Data Mining (CSCI B-565
Assignment No. 3
Masters in Data Science
Indiana University
Bloomington, IN, USA

Abhishek Rapelli
arapellii@iu.edu

October 23, 2017

All the work herein is solely mine

1 Mentioned below is the code used to retrieve data in R.

```
library(data.table)
data_frame <- fread("https://archive.ics.uci.edu/ml/machine-learning-databases/
breast-cancer-wisconsin/breast-cancer-wisconsin.data",na.strings="?" )
head(cancer_data)
colnames (data_frame) <- c("scn","clump.thickness","cell.size","cell.shape","margin
"bare.nuclei","bland.chromatin","normal.nucleoli","mitoses","class" )
head(data_frame)
table(data_frame$class)
summary(data_frame)
```

2.1 Given that the cost of biopsy is between $1000 and $5000 (including the pathologist cost). So the total cost of the biopsies for 699 instances is between $699,000 and $3,495,000.

2.2 Given that the cost of masectomy is between $15,000 and $55,000. From the summary of $data_frame, we have 2 and 4 vaules in the class coloumn and from the given data summary, 2 is for benign$
$241 = \$13,255,000$
$From the given data, out of 699 instances, 241 instances are of malignant instance and hence with assum$
$0.7 or 168 death cases.$

2.3 The data has given the following information: 1. Total 699 instances, 2. Total 241 malignant instances 3. The rest are benign. Also the costs range of biopsy and masectomy per instance are given. We are asked to find the total cost range of biopsy and masecutomy based on the given data. Also the death cases based on a given mortality rate.

2.5 Given total attributes as per the table are 11. If we now exclude the "scn" and "c" coloumns, we are left with 9 attributes.

2.6 In the given csv file, there are 16 missing values labelled as '?' and there are 16 women with no bare nuclei values. Hence it has 16 tuples.

2.7 There are 16 missing values.

2.8 The scn's are : 1057013 1096800 1183246 1184840 1193683 1197510 1241232 169356 432809 563649 606140 61634 704168 733639 1238464 1057067

```
#R code to find the scn's of missing values in the data_frame:
>data_frame[!complete.cases(data_frame),]$scn
 [1] 1057013 1096800 1183246 1184840 1193683 1197510 1241232
169356  432809  563649  606140   61634  704168  733639 1238464 1057067
```

2.9 It is better to go take re-examination of those 16 missing values women. If not 16, at least 2 malignant women must be re-examined to avoid deaths. The cost for biopsy of 2 of thse women suffering from malignant tumor is $2000 to $10,000 only, which not a big cost. Moreover we can update the results in place of missing values, which may be useful for more accurate analysis of data for future cases.

2.10 The missing data is not significant from algorithmic perspective because it is just 16 out of 699 instances, which is just 2.2288 percent.

2.11 By humanity point of view, it is better to keep the values of the diagnoised women with malignant tumor, which is more severe and chances of death is high. It is optional to keep or remove the missing values of the benign cancer women.

3.a The below script was used to create the table to store the cancer data in Postgresql

```
#R code for Histogram plots
install.packages("RPostgreSQL");
library('RPostgreSQL');

drivers = dbDriver("PostgreSQL");
connect = dbConnect(drivers, user="postgres",
password="adminpd",host="localhost", port=5432, dbname="Mydatabase");
data_frame = dbGetQuery(connect, "select * from uw_cancer_data c");
for (col in 2:ncol(data_frame))
{
hist(data_frame[,col],main = "Histogram Plot",
xlab = names(data_frame)[col] ,ylim=c(1,700))
}
dbDisconnect(connect)
```

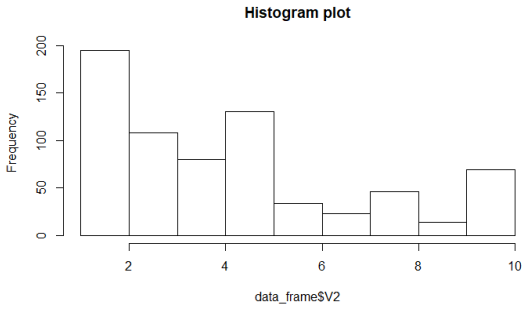
A relation is said to be Equivalence relation if it satisfies 1. Reflexive, 2. Symmetric and 3. Transitive relations.

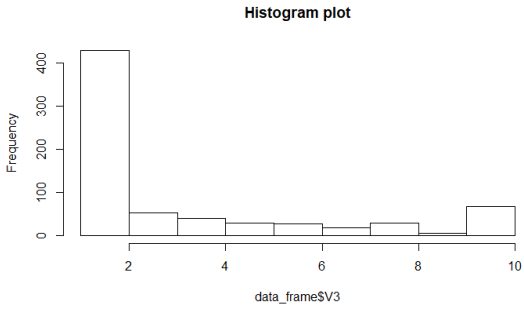
Reflexive: If a set, X is related to itself, X. [b] Symmetric: If a set X is related to another set Y, then Y is related to X. [c] Transitive: If X is related to Y and Y is related to Z then X is related to Z.
Given that x[i] = ith character and length(x) is the size of the word x such that 0 i length(x) 1.

