3803ICT

Trimester 1, 2023

DATA ANALYTICS: ASSIGNMENT 1
JOB MARKET ANALYSIS

Student: James Hudson

Student No: s5182091

Student: Todd Cooper

Student No: s2681289

Contents

Table of Figures

Table of Tables

Data Preparation and Preprocessing

Dataset Description

The *data.csv* file comprises nearly 320,000 job advertisements gathered from diverse locations across Australia. These advertisements were obtained from the renowned employment marketplace website "seek.com.au" over a span of six months, from October 2018 to March 2019. The data is organised in a tabular format, featuring structured rows and columns that establish relationships between cells. In total, the dataset consists of 318,477 entries distributed across 13 columns. Each job entry provides essential details, which are presented as Categories in Table 1.

Table 1. Data.csv categories and properties

Category	Description	Type	Range	Properties	Domain size
Id	An identification number	Int64	31671087 to 38566133	Ordinal	318477
Title	The job posting title	String	2 to 80 characters	Discrete	168065
Company	Company posting job	String	2 to 80 characters	Discrete	40628
Date	Date of posting	Datetime64	2018-10-01 to 2019-03-13	Ordinal	163
Location	Location of posting	String	3 to 36 characters	Categorical	65
Area	Sub-location of posting	String	15 to 35 characters	Categorical	19
Classification	Job sector classification	String	5 to 38 characters	Categorical	30
SubClassification	Job sub-sector classification	String	3 to 47 characters	Categorical	338
Requirement	Requirements of the Job	String	1 to 150	Discrete	234288
FullDescription	Full description on job posting	String	8 to 201580	Discrete	250901
LowestSalary	The low end of salary range	Int64	0 to 200	Categorical data	11
HighestSalary	The high end of salary range	Int64	30 to 999	Categorical data	11
JobType	The type of employment, fulltime ect.	String	9 to 15 characters	Categorical data	4

Among the categories, some contain empty data fields or undefined values. These are listed on the following page.

Table 2. NaN values in data.csv by category

Column Name	Number of NaN values	Percentage of Total Entries
Area	195819	61.5%
Location	121248	38.1%
Classification	121248	38.1%
SubClassification	121248	38.1%
FullDescription	16175	5.1%
Company	12004	3.8%
Requirement	7	0.002%

Location, Classification, and SubClassification categories are all empty of data from 2018-12-27 onward. Of the 13 categories, all except Id and Company will be used in the analysis. Id does not contain any semantic information of value. Company, while containing relevant information, is not particularly pertinent to our exploration of the data.

When deciding which attributes would be used for analysis, it was concluded that the Id column served no purpose other than to help prevent the addition of duplicate data. As other forms of duplicate validation existed, the Id column was deemed redundant. The Title and Company columns were additionally labelled as non-useful for statistical or predictive analysis. The Classification and SubClassification columns render the need for Job title as impractical, as they provided similar value of job classification with much less noise and complexity due to their broader values.

Loading Dataset using Pandas

Loading a dataset with pandas is a common task in data analysis and machine learning. Pandas is a powerful Python library that provides efficient data manipulation and analysis tools. To load the *data.csv* dataset with pandas, you can follow these steps:

1) Import the pandas library:

```
[ ]
import pandas as pd
Python
```

Here the pandas library is imported under the name *pd.* When using the pandas functions, this name will be utilised.

2) Identify the file format and location of your dataset. Pandas supports various file formats such as CSV, Excel, SQL databases, and more.

3) Use the appropriate pandas function to read the dataset. The most commonly used function is *pd.read_csv()* for reading CSV files. For example:

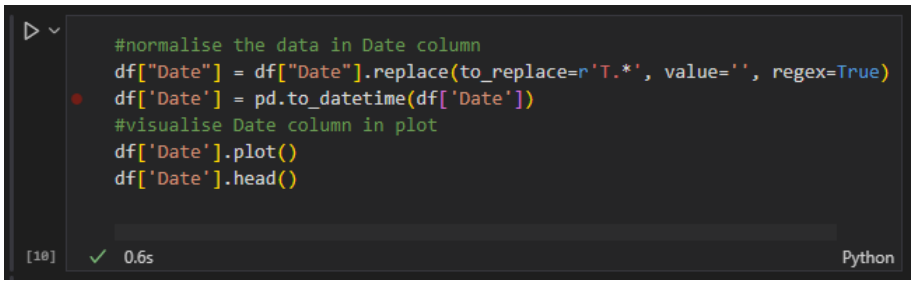
```
[ ]
df = pd.read_csv("data.csv")
Python
```

Here the *data.csv* file is imported as a pandas data frame and assigned to the variable *df (data frame).*

Data Normalisation & Cleaning

Normalisation a preprocessing step in data analysis and machine learning that aims to transform the features or variables in a dataset onto a similar scale. Normalisation is often performed to ensure that the features have comparable ranges and magnitudes, which can be beneficial for certain algorithms and models.

Normalise Values and Format encoding: Firstly, the Date column which contained the date that the job advertisement was posted needed to be normalised. The format chosen was as follows: *Year-Month-Day (YY/MM/DD).* The date data provided in the file contained redundant values for seconds as well as the standard date. The following code provided a solution:

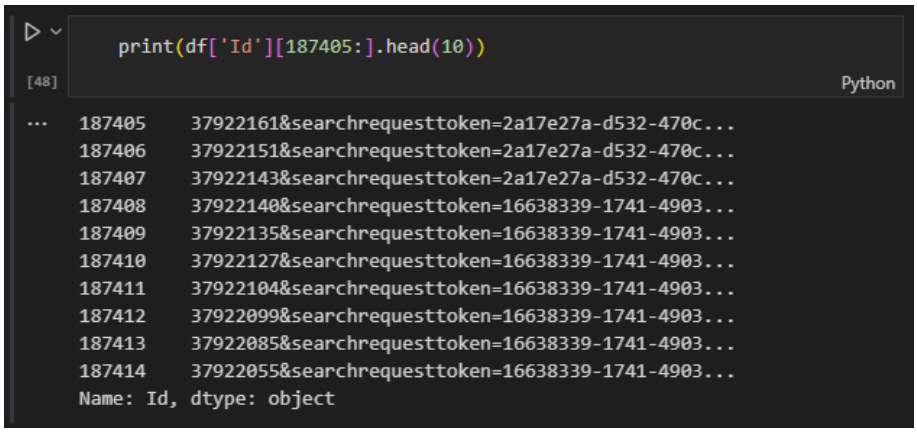


```
#normalise the data in Date column
df["Date"] = df["Date"].replace(to_replace=r'T.*', value='', regex=True)
df['Date'] = pd.to_datetime(df['Date'])
#visualise Date column in plot
df['Date'].plot()
df['Date'].head()
```

[10] ✓ 0.6s Python

The datatype was corrected to datetime64[ns, UTC] and seconds measurements were removed.

Secondly, the Id column needed to be converted to an Integer datatype. Upon doing so it was revealed that this column contained values other than integer numbers, which is invalid. From index 187,406 the Id column started to show errors.



```
print(df['Id'][187405:].head(10))
```

[48] Python

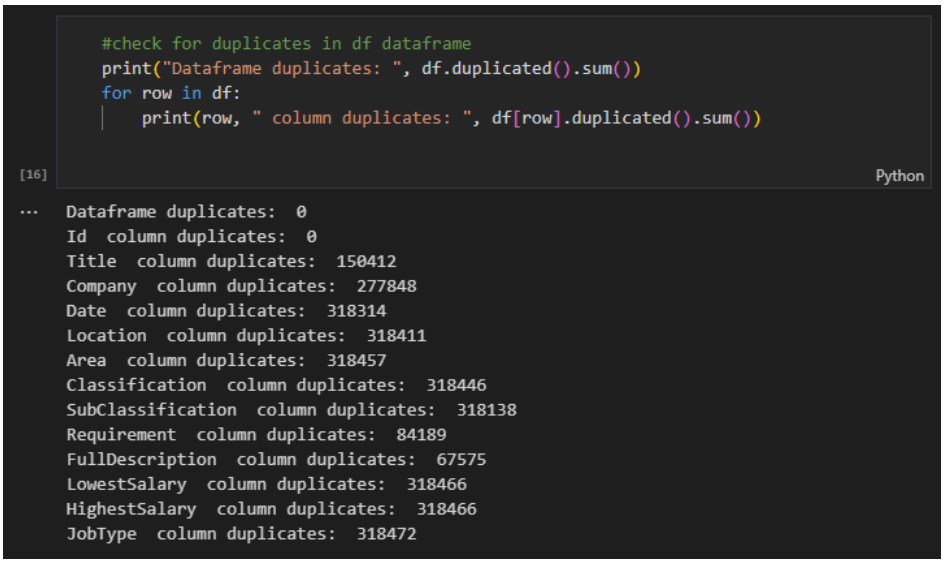
```
187405    37922161&searchrequesttoken=2a17e27a-d532-470c...
187406    37922151&searchrequesttoken=2a17e27a-d532-470c...
187407    37922143&searchrequesttoken=2a17e27a-d532-470c...
187408    37922140&searchrequesttoken=16638339-1741-4903...
187409    37922135&searchrequesttoken=16638339-1741-4903...
187410    37922127&searchrequesttoken=16638339-1741-4903...
187411    37922104&searchrequesttoken=16638339-1741-4903...
187412    37922099&searchrequesttoken=16638339-1741-4903...
187413    37922085&searchrequesttoken=16638339-1741-4903...
187414    37922055&searchrequesttoken=16638339-1741-4903...
Name: Id, dtype: object
```

This was corrected by removing the remaining Id sequence following the first encounter of a non-integer value. This method proved to be suitable as after, the Id column contained no duplicates and only integer values. The column was then set to a 64bit integer datatype.

Handle missing data: When checking the *info()* of the dataset we can view each column and the missing count of missing/non-missing values from them. The *data.csv* file revealed that Location, Classification, and Subclassification were missing 38% of their data entries. The Area column had the most missing values, with

61%. The values were unable to be corrected as removing these rows would dispose of important information seen in other columns. Instead, the problem was noted and will be accounted for when performing analysis involving them.

Remove Duplicates: When testing the dataset for duplicate entries, none were found. Additionally, each individual column was tested. The main concern would be duplicate entries within the "Id" column, as this is a unique identifier value or "key" value for each row. Testing also revealed no duplicate data being found within the Id column. Duplicates were found elsewhere, which is to be expected given the nature of the data.



```
#check for duplicates in df dataframe
print("Dataframe duplicates: ", df.duplicated().sum())
for row in df:
    print(row, " column duplicates: ", df[row].duplicated().sum())
```

[16] Python

```
Dataframe duplicates:  0
Id  column duplicates:  0
Title  column duplicates:  150412
Company  column duplicates:  277848
Date  column duplicates:  318314
Location  column duplicates:  318411
Area  column duplicates:  318457
Classification  column duplicates:  318446
SubClassification  column duplicates:  318138
Requirement  column duplicates:  84189
FullDescription  column duplicates:  67575
LowestSalary  column duplicates:  318466
HighestSalary  column duplicates:  318466
JobType  column duplicates:  318472
```

Removal of Redundant Columns: As mentioned above, the Id, Title, Company, and JobType columns provide no valuable information when performing the analytic tasks covered in this document. Due to this, they were dropped from the dataset in the cleaning stage. As a new file is created for the data once preprocessed, this discarded information may always be recovered if needed.

Hypothesis

It is hypothesized that the largest job markets will be in metropolitan areas, with Sydney and Melbourne, Australia's most populous cities, having the largest job markets. It is expected that salaries in major cities will be slightly higher than those in rural areas. Additionally, it is likely that technological sectors will see an upwards trend due to their increased demand in recent years which hasn't shown to slow down [1]. The Australian Bureau of Statistics published a statement about the financial performance of industries within Australia from the 2021-22 financial year, it showed that the Mining industry earnings grew $54.3b (32.7%) [2]. From this statistical information it seems reasonable to predict the mining sector will contain the highest paid job advertisements.

