Griffith School of Information and Communication Technology
Griffith University

3803ICT - Big Data Analysis
Trimester 1, 2022

Lab Report: Job Market Analysis

Jessy Barber, s5138877
Zac Jensen, s5153515

A report submitted in partial fulfilment of the degree Bachelor of Computer Science

TABLE OF CONTENTS

1 Data Preparation and Preprocessing

The data used in this exploratory analysis will be the provided excel spread sheet, "data.csv".

1.1 Describe the Dataset

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 318477 entries, 0 to 318476
Data columns (total 13 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Id                 318477 non-null  object
 1   Title              318477 non-null  object
 2   Company            306473 non-null  object
 3   Date               318477 non-null  datetime64[ns, UTC]
 4   Location           197229 non-null  object
 5   Area               122658 non-null  object
 6   Classification     197229 non-null  object
 7   SubClassification  197229 non-null  object
 8   Requirement        318470 non-null  object
 9   FullDescription    302302 non-null  object
 10  LowestSalary       318477 non-null  int64
 11  HighestSalary      318477 non-null  int64
 12  JobType            302379 non-null  object
dtypes: datetime64[ns, UTC](1), int64(2), object(10)
memory usage: 31.6+ MB
```

Figure 1: Categories / Domains of the Dataset

As seen in figure 1, the categories / domains of the dataset are clearly shown. Figure 1 also shows the number of non-null values that exist in each of these categories. The types of these categories are int64, which represents the lowest salary / highest salary categories, datetime64, which has been used to convert the Date category from its original object format, and the rest of the data are object file formats. The object file format represent strings since these categories contain strings describing their respective job meta data. The original job market dataset contains 13 columns of categories and contains 318'477 rows.

This report will conduct multiple vectors of analysis on this job data including analysis on the job metadata / attributes, analysis on the market by locations and analysis on the market by sectors. This analysis will then be visualised using an interactive visualiser. For the attribute analysis, the sector / sub-sectors for each job will be studied, along with the location and range of salaries for each job. The locational analysis will take a further look at the market size in each city and their hottest sectors. The range of salaries common in each city and where the employees are best paid will also be studied. Additionally, the pattern of job posts for each city will be analysed. The market's sectors will then be studied to determine which sectors keep the highest market share, which sub-sectors are of particular interest, what salary ranges are common for each sector / sub-sector, what is the market trend in terms of its sectors and which skills are required for each sector.

1.2 Describe the Steps You Used for Data Preparation and Preprocessing

```
df = pd.read_csv("data.csv", low_memory=False)
print(f"The dataset has {df.shape[0]} rows and {df.shape[1]} columns.")
df['Date'] = pd.to_datetime(df['Date']) # The data is changed to datetime var type
df.head()
```

```
The dataset has 318477 rows and 13 columns.
```

	Id	Title	Company	Date	Location	Area	Classification	SubClassification	Requirement	FullDescription	LowestSalary
0	37404348	Casual Stock Replenisher	Aldi Stores	2018-10-07 00:00:00+00:00	Sydney	North West & Hills District	Retail & Consumer Products	Retail Assistants	Our Casual Stock Replenishers pride themselves...	NaN	0
1	37404337	Casual Stock Replenisher	Aldi Stores	2018-10-07 00:00:00+00:00	Richmond & Hawkesbury	NaN	Retail & Consumer Products	Retail Assistants	Our Casual Stock Replenishers pride themselves...	NaN	0

Figure 2: Loading the Data with Pandas

As seen in figure 2, the .csv file is read in and stored as a DataFrame type in the variable df. The head of the dataframe is then printed for visualisation purposes.
The first step to begin working with the data is to load it into a DataFrame using Pandas. Pandas is a flexible and powerful open-source data analytics tool, built for use in Python. To load the csv file into a DataFrame, just call read_csv with the filename. In this assignment, an optional parameter "low_memory=False" is also provided, to allow Pandas to use enough memory to determine the correct datatype for each column due to the size of the dataset. Without this parameter, Pandas will attempt to guess the datatypes, which may lead to unexpected results. To normalize the data, the average salary is calculated for all job entries, by taking the LowestSalaray and HighestSalary columns, and placed back into the DataFrame as a new column AverageSalary. This number is then multiplied by 1,000 and formatted for easier readabilitiy. This results in an average salary looking like "15,000" instead of "15.0". Normalizing the data this way provides each job entry with a fair visualisation of salary.

The dataset also requires some cleaning for a couple of the columns. The "Id" column is how Seek keeps track of unique job entries and should be an integer number, but occasionally contains some random characters. To clean this up, a regular expression is used to remove any occurence of characters that occur after - and including - an ampersand. The "Date" column also contains extra information that is not necessary for this analysis. This column includes hours, minutes, and seconds, but only the day, month, and year are required. To clean this column, a regular expression is used to remove anything after and including a 'T' character. This results in the "Date" column only containing the necessary information in the format yyyy-mm-dd. After the data cleaning is complete, the correct dtypes are assigned to the "Id" and "Date" columns.

1.3 Hypothesis About the Analysis Outcome

The expected result is that jobs will be concentrated on the coast, with the highest number seen in the five large cities of Australia. A similar outcome is expected for the salaries of jobs; the further South you go - Brisbane, Sydney, Melbourne - the higher paying jobs you can expect to find. This is due to a large population density and higher cost of living in the Southern cities of Australia. It is also our prediction that the top performing sector will be ICT and that this sector will contribute to the majority of job listings in each city.

