

Draft

STA 210 - Project

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

Exploratory Data Analysis

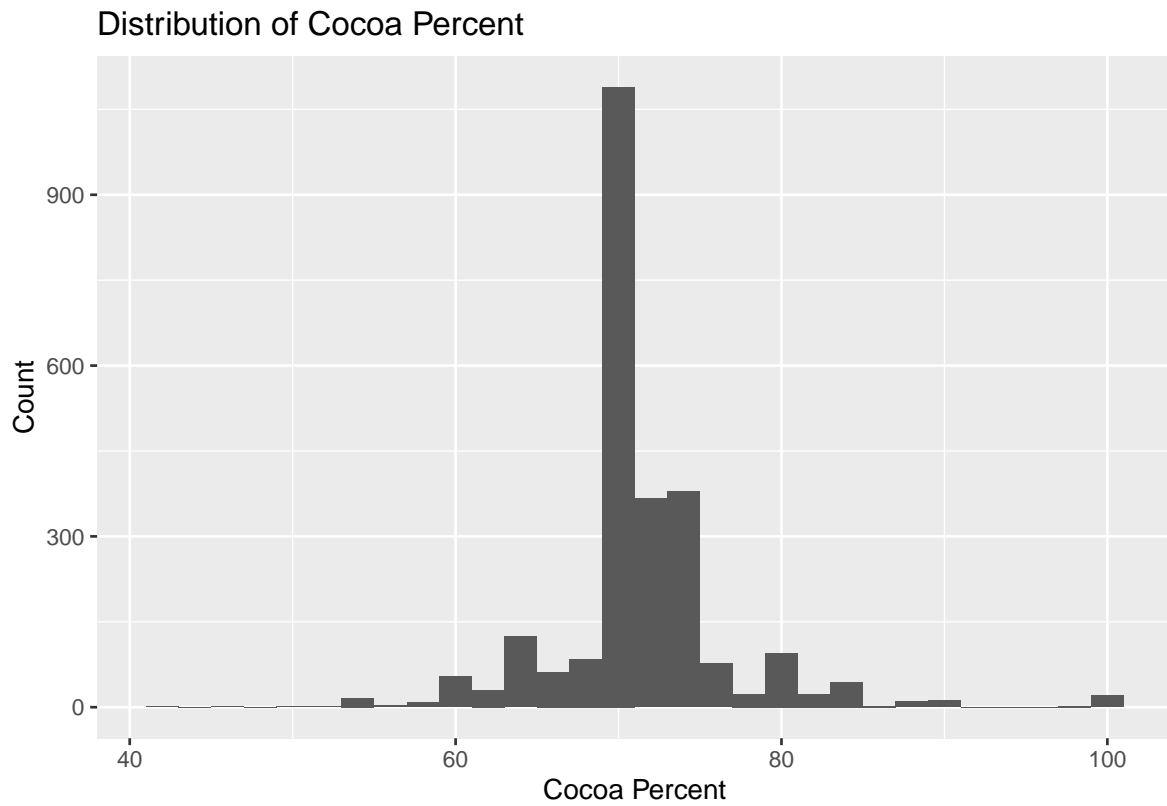
Data description

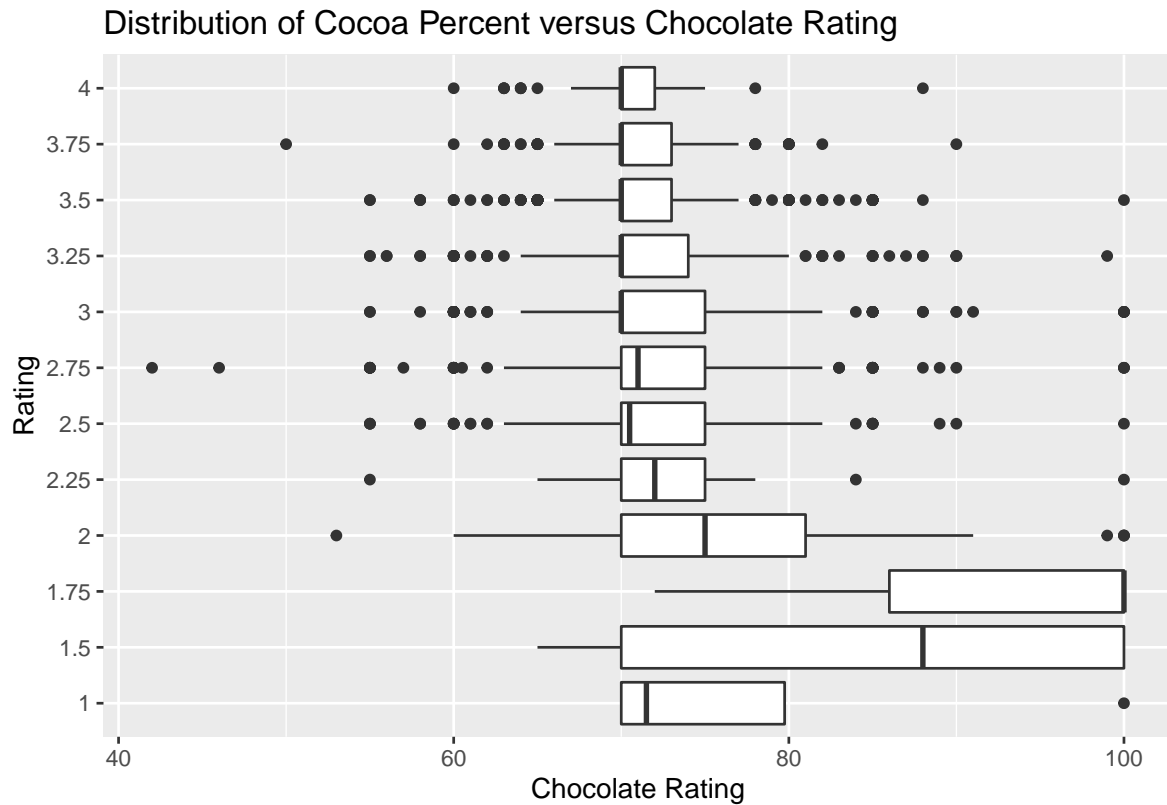
- Description of the observations in the data set:
 - The observations in this data set represent a review of general characteristics for different chocolate bars. A single observation in this data set represents a single chocolate bar.
 - The general characteristics are as follows:
 - * Company (Manufacturer) 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
- Description of how the data was originally collected (not how you found the data but how the original curator of the data collected it).
 - Data 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.

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.

Shape of Ratings (already done)