Data Analysis and Interpretation

Job Metadata

The dataset encompassed a diverse range of 31 distinct main job sectors. Within these primary sectors, a comprehensive total of 339 sub-sectors were identified. To highlight the prominent sectors that yielded significant influence in the market, the following emerged as the largest contributors:

- Information & Communication Technology (ICT)
- Trades and Services
- Healthcare and Medical
- Hospitality & Tourism
- Manufacturing, Transport & Logistics
- Administration and Office Support

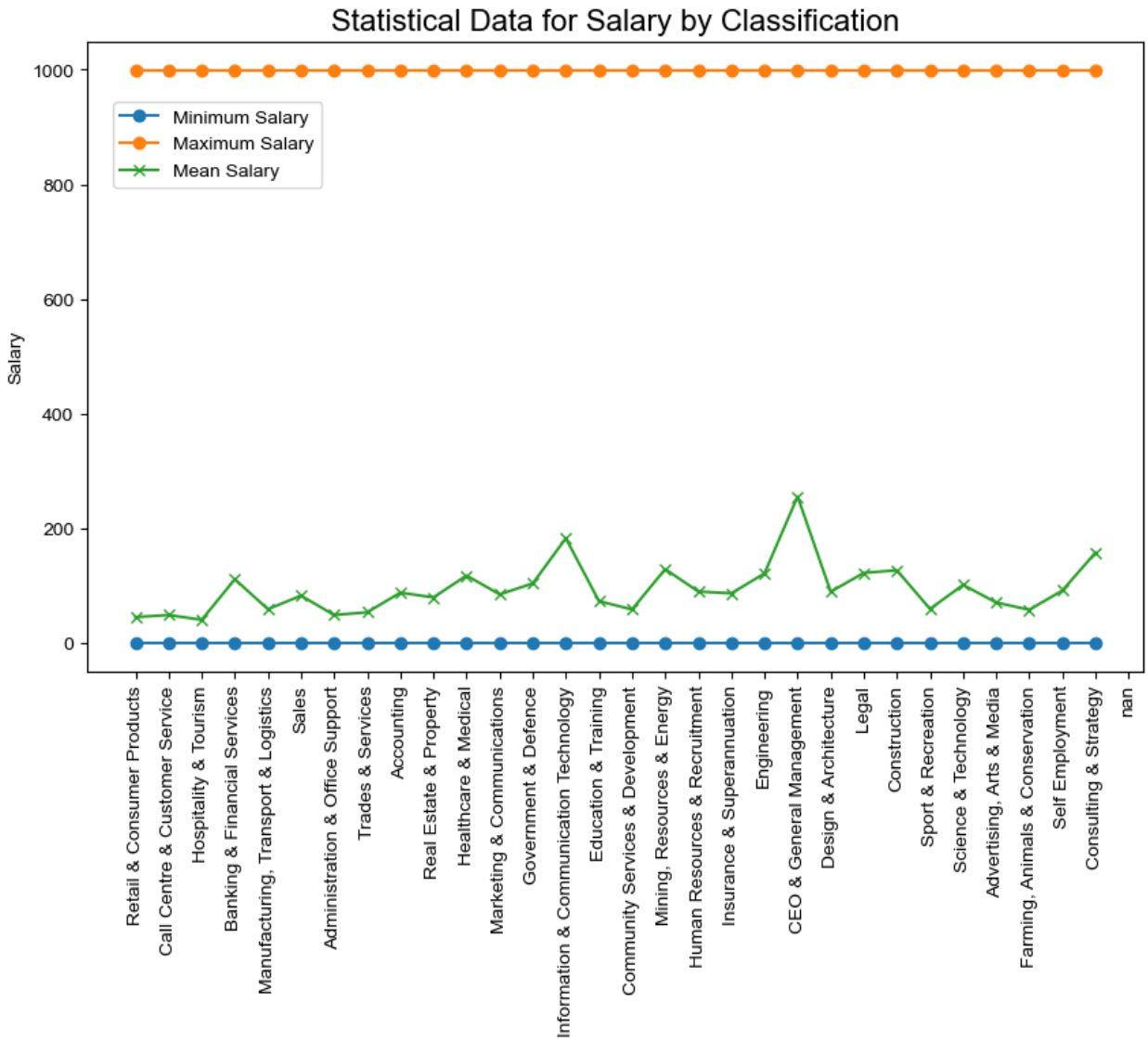
By disregarding the primary job sector from which the sub-sectors originate, we can identify the most significant sub-sectors that dominate the market. The following sub-sectors emerged as the largest in terms of size and influence:

- Management
- Chefs/Cooks
- Administrative Assistants
- Developers/Programmers
- Retail/Assistants
- Government – State

The locations of each job contained the 8 capital cities in Australia, with the addition of 58 regions of varying size totaling 66 unique locations. The distribution of jobs by location is shown in figure 1 below. Location of job advertisement was recorded with further accuracy due to the use of the column Area. This consisted of 20 unique values, the majority showing “CBD & Inner Suburbs” and “CBD, Inner West & Eastern Suburbs” as common Area values.

Analysing the salary ranges across different sector classifications reveals a consistent pattern. Each sector category exhibits the same salary range, with the minimum salary recorded as $0 and the maximum salary reaching $999k. Examining the mean salaries within each job classification allows us to identify the categories that generally offer higher pay. The results indicate that CEO & General Management, Information & Communication Technology, and Consulting & Strategy emerge as the top-paying sectors with the highest mean salaries in the market.

Figure 1. Salary Range by Classification

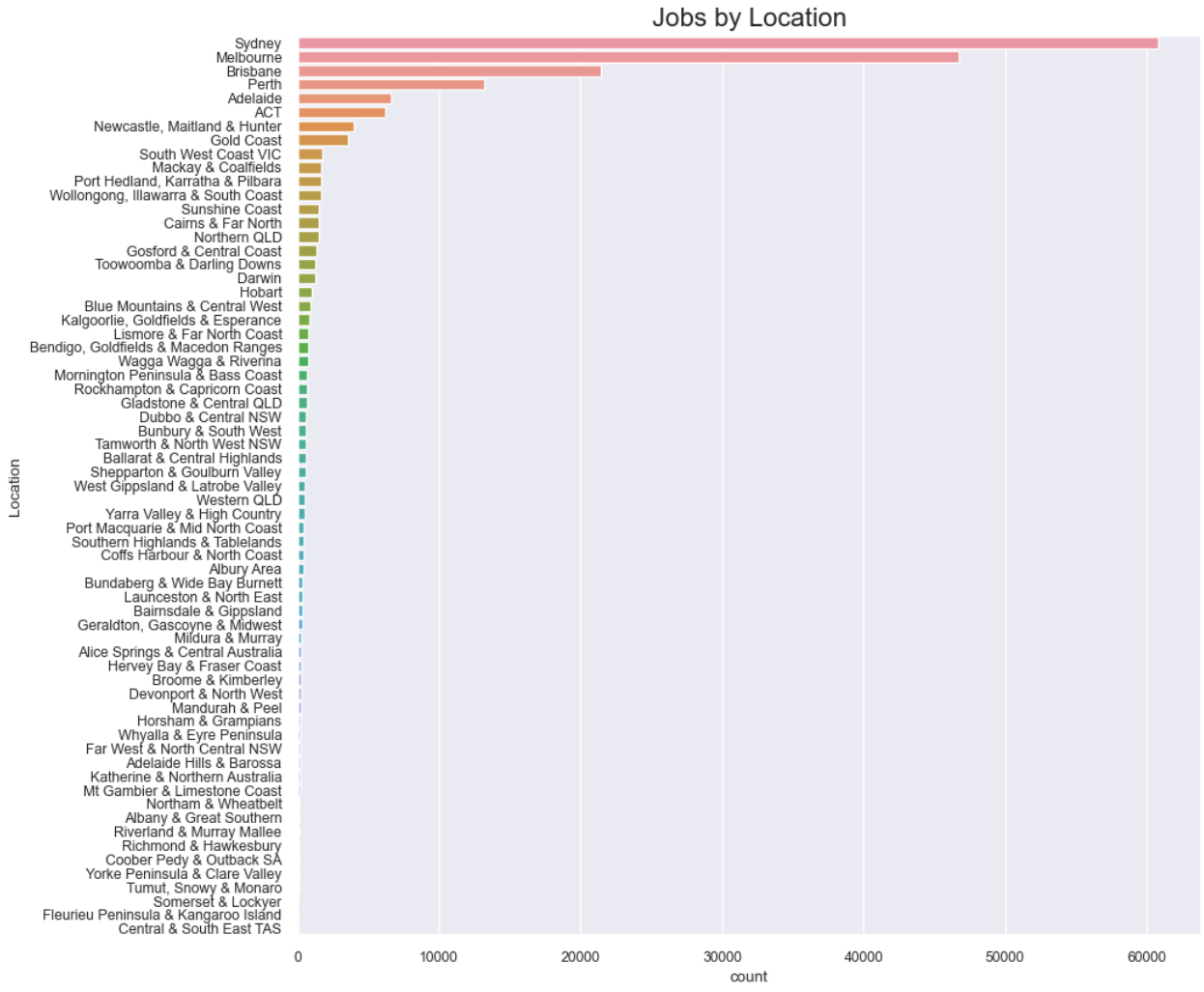


Market by Location

Geographical Distribution

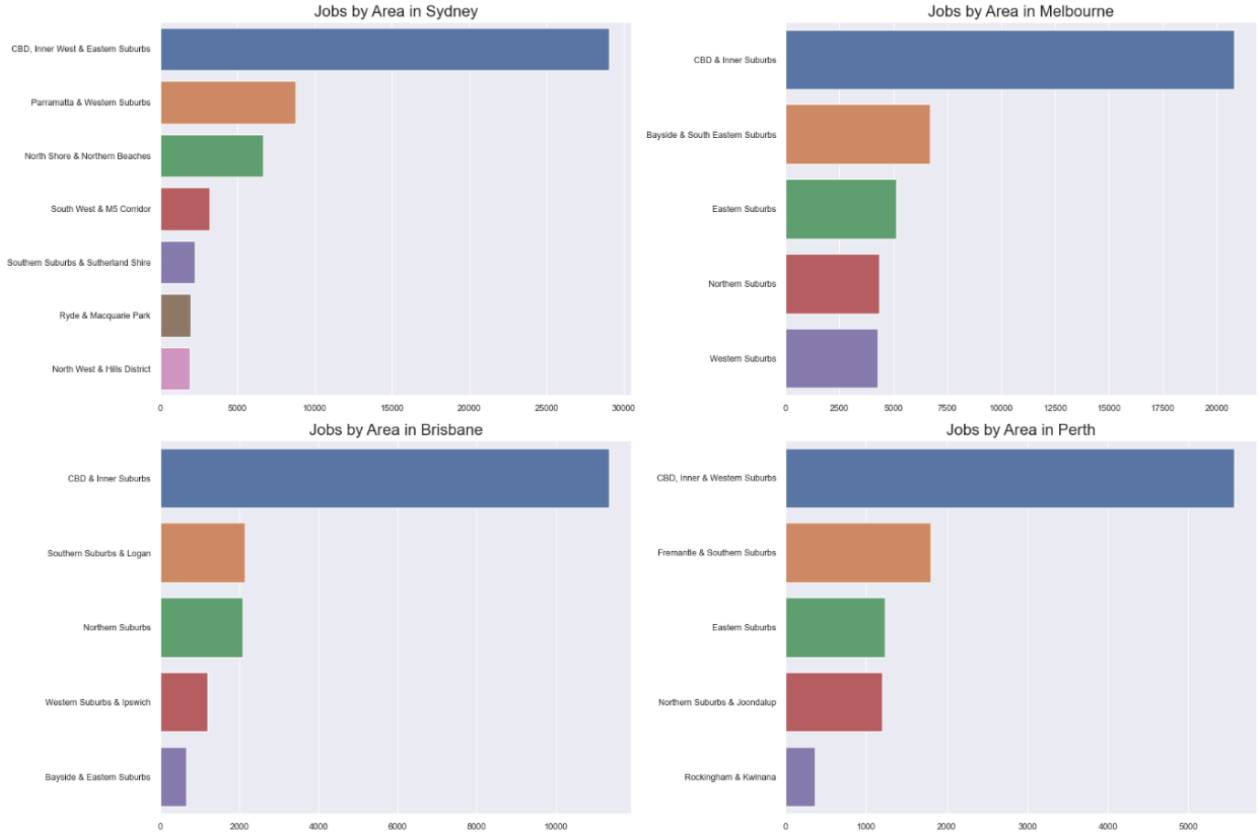
When looking at the top 5 cities in terms of market size, the dataset reveals that Sydney has the highest share of the market with 30.80%. This is followed by Melbourne at 23.67%, Brisbane at 6.74%, Perth at 4.14%, then Adelaide at 2.09%. The entire market by location can be visualised using this graph:

Figure 2. Jobs by Location



The following bar chart shows within the main cities, the majority of jobs are located in the Central Business District (CBD) and Inner Suburbs.

Figure 3. Job Population in Regions of Major Cities

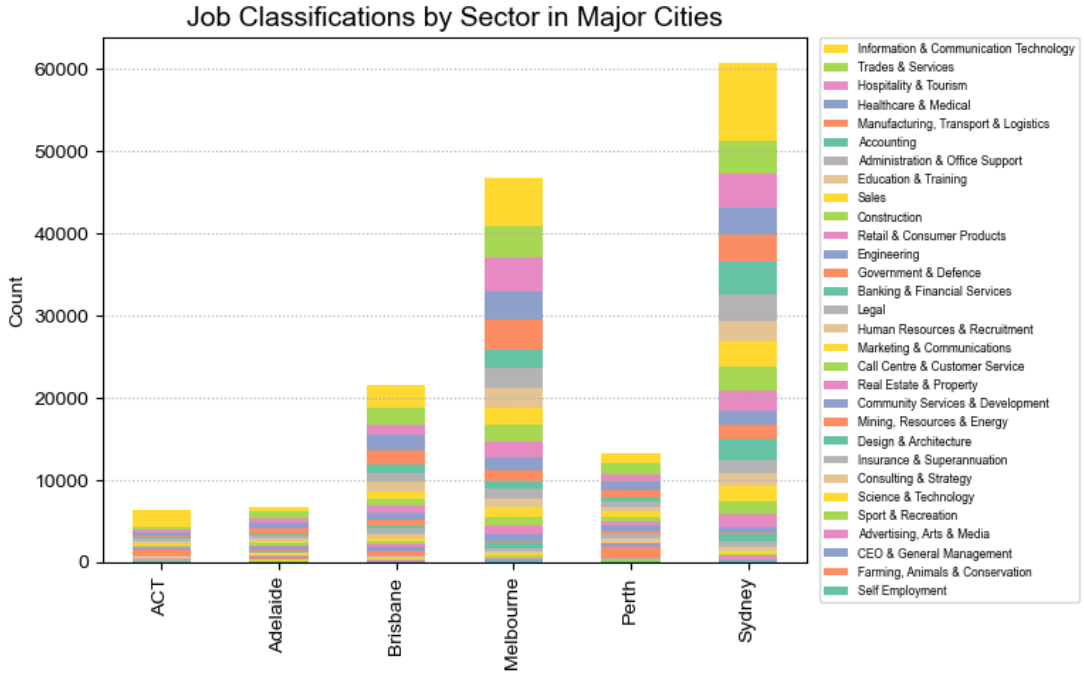


Sectors by Location

In this section we explore the market sectors in the major metropolitan areas. The six largest areas by job Classification are Sydney, Melbourne, Brisbane, Perth, Adelaide, and ACT. Sydney, Melbourne, Brisbane, and ACT all share ICT as their largest job sector, with Trades & Services taking the largest share in Perth and Adelaide.

