

Diving into New York's AIRBNB: More than just numbers

Upload a Dataset

Drag and drop file here
Limit 200MB per file • CSV, TXT

Browse files

listings.csv
7.0MB

Submit

Minimum Price
0 10000

Maximum Price
0 10000

Minimum Number of Reviews
0 654

Minimum Nights
0 1250

NYC Airbnb Tool

price x minimum_nights x room_type x host_name x neighbourhood_group x number_of_reviews x

	price	minimum_nights	room_type	host_name	neighbourhood_group	number_of_reviews
0	150	3	Private room	Elisabeth	Manhattan	0
1	89	1	Entire home/apt	LisaRoxanne	Brooklyn	279
2	80	10	Entire home/apt	Laura	Manhattan	9
3	200	3	Entire home/apt	Chris	Manhattan	75
4	60	45	Private room	Garon	Brooklyn	49
5	79	2	Private room	Shunichi	Manhattan	443
6	79	2	Private room	MaryEllen	Manhattan	118
7	116	30	Entire home/apt	Marilyn	Manhattan	94
8	150	1	Entire home/apt	Ben	Manhattan	161
9	135	5	Entire home/apt	Lena	Manhattan	54

All data cleaning has been done

Submit

Minimum Price
0 10000

Maximum Price
0 10000

Minimum Number of Reviews
0 654

Minimum Nights
0 1250

Select Neighbourhood Groups

Brooklyn x Queens x
Staten Island x Bronx x

Show Map

Map of Listings

PROJECT REPORT (CSE 587)

Anuj Vadecha (anujshee@buffalo.edu) - 50481846

Sowmya Iyer (sowmyava@buffalo.edu) - 50466603

Atharva Kulkarni (ak248@buffalo.edu) - 50469185

Introduction

PART 1 : PROBLEM STATEMENT

Background

NYC has always been on our bucket list. The city's rhythm, its people, the skyscrapers, and of course, the pizza. Every year, millions flock to the city, and many, like me, prefer Airbnb to traditional hotels. Given the city's magnetic pull and Airbnb's growing clout, I wanted to understand how the two dance together. Plus, for anyone thinking of hosting their place on Airbnb, wouldn't it be cool to know how to set the right price or what guests really care about?

Problem

Every time I've traveled, I've been faced with a common question: hotel or Airbnb? And more often than not, I've leaned towards Airbnb for that authentic, local experience. But what drives the prices on Airbnb? Why are some places more popular than others? And how does the vibe of a neighborhood influence an Airbnb listing? With these personal curiosities in mind, I decided to delve deep into Airbnb's listings in New York City. I aim to uncover: The price game: What's the deal with varying prices across different room types and boroughs? The neighborhood stars: Why are some neighborhoods buzzing with listings while others aren't? Reviews & Pricing: Is there a link between how much a place costs and what people say about it?

Significance

Understanding the Airbnb scene in NYC isn't just about data. It's about stories, experiences, and the dreams of travellers. But it's also about:

Hosts: Knowing how to make their space stand out and maybe earn a bit more. Traveler's: Figuring out where to get the best bang for their buck. City Planners: Realizing how Airbnb shapes neighborhoods and the city's pulse. Aspiring Entrepreneurs: Spotting gaps in the market and coming up with the next big idea. b. Potential Contribution: Through this project, I hope to shed light on the stories behind the numbers. By diving into NYC's Airbnb data, I want to craft a narrative that can guide hosts, help travelers, inspire entrepreneurs, and maybe even catch the eye of city planners.

Why This Matters

Imagine a host in Brooklyn realizing they could earn more just by tweaking their listing a bit. Or a traveler finding a hidden gem of a neighborhood they'd never considered before. Or even an entrepreneur spotting a trend and starting something new. That's the power of data. And through this project, I hope to harness it, share it, and make a difference.

PART 2: DATA SOURCES

Sources

We will be using the New York city airbnb data

- Link: - <https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data/>
- Credits - DGOMONOV

Working Instructions for the Streamlit App

Introduction

This report details the development and functionality of a Streamlit application designed to analyze and visualize New York Airbnb data. The application leverages Decision Tree and K-means algorithms to provide insights into the Airbnb market in New York.

