12219897

Exploratory Data Analysis on Student Performance Dataset

Dissertation submitted in fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

Velagana Balaji

Registration no **:12219897**

Section**: K22UP**

Supervisor

VED PRAKASH CHAUBEY (63982)

(Technical Institutor of EDA-Up Grad)

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled **Exploratory Data Analysis on Student Performance Dataset** in partial fulfilment of the requirement for the award of Degree for Bachelor of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor **Mr. VED PRAKASH CHAUBEY**. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Velagana Balaji
Signature of Candidate

Reg. No: 12219897

12219897

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the B.Tech Dissertation/dissertation proposal entitled "**Exploratory Data Analysis on Student Performance Dataset"**, submitted **Velagana Balaji** at **Lovely Professional University, Phagwara, India** is a Bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Name of Supervisor)

Date:

TABLE OF CONTENTS

ACKNOWLEDGEMENT

I would like to take this opportunity to express my profound gratitude to Ved Prakash Sir, whose invaluable guidance, mentorship, and unwavering support made this project on EDA on Student Performance data. From the initial conception of the idea to the final stages of this project, his vast knowledge and expertise in data science and machine learning have been instrumental in shaping the direction and depth of my research. His insightful advice and constructive feedback consistently pushed me to explore beyond the surface and helped refine my work to a higher standard.

Ved Prakash Sir's passion for teaching and his dedication to his students have truly been a source of inspiration throughout this journey. His ability to simplify complex concepts, coupled with his approachable nature, created an environment that encouraged learning and curiosity. I am immensely thankful for the time and effort he invested in reviewing my work, providing critical input, and offering continuous encouragement during challenging phases of the project. His feedback was not only crucial for the technical aspects of the project but also helped me build a stronger foundation in the principles of data analysis and machine learning.

I would also like to extend my appreciation to my institution for providing the necessary resources and a conducive learning environment that allowed me to pursue this research. My heartfelt thanks to my colleagues and classmates, who have been a constant source of motivation, encouragement, and collaboration. Their shared knowledge and perspectives contributed significantly to my understanding of the subject matter.

In conclusion, I would like to express my sincere gratitude to everyone who has contributed to the completion of this project. The learning and experiences I have gained throughout this process will undoubtedly serve as a foundation for my future endeavors in the field of data science and machine learning. This project would not have been possible without the invaluable guidance, support, and encouragement from Ved Prakash Sir, and I am truly grateful for the opportunity to learn under his mentorship.

ABSTRACT

This project conducts a comprehensive Exploratory Data Analysis (EDA) of the Student Performance Dataset to uncover patterns, trends, and correlations in academic performance. The dataset includes features such as gender, ethnic group, parental education, and marital status, providing insights into how familial and socio-demographic factors influence academic performance. For instance, the analysis investigates how parental education levels correlate with students' scores and whether the marital status of parents has an impact on their children's achievements. Additionally, lifestyle factors such as the frequency of sports participation and self study hours are explored to determine their contributions to academic success. The dataset encompasses diverse features, including demographics, parental background, lifestyle choices, and scores in mathematics, reading, and writing. By examining variables such as gender, ethnic group, parental education, marital status, weekly study hours, and participation in sports, the study explores how socio-demographic and lifestyle factors influence academic outcomes. The analysis includes data cleaning and preprocessing, followed by univariate, bivariate, and multivariate analyses to investigate distributions and relationships between variables. Key insights are visualized using bar charts, scatter plots, and heatmaps for accessibility. Specific areas of focus include the impact of parental education and marital status on scores, the effect of test preparation and lunch type, and the influence of family dynamics, such as birth order and sibling count, on performance. The findings aim to guide educators, parents, and policymakers in designing strategies to enhance study habits, promote extracurricular activities, and address educational inequalities. Understanding the factors influencing student performance is critical for designing effective educational strategies. This project undertakes a comprehensive Exploratory Data Analysis (EDA) of the Student Performance Dataset, which contains rich information on demographics, parental background, lifestyle choices, and academic scores in mathematics, reading, and writing. By examining these variables, the study aims to uncover patterns, trends, and correlations that can guide educators, parents, and policymakers in enhancing student outcomes. Finally, the study also delves into patterns related to academic engagement, such as whether being the first child in a family or the number of siblings impacts performance. The role of transportation means and school lunch types are analyzed to assess their relevance to student outcomes. By synthesizing findings from various analyses, the project offers actionable insights to diverse stakeholders, including strategies for improving study habits, promoting extracurricular activities, and addressing inequalities in education.

Understanding the factors influencing student performance is critical for designing effective educational strategies.This project undertakes a comprehensive Exploratory Data Analysis (EDA) of the Student Performance Dataset, which contains rich information on demographics, parental background, lifestyle choices, and academic scores in mathematics, reading, and writing. By examining these variables, the study aims to uncover patterns, trends, and correlations that can guide educators, parents, and policymakers in enhancing student outcomes.

INTRODUCTION

Student performance is a multifaceted phenomenon influenced by a combination of demographic, familial, and lifestyle factors. Understanding these influences is crucial for educators, policymakers, and parents striving to enhance academic outcomes. This study focuses on the Student Performance Dataset, which provides detailed information about students, including their demographics, parental background, lifestyle habits, and academic scores in mathematics, reading, and writing. By exploring this dataset, we aim to uncover trends and patterns that can provide actionable insights into factors driving student achievement. This dataset provides a comprehensive overview of student performance, encompassing various demographic, socioeconomic, and academic factors. It includes information on gender, ethnicity, parental education, lunch type, test preparation, parental marital status, sports participation, family size, transportation, study habits, and scores in math, reading, and writing. The dataset includes key features such as gender, ethnic group, parental education, and marital status, which reflect the socio-demographic context of the students. These variables offer valuable insights into how students' backgrounds affect their academic performance. For example, the level of parental education may influence access to resources or attitudes toward education, while marital status might impact the emotional and financial support available to the student. Such analyses can provide a nuanced understanding of the role of family dynamics in shaping academic success. Lifestyle factors, including participation in sports, weekly study hours, and transportation to school, are also integral to the study. These aspects highlight how students balance academics with other activities and how external conditions affect their ability to perform in school. For instance, regular engagement in sports may promote discipline and better time management, while access to private transportation could reduce stress and enhance attendance. Analyzing these variables helps to identify actionable strategies for fostering holistic student development. This dataset provides a comprehensive view of student performance, encompassing a range of demographic, socioeconomic, and academic factors. By analyzing this data, we can gain valuable insights into the factors that influence student achievement, such as parental education, family size, study habits, and extracurricular activities. These insights can be used to develop targeted interventions, inform educational policies, and ultimately improve student outcomes. This research leverages advanced data analysis techniques, including univariate, bivariate, and multivariate analyses, alongside data visualization, to present insights clearly and comprehensively.

The findings from this analysis not only shed light on the various factors influencing student performance but also provide a foundation for developing targeted interventions and policies. This study underscores the importance of a data-driven approach in addressing challenges within the education system.

PROBLEM STATEMENT

Student performance is a critical metric in assessing the effectiveness of educational systems and identifying areas for intervention to ensure academic success. The dataset under study captures a variety of factors, including demographic information, parental background, extracurricular habits, and test scores in mathematics, reading, and writing. By analyzing this dataset, the goal is to uncover relationships and trends that influence student academic achievement and highlight actionable insights for educators, policymakers, and parents. Understanding the impact of socioeconomic and demographic variables, such as gender, ethnic group, and parental education, can provide valuable information on systemic disparities in education. Additionally, analyzing variables like lunch type, transportation means, and test preparation reveals the potential influence of external support and resources on student outcomes. This information can guide targeted initiatives to bridge gaps and enhance learning outcomes. Extracurricular activities, such as practicing sports and self-study hours, are also vital aspects of student life. These habits may influence academic performance by affecting time management, cognitive development, and overall well-being. Exploring these variables allows for a better understanding of how lifestyle choices and discipline correlate with academic success. Similarly, familial factors, such as the number of siblings and being the first child, may shed light on the dynamics of resource allocation and parental attention. Using a dataset that includes variables like gender, ethnic group, parental education, lunch type, test preparation, parental marital status, sports participation, sibling count, transportation means, weekly study hours, and test scores in math, reading, and writing, this study seeks to identify the major factors influencing student performance. Finding important indicators of academic achievement and patterns or differences in performance between various student groups are the main goals. The objective of this study is to identify key factors driving performance in math, reading, and writing, and to assess the interactions between them. This analysis will not only help in predicting performance but also in designing evidence-based interventions for underperforming students .Ultimately, the findings aim to empower stakeholders in creating a supportive educational environment that fosters equity and excellence

LITERATURE

Overview:

Exploratory Data Analysis (EDA), multivariate analysis, and feature engineering are critical steps in preparing and optimizing data for machine learning models. Numerous studies emphasize these methods for analyzing high-dimensional datasets, such as Student performance datasets, where parameters like gender, ethnic group, parental education, lunch type, test preparation, parents marriage status, number of siblings for the students, and marks of mathematics, reading and writing are important factors for predicting outcomes such as reason behind student academic and non-academic performance.

Exploratory Data Analysis (EDA):

The first step in working with this dataset is to perform an in-depth EDA to understand the distribution and relationship of features.

- **Distribution of Features**:
    - Visualizing the distribution of students' test scores in mathematics, reading, and writing can provide valuable insights into their academic performance. Histograms or box plots of these scores might reveal variations across different groups. For instance, the math scores may have a bimodal distribution, indicating distinct performance levels among students

    - Exploring demographic and socioeconomic variables such as gender, ethnic group, and parental education can reveal disparities in academic performance. Box plots grouped by these factors could show differences in mean test scores, potentially indicating systemic inequities.

- **Correlation Analysis**:

    - A heatmap of correlations between features like MathScore, ReadingScore, WritingScore, and other continuous variables, such as NrSiblings, could help identify strong linear relationships. For example, math and reading scores might show a strong positive correlation, indicating that students who excel in one subject often perform well in others.

    - Categorical features like TestPrep and LunchType can be analyzed for their interaction with continuous variables like test scores using violin plots or swarm plots. For instance, students who completed test preparation courses may exhibit higher average scores, which can be visualized effectively with violin plot.

- **Missing Values**:
    - Missing values are a common issue in real-world datasets. For instance, some entries might not have taght about Is first child or about their ethnic group. Imputation techniques or simply flagging these missing values could become an important aspect of the data preprocessing phase.

2. Multivariate Analysis:

Once EDA is complete, multivariate techniques are applied to better understand how these features interact and to reduce the dimensionality of the dataset.

