COFFEE QUALITY

GENERAL OVERVIEW

Data and Goal

Data Cleaning and Pre-processing

Exploratory Data Analysis

Unsupervised Learning

Supervised Learning

```
'data.frame':   1311 obs. of  44 variables:
 $ X                    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Species              : chr  "Arabica" "Arabica" "Arabica" "Arabica" ...
 $ Owner                : chr  "metad plc" "metad plc" "grounds for health admin" "yidnekachew
dabessa" ...
 $ Country.of.Origin    : chr  "Ethiopia" "Ethiopia" "Guatemala" "Ethiopia" ...
 $ Farm.Name            : chr  "metad plc" "metad plc" "san marcos barrancas \"san cristobal
cuch" "yidnekachew dabessa coffee plantation" ...
 $ Lot.Number           : chr  "" "" "" "" ...
 $ Mill                 : chr  "metad plc" "metad plc" "" "wolensu" ...
 $ ICO.Number           : chr  "2014/2015" "2014/2015" "" "" ...
 $ Company              : chr  "metad agricultural developmet plc" "metad agricultural
developmet plc" "" "yidnekachew debessa coffee plantation" ...
 $ Altitude             : chr  "1950-2200" "1950-2200" "1600 - 1800 m" "1800-2200" ...
 $ Region               : chr  "guji-hambela" "guji-hambela" "" "oromia" ...
 $ Producer             : chr  "METAD PLC" "METAD PLC" "" "Yidnekachew Dabessa Coffee
Plantation" ...
 $ Number.of.Bags       : int  300 300 5 320 300 100 100 300 300 50 ...
 $ Bag.Weight           : chr  "60 kg" "60 kg" "1" "60 kg" ...
 $ In.Country.Partner   : chr  "METAD Agricultural Development plc" "METAD Agricultural
Development plc" "Specialty Coffee Association" "METAD Agricultural Development plc" ...
 $ Harvest.Year         : chr  "2014" "2014" "" "2014" ...
 $ Grading.Date         : chr  "April 4th, 2015" "April 4th, 2015" "May 31st, 2010" "March 26th,
2015" ...
 $ Owner.1              : chr  "metad plc" "metad plc" "Grounds for Health Admin" "Yidnekachew
Dabessa" ...
 $ Variety              : chr  "" "Other" "Bourbon" "" ...
 $ Processing.Method    : chr  "Washed / Wet" "Washed / Wet" "" "Natural / Dry" ...
 $ Aroma                : num  8.67 8.75 8.42 8.17 8.25 8.58 8.42 8.25 8.67 8.08 ...
 $ Flavor               : num  8.83 8.67 8.5 8.58 8.5 8.42 8.5 8.33 8.67 8.58 ...
 $ Aftertaste           : num  8.67 8.5 8.42 8.42 8.25 8.42 8.33 8.5 8.58 8.5 ...
 $ Acidity              : num  8.75 8.58 8.42 8.42 8.5 8.5 8.5 8.42 8.42 8.5 ...
 $ Body                 : num  8.5 8.42 8.33 8.5 8.42 8.25 8.25 8.33 8.33 7.67 ...
 $ Balance              : num  8.42 8.42 8.42 8.25 8.33 8.33 8.25 8.5 8.42 8.42 ...
 $ Uniformity           : num  10 10 10 10 10 10 10 10 9.33 10 ...
 $ Clean.Cup            : num  10 10 10 10 10 10 10 10 10 10 ...
 $ Sweetness            : num  10 10 10 10 10 10 10 9.33 9.33 10 ...
 $ Cupper.Points        : num  8.75 8.58 9.25 8.67 8.58 8.33 8.5 9 8.67 8.5 ...
 $ Total.Cup.Points     : num  90.6 89.9 89.8 89 88.8 ...
 $ Moisture             : num  0.12 0.12 0 0.11 0.12 0.11 0.11 0.03 0.03 0.1 ...
 $ Category.One.Defects : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Quakers              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Color                : chr  "Green" "Green" "" "Green" ...
 $ Category.Two.Defects : int  0 1 0 2 2 1 0 0 0 4 ...
 $ Expiration           : chr  "April 3rd, 2016" "April 3rd, 2016" "May 31st, 2011" "March 25th,
2016" ...
 $ Certification.Body   : chr  "METAD Agricultural Development plc" "METAD Agricultural
Development plc" "Specialty Coffee Association" "METAD Agricultural Development plc" ...
 $ Certification.Address: chr  "309fcf77415a3661ae83e027f7e5f05dad786e44"
"309fcf77415a3661ae83e027f7e5f05dad786e44" "36d0d00a3724338ba7937c52a378d085f2172daa"
"309fcf77415a3661ae83e027f7e5f05dad786e44" ...
 $ Certification.Contact: chr  "19fef5a731de2db57d16da10287413f5f99bc2dd"
"19fef5a731de2db57d16da10287413f5f99bc2dd" "0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660"
"19fef5a731de2db57d16da10287413f5f99bc2dd" ...
 $ unit_of_measurement  : chr  "m" "m" "m" "m" ...
 $ altitude_low_meters  : num  1950 1950 1600 1800 1950 ...
 $ altitude_high_meters : num  2200 2200 1800 2200 2200 NA NA 1700 1700 1850 ...
 $ altitude_mean_meters : num  2075 2075 1700 2000 2075 ...
```

DATA AND GOAL

The dataset was directly downloaded from James LeDoux's git hub repository which can be accessed using the following link https://github.com/jldbc/coffee-quality-database/blob/master/data/arabica_data_cleaned.csv.

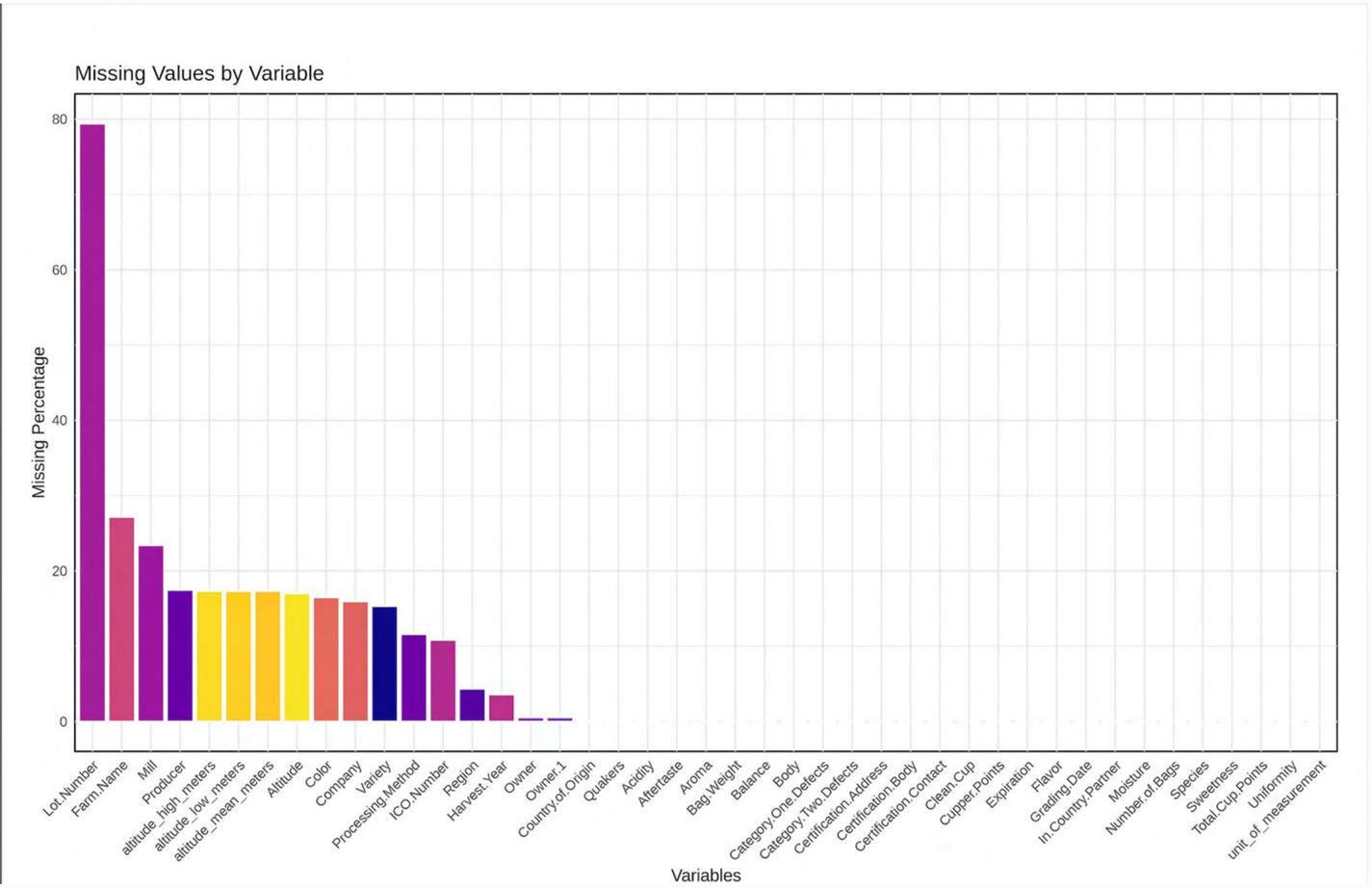
Data has 1311 observation with 44 variables, out of which 19 are numerical and 24 are categorical with records from harvest year of 2010 to 2018.

Apply statistical learning and machine learning methodologies to

1. Use and sensory & non-sensory data to find groupings
2. Understand and predict coffee quality beans
3. Determine factors that influence coffee quality

Application:

To deliver insights for growers, traders and coffee consumers on which features are important in dertermining in the quality of coffee before making decision on deciding how to produce, process or what to mainly evaluate before making purchases.



Dataset has missing values especially Lot_Number, Farm_Name, Mill and Producer.

Variables altitude_low_meter, altitude_high_meters and altitude_mean_meters are all derived from Altitude

Variable Owner.1 is derived from Owner

Based on additional preliminary inspection of the data set we will be cleaning the variables that have discrepancies and subsetting the variable we need for our analysis.

Process 1

- Simple cleaning up of data(index dropping, empty spaces and duplicates in the data)

Process 2

- Creating a new variable called Grade based on the Total.Cup.Points and as defined by the SCA

Process 3

- Removing the value 0 in Total.Cup.Points

Process 4

- Convert the variables stored in characters to factors

Process 5

- Handling all the discrepancies in the altitude_mean_meters variable data entry based unit measurement and different entry methods used

Process 6

- Creating a new variable that combines Bag.Weight and Number.of.Bags to create Total.Weight of samples

Process 7

- Dropped an non-plausible value for altitude.

Process 8

- Subsetted the data with the variables we need for analysis and excluded missing values from the analysis

We identify two different types of variable which we refer to as objective parameters (non-sensory data) and subjective parameters (sensory data)

Objective parameters: Category.One.Defects, Category.Two Defects, Quakers, Moisture, altitude_mean_meters_new, Total.Weight, Variety, Color, Processing.Method, Country.of.Origin

Subject parameters: Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean.Cup, Sweetness, Cupper.Points which determine Total.Cup.Points

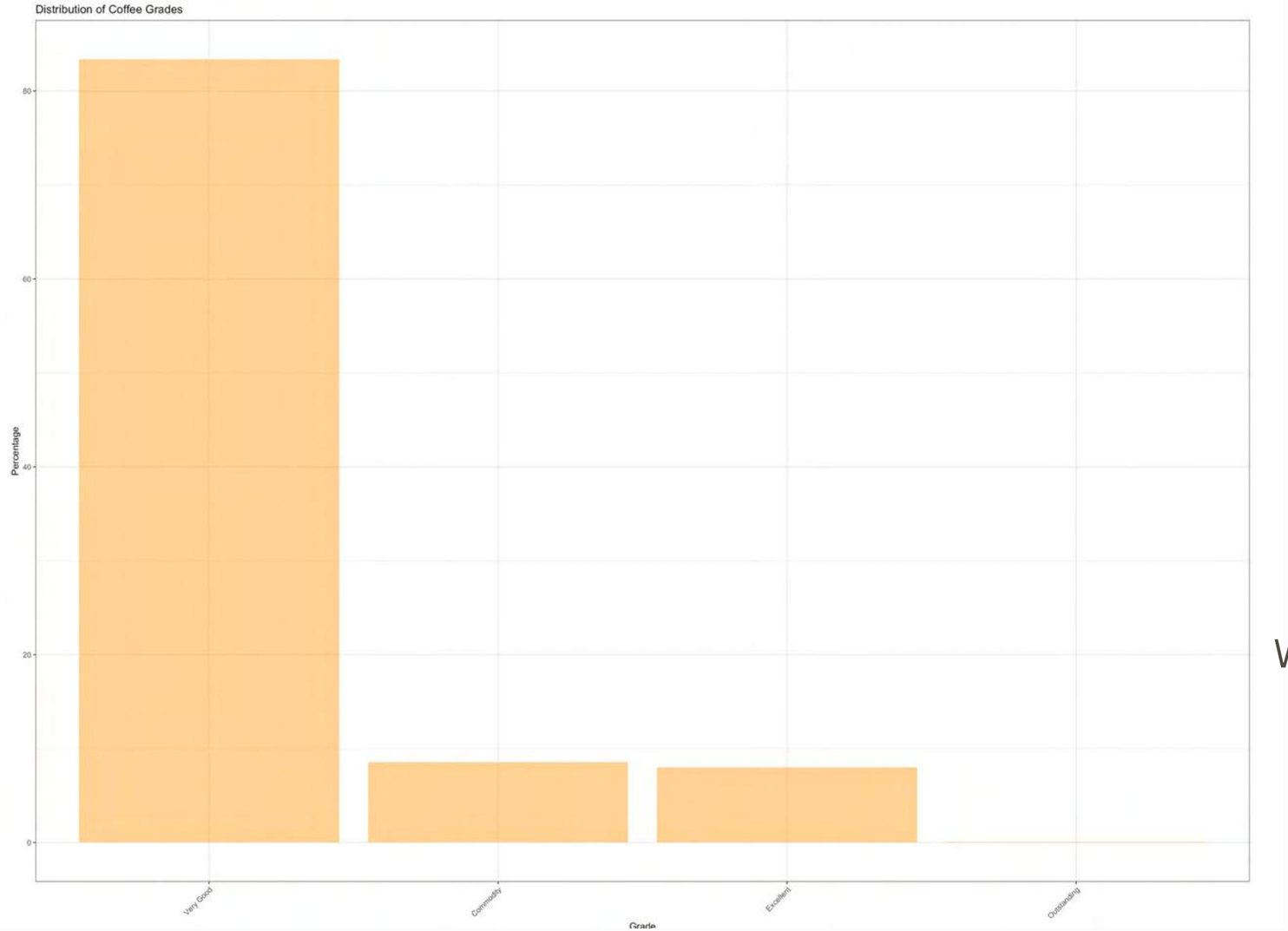
```
'data.frame':   1078 obs. of  23 variables:
 $ Grade                     : Factor w/ 3 levels "Commodity","Very Good",..: 3 3 3 3 3 3 3
 $ Total.Cup.Points          : num  89.9 89.8 89 88.8 88.7 ...
 $ Aroma                     : num  8.75 8.42 8.17 8.25 8.25 8.67 8.08 8.17 8.25 8.08 ...
 $ Flavor                    : num  8.67 8.5 8.58 8.5 8.33 8.67 8.58 8.67 8.42 8.67 ...
 $ Aftertaste                : num  8.5 8.42 8.42 8.25 8.5 8.58 8.5 8.25 8.17 8.33 ...
 $ Acidity                   : num  8.58 8.42 8.42 8.5 8.42 8.42 8.5 8.5 8.33 8.42 ...
 $ Body                      : num  8.42 8.33 8.5 8.42 8.33 8.33 7.67 7.75 8.08 8 ...
 $ Balance                   : num  8.42 8.42 8.25 8.33 8.5 8.42 8.42 8.17 8.17 8.08 ...
 $ Uniformity                : num  10 10 10 10 10 9.33 10 10 10 10 ...
 $ Clean.Cup                 : num  10 10 10 10 10 10 10 10 10 10 ...
 $ Sweetness                 : num  10 10 10 10 9.33 9.33 10 10 10 10 ...
 $ Cupper.Points             : num  8.58 9.25 8.67 8.58 9 8.67 8.5 8.58 8.5 8.33 ...
 $ Category.One.Defects      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Category.Two.Defects      : int  1 0 2 2 0 0 4 1 0 0 ...
 $ Quakers                   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Moisture                  : num  0.12 0 0.11 0.12 0.03 0.03 0.1 0.1 0 0 ...
 $ altitude_mean_meters_new: num  2075 1700 2000 2075 1635 ...
 $ Total.Weight              : num  18000 5 19200 18000 18000 18000 3000 18000 10 10 ...
 $ Variety                   : Factor w/ 19 levels "Arusha","Bourbon",..: 9 2 18 9 18 18 9 1
 $ Color                     : Factor w/ 5 levels "Blue-Green","Bluish-Green",..: 3 5 3 3 5
 $ Processing.Method         : Factor w/ 6 levels "Natural / Dry",..: 6 5 1 6 5 5 1 1 6 6 ..
 $ Country.of.Origin         : Factor w/ 27 levels "Brazil","China",..: 6 7 6 6 6 6 6 6 24 2
 $ In.Country.Partner        : Factor w/ 24 levels "Africa Fine Coffee Association",..: 14 1
```

VARIABLE DESCRIPTION

Aroma: Smell of dry coffee, wet cofee at crust break and wet coffee as it steeps
Flavor: Mid-range notes between first impression and aftertaste. Intensity, quality and complexity.
Aftertaste: Length and quality of enjoyable flavour after the coffee is swallowed
Acidity: Brightness (Good) and Sour (Bad)
Body: Tactile feeling between tongue and mouth
Balance: Balance of flavour, aftertaste acidity and body
Uniformity: Consistency of flavour across cups
Clean Cup: Lacking negative tastes from beginning to end of taste
Sweetness: Sugar flavor (Good) and astigency flavours (Bad)
Cupper Points : Holistic score by cupper
Moisture: The moisture percentage from the green bean
Category One Defect: This feature indicated the major defects in the bean. In the dataset the value ranges from 0 to 55.
Category Two Defects: Minor defects on the bean. In the data set this value ranges from 0 to 55.
Quakers: These are coffee that do not turn dark when roasted.
altitude_mean_meters_new: The average altitude of where the bean is harvested
Total Weight: The Bag Weight multiplied by the Number of Bags (the number of bags sampled for the specific bean and the weight of the bag of the coffee bean that was sampled.
Variety: Further divides into multiple varieties of coffees
Color: This represents the color of the green bean whether it is green, blueish green or blue green
Country of Origin: The country the coffee is from.
In Country Partner: The organisation that is teaming up with the Country to sample and conduct the grading process

EXPLORATORY DATA ANALYSIS

Target Variable (Grade)



Imbalance datasete

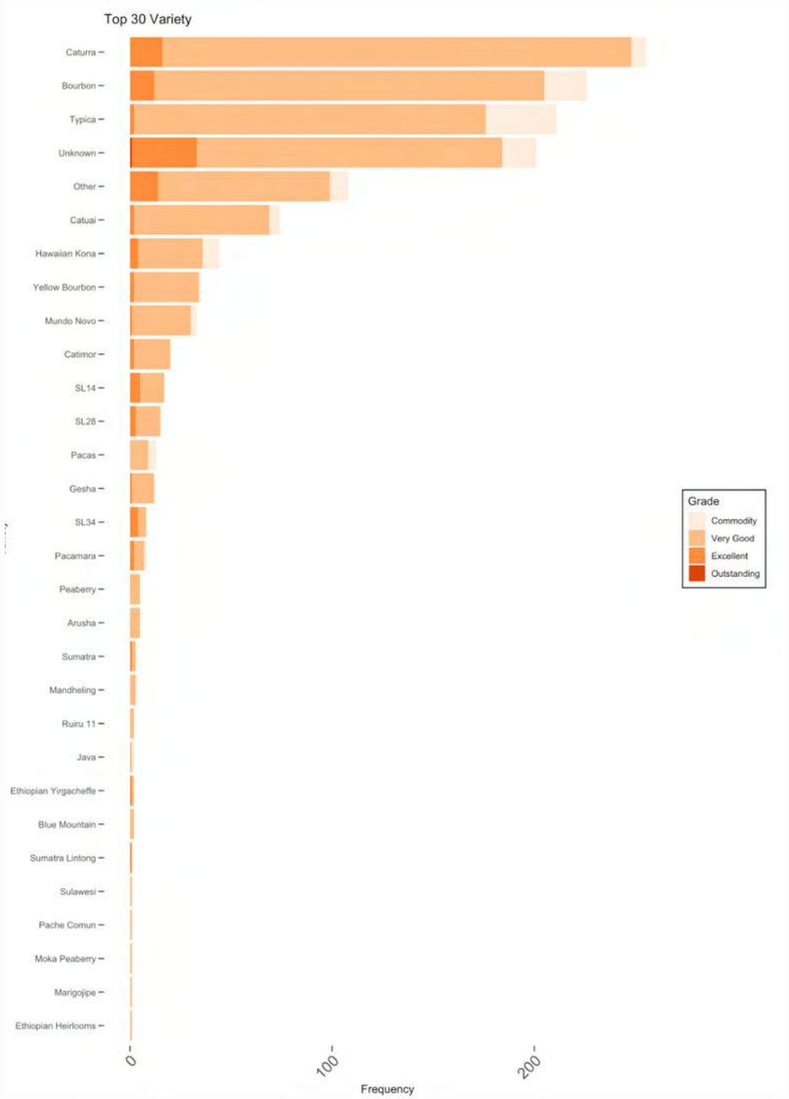
80% of the observation fall under very good

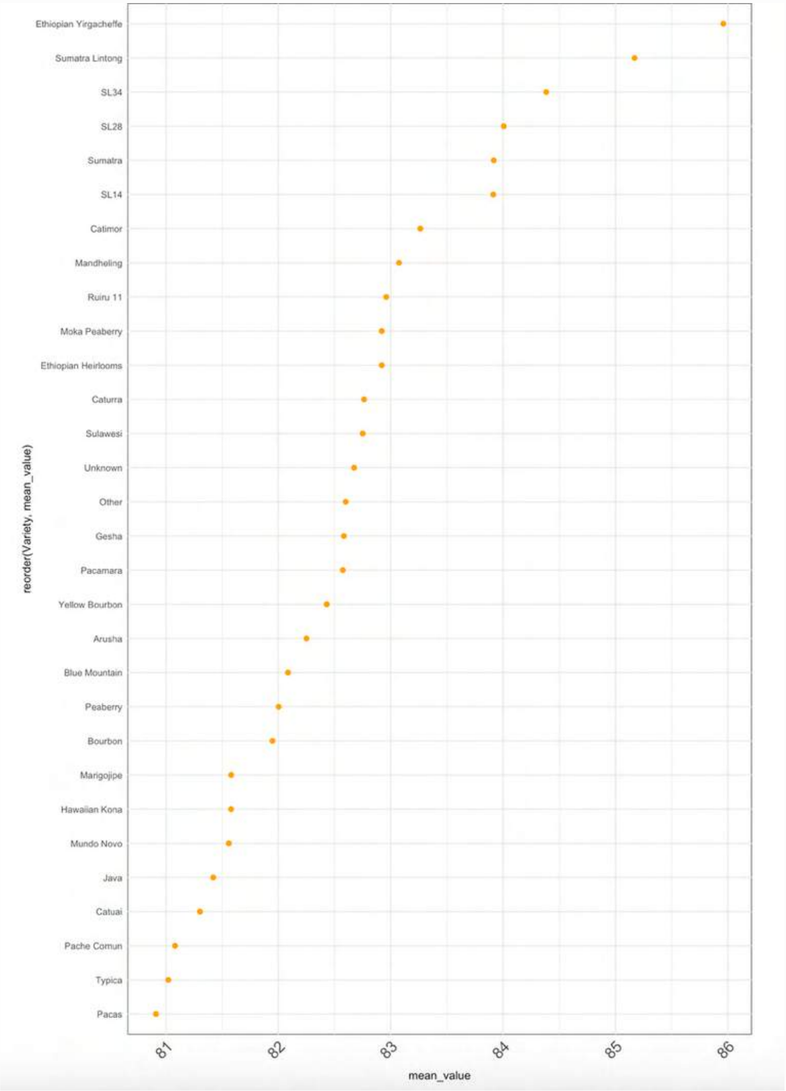
Outstanding has only one observation

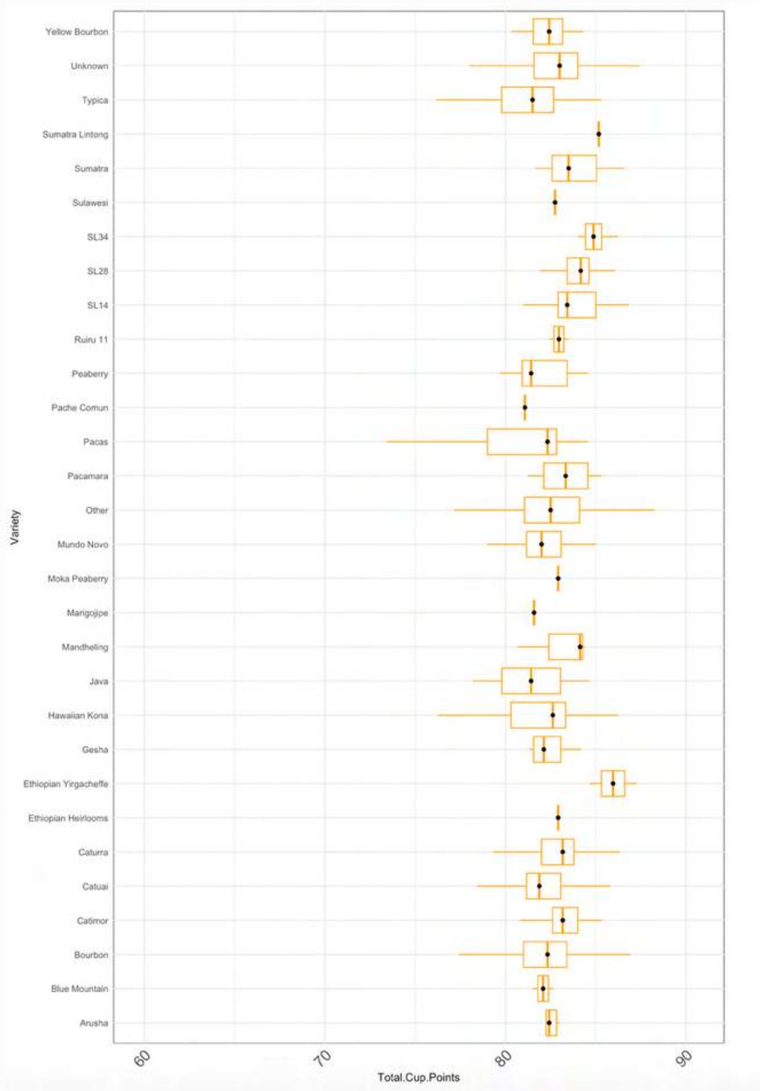
We handle this by using oversampling techniques when conducting our analysis

EXPLORATORY DATA ANALYSIS

Categorical Variables (Variety)

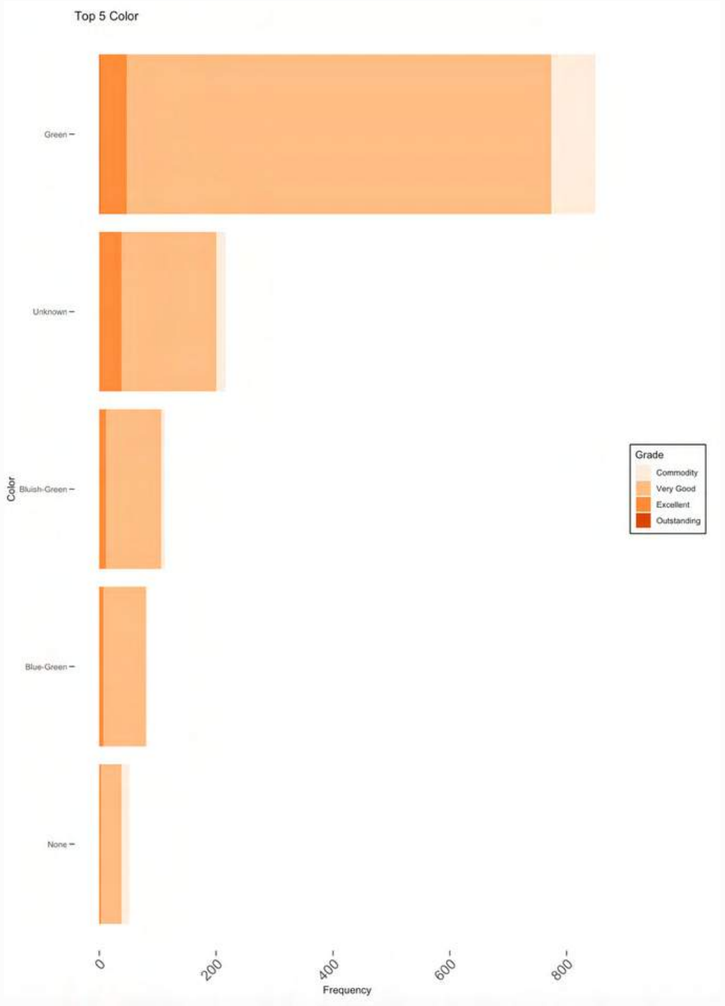


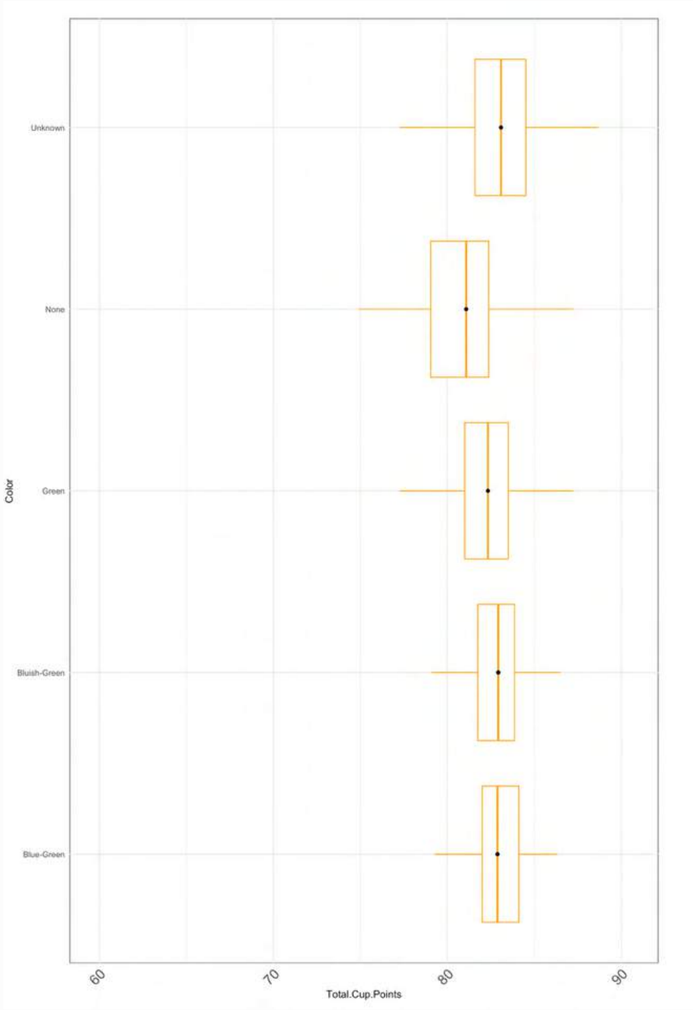


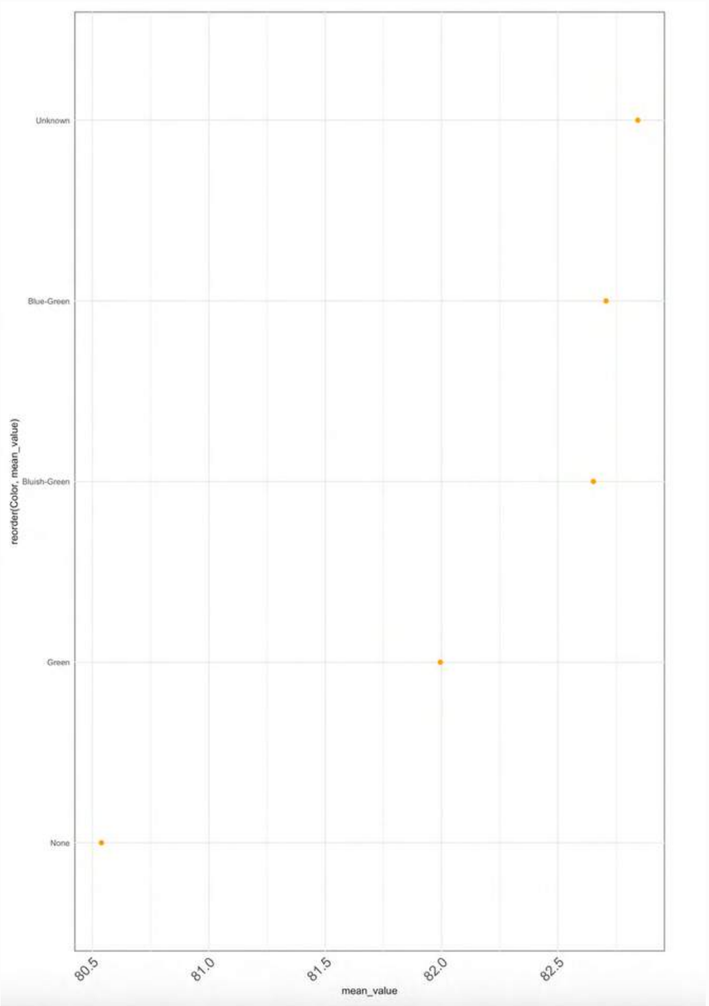


EXPLORATORY DATA ANALYSIS

Categorical Variables (Color)

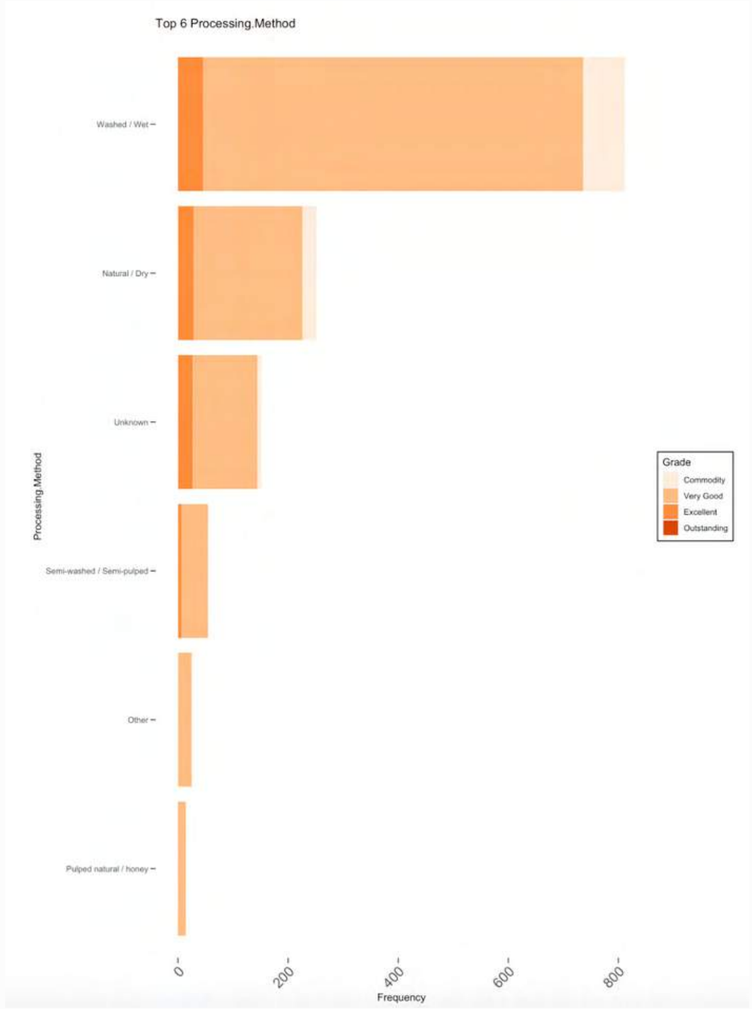


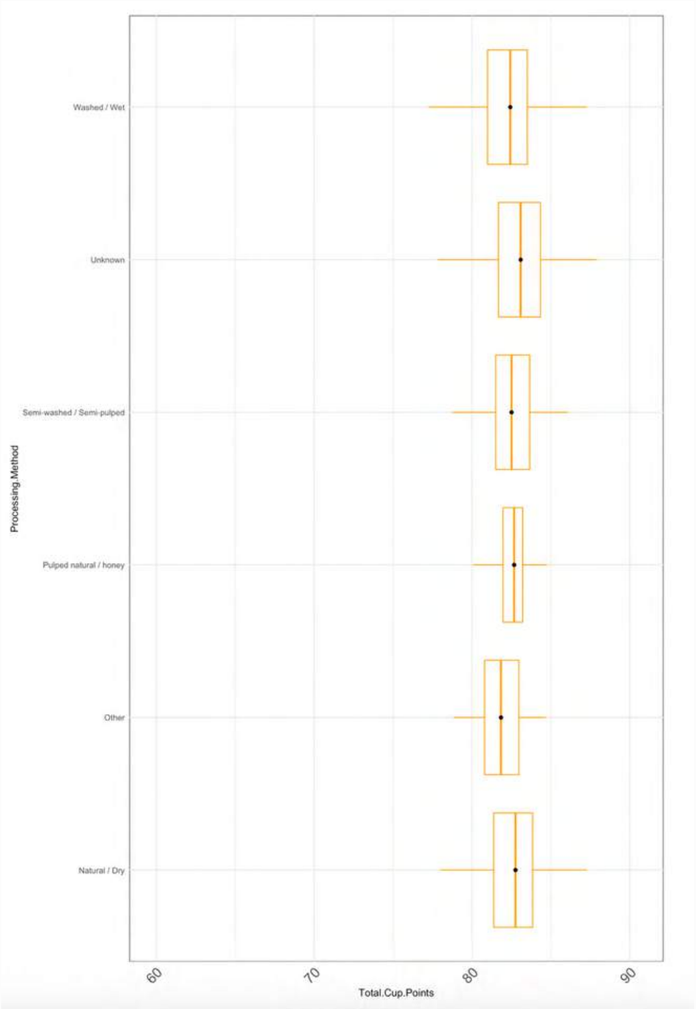


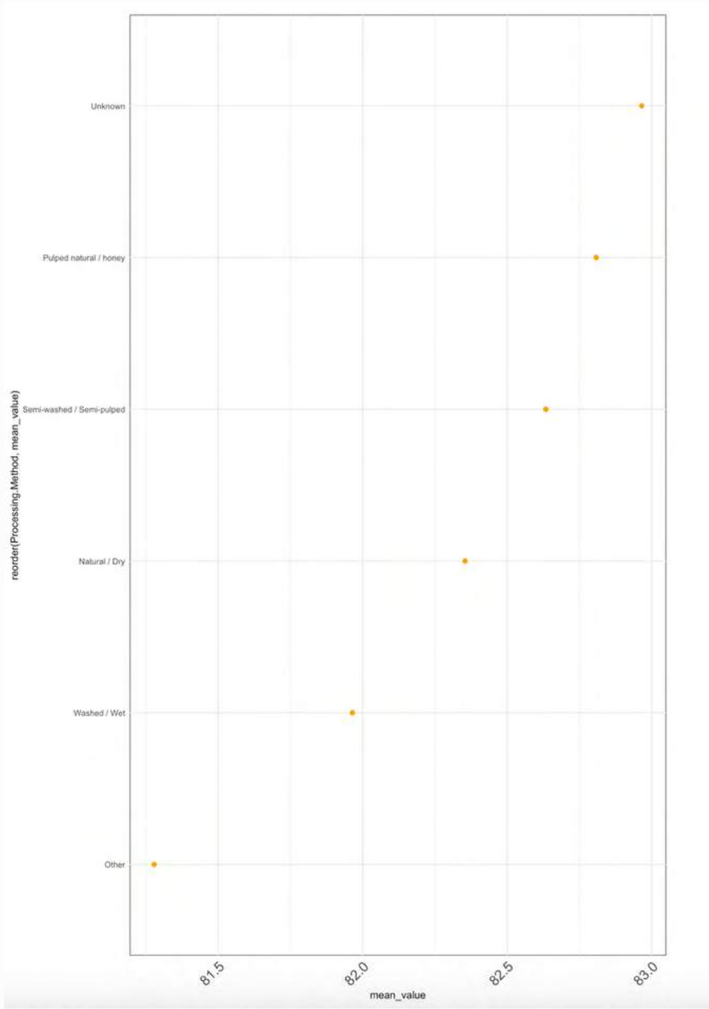


EXPLORATORY DATA ANALYSIS

Categorical Variables (Processing Methods)

