**DATA 557**

**Final Project – Assignment 1**

**Winter 2019**

**02/18/2019**

**Project title**

Recommendation Model for Sharing Economy Platforms - Case Study on Airbnb

**Group members - Group 5**

* Adrian Tullock
* Cathy Jia
* David Wei
* Monique Bi
* Naman Sharma

**Data description**

This dataset is comprised of host listing details, aggregated by city and broken down by individual listings. Users are allowed to have multiple listings, so uniqueness is determined by a numerical identifier with a matching link URL. The dataset is compiled of entries spanning individual cities from Airbnb. The Seattle data has 96 columns and well over 8000 rows. Should we see fit to include data from other cities, this number could easily double or triple in size. We expect the data to be relatively homogenous across locations. Sensitive data has been scrubbed to protect privacy, but other than that the data is pretty well filled out.

There is a broad mix of numerical, boolean, character values, web links and percentages across the data. There are some fields which explicitly have zero values and others which appear to use empty values to indicate ‘N/A’. Analysis of certain fields will likely need data extrapolation for aggregation.

The location data for the listings is a bit redundant, but it may be due to the data being split from a central source prior to archiving. Other data available includes the listing amenities, various pricing stages, average host response times, listing types, review ratings, availability and geospatial coordinates.

**Data source**

URL: <http://insideairbnb.com/get-the-data.html>

The data is provided by Inside Airbnb. According to the website, Inside Airbnb is *an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world*.   
  
The data is a collection of Airbnb data that is publicly available on its website, across multiple cities. The data is aimed at providing a 360-degree insight into Airbnb’s presence in a city. According to the source, *the data has been analyzed, cleansed and aggregated where appropriate to facilitate public discussion.* The data can be *copied, modified, distributed and performed work on,* even for commercial purposes, all without asking permission. A brief summary of assumptions and disclaimers in the dataset is available [here](http://insideairbnb.com/about.html#disclaimers).

**Data availability**

The data set is readily available and ready for further cleaning.

**Questions**

* Key factors for price and review scores
  + What is the most important single factor for getting a high price on your listing? Similarly, what’s the factor you ought to focus on to get better reviews?
  + Repeating the above with the goal of finding the secondary factors that enhance the effect of the primary factor identified.
* Factors for the number of reviews and becoming a superhost[[1]](#footnote-0)
  + As a host, what would be the important factors (among cancellation policy, days of availability, price, minimum nights, review scores, years in business, etc) that would get me more reviews to qualify for being a *superhost*.
* For travelers of group size 2, given a fixed budget of 150 dollars, what would be a good recommendation, in terms of price and rating, for 2 nights in Seattle? Why?
* Follow-up question: Expand our tests to more cities and compare the results.

**Variables**

Main variables to focus on

* *price* (daily rental price)
* *host\_is\_superhost* (T/F)
* Review Scores
  + *number\_of\_reviews*
  + *review\_scores\_rating* (*total, accuracy, cleanliness, check-ins, communication, location, value*)

Airbnb rental features to look for patterns and/or correlations

* Amenities
  + *room\_type*
  + *accommodates*
  + *bathrooms* (number of)
  + *bedrooms* (number of)
  + *beds*
  + *wifi*
  + *Cable TV*
  + *Washer/Dryer*
* Extra fee
  + *security\_deposit*
  + *cleaning\_fee*
  + *extra\_people*
* Rental policy
  + *cancellation\_policy*
  + *minimum\_nights (number of)*
* Host Verifications:
  + *email*
  + *phone*
  + *facebook*
  + *google*
  + *reviews*
* *neighborhood*
* Others (any potential factors affecting customer rating or price change)

**Analysis methods**

* ANOVA, with and without interactions
* Linear regression, against single predictor variables as well as variables, split for interaction effects

**Potential problem**

Given the objective of our hypothesis, we intend to work on as wide a variety of factors as possible. However, in this scenario, assessing the importance/significance of each factor may be a challenge for factors with a large number of groups. This problem will limit the tests to a sub-group of the available factors. We would need several attempts to decide which factors to use in our project.  
  
The response variable in the dataset also needs to be assessed, tested and reviewed before putting in use. This is caused by the fact that the available definitions for the dataset are limited in the information they provide. So, for example, in choosing an appropriate metric for *price* we would need to check the fields *listed\_price, security\_deposit,* and *cleaning\_fee*, among others, before zeroing in.

The data itself is relatively clean and well indexed however it’ll still require some cleaning, feature engineering, and reshaping.

* Missing data in key fields account for about 10% of the data and hence will have to be eliminated - which will lower the power/significance of our tests.
* Subgroups in certain fields account for too little of the population to be considered valid - they will need to be weeded out for each factor while testing.
* The amenities and owner\_verification fields are comma separated text fields that need to be normalized as they contain a plethora of possibly useful factors.

1. Superhost: <https://blog.beyondpricing.com/how-do-i-become-an-airbnb-superhost/> [↑](#footnote-ref-0)