CS57300: Homework 1 Solution

Due date: Sunday February 1, midnight (submit pdf to Blackboard)

Submit both your answers to the questions and the R code that you used for analysis. Your homework must be typed and submitted as a PDF. Use of Latex is recommended, but not required.

1 Counting (2 pts)

(a) (1) $36^6 + 36^7 + 36^8 + 36^9 + 36^{10} \approx 3.76 * 10^{15}$

(2) $92^6 + 92^7 + 92^8 + 92^9 + 92^{10} - 82^6 - 82^7 - 82^8 - 82^9 - 82^{10} \approx 3.0 * 10^{19}$

(b) (1) $92^6 + 92^7 + 92^8 + 92^9 + 92^{10} \approx 4.39 * 10^{19}$

(2) $\sum_{n=6}^{10} 92^n - 82^n - 66^n - 62^n + 10^n + 26^n + 30^n$

2 Axioms of probability (2 pts)

(a) $P(A \cup B) = P(A) + P(B \setminus (A \cap B)) \Rightarrow P(B \setminus (A \cap B)) = P(A \cup B) - P(A)$
$P(B) = P(B \setminus (A \cap B)) + P(A \cap B) \Rightarrow P(B \setminus (A \cap B)) = P(B) - P(A \cap B)$
Then we can get:
$P(A \cup B) - P(A) = P(B) - P(A \cap B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

3 Probability and conditional probability (3 pts)

(a) (i) $P($ at least once$) = P($fail at first time $) = 2/8 = 1/4$

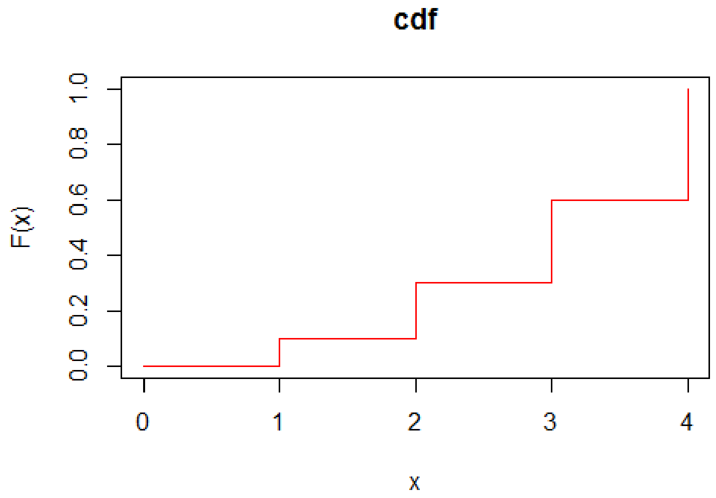
(ii) $P($ at most twice$) = P($fail at first time $) * P($ success at second time $) = \frac{2}{8} * \frac{6}{8} = \frac{3}{16}$

(b) There are 15 case that Alice win the game, if Alice want to win with number, the other number will only by 1,2,3,4, so the probability Alice won, and she rolled a 5 is 4/15.

4 Probability distributions (3 pts)

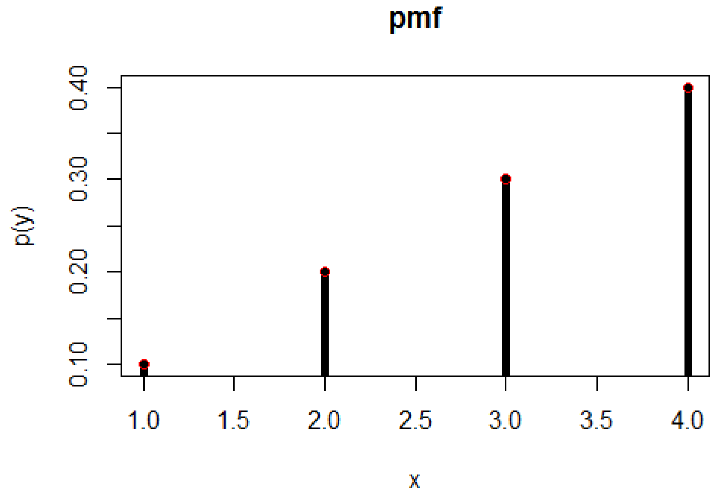
(a)

$F(x)$	1	2	3	4
	0.1	0.3	0.6	1



(b)

$p(x)$	1	2	3	4
	0.1	0.2	0.3	0.4



(c) $E(X) = 1 * 0.1 + 2 * 0.2 + 3 * 0.3 + 4 * 0.4 = 3$
$Var(X) = (1-3)^2 * 0.1 + (2-3)^2 * 0.2 + (3-3)^2 * 0.3 + (4-3)^2 * 0.4 = 1$

5 Independence (3 pts)

(a) No. Because disk 1 is red and green.

