SI 618 Homework 7

Author: Ceren Budak

Loading and Cleaning Data (5 points)

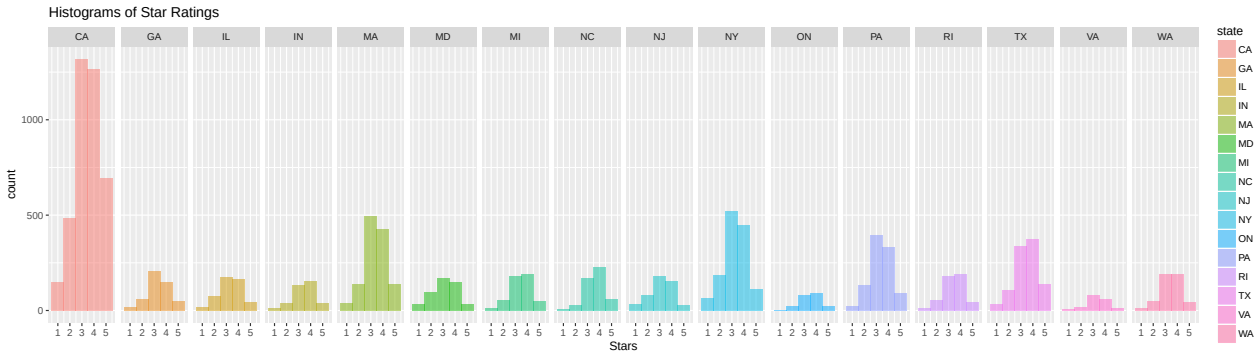
First the TSV data file created in part 1 is loaded into a R data frame using the **read.table()** function. The city, state and main_category columns should be converted to factors. Then listwise deletion (http://en.wikipedia.org/wiki/Listwise_deletion) is applied to remove records with missing data (use the **na.omit()** function). Then the data.frame is converted to a data.table. Here is the summary of the data table:

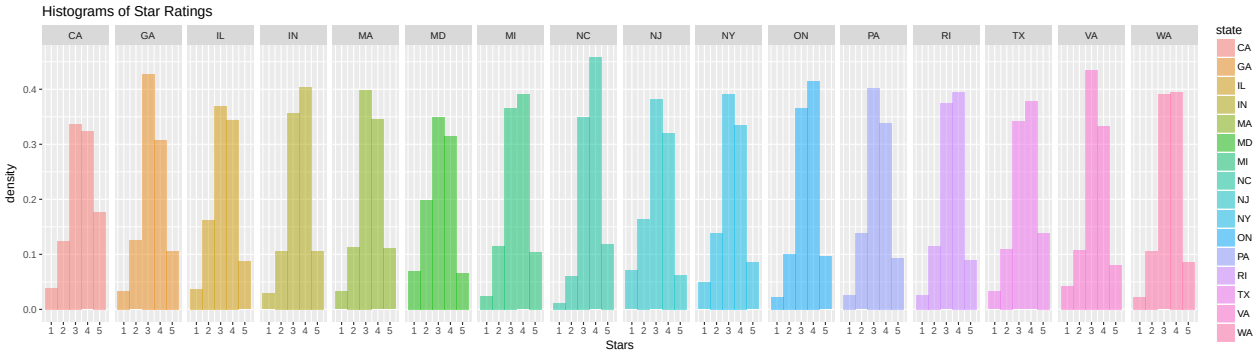
```
##      name                        city          state          stars      
##  Length:13137       Los Angeles    : 944   CA     :3917   Min.   :1.000  
##  Class :character   Cambridge      : 924   NY     :1336   1st Qu.:3.000  
##  Mode  :character   Austin         : 493   MA     :1240   Median :3.500  
##                     Houston        : 492   TX     : 987   Mean   :3.628  
##                     Berkeley       : 491   PA     : 979   3rd Qu.:4.500  
##                     San Luis Obispo: 491   NC     : 494   Max.   :5.000  
##                     (Other)        :9302   (Other):4184                  
##   review_count             main_category
##  Min.   :   2.00   Food           :1658  
##  1st Qu.:   3.00   Shopping       : 502  
##  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)        :9409
```

Histograms of Star Ratings (10 points)

Histograms of star ratings are plotted with the **qplot()** or **ggplot()** function. Both actual counts and density plot are shown. (Use **binwidth=1**)

