Data Science Job Listings in Australia

CHENG, KA WAI† LAM, CHING YAN‡ WONG, CHIN HANG§ WONG, KA CHUN¶

The Hong Kong University of Science and Technology

1 ABSTRACT

Data Science is a hot topic in current society nowadays, more and more industries started to hire data scientists to assist on the digitization of their companies as well as use advanced technology to improve their productivity. However, while data scientist is a new position, employees may not know the details of this job and how to make sure if they are suitable for this job. Therefore, we present the visualization of the data science job listings in Australia in 2019 to illustrate what insights can be found from this dataset especially for those who are looking for data scientist jobs in the job market recently. Our project visualizes some basic things job seekers need to know first and a deeper analysis between different attributes from diverse angles. Useful recommendations will be provided at the end of this report to help job seekers find a data scientist job in a more efficient and effective way.

2 INTRODUCTION

Data Scientist is an ideal job for all students who studied Big Data Technology. It is not a concrete idea to understand the content for this job category apart from the name directly. Therefore, a visualization is good for job seekers like us to explore more about it in a convenient way. In order to do so, we find a dataset from Kaggle containing around 1500 job listings from a famous job searching website in Australia named “Seek.com.au” from Jan 2019 to Feb 2020. The dataset is a collection of every search result for data scientist along with each of 25 programming languages and software applications. It provides over 50 columns of rich numeric, geographic and text data for us to explore.

There are diverse data attributes such as jobs descriptions, jobs classification, salary, skills, locations and nationality etc. This can allow us to deep dive in text analysis, geospatial analysis and time series analysis. We would like to provide answers to some basic questions like which job industry provide higher salary for data scientist jobs and which skillsets are the best for a person to equip with if he or she wants to get a data scientist job.

3 RELATED WORK

In order to visualize the data and further analyze the dataset, we need to perform data preprocess first. We use excel and Tableau to preprocess the data into a usable format. There are several types of preprocess we need to do. We need to remove the outlier and replace the null values with average values. Furthermore, we also need to revise the format from string to numerical for some attributes to further calculate as well as remove non-English characters, line breaks and double spacing. Last but not least, we need to group some attributers to more meaningful clusters to explore the relationship of the datasets.

Then we perform some data exploration to understand the dataset more and try to find out answers for some common questions.

The first goal is to find out the key words for job seekers to search for data scientists’ related job advertisements.

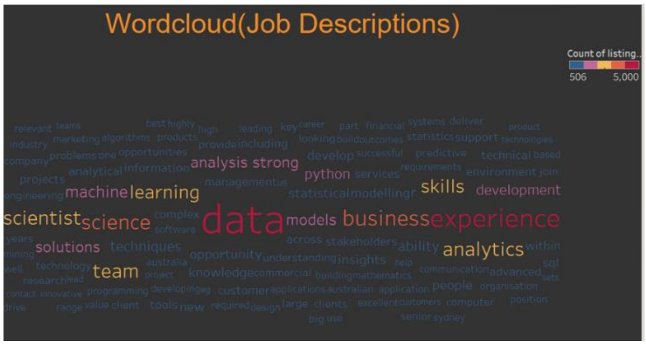


Figure 1: Wordcloud for analyzing job descriptions

The above worldcloud is use Tableau to analyze the job descriptions. Size and color encode the frequency of words appeared. From the wordcloud, we can know that data, machine, learning, python and analytics are the top 5 key words with the highest frequency appeared in the job descriptions related to data scientists’ jobs.

Next goal is to show the programming language and software requirement for data scientists’ related jobs.

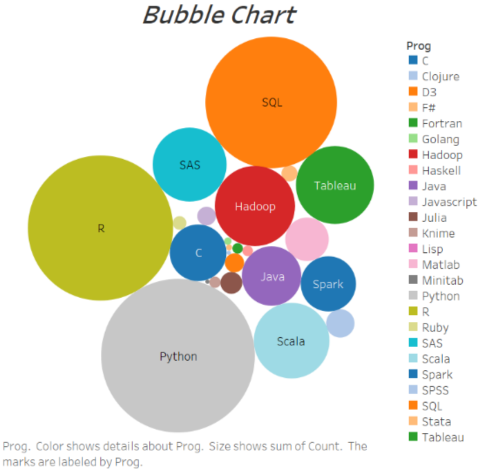


Figure 2: Bubble chart for analyzing skillsets appeared most

The above bubble chart is created using Tableau to calculate the frequency of different skills appeared in the job advertisements. Size of the bubbles refers to the frequency of words appeared while color of bubbles shows diverse programming language and software. From the chart, we find that Python, R and SQL are the

