Chocolate Bar Ratings

CS 429 Data Science

Spring 2018

Ariel Todoki and Sophia Anderson

Data Overview

The chocolate bar dataset was found on the Kaggle website. The data is from FlavorsofCacao.com which was compiled by Brady Brelinski, co-author of the site and founding member of the Manhattan Chocolate Society.

The reviews are focused on **Dark Chocolate**, and each row in the dataset represents one bar from one batch

Rating System

5= Elite (Transcending beyond the ordinary limits)

4= Premium (Superior flavor development, character and style)

3= Satisfactory (3.0) to **praiseworthy**(3.75) (well made with special qualities)

2= Disappointing (Passable but contains at least one significant flaw)

1= Unpleasant (mostly unpalatable)

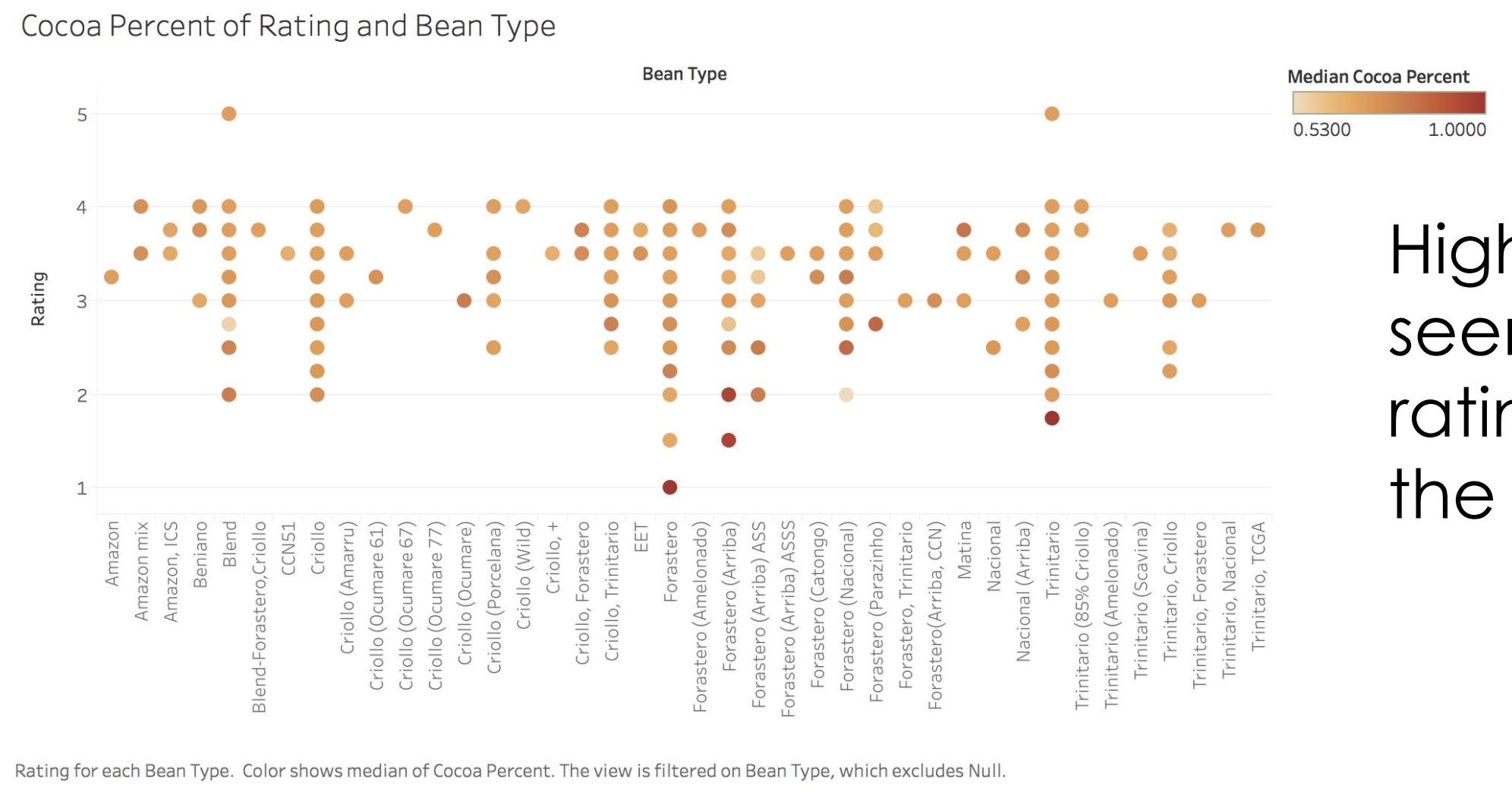
(FlavorsofCacao.com)

