greatlearning
Learning for Life

Capstone Project Final Report on:

Prediction of rentals for Airbnb, NYC - 2019

Submitted by:

Aditya Shourya
Harishanandhan K
Sachin Acharya T
Niranjan Prabhu
Momin Ashraf Ahamed

Mentor:
Mr. Jayveer Nanda
Sr. Data Scientist, Confidential

Prediction of rentals for Airbnb, NYC - 2019

Abstract

Instead of waking to overlooked "Do not disturb" signs, Airbnb travelers find themselves rising with the birds in a whimsical treehouse, having their morning coffee on the deck of a houseboat, or cooking a shared regional breakfast with their hosts.New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

Airbnb takes a unique approach toward lodging. Part of the “sharing economy,” Airbnb offers you someone’s home as a place to stay instead of a hotel. On Airbnb, you can find places to crash on your backpacking trip through Europe, or you can find a place to stay for a month during your internship in Los Angeles. Also, if you want to rent out extra space in your own home, you can host through Airbnb and make money for allowing a guest to stay the night.

The aim of our preliminary analysis and data exploration is to evaluate the actions taken by the visitors and the hosts based on their environment and predicting the visitor’s staying price. Hoping to find insights such as which hosts are the busiest and why The challenge is to track bnb’s price index , rate fluctuations and accommodation pricing trends based on location popularity , rating , number of rooms and other attributes

Keywords : Airbnb , Lodging , Visitors, Machine Learning , Predictions ,Location Popularity , Ratings , business , internet , regression analysis , world , hotels and accommodations

Acknowledgements

At the outset, we are indebted to our Mentor Mr. Jayveer Nanda, Sr. Data Scientist, Confidential, for his time, valuable inputs and guidance. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We are extremely gifted and fortunate to have Ms.Shambhavi Shukla for her in-depth knowledge coupled with her passion in delivering the subjects in a lucid manner has helped us a lot. We are thankful to her for her guidance towards entire coursework.

We are thankful to Mr.Arshad Ali for his unflinching and unabated help extended to us always.

We also thank all the course faculties of the DSE program for providing us with a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Table Of Content

Chapter 1 : Project Overview

Background

And need for study

Airbnb, Inc. is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences and it records hotel price shifts from over 650 thousand hosts operating in more than 81 thousand cities and alternative accommodations around the world in its Hotel Price Index. The challenge is to track rate fluctuations and accommodation pricing trends based on location popularity , rating, no.of rooms and other attributes.

Approach

The data is extracted from kaggle playground web platform. After processing the dataset and cleaning the inconsistencies, the numerical and categorical features used in the Airbnb price prediction model is generated. Various regression algorithms are used to predict the bnb price based on set of independent variables like neighbourhood geographical data, the day of the week, room type, availability and the nights occupied. The predictive models are also used to identify the variables that strongly influence the conversion using variable importance and probabilistic approaches. The models are evaluated using relevant model performance measures to arrive at the most robust models for prediction.

Key Learnings

1. Different hosts and areas.
2. Which hosts are the busiest and the reason for it
3. Noticeable difference of traffic among different areas and the reason for it
4. Predictions based on locations, prices, reviews, etc

What Ticks Airbnb ?

Some key overviews :

1. Airbnb is a home-sharing platform that allows home-owners and renters ('hosts') to put their properties ('listings') online, so that guests can pay to stay in them. Hosts are expected to set their own prices for their listings. Although Airbnb and other sites provide some general guidance, there are currently no free and accurate services which help hosts price their properties using a wide range of data points.

2. Paid third party pricing software is available, but generally you are required to put in your own expected average nightly price ('base price'), and the algorithm will vary the daily price around that base price on each day depending on day of the week, seasonality, how far away the date is, and other factors.

3. Airbnb pricing is important to get right, particularly in big cities like NYC where there is lots of competition and even small differences in prices can make a big difference. It is also a difficult thing to do correctly — price too high and no one will book. Price too low and you'll be missing out on a lot of potential income.

4. This project aims to solve this problem, by using machine learning to predict the price for bnb in NYC.

Problem Statement :

Airbnb Inc. is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences and it records hotel price shifts from over 650 thousand hosts operating in more than 81 thousand cities and alternative accommodations around the world in its Hotel Price Index. The challenge is to track rate fluctuations and accommodation pricing trends based on location popularity , rating, no.of rooms and other attributes

Chapter 2 : About The Dataset

Usability : 10.0 License : CC0: Public Domain Tags:Business,Analysis
Number of records : 48895 Number of features : 15 + 1 target variable

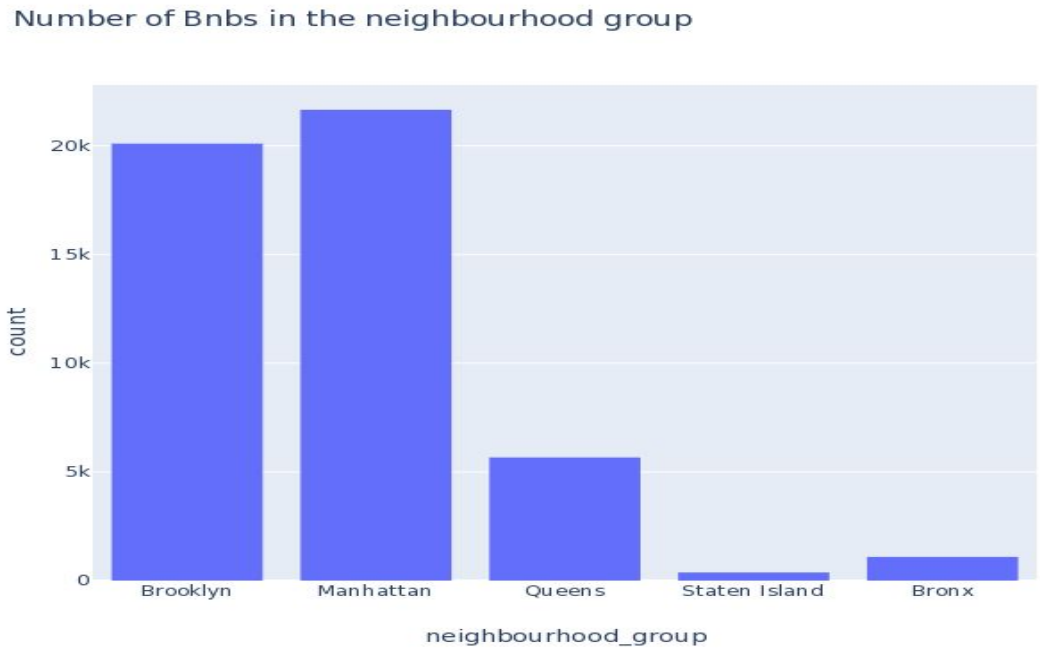
This public dataset is part of Airbnb, and the original source can be found on their website and kaggle's playground datasets.

This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

Feature Name	Feature Description	Type of data
id	listing ID	Discrete
name	listing ID	Discrete
host_id	host ID	Discrete
host_name	name of the host	Discrete
neighbourhood_group	location	Categorical
neighbourhood	area	Categorical
latitude	latitude coordinates	Discrete
longitude	longitude coordinates	Discrete
room_type	listing space type	Categorical
price	price in dollars	Numerical
minimum_nights	amount of nights minimum	Numerical
number_of_reviews	number of reviews	Numerical
last_review	latest review	Numerical
reviews_per_month	number of reviews per month	Numerical
calculated_host_listings_count	amount of listing per host	Numerical
availability_365	number of days when listing is available for booking	Numerical

About the feature columns (univariate analysis)

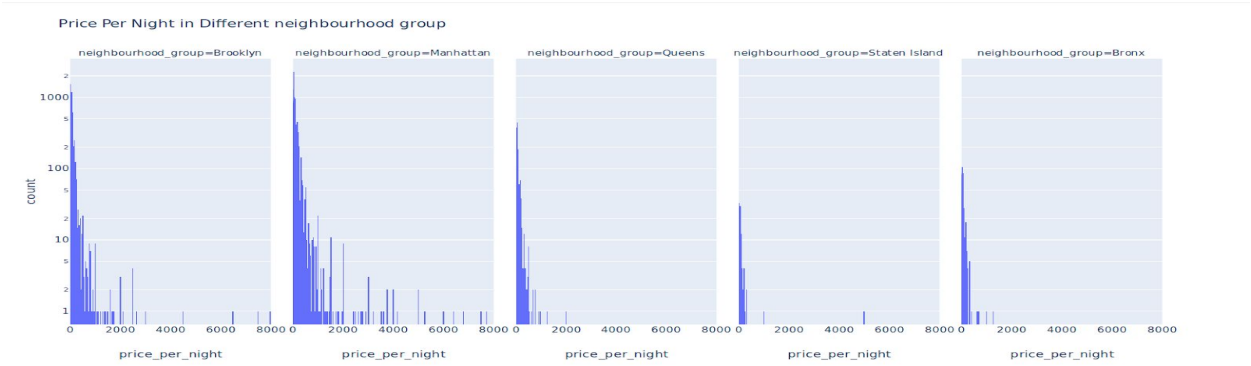
1. Number of Bnbs in the neighbourhood group



