si618_hw7_kjunwonl

Kjunwonl

10/30/2019

R Markdown

Part 1

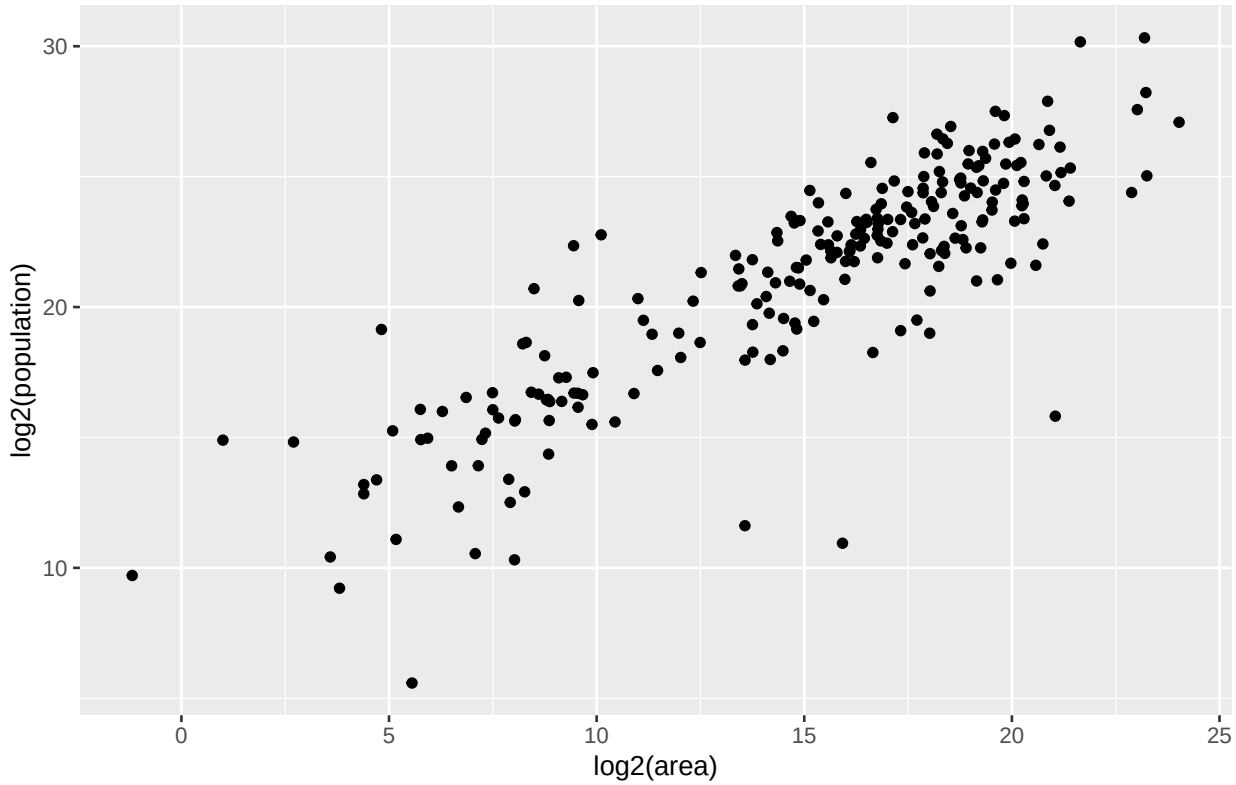
Question 1: First the provided TSV data file is loaded into R using csv.read function, seperated by column names.

```
country <- read.csv("countrydata_withregion.tsv", sep = "\t")
head(country, 15)
```

```
##                country                          region      area
## 1          AFGHANISTAN                            Asia  650230.0
## 2              ALBANIA                          Europe   28748.0
## 3              ALGERIA                          Africa 2381741.0
## 4       AMERICAN SAMOA                         Oceania     199.0
## 5              ANDORRA                          Europe     468.0
## 6               ANGOLA                          Africa 1246700.0
## 7             ANGUILLA Central America & the Caribbean      91.0
## 8  ANTIGUA AND BARBUDA Central America & the Caribbean     442.6
## 9            ARGENTINA                   South America 2780400.0
## 10             ARMENIA                            Asia   29743.0
## 11               ARUBA Central America & the Caribbean     180.0
## 12           AUSTRALIA                         Oceania 7741220.0
## 13             AUSTRIA                          Europe   83871.0
## 14          AZERBAIJAN                            Asia   86600.0
## 15        BAHAMAS, THE Central America & the Caribbean   13880.0
##    population
## 1    30019928
## 2     3002859
## 3    37367226
## 4       54947
## 5       85082
## 6    18056072
## 7       15423
## 8       89018
## 9    42192494
## 10    2970495
## 11     107635
## 12   22015576
## 13    8219743
## 14    9493600
## 15     316182
```

