Introduction to Data Analysis and Mining
Homework #2

Due on Sunday, Feb 9, 2018 10:00 p.m.

Instructor: Hasan Kurban

Kwong Yuet Michael Fadillah Wong

February 8, 2018

Contents

Problem 1

Textbook exercises, chapter 2, pages: 91-93

1. Exercise 12 **(10 points)**
   12a) Noise is not interesting nor desirable, and outlier is interesting and desirable.
   12b) Yes, Noise can be outlier
   12c) No, Noise is not always outlier, it can also modifies the data object.
   12d) No, it can be the one class that is different than other attributes.
   12e) Yes

2. Exercise 15 **(10 points)**
   15a) We select n*mi/m elements from the group
   15b) We view all the data as a whole, and select n elements from the data.

3. Exercise 16 **(10 points)**
   16a) For one document, terms have max weight, while terms have 0 weight in every document.
   16b) From this transformation, terms occur in every document cannot individualize one document from another.

4. Exercise 18 **(15 points)**
   18a) Hamming distance = 3
   Jaccard Similarity = 2/5 = 0.4
   18b) Hamming distance is more similar to the Simple Matching Coefficient, because Simple Matching Coefficient = Hamming distance / number of bits.
   The Jaccard similarity is more similar to the cosine measure, becuase they ignore matches.
   18c) Jaccard similarity, because we want to see the number matches.
   18d) Hamming distance, because we need to find the difference between genes in order to compare the genetic makeup of two organisms of the same species.

5. Exercise 19 **(15 points)**
   19a) Correlation:> $cor(c(1,1,1,1), c(2,2,2,2))$
   [1]NA
   Euclidean distance:> $sqrt(sum((c(1,1,1,1) - c(2,2,2,2))^2))$
   [1]2
   cos:(8/(2*4)) = 1
   19b) Correlation:> $cor(c(0,1,0,1), c(1,0,1,0))$
   $[1] - 1$
   Euclidean distance:> $sqrt(sum((c(0,1,0,1) - c(1,0,1,0))^2))$
   [1]2
   Jaccard: number of matches = 0.
   Therefore Jaccard Similarity = 0
   cos: numerator = 0, cosine similarity = 0
   19c) Correlation:> $cor(c(0,-1,0,1), c(1,0,-1,0))$
   [1]0
   Euclidean distance:> $sqrt(sum((c(0,-1,0,1) - c(1,0,-1,0))^2))$
   [1]2
   cos: numerator = 0, cosine similarity = 0
   19d) Correlation:> $cor(c(1,1,0,1,0,1), c(1,1,1,0,0,1))$
   [1]0.25
   Jaccard: 3/(6-1) = 0.6

cos: (3/(2*2)) = 3/4 = 0.75
19e) Correlation:> $cor(c(2,1,0,2,0,3), c(1,1,1,0,0,1))$
[1]0
cos: numerator = -2 -1 + 3 = 0, cosine similarity = 0

Problem 2

For the following data, give the best taxonomic type (interval, ratio, nominal, ordinal):

1. A section of highway on a map.
   Nominal
2. The value of a stock.
   Interval
3. The weight of a person.
   Ratio
4. Marital status.
   Nominal
5. Visiting United Airlines (`https://www.united.com`) the seating is: Economony, Economy plus, and United Business. **(10 points: each question is worth 2 points)**
   Ordinal

Problem 3

You are datamining with a column that has physical addresses in some city with the same zipcode. For example,

```
55 WEST CIR
2131 South Creek Road
Apt. #1 Fountain Park
1114 Rosewood Cir
1114 Rosewood Ct.
1114 Rosewood Drive
```

What structure would you create to mine these? What questions do you think you should be able to answer? **(10 points)**
I will divide the address into three parts; apartment number, street name, and street suffix. For example:
For 2131 South Creek Road:
Apartment number: 2131
Street name: South Creek
Street Suffix: Road
By this kind of structure, we can answer the address by part, such as how common is the street name overlapped?

Problem 4

The Wisconsin Breast Cancer data set is very famous. Here is the URL `https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)`. In the Data Folder are multiple files. Here is the beginning of an R session that allows us to read this data from the web into our local R session:

```
> install.packages("data.table")
> library(data.table)
> install.packages("curl")
> mydata <- fread("https://archive.ics.uci.edu/ml/machine-learning-databases/
                   breast-cancer-wisconsin/breast-cancer-wisconsin.data")
> head(mydata)
        V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11
1: 1000025  5  1  1  1  2  1  3  1   1   2
2: 1002945  5  4  4  5  7 10  3  2   1   2
3: 1015425  3  1  1  1  2  2  3  1   1   2
4: 1016277  6  8  8  1  3  4  3  7   1   2
5: 1017023  4  1  1  3  2  1  3  1   1   2
6: 1017122  8 10 10  8  7 10  9  7   1   4
>
```

Discussion of Data

Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?
The wisconsin breast cancer data table that examines how a factor responsible to a benign/malicient tumor. We have eleven attributes: id, Clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size bare nuclei, bland chromatin, normal nucleoli, mitoses, amd class.
We can transfrom the data into a numeric table since most of the data attributes is shown in a rank (1-10)

R Code

Using R, show code that answers the following questions:

1. How many entries are in the data set? Answer here ...

```
> nrow(mydata)
[1] 699
> nrow(mydata)*11
[1] 7689
```

2. How many unknown or missing data are in the data set? Answer here ...

```
> sum(is.na(mydata))
[1] 16
```

3. How many malignant and benign identifiers are there? Answere here ...

```
> benign <- nrow(mydata[mydata$V11 == 2,])
> benign
[1] 458
> malignant <- nrow(mydata[mydata$V11 == 4,])
```

```
> malignant
[1] 241
```

4. Make a histogram of each attribute and discuss the distribution of values *e.g.*, are uniform, skewed, normal. Place images of these histograms into the document. Show the R code that you used below and discussion below that.

```
id <- mydata$V1
hist(id)

clumpThinkness <- mydata$V2
hist(clumpThinkness)

uniformityOfCellSize <- mydata$V3
hist(uniformityOfCellSize)

uniformityOfCellShape <- mydata$V4
hist(uniformityOfCellShape)

marginalAdhesion <- mydata$V5
hist(marginalAdhesion)

SingleEpithelialCellSize <- mydata$V6
hist(SingleEpithelialCellSize)

bareNuclei <- mydata$V7
hist(bareNuclei)

blandChromatin <- mydata$V8
hist(blandChromatin)

normalNucleoli <- mydata$V9
hist(normalNucleoli)

mitoses <- mydata$V10
hist(mitoses)

dataClass <- mydata$V11
hist(dataClass)
```

Discussion of Attributes

Answer here... For v1 histogram, it can be ignored. Because v1 represent the ID.
From v2 to v10 histograms, those graphs show the frequencies of each topic occurrence.
For v11 histogram, the data separate into two columns, which are 2 and 4. 2 stands for benign, and 4 stands for malignant.

Histograms

Place images here with suitable captions.

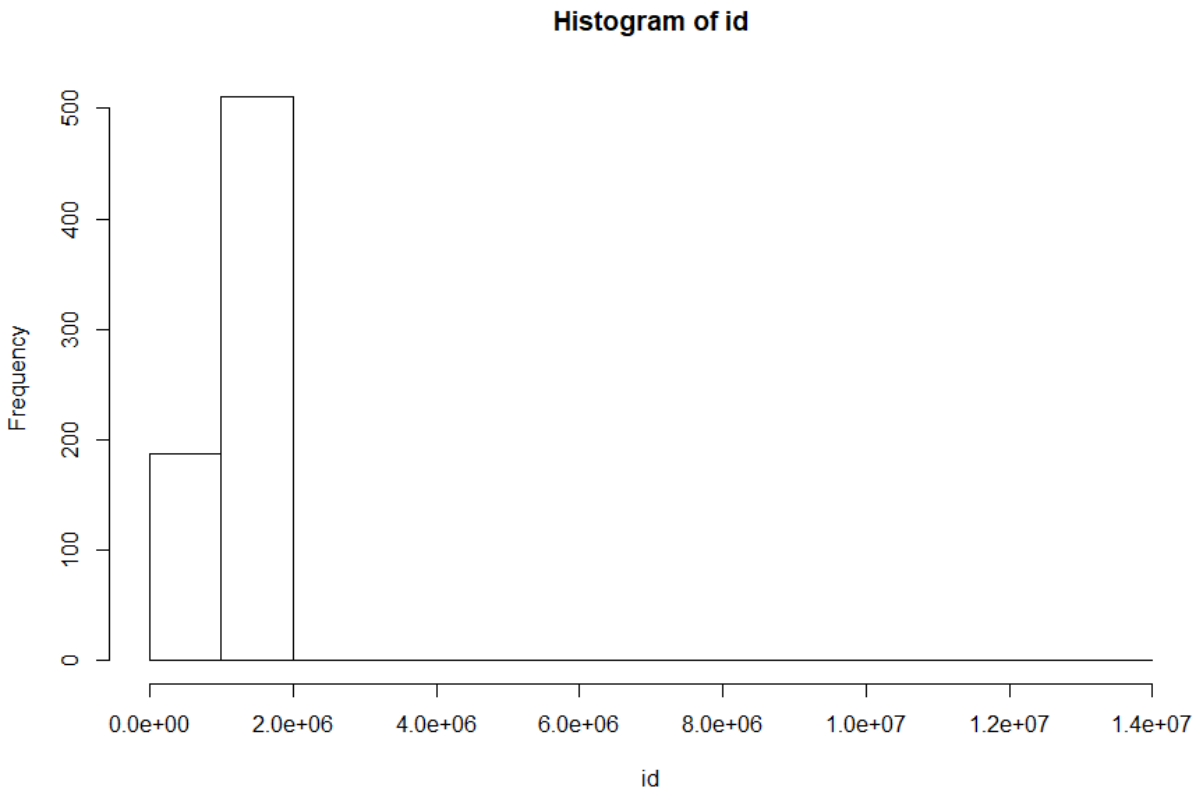
Histogram of id
Frequency
0
100
200
300
400
500
0.0e+00
2.0e+06
4.0e+06
6.0e+06
8.0e+06
1.0e+07
1.2e+07
1.4e+07
id

