Predicting Changes in Neighborhood Rent using Restaurant Reviews and Attributes

Islam Tawfik
MS in CS
Courant Institute, NYU

Sarthak Jain
MS in CS
Courant Institute, NYU

Chung-Ling Yao
MS in CS
Courant Institute, NYU

Wonik Jang
PhD in Marketing
NYU Stern School of Business

Diego Francisco Rincon Santana
MS in CS
Courant Institute, NYU

Tory Peterschild
MS in CS
Courant Institute, NYU

***Abstract*—This project explores the correlation between a neighborhood's rent prices and the reviews and attributes of that neighborhood's restaurants. Specifically, we examined neighborhoods in the five boroughs of New York City, using data from the census, Yelp, and StreetEasy.**

Keywords—analytics

I. Introduction

Using restaurant reviews scraped from Yelp, we performed sentiment analysis and a study of different writing styles contained in the reviews to extract information about the average rent price in New York City neighborhoods.

II. Motivation

Most metropolitan areas in the US are constantly changing as residents are affected by growth or decline in the economy, availability of jobs, the city's population, and rent prices. In certain cities — notably San Francisco and New York — housing has become a hot button issue as rent prices soar and long-time residents are priced out of their neighborhoods.

Any insight into these trends — and particularly the ability to predict where and how steeply rent will increase before it is predicted by other indicators — would be valuable to many parties. Developers and investors could make business decisions based on these predictions, and renters could use the information to decide when it would be wise to hold onto a lease, what neighborhoods might be affordable for them, and so on.

III. Related Work

(If anyone read related work we can summarize here — if not we can just take this section out)

IV. Design

Our initial idea was to try to predict rent in a neighborhood based on 1) the kinds of restaurants opening in that neighborhood and 2) the rate of restaurant openings in the neighborhood. Due to certain obstacles with data availability (Fig. 1 and 2) we reframed our project to focus on restaurant review data.

First we created sentiment analysis model by using CoreNLP to clean the review texts (remove punctuation, digits, and stop words such as "place," "food," etc.) for all data. Then we used "pos-tagging" to extract nouns and adjectives and generate a word cloud (Fig. 3). Then we calculated the ratio of positive and negative words to total words for each restaurant in each year, using Hu and Liu, KDD-2004's list of positive and negative words as a dictionary [1].

The second model used word2vec trained on 20% of reviews to create 20 different “writing styles.” We then ran k-means on this vectored output with k=20 using Spark ML, created an array of size 20 and grouped the reviews using the array index as the key for the review style. For a given year, the model looks one year back to see what writing styles that year contained compared the the current year. Using the vector found, the model predicts the rent increase or decrease for one year in the future. There are six categories: 1) rent rises by between $0-100; 2) rent rises by between $100-200; 3) rent rises by more than $200; 4) no change; 5) rent decreases between $0-100; 6) rent decreases by between $100-200; 7) rent decreases by more than $200.

V. Results

Because there was no historical listings data available, we used census data showing rent observations from 2010 to 2014. This was the closest substitute we could find, but we did see some unexpected results because census data represents different information from apartment listings. Apartment listings reflect the market price for rent in a given neighborhood, apartment size, etc., while census data includes observations from rent-controlled apartments, leases that have been held for a long period of time, and public housing, lowering the average rent calculated for the neighborhood.

VI. Future Work

(Future… Given time, how would you expand your analytic? Could it be applied to other areas? Etc…)

