Chocolate EDA

Manvir Kohli, Julie Song, Kelvin Wong

2022-11-18

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

```
## 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
- The `ingredients` column now has each observation as a list of ingredients so we split the list so that each ingredient is a separate column so that we can perform binary one-hot encoding on each column when we build the models. Since each ingredient will be made into a column therefore we will have 7 additional columns instead of just 1.
- We also checked our columns for null values, and found that there were 66 observations with missing values in our `num_of_ingredients` columns.

Table 1: Null Count by Feature

Feature	Null Count
company_manufacturer	0
company_location	0
review_date	0
country_of_bean_origin	0
cocoa_percent	0
num_of_ingredients	66
most_memorable_characteristics	0
rating	0
beans	0
sugar	0
sweetener_other	0
cocoa_butter	0
vanilla	0
lecithin	0
salt	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 locations with atleast 20 observations and combine all locations with less than 20 observations into "Other"

- For `country_of_bean_origin` keep only the countries with atleast 10 observations and combine all other countries 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

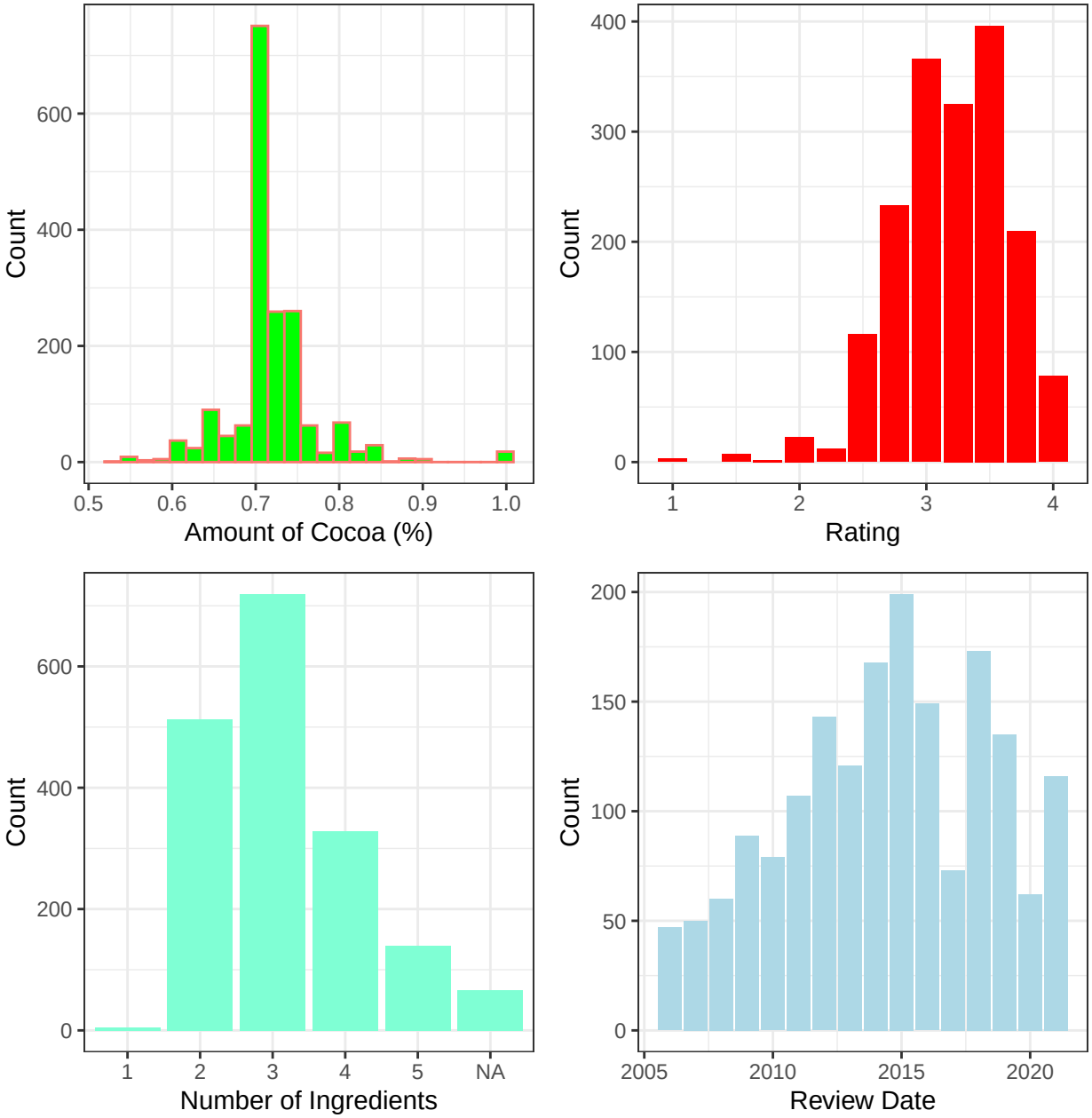
```
## Rows: 1,771
## Columns: 15
## $ company_manufacturer           <fct> "5150", "5150", "5150", "5150", "5150",~
## $ company_location               <fct> U.S.A., U.S.A., U.S.A., U.S.A., U.S.A.,~
## $ review_date                    <dbl> 2019, 2019, 2019, 2021, 2021, 2021, 202~
## $ country_of_bean_origin         <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",~
## $ 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~
## $ beans                          <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ sugar                          <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ sweetener_other                <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ cocoa_butter                   <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ vanilla                        <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ lecithin                       <fct> 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, ~
## $ salt                           <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

