INFO 304 Assignment 3
Data Modelling
DUE DATE: Midnight, Tuesday 8th October
Worth: 15% of final course mark

ELECTRICITY MODELLING (40 MARKS)

This section deals with electricity usage data for 241 households in New Zealand collected in 2009. The data is for a **single 24 hour period**, and has been **sampled every 30 minutes**. Hence there are **48 measures per household**. The first entry is at 12:30am, the second at 1am, and so on. So, for example, the 14th sample is at 7am, and the 24th sample is at midday.

Run the supplied R script **power.R,** which loads the electricity data and produces a boxplot which aggregates the 241 households over time and an example plot for one household. Note that the data is a data.frame (each row is labelled with each household identifier associated with this data), and that to plot an individual example we turned the individual 48 data points into a numeric vector (as.numeric).

1. **Examine the boxplot and briefly discuss the aggregated pattern of electricity usage.**

(5 marks)

The electricity company is interested in understanding some general patterns of usage so that they can identify different types of user and therefore target them with specific pricing structures. ***Your task is to identify 6 (SIX) overall patterns of usage.***

This problem is basically one of clustering – but to solve the problem we will first want to construct explanatory variables that characterise the pattern of usage for each household. Note that some patterns will want to be related to how the electricity usage varies over the day, as well as summary statistics such as the mean/variance/etc. electricity usage.

Examine the R file **power.kmeans.R**. This has some functions already setup that you may find useful. There is even a function that you can call to apply simple functions that take a single input.

2. Construct a table of explanatory variables, where each row is a household, and each column is an explanatory variable that you have constructed from the original usage data. *Describe the functions that you have constructed* to create these explanatory variables, and in particular *discuss how you have described (represented) the temporal structure of usage*. Think about different household electricity usage and how you can capture aspects of this behaviour with different features.

(15 marks)

3. Apply kmeans clustering with 6 centers to your final explanatory datatable. Produce two figures: one with **boxplots** (just like the original script example) showing the final clustering patterns of usage *for each of the six clusters*, the second as **6 line plots** showing the **mean usage pattern for each timestep for each cluster**. See Figures 1 and 2 below for an example (NOTE THIS USES 9 CLUSTERS, so you will have different results).

   Comment on how the patterns of usage vary across the clusters, and whether this aligns with your design regarding different household behaviour (i.e. the types of features that you have used to distinguish different behaviour).

   (15 marks)

4. Select a pair of explanatory variables (say x1 and x2) that are not highly correlated and plot x1 versus x2 for each household **colouring each point by cluster number**. Comment on how the clustering is related to these variables.

   (5 marks)

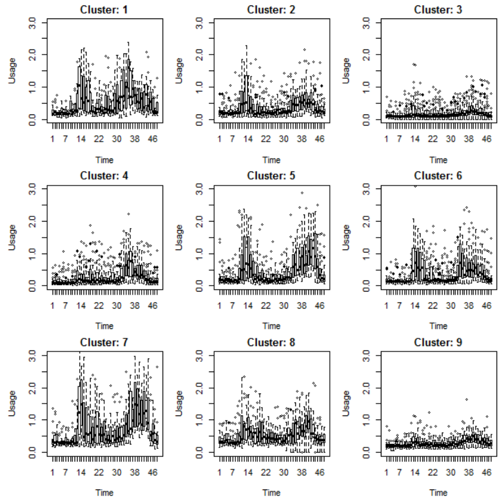


Figure 1. Boxplot of usage for each of the 9 clusters

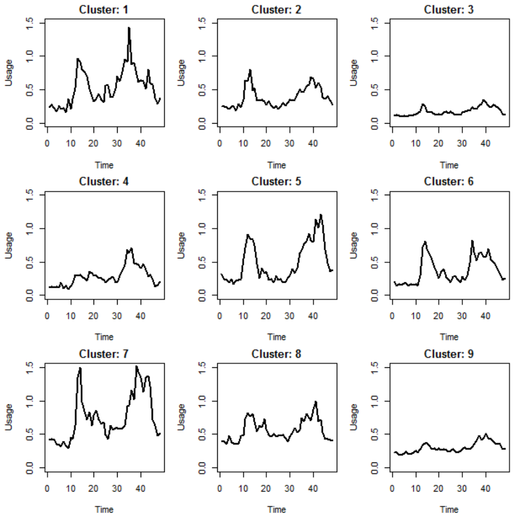


Figure 2. Mean usage for each of the 9 clusters

INVESTMENT PORTOLIO MANAGEMENT (60 MARKS)

This section deals with modelling the selection of stocks, bonds and cash to make up an investment portfolio. Load the data as follows:

```
invest <- read.table("invest.tab")
```

The ROI column is the percentage predicted return on investment, the Risk column is a measure of the risk associated with this particular investment, and the Type indicates the type of investment. Note that each row is labelled with the type of investment and a number, so that you can (if needed) refer to individual investments.