Figure 1: ID

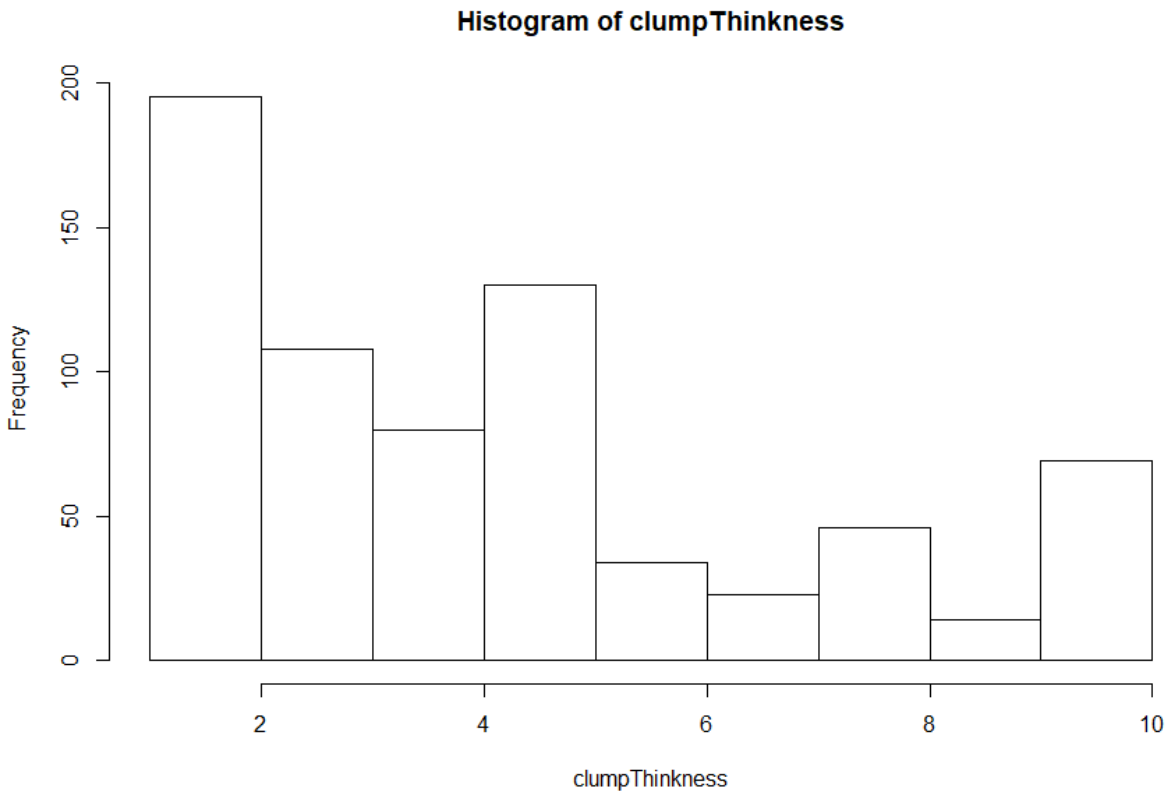

Figure 2: Clump Thickness

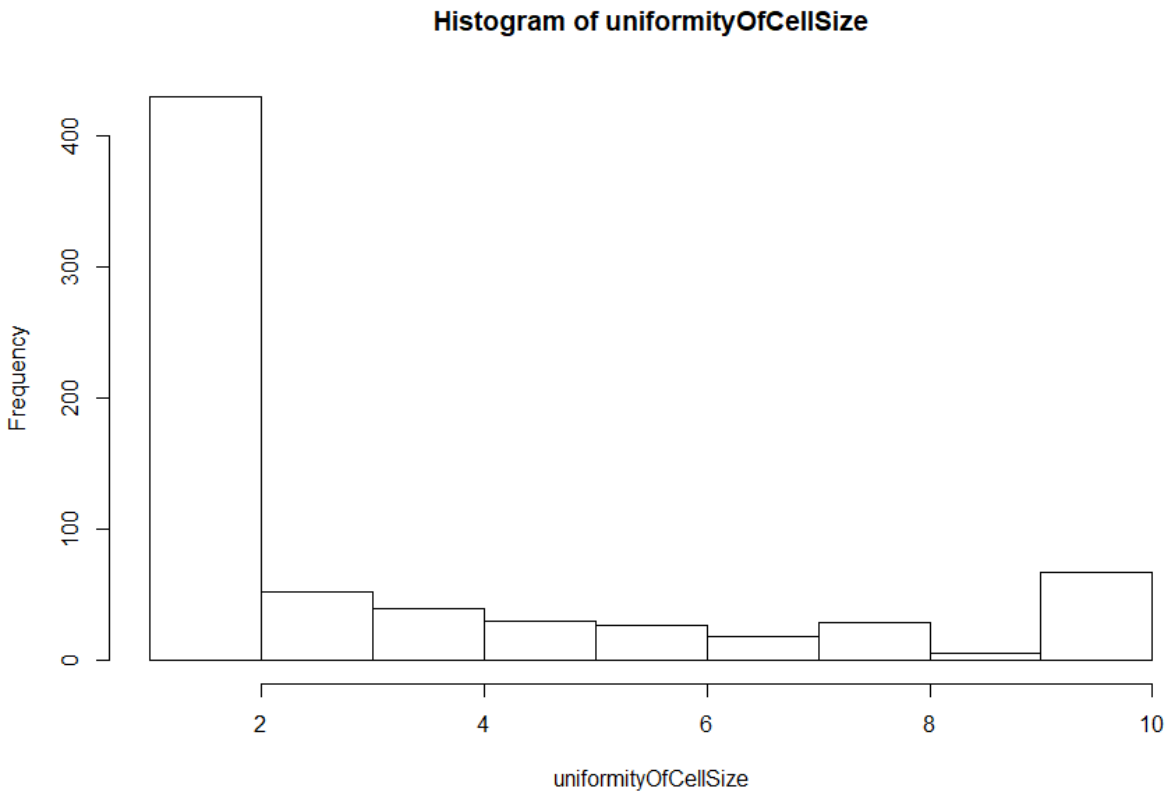

Figure 3: Uniformity of Cell Size: 1 - 10

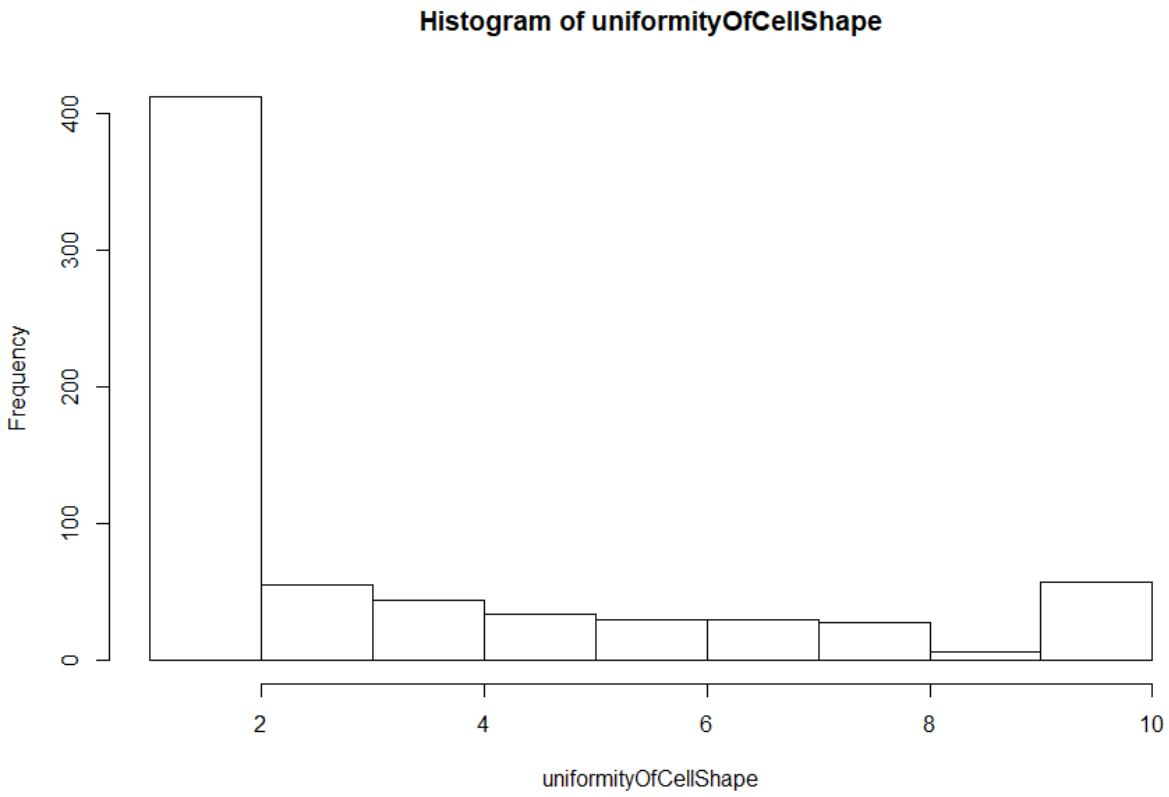

Figure 4: Uniformity of Cell Shape: 1 - 10

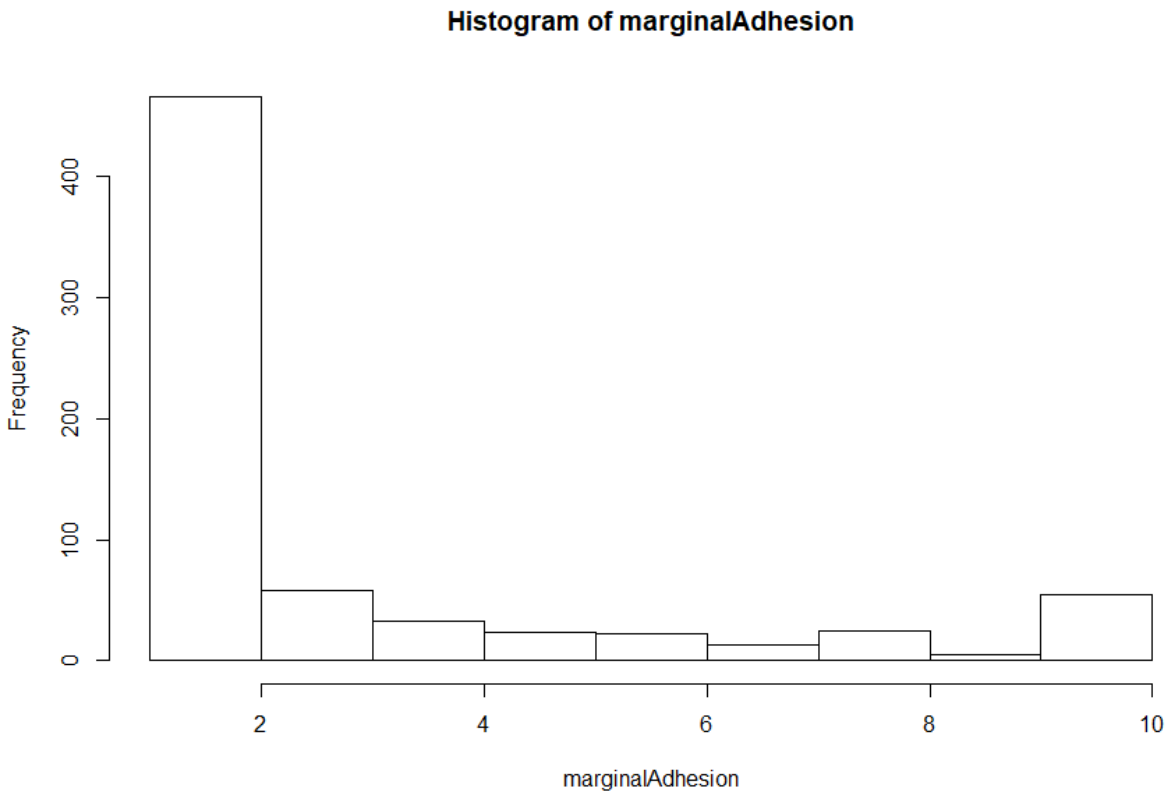

