Exploratory Data Analysis for Coffee Data

1. Introduction

This document presents an exploratory data analysis (EDA) of Coffee data. The goal is to understand the factors that impacted the coffee beans ratings, and data structure. The data comprises 29 covariates and 945 records. The final grade has been recorded under Total Cup Point variable, and it is related to 13 other review covariates.

2. Analysis of missing data

The analysis of missing data has been carried out in order to find and identify the types of missingness present in coffee data. It has shown that overall, there are 157 observations with missing data. All of these observations lack records in the three altitude covariates- Altitude Low Meters, Altitude Mean Meters and Altitude High Meters. Most of the records which miss altitude data appear to follow MAR or MNAR. It is worth nothing that 90% and 75% of records from Hawaii and Peru respectively lack information about altitude (Figure 1). There is MCAR record of Quakers and a couple of records MAR both Altitude and Harvest Year.

Subsequently, the missing data were handled by several techniques, including Listwise deletion, Mean Imputation and Predictive Mean Matching. Missing records of Altitude and Harvest.Year. are from the same plantation in Brazil. Given the insignificant number of missing values and untidiness of both categorical variables (nonspecific ranges of values and various units), listwise deletion will not have much impact on the overall statistics. Moreover, the single missing value of Quakres has been assigned the mean of the column. In order to conduct Predictive Mean Matching imputation (PMM) on the three numerical altitude variables, multicollinearity needed to be removed, leaving only one altitude covariate - Altitude Mean Meters. This stochastic method allowed to better reflect the variability in the data and led to more realistic imputations compared to deterministic methods.

3. Outlier Detection

In order to make the right modelling assumptions and assure data quality, the analysis of outliers has been conducted. Although, the obvious single error has been deleted from the data, further sensitivity analysis was undertaken w.r.t. correlation, in the next section. Modified Z-score has been computed for the numerical data, indicating numerous outliers in most of the numerical covariates. This suggests that there might be a connection between the large number of outliers and presence of clusters in the data, as the outliers may represent distinct subgroups or clusters within the data. The presence of clusters has been analyzed in the 5th and 6th paragraph.

4. Multivariate analysis including spatial data

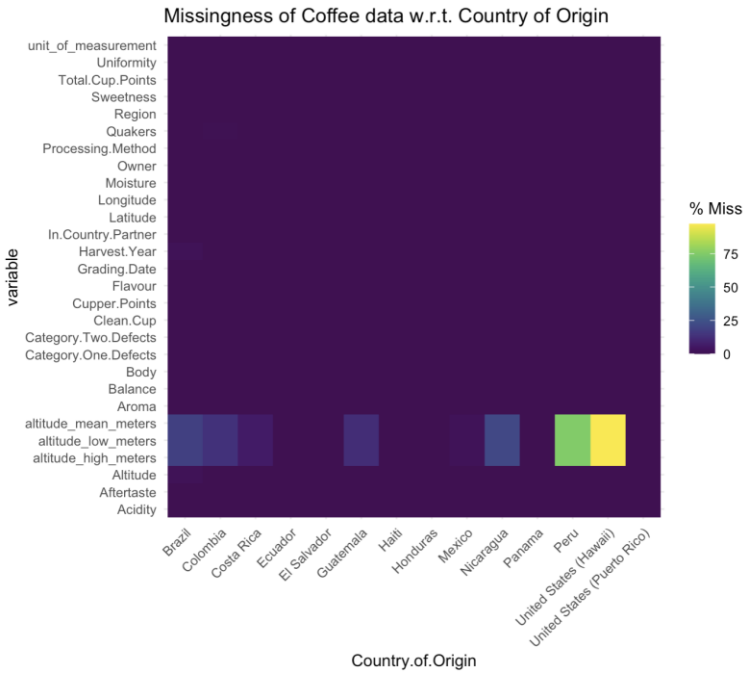

Figure 1: The plot of missing data across the Countries of Origin within the numeric altitude covariates

