**An Empirical Investigation of a Nashville Airbnb Market**

| **Project Summary:** As part of the sharing economy platforms, Airbnb has expanded significantly since its founding in 2007 and become an alternative to traditional hotels. Analyze Airbnb data in the past 4 quarters in Nashville, TN. |
| --- |
| **Time Range and Data:**  All publicly available data from the website <http://insideairbnb.com/nashville>.  Additional data sources: Census data and COVID-19 tracking data.  AirBnB Data includes quarterly detailed listing, calendar, and review data. Data set June to Dec 2021 used for the analysis. |
| **Main objective:** to gain insight into Nashville, TN Airbnb market in the year 2021. |
| **Objective 1** - ​​Please propose and design a comprehensive list of matrices to understand the local Airbnb market. Examples of such measures could be:  o Market supply: the number of active listings  o Market demand: occupancy rate, average monthly reviews per listing, etc.  § Number of reviews can be used as a proxy for number of customers, given a high percentage of customers leave reviews for Airbnb  o Customer comments: topics, sentiments, etc. |
| An airbnb listing is considered active if it has between 1 to 364 available dates in the upcoming year and also received at least 1 review in the last 8 months. For our six month period (June to Dec 2021) analysis, only availability\_365 > 1 from listing data set and available field in calendar data set are considered to identify active listings in all Nashville neighborhoods.  Results from data analysis   1. Total list data 4483 applied to above rule to derive active listings = 2351. 2. Neighborhoods data analysis from below graphs reveal that most active listings are around downtown and Grand Ole Opry. They are in the neighborhoods District 15 to 21 and District 5 to 8. 3. Full house rentals generated higher revenue, in fact they stood out compared to other property type rentals. Six months show downtown rental neighborhoods revenue higher compared to neighborhoods near Grand Ole Opry shows the effect of COVID on vacation and entertainment choices visitors made. 4. Topic model and sentimental analysis show customers had positive experience from their stay at airbnb properties.   The code for following graphs can be searched in airbnb\_nashville\_trend.R file by typing in the graph title.    The Airbnb occupancy rate represents the percentage of days your Airbnb rental property is “occupied” by guests in a year. Calendar data available field “f” is counted as listing occupied on a given day. Occupancy total and total availability of the listings in a neighborhood are calculated and applied to derive occupancy rate for each neighborhood= total number of days active listings occupied / total available days for active listings.    Average revenue is calculated by taking mean of the sum of each listed properties total six months revenue calculated from adjusted price of occupied dates in the calendar data frame, then grouped by neighborhood.    The calculation here is similar to above, revenue generated on property type instead of neighborhood. This helps us understand what type of properties are preferred by vacationers coming to Nashville.    Topics and Sentiment analysis shows consumers on District 15 to 21 and District 5 to 8 shows customers mostly had positive experience from their stay at airbnb rentals.      Topic model analysis shows customers had good experience staying at clean airbnb properties and they recommend it to others.    Below sentiment analysis show more on positive, joy, trust and anticipation. Less on fear, disgust, sadness, anger, negativity. |
| **Objective 2** - Do you observe any trend, seasonality, and neighborhood differences in your local market? |
| The Nashville Airbnb Market has increased exponentially from 2014 to 2019, dropped in 2020, and bounced back from 2021. The drop in 2020 was caused by COVID-19 pandemic. Overall, there is a positive linear growth in the Nashville Airbnb market from 2010 to 2021. The market is expected to keep growing exponentially for the 30 upcoming years according to the forecasting models prediction. |
| **Objective 3** - Do you see any association between the Airbnb market activities and the Covid-19 situation in the local market?   * + *Note: customer review data goes beyond one year* |
| [[Anulitha]] |
| **Objective 4** - What factors affect Airbnb listings’ performance?  ​​o Please propose a reasonable indicator(s) to quantify individual listings’ performance.  § For instance: average monthly occupancy rate, average monthly income, and/or average monthly reviews  Please identify suitable independent variables based on the available data sources. |
| **Data Preprocessing and Future Engineering**  The initial data file contains 49 variables which are different parameters of a listing. For the model building purpose we dropped few irrelevant columns such as **listing\_url, scrape\_id ,name, description, picture\_url, host\_url, host\_name, host\_about, license, host\_thumbnail\_url, host\_picture\_url, host\_verifications, host\_has\_profile\_pic, latitude, longitude etc.** After dropping all these variables which will not have relevance in the model, We are left with 32 variables. And we performed correlation analysis on the remembering variable. And we will use revenue as our target variable. Revenue is calculated as the sum of the adjusted price. (revenue = sum(adjusted\_price, na.rm = TRUE). Based on the VIF analysis we drop a few variables that are highly correlated with other variables. The statistical significance analysis results are as follows.  