BIG DATA ANALYSIS – SEEK JOB LISTINGS

FORBES, CONNOR

CONTENTS

PART 1 – DATA PREPARATION AND PRE-PROCESSING

Note: There is an interactive webpage which contains all information in this document along with interactive figures hosted on GitHub.

https://connorf25.github.io/Big-Data-Analysis/

1.1 - 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.

1.2 - 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 were 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').

1.3 - 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

2.1 - JOB METADATA:

After dataset cleaning has been performed, there remains 309,870 data rows. The mean salary is $89.70 with a standard deviation of 108.55. The minimum potential salary is listed as $0 with the highest maximum at $999. It should be noted that the standard deviation for 'HighestSalary' is much bigger than 'LowestSalary' with values of 177.42 and 51.00, respectively.

	Average Salary ($)	Lowest Salary ($)	Highest Salary ($)
Mean	89.70	65.46	133.93
Standard Deviation	108.55	51.00	177.42
Minimum	15.00	0.00	30.00
Median	65.00	60.00	70.00
Max	599.50	200.00	999.00

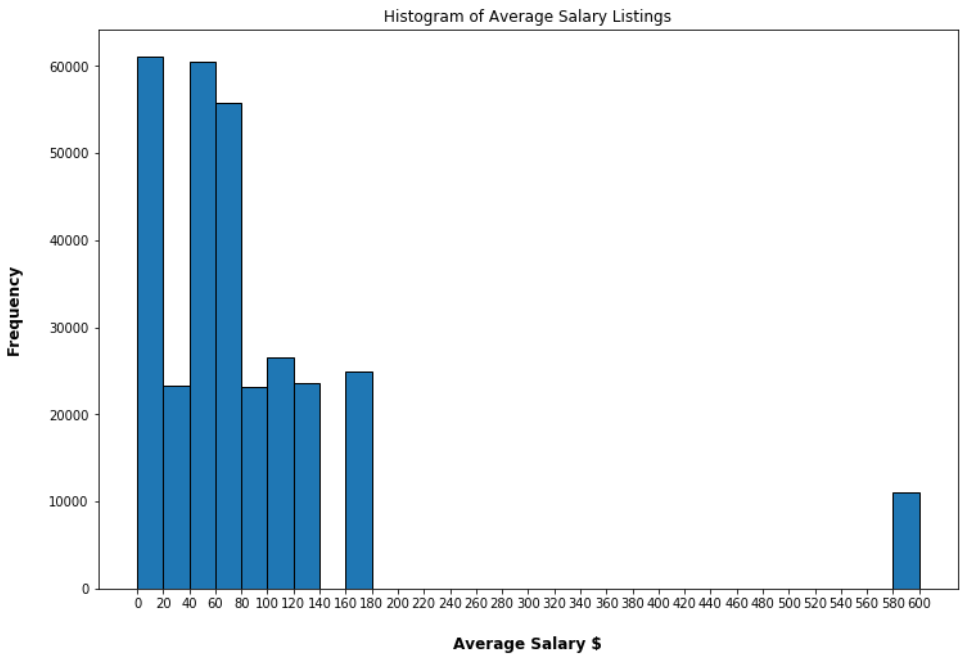
Table 1: Description of Data Columns

As for categories within the dataset, for sectors there is a total of 30 sectors with 396 total sub-sectors for classification of job listings. In addition to this there is 65 locations with the main cities (Brisbane, Melbourne, Perth and Sydney) having a total of 22 smaller locations that the cities are broken down into.

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 expected given the typical hierarchical structure of jobs with the majority of jobs being low to mid pay.

Figure 1: Histogram of Average Salary for all Listings



The histogram demonstrates a high number outlier jobs paying $580 to $600. When analysing the market by sectors, exploratory data analysis will be used on the outliers to determine which job sectors are responsible for the outliers.