VII. Conclusion

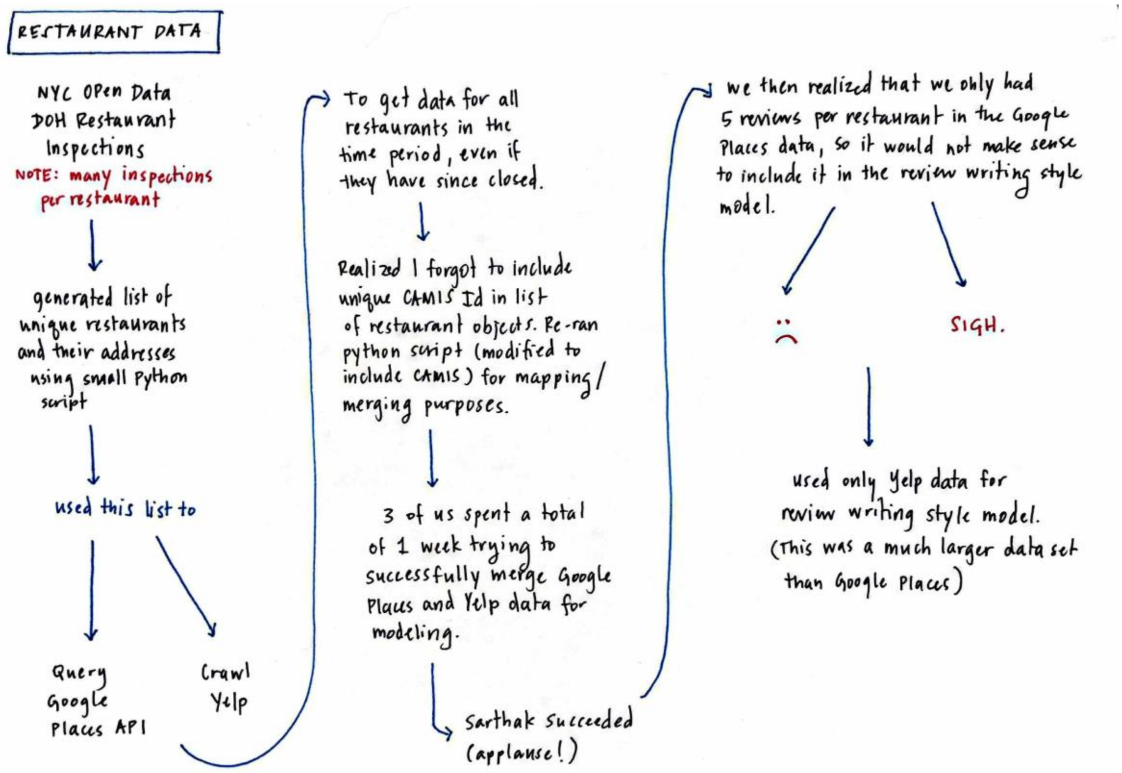
(Future… One or two paragraphs about the value/accuracy/goodness of your analytic.)

Acknowledgment

The authors of this paper would like to thank the following people and institutions for either inspiring or assisting in the development of this project and paper: Professor Anasse Bari, Yelp, Google Places, StreetEasy, NYC Open Data, US Census Data.

References

1. Hu, M., & Liu, B. (n.d.). *Mining and Summarizing Customer Reviews* [Scholarly project]. In University of Illinois at Chicago. Retrieved from https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf
2.
3.



RESTAURANT DATA

NYC OPen Data DOH Restaurant Inspections

NOTE: many inspections per restaurant

generated list of unique restaurants and their addresses using small Python script

used this list to

Query Google Places API

Crawl Yelp

To get data for all restaurants in the time period, even if they have since closed.

Realized I forgot to include unique CAMIS Id in list of restaurant objects. Re-ran python script (modified to include CAMIS) for mapping/merging purposes.

3 of us spent a total of 1 week trying to successfully merge Google Places and Yelp data for modeling.

Sarthak Succeeded (applause!)

