Linear Regression of Property Values Queens, NY

Gavin Jones

Goals

Understanding Queens Real Estate

Establish a Linear Regression that can provide insight on pricing drivers across the Borough

Understand scope of publically available data

Methodology

Data Collection

Web Scrape using BeautifulSoup to establish current pricing

Attach complimentary information collected from public databases

Methodology

Data Collection

Web Scrape using BeautifulSoup to establish current pricing

Attach complimentary information provided in public databases

Processing

Clean data into a pandas dataframe for analysis

Conduct a linear regression:

Ridge, Lasso, Polynomial Analysis

Data

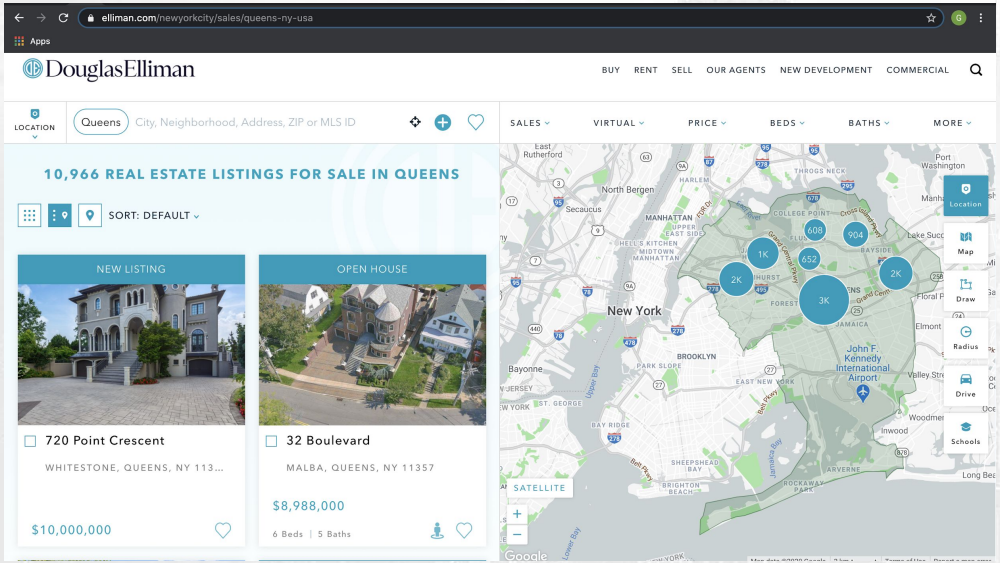
Terms and Conditions

Data

Terms and Conditions

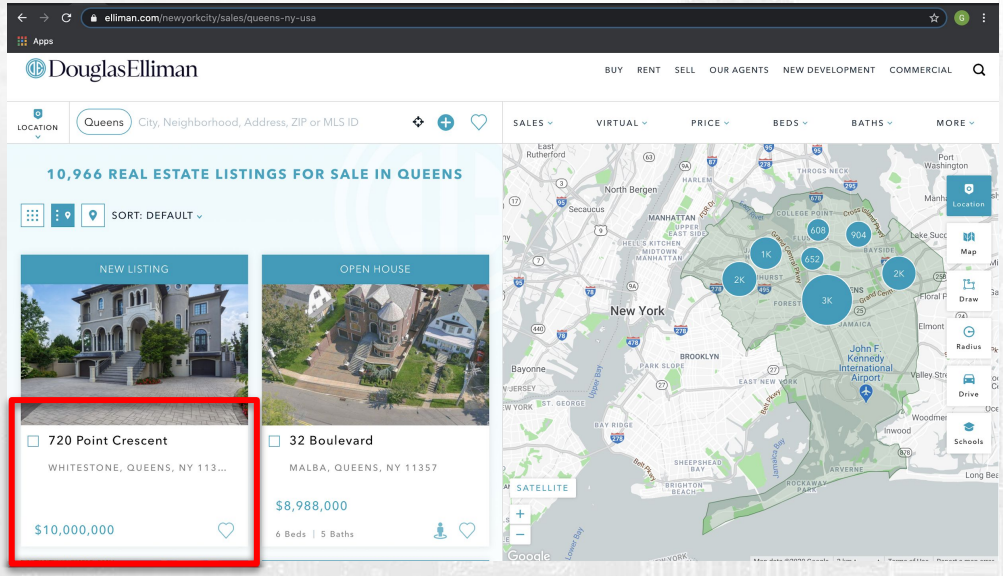
Process

Data Cleaning



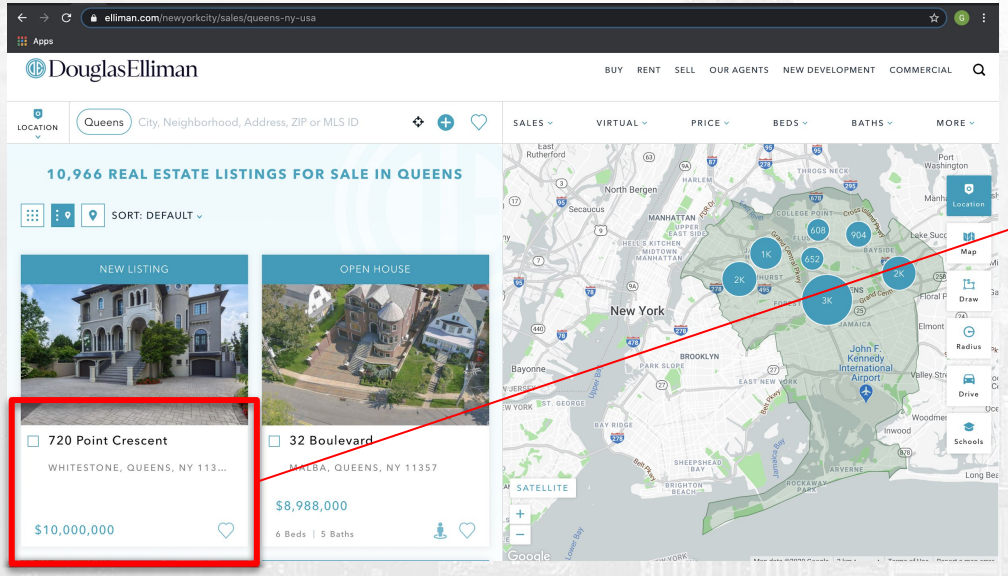
Process

Data Cleaning



Process

Data Cleaning




```
In [24]: df2.corr()
```

Out[24]:

	Price	beds	Year_Built	Bathroom	Patio	Air_Condition
Price	1.000000	0.402460	0.297136	0.591069	0.126524	0.128
beds	0.402460	1.000000	0.236975	0.626246	-0.066010	-0.094
Year_Built	0.297136	0.236975	1.000000	0.400404	-0.009970	0.101
Bathroom	0.591069	0.626246	0.400404	1.000000	0.060903	0.025
