CS57300: Homework 1

Hasini, Urala Liyanage Dona Gunasinghe

1 Counting

(a) number of all the passwords that contain lowercase and digits = $\sum_{i=6}^{10} 36^i$
number of passwords containing at least one digit = (number of all passwords containing lowercase and digits - number of passwords containing only lowercase letters)
therefore, number of passwords containing at least one digit = $\sum_{i=6}^{10} 36^i$ - $\sum_{i=6}^{10} 26^i$

(b) number of all passwords containing uppercase, lowercase, digits and special characters = $\sum_{i=6}^{10} 92^i$
number of passwords containing at least 1 digit, 1 uppercase and 1 special character = (number of all passwords containing all four types - number of passwords without any digit - number of passwords without any uppercase - number of passwords without any special character)
therefore, number of passwords containing at least 1 digit, 1 uppercase and 1 special character = ($\sum_{i=6}^{10} 92^i$) - ($\sum_{i=6}^{10} 82^i$) - ($\sum_{i=6}^{10} 66^i$) - ($\sum_{i=6}^{10} 62^i$)

2 Axioms of probability

(a) Axioms of probability:
For a sample space S with possible events:
1. For every event A: $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. $P(A_1 \cup A_2 ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$, if $A_1, A_2, ... A_n$ are pairwise mutually exclusive events.

By definition, $(A \setminus B)$, $(B \setminus A)$ and $A \cap B$ are mutually exclusive events.
From axiom 3:

$$P[(A \setminus B) \cup (B \setminus A) \cup (A \cap B)] = P(A \setminus B) + P(B \setminus A) + P(A \cap B) \quad (1)$$

By definition:

$$P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B) \tag{2}$$

From axiom 3:

$$P(A \setminus B) = P(A) - P(A \cap B) \tag{3}$$

From axiom 3:

$$P(B \setminus A) = P(B) - P(A \cap B) \tag{4}$$

From equations 1, 2, 3 and 4:

$$P(A \cup B) = P(A) + P(B) - P(A \cup B) \tag{5}$$

(b) By definition:

$$P(B \mid A, C) = \frac{P[B \cap (A \cap C)]}{P(A \cap C)} = \frac{P(A, B, C)}{P(A, C)} \tag{6}$$

like wise:

$$P(A|B, C) = \frac{P[A \cap (B \cap C)]}{P(B \cap C)} = \frac{P(A, B, C)}{P(B, C)} \Rightarrow P(A, B, C) = P(A \mid B, C).P(B, C) \tag{7}$$

Substituting 7 into 6:

$$P(B \mid A, C) = \frac{P(A \mid B, C).P(B, C)}{P(A, C)} \tag{8}$$

Again by definition:

$$P(B \mid C) = \frac{P(B \cap C)}{P(C)} = \frac{P(B, C)}{P(C)} \Rightarrow P(B, C) = P(B \mid C).P(C) \tag{9}$$

and:

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} = \frac{P(A, C)}{P(C)} \Rightarrow P(A, C) = P(A \mid C).P(C) \tag{10}$$

By substituting 9 and 10 into 8:

$$P(B \mid A, C) = \frac{P(A \mid B, C).P(B \mid C)}{P(A \mid C} \tag{11}$$

3 Probability and conditional probability

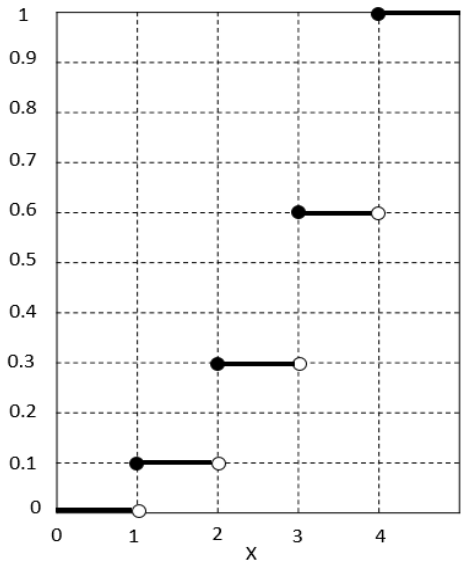
(a) (i) They will need to do it more than once iff they get all three matched in the first time. Let A be the event that they get all three-matched.
Therefore, the probability that they get all three matched in the first time P(A)= 0.25.

