Selecting the best location for a Restaurant

Samuel Schatz

1. Introduction

    1.1. Background:

    Restaurants are a big part of today's society, thus making a need for them to be around. But this means that there is a need for new restaurants to come into cities. This not only for a variety of places to choose from it also helps stimulate the economy; not only with jobs but with income from the said restaurant.

    1.2. Problem:

    The main problem is that when a restaurant is choosing an area to come into it an area it may already be populated thus forcing the new restaurant to go out of business or not thrive. In a shorter term the problem is where to place a restaurant.

    1.3. Interest:

    The interest behind his project is to see out of the 4 major cities and their surrounding county to see what restaurants would thrive given the already existing restaurants.

2. Data acquisition and cleaning

2.1. Data Sources:

For the Latitude and Longitude for each city I got it from the GeoHack tool, this is used in the mapping of the restaurants and other tools. Which can be found in the 3.1.*.1 section of data sources. For the population data I got it from Wikipedia are cities or counties, which can be found in 3.1.*.2 of the data source section. For the population statistics and Restaurant info I got from FourSquare by searching by Coordinates and Zip Codes. Then for the Restaurant Price Rating I got that using Google's Places API. The last data source was uszipcode · PyPI package which gave me information about the zip code and allowed me to search for data.

2.2. Data cleaning:

There was a lot of data cleaning in this project due to NaN values and data that would throw off the data. The lack of accurate data reduced the data set, this factored with the fact that getting data from my data source is extremely tedious due to the many inputs that it requires severely reduced the size of my dataset.

First, problem was that the Google's Places API would return a NaN value for the price point which would make graphing and categorizing that data impossible. The API returned NaN because the data wasn't available because the owner of the store did not enter into the info section on google.

Second, almost all of the data would have duplicates, so I had to make a choice which duplicates I would allow. The data that I allowed to be duplicates was the population data and the Housing data this is because I got it from two sources: uszipcode and FourSquare. This is okay because uszipcode data is broken down into each cities zip code and FourSquare is a lot more vague than the uszipcode package.

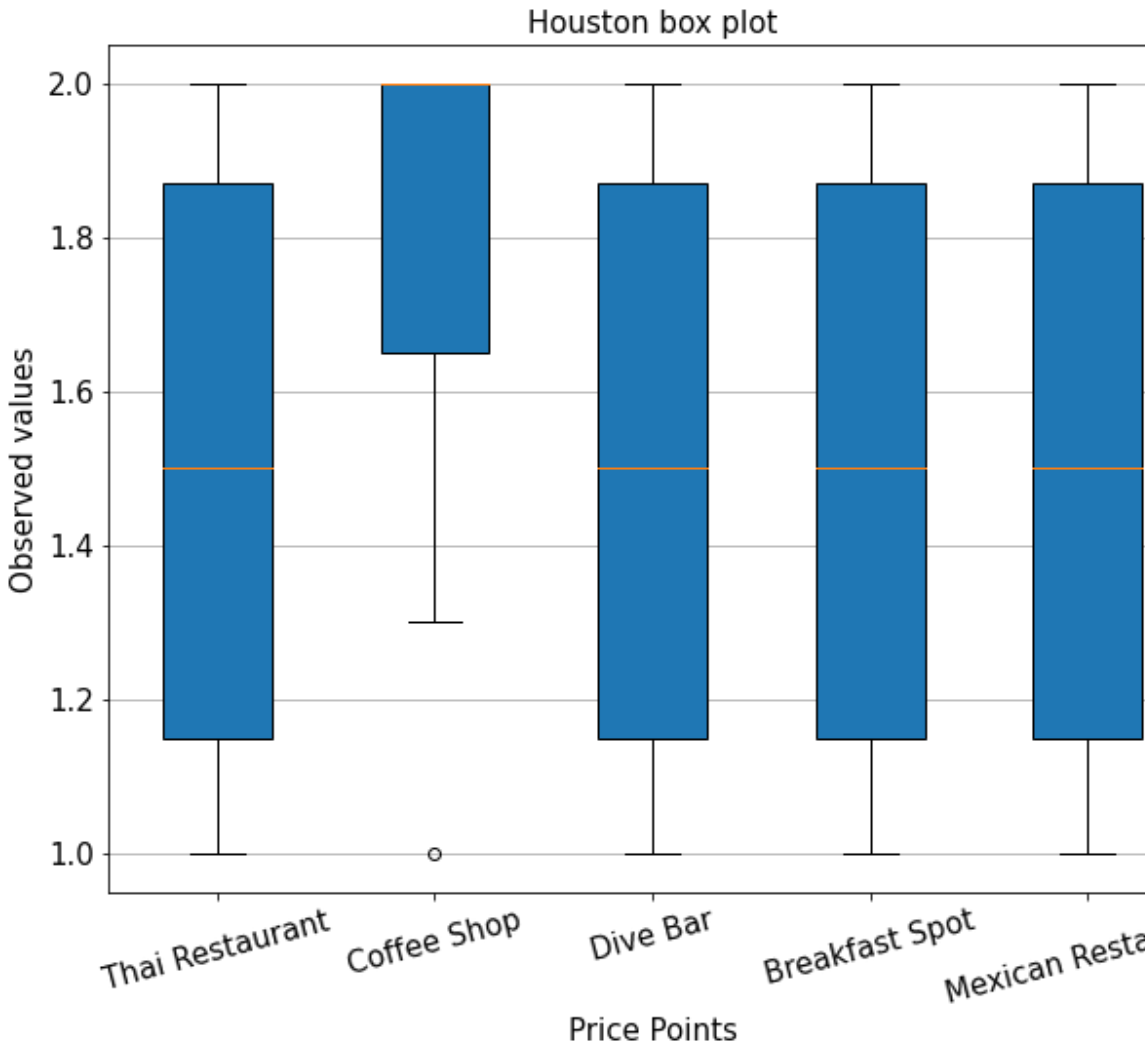
2.3. Methodology:

2.3.1. Exploratory Methods:

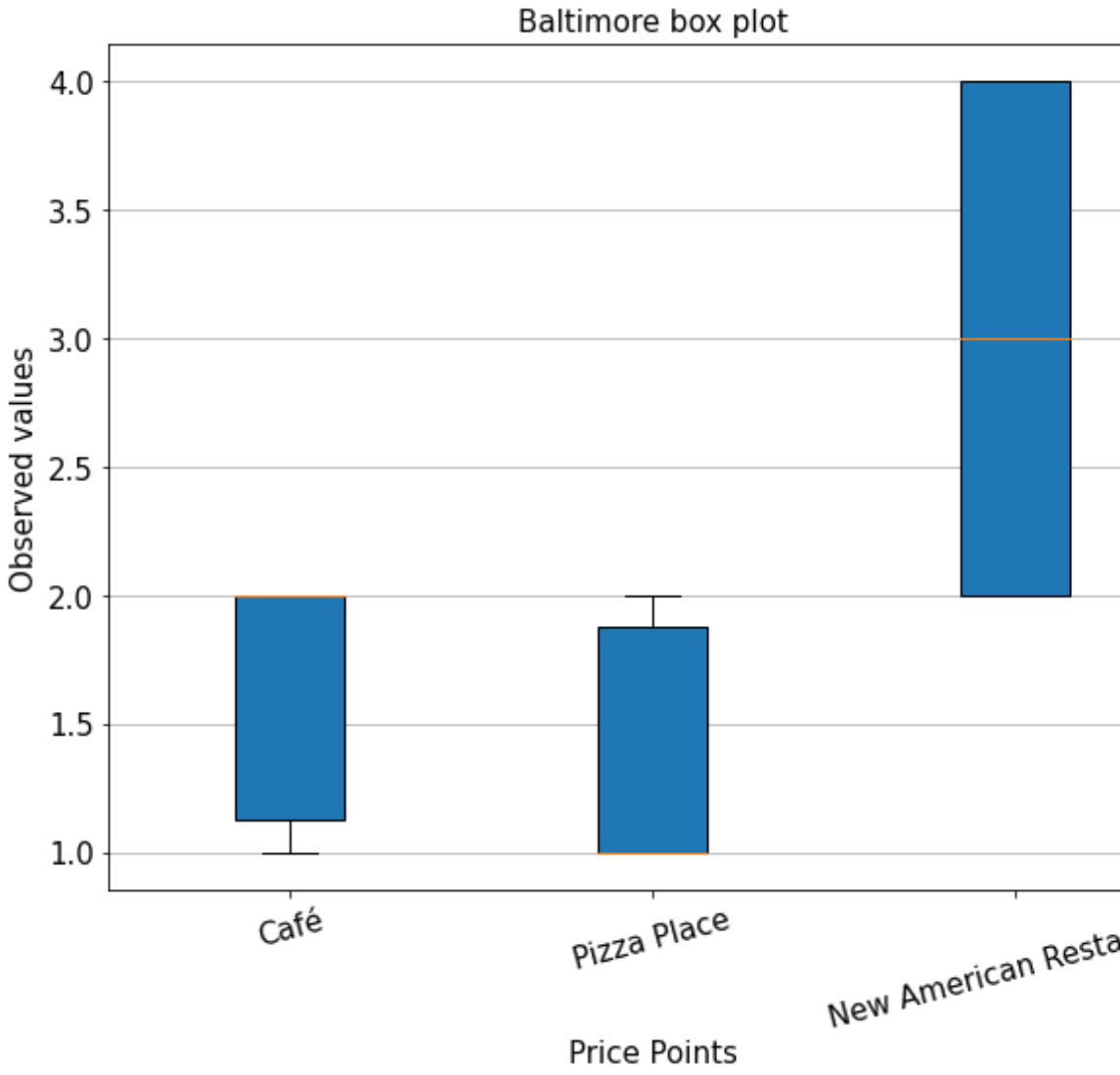
The first thing I explored was the Restaurant type by its price point for the outliers of the data set, the price point are determined by google users by rating it from a score of one to four. This was done by getting the entries that had multiple different price points throughout each city. What I learned from it that a lot of these Restaurant types only had one or two of the same type in the city.

