CS57300: Homework 1

Long Zhen, lzhen@purdue.edu

February 1, 2015

1 Counting (2 pts)

Consider a simple password system. There are 26 lowercase letters, 26 uppercase letters, 10 digits, and 30 special characters on a keyboard.

(a) Our system accepts passwords of 6-10 characters, how many unique passwords are there containing only lowercase letters and digits? How many if the system requires at least 1 digit?

Possible characters with lowercase letters and digits = 26 + 10 = 36
Unique passwords = $\sum_{i=6}^{10} 36^i$
Unique passwords with at least 1 digit = $\sum_{i=6}^{10} \binom{i}{1} \cdot 10 \cdot n^{i-1}$, where $n = 36$ if password only contains lowercase and digits, otherwise, $n = 26 + 26 + 10 + 30 = 92$

(b) Our system accepts passwords of 6-10 characters, how many unique passwords are there containing uppercase, lowercase, digits and special characters? How many if the system requires at least 1 digit, uppercase and special character?

Possible characters with uppercase, lowercase, digits and special characters = 26 + 26 + 10 + 30 = 92
Unique passwords = $\sum_{i=6}^{10} 92^i$
Unique passwords with at least 1 digit, uppercase and special character = $\sum_{i=6}^{10} {}^iP_3 \cdot 10 \cdot 26 \cdot 30 \cdot 92^{i-3}$, where P is permutation

2 Axioms of probability (2 pts)

(a) Prove that $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$ using the axioms of probability.

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- $A \cap B = \{x | x \in A \cap x \in B\}$
- $A - B = \{x | x \in A \cap x \notin B\}$
- $B - A = \{x | x \in B \cap x \notin A\}$

These three above are disjoint based on their definitions. Because of disjoint property,
$P(A - B) + P(B - A) + P(A \cap B) = P(A \cup B)$ ①
$P(A - B) + P(A \cap B) = P(A) \Rightarrow$ ① $\Rightarrow P(A) + P(B - A) = P(A \cup B)$ ②
$P(B - A) + P(A \cap B) = P(B) \Rightarrow$ ② $\Rightarrow P(A) + P(B) - P(A \cap B) = P(A \cup B)$

(b) Prove the conditional version of Bayes rule:

$$P(B|A,C) = \frac{P(A|B,C)P(B|C)}{P(A|C)}$$

$$\begin{aligned} P(B|A \cap C) &= \frac{P(B \cap (A \cap C))}{P(A \cap C)} \\ &= \frac{P(A|B \cap C)P(B \cap C)}{P(A|C)P(C)} \\ &= \frac{P(A|B \cap C)P(B|C)P(C)}{P(A|C)P(C)} \\ &= \frac{P(A|B \cap C)P(B|C)}{P(A|C)} \end{aligned}$$

3 Probability and conditional probability (3 pts)

(a) Three men toss coins to see who pays for coffee. If all three match, they toss again. Otherwise the "odd man" pays for coffee.

(i) What is the probability that they will need to do this more than once?

$$\begin{aligned} P(\textit{toss only once}) &= \frac{\binom{3}{1}\binom{2}{1}}{2^3} \\ &= \frac{3}{4} \\ P(\textit{do this more than once}) &= 1 - \frac{3}{4} \\ &= \frac{1}{4} \end{aligned}$$

(ii) What is the probability of tossing at most twice?

$$\begin{aligned} P &= P(\textit{toss only once}) + P(\textit{toss twice}) \\ &= \frac{3}{4} + \frac{1}{4} \cdot \frac{3}{4} \\ &= \frac{15}{16} \end{aligned}$$

(b) Alice and Bob are playing a simple dice game. Each rolls one dice and the one with higher number wins. If the numbers are the same, they roll again. If Alice just won, what is the probability that she rolled a '5'?

Here are all the cases when Alice won:

- Alice rolls 1, impossible to win
- Alice rolls 2, Bob rolls 1
- Alice rolls 3, Bob rolls 1 or 2
- Alice rolls 4, Bob rolls 1, 2 or 3

- Alice rolls 5, Bob rolls 1, 2, 3 or 4
- Alice rolls 6, Bob rolls 1, 2, 3, 4 or 5

P(Alice rolled 5 when winning) $= \frac{4}{1+2+3+4+5} = \frac{4}{15}$

4 Probability distributions (3 pts)

Let X be a discrete random variable such that $P(X = x) > 0$ if $x = 1, 2, 3,$ or 4 and $P(X = x) = 0$ otherwise. Suppose the CDF is $F(x) = 0.05x(1 + x)$ at the values $x = 1, 2, 3,$ or 4.

(a) Sketch the graph of the CDF.

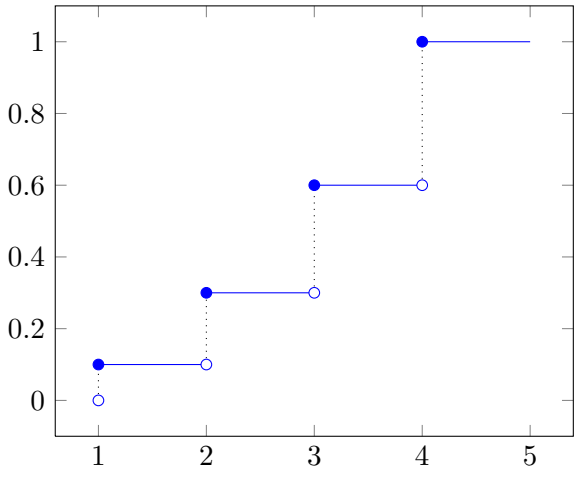

Figure 1: CDF of Question 4(a)

(b) Sketch the graph of the discrete pdf $f(x)$.

