Statistical Modelling with Python

Theo Plessas

Learning Objectives

* Accessing data using APIs
* Cleaning and transforming data using Python
* Loading data into a database using Python
* Performing EDA, including using both statistics and visualizations
* Identifying trends and patterns in data using statistical models
* Interpreting the results of the statistical models

My Goals

- investigate the relationship between the average distance each venue result is from each bike station and the distance between the bike station and the city center
- Can i model a relationship between the number of bikes available and the distances

Citybik.es

- Simple requests with No authentication
- Hamilton ON is my target
  - Location data is nested JSON
  - Flatten and filter by city
- Use the gbfs_href link to request bike data
  - Final result was 138 bike rental stations data across Hamilton
- Flatten any remaining columns
- Store as a csv

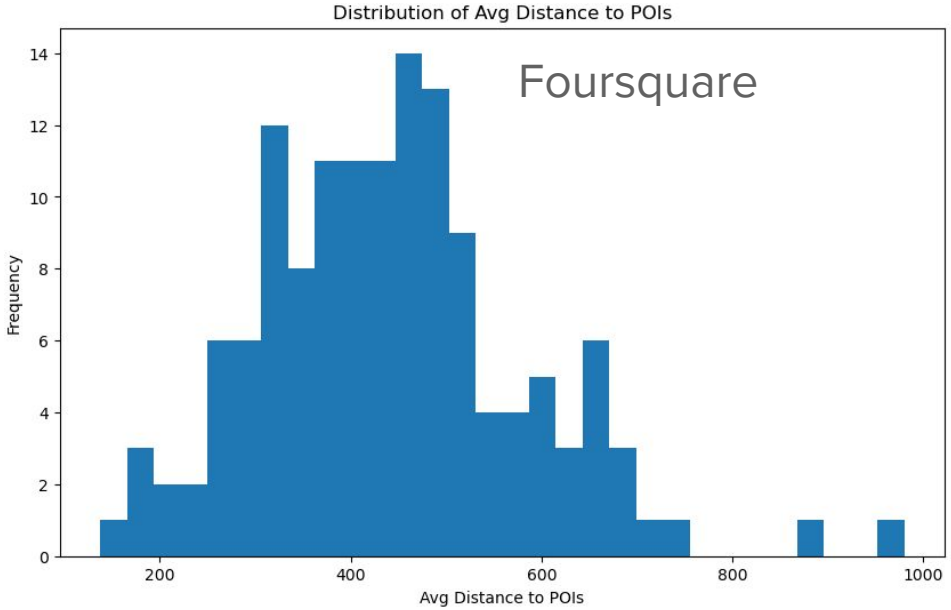
Yelp & Foursquare

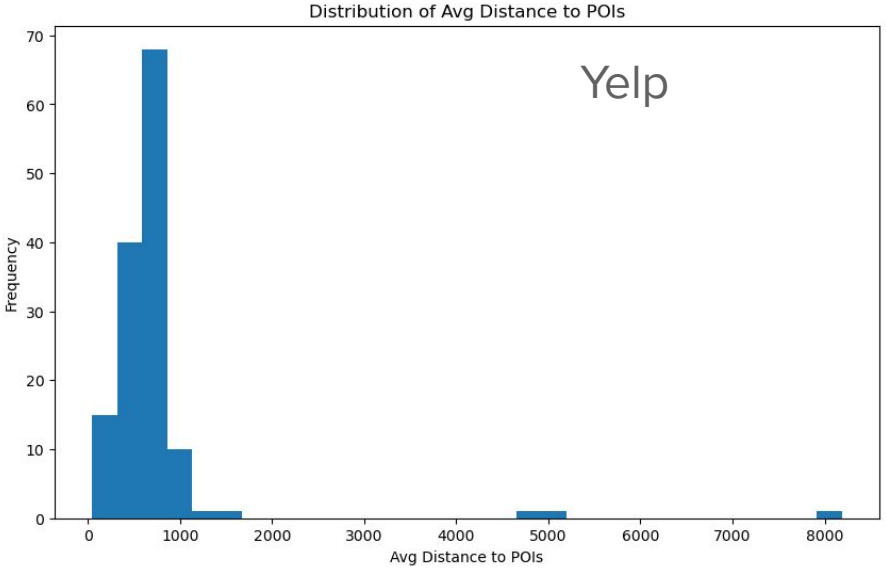
- .getenv to load my API keys into the notebook
- Define 2 python methods to request location data based on Coordinates
    - Foursquare method was used earlier in course work so no problem
    - Yelp was a modification of the FS method due to different request requirements