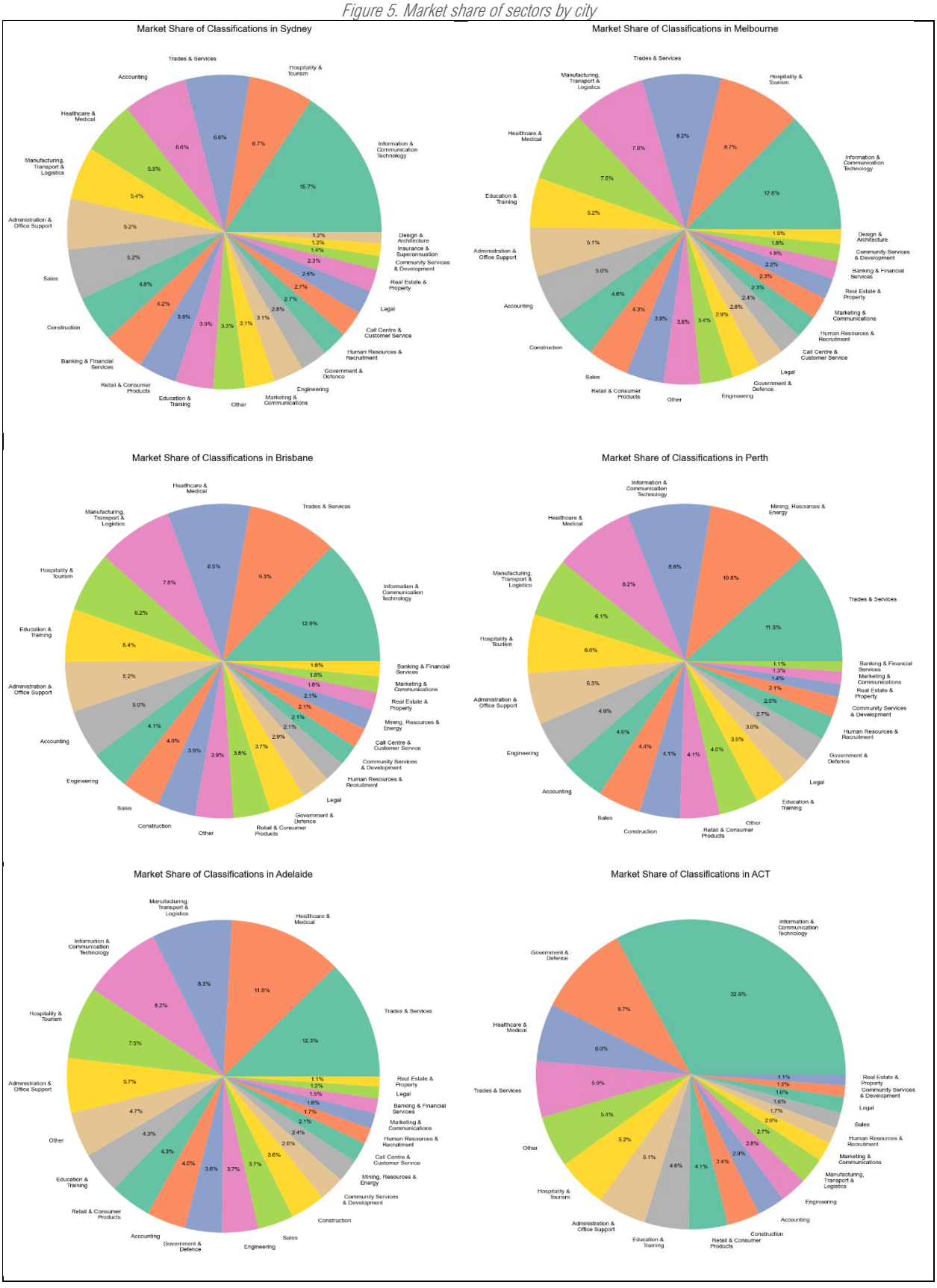
Perth has the most evenly distributed market, with their largest sector capturing only 11.5% of the market, ACT in contrast has almost a third of their market dominated by ICT Jobs at 32.9%. The influence of Perth's prominent mining industry is reflected in the data with 22.3% of their market comprising of Mining, Resources & Energy and Trades & Services. The distribution of market shares in the major cities is visualised below.

Figure 4. Job Classifications by Sector in Major Cities



The above figure gives an overview of market distribution. Information & Communication Technology, Trades & Services, and Hospitality & Tourism are consistently the most dominant or "hottest" job sectors in the top 6 majoring cities. Notably, Perth and Adelaide share Trades & Services as their largest sector. Perth's and ACT's second largest sectors are Mining, Resources & Energy, and Government & Defence respectively, reflecting their major industries. On the following page each city is represented in a figure of its own.

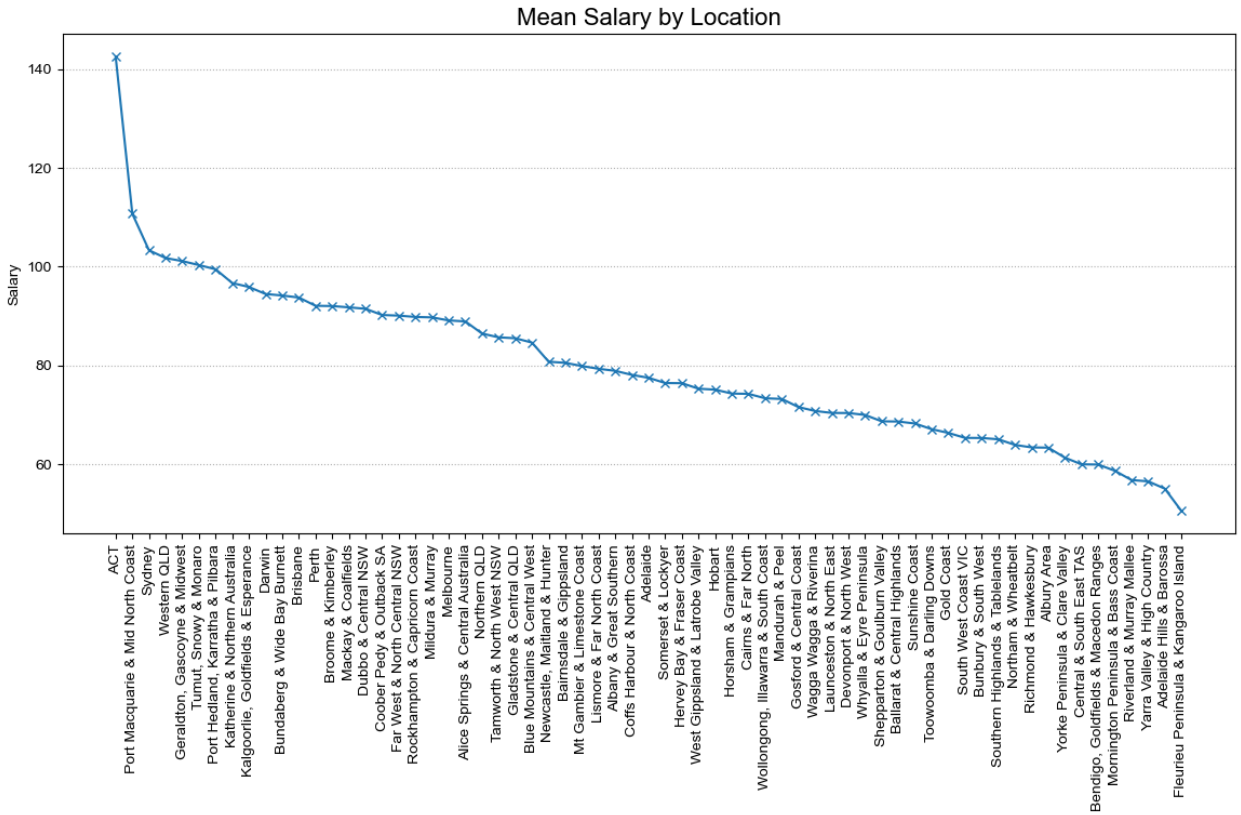
Figure 5. Market share of sectors by city



Salary By Location

For this section, a 'Salary' column is used, which is the mean of the low end and high end of a jobs advertised salary range. The high end and low end of salary ranges for each location was uniformly 999 and 0 respectively except for a few outliers. As seen below, the mean salary mostly ranges between 55k and 100k, with ACT and Port Macquarie and Mid North Coast being outliers. This appears to be due to Port Macquarie having a disproportionate number of high-end medical professionals, with 68.5% of their 200-999k salary range postings being in the Healthcare & Medical field. ACT on the other hand has a huge number of highly paid ICT professional positions with 79.3% of the 200-999k salary range postings in Information & Communication Technology. ACT and Port Macquarie and Mid North Coast have 11.3% and 7.4% of their job markets advertising in the 200-999k salary range with the prevalence of that range being only 3.6% in the overall dataset.

Figure 6. Mean salary by location



The upper end of salary ranges arbitrarily being set to 999k is skewing the mean of the data, these data points are adding salary values of 599.5k to the data set which are extreme outliers. Only 0.001% of Australians earned a taxable income of over 500k in 2021 [3]. For this reason, a second figure is presented which removes these data points, reflecting a more accurate salary market for the average Australian.

Figure 7. Mean salary by location with 200-999k excluded

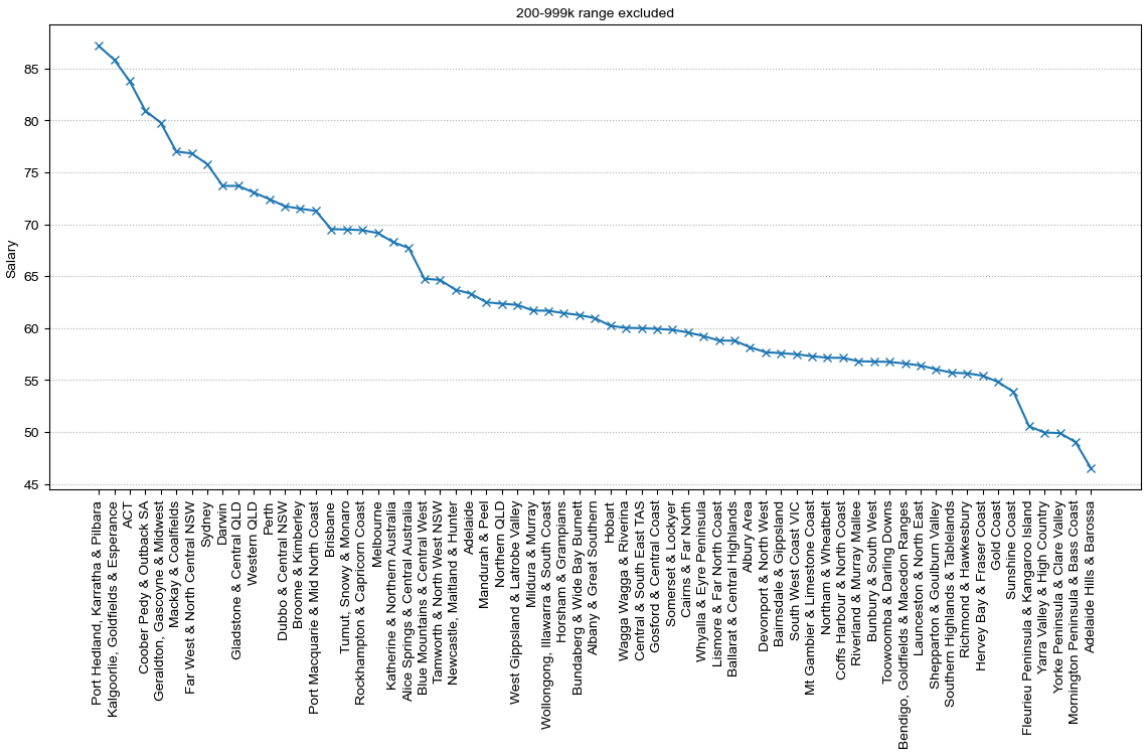
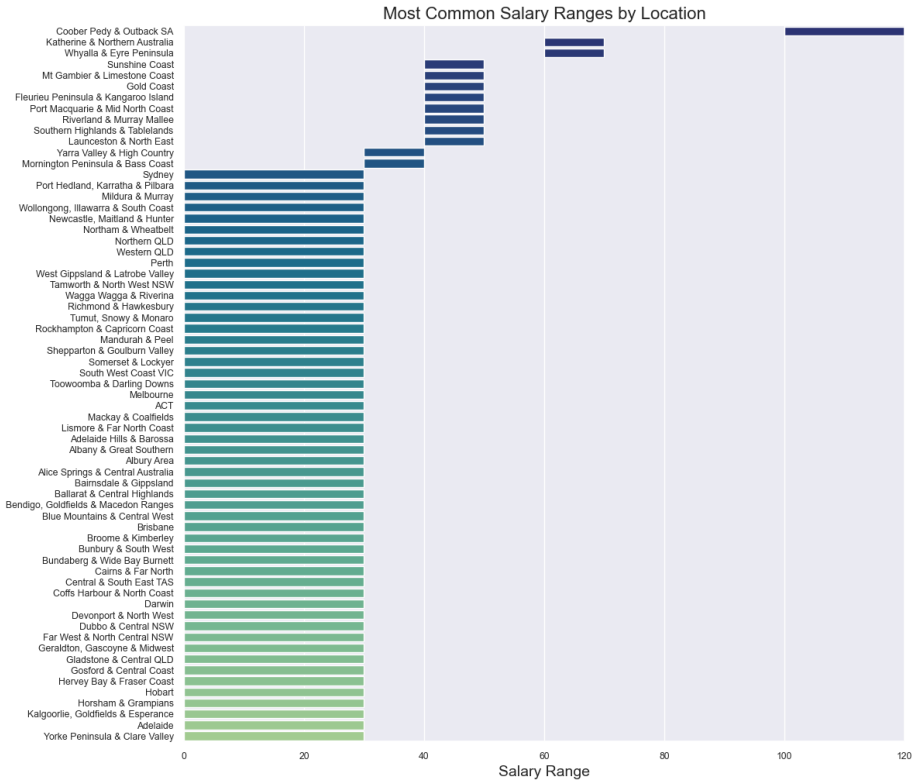


