# Part 1 – Data Preparation and Pre-processing

**Dataset Description:**

The dataset is composed of data taken from SEEK job market and is composed of a CSV file just under 900MB large. Within the CSV file there are 12 columns (excluding the ID) which carry a range of metadata about the job listing. This information includes the company, date, classification, requirements, salary and the location of the job. The data ranges over the span from October 2018 until March 2019 with 318,477 entries.

There are 3 main components of the job which will be studied over the time period, this includes classification/subclassification, location and salary (lowest/highest). Therefore the relevant columns will be Date, Location, Classification, SubClassification, Lowest Salary and Highest Salary. Optional columns to extract information from include the Company, Area, and JobType. Due to the wide variation in formatting with columns such as Title, Requirements and FullDescription, these columns will be discarded along with ID.

**Dataset Preparation and Pre-processing:**

The first step was to load in the dataset into a data frame via the pd.read\_csv() function. After this the ID was dropped and the data set was scanned for duplicate listings, of which 8607 where found. Following this the duplicate listings were dropped from the table using df.drop\_duplicates(). Additionally, the Title, Requirement and FullDescription columns were dropped as these are not useful for analysis due to the variety of formatting.

Following this the dataset was checked for null values. Any rows where a null value appeared in; Date, LowestSalary or HighestSalary were also dropped to prevent null values from interfering with calculations. Two extra columns were added, ‘AverageSalary’ and ‘RangeSalary’ which are the average and range of the highest and lowest salary values.

The date value initially was an object, as was determined upon inspection with df.dtypes. By using pd.to\_datetime() it was possible to normalize that object to a datetime64 object for time series analysis. This date was then used as the index for the graph df.set\_index(‘date’).

**Hypothesis of analysis outcome:**

It is expected that the highest paying and jobs will revolve around the field of IT and Health. The most abundant jobs but at a lower pay rate will be in the domain of retail. The major cities such as Melbourne, Sydney and Brisbane will all have the highest average pay rate in addition to the most job listings. It is hypothesised that the average salary between all job listings will be around $30/40 an hour. It is also predicted that one of the supermarket chains (Coles/Woolworths/ALDI) will have the most job listings out of any company.

# Part 2 – Data Analysis and Interpretation