NYC House Sale Price Analysis

Team Alpha
Peiying Yu, Jianxing Wan, Ziyuan Wang
04.25.2019

Outline

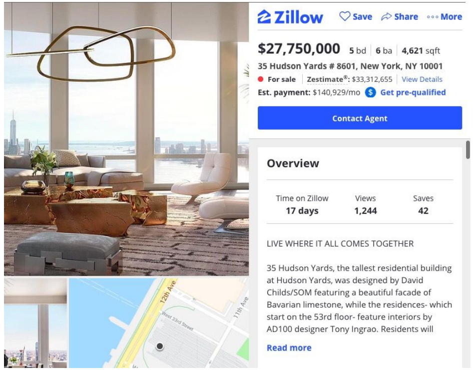
- Project Goals
- Datasets
- Exploratory Data Analysis
- Predictive Analytics
- Conclusions

Project Goals

- Identify significant features that affect house prices in NYC
- Build predictive models for home values based on features

BEST OF ZILLOW

Datasets:



- Zillow --- Web scraping
  - Build a Zillow info scraper with BeautifulSoup
  - Scraped 10,000 home for sale prices * 15 features
  - Scraped according to 12 neighborhoods and 53 zip codes in Manhattan and LIC
  - Key features: bed, bath, size, year built, zip code, latitude, longitude, overview
- Crime -- Open Data from NYPD
  - 114,673 safety complaints in total for 2018
- MTA -- Open Data from NY State official Website
  - NYC Transit Subway Entrance And Exit Data
  - Features: station name, line, longitude and latitude coordinates of entrances/exits

Data Cleaning and Processing:

- Zillow Data:
  - Drop rows with NaN in price (label)
  - Convert zip code into string and treat as dummy variables, year built into integer, etc.
  - Add price range to treat price as a categorical variable (31 categories with $0.2M increment, starting from $200,000, $1M increment if > $3M)
  - About 5,500 data points after dropping NaNs
- Crime Data:
  - Calculate total number of complaints in each zip code area

Data Cleaning and Processing:

- Subway Data:
  - Calculate the distance (in km) between each home and nearest subway station with Haversine formula
  - Convert distance between longitude and latitude into km

Distance between home and station (km)

$$d = 2r\arcsin\left(\sqrt{\operatorname{hav}(\varphi_2-\varphi_1)+\cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2-\lambda_1)}\right)$$

$$= 2r\arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)+\cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$$

where

- φ_1, φ_2: latitude of point 1 and latitude of point 2,
- λ_1, λ_2: longitude of point 1 and longitude of point 2.

Point 1: home coordinate
Point 2: station entrance coordinate
r_earth = 6371 km

Data Cleaning and Processing:

	bath	bed	latitude	logitude	price	size	year built	zipcode	distance to subway	complaints	Neighborhood	price_range
626	11	7	40.766201	-73.970397	67000000.0	13000	1910	10065	0.211896	1606.0	Upper East Side	30
628	2	2	40.765154	-73.967956	725000.0	202841	1927	10065	0.125099	1606.0	Upper East Side	3
631	4	4	40.762664	-73.964743	11950000.0	4833	1871	10065	0.108910	1606.0	Upper East Side	22
636	4	4	40.762874	-73.964476	9950000.0	4138	1871	10065	0.139596	1606.0	Upper East Side	21
639	1	1	40.763155	-73.962240	350000.0	550	1959	10065	0.325530	1606.0	Upper East Side	1
643	8	5	40.767722	-73.968977	35000000.0	9440	1940	10065	0.377638	1606.0	Upper East Side	28
647	4	4	40.764253	-73.960657	2950000.0	143590	1957	10065	0.474802	1606.0	Upper East Side	14
648	10	5	40.766178	-73.966491	19995000.0	8000	1910	10065	0.143899	1606.0	Upper East Side	25
1981	8	7	40.767501	-73.970199	59000000.0	12000	1931	10065	0.324925	1606.0	Upper East Side	29
1982	2	2	40.766137	-73.970276	1100000.0	1042	1920	10065	0.216822	1606.0	Upper East Side	5
1983	1	1	40.762880	-73.957498	499000.0	107780	1963	10065	0.545624	1606.0	Upper East Side	2
1984	2	2	40.764069	-73.964407	999000.0	158500	1963	10065	0.191409	1606.0	Upper East Side	4
1985	7	6	40.768344	-73.966579	13475000.0	8926	1910	10065	0.196246	1606.0	Upper East Side	23
1986	3	3	40.767898	-73.967463	2350000.0	1500	1924	10065	0.261894	1606.0	Upper East Side	11
1987	1	1	40.764069	-73.964407	399000.0	625	1963	10065	0.191409	1606.0	Upper East Side	1
1988	1	1	40.766921	-73.962869	675000.0	700	1965	10065	0.127669	1606.0	Upper East Side	3

Semi-final Data Frame

distance to subway	complaints	10001	...	10031	10032	10034	10036	10037	10038	10039	10044	10065	price_range
0.211896	1606	0	...	0	0	0	0	0	0	0	0	1	30
0.125099	1606	0	...	0	0	0	0	0	0	0	0	1	3
0.108910	1606	0	...	0	0	0	0	0	0	0	0	1	22
0.139596	1606	0	...	0	0	0	0	0	0	0	0	1	21
0.325530	1606	0	...	0	0	0	0	0	0	0	0	1	1
0.377638	1606	0	...	0	0	0	0	0	0	0	0	1	28
0.474802	1606	0	...	0	0	0	0	0	0	0	0	1	14
0.143899	1606	0	...	0	0	0	0	0	0	0	0	1	25
0.324925	1606	0	...	0	0	0	0	0	0	0	0	1	29
0.216822	1606	0	...	0	0	0	0	0	0	0	0	1	5
0.545624	1606	0	...	0	0	0	0	0	0	0	0	1	2
0.191409	1606	0	...	0	0	0	0	0	0	0	0	1	4

Final Data Frame (with dummy variable of zip code)

Exploratory Data Analysis:

NEW YORK STATE | Services | News | Government | Local

Department of Health | Individuals/Families | Provid

You are Here: Home Page > Appendices Menu > ZIP Code Definitions of New York City Neighborhoods

ZIP Code Definitions of New York City Neighborhoods

Manhattan	Central Harlem	10026, 10027, 10030, 10037, 10039
	Chelsea and Clinton	10001, 10011, 10018, 10019, 10020, 10036
	East Harlem	10029, 10035
	Gramercy Park and Murray Hill	10010, 10016, 10017, 10022
	Greenwich Village and Soho	10012, 10013, 10014
	Lower Manhattan	10004, 10005, 10006, 10007, 10038, 10280
	Lower East Side	10002, 10003, 10009
	Upper East Side	10021, 10028, 10044, 10065, 10075, 10128
	Upper West Side	10023, 10024, 10025
	Inwood and Washington Heights	10031, 10032, 10033, 10034, 10040

