Unlocking Airbnb Success

TEAM 3: Aman Jaglan, Aravinda Vijayaram Kumar, Jiajun Gao, Sashrika Pathuri
DATS 6103: An Introduction to Data Mining, Masters of Science Data Science
, The George Washington University

I. INTRODUCTION

In the landscape of contemporary travel, one name stands out as a transformative force, reshaping the way individuals explore and experience destinations globally - Airbnb. Established in 2008 as a pioneering American corporation, Airbnb swiftly evolved into a revolutionary platform that connects travelers with a diverse array of accommodations hosted by individuals. This paper delves into the core elements that define Airbnb's impact on the hospitality industry and explores how it has fundamentally altered the dynamics of modern travel.

From its humble beginnings, Airbnb has burgeoned into a global marketplace, operating seamlessly across numerous countries. Its success lies in offering not just conventional lodging but a spectrum of unique and unconventional spaces, ranging from apartments and houses to more eclectic choices like treehouses and castles. Airbnb's inception marked a departure from the traditional hospitality model, ushering in an era where travelers seek more personalized and immersive lodging experiences.

The platform's user-friendly interface empowers travelers to transcend the boundaries of conventional hotels, enabling them to explore and uncover the character of local neighborhoods. Airbnb embodies the spirit of choice, providing an extensive range of accommodations that cater to diverse preferences and budgets. This freedom to choose resonates with travelers seeking more than just a place to stay; it's an invitation to be part of a community and embrace the richness of local culture.

For property owners, Airbnb represents an economic opportunity to unlock the value of their spaces. Beyond mere monetization, hosts can showcase the distinctiveness of their properties, establishing a global presence and fostering a personal connection with guests. This connection goes beyond transactions; it's about creating a community where hosts become guides, sharing insider tips and making the travel experience more intimate and memorable.

Airbnb's economic impact reverberates beyond individual hosts, contributing to local economies by boosting tourism and redistributing income to a broader spectrum of communities. As hosts become ambassadors for their locales, Airbnb not only provides a platform for lodging but becomes a catalyst for authentic, community-driven travel experiences.

This exploration into Airbnb's journey unfolds against the backdrop of a paradigm shift in travel preferences. No longer confined to the predictability of hotels, travelers now seek the unique, the authentic, and the personal - qualities that Airbnb has seamlessly woven into the fabric of modern travel. As we navigate the evolution of travel, Airbnb's influence is not just in the spaces it provides but in the stories, connections, and experiences it enables.

The reason for selection of this topic is motivated by Airbnb's substantial influence on the hospitality sector, introducing unique lodging options and transforming the landscape of travel experiences. Investigating the factors influencing pricing and the attributes of high-cost listings is of significant value for both hosts and travelers. Prior research has extensively explored various facets of Airbnb, ranging from its impact on the travel industry to its economic implications for local communities and user satisfaction. This project extends existing

analyses by striving for a deeper comprehension of the factors influencing pricing decisions and the characteristics contributing to premium listings. It aspires to provide nuanced insights through a thorough and comprehensive investigation.

II. DATASET

This study relies on a robust dataset encompassing details from over 250,000 Airbnb listings spanning 10 major cities, presenting a comprehensive view of the platform's global landscape. The dataset, covering host specifics, pricing structures, geographical attributes, and room types, enables a nuanced exploration of the diverse dimensions within the Airbnb marketplace. With its extensive temporal coverage, the dataset facilitates insights into the evolving trends of host performance and guest satisfaction over time. Notably, the geographical dimension opens avenues for analyzing the impact of neighborhood and district factors on pricing dynamics and overall listing popularity.

Further enhancing its depth, the dataset includes essential metrics such as host response time, response rate, and superhost status, contributing to a comprehensive understanding of host behaviors and strategies. Moreover, the dataset captures key elements of guest satisfaction through a spectrum of review scores, including accuracy, cleanliness, check-in experience, communication, location, and value. These multifaceted attributes make the dataset an invaluable resource for researchers seeking to unravel the intricate dynamics of the Airbnb platform, providing insights into host practices, guest preferences, and the overall functioning of this global accommodation-sharing network.

This Dataset -Airbnb Listings & Reviews [1] is obtained from Kaggle.

III. OBJECTIVES

This project dives deep into Airbnb pricing, aiming to uncover what factors influence how hosts set their prices. The primary aim is to identify and analyze the determinants that significantly influence pricing in Airbnb listings. This involves an in-depth exploration of various factors, including location, property size, amenities, seasonality, and local demand, to gain a comprehensive understanding of host pricing decisions based on these determinants.

Simultaneously, the project seeks to unravel the distinctive features shared by the most expensive Airbnb listings. Through careful analysis, correlations with luxurious amenities, unique property features, or exceptional locations will be explored to shed light on the characteristics that contribute to premium pricing. Furthermore, the project aims to empower travelers with cost-effective knowledge, creating a guide or set of criteria to assist them in identifying Airbnb listings that align with their preferences and offer optimal value for money.The ultimate goal is to inform both hosts and travelers comprehensively, contributing valuable insights to the broader understanding of Airbnb's pricing dynamics.

IV. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis(EDA) refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.[2]

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, and find interesting relations among the variables.[3]

We begin by exploring the dataset, checking the columns, checking for null values, dropping unnecessary columns, and dropping the null and infinite values. We also converted all the prices from local currencies to a common price in US dollars. Once data is cleaned, we can see the statistical summary of the data-

Basic Statistics:

	listing_id	host_id	host_total_listings_count	latitude
count	2.493890e+05	2.493890e+05	249389.000000	249389.000000
mean	2.617670e+07	1.073569e+08	22.236658	17.648673
std	1.447029e+07	1.105982e+08	256.809736	32.710678
min	2.577000e+03	1.822000e+03	0.000000	-34.264400
25%	1.365794e+07	1.681687e+07	1.000000	-22.969650
50%	2.736635e+07	5.724332e+07	1.000000	40.690960
75%	3.967077e+07	1.818170e+08	4.000000	41.899010
max	4.834353e+07	3.901874e+08	7235.000000	48.904720

	longitude	accommodates	bedrooms	price
count	249389.000000	249389.000000	249389.000000	249389.000000
mean	12.532654	3.400808	1.515404	636.597188
std	74.207527	2.191669	1.153138	3537.661191
min	-99.339630	1.000000	1.000000	1.000000
25%	-43.215480	2.000000	1.000000	80.000000
50%	2.392090	3.000000	1.000000	160.000000
75%	28.991950	4.000000	2.000000	500.000000
max	151.339810	16.000000	50.000000	625216.000000

	minimum_nights	maximum_nights
count	249389.000000	2.493890e+05
mean	7.952560	2.221107e+04
std	32.004334	6.402950e+06
min	1.000000	1.000000e+00
25%	1.000000	4.400000e+01
50%	2.000000	1.125000e+03
75%	5.000000	1.125000e+03
max	9999.000000	2.147484e+09

Fig 1. Basic Statistics of the data.

