“Unlocking Academic Success:Predicting Student Performance”

Supervised By:
Md. Talib

Submitted by:
Varsha Kumari

CONTENTS

- Introduction
- Dataset Overview
- Exploratory Data Analysis
- Categorical Encoding
- Model Training & Testing
- Conclusion
- Recommendations
- Challenges
- Acknowledgements
- References

INTRODUCTION

OBJECTIVE

- Dentifying Factors Affecting Performance: Analyze various factors such as socio-economic status, demographiccharacteristics, parental involvement, study habits, attendance, extracurricular activities, and teaching methodologies to understand their impact on student performance.
- Understanding Academic Trends: Examine trends in student performance over time, across different subjects, grades, or cohorts to identify patterns, areas of improvement, and potential disparities.
- Predicting Student Outcomes: Develop predictive models to forecast student performance based on historical data and identified factors. These models can help in identifying at-risk students and implementing timely interventions to improve outcomes.
- Evaluating Interventions: Assess the effectiveness of interventions, programs, or policies implemented to improve student performance. Evaluate whether the interventions are achieving their intended outcomes and adjust strategies as needed.

PROBLEM STATEMENT

First, I wanted to list the questions I wanted to answer throughout the project.
I believe 60% success of a project depends on the understanding of the problem.
After going through the dataset thoroughly I came up with the following problems which I wanted to solve. They are:

1. Among the top scorer, are they equally good in math, reading, and writing?
2. Among male and female students who are performing better?
3. What is the distribution of reading, writing and math scores?
4. What ethnicity is standing out?
5. Does the parent's education level have any effect on their performance?
6. What effect does 'test preparation' have on their performance?
7. Does lunch type affect their performance?

DATASET OVERVIEW

Our dataset contains information about high school students' performance, influenced by several factors.It is consist of marks secured in various subjects by students in the United States.

Altogether, our dataset has a thousand rows and eight columns with no missing/null values or duplicate records

Out of eight columns,there are five categorical and three numerical columns.Aditionally,I added two more numerical columns, named as, ’total’ and ’average’.

Brief Description of Variables

Attributes related to demographics:

1.Gender: Gender of the student (e.g., male or female).

2.Race/Ethnicity: Ethnic background or race of the student (e.g., Group A, Group B, etc.).

3. Educational background: Parental Level of Education: Education level of the student's parents or guardians.

4.Lunch: Whether the student receives free/reduced-price lunch or not, which can serve as an indicator of socioeconomic status.

5. Test preparation course: Whether the student has opted for and completed test preparatory course or not.

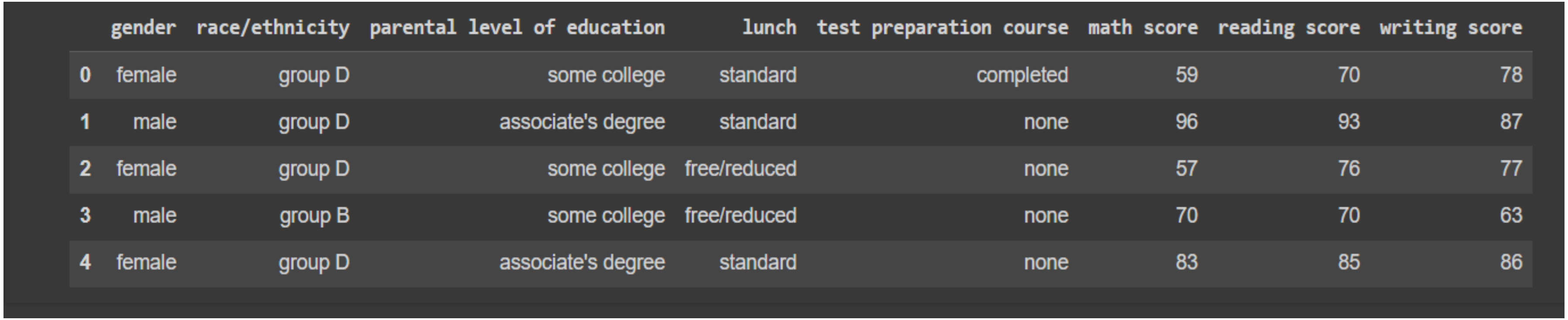
Academic performance:

6.Math Score: The score achieved by the student in the mathematics exam.

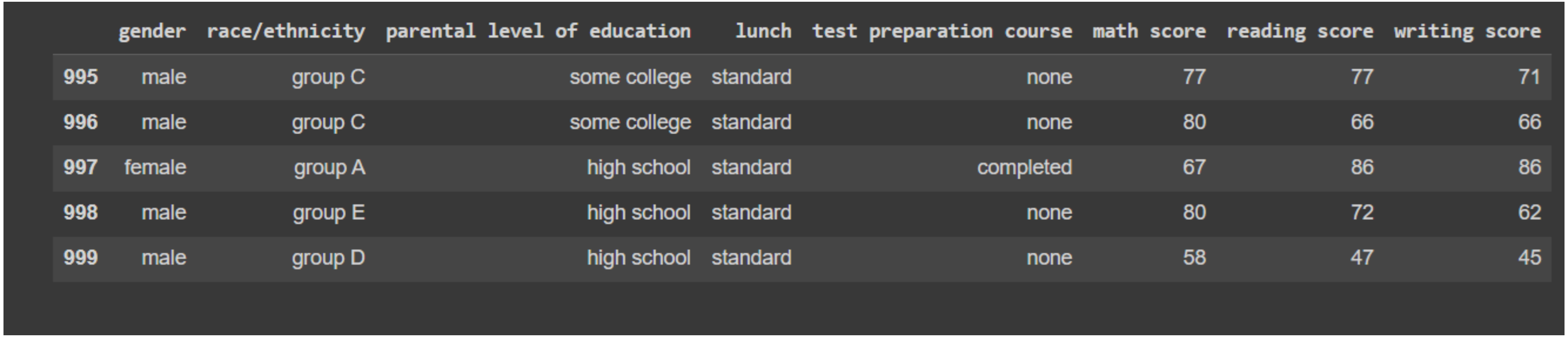
7.Reading Score: The score achieved by the student in the reading exam.

8.Writing Score: The score achieved by the student in the writing exam.

Visualization of dataset structure



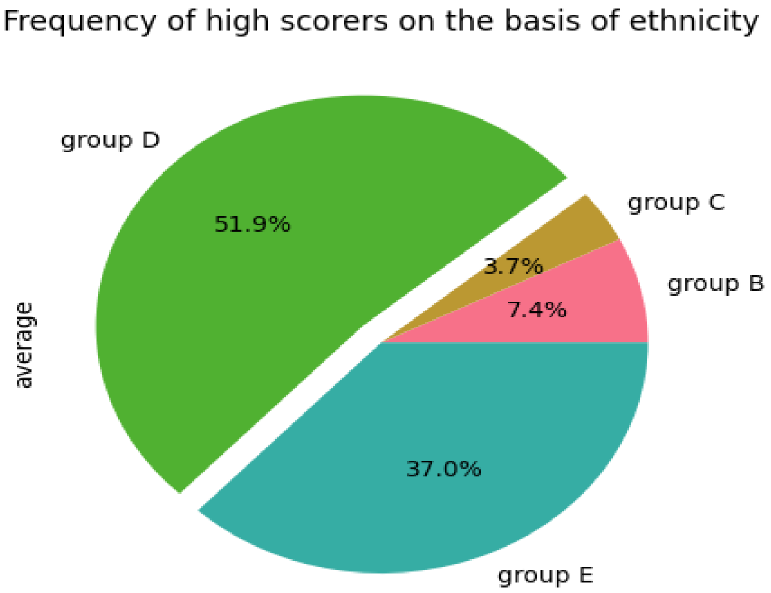
	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group D	some college	standard	completed	59	70	78
1	male	group D	associate's degree	standard	none	96	93	87
2	female	group D	some college	free/reduced	none	57	76	77
3	male	group B	some college	free/reduced	none	70	70	63
4	female	group D	associate's degree	standard	none	83	85	86



	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
995	male	group C	some college	standard	none	77	77	71
996	male	group C	some college	standard	none	80	66	66
997	female	group A	high school	standard	completed	67	86	86
998	male	group E	high school	standard	none	80	72	62
999	male	group D	high school	standard	none	58	47	45

EXPLORATORY DATA ANALYSIS

The insight(s) found from the pie chart below?

