DATA ANALYSIS PRESENTATION
COFFEE QUALITY

BY ETHAN, DANIEL, SEAN, NHI

WHY WE CHOOSE THIS DATASET?

THE WORLD CONSUMES CLOSE TO 2.25 BILLION CUPS OF COFFEE EVERY DAY.

Coffee Quality Data (CQI May-2023)

Scraped data from CQI database

- WORKS TO IMPROVE THE QUALITY AND VALUE OF COFFEE WORLDWIDE.
- TO PROMOTE COFFEE QUALITY.
- WEB DATABASE THAT SERVES AS A RESOURCE FOR COFFEE PROFESSIONALS.
- INFORMATION ON COFFEE PRODUCTION, PROCESSING, AND SENSORY EVALUATION.

41 COLUMNS, 162 ENTRIES

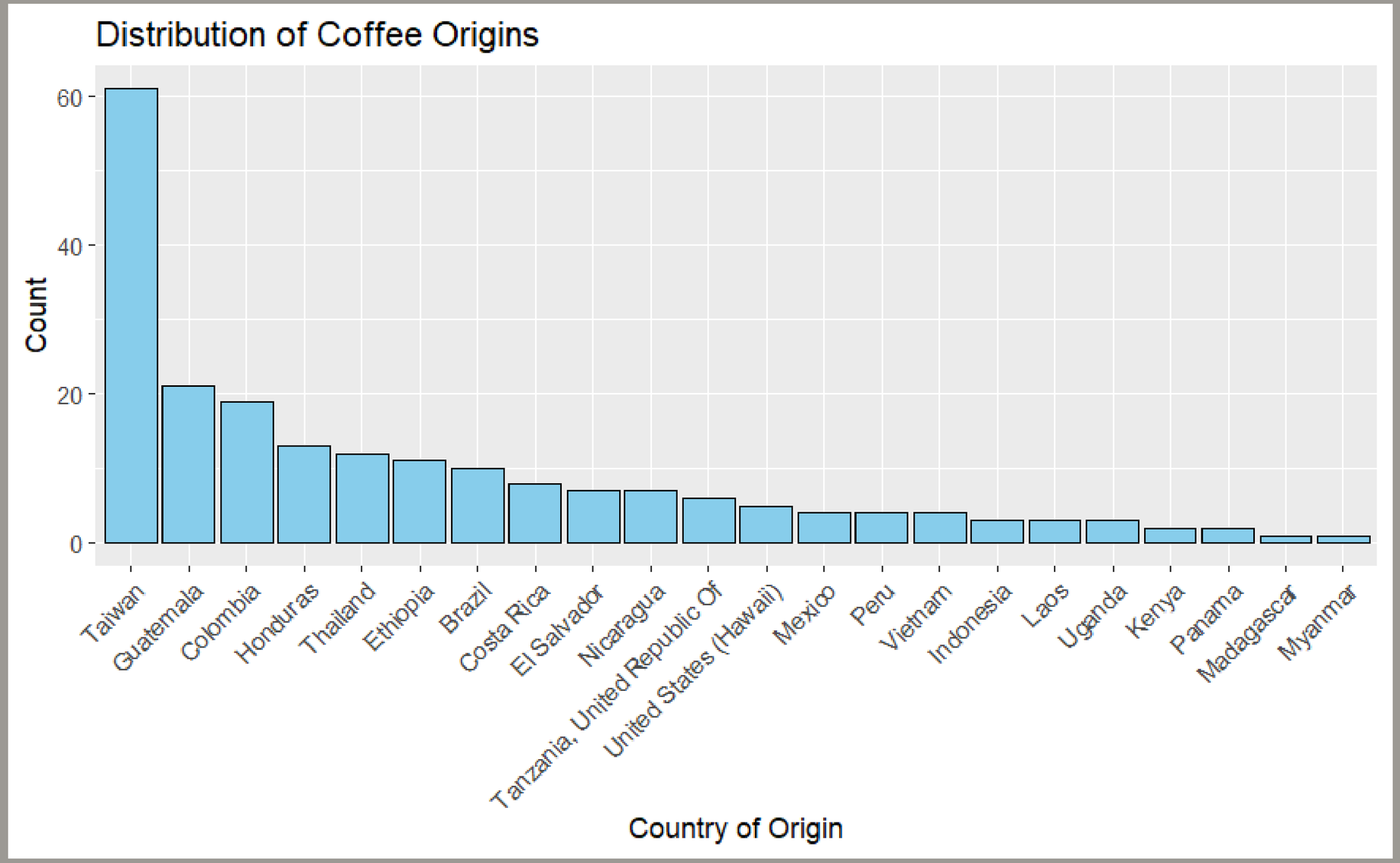
IMPORTANT COLUMNS:

- COUNTRY OF ORIGIN
- REGION
- PROCESSING METHOD
- VARIETY
- ALTITUDE
- COFFEE QUALITY SCORES (AROMA, SWEETNESS, FLAVOR..)

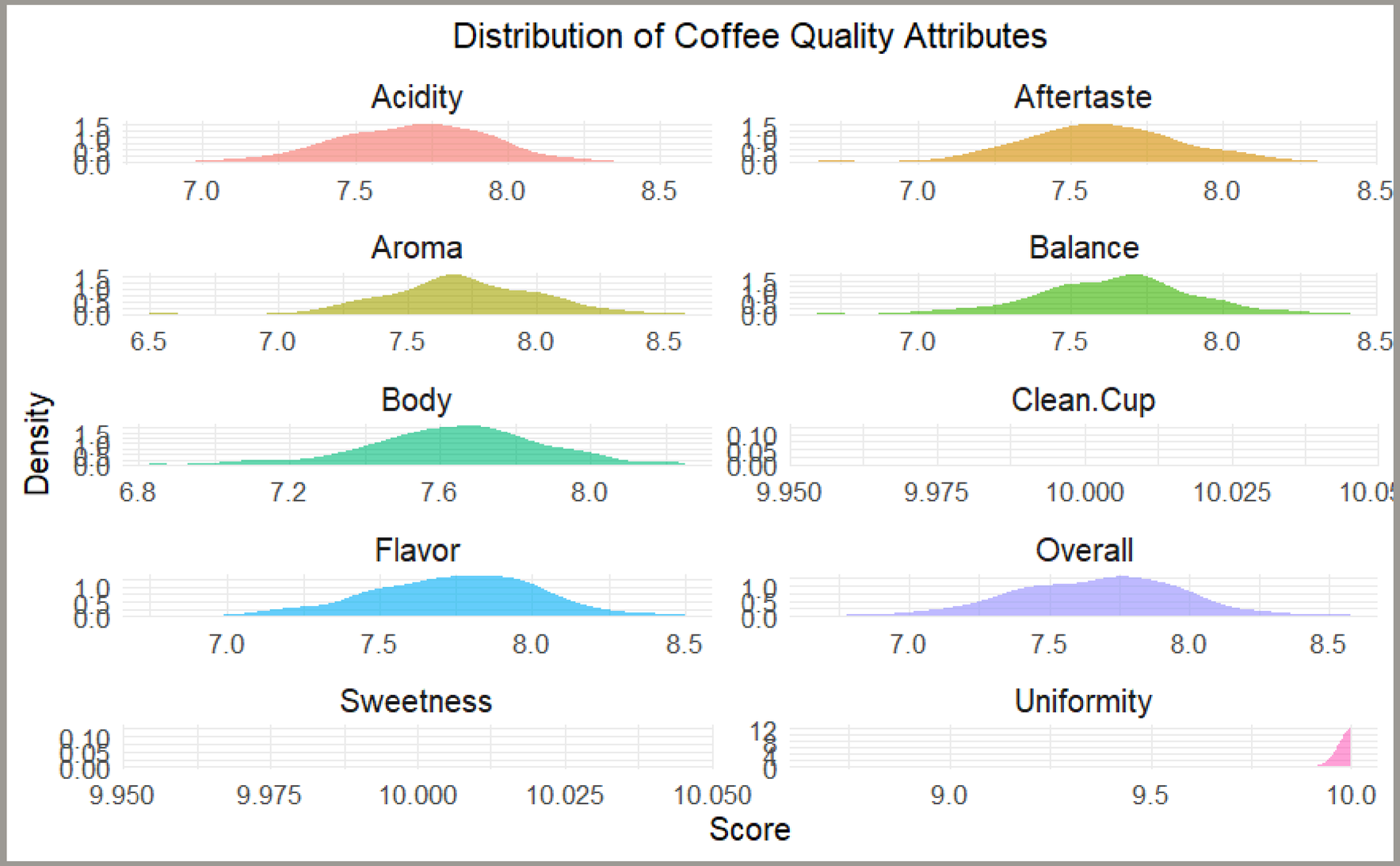
Country of Origin Which country the coffee produced in	
Taiwan	29%
Guatemala	10%
Other (125)	60%

