Utilizing Yelp Cost Estimates to Predict Neighborhood Affluence

Alex Lau | Despina Matos | Julie Vovchenko | Kelly Wu
DSI - NYC Tensors

New Light Technologies

Our work today focuses on:

- Preparing for emergencies
- Rapidly responding to emergencies
- Estimating the economic impact of disasters

Problem Statement

Can Yelp cost estimates ($, $$, $$$) determine neighborhood affluence in the Manhattan borough of New York?

Can we utilize neighborhood affluence to respond to emergencies?

Gathering and Cleaning of Data

Gathering the Data:

- Using Yelp API fusion
  - Creating a loop for the location and zip codes
- Using US Census data tables
  - Income for 2018
  - Filtering by zip codes

Cleaning the Data:

- Restaurant Businesses Only
- Dropped columns and nulls
  - Phone number and coordinates
- Isolations on columns
  - Address -> Zip Code
  - Categories -> Aliases
- Dummying columns

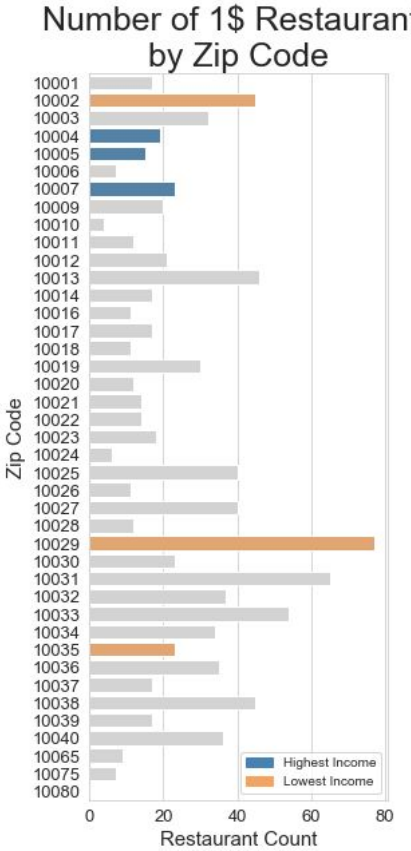
Exploratory Data Analysis

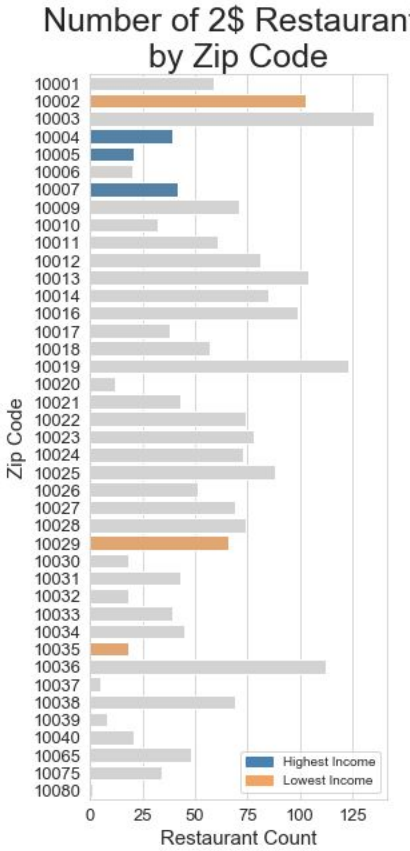
Visualizing Yelp and Census Data

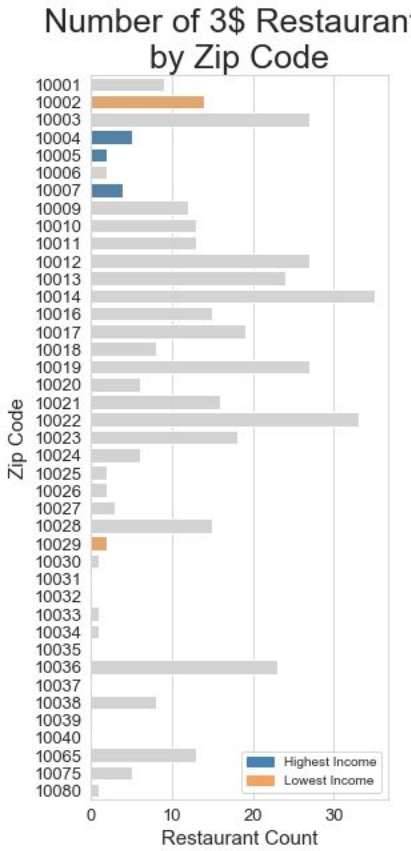
Analysis of trends and relationships between median income and restaurants' cost estimates provided by Yelp:

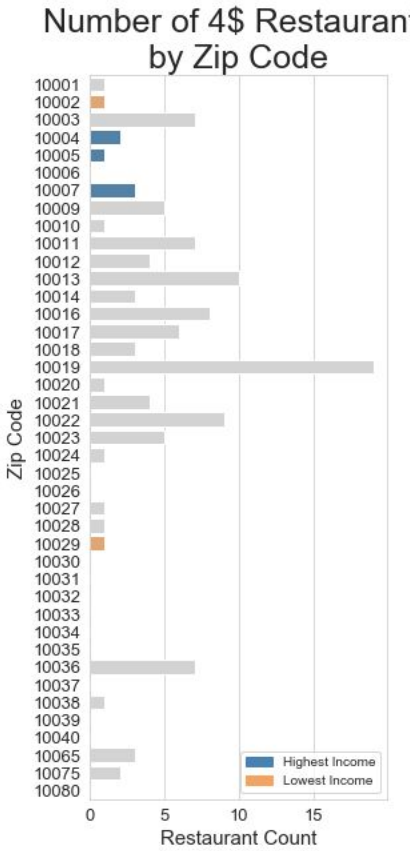
1. Three highest and lowest incomes against each cost estimate category
2. Relationship between cost estimates and incomes

Yelp Cost Estimates: Highest Versus Lowest Incomes

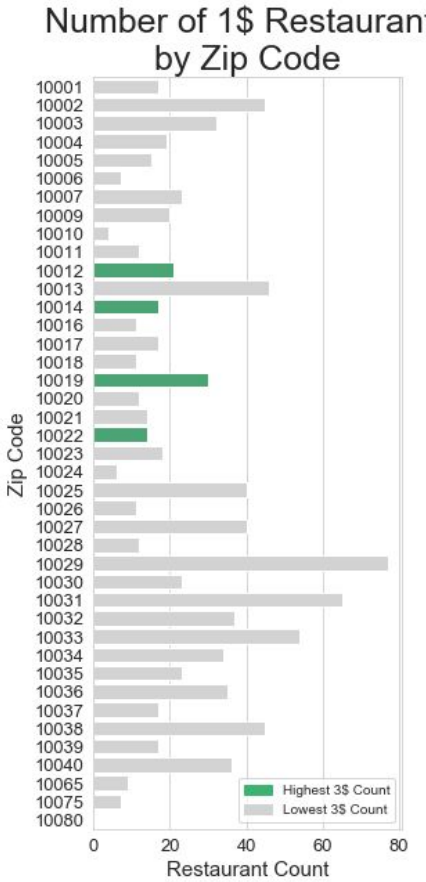


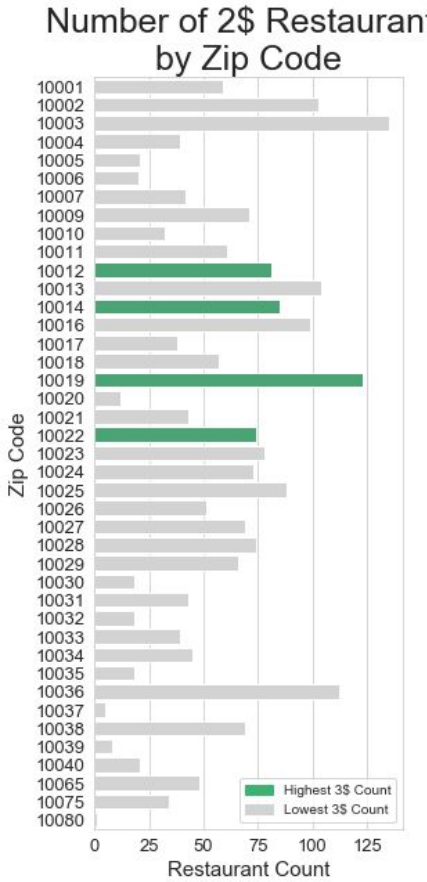


