SI 618 Homework 7

CHEN,Po-Heng (pohechen)

March 9, 2017

Loading and Cleaning Data (5 points)

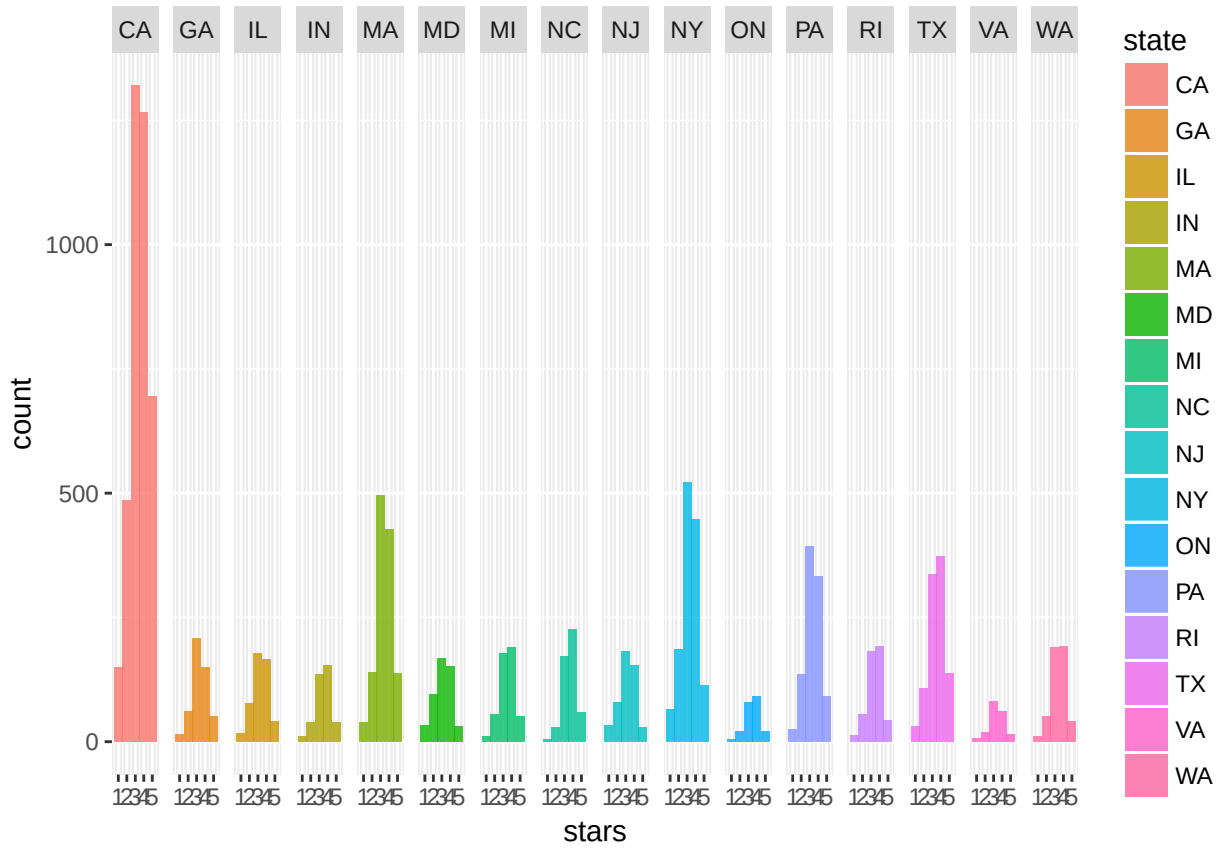
```
d = read.csv('businessdata.tsv', sep = '\t')
# I used read.csv to read because it's more robust than read.table
# and when using read.table, it creates error
d = na.omit(d)
summary(d)
```

```
##                                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
```

