Part 1 – Data Preparation and Pre-processing.

1. Describe the dataset.

The job posting website seek was used to source data for the report. The data was given in CSV format and initial investigation of the data is shown in the table below.

Data Column Name	Type	Variations
Id	int64	149999
Title	object	92115
Company	object	26318
Date	object	44
Location	object	65
Area	object	19
Classification	object	30
SubClassification	object	338
Requirement	object	121324
FullDescription	object	127676
LowestSalary	int64	11
HighestSalary	int64	11
JobType	object	4

When we look at the whole dataset there are 149999 records. The sheer number of records indicates that the dataset is quite large.

As the dataset contains a date column the range can be identified with the period starting on the 1/10/2018 to the 13/11/2018 covering 43 days.

1/10/2018	1023
2/10/2018	1980
3/10/2018	4945
4/10/2018	8506
5/10/2018	5711
6/10/2018	471
7/10/2018	1037
8/10/2018	4995
9/10/2018	6329
10/10/2018	8337
11/10/2018	1221
12/10/2018	1324
13/10/2018	196
14/10/2018	239

15/10/2018	5210
16/10/2018	6267
17/10/2018	13059
18/10/2018	8121
19/10/2018	5989
20/10/2018	422
21/10/2018	1145
22/10/2018	688
23/10/2018	827
24/10/2018	835
25/10/2018	1205
26/10/2018	1701
27/10/2018	232
28/10/2018	291
29/10/2018	4656
30/10/2018	8179
31/10/2018	8641
1/11/2018	1316
2/11/2018	2874
3/11/2018	499
4/11/2018	636
5/11/2018	7778
6/11/2018	3654
7/11/2018	6947
8/11/2018	1199
9/11/2018	3377
10/11/2018	743
11/11/2018	343
12/11/2018	5865
13/11/2018	986

A brief overview of the location contained in the data set shows there is 65 unique locations which are located around Australia. The top location with the most job postings is Sydney with 46357.

In the period for the dataset, there are 30 job sectors. The table below shows each sector the the total of job postings.

Information & Communication Technology	16661
Trades & Services	14125
Healthcare & Medical	12515
Hospitality & Tourism	11818
Manufacturing, Transport & Logistics	9608
Administration & Office Support	7636
Accounting	7075
Education & Training	7033
Retail & Consumer Products	6496
Sales	6281
Construction	6254
Government & Defence	5926
Engineering	4812
Mining, Resources & Energy	4679
Community Services & Development	3528
Banking & Financial Services	3481
Human Resources & Recruitment	3233
Call Centre & Customer Service	3127
Legal	3071
Marketing & Communications	3052
Real Estate & Property	2729
Design & Architecture	1379
Insurance & Superannuation	1144
Consulting & Strategy	905
Sport & Recreation	750
Science & Technology	696
Farming, Animals & Conservation	683
Advertising, Arts & Media	644
CEO & General Management	585
Self Employment	73

The report focused on the job sector - Information & Communication Technology and the total number of job postings for each subclassification is shown below.

Subclassification	
Developers/Programmers	3069
Business/Systems Analysts	2076
Programme & Project Management	1665
Architects	1110
Engineering - Software	1087
Help Desk & IT Support	1074
Networks & Systems Administration	927
Consultants	849
Other	742
Testing & Quality Assurance	606
Management	561
Security	524
Engineering - Network	402
Database Development & Administration	395
Web Development & Production	351
Sales - Pre & Post	349
Product Management & Development	281
Telecommunications	269
Team Leaders	141
Engineering - Hardware	94
Technical Writing	77
Computer Operators	12

The number of job postings for the salary ranges is shown in the table below.

Starting Salary Range	Highest Salary Range	
0	30	27606
30	40	12441
40	50	17708
50	60	12559
60	70	14108
70	80	12932
80	100	12582
100	120	13062
120	150	10651
150	200	11738
200	250	4612

The data contains a categorical value for job type and it includes:

- Casual/Vacation
- Contract/Temp
- Full Time
- Part Time

A closer look at the lowest and highest salaries for each job types was determined by both tables shown below.