Region	
Chiayi	6%
新竹縣	5%
Other (184)	89%

COUNTRY OF ORIGIN



COFFEE QUALITY ATRIBUTES DISTRIBUTION

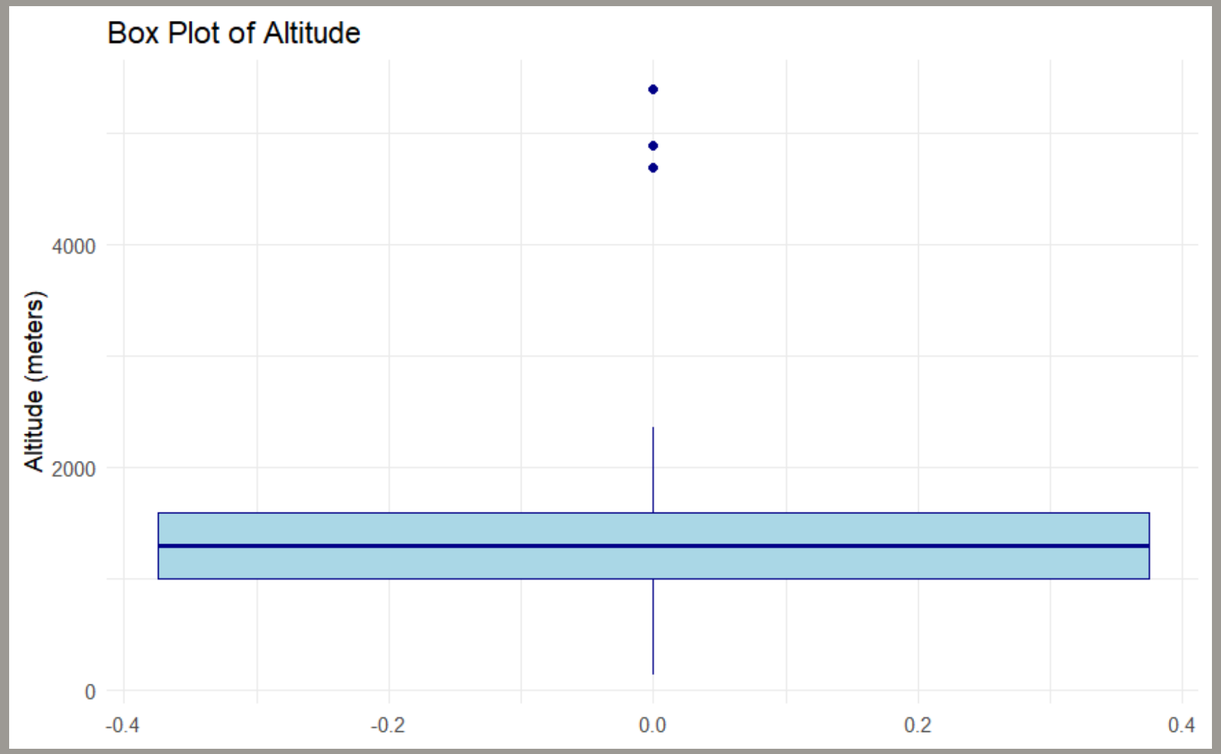


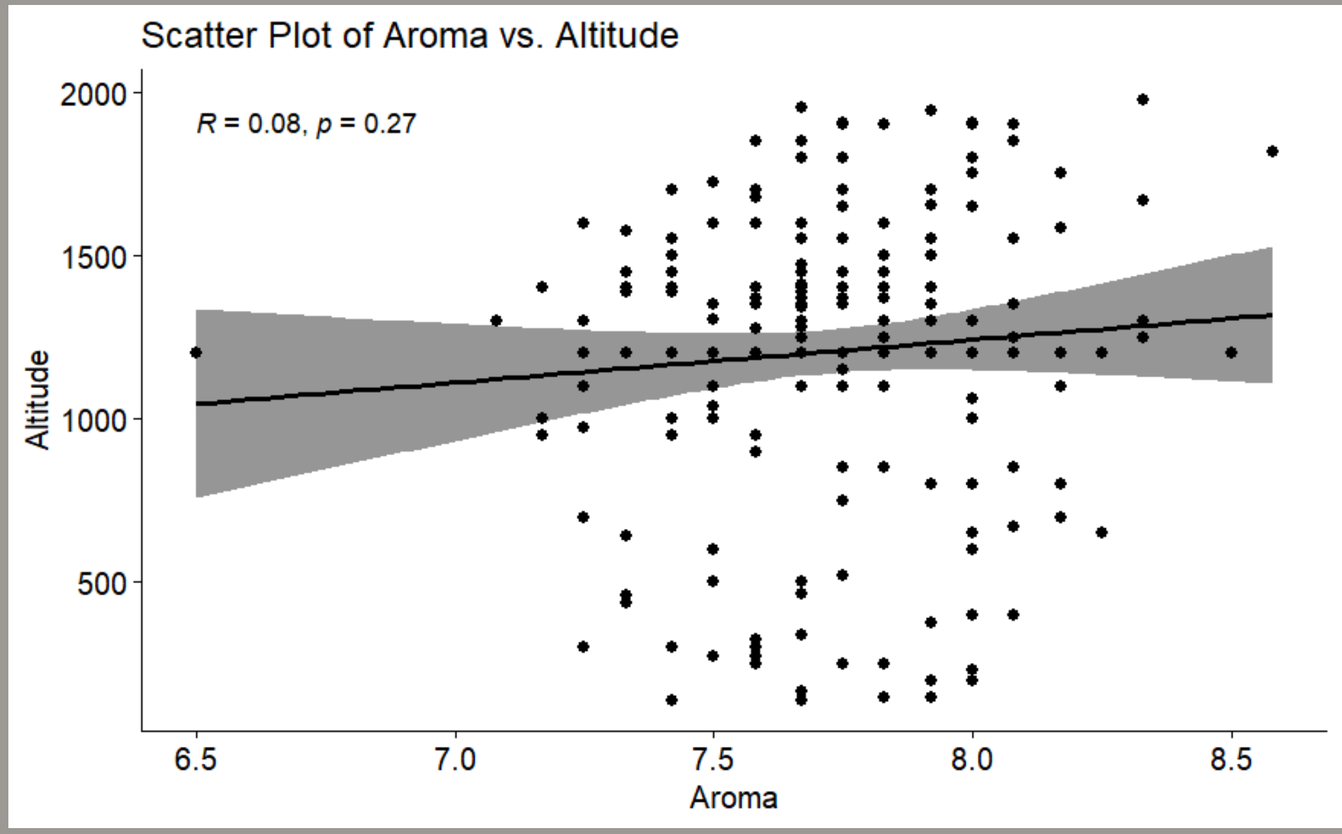
LINEAR REGRESSION

ALTITUDE AND AROMA

a relationship between Altitude and Aroma

Calculates the Pearson correlation coefficient between the ‘Altitude’ and ‘Aroma’ columns to see if there is actually a relationship between them

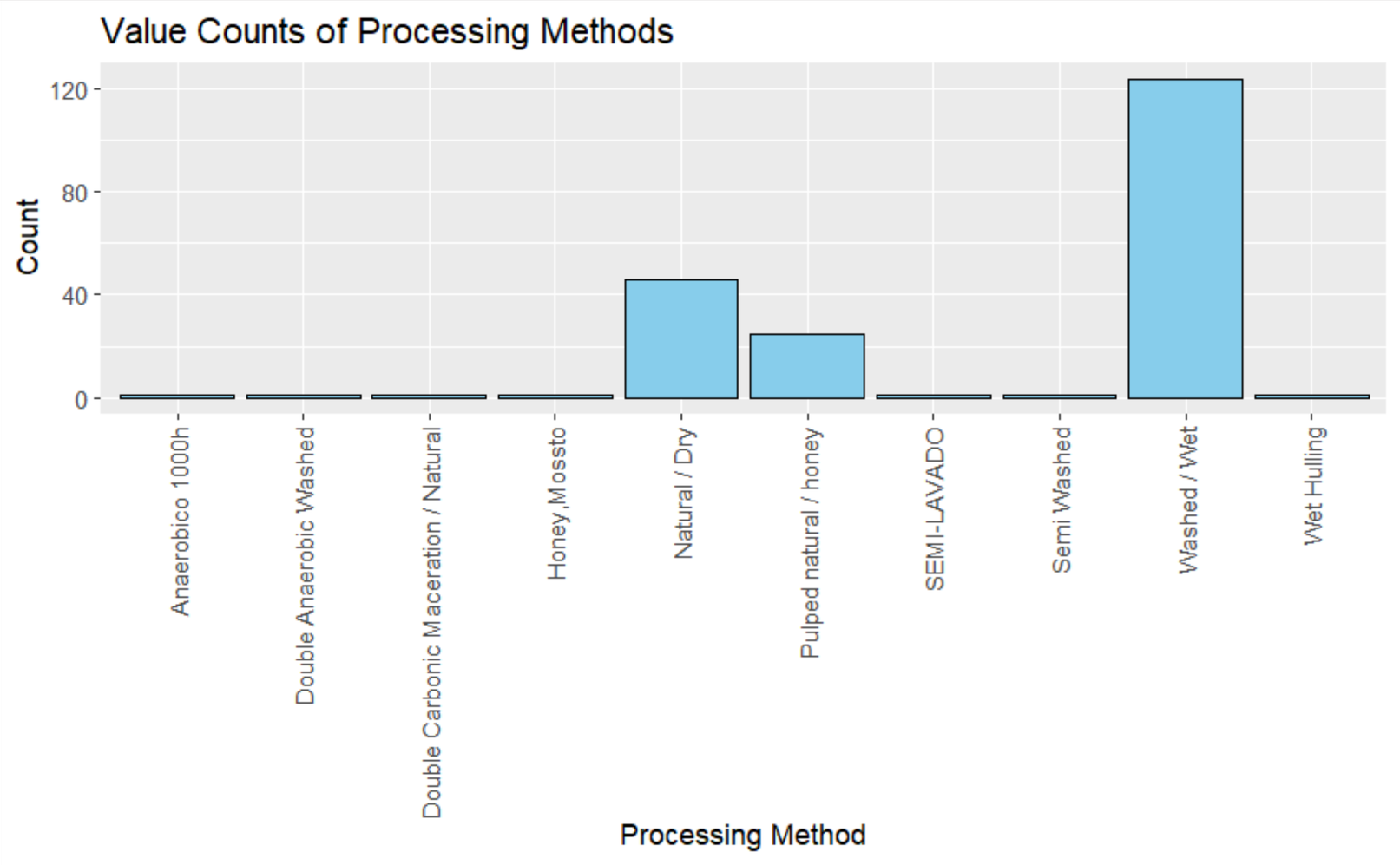




Removed 3 outliers to get better insight

LINEAR REGRESSION

PROCESSING METHOD



LINEAR REGRESSION

PROCESSING METHOD

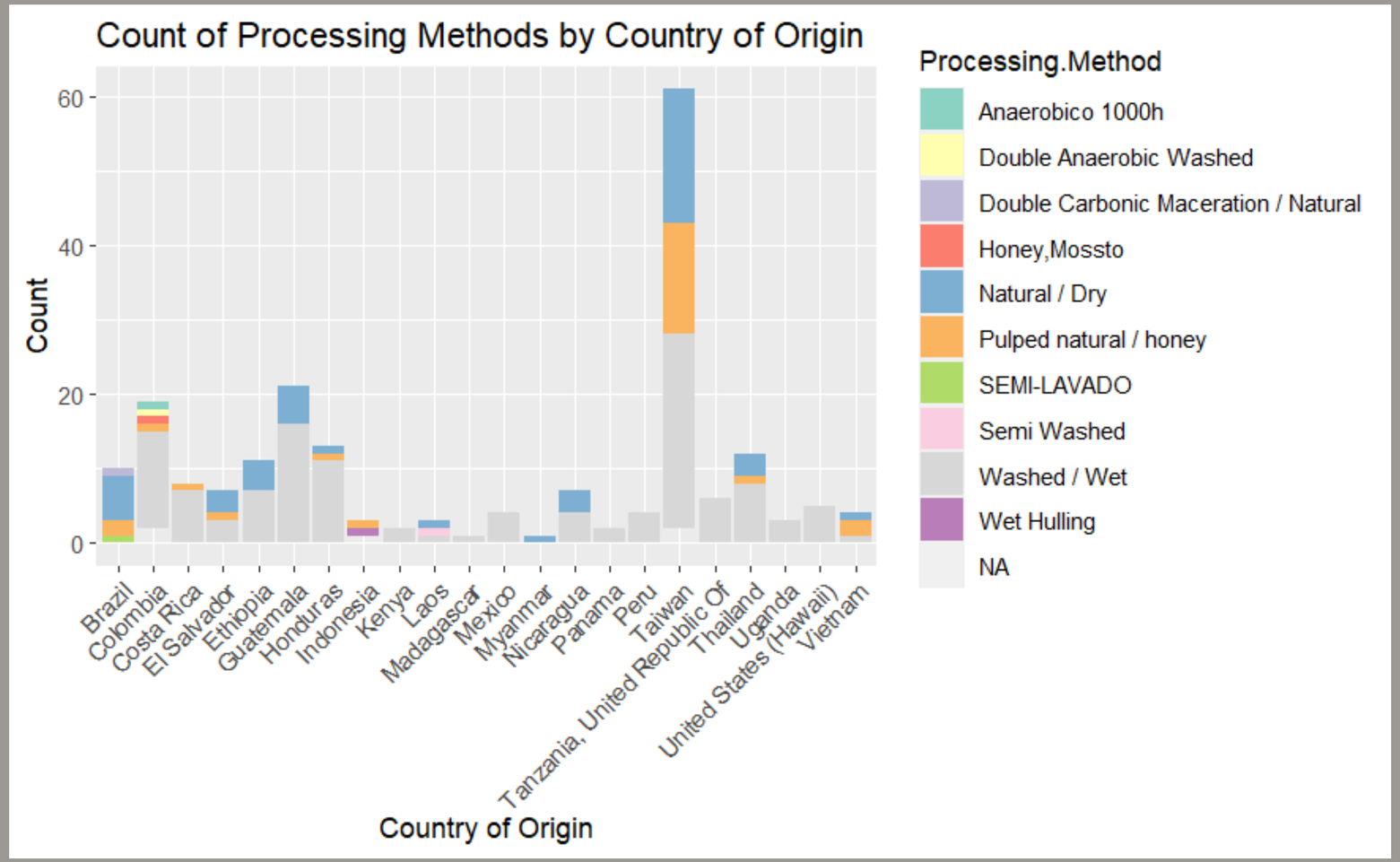
The average of all Scores, grouped by Processing Method

