# PROGRAMMING FOR BIG DATA

PROJECT REPORT

VIDEO GAMES SALES AND STUDENT PERFORMANCE

MUHAMMAD NAZIM MUSHTAQ X18132723

## 1. Objectives of the Project:

- **Video Games Sales**

In the period of technology advancement which has increased the demand of the video games among the population. In the article https://www.wepc.com/news/video-game-statistics/ which state that in next two year 2020 the video game market will touch the 90 billion of the market worth compared to 78.61 billions of 2017 and Asian pacific is emerging and biggest market for the gaming industry. Similarly https://www.statista.com/topics/868/video-games/ state that in the year 2017 the gaming industry market worth of 18.4 billions in the US and Nintendo is the most popular brand.

- **Student Performance**

The performance of the student is one of the important research field for the engineers and researchers to analyze the student dataset in order to understand the influencing factors affect the performance of the students. In the article https://www.kenyaplex.com/resources/13223-factors-that-affect-students-performance-in-high-schools-in-kenya.aspx has state that there are the six major factors affecting the performance of the students in Kenya. The factors like students background ,Lack of school fees , Friends, Disagreements and quarrels with parents, Teacher-student relationship and participating of the students in class has been discussed similarly the article https://www.mckinsey.com/industries/social-sector/our-insights/drivers-of-student-performance-insights-from-europe state that there are three major subject to analyze the student performance are math, science and reading. This survey has been taken on 72 countries for 18000 schools and 140000 parents ,110000 teachers and 540000 students has participated.

## 2.Data Types:

- **Dataset 1. Video Games**

```
> str(video_game)
'data.frame':   16719 obs. of  16 variables:
 $ Name           : Factor w/ 11563 levels "","'98 Koshien",..: 11059 9406 5573 11061 7417 9771 6693 11057 6696 2620 ...
 $ Platform       : Factor w/ 31 levels "2600","3DO","3DS",..: 26 12 26 26 6 6 5 26 26 12 ...
 $ Year_of_Release: Factor w/ 40 levels "1980","1981",..: 27 6 29 30 17 10 27 27 30 5 ...
 $ Genre          : Factor w/ 13 levels "","Action","Adventure",..: 12 6 8 12 9 7 6 5 6 10 ...
 $ Publisher      : Factor w/ 582 levels "10TACLE Studios",..: 371 371 371 371 371 371 371 371 371 371 ...
 $ NA_Sales       : num  41.4 29.1 15.7 15.6 11.3 ...
 $ EU_Sales       : num  28.96 3.58 12.76 10.93 8.89 ...
 $ JP_Sales       : num  3.77 6.81 3.79 3.28 10.22 ...
 $ Other_Sales    : num  8.45 0.77 3.29 2.95 1 0.58 2.88 2.84 2.24 0.47 ...
 $ Global_Sales   : num  82.5 40.2 35.5 32.8 31.4 ...
 $ Critic_Score   : int  76 NA 82 80 NA NA 89 58 87 NA ...
 $ Critic_Count   : int  51 NA 73 73 NA NA 65 41 80 NA ...
 $ User_Score     : Factor w/ 97 levels "","0","0.2","0.3",..: 79 1 82 79 1 1 84 65 83 1 ...
 $ User_Count     : int  322 NA 709 192 NA NA 431 129 594 NA ...
 $ Developer      : Factor w/ 1697 levels "","10tacle Studios",..: 1035 1 1035 1035 1 1 1035 1035 1035 1 ...
 $ Rating         : Factor w/ 9 levels "","AO","E","E10+",..: 3 1 3 3 1 1 3 3 3 1 ...
```

Dataset has 8 factors and 8 numerical and integer columns with 16719 observations and 16 variables.

Missing values: - the dataset consists of the 10% of the missing values. The columns contain the missing values mentioned below with the number of the missing values present in it.

| Name | Platform | Year_of_Release | Genre | Publisher | NA_Sales | EU_Sales | JP_Sales |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Other_Sales** | **Global_Sales** | **Critic_Score** | **Critic_Count** | **User_Score** | **User_Count** | **Developer** | **Rating** |
| 0 | 0 | 8582 | 8582 | 0 | 9129 | 0 | 0 |

- **Dataset 2. Student performance**

```
> str(Student_data)
'data.frame':   1000 obs. of  8 variables:
 $ gender                     : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 2 2 1 ...
 $ race.ethnicity             : Factor w/ 5 levels "group A","group B",..: 2 3 2 1 3 2 2 2 4 2 ...
 $ parental.level.of.education: Factor w/ 6 levels "associate's degree",..: 2 5 4 1 5 1 5 5 3 3 ...
 $ lunch                      : Factor w/ 2 levels "free/reduced",..: 2 2 2 1 2 2 2 1 1 1 ...
 $ test.preparation.course    : Factor w/ 2 levels "completed","none": 2 1 2 2 2 2 1 2 1 2 ...
 $ math.score                 : int  72 69 90 47 76 71 88 40 64 38 ...
 $ reading.score              : int  72 90 95 57 78 83 95 43 64 60 ...
 $ writing.score              : int  74 88 93 44 75 78 92 39 67 50 ...
>
```

Dataset has 5 factors and 3 integer variables with 1000 observations and 8 variables.

Missing Values: - there is no missing value present in the dataset. Its cleaned dataset

| gender | race.ethnicity | parental.level.of.education | lunch |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| **test.preparation.course** | **math.score** | **reading.score** | **writing.score** |
| 0 | 0 | 0 | 0 |

**3. Approach of the Project:**

- **Dataset 1. Video Games Sales**

**Introduction:** In this project we are going to use the dataset related to the video games sales and its has been taken from the public dataset forum called Kaggle. The size of the dataset is 1.5 MB and it contain 16 columns and 16719 rows.