From this first method I learned Houston had the price point from its outliers that were graphed. All five categories had a max

price point of two and a low of two. As seen below:

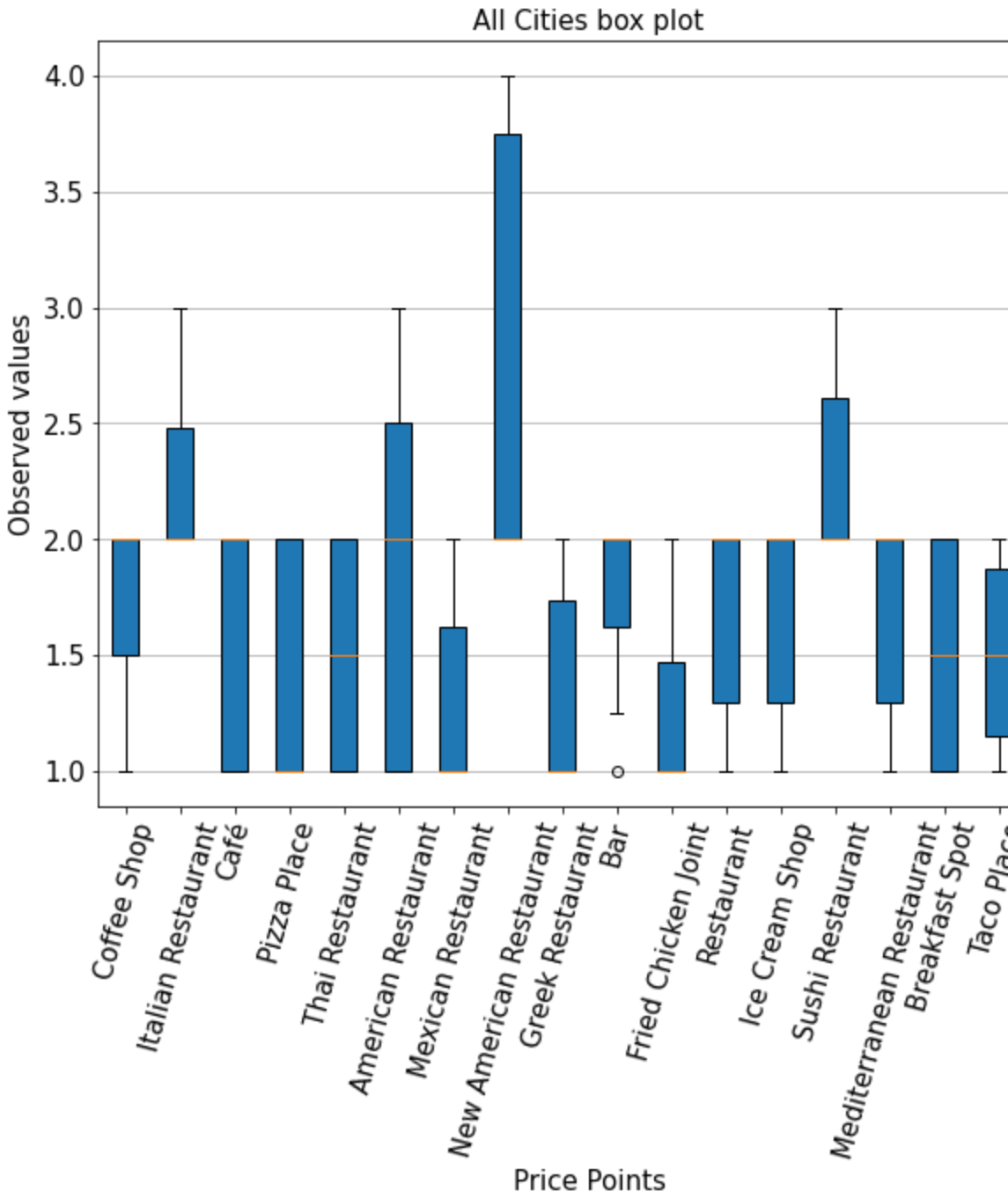


Than the category with the highest number was Baltimore with a max of four but only had three entries, as seen below:



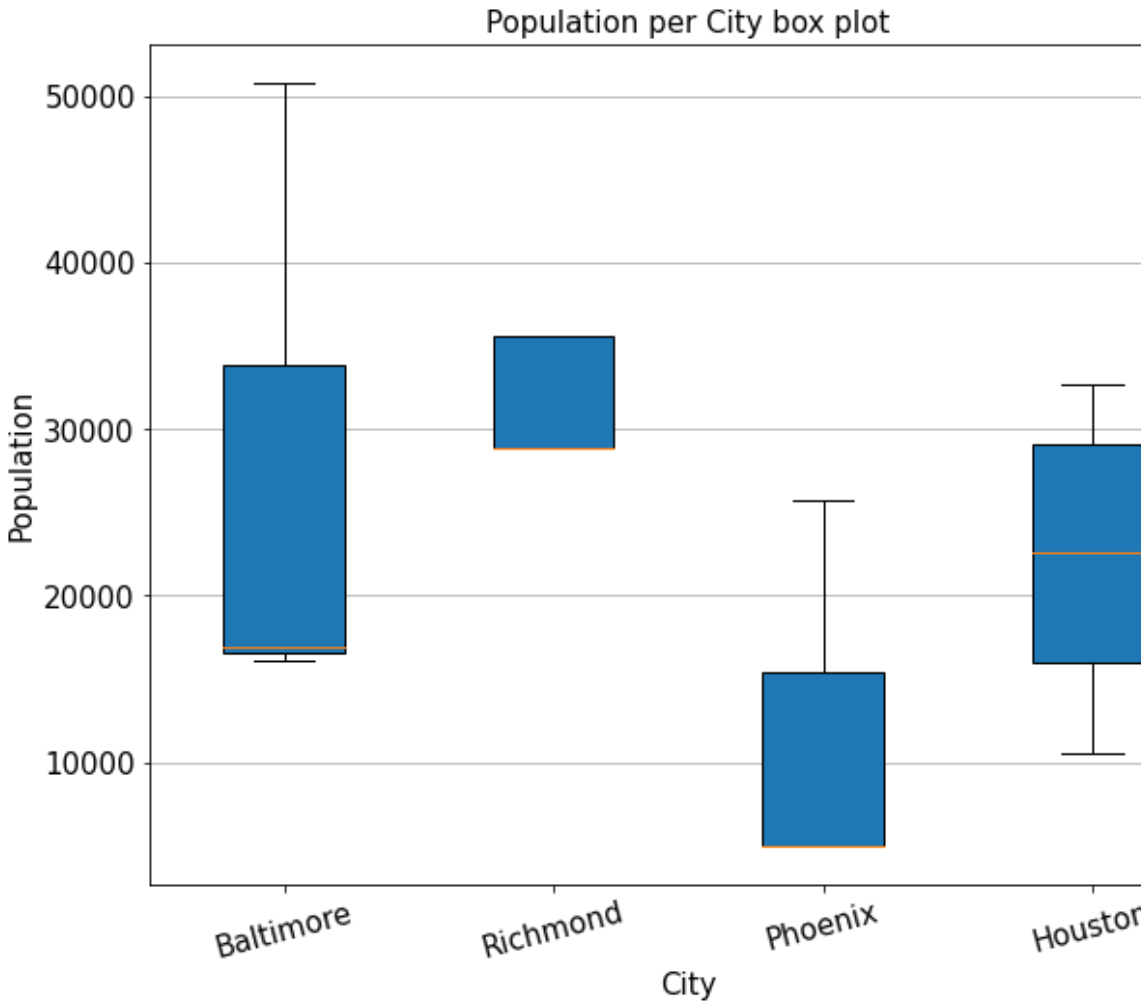
Than from all of the four cities combined the highest price point was Mexican Restaurant and the rest where pretty level as

