DATA MANAGEMENT

Note: All the insights / findings have been mentioned along with the analysis done through the course of the report.

Dataset:

- The dataset has been taken from the URL - https://www.kaggle.com/spscientist/students-performance-in-exams
- This dataset has 1000 rows with 8 columns
    - 5 of these are categorical variables
    - 3 are numerical variables
- This dataset deals with the marks obtained by 1000 students in their math, writing and reading test and we have categorical columns such as parental education, type of lunch provided, race, gender and test preparation, which will help us determine the effect of these categorical variables on the test scores
- A general description of the dataset and the way the values are stored. (We ran the str() command in R to obtain this)

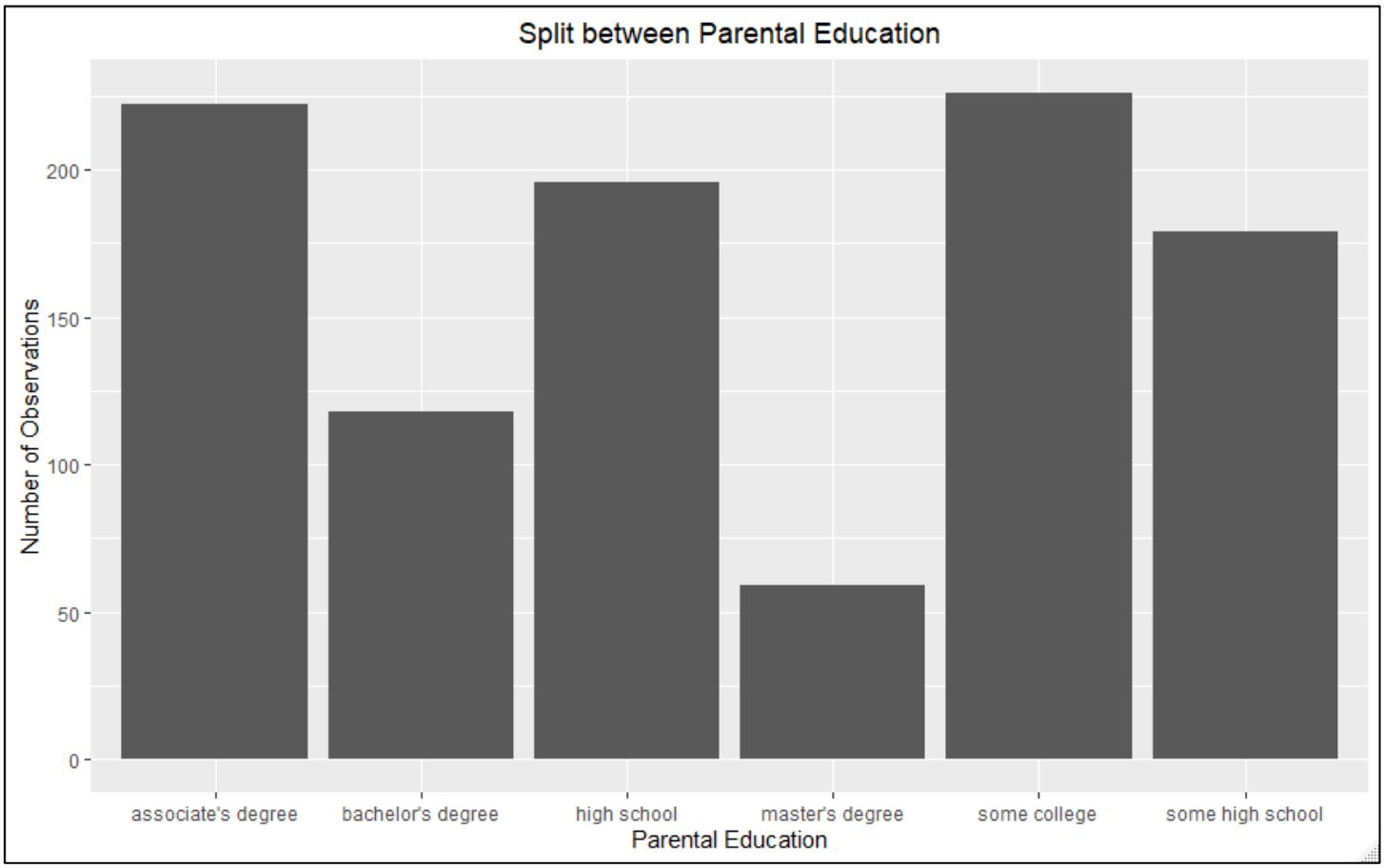
str(StudentsPerformance)

```
a) 'data.frame'    : 1000 obs. of  10 variables:
b)  $ StuID        : int  1 2 3 4 5 6 7 8 9 10 ...
c)  $ Gender       : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1
d)  $ Race         : Factor w/ 5 levels "group A","group B",..: 2 3 2 1
e)  $ Parental_Edu : Factor w/ 6 levels "associate's degree","masters"..
f)  $ Test_Prep    : Factor w/ 2 levels "completed","none": 2 1 2 2 2 2..
g)  $ Math_Score   : int  72 69 90 47 76 71 88 40 64 38 ...
h)  $ Reading_Score: int  72 90 95 57 78 83 95 43 64 60 ...
i)  $ Writing_Score: int  74 88 93 44 75 78 92 39 67 50 ...
j)  $ Overall_Score: int  218 247 278 148 229 232 275 122 195 148 ...
```

