Chocolate Rating Analysis

Rosie Bai

11/22/2020

Data Preprocessing

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
flavors_of_cacao <- read.csv('https://raw.githubusercontent.com/rouzi612/R-weekly-project/main/flavors_
names(flavors_of_cacao)
```

```
## [1] "CompanyÂ...Maker.if.known."       "Specific.Bean.Origin.or.Bar.Name"
## [3] "REF"                              "Review.Date"
## [5] "Cocoa.Percent"                    "Company.Location"
## [7] "Rating"                           "Bean.Type"
## [9] "Broad.Bean.Origin"
```

```
names(flavors_of_cacao)[1] = "maker"
names(flavors_of_cacao)[2] = "Specific.Bean.Origin"
str(flavors_of_cacao)
```

```
## 'data.frame':    1795 obs. of  9 variables:
##  $ maker               : Factor w/ 416 levels "A. Morin","Acalli",..: 1 1 1 1 1 1 1 1 1 ...
##  $ Specific.Bean.Origin: Factor w/ 1039 levels "\"heirloom\", Arriba Nacional",..: 15 494 68 16 813
##  $ REF                 : int  1876 1676 1676 1680 1704 1315 1315 1315 1319 1319 ...
##  $ Review.Date         : int  2016 2015 2015 2015 2015 2014 2014 2014 2014 2014 ...
##  $ Cocoa.Percent       : Factor w/ 45 levels "100%","42%","46%",..: 14 21 21 21 21 21 21 21 21 21 ..
##  $ Company.Location    : Factor w/ 60 levels "Amsterdam","Argentina",..: 19 19 19 19 19 19 19 19 19
##  $ Rating              : num  3.75 2.75 3 3.5 3.5 2.75 3.5 3.5 3.75 4 ...
##  $ Bean.Type           : Factor w/ 42 levels "","Â ","Amazon",..: 2 2 2 2 2 10 2 10 10 2 ...
##  $ Broad.Bean.Origin   : Factor w/ 101 levels "","Â ","Africa, Carribean, C. Am.",..: 70 80 80 80 57
```

```
flavors_of_cacao$Cocoa.Percent<-as.numeric(sub("%", "",flavors_of_cacao$Cocoa.Percent, fixed = TRUE))/1
summary(flavors_of_cacao$Cocoa.Percent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.420   0.700   0.700   0.717   0.750   1.000
```

Average Rating by Broad Bean Origin

```
# Average Rating by Broad Bean Origin
flavors_of_cacao %>%
  group_by(`Broad.Bean.Origin`) %>%
  summarise(`Rating` = mean(`Rating`)) %>%
  arrange(-`Rating`)%>%
  top_n(20)
```

```
## Selecting by Rating
```

```
## # A tibble: 20 x 2
##    Broad.Bean.Origin            Rating
##    <fct>                         <dbl>
##  1 Dom. Rep., Madagascar          4
##  2 Gre., PNG, Haw., Haiti, Mad    4
##  3 Guat., D.R., Peru, Mad., PNG   4
##  4 Peru, Dom. Rep                 4
##  5 Ven, Bolivia, D.R.             4
##  6 Venezuela, Java                4
##  7 Dominican Rep., Bali           3.75
##  8 DR, Ecuador, Peru              3.75
##  9 Peru, Belize                   3.75
## 10 PNG, Vanuatu, Mad              3.75
## 11 Ven.,Ecu.,Peru,Nic.            3.75
## 12 Venez,Africa,Brasil,Peru,Mex   3.75
## 13 South America                  3.67
## 14 Tobago                         3.62
## 15 Indonesia, Ghana               3.5
## 16 Mad., Java, PNG                3.5
## 17 Peru, Mad., Dom. Rep.          3.5
## 18 Trinidad, Ecuador              3.5
## 19 Ven., Indonesia, Ecuad.        3.5
## 20 Venezuela/ Ghana               3.5
```

Average Rating by Specific Bean Origin

```
flavors_of_cacao %>%
  group_by(`Specific.Bean.Origin`) %>%
  summarise(`Rating` = mean(`Rating`)) %>%
  arrange(-`Rating`) %>%
  top_n(20)
```

```
## Selecting by Rating
```

```
## # A tibble: 65 x 2
##    Specific.Bean.Origin                        Rating
##    <fct>                                        <dbl>
##  1 Toscano Black                                 4.17
```