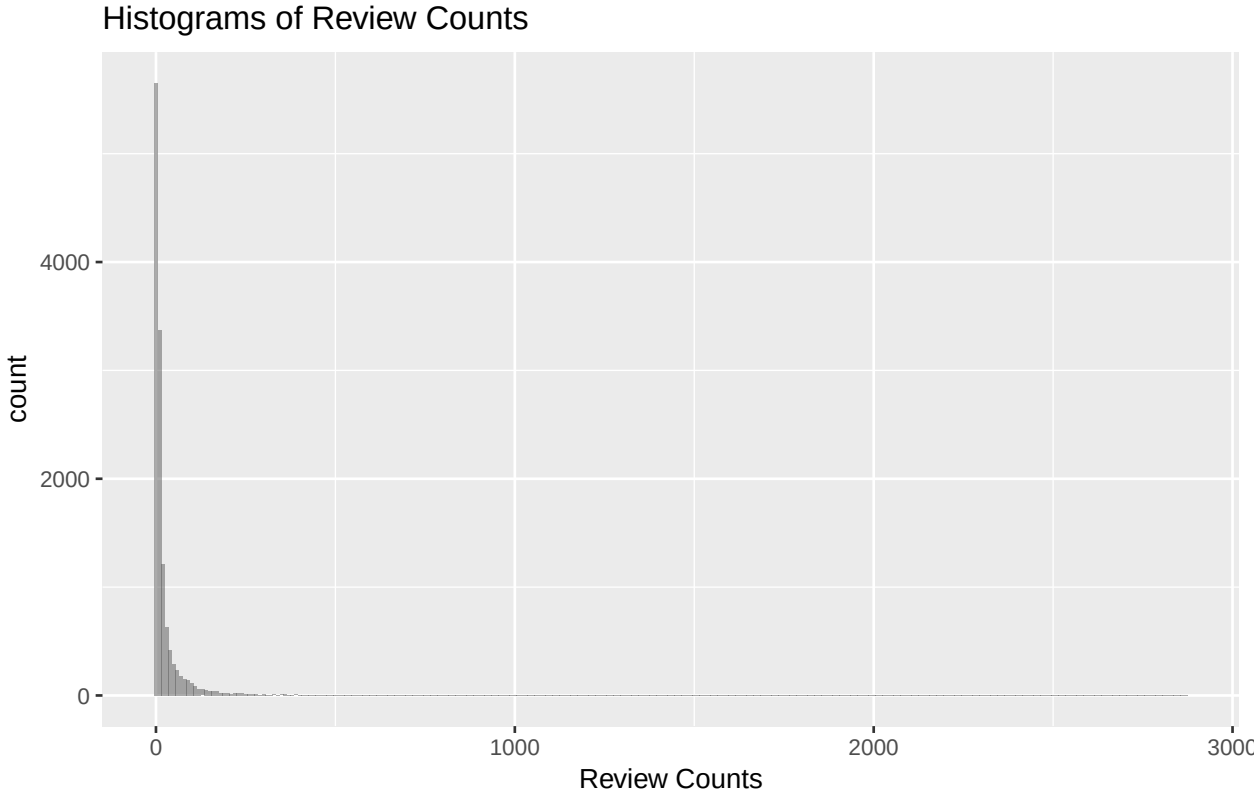
```
## Warning: package 'ggplot2' was built under R version 3.3.2
```



