CAPSTONE DATA STORY

William S. Justice

Springboard

CAPSTONE DATA STORY

The problem faced in this capstone project is analyzing the trends and dynamics of first time users in the Seattle, Washington Airbnb market. Airbnb has shown tremendous increase in popularity. Airbnb allows users to search and book privately owned accommodations according to their specifications. These specifications vary dependent on users but can include areas such as past review scores, nightly rate, room type, availability, and area. It is key variables such as these that this analysis explores to further understand the market dynamics of first time Airbnb users.

The data was obtained through the Kaggle website. All data was scraped from the Airbnb website. The data was focused on users that left reviews. Reviews for Airbnb are optional and not all users regularly provide reviews on accommodations. The data set provided reviews in a very detailed manner. When a review is made there are a large amount of options available for users to specifically review. This is the information in which other users can base their determination on booking an accommodation. The review data set offered a large amount of information on not only the host but also the space and user ratings on specific variables. These review ratings include areas such as overall rating, review score accuracy, cleanliness, check-in, host communication, location, and price value. All of the data combined gave great insight into how each individual reviewer assessed their stay and accommodation.

The data had a few limitations that were important to the problems and hypotheses that structured this project. The primary limitation within this data set was the unavailable data that clearly stated if a user was indeed a first time user or a non-initial user. This limitation had to be justified via data filtering. Another important limitation of this data set was the data provided on dates. There was no data provided on actual dates of stay at the particular accommodation. The data provided on the date was in regards to the date the user made the review. This hindered further exploration into first time user dynamics throughout the week.

The data contributed by Kaggle provided a very clean slate in which to work. There were two data sets that needed to be merged. These first data set was the Listings which had all the information on the specific accommodation. The second data set was the user reviews. These two sets were joined based on the host identification number. This resulted in a data set that had all information regarding all user and host information. This information had to be narrowed down into the specific metrics that were applicable to the problem hypotheses. The data set was filtered into host identification, host since (date), zip code, property type, room type, price, review score, comments, and date. The data set now had to be filtered in order to find first time users. This process was done via a stringr package. This package allowed certain strings to be found within the data. The specific command was str\_detect. This example of natural language processing filtered strings such as “First time experience”, “First experience”, “first time”. Multiple combinations of capital letters were used in order to detect as many first time users as possible from user review comments. Another data frame had to then be made that removed all the first time users from the original data in order to have a data frame that could compare first time users and a data frame that had none of the detected first time users.

The next step in cleaning the data was to alter the date structures of both the review date and “host since date”. This was accomplished by utilizing the as.Date function and standardize the dates into a %Y/%M/%D format. Now that the “host since date” was standardized a new column had to be made that would provide the time in years a host was on the Airbnb site. For example, host ID # 44154 has had a tenure or 7.35 years on Airbnb. This was integral to explore host tenure as a function of first time user ratings.

The hypotheses were made in order to further investigate the key areas crucial to first time users in the Seattle, Washington market. These hypotheses include 1. The more time a host has on Airbnb will result in a higher rated first time user experience. 2. First time Airbnb users are more likely to purchase accommodations at a higher price point compared to non-initial users. 3. First time Airbnb users are more interested in the Home Type (Full home/Apartment) opposed to a shared or single room.

The approach to this project changed slightly as more detail into each hypothesis was available in data. For example the initial hypotheses had no inclination to explore the geographic area. It was quickly understood that analyzing first time users based on zip code preference gave possibly the most intriguing insight into first time user preferences. Another example of how the approach needed to be changed came in the exploration of rating as a function of Airbnb host tenure length. There showed to be little correlation between host tenure length and increased user ratings. Further insight into the number of observations within a time period had to be made in order to create a visualization that showed pertinent information. The final change in approach came with further exploration of the third hypotheses (Home type preference). There was an overwhelming majority of preference shown towards entire home over private and shared rooms. To make the data tell a story it was integral to show a specific time that the bookings of these types were made. This approach showed a clear market trend throughout the year that could be utilized as a marketing tool.

The data story had continual additions over the term of this project. The story continually changed because each hypothesis lead to areas that needed further exploration. This was observable for each hypothesis. This further exploration lead to a more marketable presentation that not only answered first time user questions in regards to the hypotheses but also provided important trends and tendencies of first time Airbnb users in the Seattle, Washington area.