Processing.Method <chr>	Avg_Aroma <dbl>	Avg_Flavor <dbl>	Avg_Aftertaste <dbl>	Avg_Acidity <dbl>	Avg_Body <dbl>	Avg_Balance <dbl>	Avg_Uniformity <dbl>	Avg_Clean_Cup <dbl>	Avg_Sweetness <dbl>	Avg_Overall <dbl>
Anaerobico 1000h	7.670000	7.670000	7.580000	7.670000	7.580000	7.580000	10.000000	10	10	7.500000
Double Anaerobic Washed	8.580000	8.500000	8.420000	8.580000	8.250000	8.420000	10.000000	10	10	8.580000
Double Carbonic Maceration / Natural	7.830000	7.920000	7.750000	7.920000	7.670000	7.830000	10.000000	10	10	7.830000
Honey,Mossto	8.330000	8.330000	8.080000	8.250000	7.920000	7.920000	10.000000	10	10	8.250000
Natural / Dry	7.731739	7.744565	7.605217	7.678478	7.638043	7.641739	9.985435	10	10	7.677609
Pulped natural / honey	7.672800	7.730000	7.606400	7.676400	7.617200	7.614000	10.000000	10	10	7.637200
SEMI-LAVADO	7.250000	7.080000	6.670000	6.830000	6.830000	6.670000	10.000000	10	10	6.670000
Semi Washed	8.330000	8.420000	8.080000	8.170000	7.920000	8.170000	10.000000	10	10	8.330000
Washed / Wet	7.707984	7.733065	7.583226	7.678952	7.643548	7.640726	9.989274	10	10	7.670161
Wet Hulling	7.670000	7.670000	7.830000	7.830000	7.670000	7.750000	10.000000	10	10	7.830000

LINEAR REGRESSION

PROCESSING METHOD

the count of processing methods for each country of origin



CHI SQUARED TEST

THE COUNTRY ORIGINS

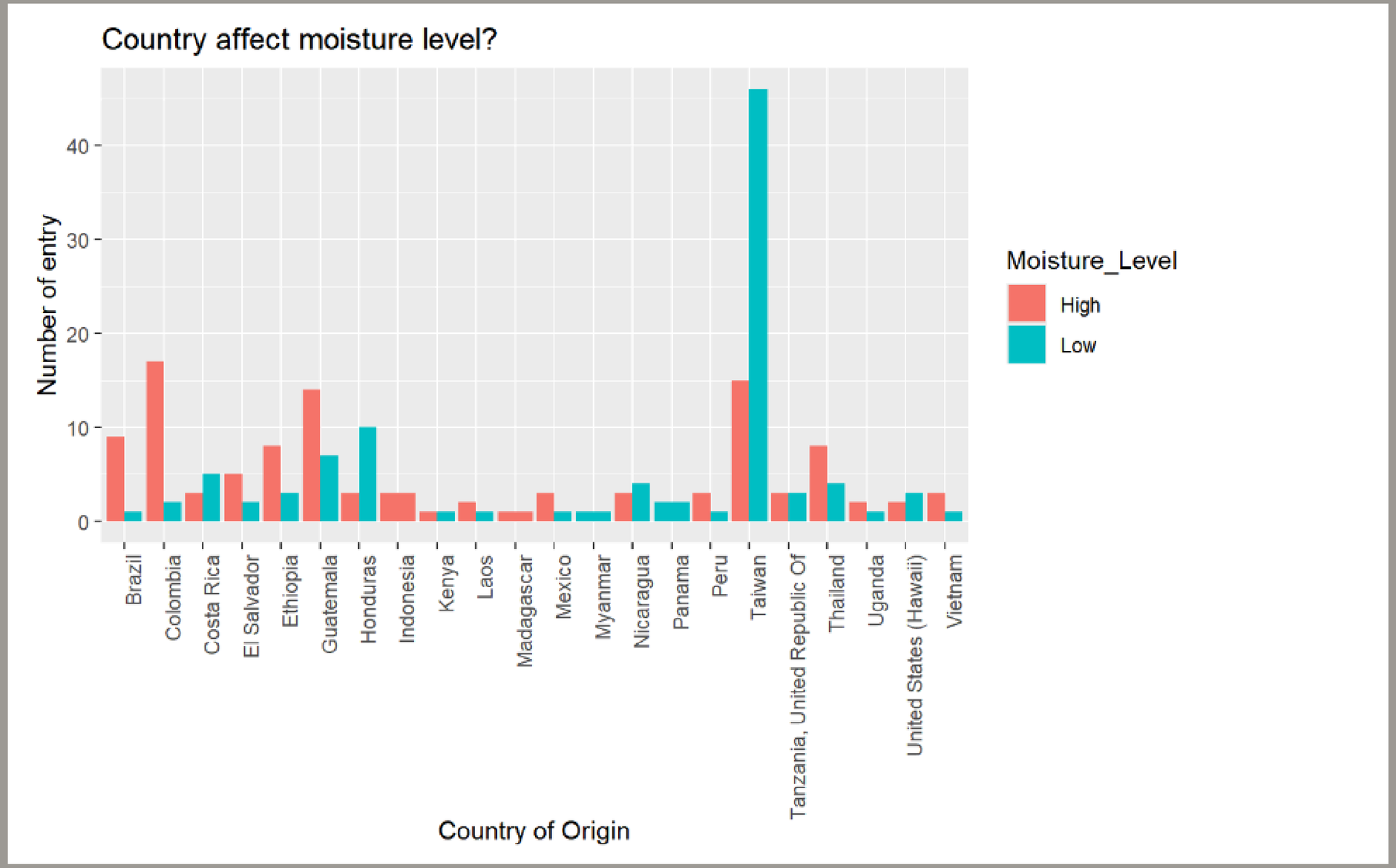
```
X-squared = 56.215, df = 21, p-value = 4.692e-05
```

With a p-value of 4.692e-05, it is significantly smaller than the chosen significance level of 0.05.

Since the p-value is less than 0.05, we reject the null hypothesis. Therefore, we conclude that there is a significant association between Country of Origin and Moisture_Level. In other words, they are not independent.

Based on this we can see that the country of origin does have an influence on the moisture level of the coffee samples in the dataset.

VISUALIZE THE RELATIONSHIP BETWEEN ML AND CO



CHI SQUARE

THE PROCESSING METHOD

Based of the Chi Square test that i did i got

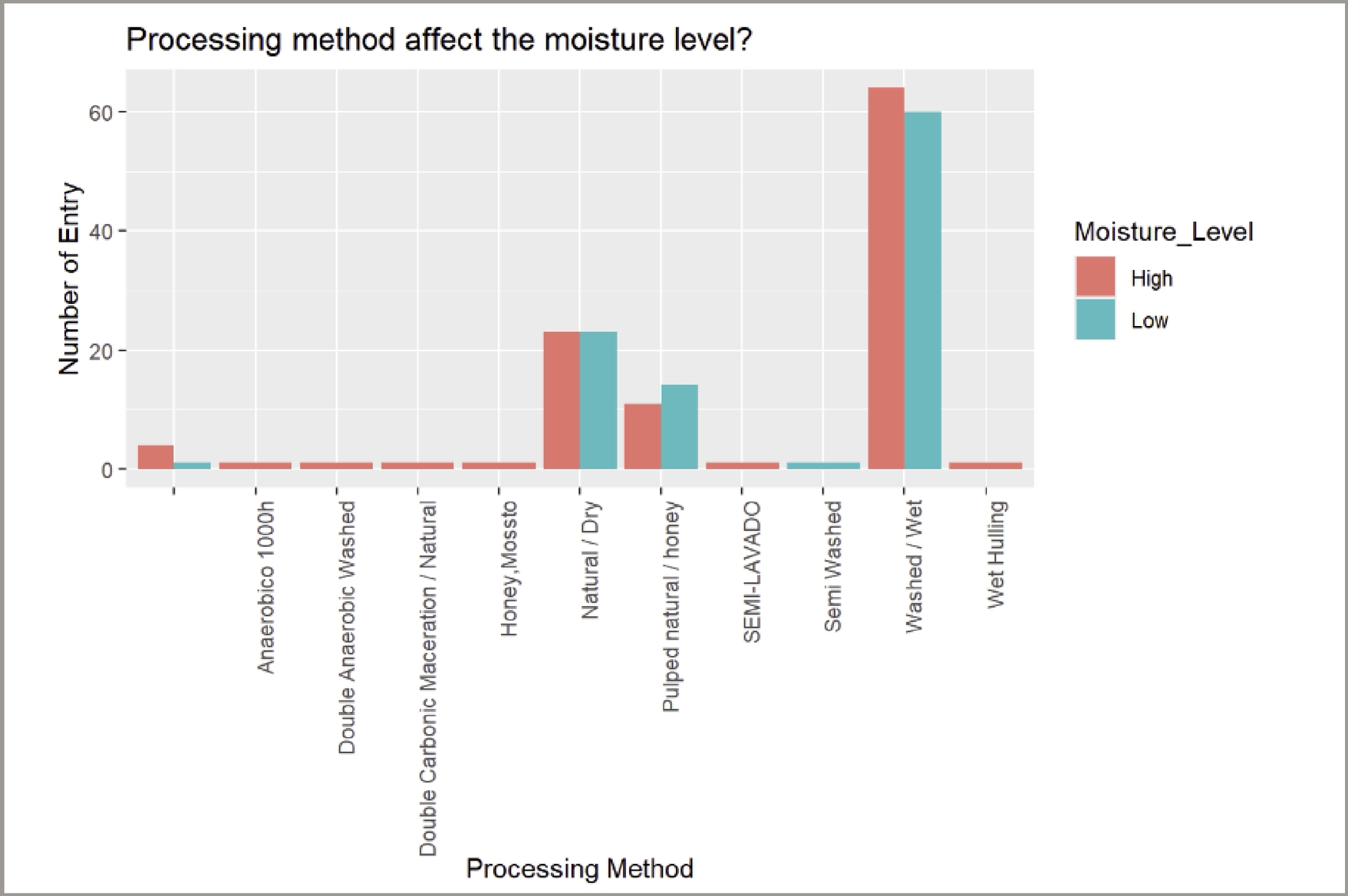
```
X-squared = 8.9146, df = 10, p-value = 0.5402
```

With a p-value of 0.5402, it is greater than the chosen significance level of 0.05.

