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. Chocolate companies that are located in the USA or a European country will have higher ratings.
3. Cocoa percentage and ingredients are the strongest predictors.
4. Country of bean origin will not be a strong predictor.

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:

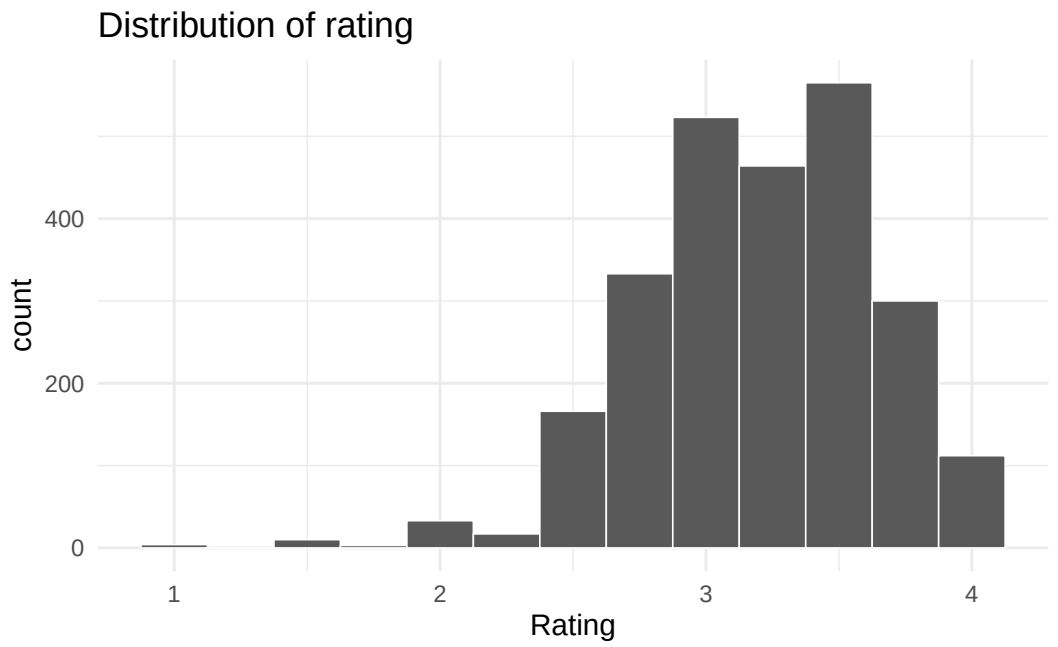
- Company (Manufacturer) lists who made the chocolate bar reviewed; the dataset also lists where this company is located under Company Location.
- The dataset characterizes the Country of Bean Origin, Specific Bean Origin or name of bar, Percentage of Cocoa within the bar for each chocolate bar.
- The 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