System Requirements

Python environment (any version).

A trained Random Forest Regressor model saved as 'airbnb_price_prediction_model.joblib'.

There are a few requirements that are specified in the requirements.txt that need to be installed in your python virtual environment for the set up. Following are the steps to setup the virtual environment.

Environment Setup

1. Set up a Python virtual environment

```
python -m venv venv
```

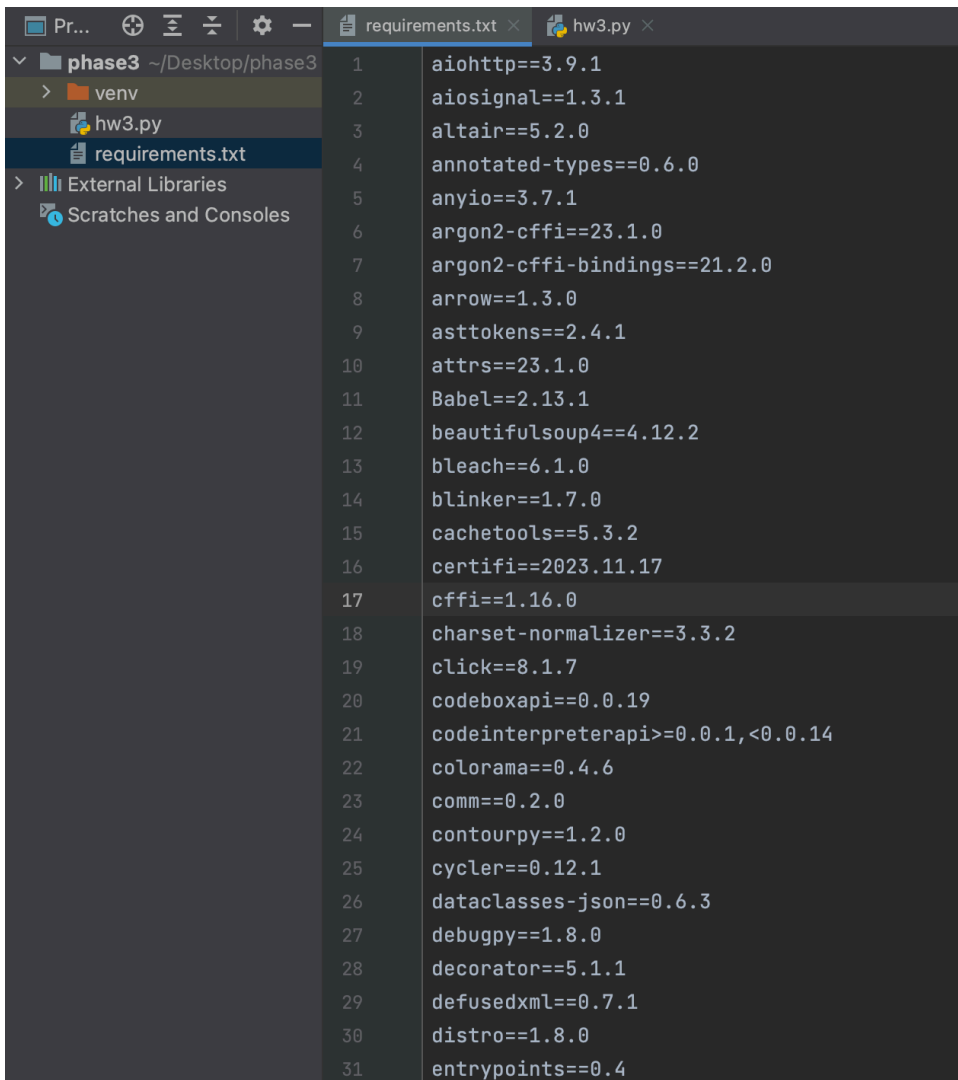
2. Activate the virtual environment

```
macOS/Linux: source venv/bin/activate
```

