Impact of the COVID-19 epidemic on AirBnB London bookings and pricing strategy

*Abstract*— Businesses established in large cities are related in some degree to the success of rental platforms and Airbnb represents a large player in this market. This project conducts a study about the effects seen with the impact of coronavirus crisis in a proportion of the rental market of London. The principal factors that relates to the characteristics of a rental pricing strategy can be explained not only using subjective analysis but also with data analysis with focus on addressing analytical questions. In order to machine learning models and text analysis will be to understand how the booking and prices.

Keywords—component, formatting, style, styling, insert (key words)

# Introduction

Airbnb is a global company responsible to create and develop new markets using a two-sided marketplace model based on guests and hosts and spans its services across over 220 countries and 7 million accommodations as of September 2020[1]. London is one of the most notable markets for Airbnb and high prices charged by agents of the hospitality sector can contribute to explain the success among its competitors.

This study aims to analyze data patterns from Airbnb listings and reviews in London and combine with data extracted from a timeline of Covid-19 pandemic in the United Kingdom. The range of effects to the sharing economy market in most global cities are unprecedent even when one considers the disruptive phenomena of short-term rentals in the tourism industry[2]. Studies conducted by DuBois reported considerable changes in Beijing weekly bookings between the beginning of 2020 and March of the same year of 96%. However, due to travel changes Airbnb had to adapt to this reality and made changes that saved the company from bankruptcy.

The research questions are motivated by the changes derived from the Covid-19 phenomena and also how Airbnb particularly hosts, is adapting to the reality.

# Analytical questions and data

Some questions are defined

This report will investigate properties of Airbnb business and the extent of changes before, during and after the start of Covid-19 crisis. In order to achieve this objective there are questions which will help to construct the foundations

Which effects of Covid-19 can be seen for the Airbnb London regions across time ?

What is the difference between the superhosts distribution before and after the disease spread ?

How has Covid-19 affected guest and host behaviour ?

What are the features that best describe the pricing strategy ?

The Airbnb dataset is originally split between listings, reviews and calendars files, but for the scope of this project the latter was not used. The dataset was collected on November 2020 and has anonymized information from guests and hosts based on Airbnb listings and reviews. To provide support part of the analytical questions geospatial libraries and a GeoJSON file was used.

The listings dataset has around 77000 observations and 75 columns and reviews dataset has more than 1 million reviews but only has 4 columns.

Both datasets contains data from 2008 until 2020 and the listings dataset has a large amount of missing values, low variance columns and columns which are not relevant for the analysis such as URL and uncertainty data about listings availability. There are columns which do not reflect the truth and one of them is the prices, which relates to the current listed price and not the . Other than that, there were features which are not present in the most updated version and this would improve the analysis, such as cleaning fees and

Diagram, table

Description automatically generated

Figure 1: Patterns of missing value in Airbnb listings

# Analysis

Before the data is ready to address the analytical questions, it was necessary to apply data processing and part of data derivation process in the raw dataset. Considering there are two main datasets and one small dataset with Covid-19 related features a plan was created. One of the most important and challenging parts of the data processing is how to apply the domain understanding to the data and how to avoid high dimensional and complexity issues. Initially, a summary of the statistics was taken from all datasets to find basic information about the columns and their distributions, null values and data types.

As part of the data processing a heuristic was followed to process and clean data. One of the first decisions was about what columns to keep instead of selecting the ones to remove. Part of the analysis was done using text contents, so the reviews dataset proved to be easier to process despite of the higher number of observations compared to the listings.

The data processing of listings started with the investigation of duplicate rows and columns with missing values for all observations and columns containing a single value were removed. At this point data conversion had to be performed to transform string variables into a native format as for example string as date converted to date and percentage rates and price converted to numerical format. Another crucial inspection was on the price variable and as observed in the picture the measures of central tendency and dispersion are affected by a small number of outliers and a careful analysis was made splitting the outliers into explanatory accommodates', 'bedrooms', 'beds', 'number\_of\_reviews

Histograms and data filtering assisted the observation of highly skewed and correlated variables and this was confirmed with results from correlation matrices. Therefore, for each group of strong correlated variables only the most explainable was kept.