seen here

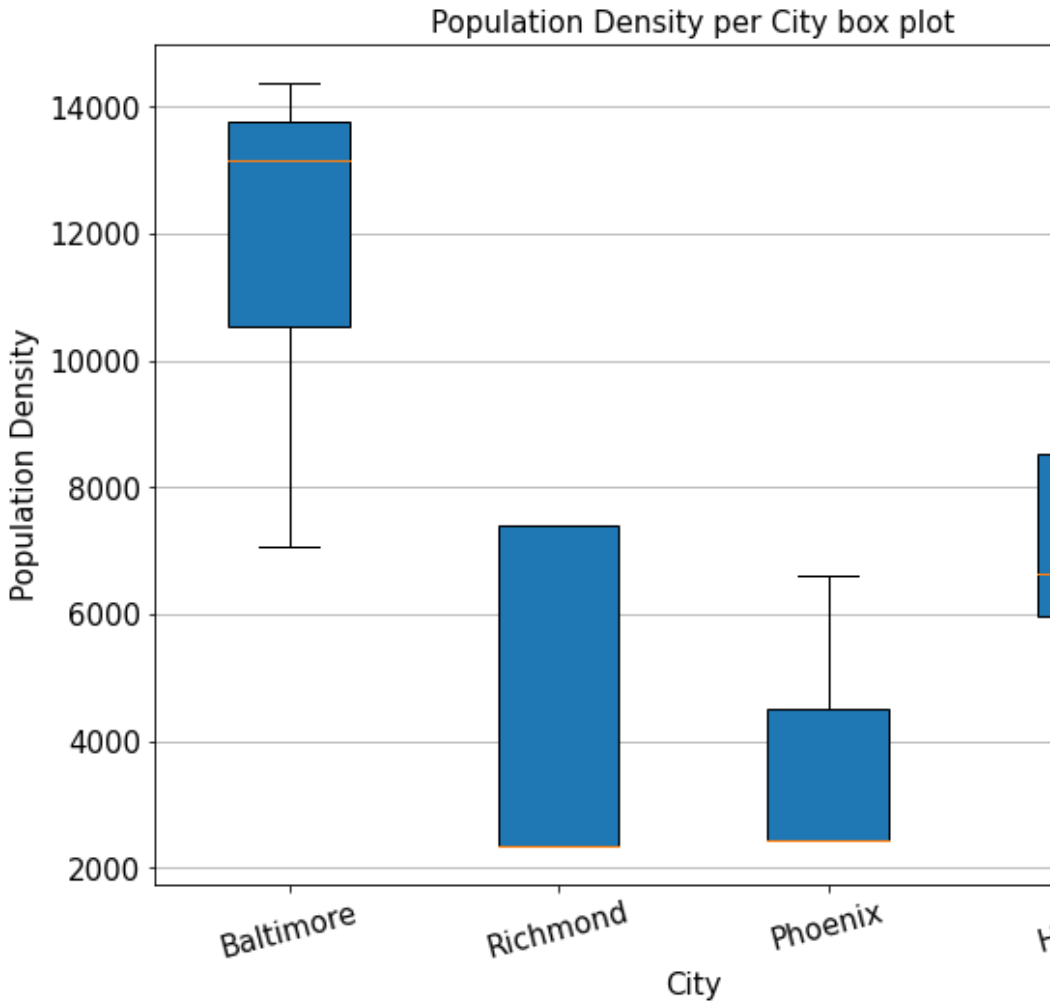

All Cities box plot
Observed values
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Coffee Shop
Italian Restaurant
Café
Pizza Place
Thai Restaurant
American Restaurant
Mexican Restaurant
New American Restaurant
Greek Restaurant
Bar
Fried Chicken Joint
Restaurant
Ice Cream Shop
Sushi Restaurant
Mediterranean Restaurant
Breakfast Spot
Taco Place
Price Points

After that I explored the Population and the Population Density of each city. The low and high of the population and population density is collected by each zip code in the city.

With the the city with the most population was Baltimore with a max of 50,846 and a low of 16,972. While Phoenix had the lowest population with a max 25,742 of and a low of 4,965. Then Richmon had the smallest distribution with a high of 35,545 and a low of 28,908. As seen below:

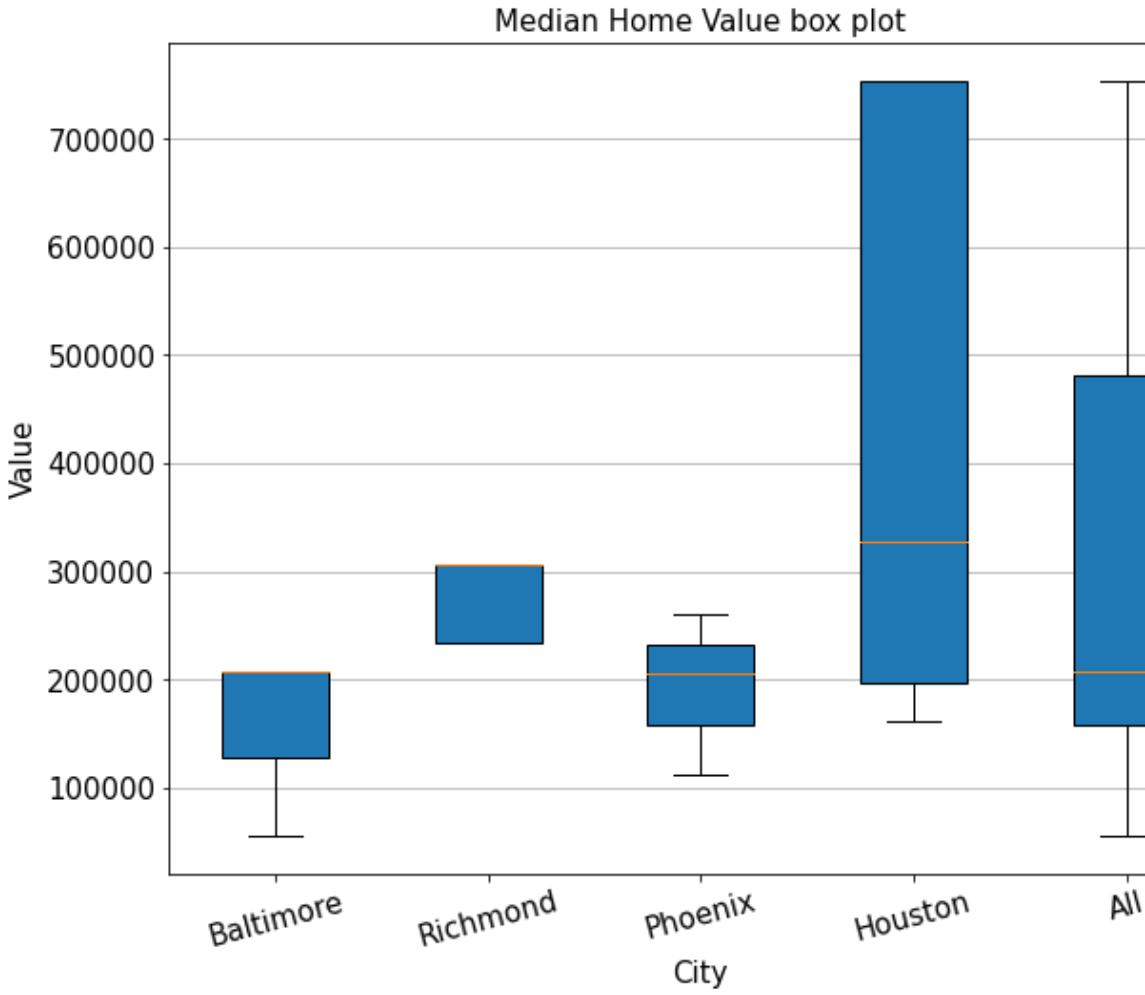


Than for population density Baltimore again had the highest with 13,158 and low of 7,883. While Phoenix and Richmand had the around the same low with around the same distribution as seen below:



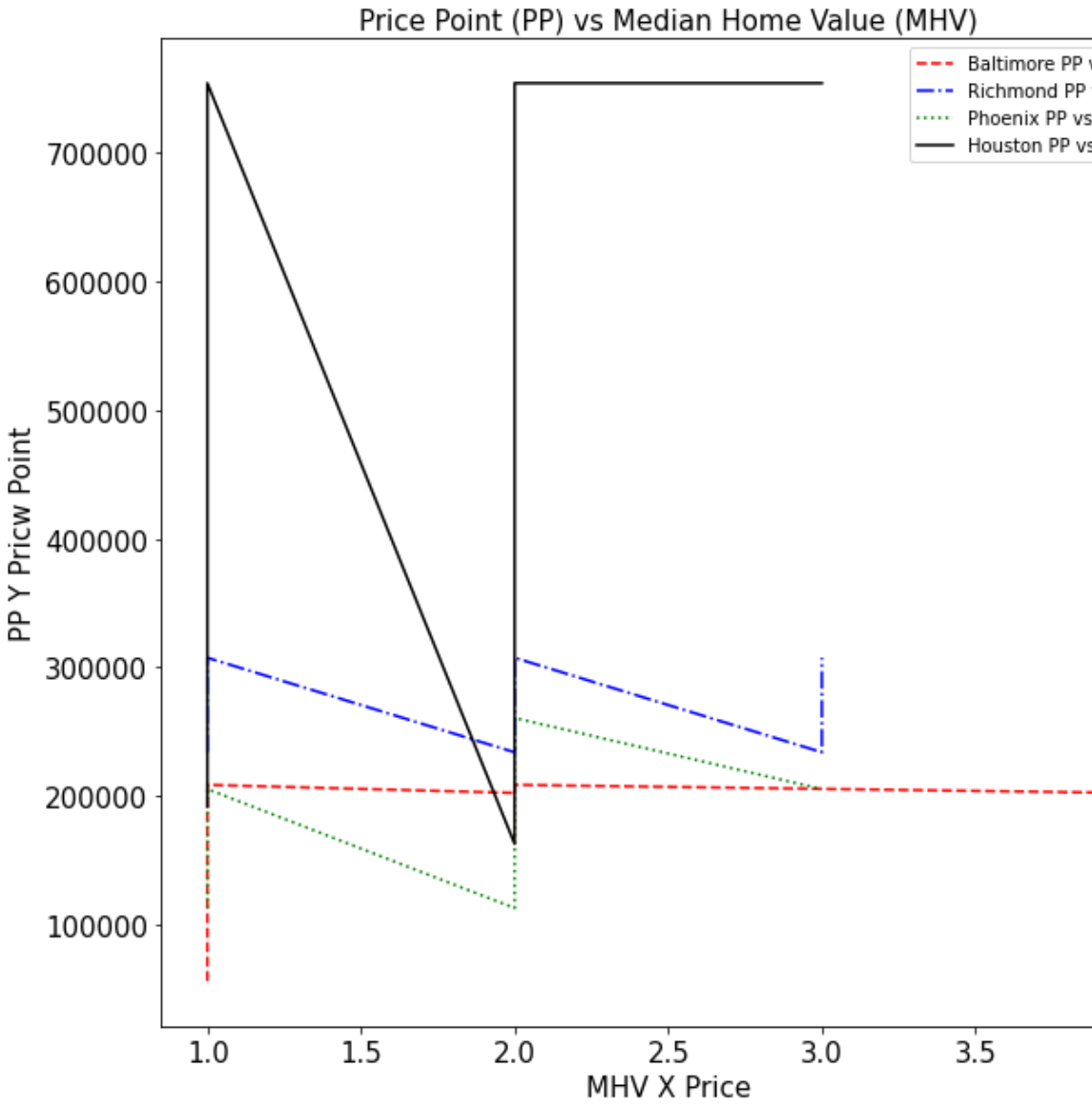
After that I explored the median home value for each city. Baltimore had the lowest with $55,900 and high of $208,500. Then Houston had the highest by almost $400,000 with a high of $754,000. Than continuing the trend Richmond had the smallest