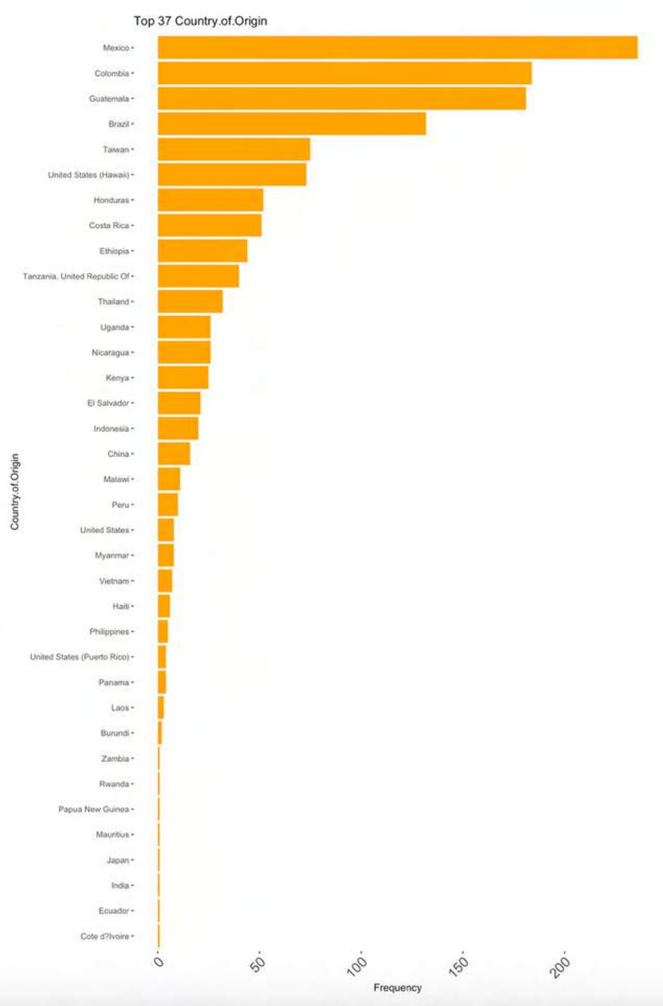


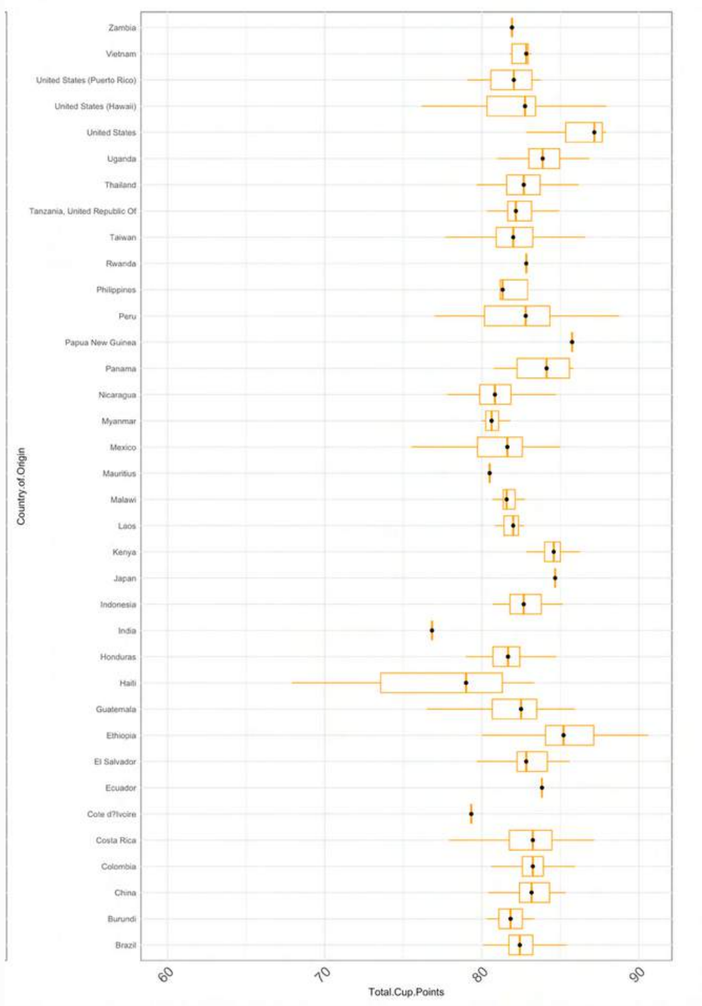


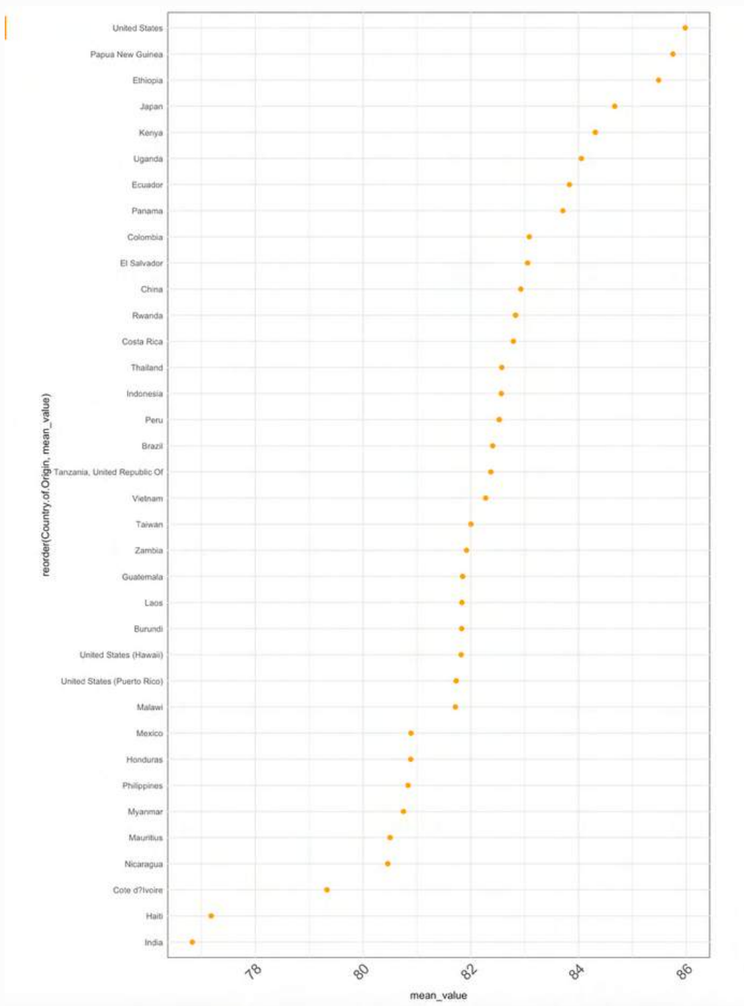


EXPLORATORY DATA ANALYSIS

Categorical Variables (Country.of.Origin)

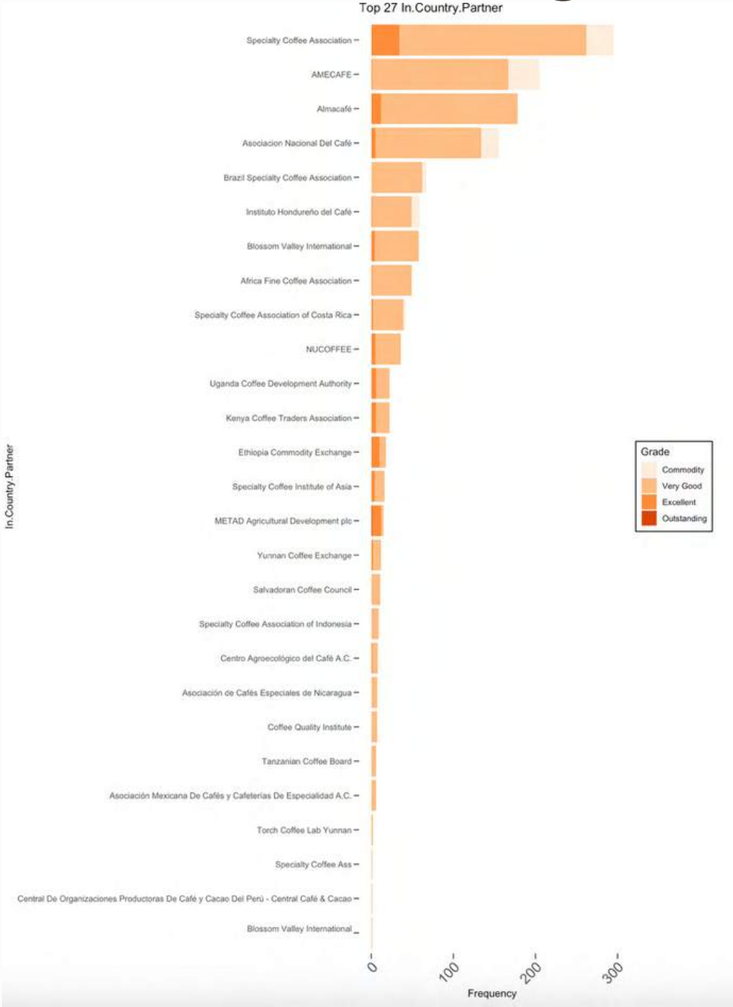


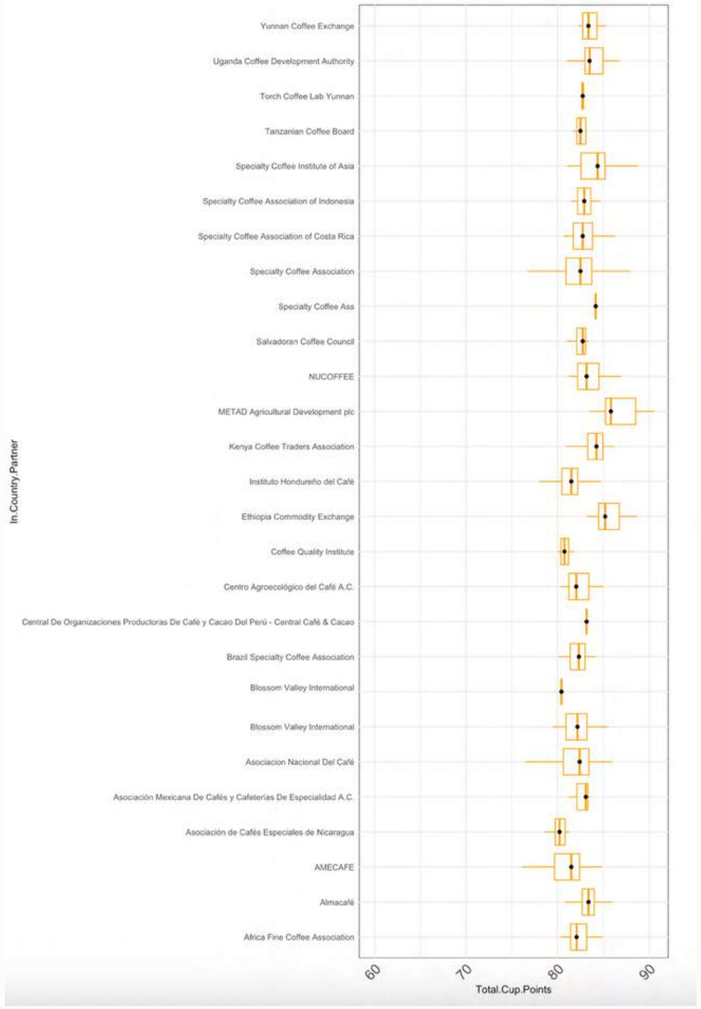


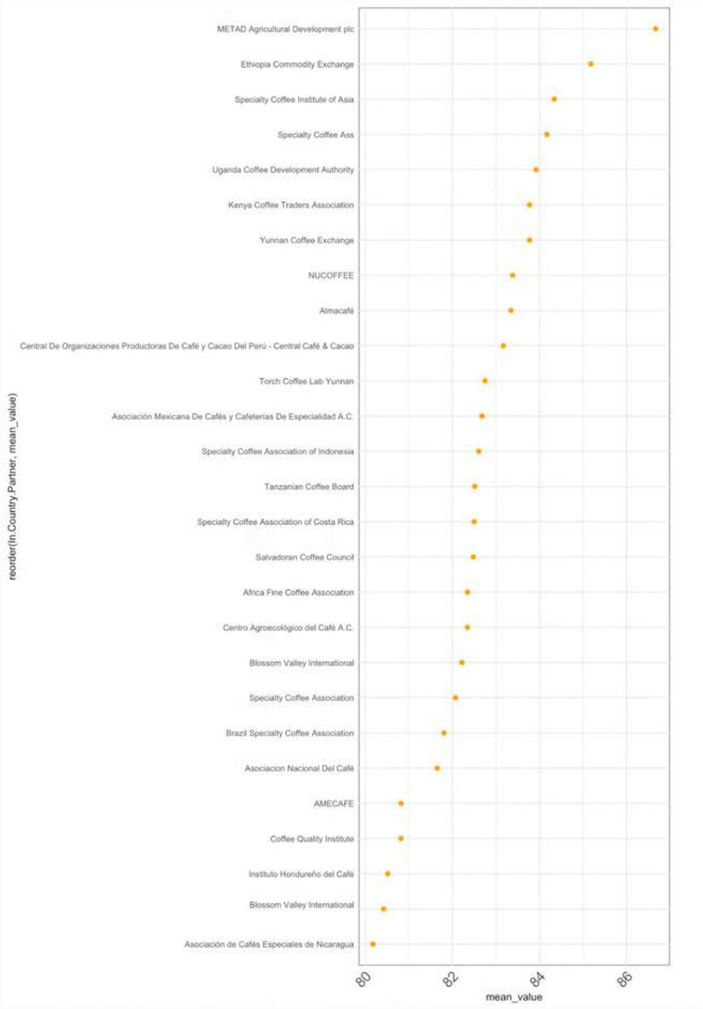


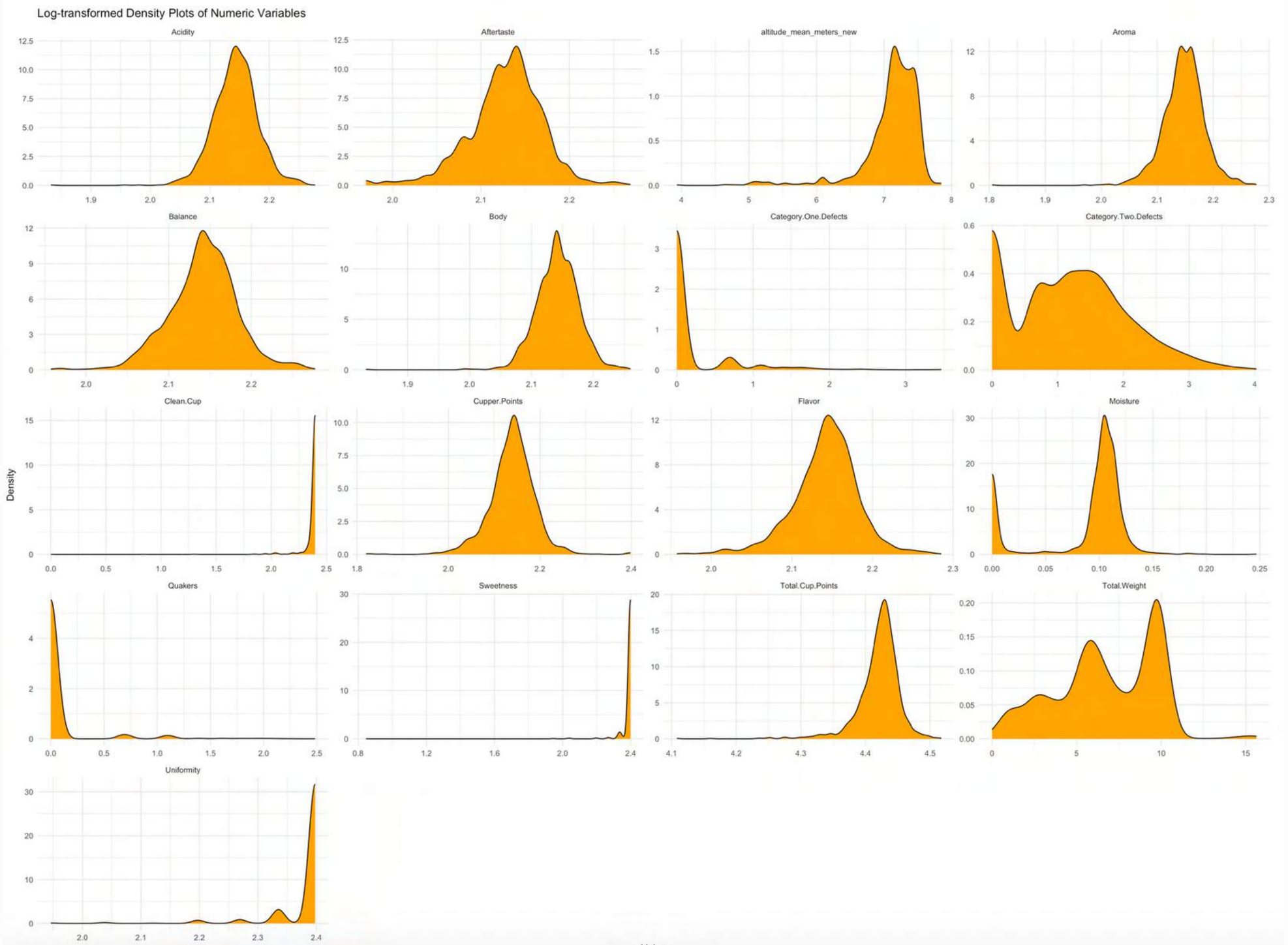
EXPLORATORY DATA ANALYSIS

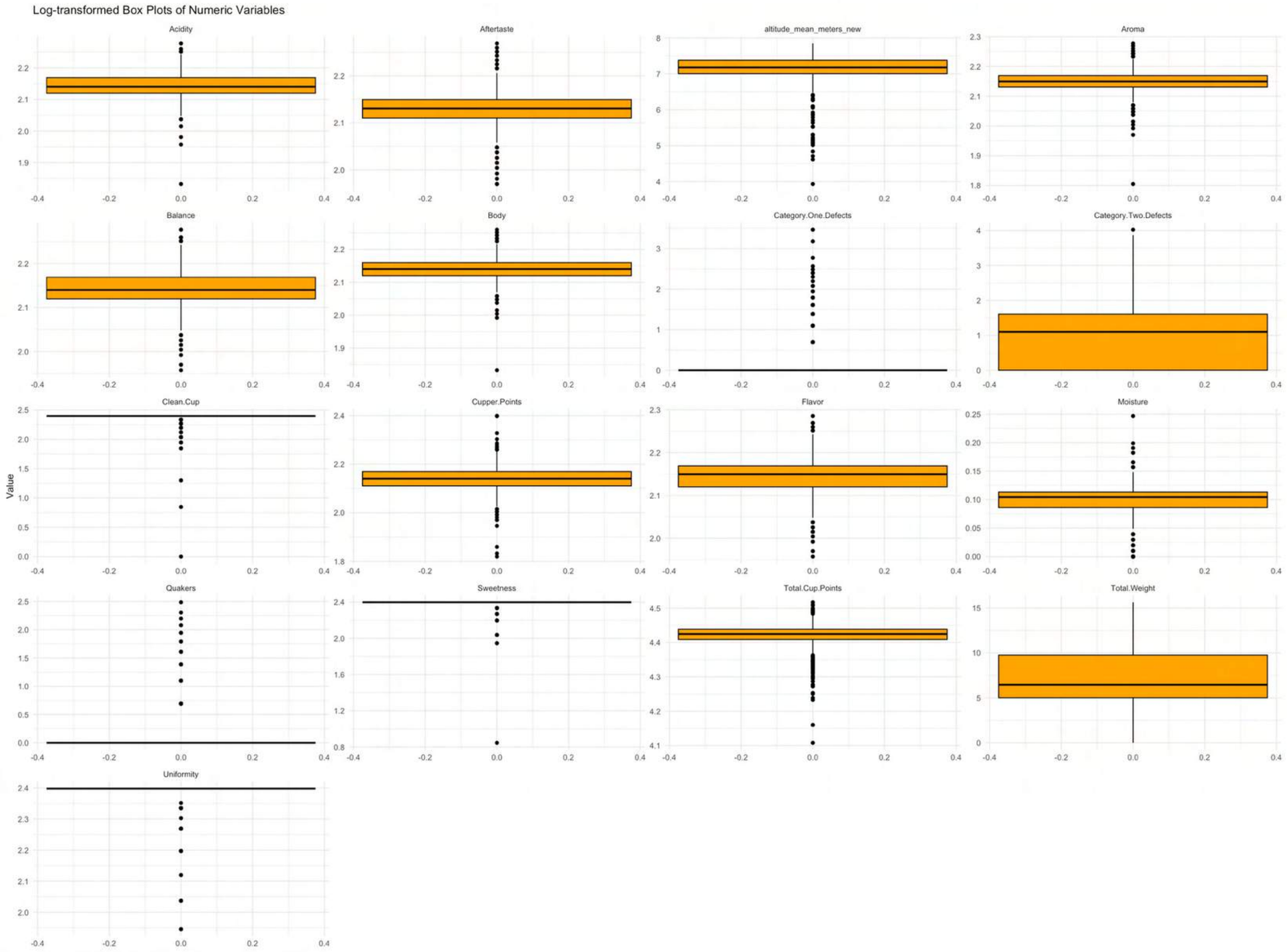
Categorical Variables (In.Country.Partner)





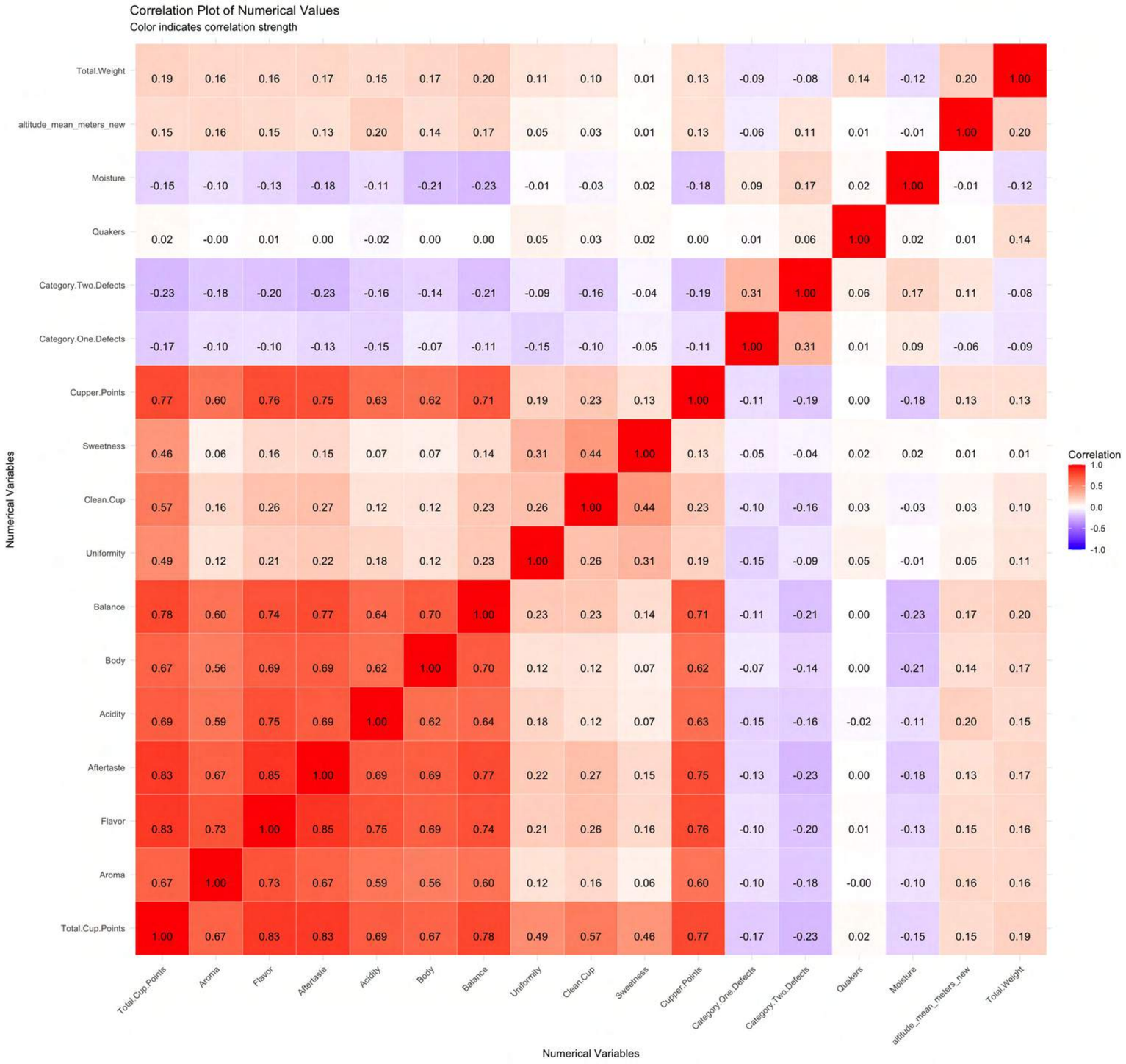



Log-transformed Density Plots of Numeric Variables
Acidity
Aftertaste
altitude_mean_meters_new
Aroma
Balance
Body
Category.One.Defects
Category.Two.Defects
Clean.Cup
Cupper.Points
Flavor
Moisture
Quakers
Sweetness
Total.Cup.Points
Total.Weight
Uniformity
Density



Outliers are present in the data set especially in the skewed variables and also total cup points.

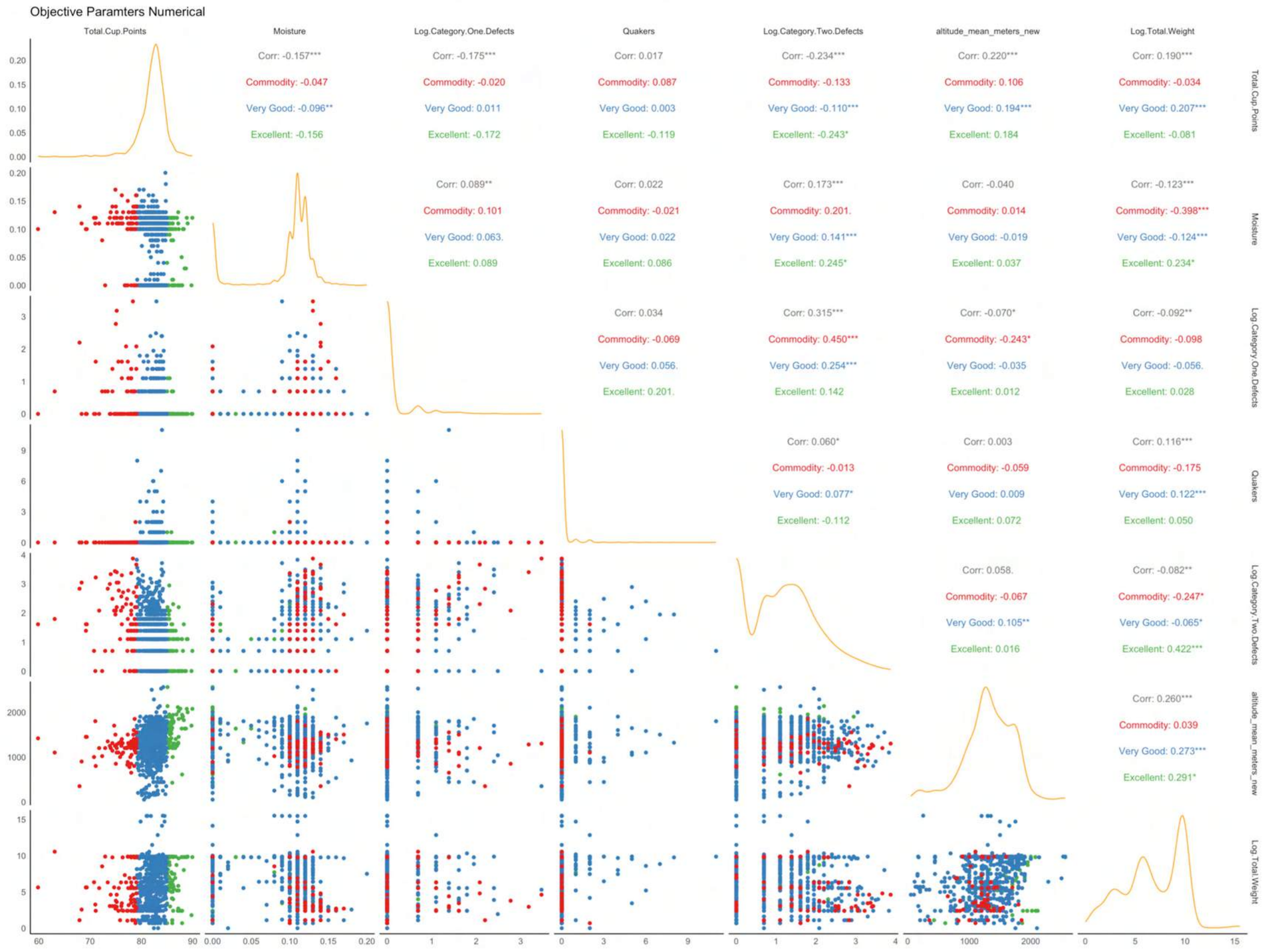
Keeping these variables is important to distinguish the different coffee grade quality and avoid loss of information.



We can see that the sensory data are highly correlated with Total.Cup.Point as the values are used to calculate it.

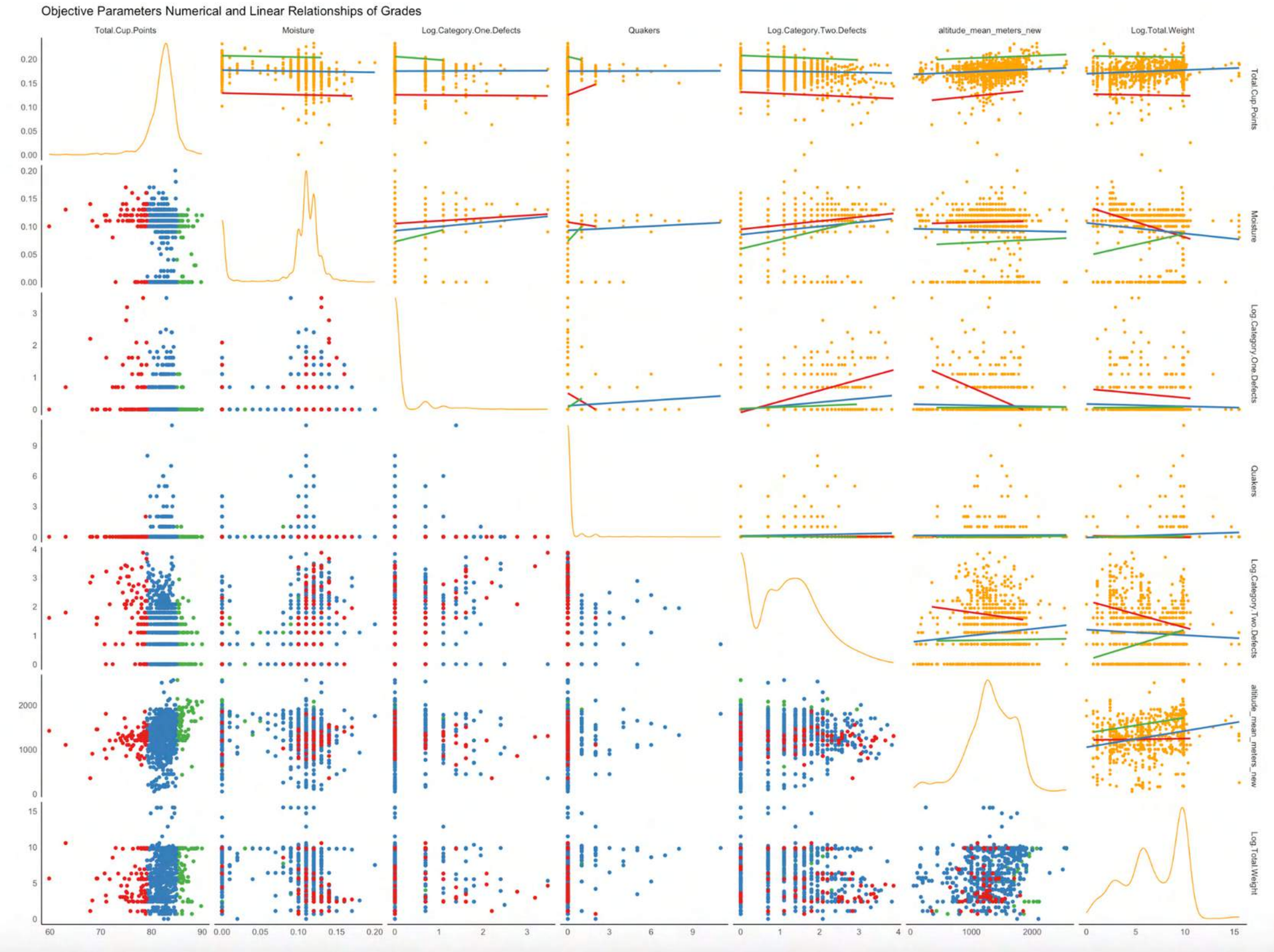
Non-sensory data we can see, Category.One.Defects, Category.Two.Defects, and Moisture are negatively weakly correlated.

No indication of multicollinearity between objective parameters.



