Data Visualisation CIA 4

Mini-Project using R and Tableau

On

Analysis of Student's performance

Submitted by

Anshika Batra (Reg No - 1920334)

Thomas Sharon (Reg No – 1920362)

Devanaboina Rohit (Reg No- 1920181)

Under the guidance of

Prof. Raghavendra N

SCHOOL OF BUSINESS AND MANAGEMENT,

CHRIST (Deemed to be University),

BANGALORE - 560 029

2020

ACKNOWLEDGEMENT

We, Anshika Batra, Thomas Sharon and Devanaboina Rohit, would like to express our profound gratitude to all those who have been instrumental in the preparation of this Data Visualisation Mini-Project Report. We wish to place on record our deep gratitude to our project guide, Prof. Raghavendra N, for guiding us through this project with valuable and timely advice.

We would like to thank Dr. (Fr). Abraham V M, Vice Chancellor Dean and Dr. Amalnathan S, HOD, School of Business and Management for their encouragement.

Last but not the least, we would like to thank our parents and friends for their constant help and support.

CONTENTS

1) Introduction of the study

Education is the process of facilitating learning and acquiring relevant knowledge. The purpose of schooling is to achieve an educational goal directed towards learning. These academic achievements are quintessential for the successful development of youth in the society. Academic performance includes various factors such as intellectual level, personality, motivation, skills, interests. Students who perform well in school tend to have an easier transition into adulthood in terms of maintaining stability through occupational and economic success. Students' academic performance is affected by various factors such as their learning ability, parental background, quality of teaching, infracture., etc. Various studies have been conducted to validate this premise and identify the key factors that contribute towards the academic performance of students under different circumstances to draw out constant determinants of the same.

In this study, we will be analysing a dataset of students with circumstantial differences, to understand the correlation between academic performance and other independent factors.

2) Statement of the Problem

The main purpose of this study is to assess student's performance keeping in mind various factors like gender, ethnicity, lunch type, etc. Another important domain in this study is the parental level of education and how it affects the simultaneous performance of the students. The performance can be divided into math score, reading score and writing score and how students belonging to different races and gender do in such assessments. The report, Analysis of Student Performance, provides detailed analysis of test preparation proficiency rates, broken down by different areas of study and parental level of education. This study aims to highlight the difference in performance of peers coming from different backgrounds and family education history.

3) Objectives of the Study

- Analysing the relative performance of Male and Female students
- Analysing the differences in academic performance of various ethnic groups
- Analysing the relationship between parents' education and childrens' academic performance

