William Tran

CS 373

PUID: 0029823770

Homework 1

Part 1:

A. Sampling with replacement. Let the events of picking the apple be A and picking an orange be B.

a. Sample space: { A, B + A, 2B + A, 3B + A…, n * B + A }

Probability of picking the apple after i[th] tosses: $\left(\frac{2}{3}\right)^{i} * \frac{1}{3}$

b. Expected number of times: $\frac{1}{3}n$ with n being the total number of picks.

To just need one apple, the student is expected to pick 3 times.

c. Set of outcomes: $E = \{2x * B + A \mid x = 0,1,2 \ldots n\}$

Probability of E: $\frac{1}{3} + \left(\frac{2}{3}\right)^{2}\left(\frac{1}{3}\right) + \left(\frac{2}{3}\right)^{4}\left(\frac{1}{3}\right) + \left(\frac{2}{3}\right)^{6}\left(\frac{1}{3}\right) \ldots + \left(\frac{2}{3}\right)^{n}\left(\frac{1}{3}\right)$

$$= \frac{1}{3}\sum_{i=0}^{\infty}\left(\frac{2}{3}\right)^{2i} = \frac{1}{3}\sum_{i=0}^{\infty}\left(\frac{4}{9}\right)^{i}$$

(using Taylor series) $= \frac{1}{3} * \left(\frac{1}{1-\frac{4}{9}}\right) = \frac{3}{5} = 0.6$

B. Events:

$$A = \{5 + x,\ x + 5 \mid x = 1,2,3,4,5,6\} - \{5 + 5\}$$

$$B = \{2,4,6,8,10,12\};\ C = \{9,10,11,12\}$$

$$P(A) = \frac{2 * 1 * 6 - 1}{36} = \frac{11}{36}\ ; P(B) = \frac{1 + 3 + 5 + 5 + 3 + 1}{36} = \frac{1}{2}$$

$$P(C) = \frac{4 + 3 + 2 + 1}{36} = \frac{5}{18}$$

a. $(A \cap B) = \{5 + x,\ x + 5 \mid x = 1, 3, 5\} - \{5 + 5\}$
$P(A \cap B) = \frac{2*3-1}{36} = \mathbf{\frac{5}{36}} = 0.1388889$

b. $P(A \cap \neg B) = \{5 + x,\ x + 5 \mid x = 2, 4, 6\} = \frac{2*3}{36} = \frac{1}{6}$
$P(A \cup \neg B) = P(A) + P(\neg B) - P(A \cap \neg B)$
$= \frac{11}{36} + \left(1 - \frac{1}{2}\right) - \frac{1}{6} = \mathbf{\frac{23}{36}} = 0.6388889$

c. $(A \cap C) = \{4 + 5, 5 + 4, 5 + 5, 6 + 5, 5 + 6\}$
$P(A \cap C) = \mathbf{\frac{5}{36}} = 0.1388889$

d. $P(A \cap \neg C) = P(A) - P(A \cap C) = \frac{11}{36} - \frac{5}{36} = \frac{1}{6}$
$P(A \cup \neg C) = P(A) + P(\neg C) - P(A \cap \neg C)$
$= \frac{11}{36} + \left(1 - \frac{5}{18}\right) - \frac{1}{6} = \mathbf{\frac{31}{36}} = 0.8611111$

e. $P(A \cap B \cap C) = \frac{1}{36}$; $P(B \cap C) = \frac{4}{36}$
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C)$
$+ P(A \cap B \cap C) = \frac{11+18+10-5-5-4+1}{36} = \frac{26}{36} = \frac{13}{18} = 0.722222$

C. <u>1 red die, 2 yellow dice and 3 blue dice.</u>
a. Probability of selecting yellow: $\frac{1}{3}$
b. A: yellow, B: 6 is observed
$P(A) = \frac{1}{3}$, $P(B) = \frac{1}{6}$, $P(A \cap B) = \frac{1}{18}$
$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{18} * 6 = \frac{1}{3}$
$P(A|B) = P(A)$ here is because these 2 events are independent.

D. $P(A) = 0.55$, $P(B) = 0.03$, $P(A \cap B) = 0.03 - 0.0055 = 0.0245$
a. $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.0245}{0.03} = 0.816667$

b. $P(\neg B|\neg A) = \frac{P(\neg B \cap \neg A)}{P(\neg A)} = \frac{0.45-0.0055}{0.45} = 0.9877778$

c. $P(B \cap A) = P(A|B) * P(B) = 0.15 * 0.03 = 0.0045$
$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{0.0045}{1-0.57} = 0.010465$

E. $P(\neg d_1) = 0.02$; $P(\neg d_2) = 0.04$; $P(\neg d_3) = 0.05$; $P(\neg d_4) = 0.07$

a. $P(W) = 1 - 0.02 * 0.04 * 0.05 * 0.07 = 0.9999972$

b. $P(A) = P(i) + P(ii) - P\big((i) \cap (ii)\big)$
$= (0.93 * 0.05) + (0.98 * 0.96) - (0.93 * 0.05 * 0.98 * 0.96)$
$= 0.9435528$

c. $P(A \mid d_4) = \frac{P(A \cap d_4)}{d_4} = \frac{(1*0.05)+(0.98*0.96)-(1*0.05*0.98*0.96)}{1} = 0.94376$

F. <u>Roll 2 standard 6-sided dice:</u>

a. $E[X] = \sum_{i=2}^{12}(r_1 + r_2) * p(r_1 + r_2)$; $(r_1 + r_2) = i$ and r_1 being the number on roll 1 and r_2 being the number on roll 2.
For $p(r_1 + r_2)$:
$p(2) = p(12) = \frac{1}{36}$; $p(3) = p(11) = \frac{2}{36}$; $p(4) = p(10) = \frac{3}{36}$;
$p(5) = p(9) = \frac{4}{36}$; $p(6) = p(8) = \frac{5}{36}$; $p(7) = \frac{6}{36}$.

$E[X] = \frac{1}{36}(2 + 12) + \frac{2}{36}(3 + 11) + \frac{3}{36}(4 + 10) + \frac{4}{36}(5 + 9) + \frac{5}{36}(6 + 8) + \frac{6*7}{36}$

$= \frac{(1+2+3+4+5) * 14 + 42}{36} = \frac{252}{36} = 7.$

Conclusion: the expected value of the sum of the rolls is **E[X] = 7**.

b. $Var(X) = E[\ (x - E[X])^2\ ]$

$= \frac{1}{36}((-5)^2 + 5^2) + \frac{2}{36}((-4)^2 + 4^2) + \frac{3}{36}((-3)^2 + 3^2) + \frac{4}{36}((-2)^2 + 2^2)$
$+ \frac{5}{36}((-1)^2 + 1^2) + 0$

$$= \frac{1*50 + 2*32 + 3*18 + 4*8 + 5*2}{36} = \frac{210}{36} = 5.83333$$

The variance of the sum of the rolls is **Var(X) = 5.833.**

c. X_{max} = 12. Expected value of X_{max}:

