3030ICT/7130ICT

Data Analytics

Assignment Specification

Trimester 1 - 2023

Instructions

- **Due**: Wednesday, 31 May 2023 at 11:59 PM (Brisbane time)
- **Marks**: 30% of your overall grade
- **Submission**: via Canvas LMS on Learning@Griffith
    - You can submit a Jupyter notebook that contains the codes, charts, and the report (written directly in the notebook).
    - Or you can submit a zip file that contains the codes, charts, and report files.
- **Data**: https://drive.google.com/file/d/1rkJpU1syooFlmwFBKh5TVthVqPYBI8ps/view
- **Late Submissions**: Late submission is allowed but Penalty applies. The penalty is defined as is the reduction of the mark allocated to the assessment item by 5% of the total weighted mark for the assessment item, for each working day that the item is late. A working day will be defined as Monday to Friday. Assessment items submitted more than five working days after the due date will be awarded zero marks.
- **Extensions**: You can request for an extension of time on one of two grounds, as follows:
    - medical
    - other (e.g., family or personal circumstances, employment-related circumstances, unavoidable commitments).

  Please note that proof documents (e.g., medical certificate) are needs for the approval.
- **Group Work**: You should complete this assignment in a group of 2 students. Group of 1 or 3 students are allowed, but an explanation should be provided.
- **Difficulty**: *(slightly difficult), ** (difficult), *** (most difficult)

Overview

In this assignment, you will need to apply data analytics, using the tools introduced during the labs. You are required to study and analyze the SEEK job market data for which a dataset is provided. The general requirements include:

- Understand data characteristics using data preparation and preprocessing techniques. Then, you will perform various data analysis techniques to gain better understandings about the dataset.

- Perform more advanced techniques to explore the dataset. You also need to propose solutions for some challenges based on real-life situations.

- The primary dataset that we would like to use is the job market dataset which is provided in CSV format (data.csv). You can try to crawl new data from seek.com.au yourself but it is optional.
- Perform data preparation and preprocessing for your analysis
- **All the chart in the assignment are used for your reference. You are free to choose your own style.**

Part 1 – Data Preparation and Preprocessing. [15 points]

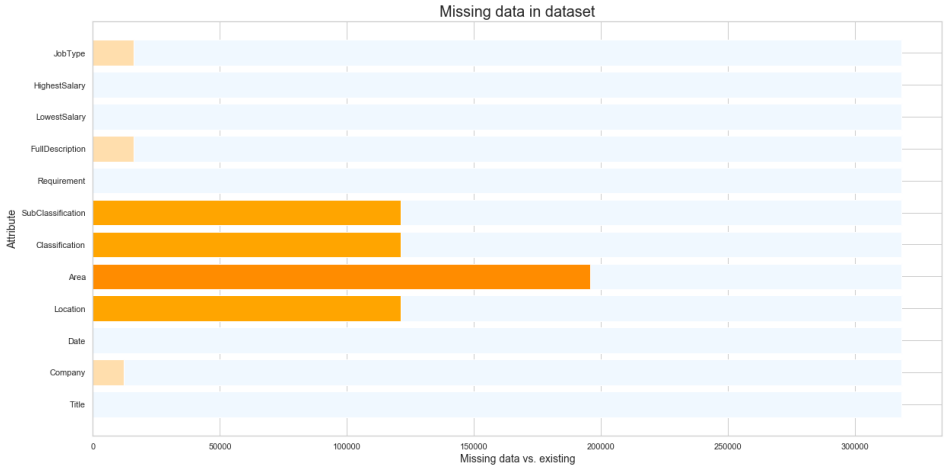
1. Describe the dataset. (8 points)

- ✓ Describe the dataset (e.g.: type of column, value range). (1 point)
- ✓ How many records are there in the dataset? (1 point)
- ✓ Which period does it cover? How many different dates have job postings? (1 point)
- ✓ How many locations does the dataset have? Which location has the most job postings? (1 point)
- ✓ How many job sectors(job classifications) are there in the dataset? List the name of each sector and its' total of job postings. (1 point)
- ✓ Choose your favorite job sector (e.g. Information & Communication Technology), how many sub-sectors are there in that sector? List the name of each sub-sector and its' job posting number. (1 point)
- ✓ List the salary ranges and their total of job postings. (1 point)
- ✓ List the job types. In each job type, what are the lowest salary and highest salary? (1 point)

