Final Project Report

Simran Shah shah.sim@northeastern.edu
Preston Reep reep.p@northeastern.edu
Nico Gaspar gaspar.n@northeastern.edu

DS 2001
Professor Strange

Problem Statement and Background

In certain geographic areas the lack of proximity to affordable and healthy food providers is restricted and sometimes nonexistent due to the absence of grocery stores that are within convenient traveling distance - these are known as food deserts. Thus, the foods that are mostly available in these areas are processed foods sold by fast food chains and small convenience stores, and therefore the easiest and sometimes the only option available to many families and individuals. According to the United States Department of Agriculture, in 2010 2.2 percent of households in the United States are over a mile away from grocery stores and do not have access to a vehicle. The consequences of long-term diet that is fueled by processed foods in any setting include higher rates of obesity, cardiovascular disease and type 2 diabetes, however are statistically higher in food deserts as nutrient-rich foods are not easily found - one study in Chicago neighborhoods found that the death rate from diabetes in food deserts was almost twice as much areas offering access to grocery stores.

In order to accurately observe where food deserts are located in Boston, we decided to use a csv file that contains information on food retailers and their respective locations throughout Boston, along with two datasets that contain the 5 year census information and median household income by census that was a csv file and shapefile respectively.

With this project we set out to find out the food accessibility of the various census tracts in Boston, as well as seeing whether or not median household income has any influence on access to grocery stores versus and access to corner stores within particular neighborhoods.

Introduction and Description of the Data

The topic of Food Deserts have repeatedly been brought up in class discussions both as an issue facing those in poverty and of how they can be quantified. We have seen diagrams

showing which areas of various cities are considered food deserts, but we wanted to look at the data on a more intimate level. To be specific, focusing on the connections of access to healthy food to the financial conditions of the nearby areas. And being that we attend University in the city of Boston, we were curious to see how this issue applies to the nearby communities.

We decided to incorporate two datasets into our project in order to look at this connection. The *Food Retailers(2017)* dataset provided a list of every store which sold food in Massachusetts and the M*edian Household Income by Tenure (Census Tracts)* dataset provided the median household income for every census tract in the state of Massachusetts. With both of these datasets, we will use median household income to define financial status, and with all the food retailer locations we can see where stores are, the distance between stores, and the density of stores within a certain region. We found both datasets on datacommon.mapc.org, project by the Metropolitan Area Planning Council (MAPC), that provides open source data in regards to geographic data. MAPC uses geographic data analysis and visualization to get a better understanding of the state of Massachusetts, and makes this information available to the public.

Although it was not a dataset in it of itself, we also incorporated a shapefile of the census tracts in Massachusetts. Later on in this report, it can be seen how this file can be interpreted as a dataframe and interact with the other two datasets.

This problem is important because food insecurity is a widespread problem in the United States that has only been amplified by the COVID-19 pandemic. In 2019, 8% of people in the State of Massachusetts reported having poor access to food and in the midst of the pandemic, this number has risen to 16% in 2020. It is also important to acknowledge that this issue does not affect all neighborhoods equally. Having a better understanding of whom might be affected by poor food access, or to see if there is a correlation between the distance from grocery stores and

property value, may give a better insight into what measures should be taken to work on this issue, and where they should be applied.

Methods

Now that we have our three data sources, we can now get into implementing them in our code and seek to address the problem at hand.

To begin, we have to work with our Food Retailer's Dataset. We first need to import pandas and write a function that reads the csv file as a pandas dataframe. This function also serves to assign each column its respective title, and uses reverse indexing to get rid of the first row, which is the row used the labels in the first place. For the sake of making our data look more presentable, we incorporate a function that takes the pandas dataframe and filters out any unwanted columns, both taking in parameters for what columns want to be kept and storing the unwanted columns into a list to all be dropped together.

The data in the filtered set of columns was then narrowed to include only data of food retailers that were both in the Boston area (municipalities in Suffolk county) and either supermarkets or convenience stores. To do this, two dataframes were created (one for supermarkets and one for convenience stores) using a function taking in a list of municipalities and a prim type. We ran this function on the filtered data twice with different prim types entered to create the different data frames. Both times the list of Suffolk county municipalities was used.

Now that we have the appropriate columns and rows which we want to look at, we import the necessary plugins: matplotlib, geopandas, and Point, Poligon from shapely.geometry. With the the parameters in line, we wrote a function which takes the pandas dataframe and converts it into a pandas dataframe, in the process applying a crs value to scale the data to the real world and

taking the latitude and longitude columns and combining them together in a geometry column, which we can now implement throughout the rest of the project.

Now that we have all of our desired coordinates together, we move on to interpreting our Census data. We reimplement the function we used to read the Food Retailers to read the Median Household income by Tenure csv file. In order to filter this census data to only cover Boston, we wrote a function which takes a specified range of census Geoid values and filters out the rest.