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

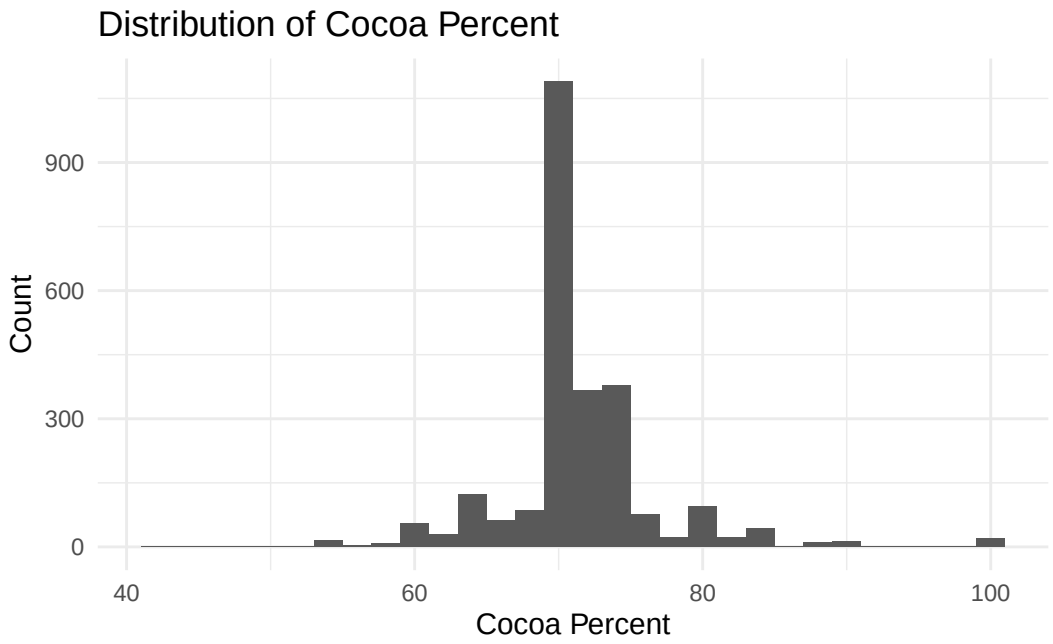
Cocoa Percent

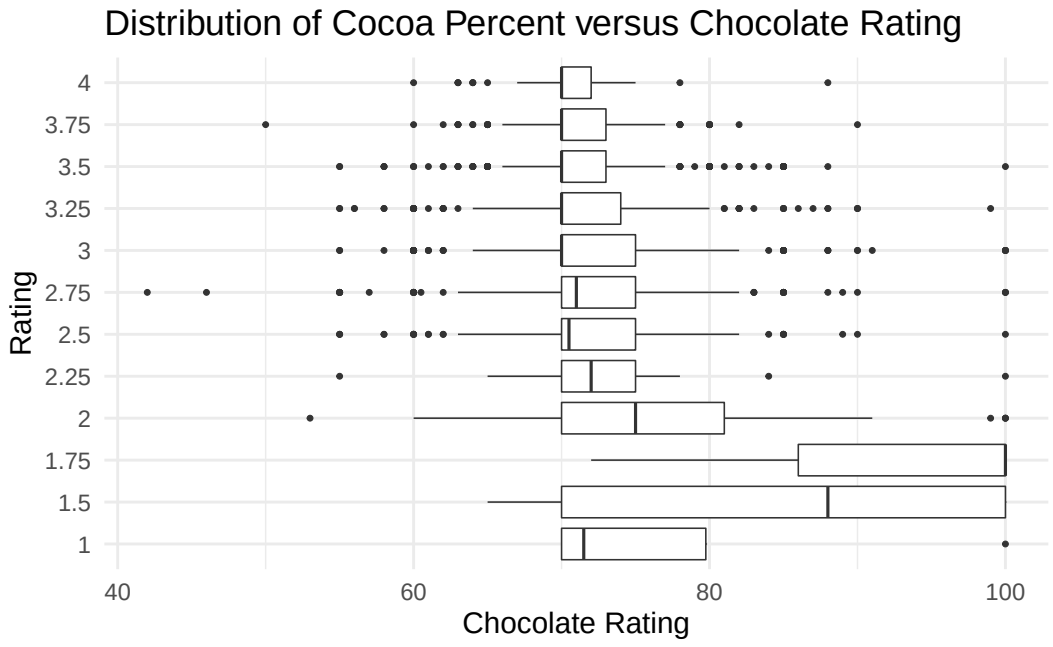
From the distribution of cocoa_percent, we see that the distribution is roughly symmetric and unimodal, and centered around 72 percent, and has apparent outliers around 55 percent and 100 percent.

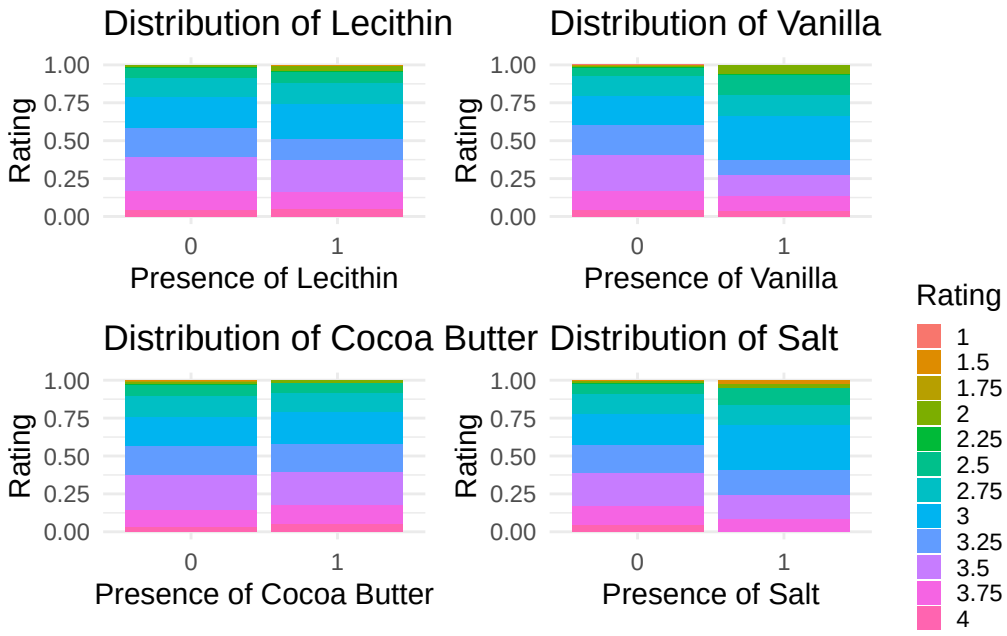
From the boxplot, we can see a general rough trend that as the median cocoa percent is lower, the rating of the chocolate bar is higher. Furthermore, there appear to be a lot of outliers in the middle ratings (2.25 - 3.75), which might be due to the fact that that is the rating for the bulk of the chocolates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


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



ratings, while the amount of high and low ratings remains roughly the same with/without the presence of cocoa butter and lecithin.

