Project 3 MVP Proposal - Songlin Qing

- **Description:** Based on user browsing and booking information for USA based users, predict the first country where new users will make their first booking (12 countries, including no booking).
- **Data source:** https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data
- **Dataset part 1:** Airbnb user booking data (213k), including:
    - User id
    - Account related information
        - Creation time
        - First active time
        - First booking time
        - Signup channel/device/browser/
    - User demographics
        - Age
        - Gender
    - Destination country – this is the outcome
- **Dataset part 2:** Web session logs (10.6m), including:
    - User id
    - Action
        - Lookup/search/click
    - Device type
    - Session duration
- **Dataset part 3:** countries (10), including geospatial info, language
- **Dataset part 4:** population breakdown into gender and age group for all destination countries
- **Model evaluation:** NDCG (Normalized discounted cumulative gain) https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings#evaluation
- **Challenges:**
    - Most of variables are categorical data with many levels, need to convert some of them to continuous variables
    - There are 12 destination countries, with majority of outcome as no booking and US (As data are for US users)
    - Many types of missing and wrong data, and it may be difficult to clean them with SQL
- **Approach:**
    - Extensive EDA to understand the outcome distribution over signup time, signup method, browsing behavior
    - Model building and selection