# BUS212f: Analyzing Big Data II Fall 2018

## Final Project: Lessons from Airbnb in Boston

### Introduction

In this final assignment, your team can combine your data mining skills to answer some questions about the relationships between the Airbnb rental properties and the neighborhoods in the City of Boston. The project requires the entire process of a data mining project: data wrangling, dimension reduction, creative thinking, supervised and unsupervised learning, visualization, and coherent reporting. In other words, you have the chance to wisely apply nearly all your skills! Your project will use some data that I provide and must also incorporate relevant data from at least one other source available from the City of Boston (see below).

### The Business Case: Guiding Questions

This assignment allows your team to select a perspective for your analysis, and your first team decision is the selection of a point of view. The perspective that you choose should shape the analysis you’ll do.

* **City Council perspective**: Imagine that the Mayor and Boston City Council (the city’s legislative body) want to know more about the impacts of Airbnb rentals on neighborhoods. For example, which areas of the city see the most Airbnb activity? Are there positive or negative implications of the activity in terms of traffic congestion, crime, property values, etc.? Which specific impact do you want to focus on as a target variable?
* **Airbnb.com perspective**: Imagine that Airbnb wants to provide business intelligence to property owners who list properties and wants to use the Boston data to develop some models. Airbnb would like to be able to make recommendations about pricing strategies, length of stay, and ways to obtain positive recommendations. Here again, the team should decide on one focal target variable.
* **Airbnb lister perspective**: Imagine that you are an independent consultant hired by owners of properties in Boston. Your clients want to list their properties with Airbnb and to maximize their Airbnb revenue by adjusting prices and length of stay. Your model should control for relevant factors like neighborhood location, taxes, crime rates, and the quality of reviews received.

### Requirements/ Criteria

Your team should agree upon one of the perspectives, and then set to work asking a few (four or fewer) research questions that make sense from your chosen point of view. Each team should select only one perspective and begin your paper by stating with perspective you have chosen. Regardless of the choice, each team’s final submission must satisfy these criteria at a minimum.

1. You must create a predictive model for one target variable. Early in your process, after becoming familiar with the data, choose one or two target variables and potential input variables and covariates. Think about which modeling methods are most appropriate to the task you have chosen. You may later change your minds and work with a different target. Let the business problem guide your choices, and subsequent decisions. Ultimately, you must choose just one target variable; resist the temptation to use more than one.
2. You must somehow incorporate information from the **Reviews** text data in your model; decide what role the reviews can play in your model. What type of new variable(s) will you introduce to summarize the reviews? (e.g. do you want to create a new numeric or categorical variable representing sentiment?). One section of your final report should include a description and pertinent output to illustrate how you created informative variables from the reviews data.
3. You must also merge in some “external” data to play a role in your model. External data can come from one or more sources **in addition to** Airbnb. As an example, I’ve posted a .csv file from the Tax Assessor’s office in Boston, listing characteristics of all properties in the city (see boston\_property.csv and BostonPropertyDictionary). Use any of the columns in the boston\_property.csv file or find different relevant data to incorporate.  
     
   Your team should think about other external data that might be useful for your questions and seek it out. Some topical areas might include weather, crime, traffic, income, and so on. See the appendix at the end of this document for potential data sources. This is a place to use your curiosity, imagination, creativity, and resourcefulness.
4. After merging all data into a single data frame, you must partition your data into a training set (70%) and test set (30%). Be selective in adding columns to final data frame, since it could be quite large and could overburden an older laptop. Consult me for suggestions on handling an especially large set of data. Given the number of reviews, and the computing required for text mining, you may want to start with a small subsample of all Airbnb data.
5. Data exploration and cleaning always consume time and resources. Make deliberate decisions about missing data, dimension reduction, transformations, etc. Consider using an unsupervised method to simplify the number of levels of important categorical variables, or PCA for correlated numerical inputs.
6. It is up to the team to apply appropriate methods to the modeling problem you have defined. Every team should train, tune, and assess several (a minimum of three) methods. Your final report should include a **summary** of your assessments, as well as selective results from your model.
7. I suggest that you structure your final report as follows:
   1. Executive Summary: Problem Statement and Key Findings (write this last)
   2. Introduction: Perspective and problem statement
   3. Data sources—identify the variables you investigated, as well as identifying their sources
   4. Data exploration and initial investigation
      1. Create and discuss one or more graphs of the variation in the target variable.
      2. Display informative summaries of the input (independent) variables, as well as correlations or other indicators of association.
   5. Discussion of data cleaning and pre-processing: transformations, data reduction, missing data, etc. What did you do and why? Include relevant code.
   6. Discussion of modeling strategy: Prose discussion of the methods considered, tuning parameters, and comparison of model performance. Provide a meaningful comparison of evaluation metrics for the models that you trained, highlighting which one did best with your test data.
   7. Code, results, and discussion of the model you finally chose. Comment on the key insights revealed by the model.
   8. Conclusions: summarize the ways in which your model addresses the problem from your chosen perspective. Where is the business value?

### Data Files (csv or zip; all available on our GitHub):

#### FROM AIRBNB:

#### NOTE: Airbnb does not provide data dictionaries, but column names are mostly self-explanatory.

* Boston Listings 2018 Case 1 Main v3.csv:   
  same file you used for Cases 1 & 2
* reviews.zip: same file as you used for Case 4

#### FROM CITY OF BOSTON (You may use these files as part of the external data, and/or supplement with other helpful data)

* boston\_property.csv: tax-relevant variables for every property in the city
  + Size: 7,341 kb zipped; 22.7 mb as csv
  + 172,841 rows – each row is 1 property address
  + 23 columns – see BostonPropertyDictionary.xlsx
* BostonPropertyDictionary.xlsx: Definitions of the 23 columns in boston\_property.csv
* MA\_occupancy\_codes.csv: code definitions for 2 columns in boston\_property (PTYPE and LAND\_USE)
* boston\_zipcodes.csv: A small table of postal codes and corresponding neighborhoods and municipalities. NOTE: In the Airbnb listings you will find a column called “**neighbourhood\_cleansed**”. This column standardizes some neighborhood names and reduces the numbers of distinct neighborhoods. For example, Cambridge is reassigned as Allston, and Chinatown is grouped with other “Downtown” locations.

### Deliverable:

Your team should prepare a pdf document created with R Notebook (markdown) reporting on your analysis and discussing your conclusions, as outlined above. No slides this time.

# APPENDIX: Some Sources of External Data

Boston has a particularly well-developed system of datasets available for public use. You will find several of them at this site:

<https://data.boston.gov/dataset>

The U.S. Federal government also makes data available through these portals:

<https://www.census.gov/>

<https://www.data.gov/>

<https://www.usa.gov/statistics>

And don’t forget the LTS Business & Economics Research Guide or the Bloomberg Lab!

<http://guides.library.brandeis.edu/friendly.php?s=business>