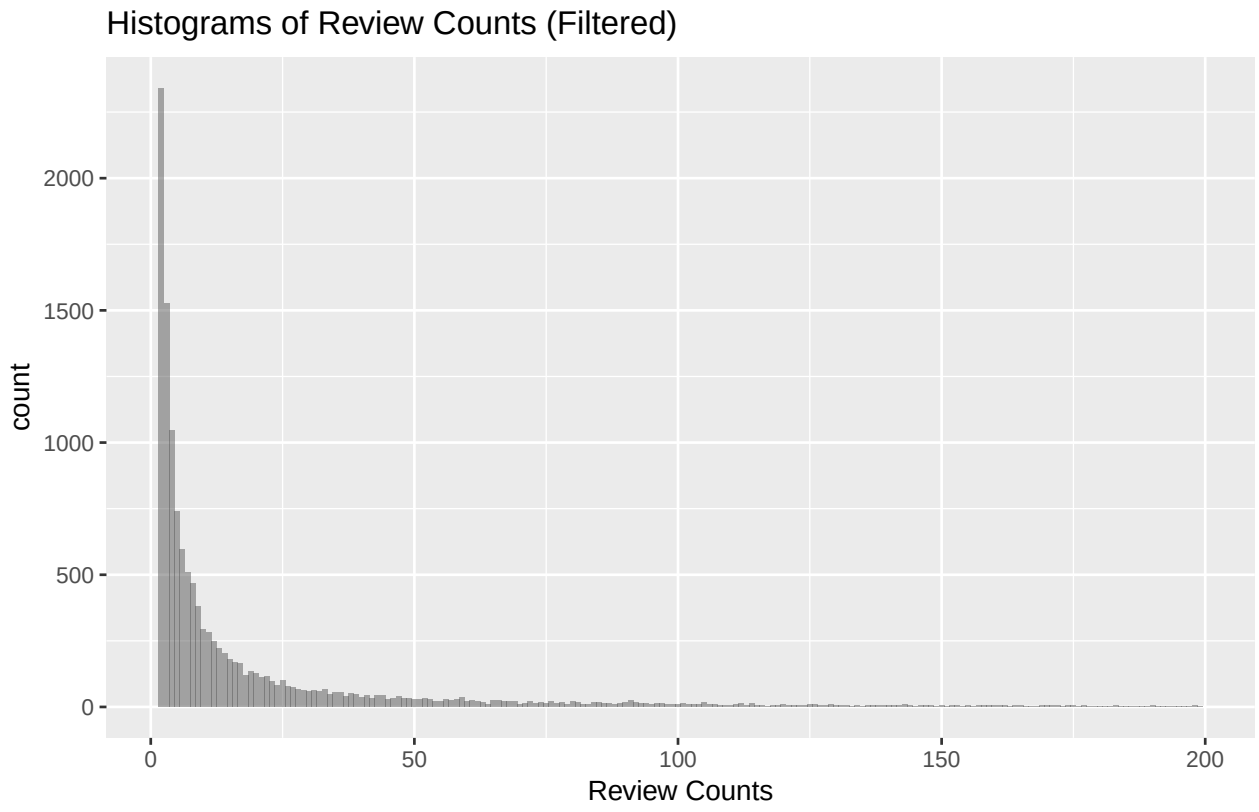


Histograms of Review Counts (10 points)

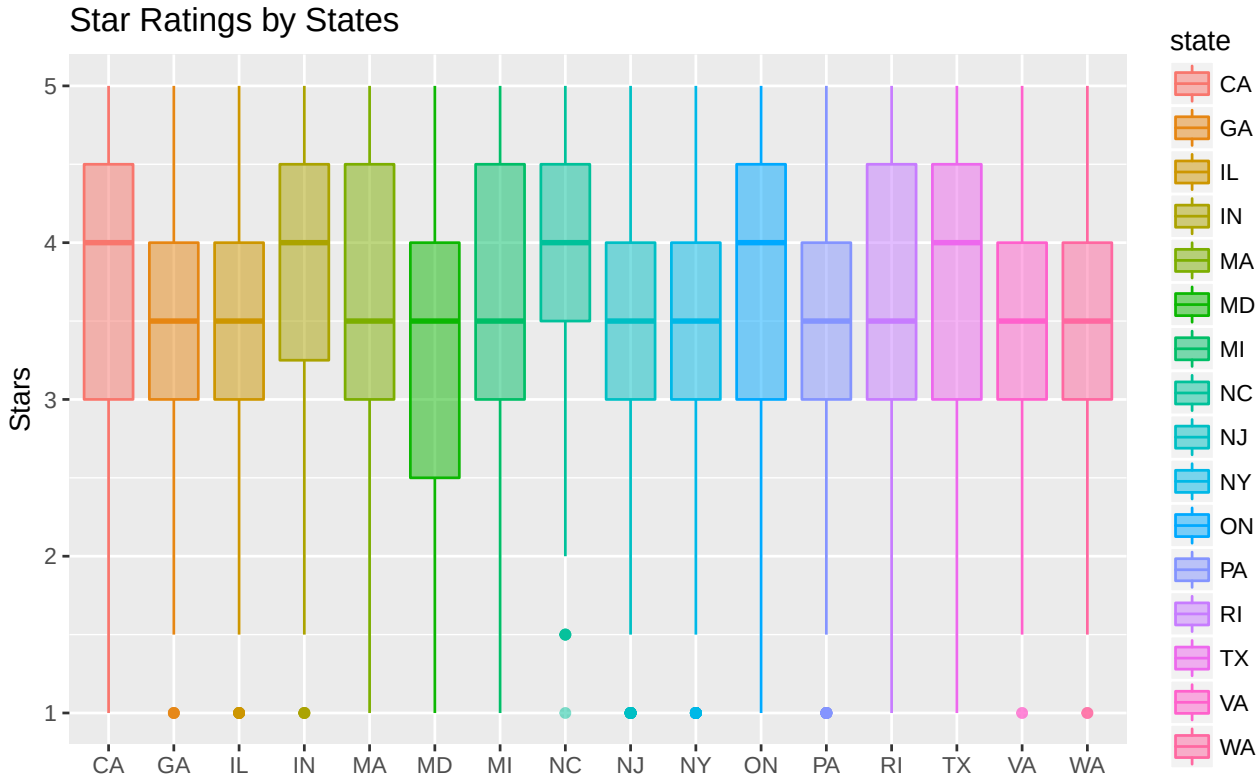
Histograms of review counts are plotted with the **qplot()** or **ggplot()** function. (Use **binwidth=10**)