After cleaning the data and all the pre-processing done, we have approximately 250,000 rows of data. Also, we can see certain characteristics of the dataset such as mean piece being around $636.59 with a standard deviation of 3537.66, the highest price is $625216 and with the lowest being $1. These are broad and to get a better understanding let us visualize the data.

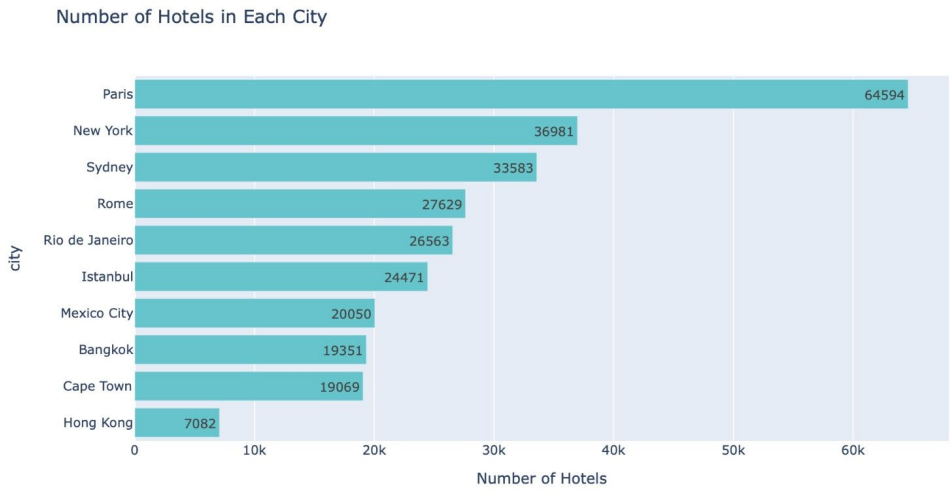

Fig2. Number of Hotels in each city

Notably, Paris emerges as the city with the highest number of hotels, boasting an impressive count of 64,594 establishments. Following closely, New York, Sydney, and Rome showcase substantial hotel figures, with each city contributing significantly to the global hospitality landscape.

Conversely, Hong Kong occupies the opposite end of the spectrum, featuring the lowest number of hotels among the represented cities, numbering 7,082. The intermediary positions are occupied by Rio de Janeiro, Istanbul, Mexico City, Bangkok, and Cape Town, each displaying varying magnitudes of hotel presence. This visual representation effectively communicates the disparity in hotel abundance across the diverse cities, shedding light on the global distribution of hospitality infrastructure.

Fig3 a. Price Comparison and Distribution of Hotels across the city Map

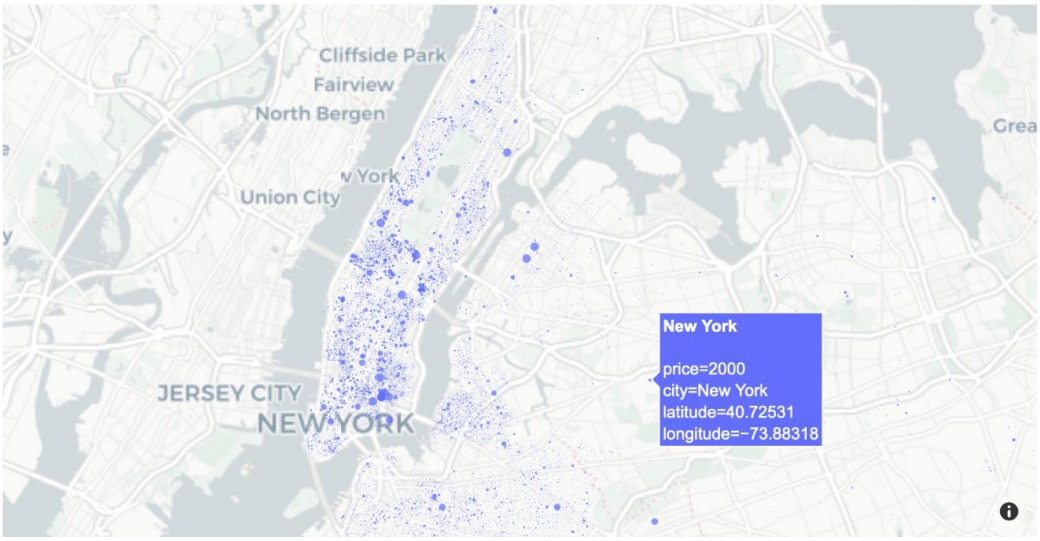

Fig3 b. Price Comparison and Distribution of Hotels across the city Map on zooming in

Fig 3 shows a world map with price distribution across the globe and on zooming in we get the map as shown in image 3b revealing a detailed scatter plot of data points within a section of New York City. Each blue dot on the map represents an Airbnb with their details such as price, location etc. The size of the dot is in correspondence to the price. Higher the price larger the size. This gives us an overview of the price distribution of the Airbnb around the world.

	neighbourhood	count
0	Ward 54	13
1	Copacabana	9
2	Fatih	5
3	Khlong Toei	3
4	Miguel Hidalgo	3
5	Barra da Tijuca	2
6	Botafogo	2
7	Guaratiba	1
8	Santa Teresa	1
9	Lagoa	1
10	Bang Rak	1
11	Ward 100	1
12	Alto da Boa Vista	1
13	Cuauhtemoc	1
14	Din Daeng	1
15	Sao Cristovao	1
16	Iztapalapa	1
17	Bang Khae	1
18	Ward 115	1
19	Ipanema	1
20	Khlong San	1

1. **Copacabana:** Rio de Janeiro, Brazil
2. **Fatih:** Istanbul, Turkey
3. **Khlong Toei:** Bangkok, Thailand
4. **Miguel Hidalgo:** Mexico City, Mexico
5. **Barra da Tijuca:** Rio de Janeiro, Brazil
6. **Botafogo:** Rio de Janeiro, Brazil
7. **Guaratiba:** Rio de Janeiro, Brazil
8. **Santa Teresa:** Rio de Janeiro, Brazil
9. **Lagoa:** Rio de Janeiro, Brazil
10. **Bang Rak:** Bangkok, Thailand
11. **Alto da Boa Vista:** Rio de Janeiro, Brazil
12. **Cuauhtemoc:** Mexico City, Mexico
13. **Din Daeng:** Bangkok, Thailand
14. **Sao Cristovao:** Rio de Janeiro, Brazil
15. **Iztapalapa:** Mexico City, Mexico
16. **Bang Khae:** Bangkok, Thailand
17. **Ipanema:** Rio de Janeiro, Brazil
18. **Khlong San:** Bangkok, Thailand

Fig 4. Neighbourhoods that have both costliest and the cheaper Airbnb

This shows the neighbourhood which has the costlier and cheaper Airbnb and their counts. The chart reveals that "Ward 54" leads the chart with the highest count at 13, followed by "Copacabana" with 9, and "Fatih" with 5. Subsequent neighbourhoods, including "Klong Toei," "Miguel Hidalgo," and "Barra da Tijuca," exhibit counts ranging from 3 to 1.

