Chocolate EDA

Manvir Kohli, Julie Song, Kelvin Wong

2022-11-18

Summary of the Data Set

The data set is provided by the Manhattan Chocolate Society, and was found and retrieved from the tidytuesday data project, specifically through this link. The data set contains observations for different types of dark chocolate, including the manufacturing company, origin of the cocoa beans used to make the chocolate, the other ingredients in the chocolate, and the amount of cocoa. They have also provided a feature column that contains descriptive words relating to the characteristics of the chocolate flavor, and a final rating.

We have split the original data set into training and testing data sets using a 70%-30% split. The following data processing and EDA analysis have been performed on the training set, which contains originally contains 1771 observations with 9 features and 1 target. After data processing and analysis, we have 7 features that we will use for modelling.

Glimpsing the Data

We have 1771 observations with 9 features and 1 target. After checking the structure and summary statistics for our data, we find the following:

- Our target variable is `rating`
- The columns `ref` and `specific_bean_origin_or_bar_name` are identifier columns and should be dropped
- The columns `company_manufacturer`, `company_location`, and `country_of_bean_origin`, are all read as character columns but should ideally be factors (i.e. categorical columns)
- `most_memorable_characteristics` is likely a text column, containing many unique words
- `cocoa_percent` is read as a character column while it should be numeric

Data Processing

We need to convert all the columns to the correct data types, but we will do this as the last step in our data processing.

The ingredients column has two components in each cell - the number of ingredients and the actual ingredients. So we can split this column into two and have two separate features. Now we have an overall total of 8 features with 1 target. We also checked our columns for null values, and found that there were 55 observations with missing values in our `ingredients` and `num_of_ingredients` columns.

Exploring Categorical Columns Further: For all the factors there are many levels. We can reduce the number of levels for different factors as follows :

- For `company_location` keep only the top 10 locations and combine all other locations into "Other"
- For `country_of_bean_origin` keep only the top 25 countries and combine all other into "Other"
- For `ingredients` keep the top 5 ingredients and combine all other into "Other"
- For `company_manufacturer`, it seems that there many distinct companies that manufacture chocolate in this data set, such that this feature acts more like an identifier. We choose to drop this feature

column, as the values are too unique, and we would have an overwhelming `Other` category even if we considered the top 50 companies.

Converting Data Types: Now we can convert our character columns into factors and also convert cocoa_percent column into a numeric column. The first few rows of our final, processed training data set is shown in Table 1.

Table 1: Final Features and Target in the Chocolate Dataset

Company Location	Review Date	Country of Bean Origin	Amount of Cocoa (%)	Number of Ingredients	Ingredients Present	Most Memorable Characteristics	Rating (1-5)
U.S.A.	2019	Dominican Republic	0.76	3	B,S,C	cocoa, vegetal, savory	3.50
U.S.A.	2019	Madagascar	0.76	3	B,S,C	cocoa, blackberry, full body	3.75
U.S.A.	2021	Other	0.68	3	B,S,C	chewy, off, rubbery	3.00
U.S.A.	2021	Other	0.80	3	B,S,C	mildly bitter, basic cocoa, fatty	3.25
U.S.A.	2021	India	0.68	3	B,S,C	milk brownie, macadamia,chewy	3.50
France	2012	Bolivia	0.70	4	B,S,C,L	vegetal, nutty	3.50

Data Distributions

Now let us examine the distributions for each of our numerical and categorical features.

Numerical and Discrete Features The only numerical feature we have is `percent_cocoa`. The `num_of_ingredients` and `review_date` features are discrete, and our target `rating` column is also discrete, as it has values between 1 and 5 in 0.25 intervals. Figure 1 shows the distributions for these features.

