3803ICT

Big Data Analysis

Assignment Specification

Trimester 1 – 2024

Instructions

- **Due**: Sunday, 26 May 2024 at 11:59 pm (Brisbane time)
- **Marks**: 35% 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/1rkJpU1syooFImwFBKh5TVthVqPYBI8ps/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. The assignment consists of 3 parts. In the first part, you will need to understand data characteristics using data preparation and preprocessing techniques. In the second part, you will perform various data analysis techniques, including exploratory, statistical, and predictive analysis. In the third part, you will need to evaluate your findings and determine appropriate future actions.

Part 1 –Data Preparation and Preprocessing [5 points]

- 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.
- Submit your Jupyter notebook in your Github repository.

1) Describe the dataset.

   For example:

   - What are the categories/domains of the dataset?
   - What is the dataset size of each variation?
   - What is dataset structure/format?
   - What are attributes/features of data you are going to use?
   - Which parts of the dataset will you use or all of them?

   [1-2 paragraphs, 2 points]
2) Describe the steps you used for data preparation and preprocessing. For example:

   - How do you load the data using Pandas?
   - How do you normalize the data?
   - How do you clean the data?

   [2-3 paragraphs, 2 points]
3) What is your hypothesis (expectation) about the analysis outcome?

   [1-2 paragraphs, 1 point]

Part 2 – Data Analysis and Interpretation [10 points]

- Perform exploratory data analysis.
- Perform statistical data analysis.
- Perform predictive data analysis.

- Submit your Jupyter notebook in your Github repository.

1) Study the job metadata. Extract the relevant information to describe the job's attributes.

   For example:
   - What is the sector, sub-sector of each job?
   - Where is the location of the job?
   - Which is the range of salaries for each job?

   [1-2 paragraphs, 2 points]
2) Study the market by locations.

   For example:
   - What is the market size in each city? Which are the hottest job sectors in each city?
   - Which range of salary is common in each city? Where are the employees more well-paid?
   - Can you detect the pattern of posting: e.g., are more jobs posted at the beginning of the month?

   [1-2 paragraphs, 3 points]
3) Study the market by sectors.

   For example:
   - Which sectors keep the highest market share?
   - In each sector, which sub-sectors are the main spotlights?
   - What is the salary range for each sector/sub-sector? Can you compare the salary range between sectors/subsectors?
   - What is the trending of the market, i.e., if a high school student asks you which subject should he/she learn in the university (to guarantee a job in the future), what is your advice?
   - Can you detect which skills are required in each sector?

[1-2 paragraphs, 3 points]

4) Visualize the results on an interactive visualization (web page, plotly, etc.)

   For example:

   - Trend analysis: visualize the number of jobs by locations, sectors, etc.
   - Compare between locations or sectors about the number of jobs, the salary, etc.
   - Present the necessary skills by sectors, by subsectors.

   [1-2 paragraphs, 2 points]

Part 3 - Evaluation [5 points]

1) What are the findings of your data analytics for the above sections?
   [2-3 paragraphs, 2 points]
2) What actions for balancing the markets do you suggest based on your findings?
   [1-2 paragraphs, 1 point]
3) How could you refine your data analytics?

   For example:

   - Could you use different data sources?
   - Could you choose different parameters?
   - Could you choose other techniques?
   - Can you think of ways to obtain more relevant data?

   [1-2 paragraphs, 1 point]
4) Are there any implications for employers and employees based on the findings you obtained? Justify your answer.
   [1-2 paragraphs, 1 point]

Part 4 – Case Studies [10 points]

- ✓ **Case study 1:** Mathew is a first year student in computer science. He wants to be an expert in the future, but he is unsure about the skillsets he should learn and improve. Based on the current job market dataset, which subjects and skills do you recommend him? Explain your choice.*** (5 points)
- ✓ **Case study 2:** 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. Propose your solution to solve this problem.*** (5 points)