By: Brian Jankowitz

Yelp Data to Predict Affluency

Brian Jankowitz

Data Analyst

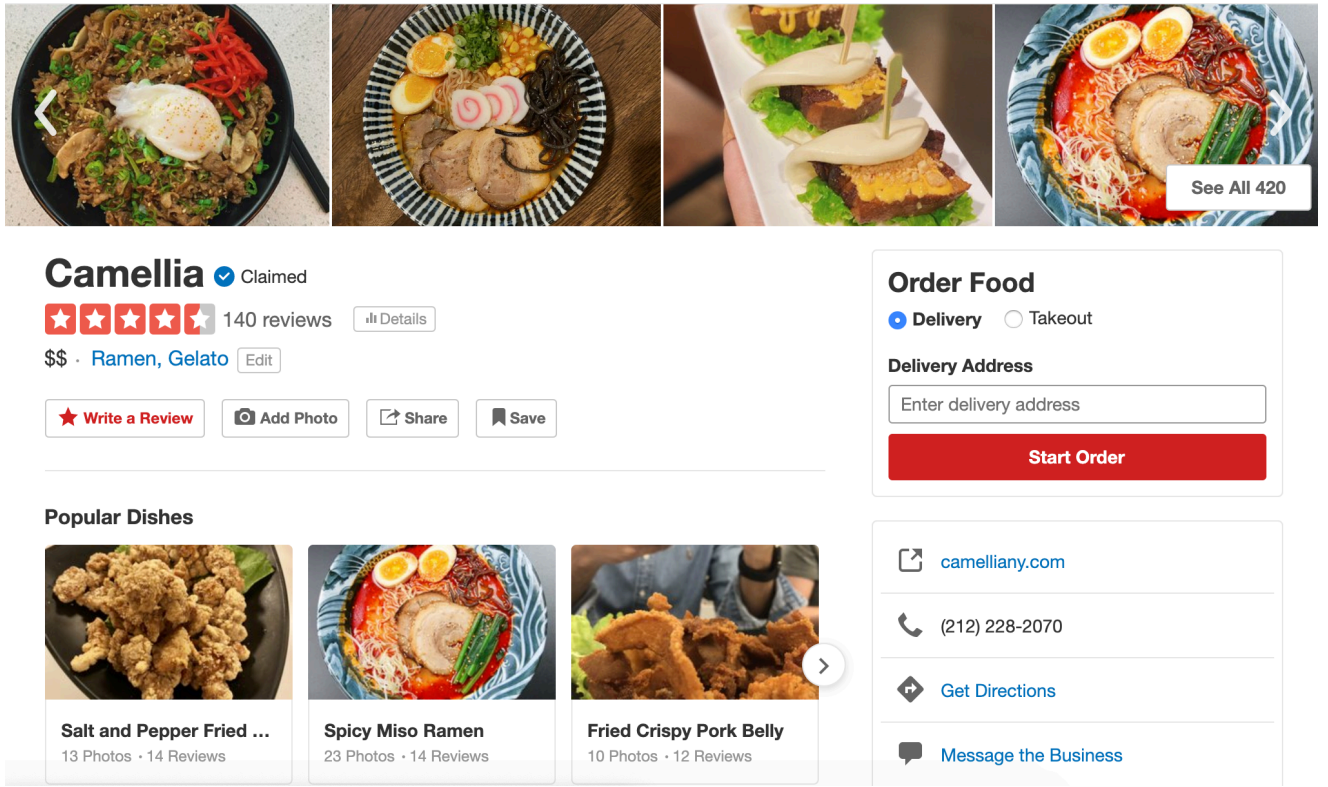
- Presented in front of audiences 300+ people
- Background data analyzation

Agenda

1. What is Yelp?
2. Problem Statement
3. Data Gathering
4. Analyzing Data
5. Modeling
6. Model Selection
7. Conclusion
8. Recommendations

What is Yelp?