Analyzing Relationships Between Test Scores

Multivariate analysis of the dataset provides an opportunity to examine relationships between test scores in mathematics, reading, and writing. A pairplot or scatterplot matrix can visualize how these scores correlate, helping identify whether students who perform well in one subject also excel in others. Strong positive correlations between these scores might suggest that underlying skills, such as comprehension or logical reasoning, influence performance across subjects.

Exploring the Impact of Categorical Variables

Analyzing how categorical variables interact with test scores reveals valuable insights into demographic and behavioral influences on academic performance. For instance, group-wise box plots can illustrate how LunchType (standard vs. free/reduced) and TestPrep (completed vs. not completed) impact scores. It's plausible that students with standard lunches score higher, reflecting the influence of nutrition and structured learning strategies. Similarly, group comparisons across EthnicGroup or ParentEduc may uncover disparities, offering critical insights for targeted interventions.

Interaction Effects of Lifestyle and Socioeconomic Factors

Exploring the combined effects of lifestyle factors, such as WklyStudyHours and PracticeSport, alongside socioeconomic variables like ParentMaritalStatus, can provide a holistic understanding of academic outcomes. Multivariate visualizations like clustered bar charts or interaction plots could help analyze how these factors jointly affect test scores. For example, students who regularly practice sports and study more than 10 hours weekly may achieve higher scores, but the effect might vary depending on their parental marital status or education level.

3. Feature Engineering:

This step involves creating new features or transforming existing ones to make the data more

useful for modeling.

- **Interaction Terms**:
  - New features could be engineered by combining existing ones to capture interactions that are not immediately obvious. For example, combining ParentEduc and EthnicGroup into a single feature could reveal patterns related to the intersection of cultural and educational influences on student performance.

- **Binarization**:
  - Categorical variables like Gender (male/female), TestPrep (completed/not completed), and LunchType (standard/free or reduced) can be converted into binary features or one-hot encoded variables to enhance their compatibility with machine learning algorithms. For instance, creating binary flags for TestPrep completion can help capture its impact on student performance across different test scores.

- **Total/Average Score**:
  - A new feature, Total/Average score, could be engineered by combining total of math, reading and writing score and dividing by three. This score might provide a more holistic view of student performance, as both the number of marks and the average marks would play a role in indicating student performance.

4. Feature Selection:

Finally, feature selection techniques help reduce the number of features, making the model simpler and faster to train without sacrificing accuracy.

- **Correlation-based Feature Selection:**
  After performing correlation analysis, redundant features, such as MathScore, ReadingScore, and WritingScore if they are highly correlated, could potentially be combined into a single composite feature (e.g., AverageScore). This reduces redundancy while preserving the essential information, allowing the model to focus on meaningful patterns without being overwhelmed by duplicate data.

- **Recursive Feature Elimination (RFE):**
  - By using RFE, less relevant features such as TransportMeans or PracticeSport (if found to have minimal impact on the target variable) could be excluded from the final model. This approach ensures the model focuses only on the most predictive variables, improving both efficiency and accuracy

METHODOLOGY

1. Overview:

In this methodology, we detail the step-by-step process used to clean, manipulate, and analyze a student performance dataset containing features such as Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, PracticeSport, WklyStudyHours, and test scores (MathScore, ReadingScore, WritingScore). The process included data preprocessing, feature engineering, and applying multivariate analysis techniques to prepare the dataset for predictive modeling and insightful analysis.

2. Data Preprocessing:

To ensure the dataset is clean and ready for analysis, the following data preprocessing steps were undertaken:

- **Handling Missing Data**:
  - Imputation: For columns with missing values (e.g., ethnic group, parent education and number of siblings), imputation methods were used. For numerical data such as Mathscore, the median was chosen to fill in missing values, as it is less sensitive to outliers. Categorical columns like parent education and lunch type were imputed using the mode (most frequent value).
  - Dropping Missing Entries: In cases where there were substantial missing values and imputation wasn't reliable, those rows were dropped.

- **Handling Outliers**:
  - Box Plot Visualization: Outliers in numerical features such as reading score,writing score , and number of siblings were identified through box plots. Outliers could potentially skew the analysis, so these extreme values were either removed or transformed, depending on their relevance.
  - Log Transformation: If features like votes had a highly skewed distribution, a log transformation was applied to reduce the effect of extreme values and normalize the distribution.

- **Standardization/Normalization**:
  - Continuous variables like number of siblings and writing score were standardized using z- score normalization to ensure that features are on the same scale, especially since some models are sensitive to the magnitude of input values.

3. Tools and Libraries:

To perform these tasks, several Python libraries were utilized:

- Pandas: For data manipulation and handling missing values.
- NumPy: For numerical operations and matrix manipulations.
- Matplotlib & Seaborn: For data visualization to explore the distribution of data, visualize outliers, and understand relationships between features (e.g., scatter plots, histograms, and heatmaps).
- Scikit-learn: For applying machine learning techniques, including PCA, clustering, and feature selection methods.

4. Multivariate Analysis Techniques:

To better understand the relationships between features and reduce data dimensionality, multivariate analysis techniques were applied:

- **Correlation Matrix**:
  - A correlation matrix was constructed to analyze relationships between numerical features such as MathScore, ReadingScore, and WritingScore. A heatmap was used to visualize these correlations and identify multicollinearity. Features with a high correlation (e.g., MathScore and ReadingScore) were flagged for potential removal or transformation to prevent redundancy and ensure that the model doesn't overemphasize certain variables. This step helps streamline the analysis by eliminating variables that provide similar information.

- **Principal Component Analysis (PCA)**:
  - **Principal Component Analysis (PCA)** was employed to reduce the dimensionality of the student performance dataset by transforming correlated features, such as MathScore, ReadingScore, and WritingScore, into principal components that capture the most significant variance in the data. This step simplified the analysis by compressing features like study hours, number of siblings, and parental education into fewer components while retaining essential information about student performance. PCA helped improve computational efficiency and model interpretability.

  - The first two principal components were visualized using scatter plots, revealing clusters of students with similar performance patterns or characteristics (e.g., students who study more regularly and perform better in multiple subjects). These visualizations helped to identify underlying patterns and outliers, enabling further analysis on how various factors, such as parental education or test preparation, influenced student achievement..

5. Feature Engineering and Transformation:

Feature engineering and transformation were key steps in enhancing model performance by creating new, more informative variables:

- **Creating New Features**:
  - Total/Average Score: A new feature, total/average score, was engineered by combining math, reading and writing score to better about student performance and understanding.
  - Interaction Terms: Interaction terms between MathScore, ReadingScore, and WritingScore features might have varying distributions, applying transformations such as a log transformation or standard scaling could help normalize the data for modeling.
  - Binarization of Categorical Features: For instance, creating binary flags for TestPrep completion can help capture its impact on student performance across different test scores. Ethnic group and number of siblings were one- hot encoded to capture the distinct types and geographical effects in the dataset.

- **Feature Transformation**:
  - Log Transformation: As mentioned earlier, log transformation was applied to highly skewed features like mathscore to reduce the effect of extreme values and bring the data closer to a normal distribution.
  - Polynomial Features: Polynomial features were created for gender and number of siblings to capture non-linear relationships, improving model complexity where needed.

RESULT AND ANALYSIS

Fig1:Importing all the required libraries

```
[3]: #importing libraries
     import pandas as pd
     import  numpy as np
     import matplotlib.pyplot as plt
     from scipy import stats
     import seaborn as sns
```

Fig2: Reading the dataset

```
[4]: df = pd.read_csv('Studentdata_eda.csv')
```

Fig3: Information of the dataset.

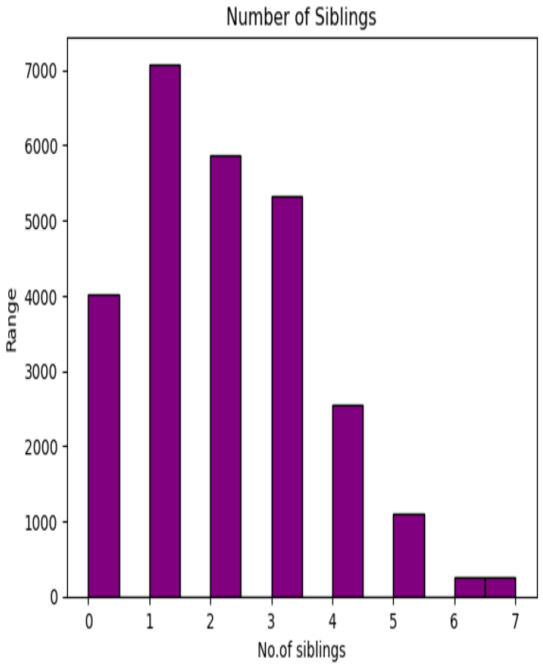
```
#### Dataset info


##### Gender: Gender of the student (male/female)
##### EthnicGroup: Ethnic group of the student (group A to E)
##### ParentEduc: Parent(s) education background (from some high school to master's degree)
##### LunchType: School lunch type (standard or free/reduced
##### TestPrep: Test preparation course followed (completed or not)
##### ParentMaritalStatus: Parent(s) marital status (married/single/widowed/divorced)
##### PracticeSport: How often the student practice sport (never/sometimes/regularly)
##### IsFirstChild: If the child is first child in the family or not yes/no)
##### NrSiblings: Number of siblings the student  (0 to 7)
##### TransportMeans: Means of transport to school (school bus/private)
##### WklyStudyHours: Weekly self-study hours(less that 5hrs; between 5 and 10hrs; more than 10hrs)
##### MathScore: maths test score(0-100)
##### ReadingScore: reading test score(0-100)
##### WritingScore: writing test score(0-100)writing test score(0-100)
```

Fig 4: Cleaning the data

```
[15]: df1.dropna(subset=['Unnamed: 0','TransportMeans','ParentMaritalStatus'], inplace=True)

[16]: df1.shape

[16]: (26445, 15)

[17]: df1.isnull().sum()

[17]: Unnamed: 0                0
      Gender                    0
      EthnicGroup            1581
      ParentEduc             1581
      LunchType                 0
      TestPrep               1563
      ParentMaritalStatus       0
      PracticeSport           536
      IsFirstChild            784
      NrSiblings             1358
      TransportMeans            0
      WklyStudyHours          834
      MathScore                 0
      ReadingScore              0
      WritingScore              0
      dtype: int64

[18]: df1['EthnicGroup'].value_counts()

[18]: EthnicGroup
      group C     7964
      group D     6458
      group B     5039
      group E     3488
      group A     1915
```

Fig 5: Histogram – Showing the number of siblings of the students

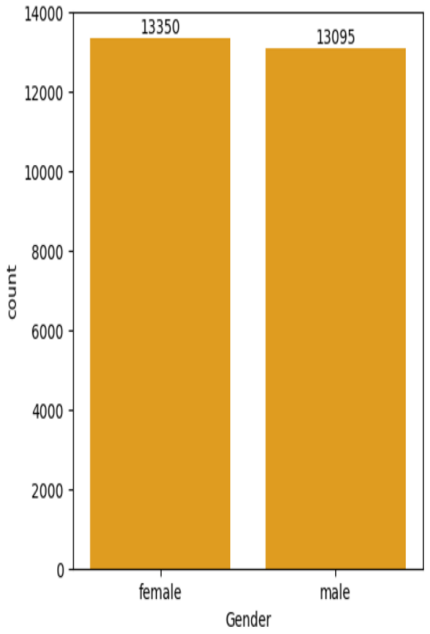
```
import matplotlib.pyplot as plt

# Histogram for a specific column
df1['NrSiblings'].plot(kind='hist', bins=14, color='purple',edgecolor='black')
plt.title('Number of Siblings')
plt.xlabel('No.of siblings')
plt.ylabel('Range')
plt.show()
```



From the above graph we can say that the single siblings are more for the most of the students and very less studenets are having 6 or 7 siblings.

