

SI 618 Homework 7

Part 1 (40 points)

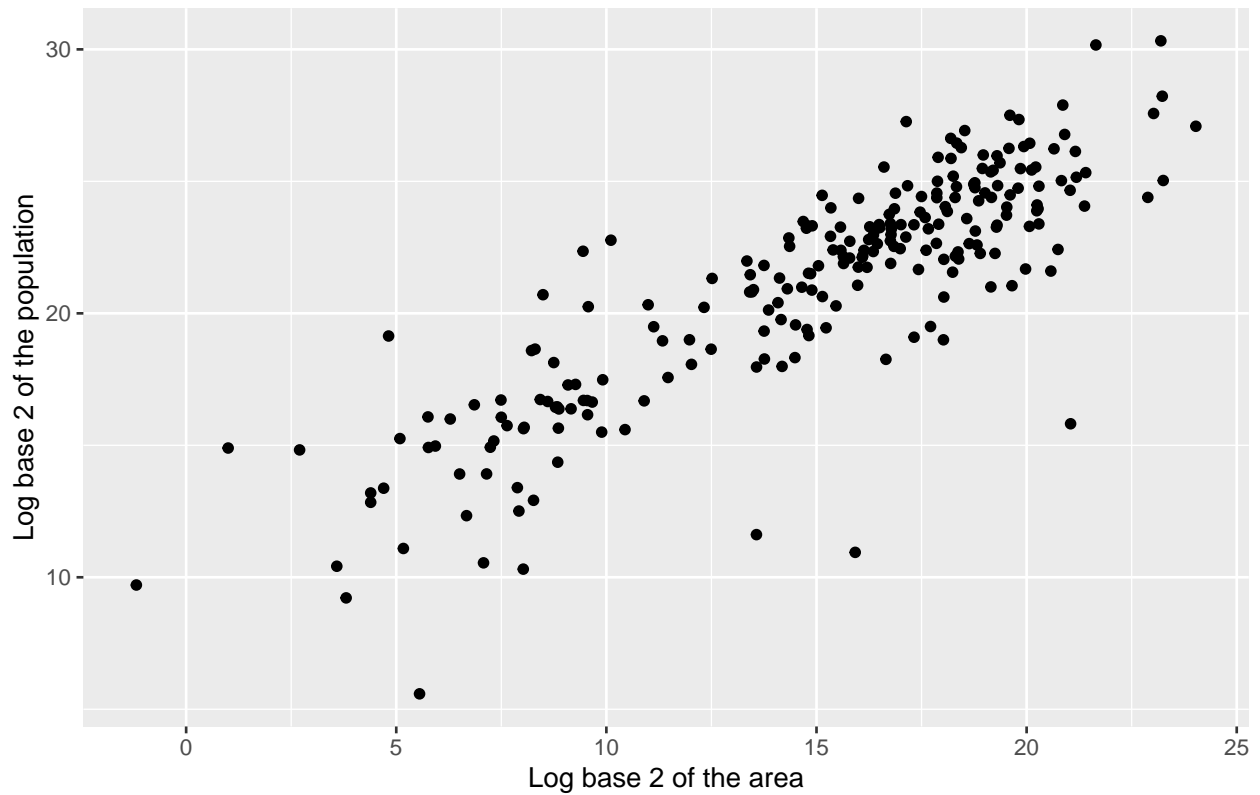
Question 1: Load country data (5 points)

First the provided TSV data file is loaded into R using the `read.table()` function. Display the first 15 rows of the data frame:

```
##           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: Scatter plot of log transformed data (5 points)

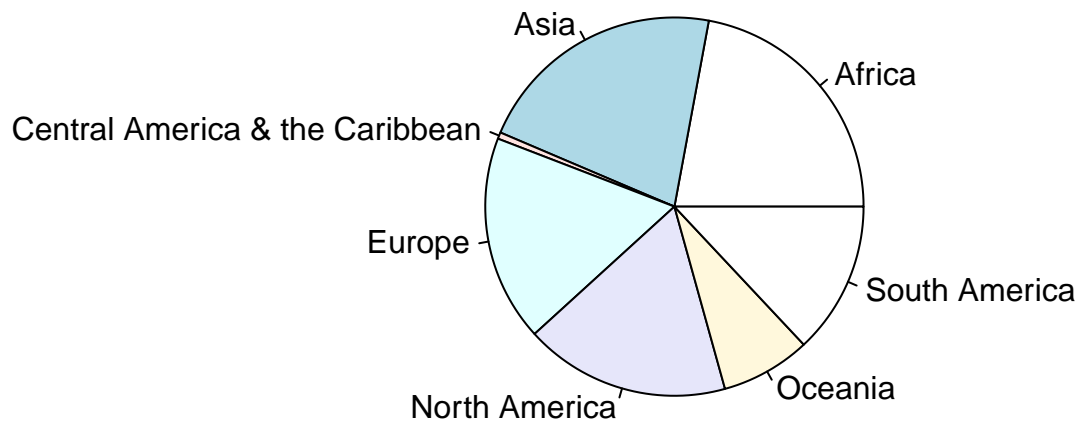
Logarithms (base 2) of the area and the population of each country are computed and used to produce the following scatter plot using the `qplot()` function. Use `{r echo=FALSE, fig.width=7}` for all of the plots.