	Lowest Salary							
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
JobType								
Casual/Vacation	13496	39.68213	30.07874	0	30	40	50	200
Contract/Temp	26238	80.7108	59.06716	0	40	70	120	200
Full Time	96410	66.42454	47.87934	0	40	60	100	200
Part Time	10622	41.59198	31.49745	0	30	40	50	200

	Highest Salary							
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
JobType								
Casual/Vacation	13496	55.04594	31.16269	30	40	50	60	250
Contract/Temp	26238	106.6857	69.67738	30	50	80	150	250
Full Time	96410	87.69059	54.5322	30	50	70	120	250
Part Time	10622	56.63246	33.05726	30	40	50	60	250

The analysis shows the minimum salary in the lowest salary range is $0 and the highest salary for each job type is $250k.

2. Normalize and clean data. (7 points)

In this section, the data was normalised and cleaned. The average salary was calculated from the data using the lowest salary column and highest salary column. Using the simple average calculation, the result was calculated using the following formula.

$$Average\ Salary = \frac{(Highest\ Salary + Lowest\ Salary)}{2}$$

The raw data showed there was no anomalies with the ‘Id’ column and it used 8 number long integers.

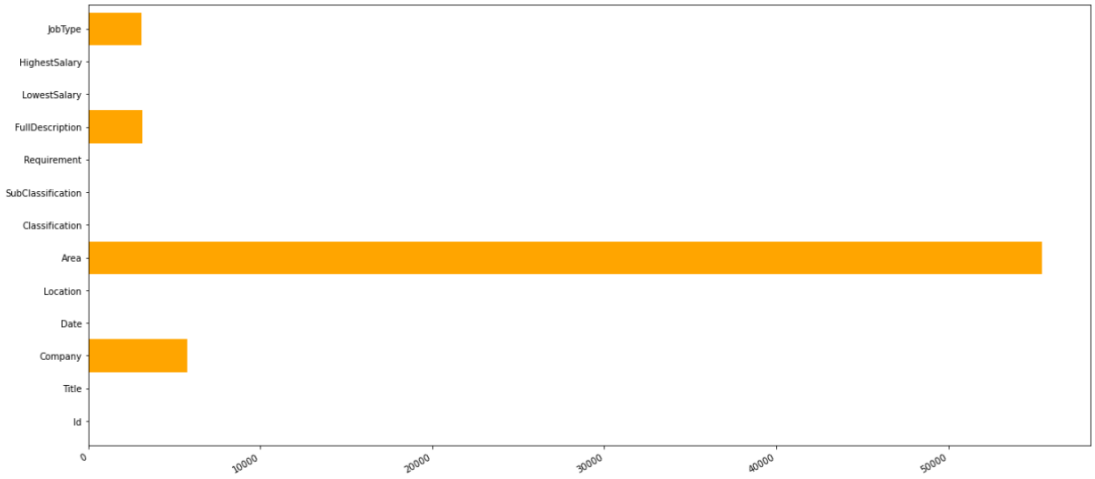
A requirement for the data was to remove the time component from the date column and leave the date element of the date value e.g. DD/MM/YYYY TT:TT:TT to DD/MM/YYYY format.

The next stage converted the “Date” column to DateTime datatype.

Duplicate data was reviewed with a bit of conjecture. If we look at the whole data set and compare each record against each other there is no duplicate data.

However, the method used in the report was to consider jobs with the same Date, Company, Location, Area, Requirement and FullDescription as duplicates. Using this assumption there was 2591 duplicates and they were removed from the dataset.