- A place for business reviews
- Crowd-sourced



Problem Statement

- How can we target people in non affluent area to help provide them health insurance
  - Find non-affluent area in a non-traditional way
  - Use of big data related to commercial activity and cost of product and services as an indicator for affluency.
  - Increase healthcare coverage
  - Market to those people
- Facts (US Census Bureau)
  - In 2018, 28 million people without health insurance
  - 72.2% of people 100% or more below that poverty line are uninsured

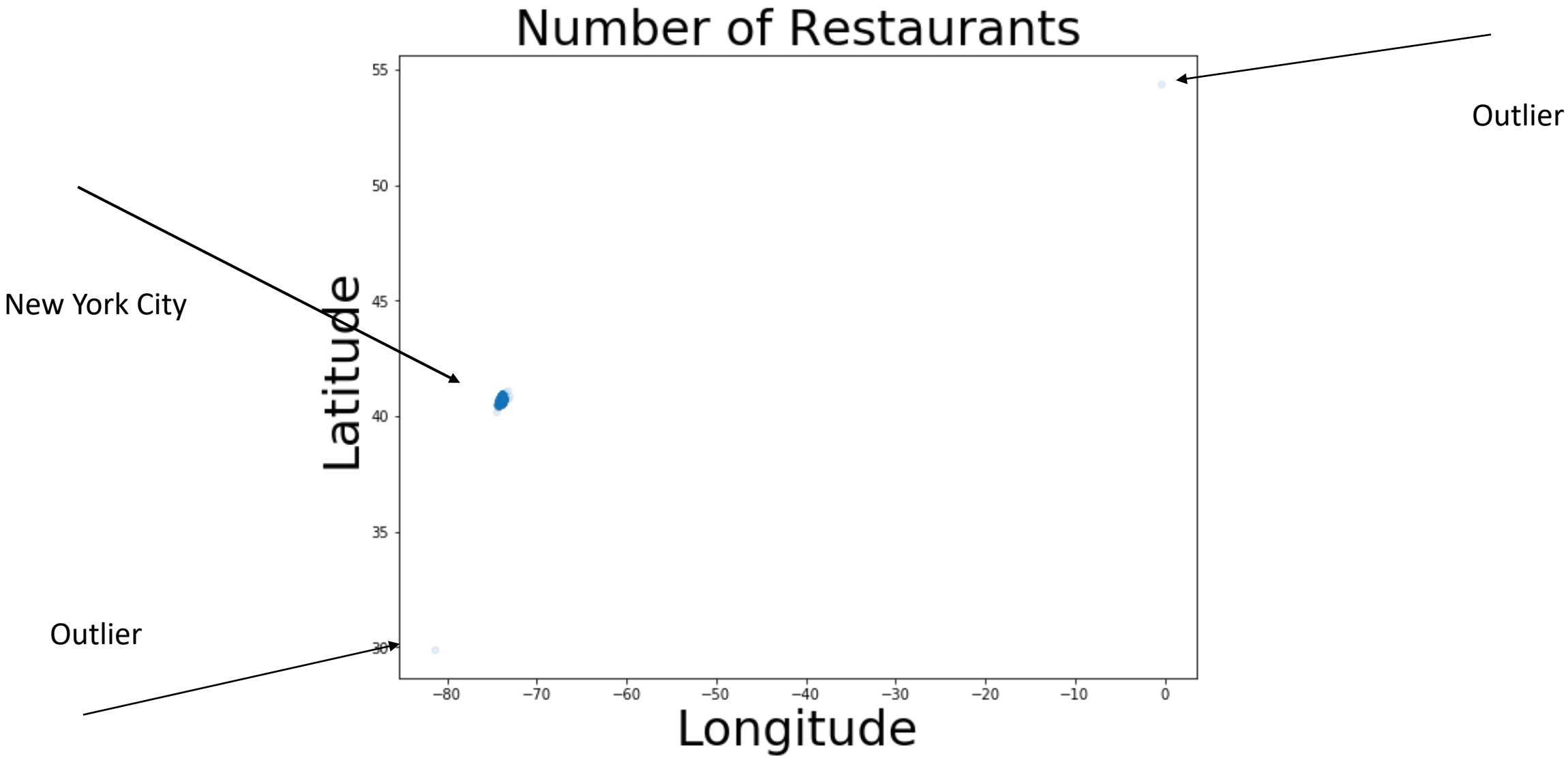
Data Gathering- Yelp

- Scrapped Yelp
- Gathered information on businesses in all 5 boroughs of New York City
- Was this information enough by itself? No
- Get more data
  - Round 1: 5000 places to only 1200 places
  - Round 2: 32000 places to only 12000 places