Neighborhoods with zip codes in Manhattan

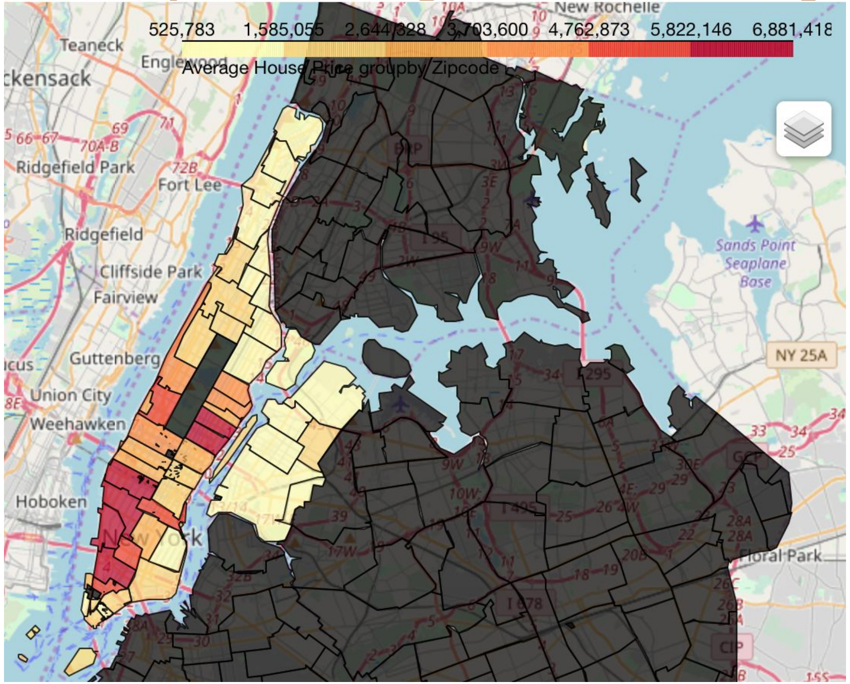
```
1 df['price'].describe()

count    5.504000e+03
mean     3.659049e+06
std      6.174104e+06
min      2.995000e+03
25%      8.250000e+05
50%      1.650000e+06
75%      3.595000e+06
max      8.800000e+07
Name: price, dtype: float64
```

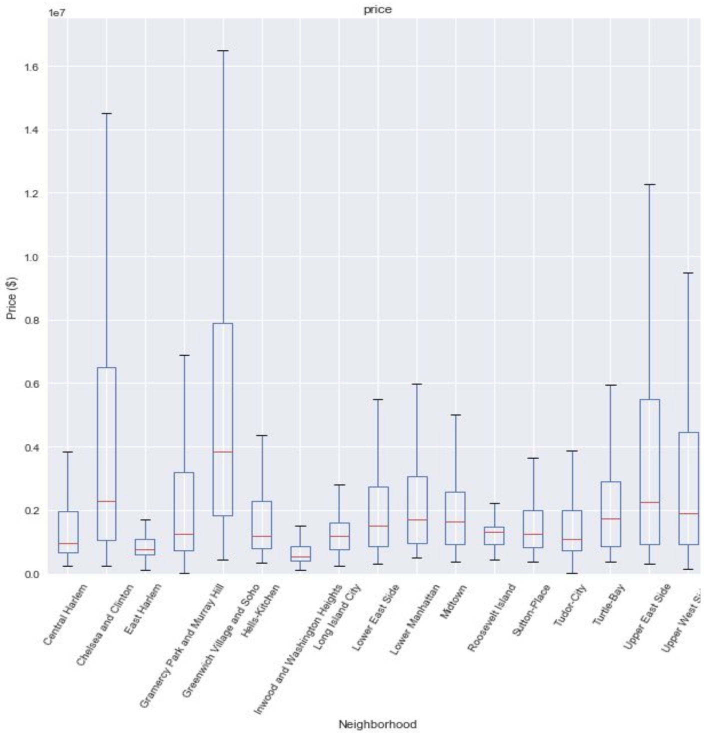
General price description

BEST OF ZILLOW

Exploratory Data Analysis:

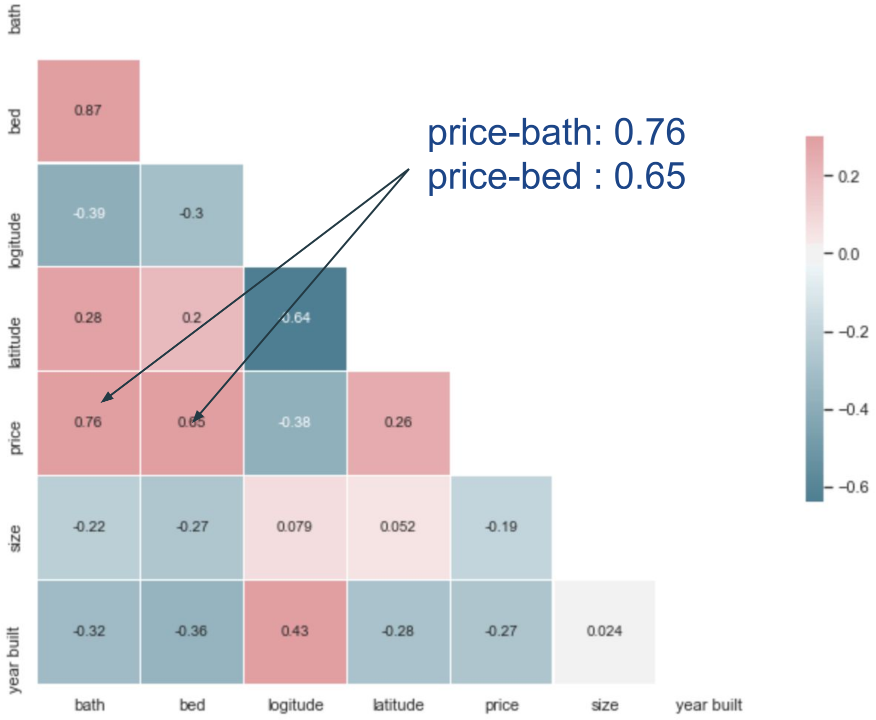


Heatmap of average price by zip code



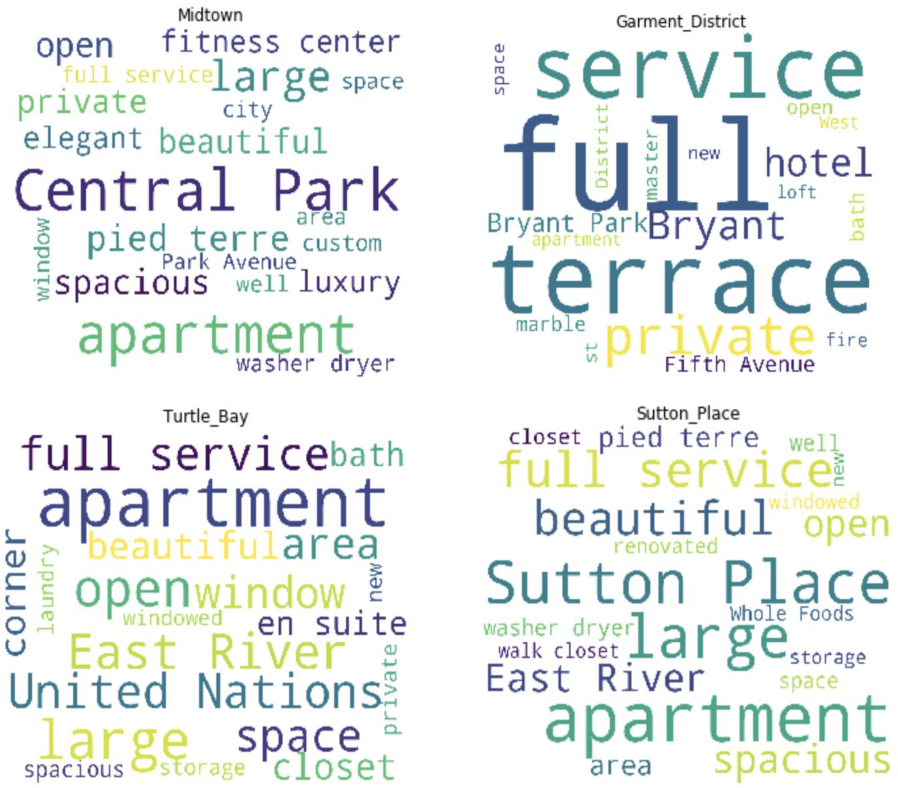
Boxplot of price for each neighborhood

Exploratory Data Analysis:

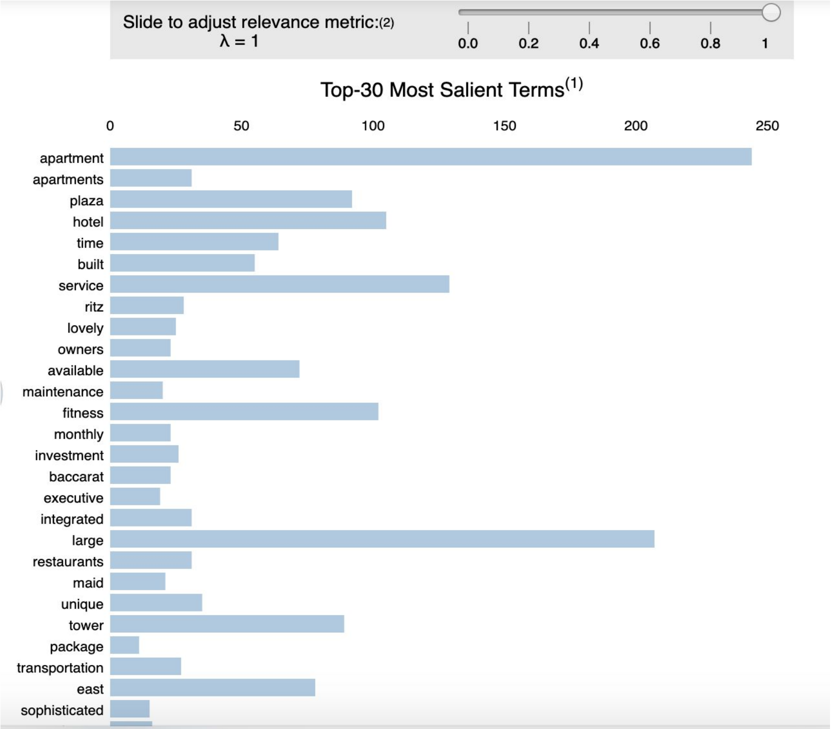


BEST OF ZILLOW

Exploratory Data Analysis:

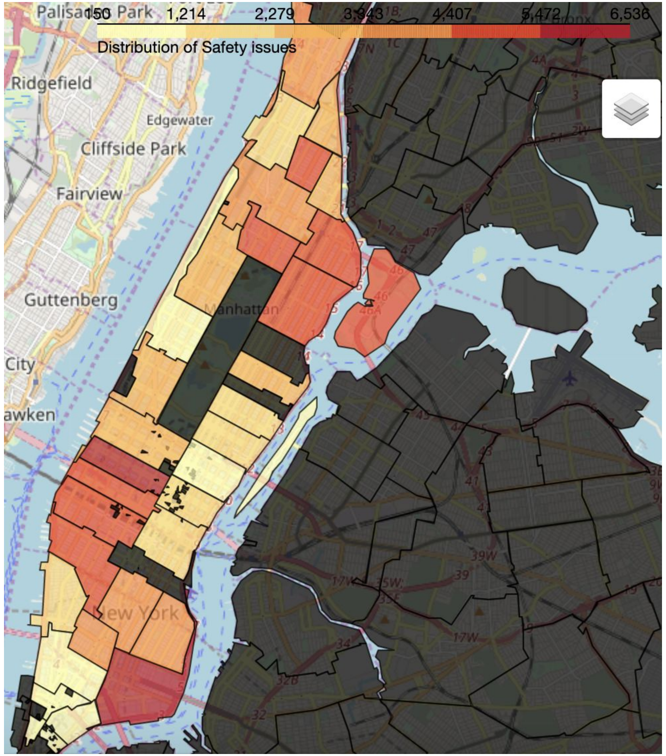


Apart from the intrinsic factors of the house itself, what are the external factors of the house price?

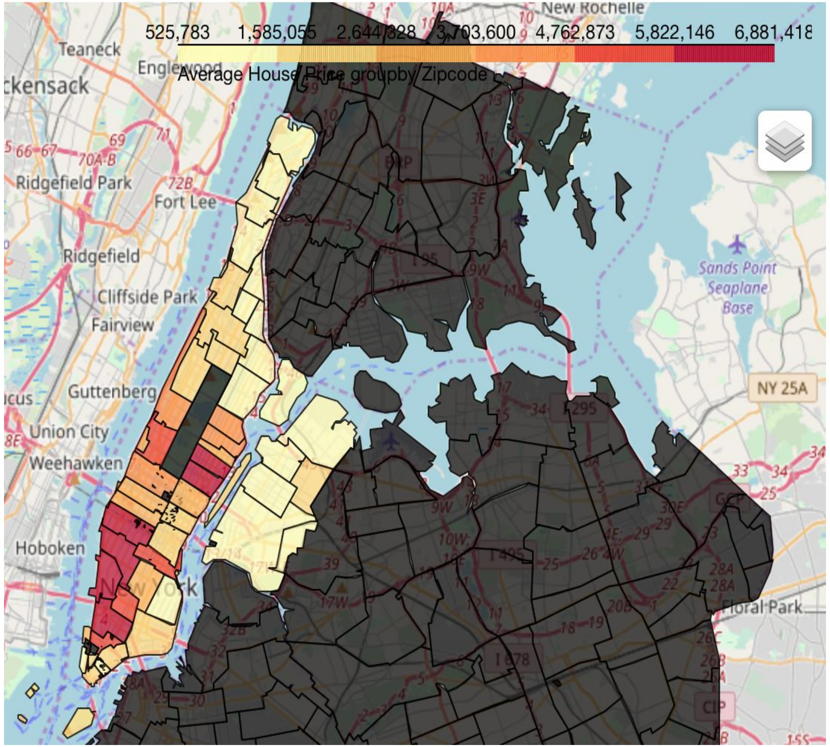


Exploratory Data Analysis:

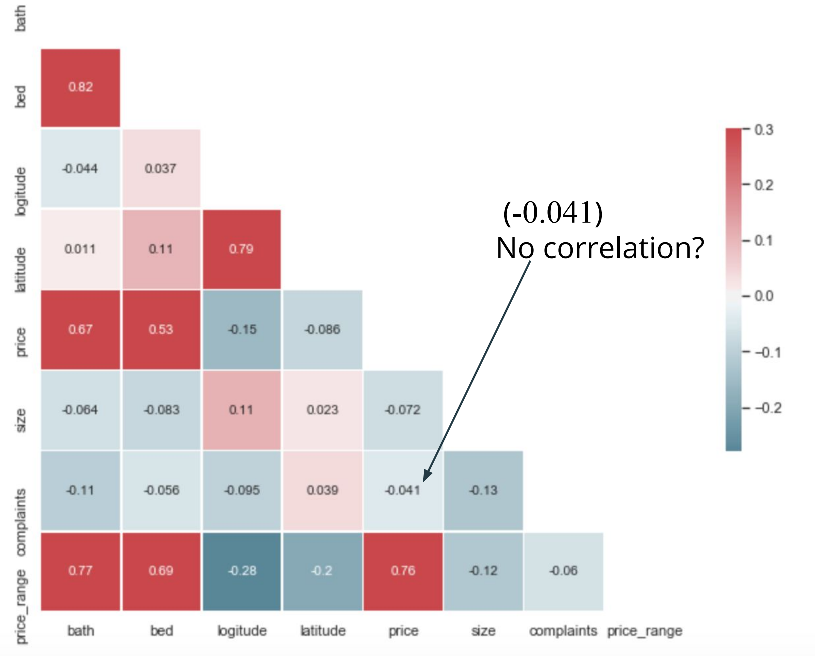
Crime Map



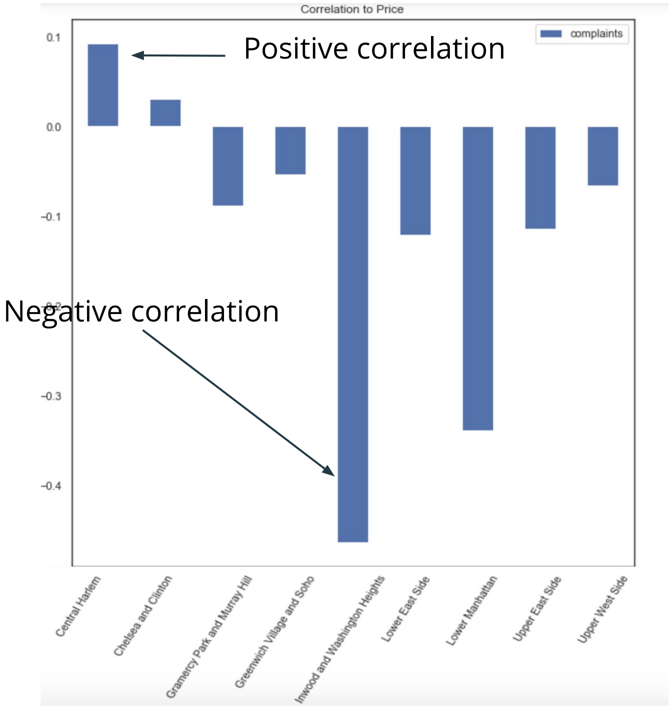
Price Map



Exploratory Data Analysis:

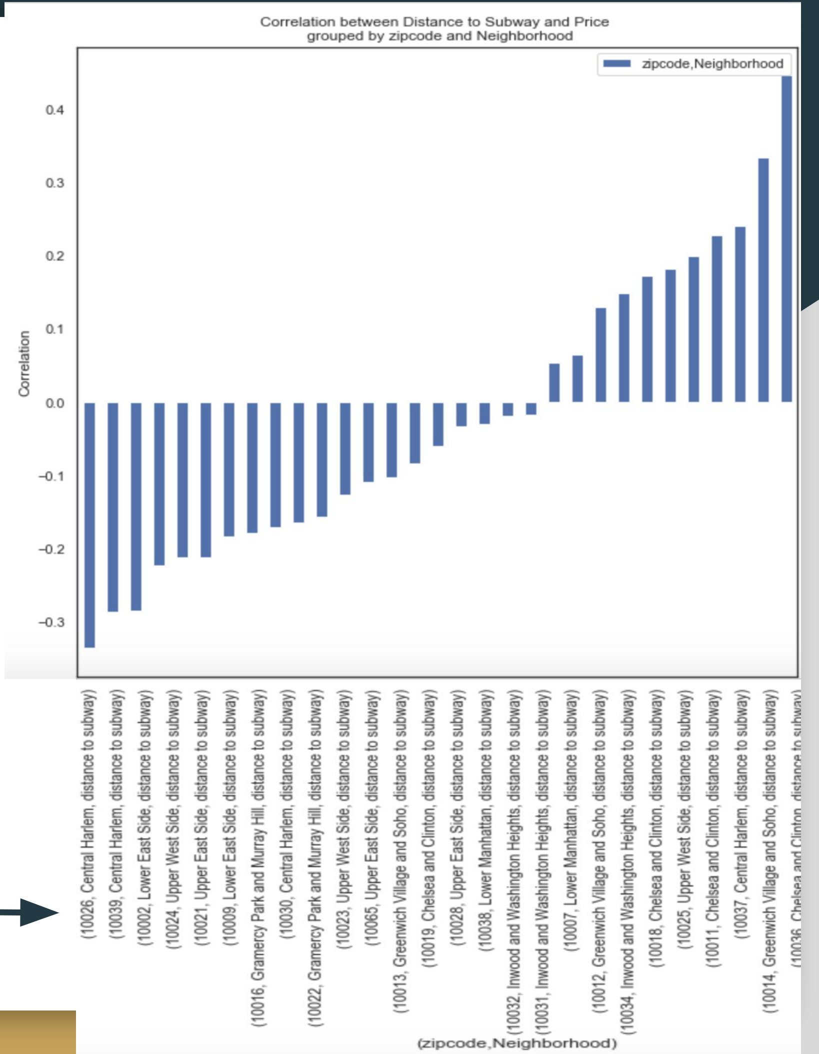


Relationship between price and number of crime complaints



Exploratory Data Analysis:

- Home price can be either positively or negatively correlated with its number of complaints, <u>conditional on zip code or neighborhood</u>

