Contents

[Part 1 – Data Preparation and Pre-processing 1](#_Toc41075917)

[Dataset Description: 1](#_Toc41075918)

[Dataset Preparation and Pre-processing: 1](#_Toc41075919)

[Hypothesis of analysis outcome: 1](#_Toc41075920)

[Part 2 – Data Analysis and Interpretation 1](#_Toc41075921)

[Job Metadata: 1](#_Toc41075922)

[Salary Distribution 1](#_Toc41075923)

[Average Salary over Time 3](#_Toc41075924)

[Number of Job Listings Over Time 4](#_Toc41075925)

# 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:

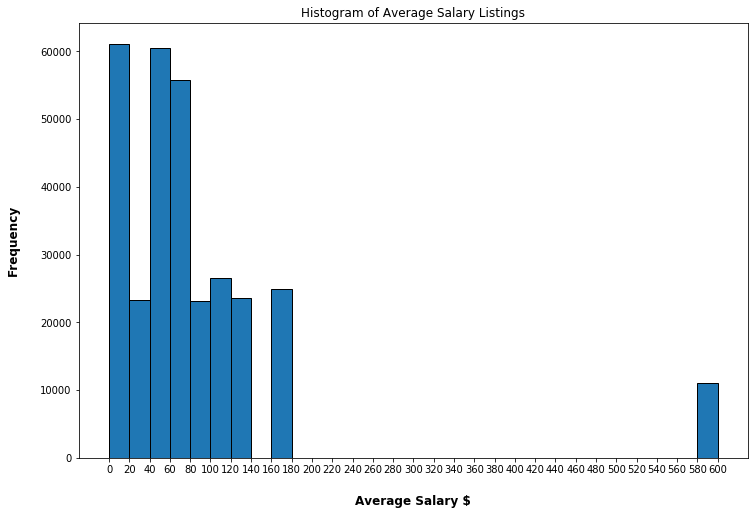
It is expected that the highest paying and jobs will revolve around the field of IT and Health. The most abundant jobs but at a lower pay rate will be in the domain of retail. The major cities such as Melbourne, Sydney and Brisbane will all have the highest average pay rate in addition to the most job 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

## Job Metadata:

### Salary Distribution

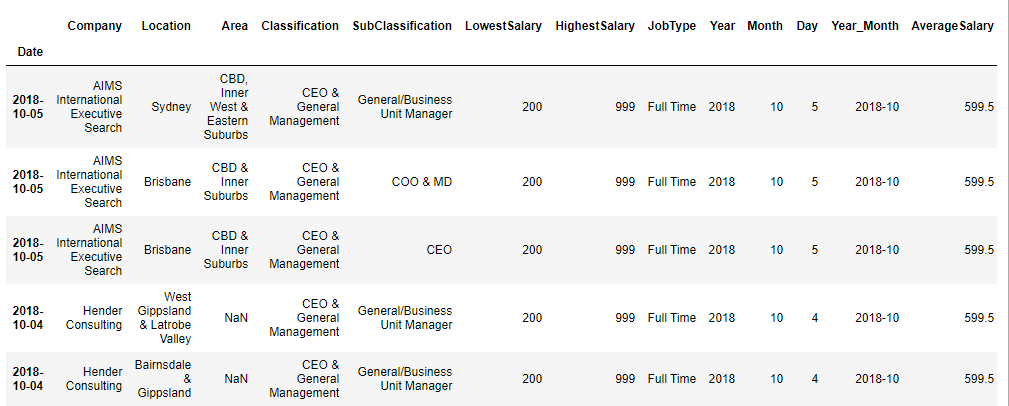
The first analysis of the Job metadata was to create a histogram of the job salaries to analyse the distribution of the salaries. As expected, lower salaries are much more common than higher salaries with the 3 most common bins by a large amount being 0-20, 40-60 and 60-80. This is to be expected however given the typical hierarchical structure of jobs with the majority of jobs being low to mid paying.



**Figure 1: Histogram of Average Salary for all Listings**

Outliers:

However, an anomaly in this histogram which is apparent is that there is a high number of outlier jobs paying $580 to $600. Using exploratory data analysis this bin will be analysed to determine what the root cause of this anomaly is.