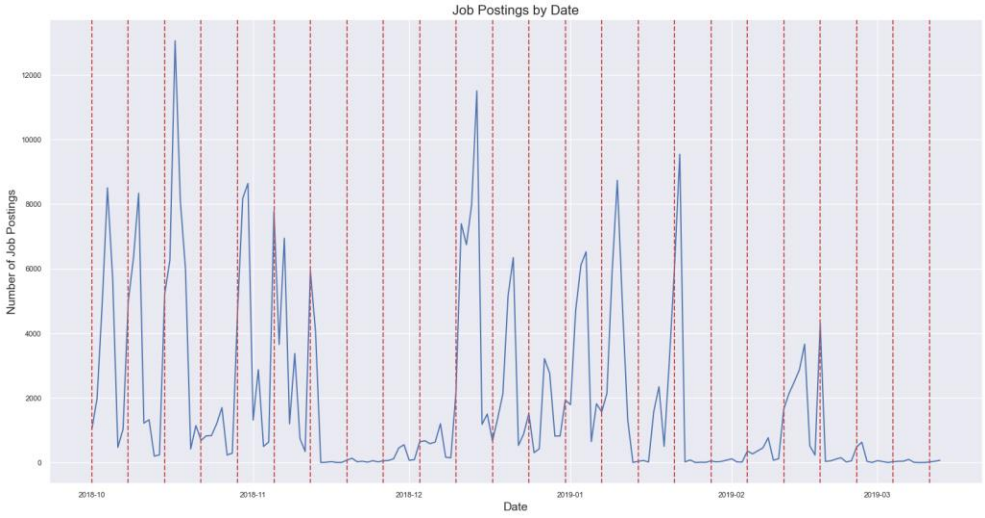
Figure 8. Most common salary ranges by location



Seasonality

Upon the initial visual inspection of the Job postings by date there appeared to be some amount of seasonality, when grid lines were added to show weekly cycles, a pattern emerged, and weekly seasonality was investigated.

Figure 9. Number of job postings by date with Monday grid lines



As seen in the below figures, the observed correlation between job posting frequencies on identical weekdays is minimal, implying a weak linear relationship. With visual inspection of the above figure, it is observed that there are periods where the postings are suppressed, such as week 7 to 11, which may obfuscate statistical seasonality. A higher correlation may be attainable by subdividing the analysis to examine periods of high and low activity separately. The Augmented Dickey-Fuller (ADF) test indicates stationarity within the series. The ADF statistic is negative and exceeds all critical values, indicating a strong rejection of the null hypothesis, thereby suggesting that the data is stationarity. The corresponding p-value of 0.007 substantiates this conclusion, falling well below the 0.05 threshold for statistical significance.

Figure 10. Autocorrelation and Partial Autocorrelation of no. of postings on weekdays

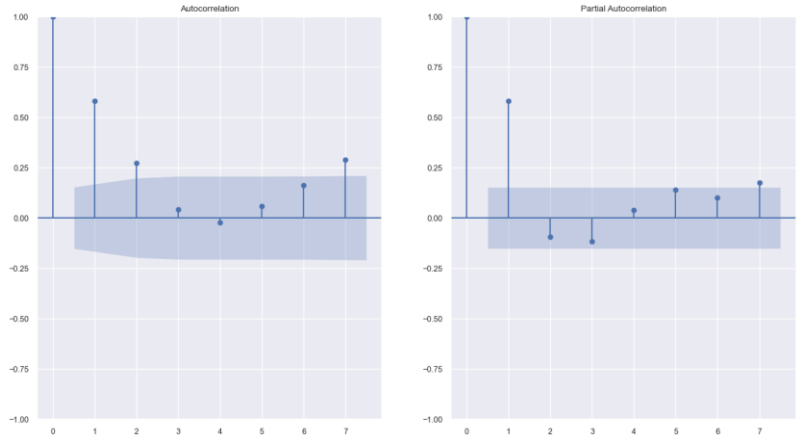
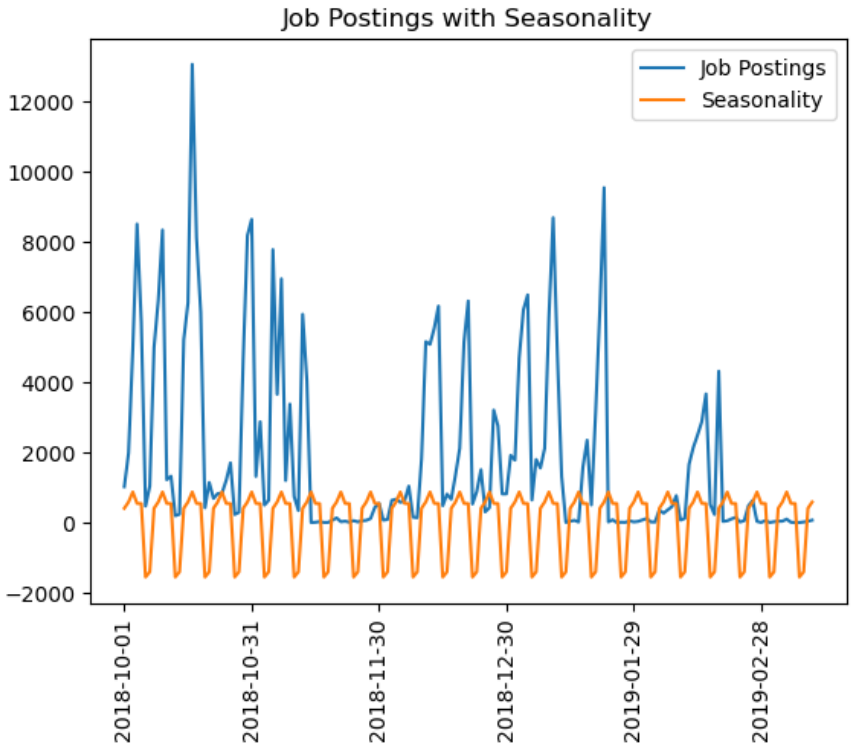


Table 3. ADF test statistics over 7-day period

ADF Statistic	-3.5
p value	0.008
Critical Values	
1%	-3.47
5%	-2.88
10%	-2.57

The *seasonal_decompose* function from the *sm.tsa* module in Python's StatsModels library is used for time series decomposition. Time series decomposition involves breaking down a time series into its individual components: trend, seasonality, and residuals (or noise). This decomposition helps in understanding and analysing the underlying patterns and characteristics of the time series data.

Figure 11. TSA Seasonality

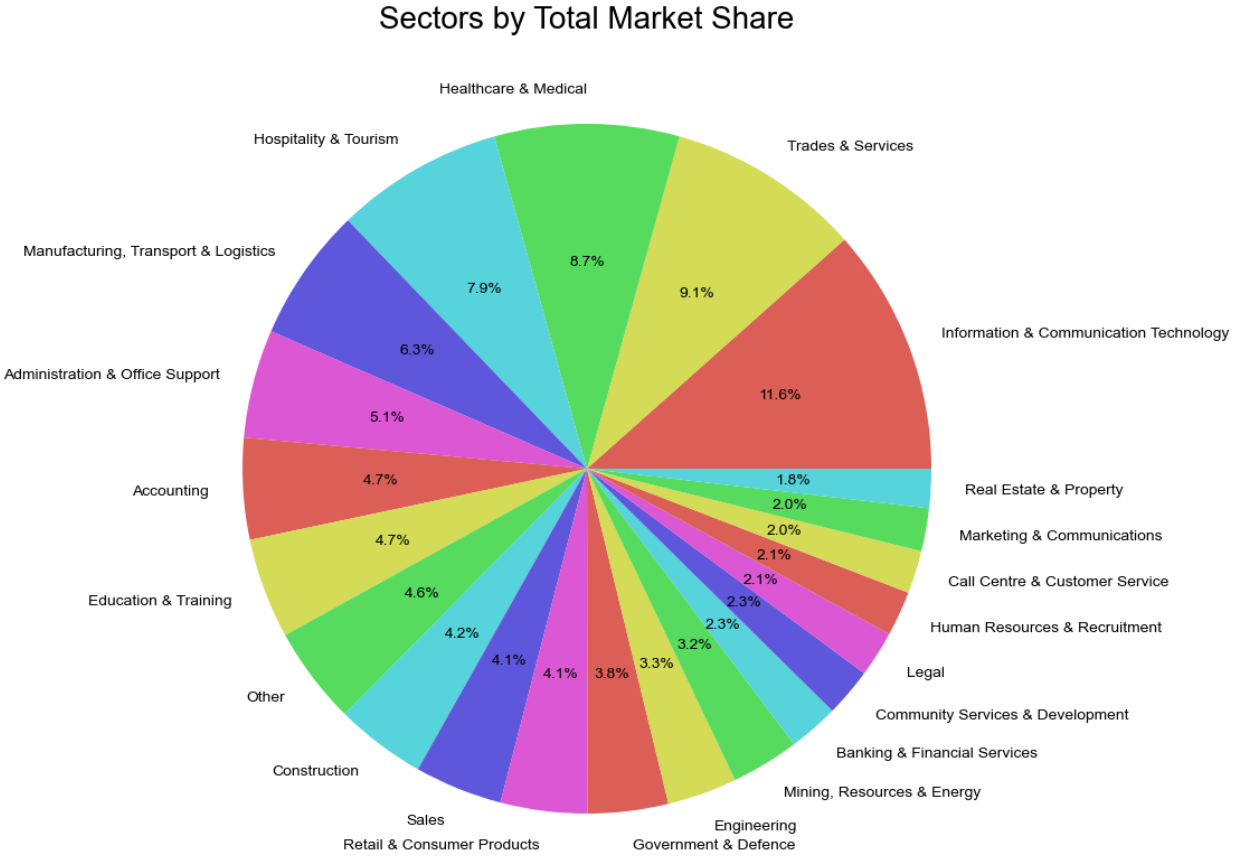


The use of the *seasonal_decompose* seasonal component provides additional context on whether job postings are seasonal or not. On a period of 7-days, the results visualised in Figure 11 show a weekly seasonality which peaks towards the middle of the cycle and drops nearing the end. Indicating that the most job postings occur within the middle of the week. With highest occurrences between Tuesday-Friday.

Market by Sectors

This section explores market data by Job Sectors, extracting relevant salary information, and the most desirable skills in specific Job subclassifications. A new column, 'ClassificationLumped', was created for this section. It groups classifications responsible for less than 1% of the total market share into an 'Other' category. Please note that this category is used only for visualisation purposes and not for statistical analysis. As seen below, six sectors are responsible for almost half of the entire job market, these sectors are Information & Communication Technology with 11.6%, Trades & Services with 9.1%, Healthcare & Medical with 8.7%, Hospitality & Tourism with 7.9%, Manufacturing, Transport & Logistics with 6.3% and Administration & Office Support with 5.1%.

Figure 12. Sectors by market share



Below, Breakdowns of each of these 6 sectors into subclassifications can be seen, exploration of all other categories can be seen in the interactive visualisation presented later in the report. The largest Subclassifications in order are Other, Management, Chefs/Cook, Administrative Assistants, and Developers/Programmers. Other is disproportionately large due to being a catch all, and Management is a general term which appears in several categories; therefore Chefs/Cooks, Administrative Assistants and Developers/Programmers are the most in demand subclassifications.

