Project I

Predicting student's performance on math exams using statistical tools and regression algorithms (Supervised Learning Machine)

Table of contents

1. Prior knowledge - data description and goal setting

The dataset used for this project is extracted from kaggle website and can be found here. It is a fictional dataset created by Mr.Royce Kimmons. It contains **30,641 rows and 15 columns**. Each row corresponds to a student enrolled in a public school, detailing his/her personal and socio-economic information along with scores attained from three tests: Mathematics and assessment of reading and writing proficiency.

DATA STRUCTURE

The majority of variables in the dataset are qualitative and just only those connected with exam scores (Math, Reading, Writing) are quantitative. Additionally, there is one variable *att1* which does not convey any meaningful information and serves only to enumerate observations.

MISSING VALUES

Among 14 variables, 9 exhibit varying amounts of missing values. Their count in each column is summarized in the Table 1.

	Variable	Number of NANs
1	EthnicGroup	1840
2	ParentEduc	1845
3	TestPrep	1830
4	ParentMaritalStatus	1190
5	PracticeSport	631
6	IsFirstChild	904
7	NrSiblings	1572
8	TransportMeans	3134
9	WklyStudyHour	955

Table 1 Missing values summary by variable

GOAL

The goal for this project is to predict Math score variable, which takes natural numbers from 0 to 100, using other significant variables. Due to the fact that the dataset provides data (labels) about the target variable, I decided to utilize supervised learning algorithms for regression. I compared the performance of three different algorithms: Linear Regression, Generalized Linear Regression and Support Vector Machine.

An extract of the dataset is presented in the Figure 1.

att1	MathScore	WklyStudyH...	Gender	EthnicGroup	ParentEduc	LunchType	TestPrep	ParentMarit...	PracticeSport	IsFirstChild	NrSiblings	TransportM...	ReadingSc...	WritingScore
0	71	< 5	female	?	bachelor's deg...	standard	none	married	regularly	yes	3	school_bus	71	74
1	69	5-10	female	group C	some college	standard	?	married	sometimes	yes	0	?	90	88
2	87	< 5	female	group B	master's degre...	standard	none	single	sometimes	yes	4	school_bus	93	91
3	45	5-10	male	group A	associate's de...	free/reduced	none	married	never	no	1	?	56	42
4	76	5-10	male	group C	some college	standard	none	married	sometimes	yes	0	school_bus	78	75
5	73	5-10	female	group B	associate's de...	standard	none	married	regularly	yes	1	school_bus	84	79
6	85	5-10	female	group B	some college	standard	completed	widowed	never	no	1	private	93	89
7	41	> 10	male	group B	some college	free/reduced	none	married	sometimes	yes	1	private	43	39
8	65	> 10	male	group D	high school	free/reduced	completed	single	sometimes	no	3	private	64	68
9	37	< 5	female	group B	high school	free/reduced	none	married	regularly	yes	?	private	59	50
10	58	5-10	male	group C	associate's de...	standard	none	?	sometimes	yes	1	private	54	52
11	40	5-10	male	group D	associate's de...	standard	none	divorced	sometimes	yes	1	school_bus	52	43
12	66	5-10	female	group B	high school	standard	none	married	regularly	no	1	private	82	74
13	80	> 10	male	group A	some college	standard	completed	single	sometimes	yes	1	private	73	71
14	48	< 5	female	group A	master's degre...	standard	none	divorced	sometimes	yes	2	private	53	58
15	69	?	female	group C	some high sch...	standard	none	married	sometimes	yes	0	private	75	78
16	88	5-10	male	group C	high school	standard	?	married	sometimes	yes	0	school_bus	89	86

Figure 1 The first few rows of the dataset with special attributes: id (*att1*) and the target (*MathScore*)

Before analyzing in Rapid Miner, I would like to concentrate on understanding how each provided variables might affect the Math Score and how they could be correlated with each other. Below are listed 13 variables with brief potential explanations.

1. Gender – male/female

No one can deny that girls and boys in different countries and cultures have varied educational opportunities, societal expectations or individual interests. Moreover, some studies suggest that males tend to perform better in math-related subjects, while female in language-related subjects.

2. Ethnic Group – group A to E

Different ethnic groups might have varying access to educational resources, cultural attitudes towards education, societal support systems.

3. Parent's Education Background – from some high school to master's degree

Parents with higher education levels might provide more academic support at home and have higher expectations for their children's educational attainment.

4. Lunch Type – standard or free/reduced

Lunch type can reflect the socio-economic status of the student's family. Students from low-income families (receiving free/reduced lunch) may face additional challenges that could affect their academic performance compared to students from more affluent backgrounds.

5. Test preparation completion – completed or none

Completion of test preparation courses might enhance a student's understanding of math concepts and test-taking strategies, leading to higher Math Score.

6. Parent's Marital Status – married / single / widowed / divorced

Family structure can impact child's academic performance through various mechanisms such as parental involvement, emotional support or stability at home.

7. Sports Practice Frequency – never, sometimes, regularly
Regular participation in sports can improve discipline, time management and overall well-being which might positively influence academic performance including Math Score.

8. Is first child – yes / no
Being the first child might mean receiving more attention or pressure from parents regarding academic performance, which could influence Math Score positively or negatively depending on individual circumstances.

9. Number of siblings – from 0 to 7
More siblings might mean less individual attention from parents, potentially affecting academic support and resources available for each child.

10. Transport Means – school bus/private
The mode of transport to school could indirectly reflect the socio-economic status of the family and might influence punctuality, stress levels and overall preparedness for learning.

11. Weekly Study Hours – less than 5h, 5-10, more than 10h
The amount of time spent studying weekly could directly affect Math Score. More study hours generally are correlated with higher scores due to increased practice and understanding of math concepts.

12. Reading Score – from 0 to 100
Strong reading abilities might indicate a student's aptitude for comprehending math problems and instructions, thus potentially positively impacting Math Score.

13. Writing Score – from 0 to 100
Similar to reading, strong writing skills may reflect critical thinking abilities and clarity of expression, which are also beneficial in mathematics.

These variables offer a multifaceted view of a student's background, habits and environment, each potentially playing a role in influencing Math Score individually or in combination with others. Understanding their correlations can provide insights into how various factors contribute to academic performance.

2. Procedure overview

The steps taken in Rapid Minor are as follows:

1. Loading the data

2. Preprocessing and EDA

   a. First, I checked if all the variables had correct types and I discovered one discrepancy. Rapid Miner assumed that *NrSiblings* attribute is numerical, however, in reality, it is categorical (classes from 0 to 7). Using "*Numerical to polynomial*" operator, I changed its type to qualitative.
   b. Then, I verified correctness of classes / values in every single variable. I noticed an error in the *WklyStudyHour* attribute, where instead of 10-May should be 5-10, indicating that student has dedicated from 5 to 10 hours per week for study. The correct values for that variable are: <5, 5-10, >10.
   c. I set the role for the *att1* attribute as an id and for *MathScore* as a label since it is our target variable
   d. Dealing with missing values – I tried two approaches. In the first one, I removed all the missing values from each column; in the second one, I replaced them with the average/mode. Ultimately, I found that the first approach provides better results and consequently, I eliminated NANs, resulting in reduced dataset from 30,641 rows to 19,242. Despite removing a considerable number of observations, the dataset remains sizable.
   e. Using the "*statistics*" operator, I extracted basic statistics related to every single attribute and performed EDA 1D and 2D. Apart from that, I determined the correlation between some attributes and I found out that *Reading Score* and *Writing Score* are highly correlated. To address multicollinearity, I removed the *Writing Score*.
   f. Then, I changed the type of some variables, by converting all non-numerical (nominal) attributes to numerical.
   g. To ensure that all variables have the same weights /scale, I normalized their values.

3. To ensure randomness, I shuffled the entire dataset using local random seed of 1992.

4. I split the dataset into a training set with 15,394 (around 80%) observations and a test set with the remaining 3,848 (around 20%) observations.

5. Using the training set, I performed cross - validation and created model.

6. I fitted the model with unseen data (test data) and compared the performance of the three models.

7. Last but not least step, was to change the name of newly added by default column, which contained predicted values for *Math Score*. What is more, I created a new column named 'residuals', which shows the difference between the true/real values and the predicted ones.

8. Finally, I verified regression assumptions.

Exploratory analysis, model comparison and verification of regression assumptions are detailed described in the next sections.

PROCESS ILLUSTRATION

Figure 2 outlines the complete process, while Figure 3 focuses on the subprocess involving Data Cleansing and Exploratory Data Analysis. Figure 4 showcases content related to Cross Validation.

