Coffee Quality Data

Aurora (Yutian) Qiu Alex Marshall Dataset: https://github.com/jldbc/coffee-quality-database Original data source: https://database.coffeeinstitute.org

Introduction

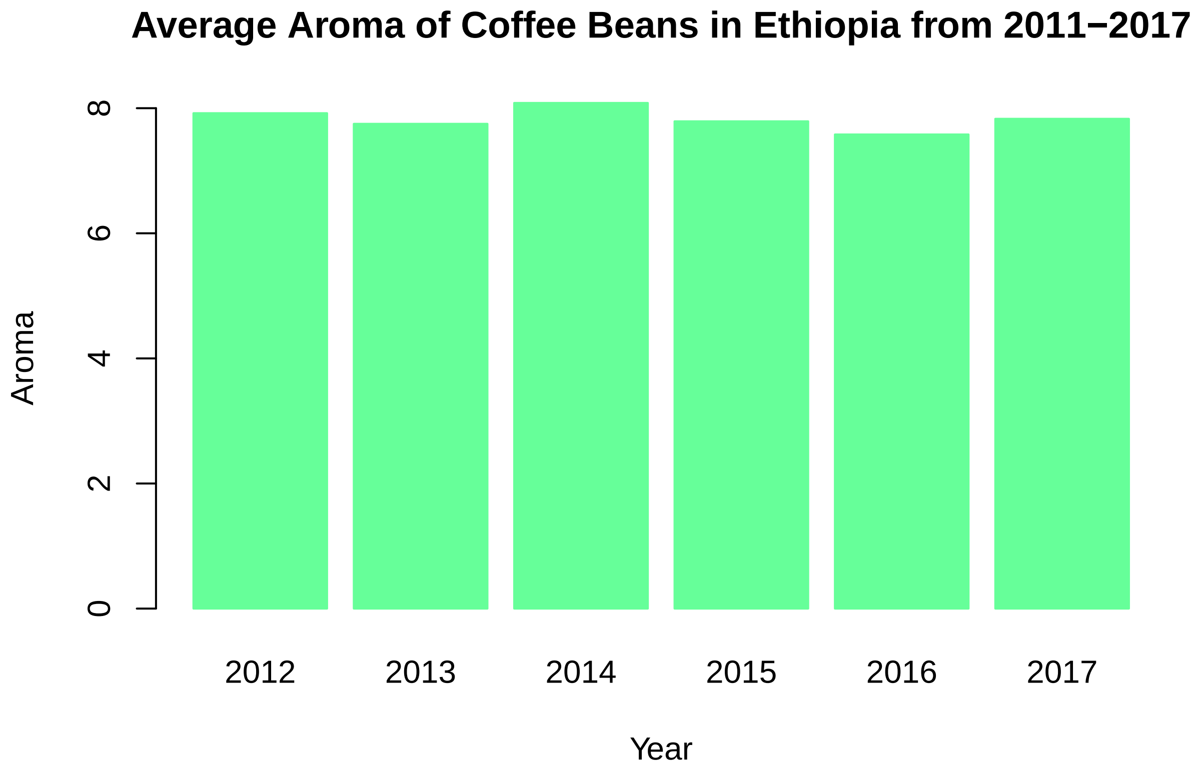
We explore a dataset scraped from Coffee Institute's Coffee Quality Database. Our dataset consists of reviewed features of different coffees as well as bean and farm metadata for 1312 sources of Arabica beans from around the world. We explore bean attributes and growing conditions across different countries, and whether the bean attributes can be used to infer the beans' provenance.