$E[X_{max}] = 12 * \frac{1}{36} = \frac{1}{3} = 0.33333$

$P(\text{sum} = 7) = \frac{6}{36} = \frac{1}{6} = 0.16667$

Part 2: R Code

- **(3) Data Import and Summarization:**

a. Summary of the data:

```
> view(yelp)
> summary(yelp)
 business_id            name           fullAddress            city              state              latitude
 Length:24813       Length:24813       Length:24813       Length:24813       Length:24813       Min.   :32.88
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:33.54
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :36.03
                                                                                                Mean   :37.53
                                                                                                3rd Qu.:40.41
                                                                                                Max.   :55.99

   longitude           stars         reviewCount         checkins         open          neighborhoods
 Min.   :-115.370   Min.   :1.000   Min.   :   3.00   Min.   :    3   Mode :logical   Length:24813
 1st Qu.:-114.977   1st Qu.:3.000   1st Qu.:   8.00   1st Qu.:   16   FALSE:3580      Class :character
 Median :-111.924   Median :3.500   Median :  18.00   Median :   48   TRUE :21233     Mode  :character
 Mean   : -97.298   Mean   :3.544   Mean   :  49.03   Mean   :  166
 3rd Qu.: -80.807   3rd Qu.:4.000   3rd Qu.:  48.00   3rd Qu.:  155
 Max.   :   8.549   Max.   :5.000   Max.   :4578.00   Max.   :14203

  categories          alcohol           noiseLevel           attire            priceRange     delivery
 Length:24813       Length:24813       Length:24813       Length:24813       Min.   :1.000   Mode :logical
 Class :character   Class :character   Class :character   Class :character   1st Qu.:1.000   FALSE:14471
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :2.000   TRUE :3093
                                                                             Mean   :1.631   NA's :7249
                                                                             3rd Qu.:2.000
                                                                             Max.   :4.000
                                                                             NA's   :903
   ambience           parking          dietaryRestrictions waiterService     smoking          outdoorSeating
 Length:24813       Length:24813       Length:24813        Mode :logical   Length:24813       Mode :logical
 Class :character   Class :character   Class :character    FALSE:6208      Class :character   FALSE:10989
 Mode  :character   Mode  :character   Mode  :character    TRUE :10351     Mode  :character   TRUE :8698
                                                           NA's :8254                         NA's :5126


   caters        recommendedFor     goodForGroups   goodForKids
 Mode :logical   Length:24813       Mode :logical   Mode :logical
 FALSE:6503      Class :character   FALSE:2054      FALSE:506
 TRUE :5932      Mode  :character   TRUE :17078     TRUE :1283
 NA's :12378                        NA's :5681      NA's :23024
```

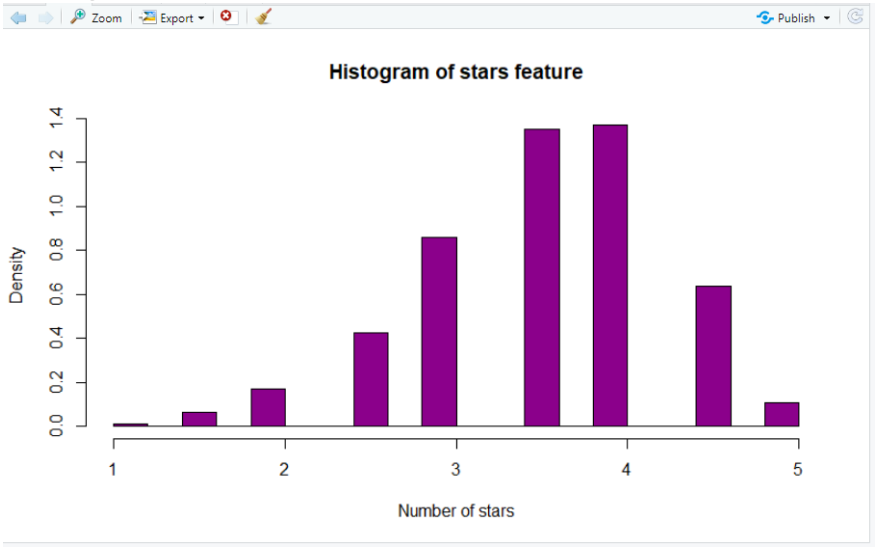
b. Names of the columns:

```
> names(yelp)
 [1] "business_id"          "name"              "fullAddress"       "city"
 [5] "state"                "latitude"          "longitude"         "stars"
 [9] "reviewCount"          "checkins"          "open"              "neighborhoods"
[13] "categories"           "alcohol"           "noiseLevel"        "attire"
[17] "priceRange"           "delivery"          "ambience"          "parking"
[21] "dietaryRestrictions" "waiterService"     "smoking"           "outdoorSeating"
[25] "caters"               "recommendedFor"    "goodForGroups"     "goodForKids"
>
```

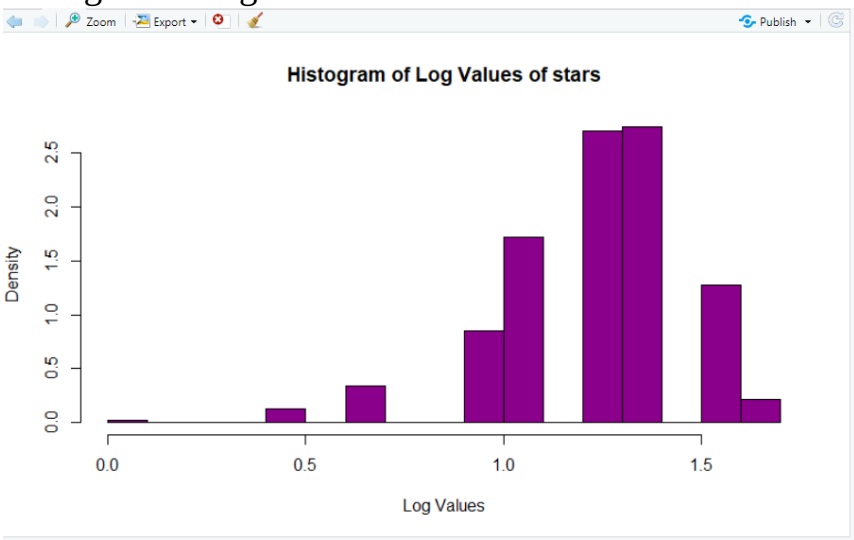
- **(4) 1D Plots:**

A. "stars" column:

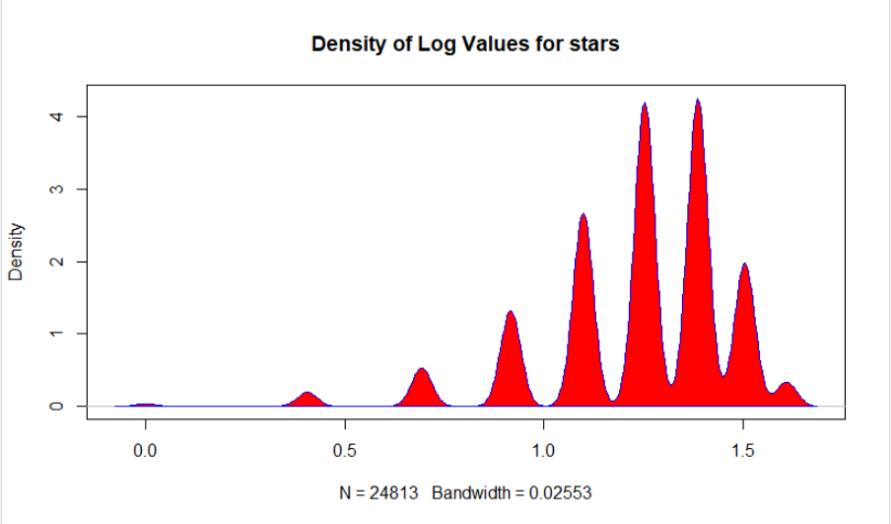
a. Histogram of stars:



b. Histogram of log values:

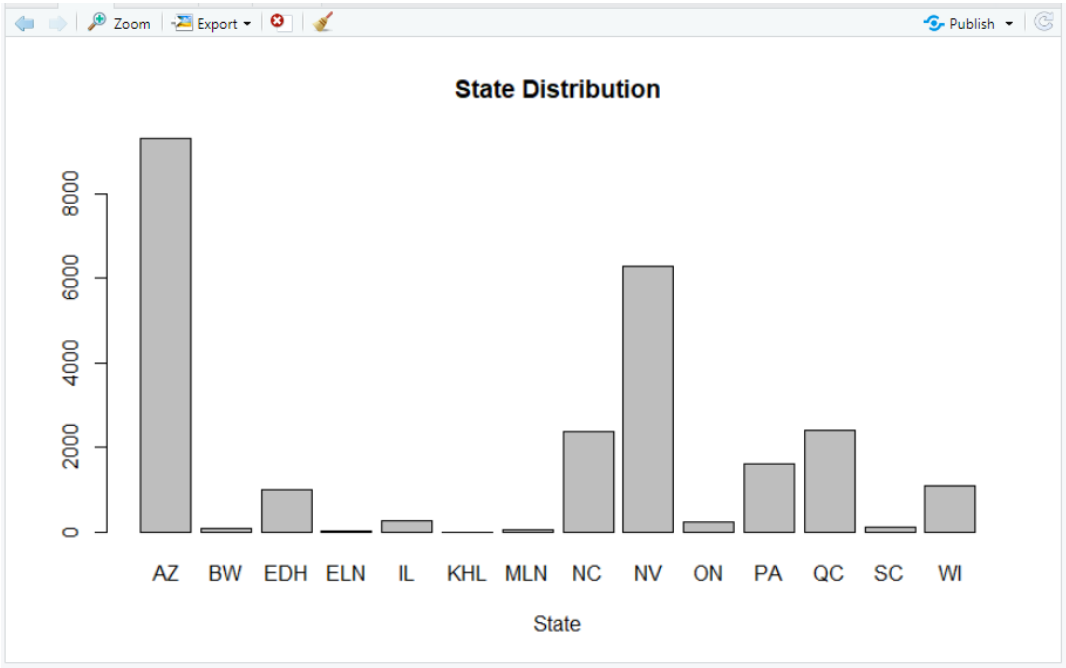


c. Density plot of log values:



d. All 3 graphs describe the distribution of the values of 'stars' attribute. They all have a similar general bell-like shape and are all left-skewed. However, the second graph is different from the first because the values are less scattered throughout the distribution. Graph 3 is different from the other two graphs because it describes density as a continuous curve, while the other two graphs have bars, describing discrete values rather than continuous values.

B. yelp$state:



This is the barplot of the ‘state’ attribute. Note: The function **names.arg** was not needed because the default of barplot() already puts the states in alphabetical order. Names of the states:

```
> names(states)
 [1] "AZ"  "BW"  "EDH" "ELN" "IL"  "KHL" "MLN" "NC"  "NV"  "ON"  "PA"  "QC"  "SC"  "WI"
```

- **(5) Sampling and transforming data:**

A. (all code is in appendix)

A fraction of the new 2 columns:

reliableReview	highStar
TRUE	FALSE
TRUE	FALSE
TRUE	TRUE
TRUE	FALSE
TRUE	FALSE
FALSE	FALSE
FALSE	FALSE
TRUE	FALSE
TRUE	TRUE
FALSE	FALSE
FALSE	FALSE
FALSE	FALSE
FALSE	FALSE
FALSE	FALSE
TRUE	FALSE
TRUE	FALSE

The name list of the columns before and after adding 2 new columns:

```
> names(yelp)
 [1] "business_id"          "name"             "fullAddress"      "city"
 [5] "state"                "latitude"         "longitude"        "stars"
 [9] "reviewCount"          "checkins"         "open"             "neighborhoods"
[13] "categories"           "alcohol"          "noiseLevel"       "attire"
[17] "priceRange"           "delivery"         "ambience"         "parking"
[21] "dietaryRestrictions"  "waiterService"    "smoking"          "outdoorSeating"
[25] "caters"               "recommendedFor"   "goodForGroups"    "goodForKids"
[29] "log_stars"
> names(yelp_new)
 [1] "business_id"          "name"             "fullAddress"      "city"
 [5] "state"                "latitude"         "longitude"        "stars"
 [9] "reviewCount"          "checkins"         "open"             "neighborhoods"
[13] "categories"           "alcohol"          "noiseLevel"       "attire"
[17] "priceRange"           "delivery"         "ambience"         "parking"
[21] "dietaryRestrictions"  "waiterService"    "smoking"          "outdoorSeating"
[25] "caters"               "recommendedFor"   "goodForGroups"    "goodForKids"
[29] "log_stars"            "reliableReview"   "highStar"
>
```

B. <u>Attribute ‘checkins’:</u>

a. Quantiles of ‘checkins’:

```
> # Part 5b
> # a/Quantiles
> quantile(yelp$checkins)
   0%   25%   50%   75%  100%
    3    16    48   155 14203
>
```

b. (code in appendix below)

c. Comparisons of the 6 attributes:

```
> summary(yelp$checkins)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      3      16      48     166     155   14203
> summary(yelp$stars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   3.500   3.544   4.000   5.000
> summary(yelp$noiseLevel)
   Length     Class      Mode
    24813 character character
> summary(yelp$priceRange)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   1.000   2.000   1.631   2.000   4.000     903
> summary(yelp$reviewCount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00    8.00   18.00   49.03   48.00 4578.00
> summary(yelp$goodForGroups)
   Mode   FALSE    TRUE    NA's
logical    2054   17078    5681
>
```

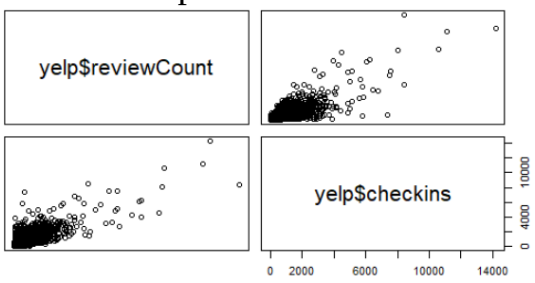
```
> summary(checkins_subset$checkins)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   5.000   8.000   8.739  12.000  16.000
> summary(checkins_subset$stars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   3.500   3.484   4.000   5.000
> summary(checkins_subset$noiseLevel)
   Length     Class      Mode
     6391 character character
> summary(checkins_subset$priceRange)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   1.000   2.000   1.674   2.000   4.000     663
> summary(checkins_subset$reviewCount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   7.000   9.004  11.000 230.000
> summary(checkins_subset$goodForGroups)
   Mode   FALSE    TRUE    NA's
logical     756    3734    1901
>
```

For the attribute ‘checkins’, the max value of the subset is the value at the first quantile of the whole data, which makes sense because the subset is taken from the first 25 percentile of ‘checkins’.

The other attributes do not have a lot of differences except for the fact that there is less data for the subset. For example, the summary of 2 ‘stars’ attributes show almost the same values, and ‘goodForGroups’ just have less TRUEs and FALSEs in the subset.

