**Multivariate Regression and Index comparison.**

This document provides details into Multivariate regression for calculating education index and comparative analysis between Single Linear regression and Multivariate regression.

**METHOD 2 (Multivariate Regression)**

**Refer below jupyter notebook file for steps 1 to 4**

[**https://cometmail-my.sharepoint.com/:u:/r/personal/hxz172830\_utdallas\_edu/Documents/BUAN6390/Final%20Presentation%20%26%20Package/Multivarreg%26Compartiveanalysis/Educationindataprep\_v1\_0424.ipynb?csf=1&web=1&e=aGl9lz**](https://cometmail-my.sharepoint.com/:u:/r/personal/hxz172830_utdallas_edu/Documents/BUAN6390/Final%20Presentation%20%26%20Package/Multivarreg%26Compartiveanalysis/Educationindataprep_v1_0424.ipynb?csf=1&web=1&e=aGl9lz)

**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.

1. Poverty\_prob\_index
2. Mental health
3. Physical health

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

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

**4.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 censustract, there will be missing values when combing final dataset, so we decided to keep these outliners.

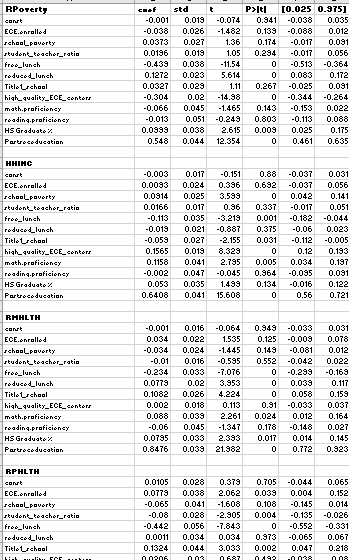
**Refer below jupyter notebook file for steps 5 and 6**

<https://cometmail-my.sharepoint.com/:u:/r/personal/hxz172830_utdallas_edu/Documents/BUAN6390/Final%20Presentation%20%26%20Package/Multivarreg%26Compartiveanalysis/Regression%20and%20correlation%20model%20for%20education%20index-zscore-0425.ipynb?csf=1&web=1&e=9JNqht>

**5.Checking collinearity and removing Education attainment**

Education attainment is 99% collinear with Post secondary education so removing the education attainment before performing multivariate regression.

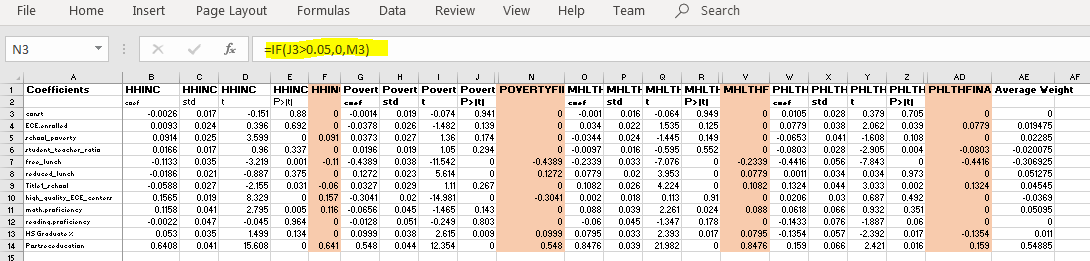
**5.Run multivariate regression with 4 target variables one by one in Python. Below are the results**



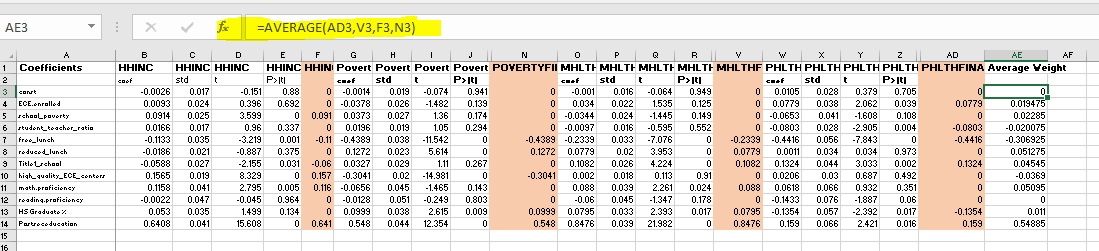
Refer below Excel file for steps 6 to 11.

