

# INFO304 Assignment 1

## Data Analysis, Clustering and Visualisation

**DUE DATE: 10pm, Monday, 3<sup>rd</sup> August**

**Worth: 10% of final course mark**

---

### INTRODUCTION

This assignment will examine a range of visualisation and analysis methods for different data types. This will involve writing some R code to explore properties of different data sets and relevant discussions regarding the observed patterns and representation.

### TIME SERIES DATA: THE SP500 INDEX (15 MARKS)

Run the script **sp500.R** that has been supplied. This will load some libraries that handle time series data and, in particular, tools for examining stock market data. A plot should be displayed that shows the last 3 months of the SP500 index data (July-Sept, 2009) as a candle plot. Note that the lower part of the plot shows the volume associated with trades.

1. Examine the upper part of the plot and **describe what the candles are showing** (HINT: change the script to show just the last “1 months” and look at the data for the last “1 months”). **(2 marks)**
2. The daily average price can be approximated by:

$$Av(t) = \frac{C(t) + H(t) + L(t)}{3}$$

where  $C(t)$ ,  $H(t)$  and  $L(t)$  are the close, high and low quotes for day  $t$  respectively.

- 2.1 Calculate the average price  $Av(t)$  for the sp500 data and add this to the sp500 table.

**Produce a plot** showing  $Av(t)$  and the **R code** used to calculate  $Av(t)$ .

**(3 marks)**

3. Daily returns are commonly used in stock market analysis. The daily return is defined as:

$$\text{return}(t) = \frac{C(t) - C(t-1)}{C(t-1)}$$

- 3.1 Calculate the return for the **last 12 months** of the sp500 data and compare the candleChart for the **last 12 months** and the daily returns. **Discuss the patterns in daily returns in relation to the sp500 price signal.** Include the **figure(s)** showing the candleChart, returns plot and **R code** to calculate returns.

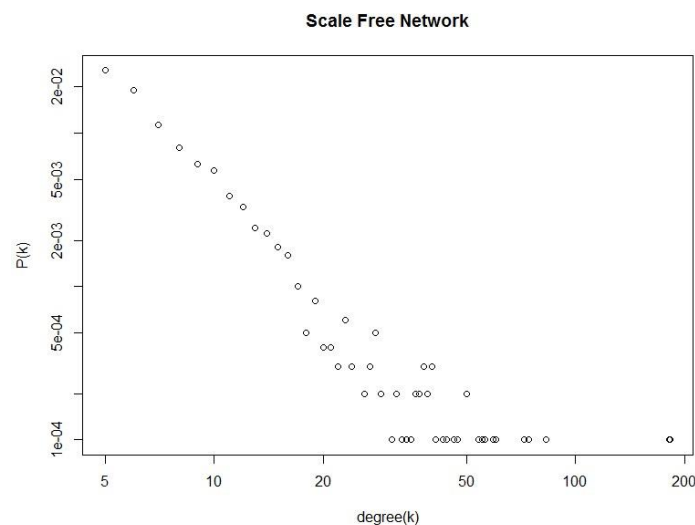
**(10 marks)**

## DEGREE DISTRIBUTION : NETWORKS (15 MARKS)

A network (graph) is defined as a set of nodes connected by a set of edges. The degree ( $k$ ) of a node is defined as the number of edges connected to a node. For example, a social network could be defined by having nodes as people, and the edges link people (nodes) based on whether they are friends. The characteristics of a network are often defined in terms of the distribution of node degree for the entire network. There are many types of networks, but we are interested in what is called scale-free networks (see [https://en.wikipedia.org/wiki/Scale-free\\_network](https://en.wikipedia.org/wiki/Scale-free_network) for additional information), where the probability of a node having degree  $k$ ,  $P(k)$ , is proportional to the degree ( $k$ ) to some constant power  $\gamma$ . Hence these types of networks are defined by  $P(k) \sim k^{-\gamma}$ .

