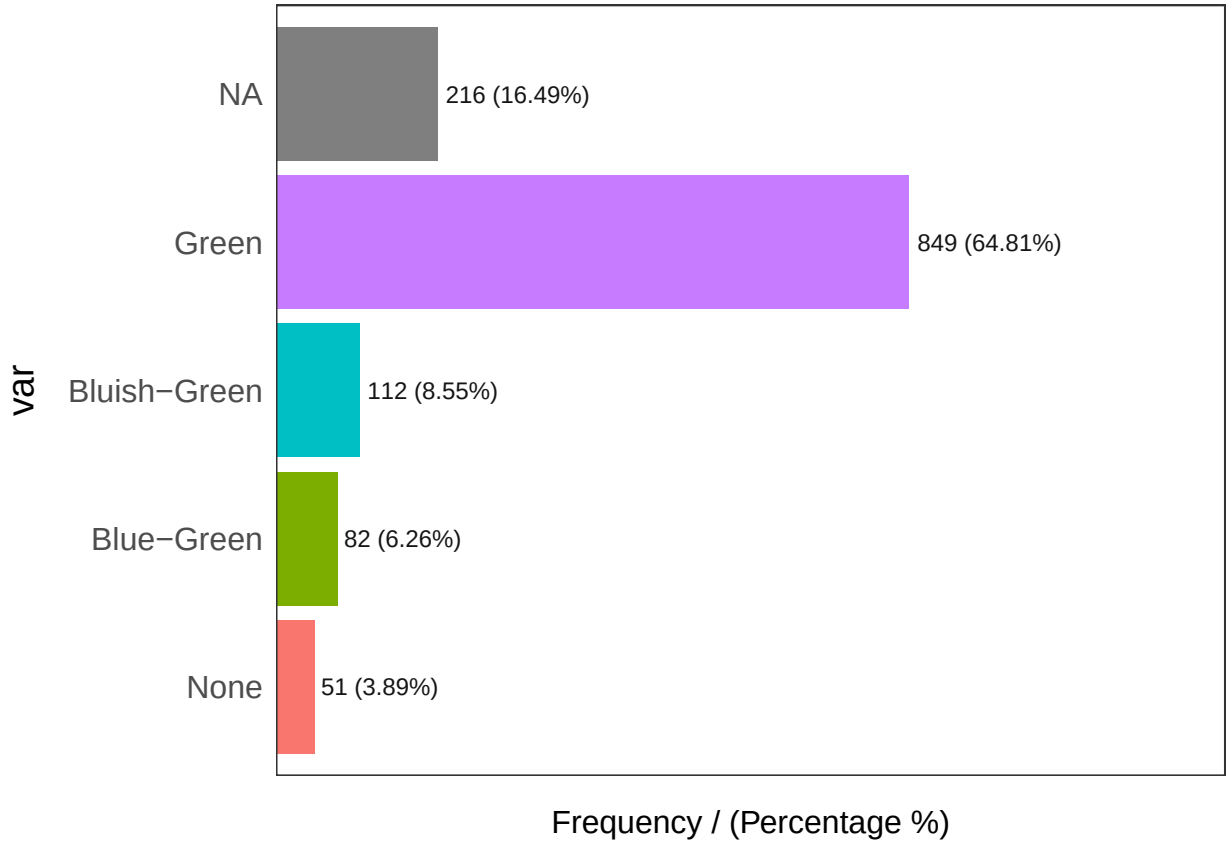
Processing Method




```
##                         var frequency percentage cumulative_perc
## 1               Washed / Wet       812      61.98           61.98
## 2              Natural / Dry       251      19.16           81.14
## 3                      <NA>       151      11.53           92.67
## 4 Semi-washed / Semi-pulped        56       4.27           96.94
## 5                      Other        26       1.98           98.92
## 6    Pulped natural / honey        14       1.07          100.00
```

Most reviews are washed/wet which is important to keep in mind when creating the final model.

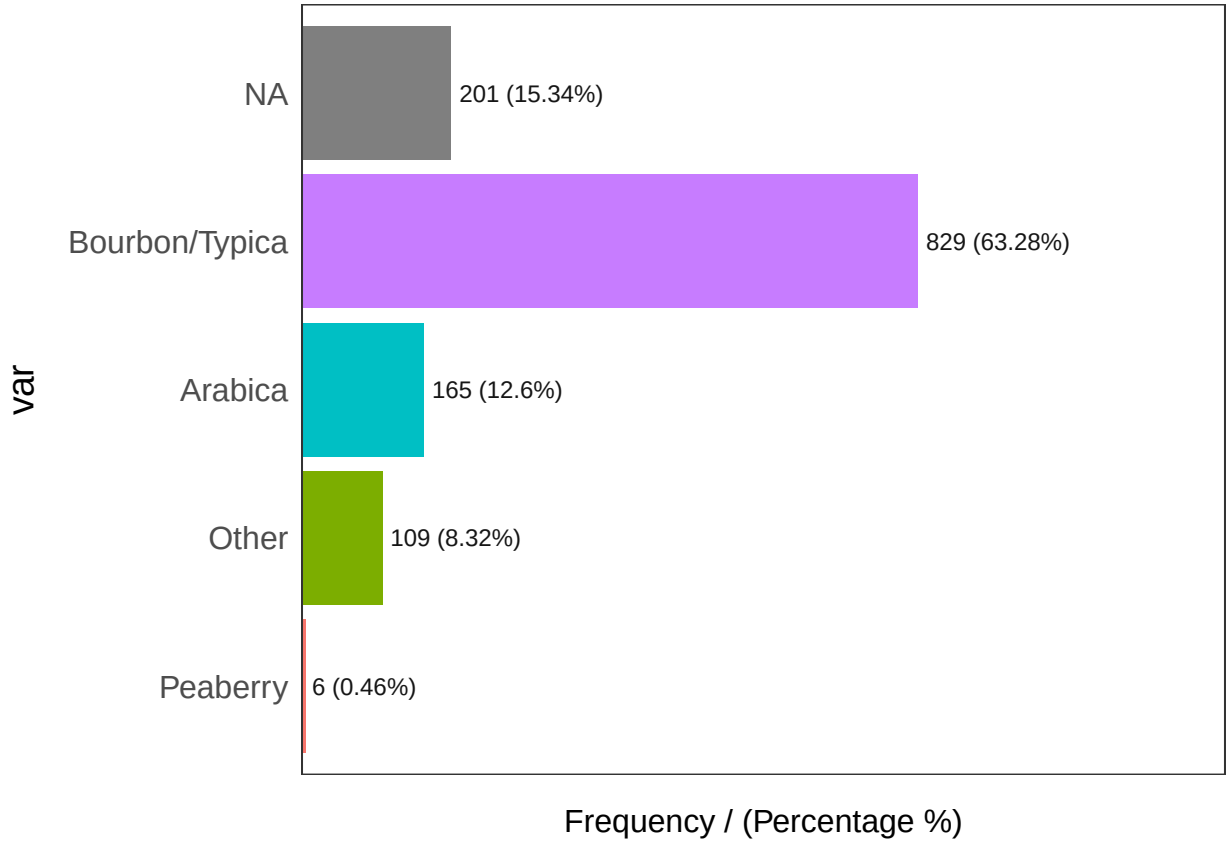
Color




```
##            var frequency percentage cumulative_perc
## 1        Green       849      64.81           64.81
## 2         <NA>       216      16.49           81.30
## 3 Bluish-Green       112       8.55           89.85
## 4   Blue-Green        82       6.26           96.11
## 5         None        51       3.89          100.00
```