Cocoa Percent

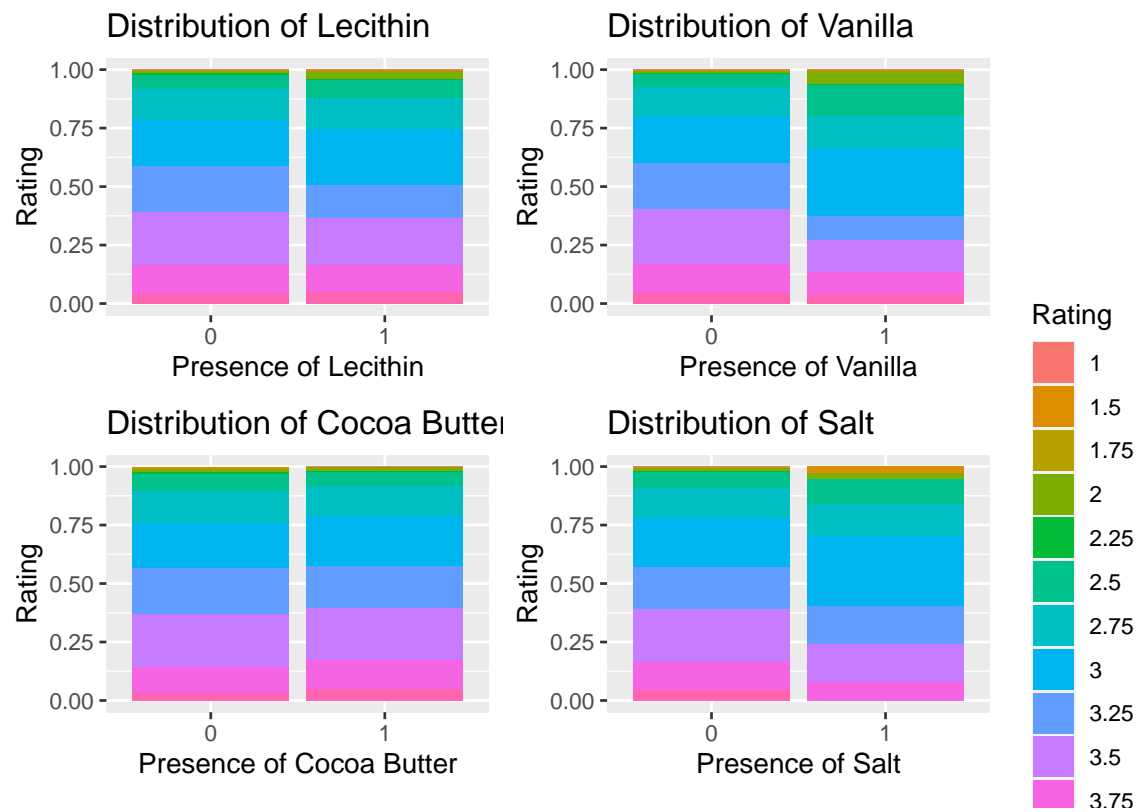




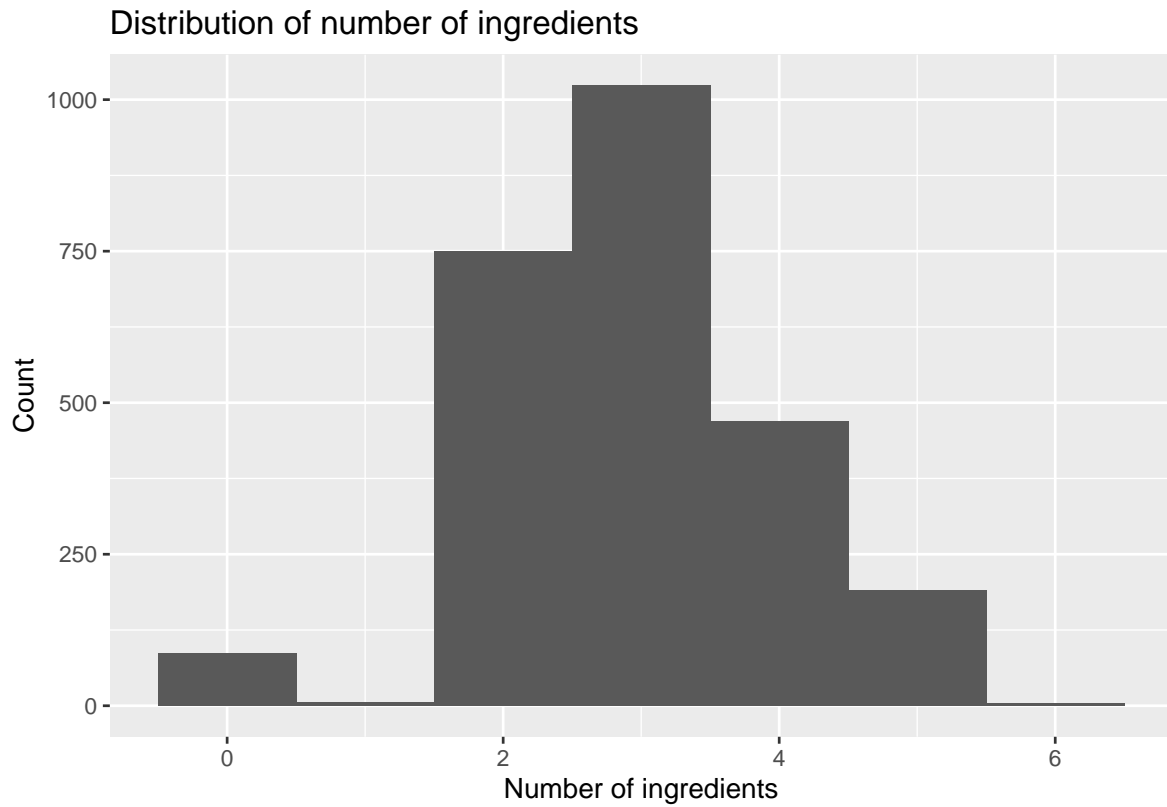
From the distribution of `cocoa_percent`, we see that the distribution is roughly symmetric and unimodal, and centered around 72 percent, and has apparent outliers around 55 percent and 100 percent.