Fig 6: Count plot – Showing the number of males and females in the dataset.

```
[39]: plt.figure(figsize= (5,5))
ax = sns.countplot(data = df1, x = "Gender",color="orange")
ax.bar_label(ax.containers[0])
plt.show()
```



From the above graph we can observe that the number of females in the dataset are more than the number of males

Fig 7: Pie Chart -We can see the portion of different Ethnic groups in the dataset

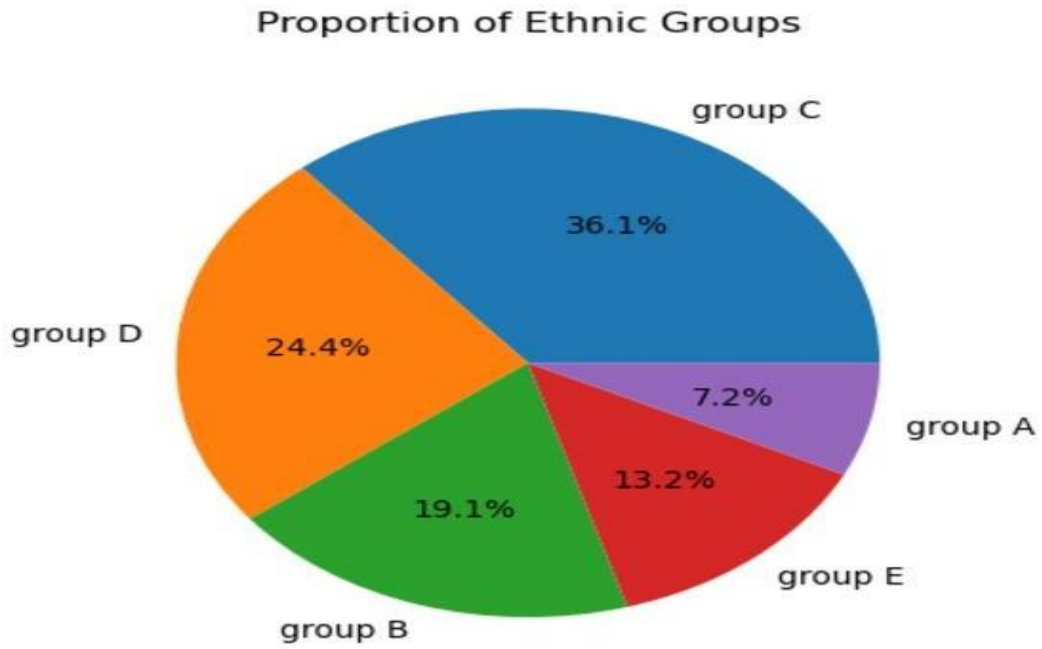


Fig8: Boxplot -Showing the outliers present in the MathScore values in the dataset. It reveals a right-skewed distribution, suggesting that a majority of students scored relatively low in Math, while a few students scored exceptionally high.

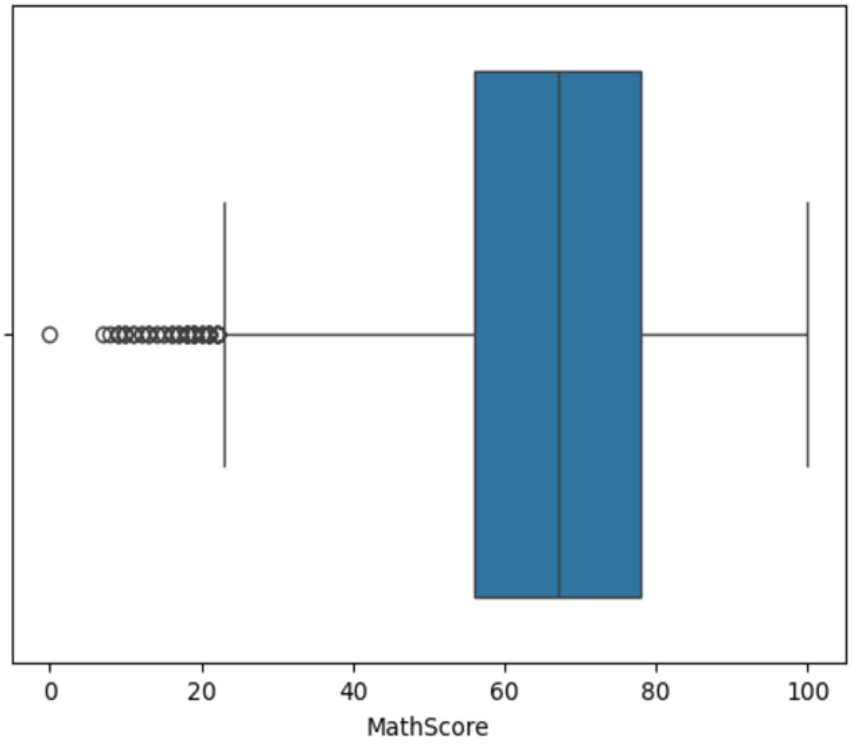


Fig9: Bar Chart - The bar plot shows the average number of siblings for students whose parents have different levels of education

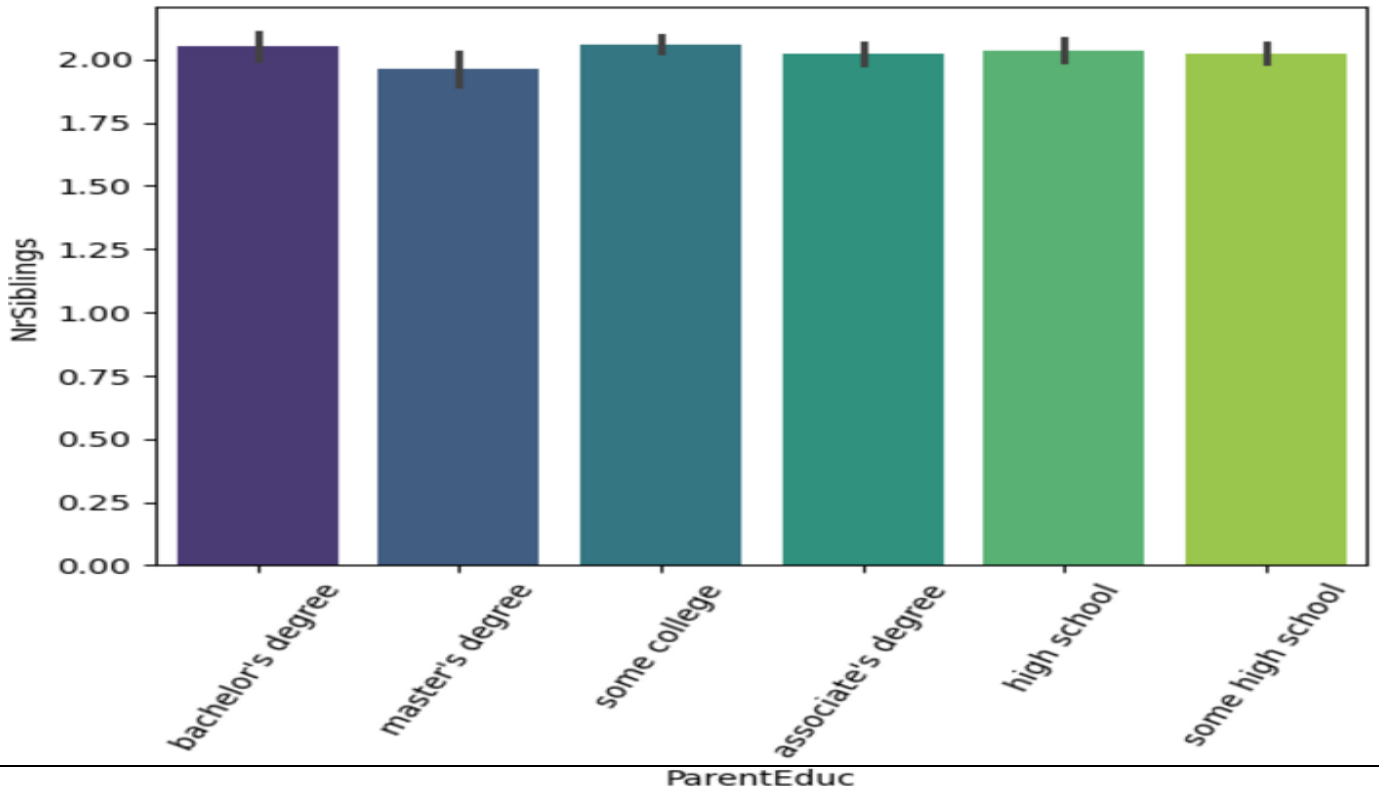
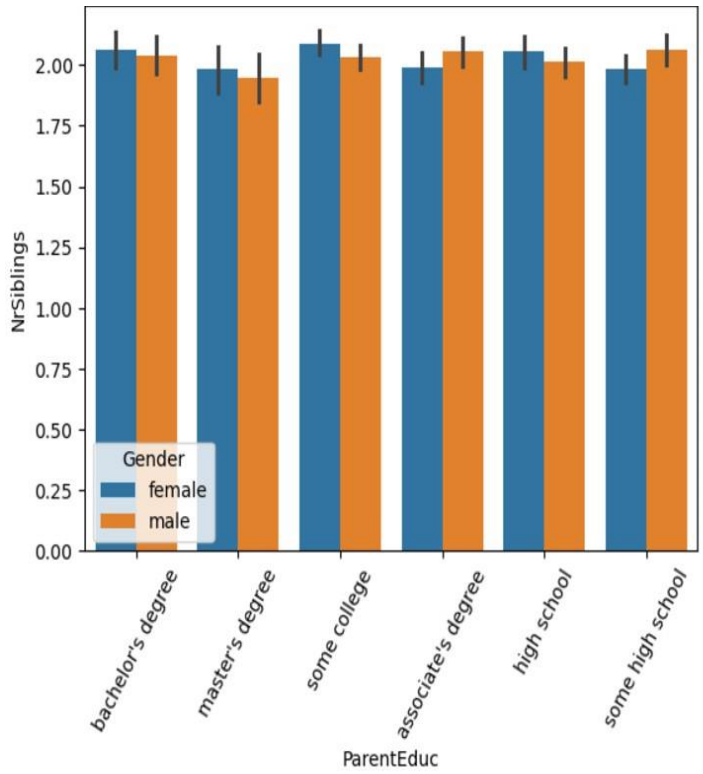
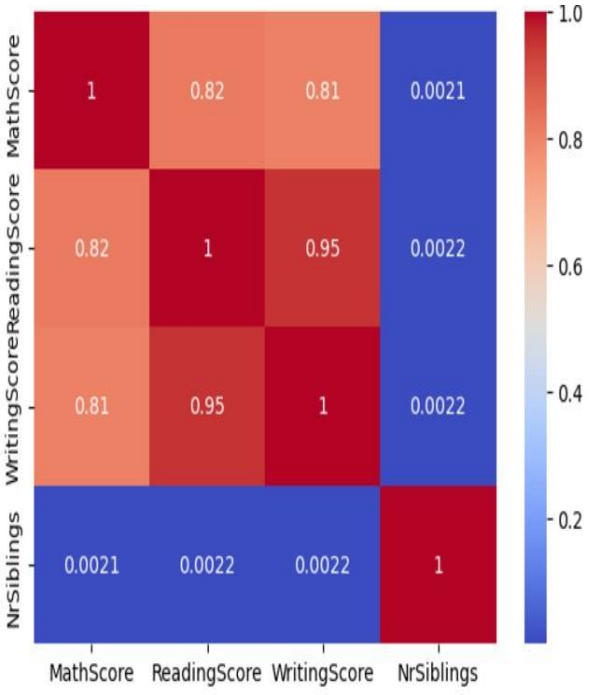


