Student Exam Performance Predictor

DATASET: Student Performance in Exam

Project Group No.04
By:Rohan Dange - A20517920
Neelam Borse - A20502809
Maithili Solanki - A20513656

BACKGROUND OF DATA

This data collection includes each student's marks in numerous disciplines as well as many other criteria that influence the student's exam performance. We have 5 categorical columns and 3 numerical columns.

General View of the Data

.Genders - Genders observed (Male & Female)

.Race/Ethnicity - Total of 5 groups divided into Group A,B,C,D,E.

.Parental level of education - Highest Level of education for parent is observed.

.Lunch - Students' usual dietary habits are observed.

.Test preparation course - Before the final exam, Student’s performance on the examination is tracked.

.Math score - Math score for each student is observed.

.Reading score - Reading score for each student is observed.

.Writing score - Writing score for each student is observed.

Purpose of this project

- Recognize variables and predict student’s exam achievement.
- Investigate the impact of parents' backgrounds, exam preparation, and other factors on student’s success.
- Determining the average grade of each student in each subject
- Using the many elements in the data set, forecast who will receive the highest score gender wise.

Data preparation and exploration summary

- Dimensions: 1000 Rows & 08 Columns.
- The missing values were found and treated by imputation of median value.
- There we no duplicate values found.
- The outliers of our response were found and treated.
- Multicollinearity was found and fixed.
- Average scores were determined group-wise utilizing each component.

Data Exploration(EDA): Boxplot

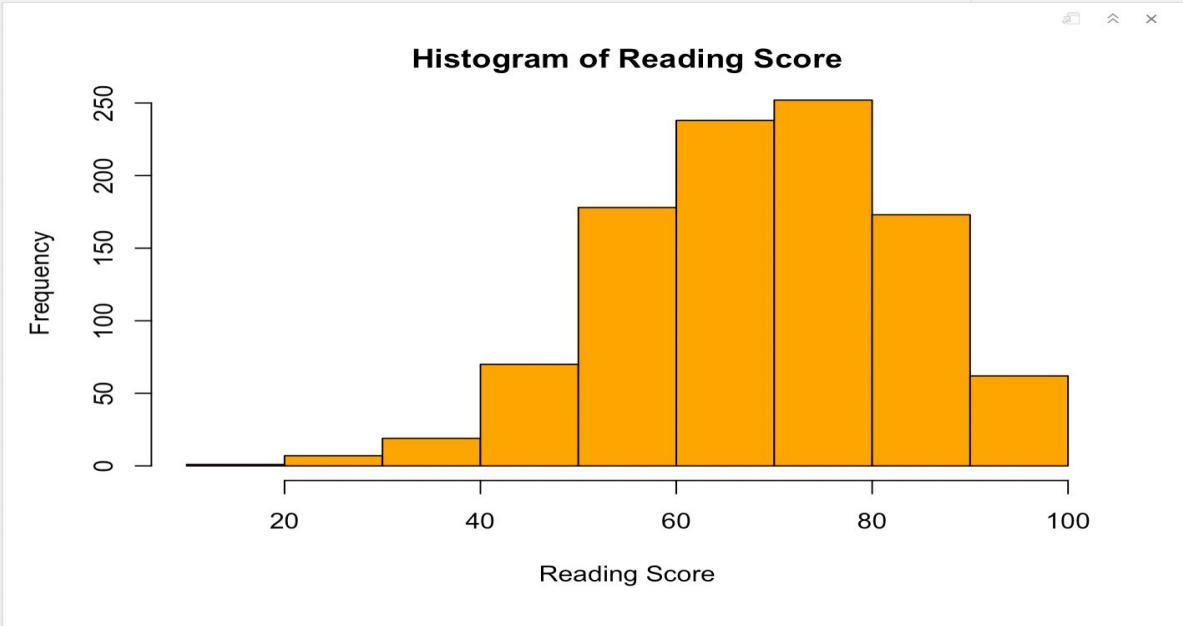
Summary for Boxplot visualizations:

1. Students who completed the prep class had better scores in all three tes
2. Female students have received better scores in reading and writing.
3. There is a presence of outliers in all three tests.

Data Exploration(EDA): Histogram

Summary for Histogram visualizations:

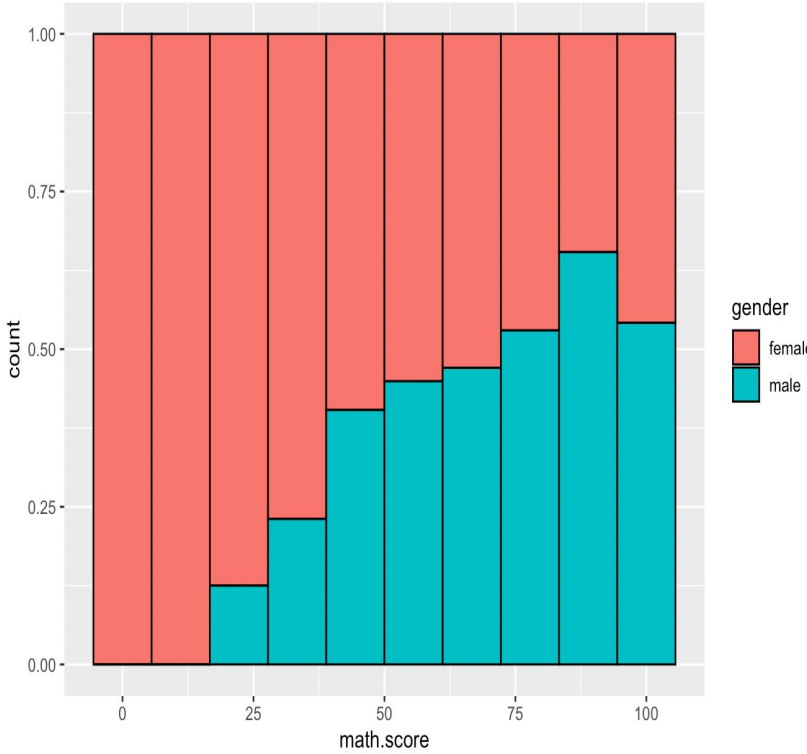
1. We have plot the histogram for reading score.
2. On X - axis we have Reading score and on Y - axis is the frequency.
3. Highest score for reading ranges from 70 to 80.



