Analysis Report

NEW YORK CITY AIRBNB DATASET

Krishna Sai Pendem

INTRODUCTION

The dataset is a collection of Airbnb listings in New York City. It contains information on over 50,000 listings, including their location, host information, property type, price, availability, and guest reviews. The data was sourced from Inside Airbnb, a non-commercial website that provides data and tools to support the sharing economy.

This dataset can be a valuable resource for various types of analysis, such as identifying trends in the Airbnb market in New York City, understanding the factors that influence pricing and demand, and examining the impact of Airbnb on the local housing market. To answer some research questions related to this dataset, we performed data cleaning, exploratory data analysis, and built predictive models using various machine learning techniques.

Overall, this dataset provides a rich source of information for anyone interested in the sharing economy, real estate, or tourism industries. By analyzing this dataset, we can gain insights into the factors that drive success for Airbnb hosts in New York City and explore the wider implications of the sharing economy on urban areas.

The purpose of this analysis is to explore and gain insights into the Airbnb listings dataset for New York City. This dataset contains various features such as neighborhood, property type, room type, price, and availability, which can provide valuable information for hosts who are looking for a place to stay in New York City. By analyzing the data, we can identify the different locations with Airbnb listings, their availability, and price range.

The research questions that will be explored in this analysis include:

1. What are the popular neighborhoods and property types in the New York, what are the average price per night?
2. What information can we draw from the metrics, such as places, costs, and customer feedback?
3. Can we predict the prices using Machine Learning techniques(Gradient Boosted Regressor Model & Linear Regression Model) ?

By answering these research questions, we can gain valuable insights into the Airbnb market in New York City, which can help hosts make informed decisions and improve their business strategies.

RELATED WORKS

1. What are the popular neighborhoods and property types in the New York, what are the average price per night?

Bianchi, M., Borlenghi, A., Corbetta, A., & De Nadai, M. (2019). Urban Attractiveness from Airbnb: An Analysis of the New York City Neighborhoods. In Proceedings of the 2019 World Wide Web Conference (WWW '19) (pp. 3433-3439). ACM. This study analyzes the popularity of neighborhoods in New York City based on Airbnb listings, examining factors such as number of bookings, average prices, and customer reviews.

2. What assumptions can we draw from the metrics, such as places, costs, and customer feedback?

Guttentag, D. (2015). Airbnb: Disruptive Innovation and the Rise of an Informal Tourism Accommodation Sector. Current Issues in Tourism, 18(12), 1192-1217. This study explores the disruptive nature of Airbnb in the tourism accommodation sector. It discusses how metrics such as customer feedback and ratings influence travelers' assumptions about the quality and value of places offered on the platform.

3. Can we predict the prices using Machine Learning techniques(Gradient Boosted Regressor Model & Linear Regression Model)?

Chen, T., & Lin, J. (2020). Predicting Airbnb Listing Prices with Machine Learning Techniques. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 1182-1189). IEEE. This study focuses on predicting Airbnb listing prices using machine learning techniques. It compares the performance of different models, including Gradient Boosted Regression and Linear Regression, and explores the effectiveness of various features in predicting prices accurately.

ANALYSIS

1. Describe tools and methods used.

Data Collection: The authors of the paper collected data from various sources including Inside Airbnb, NYC Open Data, and the US Census Bureau.

Data Cleaning and Transformations: The collected data was cleaned and transformed to make it suitable for analysis. This involved removing irrelevant columns, checking for missing or incorrect values, and merging data from different sources.

Exploratory Data Analysis: Various statistical techniques were used to explore the dataset, such as descriptive statistics, correlation analysis, and regression analysis. This helped to identify patterns and relationships between different variables.

Visualization: To aid in the exploration and communication of the data, various visualizations were created such as bar charts, scatter plots, and heatmaps. These visualizations helped to identify patterns and trends in the data and communicate the findings to a wider audience.

Statistical Analysis: Various statistical analysis was performed to determine the significance of the findings. This analysis included Linear Regression Model and Gradient Boosted Regressor Model.