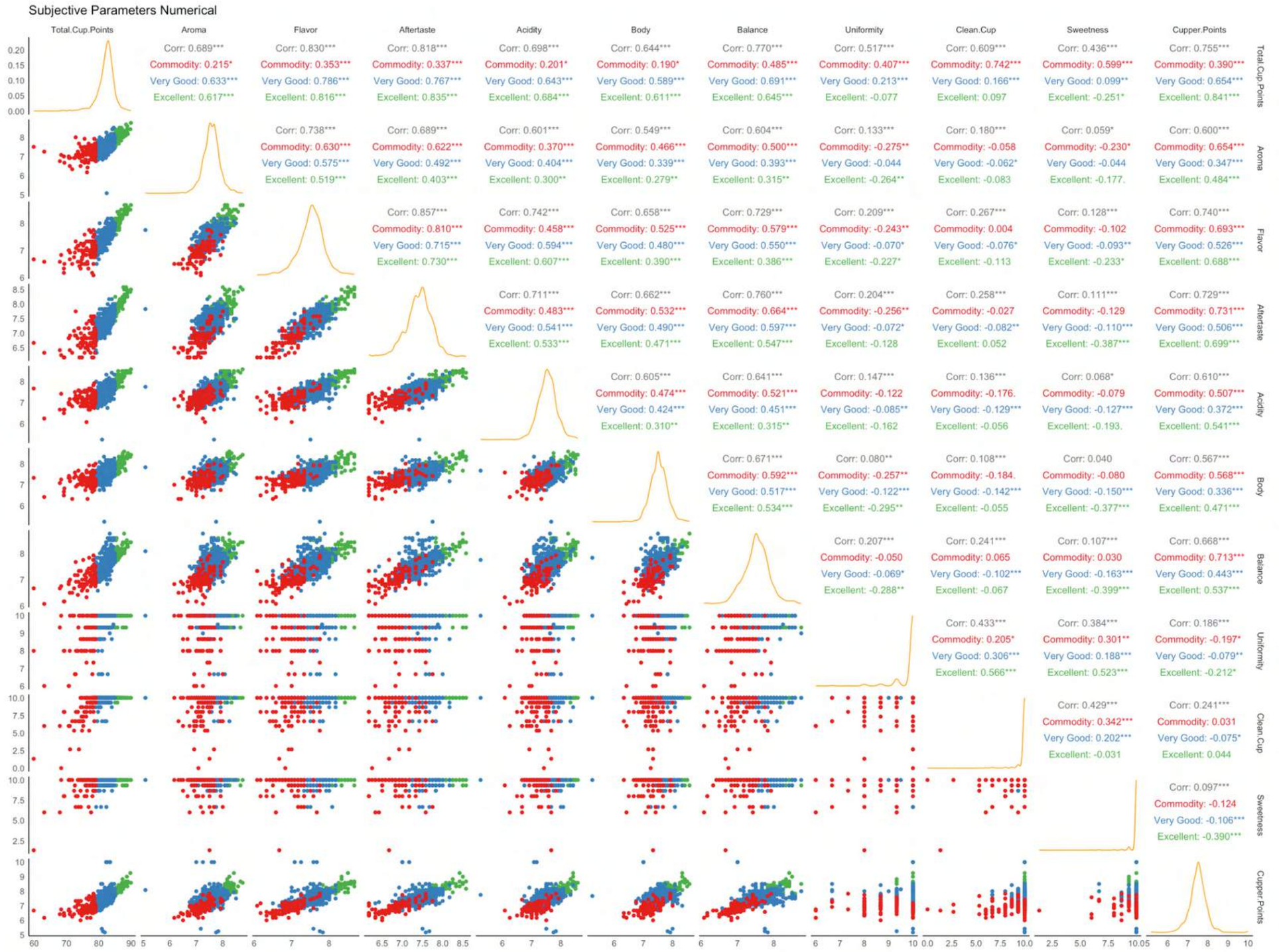
No linear relationship can be seen at first glance.

The objective parameters show weak correlation between each other and Total.Cup.Points.



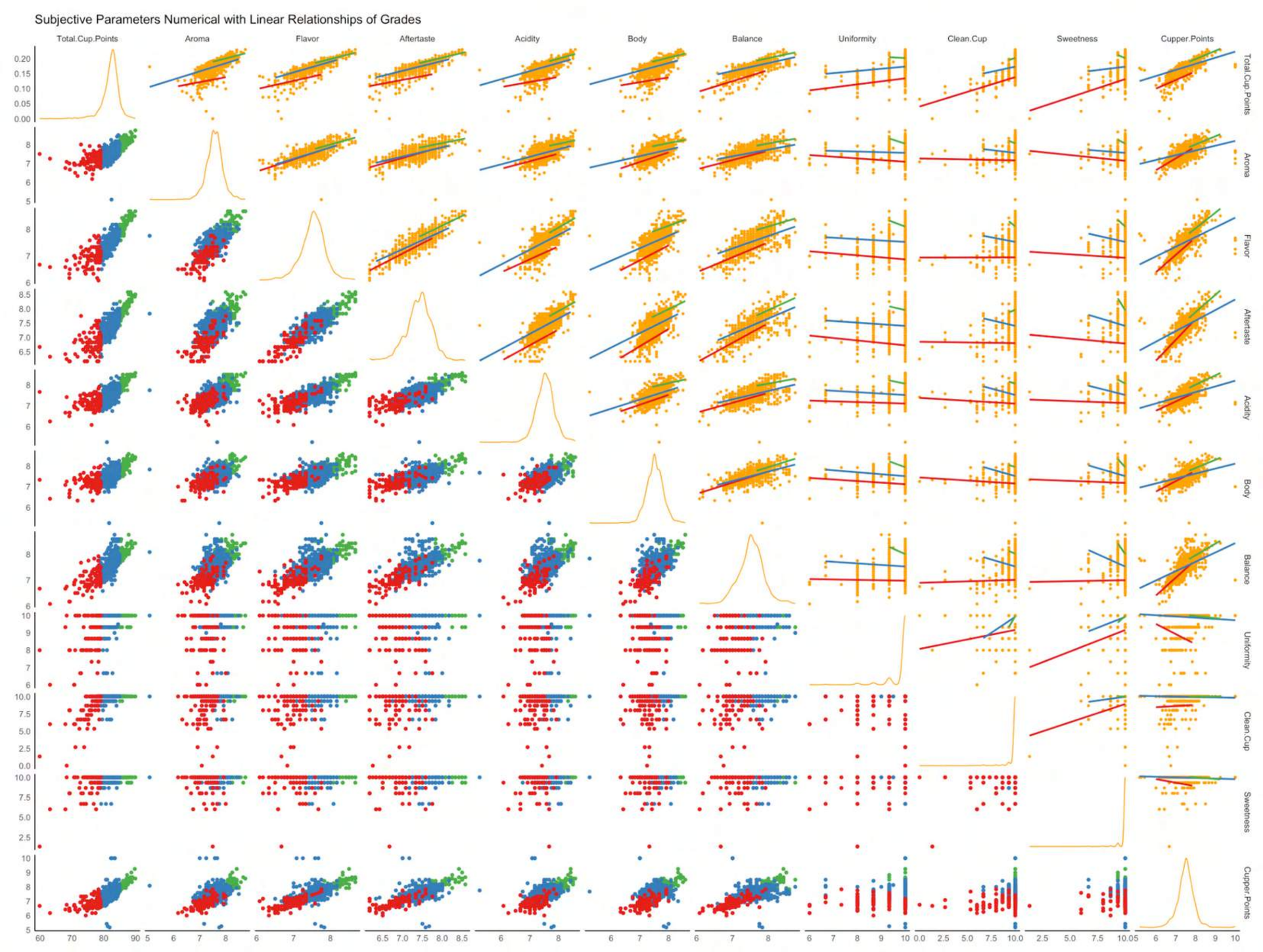
When plotting the linear relationship we can confirm there to no distinct (small) linear relationships.

We have separated it into grades to see how each of the variables impact coffee quality grade for commodity, very good and excellent.



With subjective parameters we can see a liner relationship between Total.Cup.Points and between each other.

The correlation is also positively and positively strong for a majority of the variables.



We can see that linear relationship of each of the grades is positive for most of the variables just different levels.

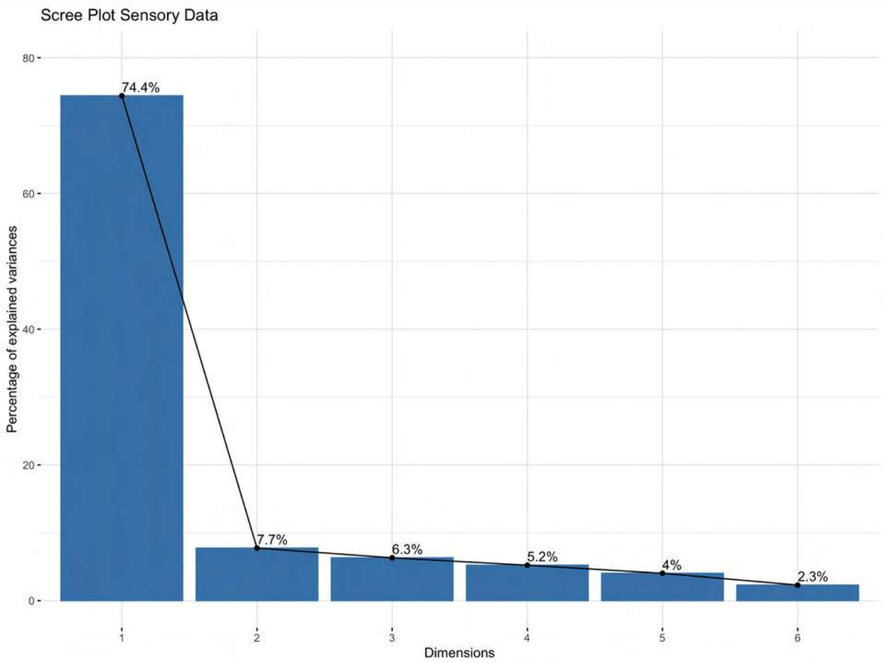
This is expected as these values are used to calculate the Tota.Cup.Points.

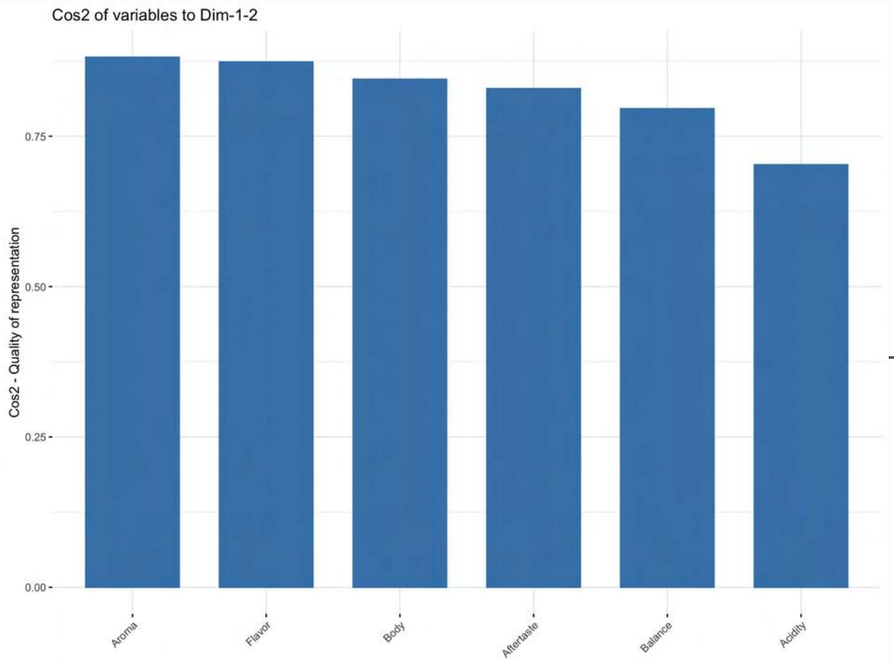
UNSUPERVISED

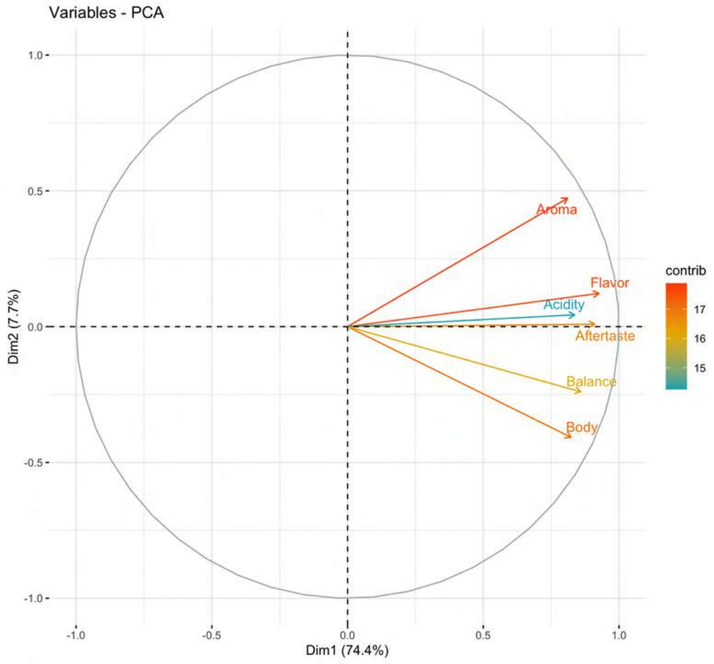
These are the questions we would like to answer:

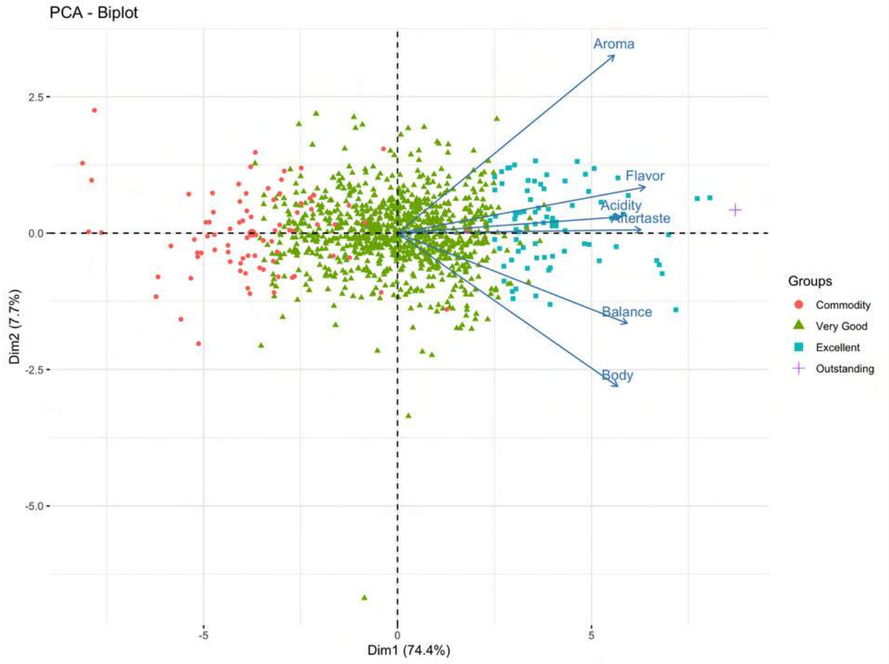
1. Can we identify distinct sensory groupings of coffee beans based on the subjective parameters?
2. Are there natural groupings based on non-sensory features like altitude, country of origin, processing methods?
2. Can we find distinct clusters based on the sensory and the non-sensory features for coffee bean grade?

SENSORY FEATURES









Conducted a PCA on the sensory attributes.

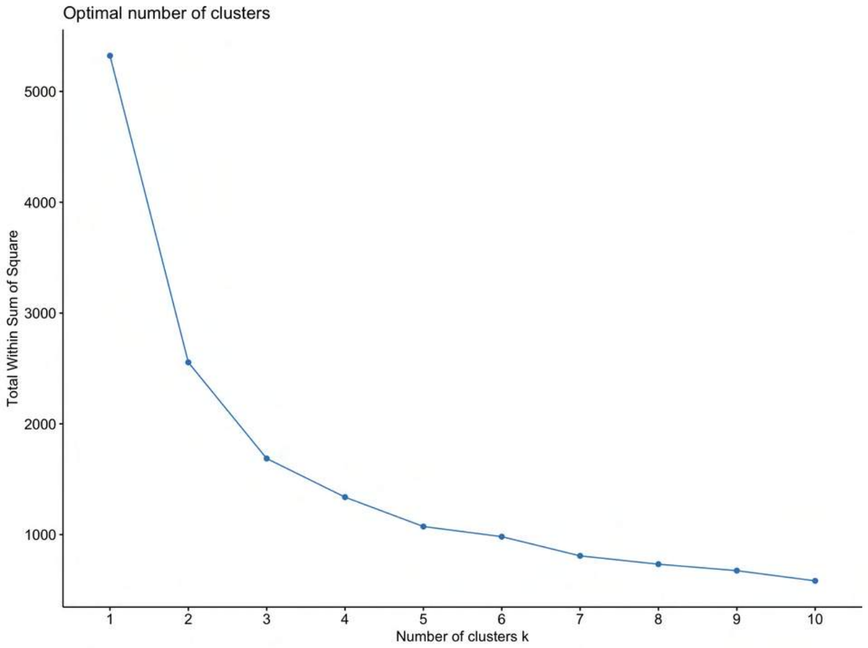
The scree plot indicated that the first principal component explains around 74% of the variance and the second component explains 7.7%.

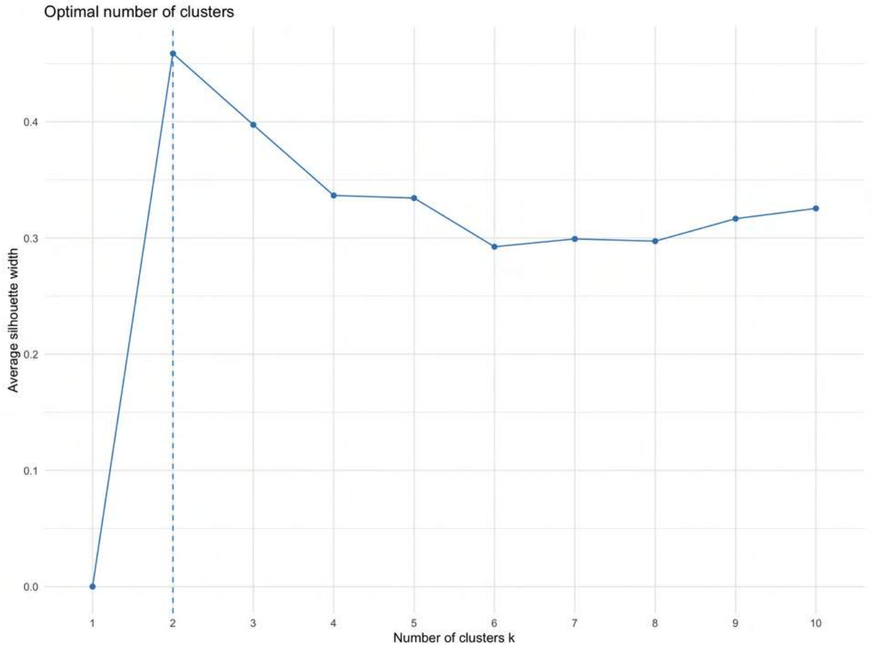
The contribution plot indicated that Aroma and Flavor, contribute significantly, followed by Aftertaste.

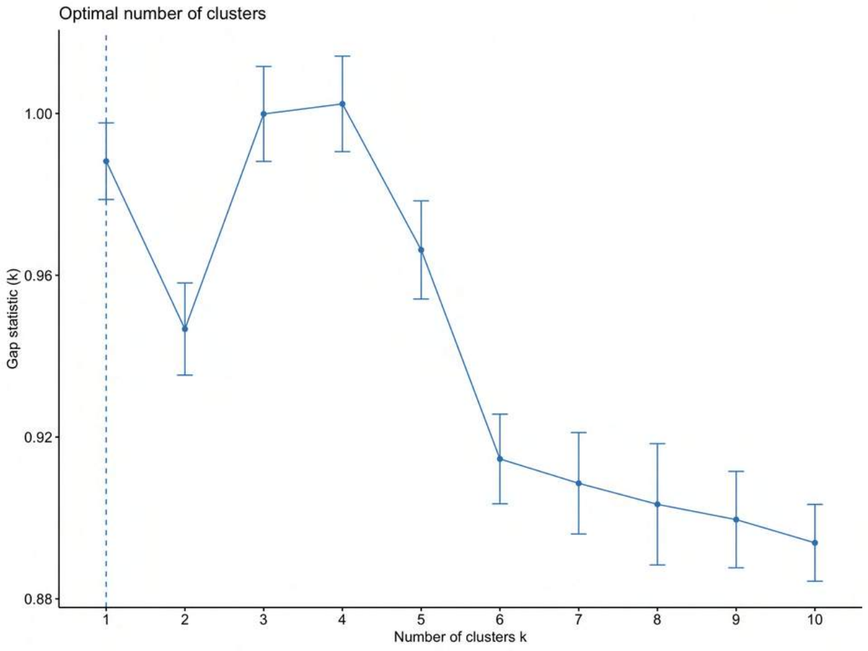
Aroma and Flavor contribute to the first dimension. Where as Aroma and Body contribute to the second principal component.

We can see from the fourth plot that the ones that are clear separation between groups indicates distinct patterns between classes.

SENSORY FEATURES







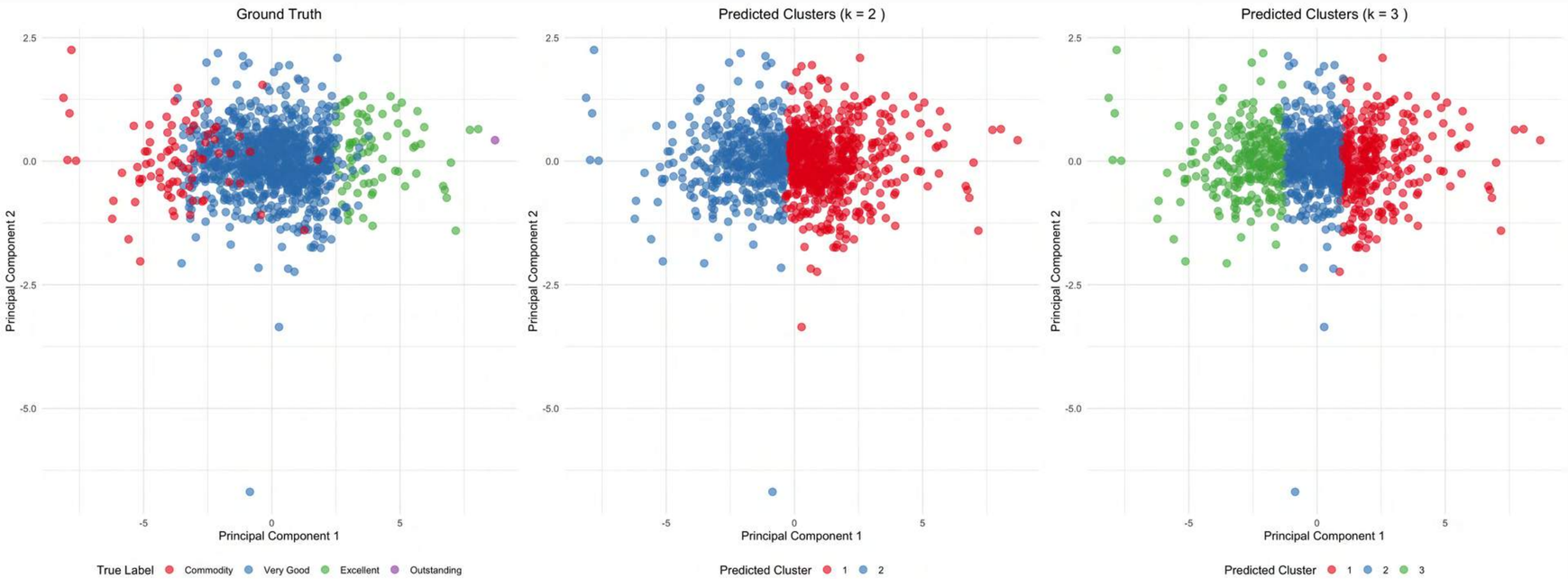
Assessing the optimal number of clusters that we need using the tree common techniques which are the “elbow method”, silhouette method and the gap statistics.

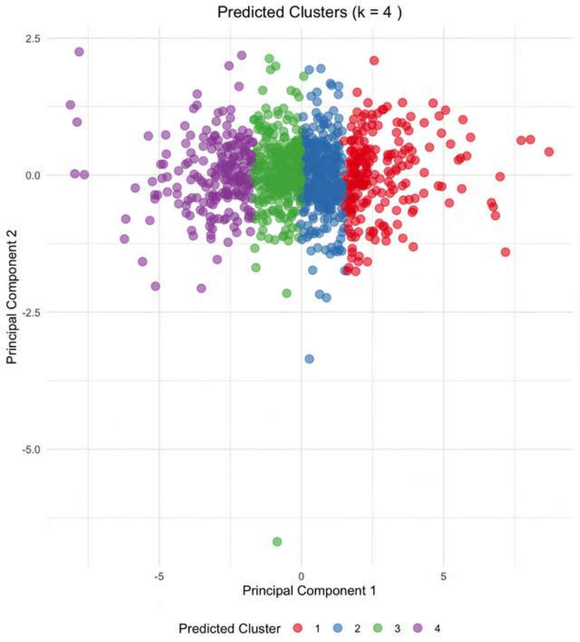
Elbow Method indicates that 3 or 4 would be the best number of clusters.

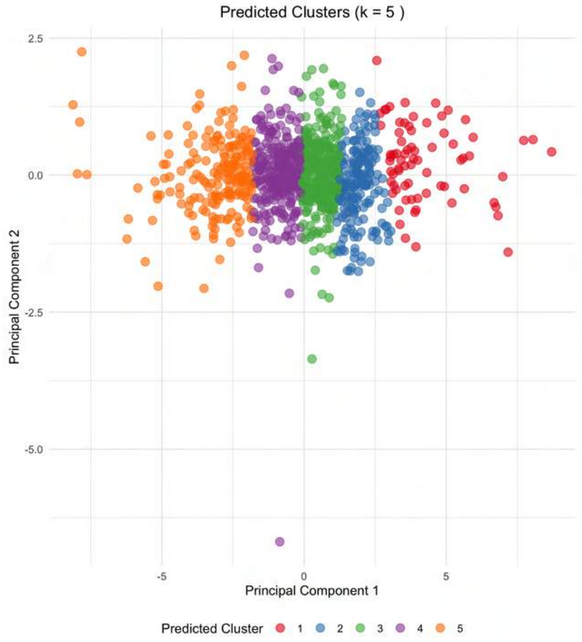
Silhouette method shows that 2 would be the best number of clusters.

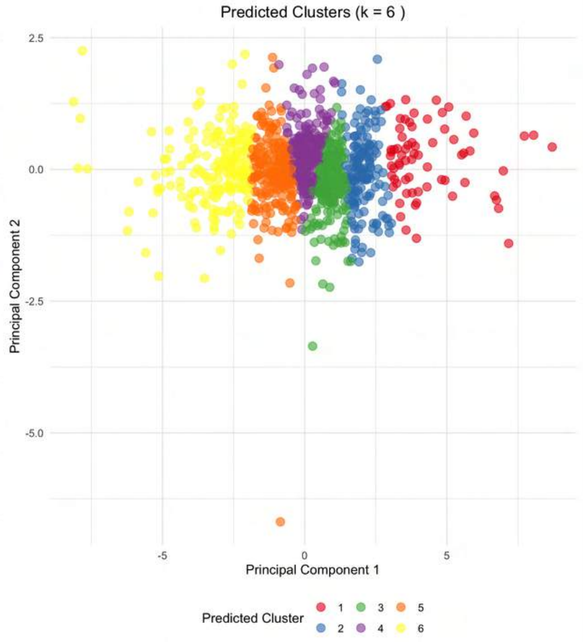
Gap statistics indicates that 4 would be the best number of clusters.

SENSORY FEATURES









Then we conducted a PAM clustering on the dimensions.

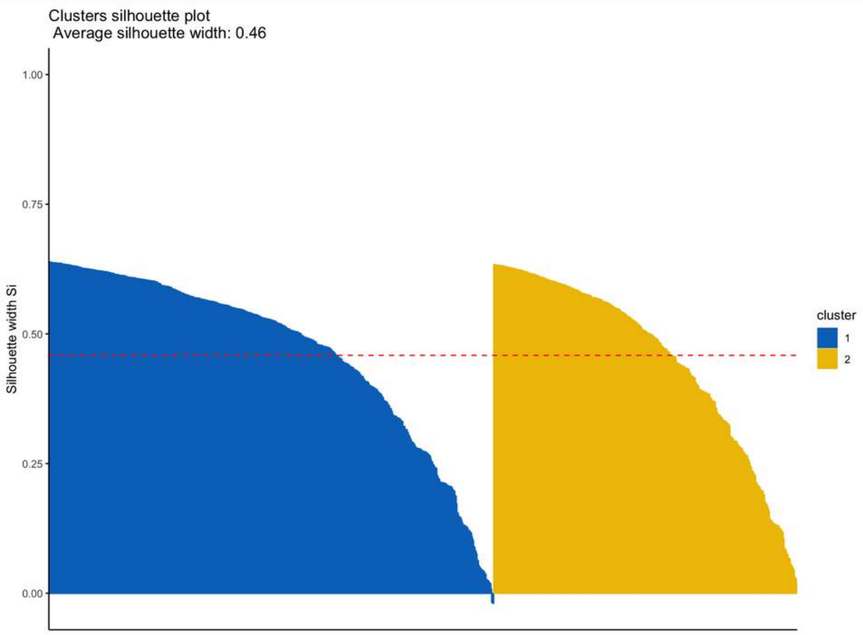
The first plot shows the ground truth.

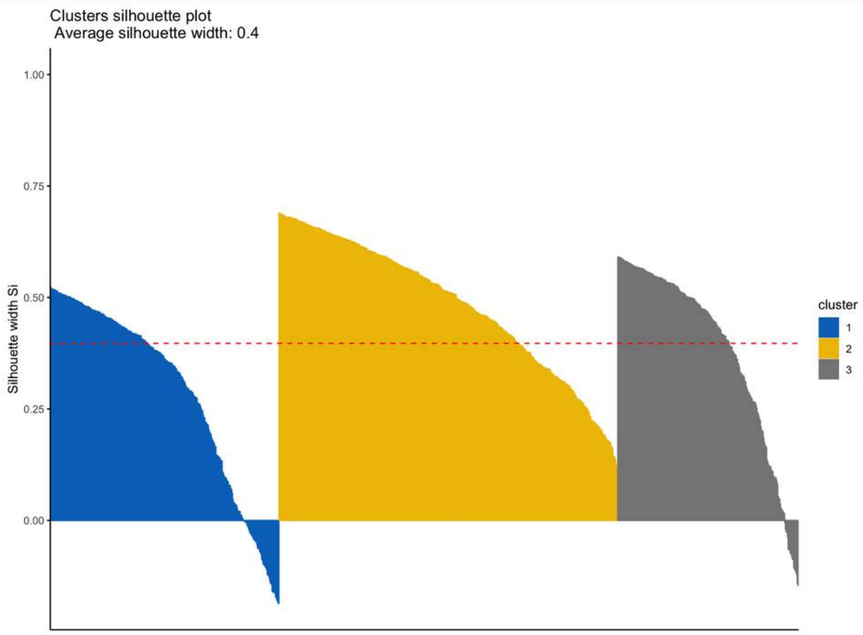
We can see that when k = 53 somewhat resembles the ground truth yet it is still far perfect.

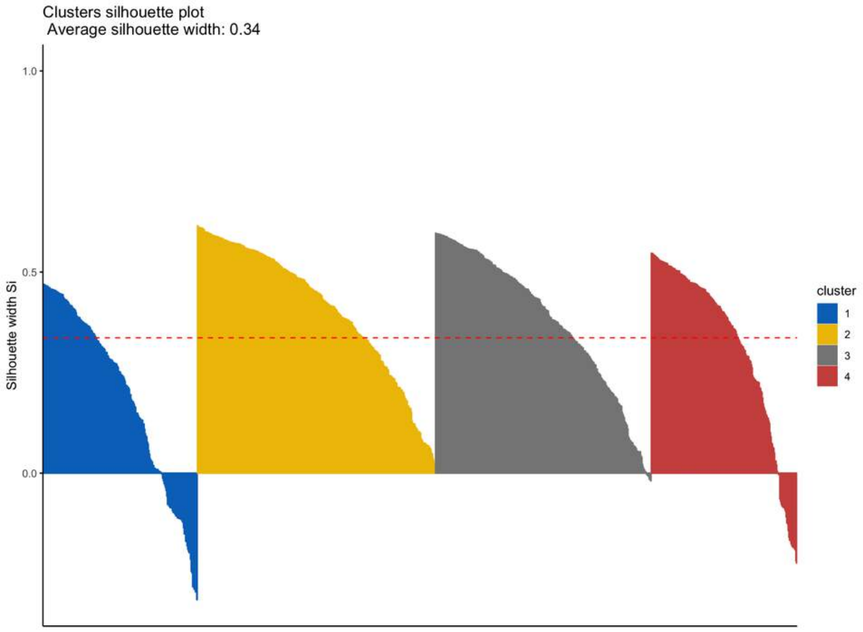
When increase the number of clusters we can see that the Commodity and Excellent are starting to be represented better.

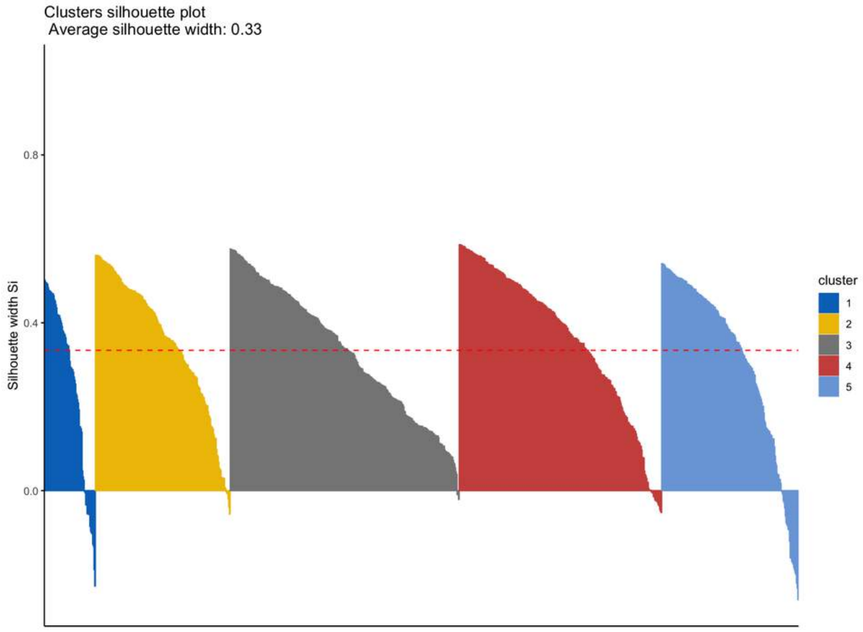
But the Very Good coffee is being split into further and smaller clusters.

SENSORY FEATURES









When we have our cluster is 2, the average silhouette is 0.44 which with some data points being clustered incorrectly for the furst cluster.

By increasing the number of clusters to k = 3, we can see that the average silhouette width decrease to 0.4 with some the number of miss clustering being present observed in cluster 1 and 3.

As we increase we can see that the width of the silhouette decreasing, dropping 0.33.

SENSORY FEATURES

Number of Clusters	Rand Index
2	0.03961925
3	0.10958356
4	0.09082945
5	0.11950625
6	0.08053726
7	0.09360433

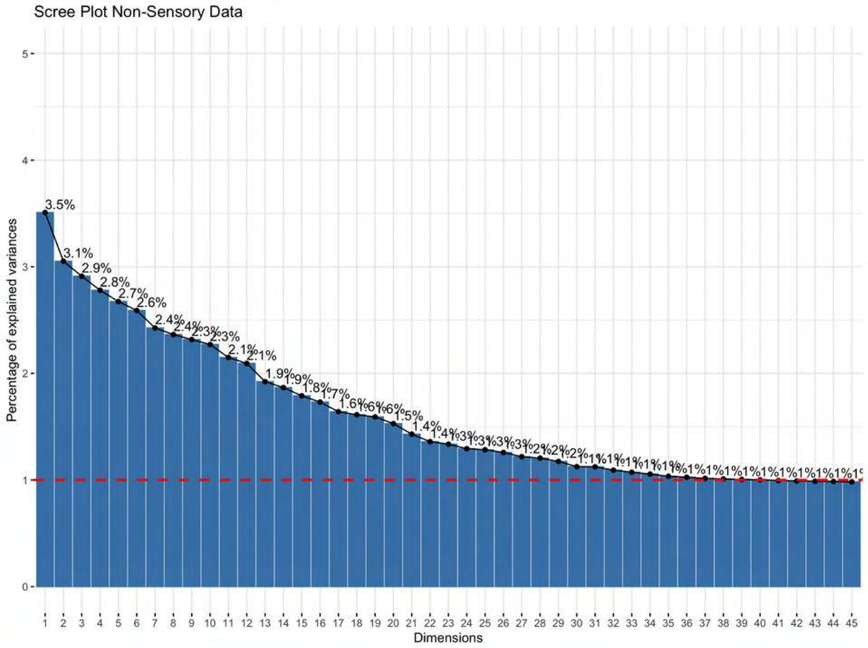
As an external validation we used Rand Index which measures the similarity between two data clustering by using true labels.

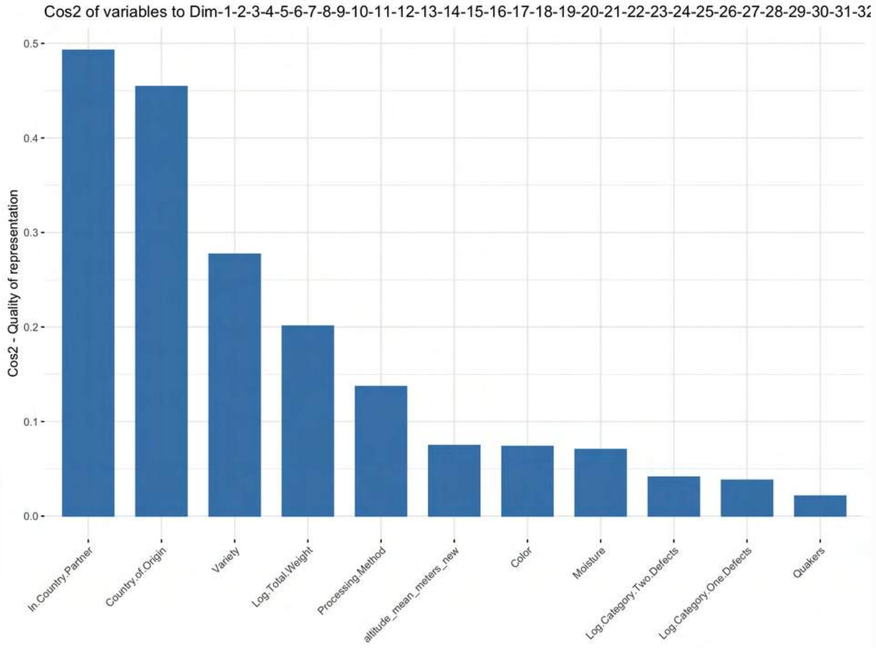
We can see that from the table when k = 5, the Rand Index is at the highest and matches the plot of the clustered plots.