Fig10: Bar Chart- The bars are grouped by gender (female and male), and the height of each bar within a group shows the average number of siblings for that parent education level and gender combination



Bachelor's Degree: Female students whose parents have a bachelor's degree tend to have slightly more siblings compared to male students with similar parental education. High School and Some High School: Male students whose parents have a high school or some high school education tend to have slightly more siblings compared to female students with similar parental education.

Fig 11: Heat map -Showing the relation between marks and number of siblings.



The heatmap reveals that Math, Reading, and Writing scores are strongly related, while the number of siblings has a negligible impact on these scores. This suggests that factors like cognitive abilities, study habits, or teaching quality might be more influential in determining student performance in these subjects than the number of siblings.

Fig 12: Scatter plot-The scatter plots off the diagonal reveal strong positive correlations between Math, Reading, and Writing scores. This suggests that students who perform well in one subject tend to perform well in the others as well. There could be underlying factors such as cognitive abilities, study habits, or access to resources that contribute to this pattern.

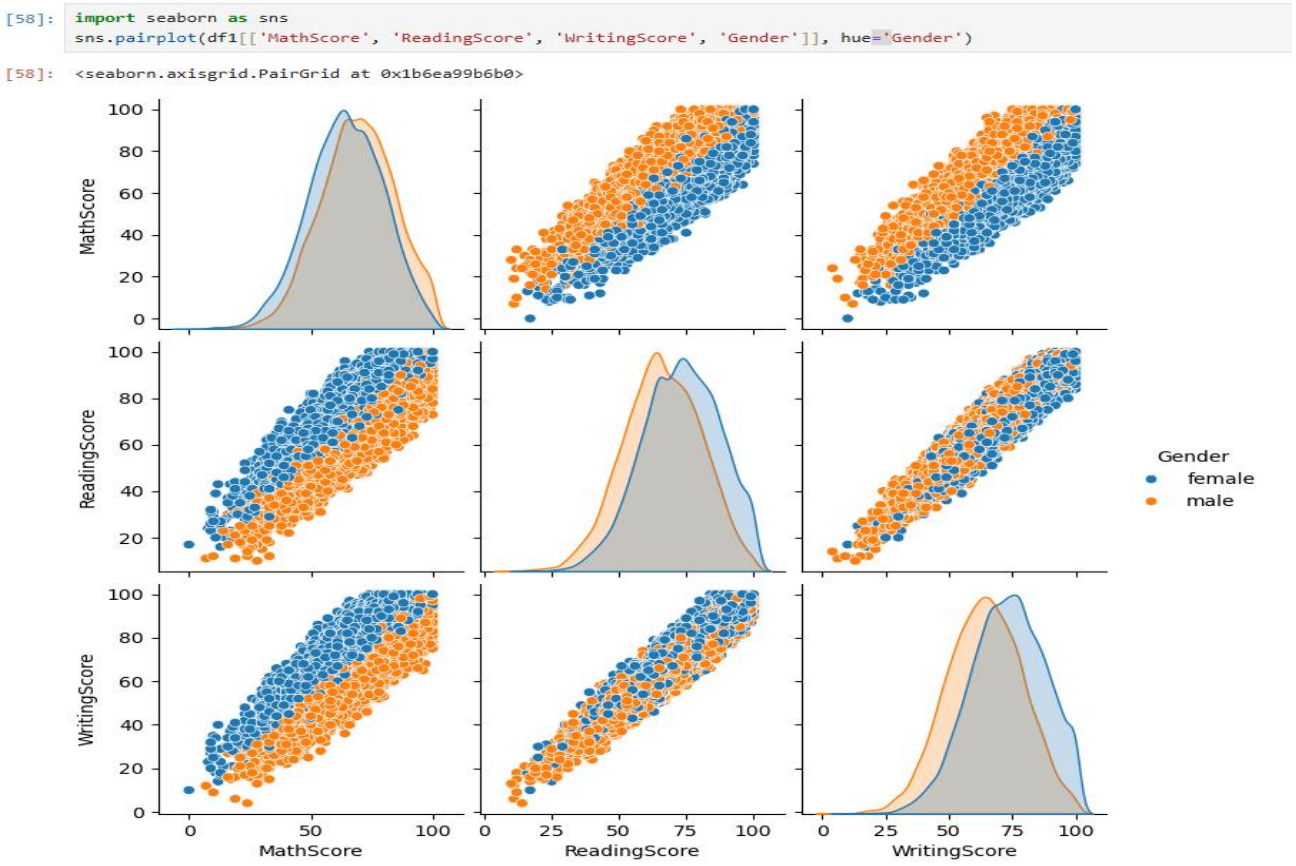
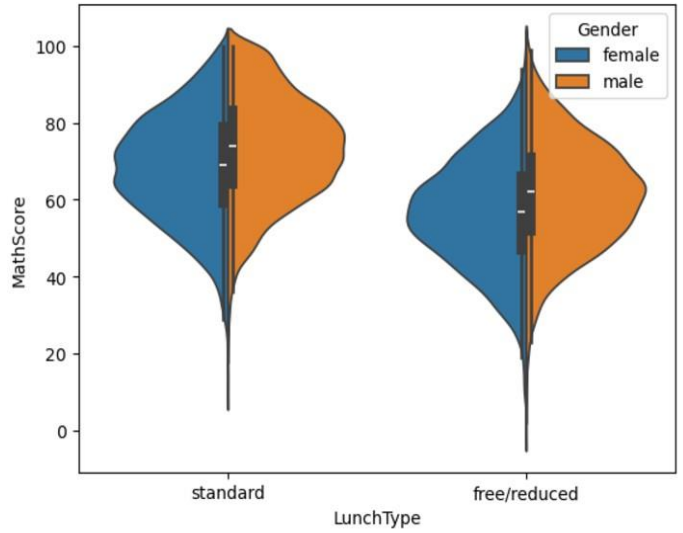
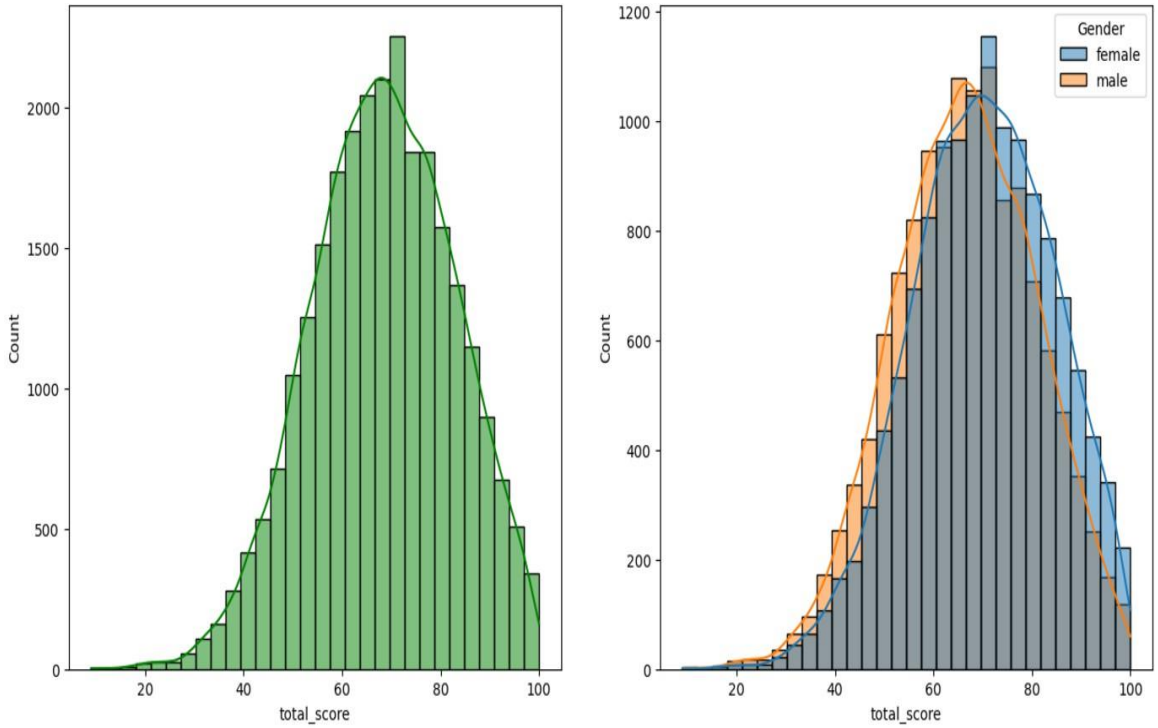


Fig 13: Violin Plot - The violin plots show that the distribution of Math scores is roughly bell-shaped for both lunch types, with some slight skewness.



Distribution Shape: The violin plots show that the distribution of Math scores is roughly bell-shaped for both lunch types, with some slight skewness.