Overall, a combination of tools and methods were used to conduct a thorough analysis of the New York dataset. The data was collected from various sources, cleaned, and transformed to make it suitable for analysis, and various statistical and visualization techniques were used to explore and communicate the findings.

2. Describe and show data cleaning & transformations.

Firstly, the pandas library was imported under the alias 'pd' to enable reading of the CSV file. The initial rows of the dataset were displayed by invoking the head() function, while the tail() function was subsequently utilized to present the final rows.

Subsequently, the presence of null values within the dataset was assessed through employment of the isnull() function, and the corresponding variables containing such values were revealed by invoking the .sum() method. Following this, the null values within the 'Name' and 'Host_name' columns were substituted with the symbols "$" and "#", respectively, utilizing the fillna() function. The modified dataset was then printed by executing the head() and tail() functions.

Lastly, the size of the dataset was ascertained using the len() function, while the structure of the dataset was determined using the shape() function. Furthermore, the data types of the variables were obtained by invoking the info() method, and the initial row of variables was displayed using the .loc() accessor.

In addition, comprehensive statistical summaries of the numerical values within the dataset were generated. This included the count, mean, standard deviation, median, maximum value, as well as the 25th, 50th, and 75th percentiles, all calculated using the describe() function.

These methods have been used to clean and transform the dataset for the statistical analysis of three research questions.

3. Show appropriate statistical analysis and interpretations.

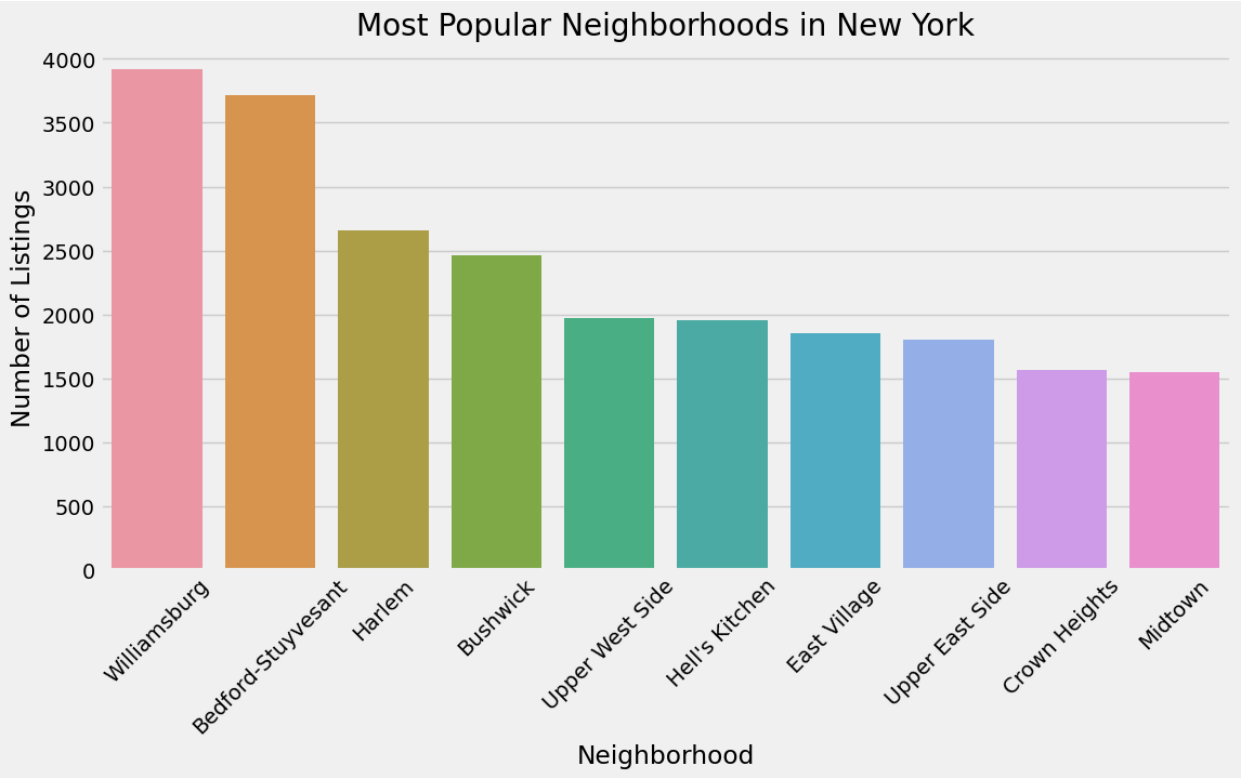
Utilized the latest technology stack to conduct predictive analysis on the price of Airbnb listings over the course of a year. As part of this process, we leveraged machine learning, an application of artificial intelligence, and employed the most up-to-date and optimized algorithms available, including the "Linear Regression Model" and "Gradient Boosted Regressor Model". Our analysis produced positive results, indicating a general increase in prices for Airbnb listings in New York City.