AVERAGE SALARY OVER TIME

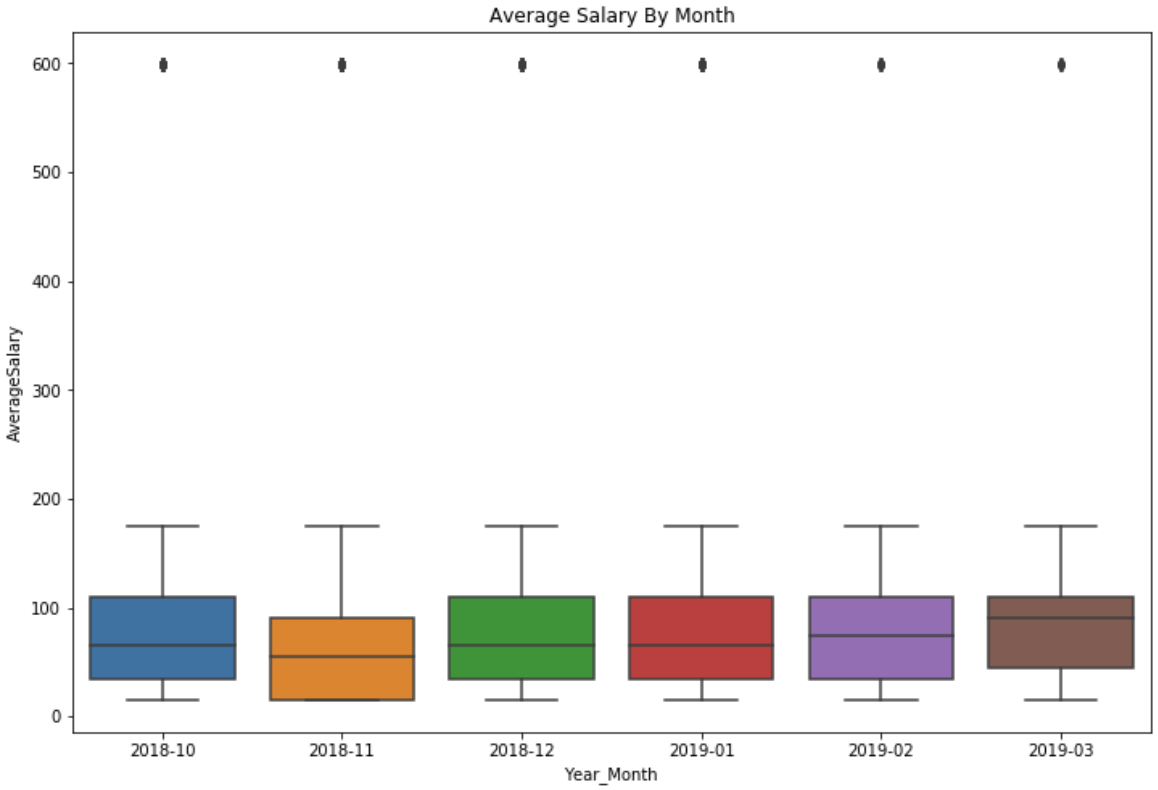


Figure 2: Average Salary by Grouped by Month Boxplot

The boxplot above for the average salary by month reveals that while salary is mostly even month to month, 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 3); that there was a huge spike in the average salary around early December.

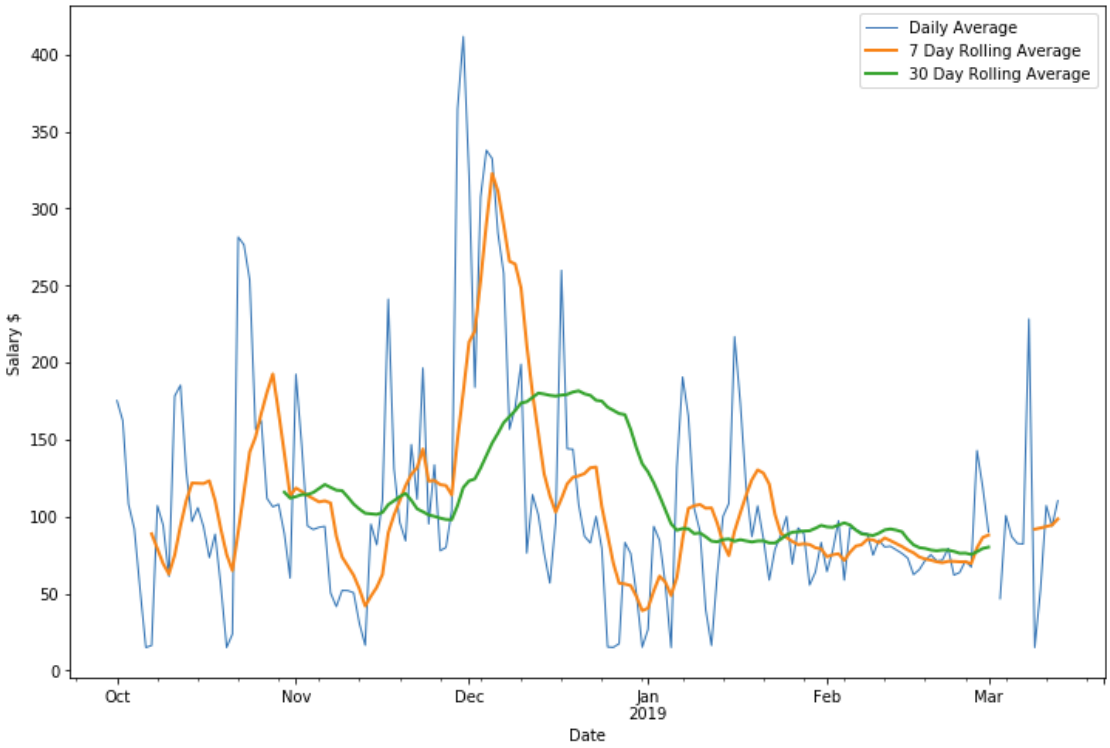


Figure 3: Time-series Graph for Salary

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 4 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 5). From the 23rd to the 28th in the month there appears to be overall a lower rate of posting.

