Introduction

Chocolate, a product cherished worldwide for its rich flavor and diverse varieties, can be found in many of the products we use today. Whether it be a casual chocolate bar or dessert cake we can find it everywhere we go in our dishes, making it interesting to know a little bit more about it. This project delves into an extensive dataset of chocolate bars, scrutinizing the subtle yet significant factors that influence their ratings and consumer preferences. With an emphasis on the types of cocoa beans and their geographic origins, the study employs sophisticated data analysis techniques, including Gradient Boosting and Linear Discriminant Analysis, to unveil patterns in the chocolate data.

In a business context, such insights are valuable as they can guide chocolate manufacturers and retailers in refining their products, targeting specific market segments, and tailoring their marketing strategies to meet consumer preferences. By applying statistical and data science techniques, this project not only enhances our understanding of what makes certain chocolates more favored than others but also provides actionable intelligence for businesses operating in the chocolate industry such as where to locate factories and what types of beans to be used.

Data Description

Preprocessing

The preprocessing stage of the analysis involved several key steps to prepare the chocolate dataset for thorough examination. Initially, the dataset was loaded and reviewed to understand its structure. Column names were then modified for clarity and ease of use. Then there was the conversion of the 'CocoaPercent' from a string to a numeric format by removing percentage signs and converting the resultant values to numbers. Furthermore, the bean types were grouped into 5 different types ensuring a focused and relevant analysis: Criollo, Trinitario, Forastero, Blend and others with the first three being 3 of the 4 major bean types currently used in the industry. Also, the columns with missing data were dropped. In our case the columns “Bean Type” and “Broad Bean Origin” missing row contained the character “Â”. The resulting dataset was then ready to be analyzed.

Data Visualization:

After preprocessing, the grouped bean types were quantified and plotted as shown in figure 1 with Trinitario being the most common bean type in our dataset indicating its variety of use for different end-product chocolate types. We also wanted to see the count of the cocoa percentage as shown in figure 2 showing us a roughly similar plot to the normal distribution centered around 70%. The boxplot in figure 3 shows us that for all the bean type does not affect the final rating where all of them had roughly similar distributions and a mean rating of around 3.5. This indicates that the rating is affected by several factors rather than bean type only.

Lastly, a correlation matrix (figure 4) was computed and visualized for the numerical columns and showed us that there was no multicollinearity between these variables (expected due to the different info nature of each).

Model Selection & Methodology

For a business, if it wants to create chocolate to sell, it is important to have a good rating. To get that one needs to know what factors affect the ratings of chocolate. 2 models we tested, the random forest and the gradient boosting method.

With the random forest we needed to know which variables to choose from to make the model with the least MSE while capturing the most variance of the data. So first, we tested all the variables (obviously some overfitting here if all included) and then reduced the number of variables to a minimum by seeing the variable importance as shown in figure 5. we ended up with having the best model being with the following variables as shown in table 1 in the appendix.

Next, we tested the gradient boosting method (gbm)and this proved to be more effective in reducing the MSE. Similarly, testing for the different variable and seeing the relative influence of each predictor, we ended up choosing BeanType and CocoaPercent as our predictors as the others could cause overfitting and poor performance for out of sample dataset. So, for rating prediction if choose gbm as my model choice for the future work.

A third model tested was the Linear Discriminant Analysis (LDA). But this time, the bean type is to be predicted. Similarly the random forest and gbm, we tested several predictors and wanted to avoid overfitting. Having company and BeanOriginBarName as part of the predictors made our prediction accuracy as close to 99 percent which clearly shows overfitting, leading them to be dropped. The predictors Rating, ReviewDate, CocoaPercent and CompanyLocation contributed little to the model, so they were also dropped leaving the final model with only BroadBeanOrigin (the origin country of the bean) and giving us a reasonable accuracy (Table 2).

The business also is interested in knowing what similar traits the chocolate has leading us to clustering. First clustering method was a Principal Component Analysis (PCA) analysis conducted after dummifying the categorical variables BeanType and used the numerical variables ReviewDate, CocoaPercent and Rating. To see the effect of the other variables we attempted to also add BroadBeanOrigin into another PCA model however it proved to be hard to analyze due to the number of categories this predictor has. Same goes with company location.

The second clustering method, kmeans clustering, was performed where similar to PCA categorical variables 'BeanType' dummies was used to facilitate numerical analysis. Company and company location were not included due to the large number of categories this adds. The dataset was then combined with numerical variables such as 'CocoaPercent', 'Rating', and 'ReviewDate'. The optimal number of clusters was determined using the Elbow method and opted for five clusters as can be deduced from figure 10. The results were visualized using various plotting methods, each highlighting different aspects of the clusters, such as the relationships between key variables and the distribution of chocolates in each cluster.

Results, Predictions & Classification

Rating Prediction

Starting with the rating prediction, we had the random forest and the gradient boosting. After testing many predictors, we settled on myforest1 for random forest as it was able to capture the data pretty well with an mse of around 0.18. But the model which performed better without being overfitted is the gradient boosting gbm_model. The mse obtained with only the predictors “CocoaPercent” and BeanType was 0.17 which is slightly lower than the best random forest. To further test it, the dataset was divided into training set and test set with a division of 50% to see how good it performs if a new chocolate dataset is to be used and received an mse of 0.2 meaning that our model is pretty solid in predicting the chocolate rating. Any of the other variables can prove to be problematic as they tend to overfit the data (figure 6 showing one predictor dominating, same happened with the other variables) and perform poorly when tested as it increases the mse in an out of training dataset. It can be deduced that the rating given to the chocolate is mainly attributed to what chocolate bean is being used and the cocoa percent the chocolate has (Figure 7).