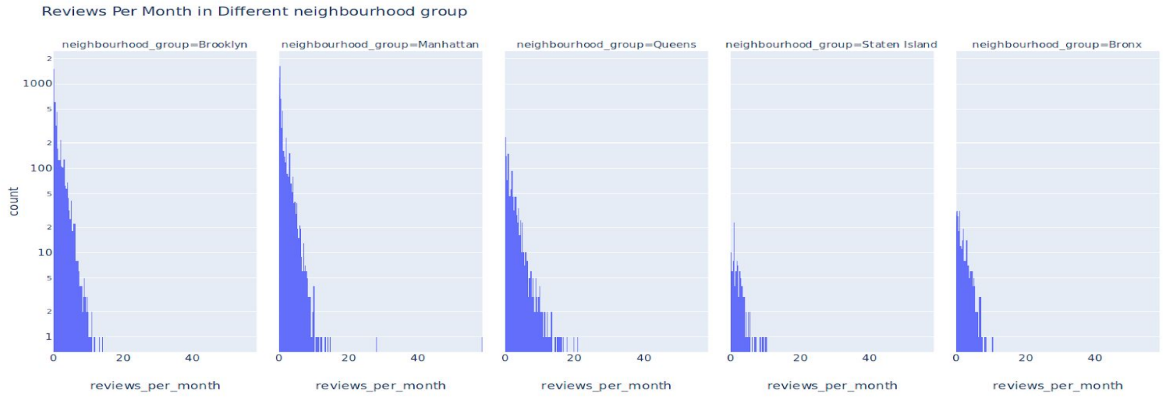
2. Hierarchy in Neighbourhood group (->) and neighbourhood. Showing top 10 neighbourhoods in their neighbourhood group

neighbourhood	price_per_night	minimum_nights	neighbourhood_group
Randall Manor	299.373434	2.578947	Staten Island
Battery Park City	248.086576	30.328571	Manhattan
Breezy Point	213.333333	1.000000	Queens
Riverdale	210.292208	5.363636	Bronx
Sea Gate	186.035714	4.142857	Brooklyn
Tribeca	183.826712	11.378531	Manhattan
Flatiron District	158.990377	6.225000	Manhattan
Jamaica Estates	139.682331	1.947368	Queens
NoHo	137.952595	5.987179	Manhattan
City Island	137.747354	2.055556	Bronx

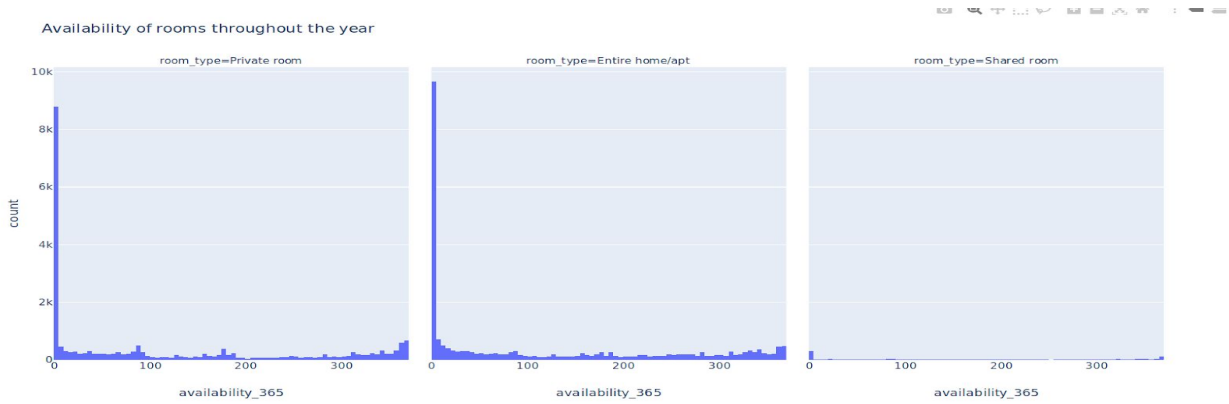
3. Price (per night) distribution



4. Reviews (per month) distributibution



5. Availability of rooms throughout the year

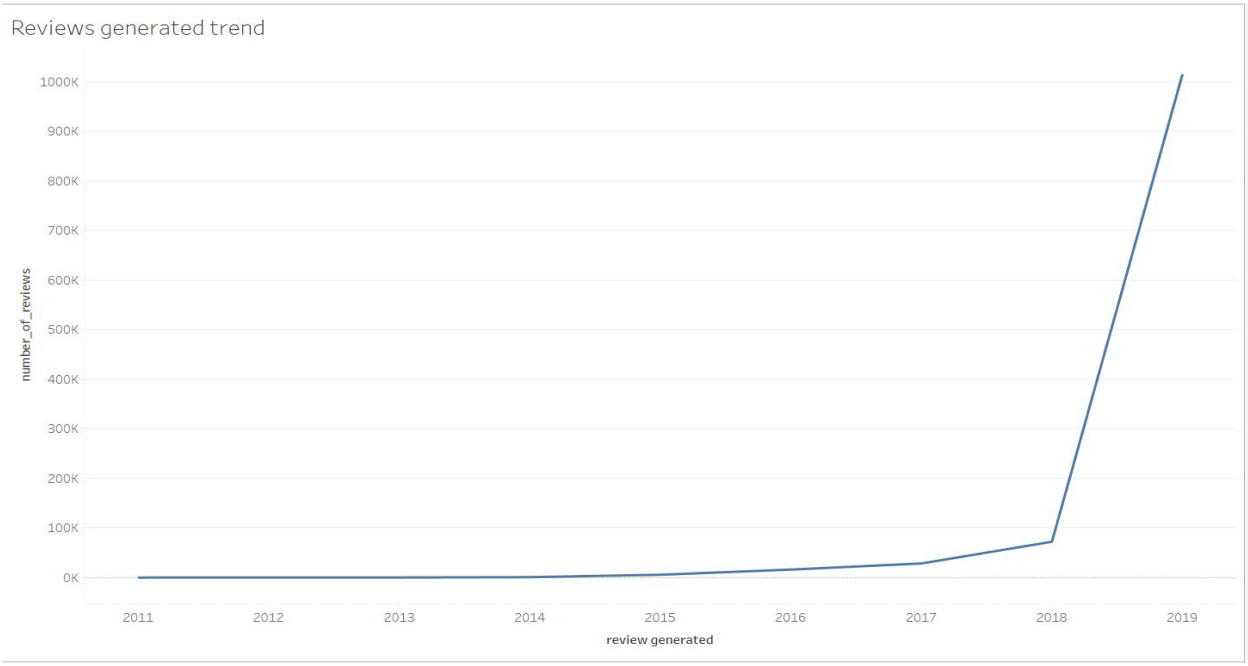


Chapter 3 : Our Tools and Techniques

Statistical methods

And testing techniques

Various regression algorithms can be used to analyze price prediction of bnbs. The model building exercise has also considered feature engineering and feature selection to ensure that the models built perform well when used for prediction.As most of the ratings data was generated in 2019 alone



Two main statistical methods are used in data analysis:

Descriptive statistics : which summarize data from a sample using indexes such as the mean or standard deviation

Inferential statistics :which draw conclusions from data that are subject to random variation

Evaluation Techniques

While training a model is a key step, how the model generalizes on unseen data is an equally important aspect that should be considered in every machine learning pipeline. We need to know whether it actually works and, consequently, if we can trust its predictions.

1. Holdout :The purpose of holdout evaluation is to test a model on different data than it was trained on. This provides an unbiased estimate of learning performance.The holdout approach is useful because of its speed, simplicity, and flexibility. However, this technique is often associated with high variability since differences in the training and test dataset can result in meaningful differences in the estimate of accuracy
2. Cross-validation is a technique that involves partitioning the original observation dataset into a training set, used to train the model, and an independent set used to evaluate the analysis.The most common cross-validation technique is k-fold cross-validation,

Evaluation Metrics

1. RMSE (Root Mean Square Error) : It represents the sample standard deviation of the differences between predicted values and observed values (called residuals).
2. MAE : MAE is the average of the absolute difference between the predicted values and observed values. The MAE is a linear score which means that all the individual differences are weighted equally in the average.
3. Adjusted R^2 : Just like R^2, adjusted R^2 also shows how well terms fit a curve or line but adjusts for the number of terms in a model.

Chapter 4 : Exploring data

Why is EDA necessary for modelling ?

The objective is to understand the problem in order to generate testable hypotheses. As such, the outcomes like the graphs and summary statistics are only to improve the understanding, not to demonstrate a relationship in the data to a general audience. This gives the agile flavor to the process.

