Draft

STA 210 - Project

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

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.

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.

Methodology

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.

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.

mean	median	sd
3.196	3.25	0.445

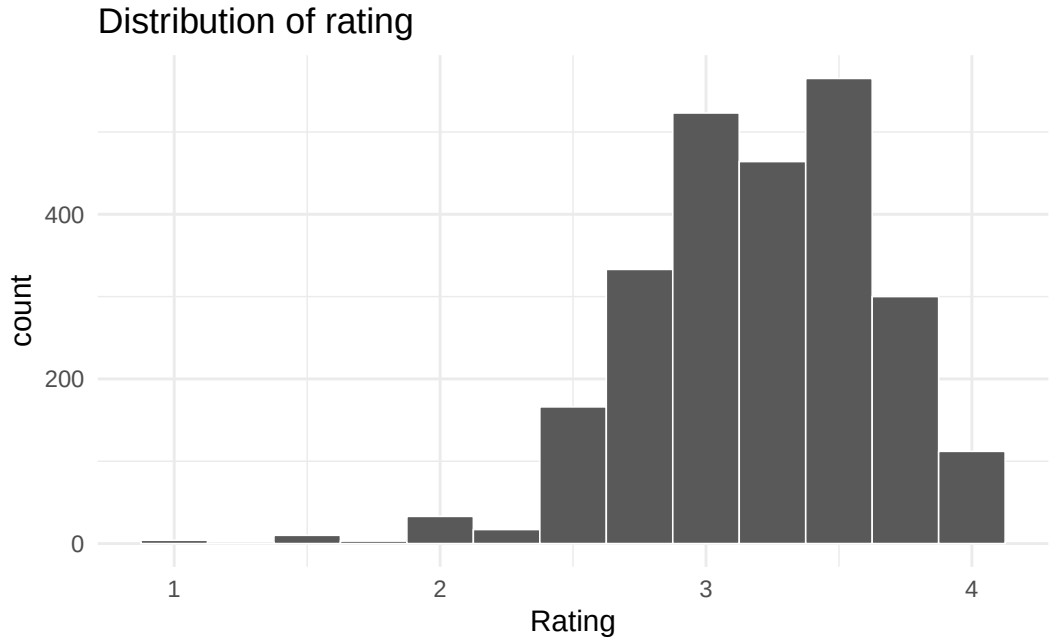

Figure 1: Ratings

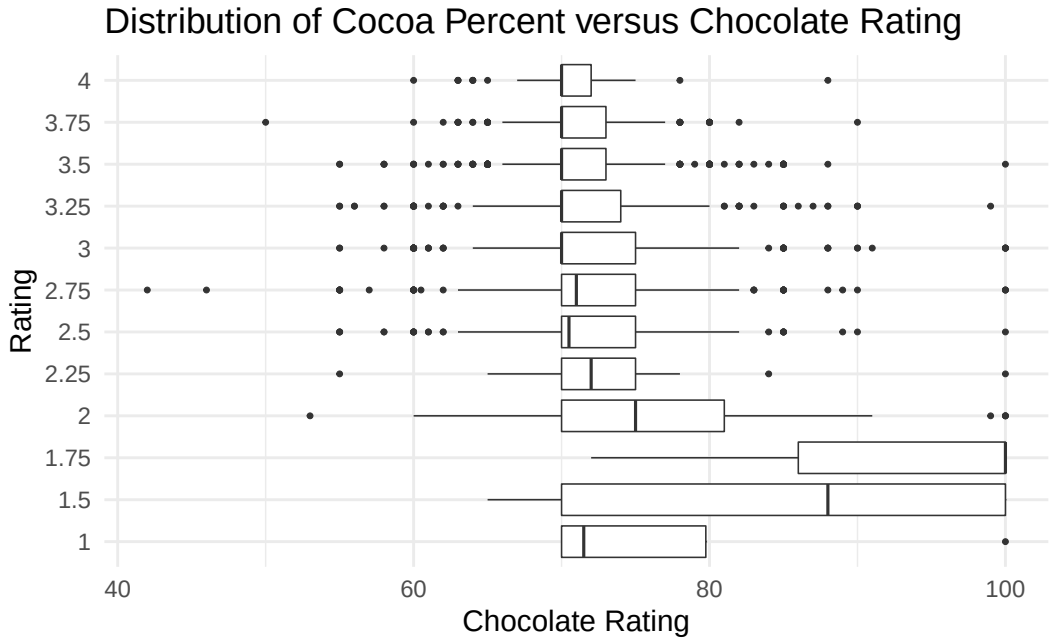

Figure 2: Cocoa Percent and Ratings

Cocoa Percent

From Figure 2, we can see a general rough trend that as the median cocoa percent is lower, the rating of the chocolate bar is higher. Furthermore, there appear to be a lot of outliers in the middle ratings (2.25 - 3.75), which might be due to the fact that that is the rating for the bulk of the chocolates tested.

