Draft

STA 210 - Project

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

Introduction and data

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

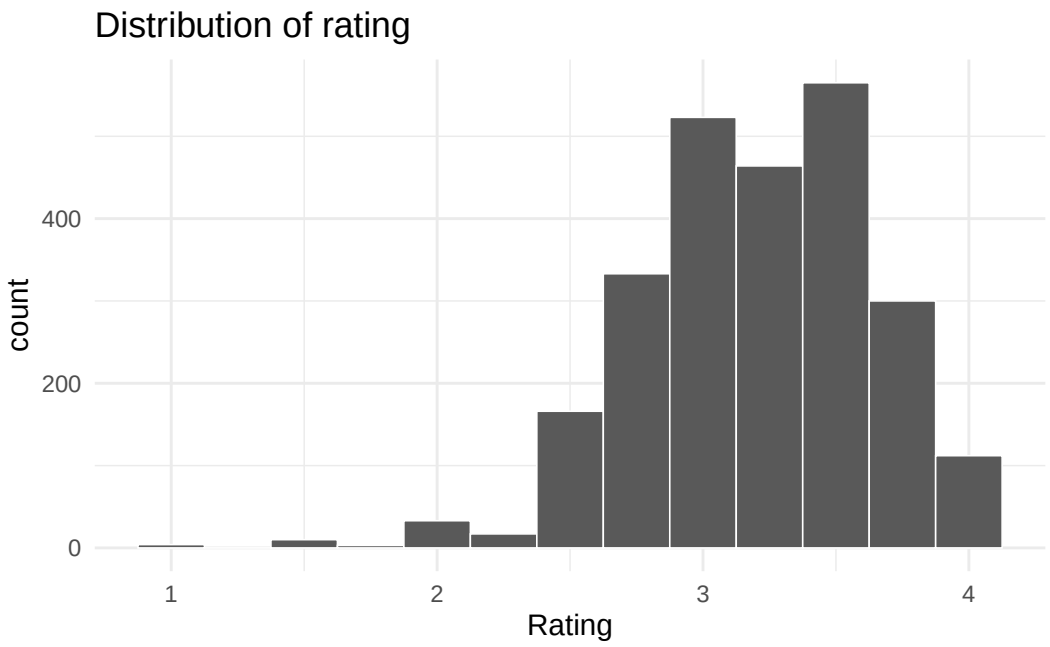
The general characteristics that will be our main interest are described as follows:

- 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.

Exploratory data analysis

Shape of Ratings

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

Cocoa Percent

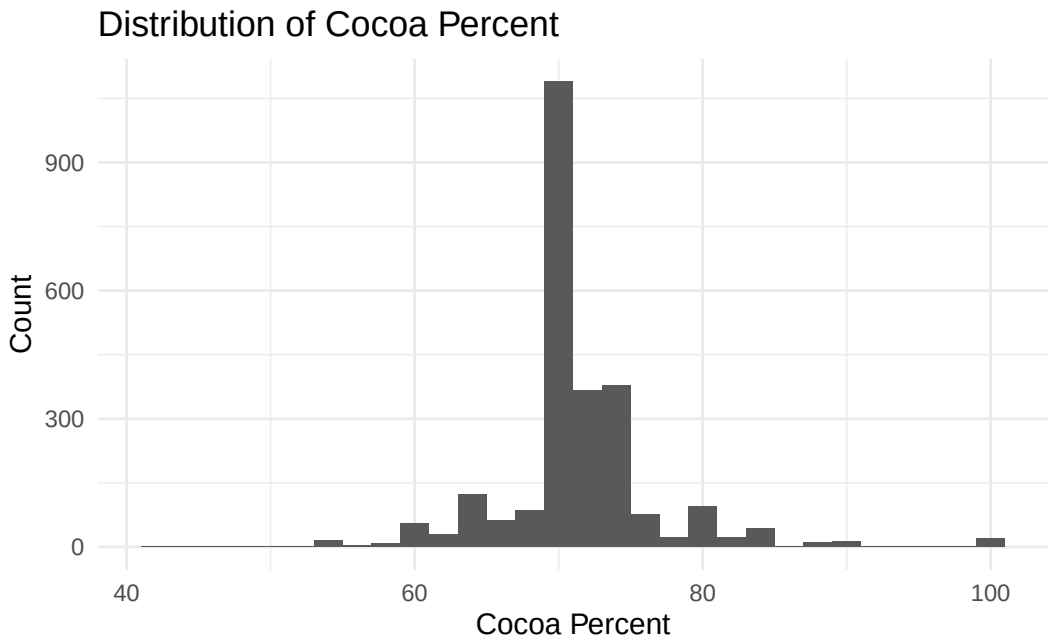
From the distribution of cocoa_percent, we see that the distribution is roughly symmetric and unimodal, and centered around 72 percent, and has apparent outliers around 55 percent and 100 percent.

