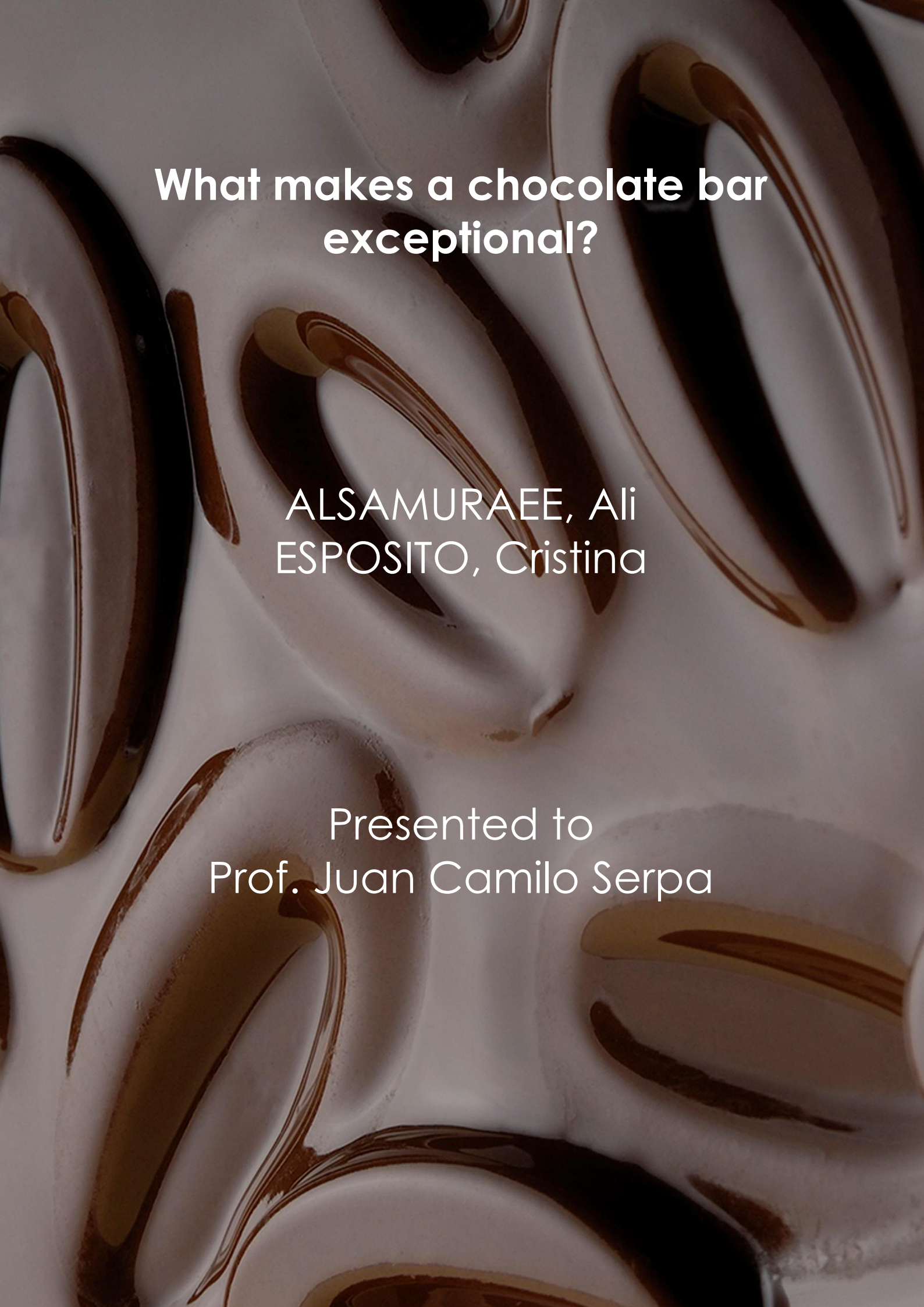


What makes a chocolate bar exceptional?