Link of the dataset: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

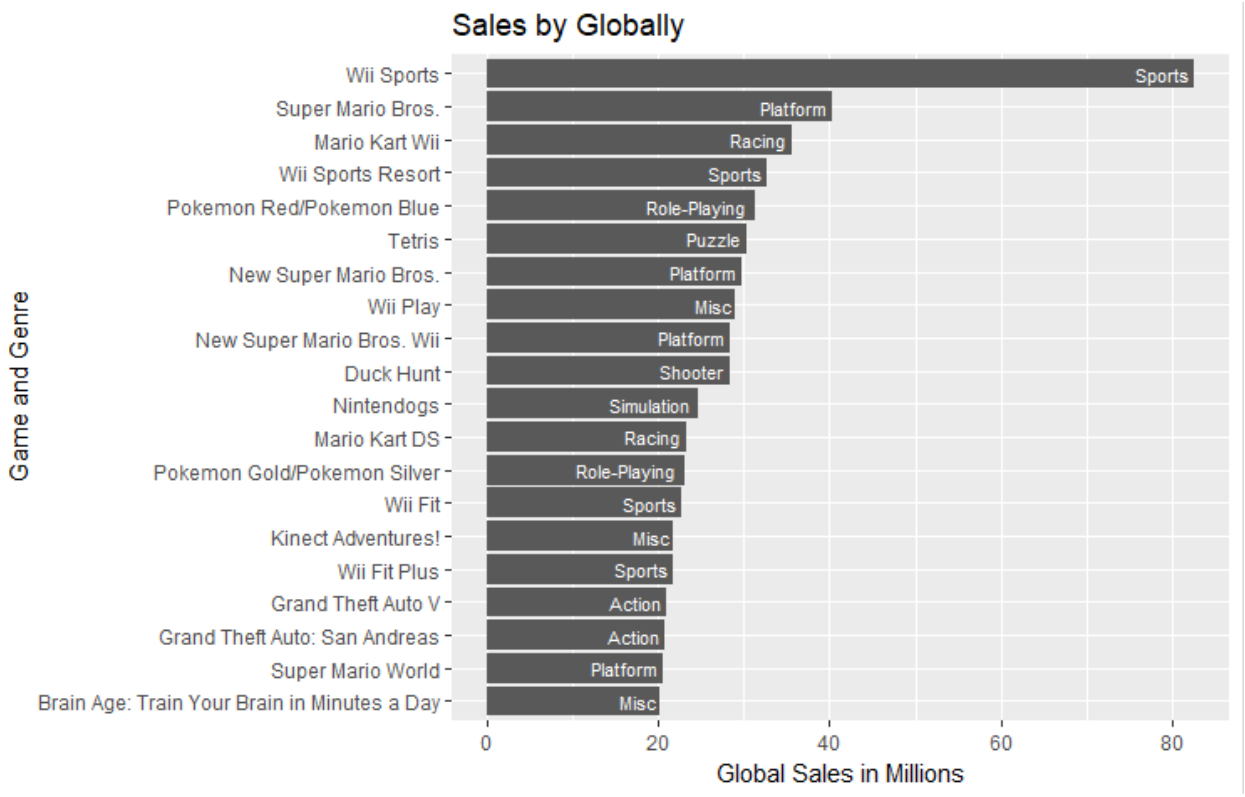
The list of the attributes present in the dataset has been given below: -

- **Data Description: -**

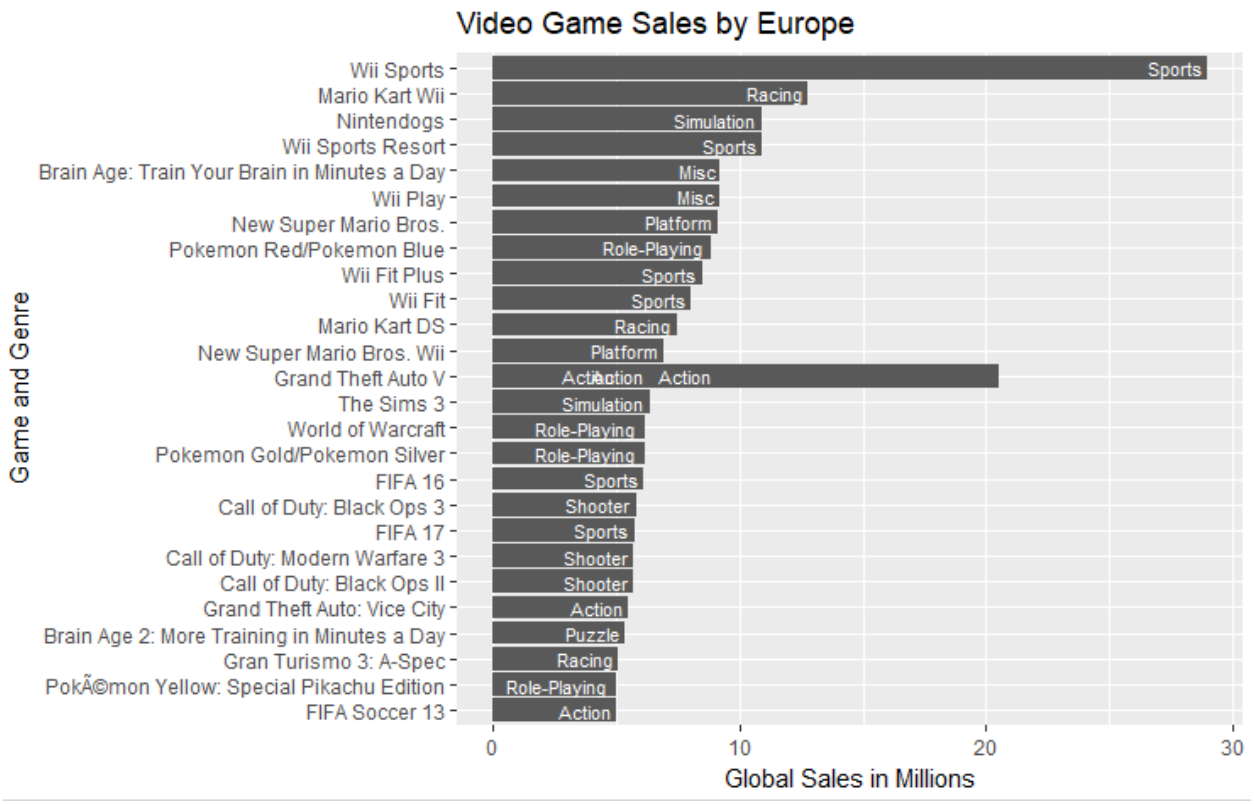
1. **Name:**
2. **Platform:**
3. **Year_of_Release:**
4. **Genre:**
5. **Publisher:**
6. **NA_Sales:**
7. **EU_Sales:**
8. **JP_Sales:**
9. **Other_Sales:**
10. **Global_Sales:**
11. **Critic_Score:**
12. **User_Score:**
13. **User_Count:**
14. **Developer:**
15. **Rating:**