2. Normalize and clean data. (7 points)

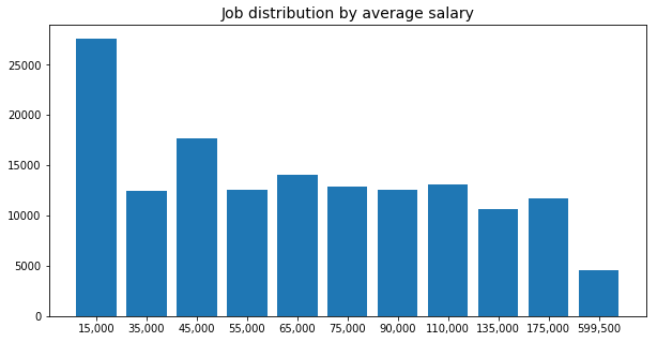
- ✓ The salaries are kept in the dataset as "HighestSalary" and "LowestSalary". You should calculate the "AverageSalary" for each job. (1 point)
- ✓ The raw dataset values of the "Id" column had inconsistencies in their representation. The Id values should have 8 number long integers only. Write code to remove unnecessary characters. (1 point)
- ✓ The "Date" column is represented in a format that contained both date and time information. However, the time is not correct and should be removed. (1 point)

- ✓ Change type of “Id” column to numeric and change type of “Date” column to DateTime. (1 point)
- ✓ Are there any duplicate data in the dataset? Provide the way you find them and your solution to fix it. ** (1.5 point)
- ✓ Check missing data and visualize them in a corresponding chart.** (1.5 point)

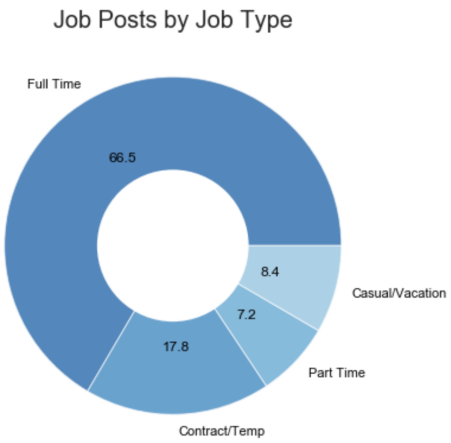


Part 2 – Data Understanding. [5 points]

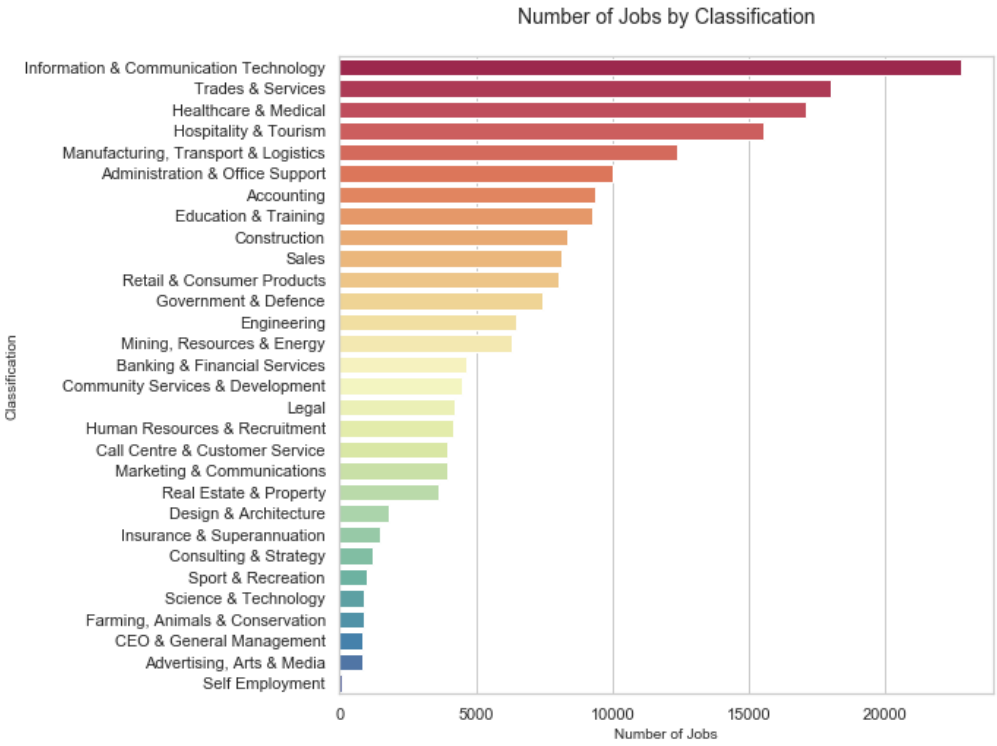
- ✓ Get the salary ranges using “AverageSalary”, the total jobs of each range and display them in the bar chart. (1 point)



