Jason Rapsinski
CS 373 HW1
10/1/2020

Part I: Basic Probability and Statistics

1.

a. The sample space for this experiment is infinite. The probability of landing on heads on the i^{th} flip is $\frac{1}{2^i}$, because there must be a sequence of i-1 flips that lands on tails and then a flip that lands on heads. Assuming a fair coin, the probability of the coin landing on heads and tails are the same, $\frac{1}{2}$.

b. The set of outcomes that belong to this event are the set of all sequences where tails is flipped an odd number times and then a head is flipped. The probability of this event occurring is:

$$\sum_{1}^{\infty} \frac{1}{2^{2n-1}} * \frac{1}{2} = \sum_{1}^{\infty} \frac{1}{2^{2n}} = \frac{1}{3}$$

2.

$$P(E) = P(Sum = 3) + P(Sum = 5) + P(Sum = 7) + P(Sum = 9) + P(Sum = 11)$$
$$= \frac{2}{36} + \frac{4}{36} + \frac{6}{36} + \frac{4}{36} + \frac{2}{36} = \frac{18}{36}$$
$$P(F) = 1 - P(\neg F) = 1 - \left(\frac{5}{6}\right)^2 = \frac{11}{36}$$
$$P(G) = P(Sum = 5) = \frac{4}{36}$$

a. The set of E∩F contains {(1,2), (1,4), (1,6), (2,1), (4,1), (6,1)}

$$P(\mathrm{E} \cap \mathrm{F}) = \frac{6}{36}$$

b. The set of EUF contains $P(E \cup F) = P(E) + P(F) - P(\mathrm{E} \cap \mathrm{F}) = \frac{18}{36} + \frac{11}{36} - \frac{6}{36} = \frac{23}{36}$

c. The set of F∩G contains {(1,4), (4,1) }

$$P(\mathrm{E} \cap \mathrm{F}) = \frac{2}{36}$$

d. The set of E∩¬F contains {(2,3), (2,5), (3,2), (3,4), (3,6), (4,3), (4,5), (5,2),(5,4),(5,6),(6,3),(6,5)}

$$P(\mathrm{E} \cap \mathrm{F}) = \frac{12}{36}$$

e. The set of E∩F∩G contains {(1,4), (4,1)}

$$P(\mathrm{E} \cap \mathrm{F}) = \frac{2}{36}$$

3.

a. $P(E) = P(d_1) * P(d_2) + P(d_2) * P(d_3) + P(d_1) * P(d_3) - P(d_1) * P(d_2) * P(d_3) = 0.01 * 0.03 + 0.03 * 0.05 + 0.01 + 0.05 - 0.01 * 0.03 * 0.05 = 0.061785$

b. $P(F) = P(d_1) + P(d_2) * P(d_3) - P(d_1) * P(d_2) * P(d_3) = 0.01 + 0.03 * 0.05 - 0.01 * 0.03 * 0.05 = 0.011485$

c. $P(F|d_3) = P(d_1) + P(d_2) - P(d_1) * P(d_2) = 0.01 + 0.03 - 0.01 * 0.03 = 0.0397$

Jason Rapsinski
CS 373 HW1
10/1/2020

4. $P(Female) = 0.52, P(CS\ Major) = 0.05, P(Female\ \cap\ CS\ Major) = 0.0055$
   a. $P(Female|CS\ Major) = \frac{P(Female \cap CS\ Major)}{P(CS\ Major)} = \frac{0.0055}{0.05} = 0.11 = 11\%$
   b. $P(CS\ Major|Female) = \frac{P(Female \cap CS\ Major)}{P(Female)} = \frac{0.0055}{0.52} = 0.01058 = 1.058\%$
   c. $P(Female\ \cap\ CS\ Major) = P(Female|CS\ Major) * P(CS\ Major) = 0.15 * 0.05 = 0.0075$

$$P(CS\ Major|Female) = \frac{P(Female\ \cap\ CS\ Major)}{P(Female)} = \frac{0.0075}{0.57} = 0.01316 = 1.316\%$$

5.
   a. Let $P(Z) = P(F \cup G)$. Because E, F, and G are all independent of each other, we know that Z and E are independent. Thus, we know that $P(E\ \cap\ Z) = P(E) * P(Z)$ and we can replace P(Z) with the original value to get $P\big(E\ \cap (F \cup G)\big) = P(E) * P(F \cup G)$.
   b. It is given that A and B are independent, so we know that $P(A\ \cap\ B) = P(A) * P(B)$
   To find $P(\neg A \cap\ B)$, we are looking for the set of events that occur in B and not in A, which can be expressed as $P(\neg A \cap\ B) = P(B) - P(A \cap\ B) = P(B) - P(A) * P(B)$ by independence. This same relationship can be expressed as
   $P(\neg A \cap\ B) = P(B)\big(1 - P(A)\big) = P(B) * P(\neg A)$
   thus ¬A and B are independent.
   c. $P(X = 0) = 0.25, P(X = 1) = 0.5, P(X = 2) = 0.25, P(Y = 0) = 0.5, P(Y = 1) = 0.5$
   The events are not independent. If the events were truly independent, then the probability that X and Y both being 1 should be:
   $$P(X = 1 \cap \mathrm{Y} = 1) = P(X = 1) * P(Y = 1) = 0.5 * 0.5 = 0.25$$
   This is not the case in practice. If X=1, then there is exactly one head flipped and one tail, and as such there is not way to have two flips of the same result. Thus, the actual probability $P(X = 1 \cap \mathrm{Y} = 1) = 0$ and the property of independence does not hold for these events.