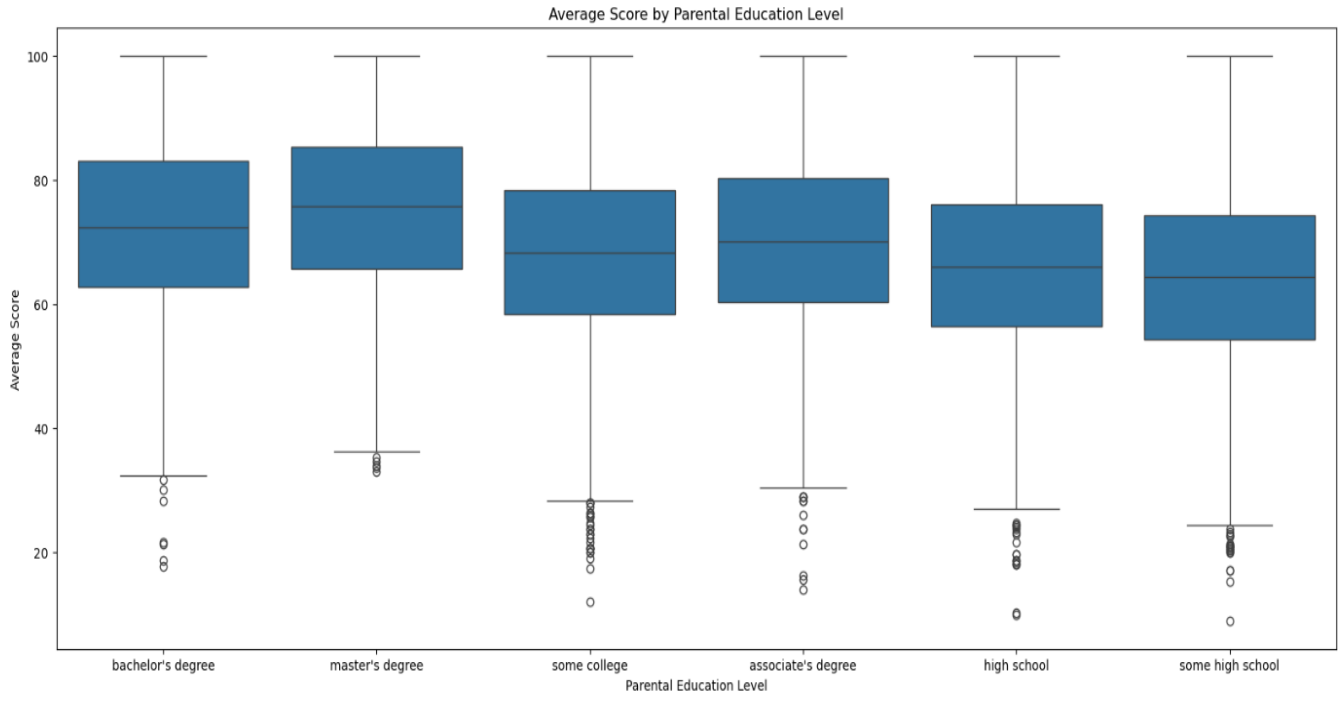
Fig 14: Histogram -The slight rightward shift in the distribution for female students could be attributed to various factors, including societal expectations, educational opportunities, or biological differences. However, further research is needed to confirm these potential explanations.



The histograms you provided show the distribution of the total score (MathScore + ReadingScore + WritingScore) for the entire dataset and separately by gender.

The histogram for the entire dataset shows a bell-shaped distribution, which is characteristic of a normal distribution. This indicates that the total scores are centered around a specific value, and most students have scores close to this average.

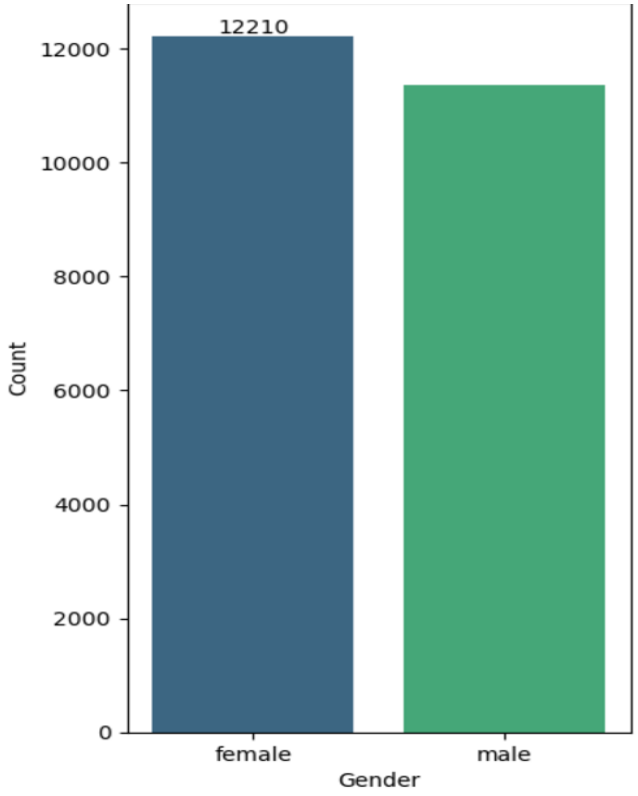
Fig 15: Box plot – The outliers among the parental education and the average score of the students.



Insights

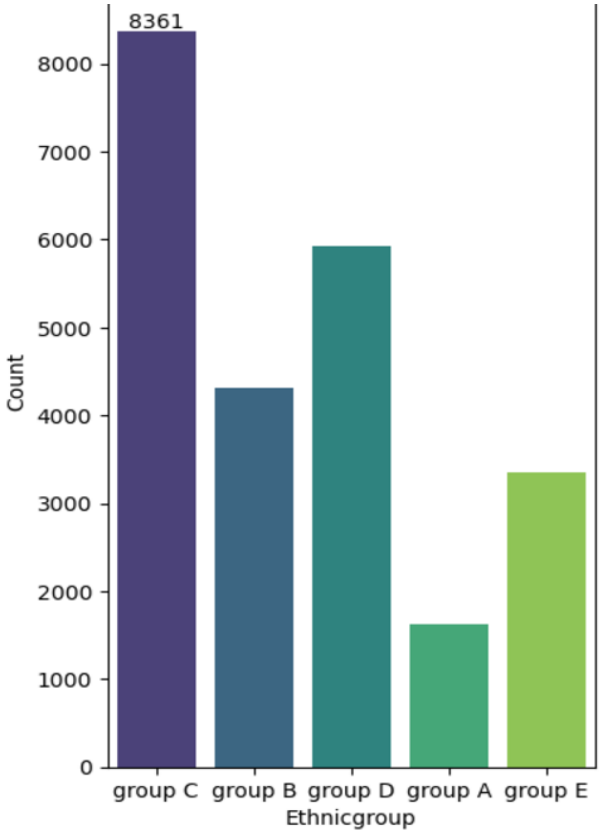
Female students have performed well than male students

Fig 16: Count plot -This graph shows who are passing more and the number of students.



This graph shows that the female students are passing more then compared to male in the student performance dataset

Fig 17: Count plot – This graph shows from which ethnic group most of the students are passing from.



This group shows that from the Ethnic Group-c people are passsing more and the least passing are from the Group-a

Fig 18: Stacked bar chart – This graph shows relation between the level of parent education and passing of students.

```
pd.crosstab(df1['ParentEduc'], df1['did_pass']).plot(kind='bar', stacked=True, colormap='viridis')
plt.title('Pass Status by Parent Education')
plt.xlabel('Parent Education Level')
plt.ylabel('Count')
plt.legend(title='Did Pass')
plt.show()
```

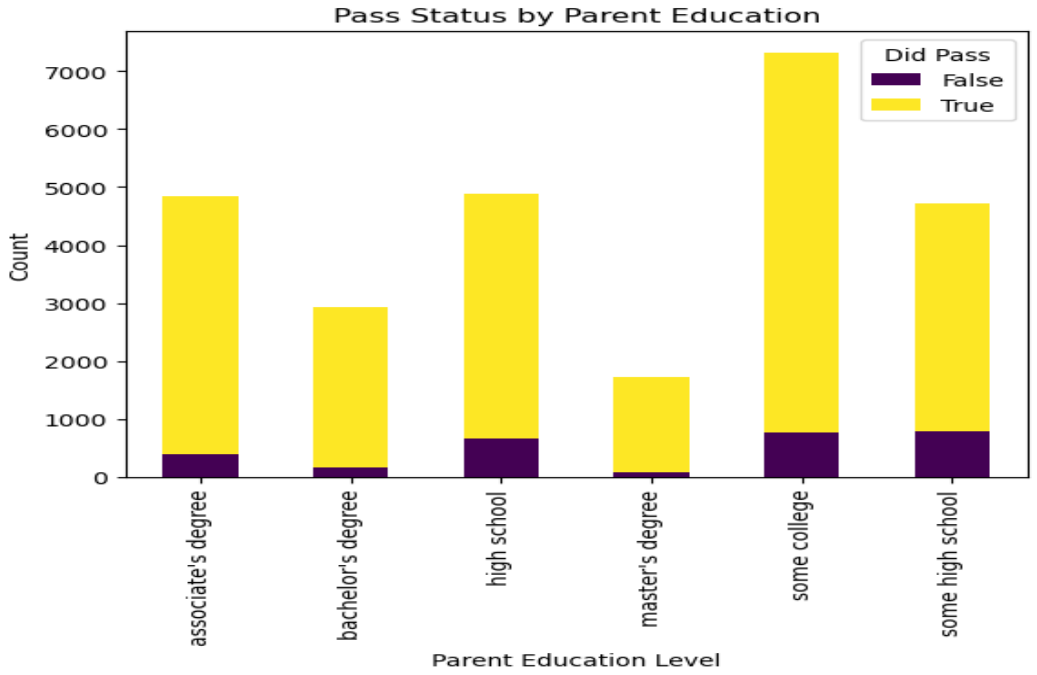
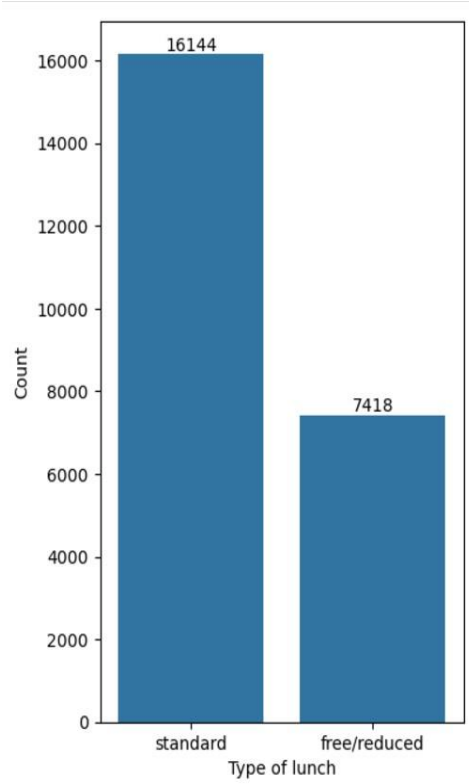


Fig 19: Count plot – This graph shows the effect of the lunch type on the students. We can see that students who are having standard lunch are passing more than the students who are having free/reduced lunch type.