(b) No. Because disk 4 is green and white

(c) Green and black, white and black

6 Conditional Expectation (2 pts)

According p(x,y)=48/(45xy),we get $P(x=2,y=4)=2/15$,$P(x=2,y=1)=8/15$, $P(x=4,y=1)=4/15$, $P(x=4,y=4)=1/15$.
According total probability $P(x=2)=P(x=2,y=4)+P(x=2,y=1)=10/15=2/3$, $P(x=4)=P(x=4,y=1)+P(x=4,y=4)=5/15=1/3$
so $P(y=1|x=2)=P(x=2,y=1)/P(x=2)=4/5$
$P(y=4|x=2)=P(x=2,y=4)/P(x=2)=1/5$
$P(y=1|x=4)=P(x=4,y=1)/P(x=4)=4/5$
$P(y=4|x=4)=P(x=4,y=4)/P(x=4)=1/5$
$E(Y|x=2)=1*P(y=1|x=2)+4*P(y=4|x=2)=4/5+4*1/5=8/5$
$E(Y|x=4)=1*P(y=1|x=4)+4*P(y=4|x=4)=4/5+4*1/5=8/5$
$E(Y^2|x=2)=1*P(y=1|x=2)+4^2*P(y=4|x=2)=4/5+16*1/5=4$
$E(Y^2|x=4)=1*P(y=1|x=4)+4^2*P(y=4|x=4)=4/5+16*1/5=4$
$Var(Y|x=2)=4-(8/5)^2=1.44$
$Var(Y|x=4)=4-(8/5)^2=1.44$

7 Correlation (5 pts)

(a)

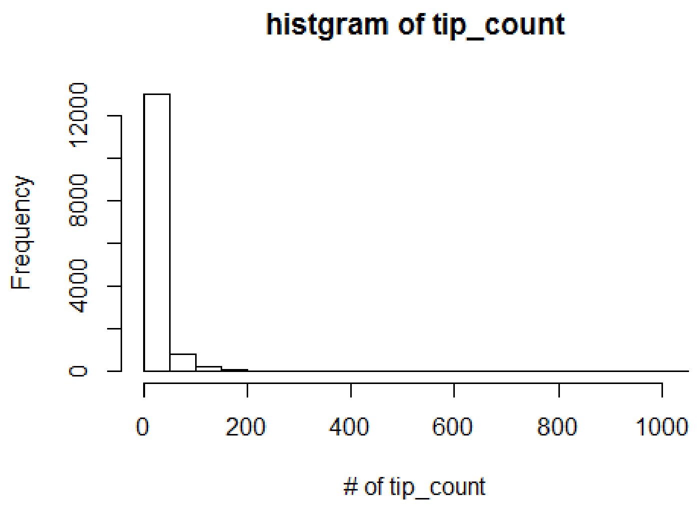
(b) Both covariance and correlation are described how two variables related, but correlation standardizes covariance by dividing through standard deviation. $Corr(X,Y)=1$ represent two variables positive related perfectly, but $Cov(X,Y)=1$ represent two variables are positive related but not so perfectly. So $corr(X,Y)=1$ will be more stronger.

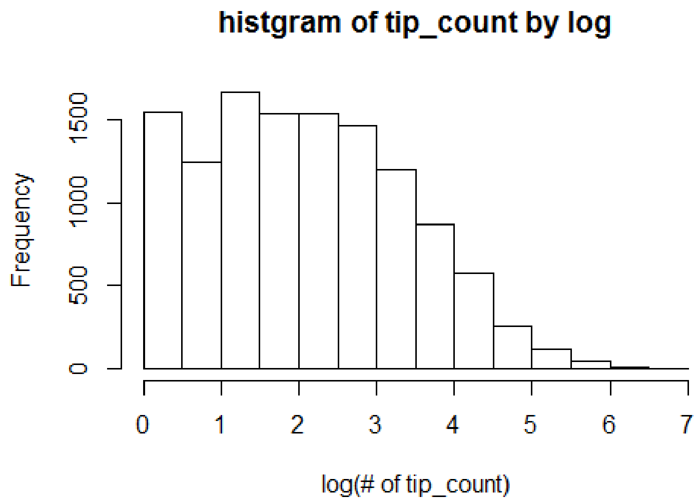
(c) Based on Proposition: $Cov(X,Y)=E(XY)-E(X)E(Y)$
$Cov(aX+b,cY+d)=E((aX+b)(cY+d))-E(aX+b)E(cY+d)=E(acXY+bcY+adX+bd)-E(aX+b)E(cY+d)=acE(XY)+bcE(Y)+adE(X)+E(bd)-(aE(X)+b)(cE(Y)+d)=acE(XY)+bcE(Y)+adE(X)+E(bd)-(acE(X)E(Y)+adE(X)+bcE(Y)+E(bd))=acE(XY)-acE(X)E(Y)=acCov(X,Y)$
$Corr(aX+b,cY+d)=\frac{Cov(aX+b,cY+d)}{\sqrt{Var(aX+b)}*\sqrt{Var(cY+d)}}=\frac{acCov(X,Y)}{|a||c|\sqrt{Var(X)Var(Y)}}$
Therefore if a,and c are opposite, $Corr(aX+b,cY+d)=-Corr(X,Y)$ and $Cov(aX+b,cY+d)=acCov(X,Y)$