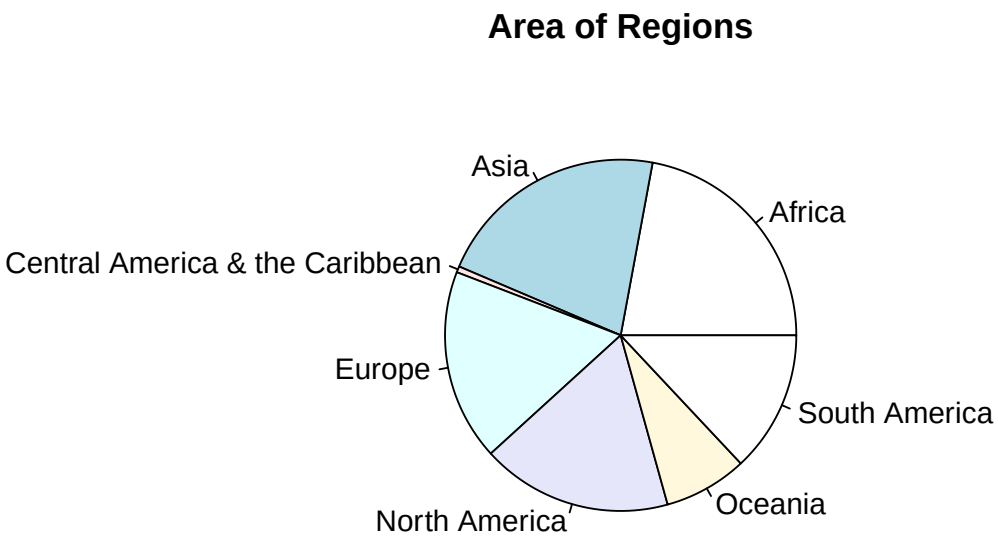
Question 2:

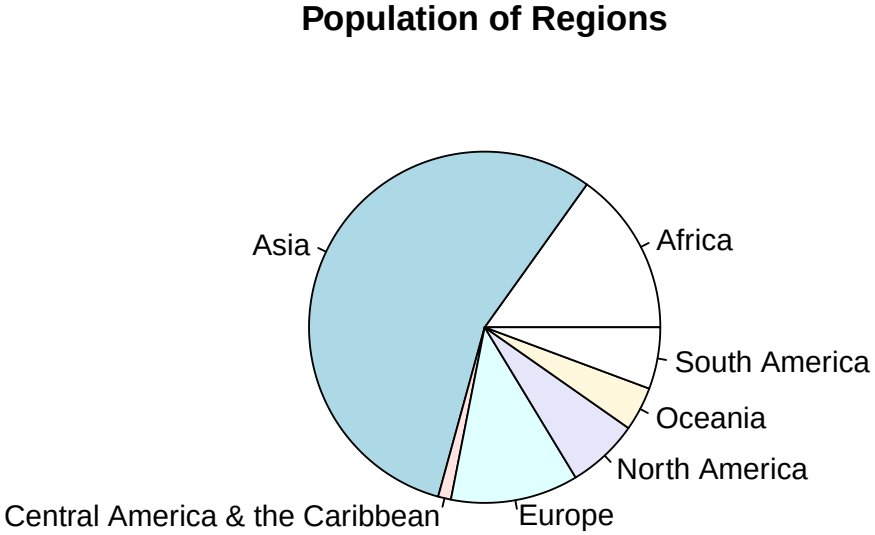
Logarithms (base 2) of the area and the population of each country are computed and used to produce the following scatter plot using ggplot() + geom_point() function.



Question 3:

The areas and populations of all countries in a region are summed up using the aggregate() function, respectively. Then the following two pie charts are created using the pie() function.