Thhis graph shows that most of the students who are having standard lunch are passing more than compared to the students who are having free/reduced lunch type

Fig 20: Density plot

Fig 20.1-Math Score: Students whose parents have a bachelor's degree, master's degree, or some college tend to have higher Math scores compared to those whose parents have an associate's degree, high school diploma, or some high school education.

Fig 20.2-Reading Score: A similar pattern is observed for Reading scores. Students whose parents have higher levels of education tend to have higher Reading scores.

Fig 20.3-Writing Score: The impact of parental education on Writing scores seems to be less pronounced compared to Math and Reading. However, there is still a noticeable trend of higher Writing scores for students whose parents have higher levels of education.

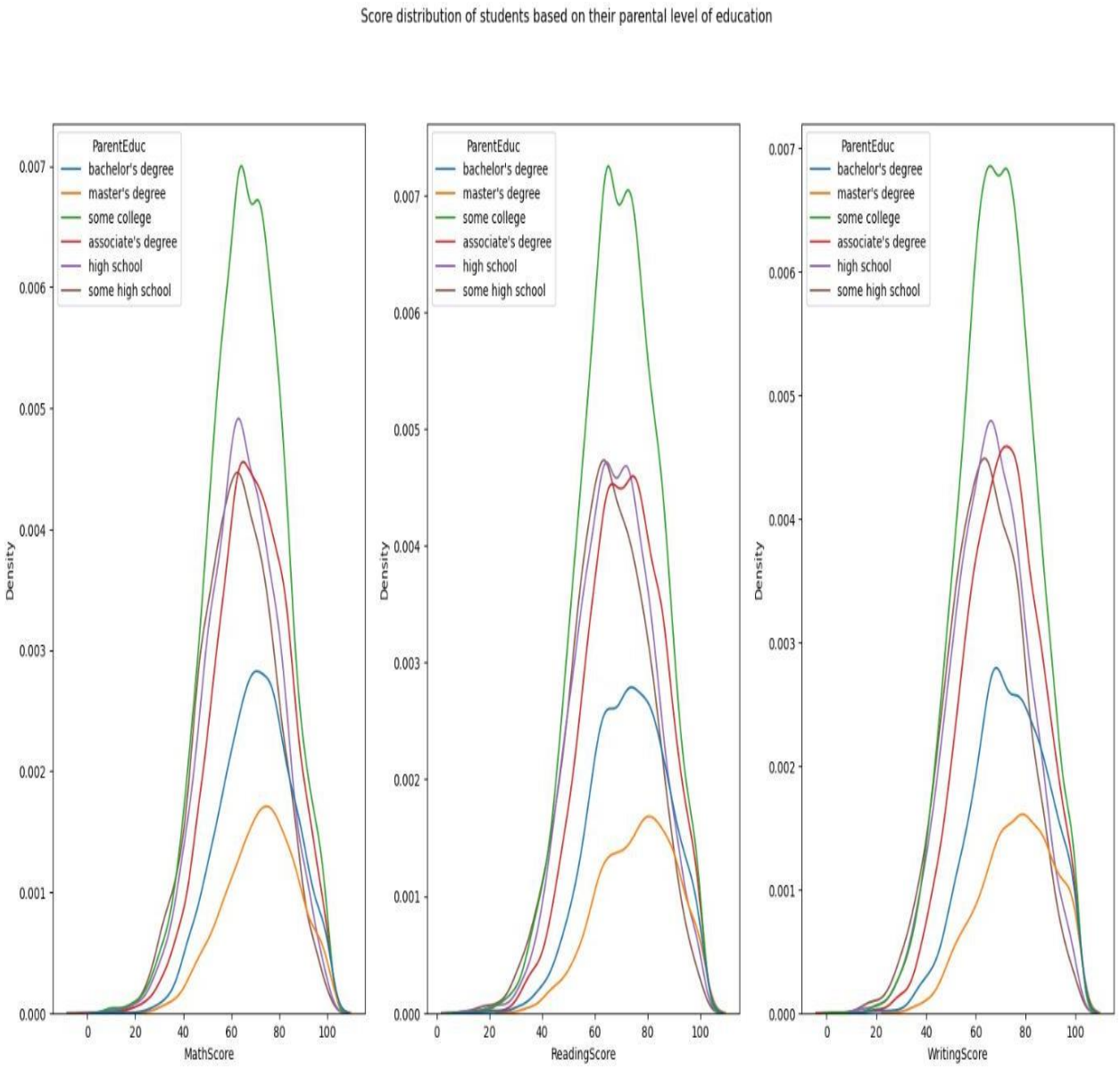


Fig 21:Density plot -This graph showing the number of siblings of the students.

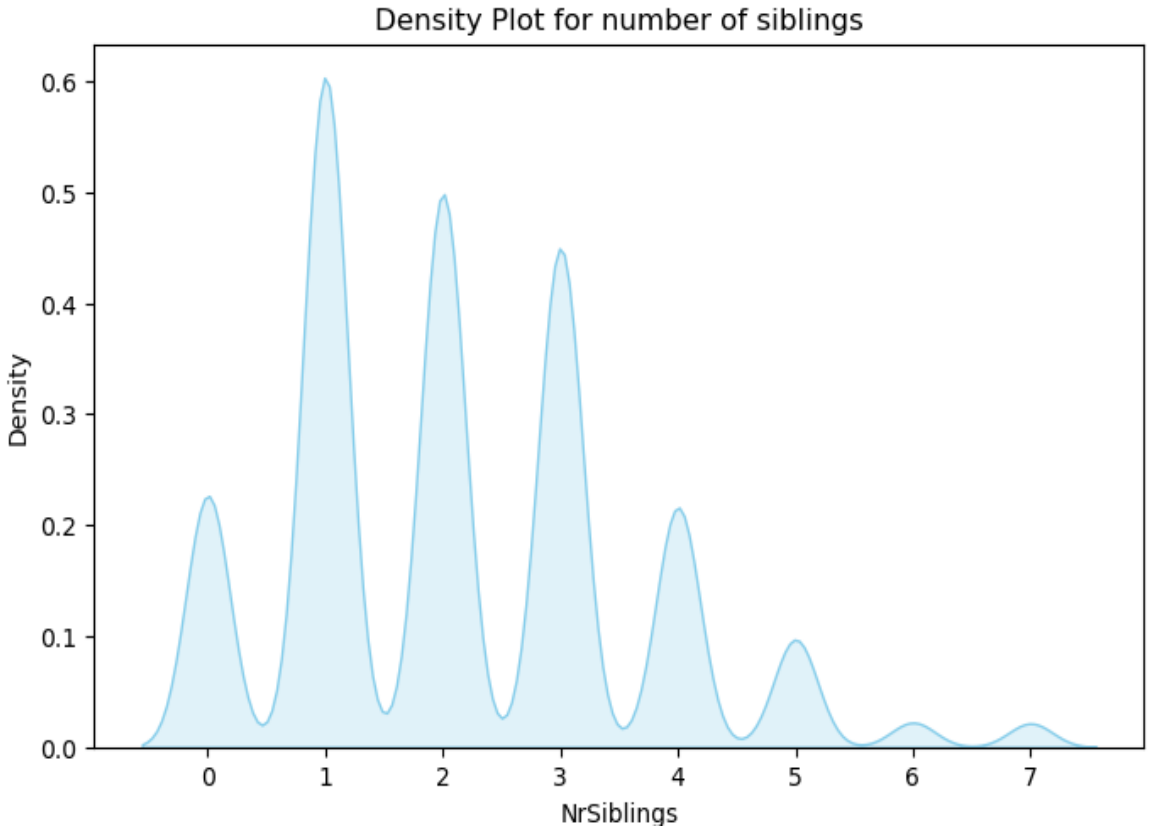


Fig 22: Advance pie chart-this graph shows the percentage of students are passing.

```
# Advanced Pie Chart
plt.figure(figsize=(8, 5))
df1['did_pass'].value_counts().plot(kind='pie', autopct='%1.1f%%', colors=['lightblue', 'lightgreen'], startangle=140)

# Adding labels and title
plt.title('Total no.of students passing')
plt.show()
```

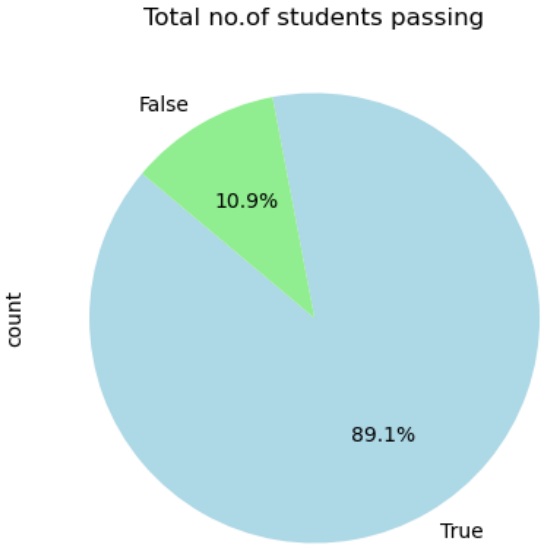


Fig 23: Box plot- This graph shows the relation between mathscore and weekly study hours.
There seems to be a positive trend between study hours and math scores. As the number of weekly study hours increases, the median math score tends to rise.

```
sns.boxplot(data=df1, x='WklyStudyHours', y='MathScore', hue='did_pass', palette='muted')
plt.title('Math Score by Weekly Study Hours')
plt.show()
```