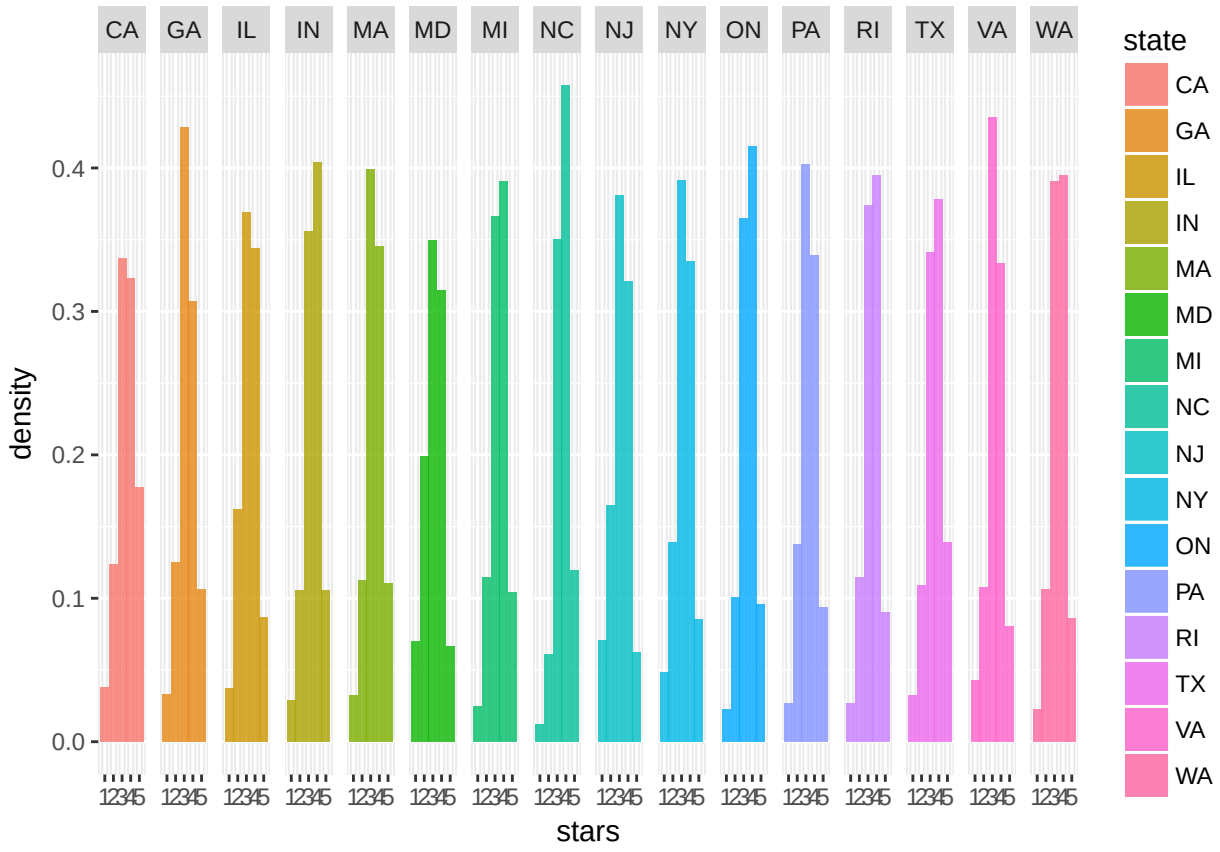
```
d = data.table(d)
```

Histograms of Star Ratings (10 points)

```
ggplot(d, aes(x = stars, fill = state)) +
    geom_histogram(binwidth = 1, alpha = 0.8) +
    facet_grid(. ~ state)
```