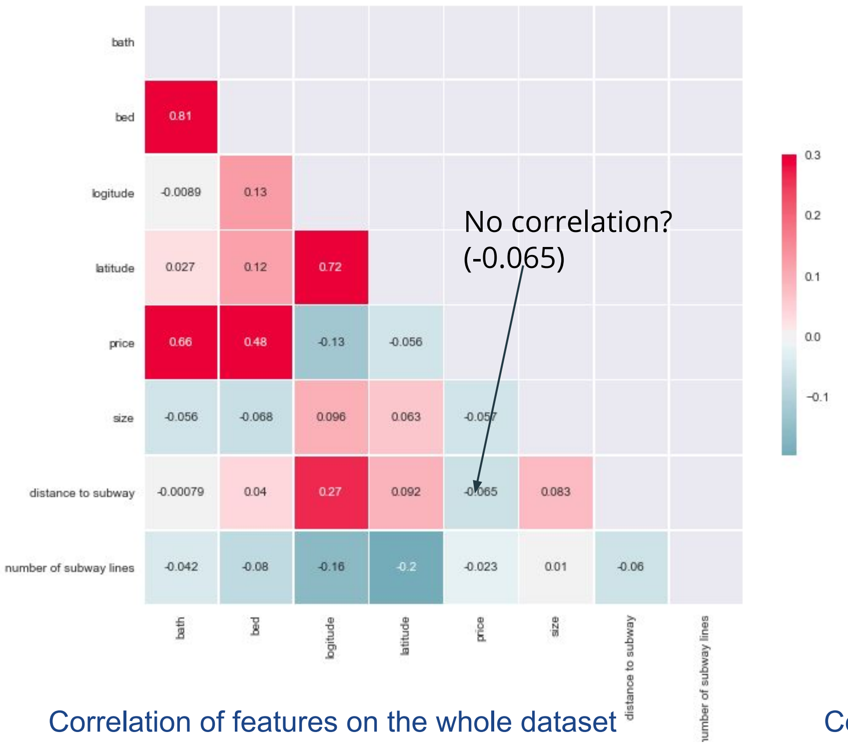


Correlation between price and distance after grouped by zip code

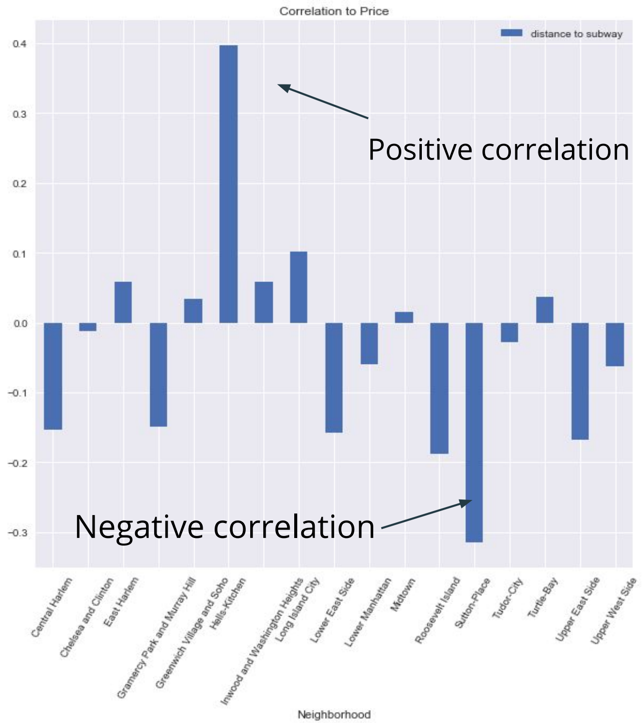
BEST OF ZILLOW

Exploratory Data Analysis:

Relationship between price and distance to the closest subway station



Correlation of features on the whole dataset

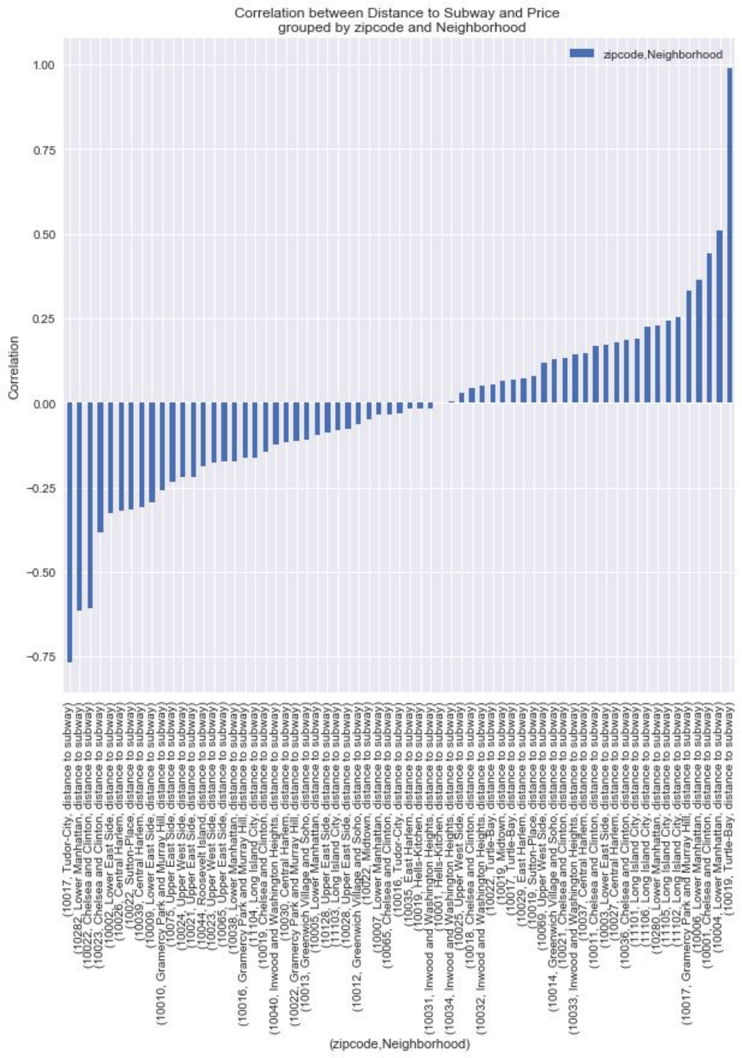


Correlation between price and distance after grouped by neighborhoods

Exploratory Data Analysis:

- Home price can be either positively or negatively correlated with its distance to the closest subway station, <u>conditional on zip code or neighborhood</u>

Correlation between price and distance after grouped by zip code



Predictive Models

3 Models:

- Linear Regression
- Random Forest
- K nearest neighborhood (K-NN)

3 Methods:

- On Price (price as continuous)
- On Price Range (price as categorical)
- On a specific neighborhood (i.e. based on zip code: 10036 Chelsea and Clinton)

Model Selection:

- Train-Test split: 70% training, 30% test
- Validation to select best depth and number of neighborhoods (k)

