BRANDEIS UNIVERSITY

**International Business School**

Data Assigned Date: **October 21st, 2019**
Report Submission Deadline: **November 7th, 11:59 PM, 2019**

Submission: Please submit your report and the code file to **dac@brandeis.edu**. Each team should only submit one time by the team leader. The subject line should be **"Team Name-Data Analytics Competition Report"**.

# Part I: Dataset Description

**Dataset:** Airbnb in Boston, MA (January 2019 - September 2019).

**Source:** The original data was scraped from Airbnb website and made available.

**Description:** There are 5 files in total. For each one of them, it will contain the information from January 2019 to September 2019. Please see the detailed file information listed below.

| File Name | Description |
|---|---|
| **listings_details.csv** | Detailed Listings data for Boston |
| **calendar.csv** | Detailed Calendar Data for listings in Boston |
| **reviews_details.csv** | Detailed Review Data for listings in Boston |
| **listings_summary.csv** | Summary information and metrics for listings in Boston (good for visualizations) |
| **reviews_summary.csv** | Summary Review data and Listing ID (to facilitate time-based analytics and visualizations linked to a listing) |

Please feel free to implement analysis using any tool you would like (Python, R, SQL, Tableau).

# Part II: Data Analysis Report Instruction

In a professional quality report, you will describe and discover structure in your project dataset, train and validate your models, and interpret the best model. For each significant step, document your results, interpretations, assumptions, and process.

**Step I: Descriptive Statistics** (R packages: tidyverse; SQL, etc.)

- Show descriptive statistics for relevant and important variables in the dataset.
    - generate analytical charts (box-and-whisker plots, scatter plots, histograms, etc) for relevant and important variables.
    - demonstrate the minimum, maximum, mean, median, and standard deviation for relevant and important variables.

**Step II: Visualization** (R packages: ggplot2, shiny, ggmap, plotly, etc.)

- Create visualizations to answer the following questions:
    - How popular has Airbnb become in Boston?
    - What geography patterns appear in the Airbnb property listings?
    - Which neighborhood is the most popular among customers?

**Step III: Data Mining(**(R packages: data.table, rpart, caret, glmnet, etc.)

- Perform Multivariate Regression Analysis
    - **Target variable**: Daily Housing Price
    - **Independent variables**:

      Considering the following question: based on the attributes given in the dataset, what are the key factors that will affect daily housing prices?

      -- Use potentials ones as regressors and include in the regression model.
    - **Model selection**: Use forward selection, backward selection, forward-and-backward selection, or best subset (exhaustive) search. To ensure that the assumptions of regression are not being violated, check the diagnostic residual plots. These are common ones that help you identify outliers or leverage points:

      a) Histogram of Residuals. The residuals should look Gaussian, a bell-shaped curve.

      b) Normal Probability Plot of Residual. The residuals should closely track the diagonal line.

      c) Residuals vs. Fitted Values. The residuals should look randomly distributed.
    - **Result interpretation**: interpret coefficients and draw conclusions on your best regression model.

- Perform Machine learning
    - Predict daily housing price using machine learning methods you are familiar with(e.g. ridge, lasso, random forest regression, etc.)

Note: Please make sure to partition your dataset into training and testing subsets to implement the analysis and show your accuracy (RMSE, R-squared, etc) in the end.

**Step IV: Insight and Recommendation**

- Based on the above analysis, summarize your key findings for the following points:
    - How is Airbnb really being used in and affecting the neighborhoods?
    - Is there any trend of using Airbnb in Boston over time?
    - What recommendation you will make to Airbnb hosts and Airbnb?

**Step V: Appendix**

- Please append your code here for review.

**References**:

Here are some resources you might find useful:

1. Regression Analysis Essentials For Machine Learning: http://www.sthda.com/english/wiki/regression-analysis-essentials-for-machine-learning
2. Model Selection Essentials in R: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/
3. Scikit Learn: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning