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

One expected outcome of the data analysis is the number of jobs based on location. 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 - Melbourne, Sydney, Adelaide - 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.

2 Data Analysis and Interpretation

2.1 Studying the Job Meta Data / Attributes

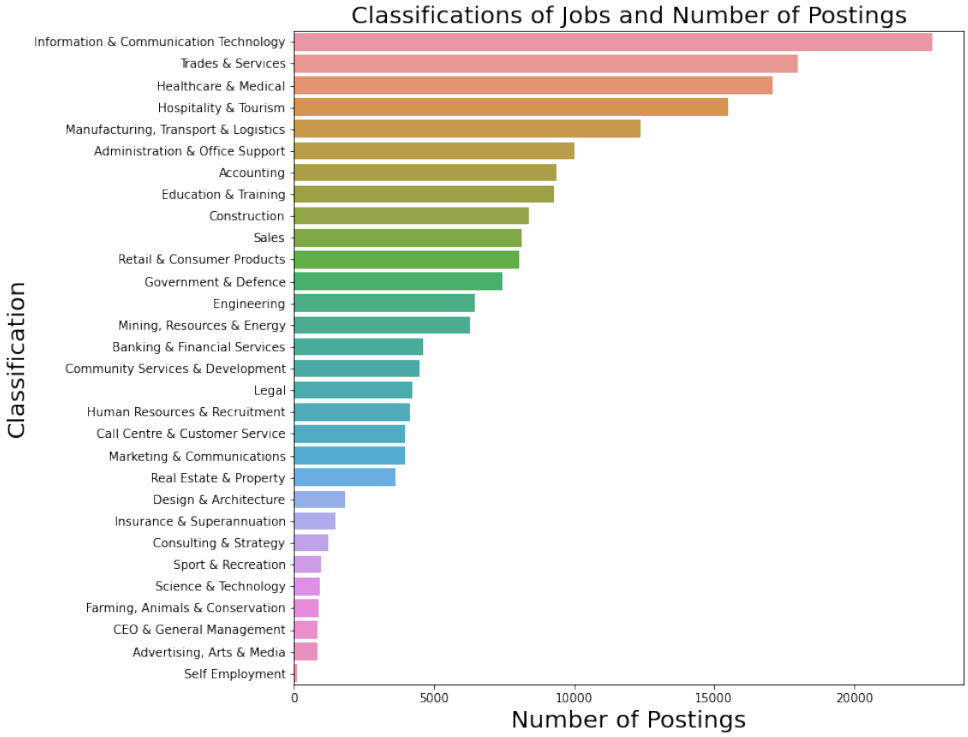

Figure 3: Classification of Jobs and Number of Postings

