MATH411 | Fall 2018 | Exam I

Paul Tomosky

Monday in class, 10/1/2018

Getting the Data

Scrape `The Ramen Rater` data from "https://www.theramenrater.com/resources-2/the-list/", then answer the following questions. Note: (1) Provide your R codes for all the problems, and (2) Save and put your work into you own folder on P drive.

```
URL = "https://www.theramenrater.com/resources-2/the-list/"


ramen = read_html(URL) %>%
 html_table() %>%
 .[[1]] %>%
 as_tibble()


ramen %>% colnames()
ramen %>% summary(Stars)
ramen %>% View()
ramen


ramen = ramen %>%
 mutate(Stars = as.numeric(Stars))


ncol(distinct(ramen(Country)))


ramen %>%
 select(Country) %>%
 distinct() %>% View()
```

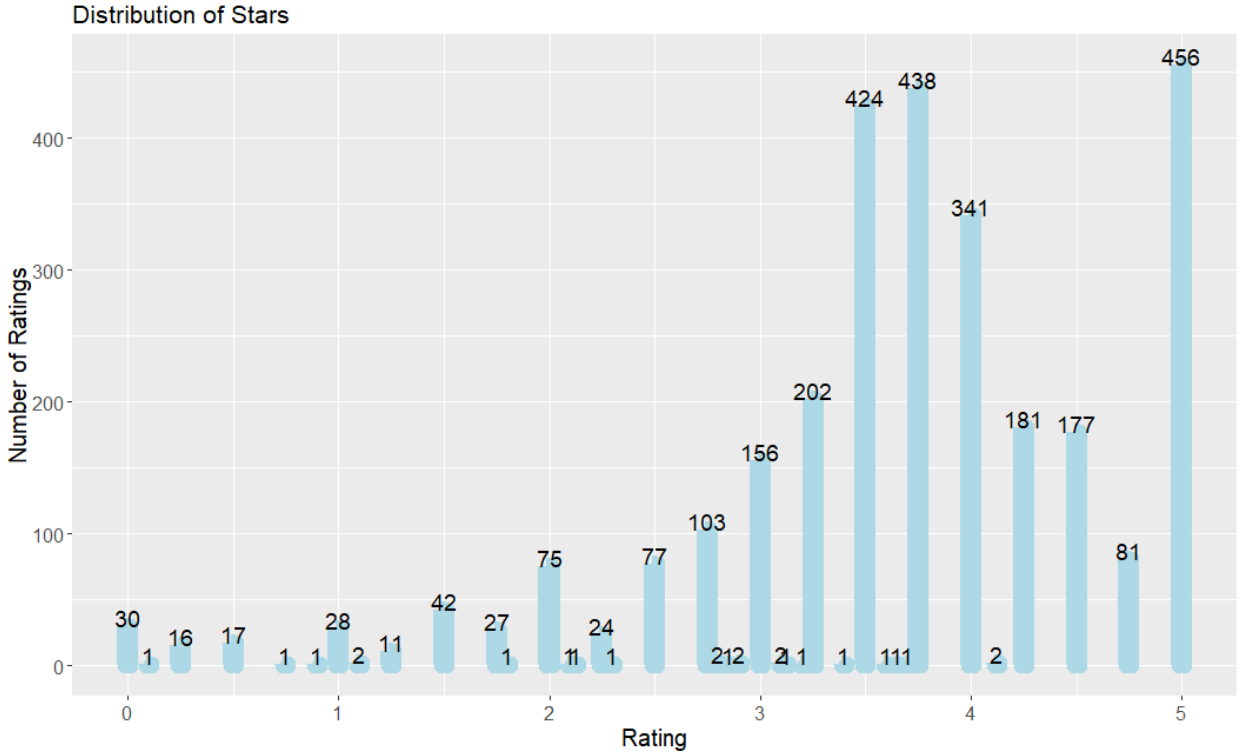
```
ramen %>%
 select(Country, Stars) %>%
 distinct() %>% View()
```

1. **How many ramen items were rated (i.e., has a `Stars` value) so far? (Note: there are 2942 items listed on the webpage, but not all of them were rated.)**

```
ramen %>%
 select(Stars) %>%
 summarise_all(funs(sum(is.na(.))))
ramen = ramen %>%
 drop_na(Stars)
```