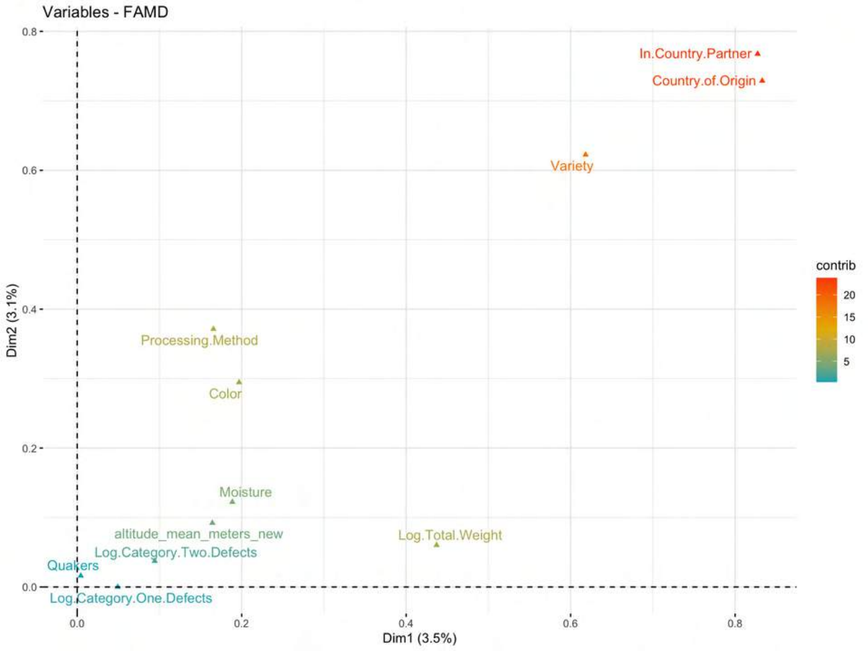
With other values of clusters, index is quite low when compared to k = 5, except for when k = 3.

This suggests that give cluster is more similar to the ground truth but we have to acknowledge that the index is still very low.

NON-SENSORY FEATURES







Conducted a FAMD on the non-sensory attributes.

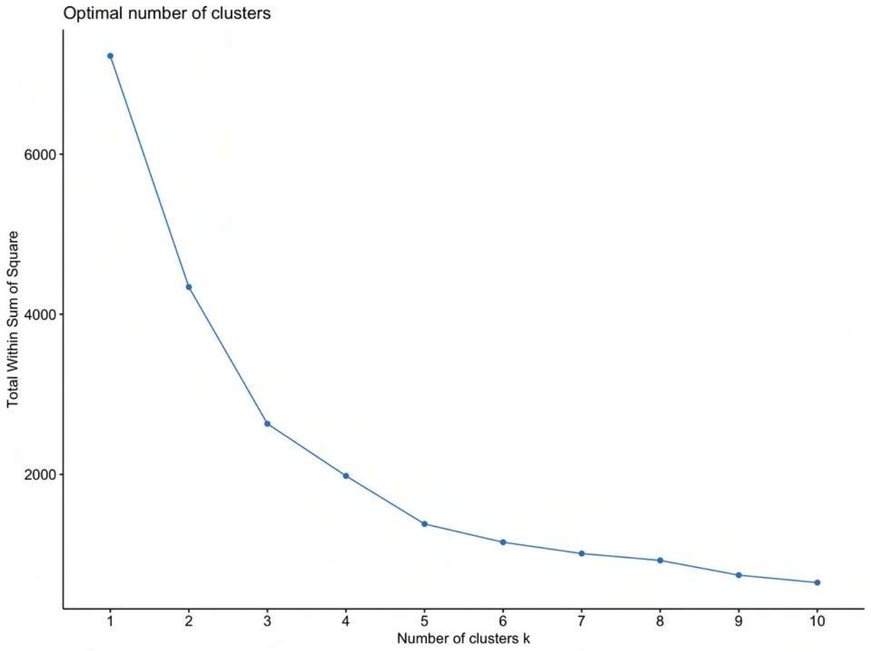
According the Kaiser we can see that the cutoff dimension is 35 principal component. In addition 80% of the variance is explained by 53 dimensions.

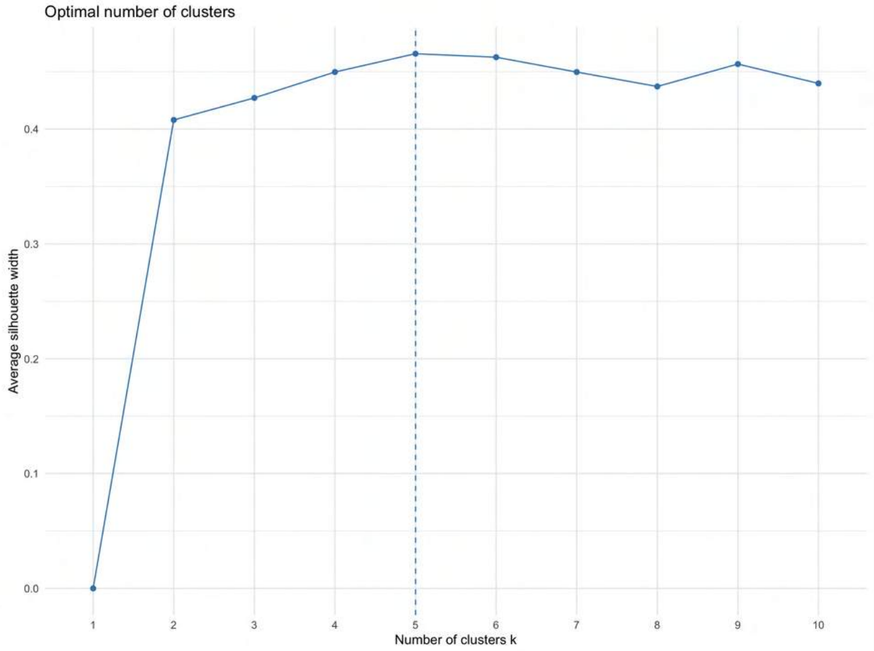
Form the cos squared plot we can see that In.Country.Partner and Country of Origin are the most important features.

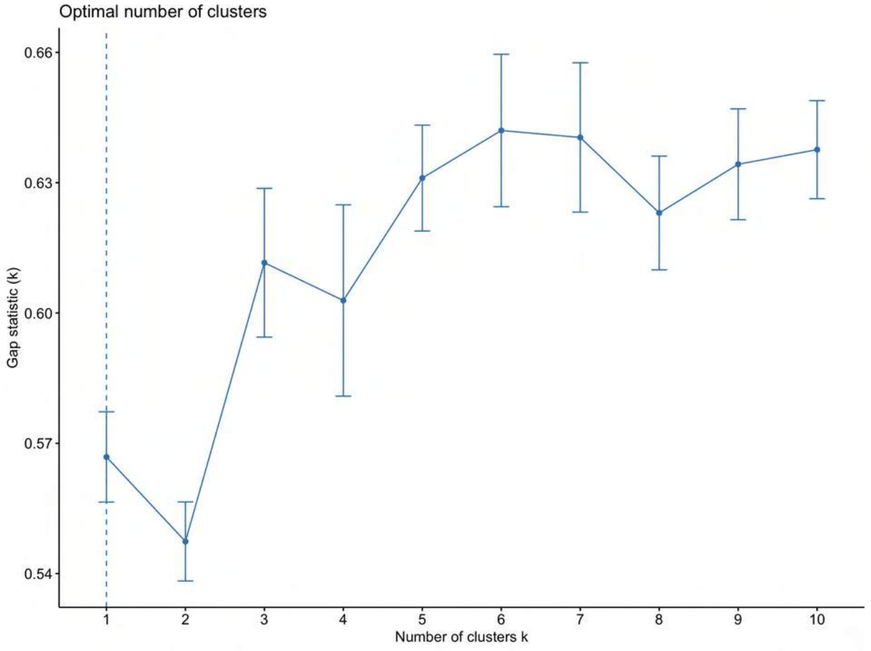
In the third plot, features such as Quakers, Log.Category.One.Defects and Log.Category.Two.Defects have low contribution but are somewhat correlated. Variety on the other hand is isolated.

When compared the sensory biplot we can see that the different grades are scattered with no clear clusteres.

NON-SENSORY FEATURES





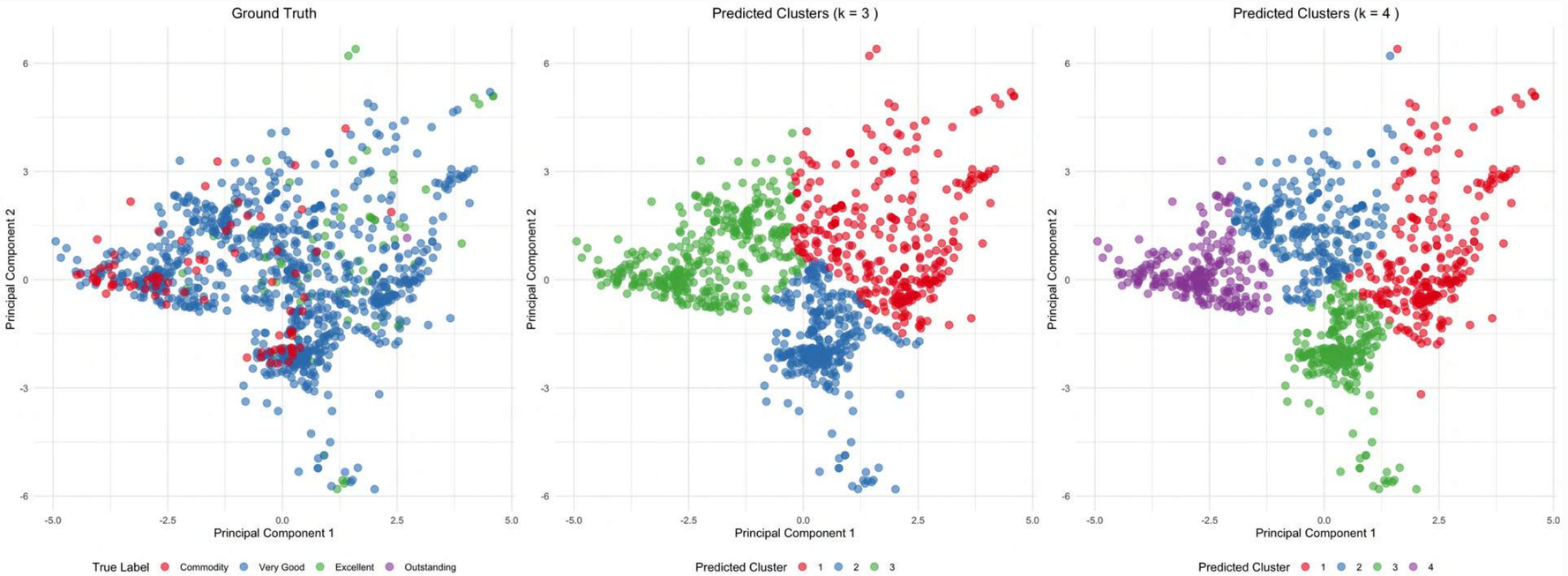


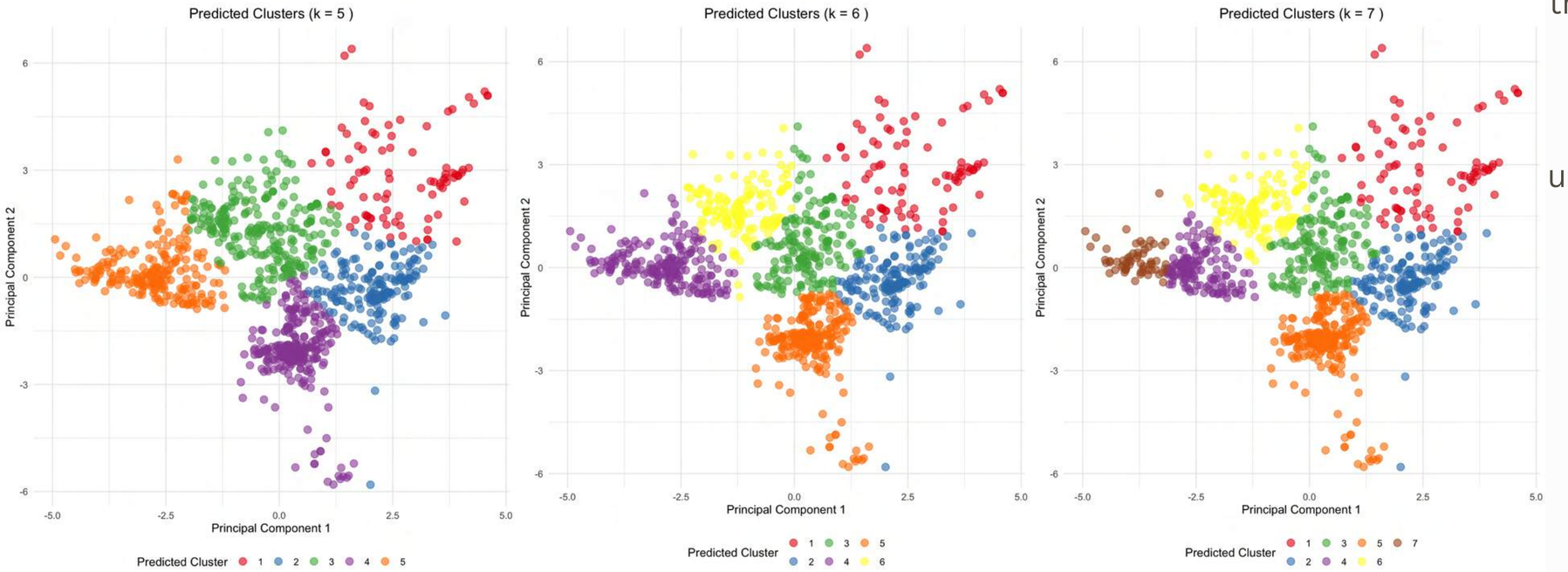
Based on the “elbow method” we can eyeball the optimal number of cluster to be around 5.

For the silhouette method the optimal number of clusters is indicated by the vertical line at 5.

On the other hand the gap statistics shows that the optimal number of clusters to have no clusters, but the highest gap statistics is achieved at 6.

NON-SENSORY FEATURES



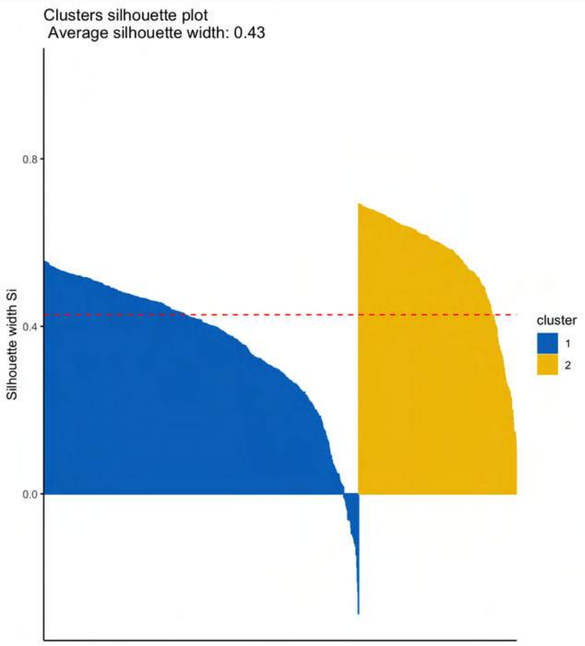


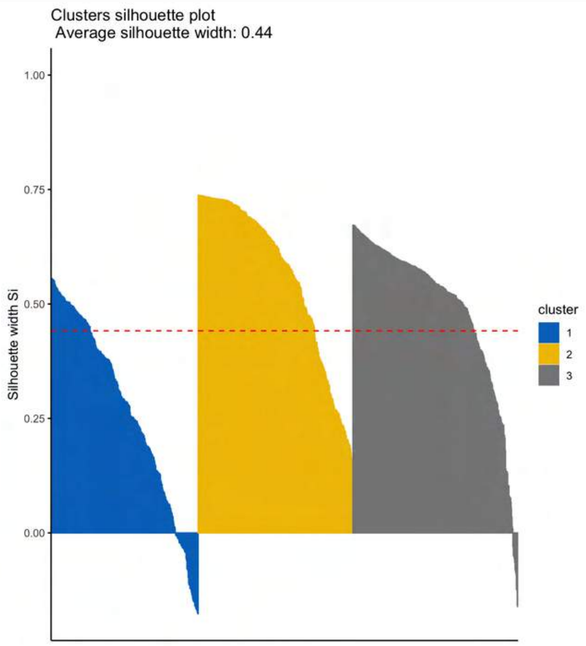
When inspecting the ground truth it is difficult to discern the different classes when compared to the clustering on sensory features.

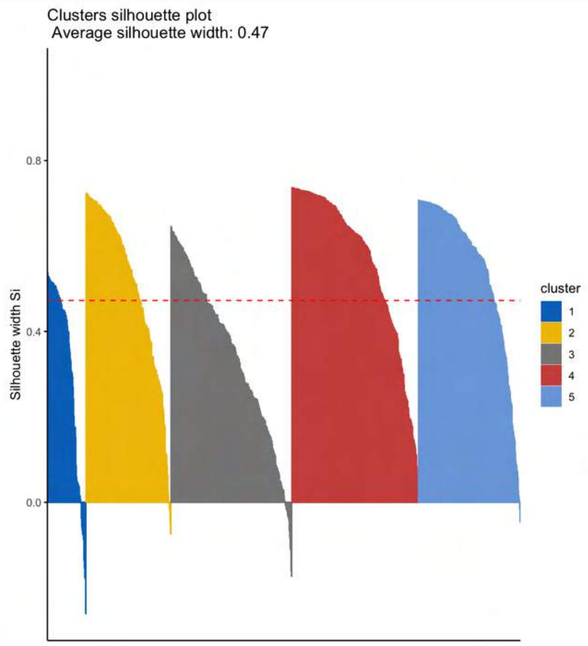
There is an overlap and between the classes that makes it difficult to determine the classes .

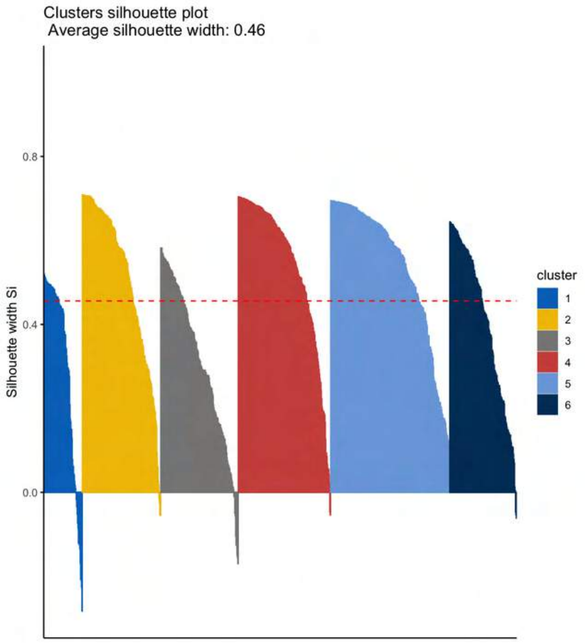
With the clustering different groups are uncovered with main variation in the clustered found to be the country, altitude, minor defects and total weights.

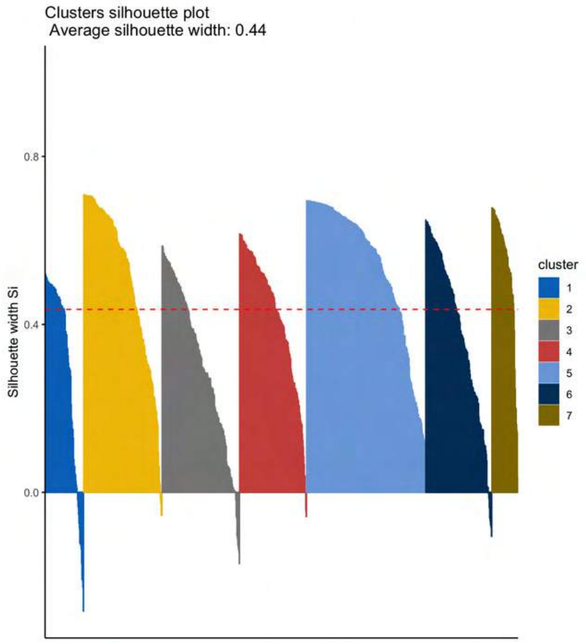
NON-SENSORY FEATURES











The average width increases as we k increases. up to the point of k = 5.

The best number of clusters up to the point of k = 5 based on the average width silhouette.

We can see also that the clusters are above the average score. However it is important to note that some instances are miss clustered in all of the clusteres.

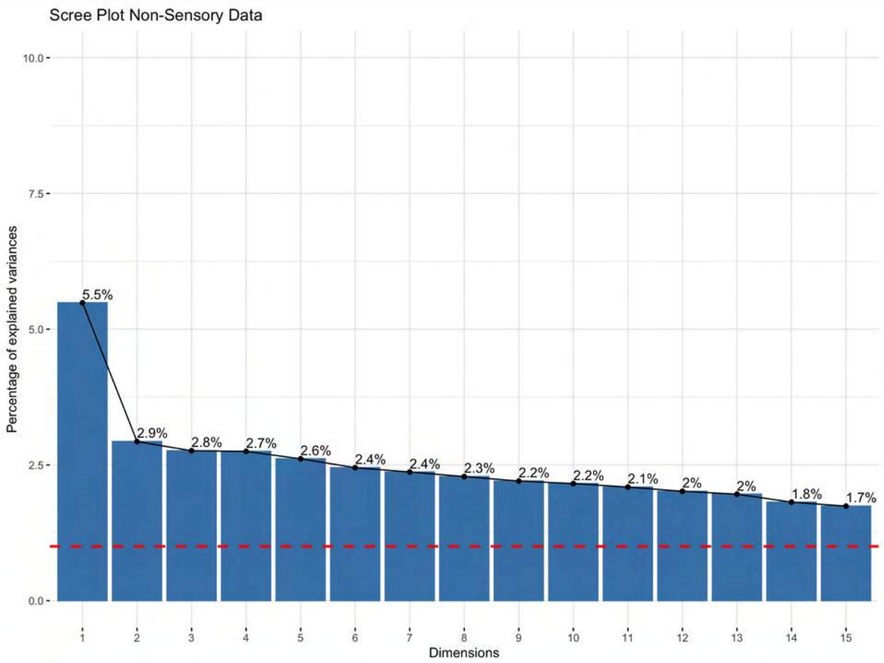
NON-SENSORY FEATURES

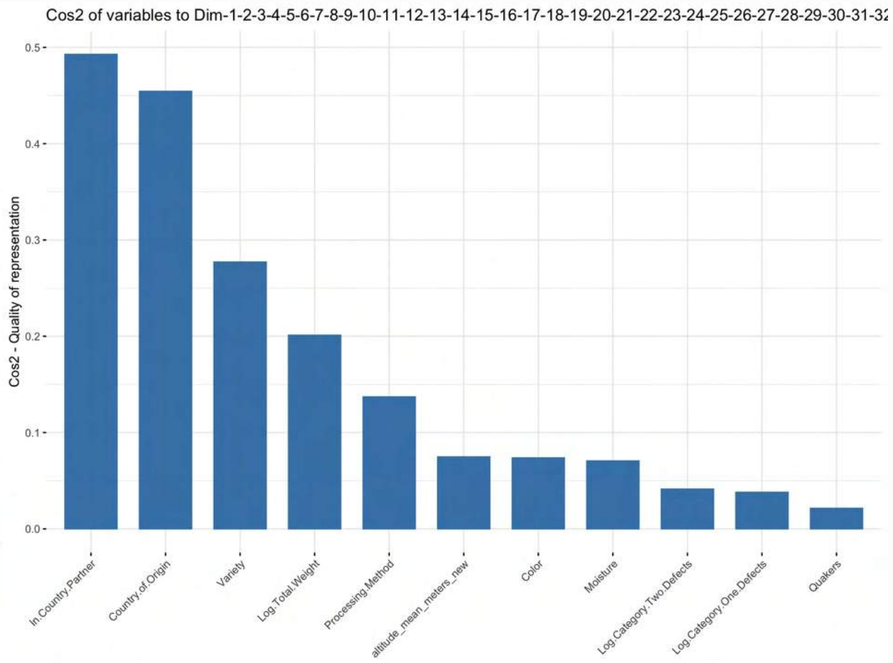
Number of Clusters	Rand Index
2	0.020092127
3	0.003965216
4	0.007000317
5	0.012151546
6	0.009004665
7	0.015780379
8	0.012689104
9	0.014473793
10	0.014473793

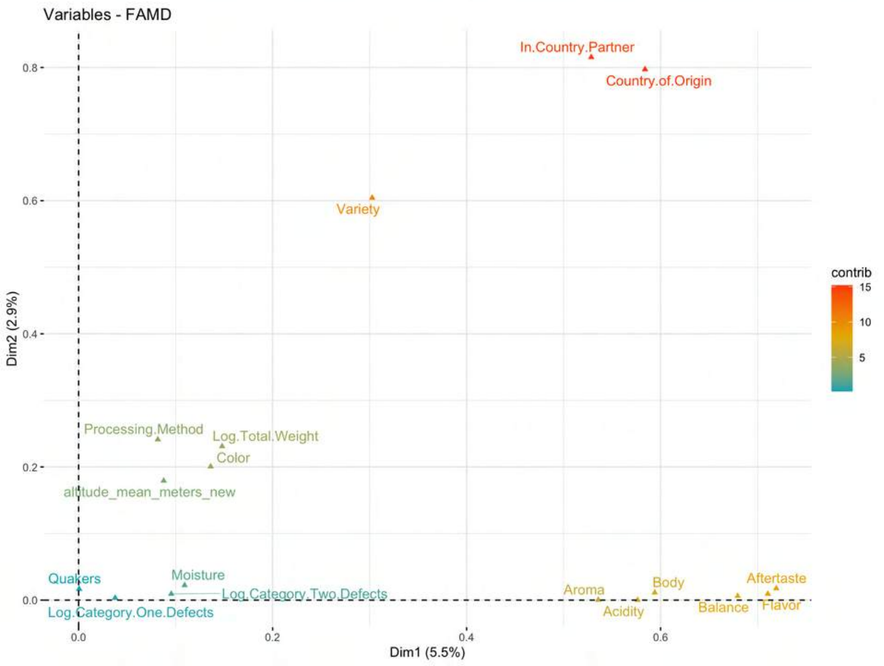
The Rand Index shows that most of the scoreas are low and that none of the clusters are a match of the ground truth.

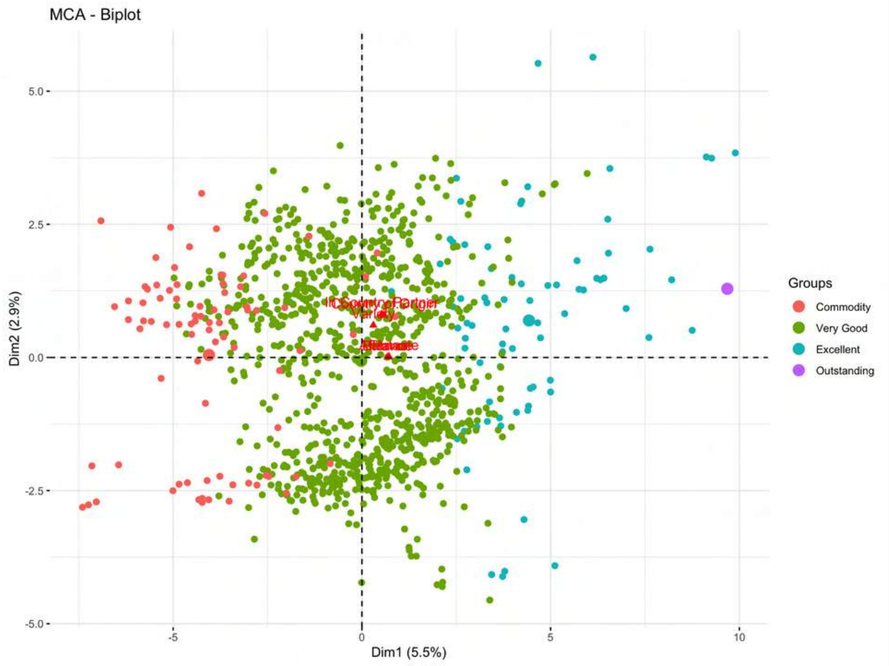
However when k = 7 in this case is relatively the highest, which means that it is closer to the ground truth when compared to other number of clusters.

SENSORY & NON-SENSORY FEATURES







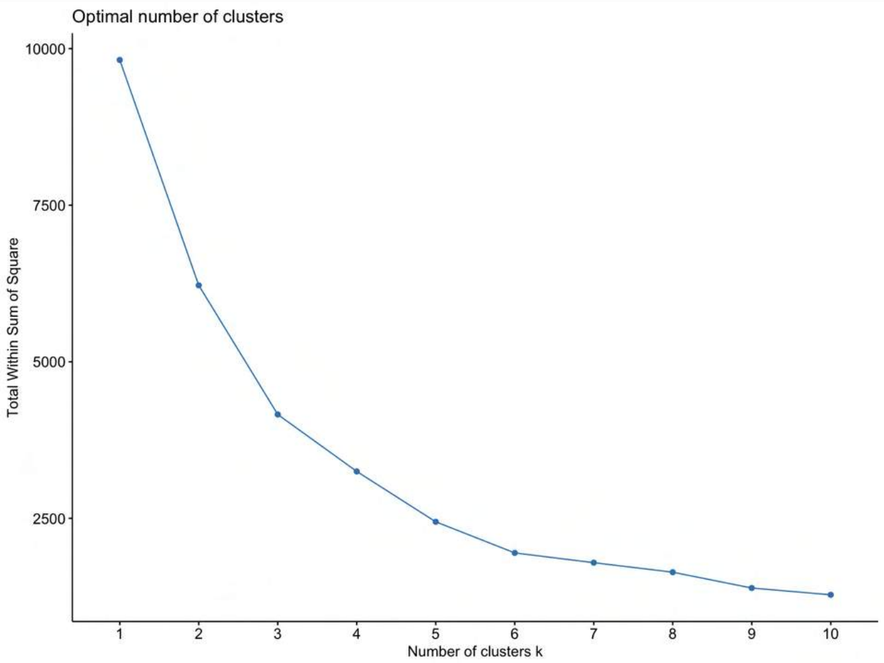


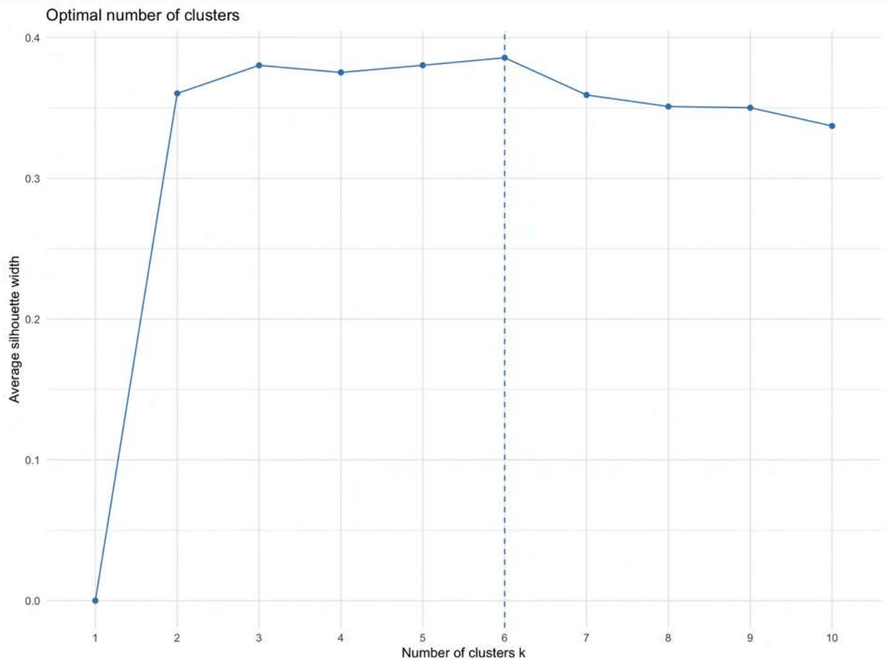
In the scree plot we can see a noticeable decline after the firt few dimensions and then the decline start to slow down.

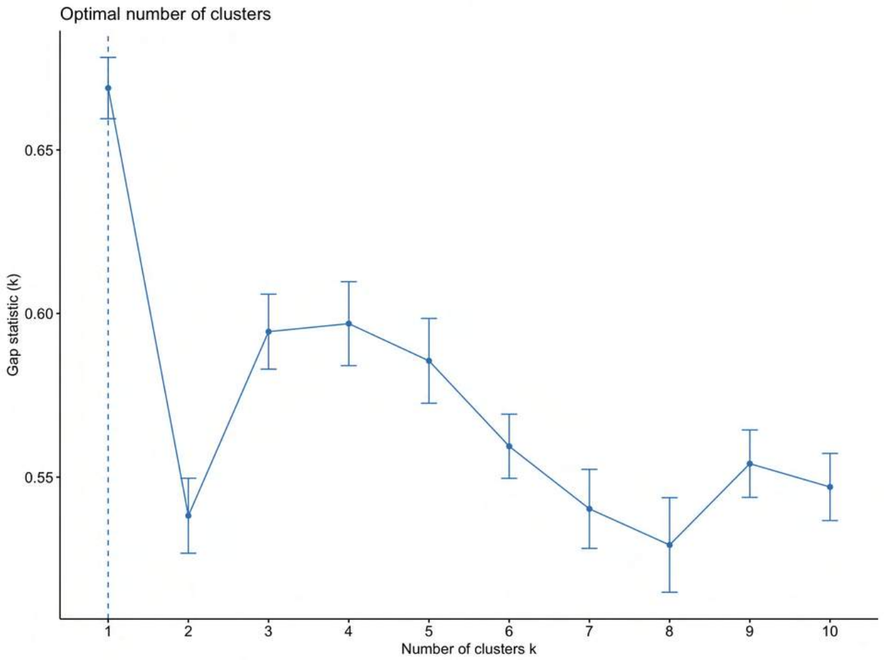
In the biplot we can see that different colors representing different grades are clustered.

When compared to the biplot with just non-sensory data we can see it does better.

SENSORY & NON-SENSORY FEATURES





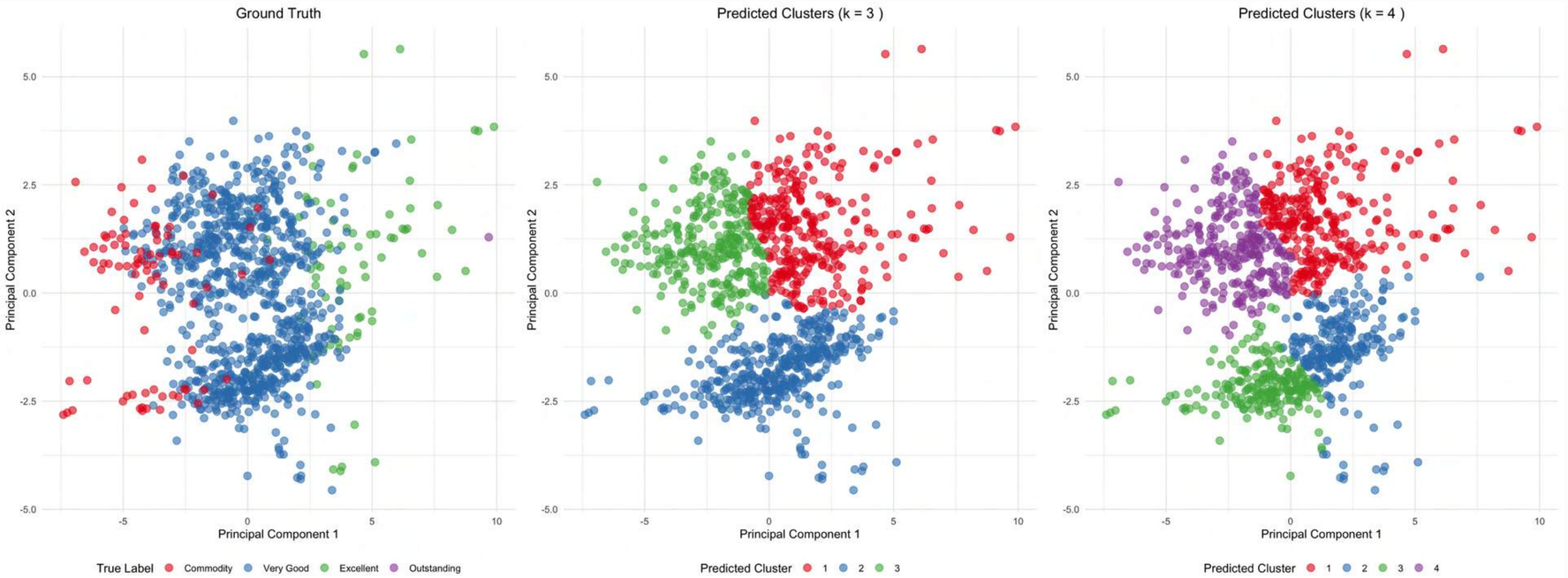


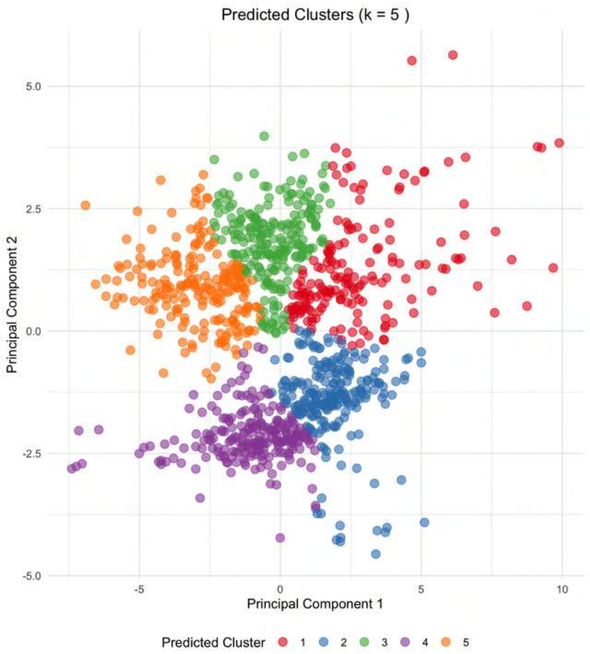
The “elbow method” indicated that the k = 5 or k = 6 could potentially be an optimal number of cluster.