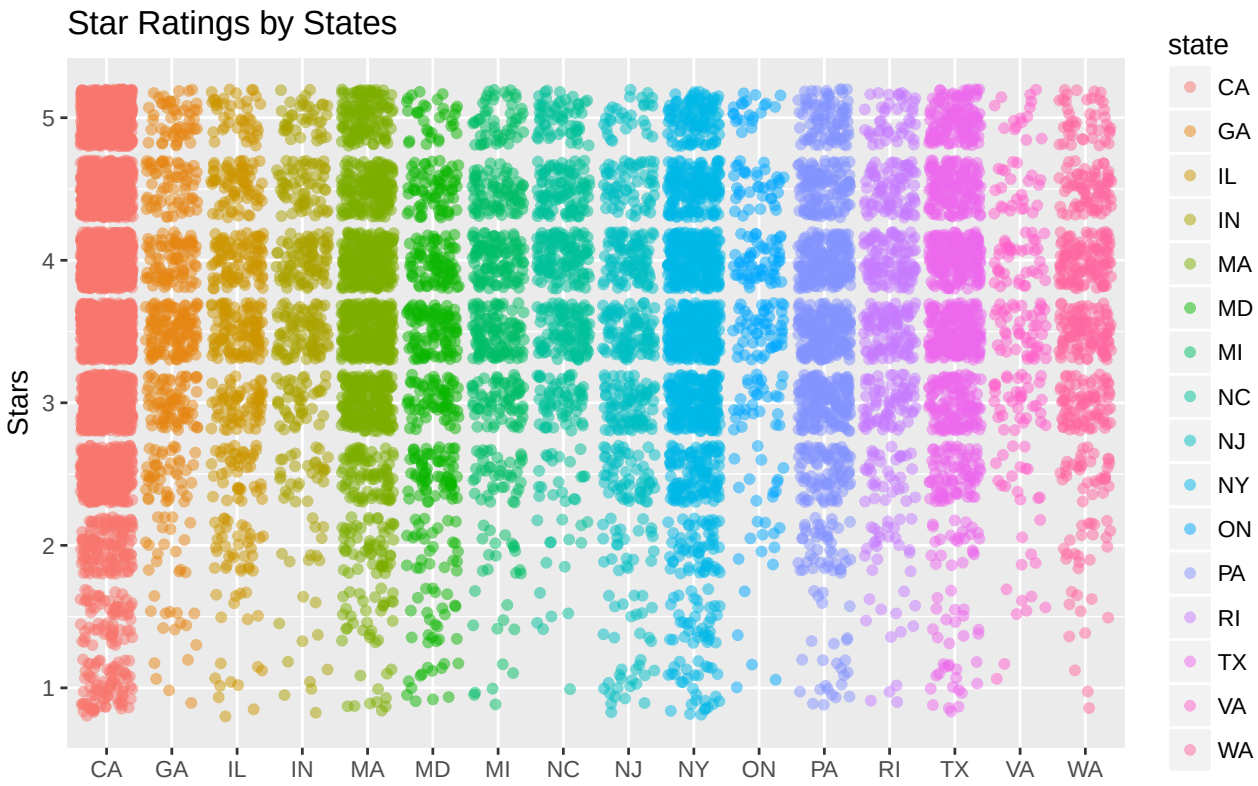
We can see that the distribution of review counts has a long tail. To zoom in on the bars to the left of the 200 mark, we use the **data.table syntax** or the **subset()** function to select just the data with review count <= 200. And then plot the histogram again with **binwidth=1**.