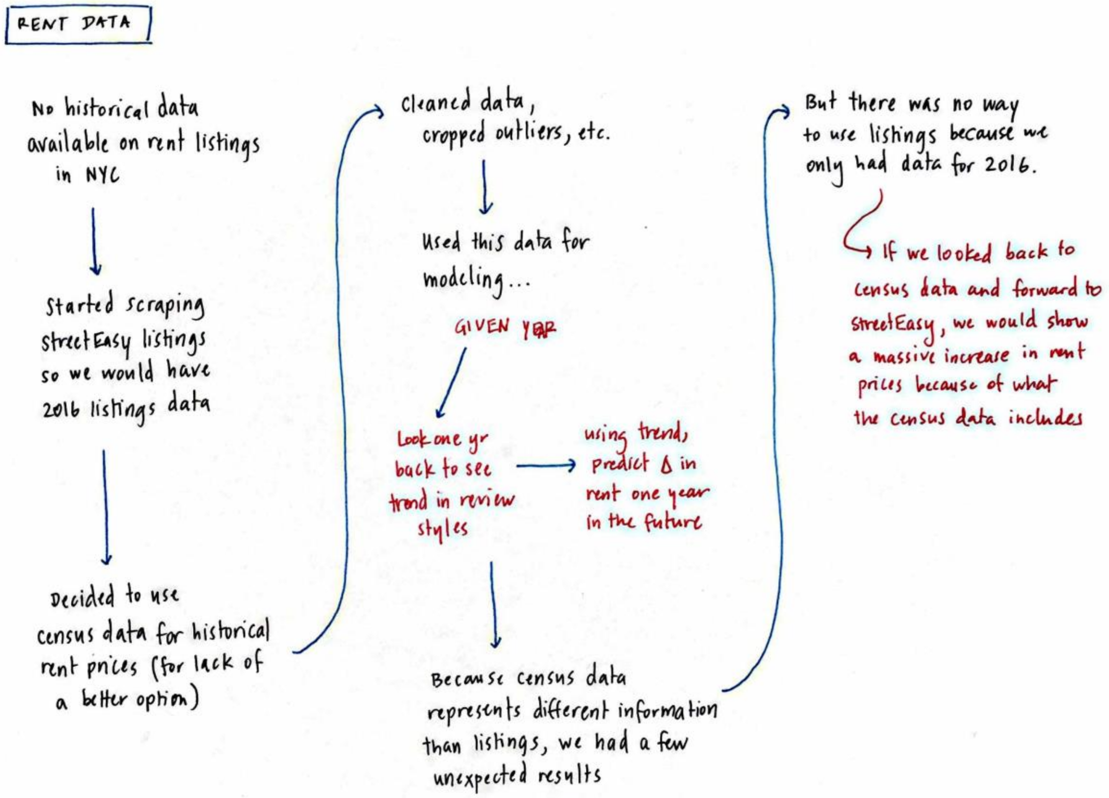
we then realized that we only had 5 reviews per restaurant in the Google Places data, so it would not make sense to include it in the review writing style model.

:(

SIGH.

used only Yelp data for review writing style model. (This was a much larger data set than Google Places)

Figure 1: Process of getting restaurant data.



RENT DATA

No historical data available on rent listings in NYC

Started scraping StreetEasy listings so we would have 2016 listings data

Decided to use census data for historical rent prices (for lack of a better option)

Cleaned data, cropped outliers, etc.

Used this data for modeling...

GIVEN YEAR

Look one yr back to see trend in review styles

using trend, predict Δ in rent one year in the future

Because census data represents different information than listings, we had a few unexpected results

But there was no way to use listings because we only had data for 2016.

If we looked back to census data and forward to StreetEasy, we would show a massive increase in rent prices because of what the census data includes

Figure 2: Process of getting rent data.