Bean Type Prediction

Next, we move to the category prediction with the lda model to predict the bean type we have on hand. As mentioned, the model was narrowed down to one predictor the bean origin (BroadBeanOrigin) as the numerical predictors didn't contribute to the model accuracy and the categorical variables tended to overfit the data to give 99% accuracy. The current model is giving us 75% accuracy giving us a good indication that for the most part, the bean type that can be obtained by the business can be mainly attributed to the country they are dealing with. This can be explained that usually in the modern world each country tends to focus on one type of bean due to weather and environmental factors surrounding the bean (its adapted to it). The confusion matrix in table 2 shows us how well the model worked for each bean type.

PCA

As mentioned, 2 PCA analyses were made to observe how the data clusters. Figure 8 shows us for the first PCA, we can make a few deductions about the predictors. Cocoa percentage and rating has little to no relation with the cluster indicated by the arrow direction being parallel to the clusters. This means that bean type has no effect on how good the chocolate is, and that the cocoa percentage is something controlled after the bean is obtained and nothing to do with the bean itself. The cocoa percentage and the rating arrows though were shown to be point opposite sides indicating a negative correlation between the two, meaning if one goes up, the other goes down. In the second PCA showed by figure 9 even though hard to understand due to the number of countries from which the beans come, we can see that as deduced from the lda that groups of countries are pointing to roughly similar direction to the bean type indicating there is a tendency that the bean type you get is attributed to a country you purchased. Ecuador's and bean type Forastero's arrow point to almost the same direction and length clearly showcase that each country specialized in certain bean types. This could be used by the business for their supply chain optimization to reduce distance traveled when buying the coco bean.

Kmeans

Another way to cluster the data is using kmeans which groups the data based on similarities. The characteristics of the five clusters that were created can be shown in table 3 where a few interesting deductions can be made. Cluster 5 shows us that the mean cocoa percent is high at around 90% but had a rating of 2.75 the lowest of the 5 clusters which can tell us that the higher the cocoa percent the less the rating the chocolate has (similar to the PCA deduction). A second notable observation is that when the mean cocoa percent was closer the 70 % the mean rating was highest around 3.3 as shown in clusters 2 and 4. As for the rest the clusters were similar in ways with each cluster having a more dominant bean type than the other.

Figure 11 and figure 12 show us the clusters were mainly visible around the cocoa percent and not clear when we look at figure 13.

Conclusion

The analysis of the chocolate dataset yields several insightful conclusions for businesses in the chocolate industry. The Gradient Boosting Model, using cocoa percent and bean type, emerged as the most effective tool for predicting chocolate ratings, suggesting these factors are pivotal in determining chocolate quality. The Linear Discriminant Analysis, centered on chocolate bean origin, successfully predicted bean types, highlighting the significant influence of geographical origin on bean characteristics where each country tends to specialize more into one type of bean. The PCA analysis indicated a limited correlation between cocoa percentage, rating, and bean type, and showed us opposite correlation between cocoa percentage and rating. The Kmeans

Clustering analysis revealed consumer preference patterns, notably a trend towards lower ratings for chocolates with higher cocoa percentages. This suggests a consumer preference for chocolates with moderate levels of cocoa. These findings can guide businesses in refining product quality, tailoring marketing strategies, and optimizing supply chain decisions.

The findings from the chocolate dataset analysis present valuable business applications. Firstly, emphasizing cocoa percent and specific bean types in product development can enhance chocolate quality, catering to consumer preferences. This insight aids in refining product offerings. Secondly, understanding the geographical influence on bean characteristics can inform sourcing strategies, ensuring the selection of beans best suited for desired chocolate profiles while reducing the supply chain costs as bean types did not affect the ratings. Finally, consumer trends towards moderate cocoa levels can guide marketing strategies, focusing on chocolates that align with these preferences, potentially boosting sales and customer satisfaction.

