Student Success : Factors & Insights

ISE-201 Math Foundations for Decision and Data Sciences

Name: Bishakha

SJSU ID: 019178354

Understanding the Problem Statement

Objective

The performance of students may be affected by several factors such as parental involvement, access to resources, extracurricular activities, sleep hours, motivation, and socio-economic background. There can be a wide variety of factors , unique to each student. The goal of the EDA Project was to understand which factors contribute most to the academic success of students at mass level.

Dataset: *https://www.kaggle.com/datasets/anassarfraz13/student-success-factors-and-insights*

EDA Plan Review

- Understanding the dataset
- Exploring and Inspecting dataset
- Handling the Missing values / Duplicates
- Exploring Patterns in data (Box Plots, Histograms, Count plots , scatter plot)
- Handling the outliers
- Transform the data (Scale, Encode , Derive ratios/Columns)
- Finding the Correlations
- Insights (Finding the answers and conclusion)

Understand , Inspect and clean the dataset

- Shape : 6607 X 20
- Categorical Features: "Parental_Involvement","Access_to_Resources","Extracurricular_Activities", "Motivation_Level","Internet_Access","Family_Income","Teacher_Quality","School_Type","Peer_Influence" ,"Learning_Disabilities","Parental_Education_Level","Distance_from_Home","Gender"
- Numerical Features: "Hours_Studied","Sleep_Hours","Previous_Scores", "Physical_Activity", "Tutoring_Sessions", "Exam_Score", "Attendance"
- dtypes: int64(7), object(13)
- Check for **duplicate** rows **:** no duplicate rows found
- Check for columns with **Null Values**
  - Teacher quality : 78 rows
  - Parental Education level : 90 rows
  - Distance from home : 67 rows
  - Use mode value to fill na values