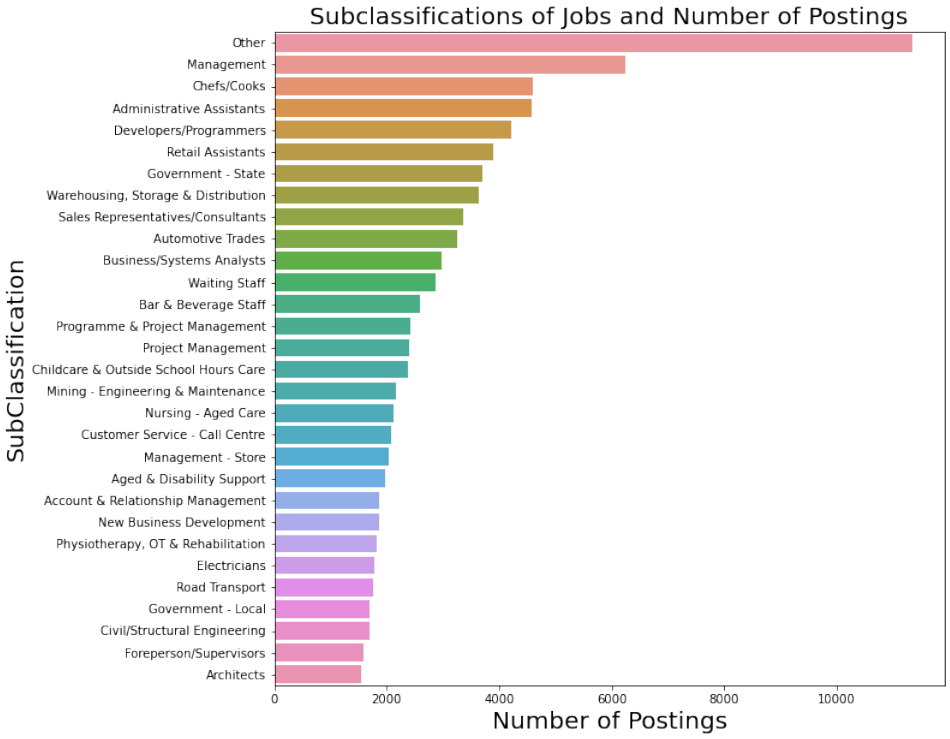

Figure 4: Subclassification of Jobs and Number of Postings

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.

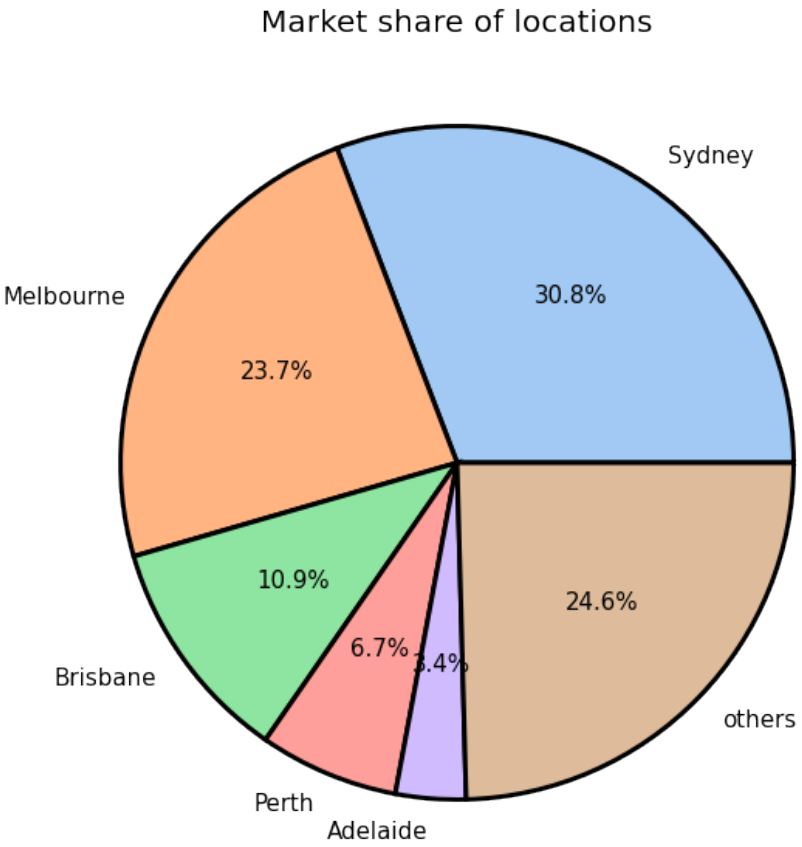

Figure 5: Top Five Cities for Market Share

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% and Melbourne in second with 23.7% of the market share. Adelaide presents the lowest market share at 3.4%.