2 Data Analysis and Interpretation

2.1 Studying the Job Meta Data / Attributes

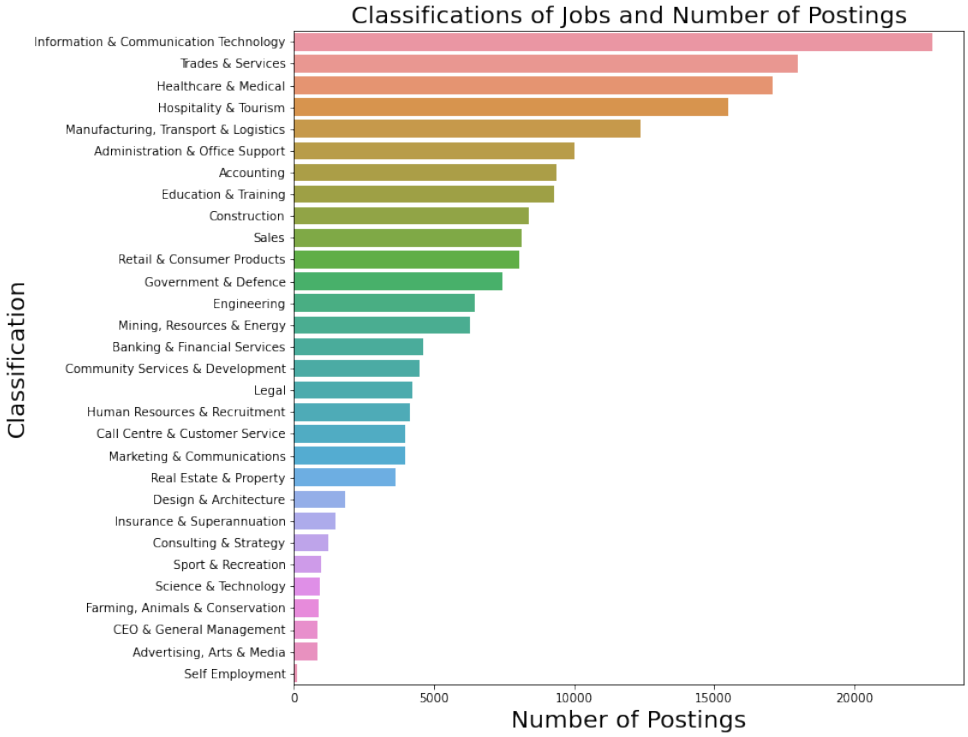

Figure 3: Classification of Jobs and Number of Postings

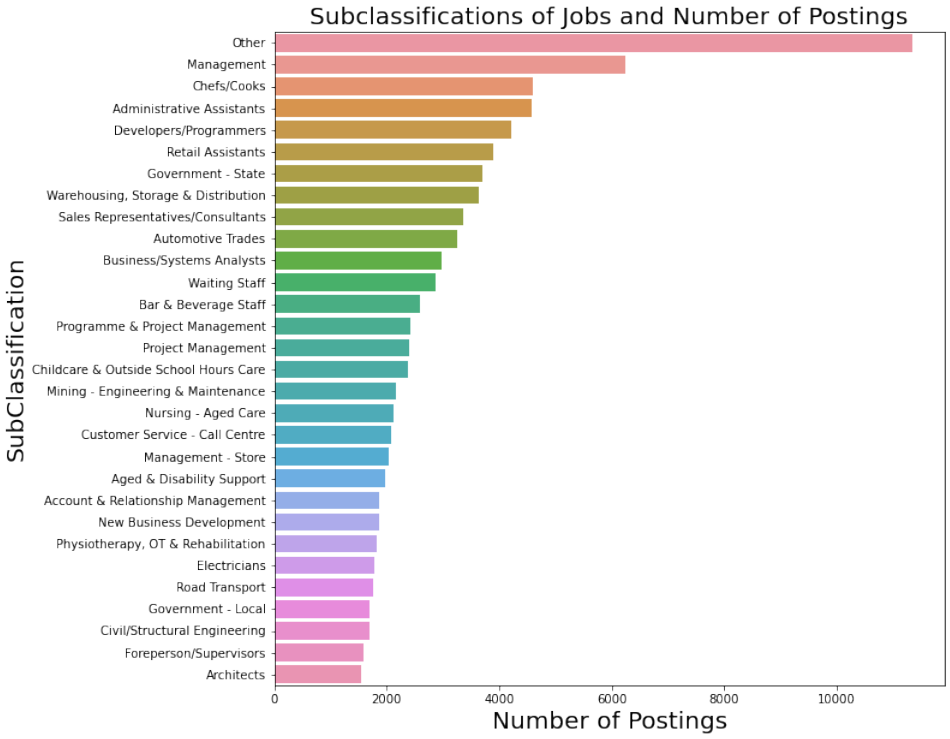

Figure 4: Subclassification of Jobs and Number of Postings

Figure 3 shows the 30 unique job classifications from the market dataset. Figure 3 also shows the posting frequency of each of these classifications with information and communication technology, trades and services, healthcare and medical, hospitality and tourism and manufacturing, transport and logistics being in the top five. Figure 4 shows the top 30 sub classifications from the market dataset. Figure 4 also shows the posting frequency of each of these sub classifications with other, management, chefs / cooks, administrative assistants and developers / programmers being in the top five.

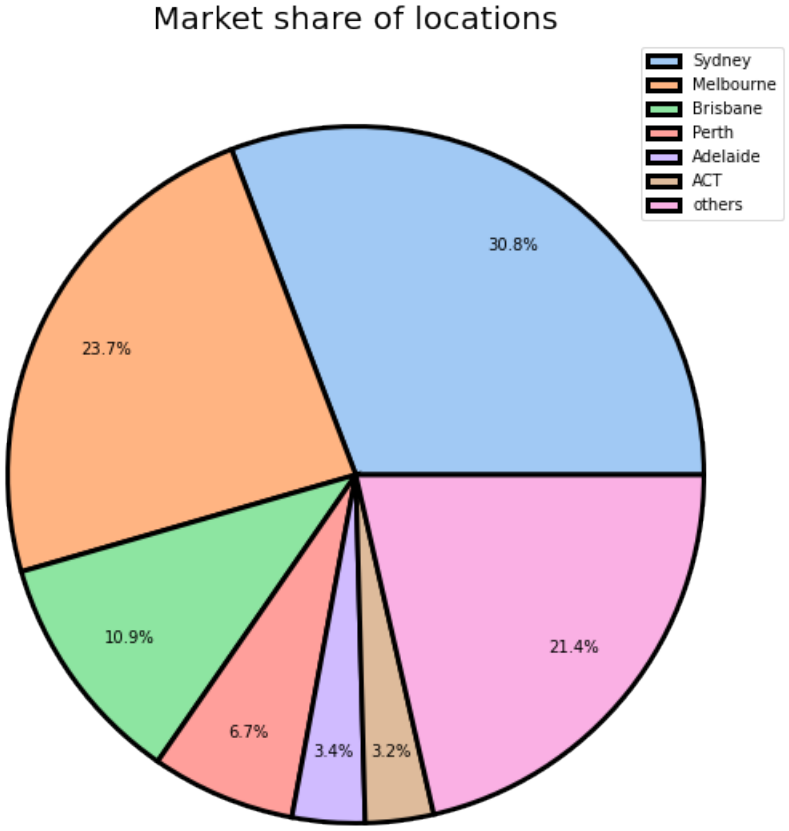

Figure 5: Top Five Cities for Market Share

Figure 5 shows the top five cities in Australia in terms of market share from the dataset. It is clear that Sydney holds the highest market share of employment at 30.8%, Melbourne in second with 23.7% of the market share and Brisbane in third with 10.9% of the market share. The other categories represents all other cities in Australia, and accounts for 24.6% of the market. Adelaide presents the lowest market share at 3.4% of the top five cities.