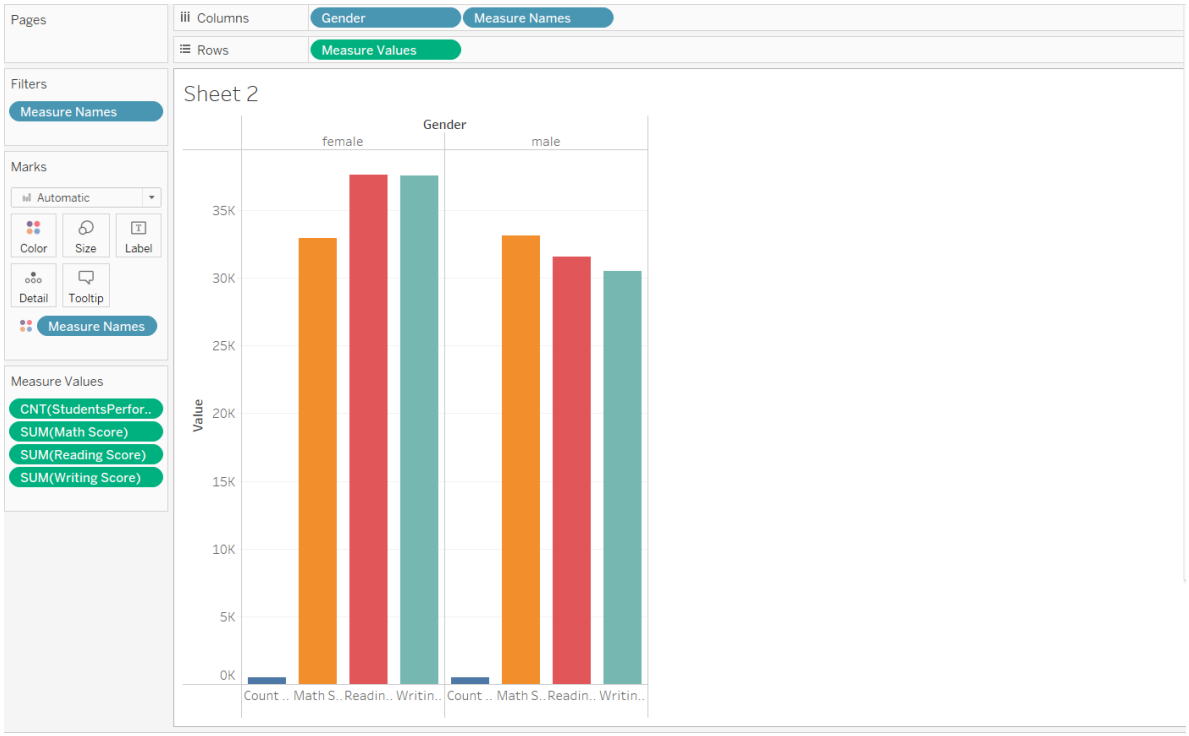
4) Data sources

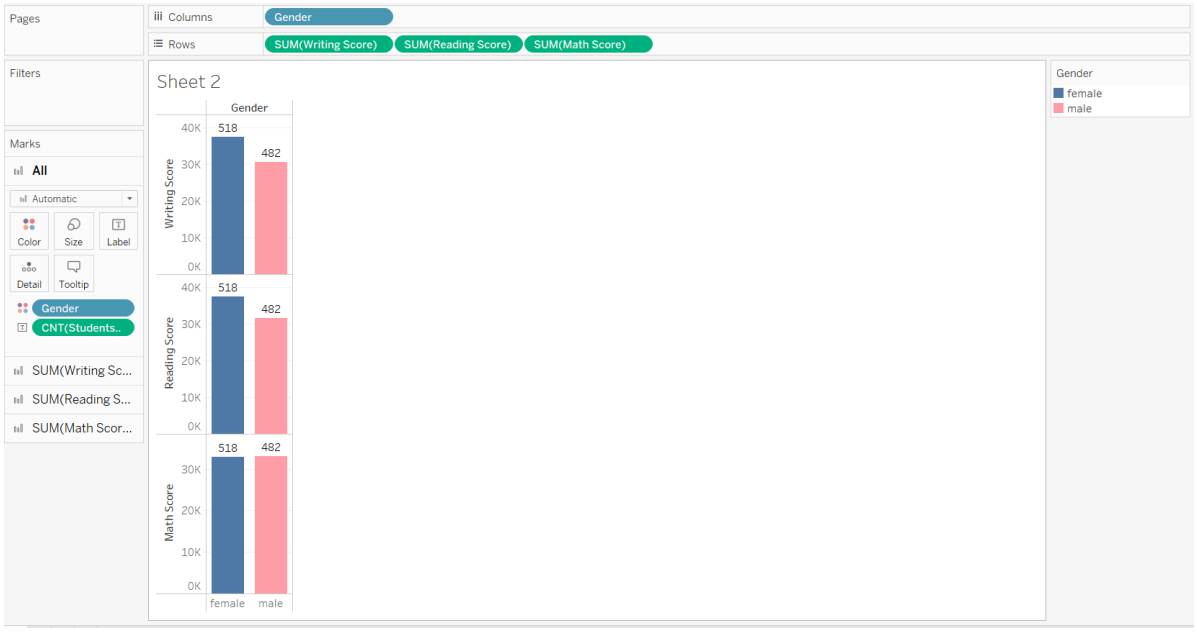
https://www.kaggle.com/spscientist/students-performance-in-exams

5) Analysis of the data using Tableau and R

Tableau Analysis

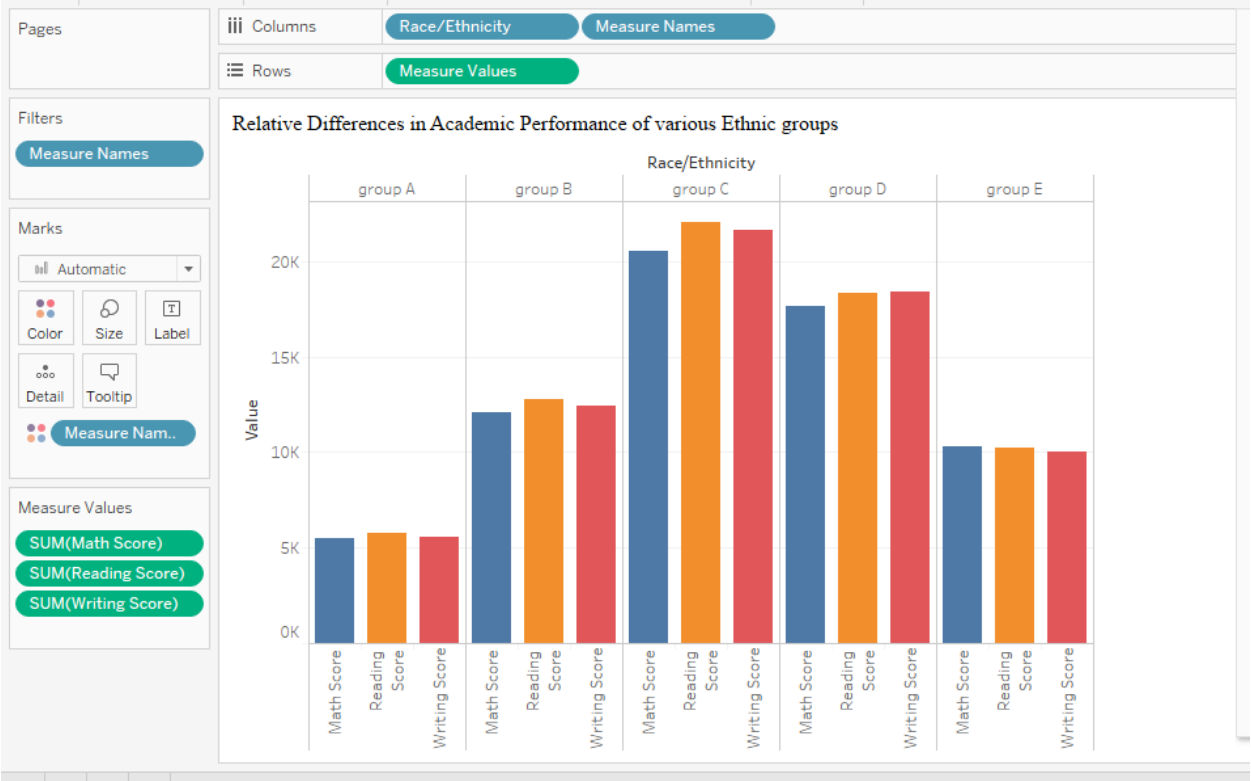
1. Analysing the relative performance of Male and Female students





