Reflections on the project

We have encountered many challenges when analysing the county level Obesity data obtained from data.gov. We did come across a fair amount of challenges in our objective to analyse Obesity data obtained from data.gov. A structured approach to problem solving has allowed us to overcome the challenges in an efficient way. The collaborative effort of the team helped us in obtaining the final result. Below are our reflections on the project:

- In the process of finalizing on a good dataset for our project, we had to go through a fair number of raw csv and json files available through data.gov. Most of them did not have good data dictionaries, or were missing key fields required for our analysis. Even though it was challenging to research for the best dataset available, we did a good job in narrowing down to the best dataset which will help us answer our questions.

- The data was spread across various sheets with considerable number of unrelated metrics. It would require reasonable amount of manual effort to collate the required metrics in a single table. In order to solve this problem, we created a python code which takes the column names required and creates a master dataset. The creation of a master dataset simplified the analysis process.

- The data had few metrics obtained through a census which was carried out every two years or so. Our team had to interpolate some of metrics to get all the required data for year of 2010, the time period of analysis for this project.

- We had missing data for obesity percentage, the metric we are trying to analyse. Since it is inappropriate to replace these missing values with any statistical measures, we removed them from our analysis. Using the regression model built on the remaining rows, we were able to predict the obesity values for these missing rows.

- We had various libraries for regression and clustering and were not sure about which library to use. We tried various techniques and then finalized on the best codes for our analysis which are also easy to understand and maintain.

- One of the major challenges was to plot a heat map at county level for different clusters. There are very few modules which have county level specifications and most of them are very complicated. We found Vincent library which is supported by IPython but there was not enough guidance on the web about how to customize the maps. We ran multiple iterations in IPython to understand the behaviour of the library.

- Finally, looking back, we felt that though we did a good job of identifying the most significant factors from those which were available to us, we could have brought in external datasets to better explain the variations in obesity rate for the counties which could have gone on to make our model even more robust.