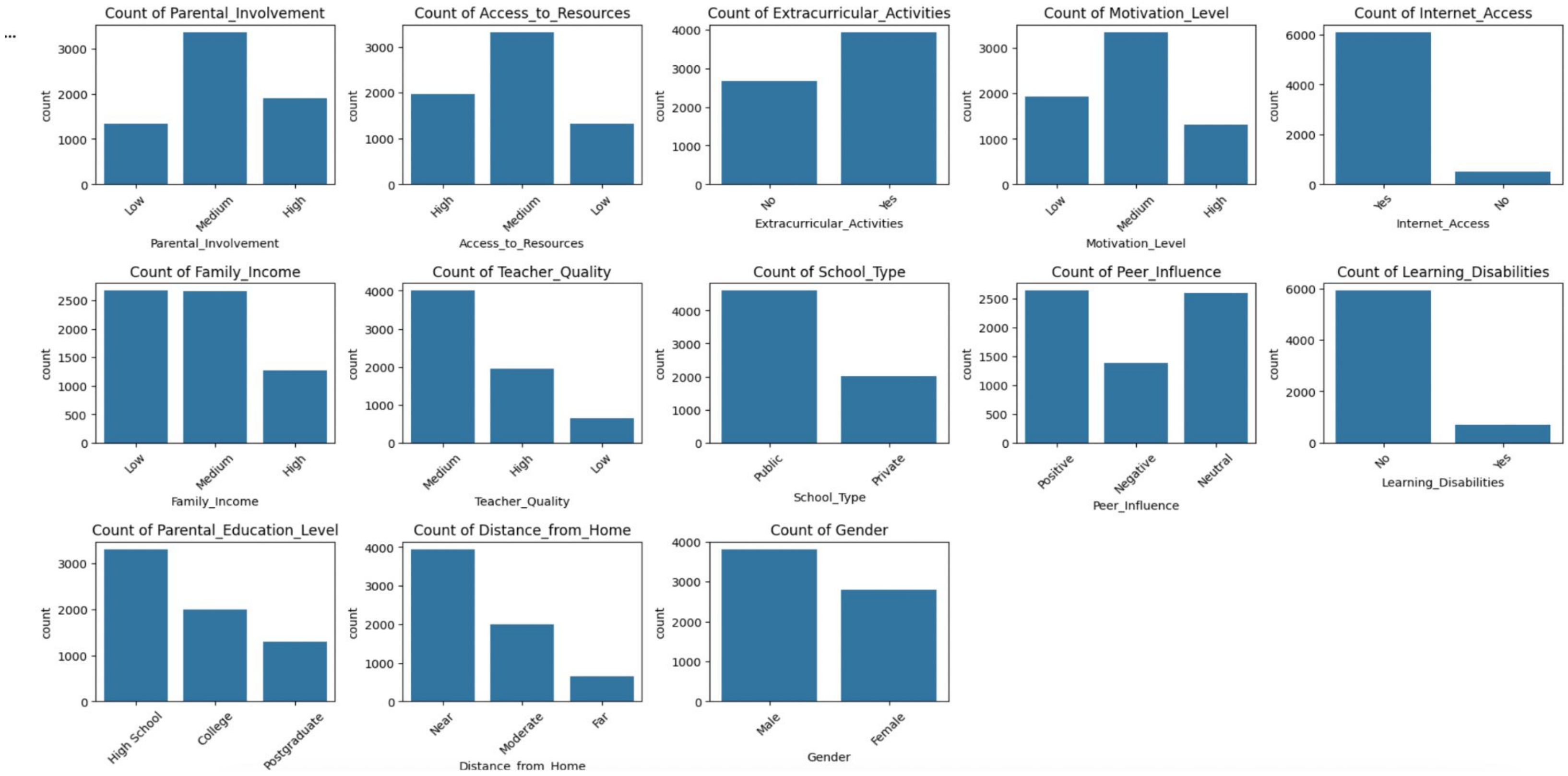
Exploring Patterns in dataset

- **Look at the distribution of data using different plots**
- **Used Box Plots , histograms for numerical features**
- **Count plot diagram for categorical features**

Features	Mean	Max	Std Dev.	Distribution
Hours_Studied	19.975329	44.000000	5.990594	Symmetric normal distribution
Attendance	79.977448	100.000000	11.547475	Symmetric normal distribution
Sleep_Hours	7.02906	10.00000	1.46812	Symmetric normal distribution
Previous_Scores	75.070531	100.000000	14.399784	Symmetric normal distribution
Tutoring_Session	1.493719	8.000000	1.230570	normal distribution with little right skewed
Physical_Activity	2.967610	6.000000	1.031231	normal distribution with little right skewed
Exam_Score	67.235659	101.000000	3.890456	normal distribution with little left skewed