Data Exploration(EDA): Overlay Histogram

Summary for Histogram visualizations:

1. We have plot the overlay histogram for math score gender wise
2. Female students tend to have good maths scores compared to male students



HYPOTHESIS TESTING

QUESTION: How does a student's reading score affect their exam performance?

- Reading Score is one attribute in our data set which indicates of whether a Student performs better in exams.
- Histogram and Boxplot indicated that female student perform better compared to male students.
- This offers us a broad sense of how a reading component could be influenced.
- According to “Annual Review of School Performance”, reading is a major problem in majority of the schools.
- The shift towards increased video learning is main concern for many schools.
- Hypothesis testing can be conducted to find the average reading score for male and female students.

Hypothesis Testing (Continued...)

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.

- We have used Two Sample T test.
- Two Sample T-test is used when we have two samples, and we want to compare mean of one of the numerical variables from both of them. So in order to get two samples, we have split the whole data by gender.
- In doing Two-Sample T-test, usually we need to have two samples with equal variances.
- The variances look different from each other, but as we know variance is the square of standard deviation so when we look at the standard deviation (another measure of variability), they are actually quite similar.

Hypothesis Testing (Continued...)

- Two Sample T-test assumes that our population data is normally distributed. So, we need to check for normality, which we have done.
- Male and Female Students Reading Score Comparison is the next what we have observed.
- Female students tend to score higher in reading rather than male students do. But to check whether it has a significant difference, we performed the Two-Sample T-test.
- Two-sample T- test used gives statistical evidence to show that female students tend to score higher in reading rather than male students and as p-value is less than the usual threshold of 0.05, so we can reject the null hypothesis.
- Altogether, these test results indicate that there is a difference between the average reading score of male and female students.

Linear Regression

- Student's Exam performance is a statistical measure of the average score an individual student is expected to get.
- Student's exam performance depends on several factors such as course taken, parent's level of education, daily diets and mock exams.
- Statistical Analysis on factors influencing Student's Exam Performance can be done using linear regression
- We aim to explore the parameters affecting the students exam performance individually.
- The model can be used to offer direction for many school across globe, as it targets several quantifiable parameters.

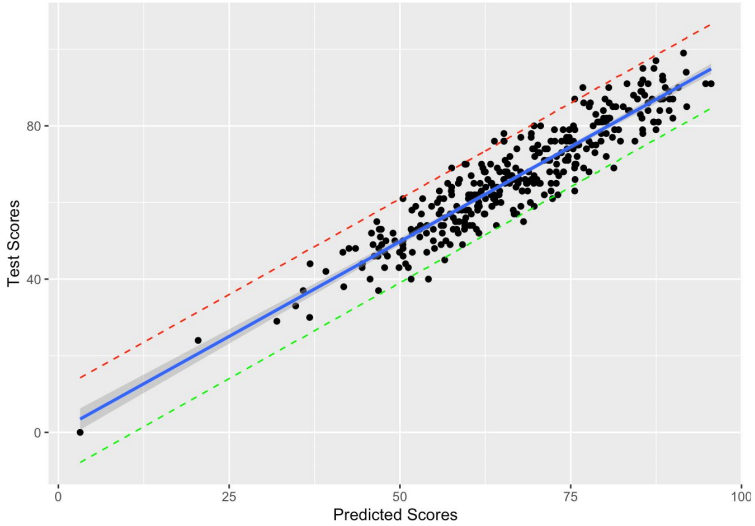
Linear Regression Summary

In this section we are going to build a linear regression model, predicting Math scores.

- Comparison between the data categories in features (Course/Parents Education/Lunch/Mock exams) was done to determine the most impactful category on Student exam performance.
- Subset Selection using Forward, Backward and Both-direction approach was done.
- Results show that, Course taken are the most impactful, based on criteria of highest adjusted R-square and lowest value of RSE.
- Here we have consider Math scores is dependent variable (Y), Writing.Score, Gender, Race, Lunch, Parent_Education, Test_Prep - independent variables (X)
- Data was split into training and testing (80:20)

Linear Regression Summary (Continued...)

1. Adjusted R Squared = 0.864 -> 86% of variation in Math scores can be explained by independent variables in our model.
2. R Squared value indicated that we created a good prediction model.
3. F-Statistics produces P-value of near to zero which indicates significance of our equation.
4. Visualization at 95% confidence and prediction intervals.



Any Questions?

THANK YOU