Ingredients

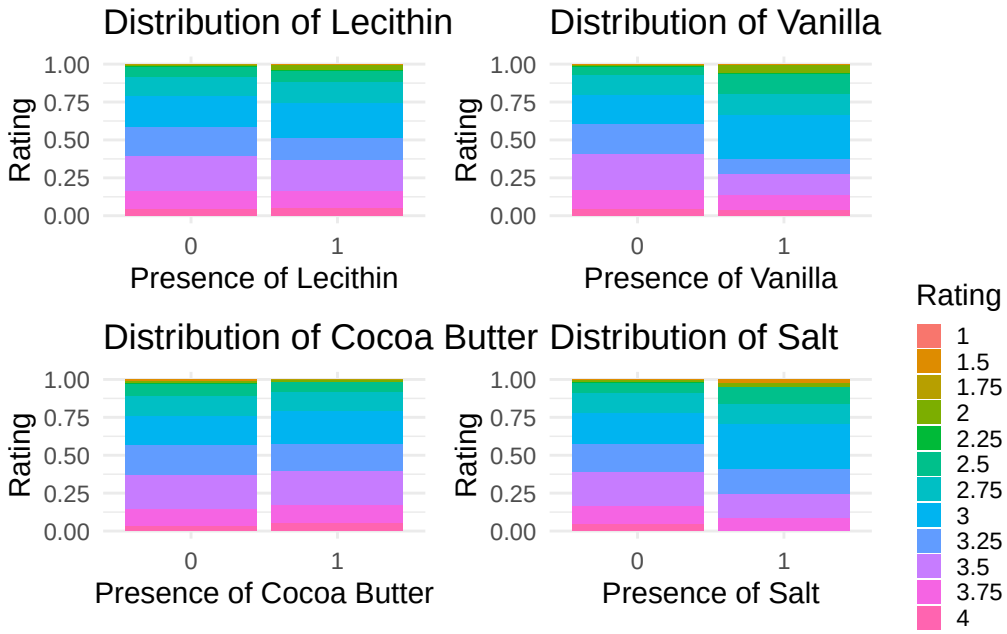

Figure 3: Rating and Ingredients

From Figure 3, 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.

In Figure 4, we regard the NA value of ingredients as 0, which means we should understand this as nonrecorded value instead of no ingredients are presented in the chocolate. This visualization showcases a right skewed distribution for the number of ingredients. The median is somewhere around 3 ingredients, and there appears to be an outlier centered around 0. This could be as many chocolate bars use at least one of the common ingredients, and it is quite rare for a chocolate bar not to have any of those ingredients.