Data Gathering- Income

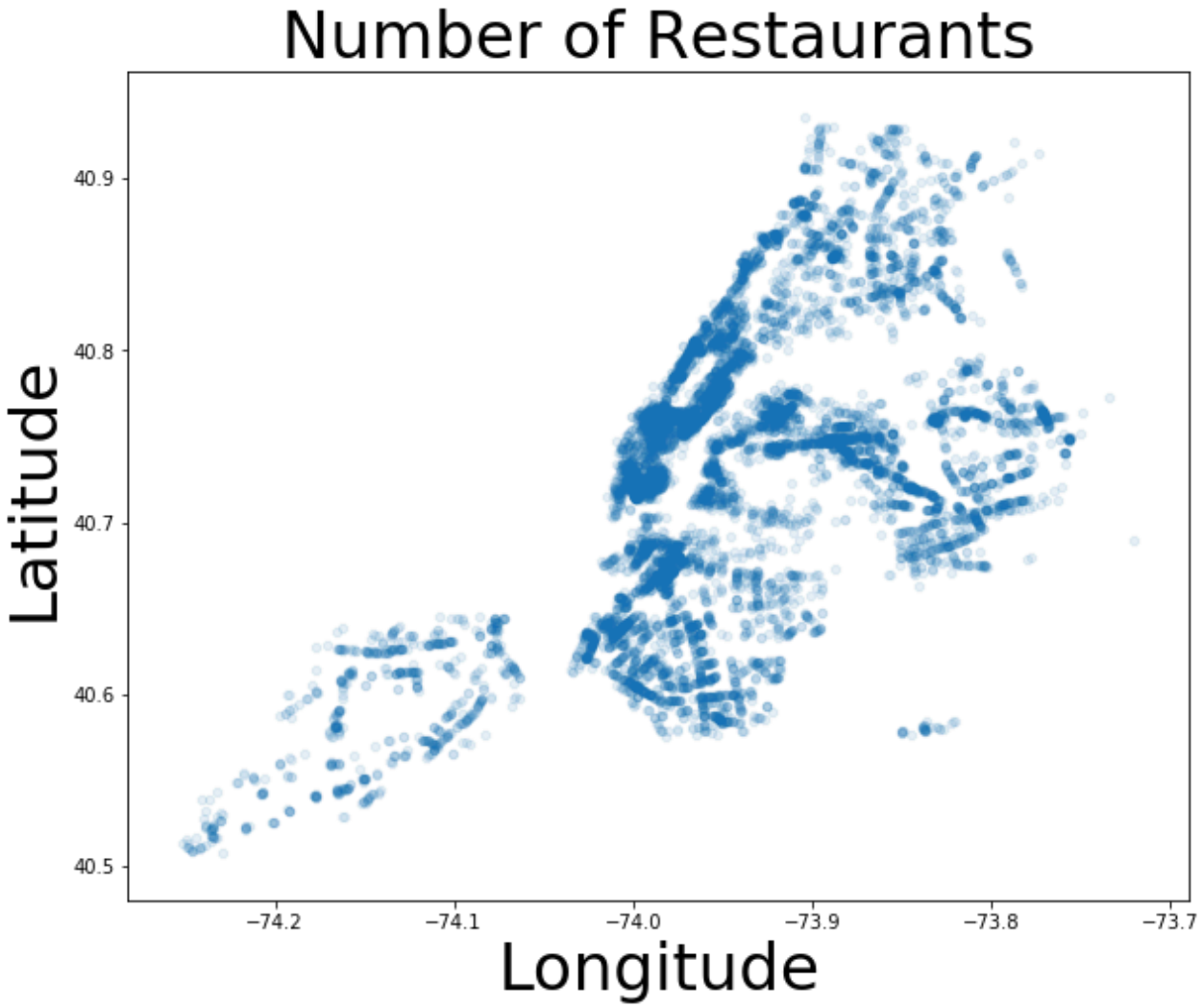
- IRS data
- Last year available 2017
- Number of returns and amount for AGI
- Why AGI and not just income?
  - To adjust for other areas
  - Each person pays different taxes
- Affluenct > 25=

