# **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 other parts appended to improve analysis. Firstly, it was found that all Classification and Location NaN records were shared, and so these values were removed. This decreased the total data entries to 197,229. In addition, rows with a Sub-Classification value of ‘Other’ was removed. This is because ‘Other’ does not tell us anything useful about the data. It decreased the total data entries to 184,841. After this, the Id and Full Description columns were removed because they provided no useful information for analysis and created problems with duplicate data. The *Id* and *Full Description* revealed 1,039 duplicate entries. After deleting these duplicates, the total data entries were lowered to 183,802. 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 the 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 was 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. It should be noted that a *low\_memory* flag was used to fix importing issues caused by columns with mixed types. 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.

Data cleaning was an important phase of the data preparation process. Firstly, we needed to get a big picture of the data by displaying its head. From this, we looked at how the data was formatted. It was noticed that the Dates column was formatted as an object type and was converted into a DateTime type. This was done to allow for easier manipulation of the dates. Next, we checked for NULL values. This gave us a clearer picture of the parts of the data that was missing. Fortunately, there were no NULL values for the LowestSalary and HighestSalary columns and so nothing was dropped. Additionally, it was found that 19.53% of jobs had a LowestSalary value of 0k, and 4.35% had a HighestSalary value of 999k (although none had both a LowestSalary of 0k and HighestSalary of 999k). These salaries appeared to be anomalies as no reasonable salary pays no nothing, nor do so many pay exactly 999k. In addition, the SEEK website allows job searches from 30k to 200k+. As such, to keep closer to these amounts 0k *LowestSalary* rows were converted to 15k and 999k *HighestSalary* rows were converted to 300k. Accounting for these anomalies will help reduce inaccurate skewing of the data. Moreover, salary values were scaled from 1 to 1,000. This allowed us to interpret the data as $1000 instead of 1k which was found to be clearer and less ambiguous. Furthermore, duplicate data removed. At first, there appeared to be no duplicate data. However, upon removing redundant Id and Full Description columns, 1039 duplicate entries were discovered and then removed. Lastly, to make the Title columns easier to read, any matching Area column that was found inside the title was removed. This affected 8,011 rows.

Normalization was also performed during data preparation. Normalisation is important as it allows us to standardise the metric data. To do this, the Sci-kit learn pre-processing library *StandardScaler* has been used. *StandardScaler* imposes a mean of 0 and a standard deviation of 1 onto the data. In application, we declared a StandardScaler object and then used it to fit and transformed our metric columns. This was then saved into a DataFrame called *df\_norm* – a clone of *df* but with normalized metric values.

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

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

Some predictions have been made regarding the analysis outcomes:

* Firstly, it is believed that high population cities will have the largest quantity of job postings. This is because with people come the demand for more jobs. It would also not make sense for a small city to have more job postings than its population.
* Additionally, it is believed that the most job postings will be in July. This is because July is the beginning of the new financial year in Australia whereby companies will have finalised their tax and hence may be more willing to hire new staff.
* Furthermore, it is believed that Healthcare & Medical classification listings will have the highest average salary. This is because medical roles such as doctors and surgeons are typically highly paid.
* Moreover, it is believed that IT classifications will be the most commonly listed. This is because IT is one of today’s most rapidly growing industries and hence would be expected to have significant demand.
* Lastly, it is believed that within the IT field the most in-demand skill will be Python. This is because Python has become one of the most popular programming languages and is used frequently in universities.

# **Part 2 – Data Analysis and Interpretation [17 points]**

## 1) Job metadata

### [1-2 paragraphs, 5 points]

## The job metadata was explored to better understand the data. Firstly, a summary of the data was collected. This showed that the mean Average Salary is $78,679. It also showed that the minimum Lowest Salary is $15,000 and the maximum Highest Salary is $300,000. Additionally, the salary ranges of the data were taken. It showed that there were 11 unique salary ranges: 15-30k, 30-40k, 40-50k, 50-60k, 60-70k, 70-80k, 80-100k, 100-120k, 120-150k, 150-200k, and 200-300k.