distribution. As seen below:



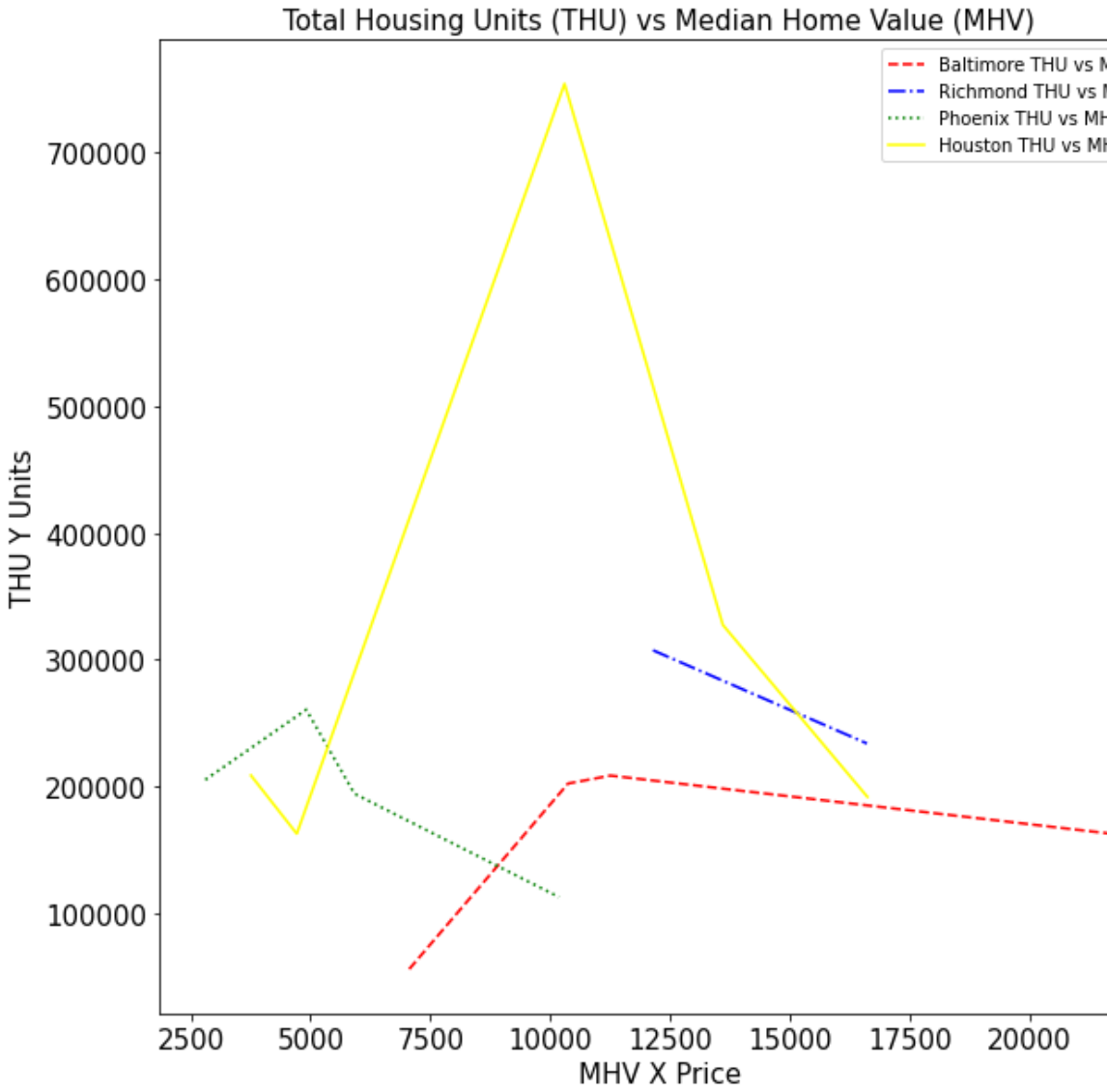
Than I explored the relationship between Price Point (PP) vs Median Home Value (MHV) and found no correlation I think that is because I did not have enough info between the two features. This