Most reviews are green by a large margin, also important to be aware of for creating the model.

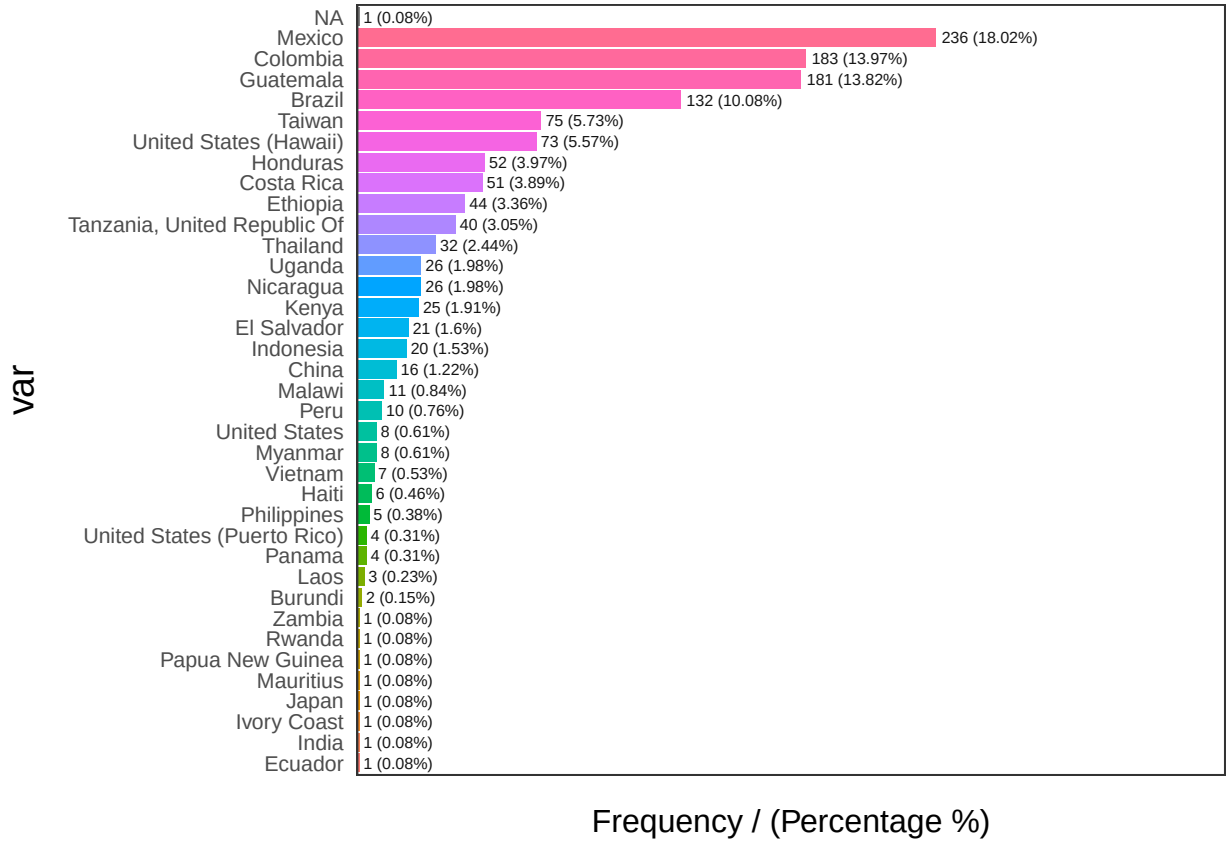
Type




```
##              var frequency percentage cumulative_perc
## 1 Bourbon/Typica       829      63.28           63.28
## 2           <NA>       201      15.34           78.62
## 3        Arabica       165      12.60           91.22
## 4          Other       109       8.32           99.54
## 5       Peaberry         6       0.46          100.00
```

The Bourbon/Typica type has the most reviews. These are arabica coffees still but worthy of being in their own category. Peaberry coffee refers to when there is only one bean in the plant, it technically does not specify the type of coffee. There are very few reviews of this type so it won't affect the analysis of other coffees and could potentially be excluded.

