Utilizing Yelp Cost Estimates for Estimating Neighborhood Affluence

DSI 10 Atlanta

Holly Bok, Houston Bova, Raven Hinson, Lisa Hwang, Julian Lafaye, Owen Swetenburg

Problem Statement

“This tool will estimate the affluence of a neighborhood based on Yelp data.

This tool will expect to get, as an input, a list of Yelp data corresponding to zip codes or names of neighborhoods and will estimate the median income.

While traditional methods typically estimate wealth of a locality based on demographic characteristics (e.g. unemployment rate), the novelty of this approach is in its use of big data related to commercial activity and cost of product and services as an indicator for affluency.”

TL;DR
Can we predict median household income based on Yelp data?

About Yelp

Yelp is a website that allows users to submit ratings and reviews for local businesses

While many reviews tend to be of restaurants, many other businesses reviewed include boutiques, grocery stores, doctors, mechanics, and more

Yelp was founded in July 2004 in San Francisco

At the end of 2019, there were 205 million cumulative reviews posted

The number of reviews makes Yelp a promising resource for neighborhood data

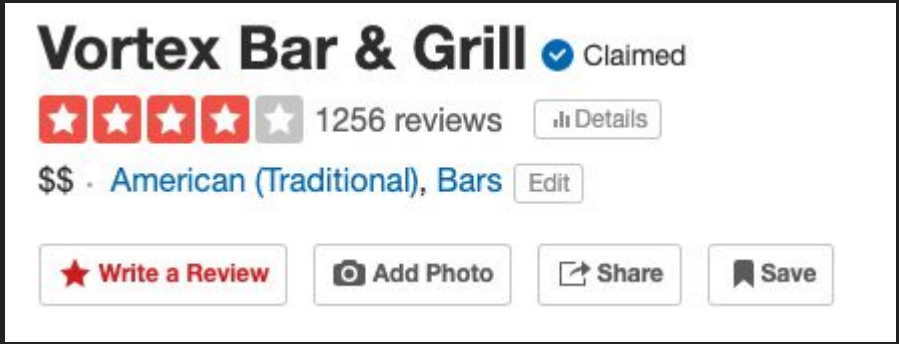
https://www.yelp-ir.com/

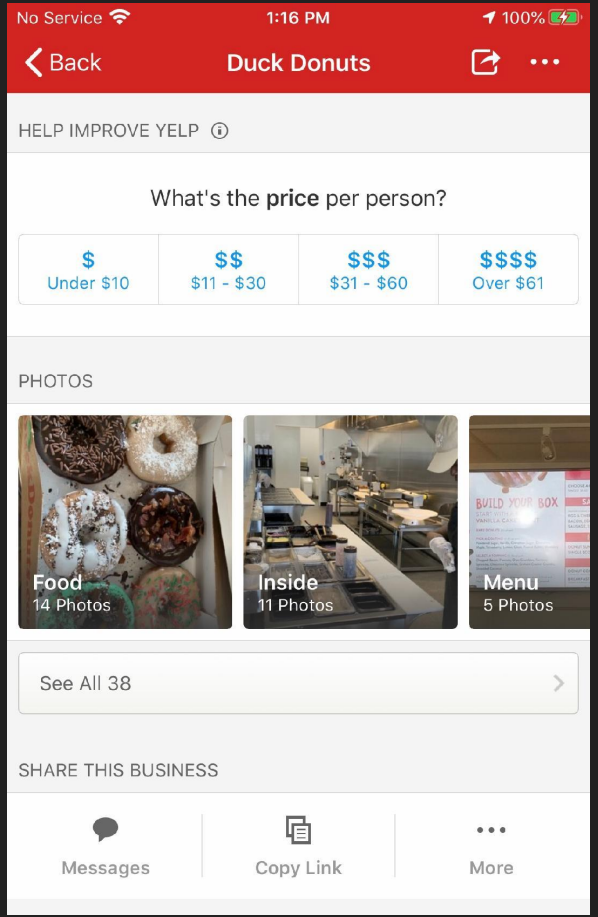
Yelp Pricing System

The average amount spent by each customer at a business is designed by one ($) to four ($$$$) dollar signs

