Airbnb Pricing Analysis

New York City

"BUAN 6340.001 – Programming for Data Science- S23"

Jindal School of Management, UT-Dallas

Professor: Thomas Lavastida

Group Members

NAME	NET ID
Harshini Parepally	HXP220004
Mohanish Pradeep	MXP210040
Navya Bhat	NXB220008
Rohit Patil	RXP220030

Abstract

This project aims to analyze the New York city airbnb dataset to identify key factors affecting pricing of Airbnb listings. Understanding the key factors that influence pricing, such as amenities, reviews, room type, and local demand and supply, will enable the hosts to develop a more strategic and effective pricing strategy. By leveraging this knowledge, hosts can optimize their pricing decisions to attract more bookings and maximize their revenue potential.

The dataset contains information on the listing demographics, cancellation rules, and reviews. We used a combination of exploratory data analysis, feature engineering, and machine learning models to identify key drivers of the pricing strategy.

Our analysis revealed that the most important factors affecting the price are the location of listing, construction year, reviews, and its availability around the year.

Introduction and Motivation

The hospitality sector undeniably contributes to the country's GDP and vice versa. Airbnb (founded in 2008) currently penetrates this market with a share of approximately 20% with over 6.1 million listings as of September 2022. What expedited them to conquer such market placement is their pricing strategy to beat the competitors like booking.com, HomeAway and many other private Hotel/Holiday apartment-style accommodation.

To help the new as well as existing hosts set a smart pricing strategy, the Airbnb's new SmartPricing software focuses on parameters like the local demand and supply, amenities, booking frequency, ratings, room type, calendar availability and similar listing searches into its algorithm. The following project tries to dig deeper into how the above parameters help Airbnb in striking the right prices for its listings in a Metropolitan like New York. In future, the scope of the project can be further extended to understand the impact of seasonality on pricing with relevant data.

Data Source and Description

This project seeks to analyze the New York Airbnb Open Data, made available by Airbnb on Kaggle. The data encompasses a comprehensive overview of listings and hosts in diverse neighborhoods within New York City, presenting a wealth of information on homestay activity in the area.

The data is presented in a CSV format, with 25 attributes and approximately 61283 unique data points. The attributes provide a detailed picture of the listings, including information on descriptions, neighborhood, neighborhood group, construction year, average review scores, host identity, and the target variable of price, among others. The dataset offers a rich source of information to delve deeper into the factors that influence the pricing of homestays in New York City.

Attribute	**Description**	**Data type**
id	Airbnb's unique identifier for the listing	integer
name	Name of the listing	string
hostid	Airbnb's unique identifier for the host	integer
host_identity_verified	True if the identity of the host has been verified. False if it has not been verified	boolean [t=true; f=false]
hostname	Name of the host. Usually just the first name(s).	string
neighborhood group	The neighborhood group geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	string
neighborhood	The neighborhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	string
lat	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	integer
long	Uses the World Geodetic System (WGS84) projection for latitude and longitude.	integer
country	country of the listing	string
country code	country code of the listing	string
instant_bookable	Whether the guest can automatically book the listing without the host requiring accepting their booking request. An indicator of a commercial listing.	boolean [t=true; f=false]
cancellation	cancellation policy	string
roomtype	Type of the room that has been listed.	string
constructionyear	Year of construction of the room	numeric
service_fee	service fee charged to the customer	numeric
price	daily price in local currency. Note, $ sign may be used despite locale	numeric

Minimum nights	minimum number of nights stay for the listing	numeric
number of reviews	The number of reviews the listing has (in the last 12 months)	numeric
last review	The date of the last/newest review	date
reviews per month	The number of reviews the listing has over the lifetime of the listing	numeric
review rate number	review rate number	numeric
calculated host listings count	The number of listings the host has in the current scrape, in the city/region geography.	numeric
availability 365	The availability of the listing x days in the future as determined by the calendar. Note a listing may be available because it has been booked by a guest or blocked by the host.	numeric
house_rules	house rules specified by the host	string
license	The license/permit/registration number	string

Project Outline

This project aims to delve into the complexities of the New York City Airbnb market. Our focus will be to understand the key factors that drive the pricing of Airbnb listings and the characteristics that define a successful host. To do so, we will utilize regression analysis and clustering techniques on the Airbnb Open Data, sourced from Kaggle.

The initial stage will include cleaning and preprocessing the data, which includes removing irrelevant features, addressing missing values, and handling outliers. We will also standardize the features to ensure comparability.

In the regression analysis, we will split the data into a training set and a testing set. Our aim is to identify the most significant factors affecting the price of a listing, such as the number of bedrooms, neighborhood groups, reviews, and amenities offered. The training set will be used to build the regression model, which will help us in determining these factors.