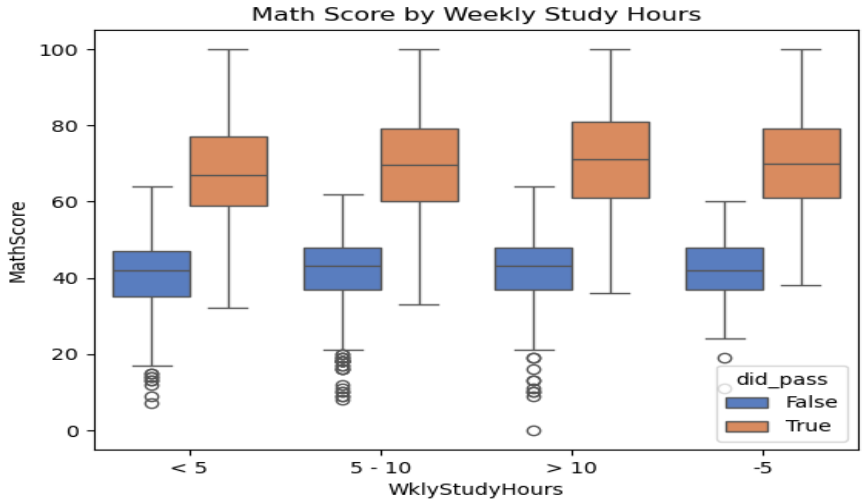


Fig 24: Count plot -This graph shows the relation between the transport means and the pass percentage of the students. The count plot suggests that students who travel privately have a higher likelihood of passing compared to those who use the school bus. However, further analysis with more data and context is needed to draw definitive conclusions about the relationship between transport means and passing rates.

```
sns.countplot(data=df1, x='TransportMeans', hue='did_pass', palette='Set2')
plt.title('Pass Rates by Transport Means')
plt.show()
```

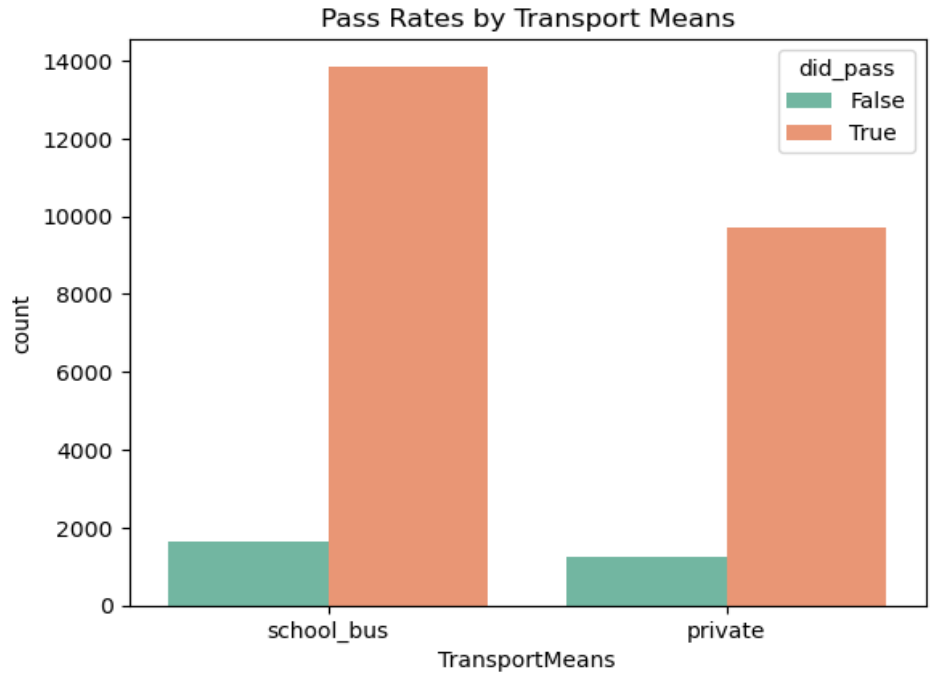


Fig 25: Violin plot – This graph shows the relation between the students who participated sports and their scores.

```
sns.violinplot(data=df1, x='PracticeSport', y='total_score', hue='did_pass', split=True, palette='pastel')
plt.title('Total Scores by Sport Practice Frequency')
plt.show()
```

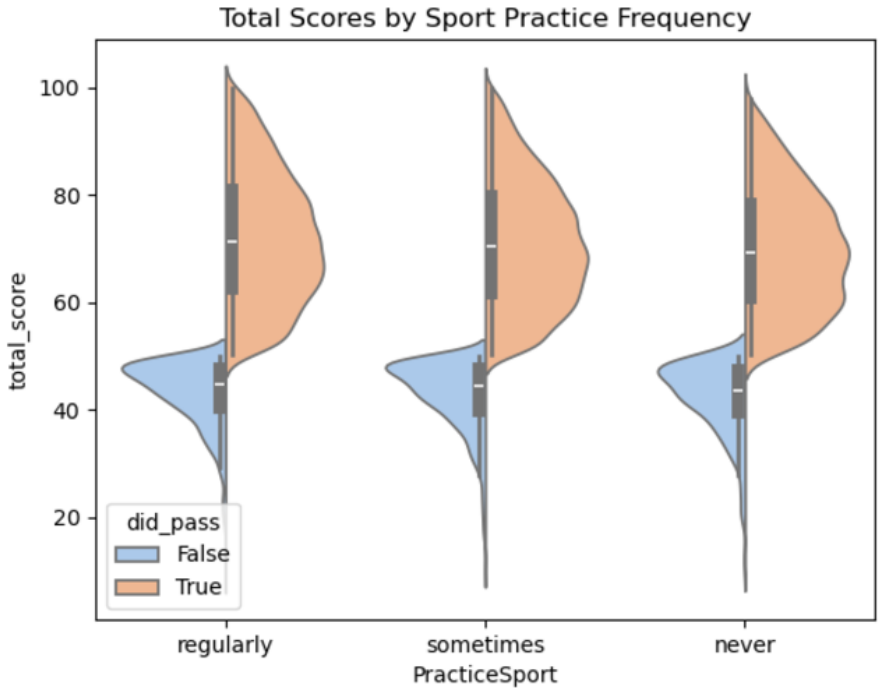
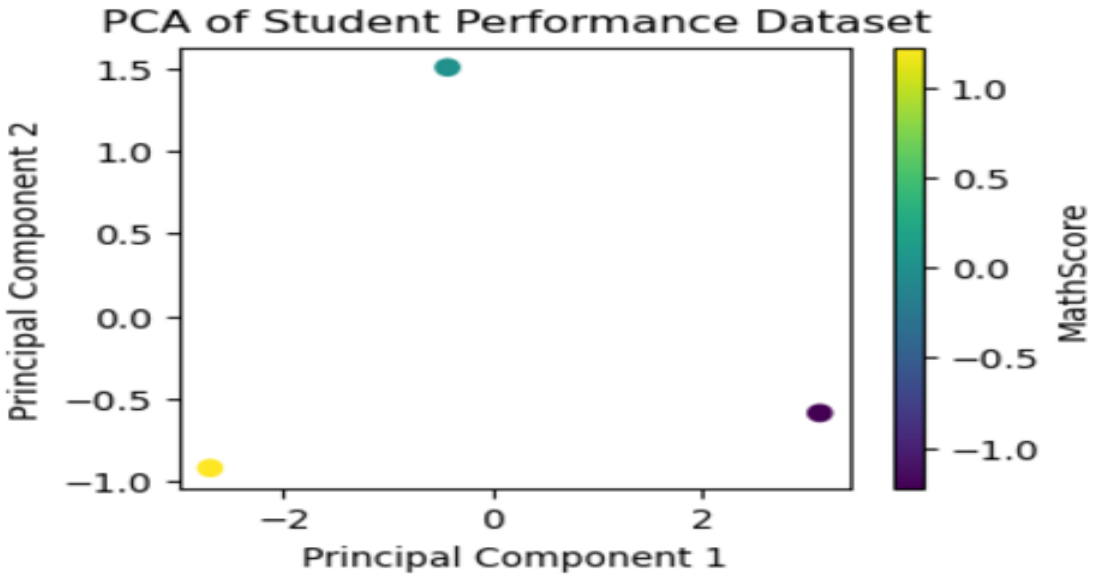


Fig 26: PCA - The color of each dot corresponds to the MathScore of the student. This color-coding helps visualize how the MathScore varies across different points in the PCA space.

```
Explained variance ratio by each principal component:
[0.83328997 0.16671003]

Cumulative explained variance:
[0.83328997 1.        ]
```



CONCLUSION

The analysis of the student performance dataset, which includes parameters such as Gender, EthnicGroup,ParentEduc,LunchType, TestPrep, PracticeSport, NrSiblings, WklyStudyHours, and test scores (MathScore, ReadingScore, WritingScore), provided critical insights into the factors influencing academic success. Using Exploratory Data Analysis (EDA), multivariate analysis, and feature engineering, key relationships between these variables and test scores were identified. Gender analysis revealed that while the dataset had more females, they also performed better overall, possibly due to societal, educational, or biological factors requiring further investigation. Parental education emerged as a significant factor, with students whose parents held a bachelor's or master's degree scoring higher in Math and Reading, although the impact was less pronounced for Writing. Strong positive correlations between Math, Reading, and Writing scores underscored the role of shared cognitive abilities, study habits, and resource access in driving performance. Conversely, variables like NrSiblings showed weak correlations, indicating minimal impact of family size on academic outcomes. Study hours and test preparation were strongly linked to improved performance, particularly in Math, suggesting actionable interventions to encourage these practices. Socioeconomic factors also played a critical role, as students receiving free or reduced lunch, often indicative of lower socioeconomic status, exhibited lower scores, emphasizing the importance of addressing resource disparities. In contrast, the marital status of parents showed minimal direct impact, pointing to the greater influence of parental involvement and support. Challenges such as outliers, particularly in EthnicGroup and ParentEduc, were mitigated through transformations like log scaling, though future work could explore more robust techniques. The findings also highlighted opportunities for predictive modeling using methods like linear regression, decision trees, or random forests to validate relationships and generalize insights to broader datasets. Such models could refine interventions and offer precise strategies for improving student outcomes. Overall, this analysis underscores the multifaceted nature of academic performance, influenced by internal factors like study habits and test preparation and external ones like socioeconomic status and parental education. By considering these elements holistically, educators and policymakers can design data-driven strategies to support student achievement, address inequalities, and foster environments where all students can thrive.

REFERENCES

1. Kaggle - A platform for datasets and data science projects.
https://www.kaggle.com
2. Towards Data Science - Articles on data science techniques and tools.
https://towardsdatascience.com
3. Medium: Analytics Vidhya - Insights on data science, machine learning, AI.
https://medium.com/analytics-vidhya
4. GeeksforGeeks - Tutorials on data science, Python, and machine learning.
https://www.geeksforgeeks.org
5. DataCamp - Learn data science with online courses and projects.
https://www.datacamp.com
6. Coursera - Data science courses from top universities.
https://www.coursera.org
7. Statista - Insights and statistics for e-commerce trends.
https://www.statista.com
8. Stack Overflow - Community discussions on coding and data analysis.
https://stackoverflow.com
9. GitHub - A repository for data science projects and open-source code.
https://github.com
10. Investopedia - Insights into financial and sales data analysis.
https://www.investopedia.com
11. edX - Data science and analytics courses from universities worldwide.
https://www.edx.org
12. Analytics Vidhya - Comprehensive guides on data science projects and tools.
https://www.analyticsvidhya.com
13 KDNuggets - Articles and tutorials on big data, analytics, and machine learning.
https://www.kdnuggets.com

GIT HUB REPOSITORY LINK :

https://github.com/balaji12v/EDAonstudentperformancedata

12219897

PROJECT RELATED QUESTIONS

CA-1

Questions basing on dataset:

1. What dataset did you choose for this project, and why?
   I have chosen Student performance data set, as I was interested in this domain.

