1. **Variables:**

Below are some variables and their sources we considered from the beginning, not all of them are kept in the final model.

1. High-quality ECE centers. Source: Table B14003, S1401 from census.gov
2. ECE enrollment. Source: Table B14003, S1401 from census.gov
3. Third grade reading proficiency. Source: <https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html>
4. Third grade math proficiency. Source: <https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html>
5. School poverty. Source: <http://data.diversitydatakids.org/dataset/17001_2_p-poverty-rate--children-aged-0-17--percent--by-race-ethnicity/resource/ce623331-3821-47db-ae56-4909b6d1a333?filters=year%3A2008-2012>
6. Adult educational attainment. Source: Table B15002 from census.gov. <https://data.census.gov/cedsci/table?q=education&g=0400000US48.970000&hidePreview=true&tid=ACSST1Y2018.S1501&t=Education>
7. Pupil/Teacher Ratio. Source: NCES self-generating table tool.
8. Free Lunch Ratio. Source: <https://cometmail-my.sharepoint.com/:x:/r/personal/hxz172830_utdallas_edu/_layouts/15/Doc.aspx?sourcedoc=%7BF0A50605-85DC-4D82-96DD-6BC355C73815%7D&file=StudentTeacherRatio.csv&action=default&mobileredirect=true>
9. In married-couple families. Source: B09005 from https://censusreporter.org/
10. Reduced Lunch Ratio. Source: http://nces.ed.gov/ccd
11. Title I School. Source: http://nces.ed.gov/ccd
12. Graduate % for High school, some college, Associate degree, Bachelor’s degree, and Graduate degree or higher, further combined into 2 groups: Bachelor’s degree or higher and High school. Source: <https://data.census.gov>, Table B15001
13. **Python API:**

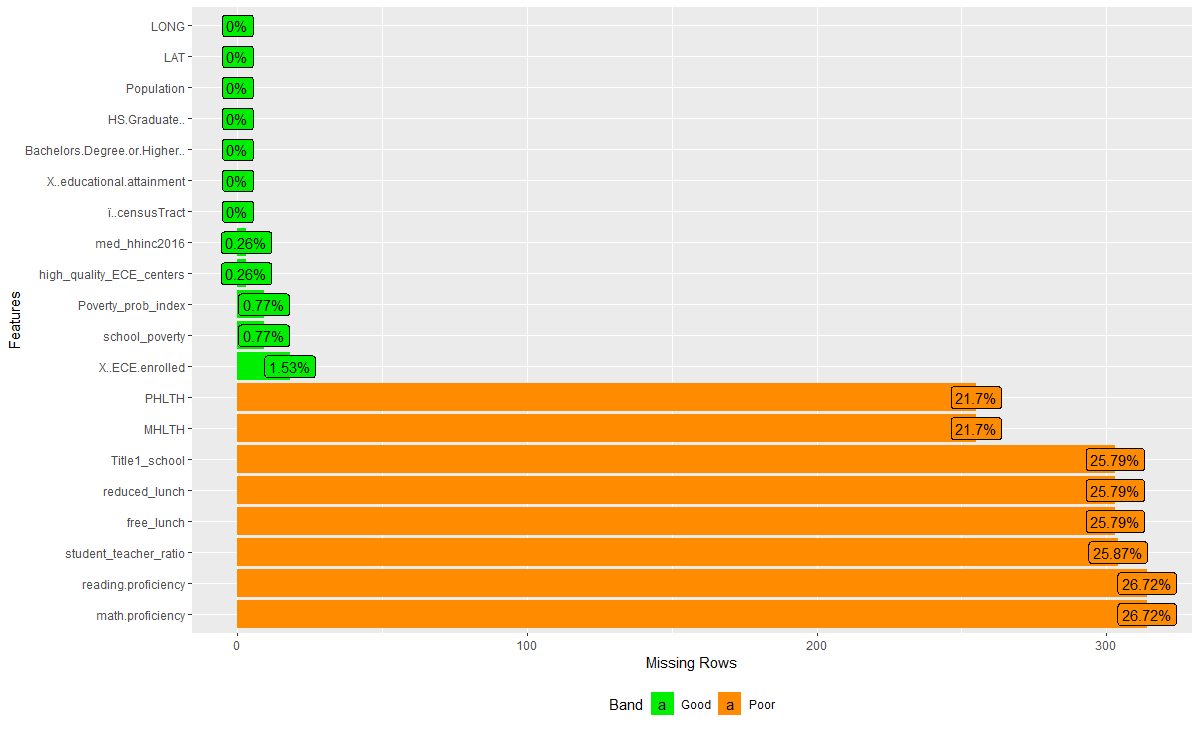
One major issue when gathering data for education is that most data is collected by individual state agencies and are at either the individual school or district level. This created an issue when trying to map this data to individual census tracts. To get around this the census website gives you the ability to look up latitudes and longitudes and get back what census tract that location falls into. Normally this is done one by one, which was not feasible with the number of schools we had in the data. A python script was created to call the website, input the latitude/longitude, and record the census tract associated with the school. This allowed us to create an index containing school identification numbers and census tracts.