Dataset

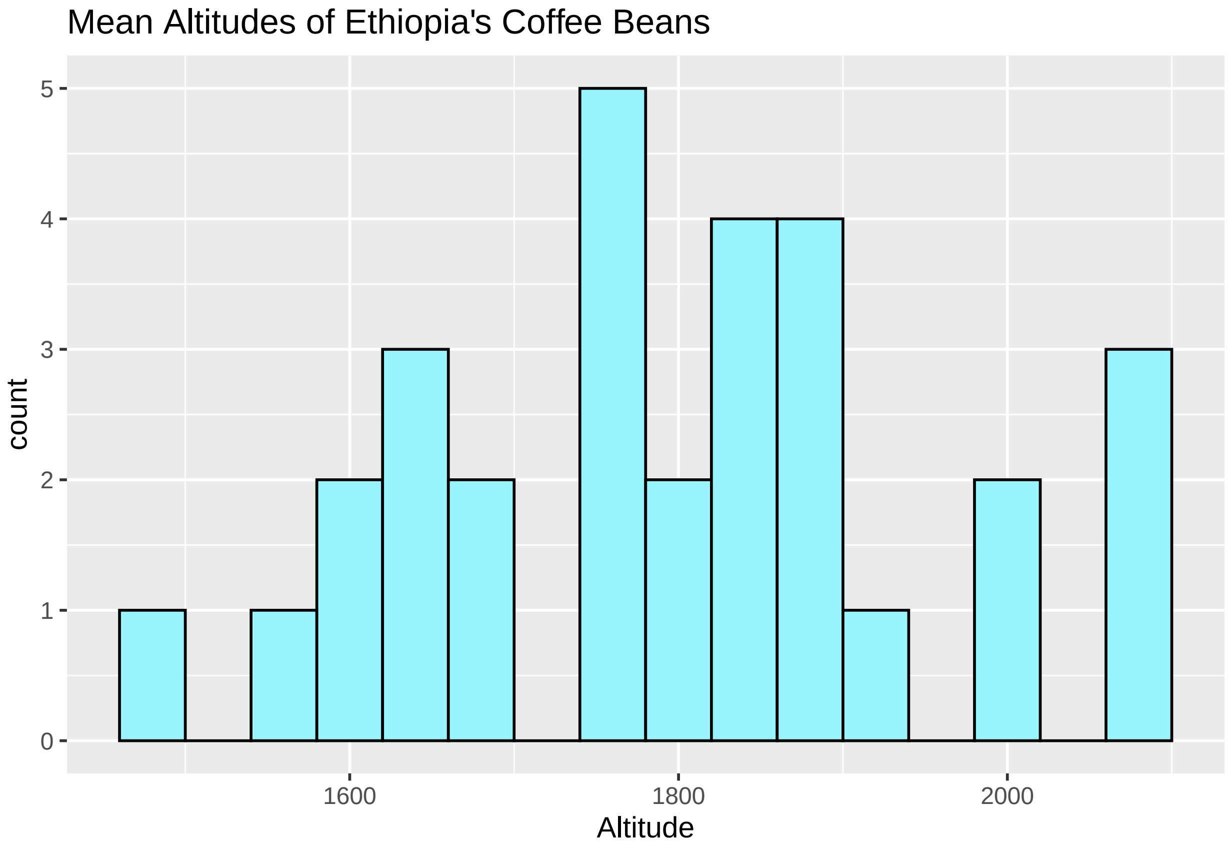
```
## 'data.frame':    1084 obs. of  10 variables:
##  $ Country   : Factor w/ 37 levels "","Brazil","Burundi",..: 10 10 11 10 10 10 10 10 10 33 ...
##  $ Year      : Factor w/ 47 levels "","08/09 crop",..: 16 16 1 16 16 41 41 16 16 16 ...
##  $ Method    : Factor w/ 6 levels "","Natural / Dry",..: 6 6 1 2 6 1 1 2 2 6 ...
##  $ Aroma     : num  8.67 8.75 8.42 8.17 8.25 8.25 8.67 8.08 8.17 8.25 ...
##  $ Flavor    : num  8.83 8.67 8.5 8.58 8.5 8.33 8.67 8.58 8.67 8.42 ...
##  $ Aftertaste: num  8.67 8.5 8.42 8.42 8.25 8.5 8.58 8.5 8.25 8.17 ...
##  $ Acidity   : num  8.75 8.58 8.42 8.42 8.5 8.42 8.42 8.5 8.5 8.33 ...
##  $ Body      : num  8.5 8.42 8.33 8.5 8.42 8.33 8.33 7.67 7.75 8.08 ...
##  $ Balance   : num  8.42 8.42 8.42 8.25 8.33 8.5 8.42 8.42 8.17 8.17 ...
##  $ Altitude  : num  2075 2075 1700 2000 2075 ...
```

Ethiopian Beans: Aroma, Altitude and Processing Method

We start by exploring some basics of the dataset. The average aroma score of coffee beans in Ethoipia from the last five years has been very stable at around 8. If year has a small or no impact on coffee qualities and has little noise, we can group years together when looking at other attributes.



Next we look at a histogram showing the distribution of mean altitudes of Ethiopian coffee beans. There's only around 30 datapoints here, but it appears to be somewhat skewed left. Most are grown at around 1750 to 1950 meters.



Ethiopia processes beans different than producers in other countries. Ethiopian beans are less commonly processed with a wash/wet method than other beans in the dataset.

```
##                Washed
## Eth             Non Washed/Wet Method Washed/Wet Method
##   Non-Ethiopia                    324               730
##   Ethiopia                         22                 8
```