Country of Origin



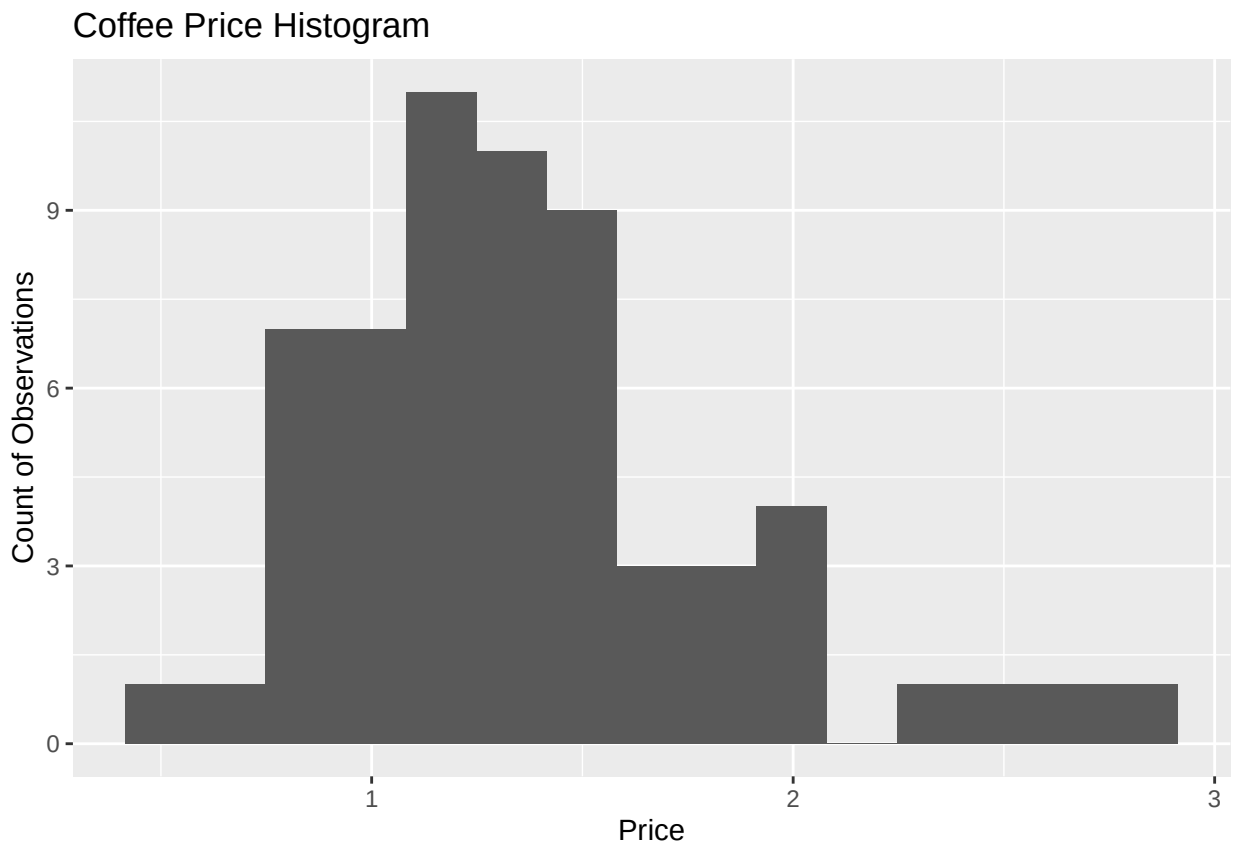

```
##                          var frequency percentage cumulative_perc
## 1                     Mexico       236      18.02           18.02
## 2                   Colombia       183      13.97           31.99
## 3                  Guatemala       181      13.82           45.81
## 4                     Brazil       132      10.08           55.89
## 5                     Taiwan        75       5.73           61.62
## 6      United States (Hawaii)        73       5.57           67.19
## 7                   Honduras        52       3.97           71.16
## 8                 Costa Rica        51       3.89           75.05
## 9                   Ethiopia        44       3.36           78.41
## 10 Tanzania, United Republic Of       40       3.05           81.46
## 11                  Thailand        32       2.44           83.90
## 12                 Nicaragua        26       1.98           85.88
## 13                    Uganda        26       1.98           87.86
## 14                     Kenya        25       1.91           89.77
## 15               El Salvador        21       1.60           91.37
## 16                 Indonesia        20       1.53           92.90
## 17                     China        16       1.22           94.12
## 18                    Malawi        11       0.84           94.96
## 19                      Peru        10       0.76           95.72
## 20                   Myanmar         8       0.61           96.33
## 21             United States         8       0.61           96.94
## 22                   Vietnam         7       0.53           97.47
## 23                     Haiti         6       0.46           97.93
```

```
## 24                  Philippines         5       0.38          98.31
## 25                       Panama         4       0.31          98.62
## 26  United States (Puerto Rico)         4       0.31          98.93
## 27                         Laos         3       0.23          99.16
## 28                      Burundi         2       0.15          99.31
## 29                      Ecuador         1       0.08          99.39
## 30                        India         1       0.08          99.47
## 31                  Ivory Coast         1       0.08          99.55
## 32                        Japan         1       0.08          99.63
## 33                    Mauritius         1       0.08          99.71
## 34            Papua New Guinea          1       0.08          99.79
## 35                       Rwanda         1       0.08          99.87
## 36                       Zambia         1       0.08          99.95
## 37                         <NA>         1       0.08         100.00
```

Most of the reviews are for coffee from Mexico, Colombia, and Guatemala.