- Using a for loop on the bike station
    - Request data for every station given its coordinates and append the JSON data to each station
    - Flatten the JSON data into individual venues and maintain which bike station id it was associated with (like a foreign key)
- Save the Yelp and FS data separately as csv

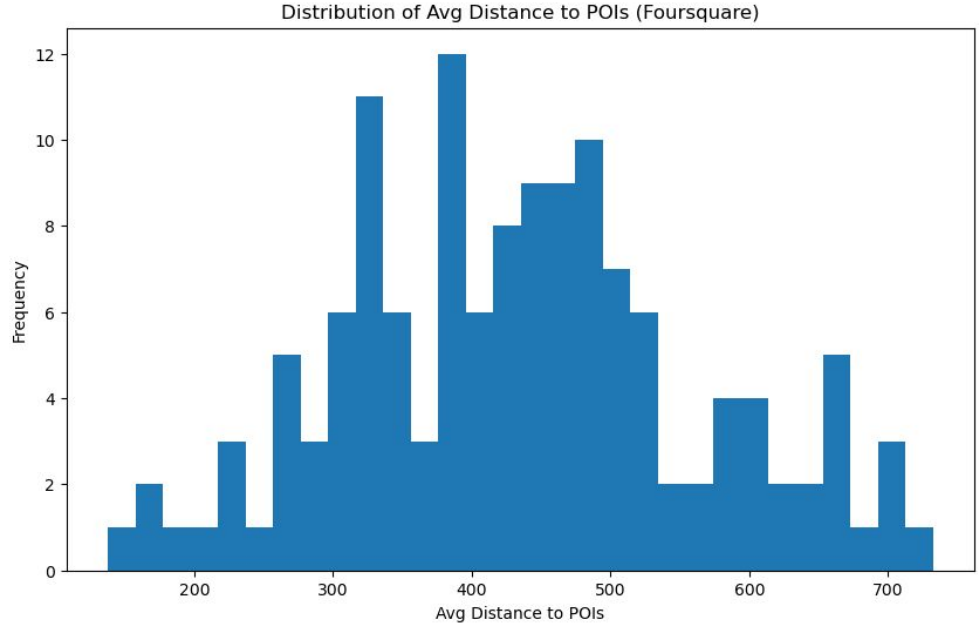
EDA and visualizations

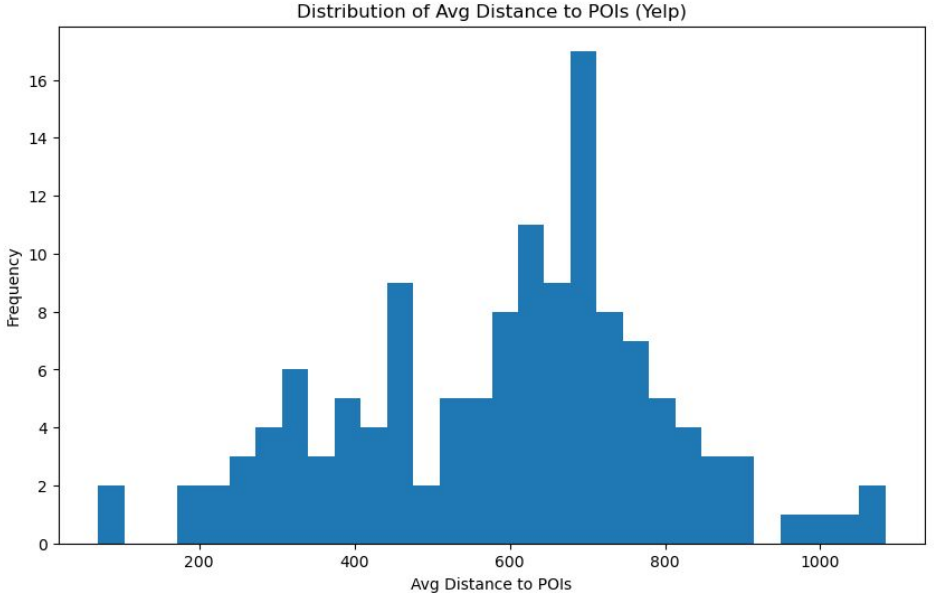
investigate the relationship between the average distance each venue result is from each bike station and the distance between the bike station and the city center