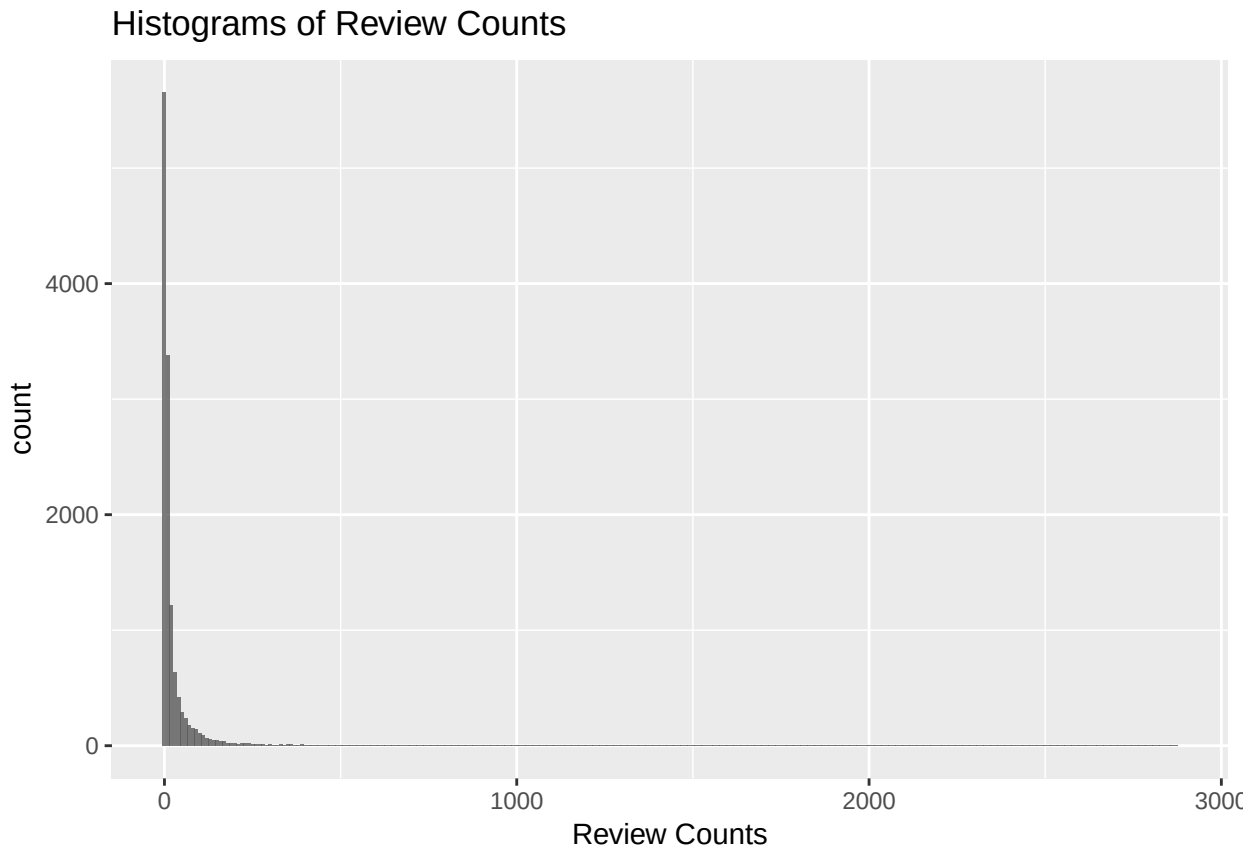



```
ggplot(d, aes(x = stars, fill = state)) +
    geom_histogram(aes(y=..density..), binwidth = 1, alpha = 0.8) +
    facet_grid(. ~ state)
```