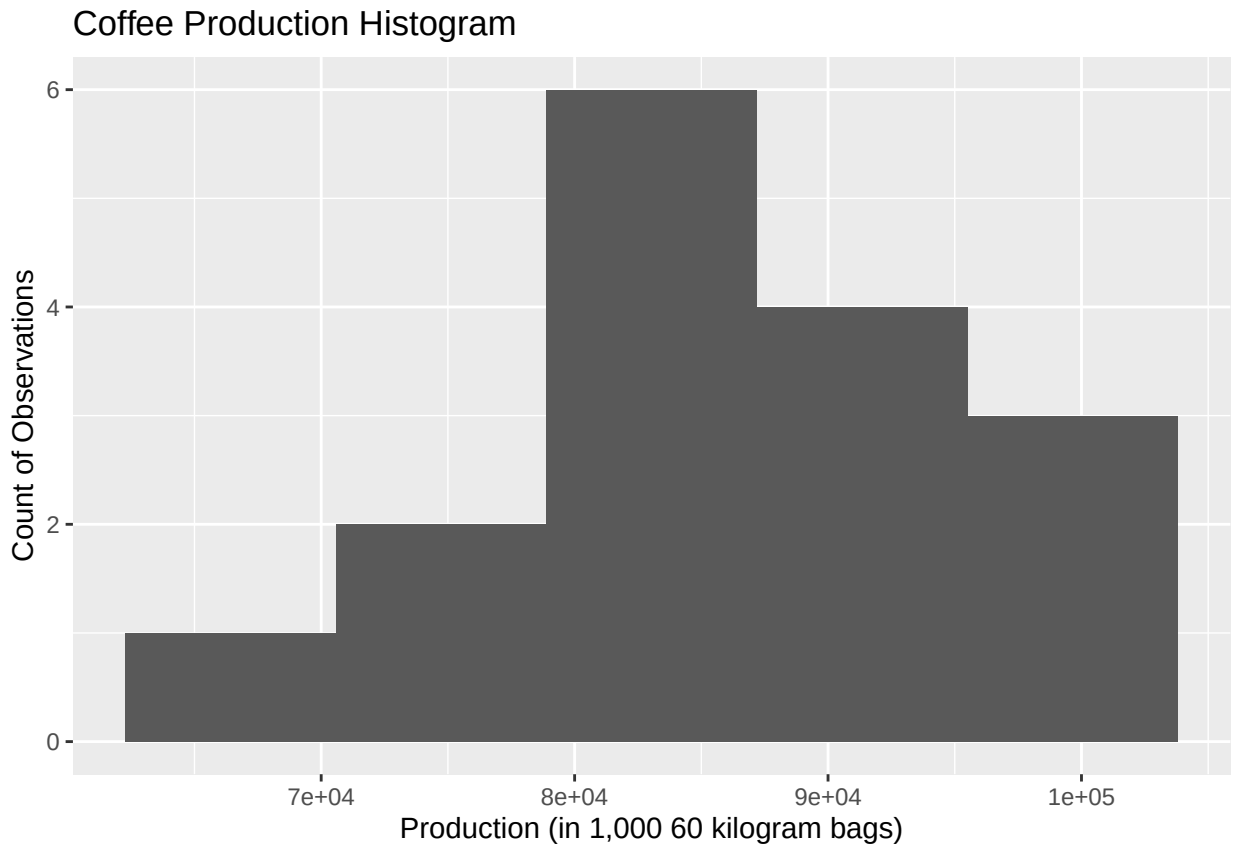
Histograms

Coffee Price



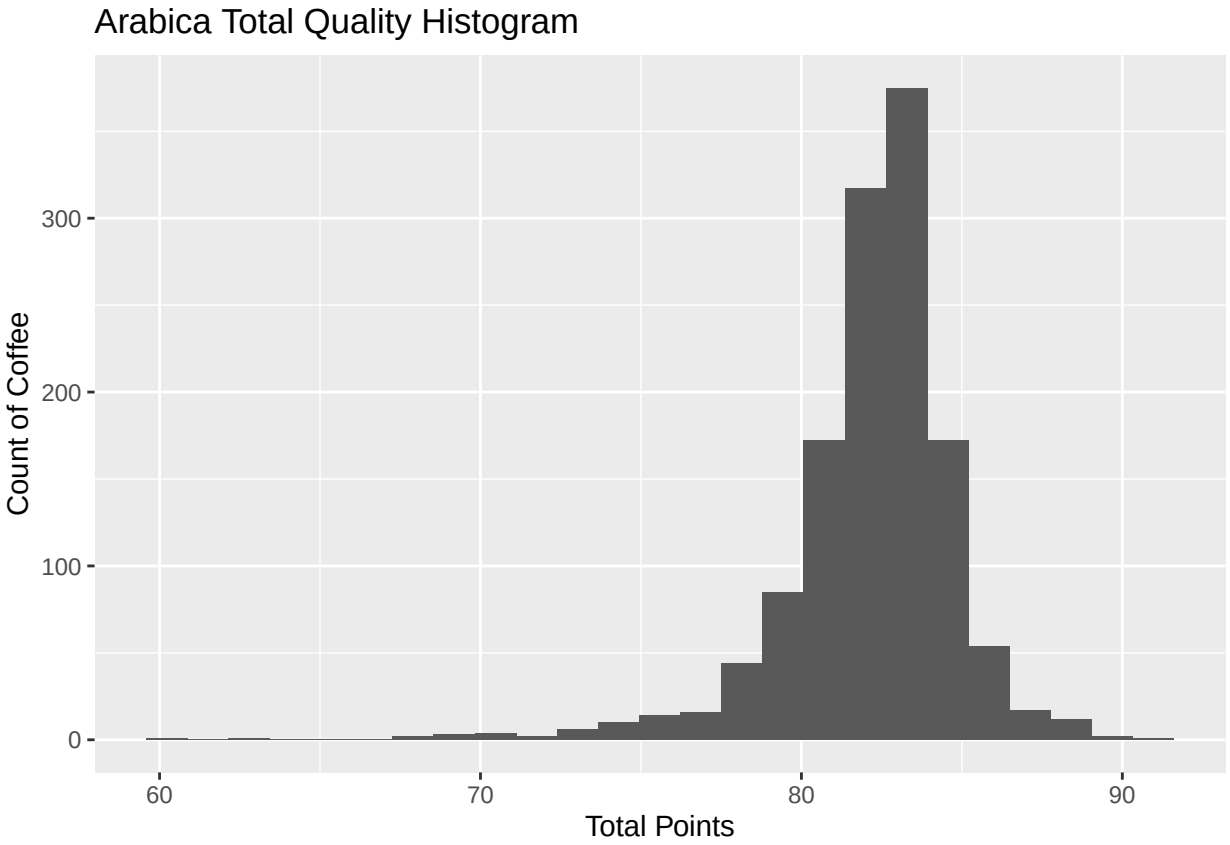
Coffee price has a somewhat normal distribution for how little data there is available.

Coffee Production



The production distribution is slightly negatively skewed but mostly normal. As with the price data, there is not a large amount of data in this distribution.

