Coffee Ratings EDA

Esslam Mumen, Sultan Alkadhi, Asma AlQahtani, Raef Salem

2022-06-22

Raef Salem

```
#Introduction

##In this report, we aim to:

###1. Introduce our data set.
###2. Describe procedures we follow to clean the data.
###3. Describe the attribute in the data.
###4. Perform some exploratory data analysis

##In this EDA, we have used the data set that was originally scraped from the Coffee Quality Institute
```

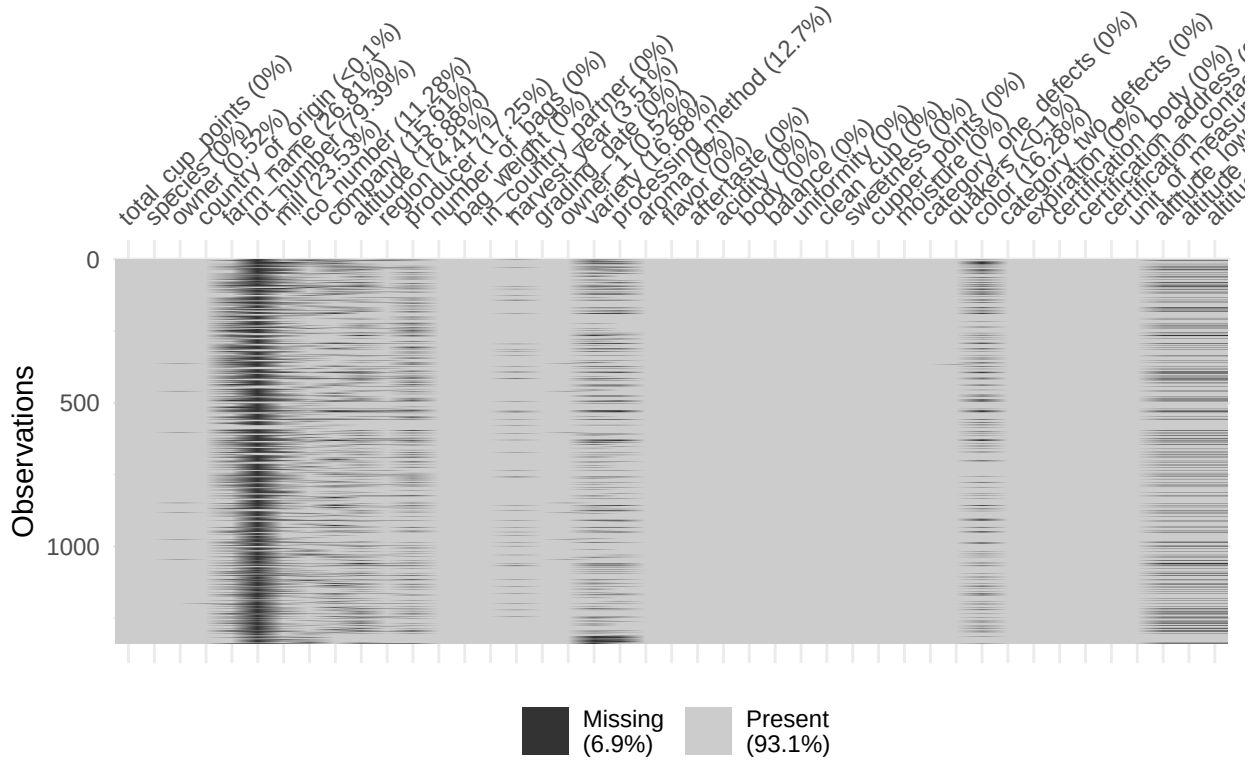
Checking for cleanliness and missing data

```
##  [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"
```

```
##      total_cup_points               species                 owner
##                     0                     0                     7
##     country_of_origin             farm_name           lot_number
##                     1                   359                  1063
##                  mill             ico_number               company
##                   315                   151                   209
##              altitude                region              producer
##                   226                    59                   231
##        number_of_bags            bag_weight    in_country_partner
##                     0                     0                     0
```

```
##          harvest_year          grading_date               owner_1
##                    47                     0                     7
##               variety     processing_method                 aroma
##                   226                   170                     0
##                flavor            aftertaste               acidity
##                     0                     0                     0
##                  body               balance            uniformity
##                     0                     0                     0
##             clean_cup             sweetness         cupper_points
##                     0                     0                     0
##              moisture  category_one_defects               quakers
##                     0                     0                     1
##                 color  category_two_defects            expiration
##                   218                     0                     0
##    certification_body certification_address certification_contact
##                     0                     0                     0
##   unit_of_measurement   altitude_low_meters  altitude_high_meters
##                     0                   230                   230
##  altitude_mean_meters
##                   230
```

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## Please use `gather()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```