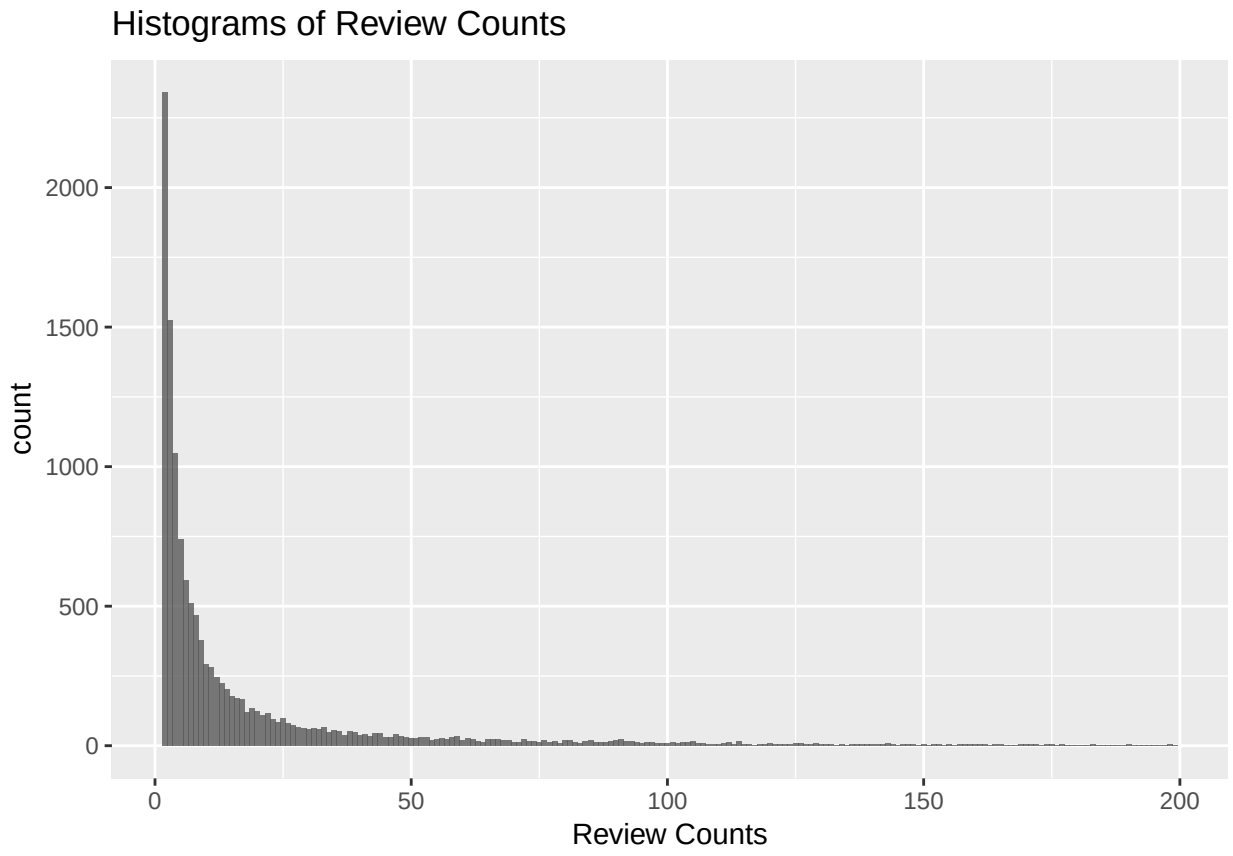


Histograms of Review Counts (10 points)

```
ggplot(d, aes(x = review_count)) +
    geom_histogram(binwidth = 10, alpha = 0.8) +
    ggtitle('Histograms of Review Counts') +
    labs(x = 'Review Counts')
```

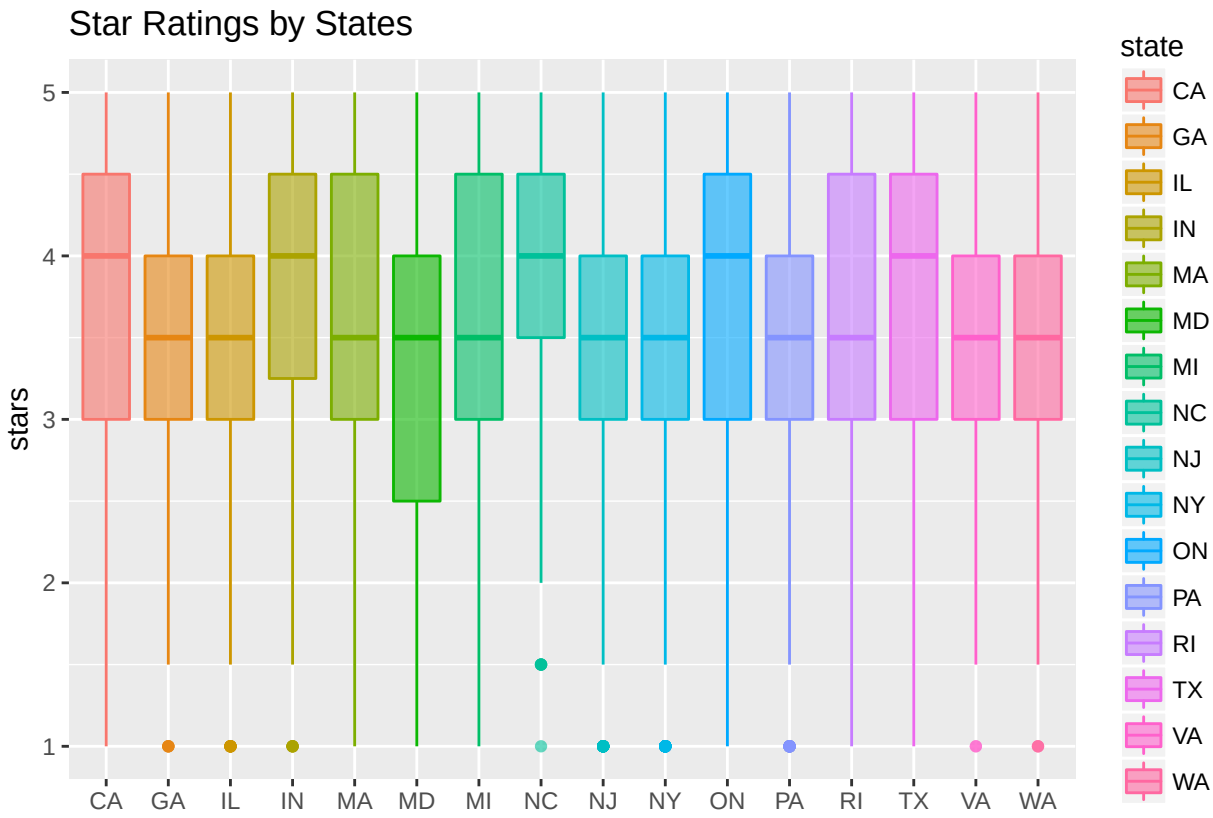



```
p = d[d$review_count <= 200, ]
ggplot(p, aes(x = review_count)) +
    geom_histogram(binwidth = 1, alpha = 0.8) +
    ggtitle('Histograms of Review Counts') +
    labs(x = 'Review Counts')
```