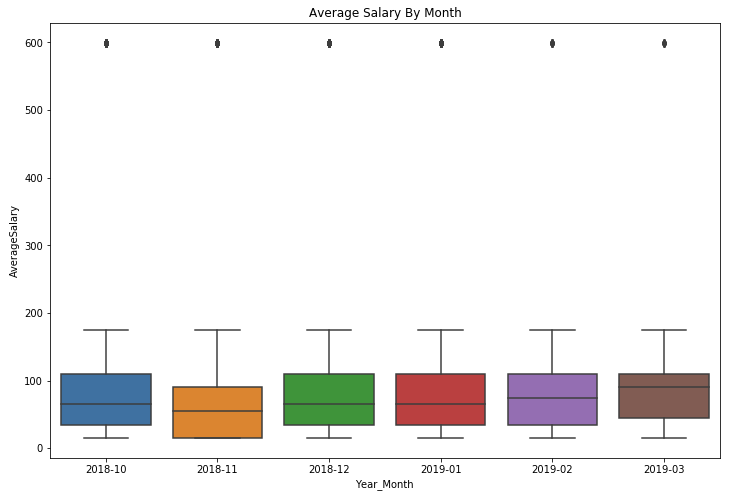


**Figure 2: First 5 Outliers where the Average Salary is 580-600**

Analysis of the outliers appear to reveal that all if them contain an average salary of $599.5. This is an average between 200 and 999, which is the low and high salary for all these jobs. The predicted reason for this same salary is due to the fact that all of these jobs are listed at the highest possible price range allowed on SEEK.

Out of these outlier jobs; ICT, Healthcare, Construction, Mining and Government were the highest paying suggesting that these jobs may be the highest paying, however this will be explored further in the data.

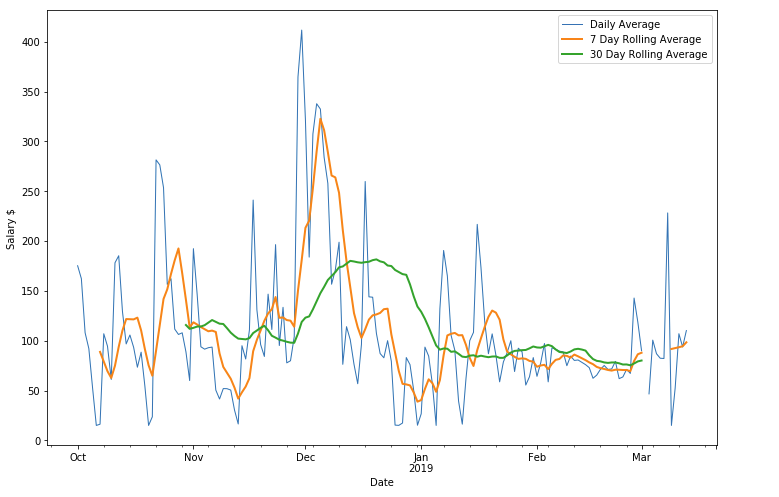
### Average Salary over Time



**Figure 3: Average Salary by Grouped by Month Boxplot**

The boxplot above for the average salary by month reveals that while salary is fairly stable, there was a sharp drop in November compared to other months. The predicted cause of this is that Christmas casual listings in retail spiked in November, causing the average salary for that month to be lower in comparison to the other months.

However, this boxplot fails to demonstrate another feature with the data which is present in the time series graph below (see Figure 4); that there was a huge spike in the average salary around early December.



**Figure 4: Time-series Graph for Salary**

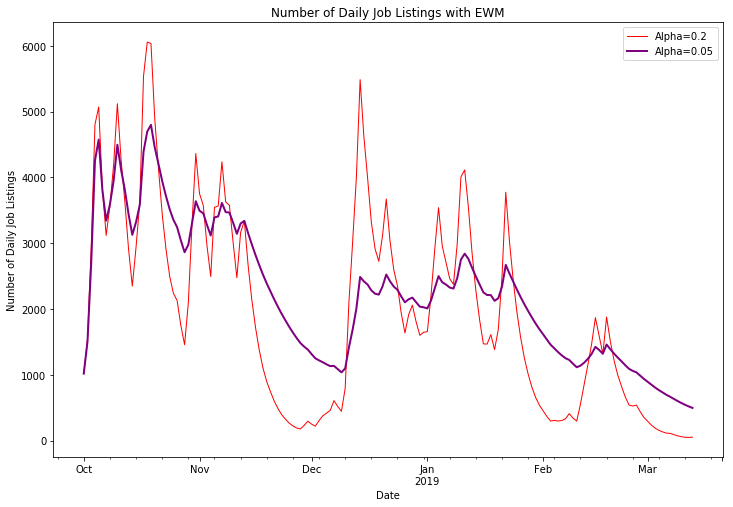
The large spike in early December is also evident on the rolling averages, however slightly delayed and less prominent. This is because rolling averages use an average of the past n-days, meaning that the data will be slower to respond to change.