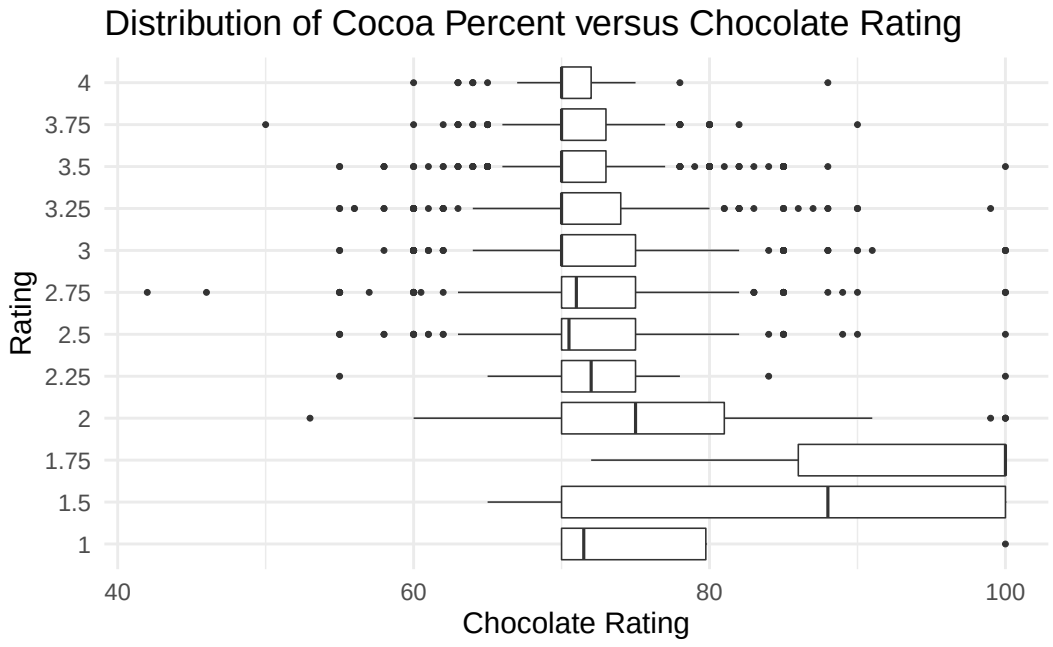
From the boxplot, we can see a general rough trend that as the median cocoa percent is lower, the rating of the chocolate bar is higher. Furthermore, there appear to be a lot of outliers in the middle ratings (2.25 - 3.75), which might be due to the fact that that is the rating for the bulk of the chocolates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

