Chocolate EDA

Manvir Kohli, Julie Song, Kelvin Wong

2022-11-18

```
## Warning: package 'kableExtra' was built under R version 4.2.2
```

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, company location, origin of the cocoa beans used to make the chocolate, the other ingredients in the chocolate, the amount of cocoa in the chocolate, and others. 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

```
## Rows: 1,771
## Columns: 10
## $ ref                              <dbl> 2454, 2458, 2454, 2542, 2546, 2546, 2~
## $ company_manufacturer             <chr> "5150", "5150", "5150", "5150", "5150~
## $ company_location                 <chr> "U.S.A.", "U.S.A.", "U.S.A.", "U.S.A.~
## $ review_date                      <dbl> 2019, 2019, 2019, 2021, 2021, 2021, 2~
## $ country_of_bean_origin           <chr> "Tanzania", "Dominican Republic", "Ma~
## $ specific_bean_origin_or_bar_name <chr> "Kokoa Kamili, batch 1", "Zorzal, bat~
## $ cocoa_percent                    <chr> "76%", "76%", "76%", "68%", "72%", "8~
## $ ingredients                      <chr> "3- B,S,C", "3- B,S,C", "3- B,S,C", "~
## $ most_memorable_characteristics   <chr> "rich cocoa, fatty, bready", "cocoa, ~
## $ rating                           <dbl> 3.25, 3.50, 3.75, 3.00, 3.00, 3.25, 3~
```

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`, `country_of_bean_origin` and `ingredients` 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. We can split this column into two separate features (`num_of_ingredients` and `ingredients).`
- Note that `ingredients` uses the following symbols, as defined by Flavors of Cacao:
  - B = Beans
  - S = Sugar
  - S* = Sweetener other than white cane or beet sugar
  - C = Cocoa Butter
  - V = Vanilla
  - L = Lecithin
  - Sa = Salt
- Thus after dropping we `ref` and `specific_bean_origin_or_bar_name` 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.

Table 1: Null Count by Feature

Feature	Null Count
Manufacturing Company	0
Company Location	0
Review Date	0
Country of Bean Origin	0
Amount of Cocoa (%)	0
Number of Ingredients	66
Ingredients Present	66
Most Memorable Characteristics	0
Rating (1-5)	0

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`, keep the top 50 manufacturers and combine all other into "Other" (50 was chosen because this categorical feature has too many levels)

Converting Data Types: Now we can convert our character columns into factors and also convert cocoa_percent column into a numeric column. Below is the glimpse of our data after converting the column data types

```
## Rows: 1,771
## Columns: 9
## $ company_manufacturer           <fct> "5150", "5150", "5150", "5150", "5150",~
## $ company_location               <fct> U.S.A., U.S.A., U.S.A., U.S.A., U.S.A.,~
## $ review_date                    <dbl> 2019, 2019, 2019, 2021, 2021, 2021, 202~
## $ country_of_bean_origin         <fct> Tanzania, Dominican Republic, Madagasca~
## $ cocoa_percent                  <dbl> 0.76, 0.76, 0.76, 0.68, 0.72, 0.80, 0.6~
## $ num_of_ingredients             <chr> "3", "3", "3", "3", "3", "3", "3", "4",~
## $ ingredients                    <fct> " B,S,C", " B,S,C", " B,S,C", " B,S,C",~
## $ most_memorable_characteristics <chr> "rich cocoa, fatty, bready", "cocoa, ve~
## $ rating                         <dbl> 3.25, 3.50, 3.75, 3.00, 3.00, 3.25, 3.5~
```

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.

