NYC AirBNB

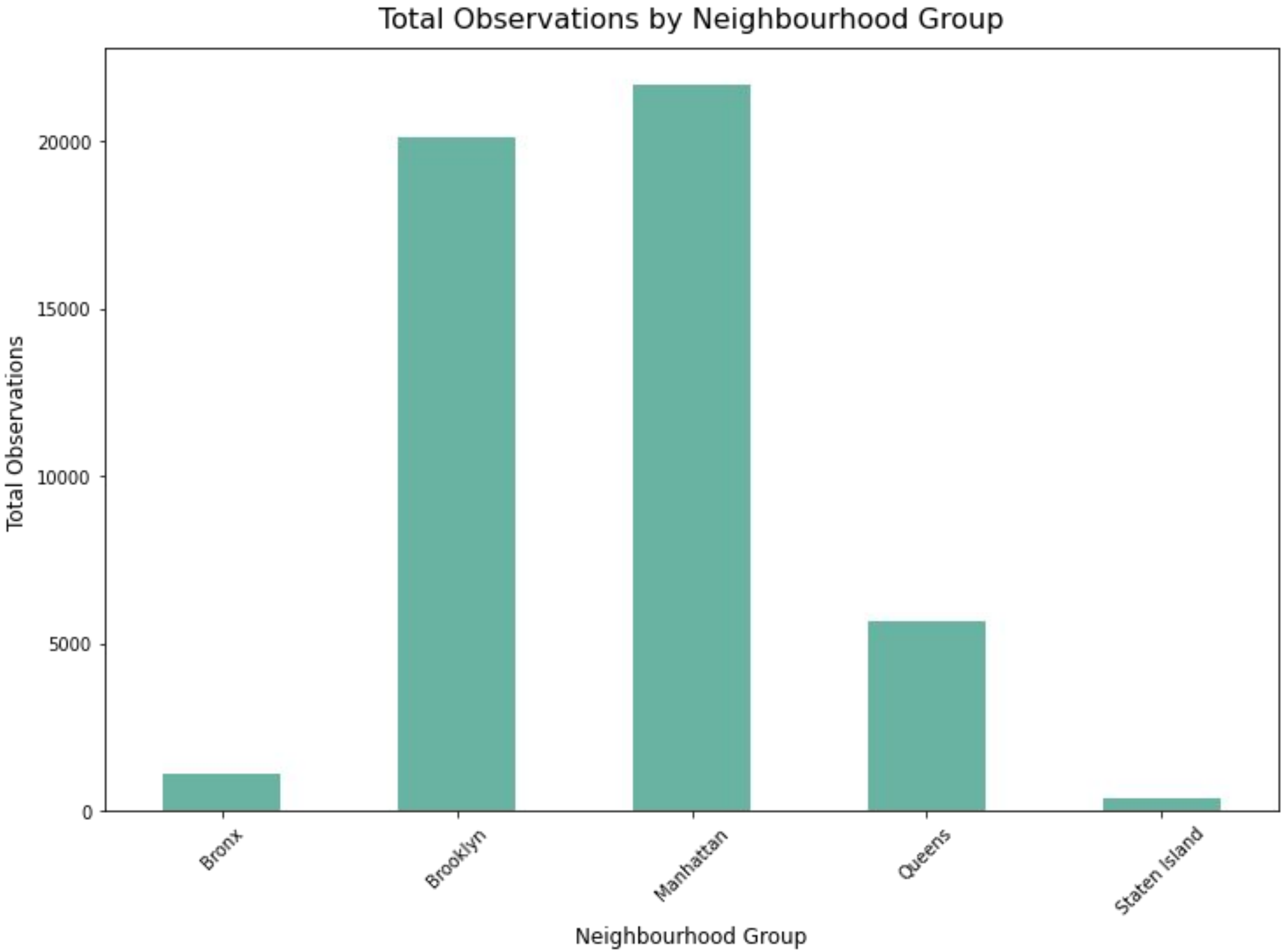
Predicting Prices of AirBNBs in NYC