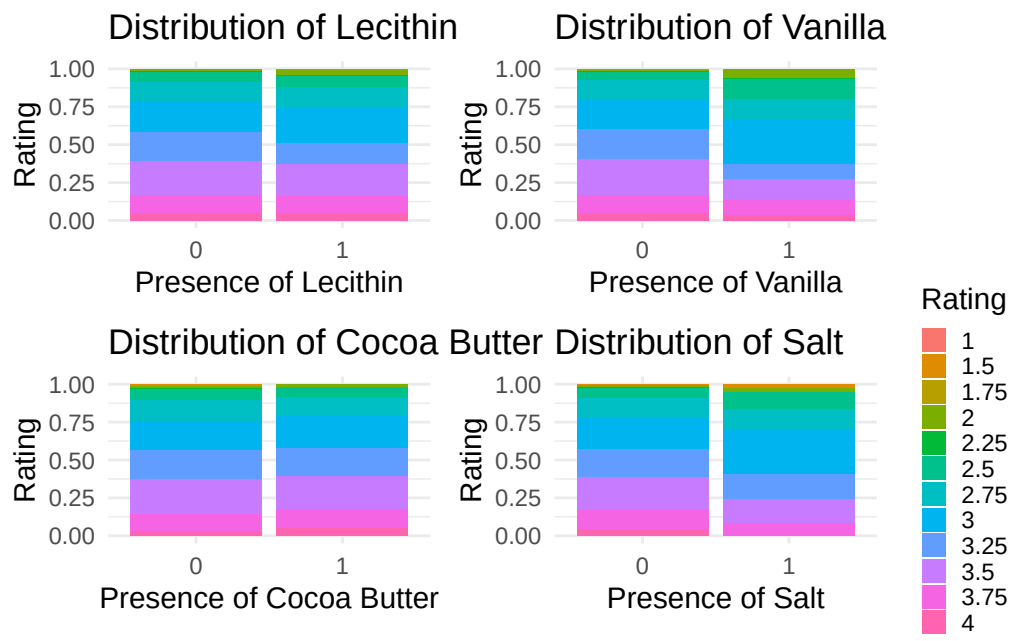
Ingredients

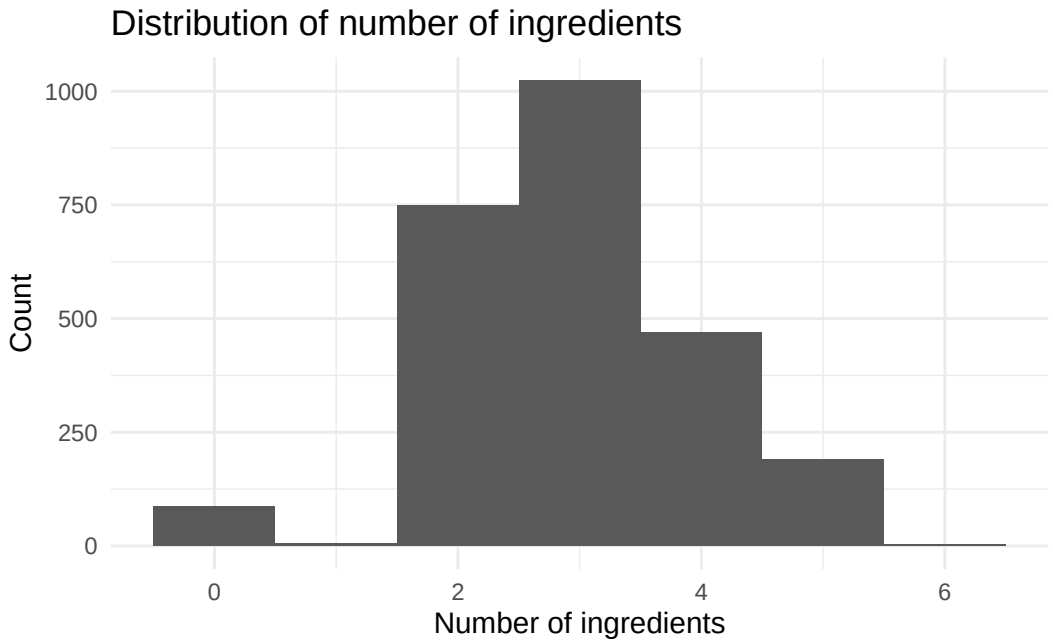
From this visualization, we can see that the presence of salt and vanilla seem to affect the rating the most out of all the predicters. The presence of salt and vanilla results in more lower ratings, while the amount of high and low ratings remains roughly the same with/without the presence of cocoa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

In this visualization 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.