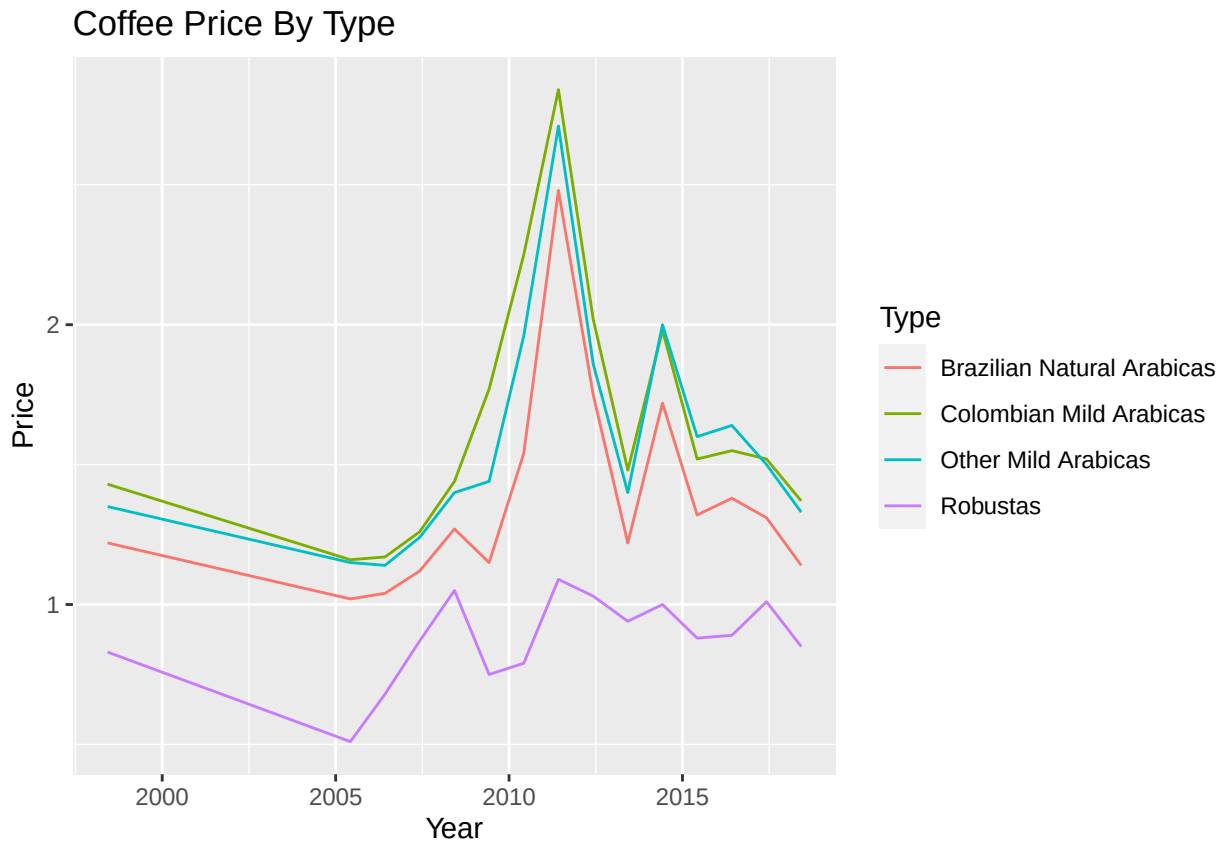
Coffee Quality



The distribution is a bit negatively skewed with only a few outlier observations having total scores in the 60's-70's. Other than that, it looks fairly normal with the most scores being between 80 - 85 total points.