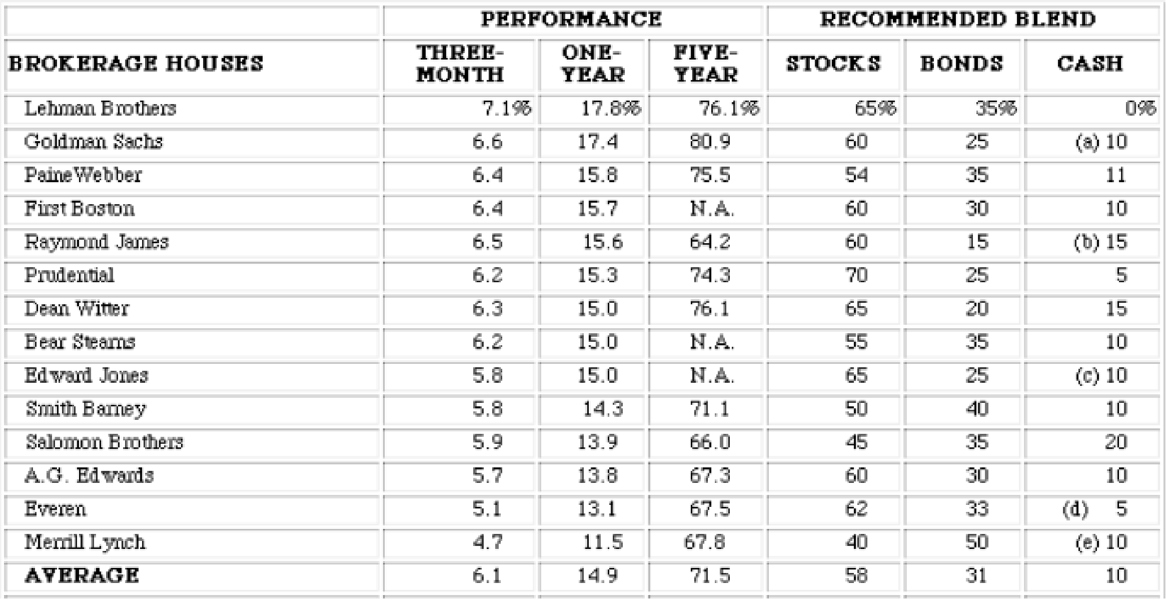
1. **Visualise and discuss the different ROI and Risk associated with Stocks, Bonds and Cash. In addition, load, visualise and give a simple interpretation for the correlation table (**“corr.tab”).

   (10 marks)

2. **Run the “invest.R” script.** This script does a multi-objective criteria analysis to determine the best mix of stocks, bonds and cash over a range of tradeoffs. **Describe in words** what the “invest.R” script is doing. In particular, state how a solution on the pareto front **is represented** and the relationship between the **solution space**, the **objective space** and the **constraints**. Include in your description **a histogram** showing the number of investments for each solution on the pareto front.

   (10 marks)

The table below shows the recommended blend of stocks, bonds and cash for a number of different brokerage houses (i.e. businesses that take your money and invest it to give you a return).