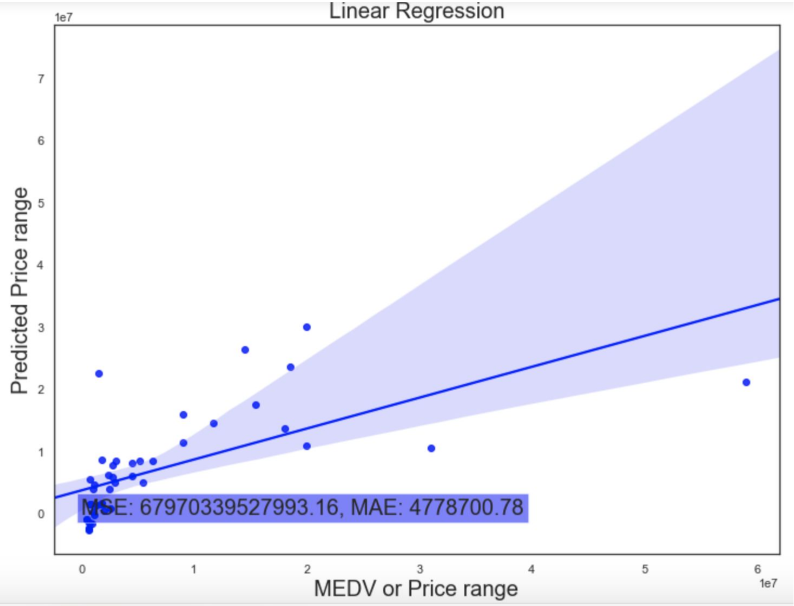
Evaluation of Models:

- MSE on test data
- R-square

Linear Regression: Price

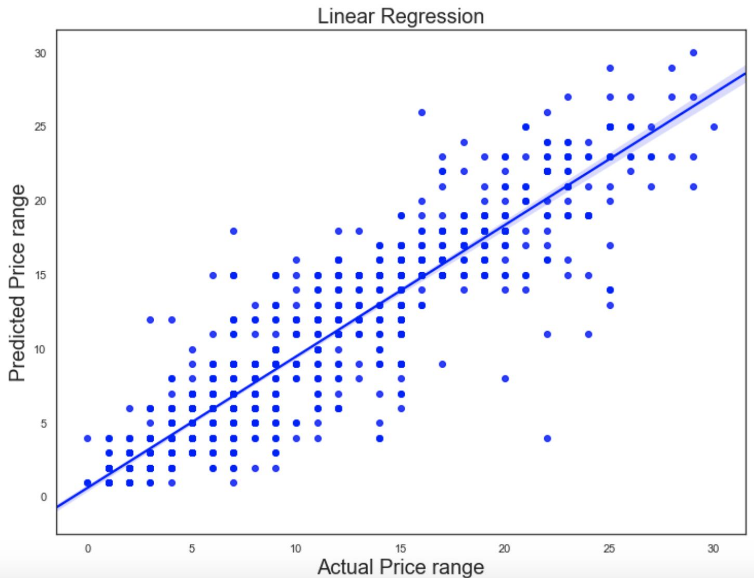
	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0056	0.039	-0.143	0.886	-0.083	0.071
bath	0.6866	0.029	23.777	0.000	0.630	0.743
bed	0.0021	0.030	0.071	0.943	-0.057	0.061
latitude	0.1327	0.155	0.858	0.391	-0.171	0.436
logitude	-0.1835	0.073	-2.518	0.012	-0.326	-0.041
size	-0.0075	0.015	-0.505	0.614	-0.037	0.022
year built	0.0478	0.017	2.867	0.004	0.015	0.080
distance to subway	-0.0202	0.019	-1.082	0.279	-0.057	0.016
complaints	0.0730	0.079	0.923	0.356	-0.082	0.228
10002	-0.0916	0.207	-0.443	0.658	-0.497	0.314
10003	-0.0220	0.158	-0.139	0.890	-0.332	0.288

```
R^2 for training data is  0.6815091193255008
R^2 for testing data is  0.3919420769960831
```



Linear Regression: Price Range

```
Text(35.0, 10.0, 'MSE: 15.8, MAE: 3.12')
```

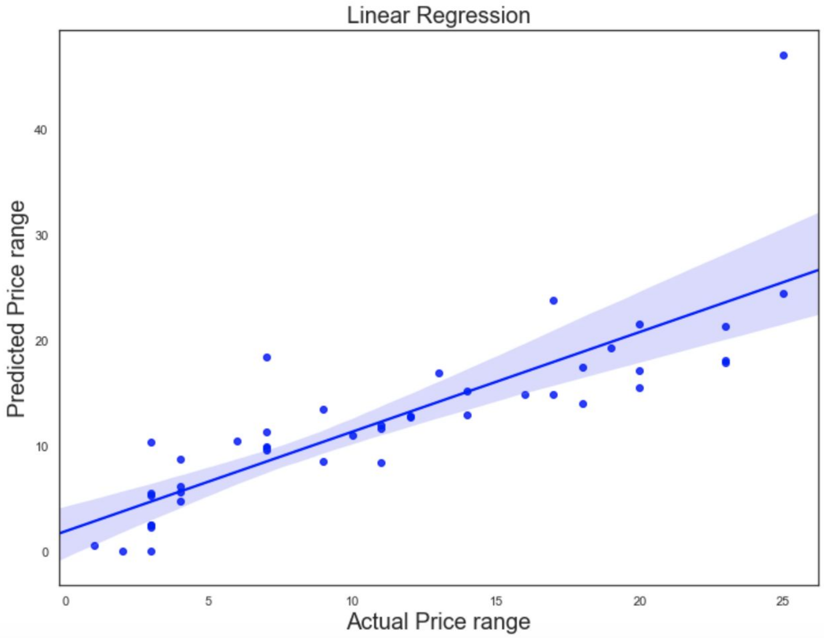


	coef	std err	t	P>|t|	[0.025	0.975]
const	-4658.0842	728.040	-6.398	0.000	-6085.683	-3230.485
bath	2.0117	0.071	28.196	0.000	1.872	2.152
bed	1.0214	0.076	13.469	0.000	0.873	1.170
latitude	-15.6507	4.932	-3.173	0.002	-25.321	-5.980
logitude	-71.3726	7.547	-9.457	0.000	-86.172	-56.574
size	-2.25e-06	8.67e-07	-2.595	0.009	-3.95e-06	-5.5e-07
year built	0.0191	0.002	9.742	0.000	0.015	0.023
zipcode	-0.0018	0.007	-0.247	0.805	-0.016	0.012
distance to subway	-2.2092	0.425	-5.195	0.000	-3.043	-1.375
complaints	-9.762e-05	5.08e-05	-1.924	0.055	-0.000	1.89e-06

```
Training R-Square 0.6808705242786378
Testing R-Square 0.6819111719170822
16.362680754869984
```

Linear Regression: Certain Region

Text(35.0, 10.0, 'MSE: 22.45, MAE: 3.02')