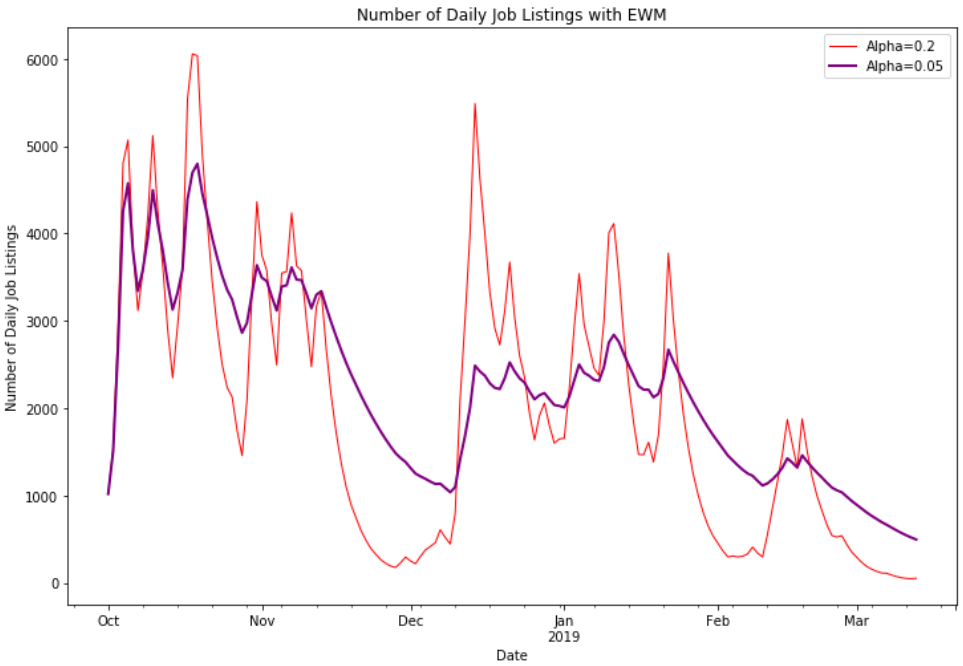


Figure 4: Number of Daily Job Listings with Exponential Weighted Smoothing

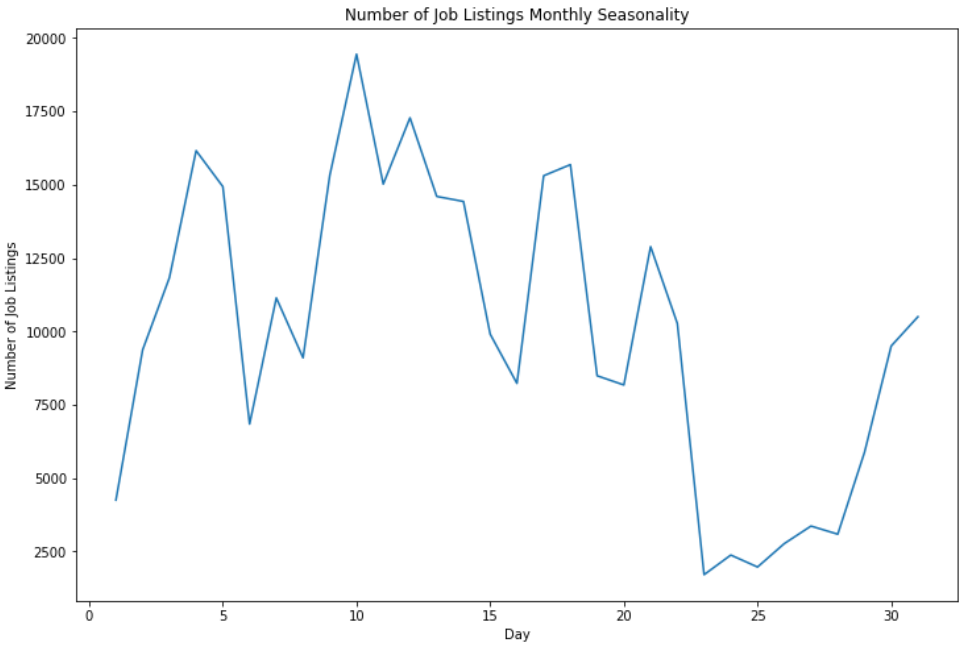


Figure 5: Number of Job Listings Grouped by Day of the Month

2.2 - MARKET BY LOCATIONS

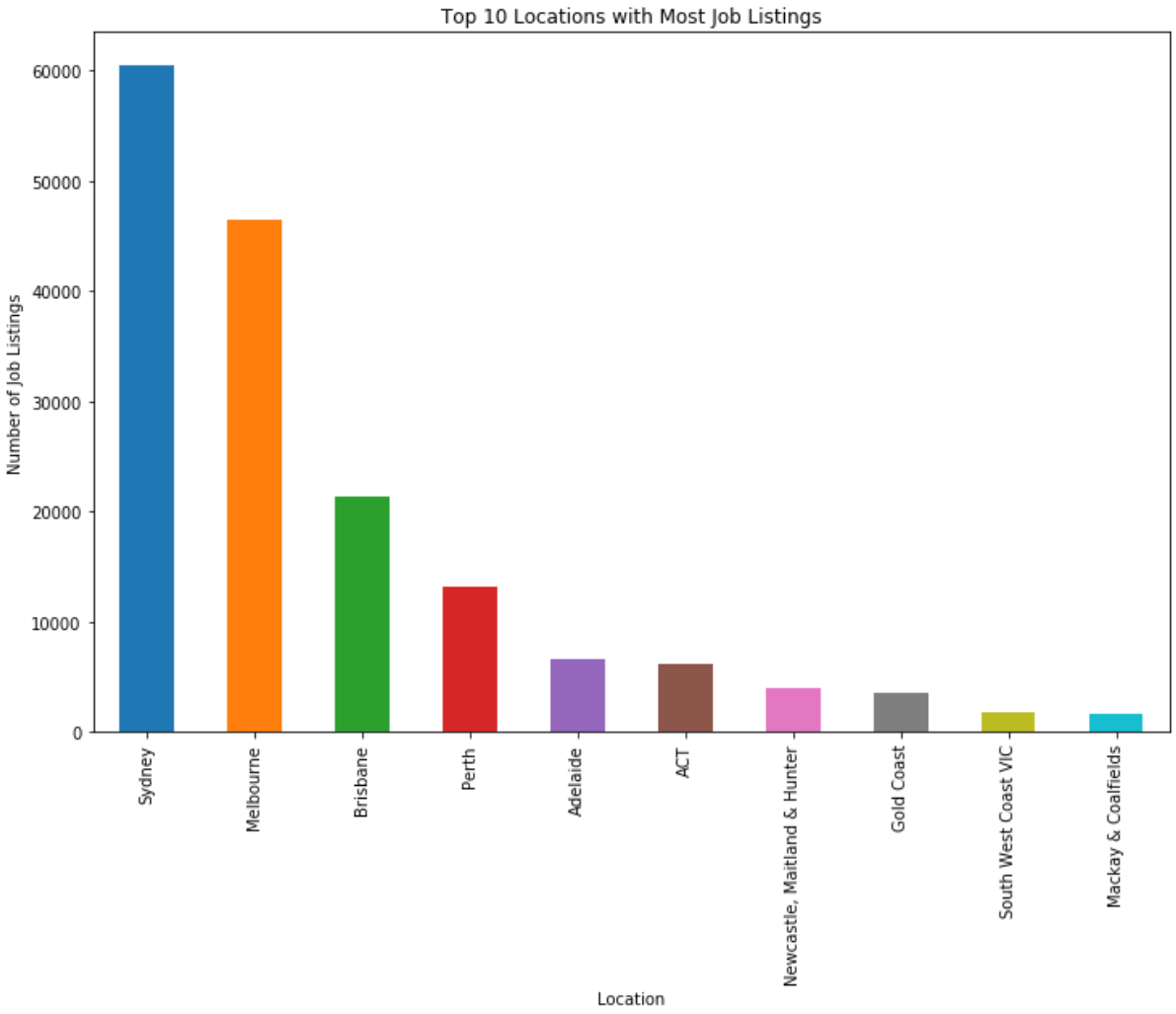


Figure 6: Top 10 Job Locations

As is expected, the main cities are where the majority of job listings appear in the major capital cities in Australia. Gold coast ranks 8th on this scale with 3581 job listings. Sydney is by far the largest with 17 times the number of listings at 60462 total listings. The number of job listings then appears to exponentially decrease when moving from the major cities to the minor cities.

Analysis of the average salary at each location revealed an unexpected result; the top paying locations aren't always the large cities. For example, the small town of Port Headland was ranked 7th with an average salary of $99.53 despite being a small country town. Exploratory analysis of the sectors in port headland should help explain this phenomenon.