Boxplot of Star Ratings by States (10 points)

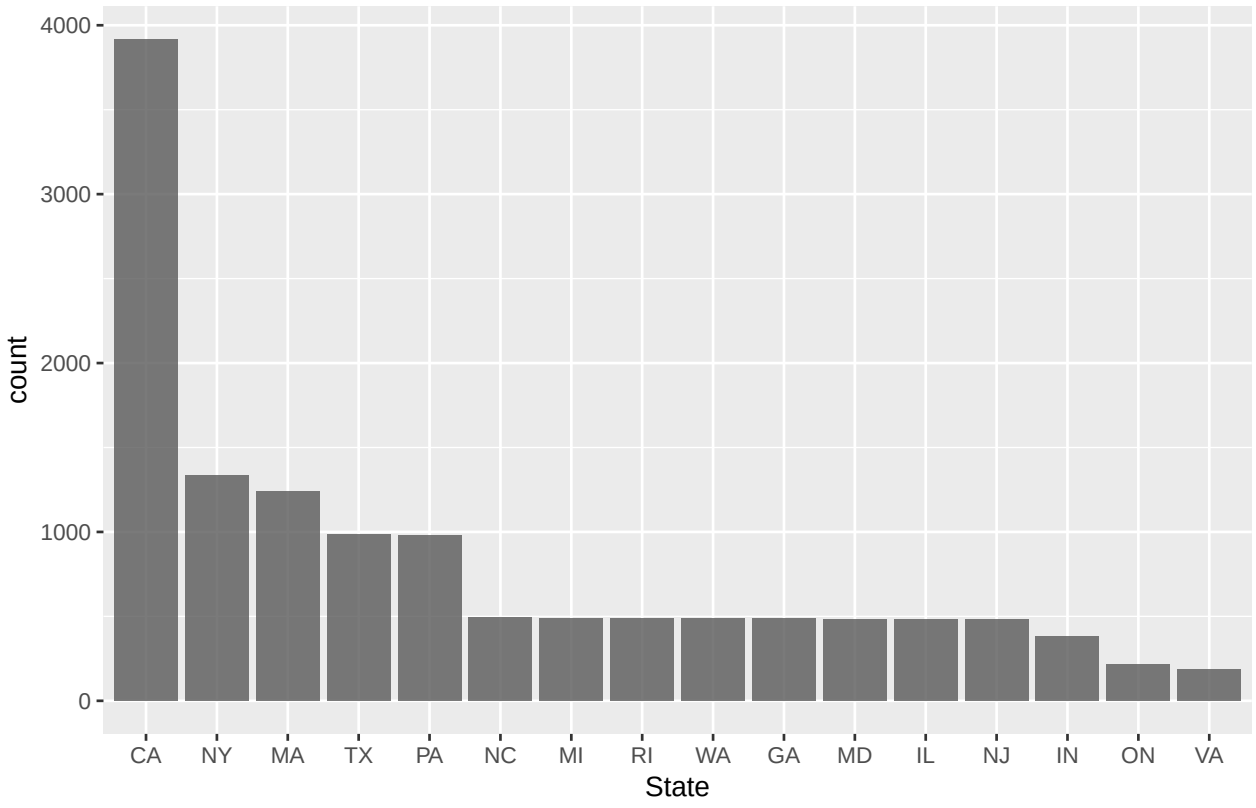


Jittered Plot of Star Ratings by States (10 points)

