Chocolate Exploratory Data Analysis

Ryan Heslin

July 15, 2024

Introduction

```
library(tidyverse)
source(here::here("R", "utils.R"))

chocolate <- readRDS(here::here("data", "chocolate.Rds"))
theme_set(theme_minimal())
```

The data for this project come from week 3 of 2022 Tidy Tuesday, available here. They consist of ratings for `nrow(chocolate)` chocolate bars. Useful features include country of origin, manufacturing company, ingredients, and words used by taste-testers to describe each bar. (How do I get that job?) I have preprocessed the data by extracting these descriptive words into columns and creating a dummy column for each ingredient.

The Response

Chocolate bar quality is rated on a scale from 1 to 4 inclusive, with increments of 0.25. The mean is 3.1963439. I will have to round predictions to that scale, and I may have to create a custom loss function to accommodate it.

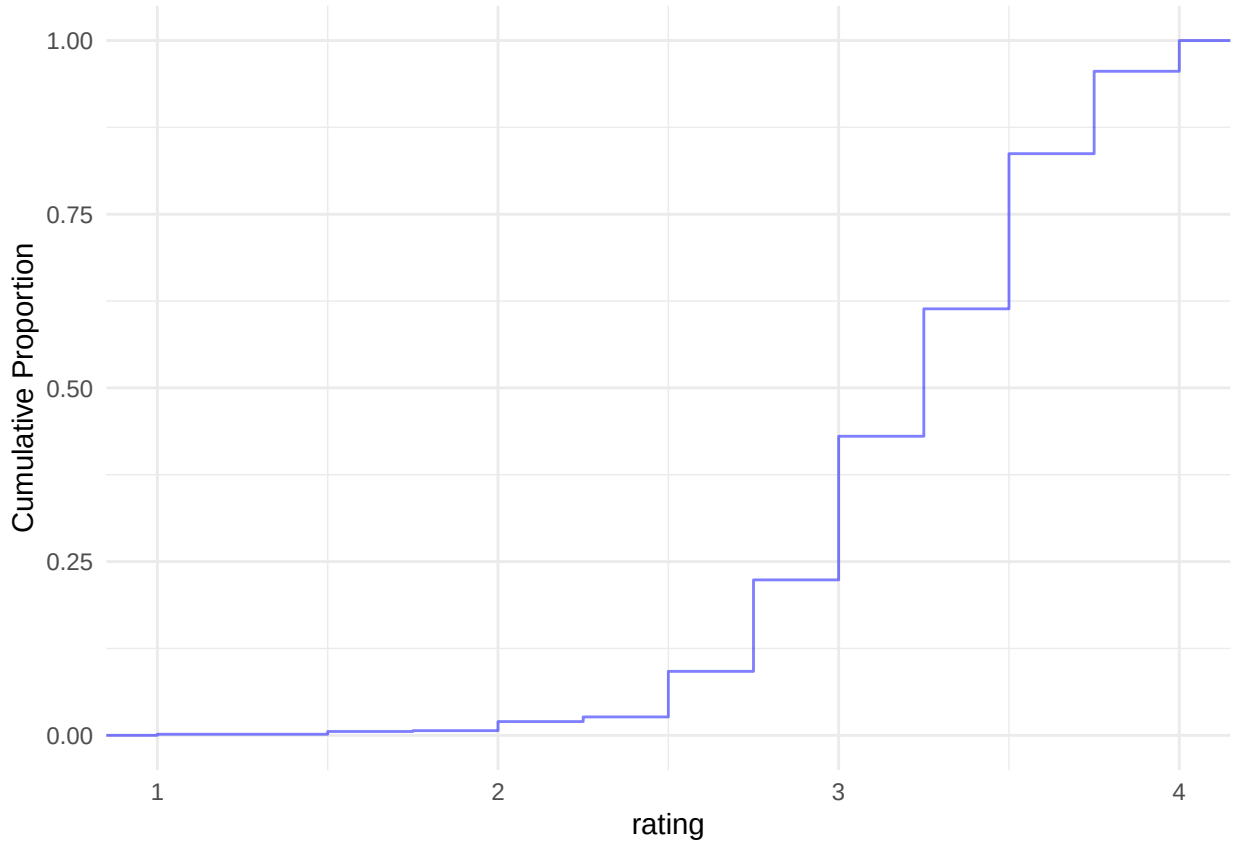
```
ratings <- vec2DF(table(chocolate[["rating"]]))
ratings
```