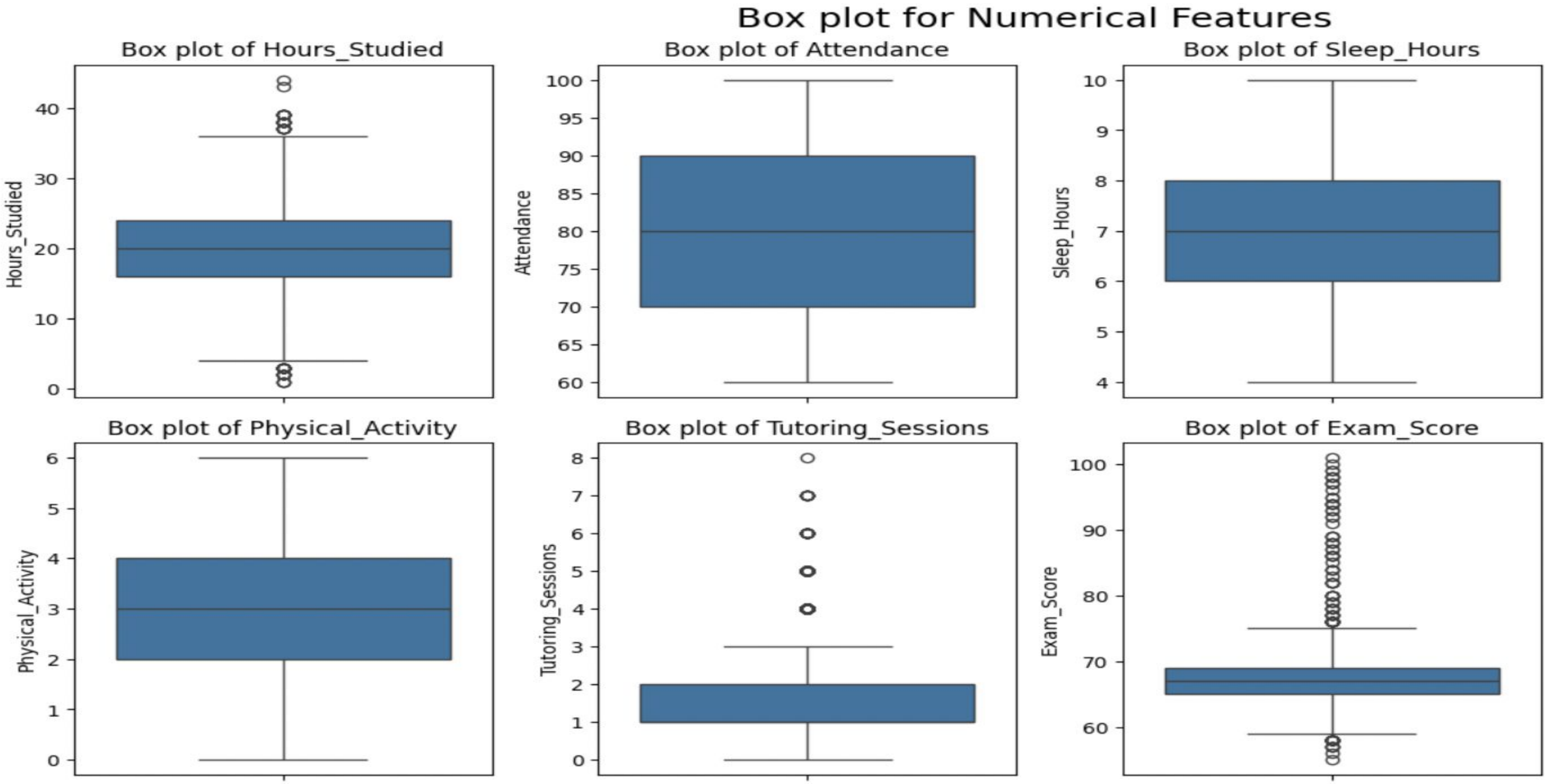
CountPlot for Categorical Features

- Normally distributed
- Few features Skewed to left and right



Box Plot for Numerical Features

- **Finding the outliers**

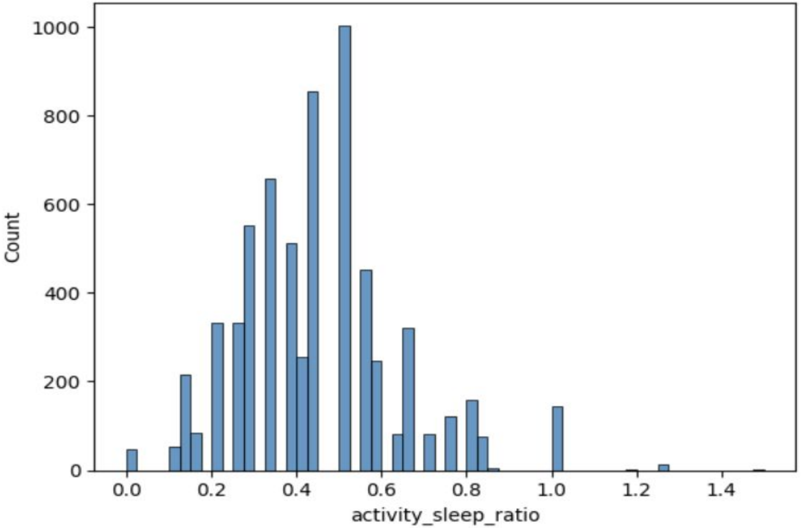


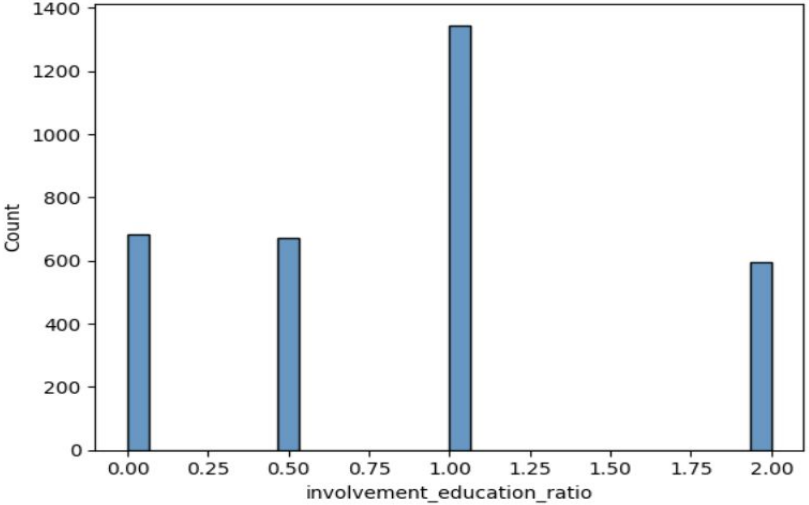
Handling the outliers

- Outliers identified using box plots
    - Exam Score : 1 student with 101 score , set to 100
    - Tutoring Session : 1266 , no need to correct
    - Hours Studied : May vary too much , so no need to handle

Transform the data

- **Scaling :** No features holds large values.
- **Encoding :** Ordinal encoding for most categorical features.
- **Ratios :** Activity_to_Sleep_ratio , Parental_Education_Level_to_Parental_Involvment_ratio