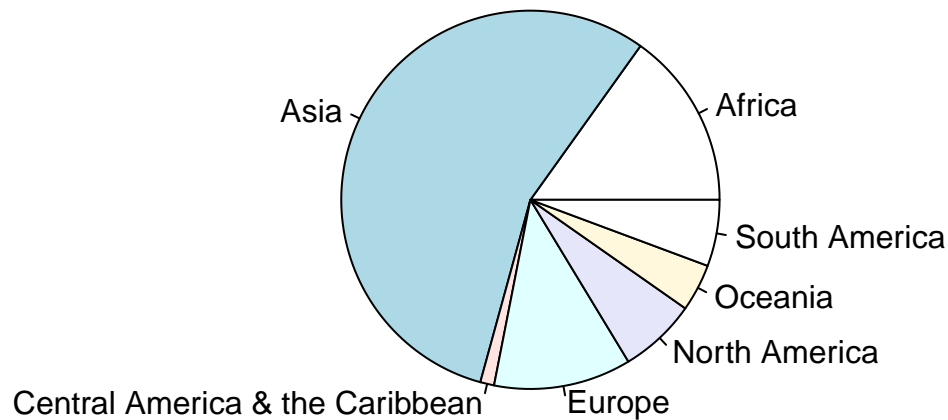
Question 3: Data aggregation by region (15 points)

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.

Area of Regions

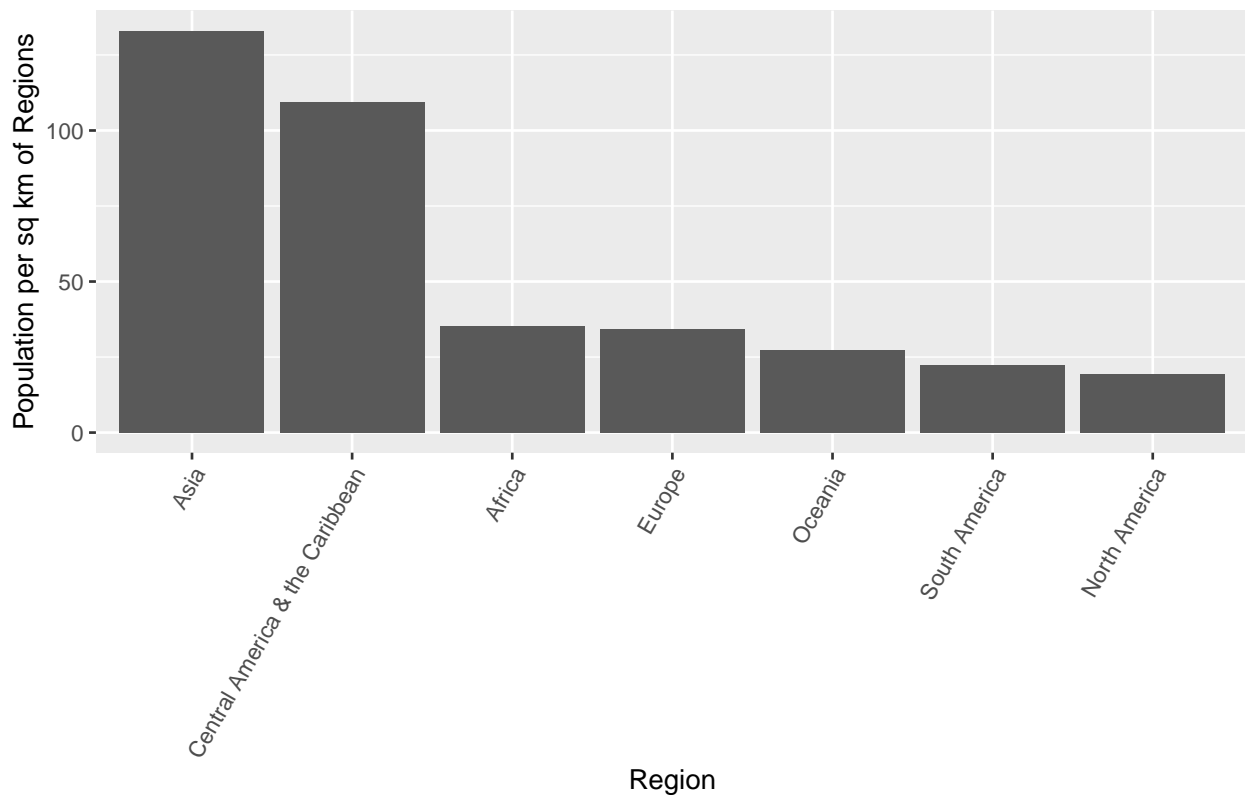


Population of Regions



Question 4: Visualization of Population per sq km of Regions (15 points)

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 the `reorder()` function. Finally, the following bar plot is created using the `qplot()` function with `geom="bar"`. In order to rotate the x-axis labels, add `+ theme(axis.text.x = element_text(angle = 60, hjust = 1))` at the end of the `qplot()` function call.



