BUSINESS DATA MINING (IDS 572)

HOMEWORK 1
DUE DATE: WEDNESDAY, SEPTEMBER 07 AT 3:00 PM

- Please provide succinct answers to the questions below.
- You should submit an electronic pdf or word file in blackboard.
- Please include the names of all team-members in your write up and in the name of the file.
- You should include all the R functions you use.
- Before importing the EXCEL data file into R, you can rename the variables in the first row and give them a shorter name so that it will be easier for data processing later.

Problem 1. Download the data hw1.xls from blackboard. Use RStudio to answer the following questions.

(a) How many attributes and instances are in this data set? How do you get this information in RStudio?

(b) Write down the measurements of all the variables.

(c) Find the overall mean, median, variance and sample standard deviation of weight variable in this data set.

(d) Does the distribution of the weight data have a symmetrical belled-shape? Justify your answer.

(e) Report the percentage distribution of the Ate Fried Food variable?

Ate Fried Food	Relative Frequency
None	_____ %
Less than 3 times	_____ %
At least 3 times	_____ %

(f) Report the percentage distribution of Exercise Per Week variable. Write your answer into a table similar to the table in part (e).

(g) Does the weight data suggest that it was from a normally distributed population? Perform a normality test and report the p-value of the test using 5% as the cutoff for decision making of the normality test.

(h) Report the mean, median and sample standard deviation of weight variable for female subjects in this data set.

Problem 2. Use hw1.xls to answer the following questions. Include all the charts with proper labels in your report. Please, for each of the chart you produced, write a sentence or two explaining what you see from the chart.

(a) Make a frequency distribution table for the gender variable to see the frequency distribution.
(b) Make a bar chart for gender variable.
(c) Make a histogram to display the distribution of the Height variable.
(d) Make a cluster bar chart (side-by-side bar chart) to examine the correlation between gender and Ate Fried Food variables.
(e) Make a scatter plot to examine the correlation between Weight and Height variables, and write a sentence to describe the trend you observed from the scatter plot.
(f) Find the 5-number summary for the Height data and make a boxplot for the Height data with mild and extreme outliers identified using inner and outer fences. Draw the boxplot.

Problem 3. To solve this question you do not need to use R.
The following table comes from a hypothetical database.

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

(a) Find the support and confidence for the rule

$$(\text{Give Birth} = \text{no}) \wedge (\text{Can Fly} = \text{yes}) \rightarrow \text{Birds}$$

(b) Using the 1-rule method discussed in class, find the relevant sets of classification rules for the target "Class" by testing each of the input attributes Blood Type, Give Birth, Can Fly, and Live in Water. Which of these three sets of rules has the lowest misclassification rate?