
Exploratory Data Analysis on Student Performance DataSet

Analysis of Factors Affecting Exam Scores

Presented by : Jobina

K v

Introduction

- **Exploratory Data Analysis is the process of analyzing datasets to summarize their key characteristics using statistical methods and visualizations.**
- **EDA is the process of checking and exploring a dataset before analysis.**

→ Main purpose of EDA is :

- **Understand dataset structure like columns, types and missing values.**
- **Identify patterns and relationships.**
- **To understand the data clearly and find useful information.**

Problem Statements

Students exam performance is affected by many factors like

- **Hours studied**
- **Sleep hours**
- **Attendance**
- **Previous scores**

So the problem is we need to find which factor is most affect the student exam scores

Proposal solution

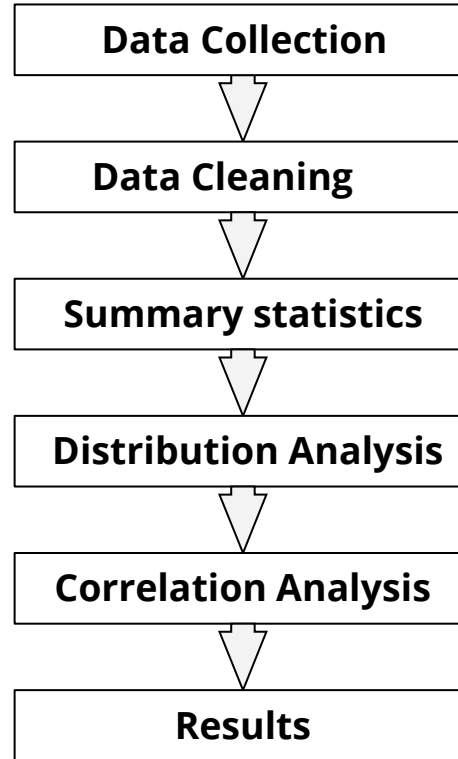
- **The problem is Students exam performance is affected by many factors. So We need to find which factors is most affect Exam Score.**
- **To solve this problem we perform Exploratory Data Analysis (EDA) on the students performance dataset to identify relationship between study routine,attendance, sleep hours, previous exam scores and current exam score.**
- **This analysis will show which factors most strongly affect exam performance.**

Dataset Overview & Structure

- The Students performance dataset includes records of 30 students with details about study routine, attendance, sleep hours, previous exam scores and current exam score.
- It is analyzed to identify the key factors influencing exam performance.

```
RangeIndex: 30 entries, 0 to 29
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  -
 0   student_id          30 non-null     object
 1   hours_studied        30 non-null     float64
 2   sleep_hours          30 non-null     float64
 3   attendance_percent   30 non-null     float64
 4   previous_scores      30 non-null     int64
 5   exam_score           30 non-null     float64
dtypes: float64(4), int64(1), object(1)
memory usage: 1.5+ KB
```

Workflow



Tools Used

- **Software:** Python, Google Colab
- **Libraries:** pandas, matplotlib

Implementation

- **Step 1: Selected the first 30 students from the dataset for analysis.**
- **Step 2: Checked data types, missing values, and duplicates to ensure data quality.**
- **Step 3: Generated summary statistics to understand the mean, median and spread of each feature.**
- **Step 4: Plotted bar charts to study the distribution of:**
 - **Study Routine**
 - **Attendance**
 - **Sleep Hours**
 - **Current Exam Score**
 - **Previous Exam Scores**
- **Step 5: Performed correlation analysis to measure the strength of relationships between these factors and the exam score.**

Dataset info

```
RangeIndex: 30 entries, 0 to 29
Data columns (total 6 columns):
#   Column              Non-Null Count  Dtype
---  -
0   student_id          30 non-null    object
1   hours_studied        30 non-null    float64
2   sleep_hours          30 non-null    float64
3   attendance_percent   30 non-null    float64
4   previous_scores      30 non-null    int64
5   exam_score           30 non-null    float64
dtypes: float64(4), int64(1), object(1)
memory usage: 1.5+ KB
```