- **Analysis: -**

1. **Sales of the Video Games Globally and in Europe**

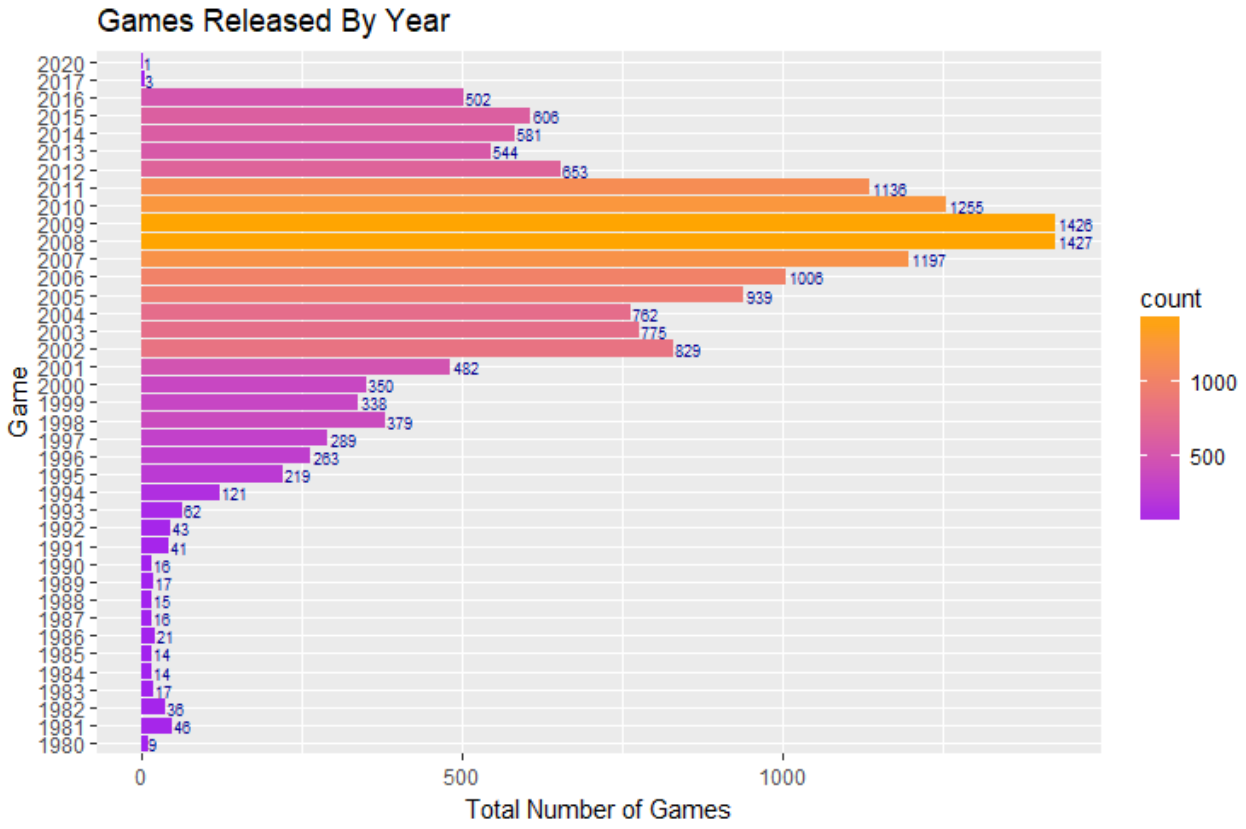


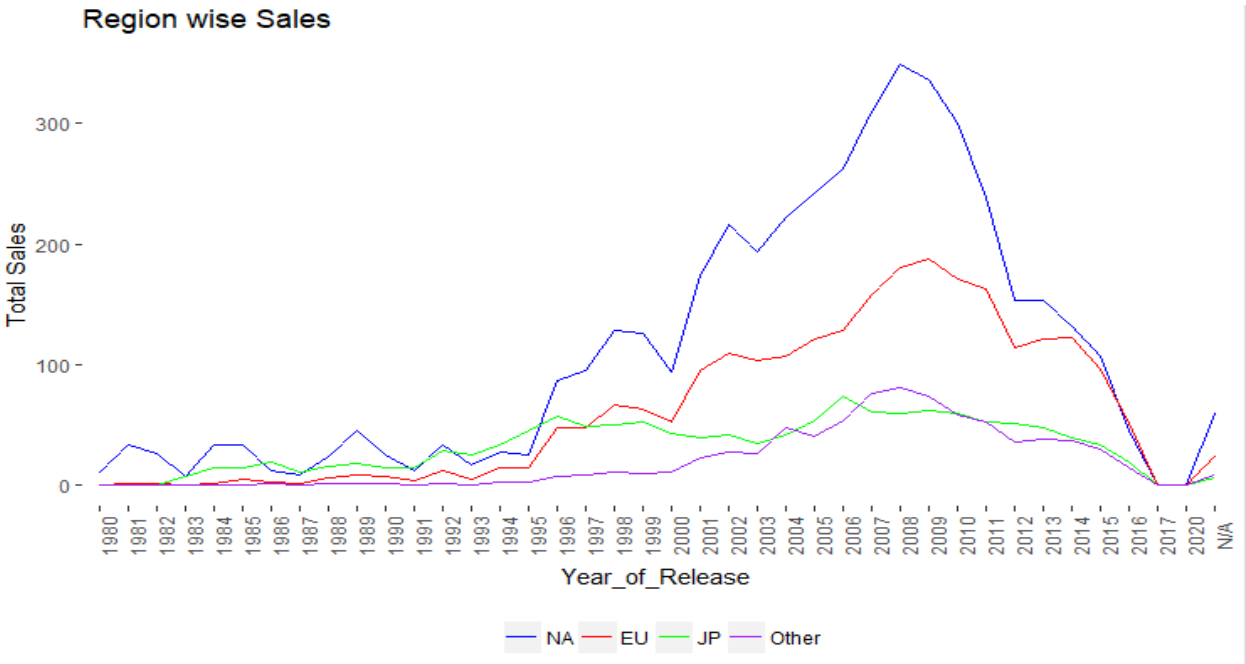

The above plot represents the sales of the video games globally above 10 million in volume. By analyze it we can see it that Wii sports has been sold out more than 10 million  globally and Super Mario Bros is second highest globally sold out game.




Similarly, In the above plot on Europe is showing that Wii sports has been sold out highest with total volume more then 5 million and Mario Kart Wii is the second highest game has been sold in the Europe.

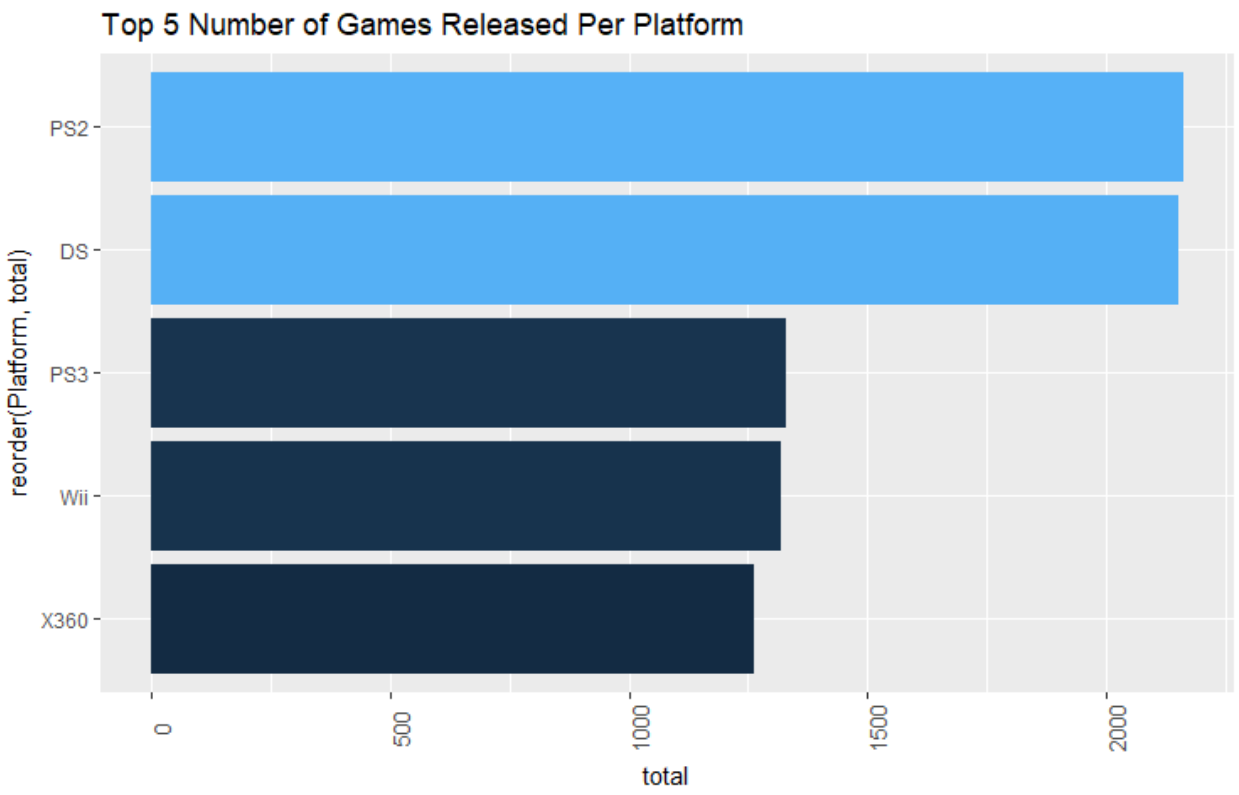
## 2. Released of the Games per Years & Regions



In this analysis we are trying to analyze from the above plot is the number of the release game by year in which we can see that year 2008 and 2009 has experienced the highest number of the release game compared to the other years whereas during the year 2016-17 has suffered huge declined in the number of the released game. In the second plot, we are analyzing the sales of the games by region which include America, Europe, north America and other regions across the world. In the plot we can see that NA has highest number of the sale of the games during the year 2008 and 2009 compared to another region. Europe is the second region which has highest number of the sales of the games among other regions.