ALSAMURAEE, Ali
ESPOSITO, Cristina

Presented to
Prof. Juan Camilo Serpa

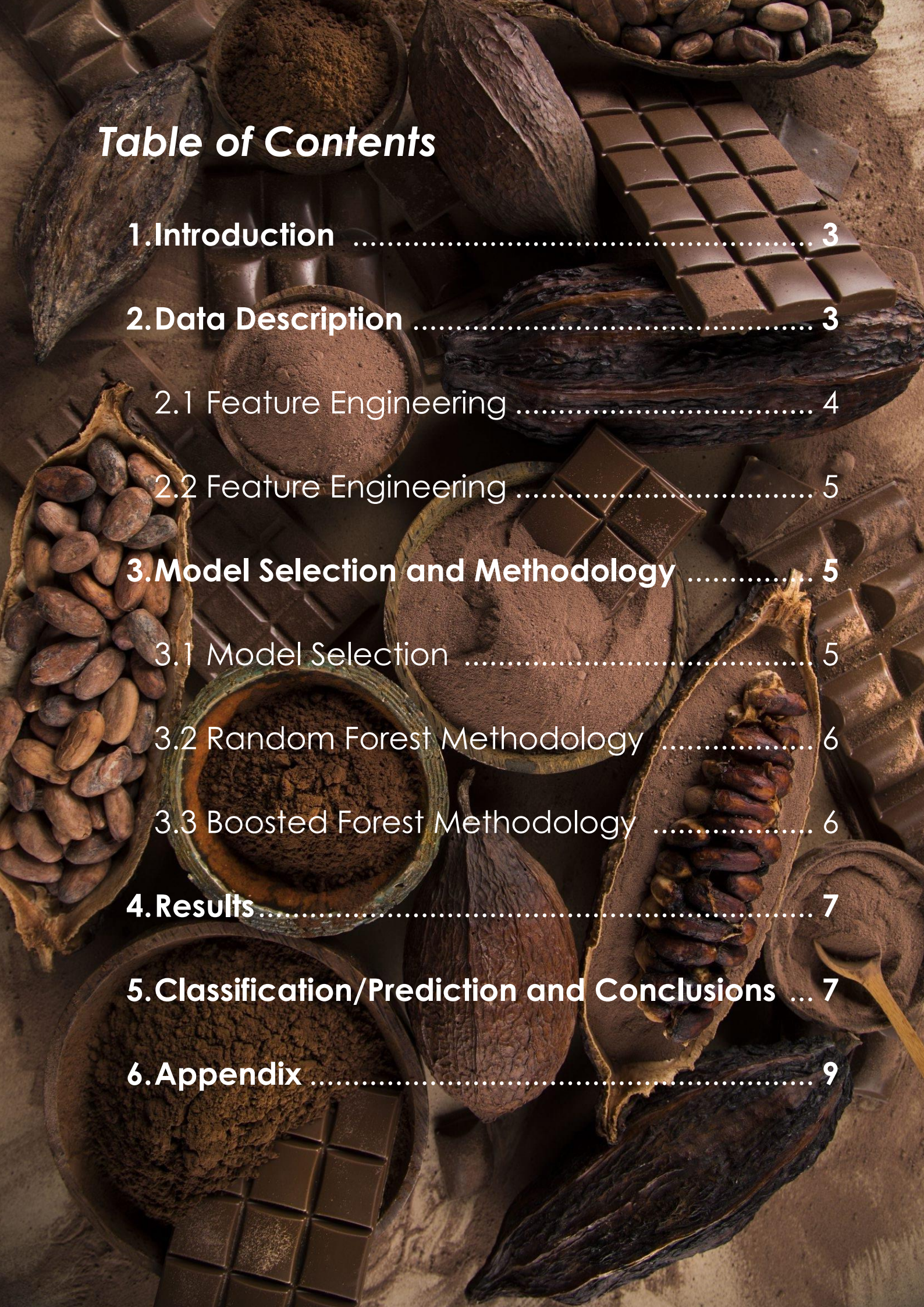


Table of Contents

1 Introduction

The history of chocolate dates back 4,000 years, to ancient Mesoamerica which included, among others, today's Mexico, where the first cocoa trees were found. The Olmec civilization, one of the oldest in Latin America, was the first to turn cocoa into chocolate. It was made into infusions that they drank during rituals and that was used as medicine. They regarded chocolate as the drink of the gods. Years later, human beings have been trying to perfect the art of making chocolate. But the question is, what does it take to make the best kind of chocolate?
To understand this, our goal is to discover what defines a good chocolate and predict chocolate ratings around the world. The dataset contains 2452 chocolate bar ratings, focusing primarily on plain dark chocolate with an aim of appreciating the flavours of the cacao when made into chocolate.

The objectives that we have set for this project are:

a) To study the dataset, get a feeling of each variable, their meaning, and how they could impact the rating of a chocolate bar
b) Explore relationships between variables, through skewness, correlation, and other interesting facts such as multicollinearity, and outliers
c) Explore relationships between the target variable (chocolate rating) and potential predictors.
d) Test a variety of tree-based models, focusing primarily on random forest and boosted trees
e) Build models based on intuition and insight gained from the data exploration phase. By generating out of bag error ratings and comparing performance we evaluate ways to enhance the model
f) Test on out-of-sample and look for overfitting issues. Continue to enhance model until there is no apparent overfitting issue

2 Data Description

2.1 Feature Engineering

In order to have a better understanding and get some quick takeaways, we decide to create the following variables:

- Company Continent → based on the Company Location Country
- Continent of bean origin → based on the Country where the bean originates from
- First characteristic → extracted from the Most memorable characteristics
- Rating Class → based on the rating (See Appendix Table 1)
- Ingredients column was broken down into numerical and categorical variables (See Appendix Table 2)

2.2 Data Exploration

As we explore the data in more depth, we notice that from all the chocolates rated, 1,088 (44.4%) were manufactured in the US, 175 (7.1%) in France and 163 (6.6%) in Canada. This could be explained by the fact that the taster is based in the US.

