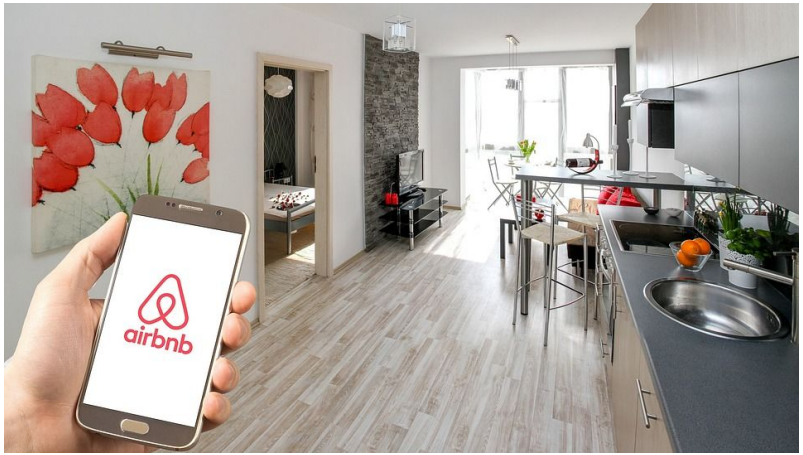
The Cost of Staying in the Big Apple

Developing a price recommender for Airbnb listings in New York

Group 7
Sidd Chauhan | Lucy Hwang | Shakirah Oladokun | Patricia Schutter | Katelyn Vincent | Yiwei Zhou

Agenda

- Project Goals
- The Dataset
- Dataset Challenges
- Model Comparisons
- Insights & Recommendations



Project Goals

Business Initiative: Investigate the feasibility of developing an accurate price recommendation model for Airbnb.

What We Did: Analyzed NYC Airbnb booking data and evaluated the application of various machine learning algorithms to determine if a model could be constructed that would achieve the desired accuracy.

The Dataset

Airbnb booking data for NYC from 2011-2019

Approximately 49k observations of 16 attributes that capture:

Listing Information	• listing name • price • ID • minimum nights • room type • number of days available
Listing Reviews	• number of reviews • last review • reviews per month
Host	• host name • host ID • count of total listings
Location	• neighborhood • neighborhood group/borough • latitude • longitude

Dataset Challenges

Missing Data

Removed rows with missing values (remainder = 38k observations)

Irrelevant Features

Eliminate columns/features that are not useful for predicting price (host name and ID, listing name & ID)

Categorical Data

Convert last review date to days since last review, create dummy variables for neighborhood group & room type