The other issue that arose through mapping school level data to census tracts is that not every census tract will have a school in it. Due to this we ended up with about 300 out of 1175 census tracts missing data. In comparing with Owen’s education index, he did not encounter this situation because he used the schools themselves as the point of interest rather than census tracts.

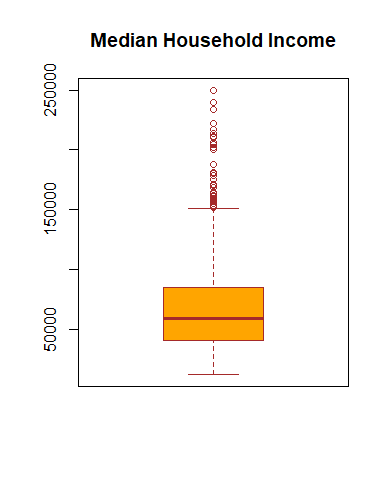
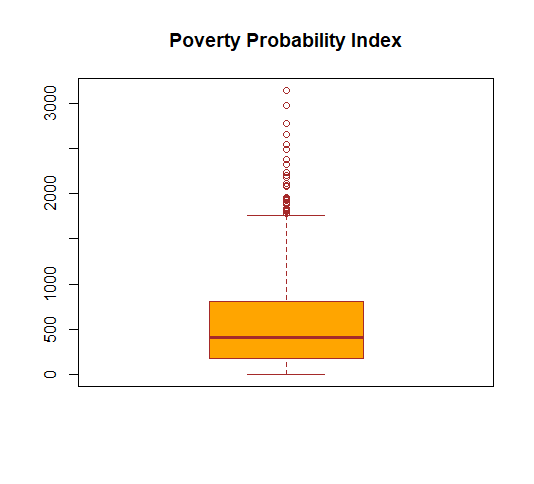
*\*Python code for this attached in the package.*

1. **Handling missing values – KNN imputation:**

* Our dataset has a total of 11 features and 4 target variables. All 4 target variables and 9 features had missing data



* We filtered out tracts that had zero population
* We analyzed the data distribution of all the missing features and target variables using a box plot. We observed there were outliers.



* We ran the KNN algorithm to get the nearest neighbors based on the latitude and longitude of the tracts.
* We found 5 nearest neighbors for our features and 4 nearest neighbors for our target variables.
* Based on the boxplot distribution we imputed the median value of the neighbors where the variables had outliers while the others were imputed with the mean of the nearest neighbors

*\*R code for this attached in the package.*

1. **Calculation for weights:**

**Method 1: Simple linear regression**

* Handling outlier by Z-score, set the threshold with 5 to identify outlier.

Figured out there are 12 outliners in the dataset, however, this is only the sub-index, if we drop these census tract, there will be missing value when combing final dataset, so we decided to keep these outliners.

* Scale data by calculating Z-Score for all variable to ensure the comparable variables

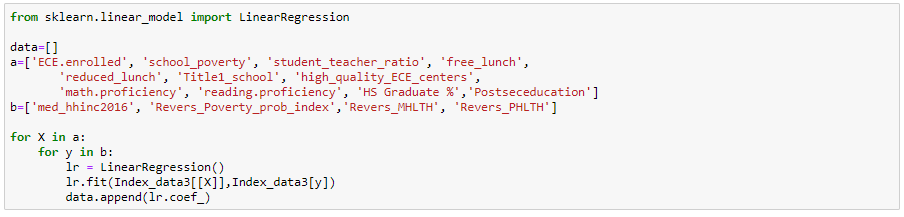
Z-score = [Xi-mean(Xi)] / std(Xi)

