Homework 3 Part 1  
William Chirciu, Amy Edwards, Julie Simms, Breanna Atwood, Jeffrey Moy

a) Record minutes of the discussion

Tuesday, 10/16/18, begin at 7:30 PM

-- discussed the data and what all the variables mean

-- answered these questions in problem 1

-- discussed some future analysis ideas

-- each chose a method of analysis for the project

Ended 8:50 pm

b) Choose a group liaison who will be responsible for interaction with me on group project issues

Amy Edwards / William Chirciu

c) Perform an initial analysis on your data and summarize the data for its content. What are the variables involved and what do they mean? Don’t just list the data dictionary from the download site. I want you to think and analyze these variables for how they might interact with others.

Answer the following questions:

• Is there an obvious parameter of interest, or are there several

-- There are several parameters that are of interest. We discussed being able to predict the economic need index or school income based on test scores or school demographics. We also talked about academic progress (in terms of Math and ELA scores) over grades 3 - 8.

• What interesting metric variables does the data have

-- Many; for example: average MATH/ELA proficiency scores, school income, economic need index, and the number of students across multiple ethnicities in grades 3-8 from each school who have scored >4 on MATH/ELA proficiency tests. We also have percent data tied to a school quality rating system involving things such as Rigorous Instructive Rating, Supportive Environment Rating, Collaborative Teachers Rating,etc. Each school has its own percentage target and an associated rating (Not Meeting Target, Approaching Target, Meeting Target, Exceeding Target). Finally, there is percent data relating to the ethnic representation within each school (%black, %white, %hispanic, etc.)

• What interesting categorical variables does the data have

-- There is the Community school variable : ‘yes’ if it is a community school and ‘no’ if it isn’t.

-- There is District/Zip Codes: we can see how location impacts economic need and test scores

-- Grades low/high give us the highest and lowest grade level within each school

• Are some of the variables ordinal (ordered but not metric like we discussed in class?)

-- -- We have a lot of ordinal variables relating to the quality of the school: rigorous instruction rating, trust rating, supportive environment rating, strong family community ties rating, and student achievement rating. These are all on a scale of {Not Meeting Target, Approaching Target, Meeting Target, Exceeding Target}.

• What missing values are there and are there any patterns that you might exploit for filling them in?

-- There are some missing values, in particular for school income. In this case, we can average the school income across zip codes and use that as the missing value. There are some missing values in economic need and we can exploit the same pattern. For categorical missing values (Like trust rating), we would probably have to deduce them by finding the most common rating for the zip code. In few instances, we will have to remove the data row completely.

• What variables look like they might interact?

-- There are a few variables like Trust % + Trust Rating, where the rating and the percent for certain parameters are provided, that will correlate. As well as “Percent Black” and “Percent Hispanic” with “Percent Black and Hispanic”. Economic need + School Income will interact because greater economic need typically arises in lower income communities. Student attendance is likely to be lower in schools that have higher percent of chronically absent. Average ELA (or Math) Proficiency may relate to Rigorous Instruction Percent - how well the curriculum and instruction engage students and align with the common core.

• For the metric variables, what units are used and are different variables measured in related but different units? (e.g. cm and mm or feet and miles)?

-- The ethnic makeup of the schools, attendance rate, and school ratings are all in percentages

-- School income is in dollars

-- Economic need index, number of students who scored > 4, and the average test scores are all unitless.

• For some of the categorical variables, what dependencies are there between any metric variables (parameters of interest?). Explore with boxplots, or if you need to investigate categorical/categorical interactions, look at correspondence analysis.

-- All of the rating parameters have a corresponding percent parameter and these may be dependent on each other.

• What obvious directions for investigation present themselves?

-- So far it is obvious that we want to explore the relationship school income and economic need index have with the other variables. We want to investigate how school quality ratings impact test scores. It would also be a good idea to see how the ethnic makeup of the school correlates with test scores and economic need.

d) Each team-member should choose some initial individual direction to accomplish for the individual part (problem 2). These do not have to be incredibly deep at this point, but should show some progress towards determining how variables relate, what kinds of latent variables might be present or interaction between metric and non-metric variables.

Amy - CA

Will - CCA

Breanna - Model Building & Multiple Regression

Jeffrey - PCA

Julie - CFA

(see problem 2 for individual responses)