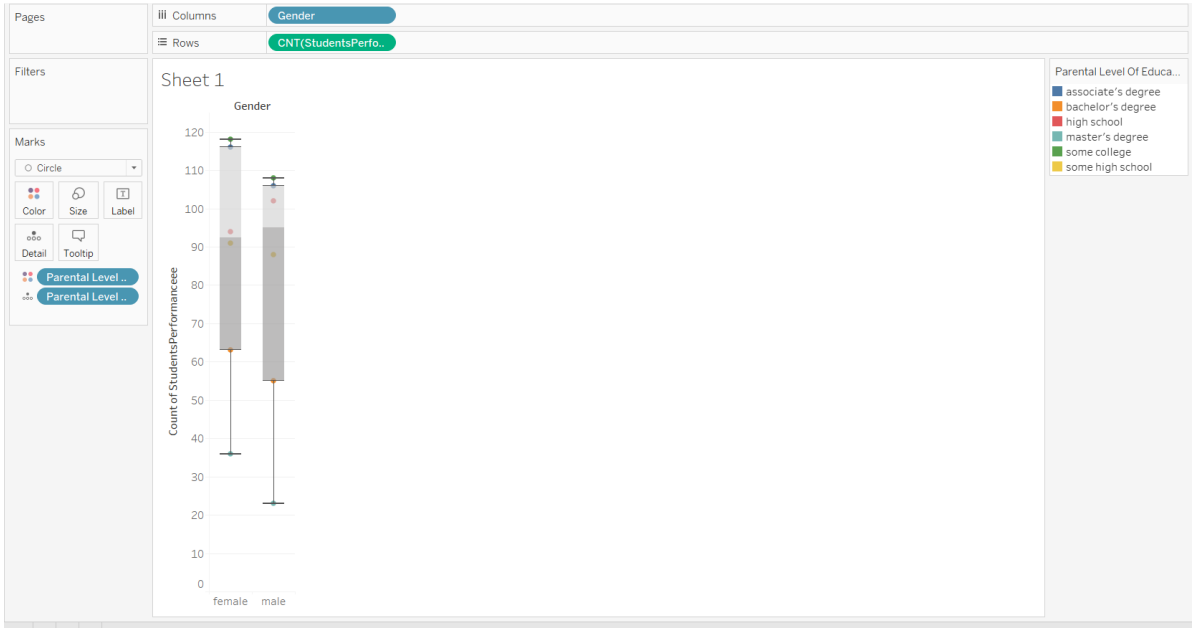
Explanation: The graph highlights the performances of 518 females and 482 males. A higher reading and writing score in females is clearly seen as compared to males in Reading and Writing. However, there is not much difference when comparing the math score of males and females.

2. Identifying the differences in academic performance of various ethnic groups



Explanation: The Graphs compares the academic performances of 5 different ethinc groups across the 3 scores taken. The analysis shows the distinctively strong performance of group C over the other ethnic groups while group A displayed the worst performance in all 3 sets of scores.There is no similarity between the ethnic groups for any of the scores taken.

3. Analysing the relationship between parents' education and childrens' performance



Explanation: The graph analyses the relationship between parental level of education and student's performance. There is a similar relationship between associate's degree and some college when comparing the student's performance based on parental level of education. Master's degree stands at the last with the least number of students under that parental level of education.

Descriptive Statistics & Hypothesis Testing using RStudio

Descriptive Statistics

```
> summary(student_df)
    gender     race.ethnicity     parental.level.of.education          lunch     test.preparation.course   math.score
 female:518   group A: 89    associate's degree:222          free/reduced:355   completed:358           Min.   :  0.00
 male  :482   group B:190    bachelor's degree :118          standard    :645   none     :642           1st Qu.: 57.00
              group C:319    high school       :196                                                     Median : 66.00
              group D:262    master's degree   : 59                                                     Mean   : 66.09
              group E:140    some college      :226                                                     3rd Qu.: 77.00
                             some high school  :179                                                     Max.   :100.00
 reading.score    writing.score
 Min.   : 17.00   Min.   : 10.00
 1st Qu.: 59.00   1st Qu.: 57.75
 Median : 70.00   Median : 69.00
 Mean   : 69.17   Mean   : 68.05
 3rd Qu.: 79.00   3rd Qu.: 79.00
 Max.   :100.00   Max.   :100.00
>
```