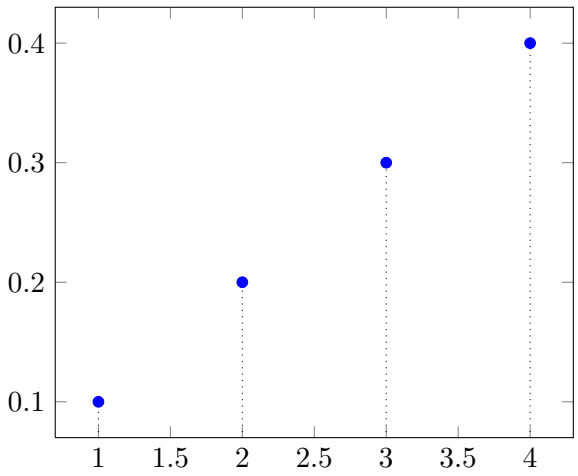

Figure 2: Discrete PDF $f(x)$ of Question 4(b)

(c) Find $E(X)$ and $Var(X)$.

$$\begin{aligned} E(X) &= 1 \times 0.1 + 2 \times 0.2 + 3 \times 0.3 + 4 \times 0.4 \\ &= 0.1 + 0.4 + 0.9 + 1.6 \\ &= 3 \\ Var(X) &= 0.1 \times 1 + 0.2 \times 4 + 0.3 \times 9 + 0.4 \times 16 - 9 \\ &= 1 \end{aligned}$$

5 Independence (3 pts)

A box contains four disks that have different colors on each side. Disk 1 is red and green, disk 2 is red and white, disk 3 is red and black, and disk 4 is green and white. One disk is selected at random from the box. Define the events as follows: A = one side is red, B = one side is green, C = one side is white, and D = one side is black.

(a) Are A and B independent events? Why or why not?

$$\begin{aligned} \because P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{1/4}{2/4} = \frac{1}{2} \\ P(A) &= \frac{3}{4} \\ \therefore P(A|B) &\neq P(A), \textit{ not independent} \end{aligned}$$

(b) Are B and C independent events? Why or why not?

$$\begin{aligned} \because P(B|C) &= \frac{P(B \cap C)}{P(C)} = \frac{1/4}{2/4} = \frac{1}{2} \\ P(B) &= \frac{2}{4} = \frac{1}{2} \\ \therefore P(B|C) &= P(B), \textit{ independent} \end{aligned}$$

(c) Are any pair of events mutually exclusive? Which ones?

Mutually exclusive pairs:

- B and D
- C and D

6 Conditional Expectation (2 pts)

If X and Y are jointly distributed random variables, then the conditional expectation and conditional variance of Y given X are given by:

$$E(Y|x) = \sum_y y \cdot p(y|x)$$

$$Var(Y|x) = E(Y^2|x) - [E(Y|x)]^2$$

Let X and Y be discrete random variables with joint pdf $p(x,y) = 48/(45xy)$ if $x = 2,4$ and $y = 1,4$, and zero otherwise. Determine $E(Y|x)$ and $Var(Y|x)$.

$$
\begin{aligned}
&When\ x = 2 \\
E[Y|x=2] &= 1 \times \frac{48/90}{60/90} + 4 \times \frac{12/90}{60/90} \\
&= 0.8 + 0.8 \\
&= 1.6 \\
Var[Y|x=2] &= 1 \times \frac{48/90}{60/90} + 16 \times \frac{12/90}{60/90} - 1.6^2 \\
&= 0.8 + 3.2 - 1.6^2 \\
&= 1.44 \\
&When\ x = 4 \\
E[Y|x=4] &= 1 \times \frac{24/90}{30/90} + 4 \times \frac{6/90}{30/90} \\
&= 0.8 + 0.8 \\
&= 1.6 \\
Var[Y|x=4] &= 1 \times \frac{24/90}{30/90} + 16 \times \frac{6/90}{30/90} - 1.6^2 \\
&= 0.8 + 3.2 - 1.6^2 \\
&= 1.44
\end{aligned}
$$

7 Correlation (5 pts)

(a) Let X and Y be independent Bernoulli random variables with $p = \frac{1}{2}$. Show that $X + Y$ and $|X - Y|$ are dependent but uncorrelated.

$$
\begin{aligned}
&\because P(X+Y=0) = \frac{1}{4} \\
&\quad P(|X-Y|=0) = \frac{1}{2} \\
&\quad P\{(|X-Y|=0) \cup (X+Y=0)\} = \frac{1}{2} \\
&\therefore P\{(|X-Y|=0) \cup (X+Y=0)\} \neq P(X+Y=0) + P(|X-Y|=0), dependent
\end{aligned}
$$

$$Corr(X+Y,|X-Y|) = \frac{Cov(X+Y,|X-Y|)}{\sigma_{X+Y}\sigma_{|X-Y|}} = \frac{E[(X+Y)|X-Y|] - E[X+Y]E[|X-Y|]}{\sigma_{X+Y}\sigma_{|X-Y|}}$$

$$When X \geq Y$$

$$\begin{aligned} Corr(X+Y,|X-Y|) &= \frac{E[X^2-Y^2] - (E[X]+E[Y])(E[X]-E[Y])}{\sigma_{X+Y}\sigma_{|X-Y|}} \\ &= \frac{E[X^2]-E[Y^2]-(E[X]^2-E[Y]^2)}{\sigma_{X+Y}\sigma_{|X-Y|}} \\ &= \frac{1/2-1/2-(1/4-1/4)}{\sigma_{X+Y}\sigma_{|X-Y|}} \\ &= 0 \end{aligned}$$

$$Similarly,\ when\ X < Y,\ we\ can\ compute\ Corr(X+Y,|X-Y|) = 0$$

$$\therefore uncorrelated$$

(b) Discuss the differences between Correlation and Covariance. How does $Cov(X,Y) = 1$ differ from $Corr(X,Y) = 1$? Which statement is stronger?

Definitely, Correlation is much stronger. Because $-1 \leq Corr(X,Y) \leq 1$, when it is equal to 1, we can confidently say that Y will increase the same scale with X as X increases. Although covariance is also a measure of how much two random variables change together. Theoretically speaking, it can be arbitrary large. Therefore, $Cov(X,Y) = 1$ definitely means less strong relation than $Corr(X,Y) = 1$.