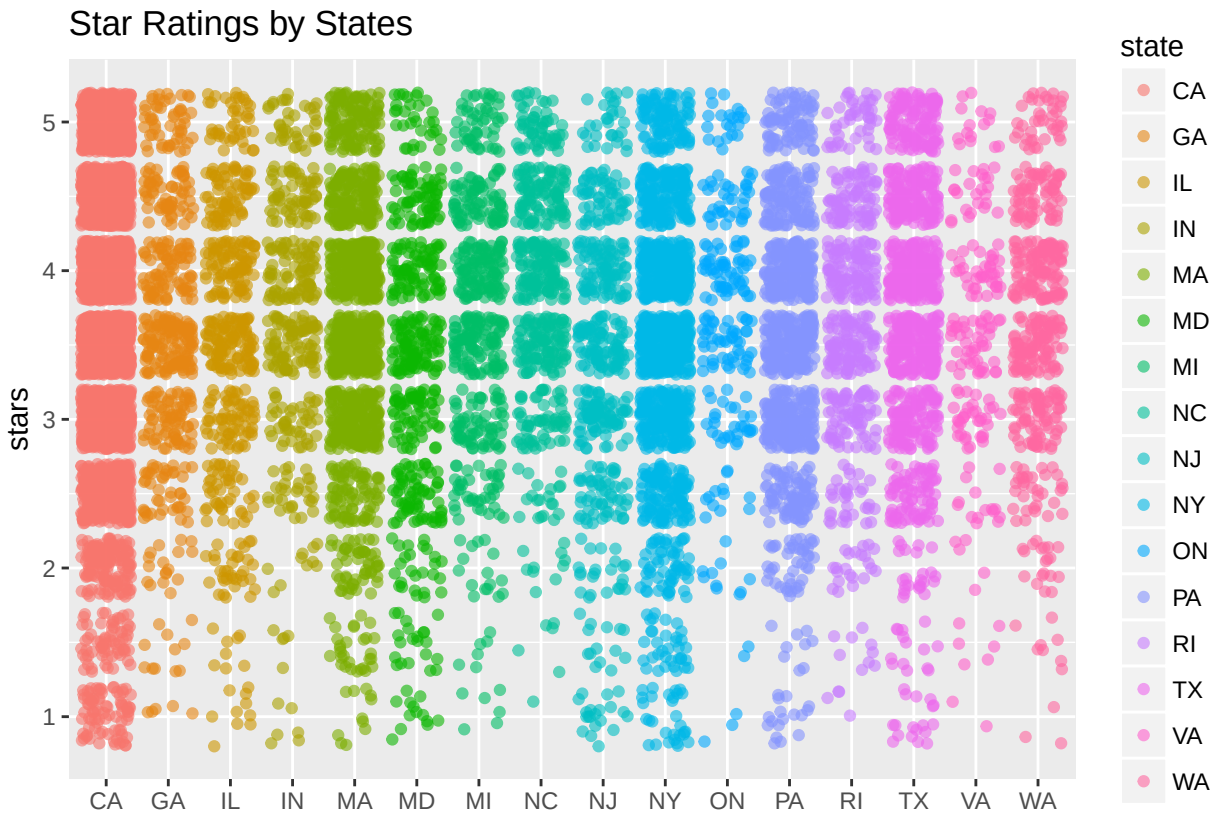
Boxplot of Star Ratings by States (10 points)

```
ggplot(d, aes(x = state, y = stars, col = state, fill = state)) +
    geom_boxplot(alpha = 0.6) +
    ggtitle('Star Ratings by States') +
    labs(x = '')
```