Data Cleaning

```
0
student_id    0
hours_studied 0
sleep_hours   0
attendance_percent 0
previous_scores 0
exam_score    0
dtype: int64
```

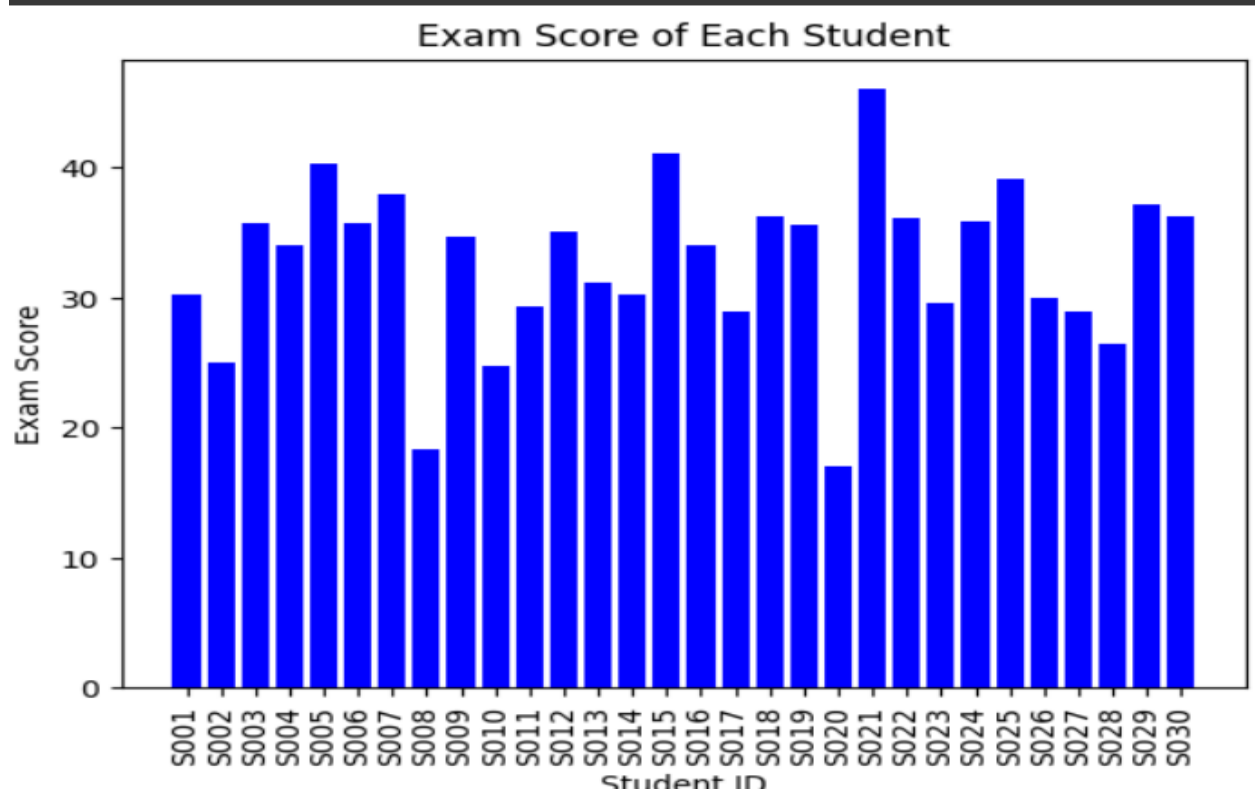
```
df1.duplicated().sum()

np.int64(0)
```

Statistics Summary

	hours_studied	sleep_hours	attendance_percent	previous_scores	exam_score
count	30.000000	30.000000	30.000000	30.000000	30.000000
mean	5.656667	6.826667	72.920000	68.433333	32.706667
std	3.318878	1.544051	13.753129	13.574020	6.340833
min	1.100000	4.300000	50.300000	41.000000	17.100000
25%	2.825000	5.600000	60.825000	58.250000	29.375000
50%	5.150000	6.750000	73.750000	70.000000	34.400000
75%	8.325000	8.275000	84.500000	80.000000	36.175000
max	11.500000	9.000000	95.100000	90.000000	46.000000

