CS373 Homework 1

Full Name: Ma, Ji

Purdue Email: ma438@purdue.edu (mailto:ma438@purdue.edu)

PUID: 28947432

1 Part I: Basic Probability and Statistics

1. (4 pts) Consider an experiment where a coin is tossed repeatedly until the first time a head is observed.

   This is a geometric distribution

   - a. What is the sample space for this experiment? What is the probability that the coin turns up heads after i tosses?

     The sample space is {H, TH, TTH, TTTH, …}

     $(1-p)^{k-1}p$

     So, the probability will be $(1-\frac{1}{2})^{i-1}\frac{1}{2}=\frac{1}{2}^{i}$

   - b. Let E be the event that the first time a head turns up is after an even number of tosses. What set of outcomes belong to this event? What is the probability that E occurs?

     $(1-p)^{k-1}p$, where k is even

     So, $1/2^2$ + $1/2^4$ + ...

     a = 1/4 , r = 1/4

     $\frac{a}{1-r}=\frac{1/4}{3/4}=\frac{1}{3}$

     The sample space is {TH, TTTH, TTTTTH…}

2. (5 pts) Two standard dice are rolled. Let E be the event that the sum of the dice is odd; let F be the event that at least one of the dice lands on 1; and let G be the event that the sum is 5. Compute the following:

   Total events = 36

   E=(1,2),(1,4),(1,6),(2,1),(2,3),(2,5),(3,2),(3,4),(3,6),(4,1),(4,3),(4,5),(5,2),(5,4),(5,6),(6,1),(6,3),(6,5)

   F=(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(3,1),(4,1),(5,1),(6,1)

   G=(1,4),(2,3),(3,2),(4,1)

   - a. P(E∩F)

     {(1, 2),(1, 4),(1, 6),(2, 1),(4, 1),(6, 1)}

     $\frac{6}{6^2}=\frac{1}{6}$

   - b. P(E∪F)

     $\frac{23}{36}$

   - c. P(F ∪ G)

     $\frac{13}{36}$

   - d. P(E ∪ ¬F)

     $\frac{31}{36}$

- e. P(E∪F∪G)

  $\frac{23}{36}$

3. (6 pts) A system is built using 3 disks d1 , d2 , d3 having probabilities of failure 0.01, 0.03 and 0.05 respectively. Suppose the disks fail independently.
   - a. Let E denote the event of loss of data, which occurs only if two or more disks fail. Compute P (E), the probability of loss of data.

   P(¬E) = 1 disk fail or no disk fail = 0.99 * 0.97 * 0.95 + 0.99 * 0.97 * 0.05 + 0.99 * 0.03 * 0.95 + 0.01 * 0.97 * 0.95 = 0.99773

   P(E) = 1 - 0.99773 = 0.00227

   - b. Instead, let F denote the event that at least one of the following happens: (i) d1 fails; (ii) d2 and d3 both fail. If loss of data only occurs when event F occurs, then what is the probability that there is loss of data?

   P(F) = P(i ∪ ii) = 0.01 * 0.97 * 0.95 + 0.99 * 0.03 * 0.05 - 0.01 * 0.03 * 0.05 = 0.010685

   - c. Considering the setting of 3b, given that d3 has failed, what is the conditional probability that event F will occur and there will be loss of data?

   G: d3 has failed

   P(F | G) = P(F∩G) / P(G) = 0.99 * 0.03 * 0.05 / (0.99 * 0.97 * 0.05) = 0.0309

4. (6 pts) 52% of the students at a particular college are female. 5% of the students in the college are majoring in computer science. 0.55% of the students are women majoring in computer science.

   F: female P(F) = 0.52

   C: CS studdent P(C) = 0.05

   P(C ∩ F) = 0.0055

   - a. If a student is selected at random, find the conditional probability that the student is female given that they are majoring in computer science. (State this as a conditional probability and show the calculation.)

   P(F | C) = P(F ∩ C) / P (C) = 0.0055 / 0.05 = 0.11

   - b. If a student is selected at random, find the conditional probability that the student is majoring in computer science given that they are female. (State this as a conditional probability and show the calculation.)

   P(C | F) = P(F ∩ C) / P (F) = 0.0055 / 0.52 = 0.0106

   - c. Now suppose that the overall proportion of female students increases to 57% and that the conditional probability from 4a changes (i.e., increases or de- creases) to 15%. Compute the updated conditional probability that a student is majoring in computer science given that they are female. (Assume that the overall proportion of students majoring in CS stays the same.)

   F: female P(F) = 0.57

   P(F | C) = 0.15

   P(C | F) = P(F ∩ C) / P(F) = P(F | C) * P(C) / P(F) = 0.15 * 0.05 / 0.57 = 0.0132

5. (6 pts) Let Xn be the random variable that equals the number of heads minus the number of tails when n coins are flipped. Each flip has a probability of p of heads, 1 - p probability of tails. Do not assume p = 1/2.

- a. What is the expected value of Xn ?

  E of heads = np, let’s assume i heads, and n-i tails

  P(i - (n - i)) = P(2i - n); E(2i - n) = 2E(i) - E(n) = 2np - n

- b. What is the variance of Xn?

  Var = E(X_n^2) - $E(X_n)^2$ = 4np(1-p)

- c. Compute the expected value and variance of X3. Plug it in, n = 3

  E(X3) = 6p - 3

  Var(X3) = 12p(1-p)

2 Part II: R

3 Data import and summarization

```
yelp = read.csv("yelp.csv",  header = TRUE, quote="\"", comment.char="")
```

a. (2 pts) Print the names of the columns in the table using names().

```
names(yelp)
```

```
##  [1] "business_id"          "name"                 "fullAddress"         
##  [4] "city"                 "state"                "latitude"            
##  [7] "longitude"            "stars"                "reviewCount"         
## [10] "checkins"             "open"                 "neighborhoods"       
## [13] "categories"           "alcohol"              "noiseLevel"          
## [16] "attire"               "priceRange"           "delivery"            
## [19] "ambience"             "parking"              "dietaryRestrictions" 
## [22] "waiterService"        "smoking"              "outdoorSeating"      
## [25] "caters"               "recommendedFor"       "goodForGroups"       
## [28] "goodForKids"
```

