SafeGraph Shop - Data Request/Download

- Download dataset by location and date range
- Once approved and downloaded, manually extract *core-poi-patterns.csv* file into one separate folder for querying later

Python Script - Data Cleaning

The purpose of this python script is to collect all data into one place and wrangle data into desired output.

- Volume import datasets into one pandas dataframe
- Drop rows with NAs and columns of less value to optimize memory performance
- In order to expand visit count column from list to rows:
    - Divide data up to 28, 29, 30, 31-days months
    - For each subset of data, explode data by visits_by_day column and attach/loop through day (28, 29, 30, 31) information
    - Using substring from the time column to reconstruct exact date for each row
- Stack up all data subsets back into one dataframe
- Export data into csv

PostgreSQL Database (DBeaver) - Data Import

- Run SQL script with column names and types; import data into empty dataframe (2GB, ~3 hours runtime)

Python Script - Data Visualization

This script was created to run data visualization for demonstration purposes.

- Use sqlalchemy to connect jupyter notebook with sql database for querying data
- Data were dissected by different categories and aggregated using mean and median, seaborn package used for line charts

PostgreSQL Database - Spatial Analysis

This script was created to perform spatial analysis on dataset, as well as assigning health units (census block group, in this case) to data. Census shapefile was downloaded from the United States Census Bureau.

- Import census shapefile to PostgreSQL database using QGIS DBManager API

- In SQL, a matching table was created joining census block group information to all unique locations that were present within the timeframe for the given location; this operation was conducted using PostGIS ST_Contains function
- The matching table was then joined back with the original visit information table, creating the ultimate base table for upcoming model creations and corresponding analysis. This data can be aggregated on store, date, census block group (or health unit) levels depending on model needs