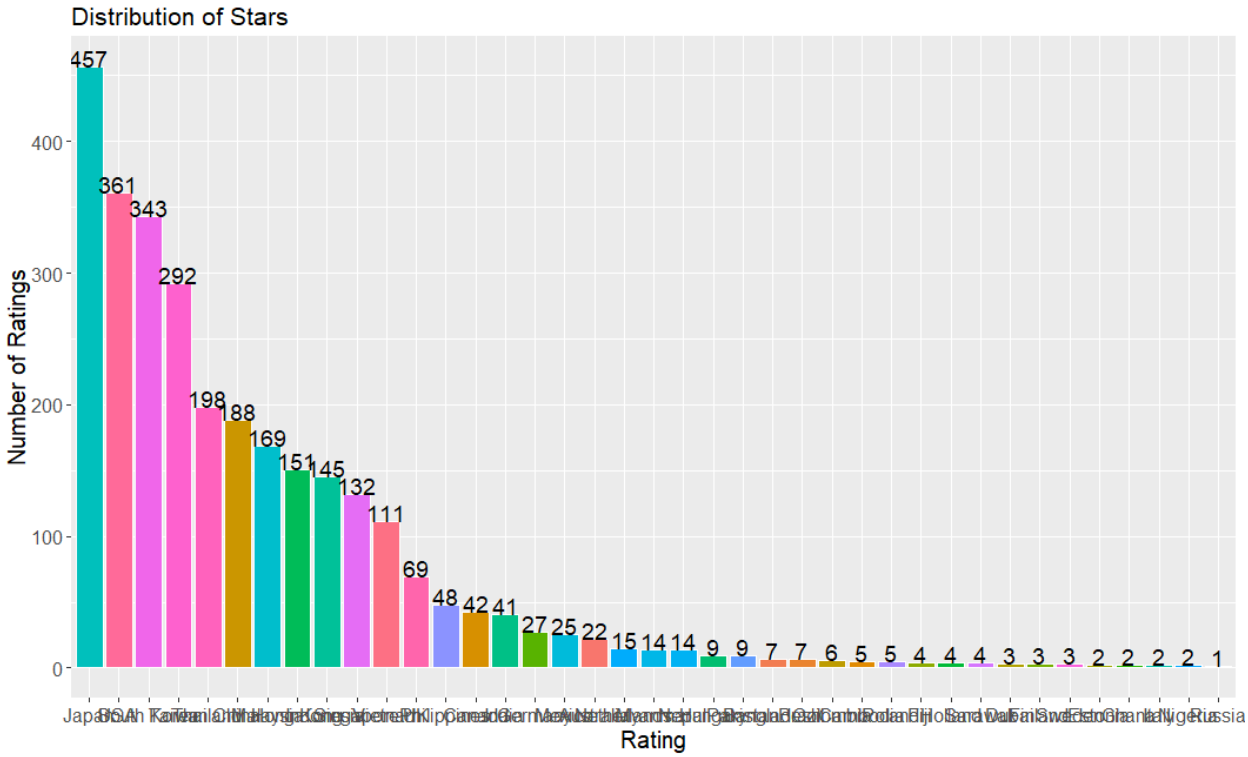
12 were unrated, there were 2932-12 = 2920 rated ramen.

2. **Visualize the distribution of the `Stars` values with an approprate graph. Note: (1) You need to convert the `Stars` values into numeric if it is not, and (2) some of the `Stars` values are in the format like `3.5/2.5`, just use the fisrt number (i.e., 3.5 in the example) as the `Stars` value.**



```
ramen %>%
  count(Stars) %>%
  rename(freq = n) %>%
  ggplot(aes(x = Stars, y = freq, fill = Stars)) +
  geom_bar(stat = "identity", color = "lightblue", size = 5) +
  scale_x_discrete(limits = seq(0, 5, 0.5))+
  guides(fill = FALSE) +
  geom_text(aes(label = freq),
            vjust = 0,
            color = "black",
            size = 5) +
  labs(title = "Distribution of Stars",
       x    = "Rating",
       y    = "Number of Ratings") +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15),
        plot.title = element_text(hjust = 0, size = 16),
        plot.subtitle = element_text(hjust = 0, size = 12))
```

3. Visualize the distribution of the vairable `Country` with an approprate graph. Note: the country names are not consistent. For example, both `United States` and `USA` are used; and some country names are mistyped, for example there are `Philippines` and `Phlippines`. Thus you need to make the country names consistent before you make the graph.