Since 6.9% of the data is missing, some filling/replacing should be made throughout the EDA.

```
##  total_cup_points   species             owner           country_of_origin
##  Min.   : 0.00    Length:1339        Length:1339        Length:1339
##  1st Qu.:81.08    Class :character   Class :character   Class :character
##  Median :82.50    Mode  :character   Mode  :character   Mode  :character
##  Mean   :82.09
##  3rd Qu.:83.67
##  Max.   :90.58
##
##   farm_name          lot_number            mill            ico_number
##  Length:1339        Length:1339        Length:1339        Length:1339
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    company            altitude            region            producer
##  Length:1339        Length:1339        Length:1339        Length:1339
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  number_of_bags    bag_weight        in_country_partner harvest_year
##  Min.   :   0.0   Length:1339        Length:1339        Length:1339
##  1st Qu.:  14.0   Class :character   Class :character   Class :character
##  Median : 175.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 154.2
##  3rd Qu.: 275.0
##  Max.   :1062.0
##
##  grading_date         owner_1            variety          processing_method
##  Length:1339        Length:1339        Length:1339        Length:1339
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      aroma            flavor       aftertaste       acidity           body
##  Min.   :0.000   Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:7.420   1st Qu.:7.33   1st Qu.:7.250   1st Qu.:7.330   1st Qu.:7.330
##  Median :7.580   Median :7.58   Median :7.420   Median :7.580   Median :7.500
##  Mean   :7.567   Mean   :7.52   Mean   :7.401   Mean   :7.536   Mean   :7.517
##  3rd Qu.:7.750   3rd Qu.:7.75   3rd Qu.:7.580   3rd Qu.:7.750   3rd Qu.:7.670
##  Max.   :8.750   Max.   :8.83   Max.   :8.670   Max.   :8.750   Max.   :8.580
##
##     balance        uniformity       clean_cup        sweetness
##  Min.   :0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:7.330   1st Qu.:10.000   1st Qu.:10.000   1st Qu.:10.000
##  Median :7.500   Median :10.000   Median :10.000   Median :10.000
```

```
##  Mean   :7.518   Mean   : 9.835   Mean   : 9.835   Mean   : 9.857
##  3rd Qu.:7.750   3rd Qu.:10.000   3rd Qu.:10.000   3rd Qu.:10.000
##  Max.   :8.750   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##  cupper_points       moisture       category_one_defects    quakers
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.0000      Min.   : 0.0000
##  1st Qu.: 7.250   1st Qu.:0.09000   1st Qu.: 0.0000      1st Qu.: 0.0000
##  Median : 7.500   Median :0.11000   Median : 0.0000      Median : 0.0000
##  Mean   : 7.503   Mean   :0.08838   Mean   : 0.4795      Mean   : 0.1734
##  3rd Qu.: 7.750   3rd Qu.:0.12000   3rd Qu.: 0.0000      3rd Qu.: 0.0000
##  Max.   :10.000   Max.   :0.28000   Max.   :63.0000      Max.   :11.0000
##                                                          NA's   :1
##     color           category_two_defects  expiration        certification_body
##  Length:1339        Min.   : 0.000       Length:1339        Length:1339
##  Class :character   1st Qu.: 0.000       Class :character   Class :character
##  Mode  :character   Median : 2.000       Mode  :character   Mode  :character
##                     Mean   : 3.556
##                     3rd Qu.: 4.000
##                     Max.   :55.000
##
##  certification_address certification_contact unit_of_measurement
##  Length:1339           Length:1339           Length:1339
##  Class :character      Class :character      Class :character
##  Mode  :character      Mode  :character      Mode  :character
##
##
##
##
##  altitude_low_meters altitude_high_meters altitude_mean_meters
##  Min.   :     1      Min.   :     1       Min.   :     1
##  1st Qu.:  1100      1st Qu.:  1100       1st Qu.:  1100
##  Median :  1311      Median :  1350       Median :  1311
##  Mean   :  1751      Mean   :  1799       Mean   :  1775
##  3rd Qu.:  1600      3rd Qu.:  1650       3rd Qu.:  1600
##  Max.   :190164      Max.   :190164       Max.   :190164
##  NA's   :230         NA's   :230          NA's   :230

## Rows: 1,339
## Columns: 43
## $ total_cup_points      <dbl> 90.58, 89.92, 89.75, 89.00, 88.83, 88.83, 88.75,~
## $ species               <chr> "Arabica", "Arabica", "Arabica", "Arabica", "Ara~
## $ owner                 <chr> "metad plc", "metad plc", "grounds for health ad~
## $ country_of_origin     <chr> "Ethiopia", "Ethiopia", "Guatemala", "Ethiopia",~
## $ farm_name             <chr> "metad plc", "metad plc", "san marcos barrancas ~
## $ lot_number            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ mill                  <chr> "metad plc", "metad plc", NA, "wolensu", "metad ~
## $ ico_number            <chr> "2014/2015", "2014/2015", NA, NA, "2014/2015", N~
## $ company               <chr> "metad agricultural developmet plc", "metad agri~
## $ altitude              <chr> "1950-2200", "1950-2200", "1600 - 1800 m", "1800~
## $ region                <chr> "guji-hambela", "guji-hambela", NA, "oromia", "g~
## $ producer              <chr> "METAD PLC", "METAD PLC", NA, "Yidnekachew Dabes~
## $ number_of_bags        <dbl> 300, 300, 5, 320, 300, 100, 100, 300, 300, 50, 3~
## $ bag_weight            <chr> "60 kg", "60 kg", "1", "60 kg", "60 kg", "30 kg"~
## $ in_country_partner    <chr> "METAD Agricultural Development plc", "METAD Agr~
```

```
## $ harvest_year          <chr> "2014", "2014", NA, "2014", "2014", "2013", "201~
## $ grading_date          <chr> "April 4th, 2015", "April 4th, 2015", "May 31st,~
## $ owner_1               <chr> "metad plc", "metad plc", "Grounds for Health Ad~
## $ variety               <chr> NA, "Other", "Bourbon", NA, "Other", NA, "Other"~
## $ processing_method     <chr> "Washed / Wet", "Washed / Wet", NA, "Natural / D~
## $ aroma                 <dbl> 8.67, 8.75, 8.42, 8.17, 8.25, 8.58, 8.42, 8.25, ~
## $ flavor                <dbl> 8.83, 8.67, 8.50, 8.58, 8.50, 8.42, 8.50, 8.33, ~
## $ aftertaste            <dbl> 8.67, 8.50, 8.42, 8.42, 8.25, 8.42, 8.33, 8.50, ~
## $ acidity               <dbl> 8.75, 8.58, 8.42, 8.42, 8.50, 8.50, 8.50, 8.42, ~
## $ body                  <dbl> 8.50, 8.42, 8.33, 8.50, 8.42, 8.25, 8.25, 8.33, ~
## $ balance               <dbl> 8.42, 8.42, 8.42, 8.25, 8.33, 8.33, 8.25, 8.50, ~
## $ uniformity            <dbl> 10.00, 10.00, 10.00, 10.00, 10.00, 10.00, 10.00,~
## $ clean_cup             <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, ~
## $ sweetness             <dbl> 10.00, 10.00, 10.00, 10.00, 10.00, 10.00, 10.00,~
## $ cupper_points         <dbl> 8.75, 8.58, 9.25, 8.67, 8.58, 8.33, 8.50, 9.00, ~
## $ moisture              <dbl> 0.12, 0.12, 0.00, 0.11, 0.12, 0.11, 0.11, 0.03, ~
## $ category_one_defects  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ quakers               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ color                 <chr> "Green", "Green", NA, "Green", "Green", "Bluish-~
## $ category_two_defects  <dbl> 0, 1, 0, 2, 2, 1, 0, 0, 0, 4, 1, 0, 0, 2, 2, 0, ~
## $ expiration            <chr> "April 3rd, 2016", "April 3rd, 2016", "May 31st,~
## $ certification_body    <chr> "METAD Agricultural Development plc", "METAD Agr~
## $ certification_address <chr> "309fcf77415a3661ae83e027f7e5f05dad786e44", "309~
## $ certification_contact <chr> "19fef5a731de2db57d16da10287413f5f99bc2dd", "19f~
## $ unit_of_measurement   <chr> "m", "m", "m", "m", "m", "m", "m", "m", "m", "m"~
## $ altitude_low_meters   <dbl> 1950.0, 1950.0, 1600.0, 1800.0, 1950.0, NA, NA, ~
## $ altitude_high_meters  <dbl> 2200.0, 2200.0, 1800.0, 2200.0, 2200.0, NA, NA, ~
## $ altitude_mean_meters  <dbl> 2075.0, 2075.0, 1700.0, 2000.0, 2075.0, NA, NA, ~
```

Checking the stats for the total cup points

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   81.08   82.50   82.09   83.67   90.58
```

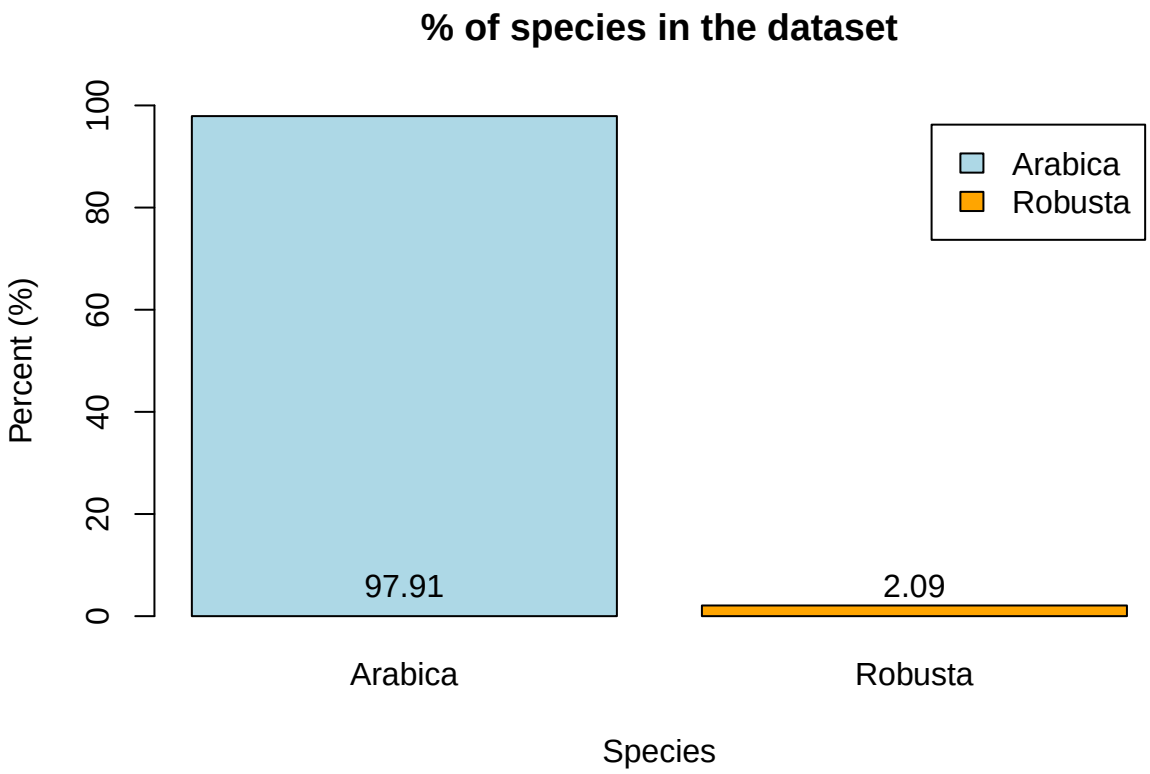
Checking for extreme values “top 10 outliers” in total cup points