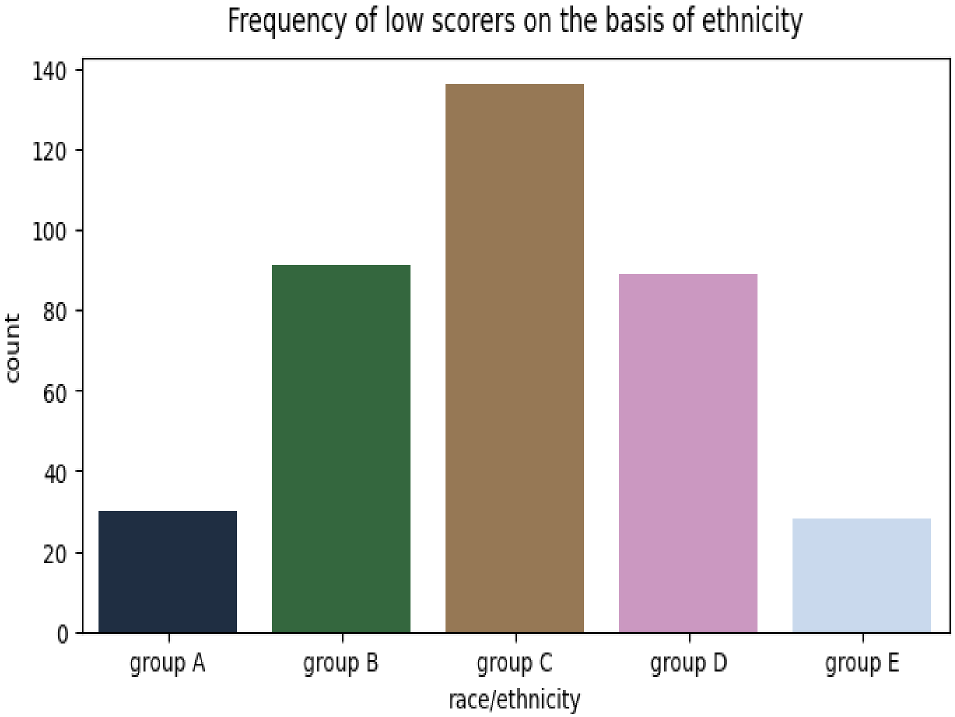


1.More than 50% of the high scorers belong to Group D followed by the high scorers of Group E. Thus we can say that students from Group D and Group E are far more likely to perform great in exams.

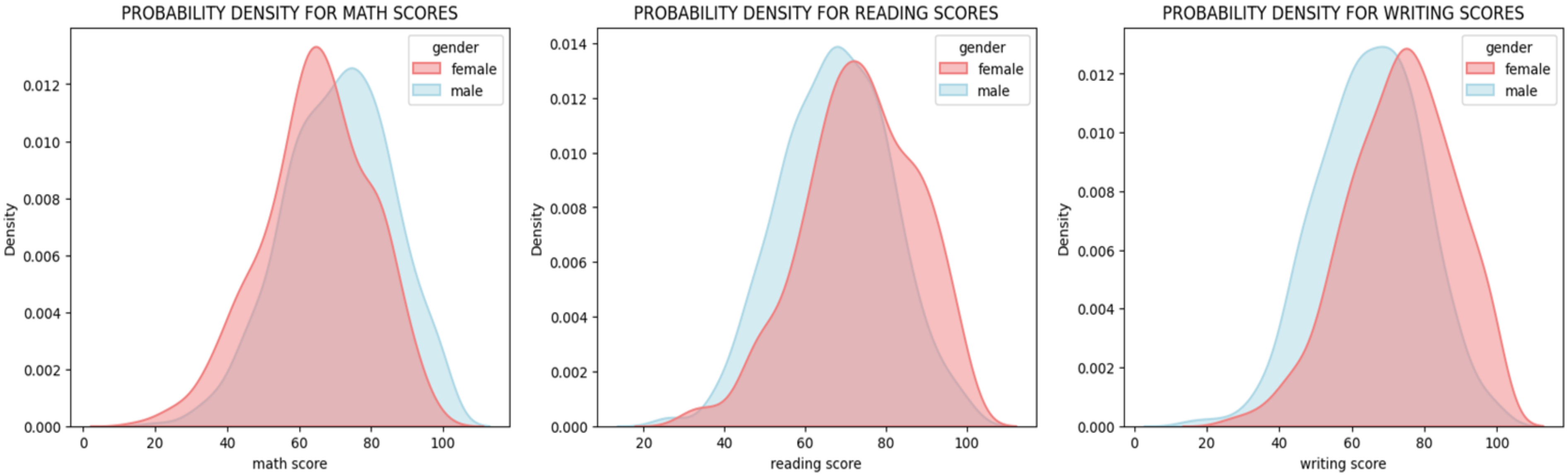
2.Group C has the least number of high scorers and Not a single student from Group A is a high scorer. Thus we can say that students from Group C are far less likely to perform great in exams whereas students from Group A has near to null possibility of performing great in exams.

The insight(s) found from the countplot below?

Greatest number of low scorers belong to Group C.
Thus we can say that students from Group C are far more likely to perform poorly in exams.



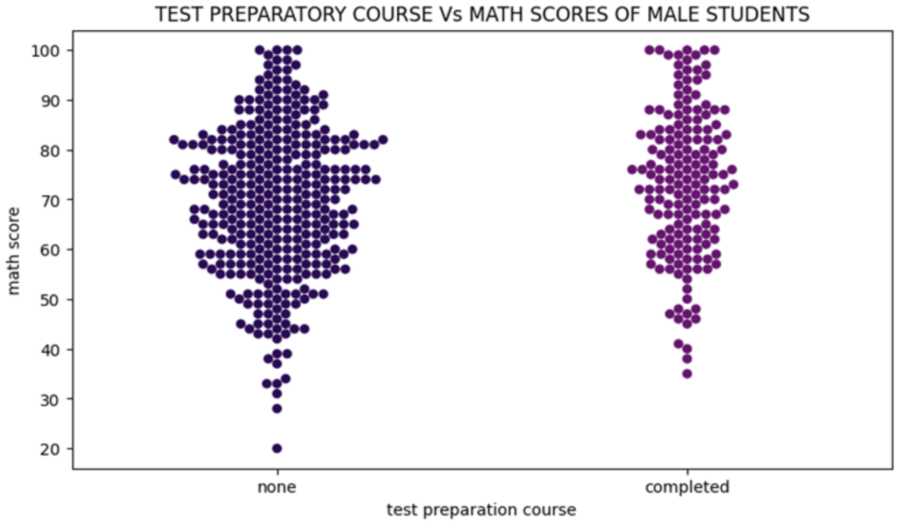
The insight(s) found from the KDEplots below?

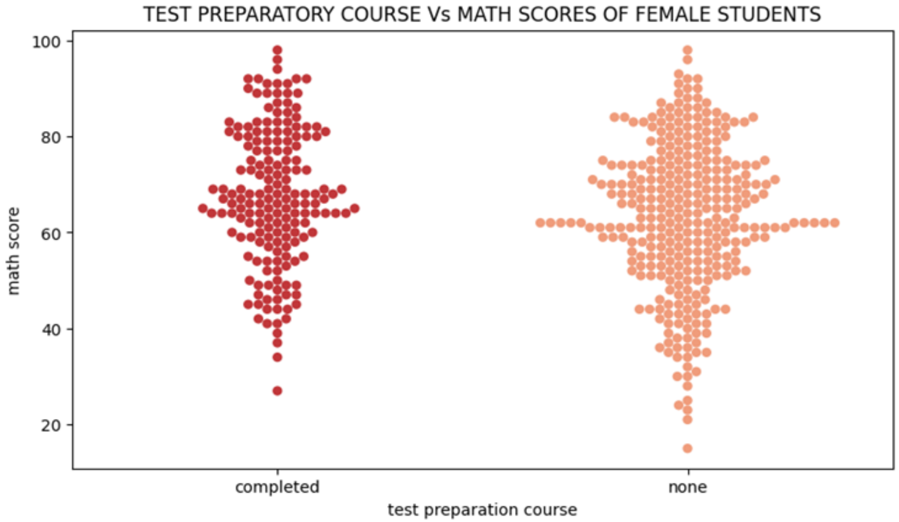


1.Male students have performed great in maths in comparison to the female students.