In the location analysis, the top three cities Sydney, Melbourne and Brisbane will be studied in terms of the market size in each city area, the hottest job sectors, the average salaries and the job posting dates for their respective job listings.

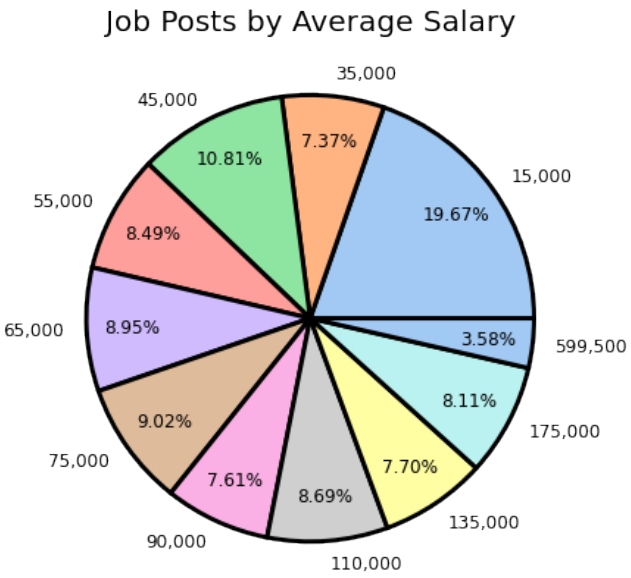

Figure 6: Average Salary Ranges of Jobs

Figure 6 shows the average salary range distribution for all of the listed jobs. As seen in this chart, the most common salary is $15,000 at 19.67%. Jobs at this salary can expect around $7.50 per hour working 40 hours a week, 50 weeks a year with 2 weeks of holidays. The highest average salary is around $599,500 at 3.58%. Jobs at this salary can expect around $299.75 per hour working 40 hours a week, 50 weeks a year with 2 weeks of holidays.

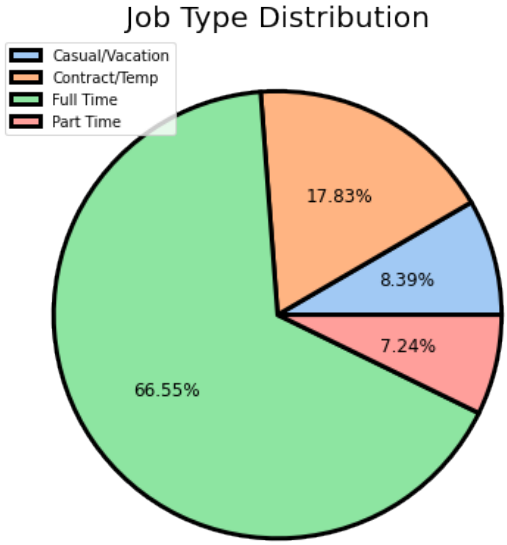

Figure 7: Distribution of Job Types

Figure 7 shows the distribution of advertised job types. From this pie chart it is clear that most jobs are under the job type "Full Time" at 66.55%, whilst the smallest number of jobs fall under the job type "Part Time" at 7.24%.