IMAGE APPENDIX

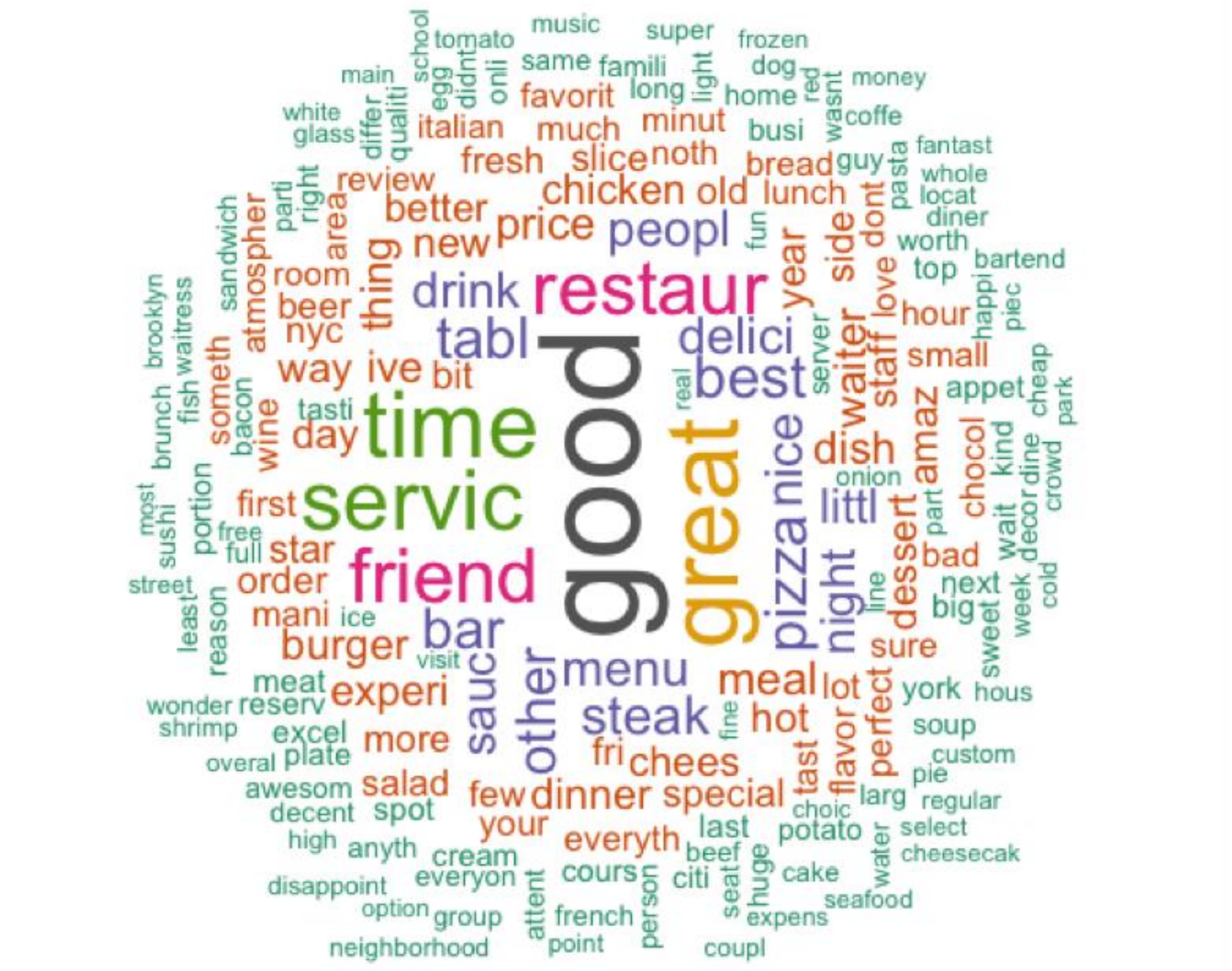