The hypothesised cause for the spike in early December is people leaving high pay, high stress jobs in order to spend time with their family over Christmas. This results in a greater demand for people in these high paying positions leading to more SEEK listings and a spike in the average salary.

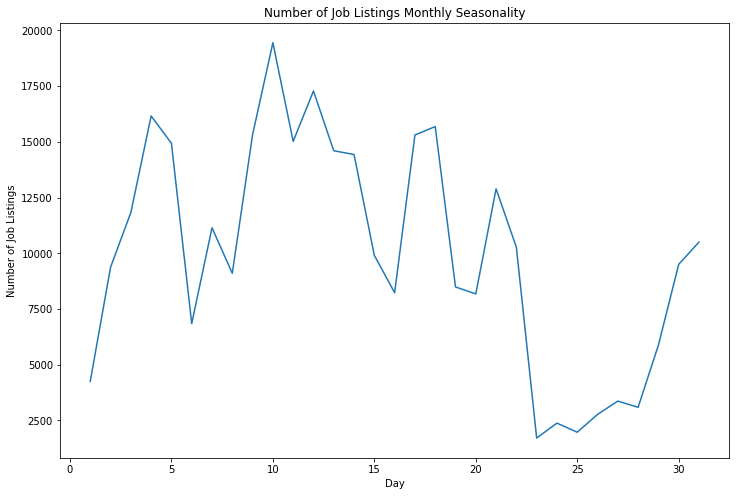
### Number of Job Listings Time Analysis

Analysing the daily job listings over time reveals a trend downwards. This is revealed in Figure 5 below which demonstrates the average number of daily job listings with exponential smoothing implemented, however this trend downwards may just be because of seasonality. Late November/December reveals a dip in the number of job listings which then spiked again around new year before dipping again.

As for the monthly pattern of posting, it appears that the majority of jobs are posted in the middle of the month (see Figure 6).