Additionally, we will group the data by neighborhood and perform clustering analysis to identify price trends in each neighborhood. This will help us understand the pattern of listings based on their price range. The results will be visualized through scatterplots and heat maps.

Datasets and Preprocessing

The dataset was sourced from Kaggle, all the applicable reference links have been provided in the reference section below.

The dataset has 26 columns (attributes) and 102599 rows (observations). The target variable is 'price' for linear regression.

	A_Variable	Levels	Datatype	Min Length	Max Length	Level_Values
0	host_identity_verified	2	int32	1	1	{1: 15086, 0: 14935}
1	neighbourhood	213	float64	15	17	{611.3635240839852: 2429, 614.1483704974271: 2...
2	lat	13502	float64	4	18	{40.71813: 14, 40.76625: 13, 40.67293: 12, 40...
3	long	10857	float64	5	18	{-73.9398: 18, -73.95419: 16, -73.95528: 15, -...
4	instant_bookable	3	float64	3	19	{1.0: 15014, 0.0: 14995, 0.4978038496076211: 12}
5	construction_year	21	float64	6	18	{2006.0: 1577, 2019.0: 1561, 2009.0: 1559, 202...
6	minimum_nights	93	int32	1	4	{1: 7653, 2: 7208, 3: 4892, 30: 2297, 4: 2114,...
7	number_of_reviews	324	int32	1	3	{0: 5716, 1: 2995, 2: 2052, 3: 1590, 4: 1226, ...
8	last_review	1530	float64	7	16	{14248.8466196111: 5703, 14230.0: 835, 14238.0...
9	reviews_per_month	16	int32	1	2	{0: 19833, 1: 3975, 2: 2683, 3: 1691, 4: 961, ...
10	review_rate_number	6	float64	3	18	{4.0: 6748, 3.0: 6710, 5.0: 6700, 2.0: 6443, 1...
11	calculated_host_listings_count	48	float64	3	18	{1.0: 19795, 2.0: 4086, 3.0: 1707, 4.0: 927, 5...
12	availability_365	427	int32	1	3	{0: 7691, 365: 581, 364: 283, 1: 257, 179: 218...
13	reviews	740	int32	1	4	{0: 19842, 1: 365, 4: 207, 16: 168, 12: 126, 1...
14	neighbourhood_group_Bronx	2	uint8	1	1	{0: 29356, 1: 665}
15	neighbourhood_group_Brooklyn	2	uint8	1	1	{0: 17618, 1: 12403}
16	neighbourhood_group_Manhattan	2	uint8	1	1	{0: 16838, 1: 13183}
17	neighbourhood_group_Queens	2	uint8	1	1	{0: 26497, 1: 3524}
18	neighbourhood_group_Staten Island	2	uint8	1	1	{0: 29775, 1: 246}
19	cancellation_policy_flexible	2	uint8	1	1	{0: 20180, 1: 9841}
20	cancellation_policy_moderate	2	uint8	1	1	{0: 19925, 1: 10096}
21	cancellation_policy_strict	2	uint8	1	1	{0: 19937, 1: 10084}
22	room_type_Entire home/apt	2	uint8	1	1	{1: 15150, 0: 14871}
23	room_type_Private room	2	uint8	1	1	{0: 15824, 1: 14197}
24	room_type_Shared room	2	uint8	1	1	{0: 29347, 1: 674}

The variables in the above table were utilized as predictors or explanatory variables in the performed regression analysis.

Below is the description of all the variables in the datasets. We have used a function that provides a detailed summary of all the independent and dependent variables.

Out[500]:

	A_Variable	Levels	Datatype	Min Length	Max Length	Level_Values
0	id	102058	int64	7	8	{6044940: 2, 6067584: 2, 6077525: 2, 6076973: ...
1	NAME	61281	object	1	248	{'Home away from home': 33, 'Hillside Hotel': ...
2	host id	102057	int64	9	11	{38729751923: 2, 3895711649: 2, 43698780331: 2...
3	host_identity_verified	2	object	3	11	{'unconfirmed': 51200, 'verified': 51110}
4	host name	13190	object	1	35	{'Michael': 881, 'David': 764, 'John': 581, 'A...
5	neighbourhood group	7	object	3	13	{'Manhattan': 43792, 'Brooklyn': 41842, 'Queen...
6	neighbourhood	224	object	3	26	{'Bedford-Stuyvesant': 7937, 'Williamsburg': 7...
7	lat	21991	float64	3	11	{40.76411: 36, 40.71813: 32, 40.76125: 28, 40....
8	long	17774	float64	3	12	{-73.99371: 44, -73.9535: 40, -73.95427: 37, -...
9	country	1	object	3	13	{'United States': 102067}
10	country code	1	object	2	3	{'US': 102468}
11	instant_bookable	2	object	3	5	{False: 51474, True: 51020}
12	cancellation_policy	3	object	3	8	{'moderate': 34343, 'strict': 34106, 'flexible...
13	room type	4	object	10	15	{'Entire home/apt': 53701, 'Private room': 465...
14	Construction year	20	float64	3	6	{2014.0: 5243, 2008.0: 5225, 2006.0: 5223, 201...
15	price	1151	object	3	7	{'206′ : 137,′1,056 ': 132, '$481 ': 129, '...
16	service fee	231	object	3	5	{'41′ : 526,′216 ': 524, '81′ : 519,′177...
17	minimum nights	153	float64	3	7	{1.0: 25421, 2.0: 23604, 3.0: 16113, 30.0: 116...
18	number of reviews	476	float64	3	6	{0.0: 15734, 1.0: 10408, 2.0: 7175, 3.0: 5375,...
19	last review	2477	object	3	10	{'6/23/2019': 2443, '6/30/2019': 2232, '7/1/20...
20	reviews per month	1016	float64	3	5	{0.03: 1666, 0.05: 1490, 1.0: 1455, 0.04: 1270...
21	review rate number	5	float64	3	3	{5.0: 23369, 4.0: 23329, 3.0: 23265, 2.0: 2309...
22	calculated host listings count	78	float64	3	5	{1.0: 63429, 2.0: 14445, 3.0: 6577, 4.0: 3552,...
23	availability 365	438	float64	3	6	{0.0: 23544, 365.0: 2500, 364.0: 1168, 89.0: 7...
24	house_rules	1976	object	3	1001	{'#NAME?': 2712, 'House Rules 1. Check-in is 4...
25	license	1	object	3	8	{'41662/AL': 2}

We have used function to get the count of null values, the house_rules and licence columns had the highest number of null values (> 50000). The above-mentioned columns have been dropped in the subsequent stages. Likewise, the ‘id’ column has been dropped as well.

Empty spaces have been replaced by ‘_’ (underscores) to make it more readable and consistent. The string types have been changed to lower case for better readability. In the minimum nights column, some of the values are negative so to make more sense of the data they have been replaced with ‘0’ wherever it was negative.

‘construction_year' column has been changed to numeric data type, the dollar sign ‘$’ and commas ‘,’ have been replaced with an empty space for better readability of the data in the columns such as ‘service_fee’ and ‘price’.

‘KNN imputer’ has been used to impute values for ‘latitude’, ‘longitude’ and ‘construction_year’ columns to replace the null values.

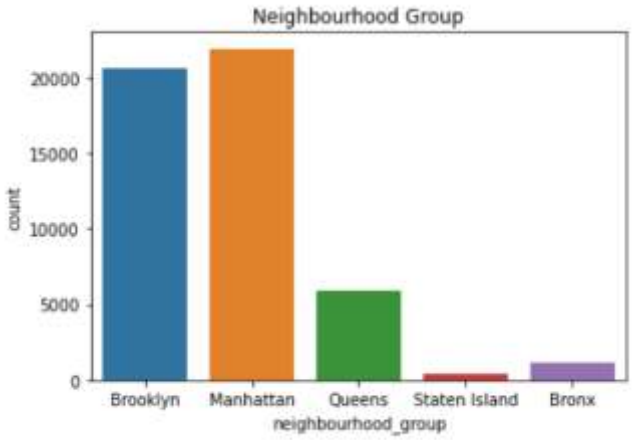
The missing values in reviews_per_month, number_of_reviews and availability column have been replaced with zero using fillna() function, the rows with null values have been dropped in ‘price’ and ‘house_rules’ column.

Duplicates in the data set have been dropped to filter repeated values. The null values in 'review_rate_number'and ‘calculated_host_listings_count’ columns of the abnb DataFrame have been filled with mean value.

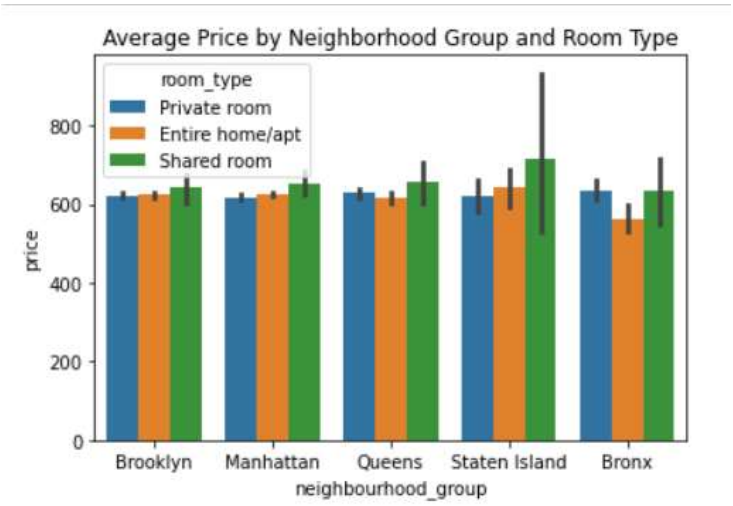
A new column was added into the dataset to account for interaction effect of ‘number_of_reviews’ and ‘reviews_per_month’ columns.