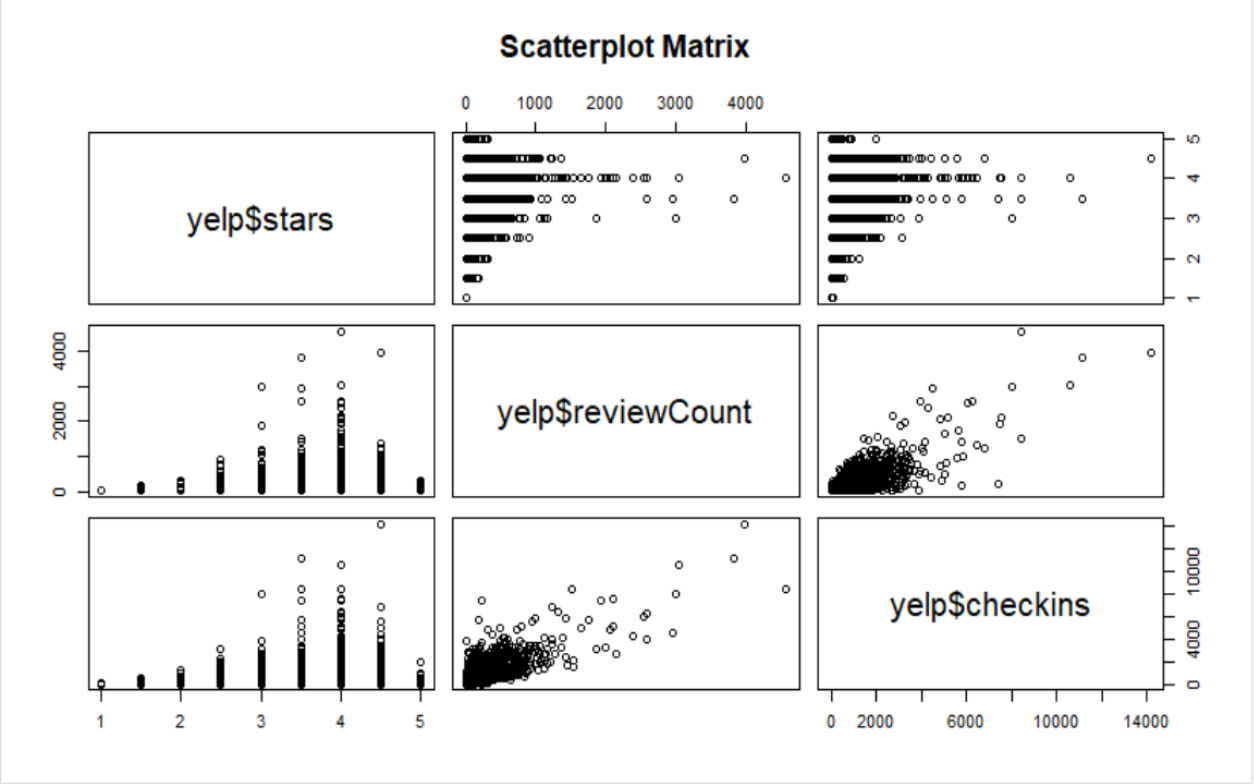
- **(6) 2D plots and correlations**

A. Scatterplot matrix of ‘stars’, ‘reviewCount’, and ‘checkins’:
Visually, ‘reviewCount’ and ‘checkins’ exhibit the most correlation because the plots between them are always linearly increasing as showed:



Between the other pairs of attributes, it is not always linearly increasing. This is expected because in reality, a restaurant only has more reviews when more customers check in to try the food.

Below is the whole scatterplot matrix for the three attributes:



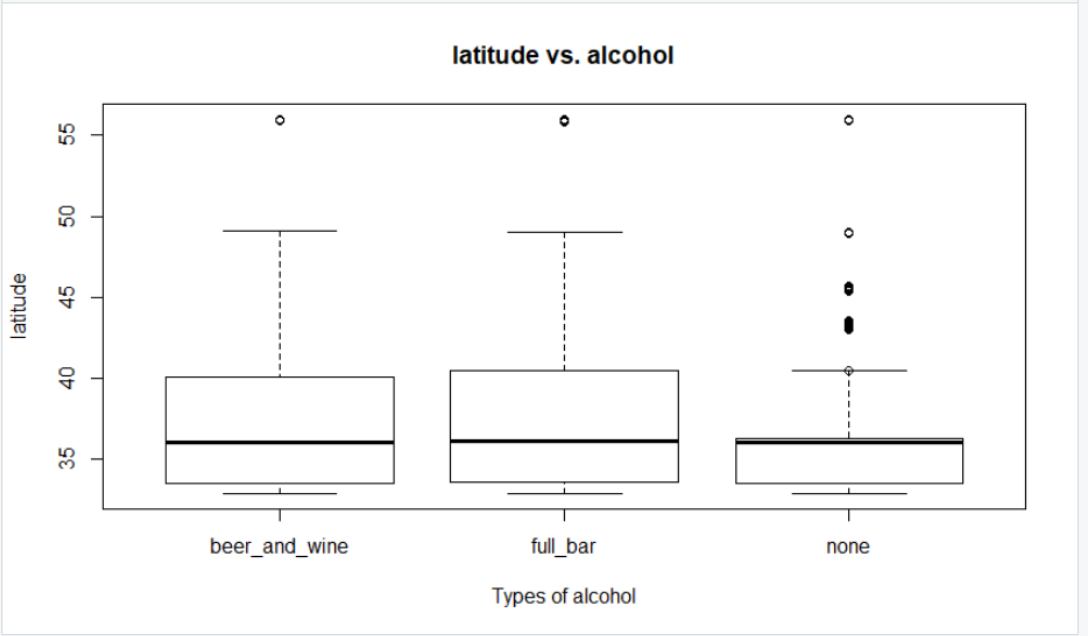
B. 9 pairwise correlations among the above three attributes:

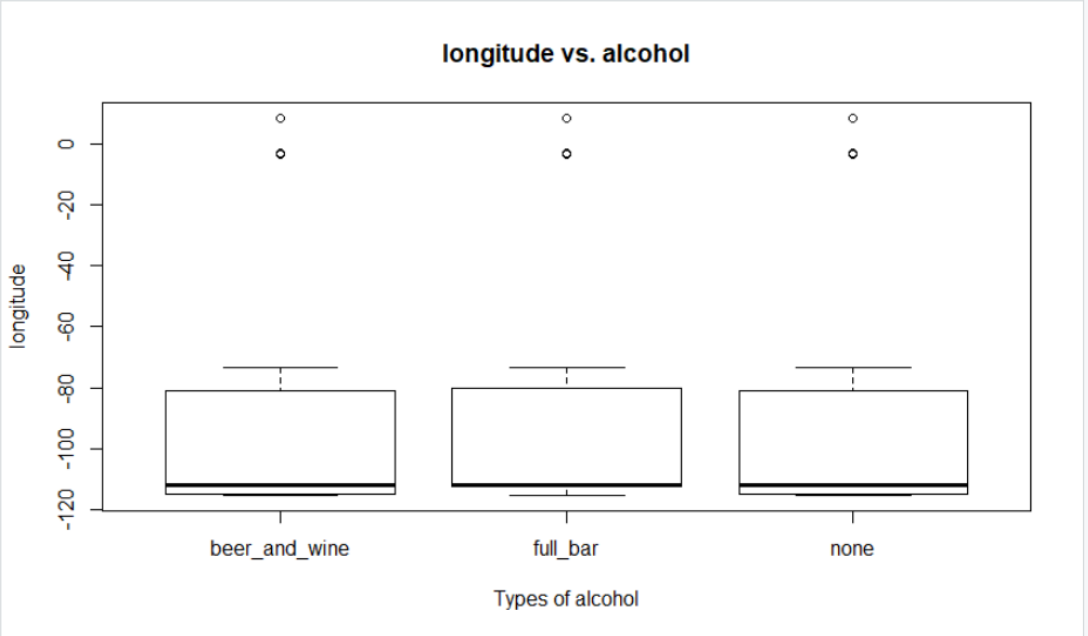
```
> # B. Correlation
> cor(yelp$stars, yelp$stars)
[1] 1
> cor(yelp$stars, yelp$reviewCount)
[1] 0.1070506
> cor(yelp$stars, yelp$checkins)
[1] 0.09440071
> cor(yelp$reviewCount, yelp$stars)
[1] 0.1070506
> cor(yelp$reviewCount, yelp$reviewCount)
[1] 1
> cor(yelp$reviewCount, yelp$checkins)
[1] 0.8274936
> cor(yelp$checkins, yelp$stars)
[1] 0.09440071
> cor(yelp$checkins, yelp$reviewCount)
[1] 0.8274936
> cor(yelp$checkins, yelp$checkins)
[1] 1
>
```