b. (2 pts) Print a summary of the data using the summary() function.

```
summary(yelp)
```

```
##                   business_id            name      
##  __etvGuL2dh_a1LOT0gNYQ:    1   Starbucks :  407  
##  __kNfrrGoUXoF-BYciMU_Q:    1   McDonald's:  275  
##  __Y2jjdCFHvq3rzSbpDBlw:    1   Subway    :  256  
##  _-1EgXrkOlKajCsmasuEgg:    1   Walgreens :  158  
##  _-6I6VXjr-NiwIBa_1uI4A:    1   Taco Bell :  148  
##  _-9pMxBWtG_x8l4rHWBasg:    1   Wendy's   :  113  
##  (Other)               :24807   (Other)   :23456  
##                                                                             full
Address
##  Bellagio Las Vegas\n3600 S Las Vegas Blvd\nThe Strip\nLas Vegas, NV 89109
:   21
```

```
##  Las Vegas, NV
:   17
##  5000 S Arizona Mills Cir\nTempe, AZ 85282
:   14
##  3131 Las Vegas Blvd. South\nThe Strip\nLas Vegas, NV 89109
:   13
##  Monte Carlo Hotel and Casino\n3770 Las Vegas Blvd S\nThe Strip\nLas Vegas, NV
89109:   13
##  2000 E Rio Salado Pkwy\nTempe, AZ 85281
:   12
##  (Other)
:24723
##          city           state         latitude       longitude
##  Las Vegas : 5256   AZ     :9301   Min.   :32.88   Min.   :-115.370
##  Phoenix   : 3072   NV     :6296   1st Qu.:33.54   1st Qu.:-114.977
##  Charlotte : 1993   QC     :2389   Median :36.03   Median :-111.924
##  Pittsburgh: 1467   NC     :2370   Mean   :37.53   Mean   : -97.298
##  Scottsdale: 1296   PA     :1613   3rd Qu.:40.41   3rd Qu.: -80.807
##  Montral   : 1267   WI     :1089   Max.   :55.99   Max.   :  8.549
##  (Other)   :10462   (Other):1755
##      stars        reviewCount         checkins        open
##  Min.   :1.000   Min.   :   3.00   Min.   :    3   Mode :logical
##  1st Qu.:3.000   1st Qu.:   8.00   1st Qu.:   16   FALSE:3580
##  Median :3.500   Median :  18.00   Median :   48   TRUE :21233
##  Mean   :3.544   Mean   :  49.03   Mean   :  166   NA's :0
##  3rd Qu.:4.000   3rd Qu.:  48.00   3rd Qu.:  155
##  Max.   :5.000   Max.   :4578.00   Max.   :14203
##
##        neighborhoods                                     categories
##  []           :15727   ['Mexican', 'Restaurants']             : 1331
##  ['The Strip']:  816   ['Food', 'Coffee & Tea']               :  844
##  ['Southeast']:  639   ['Pizza', 'Restaurants']               :  831
##  ['Downtown'] :  533   ['Chinese', 'Restaurants']             :  776
##  ['Westside'] :  526   ['Burgers', 'Fast Food', 'Restaurants']:  549
##  ['Eastside'] :  447   ['Restaurants', 'Italian']             :  509
##  (Other)      : 6125   (Other)                                :19973
##           alcohol          noiseLevel       attire        priceRange
##               :    3             : 7947          : 7005   Min.   :1.000
##  beer_and_wine: 2497   average  :10957   casual:17129   1st Qu.:1.000
##  full_bar     : 7565   loud     : 1622   dressy:  640   Median :2.000
##  none         :14748   quiet    : 3562   formal:   39   Mean   :1.631
##                        very_loud:  725                  3rd Qu.:2.000
##                                                         Max.   :4.000
##                                                         NA's   :903
##   delivery            ambience                 parking
##  Mode :logical   ['casual']:7878   ['lot']          :10348
##  FALSE:14471               :7875   []               : 6675
##  TRUE :3093      []        :6348   ['street']       : 3046
##  NA's :7249      ['divey'] : 716                    : 2456
##                  ['trendy']: 567   ['garage']       :  907
##                  ['classy']: 320   ['street', 'lot']:  364
##                  (Other)   :1109   (Other)          : 1017
##                    dietaryRestrictions waiterService      smoking
##                              :24696    Mode :logical          :21862
##  ['vegan']                   :   45    FALSE:6208      no     :  904
##  ['vegetarian']              :   23    TRUE :10351     outdoor: 1415
##  []                          :   20    NA's :8254      yes    :  632
##  ['dairy-free', 'vegetarian']:    7
##  ['vegan', 'vegetarian']     :    5
##  (Other)                     :   17
##  outdoorSeating    caters                    recommendedFor
##  Mode :logical   Mode :logical                      :7859
##  FALSE:10989     FALSE:6503      []                 :4932
##  TRUE :8698      TRUE :5932      ['lunch']          :4324
##  NA's :5126      NA's :12378     ['dinner']         :2553
##                                  ['lunch', 'dinner']:1966
##                                  ['breakfast']      :1004
##                                  (Other)            :2175
##  goodForGroups   goodForKids
##  Mode :logical   Mode :logical
##  FALSE:2054      FALSE:506
##  TRUE :17078     TRUE :1283
##  NA's :5681      NA's :23024
##
##
##
```

c. (2 pts) Print a summary of the noiseLevel attribute and the stars attribute.

```
summary(yelp$noiseLevel)
```

```
##             average      loud     quiet very_loud
##      7947     10957      1622      3562       725
```

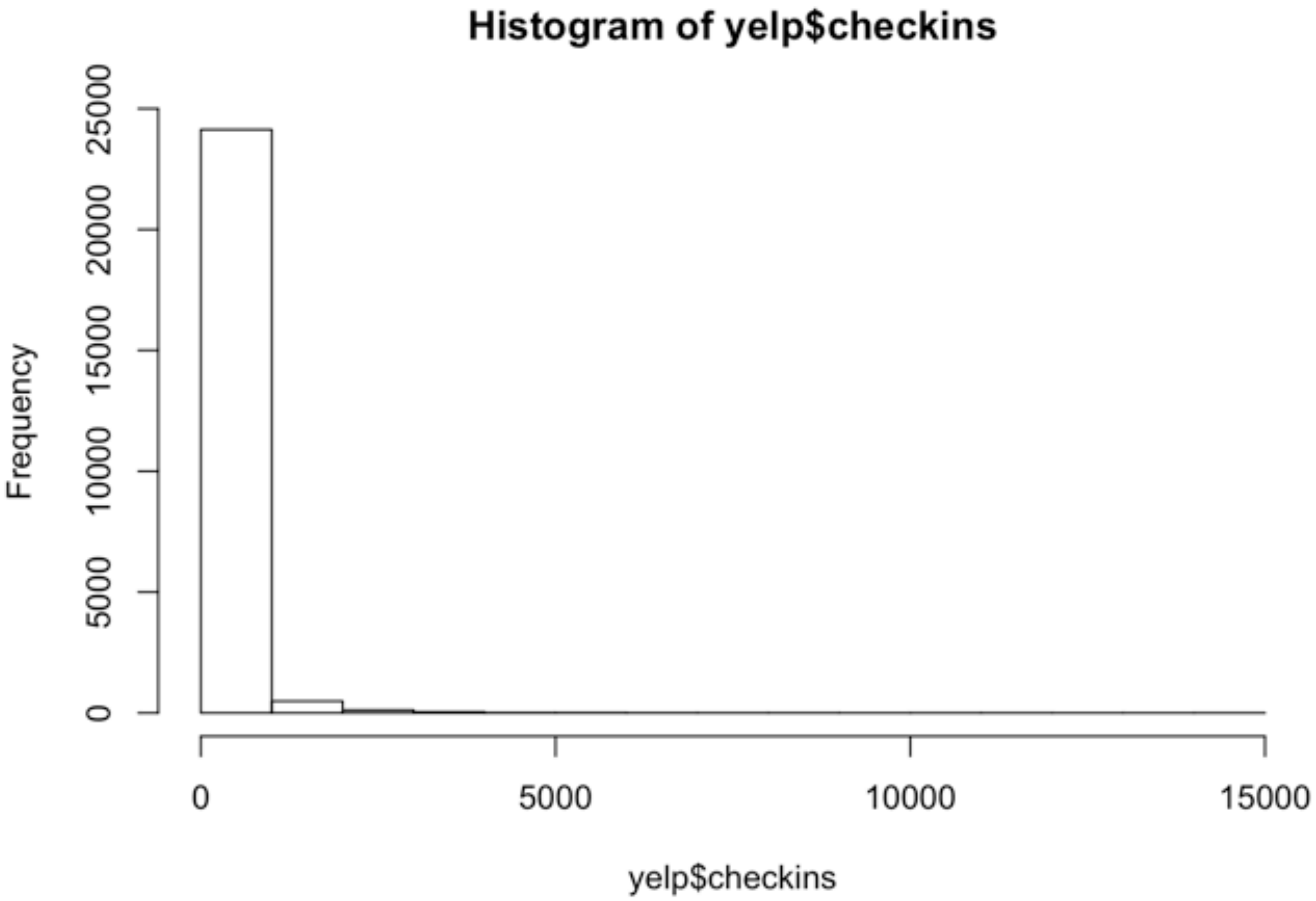
```
summary(yelp$stars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.500   3.544   4.000   5.000
```