Exploratory Data Analysis

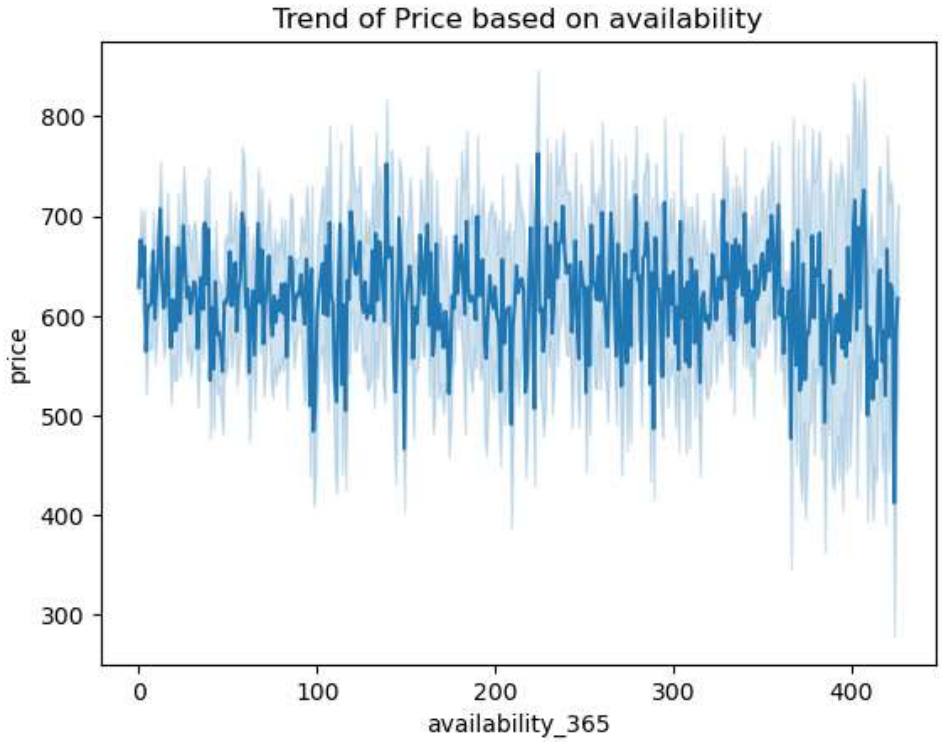
The following characteristics of Airbnb listings and price for New York City are apparent from the data:



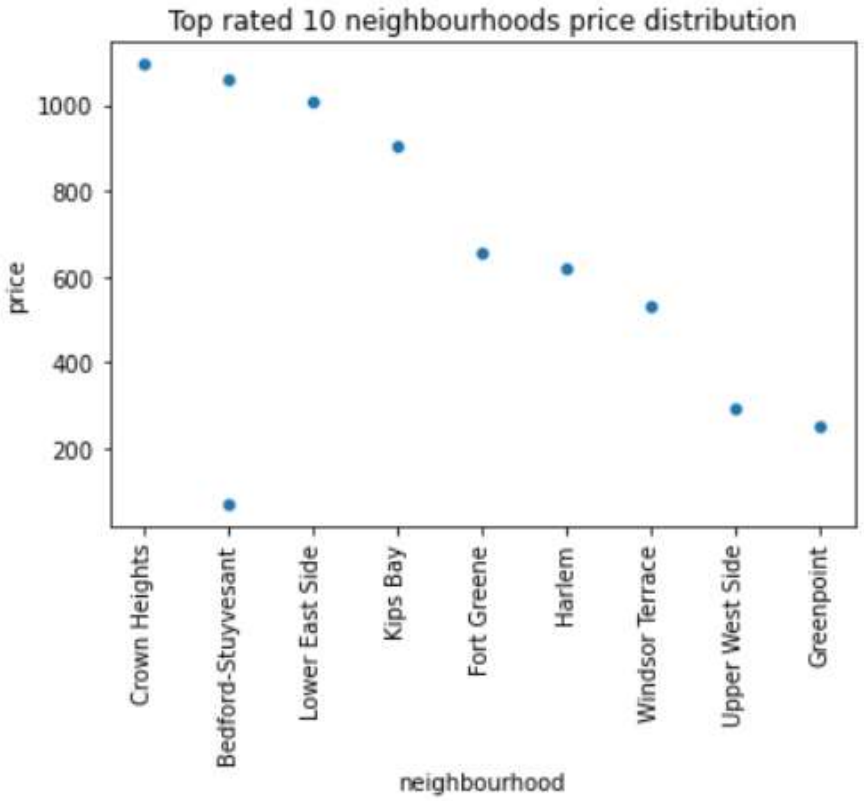
1. The Brooklyn and Manhattan districts boast an impressive inventory of over 20,000 Airbnb properties, affirming their popularity among travelers.



