COMP517–Data Analysis

ASSIGNMENT TWO

Semester 2, 2024

Student Names and ID: Provide the name and ID for all group members.

Due Date: Midnight Friday 18th Oct 2024

Total Marks: 100

Submission Guidelines:

Please ensure that the front page of your report includes the names and student IDs of all group members.

Important Notes:

- This is a **group assignment**, and each group should consist of a maximum of **two students**. It is mandatory for all students within the group to contribute to each part of the assignment.
- Each team is required to complete a peer assessment, which can be found in the assignment 2 material.
- Only one submission is accepted per group.
- Late submissions will incur a 10 marks penalty per day.

Introduction

KiwiLearn is a leading education provider based in New Zealand, specializing in providing high-quality educational resources to tertiary students. Over the years, KiwiLearn has established itself as a trusted name in the education sector. With the rise of online education, KiwiLearn has adapted and expanded its services to meet the evolving needs of students. KiwiLearn recognizes the importance of effectively reaching out to students and institutions. The Sales and Marketing department is responsible for student outreach, partnerships with educational institutions, and marketing campaigns. The IT department plays a crucial role in maintaining KiwiLearn's online platforms, ensuring smooth user experiences, and implementing technical solutions to support the educational services. KiwiLearn believes in the continuous development of its employees. They provide training hours for skill enhancement and career growth. Employees in all departments have access to ongoing training programs to stay updated with the latest trends in education and technology.

KiwiLearn is enthusiastic about examining its data to gain insights for future decision-making. Consequently, you have been brought on board as a data analyst to provide your expertise and support in this endeavour. KiwiLearn has provided you with a dataset (Employee_Performance.csv) containing the following information for its employees:

- **Employee ID:** A unique identifier assigned to each employee.
- **Department:** The specific department in which the employee is employed (e.g., Sales, Marketing, HR, IT).
- **Gender:** Designation of the employee as either Male or Female.
- **Years of Experience *:** The total number of years of professional experience possessed by the employee, ranging from 0 to 9 years.
- **Salary:** The monthly income earned by each employee.
- **Performance Rating:** A performance rating attributed to each employee, measured on a scale from 1 to 5.5. This scale represents the employee's performance, with 1 indicating poor performance and 5.5 indicating exceptional performance.
- **Training Hours:** The quantity of training hours that each employee has undergone within the past year.

* If years is greater than or equal to 8, it sets experience_category to "Senior". If years is greater than or equal to 5 (but less than 8), it sets experience_category to "Mid-level". If years is greater than or equal to 2 (but less than 5), it sets experience_category to "Junior". Otherwise, it sets experience_category to "Entry-level".

Part One: Exploring Data and Testing Hypotheses: Uncovering Insights from Dataset

To gain insights into how various departments within the organization assess and evaluate different aspects of the workplace, your task is to perform an analysis aimed at investigating potential variations in *employee performance rating* across these departments. If such variations exist, your goal is to identify the department(s) that exhibit notably higher or lower ratings.

Task 1: Data Preparation and Exploration (5 marks)

- Perform initial data exploration to understand the characteristics of the dataset. This includes summary statistics, data distribution etc.
- Perform multivariate analysis of data to visualize relationships between employees' years of experience on their performance ratings within different departments.

Task 2: Assumptions, and Hypothesis Formulation (10 Marks)

- Begin by defining your analysis's objective.
- Identify any assumptions you've made before conducting the analysis.
- Formulate your null and alternative hypotheses based on the objective of your analysis.

Task 3: Statistical Technique: Hypothesis Testing (20 marks)

- Explain the statistical method you employed and why it's suitable for this analysis.
- Perform the test using your dataset. Include the relevant outputs (F-statistic, *p*-value, critical value) in your results.
- If you find a significant difference in ratings across departments, conduct Tuky's post-hoc test to identify specific department pairs with significant differences. Include relevant statistical values and measures.
- Present the results obtained from both analyses, including any significant findings regarding variations in employee ratings among different departments.

Task 4: Discussion and Conclusion (10 marks)

- In the discussion section, elaborate on the implications of the results. Explain potential reasons for differences in ratings among departments and consider any actionable insights that can be derived from the analysis.
- Summarize the key takeaways from the analysis in your conclusion. State whether the analysis revealed significant differences in departmental ratings and which departments, if any, stood out in terms of higher or lower ratings. Ensure your conclusion is supported by the data and analysis you have conducted.

Part Two: Regression Analysis

Based on your understanding of the scenario and the employee-related information provided in the dataset, your task is to explore and uncover the underlying relationships between various parameters associated with employee performance. Utilize regression analysis to reveal these relationships and gain insights into how different factors may influence or correlate with employee performance ratings.

Task 1: Identify Potential Predictor Variables (5 marks)

- Generate ideas and identify possible independent variables within your dataset that may have a correlation with employee's performance rate. Provide explanations and reasoning to support your choices.

Task 2: Assumptions for Regression Analysis (10 marks)

- Outline the assumptions necessary for conducting regression analysis.
- Explain why these assumptions are relevant to your analysis.
- Which assumption should be tested before performing regression? Conduct the test and present the results along with explanations.

Task 3: Regression Analysis (10 marks)

- Perform multiple linear regression analysis using the identified predictor variables. Include the model output with all relevant parameters.

Task 4: Assumptions of Linear Regression (10 marks)

- Present the result of the remaining assumption to validate the reliability of the regression analysis and the accuracy of its results.

Task 5: Discussion and Conclusion (10 marks)

- Interpret the results and discuss the relationships between predictor variables and employee's performance rating. Identify which variables, if any, are significantly associated with performance rating.
- Draw a well-supported conclusion that summarizes the key insights from your analysis.
- Discuss any limitations of the analysis and suggest potential areas for further research.

There will be 10 marks for the presentation of the assignment including spelling and grammar, layout, formatting, and readability of the figures.

Submission Instructions:

Please submit the following two files as part of your assignment:

1. **Python Notebook or Code File (.ipynb, .py)**:
    - Ensure that your code is clean, well-organized, and properly commented.
    - The code must be ready to execute without errors.
2. **Report File (PDF Format)**:
    - The report file should be in PDF format. Your report must include the following elements:
        - **Title, Full Name, and Student ID**: Clearly state your title, full name, and student ID at the beginning of the report.
        - **Table of Contents**: Include a table of contents to provide an overview of the report's structure.
        - **List of Figures/Tables**: Provide a list of figures and tables used in your report for easy navigation.
        - **Answers to Questions**: Present your answers to the questions asked. Explain your findings, insights, and observations in a clear and concise manner.
        - **Figures (Plots) and Tables**: Include all relevant figures and tables that support your answers. However, DO NOT include the actual code used to generate these visualizations and tables.
        - **Informative Labels and Captions**: Ensure that all visualizations and tables have informative labels and captions with suitable resolution to help the reader understand their significance.
    - Please note that the report should focus on presenting your findings and insights, rather than including the code itself. Please refrain from including the code file in your report, as including code in the report will result in a penalty.

Note:

- Plagiarism is strictly prohibited. Ensure that you acknowledge the sources of code snippets, datasets, or ideas used in your individual analysis.
- Seek help from your instructor for any clarifications or guidance during the assignment but remember that the analysis and preprocessing tasks should be done individually.