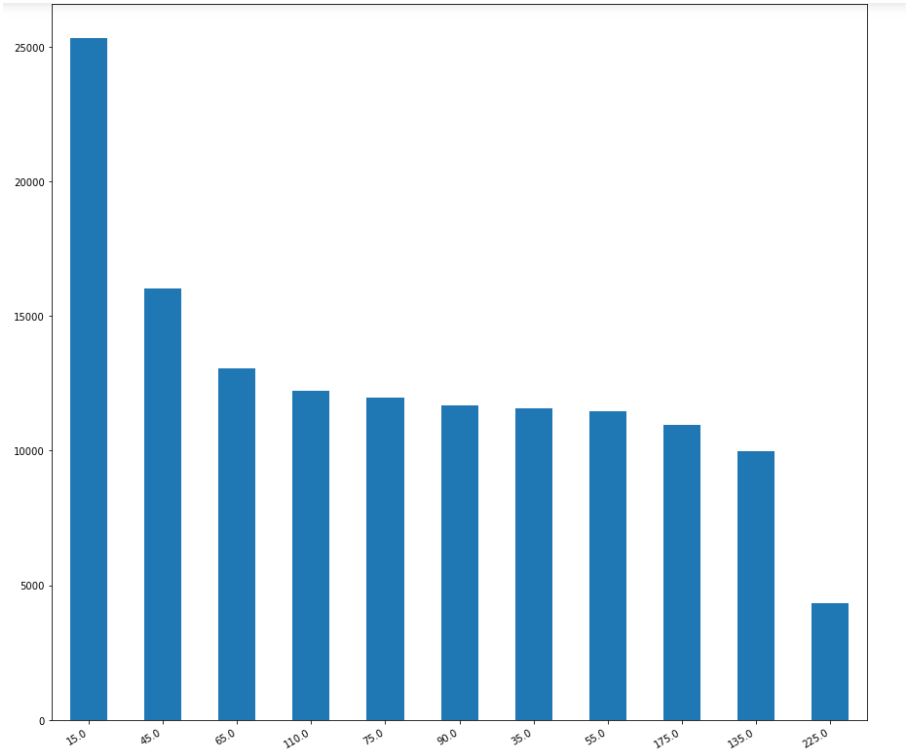
The investigation found there was a large amount of missing data for the columns job type, full description, area and company (see the graph below). To make the data useable, job postings with missing data were removed for columns including job type, full description, and company. Jobs with no suburb information were kept.



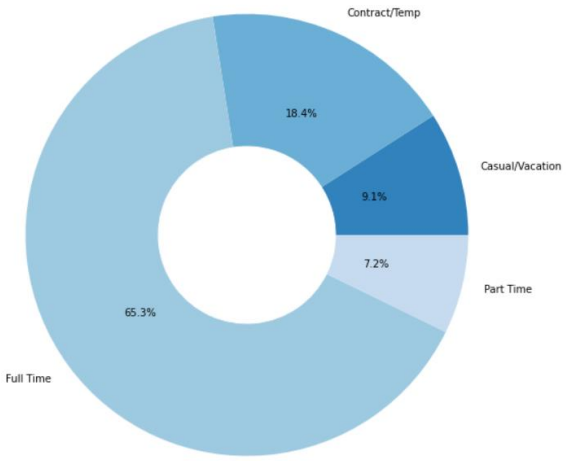
Part 2 – Data Analysis and Interpretation. [5 points]

In this part of the report, the data analysis and interpretation were completed.