We then have to interpret the shapefile containing all of the Census tracts in Boston. We did this by writing a function that reads the shapefile as a geopandas dataframe. Now that we have that dataframe, we can apply the same function we used to filter the census tracts to get the desired polygons for the city of Boston. With all of these operations completed, we are now left with a geopandas dataframe which contains all of points of grocery stores in Boston, a pandas dataframe showing the median household income in each census tract, and a geopandas dataframe that shows the polygons of each census tract in Boston.

In their application the problem in hand, we begin by writing a function which takes a pandas dataframe and a pandas dataframe, using it to merge the median household income pandas dataset with the shapefile geopandas dataframe. It converts the geopandas dataframe into a pandas dataframe, maintaining the geometry column, and combining them on the GEOID index. Once we have these merged datasets, we then use a function to convert it back into a geopandas dataframe so we can get access to the geometry column.

With the merged data frame to our disposal, we can then move on to finding out it's the density of a given store type within each census tract. To do this, we create a function which finds out which census tract each grocery store is found within. We did this by creating a function which takes a dataframe, out merged dataframe, and a geopandas dataframe of

coordinates and compares them. It initially creates a column and fills each role with an empty list. In a for loop, it then runs through each coordinate and sees if it is or is not within the countries of each shape. If it is found to be within a shape, then it is appended to its respite intersection list.

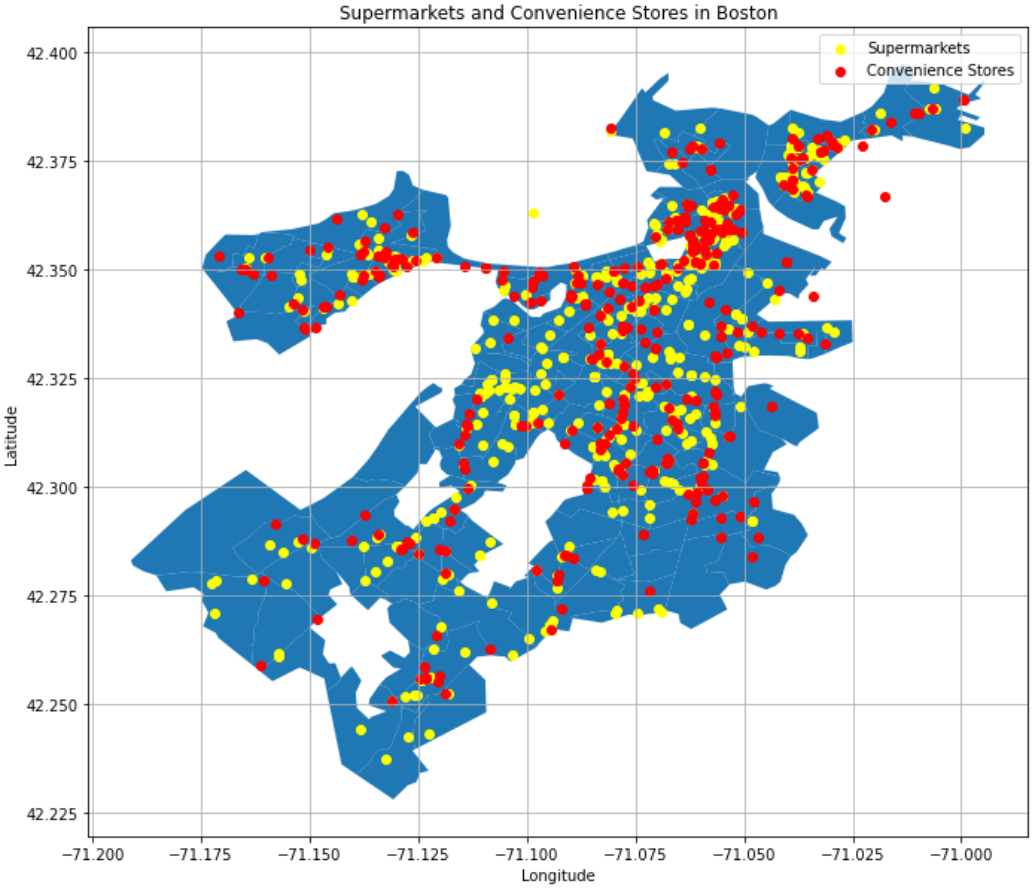
We then calculate the area. Given the polygon shape in the pandas dataframe and a prescribed CRS value, the same one we used earlier in the project, we can use .area to find the square area within the boundary in square meters.

A new column was created in both the supermarkets and convenience store dataframes for density of the types of stores in the tracts. This column was created by first determining the number of stores in each tract by finding the length of the list of points in the Intersections column. Then, the number was divided by the land area of the tract.