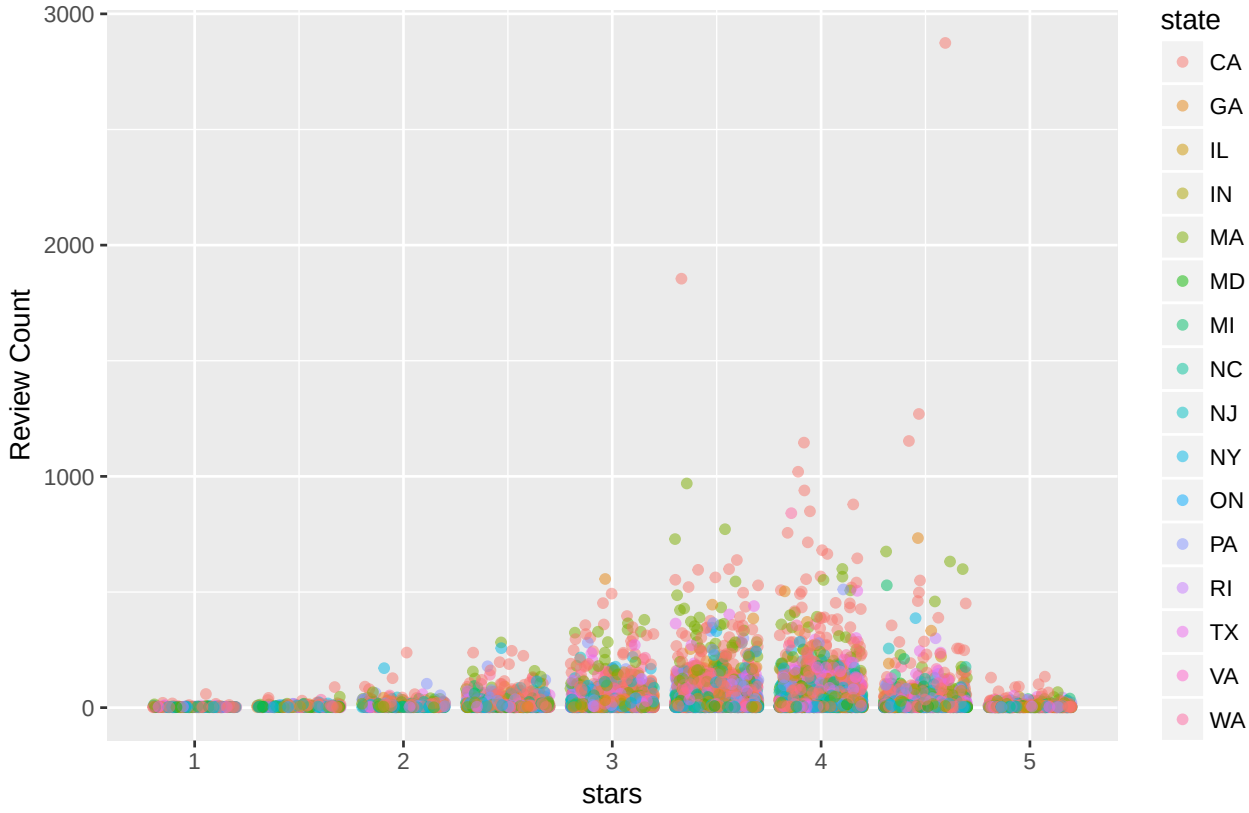


Bar Chart of Number of Businesses by State (10 points)

The states should be orderd by descreasing height of bars.



Jittered Scatterplot of Stars and Review Counts (10 points)



Slice and Dice Data using data.table syntax (or plyr)

Subsetting Data (10 points)

We first rank the business in each city for each main category. Then top 5 ranked businesses in each city for each main category are found.

