Job Market Analysis Report

Daniel Spindler - s5238197

Gabriela Almeida Monteiro - s5198626

Julio Pimentel Albores – s5172620

Griffith University

7030ICT Introduction to Big Data Analytics

Lecturer: Dr Henry Nguyen

Assessment item: Written Assignment

Word count: 1,115

Due date: 03/08/2020

Part 1 - Data Preparation and Pre-processing

1. Describe the Dataset

This report aims to analyse a dataset of the SEEK job market, which is an Australian platform where companies can post job announcements and job seekers can apply for a position. The original dataset subject of this analysis is structured with 13 columns and with attributes to identify each posted position appropriately. Some of these attributes are the title of the position, the name of the company that opened the position, the date when the ad was published, the employment location, the classification and subclassification of the type of job, among others. Classification refers to the sector of work (e.g. accounting, retail, information technology, and law). The subclassification refers to the specific areas within the classification (e.g. in information technology, there is a help desk, and developer).

Apart from the 13 columns, the dataset has 149,999 rows which makes a total number of 1,949,987 records and covers a period of 6 weeks and one day, starting from the 1st of October 2018 and finishing on the 13th of November 2018. Within this period, 44 different dates were identified for the job postings. The original dataset has 65 unique locations with Sydney representing almost 31% of all posts (46,357 posts in total), which is by far the location with the most openings.

The dataset has 30 different job classifications and Information, and Communication Technology occupies the first place with 16,661 job posts. In the second place there is Trades and Services with 14,125 posts and in the third place is Healthcare and Medical with 12,515 posts.

The sector of Information and Communication Technology, which was chosen for a more in-depth analysis in this report, accommodates 22 subsectors, with Developers/Programmers the one with the highest number of posts (3,069 posts). The minimum salary for this sector is zero, and the maximum salary is $250,000 per annum. Regarding these two values, the following assumptions are made: It is assumed that some companies opt to not include the salary in their job advertisement and that is the reason for zero values in the dataset. Additionally, since there is no further specification about this value in the dataset, it is assumed that 250 refers to $250,000, which would be the maximum annual income for some positions.

2. Normalise and Clean Data

Before analysing the data, it was necessary to do some cleaning and make some corrections. Most columns were of the generic “object” type, noting that the “Date” column needed the correct data type to be assigned. Using the correct datatype allowed for more accessible analysis and helped to ensure the data was valid.

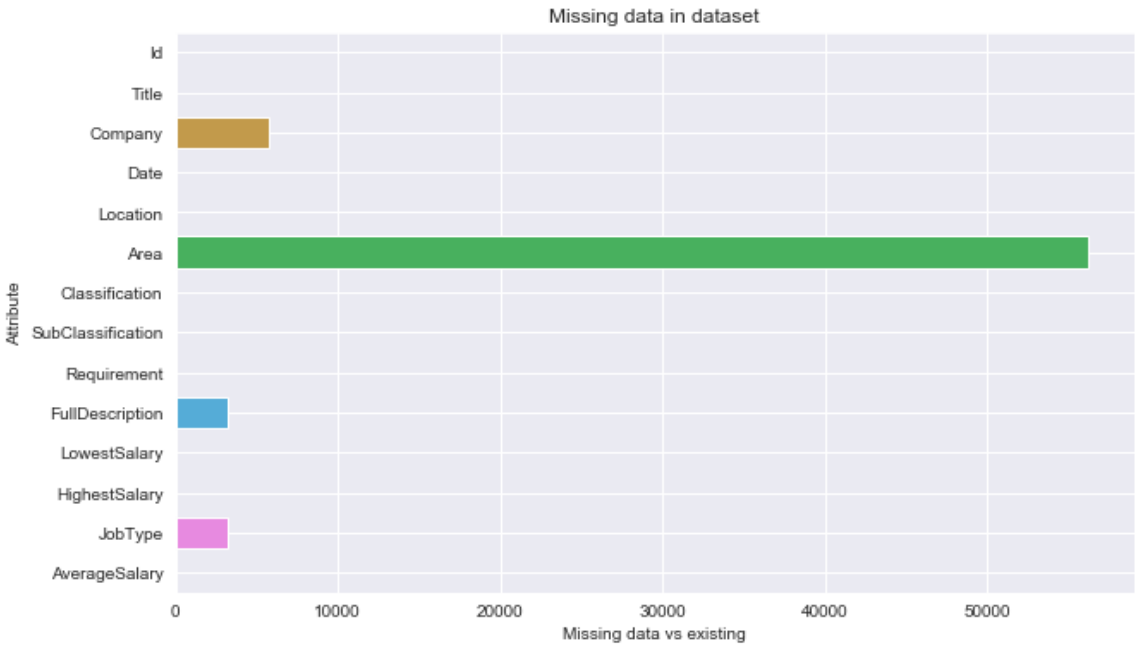
The “HighestSalary” and the “LowestSalary” columns were used to create a new column, of a numeric type, showing the Average Salary, allowing for easier salary comparison across the jobs. It should be noted that this brings the column count to 14.