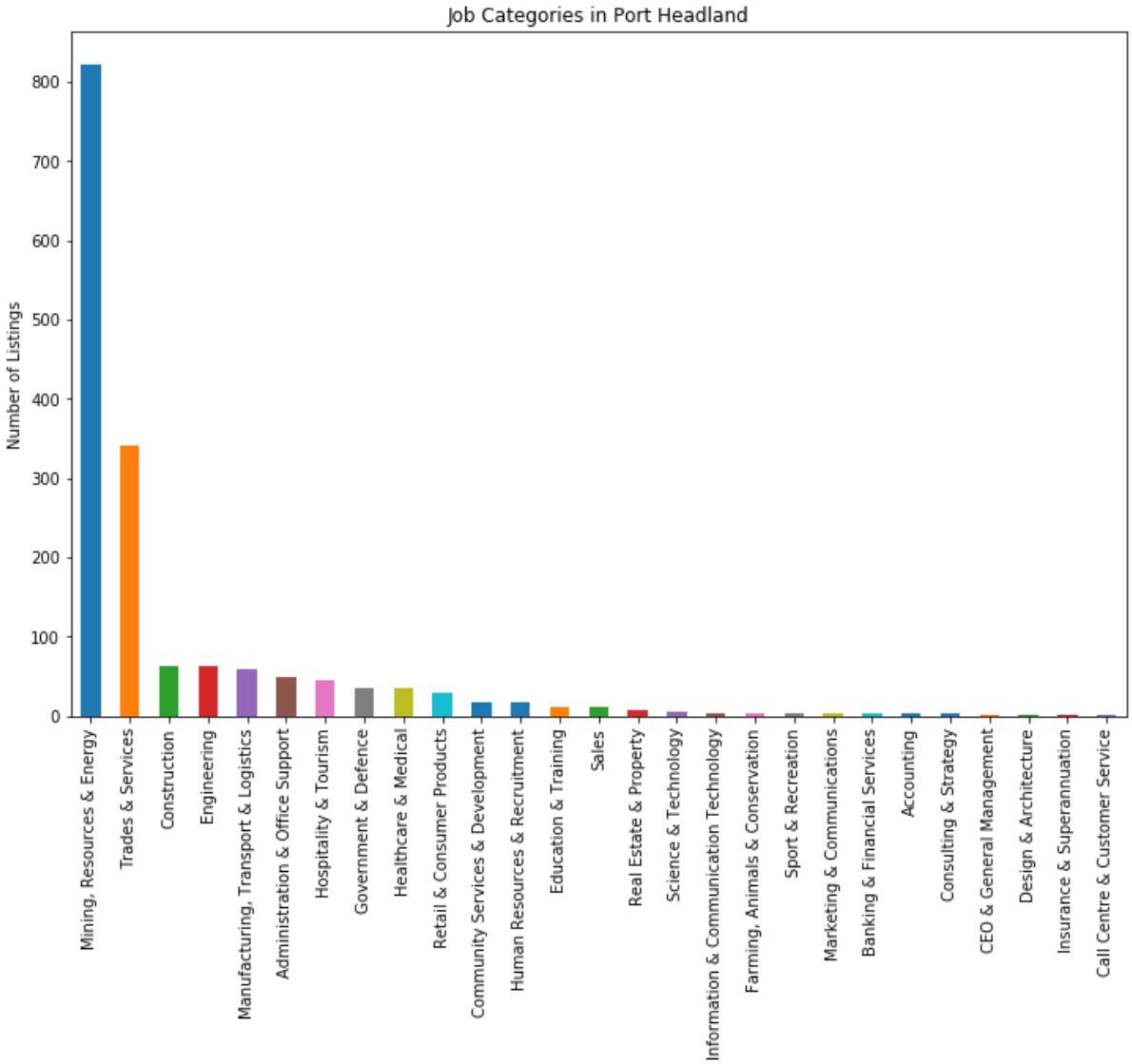


Figure 7: Job Categories in Port Headland

Analysis of the job categories in port headland reveal that the largest category is Mining, Resources & Energy. This explains the unusually high salary for this area as the average salary for mining is $128.66. It is therefore reasonable to assume that high salaries in small country towns can be explained by the abundance of mining jobs.

2.3 - MARKET BY SECTORS

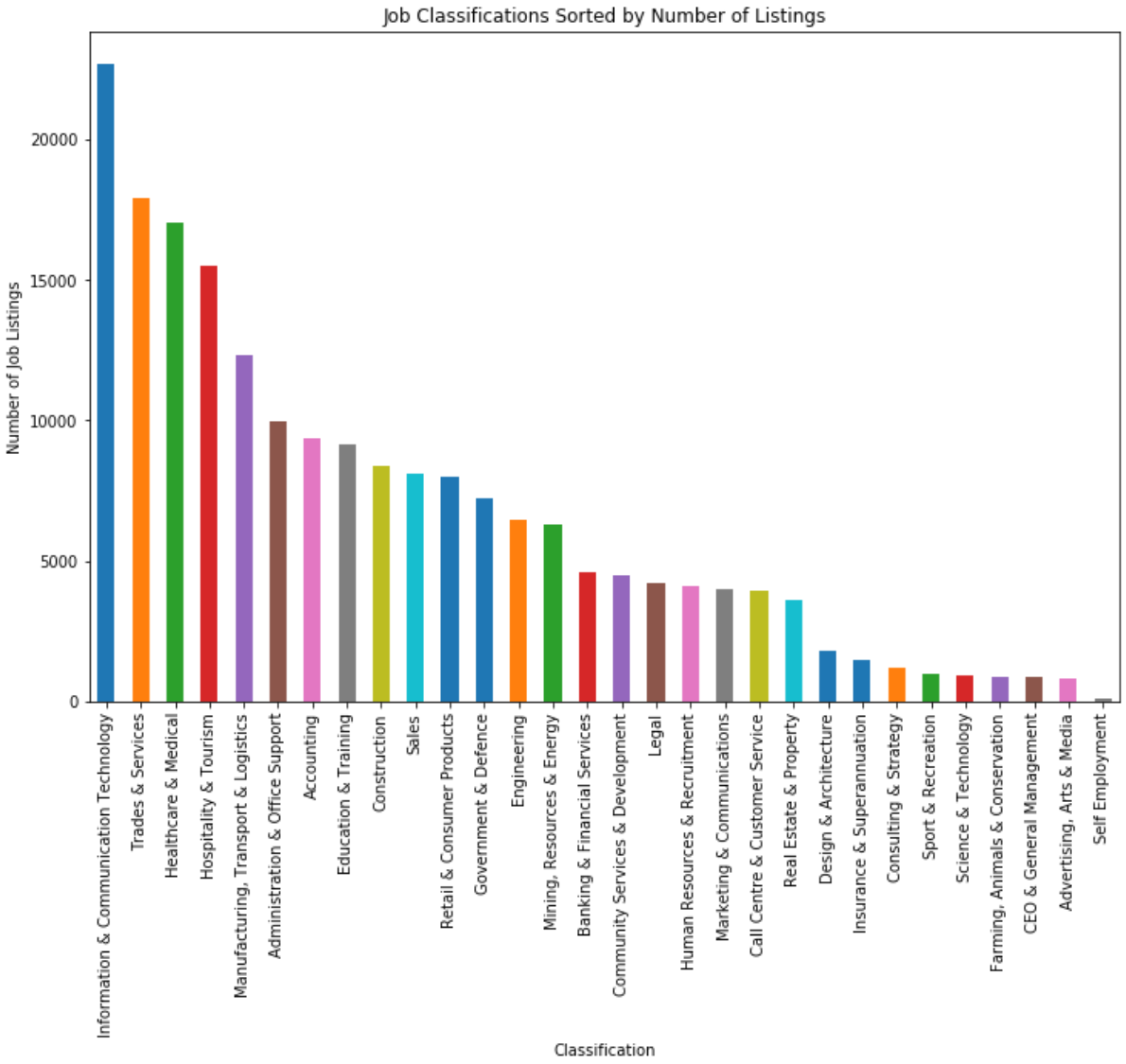


Figure 8: Number of Job Listings by Sector

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.

