**GraceFoundations Inc.**

**Statistical Analysis of Educational Data**

**Karnik Aswani**

**Data Science Graduate Student**

**Table of Contents Page**

**Introduction……………………………………………………………..1**

**Summary………………………………………………………………...1**

**Analysis……………………………………………………………….....2-3**

**Conclusions……………………………………………………………...4**

**Recommendations……………………………………………………....4**

**Summary**

**The Garces Foundation is a nonprofit organization that specializes in providing the Philadelphia immigration community with food, healthcare, and education. This organization has been operating for 10 years and it has made significant impacts on individuals that need the resources mentioned in this statement. This report will discuss the analytics tools that were used to analyze educational data.**

**Introduction**

**In this report, I will be performing several analytical techniques to answer GarceFoundations research questions. The most important research question would be the performance of student test scores in the organization. The performance of test score is crucial because it will show if the students are improving their performance in their academic performance or not. The more specific questions will be answered as the analytical techniques are being performed including hypothesis testing.**

**Analysis**

**Exploratory Data Analysis**

**The purpose of exploratory data analysis is to find patterns and trends of the dataset. Before analyzing the dataset, we must drop the columns that are not related to our analysis. Next step, we use the summary function to gather the basic statistics of the dataset. Also, we need to use the groupby function to calculate the mean of each subject level with the subject test score to compare it to the individual test score.**

**Outliers are values that can affect our results which is why we need to remove the outliers by visualizing them and extract the outliers through several functions in the outlier’s library. Distribution of the dataset is important because it will determine which statistical tests are appropriate based on how the data is distributed. All the test scores are non-normally distributed which means the values are left-skewed.**

**Associational Statistical Analysis**

**The purpose of association statistical analysis is to find relationships in the dataset. The most important techniques to find relationships between two or more variables are linear regression and multiple regression. All the linear regressions of the test scores are positive relationships since all of them have positive slopes.**

**Statistical Tests**

**The purpose of statistical tests is to perform hypothesis testing including alternative and null hypothesis.**

**The first statistical test is the Grubbs test which is used to detect outliers in a dataset. All the p- values are less than 0.05 meaning we would reject the null hypothesis and conclude there are outliers in all the subject test scores.**

**The second statistical test is the Chi – Square test of independence which is used to determine the significance of two categorical variables. All the p – values are less than 0.05 meaning we would reject the null hypothesis and conclude there are strong correlations between two variables from all the tests.**

**The third statistical test is the Shapiro test which is used to determine if the distribution is normal or non-normal. All the p – values are less than 0.05 meaning we would reject the null hypothesis and conclude all the distributions are normal.**

**The fourth statistical test is the Fligner Killen test which is used to determine homogeneity of group variances based on ranks. All the p – values are less than 0.05 except the grammar and listening scores meaning we would reject the null hypothesis of two of the tests and conclude those population variances are unequal in those test scores.**

**The fifth statistical test is the ANOVA test which is used to compare variances of the means of different groups. All the p – values are less than 0.05 meaning we would reject the null hypothesis and conclude all the population means are unequal in all the test scores.**

**Conclusion**

**We can conclude that there is a strong correlation among the three subject levels and scores and positive relationships among the three test scores. Also, we can conclude that the population variances and the population means are unequal across the groups except the relationship between the grammar and listening score in the Fligner Killen test. This means there will be low errors in the data values to show the highest accuracy of the individual test scores.**

**Recommendations**

**There are not too many recommendations to improve the test scores except performing more test sessions with the students and taking the time to teach the students to learn the material more efficiently.**