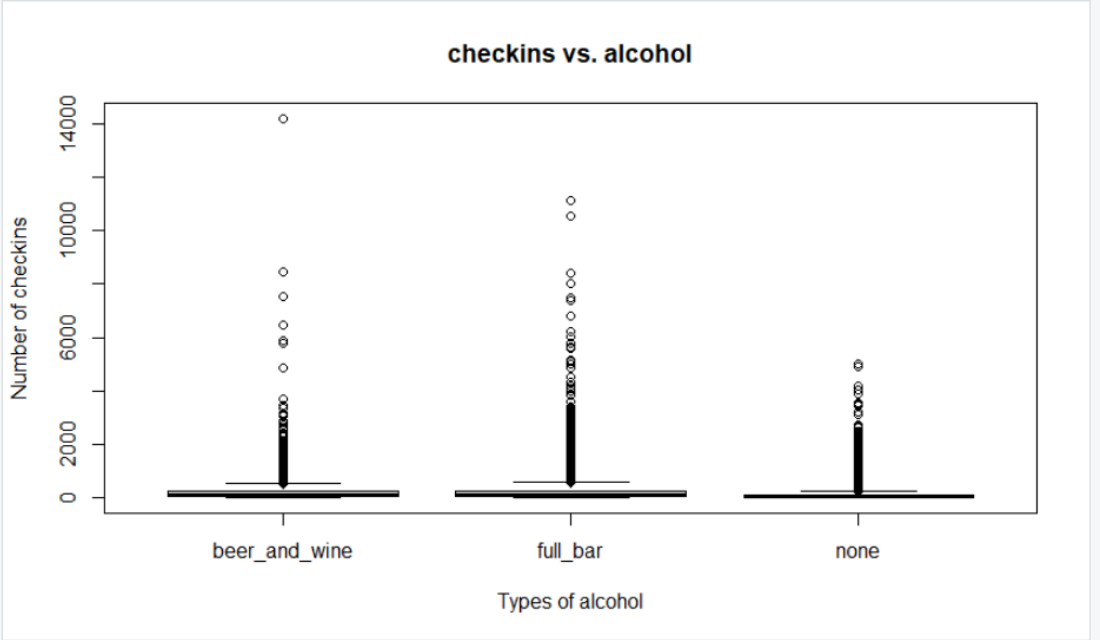
According to the numbers above, ‘reviewCount’ and ‘checkins’ exhibit the largest positive correlation (0.8274936), which matches my observation in (A) and supports the claim that these 2 have the strongest correlation. Attributes ‘stars’ and ‘checkins’ have the smallest positive correlation

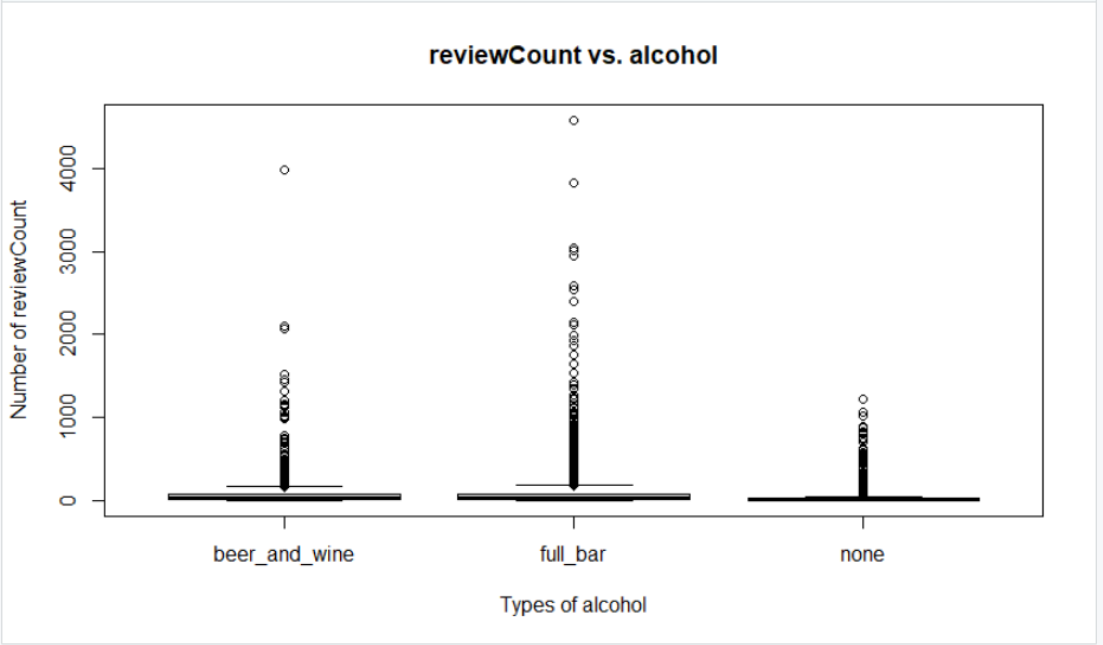
(0.09440071), which can be seen in the scatterplot matrix above because the curve for their plot is nearest to the center of the plot.

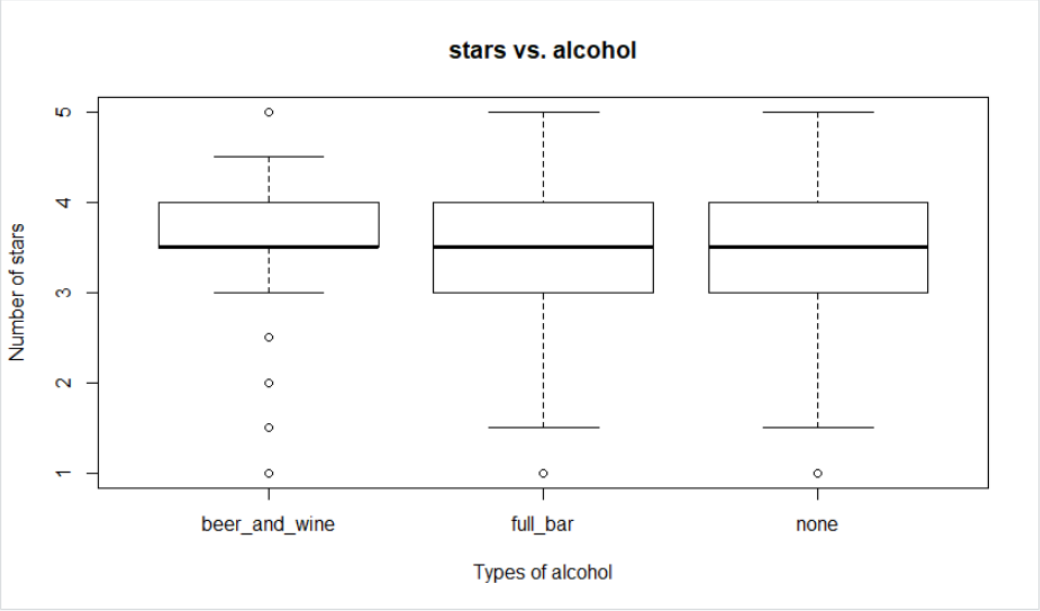
C. 5 boxplots:






checkins vs. alcohol
Number of checkins
0
2000
6000
10000
14000
beer_and_wine
full_bar
none
Types of alcohol


reviewCount vs. alcohol
Number of reviewCount
0
1000
2000
3000
4000
beer_and_wine
full_bar
none
Types of alcohol



a. The boxplot between reviewCount vs. alcohol exhibits the most association because the number of reviews increases as we go from "none" to "beer_and_wine" to "full_bar". This is expected because the more choices of alcohol there are, the more things people have to review on a restaurant.

b. Interquartile ranges:

```
> # C(b). IQRs
> none_sub <- subset(yelp, yelp$alcohol == "none")
> baw_sub <- subset(yelp, yelp$alcohol == "beer_and_wine")
> full_sub <- subset(yelp, yelp$alcohol == "full_bar")
> empty_sub <- subset(yelp, is.na(yelp$alcohol))
> quantile(none_sub$reviewCount)
  0%  25%  50%  75% 100%
   3    6   11   26 1223
> quantile(baw_sub$reviewCount)
  0%  25%  50%  75% 100%
   3   16   38   81 3984
> quantile(full_sub$reviewCount)
  0%  25%  50%  75% 100%
   3   15   36   87 4578
> quantile(empty_sub$reviewCount)
   0%   25%   50%   75%  100%
 34.0  89.0 144.0 263.5 383.0
>
```

'alcohol' categories:

➔ "none"

➔ "beer_and_wine"

➔ "full_bar"