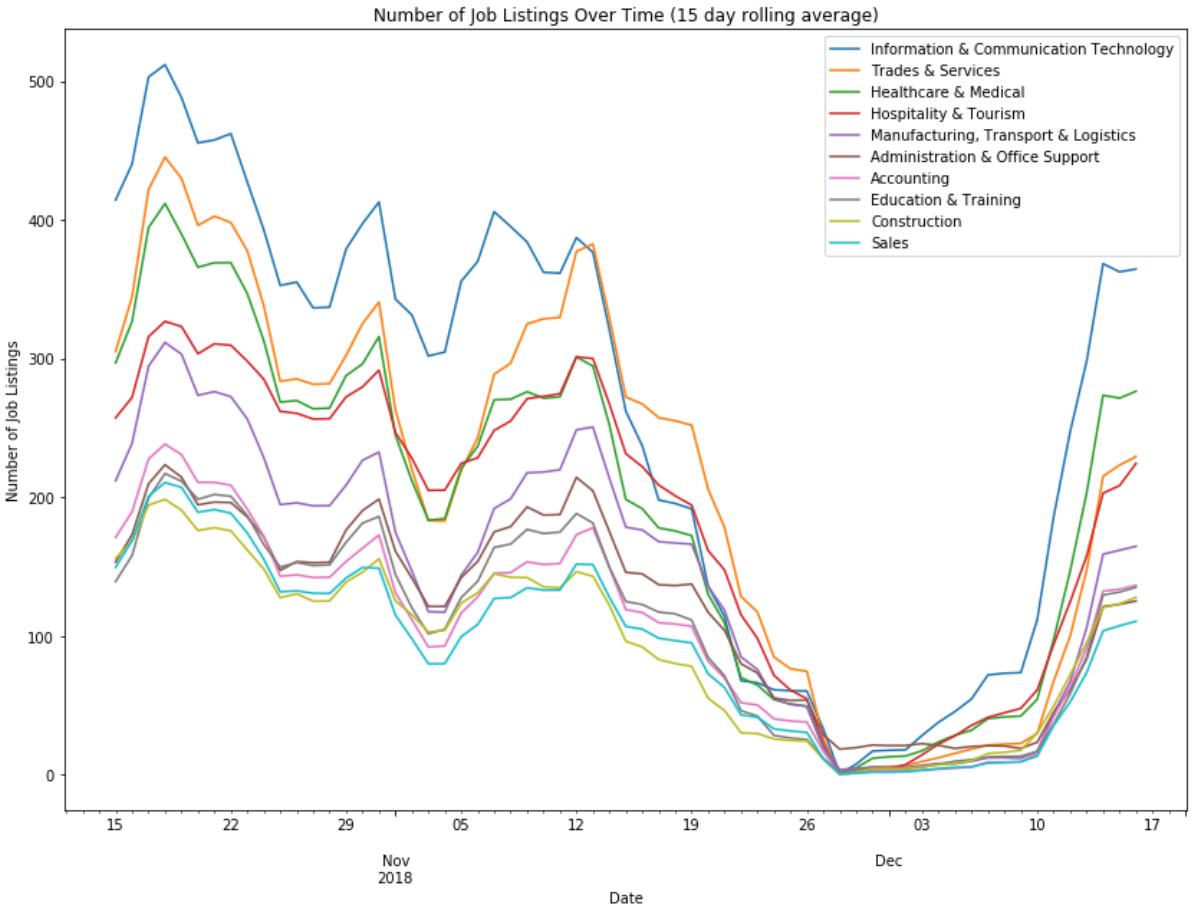


Figure 9: Number of Job Listings by Sector Over Time (30 day rolling average)

It appears in the number of job listings by sector there is continual decline until the start of December when a huge re-emergence in the number of job listings appears. Because the trend is being analysed over such a short timeframe. To assist with looking at the trend, linear regression will be implemented.

Figure 10 below demonstrates the implementation of linear regression. It is evident in the diagram that all sectors show an overall downtrend. Trades show the quickest decline, while all other fields show similar yet slower downtrends. However, because the data is only over 80 days, it is difficault to draw any solid conclusions.