- There are 1000 observations (students) with the above breakdown by
    - Gender
    - Ethnicity
    - Parental Level of Education
    - Lunch received on day of exam
    - Test Preparation Course usage
    - Math Score
    - Reading Score
    - Writing Score

Hypothesis Testing:

Hypothesis 0: Test for Normality

As most of our analysis will be based on the Math Scores of students, we must first confirm that the same is normally distributed - a prerequisite for further statistical analysis

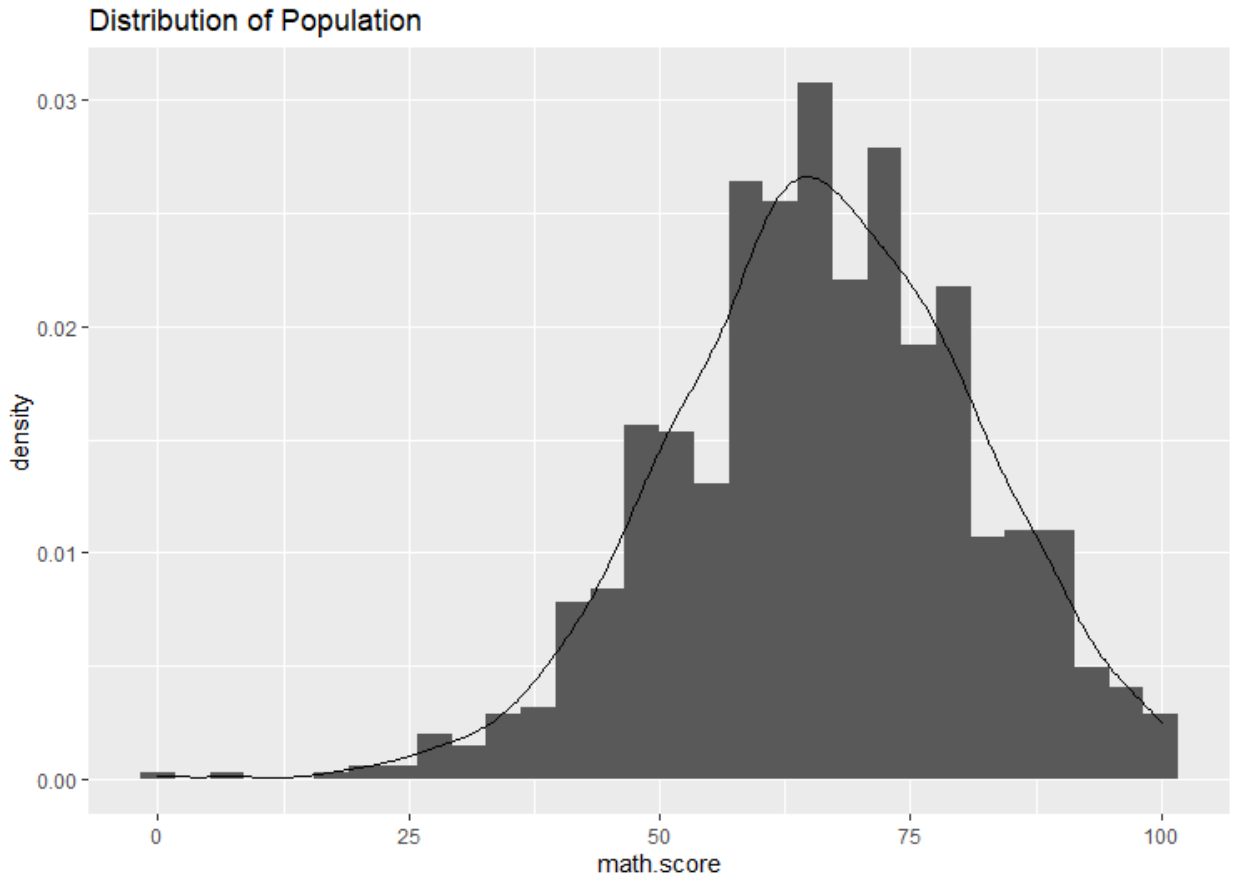
H0: There is no statistical difference between the sample mean and population mean of students' Math score

H1: There is a statistical difference between the sample mean and population mean of students' Math score.

Significance Level: 0.05 (95% confidence interval)

Analysis:

Visualising shape of Math Score data



Interpretation: Visual Inspection shows that the data is normally distributed

Comparing the Means of Population and Sample:

```
> random_sample <- student_df %>%
+   sample_n(30)
> population_mean <- mean(student_df$math.score)
> population_mean
[1] 66.089
> sample_mean <- mean(random_sample$math.score)
> sample_mean
[1] 65.83333
```

T-Test on Sample and Population:

```
> t.test(random_sample$math.score, mu = mean(student_df$math.score))

        One Sample t-test

data:  random_sample$math.score
t = -0.071123, df = 29, p-value = 0.9438
alternative hypothesis: true mean is not equal to 66.089
95 percent confidence interval:
 58.48130 73.18537
sample estimates:
mean of x
 65.83333
```

Interpretation: The means of the population and sample are very similar. The T-Test determines that P-statistic > 0.05 (significance level). Therefore, H1 is rejected and H0 is accepted.