2. How did you obtain the dataset, and what is its source?
   I have got it from Kaggle.

3. What are the main features in the dataset?
   The main features in the Student performance dataset include Gender, Ethnic Group, Parent Education, parent marriage status, Practice sports, Reading score, Writing Score and Writing score.

4. What is the shape of the dataset?
   (30148, 15)

5. What techniques did you to identify outliers in the dataset?
   IQR(boxplot)

6. How did you handle outliers in the dataset?
   Outliers in the dataset were handled by using statistical methods such as Z-score or IQR (Interquartile Range) to detect and remove or cap extreme values that could skew the analysis

7. How did you perform feature scaling on the dataset, and why is it necessary?
   Feature scaling was performed using methods like Standardization or Min-Max Scaling to normalize data ranges and ensure that all features contribute equally to the analysis and modeling.

8. What steps did you take to clean the data?
   To clean the data, I handled missing values by imputation or removal, corrected inconsistencies and errors, and removed duplicates to ensure data integrity. Additionally, I standardized data formats and normalized text entries for consistency. These steps helped prepare the dataset for accurate analysis.

9. Did you perform any transformation on the data, such as encoding categorical features? How?
   No

10. How did you handle duplicate records in the dataset?

```
df.drop_duplicates(inplace = True)
```

11. What libraries did you use for data cleaning and manipulation?
    Pandas and Numpy

12. What statistical summary did you generate for the dataset?
    The statistical summary generated for the dataset included measures such as mean, median, standard deviation, minimum, maximum, and quartiles for numerical features, as well as frequency counts for categorical features.

13. How did you identify correlations between features in the dataset?
    Correlations between features were identified using Pearson correlation coefficients for numerical variables and heatmaps to visualize the correlation matrix, revealing the strength and direction of relationships between features.

14. What visualization techniques did you use to represent the correlation between feature?
    Heatmap and scatterplot

15. How did you create histograms for numerical features?

```
plt.figure(figsize = (10,5))
plt.hist(df['NrSiblings'] , bins = 3, color = 'olive', edgecolor = 'red')
plt.xlabel('Siblings')
plt.ylabel('range')
plt.title('Histogram on Number of siblings')
    plt.show()
```

16. How did you visualize outliers in the dataset using box plots?

By using box plots, you can easily identify which data points lie significantly above or below the bulk of the data (outliers), especially for continuous features like student performance scores.

- The box represents the interquartile range (IQR), which is the middle 50% of the data.
- The line in the middle of the box is the median.
- The "whiskers" extend to the smallest and largest values that are not considered outliers.
- Points outside the whiskers are considered potential outliers.

17. How did you ensure your code exceeded 60 lines?
    As I have done data exploring, importing all required libraries, a lot of code is written while cleaning the dataset and exploring the insights and done required data visualization of univariant and bivariant analysis .

18. How many different types of visualizations did you include in the project, and why?
    I have implemented a count plot to count the number of males and females, histogram on the basis of number of siblings so that we can say the number of siblings for the students individually .

19. What is the significance of the key drive variables identified in the project? The key drive variables identified in the project are significant because they have a significant impact on

employee satisfaction and performance.

20. What challenges did you encounter while cleaning and manipulating the data? Challenges encountered while cleaning and manipulating the data include: * Handling missing data * Identifying and handling outliers.

21. How does the data cleaning and visualization process contribute to understanding the dataset better? The data cleaning and visualization process contributes to understanding the dataset better by providing insights into the distribution of the data, relationships between variables, and trends and patterns in the data.

22. How did you ensure there was no plagiarism in your report?
Plagiarism was avoided by ensuring that all code and text were original and not copied from any other source.

23. How did you handle categorical data during EDA?
Encoded categorical data using techniques like one-hot encoding and label encoding.

24. What tools or libraries did you use for data visualization in EDA?
Used Matplotlib, Seaborn and Plotly for data visualization.

25. How did you analyze the data's spread and variance?
Analyzed spread and variance with descriptive statistics and boxplots.

26. What insights did you gain from using pair plots during the analysis?
Pair plots revealed correlations, clusters, and outliers between features.

27. How did you interpret the results of your visualizations?
Interpreted visualizations by identifying trends, distributions, and anomalies.

28. How did the data cleaning process affect the outcome of your EDA?
Data cleaning ensured consistency, improving visualization clarity and analysis accuracy.

29. How did your EDA process assist in selecting features for future model building?
EDA highlighted relevant features and reduced dimensionality for model building.

30. What challenges did you face during the EDA process, and how did you address them?

Faced challenges like messy data and multicollinearity, resolved through cleaning and feature engineering.

12219897

40 questions(ca3)

1. What is the main goal of the analysis on the Student Performance Factors dataset?
• The main goal is to identify which factors affect student performance and to evaluate the accuracy of prediction models.
2. What kind of dataset was used in the analysis?
• The dataset consists of student demographic data, study habits, final grades, and features like parental involvement and education.
3. How was the data preprocessed for the analysis?
• The data was cleaned for missing values, categorical data was encoded, and numerical features were scaled before analysis.
4. What are some key features in the dataset?
• Key features include study time, parental education, gender, age, and final grade.
5. What were the primary goals of exploratory data analysis (EDA)?
• The primary goals were to understand the distribution of data, identify relationships between features, and check for missing values or outliers.
6. What did the initial exploratory data analysis reveal?
• It revealed that the dataset has a fairly even distribution of males and females, and the average study time is between 5 to 10 hours per week.
7. How did the final grades of students distribute?

• Most students had final grades between 10 and 15, with some scoring exceptionally high or low.
8. What correlation was found between study time and final grade?
• A positive correlation was found, indicating that more study time generally leads to higher final grades.
9. How did parental involvement correlate with student performance?
• Parental involvement showed a moderate positive correlation with student performance, suggesting that students with more involved parents tend to perform better.
10. Was there any significant correlation between gender and student performance?
• No significant correlation was found between gender and student performance.
11. What machine learning models were used for predicting student performance?
• Models used include Linear Regression, Random Forest Regression, Decision Tree Regression, and Support Vector Machine (SVM).
12. Which model performed best in predicting final grades?
• The Random Forest Regression model performed best, with an R-squared value of 0.68.
13. What was the R-squared value for the Linear Regression model?
• The R-squared value for Linear Regression was 0.45.
14. What does an R-squared value of 0.68 indicate for the Random Forest model?
• It indicates that the Random Forest model explained 68% of the variance in the final grades.
15. What was the Mean Absolute Error (MAE) for the Random Forest model?
• The MAE for the Random Forest model was 1.5, suggesting it was more accurate than other

models.

16. How did the Decision Tree model perform?

• The Decision Tree model had an R-squared value of 0.60 and an MAE of 1.9, performing well but less accurately than the Random Forest model.

17. How did the SVM model perform in classifying pass/fail students?

• The SVM model achieved 72% accuracy, with a good balance between precision (0.75) and recall (0.70). 18. Why was the Random Forest model chosen as the best performer?

• The Random Forest model handled non-linear relationships and interactions between features better than other models.

19. What does the Mean Absolute Error (MAE) indicate?

• MAE indicates the average error between predicted and actual values. A lower MAE indicates a more accurate model.

20. What is the significance of feature importance in the Random Forest model?

• Feature importance shows which features most strongly affect predictions, with study time, parental involvement, and parental education being the top factors.

21. What features were most important in predicting student performance?

• Study time, parental involvement, and parental education were the most important factors.

22. Why is study time such an important factor in student performance?

• Study time correlates with higher academic performance, as students who dedicate more time to studying tend to achieve better results.

23. How does parental involvement influence student performance?

• Students with more involved parents tend to perform better, as parental support plays a crucial role in a student's education.

24. How does parental education level affect student performance?

• Students with highly educated parents are more likely to perform better academically, as educated parents may offer more support and guidance.

25. Were there any significant findings regarding gender and performance?

• Gender did not significantly impact student performance, suggesting that both male and female students perform similarly.

26. How does the performance of Random Forest compare to Decision Trees?

• Random Forest performed better due to its ability to handle complex relationships and interactions between features, while Decision Trees were more prone to overfitting.

27. What metric did you use to compare the models?

• Models were compared based on R-squared values, Mean Absolute Error (MAE), and cross-validation performance.

28. How do cross-validation results support the models' reliability?

• Cross-validation results showed that the models performed consistently across different data subsets, suggesting good generalization.

29. What role did hyperparameter tuning play in improving model performance?

• Hyperparameter tuning improved the models' performance by optimizing the settings to better fit the data. 30. Why is it important to evaluate multiple models?

• Evaluating multiple models allows us to choose the one that best fits the data and meets performance goals, ensuring the most accurate predictions.

31. How did you interpret the results of the Random Forest model?
• Feature importance plots were used to interpret the results and determine the most influential factors, such as study time and parental involvement.
32. How did visualizations help in understanding model predictions?
• Visualizations like scatter plots of predicted vs. actual grades and feature importance plots helped us better understand how the model was performing.
33. What did the feature importance plot reveal?
• It revealed that study time, parental involvement, and parental education were the most influential features for predicting student performance.
34. How do visualizations contribute to model interpretability?
• Visualizations make it easier to understand the relationship between features and predictions, providing clarity on how the model reaches its conclusions.
35. Did the analysis use any specific visualization tools?
• Yes, tools like Matplotlib and Seaborn were used to create visualizations such as scatter plots and bar charts for feature importance.
36. What were the most significant insights gained from the analysis?
• Key insights included the importance of study time, parental involvement, and parental education in predicting student performance.
37. How can the findings from this analysis be applied to real-world education settings?
• Schools and educators can use these insights to create targeted interventions, such as increasing study time and encouraging parental involvement.
38. How could this analysis be expanded in future work?
• Future work could include additional features like emotional well-being, peer relationships, and teaching quality, which could further improve the model's predictions.
39. What challenges did you face while analyzing the data?
• Challenges included handling missing data, selecting the right features for the models, and tuning the hyperparameters for optimal performance.
40. What is the main takeaway from this analysis?
• The main takeaway is that both individual effort (study time) and external factors (parental involvement) significantly influence student performance, and machine learning can effectively predict academic success based on these factors.