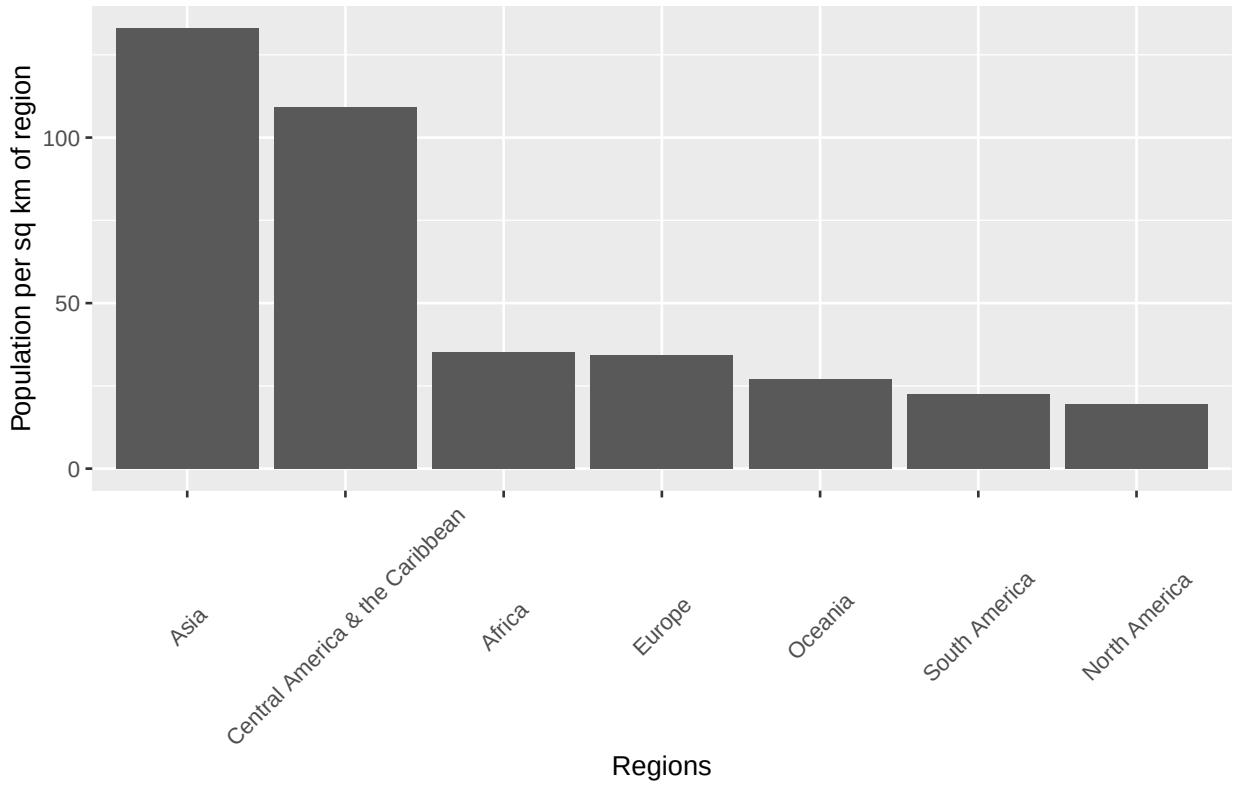




Question 4:

A new data frame is created to contain the population per sq km of each region using the data.frame()function.

The data frame is then sorted by population per sq km in decreasing order with the help of thereorder() function. Finally, the following bar plot is created using ggplot(). In order to rotate the x axis labels, I used theme(axis.text.x=element_text(angle=45,hjust=0.5,vjust=0.5)) to make it be seen more clearly.



Part 2

Question 5: First the provided TSV data file is loaded into R using csv.read function, seperated by column names. I then mutated columns city, state and main_category in order to factor them. I then used na.omit() to clear all my business data of empty values. I then created a summary of this data.

