Assessment 2

43031 - Python Programming for Data Processing

Students Grading Analysis

Somesh
Shanbhag

25525837

Table of Contents

Introduction

This report documents the steps taken to process and analyze a student performance and behaviour dataset using Python and its powerful data processing libraries. The task involved mounting a google drive to access the dataset, reading it into a pandas data frame, performing exploratory data analysis (EDA) and identifying data quality issues. The overall aim was to gain insight into student performance by applying statistical technique and correlation analysis using Python.

Dataset

The dataset used in this assignment is a Kaggle-sourced CSV file that contains student information and academic metrics. Key details of the dataset are:

- Total samples: 5000 records(Students)
- Attributes: 23 Columns, which includes demographic details (e.g., Student_ID, First_Name, Last_Name, Age), academic performance indicators (e.g. Attendance (%), Midterm_Score, Final_Score, Projects_Score, Total_Score) and additional variables (e.g., Study_Hours_per_Week, Stress_Level, Sleep_Hours_per_Night).

A brief glance at the dataset (using the top 5 and bottom 5 rows) confirmed its structure and content. This initial overview helped in understanding the variety of data types present, ranging from numeric values to categorical descriptors.

```
Top 5 rows:
  Student_ID First_Name Last_Name                    Email  Gender  Age  \
0      S1000       Omar  Williams  student0@university.com  Female   22
1      S1001      Maria     Brown  student1@university.com    Male   18
2      S1002      Ahmed     Jones  student2@university.com    Male   24
3      S1003       Omar  Williams  student3@university.com  Female   24
4      S1004       John     Smith  student4@university.com  Female   23

    Department  Attendance (%)  Midterm_Score  Final_Score  ...  \
0  Engineering           52.29          55.03        57.82  ...
1  Engineering           97.27          97.23        45.80  ...
2     Business           57.19          67.05        93.68  ...
3  Mathematics           95.15          47.79        80.63  ...
4           CS           54.18          46.59        78.89  ...

   Projects_Score  Total_Score Grade  Study_Hours_per_Week  \
0           85.90        56.09     F                   6.2
1           55.65        50.64     A                  19.0
2           73.79        70.30     D                  20.7
3           92.12        61.63     A                  24.8
4           68.42        66.13     F                  15.4

   Extracurricular_Activities Internet_Access_at_Home  Parent_Education_Level  \
0                          No                     Yes             High School
1                          No                     Yes                     NaN
2                          No                     Yes                Master's
3                         Yes                     Yes             High School
4                         Yes                     Yes             High School

  Family_Income_Level Stress_Level (1-10) Sleep_Hours_per_Night
0              Medium                   5                   4.7
1              Medium                   4                   9.0
2                 Low                   6                   6.2
3                High                   3                   6.7
4                High                   2                   7.1

[5 rows x 23 columns]
```


```
Top 5 rows:
  Student_ID First_Name Last_Name                    Email  Gender  Age  \
0      S1000       Omar  Williams  student0@university.com  Female   22
1      S1001      Maria     Brown  student1@university.com    Male   18
2      S1002      Ahmed     Jones  student2@university.com    Male   24
3      S1003       Omar  Williams  student3@university.com  Female   24
4      S1004       John     Smith  student4@university.com  Female   23

    Department  Attendance (%)  Midterm_Score  Final_Score  ...  \
0  Engineering           52.29          55.03        57.82  ...
1  Engineering           97.27          97.23        45.80  ...
2     Business           57.19          67.05        93.68  ...
3  Mathematics           95.15          47.79        80.63  ...
4           CS           54.18          46.59        78.89  ...

   Projects_Score  Total_Score Grade  Study_Hours_per_Week  \
0           85.90        56.09     F                   6.2
1           55.65        50.64     A                  19.0
2           73.79        70.30     D                  20.7
3           92.12        61.63     A                  24.8
4           68.42        66.13     F                  15.4

   Extracurricular_Activities Internet_Access_at_Home  Parent_Education_Level  \
0                          No                     Yes             High School
1                          No                     Yes                     NaN
2                          No                     Yes                Master's
3                         Yes                     Yes             High School
4                         Yes                     Yes             High School

  Family_Income_Level Stress_Level (1-10) Sleep_Hours_per_Night
0              Medium                   5                   4.7
1              Medium                   4                   9.0
2                 Low                   6                   6.2
3                High                   3                   6.7
4                High                   2                   7.1

[5 rows x 23 columns]
```