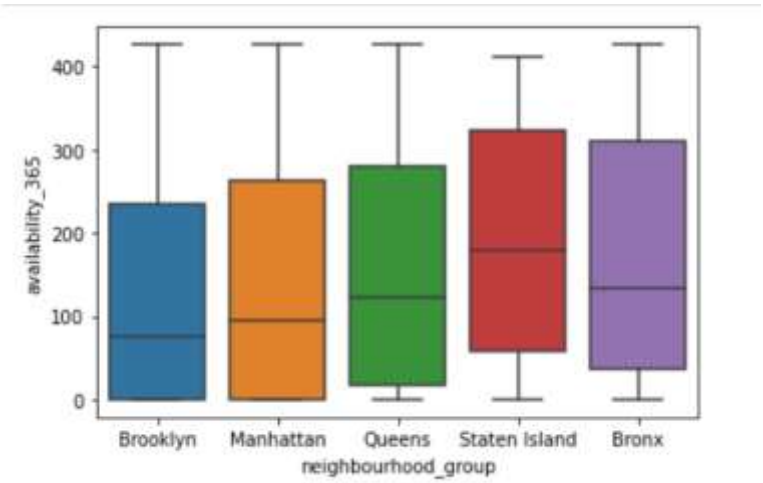
2. Interestingly, shared room rentals in Staten Island command the highest rates. The Bronx presents a unique case where the cost of renting private and shared rooms are on par.



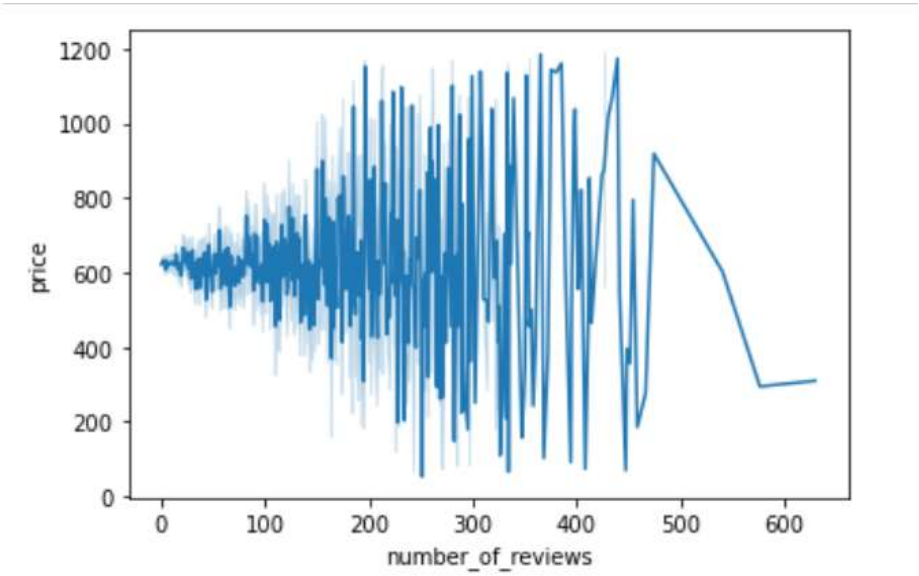
3. A closer look at the yearly availability of listings reveals a positive correlation between rates and hosts who have more than 120 days (about 4 months) of availability. Conversely, year-round availability tends to attract lower rental rates. This could be attributed to the higher demand during peak travel seasons for properties with superior ratings. However, properties with minimal availability might struggle to maintain a consistent customer base.



4. With respect to prices per square foot, Crown Heights, Bedford Stuyvesant, Lower East Side, Kips Bay, and Fort Greene stand out, each commanding prices exceeding $1,000.



5. Staten Island outshines other areas with the highest number of listings available for more than 300 days per year.



6. An intriguing trend is the direct proportional relationship between the price of listings and the number of reviews they garner.