The “Id” column datatype allows only numerical data, and the range of values show that all values are eight digits. Therefore, there are no inconsistencies in the “Id” column. The original dataset had 149,999 unique rows, with an “Id” value range of 915,340, indicating many values are not consecutive.

Date format was standardised, using python DateTime datatype with only the date portion of the existing dataset used. It is important to note that “Date” was formatted correctly, “Id” was found to be numeric and consistent, the salary data was numeric, with all other columns being composed of free text strings.

Duplicate data was found by not including the unique “Id” field in duplicate analysis and including all other columns. A new data frame was created containing all copies of duplicated rows, from this, 611 duplicated rows were observed, that should be removed from the original dataset. Interestingly, the new data frame contained 1119 rows, of which 508 were unique and 611 duplicates. The original query was then used to drop duplicates in place in the original dataset data fame – effectively reducing the row count to 149,388. Duplicate count checks confirmed duplicate issue resolved.

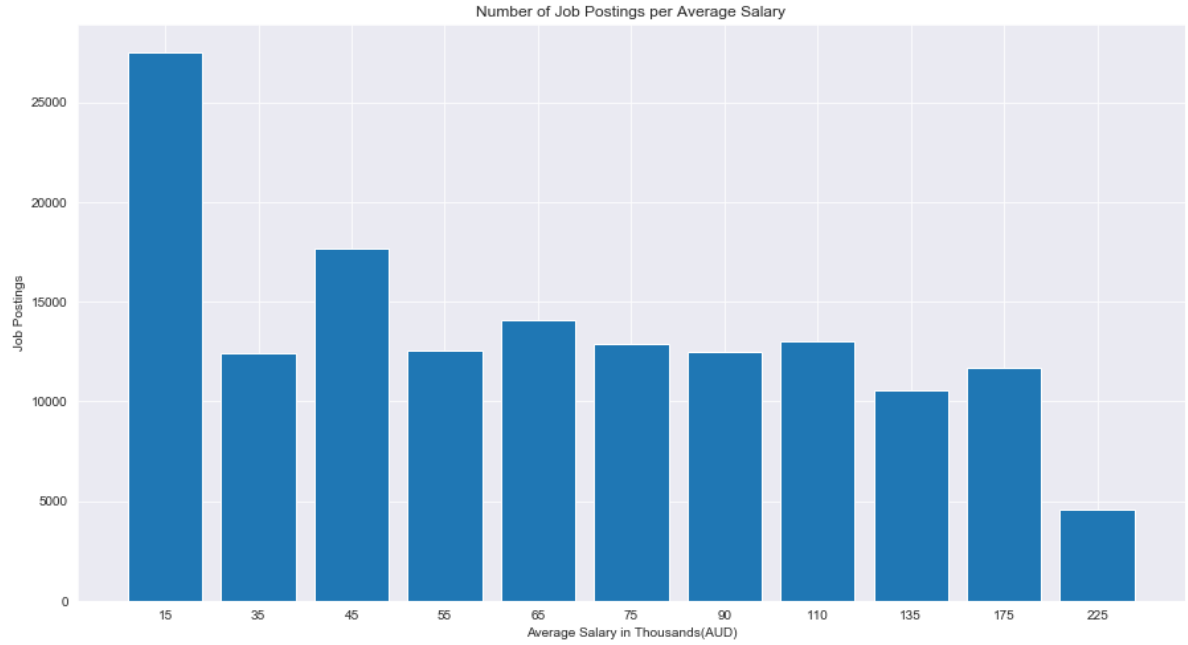
A horizontal bar chart was chosen to display the presence of null or missing values. The “Area” column contains a significant number of missing values. “Area” is a subcategory of “location” and therefore not an issue within the dataset. “Company” is likely absent due to some companies preferring not to display their company name.