Complete the following steps:

1. Load in the comma separated dataset **network.csv** into R. (*HINT:* `read.csv('network.csv',header=TRUE)`). This dataset is a network with 1000 nodes. The data shows for each node the degree (number of edges) for that node.
2. What is the maximum and minimum degree for this network? **(1 mark)**
3. Produce a **single figure showing 2 plots**: the histogram of the degree, and a line plot **sorted by degree**. Comment on what this tells you about the network. **(2 marks)**
4. Calculate  $P(k)$ , which is the probability of observing a node with degree  $k$ , and produce a **plot of  $P(k)$  versus  $k$**  (as shown below, Figure 1) **using a log scale for both the x and y axes**. In your assignment **include the R code** for calculating  $P(k)$  and producing the plot. **(10 marks)**
5. Explain why the plot from step 4 implies that the network is likely to be scale-free. **(2 marks)**



**Figure 1. A plot of  $P(k)$  versus  $k$  for network.csv**

### DATA QUALITY AND STRUCTURE (20 MARKS)

Examine the supplied script **distplots.r**. This contains a definition for a function **dist.table(...)**, that creates a table (data frame) of two columns: the *normalised distance between all pairs of rows in a dataset and the distance between the corresponding response values for each pair*. If you run **distplot.r** it produces a single plot using generated data for a simple linear model of two explanatory variables.

1. **Explain what the plot represents** and why this might be a useful method for examining the **quality of a dataset** in terms of its *potential to be modelled*.  
(5 marks)
2. Create a dataset where there is NO relationship between the explanatory and response variables and **plot the distance relationship** created by **dist.table(...)**. **Explain why you observe this pattern. INCLUDE THE R CODE TO PRODUCE THE DATASET.**  
(5 marks)
3. **Produce a figure with 2 plots** using **dist.table(...)**: one using the Boston housing dataset (**library(MASS); data(Boston)**; response variable **medv**) and one using the supplied table **bioavailability** dataset (**bioavailability.txt**; **response variable last column** (unlabelled)) . **Explain** which dataset you believe would be easier to model and why. Include a discussion on the relationship between the visualisation created by the function **dist.table** and how it might relate to a **k-nearest neighbour model**.  
(10 marks)

### VISUALISATION AND CLUSTERING (25 MARKS)

This question uses the supplied dataset “**countrystats.csv**”. This file contains the data for 178 countries, showing their population density (people per km<sup>2</sup>), income per capita (\$), purchasing power parity and **change** in gross domestic product as a percentage (%change in GDP) for 2011. You will need to read this data into “R”, change the row names to be the country names, and delete the country names column prior to working with this data.

1. Briefly state the meaning of each variable for the countrystats data. (2 marks)
2. Visualise the countrystats data to examine the relationships between (and within) countries. Which countries appear to be the similar to New Zealand? Which countries are the strongest/weakest? You will need to consider which features (explanatory variables) are important when doing this assessment, and why.

(8 marks)

3. Create two hierarchically clustered dendrograms for the country data **after scaling the data** to have a mean of zero and standard deviation of one (for each variable). Use the agglomeration methods “average” and “complete”. **Produce figures** showing each of these dendrograms, and **explain why they are different**. Finally, using the “complete” linkage dendrogram, cut the dendrogram into 50 clusters (using cutree), and state the countries that are in the same cluster as New Zealand. **Provide the R code** to do this dendrogram/cutree and analysis to get the country names.

(10 marks)

4. Use the dimensionality reduction method **t-SNE** to create a 2 dimensional plot of the country data. **Visualise (plot) the result** and compare the nearest 5 countries from New Zealand using the **t-SNE** mapping to the results in step 3 (HINT: You will need to find out which entry is New Zealand, and “zoom” the plot with text labels). Comment on why they are the same (or different). **Include the plot with your answer.**

(5 marks)

### **SUBMITTING THE ASSIGNMENT**

Ensure that you have answered all of the questions, and that the appropriate R commands and figures have been included in your document.

- Ensure that you have your NAME and STUDENT ID clearly written at the top of your assignment.
- If you use any reference material ensure that the references are included and that citations within the assignment are given at the appropriate place(s).

Submission is via the ASSIGNMENT section of BLACKBOARD. **Please only submit a single PDF file.**