Here are some of the basic objectives of EDA :

1. Suggest hypotheses about the causes of observed phenomena
2. Assess assumptions on which statistical inference will be based
3. Support the selection of appropriate statistical tools and techniques
4. Provide a basis for further data collection through surveys or experiments

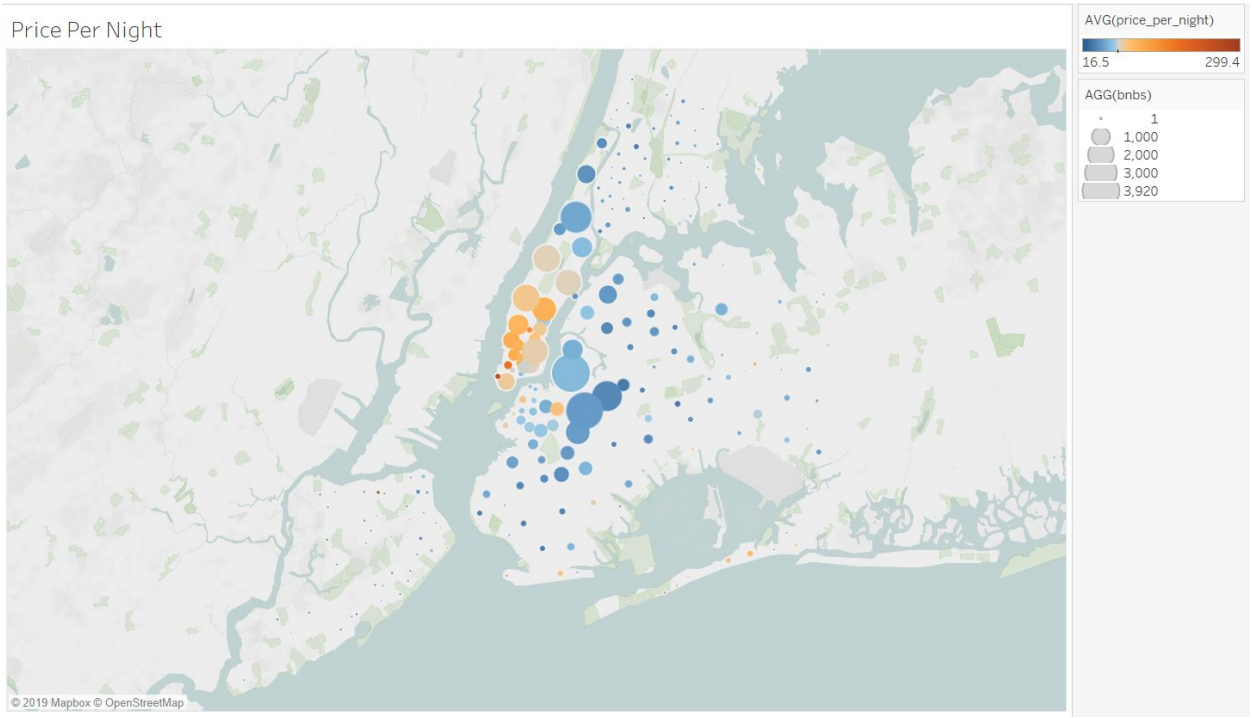
In spending time with the data up-front one can build an intuition with the data formats, values, and relationships that can help to explain observations and modeling outcomes later.

The process can be used to sanity check the data, to identify outliers and come up with specific strategies for handling them. In spending time with the data, you can spot corruption in the values that may signal a fault in the data logging process.It is called exploratory data analysis because you are exploring your understanding of the data, building an intuition for how the underlying process that generated it works and provoking questions and ideas that you can use as the basis for your modeling.

Diving Into Our dataset

Inference about our feature columns

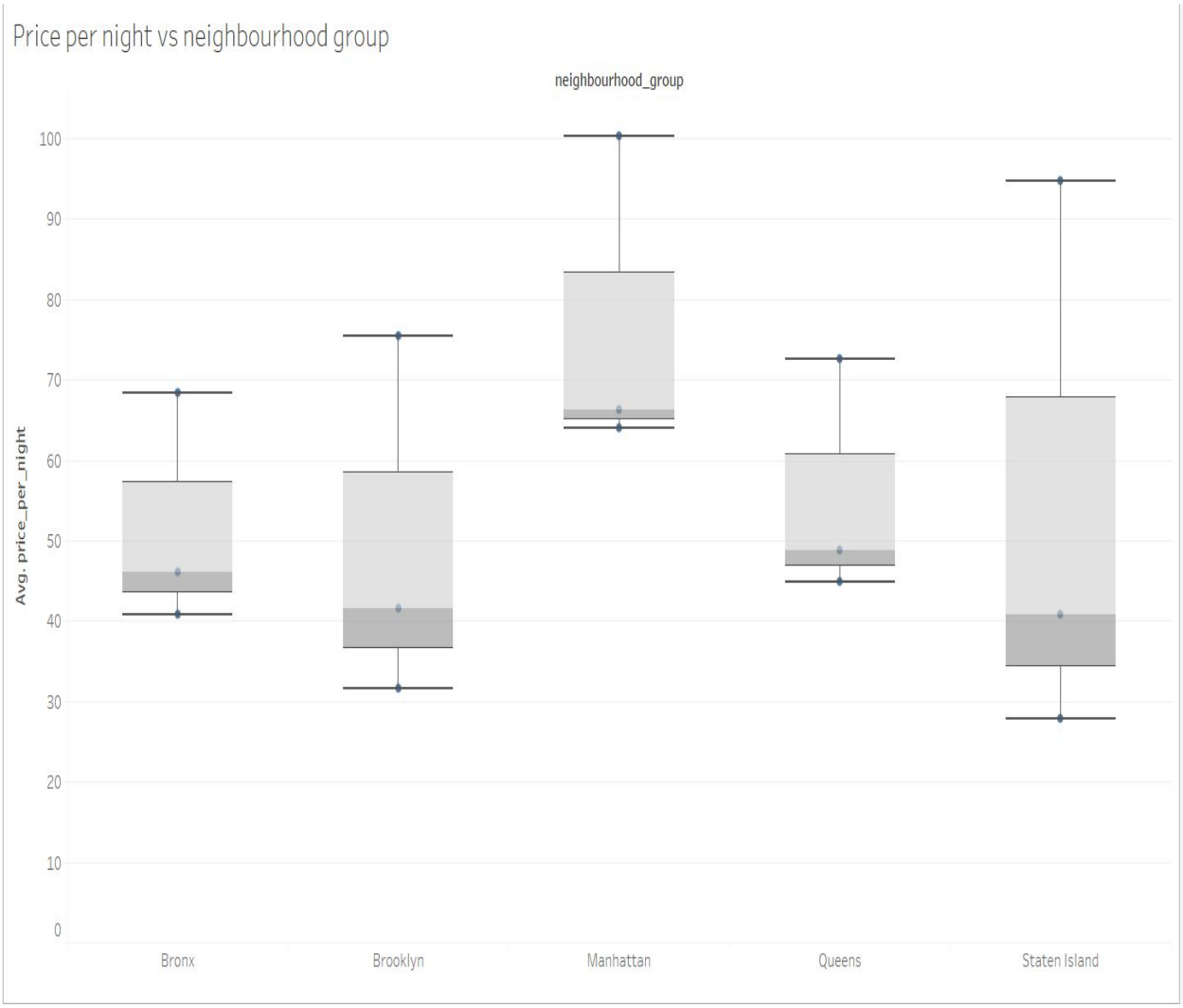
How costly is New York ?



The average price ranges somewhere from 16.5 dollars per night to 299.4 dollars per night.This plot shows the number of bnb in the neighbourhood as the size of the bubble and price per night from the diverging colour map .