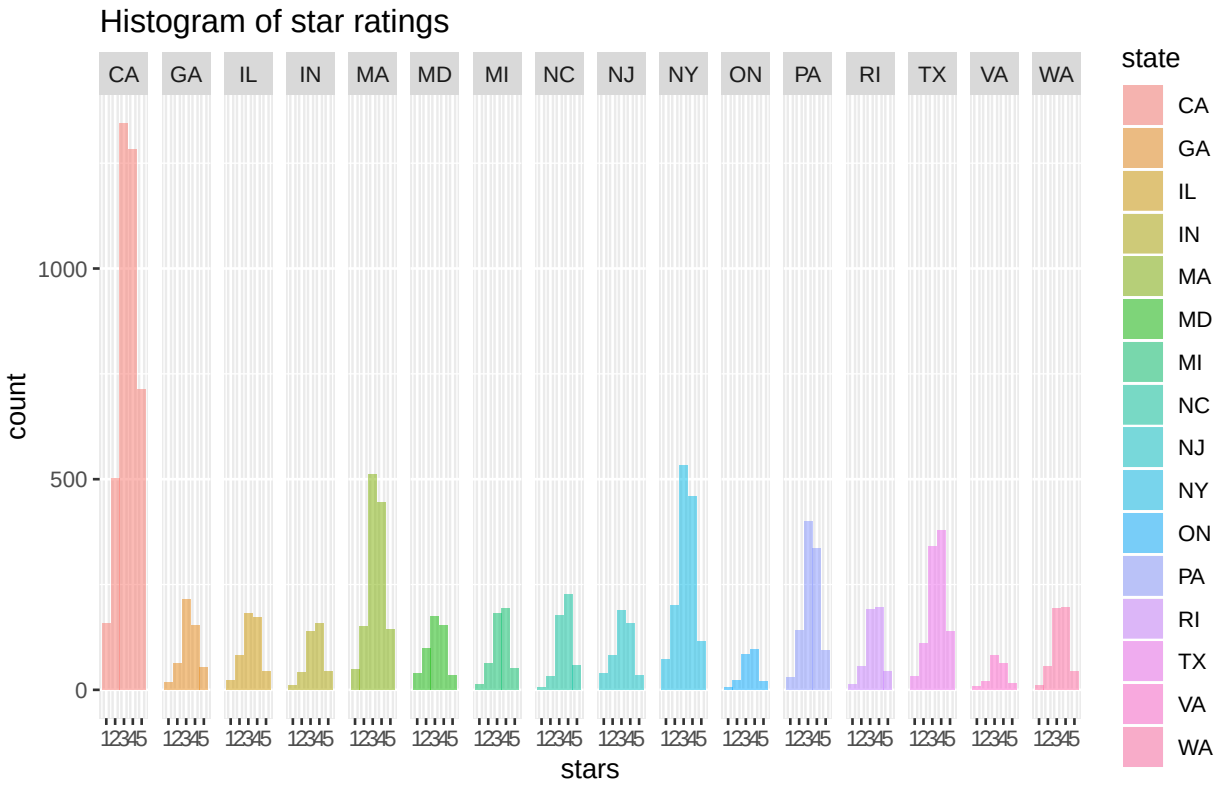
```
business <- read.csv("businessdata.tsv", sep = "\t")
newbusiness <- business %>% mutate(city = factor(city)) %>% mutate(state = factor(state)) %>% mutate(ma
finalbusiness <- na.omit(newbusiness)
summary(finalbusiness)
```

```
##                                name                   city
##  Starbucks                       :   43   Los Angeles    : 944
##  Subway                          :   39   Cambridge      : 924
##  FedEx Office Print & Ship Center:   18   Austin         : 493
##  Starbucks Coffee                :   18   Houston        : 492
##  McDonald's                      :   17   Berkeley       : 491
##  Domino's Pizza                  :   16   San Luis Obispo: 491
##  (Other)                         :12986   (Other)        :9302
##      state          stars        review_count            main_category
##  CA     :3917   Min.   :1.000   Min.   :   2.00   Food           :1658
##  NY     :1336   1st Qu.:3.000   1st Qu.:   3.00   Shopping       : 502
##  MA     :1240   Median :3.500   Median :   7.00   Local Services : 446
##  TX     : 987   Mean   :3.628   Mean   :  26.86   Active Life    : 401
##  PA     : 979   3rd Qu.:4.500   3rd Qu.:  21.00   Hair Salons    : 369
```

