Amber Pasiak

Data Analysis INFO 640

Descriptive Data Analysis

10/29/2019

For the Exploratory Data Analysis, I chose to examine the ratings of different Ramen brands and styles. The data was made available through Kaggle.com, by Aleksey Bilogur. The data was collected through ‘Ramen Rater’, which is a product that organized over 2,500 reviews by Country, Style, Brand, and Variety. The data was converted into a .csv format for reuse and exportation into other programs. After conducting a simple clean and summary of the data set, I found 7 variables, including the ones listed above there was also Stars, Top Ten, and Review Number. These variables each had multiple identifiers within them. Stars had a 1-5 rating scale, there were 38 countries in Country, 355 different Brands, and 8 different types of Styles. I was unable to locate the dates of each review; however, the dataset was updated last on Kaggle in 2017.

The summary of the data indicated that most of the variables were categorical. The numerical values were only represented under the review. This presented a problem later on when I tried to create a histogram of the 2 variables Stars and Styles. I was unable to figure out how to force the Star variable into a numeric form. I created a boxplot (Fig. 1) in order to access the information in a visual form, however the result did not show a good representation of the distribution or central tendency to the variables. I continued by trying to specify, in a scatterplot, by country and star ratings (Fig. 2) but then realized that these variables showed the average number of ratings by each country, which is not what I was looking to explore. Next I created a scatterplot that looked at the average rating of different styles from the USA (Fig. 3).

At the end of the assignment, when I was finding the minimum and maximum trends, I felt I was gaining the best understanding of the dataset. This is when I was able to figure out which brand was USA favorite; however, I was not able to get the ggplot to work properly (Fig. 4). I would like to have been able to figure out how to create a ggplot containing the mean ratings per brand, or style for USA, as well as a desc rating per brand or style for USA. I tried to conduct a range rate of ramen ratings through the minimum and maximum (…)/(…), but was also unable to complete this.

Even though the dataset is clean and organized, I believe that the reliability of the reviews cannot be established. This is the largest limitation to the dataset. It is questionable with star reviews if the reviewer is genuinely giving their response or being paid for a high mark review, resulting in a lack of ability to independently verify each response. After examining the Ramen Rater website, it also became apparent that the amount of responses for each poll, such as those on favorite flavor or brand, were not equal. As an example, there were 516 votes for preferred flavor, and only 214 on the poll for origin on favorite noodles. There is no metadata or information on how this is accounted for in the dataset. An additional possible blindspot deals with bias in the ratings. There is no way of knowing if all the reviewers that rated Japan as the best origin of noodles were Japanese or not.

Another aspect of the data to take into consideration are the variables “chosen”. In the inspirations listed for republishing this data, the creator mentions wanting to see how ramen ratings compare to other products, yet the variables were predetermined from Ramen Rater.

While keeping the boundaries of the dataset in mind, I do believe that it is possible to make fair assessments about preferred styles, brands and flavors from different countries.

*Findings:* (Fig 5.)

USA favorite brand – Yamachan

USA favorite style – Pack

**Images:**

Figure 1:

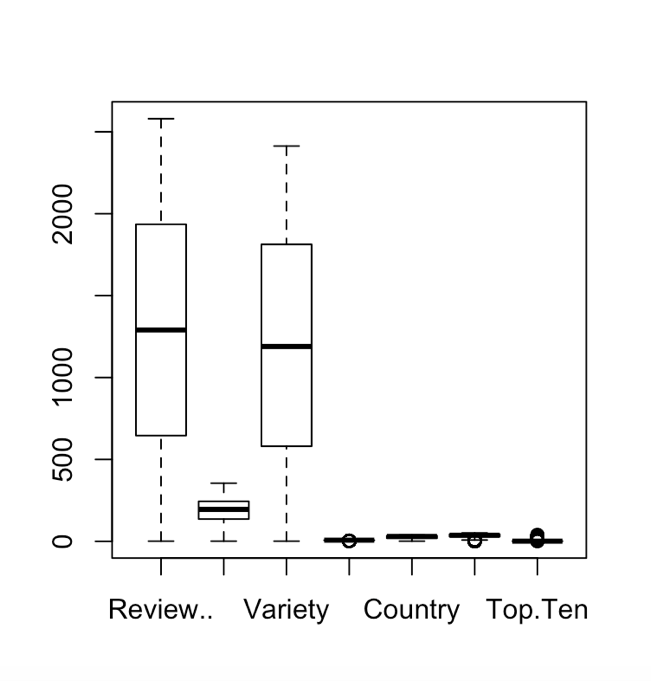
****

Figure 2:

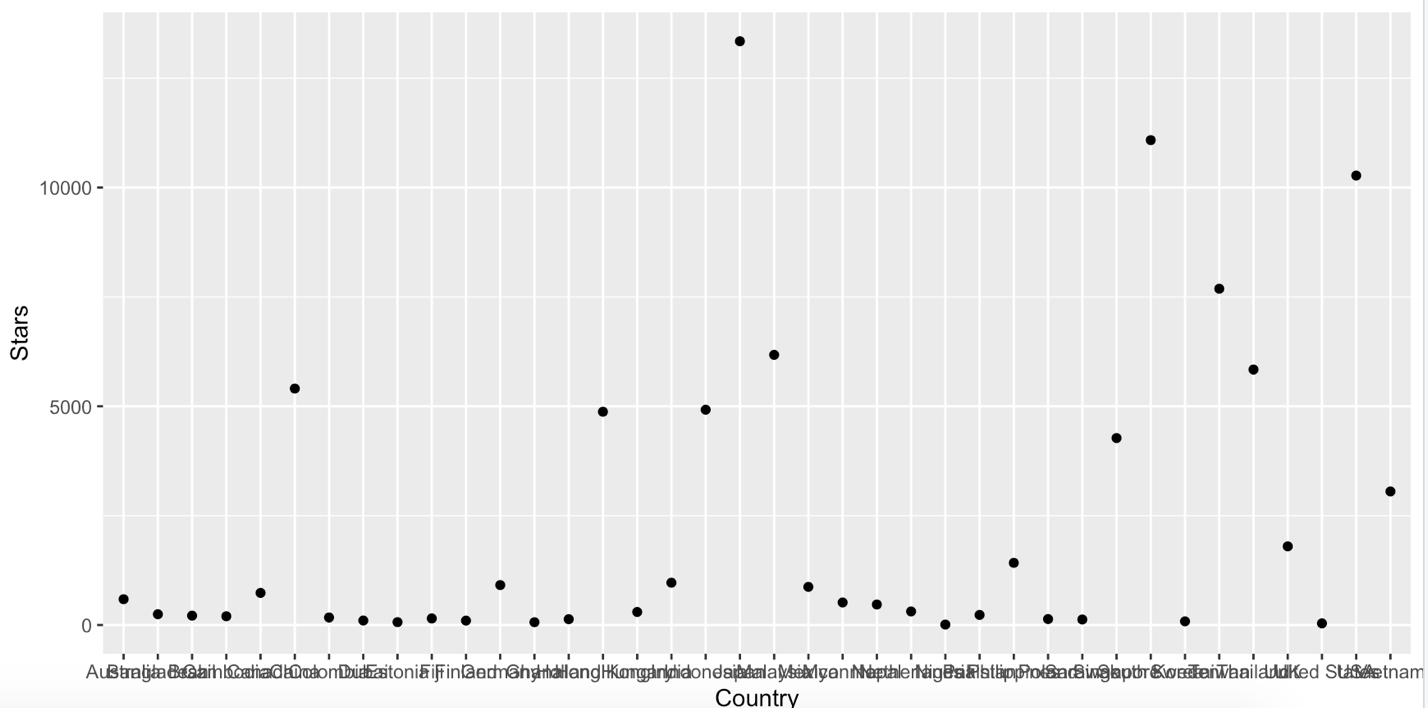
****

Figure 3:

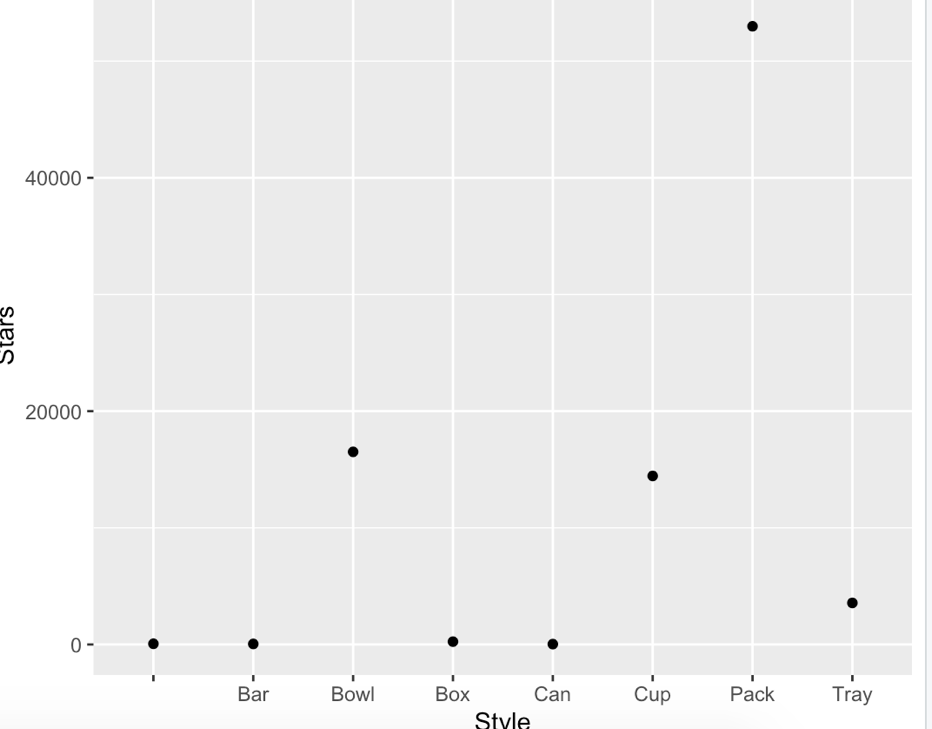
****

Figure 4:

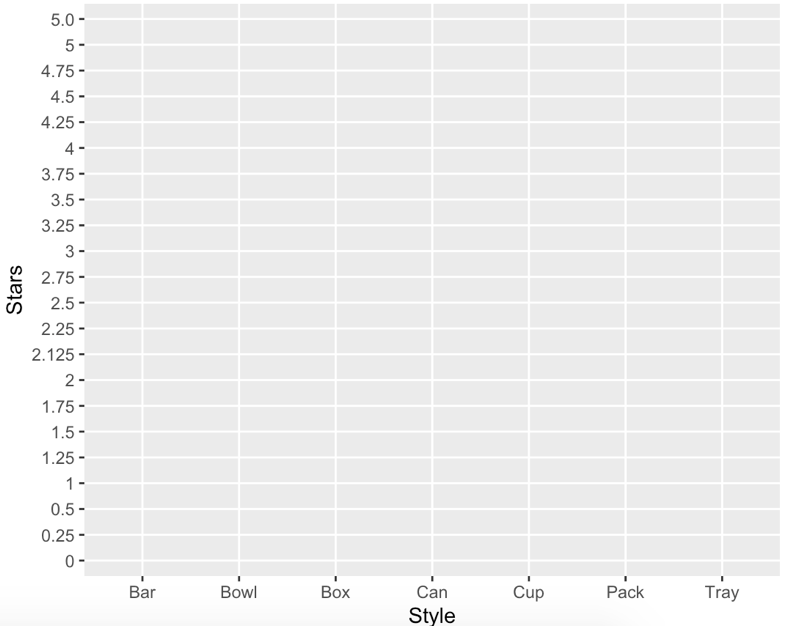
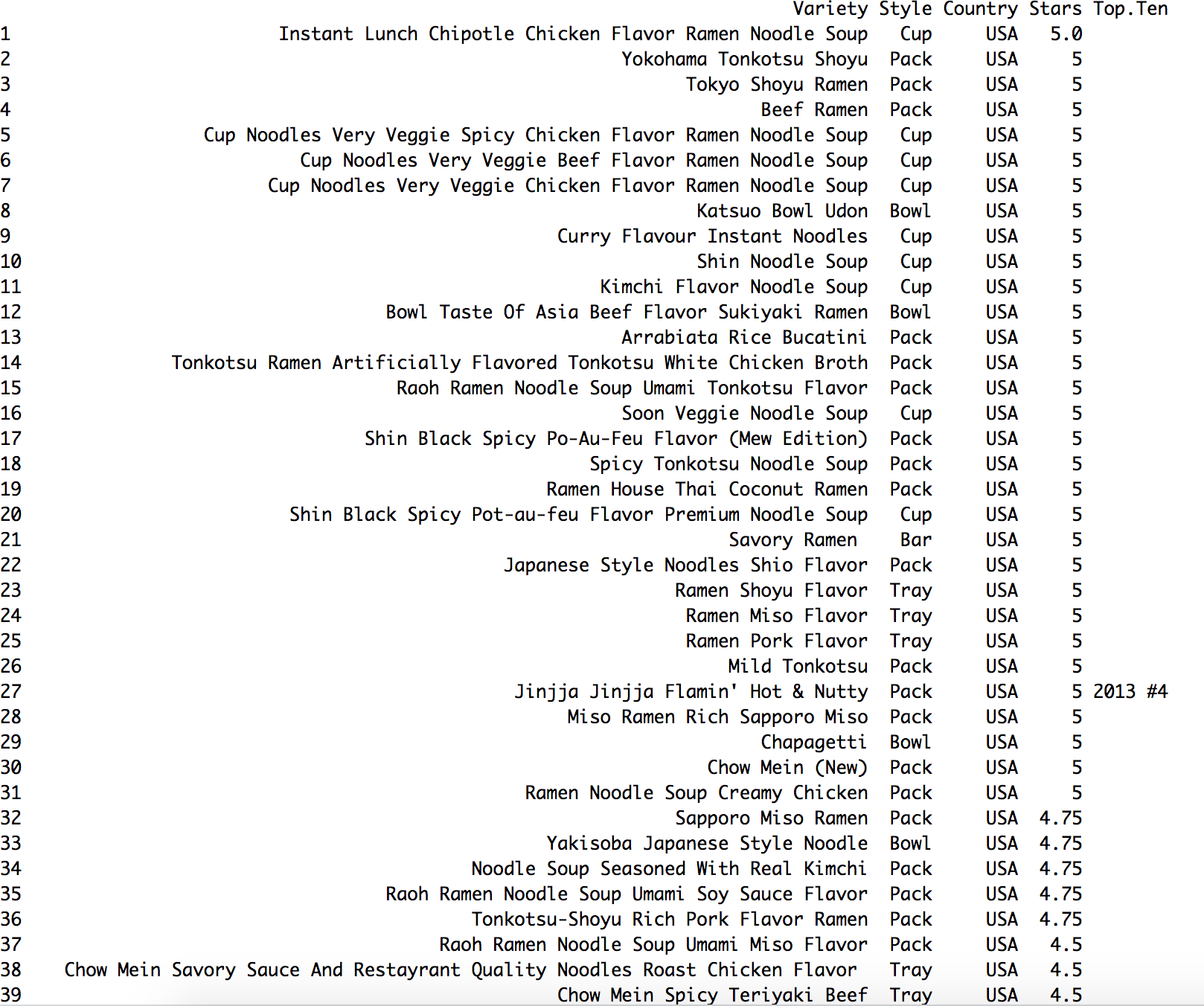


Figure 5:



**Code:**

library(tidyverse)

library(dplyr)

library(lubridate)

ramen <- read.csv ("../Desktop/ramen-ratings.csv", header = TRUE)

dim(ramen)

names(ramen)

str(ramen)

glimpse(ramen)

summary(ramen)

head(ramen)

tail(ramen)

sum(is.na(ramen))

boxplot(ramen)

ramen1 <- ramen %>% filter (Country == "USA", Stars == "5")

head(ramen1)

write.csv(ramen1, 'ramen1USAStars.csv')

ramen %>% arrange (Stars)

ramen %>% arrange (desc(Stars))