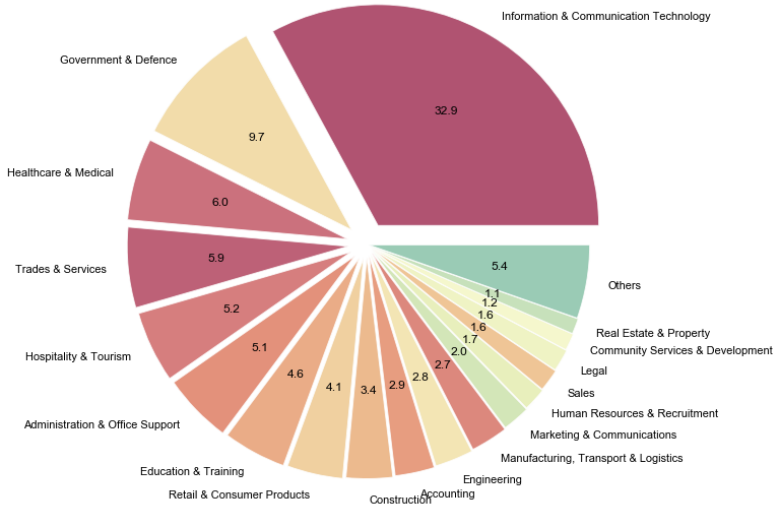
- ✓ Display the list of job types and the number of jobs of each type using pie chart. (1 point)



✓ Display the list of job sectors and the number of jobs of each type using horizontal bar chart. (1 point)



✓ Choose your favorite location. Visualize the market share of that location in pie chart. (1 point)



- ✓ Can you find the salary distribution for the top 30 cities for the number of job postings? Visualize them in the boxplot chart. (1 point)

Part 3 – Data Analysis and Visualisation. [6 points]

1. Analyse by comparison. (2 points)

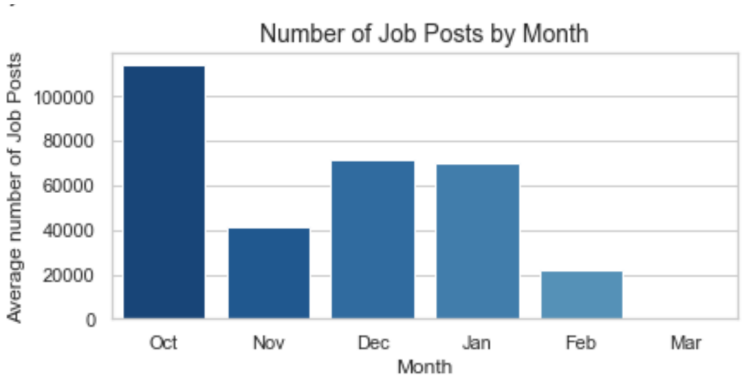
Choose your two favourite cities/locations. We will explore the difference between them by answering the following questions:

- ✓ Which city has more job? How many jobs each type (casual, fulltime, etc.) are there in each city?
- ✓ In each city, which are top 5 job sectors? How many jobs are there in each sector?
- ✓ Visualise the top 5 job sectors in pie chart for each city.*
- ✓ In each city, list the job salary range with the corresponding number of jobs. Which city is more well-paid?
- ✓ List top 5 companies in each city? Which sectors do they belong to?
- ✓ Between 2 cities, which do you think it is better for employees. Explain your choice.*

2. Analyse by time. (2 points)

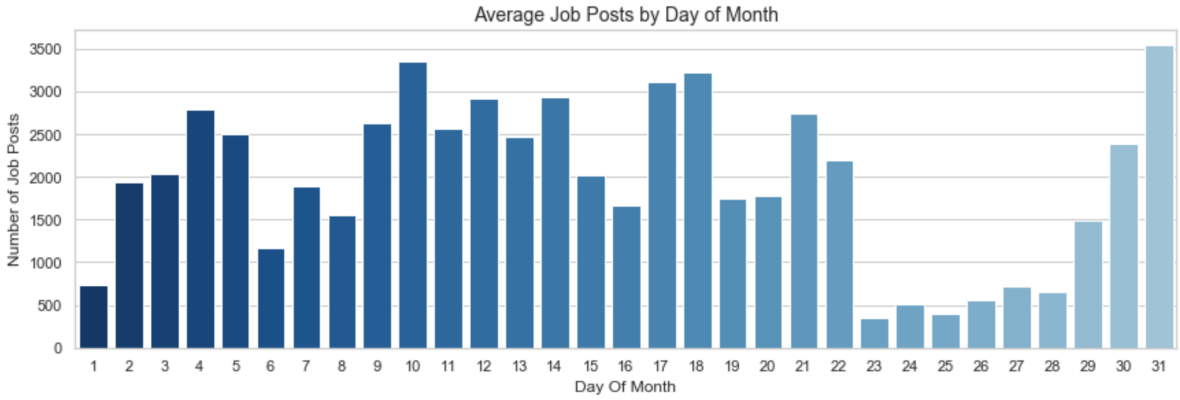
In this part, we will analyse the jobs based on the posting date.