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

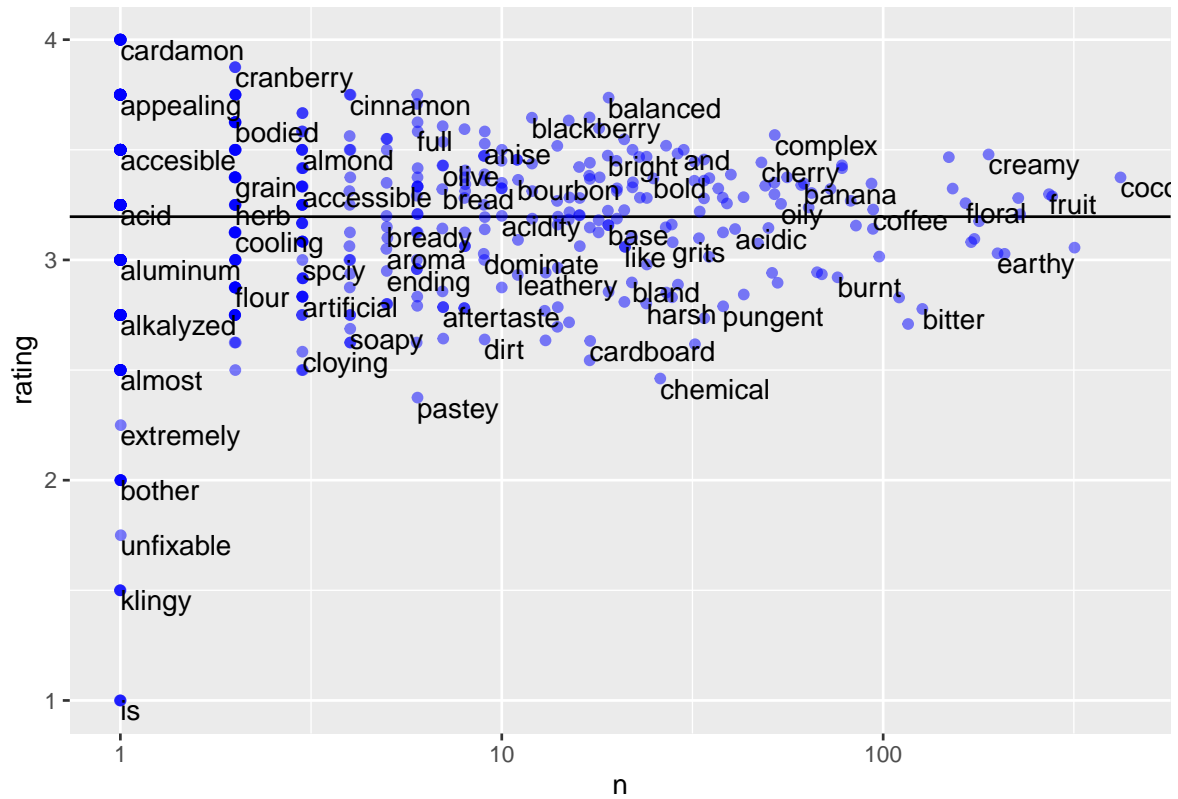


From this visualization, we can see that the presence of salt and vanilla seem to affect the rating the most out of all the predictors. The presence of salt and vanilla results in more lower ratings, while the amount of high and low ratings remains roughly the same with/without the presence of cocoa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



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.

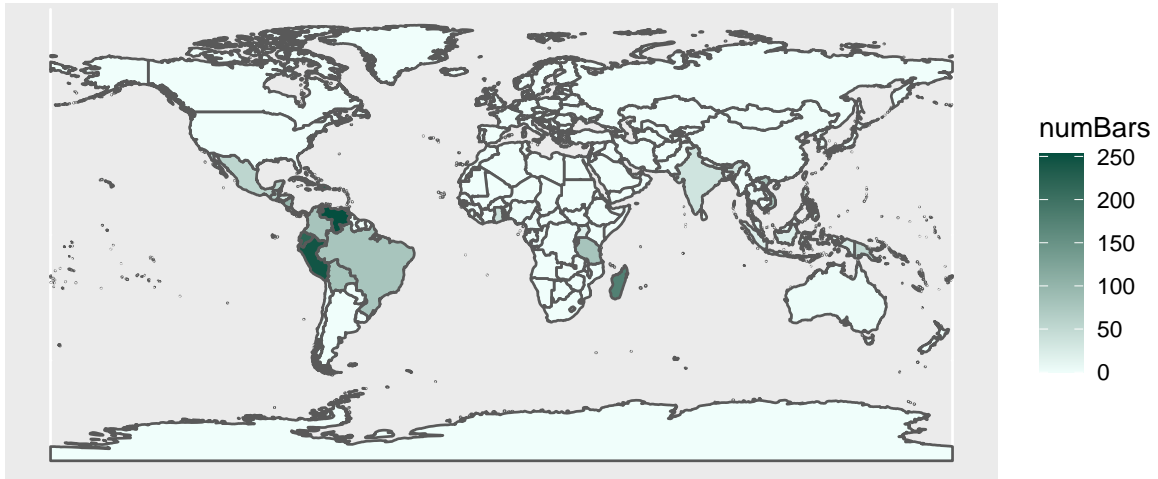
Most Memorable Characteristic (Aimi)