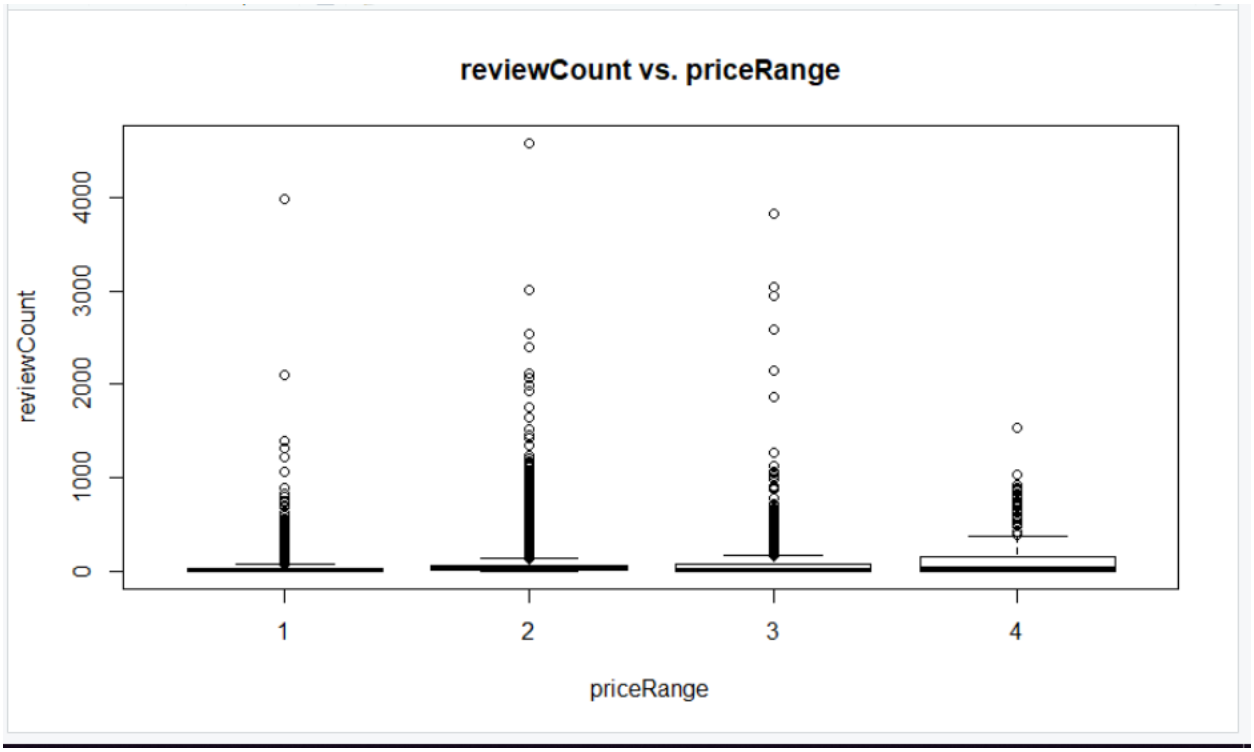
➔ NA

Overlap between NA, "none", "beer_and_wine", and "full_bar" is between 3 and 383. Between "none", "beer_and_wine", and "full_bar", there is an increase in the number restaurants with high number of reviews, which **supports my observation in part (a) above**. Between "none" and "beer_and_wine", the increase already starts around the first quantile until

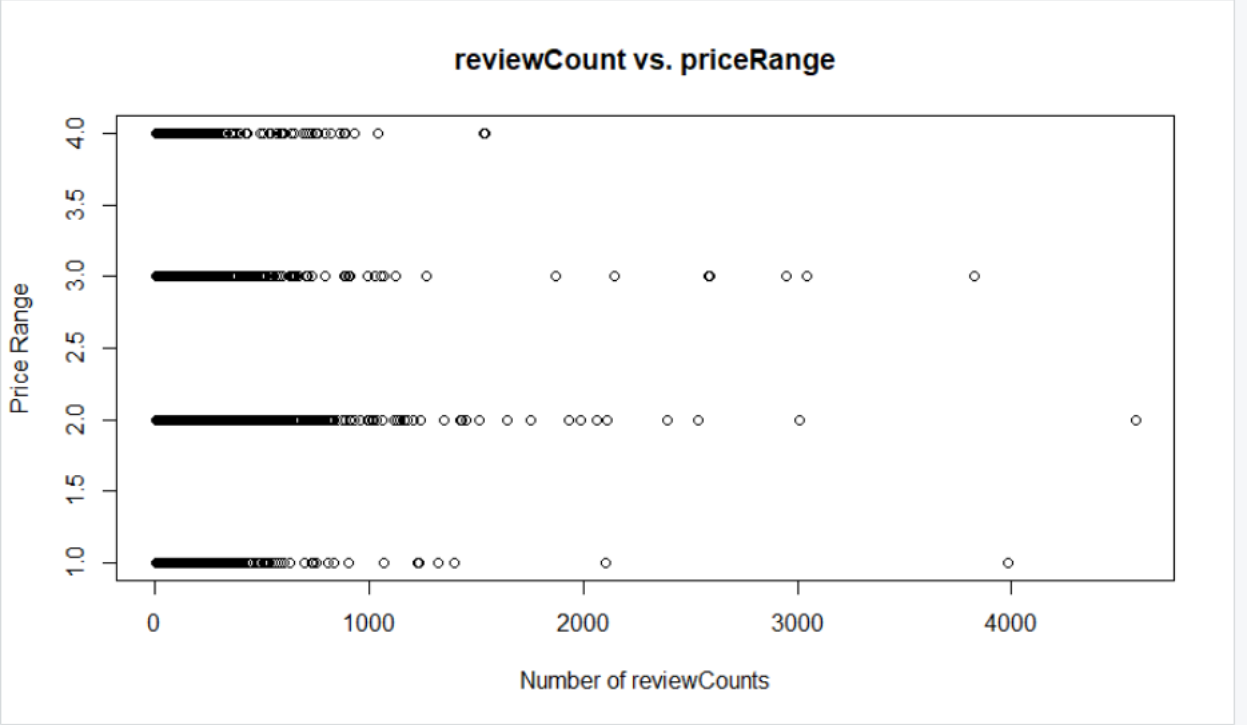
the maximum values, provided that all values except “min” of “beer_and_wine” quantiles are larger than that of “none”. The same increase happens between “beer_and_wine” and “full_bar”, suggesting a noticeable increase of reviews for restaurants as that restaurant’s alcohol choices increases.

- **(7) Identifying 2 potential hypotheses:**
  A. reviewCount vs. priceRange:

    a. Boxplot of reviewCount vs. priceRange:



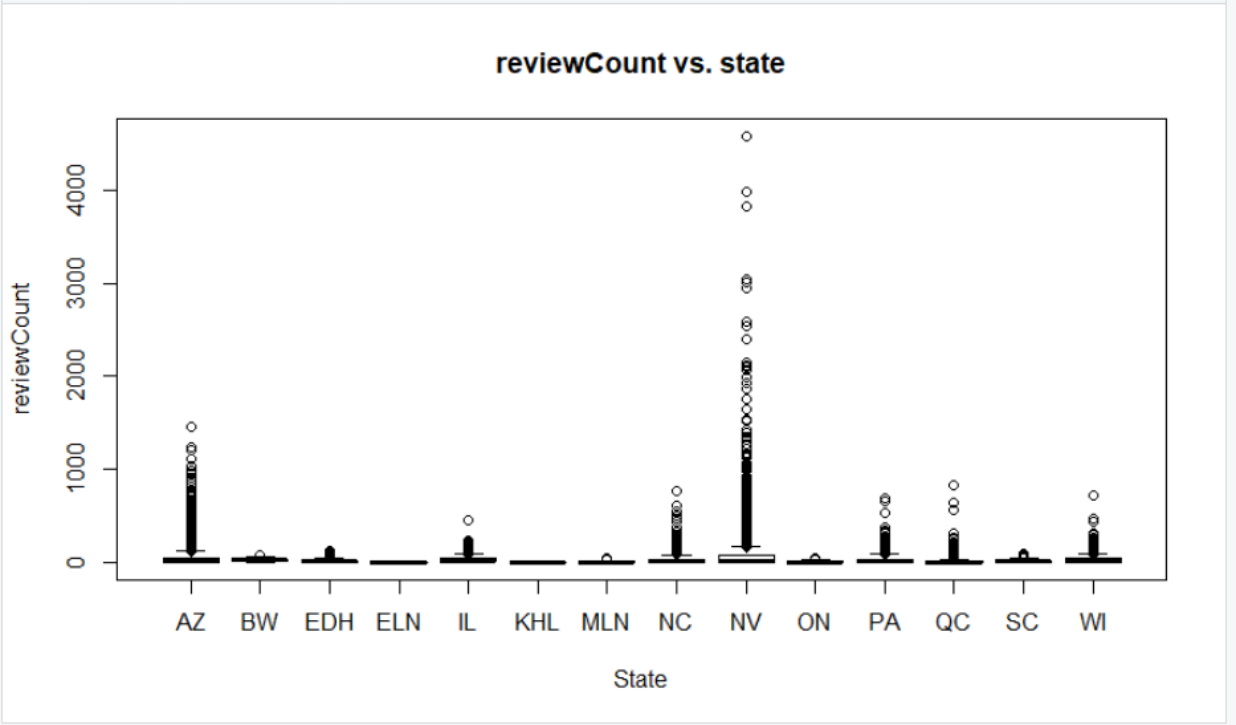
Scatterplot of reviewCount vs. priceRange:



b. Variable ‘priceRange’ is discrete because there is no value between two whole numbers. Variable ‘reviewCount’ is discrete because there is no half of a review. Since both variables are numerical with a lot of data, some kind of non-continuous plots such as scatterplot or boxplot would be sufficient to represent this data.
c. Hypothesis: Given the 2 plots above, there is no correlation between two variables reviewCount and priceRange.
d. English: There is little to no correlation between the number of reviews of a restaurant and that restaurant’s price range.
e. Type of hypothesis: Non-directional and relational.

B. reviewCount vs. state:

a. Boxplot of reviewCount vs. state:



b. As explained above, ‘reviewCount’ is a discrete variable. Variable ‘state’ is categorical, so it is also discrete. Since ‘state’ is categorical, using a boxplot to describe the data is a good choice.
c. Hypothesis: Given this box plot, since category ‘AZ’ in ‘state’ has a lot of data, the variance of reviewCount in AZ is also higher than that of other categories in ‘state’.
d. English: Since Arizona has more restaurants than mostly other states, the range of number of restaurant reviews also spread out more.
e. Type of hypothesis: Causal and directional.

APPENDIX

```
library(readr)
yelp <- read_csv("yelp.csv", quote = "\"", comment.char = "")
View(yelp)

# Part 3
summary(yelp)
names(yelp)

# Part 4a
# a/ Histogram of stars
hist(yelp$stars,
       main = "Histogram of stars feature",
       xlab = "Number of stars",
       col = "darkmagenta",
       freq = FALSE
       )

# b/ Logged stars
yelp$log_stars <- log(yelp$stars)
hist(yelp$log_stars,
    main = "Histogram of Log Values of stars",
    xlab = "Log Values",
    col = "darkmagenta",
```

```
    freq = FALSE
)

# c/ Density plot
plot(density(yelp$log_stars), main="Density of Log Values for stars")
polygon(density(yelp$log_stars), col="red", border="blue")

# Part 4b
states <- table(yelp$state)
names(states)
barplot(states, main="State Distribution",
        xlab="State")

# Part 5A
reliableReview <- ifelse(yelp$reviewCount > 10, TRUE, FALSE)
highStar <- ifelse(yelp$reviewCount > 10 & yelp$stars > 4, TRUE, FALSE)
yelp_new <- yelp
yelp_new <- cbind(yelp_new, reliableReview, highStar)
View(yelp_new)

names(yelp)
names(yelp_new)

# Part 5b
```

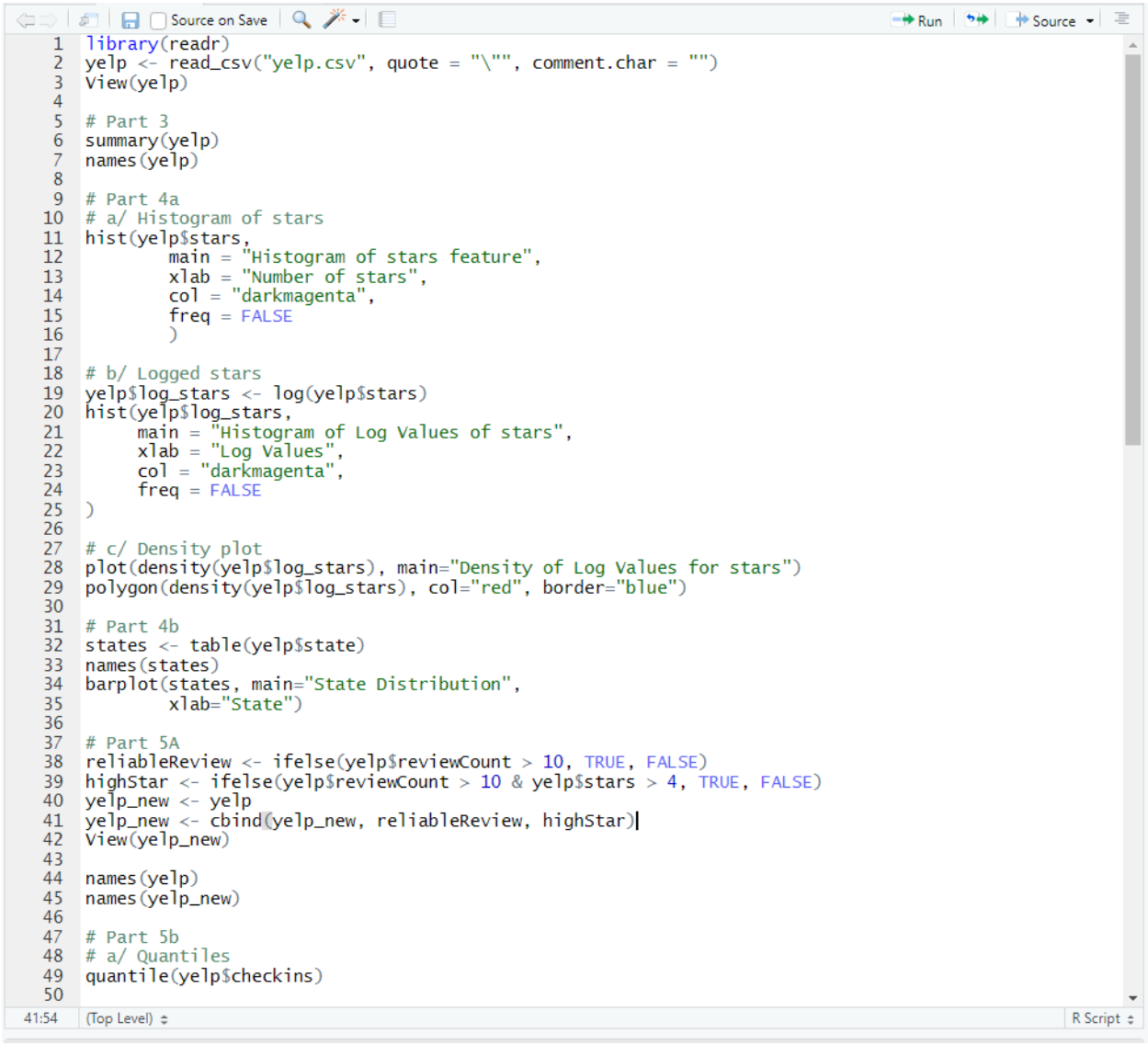
```
# a/ Quantiles
quantile(yelp$checkins)

# b/ Subset
checkins_subset <- subset(yelp, yelp$checkins <= 16)

# c/ Summary
summary(checkins_subset$checkins)
summary(checkins_subset$stars)
summary(checkins_subset$noiseLevel)
summary(checkins_subset$priceRange)
summary(checkins_subset$reviewCount)
summary(checkins_subset$goodForGroups)

summary(yelp$checkins)
summary(yelp$stars)
summary(yelp$noiseLevel)
summary(yelp$priceRange)
summary(yelp$reviewCount)
summary(yelp$goodForGroups)

# Part 6
# A. Scatterplot matrix
pairs(~ yelp$stars + yelp$reviewCount + yelp$checkins,data = yelp,
```

```
    main = "Scatterplot Matrix")

# B. Correlation
cor(yelp$stars, yelp$stars)
cor(yelp$stars, yelp$reviewCount)
cor(yelp$stars, yelp$checkins)

cor(yelp$reviewCount, yelp$stars)
cor(yelp$reviewCount, yelp$reviewCount)
cor(yelp$reviewCount, yelp$checkins)

cor(yelp$checkins, yelp$stars)
cor(yelp$checkins, yelp$reviewCount)
cor(yelp$checkins, yelp$checkins)

# C. Boxplots
boxplot(yelp$stars ~ yelp$alcohol,data = yelp, main="stars vs. alcohol",
        xlab="Types of alcohol", ylab="Number of stars")
boxplot(yelp$reviewCount ~ yelp$alcohol,data = yelp, main="reviewCount vs.
alcohol",
        xlab="Types of alcohol", ylab="Number of reviewCount")
boxplot(yelp$checkins ~ yelp$alcohol,data = yelp, main="checkins vs. alcohol",
        xlab="Types of alcohol", ylab="Number of checkins")
boxplot(yelp$longitude ~ yelp$alcohol,data = yelp, main="longitude vs. alcohol",
        xlab="Types of alcohol", ylab="longitude")
```

```
boxplot(yelp$latitude ~ yelp$alcohol,data = yelp, main="latitude vs. alcohol",
        xlab="Types of alcohol", ylab="latitude")


# C(b). IQRs
none_sub <- subset(yelp, yelp$alcohol == "none")
baw_sub <- subset(yelp, yelp$alcohol == "beer_and_wine")
full_sub <- subset(yelp, yelp$alcohol == "full_bar")
empty_sub <- subset(yelp, is.na(yelp$alcohol))

quantile(none_sub$reviewCount)
quantile(baw_sub$reviewCount)
quantile(full_sub$reviewCount)
quantile(empty_sub$reviewCount)

# Part 7
# A. reviewCount vs. priceRange
plot(yelp$reviewCount, yelp$priceRange, main="reviewCount vs. priceRange",
     xlab="Number of reviewCounts", ylab="Price Range")
boxplot(yelp$reviewCount ~ yelp$priceRange, data = yelp, main="reviewCount
vs. priceRange",
        xlab="priceRange", ylab="reviewCount")

# B.
boxplot(yelp$reviewCount ~ yelp$state, data = yelp, main="reviewCount vs.
state",
```

```
xlab="State", ylab="reviewCount")
```