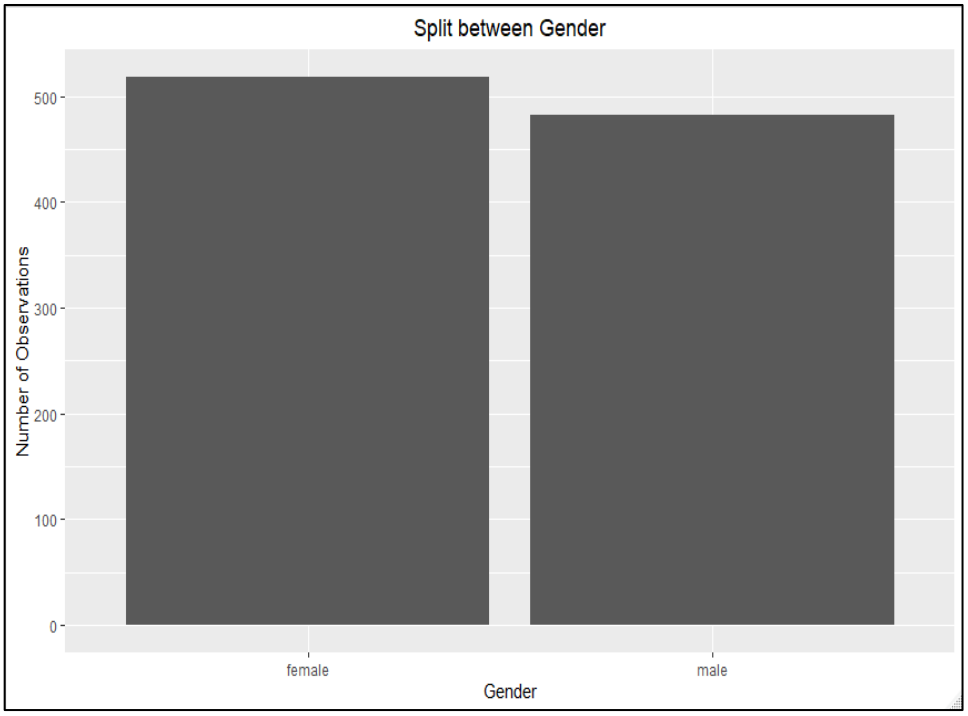
- The dataset is unique, each record belongs to one student and a simple structured dataset
    - So, we wouldn't need any normalization techniques for this dataset
    - In case we would have needed it –
        - We will break down the table into subsets and define the relationship between these entities or subsets
        - This would help us maintain data integrity and reduce data redundancy
- The data did not have many problems in terms of consistency
    - The column names and data types have been properly defined
    - There are no null values or missing values
    - There are no repeating values in the data

SQL and R have been predominantly used for the creation of this report. A few graphs have been plotted in Tableau as well

- First, we import the dataset into SQL server
- We then use bulk insert from this excel import file into a table created in the database.
- We add in a unique student id which is of the auto incrementing Identity data type
- We also alter the table structure to add in overall score, which is the sum of the Math, Reading and Writing scores
- We then proceed to check if the table has been imported correctly without any NULL values
    - We check the count to be 1000 rows without any NULL values and hence the import has been correctly done
- We proceed to check the distributions of all the categorical and numerical columns
    - We check the counts in SQL and have plotted the distributions in R

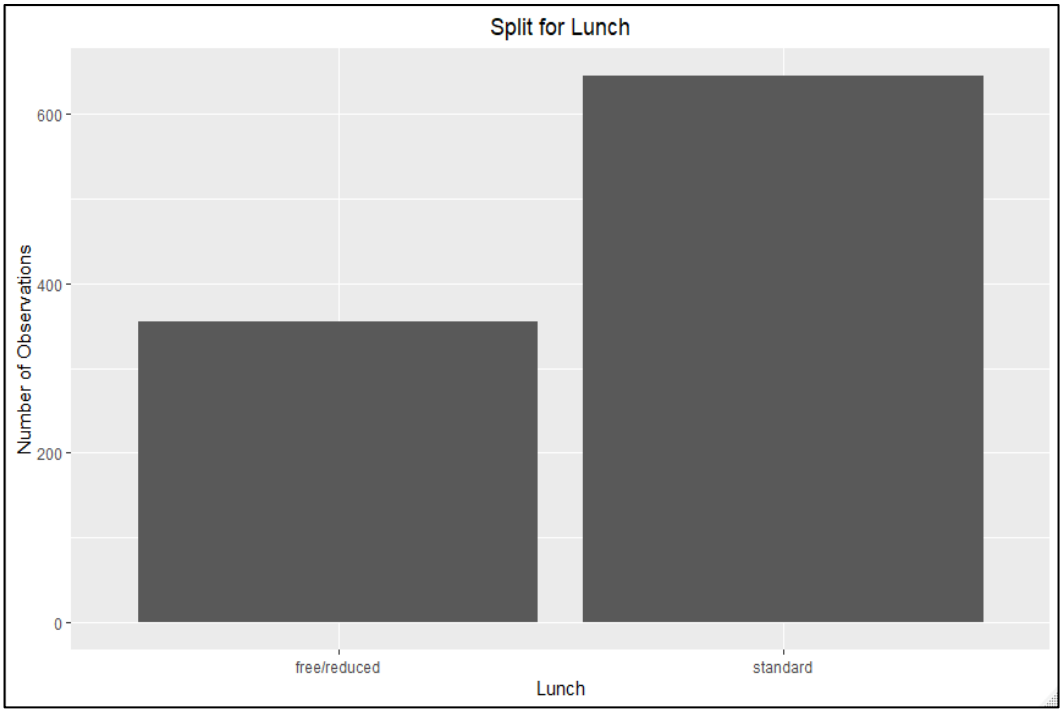


We see a majority of observations with students having parental education - "associate's degree", "some college".