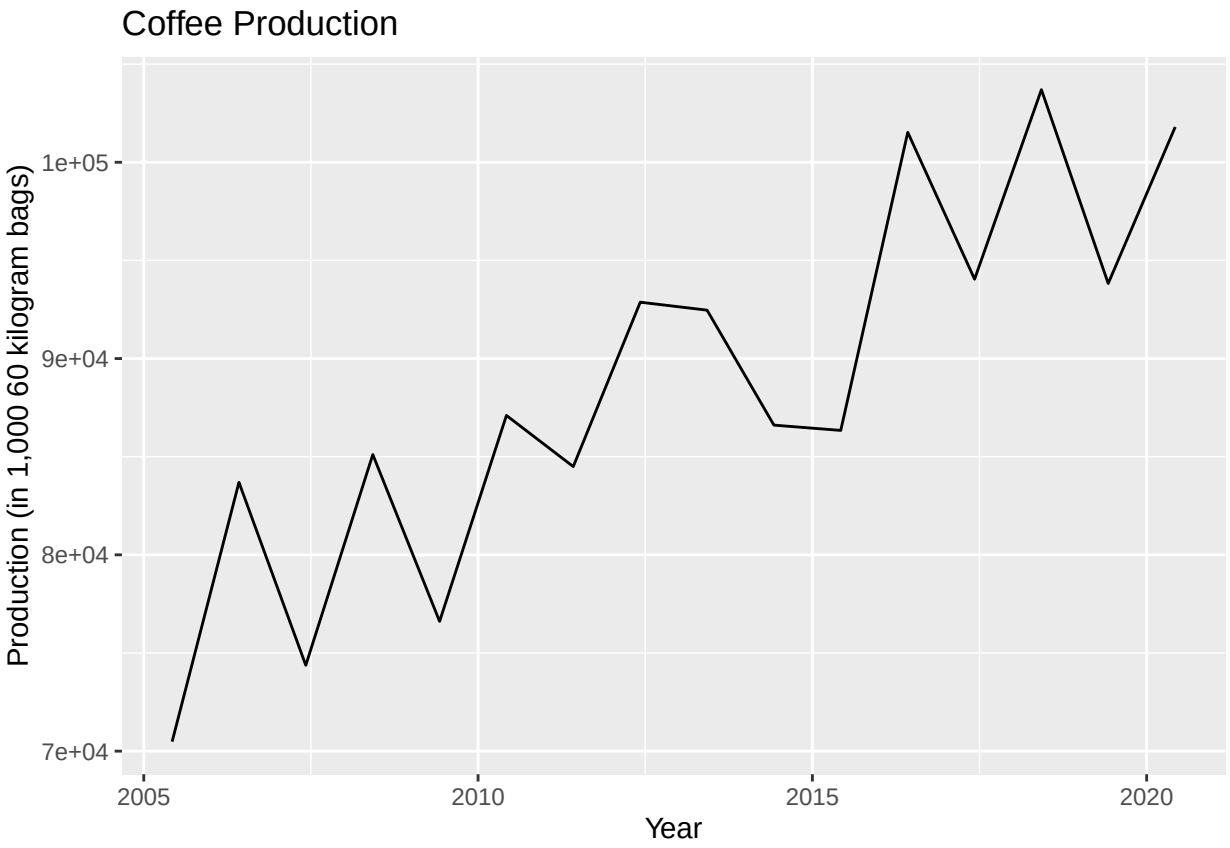
Line Charts

Coffee Price



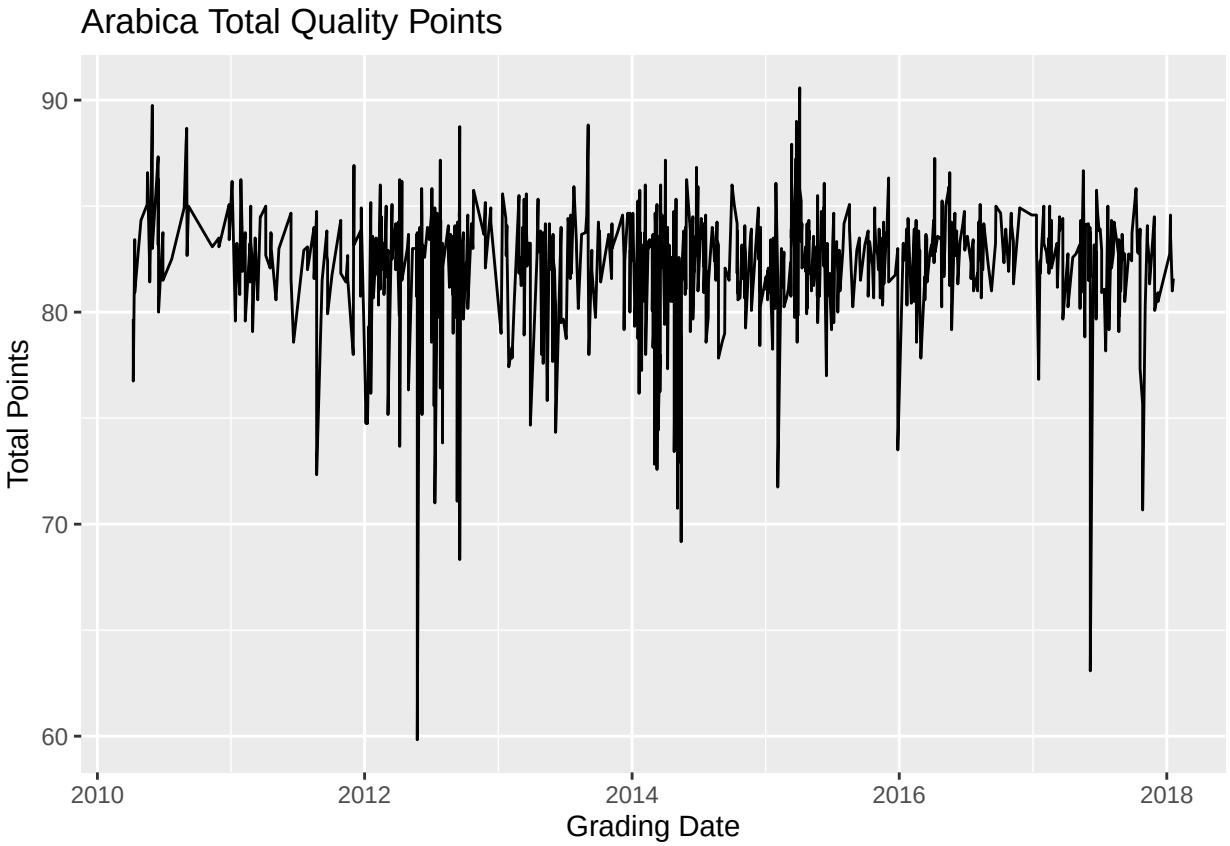
Colombian mild arabicas seem to almost always sell for higher prices, it sometimes gets overlapped by other mild arabicas. The Brazilian natural arabicas are consistently the lowest priced arabica coffee. Robustas aren't discussed in this analysis but I believe it's worth seeing that they are a much more affordable coffee bean than Arabica beans. They are also known for being of less quality taste which is why there are no robusta coffees in the review data.

Coffee Production



The trend of production is increasing but it's not happening at a steady or constant rate. The ups and downs of production numbers are quite visible in the graph.

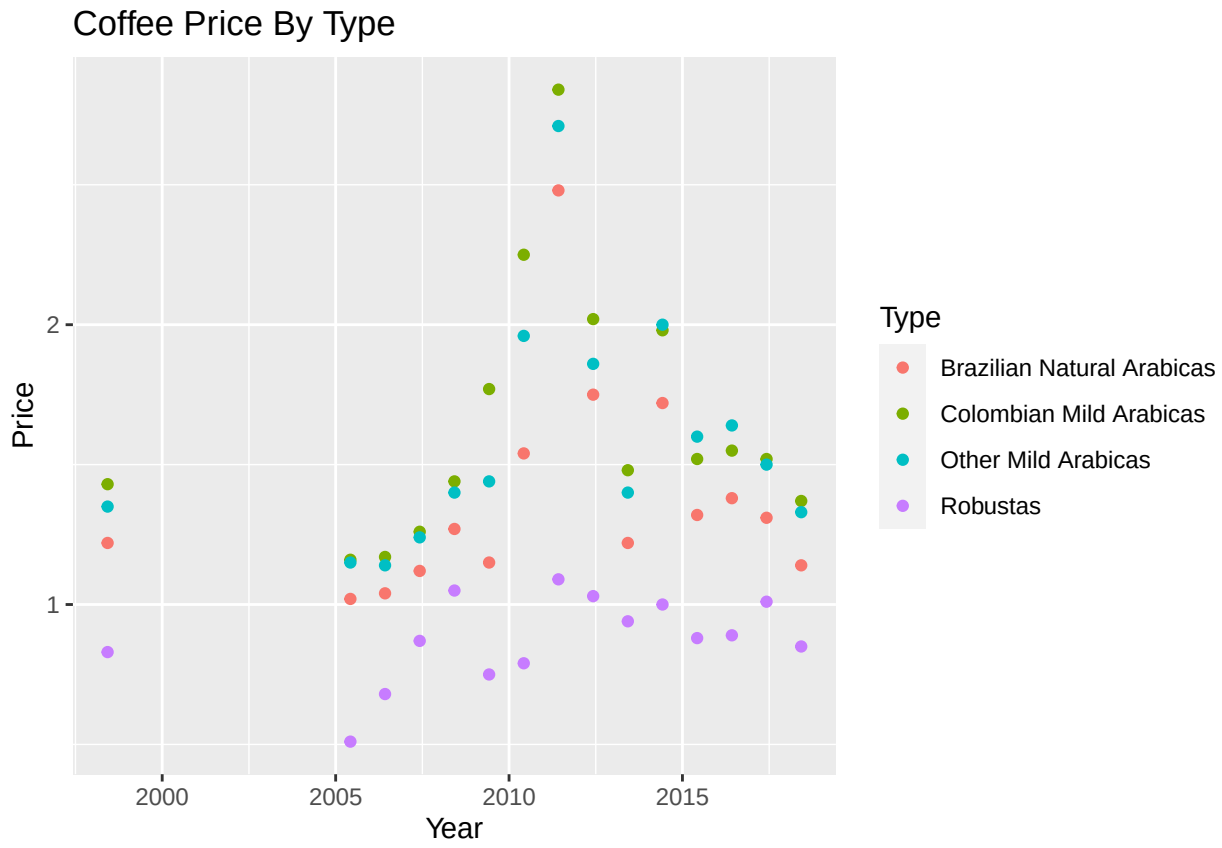
Coffee Quality



Looking at the total points by the grading date confirms that there is no general trend with the scoring. Therefore, I would assume that the method for grading coffee is consistent throughout the years. It also indicates there are quality and less quality coffees throughout the last few years. The two outliers with low points are seen in this graph as well.