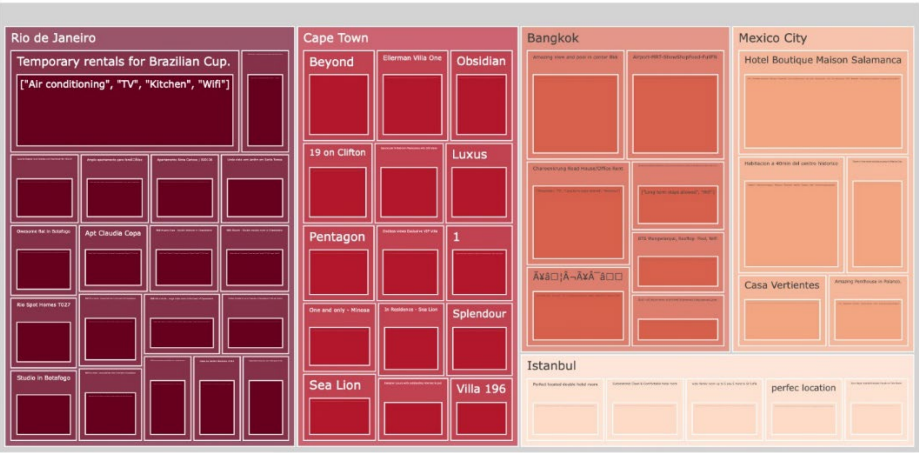

Fig5 a. Hotels that are pricier than 10000 and their amenities

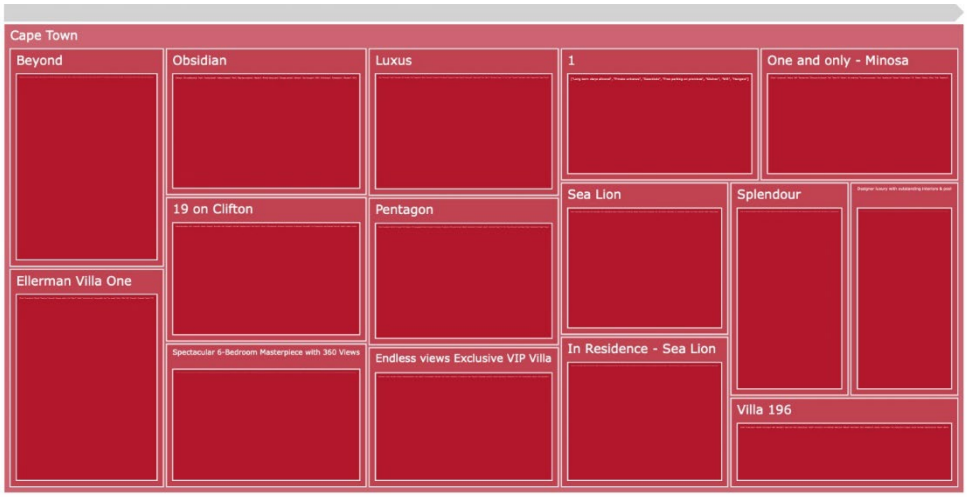

Fig5 b. On clicking a City, Airbnb options displayed

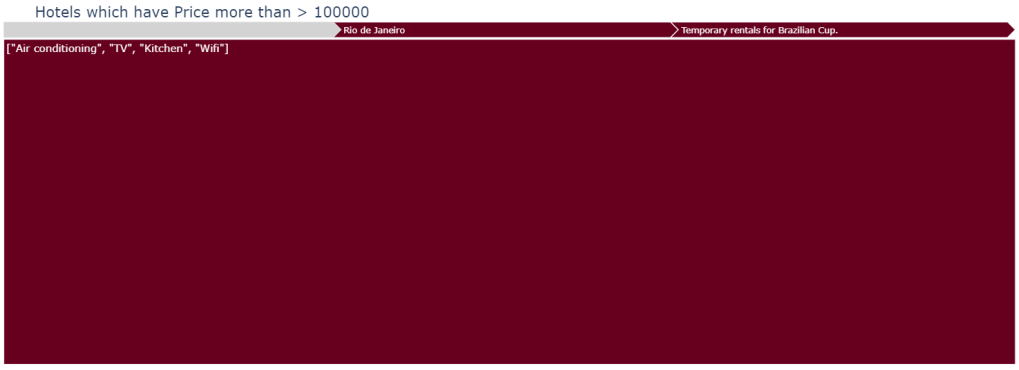

Fig5 c. On clicking an Airbnb, Amenities displayed

The figures 5 a,b,c show a plot with the cities and their Airbnb listings that are pricier than 10000 and on clicking a listing it lists the Amenities and their prices. This kind of plot gives us a better understanding of what amenities are offered at price.

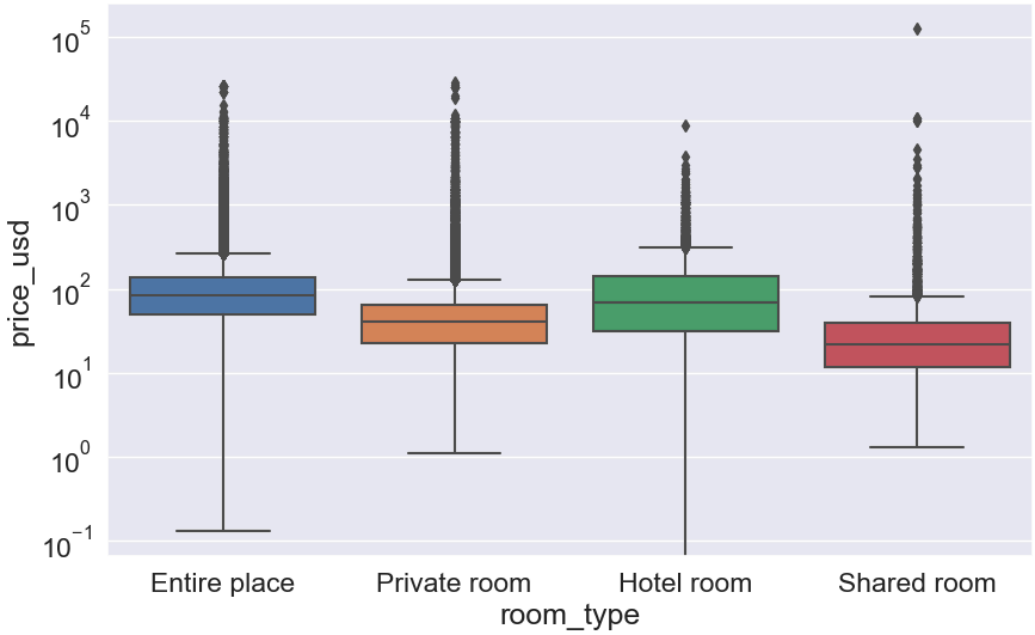

Fig 6. Box-plot of Room Type vs. price

The Fig 6 features a logarithmic box-and-whisker plot, providing a comparative analysis of accommodation

prices across four types: Entire place, Private room, Hotel room, and Shared room. The median price for each accommodation type is highlighted by the line within the corresponding box. Notably, entire places exhibit a consistently higher median price compared to other accommodation types.Private rooms show a lower median price than entire places but a higher median compared to hotel rooms and shared rooms, indicating a mid-range pricing category. The presence of points above the whiskers in all categories indicates outliers, with certain instances of exceptionally high prices compared to the main distribution of the data.The range of prices, denoted by the length of the whiskers, is notably broad for all accommodation types, especially for entire places and hotel rooms. This suggests a high degree of variability in prices within these categories.