```
##  [1]  0.00 59.83 63.08 67.92 68.33 69.17 69.17 69.33 70.67 70.75
```

Checking for the proportion of coffee species

```
##
## Arabica Robusta
##    1311      28
```

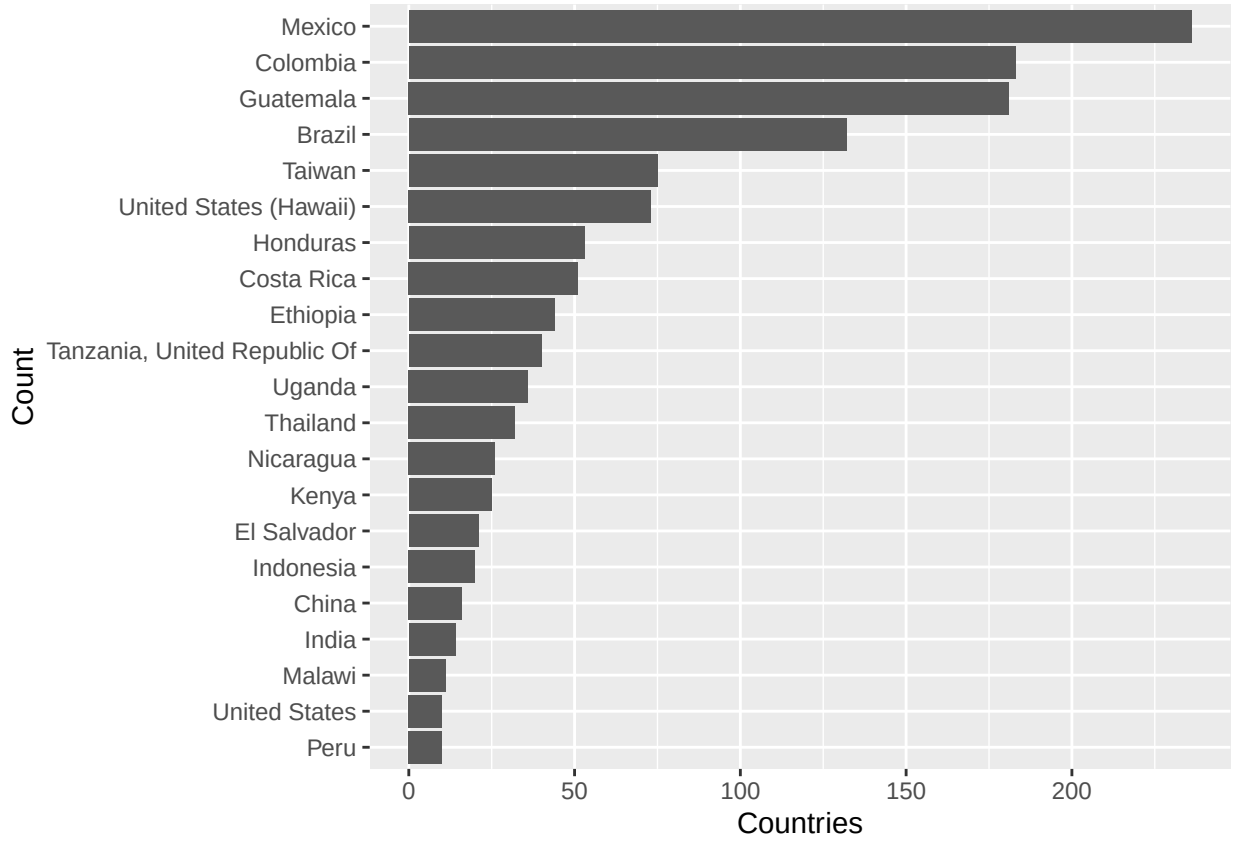
```
##
##    Arabica    Robusta
## 0.97908887 0.02091113
```



Therefore, it seems like Arabica species is much more popular than Robusta coffee

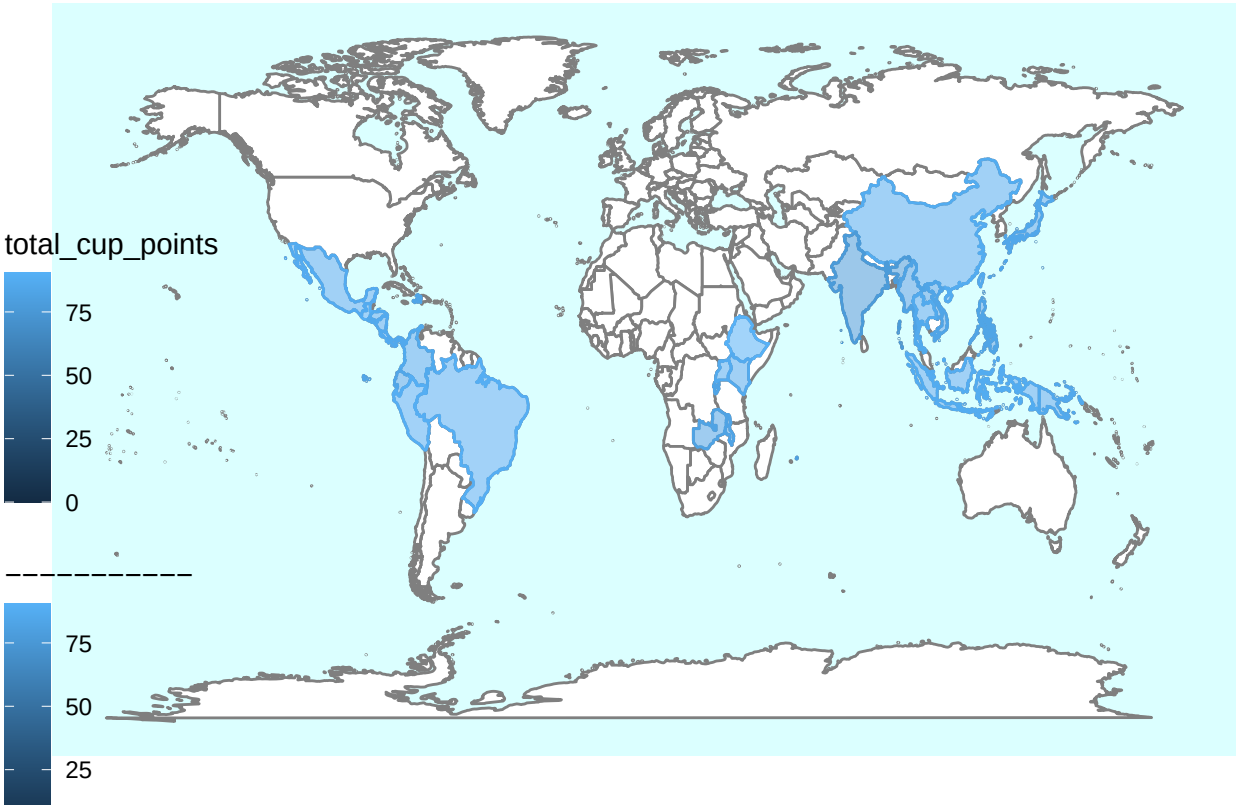
Countries Represented

```
## Selecting by n
```



Creating the world map. This is to visualize the major producers of coffee that are included in the dataset.

