Chocolate Ratings

Eszter Ari

3/18/2022

Input data

Download and save the raw data by using the `tidyverse` `R` package.

```
library(tidyverse)
library(magrittr)
chocolate <- read_csv(
  "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2022/2022-01-18/chocolate.
chocolate %>%
  write_csv(here::here("data/raw/chocolate.csv"))
```

Let's check the content.

```
chocolate %>%
  slice(1:5) %>%
  mutate_all(~paste('\\scriptsize', .)) %>%
  setNames(paste('\\scriptsize', names(.))) %>%
  pander(caption = "Head of the binding site table")
```

Table 1: Head of the binding site table (continued below)

ref	company__manufacturer	company__location	review__date	country__of__bean__origin
2454	5150	U.S.A.	2019	Tanzania
2458	5150	U.S.A.	2019	Dominican Republic
2454	5150	U.S.A.	2019	Madagascar
2542	5150	U.S.A.	2021	Fiji
2546	5150	U.S.A.	2021	Venezuela

specific__bean__origin__or__bar__name	cocoa__percent	ingredients	most__memorable__characteristics	rating
Kokoa Kamili, batch 1	76%	3- B,S,C	rich cocoa, fatty, bready	3.25
Zorzal, batch 1	76%	3- B,S,C	cocoa, vegetal, savory	3.5
Bejofo Estate, batch 1	76%	3- B,S,C	cocoa, blackberry, full body	3.75
Matasawalevu, batch 1	68%	3- B,S,C	chewy, off, rubbery	3
Sur del Lago, batch 1	72%	3- B,S,C	fatty, earthy, moss, nutty,chalky	3

```
str(chocolate)
```

```
## spec_tbl_df [2,530 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ref                            : num [1:2530] 2454 2458 2454 2542 2546 ...
##  $ company_manufacturer           : chr [1:2530] "5150" "5150" "5150" "5150" ...
##  $ company_location               : chr [1:2530] "U.S.A." "U.S.A." "U.S.A." "U.S.A." ...
##  $ review_date                    : num [1:2530] 2019 2019 2019 2021 2021 ...
##  $ country_of_bean_origin         : chr [1:2530] "Tanzania" "Dominican Republic" "Madagascar" "Fiji"
##  $ specific_bean_origin_or_bar_name: chr [1:2530] "Kokoa Kamili, batch 1" "Zorzal, batch 1" "Bejofo E
##  $ cocoa_percent                  : chr [1:2530] "76%" "76%" "76%" "68%" ...
##  $ ingredients                    : chr [1:2530] "3- B,S,C" "3- B,S,C" "3- B,S,C" "3- B,S,C" ...
##  $ most_memorable_characteristics  : chr [1:2530] "rich cocoa, fatty, bready" "cocoa, vegetal, savory
##  $ rating                         : num [1:2530] 3.25 3.5 3.75 3 3 3.25 3.5 3.5 3.75 2.75 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ref = col_double(),
##   ..   company_manufacturer = col_character(),
##   ..   company_location = col_character(),
##   ..   review_date = col_double(),
##   ..   country_of_bean_origin = col_character(),
##   ..   specific_bean_origin_or_bar_name = col_character(),
##   ..   cocoa_percent = col_character(),
##   ..   ingredients = col_character(),
##   ..   most_memorable_characteristics = col_character(),
##   ..   rating = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Let's see if the "cocoa_percent" and the "rating" correlate each other.

Converting "cocoa_percent" to numbers.

```
chocolate %<>%
  mutate(cocoa_percent = str_remove_all(cocoa_percent, "%")) %>%
  mutate(cocoa_percent = as.numeric(cocoa_percent))
```

Pearson correlation:

```
cor(chocolate$cocoa_percent, chocolate$rating)
```

```
## [1] -0.1466896
```

Linear regression:

```
lm(rating ~ cocoa_percent, data = chocolate) %>%
  summary()
```

```
##
## Call:
## lm(formula = rating ~ cocoa_percent, data = chocolate)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.21541 -0.23867  0.03459  0.28459  0.99393
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.02953    0.11209  35.949  < 2e-16 ***
## cocoa_percent -0.01163    0.00156  -7.456 1.22e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4406 on 2528 degrees of freedom
## Multiple R-squared:  0.02152,    Adjusted R-squared:  0.02113
## F-statistic: 55.59 on 1 and 2528 DF,  p-value: 1.218e-13
```

```
chocolate %>%
  ggplot(aes(x = cocoa_percent, y = rating)) +
  geom_smooth(method = lm, formula = y ~ x, color = "steelblue") + # regression
  geom_point(color = "coral", size = 3)
```