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.

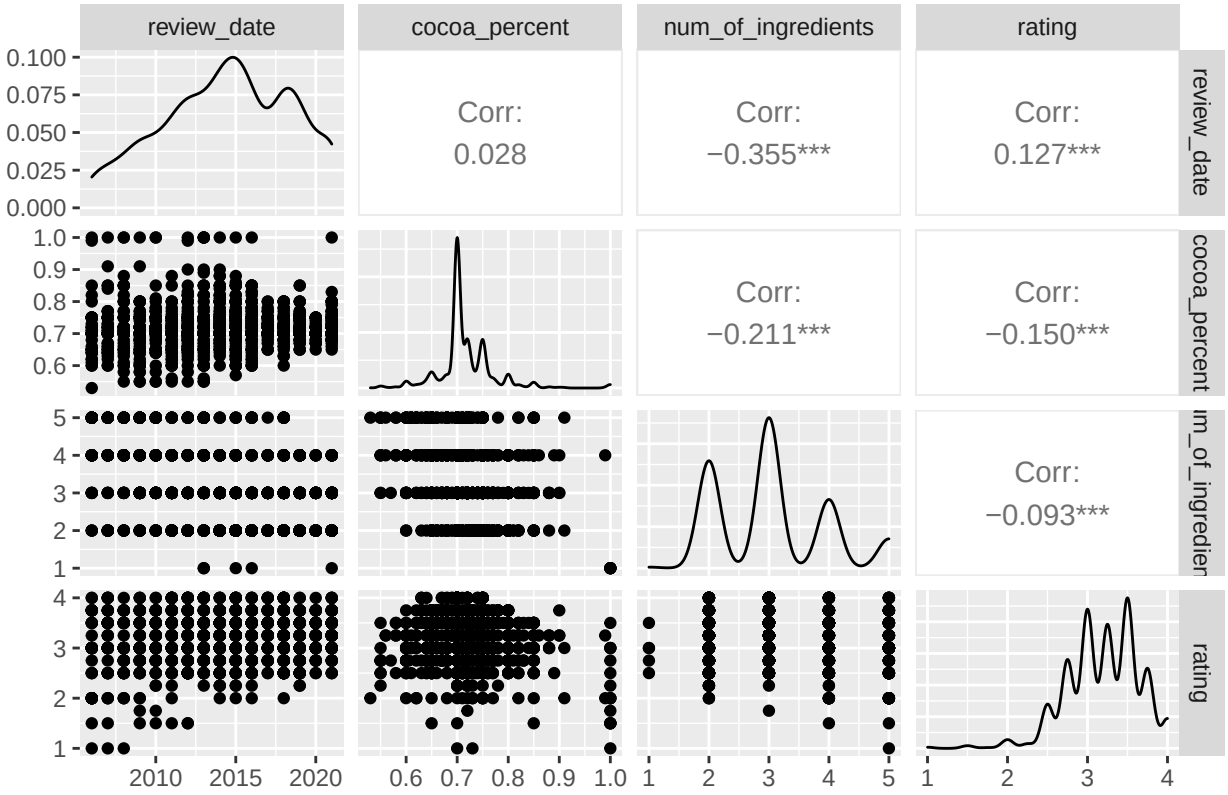
Figure 1: Numeric Plots




```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 66 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 66 rows containing missing values
```