Distribution of Cocoa Percent
Count
900
600
300
0
40
60
80
100
Cocoa Percent

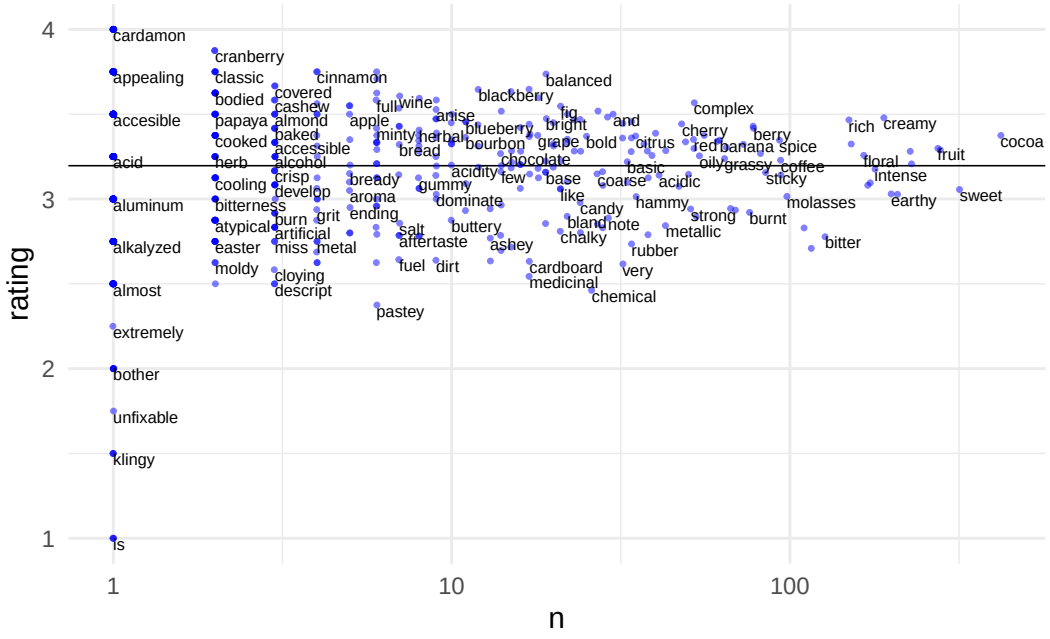

Distribution of Cocoa Percent versus Chocolate Rating
Rating
4
3.75
3.5
3.25
3
2.75
2.5
2.25
2
1.75
1.5
1
40
60
80
100
Chocolate Rating


Distribution of Lecithin
Distribution of Vanilla
Distribution of Cocoa Butter
Distribution of Salt
Rating
1.00
0.75
0.50
0.25
0.00
0
1
Presence of Lecithin
Presence of Vanilla
Presence of Cocoa Butter
Presence of Salt
Rating
1
1.5
1.75
2
2.25
2.5
2.75
3
3.25
3.5
3.75
4


Distribution of number of ingredients
Count
1000
750
500
250
0
0
2
4
6
Number of ingredients

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.

Most Memorable Characteristic



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”.