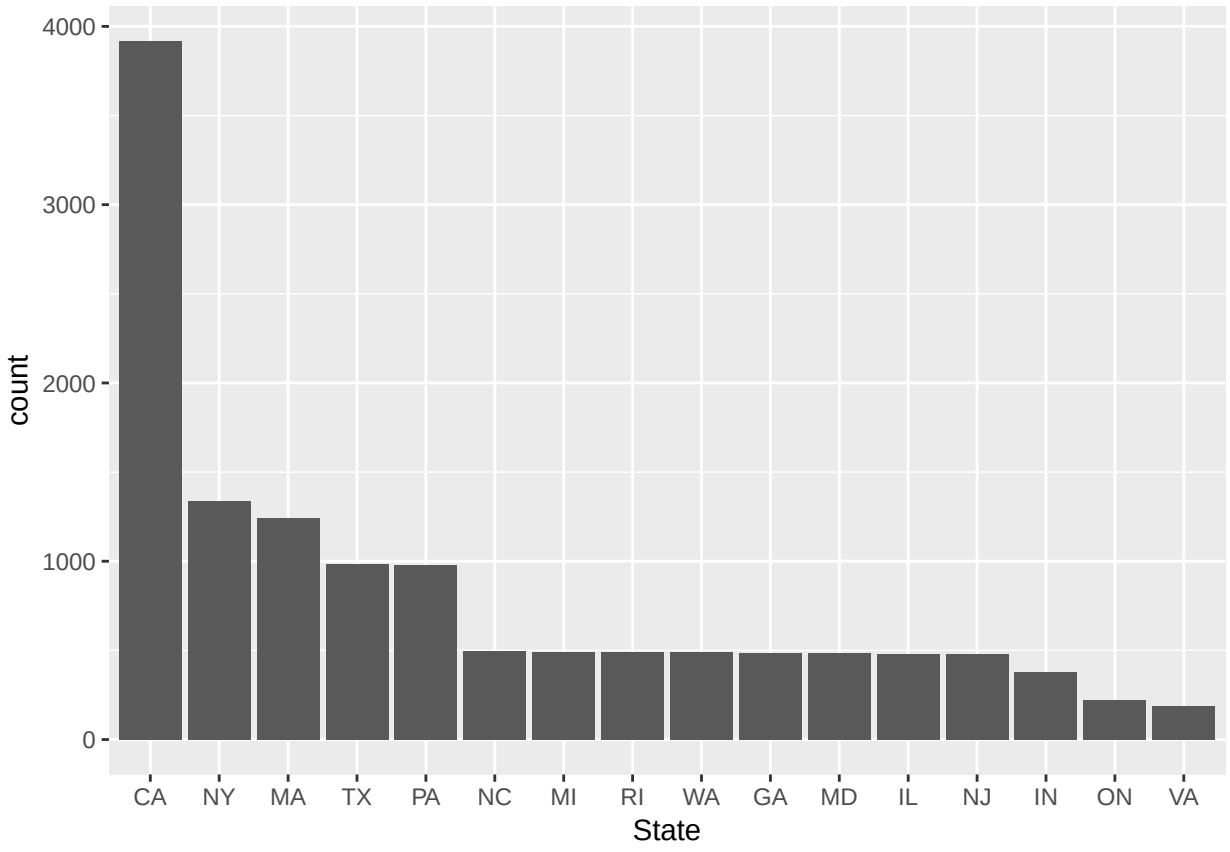
Jittered Plot of Star Ratings by States (10 points)

```
ggplot(d, aes(x = state, y = stars, col = state, fill = state)) +
    geom_jitter(alpha = 0.6) +
    ggtitle('Star Ratings by States') +
    labs(x = '')
```