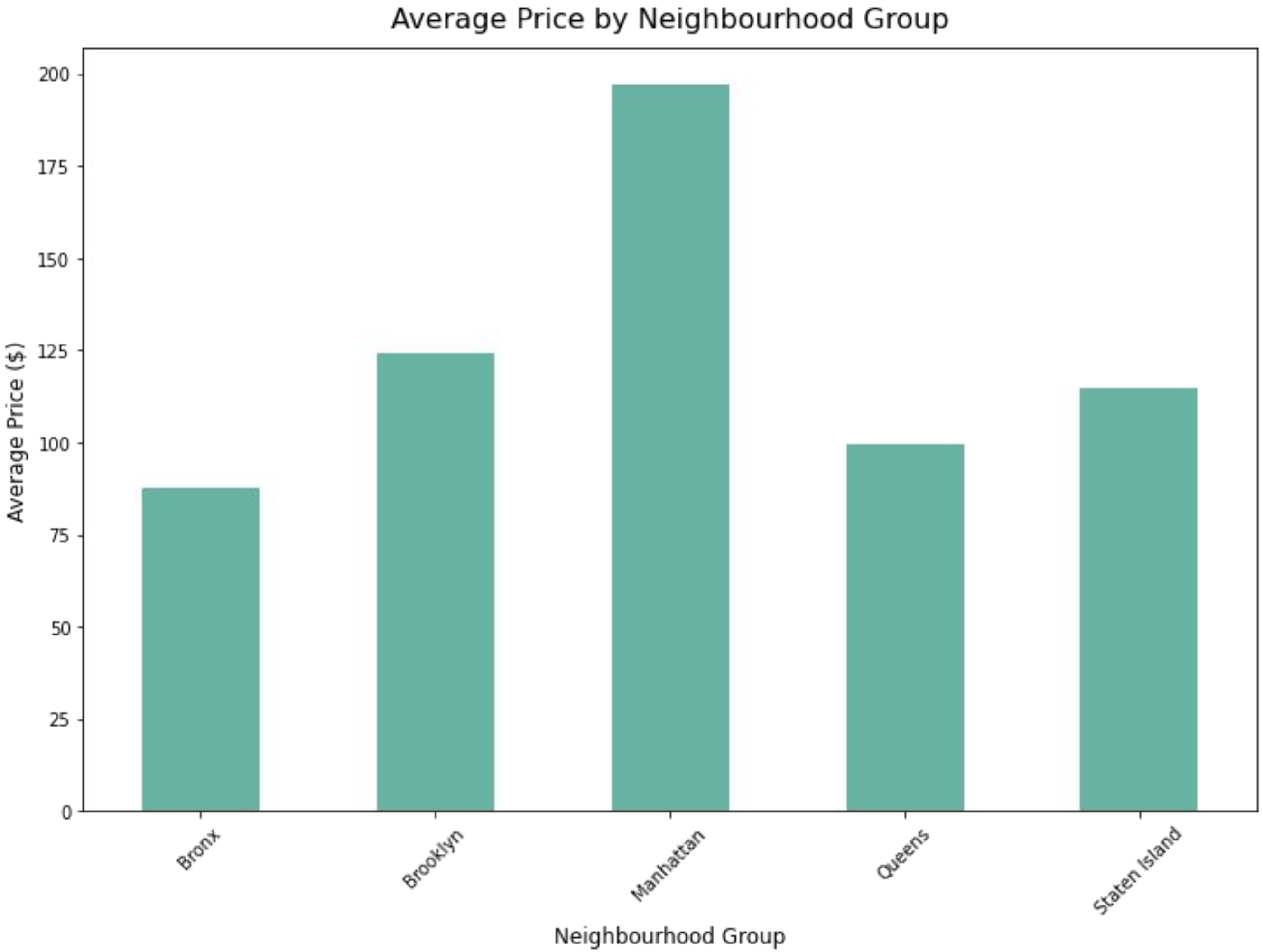
Aditi Gajjar, Divya Satrawada, Rachel Koenigsberg