Research Questions

- Are the ratings consistent enough for us to predict a bar's rating based on a few characteristics?
- Is a chocolate's rating consistent with the bean type?
- Are certain varieties of bean better than others?
- Is a bar's rating based more on the variety of bean or the company making the bar?
- Is there a relationship between cocoa solids percentage and rating?

We used iPython Notebook to do our analysis and create graphs, as well as using Tableau to create a lot of our graphs for our exploratory data analysis.

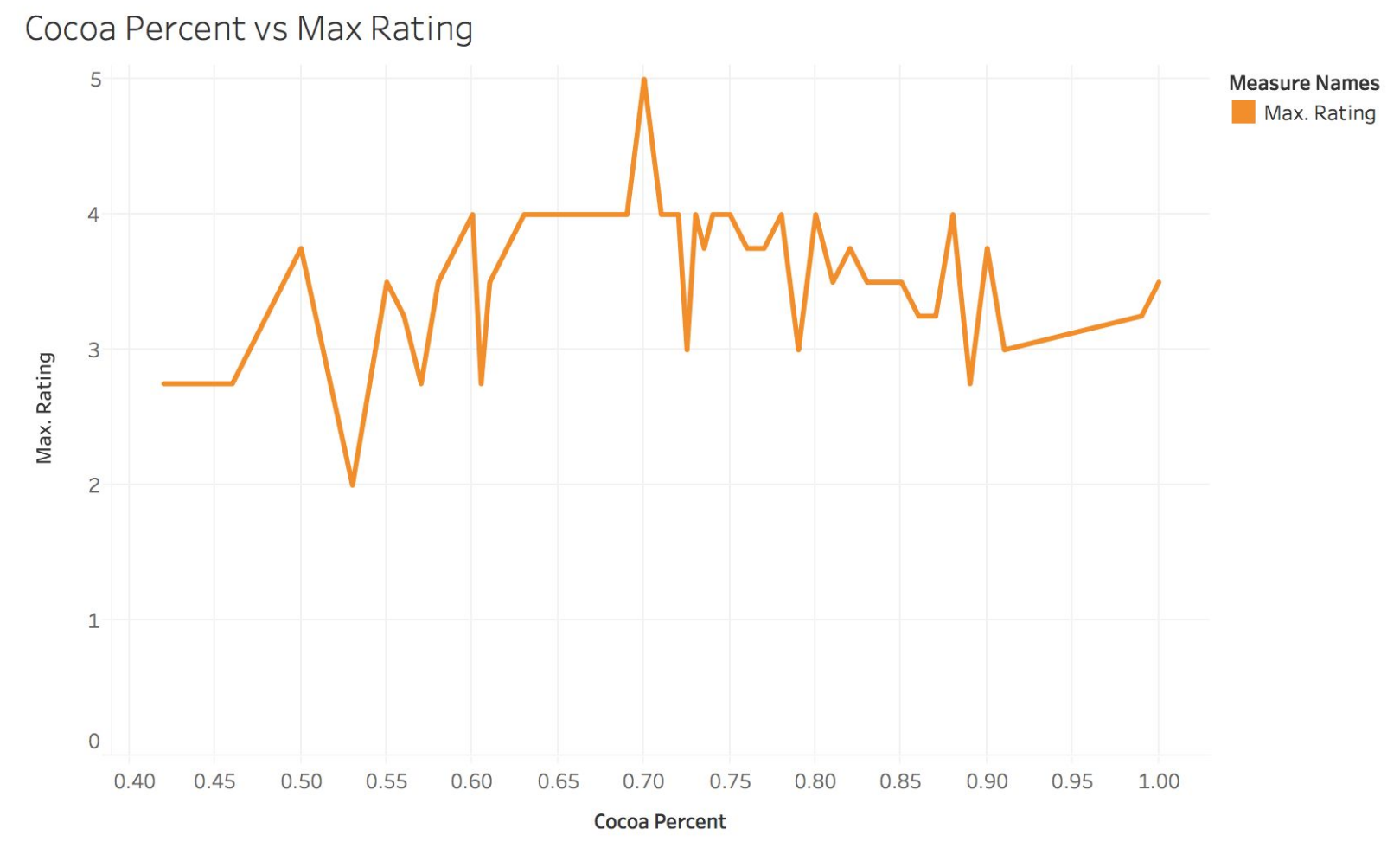
Exploratory Data Analysis



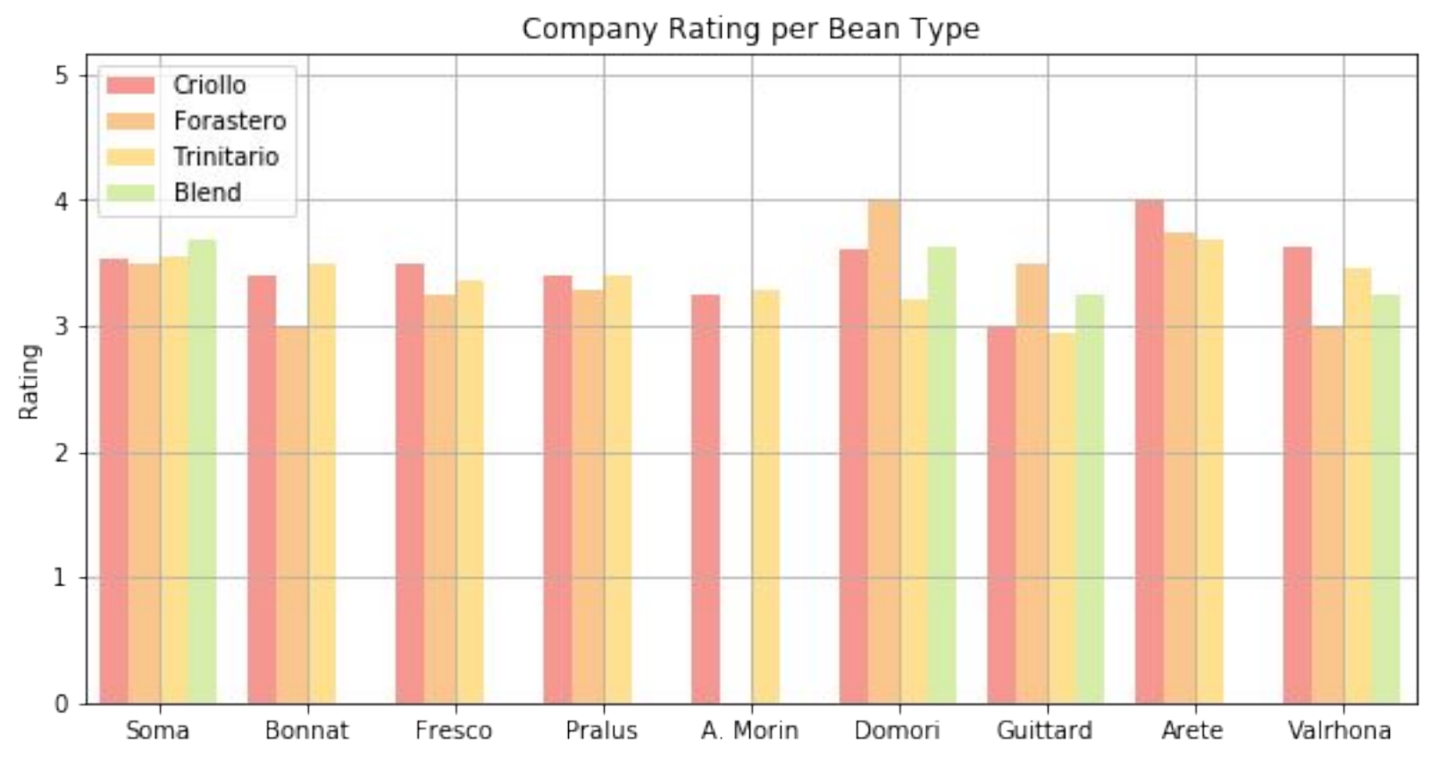
Rating for each Bean Type. Color shows median of Cocoa Percent. The view is filtered on Bean Type, which excludes Null.

Higher cocoa percent seems to have lower ratings as indicated by the dark brown dots.

Taking the best rated chocolate from each cocoa percent category, the best rated chocolate has 70% cocoa.



The trend of Max. Rating for Cocoa Percent. Color shows details about Max. Rating.

