STUDENT PERFORMANCE FACTOR

DATA ANALYSIS

By,

PRANATHA M

Introduction

The education sector continually seeks to understand factors influencing student performance. This report analyzes various attributes, including study habits, attendance, parental involvement, and socio-economic factors, to identify key influences on student exam scores.

Objectives

- To explore the relationships between different factors and student performance.
- To identify significant predictors of exam scores.
- To provide actionable insights for educators and policymakers.

Dataset Overview

The dataset contains the following attributes:

- **Hours_studied**: Hours spent studying.
- **Attendance**: Student attendance rate.
- **Parental_involvement**: Level of parental engagement.
- **Access to Resources**: Availability of educational resources.
- **Extracurricular Activities**: Participation in non-academic activities.
- **Sleep_hours**: Average hours of sleep per night.
- **Previous_Scores**: Scores from previous assessments.
- **Motivation_level**: Student motivation on a scale.
- **Internet Access**: Availability of internet for study purposes.
- **Tutoring_sessions**: Frequency of tutoring sessions attended.
- **Family_income**: Family's economic status.
- **Teacher_quality**: Perceived quality of teaching.
- **School_type**: Type of school (public/private).
- **Peer_influence**: Impact of peers on student behavior.
- **Physical_activities**: Participation in physical education.
- **Learning_disability**: Presence of any learning disabilities.
- **Parental_education_level**: Highest education level achieved by parents.
- **Distance_from_home**: Distance of student’s home to school.
- **Gender**: Student's gender.
- **Exam_Score**: Final exam score.

Data Source

The Dataset is taken from Kaggle

GET AN IMPRESSION ABOUT THE DATASET

```
select * from studentperformancefactors;

select count(*) from studentperformancefactors;
```

INTERPRETATION

Number of Rows : 6607

Contains The Following Columns

Hours_Studied	Previous_Scores
Attendance	Motivation_Level
Parental_Involvement	Internet_Access
Access_to_Resources	Tutoring_Sessions
Extracurricular_Activities	Family_Income
Sleep_Hours	Teacher_Quality
School_Type	Peer_Influence
Physical_Activity	Learning_Disabilities
Parental_Education_Level	Distance_from_Home
Gender	Exam_Score

Missing observation:

Missing observations in a dataset can pose significant challenges for analysis. Here are some common approaches to handle missing data:

- **Identify the Type of Missing Data**:
  - **MCAR (Missing Completely at Random)**: The missingness is unrelated to any observed or unobserved data.
  - **MAR (Missing at Random)**: The missingness can be explained by other observed variables.
  - **MNAR (Missing Not at Random)**: The missingness is related to the value of the missing data itself.
- **Imputation Methods**:
  - **Mean/Median/Mode Imputation**: Replace missing values with the mean, median, or mode of the column.
  - **K-Nearest Neighbors (KNN)**: Use similar observations to predict and fill in missing values.
  - **Regression Imputation**: Use regression models to predict missing values based on other variables.
  - **Multiple Imputation**: Generate several imputed datasets and combine the results to account for uncertainty.
- **Deletion Methods**:
  - **Listwise Deletion**: Remove any observations with missing values. This is simple but can lead to significant data loss.
  - **Pairwise Deletion**: Use all available data for each analysis instead of dropping entire rows.

- **Use of Algorithms That Handle Missing Data**:
  - Some machine learning algorithms can handle missing values internally, such as decision trees.
- **Modeling Techniques**:
  - Incorporate missingness as a feature if it has predictive power.
  - Use techniques like Bayesian methods that can accommodate missing data.
- **Sensitivity Analysis**:
  - Assess how different methods of handling missing data affect the results to understand the impact of your chosen approach.

```
select count(*) from studentperformancefactors

WHERE Hours_Studied IS NULL

   OR Attendance IS NULL

   OR Parental_Involvement IS NULL

   OR Access_to_Resources IS NULL

   OR Extracurricular_Activities IS NULL

   OR Sleep_Hours IS NULL

   OR Previous_Scores IS NULL

   OR Motivation_Level IS NULL

   OR Internet_Access IS NULL

   OR Tutoring_Sessions IS NULL
```

```
OR Family_Income IS NULL
 OR Teacher_Quality IS NULL
 OR School_Type IS NULL
 OR Peer_influence IS NULL
 OR Physical_Activity IS NULL;
  OR learning_Disabilities IS NULL
 OR Parental_Education_Level IS NULL
 OR Distance_from_Home IS NULL
 OR Gender IS NULL
 OR Exam_Score IS NULL;
```