Another group of variables had missing values and at the same time and specifically for beds, bedrooms and considering a discrete range

For text analysis the reviews missing data was not present, but performance of reviews text processing was slow and a subset of comments was created to address those issues. This process employed a stratified sampling method based on years as subgroups and a percentage of 20% was sampled from each subgroup. A function was created to filter and remove stop words, punctuation and other meaningless words that would not add value to the subject. Other filters were also included to apply a random sampling and start and end dates to enable filters for periods of interest. A polarity column based on comments was calculated to further analysis and finally.

When the main data analysis steps new features were conceived. Date variables from the dataset proved to be crucial for the analysis and had to be discarded before the construction of the model. To avoid loss of this feature and two new calculated variables related to the first and last review dates and date a host joined the platform.

One major concern was how to best represent categorical variables with many possible values. As a requirement it was decided to apply distinct methods depending on the variable properties. At first, neighbourhoods and amenities needed to be analyzed against their proportion and the most representative were encoded, mainly because the sparse effect created after encoding and the number of variables created. Other variables could be entirely represented such as

At this point the dataset still contained 43 variables and scatter plots and correlation matrix were used to find any high correlation between independent variables. A Variance Inflation Factors method was used to access how much variance of an independent variables is influenced by others. Originally, 16 variables were over the threshold, but only 5 were removed to avoid missing relevant variables and knowing that multicollinearity does not affect the results of prediction.

A correlation analysis was done and no other variable was I apply Principal Component Analysis and choose proper number of PC's.

Construction of the model.

In order to build the model considerations needed to be done first. The first decision based on the target variable lead to the choice of a multiple linear regression model and apart from considering that was not the best fit to the data the model offers great potential for interpretability and for simplicity reasons of adaptation to the project scope.

In the Airbnb application, both reviews score ratings and prices are relevant business metrics to decide how successful is each observation. However, review scores tend to be highly left-skewed and most values are concentrated above the 80% score. There are many reasons for that and one of them but when both distributions are compared only the prices have order to predict how the prices of listings are affected by each individual independent variable a linear regression model. Among the variety of metrics available this project focused the dataset was divided into training and test sets using cross-validation with a parameter of 10 for folds after initial assessment.

Validation of results:

The Linear Model results was measured the first time against the standardized target variable. Furthermore, it was possible to detect a cone shaped pattern in the model residuals due to heteroscedasticity because the variance of residuals was not constant across the range of measured values and as a principle residuals considers data with homoscedasticity[3]. To avoid this issue another model was tested using the original scaled features and the log normalized target variable. This time the mean square root error decreased, and coefficient of determination were higher, and residuals plot had scattered values randomly closer to the horizontal line with a small effect size of unequal variability. A careful approach should be done because the target variable is log transformed. Residual plots showed that predicted values were slightly more positive in y axis and concentrated toward the upper right part and this was interpreted as lower estimated predictions for most prices. Actual prices were compared with predictions using a density plot and as a conclusion prices closer to the mean were overestimated but prices at the end of each side had the opposite behavior.

As another form of comparison, a Q-Q plot was used to compare the residuals against a theoretical normal distribution and tails deviated from the normal. This means the model is not correctly predicting prices on both extremes but with more effect on higher prices which might be explained for the presence of outliers even after the log transformation.

The model constructed in statsmodels was used to explore the p-values and coefficients. Only 5 from 38 variables had p-values of more than 0.05 and apart from days\_hosting and host\_response rate all the others were binarized from categorical variables.

To improve the model performance, it might be necessary to transform other variables with asymmetrical distribution and evaluate the model again.

# Findings, reflection and further work

To understand how Covid-19 influenced the Airbnb London business one can ask to what extent are the effects seen for its regions across time. Those effects could be seen in reviews and listings and both figures had decreased for all boroughs between April and July of 2020, which was the period of the first lockdown in the UK. Initial investigations of reviews showed trends in adoption of Airbnb since 2013 despite of seasonal drops and spikes, and they usually followed a yearly trend specially when middle-year and end-year periods are compared between 2017 until the end of 2019, as seen in the figure. However, numbers after the 26th of March had plummeted and equally influenced hosts with and without superhost status despite their magnitude differences. There was a continuous fall that was maintained until the end of the second trimester, which suggests a moderate correlation with the pattern of travels due to social distance measures from the lockdown, that affected not only tourists but internal guests.

As expected, one of the most affected room type listings were private rooms that decreased from an average of more than 2000 reviews from the previous 3 months to approximately 100 reviews until July. A careful observation of London boroughs median prices revealed lower numbers compared to the same period in 2019 and the previous two months, but surprisingly boroughs of Camden, Lewisham and Haringey had the opposite effect. Comparatively, changes in reviews were negative for all representative boroughs and only Barnet and Hounslow had smaller differences in absolute numbers.