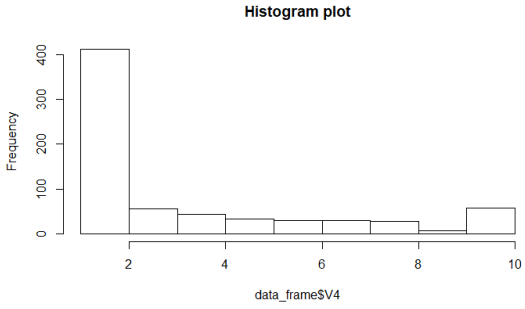
4.1 Given relation is length(x) = length(y):
For all x, y belongs to X, the lengths of strings x and y are same. If x =y, then we can say x length is same as itself
For all x,y belongs to X, and x not equal to y, we can have strings x amd y such that length(x) = lemgth(y) when both contain same number of characters. Hence this is symmetric relation. For all, x,y,z belomgs to X, and x, y and z not equal to each other, Let length(x) = length(y) and length(y) = length(z), it implies that x and y have same number of chars and y and z have same number of chars thus x and z will have same number of chars, thus this is a transitive relation property. Hence the relation is equivalence relation.

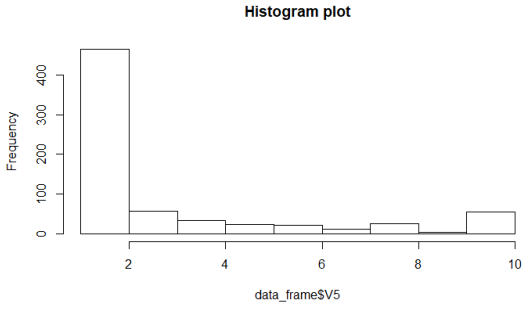
4.2 Given relation is x[0] = y[0], that is the first characters of each of x and y are same.
For all x, y belongs to X, we have x[0] = y[0] and obviously, this holds true for x=y, thus this is a reflexive property.
For all x,y belongs to X , Let y[0] = k. Now from the given relation, x[0] = y[0] this imples x[0] = k and hence we can say y[0] = x[0] = k. This proves the symmetric property. For all x,y,z belongs to X, Let y[0] = k. Now from the given relation x[0] = y[0] , we can say y[0] = k. Let z[0] = y[0] , which is k. Thus z[0] = x[0] = k. Hence we proved that if x[0] = y[0] and y[0] = z[0], then x[0] = z[0], Hence it is Equivalence relation.

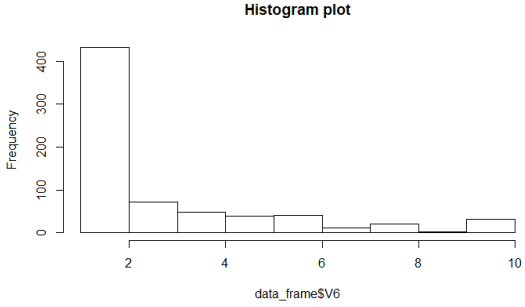
4.3 Given relation is x, y share at least one character.
For all x,y belongs to X, we have x and y share at least one character. We can sa x will have all same elements as x itself by common sense and hence we proved the relation to be Reflexive.
For all x,y belomgs to X, we have x and y share at least one character this also means the converse that y shares at least one element in common with x, Hence ths is clearly a transitive relation.
For all x,y,z belongs to X, we have x and y share at least one character say 'k'. Now let, y and z share at least one character say 'l', now there is no guarantee that k and l are same, they may be or may not be same and hence we cat say that x and z share at least one common character. Thus this is not a transitive property. Hence this is not a equivalence relation.

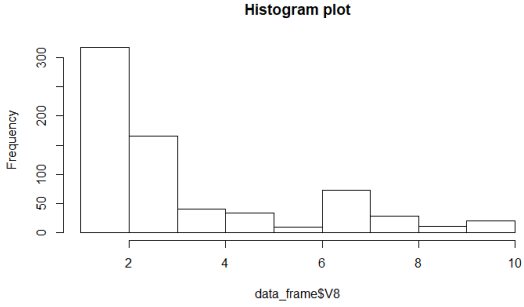

Histogram plot
Frequency
0
50
100
150
200
2
4
6
8
10
data_frame$V2

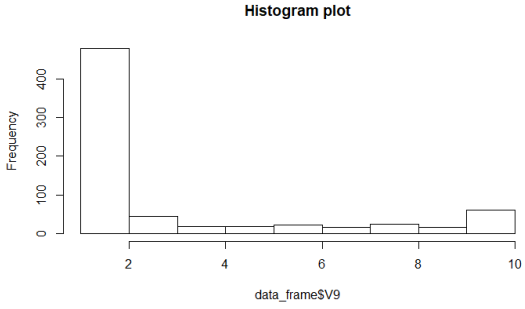

Histogram plot
Frequency
0
100
200
300
400
2
4
6
8
10
data_frame$V3

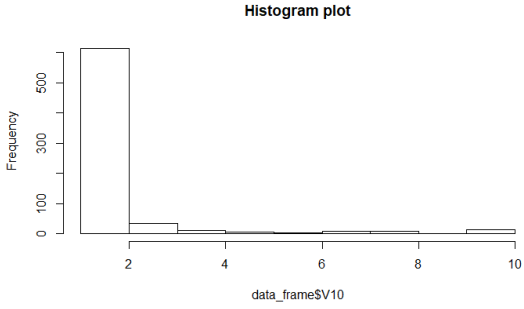

Histogram plot
Frequency
0
100
200
300
400
2
4
6
8
10
data_frame$V4


Histogram plot
Frequency
0
100
200
300
400
2
4
6
8
10
data_frame$V5


Histogram plot
Frequency
0
100
200
300
400
2
4
6
8
10
data_frame$V6


Histogram plot
Frequency
0
50
100
200
300
2
4
6
8
10
data_frame$V8


Histogram plot
Frequency
0
100
200
300
400
2
4
6
8
10
data_frame$V9


Histogram plot
Frequency
0
100
300
500
2
4
6
8
10
data_frame$V10