Exam Score Distribution

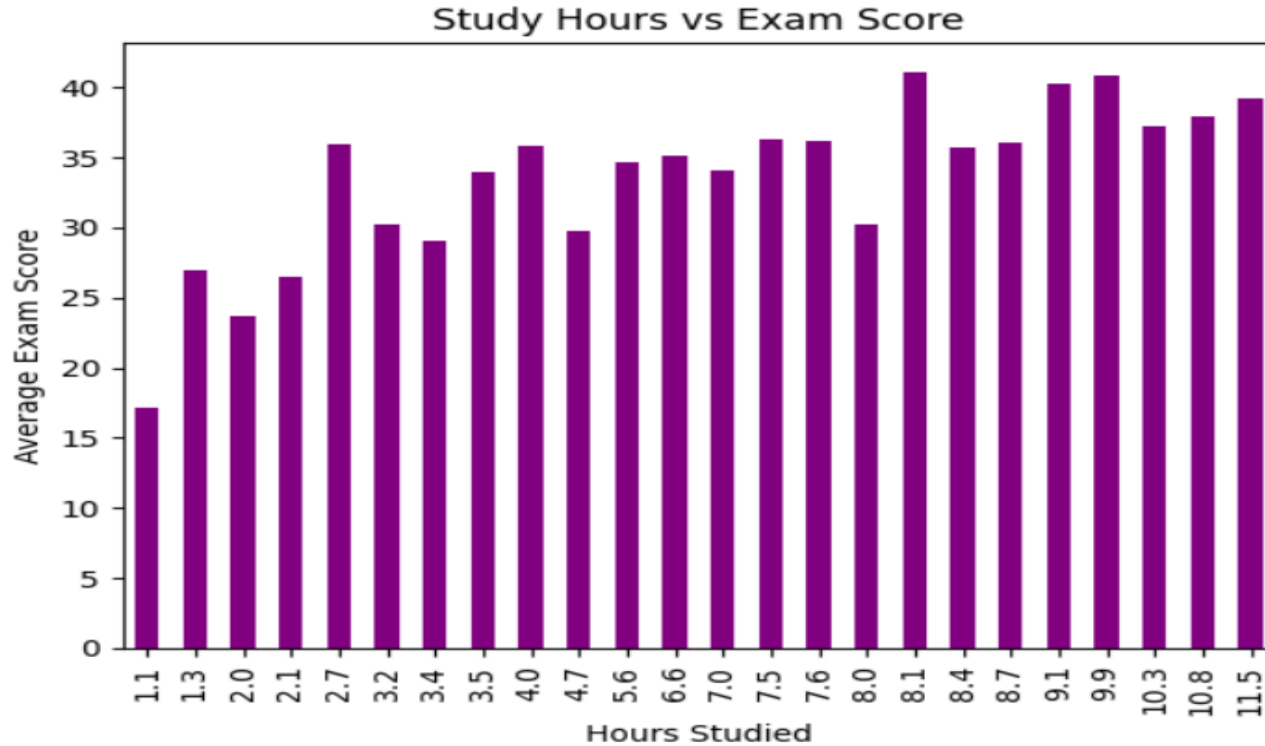


- The chart shows the exam scores of each student.
- To see how exam marks vary from one student to another.

Key Insight :

- The scores go from 17 to 46.
- Some students scored below 25, while some scored above 40.
- Overall the scores show a wide difference in Performance.
- This chart helps us to compare the exam score of each student and find who scored low or high mark.