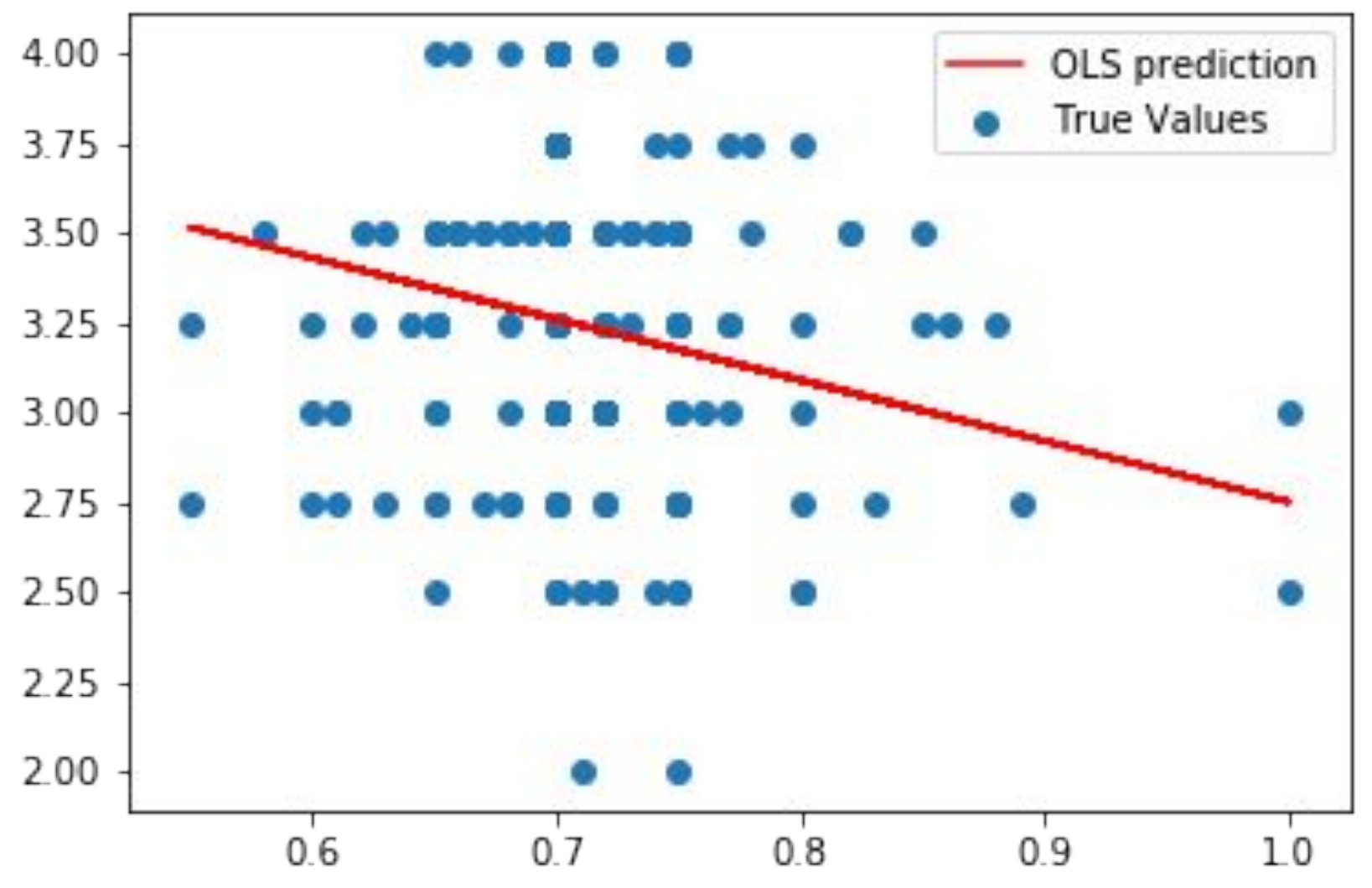


These are the top 10 companies with the most chocolate bars in the dataset. Forastero seems to always rank lowest for all companies except for ones that used Forastero (Arriba) or Forastero (Nacional) which are considered fine flavor cocoa. (Guittard, Domori, and Arete)