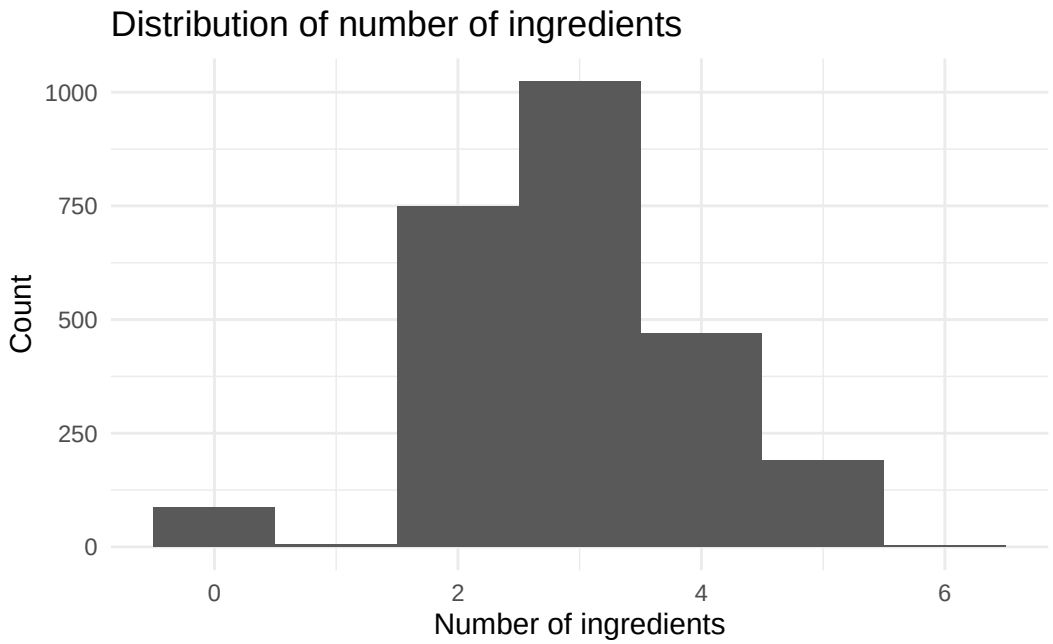
In this visualization we regard the NA value of ingredients as 0, which means we should understand this as nonrecorded value instead of no ingredients are presented in the chocolate. This visualization showcases a right skewed distribution for the number of ingredients. The median is somewhere around 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.

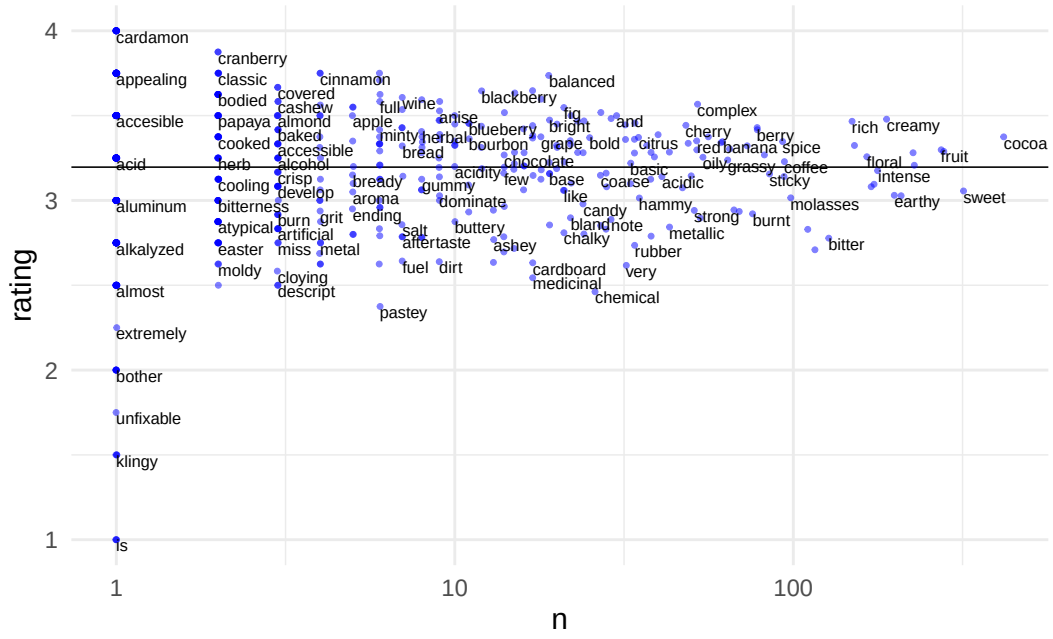
Most Memorable Characteristic

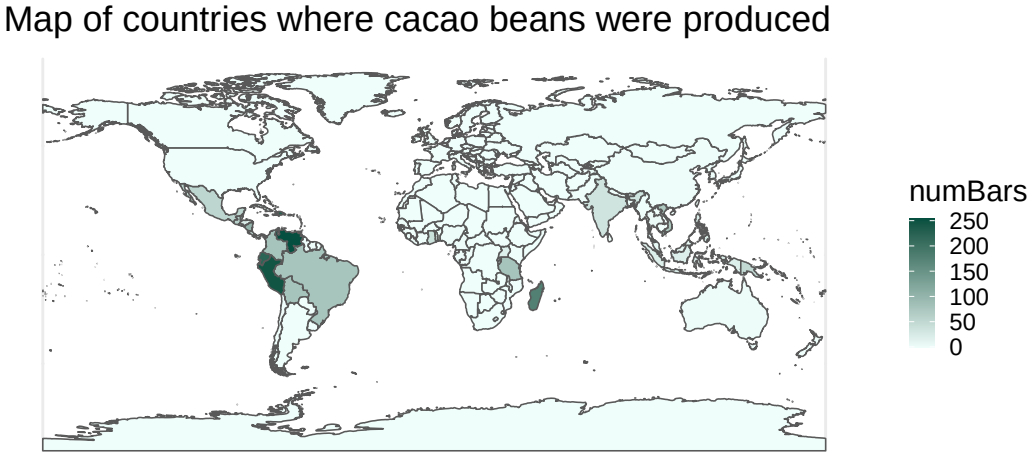
From this visualization, we can see that the phrases and most memorable charactersists that were often associated with a higher rating were “balanced” and “complex”, as well as fruity chocolate like “fruit”, “Cardamon”, “floral”.