Patio	0.126524	-0.066010	-0.009970	0.060903	1.000000	0.201
Air_Conditioning	0.128515	-0.094894	0.101310	0.025091	0.201860	1.000
Garage_-_Attached	0.196999	-0.007516	0.217963	0.091501	0.142653	0.073
Brick	0.227525	0.008049	0.299488	0.128270	0.093619	0.075
Vinyl_(F)	-0.092264	-0.016635	-0.063479	0.004886	0.049659	0.005
Interior_Sq_Ft	0.630408	0.554850	0.313898	0.622816	0.101483	0.007
Exterior_Acres	0.387505	0.160270	0.114400	0.269965	0.077829	0.136
zipcode_median	0.556748	0.152211	0.082822	0.257807	0.172313	0.121
Neighborhood_median	0.618925	0.104993	0.111912	0.246646	0.194744	0.133
Miles to Penn Station	-0.249140	-0.051031	0.092571	-0.065541	-0.073710	0.008
Fare Zone	-0.268254	-0.090366	0.100290	-0.071575	-0.059138	0.031

Process

Regression Analysis: First Pass

Target = Price

Process

Regression Analysis: First Pass

Target = Price

Attributes:

URL, Property Type, Address, Neighborhood, Zip Code, Broker's Description, Bedrooms, Bathrooms, House Sq. Ft., Property Sq. Ft., Year Built, Patio, Pool, Air Conditioning, Garage, Water View, New Construction, Stucco Siding, Brick Siding, Pre-War, Courtyard, Miles Away from Penn Station, Wood Heat, Garde, Pets Allowed, Stone Siding, Private Storage, Terrace, Aluminum Siding, Cedar Siding, Wooded Area, Wood Siding, Vinyl Siding