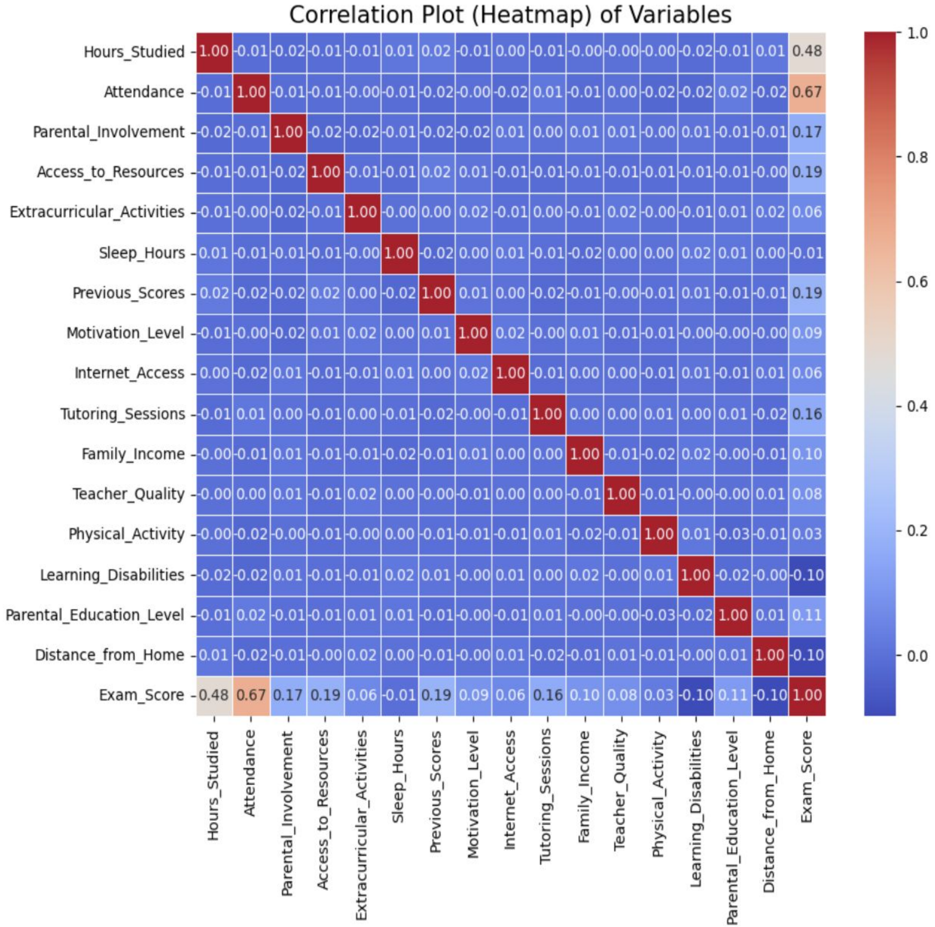


Finding the Correlations

- Used **corr(method="spearman")**
- Used **sns.heatmap** to plot the correlation
- **Attendance , hours_studied** have **highest** correlation

- Parental_Involvment , access_to_resources, Previous_Scores, Tutoring_Session, Parental_Education_Level also have significant correlation with Exam_Score

- **Distance_From_Home** and **Learning_Disabilit**y have **-ve** correlation with Exam_Score