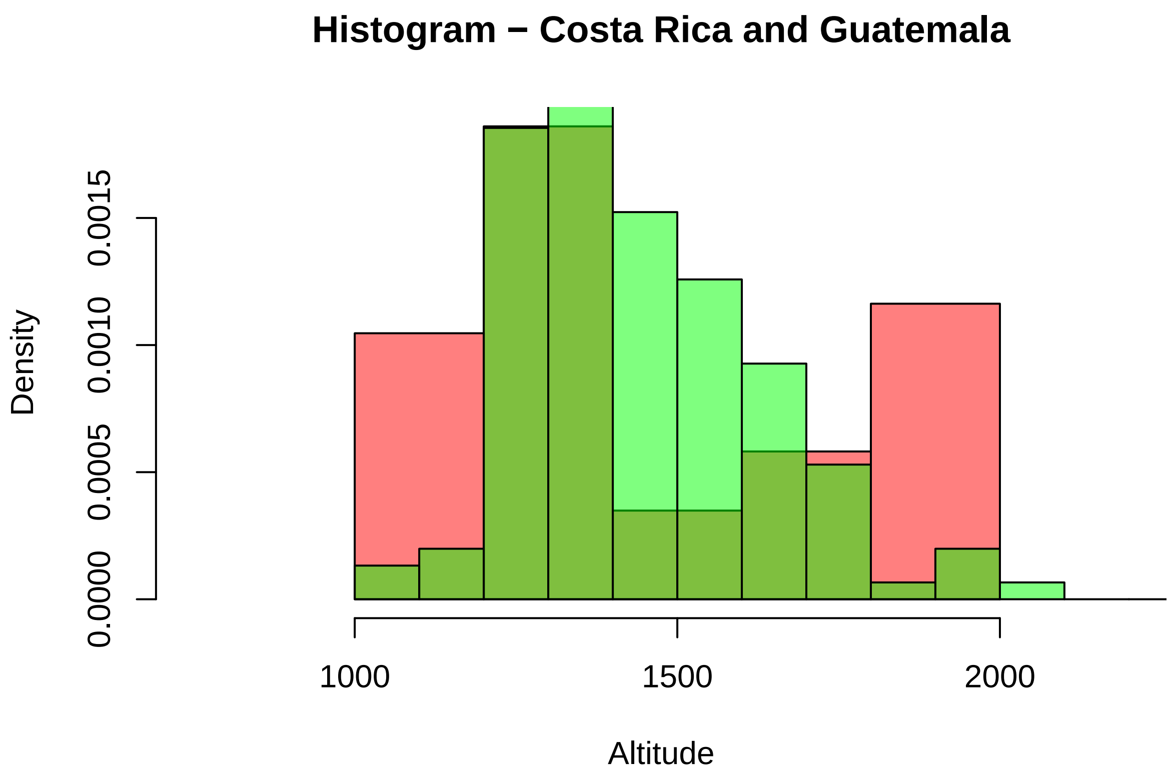
```
#confirm with Chi Sq test for independence
chisq.test(observed)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  observed
## X-squared = 22.432, df = 1, p-value = 2.177e-06
```

A χ^2 test confirms what we can see from the table. The very low p-value leads us to reject the hypothesis that being from Ethiopia is independent of the processing method.

Comparing Countries: do beans from Costa Rica and Guatemala tend to be grown at the same altitude?

These two countries are geographically close. Is there a significant difference between the means of the mean altitude of each country's coffee crop?



The distributions certainly seem different but the observed means are similar. (This is after removing outliers - Guatemala has some extremely high altitude plantations).

```
var(GC$Altitude[GC$Country=='Costa Rica'])
```

```
## [1] 72875.67
```

```
var(GC$Altitude[GC$Country=="Guatemala"])
```

```
## [1] 104300
```

```
var(GC$Altitude)
```

```
## [1] 96925.99
```

```
#Calculate the observed altitude difference by country
GAvg <- sum(GC$Altitude*(GC$Country == "Guatemala"))/sum(GC$Country == "Guatemala"); GAvg
```

```
## [1] 1466.219
```

```
CAvg <- sum(GC$Altitude*(GC$Country == "Costa Rica"))/sum(GC$Country == "Costa Rica"); CAvg
```

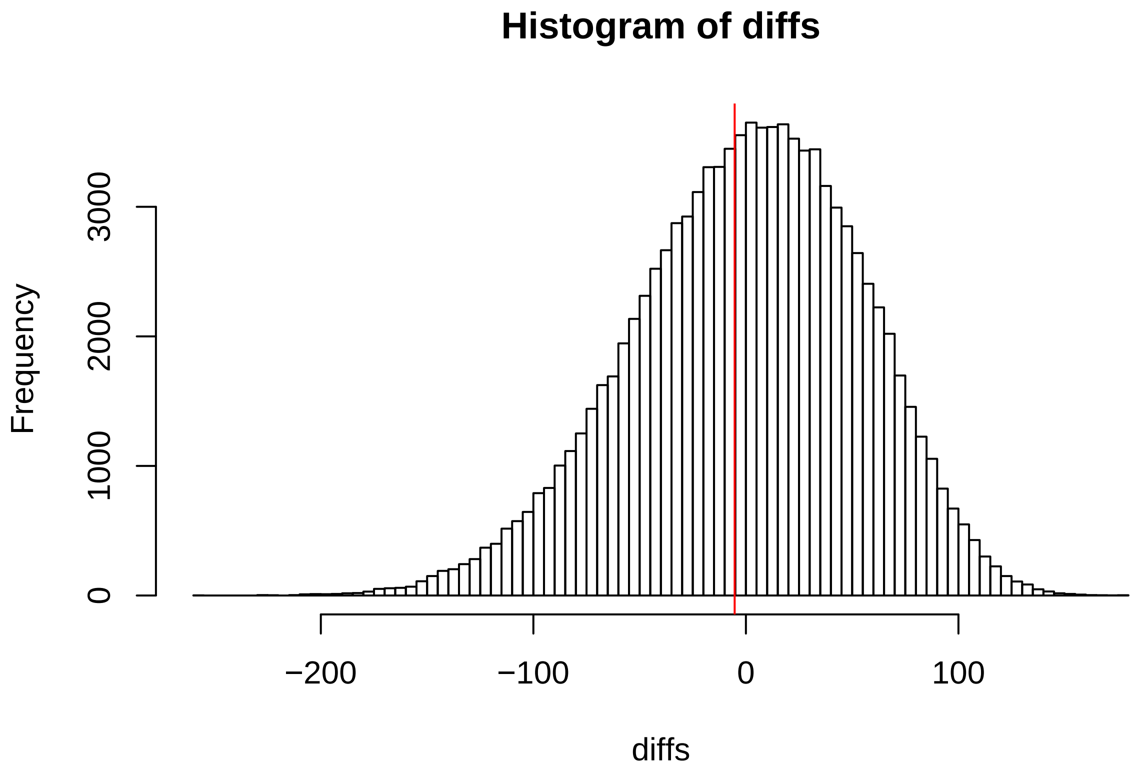
```
## [1] 1471.516
```

```
observed <- GAvg - CAvg; observed     #With outliers removed, Costa Rican beans grow at a higher mean a
```

```
## [1] -5.296851
```