Figure 1: Top and Bottom 5 Records Overview

Data Processing Outcomes

Data Preparation and Library Utilization

The data was processed using the Pandas library in python which is well-suited for data manipulation and analysis. The workflow involved:

- Mounting Google Drive: this allowed direct access to the dataset stored in a Google Drive folder.
- Reading the Dataset: The pd.read_csv fuction was used to load the CSV file into a DataFrame.
- Initial Data inspection: Commands like df.head() and df.tail() provided a quick visual check of the dataset's beginning and ending rows.

Statistical Summary

To better understand the distribution of numeric variables, a statistical summary was generated based on the filtered data frame which had limited columns and 464 records of data. This summary included:

- Count, Mean, Standard Deviation: Offering a general idea of central tendency and dispersion.
- Minimum and Maximum Values: Establishing the range of values for each attribute.
- Quartiles: Indicating the spread of data across the distribution.

For instance, within a subset of the dataset (filtered for Engineering students with attendance below 85% and study hours of 18 or less), the following insights were noted:

- Age: The mean age was approximately 21 years.
- Attendence (%): The average attendance was around 67.66% with a minimum of 50% and a maximum just under 85%.
- Final Score: Average roughly 70.67 with scores ranging from 40.62 to almost 100.
- Stress Level(1-10): This variable had a mean of 5.43 and a range 1 to 10.

	Age	Attendance (%)	Study_Hours_per_Week	Final_Score
count	464.000000	464.000000	464.000000	464.000000
mean	20.956897	67.660280	11.467241	70.669978
std	1.993043	10.097719	3.721307	17.089147
min	18.000000	50.010000	5.000000	40.620000
25%	19.000000	58.990000	8.375000	56.605000
50%	21.000000	68.095000	11.500000	71.540000
75%	23.000000	76.102500	14.400000	85.247500
max	24.000000	84.990000	18.000000	99.950000

	Stress_Level (1-10)
count	464.000000
mean	5.426724
std	2.791343
min	1.000000
25%	3.000000
50%	5.000000
75%	8.000000
max	10.000000

Figure 2: Statistical Summary Output

Correlation Analysis

A correlation table was generated to assess the relationships between key numerical variables in the filtered subset. The findings included:

- Attendance(%) and Final Score: A weak positive correlation (approximately 0.065) suggesting that within this subset higher attendance might be marginally related to higher final score.
- Study Hours per week and Final Score: An almost negligible negative correlation (around -0.030) indicating no strong linear relationship between study hours and final performance.
- Stress Level and other variables: Stress Level showed almost no significant correlation with attendance, study hours or final score.

These correlation values imply that within the selected group of Engineering students, the expected linear relationships between these academic metrics are either weak or non-existent.

The general lack of strong correlations across all pairs of numerical variables suggests several possibilities:

- Non-Linear Relationships: It is possible that relationships between variables exist but are non-linear. In such cases there are many other analytical methods (e.g. polynomial regression etc.) that might be more appropriate for uncovering these relationships.
- Influence of External Factors: The weak correlation might indicate that there are additional variables or contextual factors (such as teaching methods, curriculum

differences or extracurricular activities) that are not captured in the dataset but significantly influence student performance.

- Data Quality and Variability: The dataset itself might considerable variability or noise, particularly given some identified data quality issues (missing values in key columns). This noise could dilute any potential linear relationships.
- Assessment Metrics: The evaluation metrics (e.g. midterms, final, assignments, etc.) may measure different aspects of student performance. Their lack of strong correlation could mean that each metric is capturing distinct competencies or skills which are not directly comparable on a linear scale.