3. **Top 5 Games released per platform**




In the above plot, we are analyzing top 5 platform on which highest number of games has been released which include PS2, DS, PS3,Wii and X360 compared to the other available platforms. The platform PS2 and DS has equal number of the release of the games equal to 2350.

## 4. Statistical test

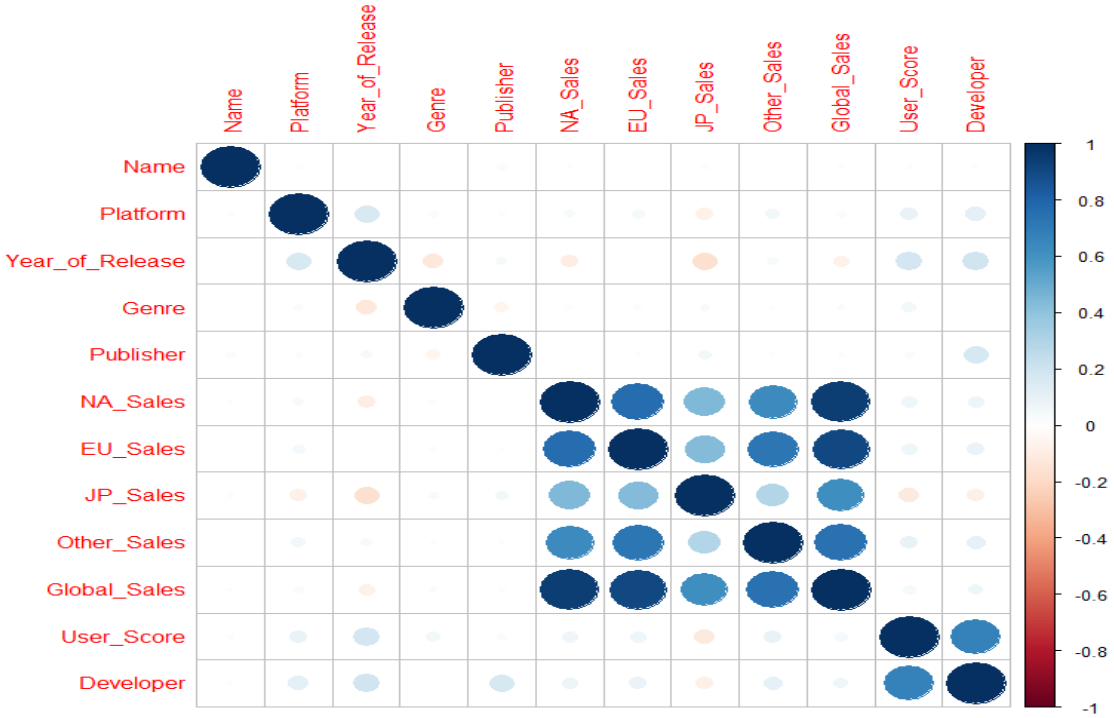
- One-way Anova test
  This will give us the summary of the anova test on the given available global sale~ platform. The degree of freedom for the Platform is 16688, mean is 48.35 and F value is 20.9 where as Pr (>F) values is less then <0.005

```
> summary(anova_test)
               Df Sum Sq Mean Sq F value               Pr(>F)
Platform       30   1451   48.35    20.9 <0.0000000000000002 ***
Residuals   16688  38607    2.31
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- T test
  This test on the variable global sales and European sales will tell us that P value is < less than 0.005.

```
> t.test(video_game$Global_Sales,video_game$EU_Sales) # where y1 and y2 are numeric

