

## Assessed Project: Learning from Data/Data Science Foundations

These data are taken from <http://tomslee.net/airbnb-data-collection-get-the-data> and are for July 2017.

### Airbnb Data Collection: Get the Data

(Last updated: August 7, 2017)

The data in the Rdata files are for different named cities. To see the names of the objects in the RData file, type 'ls()' in R.

The data for each city represents a single "survey" or "scrape" of the Airbnb web site for that city. The data is collected from the public Airbnb web site by the author of the website.

Here is a description of some of the variables and data that's provided on the website by the website author:

- **room\_id:** A unique number identifying an Airbnb listing. The listing has a URL on the Airbnb web site of [http://airbnb.com/rooms/room\\_id](http://airbnb.com/rooms/room_id)
- **host\_id:** A unique number identifying an Airbnb host. The host's page has a URL on the Airbnb web site of [http://airbnb.com/users/show/host\\_id](http://airbnb.com/users/show/host_id)
- **room\_type:** One of "Entire home/apt", "Private room", or "Shared room"
- **borough:** A subregion of the city or search area for which the survey is carried out. The borough is taken from a shapefile of the city that is obtained independently of the Airbnb web site. For some cities, there is no borough information; for others the borough may be a number. If you have better shapefiles for a city of interest, please send them to me.
- **neighborhood:** As with borough: a subregion of the city or search area for which the survey is carried out. For cities that have both, a neighbourhood is smaller than a borough. For some cities there is no neighbourhood information.
- **reviews:** The number of reviews that a listing has received. Airbnb has said that 70% of visits end up with a review, so the number of reviews can be used to estimate the number of visits. Note that such an estimate will not be reliable for an individual listing (especially as reviews occasionally vanish from the site), but over a city as a whole it should be a useful metric of traffic.
- **overall\_satisfaction:** The average rating (out of five) that the listing has received from those visitors who left a review.
- **accommodates:** The number of guests a listing can accommodate.
- **bedrooms:** The number of bedrooms a listing offers.
- **price:** The price (in \$US) for a night stay. In early surveys, there may be some values that were recorded by month.
- **minstay:** The minimum stay for a visit, as posted by the host.
- **latitude and longitude:** The latitude and longitude of the listing as posted on the Airbnb site: this may be off by a few hundred metres. I do not have a way to track individual listing locations with
- **last\_modified:** the date and time that the values were read from the Airbnb web site.