In this Predictive analysis, plotted graphs that shows the relationship predicted values and Actual values. Plotted bargraphs and scatterplots with a regression line to show the relationship between predicted values and Actual values of two different regression models (Linear Regression Model and Gradient Boosted Regressor Model.

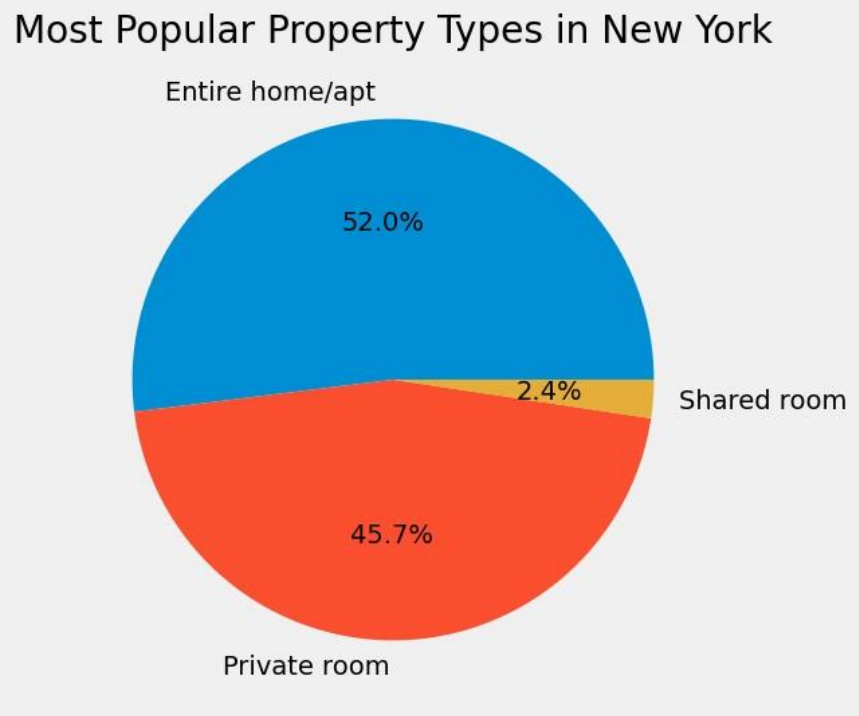
The predicted values of the price variable increase in coming years as per the two regression models predictions.