2.2 Studying the Market by Locations

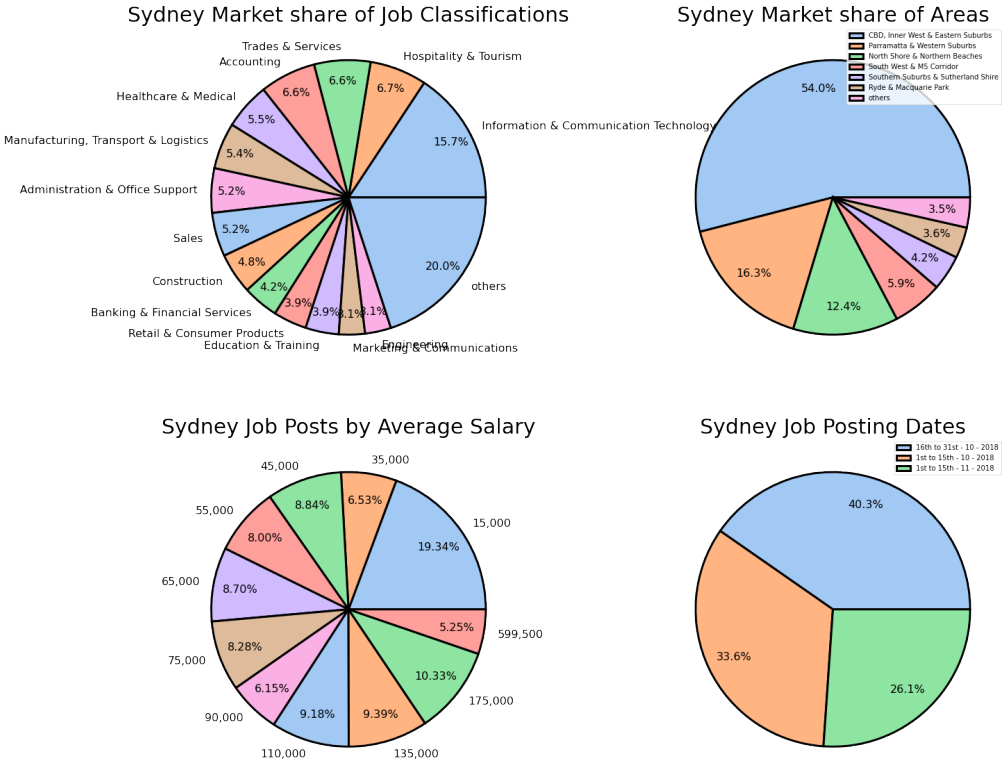

Figure 8: Sydney Locational Analysis

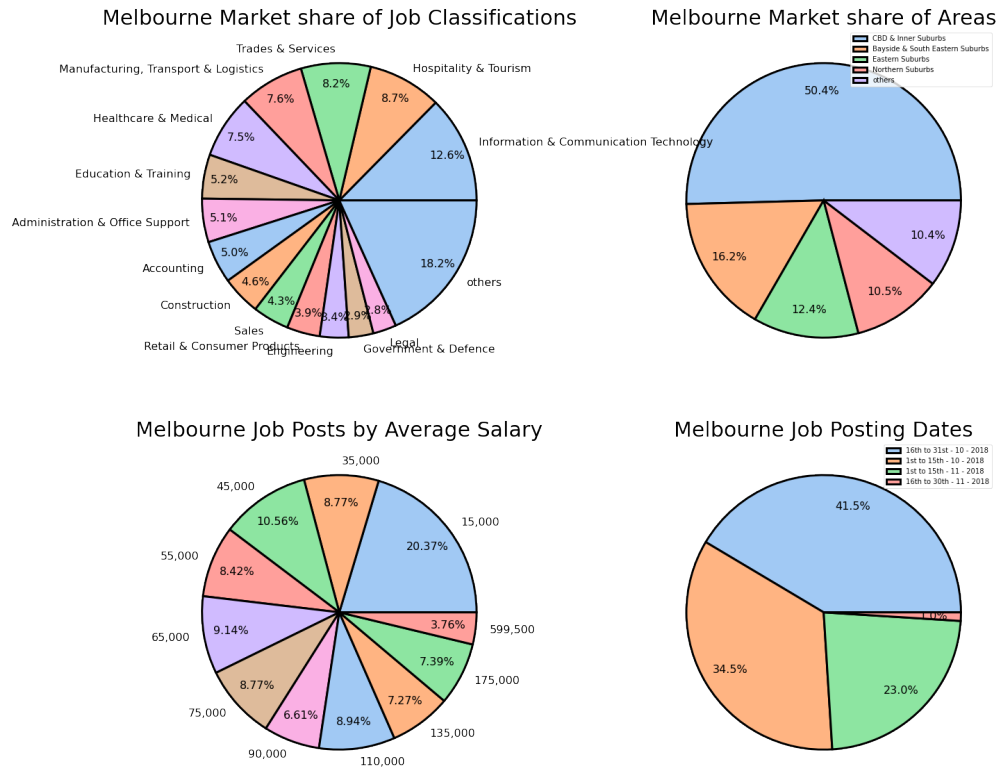

Figure 9: Melbourne Locational Analysis

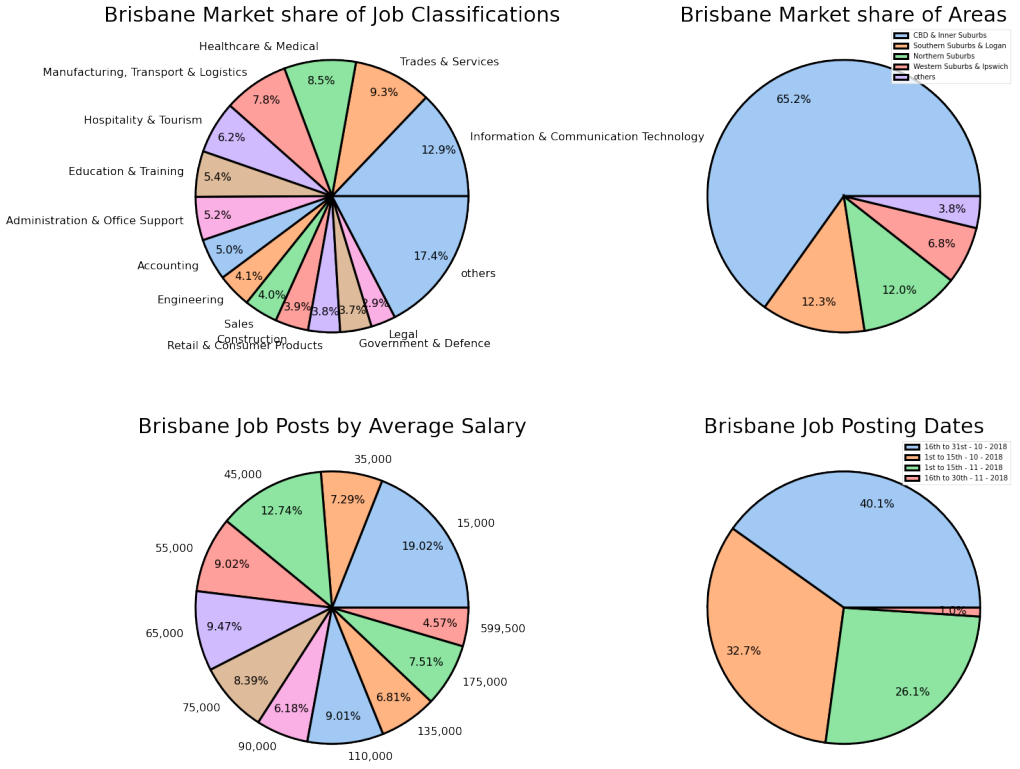

Figure 10: Brisbane Locational Analysis

As seen in figure 8, the Sydney job market is cornered by the information and communication technology sector which contributes to 15.7% of the advertised jobs. More than half of the jobs, 54%, of the market in Sydney are within the CBD, Inner West and Eastern Suburbs. The average salary for the advertised jobs in Sydney is $15,000 taking up 19.34%, of the advertised jobs. The least common average salary in Sydney is also the highest at $599,500 taking up 5.25% of advertised jobs. The majority of jobs, at 40.3%, were listed in the second half of October, and 33.6% and 26.1% of the jobs were posted in the first half of October and November respectively. The least amount of jobs were advertised in the second half of November at 0%.

