NYC Airbnb Price Prediction
Capstone 1 - Final Report

Main Objective

The dataset contained detailed information for Airbnb listings in New York City, NY. There were about 50k listings with several different features for each listing, including the location, price, host details, etc. The main objective of the project is to build a price prediction model for NYC Airbnb listings based on its different characteristics. A slide deck for overview of the project can be found here.

Business Value

Airbnb is a company that operates an online marketplace and hospitality service which is accessible via its websites and mobile apps. Members can use the service to arrange or offer lodging, primarily homestays, or tourism experiences. This project is to help us analyze which characteristics contribute to higher prices and help hosts increase their revenues. This will be useful for Airbnb as a company, hosts located in NYC as well as customers who are planning on booking accomodations in the city.

Dataset Details

The New York City Airbnb listings data was obtained from Inside Airbnb, which is sourced from publicly available information from the Airbnb website. The dataset contained 49056 listings and 96 columns with different characteristics of each listing. The data types included float, integer and object data types.
Description Provided:

> "*Inside Airbnb is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world.*
> *By analyzing publicly available information about a city's Airbnb's listings, Inside Airbnb provides filters and key metrics so you can see how Airbnb is being used to compete with the residential housing market.*"

Data Wrangling:

The data wrangling notebook is available at: Data Exploration and Wrangling Notebook

The following steps were implemented to clean up the dataset:

- Preliminary steps performed on the dataset as a whole:
  - Removed all columns with more than 50% null values.
  - Dropped columns with redundancy and non-useful information.
- Once general clean-up steps were performed, we started weeding through the remaining columns and focused on them individually:
  - These included IDs, descriptions, listing counts columns.
  - Removing outliers:
    - Listings outside NY were dropped.
    - Hosts with no prior data/listings were dropped.
- After cleanup, we were left with 48353 listings and 55 rows. The next step was to fill up any null values and clean up each of the data columns.
  - For columns with information about the listing - access, transit, interaction, etc. we converted them to simpler columns which stated whether the information was provided or not provided.
  - Descriptive text columns were combined into one.

- Host location was classified into 3 categories - in NY, in US but not in NY and outside US.
- Columns with percentage were converted to ratio columns with float values.
- Missing values:
  - To fill in the missing zip codes, neighbourhood_cleansed column was used as reference.
  - Other columns with missing values, including bathroom, bedrooms, security deposit, cleaning fee, and all different rating columns which had null values were filled in with 0.

Finally, we had a clean data set with no missing values and rid of any unwanted or redundant columns. The data has also been simplified for the further EDA and eventual modelling for price prediction.

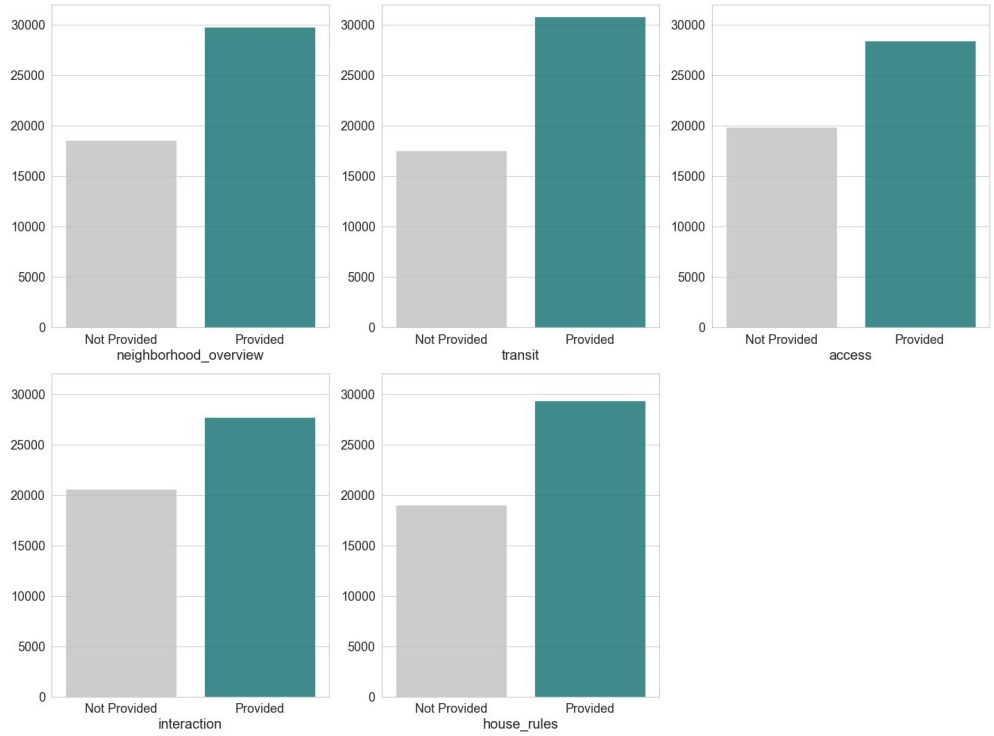
Exploratory Data Analysis:

After cleaning and wrangling the data, we were left with 55 columns and ~48k listings. The most crucial step is to visualize the data. That is the first step towards understanding the data and proceeding with further analysis. The EDA notebook is available at: [Exploratory Data Analysis Notebook]

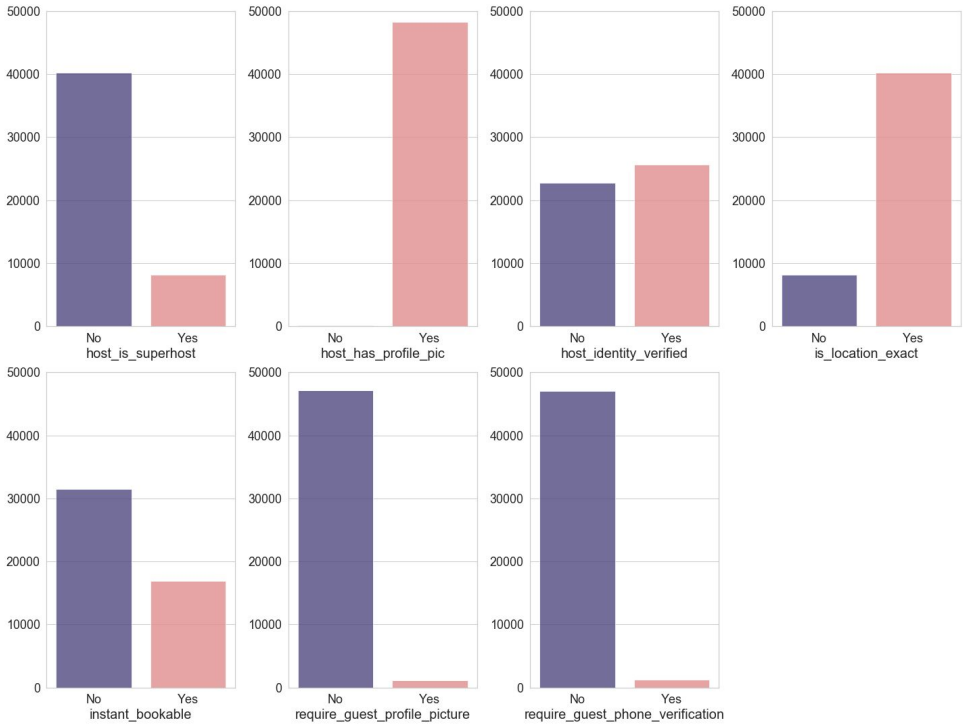
The following steps were performed for EDA:

- We started with univariate analysis to understand underlying patterns in each variable and visualize the distributions for the numeric variables.
  - Count Plots were used to visualize distribution of categorical data.
  - Violin Plots and box plots were used to visualize the distribution of numerical features (numerical as well as discreet).

Count plots for the textual features - Provided vs Not Provided



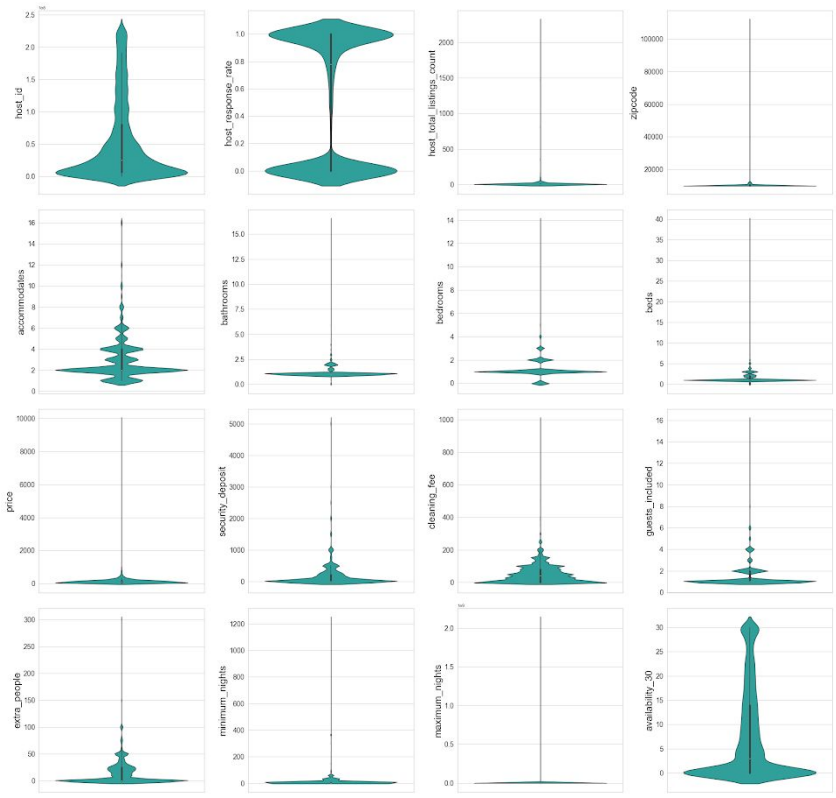
Count Plots for 'Yes' vs 'No' features



Based on the Yes/No count plots above, we can conclude the following:

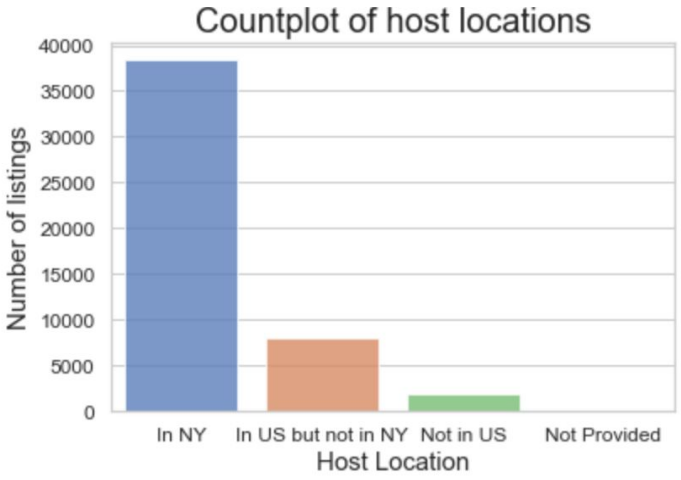
- The host_has_profile_pic has almost all values as 'Yes'. This could suggest that having a profile picture is mandatory for hosts on Airbnb to create listings.
- Also, in most cases the guest profile picture of guest phone verification are not required. Airbnb probably doesn't want their customers feel they have to follow some mandatory rules and thus keeps these requirements lenient for its customers.

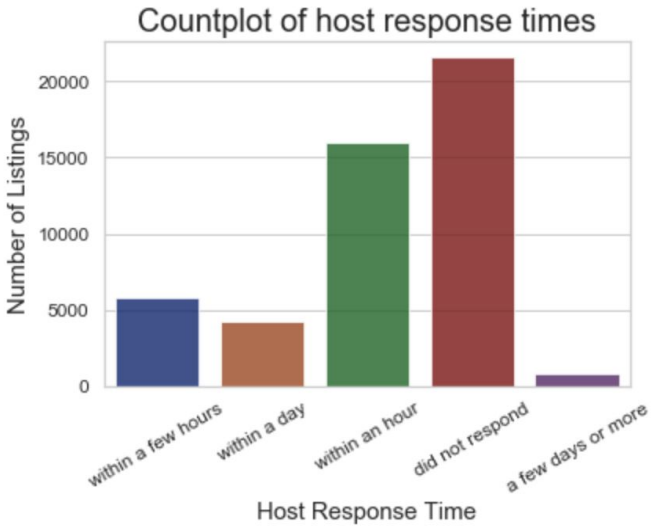
Violin plots to visualize distributions of few numeric features



Several host properties are also important features for each listing. The following count plots provide distribution of different features of the hosts.

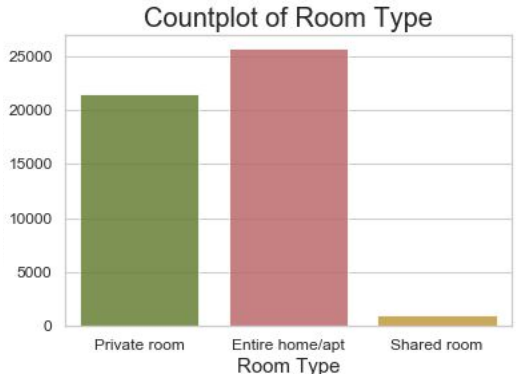
- The hosts for most of the listings are located in NY. It's more favorable to have the host in the same city as the property listed to keep the logistics uncomplicated.
- A lot of hosts did not respond, which is a bad indicator of host response rates. But, this could be also because we filled in the missing values with ‘did not respond’ and that could be a wrong assumption to make.

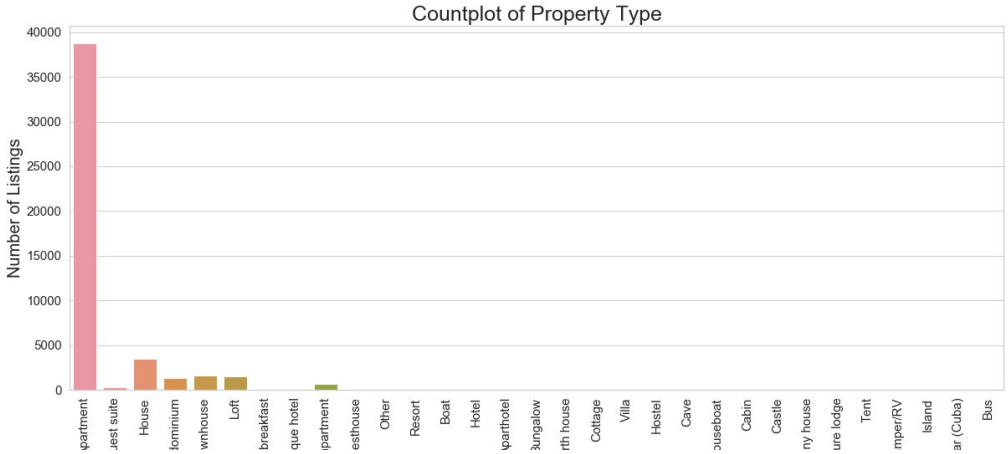




The properties of the listings including the room type, and property type are are also equally important - based on the plots below we can conclude:

- In a majority of listings, the property is an entire home/apartment or a private room.
- Most of the properties are apartments.