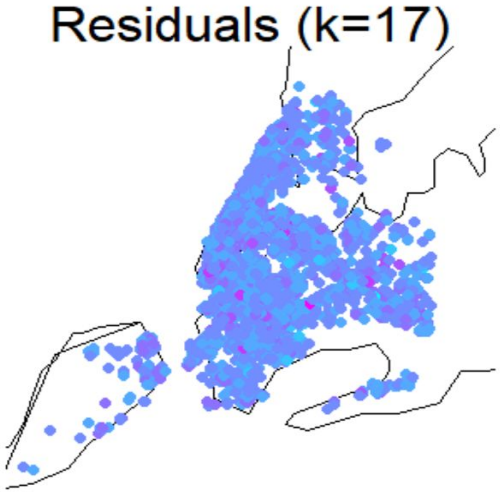
Model Comparisons

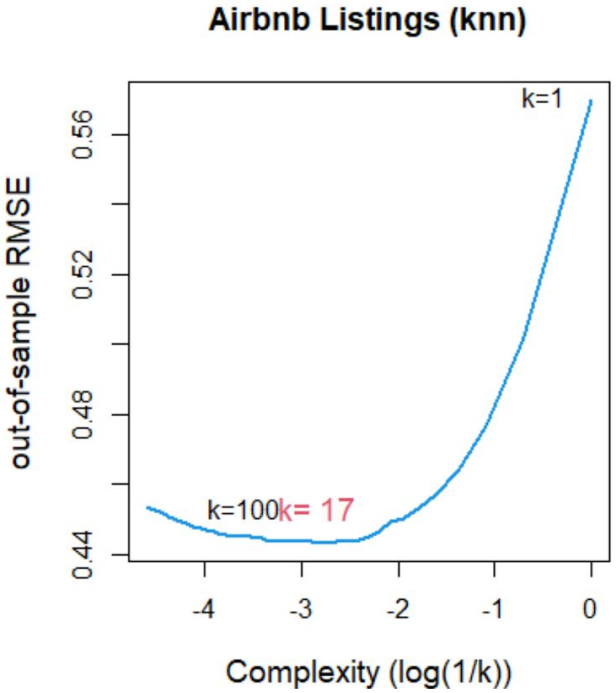
KNN

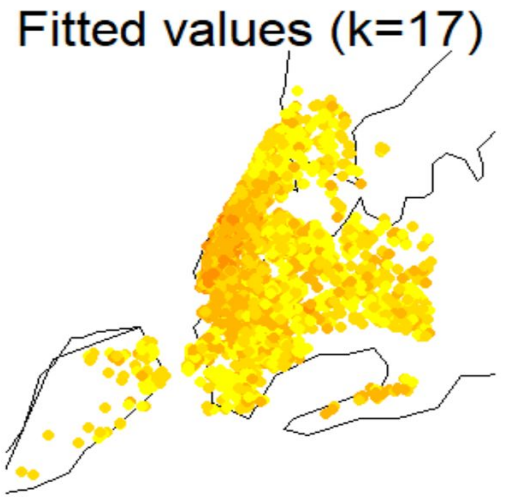
KNN Model Comparisons

RMSE Value	Variables	K-value
0.4432177	All	17
0.4517004	latitude, longitude, private room, shared room, bronx, brooklyn, queens, statisland, min nights, rev per month	32
0.4519254	latitude, longitude, private room, shared room, bronx, brooklyn, queens, statisland, min nights	51
0.4533745	latitude, longitude, private room, shared room	41
0.4543689	latitude, longitude, private room, shared room, bronx, brooklyn, queens, statisland	37
0.4743354	latitude, longitude, private room	69
0.5697155	latitude, longitude, bronx, brooklyn, queens, statisland	38
0.5702855	latitude, longitude, reviews per month	24
0.5709886	latitude, longitude	26
0.5791968	longitude, latitude, calculated host listings count	33
0.5928538	latitude, longitude, number of reviews	59
0.5962759	latitude, longitude, min nights	31
0.6003741	latitude, longitude, availability	35

(KNN) Price ~ All Independent Variables





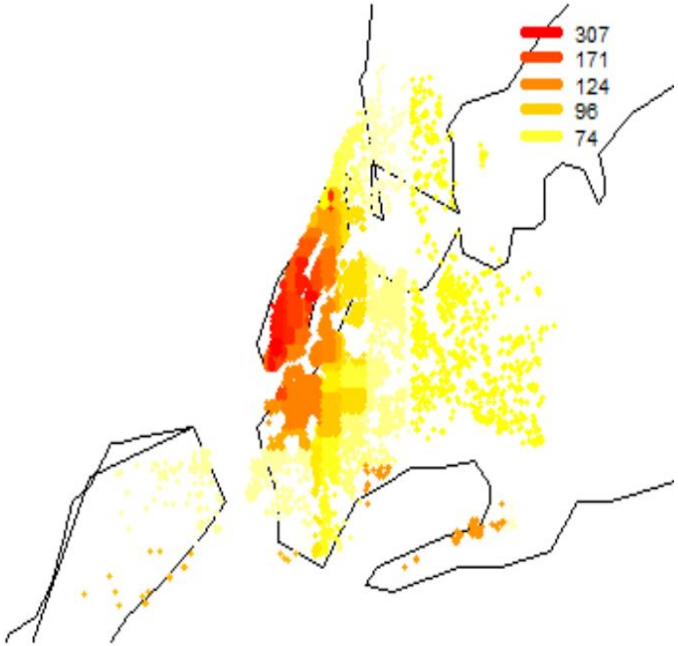


RMSE 0.443

Trees

Decision Tree

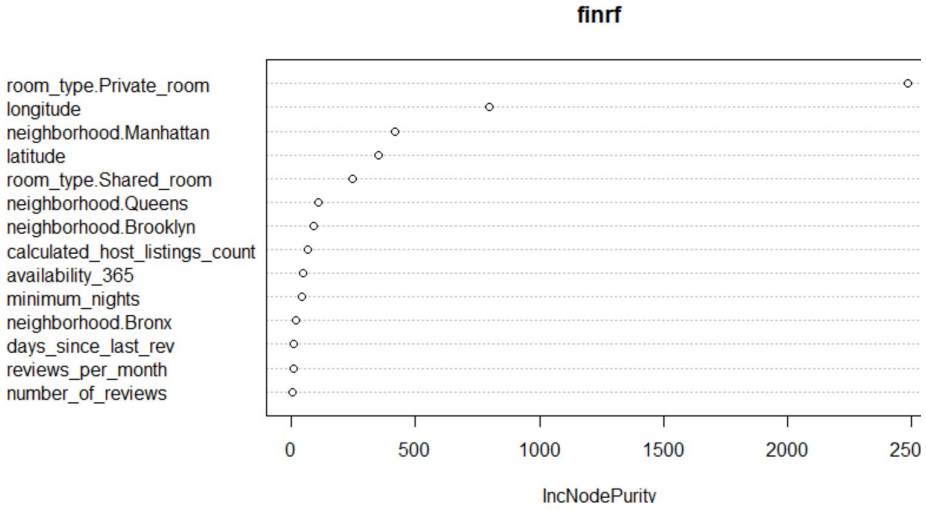
Price ~ Lat/Long



Borough	Avg. Listing Price
Manhattan	$197
Brooklyn	$124
Staten Island	$115
Queens	$100
Bronx	$87