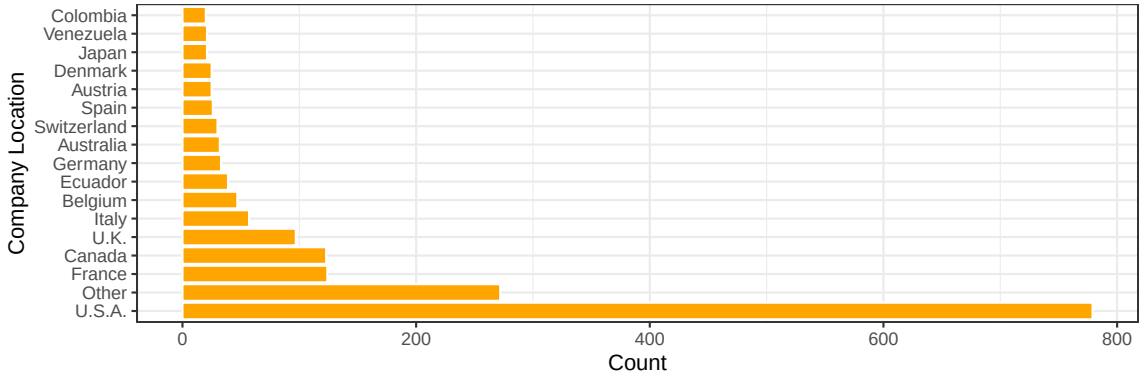
```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 66 rows containing missing values

## Warning: Removed 66 rows containing missing values (geom_point).
## Removed 66 rows containing missing values (geom_point).

## Warning: Removed 66 rows containing non-finite values (stat_density).

## Warning: Removed 66 rows containing missing values (geom_point).
```