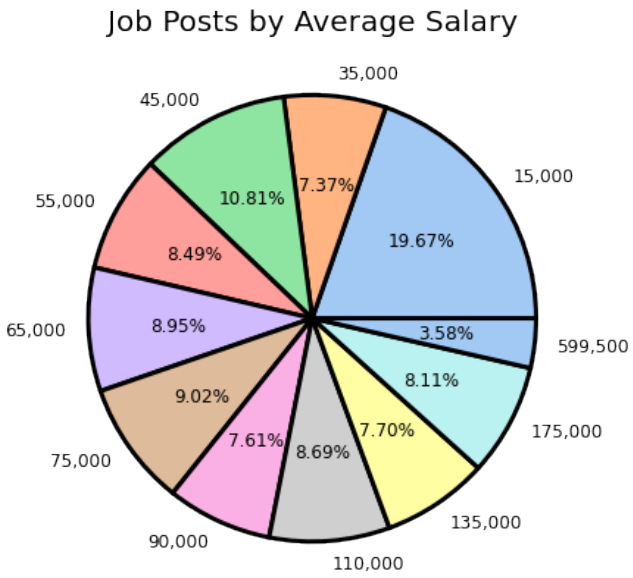

Figure 6: Average Salary Ranges of Jobs

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.

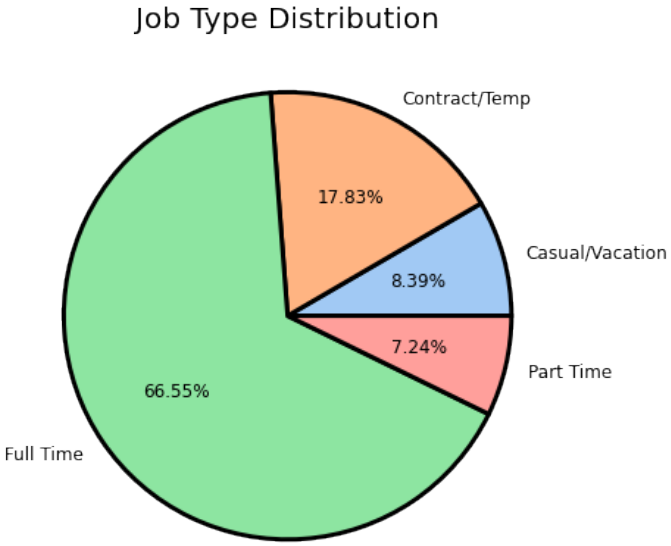

Figure 7: Distribution of Job Types

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

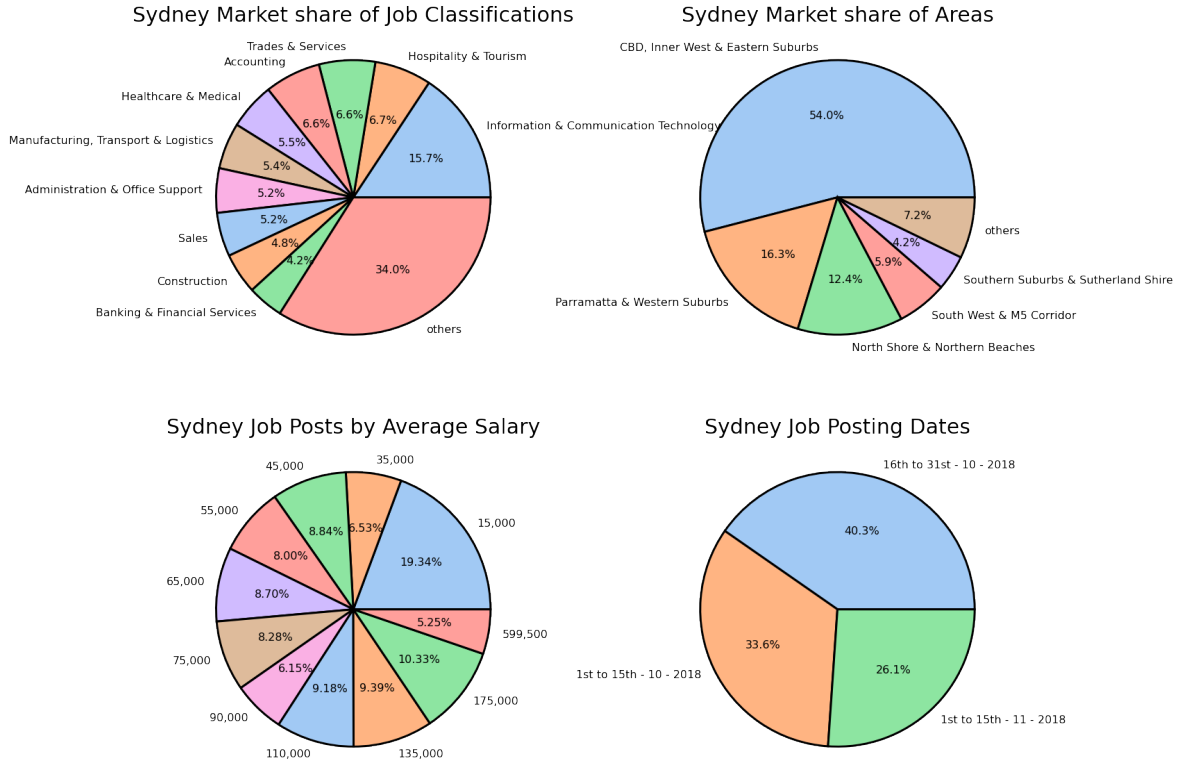

Figure 8: Sydney Locational Analysis

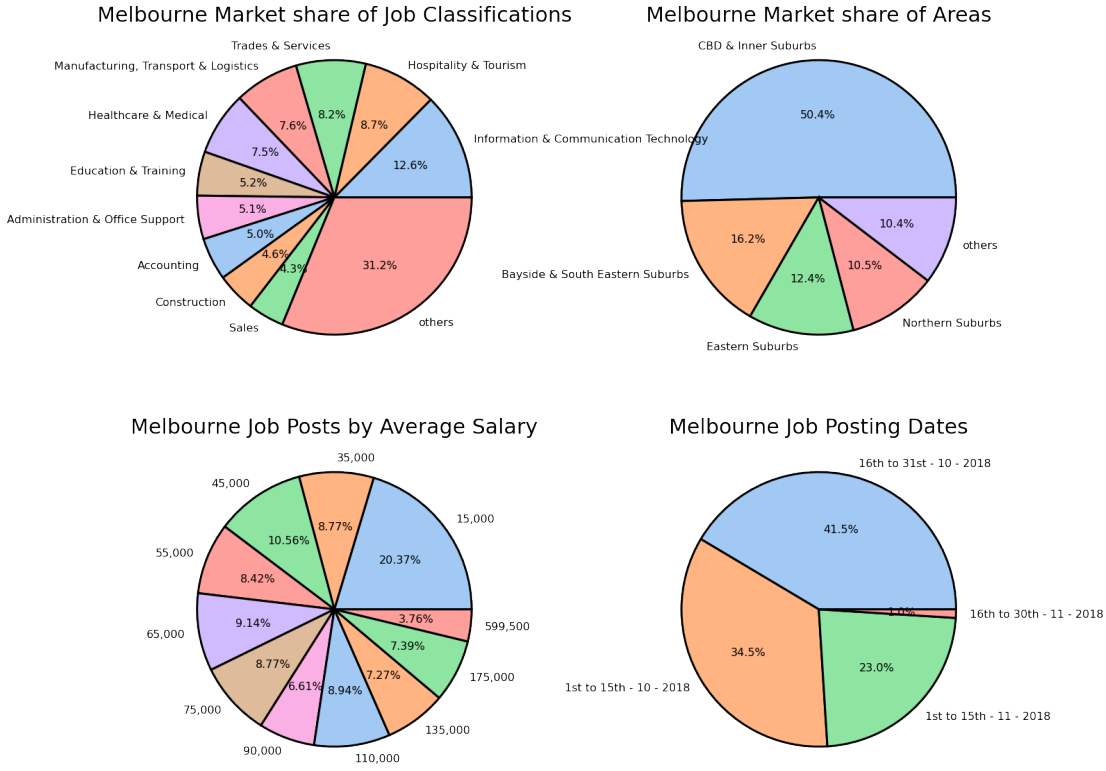