In summary, the logarithmic box-and-whisker plot effectively communicates the distribution and comparative pricing patterns across different types of accommodations. The visualization allows for insights into median prices, the presence of outliers, and the overall variability in pricing for each accommodation type, providing valuable information for those seeking to understand the range of accommodation costs.

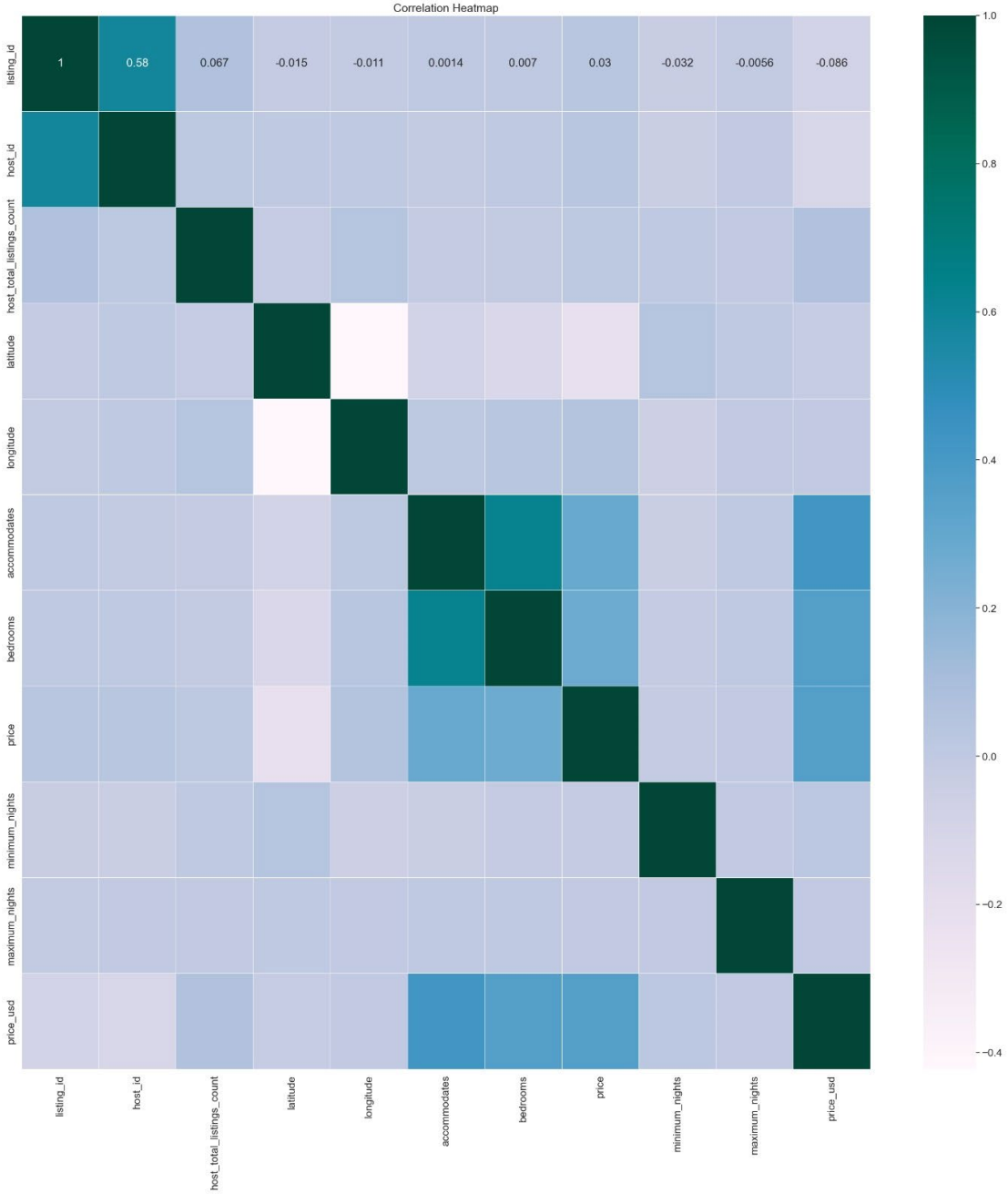

Fig 7. Correlation HeatMap

The heatmap depicts the correlation coefficients between different variables in a dataset, which could include features like location, property size, or amenities. A darker color indicates a stronger correlation, either positive or negative. The correlation analysis of accommodation features reveals insightful relationships with pricing. Both the number of accommodates and bedrooms exhibit a moderate positive correlation with the accommodation price in USD, suggesting that larger and more spacious properties command higher prices. Hosts with a greater total listings count also correlate positively with prices, indicating a potential connection between hosting expertise or reputation and pricing strategies. Conversely, the minimum and maximum nights required for booking display negligible correlations with price, implying that variations in booking durations have limited influence on overall pricing decisions. These findings contribute to a nuanced understanding of the factors shaping accommodation prices, emphasizing the significance of accommodation size and host-related characteristics in determining pricing dynamics within the lodging market.

Now let us move on to the SMART Questions

V. SMART QUESTIONS

A SMART question is a well-defined and structured inquiry that adheres to the SMART criteria, which are specific, measurable, achievable, relevant, and time-bound. It helps to ensure that the question is clear, focused, and designed to yield actionable and meaningful results. The SMART framework is commonly used in goal-setting and project management to create objectives that are precise and attainable. When applied to formulating questions, the SMART approach helps to enhance clarity and effectiveness in seeking information or solving problems.

Given the AirBnb dataset, we have formulated three SMART questions in the direction of addressing the variation of property prices across neighborhoods, the questions were designed to be specific by targeting a distinct aspect of the Airbnb marketplace. First question was aimed at analyzing how prices varied accorss different areas and what factors contribute to it, meanwhile the second question is aimed to analyze the superhosts in Airbnb and the final one is aimed to understand how review scores are affected. These are the questions-

1. How do property prices vary across different neighborhoods or districts, and what factors contribute to these differences?
2. Do certain districts attract more super hosts, and if so, how does this influence the pricing strategy and overall customer satisfaction in those areas?
3. a. Are there specific types of properties or rooms where hosts with a longer tenure tend to achieve higher review scores?
3. b. Do hosts with a higher response rate typically maintain better cleanliness and accuracy, leading to improved review scores?