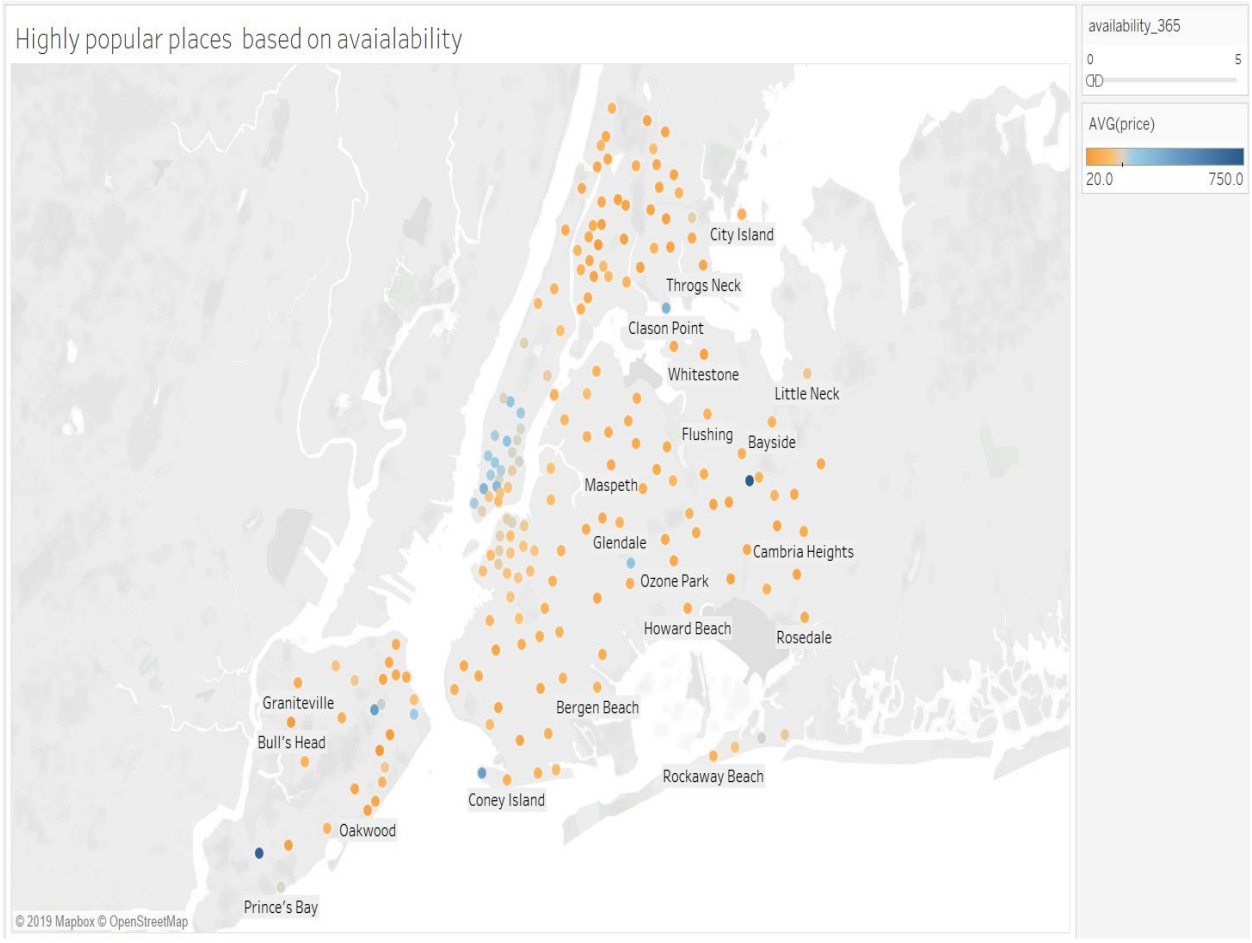
1. Most of the orange dots lie in the neighborhood group of Manhattan . Manhattan being one of the most expensive cities in new york justifies for such prices.
2.

Interesting Fact : The average price of a home in Manhattan is roughly around $2 million, with lower-priced homes going for around $1 million in places such as Williamsburg, Brooklyn.



Price distribution of the famous neighbourhood group

No surprise that manhattan leads this race with its skewed distribution. Although one can Rent Bed and breakfasts in Manhattan from $20/night these days or find unique places to stay with local hosts. But the lower whiskers starts from a whopping $65/night . As per a report by Alastair Boone from city lab [1] Airbnb was raising rents and taking housing off the rental market in Manhattan . There was a kind of increasing outcry from communities in 2018 , from housing organizations, from activists, and from elected officials that short-term rentals are having a negative impact on housing



This plot shows the most popular places in nyc based on low availability throughout the year. Here each bubble shows the neighbourhood with availability less than 5 days throughout the year most of these bnbs are cost effective about $20/night.Here are the top 5 cheapest places with high availability in a popular area . Most of these bnbs are owned by superhosts.Superhosts are experienced, highly rated hosts who are committed to providing great stays for guests.

1. Chelsea , Manhattan, average price per night $45/night , availability : 40 days
2. Windsor Terrace , Brooklyn , average price per night $42/night , availability : 76 days
3. Forest Hills , Queens , average price per night $45/night , availability : 82 days
4. Woodhaven , Queens , average price per night $ 48/night ,availability : 25 days
5. Washington Heights , Manhattan,average price per night $60/night ,availability : 30 days. But prices are on the rise, says Warburg Realty broker Joel Moss [2].

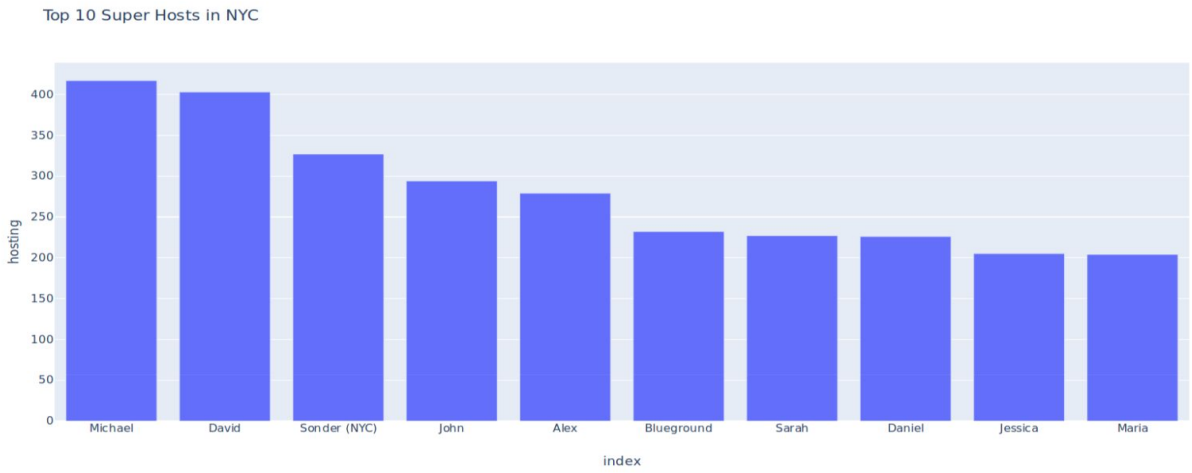
Superhosts in New York

Airbnb launched their Superhost program in 2016 to reward their most devoted hosts with a special VIP status. Beyond just a feather in the cap, They stated that hosts who achieved Airbnb Superhost status would reap benefits such as improved search placement, better booking conversions, and at the end of the day, more revenue.

What constitutes a superhost ?

The four criteria that hosts must meet to become an Airbnb Superhost are:

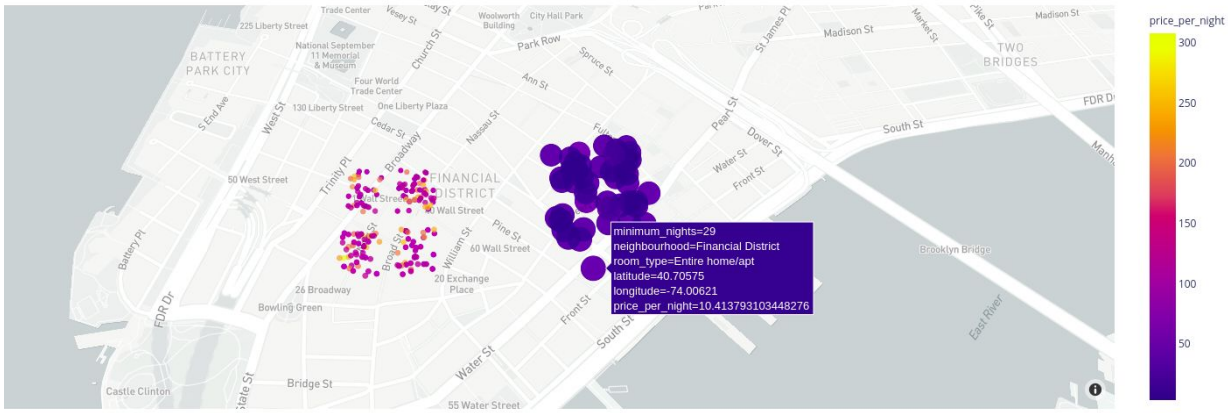
1. Host a minimum of 10 stays in a year
2. Respond to guests quickly and maintain a 90% response rate or higher
3. Have at least 80% 5-star reviews
4. Honor confirmed reservations (meaning hosts should rarely cancel)



Of these hosts:

Sonders NYC is a startup that is trying to combine the reliability of a hotel with the authenticity of a rental . In a sense it may have already succeeded. This 3rd party hosting service has been testing it concepts for the last several months through airbnb and other online booking sites , as per October 2019