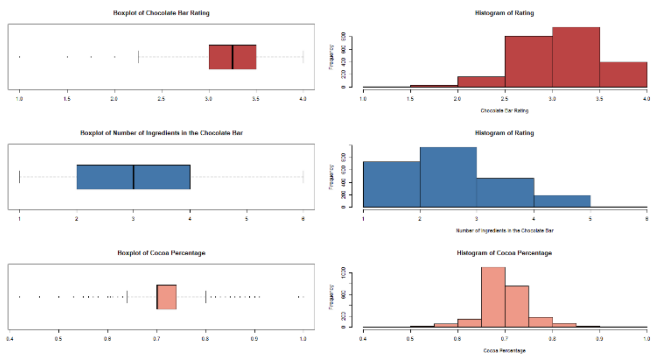


Figure 1: General overview of rated chocolate bars

Looking at general trends, we noticed that our chocolate bar ratings were on average good, as 75% of the data has a rating above 3 on a scale of 5 (skewed to the right). The number of ingredients in a chocolate bar is between 1 to 6, where half of the bars have 3 ingredients or less. It can be seen that our taster prefers dark chocolate, as only a quarter of rated chocolate[1] bars were not considered to be healthy dark chocolate, the rest being above the threshold.

Not surprisingly, cocoa beans are primarily from South America (44.1%), specifically from Venezuela, Peru, Ecuador, Bolivia, Brazil and Colombia. North America accounts for 25.1%, in which Dominican Rep. and Nicaragua take the lion share. Africa has a share of 14.2%, through Madagascar, Tanzania and Ghana, respectively. (See Appendix Figure 9)

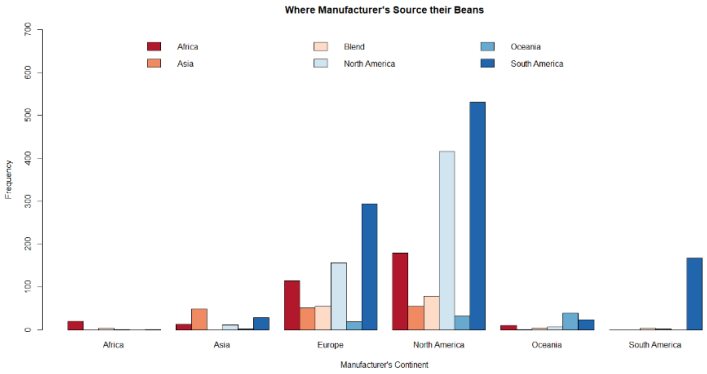


Figure 2: Manufacturers Geographical Bean Supply

To take it a step further, we have decided to understand the relationship between the origin of manufacturers and the origin continent of cocoa beans. We noticed that North American and European manufacturers source most of their beans from North & South America, whereas African and South American manufacturers source their beans primarily from their own shores.

Knowing more about the story, we were intrigued by the million-dollar question: What makes a chocolate bar exceptional?

A segment to look at is the ingredients in the chocolate bars, as they are not created the same way. Some have more ingredients than others, almost none (6 observations), have one ingredient only.

(See Appendix Figure 8) It gets more interesting when we're going from two to three ingredients, as the average rating increases from 3.22 to 3.27. As we dig deeper, 725 bars have 2 ingredients (30.66%) and 974 have 3 ingredients (41.20%), yet the difference between the disappointing results are relatively close (157 vs 169). Visually speaking, having three ingredients seem more beneficial than two. On average, three ingredients seem to be the best mix.

