****

**SKEE 1033: SCIENTIFIC PROGRAMMING**

**SESSION/SEMESTER: 2024/2025-1**

**GROUP ASSIGNMENT: DATA ANALYTICS**

**SCENARIO 3**

**SECTION:** 03

**LECTURER:** DR. MUHAMAMD AL FARABI BIN MUHAMMAD IQBAL

**GROUP MEMBERS:**

1. ISAAC CHAN XIAN YU
2. KAM YU SHENG
3. KOK QI SHENG
4. NICHOLAS LEE PAY ZHI
5. YEE DING HENG

**List of Content**

|  |  |  |
| --- | --- | --- |
| **No.** | **Title** | **Page** |
| **1.** | **Introduction** |  |
| **2.** | **Procedure** |  |
|  | Task 1: Data Cleaning Process |  |
|  | a) Examine the dataset for missing values and inconsistencies |  |
|  | b) Detect the outliers |  |
|  | Task 2: Descriptive Data Analysis |  |
|  | a) Mean, Median, Mode and Standard Deviation |  |
|  | b) Perform a bivariate analysis |  |
|  | Task 3: Data Visualization |  |
|  | a) Histogram/Box Plots |  |
|  | b) Scatter Plot Matrix |  |
|  | Task 4: Predictive Analysis |  |
|  | a) Linear Regression Model |  |
|  | b) Evaluation of predictive model using R-squared and Mean Absolute Error (MAE) |  |
| **3.** | **Work division** |  |
| **4.** | **References** |  |

**Introduction**

You are provided with a dataset containing information about university students' performance over the past three years. The dataset includes features such as attendance rate, assignment scores, study hours, participation in extracurricular activities, and final exam scores. Your objective is to analyse the data to identify key factors impacting academic performance and build a predictive model to forecast students' final exam scores based on these factors. With the knowledge of scientific programming, you are required to WRITE THE SCRIPT for each task together with comments that explain the code.

**Objectives**

**Task 1: Data Cleaning Process**

1. Examine the dataset for missing values and inconsistencies. Document how many data points are missing for each column and explain your chosen method for handling them using Python (pandas).
2. Detect and handle outliers in the study hours, assignment scores, and final exam scores columns using techniques like the Interquartile Range (IQR) or Z-score method.

**Task 2: Descriptive Data Analysis**

1. Generate descriptive statistics (mean, median, mode, standard deviation) for the key variables: attendance rate, assignment scores, study hours, and final exam scores. What insights can be drawn from these statistics?
2. Perform a bivariate analysis to investigate the relationship between attendance rate, study hours, and final exam scores. Use correlation coefficients and explain your findings.

**Task 3: Data Visualization**

1. Create visualizations such as histograms or box plots to represent the distribution of study hours and final exam scores. What patterns or trends do you observe?
2. Construct a scatter plot matrix or heatmap to illustrate the relationships between attendance rate, study hours, assignment scores, and final exam scores. Identify any notable trends or strong correlations.

**Task 4: Predictive Analysis**

1. Develop a predictive model using linear regression to estimate students' final exam scores based on attendance rate, study hours, and assignment scores. Provide a rationale for your choice of model and describe the data preprocessing steps taken.
2. Evaluate the predictive model using metrics such as R-squared and Mean Absolute Error (MAE). Discuss what these results imply about the accuracy and reliability of the model.

**Procedure**

**Work division**

|  |  |  |
| --- | --- | --- |
| **Member Name** | **Assigned Task** | **Role Description** |
| Kok Qi Sheng | Task 1 – 4: Scripting and Director | Implements all coding tasks for the project, including data cleaning, analysis, visualization, and predictive modelling. Oversees integration and project quality. |
| Yee Ding Heng | Task 1: Data Cleaning Process | Explains the data cleaning methods, including handling missing values, resolving outliers, and ensuring the dataset's integrity. |
| Nicholas Lee Pay Zhi | Task 2: Descriptive Data Analysis | Analyses and interprets the descriptive statistics and bivariate analysis, presenting insights into the relationships between key variables. |
| Kam Yu Sheng | Task 3: Data Visualization | Interprets and explains the visualizations (histograms, box plots, scatter plot matrices, heatmaps) to identify trends, patterns, and correlations in the data. |
| Isaac Chan Xian Yu | Task 4: Predictive Analysis | Reviews and explains the predictive analysis results, including the model's performance, metrics (R-squared, MAE), and implications for academic performance. |