Permutation test

We now confirm how similar the observed mean altitudes are with a permutation test. We sample our data 10^4 times and take the difference in the means to get a sampling distribution of the mean differences.



It certainly wouldn't be useful to distinguish groups of beans from these two countries based on their mean altitudes. If the beans in Guatemala and Costa Rica have their countries reassigned, only 4.7% of those assignments would have differences within −5.30 and 5.30 meters, showing that the mean altitudes are very similar.

```
cilow <- (sum(diffs <= observed)+1)/(N+1);
cihigh <-(sum(diffs <= observed*-1)+1)/(N+1)
cihigh-cilow
```

```
## [1] 0.07662923
```

Student T Test

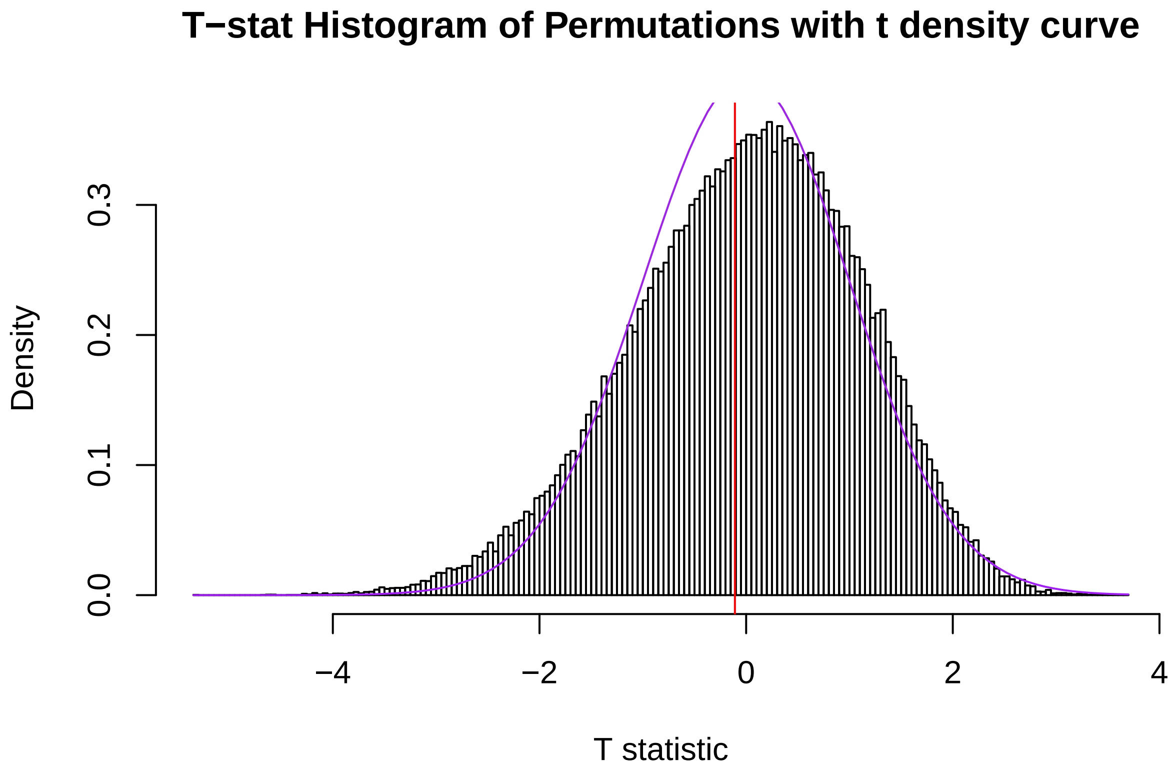
We expected to see a distribution of mean differences from permutation which closely resemble the t-distribution. However our distribution from the permutation test is skewed left.

```
skewness(diffs) # -0.3 skewed left
```

```
## [1] -0.2952549
```

A negative skew of this magnitude suggests that a t test may not be appropriate but we perform one for comparison.

```
## [1] -0.1084494
```



Our t-statistic is -0.108 which corresponds to a p-value of of 0.457

```
pvalue.t<-pt(tstat,df=192);pvalue.t #.46
```

```
## [1] 0.4568762
```

Despite the permutation distribution being different from the t distribution, the p values and result of the tests are similar. This is because the observed means are close, so we land near the middle of both the t and permutation distributions. The differences would show up more with a more extreme observed value.