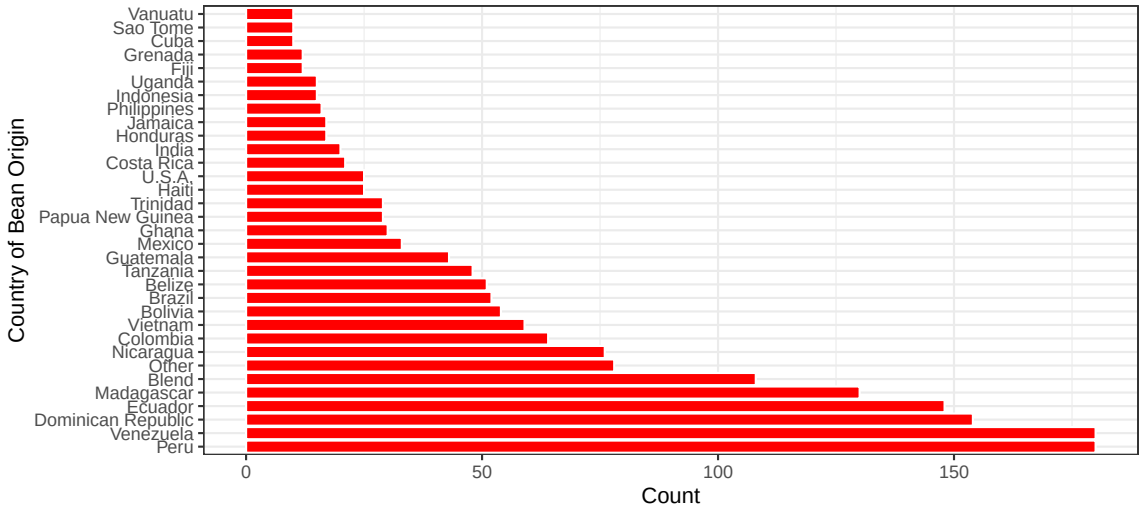
Figure 2: Pairwise Scatter Plots



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 3: Categorical Plots





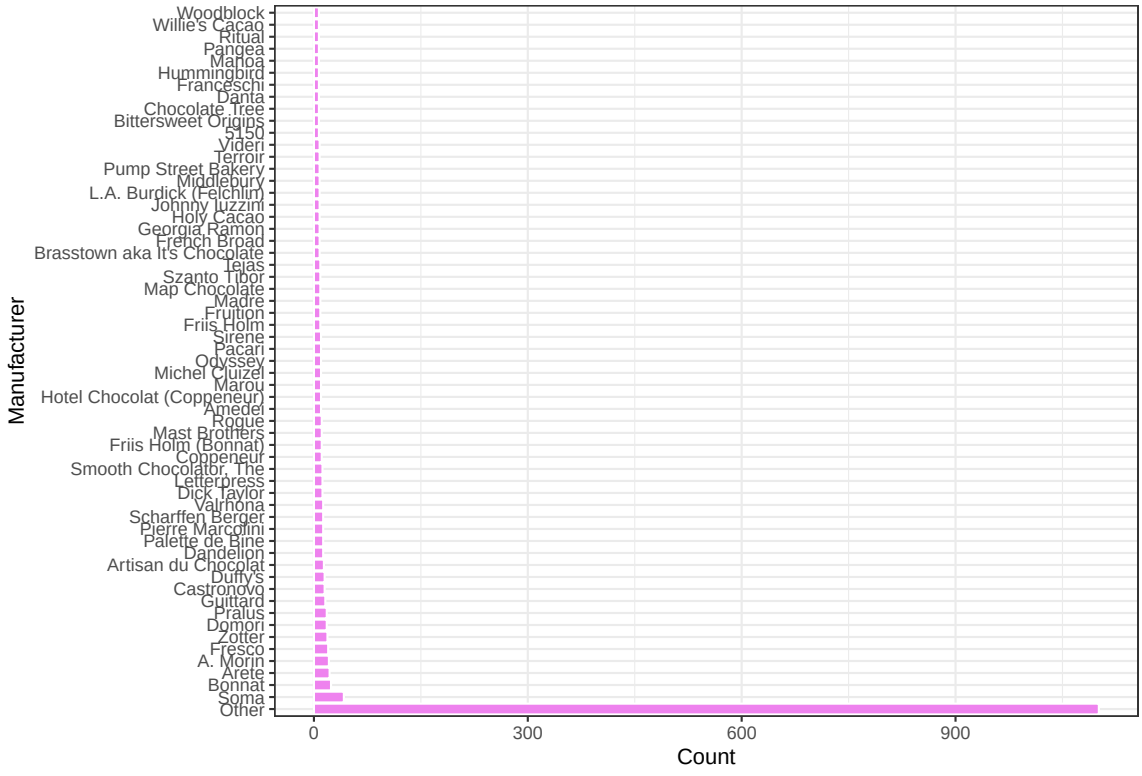
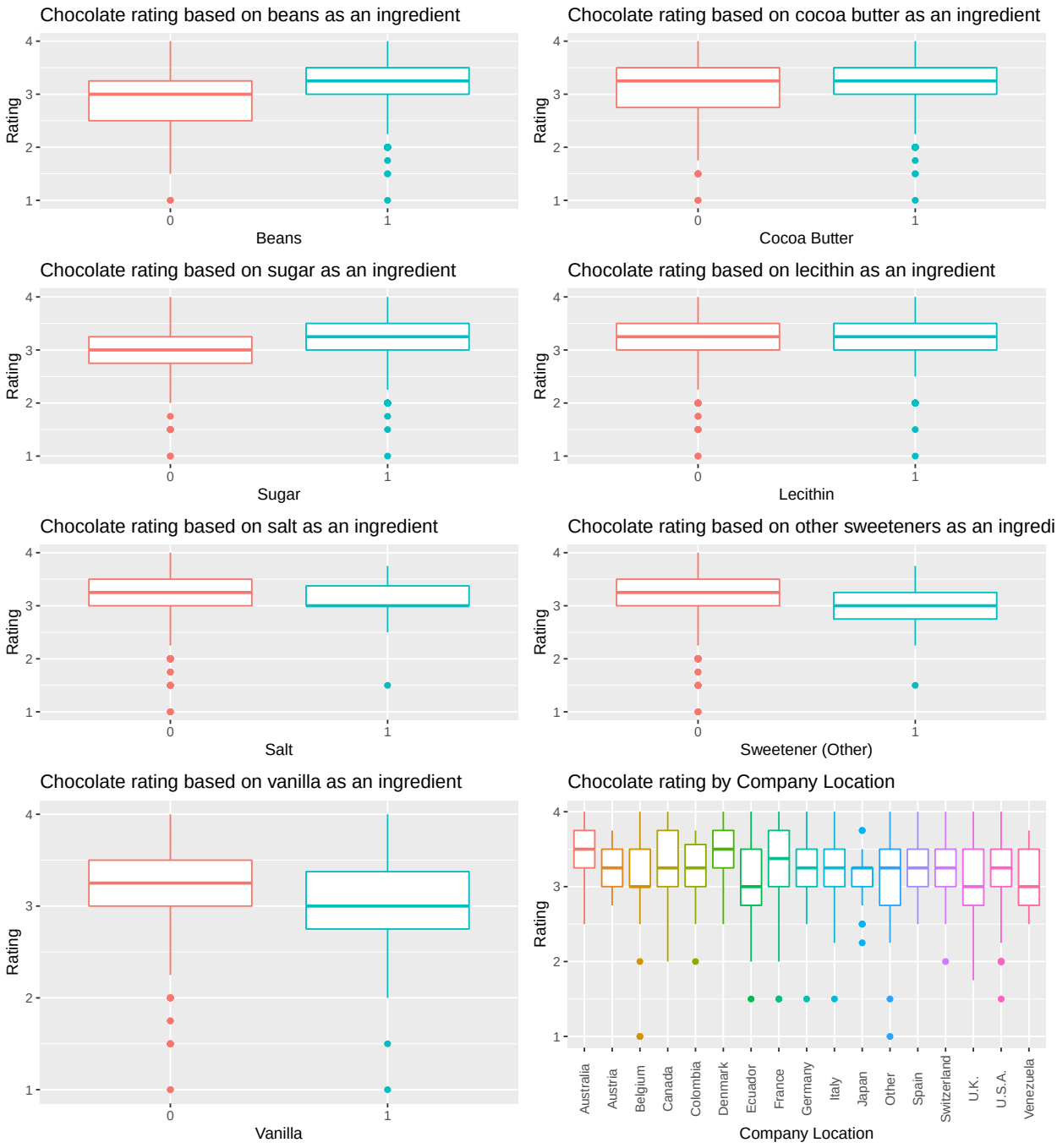