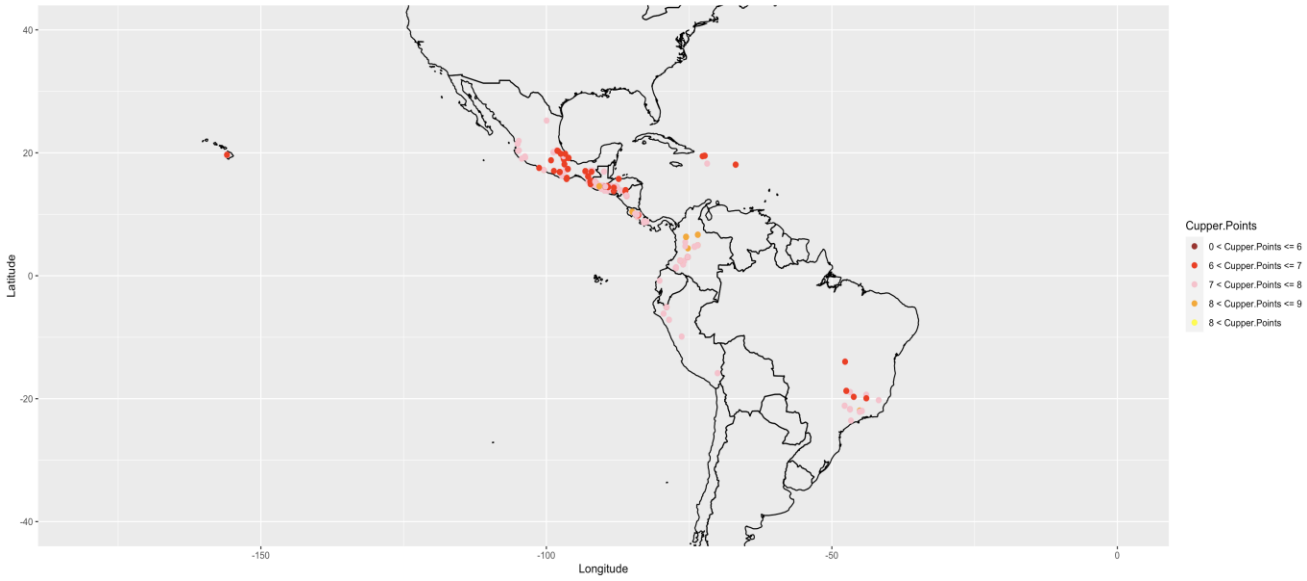

Figure 2: Cupper points distributed in space

As shown in Figure 2, the spatial analysis of Cupper Points covariate indicated that the closer to the Equator, the higher the aftertaste was rated. Presenting the data in the map is also an effortless way

to immediately determine that all the data was gathered in North and South America. Correlation between the numeric covariates has also been examined. Figure 3 shows the correlation heatmap illustrating the correlation coefficients between numerical variables from Coffee dataset. There is a very strong correlation of: 0.85 between Flavor and Aftertaste; - 0.81 and 0.82 between Cupper Points and Flavor, Aftertaste respectively; - and 0.82, 0.80, 0.80 between Total Cup Points and Flavor, Aftertaste, Cupper Points respectively and strong negative correlation of -0.79 between latitude and longitude.

This required further investigation in respect to the possible multicollinearity. Before having removed the altitude covariates and the error, correlation between them was approx. 0.99 indicating perfect multicollinearity. Also, Variance Inflation Factor indicated multicollinearity between Favour, Aftertaste and Cupper Points. However, after tidying the data, VIF analysis indicated no multicollinearity between the covariates.

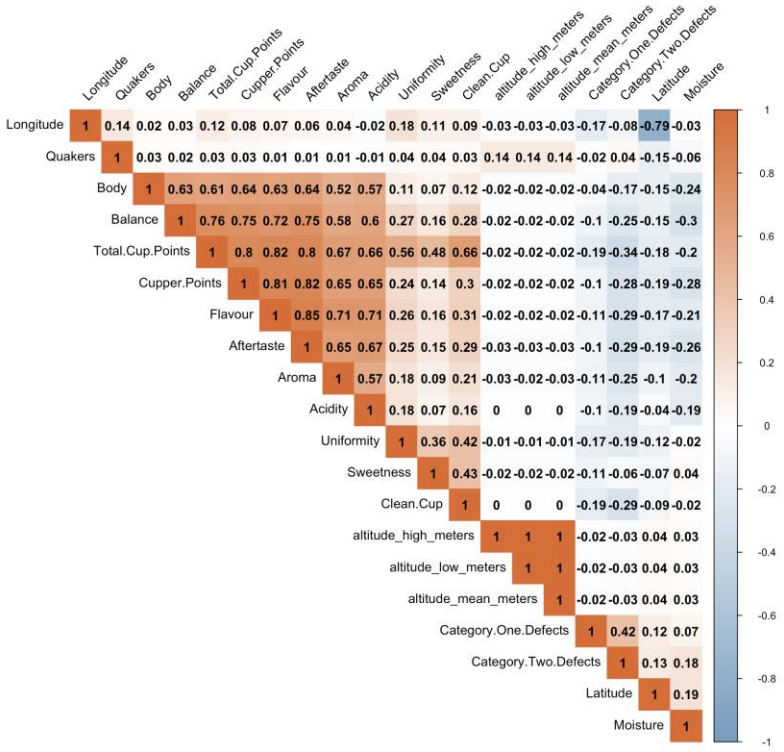

Figure 3: Correlation heatmap between the numeric coffee data

5. Clustering behaviour present in the numeric review data

At first clustering was analysed using simple tools as *ggpairs* and *contour plots*. Two clusters were recognized in all scatterplots congaing Moisture covariate, but an assumption was made that there still be some clusters masking one another.

It was demonstrated that the application of k-means clustering to the t-SNE components projection of the coffee dataset was more accurate method to detect clusters.