Problems with Data- Outliers



Coverage of New York City

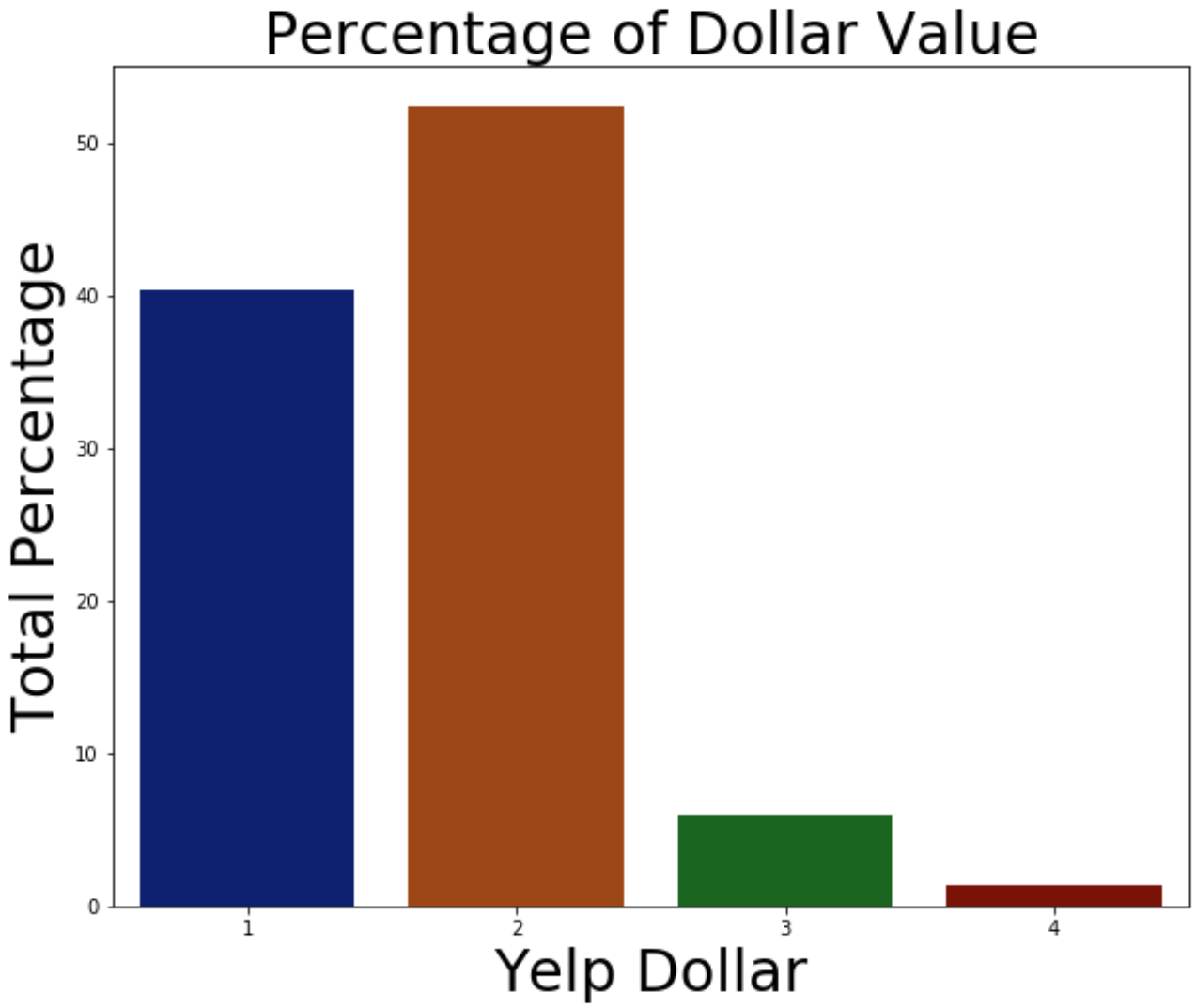
- Dark vs light colors
- Business density