4 1D plots

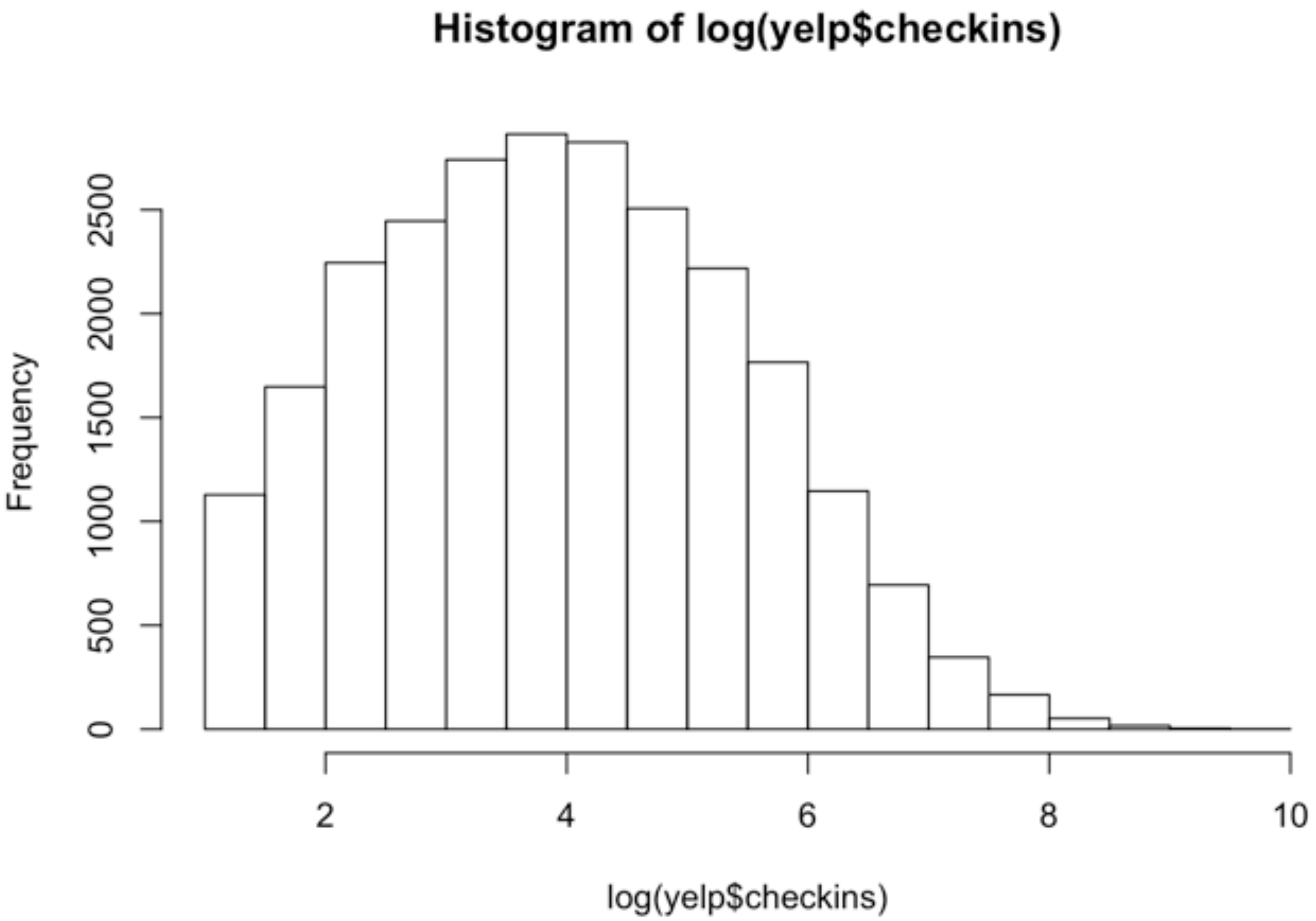
a. (4 pts) Plot a histogram of the checkins attribute. Use the hist() function with its default values and make sure to title the plot with the name of the attribute for clarity.

```
hist(yelp$checkins)
```



(b) (4 pts) Compute the logged values for checkins (you can use log() to compute the log of all the values in a vector). Plot a histogram of the logged values.

```
hist(log(yelp$checkins))
```



c. (4 pts) Discuss the differences between the two plots and the information they convey about the distribution of checkins values in the data.

The different between two plots is abvious. The first plot is skewed as hell, and the log function will help us remove the skewness and it will reveil the more helpful information.Because of the density for the low frequency is too high, so the normal histogram is sort of meaningless.

5 Sampling and transforming data

a. (4 pts) The attributes categories and recommendedFor each contain a comma separated list of values associated with each restaurant. Compute two new boolean features: isAmerican and goodForDinner with a value of TRUE if the list contains “American” (in categories), “dinner” (in recommendedFor) respectively and FALSE otherwise. You can use the function grepl(str, f$column name) to check whether the values in column name contain the string str. Append the two new columns to the original data frame, using cbind(), to increase the number of features to 32. Show the output of summary() for those two columns.

```
isAmerican = grepl("American", yelp$categories)
goodForDinner = grepl("dinner", yelp$recommendedFor)
yelp = cbind(yelp, isAmerican, goodForDinner)
summary(yelp$isAmerican)
```

```
##    Mode   FALSE    TRUE    NA's
## logical   21456    3357       0
```

```
summary(yelp$goodForDinner)
```

```
##    Mode   FALSE    TRUE    NA's 
## logical   19670    5143       0
```

b. (4 pts) Print the quantiles (using quantile()) for the reviewCount attribute.

```
quantile(yelp$reviewCount)
```

```
##   0%  25%  50%  75% 100% 
##    3    8   18   48 4578
```