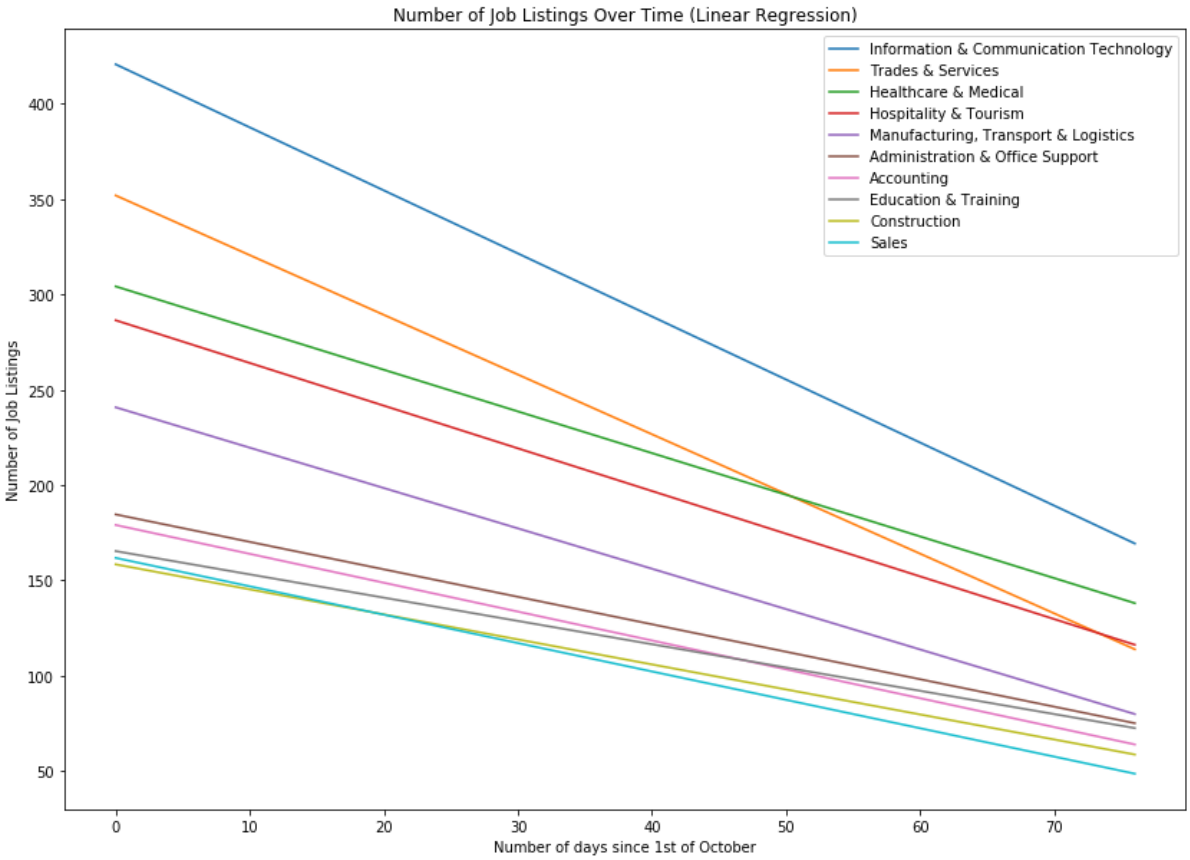


Figure 10: Trend of Job Listings with Linear Regression

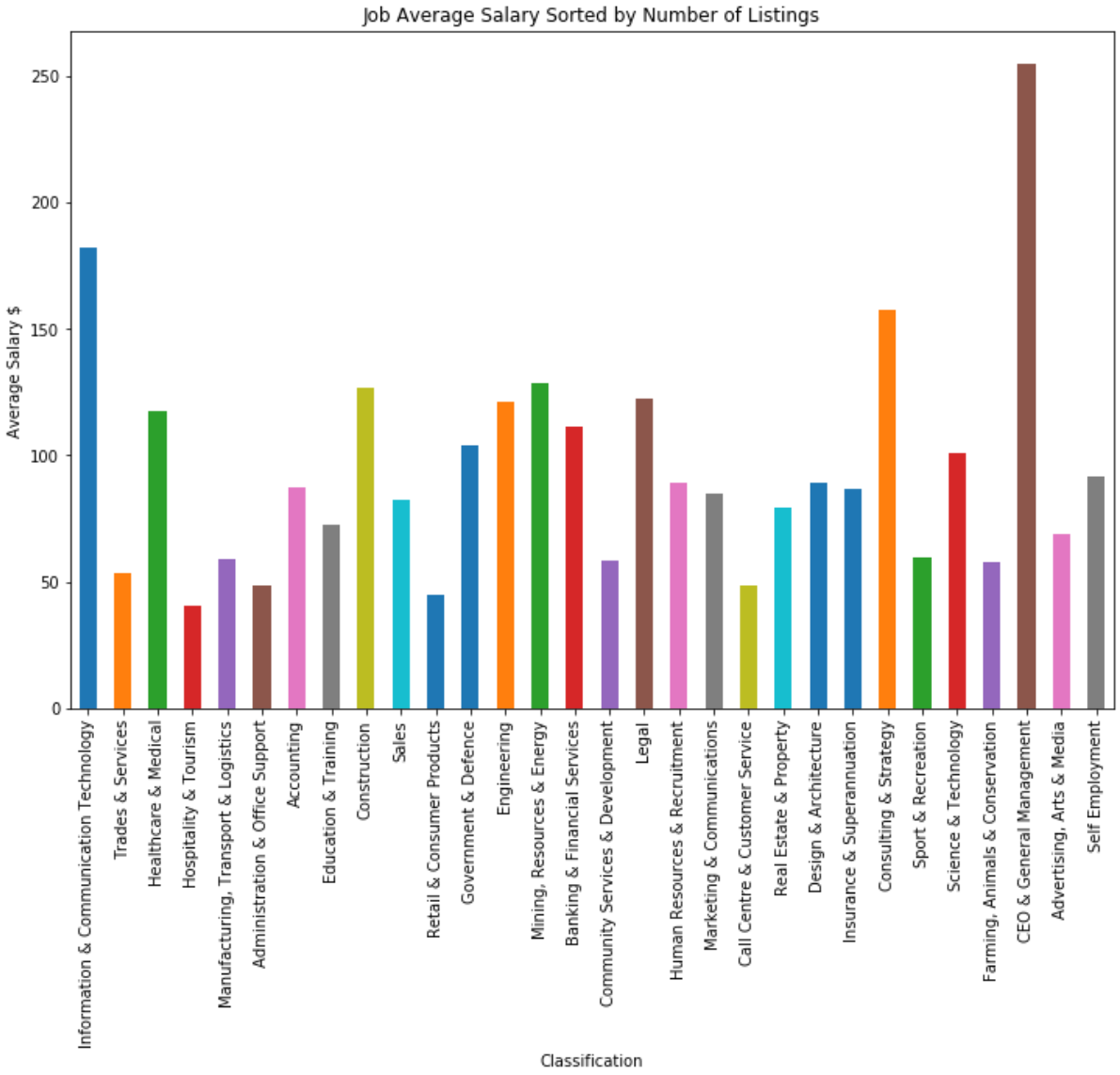


Figure 11: Average Job Salary each Sector Sorted by Number of Listings (left to right)

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.