(ii) Let B be the event that they get a non-match. P(B) = 0.75 Therefore, the probability that they do it at most twice = $P(A \cap B)$
$= P(B \mid A).P(A) = P(B).P(A) = 0.25 * 0.75 = 0.1875$

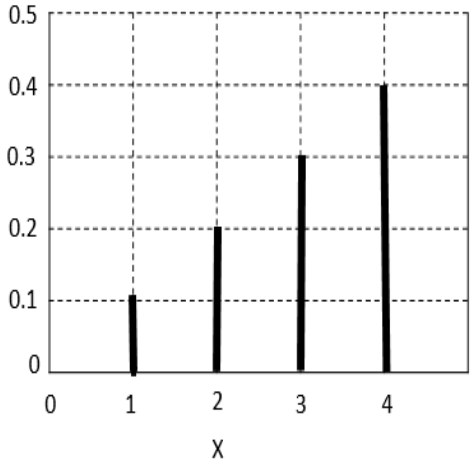
(b) Let the event that Alice wins = A
P(A) = 15/36
Let the event that Alice rolls 5 = B
P(B) = 1/6 (assuming that it is a six sided die)
$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)}{P(A)} = \frac{(4/6).(1/6)}{15/36} = \frac{4}{15}$

4 Probability Distributions

(a) F(0) = 0, F(1) = 0.1, F(2) = 0.3, F(3) = 0.6, F(4) = 1
Graph of the CDF:



(b) P(X=1) = F(1) - F(0) = 0.1
P(X=2) = F(2) - F(1) = 0.2
P(X=3) = F(3) - F(2) = 0.3
P(X=4) = F(4) - F(3) = 0.4
Graph of the PMF:



(c) $\mathrm{E(X)} = \sum_{x=0}^{4} x.P(x) = 3$

$\mathrm{E}(X^2) = \sum_{x=0}^{4} x^2.P(x) = 10$

$\mathrm{Var(X)} = \mathrm{E}(X^2)$ - $[E(X)]^2 = 1$

5 Independence

(a) A and B are not independent.
$P(A) = 3/8$
$P(A \mid B) = 0.5 \neq P(A)$

(b) B and C are not independent.
$P(B) = 2/8$
$P(B \mid C) = 0.5 \neq P(B)$

(c) The events B and D are mutually exclusive.
The events C and D are mutually exclusive.

6 Conditional Expectation

$$P(x,y) = \begin{cases} \frac{48}{45xy} & \text{if x=2,4 and y=1,4} \\ 0 & \text{otherwise} \end{cases}$$

possible x,y pairs = (2,1), (2,4), (4,1) and (4,4)

Table 1: Joint probability distribution table

x	y	P(x,y)
2	1	24/45 = 8/15
2	4	6/45 = 2/15
4	1	12/45 = 4/15
4	4	3/45 = 1/15

From Table 1:
P(X=2) = 10/15 = 0.6667
P(X=4) = 5/15 = 0.333

$P(Y \mid X) = \frac{P(X,Y)}{P(X)}$
$P(y = 1 \mid x = 2) = 0.8$
$P(y = 4 \mid x = 2) = 0.2$
$P(y = 1 \mid x = 4) = 0.8$
$P(y = 4 \mid x = 4) = 0.2$

$$E(Y \mid x = 2) = \sum_{y=1,4} y.P(Y \mid x = 2) = 1.6$$
$$E(Y^2 \mid x = 2) = \sum_{y=1,4} y^2.P(Y \mid x = 2) = 4$$
$$Var(Y \mid x = 2) = E(Y^2 \mid x = 2) - [E(Y \mid x = 2)]^2 = 4 - (1.6)^2 = 1.44$$

$$E(Y \mid x = 4) = \sum_{y=1,4} y.P(Y \mid x = 4) = 1.6$$
$$E(Y^2 \mid x = 4) = \sum_{y=1,4} y^2.P(Y \mid x = 4) = 4$$
$$Var(Y \mid x = 4) = E(Y^2 \mid x = 4) - [E(Y \mid x = 4)]^2 = 4 - (1.6)^2 = 1.44$$

7 Correlation

(a) Since X and Y are Bernoulli random variables the possible values X and Y can take are: X = 0,1 and Y = 0,1.
Let A = X + Y and B = $\mid X - Y \mid$

Table 2: Distribution of A and B

X	Y	A=X+Y	B=$\mid X-Y \mid$
0	0	0	0
0	1	1	1
1	0	1	1
1	1	2	0

From Table 2:
P(A=0) = 1/4, P(B=0) = 2/4
$P(A = 0 \mid B = 0) = (1/4)/(2/4) = 1/2 \neq P(A = 0)$
Therefore, A and B are dependent.

$\overline{A} = 1$, $\overline{B} = 0.5$
Cov(A,B) = $\frac{1}{n} \sum_{i=1}^{4}(a_i - \overline{A})(b_i - \overline{B}) = 0 \Rightarrow$ A and B are uncorrelated.