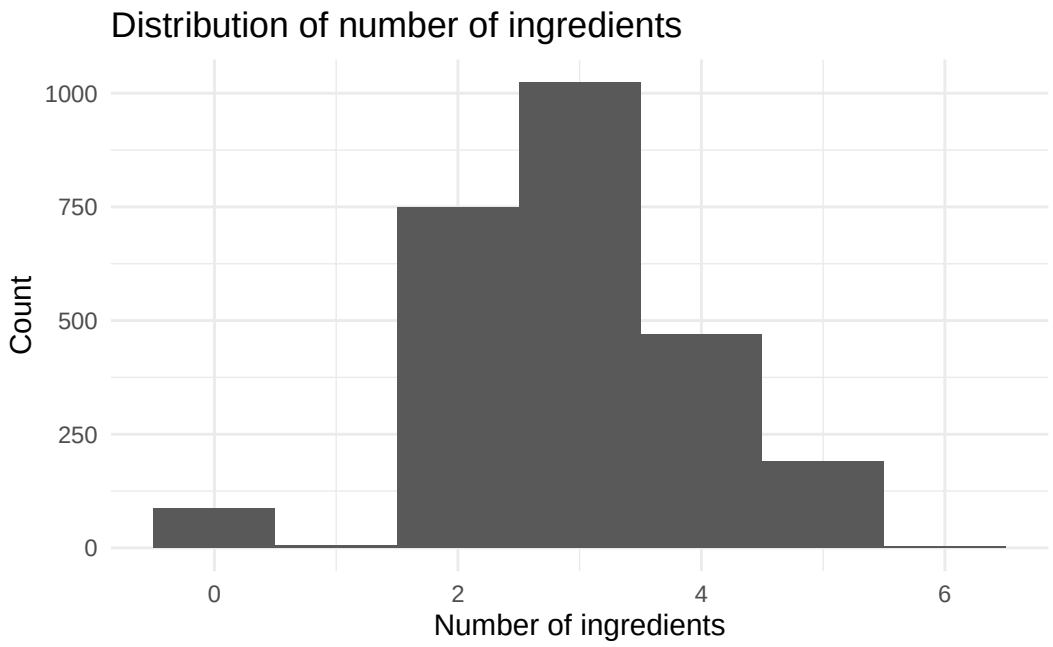

Figure 4: Number of Ingredients and Ratings

Most Memorable Characteristic

Figure 5 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".

Country Bean of Origin

Figure 6 shows that the majority of cacao beans are produced in central America, South America, Asia, and Africa.

Company Location

Figure 7 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.

