# Homework 1

BUAN 6356

**Read the instructions below before you start your analysis.**

1. Create an R Markdown document to prepare your answers. You should upload **two (2)** files on eLearning: (i) an **.RMD** file; and (ii) a **.PDF** file that is generated using "knit" in the .RMD file. Both of these files should contain the required R code, R tables and charts, and all the required explanations and answers to the questions in the homework.

2. Include your name at the top of the file.

3. **DO NOT** use an absolute directory path. I should be able to "knit" your R Markdown document to an .html/.pdf document without trying to find the input data in another directory. Test the "knit" process before uploading files on eLearning. Assume that I have the .csv file(s) mentioned below.

4. **DO NOT** change the dataset name before importing it into R. If you rename the dataset or any variable(s), use your R script to do that.

5. Label the charts appropriately. I should be able to figure out what information a chart is providing by looking at the chart title and its labels.

6. Any assignment submitted after the deadline will be considered late and will not be graded.

Homework 1

The "Utilities" dataset includes information on 22 public utility companies in the US. The variable definitions are provided below.

Fixed_charge = fixed-charge covering ratio (income/debt)

RoR = rate of return on capital

Cost = cost per kilowatt capacity in place

Load_factor = annual load factor

Demand_growth = peak kilowatthour demand growth from 1974 to 1975

Sales = sales (kilowatthour use per year)

Nuclear = percent nuclear

Fuel_Cost = total fuel costs (cents per kilowatthour)

For **Questions 1-4** below, do not scale the data.

1. Compute the minimum, maximum, mean, median, and standard deviation for each of the numeric variables using data.table package. Which variable(s) has the largest variability? Explain your answer.

2. Create boxplots for each of the numeric variables. Are there any extreme values for any of the variables? Which ones? Explain your answer.

3. Create a heatmap for the numeric variables. Discuss any interesting trend you see in this chart.

4. Run principal component analysis using *unscaled numeric variables* in the dataset. How do you interpret the results from this model?

5. Next, run principal component model after *scaling the numeric variables*. Did the results/interpretations change? How so? Explain your answers.