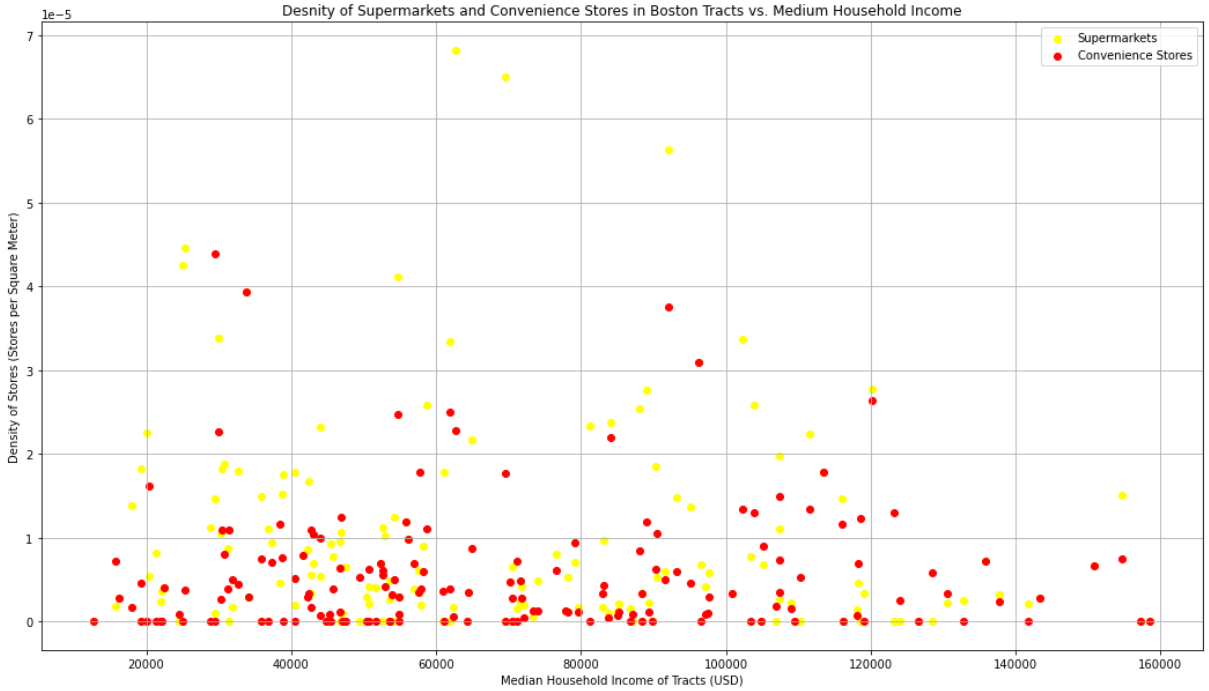
The shape file was then plotted and the points from the store datasets were plotted on top of it. The density of both types of stores was then plotted against median household income for each tract.

Results, Conclusions and Future Work

In the results of our project, we were able to generate a map which charts all of the grocery stores and all of the convenience stores in Boston.



We were then able to compare this data to the median household income in each tract, forming scatterplots where income is the independent variable and median household income is the dependent variable.



What we noticed from our observations, there tended to be a larger concentration of both grocery stores and convenience stores in the north, centring around the Back Bay, Beacon Hill area. As you start to go away from this area, the stores begin to grow more spread out especially as you go more south in Boston. It can also be seen that some areas such as Dorchester are heavily dominated by convenience stores.

The strengths of our code is that due to it consisting of many functions with many possible variables, it can work for various ranges of census geocodes. Moreover, the number of functions acts as a way of cleaning the data without necessarily having a “clean data” function as it reduces the problems that would come changing names and creating multiple data frames through hardcode. Although we do have many functions, a shortcoming is that many of these functions are built around a specific column that we have included in our geodataframes, *Geoid.* This does not make the code as efficient as possible, as we are not taking a column in as an argument, so these functions can only be used for our specific data frames that we have created

from reading the datasets we used. Apart from taking columns in as an argument in our functions if we were given more time, we would have found a way to measure and plot food desserts based on their distance from existing points.

Works Cited

Report:

https://www.bostonindicators.org/reports/report-website-pages/covid_indicators-x2/2020/october/food-insecurity
https://www.mapc.org/

Coding Resources:
https://towardsdatascience.com/geopandas-101-plot-any-data-with-a-latitude-and-longitude-on-a-map-98e01944b972
https://www.earthdatascience.org/workshops/gis-open-source-python/intro-vector-data-python/
https://stackoverflow.com/questions/48097742/geopandas-point-in-polygon#48105955
https://gis.stackexchange.com/questions/218450/getting-polygon-areas-using-geopandas
https://stackoverflow.com/questions/26645515/pandas-join-issue-columns-overlap-but-no-suffix-specified
https://geopandas.org/gallery/create_geopandas_from_pandas.html
https://gis.stackexchange.com/questions/336437/colorizing-polygons-based-on-color-values-in-dataframe-column#336455
https://www.twilio.com/blog/2017/08/geospatial-analysis-python-geojson-geopandas.html
https://transition.fcc.gov/form477/Geo/more_about_census_tracts.pdf
https://stackoverflow.com/questions/42739327/iloc-giving-indexerror-single-positional-indexer-is-out-of-bounds
https://stackoverflow.com/questions/62649240/python-geopandas-module-object-not-callable
https://geopandas.org/data_structures.html

Data sources:
https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html
https://www.census.gov/library/reference/code-lists/ansi.html#par_statelist
Median Household Income in Massachusetts by Census Tracts
Food Retailers in Massachusetts