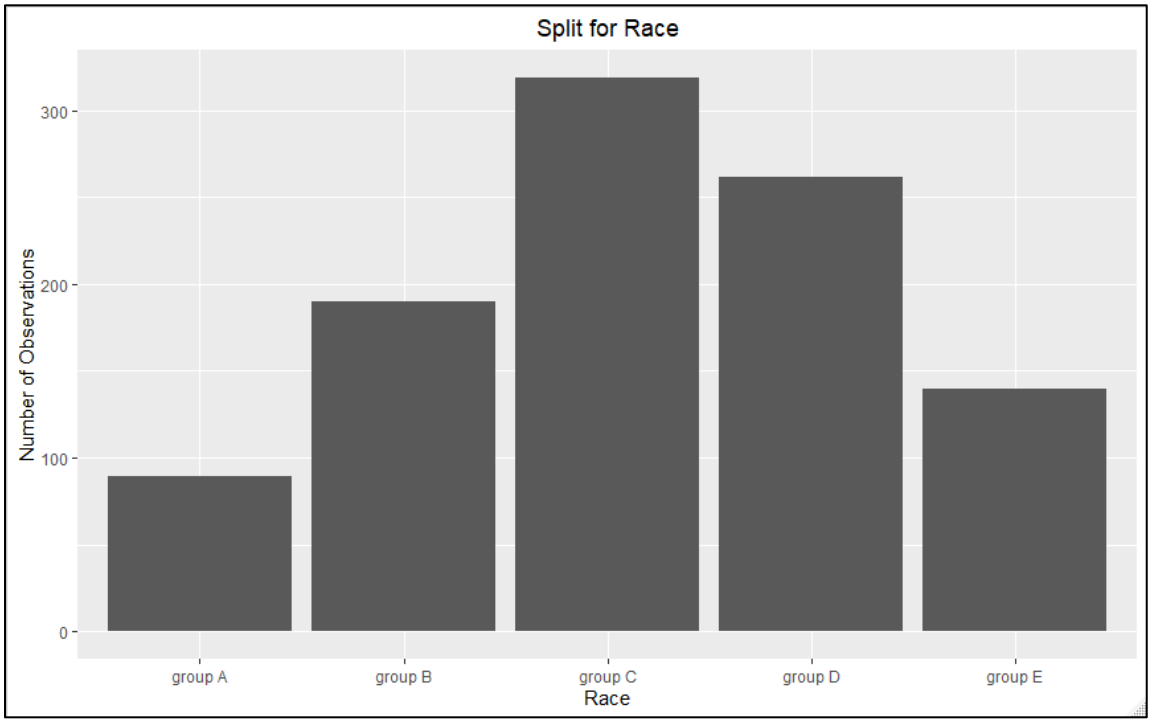


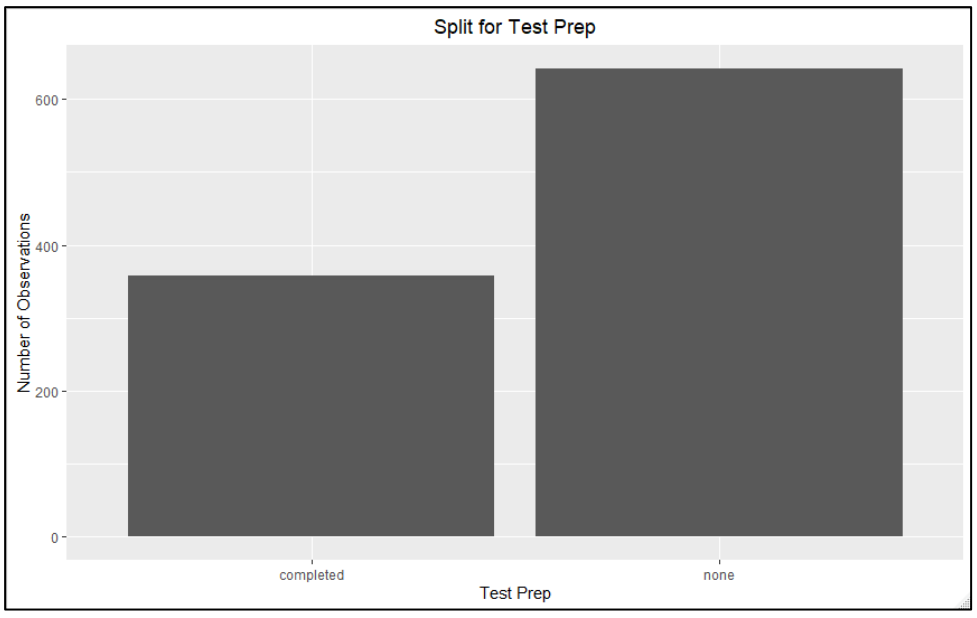
We have an almost equal number of records for both female and male, although the number of female records is slightly higher