Yelp pricing guidelines are not easily located on the site

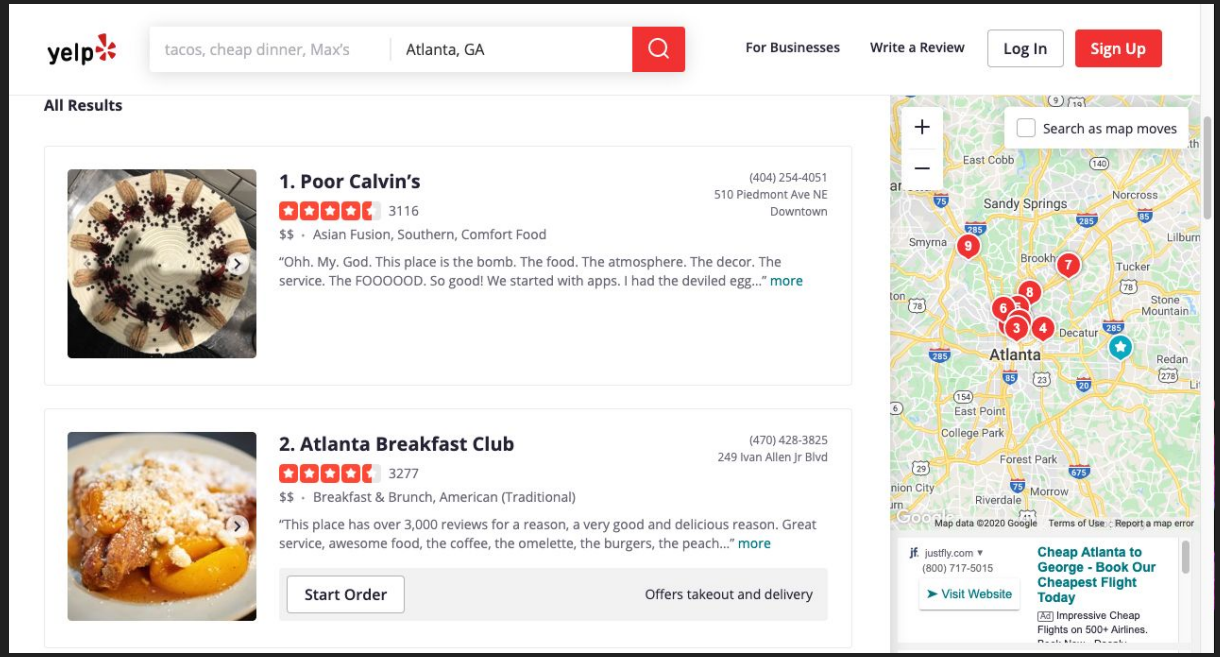
However, Yelp users are invited to vote for a business's $ assignment on the app





Yelp and Atlanta

Decided to focus on Yelp businesses in Atlanta due to our familiarity with the city and its neighborhoods



About Statistical Atlas

Located at statisticalatlas.com

Offers data from the US Census Bureau, particularly the 2010 Census and the 2012-2016 American Community Survey