Enriched analysis based on reviews was made based on geographical data and compared with periods from before, during and after the lockdown. This analysis used heatmaps and a geospatial json file of London and received inputs from latitude and longitude coordinates to project each observation into a map. The analysis confirmed what was seen and during the lockdown there were only reviews closer to the geographical centroid of the city with few points scattered around the central zone. After the end of restrictions in July the number of reviews increased with a smaller effect size compared to the first three months of the year and regions of Bromley, Wembley and Sydenham became negatively affected even after the lockdown. A similar trend was observed on the number of listings per borough with showed a large difference before and during the lockdown.

Choropleth maps displayed similar characteristics of prices in center regions before and after the lockdown with a slightly less application of higher prices during the lockdown period, which could explain hosts decided to change prices to adapt to a new market demand. Conversely, there were regions in the suburbs that had prices with range of 30/50 £ increased to 60/90 £ and one possible reason might come from people seeking to be more isolated from central locations.

One other question derives from the distinction between hosts with and without superhost status and it is originated from the criteria to participate in this recognized group which turns the analysis against the Covid-19 crisis an interesting subject of how the quality of service performs and if any interruption is noticed. In other words, to qualify for this status regular hosts need to maintain review scores of more than 90% and constantly low cancellation rates, among other requirements.

A first analysis showed similar curves of Airbnb adoption for both groups of hosts in 2020, however superhosts in previous years joined in a lower rate compared to the other group. Price figures are not different among them, however the review ratings and response rates appeared to offer a good evidence of differences between hosts groups specially when several periods were confronted. At this time a null hypothesis was formulated as part of the answer for this question and a statistical test of proportions conducted. The null hypothesis was represented as the null difference of proportions between superhosts and non superhosts with more than 90% of review ratings. The listings dataset represented the population and a sample of the listings during the first lockdown was considered. A significance level of 0.05 was adopted and from that point the proportions along with the standard error were calculated. Finally, a test statistic and the correspondent p-value were computed p-value as 7.5e-21, therefore there was evidence against the null hypothesis and the same could be rejected, then the difference in proportions between superhosts and non superhosts was statistically significant.

Another question relevant concerns to how guests and hosts respond to those changes and this project explored text analysis from both sides. Guest behavior can be measured with analysis of review comments and hosts insert text through listings and neighborhood overview descriptions. The pre preprocessed text was used to generate wordclouds based on time periods, superhost status and polarity scores could be compared against the same variables and review ratings. Wordclouds proved to be useful in finding correlation between negative comments and low ratings, but period comparison was not as relevant. As for polarity it was possible to find some patterns with the review ratings so that higher ratings usually had higher polarity scores and surprisingly 2/5 star ratings had a expressively higher polarity in 2020 than 2019.

Finally, among the predictors the model responded for a smaller set of positive coefficients than the opposite but practically the former had large p-values. On the one side, the in\_lockdown generated variable had a high negative coefficient and it is possible to explain periods of lockdown and lower prices and out of lockdown if prices were generally higher, but a p-value of . Other predictors with positive coefficients are the private room type, bedrooms, accommodates, and the boroughs of Westminster and Camden. As a suggestion for future works larger periods could be explored, particularly the ones with the second lockdown and other regression models

1. References

[1] “About Us,” *Airbnb Newsroom*. https://news.airbnb.com/about-us/ (accessed Dec. 17, 2020).

[2] M. Hossain, “The effect of the Covid-19 on sharing economy activities,” *J. Clean. Prod.*, vol. 280, p. 124782, Jan. 2021, doi: 10.1016/j.jclepro.2020.124782.

[3] “Heteroscedasticity in Regression Analysis,” *Statistics By Jim*, Aug. 13, 2017. http://statisticsbyjim.com/regression/heteroscedasticity-regression/ (accessed Dec. 19, 2020).

1. DuBois, D. (2020) Impact of the coronavirus on global short-term rental markets, https://www.airdna.co/blog/coronavirus-impact-on-global-short-term-rental-markets, last accessed 13.12.2020.

##### Word counts

|  |  |
| --- | --- |
| Abstract | 149/150 |
| Introduction | 149/150 |
| Analytical Questions | 290/300 |
| Analysis | 985/1000 |
| Findings, reflections and further work | 594/600 |