2.Female students have performed great in reading and writing in comparison to the male students.

The insight(s) found from the swarmplots below?



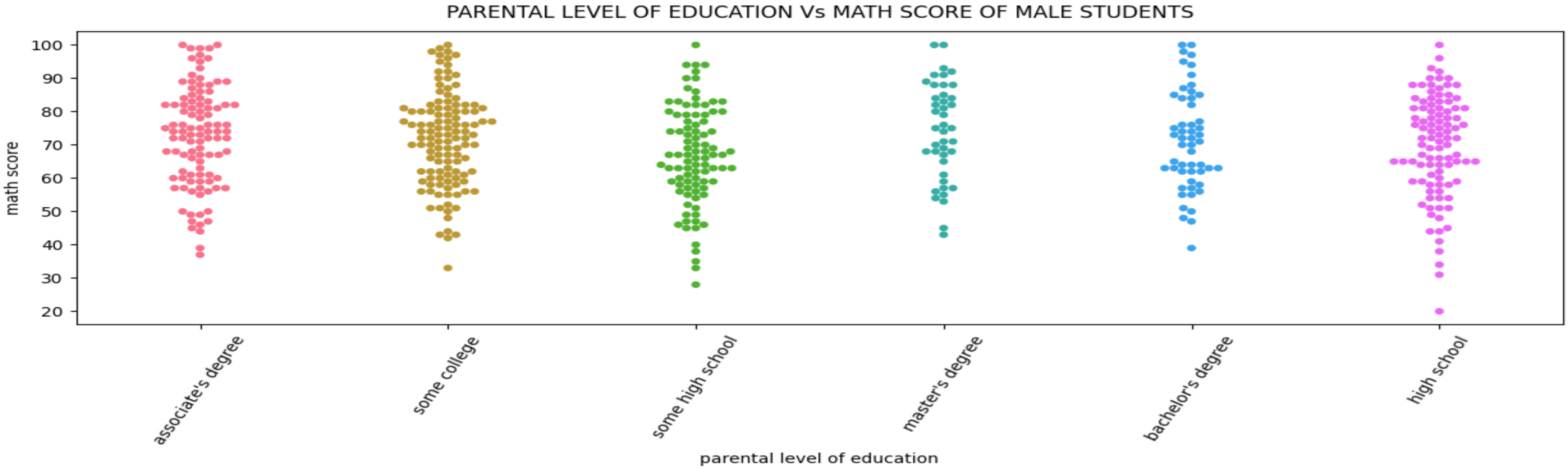


1.Test preparation course does not have much impact on the math score of students.

2.But for the ones who have completed the test preparation course they are less likely to fall in the category of bottom scorers.

3.Few of the students who did not opted for or complete test preparation course are the bottom scorers.

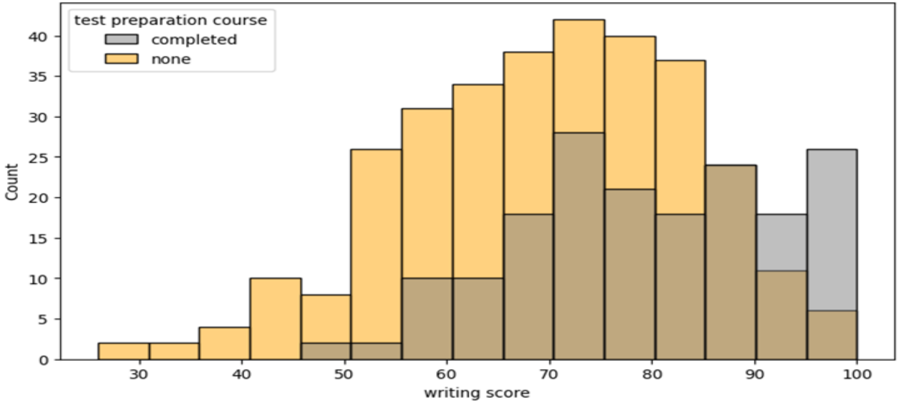
The insight(s) found from the swarmplot below?



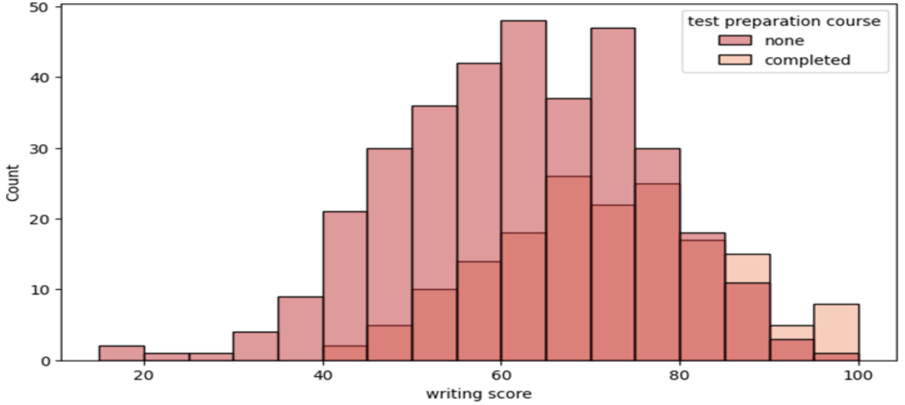
1.Male students(more in number) whose parents hold associate's degree or completed college , have fairly good math scores than others with different parental level of education.
2.Male students(less in number) whose parents hold master's degree ,have scored good, none of them is a bottom scorer in maths.
3.Male students(more in number) whose parents just completed their high school studies , are the bottom scorers in maths,though some of them have scored good marks as well.

The insight(s) found from the histplots below?

WRITING SCORES OF FEMALE STUDENTS ON THE BASIS OF TEST PREPARATORY COURSE

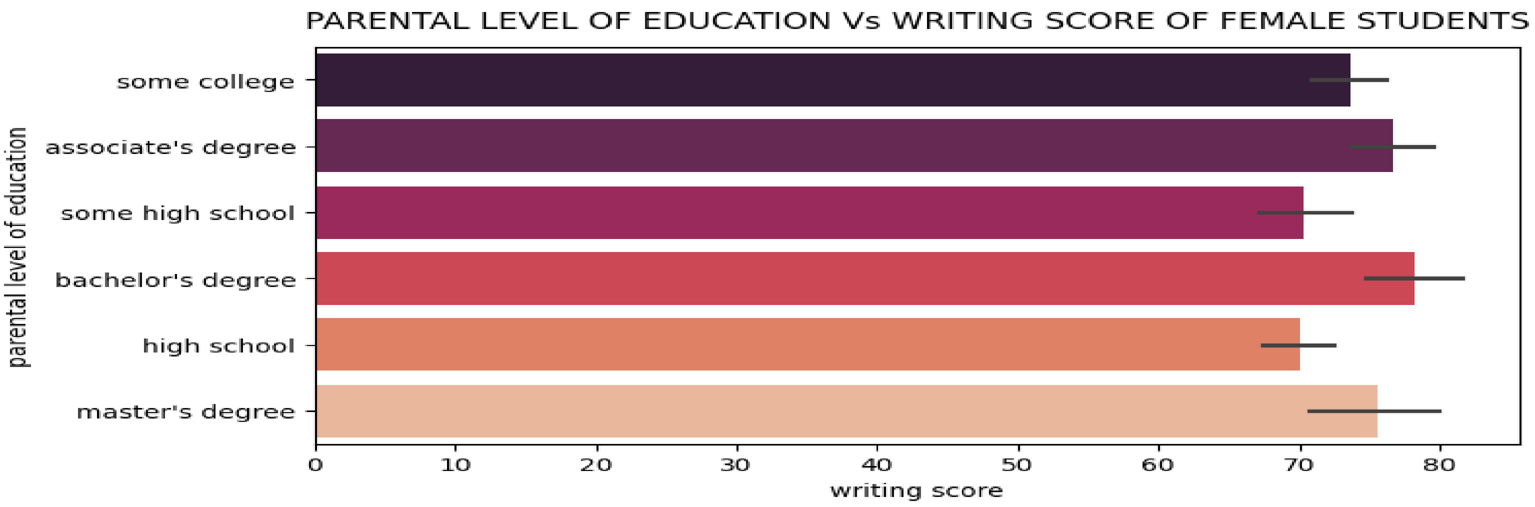


WRITING SCORES OF MALE STUDENTS ON THE BASIS OF TEST PREPARATORY COURSE



1.The students who have completed test preparatory course ,are the top scorers in reading and writing.
2.The students who have not opted for any test preparatory course ,are the mid and bottom scorers in reading and writing.
3.Female students are proportionally higher than the male students when we talk about the top scorers in reading and writing.