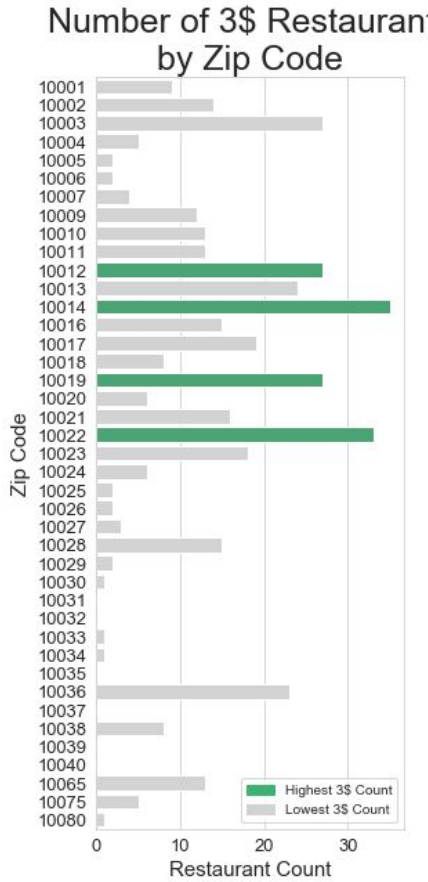


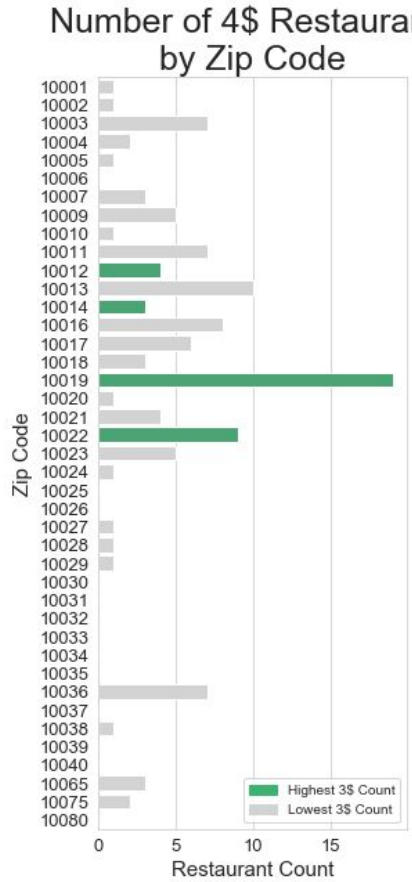


Yelp Cost Estimates: Standout Zip Codes





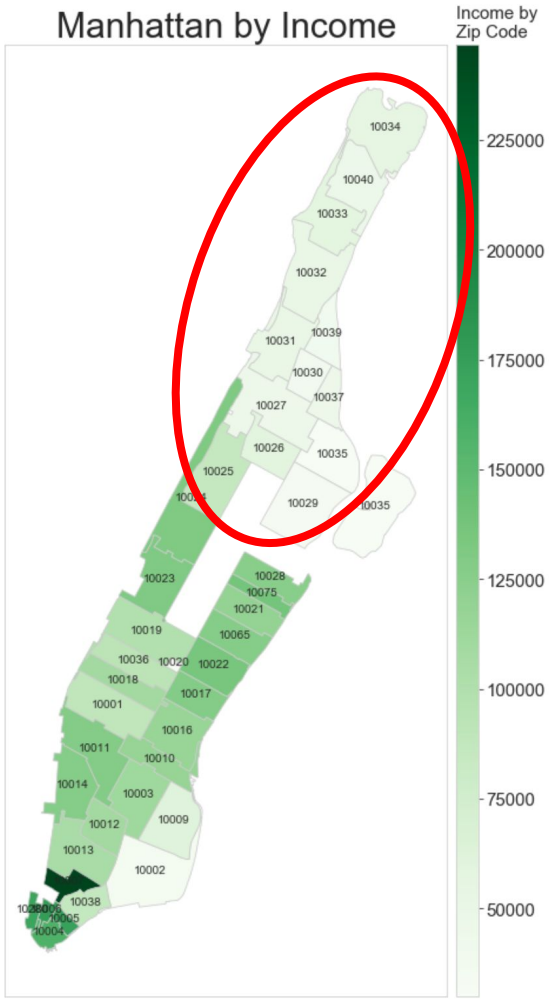


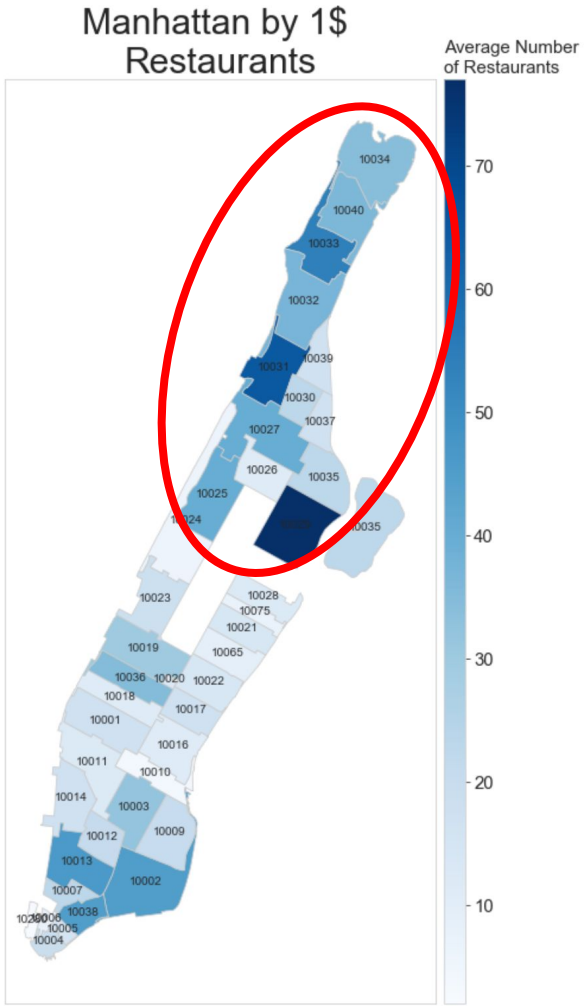


Key Takeaways

Summarizing our relationships between Yelp cost estimates and zip codes

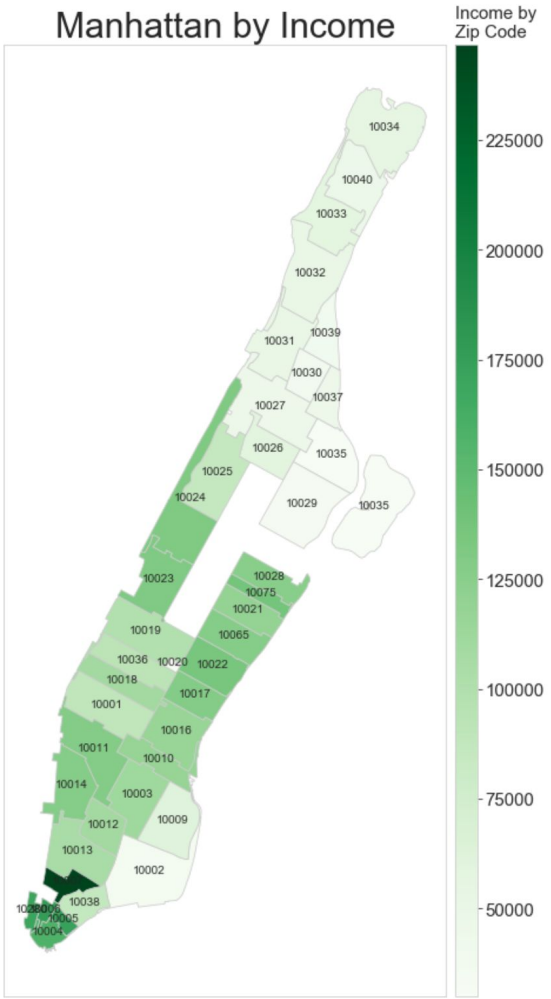
- The trend between restaurant costs follow a decreasing pattern
  - $ being the most popular
  - $$$$ being the least popular
- Lower income areas show a higher count in this trend while higher income areas show a less drastic trend
- Some zip codes don’t follow the trend at all
  - Almost displays a reverse trend

