Draft

STA 210 - Project

Ginger and Stats - Aimi Wen, Rakshita Ramakrishna, Nathan Nguyen

I. Introduction and data

Broader Context + Research Question

Chocolate is one of the most popular sweets in the world– according to the World Cocoa Foundation, more than 3 million tons of cocoa beans a year are consumed. Dark chocolate, which this dataset focuses on, has been linked to increase heart health, balance the immune system, combat diabetes, improve brain function, boost athletic performance, and reduce stress (1). While dark chocolate can be helpful to human health, arguably, its popularity is due to its taste and its ability to make us "feel good." Studies have found that the ability to make us "feel good" is due to the psychoactive chemicals it contains (2). For serious chocolate lovers, chocolate's particular chemical signature can be needed by chocolate lovers' metabolic systems, thus making the treat particularly delicious to them (3). But other than the chemical compounds in chocolate, how does taste impact chocolate's likeability? What other factors can impact chocolate's likeability? Our dataset contains different dark chocolate bars. One of the columns is chocolate ratings, which are made by members of the Manhattan Chocolate Society. Using the chocolate rating as an indication of the chocolate's likeability, our general research question, therefore, is what can predict chocolate ratings?

Based on our research question, we have the following hypotheses:

1. A lower cocoa percentage is linked to a higher rating.
2. Cocoa percentage and ingredients are the significant predictors.

From our modeling and our analysis, we came up with a model that proved both of our hypotheses as correct. Although our model provided interesting insights, our R-Squared values were relatively low (hovering at around 0.2 during cross-validation). So, we also offered suggestion for other model and data explorations.

References:

1. https://www.hopkinsmedicine.org/health/wellness-and-prevention/the-benefits-of-having-a-healthy-relationship-with-chocolate
2. https://www.bbc.com/news/health-39067088
3. https://www.acs.org/content/acs/en/pressroom/newsreleases/2007/october/news-release-study-finds-that-people-are-programmed-to-love-chocolate.html

Data description

The data is collected by members of the Manhattan Chocolate Society reviewing chocolate bars using the rating system found at http://flavorsofcacao.com/review_guide.html and adding other characteristics about the bar itself. It is being continuously collected and added to the dataset after reviewing chocolate bars - this can be seen as the first review years for chocolate bars began in 2006 and have continued until 2021. It contains 2530 observations, each represents a review of general characteristics for different chocolate bars. A single observation in this dataset represents a single chocolate bars

The general characteristics that will be our main interest are described as follows:

- Company (Manufacturer) lists who made the chocolate bar reviewed; the dataset also lists where this company is located under Company Location.
- The dataset characterizes the Country of Bean Origin, Specific Bean Origin or name of bar, Percentage of Cocoa within the bar for each chocolate bar.
- The data also shows which ingredients are used using letters, where B = Beans, S = Sugar, S* = Sweeteners other than white can or beet sugar, C = Cocoa Butter, V = Vanilla, L = Lecithin, Sa = Salt.
- Finally, the data shows the rating (which ranges from 1-5, incrementing by 0.25) given under their rating system, which is linked above, as well as the date it was reviewed on.

The data dictionary can be found here.

II. Methodology

1. Exploratory Data Analysis (EDA)

Before we began modeling, we first performed some Exploratory Data Analysis to decide how we were going to use the variables in our modeling.

a. Shape of Ratings

From Figure 1, we can see that the distribution of the rating is unimodal, centered around the value of 3 or 3.25. It is also left-skewed, with some possible outliers of value 1 or 1.5. The mean of rating is 3.196, the median is 3.25, and the standard deviation is 0.445