Random Forest

Variable Importance



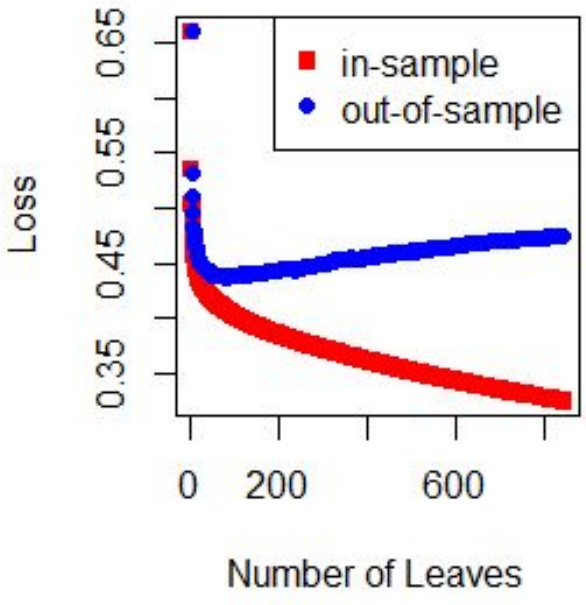
Best RMSE: 0.4912794

Gradient Boosting Machine

Relative Influence by Variable

	var	rel.inf
1	room_type.Private_room	24.92487402
2	longitude	17.86960871
3	latitude	16.01912747
4	days_since_last_rev	8.61026222
5	availability_365	7.99818630
6	reviews_per_month	6.27414159
7	room_type.Shared_room	5.30802449
8	number_of_reviews	4.09182750
9	minimum_nights	3.98320014
10	calculated_host_listings_count	2.27989556
11	neighborhood.Manhattan	2.24623393
12	neighborhood.Brooklyn	0.28265416
13	neighborhood.Queens	0.07151644
14	neighborhood.Bronx	0.04044744

Best RMSE: 0.438808



Stepwise

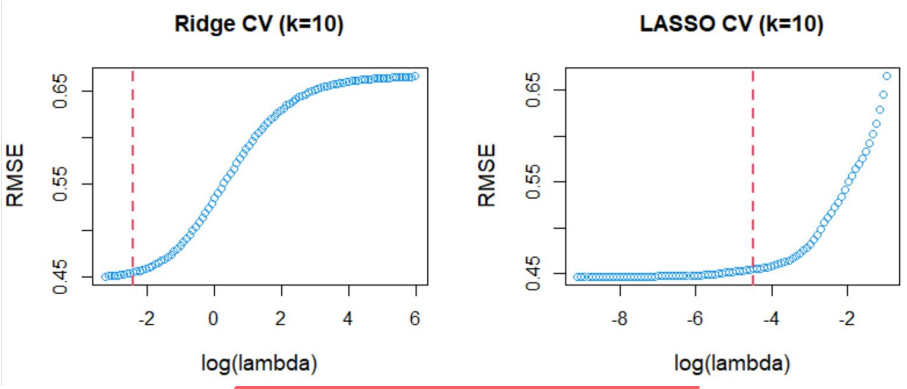
Stepwise Regression

Price Model

Price ~ room_type_Entire.home.apt + neighbourhood_group_Manhattan + availability_365 + longitude + neighbourhood_Lower.East.Side + neighbourhood_SoHo + neighbourhood_Tribeca + minimum_nights

AIC=45662.78

MSE = 2.964900165



Ridge Lambda: 0.08208

LASSO Lambda: 0.01005

Conclusion

Model Comparisons

Algorithm	Best RMSE
GBM	0.439
KNN	0.443
Stepwise & Lasso/Ridge Regression	0.45
Random Forest	0.491

Insights & Recommendations

Business Initiative: Investigate the feasibility of developing an accurate price recommendation model using Airbnb booking data.

Insight	Recommendation
Room type has the biggest impact on Airbnb pricing in NYC - specifically, private spaces	Call attention to the higher prices Airbnb hosts can command by offering private spaces instead of shared rooms
Location is also an important predictor when it comes to pricing - for example, a listing in Manhattan is associated with an increase in price	Share location and listing price relationships with higher listing count Airbnb hosts to encourage investing in Manhattan-based units
Reviews impact pricing as well, and hosts with more (and more recent) reviews tend to price their listings higher	Highlight the importance of reviews, and offer suggestions on how to engage with guests

Thank you!