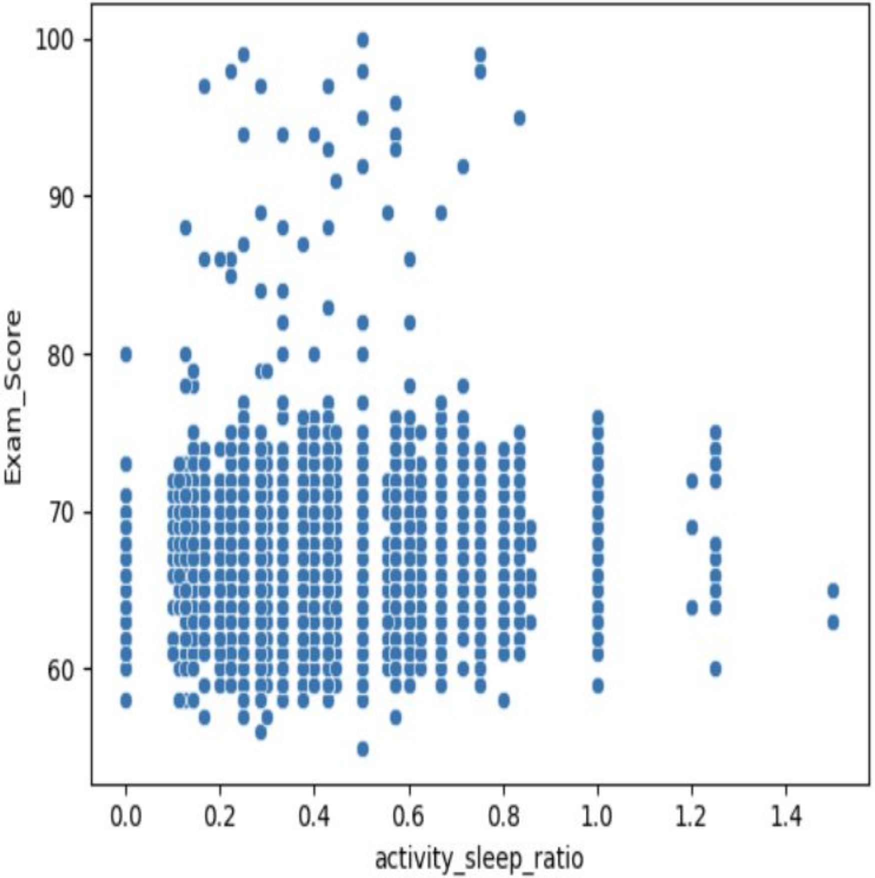
Correlation Plot (Heatmap) of Variables

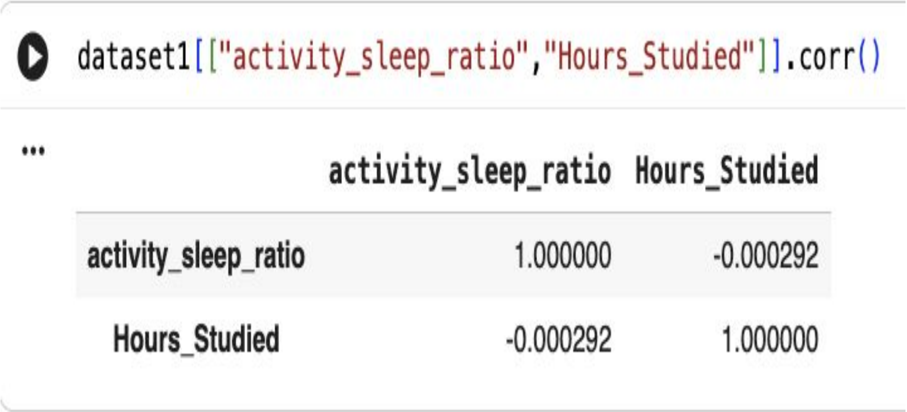
	Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Scores	Motivation_Level	Internet_Access	Tutoring_Sessions	Family_Income	Teacher_Quality	Physical_Activity	Learning_Disabilities	Parental_Education_Level	Distance_from_Home	Exam_Score
Hours_Studied	1.00	-0.01	-0.02	-0.01	-0.01	0.01	0.02	-0.01	0.00	-0.01	-0.00	-0.00	-0.00	-0.02	-0.01	0.01	0.48
Attendance	-0.01	1.00	-0.01	-0.01	-0.00	-0.01	-0.02	-0.00	-0.02	0.01	-0.01	0.00	-0.02	-0.02	0.02	-0.02	0.67
Parental_Involvement	-0.02	-0.01	1.00	-0.02	-0.02	-0.01	-0.02	-0.02	0.01	0.00	0.01	0.01	-0.00	0.01	-0.01	-0.01	0.17
Access_to_Resources	-0.01	-0.01	-0.02	1.00	-0.01	-0.01	0.02	0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.00	0.19
Extracurricular_Activities	-0.01	-0.00	-0.02	-0.01	1.00	-0.00	0.00	0.02	-0.01	0.00	-0.01	0.02	-0.00	-0.01	0.01	0.02	0.06
Sleep_Hours	0.01	-0.01	-0.01	-0.01	-0.00	1.00	-0.02	0.00	0.01	-0.01	-0.02	0.00	0.00	0.02	0.01	0.00	-0.01
Previous_Scores	0.02	-0.02	-0.02	0.02	0.00	-0.02	1.00	0.01	0.00	-0.02	-0.01	-0.00	-0.01	0.01	-0.01	-0.01	0.19
