Problem:
Many food factors affect a person's food choices and diet quality. Factors such as grocery store proximity, restaurant proximity, food prices, nutrition assistance programs, and community characteristics are not clearly understood in how they interact with each other and their influence on the general population. Thus, the purpose of this project is to better understand the complexity of these influences and identify any causal relationships in order to create more effective policy interventions.

Client:
The client is the Department of Agriculture. They are interested in understanding the relationships between food availability and choices and a population's diet quality. In doing so, the DoA provided a comprehensive data set that examines factors such as access and proximity to grocery stores, store availability, restaurant availability and expenditures, food assistance, food insecurity, food prices and taxes, local foods, health and physical activity, and socioeconomic characteristics across all races, ages, genders, and income levels. By understanding the relationships between the factors listed above, the results will be used to create better policies regarding improving the US population's health and attitude towards food and food choices.

Data:
The data is available through the Department of Agriculture, through the Food Environment Atlas. The single excel file contains a variable list, supplemental data, and individual tabs for each variable. Any changes made to the data is accounted for and written as notes. The variable list includes metadata. As noted on the excel file, any missing data is denoted with a blank cell. Data variables include data taken from 2009-2016, depending on the variable.

Methodology:

The project follows the OSEMN framework. We start by acquiring the data from the Food Environment Atlas as an excel file and process it using Python packages by reading the data directly into the program. Next, we scrub and clean the data (wrangling) to extract any missing or irrelevant values. This process will constitute the bulk of this project since improperly cleaned data can affect the results. To get started, we need to consolidate the data (excel) and standardize it so that it is consistent throughout. Some columns can be merged or split during this time. Data wrangling is done using Python.

Once the data is clean and ready to use, we use exploratory data analysis to take a closer look at the data. Numerical and categorical data are handled separately. During this process, our goal is to understand as much as possible about the data and gather supporting evidence. We will create data visualizations (bar graphs, line charts, box and whisker plots, etc.), test our hypothesis, draw inferences based on what we see, account for any pattern abnormalities and skewness, and apply normalization methods (data transformation) to skewed data. This phase relies on good use of statistical knowledge and finding statistical significance from the data.

Using our best judgement, we can infer if the variables are correlated or if one implies the causation of the other. From there, we can start making a plan to model our data.

After examining the data, we will model the data. Since this is a regression problem, this project requires a machine learning model and multiple linear regression in order to understand the relationship between multiple variables. Because the variables are plentiful, only relevant ones will be selected in order to reduce dimensionality. We will measure goodness-of-fit and use error scores (MAE, RMSE) to measure accuracy between predicted and actual data points. If there is any discrepancy between values, then we will use feature engineering and a random forest model to alleviate the problem.

Finally, after modeling the data, we can interpret the data and provide appropriate conclusions for the clients. This part requires visual and written presentations that the client can easily understand and create actions for.

Deliverables:
The deliverables in this project include a Jupyter notebook, PPT presentation, full written report, and Tableau data story.