**Small pivot update 4/21/22:**

I have successfully used the Places API to gather all my food related data. However, when cleaning the ‘groceries\_df’ by hand, I have realized that there are far too many ambiguous results that were captured. I can clean the entire dataframe by hand, but I am afraid that differentiating one market from another market is too subjective, and I want to try to keep as much subjectivity out of the data collection as possible.

So my pivot is this: my overall hypothesis and general big picture idea of my MVP is the exact same. However, for the main meat and potatoes, I am going to primarily explore the relationship between disease incidence and the availability of larger, big name grocery stores such as Kroger, Publix, Walmart, and so on, or the lack of said grocery stores and the potential food desert that remains. This way I can be sure I am only looking at big grocers that contain the typical variety of healthy produce. I will still be able to use my other data – the presence of all the local corner store market stores adds another interesting layer for food availability and health impact.

**Executive Summary**

For my Capstone Project, I will be exploring publicly available health data alongside manually sourced geolocation data obtained via Google API queries. I am curious about how the availability of various food sources has an impact on local health. This curiosity is present in my own mind as a registered nurse and my aspirations to explore or somehow work with public health as a registered nurse/data analyst in the near future. The scope of my project is going to be limited to the greater Nashville area. If I can fine tune my project well enough, I should be able to adapt my queries to encompass a larger area if desired. The primary method of presenting my data will be through a visual presentation, comprising of both a powerpoint and an interactive dashboard.

**Motivation**

I have chosen this project based upon my background as a registered nurse. Public health and epidemiology have always been interests of mine, and my desire to continue working in the healthcare field as a hybrid registered nurse/data analyst will be demonstrated by a project highlighting the exploration of pertinent data. Hypothetical ways of addressing and improving public health will be synthesized from the data involved in the project.

**Data Question**

I propose that there is a positive correlation between the incidence of adult onset DM2 (diabetes mellitus type 2) and the prevalence of fast food restaurants in a geographic area. I want to expand on this topic by also exploring relationships between the incidence of DM2, and the number of grocery stores available in a geographic area. Lastly, I want to explore the relationship between average household income in a geographic area and the incidence of DM2.

Specific – I propose that there is a positive relationship between the number of fast food restaurants, prevalence of food deserts and DM2 incidence.

Measurable – Measuring DM2 incidence and its relationships with other objective data sets (restaurant, grocery availability, household income)

Attainable – The CDC and USDA both have publicly accessible datasets that provide raw data for health and food subjects matters.

Relevance – The incidence of DM2 and potential food poverty are an impactful subject matter with direct consequences within public health.

Time – Datasets from both the CDC and USDA both provide objective data that is tracked over periods of time. Analyzing the data over time will help illustrate how the public health impact changes in relation to our variables.

An additional question I want to explore revolves around the concept of food deserts. This was brought about by the recent news of the West End Piggly Wiggly closing down. The closure of this grocery store will eliminate a source of fresh food to a large number of local residents who live in lower income housing. The elimination of a healthier food source will surely have health impacts on their lives. I want to see if I can determine the prevalence of food deserts in the Nashville area, and the possible ramifications it has on the local population.

**Minimum Viable Product (MVP)**

My MVP will consist of a Jupyter Notebook/Python code that is then synthesized into an interactive dashboard. A PowerPoint will be created to introduce the audience to the topic, providing a minimum of understanding required to fully interpret the dashboard. The dashboard will be provided in presentation format to an audience who will be provided the opportunity to ask questions at the end.

**Schedule (through 4/29/2022)**

1. Get the Data (4/4/2022)
2. Clean & Explore the Data (4/8/2022)
3. Create Presentation of your Analysis (4/15/2022)

* Should be a presentation, but could include a Jupyter Notebook or dashboard in Excel, Tableau, or PowerBI

1. Internal demos (4/22/2022)
2. Demo Day!! (4/29/2022)

**Data Sources**

I will be utilizing the Google Maps/Places API to query Google in order to obtain information on local grocery stores, restaurants, and other food establishments. The data from Google will be cleaned and organized in a geospatial environment. Utilizing a geospatial environment is important as I will be attempting to map various localities in the Nashville area with the various food resource establishments overlayed on the map. I want to somehow use geolocation geometry to perform measurements to see if I can determine the existence and prevalence of food deserts.

For my health data, the CDC offers a plethora of publicly available data. A large number of datasets are easily available via <https://healthdata.gov>. I already have a .csv that contains epidemiological data.

Data Sources:

* Google Places API
* CDC PLACES collaboration, 500 Cities datasets: https://www.cdc.gov/places/
  + 2016 release:
    - https://chronicdata.cdc.gov/500-Cities-Places/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/5mtz-k78d
  + 2019 release:
    - https://chronicdata.cdc.gov/500-Cities-Places/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/k86t-wghb

**Known Issues and Challenges**

Known Issues and Challenges:

1. I need to debug my Google API query code so that 100% of it works as intended. I am currently able to form a useable dataframe, but some of the columns are not populating as intended.

2. There are a few different definitions for a food desert. I need to determine which definition is most relevant to my needs, and see how it can be adapted to geospatial Python code. I don’t know the feasibility of this idea, but I am thinking that I can somehow use the geolocation data of grocery stores/restaurants as a ‘centroid’ of sorts and find the minimum and maximum radii cutoff for walking or driving to the food source, as it applies to the food desert definition. I think incorporating this idea into useable code is going to be one of my biggest challenges.

3. Any health data I obtain will be in the form of a .csv. All of these will be imported into Python, where I will do all of the cleaning in pandas.

4. One of my other bigger challenges is going to be turning all of the data I obtain into a presentable format. I plan on using Tableau for creating some sort of a dashboard. For as powerful as Tableau is, I find that with all of its power comes a long list of peculiarities. In regards to displaying geodata on a map, I am wondering how I will import my geospatial data that I explore in Python into a useable map format in Tableau.