The insight(s) found from the barplot below?



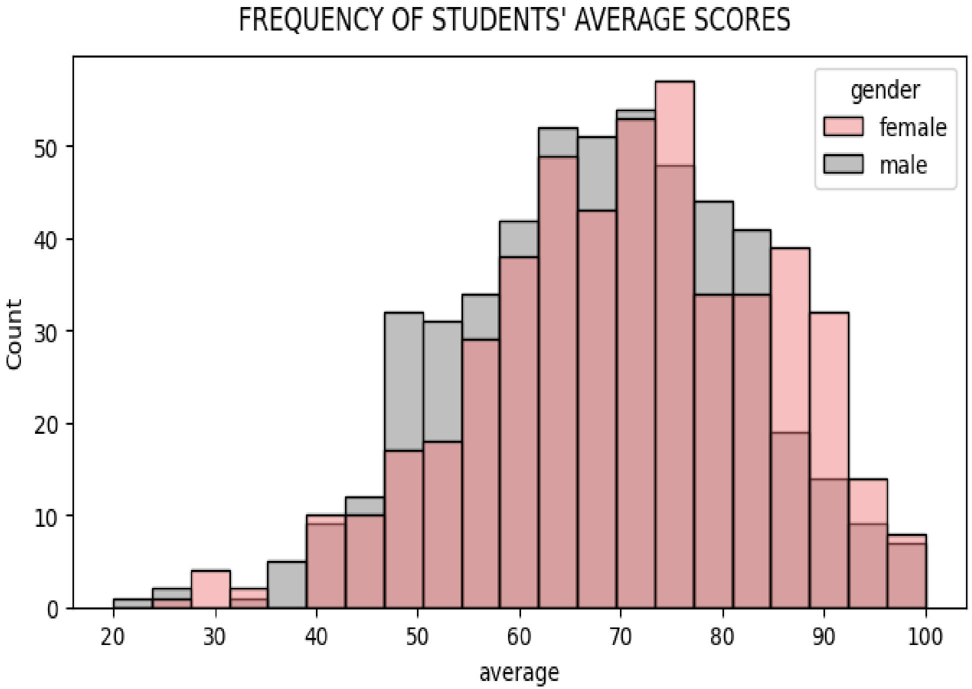
1.Female Students whose parents hold bachelor's, master's or associate's degree , are fairly the top scorers in writing and reading.

2.Female Students whose parents have just completed their high school studies , are the mid/bottom scorers in writing and reading.

The insight(s) found from the histplot below?

1.The count of female students scoring top average scores is higher than the male students.

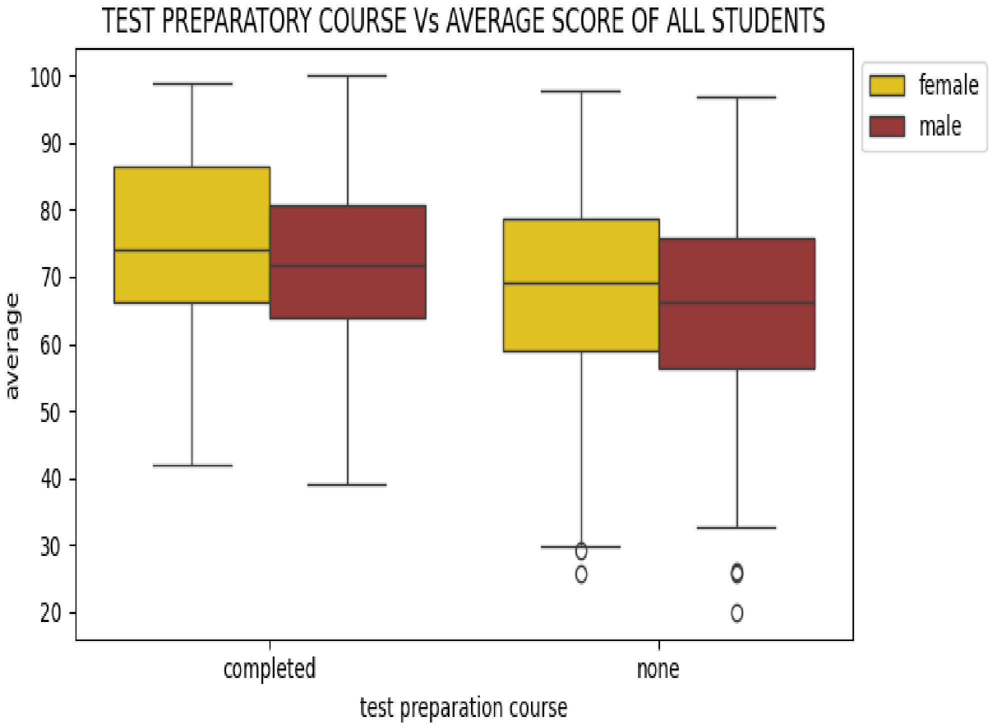
2.The count of male students scoring mid to bottom average scores is higher than female students.



The insight(s) found from the boxplot below?

1.Students who completed test preparation course , have secured higher average scores than the ones who did not opted for the course itself.

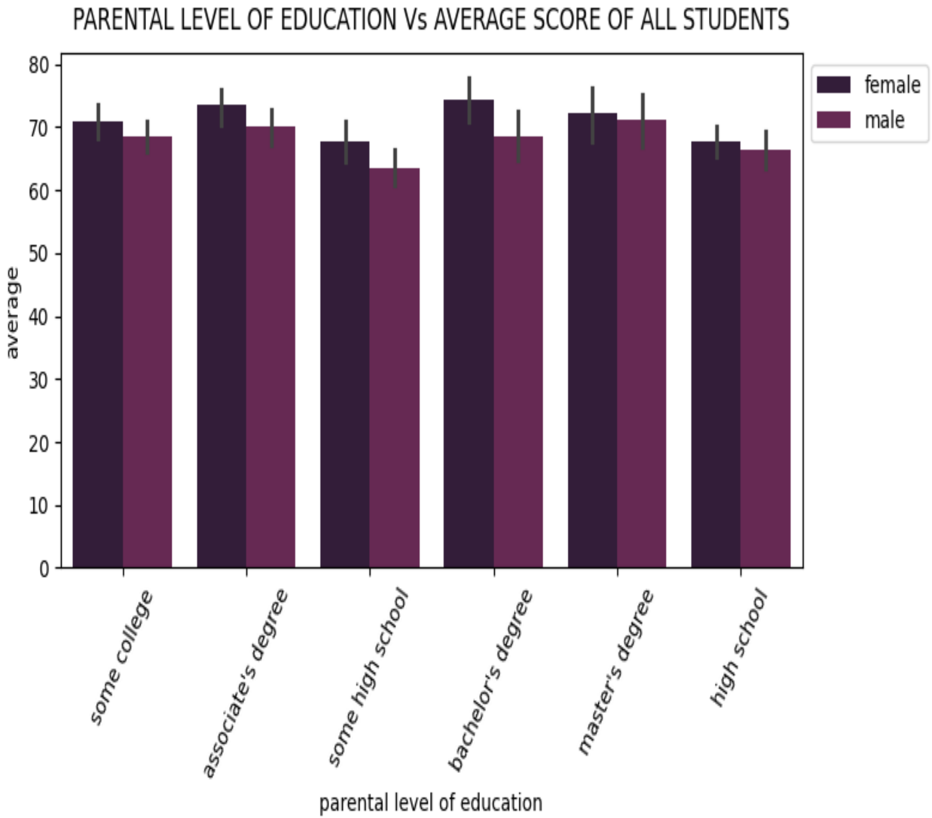
2.Female students outshine male students in scoring higher average scores in both the conditions.