        Welch Two Sample t-test

data:  video_game$Global_Sales and video_game$EU_Sales
t = 30.863, df = 20213, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3638437 0.4131920
sample estimates:
mean of x mean of y
0.5335427 0.1450248
```

- **Correlation Plot**




By correlation plot, we can clearly analyze that NA_Sales is highly correlated with Global_Sales and EU_Sales is highly correlated with the Global_Sales.

- **Dataset 2: Student Performance dataset**

**Introduction:** Student performance dataset has been used for the to analysis the performance of the students on the different parameters which includes Math, reading and writing. The dataset has been taken from the public dataset forum Kaggle which consists of the 1000 rows and 8 columns and size of 70.3 KB.

Link of the dataset: https://www.kaggle.com/spscientist/students-performance-in-exams

The lists of the attributes of the dataset have been mentioned in the data description.

Dataset description:

1. Genders
2. Race.Ethnicity
3. Parental.level.of.education
4. Lunch

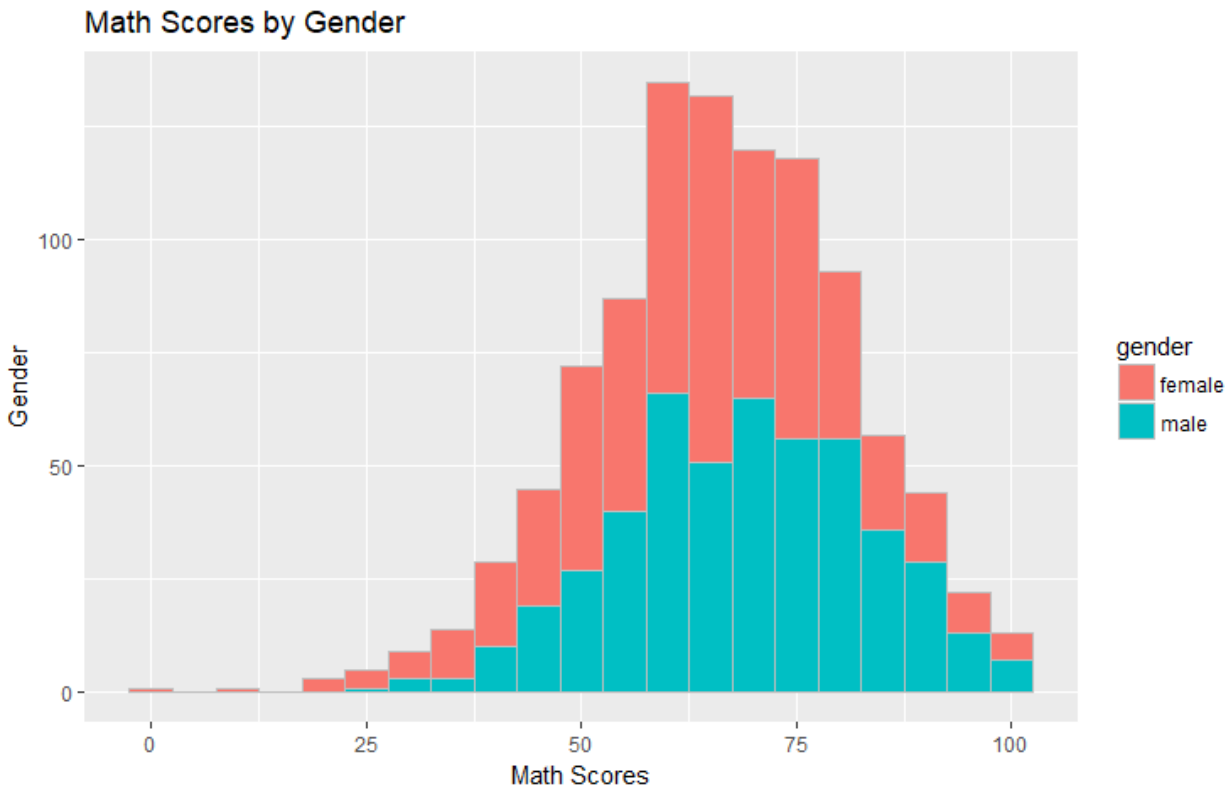
5. Test.preparation.course
6. Math.score
7. Reading.score
8. Writing.score

- **Analysis**

**1. Score by Genders**

1.1 Math score by gender