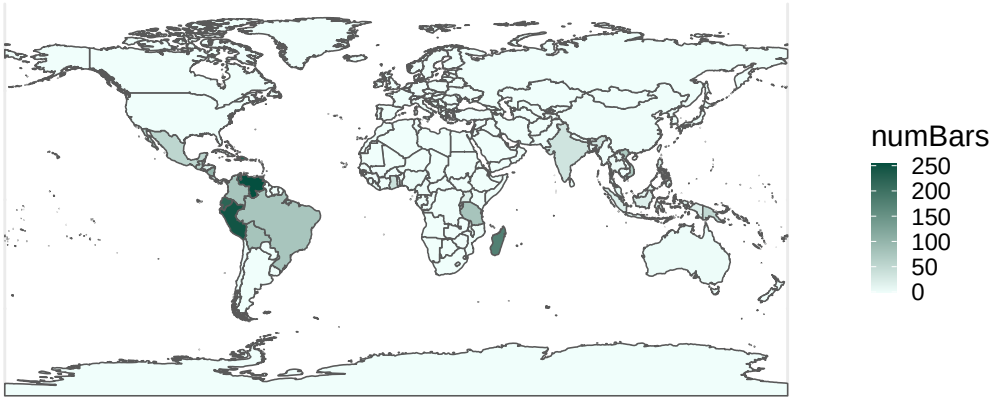
Country Bean of Origin

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

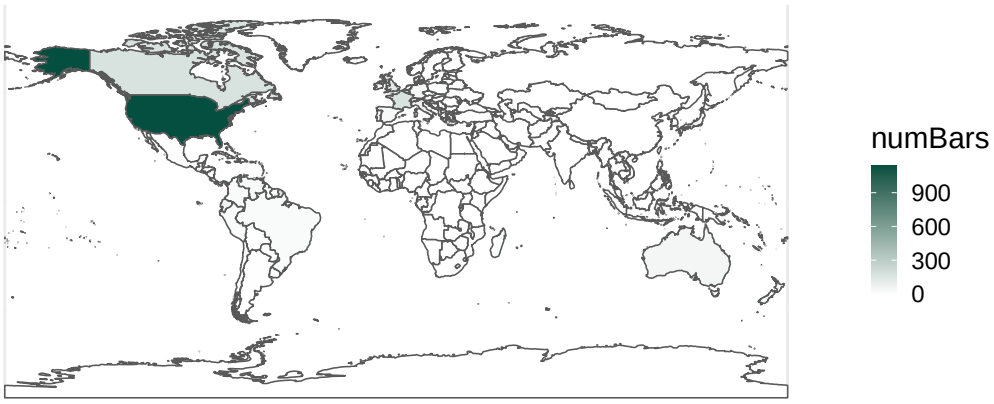
Company Location

This map 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.

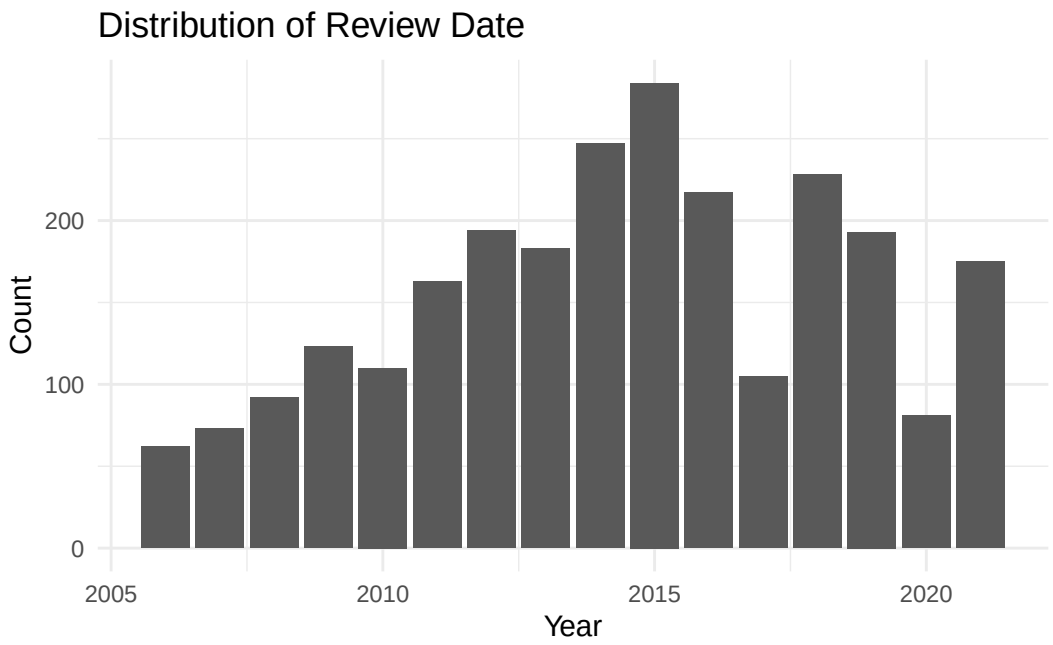
Map of countries where cacao beans were produced