Figure 13. SubClassifications within Information & Communication Technology

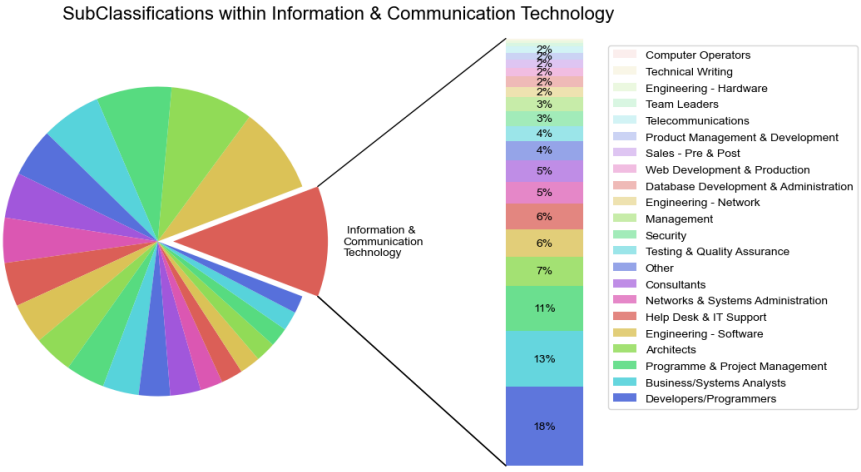


Figure 14. SubClassifications within Trades & Services

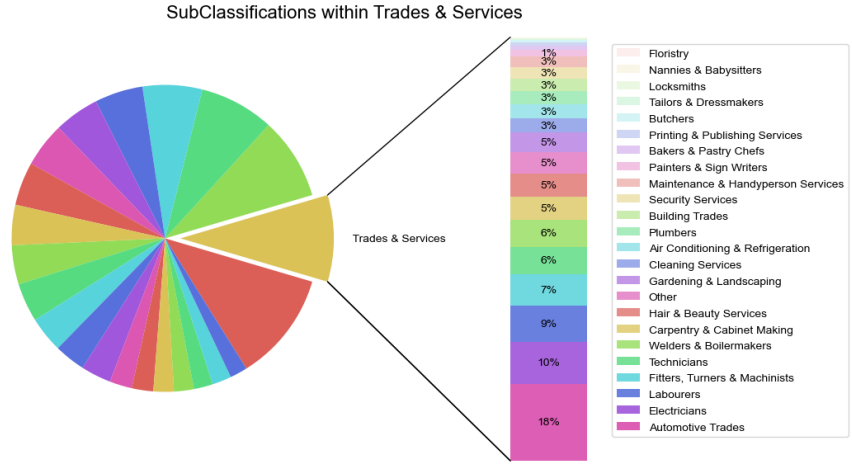


Figure 15. SubClassifications within Healthcare & Medical

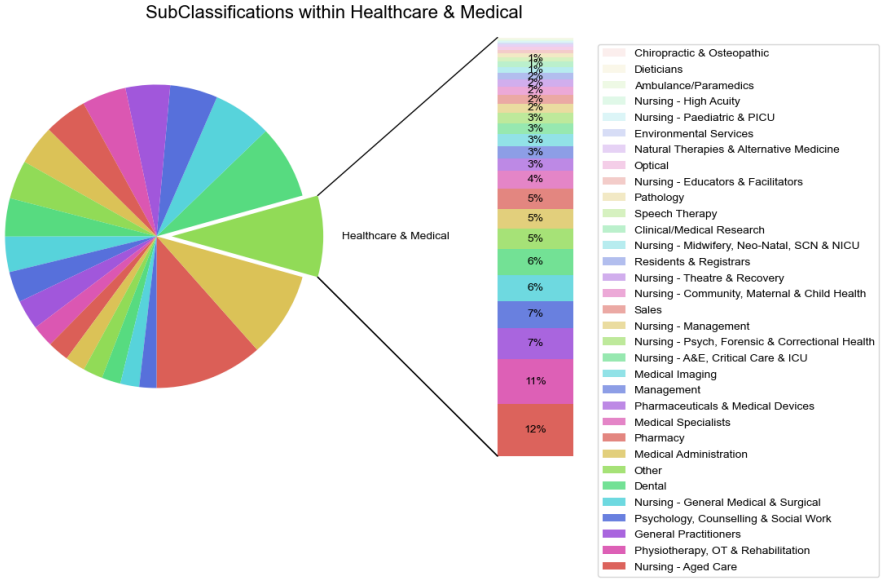


Figure 16. SubClassifications within Hospitality & Tourism

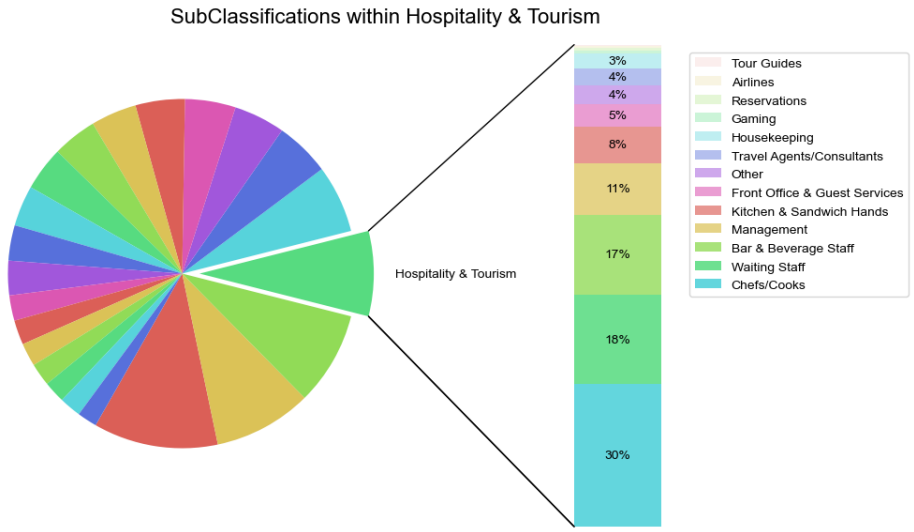


Figure 17. SubClassifications within Manufacturing, Transport & Logistics

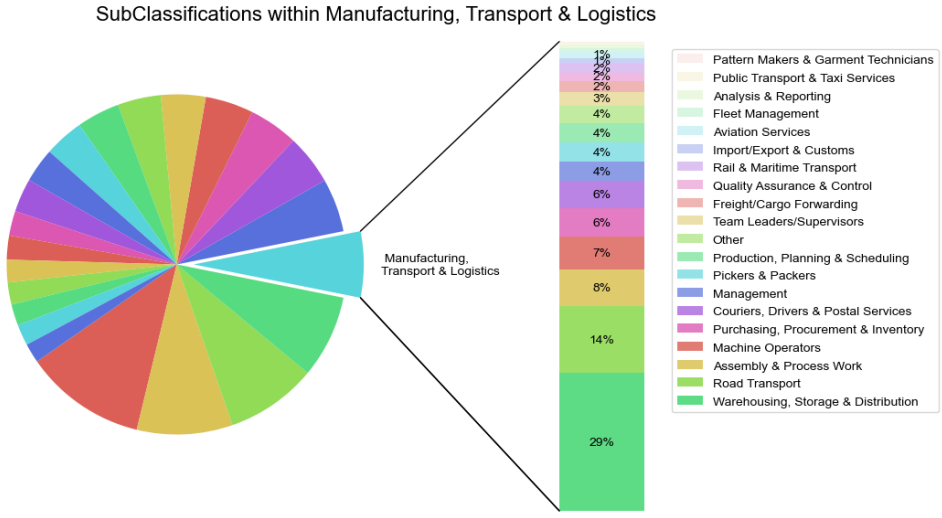
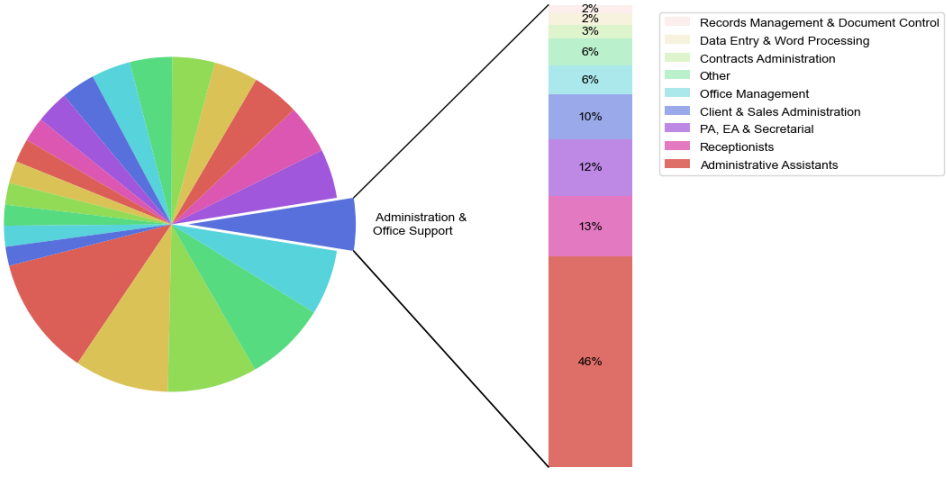


Figure 18. SubClassifications within Administration & Office Support



Salary by Sector

As seen below the mean salary by sector ranges widely, with Hospitality & Tourism having the lowest value at 40.2k and CEO & General Management the highest with 255.4k. Similarly to mean salary by Location, the sectors which have larger proportions of posting with salaries of 200-999 will heavily weigh the means upward due to the arbitrary upper limit, however in this case, this data is pertinent, if overweighted, so will not be removed. Over 29% of postings for CEO & General Management have the maximum salary range, and it is the most common range for that sector.