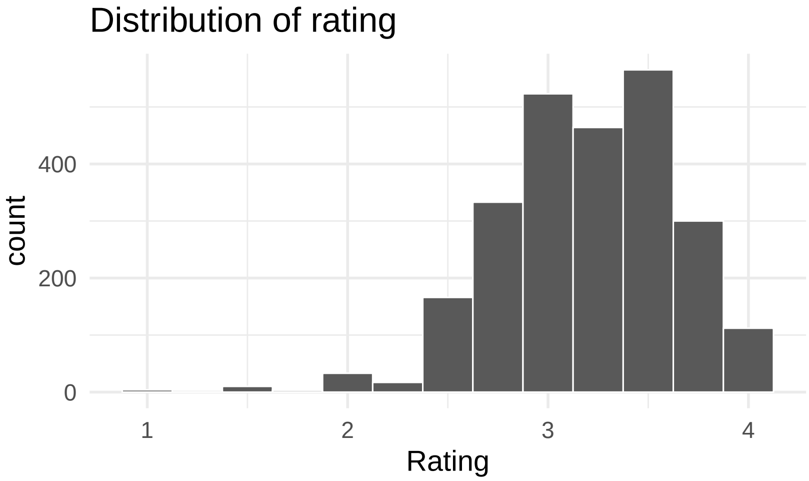

Figure 1: Ratings

b. Ingredients

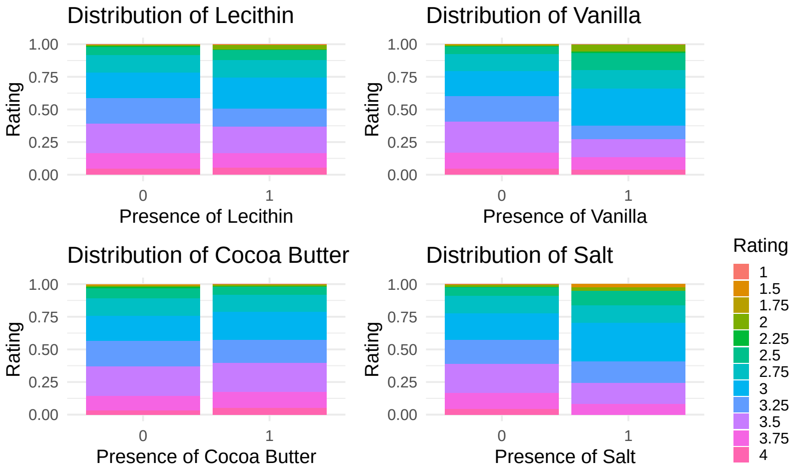

Figure 2: Rating and Ingredients

From Figure 2, we can see that the presence of salt and vanilla seem to affect the rating the most out of all the predicters. The presence of salt and vanilla results in more lower ratings, while the amount of high and low ratings remains roughly the same with/without the presence of cocoa butter and lecithin.

d. Most Memorable Characteristic

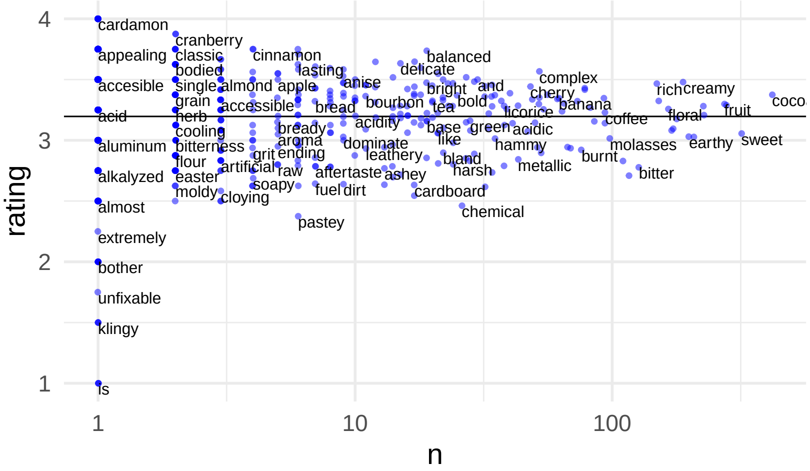

Figure 3: Memorable Characteristics

Figure 3 is taken from https://juliasilge.com/blog/chocolate-ratings/. From this visualization, we can see that the phrases and most memorable charactersists that were often associated with a higher rating were “balanced” and “complex”, as well as fruity chocolate like “fruit”, “Cardamon”, “floral”.

e. Country Bean of Origin

Figure 4 shows that the majority of cacao beans are produced in central America, South America, Asia, and Africa.

f. Company Location

Figure 5 shows that the majority of countries that chocolate companies are located in are concentrated in North America and Europe, and that the US is host to the largest amount of chocolate companies.