Interpretation

No missing values In Our DataSet

Data types

```
SELECT DATA_TYPE
FROM information_schema.columns
WHERE table_name = 'studentperformancefactors'
AND column_name = 'Hours_Studied';
SELECT DATA_TYPE
FROM information_schema.columns
WHERE table_name = 'studentperformancefactors'
AND column_name = 'Gender';
SELECT DATA_TYPE
FROM information_schema.columns
WHERE table_name = 'studentperformancefactors'
AND column_name = 'Attendance';
```

```
SELECT DATA_TYPE
FROM information_schema.columns
```

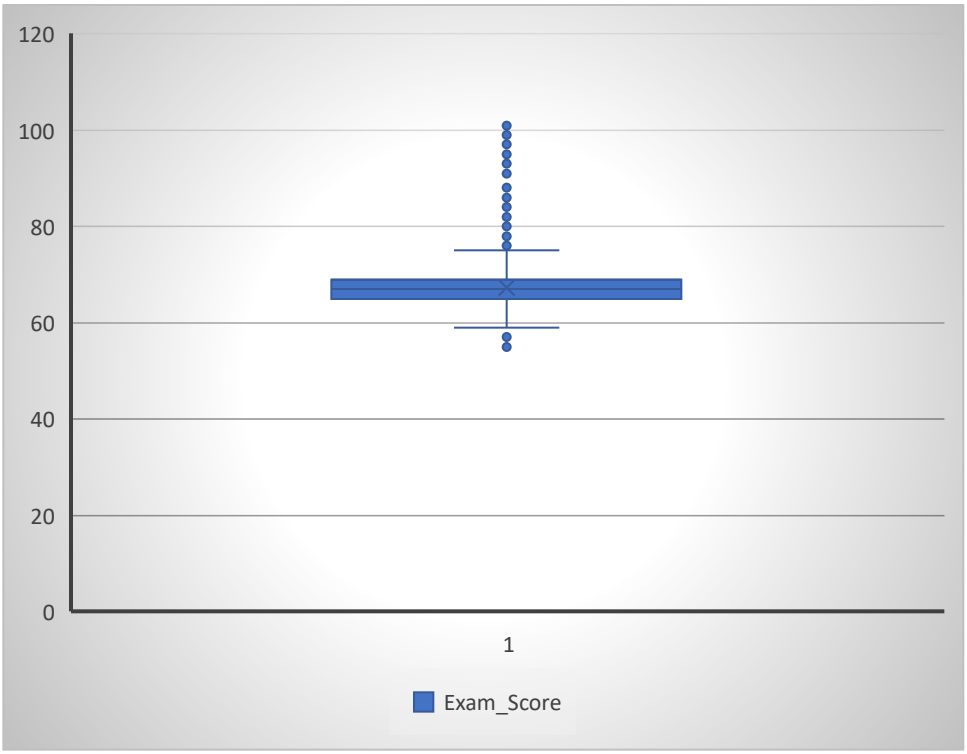
```
WHERE table_name = 'studentperformancefactors'
AND column_name = 'Parental_involvement';
SELECT DATA_TYPE
FROM information_schema.columns
WHERE table_name = 'studentperformancefactors'
AND column_name = 'Access_to_Resources';
SELECT DATA_TYPE
FROM information_schema.columns
WHERE table_name = 'studentperformancefactors'
AND column_name = 'Extracurricular_Activities';
SELECT DATA_TYPE
FROM information_schema.columns
WHERE table_name = 'studentperformancefactors'
AND column_name = 'Sleep_Hours';
SELECT DATA_TYPE
FROM information_schema.columns
WHERE table_name = 'studentperformancefactors'
AND column_name = 'Previous_Scores';
SELECT DATA_TYPE
FROM information_schema.columns
```

```
WHERE table_name = 'studentperformancefactors'

AND column_name = 'Motivation_Level';

SELECT DATA_TYPE

FROM information_schema.columns

WHERE table_name = 'studentperformancefactors'

AND column_name = 'Internet_Access';
```