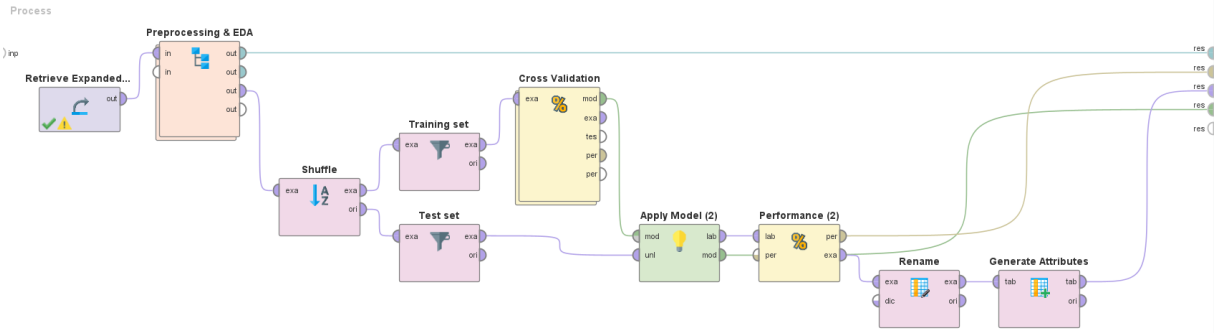


Figure 2 Overall Process

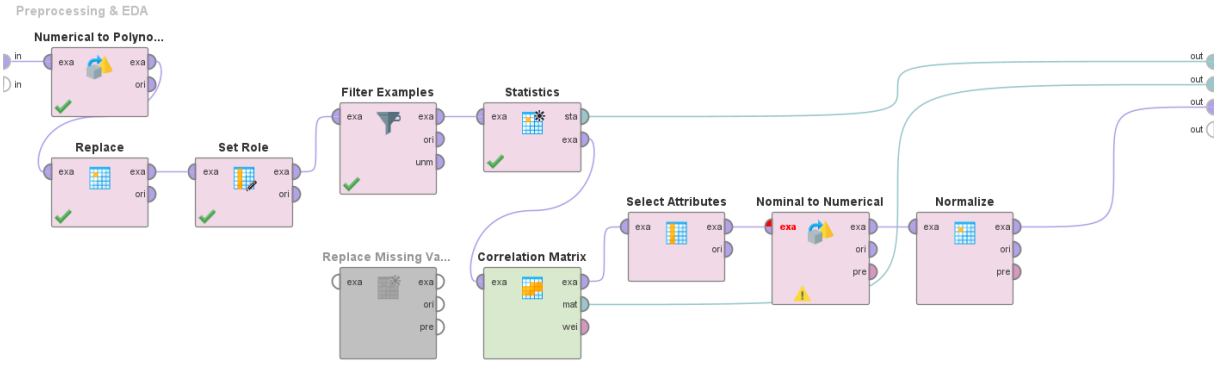


Figure 3 Data cleaning and Exploratory Data Analysis

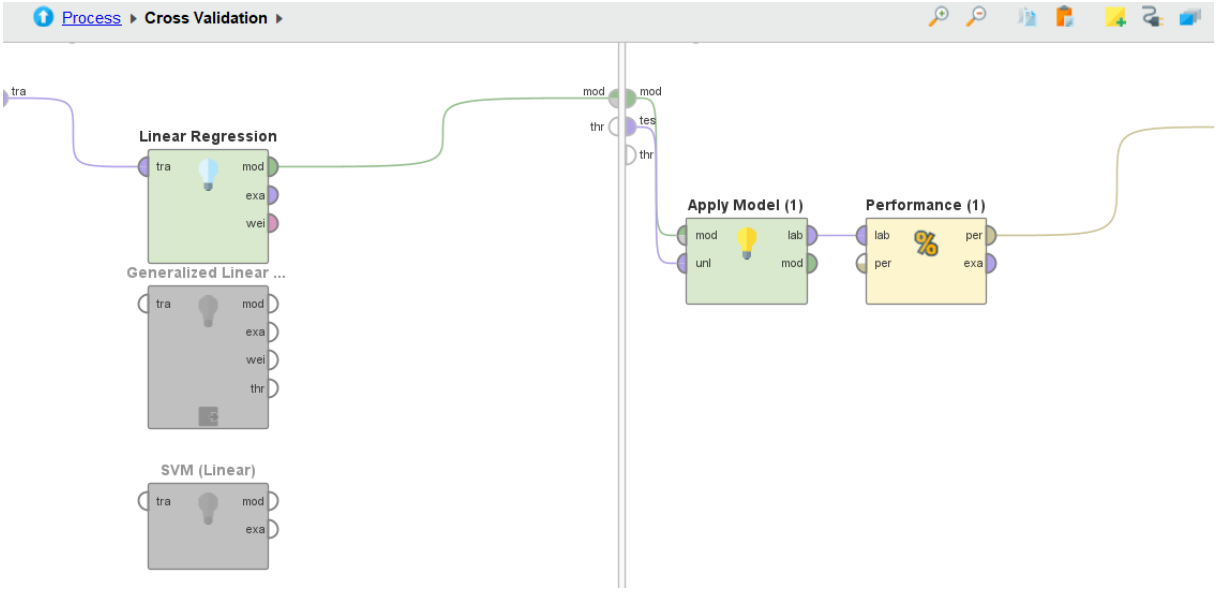


Figure 4 Cross Validation details

3. Exploratory analysis

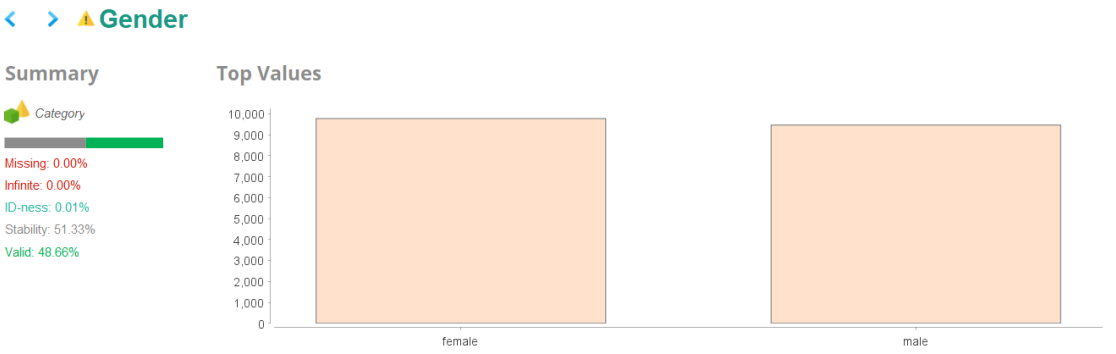
This section delves into the initial exploration of the dataset to understand its main characteristics, uncover patterns and trends. The first part focuses on one-dimensional analysis, where each variable was examined individually. The second part covers two-dimensional analysis and includes the discovery of relationships between pairs of variables. Table 2 below outlines the approaches taken for numerical and categorical variables across both univariate and bivariate contexts.

	univariate (EDA 1D)		**bivariate (EDA 2D)**	
	numerical	categorical	numerical - numerical	numerical - categorical
statistics	minimum, maximum, average, standard deviation	absolute frequency (counts), relative frequency	correlation (Rapid miner also calculates Pearson's coefficient for categorical variables with 2 classes (binominals))	
visualization	histogram	bar plot	scatter plot	box plot

Table 2 Univariate vs. bivariate analysis for both numerical and categorical variables

3.1 One – dimensional exploratory data analysis

1. Gender – male/female



2 Distinct Values:

Value	Count	Percentage
female	9,775	50.80%
male	9,468	49.20%

The gender distribution among students reveals a nearly equal split, with slightly more females than males.

2. Ethnic Group – group A to E

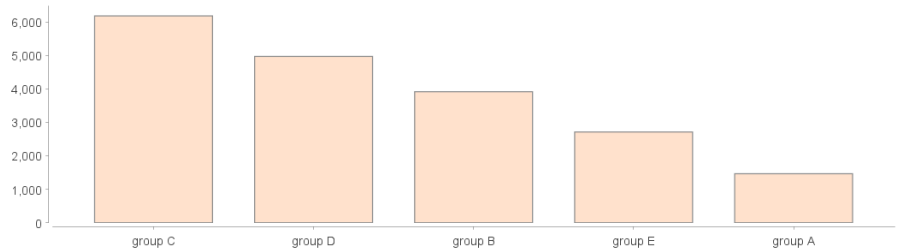
EthnicGroup

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 32.99%
Valid: 66.99%

Top Values



5 Distinct Values:

Value	Count	Percentage
group C	6,181	32.12%
group D	4,970	25.83%
group B	3,915	20.35%
group E	2,712	14.09%
group A	1,465	7.61%

Among all the ethnic groups, group C appears most often (around 6,000 observations), while group A the least (around 1,500 observations).