1	1.5	1.75	2	2.25	2.5	2.75	3	3.25	3.5	3.75	4
4	10	3	33	17	166	333	523	464	565	300	112

```
write_csv(ratings, here::here("outputs", "ratings.csv"))
```

The empirical cumulative distribution of ratings.

```
ggplot(chocolate, aes(x = rating)) +
  stat_ecdf(geom = "step", color = "blue", alpha = .5) +
  labs(y = "Cumulative Proportion")
```



Countries of Origin

The vast majority of companies are American, which may reflect bias in the data.

The cacao beans used to make the chocolate bars come from just six countries.

```
table_head(chocolate[["company_location"]])
```

```
 U.S.A.  Canada  France    U.K.   Italy Belgium
   1136     177     176     133      78      63
```

```
table_head(chocolate[["country_of_bean_origin"]]) |>
  vec2DF()
```

Venezuela	Peru	Dominican Republic	Ecuador	Madagascar	Blend
253	244	226	219	177	156

Flavor Description

I created five dummy columns to hold each word or phrase used to describe each bar. Many bars had fewer terms, so I filled the columns with `NA`.

Most of the words seem to describe bar flavors ("oily", "sweet", etc.); fewer convey direct sentiments. They may still be useful as predictors if the testers consistently preferred some combinations of ingredients to others.

The worst-rated bars have some amusing words.

```
chocolate <- select(chocolate, -characteristic5)

filter(chocolate, rating == 4) |>
  select(starts_with("characteristic")) |>
  head()
```

characteristic1	characteristic2	characteristic3	characteristic4
oily	nut	caramel	raspberry
sweet	cocoa	tangerine	NA
delicate	hazelnut	brownie	NA
light color	fruit	yogurt	NA
strong spice	intense pepper	NA	NA
tart	lemon	smoke	NA

```
filter(chocolate, rating == 1) |>
  select(starts_with("characteristic"))
```

characteristic1	characteristic2	characteristic3	characteristic4
bitter	cocoa	NA	NA
chalky	musty	very bitter	NA
this is not chocolate	NA	NA	NA
pastey	strong off flavor	NA	NA

```
all_words <- select(chocolate, starts_with("characteristic")) |>
  unlist() |>
  na.omit()

length(all_words)
```

```
[1] 7060
```

```
library(tidytext)

sum(all_words %in% (get_sentiments("bing")$word))
```

```
[1] 1357
```

Of the 7060 total words recorded, only a fraction appear in the Bing sentiment lexicon, though that's not too bad. Many are repeated, with only 971 distinct values.

The most common are generic flavor descriptors.

```
ingredients <- table_head(all_words, .n = 8) |>
  vec2DF()
ingredients
```

sweet	nutty	cocoa	roasty	earthy	creamy	sandy	fatty
273	261	252	213	190	188	170	166

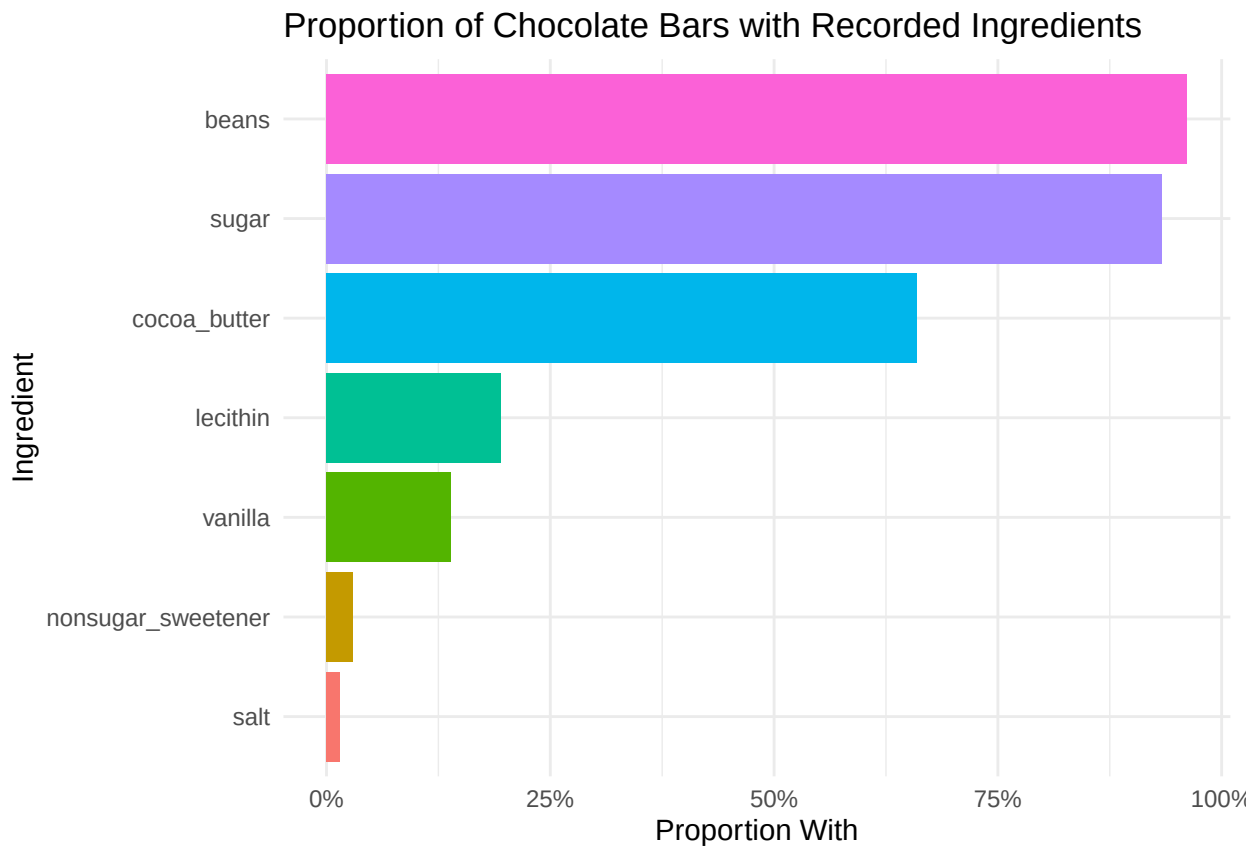
```
ingredient_words <- names(table_head(all_words, .n = 10))

write_csv(ingredients, here::here("outputs", "ingredients.csv"))
```

Ingredients

As we'd expect, almost all chocolate bars have cacao and sugar. The other ingredients are much rarer.

```
ingredients_bar <- chocolate |>
  select(beans:last_col()) |>
  colMeans() |>
  vec2DF() |>
  pivot_longer(everything(),
    names_to = "Ingredient",
    values_to = "Proportion With"
```

```
  ) |>
  mutate(Ingredient = factor(Ingredient, levels = unique(Ingredient)[order(`Proportion With`)])) |>
  ggplot(aes(y = Ingredient, x = `Proportion With`, fill = Ingredient)) +
  geom_col() +
  scale_x_continuous(labels = scales::percent) +
  labs(title = "Proportion of Chocolate Bars with Recorded Ingredients") +
  theme(legend.position = "none")
ingredients_bar
```

