For this project I would like you to choose a problem which involves answering one or more questions, finding one or more interesting insights, and applying machine learning to a problem you are personally interested in. To do this, you will apply some of the techniques we have seen in class to one or more data sets. The problem you will work on is entirely of your own choosing. You will be responsible for finding an appropriate set of data, performing the analysis, and making a presentation of your findings.

For your proposal you should write a short intro/background to your problem and answer the following questions:

1) What are the key questions you hope to answer, what are the key insights you are trying to find, and how do you intend to apply machine learning?

2) What data set(s) will you use, and where can they be obtained?

3) What are your anticipated methods?

The purpose of this proposal is simply to ensure you choose an appropriate project. Your proposal should be about 1/2 a page. Please write in the Canvas text editor for this assignment - no document upload is necessary here.

In my FSEM, we studied how race continues to play a big role in present day America. One of the books claimed that this pattern could be best seen in education. I did a project which seemed to suggest that high schools in Prince William County with less minorities tended to have better test scores. My hypothesis was that race itself doesn’t have an impact, but minorities tend to be less economically well-off. With machine learning, I want to increase the research to include all public schools in Virginia. The key question I hope to find is whether the percentage of minorities has any correlation with the percentage of disadvantaged students. If it does, is the percentage of minorities is a good predictor of the passing test rate? The key insights I am trying to find if race still affects the education system through financial well-being. I intend to use data that includes: the number of students in each school broken down by race, the number of disadvantaged students in each school, and the passing test rates of each school. All of these data sets can be found on the Virginia Department of Education Website (<http://doe.virginia.gov/statistics_reports/index.shtml)>.

I intend to use the 2015 data as my training data and run my model on the 2016 data. Since the raw data sets gives me the number of students, I will need to change this into the percentage of students. I will probably need to manipulate the panda dataframe to do this. I think I will use basic regression because I am only comparing the relationship between two variables at a time. To determine if there is a correlation, I will look at the Mean Absolute Error and R2.