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 (Rakshita)

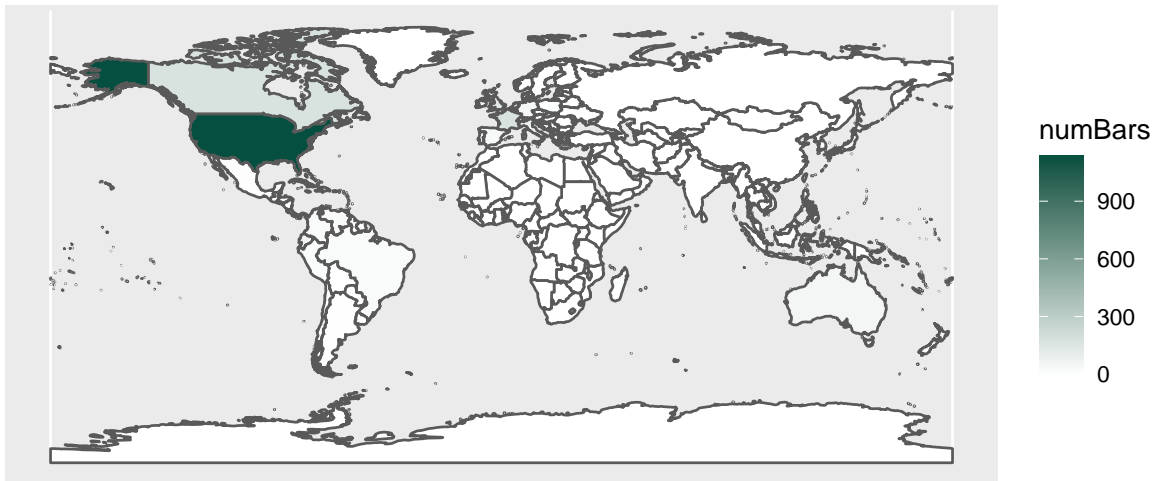
Map of countries where cacao beans were produced



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

Company Location (Rakshita)

Map of countries where companies are located

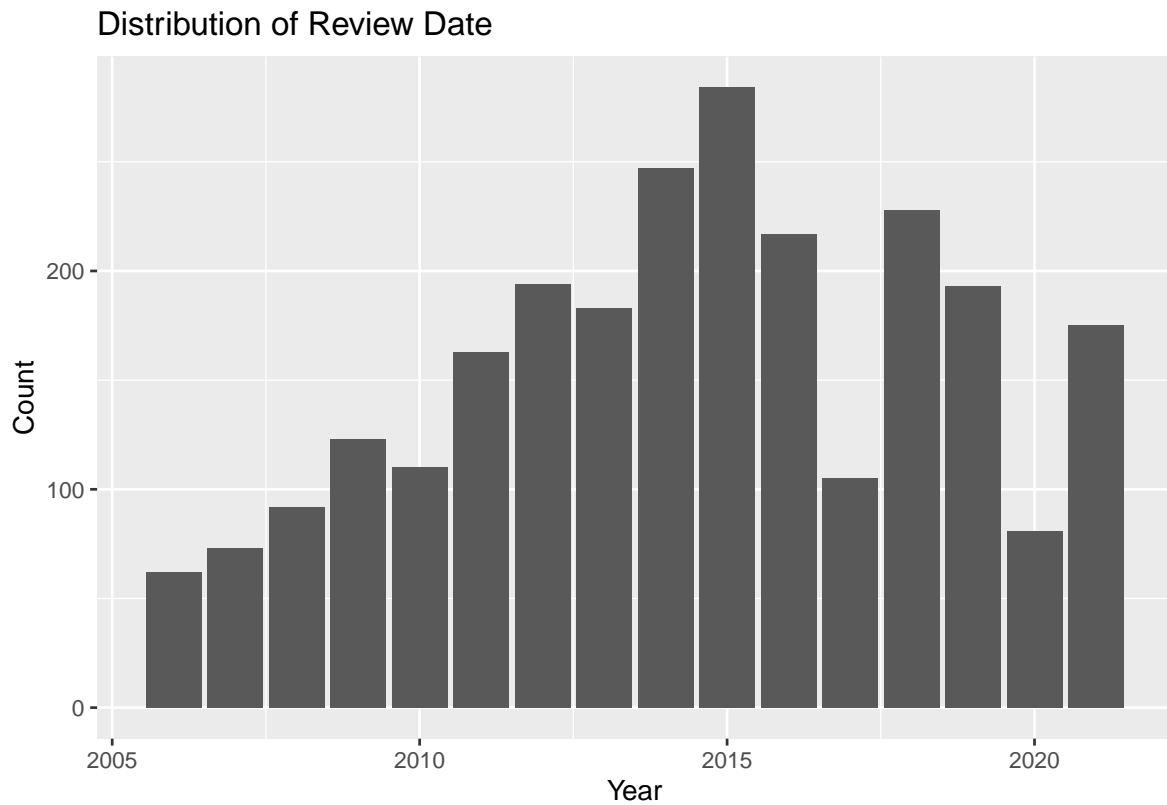


```
# A tibble: 67 x 2
  company_location    n
  <chr>             <int>
1 U.S.A.           1136
2 Canada            177
3 France            176
4 U.K.              133
5 Italy             78
6 Belgium           63
7 Ecuador           58
8 Australia         53
9 Switzerland       44
10 Germany          42
# ... with 57 more rows
```

