**Data Mining on Chocolate Bar**

Wong, Ho Sum Student

The Open University of Hong Kong

**Introduction**

Since chocolate bar is one of the famous snacks in the world, and most of the people will use chocolate as a gift. But somebody will have an allergy to chocolate, so they cannot choose a good chocolate for a gift to other. Therefore, if will can predict the rating of a chocolate bar it will be very useful for them. And we are aiming for build up a model for predicting the chocolate rating. In this project, I am going to use the data set collected from 2006 to 2020 by Brady Brelinski, founding member of the Manhattan Chocolate Society.

**Exploratory and descriptive analysis of data**

This data set consist of 21 variables and 2224 rows observations with no missing value. The list of variables shown on Fig.1. In this project we are looking the rating of chocolate bar, the distribution of chocolate bar can be seen on Fig.2. The figure shown that most of the rating of chocolate bar are higher than 2.5. Since not all the variables will be consider as variable X, so we will drop the unuse variables and the list are shown on Fig.3. After choosing the variable we need, we can see the relationship between those variables on Fig.4, as we can see most of the variables X do not have high correlation with Y which is the rating of chocolate bar. And some of the variables X got weird distribution such as ‘salt’, because of most of the observations did not using salt as ingredients (The distribution shown on Fig.5).

**Analysis of Data & Finding**

Before we are doing analysis of data, we will split the data into two parts randomly. And the training data consist of 80% of data and testing data consist of 20% of data, which is base on Pareto’s Principle. Pareto’s Principle rules are telling us “20 percent of something always are responsible for 80 percent of the results” (John, 2005). And we will split the variable Y in to 3 classes and 2 classes (The distribution of variable Y shown on Fig. 6 & 7). We will perform 5 models in total, which is SVM with 3 labels, SVM with 2 labels, Logistics regression, decision tree with 2 labels and Random Forest with 2 labels. We will compare the accuracy of these five models. The result of these five models is shown on Fig.8. From the result we can see that SVM with 3 label got very low accuracy, so we will not be using it. SVM with 2 label and logistics regression model got 63% accuracy which is a not bad result. On the other hand, Decision Tree and Random forest got 59% and 61% accuracy respectively.

If will only consider of accuracy of these four models, we cannot choose the best classification model. So, we will consider of class accuracy. From Fig.8 we can know that both models can have good performance to predict class 1 which mean good chocolate, but Decision Tree and Random Forest only have 32% and 28% accuracy for predicting class 0.

**Conclusion**

We cannot choose the best model base on the total accuracy and class accuracy, but we know that SVM with 2 label and Logistics Regression model have better result in classification model. In this project, there are lot of limitation. For examples, most of the answer of “Salt” is with No, it is mean that we do not have enough observation for chocolate bar with salt. Also, in this project we did not consider with variable of the location of the company who made the chocolate or the origin of the bean. It is because variables can have many options and some country have very few observations, it is hard for us to analysis. If we want to include this in our analysis, we should focus on some famous country, to find out is the country affecting the rating of chocolate or not. But it is not our objective, so we do not consider about the country this time. On the other hand, our best two classification model only have 63% accuracy in total, so it means that maybe classification is not the best way to analysis this data set. We can try regression model such as PLS, PCR or multiple polynomial regression model next time. But our objective is built up a model for personal use, so I think it is an acceptable for personal use, if for business, it is better to build up another model.