c. (6 pts) Select a subset of the data with reviewCount value ≤ 1st quartile (25th percentile). You can use subset() or select from the data frame with [] operations. Print a summary of the above subset for the following attributes: reviewCount, stars, attire, priceRange, delivery, goodForKids, and compare them to their summary for the full dataset. Discuss any differences in the distributions of the numerical attributes that you find.

```
lowReviewCount = subset(yelp, yelp$reviewCount <= quantile(yelp$reviewCount, 0.25)
)
#instead of give summary for individual, I print out the general summary and look
at the reviewCount, stars, attire, priceRange, delivery, goodForKids individually.
summary(lowReviewCount)
```

```
##                   business_id            name     
##  __Y2jjdCFHvq3rzSbpDBlw:   1   Subway     : 213  
##  _-1EgXrkOlKajCsmasuEgg:   1   McDonald's : 174  
##  _-6I6VXjr-NiwIBa_1uI4A:   1   Starbucks  : 137  
##  _-EB8tQzBlM_jLkgtRw4Rg:   1   Walgreens  : 115  
##  _04PNAespgMZVXBJrkmbNA:   1   Taco Bell  :  99  
##  _0DI4UXAaFC6hOYpBadtIw:   1   Burger King:  80  
##  (Other)               :6954   (Other)    :6142  
##                                          fullAddress            city     
##  5000 S Arizona Mills Cir\nTempe, AZ 85282     :   7   Las Vegas :1189  
##  1300 W Sunset Rd\nHenderson, NV 89014         :   6   Phoenix   : 819  
##  Las Vegas, NV                                 :   6   Charlotte : 553  
##  11025 Carolina Place Pkwy\nPineville, NC 28134:   5   Montral   : 514  
##  138, avenue Atwater\nMontreal, QC H4C 2G3     :   5   Pittsburgh: 386  
##  4300 Meadows Ln\nWestside\nLas Vegas, NV 89107:   5   Montreal  : 347  
##  (Other)                                       :6926   (Other)   :3152  
##      state         latitude       longitude            stars      
##  AZ     :2400   Min.   :32.88   Min.   :-115.352   Min.   :1.000  
##  NV     :1479   1st Qu.:33.58   1st Qu.:-112.264   1st Qu.:3.000  
##  QC     :1129   Median :36.08   Median :-111.823   Median :3.500  
##  NC     : 701   Mean   :38.30   Mean   : -94.056   Mean   :3.418  
##  PA     : 440   3rd Qu.:43.07   3rd Qu.: -79.998   3rd Qu.:4.000  
##  EDH    : 299   Max.   :55.99   Max.   :   8.485   Max.   :5.000  
##  (Other): 512                                                     
##   reviewCount       checkins          open                neighborhoods 
##  Min.   :3.000   Min.   :  3.00   Mode :logical   []             :4839
##  1st Qu.:4.000   1st Qu.:  7.00   FALSE:887       ['Southeast']: 144
##  Median :5.000   Median : 13.00   TRUE :6073      ['Downtown'] : 140
##  Mean   :5.247   Mean   : 24.78   NA's :0         ['The Strip']: 136
##  3rd Qu.:7.000   3rd Qu.: 29.00                   ['Eastside'] : 133
##  Max.   :8.000   Max.   :694.00                   ['Westside'] : 119
##                                                   (Other)      :1449
##                                   categories           alcohol
##  ['Burgers', 'Fast Food', 'Restaurants']: 310                :   0
##  ['Food', 'Grocery']                    : 293   beer_and_wine: 266
##  ['Food', 'Coffee & Tea']               : 285   full_bar     : 968
##  ['Fast Food', 'Restaurants']           : 278   none         :5726
##  ['Mexican', 'Restaurants']             : 274
##  ['Pizza', 'Restaurants']               : 262
##  (Other)                                :5258
##      noiseLevel      attire       priceRange     delivery
##           :4096         :3248   Min.   :1.000   Mode :logical
##  average  :1549   casual:3581   1st Qu.:1.000   FALSE:2899
##  loud     : 324   dressy: 107   Median :1.000   TRUE :693
##  quiet    : 836   formal:  24   Mean   :1.546   NA's :3368
##  very_loud: 155                 3rd Qu.:2.000
##                                 Max.   :4.000
##                                 NA's   :825
##         ambience                parking
##             :4104   []                :3518
##  []         :2550                     :1897
##  ['casual'] : 226   ['lot']           :1059
##  ['divey']  :  36   ['street']        : 333
##  ['hipster']:   9   ['garage']        :  59
##  ['trendy'] :   7   ['street', 'lot']:  39
##  (Other)    :  28   (Other)           :  55
##                                                                      dietaryRe
strictions
##
:6955
##  ['vegan']
:   3
##  ['vegetarian']
:   1
##  []
:   1
##  ['dairy-free', 'gluten-free', 'vegan',  halal, 'soy-free', 'vegetarian']
:   0
##  ['dairy-free', 'gluten-free', 'vegan', kosher, halal, 'soy-free', 'vegetarian'
]:   0
##  (Other)
:   0
##  waiterService      smoking     outdoorSeating    caters
##  Mode :logical          :6578   Mode :logical     Mode :logical
##  FALSE:1323      no      : 111   FALSE:2672       FALSE:1040
##  TRUE :1729      outdoor: 157   TRUE :1370       TRUE :620
##  NA's :3908      yes    : 114   NA's :2918       NA's :5300
##
##
##
```

```
##              recommendedFor goodForGroups   goodForKids
##                     :3832   Mode :logical   Mode :logical
##  []                 :2532   FALSE:704       FALSE:15
##  ['lunch']          : 249   TRUE :3471      TRUE :31
##  ['breakfast']      :  93   NA's :2785      NA's :6914
##  ['lunch', 'dinner']:  76
##  ['dinner']         :  54
##  (Other)            : 124
##  isAmerican      goodForDinner
##  Mode :logical   Mode :logical
##  FALSE:6452      FALSE:6796
##  TRUE :508       TRUE :164
##  NA's :0         NA's :0
##
##
##
```

```
summary(yelp)
```

```
##                   business_id            name
##  __etvGuL2dh_a1LOT0gNYQ:    1   Starbucks :  407
##  __kNfrrGoUXoF-BYciMU_Q:    1   McDonald's:  275
##  __Y2jjdCFHvq3rzSbpDBlw:    1   Subway    :  256
##  _-1EgXrkOlKajCsmasuEgg:    1   Walgreens :  158
##  _-6I6VXjr-NiwIBa_1uI4A:    1   Taco Bell :  148
##  _-9pMxBWtG_x8l4rHWBasg:    1   Wendy's   :  113
##  (Other)               :24807   (Other)   :23456
##                                                                            full
Address
##  Bellagio Las Vegas\n3600 S Las Vegas Blvd\nThe Strip\nLas Vegas, NV 89109
:   21
##  Las Vegas, NV
:   17
##  5000 S Arizona Mills Cir\nTempe, AZ 85282
:   14
##  3131 Las Vegas Blvd. South\nThe Strip\nLas Vegas, NV 89109
:   13
##  Monte Carlo Hotel and Casino\n3770 Las Vegas Blvd S\nThe Strip\nLas Vegas, NV
89109:   13
##  2000 E Rio Salado Pkwy\nTempe, AZ 85281
:   12
##  (Other)
:24723
##          city           state         latitude       longitude
##  Las Vegas : 5256   AZ     :9301   Min.   :32.88   Min.   :-115.370
##  Phoenix   : 3072   NV     :6296   1st Qu.:33.54   1st Qu.:-114.977
##  Charlotte : 1993   QC     :2389   Median :36.03   Median :-111.924
##  Pittsburgh: 1467   NC     :2370   Mean   :37.53   Mean   : -97.298
##  Scottsdale: 1296   PA     :1613   3rd Qu.:40.41   3rd Qu.: -80.807
##  Montral   : 1267   WI     :1089   Max.   :55.99   Max.   :   8.549
##  (Other)   :10462   (Other):1755
##      stars        reviewCount         checkins         open
##  Min.   :1.000   Min.   :   3.00   Min.   :    3   Mode :logical
##  1st Qu.:3.000   1st Qu.:   8.00   1st Qu.:   16   FALSE:3580
##  Median :3.500   Median :  18.00   Median :   48   TRUE :21233
##  Mean   :3.544   Mean   :  49.03   Mean   :  166   NA's :0
##  3rd Qu.:4.000   3rd Qu.:  48.00   3rd Qu.:  155
##  Max.   :5.000   Max.   :4578.00   Max.   :14203
##
##        neighborhoods                                          categories
##  []           :15727   ['Mexican', 'Restaurants']             : 1331
##  ['The Strip']:  816   ['Food', 'Coffee & Tea']               :  844
##  ['Southeast']:  639   ['Pizza', 'Restaurants']               :  831
##  ['Downtown'] :  533   ['Chinese', 'Restaurants']             :  776
##  ['Westside'] :  526   ['Burgers', 'Fast Food', 'Restaurants']:  549
##  ['Eastside'] :  447   ['Restaurants', 'Italian']             :  509
##  (Other)      : 6125   (Other)                                :19973
##           alcohol          noiseLevel       attire        priceRange
##                :    3            : 7947          : 7005   Min.   :1.000
##  beer_and_wine: 2497   average  :10957   casual:17129   1st Qu.:1.000
##  full_bar     : 7565   loud     : 1622   dressy:  640   Median :2.000
##  none         :14748   quiet    : 3562   formal:   39   Mean   :1.631
##                        very_loud:  725                  3rd Qu.:2.000
##                                                         Max.   :4.000
##                                                         NA's   :903
##   delivery             ambience                  parking
##  Mode :logical   ['casual']:7878   ['lot']          :10348
##  FALSE:14471               :7875   []               : 6675
##  TRUE :3093      []        :6348   ['street']       : 3046
##  NA's :7249      ['divey'] : 716                    : 2456
##                  ['trendy']: 567   ['garage']       :  907
##                  ['classy']: 320   ['street', 'lot']:  364
##                  (Other)   :1109   (Other)          : 1017
##                    dietaryRestrictions waiterService      smoking
##                              :24696    Mode :logical          :21862
##  ['vegan']                   :   45    FALSE:6208      no     :  904
##  ['vegetarian']              :   23    TRUE :10351     outdoor: 1415
##  []                          :   20    NA's :8254      yes    :  632
##  ['dairy-free', 'vegetarian']:    7
##  ['vegan', 'vegetarian']     :    5
##  (Other)                     :   17
##  outdoorSeating    caters                  recommendedFor
##  Mode :logical   Mode :logical                      :7859
##  FALSE:10989     FALSE:6503      []                 :4932
##  TRUE :8698      TRUE :5932      ['lunch']          :4324
##  NA's :5126      NA's :12378     ['dinner']         :2553
##                                  ['lunch', 'dinner']:1966
##                                  ['breakfast']      :1004
##                                  (Other)            :2175
##  goodForGroups   goodForKids     isAmerican      goodForDinner
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:2054      FALSE:506       FALSE:21456     FALSE:19670
##  TRUE :17078     TRUE :1283      TRUE :3357      TRUE :5143
##  NA's :5681      NA's :23024     NA's :0         NA's :0
##
##
```