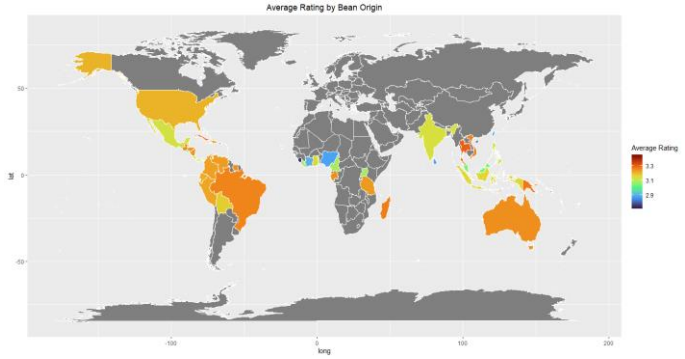


Figure 3: Average Rating by Bean Origin

Putting into perspective bean's origin and its rating, we noticed that western African beans generate below average ratings, while south Asian beans are slightly better. South American beans are above average and have very low variance between them, the same can be said for central America. At this point, African countries besides Tanzania and Madagascar seem to have beans of lower quality.

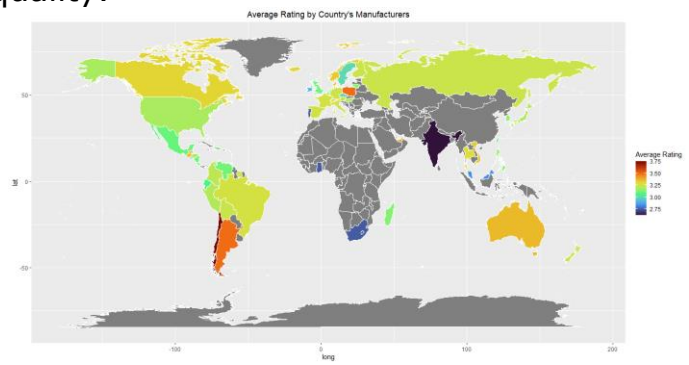


Figure 4: Average Rating by Country's Manufacturers

Canadian manufacturers have a better influence than American ones. In fact, Canadian manufacturers seem to be ranked as one of the best in the world, only behind Chile, Argentina, Poland and Australia. India produces the worst chocolate bars. However, this is only one bar that was rated. This is a clear example of bias due to a small sample, which we will not overlook.
We also did sentiment analysis based on the characteristics given by the taster first. To do this, we used Word cloud , in which, literally, an image describes a thousand words. This could be helpful to understand the qualitative aspect of chocolates and their rating.

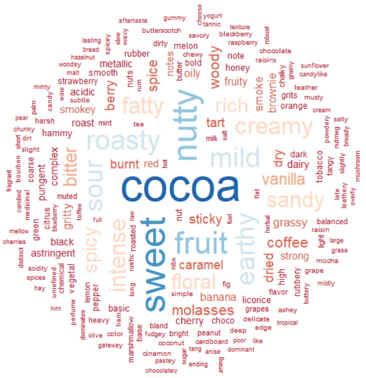


Figure 5: Word Cloud based characteristics

Furthermore, a correlation matrix was generated to quantify the relationship between predictors to analyze collinearity issues (see Appendix Figure 10).

The relationships in blue indicate a positive correlation and in red a negative correlation. The matrix determines that some predictors are excessively correlated (absolute value exceeding 0.7). Sugar and sweetener are logically negatively correlated, as there's no need to have both in the same chocolate bar. The number of ingredients is correlated with cocoa butter, vanilla, and lecithin. Intuitively speaking, when adding an ingredient, the number of ingredients increases, especially if it is a primary ingredient. We will have to consider this during our modelling process, as collinearity is present in this dataset.

3

Model Selection and Methodology

3.1 Model Selection

The two models we will be using for comparison are random forests and boosted trees. We decided not to incorporate regular decision trees, as this model tends to overfit the data. The goal is to select the best model that will provide the lowest mean squared error (MSE) from out of bag performance when predicting the chocolate bar ratings.
Since our dataset contains new features that are highly interconnected with their original features, we decided to split up our variables into two different groups:

- Feature Group 1: Company Continent, Bean Origin Continent, Cocoa Percentage, and Ingredients

- Feature Group 2: Company Continent, Bean Origin Continent, Cocoa Percentage, Number of Ingredients Used, if the chocolate bar contains sugar, if it contains cocoa butter, if it contains lecithin, if it contains salt, if it contains sweetener, and if it contains vanilla

We did not include the country for the manufacturer and the bean origin because both variables contain more than 55 levels, which is a limitation of the tree models we will be using. In addition, we also did not include if the chocolate bar contains beans, as this variable was completely univariate (i.e. all chocolate bars contained beans). Lastly, any records that contained blanks for Number of Ingredients were dropped, as we gage that this will be an important variable to include in the models.

3.2 Random Forest Methodology

Our approach with the random forest model was to first run a model for each of the feature groups with 500 trees and plot the MSE results to help us determine which number of trees to use. The number of trees to use impacts the performance of the model. Specifically, the higher the number of trees you use, the higher the possibility that the model will be overfitting. When looking at the plots for both feature groupings, we can see that the best number of trees is 40 (See Figure 6) and 30 (See Figure 7) respectively.