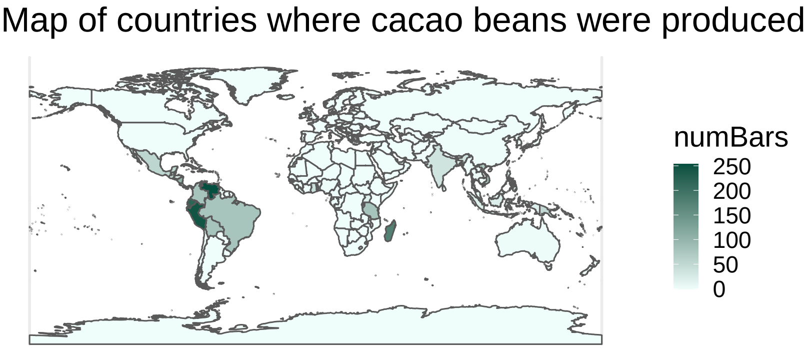

Figure 4: Country of Bean Origin

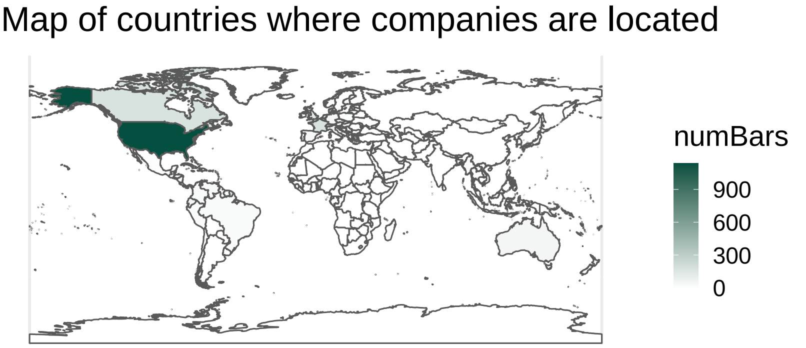

Figure 5: Chocolate Company Location

2. Data Cleaning

To get our data ready for modeling, we first performed some data cleaning.

1. One of the variables that we are using in our modeling is a list of most memorable characteristics by the rater. To organize this variable in a way that can be used for our model, we assumed that the first characteristic listed was the dominant characteristic and made the biggest lasting impression. So, we only kept the first characteristic. From there, because there are a variety of characteristics, we decided to group them into some general groups: fatty_smooth, roast, strong_sweet, rough_texture, nutty, greasy, spiced, floral, fruity, complex, and other. For example, we put "cream" and "dairy" in the category of "fatty_smooth". Another example is that characteristics that contained the word "fruit" or "berry" were grouped together into "fruity".
2. Next, we also decided to simplify the locations of cocoa bean production. From our EDA, we learned that cocoa bean production locations are mostly based in South America, Asia, and Africa. So, we categorized the countries of cocoa bean production by the most popular continent categories: South America, Africa, Asia, and Other.
3. Similarly, from our EDA, we learned that the company locations are mostly based in North America and Europe. So, we categorized the countries of cocoa bean production by the most popular continent categories: North America, Europe, and Other.
4. We created a new variable that was the number of ingredients. Since there are NA value in the ingredients, we mark these values with 0, though it is kind of unreasonable to assume that a chocolate doesn't have any ingredients composed to it. The median and mean are around 3.
5. In addition to the number of ingredients, we created 2 new variables: vanilla and salt. These variables indicated whether their specified ingredient was listed in the ingredients list. From our EDA, we learned that the presence of salt and vanilla seems to affect the rating the most out of all the ingredients.

3. Modeling

Our main goal of this analysis is to understand how the characteristics of a chocolate can explain its rating. Although the rating only goes from 1 to 5 in 0.25 increments, we treated rating as a quantitative continuous variable. Therefore, we used a linear regression model to fit and predict the rating from the features of a chocolate. We decided compare 2 models: a full model that had all the explanatory variables that we were interested in (location of cocoa bean production, location of chocolate company, number of ingredients, presence of vanilla, presence of salt, top memorable characteristics, and cocoa percentage) and a model with just the "taste" predictors (number of ingredients, presence of vanilla, presence of salt, top memorable characteristics, and cocoa percentage). We defined Model 1 as our model with just the "taste" predictors and Model 2 as our full model.

To evaluate which model performed better, we decided to perform a cross-validation and compare R-squared and RMSE values instead.