Conclusion: There is no statistical difference between the population and sample sets of Math Scores → **The Data is normally distributed**

Hypothesis 1 - Male-Female performance difference in Reading

H0: There is no difference between the average reading score of male and female students.

H1: There is a difference between the average reading score of male and female students.

Significance Level: 0.05

Analysis:

Creating Male and Female Subsets

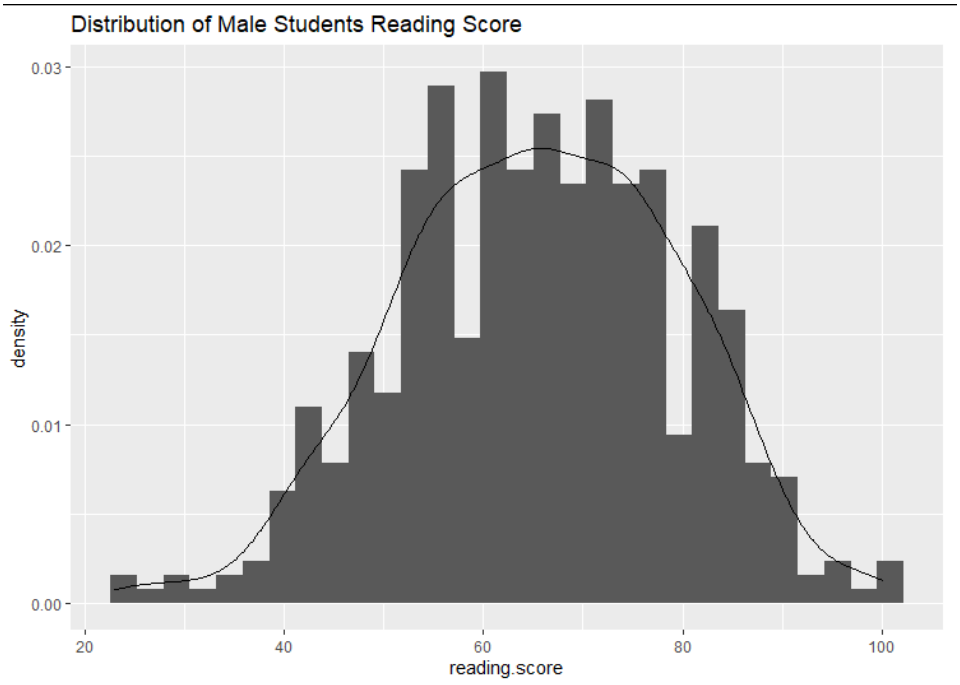
```
> male_students <- student_df %>%
+   filter(gender == "male")
>
> female_students <- student_df %>%
+   filter(gender == "female")
```

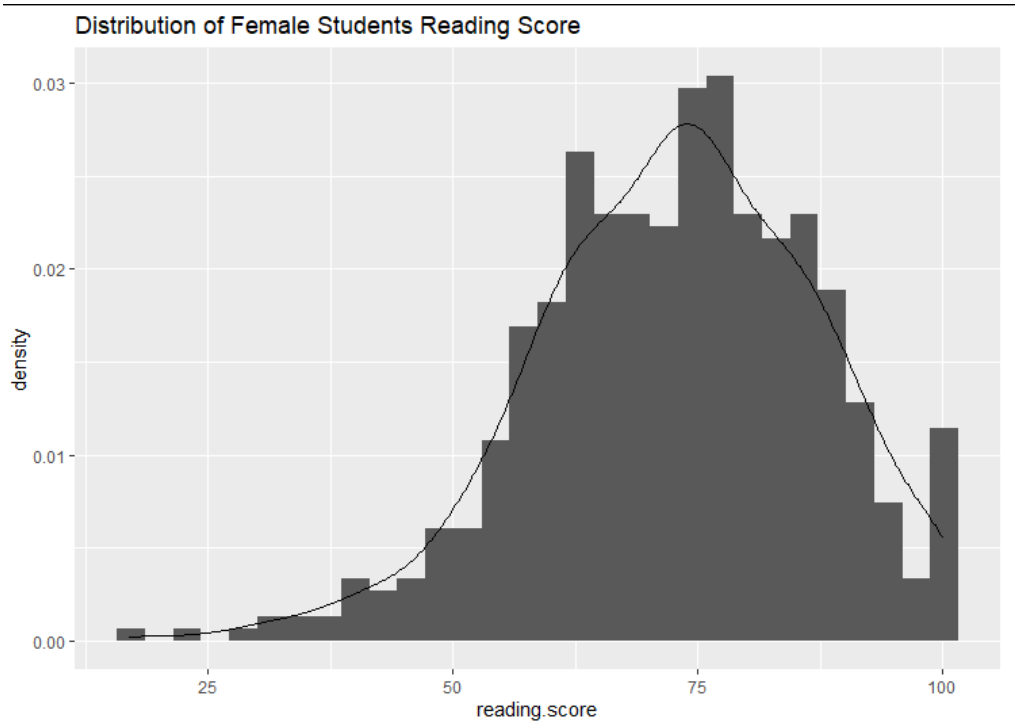
Comparing the Variances/SD of the two subsets

```
> var(male_students$reading.score)
[1] 194.0959
> var(female_students$reading.score)
[1] 206.7339
> sd(male_students$reading.score)
[1] 13.93183
> sd(female_students$reading.score)
[1] 14.37825
>
```