 # More observations for standard lunch as compared to free/reduced



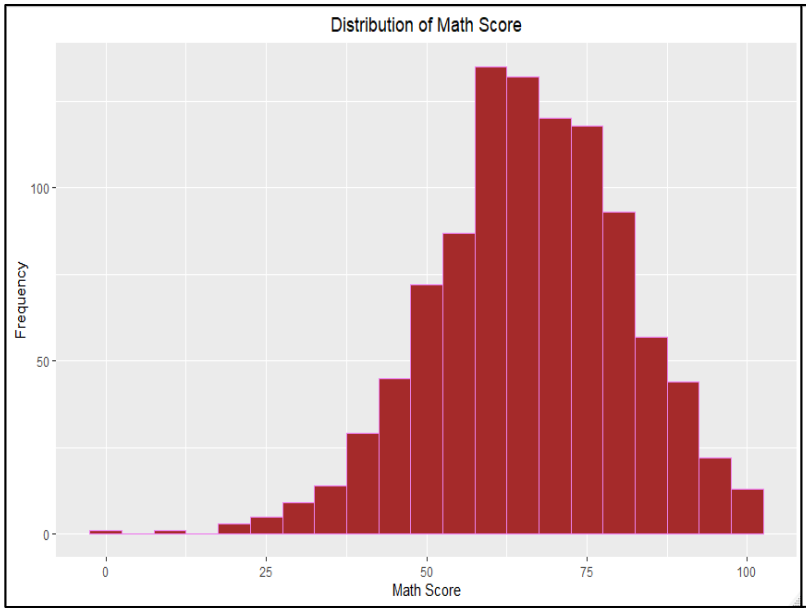
Highest number of observations present for Race C followed by Race D and Race B

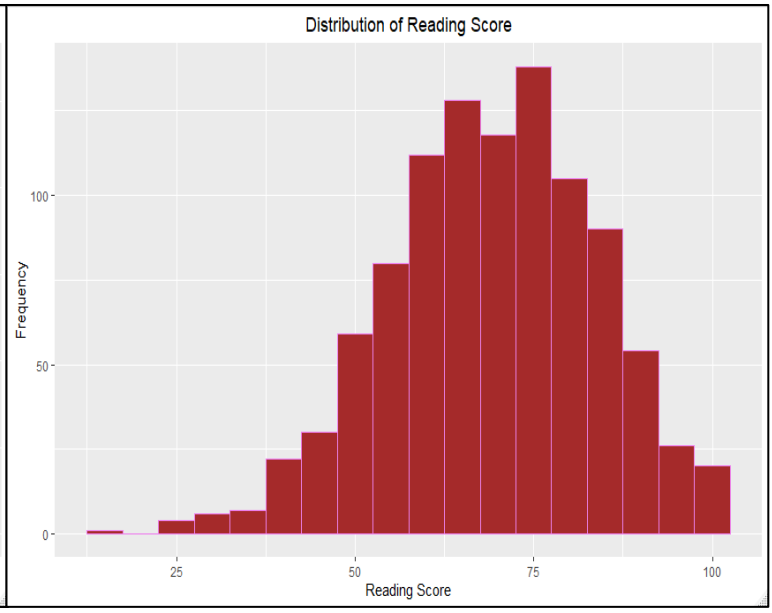


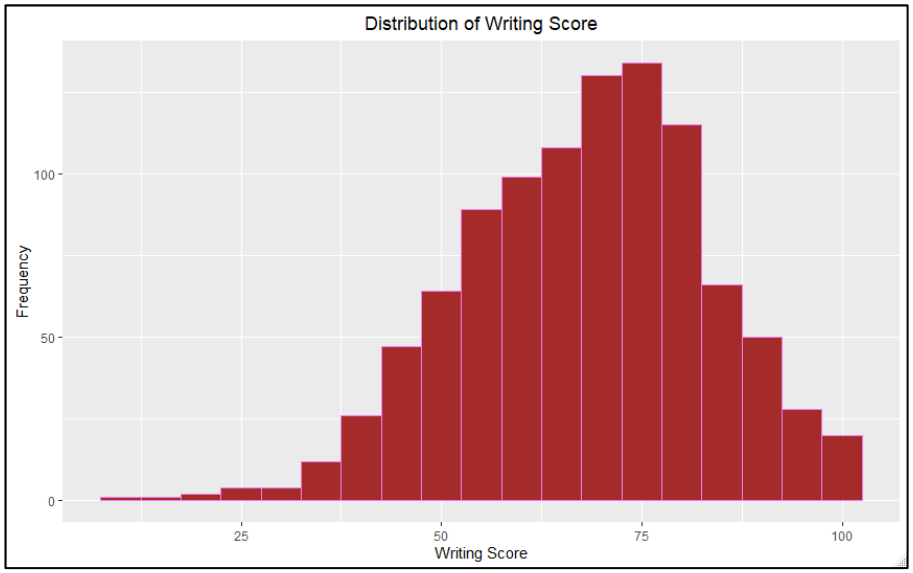


A lot of students in this sample did not complete the prep before taking the test.

Now, let us look at the numerical variables –







We see that all 3 of the scores are slightly skewed towards the left i.e. they have a longer tail towards the left.