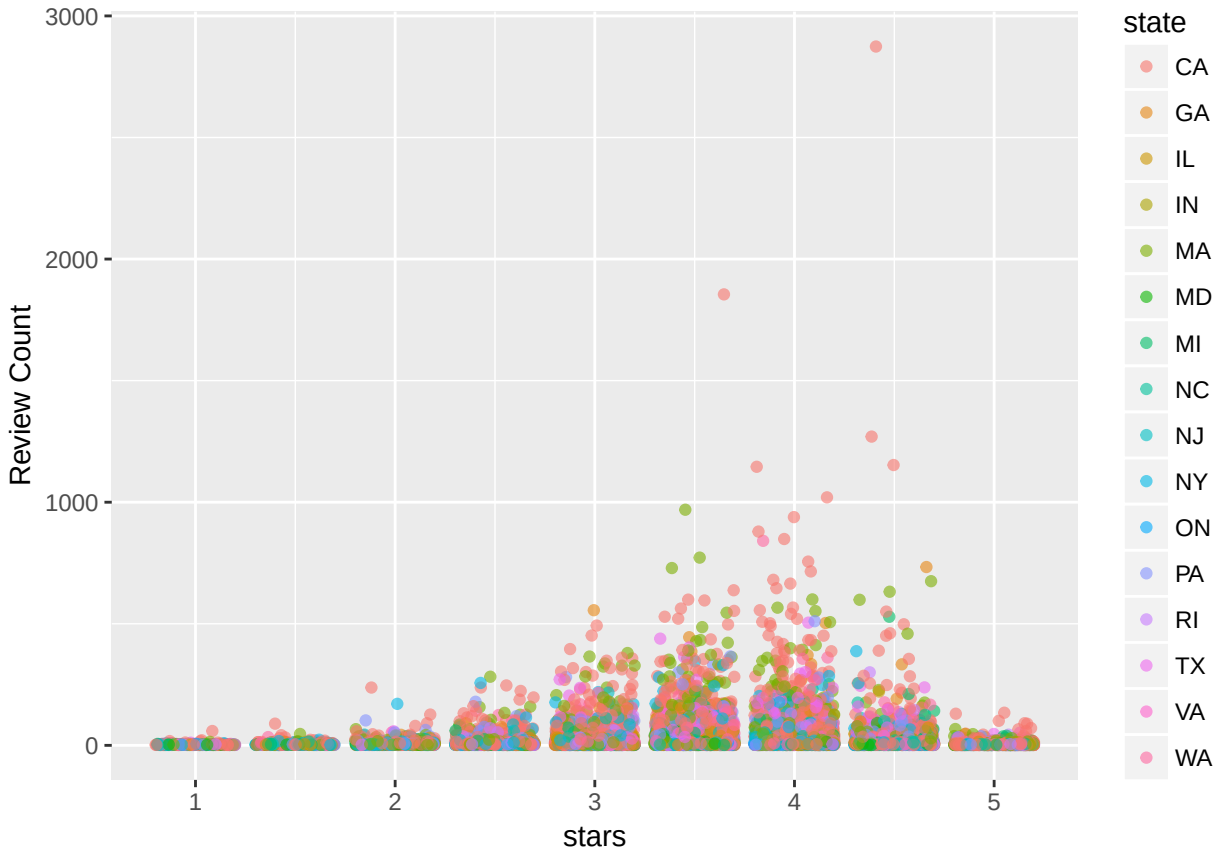
Bar Chart of Number of Businesses by State (10 points)

```
ggplot(d, aes(x = reorder(state, state, function(x)-length(x)))) +
    geom_bar() +
    labs(x = 'State')
```



Jittered Scatterplot of Stars and Review Counts (10 points

```
ggplot(d, aes(x = stars, y = review_count, col = state, fill = state)) +
    geom_jitter(alpha = 0.6) +
    labs(y = 'Review Count')
```



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

Subsetting Data (10 points)

```
tmp = d[, rank:= rank(-stars, ties.method = 'first'), by = .(city, main_category)]
print(tmp)
```

```
##                                     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
##    ---
```