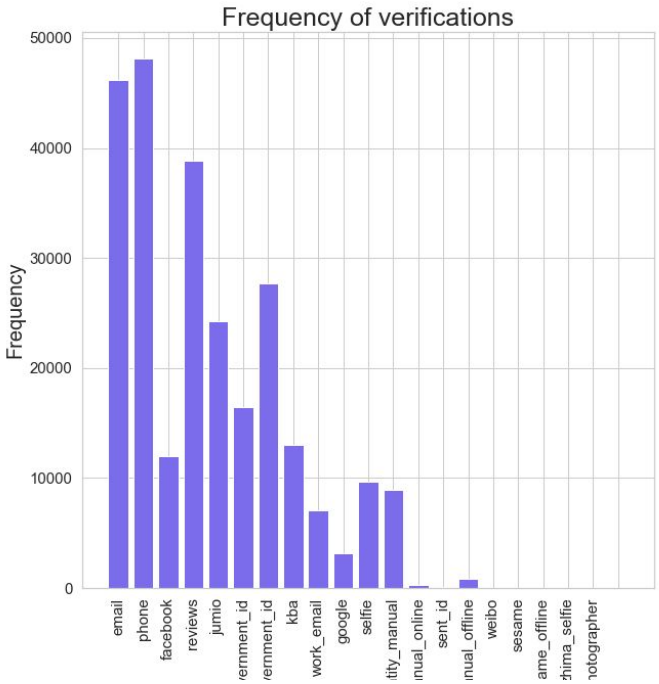




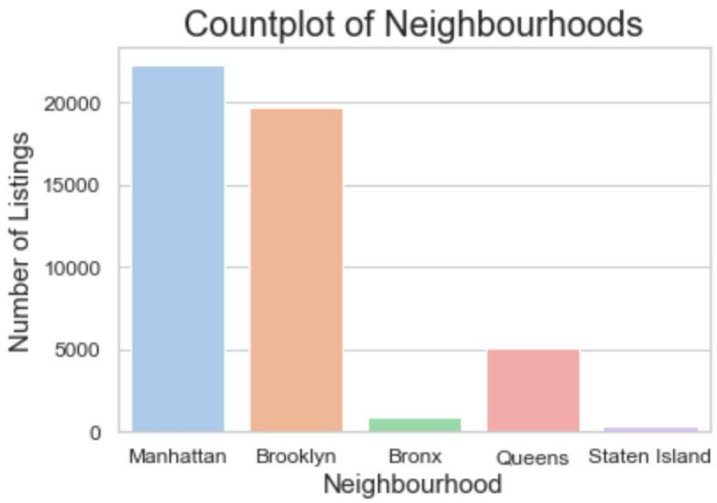
We also looked at what amenities were being provided in each of the listings, which had a diverse kinds of amenities available. The top amenities, i.e the amenities most frequently provided were:

- ★ Wifi
- ★ Heating
- ★ Kitchen
- ★ Smoke Detector
- ★ Air conditioning
- ★ Hangers
- ★ Television
- ★ Shampoo
- ★ Hairdryer
- ★ Workspace

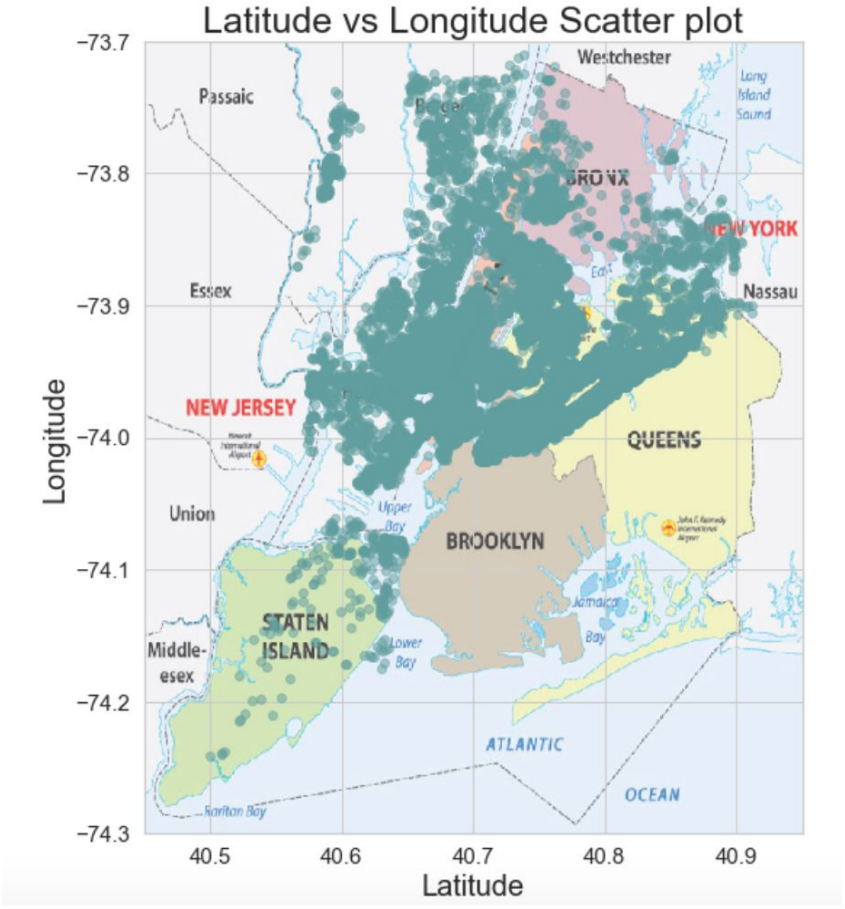
Host verifications were also analyzed to determine which were the most popular and favored forms of verification for hosts. Email and phone seem to be the most widely used forms of verification which was expected. It is possible that each host needs to provide email and phone verification, other verifications are optional.