Hours_Studied.	int(11)
Attendance.	int(11)
Parental_Involvement.	text
Access_to_Resources	text
Extracurricular_Activities	text
Sleep_Hours	int(11)
Previous_Scores	int(11)
Motivation_Level	text
Internet_Access	text
Tutoring_Sessions	int(11)
Family_Income	text
Teacher_Qualiy	text
School_Type	text
Peer_Influence	text
Physical_Activity	int(11)
Learning_Disabilities	text
Parental_Education_Level	text
Distance_from_Home	text
Gender	text
Exam_Score	int(11)

Creating a box plot to visualize the distribution of a dataset, including its outliers, can be an effective way to understand the data. Shows the median, quartiles, and potential outliers, giving a clear overview of the data's spread.

Boxplot



Interpretation

The presence of outliers suggests variability in student performance. This could be due to factors such as differences in study habits, test anxiety, or external circumstances affecting exam performance.

Univariate Analysis

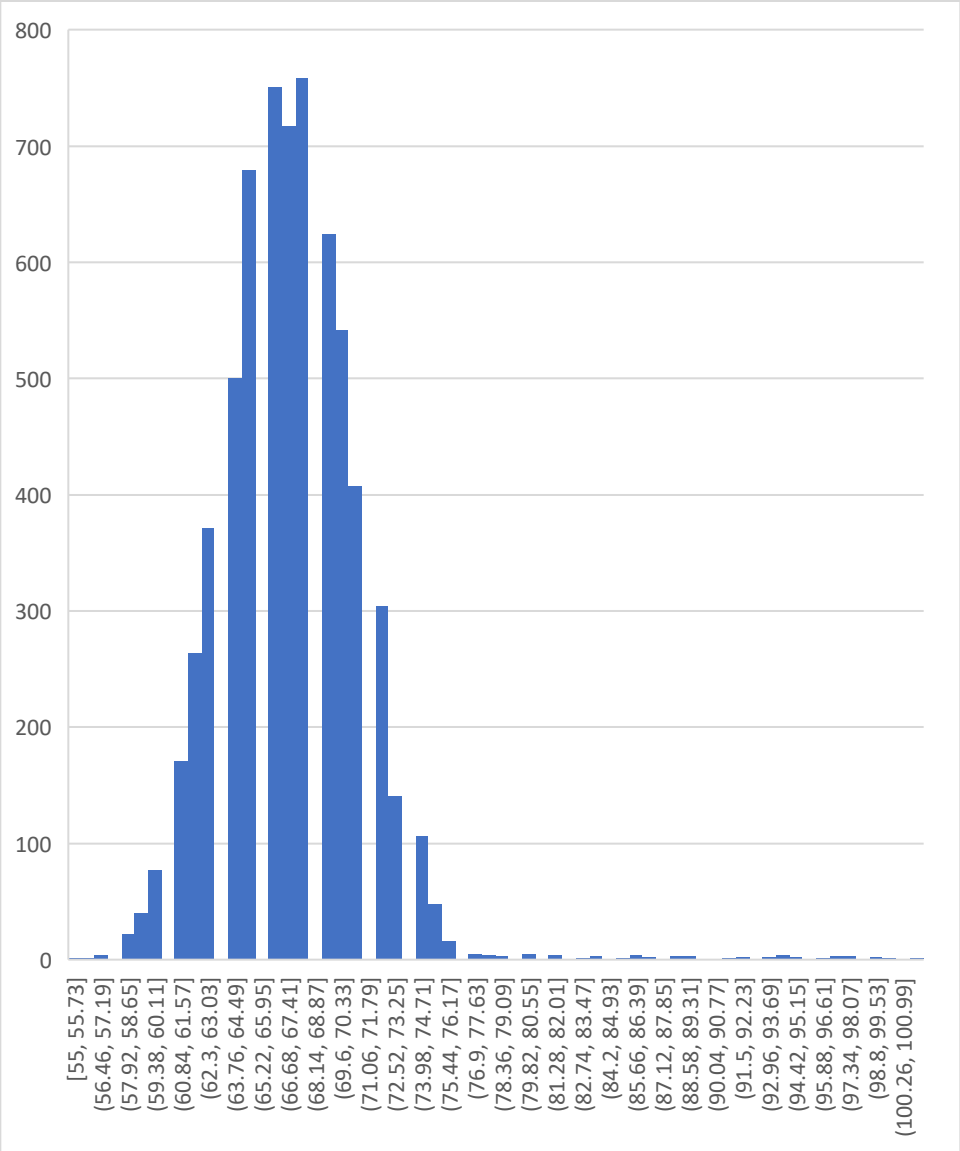
```
SELECT
   COUNT(*) AS total_students,
    AVG(Exam_Score) AS average_exam_score,
    AVG(Hours_studied) AS average_hours_studied,
     AVG(Attendance) AS average_attendance,
    AVG(Parental_involvement) AS average_parental_involvement
     FROM
     studentperformancefactors;
```