Country Bean of Origin

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


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


rating
4
3
2
1
n
1
10
100
cardamon
cranberry
appealing
classic
cinnamon
balanced
bodied
covered
blackberry
cashew
full
wine
fig
complex
accesible
papaya
almond
apple
anise
and
bright
cherry
berry
rich
creamy
cooked
baked
minty
herbal
blueberry
bourbon
grape
bold
citrus
red
banana
spice
cocoa
acid
accessible
bread
chocolate
oily
grassy
coffee
floral
fruit
herb
alcohol
acidity
few
base
coarse
basic
acidic
sticky
intense
cooling
crisp
develop
bready
gummy
like
hammy
molasses
earthy
sweet
aluminum
bitterness
aroma
dominate
candy
strong
burnt
grit
ending
atypical
burn
artificial
buttery
blandness
metallic
alkalyzed
easter
miss
metal
salt
aftertaste
ashey
chalky
rubber
bitter
moldy
fuel
dirt
cardboard
very
almost
cloying
descript
medicinal
chemical
pastey
extremely
bother
unfixable
klingy
is



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.

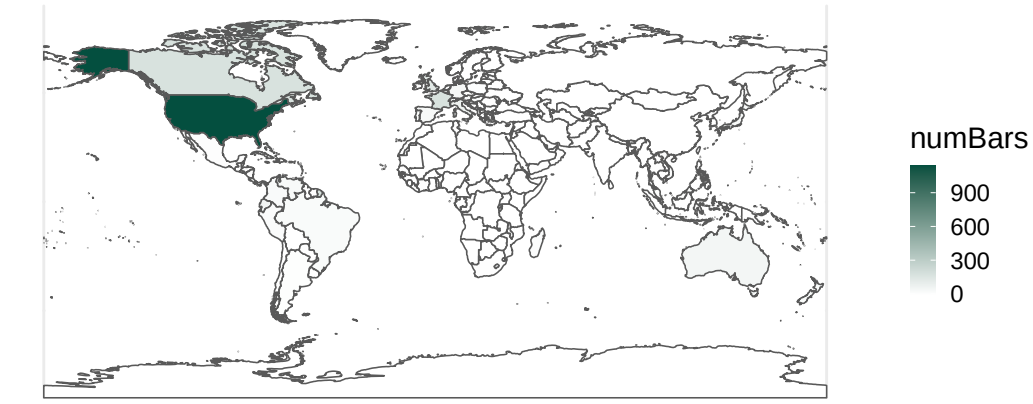
Review Date

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.

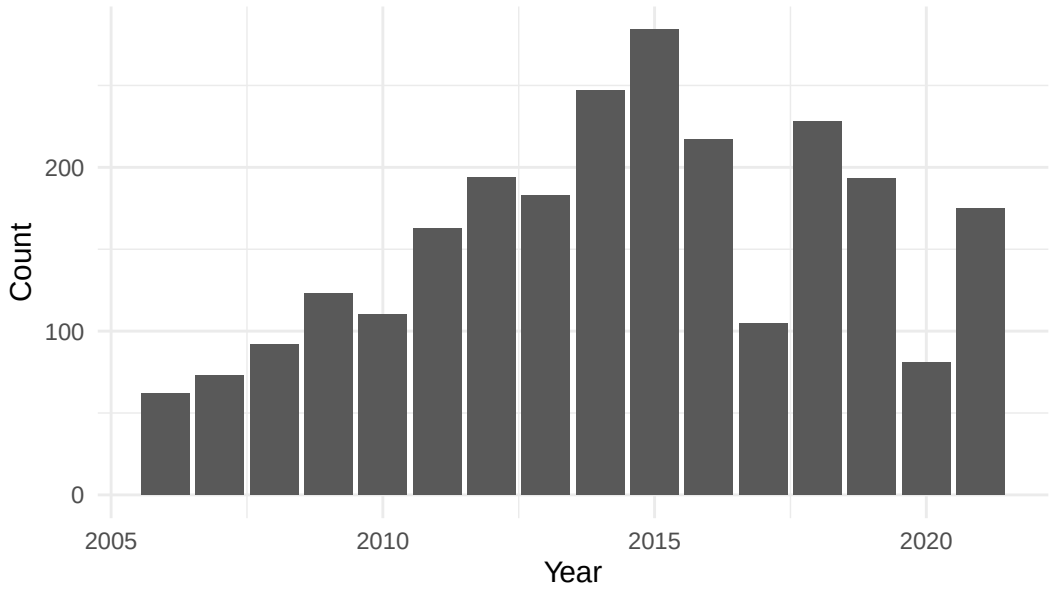
```
# A tibble: 1 x 3
   mean median    sd
  <dbl>  <dbl> <dbl>
1 2014.   2015  3.97
```