Appendix

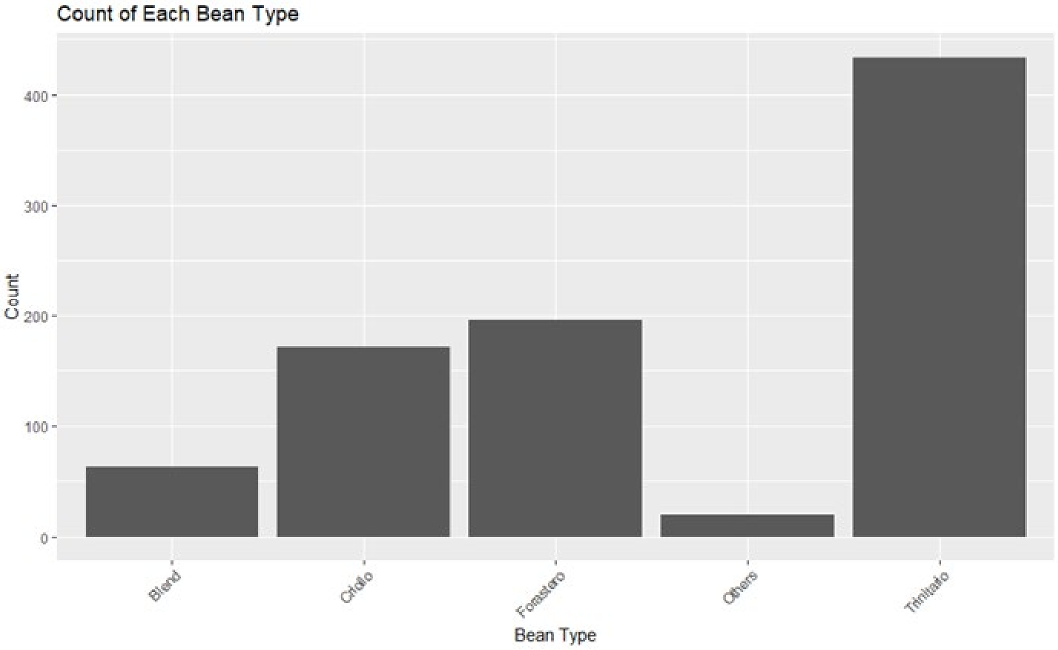

Figure 1 Bean type count

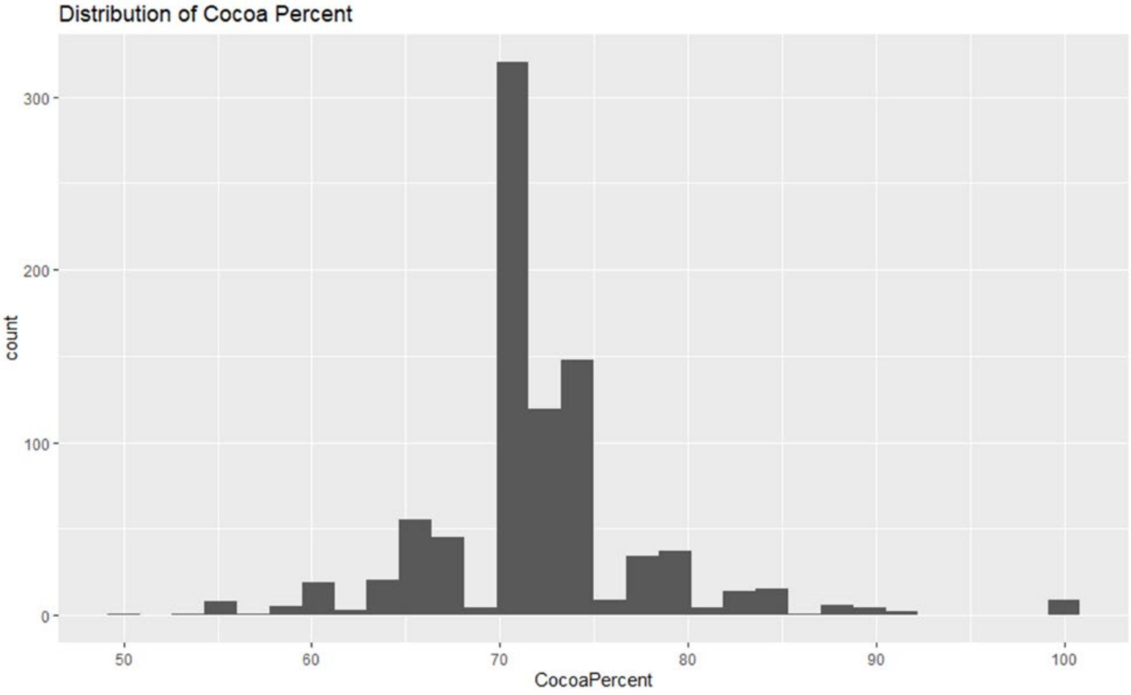

Figure 2 Distribution of cocoa percent

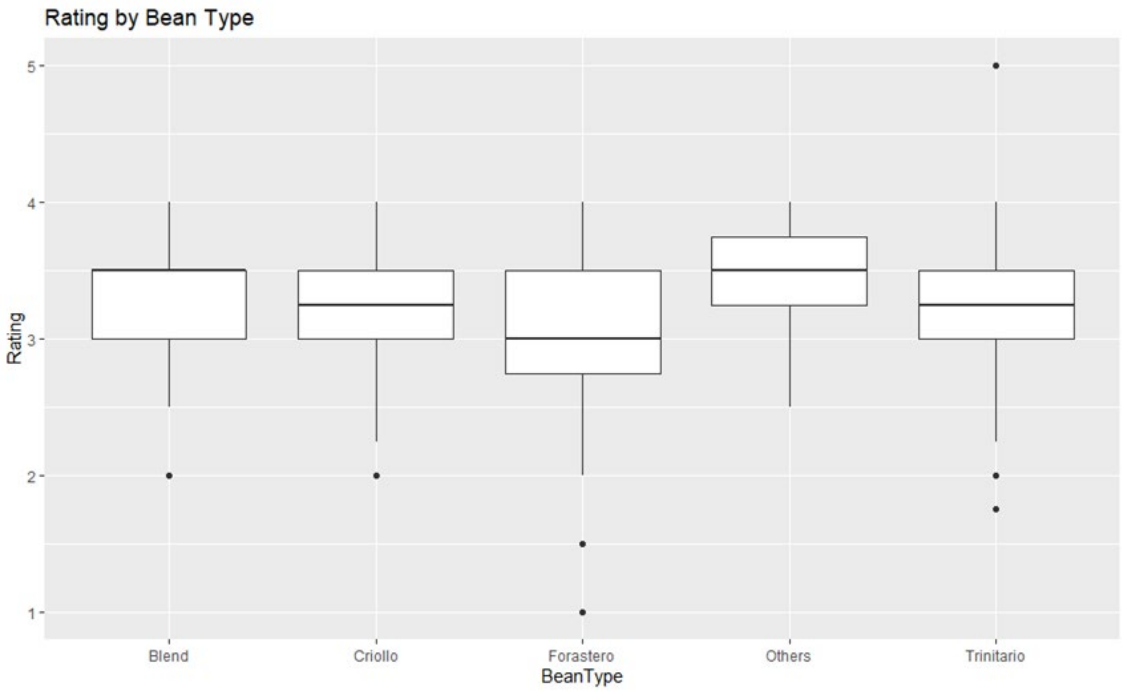

Figure 3 Box plot of rating vs bean type

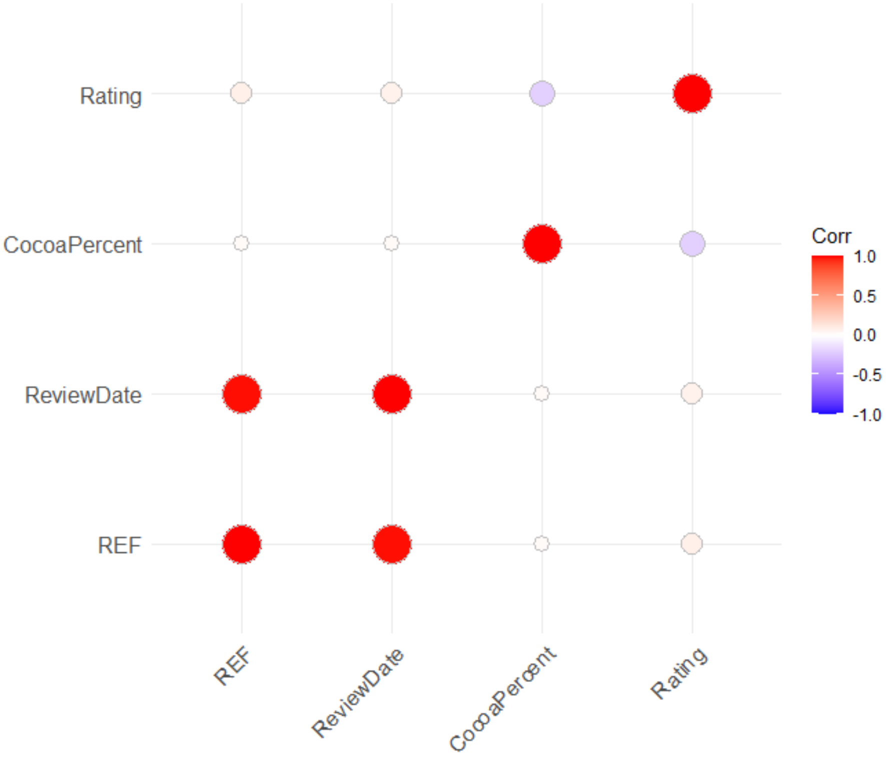

Figure 4 Correlation matrix of numerical variables

Myforest1 = randomForest(Rating ~ Company + CocoaPercent +CompanyLocation +BeanType, ntree=1000, data=chocolate_data, importance=TRUE, na.action = na.omit)
gbm_model = gbm(Rating ~ CocoaPercent + BeanType, data = chocolate_data, distribution = "gaussian", n.trees = 10000, interaction.depth = 4)
lda_model = lda(BeanType ~BroadBeanOrigin, data = chocolate_data)

Table 1 Finals models for Random Forest, Gradient Boosting and Linear Discriminant Analysis