Figure 4: Boxplots by Rating



EDA Findings:

- The distributions of our numerical plots are not heavily skewed which suggests that our numerical features do not have a lot of outliers and hence are suitable for modeling
- The pairwise scatter plots show that there is no strong relationship between any of our numerical features.

- The categorical plots show that most of the companies manufacturing chocolates in our dataset are located in USA,. The top 3 countries where beans are sourced from are Peru, Venezuela and Dominican Republic.
- The last categorical plot showing the count of manufacturers suggests that there many distinct chocolate manufactures 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.
- From the boxplots we can infer that chocolates which have beans are rated better on average than chocolates without beans. The same is true for chocolates that have sugar as the sweetener. On the other hand chocolates that do not contain Vanilla or any other sweeteners are rated better which suggests that using Sugar as the sweetener is likely to result in the chocolate getting a better rating,
- The last boxlot showing chocolate rating by company location shows that chocolates produced by companies in Austria and Denmark tend to have higher ratings on average whereas chocolates produced by companies in Venezuela, Ecuador and the UK have lower ratings on average than the rest of the group

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
country_of_bean_origin	Factor
cocoa_percent	Numeric
num_of_ingredients	Numeric
most_memorable_characteristics	Character(Text)
beans	Factor (Binary)
sugar	Factor (Binary)
sweetener_other	Factor (Binary)
cocoa_butter	Factor (Binary)
vanilla	Factor (Binary)
lecithin	Factor (Binary)
salt	Factor (Binary)

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

Table 3: Preview of Final Dataset

company_location	review_date	country_of_bean_origin	cocoa_percent	num_of_ingredients	most_memorable_characteristics	rating	beans	sugar	sweetener_other	cocoa_butter	vanilla	lecithin	salt
U.S.A.	2019	Tanzania	0.76	3	rich cocoa, fatty, bready	3.25	1	1	0	1	0	0	0
U.S.A.	2019	Dominican Republic	0.76	3	cocoa, vegetal, savory	3.50	1	1	0	1	0	0	0
U.S.A.	2019	Madagascar	0.76	3	cocoa, blackberry, full body	3.75	1	1	0	1	0	0	0
U.S.A.	2021	Fiji	0.68	3	chewy, off, rubbery	3.00	1	1	0	1	0	0	0
U.S.A.	2021	Venezuela	0.72	3	fatty, earthy, moss, nutty,chalky	3.00	1	1	0	1	0	0	0
U.S.A.	2021	Uganda	0.80	3	mildly bitter, basic cocoa, fatty	3.25	1	1	0	1	0	0	0
U.S.A.	2021	India	0.68	3	milk brownie, macadamia,chewy	3.50	1	1	0	1	0	0	0
France	2012	Bolivia	0.70	4	vegetal, nutty	3.50	1	1	0	1	0	1	0
France	2012	Peru	0.63	4	fruity, melon, roasty	3.75	1	1	0	1	0	1	0
France	2013	Colombia	0.70	4	burnt rubber,alkalyzed notes	2.75	1	1	0	1	0	1	0

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.