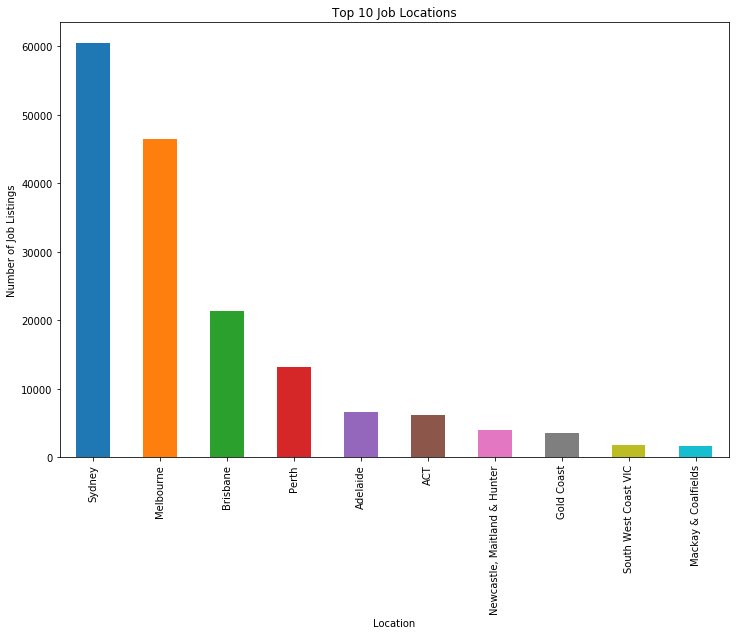


**Figure 5: Number of Daily Job Listings with Exponential Weighted Smoothing**



**Figure 6: Number of Job Listings Grouped by Day of the Month**

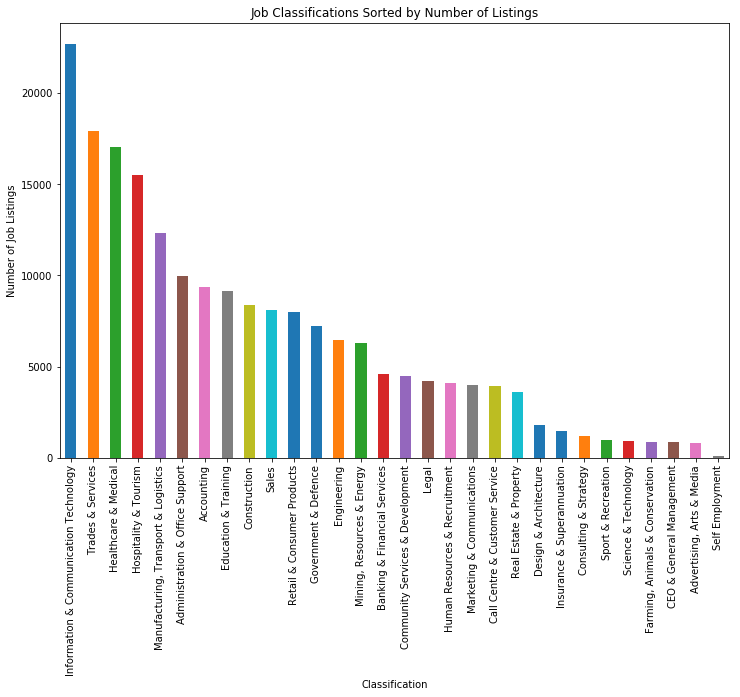
## Market by Locations



**Figure 7: Top 10 Job Locations**

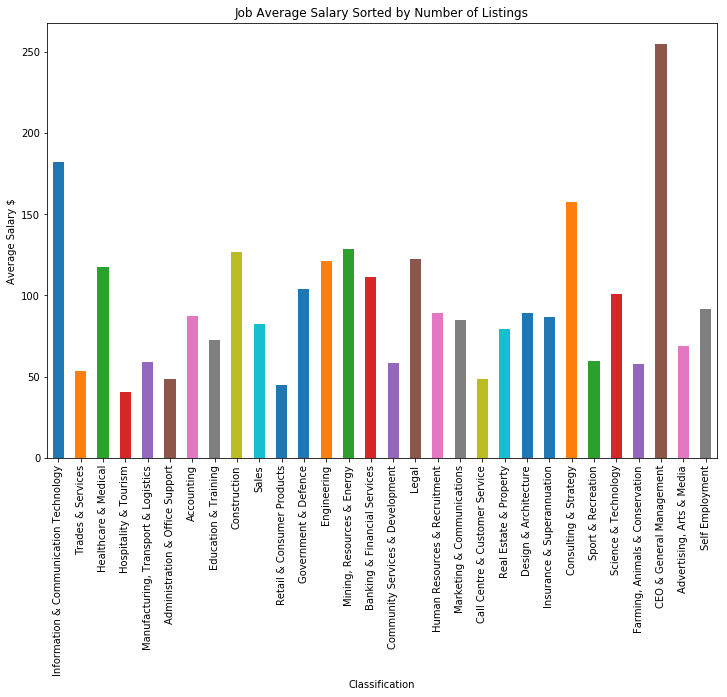
As is expected,

## Market by Sectors



**Figure 8: Number of Job Listings by Sector**

Analysis of the above bar chart reveals the sectors that are in high demand. The largest sector by some margin is ICT with 22715 total job listings in the dataset. Following this is trades & services along with healthcare & medial. Self employment was the least common category which is unsuprising considering SEEK is designed to allow companies to find employees.



**Figure 9: Average Job Salary each Sector Sorted by Number of Listings (left to right)**

Unsurprisingly, the highest paying classification is CEO & General Management followed by ICT. However, one thing to note is the difference between job availability for the two classifications as is shown in the number of job listings per sector (see Figure 8). This puts ICT in a better position for going into as a potential career path due to the higher demand.