Interpretation: The Variances and Standard Deviations of the two sets are similar enough to conduct a Two-Tailed T-Test

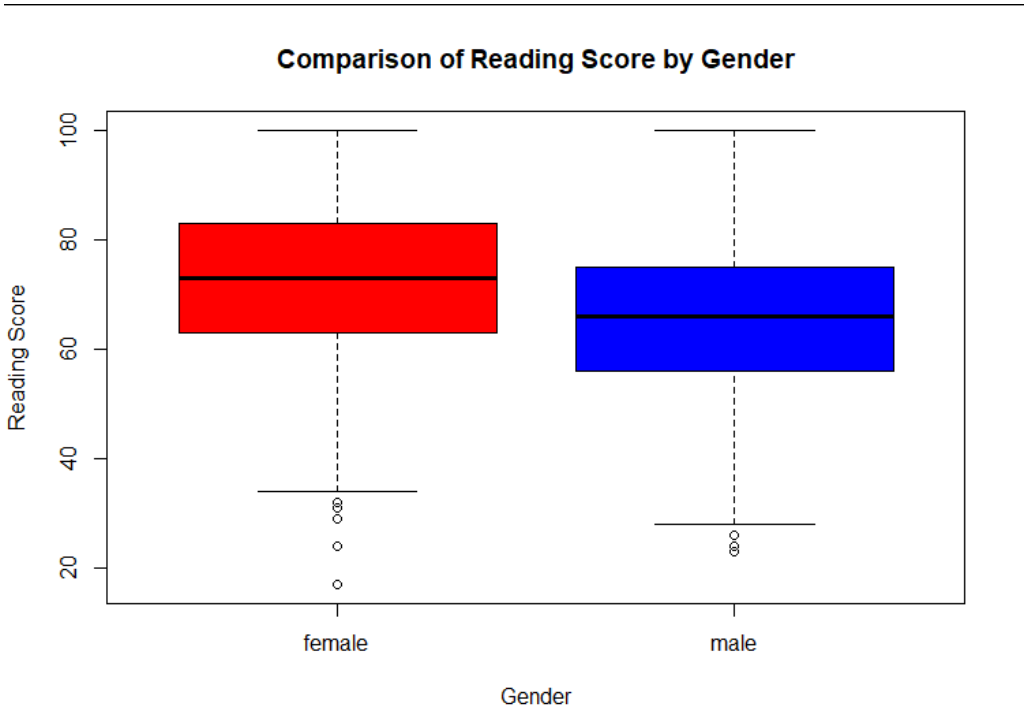
Checking for Normality of the two Sets





Interpretation: Visual Inspection shows that the data is normally distributed

Comparing Reading Scores



Interpretation: The bar plot shows female scores > male scores. We must conduct a T test to confirm.

T-Test to statistically prove hypothesis:

```
> t.test(reading.score ~ gender, data = student_df)

        Welch Two Sample t-test

data:  reading.score by gender
t = 7.9684, df = 996.36, p-value = 4.376e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.377941 8.892218
sample estimates:
mean in group female   mean in group male
            72.60811             65.47303
```

Interpretation: As p-value is > 0.05, so we can reject the null hypothesis. There exists a significant difference between the performance of male and female students vis-a-vis reading score.