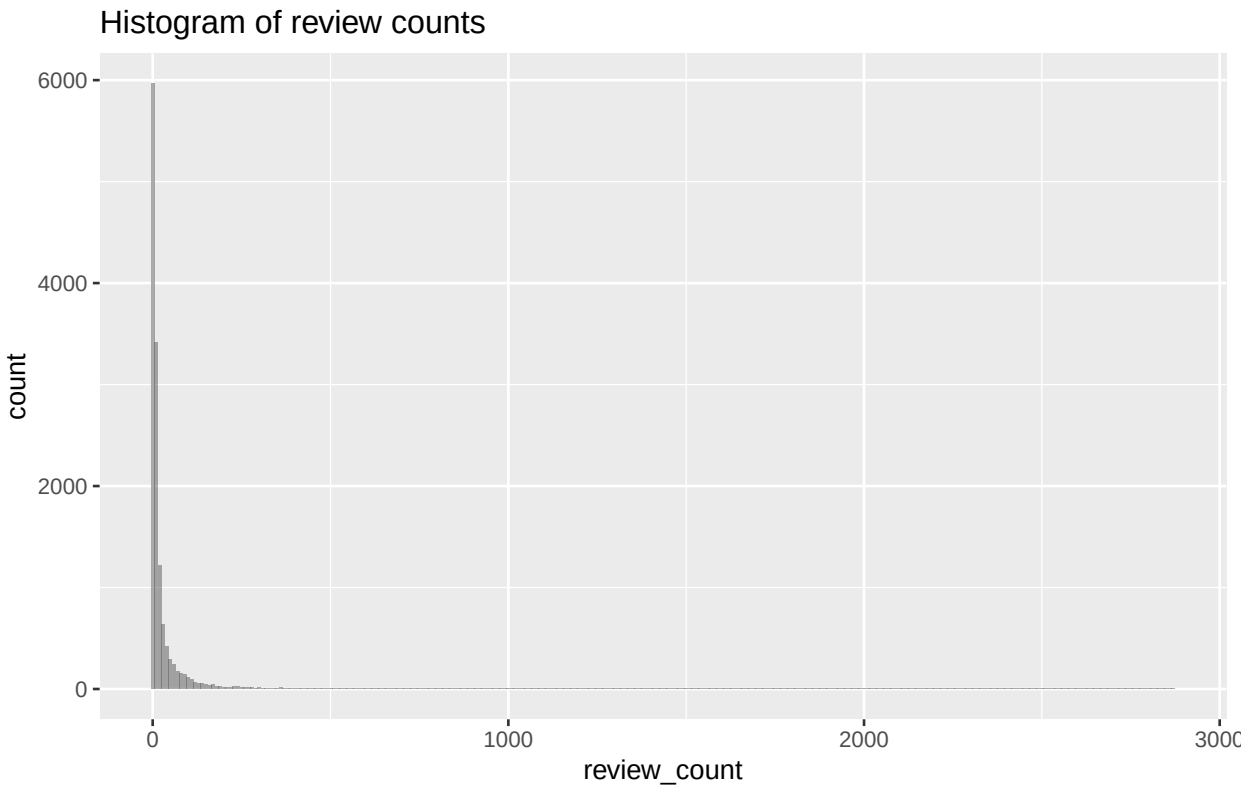
```
##  NC      : 494   Max.   :5.000   Max.   :2874.00   Hotels & Travel: 352
##  (Other):4184                                     (Other)        :9409
```

Question 6:

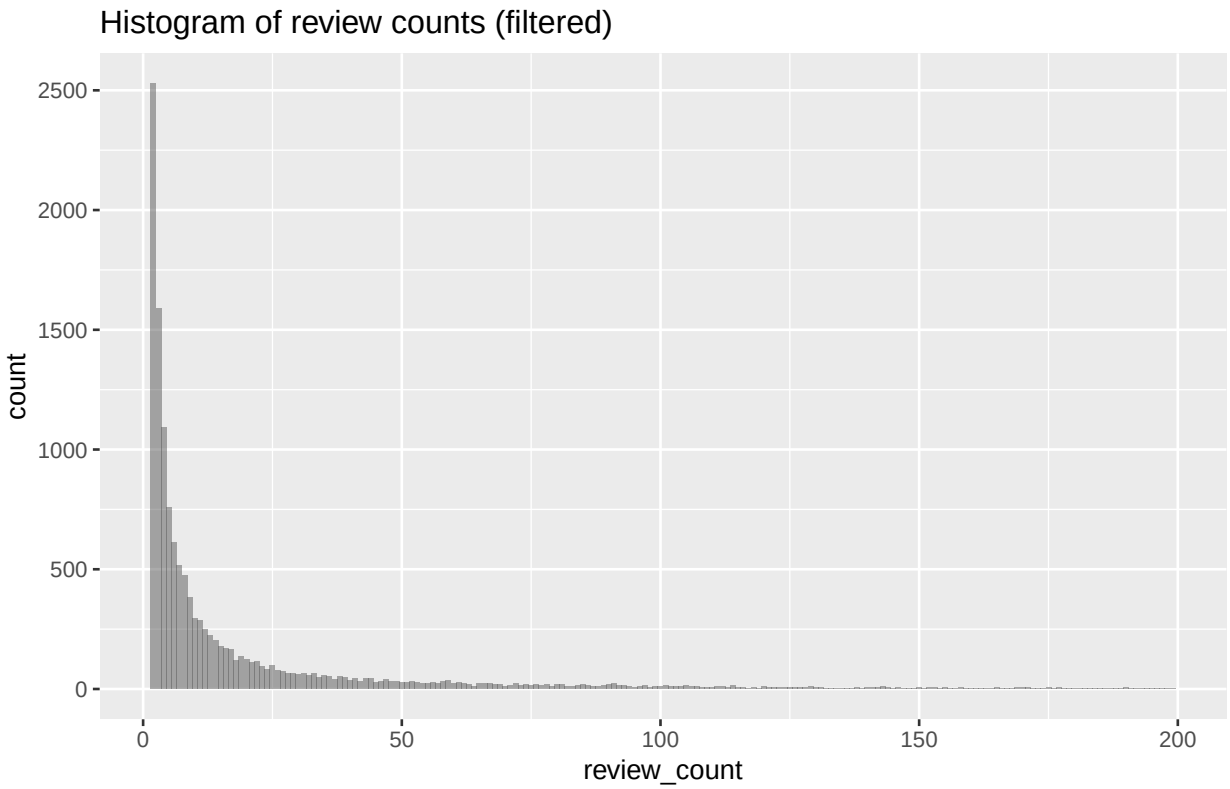
Histograms for star ratings using ggplot() is shown below with a binwidth = 1.



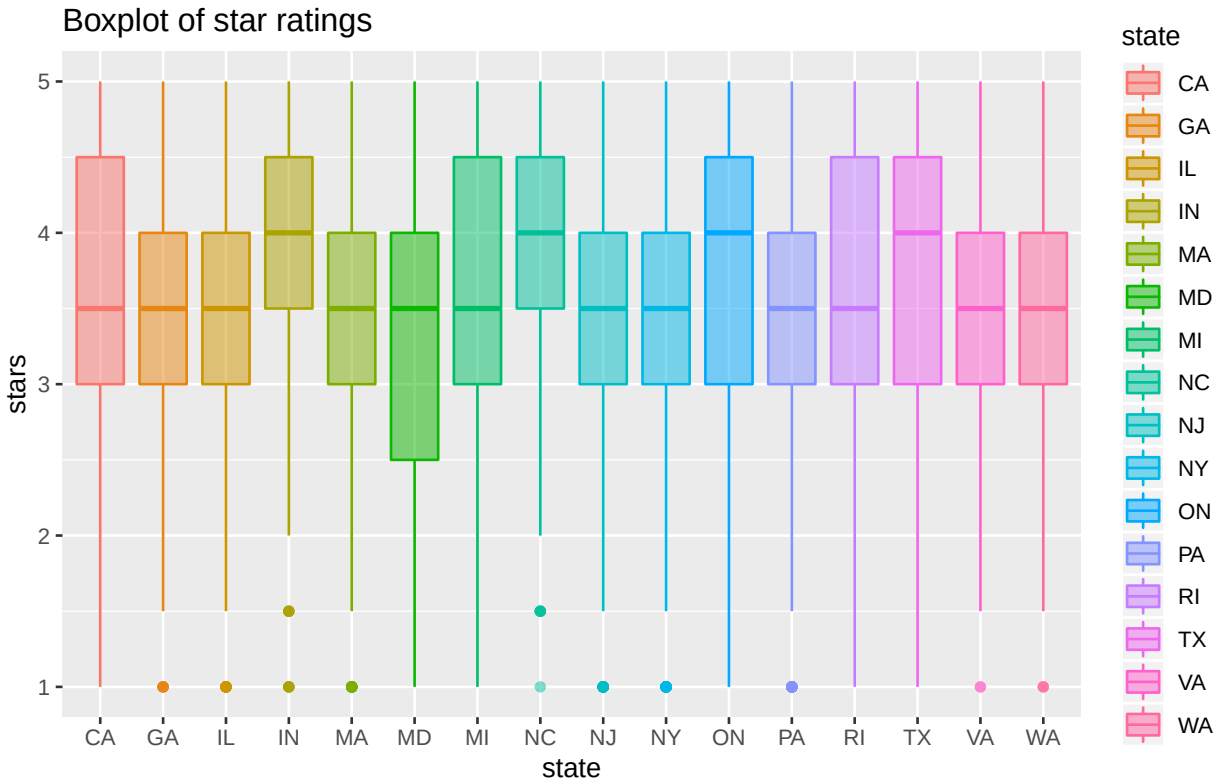
Question 7: Histograms of review counts are plotted with ggplot() function and a binwidth of 10