Date	Company	Location	Area	Classification	SubClassification	LowestSalary	HighestSalary	JobType	Year	Month	Day	Year_Month	AverageSalary
2018-10-05	AIMS International Executive Search	Sydney	CBD, Inner West & Eastern Suburbs	CEO & General Management	General/Business Unit Manager	200	999	Full Time	2018	10	5	2018-10	599.5
2018-10-05	AIMS International Executive Search	Brisbane	CBD & Inner Suburbs	CEO & General Management	COO & MD	200	999	Full Time	2018	10	5	2018-10	599.5
2018-10-05	AIMS International Executive Search	Brisbane	CBD & Inner Suburbs	CEO & General Management	CEO	200	999	Full Time	2018	10	5	2018-10	599.5
2018-10-04	Hender Consulting	West Gippsland & Latrobe Valley	NaN	CEO & General Management	General/Business Unit Manager	200	999	Full Time	2018	10	4	2018-10	599.5
2018-10-04	Hender Consulting	Bairnsdale & Gippsland	NaN	CEO & General Management	General/Business Unit Manager	200	999	Full Time	2018	10	4	2018-10	599.5

Figure 12: First 5 Outliers where the Average Salary is 580-600

Analysis of the outliers appear to reveal that all of 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 identical salary is that all jobs are listed at the highest possible price range allowed on SEEK.

Out of these outlier jobs, ICT, Healthcare, Construction, Mining and Government are all prevalent sectors in the data further supporting the idea that these are the highest paying job categories.

2.4 - INTERACTIVE RESULTS

https://connorf25.github.io/Big-Data-Analysis/website/interactive.html

This GitHub webpage features interactive figures the user can interact with. It works through implementation of d3 JS to allow embedded data documents. They are hosted on ObservableHQ and embedded onto the webpage.

The first diagram is a collapsible tree which allows the user to browse the subsectors for each category in an easy to visualize format.

The second diagram is another collapsible tree but with the data separated by Location rather than sectors.

PART 3 – EVALUATION

3.1 - FINDINGS

Through analysis of the data analytics, it has been established that there is an overall downtrend in the number of market listings across all sectors, while the average salary remains relatively stable. The large majority of salaries are listed at between $20-80. With quite a few outliers all listed at $200-999 which most likely correspond to the maximum possible values possible on SEEK.

There was a dip in salaries in mid-November most likely corresponding to a listing of casual Christmas positions in retail. In early December there was a surge in salaries most likely related to people leaving high paying positions to spend time with family instead. For people searching for a high paying job, this corresponds to the best time to apply.

Based off the number of listings in each sector, and the average salary in each, ICT appears to be the best field for balancing a high demand with a good salary (highest demand and 2nd highest average salary). While CEO & General Management offer the highest salary, it is not advisable to enter this sector looking for a job due to the low availability.

3.2 - BALANCING THE MARKETS

For balancing the job market, it is suggested that more workers choose to go into sectors such as ICT, Trades. Healthcare and Hospitality. This will help lower the number of active listings on SEEK for these industries.

When balancing the locations in terms of employment, it is suggested that employers move more towards remote working to prevent the oversaturation of jobs which are exclusive in cities.

3.3 - REFINEMENTS

The main refinement which could be made to the dataset is a longer timeframe than a few months. If for instance, the timeframe was expanded to a few years; seasonality could be taken into account to allow for looking at long term trends. Without any big disruptions, the job market tends to remain stable over the timeframe of months and hence, very little can be extracted from the data by looking at trends over time.

In addition to this, the 2019 data is missing multiple fields including the sector and subsector of each job. This makes time series analysis for the data broken down into sectors even harder as there is only 3 months to look at.

Other job listing websites should be explored and if international websites are used, the international data could be compared to Australian data to compare the markets in different countries. Perhaps this could even be used to make predictions about where to move depending on the decided sector as a career path.

3.4 - IMPLICATIONS FOR EMPLOYEES AND EMPLOYERS

The main implication for employees within the findings of the data analysis is that to land a casual retail job, it is best to look around mid-November and to find a high paying job it is best to look around early/mid-December. As for which sector to target, for the best job opportunities it is recommended to live in a major city and choose to do ICT. This will also provide a competitive salary. For a high salary with out the skill investment needed for ICT, it is possible to chose to do mining work in a small country town such as Port Headland.

For employers, it is recommended to create listings for Christmas casuals in early November if it is desired to get in early and have the competitive advantage over other businesses. The trade off for this is that the employee will have to remain hired for a longer time compared to listing it later.

ONLINE DATA STORY

https://connorf25.github.io/Big-Data-Analysis/