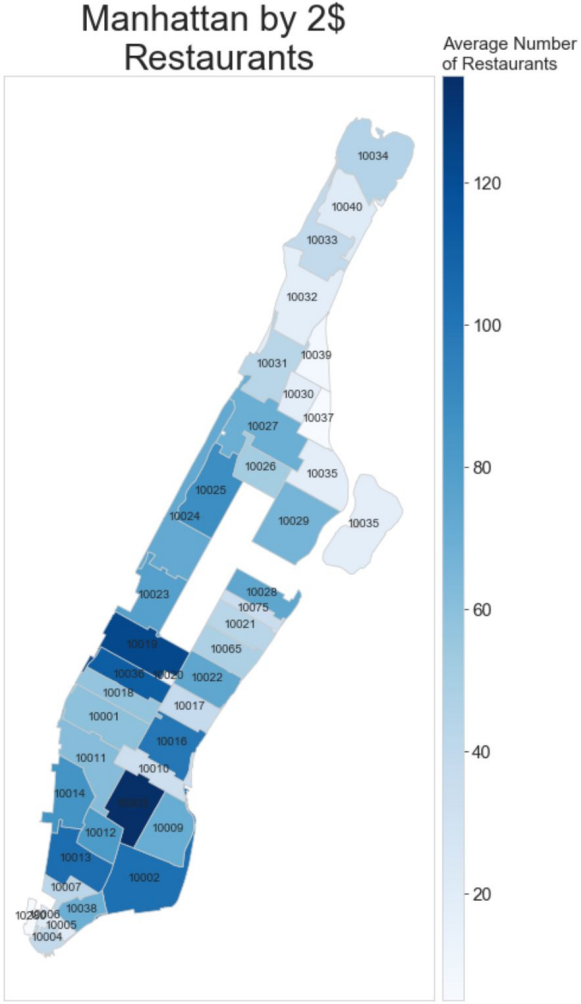




Income vs Yelp $ Cost

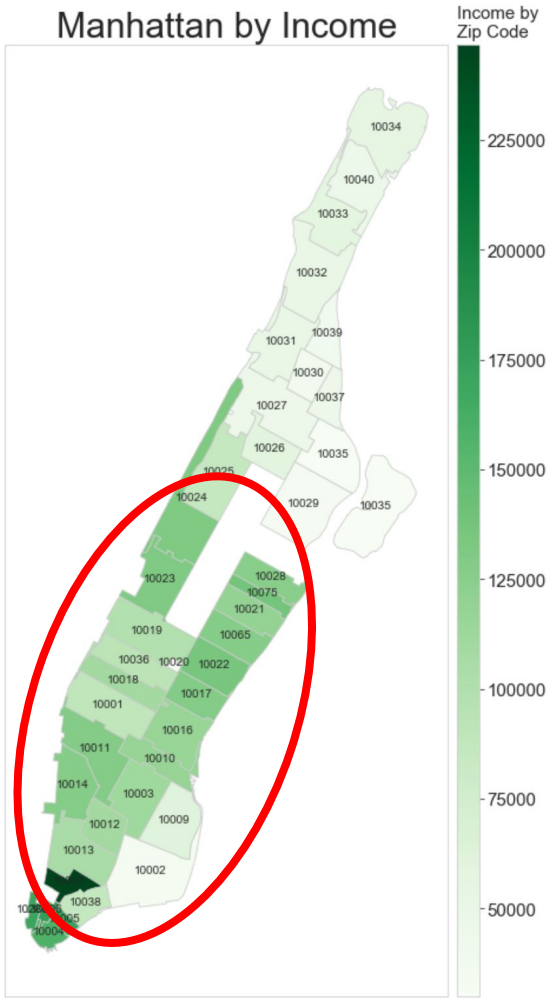
Lower income areas have a larger number of restaurants with a $ cost.