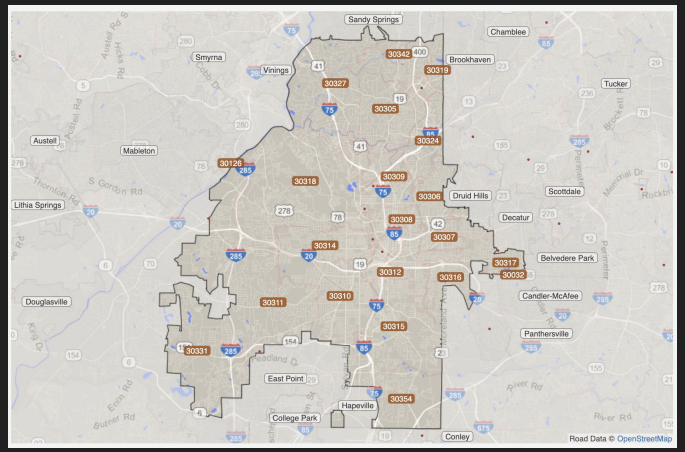
https://statisticalatlas.com/about

Atlanta Zip Codes

Statistical Atlas includes zip code information for cities

There are 30 zip codes for Atlanta

One zip code may contain several neighborhoods



State:	Georgia
Counties:	Cobb, DeKalb, Fulton
Metro Area:	Atlanta Area
ZIP Codes:	30030, 30032, 30126, 30303, 30305, 30306, 30307, 30308, 30309, 30310, 30311, 30312, 30313, 30314, 30315, 30316, 30317, 30318, 30319, 30324, 30326, 30327, 30331, 30332, 30334, 30336, 30342, 30344, 30354, 30363
Unified School Districts:	Atlanta, Cobb County, Fulton County

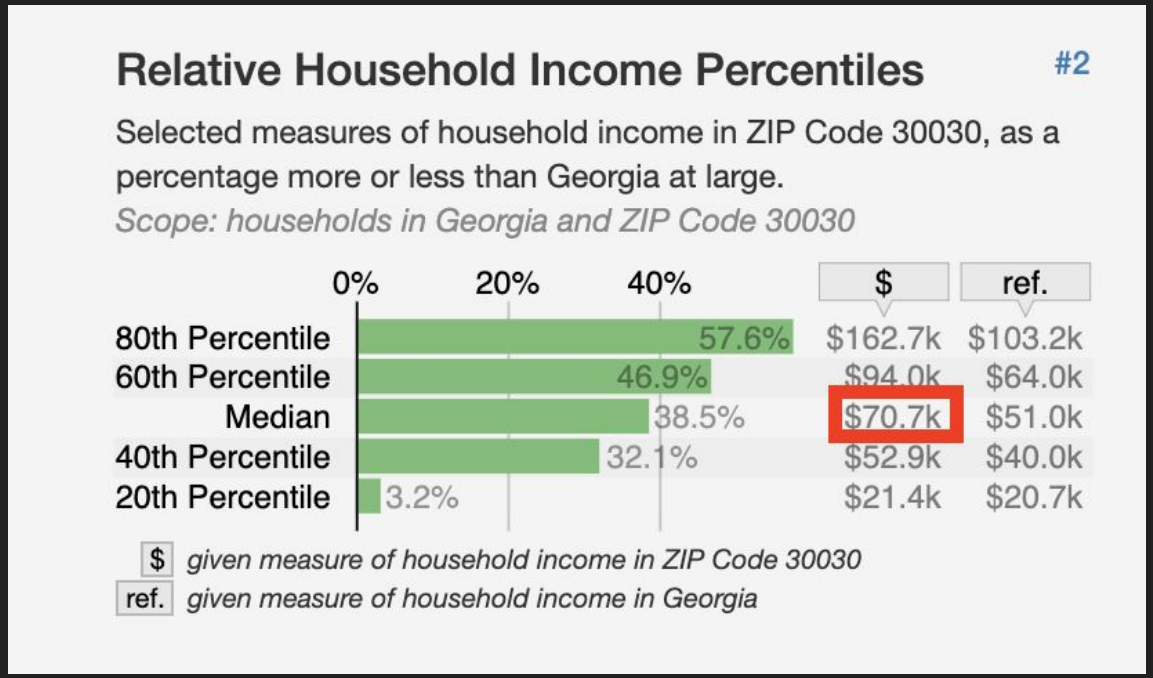
Atlanta, Georgia (City)