As seen in figure 9, the Melbourne job market is also made up with a majority of listings in information and communication technology at 12.6% of total job listings. The clear majority of job listings in Melbourne are in the CBD and Inner Suburbs of the city. The highest average salary from the job listings in Melbourne is $15,000 which take up 20.37% of the advertised jobs. The least common average salary is $599,500 which consist of 3.76% of the advertised jobs. Most of the jobs advertised in Melbourne were added in the second half of October at 41.5% of listings and the first half of October at 34.5%. The least amount of jobs were listed in the second half of November at 1%.

As seen in figure 10, the highest advertised job in Brisbane is again, information and communication technology at 12.9% of total job listings. The majority of the advertised jobs in Brisbane are located within the CBD and inner suburbs, accounting for 65.2% of listings. The average salary in Brisbane is also $15,000, at 19.02% of listings, with $599,500 once again being the least common listed salary at 4.57% of total listings. Most of the jobs in Brisbane were listen in the second half of October at 40.1% and the first half of October at 32.7%. The least amount of jobs were listed in the second half of November at 1%.

2.3 Studying the Market by Sectors

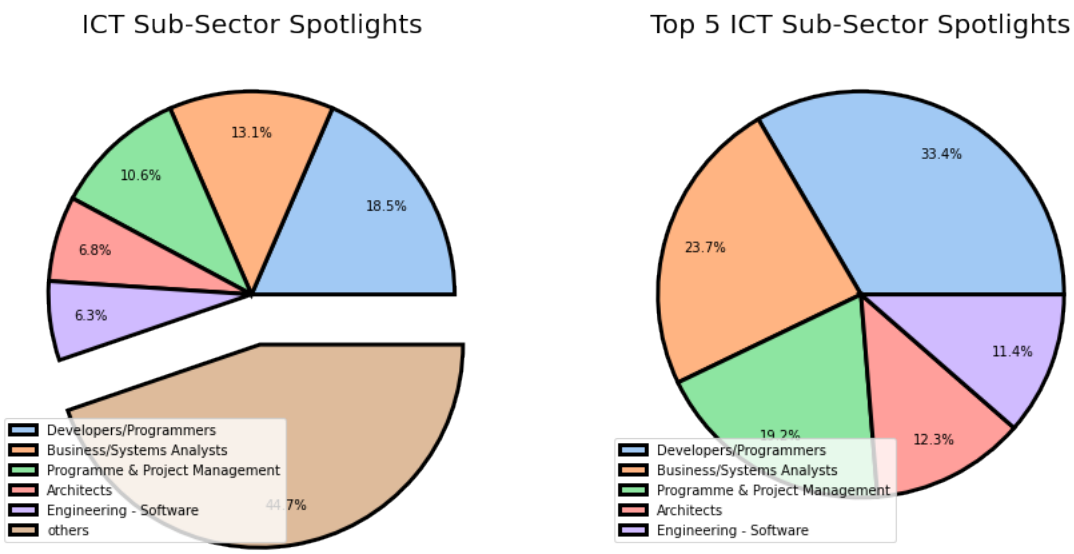

Figure 11: ICT Sub-Sector Spotlights

Figure 11 displays the top five sub-sector spotlights for the ICT job sector. As seen, developers/programmers is the highest listed sub-sector, accounting for 33.4% of the job listings within ICT. A close second and third are business/systems analysis and programme & project management, taking up 23.7% and 19.2% of the job listings in ICT respectively.

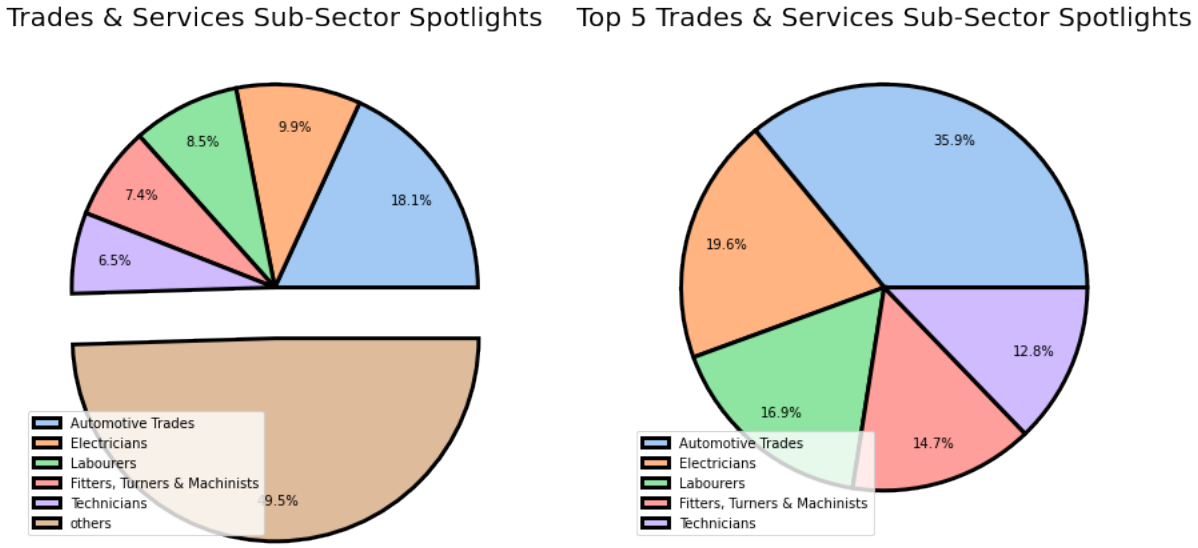

Figure 12: Trades and Services Sub-Sector Spotlights

Figure 12 displays the top five sub-sector spotlights for the trades and services job sector. As seen, automotive trades is the highest listed sub-sector, accounting for 35.9% of the job listings within trades and services. Second and third are electricians and labourers, taking up 19.6% and 16.9% of the job listings in trades and services respectively.