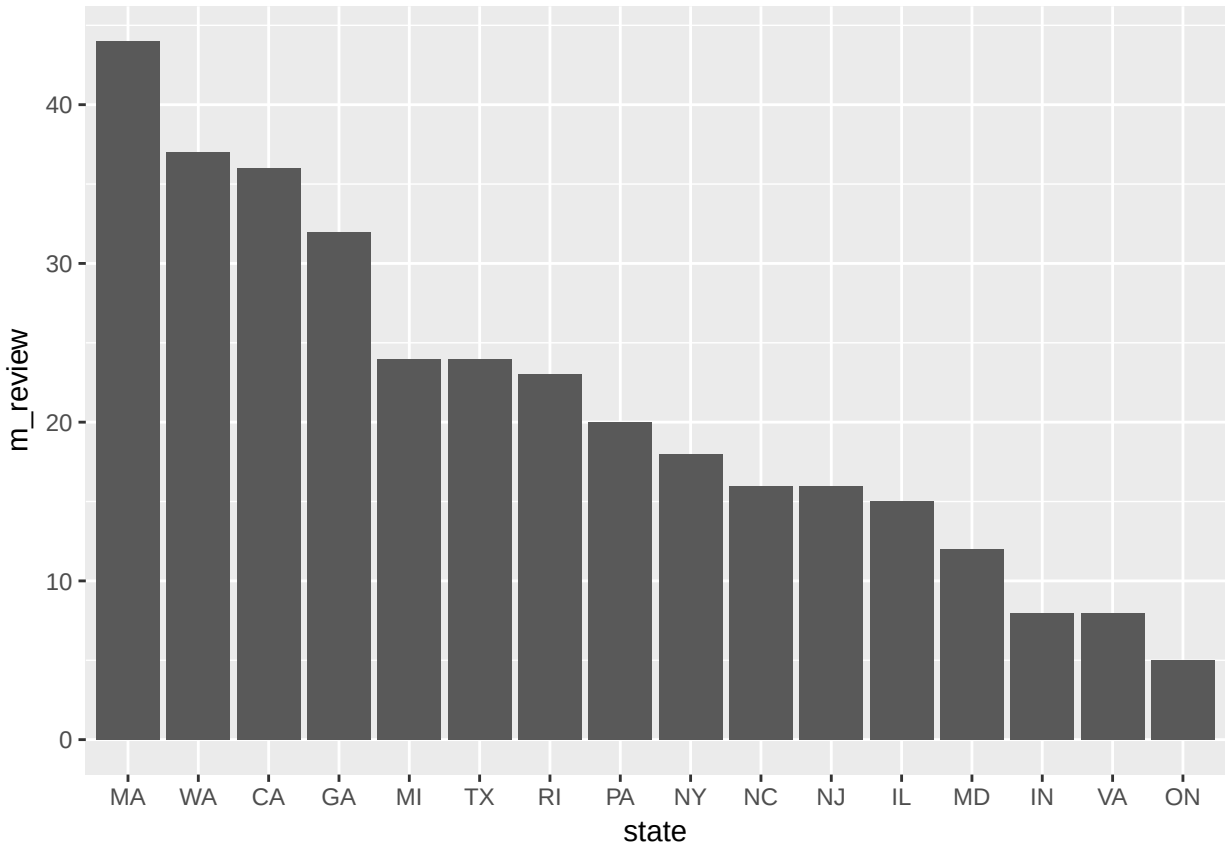
```
## 13133:           65            Food   55
## 13134:            8         Tanning    2
## 13135:            2            Bars   13
## 13136:           29  Specialty Food    2
## 13137:           35        Japanese    1
```

```
tmp = tmp[rank %in% 1:5 & main_category == 'Chinese', .(city, name, rank, stars) ]
tmp = tmp[order(city, rank),]
print(tmp)
```

```
##                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)

```
tmp = d[, .(m_review = round(mean(review_count))) ,by = .(state)]
tmp = transform(tmp, state = reorder(state, -m_review)) ## magic trick
ggplot(tmp, aes(x = state, y = m_review)) +
    geom_bar(stat = "identity")
```


m_review
40
30
20
10
0
MA
WA
CA
GA
MI
TX
RI
PA
NY
NC
NJ
IL
MD
IN
VA
ON
state