<https://cometmail-my.sharepoint.com/:u:/r/personal/hxz172830_utdallas_edu/Documents/BUAN6390/Final%20Presentation%20%26%20Package/Multivarreg%26Compartiveanalysis/Regression%20and%20correlation%20model%20for%20education%20index-zscore-0425.ipynb?csf=1&web=1&e=9JNqht>

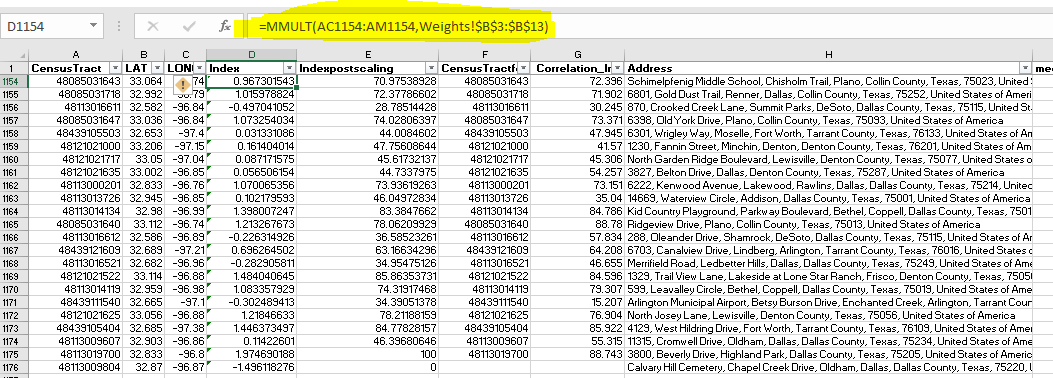
**6.Remove the non significant coefficients having P value >0.05**



**7. Average the coefficients to calculate the final weights**

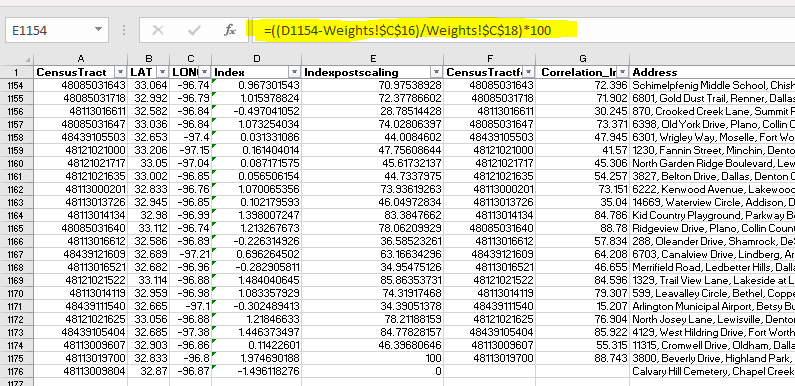


**8. Multiplied Independent variables with weights to calculate index**

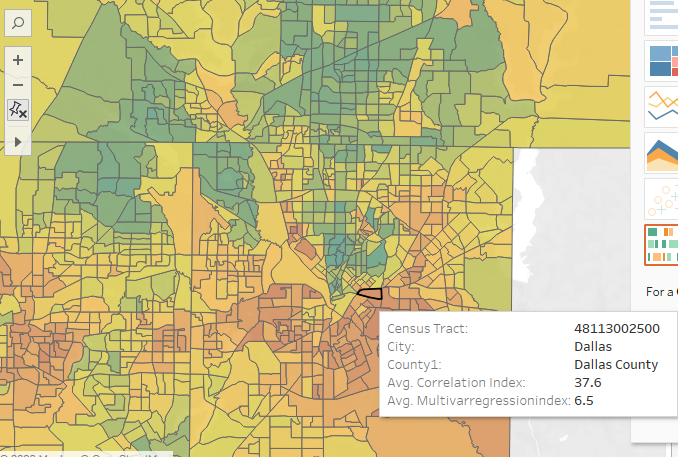


**9.Scaled Index from 1 to 100**

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



**10. Mapped the scaled index on the Tableau dashboard**



**11.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.**