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 (Nathan)



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 significant 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
```

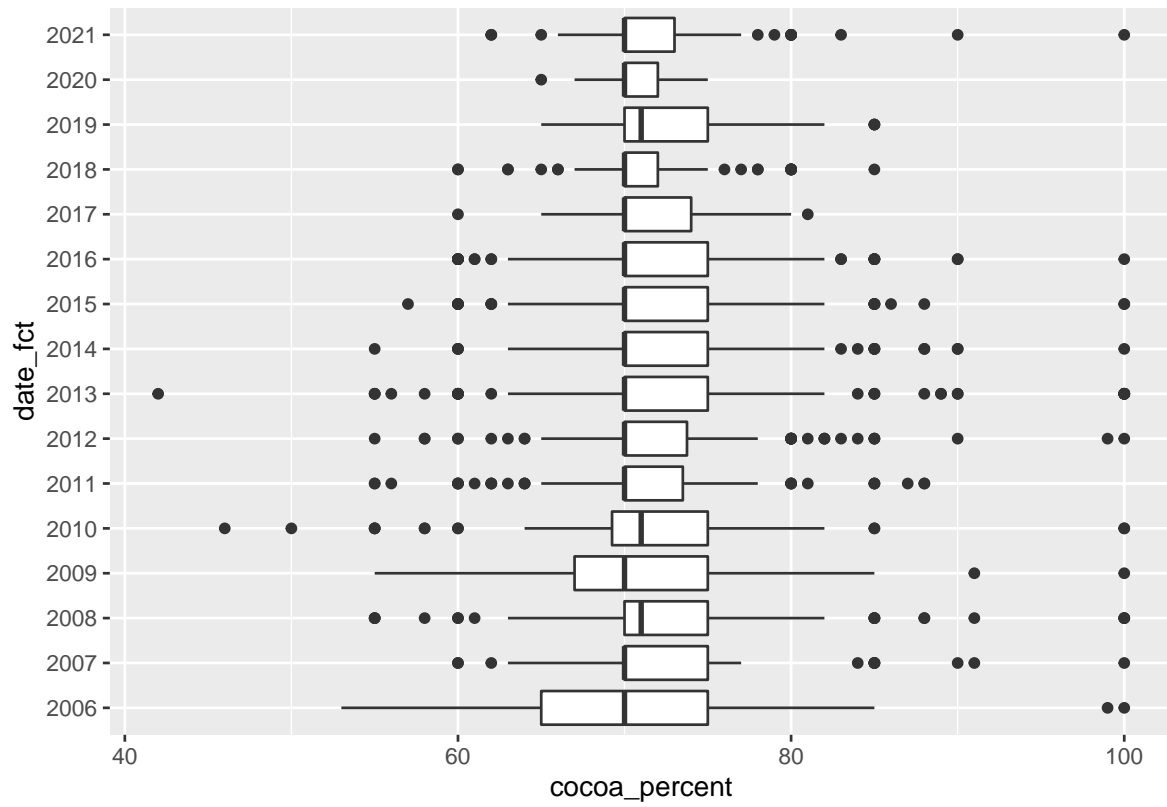
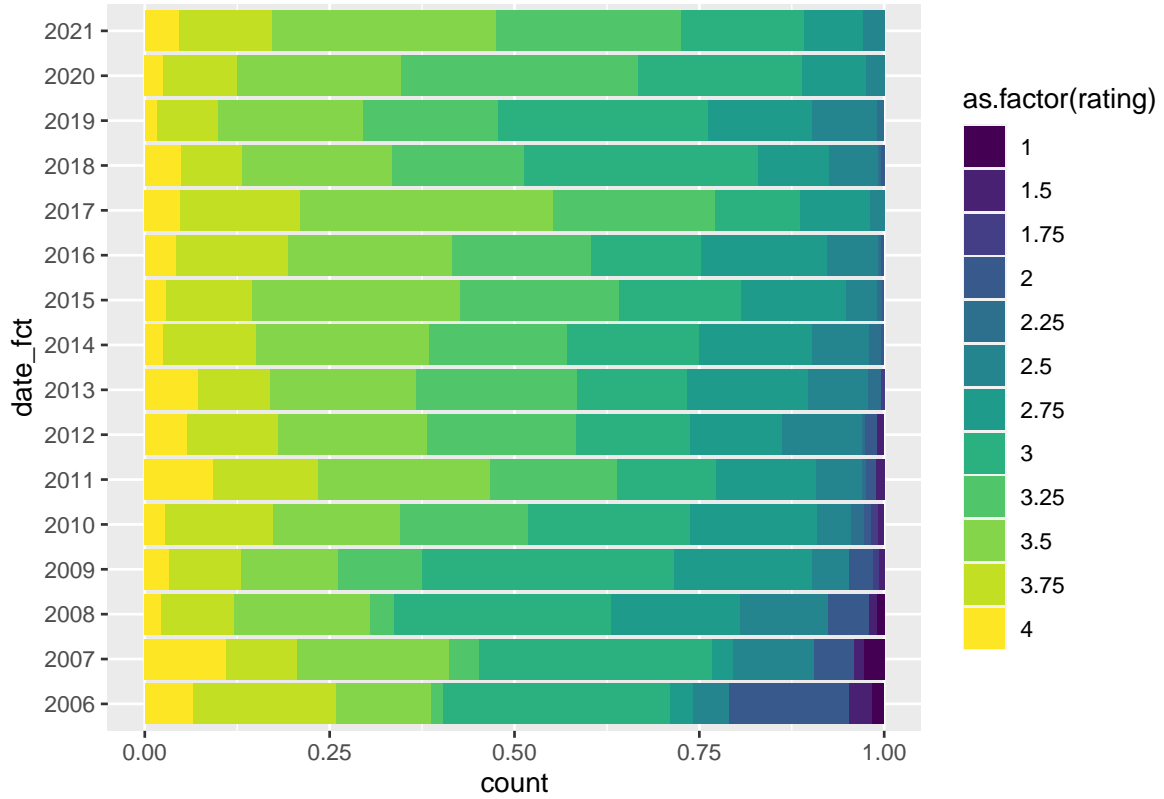


Table 1: Model 1