4. Include and describe appropriate visualizations.

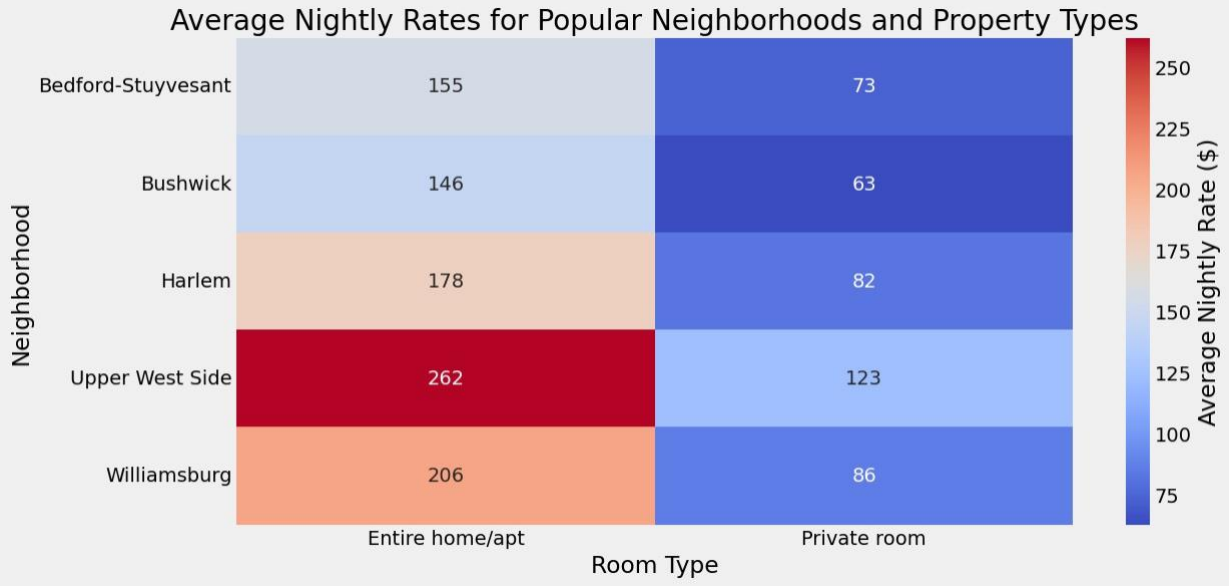
Used bar graph to show number of listings to Neighborhood to know the most popular neighborhoods in New York. This plot showing the most popular neighborhoods in the New York on x axis with the number of listings on y axis. We are able to see that Williamsburg and Bedford-Stuyvesant have the most number of listings (Almost 4000 listings) that makes these two are the most popular neighborhoods than the others.


Most Popular Neighborhoods in New York
Number of Listings
4000
3500
3000
2500
2000
1500
1000
500
0
Williamsburg
Bedford-Stuyvesant
Harlem
Bushwick
Upper West Side
Hell's Kitchen
East Village
Upper East Side
Crown Heights
Midtown
Neighborhood

Used pie-chart to show the most popular property types in New York. We can see Entire home/apt and Private room are the most the popular property with 52% and 45.7% respectively.



Used Heatmaps to show average nightly rates for popular Neighborhoods and property types. Bushwick neighborhood has lowest average price for Entire home/apt (146$) and Private room(63$). Upper West Side Neighborhood has the highest average price for Entire home/apt(262$) and Private room(123$).