Now let us analyze and answer these questions

1. ***How do property prices vary across different neighborhoods or districts, and what factors contribute to these differences?***

A comprehensive analysis of factors influencing price variation in the accommodation sector was conducted using a diverse set of predictive models, including linear regression, decision tree, random forest[4], lasso[5] , gradient boosting, xgboost[6], and catboost [7], applied to the entire dataset. To gain a deeper understanding of the nuances within the data, a focused investigation was undertaken specifically for two prominent cities, New York and Sydney. The selected features considered for the entire dataset encompassed critical parameters such

as 'host_is_superhost', 'room_type', 'accommodates', 'bedrooms', 'neighbourhood', 'minimum_nights', 'instant_bookable', and 'amenities'. Pre-processing of features included log transformation for 'minimum_nights', the application of RareLabelEncoder to limit categories for 'neighbourhood', 'room_type', 'host_is_superhost', and 'instant_bookable', and one-hot encoding for categorical variables. StandardScaler was employed to normalize 'accommodates' and 'bedrooms'. This meticulous approach aims to uncover the key determinants of price fluctuations, particularly within the unique contexts of New York and Sydney, shedding light on the significance of various features in influencing accommodation pricing.

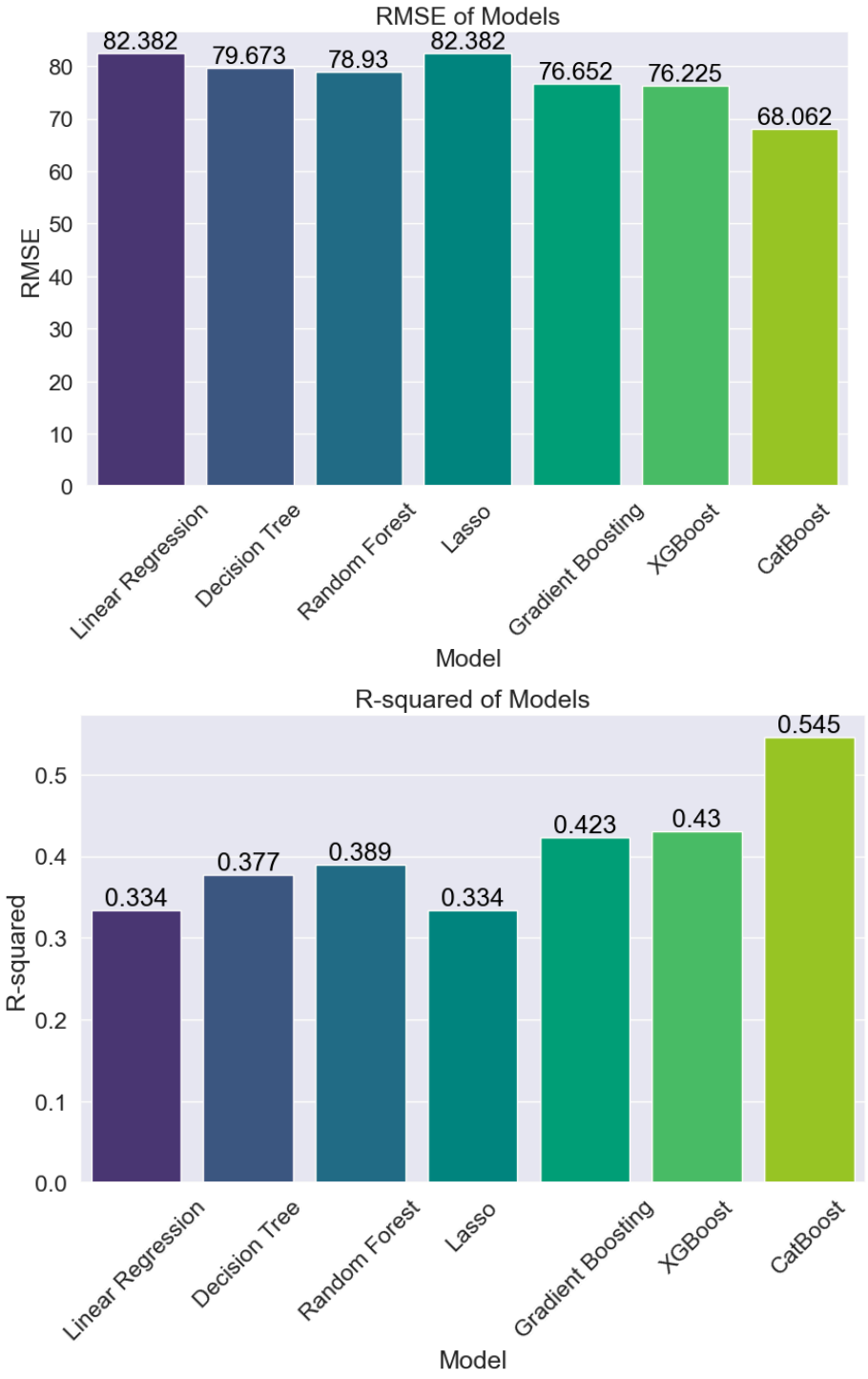

Fig 8. RMSE and R Squared for the Models Performance

For the whole dataset, the models performing vary differently. Linear regression model reached a RMSE at

82.382, and an R-Squared at 0.334. It is the poorest model performing in all of the models. Catboost regression model which is an optimization of tree model XGBoost regression model reached a RMSE at 68.062 and an R-Squared at 0.545.

Among the models assessed, CatBoost emerges as the most effective, boasting the highest R-squared value of 0.545. This signifies that CatBoost provides the most comprehensive explanation of the variability in the dataset around its mean. In contrast, the other models, including Linear Regression, Decision Tree, Random Forest, Lasso, and Gradient Boosting, exhibit comparatively lower R-squared values, indicating a lesser ability to capture the variance within the data. It is the best model performing in all of the models. These results shows that in terms of the whole data set, these variables have an impact on Airbnb price.