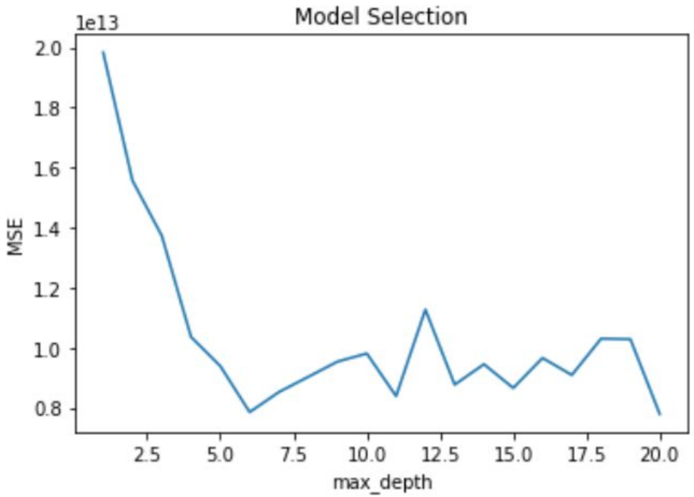


Linear regression based on zip code

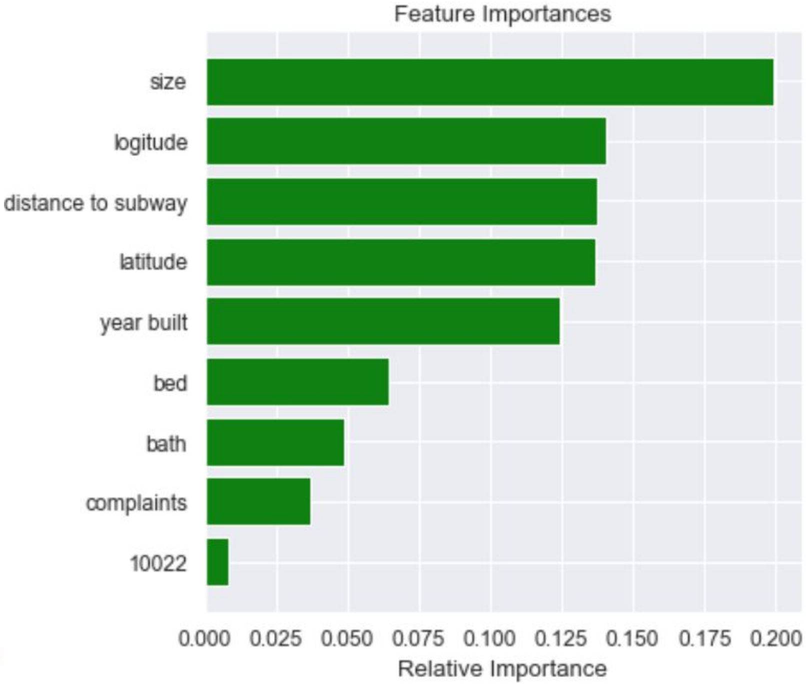
	coef	std err	t	P>|t|	[0.025	0.975]
bath	1.5519	0.387	4.012	0.000	0.784	2.320
bed	1.1628	0.468	2.486	0.015	0.234	2.092
latitude	110.8840	271.862	0.408	0.684	-428.980	650.748
logitude	-419.8556	206.948	-2.029	0.045	-830.812	-8.899
size	-3.441e-06	4.19e-06	-0.821	0.414	-1.18e-05	4.88e-06
year built	0.0018	0.015	0.119	0.905	-0.029	0.033
distance to subway	0.2264	3.903	0.058	0.954	-7.525	7.978
complaints	-22.1509	7.487	-2.959	0.004	-37.019	-7.283

R^2 for training data is 0.768634555822183
R^2 for testing data is 0.7261922497865134

Random Forest:



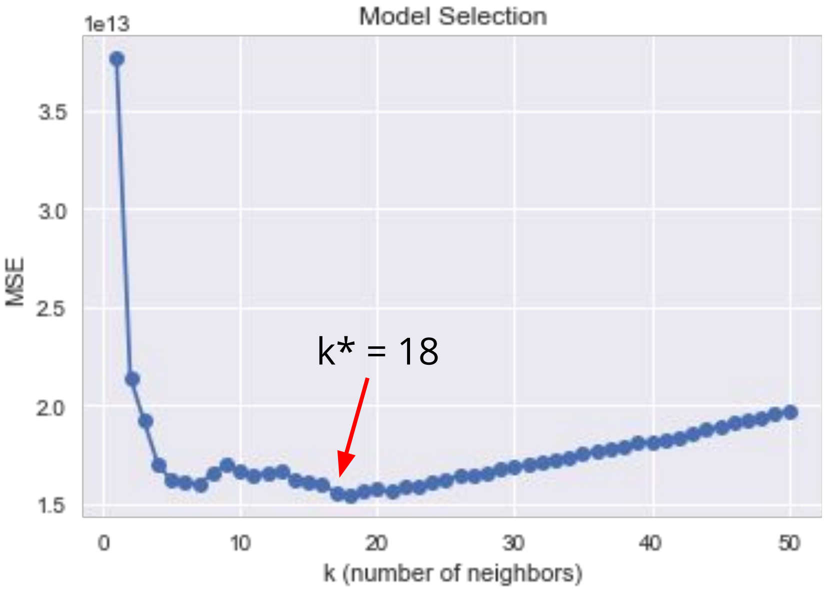

```
Training R-Square 0.8018855891606547
Testing R-Square 0.7198827745842173
MSE 9238721559846.172
```