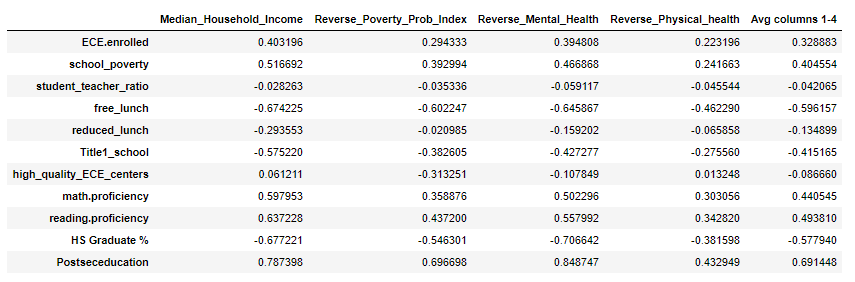
* Calculating the weights: Calculating weights using simple linear regression.

The weight of each sub-index is calculated based on the correlation coefficient of sub-index with the four target variables: Median Household Income, Reverse Poverty Probability Index, Reverser Mental Health, Reverse Physical Health.

The correlation coefficients are obtained from the simple linear regression models, we can avoid the multi-correlation between independent variables, because each time there is only one X variable regressed against one Y variable.

Yi= a + rj\*Xi + ei



1. In order to facilitate this process, we built a LOOP function in Python in which we regress each target variable (4) with independent variables (11) and output the results are below (Refer the file Practicum Python Code): 
2. And the following steps show how we calculate the weights for each sub-indicators:

* Calculate the average of correlation coefficients (rj)
* Rescale average correlation coefficients so that they sum up to the number of indicators:

Rj= rj\*D/S

Where:

D is the number of sub-indicators (11)

S is the sum of the average correlation coefficients (rj)

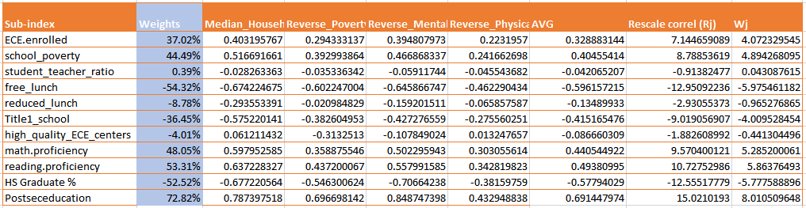
* Standardize Rescaled average correlation coefficients (Rj): We averaged correlation coefficients with a constant unity-weights and divide the resulting sum by 2. The resulting weights sum up to the number of sub-indicators.

Wj= (Rj +1)/2

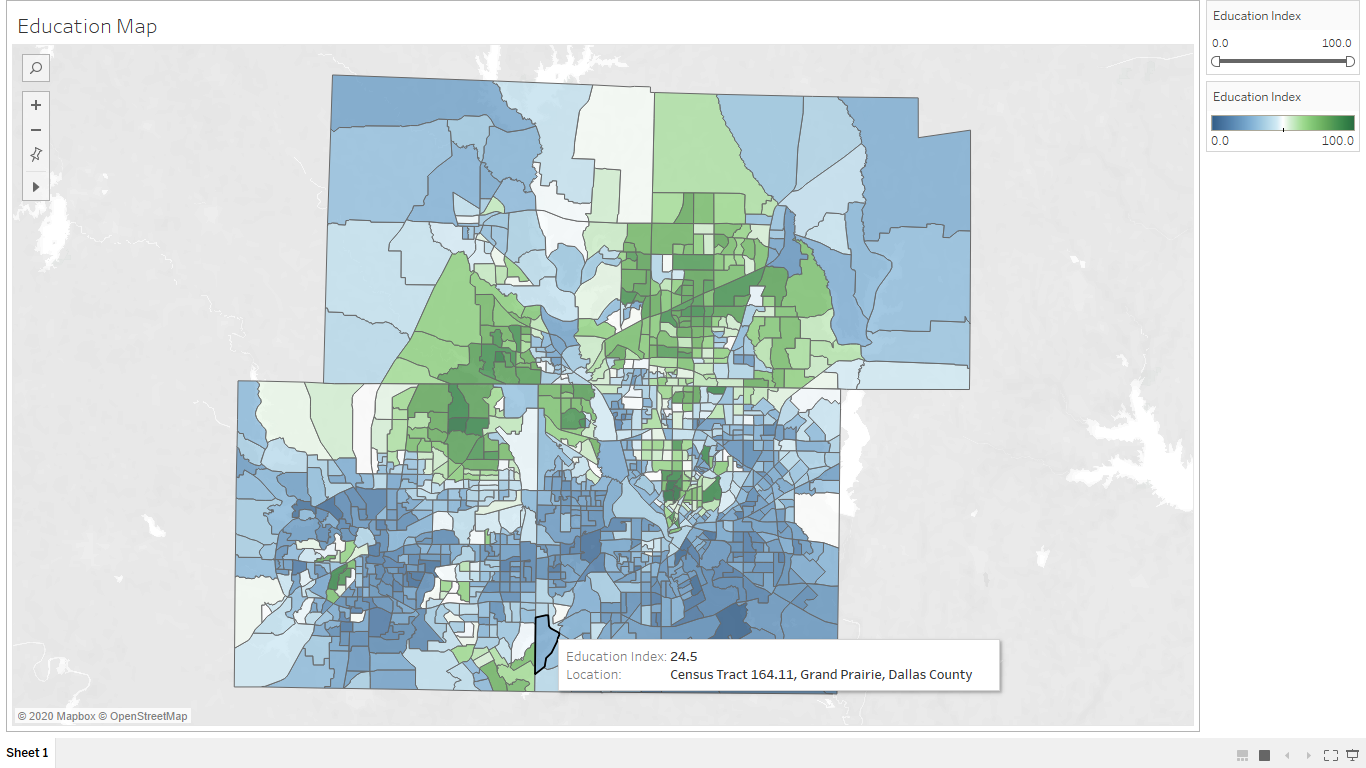
* Final weights: we rescaled the weights so that they sum up to one