total_students	average_exam_score	average_hours_studied	average_attendance	average_parental_involvement
6607	67.2357	19.9753	79.9774	0

Exam Score

Mean	67.271
Standard Error	0.1265131
Median	67
Mode	66
Standard Deviation	4.00069551
Sample Variance	16.0055646
Kurtosis	14.8536734
Skewness	2.34339384
Range	42
Minimum	58
Maximum	100
Sum	67271
Count	1000

Distribution of Exam Score



1. **Shape of the Distribution**:
    - **Normal Distribution**: If the histogram resembles a bell curve (symmetrical with a peak in the middle), the data may follow a normal distribution.
    - **Skewness**:

        - **Right-Skewed (Positive Skew)**: A longer tail on the right side indicates that most scores are lower, with some higher outliers.
        - **Left-Skewed (Negative Skew)**: A longer tail on the left suggests that most scores are higher, with some lower outliers.
    - **Uniform Distribution**: If the bars are roughly equal in height, the data is uniformly distributed.
2. **Central Tendency**:
    - Look for the peak of the histogram, which indicates the mode (the most frequently occurring score).
    - The center of the data can also be inferred; if the histogram is symmetrical, the mean and median will be close to the mode.
3. **Spread (Variability)**:
    - The width of the histogram shows the range of scores. A wider histogram indicates greater variability in student performance.
    - Look at how spread out the scores are and identify the interquartile range, which can be inferred from the histogram.
4. **Outliers**:
    - Check for bars that are isolated from the rest of the distribution. These can indicate outliers in the dataset, which may require further investigation.
5. **Frequency**:
    - The height of each bar represents the number of occurrences (frequency) of scores within each bin. Consider how many students fall into each score range.

Interpretation

- **Shape**: The histogram shows a Bell Curve.Which Means ExamScore is Normally Distributed.
- **Central Tendency**: The peak is around 60 suggesting that this is
- the most common score among students.

By interpreting the histogram, you can gain insights into student performance, such as identifying common scores, assessing variability, and spotting any outliers that may warrant further analysis. This understanding can inform educational strategies, help identify areas where students may need additional support, or celebrate high achievers.

Correlation Analysis

Correlation is a statistical measure that describes the strength and direction of a relationship between two variables.

1. **Direction**:
    - **Positive Correlation**: When one variable increases, the other variable tends to also increase. For example, a positive correlation might exist between study hours and exam scores.
    - **Negative Correlation**: When one variable increases, the other variable tends to decrease. For example, a negative correlation could exist between stress levels and exam performance.
    - **Zero Correlation**: There is no discernible relationship between the two variables. Changes in one variable do not predict changes in the other.
2. **Strength**:
    - Correlation coefficients range from -1 to +1.
        - **+1** indicates a perfect positive correlation.
        - **-1** indicates a perfect negative correlation.
        - **0** indicates no correlation.
    - Values closer to +1 or -1 indicate a stronger relationship, while values closer to 0 indicate a weaker relationship.
3. **Types of Correlation Coefficients**:
    - **Pearson Correlation Coefficient**: Measures linear correlation between two continuous variables. It assumes that the data is normally distributed.
    - **Spearman's Rank Correlation Coefficient**: A non-parametric measure that assesses how well the relationship between two variables can be described by a monotonic function. It’s used for ordinal data or when the assumptions of Pearson’s correlation are not met.

Applications of Correlation

- **Data Analysis**: Helps to identify relationships in datasets, which can inform decisions in fields such as finance, psychology, and social sciences.
- **Predictive Modeling**: Correlation can be a precursor to building predictive models, guiding which variables might be included in analyses.
- **Research**: Used to support hypotheses in scientific research by showing relationships between variables.