(b) Covariance depends on the ranges of X and Y whereas Correlation standardizes the covariance.
Covariance has dimension which is the product of the units of X and Y, whereas Correlation is dimensionless.
Therefore covariance of two variables is hard to compare with the covariance of two other variables with different measurements and scales, and conclude which two variables covariate better.
On the other hand, since the correlation normalizes the covariance (and hence is assured to be between +1 and -1), correlation of two pairs of variables can be compared and understand whether two variables covariate more than the other two variables.
Therefore, the statement Corr (X,Y) = 1 is stronger than the statement Cov (X,Y)=1.

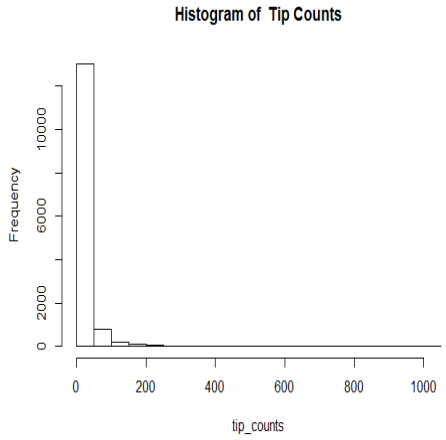
(c) By definition: Corr(X,Y) = $\frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = \frac{E((X-\overline{X})(Y-\overline{Y}))}{\sigma(X)\sigma(Y)} = \frac{E(X,Y)-E(X).E(Y)}{\sigma(X)\sigma(Y)}$
LHS of the given equation to prove = Corr(aX + b, cY + d) = $\frac{Cov(aX+b,cY+d)}{\sqrt{Var(aX+b).\sqrt{Var(cY+d)}}}$
=$\frac{E[(aX+b)(cY+d)]-E(aX+b).E(cY+d)}{a\sqrt{Var(X)}.c\sqrt{Var(Y)}}$
=$\frac{E[acXY+adX+bcY+bd]-[aE(X)+b][cE(Y)+d]}{ac\sigma(X)\sigma(Y)}$
=-$\frac{E(X.Y)-E(X).E(Y)}{\sigma(X)\sigma(Y)}$ (since A and C have opposite signs,)
=-Corr(X,Y) = RHS.

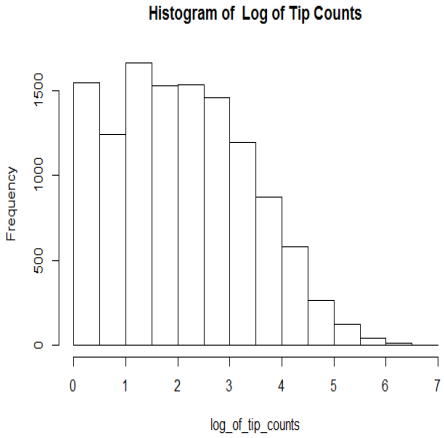
8 Exploratory Data Analysis

(a)
Code for reading the table and summary:

```
#**********print summary******************
w = read.table("C:/HASINI/Purdue/LectureNotes/DataMining/HW1/yelp.dat.txt",header=TRUE,sep=";",quote="\"",comment.char="")
print(summary(w))
```

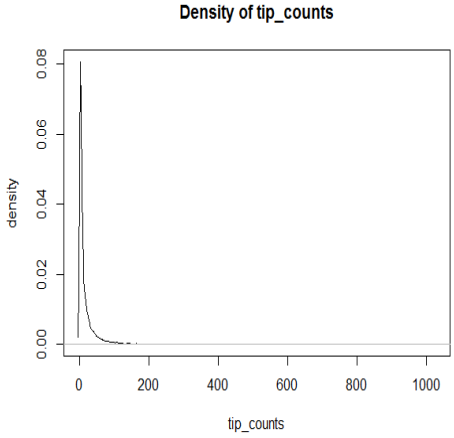
```
#**********print histogram*****************
tip_counts=w["tip_count"]
hist(tip_counts,main="Histogram of Tip Counts")

log_of_tip_counts = log(tip_counts)
hist(log_of_tip_counts, main="Histogram of Log of Tip Counts",xlab="log(tip_counts)")
```