The permutation test is more appropriate here. The underlying data is fairly small, irregular, unbalanced and they have different variances which may cause slow or failed convergence to the t distribution. One advantage of the t test is that we can find the distribution of t statistics created from underlying distributions with different variances. However this advantage does not outweigh its limitations here.

Confidence Interval for Guatemalan Mean Altitude

Next we find a range for the mean of the mean altitudes of coffee grown in Guatemala.

```
## [1] 1466.219
```

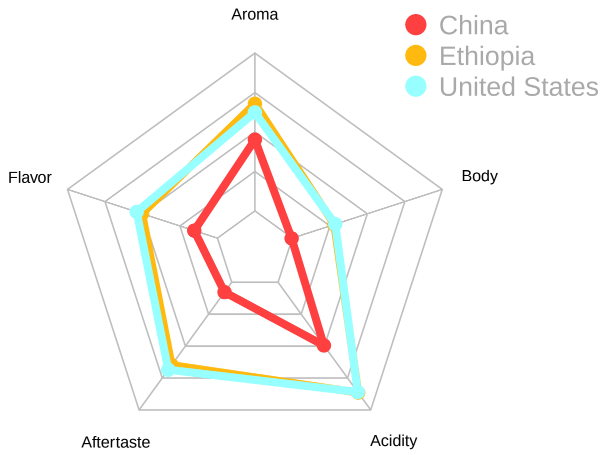
```
## [1] 1414.289
```

```
## [1] 1518.15
```

This shows that our population mean estimate of the mean altitudes for coffee in Guatemala lies within a 95% confidence interval between 1414.29 and 1518.15 meters. Just a short hike!

Visualize the Performance of Coffees from Different Countries

We compared three countries: China, Ethiopia, and the United States as they represent distinct geographical regions in the world of coffee been producers. We charted the average score of the ratings in five interesting categories of characteristics.

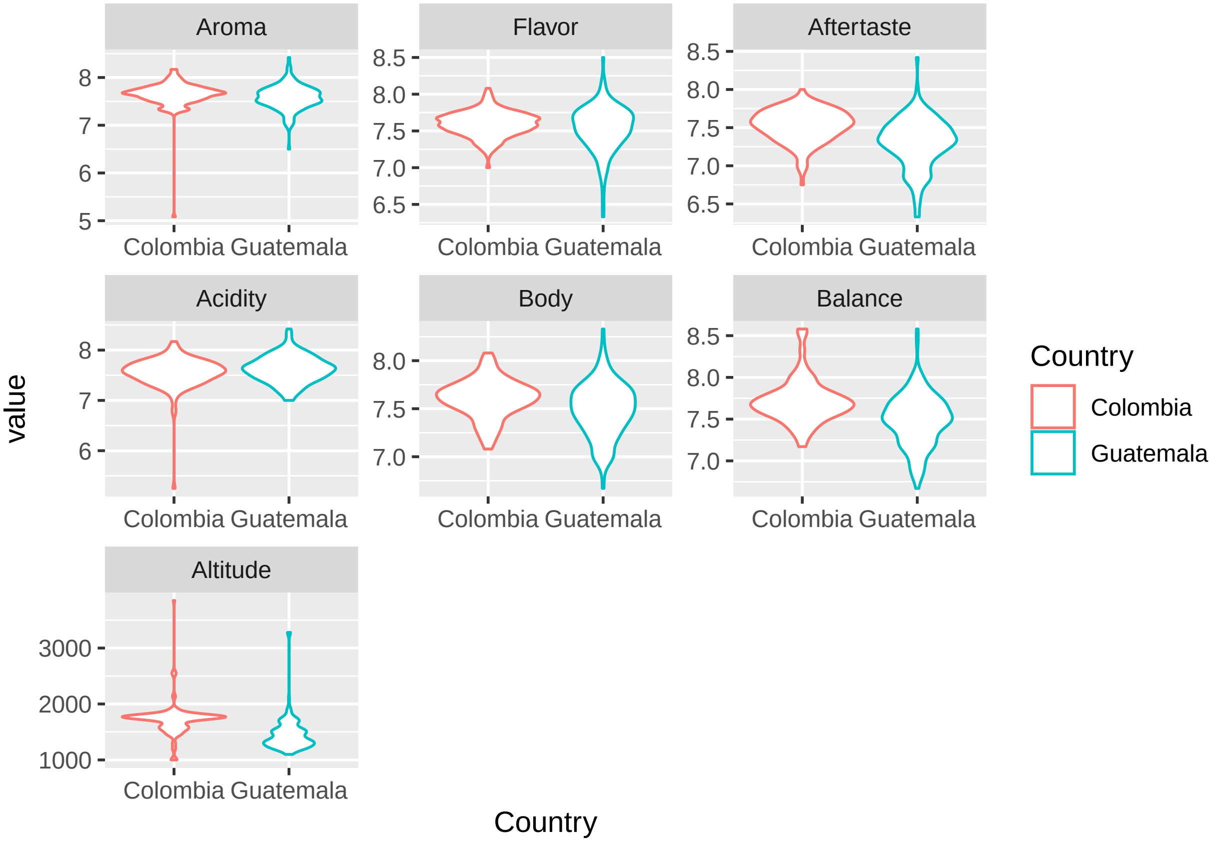


Ethiopia and United States coffee beans have similar performance, while China has relatively lower scores in all five categories; especially flavor, body and aftertaste.