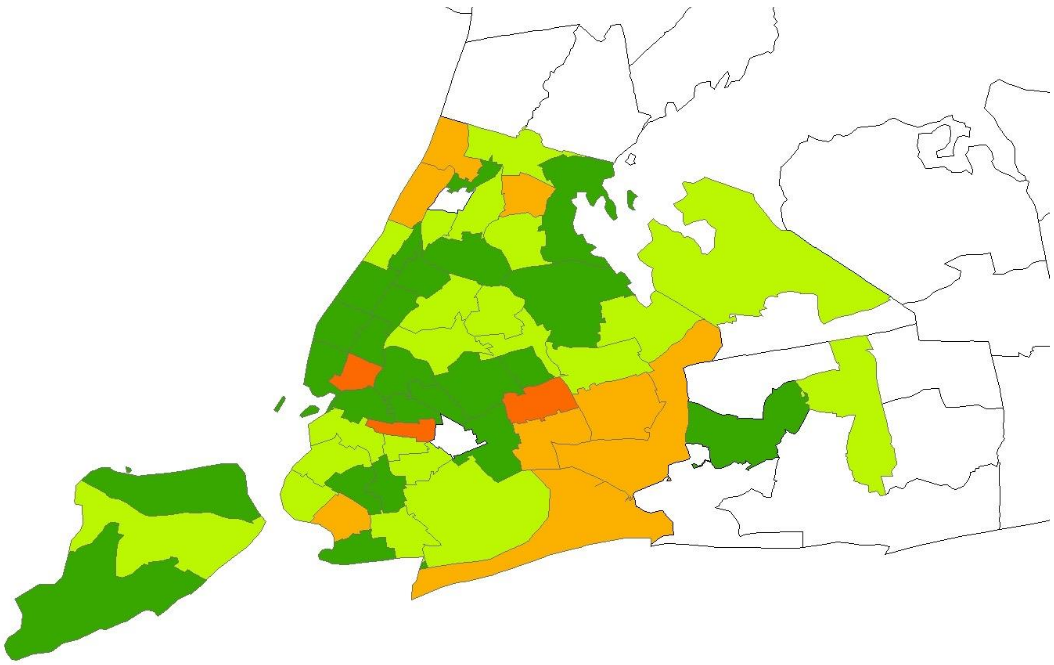
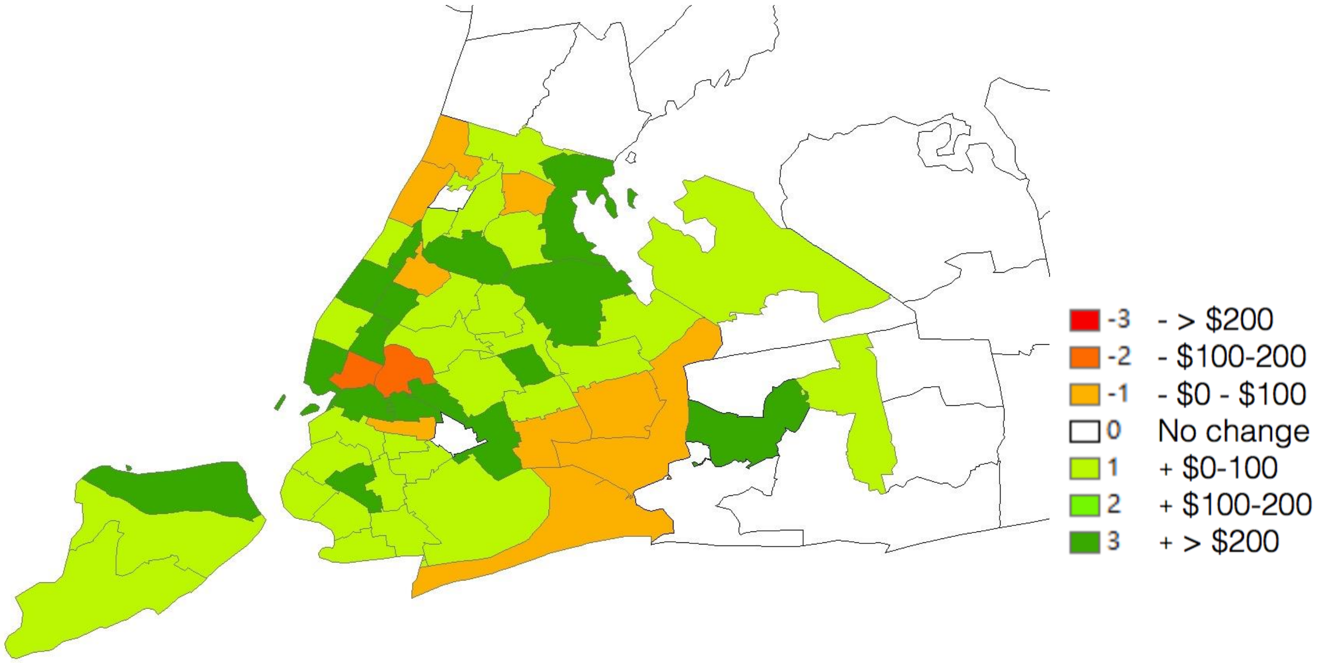


