Gregory Lull
October 2nd 2019
Sea19_ds10

Project 2 Proposal
Predicting NYC Rent and Sale Prices

Domain

I would like to predict rent and/or sales price in New York City, which will include all five boroughs. My personal interest in this is for a couple of reasons:

1. I lived in NY for 4 years and have moved three times, each time my criteria changed and each year the pricing for different neighborhoods *feel* different, and each time it was difficult to pinpoint which neighborhood will fit my budget
2. there is a lot of data hosted on StreetEasy.com, this website has pagination and easy HTML identifiers for the different variables I want to examine
3. There are also historical data (previous rent prices) listed on the same page for the different units, I think I can use this data to see if my "predictions" work well, by using historical data to predict current listed prices.
4. If this type of information is generalizable for NYC, I think future work could include other cities.

Ultimately the question I want to ask is if NYC real estate is a good investment-property (assuming you have the down payment). If rent stays consistent or increases year-over-year, it may make sense to purchase a place and either live there or rent it out.

Data

Variables	Type	Description
Rental Price	int	The target I am trying to predict.
Sales Price	int	Possibly another target I am trying to predict, I will scrape this information at the same time.
Bedroom count	int	This can be a count from 0+
Bathroom count	int	This can be a count in increments of 0.5
Amenities	Dummy variables	There is a listing of common amenities, I think these might be

		standardized, or I can use very simple text strings to search, for example: ‘laundry in building’, ‘hardwood floors’
Building stories	integer	High rises may offer more amenities or be more modern
Building built date	datetime	Older buildings like ‘brownstones’ and prewar are quaint and cute, may serve different markets.
Rental history - date - same unit (new row)	datetime	If there’s rental history listed i think i can create a new row with identical data and then list each row
Rental history - price - same unit (new row)	price	If there’s rental history listed i think i can create a new row with identical data and list the previous price
address	string	So i can locate where this unit is on a map
Address - geo	String of lat, lng	If there’s no physical address there may be a lat,lng (there is a google map on the bottom)
Building facts - sometimes units are connected with a Building Page which lists other information such as school districts, available floor plans, past sales and past rentals	Dummy variables (maybe)	There are some information here that I would like, for example school districts, but if there’s no detailed “building page” i may be able to find such information using another webpage or API
Nearby subway stations count	int	Not sure what this looks like yet. Could be a number for total count
Nearby subway stations distance	float	This could be an average, or the closest subway station, not sure yet. Or mabye multiple columns such as ‘average distance’ and closest distance. This variable is important for work, and winter.
Number of pictures shown	int	I don’t know how to qualify the quality of pictures, but having more nicer pictures might relate to how quickly something is off the market. Or if it looks nicer the

		perceived value may be higher.
Parking	dummy	This may be included in amenities, or street parking, which could have different effects. Areas that are more dense will likely have less parking.

Known Unknowns

Some other variables that I can’t get from StreetEasy relate to the neighborhoods themselves. For example I think restaurant density and quality, family oriented services (day care), noise complaints, crime statistics, and proximity to “downtown” areas will all influence. I think if there is more time in the project I can scrape (or use the API) for Yelp, some government websites for noise and crime, and get a list of nearby parks etc.