Figure 19. Mean salary by sector

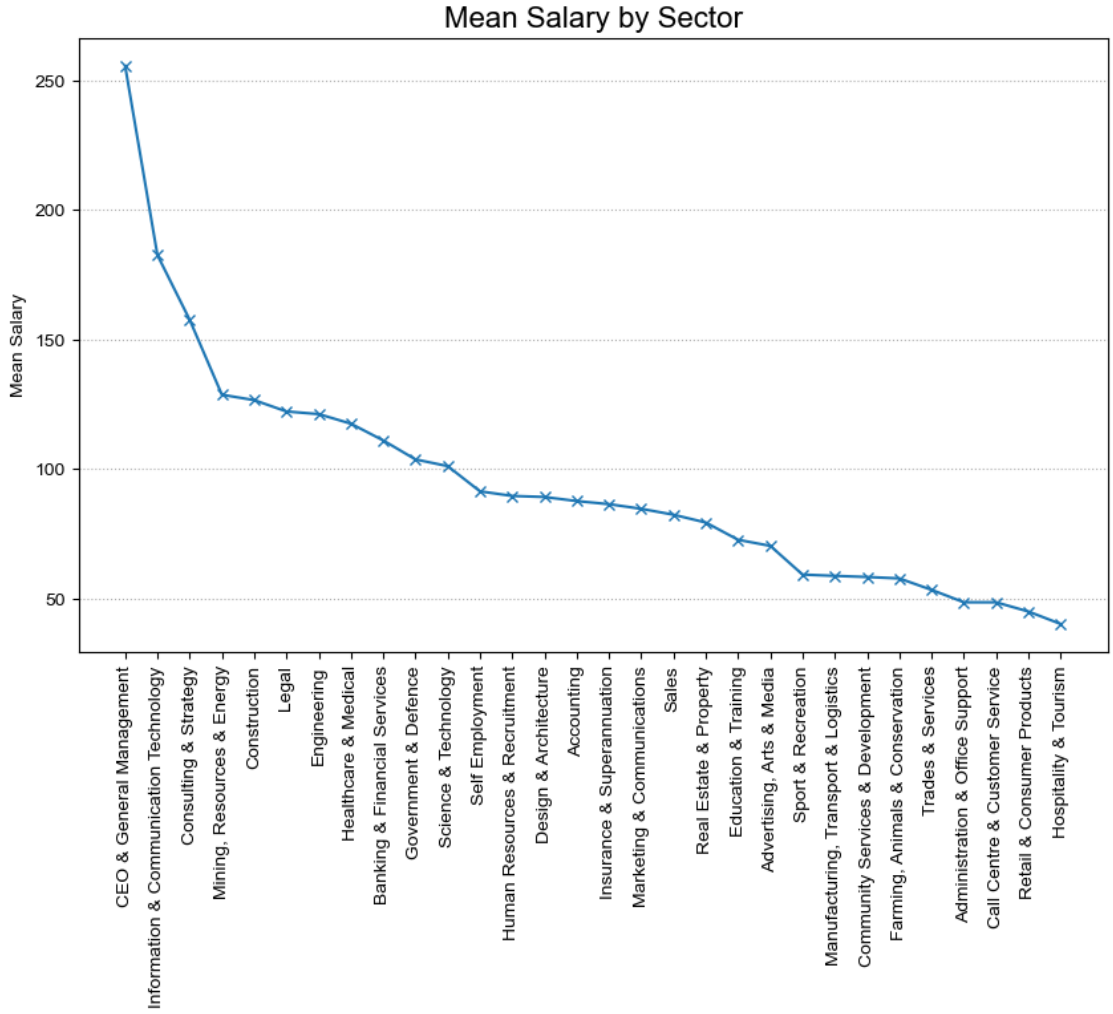


Figure 20. Most common salary range by sector

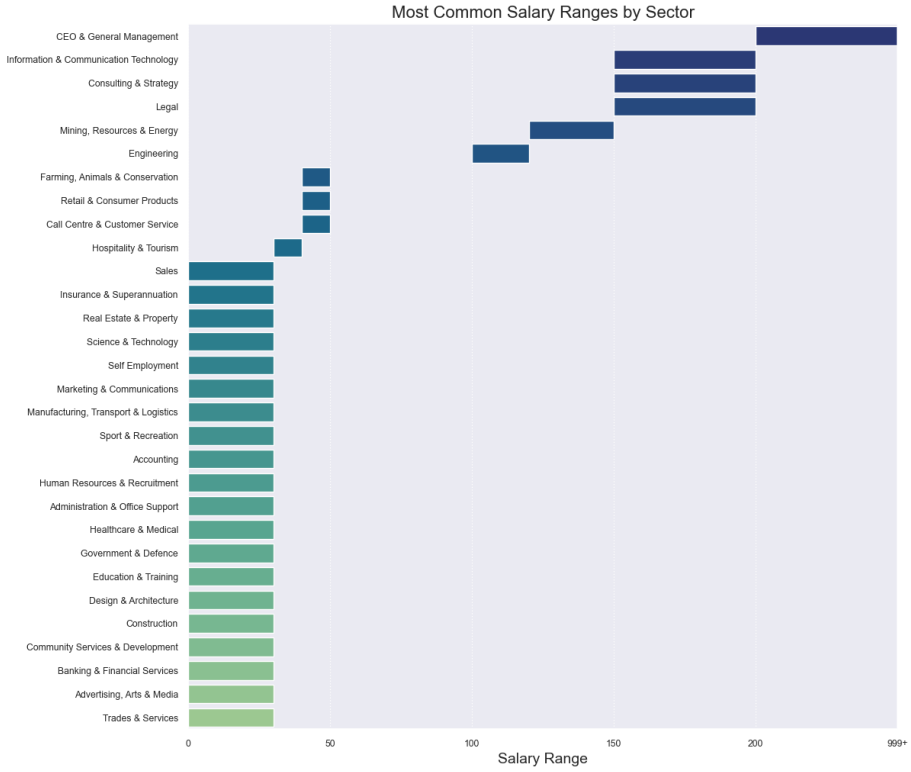
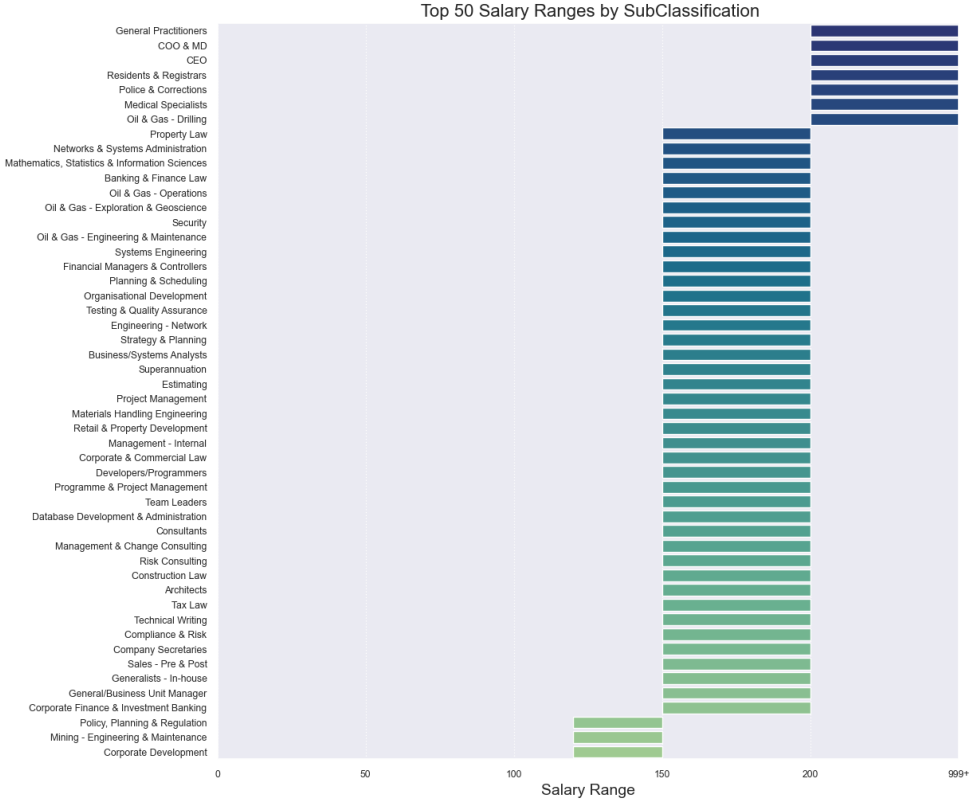


Figure 21. Top 50 highest common salary ranges by Subclassification



Among the 50 top salary ranges in subclassifications, ICT, Legal, Mining, Resources & Energy, and Consulting & Strategy are the most heavily represented sectors. Notably, ICT accounts for 13 of the 50.

Table 4. Number of Subclassifications from Sectors in top 50 highest average salary ranges

Sector	Count
Information & Communication Technology	13
Legal	6
Mining, Resources & Energy	5
Consulting & Strategy	3
CEO & General Management	3
Engineering	3
Healthcare & Medical	3
Government & Defence	2
Banking & Financial Services	2
Accounting	2
Insurance & Superannuation	2
Human Resources & Recruitment	2
Construction	2
Real Estate & Property	1
Science & Technology	1

Trend of Market

Given the fact that the job postings in the dataset have a 6-month range, determining a trend is difficult. To generally trend detection requires the use of a long-term pattern in data over time. Given the data.csv file, each individual job classification and the overall job market is analysed.

Figure 23. Market Trend

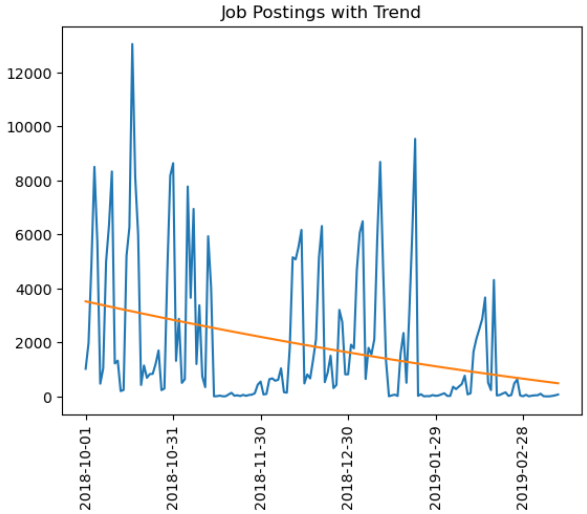
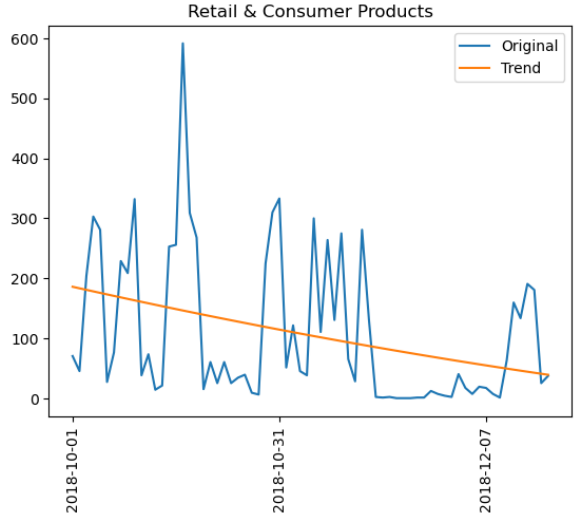


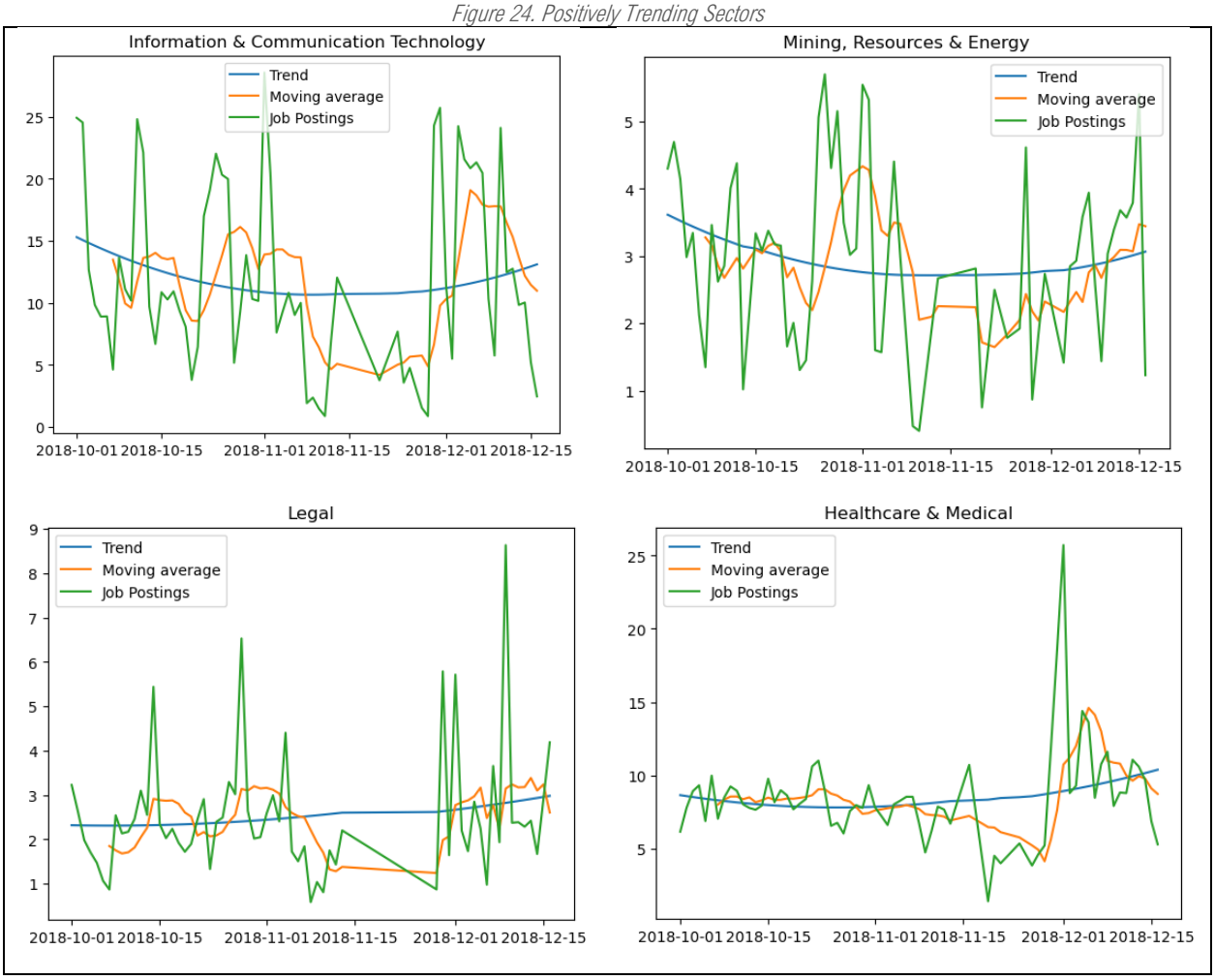
Figure 22. Retail & Consumer Products Trend - Market Influence

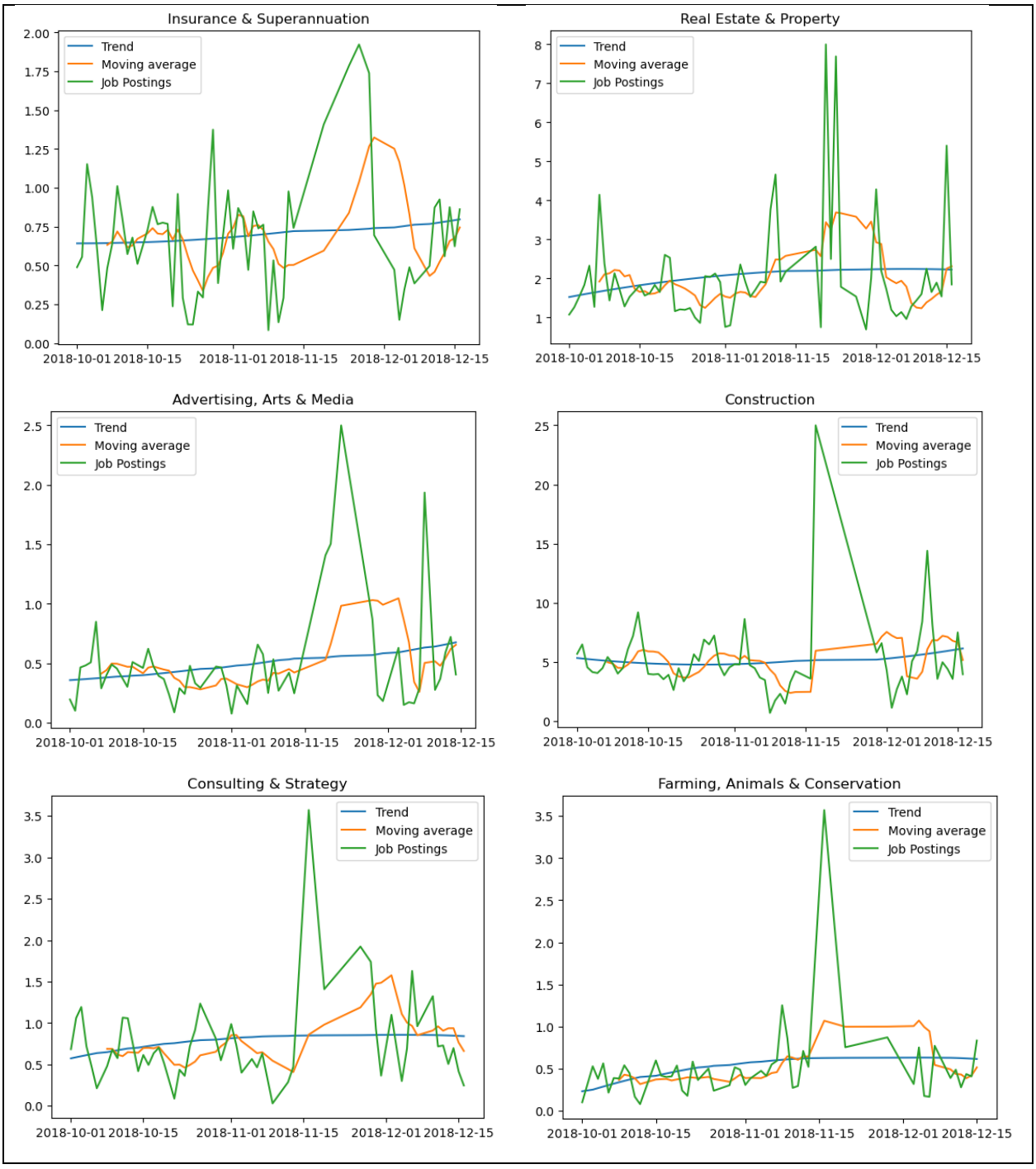


First, the number of job postings was looked at. The overall market showed an overall downward trend. Keep in mind that this may be due to the short period in which the job postings were recorded. The beginning of the year may generally have less job postings, which may be the cause this downward trend as the job posting range is from 10/2018 – 02/2019. Individual sectors reflected this downward trend as each job classification was influenced by the overall trend of the market.