3. Parent's Education Background – from some high school to master's degree

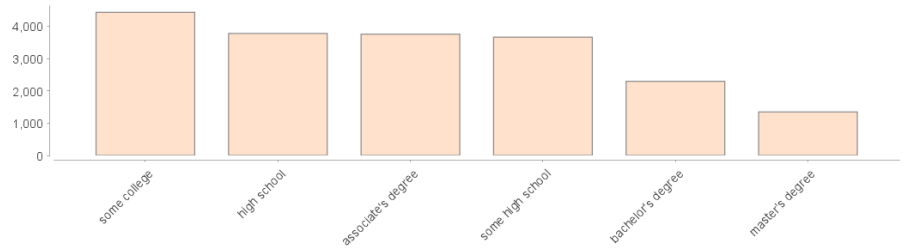
ParentEduc

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 23.05%
Valid: 76.91%

Top Values



6 Distinct Values:

Value	Count	Percentage
some college	4,425	23.00%
high school	3,773	19.61%
associate's degree	3,749	19.48%
some high school	3,658	19.01%
bachelor's degree	2,293	11.92%
master's degree	1,345	6.99%

The education background of parents among the students is diverse. The largest portion, equals to 23%, have completed some college, followed by those with a some high school, high school and associate's degree, each representing around 19% of the total. Only approximately 12% of parents have attained a bachelor's degree and just 7% a master's degree. This suggests that a considerable proportion of parents do not have higher education qualifications, which can influence the support

and resources available to students at home and potentially impact their academic performance and aspirations.

4. Lunch Type – standard or free/reduced

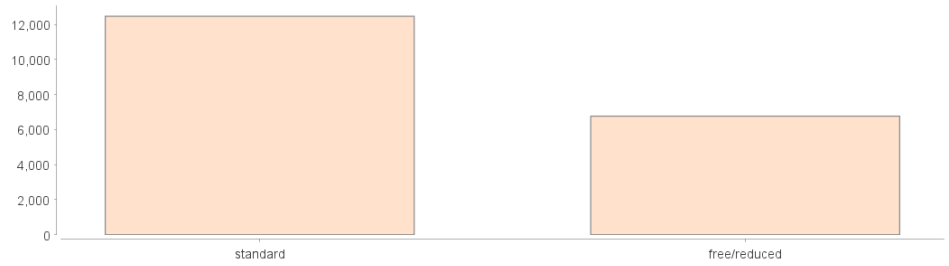
LunchType

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 65.00%
Valid: 34.98%

Top Values



2 Distinct Values:

Value	Count	Percentage
standard	12,472	64.81%
free/reduced	6,771	35.19%

The majority of students, approximately 65%, have a standard lunch type, while around 35% are included in free or reduced lunch program.

5. Test preparation completion – completed or none

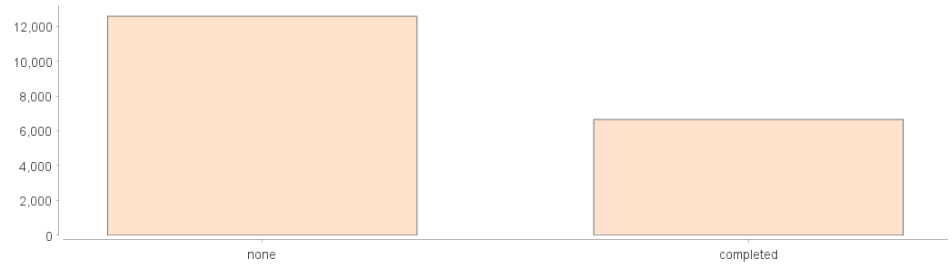
TestPrep

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 65.54%
Valid: 34.45%

Top Values



2 Distinct Values:

Value	Count	Percentage
none	12,587	65.41%
completed	6,656	34.59%

Among the students, approximately 65% reported having no test preparation, while around 35% indicated completing test preparation.

6. Parent's Marital Status – married / single / widowed / divorced

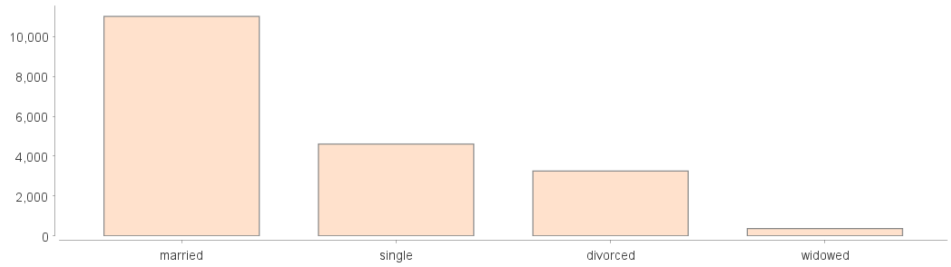
ParentMaritalStatus

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.02%
Stability: 57.47%
Valid: 42.51%

Top Values



4 Distinct Values:

Value	Count	Percentage
married	11,009	57.21%
single	4,608	23.95%
divorced	3,256	16.92%
widowed	370	1.92%

The parental marital status distribution reveals that the majority of students, around 60%, come from married households and around 24% come from single-parent households. Divorced parents represent about 17% of the total, while widowed parents make up around 2%.

7. Sports Practice Frequency – never, sometimes, regularly

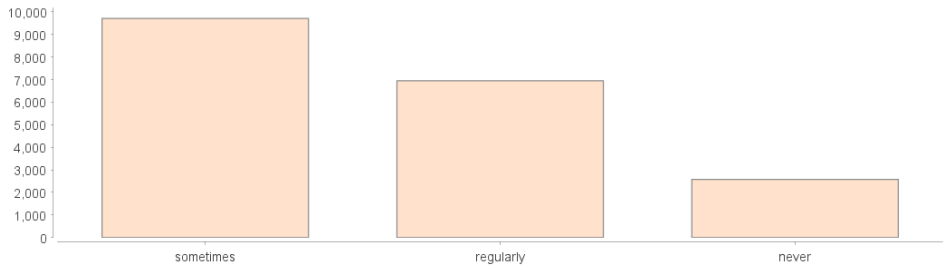
PracticeSport

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.02%
Stability: 50.01%
Valid: 49.97%

Top Values



3 Distinct Values:

Value	Count	Percentage
sometimes	9,715	50.49%
regularly	6,950	36.12%
never	2,578	13.40%

The involvement of students in sports activities varies. Approximately 50% of them do sport sometimes, 36% regularly. Conversely, around 13% of students declare that never participate in sports. Understanding these patterns can shed light on the physical activity levels of students and have significant impact on overall health and academic achievements.

8. Is first child – yes / no

IsFirstChild

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 64.53%
Valid: 35.46%

Top Values

2 Distinct Values:

Value	Count	Percentage
yes	12,417	64.53%
no	6,826	35.47%

Similarly to proportion of classes of variable Lunch Type and Test Preparation, around 65% of students are reported as being the first child, while almost half that percentage, around 35% are not. This discrepancy suggests that the majority of students is the eldest in their families.

9. Number of siblings – from 0 to 7

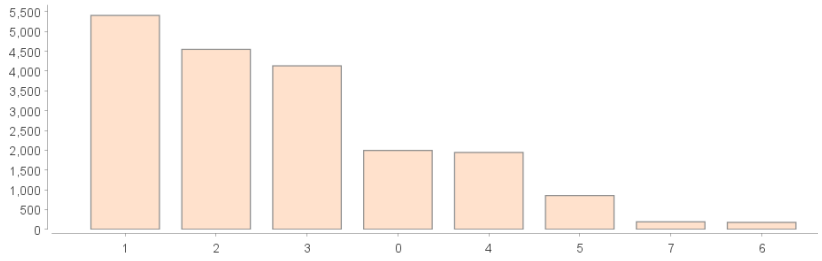
NrSiblings

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.04%
Stability: 27.86%
Valid: 72.10%

Top Values



8 Distinct Values:

Value ↑	Count	Percentage
0	1,993	10.36%
1	5,407	28.10%
2	4,549	23.64%
3	4,132	21.47%
4	1,943	10.10%
5	853	4.43%
6	176	0.91%
7	190	0.99%

The distribution of the number of siblings among students indicates a range of family sizes. The highest percentage of respondents, around 28%, have just one sibling, around 24% have two and 21% have three. Fewer students have larger family sizes. Data also show that approximately 10% of students are an only child. Understanding these family dynamics can provide insights into student's social environments and may have an impact on their academic experiences and outcomes.