†e-mail: kwchengaj@connect.ust.hk
‡e-mail: cylambg@connect.ust.hk
§e-mail: chwongar@connect.ust.hk
¶e-mail: kcwongbs@connect.ust.hk

top three programming language that a data scientist job requires probably because of their various usage and freely usability.

The last goal is to find out the top 3 job industries that provide the most data science related jobs

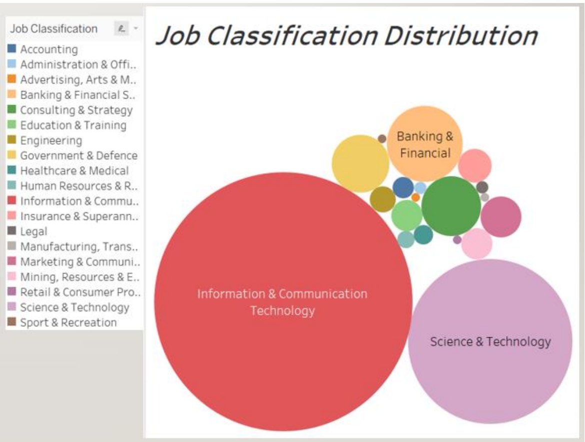


Figure 3: Bubble chart analyzing the job classification

The above bubble chart is created using Tableau to analyze the attribute of job classification. Size of bubbles encodes the frequency of job classification appeared and color of bubbles show a variety of jib industries. The results show us that IT, science and technology and Banking & Finance are the top 3 industries providing the most data scientist jobs. Job seekers can search these industries in order to find out more opportunities.

4 TASK AND REQUIREMENT ANALYSIS

Task 1: What factors affect salary?

The goal of the first task is to answer the most asked question by job seekers: What factors affect salary?

We picked the 3 most related attributes, namely seniority, job classification, and years of working experience, to visualize their impact on salary.

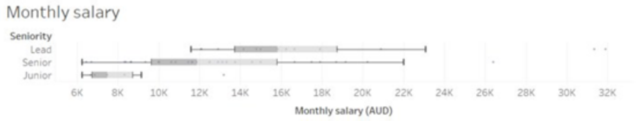


Figure 1.1: The Box and whisker plot of monthly salary against seniority

The Box and whisker plot above shows the monthly salary against seniority. It also shows a general trend of higher seniority, better the salary.

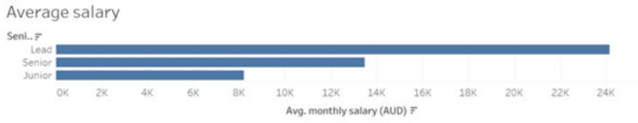


Figure 1.2: The bar chart of average monthly salary against seniority

The bar chart shown in Figure 1.2 shows the clear positive correlation between seniority and avg salary.

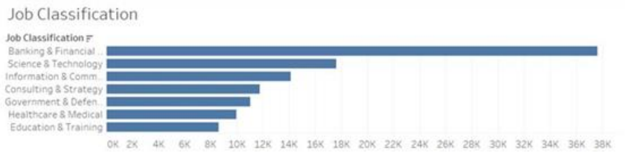


Figure 1.3: The bar chart of average monthly salary against job classification

The above bar chart shows that the banking sector has far more avg salary than other sectors.

