Proposal

STA 210 - Project

Ginger and Stats - Aimy Wen, Rakshita Ramakrisna, Nathan Nguyen, Bryan Pan

```
library(tidyverse)
library(tidymodels)
library(kableExtra)

chocolate <- read_csv("../data/chocolate.csv")

glimpse(chocolate)
```

```
Rows: 2,530
Columns: 11
$ ...1                            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12~
$ 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~
```

Introduction

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.

Citations:

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

- 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.

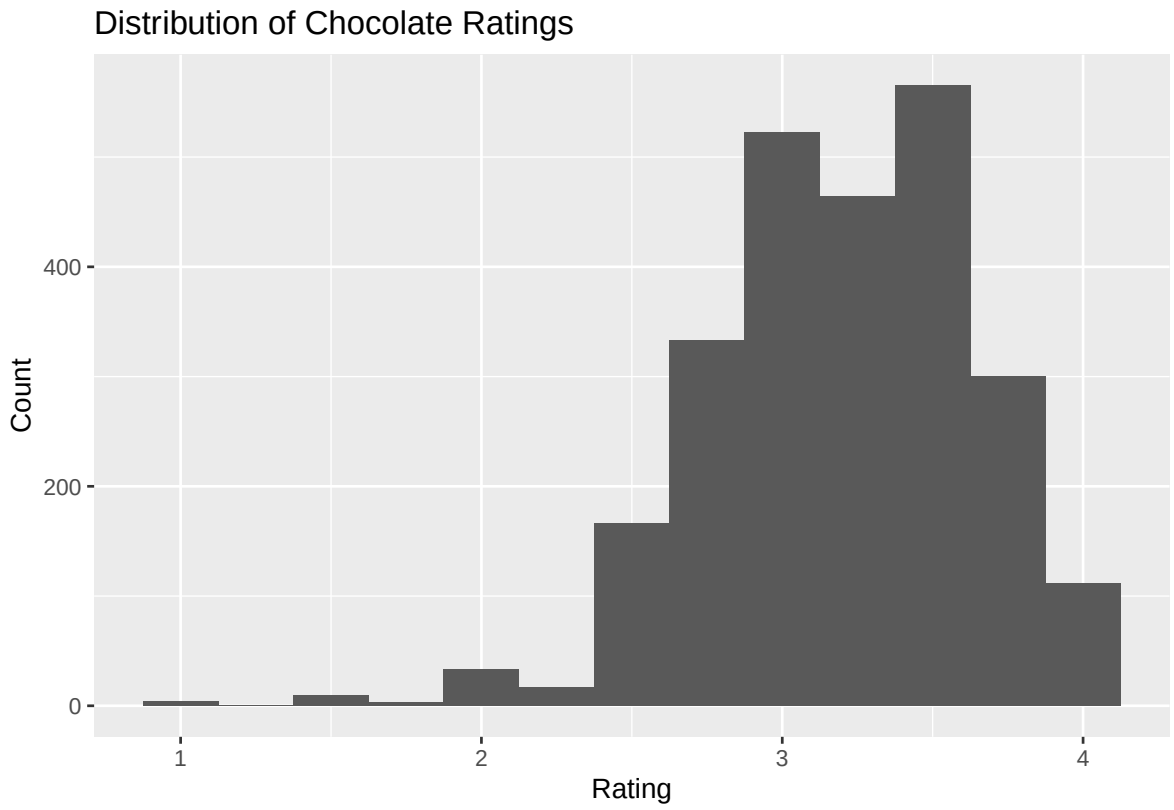
The data is collected by members of the Manhattan Chocolate Society reviewing chocolate bars using the rating system found at http://flavorsofcacao.com/review_guide.html and adding other characteristics about the bar itself.

...

Analysis approach

- Description of the response variable.
  - Our response variable is the ‘ rating ‘ variable. As a summary, the flavors of cacao rating scale is:
    * 4.0-5.0 = Outstanding
    * 3.5-3.9= Highly Recommended
    * 3.0-3.49= Recommended
    * 2.0-2.9= Disappointing
    * 1.0-1.9= Unpleasant
- Visualization and summary statistics for the response variable.

```r
ggplot(data = chocolate, mapping = aes(x = rating)) +
  geom_histogram(binwidth = .25) +
  labs(
  title = "Distribution of Chocolate Ratings",
  x = "Rating",
  y = "Count"
)
```




```r
summary_stats <- chocolate %>%
  summarize(
    mean_rating = mean(rating),
    sd_rating = sd(rating),
    median_rating = median(rating),
    iqr_rating = IQR(rating)
  )
```

```
summary_stats %>%
  kable()
```

mean_rating	sd_rating	median_rating	iqr_rating
3.196344	0.4453213	3.25	0.5

```
chocolate %>%
  count(rating)
```

```
# A tibble: 12 x 2
   rating     n
    <dbl> <int>
 1   1        4
 2   1.5     10
 3   1.75     3
 4   2       33
 5   2.25    17
 6   2.5    166
 7   2.75   333
 8   3      523
 9   3.25   464
10   3.5    565
11   3.75   300
12   4      112
```

- The distribution is left-skewed. Therefore, the median and IQR would be a better estimate of its center, (centered around a 3.25 rating) and spread (the distribution has a spread of 0.5 rating points).

  List of variables that will be considered as predictors:

  - `country_of_bean_origin`
    - It would be interesting to understand whether beans produced in Asia or South America tend to result in differing chocolate ratings
  - `company_location`
    - We think there could be an association between company locating and rating, specifically European based companies having higher chocolate ratings
    - We could mutate this variable to tell us the continent the company is located in.
  - `cocoa_percent`

    - Do the Manhattan Chocolate Society members favor milkier chocolates (with lower cocoa percentages) over darker chocolates? Due to evolution, humans tend to prefer milkier and sugarier food items, so it would be interesting to see if a lower cocoa percent is linked to a higher rating.
  - `ingredients`
    - Does a decrease in the number of ingredients translate to a higher chocolate rating?
  - We would also be interested in exploring the most memorable rating
    - it would be cool to see if for each rating if a specific memorable rating keyword comes up often
    - Is there a way to group roast and smoke in the same category? Or berry and fruity? Or peanut buttery and nutty?
- Regression model technique (multiple linear regression and logistic regression)
  - We plan to use multiple linear regression for our data analysis as our response variable that we are measuring (rating) is not binary, and since overall none of our variables in the dataset are binary.

Data dictionary

The data dictionary can be found here.

Explicitly, one can copy and paste the this link https://github.com/sta210-s22/project-ginger_and_stats/tree/main/data