Important Considerations

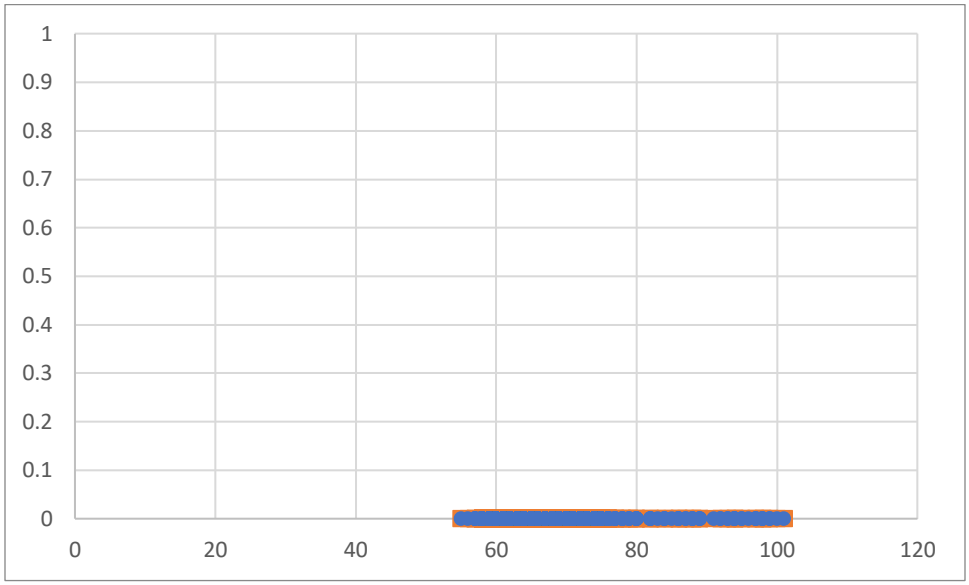
- **Correlation Does Not Imply Causation**: Just because two variables are correlated does not mean that one causes the other. For instance, ice cream sales and drowning incidents might be correlated due to both being related to summer, but one does not cause the other.

Assess the relationship between Hours studied and Exam scores

```
SELECT
   Hours_studied,
   AVG(Exam_Score) AS avg_exam_score
FROM
   studentperformancefactors
GROUP BY
   Hours_studied
ORDER BY
   Hours_studied;
```

Hours_studied,	avg_. exam_score
1	71.0000
2	63.0000
3	61.3333
4	61.6471
5.	62.8571
6	63.4706
7	64.3529
8	64.1552
9	64.1628
10	64.3936
11	64.9795
12	64.7656
13	64.7248
14	65.5762
15	65.5905
16	66.1823
17	66.3281
18	67.0424
19	66.9592
20	66.9505
21	67.6821
22	67.4801
23	68.1800
24	68.1877
25	68.8478
26	68.6654
27	69.4629
28	69.6082
29	70.2836
30	70.6179
31	70.6883

32	70.9444
33	70.0250
34	70.7241
35	71.8000
36	71.1818
37	73.3333
38	72.7143
39	74.7143
43	78.0000
44	71.0000



Scatterplot Shows Moderate Positive Correlation Between ExamScore and Study Hours