6.
   a. $E(X_n) = \sum_1^n 1 * P(heads) - 1 * P(tails) = \sum_1^n 1 * p - 1 * (1 - p) = \sum_1^n 2p - 1 = n * (2p - 1)$
   b. $VAR(X_n) = \frac{1}{n}\sum_{i=1}^n \big(x(i) - E(x_n)\big)^2 =$

$$\frac{1}{n}\sum_{i=1}^{n}(1 * p - 1 * (1 - p) - n * (2p - 1))^2 = \frac{1}{n}\sum_{i=1}^{n}(p - 1 + p - n(2p - 1))^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n}\big((2p - 1) - n(2p - 1)\big)^2 = \frac{1}{n}\sum_{i=1}^{n}\big((n - 1) * (2p - 1)\big)^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n}(n^2 - 2n + 1) * (2p - 1)^2 = \frac{1}{n}(n - 1)^2 * n * (2p - 1)^2 = (n - 1)^2 * (2p - 1)^2$$

   c. $E(X_3) = 3 * (2p - 1), VAR(X_3) = (3 - 1)^2 * (2p - 1)^2 = 4 * (2p - 1)^2$

Jason Rapsinski
CS 373 HW1
10/1/2020

7. $E(X) = \sum_1^{25} 2 * 0.8 + 0 * 0.2 + \sum_1^{35} 1 * 0.75 + 0 * 0.25 + \sum_1^{5} 3 * 0.65 + 0 * 0.35 = 40 + 26.25 + 9.75 = 76$ points for the base case
   $E(X) = \sum_1^{25} 2 * 0.8 + (-0.5) * 0.2 + \sum_1^{35} 1 * 0.75 + 0 * 0.25 + \sum_1^{5} 3 * 0.65 + 0 * 0.35 = 37.5 + 26.25 + 9.75 = 73.5$ points for the case where a wrong True/False answer incurs a 0.5-point penalty

8. If the random variables X and Y are defined on the same sample space S, then they can be said to be jointly distributed over some function $f_{x,y}(X, Y)$, which means the expectation can be represented as:
$$E(X+Y) = \sum_i \sum_j (x_i + y_j) * f_{x,y}(X,Y) =$$
$$\sum_i \sum_j x_i * f_{x,y}(X,Y) + \sum_i \sum_j y_j * f_{x,y}(X,Y)$$
   By summing over the full set for a variable, we can find the distribution function for the other variable. Thus, we get the following function
$$\sum_i \sum_j x_i * f_{x,y}(X,Y) + \sum_i \sum_j y_j * f_{x,y}(X,Y) =$$
$$\sum_i x_i * f_x(X) + \sum_j y_j * f_y(Y) = E(X) + E(Y)$$

Jason Rapsinski
CS 373 HW1
10/1/2020

Part II: R

3.

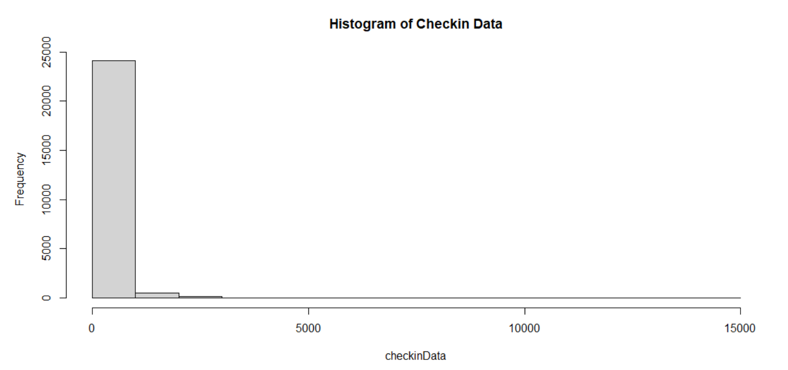
```
> summary(d)
 business_id            name            fullAddress            city          
 Length:24813       Length:24813       Length:24813       Length:24813      
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  



    state              latitude       longitude            stars        reviewCount     
 Length:24813       Min.   :32.88   Min.   :-115.370   Min.   :1.000   Min.   :   3.00  
 Class :character   1st Qu.:33.54   1st Qu.:-114.977   1st Qu.:3.000   1st Qu.:   8.00  
 Mode  :character   Median :36.03   Median :-111.924   Median :3.500   Median :  18.00  
                    Mean   :37.53   Mean   : -97.298   Mean   :3.544   Mean   :  49.03  
                    3rd Qu.:40.41   3rd Qu.: -80.807   3rd Qu.:4.000   3rd Qu.:  48.00  
                    Max.   :55.99   Max.   :   8.549   Max.   :5.000   Max.   :4578.00  

    checkins         open         neighborhoods       categories          alcohol         
 Min.   :    3   Mode :logical   Length:24813       Length:24813       Length:24813      
 1st Qu.:   16   FALSE:3580      Class :character   Class :character   Class :character  
 Median :   48   TRUE :21233     Mode  :character   Mode  :character   Mode  :character  
 Mean   :  166                                                                           
 3rd Qu.:  155                                                                           
 Max.   :14203                                                                           

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

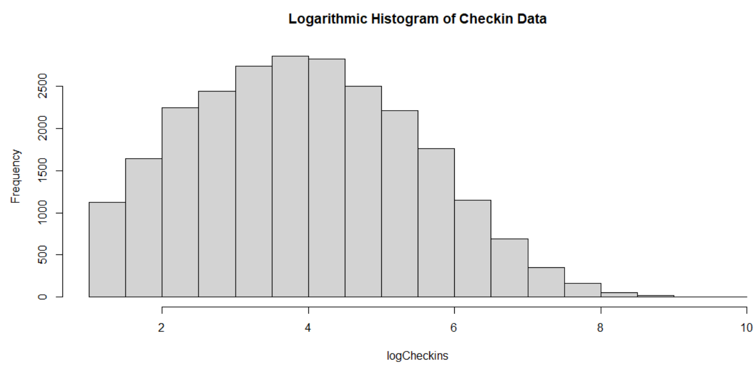
```
> names(d)
 [1] "business_id"         "name"                "fullAddress"        
 [4] "city"                "state"               "latitude"           
 [7] "longitude"           "stars"               "reviewCount"        
[10] "checkins"            "open"                "neighborhoods"      
[13]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

