Data Science Project Proposal

1. **Topic**: **Boston Airbnb Open Data**

2. **Group Members:**
   a. Harsh Shinde (Nuid: 001822871)
   b. Varad Desai (Nuid: 001465732)
   c. Anshul Balamwar (Nuid: 001400625)

3. **Background or Introduction or Overview**

Since 2008, guests and hosts have used Airbnb to travel in a more unique, personalized way. As part of the Airbnb Inside initiative, this dataset describes the listing activity of homestays in Boston, MA.
The following Airbnb activity is included in this Boston dataset: * Listings, including full descriptions and average review score * Reviews, including unique id for each reviewer and detailed comments * Calendar, including listing id and the price and availability for that day

4. **Motivations or Goals or Objectives**
   a. To find optimal price of an accommodation with regards to the features.
   b. To classify the no of people with regards to the features.
   c. To cluster the no of people with regards to the specification of the house.

5. **Methodology or Algorithms and Methods**
   a. Exploration of data
      i. Print shape (no. of rows and columns)
      ii. Determine feature variable (X) and target variable (y = price)
      iii. Find correlation with correlation matrix
   b. Feature Engineering
      i. Train Test Split
      ii. Cross validation of feature variable(X) via GridSearchCV/K-Fold cross validation
   c. Predicting house prices in Boston Airbnb
      i. Linear Regression
      ii. Random forest / Decision Tree for predicting accuracy
   d. Classify no of people for suitable accommodations features
      i. Naive Bayes Classifier
      ii. Support Vector Machine
   e. Matching number of people with the suitable accommodation using clustering algorithms
      i. K-means
      ii. Hierarchical Clustering

6. **Include: description of dataset, file, or task, etc.**
    a. <u>calendar.csv</u>**:** This dataset contains availability of a listing on a date
    b. <u>listings.csv</u>**:** This dataset contains information of a listing
    c. <u>reviews.csv</u>**:** This dataset contains information of reviews of a listing provided by the user

7. **Data Sources:**
    a. www.kaggle.com
    b. www.airbnb.com

8. **Reference (if needed in the proposal)**

    a. https://www.kaggle.com/airbnb/boston
    b. http://insideairbnb.com/boston/