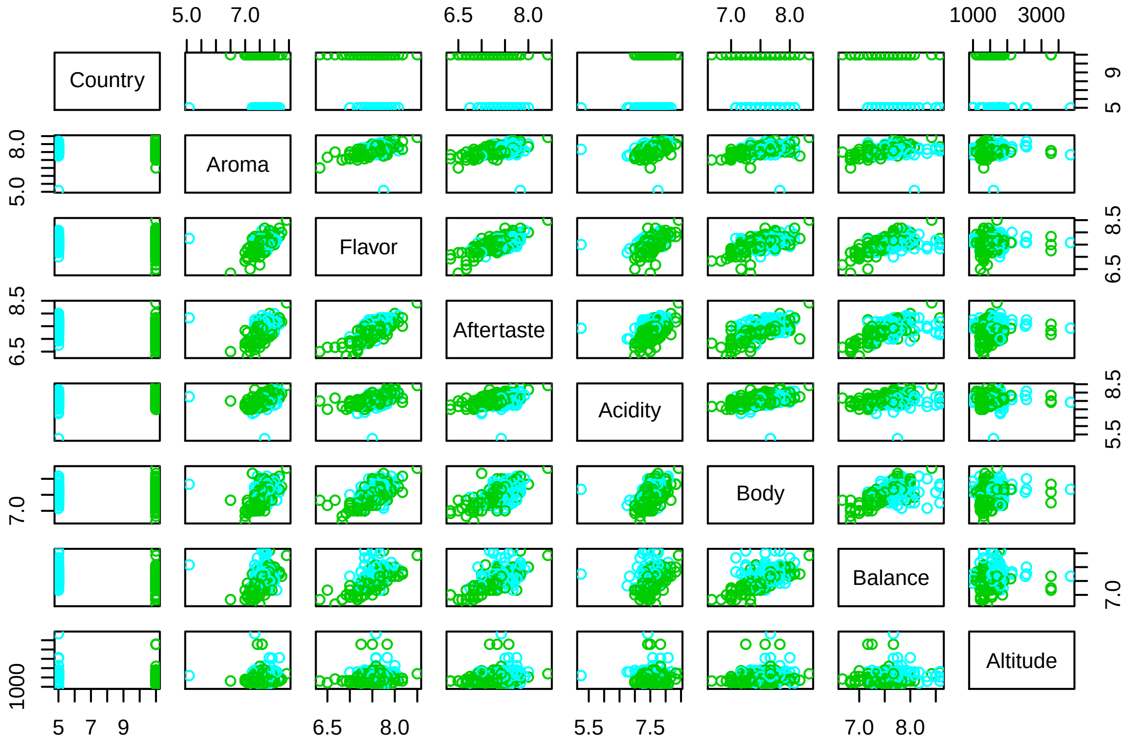
Predicting Altitude and Country from Rating Scores

From the previous analysis, it's clear that coffees from different countries have different characteristics. We explored how well we could predict the altitude and the country of origin from the ratings data. We compared Colombia and Guatemala, the countries with the largest number of beans in the dataset.

```
## Using Country as id variables
```



The average of their ratings is similar but Guatemala tends to have more variance. It seems like aftertaste may be the best predictor of country as Colombia tends to score higher there.Aftertaste, body and balance produce the most separation with country in the scatterplots but none of the attributes seem to correlate well with altitude.