Part 2 (60 points)

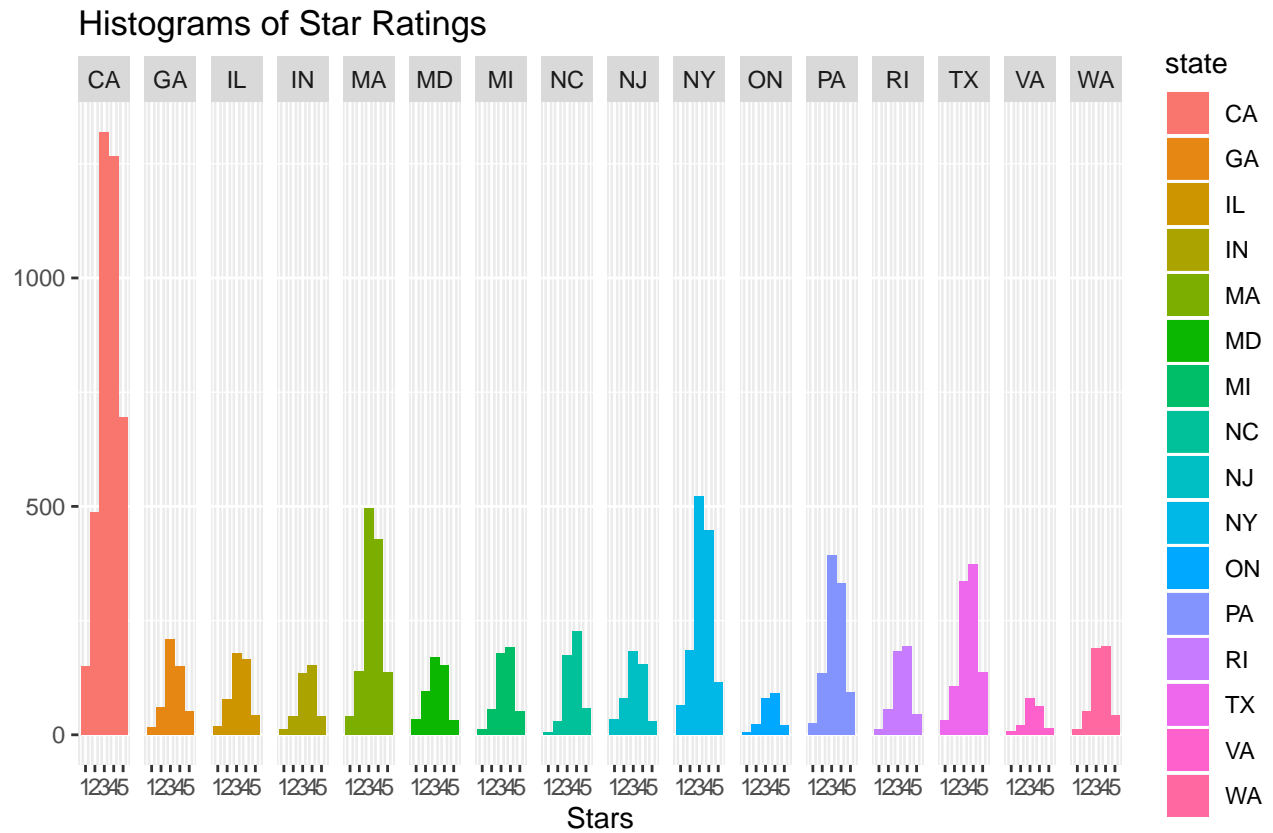
Question 5: Load yelp data & generate summary (10 points)

Load the TSV data file: `businessdata.tsv` into a R data frame using the `read.table()` function. The `city`, `state` and `main_category` columns should be converted to factors. Then listwise deletion (http://en.wikipedia.org/wiki/Listwise_deletion) is applied to remove records with missing data (use the `na.omit()` function). Then the `data.frame` is converted to a `data.table`. Here is the summary of the data table:

```