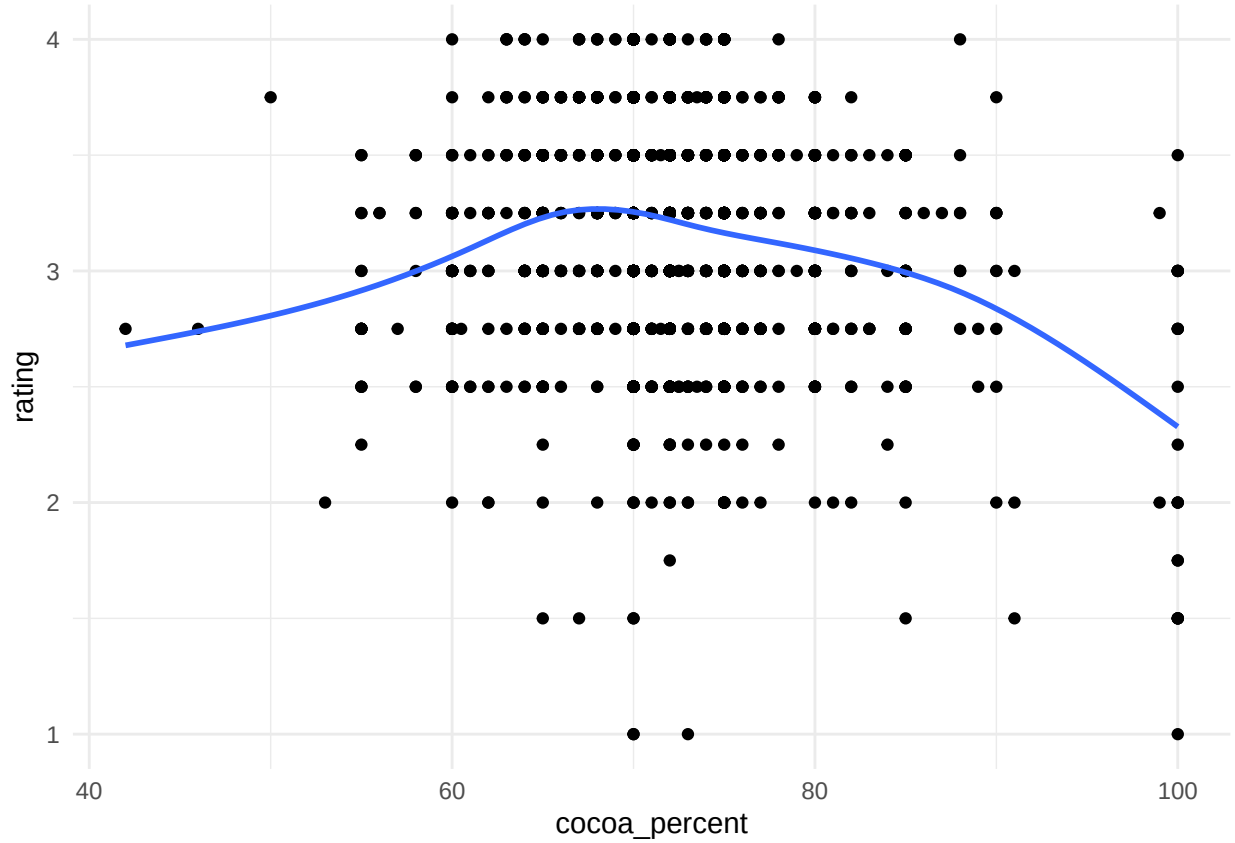



```
ggsave(here::here("outputs", "ingredients-bar.png"), ingredients_bar)
```

Cocoa Percentage

There doesn't seem to be a linear relationship between cocoa percentage and ratings.

```
ggplot(chocolate, aes(x = cocoa_percent, y = rating)) +
  geom_point() +
  geom_smooth(se = FALSE)
```



Conclusion

`country_of_bean_origin` and `company_location` would be useful features as is. It will take more work to wring usable data out of ingredients and flavor terms. Text could be transformed into sentiments, but that would not be useful for words describing flavors or ingredients.