Wj/sum(Wj)

Below is the final weight (Refer file: Weights calculation 2)



Refer file: Education Index M1.twbx



**Method 2: Multiple linear regression**

1. Used the data with no missing values post imputing the missing values with KNN.
2. Aligning Target variables.

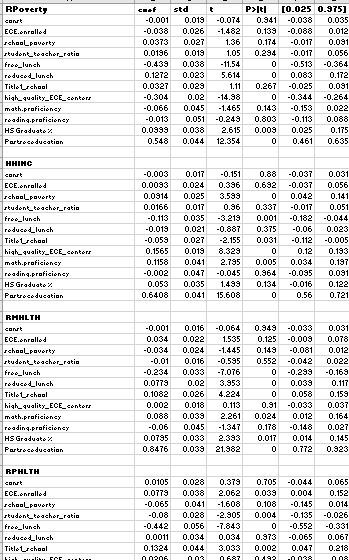
HHINC higher value means better score for the tract while higher value Poverty\_prob\_index ,Mental health and Physical health means a bad score for the tract. Multiplied below target variables by -1 to ensure all scores are aligned.

* Poverty\_prob\_index
* Mental health
* Physical health

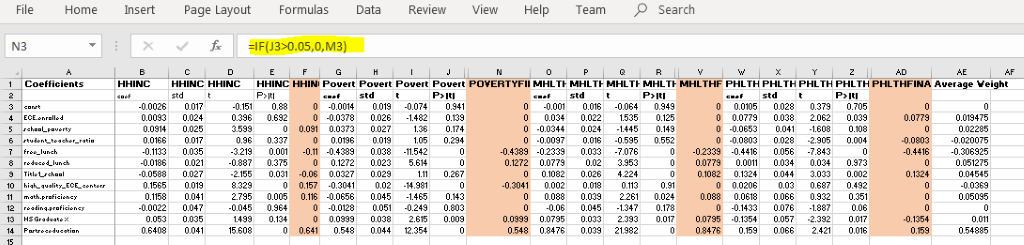
1. Scale data by calculating Z-Score for all variable to ensure the comparable variables

Z-score = [Xi-mean(Xi)] / std(Xi)

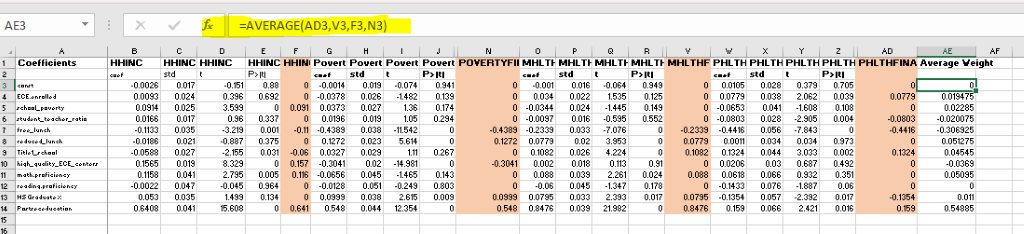
1. Handling outlier by Z-score, set the threshold with 5 to identify outlier. Figured out there are 12 outliners in the dataset, however, this is only the sub-index, if we drop these census tracts, there will be missing value when combing final dataset, so we decided to keep these outliners.
2. Checking collinearity and removing Education attainment Education attainment is 99% collinear with Post-secondary education so removing the education attainment before performing multivariate regression.
3. Run multivariate regression with 4 target variables one by one in Python. Below are the results