Location of the listing is also a very important factor, especially in a city like NYC. The city has been divided into five main borough and the countplot below indicates that the top three boroughs with the most listing are - Manhattan, Brooklyn and Queens.



The following plot shows the map of NYC with the five boroughs indicated and overlays the scatter plot of longitude vs latitude on the map to show that the listings are clustered based on the boroughs and we see high number of listings in certain areas.

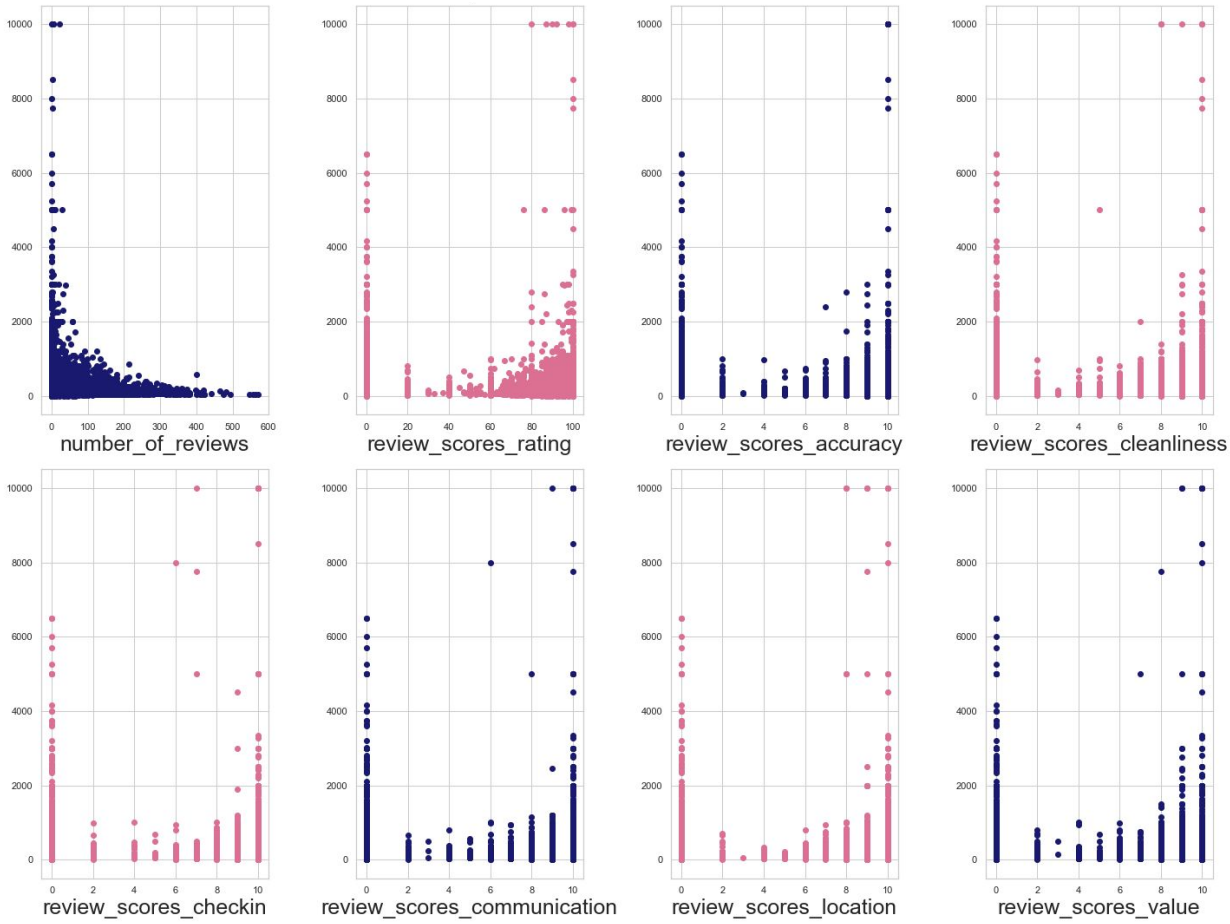


The neighborhoods with highest and lowest prices are as follows:

Highest Priced Neighborhoods	Lowest Priced Neighborhoods
➢ Fort Wadsworth	➢ Little Neck
➢ Woodrow	➢ Concord

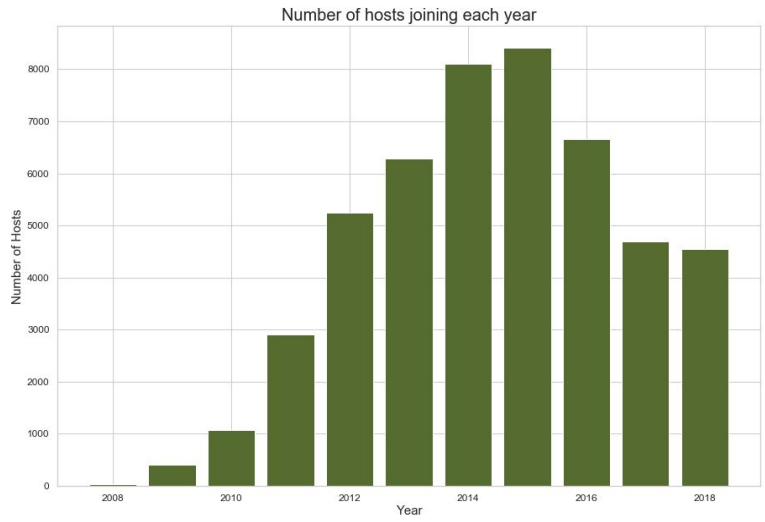
- Sea Gate
- College Point
- Tribeca
- Soundview
- Bull's Head
- Hunts Point

Scatter plot of Review scores vs Price



The number of reviews is a continuous numerical variable while all the review scores are discreet integer values from 1-10. In this case, it is important to note that whenever review score wasn't provided, it was filled in with 0. So, 0 includes count of all missing values as well.

Next, it was interesting to see how many hosts joined over a span of ten years. Airbnb was established in August 2008.That explains the lower number of hosts signing up in the first few years while Airbnb was getting more established. We notice a sudden dip in the number of hosts signing up after 2015, which could be attibuted to the fact that the market was already saturated by that point.



Plot of hosts signing up per month in each year

Visual EDA provided us meaningful insights into the distribution of different features - numerical as well as categorical. We also saw patterns between different features with bivariate and multivariate analysis.

Inferential Statistics:

The Inferential statistics notebook is available at: [Inferential Statistics Notebook]

Inferential statistics will help us dive deeper into the trends of the data and statistically draw conclusions. We performed different hypothesis testing on different features in the dataset to find statistical relevance. The alpha value fixed at 0.5 for all the tests and the p-values were calculated. The following table provides a summary of the tests performed and the outcomes of those.

Hypotheses	**P-value**	**Final Verdict**
$\mathbf{H_0}$ → Mean price of Manhattan listings = Mean price of Brooklyn listings $\mathbf{H_a}$ → Mean price of Manhattan listings ≠ Mean price of Brooklyn listings	1.959e-227	The null hypothesis was rejected.indicating that we have enough evidence to conclude that the mean price of Manhattan listings are statistically different from mean price of Brooklyn listings. Manhattan Airbnbs seem to be in general pricier than Brooklyn ones.
$\mathbf{H_0}$ → Mean communication score when host responds within hours - Mean communication score when host takes longer = 0 $\mathbf{H_a}$ → Mean communication score when host responds within hours - Mean communication score when host takes longer > 0	0.5068	The null hypothesis could not be rejected, so we concluded that the time taken by the host to respond does not have an effect on the review communication scores.
$\mathbf{H_0}$ → Mean price of listings with <=13 amenities = Mean price of listings with >=25 listings $\mathbf{H_a}$ → Mean price of listings with <=13 amenities ≠ Mean price of listings with >=25 listings	1.0207e-44	The null hypothesis can be rejected, indicating that we have enough evidence to conclude that the number of amenities does affect the price of the listing. Listing with >=25 amenities seems to be more expensive than listings with <=13 amenities.
$\mathbf{H_0}$ → Mean price of listings rating <=78 = Mean price of listings with rating >=99 $\mathbf{H_a}$ → Mean price of listings rating <=78 ≠ Mean price of listings with rating >=99	7.238e-09	The null hypothesis can be rejected, indicating that we have enough evidence to conclude that the review scores rating does affect the price of the listing. Listing with <=78 rating score seems to be more expensive than listings with >=99 rating score.
$\mathbf{H_0}$ → Mean reviews per month for host 1 (107434423) = Mean reviews per month for host 2(30283594) $\mathbf{H_a}$ → Mean reviews per month for host 1 (107434423) ≠ Mean reviews per month for host 2(30283594)	0.0019	The null hypothesis can be rejected, indicating that we have enough evidence to conclude that the reviews per month for the top two hosts are significantly different. Host 2 seems to have more reviews per month as compared to Host 1.