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1   * review\_scores\_rating ‘\*\*\*’ * review\_scores\_accuracy ‘\*\*\*’ * review\_scores\_cleanliness ‘\*\*\*’ * review\_scores\_checkin ‘\*\*\*’ * review\_scores\_communication ‘\*\*\*’ * review\_scores\_location ‘.‘ * review\_scores\_value ‘\*\*\*’ * Reviews\_per\_month ‘\*’ * host\_response\_rate ‘\*\*\*’ * host\_acceptance\_rate ‘\*\*\*’ * host\_listings\_count ‘\*’ * host\_has\_profile\_pic1 ‘.’ * host\_since ‘\*\*\*’   The variables we will be using for analysis review\_scores\_rating   * review\_scores\_accuracy * review\_scores\_rating * review\_scores\_cleanliness * review\_scores\_checkin * review\_scores\_communication * host\_acceptance\_rate * host\_since   The VIF table shows no multicollinearity between these features.  Necessary preprocessing was taken before model building. This includes:   * Review of variables to figure if there are privacy or ethical concerns * Visual analysis to identify variables that may not have correlation to the target * Reviewing each variable purpose to better understand if they have any meaningful relevance to the target * Combined all data frames for each quarter and only analyzed active listings for the combined data frame. * Reviewing null values for very high ratios of missing values * Reviewing distribution of data and finding potential outliers * Checking variable types and factoring those that are categorical * For the dataframe for the final model we joined both listing data and calendar data using inner join based on listing id. And also included a filter which will return listings occupied at least one day in a year. This inner join gives the active listing dataframe for the model.   + active\_listings <- inner\_join(listings\_data, calendar\_data, by = "listing\_id") %>%   filter( (available=='f') & (availability\_365 > 1))   * **active\_listings** is then grouped based on listing Id. And the mean score is calculated for all the review ratings that are used for the model.      * MEAN imputation was done on the data as some of our variables in our data have missing values, it was safe to impute.      * The final data frame (active\_listings\_for\_model) consists of 10 variables with only active listing Id. And all the review scores are averaged. And imputed all the variables to make sure missing values will not affect the model performance. In summary, we carry forward the following variables for modeling.**review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_communication , host\_acceptance\_rate, host\_since, review\_scores\_checkin.** |
| [Hamed] - Modeling  Using the variables identified above, we create a target variable that shows whether revenue is greater than average. We then drop the listing id and revenue fields - as well as price that was used to calculate revenue - and carry forward the variables shown below for the forest model.    A histogram of input variables as well as our newly created target is shown below.  **Target Variable**  We define revenue as performed in the data processing above as a measure to evaluate performance of a listing. For our modeling purposes, we define a new target variable that equals 1 if revenue of a listing is greater than median and 0 if less than it.    We also reject listing\_id and price as well as revenue as the listing\_id has no meaningful information to predicate our target and the other two variables are already used to derive the target.  After rejecting the variables:    The median revenue calculated from the final dataframe equals $14,801. VIF also shows our variables do not have any multicollinearity:      **Running the Model**  We partition the data using a 70% vs 30% proportion for training and test datasets.  After partitioning, we run a default random forest model first (with default values and number of trees = 20):    The default mtry value = 4.  Then we run a more detailed model by changing the parameter of mtry, increasing the number of trees, and introducing the importance argument.    We observe the selected mtry given the optimal model on accuracy is achieved by mtry = 3 and the accuracy has increased.    This means, 3 variables are randomly selected each time a tree is selected in the random forest and in each tree 1432 random observations are selected from the data.    And by evaluating the performance of the model on validation data, we observe the following that shows confusion matrix as well as accuracy metrics:    We then evaluate the results using the test dataset and achieve an accuracy of the following:    **Importance of the Variables**    This suggests the most important variable is checking score and then other similar scores including history of being in the market and cleanliness as well as check-in scores.  It’s also noteworthy that communication score is the least important factor in revenue which suggests customers care most probably about other quality factors than communication and issues related to communication (complaints, etc.) are minimal when quality of service is high. |
| **References**   1. <https://www.census.gov/quickfacts/fact/table/davidsoncountytennessee#>   <https://www.census.gov/programs-surveys/popest/data.html>  https://www.census.gov/data/tables/2020/demo/popest/2020-demographic-analysis-tables.html  https://www.census.gov/quickfacts/ashevillecitynorthcarolina  not able to access Nashville TN on this link  <https://www.census.gov/data/datasets/time-series/demo/popest/2020s-total-housing-units.html> - CO-EST2021-HU-47.xlsx file  <https://www.census.gov/data/datasets/time-series/demo/popest/2020s-total-cities-and-towns.html> - sub-est2021\_all.csv  area of census.gov that you are trying to access is currently unavailable due to maintenance   1. <http://insideairbnb.com/nashville> |