Report: Airbnb Business Data Analysis.

1. Research question

What properties are most commonly listed on Airbnb in New York City in 2019, and how does the property type affect its price and availability?

2. Background

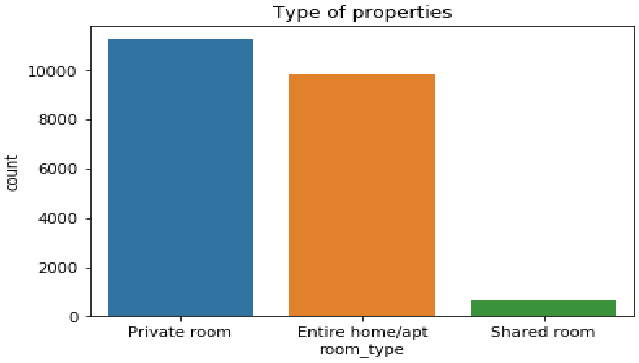
The Airbnb business model relies on hosts offering up their properties to guests. Since its founding in 2007, millions of travellers have benefited from the service (Airbnb, 2023). In 2019, New York City received over 66.6 million visitors (Office of the New York State Comptroller, 2021), presenting a significant business opportunity. An analysis of the dataset logged by the company in 2019 is being used to identify trends and thus enable Airbnb to increase its market share. Executives propose that following the adoption of the necessary data-driven decisions, it would lead to an increasing number of guests choosing the service over hotels. Conversely, there are concerns about the potential of this approach to increase the average rental price, which would likely deter guests.

3. Results and Analysis

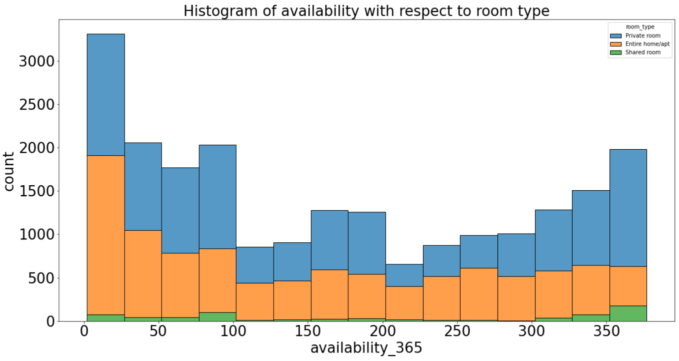
The dataset contains 48895 rows and 16 columns of numerical and categorical variables. In attempting to provide comprehensive answers to the research question, the focus was primarily on the variables ‘price’, which represents the daily rental price of each property and ’availability_365’, which indicates the number of days a residence is available per year, combined with the ‘room_type’, which refers to the type of the property.

3.1 Exploratory Data Analysis

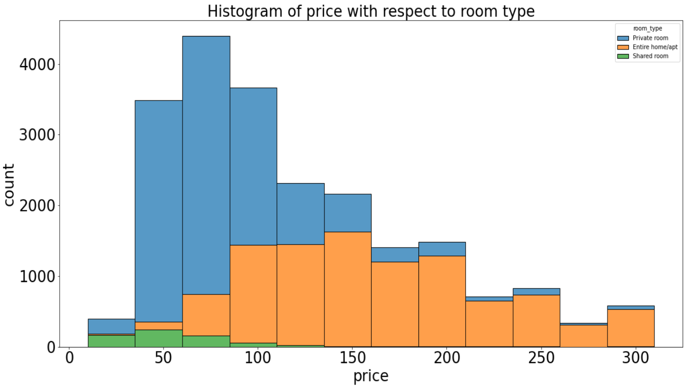
At the initial stages of exploratory data analysis, no missing values were identified within the variables of interest. Consequently, variables with missing values were excluded entirely from the examined dataset to simplify the analysis. In general, all records that included outliers or other anomalies were also excluded from the dataset (Harmadi, 2021).



The value count of room type was displayed as shown in the figure on the left, revealing a relative balance between entire home and private room listings, compared with a significantly smaller number of shared room listings. This finding may be a determining factor in the decision of hosts in relevance to potential future investments, given the apparent saturation of private rooms and entire homes.



In addressing the second part of the research question, a histogram was used to identify any likely impact of the property type in terms of the price and the availability of the corresponding residence. The histogram, presented in the figure to the right, did not reveal any trends in terms of availability. In general, there is a broad spread in availability, regardless of room type. Nevertheless, it should be emphasised that although shared rooms comprise the minority of offerings, most are available throughout the year.

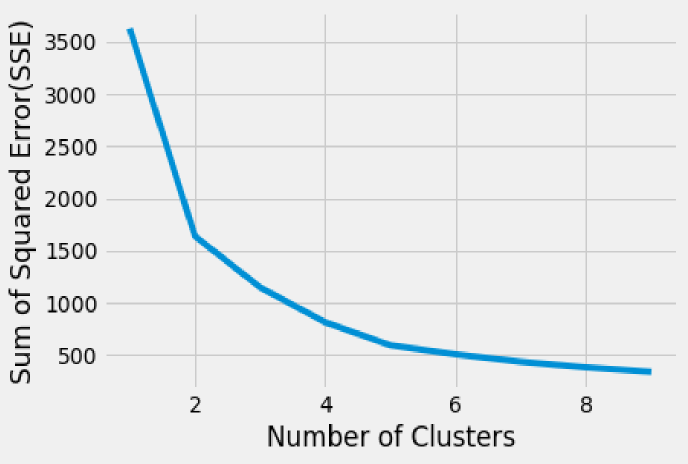


On the other hand, there is a clear trend in the range of prices. As displayed in the figure to the right, shared and private rooms are considerably cheaper in comparison to entire homes. More precisely, most private rooms and shared rooms range between 30 to 100 dollars in rent price, in contrast to entire homes, with most of them ranging between 100 to 200 dollars in rental price. This fact is reasonable since entire homes generally provide additional amenities and offer greater privacy.