- ✓ Visualise the number of job posts by month

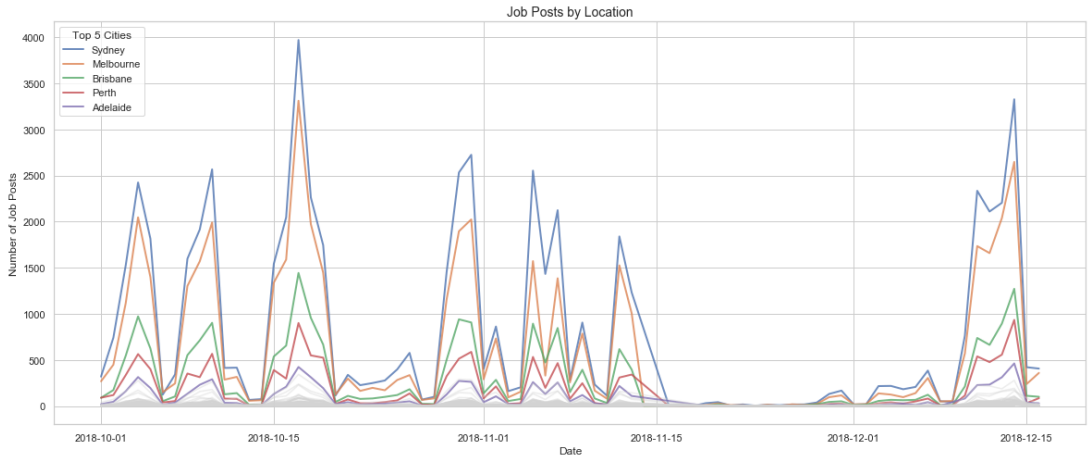


- ✓ Visualise the number of job posts by day of week.

- ✓ Visualise the number of job posts by day of the month.



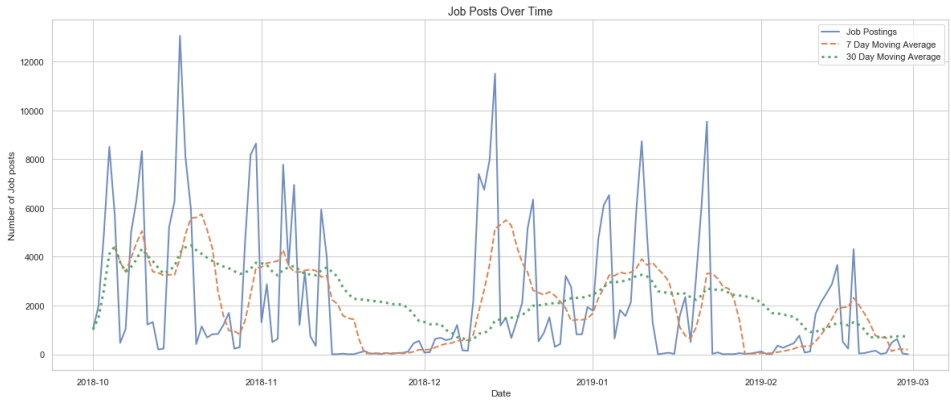
✓ Visualise trending of the job postings for the big cities.*



✓ Based on the above charts, provide your observation about the number of job posts over time. Give your own conclusion.

3. Forecasting and skill extractions. (2 points)

✓ Using moving average for 7 days and 30 days to predict the number of job postings and visualise them in line chart. Which one creates a better prediction? Explain your choice.**



- ✓ Choose your favourite job sector/sub-sector, then use TF/IDF to extract important keywords. Visualise them in word cloud chart (hint: you can use the online tool https://wordart.com/create or similar websites)**.

Part 4 – Discussion. [4 points]

- ✓ **Scenario 1:** Tom, a grade 12 student in Brisbane, has good results in all of the subjects in his school. He is finding a major in the University which can guarantee a job in his state (Queensland) with a good income (>=80K) in the future. Based on the current job market dataset, which major and the related subjects do you recommend him? Explain your choice.*** (Max 750 words)
- ✓ **Scenario 2:** To gain a better reputation, Griffith University wants to attract more students by enhancing the employability of the graduates. As a data analyst, you need to give the recommendation for the management board for the changing of the admission numbers in each major. Which majors should be extended, and which majors should be reduced (Assume that we only need to use job market data, regardless of other factors such as social impact, economic impact, etc.)? Justify your recommendation.*** (Max 750 words)