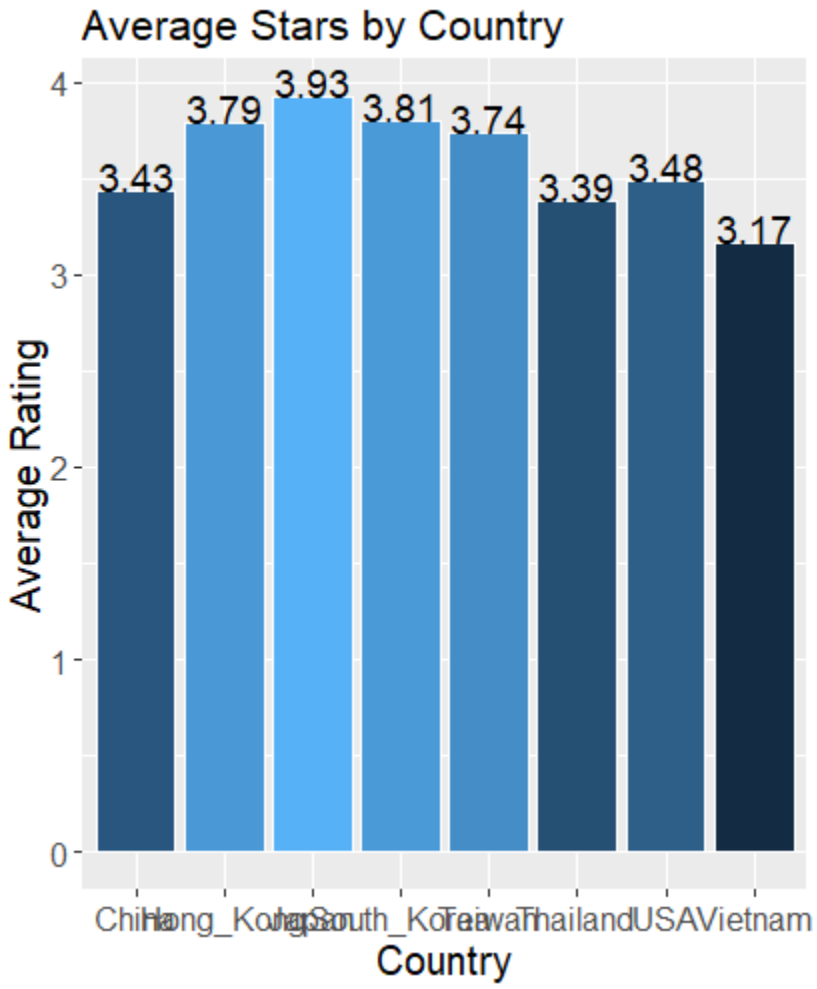

```
ramen = mutate_if(ramen,
                  is.character,
                  str_replace_all, pattern = "United States", replacement = "USA")
ramen = mutate_if(ramen,
                  is.character,
                  str_replace_all, pattern = "Phlippines", replacement = "Philippines")
ramen = mutate_if(ramen,
                  is.character,
                  str_replace_all, pattern = "Hong Kong", replacement = "Hong_Kong")
ramen = mutate_if(ramen,
```

```
            is.character,
            str_replace_all, pattern = "South Korea", replacement = "South_Korea")

    # display ratings by county

ramen %>%
  count(Country) %>%
  rename(freq = n) %>%
  ggplot(aes(x = reorder(Country, -freq), y = freq, fill = Country)) +
  geom_bar(stat = "identity", color = "white") +
  guides(fill = FALSE) +
  geom_text(aes(label = freq),
            vjust = 0,
            color = "black",
            size = 5) +
  labs(title = "Distribution of Stars",
       x    = "Rating",
       y    = "Number of Ratings") +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15),
        plot.title = element_text(hjust = 0, size = 16),
        plot.subtitle = element_text(hjust = 0, size = 12))
```

4. Make an appropriate plot to visualize the distributions of the `Stars` value between the countries `USA`, `China`, `Japan`, `South Korea`, `Taiwan`, `Thailand`, `Hong Kong` and `Vietnam`.




```
AverageStar = ramen %>%
  group_by(Country) %>%
  summarise(avgscore = mean(Stars)) %>%
  drop_na()
```

```
AverageStar = filter(AverageStar, (Country == "USA" | Country == "China"| Country == "Japan"| Country == "South_Korea" |  Country == "Taiwan"| Country == "Thailand"| Country == "Hong_Kong"| Country == "Vietnam" ))


AverageStar %>%
  count(Country, avgscore) %>%
  rename(freq = n) %>%
  ggplot(aes(x = reorder(Country, -freq), y = avgscore, fill = avgscore)) +
  geom_bar(stat = "identity", color = "white") +
  guides(fill = FALSE) +
  geom_text(aes(label = round(avgscore, 2)),
            vjust = 0,
            color = "black",
            size = 5) +
  labs(title = "Average Stars by Country",
       x    = "Country",
       y    = "Average Rating") +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_text(size = 15),
        axis.title.y = element_text(size = 15),
        plot.title = element_text(hjust = 0, size = 16),
        plot.subtitle = element_text(hjust = 0, size = 12))
```