- We tried to get some other counts in addition to the basic ones

Gender	Min_Overall	Avg_Math_Score	Avg_Writing_Score	Avg_Reading_Score
male	69	68	63	65
female	27	63	72	72

- We see that the minimum overall score is less for females as compared to males – In other words, the lowest overall score in the observations we have belong to females

- The average reading and writing scores are better for females as compared to males, but is not the case with math score

StuID	Race	Parental_Edu	Test_Prep	Math_Score	Reading_Score	Writing_Score
60	group C	some high school	none	0	17	10

- The student with lowest math, reading and writing score belongs to the race group C and has not completed any test prep before the exam

StuID	Race	Parental_Edu	Test_Prep	Math_Score	Reading_Score	Writing_Score
459	group E	bachelor's degree	none	100	100	100
917	group E	bachelor's degree	completed	100	100	100
963	group E	associate's degree	none	100	100	100

- We see that all the 3 top scorers are from the same race and 2 out of the 3 students actually scored the highest without completing the test prep

Now, Let us look at average overall_score by parental education

Parental_Edu	Overall_Average
master's degree	220
bachelor's degree	215
associate's degree	208
some college	205
some high school	195
high school	189

- This is in line with our hypothesis, that parental education influences the students' performance
- We can see that the students with parental education of masters have the highest average when compared to other levels of education

Look at the number of students who get free/ reduced lunch and see the division by parental education

Parental_Edu	Lunch	Number_Observations1	Number_Observations2
associate's degree	free/reduced	77	222
bachelor's degree	free/reduced	44	118
high school	free/reduced	70	196
master's degree	free/reduced	24	59
some college	free/reduced	79	226
some high school	free/reduced	61	179

- The percentage is pretty constant and does not give much insight.

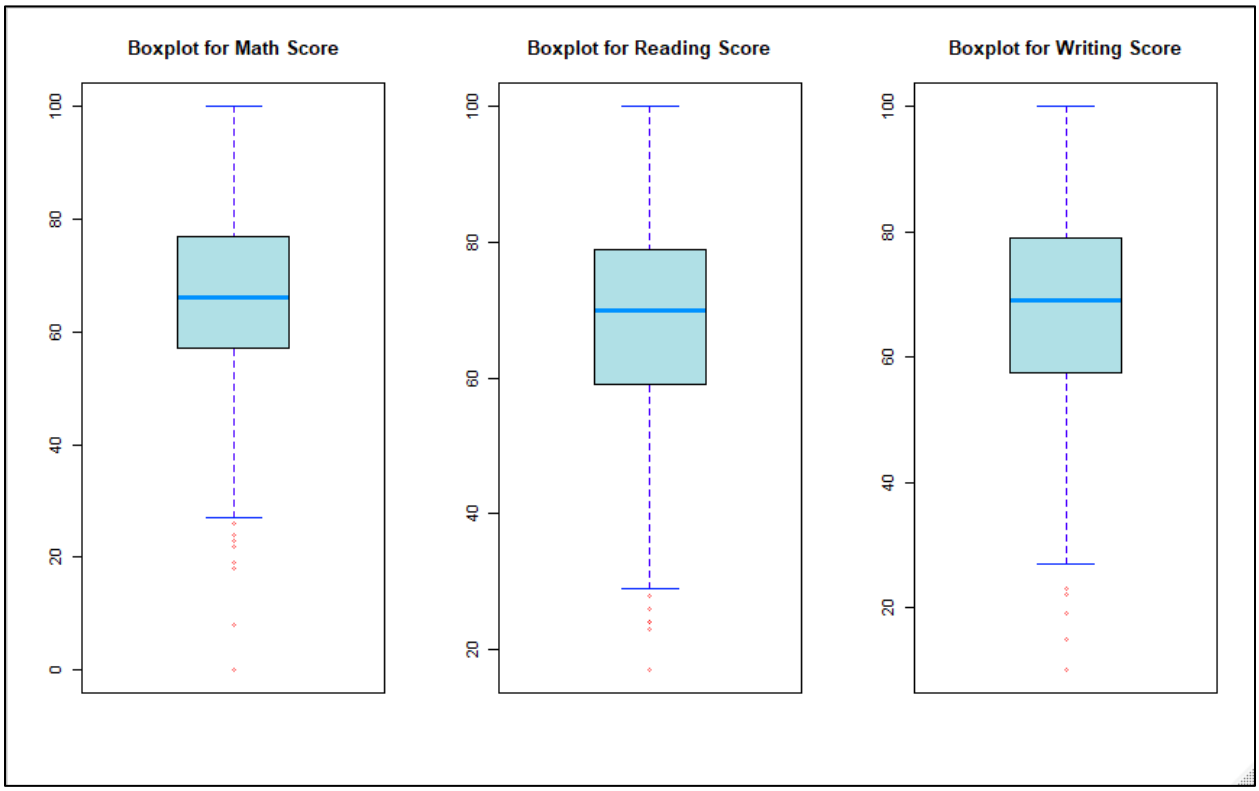
Now, let us shift to R to understand if there are any outliers, outlier treatment and fit a regression model.