Motivation_Level	-0.01	-0.00	-0.02	0.01	0.02	0.00	0.01	1.00	0.02	-0.00	0.01	-0.01	-0.01	-0.00	-0.00	-0.00	0.09
Internet_Access	0.00	-0.02	0.01	-0.01	-0.01	0.01	0.00	0.02	1.00	-0.01	0.00	0.00	-0.01	0.01	-0.01	0.01	0.06
Tutoring_Sessions	-0.01	0.01	0.00	-0.01	0.00	-0.01	-0.02	-0.00	-0.01	1.00	0.00	0.00	0.01	0.00	0.01	-0.02	0.16
Family_Income	-0.00	-0.01	0.01	-0.01	-0.01	-0.02	-0.01	0.01	0.00	0.00	1.00	-0.01	-0.02	0.02	-0.00	-0.01	0.10
Teacher_Quality	-0.00	0.00	0.01	-0.01	0.02	0.00	-0.00	-0.01	0.00	0.00	-0.01	1.00	-0.01	-0.00	-0.00	0.01	0.08
Physical_Activity	-0.00	-0.02	-0.00	-0.01	-0.00	0.00	-0.01	-0.01	-0.01	0.01	-0.02	-0.01	1.00	0.01	-0.03	-0.01	0.03
Learning_Disabilities	-0.02	-0.02	0.01	-0.01	-0.01	0.02	0.01	-0.00	0.01	0.00	0.02	-0.00	0.01	1.00	-0.02	-0.00	-0.10
Parental_Education_Level	-0.01	0.02	-0.01	-0.01	0.01	0.01	-0.01	-0.00	-0.01	0.01	-0.00	-0.00	-0.03	-0.02	1.00	0.01	0.11
Distance_from_Home	0.01	-0.02	-0.01	-0.00	0.02	0.00	-0.01	-0.00	0.01	-0.02	-0.01	0.01	-0.01	-0.00	0.01	1.00	-0.10
Exam_Score	0.48	0.67	0.17	0.19	0.06	-0.01	0.19	0.09	0.06	0.16	0.10	0.08	0.03	-0.10	0.11	-0.10	1.00