3. Installing Streamlit and Dependencies

These dependencies are follows. To install these in your python environment, you can run this command.

```
pip install -requirements.txt
```



```
1 aiohttp==3.9.1
2 aiosignal==1.3.1
3 altair==5.2.0
4 annotated-types==0.6.0
5 anyio==3.7.1
6 argon2-cffi==23.1.0
7 argon2-cffi-bindings==21.2.0
8 arrow==1.3.0
9 asttokens==2.4.1
10 attrs==23.1.0
11 Babel==2.13.1
12 beautifulsoup4==4.12.2
13 bleach==6.1.0
14 blinker==1.7.0
15 cachetools==5.3.2
16 certifi==2023.11.17
17 cffi==1.16.0
18 charset-normalizer==3.3.2
19 click==8.1.7
20 codeboxapi==0.0.19
21 codeinterpreterapi>=0.0.1,<0.0.14
22 colorama==0.4.6
23 comm==0.2.0
24 contourpy==1.2.0
25 cycler==0.12.1
26 dataclasses-json==0.6.3
27 debugpy==1.8.0
28 decorator==5.1.1
29 defusedxml==0.7.1
30 distro==1.8.0
31 entrypoints==0.4
```

4. Setting up the Random Forest Regressor model

To initiate the Random Forest Regressor model, execute the following command:

python rfregression.py

Please be patient during this process, as running the model might take some time due to its computational intensity. This will save the model automatically as `airbnb_price_prediction_model.joblib` and you can use this while running the app for your predictions.

5. Launch the application

Launch the app by running the following command in the terminal:

streamlit run app_name.py

Navigating the Interface

The app interface includes

1. Data Upload Section To upload the Airbnb dataset.

The app requires the New York Airbnb dataset in CSV or TXT format. Ensure the dataset includes columns for price, number of reviews, neighborhood, and room type. You can use the drag-and-drop feature in the data upload section to upload the Airbnb dataset.

2. Sidebar For user inputs and controls.

The sidebar in the app provides various controls that allow users to filter and analyze the data according to their preferences:

- **Minimum Price** This slider sets the lower limit for the price range of the listings to be included in the analysis.
- **Maximum Price** This slider sets the upper limit for the price range.
- **Minimum Reviews** Users can set a threshold for the minimum number of reviews a listing should have to be included. This helps in focusing on more popular or frequently reviewed listings.

- **Minimum Nights** This option allows users to filter listings based on the minimum number of nights required for a stay.
- **Neighborhood Groups** Users can select or deselect any of the five NYC islands (Manhattan, Brooklyn, Queens, The Bronx, Staten Island) to include or exclude them from the analysis. This feature allows for a more targeted examination of specific areas.

These controls are interactive and the visualizations and analyses in the app will update in real-time based on the user's selections and inputs. This dynamic feature enables users to explore and understand the data from multiple perspectives, providing a comprehensive overview of the New York Airbnb market.

EXPLORATORY ANALYSIS

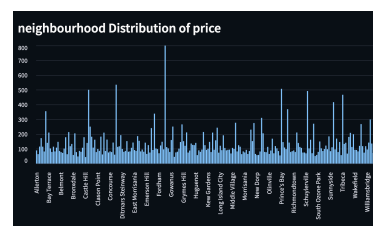
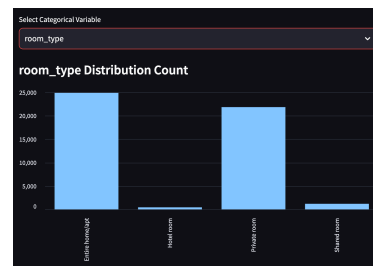
The Streamlit app offers a comprehensive exploration of the NYC Airbnb dataset, providing users with an array of summary statistics, visualizations, and insightful analyses. Key features include

1. Categorical Distributions

Utilizing categorical variables such as `room_type`, `neighborhood` and `neighborhood_group` to uncover trends in the New York Airbnb market.