THE DATASET

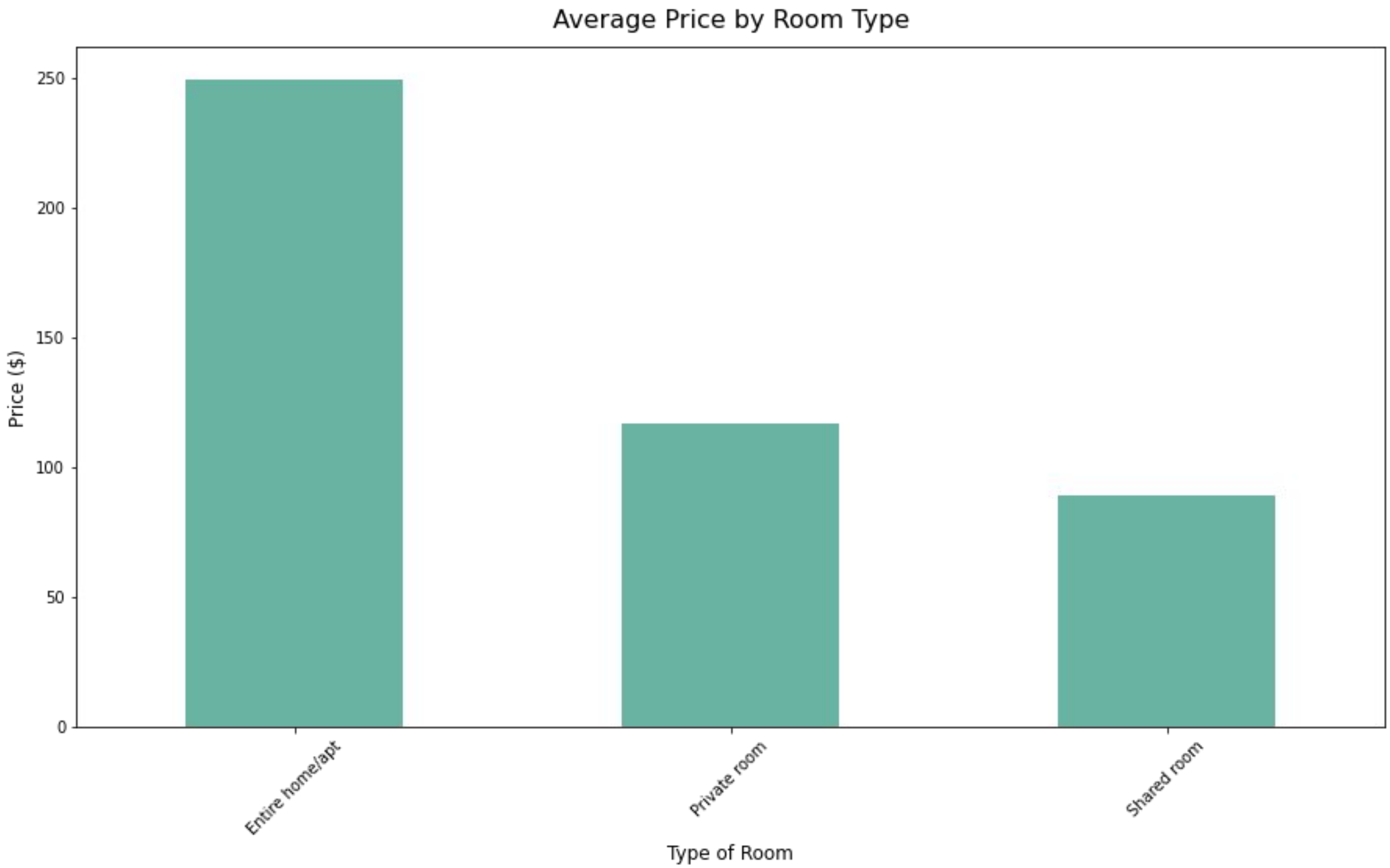
- Kaggle, ~50,000 observations of NYC Airbnb listings in 2019
- 16 variables
  - ID variables: listing ID, host ID, listing name, host name
  - Coordinates
  - Neighborhood (major and minor)
  - Listing details

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_mont
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.2
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.3
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	Na
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.6
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.1

