# Project Proposal

1. **What are we trying to accomplish in this project?**

We want to predict the price range of potential new listings, and to analyze the current Airbnb market in New York City and come up with interesting insights regarding the relations between types of rentals and prices in different locations.

1. **Project description/Abstract**

We chose this dataset because we can relate to it on both personal and professional level. We all have used Airbnb before, either as hosts and/or guests, and love the concept and usability of it. New York City is one of Airbnb’s most complex markets, as the housing regulations here are very unique and it’s illegal to own or lease an apartment for the sole purpose of posting it on Airbnb. In this project we will dive deep into NYC’s Airbnb market data and come up with conclusions about the different segments in this market, and what determines the prices. When debating where to live next, the potential of money that can be gained through Airbnb can be a factor in the decision process.

1. **Introduction**

In this project we will focus a lot on what determines listings’ prices, while considering where they are located and the listing type. We want to know how many of NYC’s listings are meant for full-time renting harming the city housing market, and how many are meant for the original concept of Airbnb - taking advantage of the market inefficiency and allowing people rent out their rooms/apartments/houses when they are away.

We want to see which type of listings are located in which borough, and how are they being priced out. We are going to check the correlation between listings prices and different features in order to be able to predict the prices of potential new listings.

Potential questions to be answered:

* What is the distribution of different listing types (entire apt. / private room / shared room) and availability in each borough and neighborhood?
* Which boroughs and neighborhoods have the most hosts with multiple listings?
* What is the relationship between price and number of reviews, availability, number of host’s listings, borough, neighborhood?
* What is the most expensive type of listing considering the location, number of reviews, last review?
* What is the relationship between a listing type and borough / neighborhood?
* What is the relationship between minimum nights and location, price, and availability?

1. **Dataset being used and description of the dataset**

We use “New York City Airbnb Open Data” from the following link: <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data#AB_NYC_2019.csv>

The dataset contains 48,895 observations about the following features of NYC’s Aribnb market, last updated in August of 2019:

**Id-** listing ID

**Name –** name of the listing

**host\_id -** host ID

**host\_name -** name of the host

**neighborhood\_group –** Borough

**neighborhood –** neighborhood

**latitude –** latitudecoordinates

**longitude -** longitudecoordinates

**room\_type -** listing space type

**price -** price in dollars

**minimum\_nights -** amount of nights minimum

**number\_of\_reviews -** number of reviews

**last\_review** - latest review

**reviews\_per\_month** - number of reviews per month

**calculated\_host\_listings\_count** - amount of listing per host

**availability\_365** - number of days when listing is available for booking

1. **Literature Survey (existing work on this dataset)**

We came across few different articles that used the exact same dataset, and some more that used similar ones that included more responses, such as neighborhood average income and listings’ ratings. We focus on problems that we can solve using our dataset and that might have not been addressed in the articles we read. From surveying relevant literature, we found that the borough of Manhattan has the most entire apt. listings, and that Brooklyn and Queens have more private room listings. Across all boroughs the shared room listings were significantly rarer.

Another borough specific hypothesis we came across was that Brooklyn and Queens have the most hosts with multiple listings, and we wonder why this is. We also read that two thirds of listings come from hosts with only one listing, and don’t have multiple ones. We want to test if it’s true that there is a relation between illegal rentals and the availability in certain neighborhoods.

1. **Project methodology**

* We’ll be using **classification** method to divide the data into different categories in order to check relations between a specific response and different predictors.
* We’ll use **clustering** to group quantitative data and examine relations between them.
* We’ll try to **predict** price range of listings for different neighborhoods and listing types.
* We’ll use **multiple linear regression** in order to check relations of quantitative types of data.

For example: examine the relations between price and different quantitative predictors.

* We’ll use **decision tree** to check if hosts post listings to function as a hotel vs. listings that are actually meant to fill short periods of time of absence.
* We’ll use **logistic regression** to examine relations of qualitative types of data.

For example: which neighborhoods have what percentage of each listing types and what number of hosts’ listings.

1. **(9) References**

* <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/kernels>
* <http://insideairbnb.com/new-york-city/>
* <https://towardsdatascience.com/a-demonstration-of-carrying-data-analysis-new-york-city-airbnb-open-data-a02d0ce8292d>