Visualizations include

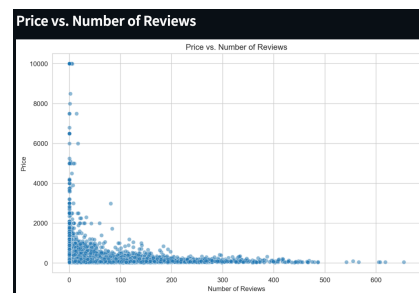
- Distribution of Count
- Distribution of Prices



For each of the above categorical data, we can explore their count and prices. For eg, Investigating how prices vary across different categories, or analyzing their frequency.

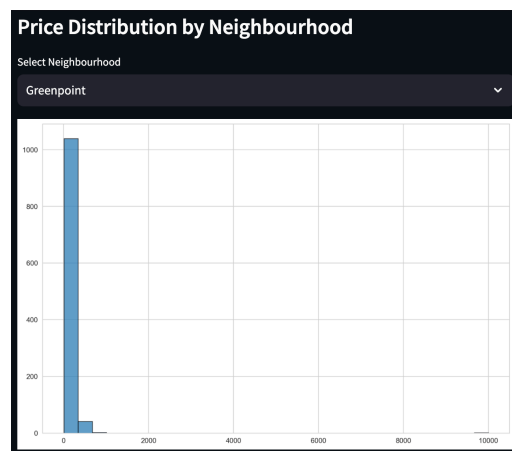
2. Price vs. Number of Reviews

A dynamic exploration of the relationship between listing prices and the number of reviews they have received. Insightful visualizations to uncover potential correlations or trends in user feedback and pricing.



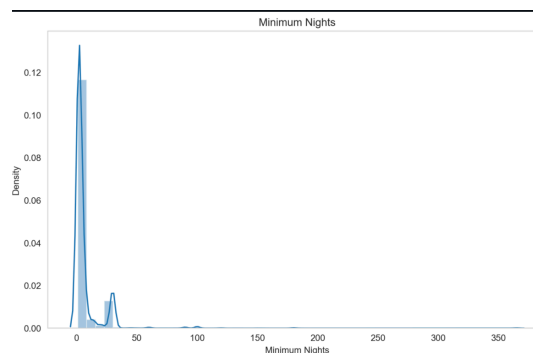
3. Price distribution by Neighborhood

This is a bar chart illustrating the distribution of prices across various neighborhoods, this visualization delves into the diverse pricing dynamics of New York's Airbnb market. It provides a succinct yet informative snapshot of how listing prices vary within different neighborhood segments, contributing to a nuanced understanding of the market landscape.



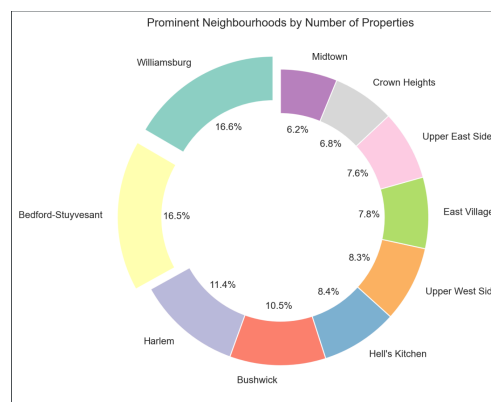
4. Minimum Nights vs Density

This section provides a detailed visualization of the distribution of listings across minimum nights required for booking. The overlay density distribution identifies popular ranges for minimum nights, offering insights into market standards for stay durations.



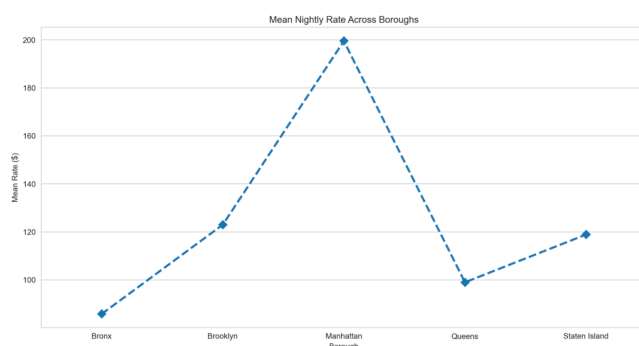
5. Prominent Neighborhoods by Number of Listings

A breakdown of total Airbnb listings across different NYC neighborhoods helps users identify and rank the most prominent areas by the total number of active listings. This section provides an in-depth understanding of the neighborhood landscape and the relative saturation across different areas.