Linear Regression Model

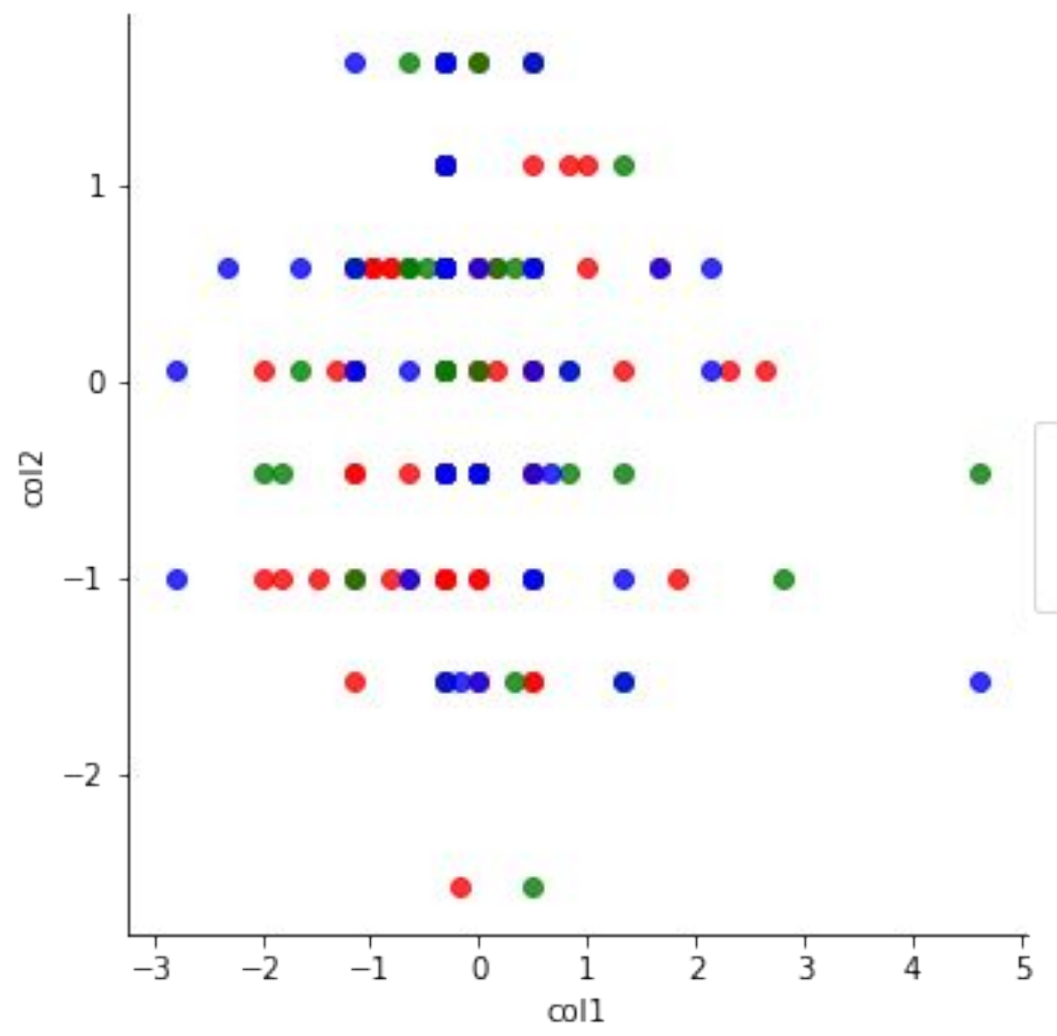
Rating = -1.7002 (Cocoa Percent) + 4.4494

To see if it's possible to predict the rating from the cocoa percent, we fit our data to a linear regression model. We found that while it does show how higher percentage of cocoa seems to get lower ratings, it does not seem to explain the peak in ratings around the 70% cocoa range.