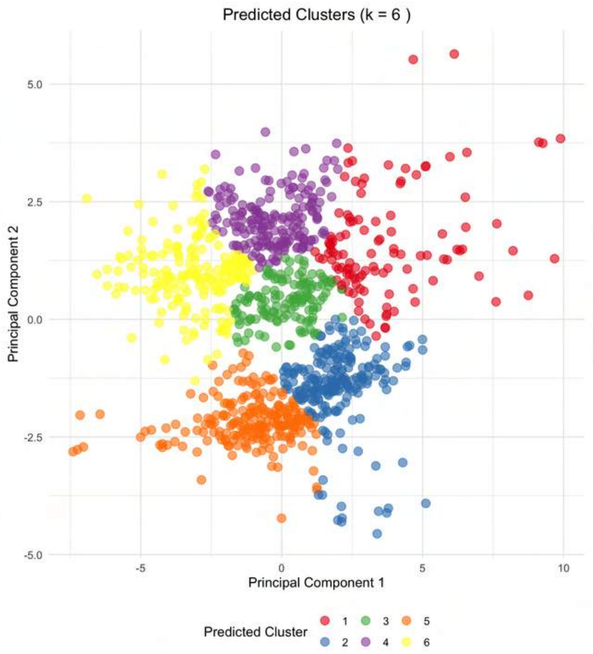
When looking at the silhouette method we can see the optimal number of cluster is represented by 6 number of cluster.

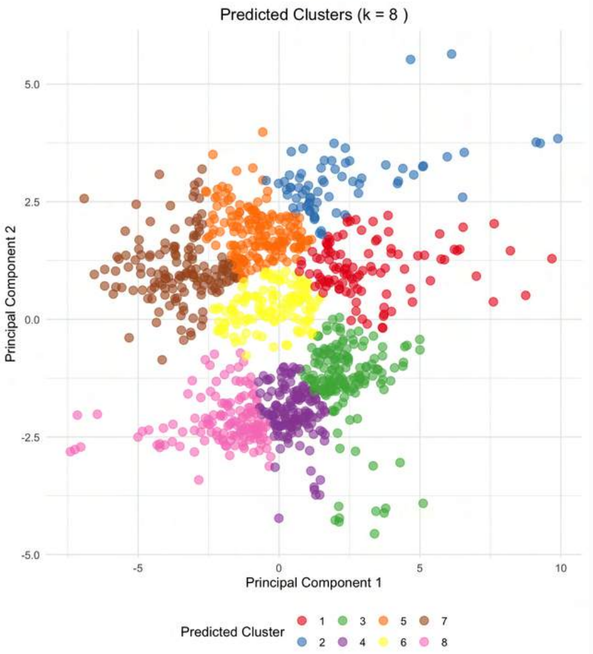
The gap statistics suggests no clusters with. a decreasing gap statistic as k increases.

SENSORY & NON-SENSORY FEATURES





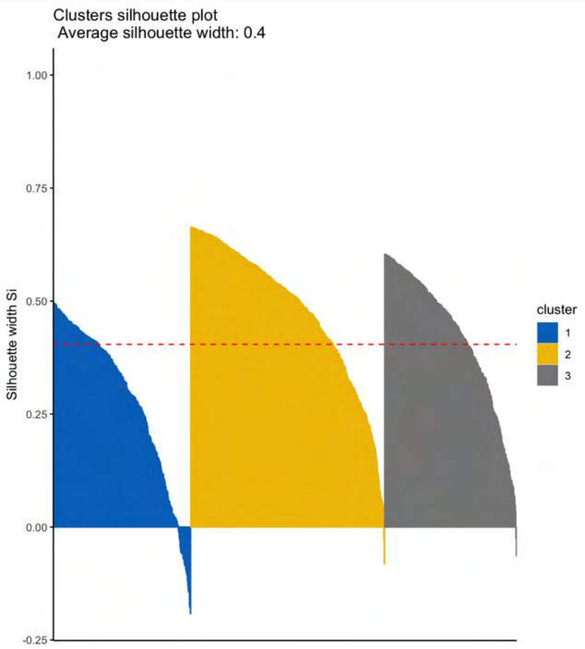


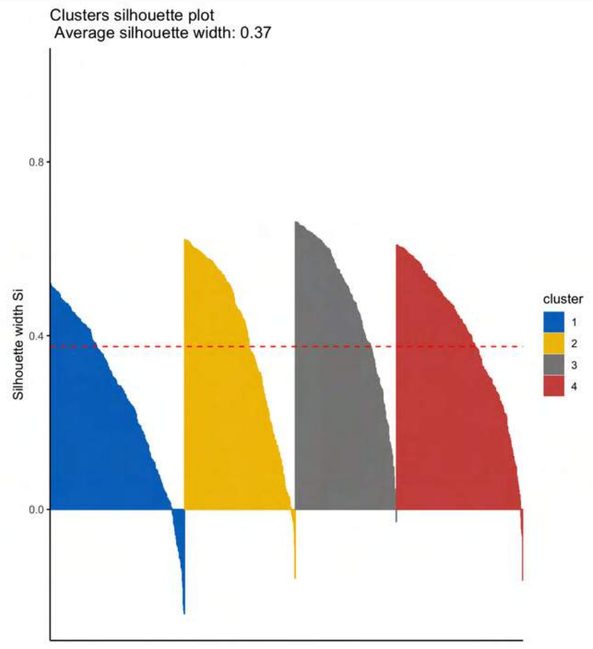


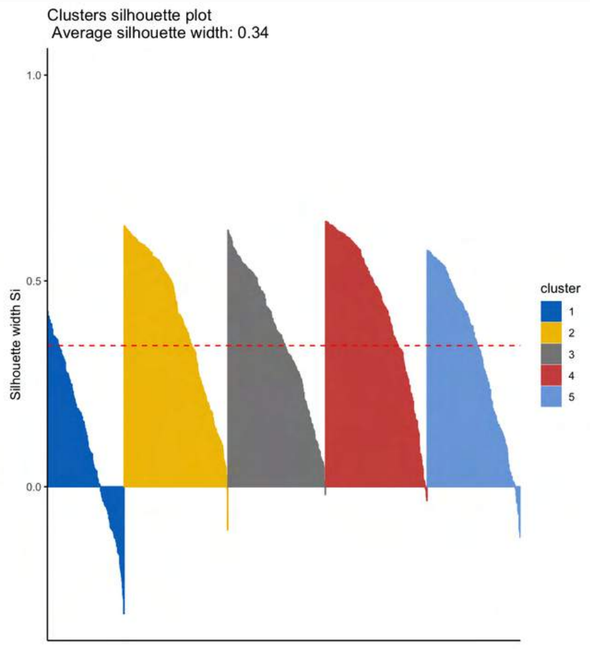
When compared to the ground truth we can see some inconsistencies in how they are being clustered.

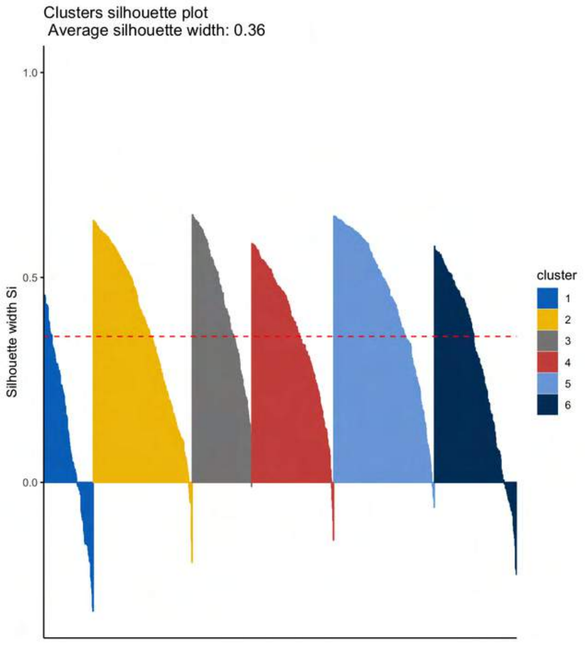
As we increase the number of clustering we can see new groupings being discovered.

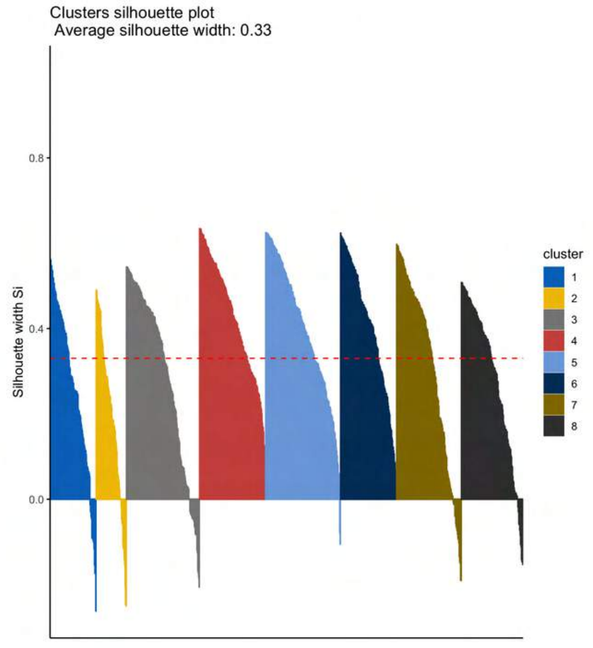
SENSORY & NON-SENSORY FEATURES











When the clusters are two or three the width are very low with 0.28 and 0.35 respectively. This is the same as the case when we build the clustering on only the Sensory.

However for the fourth and the fifth clustering we can a slight decrease in the average width silhouette.

Another important thing to note is that in all of the stages of the clustering there are data point that are being classified.

Especially for when k = 6 we can see that there are many data points that are missclassifed in the second cluster.

SENSORY & NON-SENSORY FEATURES

Number of Clusters	Rand Index
2	-0.007564600
3	0.024357062
4	0.005681346
5	0.026914014
6	0.031724289
7	0.026168363
8	0.023685378
9	0.020269551
10	0.024015375

The Rand Index shows that most of the scoreas are low and that none of the clusters are a match of the ground truth.

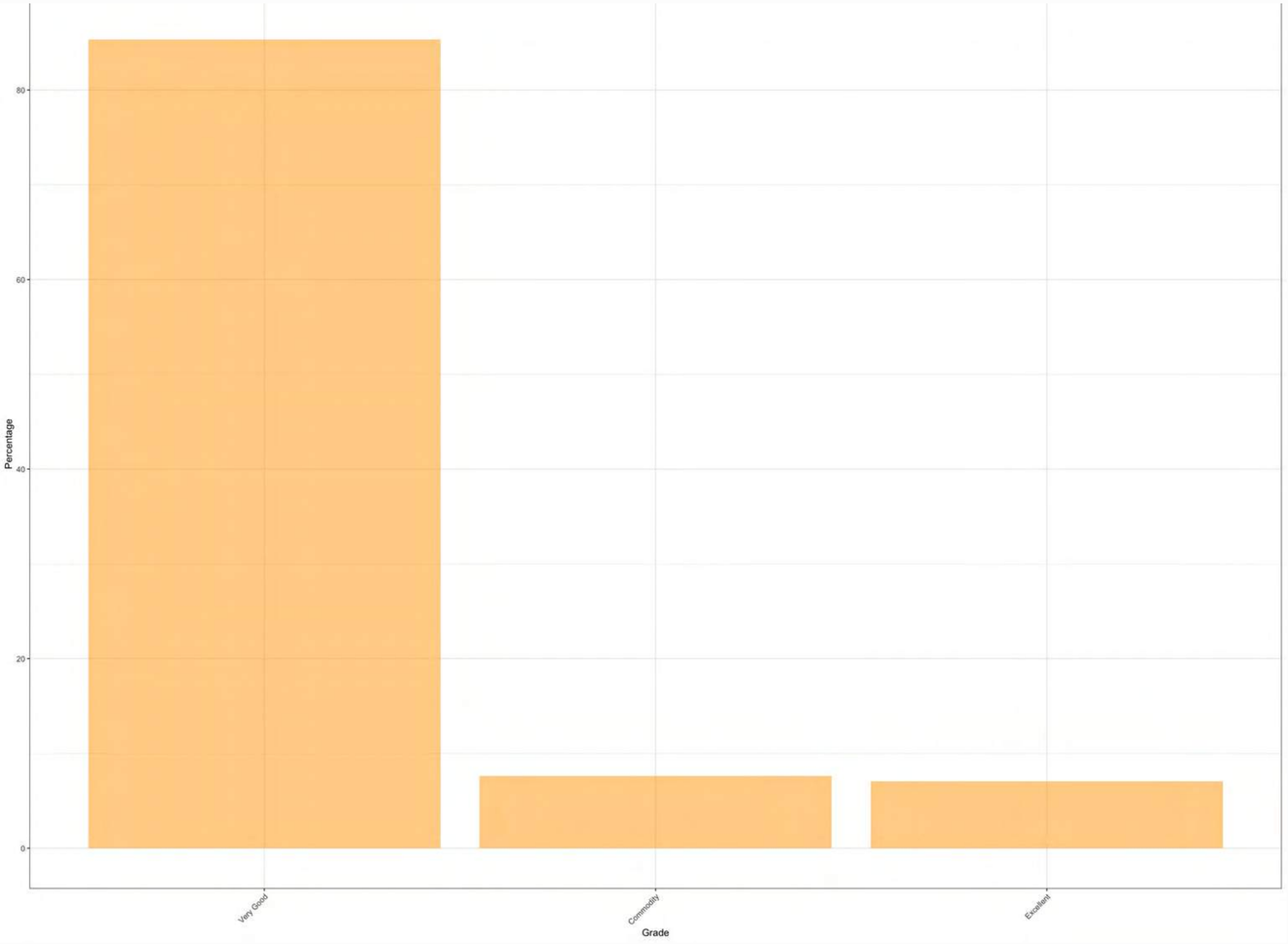
However when k = 6 in this case is relatively the highest, which means that it is closer to the ground truth when compared to other number of clusters.

SUPERVISED

These are the questions we would like to answer:

1. Determine the quality of coffee using non-sensory data that are based on objective determinants of coffee.
2. Identifying the most important non-sensory characteristics that contribute to coffee quality of beans?
3. Is there a model that can accurately predict the grade of coffee bean based on the characteristics?

SUPERVISED



We shuffle the indices then split the trining data with 70% of the data accounting for the training set and 30% for the test data.

Since our dataset is imbalanced we used oversampling, SMOTE and ROSE for the minority class levels, Commodity and Excellent.

OVERSAMPLING

Oversampling/Grade	**Commodity**	**Very Good**	**Excellent**
Initial	58	644	54
Over	644	644	644
SMOTE	594	655	425
ROSE	528	644	484

The Initial model is included for comparison and represents the original data set with imbalanced data set.

The Over technique uses the upSample function where it creates instances for the dataset by randomly sampling from the minority class until it matches the same number as the majority class.

The SMOTE technique creates synthetic instances from k nearest neighbours of minorities instances.

The ROSE oversampling technique creates a synthetic data point from the minority it selected by introducing a random perturbation based on the standard deviation of the features.

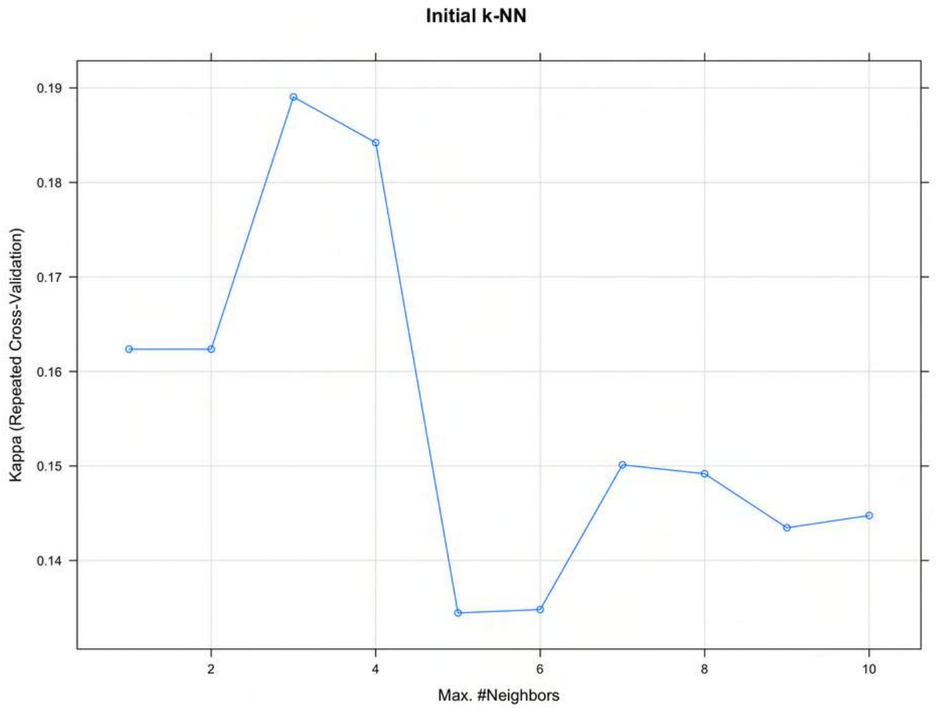
K - N N

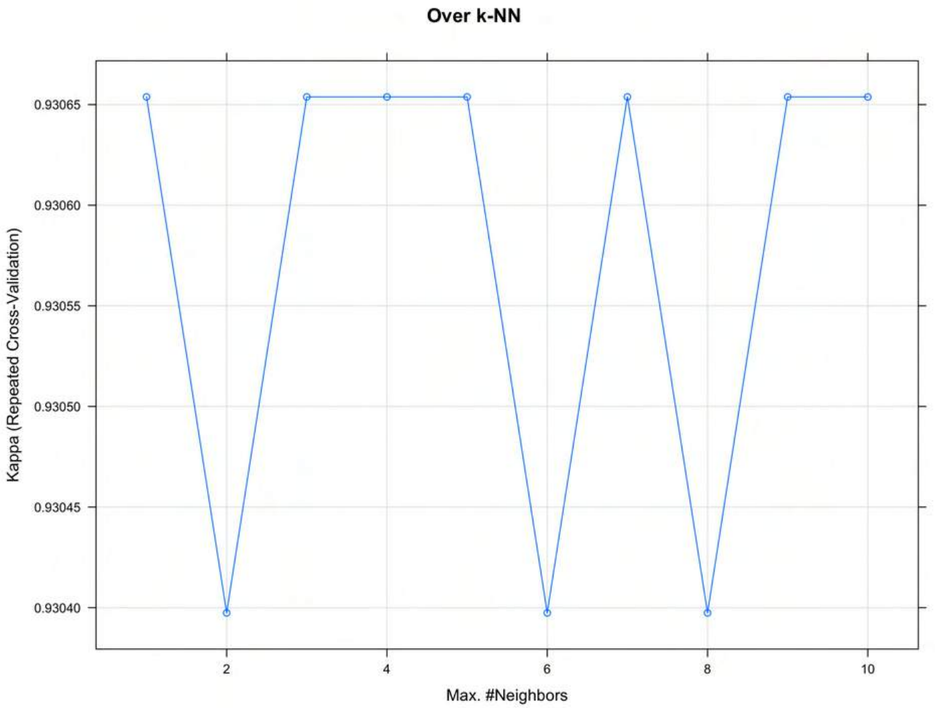
Our first algorithm is a lazy learner with k-Nearest Neighbour algorithm to classify the quality of the coffee beans.

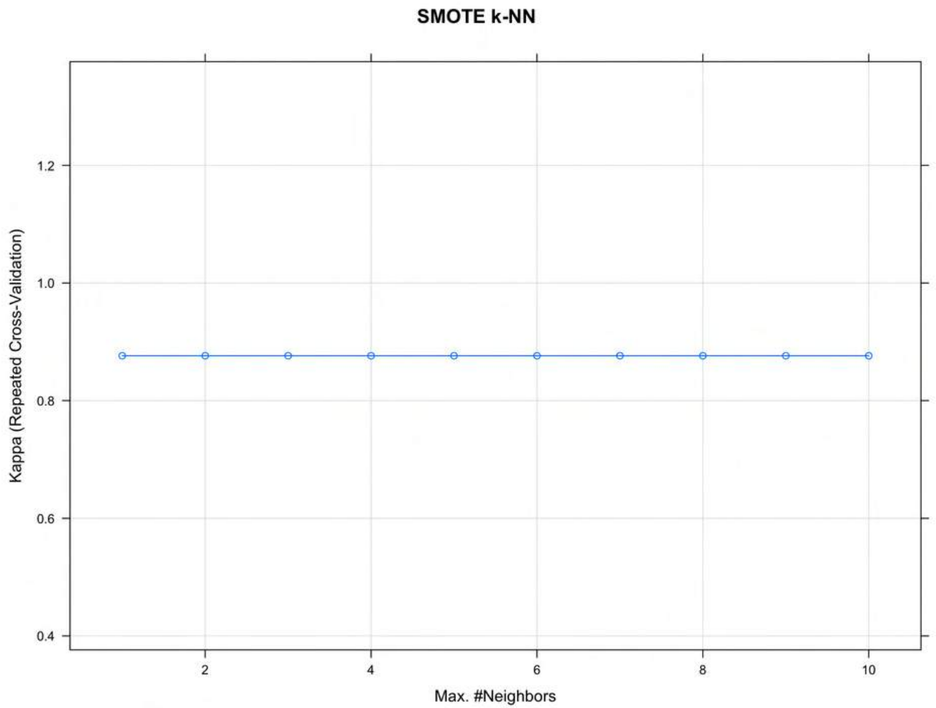
This means we will be using the nearest data points in the feature space to make a classification.

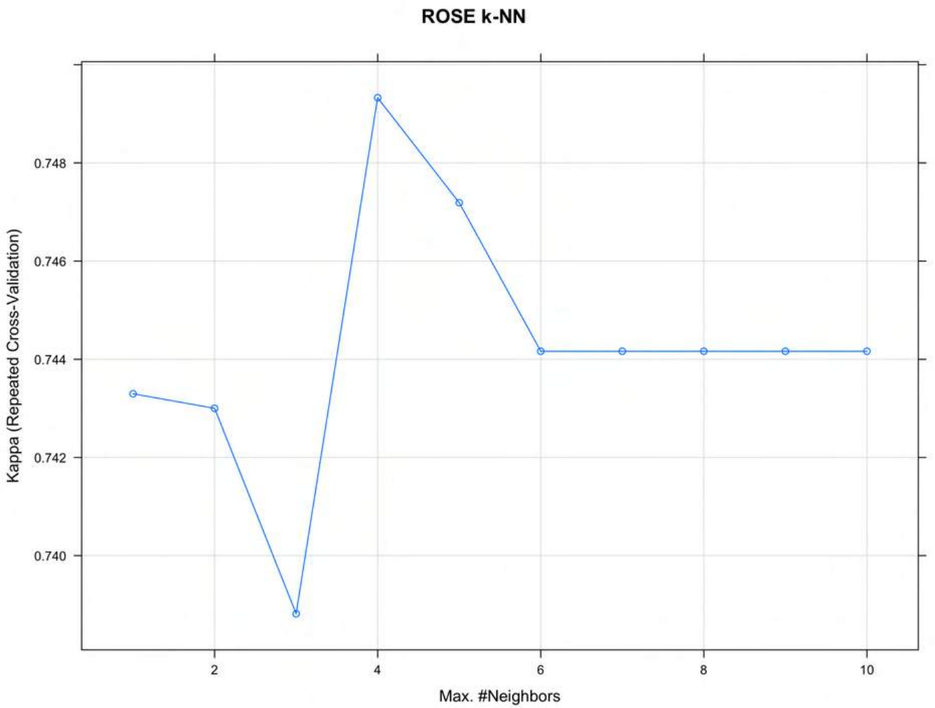
We have used performed a grid search over the ranges to determine the best ‘k’ values for each of the data set including the imbalanced data set and the oversampled data.

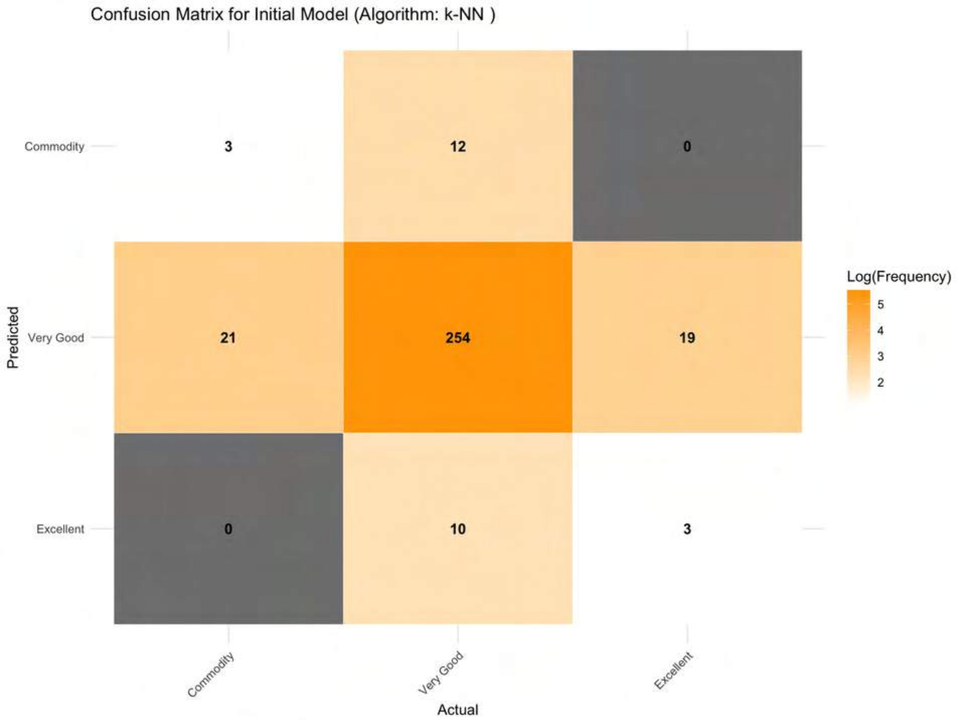
We have evaluated the models using repeated cross-validation.

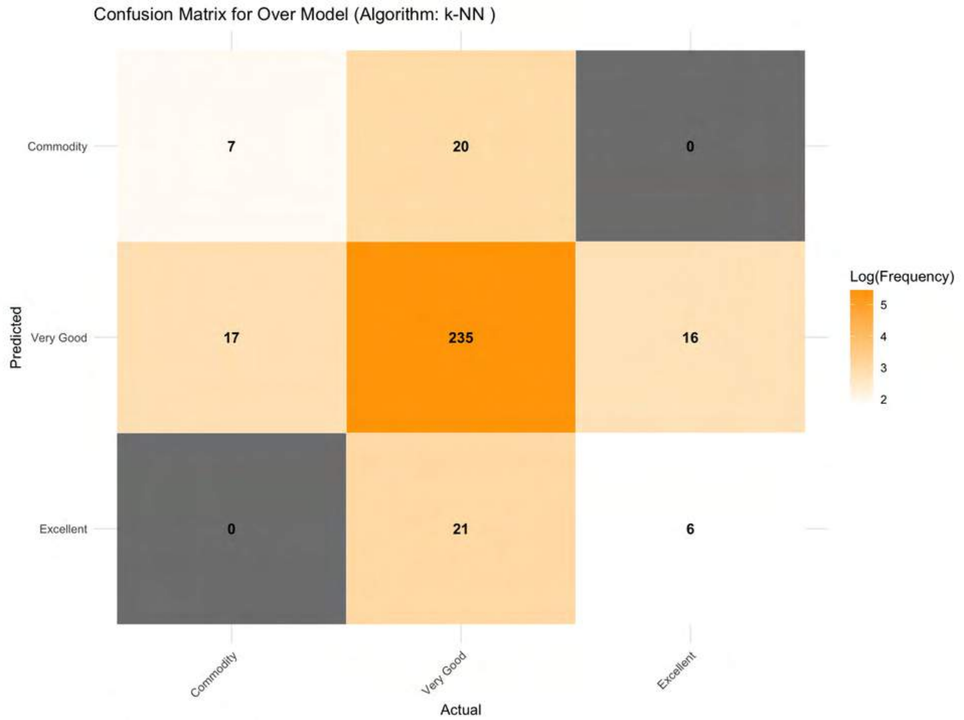

Initial k-NN
Kappa (Repeated Cross-Validation)
0.19
0.18
0.17
0.16
0.15
0.14
2
4
6
8
10
Max. #Neighbors

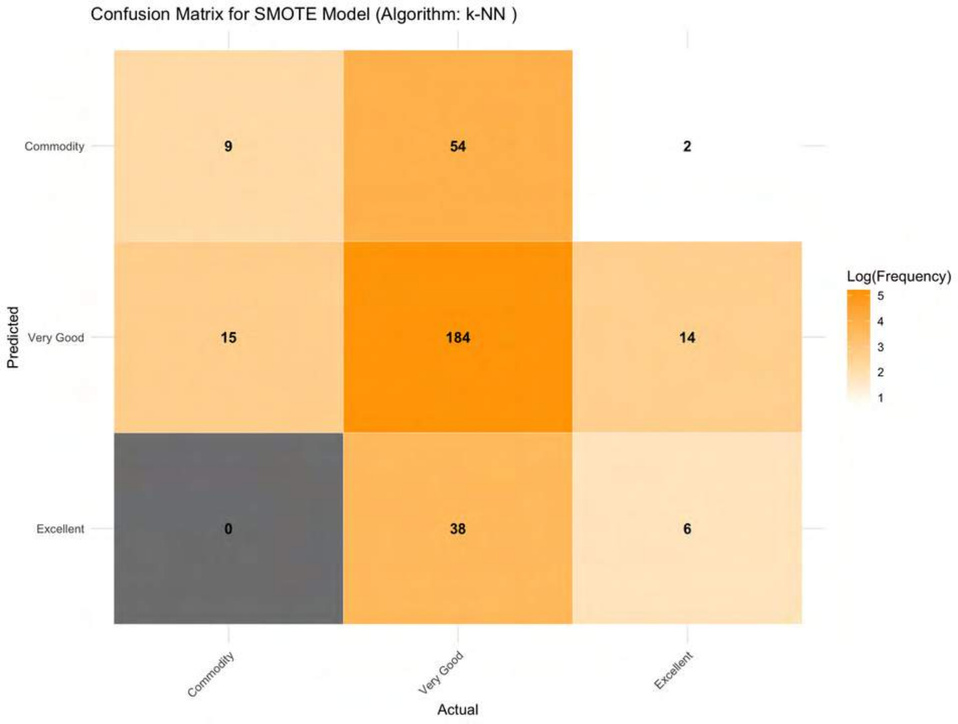

Over k-NN
Kappa (Repeated Cross-Validation)
0.93065
0.93060
0.93055
0.93050
0.93045
0.93040
2
4
6
8
10
Max. #Neighbors

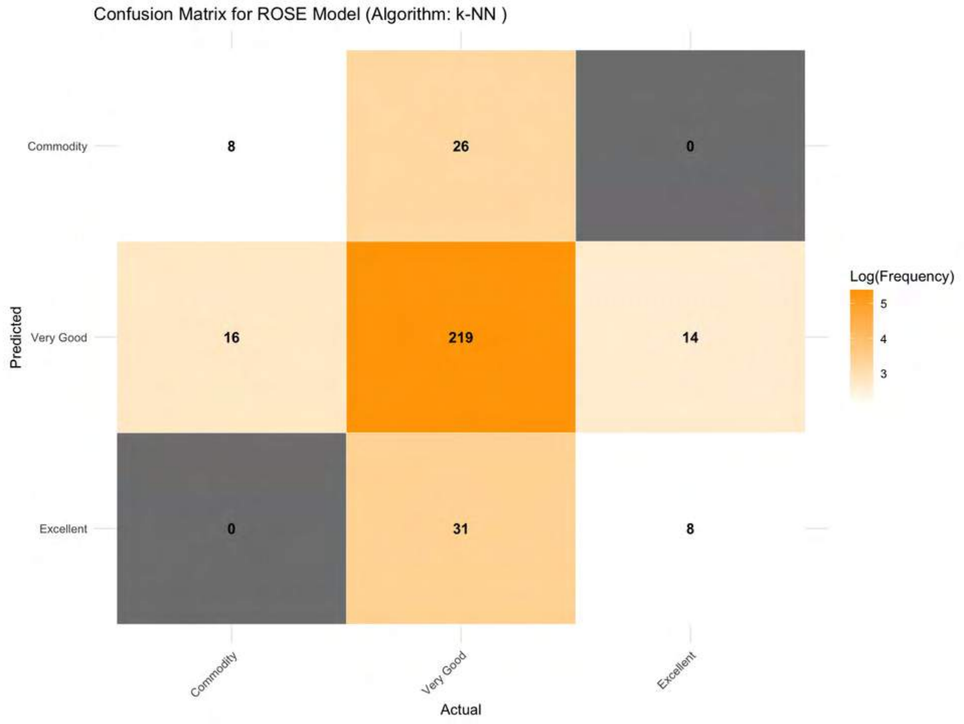

SMOTE k-NN
Kappa (Repeated Cross-Validation)
1.2
1.0
0.8
0.6
0.4
2
4
6
8
10
Max. #Neighbors


ROSE k-NN
Kappa (Repeated Cross-Validation)
0.748
0.746
0.744
0.742
0.740
2
4
6
8
10
Max. #Neighbors


Confusion Matrix for Initial Model (Algorithm: k-NN)
Commodity
Very Good
Excellent
Predicted
3
12
0
21
254
19
0
10
3
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2

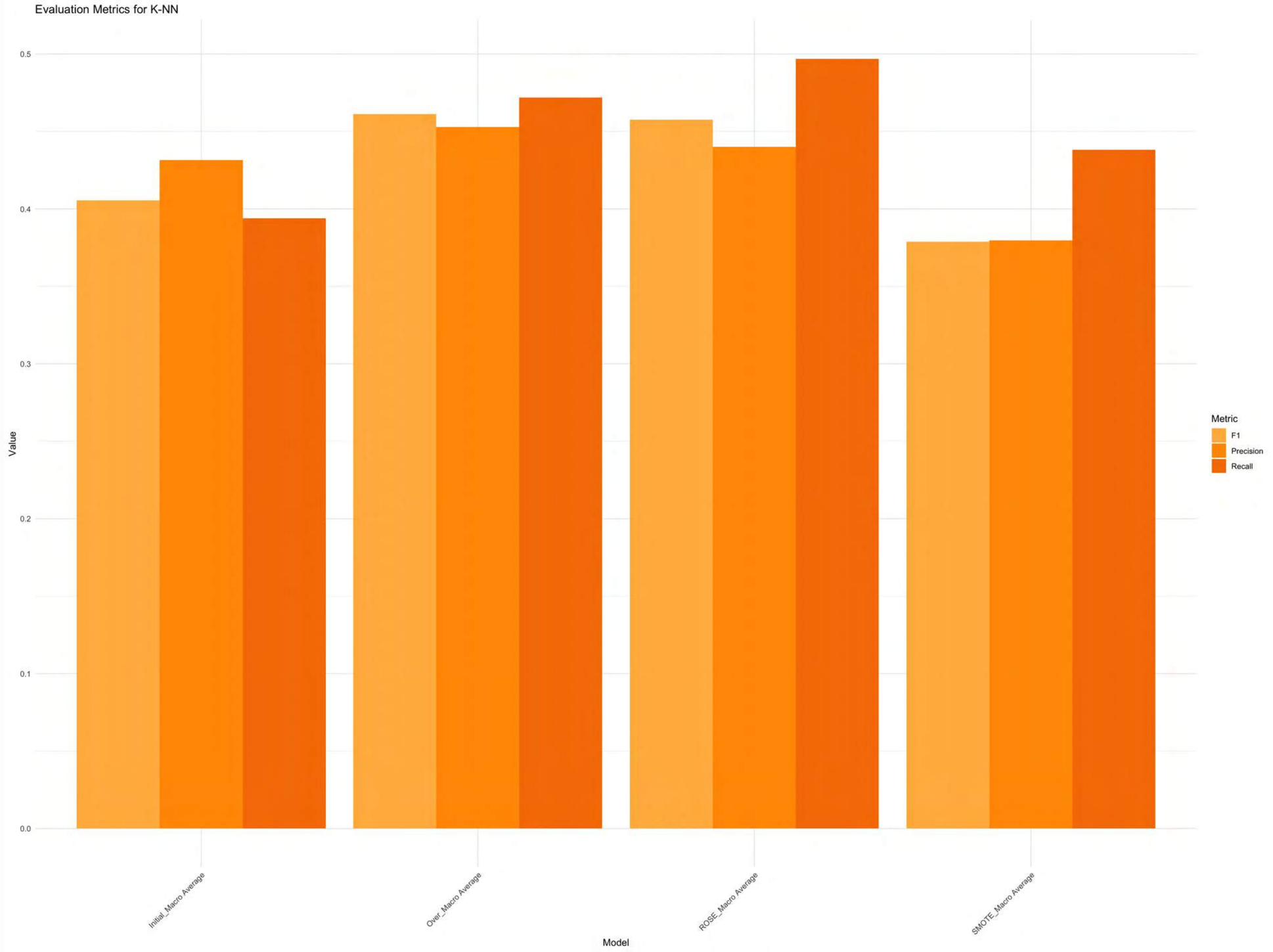

Confusion Matrix for Over Model (Algorithm: k-NN)
Commodity
Very Good
Excellent
Predicted
7
20
0
17
235
16
0
21
6
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2