In conclusion, we have a cleaned up dataset and we have analyzed the data visually as well as statistically. This helped us understand different underlying patterns in the data and uncover correlations. The data is ready for further feature engineering.

Machine Learning:

The Machine Learning notebook is available at: [Machine Learning Notebook]

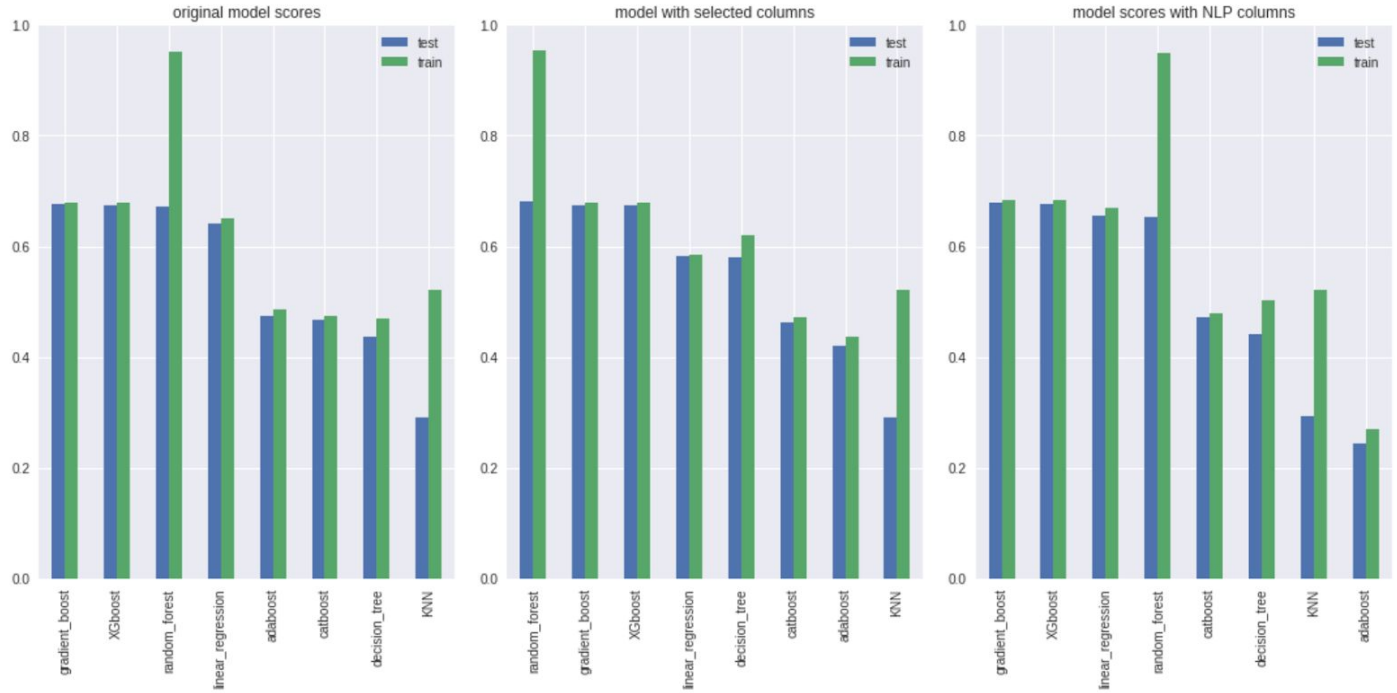
The final goal of the project was to be able to predict the Airbnb price of NYC listings based on different features provided - including reviews, host information, location, etc. The first thing that we needed to do was feature engineering of the different kinds of features - categorical, numerical, datetime, string as the machine learning models cannot work with such variables directly.

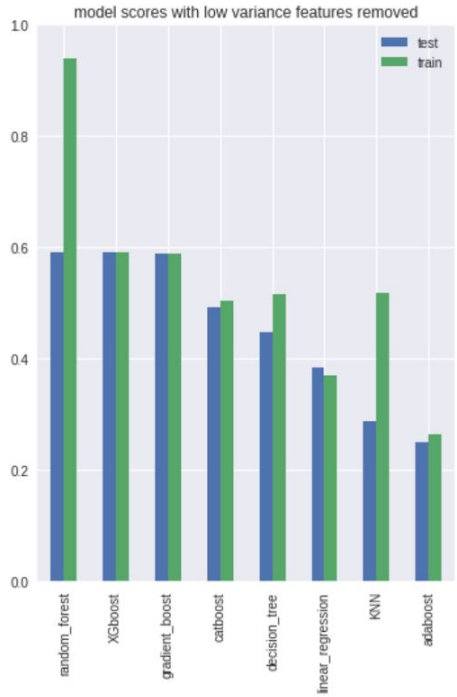
- Used one hot encoding to engineer all the categorical features in the dataset.
- Scaled all continuous numerical data points to 100 to maintain consistency.
- Preprocessed the text data in the description column and Used CountVectorization on it to extract useful text-related features.