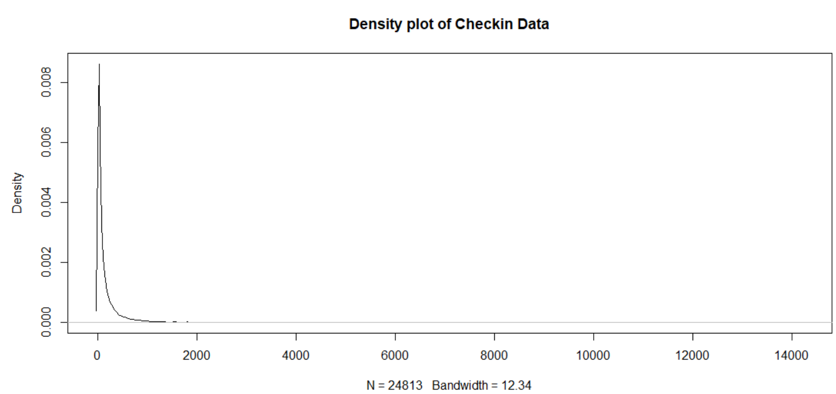
Jason Rapsinski
CS 373 HW1
10/1/2020

4.

A.



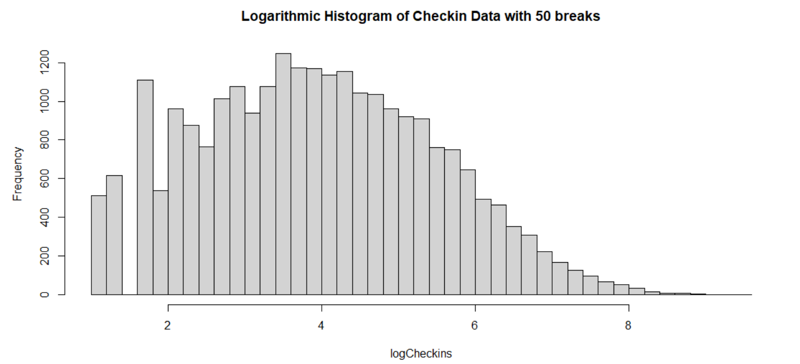


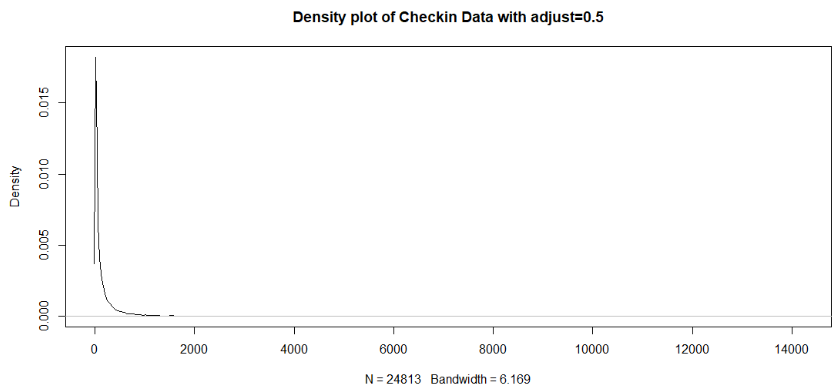


All three graphs depict the data from the “checkins” column of the yelp data. The histogram and density plot show that the distribution of data is tightly concentrated in the regions below 2000. The major difference between these two graphs is that the histogram is discrete and models the explicit values of the data whereas the density plot is continuous and provides information on the specific value with the highest density in the data. The logarithmic histogram provides a clean bell curve of the data which is useful as it makes it easier to approximate the data using normal distributions.

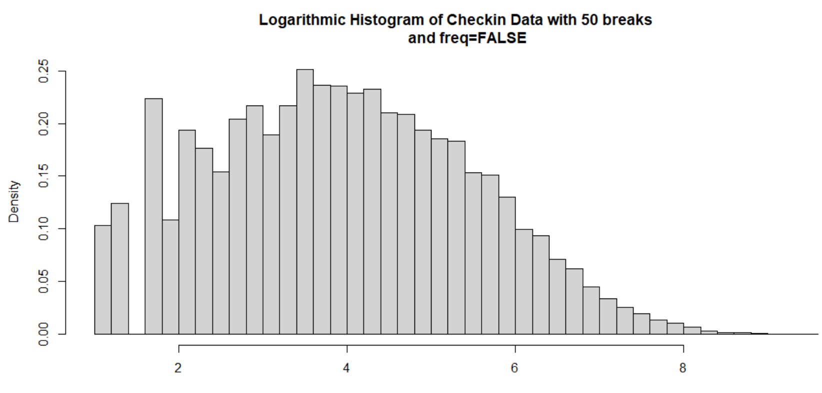
Jason Rapsinski
CS 373 HW1
10/1/2020

B.