Insights

- The data for most of the Numerical features is *normally distributed* , with a few skewed towards right for eg : Tutoring_Session, Physical_Activity
- Few **categorical Features** are also ***Right-Skewed*** , For eg: **Parental education level ,distance from home , teacher quality**.
- Most of the students score between **65-75** in Exam
- Most of the students have **3-4 hours** of **physical activity**
- Most of the students **sleep** for **7-8 hours**, which is pretty good.
- Most students, close to **6000** have **internet access**.
- Almost **2/3rd** Students go to the **public school**.
- The activity to sleep ratio the ratio is almost distributed normally with little skewed to the Right.
- Looking at the **correlation** and **Scatter plot** between **activity_sleep_ratio** and **exam_score** , we can see that correlation is not significant. But in the distribution chart we see that most of the top scorers have either a balanced or sedentary lifestyle , which makes sense .

- **Scatter Plot b/w Exam_Score and activity_sleep_ratio**
- **Correlation b/w activity_sleep_Ratio and Hours_Studied**





```
dataset1[["activity_sleep_ratio","Hours_Studied"]].corr()
```

	activity_sleep_ratio	Hours_Studied
activity_sleep_ratio	1.000000	-0.000292
Hours_Studied	-0.000292	1.000000

- The correlation between **activity_sleep_ratio** and **Hours_Studied** is a **-ve** value. Though the value is not significant , it suggests that as the activity_sleep_ratio increases , which means physical activity increases, the hours studied decreases . This also holds true in a practical sense.

- ***Attendance* and *Hour studied*** have the highest correlation with *exam score*.

- ***Distance from home*** and ***learning disability*** has ***negative correlation*** with **exam score**, which indicates that as the learning disability increases , the exam score decreases , which holds true for real life. Also as the distance from home increases , the exam score decreases ,because students may be travelling for long time which affects their performance.

- ***Parental Involvement , Access to Resources , Previous Score, Tutoring Session , Family Income , Parental Educational Level*** also have significant correlation with exam score, which clearly indicates that all these factors contribute to the success of a child in academics .

- ***Parental Involvement* and *Parental Educational Level*** have **-ve correlation** ,which means that parents with higher degrees spend less time with kids.

Learning Beyond Original Questions

After going through the dataset and performing the EDA , it looked like the dataset is realistic , but not enough . More samples of the dataset is needed to get bigger and accurate numbers. Also there is no clarity about whether the sample represents the targeted population.

Also I found that it's hard to relate , process and combine categorical features. It's hard to interpret these columns and reach a conclusion. For eg in case of involvement_to_education ratio , both these columns were ordinal encoded , still the ratio didn’t make any sense.

Though the direction of the numbers supports my hypothesis about the factors affecting students performance and the summary of the EDA seems real and authentic , more dataset is needed to reach precise and clear conclusions.

Suggestions

- Quality of sample matters and it should represent the targeted population.
- Number of sample matters too, more the data , the more clear and precise the answers.
- Quality of data matters too. Dataset should have more numerical values and less ordinal or nominal values , as they can be hard to relate and conclude.
- Taking multiple samples from the same population can give even more accurate results and conclusions.
- Data should be cleaned properly , including decisions like whether to drop rows with na or fill the null values with mean or mode , depending upon the criticality of data .
- Removing the duplicity in data and also scaling the dataset , to remove inclination towards features with larger values.
- Proper encoding method should be used for encoding the ordinal and nominal features if they have to be used.

All these points are very important and should be taken care when selecting a dataset and performing EDA.