Chocolate!

Kimberly Cable

Belleview University

DSC 550: Data Mining

Brett Werner

August 13, 2022

Chocolate!

Chocolate. Who doesn't like it? It's been around since the 19th century BCE. But what makes chocolate highly desirable?

Chocolate is loved by many worldwide. But what makes good chocolate? Chocolatiers, home cooks, and lovers of chocolate eat, use, and make chocolate. Knowing where to get good chocolate could make or break a recipe or even a good relationship.

The data I will use comes from the Flavors of Cacao website and their Chocolate Bar rating table. The data includes features such as cocoa content, where it was made, bean origin, the number of ingredients, and its characteristics and rating. I will also get the USA Craft Makers table and merge it with the Chocolate Bar Ratings to look more closely at the US Chocolate Companies.

I plan to look to see if I can predict the ratings for chocolate in the United States and see if any features play a part in good chocolate.

I will look at questions such as:

- Which states have the best chocolate?
- Which ingredients make the best chocolate?
- How much cocoa makes for good chocolate?
- What are the key characteristics of good chocolate?

I do not see any major ethical implications for my study, but it does have subjective attributes that some may disagree with.

Some of the challenges I see are cleaning the dataset and separating out some of the columns. Also, the model building may be a challenge as I have never used many categorical methods before.

Data

I will be using data from the Manhattan Chocolate Society, Flavors of Cacao website.[1]

Data Dictionary

This is a list of the data I will be working with and its definition

Chocolate Bar Ratings:

Company (Manufacturer): Company that made the chocolate
Company Location: Location of the Company (country)
Review Date: Year the chocolate was reviewed
Country of Bean Origin: Country the bean originated in
Cocoa Percent: Percentage of cocoa in chocolate
Ingredients: Number and ingredients
- B: Beans
- S: Sugar
- S*: Sweetener other than white cane or beet sugar
- C: Cocoa Butter
- V: Vanilla
- L: Lecithin
- Sa: Salt

Most Memorable Characteristics: A summary review of the most memorable characteristics of that bar.
Rating: Each chocolate is evaluated from a combination of both objective qualities and subjective interpretation. A rating here only represents an experience with one bar from one batch.

Rating Scale
- 4.0 - 5.0 = Outstanding
- 3.5 - 3.9 = Highly Recommended
- 3.0 - 3.49 = Recommended
- 2.0 - 2.9 = Disappointing
- 0 - 1.9 = Unpleasant

USA Craft Makers

Company Name: Company that made the chocolate
State: Company's state of location

Exploratory Data Analysis

In the United States, most chocolate is rated 'Recommended' and 'Highly Recommended' (see Figures 1 and 2). This should indicate that your chances of picking good chocolate are pretty good.