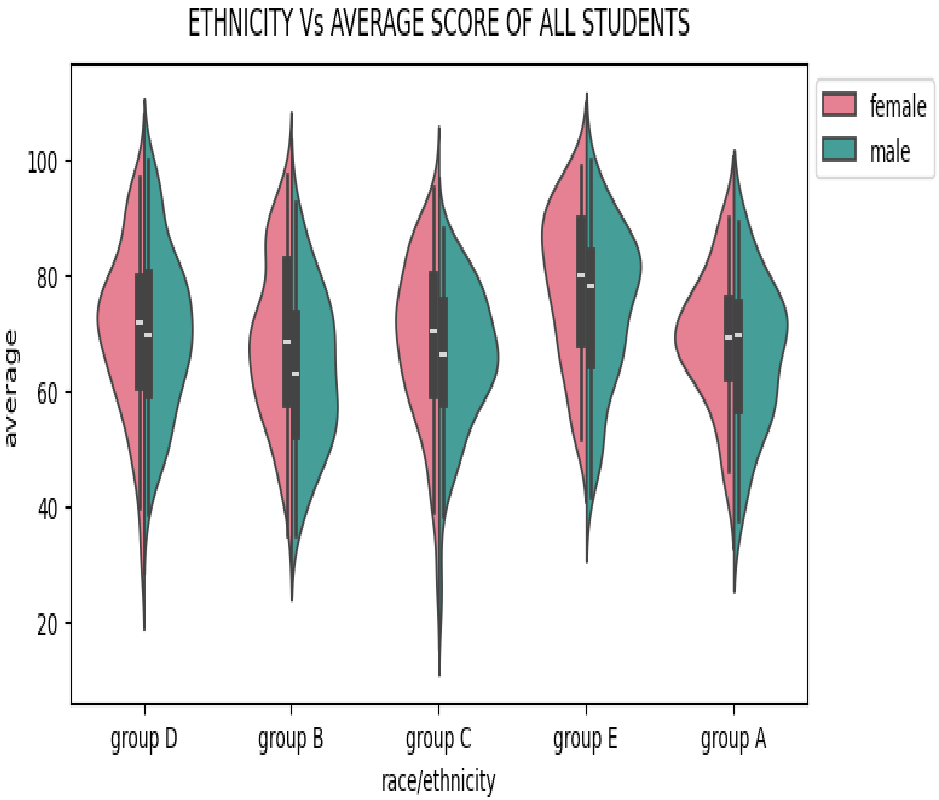
The insight(s) found from the barplot below?

1.Students whose parents hold associate's or bachelor's degree have secured higher average scores than others.

2.Students whose parents have just come from high school , have relatively lower average scores than others.



The insight(s) found from the violinplot below?



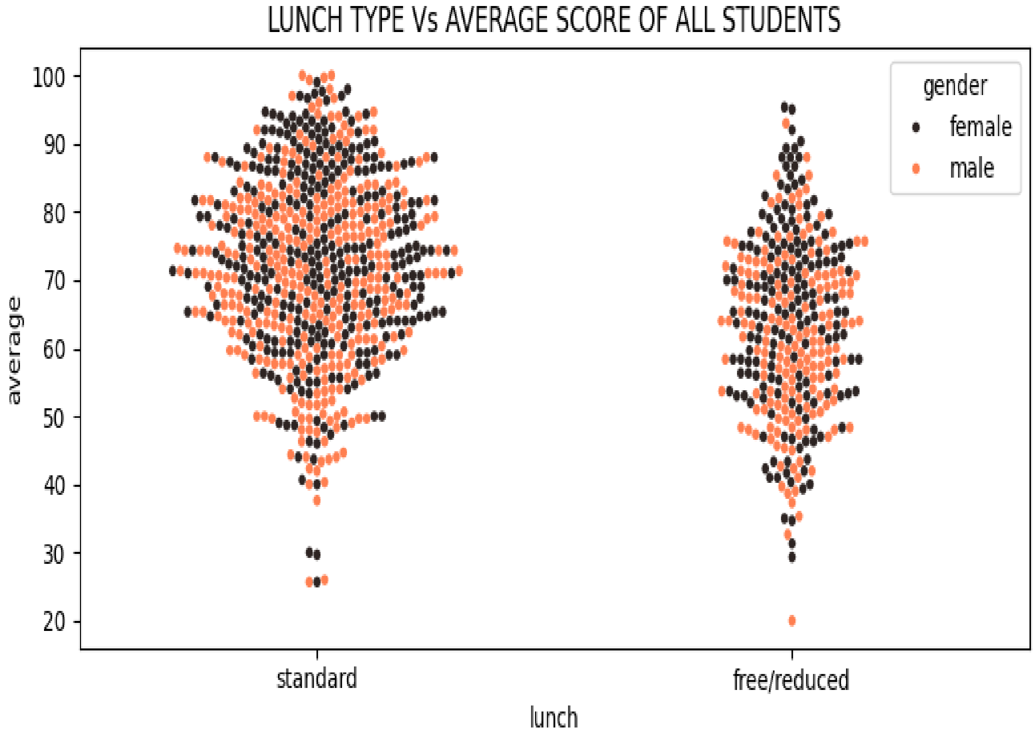
1.Students belonging to Group D and Group E have secured higher average scores than others.

2.Students belonging to Group C have secured relatively lower average scores than others.

The insight(s) found from the swarmplot below?

1.The density of Students who had standard lunch and secured higher average scores is more than the ones who had free/reduced lunch.

2.The density of students taking free lunch and scoring better average scores is little lower than the ones who had standard lunch
we can say that the type of lunch has no significant impact on average scores of students.

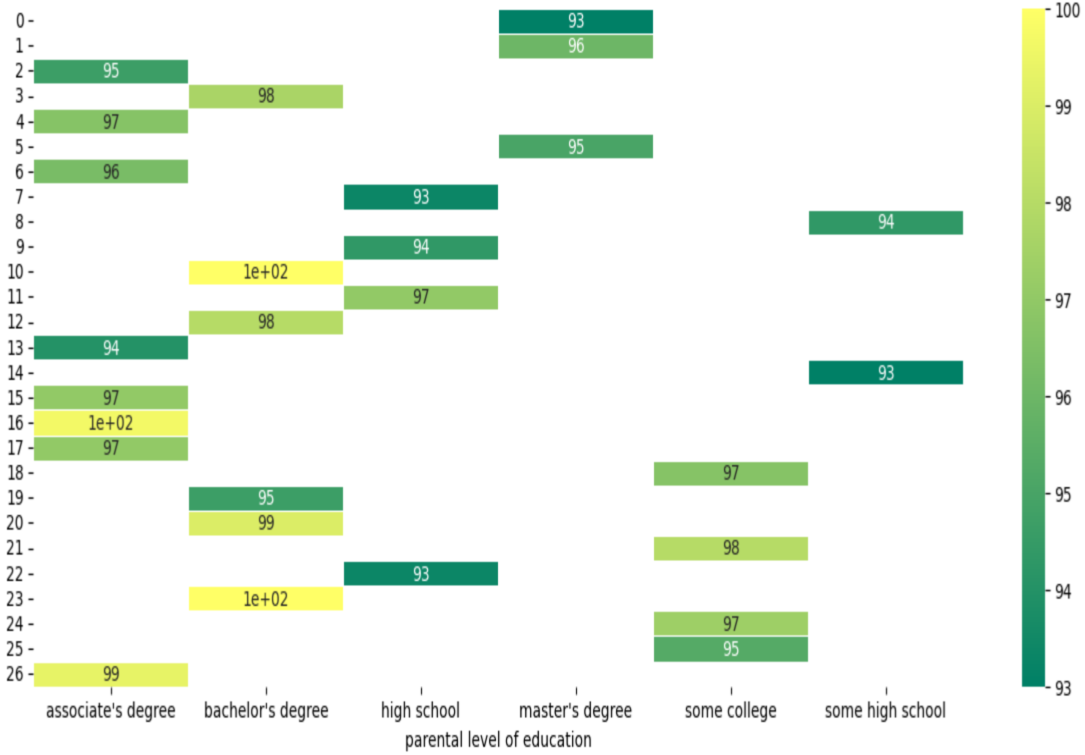


The insight(s) found from the heatmap below?