Hypothesis 2 - Ethic Groups

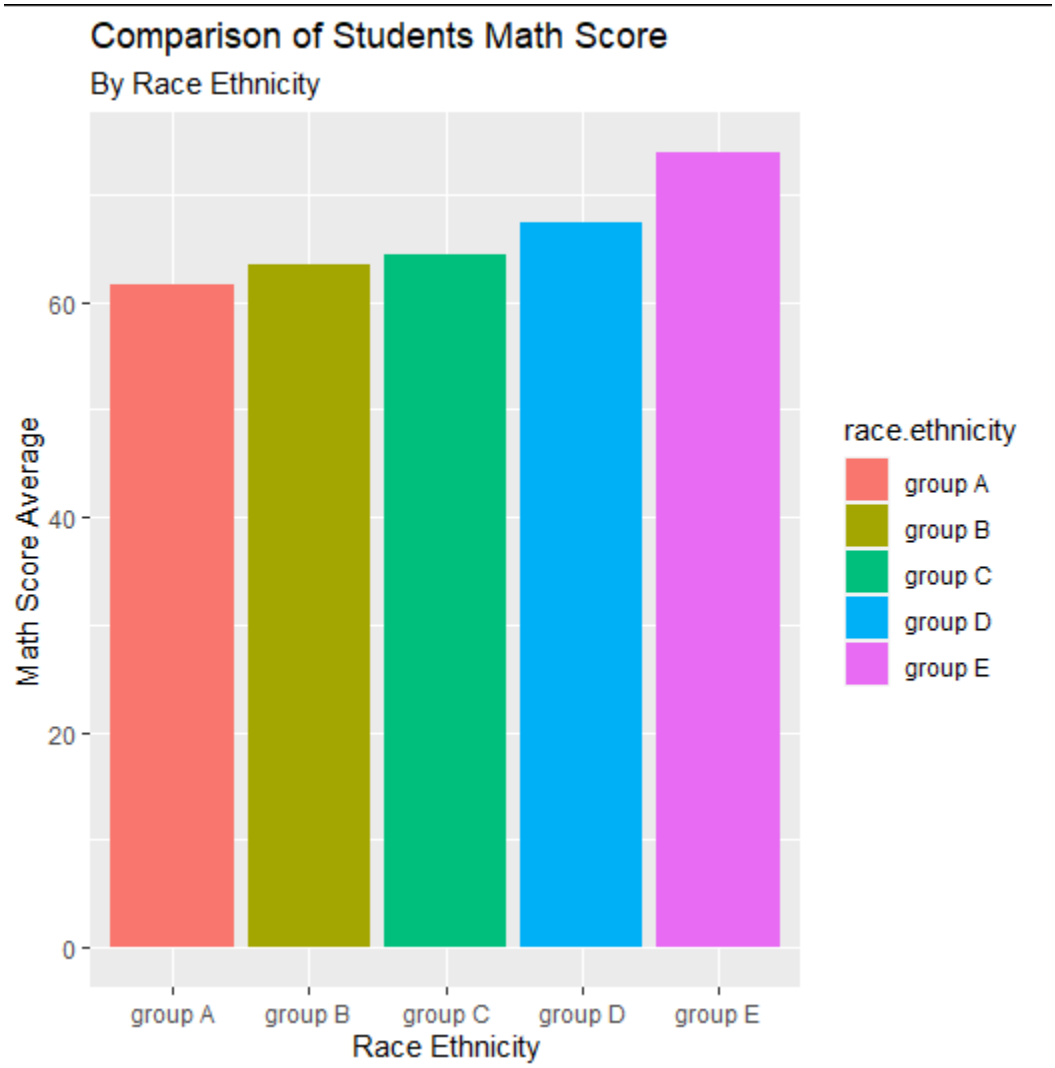
H0: There is no significant difference between the means of each ethnic group's math scores

H1: There is a significant difference between the means of each ethnic group's math scores

Significance Level - 95%

Analysis:

```
> by_race_ethnicity <- student_df %>%
+   group_by(race.ethnicity) %>%
+   summarise(math_score_mean = mean(math.score), math_score_var = var(math.score), math_score_sd = sd(math.score)) %>% arrange(desc(math_score_mean))
>
> by_race_ethnicity
# A tibble: 5 x 4
  race.ethnicity math_score_mean math_score_var math_score_sd
  <fct>                    <dbl>          <dbl>         <dbl>
1 group E                   73.8           241.          15.5
2 group D                   67.4           190.          13.8
3 group C                   64.5           221.          14.9
4 group B                   63.5           239.          15.5
5 group A                   61.6           211.          14.5
>
```



```
> aov(math.score ~ race.ethnicity, data = student_df)
Call:
   aov(formula = math.score ~ race.ethnicity, data = student_df)

Terms:
                race.ethnicity Residuals
Sum of Squares         12728.82 216960.26
Deg. of Freedom               4       995

Residual standard error: 14.76653
Estimated effects may be unbalanced
>
```

```
> summary(aov(math.score ~ race.ethnicity, data = student_df))
                Df Sum Sq Mean Sq F value   Pr(>F)
race.ethnicity   4  12729    3182   14.59 1.37e-11 ***
Residuals      995 216960     218
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Interpretation:

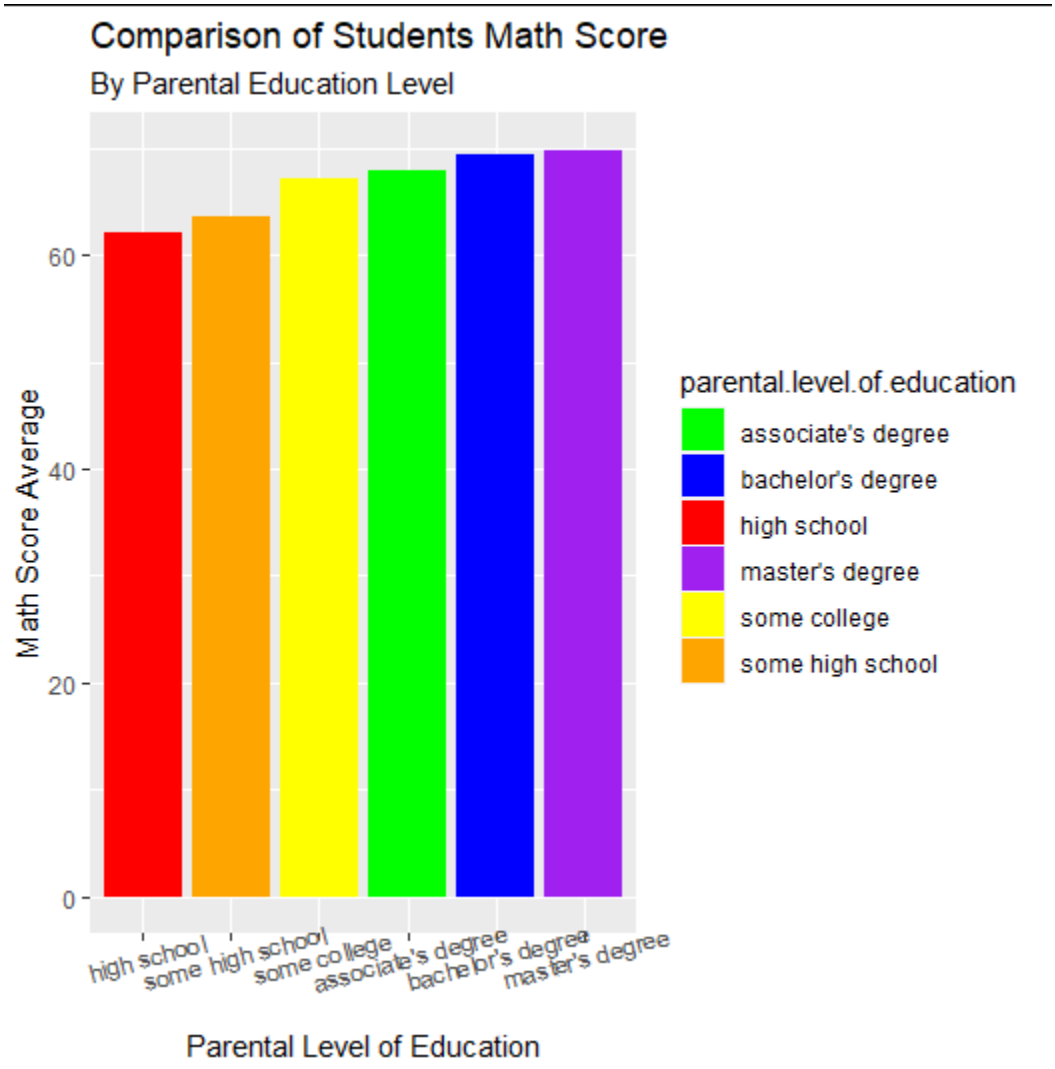
P-value is 1.37e-11 (which is much less than 0.05). Hence the Null hypothesis is rejected and H1 is accepted. There is a significant difference between the means of each ethnic group's math scores

Hypothesis 3 - Parents' Educational Background

H0: There is no significant difference between the means of each educational group's math scores

H1: There is a significant difference between the means of each educational group's math scores

Analysis




