I. Problem

Airbnb's vast user base, seasonality, and broad reach (both across the US and internationally) make discerning signal from noise very difficult. Which variables are important – that is, which variables have an impact on bookings? Does one medium (iOS vs. Android vs. Desktop) have a higher probability of booking than the others? Are longer session times more likely to result in a booking or a message? How can Airbnb utilize user data to understand users' sessions, improve user experience, and ultimately improve booking requests?

Using sample data from user sessions, I'm going to explore the data in order to understand how the variables impact and affect the probabilities of what users end up doing. I will look at individual user and session data to probabilize which users are most likely to result in a (1) a booking request, (2) repeat booking user, (3) user who doesn't book in this session, but may in the future, and (4) user who never books, as well as users' searching and messaging history.

II. Client and Use Case

The client is Airbnb. With the data that I provide, Airbnb will be able to more accurately understand what variables can tell them about their customers and what those variables mean in terms of probabilizing searches, meesages, and bookings . With further machine learning experience, I would be able to predict which users will end up where. Airbnb will be able to customize the user experience based on my data to both maximize revenue from current "high-value" customers while attempting to keep or sell "lower-value" or "no-value" customers. Ideally this would increase the percent of sessions that end with a booking and, perhaps, increase the average revenue per booking.

III. Data Source

The data will be pulled from: http://databits.io/challenges/airbnb-user-pathways-challenge

The data provided is a .txt delimiter separated data file. The .txt file is separated by '|' characters and can be read in using the readr package (read_delim function):

- Number of records in data: 7756
- Date span of the data: ['2014-05-05', '2015-04-23']
- Number of unique users in data: 630
- Number of unique sessions in data: 7756
- Percent of sessions with search: 15.9%
- Percent of sessions with sent message: 16.5%
- Percent of sessions with booking request: 1.9%

The data also includes a data dictionary with each variable (column) name and a short definition of the data. The data dictionary is a .rtf file, though I converted it to a .xlsx file and read it into R using the readxl package (read_excel function) – this table may or may not be necessary to the overall analysis, but it is helpful to have it in the same R project.

IV. Approach Outline

There are 7,756 distinct records (sessions) in the data comprising 630 different users. There are some obvious early problems with the data. Listed below are the problems and proposed solutions, as well as some other ideas to clean, merge, and deal with the initial data set:

- Problem: Users without a 'Next Session' have all next_xx fields listed as the text value 'NULL' rather than the R value of NA; Solution: read in data using read_delim and converting 'NULL' values to NA using the na argument
- Problem: Session times are not shown; Solution: calculate session times (in seconds) using difftime function and add in 2 new columns (session_time and next_session_time)
- Problem: Next session data is reduntant with other rows and only listed for the very next session; Solution: Remove all next-session data as it is redundant, replace with a binary or couple of binaries as previous_session (yes/no) and next_session (yes/no).
- Problem: dim_user_agent has blank values as well as '-unknown-' values; Solution: standardize 'Other' or 'Unknown' value groupings
- Problem: dim_user_agent and dim_device_app_combo grouped in character strings; Solution: create binaries for each grouping
- Problem: Session times seem heavily positively skewed and also a large number of 0 second session times; Solution: explore the possibility of removing 0 second session times as they don't constitute 'using' the application

There are likely to be more data and cleaning issues as the project moves along, but these are my initial thoughts after doing some early exploratory analysis into the data.

V. Deliverables

Deliverables for this project will be:

- Code
- Paper
- Slide deck