Secondly, Job classifications are plotted in relation to the current market. Meaning what percentage of the current market do they share. This method will give a fair and reliable result when comparing the growth of a job sector within the overall market. Two formulations were made, one tested daily job market data and the other weekly. The sectors which showed upwards trends are displayed below:

Figure 24. Positively Trending Sectors





Each graph uses a polynomial trendline due to their improved fit to data and visual representation. It is observed that ICT and Mining were affected by the suppression of the macro market when job postings dropped, from which they recovered faster than other markets. The remaining examples show either a lessened effect of the market suppression due to their significantly smaller market share or show signs of steady growth. Based off the trend seen with the 6-month period, these jobs are predicted to show growth in the near future.

Career Recommendation

The question asks to generate a recommendation on which market a student should base their university degree choice off to guarantee a job in the future. The 10 industries which show an upwards trend for growth are listed above. To make a recommendation, the share in the market should additionally be considered, along with following the trend.

Table 5. Market share of upward trending sectors

Job Classification	Average Market Share (%)
Information & Communication Technology	11.936
Mining, Resources & Energy	2.952
Legal	2.513
Healthcare & Medical	8.474
Insurance & Superannuation	0.695
Real Estate & Property	2.034
Advertising, Arts & Media	0.490
Construction	5.117

The Information & Communication Technology sector emerges as a dominant force in the market, representing an average share of 11.94%. Considering its expansive presence and the significant size of the job market it encompasses, it is justifiable to recommend pursuing broad ICT degrees such as Information Technology or Computer Science. These fields align with the dynamic and expanding nature of the sector, ensuring a promising range of opportunities.

Skills by Sector

To make good use of the information extracted from the dataset, actionable knowledge must be found. In service of this cause, the Requirement section of the dataset was analysed with the purpose of finding the most sought-after skills and attributes by employers in each sector. To retrieve this information, Term Frequency (TF) matrices were generated for the Requirements stated by employers in each sector. The TF matrices with the incorporation of Inverse Document Frequency weights provided the most important words in the Requirement corpus for each Sector. After the first iteration of TF-IDF's were found, they were visually inspected, and an additional list of custom stop words was added to the nltk standard list which was originally used. This increased the likelihood of the resultant words reflecting desirable skills or attributes. This method, while not perfect, did illuminate some overall employer dispositions and market demands. The table below displays the 20 most important words in each sector. TF-IDF matrices were computed for all Sectors and Subclassifications. Further Automation of this process and displaying the information becomes difficult from this point as there are over 30 Sectors, and hundreds of Subclassifications. However, some relevant information discerned from the largest sectors can be found with visual and critical analysis of the data and will be displayed. The large body of text information can potentially be further investigated in the future.

Figure 25. TF-IDF most common words by Sector

Retail & Consumer Products	Call Centre & Customer Service	Hospitality & Tourism	Banking & Financial Services	Manufacturing, Transport & Logistics	Sales	Administration & Office Support	Trades & Services	Accounting	Real Estate & Property	Healthcare & Medical	Marketing & Communications	Government & Defence	Information & Communication Technology	Education & Training
store	customer	cafe	financial	driver	sale	administration	required	accountant	property	care	marketing	service	contract	teacher
retail	service	casual	service	forklift	business	assistant	service	business	manager	nurse	communication	officer	project	school
customer	sale	restaurant	business	operator	manager	office	qualified	finance	management	health	manager	project	business	early
manager	centre	cook	manager	warehouse	account	support	maintenance	account	real	registered	digital	support	developer	educator
sale	call	barista	risk	required	customer	service	fitter	financial	estate	practice	brand	government	analyst	centre
assistant	business	bar	client	client	development	administrator	technician	officer	sale	service	business	management	client	education
service	skill	cafÃ©	bank	ongoing	service	receptionist	electrician	firm	agency	aged	medium	health	service	childhood
brand	client	chef	customer	shift	industry	customer	mechanic	senior	portfolio	permanent	strategy	permanent	senior	learning
casual	officer	waiter	leading	day	leading	officer	trade	service	development	medical	coordinator	community	manager	passionate
fashion	contact	waitress	analyst	production	consultant	business	apprentice	manager	office	hospital	leading	manager	cbd	student
passionate	professional	hand	banking	truck	global	administrative	labourer	payroll	support	clinical	campaign	council	engineer	teaching
lead	support	kitchen	management	casual	growing	admin	project	organisation	leading	required	event	contract	organisation	service
retailer	inbound	manager	finance	leading	client	client	workshop	leading	commercial	dental	contract	employment	government	child
people	people	service	senior	manufacturing	leader	contract	client	accounting	fantastic	assistant	global	provide	support	qualified
want	face	food	sydney	service	market	professional	year	client	project	manager	product	type	data	support
fun	environment	melbourne	compliance	business	growth	executive	industry	contract	assistant	casual	growing	lead	large	care
christmas	consultant	sydney	global	worker	product	project	leading	cbd	business	community	social	classification	leading	program
business	growing	hotel	support	operation	professional	cbd	commercial	management	senior	leading	organisation	senior	solution	research
environment	cbd	bartender	professional	transport	management	organisation	vehicle	growing	agent	professional	australia	development	sydney	casual
management	fantastic	hospitality	cbd	suburb	solution	sale	electrical	payable	facility	australia	project	department	system	year
Insurance & Superannuation	**Engineering**	**CEO & General Management**	**Design & Architecture**	**Legal**	**Construction**	**Sport & Recreation**	**Science & Technology**	**Advertising, Arts & Media**	**Farming, Animals & Conservation**	**Self Employment**	**Consulting & Strategy**	**Mining, Resources & Energy**	**Human Resources & Recruitment**	**Community Services & Development**
claim	project	manager	design	firm	project	fitness	data	medium	farm	earn	business	mining	hr	support
insurance	engineer	service	project	law	manager	personal	laboratory	digital	animal	calendar	project	roster	recruitment	care
manager	engineering	leadership	designer	legal	construction	trainer	environmental	agency	hand	christmas	change	operator	business	community
management	design	business	architect	lawyer	commercial	club	research	manager	manager	extra	manager	mine	consultant	worker
service	civil	lead	senior	leading	builder	training	project	account	required	income	senior	fifo	manager	service
cbd	consultancy	operation	interior	senior	site	manager	scientist	content	nursery	period	strategy	required	service	people
leading	senior	executive	practice	commercial	required	motivated	support	client	operation	pop	management	project	organisation	disability
portfolio	manager	general	graphic	practice	civil	people	technical	creative	dog	provides	analyst	fitter	professional	life
customer	service	organisation	working	client	contract	coach	science	leading	industry	retailer	consultant	perth	contract	home
life	sydney	management	sydney	secretary	sydney	sport	product	brand	working	largest	client	underground	support	provide
worker	leading	ceo	revit	national	tier	passionate	global	service	fantastic	australia	service	operation	people	client
client	electrical	growth	leading	associate	leading	health	leading	business	casual	club	contract	engineer	leading	family
sydney	mechanical	leader	studio	partner	contractor	highly	manager	production	groomer	business	leading	site	development	make
professional	infrastructure	community	contract	year	senior	sale	contract	advertising	management	broker	lead	maintenance	sydney	employment
industry	system	strategic	creative	sydney	estimator	group	client	australia	day	mortgage	strategic	contract	client	program
insurer	lead	leading	firm	cbd	operator	instructor	service	event	salon	next	organisation	permanent	culture	child
manage	melbourne	responsible	digital	corporate	supervisor	want	senior	talented	worker	lending	program	plant	growing	difference
compensation	development	support	residential	litigation	residential	centre	industry	marketing	production	level	policy	service	partner	organisation
business	contract	project	melbourne	global	administrator	enthusiastic	quality	journalist	service	supercharge	across	diesel	cbd	passionate
working	global	director	ux	property	building	member	business	growing	environment	take	transformation	leading	resource	young

Table 6. Words of Interest for Sectors extracted from TF-IDF

Information & Communication Technology	Trades & Services	Healthcare & Medical	Hospitality & Tourism	Manufacturing, Transport & Logistics
contract	required	care	cafe	driver
project	service	nurse	casual	forklift
business	qualified	health	restaurant	operator
developer	maintenance	registered	cook	warehouse
analyst	fitter	practice	barista	required
client	technician	service	bar	client
service	electrician	aged	cafe	ongoing
senior	mechanic	permanent	chef	shift
manager	trade	medical	waiter	day
cbd	apprentice	hospital	waitress	production
engineer	labourer	clinical	hand	truck
organisation	project	required	kitchen	casual
government	workshop	dental	manager	leading
support	client	assistant	service	manufacturing
data	year	manager	food	service
large	industry	casual	melbourne	business
leading	leading	community	sydney	worker
solution	commercial	leading	hotel	operation
sydney	vehicle	professional	bartender	transport
system	electrical	australia	hospitality	suburb

From the above TF-IDF figure some sought-after skills in the five sectors can be deduced.

Information & Communication Technology: Desirable skills in this sector include project management, data analytics, management, tech support, and system design.

Trades & Services: Skilled tradesmen who are fully qualified and proficient in fitting and maintenance are highly sought after. Electricians and mechanics are particularly in demand. Additionally, commercial experience, client-facing skills, and access to transportation may be necessary for many positions

Healthcare & Medical: In the medical field, registered nurses and individuals with experience in aged care and hospital environments are in high demand. Dental assistants and clinic managers are also sought after.

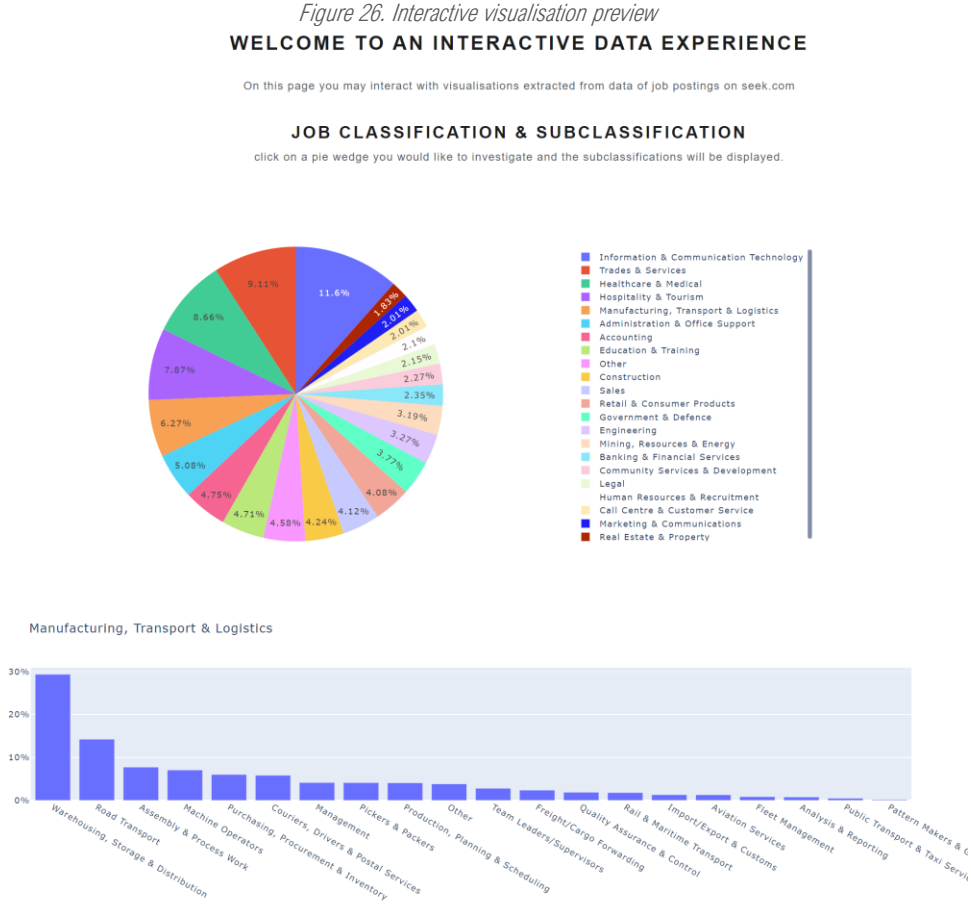
Hospitality & Tourism: The sector requires experienced kitchen staff, wait staff, and bartenders for cafes, restaurants, and hotels. Skills in food service and management are frequently mentioned as well.