Hours Studied vs Exam Score

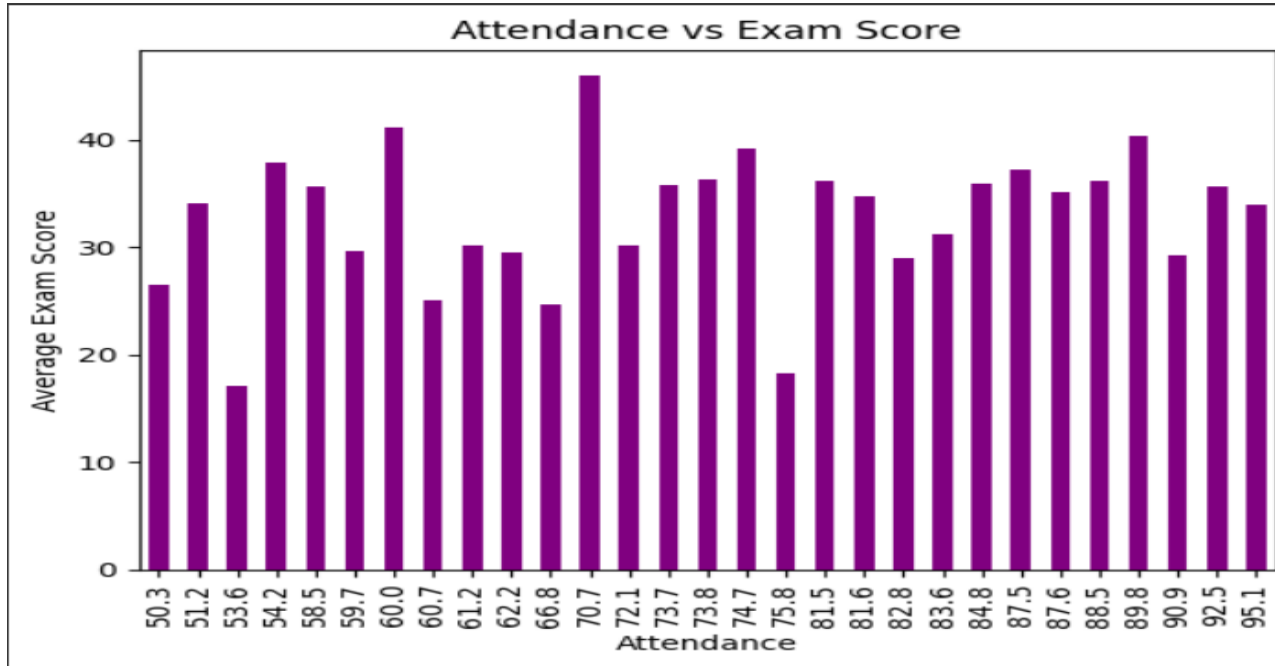


- **This chart compares the average exam score of students with the number of hours they studied each day.**
- **It helps us see how study time affects exam performance.**
- **To check if more study time leads to higher marks.**

Key Insight :

- **Students who studied 8 to 11 hours a day**
- **This means that spending more time studying usually leads to better exam scores.**