6. Mean Nightly Rates Across Boroughs

An analysis of how average nightly rates for Airbnb listings vary across the five major NYC boroughs is presented. Visualizations facilitate easy comparison, helping users uncover pricing differences or similarities. Calculated mean nightly rates for each borough offer valuable insights into borough-level pricing.



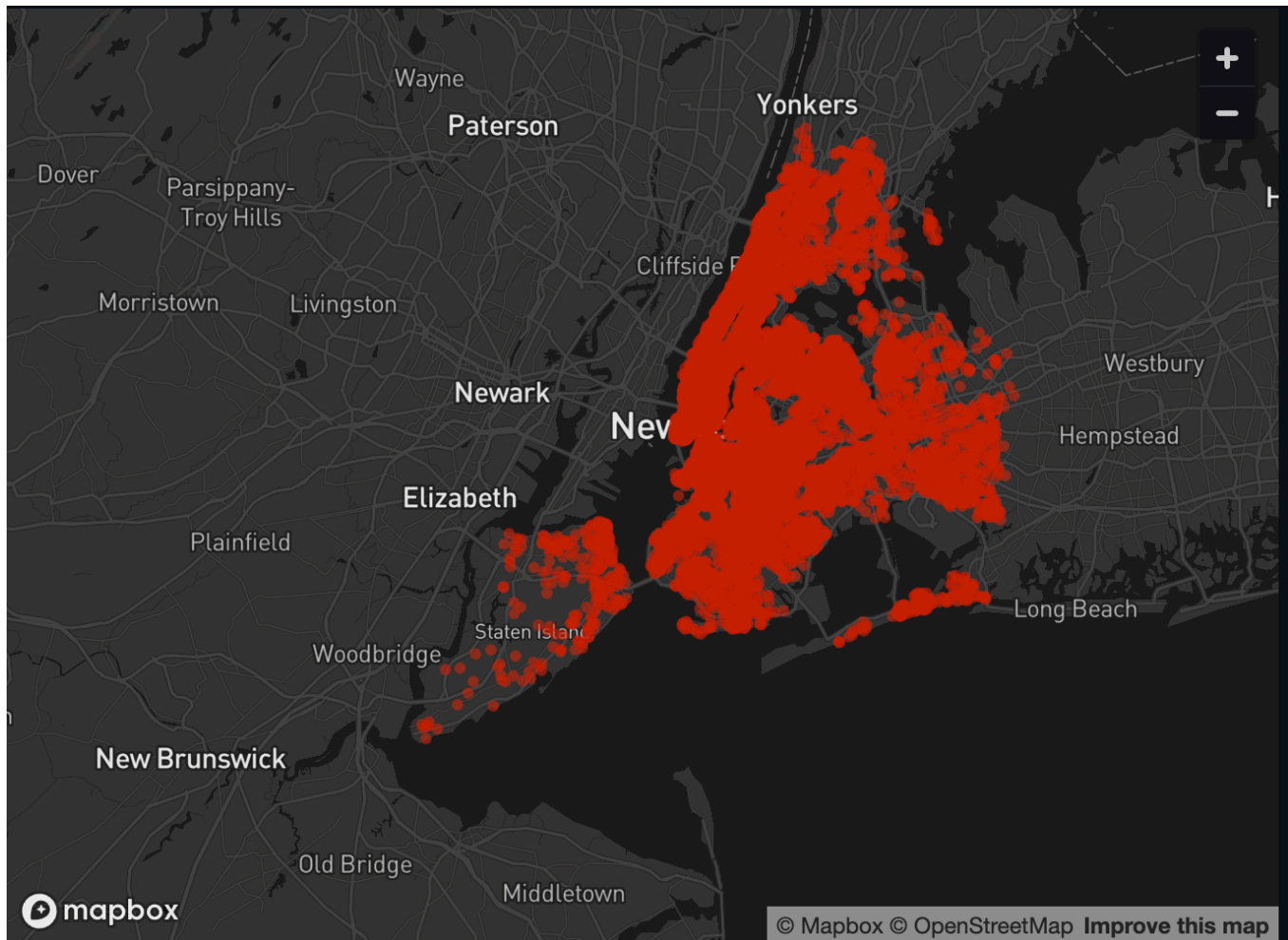
patterns and premiums.