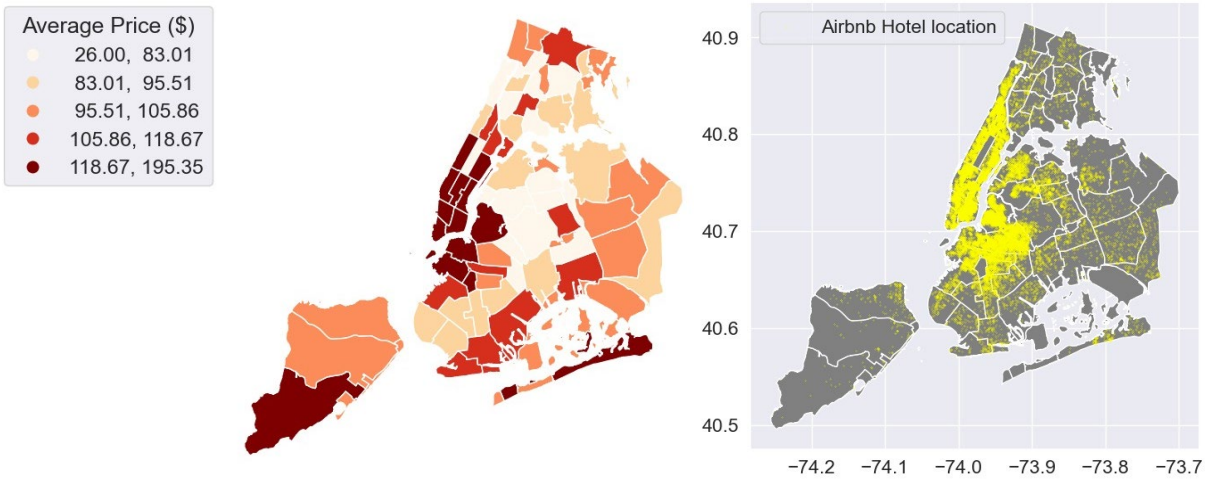

Fig 9. Price Distribution in New York

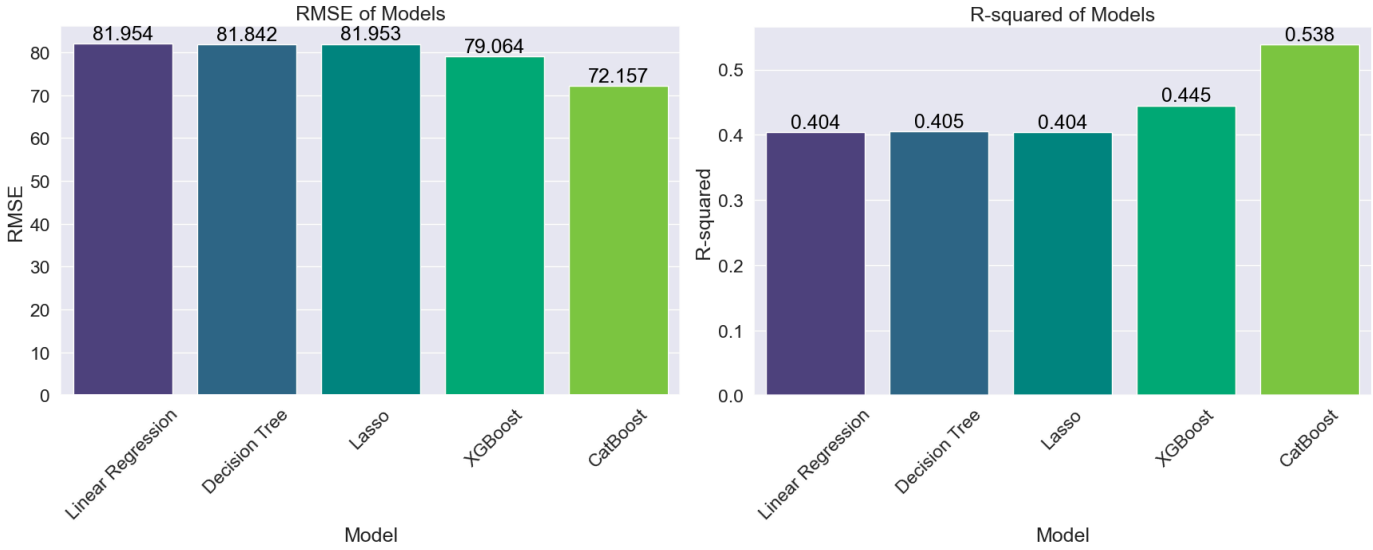

Fig 10. Model Performance for New York

Fig 9 shows a choropleth map, which uses differences in shading or coloring within predefined areas to indicate the average price in dollars in the city of New York. The map uses color gradients, with yellow representing higher concentrations of Airbnb hotels and grey indicating areas with fewer or no Airbnb locations. The densest concentration of Airbnb hotels is evident in Manhattan, particularly prominent in Midtown and Lower Manhattan. Additionally, clusters are observable in parts of Brooklyn and Queens, contributing to the overall density pattern. The use of latitude and longitude coordinates enhances the map's precision, allowing for accurate identification of specific locations. This geospatial visualization offers valuable insights into the spatial

distribution of Airbnb hotels in New York City, highlighting areas of heightened activity and providing a comprehensive overview of the platform's presence within the metropolitan area.

We also ran the same models for the new York and the performance is again catBoost performs the best with a R squared of 0.538 and RMSE is also lower compared to other models.