10. Transport Means – school bus/private

TransportMeans

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 58.78%
Valid: 41.21%

Top Values

2 Distinct Values:

Value	Count	Percentage
school_bus	11,280	58.62%
private	7,963	41.38%

Approximately 60% of students are using the school bus and about 40% are using private transportation.

11. Weekly Study Hours – less than 5h, 5-10, more than 10h

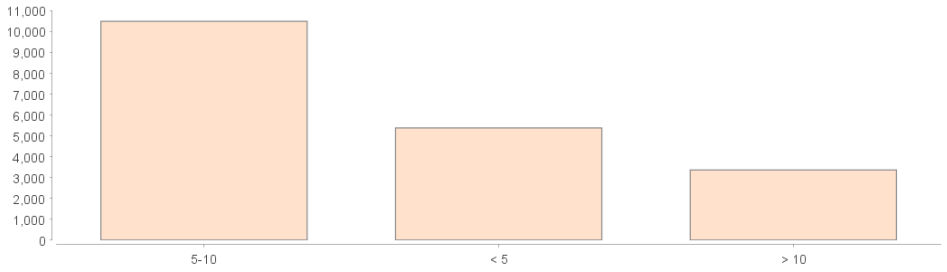
WklyStudyHours

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.02%
Stability: 54.85%
Valid: 45.14%

Top Values



3 Distinct Values:

Value	Count	Percentage
5-10	10,499	54.56%
< 5	5,381	27.96%
> 10	3,363	17.48%

Students' weekly study habits exhibit a varied distribution. Roughly 55% allocate between 5 to 10 hours per week for studying, about 28% spend less than 5 hours and around 18% dedicate more than 10 hours.

12. Reading Score – from 0 to 100

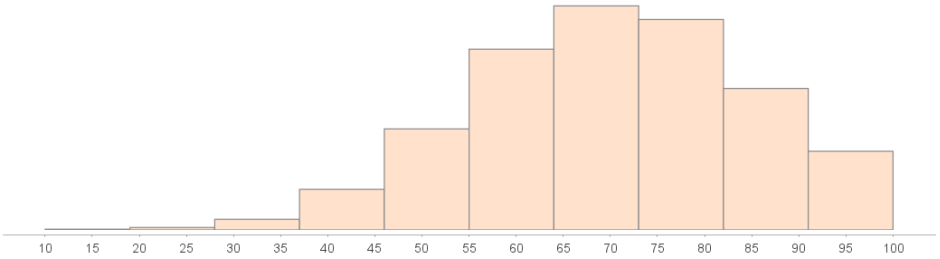
ReadingScore

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.46%
Stability: 3.20%
Valid: 96.34%

Distribution



Statistics

Name	Value
Minimum	10
Maximum	100
Average	69.534
Standard Deviation	14.786

The reading scores range from a minimum of 10 to a maximum of 100, showcasing the wide diversity in student performance. On average, students scored around 70 what gives us an idea of the typical reading level. The standard deviation of around 15 illustrates how much scores differ from the average.

13. Writing Score – from 0 to 100

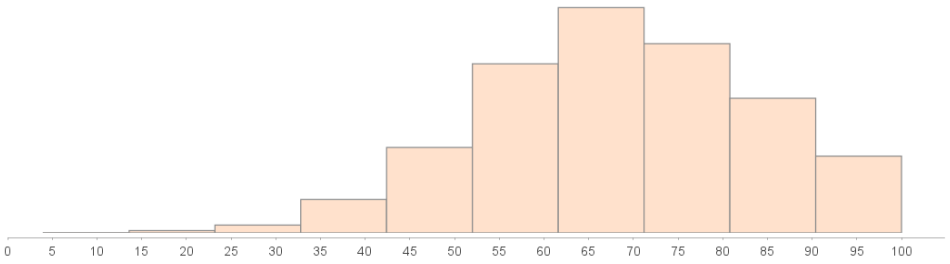
WritingScore

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.46%
Stability: 2.79%
Valid: 96.75%

Distribution



Statistics

Name	Value
Minimum	4
Maximum	100
Average	68.603
Standard Deviation	15.482

The minimum score obtained by a student on the writing exam is 4, while the maximum is 100. The standard deviation of around 16 indicates the extent to which scores deviate from the average, which, in our case, equals around 69.

14. Math Score – from 0 to 100

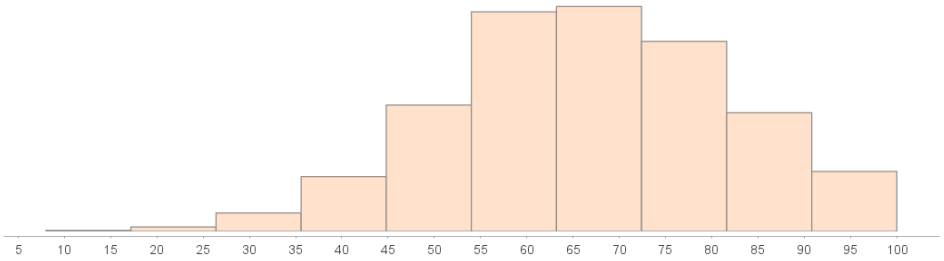
MathScore

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.47%
Stability: 2.80%
Valid: 96.74%

Distribution



Statistics

Name	Value
Minimum	0
Maximum	100
Average	66.636
Standard Deviation	15.362

When it comes to the math exam, students received results ranging from 0 to 100. This indicates a wide span of student performance. On average, students scored approximately 66, with a standard deviation of around 16. Comparing the results of three exams - math, reading, writing – it is evident that they exhibit very similar distributions, with almost identical averages and standard deviations.

3.2 Two – dimensional exploratory data analysis

CORRELATION TABLE

Figure 5, which presents a correlation table, indicates that *Reading Score* and *Writing Score* are highly positively correlated with each other and with *Math Score*. To address multicollinearity, I have opted to remove the *Writing Score* attribute. The rest of variables (quantitative / qualitative binominal) is not correlated with each other and with independent variable, as evidenced by the very low values of Pearson's correlation.

Attributes	WklyStu...	NrSiblin...	att1	Gender	EthnicG...	ParentE...	LunchT...	TestPrep	Parent...	Practice...	IsFirstC...	Transp...	MathSc...	Readin...	Writing...
WklyStud...	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?
NrSiblings	?	1	?	?	?	?	?	?	?	?	?	?	?	?	?
att1	?	?	1	-0.003	?	?	0.003	-0.010	?	?	0.005	-0.004	0.000	-0.001	0.001
Gender	?	?	-0.003	1	?	?	0.000	-0.009	?	?	-0.006	-0.010	0.157	-0.243	-0.293
EthnicGr...	?	?	?	?	1	?	?	?	?	?	?	?	?	?	?
ParentEd...	?	?	?	?	?	1	?	?	?	?	?	?	?	?	?
LunchType	?	?	0.003	0.000	?	?	1	-0.000	?	?	0.005	-0.001	-0.373	-0.262	-0.279
TestPrep	?	?	-0.010	-0.009	?	?	-0.000	1	?	?	0.002	0.017	0.143	0.213	0.294
ParentMa...	?	?	?	?	?	?	?	?	1	?	?	?	?	?	?
PracticeS...	?	?	?	?	?	?	?	?	?	1	?	?	?	?	?
IsFirstChild	?	?	0.005	-0.006	?	?	0.005	0.002	?	?	1	-0.006	-0.008	-0.007	-0.002
Transport...	?	?	-0.004	-0.010	?	?	-0.001	0.017	?	?	-0.006	1	-0.006	-0.001	-0.001
MathScore	?	?	0.000	0.157	?	?	-0.373	0.143	?	?	-0.008	-0.006	1	0.819	0.809
ReadingS...	?	?	-0.001	-0.243	?	?	-0.262	0.213	?	?	-0.007	-0.001	0.819	1	0.953
WritingSc...	?	?	0.001	-0.293	?	?	-0.279	0.294	?	?	-0.002	-0.001	0.809	0.953	1

Figure 5 Correlation table

The positive correlation between *Writing Score* and *Reading Score*, as previously indicated in the correlation table, is visually confirmed by the scatter plot presented in Figure 6. This plot demonstrates that as the *Reading Score* increases, so does the *Writing Score* and vice versa.