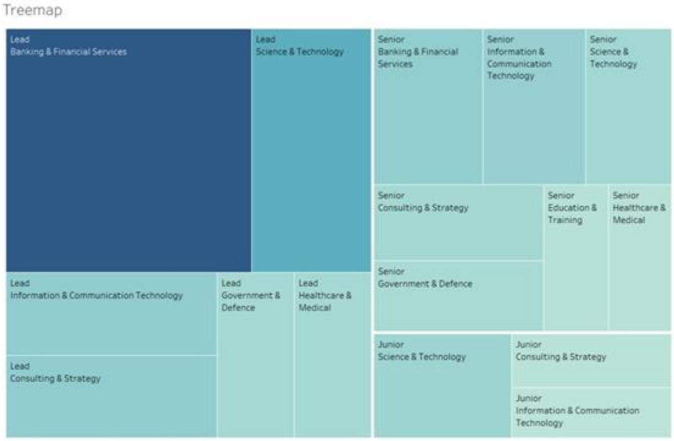


Figure 1.4: Tree map of monthly salary, job classification, and seniority

We use a tree map to visualize the combination of average salary, seniority, and job sectors. Left hand side of the tree map shows the group of Leader. Right hand side of the tree map shows the group of Senior and junior. Saturation of color encode average salary. The tree map is showcasing some findings. First, being a team leader in the banking sector is a huge privilege. Second the salary of team leader differs a lot among business sectors. Moreover, being a senior in the banking sector has higher average salary than being a team leader in government and healthcare.

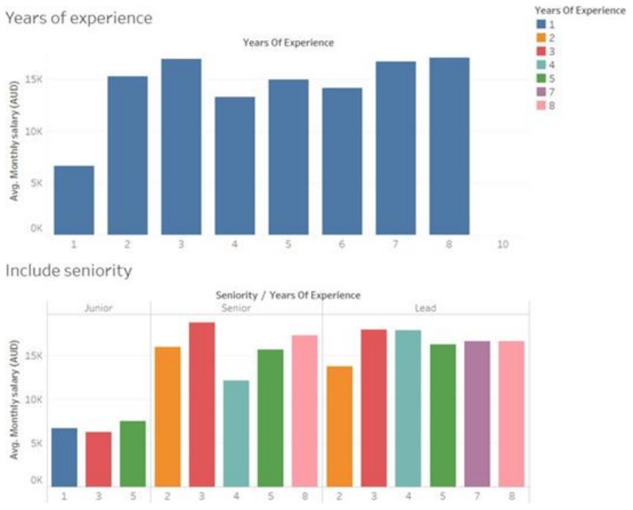


Figure 1.5: Bar charts of years of working experience and salary

The above bar charts show how years of working experience affect salary. Surprisingly, as shown from the bar charts, the years of working experience have no significant impact on average salary. The bar charts also shows that average salary has no significant difference within each seniority group.

In short, the data tells us putting effort in ranking up to be a team leader is more important than accumulates years of working experience. And developing career in Banking sector is a smart move.

Task 2: Where can job seekers find the job?

The goal of the second task is to answer another frequently asked question by job seekers: Where can they find the job?

We have built two dashboards by using Tableau to illustrate the salary and job opportunity by region. The first one is about the average monthly salary (AUD) from State by work type and job category. Another dashboard is to visualize the in-demand skills in specific state.

Firstly, we select the salary, work type, job category as the attribute in the dashboard to see how they are being affected by location.

Attribute	Values	Encoding Schema
Salary	AUD 7,000 – 15,000	Darkness of Colour
Work Type	Full Time, Part Time, Contract/Temp, Casual/Vacation	N/A (As Filter)
Job Category	Junior, Senior, Lead, Research, Null	Pie Chart Ratio (As Filter)

Figure 2.1: Table showing the encoding schema for figure 2.5

Based on the graph in upper part shown, more high-salary job opportunity is offered on New South Wales State, Victory State and Australian Capital Territory State. Job Seekers can select the work type on their need. The graph in bottom part shown the average monthly salary by job category. Job seekers can select the range of salary to find a job that can meet their expected salary.

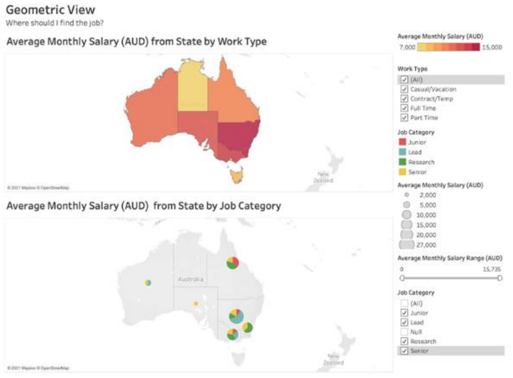
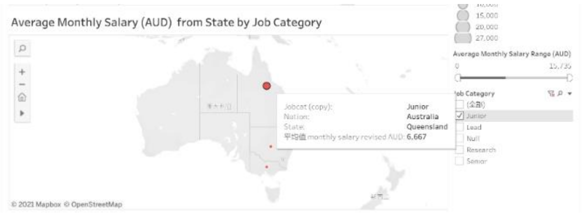