This analysis provides insightful patterns and trends in Airbnb listings across different neighborhoods, which can help both hosts and travelers make informed decisions.

Modelling

In our analysis, we have considered the linear regression model to predict the price, which is our target variable. Linear regression is a normally used and widely understood model for predicting continuous numerical values, making it a suitable choice for this dataset. The linear regression model assumes that there is a linear relationship between the input features (predictors) and the target variable(price). It estimates the coefficients of the predictors to create a linear equation that can predict the target variable based on the given input. This model is interpretable and provides insights into the impact of each predictor on the target variable.

However, after evaluating the results, which showed a low R-squared value of 0.0086, a high mean squared error of 109331.7836, and a mean absolute error of 285.4882, we concluded that the linear regression model did not provide satisfactory predictive performance. As the model did not have

Given these limitations, we explored other models, such as decision trees and K-Means Clustering (KMeans), which might better capture the complex relationships in the data. These models can

potentially provide more accurate predictions and improve our understanding of the important features affecting the price.

Results:

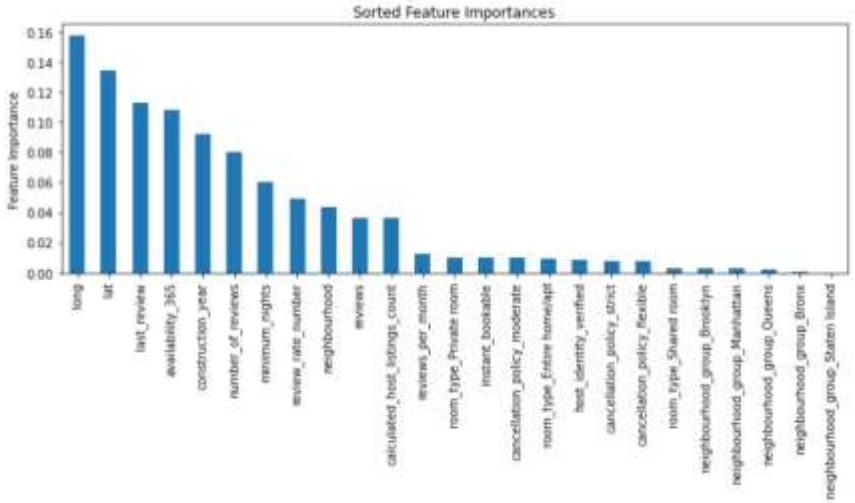
Based on the linear regression results, the model shows a low R-squared value of 0.0086, indicating that only a small portion of the variance in the target variable can be explained by the selected features. The R-squared value measures the goodness of fit of the model to the data, with higher values indicating a better fit. In this case, the low R-squared value suggests that the linear regression model may not be well-suited for capturing the underlying relationship between the features and the target variable.

Additionally, the mean squared error (MSE) value is 109331.7836, which represents the average squared difference between the predicted values and the actual values. A higher MSE value indicates larger prediction errors, further suggesting that the linear regression model may not be accurately capturing the patterns in the data.

The mean absolute error (MAE) value is 285.4882, which represents the average absolute difference between the predicted values and the actual values. This metric provides a measure of the average magnitude of the errors in the predictions. A lower MAE value indicates better accuracy, but the reported value suggests that the linear regression model is still associated with large errors.

Considering these results, it is reasonable to explore alternative models such as decision trees and K-Means Clustering (KMeans) that may better capture the underlying patterns and relationships in the data. These models could potentially provide more accurate predictions and improved performance compared to the linear regression model.

Decision tree Analysis:



We have identified several important features that significantly influence the prediction of price in the dataset. These features include:

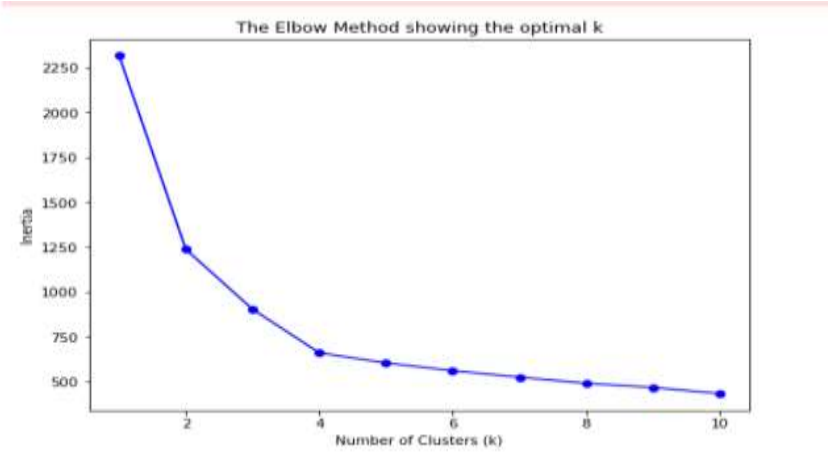
Longitude and Latitude: The geographical coordinates of a property, represented by longitude and latitude, play a crucial role in determining its price. The location of a property is a key factor in real estate valuation. Certain areas or neighborhoods may have higher demand, better amenities, or desirable characteristics, leading to variations in property prices.