Figure 5: Marginal Adhesion: 1 - 10

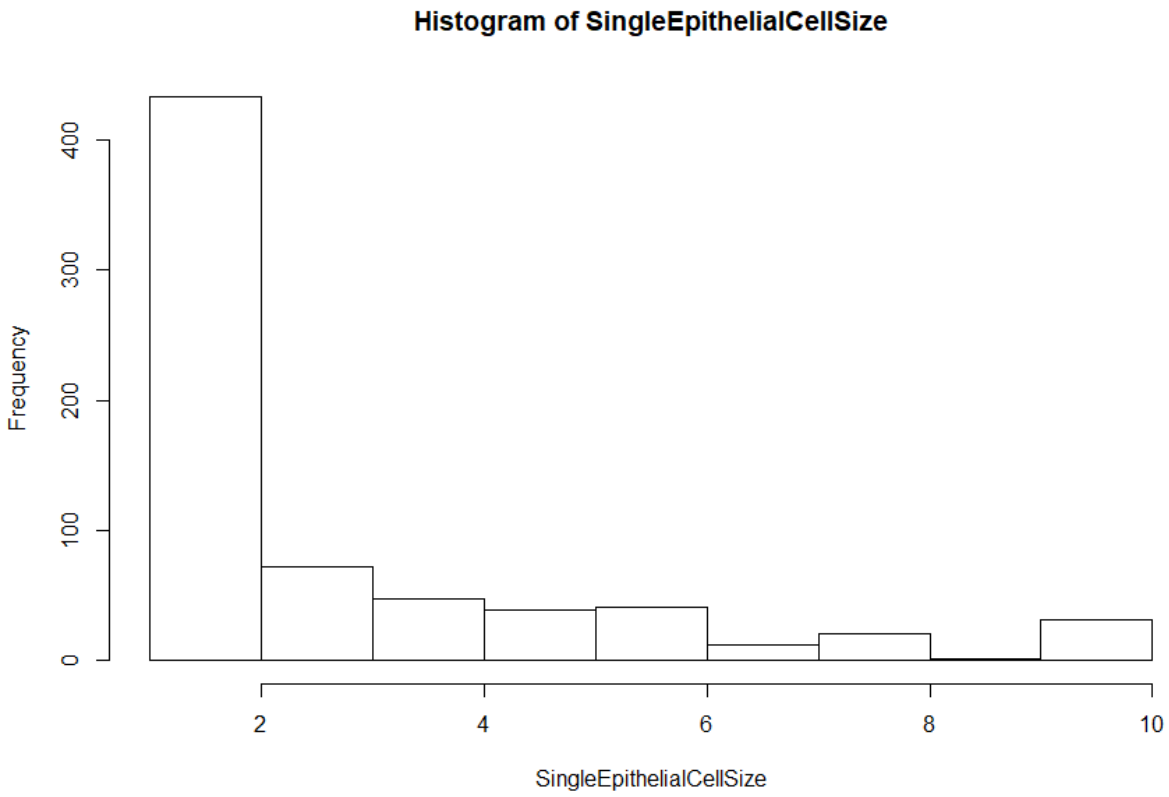

Figure 6: Single Epithelial Cell Size: 1 - 10

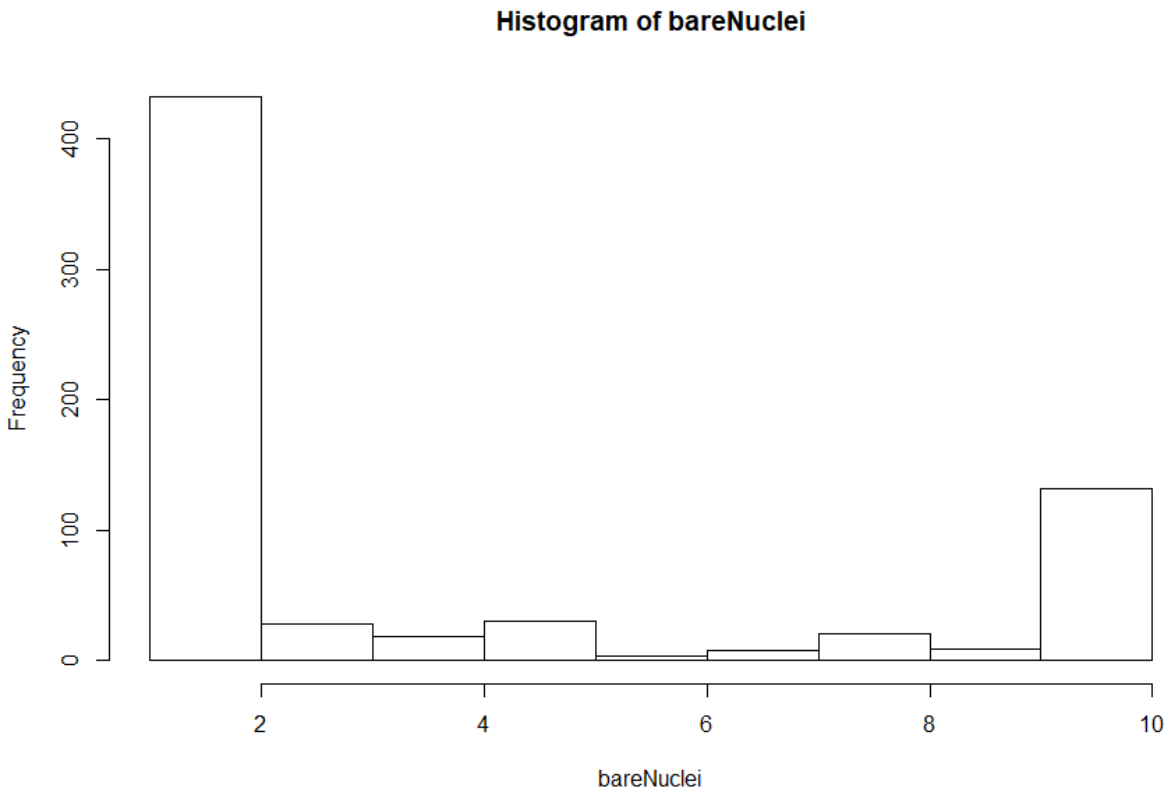

Figure 7: Bare Nuclei: 1 - 10

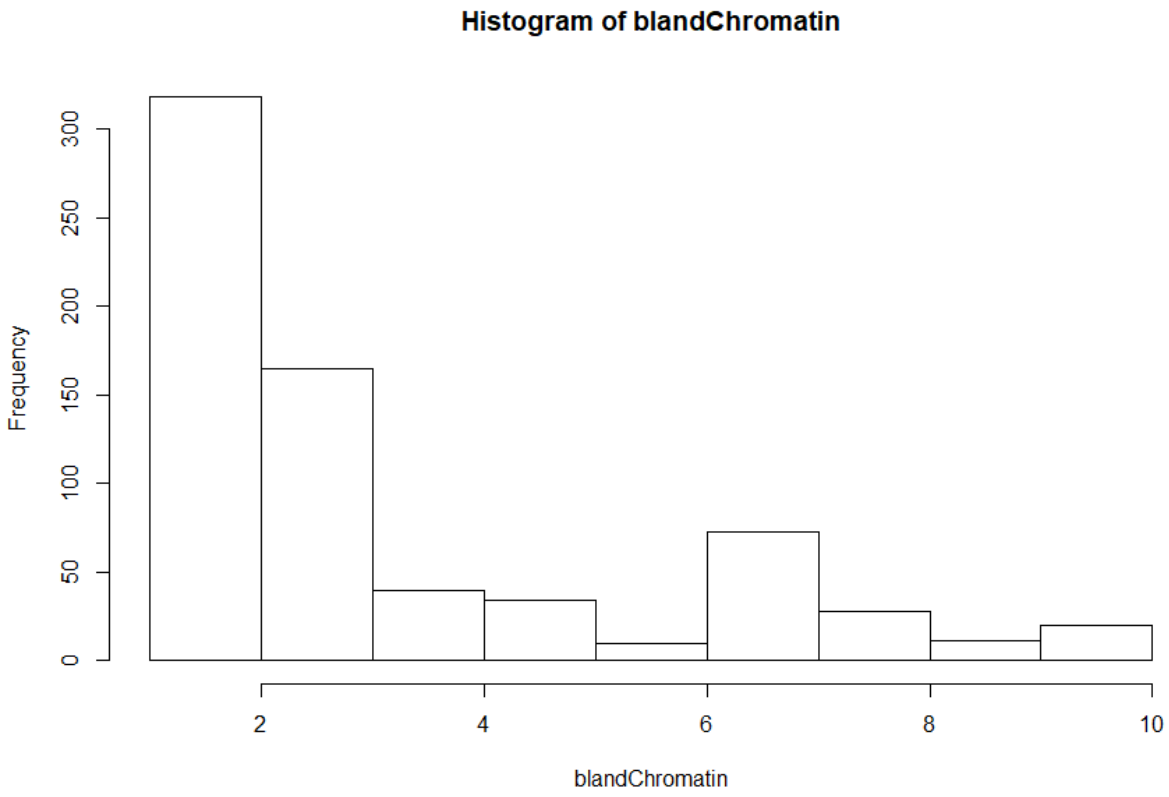

Figure 8: Bland Chromatin: 1 - 10