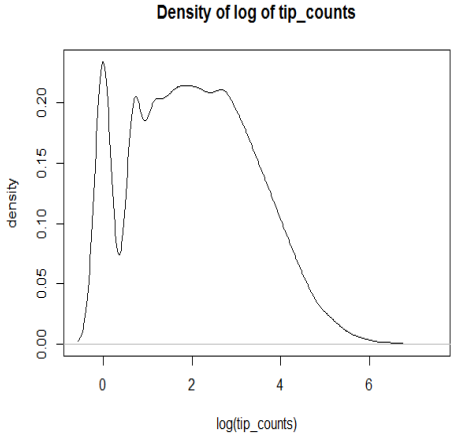



Histogram of Log of Tip Counts
Frequency
0
500
1000
1500
0
1
2
3
4
5
6
7
log_of_tip_counts

(b)

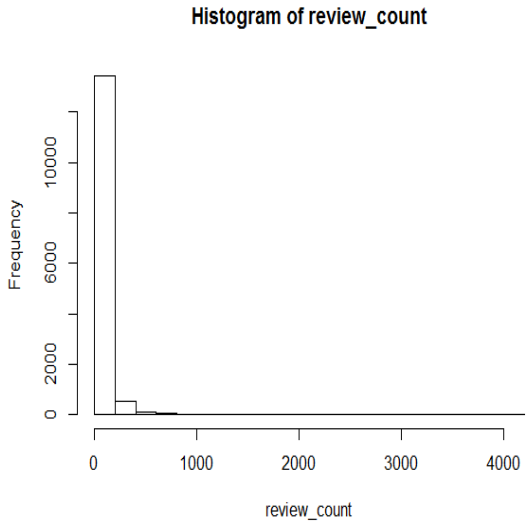
```
#*******print density plots*****************
d = density(tip_counts)
plot(d,main="Density of tip_counts",xlab="tip_counts", ylab="density")

f = density(log_of_tip_counts)
plot(f,main="Density of log of tip_counts", xlab="log(tip_counts)", ylab="density")
```





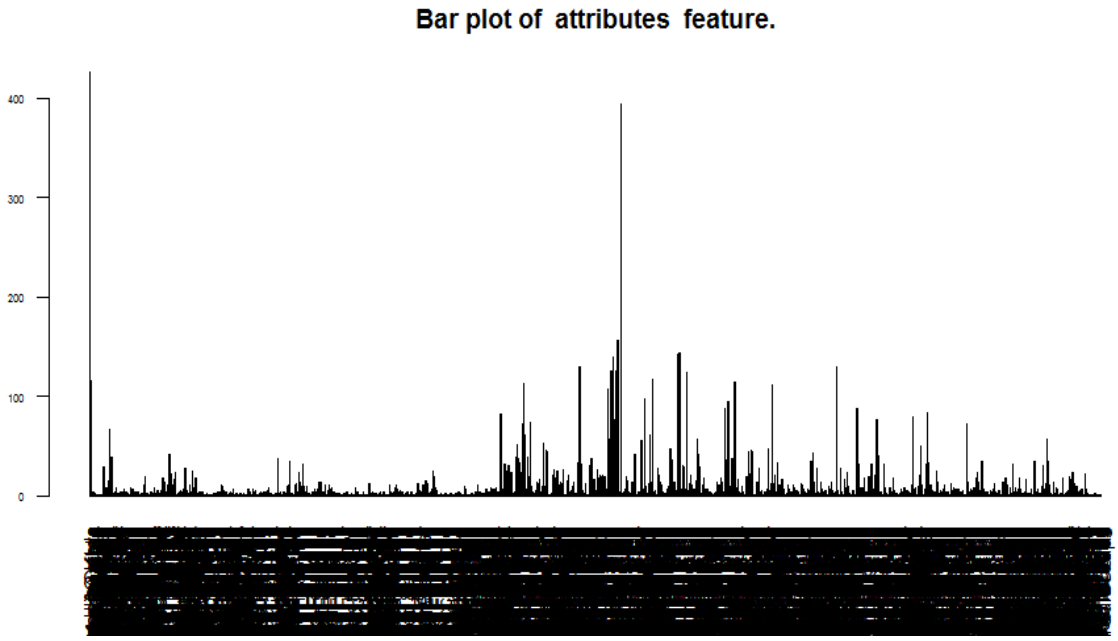
(c)

```
########check the structure of the data frame
str(w)

###########plot histogram of continuous attribute with max range#################
k <- sapply(w,is.factor)
l=which(!k)

cont_attr = w[,l]
cols = colnames(cont_attr)
max_range = 0
max_range_attr = ""
for(col in cols){
  cont_attr_vector = cont_attr[,col]
  r = range(cont_attr_vector)
  d = diff(r)
  if(d>max_range){
    max_range=d
    max_range_attr = col
  }
}
hist(cont_attr[,max_range_attr],main=paste("Histogram of",max_range_attr),xlab=max_range_attr)
```