In the given plot we can analyze that female compared to the male has maximum distribution of the marks between 45-80 as well as female has highest marks then male in the math.

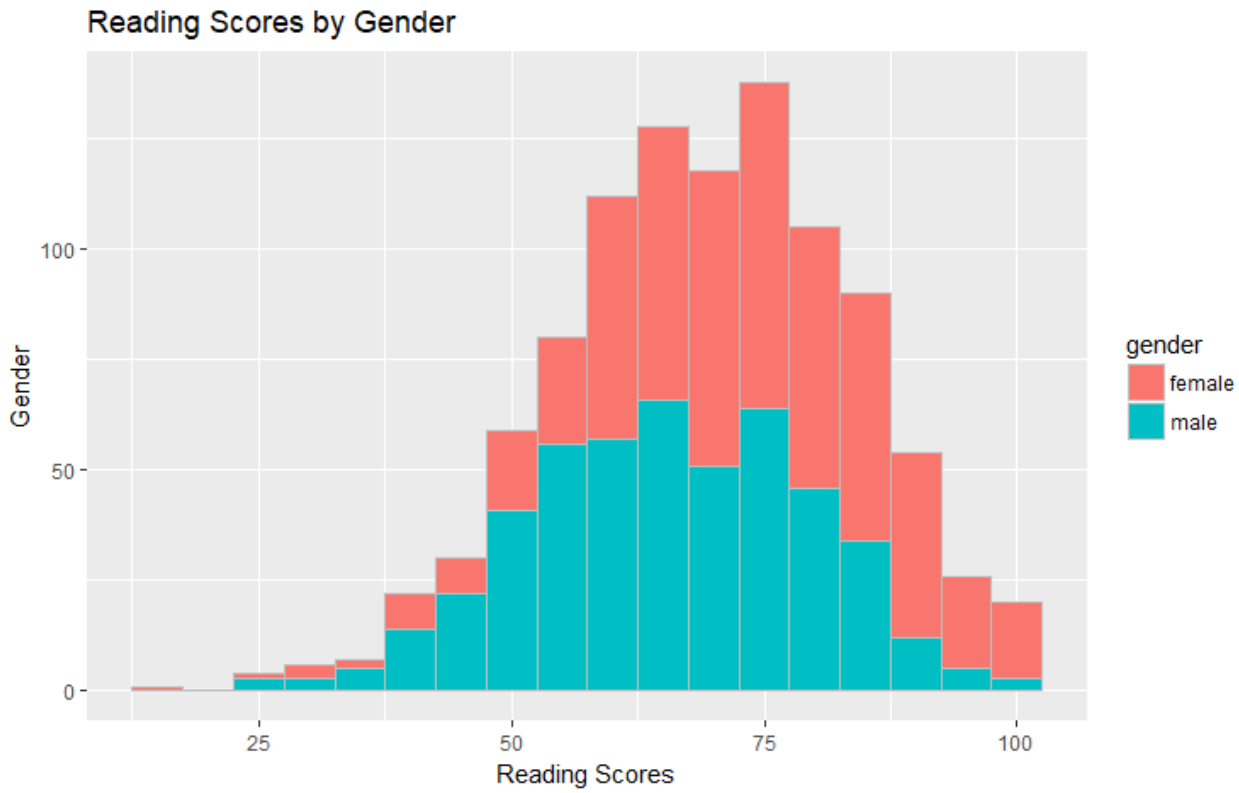



**1.2 Write Score by Gender**

In the given plot we can analyze that female compared to the male has maximum distribution of the marks between 55-80 as well as female has highest marks then male in the written scores.


### 1.3 Reading Score by Gender

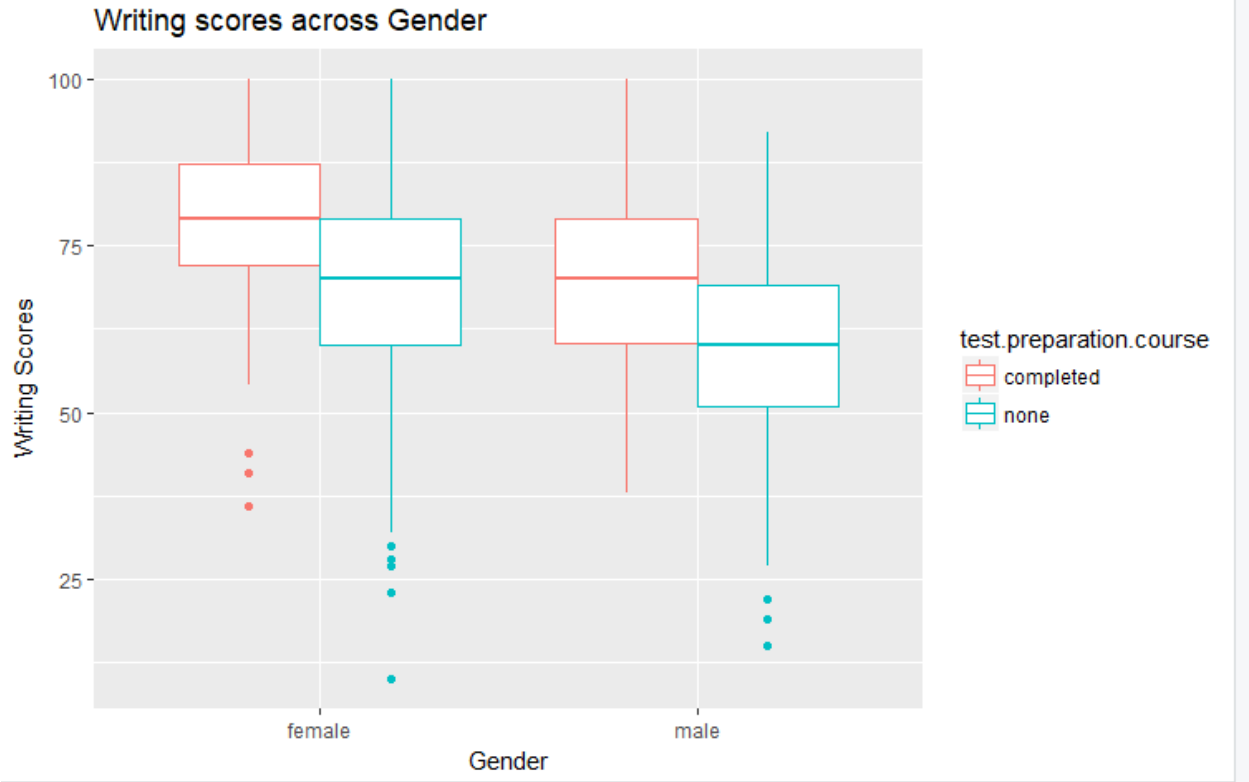
In the given plot we can analyze that female compared to the male has maximum distribution of the marks between 60-80 as well as female has highest marks then male in the reading scores.




## 2. Scores and Test Prep by Gender

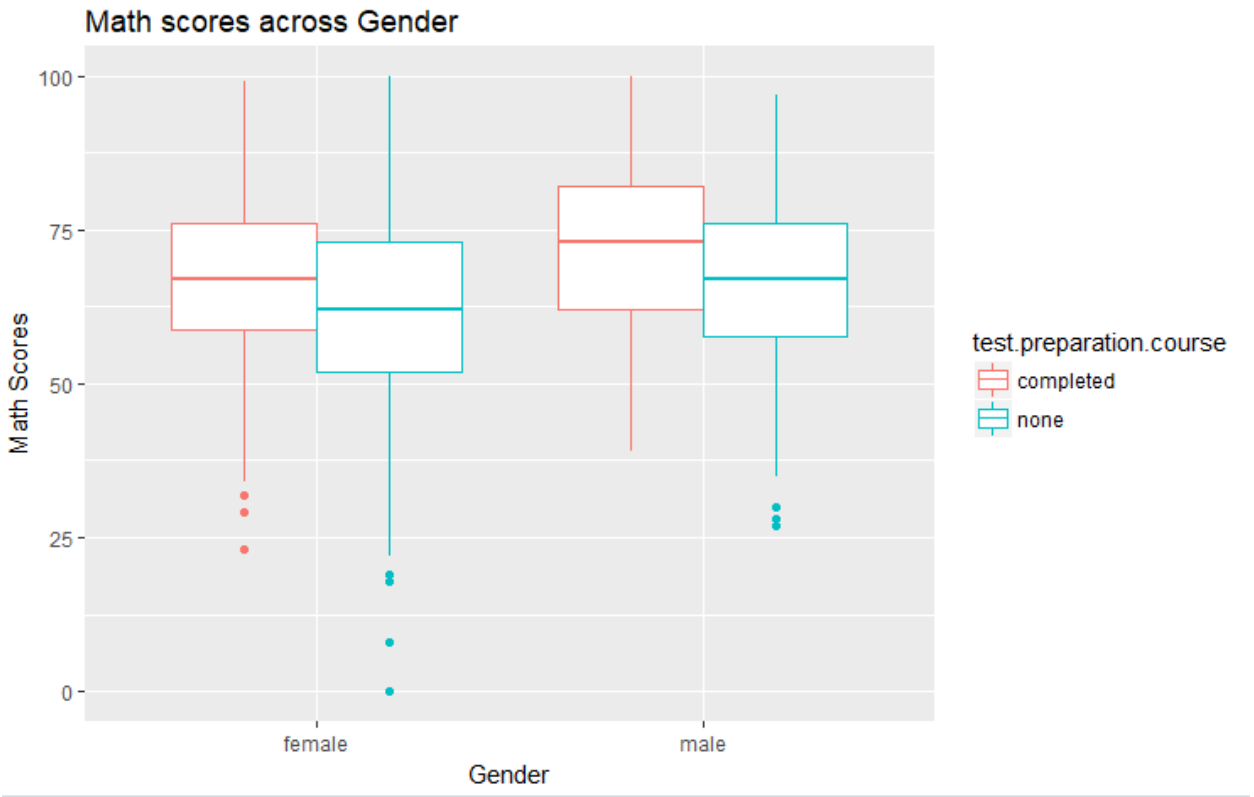
### 2.1 Writing

In the given plot we can analyze that female compared to the male has higher test competence in writing.




### 2.2 Math

In the given plot we can analyze that female compared to the male has lesser test competence in math.

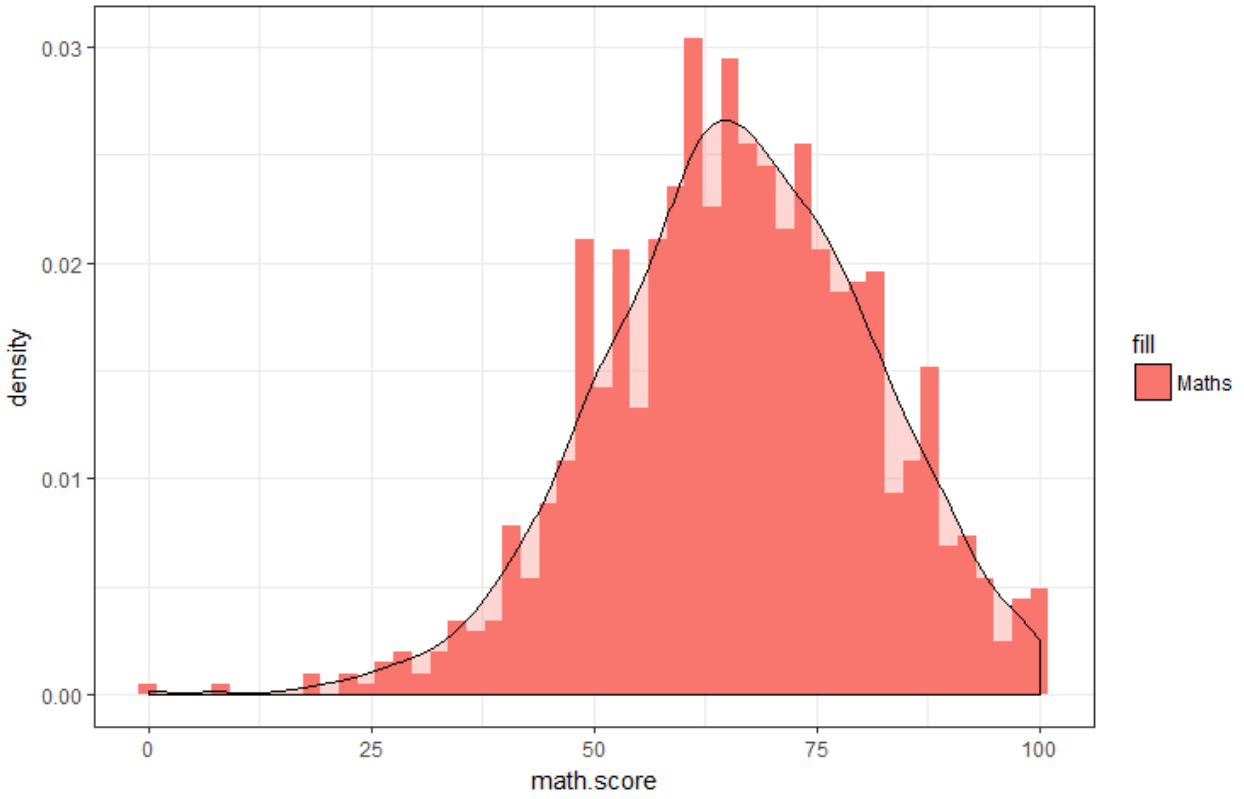



### 2.3 Reading

In the given plot we can analyze that female compared to the male has higher test competence in reading.

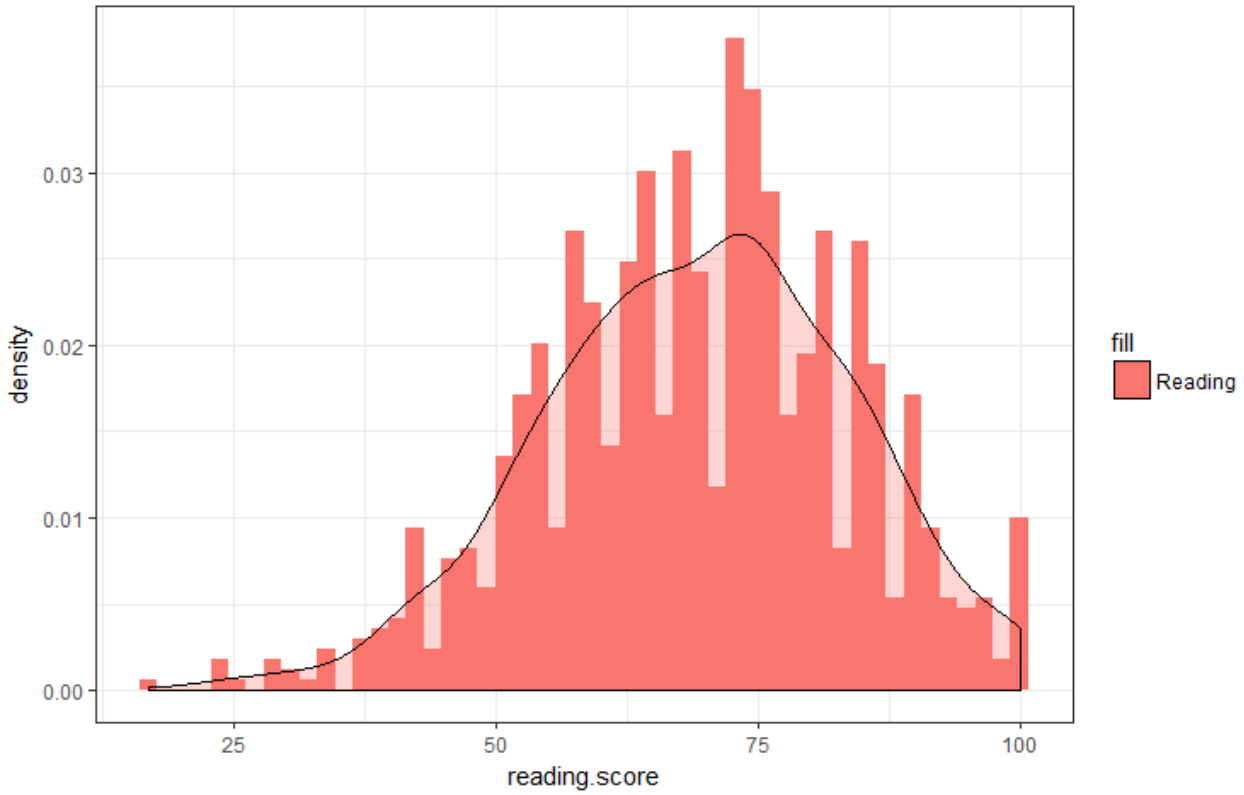

## 3. Normal distribution

### 3.1. Math Score



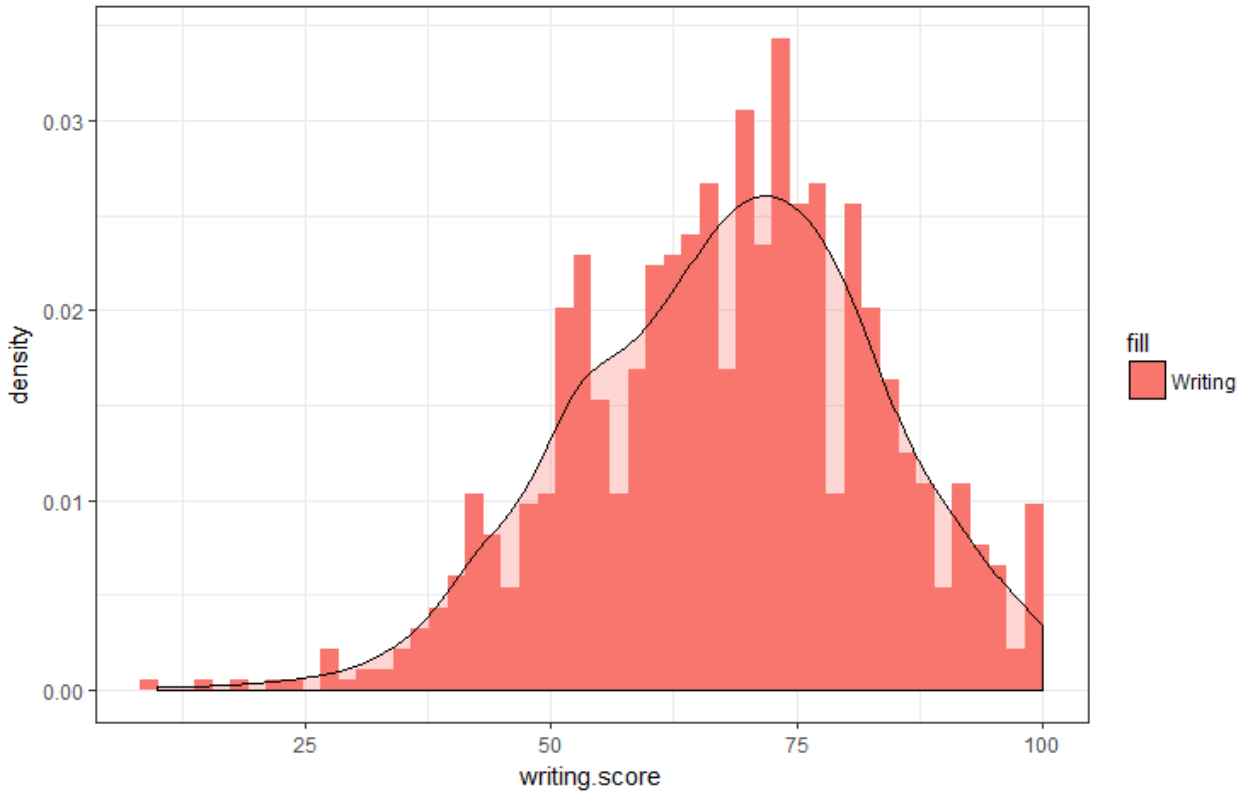
