**Research Question**

The research that is going to guide this analysis is if we can predict math scores based on a variety of different variables. This research question can be justified by if a school district needed to know the quickest and most likely way to increase math scores, they can do that with this analysis. The context in which the research question exists could be an education consulting firm was hired by this school or district to analyze some score data to see how they can best improve overall math scores. My hypothesis is: The variables gender, race/ethnicity, parental level of education, lunch, test rep, reading and writing scores can with statistical significance predict math scores.

**Data Collection**

The data collected is an Excel csv file containing information about high school students from the United States Any identifying factors have been removed for the analysis. The csv file has 1001 rows with 8 columns. The target variable is going to be the math scores having an integer datatype. The predictor variables are going to be gender (male/female), race/ethnicity (categorized as A-E), parental level of education (high school, some college, associates, bachelors, masters), lunch (standard/free and reduced), test prep (completed, none), reading score and writing score (numerical value from 1-99. The variables gender, race/ethnicity, parental level of education, and test prep are all categorical variables which will have string datatypes. The reading and writing scores are going to be integer datatypes.

One advantage of gathering my data this way is that I had a complete dataset with plenty of predictor variables for my target variables put together nicely for me into a csv file already. One disadvantage is that if I thought there were more factors or maybe factors that could predict math scores even better, but they weren’t collected as part of the data collection process for this set 1of data then I am out of luck. I could avoid that in the future by collecting my own data and specifying which variables I want to use instead of using a data set from Kaggle or any other open-source platform.

To gather the data, I used the Kaggle platform. Kaggle allows users to find public datasets and perform various analyses on them.

**Data Extraction and Preparation**

To extract my data, I used Kaggle.com. Once I had found a data set, I wanted to use I needed to download it to my computer first.

Graphical user interface, diagram

Description automatically generated

**After clicking download a bar pops up on the bottom of the screen with the file containing the csv file you just downloaded.**

Chart

Description automatically generated

**After downloading the csv, I needed to import it into jupyter notebooks, the environment in which the analysis is going to be performed. After opening jupyter notebook you need to click upload to bring the csv into the environment.**

A picture containing table

Description automatically generated

**Next you create a new notebook and load in your csv with the following code.** Graphical user interface, text, application, email

Description automatically generated

**To start cleaning the data I need to make sure we don’t have any missing values in the dataset.**

Graphical user interface, text, application

Description automatically generated

**As you can see from the image above the dataset has no missing values. Next, we want to look at the distribution of each of the variables individually.**

Chart, histogram

Description automatically generated

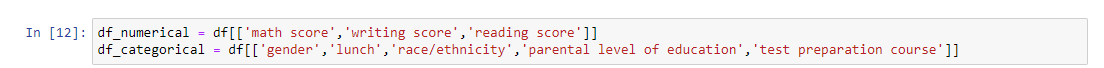
Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

**To get our data ready for analysis we need to scale the numerical variables and create dummies for the categorical variables. The first step in doing this is to split the dataset into two new ones, df\_numerical and df\_categorical.**



**The next few pictured cells will show the transformation the numerical data goes through when being scaled. We use StandardScaler to do this. When it comes to the categorical data, we are going to use Pandas’ get dummies to replace the words with a 0 or 1. The 1 meaning the column is true or 0 for false.**

Table

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

**Now that we have the data in the correct format for the analysis, we need to combine them back into a data frame to perform the analysis on.**

Graphical user interface, table

Description automatically generated

**As you can see, we have a new data frame with the correct column headings. We have a completed data frame that has the data in the correct formats labeled correctly. We are now ready for the analysis. Python is the language that I decided to use because it has an abundance of libraries that make running a multiple regression model easy. These libraries include pandas, seaborn, matplotlib.pyplot, sklearn.preprocessing, and statsmodels.api. One advantage is that I can easily check the perform of my analysis by printing the results summary. To create the model, run it, and then return the results is three lines of code in total. One disadvantage would be if the data is incomplete leading to a relationship or pattern observed that is incorrect. Incomplete data could be a data frame that has a lot of null or missing values, a data frame that has a lot of outliers or data that is in the incorrect formats for the analysis being performed.**

**Analysis**

  The data analysis technique used in this analysis is multiple regression. This type of model is good for when you have one variable you want to predict with multiple predictor variables. The evaluation metric I used for my model is the r-squared. The closer to 1 the r-squared is the better and more reliably the model can be used. When trying to find which predictor variables are statistically significant in predicting math scores, we will be looking at their p-value.

Graphical user interface, table

Description automatically generated

You can see the initial model has a R-squared value of 0.877 with a statistically significant variable being masters with a p value of 0.019. While these numbers look good it is always best practice to try and adjust the model to get the best results possible. To try and create the best model I then looked at the VIF or variance inflation factor of each of my predictor variables.

Graphical user interface, text

Description automatically generated with medium confidence

When looking at the VIF scores I decided to remove the variables Race C, writing score, reading score, and Race D. The results of the reduced model are shown below.

Graphical user interface, text, application

Description automatically generated

In the reduced model we have a R-squared value of 0.245 and based off p scores we have 0 statistically significant variables.

One advantage of running a multiple regression model of this particular set of data is being able to determine exactly the influence each predictor variable has on the target variable. Through this analysis technique I will be able to tell exactly which variables influence the target and which ones do not. One disadvantage of multiple regression is that is the data is not cleaned correctly and outliers are still present in the data at the time of the analysis they will weigh heavy into the analysis and give you a skewed analysis that could lead to improper business decisions .

**Data Summary and Implications**

E.  Summarize the implications of your data analysis by discussing the results of your data analysis in the context of the research question, including one limitation of your analysis. Within the context of your research question, recommend a course of action based on your results. Then propose **two** directions or approaches for future study of the data set.

After the analysis we can say with a solid degree of certainty that we could use the initial model above to help predict or show how we could influence math scores. The initial model had an r-squared of close to 0.9 and the reduced model had an r-squared of 0.245 which is a drastic difference. One limitation of my analysis could be my own personal biases on what I think could influence math scores or not influence math scores. One thing that could be done to further study this data set is by adding more variables like science scores or maybe even attendance. Another thing that could be done is run a side-by-side analysis of this data with data from other districts in the state or country to see how your school or district compares to others.

Online Sources:

*Machine learning - multiple regression*. Python Machine Learning Multiple Regression. (n.d.). Retrieved February 12, 2022, from https://www.w3schools.com/python/python\_ml\_multiple\_regression.asp

Katari, K. (2020, October 15). *Multiple linear regression model using Python: Machine Learning*. Medium. Retrieved February 12, 2022, from https://towardsdatascience.com/multiple-linear-regression-model-using-python-machine-learning-d00c78f1172a

Microsoft. (n.d.). Bing. Retrieved February 12, 2022, from https://www.bing.com/videos/search?q=multiple%2Bregression%2Bmodel%2Bpython&docid=608033216013542878&mid=CE5881A65195D0EBCDE8CE5881A65195D0EBCDE8&view=detail&FORM=VIRE