(d)

```
##########bar plot of the descrete attribute with max num of values##############
m=which(k)
#print(k)

discrete_attr = w[,m]
cols = colnames(discrete_attr)
max_levels = 0
max_levels_attr=""
id_cols = c("business_id","name","address")
for(col in cols){
  if(col %in% id_cols){
    next
  }
  f = factor(w[,col])
  l = nlevels(f)
  print(col)
  print(l)
  if(l>max_levels){
    max_levels = l
    max_levels_attr = col
  }
}
print(max_levels_attr)
barplot(table(w[,max_levels_attr]),width=10,space=1/10,beside=TRUE,main=paste("Bar plot of ",max_levels_attr," feature."),cex.axis = 0.5,las=2)
```



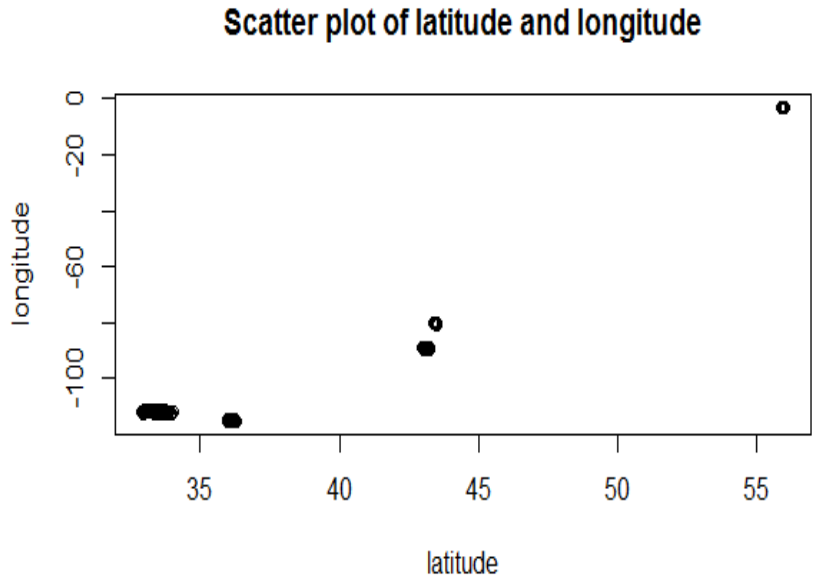
(e)

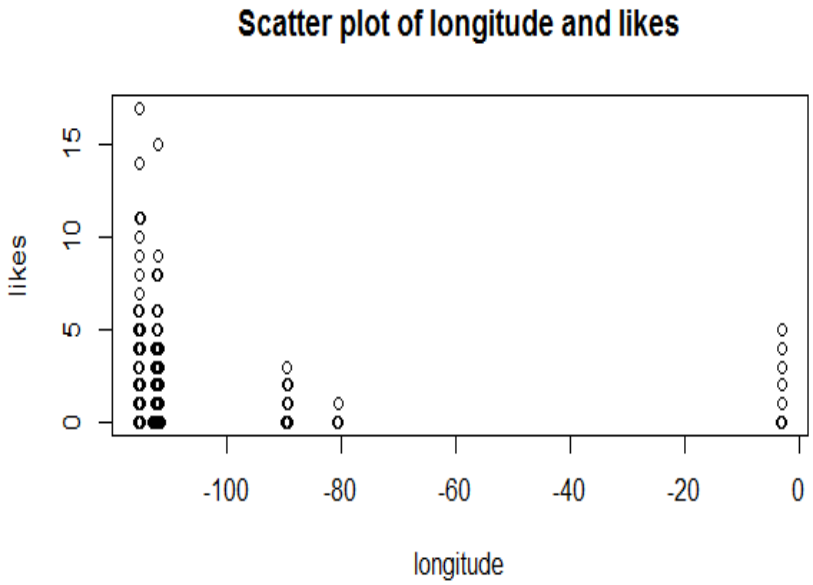
```
##########################compute pairwise correlation#############################
lat_long = cor(w[,"latitude"],w[,"longitude"])
print("pairwise correlation of latitude and longitude:")
print(lat_long)

lat_stars = cor(w[,"latitude"],w[,"stars"])
print("pairwise correlation of latitude and stars:")
print(lat_stars)

lat_likes = cor(w[,"latitude"],w[,"likes"])
print("pairwise correlation of latitude and likes:")
print(lat_likes)

long_stars=cor(w[,"longitude"],w[,"stars"])
print("pairwise correlation of longitude and stars:")
print(long_stars)

long_likes=cor(w[,"longitude"],w[,"likes"])
print("pairwise correlation of longitude and likes:")
print(long_likes)

stars_likes=cor(w[,"stars"],w[,"likes"])
print("pairwise correlation of stars and likes:")
print(stars_likes)
```

```
"pairwise correlation of latitude and longitude:"
0.9555044
"pairwise correlation of latitude and stars:"
0.1306059
"pairwise correlation of latitude and likes:"
-0.04086409
"pairwise correlation of longitude and stars:"
0.140871
"pairwise correlation of longitude and likes:"
-0.07586773
"pairwise correlation of stars and likes:"
0.1215371
```