Appendix

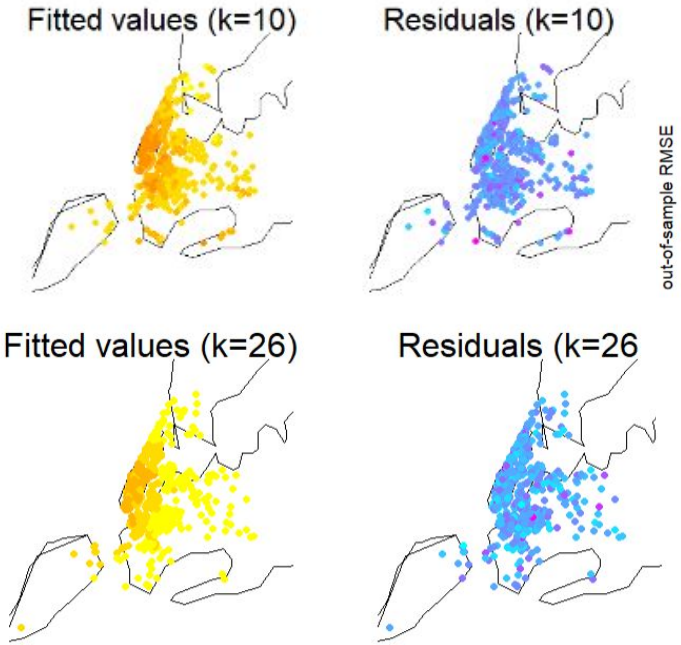
Appendix

- KNN
- Decision Trees
- Lasso and Ridge Tests
- Improvements
- Avg prices for listings

