W O M Fernando

Student No:  COMScDS241P-019

PRINCIPLES OF DATA SCIENCE

PORTFOLIO & GROUP ASSIGNMENT

Table of Contents

[Question 01 3](#_Toc191824916)

[Question 02 3](#_Toc191824917)

[References 5](#_Toc191824918)

# Question 01

1. **Data Collection:**

This dataset contains 5000 records from a private learning provider, and this includes key attributes to explore data, calculate correlations and insights related to academic performance.

A screenshot of a test

Description automatically generated

1. **Exploratory Data Analysis (EDA):**

Data cleaning and preprocessing

Detailed description of the dataset:

A table with numbers and a number of people

Description automatically generated with medium confidence

As per the above descriptive detail’s dataset contained few missing values for the Attendance and Assignments Avg columns.

Following columns are identified as categorical data in the dataset:

* Gender
* Department
* Grade
* Extracurricular Activities
* Internet Access at Home
* Parent Education Level
* Family Income Level

Following columns are identified as numerical data in the dataset:

* Age
* Attendance (%)
* Midterm Score
* Final Score
* Assignments Avg
* Quizzes Avg
* Participation Score
* Projects Score
* Total Score
* Study Hours per Week
* Stress Level (1-10)
* Sleep Hours per Night

Following are the number of categories and their counts:

* Gender
  + Male – 2551
  + Female – 2449
* Department
  + CS – 2022
  + Engineering – 1469
  + Business – 1006
  + Mathematics – 503
* Grade
  + A – 1495
  + B – 978
  + C – 794
  + D – 889
  + F – 844
* Extracurricular Activities
  + No – 3493
  + Yes – 1507
* Internet Access at Home
  + Yes – 4485
  + No – 515
* Parent Education Level
  + PhD – 820
  + Bachelor’s – 810
  + High School – 796
  + Master’s – 780
* Family Income Level
  + Low – 1983
  + Medium – 1973
  + High - 1044

Identified Null values in the following columns:

Impute the missing numerical values with the mean. And impute the missing categorical values with the mode.

Missing numerical: Attendance (%), Assignments Average, Parent Education Level

A screenshot of a computer program

Description automatically generated

Descriptive statistics and visualizations

Descriptive statistics after cleaning the dataset

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **50%** | **75%** | **Max** |
| Gender | 5000 | 0.5 | 0.50 | 0 | 0.0 | 1.0 | 1.0 | 1 |
| Age | 5000 | 21.0 | 1.99 | 18 | 19.0 | 21.0 | 23.0 | 24 |
| Department | 5000 | 1.3 | 0.90 | 0 | 1.0 | 1.0 | 2.0 | 3 |
| Attendance (%) | 5000 | 75.4 | 13.61 | 50 | 64.7 | 75.4 | 86.2 | 100 |
| Midterm\_Score | 5000 | 70.3 | 17.21 | 40 | 55.5 | 70.5 | 85.0 | 100 |
| Final\_Score | 5000 | 69.6 | 17.24 | 40 | 54.7 | 69.7 | 84.5 | 100 |
| Assignments\_Avg | 5000 | 74.8 | 13.65 | 50 | 63.7 | 74.8 | 85.6 | 100 |
| Quizzes\_Avg | 5000 | 74.9 | 14.50 | 50 | 62.5 | 74.7 | 87.6 | 100 |
| Participation\_Score | 5000 | 5.0 | 2.89 | 0 | 2.4 | 5.0 | 7.5 | 10 |
| Projects\_Score | 5000 | 74.9 | 14.42 | 50 | 62.3 | 75.0 | 87.4 | 100 |
| Total\_Score | 5000 | 75.1 | 14.40 | 50 | 62.8 | 75.4 | 87.7 | 100 |
| Study\_Hours\_per\_Week | 5000 | 17.7 | 7.28 | 5 | 11.4 | 17.5 | 24.1 | 30 |
| Extracurricular\_Activities | 5000 | 0.3 | 0.46 | 0 | 0.0 | 0.0 | 1.0 | 1 |
| Internet\_Access\_at\_Home | 5000 | 0.9 | 0.30 | 0 | 1.0 | 1.0 | 1.0 | 1 |
| Parent\_Education\_Level | 5000 | 2.0 | 1.15 | 0 | 1.0 | 3.0 | 3.0 | 3 |
| Family\_Income\_Level | 5000 | 1.2 | 0.75 | 0 | 1.0 | 1.0 | 2.0 | 2 |
| Stress\_Level (1-10) | 5000 | 5.5 | 2.86 | 1 | 3.0 | 5.0 | 8.0 | 10 |
| Sleep\_Hours\_per\_Night | 5000 | 6.5 | 1.45 | 4 | 5.2 | 6.5 | 7.7 | 9 |

Histograms for numerical features:

A group of blue bars with numbers and lines

Description automatically generated with medium confidence

Correlation Heatmap:

A blue and red chart with red squares

Description automatically generated

Scatter plots for key relationships:

A screenshot of a graph

Description automatically generated

Box plots to identify outliers:

A screenshot of a graph

Description automatically generated

1. sdfsdfs

# Question 02

# References