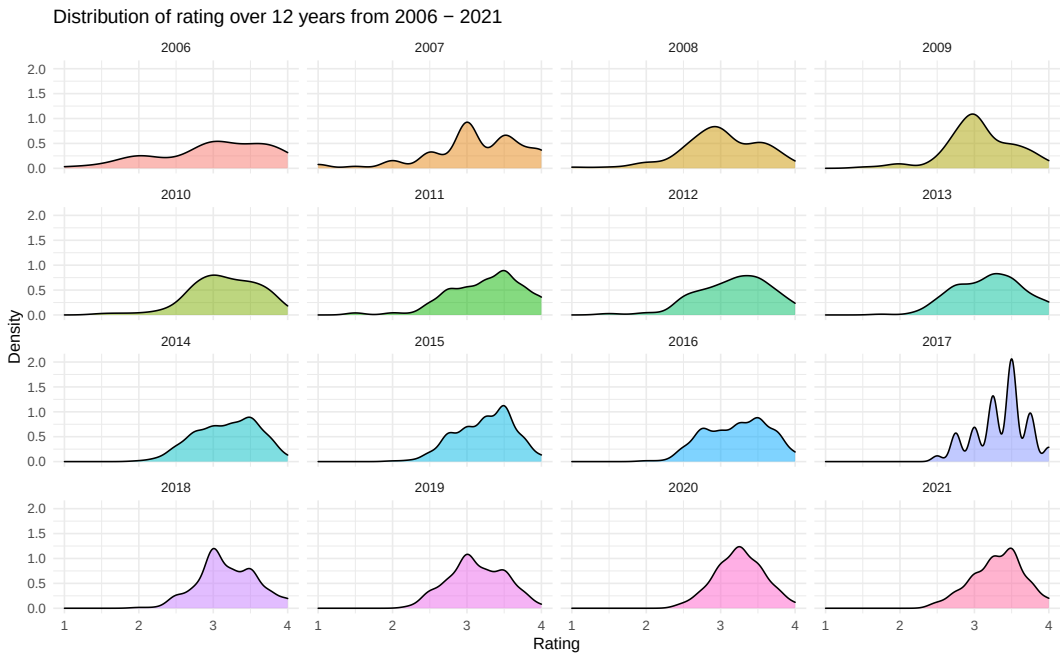
This visualization showcases the distribution of ratings for each review year. There is no apparent change or pattern to the change in ratings of years, and it appears that ratings from 2.5 - 3.25 compose the bulk of the ratings each year.

Map of countries where companies are located



Distribution of Review Date





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".

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 most noticeable 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".
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.

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

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.
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.