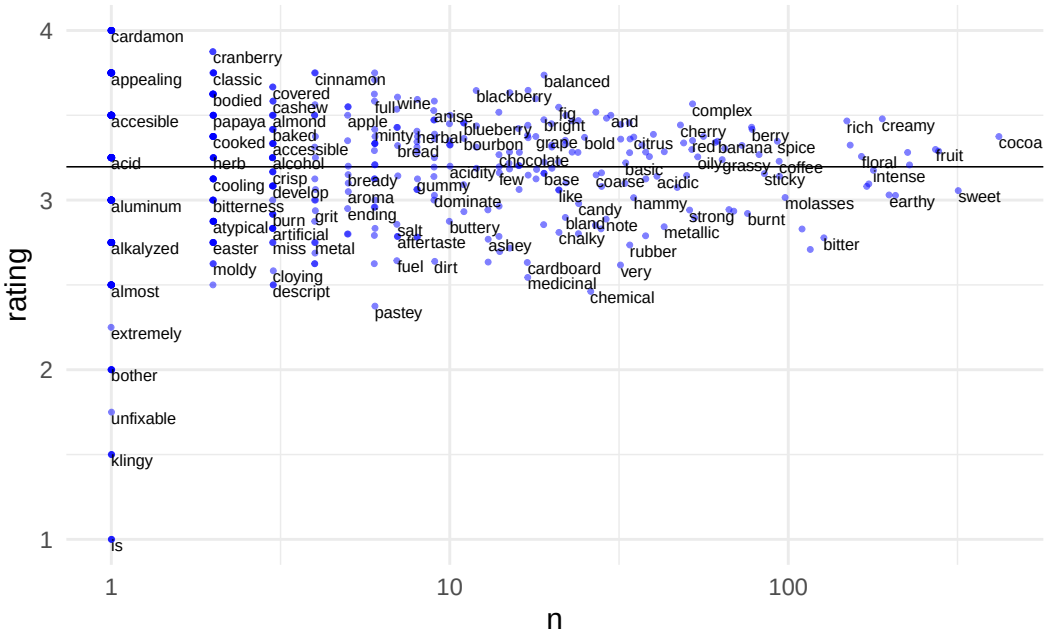

Figure 5: Memorable Characteristics

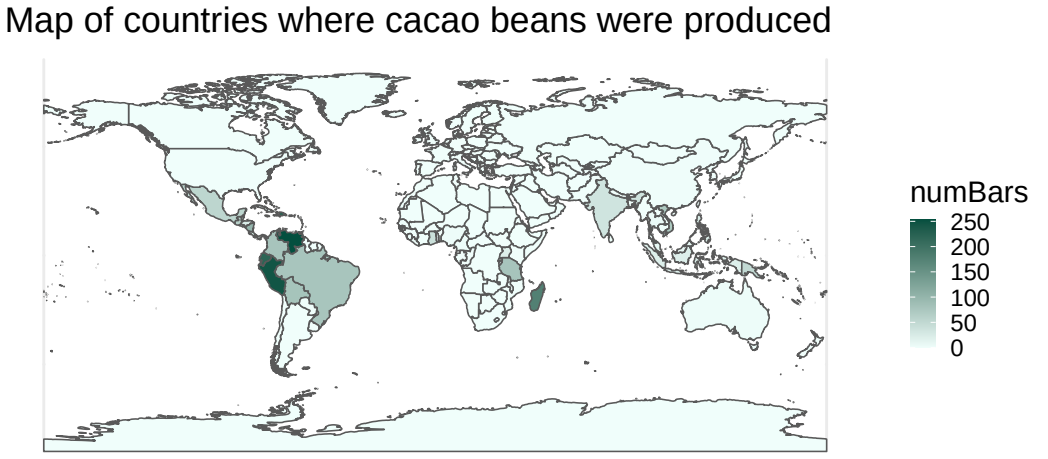

Figure 6: Country of Bean Origin

Figure 7: Chocolate Company Location

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.

Table 1: Model 1

Fold	RMSE	R-squared
Fold1	0.425	0.117
Fold2	0.386	0.111
Fold3	0.415	0.139
Fold4	0.418	0.119
Fold5	0.385	0.101

4. We created a new variable that was the number of ingredients.
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.

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.

Results

Ratings vs cocoa percent, ingredients, most memorable characteristics

All predictors

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 **?@fig-model**), we will choose the first model as it has fewer

Table 2: Model 2

Fold	RMSE	R-squared
Fold1	0.425	0.114
Fold2	0.388	0.104
Fold3	0.415	0.139
Fold4	0.419	0.113
Fold5	0.386	0.099