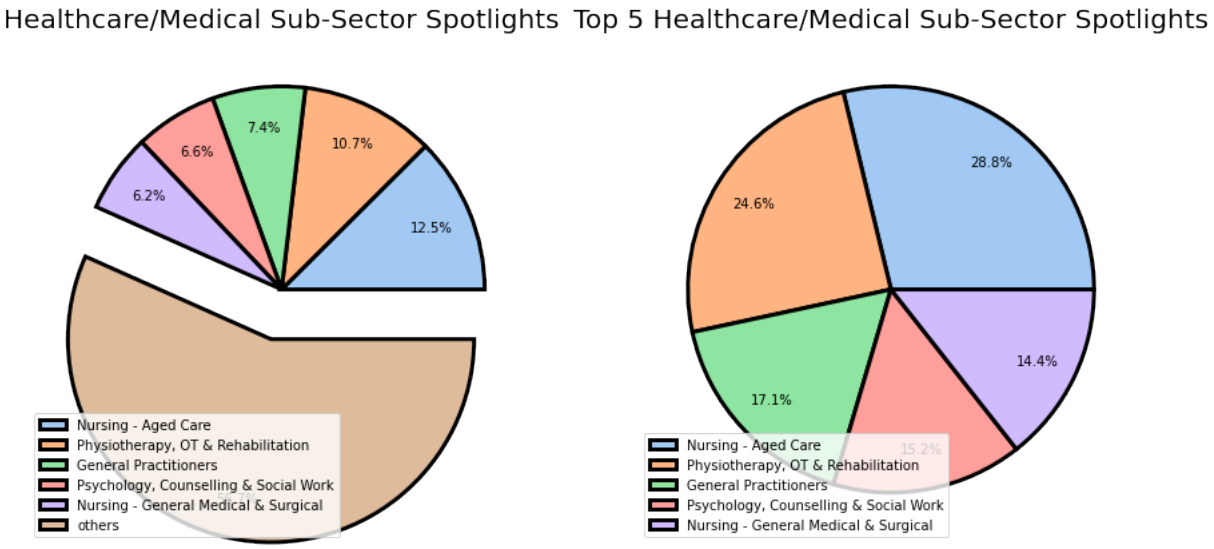

Figure 13: Healthcare and Medical Sub-Sector Spotlights

Figure 13 displays the top five sub-sector spotlights for the healthcare and medical job sector. As seen, nursing/aged care is the highest listed sub-sector, accounting for 28.8% of the job listings within healthcare and medical. A close second and third are physiotherapy, OT & rehabilitation and general practitioners, taking up 24.6% and 17.1% of the job listings in healthcare and medical respectively.

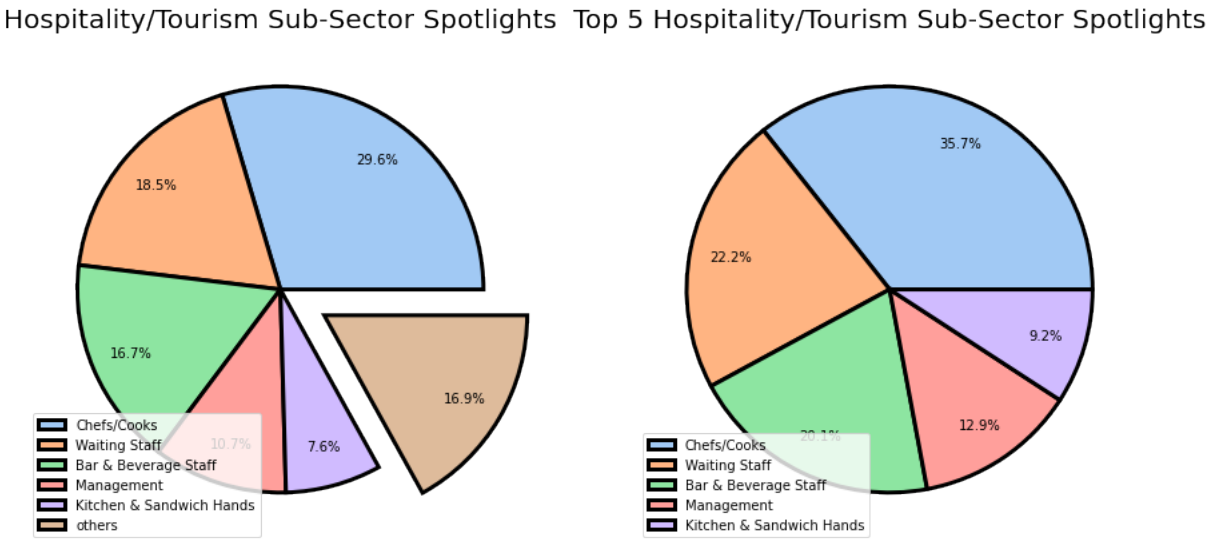

Figure 14: Hospitality and Tourism Sub-sector Spotlights

Figure 14 displays the top five sub-sector spotlights for the hospitality and tourism job sector. As seen, chefs/cooks is the highest listed sub-sector, accounting for 35.7% of the job listings within hospitality and tourism. A close second and third are waiting staff and bar & beverage staff, taking up 22.2% and 20.1% of the job listings in hospitality and tourism respectively.