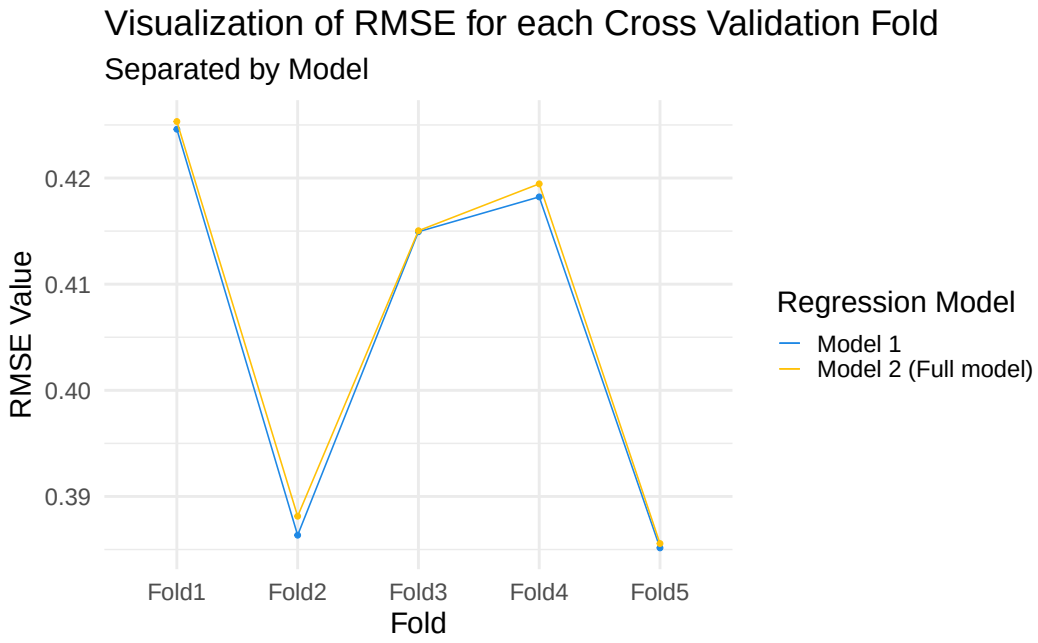
Modeling

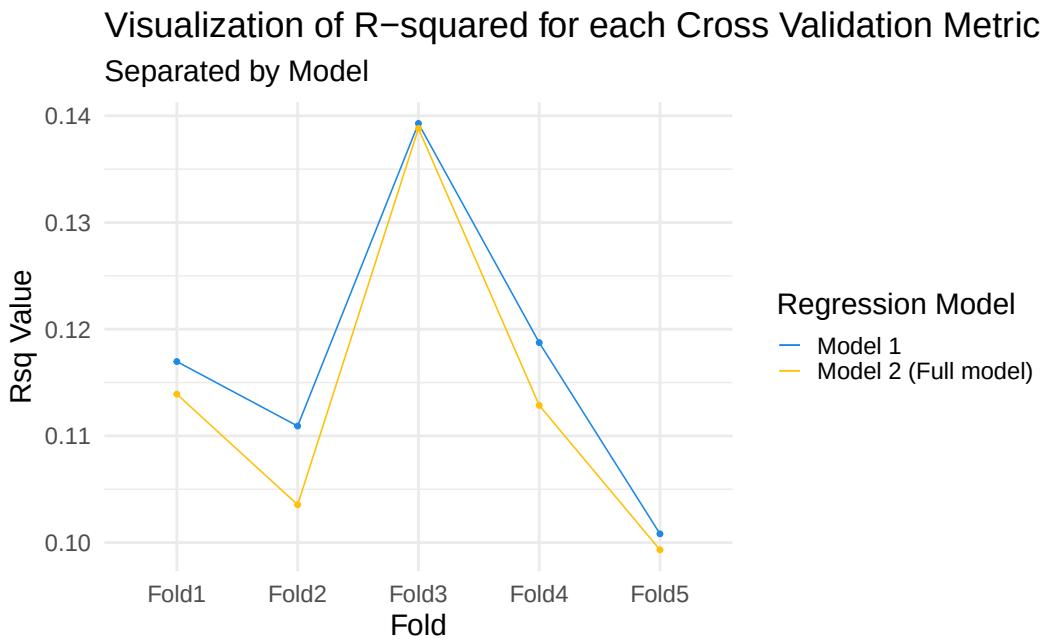
Our main goal of this analysis is to understand how the characteristics of a chocolate can explain its rating. Although the rating only goes from 1 to 5 in 0.25 increments, we treated rating as a quantitative variable. Therefore, we used a linear regression model to fit and predict the rating from the 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.

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





Results

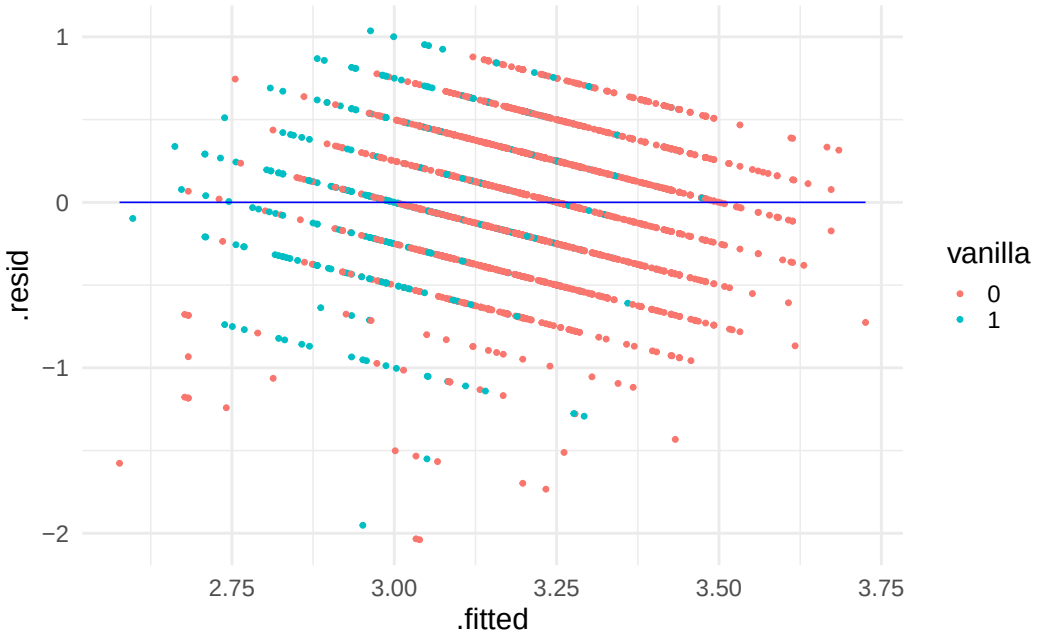
Ratings vs cocoa percent, ingredients, most memorable characteristics