Accordingly, the largest positive correlation is shown among latitude and longitude. The largest negative correlation is shown among longitude and likes. Given the fact that all most all the business locations are in North American continent, the positive correlation among latitude and longitude is expected. The positive correlation among likes and stars is also expected. However, we can not interpret any meaningful relationship from the correlation of the other pairs of attributes as they are not related at all.

The scatter plots of the two attributes with the largest positive correlation and the largest negative correlation is shown below respectively.




Scatter plot of longitude and likes
likes
0
5
10
15
-100
-80
-60
-40
-20
0
longitude

(f)

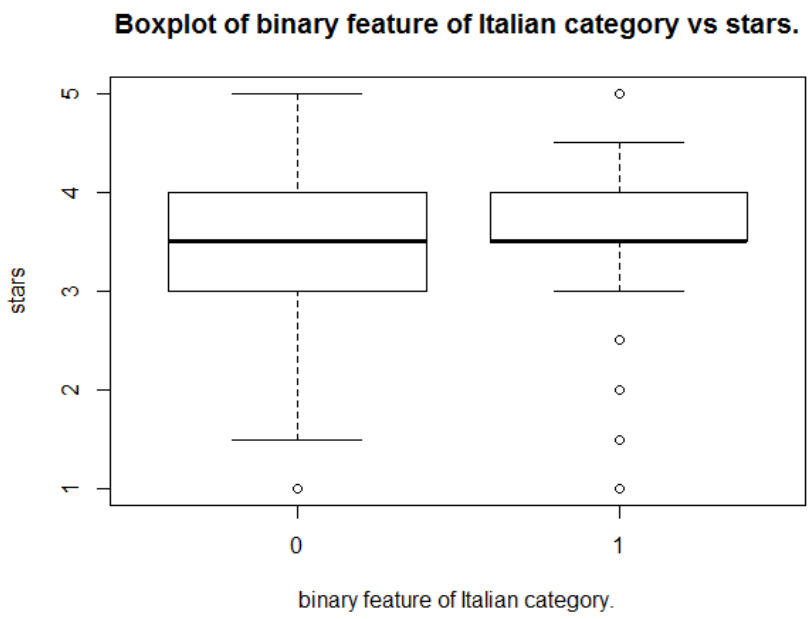
It was observed that all the businesses in the given dataset are Restaurants. Therefore, I felt it would be interesting to observe how stars/likes vary based on the different types of restaurants.

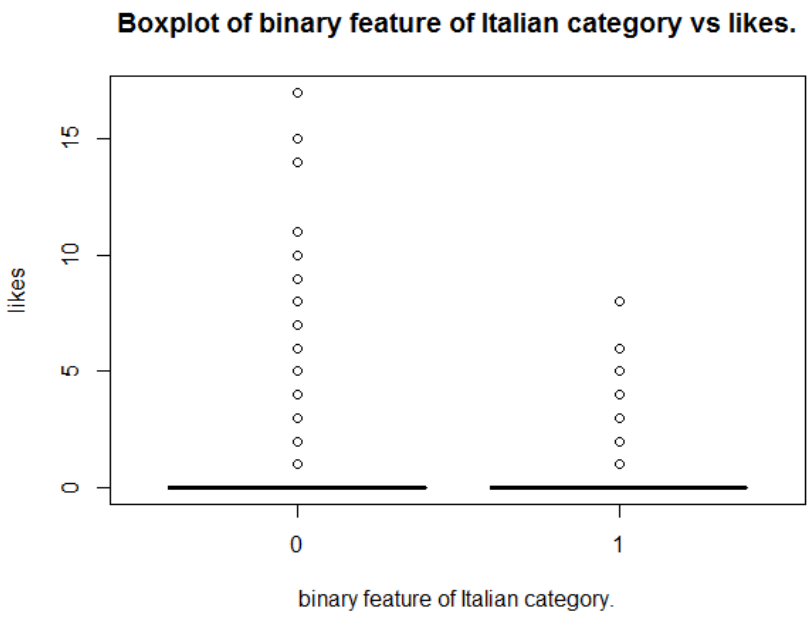
Category based on which the first binary feature was created is: Italian.