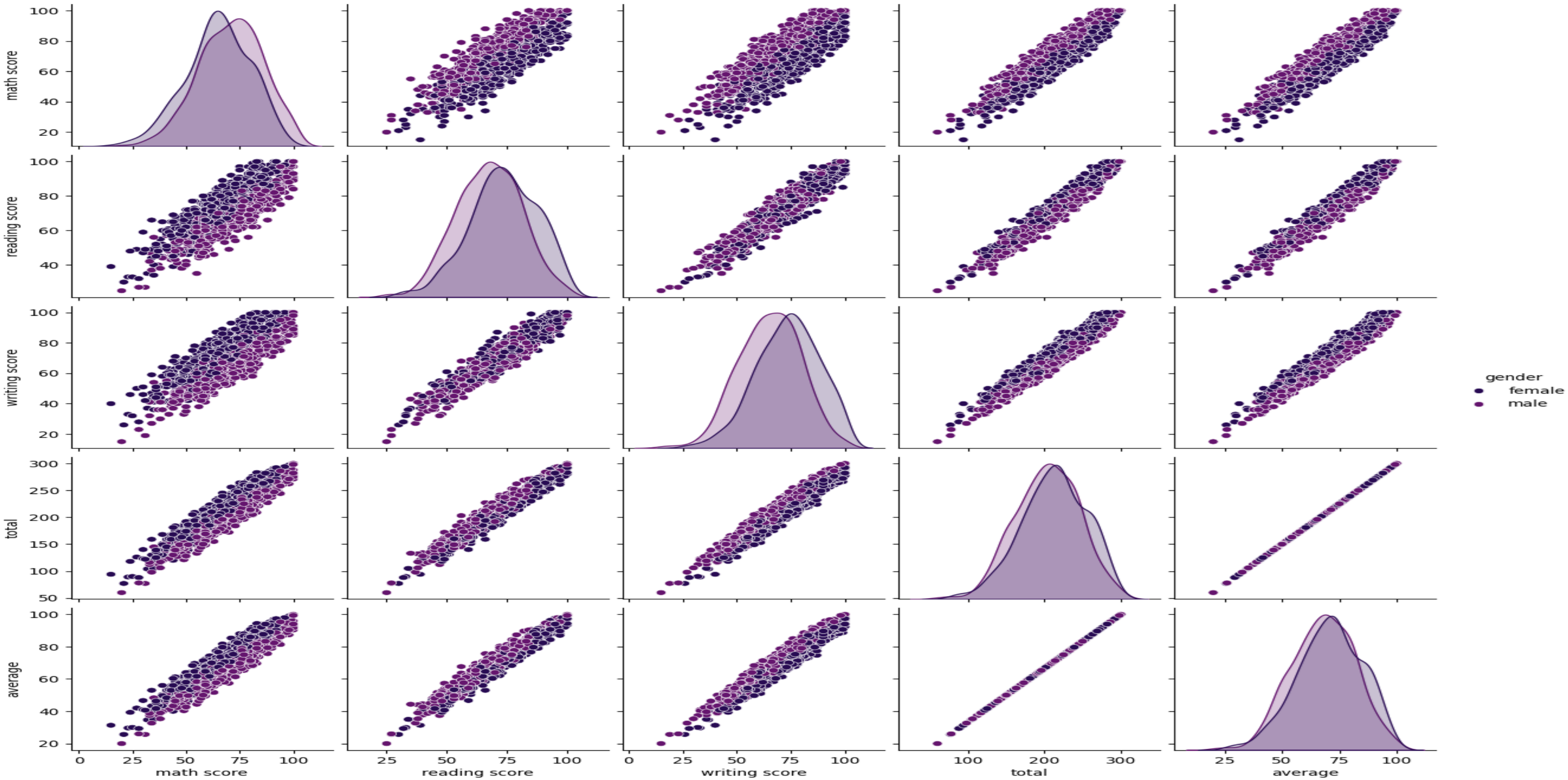
Average Scores of top scorers on the basis of parental level of Education

1.Students whose parents hold bachelor's or associate's degree are more likely to score greater average scores than others.

2.Students whose parents have just come from some high school,are a little more likely to score lesser average scores than others.



Summing up , we have multivariate analysis through the pairplot below.



EDA INSIGHTS

Achievement Disparity: A small proportion of students excelled in all three subjects despite performing well in at least one subject, indicating a notable discrepancy in subject mastery among students.

Gender Disparity: Female students generally outperformed male students overall, though males tended to excel more in mathematics.

Math Proficiency by Gender: Male students demonstrated stronger proficiency in mathematics compared to female students.

Performance by Group: Group D exhibited the highest performance rate among all groups, suggesting potential disparities in resources or educational support across different groups.

Group A's Struggles: Group A showed comparatively lower performance across subjects compared to other groups, indicating potential areas for targeted intervention or support.

Parental Influence: Students with parents holding bachelor’s degrees, tended to perform better academically. This suggests a correlation between parental education level and student achievement, possibly influenced by parental involvement and support in education.

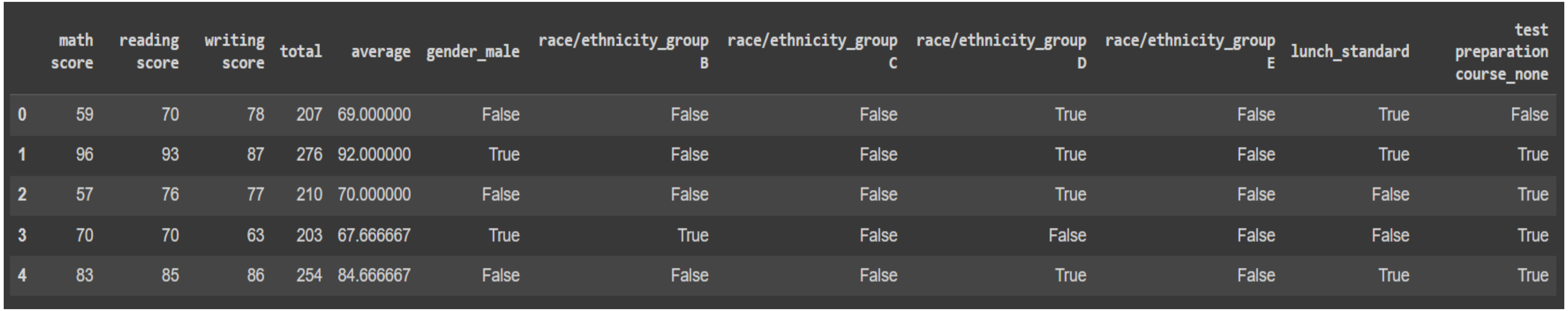
Impact of Test Preparation: Students who completed test preparation courses performed significantly better than those who did not, highlighting the effectiveness of such programs in enhancing academic performance.

Insignificant Role of Lunch Type: The type of lunch consumed by students did not appear to have a substantial impact on their academic performance, suggesting that factors outside of nutrition may have greater influence on academic outcomes.

CATEGORICAL ENCODING

1.I have used ordinal encoding on the column 'parental level of education' because it has an inherent order or ranking among its categories.

2.I have used pandas library functions for encoding on the columns viz., 'gender' , 'race/ethnicity' , 'lunch' , 'test preparation course' because they are nominal variables whose categories represent distinct states or labels with no inherent order.