Process

Regression Analysis: First Pass – Linear Regression

```
cv_results(df_lin_X, df_lin_y, model=LinearRegression(normalize=True), scoring='r2',

Test r2 :                 Mean = 0.4321    Range = (-0.0278, 0.6264)
Train/Test r2 Ratio :     Mean = -4.0007   Range = (-25.2899, 1.8139)


Test Scores: [ 0.62640696  0.59014008  0.5847681   0.38714997 -0.02781025]
Train Scores: [0.68312346 0.70025407 0.69907071 0.70226339 0.70331868]
```

```
cv_results(df_lin_X, (
Test r2 :
Train/Test r2 Ratio :


Test Scores: [ 0.62640696  0.59014008  0.5847681   0.38714997 -0.02781025]
Train Scores: [0.68312346 0.70025407 0.69907071 0.70226339 0.70331868]

Best Model Feature coefficient results:
beds:              10095.30
Year_Built:        566.74
Bathroom:          116659.31
Patio:             -47617.62
Air_Conditioning:  56278.20
Garage_-_Attached: 13430.62
Brick:             51616.49
Vinyl_(F):         2376.56
Interior_Sq_Ft:    112.39
Exterior_Acres:    1272619.43
zipcode_median:    0.52
Neighborhood_median: 0.48
```

Process

Regression Analysis: First Pass – Polynomial

```
poly_cv_results(X_poly,df_poly_y, LinearRegression(normalize=True))
```

```
Test r2 :                 Mean = 0.4985    Range = (0.4253, 0.6081)
Train/Test r2 Ratio :     Mean = 1.6739    Range = (1.2964, 1.8998)


Test Scores: [0.49024075 0.60808416 0.52623665 0.44266605 0.42526032]
Train Scores: [0.82030037 0.78834214 0.88718207 0.80297006 0.80791948]
```

Process

Regression Analysis: Feature Engineering

Attributes:

URL, Property Type, Address, Neighborhood, Zip Code, Broker's Description, Bedrooms, Bathrooms, House Sq. Ft., Property Sq. Ft., Year Built, Patio, Pool, Air Conditioning, Garage, Water View, New Construction, Stucco Siding, Brick Siding, Pre-War, Courtyard, Miles Away from Penn Station, Wood Heat, Garde, Pets Allowed, Stone Siding, Private Storage, Terrace, Aluminum Siding, Cedar Siding, Wooded Area, Wood Siding, Vinyl Siding

DON'T NEED

Process

Regression Analysis: Feature Engineering

Attributes:

URL, Property Type, Address, Neighborhood, Zip Code, Broker's Description, Bedrooms, Bathrooms, House Sq. Ft., Property Sq. Ft., Year Built, Patio, Pool, Air Conditioning, Garage, Water View, New Construction, Stucco Siding, Brick Siding, Pre-War, Courtyard, Miles Away from Penn Station, Wood Heat, Garde, Pets Allowed, Stone Siding, Private Storage, Terrace, Aluminum Siding, Cedar Siding, Wooded Area, Wood Siding, Vinyl Siding

CONTINUOUS

Process

Regression Analysis: Feature Engineering

Attributes:

URL, Property Type, Address, Neighborhood, Zip Code, Broker's Description, Bedrooms, Bathrooms, House Sq. Ft., Property Sq. Ft., Year Built, Patio, Pool, Air Conditioning, Garage, Water View, New Construction, Stucco Siding, Brick Siding, Pre-War, Courtyard, Miles Away from Penn Station, Wood Heat, Garde, Pets Allowed, Stone Siding, Private Storage, Terrace, Aluminum Siding, Cedar Siding, Wooded Area, Wood Siding, Vinyl Siding