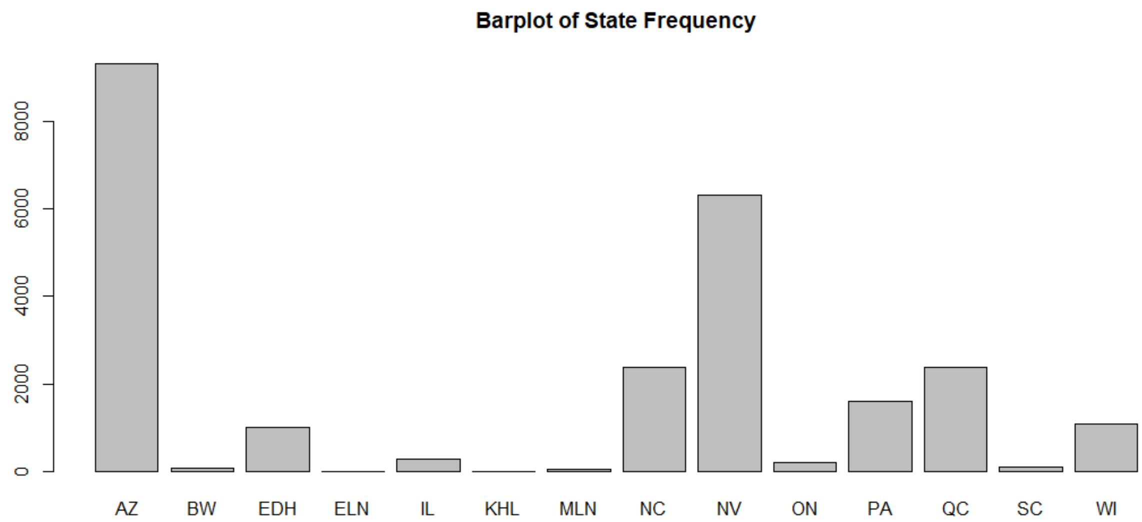
The logarithmic histogram has an increased number of bars across the data which gives a higher resolution look at the data over the domain. The density plot looks very similar to the previous version, but the peak density seems to have almost doubled from 0.008 to 0.015.



When freq=FALSE, the histogram changes from displaying the frequency with which a number appears in the data to displaying the density of each number in the data (i.e. the percentage of results that fall in the bar range).

Jason Rapsinski
CS 373 HW1
10/1/2020

C.



D.

I personally feel like the logarithmic histogram with 50 bars is the most useful because of its similarity to normal distributions makes it easier to model and the existence of the small spike and empty bar on the left act as outliers that can provide useful information.

Jason Rapsinski
CS 373 HW1
10/1/2020

5.

A.

```
> #5A.a transform alcohol and noiseLevel to ordered numeric features
> #   by using factor()
> alcoholLevels <-factor(d$alcohol,
+                        levels=c("full_bar", "beer_and_wine", "none", ""))
> alcoholInts <- as.integer(alcoholLevels)
> noiseLevels <- factor(d$noiseLevel,
+                       levels=c("quiet", "average", "loud", "very_loud", ""))
> noiseInts <- as.integer(noiseLevels)
> #5A.b append new columns to the original table
> d <- cbind(d, alcoholInts)
> d <- cbind(d, noiseInts)
```

Data	
d	24813 obs. of 30 variables
Values	
alcoholInts	int [1:24813] 3 1 1 1 1 1 1 3 2 1 ...
alcoholLevels	Factor w/ 4 levels "full_bar","beer_and_wine",..: 3 1 1 1 1 1 1 3 2 1 ...
noiseInts	int [1:24813] 4 3 1 1 3 5 2 2 2 2 ...
noiseLevels	Factor w/ 5 levels "quiet","average",..: 4 3 1 1 3 5 2 2 2 2 ...