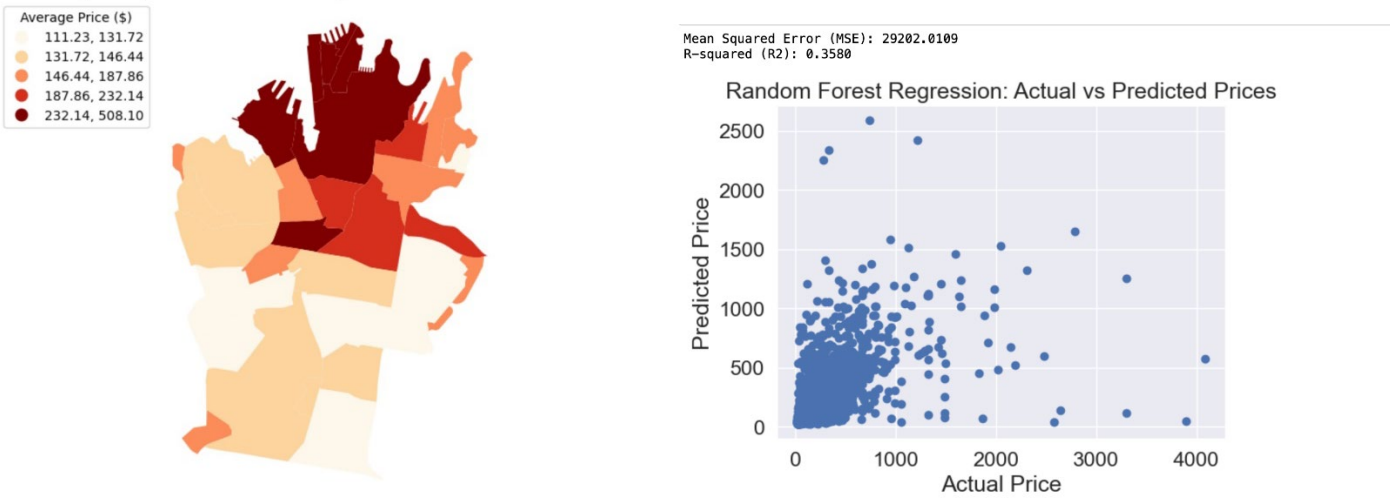

Fig 10. Performance for Sydney

OLS Regression Results

Dep. Variable:	price_usd	R-squared:	0.352
Model:	OLS	Adj. R-squared:	0.351
Method:	Least Squares	F-statistic:	2407.
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	0.00
Time:	14:36:21	Log-Likelihood:	-2.0455e+05
No. Observations:	31075	AIC:	4.091e+05
Df Residuals:	31067	BIC:	4.092e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.136e+04	1448.572	-21.651	0.000	-3.42e+04	-2.85e+04
accommodates	27.4691	0.894	30.720	0.000	25.716	29.222
bedrooms	60.6406	1.744	34.772	0.000	57.222	64.059
latitude	221.6007	13.492	16.425	0.000	195.156	248.045
longitude	308.6696	11.330	27.244	0.000	286.463	330.877
room_Entire place	-7855.5244	362.384	-21.677	0.000	-8565.811	-7145.237
room_Private room	-7851.2038	361.967	-21.690	0.000	-8560.673	-7141.735
room_Hotel room	-7804.5923	362.423	-21.534	0.000	-8514.956	-7094.229
room_Shared room	-7851.3869	362.053	-21.686	0.000	-8561.026	-7141.748

Omnibus:	42203.994	Durbin-Watson:	1.508
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15227618.992
Skew:	7.715	Prob(JB):	0.00
Kurtosis:	110.343	Cond. No.	2.81e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 9.46e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
Mean Squared Error (MSE): 27501.7635
R-squared (R2): 0.3954

Fig 11. Regression Model for Sydney

Similarly, for Sydney as well we used the Linear regression and the Random forest model. In the Linear regression summary shows the R-squared value of 0.352, suggesting a moderate fit, meaning the model explains about 35.2% of the variability of the response data around its mean. Also it has a mean square error of 27502. Meanwhile the random forest model gives a R Squared of about 0.358 slightly higher than the linear model but a higher mean square error

In conclusion, the comprehensive analysis of factors influencing accommodation prices across different neighborhoods or districts, particularly in New York and Sydney, revealed notable insights. The predictive modeling, employing various algorithms such as linear regression, decision tree, random forest, lasso, gradient boosting, xgboost, and catboost, highlighted the significance of features like host-related attributes, room type, accommodation capacity, and neighborhood in explaining price variations. The meticulous pre-processing steps, including log transformations, categorical encoding, and standardization, aimed to enhance the models' interpretability and accuracy.

The standout performer among the models was CatBoost, consistently demonstrating the highest R-squared value, indicating its superior ability to explain the variability in accommodation prices. The choropleth map for New York visually reinforced the concentration of Airbnb hotels, particularly in Manhattan, offering a clear spatial understanding of the platform's distribution within the city.

The subsequent application of the models specifically to New York and Sydney reaffirmed CatBoost's dominance, consistently outperforming other models in terms of R-squared values and lower root mean square error (RMSE). This consistency across different cities underscores the robustness and effectiveness of CatBoost in capturing and explaining the intricate dynamics of accommodation pricing.

In Sydney, the comparison between Linear Regression and Random Forest models revealed nuanced trade-offs between model fit and predictive accuracy. While the Random Forest model exhibited a slightly higher R-squared value, it incurred a higher mean square error compared to the Linear Regression model. This emphasizes the importance of considering both fit and accuracy metrics when evaluating model performance.

In summary, the findings underscore the complex interplay of factors influencing accommodation prices across diverse neighborhoods. The combination of advanced predictive modeling techniques and geospatial visualizations offers a comprehensive understanding of the dynamics at play, providing valuable insights for stakeholders in the real estate and hospitality industries, urban planning, and beyond.

2. ***Do certain districts attract more super hosts, and if so, how does this influence the pricing strategy and overall customer satisfaction in those areas?***

This question delves into the influence of superhosts on Airbnb, examining how their presence in districts impacts pricing and guest satisfaction. Through detailed visualizations and statistical analyses, we investigate whether superhost density affects lodging costs and guest happiness. The ultimate goal is to shed light on the role superhosts play within specific neighborhoods, empowering both hosts and travelers with valuable insights on pricing strategies and overall Airbnb experiences.

To answer this particular question, let us visualize a distribution of superhosts accorss the differnt districs in the dataset and perform statistical tests such as T-test and Anova.

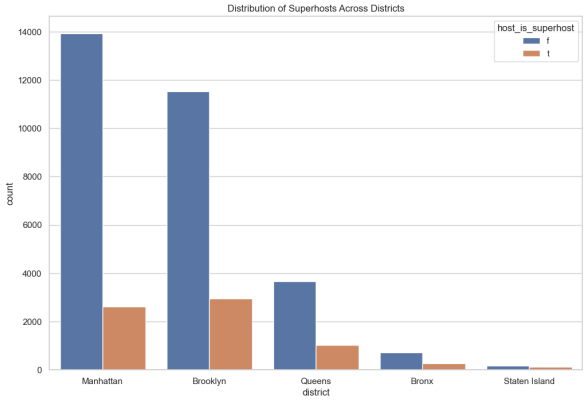

Fig 12. Distribution of Superhosts Across Districts

The barplot clearly shows that certain districsts such as Manhattan and Brroklyn have a higher number of super hosts while other like Staten island and Bronx have very few. This is a clear indication that certain districts get more superhosts.

In the statistical analysis of superhosts' impact on pricing and customer satisfaction, the null hypothesis (H0) posited no effect of superhost status on prices and review scores, while the alternative hypothesis (H1) suggested significant differences. The resulting p-values, nearly zero in both tests, strongly reject the null hypothesis, indicating substantial evidence for the presence of significant differences.

T-test Results for Price:
T-statistic: 3.987222950284716
P-value: 6.686817236530981e-05

The T-test has a T-statistic 3.99 and a p-value of 6.69e-05. The small p-value signifies a noteworthy distinction in prices between properties with superhosts and those without. The positive t-statistic indicates that, on average, prices for superhosts are higher than for non-superhosts.

ANOVA Results for Review Scores:
F-statistic: 7944.7561438765615
P-value: 0.0

Similarly the ANOVA test also resulted in a F-statistic of 7944.76 and a p-value of ~0.0. The extremely low p-value from the ANOVA indicates significant differences in review scores across various superhost categories, suggesting that the average review scores are not uniform for all superhost categories.

In conclusion these results provide compelling statistical evidence supporting the analysis of whether certain districts attract more superhosts and how this impacts pricing and customer satisfaction. The t-test implies that superhosts likely command higher prices, while the ANOVA results suggest variations in average review scores for different superhost categories.

3. ***a. Are there specific types of properties or rooms where hosts with a longer tenure tend to achieve higher review scores?***

This question is intended to analyse wheather long time hosts recieve better reviews and does the properties or roomtypes have an effect on the review. For answering this question let us visualize a scatterplot of Host tenure (calculated using host since) and review scores by room_type. Then let us perfom and Anova Test.