**Appendices**

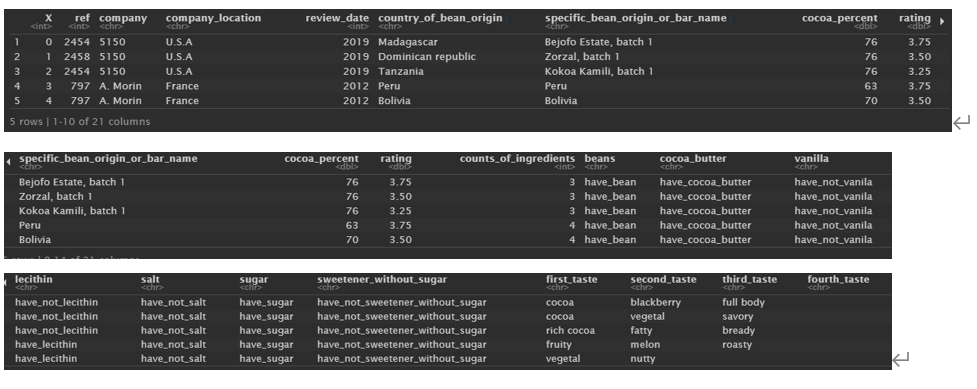


Fig. The list of variables

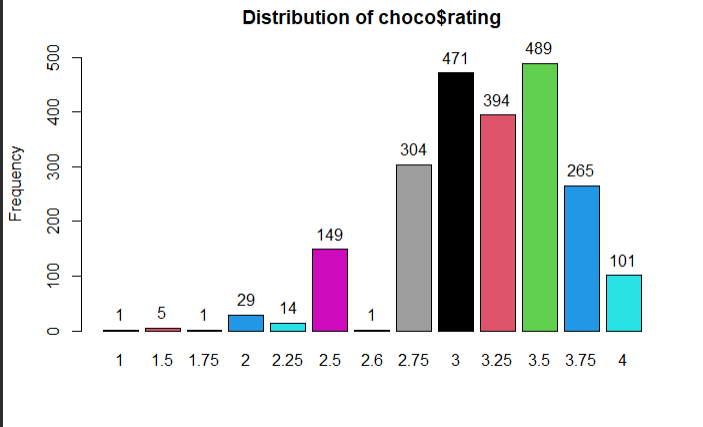


Fig. 2

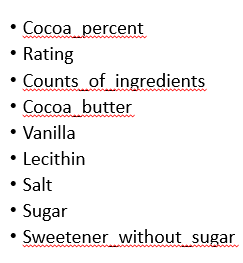


Fig. 3

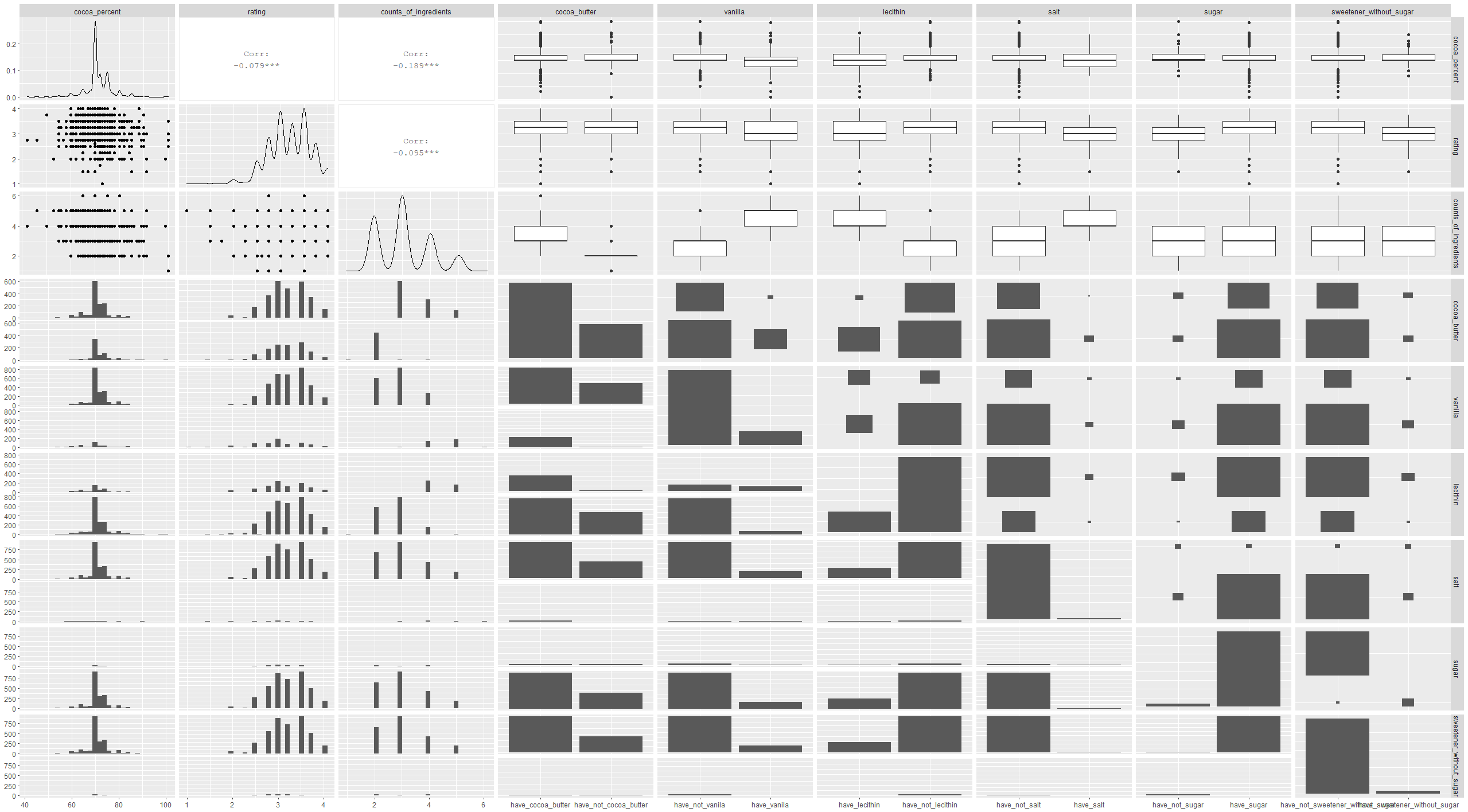


Fig. 4

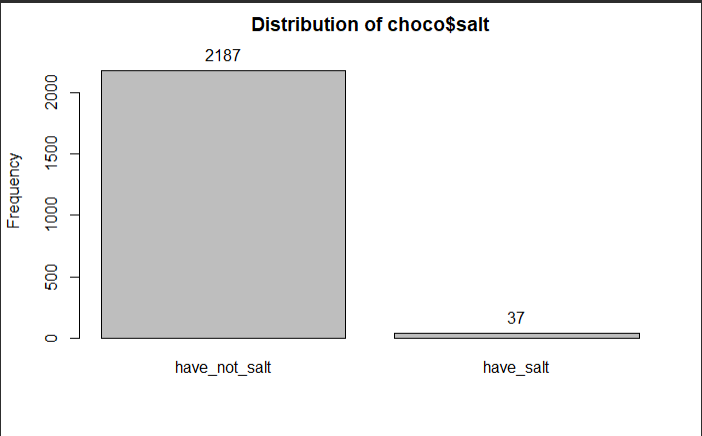


Fig. 5

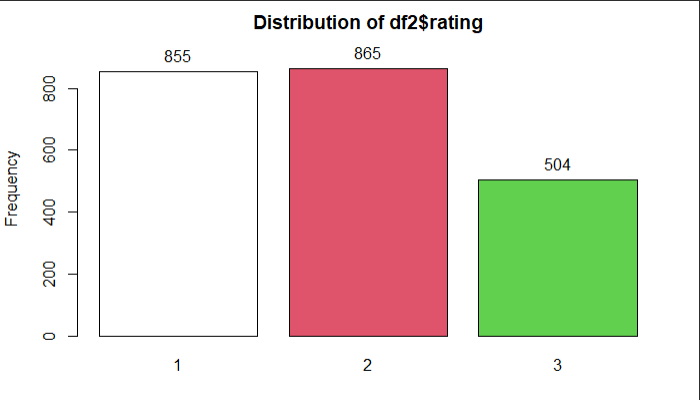