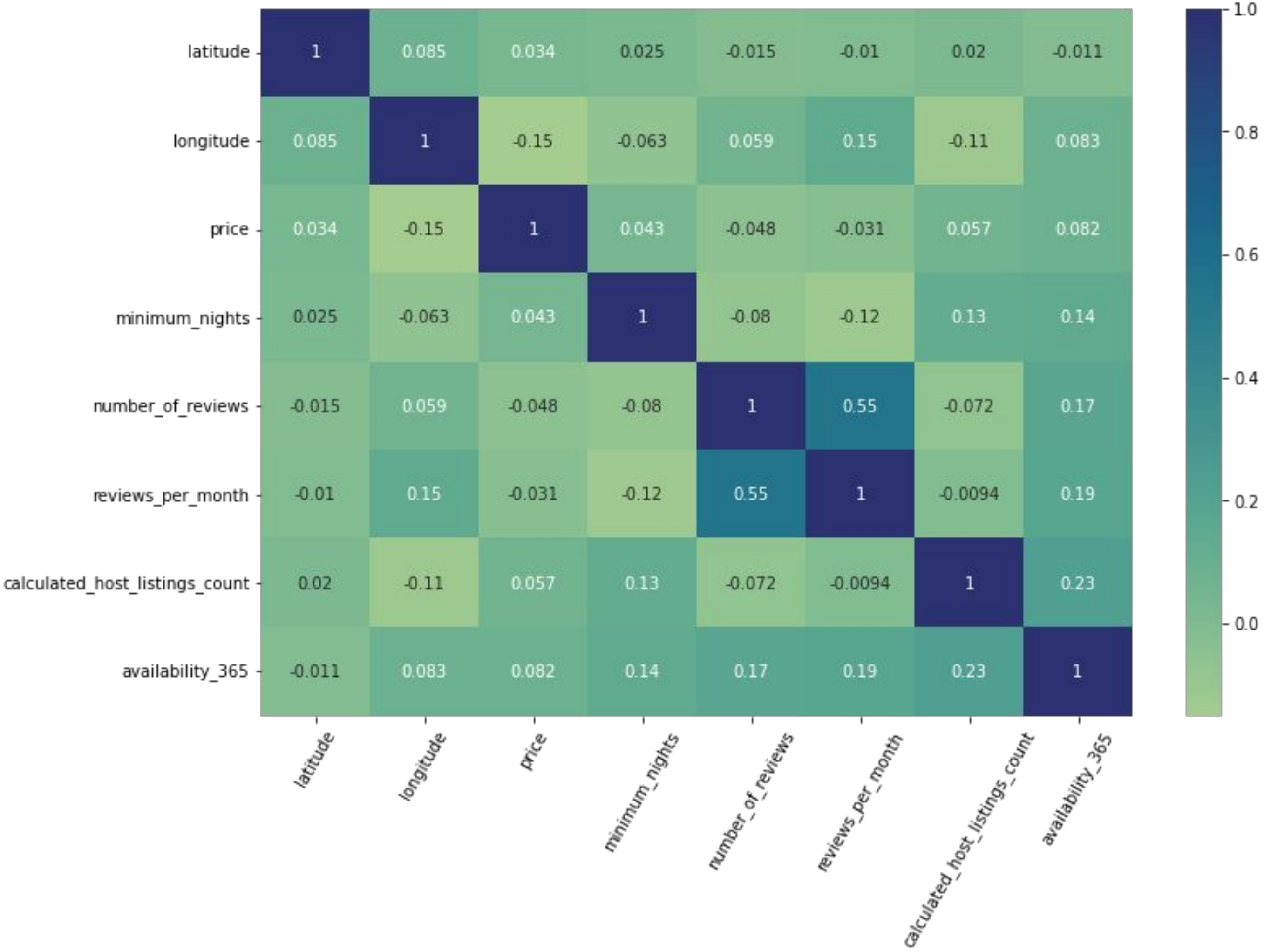

Total Observations by Neighbourhood Group
Total Observations
0
5000
10000
15000
20000
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Neighbourhood Group


Average Price by Neighbourhood Group
200
175
150
125
100
75
50
25
0
Average Price ($)
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Neighbourhood Group


Average Price by Manhattan Neighbourhood
Price ($)
0
100
200
300
400
500
Battery Park City
Chelsea
Chinatown
Civic Center
East Harlem
East Village
Financial District
Flatiron District
Gramercy
Greenwich Village
Harlem
Hell's Kitchen
Inwood
Kips Bay
Little Italy
Lower East Side
Marble Hill
Midtown
Morningside Heights
Murray Hill
NoHo
Nolita
Roosevelt Island
SoHo
Stuyvesant Town
Theater District
Tribeca
Two Bridges
Upper East Side
Upper West Side
Washington Heights
West Village
Neighbourhood


Average Price by Room Type
Price ($)
250
200
150
100
50
0
Entire home/apt
Private room
Shared room
Type of Room

Correlation Heatmap



METHODOLOGY

Feature Extraction

- Dropped irrelevant features (ex: host id)
- Queried for listings in Manhattan
- Converted categorical features into numerical variables (neighborhood)

Linear Regression

- We used a regression model because our target value is continuous
- 80-20 Train/Test Split
- Predictive Analysis: using features from past Airbnb data to predict future listings