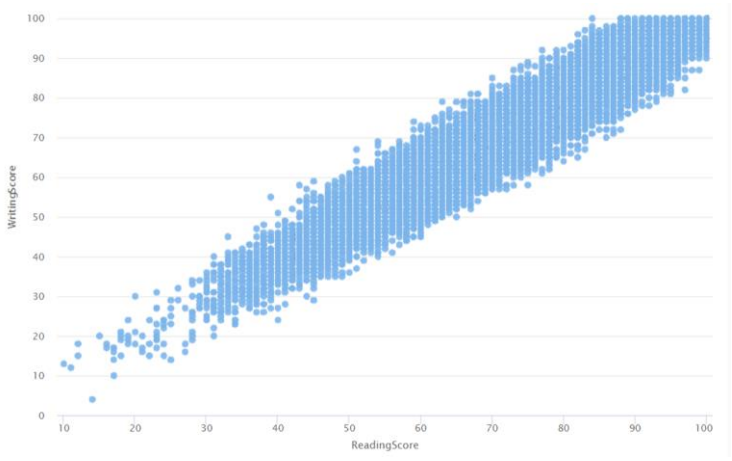


Figure 6 Scatter plot – Writing Score versus Reading Score

The positive relationship observed in the scatter plots in Figure 7 corroborates the findings from the correlation table previously presented in Figure 5. These scatter plots illustrate the connection between *Math Scores* and *Reading/ Writing Score*s differentiated by gender. Higher results from

writing / reading exams correspond to higher math scores. Additionally, the plots suggest a trend where males tend to perform better compared to females.

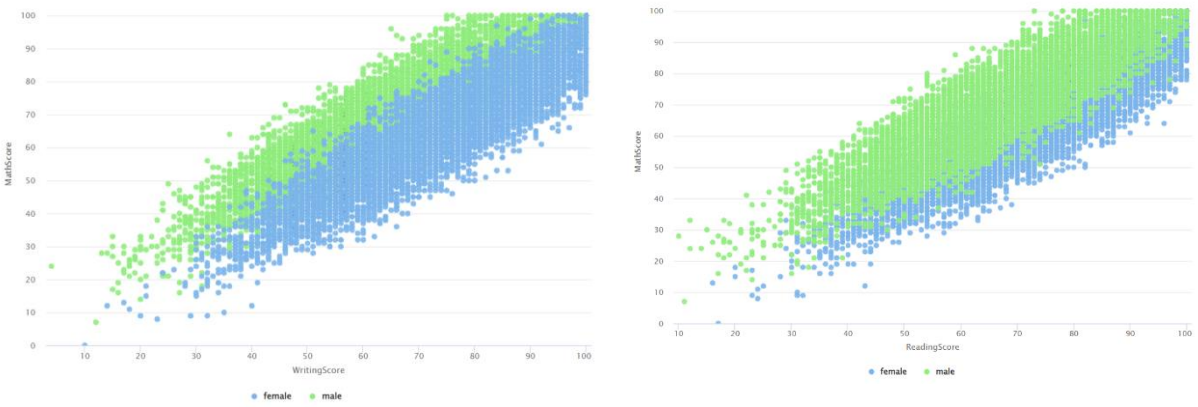
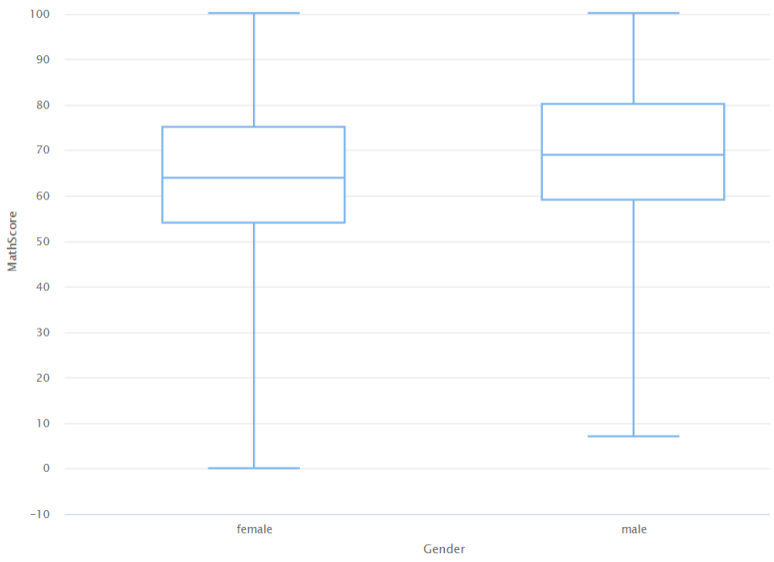


Figure 7 Scatter plots – Writing/Reading Score versus Math Score

BOX PLOTS

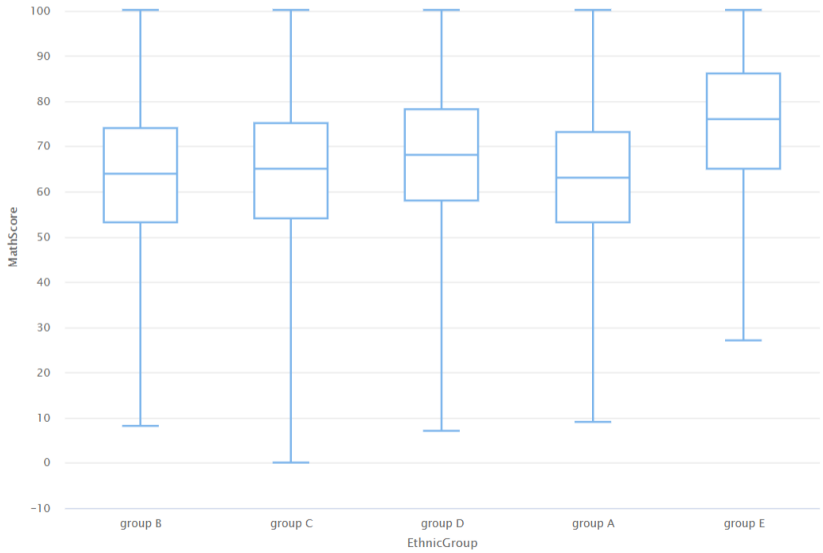
In this part, box plots were employed to explore the relationship between the dependent quantitative variable, Math Score, and the qualitative independent variables. Box plots provided a visual representation of the distribution, central tendency (median) and variability of dataset.

1. Gender – male/female



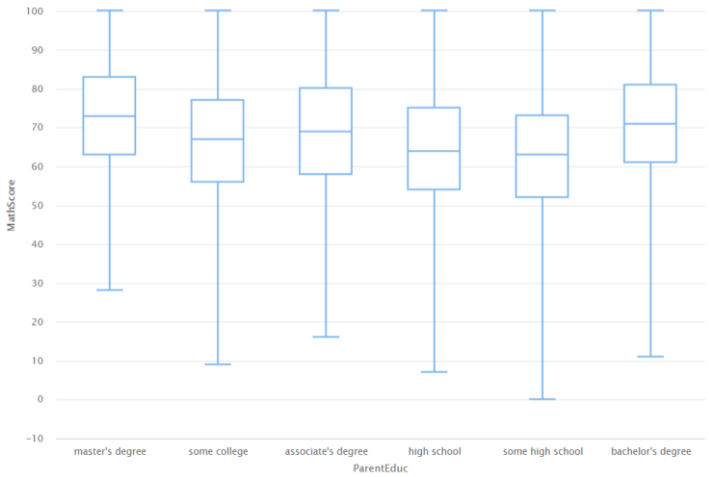
The box plots illustrate a difference in median math score, for female median is close to 65, while for males close to 70. Moreover, the range for males is narrower, suggesting that males tend to achieve slightly better results compared to females. However, the IQRs for both groups are quite similar.

2. Ethnic Group – group A to E



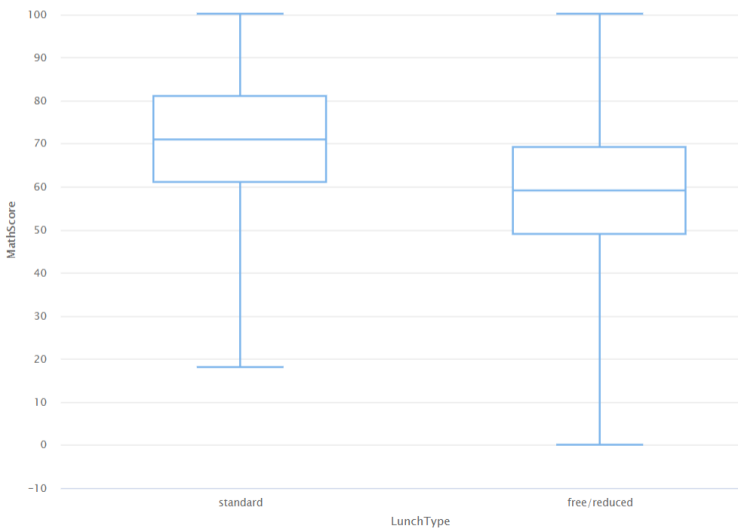
The medians show a gradual increase from Group A to Group E. Group E stands out from others due to the fact that has considerably narrower range of math scores (from around 30 to 100) and significantly higher median. However, the box lengths are relatively similar across all categories and each box generally displays symmetrical distribution, what suggests comparable variability with respect to median.