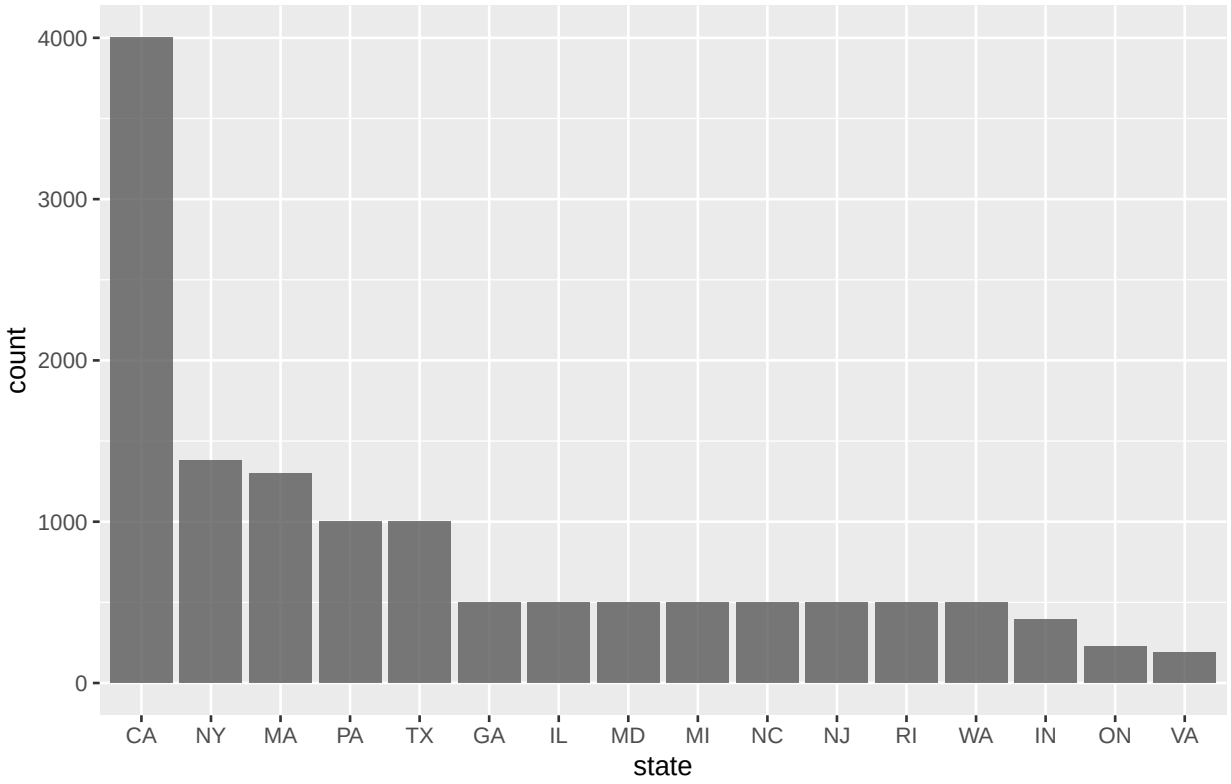
We can see that the distribution of review counts has a long tail. To zoom in on the bars to the left of the 200 mark, we use the data.table syntaxor the subset() function to select just the data with review count<= 200. Afterwards, I plotted the histogram again with binwidth = 1.



Question 8:



Question 9:


4000
3000
2000
1000
0
count
CA
NY
MA
PA
TX
GA
IL
MD
MI
NC
NJ
RI
WA
IN
ON
VA
state