Workshop - Week 4: COMP20008

1. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

   - What is the mean of the data? What is the median?
   - What is the first quartile (Q1) and the third quartile (Q3) of the data?
   - What is the interquartile range of the data?
   - Show a boxplot of the data.

2. We are given some values: 1,2,3,4,5, 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70, 70, 300, 900, 10000. Using the Tukey boxplot method outlined in lectures, which of the following values would be

   - Suspected outliers
   - Outliers

3. Given two instances represented by the tuples (22, 2, 3, 1, 42, ?,10,?) and (?,16,?, 17,20, 0, 36, 8):

   - Compute the Euclidean distance between the instances using mean imputation (Method 1 in lectures).
   - Compute the Euclidean distance between the instances using scaling (Method 2 in lectures).
   - What are advantages and disadvantages of each?
   - Describe a scenario where the scaling method might give unintuitive results.

4. Recommender systems are challenged by the "cold start" problem - how to make recommendations to new users, about whom little is known. Suggest three strategies that might be used to address this.

5. Recommender systems are sometimes criticised for over-recommending popular items to users and under-recommending rarer items. Why do you think this happens? How might it be addressed?

6. Download, open and study each of these data-sets:

   - Australian federal election voting statistics by division

- CDC trend table of diabetes prevalance and glycemic control in the US, 2011
- Rainfall/temperature measurements since 1929 at Essendon Airport weather station

For each one identify the format of the data and in your own words (not the file title) describe what the data-set appears to represent. Think of a research question that you could explore using the variables that available in them. Finally, list all the numerical and categorical variables.

7. Download, open and study the file smoking_data_us_1995_2010-fixed.csv, showing United States population smoking data from 1995 to 2010. In the first twenty rows, there are seven errors that all fall into one of the following categories:

   - Semantic
   - Range errors
   - Format errors

   Identify the errors and what category they fall into. Where possible fix the errors manually and save the new spreadsheet as *smoking-info-corrected.csv* Write notes on how you would write a program to detect them.

8. - Create a blank jupyter notebook and then write Python code for the following
   - Import your file *smoking-info-corrected.csv* into a pandas data frame
   - Remove the percentage symbols from the data. For removing/replacing characters, see http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.replace.html and after the removals you will also need to convert all the strings to numeric values (*XXX.apply(pd.to_numeric)*)
   - Using *Dataframe.plot.scatter* create four individual scatter-plots for a time-progression from 1995 to 2010 in Alaska across the four smoking categories (hint: you need to filter the data using state name, then create a chart of year against the smoking characteristic column).
   - Using these charts identify any obvious clusters or trends. Can you characterise these trends into "if ... then" rules (it doesn't have to be too specific)?