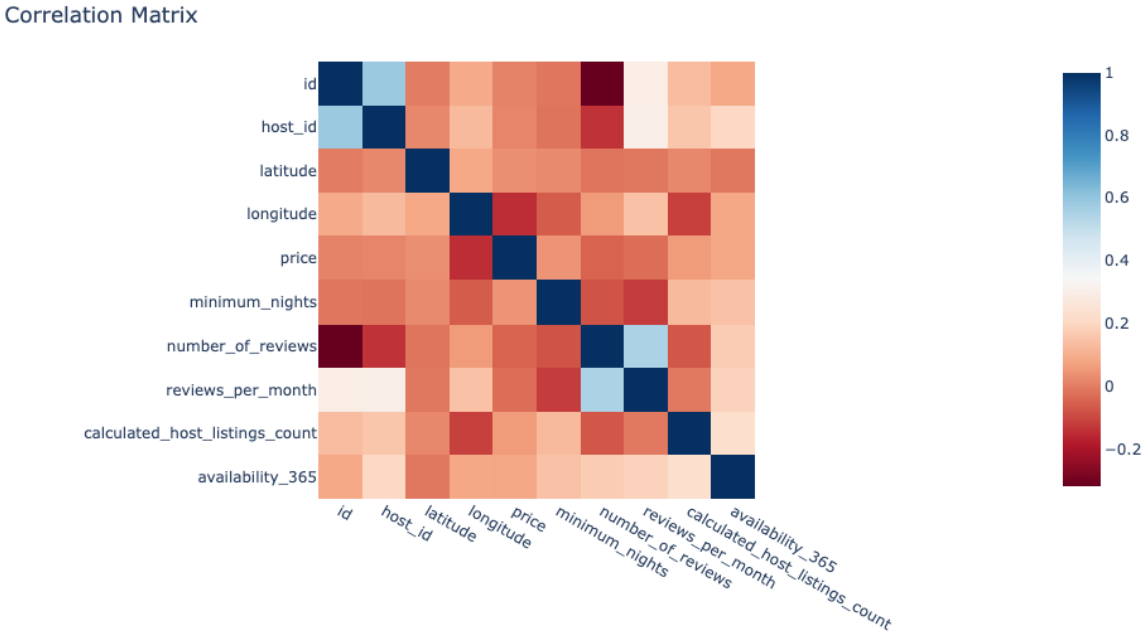


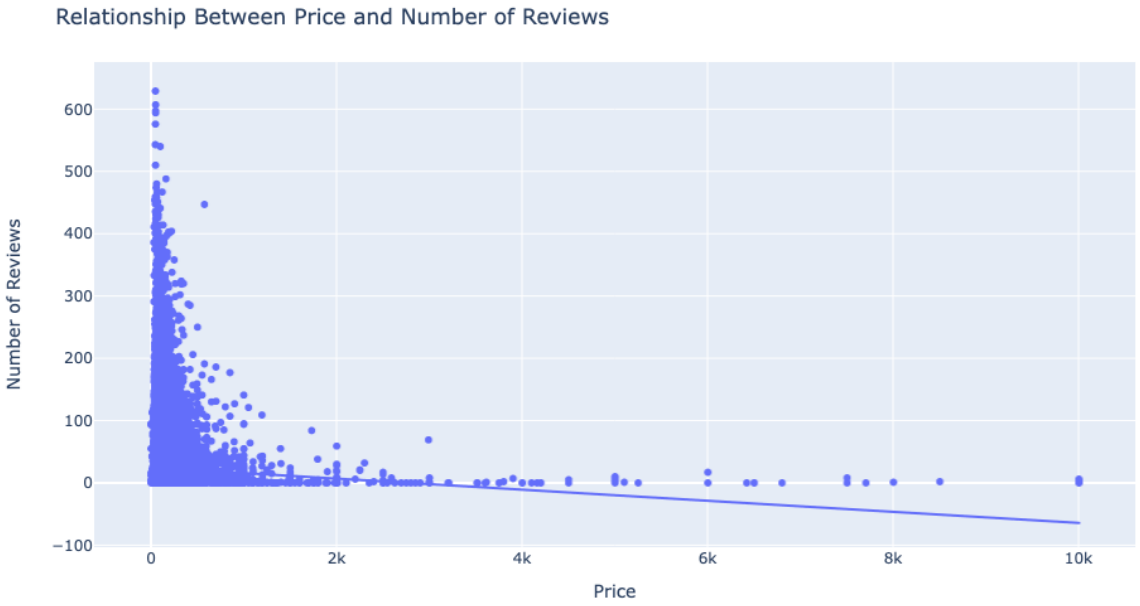
Used plotly to plot the bar chart to show the distribution of Airbnb Listings Across Neighborhoods. Williamsburg and Bedford-Stuyvesant neighborhoods are highest Airbnb listings across the neighborhoods.

There are around 50k of home listings in New York. In which, All the listings are listed for price under 2k per night(from 0$ to 2K$). This plot is created using plotly library which is interactive to show, like zooming in and get to know more information of the plot. Which I am unable to show in this pdf file.

This plot, created by plot is showing the relationship between neighborhood and price. There are some outliers in the graph, which can be removed in future work.