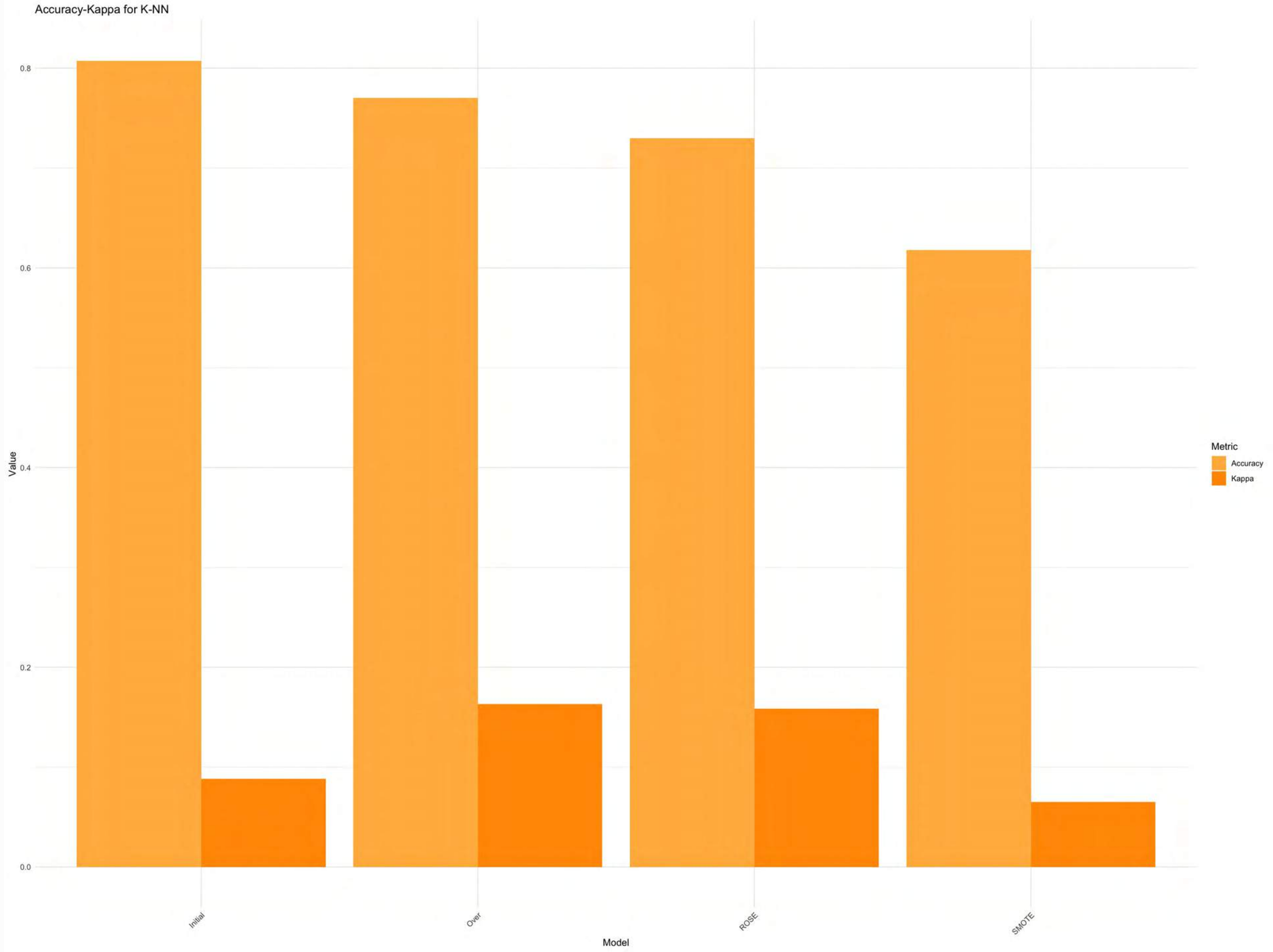

Confusion Matrix for SMOTE Model (Algorithm: k-NN)
Commodity
Very Good
Excellent
Predicted
9
54
2
15
184
14
0
38
6
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2
1

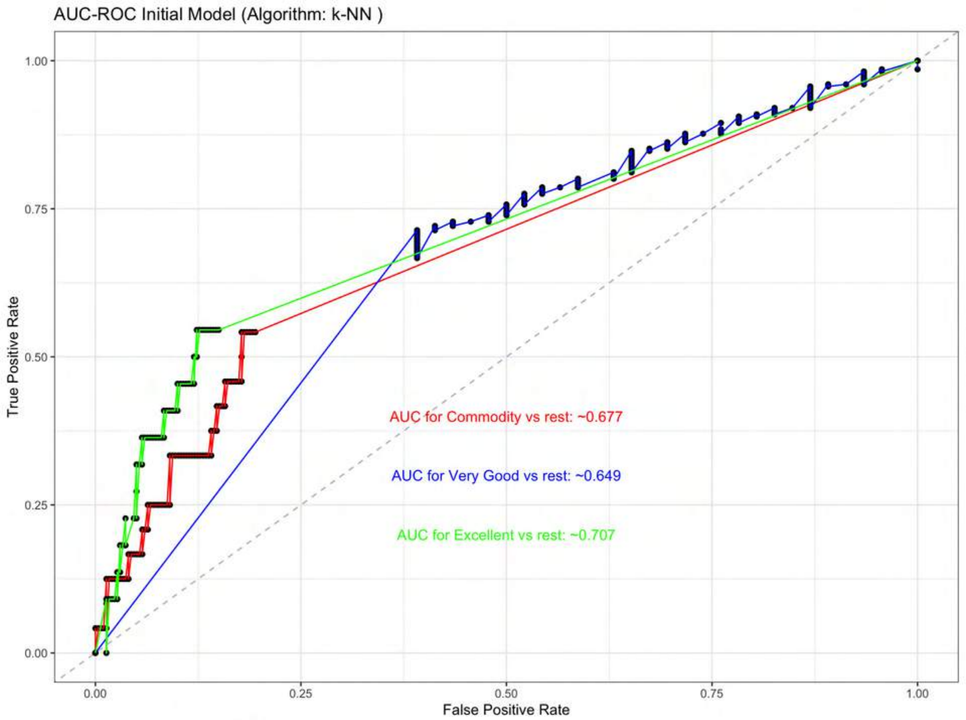

Confusion Matrix for ROSE Model (Algorithm: k-NN)
Commodity
Very Good
Excellent
Predicted
8
26
0
16
219
14
0
31
8
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3

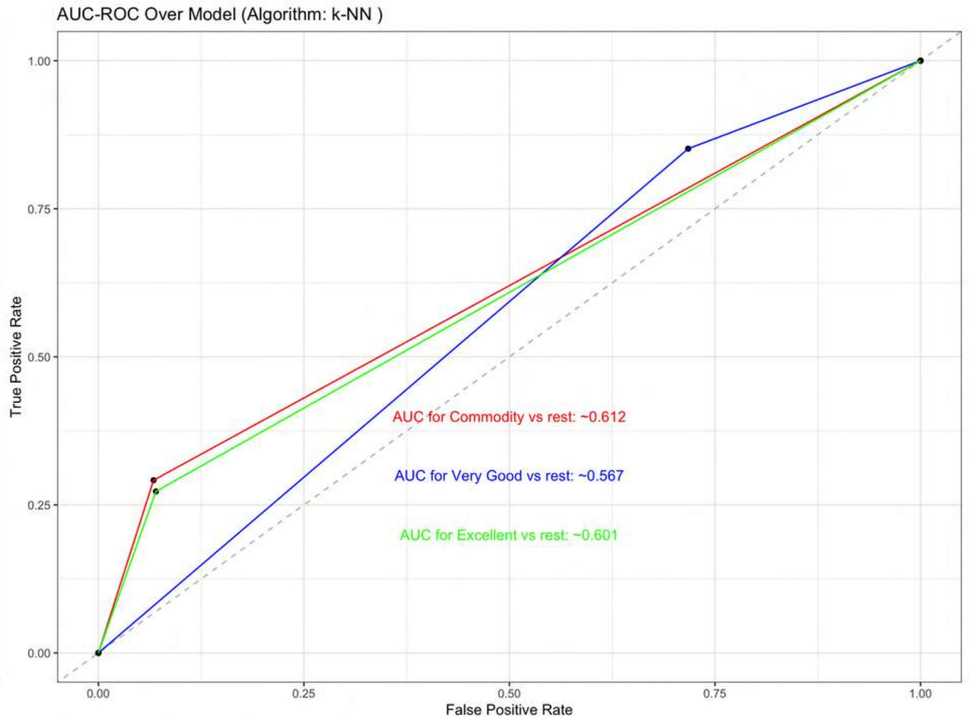
K-NN EVALUATED METRICS

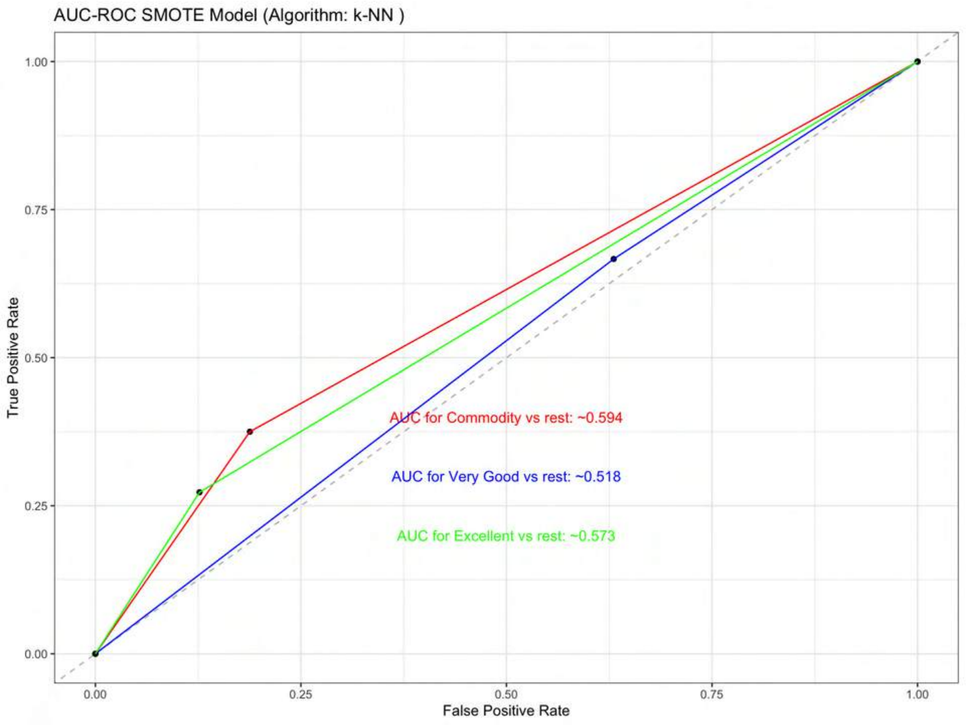
Model/Metric	Macro Precision	Macro Recall	Macro F1	Acuracy	Kappa
Initial	0.4315716	0.3938845	0.4055009	0.8074534	0.08815201
Over	0.4527824	0.4719477	0.4611261	0.7701863	0.16316640
SMOTE	0.3795583	0.4381313	0.3788739	0.6180124	0.06519071
ROSE	0.4399801	0.4968160	0.4574810	0.7298137	0.15838491

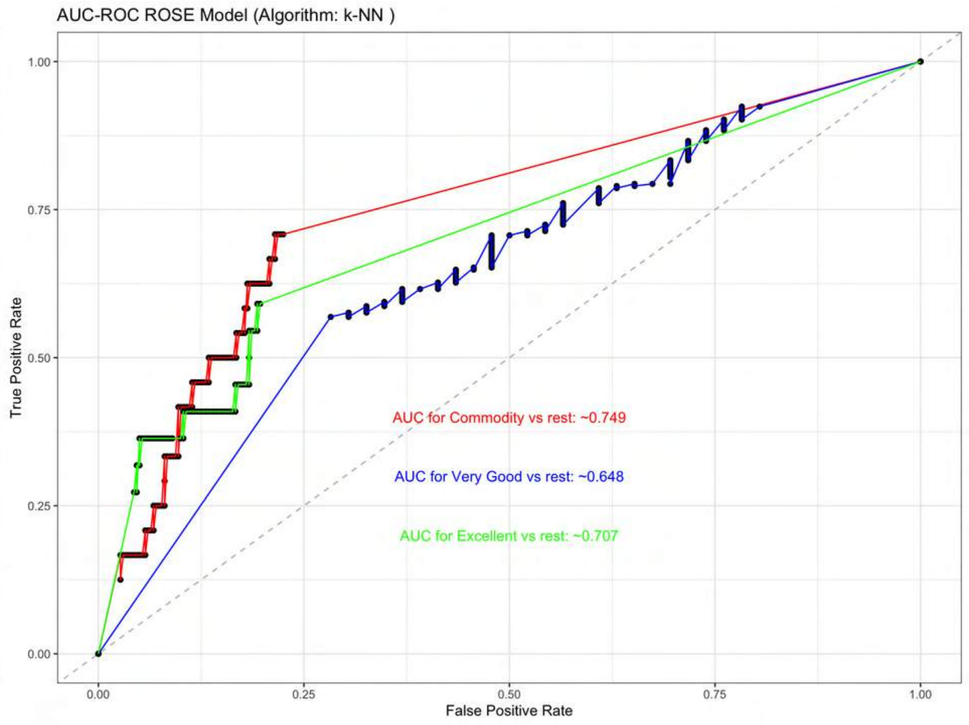

Evaluation Metrics for K-NN
0.5
0.4
0.3
0.2
0.1
0.0
Value
Initial_Macro Average
Over_Macro Average
ROSE_Macro Average
SMOTE_Macro Average
Model
Metric
F1
Precision
Recall


Accuracy-Kappa for K-NN
0.8
0.6
0.4
0.2
0.0
Value
Initial
Over
ROSE
SMOTE
Model
Metric
Accuracy
Kappa









MEAN AUC-ROC

INITIAL: 0.67

OVER: 0.59

SMOTE: 0.56

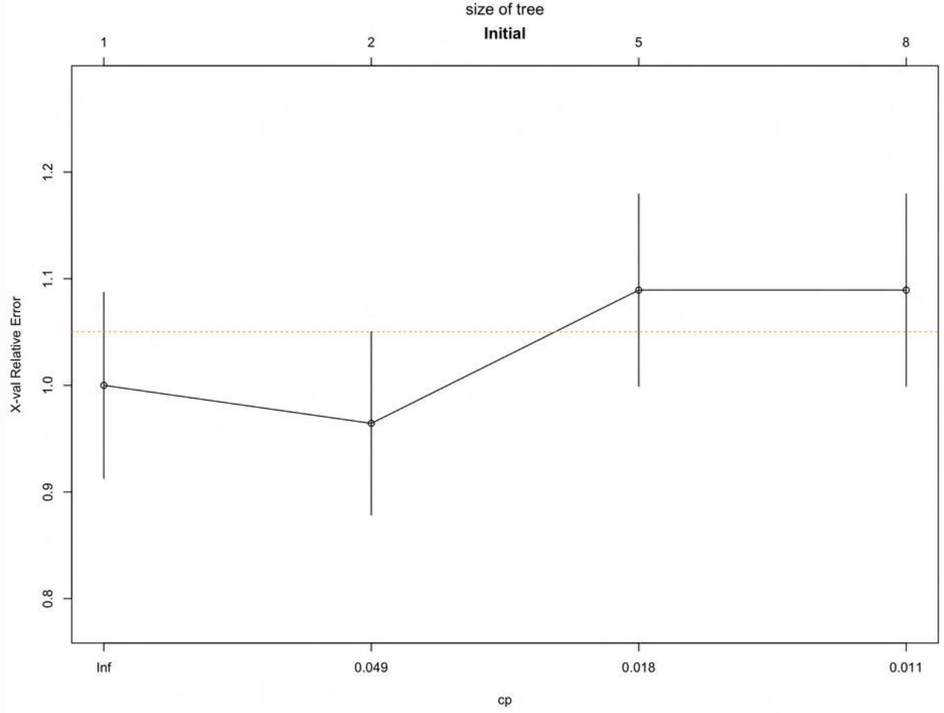
ROSE: 0.70

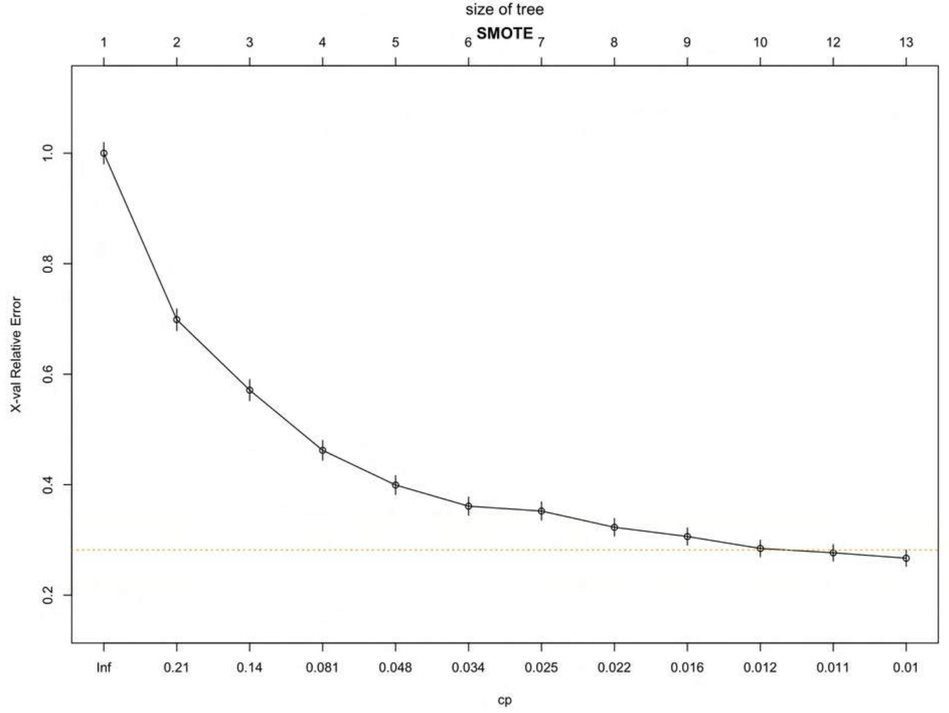
DECISION TREE

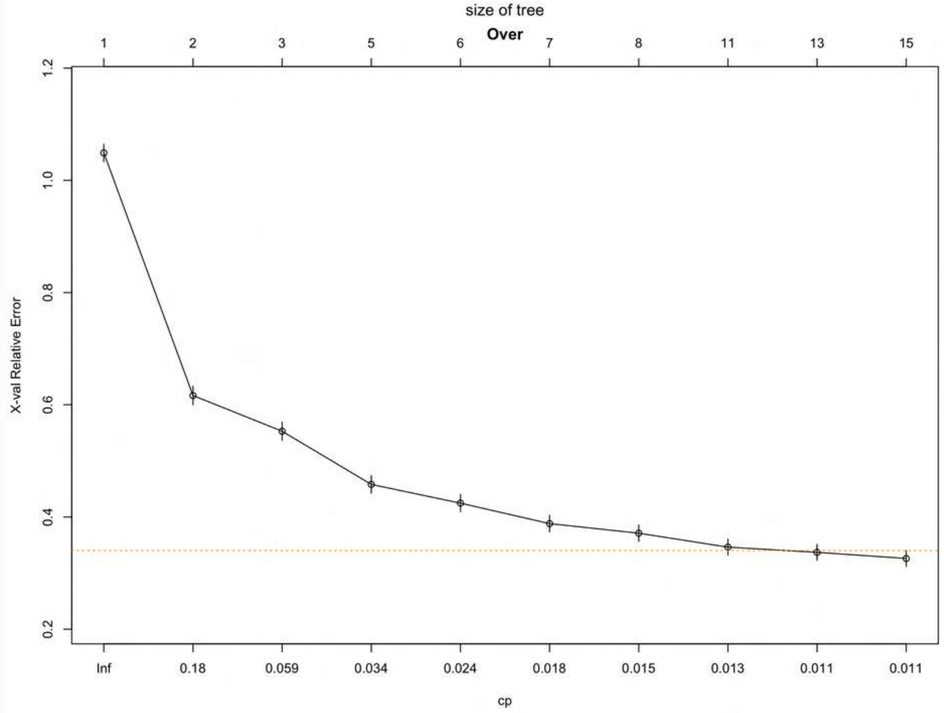
Our second algorithm is decision tree to classify the grade of the coffee.

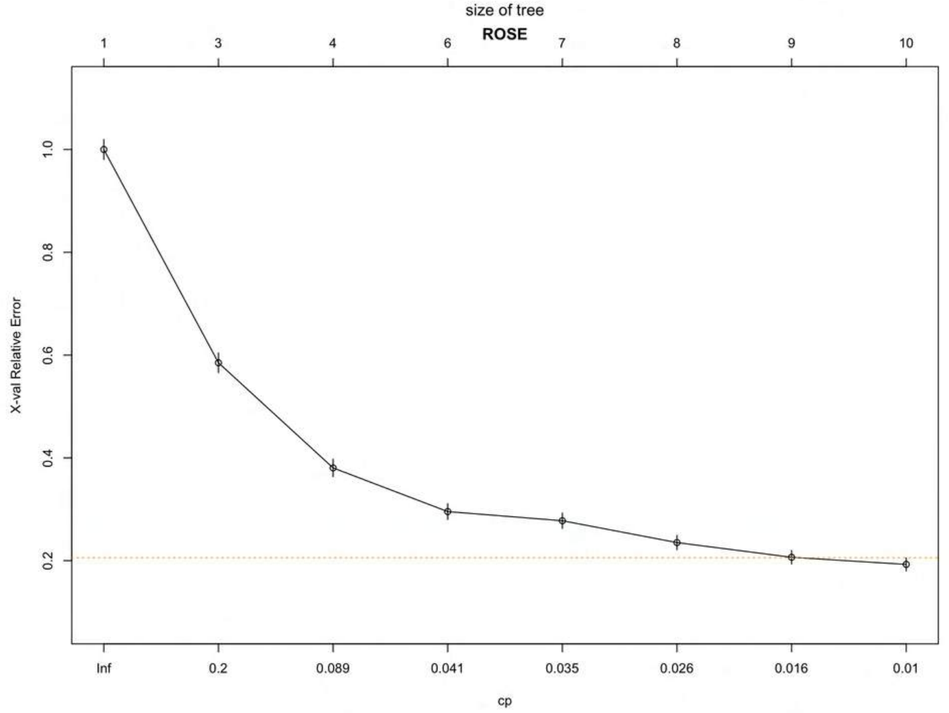
This allowed us to determine the important features and could be potentially useful when making decisions for traders and roasters to gain insight on different grades of coffee.

Since decision trees are prone to overfitting we have built the tree and used the complexity parameter (cp) to prune our model by minimizing the cross-validation error in the model's complexity table.

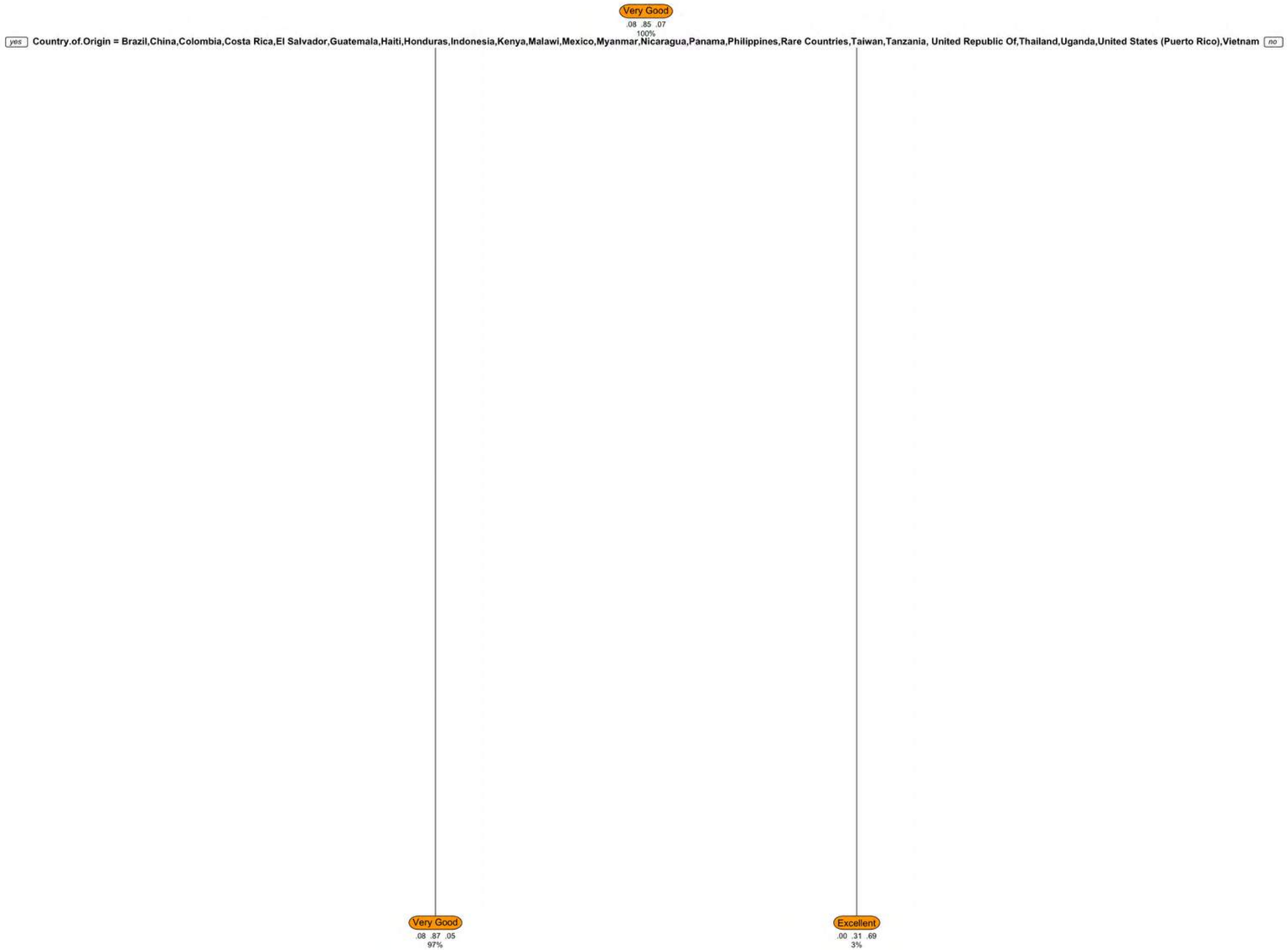

size of tree
Initial
1
2
5
8
X-val Relative Error
0.8
0.9
1.0
1.1
1.2
Inf
0.049
0.018
0.011
cp


size of tree
SMOTE
1
2
3
4
5
6
7
8
9
10
12
13
X-val Relative Error
0.2
0.4
0.6
0.8
1.0
Inf
0.21
0.14
0.081
0.048
0.034
0.025
0.022
0.016
0.012
0.011
0.01
cp

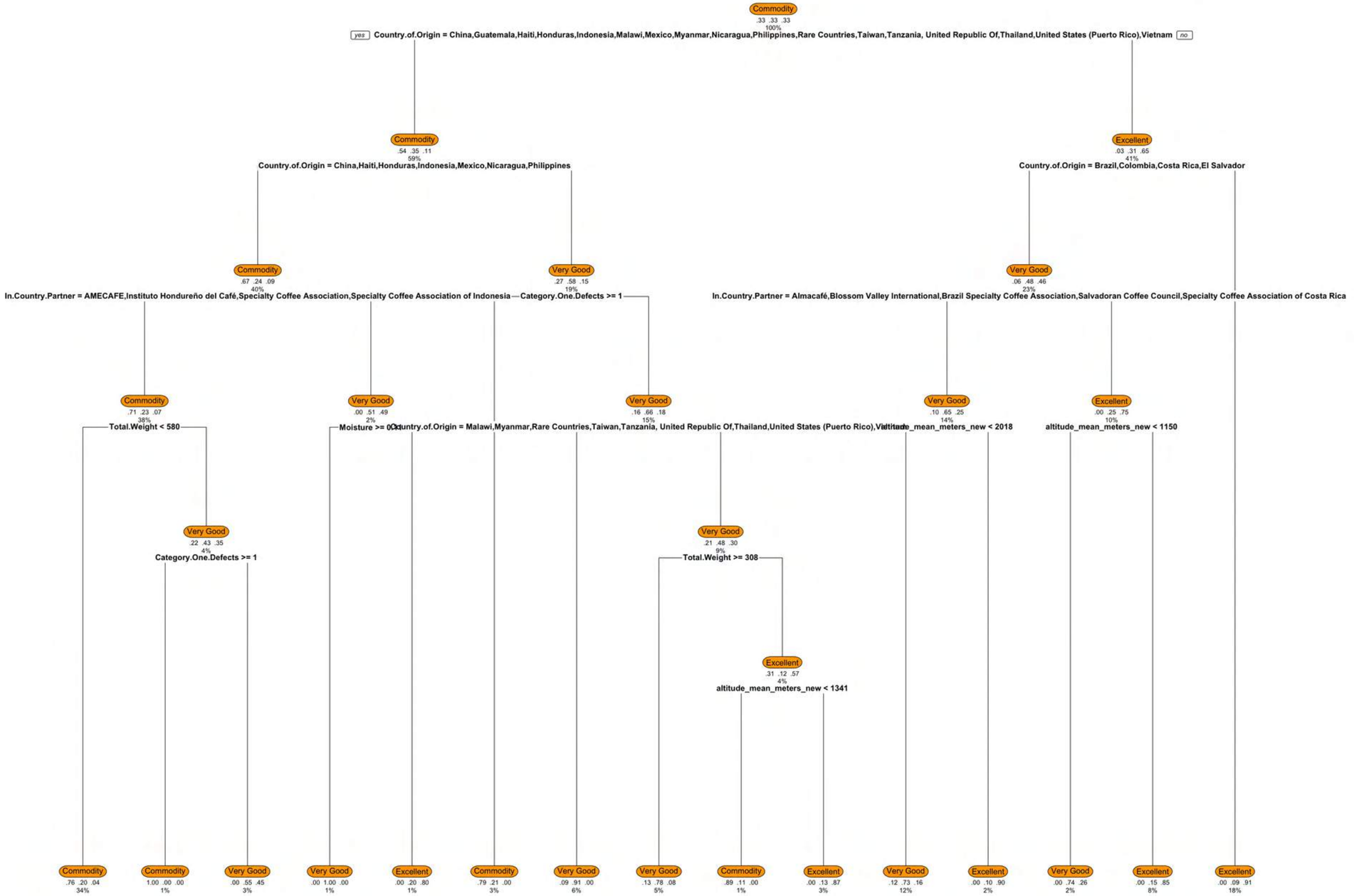

size of tree
Over
1
2
3
5
6
7
8
11
13
15
X-val Relative Error
0.2
0.4
0.6
0.8
1.0
1.2
Inf
0.18
0.059
0.034
0.024
0.018
0.015
0.013
0.011
0.011
cp


size of tree
ROSE
1
3
4
6
7
8
9
10
X-val Relative Error
0.2
0.4
0.6
0.8
1.0
Inf
0.2
0.089
0.041
0.035
0.026
0.016
0.01
cp

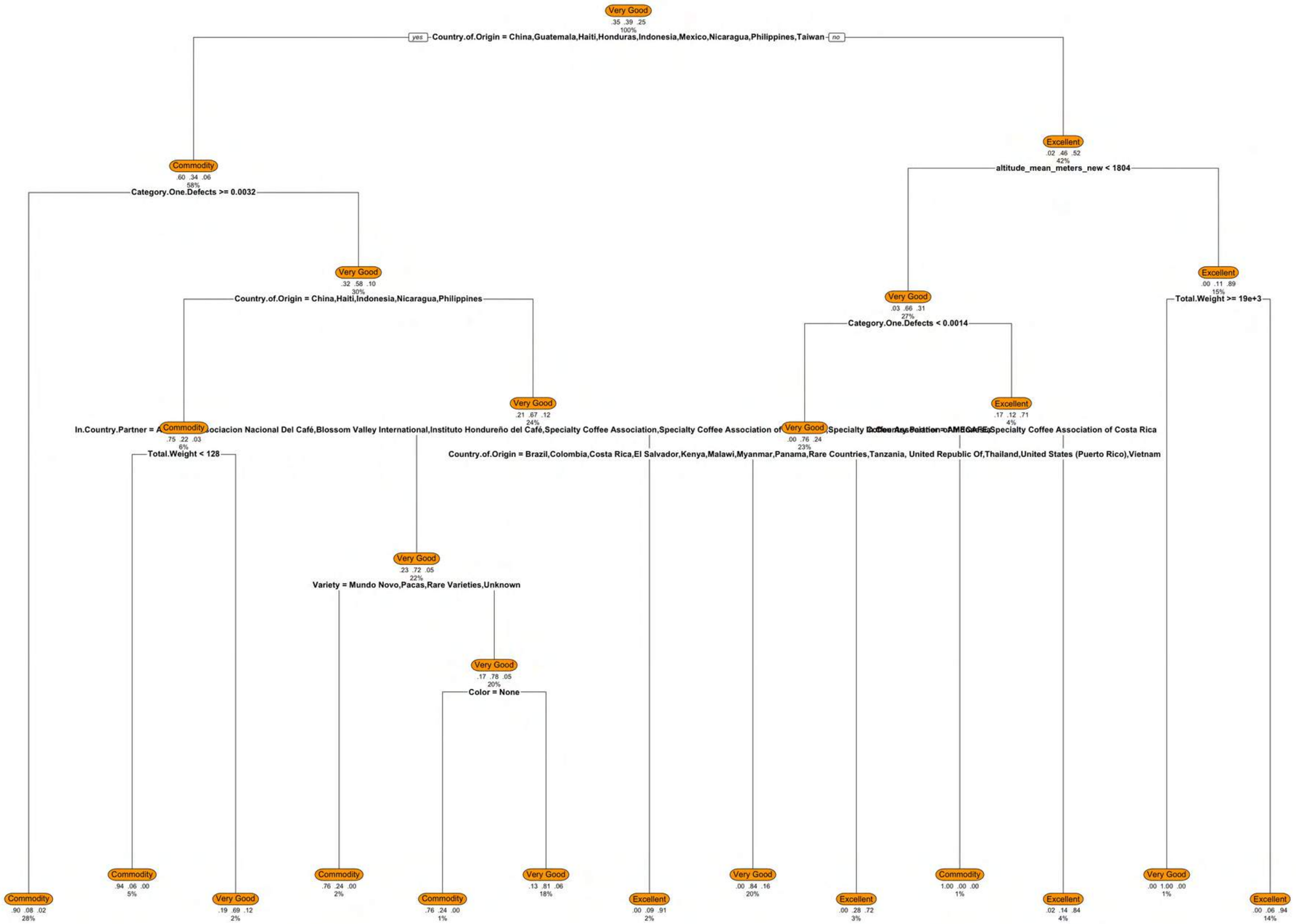
Initial DT (Pruned)