III. Results

Ratings vs cocoa percent, ingredients, most memorable characteristics

Table 1: Model 1

Fold	RMSE	R-squared
Fold1	0.416	0.064
Fold2	0.419	0.097
Fold3	0.376	0.145
Fold4	0.438	0.132
Fold5	0.421	0.125

All predictors

Table 2: Model 2

Fold	RMSE	R-squared
Fold1	0.417	0.062
Fold2	0.421	0.090
Fold3	0.381	0.124
Fold4	0.439	0.126
Fold5	0.423	0.117

As both models have similar RMSE and R-squared values for each fold in cross-validation (as seen in Table 1 and 2 and in Figure 6), we will choose the first model as it has fewer predictor variables and aligns with the goals of parsimony, as it is a less complicated model yet doesn't sacrifice predictive capability.

Because our R-squared values are so low, we wanted to explore other options. From https://juliasilge.com/blog/chocolate-ratings/, Dr. Silge made a model predicting a model to predict ratings from most-memorable characteristics. Using her model as a footprint, we decided to add to the most memorable characteristic and make a new model. Her model found which words were more associated with higher or lower ratings. We decided to take her words and create variables out of them. For example, we have a categorical variable detecting whether or not "cocoa" appears in most-memorable characteristics. We ended looking for the following words: cocoa, off, chemical, fruit, creamy, complex and bitter.

So, we created a 3rd model with the variables: number of ingredients, presence of vanilla, presence of salt, top memorable characteristic, cocoa percentage, presence of "cocoa" in most memorable characteristics, presence of "off" in most memorable characteristics, and so on with the other words/characteristics mentioned above.

Table 3: Model 3

Fold	RMSE	R-squared
Fold1	0.405	0.130
Fold2	0.390	0.211
Fold3	0.369	0.184
Fold4	0.406	0.265
Fold5	0.406	0.188

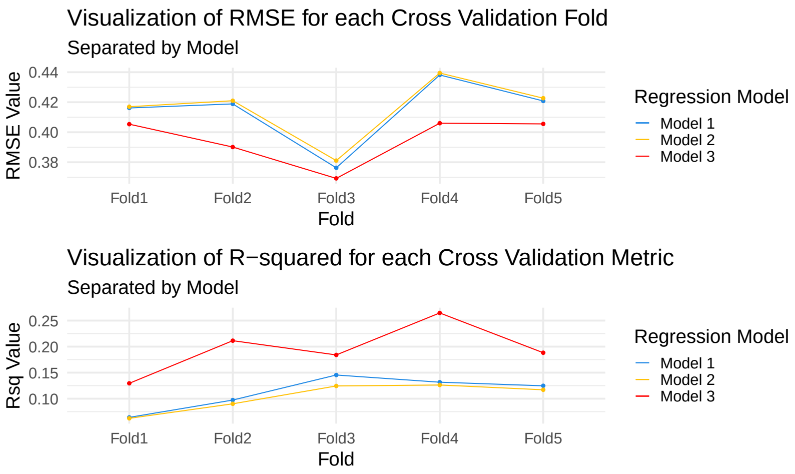

Figure 6: Comparing Three Models

We can see that the third model gives the lowest RMSE and highest R-squared value over all 5 folds. So we decided to select the third model for prediction and further analysis of our hypotheses.

IV. Discussion

Here is the result of the third model:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.210	0.023	140.172	0.000	3.165	3.255
cocoa_percent	-0.009	0.003	-2.654	0.008	-0.015	-0.002
num_ingres	0.043	0.021	2.091	0.037	0.003	0.084
isCocoa	0.218	0.044	4.892	0.000	0.130	0.305
isOff	-0.134	0.070	-1.920	0.055	-0.271	0.003
isChemical	-0.619	0.151	-4.084	0.000	-0.916	-0.321
isFruit	0.185	0.052	3.568	0.000	0.083	0.287
isCreamy	0.223	0.067	3.344	0.001	0.092	0.354
isComplex	0.194	0.114	1.697	0.090	-0.030	0.418
isBitter	-0.339	0.089	-3.792	0.000	-0.515	-0.163
vanilla_X1	-0.332	0.060	-5.504	0.000	-0.450	-0.213
salt_X1	-0.207	0.125	-1.656	0.098	-0.452	0.038

$$\begin{aligned}\widehat{rating} =&3.210 - .009 \times CocoaPercent + .043 \times NumIngredients\\ &+ 0.218 \times isCocoa - 0.134 \times isOff - .619 \times isChemical\\ &+ .185 \times isFruit + .233 \times isCreamy + .194 \times isComplex\\ &- .339 \times isBitter - .332 \times Vanilla_{X1} - .207 \times Salt_{X1}\end{aligned}$$

