# **Part 1 – Data Preparation and Pre-processing [8 points]**

## 1) Describe the dataset.

### For example:

### What are the categories/domains of the dataset? Id, Title, Company, Date, Location, Area, Classification, Sub-classification, Requirement, Full Description, Lowest Salary, Highest Salary, Job Type.

### What is the dataset size of each variation?

### What is dataset structure/format?

### What are attributes/features of review data you are going to use?

### What are attributes/features of product data you are going to use?

### Which parts of the dataset will you use or all of them?

### [1-2 paragraphs, 3 points]

This project will explore and analyse SEEK job market data. The dataset contains 318,477 data entries. It includes 13 categories: Id, Title, Company, Date, Location, Area, Classification, Sub-classification, Requirement, Full Description, Lowest Salary, Highest Salary, and Job Type. All categories were formatted as a string except for the *HighestSalary* and *LowestSalary* categories which were formatted as integers.

Some parts of the dataset were excluded and others parts appended in order to improve analysis. Firstly, the *Id* and *Full Description* columns were removed because they provided no useful information for analysis and created problems with duplicate data. Removing the *Id* column revealed 8,607 duplicate entries and removing the *Full Description* revealed 1384. After deleting these duplicates, the total data entries was lowered to 308,486. Additionally, an *AvgSalary* column was appended. The column values are derived from the mean of the *HighestSalary* and *LowestSalary* columns. It was added to allow for an easier way to rank job salaries later in analysis.

## 2) Describe the steps you used for data preparation and pre-processing.

### For example:

### How do you load the data using Pandas?

### How do you normalize the data?

### How do you clean the data?

### [2-3 paragraphs, 4 points]

Data preparation and pre-processing has been done in a Jupyter notebook using the Pandas library. Because the dataset has been formatted as a CSV, the Pandas *read\_csv( <filename> )* function has been used. [Mentioning low\_memory=False flag because had errors maybe?] From this, we could read the CSV file into a Pandas DataFrame. The name of the DataFrame is *df* as it is short and an easy way of referring to the data.

Cleaning:

* Looked at the heads to get a picture of the data.
* Looked at the data types. Saw that we could make the *Date* category a bit better so converted it to a datatime.
* Checked for NULL values
* Discovered Id and Full Description created duplicates and weren’t useful -> Deleted them
* Removed duplicates.
* Looking at jobs with the maximum HighestSalary (999) and minimum LowestSalary (0)
* Changed lowest salary to be 10k because It was a fair assumption. (Might change the other one to 400k?)
* Changed our salary to be from 1k -> 1000 as it was easier to understand that way.
* Removing Area from titles to make it clearer to read.

Normalising:

* The data was normalised.
* We normalise because it makes it easier to represent and graph data later on (I think?)
* To do this we used the sci-kit lean pre-processing library StandardScaler.
* We normalised our three integer type categories LowestSalary, HighestSalary, AvgSalary.
* To do this we defined a StandardScaler, then fit and transformed the columns. This was done into a new dataframe called df\_norm. It is a normalised version of df.

## 3) What is your hypothesis (expectation) about the analysis outcome?

### [1-2 paragraphs, 1 point]

Hypothesis:

* What cities might have the most well-paying salaries?
* What are the hottest job sectors in each city?
* Will there be more jobs posted at the beginning of the month?
* Which sectors will keep the highest market share
* (e.g. stuff we will be exploring in data analysis phase)