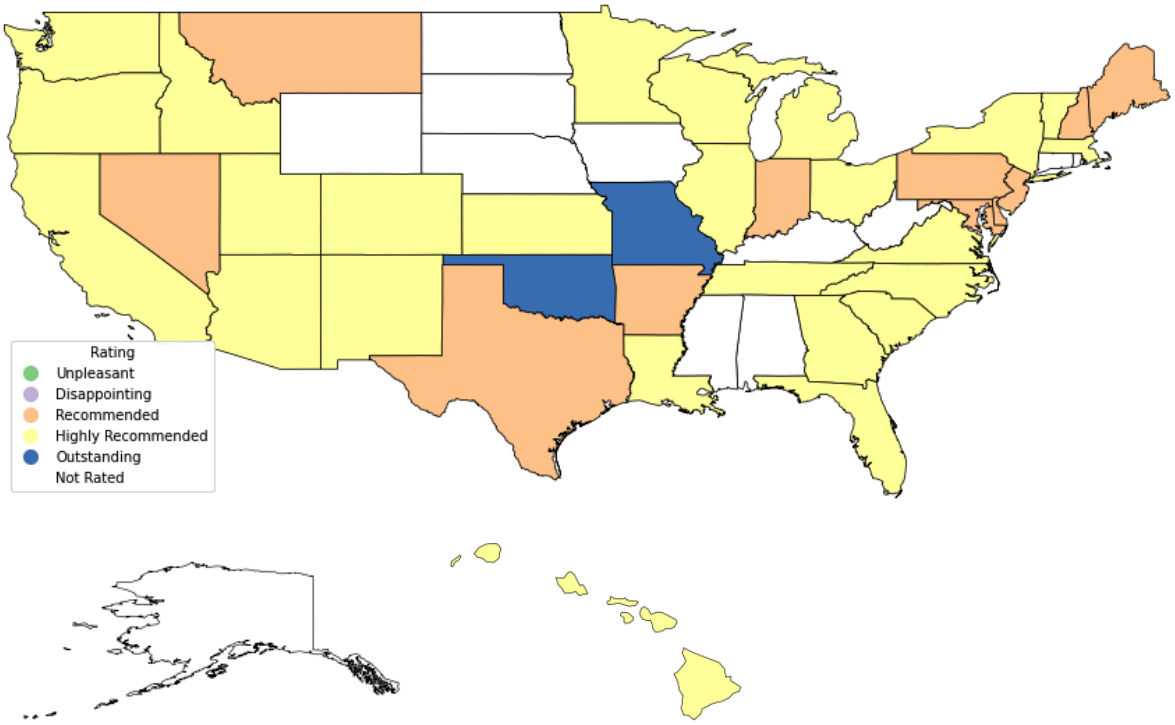

Figure 1: United States chocolate ratings per state.

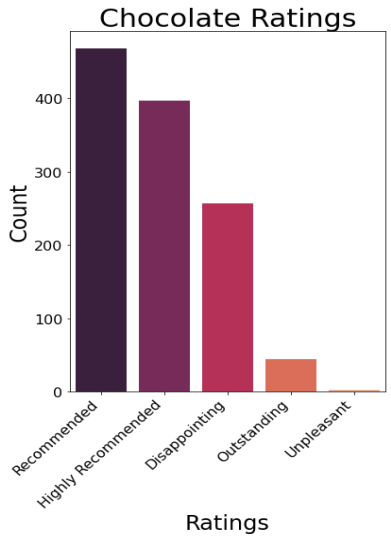

Figure 2: United States chocolate ratings.

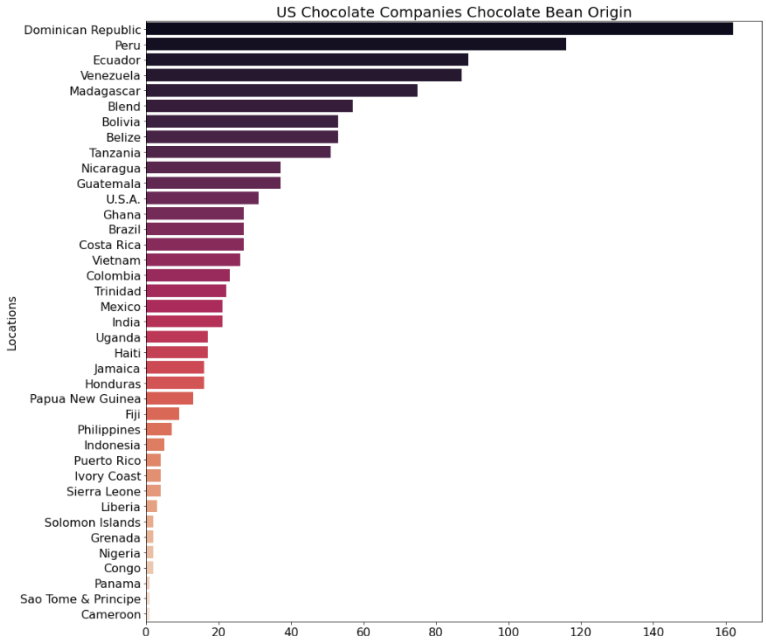

Figure 3: United States chocolate makers cocoa bean origins

The predominant chocolate bean United States chocolate makers use is from the Dominican Republic (see Figure 3).

