Final Project Proposal - NYC Rental Predictions

Joel Carlson

June 14, 2016

High Level description

The Gentrificationometer!

I wish to address the issue of understanding when are area is gentrifying and some of the factors that lead up to this. It would be ideal to be able to know when rents or sale prices in your neighborhood may go up (or even down!).

My prediction target for this would be median rent in a given neighborhood in a given year. Furthermore, I would look to predict the difference between listing and sale price for a given unit in a given area. I will accomplish this by integrating a number of data sets from a number of sources, and using time series modeling techniques to find trends. This will all be accomplished for the city of NY.

Process

I will build a model that predicts the median rental price (or sale price) in an area given a number of external factors such as:

- The number, and change in number of taxi pickups and dropoffs in the neighborhood over a number of years
- Changes in the number of business or liquor licenses awarded in an area
- Proximity to a number of amenities

The median rental and sale prices are provided in two forms:

- Zillow open research data
    - Sold prices going back to the mid 90's
    - Listing prices to 2010
- NYC valuation and tax assessment data

Due to the number of sources involved, I expect that significant time will be spent cleaning and integrating the data:

Data Pipeline

Zillow Data

Zillow provides median selling, listing, and rental prices for apartments, condos, and homes for a number of cities. This data is relatively clean, but there are large sections of missing data. I will interpolate missing values where possible by modeling the trends in the data.

The data is interesting, and there appear to be interesting trends in the median selling and listing prices over the past several years:

Difference in Median Listing and Sale Price for All NY Areas
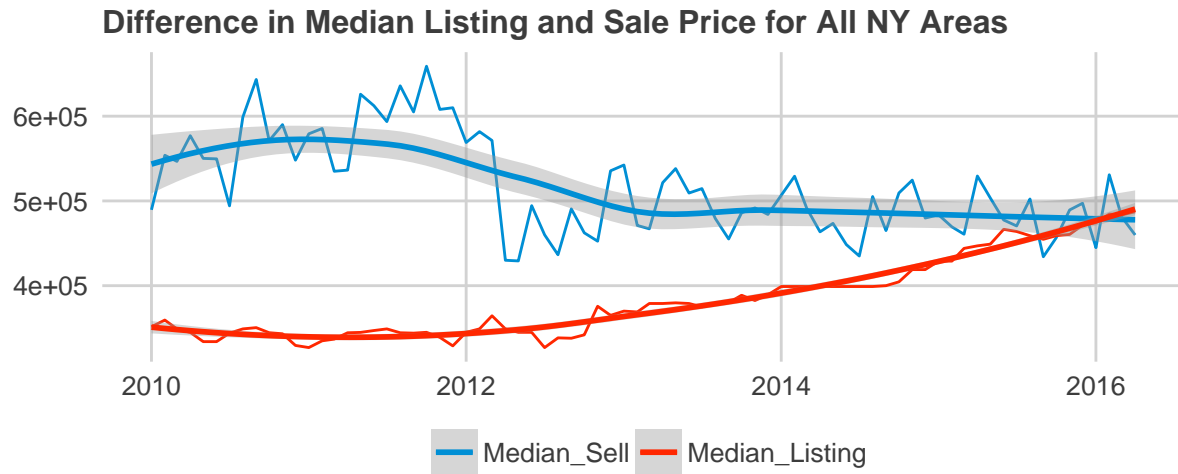
I have seen already in the data that certain regions have different rates of change in their median rent prices than other regions. This will be a good indicator of gentrification!

NYC Valuation Data

The NYC valuation data is stored in several large MongoDB databases. I have not yet explored it

- Source 1
- Source 2

Taxi Data

I have acquired ~100 GBs of taxi pickup and dropoff location for the city of NY. It will be an extreme challenge to process this data, but I believe that once it is aggregated, it will be of reasonable size.

Liquor Licenses

I am in the process of scraping this data. I have already brought down the server once by mistake, but my script has been chugging along nicely for the past several hours. If that doesn't work out, then this will not be used as a source of data.

At this moment I have enough data to approximate a map of NYC with the coordinates, at least:

Amenities Data

I have downloaded a series of geolocations in XML format of the locations of desirable amenities in new york (public parks, swimming pools, rec-centers, tennis courts, etc). I would like to create a function that calculates the minimum distance to each amenity for each location (on a discretized map of NY). These distances would then be used as features in the model.

Problems

As stated above, cleaning and combining the data will be a challenge. From what I have seen the data is reasonable enough, and I am confident that I will be able to combine it eventually.

The taxi data is extremely large, finding a way to process the data will be difficult - some sort of streaming solution will likely be the best way to tackle it.

Deliverable

I will create a working web app which would present a map of a city and allow the user to adjust the year that is being displayed. There would be an overlay of predicted rent prices and the rate of change (indicating increased gentrification) by area that would change with the year, and in this way the "flow" of money and gentrification in a city may be visualized.

Furthermore, I would like there to be an input box where a user can enter an area, and a listing price of a home, and the output would be the predicted difference between listing and selling price. This would help prospective buyers price out their offers.

Data

Links to all data are here:

https://github.com/joelcarlson/CityPredictions