How Superhosts avoid NYC’s short term rental laws

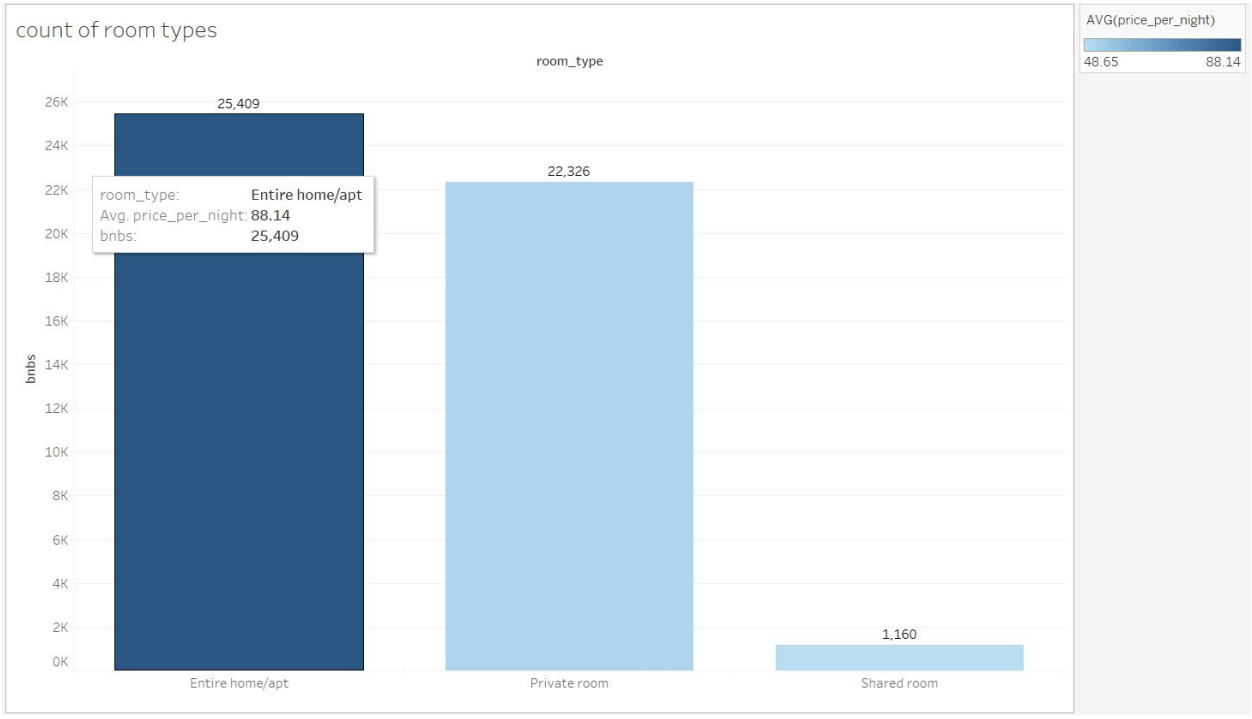


Hospitality company Sonder is opening its first New York location.The plot shows Sonder’s listing near the wall street , the size of the bubble represents minimum nights stayed and color represents the average price per night of the bnb.Normally all bubbles should be of the size as that of the blue bubble (20-30 days) .The company inked a master lease for 169 units at Metro Loft Management’s office-to-apartments conversion 20 Broad Street, where it will occupy the lower eight residential floors. It expects the units to open this quarter.San Francisco-based Sonder, founded in 2012, leases apartments and turns them into furnished short-term rentals that compete with hotels. In August the company raised $85 million in a Series C funding round led by Greenoaks Capital that brought its total funding to $135 million. It claims to manage around 3,000 units in four countries.To get around New York’s strict limits on commercial short-term rentals, Sonder goes for properties that meet the zoning and building requirements of a hotel. The units “look like residential apartments, but are actually built out as hotel suites,” said the company’s New York City general manager Arthur Shmulevsky. The units will also get dedicated elevators.New York’s regulations are a big reason it took Sonder so long to open its first location here, said CEO Francis Davidson. “New York real estate is notoriously complicated,” he said.

Where superhosts status matters most ?

Our assumption was that Superhost status would matter most in large markets such as Manhattan where hosts needed every advantage possible to stand out from the crowd. We were surprised to discover there is no correlation between Superhost hostings and the number of short-term rentals in a city.

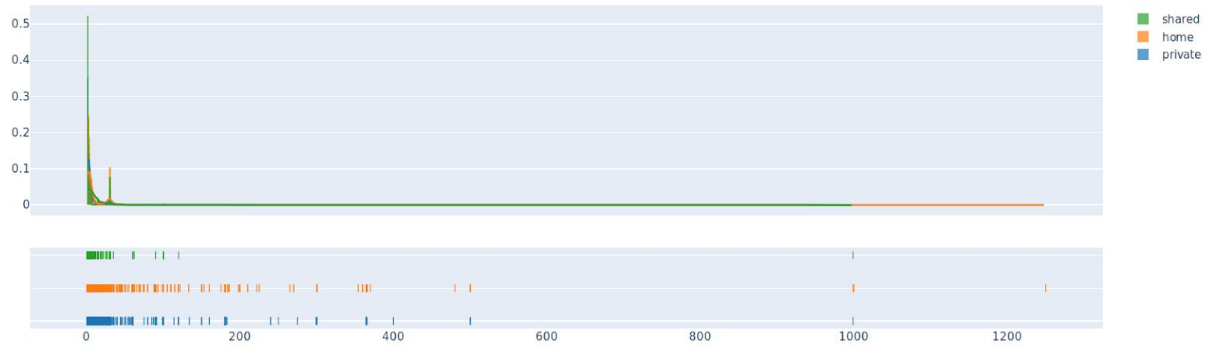
Types Of Rentals



Airbnb listings are categorized into the following home types:

1. Entire place: Guests have the whole place to themselves. This usually includes a bedroom, a bathroom, and a kitchen. Hosts should note in the description if they'll be on the property (ex: "Host occupies first floor of the home")
2. Private room: Guests have their own private room for sleeping. Other areas could be shared.
3. Shared room: Guests sleep in a bedroom or a common area that could be shared with others.

Types of guest as seen from nights stayed

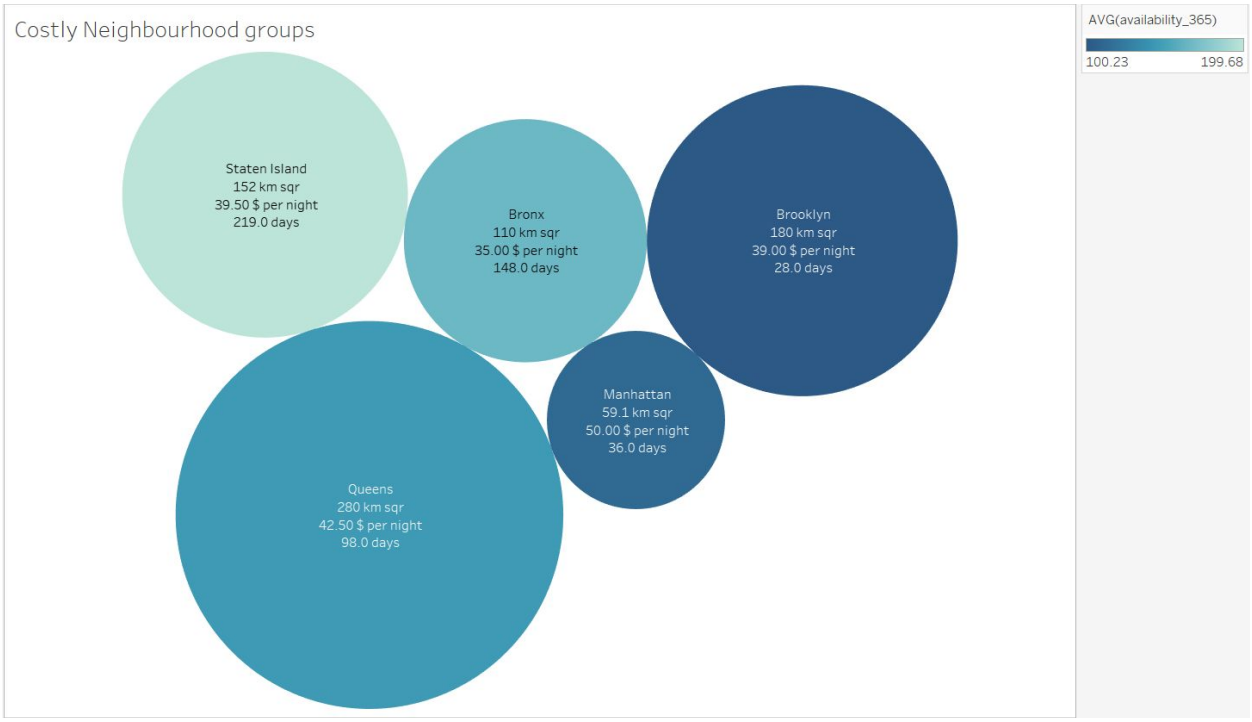


This histogram shows the type of rooms and nights stayed by the guests . People usually like to rent a home if they are planning to stay longer than a year or so .We can infer from the rug that guests can be broadly classified on the basis of nights stayed. Further analysis showed that bnbs on the left half are mostly from private rooms and shared rooms and are less available compared to rentals which give out rooms for more than 1 year.When it comes to vacation rentals, no two guests are alike. Every Airbnb guest has a different idea of where the best place to travel to is, the best time of year to go, who they'd like to travel with

4 types of vacation rental guests , we think can be classified based on number of nights stayed

1. Business Travellers (< nights stayed).
2. Cultural Enthusiast (< nights stayed).
3. Family with children (> nights stayed).
4. Couples and travellers with disabilities (> nights stayed)

