**Griffith School of Information and Communication Technology**
**Griffith University**

**3803ICT - Big Data Analysis**
**Trimester 1, 2022**

# Lab Report: Job Market Analysis

**Jessy Barber, s5138877**

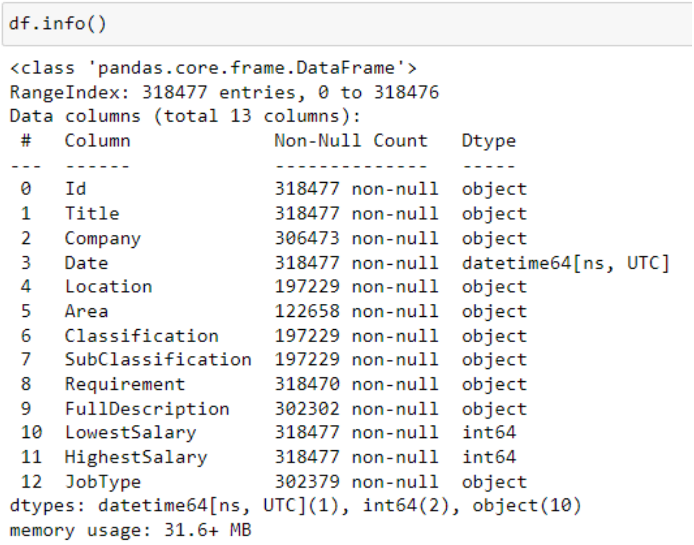
*A report submitted in partial fulfilment of the degree of Bachelor of Electrical Engineering (Honours) / Bachelor of Computer Science*

# 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
```

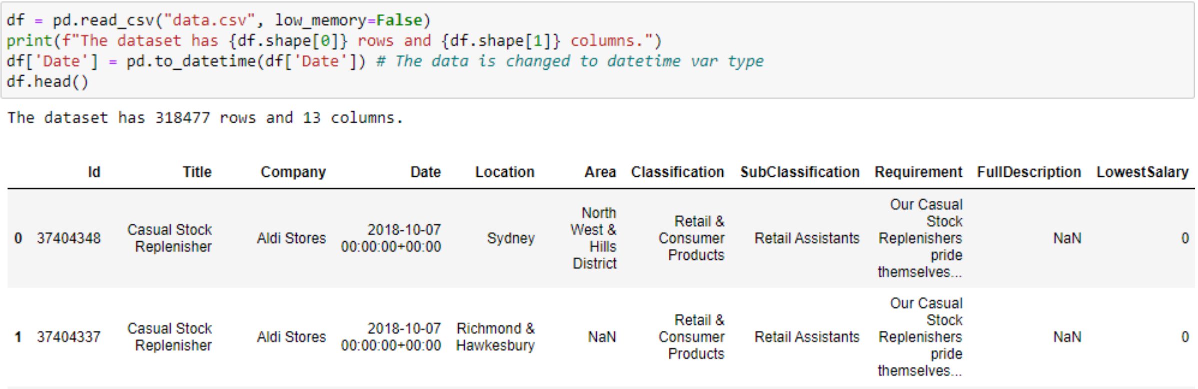
Figure 1: Categories / Domains of the Dataset

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

The data is loaded using pandas as per the following code:



```
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.

HOW DO YOU NORMALISE THE DATA:

HOW DO YOU CLEAN THE DATA:

## 1.3 Hypothesis About the Analysis Outcome

# 2 Data Analysis and Interpretation

## 2.1 Studying the Job Meta Data / Attributes

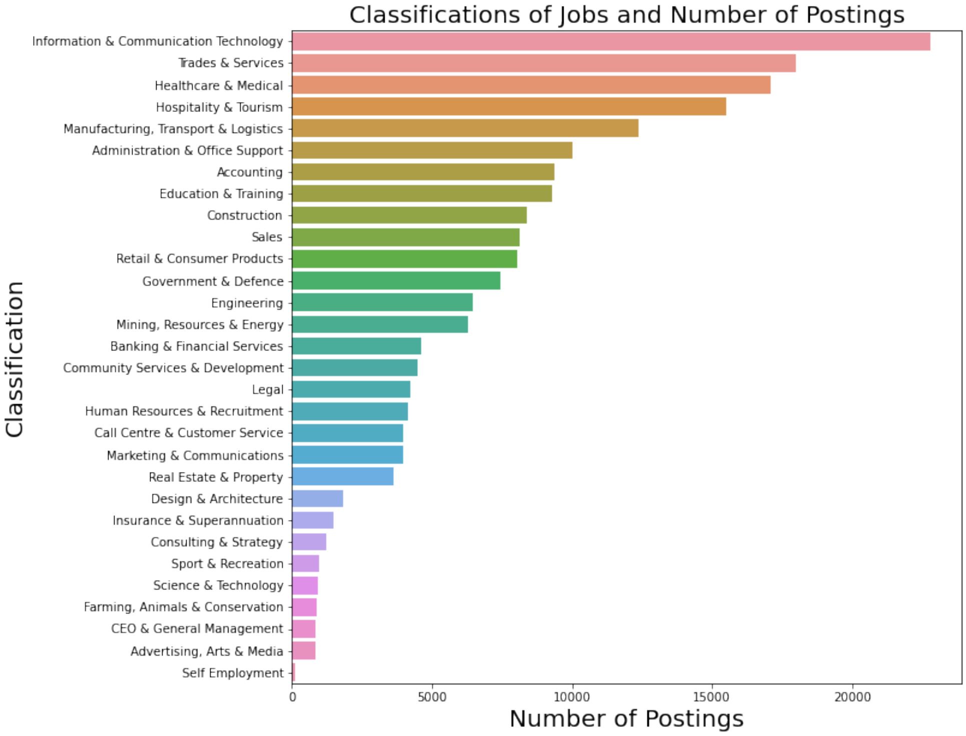


Figure 3: Classification of Jobs and Number of Postings

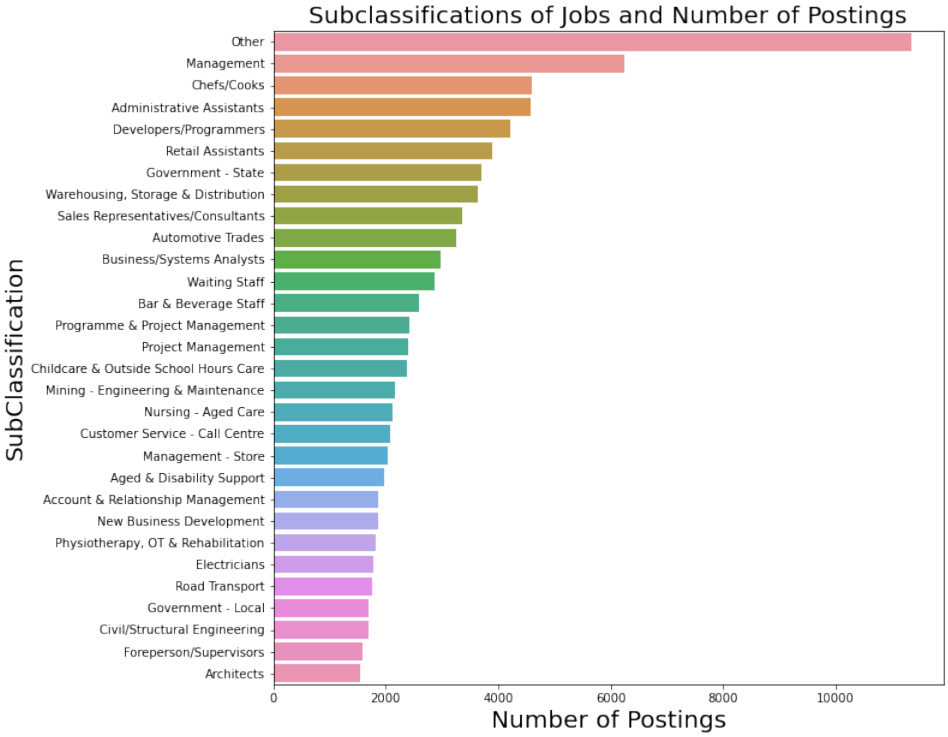


Figure 4: Subclassification of Jobs and Number of Postings

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.