Fig 13. Host Tenure vs. Review Scores by Room Type

The scatterplot shows that the room_type 'entireplace' has better reviews as compared to 'private room' and 'Hotel' etc. The bottom right of the graph is quite less dense and hence we can deduce that hosts with a longer tenure have lesser number lower review scores. Also we can hypothesize that hosts with experience genrally perform better as they are aware of the customer need and satisifaction.

Anova Test-
F-statistic: 13.234750673110252 P-value: 1.789096835339465e-06

The Anova test conducted has obtained F-statistic of 13.23 and a remarkably low p-value of 1.79e-06 indicate that there are statistically significant differences in review scores across various room types. This implies that certain properties or room types tend to have better review scores than others as evidenced the plot as well, suggesting a correlation between host tenure, property characteristics, and overall customer satisfaction. Further analysis, including pairwise comparisons using the Tukey-Kramer test, would identify specific room types that differ significantly in terms of review scores.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Entire place	Hotel room	-0.2532	0.0	-0.2975	-0.209	True
Entire place	Private room	0.0005	0.9997	-0.0135	0.0146	False
Entire place	Shared room	-0.114	0.0	-0.1713	-0.0566	True
Hotel room	Private room	0.2538	0.0	0.2085	0.299	True
Hotel room	Shared room	0.1393	0.0	0.0676	0.2109	True
Private room	Shared room	-0.1145	0.0	-0.1726	-0.0564	True

Fig14. Tukey-Kramer test results

The results of the Tukey-Kramer test further illuminate the differences in mean review scores among specific room types. The test identified statistically significant distinctions in mean review scores between "Entire place" and "Hotel room," "Entire place" and "Shared room," "Hotel room" and "Shared room," as well as "Private room" and "Shared room." However, there was no statistically significant difference in mean review scores between "Entire place" and "Private room."

These findings emphasize that certain room types, such as "Entire place," may generally receive higher or lower

review scores compared to others like "Shared room" or "Hotel room." Understanding these differences is crucial for hosts and travelers to make informed decisions regarding Airbnb properties. By recognizing the factors influencing customer satisfaction, hosts can enhance their offerings, and travelers can make more informed choices aligned with their preferences.

b. Do hosts with a higher response rate typically maintain better cleanliness and accuracy, leading to improved review scores?

This question is intended to know how Airbnb hosts' quick response to inquiries might link to better cleanliness, accuracy, and overall guest satisfaction. The idea is that hosts who respond promptly tend to maintain cleaner and more accurate spaces, ultimately earning higher reviews from guests. To investigate, we used visualizations and statistical measures to untangle the relationships among response rates, cleanliness, accuracy, and overall review scores.

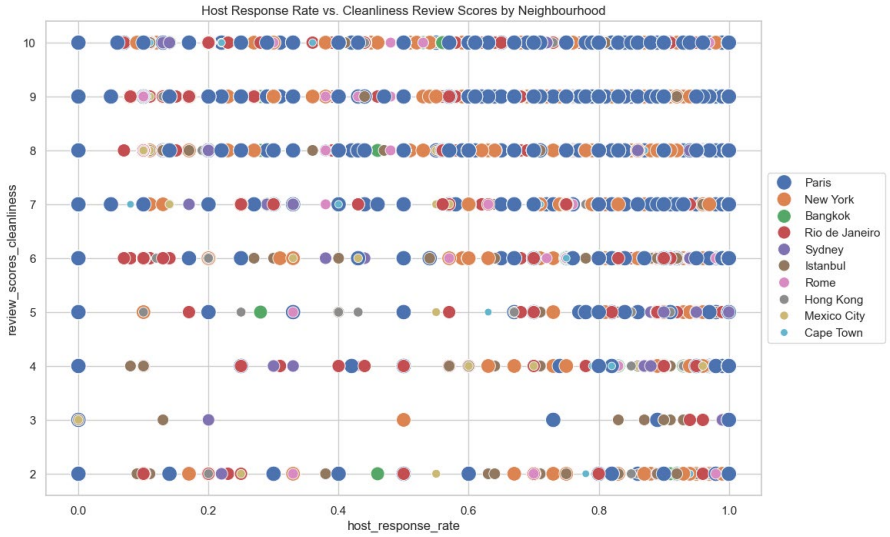

Fig 15. Host Response Rate vs. Cleanliness Review Scores by City

The Scatterplot of cleanliness score vs host response rate accross different cities shows that the hosts who respond quickly most of the times have a better score. The top right of the scatter plot is denser compared to the bottom left indicating that faster response of hosts is related to their cleanliness scores.

Pearson correlation: coefficient- 0.09677500837202732
P-value: 4.6680641532903e-26

The Pearson test conducted gives the Pearson correlation coefficient, calculated as approximately 0.097, indicates a weak positive correlation between cleanliness review scores and host response rate. This implies that, on average, as host response rates increase, cleanliness review scores also tend to increase, suggesting a modest association between the two factors.

Furthermore, the p-value associated with the correlation coefficient is very close to zero (4.67e-26), signifying statistical significance. Therefore, there is strong evidence to reject the null hypothesis, indicating that there is indeed a significant correlation between cleanliness review scores and host response rate. In practical terms, this implies that hosts with higher response rates may, on average, maintain better cleanliness, leading to improved review scores for cleanliness. Hosts who prioritize timely and effective communication may contribute to a positive guest experience, as reflected in their cleanliness review scores.

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, our project conducted a thorough analysis of global Airbnb hotel data, unveiling pivotal insights into the determinants of hotel pricing. Through extensive exploratory data analysis (EDA), we discovered that the selection of amenities and bedroom types significantly influences hotel prices, even in the presence of positive reviews. Transitioning to city-specific analyses, we observed heightened correlations between latitude, longitude, and pricing, emphasizing the importance of localized factors in shaping pricing dynamics. These findings have profound implications for pricing models and recommendations, suggesting the need for tailored strategies based on city attributes. As we conclude this phase, our project lays the groundwork for future explorations, inviting deeper dives into specific city characteristics and user demographics to refine our understanding of the complex landscape of Airbnb hotel pricing.

VII. REFERENCES

[1] Airbnb Listings & Reviews, https://www.kaggle.com/datasets/mysarahmadbhat/airbnb-listings-reviews

[2] Prasad Patil, Published in Towards Data Science (Mar 23,2018), Retrieved on Oct 31,2023; https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[3] What is exploratory data analysis?, IBM, Retrieved on Oct 31,2023; https://www.ibm.com/topics/exploratory-data-analysis

[4] Towards Data Science. (n.d.). Random Forests: The Ensemble of Decision Trees. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[5] Towards Data Science. (n.d.). LASSO Regression: Regularization to Prevent Overfitting. https://towardsdatascience.com/lasso-regression-algorithm-what-you-should-know-about-it-7947782375b1

[6] XGBoost. (n.d.). A Scalable and Accurate Implementation of Gradient Boosting. https://xgboost.ai/

[7] CatBoost. (n.d.). Diving into the Details. https://catboost.ai/