Pictures of code:



```r
library(readr)
yelp <- read_csv("yelp.csv", quote = "\"", comment.char = "")
View(yelp)

# Part 3
summary(yelp)
names(yelp)

# Part 4a
# a/ Histogram of stars
hist(yelp$stars,
        main = "Histogram of stars feature",
        xlab = "Number of stars",
        col = "darkmagenta",
        freq = FALSE
        )

# b/ Logged stars
yelp$log_stars <- log(yelp$stars)
hist(yelp$log_stars,
     main = "Histogram of Log Values of stars",
     xlab = "Log Values",
     col = "darkmagenta",
     freq = FALSE
)

# c/ Density plot
plot(density(yelp$log_stars), main="Density of Log Values for stars")
polygon(density(yelp$log_stars), col="red", border="blue")

# Part 4b
states <- table(yelp$state)
names(states)
barplot(states, main="State Distribution",
        xlab="State")

# Part 5A
reliableReview <- ifelse(yelp$reviewCount > 10, TRUE, FALSE)
highStar <- ifelse(yelp$reviewCount > 10 & yelp$stars > 4, TRUE, FALSE)
yelp_new <- yelp
yelp_new <- cbind(yelp_new, reliableReview, highStar)
View(yelp_new)

names(yelp)
names(yelp_new)

# Part 5b
# a/ Quantiles
quantile(yelp$checkins)

```

```r
# b/ Subset
checkins_subset <- subset(yelp, yelp$checkins <= 16)

# c/ Summary
summary(checkins_subset$checkins)
summary(checkins_subset$stars)
summary(checkins_subset$noiseLevel)
summary(checkins_subset$priceRange)
summary(checkins_subset$reviewCount)
summary(checkins_subset$goodForGroups)

summary(yelp$checkins)
summary(yelp$stars)
summary(yelp$noiseLevel)
summary(yelp$priceRange)
summary(yelp$reviewCount)
summary(yelp$goodForGroups)

# Part 6
# A. Scatterplot matrix
pairs(~ yelp$stars + yelp$reviewCount + yelp$checkins,data = yelp,
      main = "Scatterplot Matrix")

# B. Correlation
cor(yelp$stars, yelp$stars)
cor(yelp$stars, yelp$reviewCount)
cor(yelp$stars, yelp$checkins)

cor(yelp$reviewCount, yelp$stars)
cor(yelp$reviewCount, yelp$reviewCount)
cor(yelp$reviewCount, yelp$checkins)

cor(yelp$checkins, yelp$stars)
cor(yelp$checkins, yelp$reviewCount)
cor(yelp$checkins, yelp$checkins)

# C. Boxplots
boxplot(yelp$stars ~ yelp$alcohol,data = yelp, main="stars vs. alcohol",
        xlab="Types of alcohol", ylab="Number of stars")
boxplot(yelp$reviewCount ~ yelp$alcohol,data = yelp, main="reviewCount vs. alcohol",
        xlab="Types of alcohol", ylab="Number of reviewCount")
boxplot(yelp$checkins ~ yelp$alcohol,data = yelp, main="checkins vs. alcohol",
        xlab="Types of alcohol", ylab="Number of checkins")
boxplot(yelp$longitude ~ yelp$alcohol,data = yelp, main="longitude vs. alcohol",
        xlab="Types of alcohol", ylab="longitude")
boxplot(yelp$latitude ~ yelp$alcohol,data = yelp, main="latitude vs. alcohol",
        xlab="Types of alcohol", ylab="latitude")

# C(b). IQRs
```

```r

cor(yelp$checkins, yelp$stars)
cor(yelp$checkins, yelp$reviewCount)
cor(yelp$checkins, yelp$checkins)

# C. Boxplots
boxplot(yelp$stars ~ yelp$alcohol,data = yelp, main="stars vs. alcohol",
        xlab="Types of alcohol", ylab="Number of stars")
boxplot(yelp$reviewCount ~ yelp$alcohol,data = yelp, main="reviewCount vs. alcohol",
        xlab="Types of alcohol", ylab="Number of reviewCount")
boxplot(yelp$checkins ~ yelp$alcohol,data = yelp, main="checkins vs. alcohol",
        xlab="Types of alcohol", ylab="Number of checkins")
boxplot(yelp$longitude ~ yelp$alcohol,data = yelp, main="longitude vs. alcohol",
        xlab="Types of alcohol", ylab="longitude")
boxplot(yelp$latitude ~ yelp$alcohol,data = yelp, main="latitude vs. alcohol",
        xlab="Types of alcohol", ylab="latitude")

# C(b). IQRs
none_sub <- subset(yelp, yelp$alcohol == "none")
baw_sub <- subset(yelp, yelp$alcohol == "beer_and_wine")
full_sub <- subset(yelp, yelp$alcohol == "full_bar")
empty_sub <- subset(yelp, is.na(yelp$alcohol))

quantile(none_sub$reviewCount)
quantile(baw_sub$reviewCount)
quantile(full_sub$reviewCount)
quantile(empty_sub$reviewCount)

# Part 7
# A. reviewCount vs. priceRange
plot(yelp$reviewCount, yelp$priceRange, main="reviewCount vs. priceRange",
     xlab="Number of reviewCounts", ylab="Price Range")
boxplot(yelp$reviewCount ~ yelp$priceRange, data = yelp, main="reviewCount vs. priceRange",
        xlab="priceRange", ylab="reviewCount")

# B.
boxplot(yelp$reviewCount ~ yelp$state, data = yelp, main="reviewCount vs. state",
        xlab="State", ylab="reviewCount")

```

41:54 (Top Level) R Script