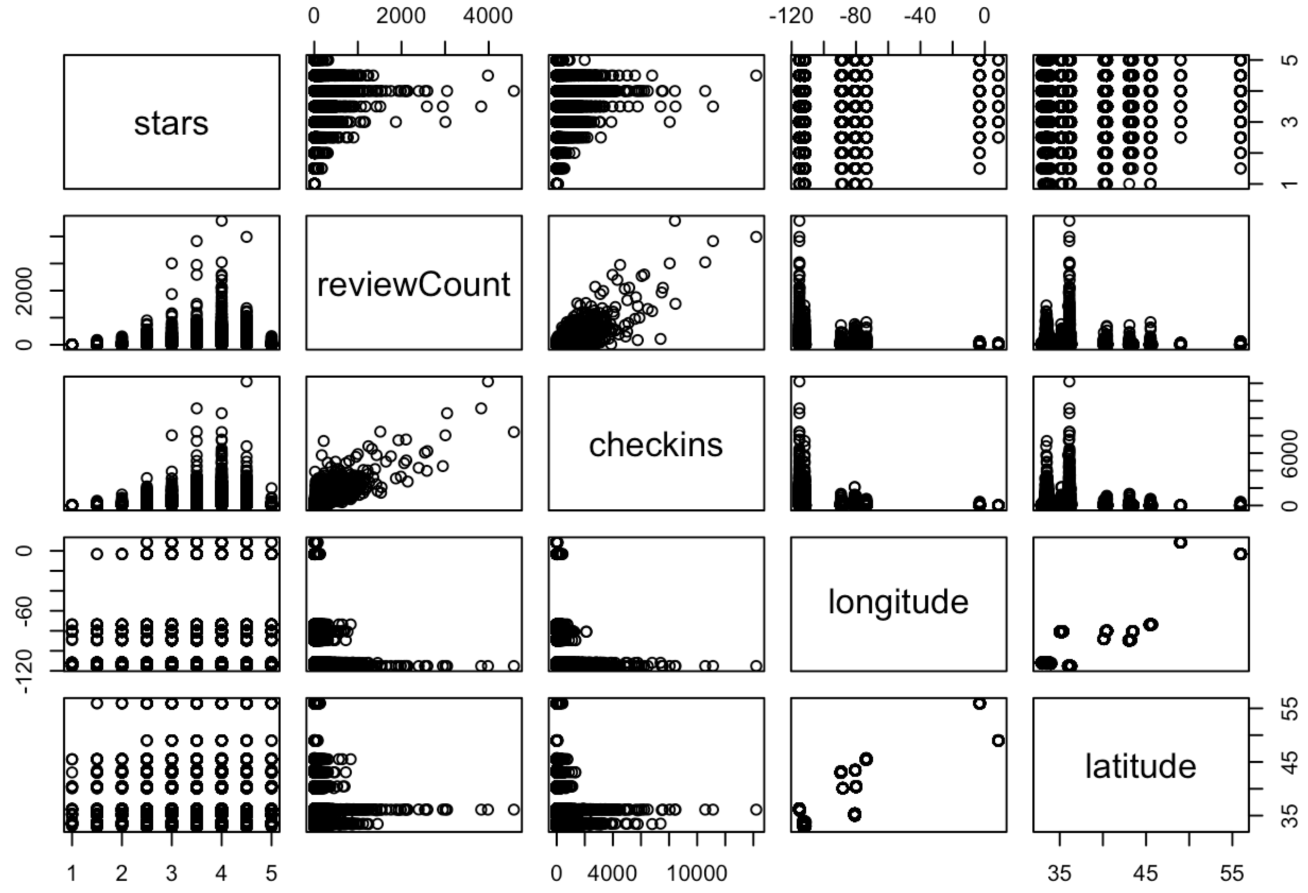
```
##
```

According to the different result showed up in the above summary stats: reviewCount: of course the mean review count in the 25 quantile is 5.247, which is much lower than the mean of whole dataset 49.03 stars: The mean of quantiled stars is 3.418 which is similar to the mean of whole dataset 3.544. attire: Although the amount of casual, dressy and formal are different from original dataset to 25 quantile daraset, the general porportion in side of the attire is matched up. priceRange: We could see that the mean price range in 25 quantile is 1.546 which is slightly less tham original data's 1.631. However, the NA's in the 25 quantile data set is siginicant greater portion compare to the original dataset. delivery: The 25 quantile data has similar pattern with the original data in terms of the True, False and NAs distribution inside of the dataset. goodForKids: The NAs in the 25 quantile data is significant higher proportion than the original dataset.

6 2D plots and correlations

a. (7 pts) Plot a scatterplot matrix (using pairs()) for the five attributes: stars, reviewCount, checkins, longitude, latitude. • Identify which pair of attributes exhibit the most association (as you can determine visually) and discuss if this is interesting or expected, given your domain knowledge.

```
pairs(~ stars + reviewCount + checkins + longitude + latitude, data = yelp)
```



ReviewCount is very related to checkins, which makes a lot of sense in terms of for those customer who reviewed a certain place, they definately tends to checked in already. Another interesting relationship is longitude and latitude, they are strongly associated, I think it makes sense here because we all know that longitude and latitude are related from our instinct.

b. (7 pts) Calculate the pairwise correlation among the above five attributes using cor(). • Identify the pair of attributes with largest positive correlation and the pair with largest negative correlation. Report the correlations and discuss how it matches with your visual assessment in part (a).

```
cor(yelp[,c('stars', 'reviewCount', 'checkins', 'longitude', 'latitude')])
```

```
##                  stars reviewCount    checkins  longitude    latitude
## stars       1.00000000  0.10705060  0.09440071  0.1174446  0.12116308
## reviewCount 0.10705060  1.00000000  0.82749365 -0.1294142 -0.09850936
## checkins    0.09440071  0.82749365  1.00000000 -0.1789531 -0.15260462
## longitude   0.11744458 -0.12941420 -0.17895315  1.0000000  0.88110176
## latitude    0.12116308 -0.09850936 -0.15260462  0.8811018  1.00000000
```

Largest pos correlation except the diagnal is latitude and longtitude. It definately match up with the graph since they are have a strong correlation. Largest neg correlation is between longitude and checkins, I think it is also true in the graph due to the meaningless cluster of points.

c. (7 pts) Plot a boxplot (using boxplot()) for each of the following four attributes (checkins, reviewCount, longitude, latitude) vs. the goodForGroups attribute. Omit outliers using the outline argument. Make sure to label both axes of the plot with the appropriate attribute names. • Identify the attribute that exhibits the most association with goodForGroups (as you can determine visually) and discuss whether this is interesting or expected, given your domain knowledge. It seems like both checkins and reviewCount are kinda associate with goodForGroups from the boxplot. I'll choose to go with Checkins. And I found that both of the attributes are some what interesting. I can't find any direct relationship of how checkins and reviewCount related to weither a place is good for group or not. I mean they are definiately a good decision of weither a place is good or not, but that is not related to group feature from my domain knowledge. • For the attribute identified above, calculate its interquartile range for each value of goodForGroups (i.e., a separate IQR for the TRUE instances and the FALSE instances). You can do this with subset() and quantile(). Calculate the overlap between the two IQRs. Discuss whether these results support the conclusion you made based on visual inspection.

```
checkinGroup = subset(yelp, goodForGroups==TRUE, select = c(checkins))
checkinNotGroup = subset(yelp, goodForGroups==FALSE, select = c(checkins))
quantile(checkinGroup$checkins)
```