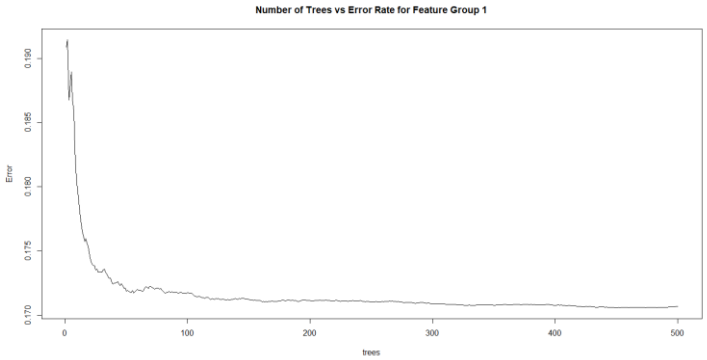


Figure 6: Number of Trees vs. Error Rate for Feature Group 1

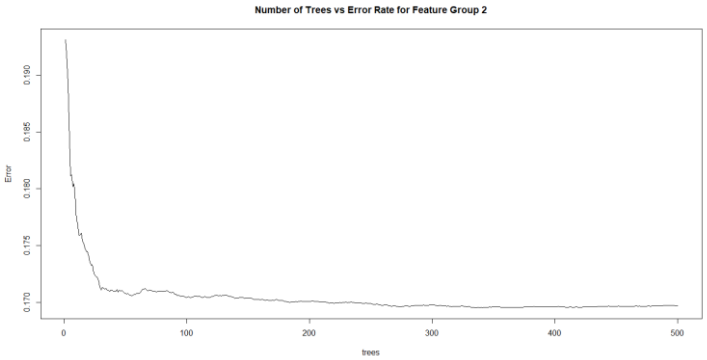


Figure 7: Number of Trees vs. Error Rate for Feature Group 2

Once the optimal number of trees is chosen, we rerun a new model to help us determine the variable importance. For feature group 1, the most important variables are the cocoa percentage and the ingredients in the chocolate bar (See Appendix Figure 11).

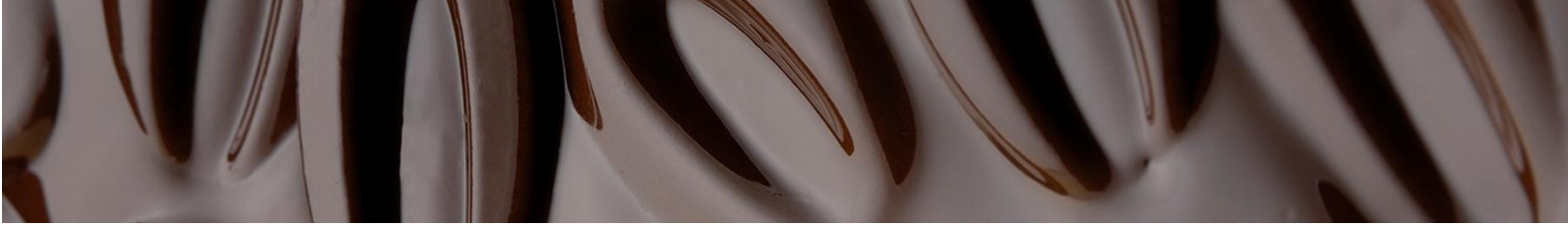
For feature group 2, the most important variables are cocoa percentage, the continent the manufacturer is from, if the chocolate bar contains vanilla, and number of ingredients in the chocolate bar (See Appendix Figure 12).
Lastly, once the most important variables were chosen, we build a final model with the chosen variables. The entire process will result with three models per feature group, totalling six different random forest models.

3.3 Boosted Forest Methodology

Our approach with the boosted forest was to apply the same steps that were used with the random forest model. The first step was to run a boosted model with all predictors for both feature groups with 500 trees and obtain a plot with the MSEs to help us determine the best number of trees to use. When looking at the plots for both feature groupings, we can see that the best number of trees is 18 (See Appendix Figure 13) and 21 (See Appendix Figure 14) respectively.

Once the optimal number of trees is determined, we produce a variable importance plot to determine the most important variables. For feature group 1, the most important variables are the cocoa percentage and the ingredients in the chocolate bar. For feature group 2, the most important variables are cocoa percentage, if the chocolate bar contains vanilla, and the continent the beans are from. (See Appendix Figures 15 & 16)

Lastly, a final model is produced with the best features. Similar to the random forest, the entire process will result with three models per feature group, totalling six different boosted forest models.