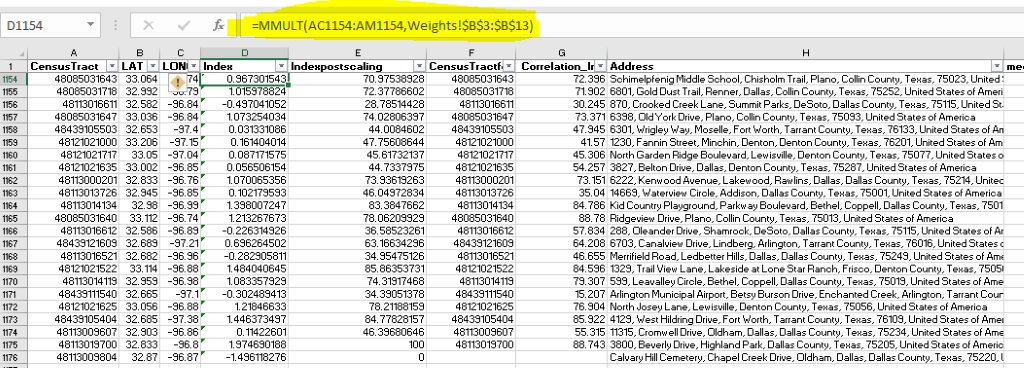
1. Remove the non-significant coefficients having P value >0.05



1. Average the coefficients to calculate the final weights

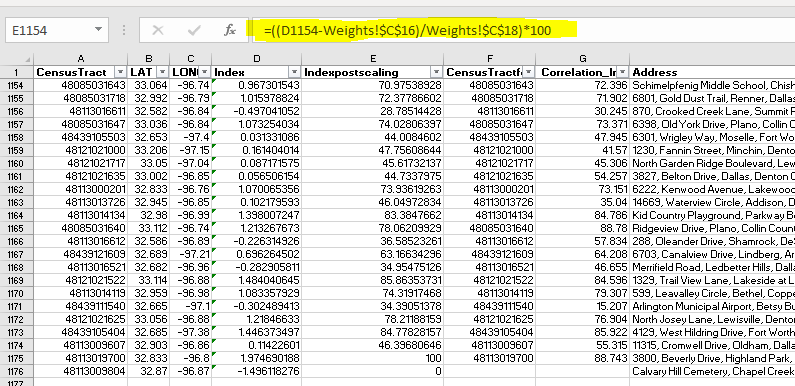


1. Multiplied Independent variables with weights to calculate index

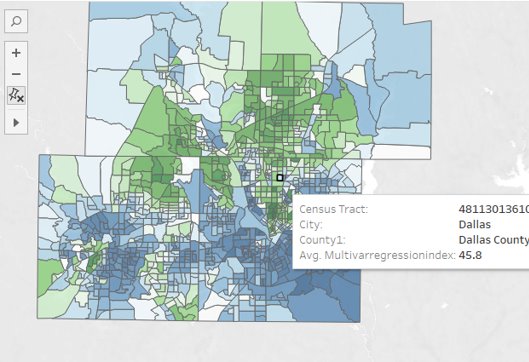


1. Scaled Index from 1 to 100

Scaled value =(X-Min/Range)\*100



1. Mapped the scaled index on the Tableau dashboard



1. Index comparison

Comparing the index obtained using method 1 and 2 to analyze the difference and fine tune the modelling.

* SEI=Single Var regression education index
* MEI= Multi Var regression education index
* Step 1 : F test between SEI and MEI to check the variance.
* Result: Variance was not statistically different
* Step 2: 2 sample t test with equal variance between SEI and MEI
* Result: SEI and MEI are statistically different
* Step 3: Compared MEI with other indexes published by other groups Sub-indices ECO INDEX, COMM INDEX, FAMILY INDEX from other group for total absolute average diff as by multiple other research have found that that education is closely correlated with Economic, Community and Family Index.
* **Result: MEI abs avg difference with other comparative indicators is less compared to SEI and thus MEI was chosen over SEI.**