Data Mining: Practical Assignment #1

Due on Thu & Fri, April 10-11 2014, 10:15am-13:15 & 14:15am-17:15

Task 1

Get familiar with Matlab/Octave:

i.) Which expression generates the column vector $A = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$?

ii.) Given a matrix $B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$.

Which of the following expressions would give the matrix $C = \begin{bmatrix} 2 & 3 \\ 5 & 6 \\ 8 & 9 \end{bmatrix}$?

1. `B(:,:)`
2. `B(:,2)`
3. `B(:,2:3)`
4. `B(2:3,:)`

iii.) Suppose you wish to generate a 3x1 vector D that contains the number 5 in every position.
Which of the following expressions will accomplish this task?

1. `eye(3)*5`
2. `ones(3)*5`
3. `ones(3,1)*5`
4. `fives(3,1)`

iv.) Which expression allows to create a new matrix E by appending the column vector D to the matrix B?

v.) Suppose you wish to generate a 2x3 matrix F that contains only zeros.
Which of the following expressions will achieve this goal?

1. `zeros(2)`
2. `zeros(3)`
3. `zeros(2,3)`
4. `zeros(3,2)`
5. `eye(2,3)`
6. `[0 0 0; 0 0 0]`
7. `[0 0; 0 0; 0 0]`

Present and discuss your solutions.

Task 2

In the lecture you encountered different data types. Think about some real-world data set examples containing:

1. Nominal data
2. Ordinal data

List 3 possible sets each or think about data sets combining the attributes. Of course, the examples from the lectures do not count as answers.

Task 3

The rent for appartments in your town is rising higher and higher. So, you decide to buy a condo (German: Eigentumswohnung) and pay the bank instead. The currently available flats are priced as follows:

1. 150000,00 Euro
2. 152000,00 Euro
3. 153000,00 Euro
4. 155000,00 Euro
5. 156000,00 Euro
6. 160000,00 Euro
7. 161000,00 Euro
8. 165000,00 Euro
9. 600000,00 Euro
10. 1000000,00 Euro
11. 2000000,00 Euro

You do not know about that, but your friend tells you that you have to pay on average 427000,00 Euro. Would you go ahead and check on your own or do you stay in your flat share?

Task 4

Consider this sequence of numbers as a datastream:

41 46 7 46 32 5 14 28 48 49 8 49 48 25 41 8 22 46 40 48 33 2 43 47 34 38 38 20 33
9 36 2 14 3 5 42 35 16 48 2 22 20 39 40 10 25 23 33 36 38 14 34 33 9 6 25 48 18

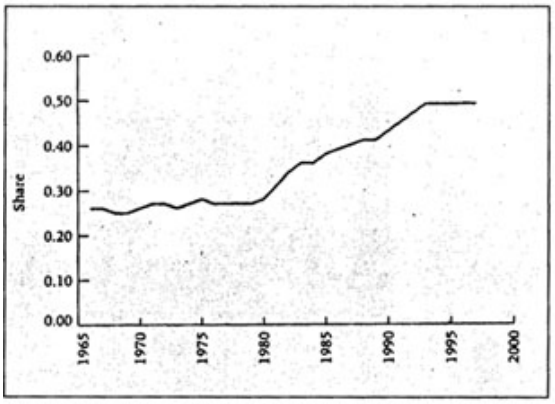
i.) Determine the following measurements:

1. Min- and max-values
2. Mean
3. Median
4. Q1 and Q3

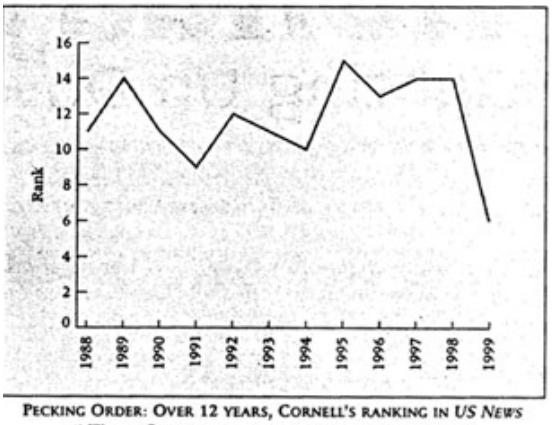
ii.) Based on that, draw a boxplot and explain the significant measurements reflected in such a plot.

Task 5

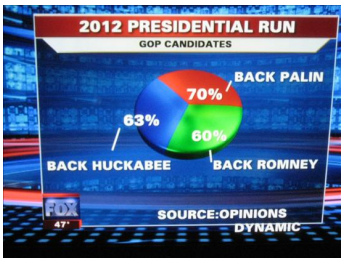
Data visualisation needs also data interpretation. Below, you find examples for (maybe) misleading data preparation and statistics. So, have a closer look at the figures below and discuss, what could be problematic with their representation or interpretation.



BY THE NUMBERS: OVER 35 YEARS, CORNELL'S TUITION HAS TAKEN AN INCREASINGLY LARGER SHARE OF ITS MEDIAN STUDENT FAMILY INCOME.



PECKING ORDER: OVER 12 YEARS, CORNELL'S RANKING IN *US NEWS & WORLD REPORT* HAS RISEN AND FALLEN ERRATICALLY.



Task 6

Data Sets: Load the data files from the data.zip into your Matlab workspace. Use the well-documented Matlab help to inform yourself about the possible visualization methods listed below. Decide, which data visualization method would be appropiate for which data set. We propose the following methods:

1. Scatter plot
2. Bar chart
3. Boxplot
4. Time series
5. Pie chart

Present and discuss your choice.