From the box plot of the binary feature vs stars, it is interesting to observe that Italian restaurants have a minimum stars of 3 with compared to all the other non-italian restaurants which have minimum lower than that. However, the median stars are equal in both Italian and non-italian restaurants.

From the box plot of the binary feature vs likes, it can be observed that median and the inter quantile ranges of both the categories lies around zero.

```
###################binary features############################
cat = "Italian"
cat_col = w[,"categories"]
i=1
binary_rest = vector('integer');
for(cell in cat_col){
  if(grepl(cat,cell)){
    binary_rest[i] = 1
  } else {
    binary_rest[i] = 0
  }
  i=i+1
}
ww = cbind(a=binary_rest,w)
boxplot(stars~a,data=ww,main="Boxplot of binary feature of Italian category vs stars.", xlab="binary feature of Italian category.",ylab="stars")
boxplot(likes~a,data=ww,main="Boxplot of binary feature of Italian category vs likes.", xlab="binary feature of Italian category.",ylab="likes")

cat1 = "Indian"
cat2 = "Chinese"
cat3 = "Fast Food"

i=1
binary_rest = vector('integer');
for(cell in cat_col){
  if(grepl(cat1,cell)){
    binary_rest[i] = 1
  } else {
    binary_rest[i] = 0
  }
  i=i+1
}
ww = cbind(b=binary_rest,w)
boxplot(stars~b,data=ww,main="Boxplot of binary feature of Indian category vs stars.", xlab="binary feature of Indian category.",ylab="stars")
```


Boxplot of binary feature of Italian category vs stars.
stars
1
2
3
4
5
0
1
binary feature of Italian category.


Boxplot of binary feature of Italian category vs likes.
likes
0
5
10
15
0
1
binary feature of Italian category.

(g)

As mentioned in the answer of part (g), I selected another three restaurant categories: Chinese, Indian and Fast Food in order to construct the binary features from and compare the variation of the statistics of stars attribute with the categorization based on such binary features.

The corresponding box plots are shown below.

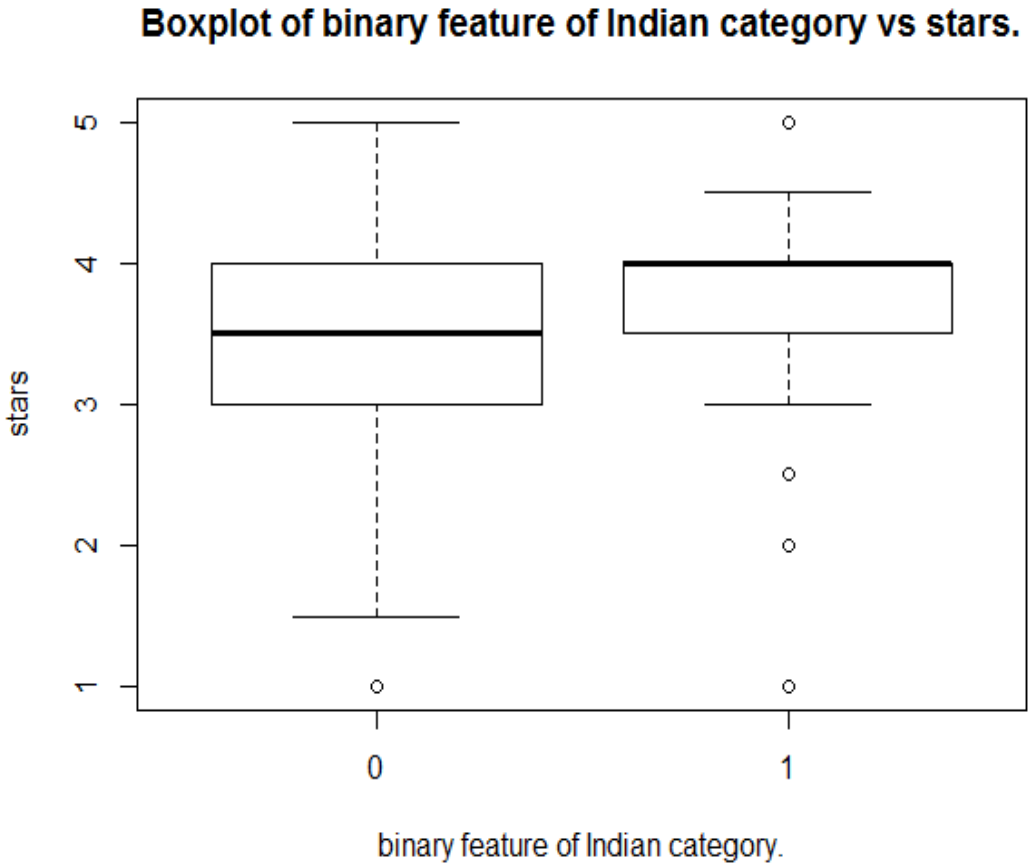
Based on the these box plots we can observe that:

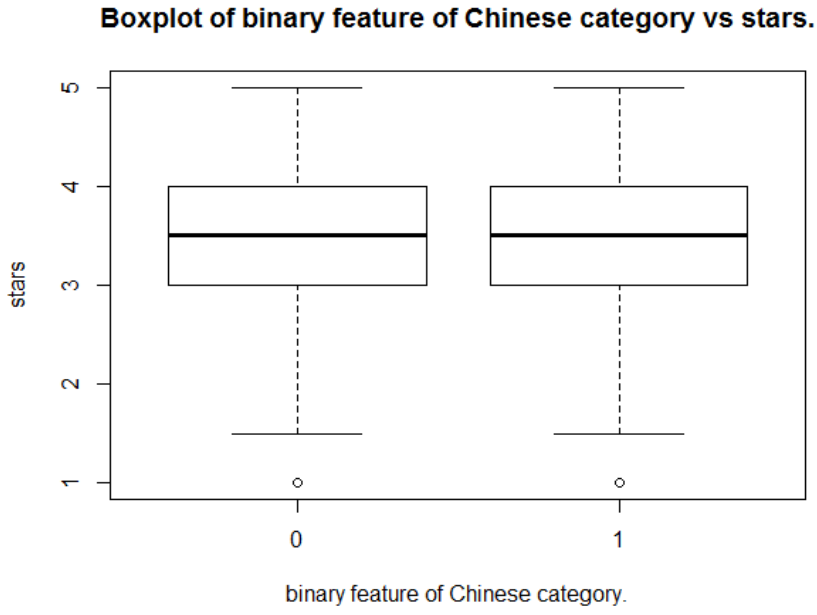
1)Indian restaurants have a overall higher median stars compared to the non-Indian restuarants. Their overall minimum stars is also higher than the non-Indian counterparts.

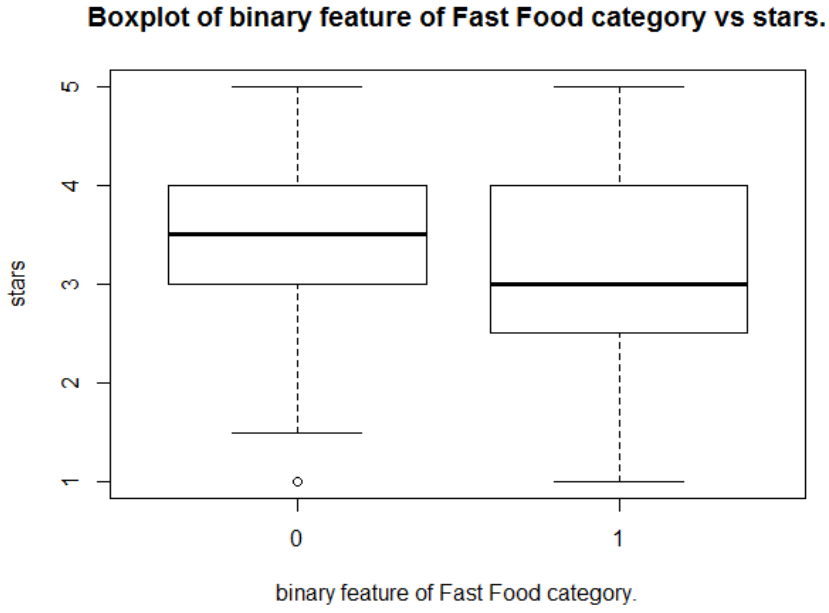
2)Chinese restaurants have got a similar star distribution as the non-Chinese restaurants.

3)Fast food restaurants have got an overall median and a minimum which is lower than the non-fast food restaurants.

I think the first and third observations above are interesting and expected considering the general popularity of those types of restaurants among our communities as well.


Boxplot of binary feature of Indian category vs stars.
stars
1
2
3
4
5
0
1
binary feature of Indian category.


Boxplot of binary feature of Chinese category vs stars.
stars
1
2
3
4
5
0
1
binary feature of Chinese category.


Boxplot of binary feature of Fast Food category vs stars.
stars
1
2
3
4
5
0
1
binary feature of Fast Food category.