DUMMY COLUMNS

Process

Regression Analysis: Feature Engineering

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.469e+06	7.38e+05	-3.346	0.001	-3.92e+06	-1.02e+06
beds	3480.9735	7546.732	0.461	0.645	-1.13e+04	1.83e+04
Year_Built	1012.7243	386.828	2.618	0.009	253.831	1771.617
Bathroom	1.048e+05	1.06e+04	9.853	0.000	8.39e+04	1.26e+05
Patio	-3.459e+04	2.13e+04	-1.625	0.104	-7.63e+04	7174.386
Air_Conditioning	4.844e+04	1.89e+04	2.556	0.011	1.13e+04	8.56e+04
Garage_-_Attached	-1.208e+04	2.44e+04	-0.496	0.620	-5.99e+04	3.57e+04
Brick	2.831e+04	2.04e+04	1.386	0.166	-1.18e+04	6.84e+04
Vinyl_(F)	-3786.3677	2.91e+04	-0.130	0.897	-6.1e+04	5.34e+04
Interior_Sq_Ft	154.2665	12.725	12.123	0.000	129.303	179.230
Exterior_Acres	1.877e+06	1.94e+05	9.654	0.000	1.5e+06	2.26e+06
zipcode_median	0.2545	0.062	4.077	0.000	0.132	0.377
Neighborhood_median	0.6668	0.044	15.041	0.000	0.580	0.754
Miles to Penn Station	6064.2984	3513.765	1.726	0.085	-829.133	1.3e+04
Fare Zone	-7.588e+04	1.7e+04	-4.462	0.000	-1.09e+05	-4.25e+04

P>|t|
0.001
0.645
0.009
0.000
0.104
0.011
0.620
0.166
0.897
0.000
0.000
0.000
0.000
0.085
0.000

Process

Regression Analysis: Second Pass

Process

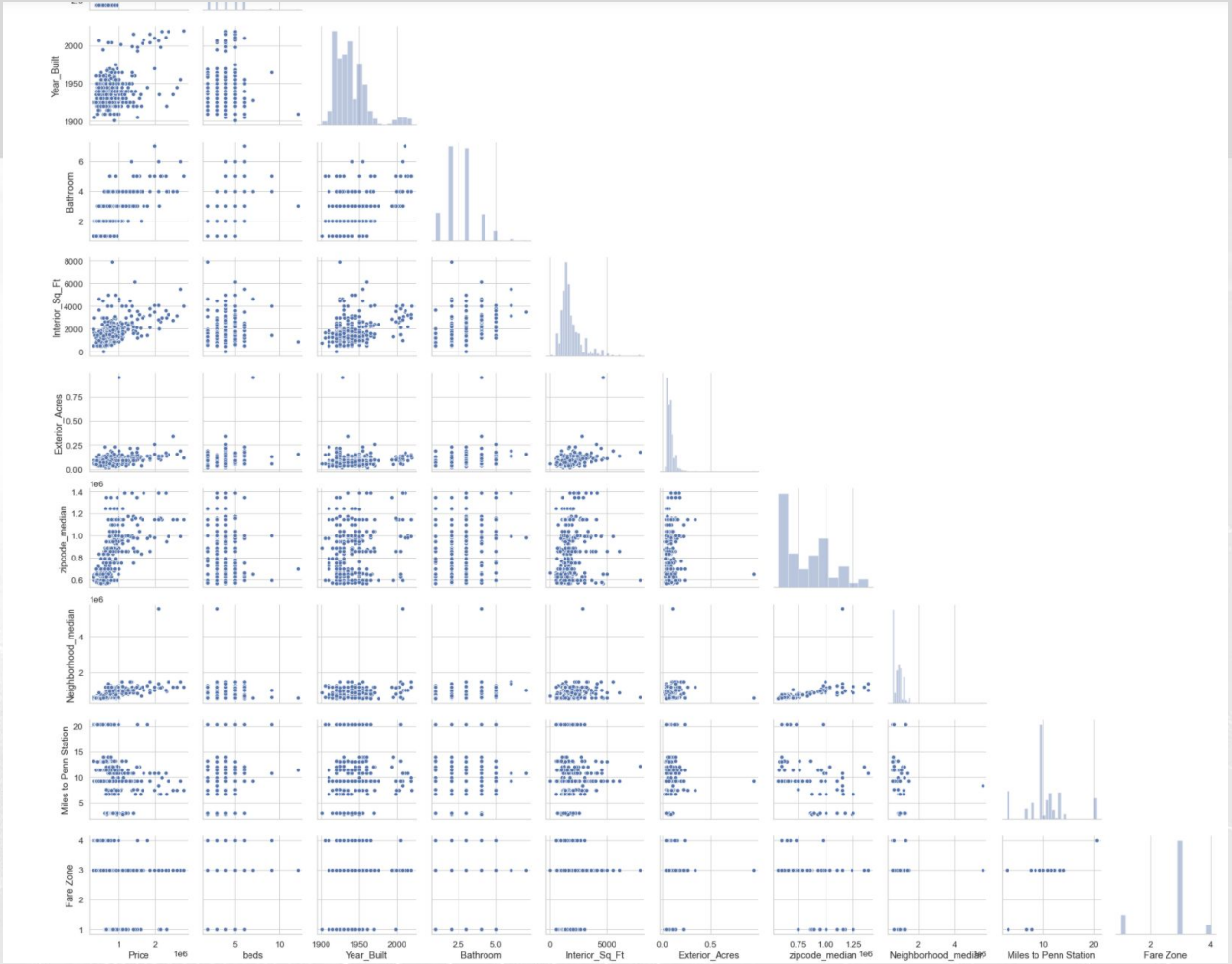
Regression Analysis: Second Pass

```
In [245]: poly_cv_results(X_poly, df_poly_y, model=LassoCV(
```

```
Fare Zone:        -69844.02
beds^2:           -0.00
beds Year_Built: -14740.33
beds Bathroom:    -0.00
beds Patio:       0.00
beds Air_Conditioning: 3350.91
beds Garage_-_Attached: -0.00
beds Brick:       -0.00
beds Vinyl_(F):   0.00
beds Interior_Sq_Ft: 0.00
beds Exterior_Acres: -0.00
beds zipcode_median: -0.00
beds Neighborhood_median: -0.00
beds Miles to Penn Station: 0.00
beds Fare Zone:   11196.18
Year_Built^2:     16792.86
Year_Built Bathroom: 0.00
Year_Built Patio: 3324.57
Year_Built Air_Conditioning: 24657.69
Year_Built Garage_-_Attached: 0.00
```