3. Parent's Education Background – from some high school to master's degree



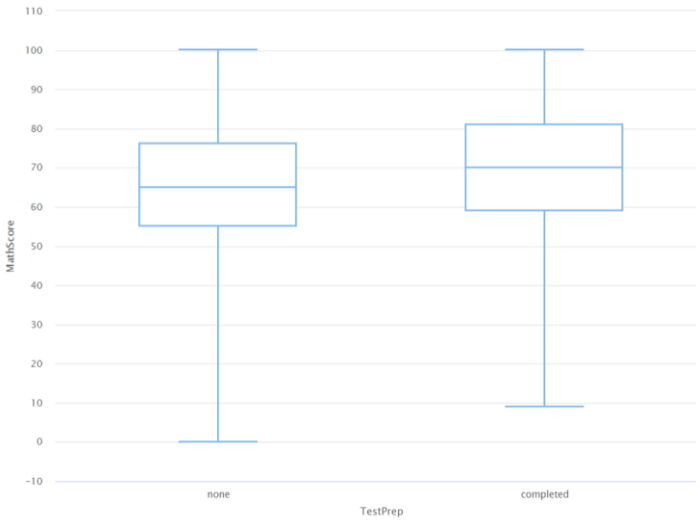
The data shows a general trend of higher median math scores associated with higher levels of parent education, with master's degree and bachelor's degree showing the highest median scores. The shortest range of scores, from around 30 to 100, is observed in the master's degree category. Instead, the widest range, from 0 to 100, refers to some high school what indicates high variability within that category. Overall, while parent education level appears to correlate to some extent, it is not the sole determinant of performance in math exam.

4. Lunch Type – standard or free/reduced



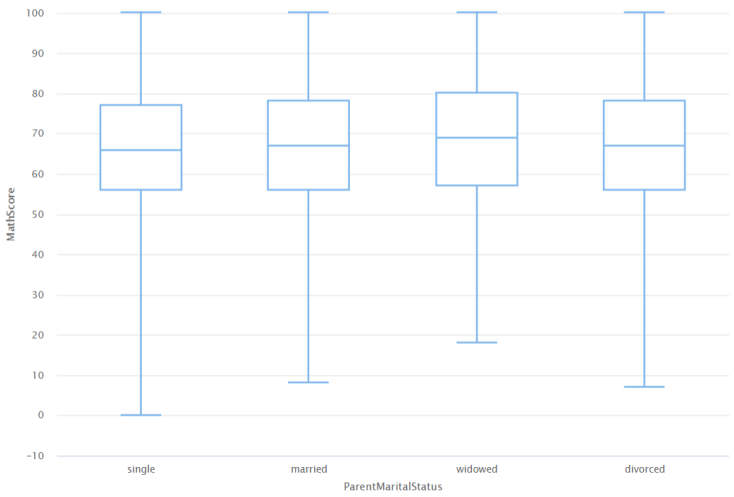
There is a noticeable difference, of around 10 units, in medians between students receiving standard lunch and those receiving free or reduced lunch, with a higher median score for standard lunch. While the variability in scores is greater for free/reduced lunch class, the interquartile ranges are comparable between two groups. These findings suggest that lunch type may have an impact on math performance and students who pay regular fee shows better results.

5. Test preparation completion – completed or none



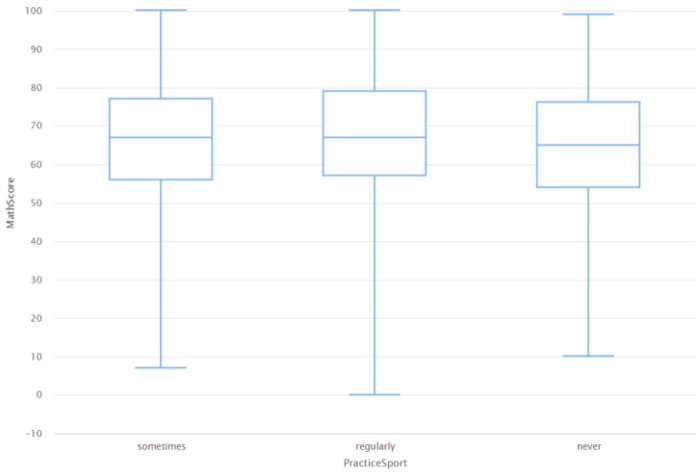
In summary, the box plots show that students who completed test preparation tend to have slightly higher median math scores compared to those who did not. Both categories exhibit similar variability in scores, comparable interquartile ranges and symmetrical distributions. Students with completed test preparation gained results from 9 to 100, while those without it scored from 0 to 100.

6. Parent's Marital Status – married / single / widowed / divorced



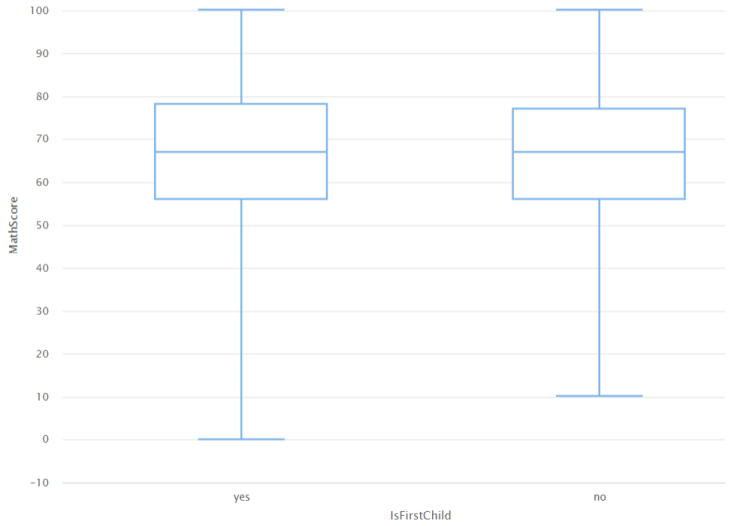
The medians math scores are quite similar across all categories, with the highest median for widowed and the lowest median for single parents. The IQRs appears to be consistent for marital status classes, with no significant outliers observed. However, the interesting fact is that students whose one parent passed away, have a minimum score of around 19 while for other categories a minimum score is equal to 0 and close to 10.

7. Sports Practice Frequency – never, sometimes, regularly



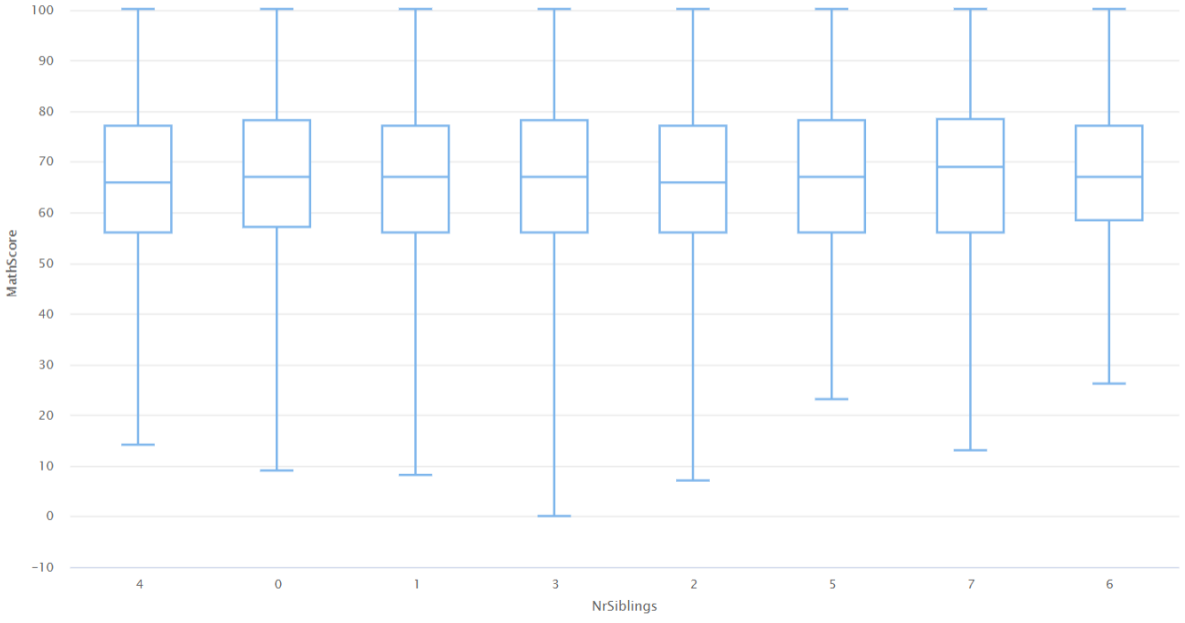
The box plots shows minimal differences in medians math scores between students who practice sports sometimes, regularly or never. The overall range of scores varies slightly, with the widest range in regularly practicing sports. The IQRs are similar across all three categories, indicating comparable spreads of math scores. Overall, these boxplots suggest that the frequency of sports practice may not have significant impact on math performance.