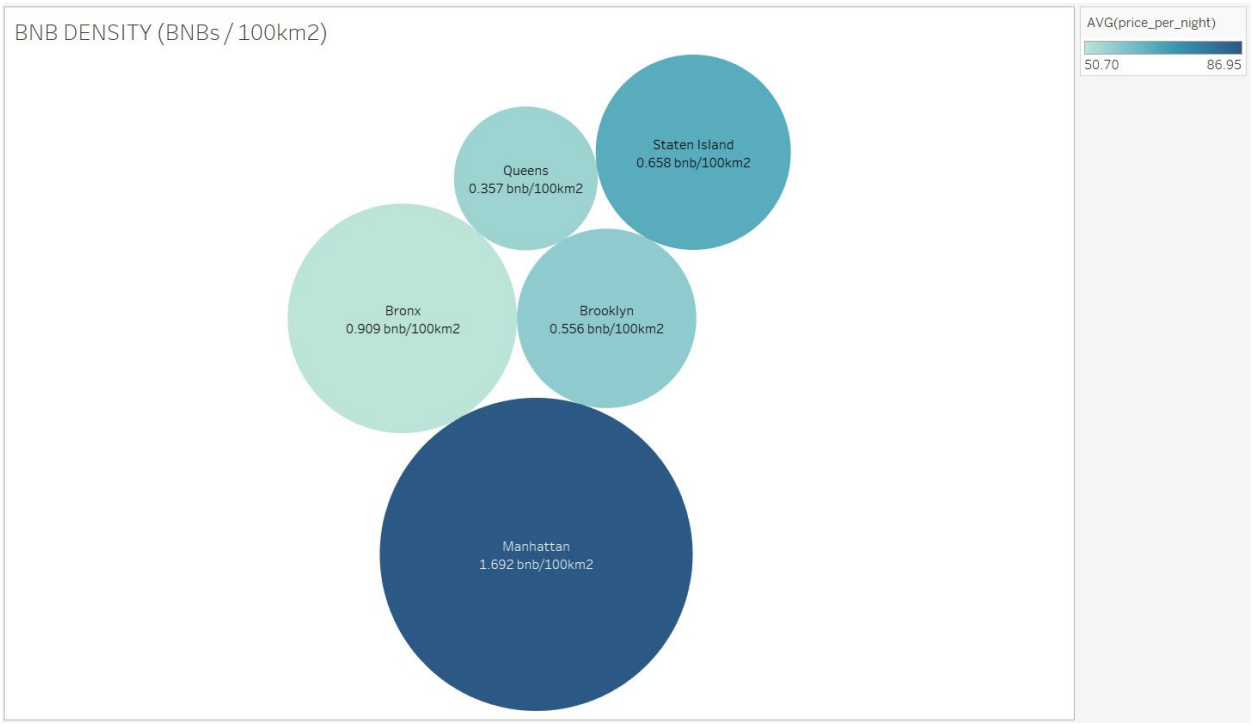
Manhattan Needs its own section



This plot shows the costly neighbourhood groups in New York .The size of the bubble represents the area of the neighbourhood and the availability is shown by the reversed blue-teal colour map.Manhattan is the costliest and only second to Brooklyn in terms of low availability.

According to Luis Ferré-Sadurní of New York Times [3] ,Airbnb's growing influence caused rents to increase significantly in tourist areas and gentrifying neighborhoods in Manhattan and Brooklyn, where the majority of the company's rentals are concentrated .

In Manhattan's Hell's Kitchen and Chelsea neighborhoods and the Midtown Business District, which accounted for about 11 percent of all Airbnb listings in New York City in 2016, average monthly rents increased by $398 between 2009 and 2016, of which $86, or 21.6 percent, was a result of Airbnb's presence.Most New Yorkers use Airbnb as a source of extra income to make ends meet and that entire apartments are rented for a median of 60 nights a year. But Airbnb has long been scrutinized by officials because some landlords use Airbnb to effectively run **illegal hotels** in residential buildings.



This plot shows the costly neighbourhood groups in New York .The size of the bubble represents number of bnbs per 100 km^2 of the area. And the price is shown by the blue-teal colour map.Manhattan has the highest bnb density.

For years, New Yorkers have felt the burden of rents that go nowhere but up, and Airbnb is one reason why, the city comptroller, Scott M. Stringer, said in an interview.

"It's just simply supply and demand. Fewer apartments to rent means higher prices, and that's the Airbnb effect." - M.Stringer

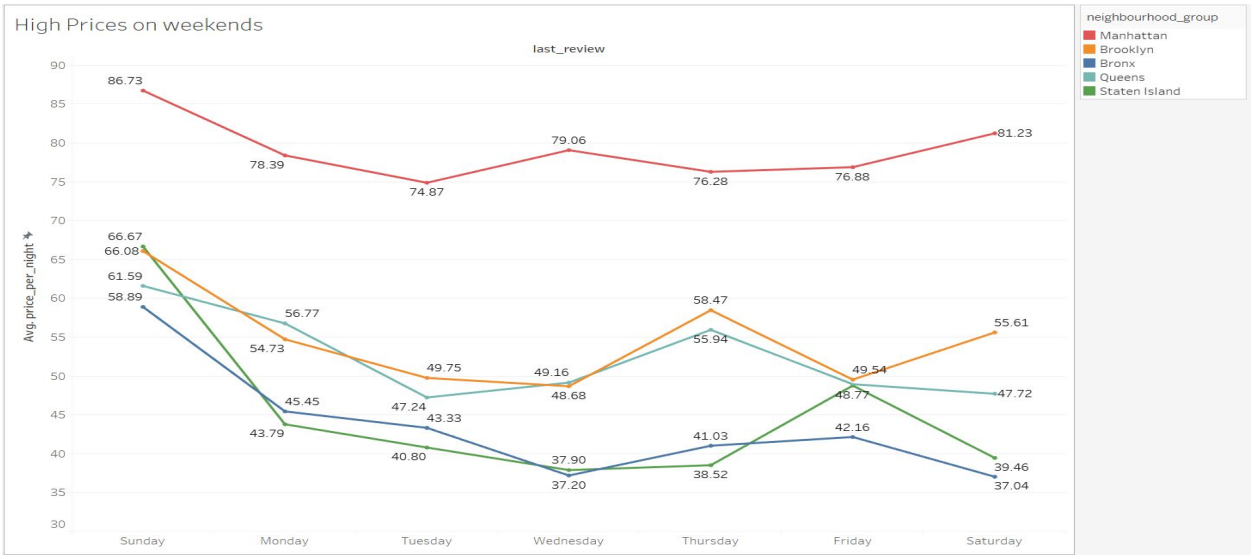
---

Scott M. Stringer is the 44th and current New York City Comptroller and a New York Democratic politician who previously served as the 26th Borough President of Manhattan.

A study drew (a kaggle kernel) on data scraped from Airbnb listings (our dataset,mixed distribution) and used regression analysis to compare what rents would have been across 55 neighborhoods if thousands of units had not been listed on the home-sharing website.

Weekdays and weekends pricing

Airbnb superhosts never draw blanks when it comes to proper pricing – It really is one of the most challenging aspects. Things like market demand and competitor behaviors are very carefully examined and almost every host comes up with a variable listing model usually based on weekdays and holidays.



This plot shows the trend line of average price per rooms vs the weekdays.The increase in price on weekends (Friday and Saturday) are more clearly evident on costlier neighbourhood group such as Manhattan and brooklyn.

According to Airbnb's website :

Custom pricing model : Any custom weekly and monthly prices the host sets after turning on weekend pricing will not be overridden. However, if the host decides to turn on weekend pricing after setting custom prices, weekend pricing will replace all Friday and Saturday nightly prices in the listing.

Chapter 5 : Feature Engineering

Feature engineering is the process of using domain knowledge of the data to create features that makes more sense and easier for a machine learning algorithms This input data comprise features, which are usually in the form of structured columns. Algorithms require features with some specific characteristics to work properly.According to a survey in Forbes, data scientists spend 80% of their time on data preparation:

According to Luca Massaron , a data scientist and marketing research director for packt publication ,

---

"The features you use influence more than anything else the result. No algorithm alone, to my knowledge, can supplement the information gain given by correct feature engineering." — Luca Massaron

---

We applied feature engineering on :

1.Datetime column ———> weekdays————-> if_weekend or if holiday (boolean)

2.Price and nights stayed————————-> price per night

3.Number of bnb and area of neighbourhood (gathered from internet) ———> Bnb density

4.Bnb density ————-> popularity grade of neighbourhood

Feature ranking :

	feature	importance	Rankings
0	host_id	0.232927	4
1	neighbourhood	0.114782	2
2	minimum_nights	0.130359	3
3	number_of_reviews	0.145520	1
4	reviews_per_month	0.165666	1
5	calculated_host_listings_count	0.057896	1
6	availability_365	0.139550	1
7	neighbourhood_group_Brooklyn	0.001146	1
8	neighbourhood_group_Manhattan	0.001453	1
9	neighbourhood_group_Queens	0.001183	1
10	neighbourhood_group_Staten Island	0.000679	1
11	room_type_Private room	0.007574	1
12	room_type_Shared room	0.001265	1

Rfe_score : Feature ranking with recursive feature elimination.

Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features.

Importance: An extra-trees regressor.