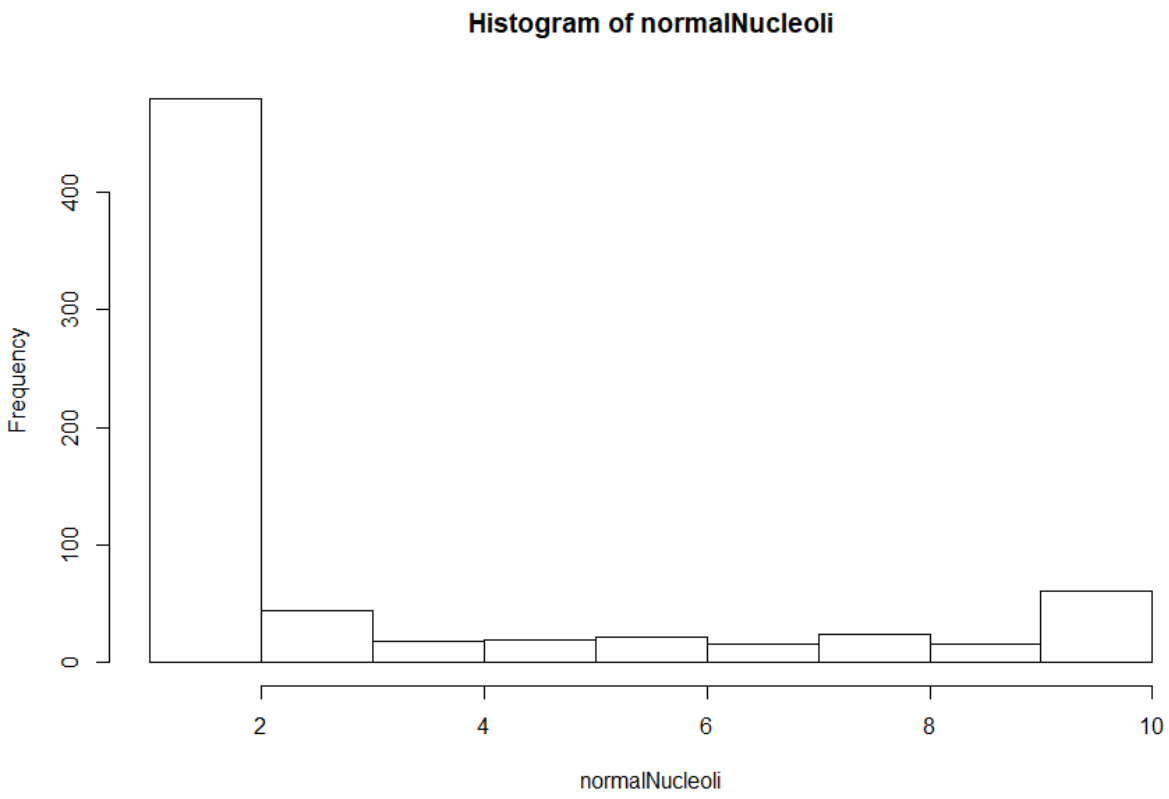

Figure 9: Normal Nucleoli: 1 - 10

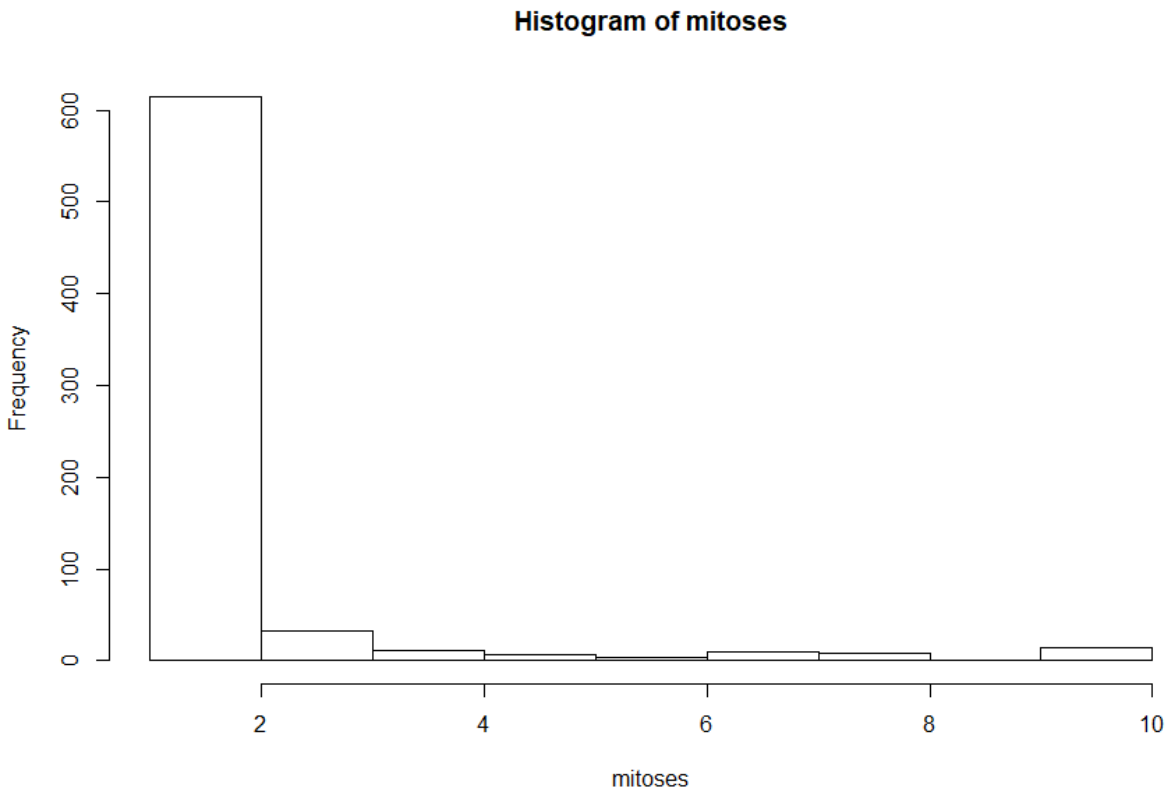

Figure 10: Mitoses: 1 - 10

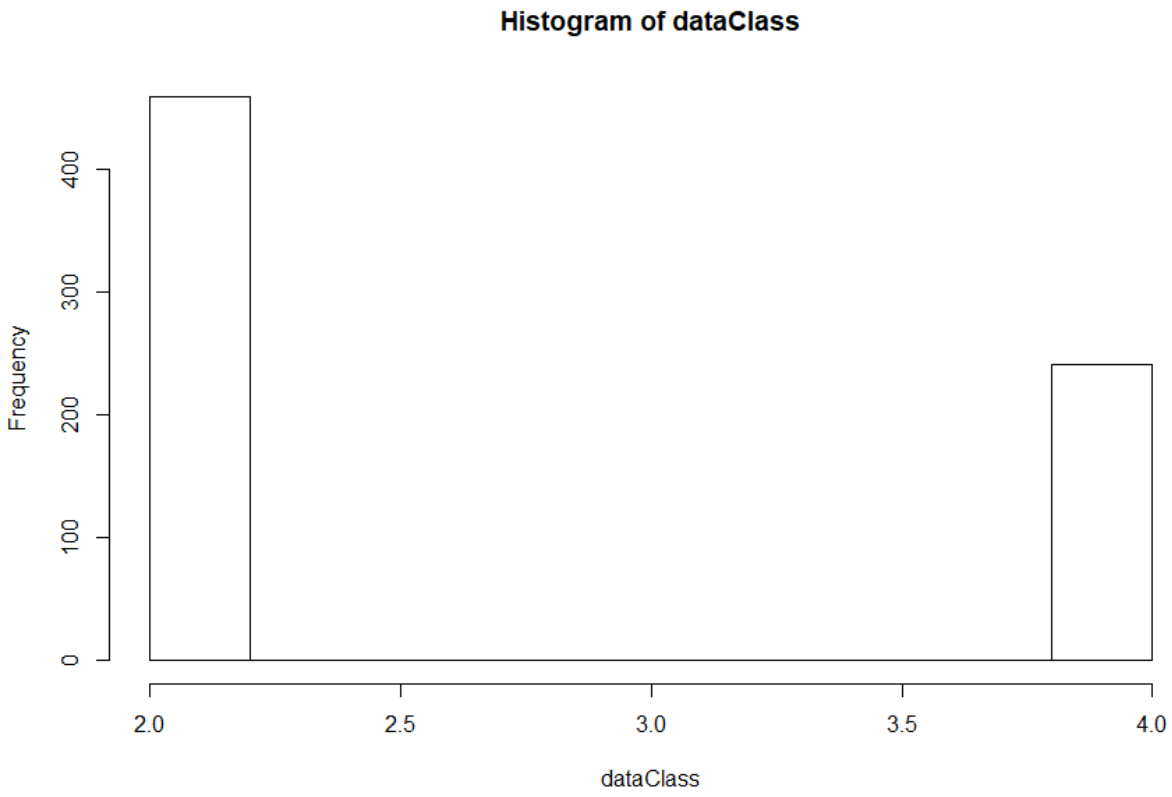

Figure 11: Class: (2 for benign, 4 for malignant)

Discussion of simply removing tuples

Quantify the affect of simply removing the tuples with unknown or missing values. What is the cost in human capital? **(20 points)**
When the missing value is removed, the data on the other attributes will also be removed, so we might lose a lot of information that is useful. Instead of removing missing value, we can use other algorithm, such as replace missing values with mean or median.

What to Turn-in

Please follow the syllabus guidelines in turning in your homework. I am providing the LaTeX of this document too. This homework is due Friday, Feb 9, 2018 10:00p.m. **OBSERVE THE TIME**. Absolutely no homework will be accepted after that time. All the work should be your own. Submit a .zip file that includes the files below. Name the .zip as "usename-section number", i.e., hakurban-B365.

1. The *tex and *pdf of the written answers to this document.