Optimal Tree Level= 5

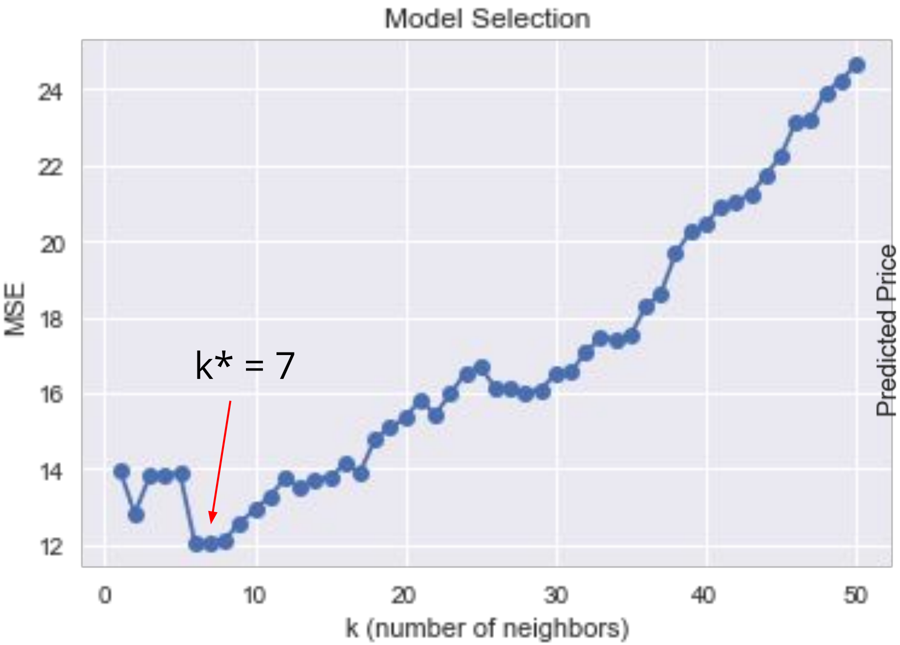
K-NN: Regressor for price

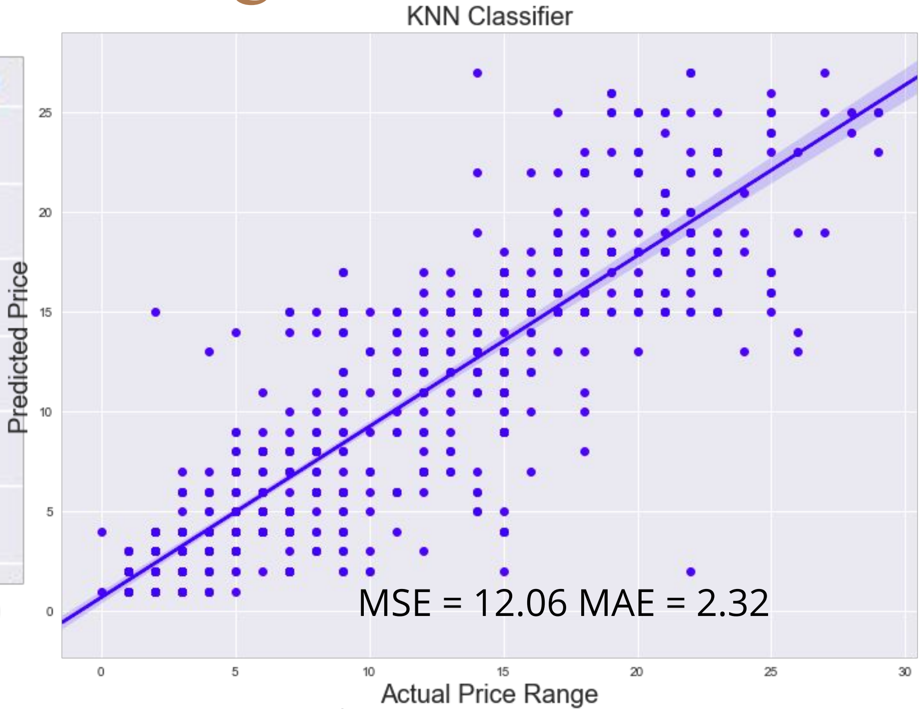
Features used: 'latitude','longitude','bath','bed','distance to subway','complaints'



Performance on test data

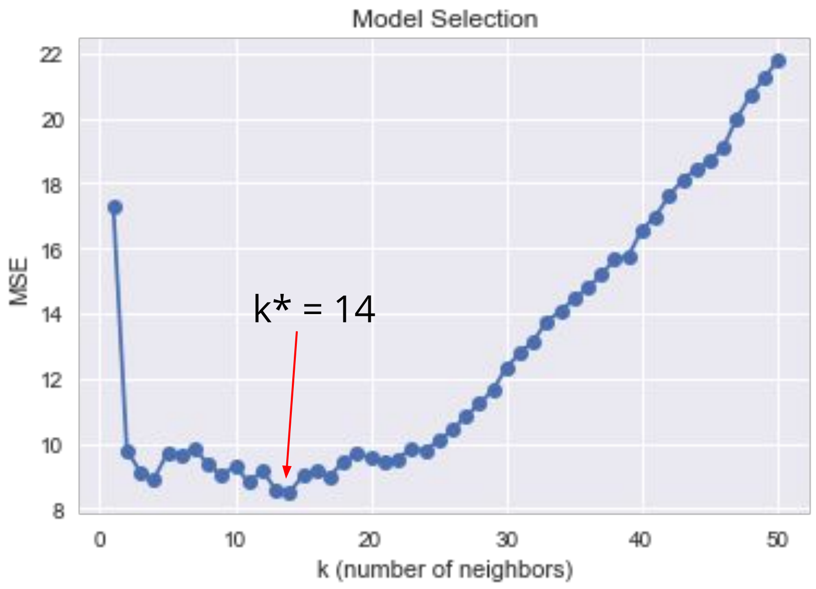
K-NN: Classifier for price range

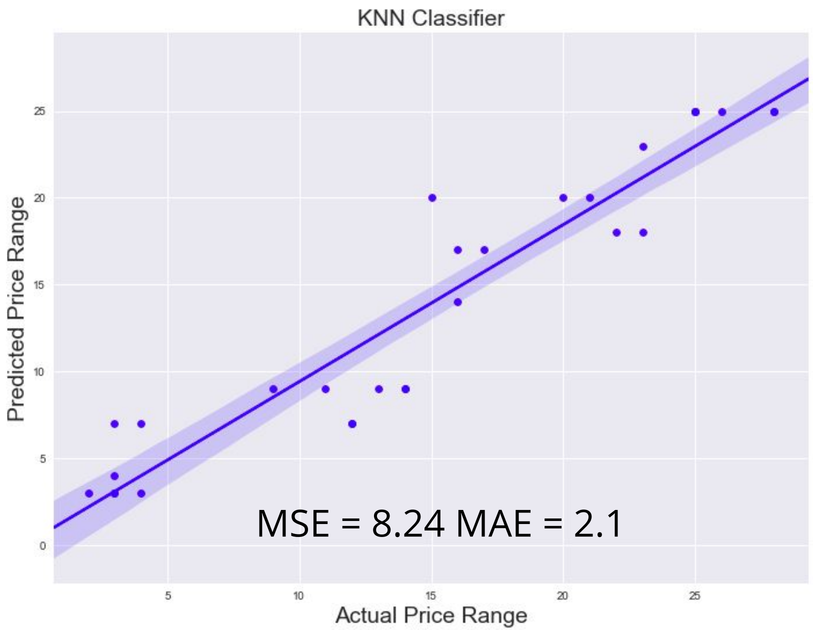




Performance on test data

K-NN: Certain Region





Performance on test data

Conclusions:

	MSE on Test Data		
Model	Price	Price Range	Region
Linear Regression	6.7e+13	15.99	12.72
Random Forest	**9.2e+12**	14.72	13.56
KNN	2.4e+13	**12.06**	**8.24**

- Prediction based on regions works better than overall dataset
- Factors that have significant impact on home price:

  (1) Inherent: number of bedrooms and bathrooms (correlation and linear regression), ‘Service’, ‘Central Park’ , ‘Fifth Avenue’ (word could)

  (2) External: # Crime Complaints, Distance from home to subway station entrance (correlation and linear regression)

BEST OF ZILLOW

References:

- https://www.zillow.com/
- https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm
- https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243
- https://data.ny.gov/Transportation/NYC-Transit-Subway-Entrance-And-Exit-Data/i9wp-a4ja
- https://en.wikipedia.org/wiki/Haversine_formula

Q&A