- Let us check the summary of the dataset that has been imported using the ODBC connection

```
> summary(StudentsPerformance$Math_Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   57.00   66.00   66.09   77.00  100.00
> summary(StudentsPerformance$Reading_Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.00   59.00   70.00   69.17   79.00  100.00
> summary(StudentsPerformance$Writing_Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   57.75   69.00   68.05   79.00  100.00
> summary(StudentsPerformance$Overall_Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   27.0   175.0   205.0   203.3   233.0   300.0
```

- We see that there are no missing values
- We do not have any negative values for subject scores - all values are between 0 and 100

Let us see if there are any outliers by using the boxplot:



We see some outliers on the lower end for all the 3 columns – let us start looking at how many rows we lose because of this –

- For Math score we see that there are 8 observations which can be considered as outliers according to the $Q_{0.25}$ - 1.5*IQR method (code attached separately), so we remove them
- When we check for outliers in the Reading Score with the same method, we see that there are 6 outlying observations
- Then, we proceed to check for outlying values in the writing score column, we now see that there are no outliers. This is because we must have already removed the outliers in this column as well through the prior 2 steps

- We then proceed to fit a regression model to estimate overall score with the help of categorical variables

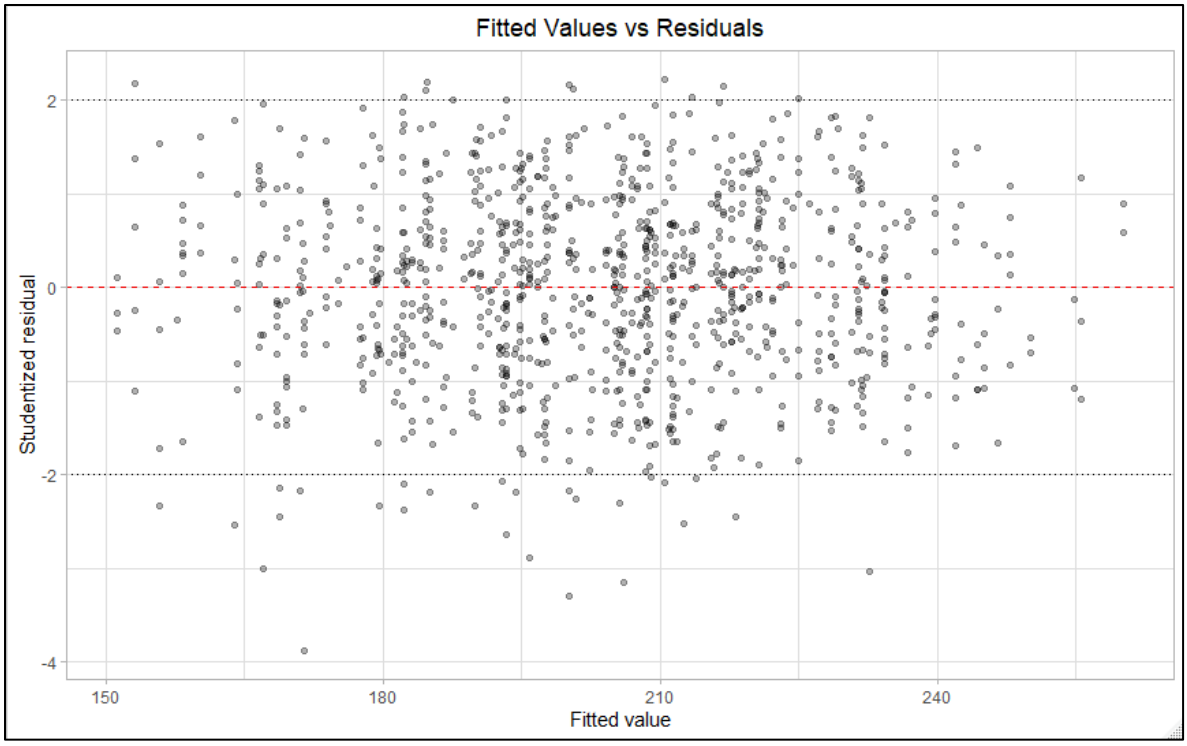
- We used 3 approaches – Forward Selection, Backward Elimination and Step Wise Selection, all the 3 approaches gave us the same model below. We included all the categorical variables and the also included up to 3-way interactions. All these 3 models use BIC as the selection criterion to select the best model

```
Overall_Score ~ Lunch + Test_Prep + Gender + Parental_Edu
```

Model Stats:

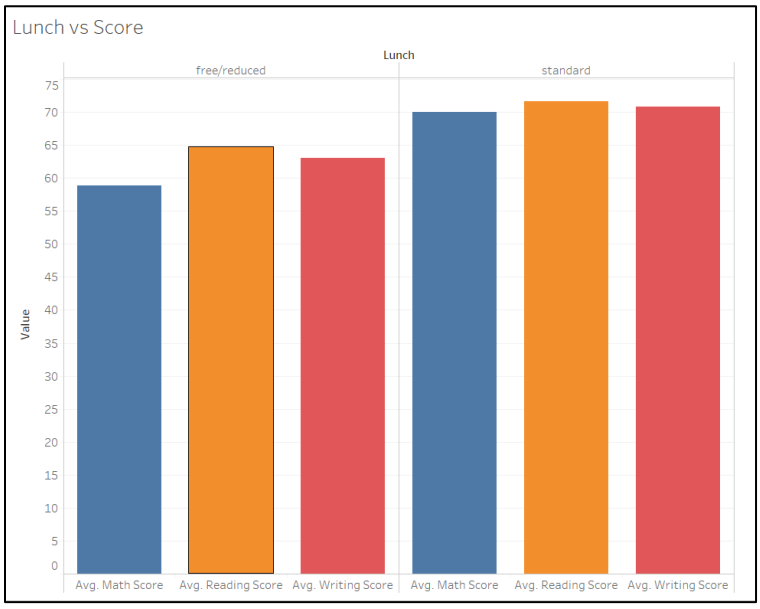
	fit_be	fit_fs	fit_step
AIC	9881.012	9881.012	9881.012
BIC	9929.949	9929.949	9929.949
adjR2	0.201	0.201	0.201
RMSE	36.095	36.095	36.095
PRESS	1296693.332	1296693.332	1296693.332
nterms	9.000	9.000	9.000