Fig. 6 Rating with 3 labels

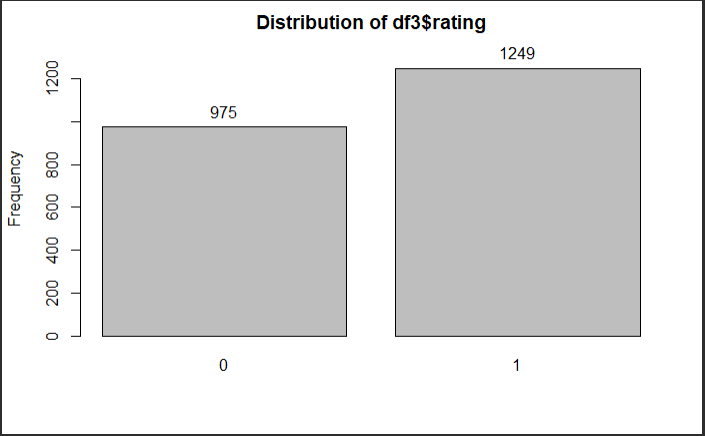


Fig. 7 Rating with 2 labels



Fig. 8

**References**

John, R. F. (2005, August). *Pareto’s Principle-The 80-20 Rule*. ProQuest. https://search-proquest-com.ezproxy.lib.ouhk.edu.hk/docview/230152185/E49922D103BD4875PQ/2?accountid=16720