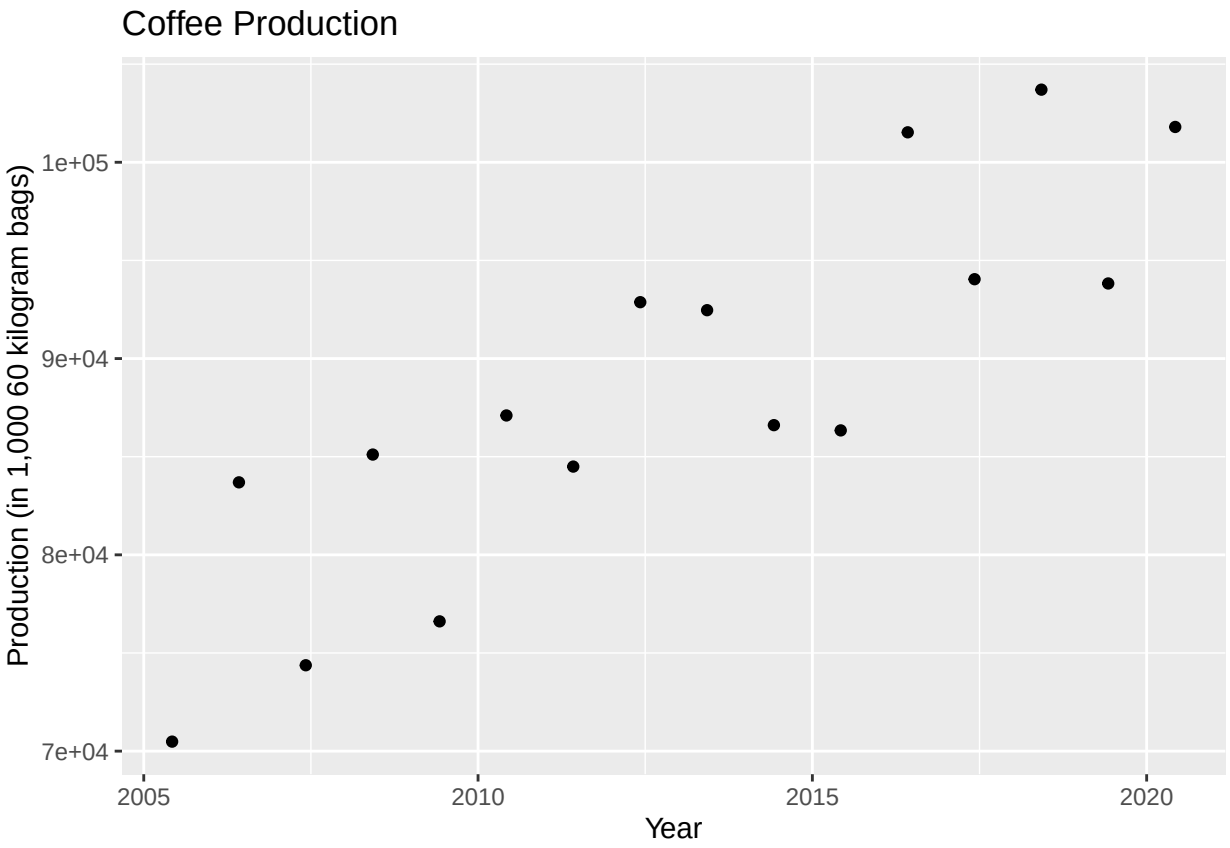
Scatter Plots

Coffee Price



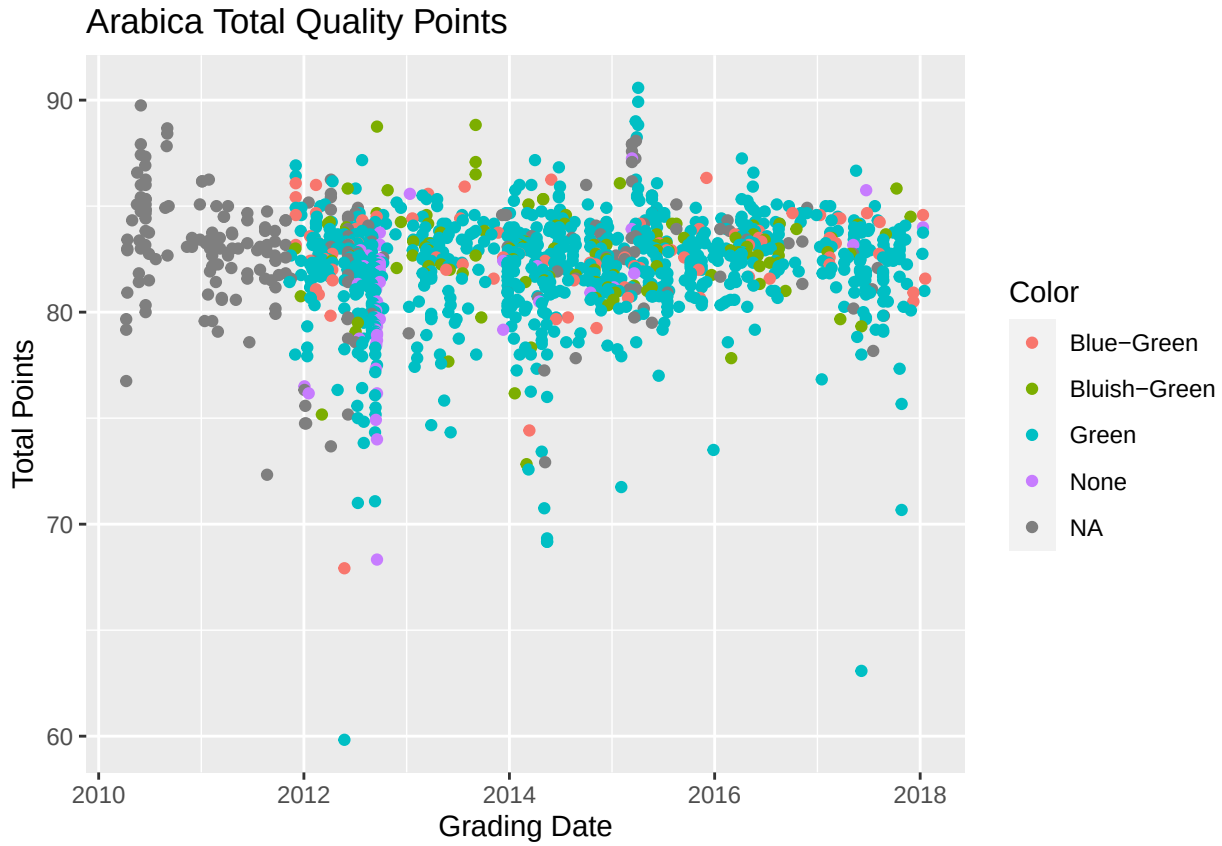
There doesn't seem to be linear with the price for any type of coffee. There was a jump around 2012 and then prices dropped back down.

Coffee Production



Production appears to be linear but there are very few data points.