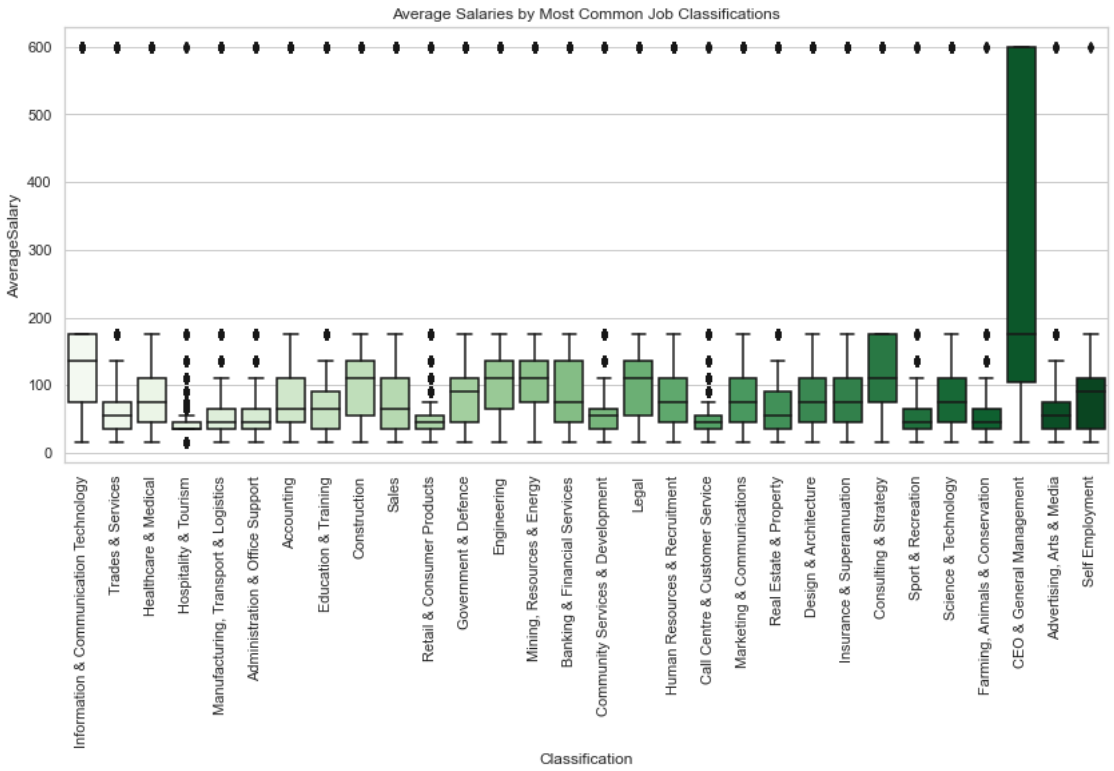

Figure 15: Average Salary Range for Job Sectors

Figure 15 shows the breakdown of average salaries in each job classification, sorted from most common to least common jobs. As seen in the boxplot, ICT is the most common job and has the highest median salary, besides CEO & General Management. Also present in the boxplot are the outliers in every classification. This is likely due to how some jobs may be listed as contract work and specify their rate in hourly or daily pay, which affects the data when pulled from the website.

As seen in figure 3, the highest sector of advertised jobs is the information & communication systems sector. Information & communication technology is the most listed job sector in the three biggest cities of Australia, Sydney, Melbourne and Brisbane, as evidenced in figures 8, 9 and 10, contributing to 15.7%, 12.6% and 12.9% of the job listings respectively. In this job sector, the hottest sub-sectors are developers/programmers, business/systems analysts and programme & project management as seen in figure 11, contributed to 33.4%, 23.7% and 19.2% of the job listings respectively. The information & communication systems sector is also one of the best paying sectors with a median average pay of around $150,000 as seen in figure 15. This is much higher than every other job sector except for CEO and general management. Therefore, since there is such a strong market trend in favour of ICT employability, it is recommended that a high school student considers a career in this sector to guarantee a job in the future. They should consider learning ICT relevant degrees in university such as IT, computer science or web development to acquire suitable skill sets to enter this job market.

3 Evaluation

3.1 Findings of Data Analytics

After conducting data analysis on the job meta data, locational analysis on the market and studying the market by sectors, there are clear trends in the job advertising in Australia. Looking at the market share of job listings in Australia, the top 5 cities were Sydney, Melbourne, Brisbane, Perth and Adelaide at 30.8%, 23.7%, 10.9%, 6.7% and 3.4% respectively, as seen in figure 5. The top five job posts by average salary were at $15,000, $45,000, $75,000, $65,000 and $110,000 at 19.67%, 10.81%, 9.02%, 8.95% and 8.69% respectively. For the sake of data analysis, all of these salaries were assumed to be yearly salaries. As seen in figure 7, the clear majority of advertised jobs were for full time work at 66.55%, followed by contract/temp work at 17.83%. The least amount of advertised jobs were of the part time type at 7.24%.

Locational data analysis was conducted for the top three cities in Australia, Sydney, Melbourne and Brisbane since they occupy a collective 65.4% of the market share for job advertisements. It was seen in figure 8 that most of the jobs in Sydney were located in the CBD, inner west and eastern suburb areas at 54%. As seen in figure 9, most of the jobs in Melbourne were also located in the CBD and inner suburbs at 50.4%, while the same area for Brisbane occupies 65.2% of job listings, as seen in figure 10.

In Sydney, the most advertised sector was information & communication technology, contributing 15.7% of the job listings. This was followed by hospitality & tourism, trades & services and accounting at 6.7%, 6.6%, and 6.6% respectively. In Melbourne, the most advertised sector was also information & communication technology which contributed to 12.6% of the advertised job listings, followed by hospitality & tourism, trades & services and manufacturing, transport & logistics at 8.7%, 8.2% and 7.6% respectively. In Brisbane, information & communication technology is again the most advertised sector at 12.9%, followed by trades & services, healthcare & medical and manufacturing, transport & logistics at 9.3%, 8.5% and 7.8% respectively.

The average salary for job postings in Sydney were predominately at $15,000 which contributed to 19.34% of job listings, followed by $175,000, $135,000 and $110,000 at 10.33%, 9.39% and 9.18% respectively. The average job postings in Melbourne were also mostly at $15,000, contributing to 20.37% of advertised jobs, followed by $175,000, $135,000 and $110,000 at 10.33%, 9.39% and 9.18% respectively. The average job postings in Brisbane were mostly $15,000, followed by $45,000, $65,000 and $55,000 at 12.74%, 9.47% and 9.02% respectively. In Sydney, Melbourne and Brisbane, the lowest advertised job had the highest salary at $599,500, contributing to 5.52%, 3.76% and 4.57% of advertised jobs respectively in each city. It is clear that the job market in Sydney and Melbourne pays much higher than the jobs in Brisbane on average.

40.3% of jobs in Sydney were posted in the first half of October 2018, whilst 33.6% and 26% of the jobs were posted in the first halves of October and November respectively. 41.5% of jobs in Melbourne were posted in the second half of October, with 34.5% and 23% of jobs posted in the second halves of October and November respectively. 40.1% of jobs in Brisbane were posted in the second half of October, with 32.7% and 26.1% of jobs advertised in the second halves of October and November respectively. There is a clear trend in the job advertisement dates for these three cities as they have very similar distributions of postings. This means that there is a surge in advertisement for jobs towards the end of the year, but drop off dramatically in the second half of October.