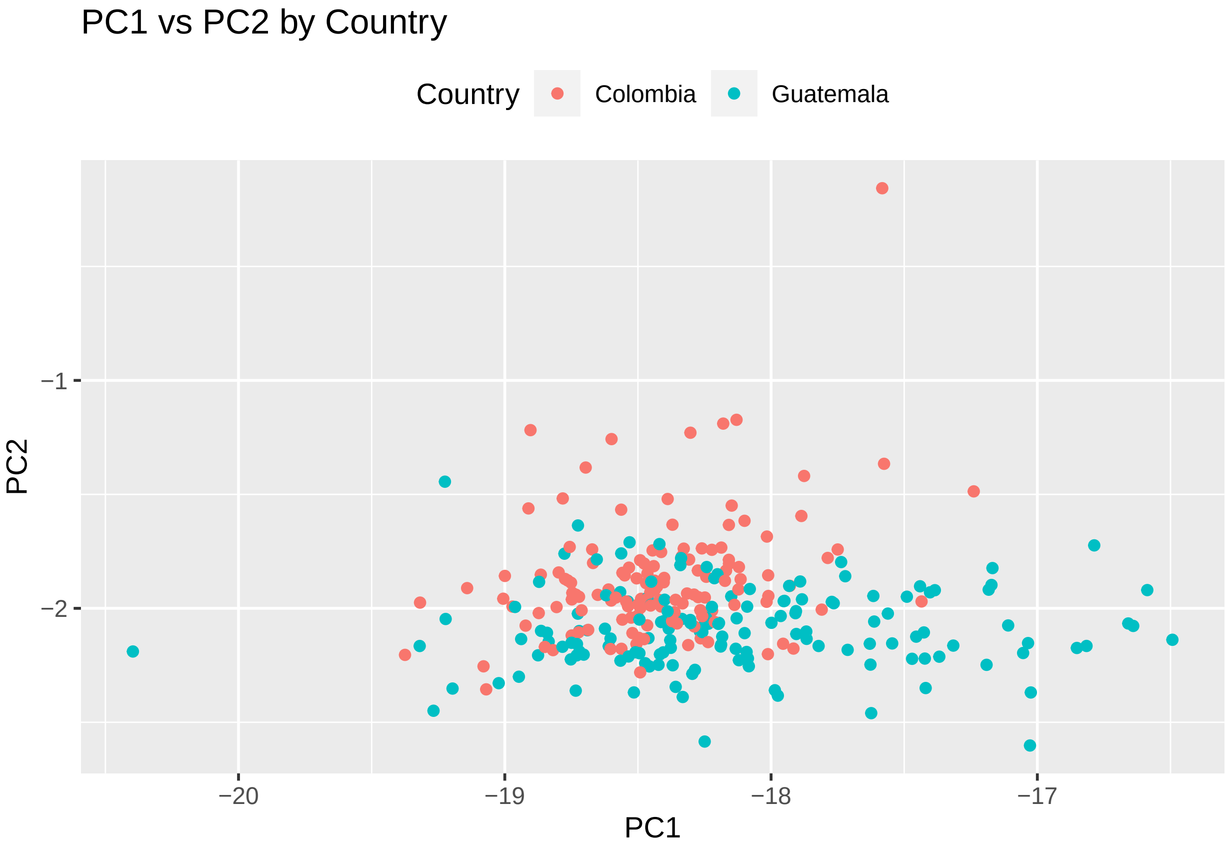


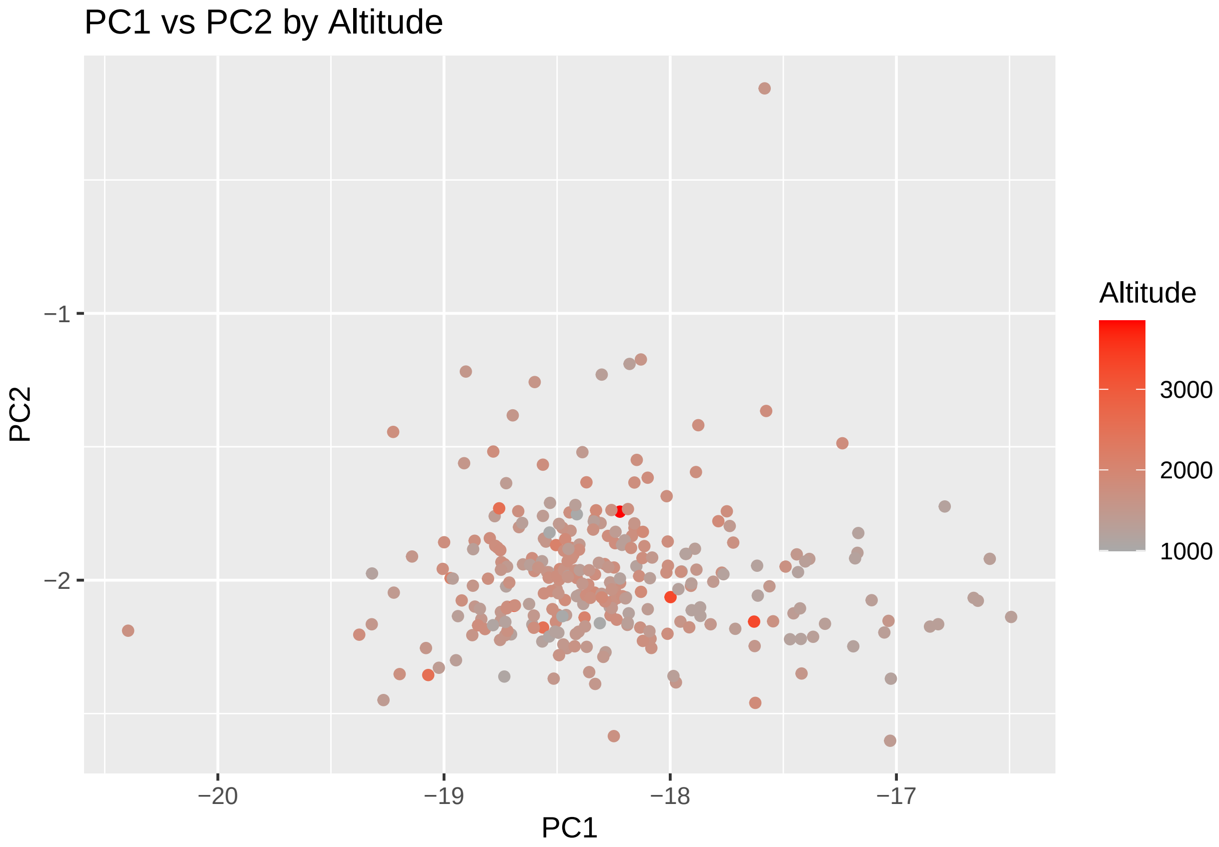
In the Pearson correlation matrix, the rating values correlate with eachother but only weakly with Altitude

```
##            Aroma Flavor Aftertaste Acidity  Body Balance Altitude
## Aroma      1.000  0.545      0.467  0.3270 0.420   0.357   0.1080
## Flavor     0.545  1.000      0.733  0.5520 0.624   0.581   0.1130
## Aftertaste 0.467  0.733      1.000  0.4560 0.680   0.641   0.2260
## Acidity    0.327  0.552      0.456  1.0000 0.459   0.321   0.0111
## Body       0.420  0.624      0.680  0.4590 1.000   0.596   0.1690
## Balance    0.357  0.581      0.641  0.3210 0.596   1.000   0.1210
## Altitude   0.108  0.113      0.226  0.0111 0.169   0.121   1.0000
```