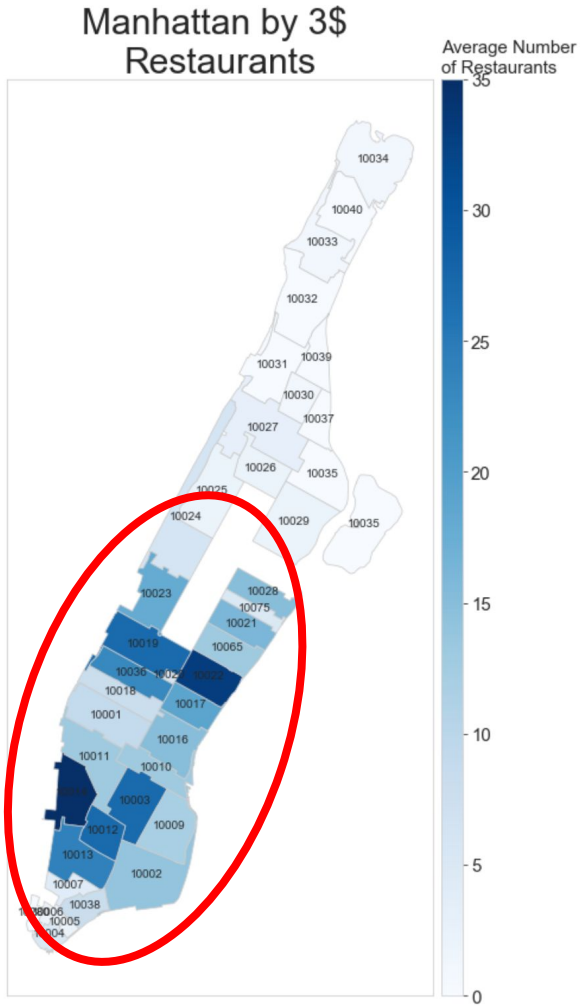




Income vs Yelp $$ Cost

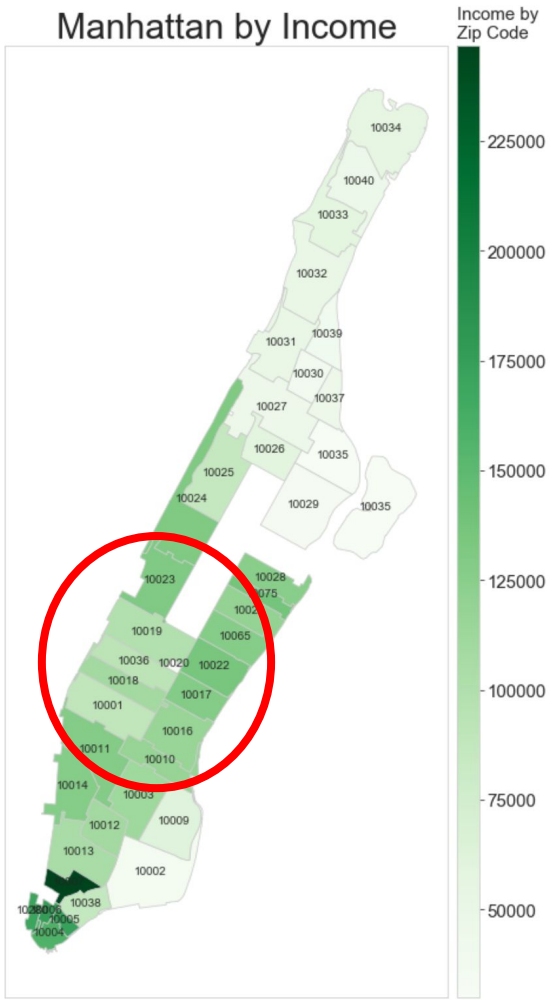
Higher cost restaurants begins to be more apparent in higher income areas.

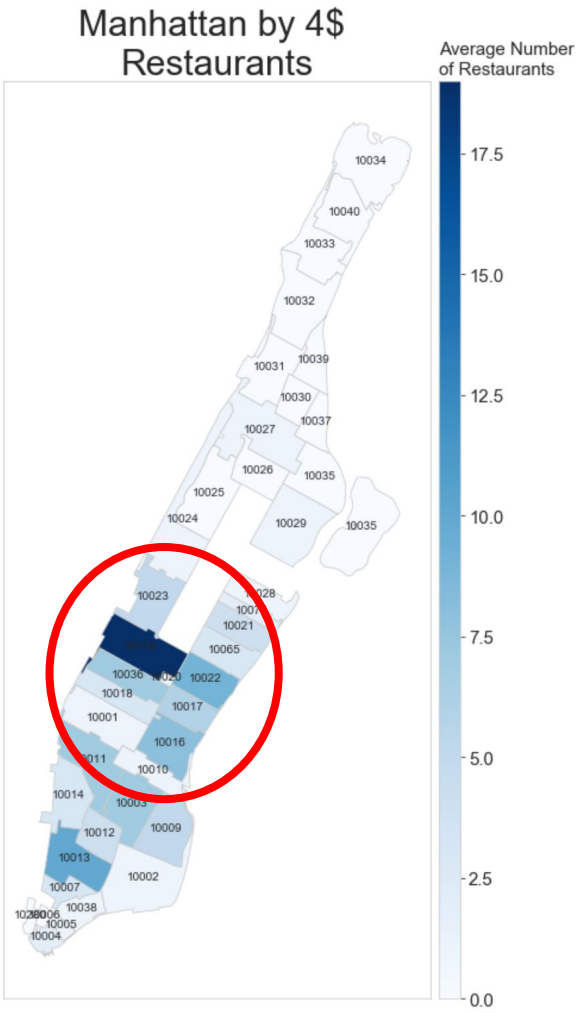




Income vs Yelp $$$ Cost

The spread of $$$ restaurants are concentrated in higher income areas.





Income vs Yelp $$$$ Cost

Lower income areas have practically no $$$$ restaurants.