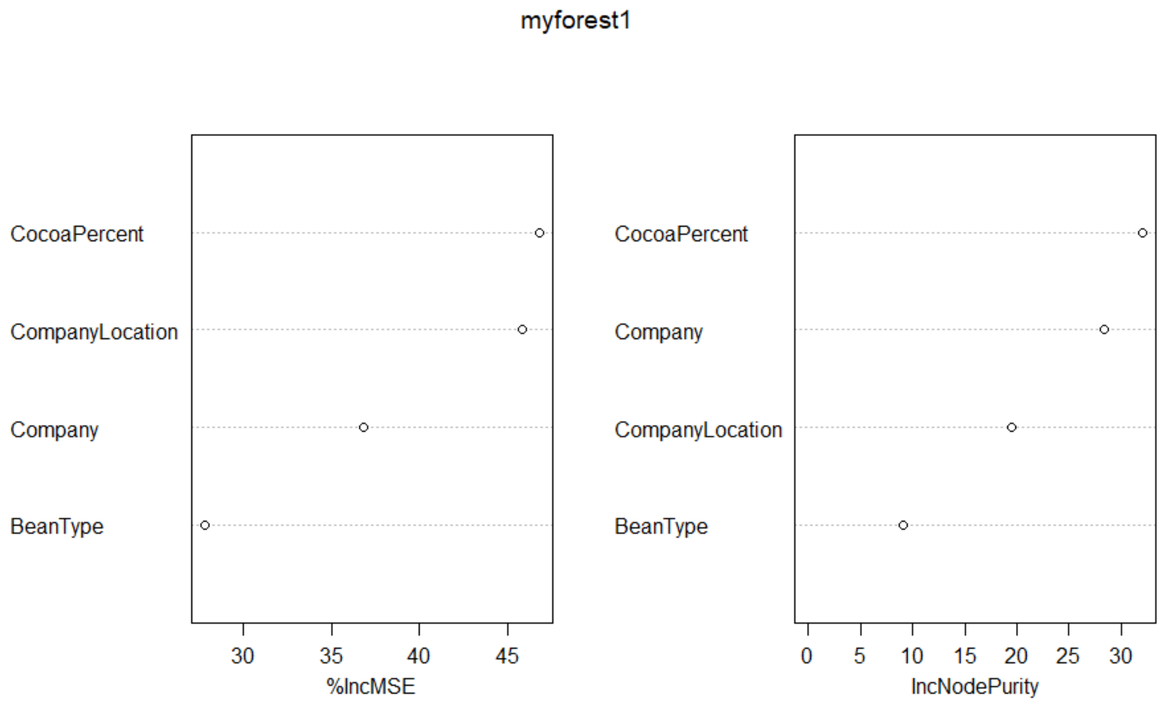

Figure 5 Variable importance for random forest model myforest1

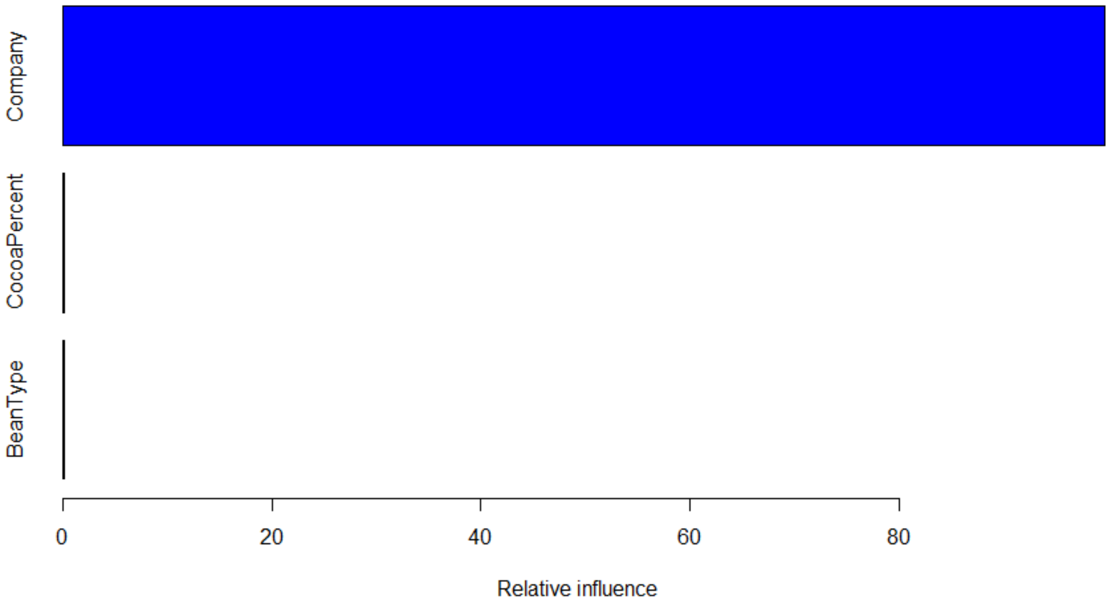

Figure 6 Overfitted gradient boosting method variable influence with company as a predictor

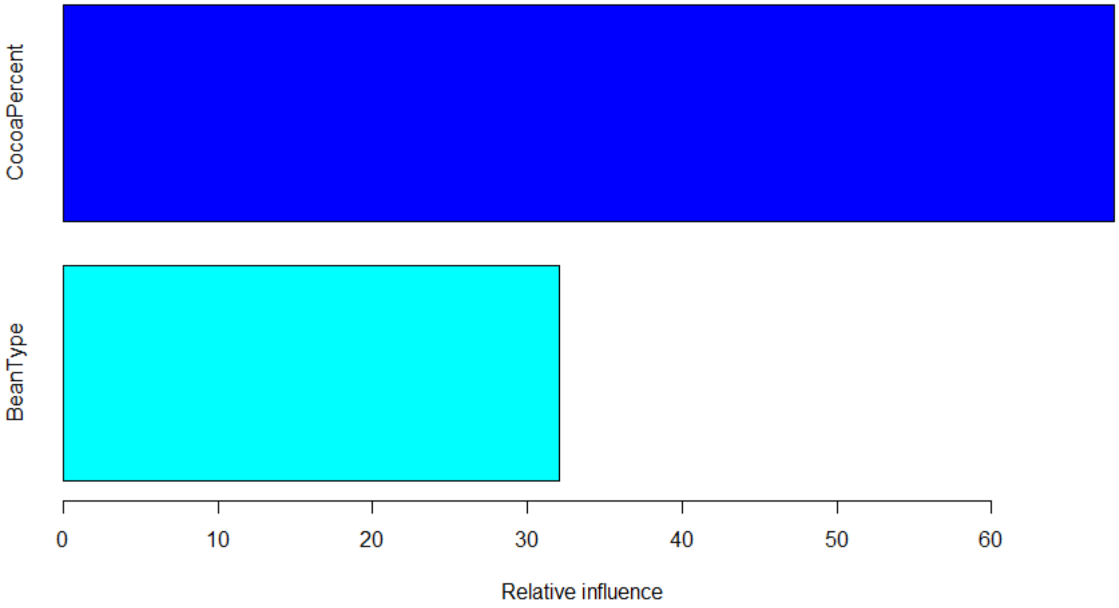

Figure 7 Final gradient boosting method variable influence.