## Furthermore, it was seen there were 297,288 rows in the dataset. Of these rows, there were 66 unique locations. The frequency of listings for these locations appeared exponential. Here high population cities such as Sydney, Melbourne, and Brisbane had the largest number of job listings whereas smaller population cities like Darwin had significantly fewer. Additionally, there were 31 unique classifications. The frequency of listings for these classifications appeared linear. Here IT classifications had the largest frequency and self-employment classifications had the lowest frequency. Moreover, there existed 337 unique sub-classifications. The frequency of these appeared linear and the most popular of the listings was the management sub-classification. The spread of sub-classifications per classification was also examined. This also appeared linear. The classification with the most sub-classifications was Healthcare & Medical with 32 sub-classifications.

## 2) Locations

### [1-2 paragraphs, 4 points]

…

## 3) Classifications

### [1-2 paragraphs, 4 points]

The various job classifications and sub-classifications were analysed to outline trends in the data. Firstly, it was found that Information & Communication Technology (IT) jobs held the highest market share, being 11.73% of the job listings. This is was followed by Trades & Services (9.2%) and Healthcare & Medical with (8.75%). Some of the lowest are CEO & General Management (0.41%), Advertising, Arts & Media (0.39%) and Self Employment (0.05%). Overall, the distribution of market share appeared to be linearly spread. In contrast, it appeared the classification with the highest average salary was CEO & General Management, having an average salary of $158,055. This was followed by IT ($131,041) and Consulting & Strategy ($122,655). The lowest average salary was held by Hospitality & Tourism ($41,1374).

From this, IT classifications were specifically examined. It was found that the most common IT sub-classification was Developers/Programmers. This was followed by Business/Systems Analysts and Programme & Project Management. The fewest listings were held by Hardware Engineering, Technical Writing, and then Computer Operators. The highest paying IT jobs were Architects, earning an average salary of $168,198. Following this was Programme & Project Management ($156,651) and Security ($141,269). Developers/Programmers job listings had a midrange average salary of $126,880. Help Desk & IT Support had the lowest average salary of $61,622. Finally, a set of popular technologies were queried against the requirements of the IT job listings. It found that Java was the most mentioned technology, mentioned by 353 unique classifications. This was followed by AWS (248 mentions), and SQL (215 mentions). Python had midrange mentions of 101. Rust, Matlab, Objective-C, and C# did not have any mentions.

# **Part 3 – Evaluation [5 points]**

## 1) What are the findings of your data analytics?

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

## 2) What actions for balancing the markets do you suggest based on your findings?

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

? – sent email to Henry asking what this means.

## 3) How could you refine your data analytics?

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

There are a variety of refinements that could be made to improve the quality of the data analysis. Firstly, it would have been beneficial to have used SEEK data from a wider range of time. The dataset provided only included data from the end of 2018 and the start of 2019 and misses half of the months (April, May, June, July, August, September). Since July is not included, it was impossible to see if the Australian end of the financial year impacted the number of job listings. In addition to this, the 2019 data contained Classifications and Location columns with NaN values and thus could not contribute to a large part of the analysis. With a more robust and wider-ranging dataset it would have been possible to analyse the market trends over time. For example, it might have allowed us to investigate the popularity of certain IT technologies over time, illustrating if some are increasing or some are decreasing in popularity (perhaps Java is slowly decreasing, but Python is quickly increasing?).

While analysing the most popular IT technologies, it would have been an improvement to use a separate and more comprehensive dataset containing the names of technologies. In the analysis a self-composed list of 30 technologies was used. Because it was self-composed and small in size there is a possibility that some technologies have been left out. If this is the case, the data may be misleading as to what technologies should be studied. Additionally, the method of finding these keywords could be improved. There are some cases where technologies may be spelt differently such as “Objective-C” and “Objective C” or “SQL” and “MySQL”. To improve this, more advanced Natural Language Tool Kit (NLTK) processing should be implemented such as stemming.

## 4) Are there any implications for employers and employees based on the findings you obtained? Justify your answer.

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