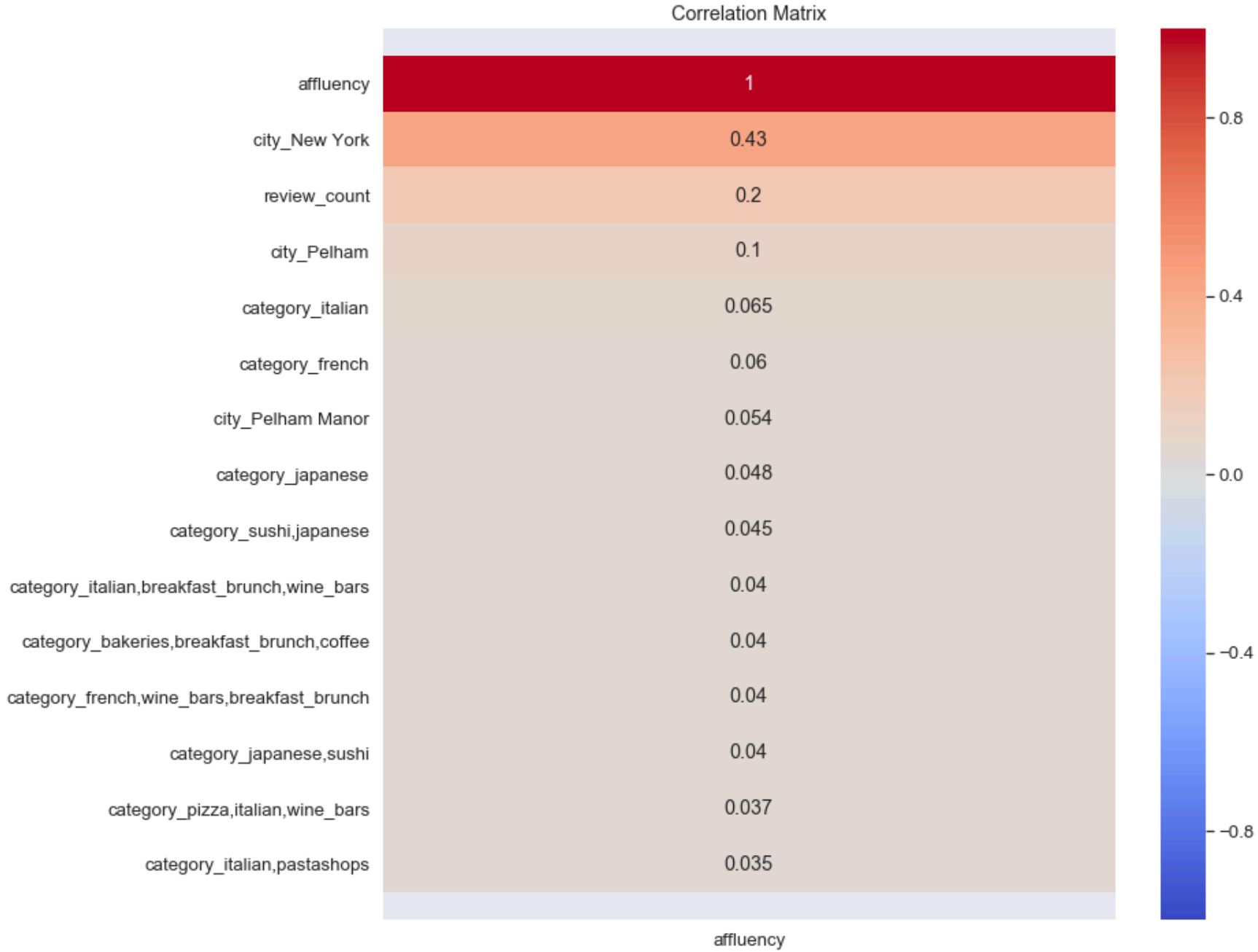


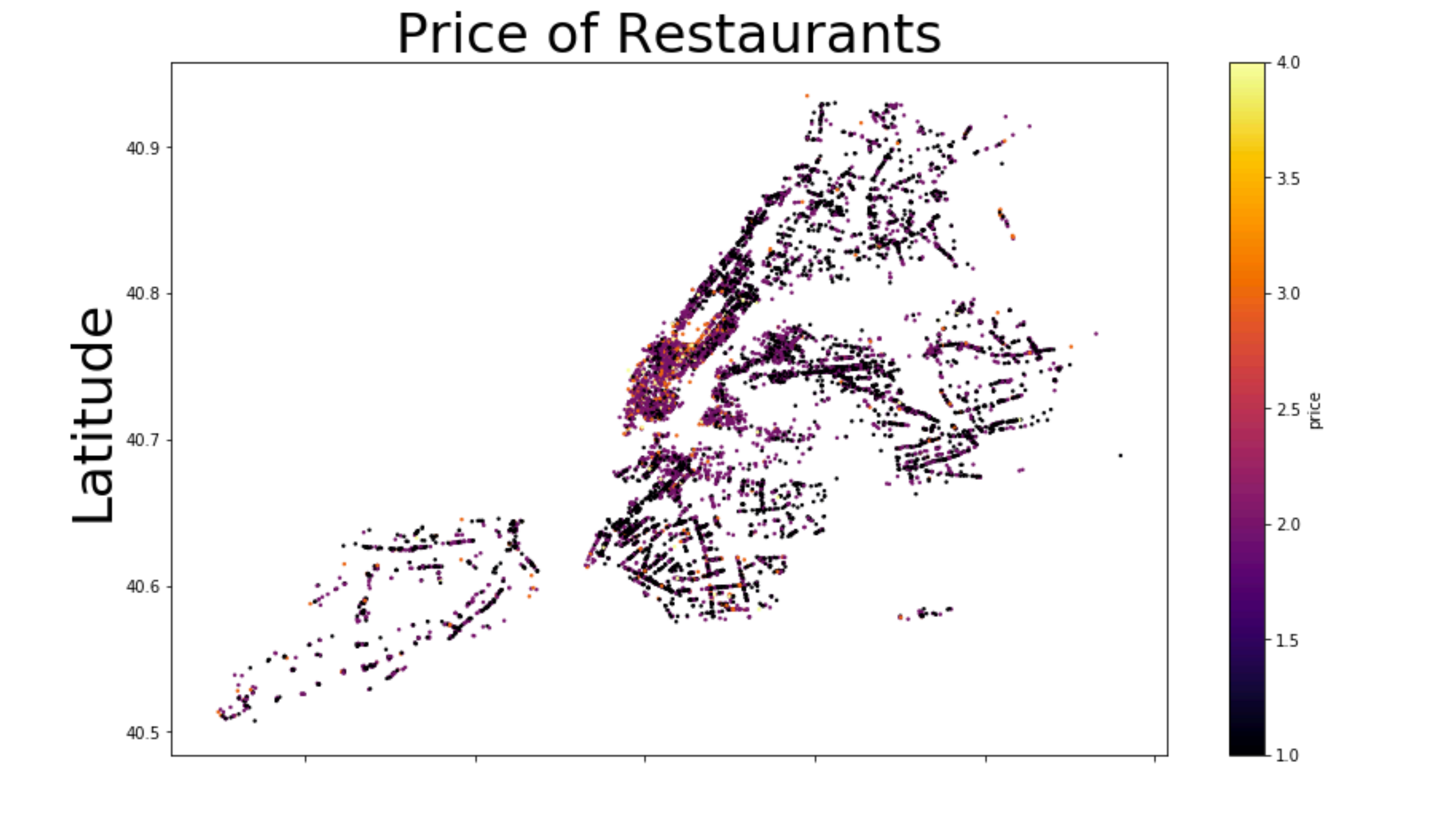
Borough	Population (2018)
The Bronx	1,432,132
Brooklyn	2,582,830
Manhattan	2,278,906
Queens	2,278,906
Staten Island	476,179

EDA

- 3($$$) and 4($$$$) less than 10%
- 1($) and 2($$) more than 90%




Correlation Matrix
affluency
1
city_New York
0.43
review_count
0.2
city_Pelham
0.1
category_italian
0.065
category_french
0.06
city_Pelham Manor
0.054
category_japanese
0.048
category_sushi,japanese
0.045
category_italian,breakfast_brunch,wine_bars
0.04
category_bakeries,breakfast_brunch,coffee
0.04
category_french,wine_bars,breakfast_brunch
0.04
category_japanese,sushi
0.04
category_pizza,italian,wine_bars
0.037
category_italian,pastashops
0.035
affluency
0.8
0.4
0.0
−0.4
−0.8