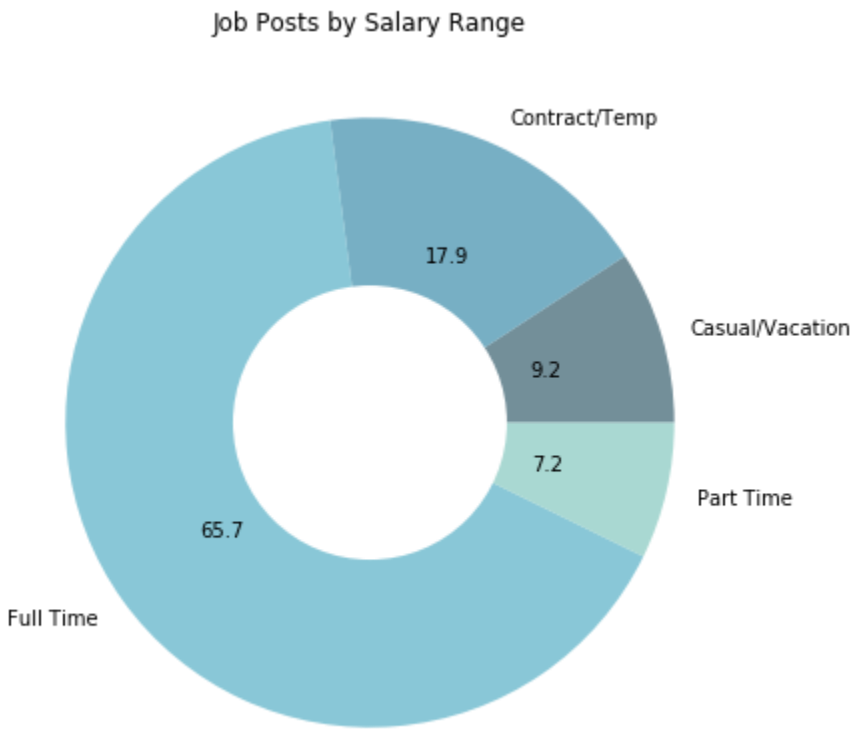
Part 2 - Data Analysis and Interpretation

The histogram was selected to analyse the average salary distribution by job postings. In this graph, it can be inferred that the average salary with more job postings is $15,000, and the average salary with fewer job postings is $225,000. It is expected behaviour in the job market because of actual organisation structures require more employees in the base and fewer employees in the top.

Another insight from this graph is that the average salary from $35,000 to $175,000 has similar job postings, except for the average salary of $45,000.

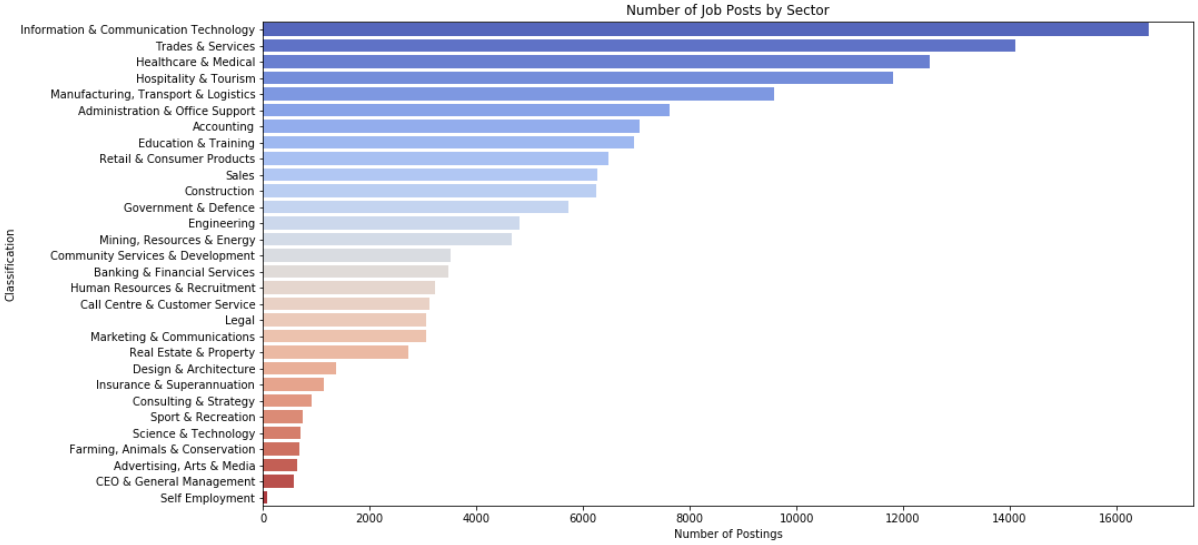


The pie chart was selected to analyse the distribution of job postings by job type. The “Full Time” category has more job postings, and the “Part-Time” category has fewer job postings in this dataset.