This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Chapter 6 : Assumptions for modelling

Model : A simpleton linear regressor

OLS Model (Ordinary Least Square)

Model Summary :

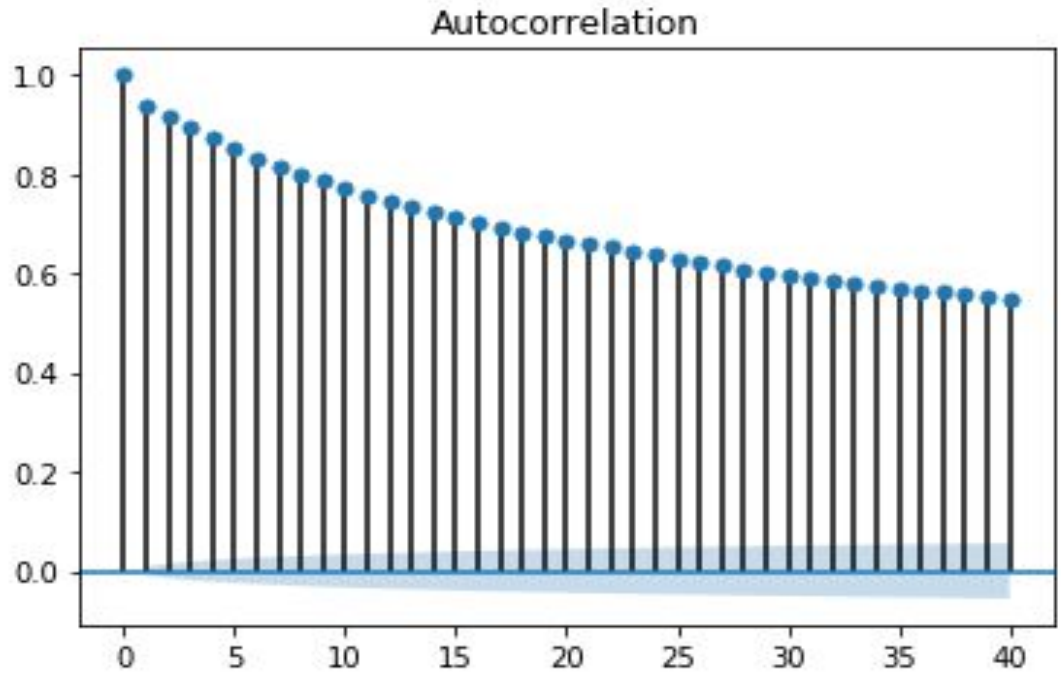
Dep. Variable:	price	**R-squared:**	0.078
Model:	OLS	**Adj. R-squared:**	0.078
Method:	Least Squares	**F-statistic:**	457.9
Date:	Mon, 11 Nov 2019	**Prob (F-statistic):**	0.00
Time:	07:35:58	**Log-Likelihood:**	-3.3541e+05
No. Observations:	48895	**AIC:**	6.708e+05
Df Residuals:	48885	**BIC:**	6.709e+05
Df Model:	9		
Covariance Type:	nonrobust		

1. Method : Least squares : sum of squares of residues
2. DF Model : Degree of freedom (number of features -1)
3. Covariance type : Standard Errors assume that the covariance matrix of the errors is correctly specified.
4. R-squared: proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.
5. Adjusted R-squared : The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.
6. F-statistic and prob F-statistic: The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. Its value will range from zero to an arbitrarily large number. The value of Prob(F) is the probability that the null hypothesis for the full model is true. Our model failed!!

Assumption : 1

Linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of y(x+1) is not independent from the value of y(x).

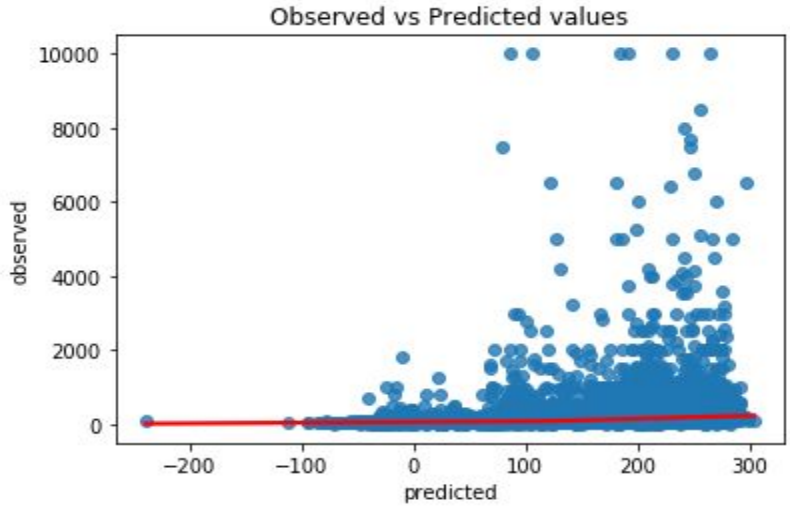
While a scatterplot allows you to check for autocorrelation, you can test the linear regression model for autocorrelation with the Durbin-Watson test. Durbin-Watson's d tests the null hypothesis that the residuals are not linearly auto-correlated. While d can assume values between 0 and 4, values around 2 indicate no autocorrelation. As a rule of thumb values of 1.5 < d < 2.5 show that there is no auto-correlation in the data. However, the Durbin-Watson test only analyses linear autocorrelation and only between direct neighbors,which are first order effect.

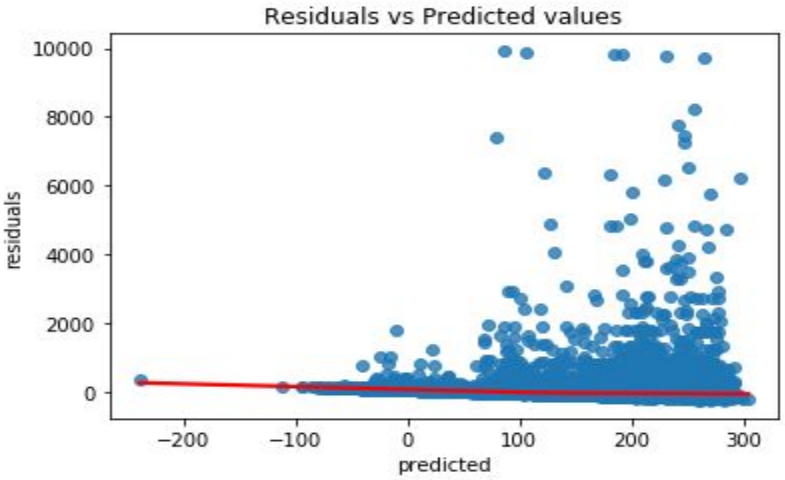


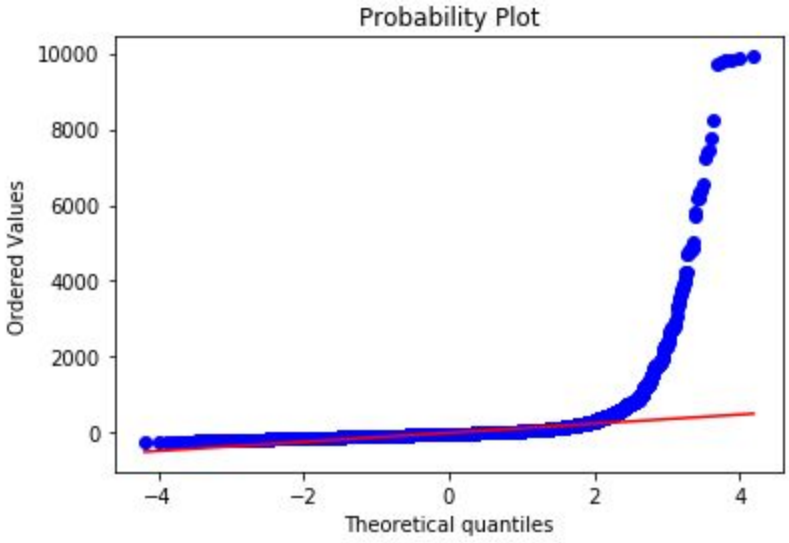
Highly positive autocorr. Durbin watson : 0.092

Assumption : 2

Linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatter plots,



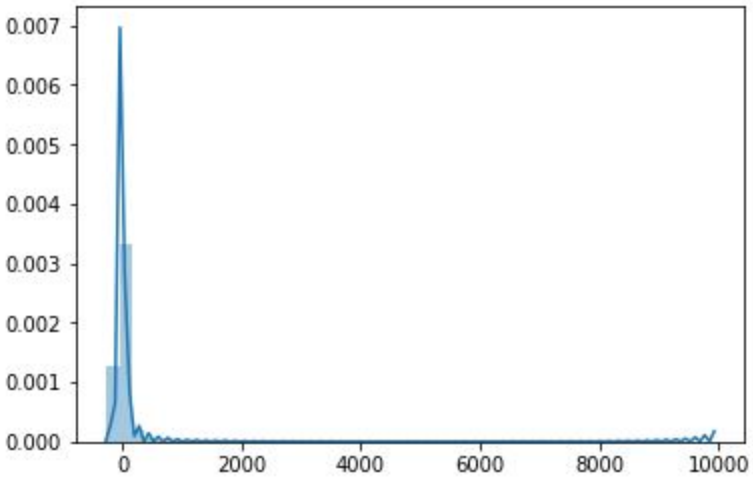




The data is highly non-linear

Assumption : 3

The linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot. Normality can be checked with a goodness of fit test, e.g., the Kolmogorov-Smirnov test. When the data is not normally distributed a non-linear transformation (e.g., log-transformation) might fix this issue.