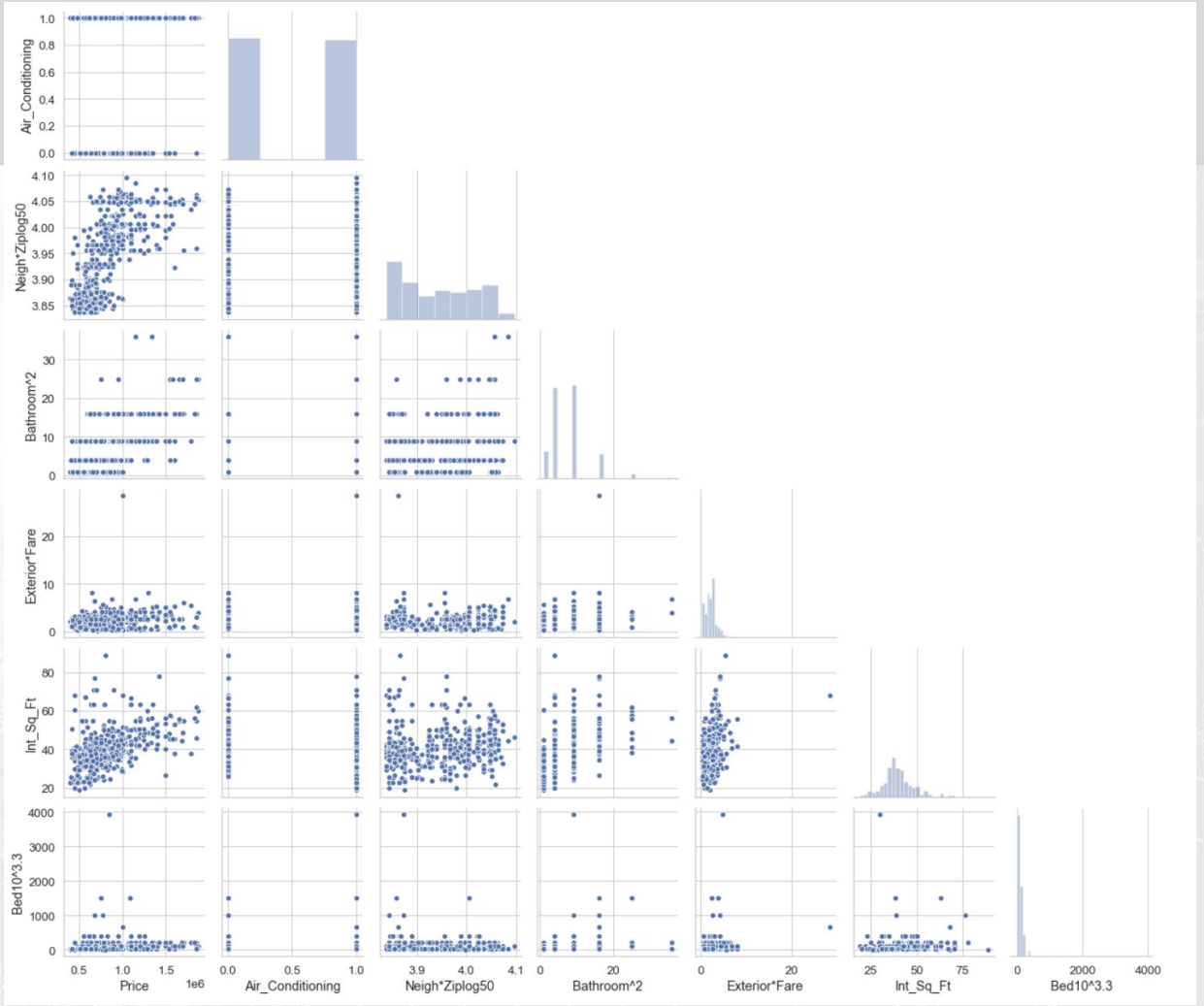
Process

Pre Engineering

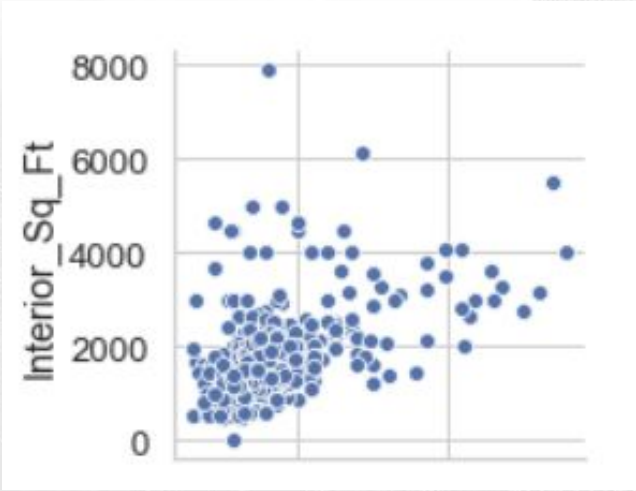


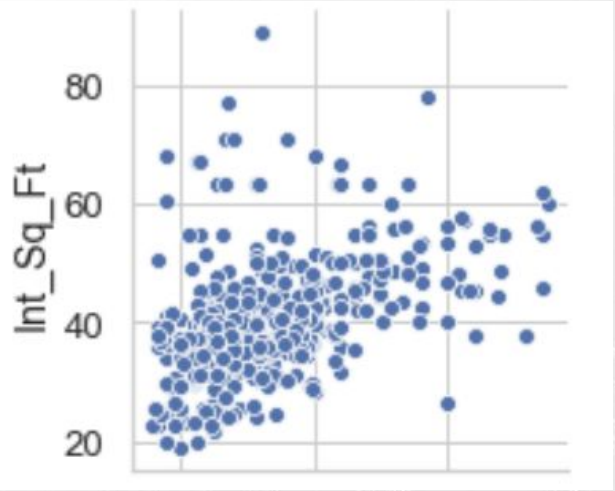
Process

Post Engineering

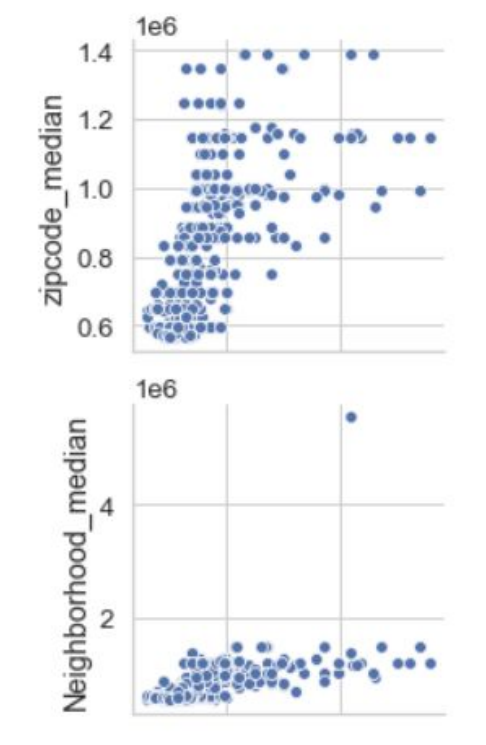


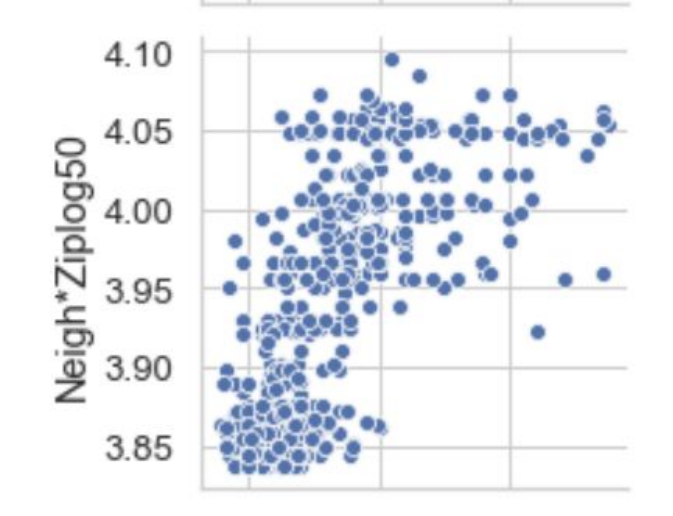
Process



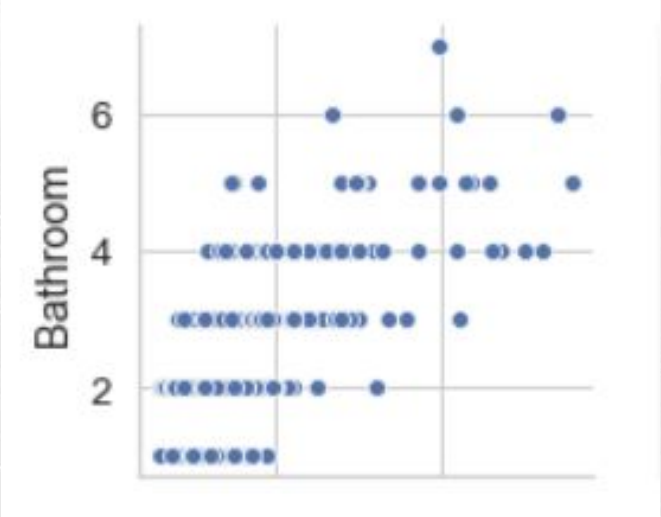


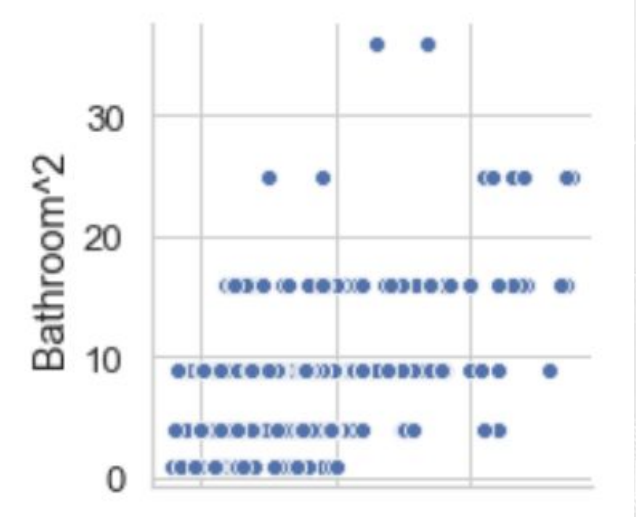
Process


1e6