Since the x^2-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, we conclude that there is not enough evidence to suggest a significant association between `Processing Method` and `Moisture Level`. In other words, they are considered independent.

This suggests that the processing method does not have a significant influence on the moisture level of the coffee samples in the dataset.

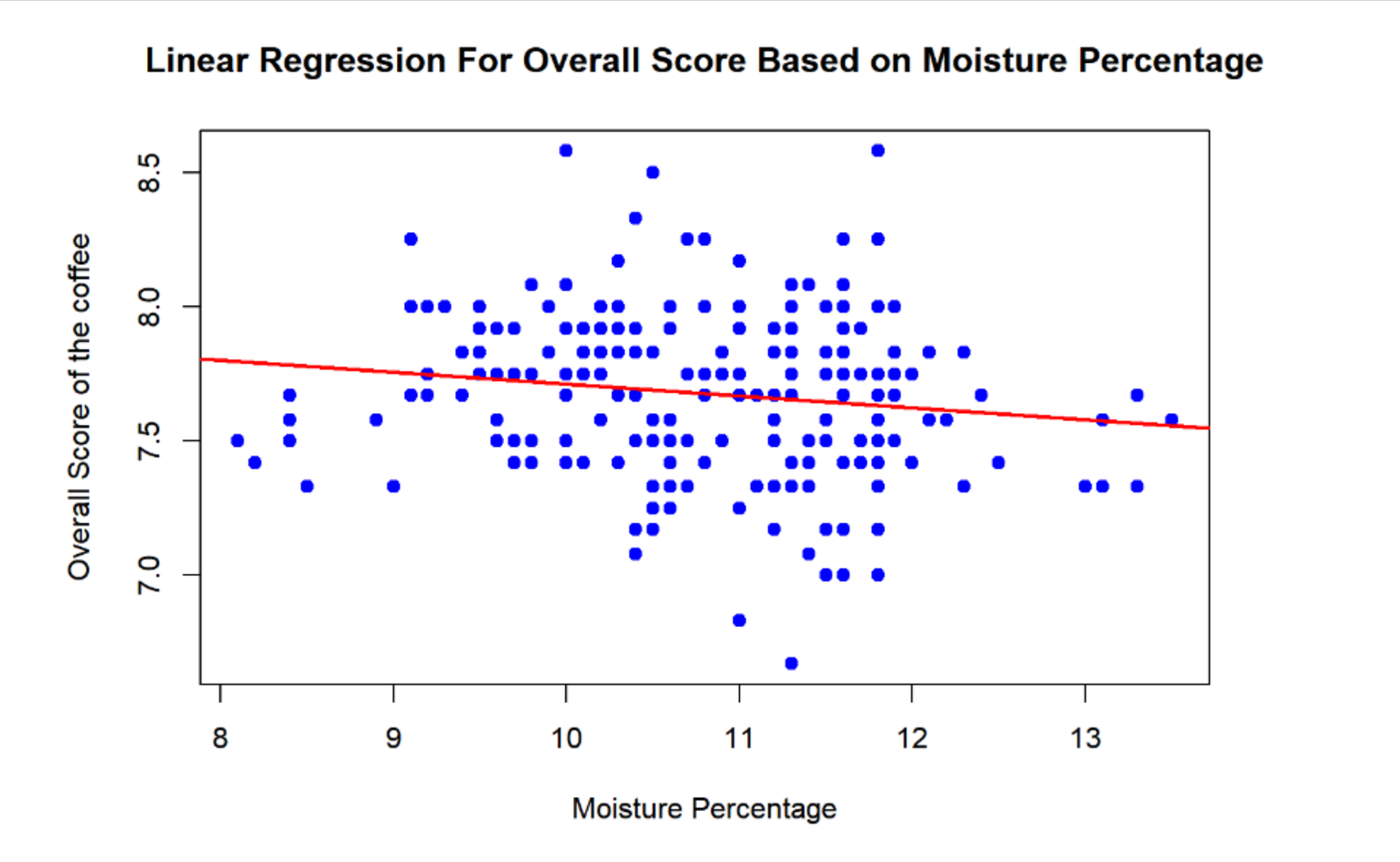
VISUALIZE THE RELATIONSHIP BETWEEN ML AND PM



LINEAR REGRESSION

THE MOISTURE PERCENTAGE

Moisture percentage can predict the Overall score of a coffee



Does the average ‘overall’ score of coffee change with region?

Variables Used:

- Country of Origin
- Overall (1-10 rating of the coffee)

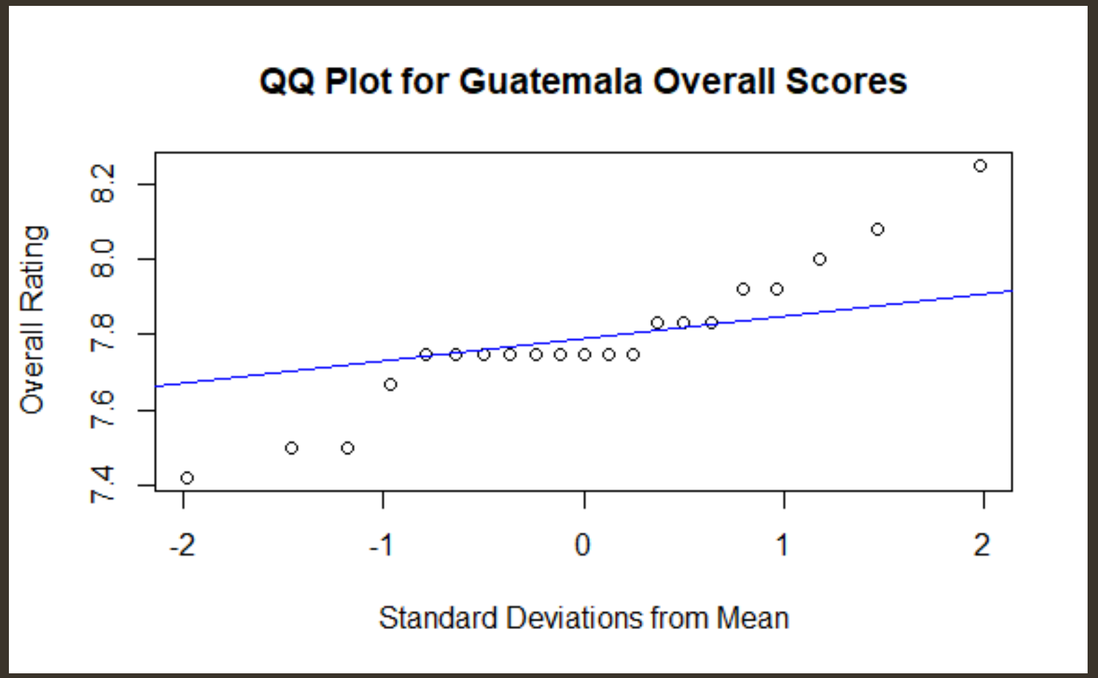
Choosing Regions

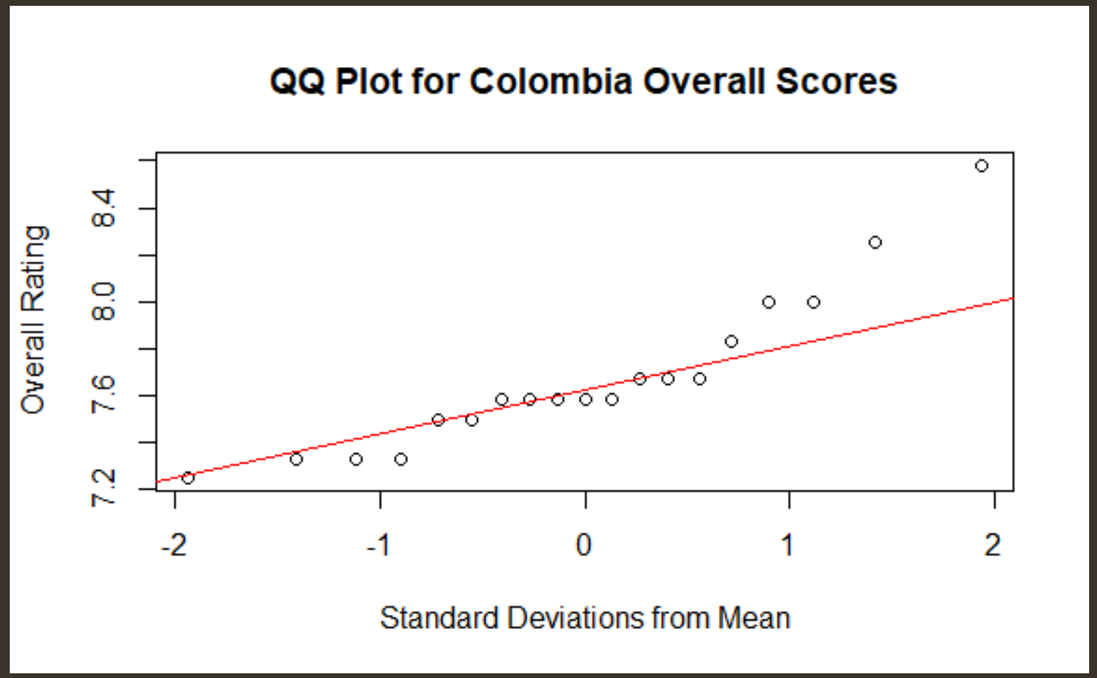
From this we chose to compare coffee from South America and Asia.

To the right is a count of how many coffee's came from a specific country. -->

Country.of.Origin <chr>	n <int>
Taiwan	61
Guatemala	21
Colombia	19
Honduras	13
Thailand	12
Ethiopia	11
Brazil	10
Costa Rica	8
El Salvador	7
Nicaragua	7

Country.of.Origin <chr>	n <int>
Tanzania, United Republic Of	6
United States (Hawaii)	5
Mexico	4
Peru	4
Vietnam	4
Indonesia	3
Laos	3
Uganda	3
Kenya	2
Panama	2
Madagascar	1
Myanmar	1

