# Regression Modelling Project

## Project Overview

In this project you will analyse and explore a large data set based on the daily closing stock prices of 31 large companies. In particular, you will build and assess linear regression models that explain variability in the daily stock price of the company Google using stock price data from other companies.

You will be prompted to explore the project during the weekly lab sessions, but are also expected to work on both the practical analysis and written report outside of class times. You may use the Friday Drop in sessions in SW601a for this purpose.

## Getting Started

The data consist of time series of the closing prices of 26 different stocks on the New York Stock Exchange (NYSE) between 4th January 2010 and 16th December 2015. Note that with the exception of Google's closing prices which are in cents, the other 25 closing stock prices have been scaled, centred and detrended and therefore no longer have measurement units. The `nyse` data are available to download from the class MyPlace page and are stored in `.RData` format. To open the data in R, either double click the `nyse.RData` file or set the working directory to the data location and type the following:

```r
# load the New York Stock Exchange data set
load("nyse.RData")
```

The data will now be accessible in a data frame called `nyse`. You can quickly visualise the data columns by printing the first few rows of data using the `head()` function

```r
# Print first rows of the NYSE data
head(nyse)
```

Full descriptions of the column contents are provided on the final page of this document.

## Project tasks

### Part 1: Model fitting and interpretation (25 marks)

(i) Calculate the sample correlation coefficient between the closing price of each stock and that of Google and use a single plot to summarise these.

(ii) Find the 5 companies whose stock prices have the strongest correlation (in absolute value) with Google's. Fit and briefly interpret a linear regression for Google's stock price using these five companies prices as independent variables.

(iii) Using an appropriate variable selection technique and any transformations of the independent variables, build an improved model for Google's daily closing price.

(iv) Using your final model from part (iii), check the regression assumptions using appropriate summary plots, and comment on whether you think that these are valid.

**Part 2: Prediction and validation** **(10 marks)**

(i) Write an R function that uses your model from Part 1 to make future predictions for Google's daily closing price for new rows of independent variables. Check the prediction accuracy by uploading it to the class leader board. This will go live during the lecture on *Monday 12th November*.

(ii) Try to improve the prediction accuracy of your model by trying different combinations of independent variables. Assess the models using cross-validation to ensure that your improved model is optimal for prediction.

**Presentation of report, clarity of language** **(5 marks)**

**Total** **(40 marks)**

## Report Structure, content and submission

The report itself does not need to follow any particular structure. For example, separate sections that briefly document the work of *Part 1* and *Part 2* are fine. However, you should ensure that any analysis you carry out is clearly interpreted using full sentences. You should write as though your audience were your statistics classmates, but that they were not familiar with the data. Background knowledge of finance is not required, although it may help in building and criticising your models to consider basic factors taht may contribute to the day-to-day variability in the value of a company.

- The report should have a cover page with your name and student ID clearly marked
- The report should be a maximum of 4 pages in length including graphics and tables, but excluding the cover page
- Graphs should be suitably labelled, sensibly scaled and cropped
- Numerical R outputs should be neatly presented in tables or presented in the text
- **Please submit your reports to myplace by midday on Monday 3rd December. Late submissions will be penalised!**
- In addition please e-mail the R script containing commands to reproduce the analysis to me (**kate.pyper@strath.ac.uk**)

## Data description

The columns in the `nyse` data are summarised briefly below. Note that the data for all of the closing stock prices have been centred, scales and deternded to avoid strong collinearity and therefore no longer have any units. Data for Google is unmodified, and is measured in US Cents (¢)

| Column number | Column name | Description |
|---|---|---|
| 1 | Date | Date of closing prices in YYYY-mm-dd format |
| 2 | Year | Year YYYY format $\in \{2010, \ldots, 2016\}$ |
| 3 | Month | Month mm format $\in \{1, \ldots, 12\}$ |
| 4 | Weekday | Day of the week $\in \{\text{Monday}, \ldots, \text{Friday}\}$ |
| 5 | GOOGL | Google |
| 6 | AAPL | Apple |
| 7 | AMZN | Amazon |
| 8 | AZN | Astrazeneca |
| 9 | BP | British Petroleum |
| 10 | C | Citigroup |
| 11 | CDE | Couer Mining |
| 12 | DAL | Delta Airlines |
| 13 | DPZ | Domino's Pizza |
| 14 | F | Ford |
| 15 | GIL | Gidlan Activewear |
| 16 | JPM | JP Morgan |
| 17 | K | Kellogs |
| 18 | KO | Coca-Cola company |
| 19 | M | Macy's Inc |
| 20 | MSFT | Microsoft |
| 21 | NOK | Nokia |
| 22 | PG | Procter & Gamble |
| 23 | RBS | Royal Bank of Scotland |
| 24 | SAM | Boston Beer Company |
| 25 | SPGI | S&P Global Inc |
| 26 | T | AT & T |
| 27 | V | Visa Inc. |
| 28 | WMT | Walmart |
| 29 | WHR | Whirlpool Inc. |
| 30 | XIN | Xinyuan Real Estate Co. Ltd. |