Map of countries where companies are located



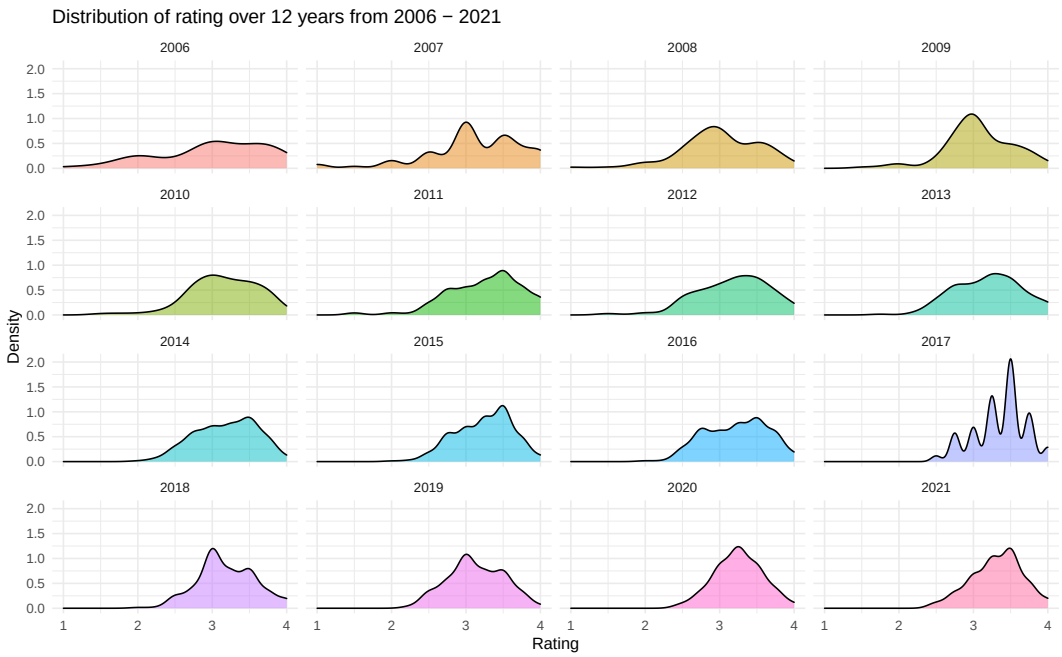
Review Date



Here, we can see that the distribution of chocolate bars reviewed over time has a roughly unimodal distribution with a peak around 2015. Furthermore there was a signficant dip in 2020, probably due to the COVID-19 Pandemic, as well as a dip in 2017, due to unknown reasons. The distribution is centered around 2014 and is roughly symmetric.

```
# A tibble: 1 x 3
   mean median    sd
  <dbl>  <dbl> <dbl>
1 2014.   2015  3.97
```

This visualization showcases the distribution of ratings for each review year. There is no apparent change or pattern to the change in ratings of years, and it appears that ratings from 2.5 - 3.25 compose the bulk of the ratings each year.



For cleaning the most memorable characteristic column, we assumed that the first word in the column was the dominant memorable characteristic. From there, we created groups out of the most popular words. For example, characteristics that contained the word "fruit" or "berry" were grouped together into "fruity".

Methodology

Our main goal of this analysis is to understand how the characteristics of a chocolate can explain its rating. Since the rating is treated as a quantitative variable, we will perform a linear regression model to fit and predict the rating from the features of a chocolate. We also wish to examine which combination of predictors would make the best model for prediction, so here we examine three different scenarios: - The rating of a chocolate might depend on their percentage of cocoa, their ingredients, and their most memorable characteristics recorded. - The rating of a chocolate might depend on more predictors besides those list above, for example, on the number of ingredients presented in the chocolate, the origin of the company (divided into continents), and the origin of the cocoa bean (divided into continents). - It is possible that the rating of a chocolate might depend only on the company location and the origin of the bean.

So we will perform a linear regression on these three scenarios, and evaluate which model perform the best. Our initial approach is to use a simple regression and compared adjusted R-squared, AIC, and BIC, as well as checking their conditions, but the differences of those

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

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

statistics between the two models are almost the same which does not tell much, so we perform a cross-validation and compare r-squared and rsme instead.

(possible, if we decide to add it in) We also consider a scenario where there are some strong interactions between the predictors,