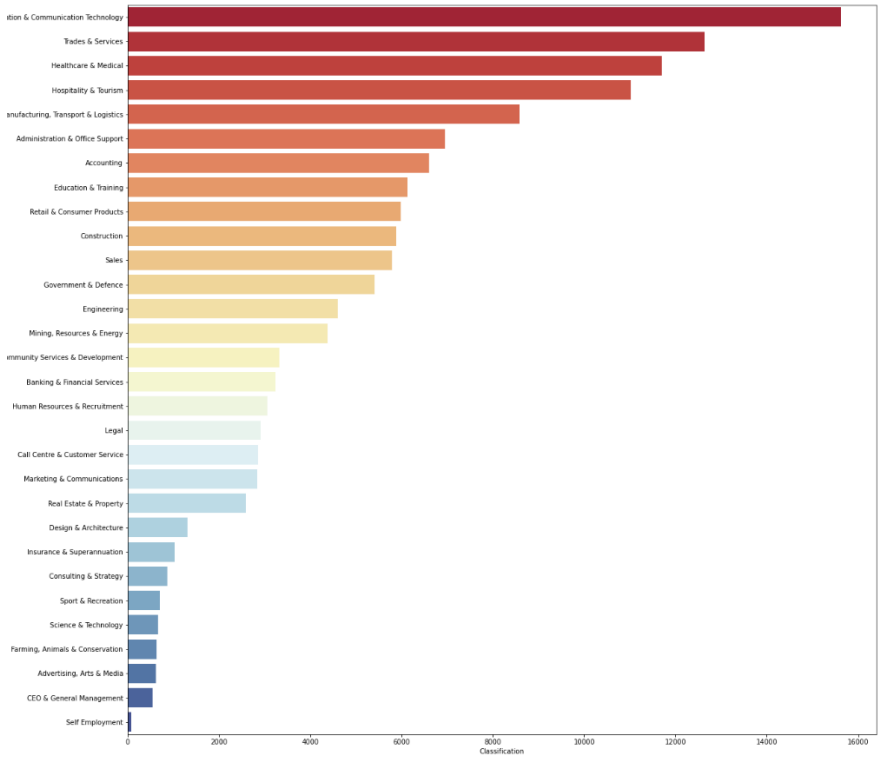
The following graph shows the total jobs for each average salary range in the bar chart.



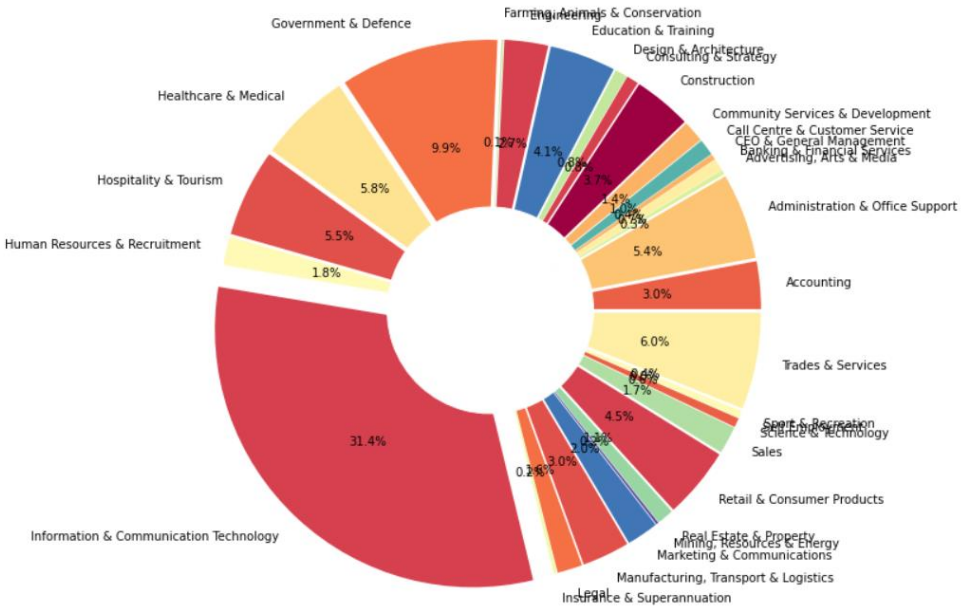
A pie chart was created to show a breakdown of the total jobs by job type.



The next part of the investigation was to use a horizontal bar chart to show the number of jobs by job classification.



The next part of the analysis was to create a pie chart looking at the breakdown of job classifications for the ACT.



Lastly the salary distribution for the top 30 cities by job posting was found using a boxplot (see below).