8. Is first child – yes / no



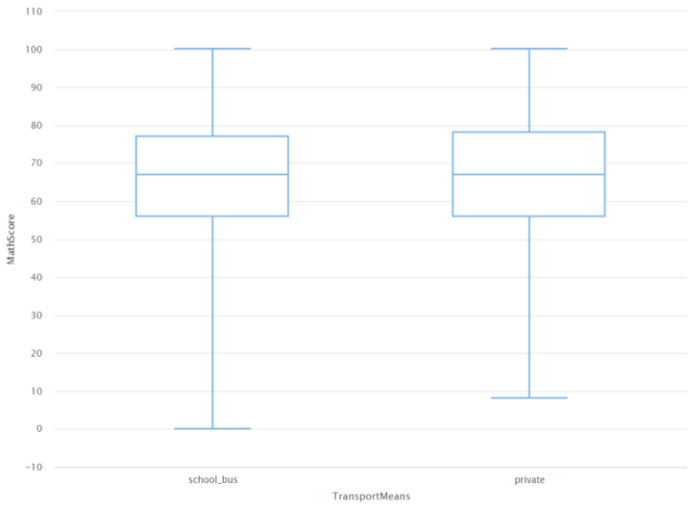
The central tendency (median) is similar for both groups and is equal to around 68. This means that 50% of respondents scored 68 or below on the math exam no matter if that student was first child or not. Additionally, the IQRs are also similar and the median lines are positioned close to the center, indicating a symmetrical distribution. The only aspect exhibiting diversity is the range of scores: first child students achieved scores from 0 to 100, whereas non first child scored from 10 to 100 (a narrower range).

9. Number of siblings – from 0 to 7



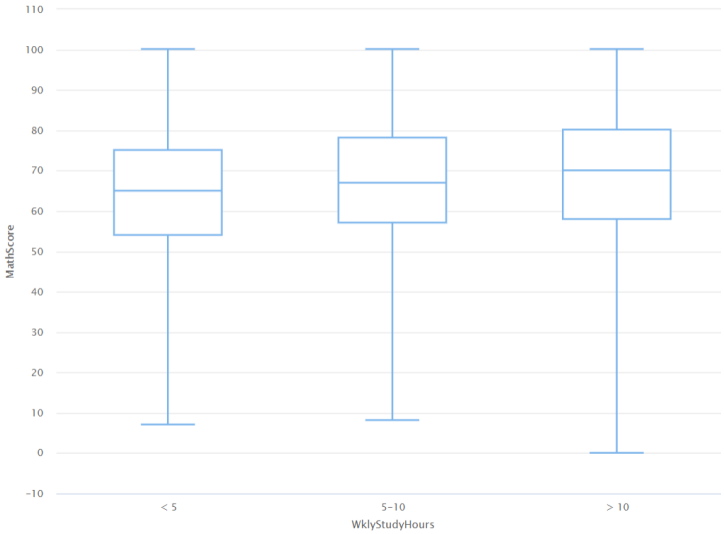
They are some fluctuations in medians math scores across different number of siblings, however medians do not differ significantly from each other. The box plots for classes 0 and 1 exhibit almost identical distributions. It is noteworthy that the range of scores extends from 0 to 100 only for the number of siblings equals 3. Moreover, students with 5 or 6 siblings achieved scores exceeding 20.

10. Transport Means – school bus/private



Both categories have similar median math scores. The box length for private transport is slightly longer what suggests slightly greater variability in scores compared to the school bus. There are probably no outliers in both categories. The overall range of scores is wider for the school bus and it includes all possible results from the math exam (from 0 to 100).

11. Weekly Study Hours – less than 5h, 5-10, more than 10h



The median math scores increases as the weekly study hours increase, indicating a positive relationship between study time and median scores. The box length is similar for all three categories, what suggests that the variability in math scores within each category is relatively consistent, despite the differences in weekly study hours. It is also interesting that there are some students who declared that they studied more than 10 hours per week and in final math exam they obtained the score close to 0, while for other categories minimum value was close to 10. It may show that more does not mean always better or that there are potential outliers in the third category.

4. Regression model comparison

I considered three algorithms for predicting a quantitative variable *Math Score*: **Linear Regression, Generalized Linear Regression and Support Vector Machine**. I conducted performance comparisons for each algorithm on both training and test dataset across three iterations.

In the first iteration, I excluded the *Writing Score* variable from analysis.

For the second iteration, I aimed to refine the model based on insights from the Linear Regression coefficient table for training dataset presented in Figure 8. In Linear Regression, coefficients indicate the strength and direction of the relationship between the independent variable and dependent variable. A higher coefficient means a stronger relationship. What is more, positive coefficients shows a positive relationship with the outcome , while negative coefficients signify the opposite.

Additionally, I considered variables with p-value less than 0.05 to be statistically significant. To enhance model performance, I removed all variables with both p-value greater or equal to 0.05 and those lacking a star in the Code column. Subsequently, I tested all three algorithms again.

The removed variables include: *NrSiblings, ParentEducation, ParentMaritalStatus*.

Attribute ↑	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
(Intercept)	66.657	∞	?	?	0	1	
EthnicGroup = gro...	-0.389	0.120	-0.026	0.994	-3.237	0.001	***
EthnicGroup = gro...	-0.430	0.122	-0.028	0.989	-3.512	0.000	****
EthnicGroup = gro...	-0.581	0.122	-0.038	0.990	-4.749	0.000	****
EthnicGroup = gro...	0.067	0.122	0.004	0.997	0.550	0.583	
EthnicGroup = gro...	1.491	0.126	0.097	0.937	11.825	0	****
Gender = female	-2.799	0.123	-0.183	0.971	-22.681	0	****
Gender = male	2.810	0.123	0.183	0.971	22.770	0	****
LunchType = free/r...	-1.153	0.134	-0.075	0.833	-8.635	0	****
LunchType = stan...	1.167	0.134	0.076	0.833	8.741	0	****
NrSiblings = 0	-0.065	0.121	-0.004	1.000	-0.535	0.593	
NrSiblings = 1	-0.025	0.121	-0.002	1.000	-0.209	0.834	
NrSiblings = 2	-0.101	0.122	-0.007	1.000	-0.830	0.406	
NrSiblings = 3	0.033	0.122	0.002	1.000	0.272	0.786	
NrSiblings = 4	-0.076	0.121	-0.005	1.000	-0.627	0.530	
NrSiblings = 5	-0.012	0.122	-0.001	1.000	-0.100	0.920	
NrSiblings = 6	0.050	0.124	0.003	1.000	0.406	0.685	
NrSiblings = 7	-0.059	0.119	-0.004	1.000	-0.496	0.620	

ParentEduc = ass...	0.145	0.122	0.009	0.996	1.189	0.234	
ParentEduc = bac...	0.199	0.123	0.013	0.988	1.624	0.104	
ParentEduc = high...	-0.114	0.122	-0.007	0.992	-0.933	0.351	
ParentEduc = mas...	0.007	0.123	0.000	0.987	0.054	0.957	
ParentEduc = som...	0.050	0.122	0.003	1.000	0.408	0.683	
ParentEduc = som...	-0.195	0.122	-0.013	0.981	-1.596	0.110	
ParentMaritalStatu...	0.015	0.122	0.001	1.000	0.124	0.901	
ParentMaritalStatu...	0.004	0.122	0.000	1.000	0.035	0.972	
ParentMaritalStatu...	-0.030	0.122	-0.002	1.000	-0.250	0.803	
ParentMaritalStatu...	0.034	0.120	0.002	1.000	0.287	0.774	
PracticeSport = ne...	-0.570	0.122	-0.037	0.997	-4.676	0.000	****
PracticeSport = re...	0.257	0.122	0.017	0.998	2.116	0.034	**
PracticeSport = so...	-0.139	0.122	-0.009	1.000	-1.141	0.254	
ReadingScore	13.004	0.129	0.848	0.888	100.575	0	****
TestPrep = compl...	-0.270	0.123	-0.018	0.973	-2.198	0.028	**
TestPrep = none	0.263	0.123	0.017	0.973	2.140	0.032	**
WklyStudyHours ...	0.087	0.122	0.006	1.000	0.719	0.472	
WklyStudyHours ...	-0.453	0.122	-0.030	0.992	-3.708	0.000	****
WklyStudyHours ...	0.495	0.122	0.032	0.994	4.067	0.000	****

Figure 8 A regression coefficient table for training set

The third iteration involved a significant reduction in the number of regressors, with only the *Reading Score* retained. I did it mainly because the correlation table and the box plots did not reveal strong connection between independent variables and the *Math Score*.