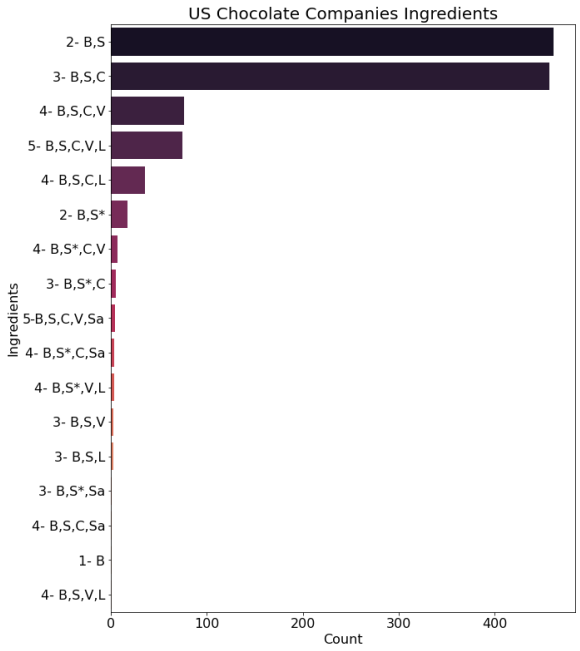

Figure 4: United States chocolate makers ingredients in their chocolates.

In looking at ingredients, most United States chocolates are made from butter and sugar and Butter, Sugar, and Cocoa Butter (see Figure 4)

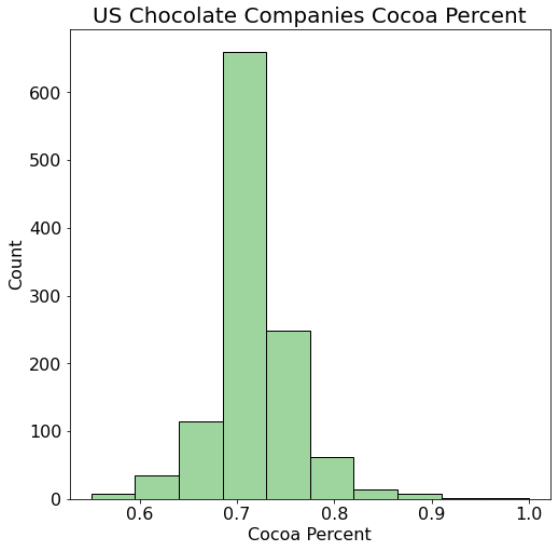

Figure 5: United States chocolate makers amount of cocoa in their chocolates.

The amount of cocoa in United States chocolates (see Figure 5). Most of the chocolates are made with around 70% cocoa.

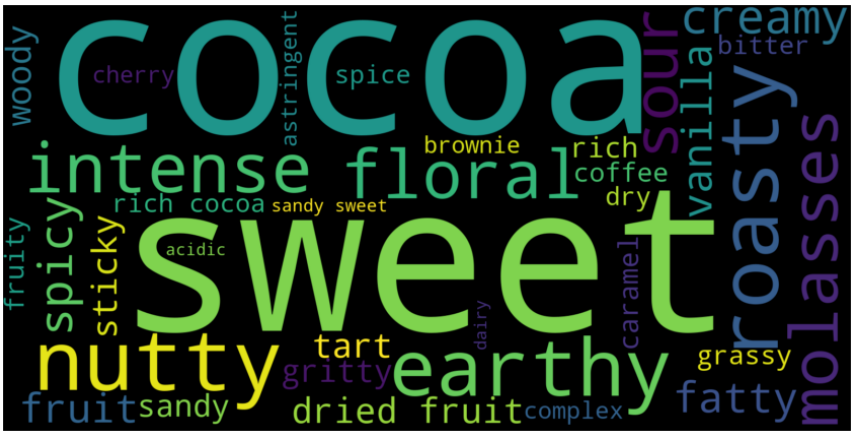

Figure 5: The top 20 most memorable characteristics from the reviews in United States chocolates.

The top 20 most memorable characteristics the reviewers said about United States chocolates (see figure 5).

Data Preparation

The data preparation steps I took were the following:

1. Split out the chocolate ratings into United States chocolates and the rest of the world. We will be looking just at United States chocolates.
2. Rating: I converted the ratings to their proper scale. This helped group the ratings into their proper bin.

- 4.0 - 5.0 = Outstanding
- 3.5 - 3.9 = Highly Recommended
- 3.0 - 3.49 = Recommended
- 2.0 - 2.9 = Disappointing
- 0 - 1.9 = Unpleasant

3. Most Memorable Characteristics: Many of the characteristics were only mentioned a couple of times so I chose only the top 20 characteristics mentioned in the United States chocolates.
4. Ingredients: With some of the ingredients, their combination was only used in a few chocolates, so I used ingredient combinations used in more than 10 chocolates. I also removed chocolates that did not have ingredients.
5. Cocoa Bean Origin: With some of the bean origins, they were only used in a few chocolates, so I used the bean origins that were used in more than 10 chocolates.
6. I dropped columns that were not needed in predicting good chocolate.