Jason Rapsinski
CS 373 HW1
10/1/2020

B.

```
> quantile(d$reviewCount)
  0%  25%  50%  75% 100% 
   3    8   18   48 4578 
```

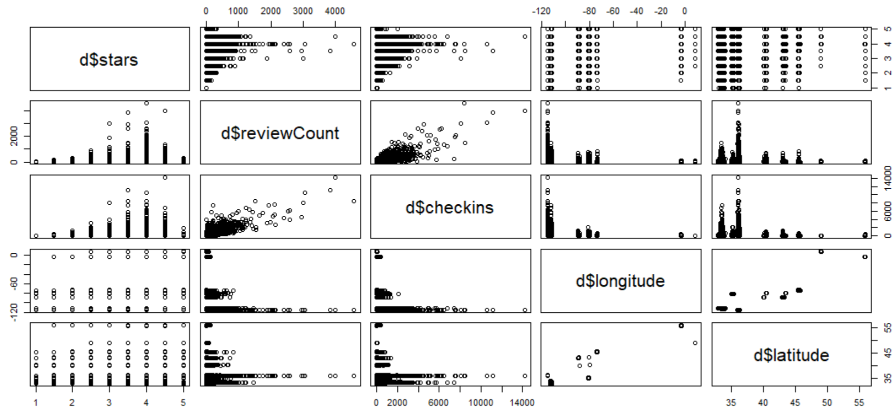
```
 business_id            name            fullAddress             city          
 Length:6960        Length:6960        Length:6960        Length:6960        
 Class :character   Class :character   Class :character   Class :character   
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   

    state              latitude       longitude             stars        reviewCount   
 Length:6960        Min.   :32.88   Min.   :-115.352   Min.   :1.000   Min.   :3.000  
 Class :character   1st Qu.:33.58   1st Qu.:-112.264   1st Qu.:3.000   1st Qu.:4.000  
 Mode  :character   Median :36.08   Median :-111.823   Median :3.500   Median :5.000  
                    Mean   :38.30   Mean   : -94.056   Mean   :3.418   Mean   :5.247  
                    3rd Qu.:43.07   3rd Qu.: -79.998   3rd Qu.:4.000   3rd Qu.:7.000  
                    Max.   :55.99   Max.   :   8.485   Max.   :5.000   Max.   :8.000  

    checkins         open         neighborhoods       categories          alcohol         
 Min.   :  3.00   Mode :logical   Length:6960        Length:6960        Length:6960        
 1st Qu.:  7.00   FALSE:887       Class :character   Class :character   Class :character   
 Median : 13.00   TRUE :6073      Mode  :character   Mode  :character   Mode  :character   
 Mean   : 24.78                                                                            
 3rd Qu.: 29.00                                                                            
 Max.   :694.00                                                                            

  noiseLevel           attire            priceRange     delivery         ambience        
 Length:6960        Length:6960        Min.   :1.000   Mode :logical   Length:6960       
 Class :character   Class :character   1st Qu.:1.000   FALSE:2899      Class :character  
 Mode  :character   Mode  :character   Median :1.000   TRUE :693       Mode  :character  
                                       Mean   :1.546   NA's :3368                        
                                       3rd Qu.:2.000                                     
                                       Max.   :4.000                                     
                                       NA's   :825                                       
   parking          dietaryRestrictions waiterService     smoking          outdoorSeating 
 Length:6960        Length:6960         Mode :logical   Length:6960        Mode :logical  
 Class :character   Class :character    FALSE:1323      Class :character   FALSE:2672     
 Mode  :character   Mode  :character    TRUE :1729      Mode  :character   TRUE :1370     
                                        NA's :3908                         NA's :2918     

   caters        recommendedFor     goodForGroups   goodForKids      alcoholInts   
 Mode :logical   Length:6960        Mode :logical   Mode :logical   Min.   :1.000  
 FALSE:1040      Class :character   FALSE:704       FALSE:15        1st Qu.:3.000  
 TRUE :620       Mode  :character   TRUE :3471      TRUE :31        Median :3.000  
 NA's :5300                         NA's :2785      NA's :6914      Mean   :2.684  
                                                                    3rd Qu.:3.000  
                                                                    Max.   :3.000  

   noiseInts    
 Min.   :1.000  
 1st Qu.:2.000  
 Median :5.000  
 Mean   :3.736  
 3rd Qu.:5.000  
 Max.   :5.000  
```

The number of entries for each attribute decreased from 24813 data points in the original summary to 6960 data points in the 1st quantile summary. This makes sense as the quantile dataset should contain slightly more than 25% of the datapoints of the original dataset. The distribution of the numbers are overall similar, except for the reviewCount and checkins attributes, which were markedly decreased between the original and quantile dataset.

Jason Rapsinski
CS 373 HW1
10/1/2020

6.

A.



The strongest visual relationship is between review count and checkins, which makes sense as both are likely correlated with overall popularity of the restaurant. Additionally, both review count and checkins were roughly correlated with the number of stars on a restaurant, with both peaking for restaurants with 4 stars. Additionally, latitude and longitude seem to have a roughly positive correlation.

B.

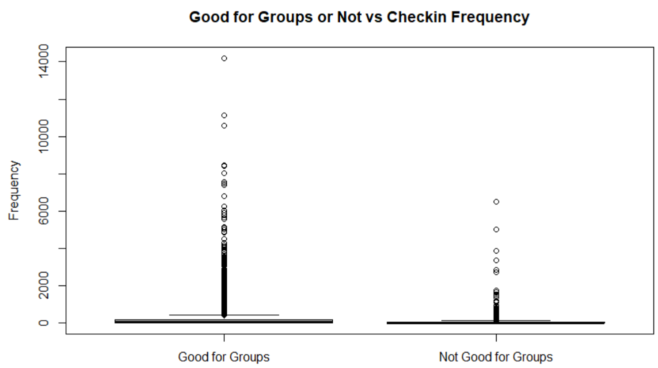
Pair-wise Correlation Table

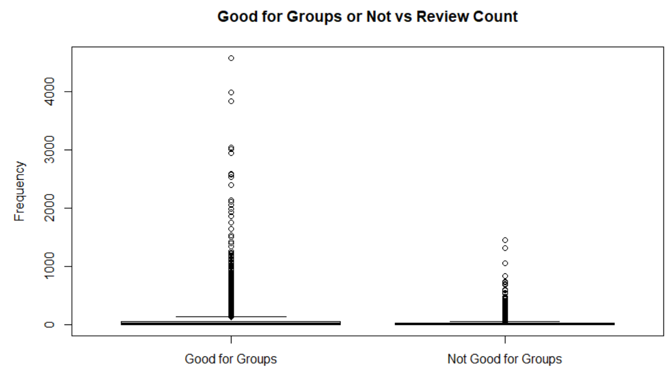
Stars				
0.0107050	Review Count			
0.0944007	0.8274936	Checkins		
0.1174446	-0.1294142	-0.1789531	Longitude	
0.1211631	-0.09850094	-0.1526046	0.8811018	Latitude