Attendance vs Exam Score

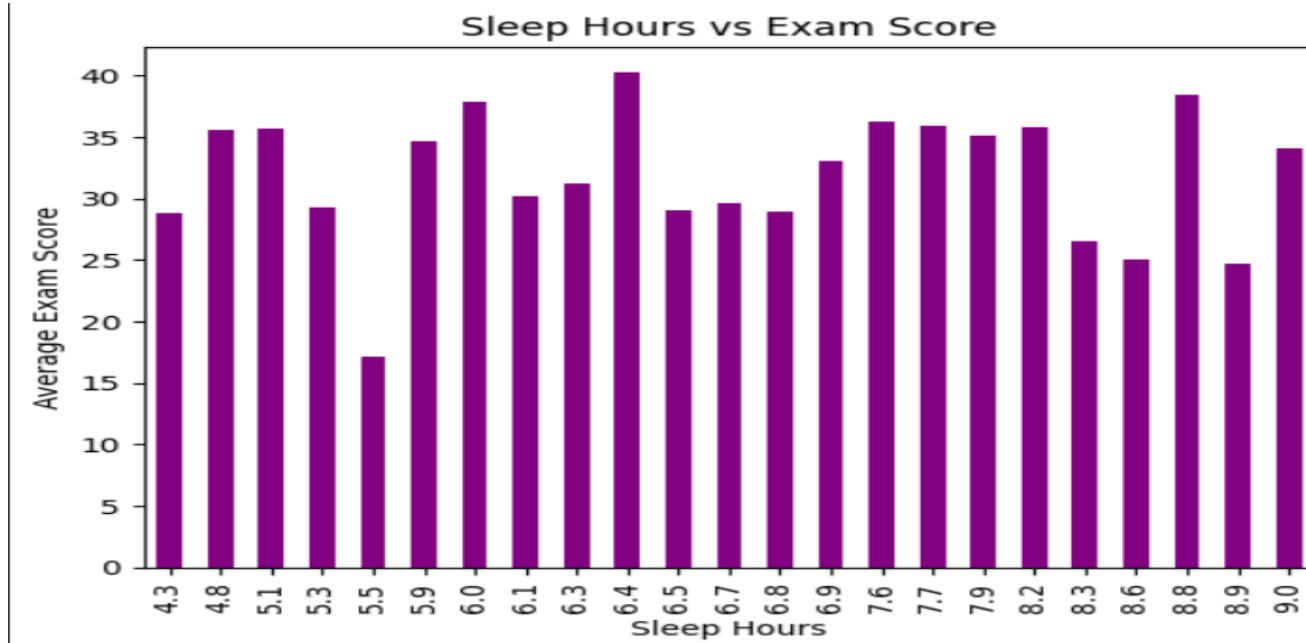


- **This chart shows average exam scores of students based on their attendance percentage.**
- **It helps us understand how attendance affects exam performance.**
- **To understand if regular class attendance improves marks.**

Key Insight :

- **Students with high attendance (Greater than 85%) often scored above 35, but some students with low attendance (around 60%) also scored above 40.**
- **Attendance helps a little, but it is not the strongest factor.**

Sleep Hours vs Exam Score

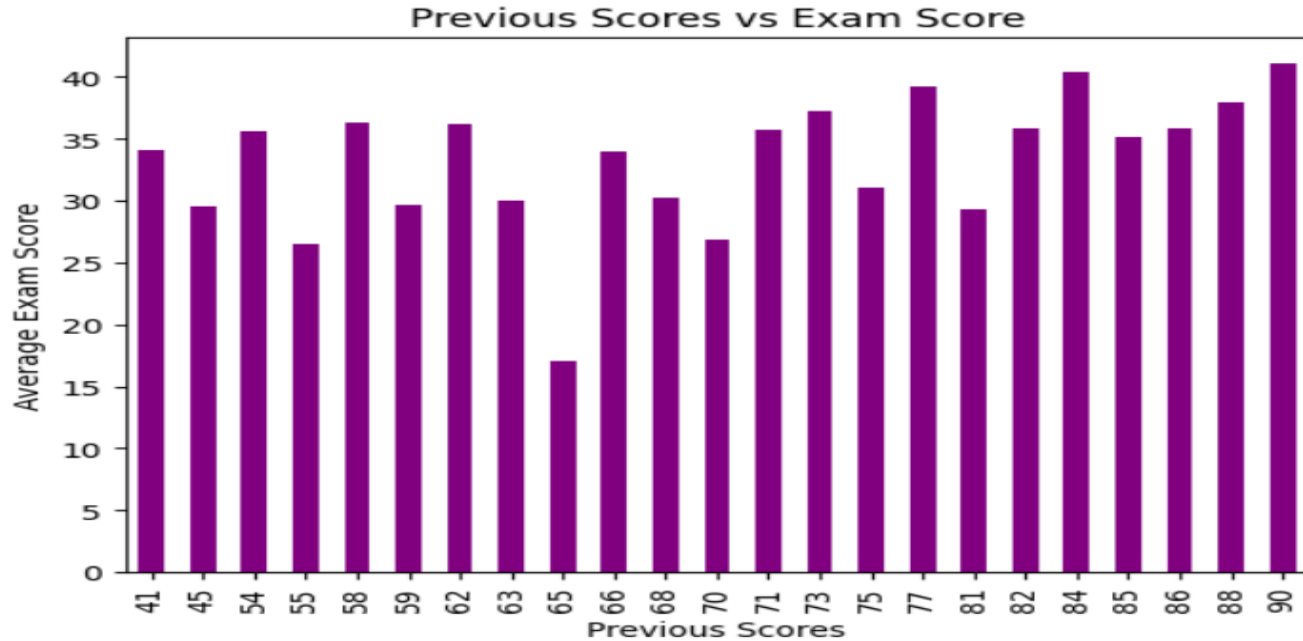


- **Bar chart showing Exam Score vs Sleep Hours.**
- **To see if the amount of daily sleep affects marks.**