Manufacturing, Transport & Logistics: The most sought-after roles in this sector are truck drivers, forklift operators, and individuals with experience in warehouse and production roles. Having personal transportation is often a requirement.

Interactive Visualisation

An visualisation which allows the user to explore the data set interactively has been created, this visualisation has two sections, the first displays a pie chart with the sectors and their respective market shares, when a wedge is clicked that sectors subclassifications are displayed by their market share of the sector below in a bar chart. The second section the same format, with Locations displayed by the bar chart, and with subclassification details given when locations selects. This data was chosen to be displayed interactively as it could not practically be displayed in its entirety in image form within this document.

Figure 26. Interactive visualisation preview



To view the interactive visualisation simply double click the 'Run Visualisation.bat' file in the local directory, or run the 'interactive.py' file directly in the terminal. It may be necessary to install the dash libraries before viewing if they are not installed, this can be achieved using the 'pip install dash' and 'pip install dash-bootstrap-components' commands in the terminal. Once you have run the script either in terminal or using the .bat file wait until a window pops up and you may click the hyperlink to view the visualisation, alternatively it is hosted at http://127.0.0.1:8050/. The script will take 10-30 seconds as it needs to load the data set. Closing the terminal will terminate the visualisation.

Evaluation

Findings

This report has found that the according to job posting data from seek.com in the period of 1/10/2018 13/03/2019 the overall Job market in Australia is dominated by 6 Sectors which make up almost 50% of the market, ICT, Trades & Services, Healthcare & Medical, Hospitality & Tourism, Manufacturing, Transport & Logistics, and Administration & Office Support. Similarly, Australia's two largest metropolitan areas, Sydney and Melbourne accounted for 54% of postings with available location data. ACT was found to have a market dominated by demand for ICT professionals with almost a third, 32.9%, of the postings advertised in the ICT sector. Perth, alternatively, had a relatively balanced market with the largest sector capturing only 11.5% of the market. Analysis found some evidence for the weekly seasonality of job postings.

The hypothesis that major cities will have higher salaries was partially substantiated. Of the major cities only Adelaide is in the lower half of mean salaries, with ACT having the highest and Sydney, Brisbane, and Perth all within the top 13. The hypothesis that Mining, Resources and Energy, would have high salaries was also confirmed, with the sector having the 4th highest mean salary and accounting for 5 of the top 50 subclassifications by average salary. Overall, the ICT industry is the largest job market, has the second largest mean salary and accounts for 13 of the 50 subclassifications with the highest most common salary range. The lowest salaries were in unskilled sectors such as Retail, Hospitality and Call centers.

Market Balancing

Market Balancing can be achieved through either influencing market demands or supply. In the case of the job market, the demand is beholden to larger economic and societal movements which are not controllable. However, with market knowledge discerned from the supplied data some action can be taken to address the supply. It has been found that 6 sectors account for almost 50% of demand, therefore focusing resources on the education and training of professionals in these areas would have a balancing effect on the market as well as positive results for employees and employers. Some of the markets which are most in demand also have the highest salaries available, this is information which is of interest to those entering the job market but isn't highly visible. Making current market data highly visible and available to those to whom it is of most interest, e.g, schools, and higher education centers, will naturally incentivise individuals to specialise in sectors with demand, as the promise of a welcoming job market is a natural draw. Furthermore, marketing campaigns for the most in need markets can help to create knowledge of sectors which employees can gravitate towards or upskill to join. Finally, government incentives to increase the attractiveness of unappealing and in need markets could help to fill roles in struggling locations/sectors.

Refining Analytics

The data source that was used in this analysis off a sufficient size, with over 300,000 entries, however there were several ways the data set could be improved to provide a more robust overview of the Australian Job Market.

The data set was limited to job postings on seek.com. While seek is a large job platform, it is only one of many similar sites that are popular in Australia. Using a single source introduces uncertainty about how accurately the data reflects the true markets as certain industries may favor different web-based job markets, for example, the mining industry may favor a different service where ICT favors seek.com. This creates an inaccurate representation of the overall market. To solve this issue, several different job market datasets could be collected, and merged into a unified dataset to provide a more accurate representation of the market. Furthermore, the temporal range of the dataset was not large enough to extrapolate any meaningful trend or seasonality information. Greater insights would be achievable with a several year time span of data.

Advertised salary ranges may not be an ideal parameter to judge financial outcomes for people in the job market as they do not necessarily reflect what successful candidates are paid and are often used loosely or excluded from job postings all together. A preferable statistic to forecast financial outcomes for employment in sectors and subsectors would be to use current salaries in those areas from reputable sources such as the Australian Bureau of Statistics.

The method employed to determine the most in demand skills in market sectors proved to be ineffective. A more in-depth filtering system of the words is required to gather a higher quality of data. Despite efforts to manually remove noise (non-skill related words) from the data set the results were less than ideal. Options to refine the process might include some POS tagging to remove unwanted words, however this represents its own challenge; the classifications available for POS tagging aren't specifically helpful for targeting 'skill' words. A skill word list could be generated and only these words investigated, however the resources to create such a list are prohibitive for the scope of this report.

Potential Implications for Employers and Employee's

For Employees the most pertinent information found through the exploration of this data is that the sectors with the highest salaries are those in which the employees are highly skilled. Highly skilled workers are those who have obtained technical skills in a specific field, ranging from ICT to technical trades or business management. Being highly trained was shown to be a requirement in the fields with the best incomes. The largest job markets are in the major cities, with Perth in great need of ICT professionals. For those seeking jobs within trade professions, if salary is a determining factor in job selection, the mining industry appears to be the most suitable choice.

For Employers it may be beneficial to focus recruitment efforts in the sectors with high demand by offering more competitive salaries and accounting for larger employee acquisition times. Additionally, partnering with education or training centers by offering opportunities/internships for students may be worthwhile. This provides access to early talent and the opportunity to develop employees from within. To address skill shortages, Employers may also look to offer current employees paid upskilling opportunities.

Case Studies

Case 1 – Mathew

The three most in demand subclassifications within the ICT industry are Developers/Programmers, Business/Systems Analysts, and Programme & Project Management. These account for 41% of ICT job postings, so focusing on skills critical to one or all of these jobs will maximise his chances of finding employment after graduation. These jobs are also in the top 50 highest common salary ranges, making them attractive financially.

Figure 27. Top 20 TF-IDF words for most common ICT subclassifications

Developers/ Programmers	Programme & Project Management	Business/Systems Analysts
organisation	organisation	organisation
contract	contract	contract
cbd	cbd	cbd
project	project	project
government	government	government
large	large	large
senior	senior	senior
client	client	client
leading	leading	leading
sydney	management	data
engineer	service	business
software	business	technical
technology	change	analyst
application	required	service
developer	program	system
stack	manager	program
service	transformation	ba
development	delivery	required
working	implementation	financial
java	support	process

Referring to the TF-IDF's 20 most common words for each subclassification, it is observed that 9 of the words are common to all three subclassifications. Of these, only some give insight into useful skills. Organisation and strong client relationship skills are highly desirable. Cbd may refer to Component-based development but is more likely in the list due to mentions of a central business district in many of the postings. Programming is shown to be a necessity, with mentions of software engineering/development, the Java coding language, applications, and a 'stack'. Knowledge of data analysis and delivering work on time also seem to be of interest.

Based on the data I would advise Mathew to ensure he takes subjects which teach the fundamentals of software programming with a focus on gaining knowledge of a particular software stack. Any subjects which cover web or app development, particularly with a focus on Java as the language would be beneficial. To bolster his fundamental organisational and project management skills he should take some software development courses which focus on different styles of the design, planning and modelling of software.

Case 2 – TalentFinders

Overview

TalentFinders is an agent who helps to match the employee CV with the company requirements based on job sector, skills, experience, ect. As a data scientist, you are hired to build a recommender system to provide top 10 jobs using a job market dataset suitable to a candidate's profile.

Proposed Solution

Generally, a resume or curriculum vitae (CV) typically contains the following information:

- Contact Information
- Summary or Objective Statement
- Education
- Work Experience
- Skills
- Certifications

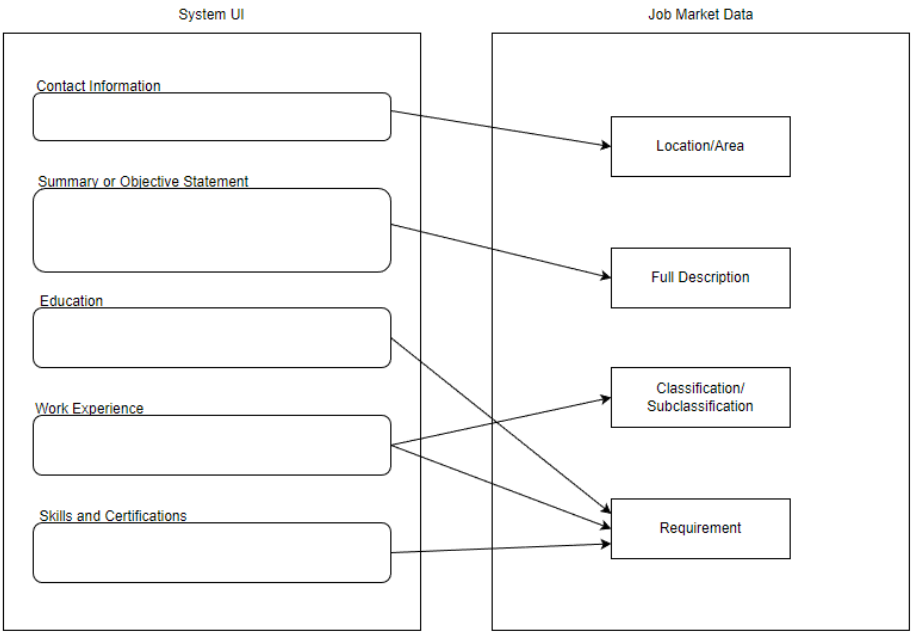
The job recommendation model should be built around these constraints or inputs. As the resume comprises of very few numerical values, a Content-based Recommendation system that follows an information retrieval approach using Natural Language Processing (NLP) would be appropriate. The proposed solution involves extracting keywords from selected sections of the resume and job market dataset, then computing their cosine similarity.

The first step involves separating sections of the resume and the dataset, then matching them accordingly. The easiest solution would be to have input boxes for each required section of the resume, so that they could be better identified. Determining the required data from the resume would be based on the information contained in the dataset. From the data.csv file, job information is comprised of:

- Title
- Company
- Location/Area
- Classification/Subclassification
- Requirement
- Full Description

The dataset contains other classifications/columns, but for the given task, few of them are necessary for comparison as seen below. For example, a section would be created where the user would put the Skills and Certifications from the resume, which the system would compare with Requirement section of the dataset. An illustration of all comparisons is shown in the figure below.

Figure 28. Recommendation System Framework



Preprocessing is the first step in Natural Language Processing, which will be used to find key words for the comparison.

1) Tokenisation would be the first step of preprocessing. Tokenisation involves taking the string input and splitting it by the individual words. Once done, any form of punctuation should be removed.
2) Using the tokenised set of words, stop words are next removed. Stop words are a set of commonly used words that add no value to the text, such as “and” or “the”. It may be important to convert the tokenised set of words to all lowercase before removing stop words, as depending on the library It may not recognise certain words due to the format differentiation.
3) Next, either Stemming or Lemmatisation should be applied. Both are text normalisation techniques which reduce words to a common base root. Stemming removes the last few characters from a word, where Lemmatisation does a similar act, but takes into account the context. Due to Lemmatisations generally higher accuracy, it will be selected for this application.

From here the data should be completely processed. The second step would be to extract keywords from the processed data. For this, techniques such as TF-IDF or Bag of Words (BOW) are algorithms which can be used to identify the most signification terms in each document (resume and dataset). Both provide the frequency of each word in a document, but TF-IDF uses a weighted factor, which is the ratio of documents that include the specific word compared to the overall number of documents.

Using the TF-IDF of the resume section with its corresponding TF-IDF of the dataset classification, a comparison is made between the resume and each individual job. This comparison can be completed with the use of algorithms such as Cosine Similarity or Jaccard. Cosine Similarity will be used for this case as it compares the entire vector of strings, resulting in higher accuracy.

A numerical value will be returned which corresponds to the similarity seen between each section of the resume and the attributes of each job advertisement. For each job advertisement, the summation of this value will represent its similarity to the resume. The 10 jobs with the highest level of correspondence will be returned.

References

[1] Consultancy.com.au. (2021, August 29). 10 charts on Australia's fast growing technology sector. https://www.consultancy.com.au/news/3869/10-charts-on-australias-fast-growing-technology-sector

[2] Australian Bureau of Statistics. (2021). Population: Census. ABS. https://www.abs.gov.au/statistics/people/population/population-census/2021.

[3] Martin, P. (2021, June 9). Other Australians earn nothing like what you think. If you're on $59,538, you're typical. ABC News. https://www.abc.net.au/news/2021-06-09/typical-australian-wage-less-than-you-might-think-typical/100198488