Figure 9: Melbourne Locational Analysis

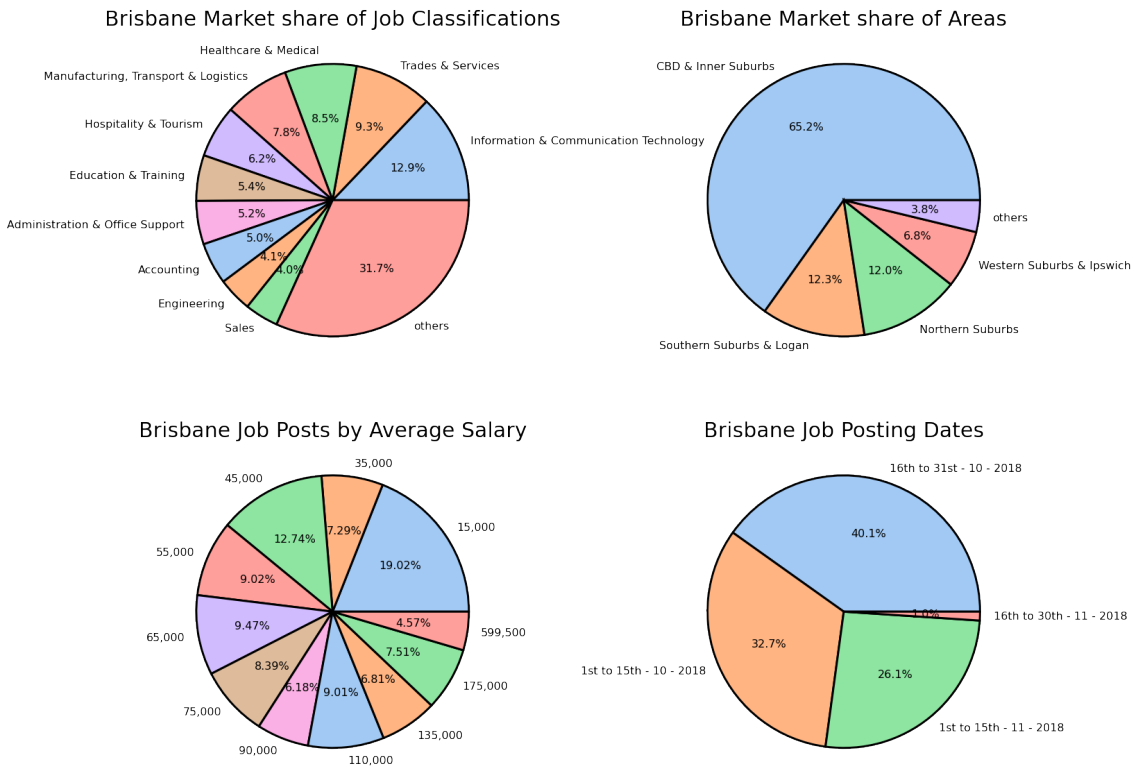

Figure 10: Brisbane Locational Analysis

ADD MORE POINTS TO PIE CHARTS BY CHANGING CUT OFF NUMBER FOR OTHERS.

What is the market size in each city? Which are the hottest job sectors in each city? Somewhere compare the max percent of sal with another city

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.

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.

2.3 Studying the Market by Sectors

As seen in

2.4 Interactive Visualisation of Results

3 Evaluation

3.1 Findings of Data Analytics

3.2 Balancing the Market

3.3 Refining the Data Analytics

3.4 Implications for Employers and Employees

4 Case Studies

4.1 Case Study 1

4.2 Case Study 2