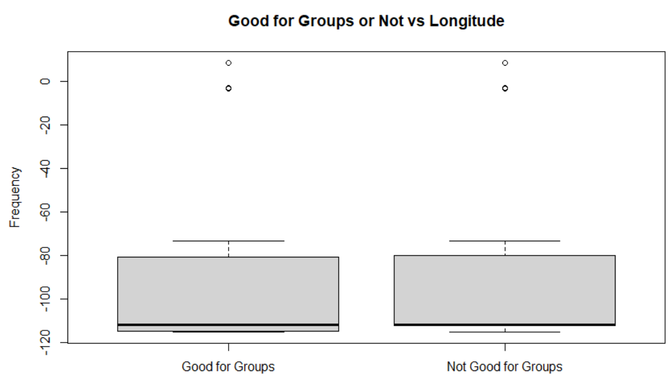
The pair of attributes with the strongest positive correlation are Longitude and Latitude, with the pair Review count and Checkins being a close second. This matches the visual assessment that was performed earlier. The pair of attributes with the strongest negative correlation was

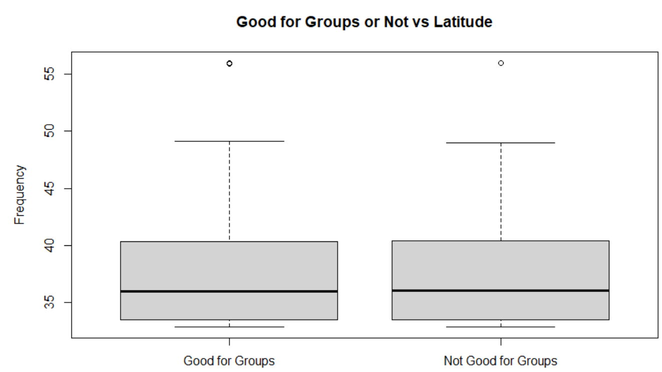
Jason Rapsinski

C.









Based on visual inspection, the graph that looks like it has the strongest association for the 'goodforgroup' attribute is the latitude graph. This is most likely due to its lack of outliers which makes the box plot much easier to read. If we look at ranges, the strongest correlation would most likely be in the checkin frequency plot. This makes some sense as people in groups are more likely to check in at a restaurant as they communicate with friends about their location.

Jason Rapsinski
CS 373 HW1
10/1/2020

Interquartile Ranges for Attributes when ‘goodforgroups’=TRUE and FALSE

Attributes	Good for Groups IQR	Not Good for Groups IQR
Checkins	> quantile(good$checkins) 0% 25% 50% 75% 100% 3 19 59 181 14203 IQR =162	> quantile(notGood$checkins) 0% 25% 50% 75% 100% 3 11 25 66 6485 IQR=55
Review Count	> quantile(good$reviewCount) 0% 25% 50% 75% 100% 3 10 24 61 4578 IQR=51	> quantile(notGood$reviewCount) 0% 25% 50% 75% 100% 3 7 13 30 1453 IQR=23
Longitude	> quantile(good$longitude) 0% 25% 50% 75% 100% -115.36973 -115.04307 -111.92574 -80.82606 8.54856 IQR=34.22	> quantile(notGood$longitude) 0% 25% 50% 75% 100% -115.328981 -112.152874 -111.840497 -80.018910 8.410954 IQR=32.13
Latitude	> quantile(good$latitude) 0% 25% 50% 75% 100% 32.87687 33.53849 36.02708 40.36092 55.99042 IQR=6.822	> quantile(notGood$latitude) 0% 25% 50% 75% 100% 32.87918 33.51192 36.04116 40.45204 55.97743 IQR=6.94

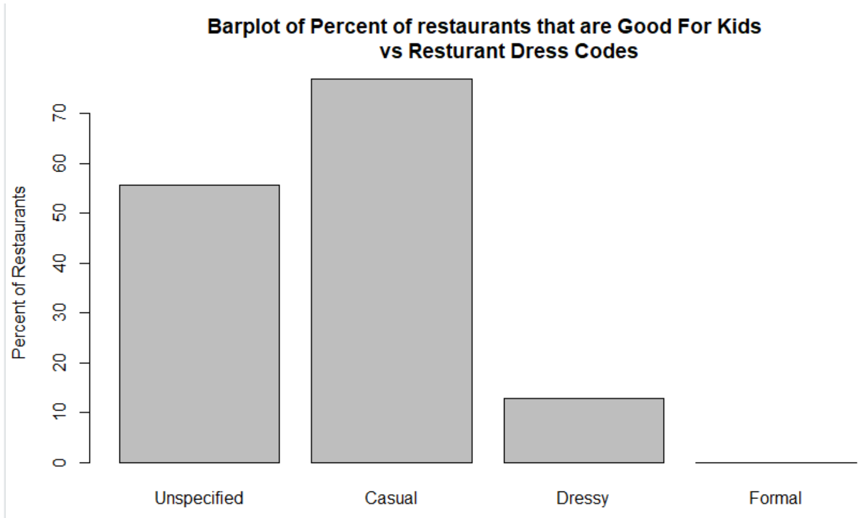
The interquartile ranges support my visual analysis of the checkins box plot, but not my analysis of the latitude box plot. Additionally it would seem that there is some correlation between the ‘goodforgroups’ attribute and the review count.