(c) Show that $Corr(aX+b, cY+d) = -Corr(X,Y)$ when a and c have the opposite sign.
Make sure to include proofs for any identities or shortcuts that you use.
Discuss how this property would change if Covariance were used instead.

$$\begin{aligned} Cov(aX+b,cY+d) &= E\{[aX+b-E[aX+b]][cY+d-E[cY+d]]\} \\ &= E\{[aX+b-aE[X]-b][cY+d-cE[Y]-d]\} \\ &= E\{a(X-E[X])c(Y-E[Y])\} \\ &= acCov(X,Y) \\ Var(aX+b) &= E[(aX+b-E[aX+b])^2] \\ &= E[(aX+b-aE[X]-b])^2] \\ &= E[a^2(X-E[X])^2] \\ &= a^2Var(X) \\ \sigma_{aX+b} &= \sqrt{Var(aX+b)} = |a|\sigma_X \\ Corr(aX+b,cY+d) &= \frac{Cov(aX+b,cY+d)}{\sigma_{aX+b}\sigma_{cX+d}} \\ &= \frac{acCov(X,Y)}{|a||c|\sigma_X\sigma_Y} \\ &= -Corr(X,Y),\ when\ a,c\ have\ the\ opposite\ sign \end{aligned}$$

8 Exploratory Data Analysis (15 pts)

In this section, you will use the R statistical package to begin exploring, transforming, and analyzing data. To get started, do the following:

1. Download and install R from:
   `http://cran.r-project.org/`
   Links to a quick intro to the R programming language and a short reference card are below.
   `http://www.stat.cmu.edu/~larry/all-of-statistics/=R/Rintro.pdf`
   `http://cran.r-project.org/doc/contrib/Short-refcard.pdf`

2. Download the Yelp dataset from the course page.
   This data set is part of the Yelp academic dataset and consists of data about 14,192 restaurants. The datafile *yelp-data.csv* contains 44 attributes: 35 numeric and 9 discrete attributes.

   The first row of the data file is a header row with the names of the attributes, the values are separated by a ";" delimiter. The `categories` attribute is a list of local classification of the restaurants (e.g., *Pizza, Fast Food*).

3. Read the data into R using the `read.table()` function. Print a summary of the data using the `summary()` function. *Make sure to consider the arguments for the `read.table()` function. You will need to use the `comment.char=""` argument to avoid errors.*

R Questions

(a) Plot a histogram of the `tip_count` attribute. Use the `hist()` function with its default values and make sure to title the plot with the name of the attribute for clarity. Next plot a histogram using the log values of tip_count.

```
yelp_dat = read.csv("~/Desktop/yelp.dat.csv", header=TRUE, sep=";")
hist(yelp_dat[, 'tip_count'], main = "Tip Count", xlab = "tip_count")
hist(log(yelp_dat[, 'tip_count']), main = "Log Value of Tip Count",
xlab = "log(tip_count)")
```

(b) Plot the `tip_count` attribute again but this time use the `density()` function in the plot, for both the original and the logged values.

```
plot(density(yelp_dat[, 'tip_count']), main =
"Tip Count with Density", xlab = "density(tip_count)")

plot(density(log(yelp_dat[, 'tip_count'])), main =
"Log Value of Tip Count with Density",
xlab = "density(log(tip_count))")
```

(c) Find the continuous attribute with **largest** range and plot a histogram of the values. Make sure to title the plot with the name of the attribute for clarity.

With the help of summary() function, we can tell that the continuous attribute with **largest** range is **review_count** (from min 3.0 to max 4084.0).