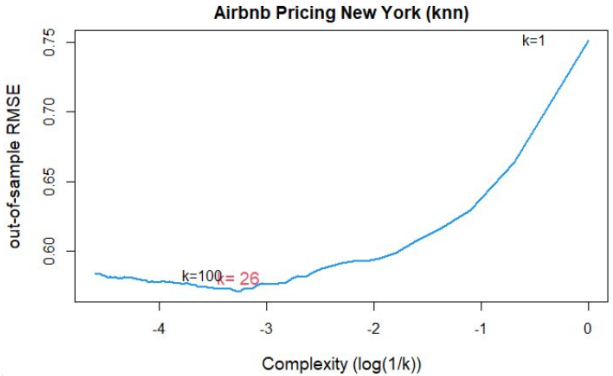
KNN Test

Price ~ Long/Lat

KNN



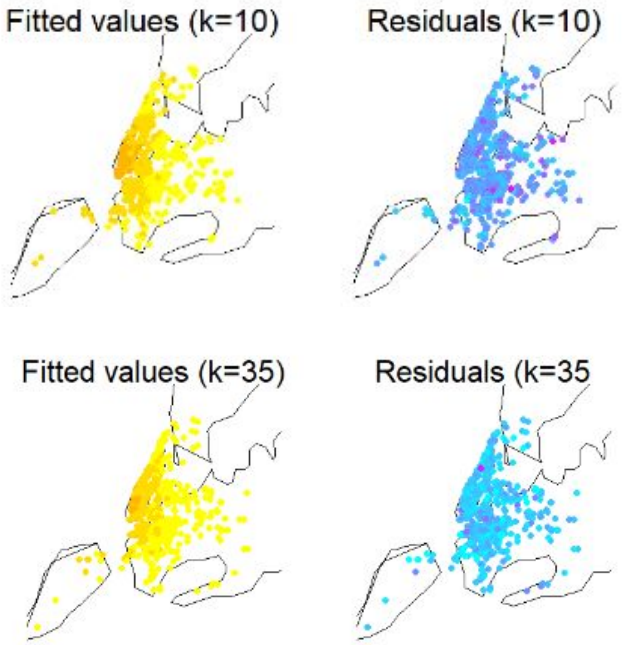
RMSE vs Complexity



RMSE 0.571

Price ~ Long/Lat, Availability 365

KNN

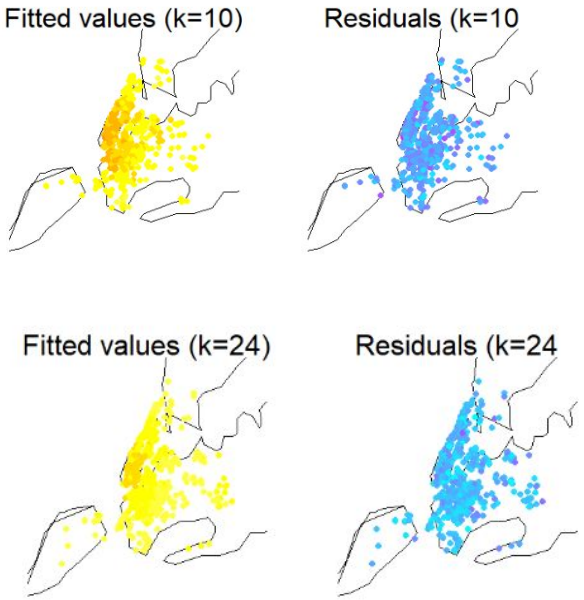


RMSE vs Complexity

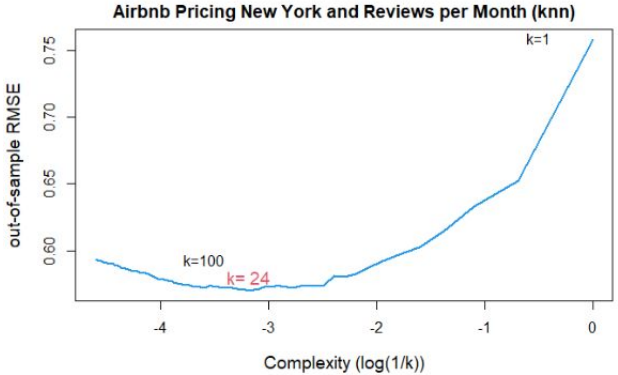
RMSE 0.600

Price ~ Long/Lat, Reviews per Month

KNN



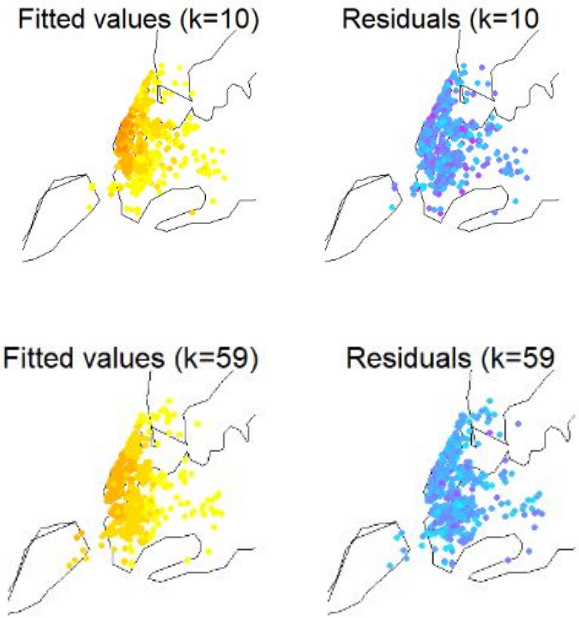
RMSE vs Complexity



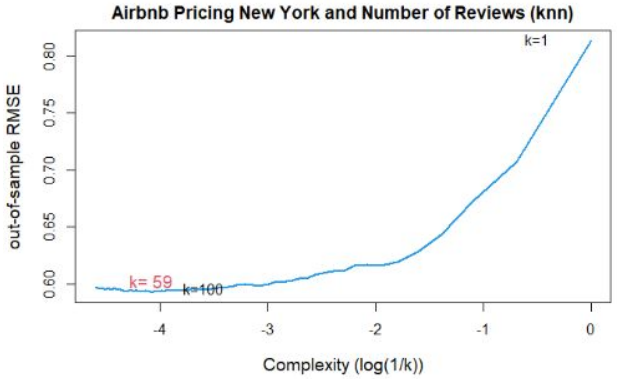
RMSE 0.570

Price ~ Long/Lat, Number of Reviews

KNN



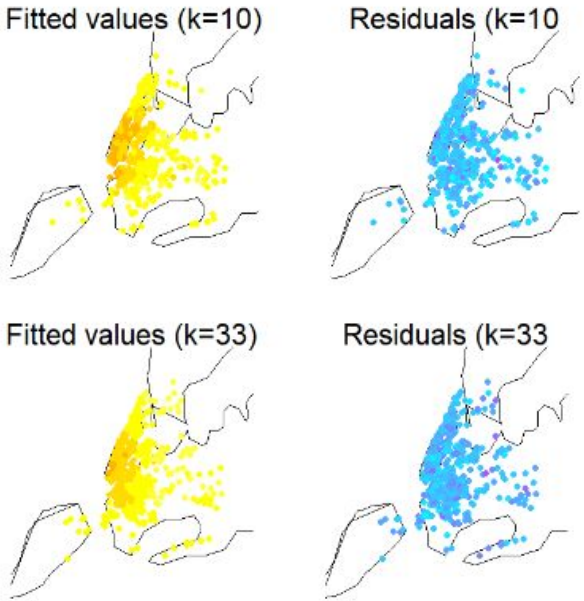
RMSE vs Complexity



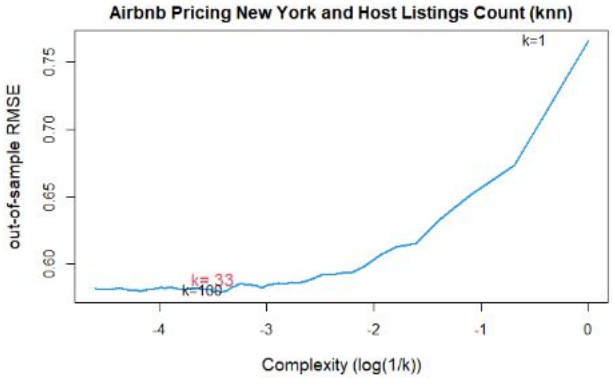
RMSE 0.593

Price ~ Long/Lat, Calculated Host Listings

KNN

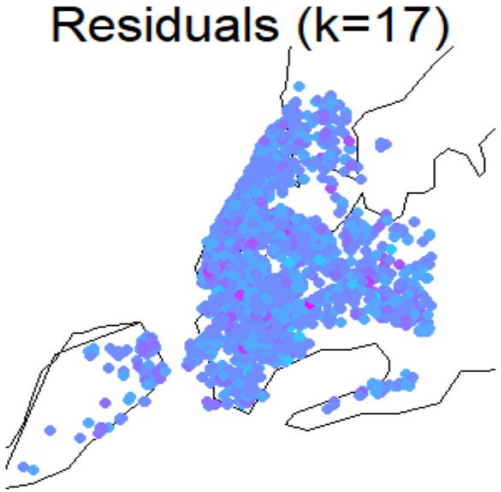


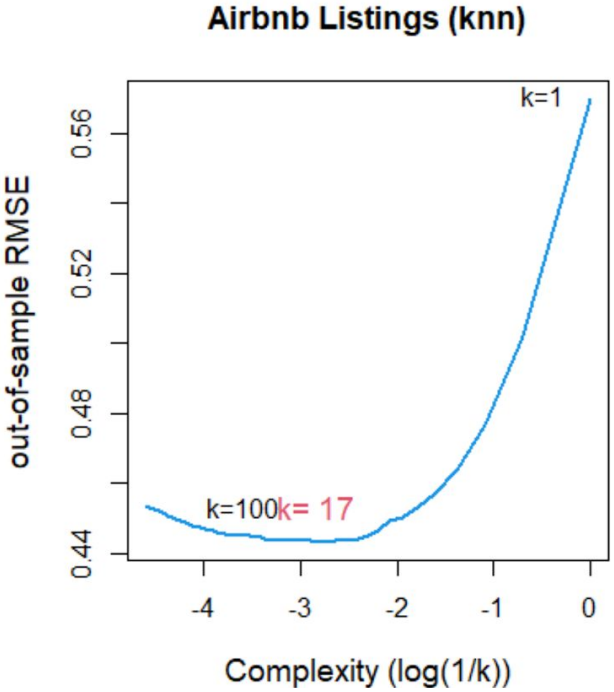
RMSE vs Complexity