```
## Warning: Ignoring unknown aesthetics: x, y
```



We can see here that coffee grows mainly in tropical climates

Checking for extreme values

```
##  [1] 190164 190164 110000  11000   4287   4001   3850   3845   3825   3800
```

Coffee can survive and grow in altitudes as high as 190000 feet and as low as 3800 feet.

Here, we're going to analyze and compare the countries of origin within the scope of total cup points.

To make it more organized, let's sort this in descending order based on the "avg" column

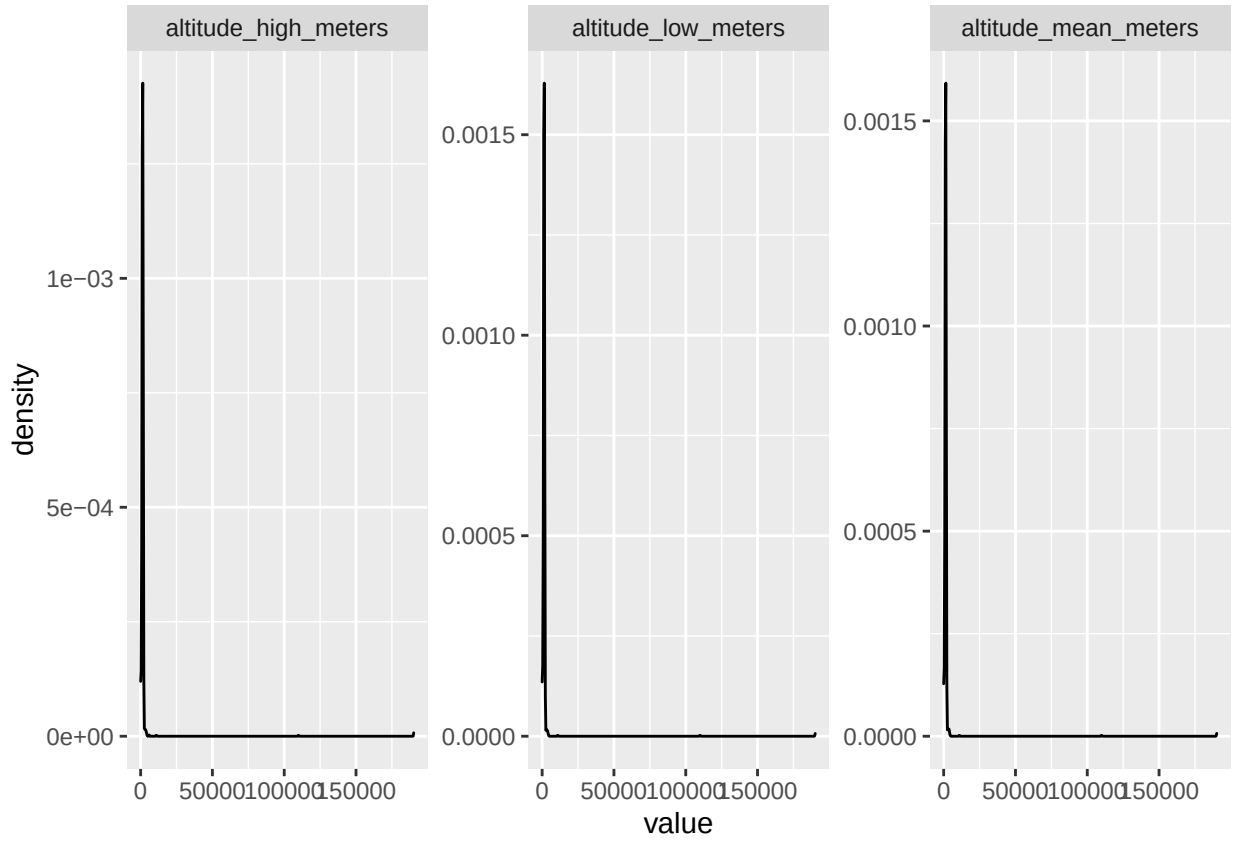
```
##                country_of_origin      avg      stdev   min   max range
## 24             Papua New Guinea 85.75000         NA 85.75 85.75  0.00
## 9                     Ethiopia 85.48409  2.3373542 80.00 90.58 10.58
## 15                       Japan 84.67000         NA 84.67 84.67  0.00
## 32                United States 84.43300  4.0108909 77.17 87.92 10.75
## 16                       Kenya 84.30960  1.4597331 79.83 86.25  6.42
## 23                      Panama 83.70750  2.4070504 80.75 85.83  5.08
## 31                      Uganda 83.45194  1.6135742 80.50 86.83  6.33
## 4                     Colombia 83.10656  1.4132733 72.83 86.00 13.17
## 8                  El Salvador 83.05286  1.5960299 79.67 85.58  5.91
## 3                        China 82.92750  2.3344107 78.00 87.25  9.25
## 27                      Rwanda 82.83000         NA 82.83 82.83  0.00
## 5                   Costa Rica 82.78902  2.3828833 71.75 87.17 15.42
## 30                    Thailand 82.57375  1.5463688 79.67 86.17  6.50
## 14                  Indonesia 82.56550  2.4171067 76.17 87.42 11.25
## 25                       Peru 82.52600  3.4431936 77.00 88.75 11.75
```