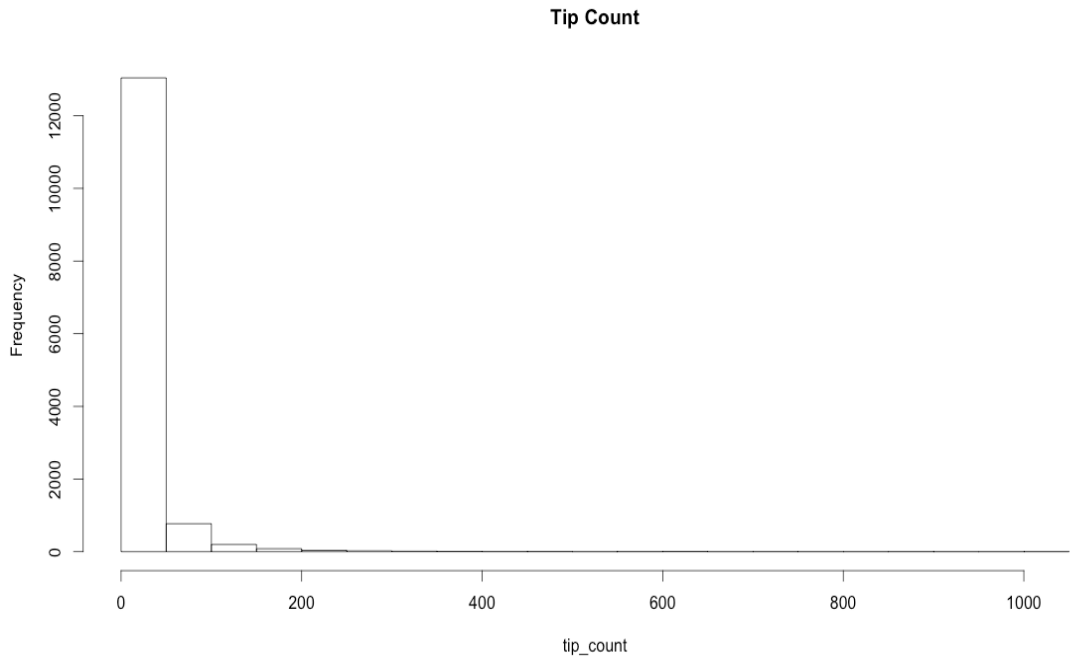

Figure 3: Question(a) the histogram of the `tip_count` attribute

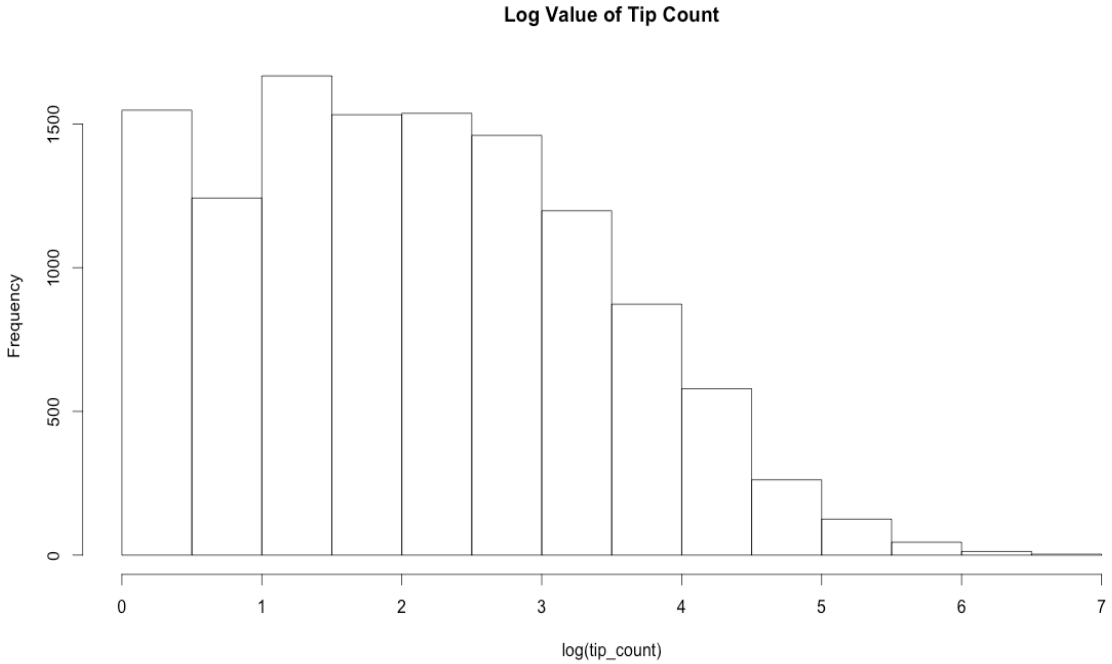

Figure 4: Question(a) the histogram of the log values of tip_count

```
hist(yelp_dat[, 'review_count'], main = "Review Count",
xlab = "review_count")
```

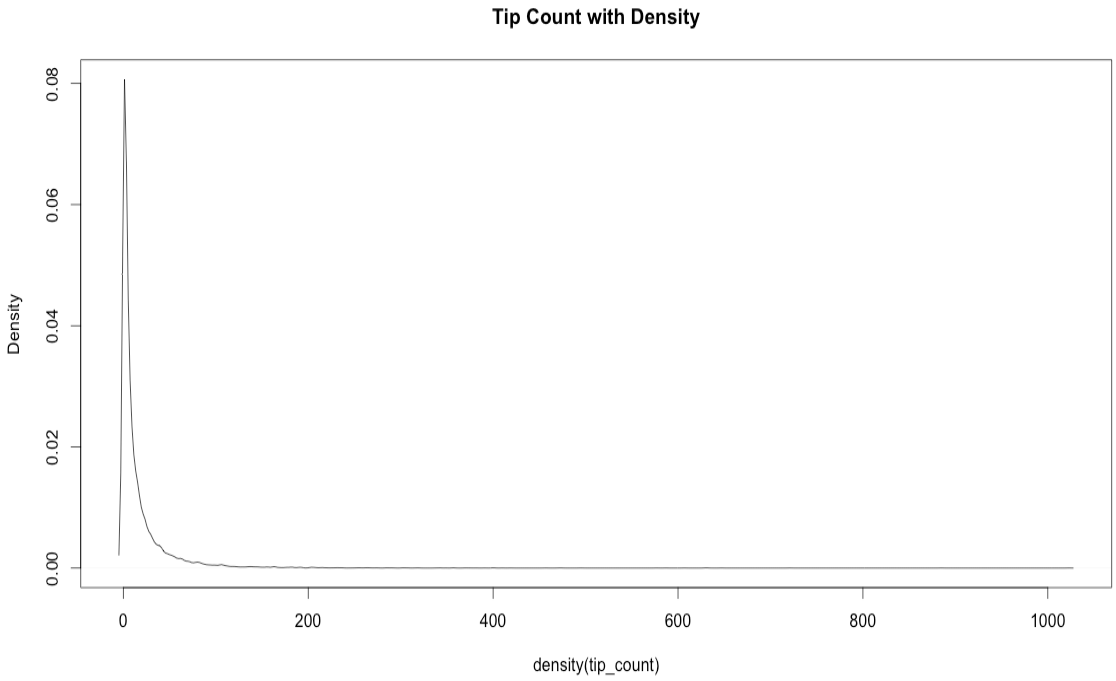

Figure 5: Question(b) plot `tip_count` with density

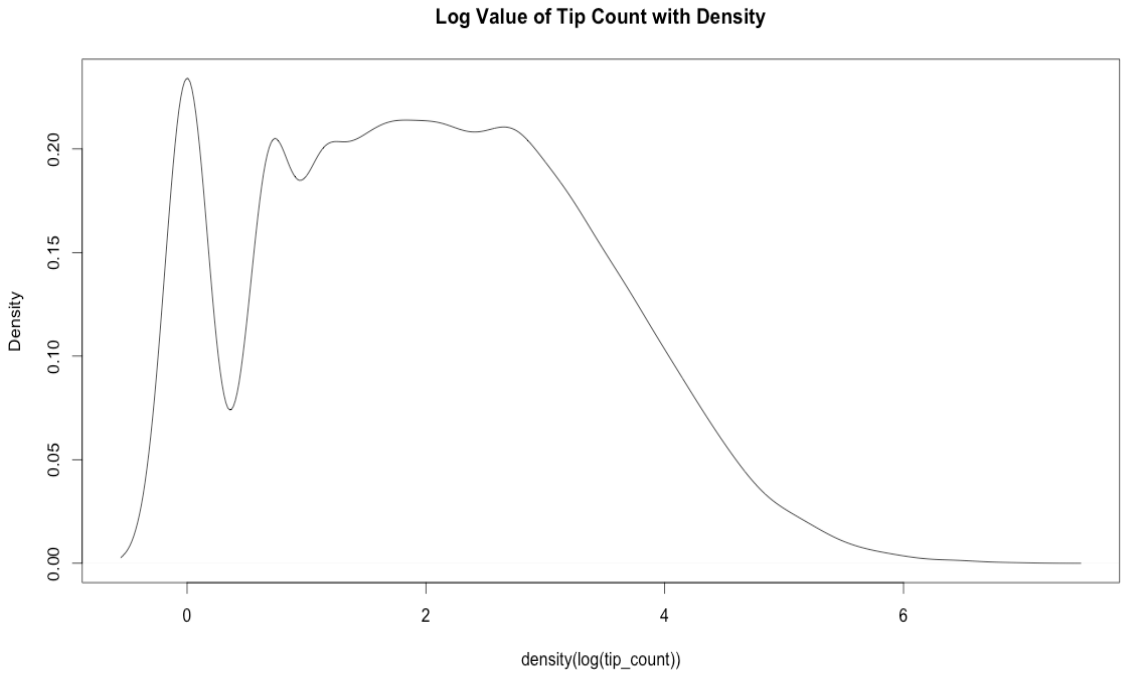

Figure 6: Question(b) plot log value of `tip_count` with density

(d) Find the discrete attribute (that is not a unique identifier) with the **maximum** number of values and plot a barplot to show the frequency of each value. Note that this will look like a histogram but for nominal values. Again, make sure to title the plot with the name of the attribute for clarity.