##      name                city      state      stars
## Length:13134    Los Angeles  : 944    CA      :3916    Min.      :1.000
## Class :character    Cambridge  : 924    NY      :1336    1st Qu.:3.000
## Mode  :character    Austin    : 493    MA      :1240    Median :3.500
##                      Berkeley   : 491    TX      : 986    Mean   :3.628
##                      Houston    : 491    PA      : 979    3rd Qu.:4.500
##                      San Luis Obispo: 491    NC      : 494    Max.   :5.000
##                      (Other)    :9300    (Other):4183
## review_count      main_category
## Min.      : 2.00    Food      :1657
## 1st Qu.: 3.00    Shopping  : 501
## Median : 7.00    Local Services : 446
## Mean   : 26.86    Active Life  : 401
## 3rd Qu.: 21.00    Hair Salons  : 369
## Max.   :2874.00    Hotels & Travel: 352
##                      (Other)    :9408
```

Question 6: Histogram of Star Rating (10 points)

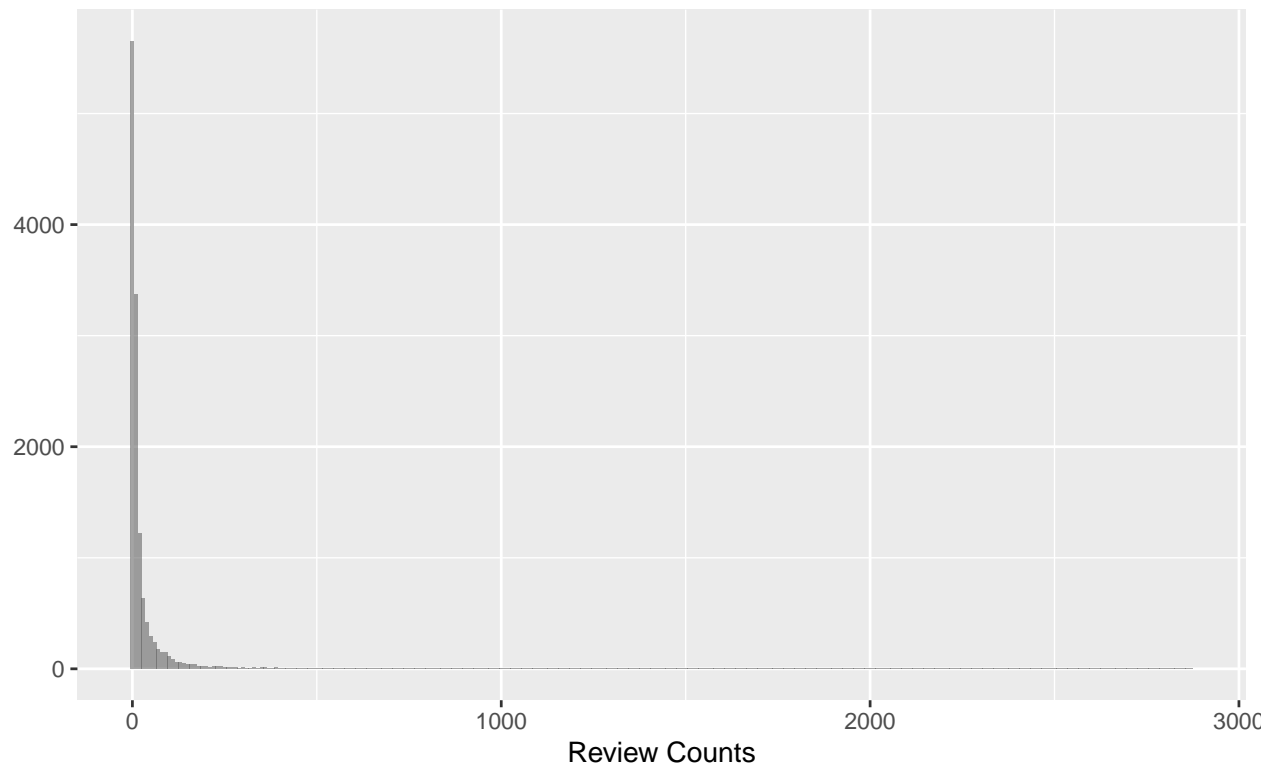
The Histogram of star ratings is plotted with the `qplot()` or `ggplot()` function. The actual counts plot is shown. (Use `binwidth=1`)



Question 7: Histograms of Review Counts (10 points)

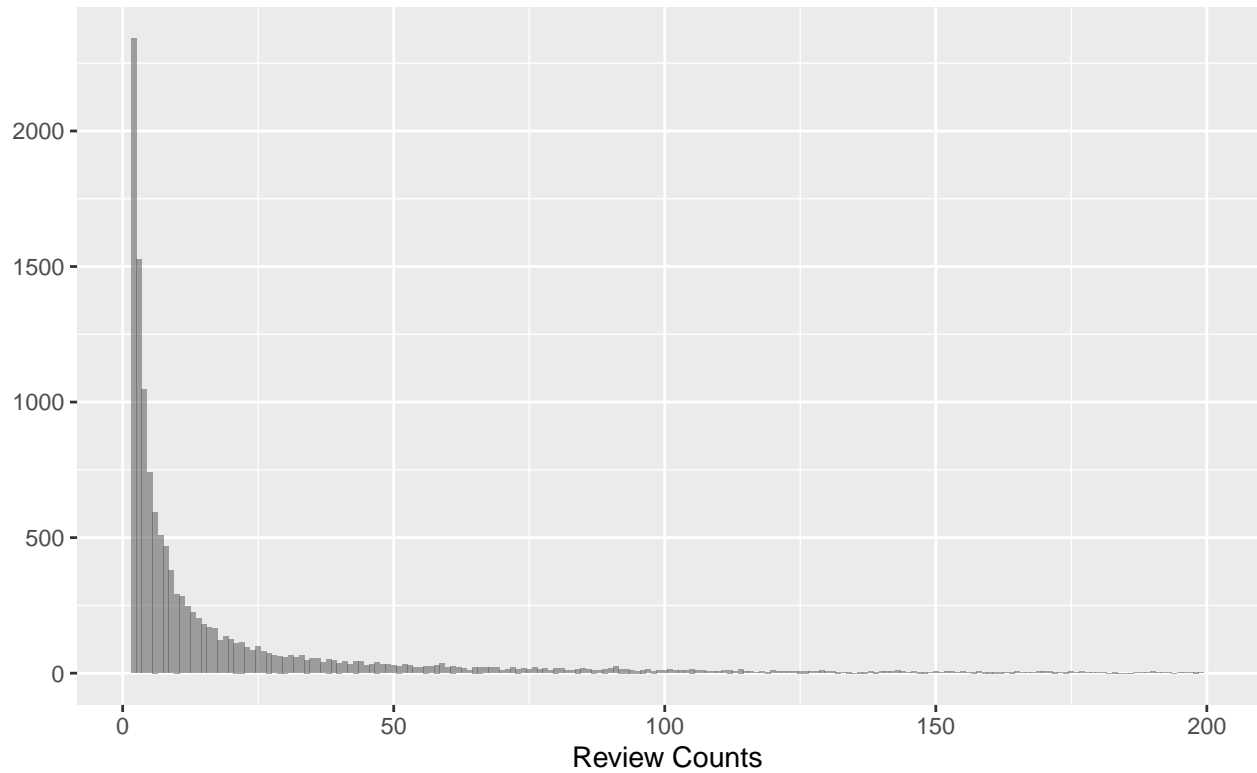
Histograms of review counts are plotted with the `qplot()` or `ggplot()` function. (Use `binwidth=10`)