Jason Rapsinski
CS 373 HW1
10/1/2020

7.

Hypothesis 1:

a.

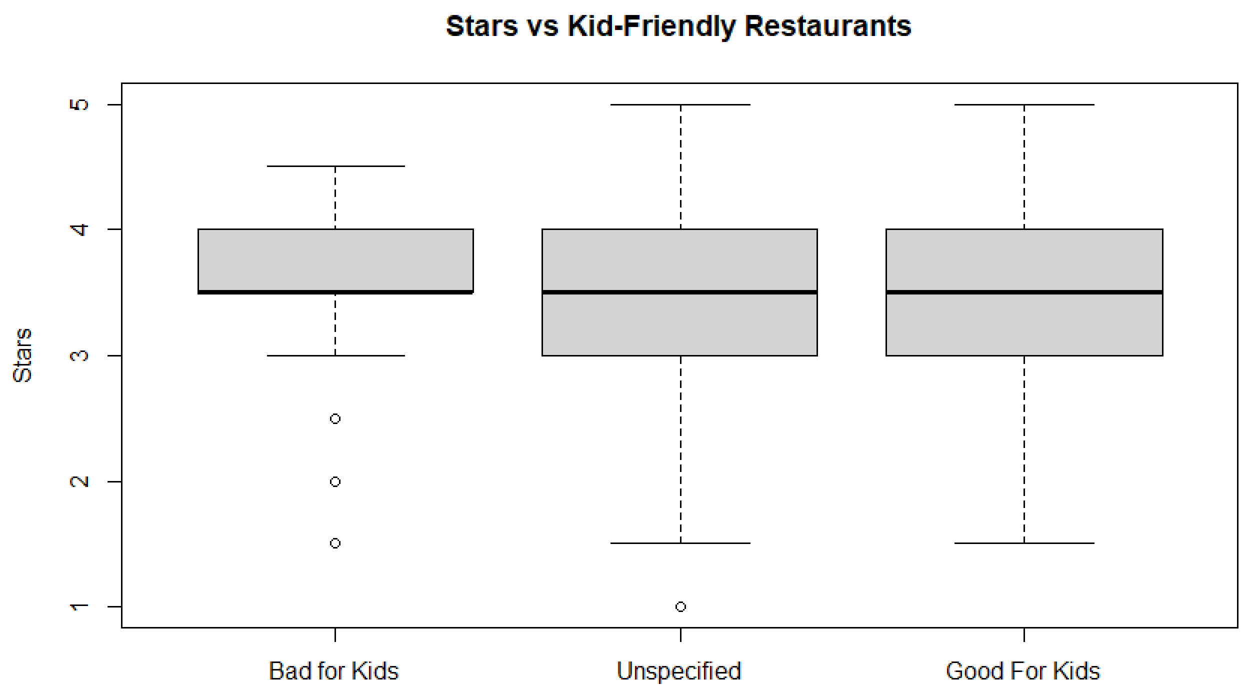


b. Both variables are discrete. There isn't really a good way to directly compare discrete variables, so I modified the data to find the percentage of restaurants that were good for kids with a given dress code.

A Hypothesis for this data set might be "Restaurants with fancier dress codes are less likely to be considered child-friendly because they are targeting more mature clientele." This hypothesis outlines a causal relationship as it proposes a mechanism by which the variables are correlated

Hypothesis 2:

a.



b. Both variables are discrete. Stars has 9 numeric values which means that it could be used to find distribution averages by using a boxplot.

A Hypothesis for this relationship might be, “Reviews that specified a restaurant was not child friendly had a narrower distribution than reviews that specified a restaurant was child friendly or did not specify.” This would be a descriptive hypothesis because it is just describing some nature of the relationship and posits neither a relationship, nor a causal link between the variables.

Code

```
#read data from yelp.csv and store in variable 'd'

d <- read.table("C:\\Users\\Jason\\Desktop\\CS 373\\hw1\\yelp.csv", sep = ",", header = TRUE,

                quote = "\"", comment.char="")


#3.a print a summary of the data using summary()
```

```
summary(d)

#3.b print names of the columns using names()

names(d)


#4A.a plot a histogram of checkin data

checkinData <- d$checkins

hist(checkinData, main="Histogram of Checkin Data")


#4A.b plot a logarithmic histogram of checkin data

logCheckins <- log(d$checkins)

hist(logCheckins, main="Logarithmic Histogram of Checkin Data")


#4A.c create a density plot of checkin data

checkinDensity <- density(checkinData)

plot(checkinDensity, main = "Density plot of Checkin Data")


#4B.a plot a logarithmic histogram of checkin data with breaks=50

hist(logCheckins, main = "Logarithmic Histogram of Checkin Data with 50 breaks",

    breaks=50)


#4B.b create a density plot of checkin data with adjust = 0.5

checkinDensity <- density(checkinData, adjust=0.5)

plot(checkinDensity, main = "Density plot of Checkin Data with adjust=0.5")


#4B.d plot a logarithmic histogram of checkin data with breaks=50

#   and freq = FALSE

hist(logCheckins, main = "Logarithmic Histogram of Checkin Data with 50 breaks

     and freq=FALSE",

    breaks=50, freq=FALSE)
```