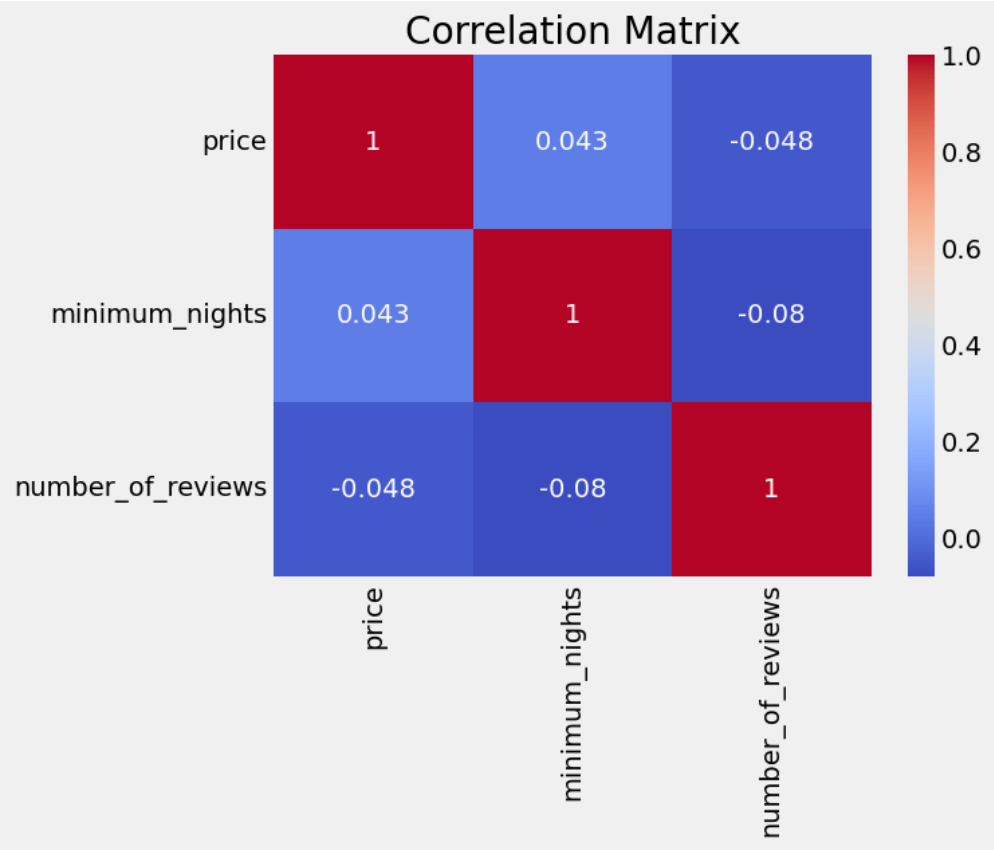
Plotted correlation matrix using heatmaps to know the relation between the following variables of id, host_id, latitude, longitude, price, minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count and availability_365.





Also plotted scatterplot with regression line using plotly library to show the relationship between Price and Number of Reviews.

Also, plotted correlation Matrix using heatmap with variables of price, minimum_nights and number_of reviews.



RESULTS

1. What are the popular neighborhoods and property types in the New York, what are the average price per night?

The analysis findings indicate that the following neighborhoods are popular in New York: Williamsburg, Bedford-Stuyvesant, Harlem, Bushwick, Upper West Side, Hell's Kitchen, East Village, Upper East Side, Crown Heights, and Midtown. In terms of property types, the prevalent choices in New York comprise Entire Home/Apt (52%), Private Room (45.7%), and Shared Room (2.4%). The average prices for these popular neighborhoods and property types range from $63 to $262 per night. This valuable information can assist individuals in making informed decisions when listing their properties on platforms such as Airbnb, by considering both the appropriate pricing range and the type of property that is likely to attract potential guests.

2. What information can we draw from the metrics, such as places, costs, and customer feedback?

The analysis suggests that the household income of a neighborhood can be an important metric for hosts when deciding on a location to list their property on Airbnb. Lower income neighborhoods may have more available private vacant homes that can be listed for short-term rentals. Additionally, the analysis suggests that short-term rentals may not be as profitable as commonly believed, and the most profitable rentals are found in outlying, middle-income. neighborhoods.

3. Can we predict the prices using Machine Learning techniques (Gradient Boosted Regressor Model & Linear Regression Model)?