Model Building and Evaluation

Looking at the data and wanting to predict good chocolate, I decided to look at three models: Random Forest Classifier, Logistic Regression, and KNN Classifier. All three were chosen due to the fact they were good at predicting a multi-class target. What we are looking for is a rating of Recommended, Highly Recommended, or Outstanding.

The first step was to look at all the features I prepared and determine the best model that predicted chocolate in the United States to be good. Using a grid search, I found that the Random Forest Classifier faired the best with an accuracy score of 54%. It correctly predicted a rating of "Disappointing" 29% of the time, "Recommended" 57%, and "Highly Recommended" 64% as indicated by the f1-score for each. It did not predict "Unpleasant" and "Outstanding" at all with an f1-score of 0%. This was probably due to the fact we did not have enough samples at the lower and upper ratings to successfully predict chocolates at that level (see Table 1).

	precision	recall	f1-score	support
2.0	0.00	0.00	0.00	1
3.0	0.67	0.19	0.29	54
3.5	0.53	0.62	0.57	92
4.0	0.54	0.78	0.64	74
5.0	0.00	0.00	0.00	9
accuracy			0.54	230
macro avg	0.35	0.32	0.30	230
weighted avg	0.54	0.54	0.50	230

The precision scores indicate a good ability to correctly predict each rating for the three middle ratings, but it had trouble with correctly predicting positive "Recommended" ratings as indicated by the low recall score of 19%.

Table 1: Classification Report using all features.

Next, I looked at reducing the number of features that did not help predict the best chocolates using Principal Components Analysis. Looking at the same three models and using a grid search, I found that the Logistic Regression model predicted chocolates the best. With the reduction in features, the accuracy score was 51%. It correctly predicted a rating of Disappointing 45% of the time, Recommended 49%, and Highly Recommended 58%. It did not predict Unpleasant and Outstanding at all with a score of 0%. This indicated that we did not have enough samples at the lower and upper ratings to successfully predict chocolates at that level. (see Table 2).

	precision	recall	f1-score	support
2.0	0.00	0.00	0.00	1
3.0	0.51	0.41	0.45	54
3.5	0.52	0.47	0.49	92
4.0	0.50	0.70	0.58	74
5.0	0.00	0.00	0.00	9
accuracy			0.51	230
macro avg	0.31	0.32	0.31	230
weighted avg	0.49	0.51	0.49	230

The precision scores indicate a good ability to correctly predict each rating for the three middle ratings, but it also had trouble with correctly predicting positive "Recommended" ratings as indicated by the low recall score of 19%.

Table 2: Classification Report using Principle Components Analysis.

Conclusion

Rating	All Features	Using PCA
	F1-Score	
Outstanding	0.00	0.00
Highly Recommended	0.64	0.58
Recommended	0.57	0.49
Disappointing	0.59	0.45
Unpleasant	0.00	0.00
Accuracy	0.54	0.51
Weighted Average	0.50	0.49

Table 3: Summary of Classification Report of both models

In predicting chocolate in the United States, I believe personal taste is the true measure. Trying to predict good chocolate, and shown, can be tricky. Aside from that, looking strictly at accuracies, the Random Forest Classifier fared slightly better than using Logistic Regression, but both were above 50% (see Table 3). For me, the most important indicators were the F1 scores of each of the ratings. Using all features and the Random Forest Classifier, chocolates of the three middle classifiers (Disappointing, Recommended, and Highly Recommended) were correctly predicted 57 to 74% of the time. The weighted average was also 54% correctly predicting a rating. I believe using all features and the Random Forest Classifier would help initially get you good chocolate. Then, you can personally choose your tastes and preferences from the initial choices the model predicted.

The next step I would take to see if I could accurately predict good chocolate is to include all chocolates not just look at United States Chocolates. This may give us more data to work with, especially in the bottom and top classifiers. Also, looking at more reviews not just from the Manhattan Chocolate Society.

References

Flavors of Cacao. (n.d.). Retrieved from Chocolate Bar Ratings: http://flavorsofcacao.com/