Results

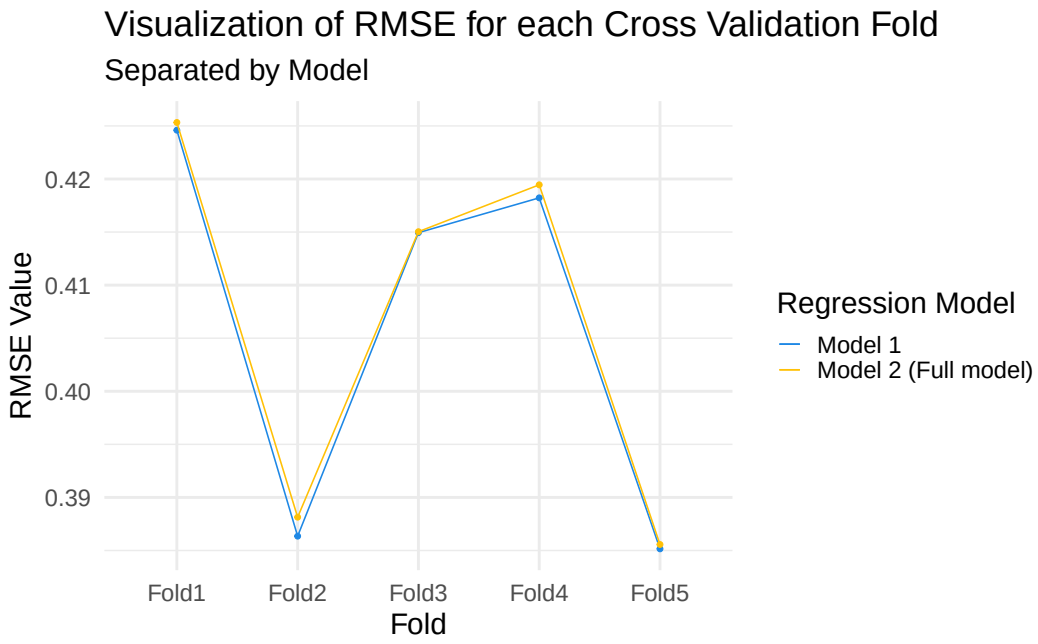
Ratings vs cocoa percent, ingredients, most memorable characteristics

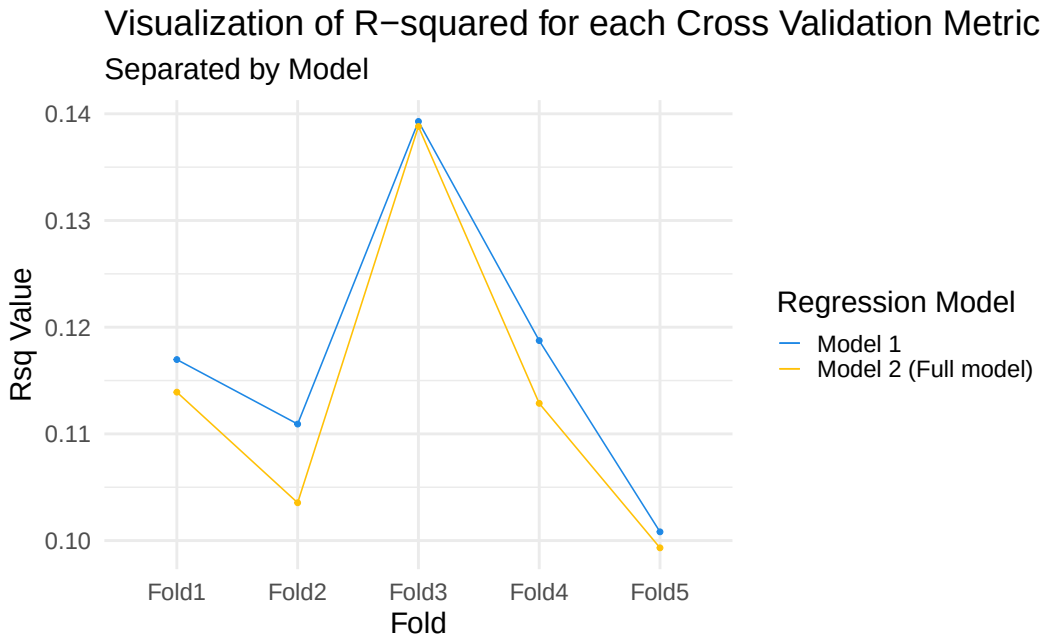
All predictors

As both models have similar RMSE and R-squared values for each fold in cross-validation we will choose the first model as it has fewer predictor variables and aligns with the goals of parsimony.

Data

The data dictionary can be found here.


Visualization of RMSE for each Cross Validation Fold
Separated by Model
RMSE Value
0.42
0.41
0.40
0.39
Fold1
Fold2
Fold3
Fold4
Fold5
Fold
Regression Model
Model 1
Model 2 (Full model)


Visualization of R−squared for each Cross Validation Metric
Separated by Model
Rsq Value
0.14
0.13
0.12
0.11
0.10
Fold1
Fold2
Fold3
Fold4
Fold5
Fold
Regression Model
Model 1
Model 2 (Full model)