```
## 1                         Brazil 82.40591  2.1329018 70.67 88.83 18.16
## 29 Tanzania, United Republic Of 82.36950  1.2821195 80.33 86.50  6.17
## 28                        Taiwan 82.00133  1.7715029 77.67 86.58  8.91
## 36                        Zambia 81.92000         NA 81.92 81.92  0.00
## 10                     Guatemala 81.84657  2.8851687 59.83 89.75 29.92
## 17                          Laos 81.83333  0.9312536 80.83 82.67  1.84
## 2                         Burundi 81.83000  2.1213203 80.33 83.33  3.00
## 33       United States (Hawaii) 81.82041  2.9897568 72.58 87.92 15.34
## 34  United States (Puerto Rico) 81.72750  2.0927076 79.08 83.75  4.67
## 18                         Malawi 81.71182  0.5931917 80.67 82.75  2.08
## 35                        Vietnam 81.20875  3.1707567 73.75 83.17  9.42
## 13                          India 81.08286  2.4178835 75.08 83.50  8.42
## 20                         Mexico 80.89008  2.7424133 68.33 87.17 18.84
## 26                    Philippines 80.83400  2.9198339 75.83 82.92  7.09
## 21                        Myanmar 80.75000  0.6417387 80.00 81.83  1.83
## 19                      Mauritius 80.50000         NA 80.50 80.50  0.00
## 22                      Nicaragua 80.45808  4.0807450 63.08 86.58 23.50
## 7                         Ecuador 80.22000  3.1442487 78.08 83.83  5.75
## 12                       Honduras 79.35755 11.6549567  0.00 86.67 86.67
## 6                  Cote d?Ivoire 79.33000         NA 79.33 79.33  0.00
## 37                           <NA> 79.08000         NA 79.08 79.08  0.00
## 11                         Haiti 77.18000  5.9750280 67.92 83.33 15.41
```

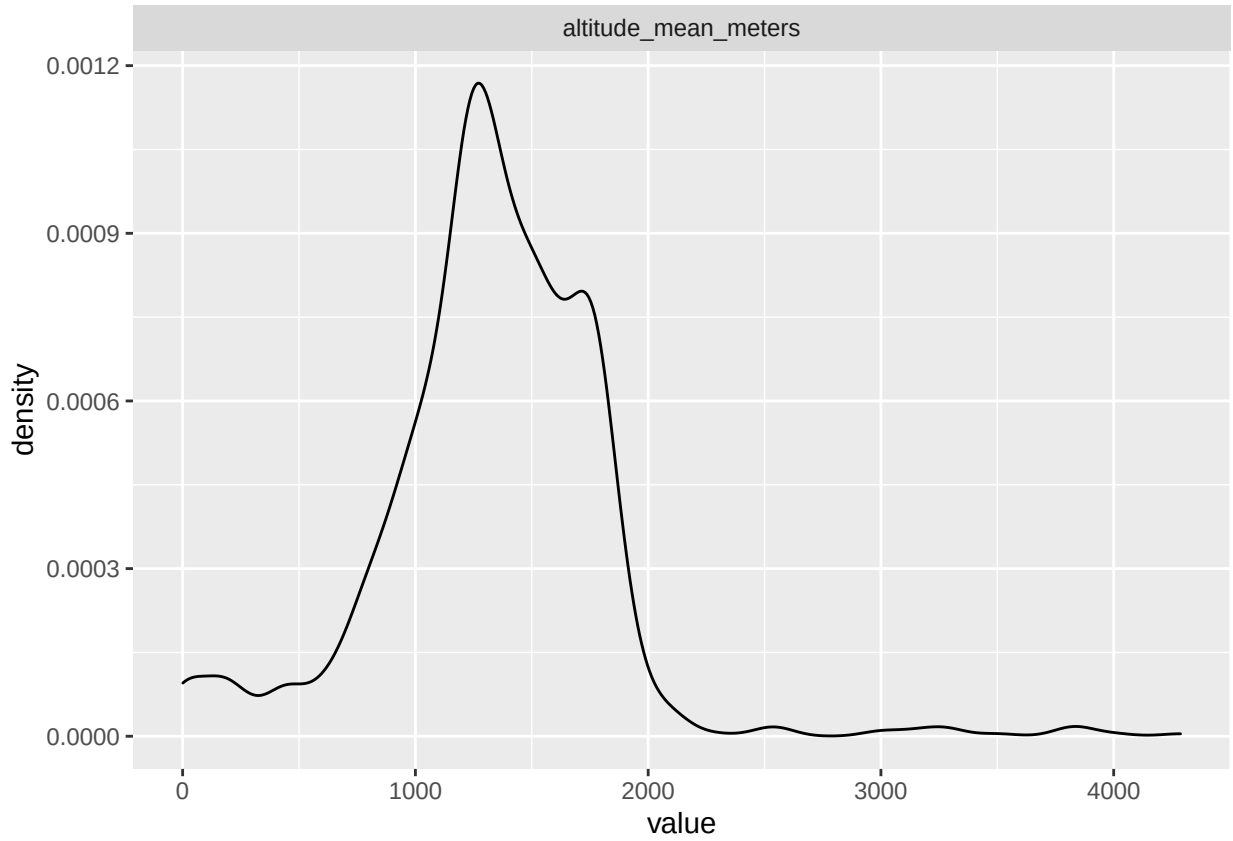
```
## [1] 85.75
```

#Papa New Guinea has the highest average total cup points with a score of 85.75.

Sultan Alkadhi Checking for altitude distributions

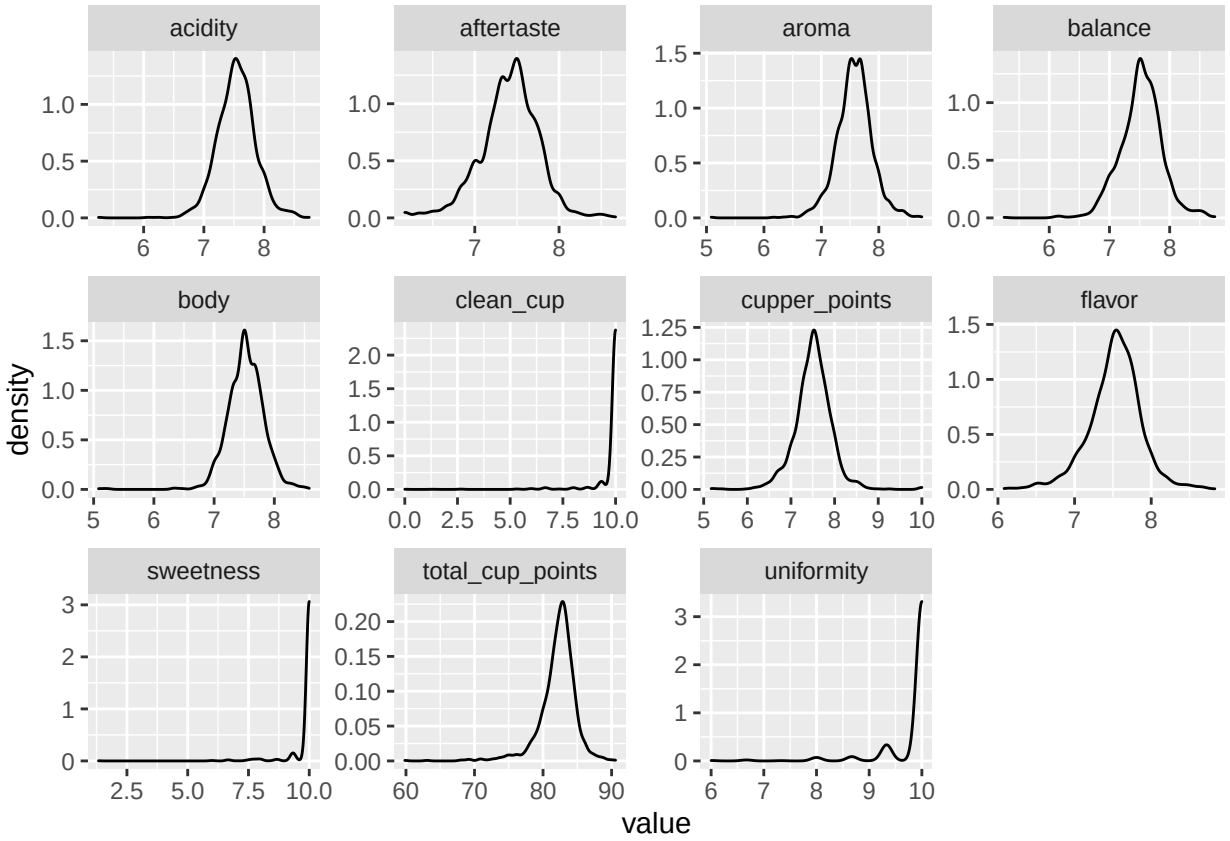


After removing erroneous values



Checking the distribution of grades

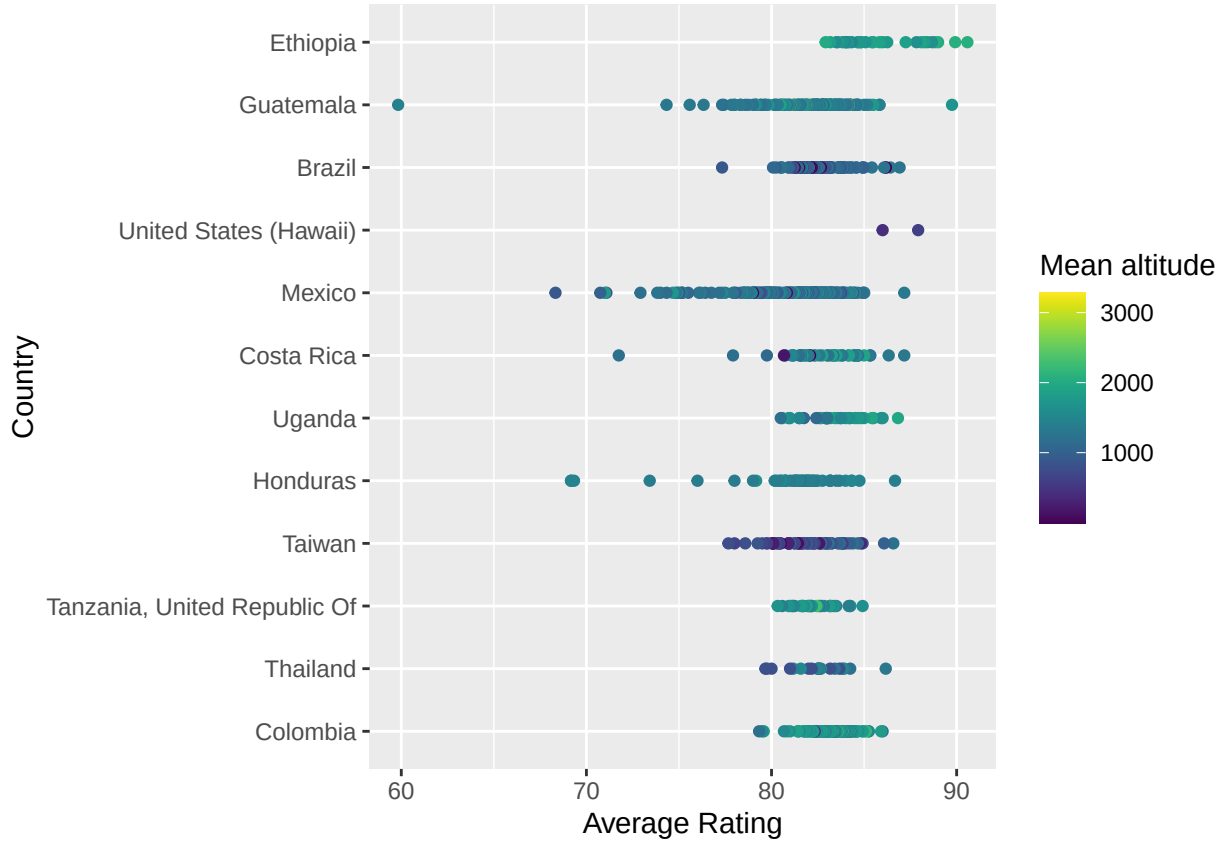
```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(grades)‘ instead of ‘grades‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```