Features Included

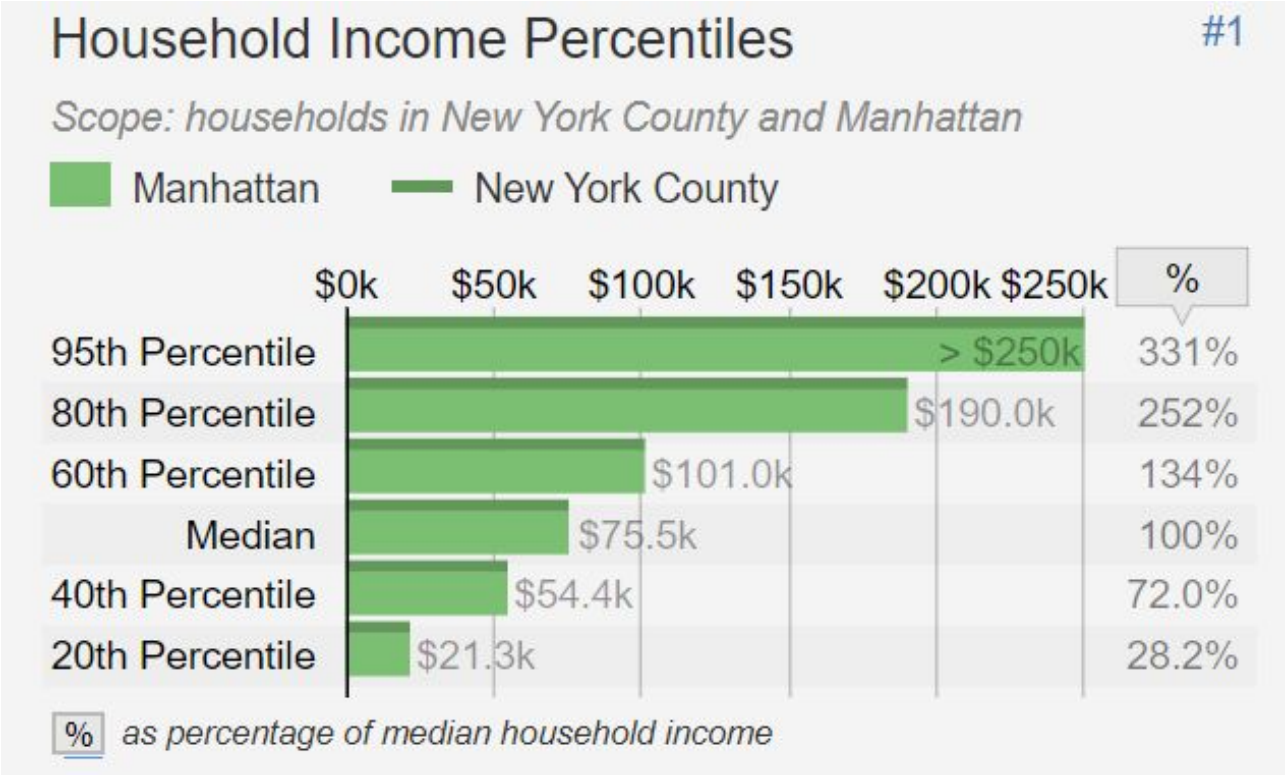
- Categorized each Yelp $ price group, gathered counts for each by zip code
- Created average price per zip code column
- Only used restaurant data

Tableau Link

https://public.tableau.com/profile/alex.lau1352#!/vizhome/MappingZipCodeAffluenceusingYelpPriceEstimates/Dashboard

Feature Engineering

Finding the average Yelp cost estimate



Affluence Categories

20 = 1 40 = 2 Median = 3 60 = 4 80 = 5 95 = 6

https://statisticalatlas.com/county-subdivision/New-York/New-York-County/Manhattan/Household-Income

Affluence Classification Values

Affluence is measured as the state of wealth, but we chose the next best data set that was easily accessible:

Income

Modeling

Determining affluence through classification

- Features
  - Yelp Cost Estimates
  - Average Yelp Cost Estimate
- Models
  - Decision Tree
  - Random Forest
  - AdaBoost
  - Gradient Boost
  - Voting
- Baseline Score: 50%

Voting Model Construction

- Decision Tree
  - Max Depth = 5
- Random Forest
  - N_Estimators = 125
- AdaBoost
  - N_Estimators = 50
- Gradient Boost
  - N_Estimators = 100