	PERFORMANCE			RECOMMENDED BLEND		
BROKERAGE HOUSES	THREE-MONTH	ONE-YEAR	FIVE-YEAR	STOCKS	BONDS	CASH
Lehman Brothers	7.1%	17.8%	76.1%	65%	35%	0%
Goldman Sachs	6.6	17.4	80.9	60	25	(a) 10
PaineWebber	6.4	15.8	75.5	54	35	11
First Boston	6.4	15.7	N.A.	60	30	10
Raymond James	6.5	15.6	64.2	60	15	(b) 15
Prudential	6.2	15.3	74.3	70	25	5
Dean Witter	6.3	15.0	76.1	65	20	15
Bear Stearns	6.2	15.0	N.A.	55	35	10
Edward Jones	5.8	15.0	N.A.	65	25	(c) 10
Smith Barney	5.8	14.3	71.1	50	40	10
Salomon Brothers	5.9	13.9	66.0	45	35	20
A.G. Edwards	5.7	13.8	67.3	60	30	10
Everen	5.1	13.1	67.5	62	33	(d) 5
Merrill Lynch	4.7	11.5	67.8	40	50	(e) 10
AVERAGE	6.1	14.9	71.5	58	31	10

Table 1. Major brokerage house investment blends (1996). From: http://pages.stern.nyu.edu/~adamodar/New_Home_Page/invphillectures/port.html

3. Using the result of the nsga2 model you have previously run, examine and present the blend of stocks, bonds and cash for a **low risk**, **moderate risk and high risk investment blend (just pick one from each general category)**. *Discuss, in relation to Table 1,* the level of risk that seems to be taken by the brokerage houses and whether the one year return performance is related to the associated risk of the brokerage house.

(10 marks)

4. **Examine the plot shown in Figure 3.** This shows how the percentage of bonds, stocks and cash vary as you move along the pareto front from the least to greatest percentage return. **Outline the approach** (set of steps, algorithm, pseudo code, or R code…) that would be required to produce this figure given the output from nsga2.

(10 marks)

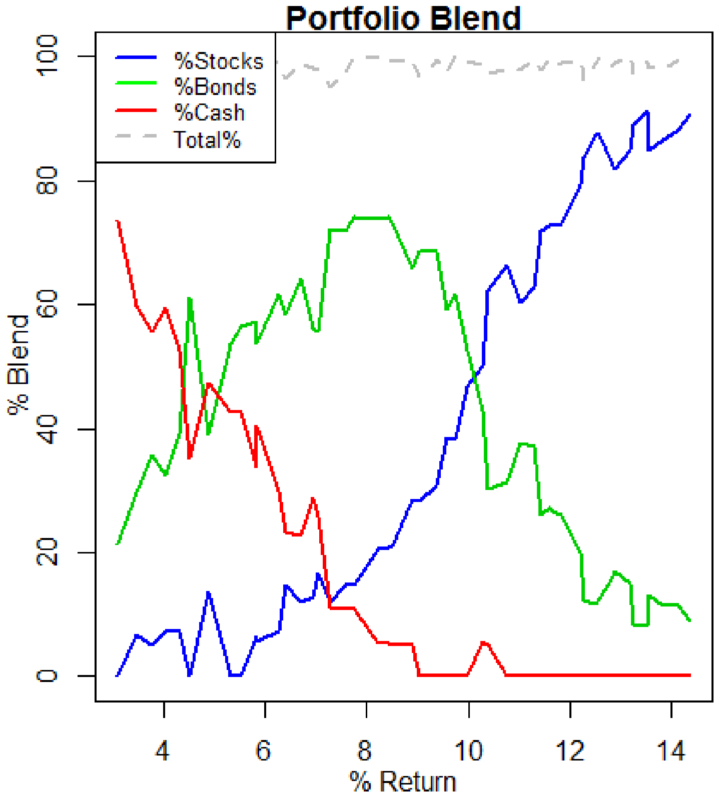


Figure 3. %Return versus investment blend as a percentage across the pareto front.

5. A matrix representing the estimated correlation between any two investments in terms of their return behaviour is given in the table **corr.tab**. For our example with 30 investments this is a 30x30 table. A good portfolio should aim for investments where correlation of behaviour is minimised, so that if some of the investments are decreasing, others in the portfolio are increasing. This reduces the risk of the portfolio as a whole losing money and results in more stable returns.

   Extend the model in **invest.R** to have an **additional objective** which is to *minimise the correlation for the investment blend*. *Describe in words* the approach you have chosen and explain the rational for the model. *Include your R code* for the function that calculates this objective. Produce a plot showing the objectives correlation versus return (x axis correlation, y axis return) for the pareto solutions, and discuss what type of return is likely to be best for minimising correlation (and probably therefore the safest combination to consider).

(15 marks)

6. Find out some historical information on the performance of **2 of the riskiest brokerage houses** from **Table 1** and comment on whether they still exist, etc.

(5 marks)