is proven bellow:



After that I explored the relationship between Total Housing Units (THU) vs Median Home Value (MHV) and the results from the

graph was inconclusive. As seen below:



3. Results:

The results of this project can be inferred and be used to make an informed decision on where to open a restaurant based on the restaurant type you want to open. So in my case the project was successful because of how the data can be used for a variety of people and use cases. Although I wanted to use machine learning the data size was not large enough or good enough to do so. I found that disappointing to me.

4. Discussion:

I observed that the higher the median house value does not correlate to the restaurant pricing in the area, which can be seen in Houston. But is the opposite as seen in Baltimore which had a higher population and higher restaurant pricing. This means that pricing and population correlate with both population and population density. This means a restaurant owner might be better opening in a higher density area rather than a higher median house value area.

5. Conclusion:

In this project I analyzed factors that a restaurant owner would take into consideration when opening a restaurant. I identified many factors such as; Price Point, Population, Population Density, and many housing and income factors. These graphs and data can be used to make a decision in which of the four cities to open a restaurant.

6. Sources:

6.1. Data Sources:

6.1.1. Baltimore:

6.1.1.1. Coordinates: https://geohack.toolforge.org/geohack.php?pagename=Baltimore&params=39_17_N_76_37_W_type:city(650000)_region:US-MD

6.1.1.2. Population: https://en.wikipedia.org/wiki/Baltimore#Population

6.1.2. Baltimore County:

6.1.2.1. Coordinates: https://geohack.toolforge.org/geohack.php?pagename=Baltimore_County,_Maryland&params=39_24_N_76_36_W_type:adm2nd_region:US-MD

6.1.2.2. Population: https://en.wikipedia.org/wiki/Baltimore_County,_Maryland

6.1.3. Richmond:

6.1.3.1. Coordinates: https://geohack.toolforge.org/geohack.php?pagename=Richmond,_Virginia&params=37_32_N_77_28_W_

6.1.3.2. Population: https://en.wikipedia.org/wiki/Richmond,_Virginia

6.1.4. Henrico County:

6.1.4.1. Coordinates: https://geohack.toolforge.org/geohack.php?pagename=Henrico_County,_Virginia&params=37.55_N_77.4_W_type:adm2nd_region:US-VA_source:UScensus1990

6.1.4.2. Population: https://en.wikipedia.org/wiki/Henrico_County,_Virginia

6.1.5. Phoenix:

6.1.5.1. Coordinates: https://geohack.toolforge.org/geohack.php?pagename=Phoenix,_Arizona&params=33_27_N_112_04_W_region:US-AZ_type:city(1445632)

6.1.5.2. Population: https://en.wikipedia.org/wiki/Phoenix,_Arizona

6.1.6. Maricopa County:

6.1.6.1. Coordinates: https://geohack.toolforge.org/geohack.php?pagename=Maricopa_County,_Arizona&params=33_30_50_N_112_28_33_W_region:US-AZ_type:adm2nd

6.1.6.2. Population: https://en.wikipedia.org/wiki/Maricopa_County,_Arizona

6.1.7. Houston:

6.1.7.1. Coordinates: https://geohack.toolforge.org/geohack.php?pagename=Houston&params=29_45_46_N_95_22_59_W_region:US-TX_type:city(2099451)

6.1.7.2. Population: https://en.wikipedia.org/wiki/Houston

6.1.8. Harris County:

6.1.8.1. Coordinates: https://geohack.toolforge.org/geohack.php?pagename=Harris_County,_Texas&params=29.86_N_95.39_W_type:adm2nd_region:US-TX_source:UScensus1990

6.1.8.2. Population: https://en.wikipedia.org/wiki/Harris_County,_Texas