Jason Rapsinski
CS 373 HW1
10/1/2020

```
#4C create a frequency barplot for the state attribute
stateData <- table(d$state)
stateNames <- names(stateData)
barplot(stateData, main = "Barplot of State Frequency", names.arg = stateNames)

#5A.a transform alcohol and noiseLevel to ordered numeric features
#   by using factor()
alcoholLevels <-factor(d$alcohol,
                 levels=c("full_bar", "beer_and_wine", "none", ""))
alcoholInts <- as.integer(alcoholLevels)

noiseLevels <- factor(d$noiseLevel,
                levels=c("quiet", "average", "loud", "very_loud", ""))
noiseInts <- as.integer(noiseLevels)

#5A.b append new columns to the original table
d <- cbind(d, alcoholInts)
d <- cbind(d, noiseInts)

#5B.a compute quantiles for the reviewCount attribute
quantile(d$reviewCount)
#5B.b create a subset that is the first quantile of reviewCount
firstQuantile <- subset(d, d$reviewCount <=8)
#5B.c create a summary of the first quantile and compare
summary(firstQuantile)

#6A.a create a scatterplot with the following attributes
#     stars, reviewCount, checkins, longitude, and latitude
```

Jason Rapsinski
CS 373 HW1
10/1/2020

```
scatterData <- data.frame(d$stars, d$reviewCount, d$checkins,
               d$longitude, d$latitude)
pairs(~ d$stars + d$reviewCount + d$checkins + d$longitude +
    d$latitude, data= scatterData)

#6B.a use cor() to calculate the pairwise correlation for all pairs
#   I couldn't get a cleaner implementation to work
cor(d$stars, d$reviewCount)
cor(d$stars, d$checkins)
cor(d$stars, d$longitude)
cor(d$stars, d$latitude)
cor(d$reviewCount, d$checkins)
cor(d$reviewCount, d$longitude)
cor(d$reviewCount, d$latitude)
cor(d$checkins, d$longitude)
cor(d$checkins, d$latitude)
cor(d$longitude, d$latitude)

#6C.a use boxplot() to model checkins, reviewcount, longitude, and latitude vs.
#    the 'goodforgroups' attribute
good<- subset(d, d$goodForGroups==TRUE)
notGood <- subset(d, d$goodForGroups==FALSE)
boxplot(good$checkins, notGood$checkins,
    main = "Good for Groups or Not vs Checkin Frequency",
    names = c("Good for Groups","Not Good for Groups"),
    ylab = "Frequency")
boxplot(good$reviewCount, notGood$reviewCount,
    main = "Good for Groups or Not vs Review Count",
    names = c("Good for Groups","Not Good for Groups"),
```

```
    ylab = "Frequency")
boxplot(good$longitude, notGood$longitude,
    main = "Good for Groups or Not vs Longitude",
    names = c("Good for Groups","Not Good for Groups"),
    ylab = "Frequency")
boxplot(good$latitude, notGood$latitude,
    main = "Good for Groups or Not vs Latitude",
    names = c("Good for Groups","Not Good for Groups"),
    ylab = "Frequency")

#6C.c Checking interquartile ranges for checkins, reviewCount, Longitude, and
#     Latitude for 'goodforgroups'=TRUE and FALSE
quantile(good$checkins)
quantile(notGood$checkins)
quantile(good$reviewCount)
quantile(notGood$reviewCount)
quantile(good$longitude)
quantile(notGood$longitude)
quantile(good$latitude)
quantile(notGood$latitude)

#7A compare attire with presence of goodForKids after excluding null options
dKids <- subset(d, d$goodForKids!="")

uAttire <- subset(dKids, dKids$attire=="")
uAttireKids <- subset(uAttire, uAttire$goodForKids==TRUE)
uProb <- 100*as.double(nrow(uAttireKids))/as.double(nrow(uAttire))

cAttire <- subset(dKids, dKids$attire=="casual")
```

Jason Rapsinski
CS 373 HW1
10/1/2020

```
cAttireKids <- subset(cAttire, cAttire$goodForKids==TRUE)

cProb <- 100*as.double(nrow(cAttireKids))/as.double(nrow(cAttire))


dAttire <- subset(dKids, dKids$attire=="dressy")

dAttireKids <- subset(dAttire, dAttire$goodForKids==TRUE)

dProb <- 100*as.double(nrow(dAttireKids))/as.double(nrow(dAttire))


fAttire <- subset(dKids, dKids$attire=="formal")

fAttireKids <- subset(fAttire, fAttire$goodForKids==TRUE)

fProb <- 100*as.double(nrow(fAttireKids))/as.double(nrow(fAttire))


data <- table(c(uProb, cProb, dProb, fProb))

dataNames <- c("Unspecified", "Casual", "Dressy", "Formal")

barplot(c(uProb, cProb, dProb, fProb),

    main = "Barplot of Percent of restaurants that are Good For Kids

    vs Resturant Dress Codes",

    names.arg = dataNames,

    ylab = "Percent of Restaurants")


#7B compare attire with reviewCount

badKid <- subset(d, d$goodForKids==FALSE)

unspecifiedKid <- subset(d, is.na(d$goodForKids))

goodKid <- subset(d, d$goodForKids==TRUE)


boxplot(badKid$stars, unspecifiedKid$stars, goodKid$stars,

    main = "Stars vs Kid-Friendly Restaurants",

    names = c("Bad for Kids","Unspecified", "Good For Kids"),

    ylab = "Stars")
```