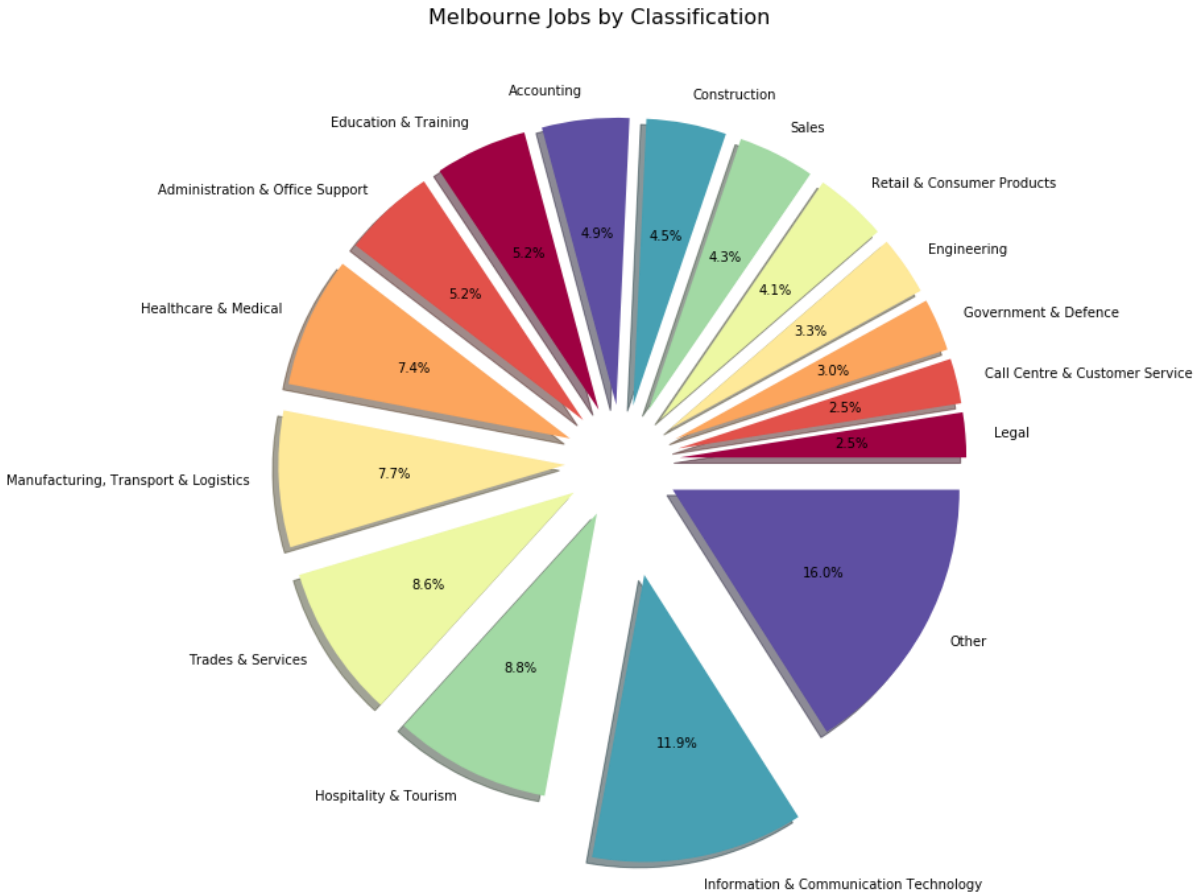


The bar chart was selected to analyse the job posts by sector. The top five sectors in this dataset are "Information & Communication Technology", "Trades & Services", "Healthcare & Medical", "Hospitality & Tourism", and "Manufacturing, Transport & Logistics". The bottom five sectors in this dataset are "Self-Employment", "CEO & General Management", "Advertisement, Arts & Media", "Farming, Animals & Conservation", and "Science & Technology".

Job postings of “CEO & General Management” seem to have a relationship between the frequency of job postings with an average salary of $225,000.



Melbourne was the location selected to visualise the market share in a pie chart. The top five sectors in Melbourne are "Information, Communication & Technology", "Hospitality & Tourism", "Trade & Services", "Manufacturing, Transport & Logistics", and "Healthcare & Medical". The top five sectors in Melbourne are consistent with the top five sectors in Australia, with minor differences in the order of the ranking. Sectors with little significance in the dataset are grouped as “Other”.



The box-and-whisker plot was selected to analyse the salary distribution for the top 30 locations for job posts. The locations “Port Headland, Karratha & Pilbara”, “Kalgoorlie, Goldfields & Esperance” and “ACT” have the most significant disparity between the highest salaries. The first two locations are likely due to mining-related jobs, whereas the last location offers Federal Government employment. On the other hand “Sunshine Coast”, “Gold Coast”, and “Mornington Peninsula & Bass Coast” have a lower disparity with the average salary, in part related to a larger tourism sector and a lack of white collar industry.