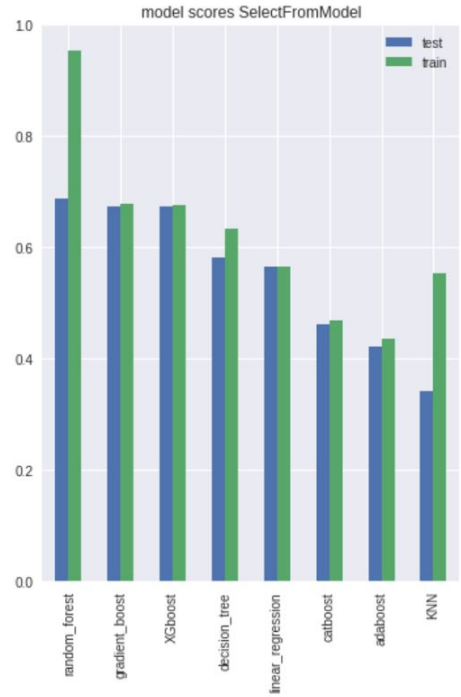
Once the feature engineering was done, we tried to fit different regression models into the data. 6 different modifications of the dataset were used:

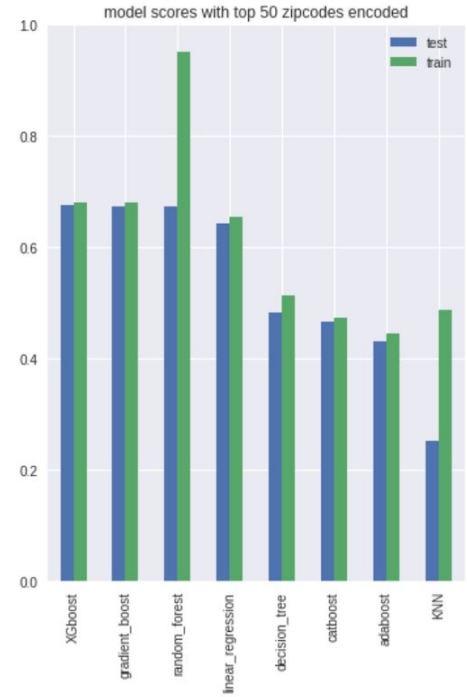
1. Dataframe with all original features engineered.
2. Dataframe with selected columns (dropping columns with large number of unique values which added lot of features during one hot encoding).
3. Dataframe with text features included.
4. Removing low variance features based on threshold of 0.5, and using the remaining features
5. Using SelectFromModel to select 150 top features.
6. Feature engineering the zipcodes based on their counts.

Test and Train scores of different models on different feature selections









Based on these visualizations, we find that the top 3 performing models were : XGboost, random forest and gradient boosting. Random forest seems to be over fitting in most cases. But we can continue with these models and try and increase the score by changing hyperparameters.

- Trying to optimize the score using Random Forest - Random Search CV and Grid Search CV
  We used random search cross validation to choose the most optimal parameters. Tested for different values of n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf. Based on the hyperparameters optimized in the random search, we targeted more specific variations in those variables using grid search cross validation.
  The table below shows the testing scores we got originally and with the optimizations.

Optimization	Score
Original Score	0.687
Random Search CV	0.684
Grid Search CV	0.685

  The score for random forest regressor did not improve much even after trying to find optimal parameters using grid search.

- Optimize the parameters for gradient boosting
  For gradient boosting hyperparameter optimization, we used the hyperopt library functions. We again tested different values for n_estimators, learning_rate, max_depth, max_features, min_samples_leaf.
  The testing score after hyperparameter optimization was 0.64 and the mean absolute error was estimated to be 20.25. We were unable to optimize gradient boosting regressor using this method.

In conclusion, random forest regressor gave the best score for the price prediction ~69%.

The top 20 features that contributed most to the price prediction and had the highest feature importances involve:

★ Entire home/apt	★ Longitude	★ Latitude	★ Accommodates	★ Cleaning Fee
★ Zip Code	★ Bedroom	★ Total Listing Count	★ Minimum Nights	★ Manhattan
★ Extra people	★ Guests Included	★ Bathroom	★ Reviews per month	★ Availability 30
★ Availability 90	★ Amenity count	★ Williamsburg	★ Review scores rating	★ Shared room

Most of these features do seem relevant in predicting price of a given place - the location, reviews, amenities, cost and availability.

Conclusion and Future Directions:

We started with a raw dataset which was obtained from the Airbnb NYC website and started with cleaning up the dataset and inferred the patterns and underlying correlations in the data set using EDA and inferential statistics. We gained alot of meaningful insights into data, and made decision on what to keep for further analysis. Following that, we performed feature engineering and rand different machine learning regression models on the dataset. The best performing model was random forest regressor and we were able to predict the Airbnb NYC listings prices with an accuracy of 69%. There were a lot of features and selecting the most important features was the main part in this dataset. The prediction can be improved if only the most important features are chosen and the regression models are run again. But overall, we were able to establish was we set out to, and there is still plenty more that can be done!