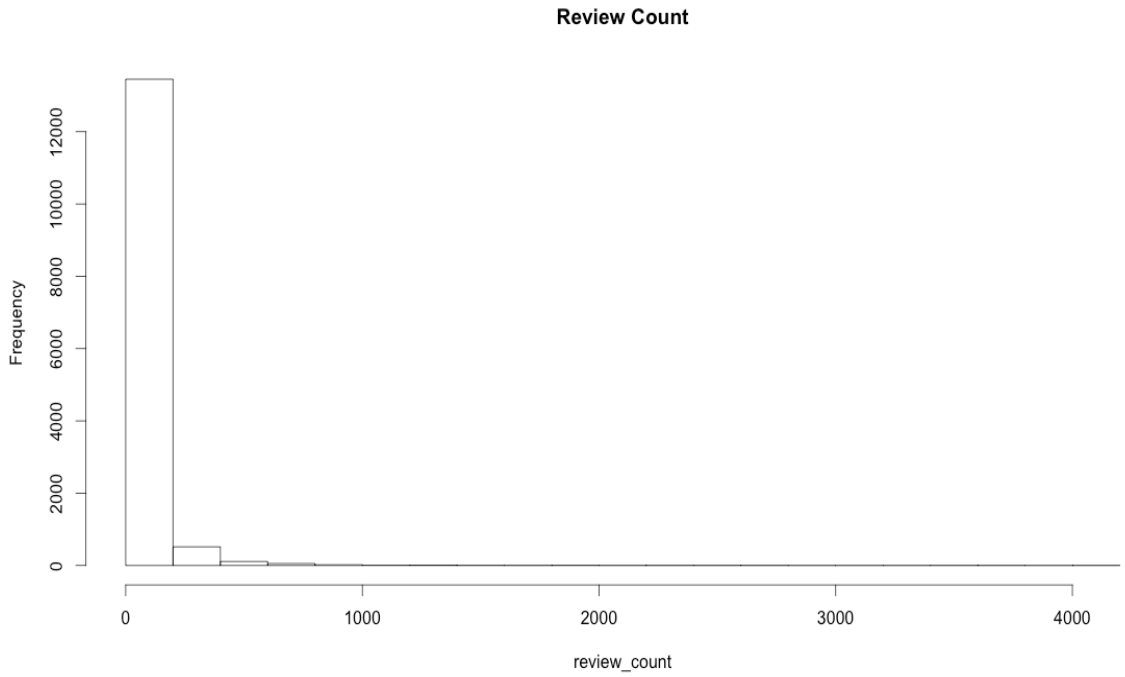

Figure 7: Question(c) the histogram of review count, the continuous attribute with **largest** range

By using the following code:

```
length(unique(yelp_dat[, "attributes"]))
```

apply it to all the discrete attributes. We can find that **attributes** attribute is the one we want.

```
barplot(table(yelp_dat[, 'attributes']), main = "Attributes Barplot")
```

(e) Consider the four continuous attributes: `latitude, longitude, stars, likes`. Calculate the pairwise correlations among these four attributes. Plot scatterplots for the pair of attributes with largest positive correlation and the pair of attributes with largest negative correlation. Make sure to label both axis of the plot with the attribute names. Report the correlations and discuss whether the correlations are interesting or expected, given your domain knowledge.

```
corr_dat = yelp_dat[c('latitude', 'longitude', 'stars', 'likes')]
corrs = cor(corr_dat)
print(corrs)
```

```
plot(yelp_dat[, 'latitude'], yelp_dat[, 'longitude'],
main = "the pair of attributes with largest positive correlation",
xlab = "latitude", ylab = "longitude")

plot(yelp_dat[, 'likes'], yelp_dat[, 'longitude'],
main = "the pair of attributes with largest negative correlation",
xlab = "likes", ylab = "longitude")
```

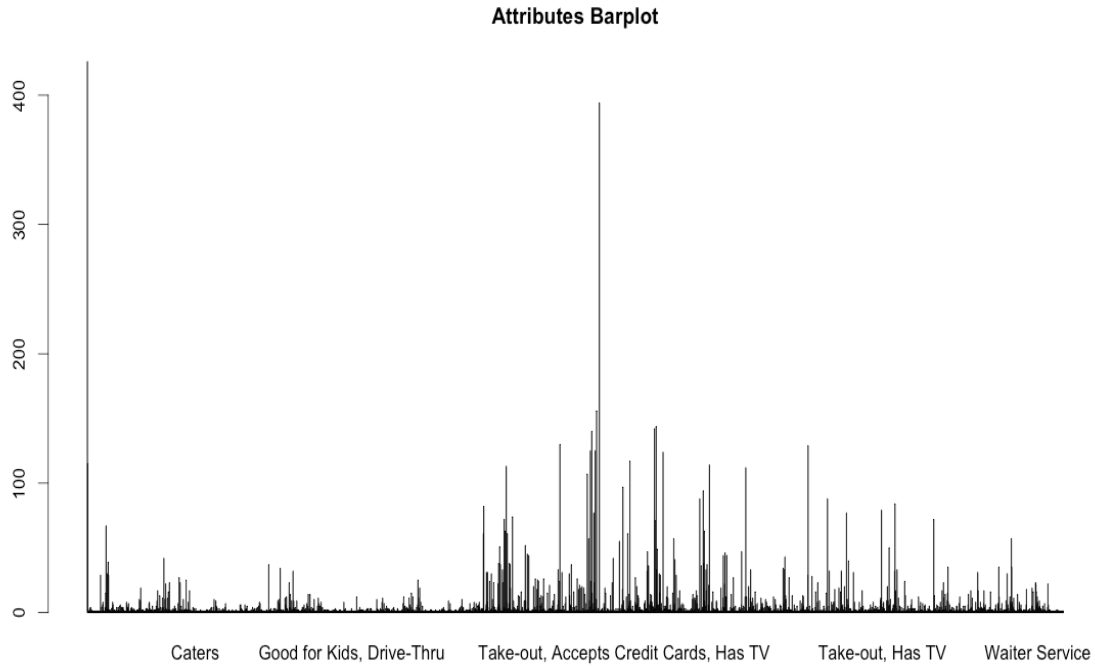

Figure 8: Question(d) the histogram of the discrete attribute with the **maximum** number of values

	latitude	longitude	stars	likes
latitude	1.00000000	0.95550444	0.1306059	-0.04086409
longitude	0.95550444	1.00000000	0.1408710	-0.07586773
stars	0.13060593	0.14087103	1.0000000	0.12153707
likes	-0.04086409	-0.07586773	0.1215371	1.00000000