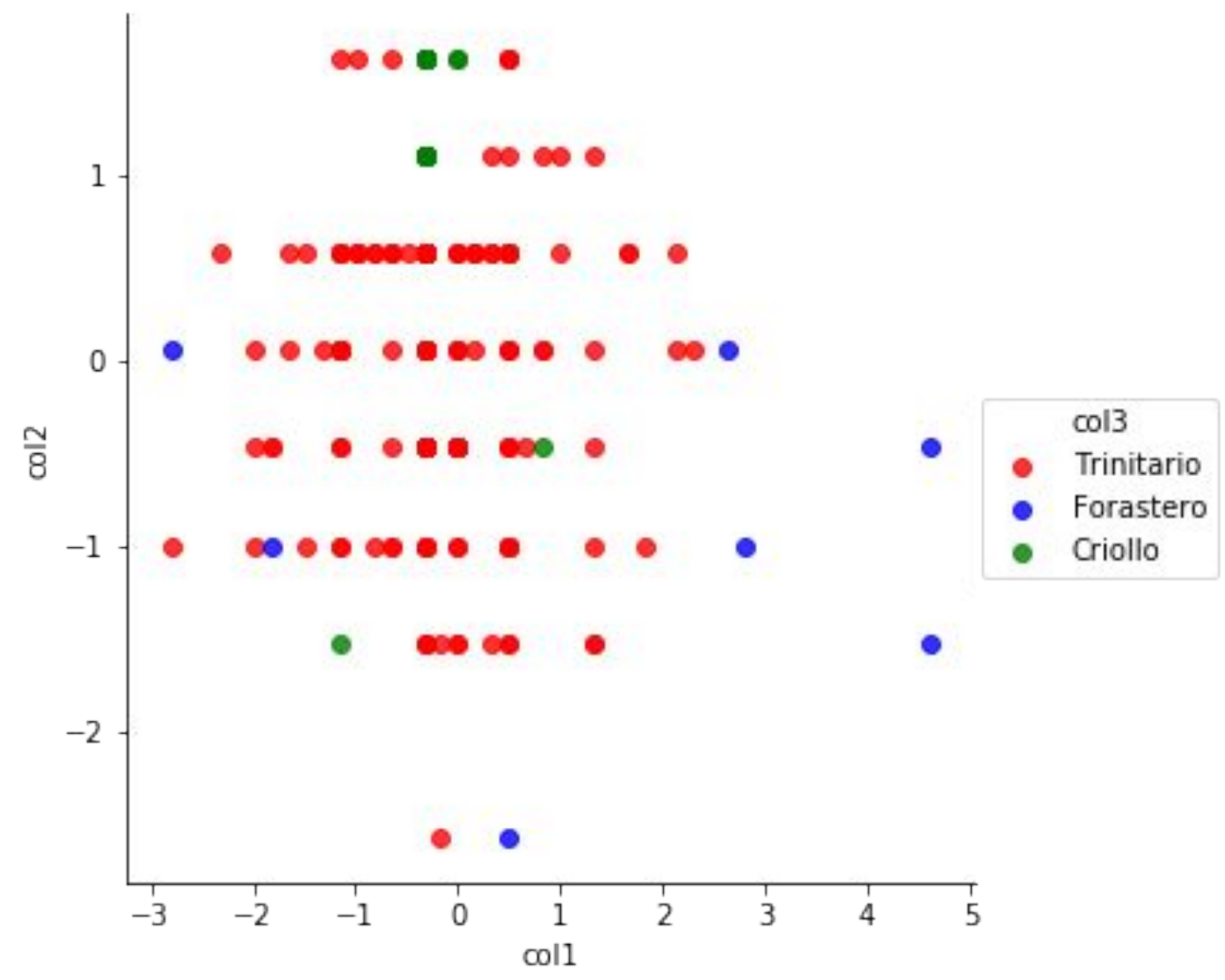


K-Nearest Neighbor Algorithm

Test Set
(Correct Categorization)



Test Set
(Predicted Categorization)



We wanted to see if you would be able to predict the type of cacao bean based on the cocoa percent and rating. After plotting the different k values and their mean error, we decided to make our k value 23. We found there is no clear pattern and that we cannot create a strong model using the data we have.

Bean Type	Precision	Recall
Criollo	0.27	0.14
Forastero	0.29	0.05
Other	0.0	0.0
Trinitario	0.49	0.84

Conclusion

There is no linear relationship between the cocoa percent and the rating of the chocolate. There is also no clear clustering of bean types when looking at their rating and cocoa percent. However, our EDA suggests that the highest rated chocolate are the ones with 70% cocoa, and that chocolate with close to 100% cocoa have lower ratings. It also suggests that chocolate that uses Forastero beans (used in mass chocolate production) ranks lower than the Criollo, Trinitario, and Blend types across companies, unless using the rarer Forastero (Arriba) or Forastero (Nacional) types.

Resources

https://www.kaggle.com/rtatman/chocolate-bar-ratings

http://flavorsofcacao.com/index.html

Tableau