The data is not normal.

Assumption 4:

The assumption of the linear regression analysis is homoscedasticity. The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line).

The Goldfeld-Quandt Test can also be used to test for heteroscedasticity. The test splits the data into two groups and tests to see if the variances of the residuals are similar across the groups. If homoscedasticity is present, a non-linear correction might fix the problem.

The F-statistic from Goldfeld-Quandt Test is 277.159 and p-value is 0.00 which makes the data heteroscedastic

Assumption 5:

linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.

Multicollinearity may be tested with three central criteria:

1. Correlation matrix – when computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1.

2. Tolerance – the tolerance measures the influence of one independent variable on all other independent variables; the tolerance is calculated with an initial linear regression analysis. Tolerance is defined as $T = 1 – R^2$ for these first step regression analysis. With $T < 0.1$ there might be multicollinearity in the data and with $T < 0.01$ there certainly is.

3. Variance Inflation Factor (VIF) – the variance inflation factor of the linear regression is defined as VIF = 1/T. With VIF > 5 there is an indication that multicollinearity may be present; with VIF > 10 there is certainly multicollinearity among the variables.

If multicollinearity is found in the data, centering the data (that is deducting the mean of the variable from each score) might help to solve the problem. However, the simplest way to address the problem is to remove independent variables with high VIF values.

	vif
host_id	1.276040
neighbourhood_group	1.034196
neighbourhood	1.022862
room_type	1.032809
minimum_nights	1.059677
number_of_reviews	1.777338
reviews_per_month	1.792588
calculated_host_listings_count	1.109140
availability_365	1.174826

All the vif scores represent that there is no multicollinearity among the features.

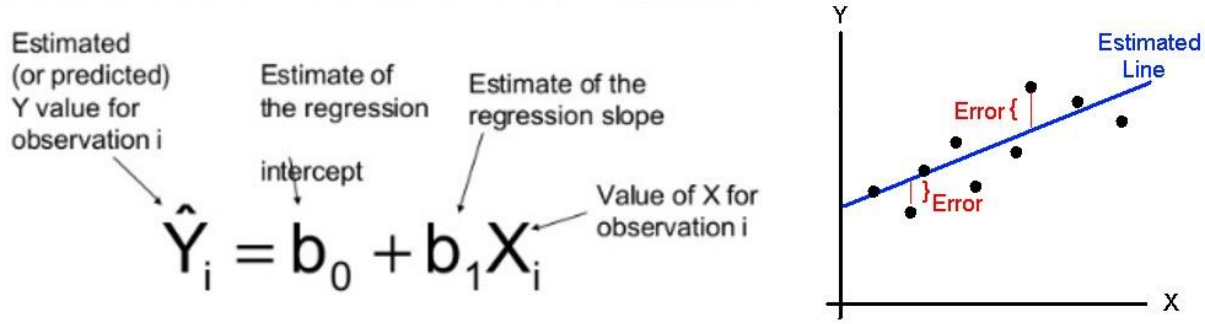
Chapter 7 : Modelling

Linear Modelling:

It's a method to predict a target variable by fitting the *best linear relationship* between the dependent and independent variable.

This method uses a single independent variable to predict a dependent variable by fitting a best linear relationship.

This method uses a single independent variable to predict a dependent variable by fitting a best linear relationship.



The model with Airbnb was very poor in both training as well as testing. The training score was 0.037 and at testing it was 0.050

The need to reconsider the features based on feature ranking is required. So we have used ExtraTreesClassifier to rank our features.

ExtraTreesClassifier

ExtraTrees is named for Extremely Randomized Trees

ExtraTreesClassifier is an ensemble learning method fundamentally based on decision trees. ExtraTreesClassifier, like RandomForest, randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting. Let's look at some ensemble methods ordered from high to low variance, ending in ExtraTreesClassifier.

Decision Tree (High Variance)

A single decision tree is usually overfits the data it is learning from because it learn from only one pathway of decisions. Predictions from a single decision tree usually don't make accurate predictions on new data.

Random Forest (Medium Variance)

Random forest models reduce the risk of overfitting by introducing randomness by:

1. Building multiple trees (n_estimators)
2. Drawing observations with replacement (i.e., a bootstrapped sample)
3. Splitting nodes on the best split among a random subset of the features selected at every node

Extra Trees (Low Variance)

Extra Trees is like Random Forest, in that it builds multiple trees and splits nodes using random subsets of features, but with two key differences: it does not bootstrap

observations (meaning it samples without replacement), and nodes are split on random splits, not best splits. So, in summary, ExtraTrees:

1. Builds multiple trees with bootstrap = False by default, which means it samples without replacement
2. Nodes are split based on random splits among a random subset of the features selected at every node

In Extra Trees, randomness doesn't come from bootstrapping of data, but rather comes from the random splits of all observations.

Extra Tree Classifier is used in our Airbnb dataset in order to rank the important features including the derived, encoded features. Selection of important features are necessary as the accuracy with the base model is very less and cant be considered.

	feature	importance
0	host_id	0.229774
4	reviews_per_month	0.166528
3	number_of_reviews	0.144377
6	availability_365	0.140986
2	minimum_nights	0.129443
1	neighbourhood	0.120734
5	calculated_host_listings_count	0.054340
11	room_type_Private room	0.007052
8	neighbourhood_group_Manhattan	0.001797
12	room_type_Shared room	0.001499
7	neighbourhood_group_Brooklyn	0.001425
9	neighbourhood_group_Queens	0.001414
10	neighbourhood_group_Staten Island	0.000631

Regularization of the model

It's a way to prevent overfitting, and thus, improve the likely generalization performance of a model by reducing the complexity of the final estimated model.

Ridge Regression or L2 Regularization:

$$RSS_{RIDGE}(\boldsymbol{w}, b) = \sum_{\{i=1\}}^{N} (\boldsymbol{y}_i - (\boldsymbol{w} \cdot \boldsymbol{x}_i + b))^2 + \alpha \sum_{\{j=1\}}^{p} w_j^2$$

As shown in the above image,the RSS modified by imposing that sum of squares penalty on the size of the W coefficients. The super power of Ridge Regression is that it minimize the RSS by enforce the W coefficients to be lower, but it does not enforce them to be zero-minimize their impact on the trained model to simplify the statistical model.

After Ridge Regularization, the model has a r-squared value of 0.037 for training data and 0.0502 with test data.

Lasso Regression or L1 Regularization:

$$RSS_{LASSO}(w,b) = \sum_{\{i=1\}}^{N} (y_i - (w \cdot x_i + b))^2 + \alpha \sum_{\{j=1\}}^{p} |w_j|$$

Another kind of regularized regression that you could use instead of ridge regression is called Lasso Regression or L1 Regularization.Like ridge regression, lasso regression adds a regularisation penalty term to the ordinary least-squares objective.but the results are noticeably different.

With lasso regression, a subset of the coefficients are forced to be precisely zero. Which is a kind of automatic feature selection, since with the weight of zero the features are essentially ignored completely in the model. This sparse solution where only a subset of the most important features are left with non-zero weights, also makes the model easier to interpret which is a huge advantage.

After Lasso Regularization, the model has a r-squared value of 0.037 for training data and 0.0502 with test data.

So there is a need to look out for other regression models.

Model for price prediction:

A **decision tree** is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and

utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression). Decision Tree algorithms are referred to as **CART (Classification and Regression Trees)**.

Common terms used with Decision trees:

1. **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
4. **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
6. **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
7. **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

Decision tree builds regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

The Decision Tree regressor for our dataset gives a r-squared value of 0.24 with train data and 0.19 for test data

Chapter 8 : Conclusion

We didn't get all the ingredients

1. We feel that some features are missing to accurately predict the price of the bnbs.Such as most bnb charge an extra $25 /day per person . Our dataset is missing the number of people in the party .
2. Since we don't have information on the booking date . Hard assumptions such as booking_date = date of review - nights stayed , we are afraid that it might increase the cost of the model as it directly affects the number of weekends and holidays during the guest's stay.
3. Price model varies between hosts and superhosts. Since our dataset is missing the rating its difficult cluster the hosts and the superhosts. However we can cluster them on the other bases.
4. Airbnb gives all their hosts have full control on their pricing model such as price of the bnb on the weekend , or a holiday ,management charges , damaged cause to property etc. which could mean each record in our dataset could very possibly have a unique custom pricing model with different coefficients attached to the feature columns. However activity of superhosts can be predicted with more confidence as they try to maintain the same model , varying only with the neighbourhood .

What we have achieved:

1. We Studied the difference of activity of top superhosts and normal hosts, activity of superhosts can be predicted with more confidence as they try to maintain the same model, varying only with the neighbourhood .
2. Clustered neighbourhood in terms of how costly the neighbourhood is.
3. Tried different types of regressors such as Decision-tree regressor : R2 : 0.25 (train) and R2 : 0.194 (test)