Table 1: correlations among `latitude, longitude, stars, likes`

The pair of attributes with largest positive correlation are latitude and longitude. Therefore, the scatterplots are as my expected like a map. Also, the points are focusing at the left bottom corner as my expected. Because if we look at the state information, we can see that most of these restaurants are located in Nevada and Arizona.
The pair of attributes with largest negative correlation are likes and longitude. This really surprised me a little bit. I did not expect such relation existed. Since the value is very small, I assume there indeed is little relation between them.

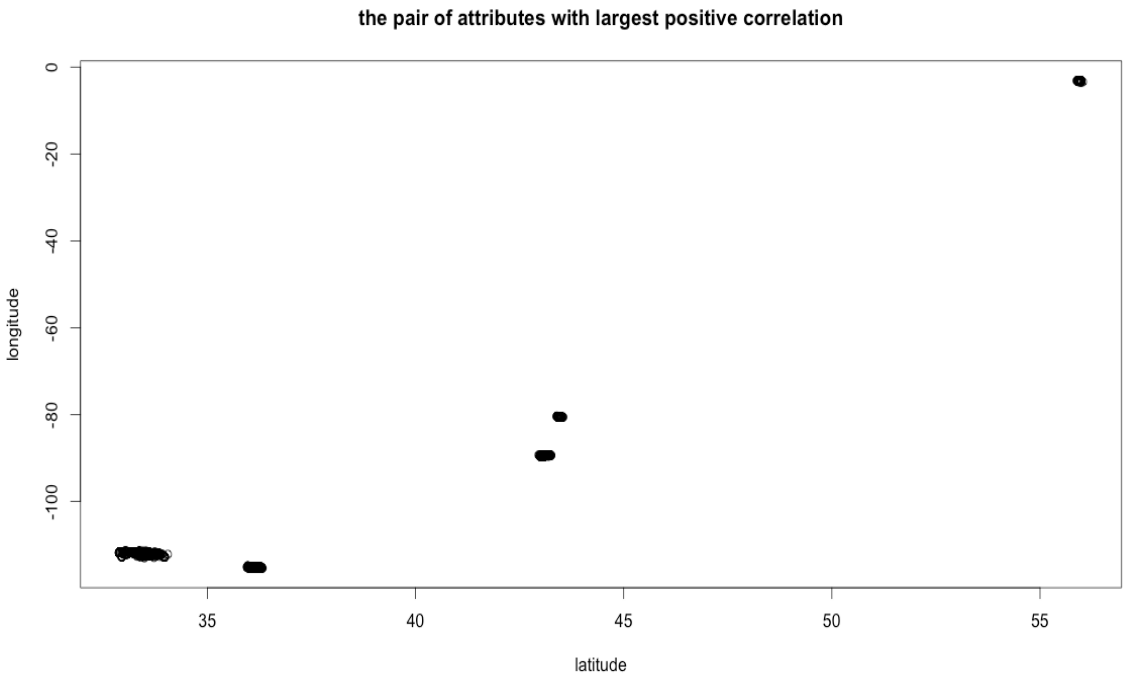

Figure 9: Question(e) scatterplots for the pair of attributes with largest positive correlation

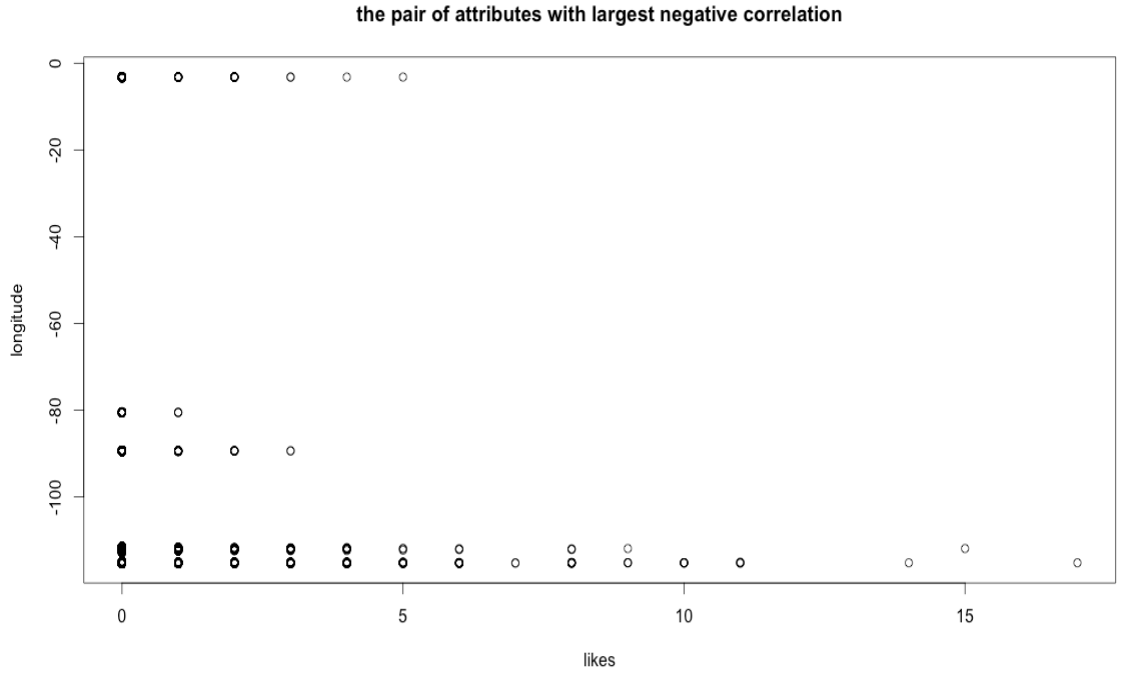

Figure 10: Question(e) scatterplots for the pair of attributes with largest negative correlation