8 Exploratory Data Analysis (15 pts)

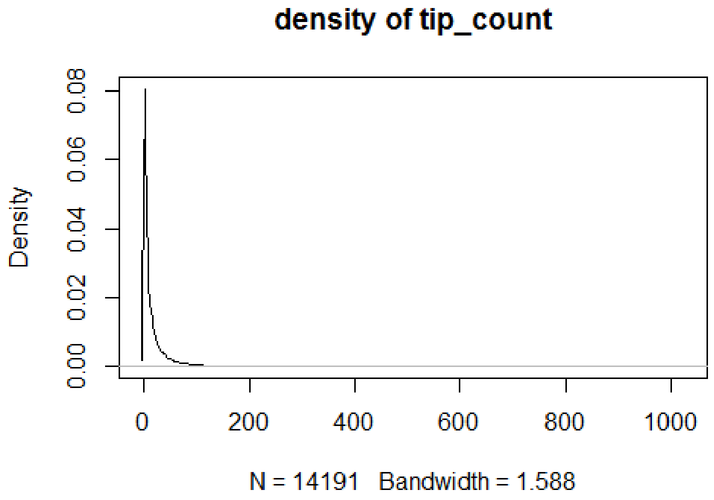
R Questions

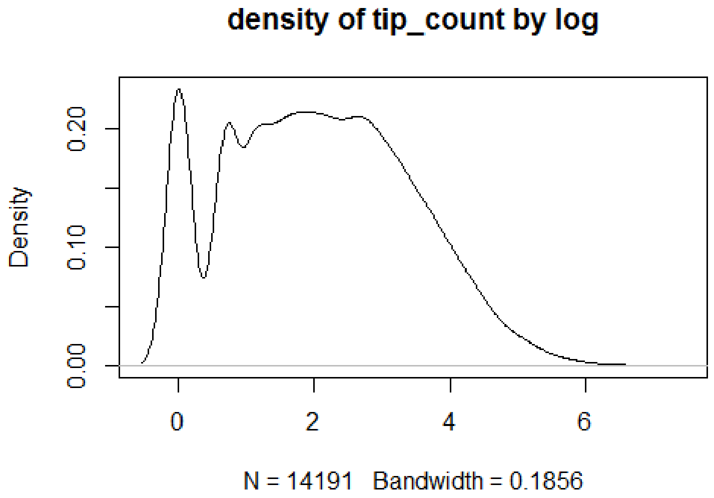
(a) Plot a histogram of the `tip_count` attribute. Use the `hist()` function with its default values and make sure to title the plot with the name of the attribute for clarity. Next plot a histogram using the log values of tip_count.



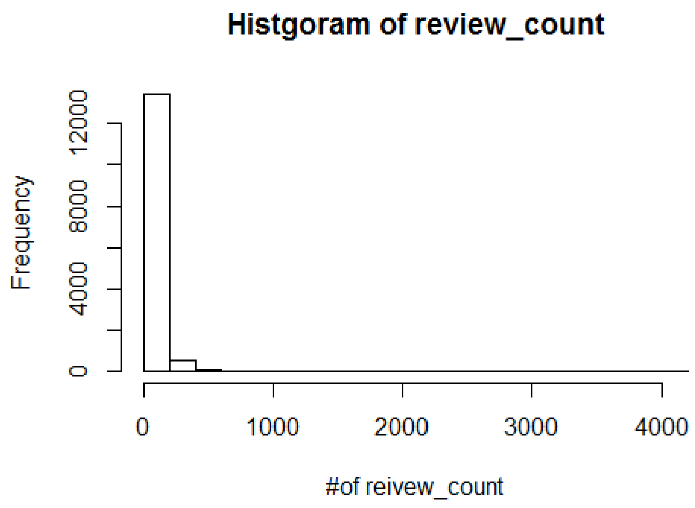


(b) Plot the `tip_count` attribute again but this time use the `density()` function in the plot, for both the original and the logged values.