456,378
189,343

- **Overview**
- Population
- Age and Sex
- Race and Ethnicity

https://statisticalatlas.com/place/Georgia/Atlanta/Overview

Median Household Income



For each zip code, Statistical Atlas shows several graphs for household income

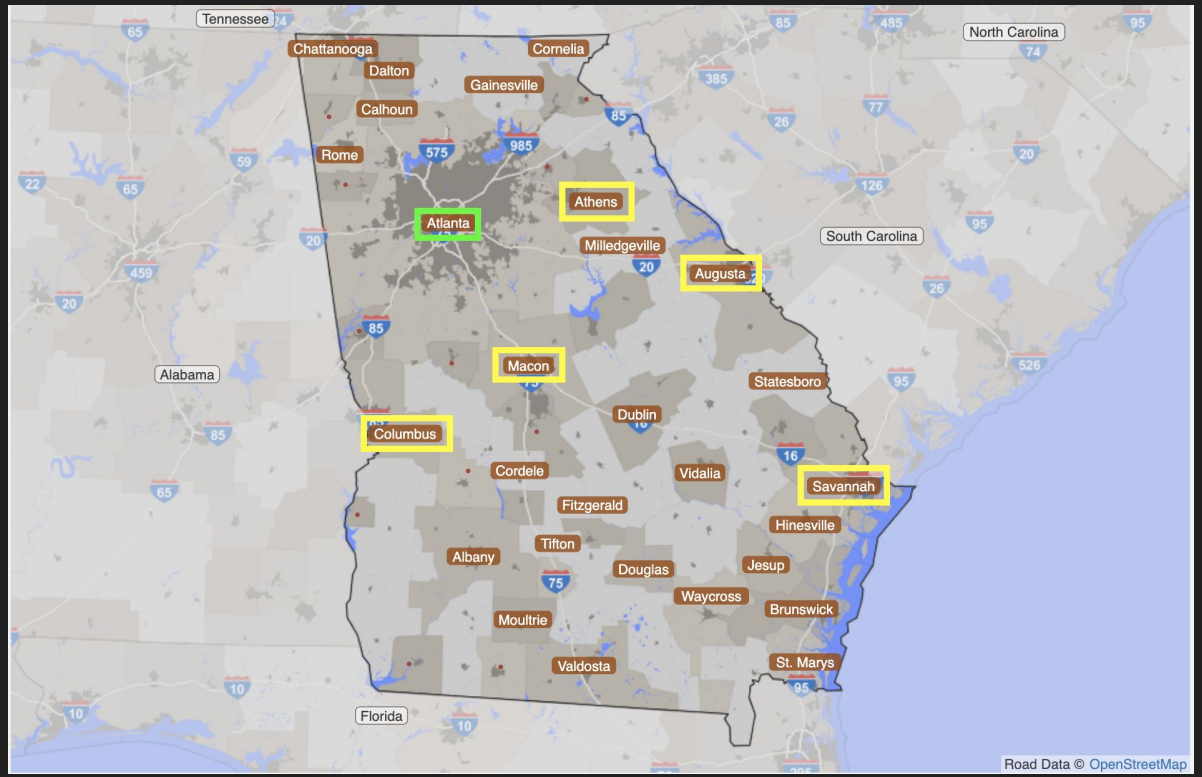
Median household income by zip code was obtained

https://statisticalatlas.com/zip/30030/Household-Income

Adding Additional Cities

For more data and modeling purposes, we added the following cities in Georgia:

- Athens
- Augusta
- Columbus
- Macon
- Savannah



https://statisticalatlas.com/state/Georgia/Overview

Data Collection

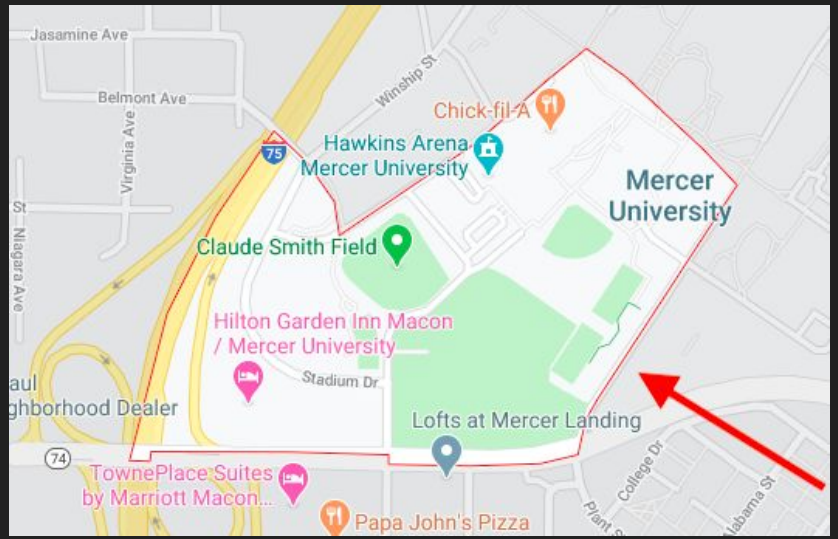
In order to help predict the affluence of a neighborhood, we scraped data from Yelp using an API.