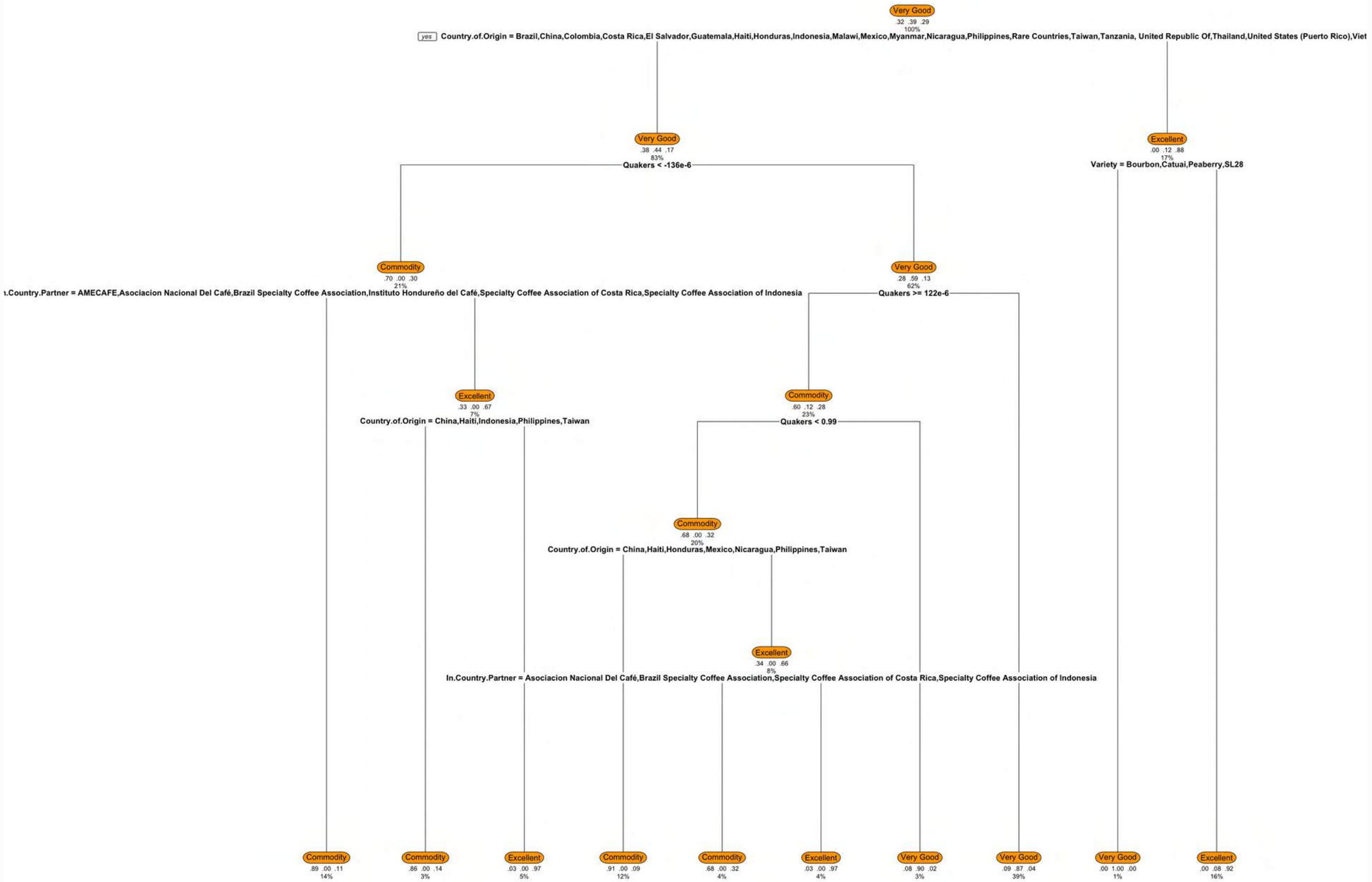
Over DT (Pruned)

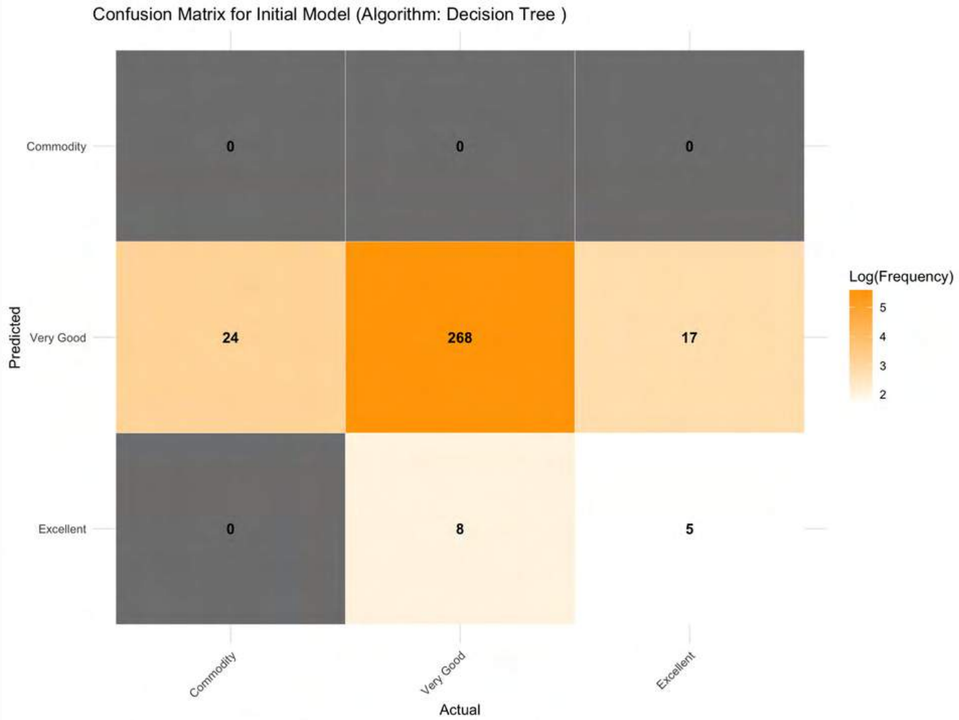


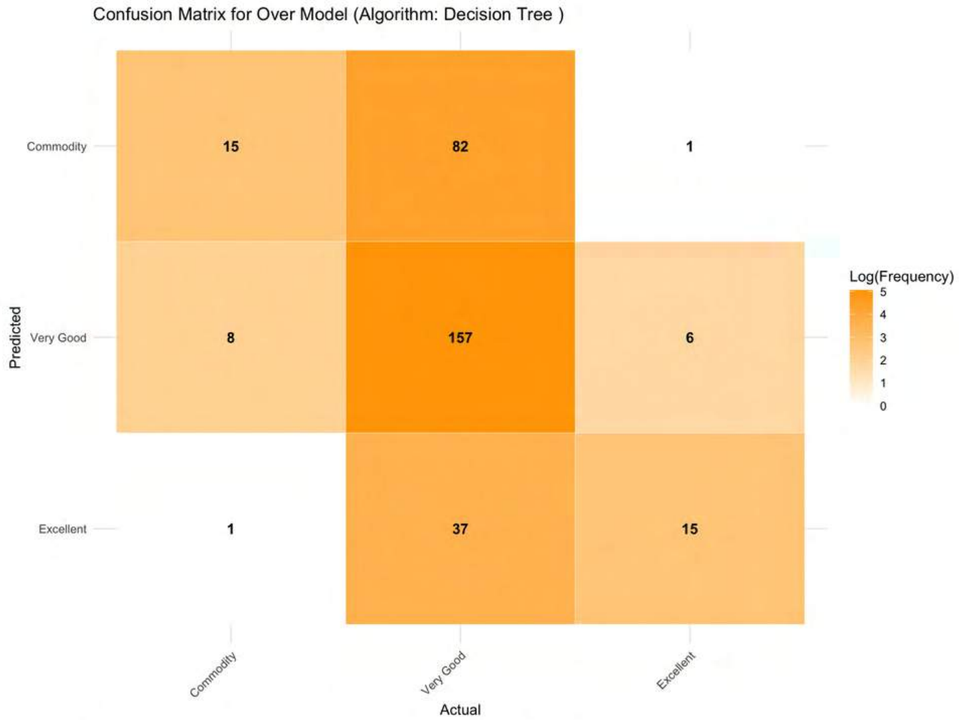
SMOTE DT (Pruned)

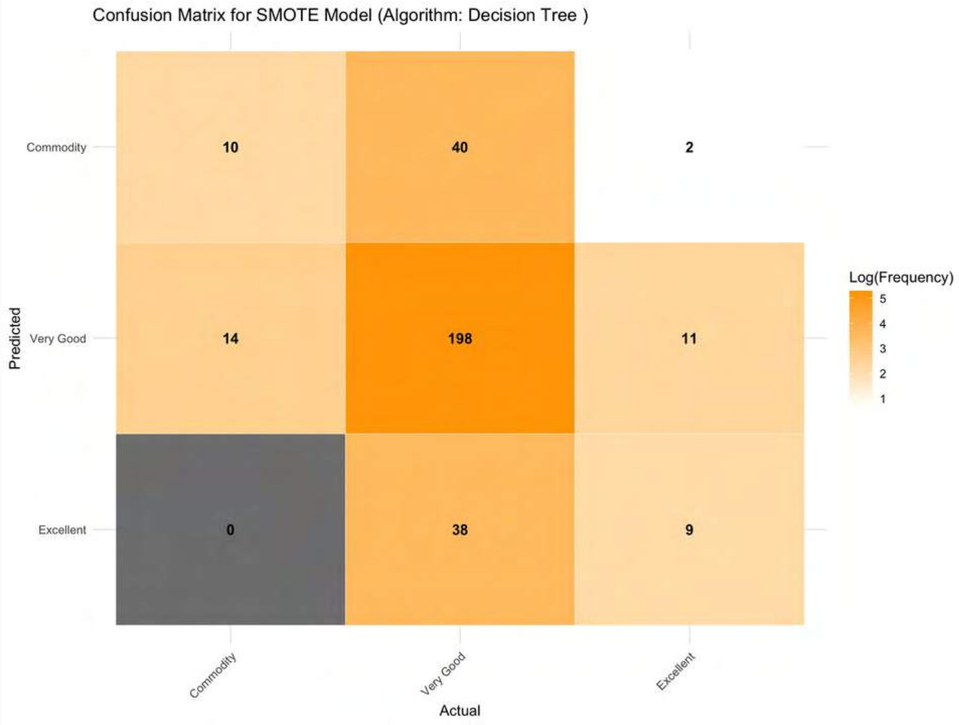


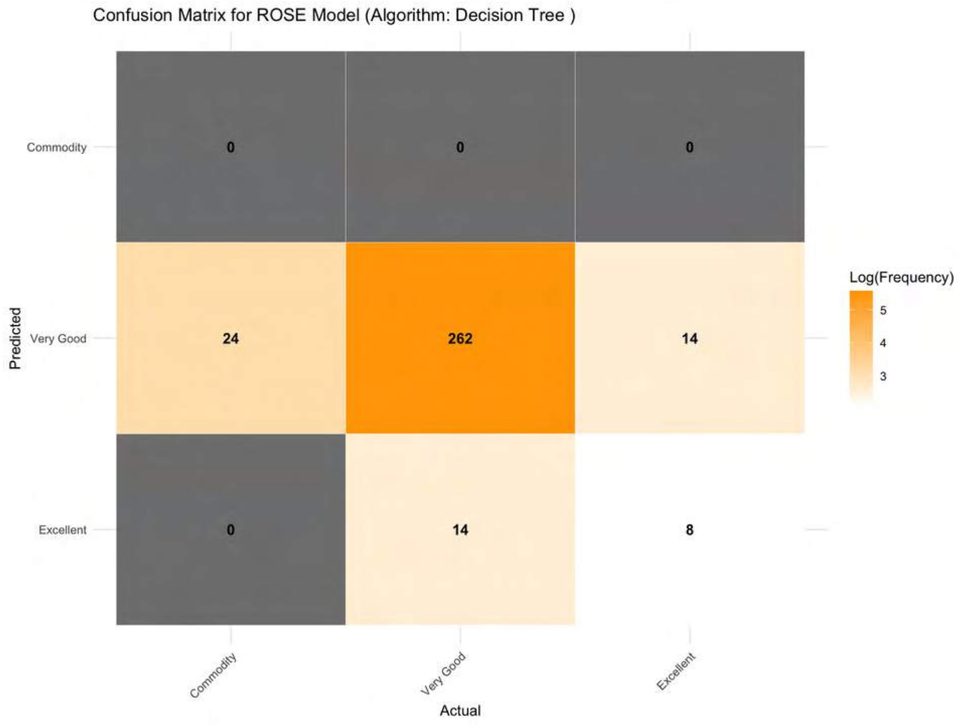
ROSE DT (Pruned)




Confusion Matrix for Initial Model (Algorithm: Decision Tree)
Commodity
Very Good
Excellent
Predicted
0
0
0
24
268
17
0
8
5
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2

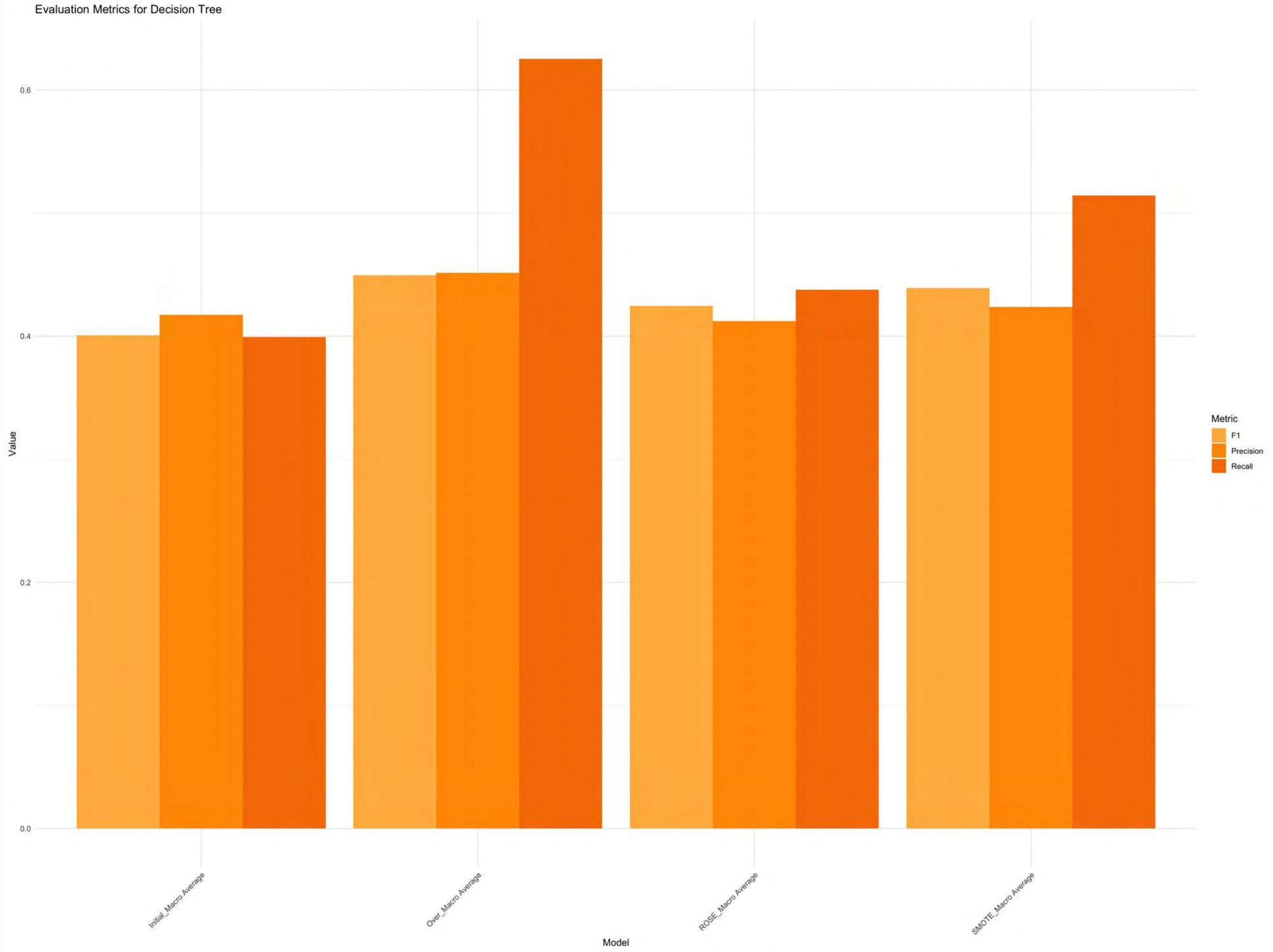

Confusion Matrix for Over Model (Algorithm: Decision Tree)
Commodity
Very Good
Excellent
Predicted
15
82
1
8
157
6
1
37
15
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2
1
0

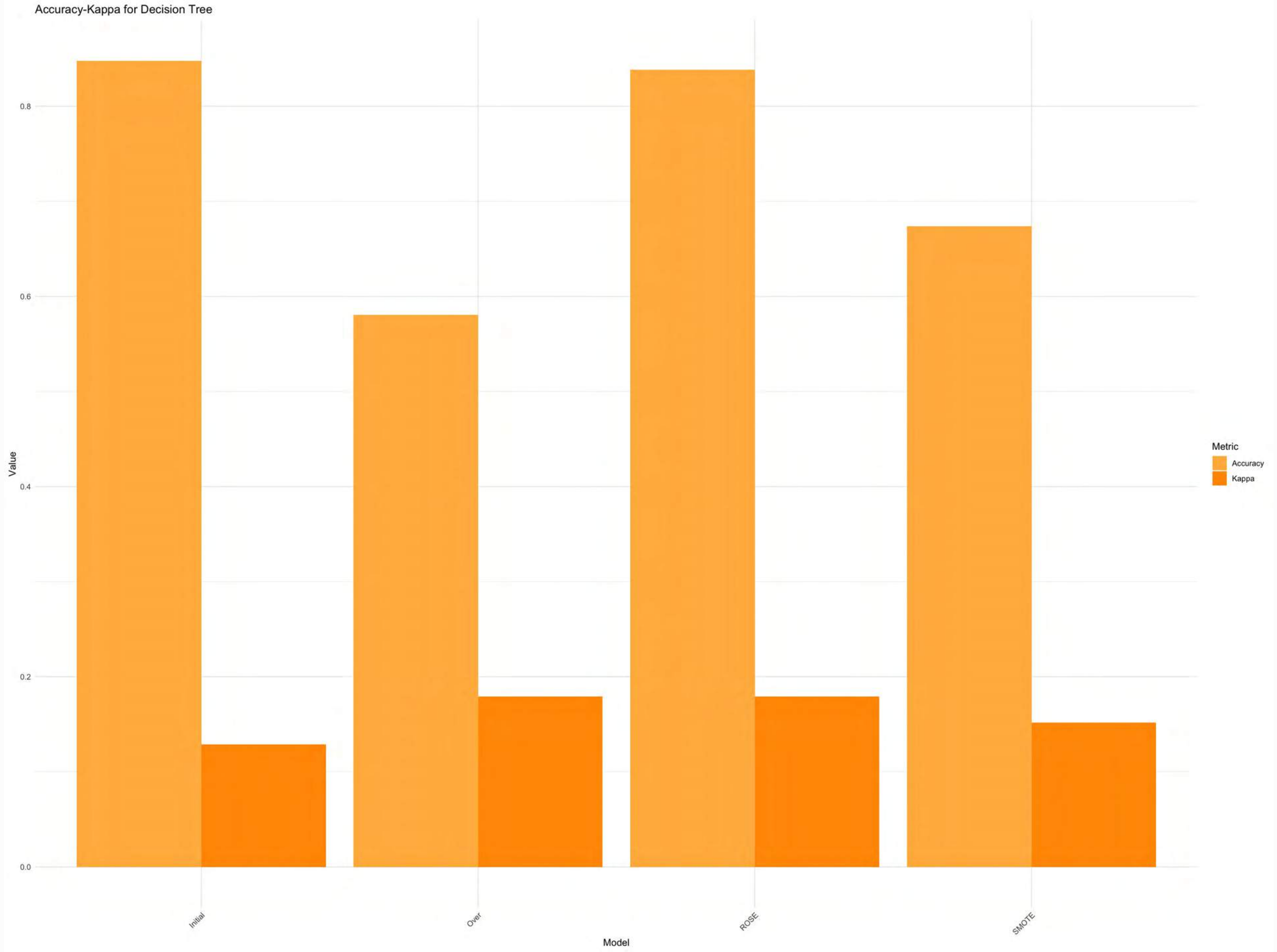

Confusion Matrix for SMOTE Model (Algorithm: Decision Tree)
Commodity
Very Good
Excellent
Predicted
10
40
2
14
198
11
0
38
9
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2
1

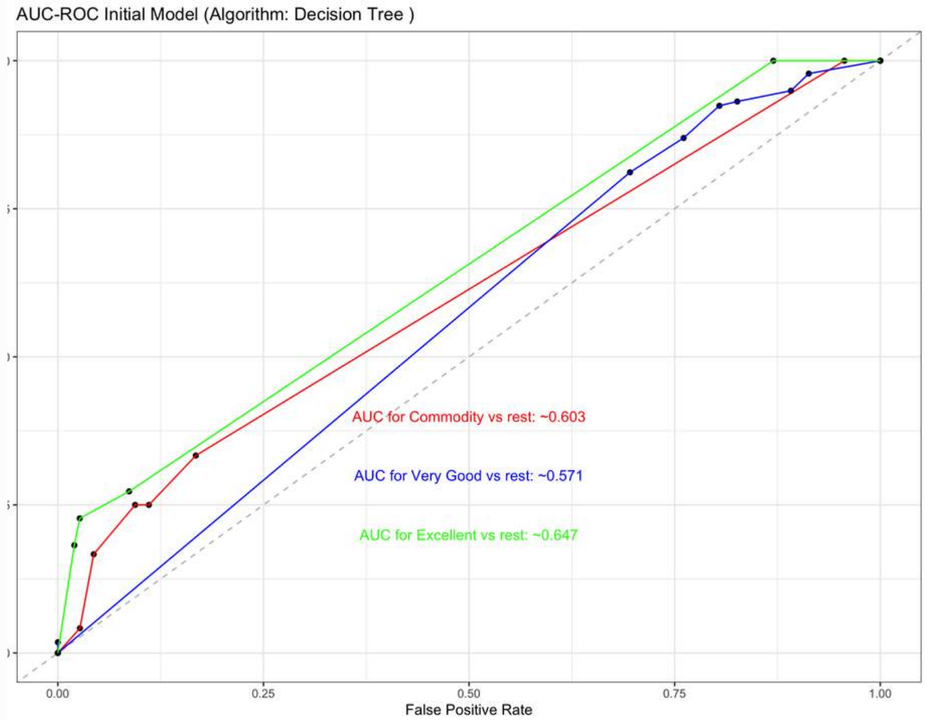

Confusion Matrix for ROSE Model (Algorithm: Decision Tree)
Commodity
Very Good
Excellent
Predicted
0
0
0
24
262
14
0
14
8
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3

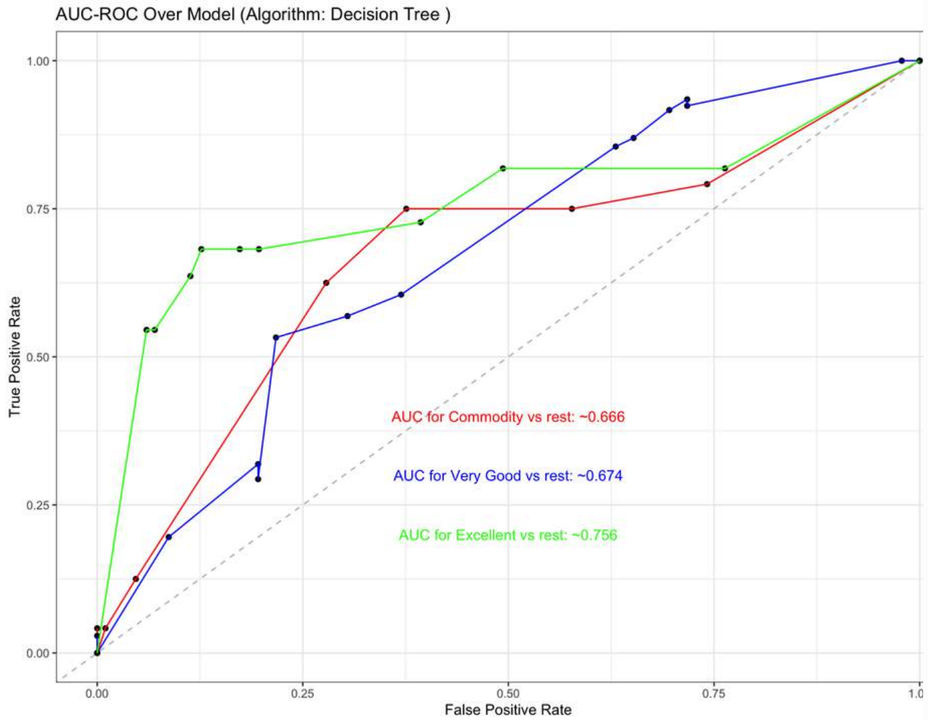
DECISION TREE EVALUATION

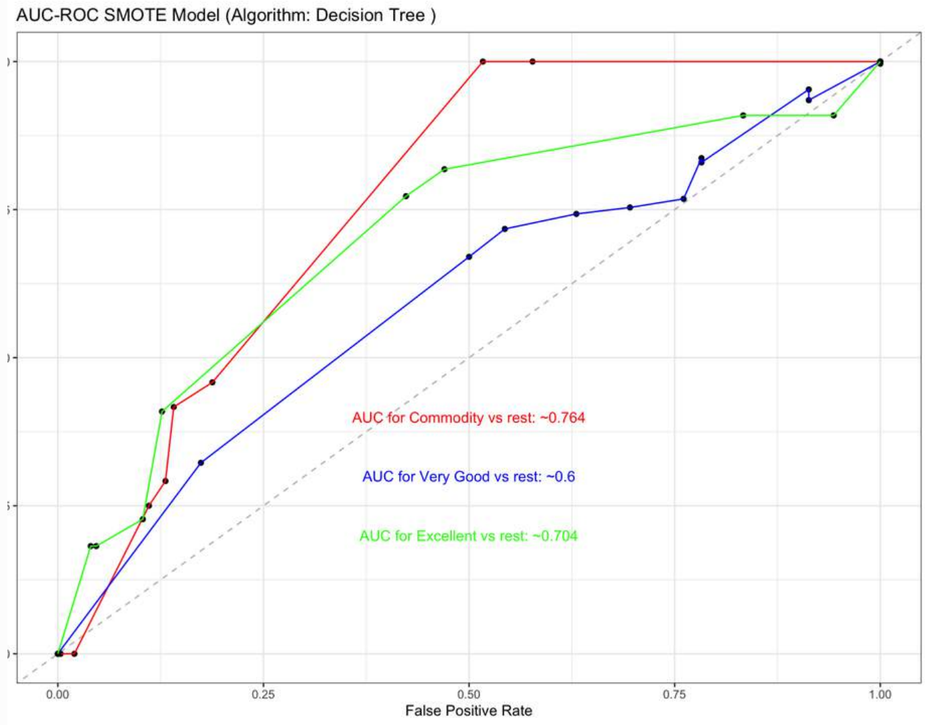
Model/Metric	Macro Precision	Macro Recall	Macro F1	Acuracy	Kappa
Initial	0.4173098	0.3994291	0.4006512	0.8478261	0.1289610
Over	0.4514029	0.6252196	0.4494542	0.5807453	0.1793468
SMOTE	0.4238965	0.5143830	0.4392049	0.6739130	0.1516535
ROSE	0.4123232	0.4376372	0.4244529	0.8385093	0.1792157

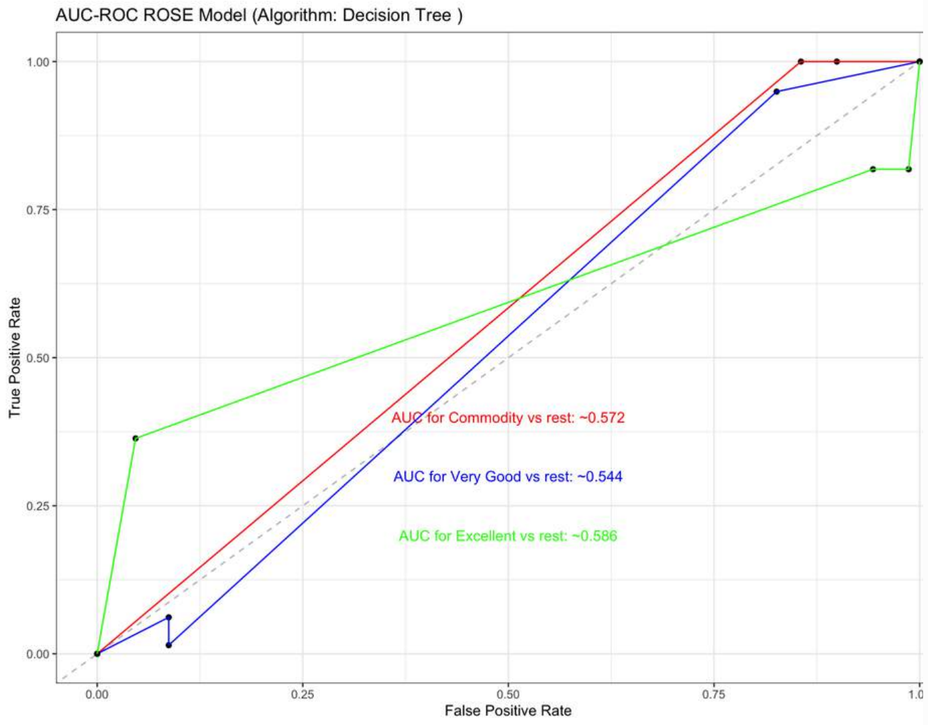

Evaluation Metrics for Decision Tree
0.6
0.4
0.2
0.0
Value
Initial_Macro Average
Over_Macro Average
ROSE_Macro Average
SMOTE_Macro Average
Model
Metric
F1
Precision
Recall


Accuracy-Kappa for Decision Tree
0.8
0.6
0.4
0.2
0.0
Value
Initial
Over
ROSE
SMOTE
Model
Metric
Accuracy
Kappa









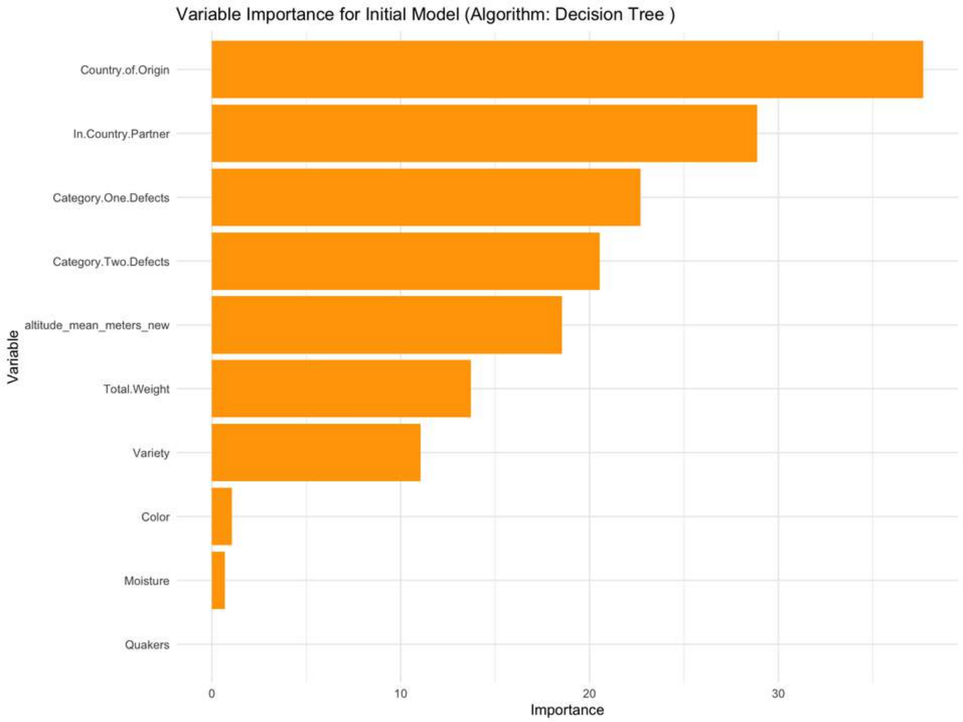
MEAN AUC-ROC

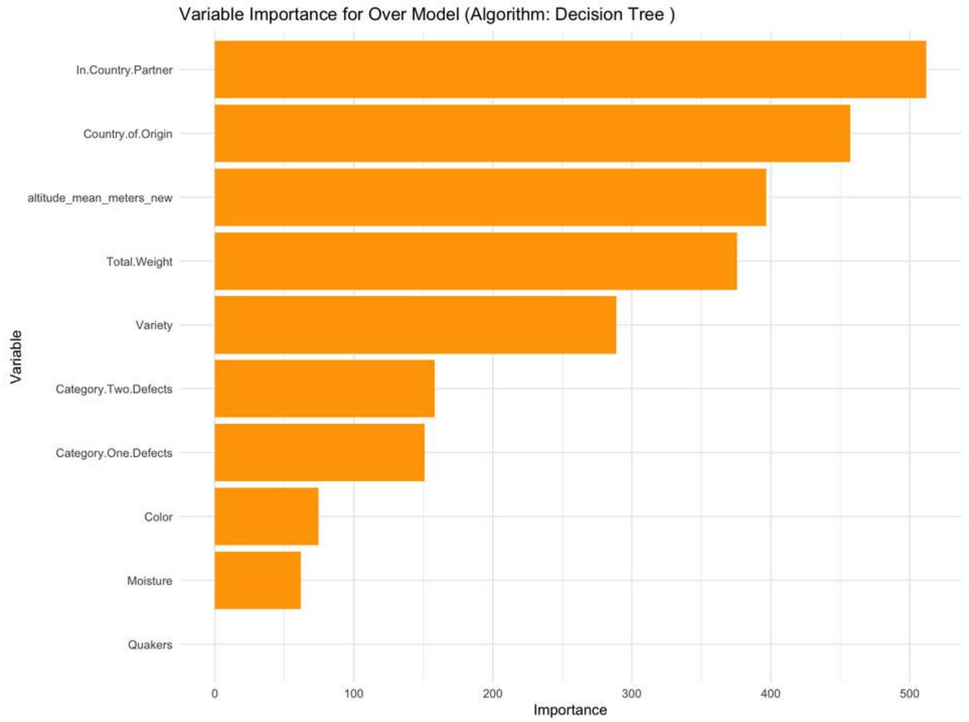
INITIAL: 0.61

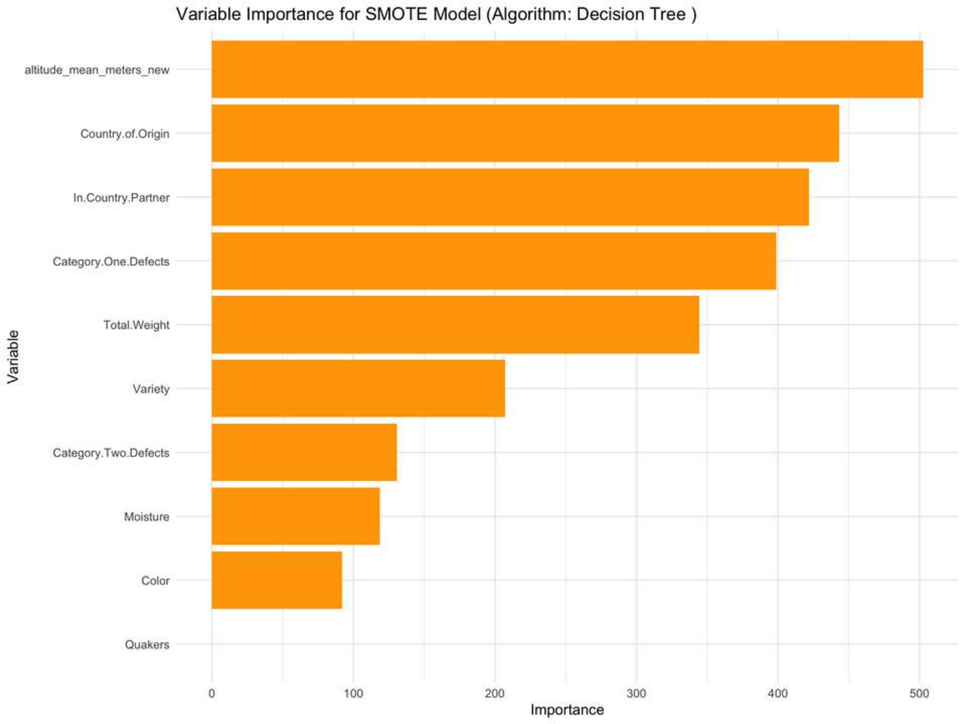
OVER: 0.69

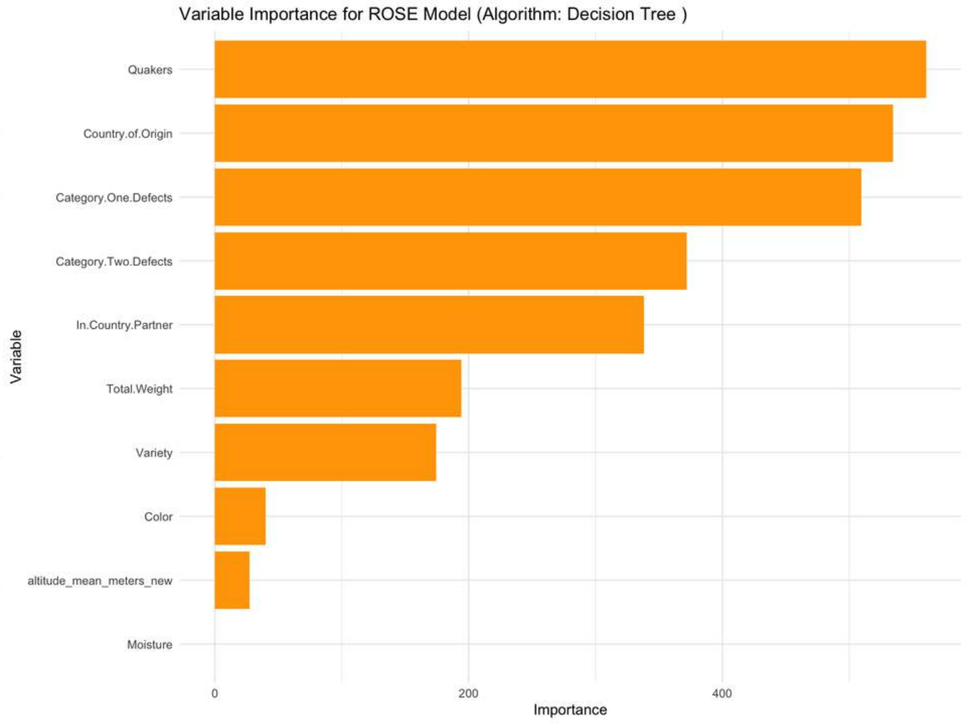
SMOTE: 0.69

ROSE: 0.57


Variable Importance for Initial Model (Algorithm: Decision Tree)
Country.of.Origin
In.Country.Partner
Category.One.Defects
Category.Two.Defects
altitude_mean_meters_new
Total.Weight
Variety
Color
Moisture
Quakers
Variable
0
10
20
30
Importance