```
> summary(aov(math.score ~ parental.level.of.education, data = student_df))
                            Df Sum Sq Mean Sq F value   Pr(>F)
parental.level.of.education   5   7296  1459.1   6.522 5.59e-06 ***
Residuals                   994 222394   223.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Interpretation:

P-value is 5.59e-06 (which is much less than 0.05). Hence the Null hypothesis is rejected and H1 is accepted. There is a significant difference between the means of each educational group's math scores

6) Findings, suggestions and conclusion

- The data analysed is Normally Distributed, making it suitable for statistical analysis and hypothesis testing
- The Math & Reading Scores of students have significant relationships with their background
    - gender, ethnicity and family education backgrounds were the 3 variables tested, all of which proved to have a significant effect of student scores
- Female students have been shown to perform better in Reading than Male students, with significance proven through T-Test.
    - Females perform on average 10.5% better than males in Reading
- Certain Ethnic Groups perform better in Math than others
    - Group E has the highest average score at 73.82/100, and Group A the lowest at 61.62/100 - a nearly 20% difference. This indicates that Ethnicity is a strong predictor of test performance, although the relationship between the two has not been established in the above analysis
- Family education has a major influence on Math scores, with 13.8% difference between the best performing group - Master's Degree at 69.8 - and the worst performing group - High School at 62.1.
- The order of importance is Ethnicity > Family Education > Gender when it comes to students' performance in reading and math tests.

7) References

[1] One-Sample T-test in R - Easy Guides - Wiki - STHDA. (n.d.). STHDA. http://www.sthda.com/english/wiki/one-sample-t-test-in-r

[2] One-Way ANOVA Test in R - Easy Guides - Wiki - STHDA. (n.d.). STHDA.COM. http://www.sthda.com/english/wiki/one-way-anova-test-in-r

[3] Seshapanpu, J. (2018, November 9). *Students Performance in Exams*. Kaggle. https://www.kaggle.com/spscientist/students-performance-in-exams

[4] STHDA. (n.d.-a). *ggplot2 - Essentials - Easy Guides - Wiki - STHDA*. SDHTA.COM. http://www.sthda.com/english/wiki/ggplot2-essentials

[5] STHDA. (n.d.-b). *ggplot2 Box Plot Guide*. STHDA.Com. http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization

[6] Unpaired Two-Samples T-test in R - Easy Guides - Wiki - STHDA. (n.d.). STHDA.COM. http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r

8) Annexure

Dataset: https://www.kaggle.com/spscientist/students-performance-in-exams