In the above plot, the normalization of the math score has been shown in which we can analyze that maximum peak distribution of the dataset has been concentrate between the score of 50 – 75.

### 3.2 Reading




In above plot, we are plotting a normalization chart for the reading score. We can clearly analyze that maximum distribution of the reading score has been concentrated between 65-80.

### 3.3 Writing




In the above plot, we are plotting the normalization plot for the writing score. We can analyze from the plot that maximum concentration of the score in writing has been between the range of 55-80.

## 4. Statistical test

### 4.1.Linear Regression

Liner regression define the relationship between the independent and dependent variable to understand the influence of the independent variable on the dependent variable and build a prediction line to make best fit in terms of prediction.

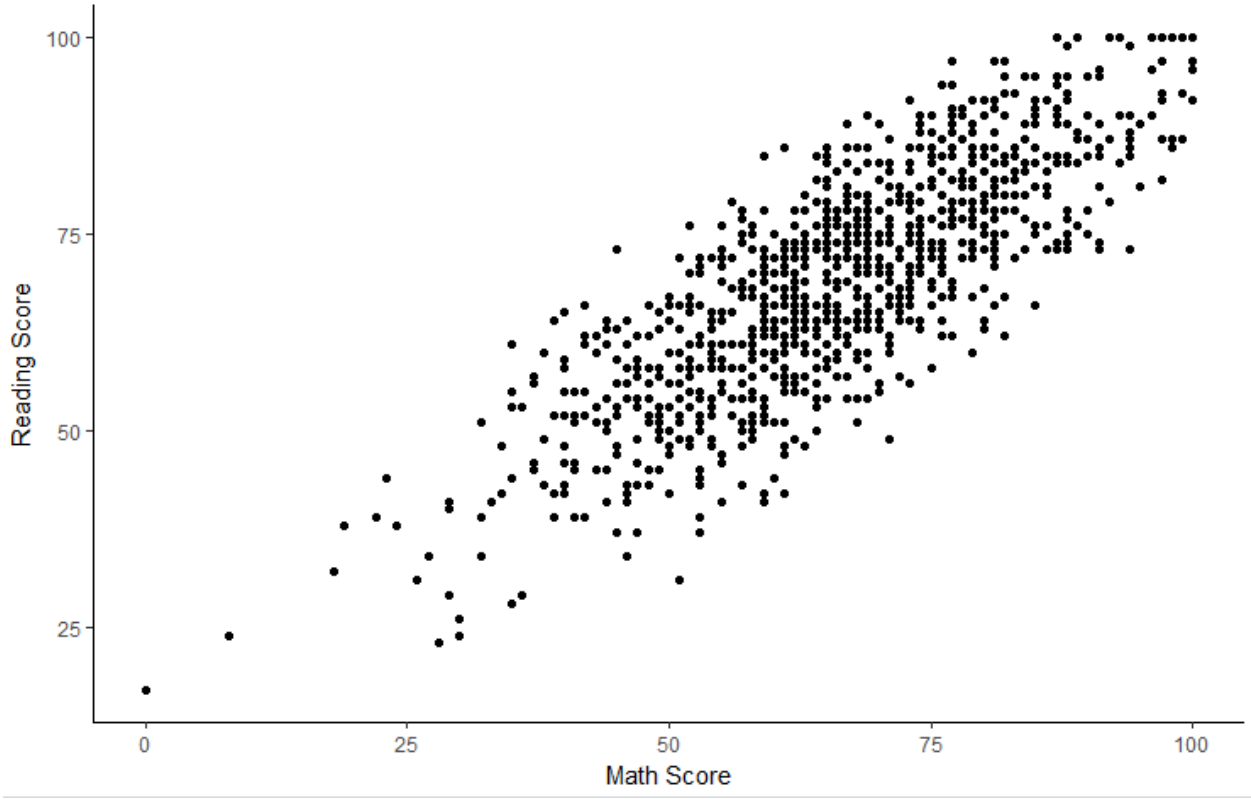
$$Y = a + bX,$$

Where Y is the dependent variable, X is the independent variable and a,b are the slope and intercept.

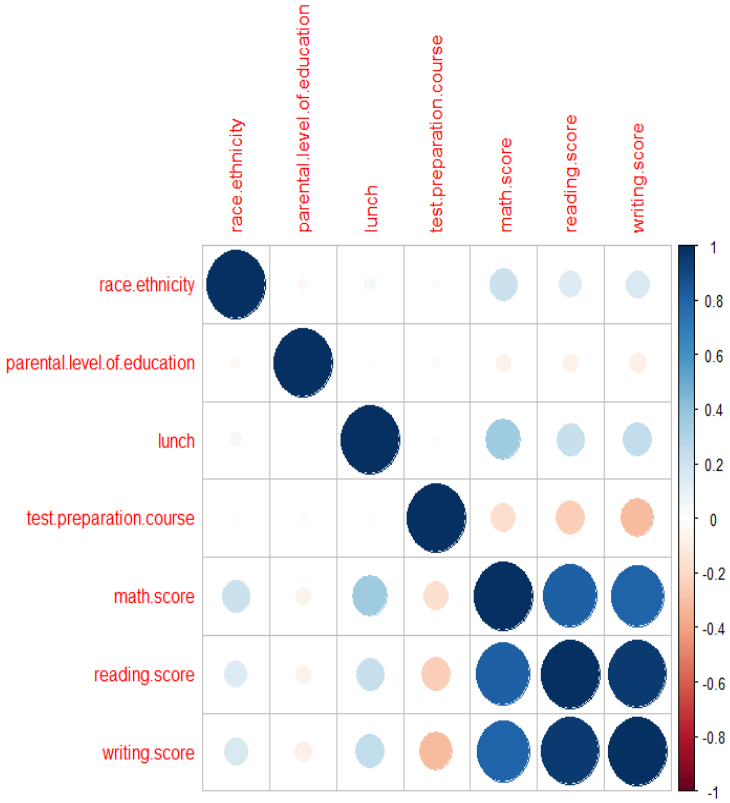
In this project, we have built the linear regression model between math score and reading score where we get the intercept score of 7.3 and reading score of 0.84 with the p value.