1.4
1.2
1.0
0.8
0.6
zipcode_median
1e6
4
2
Neighborhood_median

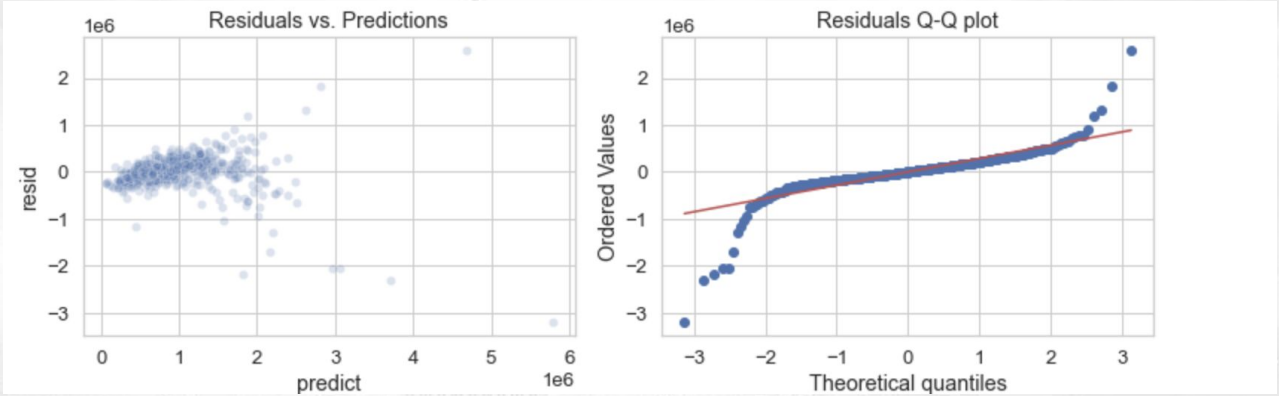

4.10
4.05
4.00
3.95
3.90
3.85
Neigh*Ziplog50

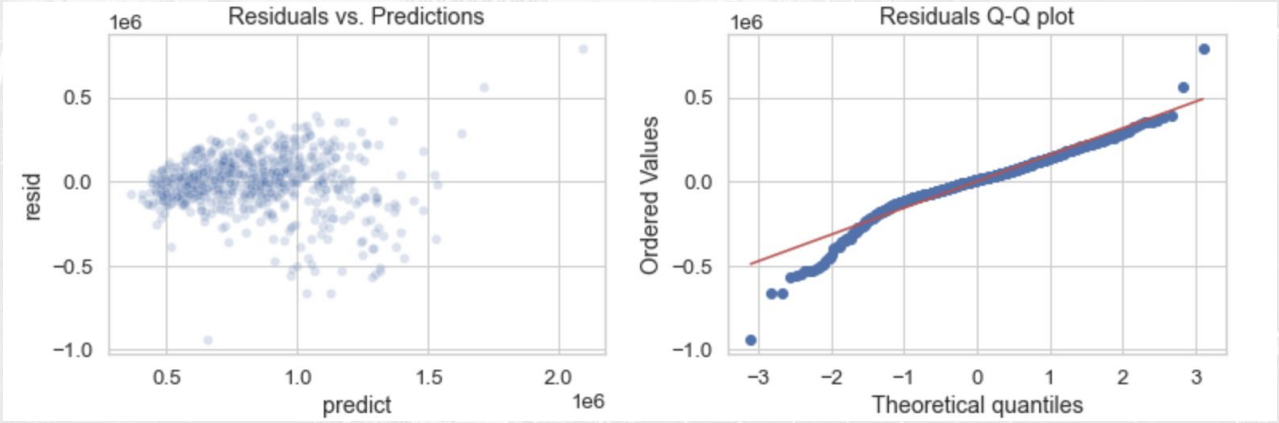
Process





Process


Residuals vs. Predictions
1e6
resid
predict
1e6
Residuals Q-Q plot
1e6
Ordered Values
Theoretical quantiles


Residuals vs. Predictions
1e6
resid
predict
1e6
Residuals Q-Q plot
1e6
Ordered Values
Theoretical quantiles

Results

```
Test r2 :                 Mean = 0.5747    Range = (0.505, 0.6579)
Train/Test r2 Ratio :     Mean = 1.2097    Range = (1.0212, 1.3832)


Test Scores: [0.57066842 0.54891816 0.50501156 0.59084961 0.65792041]
Train Scores: [0.69314508 0.69181842 0.6985498  0.69090526 0.67186051]

Best Model Feature coefficient results:
Air_Conditioning: 68981.66
Neigh*Ziplog50:  2136673.85
Bathroom^2:       13604.46
Exterior*Fare:    15122.92
Int_Sq_Ft:        7767.76
Bed10^:           57.54
```

Further Analysis

Need More Data:

1. Tax Data
2. Location Data
3. Commuting Data
4. Entertainment/Zoning
5. Queens is diverse – not just in demographics. Segmenting could help!


Thank You!