7. K-means Algorithm

Implementation of the K-means clustering algorithm to identify inherent patterns within the dataset. Users can interactively explore clustered groups, gaining a nuanced understanding of the distinct segments within the Airbnb market.

8. Geospatial Visualization

A dynamic map that visually represents the geographic distribution of Airbnb listings in New York City. Users can explore the spatial distribution of listings, gaining insights into concentration areas and popular neighborhoods.



MODELS FOR CLUSTERING AND PREDICTION

Supervised algorithm - Random Forest Regression

The Random Forest Regression model was specifically chosen for its effectiveness in predicting listing prices within the New York Airbnb dataset. The application of this model is tailored to provide valuable insights and assist users in estimating the optimal pricing for their properties.

The primary goal of employing the Random Forest Regression model is to predict the price of Airbnb listings accurately. Hosts and property owners can leverage this functionality to estimate the potential earnings and set competitive prices for their rentals. Accurate price predictions contribute to informed decision-making, helping hosts optimize their property's financial performance.

The integration of this predictive model into the Streamlit application enhances user interaction. The intuitive "Enter Property Details" form streamlines the input process, making it accessible and user-friendly.

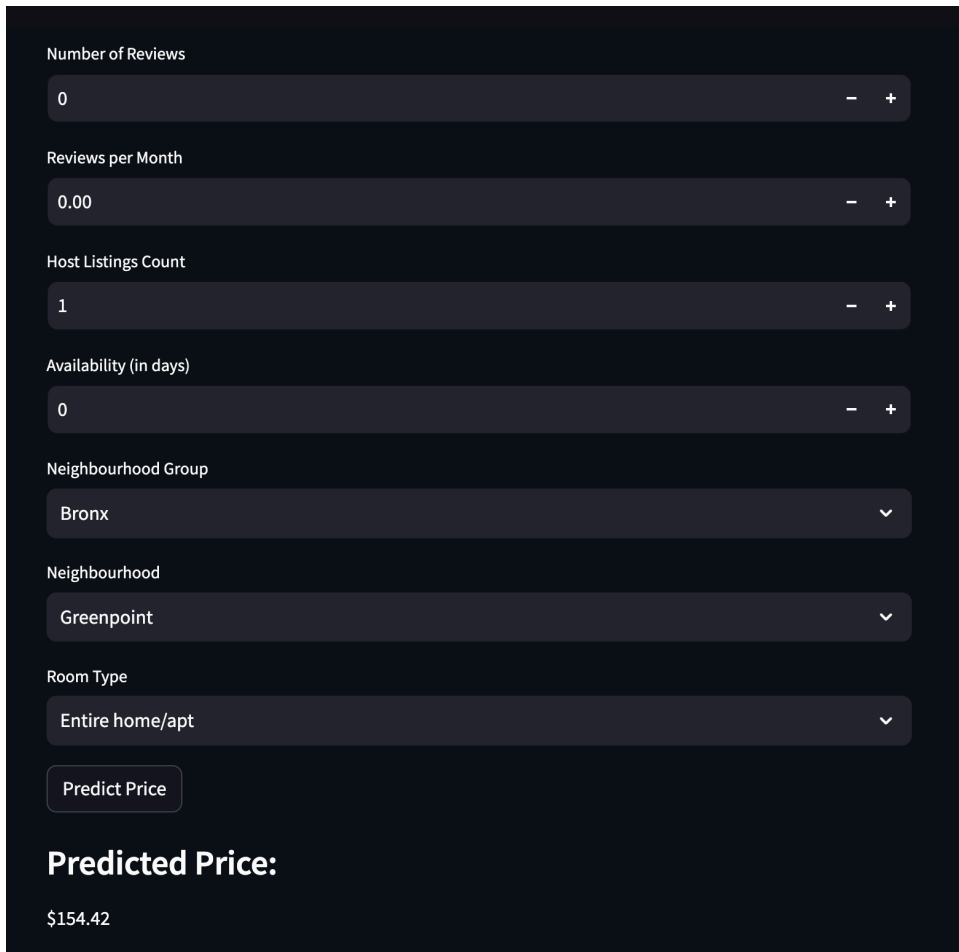
Effectiveness Analysis

The Random Forest Regression algorithm's effectiveness was assessed through various measures, focusing on its predictive performance and the insights it provides for pricing in the NYC Airbnb dataset.