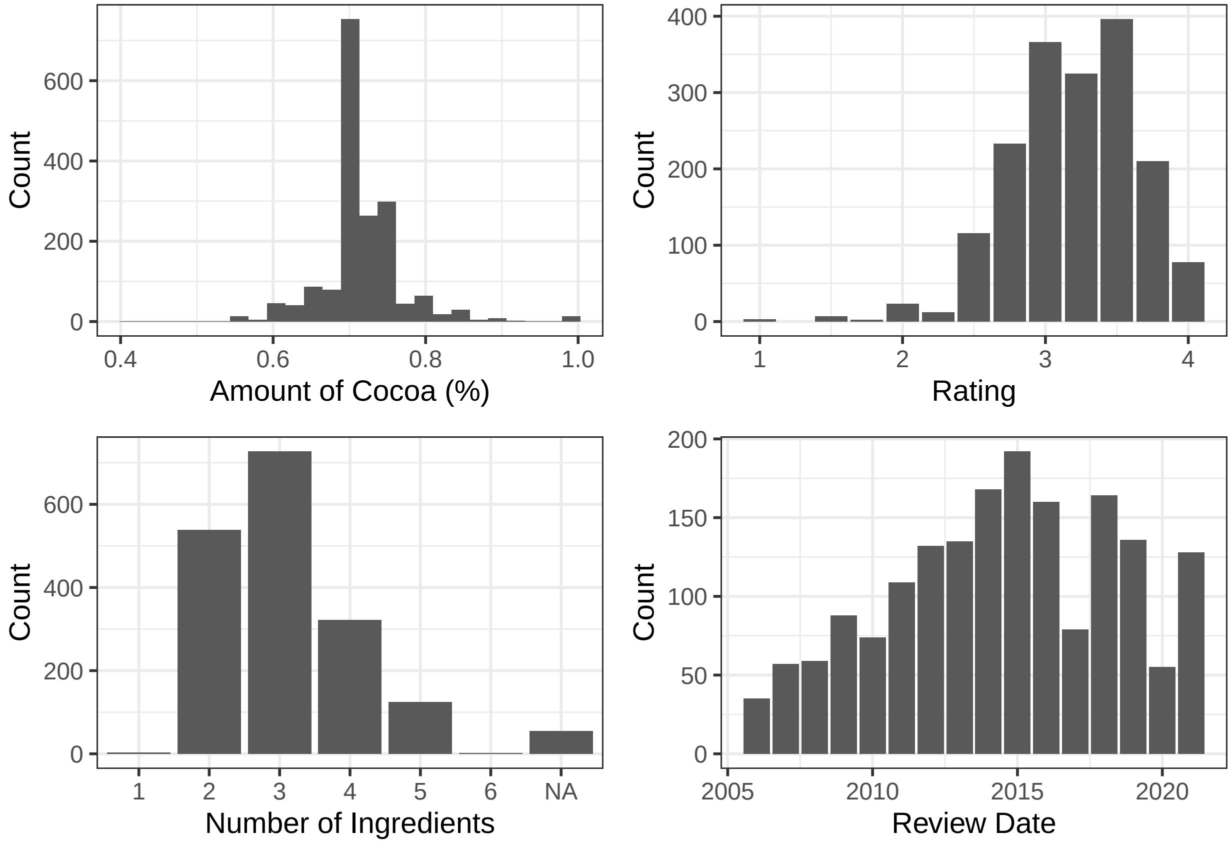

Figure 1, Distributions for numerical and discrete features in the training data set.

Categorical Features The `company_manufacturer`, `company_location`, `country_of_bean_origin`, and `ingredients` features are all categorical features. Figure 2 shows these feature distributions.

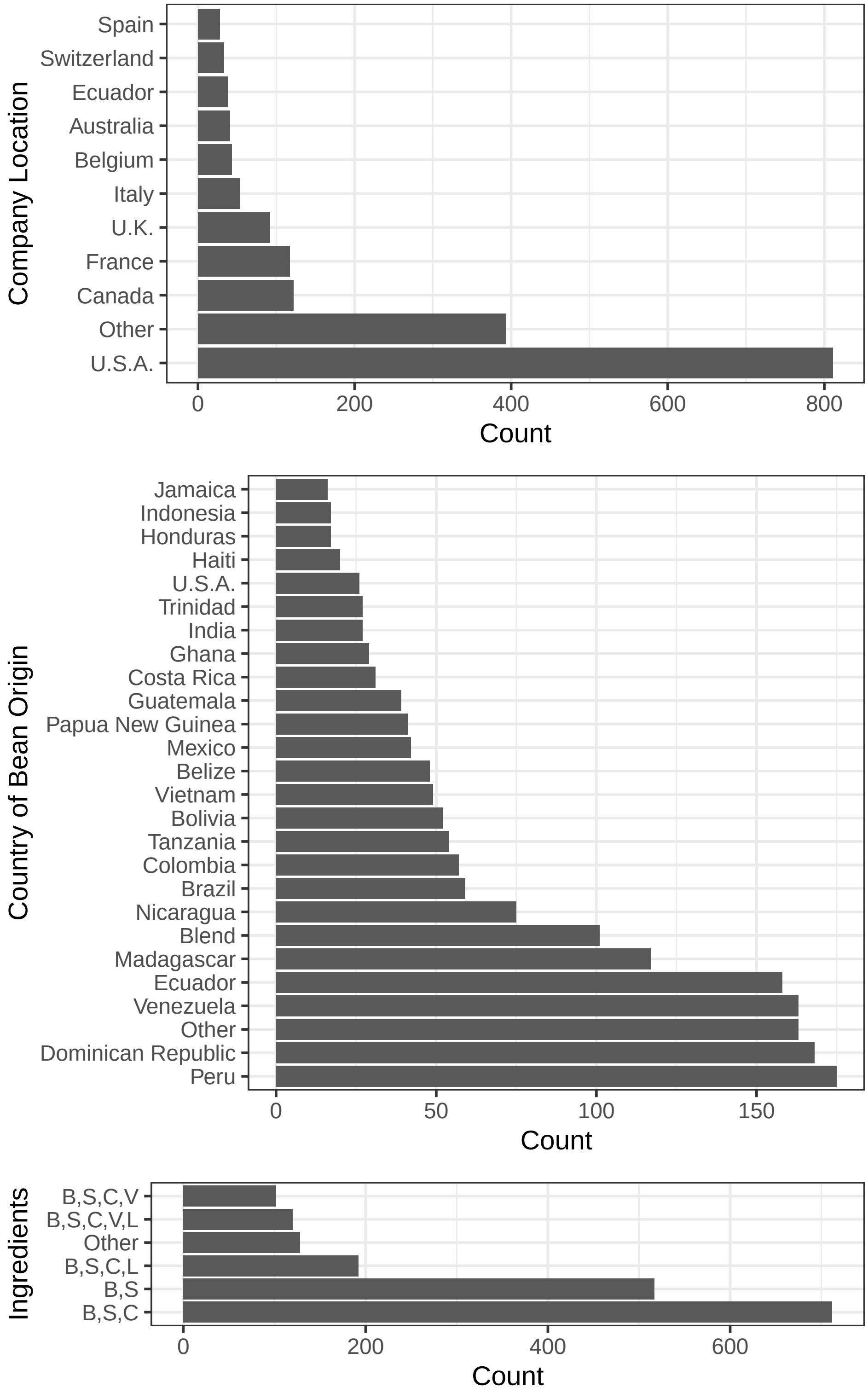

Figure 2, Distributions for categorical features in the training data set.