This analysis provides a clear indication that while the dataset is comprehensive, the relationships among the various academic and lifestyle measures are complex and likely require more advanced statistical techniques to be fully understood.

```
Correlation table:
                     Attendance (%)  Study_Hours_per_Week  Final_Score  \
Attendance (%)             1.000000             -0.057134     0.064881
Study_Hours_per_Week      -0.057134              1.000000    -0.030125
Final_Score                0.064881             -0.030125     1.000000
Stress_Level (1-10)        0.007832             -0.016741    -0.033999

                     Stress_Level (1-10)
Attendance (%)                  0.007832
Study_Hours_per_Week           -0.016741
Final_Score                    -0.033999
Stress_Level (1-10)             1.000000
```

Figure 3: Sub dataframe Correlation Table Output

Data Quality

Missing Data and Integrity Challenges

An important aspect of the analysis was assessing the quality of the dataset. The following issues were identified:

- Attendance(%) and Assignments: These columns showed a high number of missing values (516 and 517 missing entries respectively), indicating that more than 105 of the dataset might have incomplete data in these fields.
- Parent_Education_Level: This column had 1794 missing values which is significant and may affect any analysis that considers parental background.

The missing values are also not uniformly distributed across all columns. While some columns (e.g. Student_ID, First_Name) are fully complete, others have a significant amount of missing information. This inconsistency can lead to difficulties in integrating and comparing different aspects of student performance.

The missing data could be due to incomplete survey responses or errors during data collections. This challenge necessities careful handling such as considering data imputation techniques or omitting missing values in certain analyses to ensure that subsequent results and interpretation remain valid.

```
Columns with Data missing:
Student_ID                       0
First_Name                       0
Last_Name                        0
Email                            0
Gender                           0
Age                              0
Department                       0
Attendance (%)                 516
Midterm_Score                    0
Final_Score                      0
Assignments_Avg                517
Quizzes_Avg                      0
Participation_Score              0
Projects_Score                   0
Total_Score                      0
Grade                            0
Study_Hours_per_Week             0
Extracurricular_Activities       0
Internet_Access_at_Home          0
Parent_Education_Level        1794
Family_Income_Level              0
Stress_Level (1-10)              0
Sleep_Hours_per_Night            0
dtype: int64
```

Figure 4: Data Quality (Missing Data) Output

A check if any records had duplicate student ID and email in the dataset was done as well. Ideally this was done specifically in student ID because the expectation is that a student ID and email should be unique for each student. However, there were no duplicates there, which is the best case scenario.

```
Number of duplicates in 'Student_ID': 0

Number of duplicates in 'E-mail': 0
```

Figure 5: Duplicate data Output

A check was also done to see if there were any invalid datatypes for any of the columns.

```
Check for Columns with incorrect data types:
Student_ID                      object
First_Name                      object
Last_Name                       object
Email                           object
Gender                          object
Age                              int64
Department                      object
Attendance (%)                 float64
Midterm_Score                  float64
Final_Score                    float64
Assignments_Avg                float64
Quizzes_Avg                    float64
Participation_Score            float64
Projects_Score                 float64
Total_Score                    float64
Grade                           object
Study_Hours_per_Week           float64
Extracurricular_Activities      object
Internet_Access_at_Home         object
Parent_Education_Level          object
Family_Income_Level             object
Stress_Level (1-10)              int64
Sleep_Hours_per_Night          float64
dtype: object
```

Figure 6: Incorrect Datatype check Output

The only initial concern here that popped us was student ID having the datatype object but on a closer inspection an example of a student ID is “S1001” which is an alphanumeric value which is why it is being stored as a string. There are no other specific concerns raised from this check.

Conclusion

In summary, the student grading dataset provided a rich source of information with 5000 samples and 23 attributes enabling a comprehensive analysis of student performance metrics. The data processing workflow utilizing Pandas to load, inspect and subset the data, followed by generating statistical summaries and correlation matrices. Despite the robust analytical approach, the dataset revealed significant data quality issues with missing values in key columns, highlighting the need for further data cleaning before more advanced analytics could be performed.

Insights drawn from the statistical summary and correlation analysis offer a foundational understanding of student performance, though the weak correlation indicates that other, perhaps non-linear factors may influence academic outcomes. Overall, this analysis not only demonstrates proficiency in python-based data processing but also underscores the importance of addressing data quality challenges to improve the integrity and reliability of data driven conclusion.