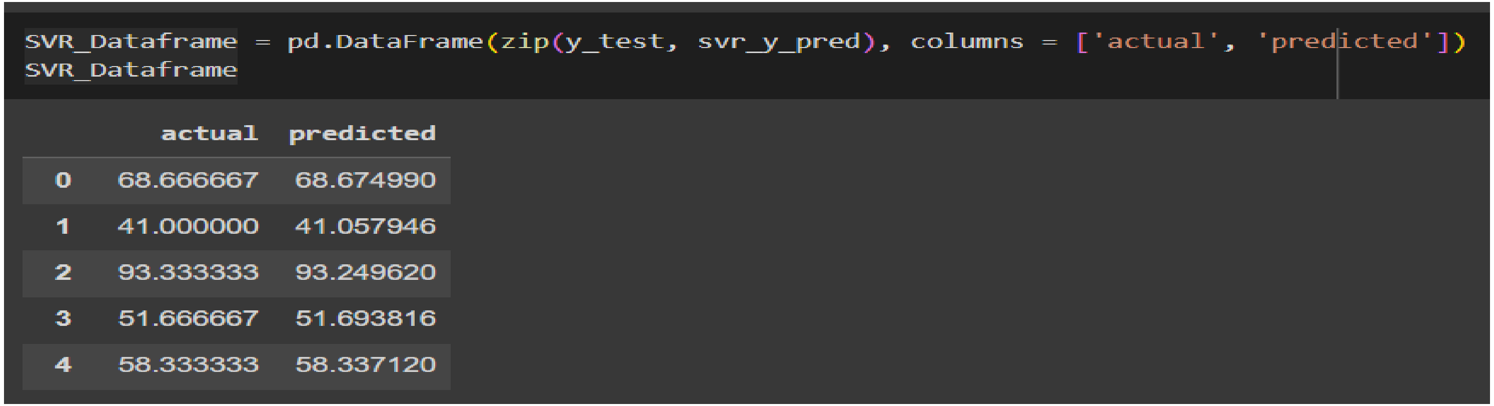


	math score	reading score	writing score	total	average	gender_male	race/ethnicity_group B	race/ethnicity_group C	race/ethnicity_group D	race/ethnicity_group E	lunch_standard	test preparation course_none
0	59	70	78	207	69.000000	False	False	False	True	False	True	False
1	96	93	87	276	92.000000	True	False	False	True	False	True	True
2	57	76	77	210	70.000000	False	False	False	True	False	False	True
3	70	70	63	203	67.666667	True	True	False	False	False	False	True
4	83	85	86	254	84.666667	False	False	False	True	False	True	True

MODEL TRAINING & TESTING

I have used 4:1 as the splitting ratio for training data(=80%) and testing data(=20%) respectively. Since the dataset is not quite large ,thus I choose this ratio as appropriate.

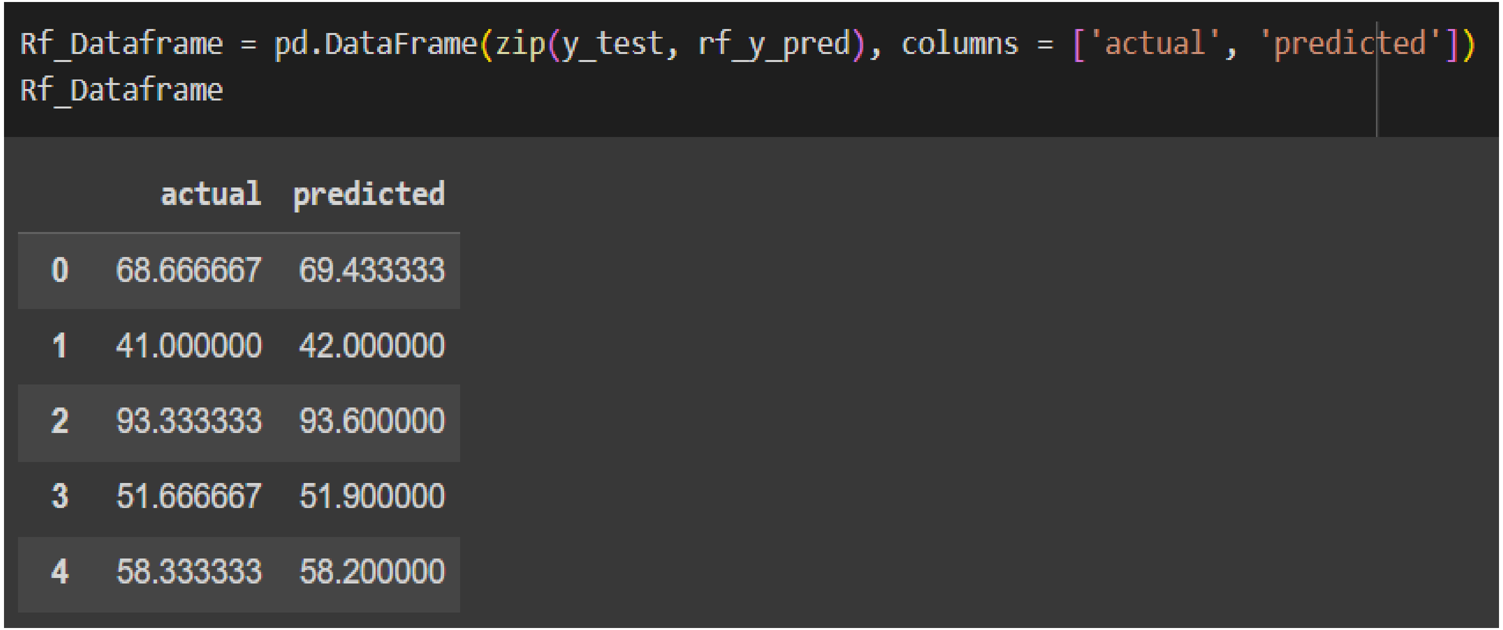
MODEL-1: SUPPORT VECTOR REGRESSION



```
SVR_Dataframe = pd.DataFrame(zip(y_test, svr_y_pred), columns = ['actual', 'predicted'])
SVR_Dataframe
```

```
      actual  predicted
0  68.666667  68.674990
1  41.000000  41.057946
2  93.333333  93.249620
3  51.666667  51.693816
4  58.333333  58.337120
```

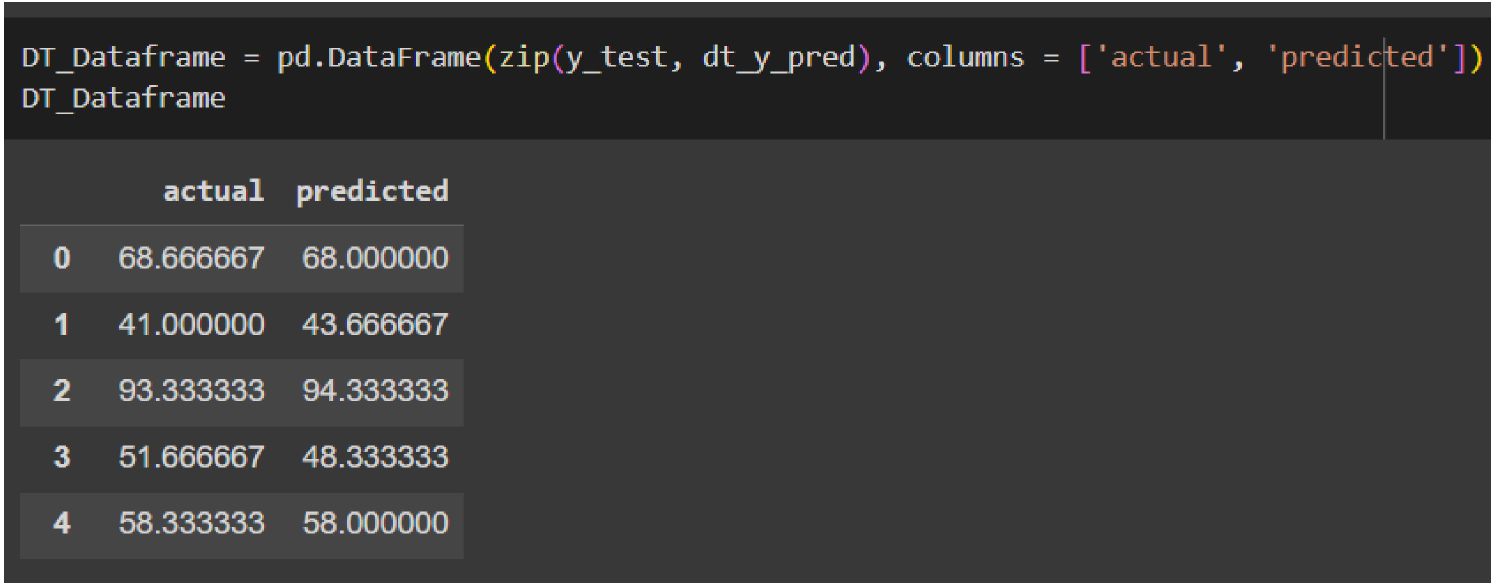
MODEL-2:RANDOM FOREST REGRESSOR



```
Rf_Dataframe = pd.DataFrame(zip(y_test, rf_y_pred), columns = ['actual', 'predicted'])
Rf_Dataframe
```

	actual	predicted
0	68.666667	69.433333
1	41.000000	42.000000
2	93.333333	93.600000
3	51.666667	51.900000
4	58.333333	58.200000