Principal Component Analysis

We apply Principal Component Analysis to get uncorrelated components of maximum variation (leaving out altitude, which we will try to predict). As we saw above, we get pretty good separation by country but not by altitude.


PC1 vs PC2 by Country
Country
● Colombia
● Guatemala
PC2
−1
−2
PC1
−20
−19
−18
−17



Use the First Two Principal Componenets to Predict Altitude

As expected, the result is not a good fit. We cannot accurately infer altitude with PC regression from these attributes.

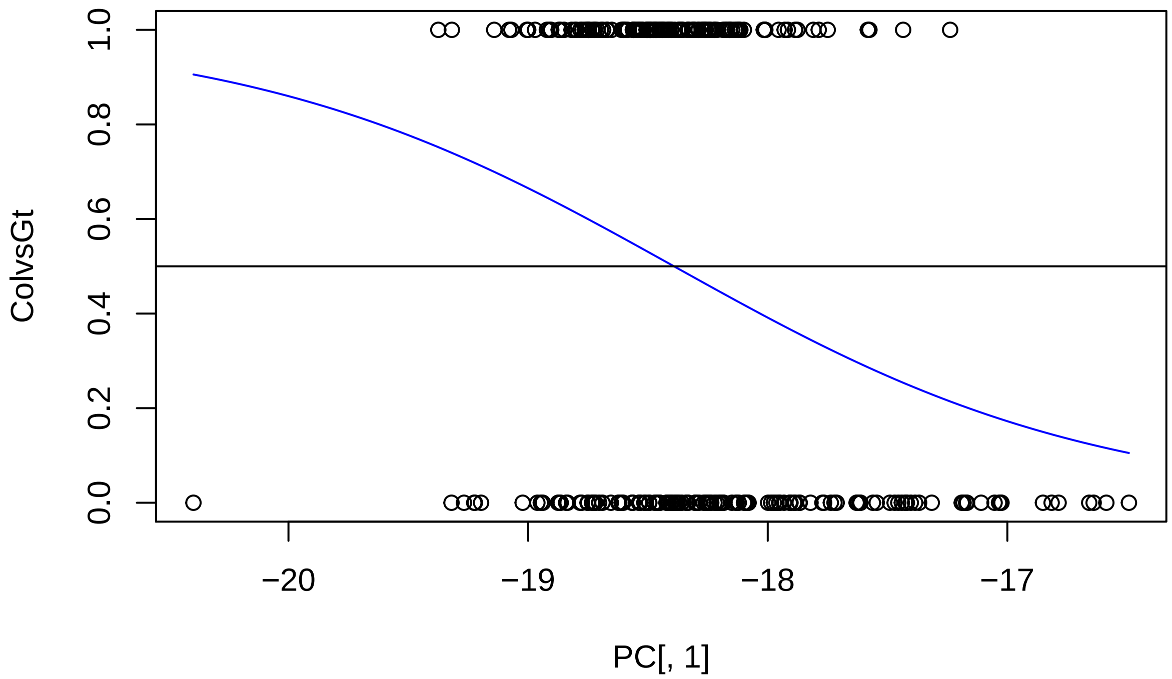
```
##          [,1]
## m1  -78.91548
## m2 -104.72197
## m3  125.99683
```

```
##            [,1]
## [1,] 0.03734761
```

Use the first principal component to predict country

We attempt to classify 1 as Colombia or 0 as Guatemala with logistic regression, using 0.5 as the threshold.

```
##      alpha       beta
## -20.739295  -1.127674
```



```
##            ColvsGt
## Predictions  0  1
##       FALSE 95 58
##       TRUE  56 77
```

```
## [1] 0.6013986
```

The accuracy is not bad and suggests that different coffee origins have different qualities.

Conclusion

We discovered that the growing and processing conditions as well as the attribute ratings of coffee beans vary by country (though not much by year). We've learned a lot about coffee in different countries but there is still a lot more that can be done with this dataset to complete the picture.