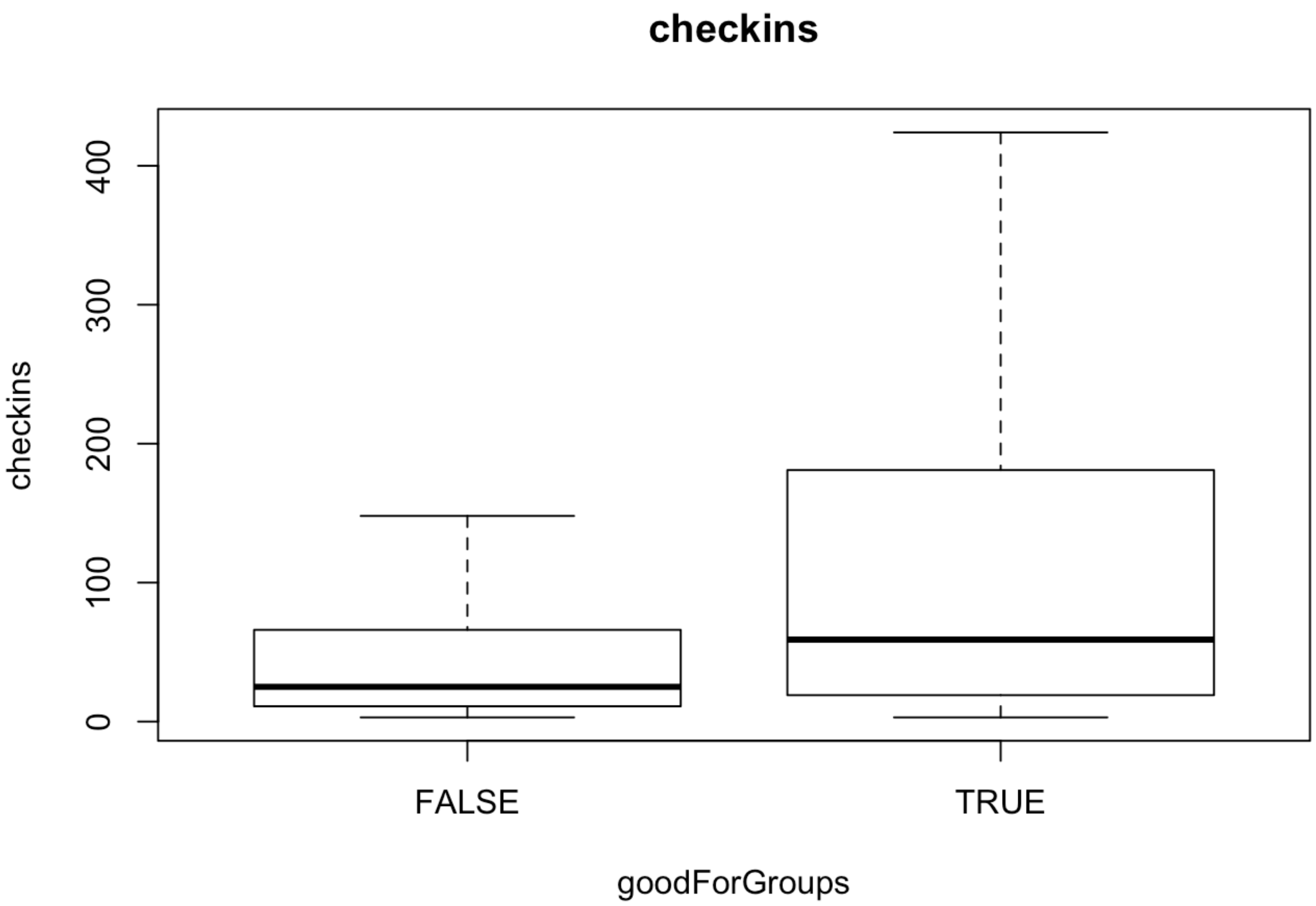
```
##    0%   25%   50%   75%  100%
##     3    19    59   181 14203
```

```
quantile(checkinNotGroup$checkins)
```

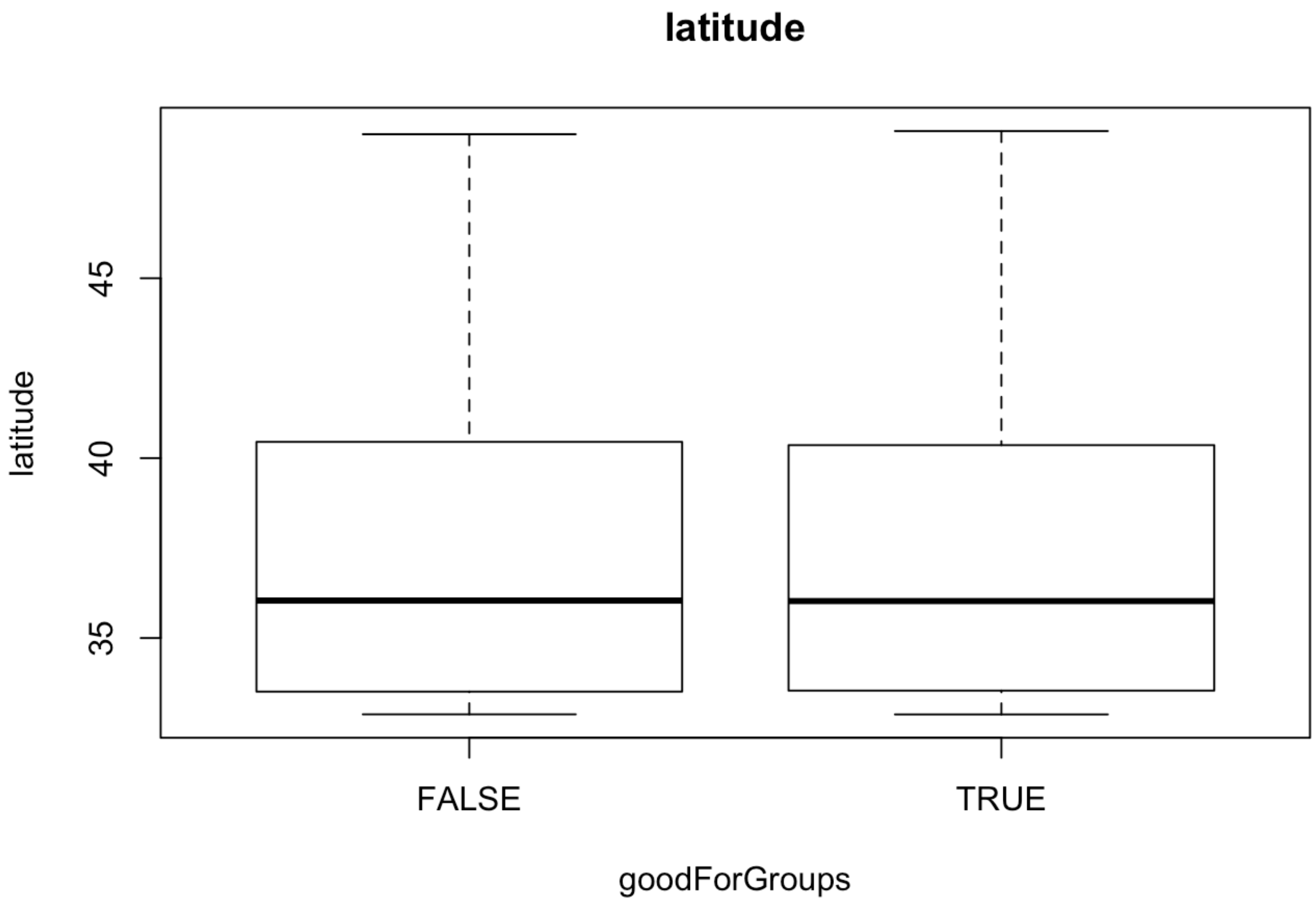
```
##   0%  25%  50%  75% 100%
##    3   11   25   66 6485
```

The overlap between those two IQR is 66 - 19 = 47, which is very minimal compare to the IQR that distinguish the TRUE goodforgroup from false. So, it support the conclusion that I made based on the visual observation.