Figure 2.2: Dashboard of Average Monthly Salary (AUD) of Job by Geometric View

The colour of pie chart located on map represent the ratio of job category on the specific state. For example, there are apparently higher ratio of junior jobs (in red colour) advertised on Queensland State compared to other states.


Average Monthly Salary (AUD) from State by Job Category
15,000
20,000
27,000
Average Monthly Salary Range (AUD)
0
15,735
Job Category
☐ (全部)
☑ Junior
☐ Lead
☐ Null
☐ Research
☐ Senior
© 2021 Mapbox © OpenStreetMap

Jobcat (copy): Junior
Nation: Australia
State: Queensland
平均值 monthly salary revised AUD: 6,667

Figure 2.3: Dashboard of Average Monthly Salary (AUD) of Junior Job by Geometric View

Secondly, we select the skills and job count as the target attribute to illustrate the relationship between job openings and skillset by location.

Attribute	Values	Encoding Schema
States	Australian Capital Territory, Northern Territories, South Australia, etc.	Colour
Skills	Java, Tableau, Spark, etc.	N/A (As Filter)
Job Count (by Ranking)	1 - 8	Ranking (Sorted by 1 – 8) 1: The States with high demand on specific skill 8: The States with less demand on specific skill

Figure 2.4: Table showing the encoding schema for figure 2.5

The dashboard below allows Job Seekers to select the ranking range of specific skill to know the in-demand skill on specific state. Then they can more easily to know where to find the job opportunity if they equip the skills.

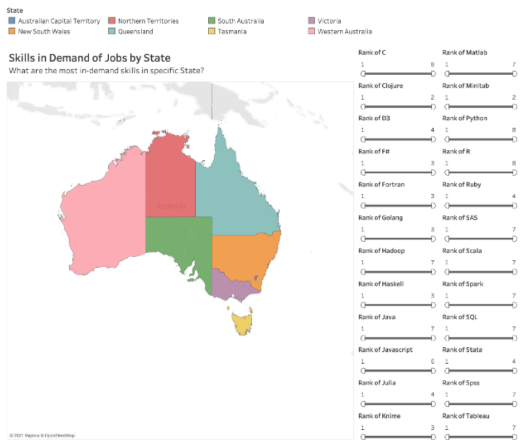


Figure 2.5: Dashboard of Skills in Demand of Jobs by Geometric View

For example, Python is one of the most competitive skill in Data Scientist Job, and it is relatively high demanded on New South Wales State, Queensland State and Victoria State. In other words, the job seeker who have expertise in python can apply more jobs on these states.

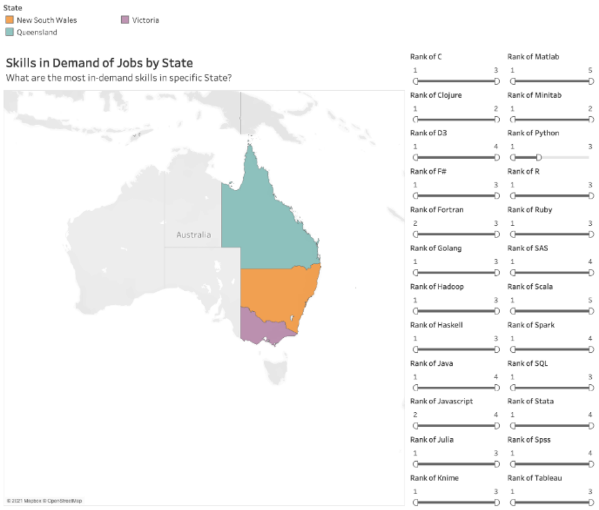


Figure 2.6: Dashboard of Skills in Demand of Jobs of Python by Geometric View

It is quite interesting to find that New South Wale is always the top 3 states in all skills in demanded. New South Wales, the Australia's most populous state with over 8.1 million population including the capital city Sydney, and it is reasonable that there are more job openings and relatively high demand for Data Scientist with different skillsets.