It's easy to score high on clean cup, uniformity, and sweetness? "Uniformity refers to consistency of flavor of the different cups of the sample tasted." "Clean Cup refers to a lack of interfering negative impressions from first ingestion to final aftertaste" "Sweetness" no longer part of the CQI scale? "Acidity is often described as"brightness" when favorable or "sour" when unfavorable. At its best, acidity contributes to a coffee's liveliness, sweetness, and fresh- fruit character and is almost immediately experienced and evaluated when the coffee is first slurped into the mouth"

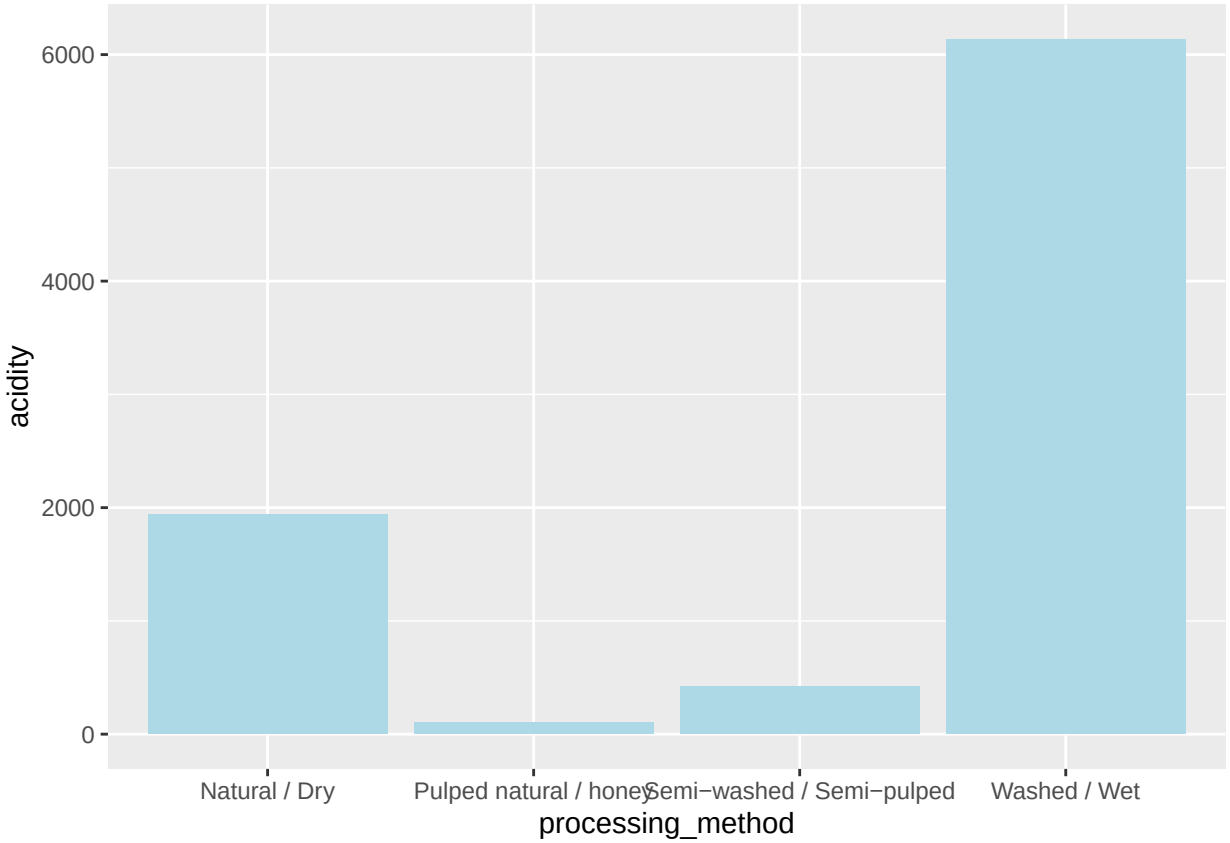
```
## Selecting by n
```

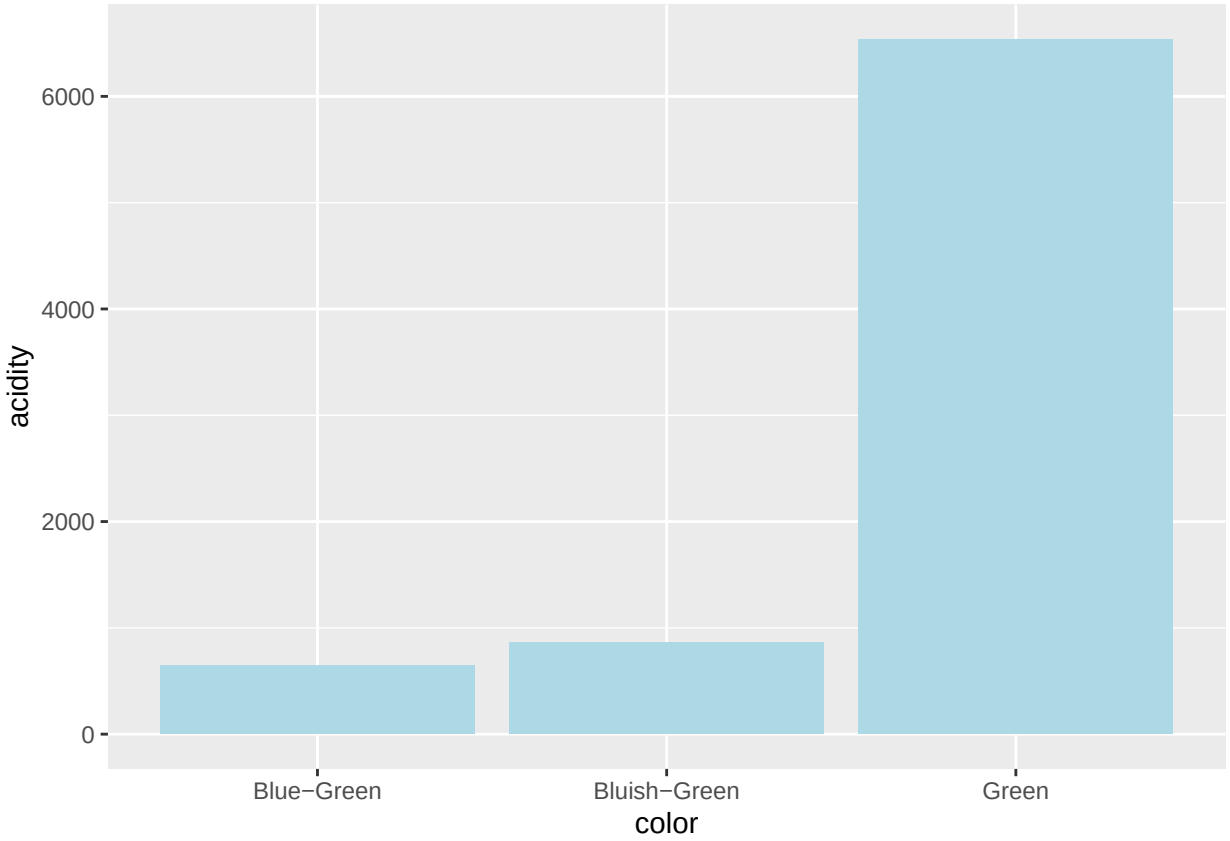
Order countries by rating and highlight altitude

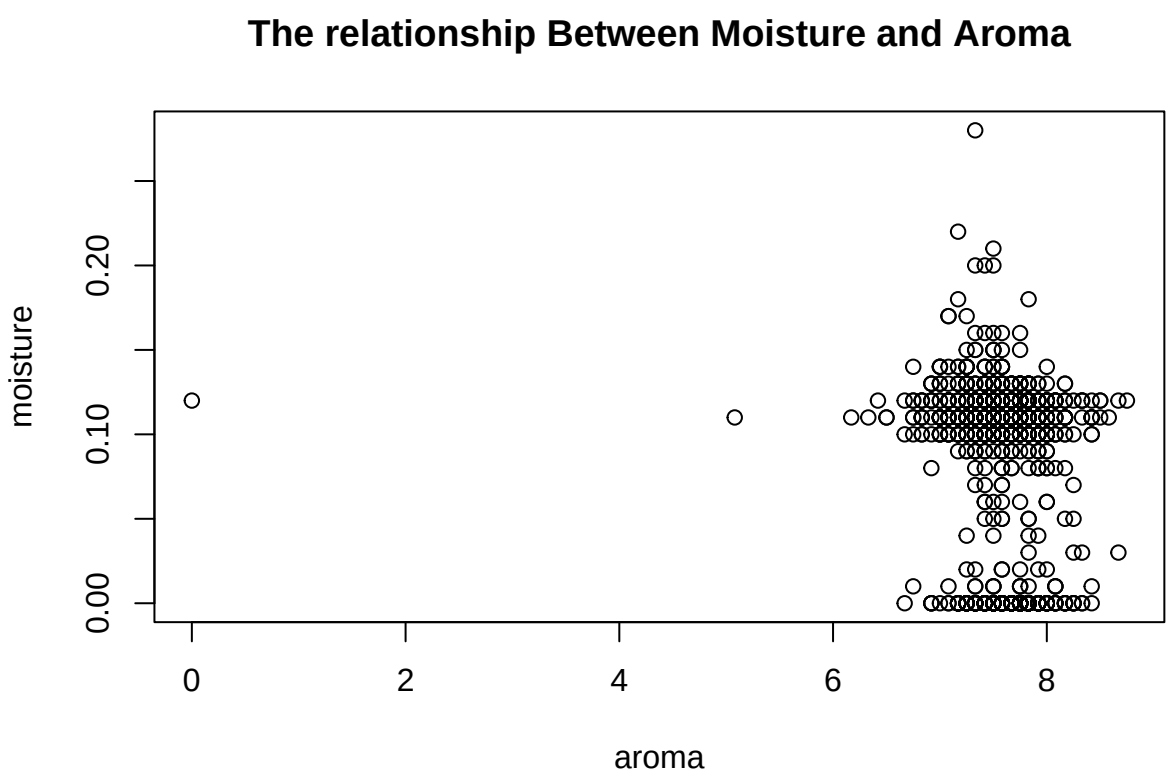


Model total score vs different grades

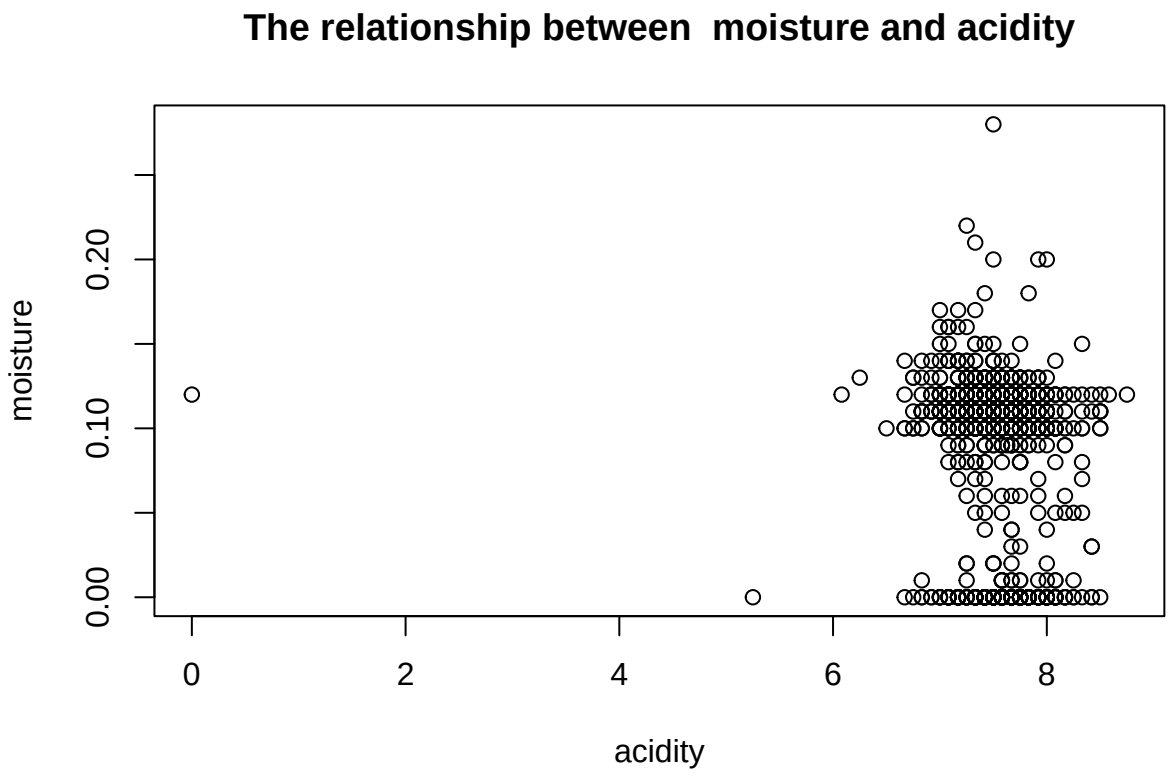
Asma AlQahtani


6000
4000
2000
0
acidity
Natural / Dry
Pulped natural / honey
Semi−washed / Semi−pulped
Washed / Wet
processing_method