```
##  2 ABOCFA Coop                                  4
##  3 Alto Beni, Cru Savage                        4
##  4 Asante                                       4
##  5 Bali, Sukrama Bros. Farm, Melaya, 62hr C     4
##  6 Bellavista Coop, #225, LR, MC, CG Exclusive  4
##  7 Cabosse                                      4
##  8 Carenero Superior, Urrutia, Barlovento       4
##  9 Chuao, #217, DR, MC                          4
## 10 Claudio Corallo w/ nibs                      4
## # ... with 55 more rows
```

Average Rating by Country

```
flavors_of_cacao %>%
  group_by(`Company.Location`) %>%
  summarise(`Rating` = mean(`Rating`)) %>%
  arrange(-`Rating`) %>%
  top_n(10)
```

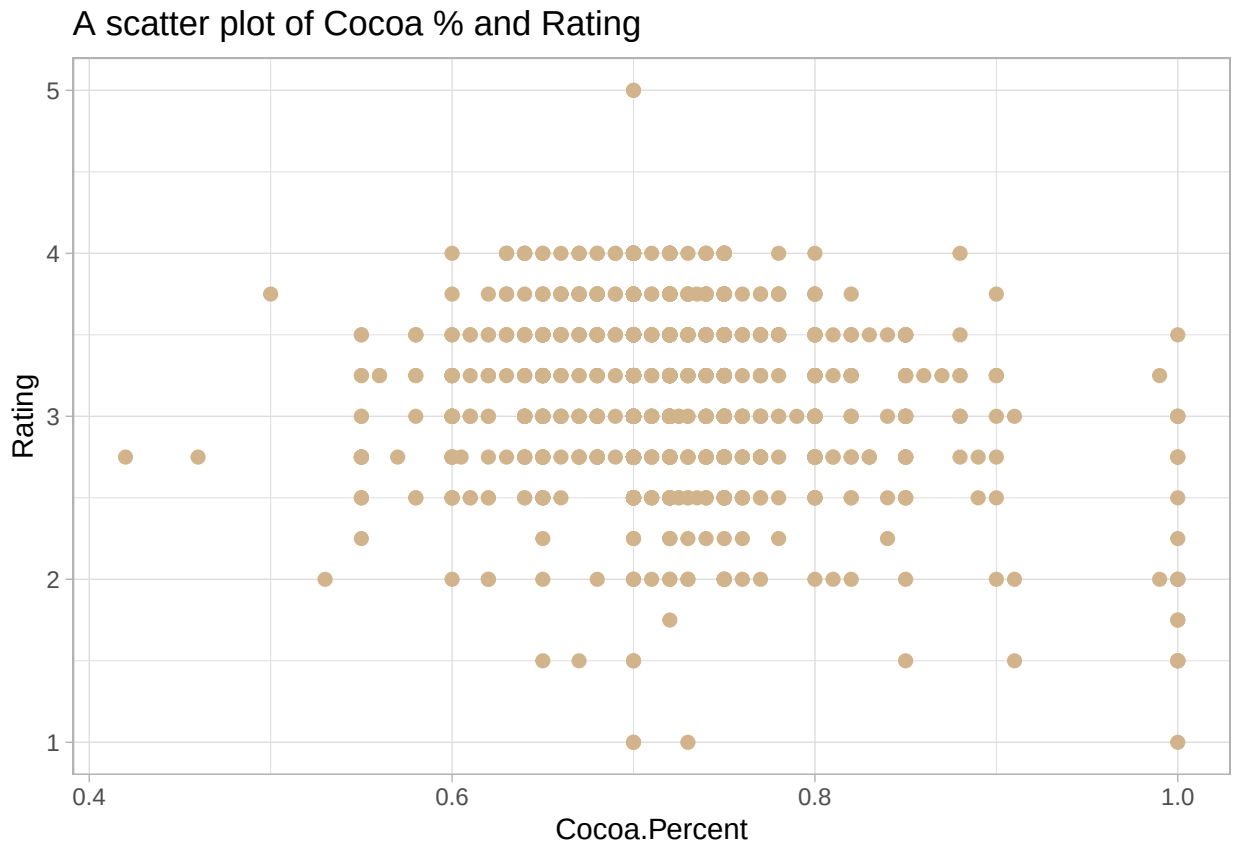
```
## Selecting by Rating
```