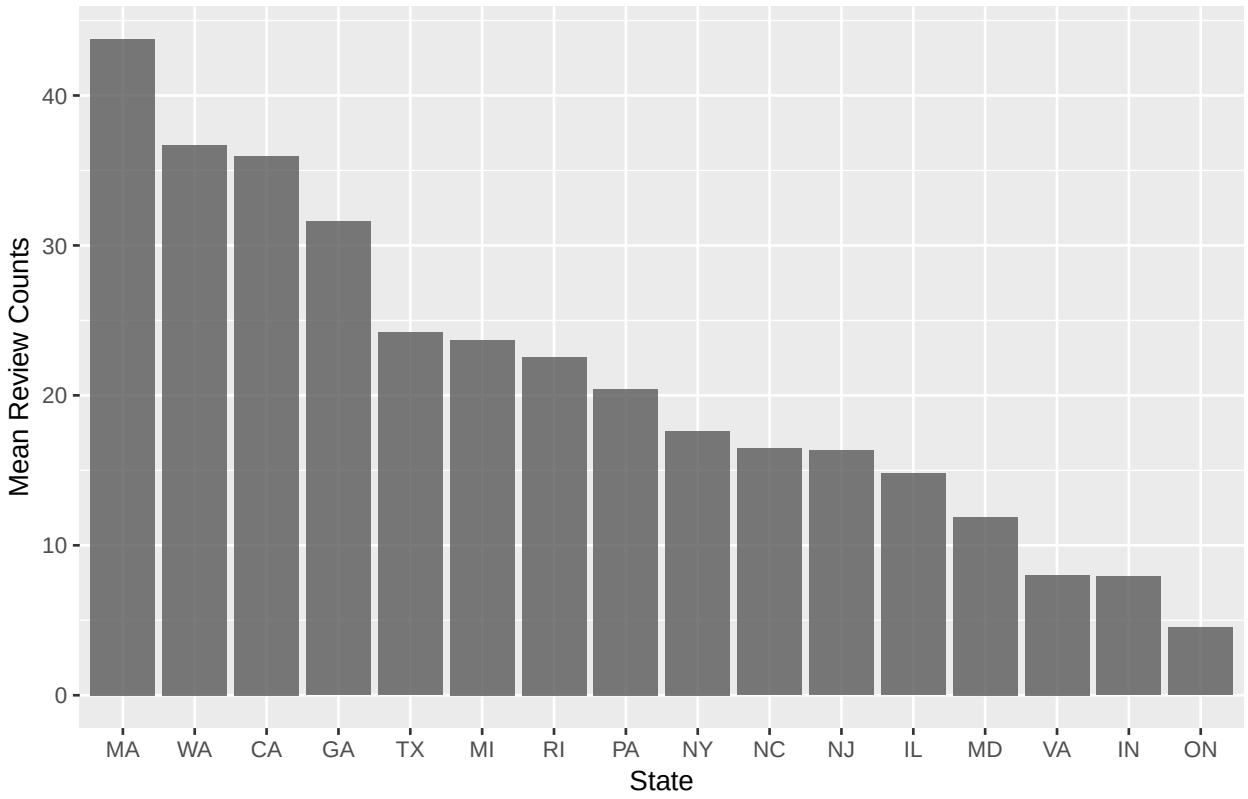
```
##                                     name           city state stars
##     1: Southern California Medical Group    Los Angeles    CA   3.5
##     2:             Harvard Square Shiatsu      Cambridge    MA   4.0
##     3:          Faith & Glory Collective      Kitchener    ON   4.0
##     4:           Von's Records & Posters West Lafayette    IN   3.5
##     5:                         JP's Java         Austin    TX   3.5
##    ---
## 13133:                        Yogurtland    Los Angeles    CA   4.0
## 13134:                     Bronz Body Tan    Los Angeles    CA   3.5
## 13135:                   The Metro Cafe      Ann Arbor    MI   3.5
## 13136:                  Follow The Honey      Cambridge    MA   4.5
## 13137:                     Lavaca Teppan         Austin    TX   3.5
##        review_count   main_category rank
##     1:            2 Medical Centers    3
##     2:            4         Massage    8
##     3:            2          Tattoo    1
##     4:            3    Music & DVDs    3
##     5:           85            Food   33
##    ---
## 13133:           65            Food   55
## 13134:            8         Tanning    2
## 13135:            2            Bars   13
## 13136:           29  Specialty Food    2
## 13137:           35        Japanese    1
```

Next, we are interested in is the top 5 business with main category of "Chinese" in each city. The data should be ordered by city names, and then by ranks. The result is listed below.

```
##                city                       name rank stars
##   1:        Amherst      Amherst Chinese Food    1   4.0
##   2:        Amherst             China Dynasty    2   2.5
##   3:      Ann Arbor                Kai Garden    1   3.5
##   4:      Ann Arbor     China Gate Restaurant    2   3.0
##   5:      Ann Arbor                    TK Wu    3   3.0
##  ---
## 138: West Lafayette           Szechuan Garden    1   3.5
## 139: West Lafayette               Happy China    2   3.0
## 140: West Lafayette          China One Buffet    3   3.0
## 141: West Lafayette Fu Lam Chinese Restaurant    4   3.0
## 142: West Lafayette                Rice Cafe    5   2.5
```

Summarize Data (10 points)

Next, we compute the mean review counts of all businesses for each state and plot the bar chart below.


Mean Review Counts
40
30
20
10
0
MA
WA
CA
GA
TX
MI
RI
PA
NY
NC
NJ
IL
MD
VA
IN
ON
State