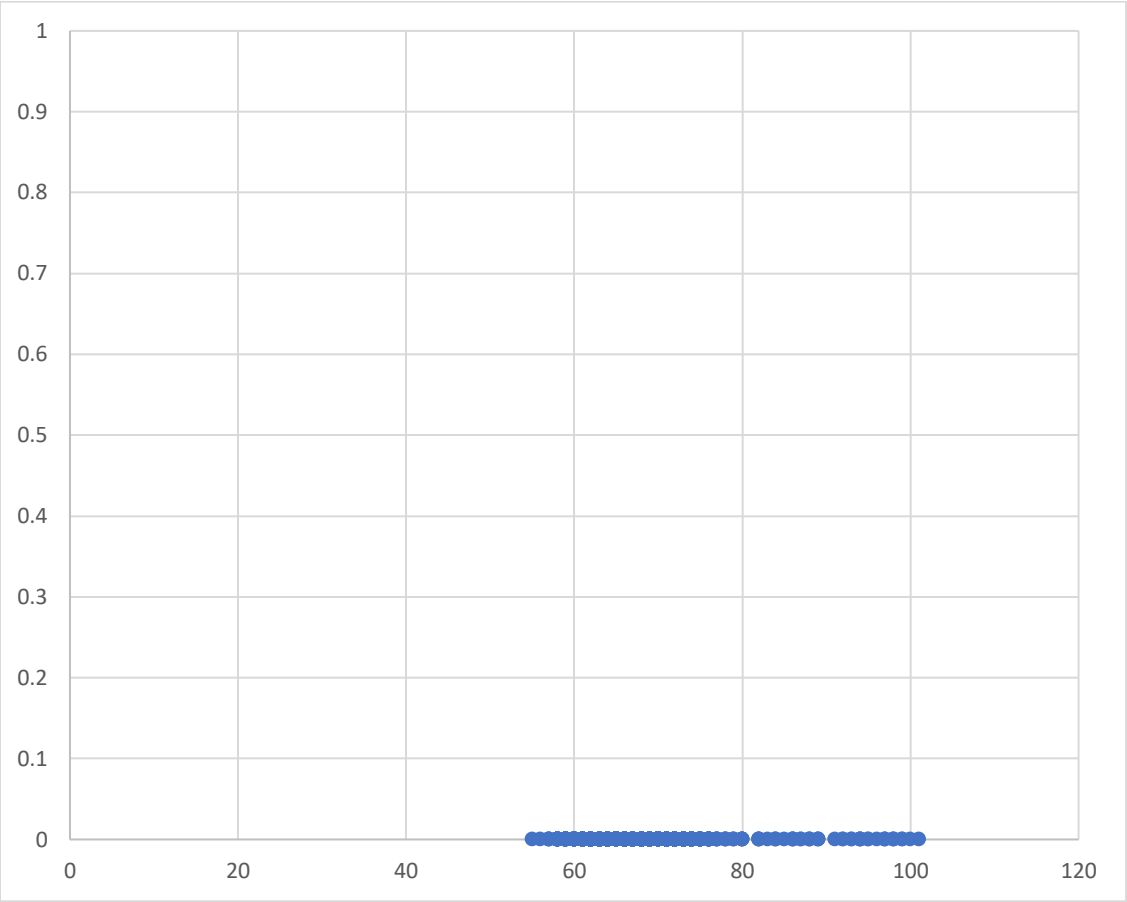
Interpretation

The correlation between Exam_Score and Hours_Studied
Which means Moderate Positive Correlation.

correlation between sleep hours and motivation levels with exam scores

```
SELECT
   sleep_hours,
   AVG(Exam_Score) AS avg_exam_score
FROM
   studentperformancefactors
GROUP BY
   sleep_hours
ORDER BY
   sleep_hours;
```

sleep_hours	avg_exam_score
4	67.6278
5	67.2978
6	67.1948
7	67.243
8	67.2187
9	67.1497
10	67.1378



Interpretation

The correlation between Exam_Score and SleepHours is Moderate Positive Correlation

For motivational level

```
SELECT
   Motivation_level,
   AVG(Exam_Score) AS avg_exam_score
FROM
   studentperformancefactors
GROUP BY
   Motivation_level
ORDER BY
   Motivation_level;
```

Motivation_level	avg_exam_score
High	67.7043
Low	66.7522
Medium	67.3306

Interpretation

Motivation Level and ExamScore is Not Correlated

Explore how extracurricular activities correlate with exam performance

```
   SELECT
   Extracurricular_Activities,
   AVG(Exam_Score) AS avg_exam_score,
   COUNT(*) AS student_count
FROM
   studentperformancefactors
GROUP BY
   Extracurricular_Activities
ORDER BY
   Extracurricular_Activities;
```

Extracurricular_Activities	avg_exam_score	student_count
No	66.9314	2669
Yes	67.4418	3938

Interpretation:

Extracurricular Activities is not Correlated With ExamScore

Student Count and Exam Score

```
SELECT
    Exam_Score,
    COUNT(*) AS student_count
    FROM
    studentperformancefactors
  GROUP BY
  Exam_Score
ORDER BY
  Exam_Score;
```

Exam_Score	student_count
55	1
56	1
57	4
58	22
59	40
60	77
61	171
62	264
63	371
64	501
65	679
66	751
67	717
68	759
69	624
70	542