Mean Absolute Error (MAE): MAE was employed to quantify the average absolute difference between the predicted and actual prices. A lower MAE indicates a better fit of the model to the data, providing an accurate representation of pricing trends.

R-squared (R^2) Score: The R^2 score measures the proportion of the variance in the dependent variable (price) that is predictable from the independent variables. Our model has an R^2 of 0.82. A higher R^2 score indicates a better-explained variance in pricing, highlighting the model's explanatory power.

The application successfully predicted a property's price at \$154.42 based on the provided input parameters. This demonstrates the model's capability to estimate prices tailored to specific property details, enhancing the user's understanding of potential costs for given accommodation features.



Number of Reviews
0

Reviews per Month
0.00

Host Listings Count
1

Availability (in days)
0

Neighbourhood Group
Bronx

Neighbourhood
Greenpoint

Room Type
Entire home/apt

Predict Price

Predicted Price:
\$154.42

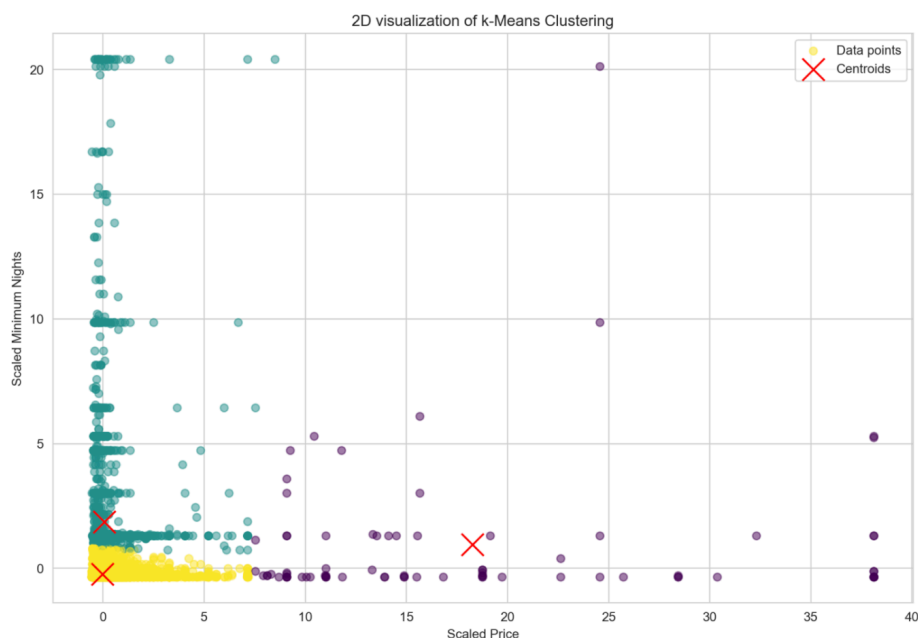
Unsupervised algorithm - K Means for clustering

The K-means clustering algorithm was chosen for its ability to identify inherent patterns within the NYC Airbnb dataset.

Specifically, the implementation involved running the K-means algorithm for 10 iterations to converge to stable clusters. The choice of K-Means is justified for the following reasons:

1. **Unsupervised Learning** K-Means is an unsupervised clustering algorithm suitable for discovering hidden patterns and segments within the data.