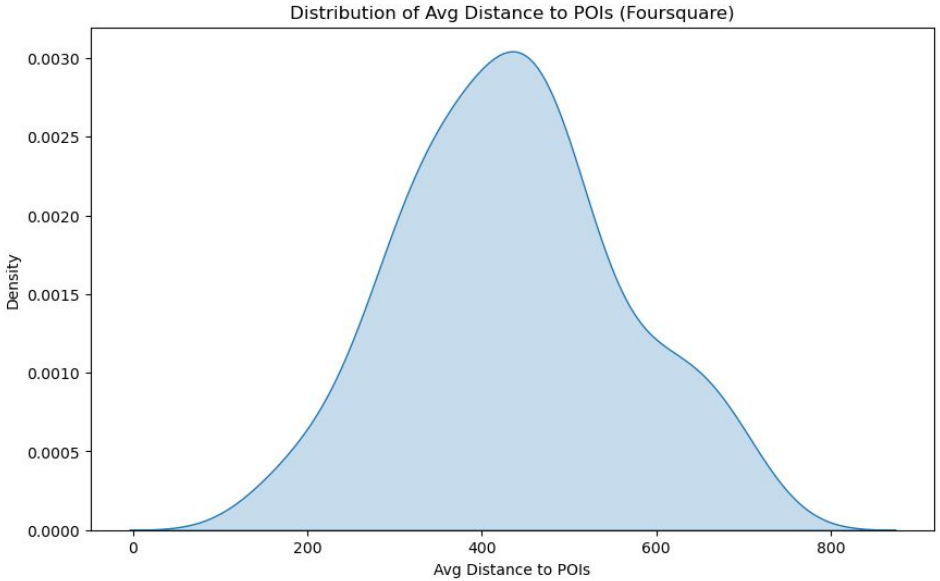


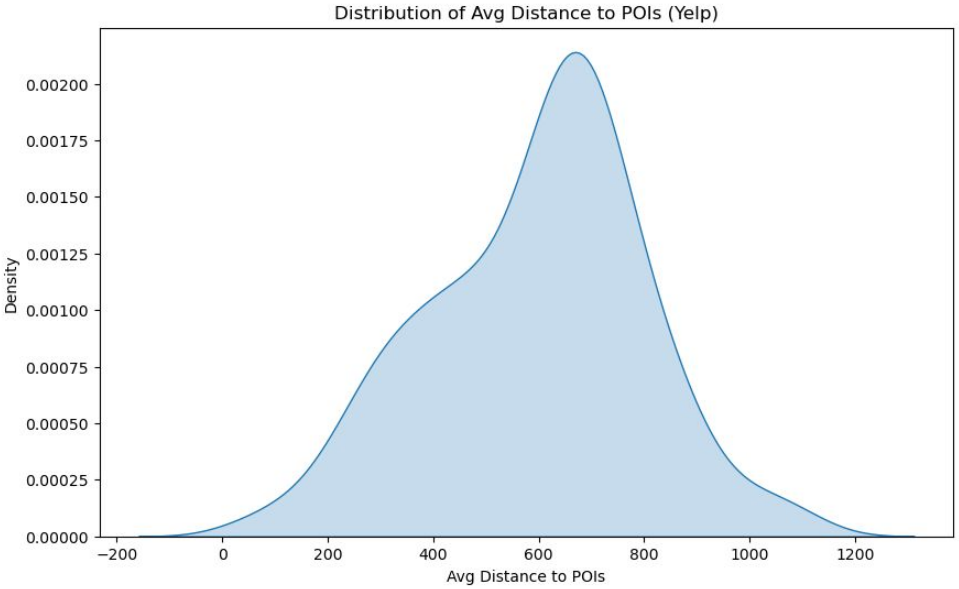
EDA and visualizations





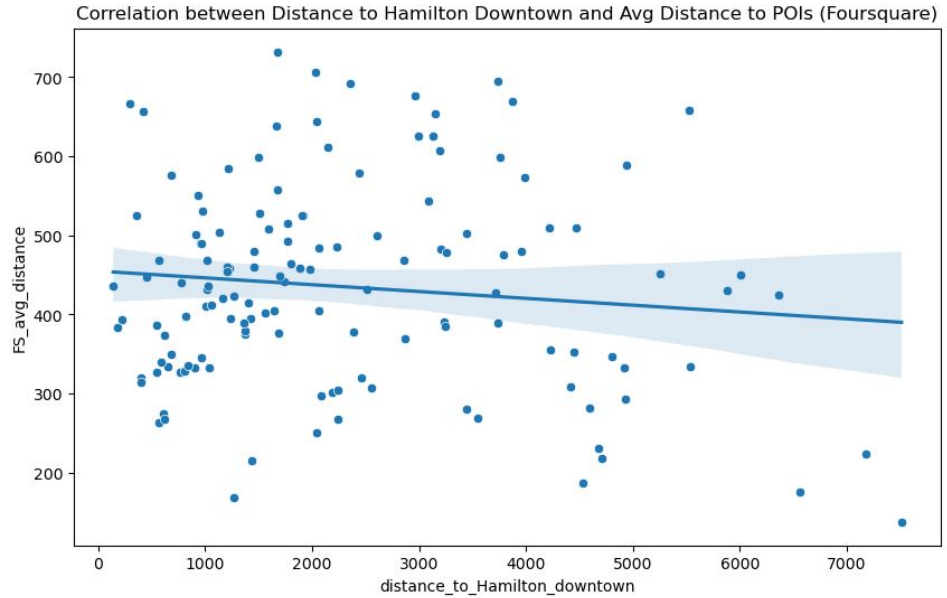
EDA and visualizations

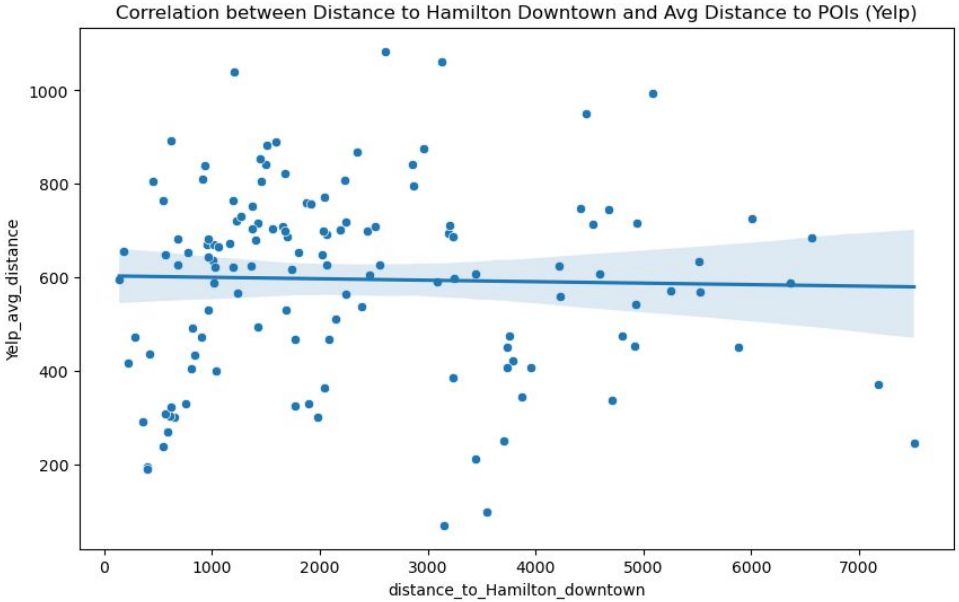




EDA and visualizations

investigate the relationship between the average distance each venue result is from each bike station and the distance between the bike station and the city center