4 Results

After running each possible combination, we get the following results:

Tested model	Random Forest MSE
Feature Group 1, 500 trees	0.1707
Feature Group 1, 40 trees	0.1725
Feature Group 1 (best features), 40 trees	0.1718
Feature Group 2, 500 trees	0.1697
Feature Group 2, 30 trees	0.1711
Feature Group 2 (best features), 30 trees	0.1758

Tested model	Boosted Forest MSE
Feature Group 1, 500 trees	0.1430
Feature Group 1, 18 trees	0.1666
Feature Group 1 (best features), 18 trees	**0.1684**
Feature Group 2, 500 trees	0.1458
Feature Group 2, 21 trees	0.1683
Feature Group 2 (best features), 21 trees	0.1702

Although the best MSE is a boosted forest with 500 trees and includes all variables from feature group 1, we feel that the best model is a boosted forest with 18 trees and only the best features from feature group 1. There are a few reasons why we chose this model. The first reason is that the boosted model with 500 trees is most likely overfitting the data, as the difference in MSE is not significant. The second reason is that the difference in MSE between a boosted forest with 18 trees and all features from group 1 is not significant with the chosen model, meaning that the two variables that were selected from the feature importance plot can produce a very similar result.

Of course, the more predictors one includes in a model, the more complex it becomes, as well as the possibility of overfitting. One may argue that selecting the model with all features from group 1 should still be sufficient because it only contains four variables (as opposed to the two from the selected features). However, if two features can produce similar results as a model with four features, this means that there is noise coming from these other variables.

5 Classification/Prediction and Conclusions

Two MMA students are aiming to open their chocolate factory in the next few months. They are now in their testing phase to bring exotical flavours from all around the world to their store Wonka Willy. Their first product is called Alietto and the second is Cristinez. Having limited funds, they are faced with a tough decision, they can only afford to promote one chocolate bar.

Both chocolate bars have different characteristics and through our model, we aim to have a superstar bar.

Name of bar	Manufacturer's Country	Company's Continent	Country Bean's Origin
Alietto	Canada	North America	Madagascar
Cristinez	Canada	North America	Brazil

Name of bar	Continent's Bean's Origin	Specific Bean's Origin	Cocoa (%)	Ingredients
Alietto	Africa	Mava Sa	0.82	3- B,S,C
Cristinez	South America	Bahia Superior	0.68	4- B,S,C,L

According to our model and taking into account the manufacturer's continent, the continent's bean's origin, the cocoa percentage and the ingredients, Cristinez seems to outrate Alitetto.

Name of bar	Prediction
Alietto	3.14
Cristinez	3.29

Based on these results, Wonka Willy should focus on promoting Cristinez. They could also continue to test other flavours and get an even better score. Despite the fact that this model can guide through different decision, we believe that we can’t solely base ourselves on this model for several reasons. First, our model is based on ratings from one individual only. Although it creates less homogeneity, it is a clear bias. To be able to get comfortable with a model, we need several individuals to participate. Second, the ratings are mostly based on dark chocolate, thus the confidence of a prediction regarding non-dark chocolate (less than 70%) is low. As this stand, the Cristinez result is slightly questionable. Third, people tend to change in time and this individual taster is not an exception. Also, he has been rating chocolate at a growing pace and peaked in 2015, which in some sense could alter the results. (See Appendix Figure 17)

In conclusion, our model can be considered accurate, considering its MSE. It could be complementary to other tools to take key business decisions. No model is perfect and ours is no different for the reasons stated above. Yet, by taking its weaknesses we could improve it and bring it to another level.

6 Appendix

Data dictionary

The dataset we will use contains the following 10 variables:

- Reference number (REF)
- Company (Manufacturer)
- Company Location Country
- Reviewed Date
- Country where the bean originates from
- Specific origin of the beans or bar name
- Coco percent
- Ingredients in the chocolate bar
- Most memorable characteristics
- Chocolate bar rating

Table 1: Rating Class Buckets

Grade	Category
1.0-1.9	Unpleasant
2.0-2.9	Disappointing
3.0-3.49	Recommended
3.5-3.9	Highly Recommended
4.0-5.0	Outstanding

Table 2: Division of ingredients into "new" variables

Created variables by ingredients
Number of ingredients
If it contains beans
If it contains sugar
If it contains sweetener
If it contains cocoa butter
If it contains vanilla
If it contains lecithin
If it contains salt