```
> liner_model=lm(math.score~reading.score, data=Student_data)
> summary(liner_model)

Call:
lm(formula = math.score ~ reading.score, data = Student_data)

Residuals:
     Min       1Q   Median       3Q      Max
-24.3419  -6.3419  -0.0221   6.2713  24.6581

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)    7.35759    1.33818   5.498          0.0000000487 ***
reading.score  0.84910    0.01893  44.855 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.736 on 998 degrees of freedom
Multiple R-squared:  0.6684,    Adjusted R-squared:  0.6681
F-statistic:  2012 on 1 and 998 DF,  p-value: < 0.00000000000000022
```

## 5. Correlation



Correlation between math and reading score shows that there is high correlation for the score between math and reading between 50-75 which we can easily see in the second below plot.




In the above plot, we can clearly see that there is a high correlation between reading score and writing score similarly there is a high correlation between math score and writing score.

**Challenges:**

In this project, searching a big dataset was one of the challenges with understanding of dataset to build a project on it as well as the functionality of each libraries which has been used in the project.

**References:**

1. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

2. http://www.sthda.com/english/articles/32-r-graphics-essentials/128-plot-time-series-data-using-ggplot/

3. https://www.statmethods.net/stats/correlations.html

4. https://www.statmethods.net/stats/anova.html

5. https://www.statmethods.net/stats/ttest.html

6. https://www.kaggle.com/monkey09/studentperformance-exploratory-predictions