Variable Importance for Over Model (Algorithm: Decision Tree)
In.Country.Partner
Country.of.Origin
altitude_mean_meters_new
Total.Weight
Variety
Category.Two.Defects
Category.One.Defects
Color
Moisture
Quakers
Variable
0
100
200
300
400
500
Importance

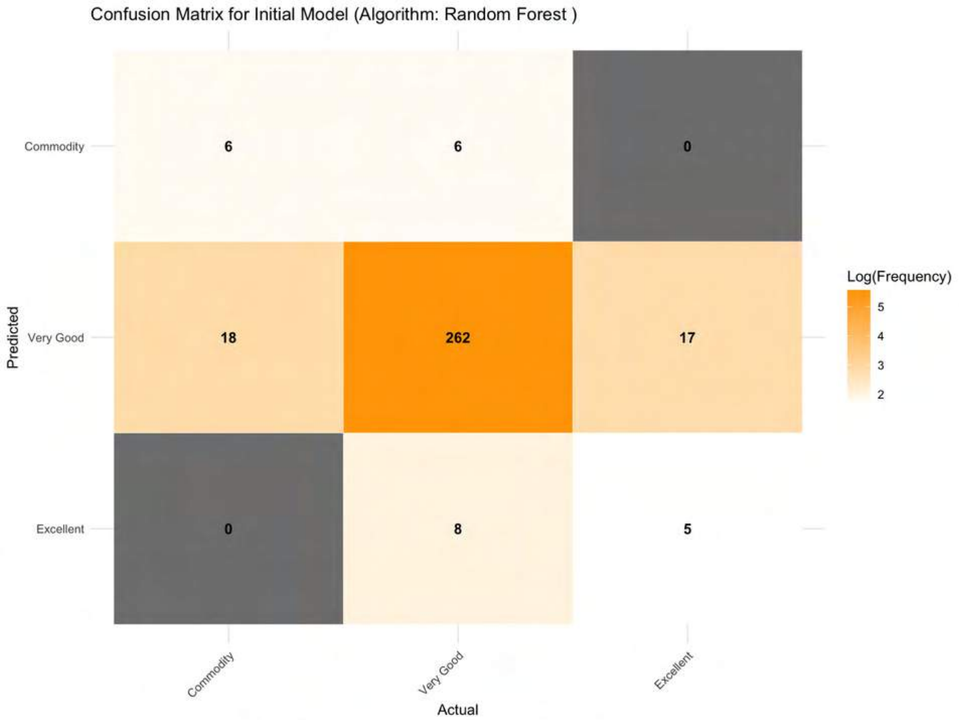

Variable Importance for SMOTE Model (Algorithm: Decision Tree)
altitude_mean_meters_new
Country.of.Origin
In.Country.Partner
Category.One.Defects
Total.Weight
Variety
Category.Two.Defects
Moisture
Color
Quakers
Variable
0
100
200
300
400
500
Importance

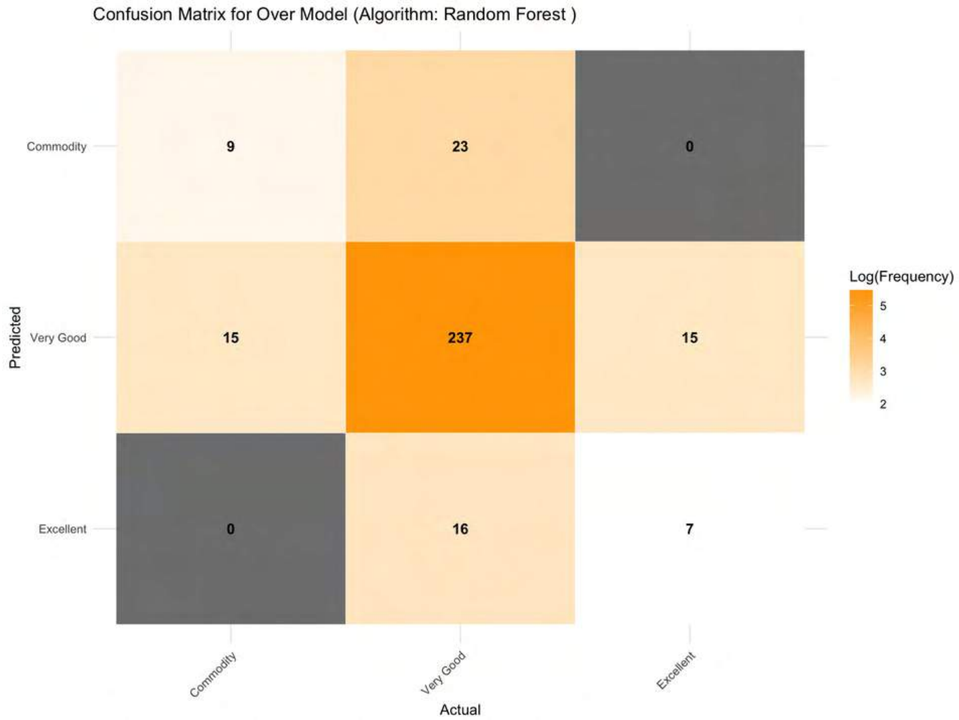

Variable Importance for ROSE Model (Algorithm: Decision Tree)
Quakers
Country.of.Origin
Category.One.Defects
Category.Two.Defects
In.Country.Partner
Total.Weight
Variety
Color
altitude_mean_meters_new
Moisture
Variable
0
200
400
Importance

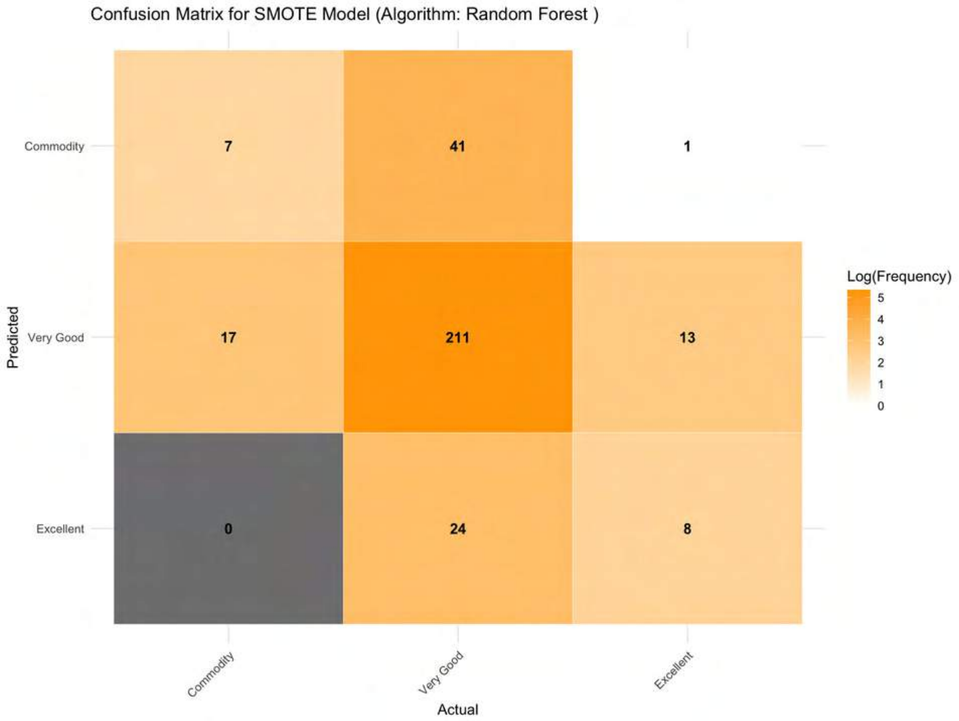
RANDOM FOREST

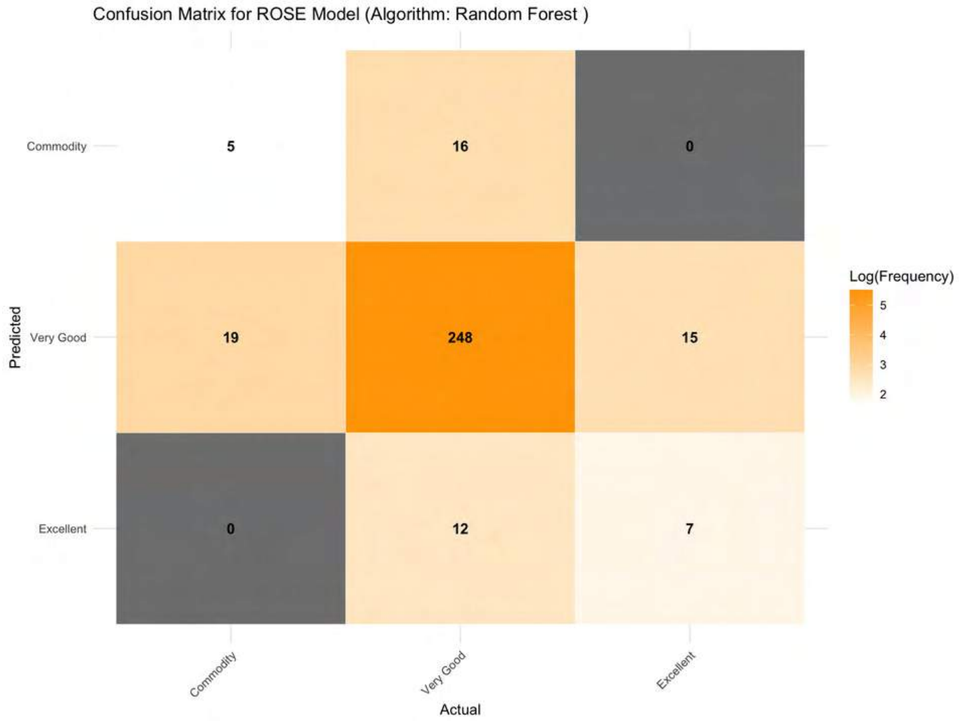
Our last algorithm is the Random Forest which builds on the concept of Decision Tree.

We have tuned for the number of features to consider at each split. We relied on a 10 fold cross validation for estimating the number of features to consider at each split.


Confusion Matrix for Initial Model (Algorithm: Random Forest)
Commodity
Very Good
Excellent
Predicted
6
6
0
18
262
17
0
8
5
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2

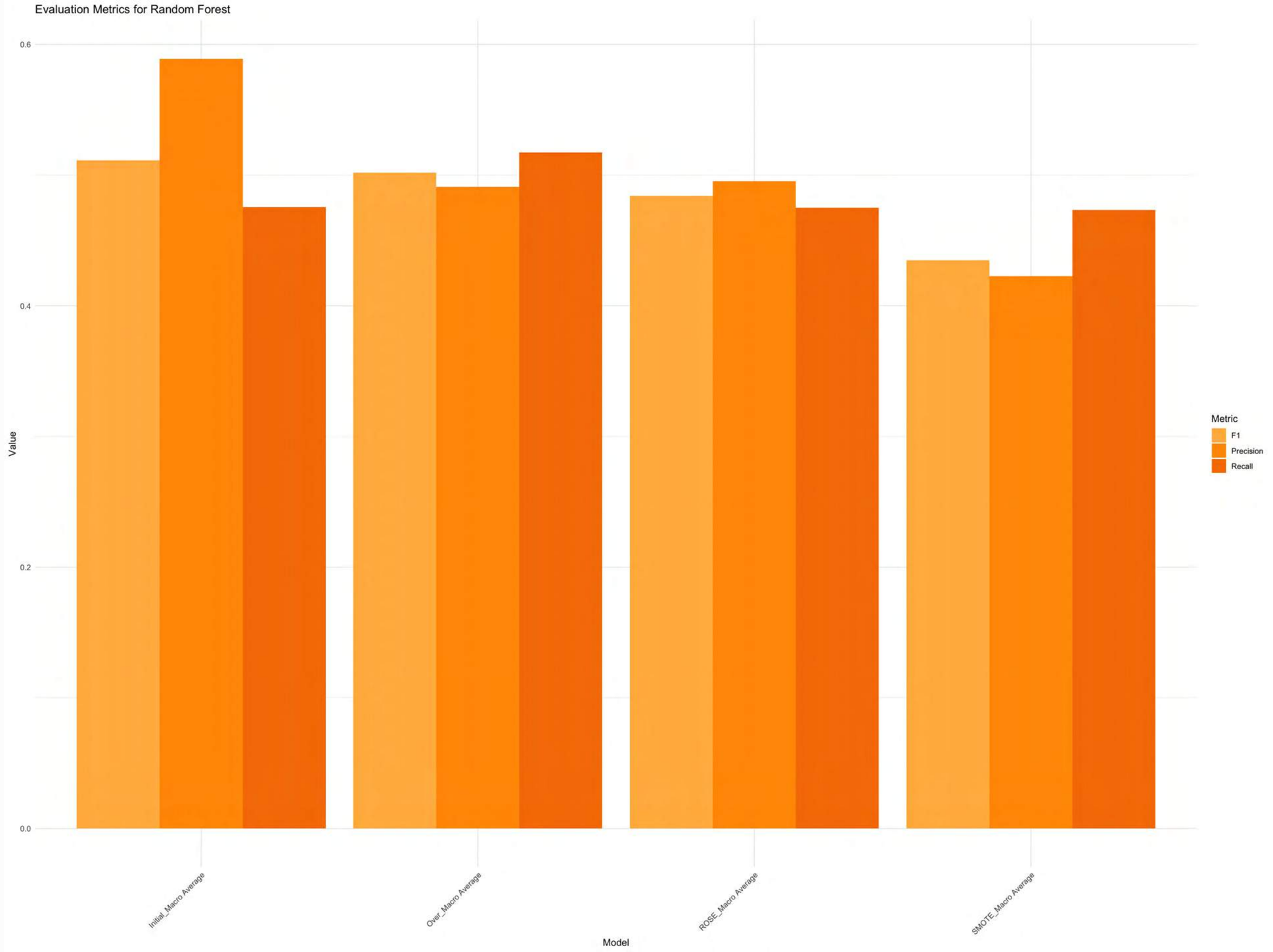

Confusion Matrix for Over Model (Algorithm: Random Forest)
Commodity
Very Good
Excellent
Predicted
9
23
0
15
237
15
0
16
7
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2

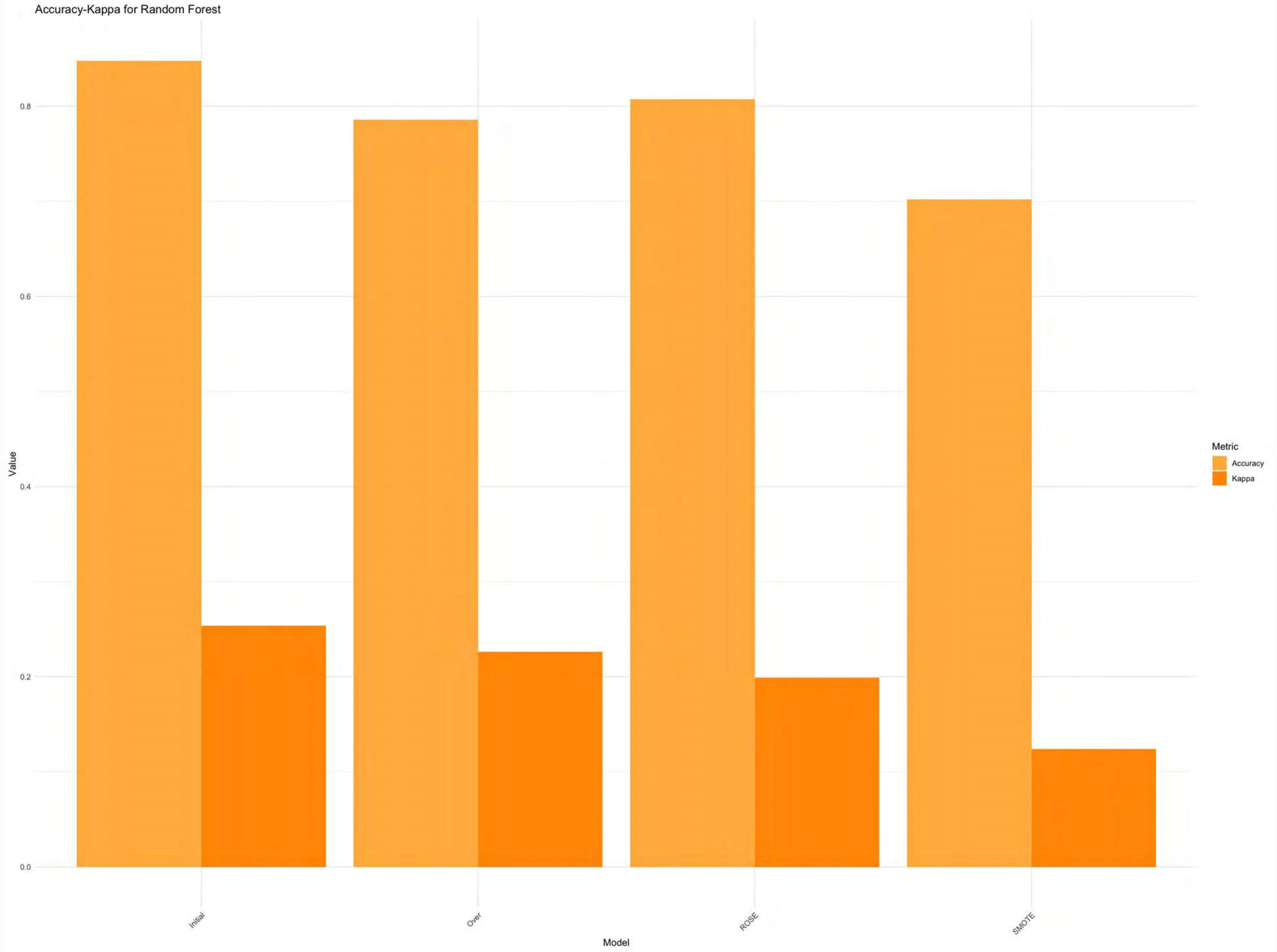

Confusion Matrix for SMOTE Model (Algorithm: Random Forest)
Commodity
Very Good
Excellent
Predicted
7
41
1
17
211
13
0
24
8
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2
1
0

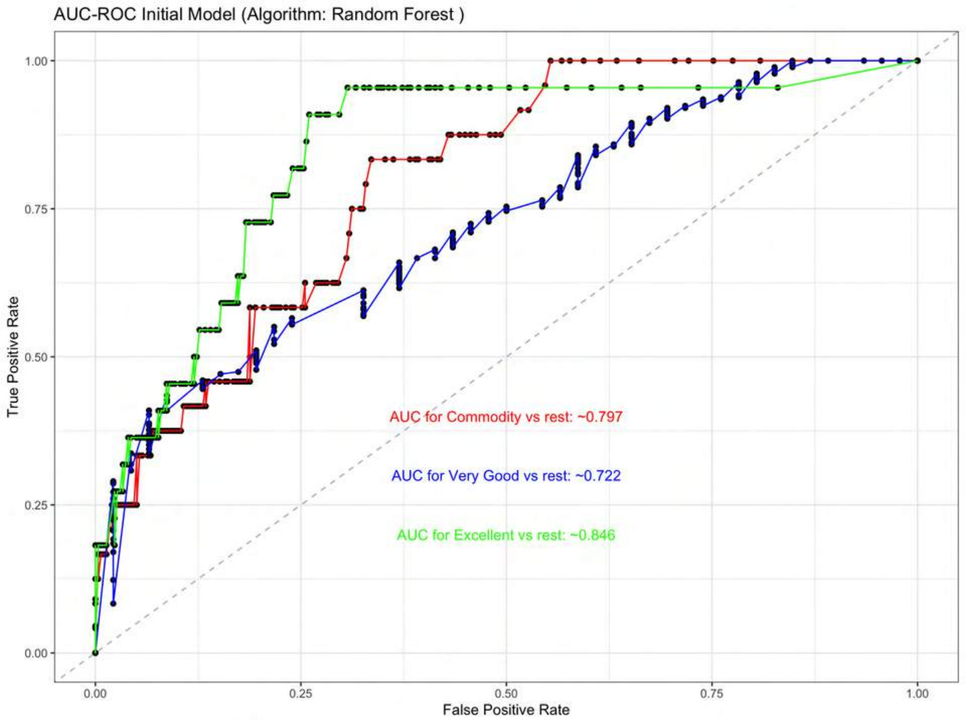

Confusion Matrix for ROSE Model (Algorithm: Random Forest)
Commodity
Very Good
Excellent
Predicted
5
16
0
19
248
15
0
12
7
Commodity
Very Good
Excellent
Actual
Log(Frequency)
5
4
3
2

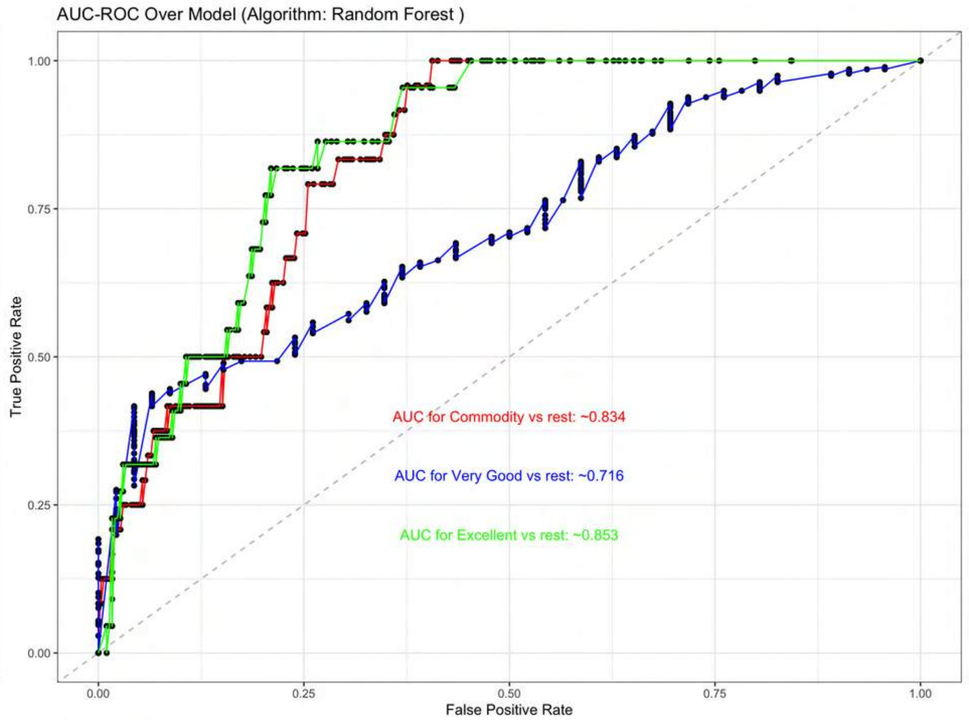
RANDOM FOREST EVALUATION

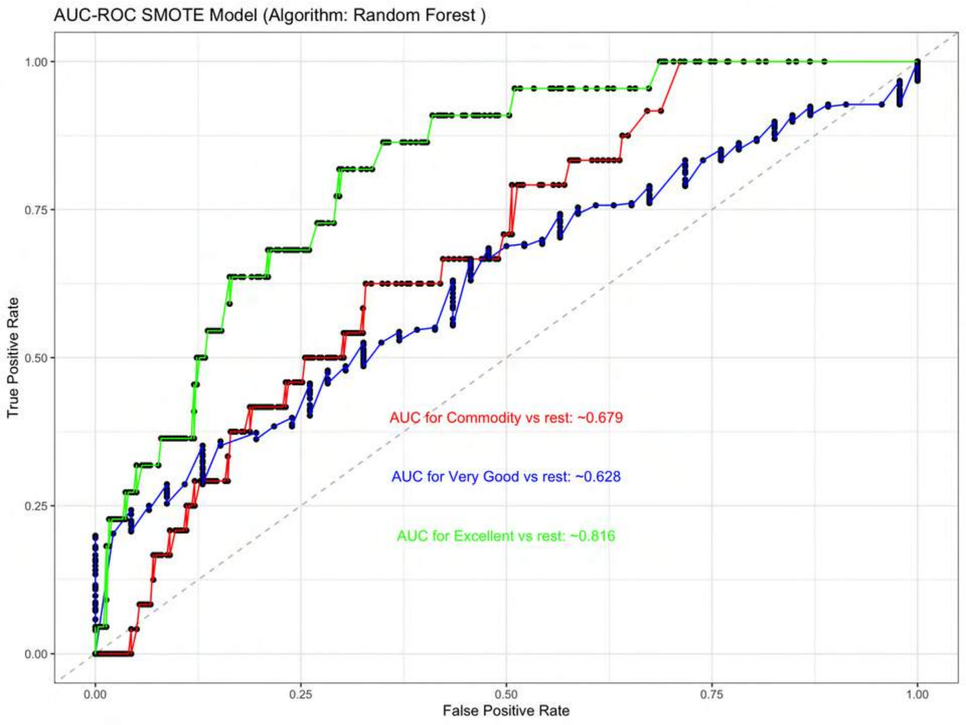
Model/Metric	Macro Precision	Macro Recall	Macro F1	Acuracy	Kappa
Initial	0.5889234	0.4755160	0.511177	0.8478261	0.2535718
Over	0.4910794	0.5172925	0.5018226	0.7857143	0.2263389
SMOTE	0.4227919	0.4732653	0.4347749	0.7018634	0.1240082
ROSE	0.4953163	0.4750220	0.4841915	0.8074534	0.1991978

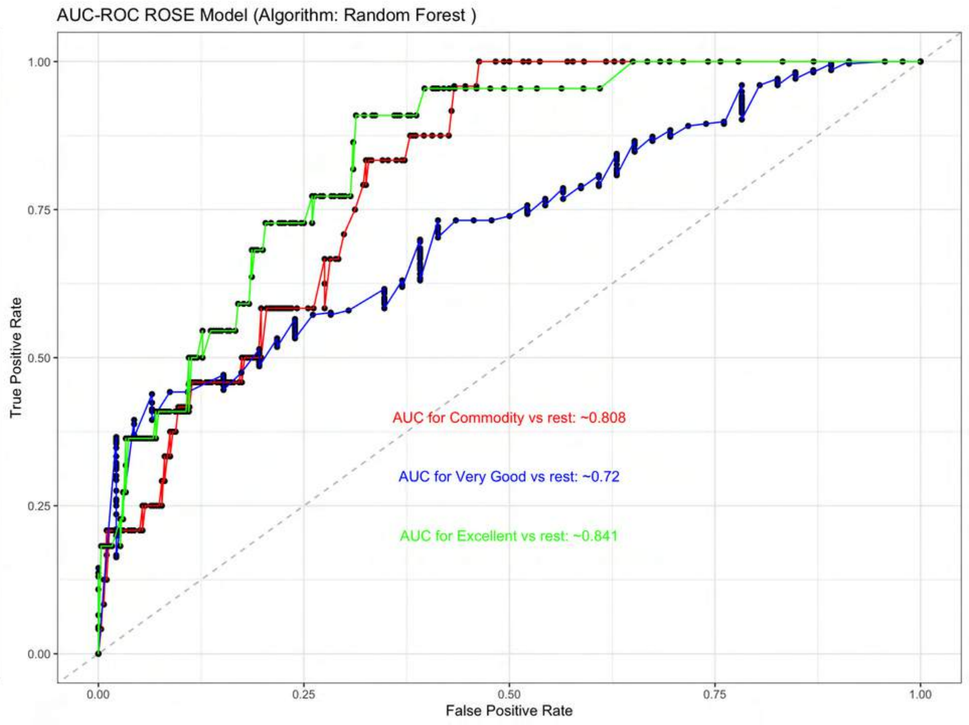

Evaluation Metrics for Random Forest
0.6
0.4
0.2
0.0
Value
Initial_Macro Average
Over_Macro Average
ROSE_Macro Average
SMOTE_Macro Average
Model
Metric
F1
Precision
Recall


Accuracy-Kappa for Random Forest
Value
0.0
0.2
0.4
0.6
0.8
Initial
Over
ROSE
SMOTE
Model
Metric
Accuracy
Kappa









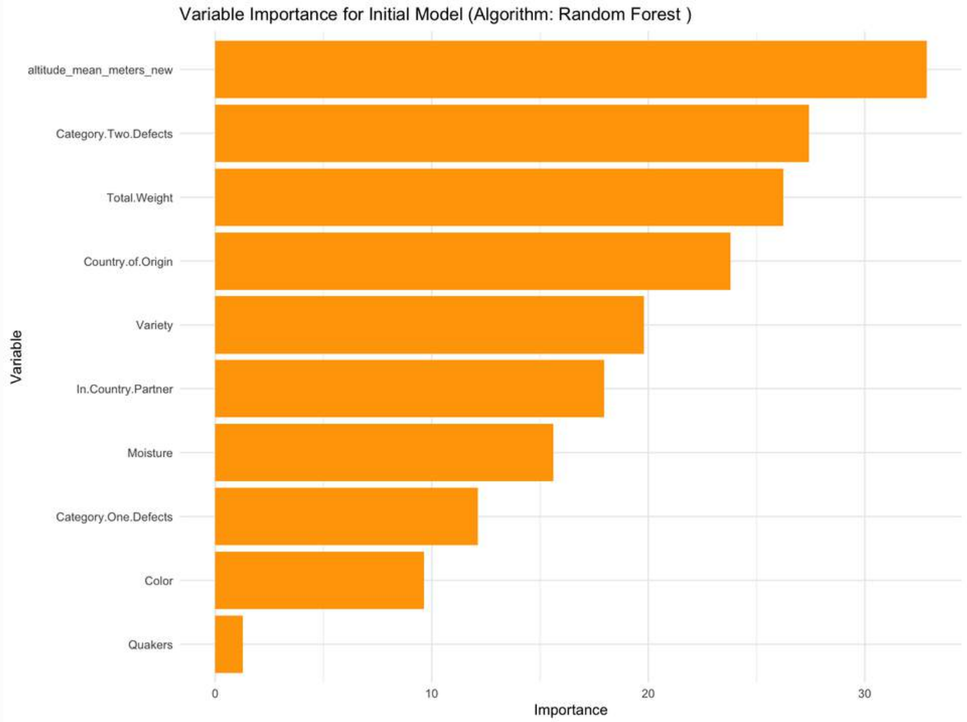
MEAN AUC-ROC

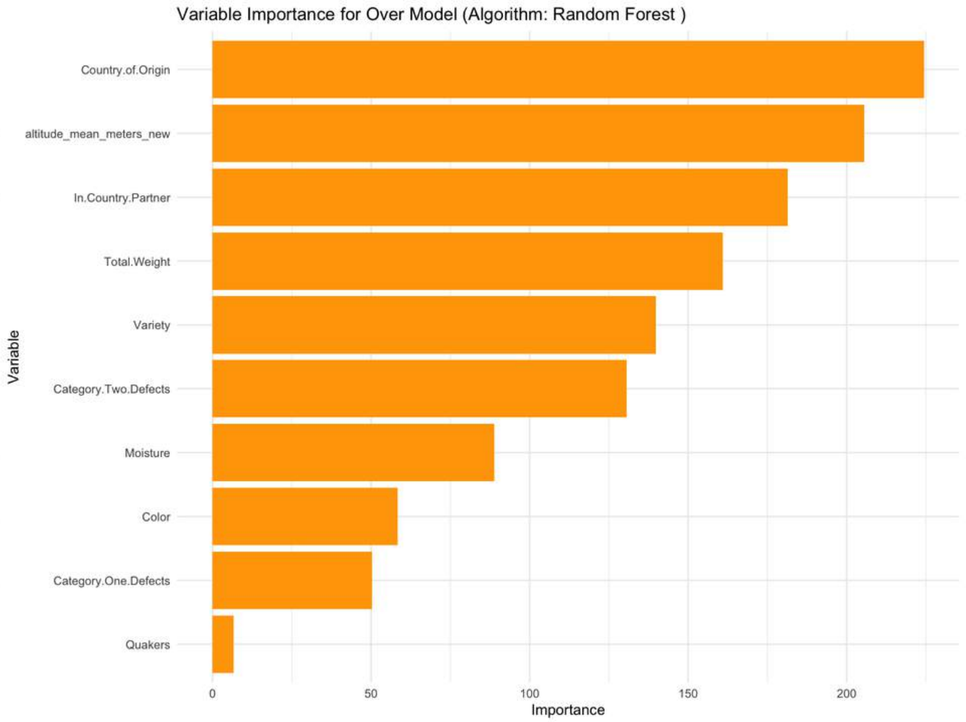
INITIAL: 0.79

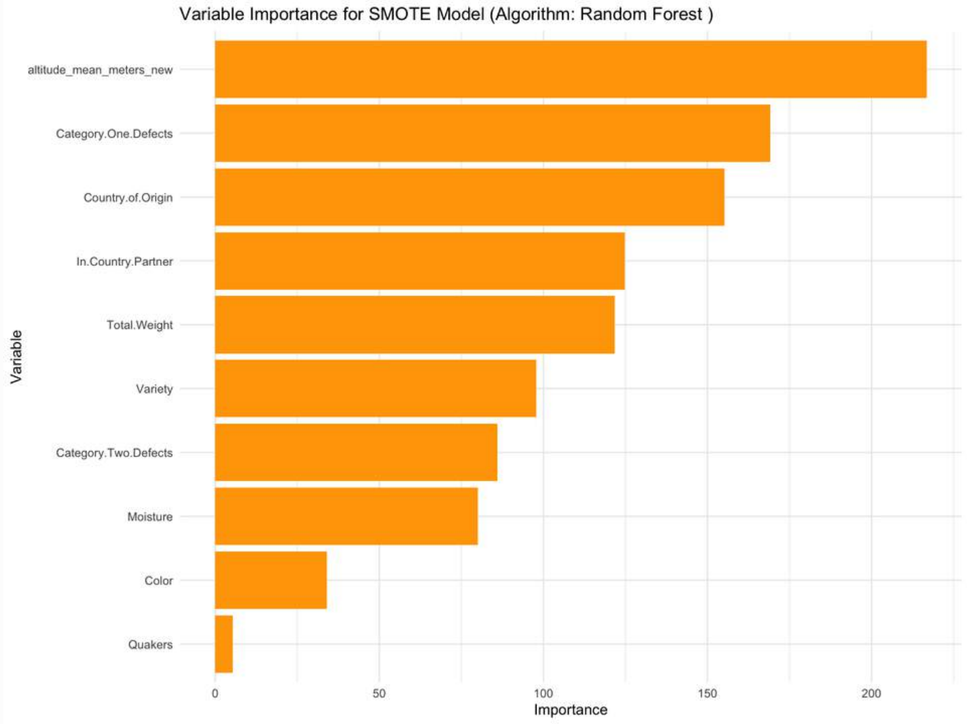
OVER: 0.80

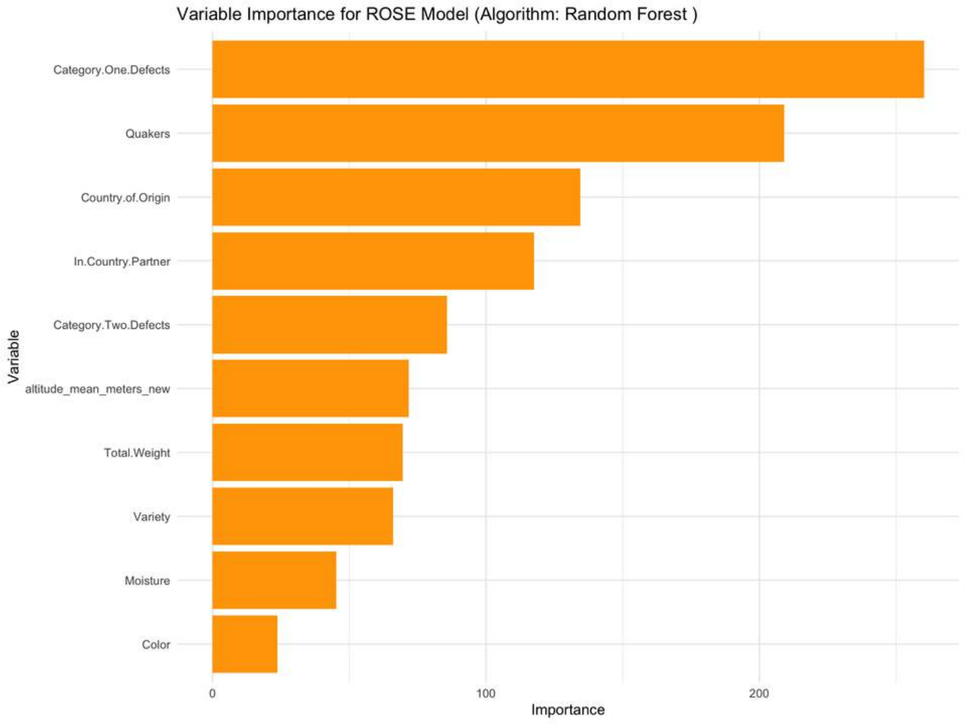
SMOTE: 0.75

ROSE: 0.79


Variable Importance for Initial Model (Algorithm: Random Forest)
altitude_mean_meters_new
Category.Two.Defects
Total.Weight
Country.of.Origin
Variety
In.Country.Partner
Moisture
Category.One.Defects
Color
Quakers
Variable
0
10
20
30
Importance


Variable Importance for Over Model (Algorithm: Random Forest)
Country.of.Origin
altitude_mean_meters_new
In.Country.Partner
Total.Weight
Variety
Category.Two.Defects
Moisture
Color
Category.One.Defects
Quakers
Variable
0
50
100
150
200
Importance


Variable Importance for SMOTE Model (Algorithm: Random Forest)
altitude_mean_meters_new
Category.One.Defects
Country.of.Origin
In.Country.Partner
Total.Weight
Variety
Category.Two.Defects
Moisture
Color
Quakers
Variable
0
50
100
150
200
Importance


Variable Importance for ROSE Model (Algorithm: Random Forest)
Category.One.Defects
Quakers
Country.of.Origin
In.Country.Partner
Category.Two.Defects
altitude_mean_meters_new
Total.Weight
Variety
Moisture
Color
Variable
0
100
200
Importance

THANK YOU!