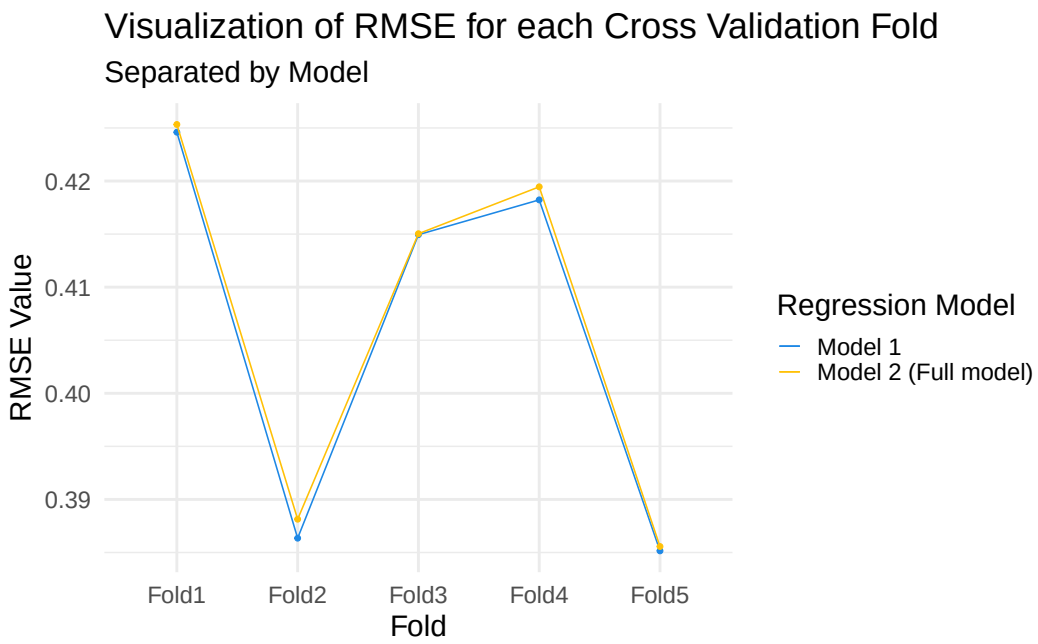

Figure 8: Comparing Model 1 and Model 2

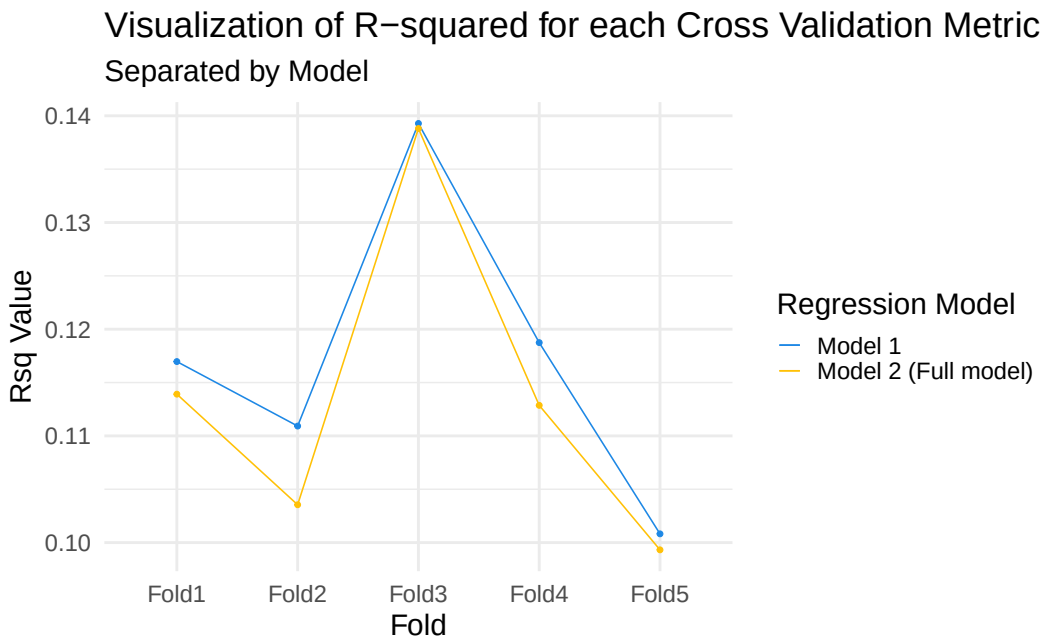

Figure 9: Comparing Model 1 and Model 2

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.

From Table 4 and **?@fig-model-comp2**, we can see that the model has improved.

Table 3: Model 1 Fit

term	estimate	std.error	statistic	p.value
(Intercept)	4.286	0.139	30.740	0.000
cocoa_percent	-0.012	0.002	-7.803	0.000
vanilla1	-0.317	0.031	-10.359	0.000
salt1	-0.277	0.070	-3.930	0.000
num_ingres	0.053	0.010	5.133	0.000
top_memorablefatty_smooth	-0.174	0.080	-2.179	0.029
top_memorablefloral	-0.388	0.086	-4.501	0.000
top_memorablefruity	-0.134	0.082	-1.644	0.100
top_memorablegreasy	-0.357	0.085	-4.195	0.000
top_memorablenutty	-0.285	0.085	-3.369	0.001
top_memorableother	-0.415	0.078	-5.342	0.000
top_memorableroast	-0.421	0.081	-5.217	0.000
top_memorablerough_texture	-0.521	0.080	-6.549	0.000
top_memorablespiced	-0.297	0.084	-3.525	0.000
top_memorablestrong_sweet	-0.328	0.079	-4.145	0.000

Table 4: Model 3

Fold	RMSE	R-squared
Fold1	0.394	0.198
Fold2	0.386	0.154
Fold3	0.412	0.220
Fold4	0.374	0.177
Fold5	0.393	0.148