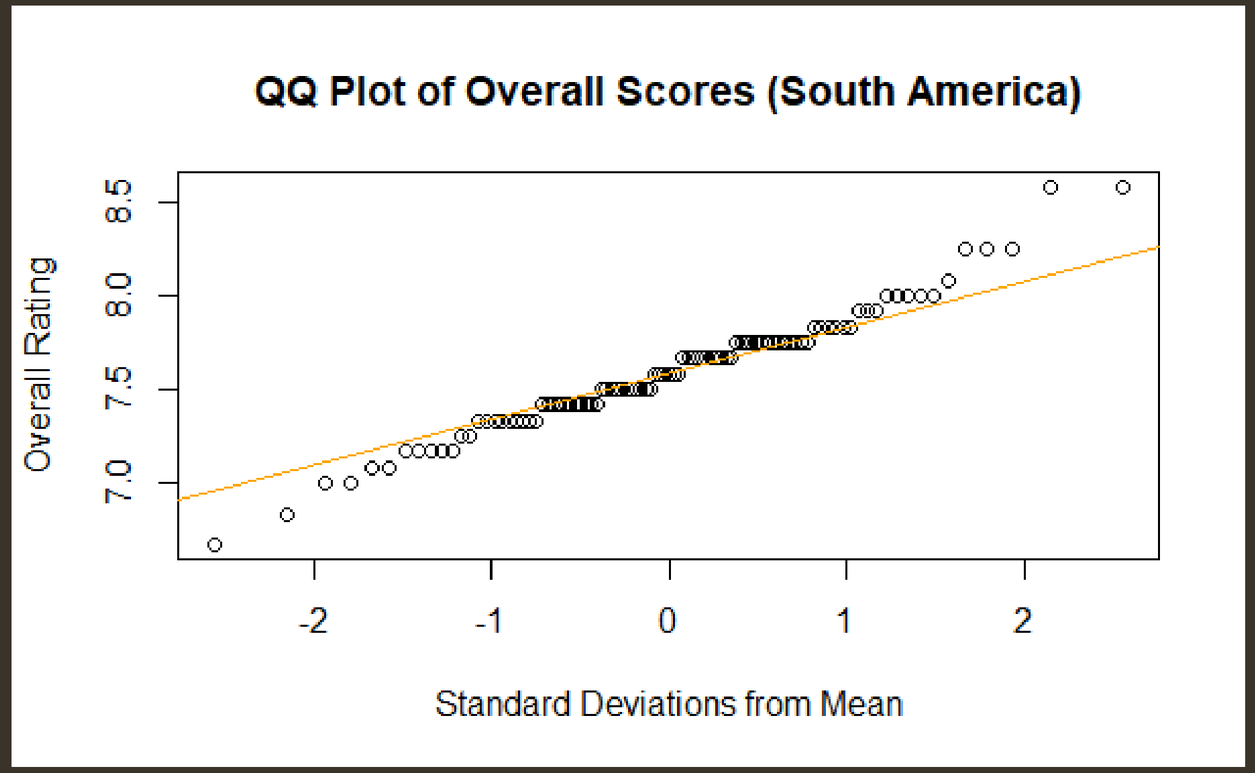


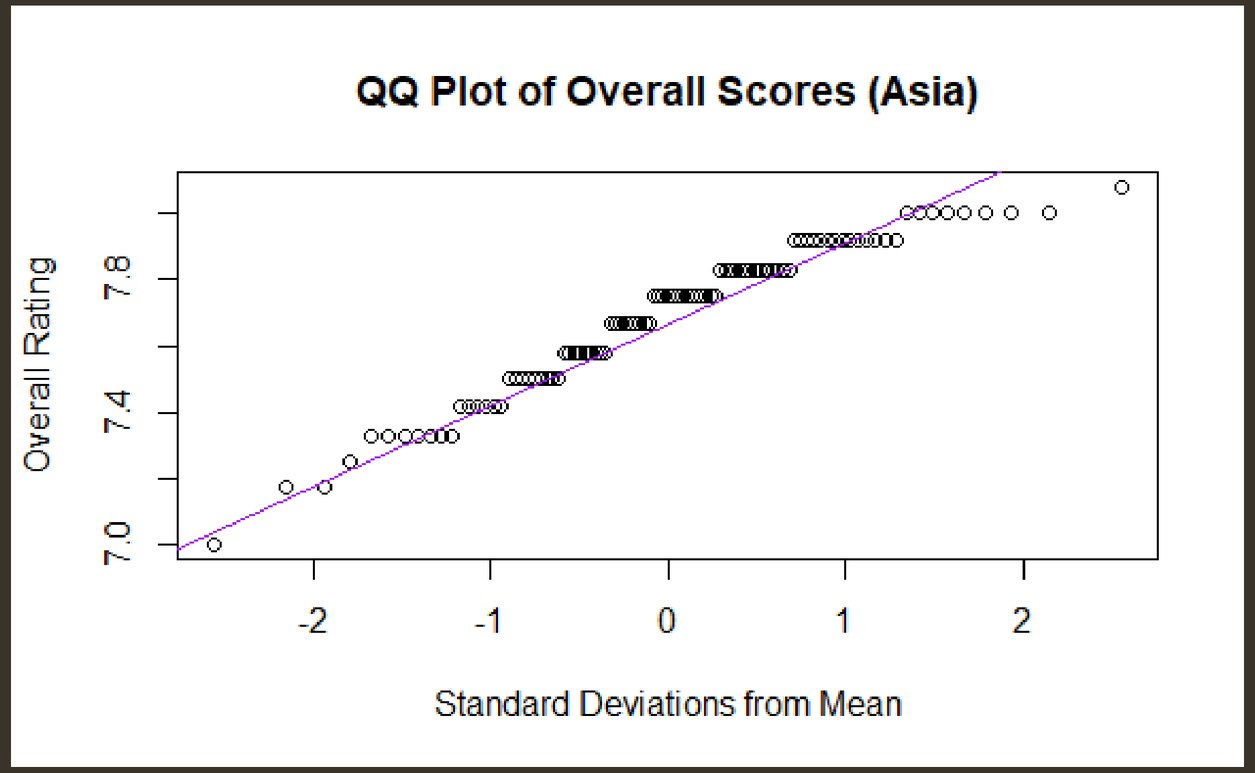



```
        Welch Two Sample t-test

data:  df_guatemala$Overall and df_colombia$Overall
t = 1.2767, df = 27.718, p-value = 0.2123
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06748249  0.29049001
sample estimates:
mean of x mean of y
 7.785714  7.674211
```

Confirming that the distribution is normal:

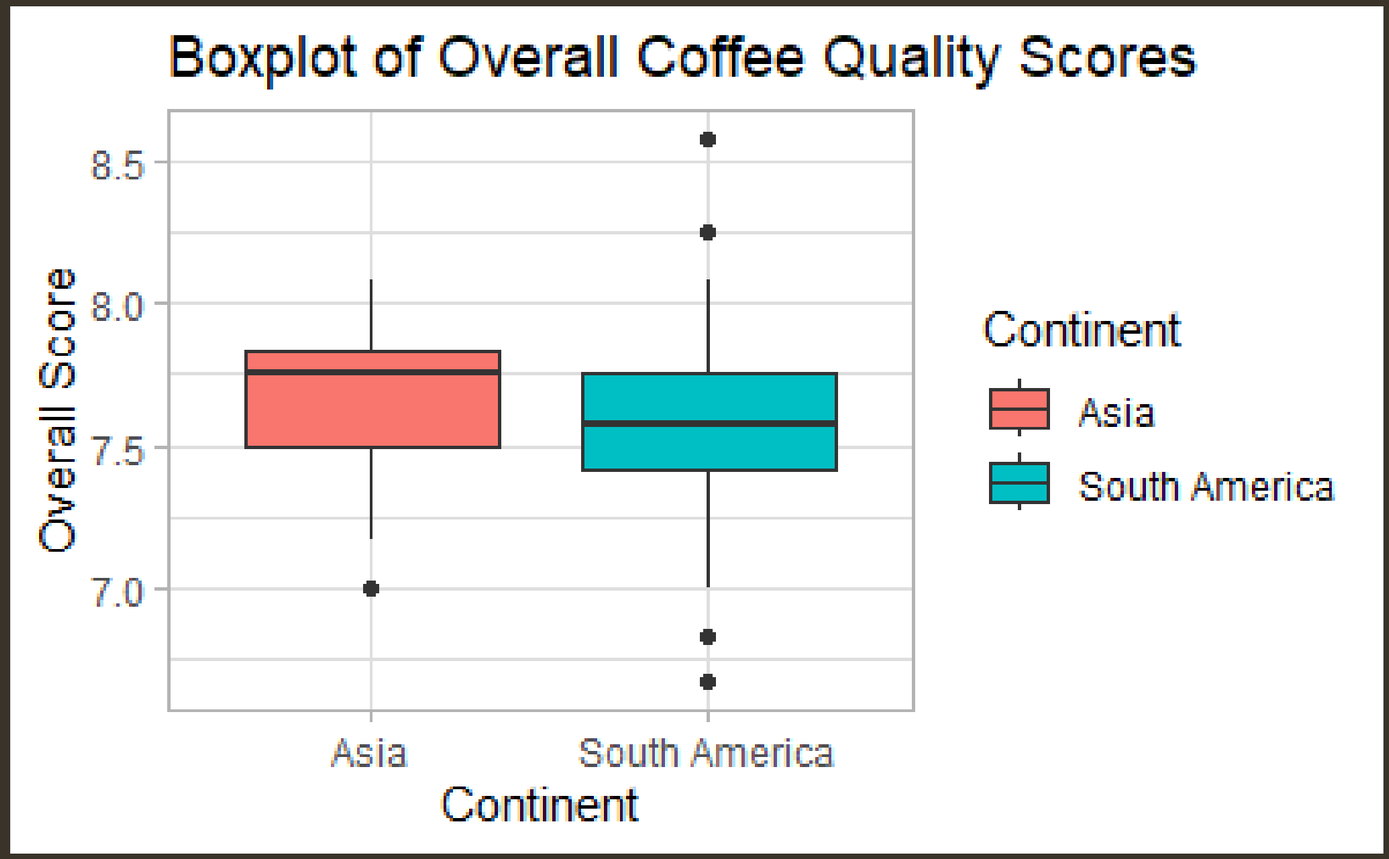




*Both data sets have 95 data points.

Observations:

- Varience between Continents arent equal
- Overall rating is higher in Asia on Average
- South America has the best and worst rated coffee.



T-Test Resaults:

```
        Welch Two Sample t-test

data:  df_south_america$Overall and df_asia$Overall
t = -2.3144, df = 168.02, p-value = 0.02185
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.17691149 -0.01403588
sample estimates:
mean of x mean of y
 7.591368  7.686842
```

H0(Null): The means of the two groups are equal.

H1(Alternative):The means of the two groups are not equal.

Given the resaults of the T-test, we reject H0. This implies that on average the Average overall rating of coffee from Asia is higher than that from of South America.