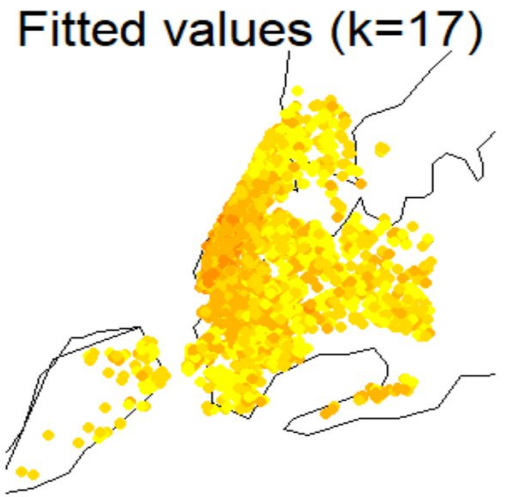


RMSE 0.579

(KNN) Price ~ All Independent Variables



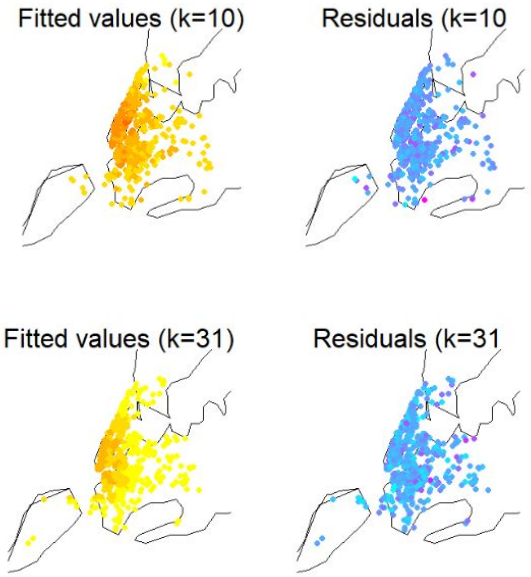




RMSE 0.443

Price ~ Long/Lat, Minimum Nights

KNN



K-fold

RMSE 0.579

KNN Model Comparisons

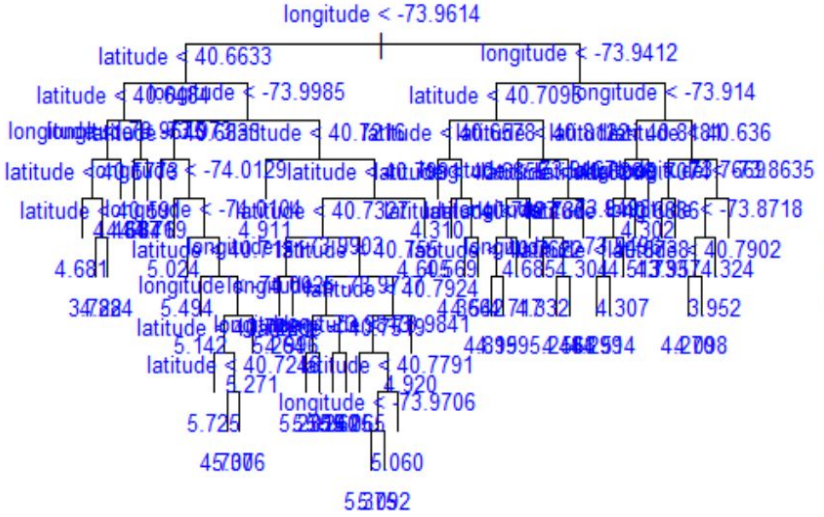
RMSE Value	Response	Predictor(s)	K
0.4432177	Price	All	17
0.5702854548	Price	longitude, latitude. reviews per month	24
0.5709886163	Price	Latitude, Longitude	26
0.5791967714	Price	longitude, latitude, calculated host listings count	33
0.5928537762	Price	longitude, latitude. number of reviews	59
0.5962759428	Price	longitude, latitude. min nights	31
0.6003741334	Price	longitude, latitude. availability	35

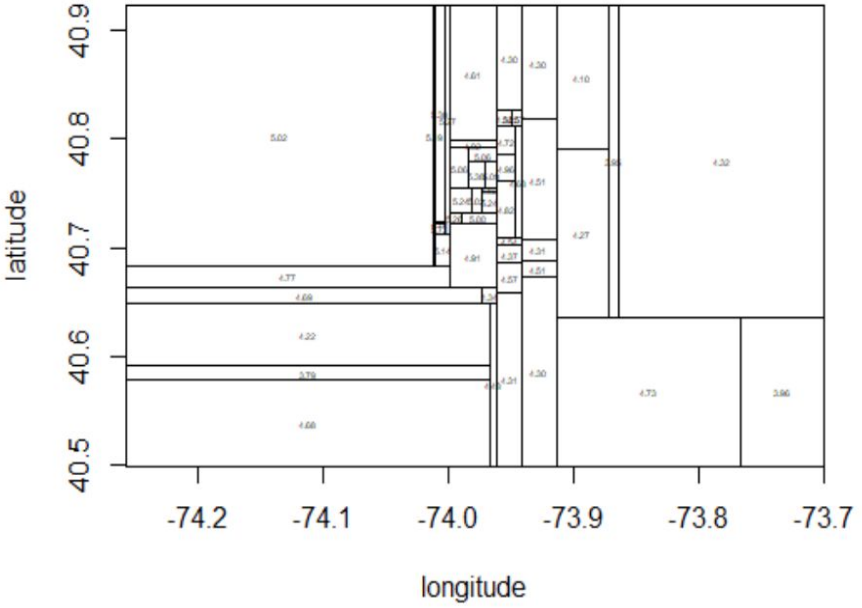
Best Algorithms and their variables

Algorithm	Independent Variables (Predictors)
KNN	All
Random Forest	room_type.Private_room, longitude, latitude, days_since_last_rev, availability_365, reviews_per_month, room_type.Shared_room, number_of_reviews, minimum_nights, calculated_host_listings_count, neighborhood.Manhattan, neighborhood.Brooklyn, neighborhood.Queens, neighborhood.Bronx
GBM	room_type.Private_room, longitude, latitude, days_since_last_rev, availability_365, reviews_per_month, room_type.Shared_room, number_of_reviews, minimum_nights, calculated_host_listings_count, neighborhood.Manhattan, neighborhood.Brooklyn, neighborhood.Queens, neighborhood.Bronx
Stepwise Regression	room_type_Entire.home.apt, neighbourhood_group_Manhattan, availability_365, longitude, neighbourhood_Lower.East.Side, neighbourhood_SoHo, neighbourhood_Tribeca, minimum_nights

