Please give me a full (at least one full page) specification of your proposed project.  This is not a software engineering course, so there is no defined format in which it needs to be written, but you may use one (UML, Z notation, etc) if you wish.  The document should be specific enough that I can understand completely the work you will be doing.  You should discuss the specifics you will use to create the program or programs.  Talk about how the interface will work, how the visualization will be incorporated, and what you will do with the data.  You will be, for the most part, bound to this description; this will be the program you write.  It will be used as a component to grade your final submission (how well you achieved what you propose).

Introduction:

Because personally I am interested in the way to reduce the prevalence of chronic diseases in America, I always want to study what are the factors that can impact patients physical and mental wellness. For my final project, I just want to demonstrate very simple linear relationship among those factors and come up with a conclusion that might be generalized to entire patient population.

Data Collection:

The data I am going to use come from data.gov. First of all, it is a website that is run by the US government. Therefore, most of those datasets are accurate and have high quality. In the meantime, there are around 200k datasets to choose. I can always find what I want from there. The three datasets I found are all pretty organized. It saves me time to tidy and transform the data. I want to combine the factors of interests from these three datasets, and create a master datasets that contain both outcome variables and multiple independent variables. I am going to use a lot of skills on handling python data frame to achieve that.

Data Visualization:

I am going to create a lot of scatter plots to investigate the association between explanatory and response variables. If scatter plots indicate there exist linear relationship, I will go ahead and build linear relationship model and draw linear regression line. I am also interested in the comparison among different states in America to see if there is any difference in term of diet, exercise and lifestyle et cetera. I am going to show the result in bar plot. All the visualization will be done using matplotlib package. And the graphs will be created in line with the actual code and write up.

Data Analysis:

Since most of the data I am going to use are numerical data, the major statistical method I will be using will be multivariable linear regression. It is the model that is probably most commonly used in real life and it is a very powerful model as well. With this model, I am able to include a lot of variables into my model. In addition, it is also very useful to make prediction using the model.

Goals:

My project will have all the analysis in a clear and organized way on one ipython file. The purpose of that being done is show differences of people's diet, exercise, and lifestyle have significant changes on their physical and mental wellness. If there are changes, with 1 unit increase or decrease of those explanatory variables, how much influences it will have on the response variable. Of course, I will also practice my python coding skills by using multiple python modules to build my report.