Best Model
Voting Classifier

Training Score: 100%
Testing Score: 71.43%

Standout Zip Codes

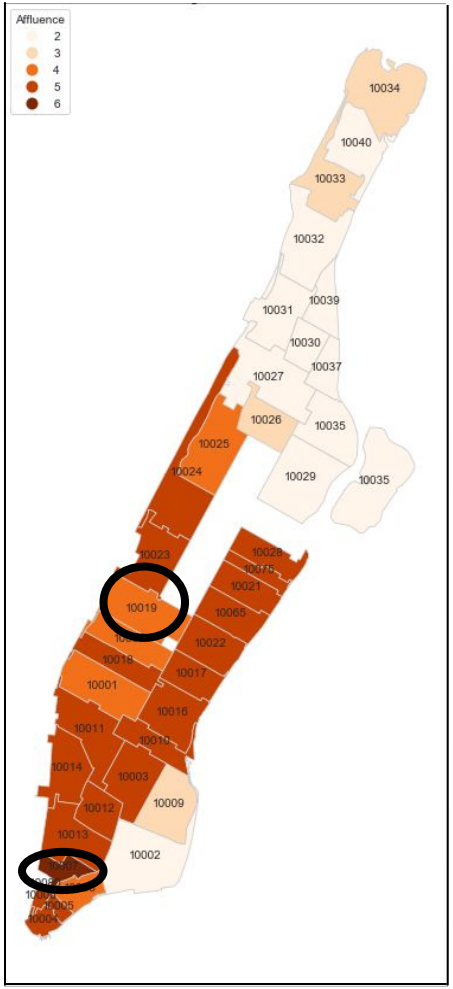
10007	Predicted: 2	Actual: 6
10019	Predicted: 5	Actual: 4

Standout Zip Codes

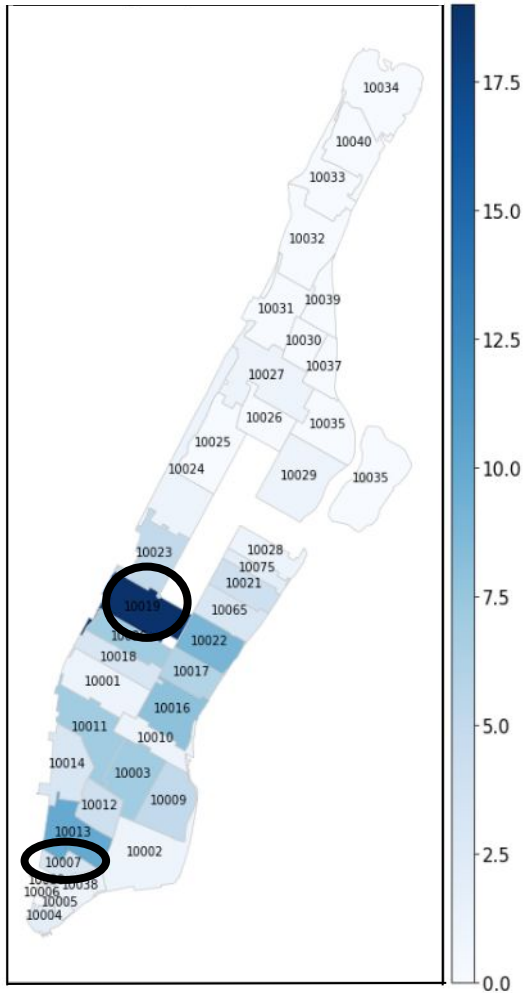
10007
- Highest Income
  - ~ $246,000
- Few Restaurants
  - 72 Restaurants Total

10019
- Cusper Income
  - ~ $99,000
- Most High End Restaurants
  - 46 ($$$) and ($$$$) Restaurants
  - 19 ($$$$) Restaurants

Manhattan by Affluence



Manhattan by 4$ Restaurants



Conclusions

Yelp cost estimates can determine a neighborhood's affluence. It isn't perfect as there may be other underlying factors that can affect affluence more than the Yelp cost estimates of restaurants.

Since our model quite accurately predicts affluence of neighborhoods based on public and easily accessible data, we can allocate resources based on affluency during emergencies; primarily focusing on less affluent neighborhoods.

Limitations

- Manhattan Dataset
  - Model may not perform well with less density areas or cities with few restaurants
- Yelp Restaurant Cost Estimates
  - Possibly missing better affluence indicators from other features
- Affluence Indicator
  - Affluence is the state of wealth - not simply median income

Further Exploration

Utilizing More Features

*Original Model Scores

Added Features

Top Categories	Restaurant Count
Italian	281
American	267
Coffee Shop	261
Pizza	256
Chinese	236

Score Comparison

Training: 100% | 100%*

Testing: 57.14% | 71.43%*

Recommendations

- Going beyond Yelp cost estimates data
- Consider using real estate data to measure affluence

Further Considerations

- Comparing commercial areas' revenues
- Relationship between restaurant count and population density

Questions?