Figure 3: Word cloud generated from restaurant reviews.





🟥 -3 - > $200
-2 - $100-200
-1 - $0 - $100
□ 0 No change
1 + $0-100
2 + $100-200
🟩 3 + > $200

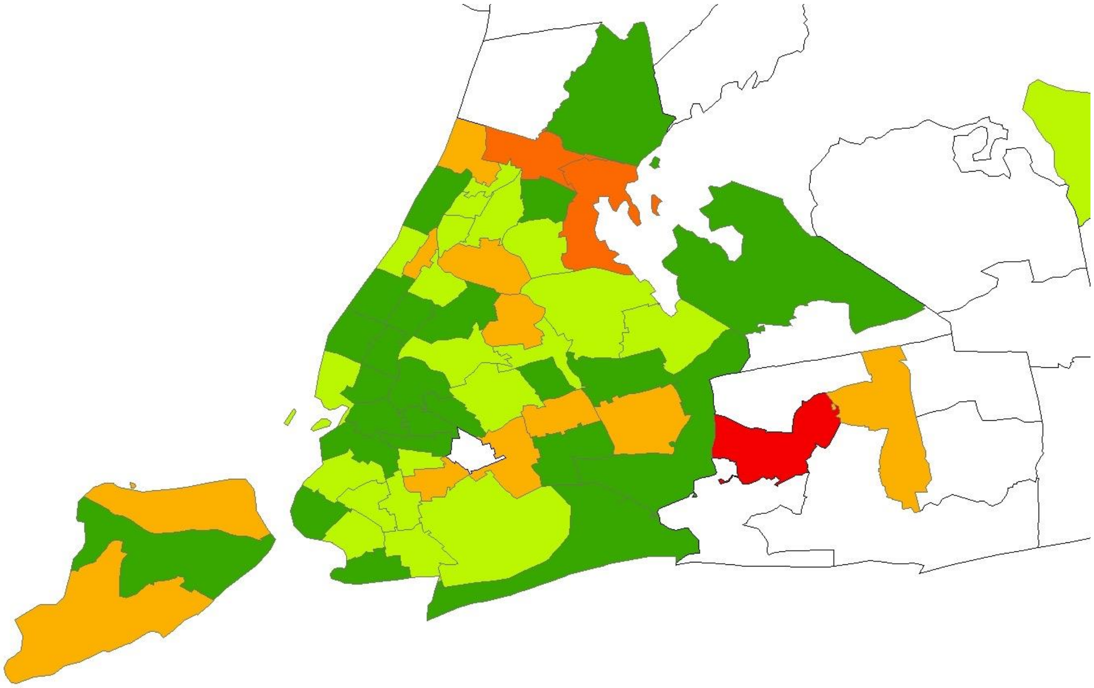


Figure 6: Predicted Labels 2013.

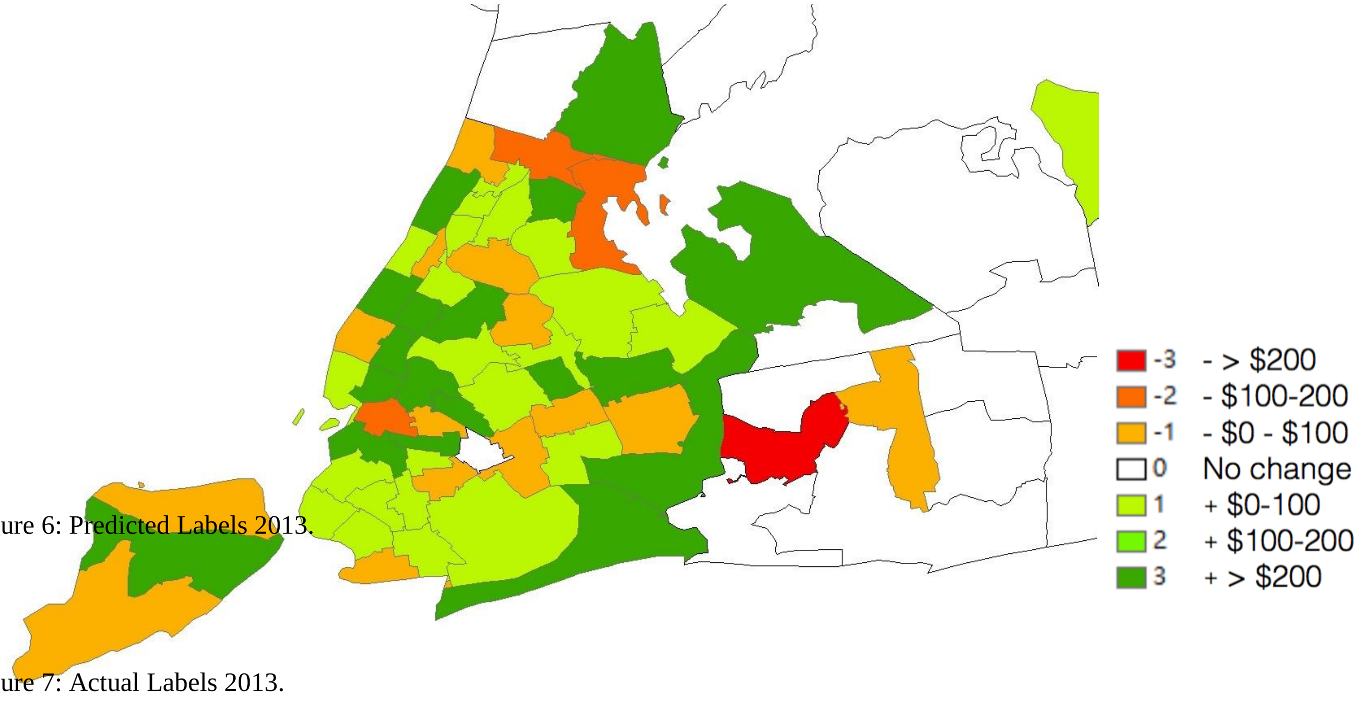


Figure 7: Actual Labels 2013.