For a chocolate bar that has 72 cocoa percent, 3 number of ingredients, does not have the words “cocoa,” “off”, “chemical”, “fruit”, “creamy”, “complex” and “bitter” listed in its most memorable characteristics, and does not contain vanilla or salt in its ingredients, the predicted rating is expected to be 3.210, on average.

Before we started our analysis, we had 2 hypotheses: a lower cocoa percentage is linked to a higher rating and cocoa percentage and number of ingredients are significant predictors. For our first hypothesis, we found that this is true. Because if we interpret the coefficient of the cocoa percentage: for each percentage increase in cocoa percentage, we expect the rating to decrease on average by 0.009, all else held constant.

For our second hypothesis, we found that the p-value for cocoa-percent is 0.008 and the p-value for number of ingredients is 0.008. Because the p-values are less than 0.05, we do have enough evidence to reject the null hypotheses.

$$H_0 : \beta_{cocoa\_percent} = 0$$
$$H_1 : \beta_{cocoa\_percent} \neq 0$$

$$H_0 : \beta_{number\_ingredients} = 0$$
$$H_1 : \beta_{number\_ingredients} \neq 0$$

Beyond our hypotheses, our model has revealed other findings. For example, out of the characteristics in memorable characteristics, the word “creamy” has the magnitude in change for increasing ratings: all else held constant, if “creamy” is included in most memorable characteristic, we expect the predicted rating to increase by 0.223, on average. On the opposite end, the word “chemical” has the largest magnitude in change for decreasing ratings. All else held constant, if “chemical” is included in the most memorable characteristic, we expect the predicted rating to decrease by 0.619, on average. Interestingly, though increasing the number of ingredients is expected to increase the predicted rating, having vanilla or having salt as an ingredient is expected to decrease the predicted rating. This raises the question of what can be done in future work. Are there other ingredients that would increase ratings? Would ingredients’ effects on the chocolate offset the expected increase in predicted ratings from the number of ingredients? Additionally, although our model offered insight into what could increase or decrease chocolate ratings, our R-square values are relatively low (see Table 4), leaving us with the question of why. What other variables should be considered? What can be done with our model to improve it? Or, is taste in chocolate truly subjective? Can there be a model that predicts ratings with more accuracy?

Although we did improve our model in the 3rd model and answered our hypotheses, there are additional items for the future. One of the most obvious ones is trying ordinal logistic regression. For our project, we chose to use linear regression in large part due to the fact that we have not covered ordinal logistic regression. Because our response variable (rating) is not truly continuous, this has influenced the shape of our residuals (as seen in Figure 7), making it hard to verify if conditions have been met for linear regression. While ordinal logistic regression would not guarantee a better model, our response variable could perhaps fit the conditions of the ordinal logistic regression better. In addition to trying a different type of model, getting more expansive data (taster’s chocolate preferences, quality of chocolate, etc.) could also potentially improve the model.

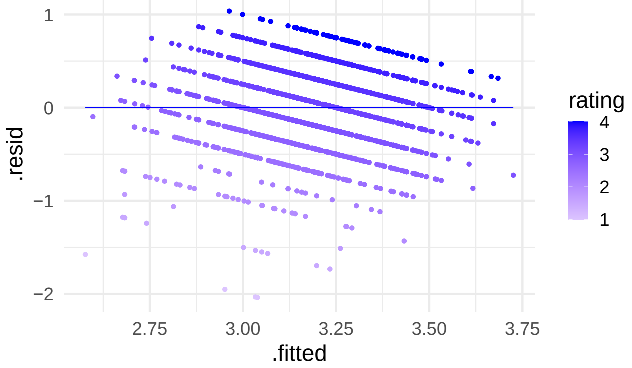

Figure 7: Residuals and Ratings