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



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.

Table 2: Model 2

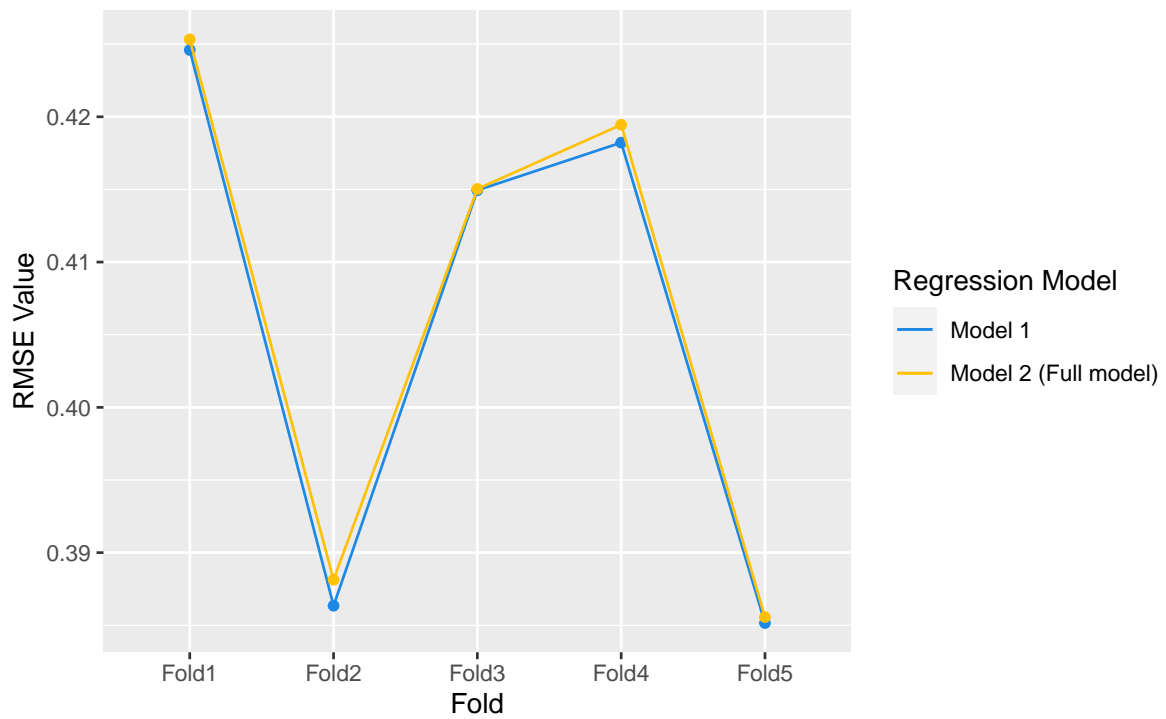
Fold	RMSE	R-squared
Fold1	0.425	0.114
Fold2	0.388	0.104
Fold3	0.415	0.139
Fold4	0.419	0.113
Fold5	0.386	0.099

Analysis approach

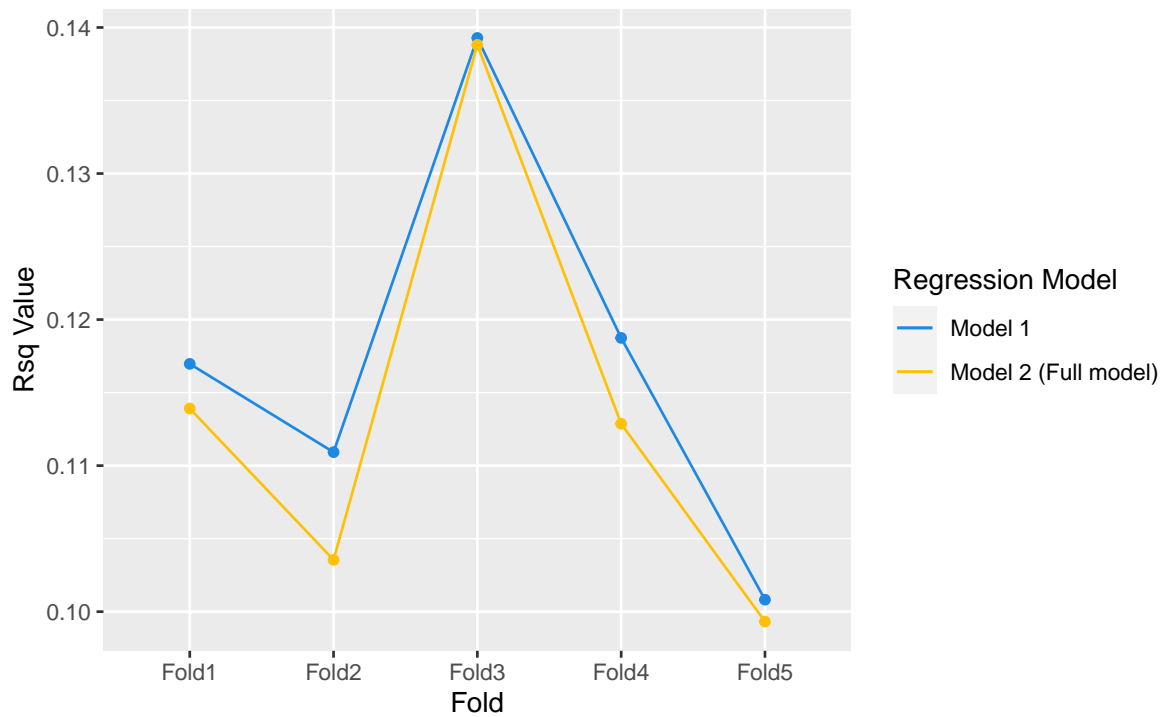
Ratings vs cocoa percent, ingredients, most memorable characteristics

All predictors

Visualization of RMSE for each Cross Validation Fold
Separated by Model



Visualization of R-squared for each Cross Validation Metric
Separated by Model



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](#).