**Capstone Project 1 - Data Wrangling**

**Importing the dataset to Jupyter notebook:**

* The raw data is on an Excel file with multiple sheets, each sheet contains variables from different domain, related to food and environment. For example: Stores data, Accessibility to food stores data, health data, etc.
* The data points are all the different counties in the USA. Each county is paired with a unique numerical identifier, called FIPS.
* The first step of importing the data was to merge all the different Excel sheets to one Pandas DataFrame. That was done by using the FIPS column, together with the State name and County name, as the indexes upon which the data was merged on.
* After merging all the Excel sheets to one DataFrame, I received a table with 3150 rows (*Counties*) and 219 columns of variables.

**Dealing with missing values:**

* The raw data contained missing values. I decided to filter out these missing values, because filling them via some interpolation process did not seem like the best approach in the context of these specific data samples (*Counties*).
* Filtering out rows with any missing values led to filtering out more than 96% of the data points, so a different approach was needed.
* I decided to first filter out columns with multiple missing values, followed by filtering out rows with any remaining missing values.
* After some tweaking with the number of missing values in a column, for that column to be filtered out, I decided on a threshold of 100, meaning that for a column to be filtered out, it should have at least 100 Null values.
* After filtering out columns with 100 or more missing values, I filtered out rows with any remaining missing values.
* The above procedure left me with a DataFrame of 2935 rows and 167 columns.

**Dealing with outliers:**

* I decided to leave outliers, if there are any, as they were. I did not finalize what will be the predictive objective of the project, so I decided not to handle with outliers at the moment.

**Further data wrangling steps:**

* My future analysis will require only numeric data, but the DataFrame contained several non-numeric data columns. I tried to convert these columns to numeric columns, and the columns that could not be converted to numeric columns, I filtered out. That left me with a DataFrame of 2935 rows, by 165 columns.
* I converted the values in the DataFrame to z-scores (standardized each column), for future machine learning algorithms that will run on the data.