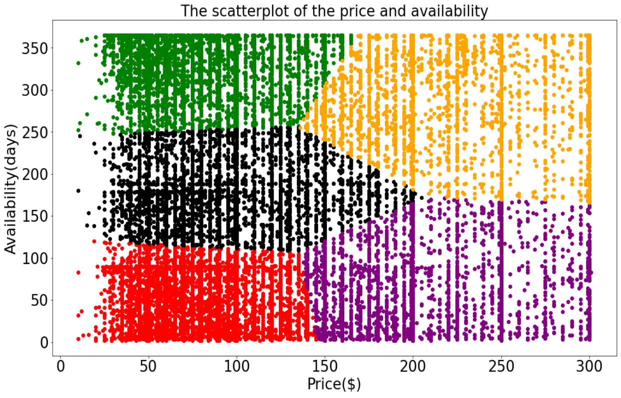
3.2 Data Clustering

For further analysis of the dataset, a K-means clustering algorithm was utilised once again on the numerical variables of interest, price and availability. On a general basis, cluster analysis is considered to be a powerful analytical technique which contributes to discovering hidden patterns, identifying segments and leading to relevant data-driven decisions (Tibco, 2023).

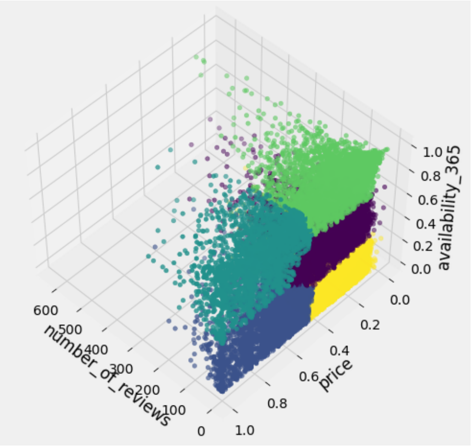
However, when utilising clustering, a suitable selection of the number of clusters is required in advance. In an attempt to determine the figure which maximises the algorithm's performance, the Sum of Squared Errors were calculated for different numbers of clusters, as indicated in the line chart. Eventually, it was concluded that the ideal quantity is five since it appears to be where both the number of clusters and the Sum of Squared Errors are maintained at low levels (Elbow Rule). A further factor reinforcing this selection was the calculation of the silhouette coefficients for a range of possible clusters. To expand on that, the silhouette coefficient is a value which measures cluster cohesion and separation and thus indicates the measure of proximity of the data points within a cluster and simultaneously points out how far away the data points of different clusters are located as well (Qazi, 2023). Therefore, the maximum value of the silhouette coefficient appeared when choosing the number of clusters to be equal to 2, 3 and 5.



The scatterplot underneath displays the price and availability variables. Each colour represents one cluster with the corresponding centroids to be displayed with a '+' symbol. It should be highlighted that the variables must be standardised on the same scale, thus avoiding any bias that might occur during the algorithm procedure.



As displayed below, a 3D plot of the price, availability and number of review variables was created for further analysis. As indicated, clusters that are less expensive and offer greater availability tend to be associated with a greater number of reviews. This could indicate that cheaper properties which are more frequently available for guests, are also more desired since, generally, the majority of reviews are positive (Sharpen, 2019). An explanation for this event can be found by considering the flexibility of lower price listings which are not limited to specific dates. Consequently, this can constitute a crucial factor for owners who intend to increase their property listings on Airbnb.



4. Conclusion

To summarise, the report has provided comprehensive answers to all the separate parts of the research question. Furthermore, it can consist of an important input as the basis for a strategic plan to increase the profitability of the Airbnb business. Therefore it can be utilised to provide the impetus for future studies and investigations, facilitating evidence-based decisions in the field of Airbnb.

5. Appendices

Project Jupyter Notebook

DevelopmentTeamProject.ipynb

References:

About us (publication date unknown) Airbnb Newsroom. Available from: https://news.airbnb.com/about-us/ [Accessed 9 June 2023]

Office of the New York State Comptroller (2021) The Tourism Industry in New York - Reigniting the Return. Available from: https://www.osc.state.ny.us/reports/osdc/tourism-industry-new-york-city [Accessed 9 June 2023]

Harmadi, A. (2021) 10 Things to do when conducting your Exploratory Data Analysis (EDA). Available from: https://medium.com/data-folks-indonesia/10-things-to-do-when-conducting-your-exploratory-data-analysis-eda-7e3b2dfbf812 [Accessed 9 June 2023]

Qazi, N. (2023) *Clustering* [Lecturecast] ML_PCOM7E MAY 2023 Machine Learning May 2023. University of Essex Online.

Sharpen (2019) The Right Mix of Positive and Negative Feedback is a Vehicle for Growth and Improved Agent Performance in Your Contact Center. Available from: https://sharpencx.com/blog/positive-and-negative-feedback-ratio/#:~:text=According%20to%20research%2C%20the%20ideal,in%20high%2Dperforming%20business%20teams [Accessed 10 June 2023]

Tibco (2023) What is Cluster Analysis? Available from: https://www.tibco.com/reference-center/what-is-cluster-analysis#:~:text=Cluster%20analysis%20is%20a%20data,is%20an%20unsupervised%20learning%20method [Accessed 10 June 2023]