MODEL-3:DECISION TREE REGRESSOR

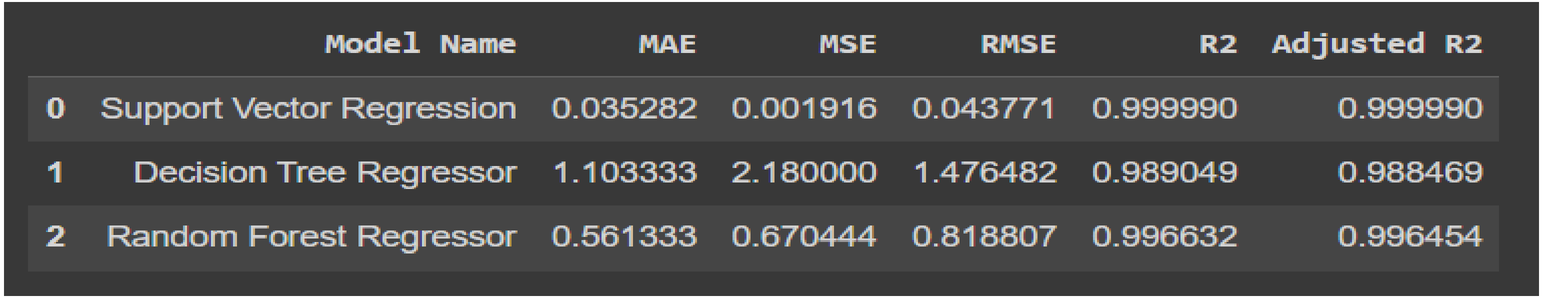


```
DT_Dataframe = pd.DataFrame(zip(y_test, dt_y_pred), columns = ['actual', 'predicted'])
DT_Dataframe
```

	actual	predicted
0	68.666667	68.000000
1	41.000000	43.666667
2	93.333333	94.333333
3	51.666667	48.333333
4	58.333333	58.000000

CONCLUSION

EVALUATION METRICS SCORE CHART



	Model Name	MAE	MSE	RMSE	R2	Adjusted R2
0	Support Vector Regression	0.035282	0.001916	0.043771	0.999990	0.999990
1	Decision Tree Regressor	1.103333	2.180000	1.476482	0.989049	0.988469
2	Random Forest Regressor	0.561333	0.670444	0.818807	0.996632	0.996454

After above observations, I choose the final model , Support Vector Regression with lowest MAE,MSE,RMSE and highest R2 score values that performs exceptionally well in predicting students' performance.

RECOMMENDATIONS

Tailored Academic Support Programs:

Develop customized academic support programs aimed at helping students who excel in one subject to improve their performance in other subjects as well. This could involve personalized tutoring, study groups, or additional resources targeted at addressing specific subject weaknesses.

Gender-Sensitive Educational Initiatives:

Implement gender-sensitive teaching methods and curriculum adaptations to ensure equitable learning opportunities for all students. Provide targeted support for male students in subjects where they may struggle compared to their female counterparts, such as reading ,writing,etc.

Math Enrichment Opportunities:

Offer specialized math enrichment programs or workshops to further develop the skills of male and female students alike. Encourage participation in math competitions or extracurricular activities to foster interest and proficiency in mathematics.

Performance Analysis for Group D Success:

Investigate the factors contributing to the success of Group D and replicate successful strategies across other groups.

Intervention Strategies for Group A:

Implement targeted intervention programs for Group A to address underlying issues impacting their performance. Provide additional academic support, mentoring, or remedial classes to help students in this group improve their academic outcomes.

Parental Engagement Initiatives:

Develop initiatives to engage parents in their children's education, regardless of their level of education. Offer workshops, seminars, or informational sessions to empower parents to support their children's academic success effectively.

Promotion of Test Preparation Courses:

Highlight the benefits of test preparation courses to students and parents, emphasizing their positive impact on academic performance. Offer incentives or subsidies to encourage more students to enroll in these programs and improve their chances of success.

Focus on Holistic Student Support:

While the type of lunch may not directly impact academic performance, prioritize holistic support for students by addressing factors such as nutrition, mental health, and overall well-being. Consider implementing wellness programs or access to counseling services to support students' overall academic success.

CHALLENGES

It was troublesome to get more valuable insights from such a small dataset.

If the dataset had more features like the teaching method of school, hours spent on studying, hours of sleep, etc, then it would be more helpful to make a list of variables that can influence student performance.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following individuals and organizations who have contributed to the success of this project:

Md. Talib(Faculty In-charge): For his guidance, expertise, and unwavering support throughout the duration of this project. His valuable insights and feedback have been instrumental in shaping the direction and outcome of this analysis.

support@cantileverlabs.com: We extend our appreciation to the organization who provided the dataset used in this analysis. Their efforts in collecting and sharing this data have been invaluable to our research.

Chitkara University ,Punjab: We are grateful to our university for providing the necessary resources, facilities, and support for conducting this analysis. Their commitment to fostering a conducive research environment has been instrumental in our endeavors.

This project would not have been possible without the collective efforts and support of these individuals and organizations. We are truly grateful for their contributions.

REFERENCES

- https://colab.research.google.com/drive/1ksmroQtN_KoCeJzpzPgGAbLv0UG_6G_a?usp=sharing#scrollTo=nBYaR4P0HZln
- https://colab.research.google.com/drive/14TP6tNzUT5M0YfgzwTMF_6WBuQkLUgXp?usp=sharing#scrollTo=OpSwoluyz5Wa
- https://colab.research.google.com/drive/1_Mk2NWYBzNxICokEtJFxsf6haq_1YIDA?usp=sharing#scrollTo=TtGLuTLMqIeW
- https://colab.research.google.com/drive/1Hgb530U11qbP4iSf1I2Le0BRpgGUpb16#scrollTo=TtzGW880z5Ul
- https://colab.research.google.com/drive/1VDRu_jTyDkjnhbQXxmF2-4NjihbdSNfb
- https://colab.research.google.com/drive/1LgX4oHckDqRUdLoEpQua2A9Iz8_IgDgT
- https://colab.research.google.com/drive/1D3445tyCOcZmXBIVPX2rnQE-x15mk0r_#scrollTo=StK0mYCHudKC
- https://colab.research.google.com/drive/1We0qcY_28wr1q44boqVsK1-cdRpSL9VY#scrollTo=qAOV1VbS1c0o
- https://colab.research.google.com/drive/1OPd_XzW77o-CSfYLqi92Q2PsmTAwsIo4#scrollTo=EzamsUNXFJAn
- https://medium.com/aiskunks/categorical-data-encoding-techniques-d6296697a40f#:~:text=It%20refers%20to%20the%20process,with%20text%20or%20categorical%20variables.

QUESTIONS & DISCUSSION

Do reach out to us at varsha2514.be22@chitkara.edu.in for additional questions or discussions after the presentation.

We are available to clarify any points or provide further explanation on specific topics in context of this project if needed.

THANK YOU

We express our gratitude to the audience for their attention and participation, and acknowledge the importance of your feedback in furthering the project's objectives.