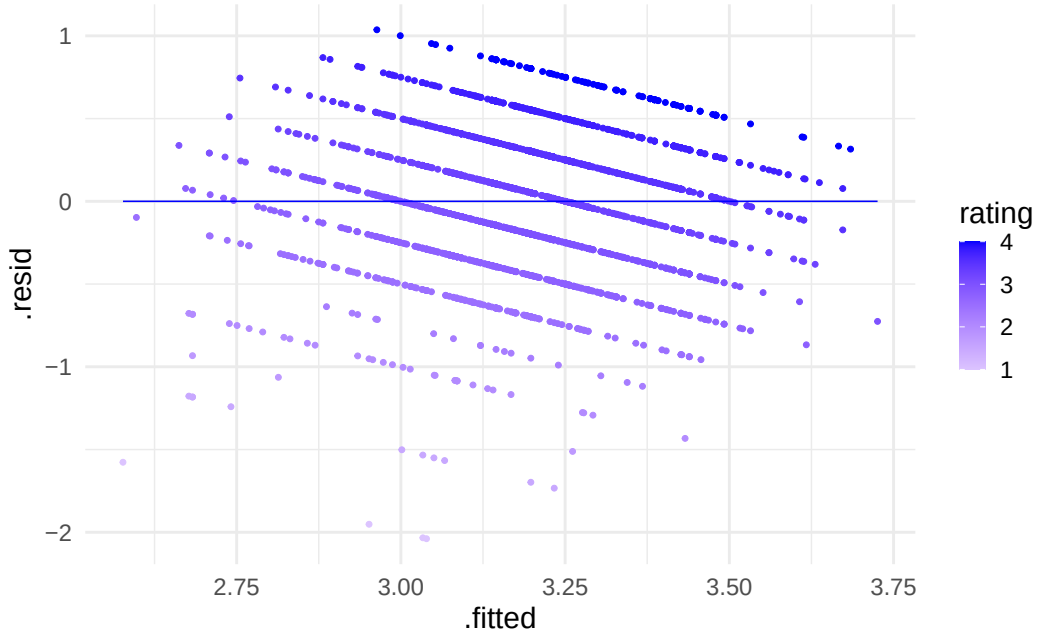
All predictors

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, as it is a less complicated model yet doesn't sacrifice predictive capability.

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




1
0
−1
−2
.resid
2.75
3.00
3.25
3.50
3.75
.fitted
rating
4
3
2
1

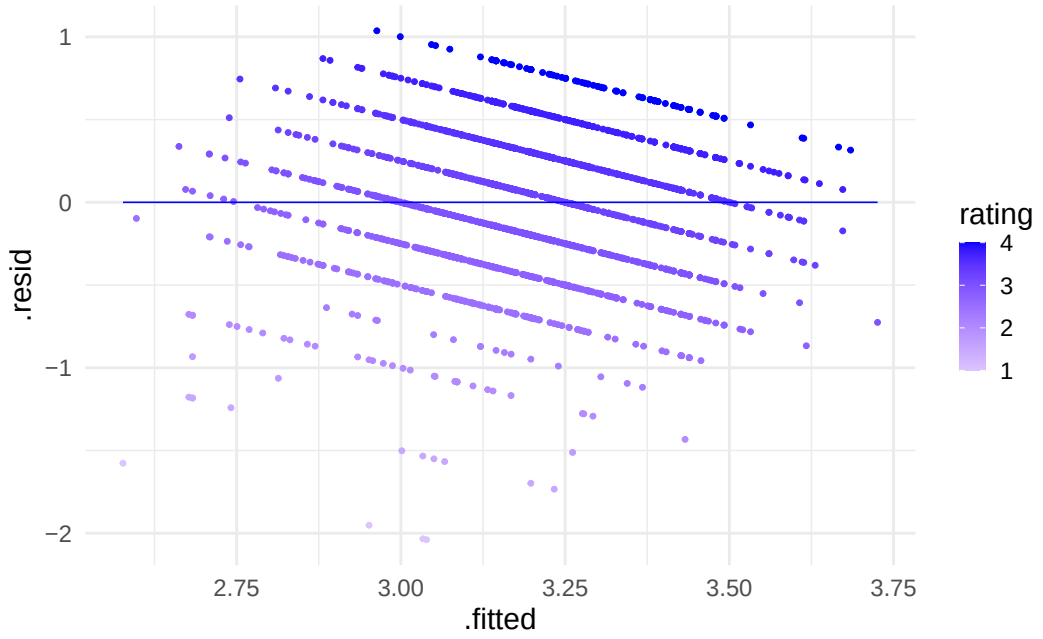
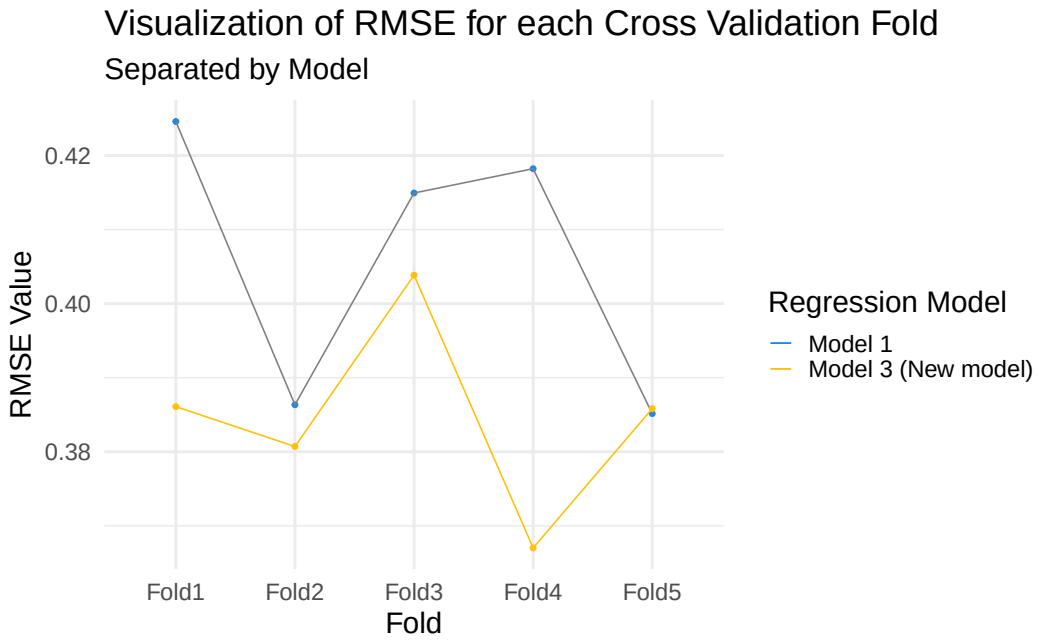
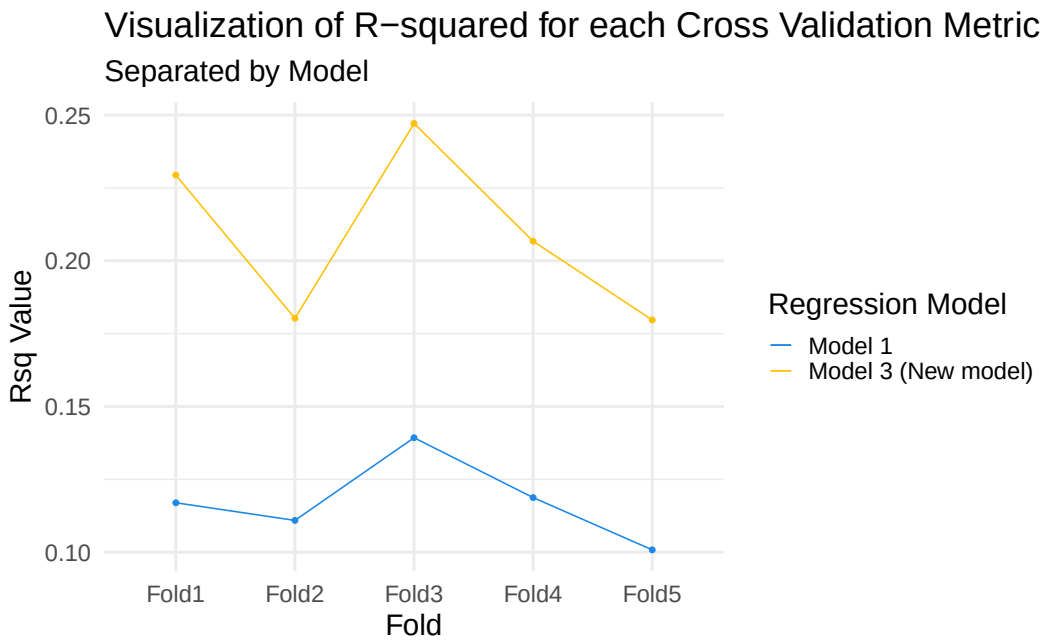