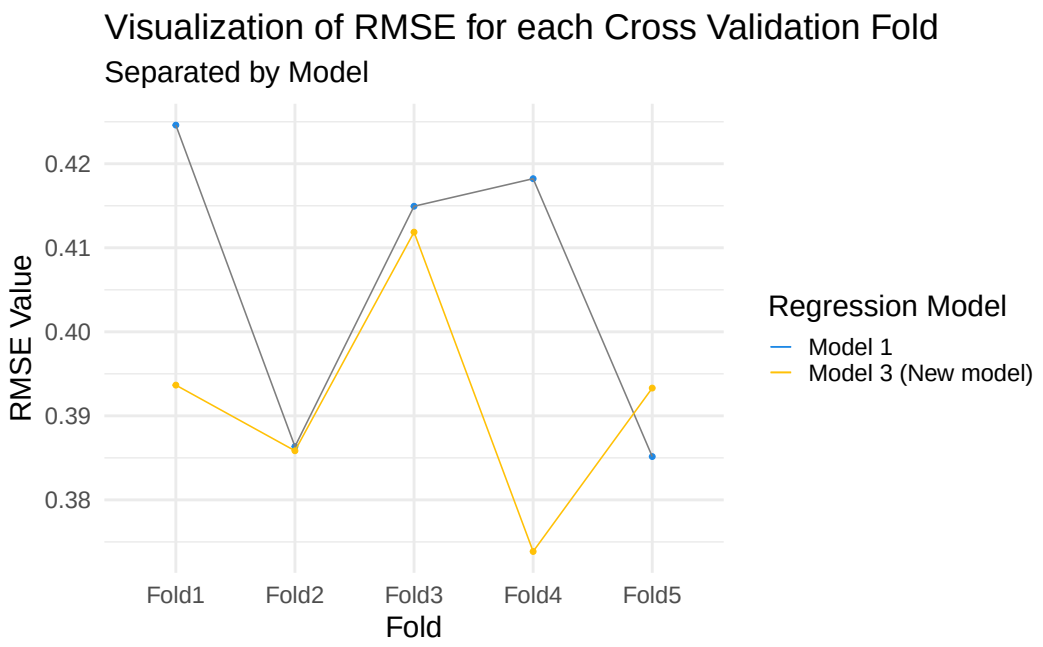

Figure 10: Comparing Model 1 and 3

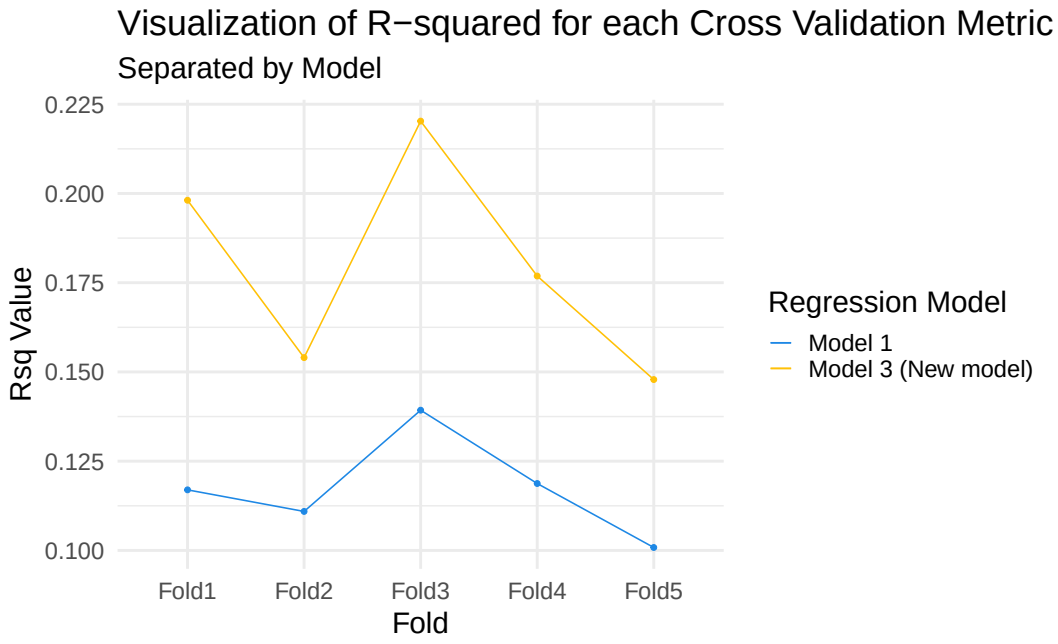

Figure 11: Comparing Model 1 and 3

Discussion

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.216	0.024	134.799	0.000	3.169	3.263
cocoa_percent	-0.007	0.004	-1.697	0.090	-0.014	0.001
num_ingres	0.027	0.021	1.285	0.199	-0.014	0.069
isCocoa	0.149	0.043	3.453	0.001	0.064	0.233
isOff	-0.232	0.057	-4.107	0.000	-0.343	-0.121
isChemical	-0.792	0.150	-5.292	0.000	-1.086	-0.498
isFruit	0.121	0.052	2.313	0.021	0.018	0.223
isCreamy	0.361	0.067	5.393	0.000	0.229	0.492
isComplex	0.387	0.124	3.118	0.002	0.143	0.630
isBitter	-0.548	0.090	-6.120	0.000	-0.724	-0.372
vanilla_X1	-0.305	0.063	-4.817	0.000	-0.430	-0.181
salt_X1	-0.122	0.127	-0.959	0.338	-0.371	0.128

$$\begin{aligned}\hat{rating} =& 3.277 - .006 \times CocoaPercent + .027 \times NumIngredients \\ &+ 0.128 \times isCocoa + 0.212 \times isOff - .811 \times isChemical \\ &+ .060 \times isFruit + .339 \times isCreamy + .359 \times isComplex \\ &- .560 \times isBitter - .291 \times Vanilla_x1 - .165 \times Salt_X1 \\ &- .019 \times topMemFattySmooth - .023 \times topMemFloral \\ &+ .176 \times topMemFruity + .148 \times topMemGreasy + .051 \times topMemNutty \\ &- .122 \times topMemOther - .137 \times topMemRoast - .201 \times topMemRough \\ &- .044 \times topMemSpiced - .012 \times topMemStrongSweet\end{aligned}$$

Interpretation of a few coefficients:

Question: How do you interpret memorable characteristics in an intercept?

Cocoa Percent: For every additional percentage in cocoa percent, the predicted rating is expected to decrease by .006 points, on average, with all else held constant.

TopMemorableFloral: If a chocolate bar is noted as having the memorable characteristic of Floral Notes, the predicted rating is expected to decrease by -.023 rating points compared to chocolate bars that don't have this memorable characteristic, on average, with all else held constant.

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 12

), 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.

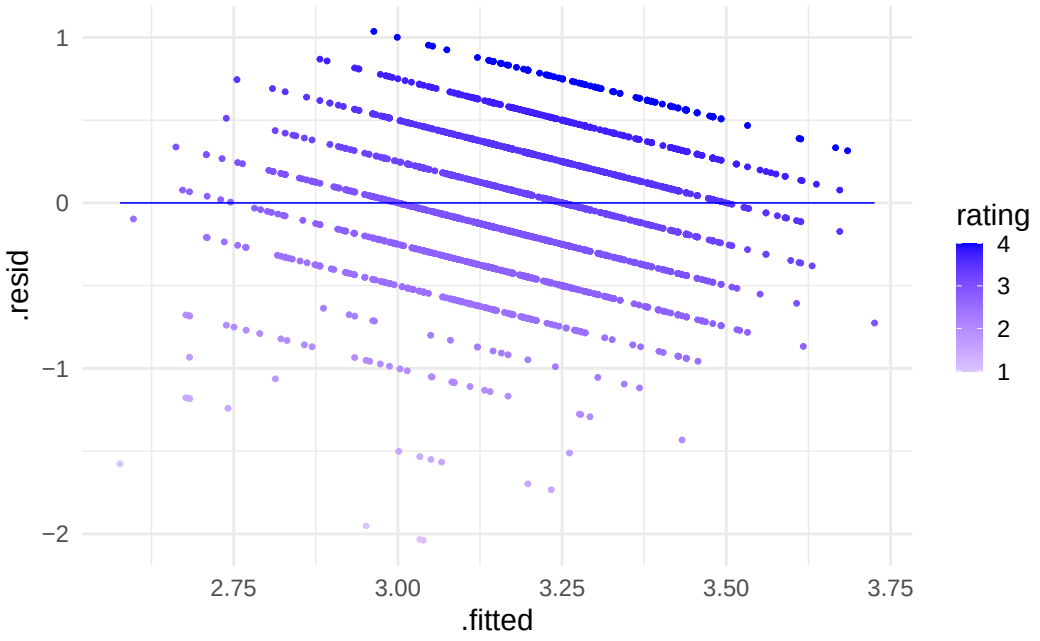

Figure 12: Residuals and Ratings