Final Project Notes

# Final Project, Project Plan

1. Acquire data
   1. Download school proficiency data as .csv from [www.mischooldata.org](http://www.mischooldata.org)
   2. Download entity demographic data as .csv from [www.mischooldata.org](http://www.mischooldata.org)
   3. Download entity data (including both state and federal district codes) from [www.mi.gov/eem](http://www.mi.gov/eem)
   4. Query API data from [www.census.gov](http://www.census.gov)
   5. Download 2016 poverty threshold data from <https://www.census.gov/library/publications/2017/demo/p60-259.html>
2. Load the data into R
3. Explore the data
4. Tidy (clean and transform) the data
5. Join the data
   1. Either in R or export to MySQL

# Issues with Data Acquisition

1. The proficiency data publically published is partitioned by assessment, school, grade, and subject. The combination of this detailed of partitioning combined with suppression rules necessary for student privacy, result in meaningful loss of data for smaller schools.
   1. Alternative, request from MDE a dataset partitioned only by school and subject (remove the assessment and grade partitions). This will result in less data loss.
2. The data pull from Michigan’s Educational Entity Master ([www.mi.gov/eem](http://www.mi.gov/eem)) only has current (December 2018) school to district relationship mappings. Schools can and do move districts from year to year (both operationally and geographically). Thus, the school to district mappings need to be pulled from a snapshot of the year matching the proficiency and census data being used which is 2016.
   1. This is particularly true for Detroit schools as the Detroit was reconstituted and given a new district code, thus more than 90 schools in Detroit alone will have a different district codes reported currently as compared to a snapshot of who was their district in 2016.
   2. Alternative, request from MDE school-to-district mappings pulled from the 2016 end-of-year snapshot of the Educational Entity Master (EEM).
3. The original return of 2015-16 Entity Data file (containing snapshot of Michigan Educational Entity) was .csv. Unfortunately it had a field [ActualGradeListSearchable] which contained commas “,” within that field. Thus, when read into R using sep = “,” the data became corrupted.
   1. To correct this, changed the request to an output type of text tab delimited and loaded into R with sep = ‘\t’ (tab). This resolved the issue
4. The 2015-16 Entity Data file had a field [BuildingNCESId] which is a numeric field with 12 digits. When loading this into R using read.csv the field was unintentionally converted to scientific notation which lost needed values.
   1. After some reading through documentation and discussion boards it became clear that the state option(scipen = 999) needed to be run before the read.csv() command. This would disable scientific notation.
5. 2016 US poverty thresholds are only given in 97-2003 excel (.xls) format. The r packages to load this type of file into R require a local copy of the file (they will now load a file referenced by URL).
   1. Quickest work around was to download the .xls file, open it, and save as a .csv file, then post that .csv file to the project GitHub repository. This introduced an manual element to the project but was by far the quickest and least complicated solution.

# Data cleansing issues

1. The peculiarities of converting factors to numeric caused a sizable portion to understand
   1. After some struggle, found that factors must first be converted to character than string
   2. Attempted to initially do this in later steps, to avoid NAs during forced coercion. However, this was harder to follow and was not resulting in the desired outcome. Thus, conversion was done earlier.

# Training on how to use the Census APIs

1. Go to [www.census.gov](http://www.census.gov)
2. Hover over “Explore Data” in horizontal menu along the top
3. Click on “Census Academy”
4. Click “Webinars” in the vertical menu on the left
5. Click on “Recorded Webinars” in the vertical menu on the right
6. Scroll down to and click on “American Community Survey”
7. Scroll down to and click on “Using the Census API with the American Community Survey Webinar” (June 2018)
8. Review the recording, the slides, and transcript

Get census data

1. Go to [www.census.gov](http://www.census.gov)
2. Hover over “Explore Data” in horizontal menu along the top
3. Click on “Developers”
4. Click on “Available APIs” in the vertical menu along the left
5. Click on “Discovery Tool page” in the page introductory paragraphs
6. Click on “api.census.gov/data.html”
7. Search for “2016”
   1. control F, enter “2016” then click next until the desired record is found
   2. Continue until finding the ‘ACS 5-Year Data Profile’
   3. Using profile rather than subject when possible as profile is a subset of variables and thus much quicker to search
   4. For this project, profile has the desired variables
8. Use the “examples” link to help build the API query for the desired location
   1. For this project Michigan Geographic School Districts
9. Use the “variables” link to identify the coding for the desired variables
   1. DP05\_0028E, which gives the estimated total population across all races
   2. DP03\_0062E, which gives the median household income (2016 inflation adjusted)
10. Build the full api query
    1. Base site: <https://api.census.gov/data/2016/acs/acs5/profile>
    2. Full query [https://api.census.gov/data/2016/acs/acs5/profile?get=DP05\_0028E,DP03\_0062E,NAME&for=school%20district%20(unified):\*&in=state:26](https://api.census.gov/data/2016/acs/acs5/profile?get=DP05_0028E,DP03_0062E,NAME&for=school%20district%20(unified):*&in=state:26)

API query to get [#total population] and [median household income] by MI school district

1. Map charters to geo district
2. Map median income to geo districts
3. Calculate ratio of median income to poverty line (24,250)
4. Map pct prof (by subject and across subjects) for each school
5. Map pct ED for each school
6. Perform regression on
   1. pct prof and pc ted
   2. Pct prof and raito of med\_income:poverty\_line
7. Calculate a regression residual for each model
8. Standardize the regression residuals
9. Look for cases where the regression residuals tell conflicting stories
10. Pull lat/long for each school and map