Histograms of Review Counts

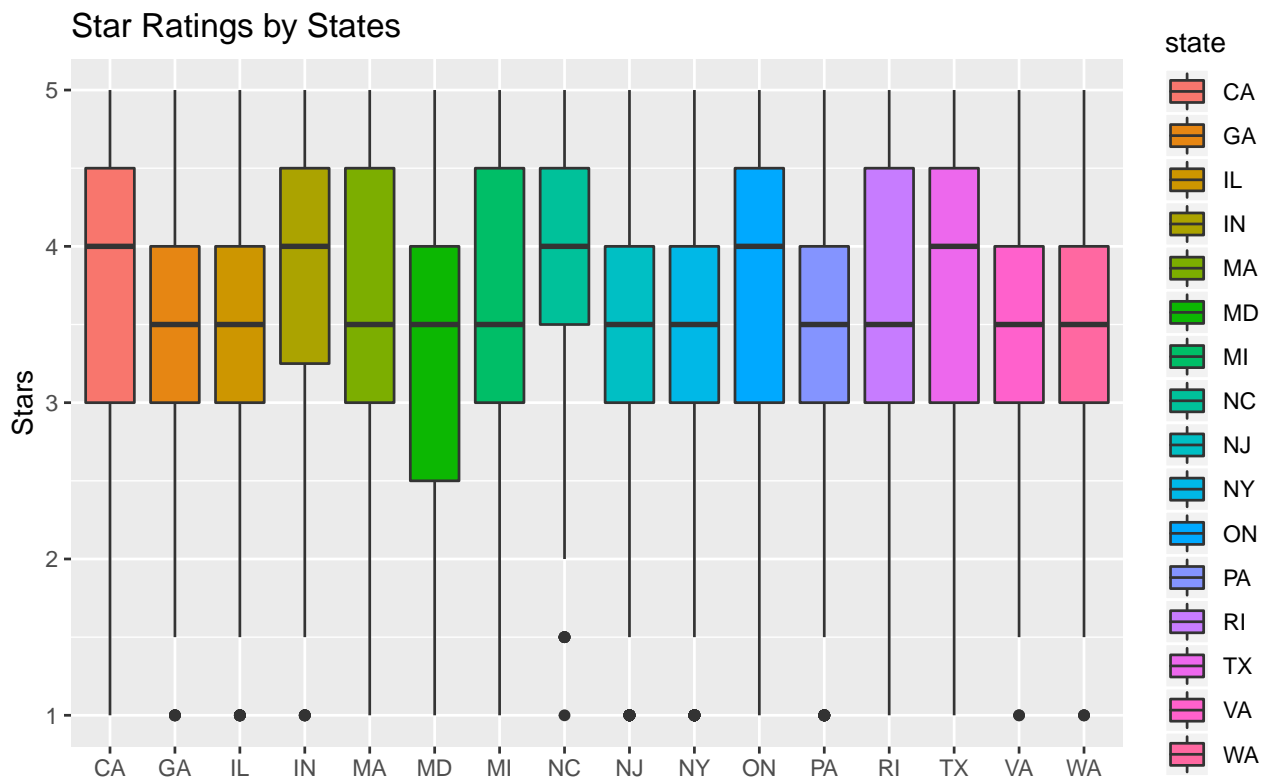


We can see that the distribution of review counts has a long tail. To zoom in on the bars to the left of the 200 mark, we use the **data.table** syntax or the **subset()** function to select just the data with review count ≤ 200 . And then plot the histogram again with **binwidth=1**.

Histograms of Review Counts (Filtered)



Question 8: Boxplot of Star Ratings by States (10 points)



Question 9: Bar Chart of Number of Businesses by State (10 points)

The states should be ordered by decreasing height of bars. Use the **reorder()** function.