Using the website statisticalatlas.com and BeautifulSoup, we scraped the median income of each zip code in Atlanta, Athens, Augusta, Columbus, Macon, and Savannah.

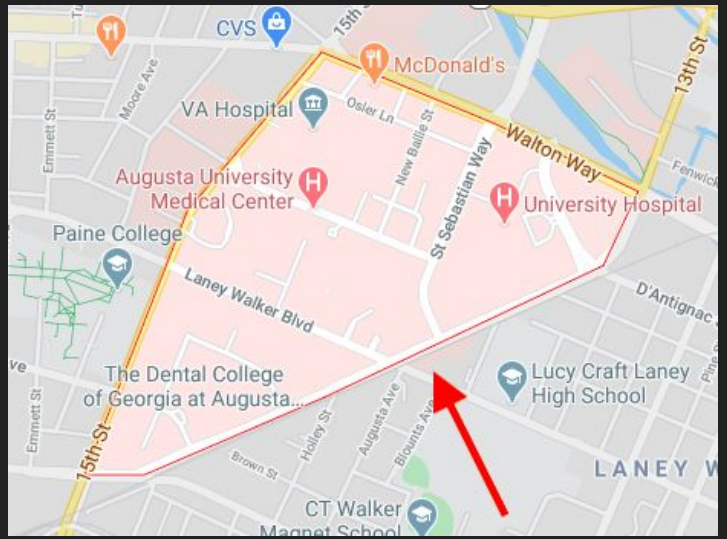
Nine zip codes presented problems due to page formatting and lack of income information. 78 zip codes are used in our model.

Zip codes with no income information:

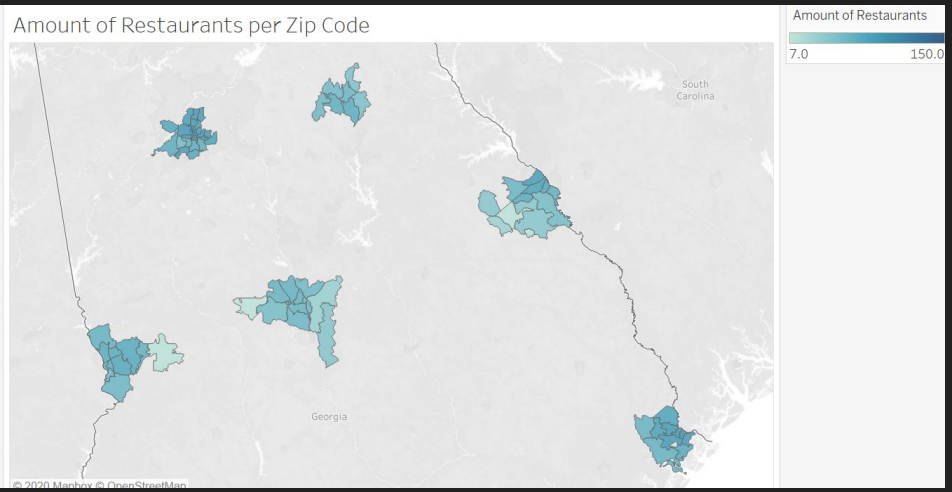
31207 in Macon

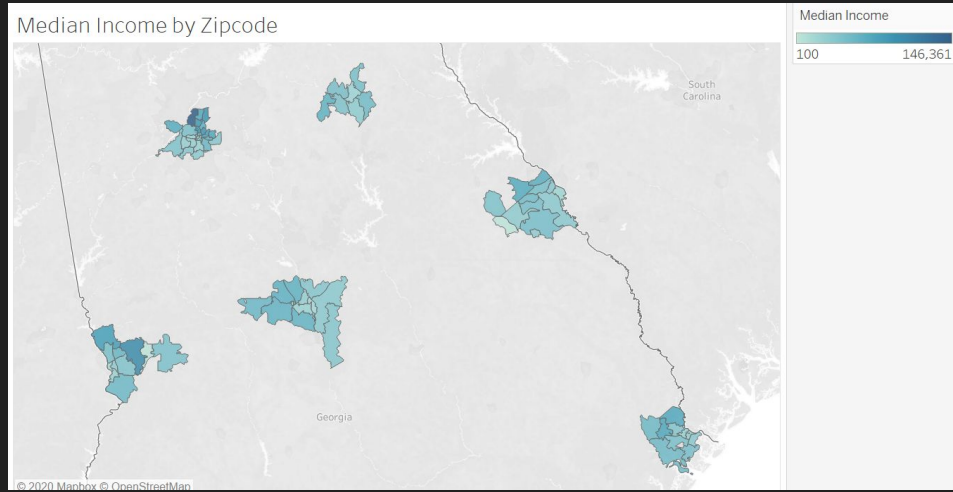


30912 in Augusta

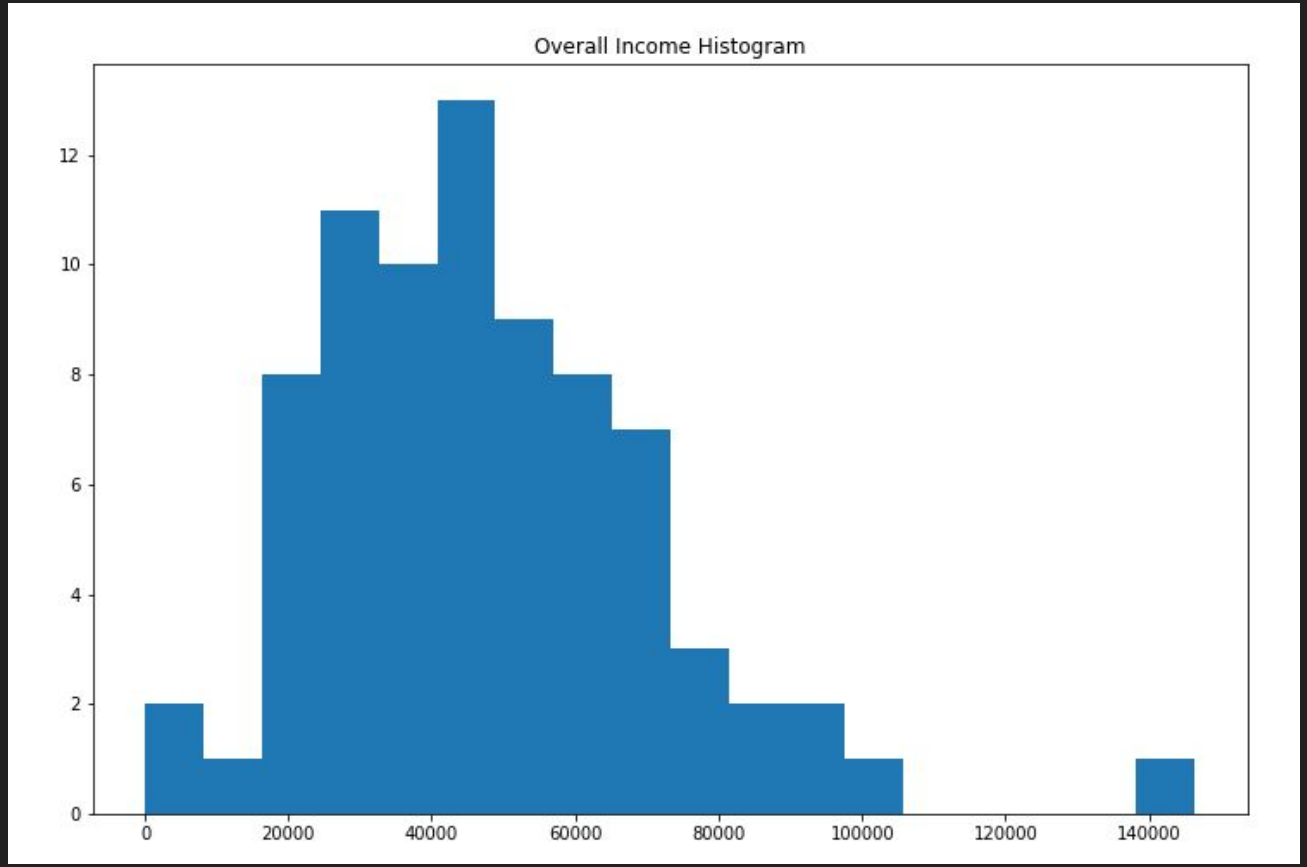


Comparing Amount of Restaurants and Median Income by Zip Code

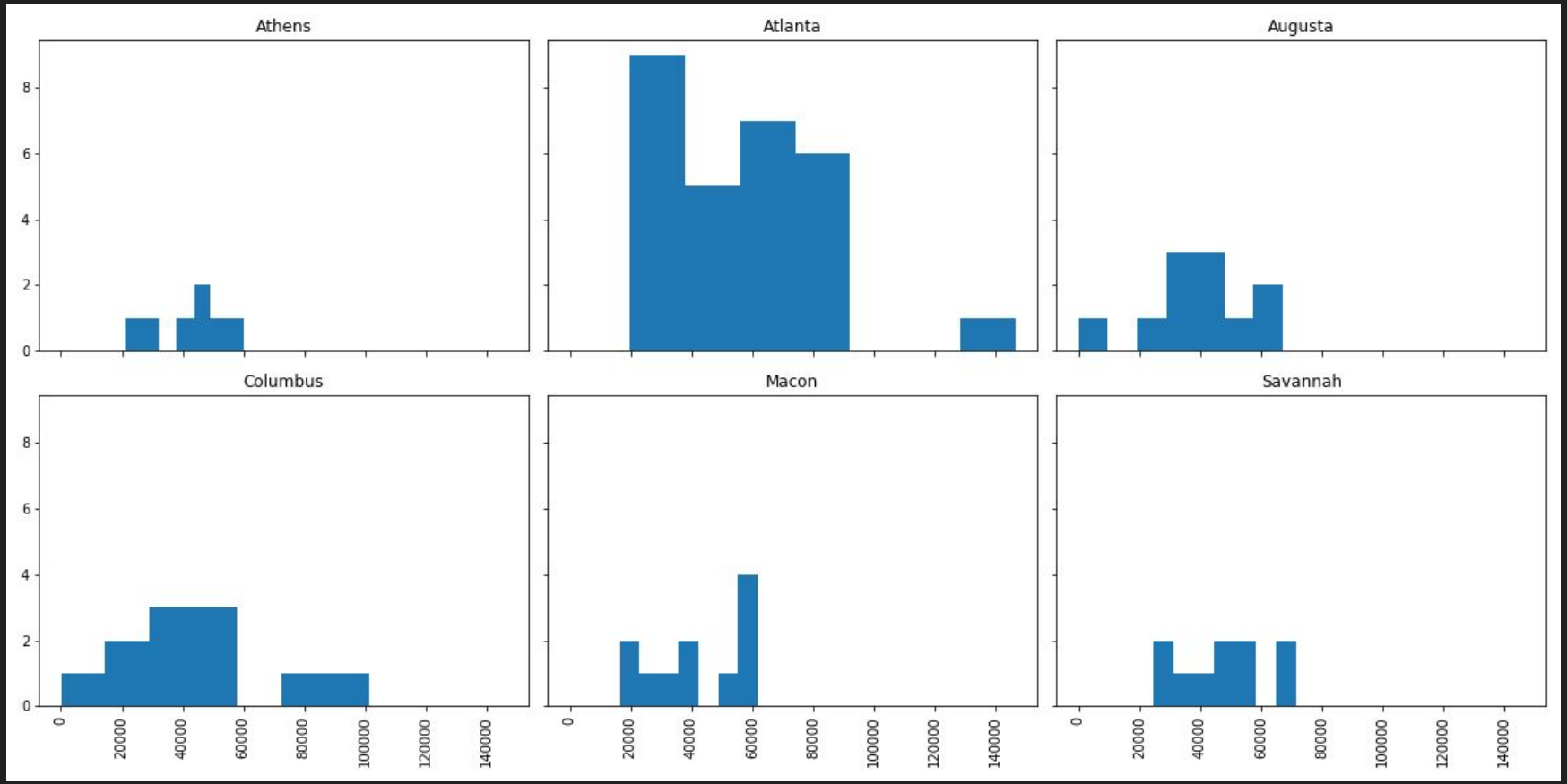




Overall Income Distribution:



Income Distribution by City:



Modeling

Linear Regression to look for coefficient trends

After seeing trends, it was passed through a Lasso model to see if they were substantial enough to build a model off of.

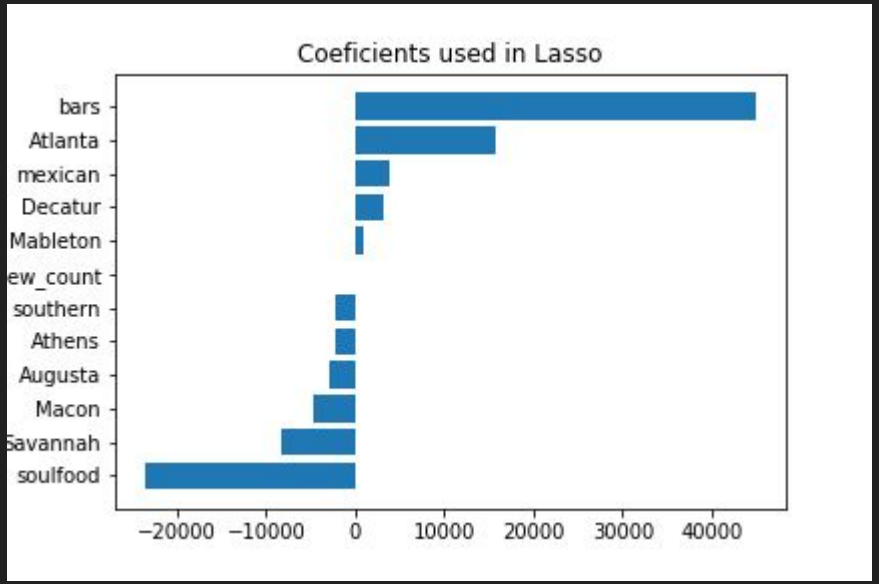
Lasso was the best performing model overall

The result was R^2 = .33, which isn't the best but better than 0.

Tried SVR and 8-layer Neural Nets model on all the data- still didn't work

Bar Graph of Categories

Graphs showing which categories were most telling of neighborhood affluence predicted using LASSO model against Data Collected



Limitations

Lack of Yelp reviews for regions with fewer business such as rural areas

Quality of Yelp data (duplicates, missing price information)

Inconsistency across cities in the Yelp data

Not much of a correlation between restaurant price and affluence of a neighborhood

Conclusions

Yelp data are not ideal for predicting neighborhood affluence

LASSO regularization might be the best option for this problem

Yelp data are geared towards apps and not data analysis

Interesting the best predictors where the categories of restaurants

Next Steps/Future Directions

- Collect more data (entire US) to improve neural net performance
- Combine Yelp data with other datasets such as housing sales for better insights
- How does the size of a zip code (sq mi) factor in?
- How does housing price data in zip codes correlated with the affluence of an area of town?

Thanks!