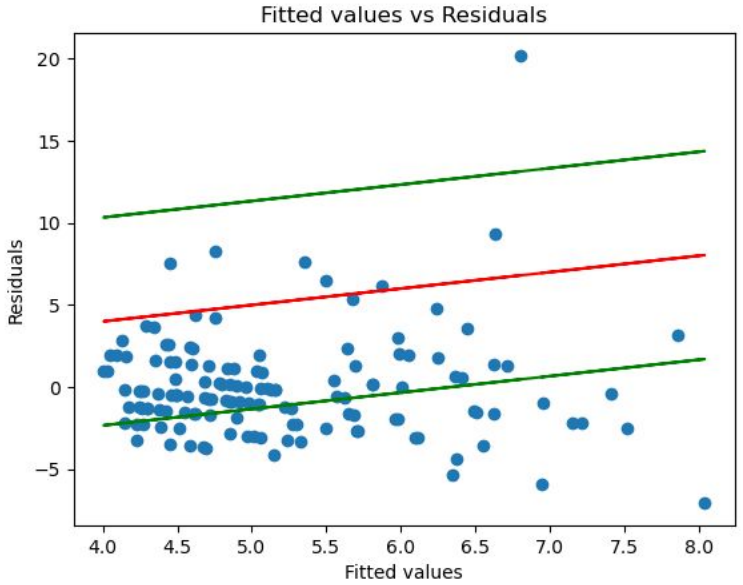




Model Building

Can i model a relationship between the number of bikes available and the distances described above i.e. distance away from the center, avg Foursquare, and Yelp venue distances from the Bike stations.

- Backwards elimination
- OLS Regression
- Final model
    - Only 1 feature remained



Thank you!