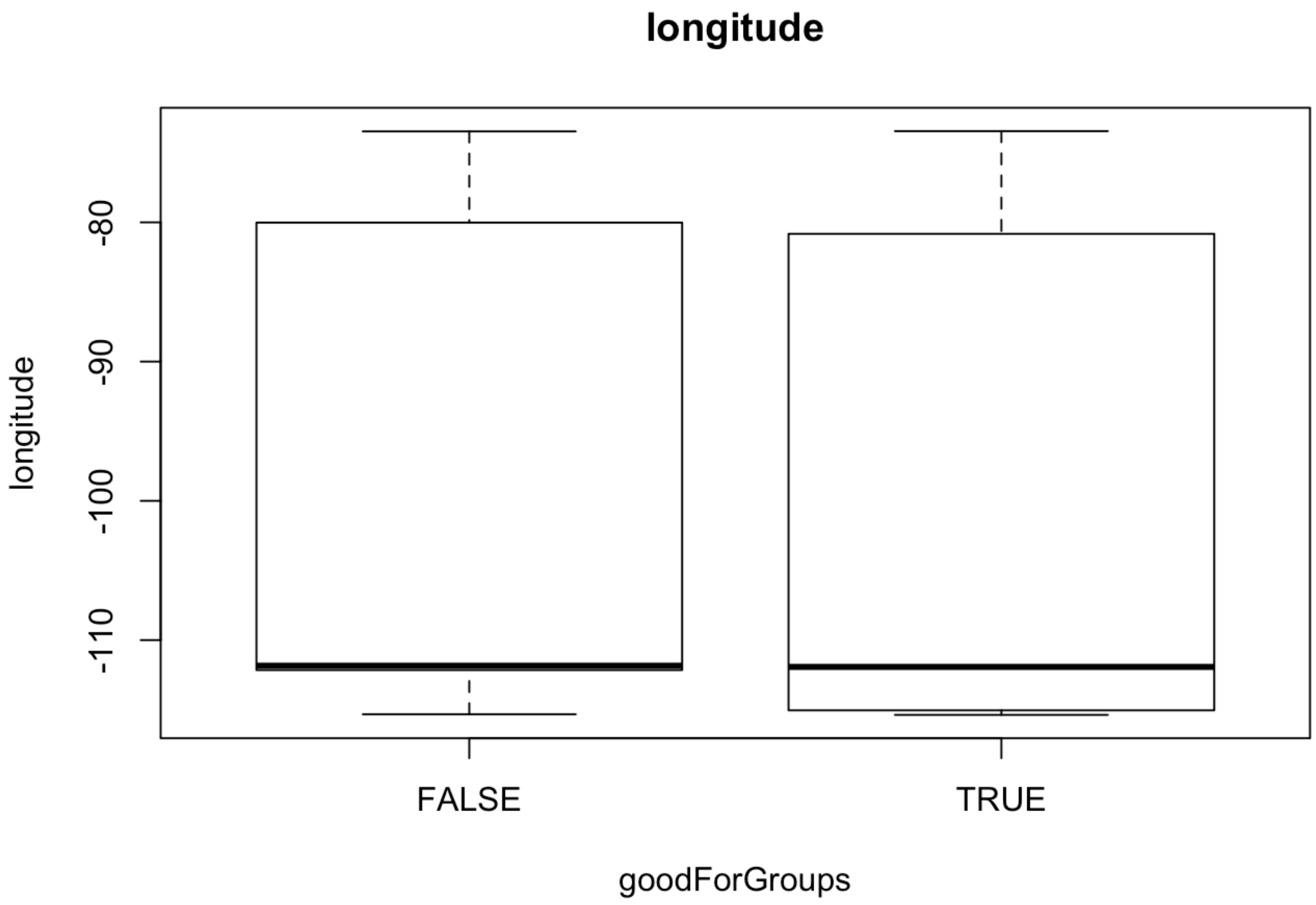
```
boxplot(checkins ~ goodForGroups, data = yelp, outline = FALSE, main="checkins", x
lab= "goodForGroups", ylab="checkins")
```



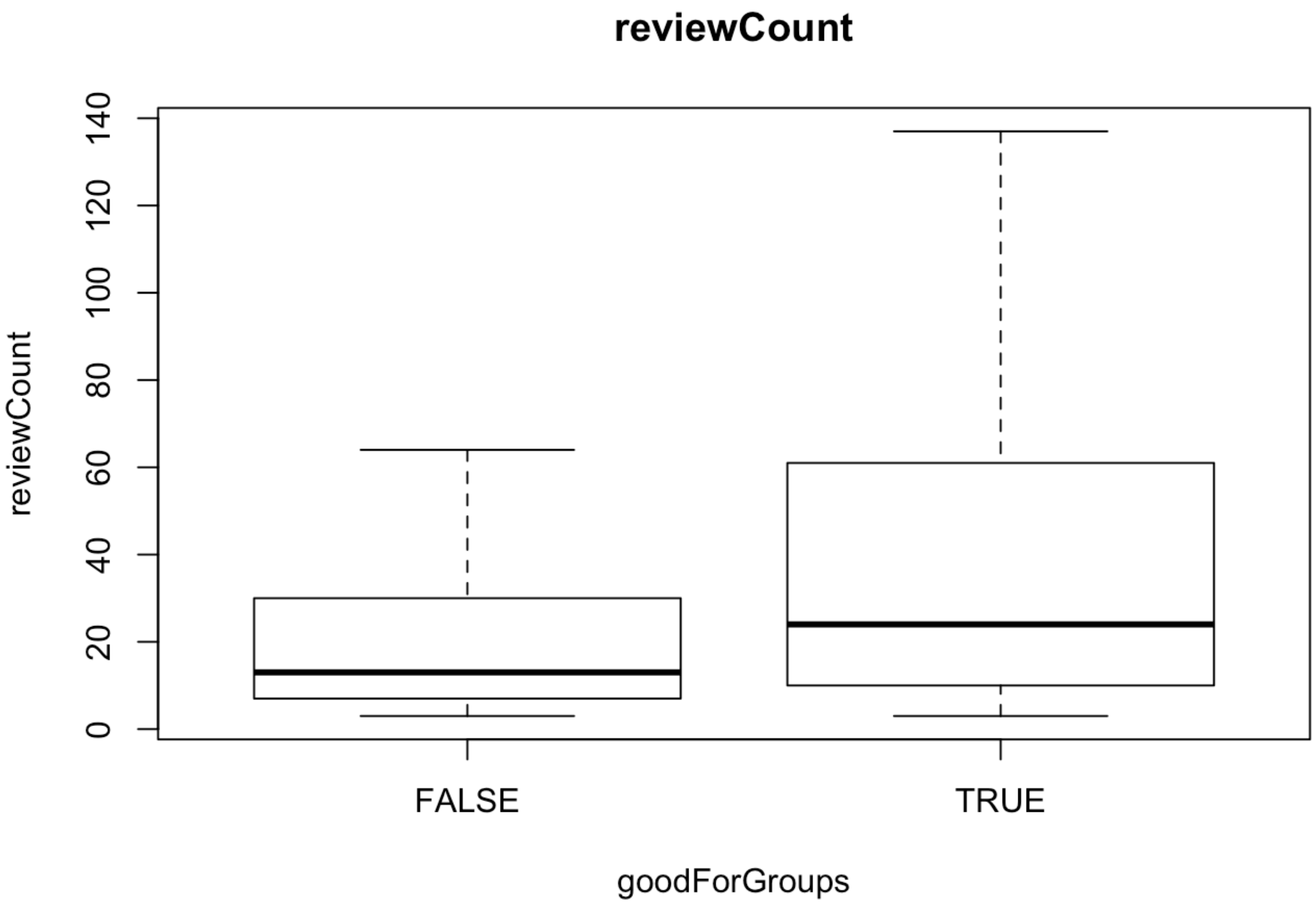
```
boxplot(latitude ~ goodForGroups, data = yelp, outline = FALSE, main="latitude", x
lab= "goodForGroups", ylab="latitude")
```



```
boxplot(longitude ~ goodForGroups, data = yelp, outline = FALSE, main="longitude",
xlab= "goodForGroups", ylab="longitude")
```



```
boxplot(reviewCount ~ goodForGroups, data = yelp, outline = FALSE, main="reviewCou
nt", xlab= "goodForGroups", ylab="reviewCount")
```