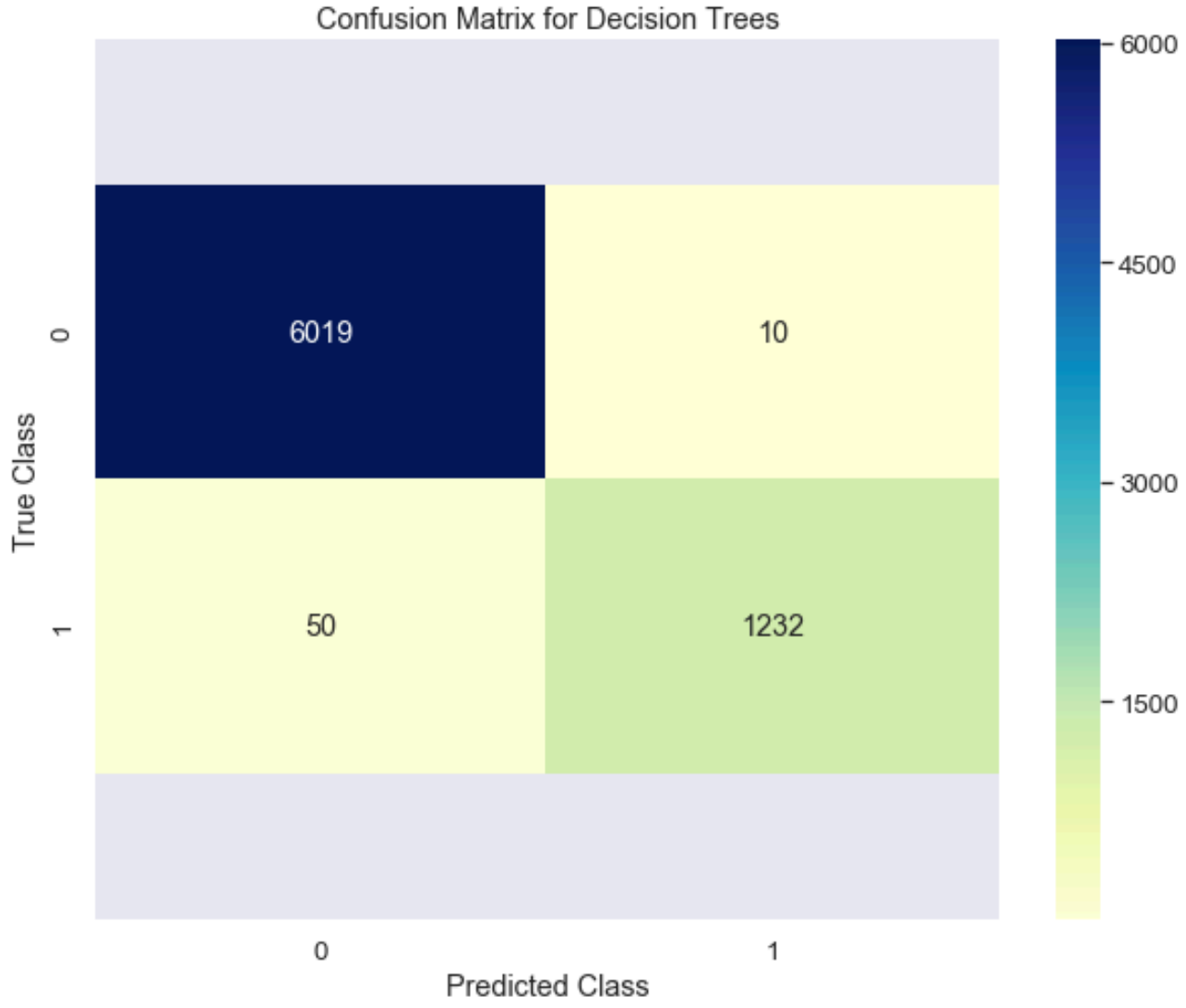

Price of Restaurants
Latitude
40.9
40.8
40.7
40.6
40.5
price
4.0
3.5
3.0
2.5
2.0
1.5
1.0

Modeling

Model	Train	Test	Sensitivity	Specificity
Logistic Regression	83.1%	83%	22%	96%
KNN	97.8%	82%	96.4%	96.1%
Decision Trees	100%	99.9%	21.9%	96.1%
Bagging	99.9%	99%	21.9%	96.1%
Random Forest	99.9%	99.9%	21.9%	96.1%

Baseline Train	Baseline Test
82.3%	82.3%

Model Selection



Conclusion

- We can predict the affluency of an area
  - High true positives
- There are some questions about the data
  - How accurate is this?
  - Tribecca
    - Richest zip code in New York City (source: Business Insider)
    - Don Bella Pizza rated $
  - More research needs to be done

Recommendations

- More time should be allocated to this
- Look at more expensive areas
  - Times Square and World Trade Center
  - Plain slice of pizza $2, Times Square $4+
- Use Big data in other ways
  - Apartment prices
  - Express trains
  - How close to train (only certain areas)
- Use big group to collaborate, see if $$ can be used to supplement another model for predicting prices