After sectional analysis on the job markets, it is more clear which sectors are more fre-

quently advertised. Sub-sector analysis was conducted for the top four sectors, information & communication technology, trades & services, healthcare/medical and hospitality/tourism. As seen in figure 11, the top five sub-sectors in the information & communication technology sector are developers/programmers, business/systems analysts, programmer/project management, architects and engineering - software, at 33.4%, 23.7%, 19.2%, 12.3% and 11.4% of the top five sub-sectors respectively. As seen in figure 12, the top five sub-sectors in the trades & services sector are automotive trades, electricians, labourers, fitters turners & machinists and technicians, at 35.9%, 19.6%, 16.9%, 14.7% and 12.8% of the top five sub-sectors respectively. As seen in figure 13, the top five sub-sectors in the healthcare/medical sector are nursing - aged care, physiotherapy OT & rehabilitation, general practitioners, psychology counselling & social work and nursing general medical & surgical, at 28.8%, 24.6%, 17.1%, 15.2% and 14.4% of the top five sub-sectors respectively. As seen in figure 14, the top five sub-sectors in the hospitality/tourism sector are chefs/cooks, waiting staff, bar & beverage staff, management and kitchen & sandwich hands, at 35.7%, 22.2%, 20.1%, 12.9% and 9.2% of the top five sub-sectors respectively. After this analysis, the top sub-sectors of the top four sectors are developers/programmers, automotive trades, nursing-aged care and chefs/cooks.
Figure 15 shows a box plot for average salaries of each advertised job sector. It is clear that the job with the highest variance is the CEO & general management sector. These jobs start at just over $100,000 and reach up to around $600,000. Compared with the sector analysis, the information & communication technology sector is well paid starting at around $80,000 and reaching around $180,000. The median of this sector is around $150,000, which is the highest median other than the CEO & general management sector.The second most advertised sector, trades & services is a relatively low paying field with its highest paying positions reaching the median of an ICT job. The third most advertised sector, healthcare/medical has a lower median average pay than ICT jobs, but has similar top paying jobs. The fourth most advertised sector of hospitality/tourism has significantly lower paid job opportunities, but with high paid outliers. These outliers are most likely due to contract or short term work however and are not necessarily indicative of a yearly salary.
The data analysis has indicated that the ICT sector is a clear best choice for a future job since it is the most advertised sector in the biggest three cities in Australia, and has the second highest median average pay.

3.2 Balancing the Market

As seen in various figures throughout this report, there are a disproportionate number of jobs in only a few cities in Australia. One action that could help balance the market is to allow more growth and job opportunities in more rural areas across the country. Doing so would provide employees with more options for where they can choose to reside. Another issue is the large number of jobs with average salaries of only $15,000. Potential employees viewing jobs with such low salaries are likely to be discouraged and seek alternate paths. Further incentivising higher education would allow salaries to be raised, increasing the likelihood of employees entering careers that they are passionate about; in turn improving work performance overall.

3.3 Refining the Data Analytics

There are several steps that can be taken to ensure that the data analytics is refined as possible. First of all, the major flaw in this data analysis is that the job data is only coming from one website, Seek.com. Therefore, a way to refine the data analytics would be to pull data from additional job market sites in Australia such as LinkedIn. A better prediction model could be used to suggest relevant high school subjects, university courses by using the bag of words library on the requirements column. Here relevant skills could be extracted for the top performing sectors such as ICT and recommended to young job-lookers. Another way to refine the data analytics would be to compare these data analytics with more time frames throughout the year. The analytics conducted in this report only consider the job market over October and November and so a longer time frame may be more representative of the true values.

3.4 Implications for Employers and Employees

There are various implications for both employers and employees based on the evidence found in this data. One such implication is the concentration of jobs based on location. As seen in figure 5, Sydney, Melbourne and Brisbane account for almost two-thirds of the entire job market in this dataset. This is a very limiting factor when it comes to an employee choosing where to live in Australia, as picking outside of these major cities puts them at a disadvantage for job opportunities. Another implication is the disproportionate pay gap for CEO's & General Management. According to figure 15, employees in this field can earn between $100k-$600k. These huge salaries may lead to people attempting to get into CEO & General Management solely for the pay. Poor management may ensue, resulting in damage and losses for the company and employees. These occurrences have been seen many times throughout various companies, often having disastrous results.

4 Case Studies

4.1 Case Study 1

The best choice for Mathew is the information & communication technology sector as his future career pathway since it is the most advertised sector in Australia, as seen in figure 3. The ICT sector contributes to 15.7%, 12.6% and 12.9% of the job listings in Sydney, Melbourne and Brisbane respectively as seen in figures 8 - 10. Looking at the requirement column for the information & communications technology sector, skills that immediately stand out include programming languages such as Python, C, C# and Java. As seen figure 11, the top sub-sector for ICT is developers/programmers at 33.4%. The second and third highest sub-sectors are business/systems analysts and programme & project management at 23.7% and 19.2% of job listings respectively. As seen in figure 15, ICT has the second highest average median salary out of all of the listed jobs at around $150,000.

Since there is such a strong market trend in favour of ICT employability, Mathew should consider pursuing a career in one of the top three sub-sectors of ICT, developers/programmers, business/systems analysts or programme & project management. Pursuing one of these career pathways will ensure that Mathew is not only highly employable but also paid well. In order for Mathew to make the best of these potential job markets, he needs to develop the appropriate skills through his computer science degree. This means learning as many programming languages as possible such as C, C++, C#, Python and Java. Mathew should also consider majoring in software development so that he can cover all of these languages in his classes. Mathew also needs to develop his analytic and problem solving skills if he wants to be successful in a future career in one of these ICT sub-sectors.

4.2 Case Study 2

The first step that should be taken is to perform any necessary data preparation and preprocessing on the job market dataset. This will ensure the data is in a clean and usable format, helping prevent any issues down the line. The next step would be to extract the key information from an employee's CV. This could be done using several methods: bag of words, term frequency-inverse document frequency (TF-IDF), and other methods. These keywords will be stored for later comparison. The next step would be to get the employee's geographical location, and eliminate any jobs that are too far away. Reducing an employee's potential travel time would increase the likelihood of choosing a job and improve their work-life balance. The last step would be to take the keywords extracted from the employee's CV, and compare them with the remaining potential jobs in the database. Using the requirements and skills from each job listing, give each job a ranking based on the similarity between keywords from the employee and the requirements/skills. Finally, rank these jobs in descending order and provide the top 10 jobs to the employee.