Cross Validation

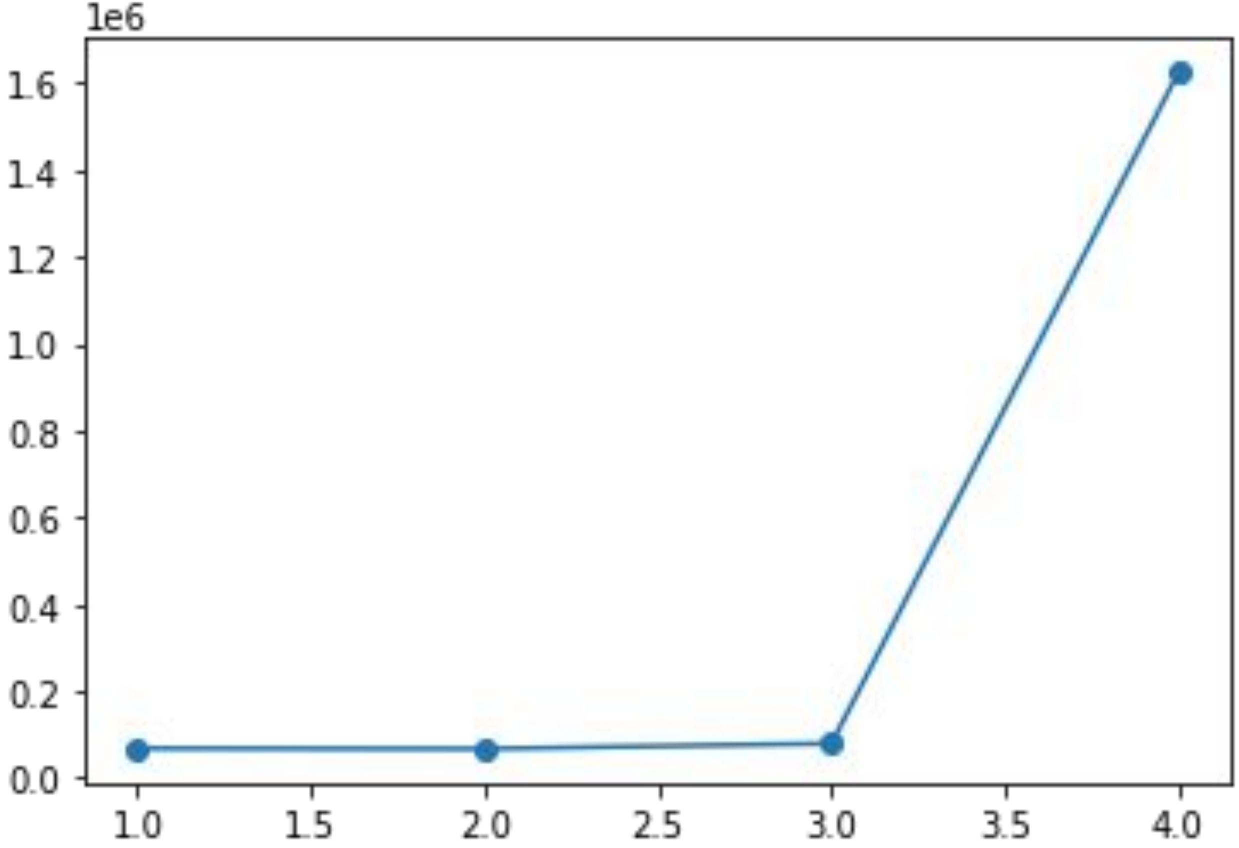
- Used to find most optimal degree for polynomial basis function
- Average accuracy of 10 ‘experiments’
  - 90/10 split
- Polynomial basis with degree 2 had the highest accuracy

```
split_size = int(len(X_train) * 0.1)
acc = []
for degree in range(1, 5):
  avgAcc = []
  print(degree)
  for split in range(split_size * 9, -1, split_size * -1):
    x_tr = X_train[0:split]
    y_tr = Y_train[0:split]
    x_ts = X_train[split:split + split_size]
    y_ts = Y_train[split:split + split_size]
    x_tr = np.vstack((x_tr, X_train[split + split_size:]))
    y_tr = pd.concat([y_tr, Y_train[split + split_size:]])
    poly_model = make_pipeline(PolynomialFeatures(degree, include_bias=False),LinearRegression(fit_intercept=True))
    poly_model.fit(x_tr, y_tr)
    y_pred = poly_model.predict(x_ts)
    avgAcc.append(mean_squared_error(y_ts, y_pred))
  acc.append(np.mean(avgAcc))
print(acc)
```