Predicted/Actual	Blend	Criollo	Forastero	Others	Trinitario
Blend	32	6	0	0	10
Criollo	7	135	27	0	75
Forastero	1	2	154	7	16
Others	2	8	2	11	3
Trinitario	21	20	13	2	330

Table 2 Confusion matrix of the lda model

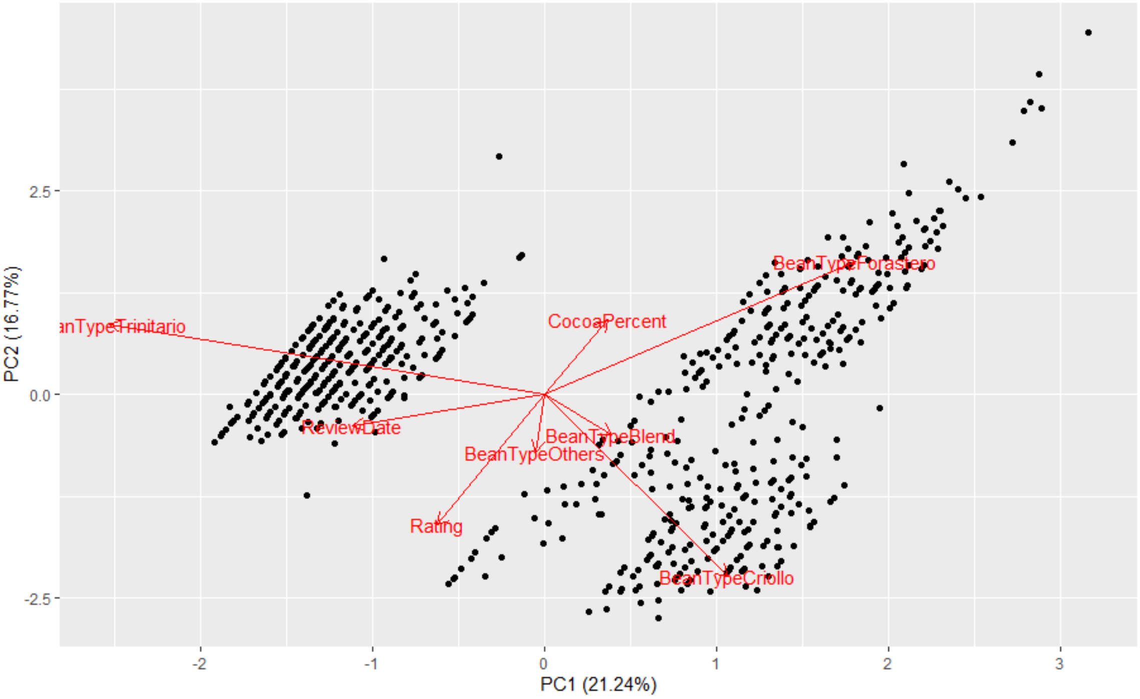

Figure 8 Table 3 PCA model 1 autoplot

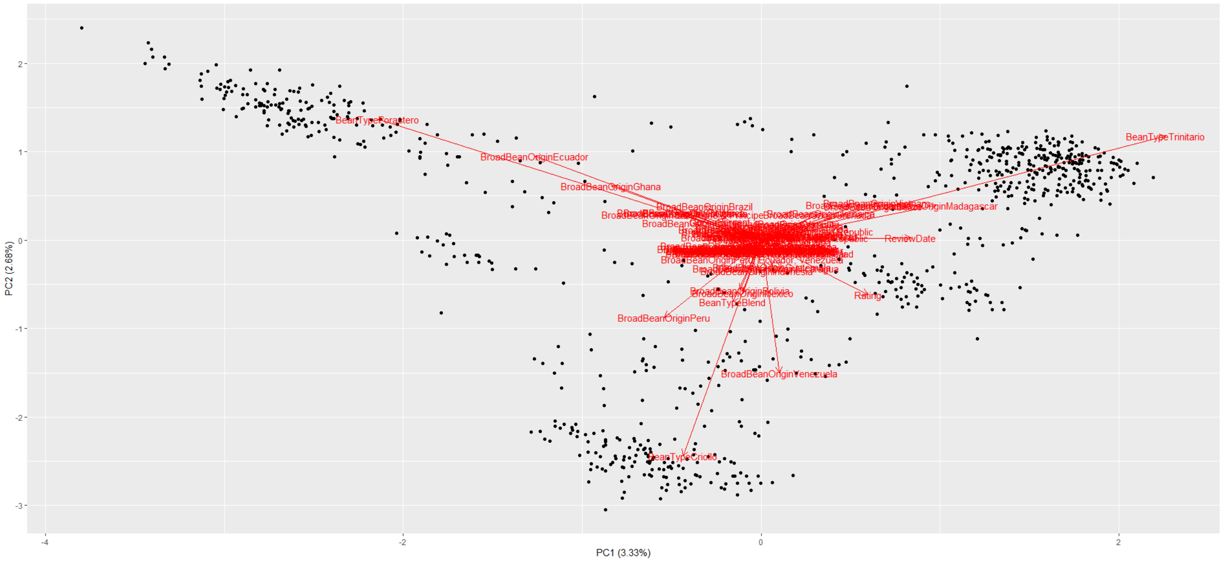

Figure 9 PCA model 2 autoplot

Cluster	Cocoa Percent	Rating	Review Date	Blend	Criollo	Forastero	Others	Trinitario
1	62.32	3.17	2009.98	0.03	0.16	0.32	0	0.48
2	71.51	3.28	2008.44	0.08	0.25	0.23	0.02	0.43
3	76.95	3.15	2012.84	0.07	0.19	0.2	0.04	0.51
4	70.04	3.31	2014.13	0.08	0.18	0.17	0.03	0.55
5	89.97	2.75	2011.24	0.11	0.14	0.46	0	0.3