Key Insight :

- **No clear pattern.**
- **Good scores appear in both low (4–6 hrs) and high (8–9 hrs) sleep groups.**
- **Sleep hours do not strongly affect exam performance in this dataset.**

Previous Scores vs Exam Score

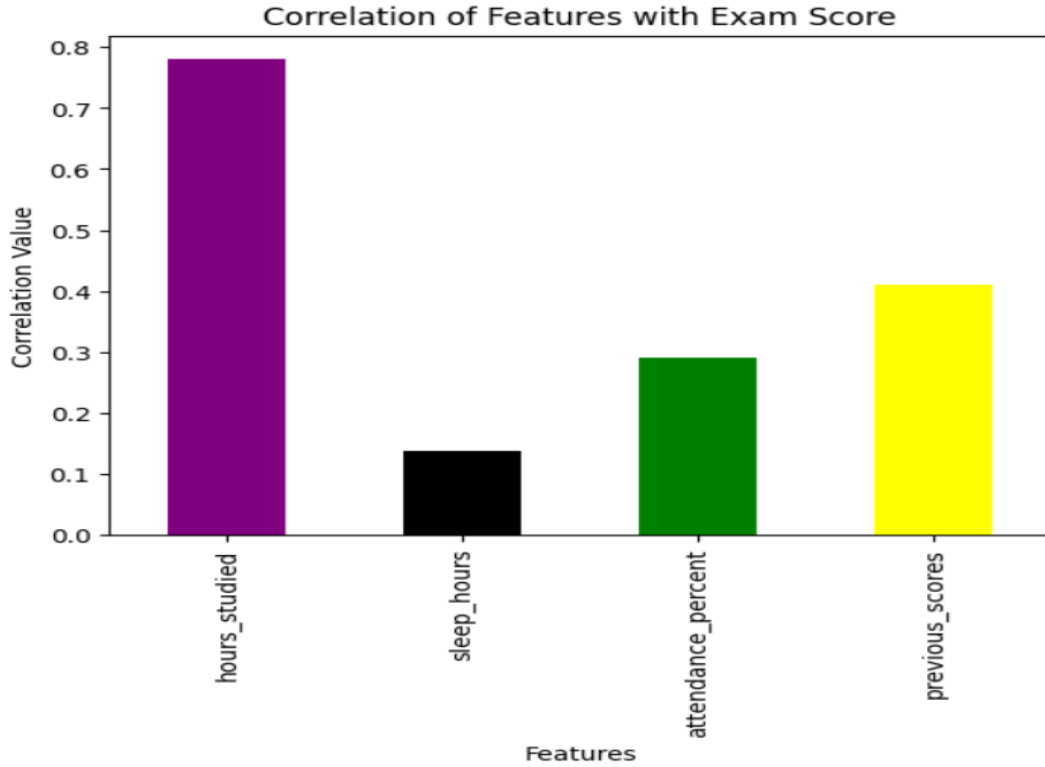


- **This Bar chart showing the Previous marks vs Current marks.**
- **To see if students who did well in past exams also get good marks now.**

Key Insight :

- **Students with past scores above 80 usually scored above 35 again.**
- **Better past performance often leads to better current scores, but not every time.**

Correlation of Features with Exam Score



- **To measure which factor is most related to exam marks.**

Key Insight :

- **Hours Studied has the strongest positive correlation with exam score.**
- **Previous Scores also show a positive link.**
- **Attendance and Sleep Hours have weaker correlations.**

Result

- **Study hours are the most important factor for higher marks.**
- **Previous performance also matters.**
- **Attendance and sleep have smaller effects.**

Conclusion

- **This Exploratory Data Analysis helped us understand the main factors affecting student exam performance.**
- **In this dataset we found that**
 - **Study hours and previous exam scores are the key factors linked to higher marks.**
 - **Attendance and sleep show some influence but are less important.**
- **These findings can help students, teachers, and parents focus on good study habits and regular preparation to improve exam results.**

Reference

- **Student Performance Dataset – Kaggle**
- **Python & Library References – Pandas, Matplotlib**

THANK YOU