How is Altitude correlated to the scores received by each coffee variety

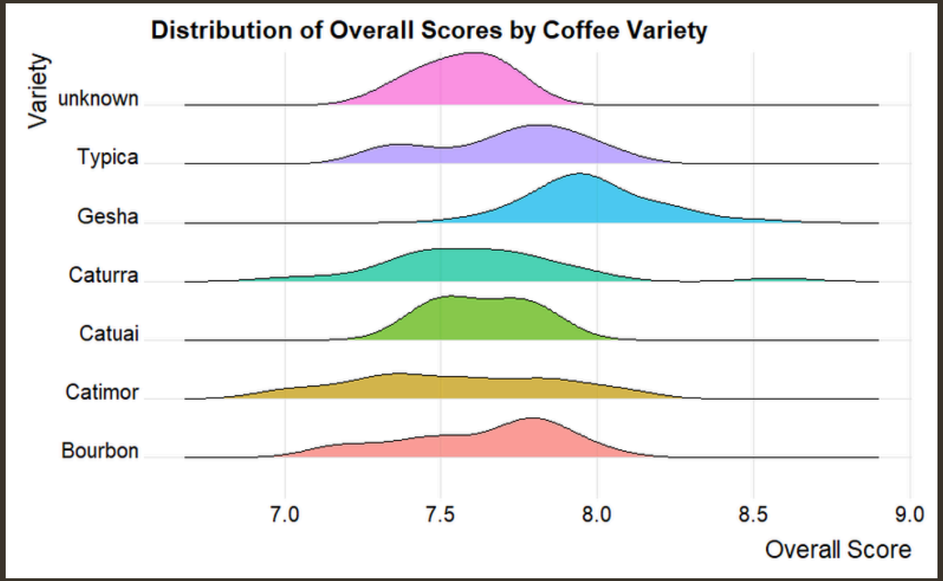
Variables Used:

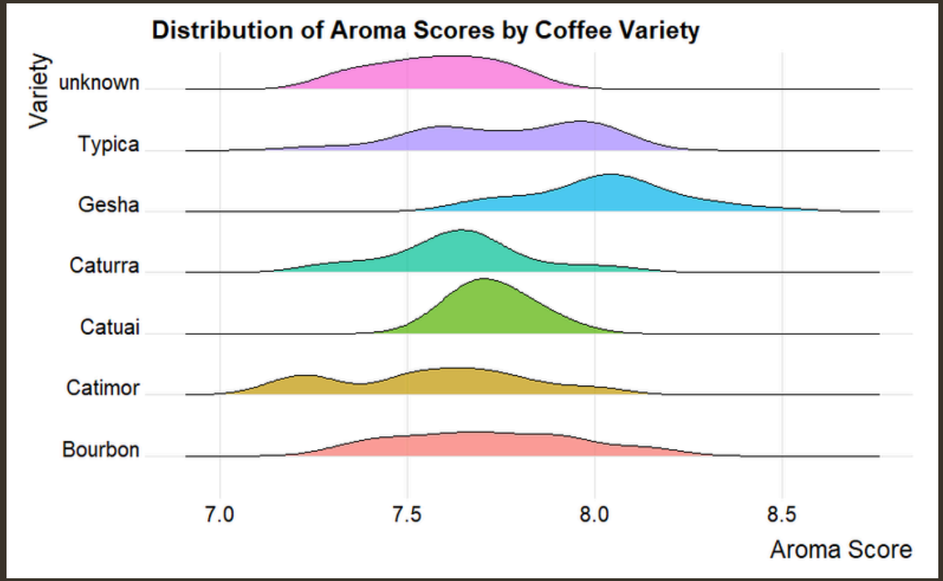
- Variety
- Altitude
- Overall score
- Flavor
- Aroma
- Acidity
- Body
- Balance
- Aftertaste

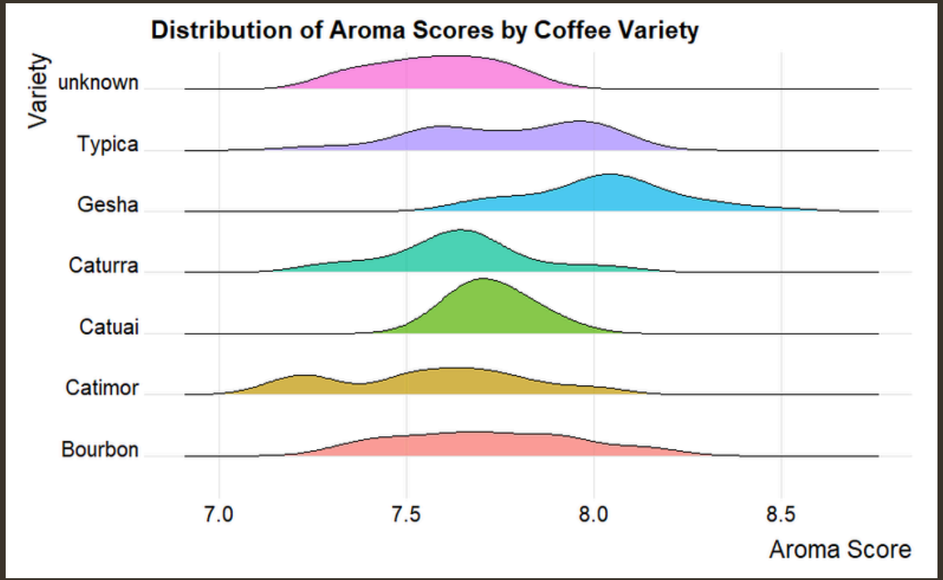
Subsetting and cleaning of Data

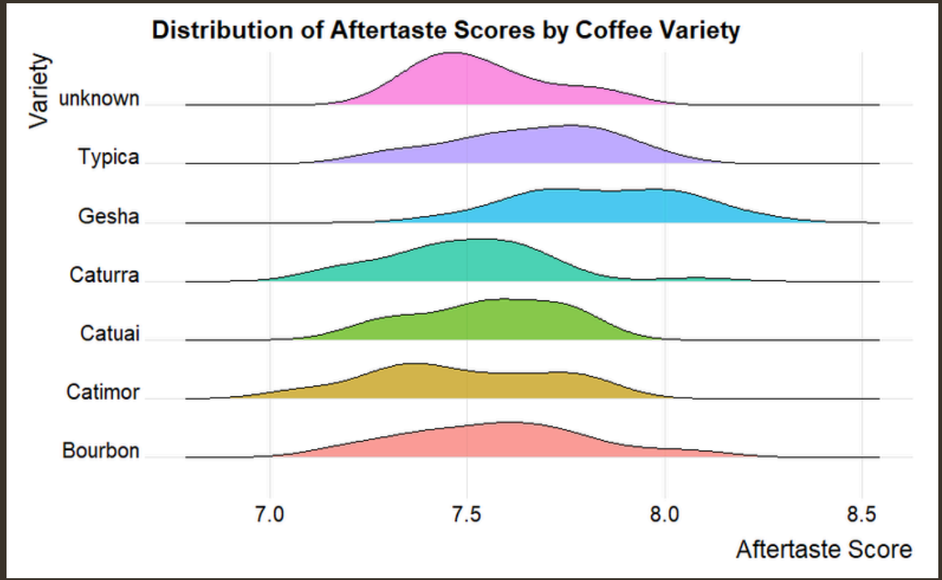
	Altitude <dbl>	Variety <chr>	Aroma <dbl>	Flavor <dbl>	Aftertaste <dbl>	Acidity <dbl>	Body <dbl>	Balance <dbl>	Overall <dbl>
2	1200	Gesha	8.50	8.50	7.92	8.00	7.92	8.25	8.50
4	1900	Gesha	8.08	8.17	8.17	8.25	8.17	8.08	8.25
6	1668	Gesha	8.33	8.33	8.25	7.83	7.83	8.17	8.25
7	1250	Gesha	8.33	8.17	8.08	8.00	7.83	8.25	8.25
10	1550	Bourbon	8.08	8.17	8.08	8.17	8.00	8.00	8.00
12	2000	Gesha	8.08	8.00	8.00	7.75	8.25	8.17	8.00