In order to reduce dimensionality of numeric covariates and find the hidden clusters, t-SNE embedding was implemented, since it is capable of preserving local and global structure of the data, so that points that are close together in the original dimension tend to be close also in the low dimension. The total within-cluster sum of squares (WCSS) was plotted to find the true number of clusters. The implementation of K-means optimisation procedure along with the Elbow Method indicated 4 clusters (Figure 4).

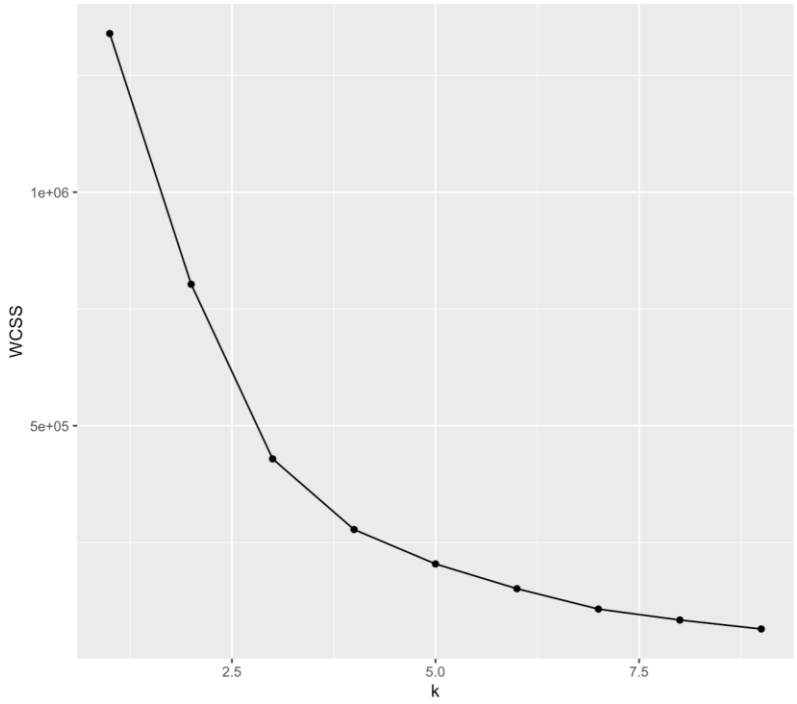

Figure 4: The plot of within-cluster sum of squares ot perform the Elbow Method

To partition the coffee data into groups based on the inherent structure of the data, the unsupervised clustering method -kmeans was applied. After setting the 'centres' argument to 4, the plot of clusters was created (Figure 5). The second validation techniques, silhouette analysis was applied to confirm the true number of clusters and it confirmed it was 4.

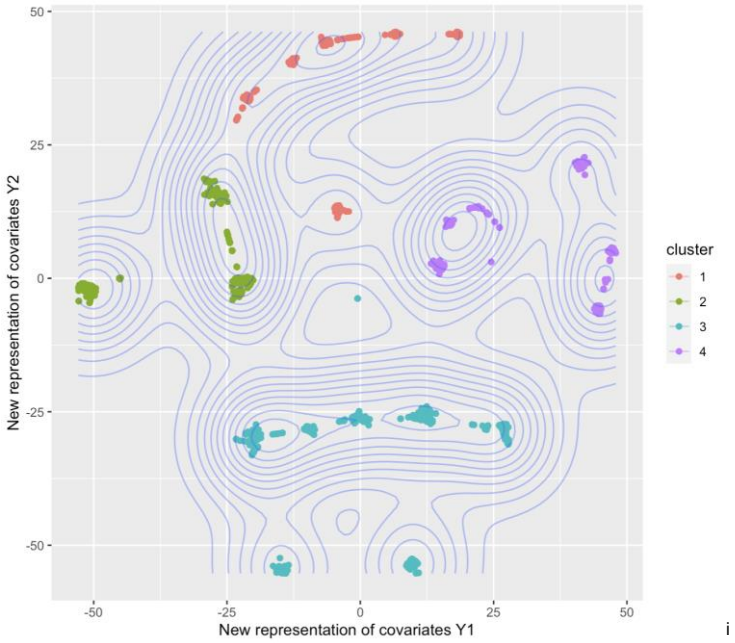

Figure 5: Four clusters computed by kmeans

6. Conclusions

The EDA has shown that the coffee data has some missing records which are MCAR, MAR and MNAR. Several different techniques have been used to remove them. The Total Cup Points has strong positive correlation with Cupper Point, Favour and Aftertaste, which could potentially be good predictors for the total grade.

There are many outliers in the data, which are, among others, the first indication that there are clusters among our data. These outliers occur because they belong to a different underlying population or have unique characteristics that distinguish them from the rest of the data. The applied unsupervised clustering method indicated that there are four clusters in the data, which can be linked to another variable in the dataset. The number of clusters has been confirmed by silhouette analysis.

Further EDA is recommended to fully understand the nature of the data. The map of the word could be upgraded by fitting a linear model and drawing the residuals into the same map, which would confirm the discovered trend.

It is also advised to apply Cross-validation to find the perfect perplexity hyperparameter.

---

i