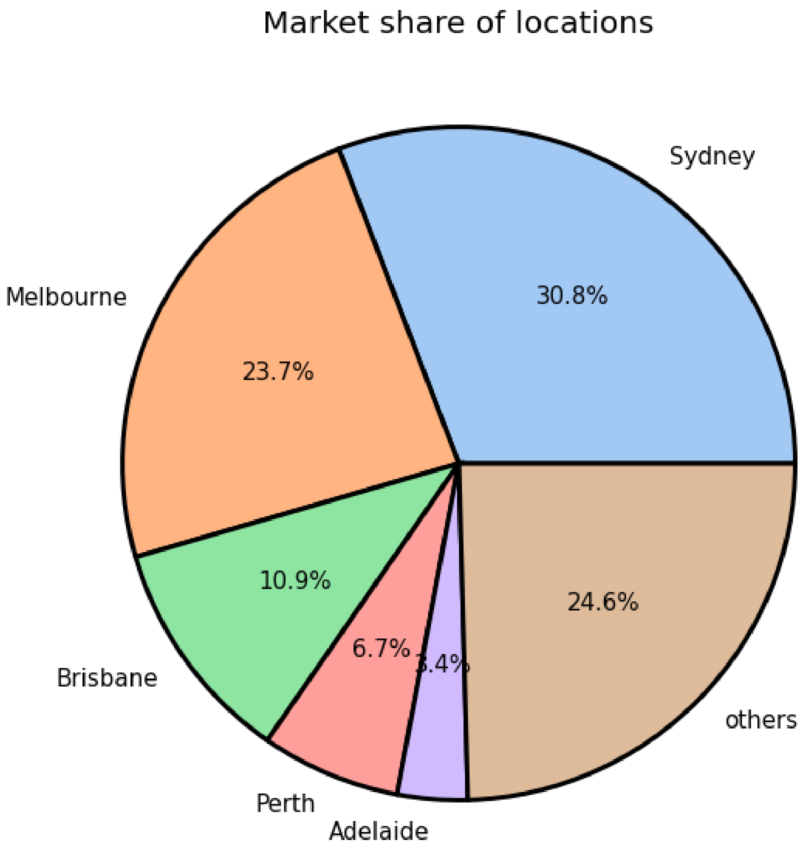


Figure 5: Top Five Cities for Market Share

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%.

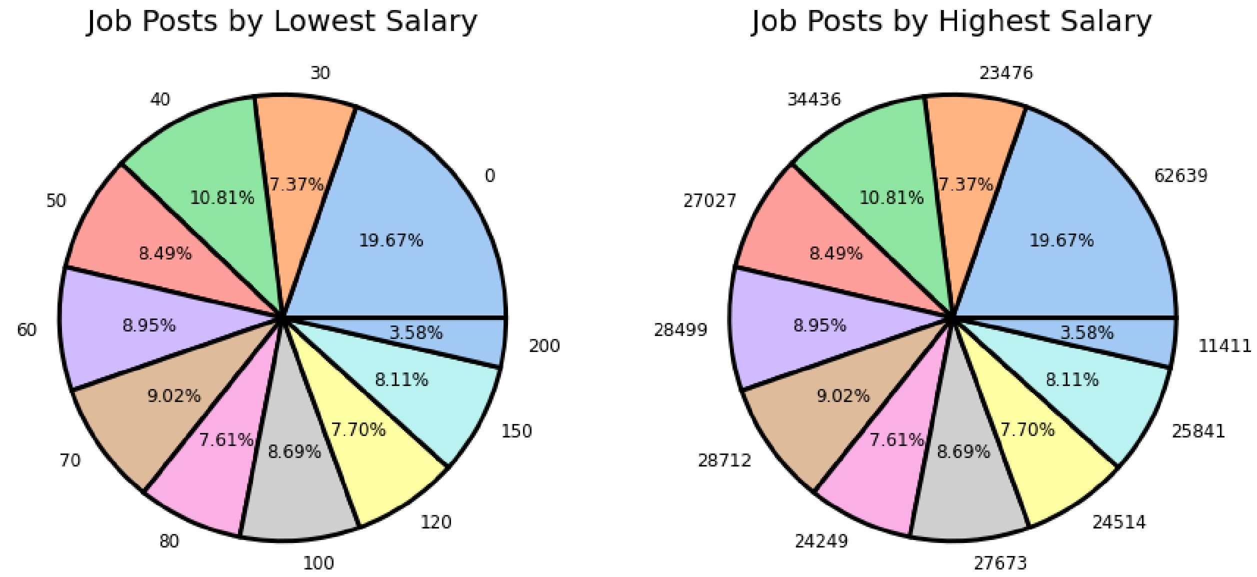


Figure 6: Salary Ranges of Jobs

Figure 6 shows the highest and lowest salary ranges for each job. WHAT DOES 0 REPRESENT -¿ UNEMPLOYMENT OR NAN. TALK ABOUT TOP LOWEST AND TOP HIGHEST SALARIES.