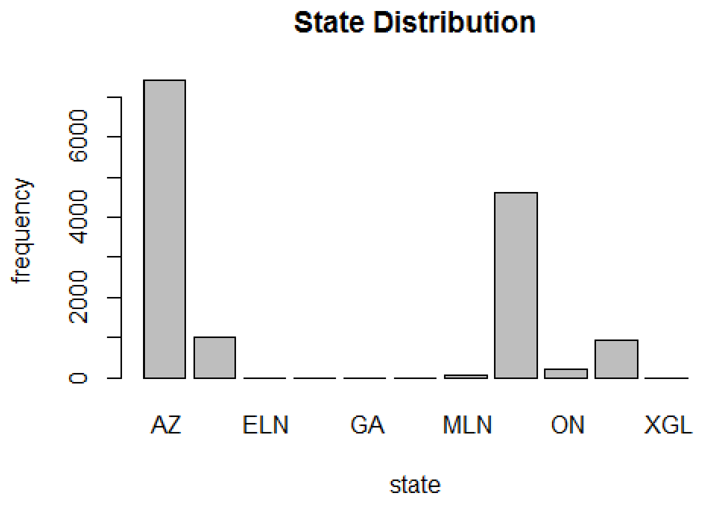




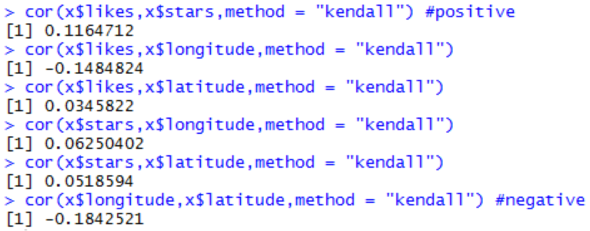
(c) Find the continuous attribute with **largest** range and plot a histogram of the values. Make sure to title the plot with the name of the attribute for clarity.



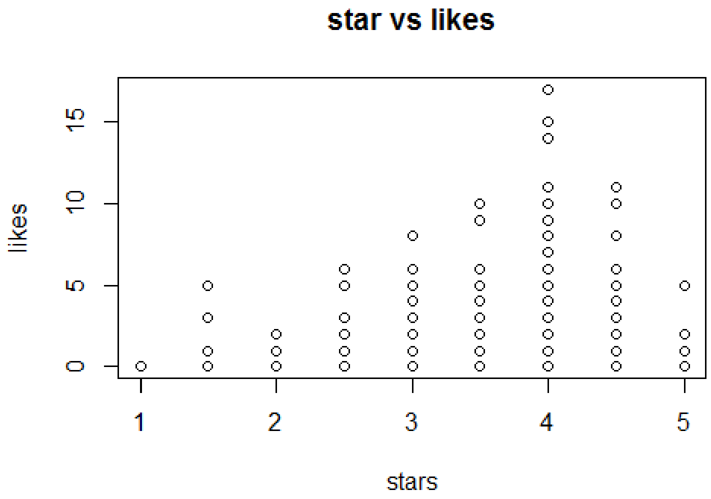
(d) Find the discrete attribute (that is not a unique identifier) with the **maximum** number of values and plot a barplot to show the frequency of each value. Note that this will look like a histogram but for nominal values. Again, make sure to title the plot with the name of the attribute for clarity.