Last_review: The last_review feature represents the date of the most recent review for a listing. The timeliness of reviews can impact the perceived value of a property. Listings with more recent positive reviews may be seen as more attractive and reliable, potentially commanding higher prices.

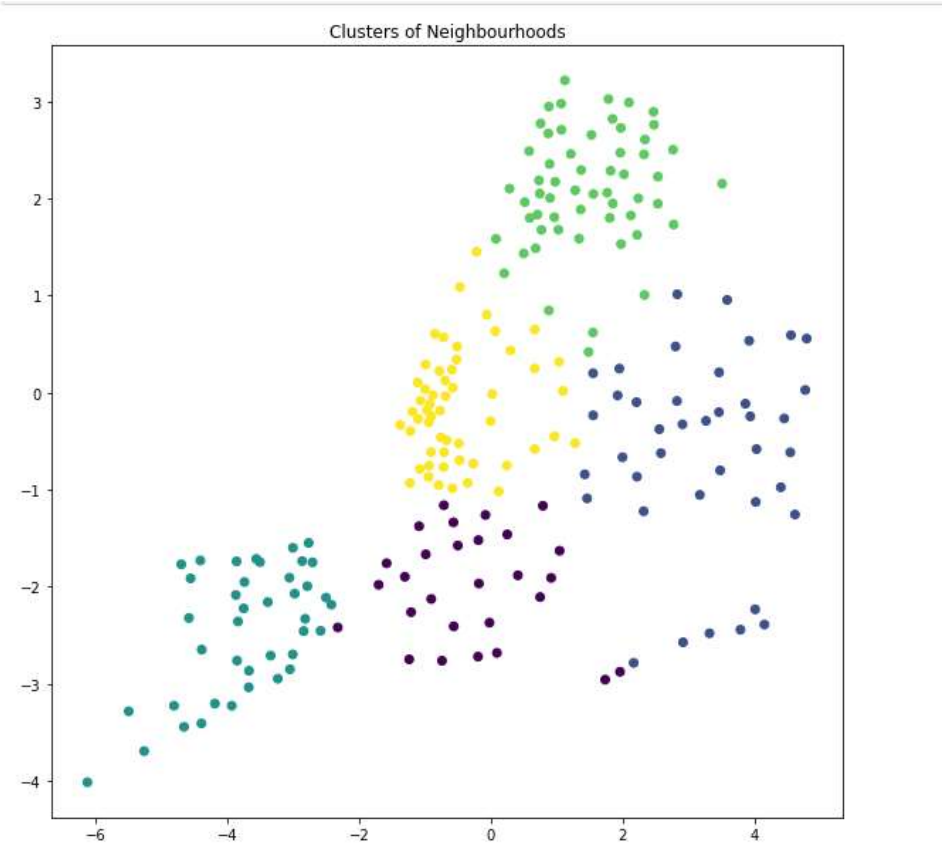
Availability_365: When the availability of the listings over the course of the year was examined, the hosts with availability of more than 120 days had higher rates, while hosts with year-round availability had lower costs. This can be explained by the increased demand for higher-rated Airbnb properties during the busiest travel seasons. Additionally, because such hosts receive frequent visits from visitors, they should often be less accessible throughout certain times of the year. However, having a very low availability could also result in fewer returning clients.

Construction_year: The construction year feature indicates the year when the property was built. The age of a property is an important factor in real estate pricing. Newer constructions often come with modern amenities, updated features, and improved infrastructure, which can justify higher prices compared to older properties.

K Means Clustering:



We decided to go ahead with 5 clusters based on the results from the elbow method.



Cluster 0: The mean price for this cluster is approximately 0.051, which indicates that on average, prices in this cluster are higher than those in the other clusters, given that prices were standardized. The standard deviation (std) of 0.449 suggests a wider spread of prices in this cluster, which could indicate a diverse mix of listings in terms of price. The maximum price in this cluster is the highest among all clusters, reinforcing this observation.

Cluster 1: The mean price for this cluster is approximately -0.014, which is relatively low compared to other clusters. This might represent neighborhoods with more budget-friendly listings. The standard deviation (std) of 0.331 indicates a relatively moderate spread of prices within this cluster.

