Pre-requisite Knowledge

0.1 Wage, Stock and Gene Data

Before delving into the hardcore statistical learning we must first establish important datasets that will be used in the examples going forward. The three datasets introduced in the book are:

- Wage Data

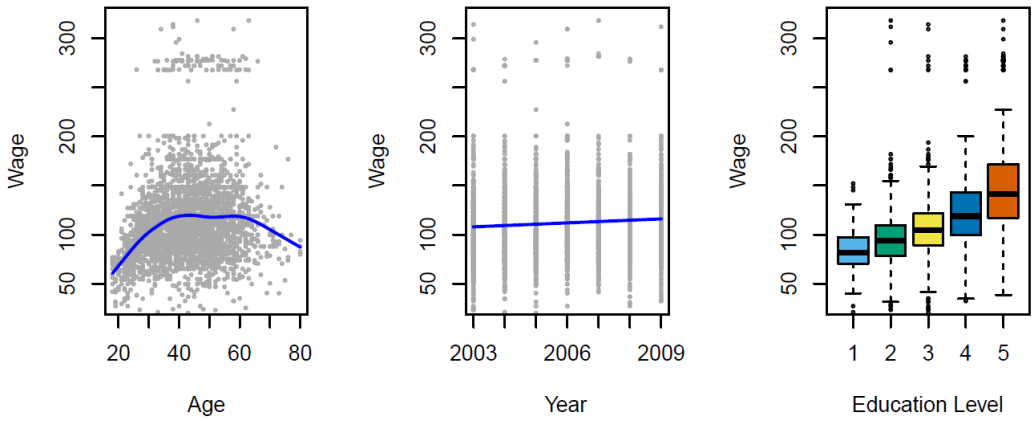

Figure 1: Wage Data

The wage data includes the association of employee wages with their age, education level and the calendar year. The different data points are plotted on the graph and the blue curve represents the average.
Wages increase with employee age up to the 40 years old then stays constant till 60 years after which it decreases. A linear relationship is seen between wages and calendar year. Employees with higher education level (level 5) are paid considerably higher than the employees with lower education level.
In this sort of data we are expected to predict employee wages combining all of these factors. The output here will be a numerical value which is continuous. This type of problem is solved by regression. We shall discuss linear regression later on.

- Stock Data

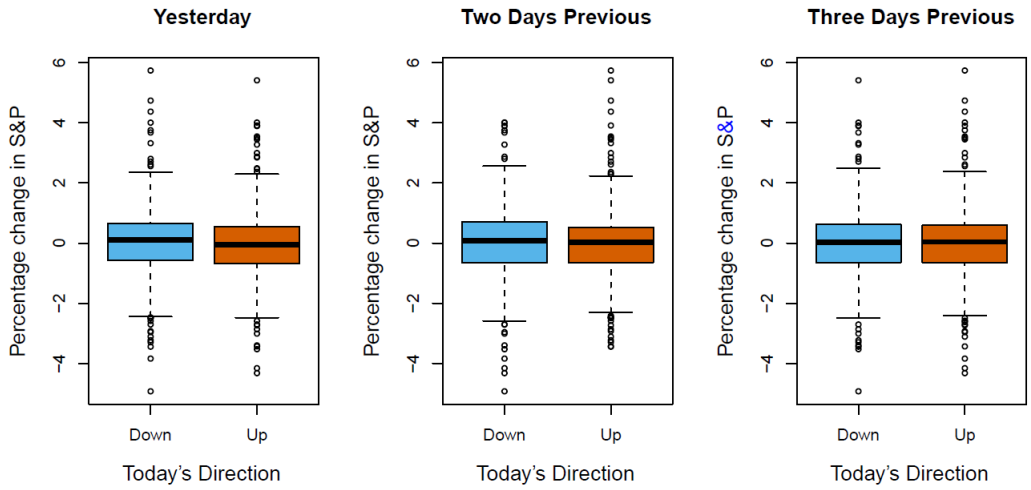

Figure 2: Stock Data

Stock data is difficult to analyse. Evidence of this fact can be found in the above figure, the boxplots given have no clear relation between them, the prices of yesterday do not necessarily dictate the prices of today. The statistical methods applied on this dataset is more categorical or qualitative. We do not need to find a numerical output value instead we need to label whether the price of a stock will go UP or DOWN. This is a standard classification problem. We will later learn about methods that predict the direction of stock movement based on the weak trends in the dataset.

- Gene Expression Data

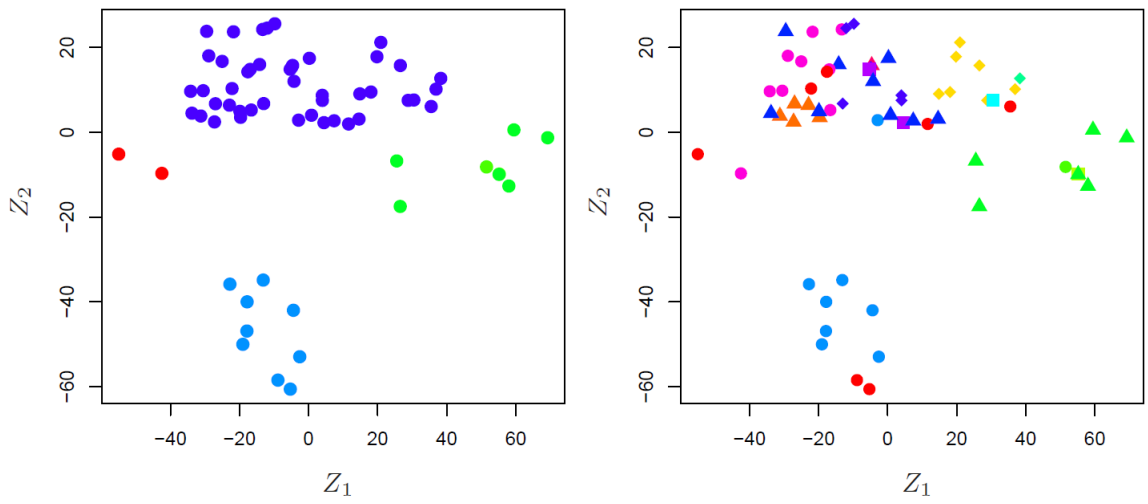

Figure 3: Gene Expression Data

Wage data introduced us to problems solved by regression, stock data analysis is a classification problems and Gene Expression data problems are of the clustering type. For example, if you want to figure out how different types of customers are

similar to each other based on their observed characteristics, you group them and perform analysis.
The resource considers the NCI60 dataset which has 6830 gene expression measurements for each of 64 cancer cell lines. Since this is a lot of data to visualize we perform a dimensional reduction on the data by only plotting two numbers Z_1 and Z_2, which are also called the principal components. Now this data can be used to find evidence of clustering.