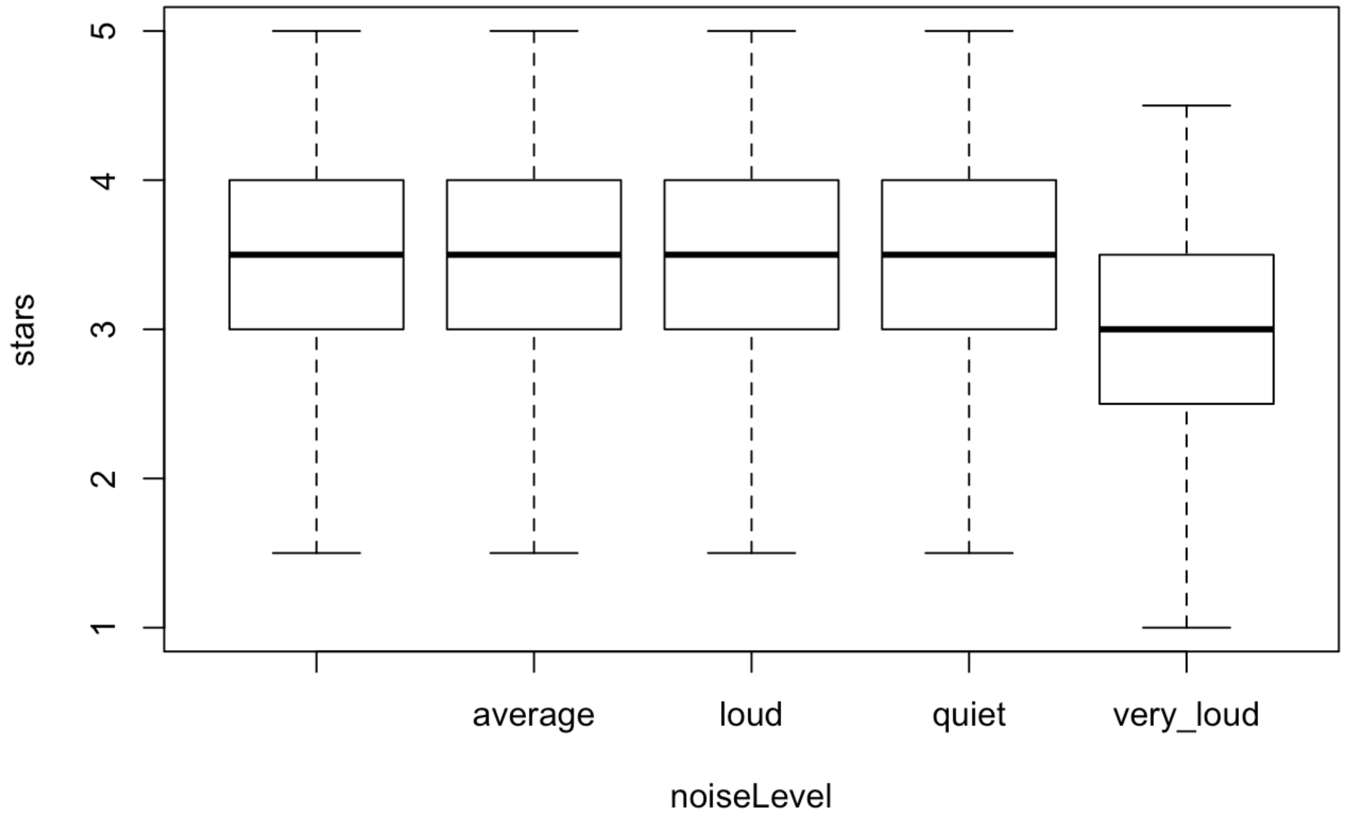
7 Identifying potential hypotheses (20 pts)

During your exploration above, investigate other aspects of the data. Explore relation- ships between variables by assessing plots, computing correlation, or other numerical analysis. Identify TWO possible relationships in the data (other than the ones specified in earlier questions) and formulate hypotheses based on the observed data. For each of the two identified relationships:

Relationship A:

a. Include a plot illustrating the observed relationship (between at least two vari- ables).

```
boxplot(stars ~ noiseLevel, data = yelp, outline = FALSE, xlab = "noiseLevel", yla
b = "stars")
```



b. State whether the variables are discrete or continuous and what type of plot is relevant for comparing these two types of variables. The star is continuous numericle variable, and the noiselevel is discrete catagorical variable. The boxplot that I used before is good for this data.

c. Formulateahypothesisabouttheobservedrelationshipasafunctionoftworandom variables (e.g., X is associated with Y). noiseLevel is assosiated with stars.

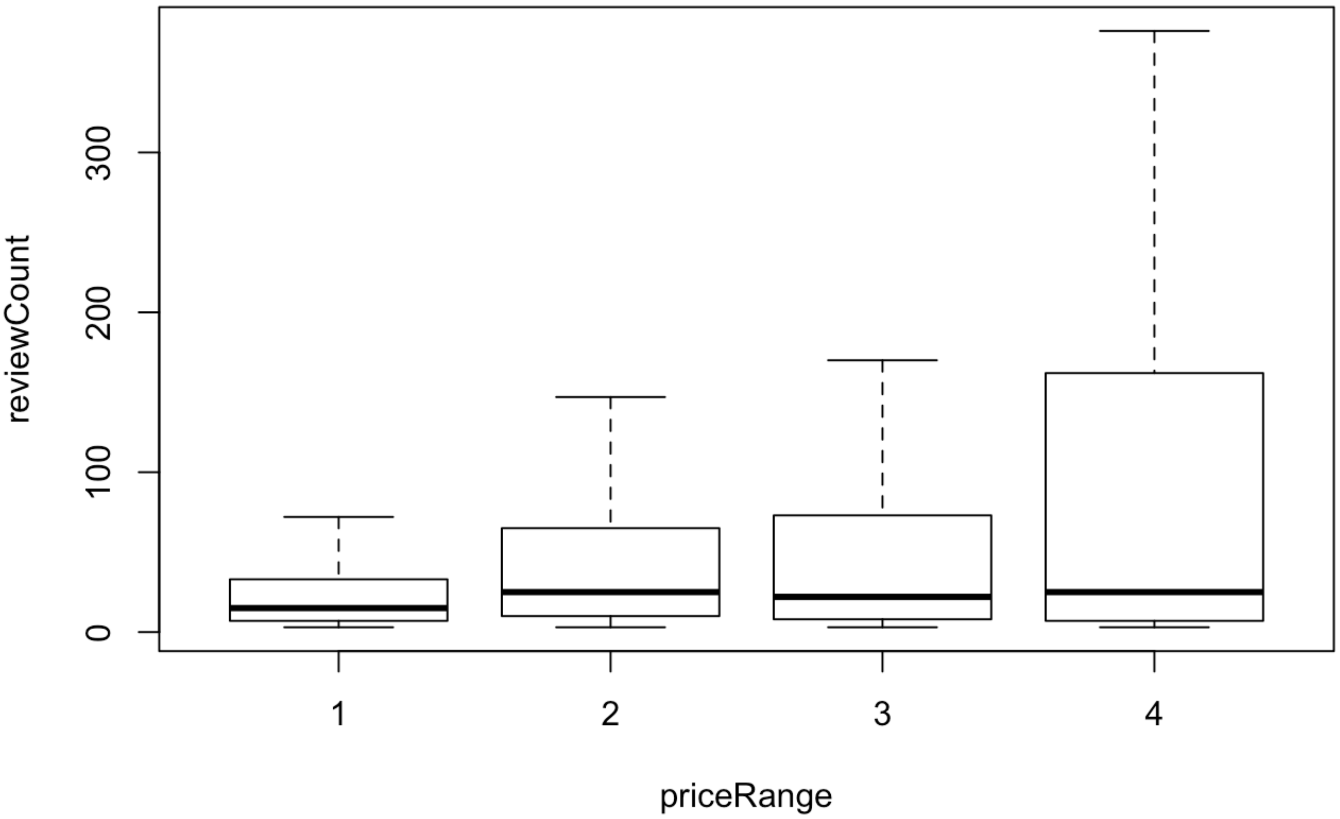
d. Write the hypothesis as a claim in English, relating it to the attributes in the data. The noiseLevel, particularly the very loud noiseLevel will have negative effect on the users review reflecting on the stars of the restauraunt.

e. Identify the type of hypothesis. Directional-relational

Relationship B:

a. Include a plot illustrating the observed relationship (between at least two vari- ables).

```
boxplot(reviewCount ~ priceRange, data = yelp, outline = FALSE, xlab = "priceRange
", ylab = "reviewCount")
```



b. State whether the variables are discrete or continuous and what type of plot is relevant for comparing these two types of variables. reviewCount is a continuous numerical variable, priceRange seems to be a numerical var, but it is actually a discrete catagorical var.

c. Formulateahypothesisabouttheobservedrelationshipasafunctionoftworandom variables (e.g., X is associated with Y). priceRange is assosiated with reviewCount

d. Write the hypothesis as a claim in English, relating it to the attributes in the data. Higher priceRange of the restaurant will tends to have more reviewCount.

e. Identify the type of hypothesis. Directional-relational