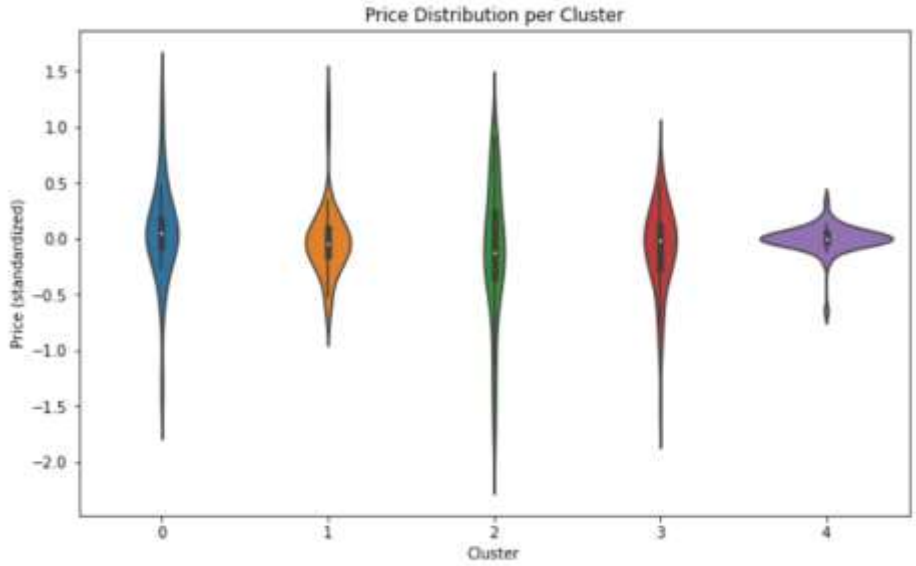
Cluster 2: The mean price for this cluster is approximately -0.113, the lowest among all clusters, which might represent neighborhoods with the most budget-friendly listings. The standard deviation (std) of 0.610 is the highest among all clusters, indicating a wide spread of prices, perhaps due to a mix of budget-friendly and more expensive listings.

Cluster 3: The mean price for this cluster is approximately -0.094, which is relatively low

compared to other clusters. The standard deviation (std) of 0.396 indicates a relatively wide spread of prices within this cluster.

Cluster 4: The mean price for this cluster is approximately -0.002, which is relatively average compared to other clusters. The standard deviation (std) of 0.123 is the lowest among all clusters, indicating a narrow spread of prices within this cluster, perhaps due to more homogeneous listing prices. These interpretations are made under the assumption that the 'price' variable was standardized before clustering, which means that a mean of 0 represents the average price across all listings, and a standard deviation of 1 represents the average variability in price.

Violin plot:



A violin plot combines the benefits of a box plot and a kernel density plot, which is a smoothed histogram. It's a great tool to visualize the distribution of numeric data and its probability density.

Interpretation:

By looking at the violin plot of price across different clusters, we can compare the distribution of price within each cluster. We can identify the median price, see the spread of the price, and identify which clusters have a similar price range and which ones differ.

Cluster 4: This cluster exhibits a broad and thick shape, suggesting that prices within this cluster are quite uniform and tend to be concentrated around a specific value. It has smaller whiskers, indicating fewer price outliers.

Cluster 2: The violin for this cluster is noticeably narrow, indicating a high degree of variability

in prices. Its longer whiskers signify the presence of outliers in the price data.

Cluster 1: The width of this cluster's violin is slightly larger than that of Cluster 2, suggesting a slightly higher uniformity in prices than Cluster 2. A larger upper whisker indicates more outliers are in the higher price range.

Cluster 3 & Cluster 0: These clusters have similar shapes, indicating similar price distributions. Their violins are less wide, suggesting less uniformity in prices. The longer whiskers at both ends indicate the presence of outliers at both the lower and upper ends of the price spectrum.

In interpreting these violin plots, the width of the "violin" at any given point is proportional to the estimated kernel density of prices at that level. In other words, a wider segment corresponds to a higher concentration of prices. The "whiskers" of the plot, typically delineate the overall range of the data. Data points that fall beyond these whiskers are often interpreted as outliers.

References

1. Airbnb's directory of Data for all the cities across USA http://insideairbnb.com/get-the-data/

2. Dataset implemented in project: https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata?select=Airbnb_Open_Data.csv

3. EDA inspiration:

    a. https://matplotlib.org/stable/gallery/index.html

    b. https://www.kaggle.com/code/thecansin/eda-and-data-visualization-ny-airbnb

4. Understanding how the housing prices work: https://www.momondo.com/discover/how-hotel-pricing-works

5. Research paper by Tao Hu and Haoyu Song:Citation on Analysis of factors determining hotel prices