```
## # A tibble: 10 x 2
##    Company.Location Rating
##    <fct>             <dbl>
##  1 Chile              3.75
##  2 Amsterdam          3.5
##  3 Netherlands        3.5
##  4 Philippines        3.5
##  5 Iceland            3.42
##  6 Vietnam            3.41
##  7 Brazil             3.40
##  8 Poland             3.38
##  9 Australia          3.36
## 10 Guatemala          3.35
```

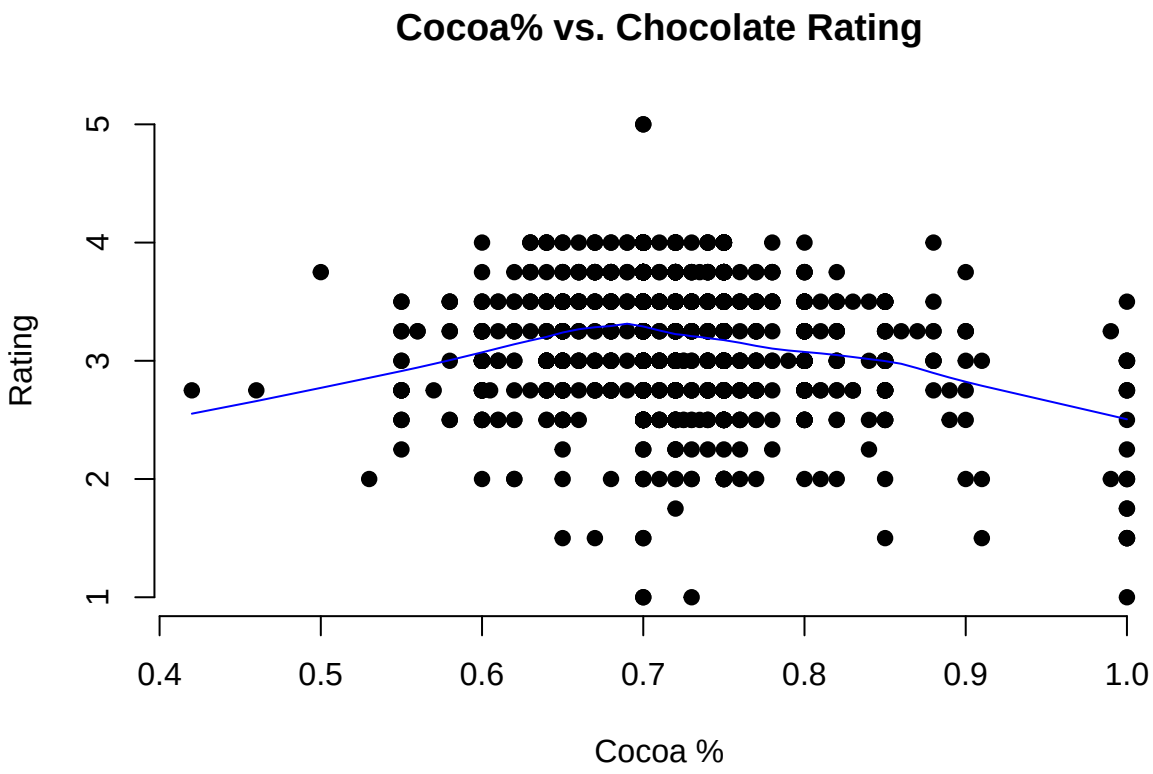
```
ggplot(flavors_of_cacao, aes(x=`Cocoa.Percent`, y=`Rating`)) +
  geom_point(size=2, color = "tan") +
  theme_light() +
  geom_smooth(method='lm',formula=flavors_of_cacao$Rating~flavors_of_cacao$Cocoa.Percent) +
  ggtitle("A scatter plot of Cocoa % and Rating")
```

```
## Warning: 'newdata' had 80 rows but variables found have 1795 rows
```

```
## Warning: Computation failed in `stat_smooth()`:
## arguments imply differing number of rows: 80, 1795
```




```
plot(x = flavors_of_cacao$Cocoa.Percent, y = flavors_of_cacao$Rating, main = "Cocoa% vs. Chocolate Rati
     xlab = "Cocoa %", ylab = "Rating",
     pch = 19, frame = FALSE)
lines(lowess(x = flavors_of_cacao$Cocoa.Percent, y = flavors_of_cacao$Rating), col = "blue")
```



Conclusion: 70% Cocoa percentages tend to have higher rating