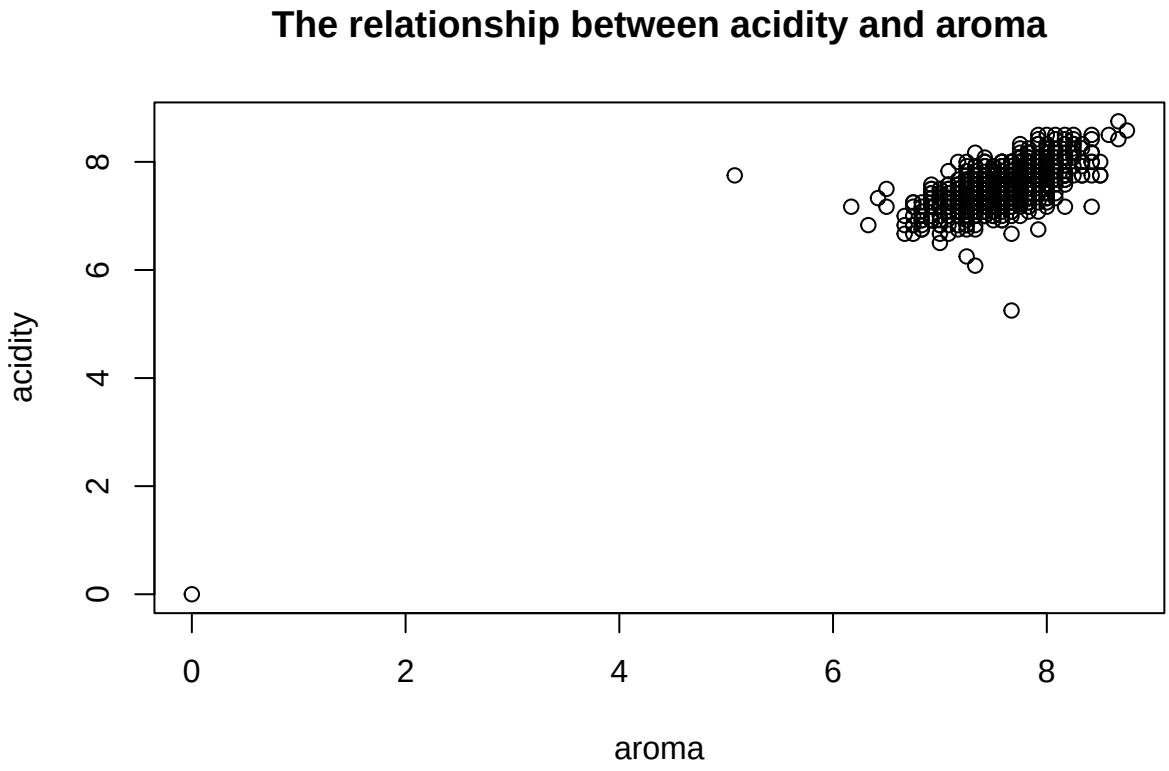

6000
4000
2000
0
acidity
Blue−Green
Bluish−Green
Green
color


The relationship Between Moisture and Aroma
moisture
0.00
0.10
0.20
0
2
4
6
8
aroma

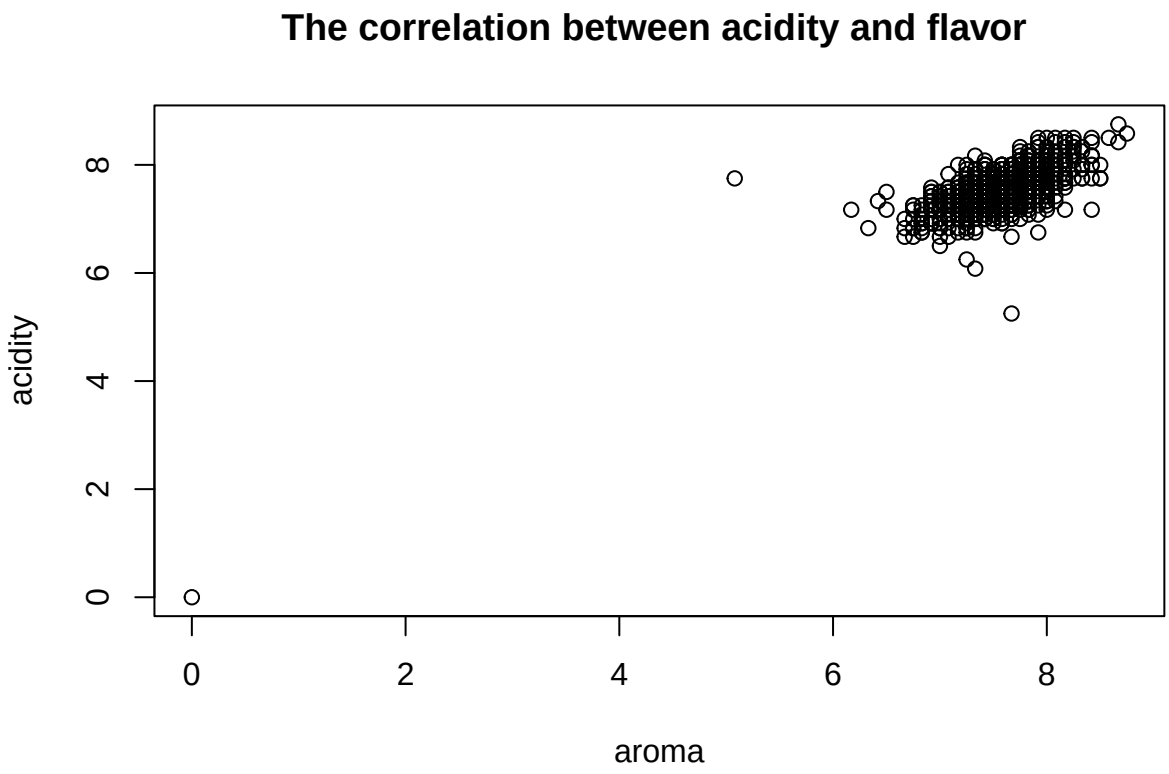
```
## [1] -0.1341272
```


The relationship between moisture and acidity
moisture
0.00
0.10
0.20
0
2
4
6
8
acidity

```
## [1] -0.1290795
```


The relationship between acidity and aroma
acidity
0
2
4
6
8
aroma
0
2
4
6
8

```
## [1] 0.7235822
```




```
## [1] 0.8166883
```

Raef Salem

```
#Conclusion

##So far, we have discovered that:

###1. Coffee is mainly grown in tropical regions
###2. Mexico contains the highest amount of data in the data set with more than 236 coffee cups.
###3. The highest score given to a cup of coffee is 90.58, and the lowest is 0.
###4. Arabica is a more popular species than Robusta, with its coffee samples taking over 98% of the da
###5. Coffee can grow in extremely high altitudes.
###6. Papa New Guinea has the highest average total cup points with a score of 85.75.
###Flavor is the highest contributor to the overall score.
###The highest scoring country (Ethiopia) has the highest altitude.
```

```
#Thank You!

##Q&A
```