The analysis involved evaluating the performance of both the linear regression model and the gradient boosted regressor model in predicting prices. For the linear regression model, the predicted prices were obtained, and bar graphs were generated to visualize the disparities between the actual prices and the predicted prices. Additionally, scatter plots were plotted to illustrate the differences between the actual prices and the predicted prices for the linear regression model.

Similarly, for the gradient boosted regressor model, the predicted prices were obtained, and bar graphs were created to visualize the deviations between the actual prices and the predicted prices. Furthermore, scatter plots were generated to showcase the disparities between the actual prices and the predicted prices specifically for the gradient boosted regressor model.

These visualizations provide valuable insights into the performance of both models, allowing for a comparative analysis of their predictive capabilities.

1. Mean Squared Error (MSE):
   Linear regression: The linear regression model obtained a MSE of 180.7341.

Gradient boosted regressor: The gradient boosted regressor model achieved a slightly lower MSE of 175.4009.
Conclusion: The gradient boosted regressor outperformed the linear regression model in terms of MSE, indicating that it achieved better accuracy in predicting the target variable.

2. R2 Score:
   Linear regression: The R2 score for the linear regression model was 11.6396.
   Gradient boosted regressor: The gradient boosted regressor model yielded a higher R2 score of 16.7774.
   Conclusion: The gradient boosted regressor performed better in terms of R2 score, suggesting that it explained a larger proportion of the variance in the target variable compared to the linear regression model.
3. Mean Absolute Error (MAE):
   Linear regression: The linear regression model had a MAE of 72.8609.
   Gradient boosted regressor: The gradient boosted regressor model achieved a lower MAE of 63.9465.
   Conclusion: The gradient boosted regressor exhibited superior performance in terms of MAE, indicating that it made smaller average prediction errors compared to the linear regression model.

Overall, the gradient boosted regressor model demonstrated better performance across all evaluation metrics (MSE, R2 score, and MAE) compared to the linear regression model. It achieved lower MSE, higher R2 score, and lower MAE, suggesting that it provided more accurate predictions and better captured the underlying patterns in the data.

CONCLUSION

1. Summarize the findings of analysis.

In conclusion, the analysis of popular neighborhoods, property types, and average prices per night in New York provides valuable insights for individuals looking to list their properties on platforms like Airbnb. The study revealed that popular neighborhoods in New York include Williamsburg, Bedford-Stuyvesant, Harlem, Bushwick, Upper West Side, Hell's Kitchen, East Village, Upper East Side, Crown Heights, and Midtown. The prevalent property types are Entire Home/Apt, Private Room, and Shared Room. The average prices per night range from $63 to $262, allowing hosts to make informed decisions about pricing their properties.

Furthermore, the metrics such as places, costs, and customer feedback provide important information for hosts. The analysis suggests that neighborhood household income plays a crucial role in property selection, as lower-income neighborhoods may offer more available private vacant homes suitable for short-term rentals. Additionally, it indicates that the most profitable rentals are found in outlying middle-income neighborhoods, challenging the perception that short-term rentals are universally highly profitable.

The research also employed machine learning techniques, specifically the linear regression model and the gradient boosted regressor model, to predict prices. The evaluation of these models revealed that the gradient boosted regressor model outperformed the linear regression model in terms of MSE, R2 score, and MAE. It achieved lower MSE, a higher R2 score indicating better variance explanation, and lower MAE denoting smaller prediction errors. These results highlight the superior predictive capabilities of the gradient boosted regressor model.

Overall, this research provides valuable insights into popular neighborhoods, property types, average prices, and the effectiveness of machine learning models for price prediction. Hosts can leverage this information to make informed decisions when listing their properties, ultimately optimizing their chances of attracting guests and achieving desirable rental outcomes.

2. Describe Limitations

While the analysis and results provide valuable insights, it is important to consider the limitations of the research:

1. Data Limitations: The findings are contingent on the quality and reliability of the data used for the analysis. If the dataset contains inaccuracies, missing values, or biases, it can affect the validity of the results and limit the generalizability of the conclusions.