Prediction

Implemented a pipeline:

- takes the input data
- adds a polynomial basis of degree 2 (calculated optimal degree)
- fits a linear regression curve with intercept



Distances: [67811.39872231337, 66928.02405712218, 79319.32421261075, 1625535.2531644276]

MAJOR OBSTACLES

Two Major Hurdles

1. Determining a degree for our polynomial basis
   a. Implement cross-validation
2. Too many features in pipeline = CRASH
   a. reduce variables, increase the precision of model -> focus on a specific neighborhood group

RESULTS

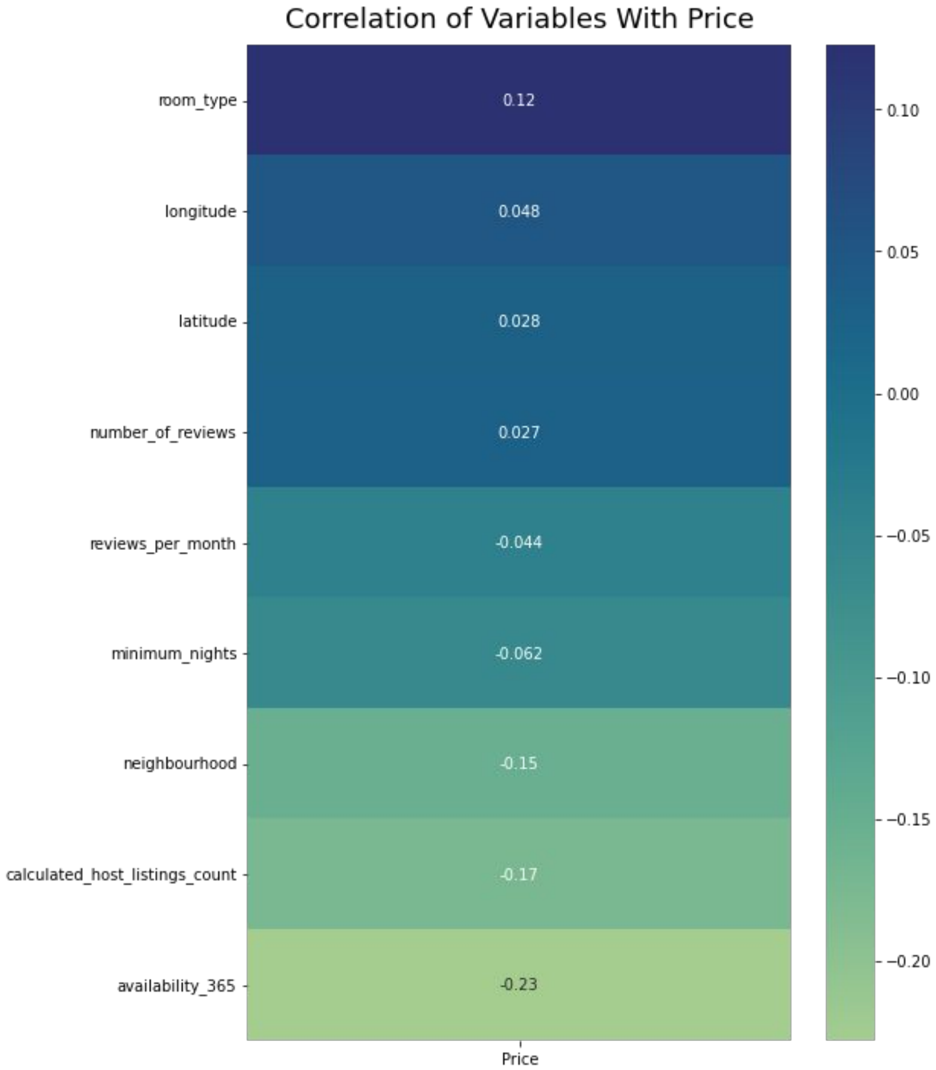
Distance and R^2

Average distance between the predicted and the actual price for each listing: ~$98.85

R-squared value: ~0.62

Difference b/w predicted & actual Airbnb prices small, and not always too high/ too low

WHAT WE LEARNED


Correlation of Variables With Price
room_type
longitude
latitude
number_of_reviews
reviews_per_month
minimum_nights
neighbourhood
calculated_host_listings_count
availability_365
0.12
0.048
0.028
0.027
-0.044
-0.062
-0.15
-0.17
-0.23
Price
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20