- **AdJ. R^2** is very low which means only 20% of the variability in overall score is being explained by the categorical variables
- Let us look at residual's vs fitted plot to understand how the residual plots look –

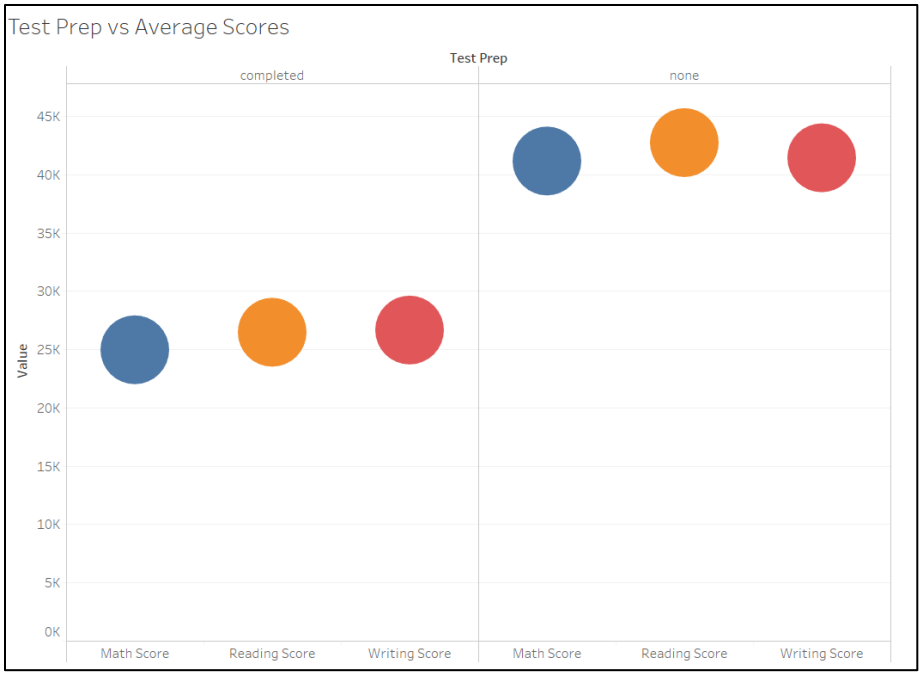


- We clearly see that there is no mean structure present nor there is any non – constant variance present
    - This suggests that even though we apply transformations – we do not get a b etter model with the current variables
    - That is – the variability of the model that can be explained cannot be increasi ng any further with the current explanatory variables

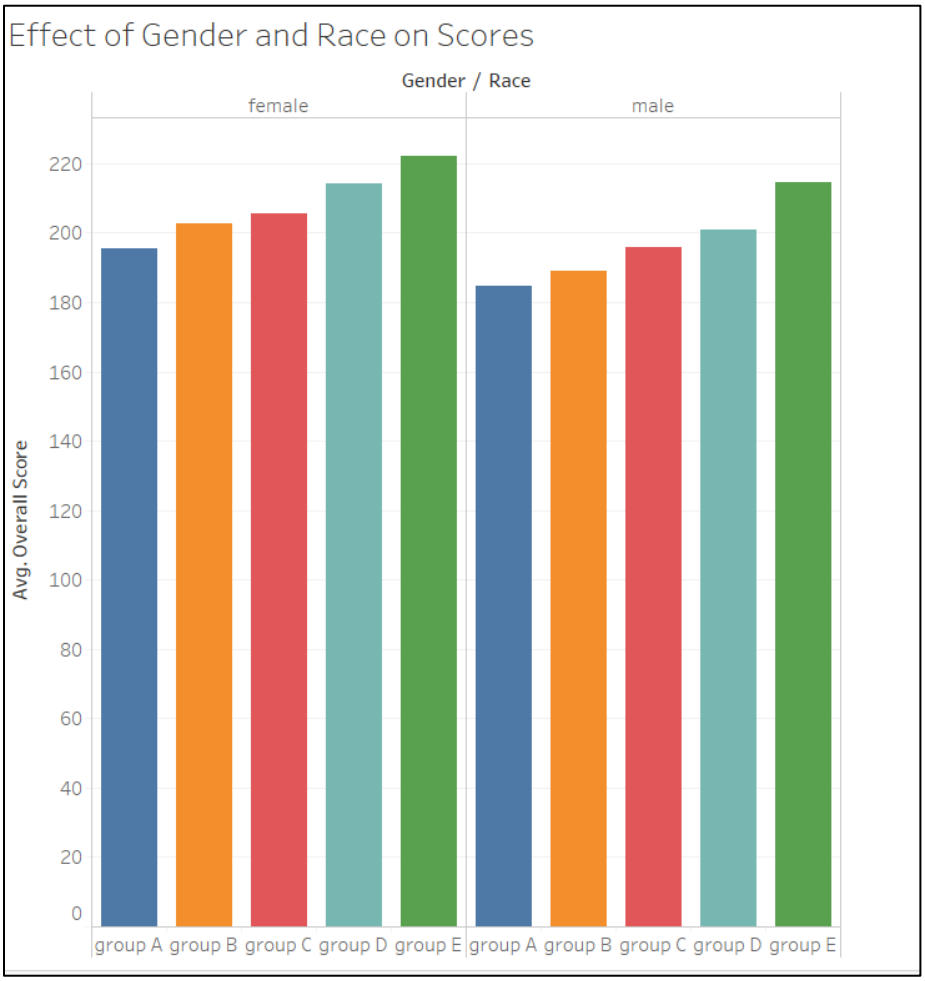
We also plotted some graphs in Tableau to be able to get some insights –



