Guide to BPDA Main Streets

Table of Contents

Data

Data Sources for Analysis

Google

Largest source for data outside InfoUSA and the Main Street Directors' list. Contains duplicate and bad entries. Relatively expensive data compared to other online sources. Does not return all businesses when requesting data for a given area.

Bing

New source of data. Similar scope to Google and completely free.

Yelp

Data source used mainly for restaurants. Does not return all businesses when requesting data for a given area. Limited amount of free requests per month.

InfoUSA

Expensive database that we are trying to replace. Data can be unreliable for a given area or for closed businesses.

Main Street Directors' List

Data collected by surveys sent out by hand form the City of Boston. Also a source we are trying to replace as it is time consuming. Data has different formats and should be looked at for undesired results.

SafeGraph

New source of data (for data validation only). Contains the data regarding the business performance based on the foot-traffic pattern data.

OpenStreetMaps

Discarded since it returns limited amount of data.

YellowPages

Discarded since it returns limited amount of data. Data can be unreliable for a given area or for closed businesses.

Data Sources to Assist Analysis

SAM Addresses

This data source is a comprehensive data set with all addresses inside of Boston. Used to run address-based requests to various web APIs.

Main Streets Shapefile

Main shapefile used for plotting and filtering for merges. This shapefile contains all Main Street regions.

City of Boston Shapefile

Boundary of the City of Boston. Used for better plotting to visualize where data is.

Boston Street Segments Shapefile

Contains all streets for the city of boston. When combined with the City of Boston Shapefile this creates a detailed map for seeing data points on a map.

Libraries

Pandas

Main library for data manipulation. The pandas dataframe is central to almost every function and demand of the project.

GeoPandas

Library that extends pandas dataframes to be able to plot and manipulate geospatial data. Very helpful plotting is built in. The function 'create_gpd' is used to directly create a geopandas dataframe from a pandas dataframe.

Requests

Simple web request library. Central to data gathering from the API's in the next section.

APIs

Google

https://developers.google.com/places/web-service/intro

Google Places API that is our main source of data at the moment. Does not return all data inside a given area, so overlap is almost necessary for complete data.

Yelp

https://www.yelp.com/developers

Can get given places for a radius and by name. Limited free requests per month.

Geonames

http://www.geonames.org/export/web-services.html

Used to convert addresses to latitude and longitude and vice versa. Free but has timeouts.

LocationIQ

https://locationiq.com

Used to replace Geonames due to speed and accuracy. Does similar geocoding.

Bing

Our new data API, businesses data searchable by radius and completely free.

Python Files

Intersect.py

Takes a spreadsheet with latitudes and longitudes and a shapefile and returns the spreadsheet with only points in the shapefile.

Filter.py

Takes a spreadsheet with latitudes and longitudes and a spreadsheet of points (plus a radius) and returns the spreadsheet with only the points inside the radius.

Cluster.py

Takes a spreadsheet and performs a merge on itself. In doing this, it replaces duplicates while simultaneously finding addresses with multiple businesses. Outputs two files, one with the business clusters and one with duplicates removed.

Useful_functions.py

This file contains many useful functions and should be looked over before attempting a data manipulation or calculation.

Jupyter Notebooks

API request.ipynb

This file contains the method which uses the Google Places API, Bing Maps Locations API, OpenStreetMaps API to query a radius around given points. This file is also a good example of some data manipulations.

Merge.ipynb

This file contains a method to merge all of our data resources into one list.

Plotting.ipynb

This file shows various plotting and filtering of points to use for intersect.py. It has lots of good examples of the power of plotting through geopandas.