Task 3: What is the relationship of time and availability / competitiveness of jobs?

In view of time, one of the concerns that most job seekers care a lot about is what is the best time to look for a new job so they are able to reach more jobs and have a higher probability to find a suitable job. For example, job seekers may want to know what is the best time in a day or a week to browse job ads so they are able to view more job ads and have more jobs to select and apply. On the other hand, job seekers may want to know what is the best time in a year for them to resign from the current jobs and look for new opportunities. Graphs can be built to show numbers of job listings against different time units to discover its correlation with time.

Another concern of job seekers is the competitiveness of jobs in a domain or a particular set of skills. One clue we can find from the dataset is how early a job ad is removed before the expiry date. The earlier the job ad is removed, the easier for the employers to find a right candidate and thus the more competitive the job is. Graphs can be built to show the trends of early removed jobs ads based on different skills or other criteria and find out if there are any kinds of jobs trend to be removed early.

To show the correlation between numbers of job listings and time, a heat map and a stream graph were created. Figure 3.1 shows the heat map of numbers of job listings in every hour of a weekday. The color in the heat map indicates the number of job listings in a time slot of the day. The darker the color is, the more job ads are posted in the particular time slot. Figure 3.2 shows the stream graph of numbers of job listings in the aspects of different skills throughout the year. The color in the graph indicates different skills listed in job ads. The size of the stream indicates the number of job listings for the corresponding skill. The wider the stream is, the more job ads are posted in the particular period of a year for the corresponding skill.

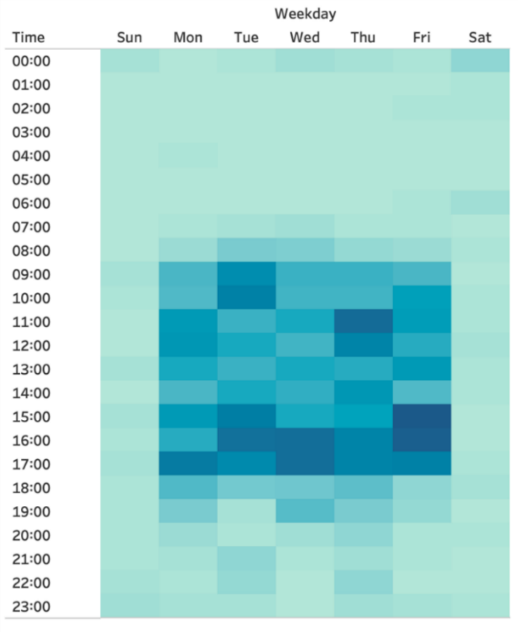


Figure 3.1: Heat map of numbers of job listings in every hour of a weekday

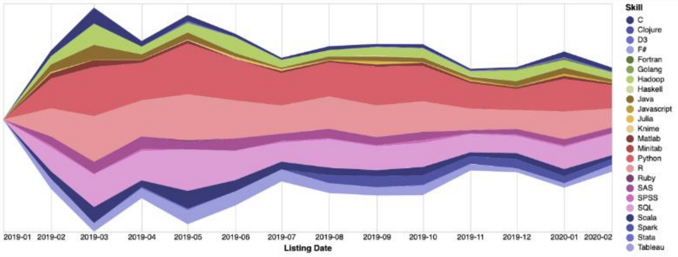


Figure 3.2: Stream graph of numbers of job listings in the aspects of different skills throughout the year

To show the trends of early removed jobs ads based on different criteria, a box plot and an area chart were created. Figure 3.3 shows the box plot of the distribution of days of early removal of job ads for different skills. The dispersion of the data is displayed by the five-number summary of the data as indicated in the chart. Only popular skills which were mentioned in more than 100 job ads are shown in the figure for an unbiased comparison. Figure 3.4 shows the area chart of average salary against days of early removal of job ads. The colored area in the chart indicates the changes of average salary as days of early removal of job ads increase.