1
0
−1
−2
.resid
2.75
3.00
3.25
3.50
3.75
.fitted
rating
4
3
2
1

Table 4: Model 3

Fold	RMSE	R-squared
Fold1	0.386	0.229
Fold2	0.381	0.180
Fold3	0.404	0.247
Fold4	0.367	0.207
Fold5	0.386	0.180





term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.277	0.254	12.902	0.000	2.778	3.775
cocoa_percent	-0.006	0.004	-1.541	0.124	-0.013	0.002
num_ingres	0.027	0.021	1.276	0.202	-0.014	0.068
isCocoa	0.128	0.043	2.933	0.003	0.042	0.213
isOff	-0.212	0.056	-3.811	0.000	-0.322	-0.103
isChemical	-0.811	0.149	-5.446	0.000	-1.103	-0.518
isFruit	0.060	0.053	1.139	0.255	-0.044	0.164
isCreamy	0.339	0.089	3.821	0.000	0.165	0.514
isComplex	0.359	0.207	1.731	0.084	-0.048	0.765
isBitter	-0.560	0.089	-6.320	0.000	-0.735	-0.386
vanilla_X1	-0.291	0.063	-4.628	0.000	-0.414	-0.167
salt_X1	-0.165	0.128	-1.288	0.198	-0.416	0.086
top_memorable_fatty_smooth	-0.019	0.257	-0.074	0.941	-0.524	0.486
top_memorable_floral	-0.023	0.264	-0.086	0.931	-0.542	0.496
top_memorable_fruity	0.176	0.259	0.682	0.496	-0.332	0.685
top_memorable_greasy	0.148	0.258	0.572	0.568	-0.360	0.655
top_memorable_nutty	0.051	0.264	0.194	0.846	-0.468	0.570
top_memorable_other	-0.122	0.253	-0.482	0.630	-0.618	0.375
top_memorable_roast	-0.137	0.258	-0.530	0.596	-0.643	0.370
top_memorable_rough_texture	-0.201	0.257	-0.784	0.433	-0.706	0.303
top_memorable_spiced	-0.044	0.262	-0.167	0.868	-0.558	0.471
top_memorable_strong_sweet	-0.012	0.256	-0.046	0.963	-0.514	0.491

We have assumed that this data fits the condition