(f) Choose a particular category (e.g., Nightlife) and create a new binary feature for each example that records whether the example contains the chosen category (e.g., Nightlife vs. not-Nightlife). You can use the `regexpr()` function to test whether the list contains a particular string. Plot a boxplots of your new binary feature vs. stars and likes (i.e., *feature vs. stars* and

feature vs. likes). Make sure to label both axes of the plot with the attribute/feature names.

```
night = regexpr("Nightlife", yelp_dat[, 'categories'],
ignore.case = TRUE) != -1

boxplot(yelp_dat[, 'stars']~night, main = "feature vs. stars",
xlab = "Nightlife Category", ylab = "Stars")
boxplot(yelp_dat[, 'likes']~night, main = "feature vs. likes",
xlab = "Nightlife Category", ylab = "Likes")
```

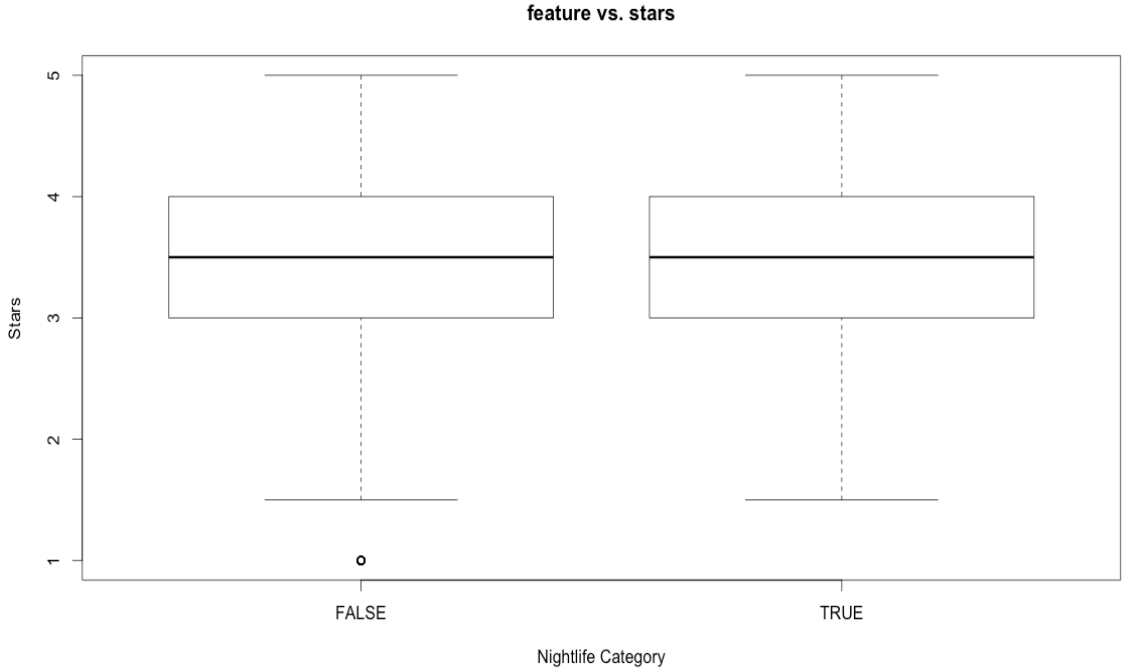

Figure 11: Question(f) boxplots for new feature vs. stars

(g) Continue with the same approach you used above to explore several categories (e.g., *Bars, Diners*). Construct at least two new binary features from those categories that exhibit a difference in the star ratings (between categories). Plot the boxplots and discuss whether the relationship is interesting or expected, given your domain knowledge.

```
bars = regexpr("Bars", yelp_dat[, 'categories'],
ignore.case = TRUE) != -1
diners = regexpr("Diners", yelp_dat[, 'categories'],
ignore.case = TRUE) != -1

boxplot(yelp_dat[, 'stars']~bars, main = "bars vs. stars",
xlab = "Bars Category", ylab = "Stars")
boxplot(yelp_dat[, 'stars']~diners, main = "diners vs. stars",
xlab = "Diners Category", ylab = "Stars")
```

From the above boxplots, we can find several very interesting relations:

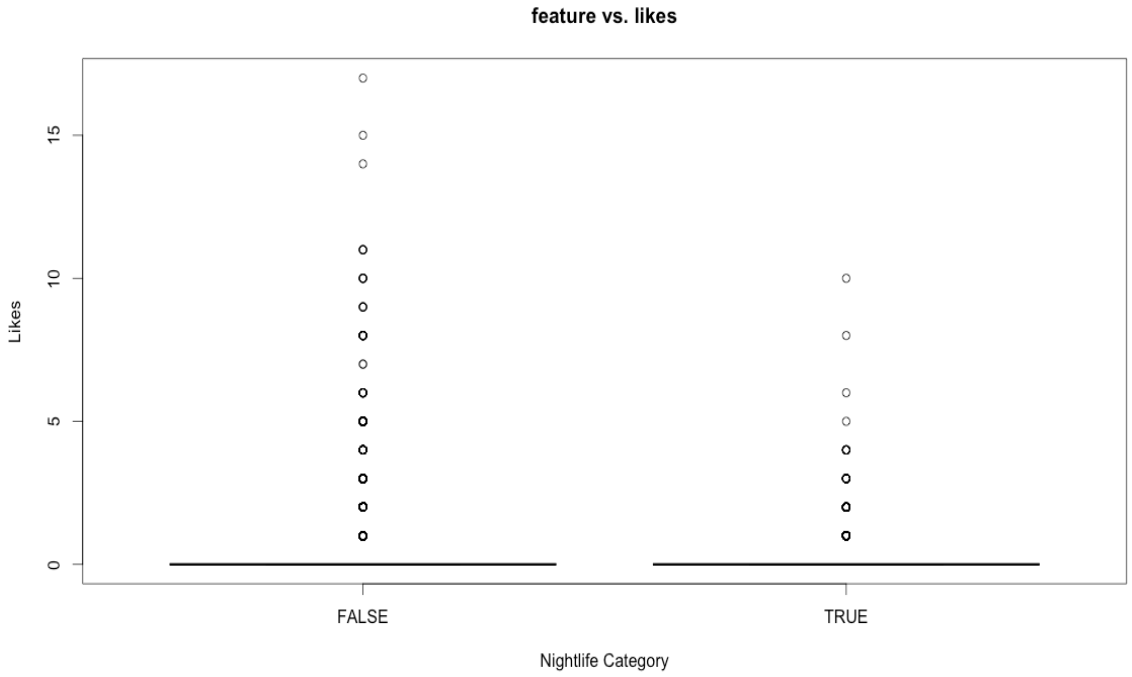

Figure 12: Question(f) boxplots for new feature vs. likes

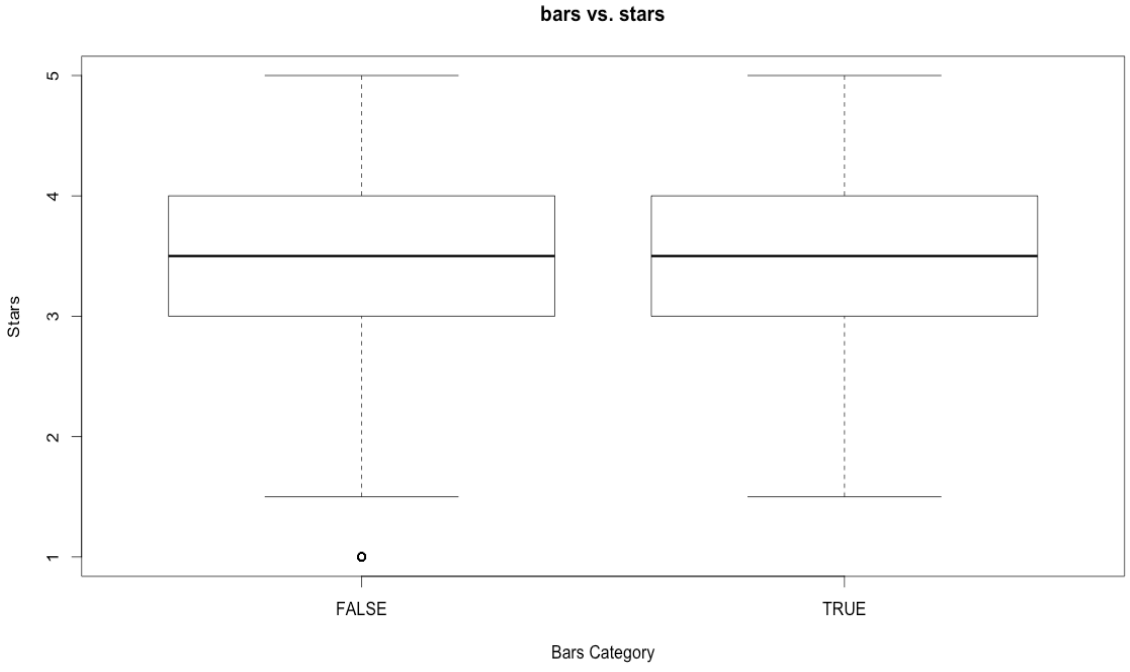

Figure 13: Question(g) boxplots for bars vs. stars

- No matter what categories the restaurants belong to, they all share very similar median, between 3 and 4, which makes sense to me. Most people are willing to give average rating, not too high, not too low unless something really bad or good happened.

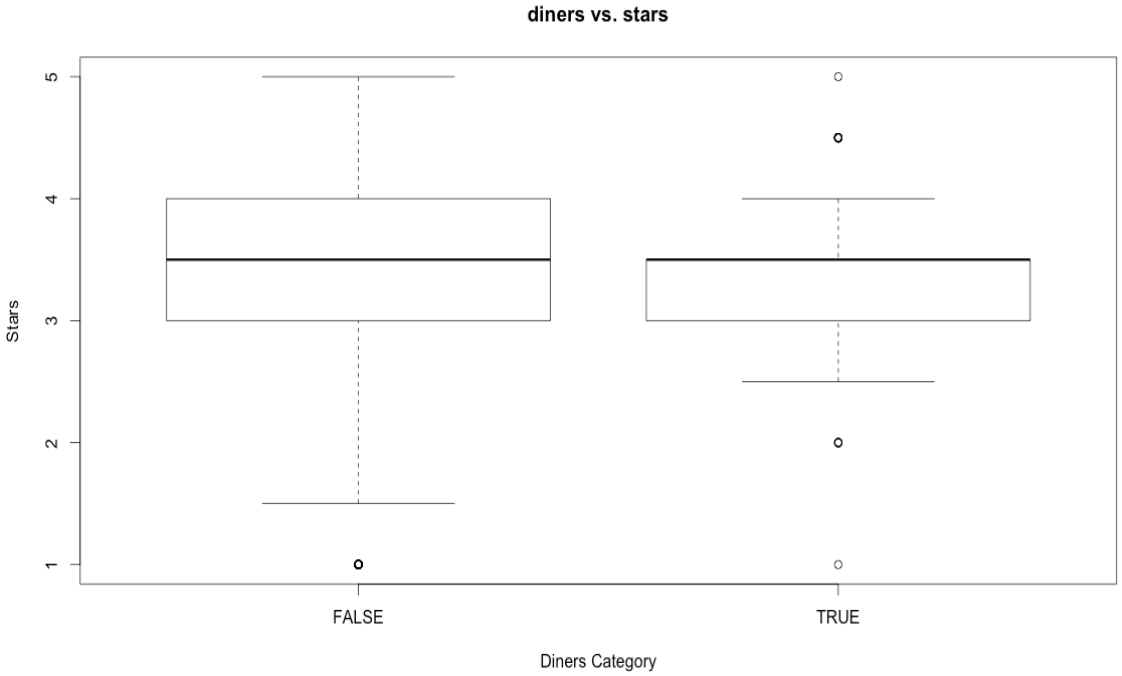

Figure 14: Question(g) boxplots for diners vs. stars

- For those belonging to diner category, we have a very interesting relation. The third quartile and the median are very close, which mean one forth people are give very similar reviews that is close to the median. Comparing this to the bar category, we can find very different relations. From my own understanding, this is because people who go to diner are willing to get great meals and enjoy the nice diner environment with their friends, lovers or family. Therefore, they will have much more strict standards. On the contrary, people who go to bars are more focus on events they will do there. They may go there for a date, a party or just wanna get drunk. They seldom leave serious reviews on Yelp afterwards.