Coffee Quality

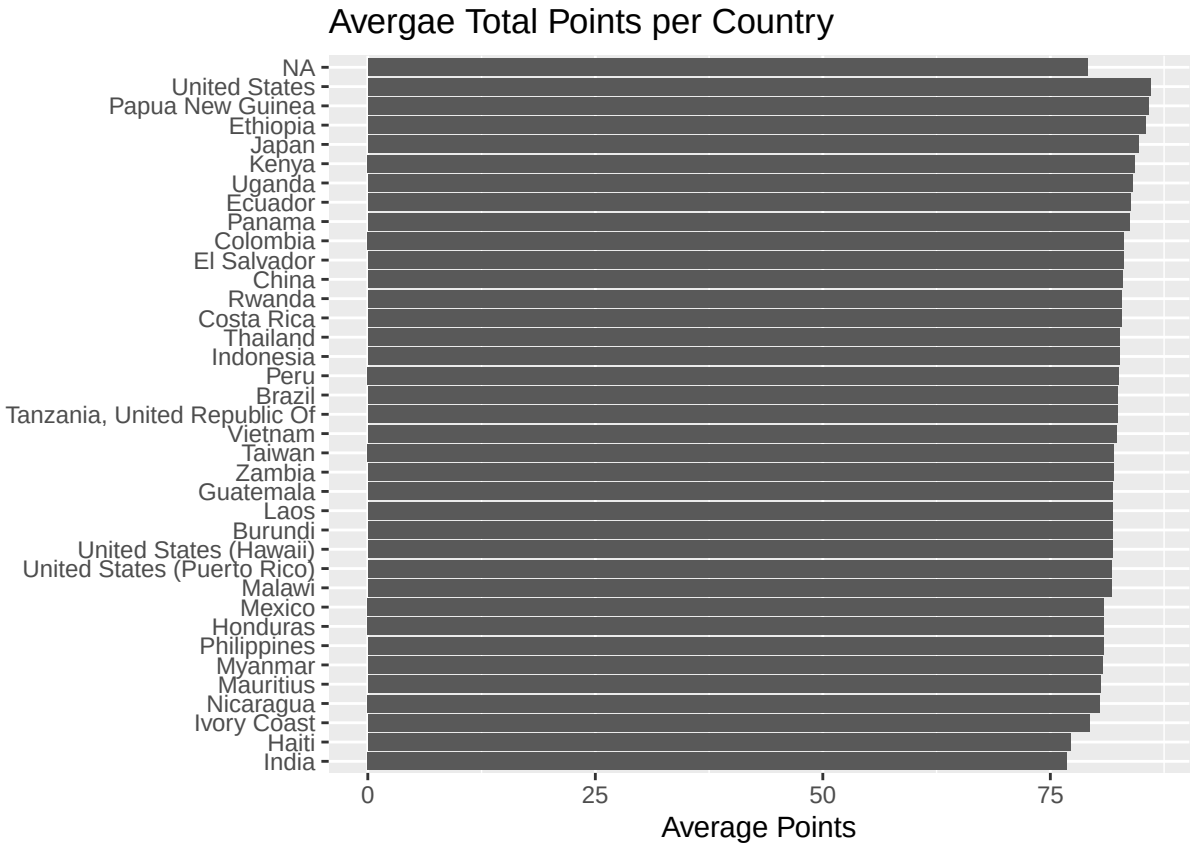


The points for each color is pretty spread out across the board. It does though appear that most of the NAs are observed in the earliest years so the more recent reviews have better data in this category. I would suggest cutting the date to only go back to 2012 for building the model based off the color since that's where the data really starts for this variable. It also might be worth combining blue-green and blueish-green into one group.

Bar Charts

Coffee Quality

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



Since some countries have far more observations than others I wanted to see how they stacked up based on the average points. Mexico has the most observations in the data by far yet they fall towards the bottom half of the group with their average total points. The United States not including Hawaii has the highest average.

Insights

As a result of the previous analysis, it appears that coffee production and prices are generally increasing with time, production more so than prices. The coffee quality total points almost look like white noise on their line chart. There are no years that seem to have generally higher or lower quality ratings. It seems that quality is potentially not improving with time, it's more constant. Therefore I would suggest focusing on other variables such as location and altitude.

Reccomendation

I recommend running a multiple regression on this data to predict total coffee quality. The distributions are very close to normal so removing outliers or transforming the data would prepare it for regression analysis. It does not initially look like processing method or color contribute significantly to quality, mostly because such a large portion of the data fall into one group from each of those categories. The average total points did vary between countries though so I would suggest including different location metrics in the analysis. I also think that the different variants of arabica coffee could play a role in total quality.

After analyzing the total points I would suggest diving into specific metrics as well. There are a number of measures of quality that go into the total points so it would be interesting to see if certain measures affect the total points more so than others. This could potentially be done with linear regression.

Implications

My biggest takeaway from this analysis is that location might affect coffee quality. I would need to perform more concrete statistical tests to either confirm or deny this but after looking at many different descriptive variables this one seems to be the most promising. I would like to take this project to the next phase and build out a multiple regression analysis on coffee quality. To begin, the data would need to be transformed and all the assumptions would need to be double checked to be sure I use the appropriate regression model. Then I would go through a variable selection process, but I think I have a good start already. Once I built out the model I would check the residuals and test for goodness of fit, significance, and accuracy. I would ideally be left with a model that could be used to build out a prediction for coffee quality based on certain known factors like the variant, location, and harvest time.

Limitations

The data I have on coffee is not robust enough to fully understand how coffee quality changes over time or by location on a global scale. My data consists of more recent years of overall production and I am only analyzing reviews from a handful of coffee variants. Because of this I will only partially address coffee quality. I also think this analysis could be improved in two areas. The first being more data. I would like to have more details on the harvest time of year, and more details on the location. Second, I have a general understanding of analyzing data but someone with more experience could know more ways to look at the data. There could be key insights that I'm missing and if there was one change made or one different style of a visual added, those would become more obvious. Additionally, building out a model would greatly improve the analysis.

Conclusion

The previous analysis on coffee quality has broadened my knowledge of different variations of coffee across the globe. Not all coffee is alike and I now understand just how many differences there are. There are naturally occurring beans, hybrid beans, and a number of different environments that can grow coffee beans differently. For others reading this analysis, I hope you too feel as though you've gained some understanding of where coffee comes from, how coffee varies, and what might cause certain coffee beans to produce better tasting coffee than others.

References

- Conway, J. (2020, November 26). Average price of coffee worldwide from 1998 to 2018, by type of coffee. Retrieved May 16, 2021, from https://www-statista-com.ezproxy.bellevue.edu/statistics/250186/average-price-of-coffee-worldwide-by-coffee-type/.
- Conway, J. (2020, November 26). World Arabica Coffee Production from 2005/06 to 2020/21. Retrieved May 16, 2021, from https://www-statista-com.ezproxy.bellevue.edu/statistics/225400/world-arabica-coffee-production/.
- LeDoux, J. (2018, June 16). Coffee-quality-database. Retrieved May 16, 2021, from https://github.com/jldbc/coffee-quality-database