71	408
72	304
73	141
74	106
75	48
76	16
77	5
78	4
79	3
80	5
82	4
83	1
84	3
85	1
86	4
87	2
88	3
89	3
91	1
92	2
93	2
94	4
95	2
96	1
97	3
98	3
99	2
100	1
101	1

Analyze how parental involvement and education level Effect exam scores

```
SELECT
   Parental_involvement,
   AVG(Exam_Score) AS avg_exam_score,
   COUNT(*) AS student_count
FROM
   studentperformancefactors
GROUP BY
   Parental_involvement
ORDER BY
   Parental_involvement;

   #for parenteducationlevel
SELECT
   parental_education_level,
   AVG(Exam_Score) AS avg_exam_score,
   COUNT(*) AS student_count
FROM
   studentperformancefactors
GROUP BY
   parental_education_level
ORDER BY
   parental_education_level;
```

Parental_involvement	avg_exam_score	student_count
High	68.0928	1908
Low	66.3583	1337
Medium	67.0982	3362

Interpretation:

1908 Students have High Parental Involvement in their studies and their ExamScore is High Compared to Low and Medium Parental Involvement Students. .But there is only small value difference between Exam Score of Students whatever there Parental Involvement.Large Number of Students get Medium Parental Involvement.

parental_education_level	avg_exam_score	student_count
	67.0556	90
College	67.3157	1989
High School	66.8936	3223
Postgraduate	67.9709	1305

Interpretation:

Parental Education Level Not Effect The ExamScore of Students

Evaluate exam performance by gender

```
SELECT
   Gender,
   AVG(Exam_Score) AS avg_exam_score,
   COUNT(*) AS student_count
FROM
   studentperformancefactors
GROUP BY
   Gender;
```

Gender	avg_exam_score	student_count
Female	67.2449	2793
Male	67.2289	3814

Interpretation:

There is Only Slight Difference Between Avg ExamScore of Male And Female Student.Gender of Student Does not Effect there Mark.

Investigate how distance from home and family income Effect exam performance.

```
SELECT
   Distance_from_home,
   AVG(Exam_Score) AS avg_exam_score
FROM
   studentperformancefactors
GROUP BY
   Distance_from_home
ORDER BY
   Distance_from_home;
```

Distance_from_home	avg_exam_score
	66.4328
Far	66.4574
Moderate	66.9815
Near	67.5121

For family income

```
SELECT
   Family_income,
   AVG(Exam_Score) AS avg_exam_score
FROM
   studentperformancefactors
GROUP BY
   Family_income
ORDER BY
   Family_income  ;
```

Family_income	avg_exam_score
High	67.8424
Low	66.8484
Medium	67.335

Interpretation:

Distance From Home and Family income effect Avg Exam Score but not that much involvement in ExamScore only have small value difference in Avg Marks.

TOOLS USED

SQL for querying , Exel for Visualization of Graphs

References

Kaggle, Fundamentals of Mathematical Statistics(S C Gupta)

Sql project

```
SELECT
   COUNT(*) AS total_students,
   AVG(Exam_Score) AS average_exam_score,
   AVG(Hours_studied) AS average_hours_studied,
   AVG(Attendance) AS average_attendance,
   AVG(Parental_involvement) AS average_parental_involvement
FROM
   studentperformancefactors;
```

total_students	average_exam_score	average_hours_studied	average_attendance	average_parental_involvement
6607	67.2357	19.9753	79.9774	0

```
SELECT
   Exam_Score,
   COUNT(*) AS student_count
FROM
   studentperformancefactors
GROUP BY
   Exam_Score
ORDER BY
   Exam_Score;
```

Exam_Score	student_count
55	1
56	1
57	4
58	22
59	40
60	77
61	171
62	264
63	371
64	501
65	679
66	751
67	717
68	759
69	624
70	542
71	408
72	304
73	141
74	106
75	48
76	16
77	5
78	4
79	3
80	5
82	4
83	1
84	3
85	1
86	4

87	2
88	3
89	3
91	1
92	2
93	2
94	4
95	2
96	1
97	3
98	3
99	2
100	1
101	1

Assess the relationship between hours studied and exam scores

```
SELECT
   Hours_studied,
   AVG(Exam_Score) AS avg_exam_score
FROM
   studentperformancefactors
GROUP BY
   Hours_studied
ORDER BY
   Hours_studied;
```

