Education Analysis Final Project

Logan Preuss

South Dakota State University

Fall 2020

Abstract

This report is on an education data set. I will be using exploratory based analysis on the data set using the tool R programming. My goal is to find out what ethnicity group is best and why.

**Education Analysis Final Project**

**Background**

The data I am using comes from a Excel file from the website Kaggle. The data is 7 columns across and 1000 lines down with each line being data on a singular student. Each student has a gender, a parental level of education, an exam prep completion, a race group, a math score, a reading score, and a writing score. The location of this data is unknown. The age range for the data is also unknown but it is most likely high school level students.

# Objective

My main objective for this project is to find out what ethnicity has the best students meaning they have the overall best scores. I am going to do this by examining the math, reading, and writing scores of each ethnicity group. After I investigate the ethnicity groups I am going to try and figure out what factors make that ethnicity group better. I am also going to make a model based on the three scores.

## Data and Methods

The first thing I did was examine the data for missing values. After examining the data set, I found that there were no missing data points. I then used the R programming language to read in my data set into the program. Next, I created some new columns so I could assign each score a letter grade and to assign each student a GPA. I did this by using a function to assign a letter to a selected range of values and putting the result in a new column. I did this for all three of the scores. Now I have a letter grade for each of the scores for each student. Next, I started to work on the GPA. I first took each letter grade and assigned it a weight. Then I combined the three weights into one final GPA and added that value to its own column. Now each student has a letter grade and a GPA added to their data.

Now with my data cleaned and prepped I can start working on my question. First, I wanted to see if the scores were correlated or not. I generated a correlation plot using the tool R. I was able to see that all three of the scores were very strongly correlated. Reading and writing were more correlated than the math scores but still they were all very positive. This told me that if a student does well on one score, they most likely are going to do well on the others as well.

Then I moved on to my big question. What ethnicity group performed the best? I generated a bar chart using R. My x axis was the ethnicity groups, and my y axis was the total number of each letter grade for math scores. This is shown in figures at the bottom of this report. I was able to see that group E had the highest number of A’s in proportion to the other groups. I then calculated the mean scores for each group and group E was 6 points higher than any other group. I then created a box plot so I could see the data without the effect of having a different number of students in an ethnicity group. Group E was still better than the other groups as suspected. Next, I used the ANOVA test. This test uses statistics to see if the means of the groups are the same. I used R to run the ANOVA test. My null hypothesis was that the means would be the same. My ANOVA test told me to reject the null hypothesis meaning that statistically the means were not the same so I could conclude that the means were not the same and it was not a sampling error. Based on my findings I concluded that group E had the best mean for math scores.

Next, I took a look at reading scores. I generated a bar chart using R. My x axis was the ethnicity groups, and my y axis was the total number of each letter grade for reading scores. This is shown in figures at the bottom of this report. I was able to see that group E had the highest number of A’s in proportion to the other groups. I then calculated the mean scores for each group and group E was 3 points higher than any other group. I then created a box plot so I could see the data without the effect of having a different number of students in an ethnicity group. Group E was still better than the other groups as suspected. Next, I used the ANOVA test. This test uses statistics to see if the means of the groups are the same. I used R to run the ANOVA test. My null hypothesis was that the means would be the same. My ANOVA test told me to reject the null hypothesis meaning that statistically the means were not the same so I could conclude that the means were not the same and it was not a sampling error. Based on my findings I concluded that group E had the best mean for reading scores.

Next, I moved on to writing scores. I generated a bar chart using R. My x axis was the ethnicity groups, and my y axis was the total number of each letter grade for writing scores. This is shown in figures at the bottom of this report. I was able to see that group E had the highest number of A’s in proportion to the other groups. I then calculated the mean scores for each group and group E was 1 point higher than any other group. I then created a box plot so I could see the data without the effect of having a different number of students in an ethnicity group. Group E was still better than the other groups as suspected. Next, I used the ANOVA test. This test uses statistics to see if the means of the groups are the same. I used R to run the ANOVA test. My null hypothesis was that the means would be the same. My ANOVA test told me to reject the null hypothesis meaning that statistically the means were not the same so I could conclude that the means were not the same and it was not a sampling error. Based on my findings I concluded that group E had the best mean for writing scores but not by a very large margin.

Based on group E having the best mean scores in all three categories and using the ANOVA test I was able to conclude that they were the group with the best students. My next question was trying to figure out why they were the best group. I created another bar chart to see if group E had more students to complete the test preparation course. I had the ethnicity groups on the x axis and the test course data on my y axis. After examining the graph, I could see that all five of the ethnicity groups had the same proportion of completed test preparation courses. This told me that the test prep course did not affect group E more than the other groups.

My next question was to see if a certain group had more male or female students. I created another bar graph using R with the ethnicity groups on the x axis and the ration of male to female on the y axis. After examining the graph, I could see that each group had the same ratio of male to female. I was able to conclude that gender did not have an affect on why group E scored higher.

Next, I moved on to parent level of education. Like before, I generated a bar graph using R with ethnicity on the x axis and the ratio of parent level of education on the y axis. I could see from the graph that group E had the lowest ratio of high school level education and the highest level of college education compared to the other groups. The difference was big enough to have an impact on the scores. This told me that if the students’ parents were more educated that their children have a better chance to score better on their tests.

Based on the data I was given, parental level of education was the only difference in the ethnicity groups. Since group E had overall the highest scores and the only difference was that they had a higher parental level of education I can conclude that parental level of education was why group E scored higher than the other groups. There might be other factors on why they scored better but based on the data I was given that is my conclusion.

Next, I wanted to generate a linear model to predict a student’s GPA based on one of their test scores. I decided to use a linear model because my data was normal, and the test scores followed a linear pattern. I then used R to create a linear model for each of the scores, math, reading, and writing. All three of the models had very low p-values and high r squared values. Meaning that all of the models fit the data well. The model that edged the other two slightly was reading. That model preformed the best out of the three so I would choose that model to predict a student’s GPA based on their reading score.

### Discussion

I have a couple of concerns with my findings. Based on the dataset I was able to conclude that group E performed better because they had a higher parental education level. I know that in the real world this is not the only factor that drives a student’s performance. I would like to have a bigger data set to see if there is a different factor that affects a certain ethnicity group more than parental level of education. With the data I was given however, parental level of education was the only difference between the groups and that is why they scored better in this sample.

**Conclusion**

Overall, I was able to conclude the group E performed the best out of the five ethnicity groups. Based on the data I was given I was able to conclude that parental level of education is one of the main reasons why group E outperformed the other groups. If the opportunity presents itself, I could look into what level of education affects the scores the most. Overall group E was the best ethnicity group.

References

https://ggplot2.tidyverse.org/

https://rstudiopubsstatic.s3.amazonaws.com/462335\_ee8bc0d5d7cc4ceca401becb96a4553f.html

https://www.scribbr.com/statistics/linear-regression-in-r/

**Figures**

Chart, bubble chart

Description automatically generated

Chart, bar chart

Description automatically generated

## race.ethnicity

## group A group B group C group D group E

## 61.62921 63.45263 64.46395 67.36260 73.82143

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

## race.ethnicity

## group A group B group C group D group E

## 64.67416 67.35263 69.10345 70.03053 73.02857

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

## race.ethnicity

## group A group B group C group D group E

## 62.67416 65.60000 67.82759 70.14504 71.40714

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -3.1196746 0.0549680 -56.75 <2e-16 \*\*\*

## reading.score 0.0744942 0.0007776 95.80 <2e-16 \*\*\*