Table 3 Kmeans cluster characteristics

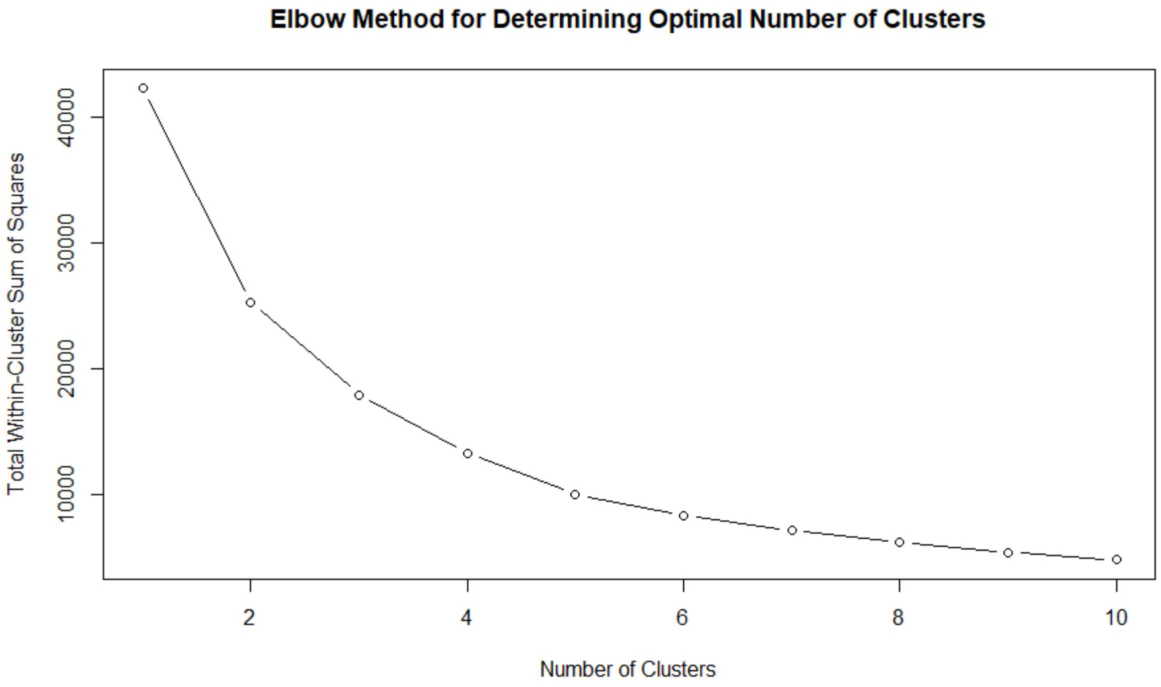

Figure 10 Kmeans Elbow Method for determining optimal number of clusters

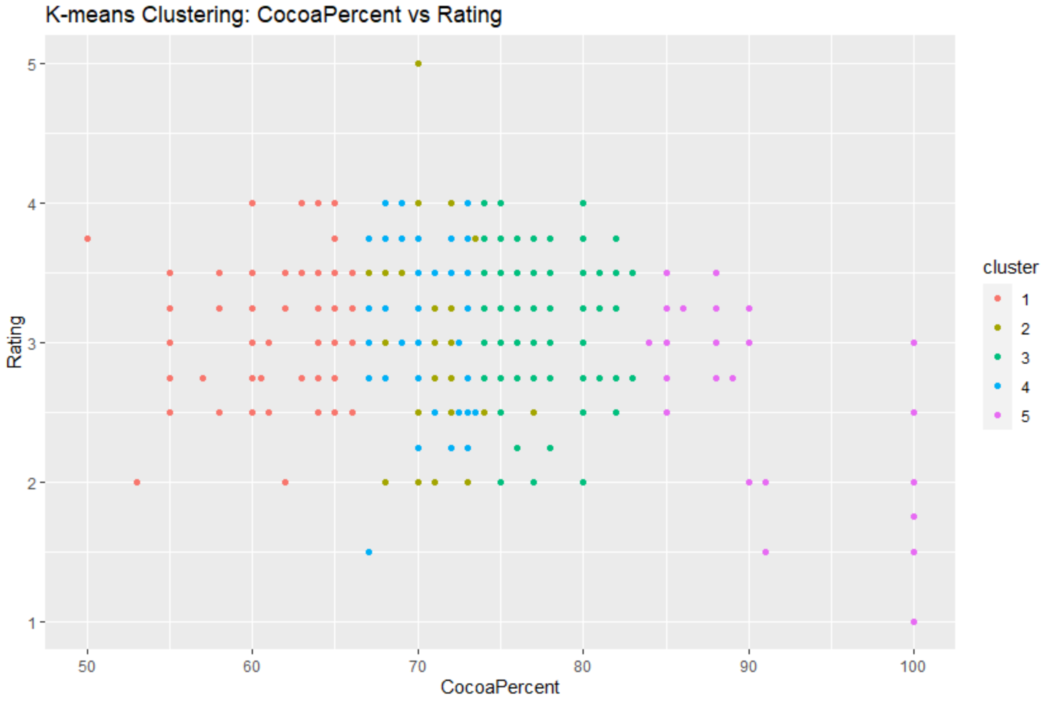