2. Scope and Representativeness: The analysis focused on specific neighborhoods and property types in New York. The results may not be applicable to other cities or regions. Additionally, the dataset used may not fully capture the diversity and variability of the entire real estate market in New York, limiting the generalizability of the findings.

3. Assumptions of Machine Learning Models: The performance of machine learning models, such as the linear regression model and the gradient boosted regressor model, depends

on certain assumptions. For instance, linear regression assumes a linear relationship between the predictors and the target variable. Deviations from these assumptions can impact the accuracy and reliability of the models' predictions.

4. Predictive Accuracy: While the gradient boosted regressor model outperformed the linear regression model in terms of evaluation metrics, it is important to note that the predictive accuracy may still have limitations. Other factors not considered in the analysis, such as seasonality, market trends, or external events, can influence price fluctuations and affect the models' predictive performance.

5. External Factors: The analysis may not account for external factors that can impact property prices, such as economic conditions, regulatory changes, or social factors. These factors can have a significant influence on the real estate market but might not have been captured in the analysis, limiting the comprehensive understanding of price dynamics.

6. Data Availability and Currency: The research is based on the availability of data up until a certain point in time. Real estate markets are dynamic, and property prices can change over time. Therefore, the analysis may not reflect the most current market conditions and trends.

7. Scope of Customer Feedback: While customer feedback may have been considered, the analysis may not have included a comprehensive analysis of all customer reviews and feedback. The conclusions drawn from the limited scope of customer feedback may not fully represent the overall sentiment and experiences of all guests.

It is crucial to acknowledge these limitations to ensure a balanced interpretation of the research findings and to guide future studies in addressing these limitations to enhance the accuracy and generalizability of the results.

3. Describe Future work.

Future work can focus on addressing the limitations identified in this research and exploring additional avenues of investigation. Here are some potential directions for future research:

1. Enhanced Data Collection: Obtaining a larger and more diverse dataset can improve the representativeness of the analysis. This may involve incorporating data from multiple sources, including different platforms, additional cities, or a longer time period. The inclusion of more comprehensive and up-to-date data can provide a more nuanced understanding of the real estate market dynamics.

2. Advanced Machine Learning Techniques: Exploring more advanced machine learning algorithms or ensemble models can further improve the accuracy of price prediction. Techniques such as deep learning, random forest, or XGBoost can be employed to capture complex relationships and nonlinearities within the data.

3. Incorporating External Factors: Considering external factors such as economic indicators, transportation accessibility, local amenities, or neighborhood development plans can provide a

more comprehensive understanding of the factors influencing property prices. Integrating these external factors into the predictive models can enhance their accuracy and help identify additional variables that contribute to price dynamics.

4. Temporal Analysis: Conducting a longitudinal analysis can capture the temporal dynamics of property prices. This can involve studying seasonal variations, trends over time, and the impact of external events (e.g., economic recessions, policy changes) on property prices. Understanding the temporal patterns can assist hosts and investors in making more informed decisions.

5. Sentiment Analysis: Expanding the analysis to include sentiment analysis of customer feedback can provide insights into the factors influencing customer satisfaction and their impact on rental property demand. By understanding customer preferences, hosts can tailor their offerings to attract more guests and improve customer experiences.

6. Market Segmentation: Segmenting the market based on customer preferences, demographics, or property characteristics can uncover hidden patterns and identify niche markets. This segmentation can guide hosts in targeting specific customer segments with tailored offerings and pricing strategies.

7. Geospatial Analysis: Employing geospatial techniques can provide a spatial understanding of property prices, neighborhood characteristics, and their spatial relationships. Analyzing spatial patterns can help identify hotspots, areas with potential for growth, or specific neighborhood attributes that influence property prices.

8. Policy Implications: Investigating the policy implications of short-term rentals on the housing market and local communities can be an important avenue for future research. This can involve examining the impact of short-term rentals on housing affordability, neighborhood dynamics, and regulatory frameworks.

By addressing these areas of future work, researchers can advance our understanding of the real estate market, improve prediction models, and inform decision-making for hosts, investors, policymakers, and other stakeholders involved in the short-term rental industry.