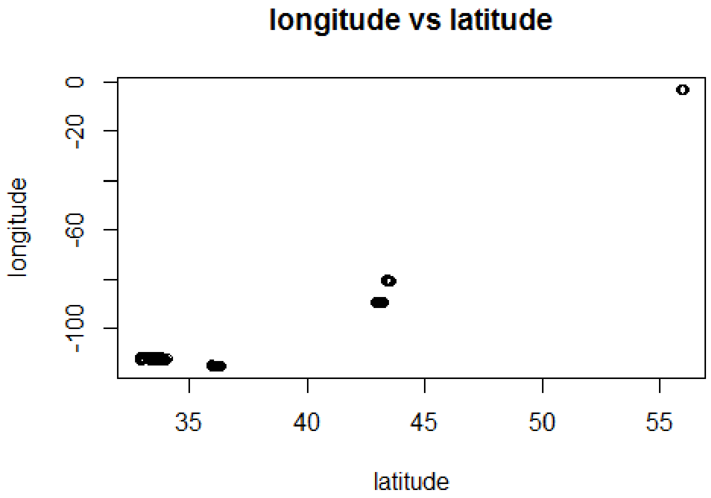
(e) Consider the four continuous attributes: `latitude, longitude, stars, likes`. Calculate the pairwise correlations among these four attributes. Plot scatterplots for the pair of attributes with largest positive correlation and the pair of attributes with largest negative correlation. Make sure to 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.



```
> cor(x$likes,x$stars,method = "kendall") #positive
[1] 0.1164712
> cor(x$likes,x$longitude,method = "kendall")
[1] -0.1484824
> cor(x$likes,x$latitude,method = "kendall")
[1] 0.0345822
> cor(x$stars,x$longitude,method = "kendall")
[1] 0.06250402
> cor(x$stars,x$latitude,method = "kendall")
[1] 0.0518594
> cor(x$longitude,x$latitude,method = "kendall") #negative
[1] -0.1842521
```

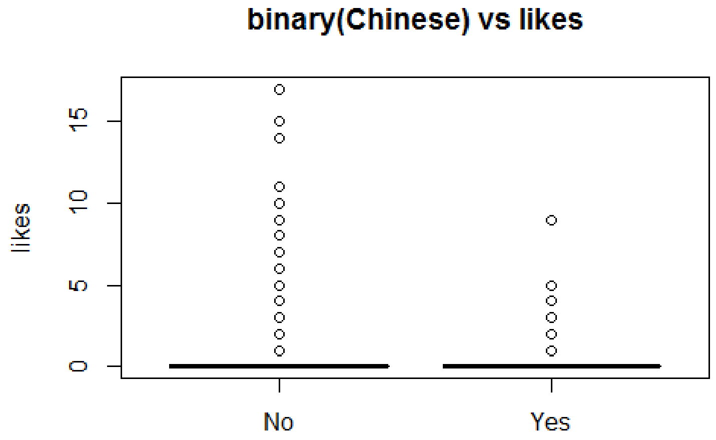


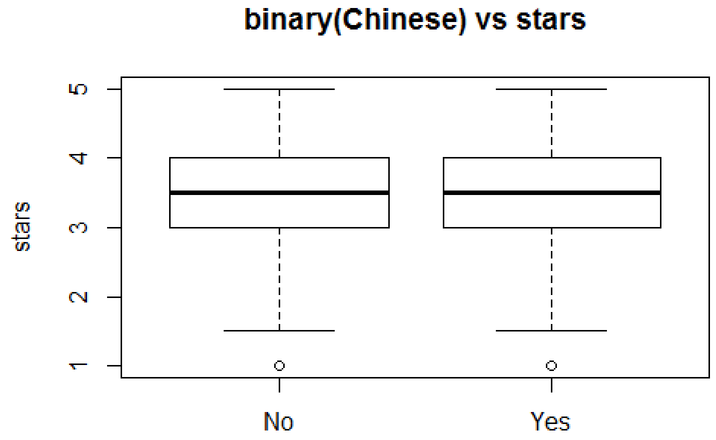
vs likes.png



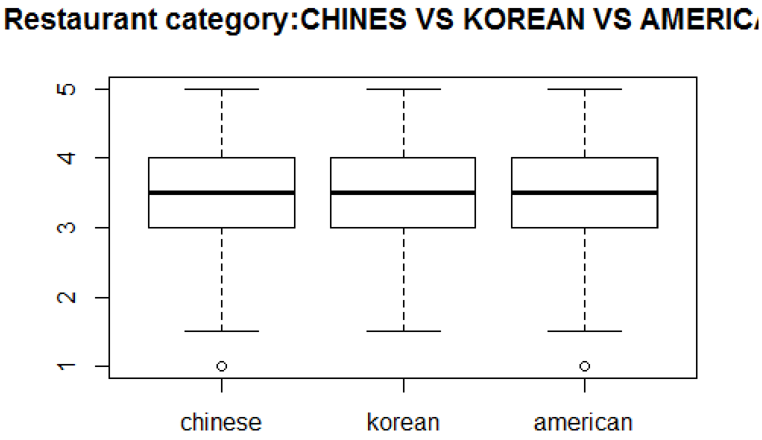
vs latitude.png
stars and likes has positive correlation, it is because if people like the restaurant, they will rate higher stars for that restaurant.

(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.





(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.



According the boxplot, we can conclude that the korean restaurant, american restaurant,and chinese restaurant are all relative good, because they have similar score. That's can proved by those type of restaurants are very easy to find in our life.