Figure 11 Kmeans Clustering Cocoa percent vs Rating

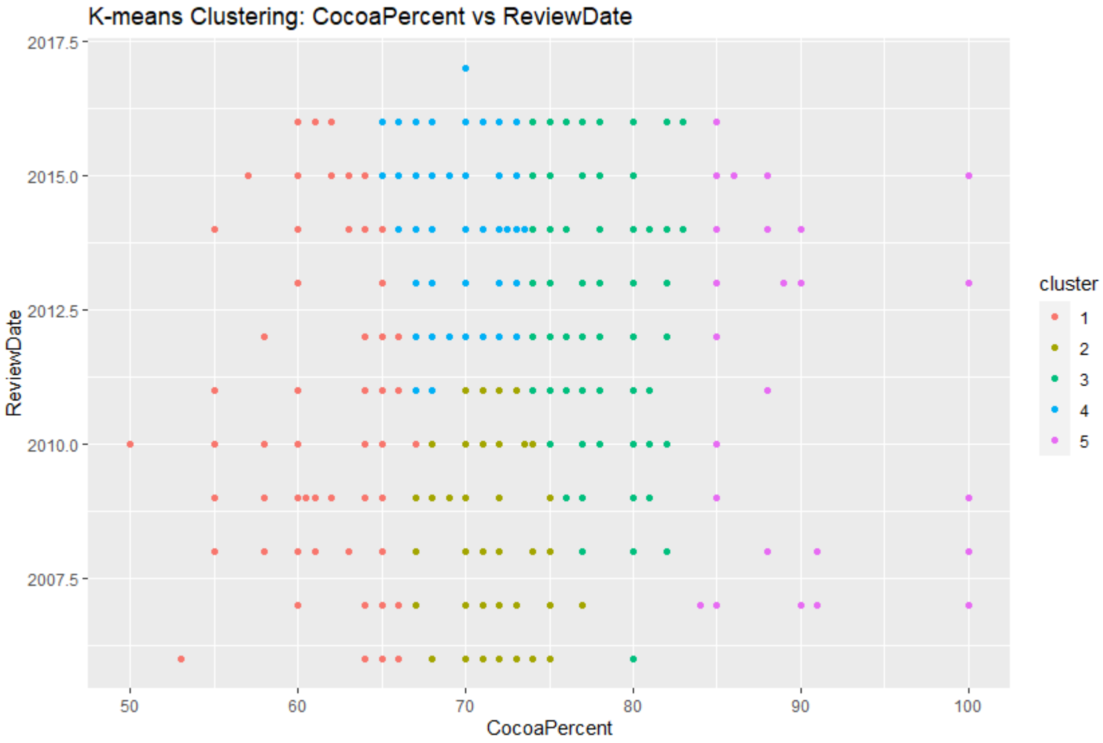

Figure 12 Kmeans Clustering Cocoa percent vs Review Date

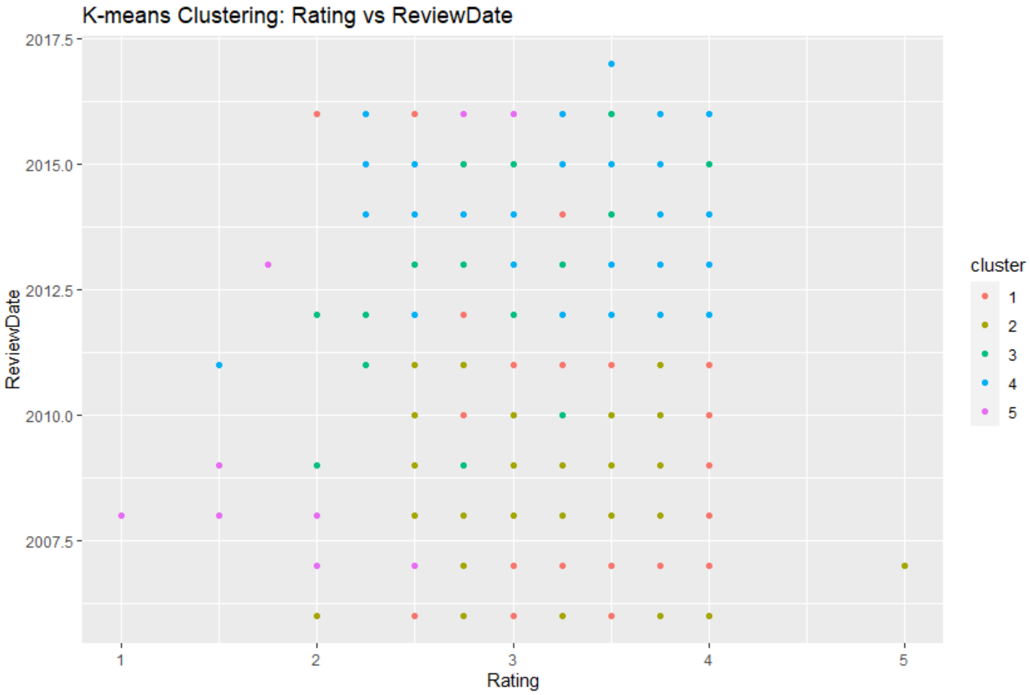

Figure 13 Kmeans Clustering Rating vs Review Date