Decision Trees

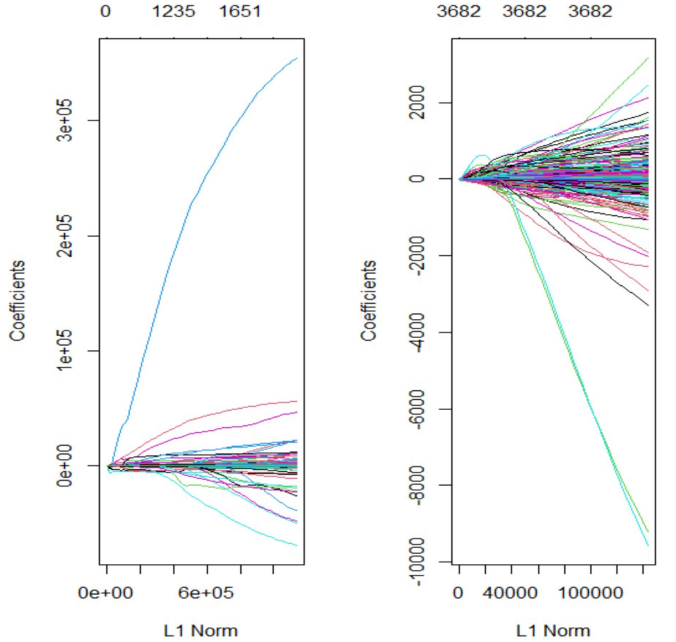
(Decision Tree)Price~Lat/Long

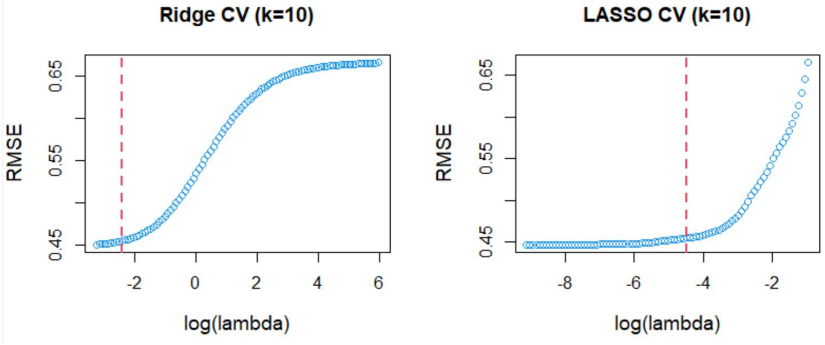




Lasso & Ridge

Lasso & Ridge Graphs





Improvements

Improvements to be made

We included substantial amount of data columns after converting categorical variables to dummies, turning the process of data analysis burdensome.

This might also cause potential complexity in the data. We might be using less categorical variables for further research.

Avg Prices for Listing

	Avg Price			
Borough	Entire home/apt	Private room	Shared room	Overall
Manhattan	$249	$117	$89	$197
Brooklyn	$178	$77	$51	$124
Staten Island	$174	$62	$57	$115
Queens	$147	$72	$69	$100
Bronx	$128	$67	$60	$87
Overall	**$212**	**$90**	**$70**	**$153**

A deeper dive into the similarities of listings at various prices also reveals some interesting insights. The majority of listings under $150 are private rooms (64%) followed by entire units (32%), whereas the overwhelming majority (89%) of higher-priced listings are for entire units.

Row Labels	Price < $150	Price $150+	Overall
Entire home/apt	**32%**	**89%**	**52%**
Bronx	3%	1%	1%
Brooklyn	48%	31%	38%
Manhattan	35%	64%	52%
Queens	13%	5%	8%
Staten Island	1%	0%	1%
Private room	**64%**	**11%**	**46%**
Bronx	3%	1%	3%
Brooklyn	47%	24%	45%
Manhattan	33%	69%	36%
Queens	16%	6%	15%
Staten Island	1%	0%	1%
Shared room	**3%**	**0%**	**2%**
Bronx	5%	4%	5%
Brooklyn	37%	23%	36%
Manhattan	40%	61%	41%
Queens	17%	11%	17%
Staten Island	1%	1%	1%
Overall	**100%**	**100%**	**100%**