RESULTS AND INSIGHTS – ROOT MEAN SQUARED ERROR

Variables number modification	Linear Regression	Generalized Linear Regression	Support Vector Machine
Exclusively eliminated the *Writing Score* attribute	5.904	5.909	5.936
Removed attributes: *Writing Score, NrSiblings, ParentEducation, ParentMaritalStatus*	5.919	5.923	5.947
All variables except *Reading Score* were removed	8.699	8.700	9.490

Table 3 Root mean squared error – **Training dataset performance**

Variables number modification	Linear Regression	Generalized Linear Regression	Support Vector Machine
Exclusively eliminated the *Writing Score* attribute	5.983	5.992	6.020
Removed attributes: *Writing Score, NrSiblings, ParentEducation, ParentMaritalStatus*	5.991	5.999	6.019
All variables except *Reading Score* were removed	8.771	8.775	9.610

Table 4 Root mean squared error – **Test dataset performance**

Insights:

1. As additional attributes were removed in each iteration, the algorithms consistently demonstrated better performance, as evidenced by higher root mean squared error value.

2. Across all iterations, Support Vector Machine outperformed Generalized Linear Regression and Linear Regression.

3. The results are quite unusual and troublesome because it turned out that they are better for test dataset and not for the training. It could be due to differences in data splits, used cross-validation or mistakes in feature engineering.

4. In terms of computation cost, SVM, despite delivering superior results, proved to be the slowest, probably due to its complexity. Conversely, Generalized Linear Regression was the fastest, while still achieving satisfactory results.

5. Assessment of regression assumptions

As SVM performed the best out of all three algorithms, I verified linear regression assumptions based on its provided results in the third iteration for the test data (Figure 9).

Row No.	att1	MathScore	MathScorePredicted	ReadingScore	residuals
1	127	76	75	0.911	-1
2	128	94	79	1.384	-15
3	129	62	70	0.370	8
4	130	91	78	1.317	-13
5	131	63	66	-0.104	3
6	133	89	76	1.046	-13
7	134	59	67	0.099	8
8	141	81	79	1.452	-2
9	142	57	62	-0.577	5
10	143	51	54	-1.456	3
11	144	83	80	1.519	-3
12	146	74	65	-0.171	-9
13	149	69	69	0.234	0
14	150	56	60	-0.712	4
15	154	65	63	-0.442	-2
16	160	63	66	-0.104	3
17	162	63	65	-0.239	2

Figure 9 Extract of the final dataset with variables: Math Score, Math Score Predicted, residuals and Reading *Score*

OLS (ORDINARY LEAST ASSUMPTIONS) REGRESSION'S ASSUMPTIONS

1. **Linearity** - the relationship between the independent variables and the dependent variable should be linear (the change in the dependent is proportional to the change in the independent variables). In our case, it is true for quantitative variables: *Reading/Writing Score* versus *Math Score*, what was confirmed in Figure 7 in chapter 3.2 EDA 2D .

2. **Homoscedasticity** - the variance of the errors should be constant. Figure 10 displays *residuals* against *Math Score Predicted*. This plot reveals some noteworthy patterns. Initially, for math scores below 40, the residuals are always positive. Subsequently, as the math score increases, the range of residuals widens considerably. Between predicted score of approximately 70 to 85 , the range of residuals narrows. These observations suggest potential violations of the homoscedasticity assumption, as the spread of residuals appears to vary across different math scores.

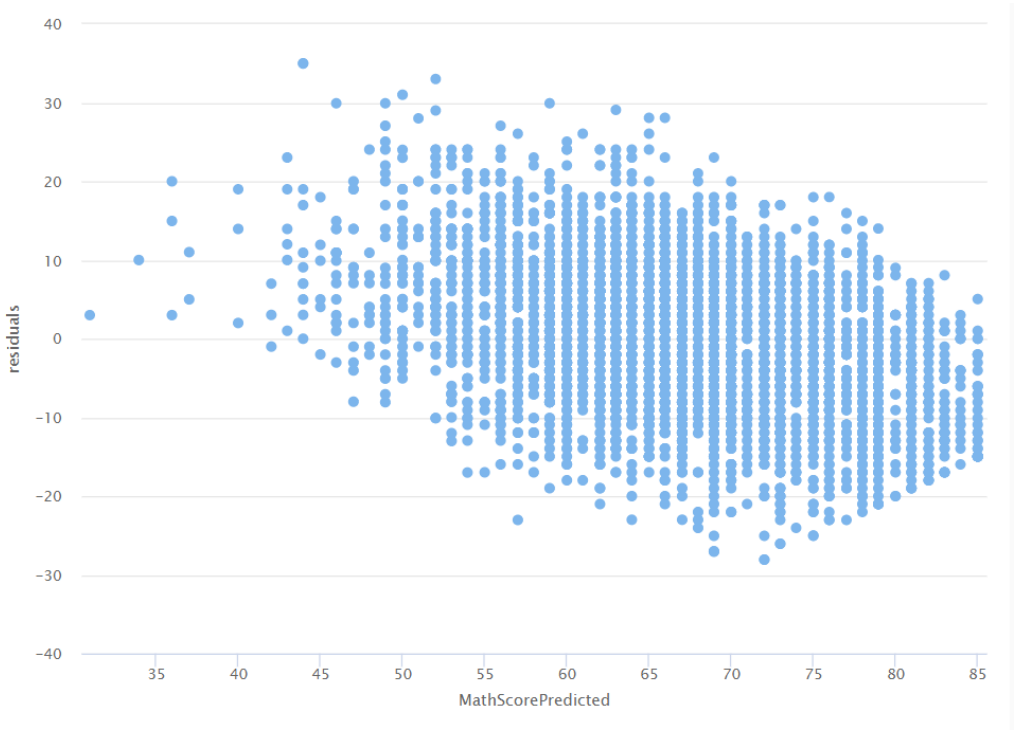


Figure 10 Scatter plot – residuals vs MathScorePredicted

3. **Normality of errors** - the errors are almost normally distributed as it is displayed in Figure 11.

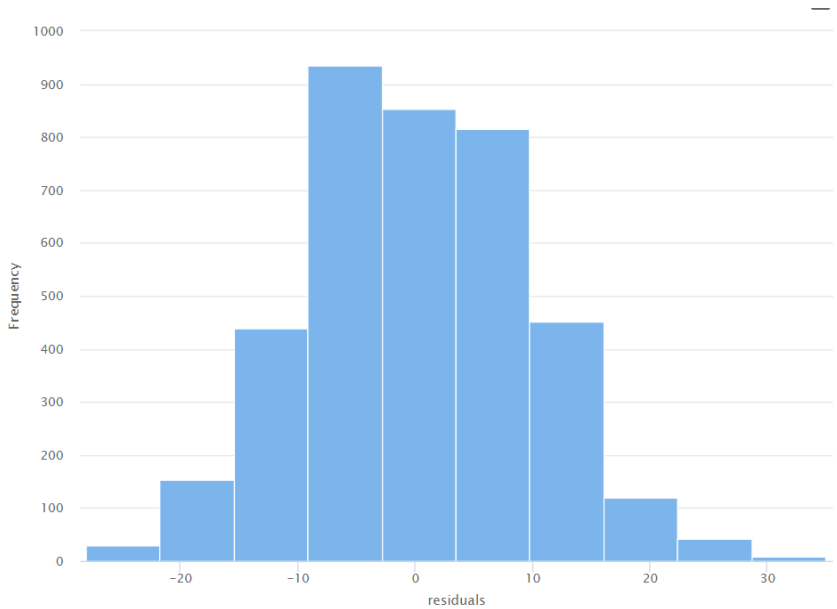


Figure 11 Histogram of residuals

4. **No perfect multicollinearity** - Multicollinearity is a phenomenon that occurs when two or more independent variables are highly correlated with each other. In other words, multicollinearity exists when there is a linear relationship between independent variables. As it was mentioned before, correlation table in Figure 5 showed that there was strong correlation between *Writing Score* and *Reading Score*. By simple removing one of them, the multicollinearity problem was solved.