*#filter & arrange*

cnt <- ramen %>% filter (Country == "USA", Style == "Pack")

head(cnt)

*#result:*

*#Review.. Brand Variety Style Country Stars*

*#1 2569 Yamachan Yokohama Tonkotsu Shoyu Pack USA 5*

*#2 2563 Yamachan Tokyo Shoyu Ramen Pack USA 5*

*#3 2559 Jackpot Teriyaki Beef Ramen Pack USA 5*

*#4 2557 Yamachan Sapporo Miso Ramen Pack USA 4.75*

*#5 2470 Myojo Udon Japanese Style Noodles With Soup Base Hot & Sour Flavor Pack #USA 3.75*

*#6 2412 Shirakiku Sanukiya Udon Shrimp Flavor Pack USA 3.25*

ramen1 %>%

filter (Counrty == "USA") %>%

arrange (desc(Stars))

by\_stars <- ramen %>%

group\_by(Country) %>%

summarize(Stars = sum(as.numeric(Stars)),

meanStars = mean(Stars))

ggplot (by\_stars, aes(x=Country, y= Stars)) +

geom\_point()

by\_stars <- ramen %>%

group\_by(Style) %>%

summarize(Stars = sum(as.numeric(Stars)),

meanStars = mean(Stars))

ggplot (by\_stars, aes(x=Style, y= Stars)) +

geom\_point()

by\_stars <- ramen %>%

filter (Country=="USA")

group\_by(Style) %>%

summarize(Stars = sum(as.numeric(Stars)),

meanStars = mean(Stars))

ggplot (by\_stars, aes(x=Style, y= Stars)) +

geom\_point()

Styles<- ramen$Style

hist(Style)

*#1*

summary(ramen)

*#2*

ramenUSA <- ramen %>% filter(Country == "USA")

ramenUSA[which.min(ramenUSA$Stars),]

ramenUSA[which.min(ramenUSA$Style),]

ramenUSA[which.max(rameUSAn$Stars),]

ramenUSA[which.max(ramenUSA$Style),]

*#Results:*

*Review.. Brand Variety Style Country Stars Top.Ten*

*14 2426 Dr. McDougall's Vegan Pad Thai Noodle Soup Cup USA 0*

*> ramenUSA[which.min(ramenUSA$Style),]*

*Review.. Brand Variety Style Country Stars Top.Ten*

*120 1155 Komforte Chockolates Savory Ramen Bar USA 5*

*> ramenUSA[which.max(ramenUSA$Stars),]*

*Review.. Brand Variety Style Country Stars Top.Ten*

*44 2132 Maruchan Instant Lunch Chipotle Chicken Flavor Ramen Noodle Soup Cup USA 5.0*

*> ramenUSA[which.max(ramenUSA$Style),]*

*Review.. Brand Variety Style Country Stars Top.Ten*

*54 1958 IbuRamen Mi Goreng Bento Noodles In A Box Curry Masala Flavour Tray USA 4*

*>*

*#3*

ram\_grouped <- ramen %>%

group\_by (Country, Style) %>%

summarize (MeanStyles = mean(Styles))

meanstars<- mean(as.numeric(Stars))

head(meanstars)

summary(ram\_grouped)

ggplot(ram\_grouped, aes (x=Style, y= meanstars)) +

geom\_line()

*#did not work, no data on line plot*

*#4*

ramenUSA <- ramen %>% filter(Country == "USA")

ggplot(ramenUSA, aes (x=Style, y= Stars))

geom\_point()

*#5*

cnt <- ramen %>% filter (Country == "USA")

head(cnt)

write.csv (cnt, 'ramen.csv')

cnt %>% arrange(Stars)

cnt %>% arrange (desc(Stars))

head(cnt)

star5 <- ramen %>% filter (Stars == "5")

**Source:**

<https://www.kaggle.com/residentmario/ramen-ratings>