Distribution of the variables with respect to the variety of coffee

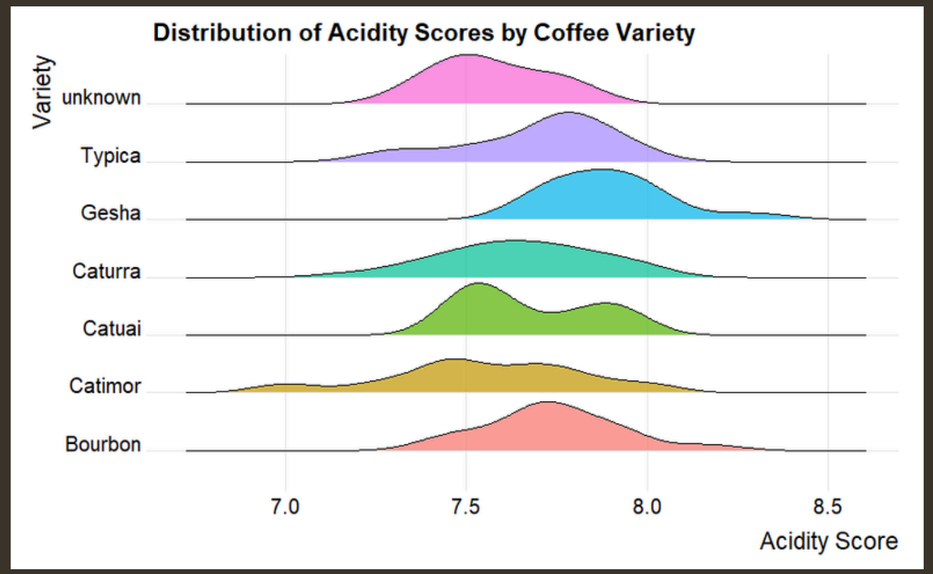


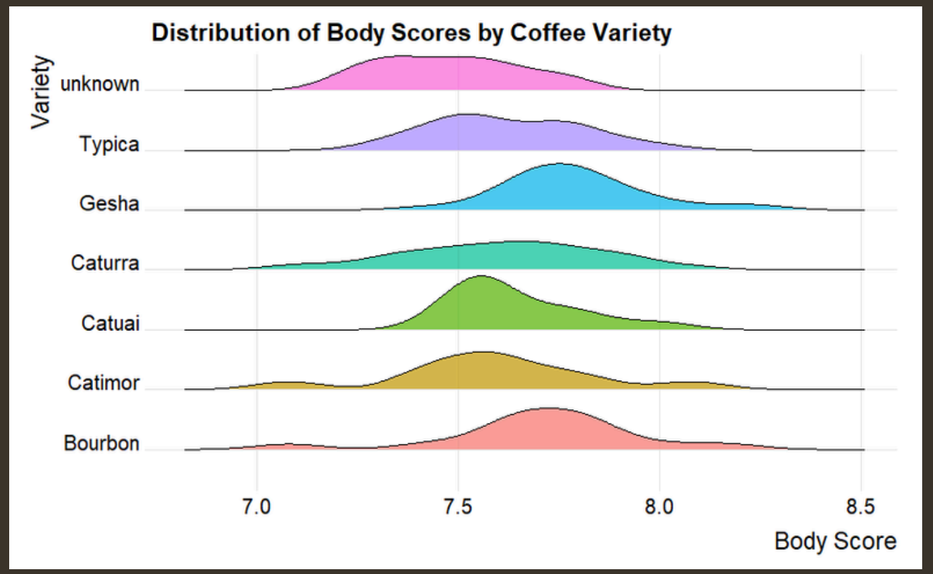


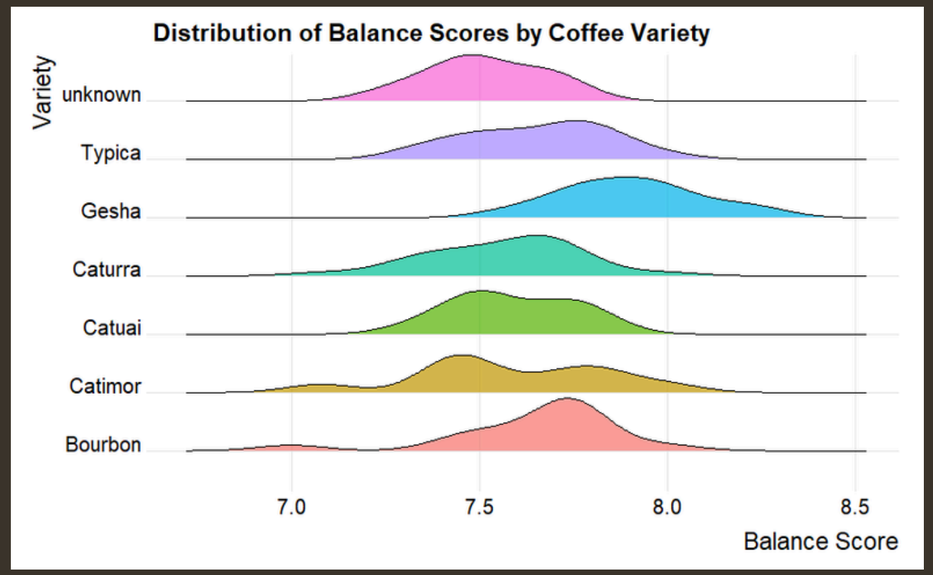




Distribution of the variables with respect to the variety of coffee



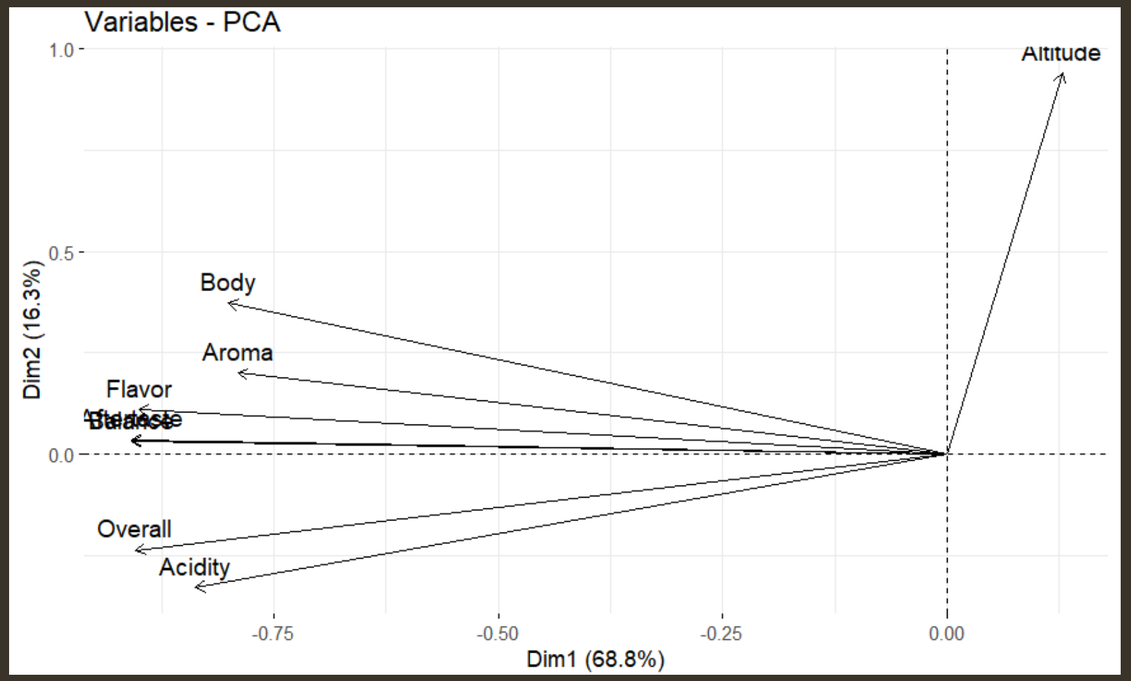


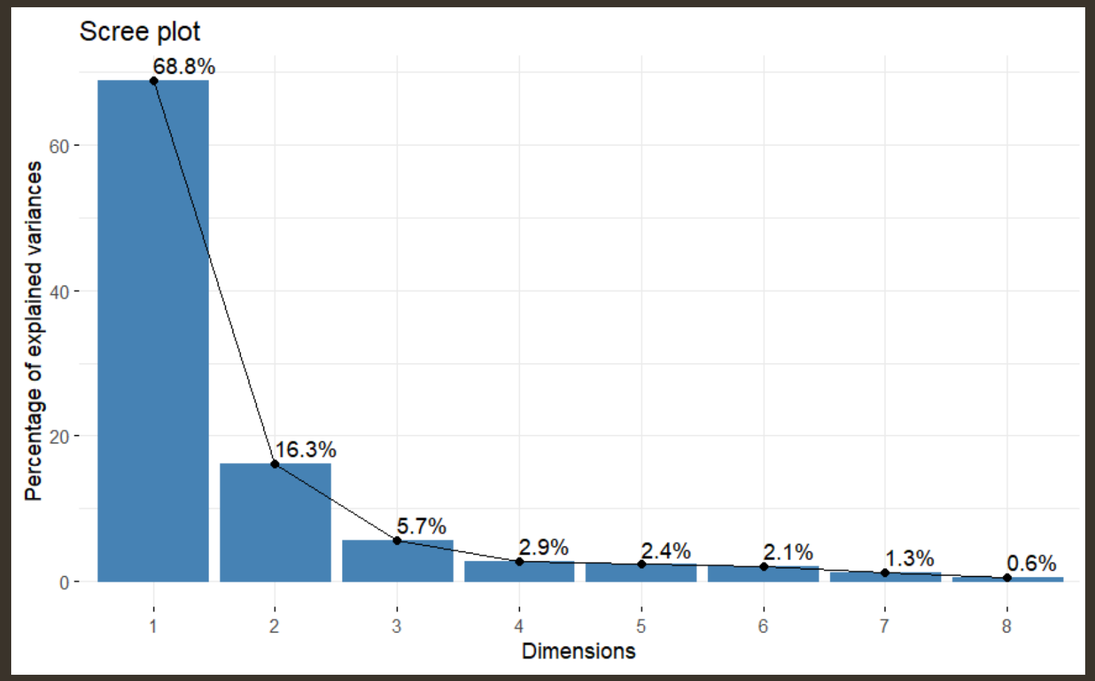


PCA for Typica

```
      Altitude     Aroma    Flavor Aftertaste   Acidity      Body   Balance   Overall
22  0.47474717 1.3625618 0.9836170 1.27831707 1.0903720 2.0632483 0.9354839 0.8909519
30 -0.02384575 1.0133816 0.5927196 1.66968203 0.6469397 0.2590945 1.8453700 1.2296609
31 -0.82159442 1.0133816 0.9836170 0.83803149 0.6469397 0.6964652 1.4171883 1.5683699
32 -0.82159442 0.6642014 0.9836170 0.83803149 1.0903720 1.1338358 0.9354839 1.2296609
41 -0.42272008 1.0133816 0.5927196 0.83803149 1.0903720 0.6964652 0.5073022 0.8909519
42  0.77390292 1.0133816 1.3745145 0.05530157 0.2527778 1.6258777 0.9354839 0.5099043
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Comp.7     Comp.8
Standard deviation     2.2942587 1.1154414 0.65759043 0.46710222 0.42966598 0.40357178
0.31405888 0.21821080
Proportion of Variance 0.6878598 0.1625956 0.05651011 0.02851275 0.02412554 0.02128417
0.01288954 0.00622254
Cumulative Proportion  0.6878598 0.8504554 0.90696546 0.93547821 0.95960375 0.98088792
0.99377746 1.00000000
                Comp.1      Comp.2
Altitude     0.05608669  0.84381882
Aroma       -0.34433139  0.17882435
Flavor      -0.39263247  0.09727396
Aftertaste  -0.39598903  0.03167017
Acidity     -0.36490388 -0.29654868
Body        -0.34918727  0.33385705
Balance     -0.39599655  0.02690286
Overall     -0.39423692 -0.21307514
```

PCA for Typica





PCA for Gesha

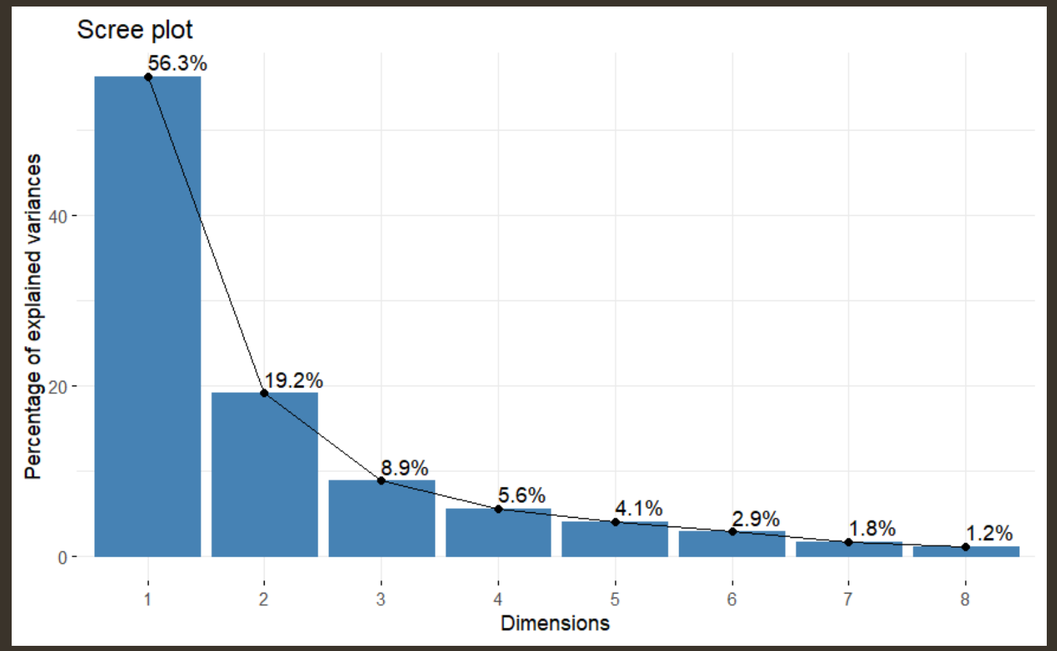
	Altitude	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Overall
2	-0.17478354	2.3533446	2.48911286	0.3538694	0.6633908	0.7350184	1.8950926	2.68612694
4	1.40127034	0.2618926	0.81403767	1.5481787	2.2021839	2.1525540	0.9627801	1.38917724
6	0.87892105	1.5068045	1.62619534	1.9303577	-0.3829885	0.2247056	1.4563573	1.38917724
7	-0.06220826	1.5068045	0.81403767	1.1182274	0.6633908	0.2247056	1.8950926	1.38917724
12	1.62642090	0.2618926	-0.04887986	0.7360484	-0.8754023	2.6061653	1.4563573	0.09222753
13	-0.06220826	0.2618926	-0.04887986	1.1182274	1.1558046	1.1886298	0.5240449	0.09222753