**References**

1. **YouTube, *Python and Pandas with Reuven Lerner***:  
   Interpolating missing values (NaN) in Pandas DataFrames.  
   <https://youtu.be/zTBiqca-32k?si=1viKal6lFA_kAw38>
2. **Akash Srivastava, *Medium***:  
   Detecting anomalies with Z-scores: A practical approach.  
   [https://medium.com/@akashsri306/detecting-anomalies-with-z-scores-a-practical-approach-2f9a0f27458d](https://medium.com/@akashsri306/detecting-anomalies-with-z-scores-a-practical-approach-2f9a0f27458d%20)
3. **Learn Statistics with Bran**:  
   Outliers in data analysis and how to deal with them.  
   <https://www.youtube.com/watch?v=3lQydBqWYk0>
4. **Geeks for Geeks**:  
   Use Pandas to Calculate Stats from an Imported CSV File.  
   <https://www.geeksforgeeks.org/use-pandas-to-calculate-stats-from-an-imported-csv-file/>
5. **YouTube, *The Math Channel***:  
   Using a bivariate scatterplot.  
   <https://youtu.be/nyDQV-LE8B8?si=OBxEDQHdLR5hYn7f>
6. **Upendra Kumar, *Kaggle***:  
   Bivariate and Multivariate Analysis.  
   <https://www.kaggle.com/code/imkushwaha/bivariate-multivariate-analysis>
7. **100 Days of ML**:  
   In-depth analysis of histograms and box plots in Python.  
   <https://100daysofml.github.io/Week_04/Lesson_18.html>
8. **Muhammed Tausif, *Kaggle***:  
   Scatterplot matrix with Seaborn.  
   <https://www.kaggle.com/code/muhammedtausif/scatterplot-matrix-with-seaborn>
9. **Caleb Curry**:

*Machine Learning Tutorial 1*: Introduction to Machine Learning and AI

*Machine Learning Tutorial 2*: Introduction to Predictive Data Analytics

*Machine Learning Tutorial 3*: Introduction to Models

<https://www.youtube.com/watch?v=ZhNpTiY4NcU&list=PL_c9BZzLwBRIPaKlO5huuWQdcM3iYqF2w>

1. **Infinite Codes**:

All machine learning algorithms explained in 17 minutes

<https://youtu.be/E0Hmnixke2g?si=I7slc1BE7ThMasIN>

1. **Alejandro AO - Software & AI**:

Linear regression in Python - Full project for beginners

<https://youtu.be/O2Cw82YR5Bo?si=iufGErrHFiCIzlb9>

1. **Rob Mulia**:

Exploratory Data Analysis with Pandas Python

<https://youtu.be/xi0vhXFPegw?si=H-eA1HQrKLq46LMW>

1. **Learn with Ankith**:

Data cleaning/Data preprocessing before building a model - A comprehensive guide

<https://youtu.be/GP-2634exqA?si=9SiMiNRA0gUCKWLe>

1. **Codersarts**:

Robust Scaler Normalization with Pandas and Scikit-learn

<https://youtu.be/q9yFNcU4zR0?si=6FnOs-ixgqtt6Y9_>

1. **StatQuest with Josh Starmer**:

R-squared, clearly explained

<https://youtu.be/bMccdk8EdGo?si=b3U3lJ53ExnBtj_A>

Linear regression, clearly explained

<https://youtu.be/nk2CQITm_eo?si=3ngAJaoSmgDIY0il>

1. **Greg Hogg**:

What is the Mean Squared Error (MSE) in Machine Learning?

<https://www.youtube.com/shorts/6MzKADkfxUI>