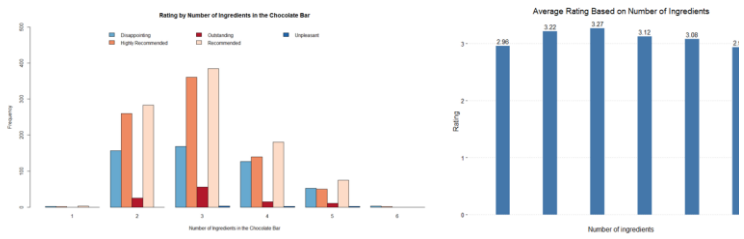


Figure 8: Number of ingredients fragmentation

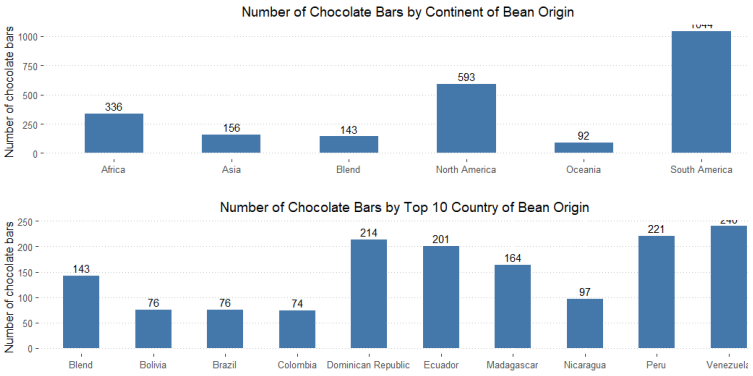


Figure 9: Chocolate Bars Classification based on Bean Origin

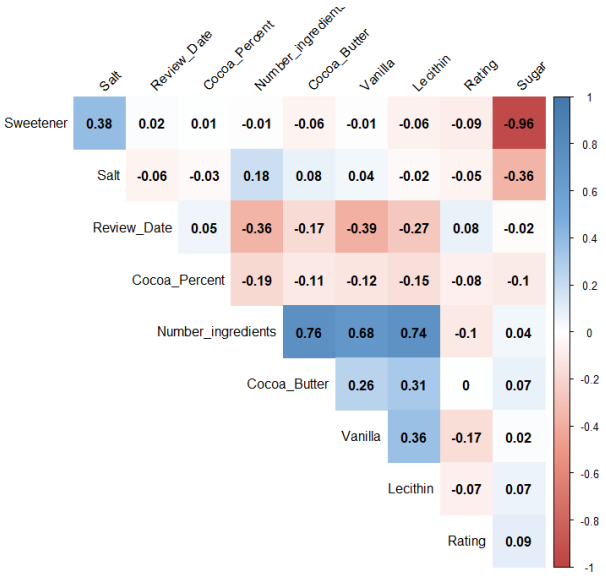


Figure 10: Correlogram of variables

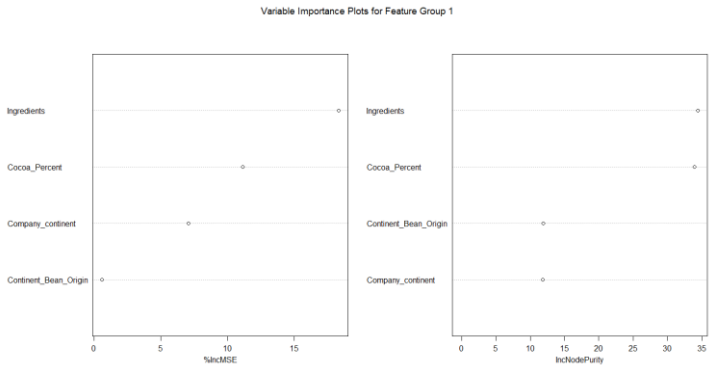


Figure 14: Variable Importance Plot for Feature Group 1 for Random Forest Model

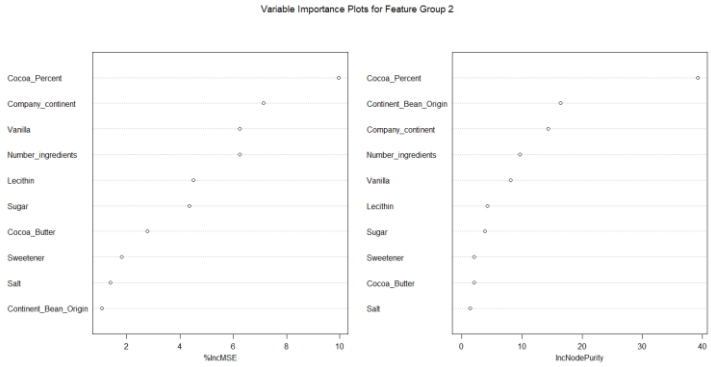


Figure 15: Variable Importance Plot for Feature Group 2 for Random Forest Model

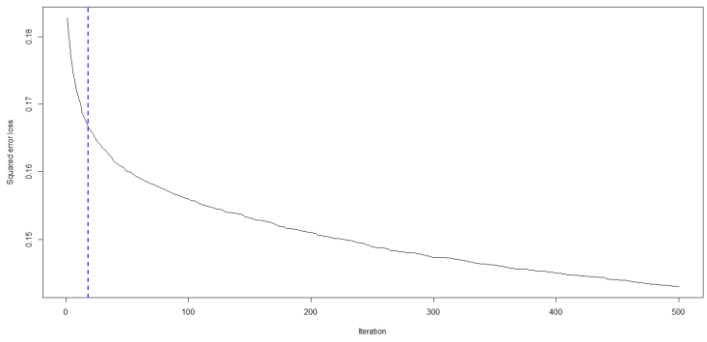