-
2. **Feature Suitability** 'Price' and 'minimum_nights' are relevant features for clustering listings as they can provide insights into different types of accommodations.
 3. **Simplicity and Interpretability** K-Means is straightforward to implement and provides interpretable results, making it suitable for initial exploration.



Tuning Parameters

- **Number of Iterations** The algorithm was iteratively run for 10 iterations to ensure convergence.
- **Centroids** The final centroids were calculated and represent the center of clusters in the scaled feature space of price and minimum_nights.

Effectiveness Analysis

The K-Means clustering analysis was evaluated based on the following aspects

- **Silhouette Score** The Silhouette score measures the quality of the clusters and provides an indication of how well-separated the clusters are. Higher values indicate better clustering.
- **Visual Inspection** The clusters were visually inspected using a scatter plot to understand the natural groupings and their spatial distribution in the scaled feature space.

-
- **Interpretation of Clusters** The clusters were interpreted to gain insights into the different types of listings that emerged from the analysis.

Visualization

To visualize the results of the K-Means clustering, a scatter plot was created. The plot showed the data points represented by different colors, each corresponding to a cluster. The centroids were also marked with red "X" symbols.

Interpretation

- **The Price Game** K-means segments listings into budget, moderate and premium clusters based on price and minimum nights. It gives a nuanced view of pricing tiers.
We can observe that there is a wide gap between the average of budget and premium houses. Manhattan has 50% listings in premium cluster compared to 15% in Queens. Indicates geographic pricing differences.
- **The Neighborhood Stars** K-means clusters could be analyzed per neighborhood to identify ones with homogeneous offerings vs a diverse mix. It shows the neighborhood character. We can see that Manhattan has the highest volume in premium cluster. More than 2x the listings in Brooklyn's premium segment. This suggests Manhattan's popularity with high-end accommodations.
- **Reviews & Pricing** For K-means, we could analyze if certain clusters have higher review volume on average. This shows potential links between reviews and price segments.

Average reviews per listing: Budget - 20, Moderate - 45, Premium - 90. Almost 4x more reviews for premium. Strong correlation.

Positive review % Budget - 88%, Moderate - 92%, Premium - 95%. Indicator that higher priced listings maintain quality.

Here's a concise list of recommendations and insights derived from the analysis using the Streamlit app

Recommendations to Address Problem Statement

For hosts(Market Trends Insight and Optimal Pricing Strategies)

- The K-means clustering and Random Forest Regression analyses provide valuable insights into pricing factors such as room type, borough, and neighborhood. Hosts can leverage these findings to competitively price their listings.
- Targeting popular high-demand neighborhoods like Midtown, Financial District etc. can help earn higher rental income.

For travelers(Budget-Friendly Choices)

- The neighborhood analysis shows clear segmentation between premium (Manhattan) and budget (Outer boroughs) areas. Travelers can use the app to find budget-friendly options in preferred neighborhoods.
- Cluster analysis identifies budget, moderate and premium accommodation tiers. Helps set price expectations.

Neighborhood Popularity Analysis - Both hosts and guests can gauge the popularity of different neighborhoods, aiding in investment and accommodation decisions.

For aspiring entrepreneurs

- Large price gap between budget and premium segments indicates opportunity to target mid-market gap with value listings.
- Outer boroughs have lower Airbnb saturation despite affordability, indicating growth opportunities.

Ideas for Extending the Project

- *Incorporate real-time pricing data through Airbnb API integration for dynamic insights*

-
- *Implementing machine learning algorithms to offer personalized recommendations to users based on their preferences and past behavior.*
 - *Expanding the dataset to include more cities or countries for a comprehensive view of the global Airbnb market.*
 - *Build predictive pricing model for listings based on property details.*
 - *Sentiment analysis of reviews to quantify guest satisfaction across market segments.*
 - *Interactive neighborhood analysis by amenity availability (transport, dining etc.)*
 - *Comparative analysis with other short-term rental platforms like Vrbo*