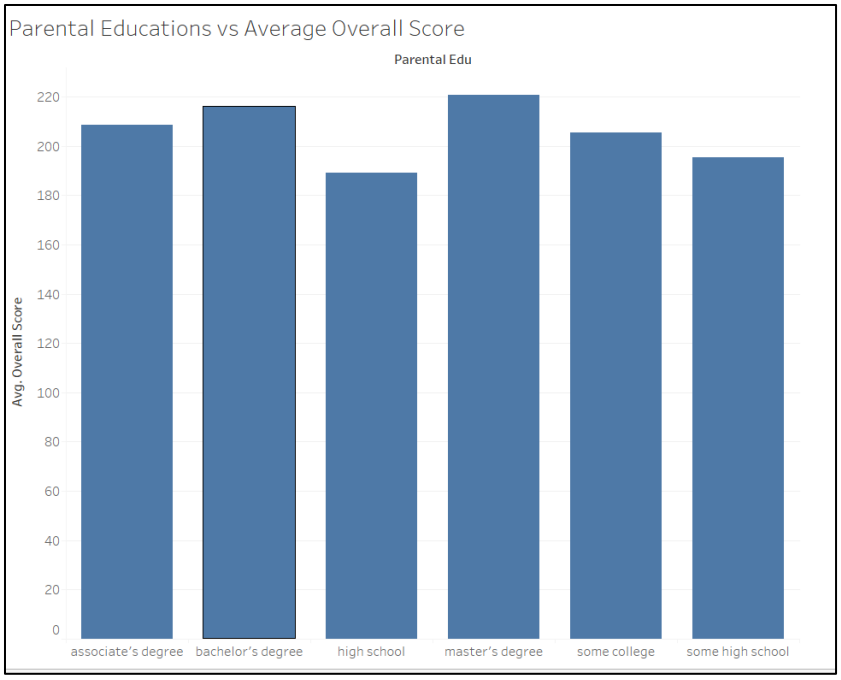
We see that the average scores are higher when the type of lunch prov ided is standard as compared to fre e/ reduced lunch.



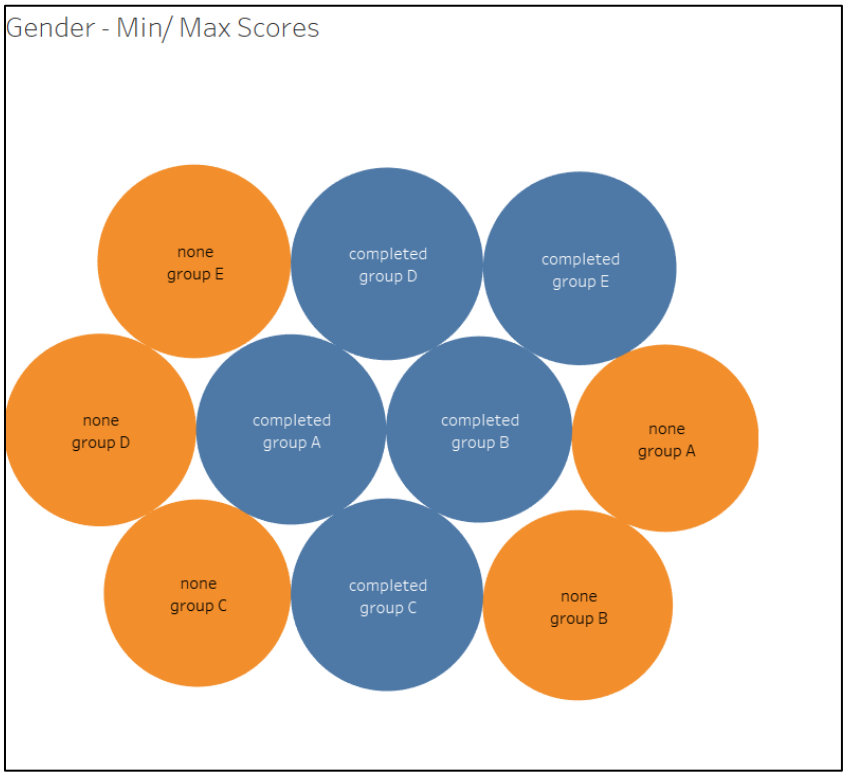
This is a surprising insi ght, where we see that the average scores are higher for people who have not completed th e test prep.



We see that both gend ers have the highest score for people who belong to the race – group E



There is a correlation between parents' education and the average overall score in the tests of the students



The highest score is obtained by students who have not completed test prep and from race E. Similar scores were also seen for students who have completed test prep and belong to group D or group E.

Finally, we can conclude that the variance in overall score cannot be explained with the current dataset. However, we have some interesting insights which we mentioned throughout the course of the report.