Hours_studied,	avg_. exam_score
1	71.0000
2	63.0000
3	61.3333
4	61.6471
5.	62.8571
6	63.4706
7	64.3529
8	64.1552
9	64.1628
10	64.3936
11	64.9795
12	64.7656
13	64.7248
14	65.5762

15	65.5905
16	66.1823
17	66.3281
18	67.0424
19	66.9592
20	66.9505
21	67.6821
22	67.4801
23	68.1800
24	68.1877
25	68.8478
26	68.6654
27	69.4629
28	69.6082
29	70.2836
30	70.6179
31	70.6883
32	70.9444
33	70.0250
34	70.7241
35	71.8000
36	71.1818
37	73.3333
38	72.7143
39	74.7143
43	78.0000
44	71.0000

Analyze how parental involvement and education level affect exam scores

```
SELECT
   Parental_involvement,
   AVG(Exam_Score) AS avg_exam_score,
   COUNT(*) AS student_count
FROM
   studentperformancefactors
GROUP BY
   Parental_involvement
ORDER BY
   Parental_involvement;

   #for parenteducationlevel
```

```
SELECT
   parental_education_level,
   AVG(Exam_Score) AS avg_exam_score,
   COUNT(*) AS student_count
FROM
   studentperformancefactors
GROUP BY
   parental_education_level
ORDER BY
   parental_education_level;
```

Parental_involvement	avg_exam_score	student_count
High	68.0928	1908
Low	66.3583	1337
Medium	67.0982	3362

For parent education level

```
SELECT
   parental_education_level,
   AVG(Exam_Score) AS avg_exam_score,
   COUNT(*) AS student_count
FROM
   studentperformancefactors
GROUP BY
   parental_education_level
ORDER BY
   parental_education_level;
```

parental_education_level	avg_exam_score	student_count
	67.0556	90
College	67.3157	1989
High School	66.8936	3223
Postgraduate	67.9709	1305

Explore how extracurricular activities and access to resources correlate with exam performance

```
   SELECT
   Extracurricular_Activities,
   AVG(Exam_Score) AS avg_exam_score,
   COUNT(*) AS student_count
```

```
FROM
   studentperformancefactors
GROUP BY
   Extracurricular_Activities
ORDER BY
   Extracurricular_Activities;
```

Extracurricular_Activities	avg_exam_score	student_count
No	66.9314	2669
Yes	67.4418	3938

Evaluate exam performance by gender

```
SELECT
   Gender,
   AVG(Exam_Score) AS avg_exam_score,
   COUNT(*) AS student_count
FROM
   studentperformancefactors
GROUP BY
   Gender;
```

Gender	avg_exam_score	student_count
Female	67.2449	2793
Male	67.2289	3814

Check if there's a correlation between sleep hours and motivation levels with exam scores

```
SELECT
   sleep_hours,
   AVG(Exam_Score) AS avg_exam_score
FROM
```

```
   studentperformancefactors
GROUP BY
   sleep_hours
ORDER BY
   sleep_hours;
```

sleep_hours	avg_exam_score
4	67.6278
5	67.2978
6	67.1948
7	67.243
8	67.2187
9	67.1497
10	67.1378

For motivational level

```
SELECT
   Motivation_level,
   AVG(Exam_Score) AS avg_exam_score
FROM
   studentperformancefactors
GROUP BY
   Motivation_level
ORDER BY
   Motivation_level;
```

Motivation_level	avg_exam_score
High	67.7043
Low	66.7522
Medium	67.3306

Investigate how distance from home and family income affect exam performance.

```
SELECT
   Distance_from_home,
   AVG(Exam_Score) AS avg_exam_score
FROM
   studentperformancefactors
GROUP BY
   Distance_from_home
ORDER BY
   Distance_from_home;
```

Distance_from_home	avg_exam_score
	66.4328
Far	66.4574
Moderate	66.9815
Near	67.5121

For family income

```
SELECT
   Family_income,
   AVG(Exam_Score) AS avg_exam_score
FROM
   studentperformancefactors
GROUP BY
   Family_income
ORDER BY
   Family_income  ;
```

Family_income	avg_exam_score
High	67.8424
Low	66.8484
Medium	67.335

After executing these queries, summarize the key insights. For example:

- How do study habits correlate with exam scores?
- Is there a significant gender gap in performance?
- What factors appear to have the most influence on academic success?