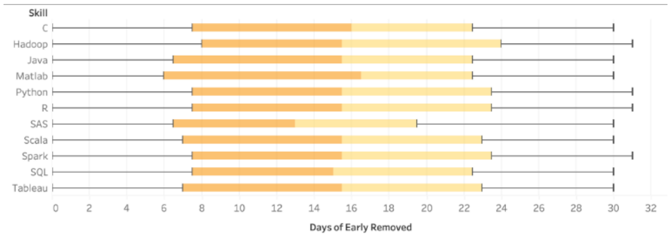


Figure 3.3: Box plot of the distribution of days of early removal of job ads for different skills

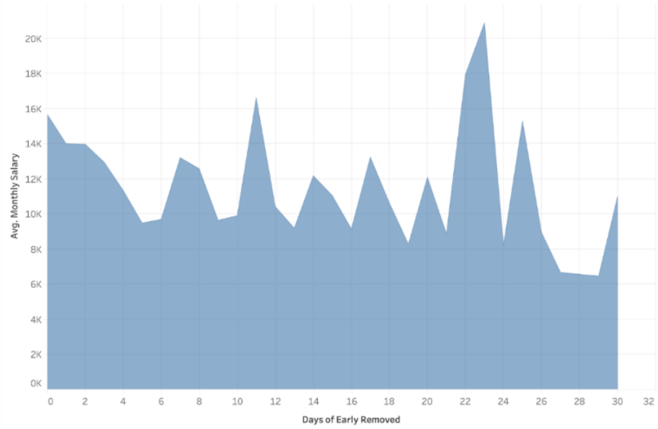


Figure 3.4: Area chart of average salary against days of early removal of job ads

From Figure 3.1, it is found that most job ads are posted in the afternoon of a working day. Particularly, the number of job listings in Friday afternoon is higher than the other time slots in the afternoon. Interestingly, there is a special case that the number of job listings at 11am of Thursday morning is also considerably high. Therefore, job seekers may consider to browse job ads when they are off work in working days as well as at noon of Thursday so that they can reach more job ads and have more choices when selecting jobs to apply.

From Figure 3.2, it is found that most job ads of different skills share similar trends, although some skills are more popular than the others such as Python, R, SQL, etc. There are more job ads posted in the first half of the year, particularly in March and May. However, job ads posted near the end of the year are fewer, particularly in November and December. Job seekers may therefore start to think of resigning the current jobs and looking for new jobs in March or May or other months in the first half of a year.

When looking at the mediums of days of early removal of job ads in the box plot of Figure 3.3, it is found that jobs of Matlab are the most competitive because the medium of Matlab is the highest among all the skills, i.e. job ads of Matlab are the quickest to be removed on average. Besides, the distribution of jobs of Matlab is left-skewed, i.e. most of job ads of Matlab are removed early. However, jobs of SAS are the least competitive because the medium of SAS is significantly less than the mediums of other skills. Therefore, for job seekers looking for jobs requiring Matlab, they may have to prepare themselves better because jobs of Matlab are the most competitive. Meanwhile, for job seekers looking for jobs requiring SAS, they may not have to be as anxious as others because jobs of SAS are the least competitive.

As for comparing average salary against days of early removal of job ads in Figure 3.4, it is found that the average salary is inversely proportional to the days of early removal of job ads, although there are special cases such that the line in the chart is not smooth. Jobs offering low salary therefore tend to be removed early. It may be due to the reason that there are more candidates of jobs offering low salary.

5 Conclusion and Future Works

In conclusion, we analyse the data scientists job listings dataset under three aspects and provide diverse visualizations to support our findings. We try to answer some common questions that all job seekers must be interested in and provide some suggestions to them so that they can search data scientist jobs more easily and effectively.

For future work, we find that visualization using Tableau still has some rooms to improve as Tableau cannot handle complex visualizations. We would like to explore other technique like d3.js and other business intelligence tools to try more sophisticated visualizations. Besides, we want to use larger datasets as the current one has not much data. We may try to combine with the job listings from other website over the world like Jobsdb in HK to do a globalized analytic visualization. Last but not least, we also want to find out more information for handling the null values in the dataset since we found that there are quite a lot of null values for different attributes and the current method how we handle the null data is still very simple.

6 Acknowledgements

The authors wish to thank Professor Huamin Qu and Teaching Assistant Leo Lo for their feedbacks on our related studies as well as their advice on the visual design along the whole period of project.