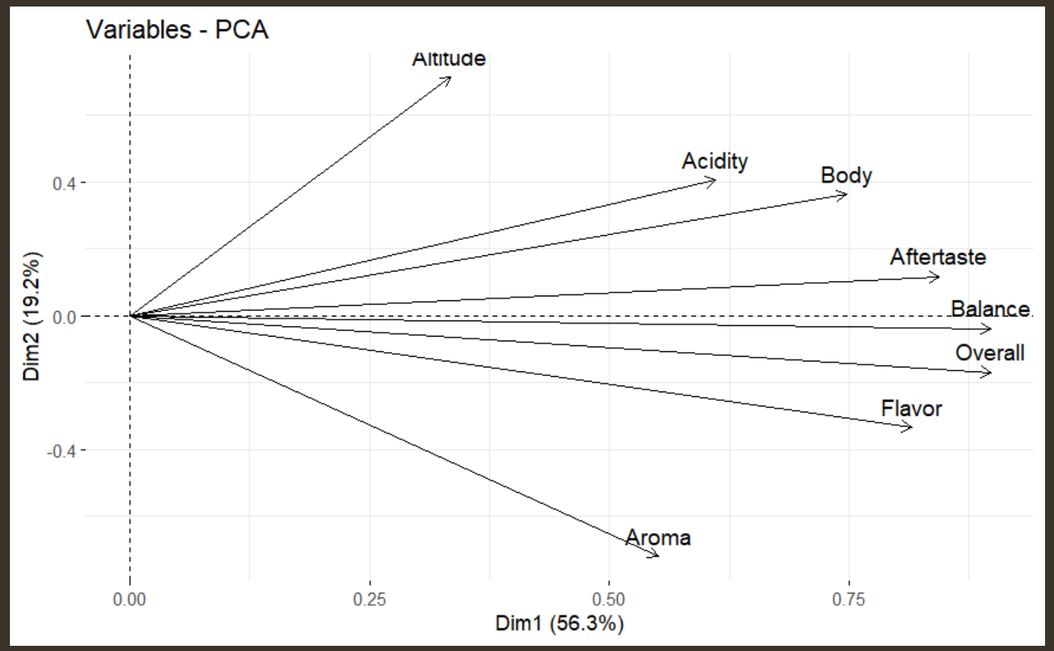
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	2.0825752	1.2160046	0.82784788	0.65797491	0.56393832	0.47485385	0.3699524	0.2987920
Proportion of Variance	0.5629914	0.1919424	0.08896138	0.05619777	0.04128228	0.02926984	0.0177661	0.0115888
Cumulative Proportion	0.5629914	0.7549338	0.84389521	0.90009298	0.94137526	0.97064510	0.9884112	1.0000000

Importance of components:

	Comp.1	Comp.2
Altitude	0.1607020	0.58678021
Aroma	0.2646891	-0.59213891
Flavor	0.3913269	-0.27442351
Aftertaste	0.4053864	0.09589532
Acidity	0.2934463	0.33350188
Body	0.3588003	0.29807711
Balance	0.4310813	-0.03174517
Overall	0.4312311	-0.13955321

PCA for Gesha





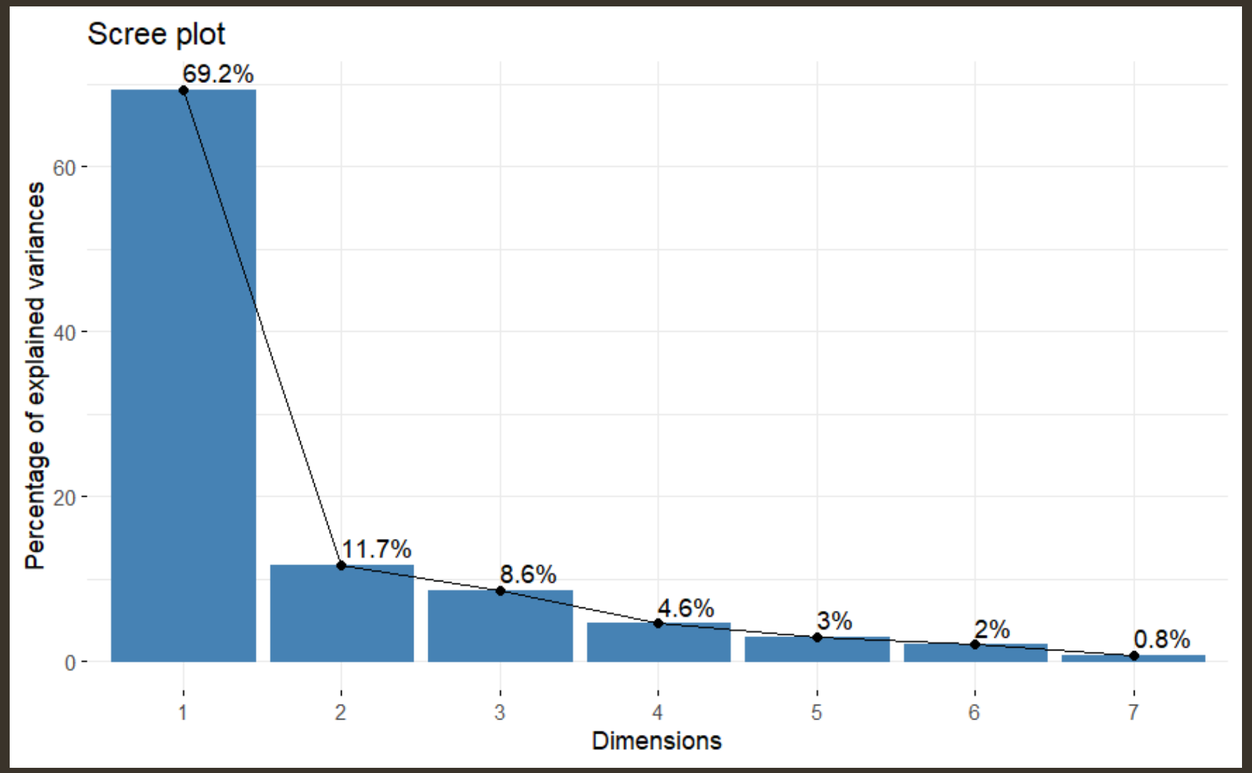
PCA for Caturra

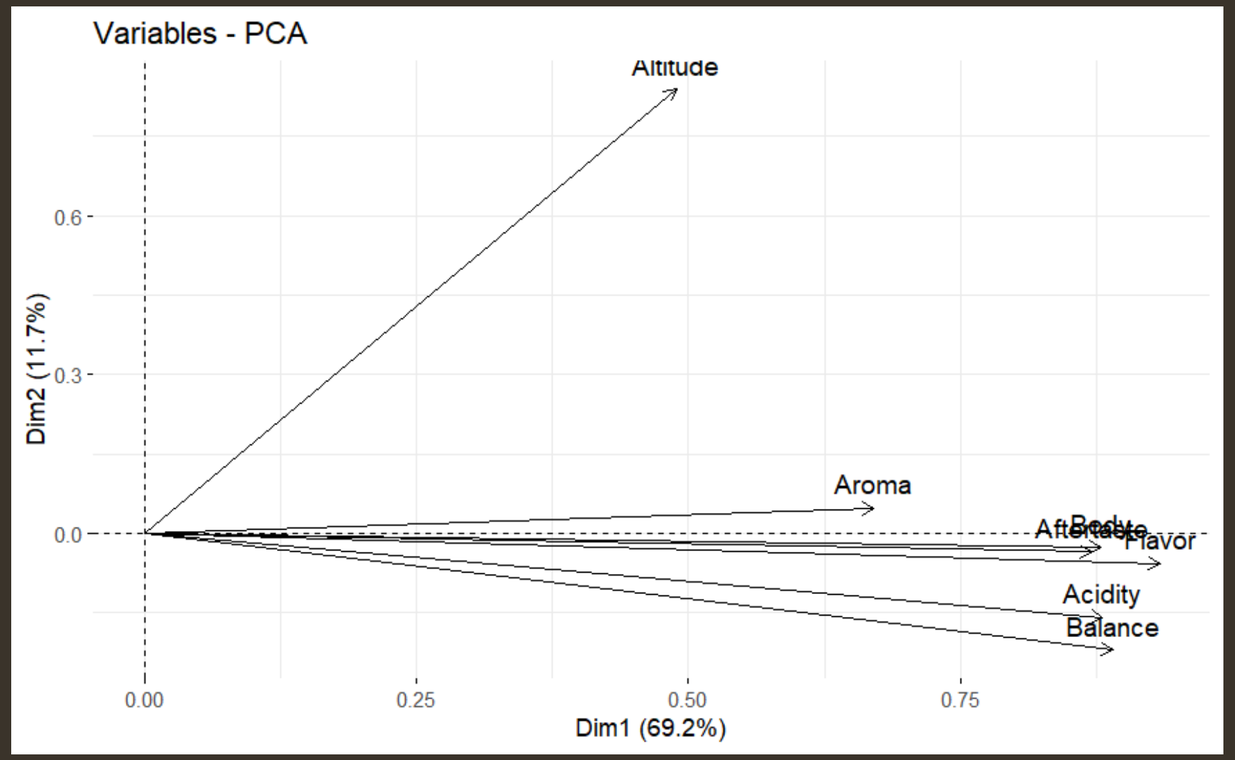
```
    Altitude      Aroma    Flavor Aftertaste   Acidity      Body   Balance
15 1.0855526  2.3039289 1.2813654 2.85948143 1.2875539 1.9362749 2.1774578
47 0.2597958  0.1664283 1.6372052 0.84746347 1.6610702 1.2829980 1.3339553
49 0.2597958  1.0005749 0.9255256 1.24005234 1.6610702 0.5888914 0.9370129
60 0.2597958  0.1664283 1.2813654 0.84746347 0.8673480 1.2829980 0.5400706
62 1.0855526 -0.3027792 0.5252059 0.01321213 0.1203154 0.9155298 0.5400706
84 0.2597958 -0.3027792 0.9255256 0.40580099 0.8673480 0.5888914 0.5400706
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Comp.7
Standard deviation     2.1588624 0.8864852 0.7600750 0.55842581 0.45115371 0.37085666
0.231513792
Proportion of Variance 0.6924449 0.1167558 0.0858318 0.04633042 0.03024018 0.02043372
0.007963226
Cumulative Proportion  0.6924449 0.8092007 0.8950324 0.94136287 0.97160305 0.99203677
1.000000000
               Comp.1      Comp.2
Altitude    0.2264452  0.94702384
Aroma       0.3106301  0.05391273
Flavor      0.4324520 -0.06451817
Aftertaste  0.4035070 -0.03929772
Acidity     0.4080186 -0.17856068
Body        0.4070658 -0.02957633
Balance     0.4125739 -0.24854312
```

PCA for Caturra

```
    Altitude      Aroma    Flavor Aftertaste   Acidity      Body   Balance
15 1.0855526  2.3039289 1.2813654 2.85948143 1.2875539 1.9362749 2.1774578
47 0.2597958  0.1664283 1.6372052 0.84746347 1.6610702 1.2829980 1.3339553
49 0.2597958  1.0005749 0.9255256 1.24005234 1.6610702 0.5888914 0.9370129
60 0.2597958  0.1664283 1.2813654 0.84746347 0.8673480 1.2829980 0.5400706
62 1.0855526 -0.3027792 0.5252059 0.01321213 0.1203154 0.9155298 0.5400706
84 0.2597958 -0.3027792 0.9255256 0.40580099 0.8673480 0.5888914 0.5400706
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Comp.7
Standard deviation     2.1588624 0.8864852 0.7600750 0.55842581 0.45115371 0.37085666
0.231513792
Proportion of Variance 0.6924449 0.1167558 0.0858318 0.04633042 0.03024018 0.02043372
0.007963226
Cumulative Proportion  0.6924449 0.8092007 0.8950324 0.94136287 0.97160305 0.99203677
1.000000000
               Comp.1      Comp.2
Altitude    0.2264452  0.94702384
Aroma       0.3106301  0.05391273
Flavor      0.4324520 -0.06451817
Aftertaste  0.4035070 -0.03929772
Acidity     0.4080186 -0.17856068
Body        0.4070658 -0.02957633
Balance     0.4125739 -0.24854312
```

PCA for Caturra





THANK YOU

30 MAY 2024