Figure 16: Number of Trees vs. Error Rate for Feature Group 1 Boosting Trees Model

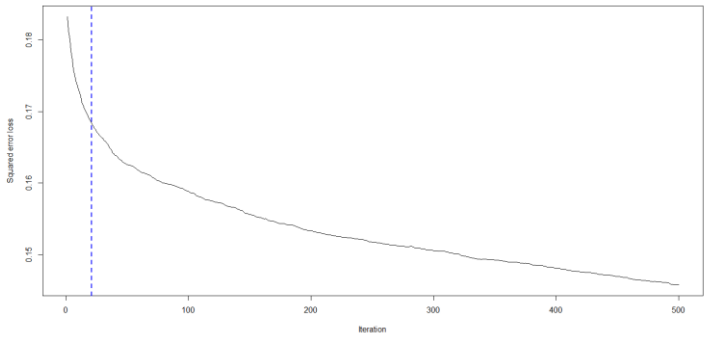


Figure 17: Number of Trees vs. Error Rate for Feature Group 2 Boosting Trees Model

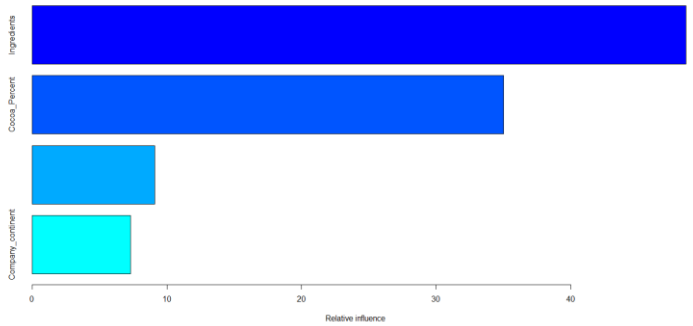


Figure 11: Feature Importance Graph for Feature Group 1 for Boosted Forest Model

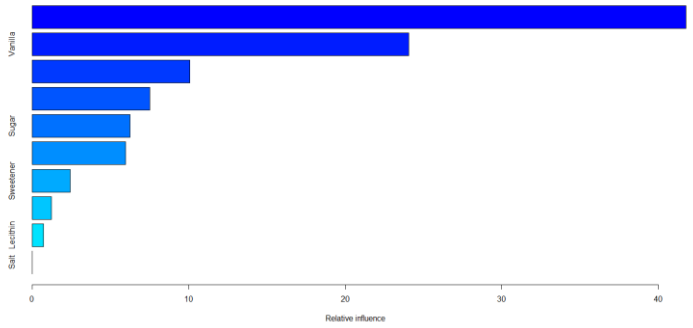


Figure 12: Feature Importance Graph for Feature Group 2 for Boosted Forest Model

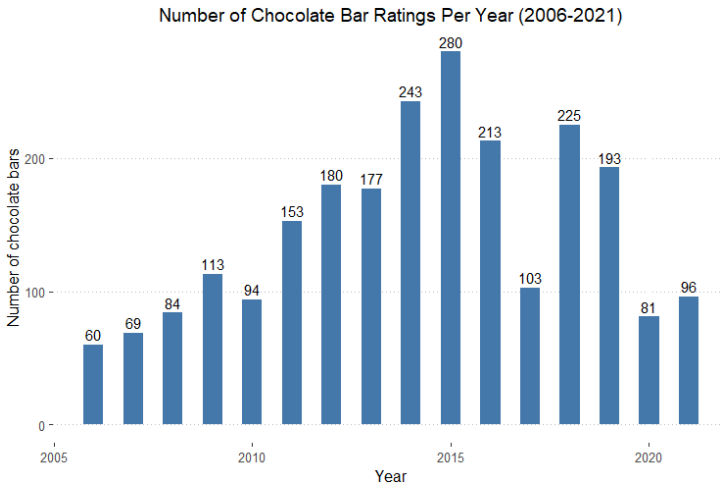


Figure 13: Number of chocolate bar per year (2006-2021)