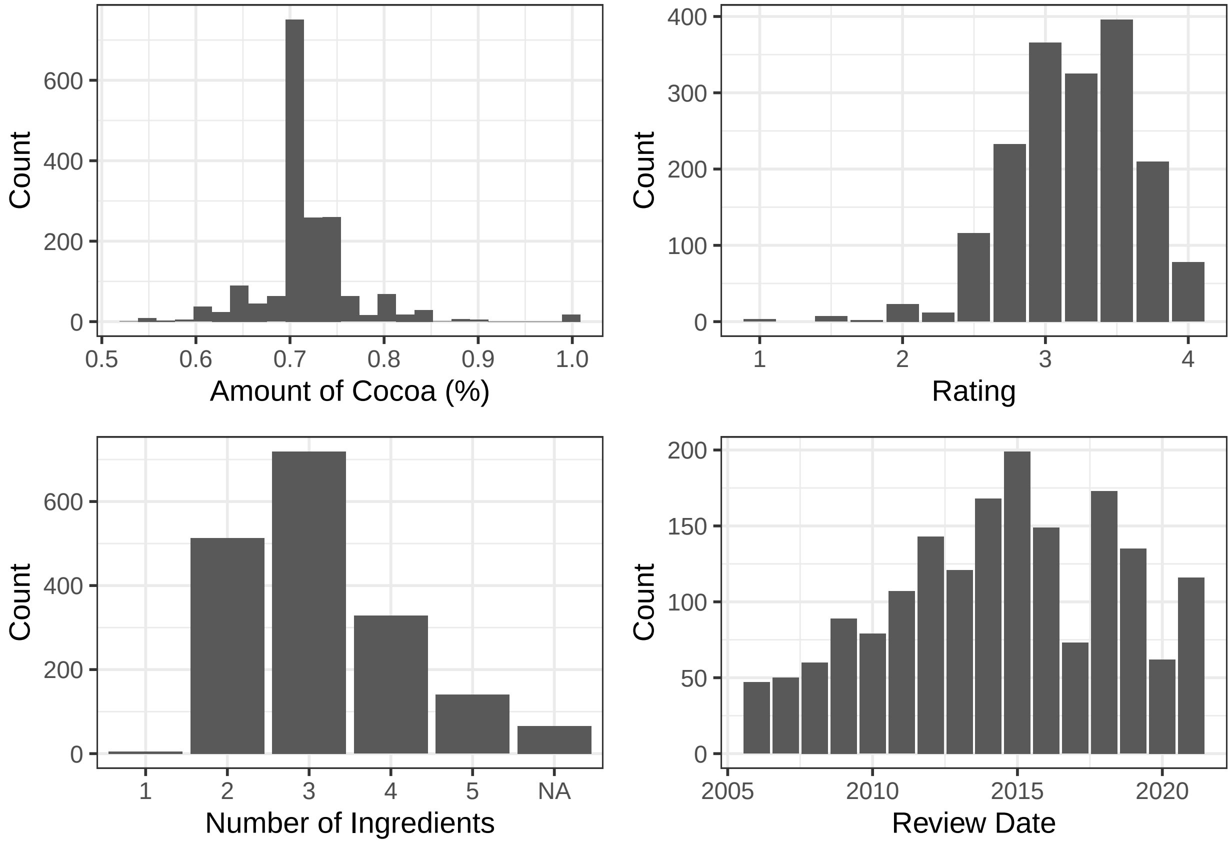

Figure 1: Numeric Plots

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

It seems that these numerical features are fairly well-distributed, and suitable for modelling.

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. ##

Based on the last plot above, it seems that there many distinct companies that manufacture chocolate in this data set, such that this feature acts more like an identifier. Therefore we can 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.

Table 2 shows the final features and their corresponding data types in our final data set, with "Rating" as our target:

Table 2: Final Features and Data Types

Feature	Type
Company Location	Factor
Review Date	Numeric - Continuous
Country of Bean Origin	Factor
Amount of Cocoa (%)	Numeric - Continuous
Number of Ingredients	Numeric - Discrete
Ingredients Present	Factor
Most Memorable Characteristics	Character(Text)

And Table 3 below shows the first 10 rows of our final processed training data set:

`{<- <-}`

References

The Manhattan Chocolate Society, 2022, "Chocolate Bar Ratings", Flavors of Cacao [Online]. Available: http://flavorsofcacao.com/chocolate_database.html

Thomas Mock (2022). Tidy Tuesday: A weekly data project aimed at the R ecosystem. https://github.com/rfordatascience/tidytuesday.