**Factors behind local rental properties pricing in New York City**

**Santiago Rodrigues Manica**

**27th, June 2020**

**Introduction**

**Background**

Imagine you have the luck of owning real estate in New York City, which is available for renting. Since you live in the 8th most visited city in 2019 (<https://edition.cnn.com/travel/article/most-visited-cities-euromonitor-2019/index.html>), you know there is a high demand for accommodation. Since traditional hotels are especially expensive in the USA more are more tourist are looking for local and affordable accommodation. There is no surprise until 2019 companies like AirBNB kept growing ( <https://news.airbnb.com/airbnb-2019-business-update/>). And there is no secret, for good and for bad, that renting to short-term tourism can bring higher revenues versus having income from a long-term tenant.

**Problem**

In this case you may be a clueless owner trying to guess what factors may influence your expected pricing. Since it may be, for example, due to the neighbourhood where the rental property is located, or may be affected by convenience factors, such as proximity to restaurants and entertainment.

**Interest**

In case you are a client curious about which factors may affect the price you could charge your guests, we will explore these data.

**Methods**

City of interest for this project: **New York City**, NY, USA

**Datasets of interest**

**New York’s AirBNBs** (csv): <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data?select=AB_NYC_2019.csv> a freely and publicly available dataset on Kaggle, which I stored as a csv file in my GitHub account. <https://github.com/RM-Santiago/Coursera_Capstone/blob/master/AB_NYC_2019.csv>

**Foursaquare application programming interface** (API): <https://developer.foursquare.com/docs/places-api/> This API will be used to obtain the venues around the rental properties and will be useful for both exploratory data analysis (EDA) and inferential analysis.

**Research questions and statistical methods:**

After proper data management and EDA (including mapping and clustering), the project will try to answer to the following questions using libraries that allow data frame analysis and statistical testing (eg; Pandas):

* 1. **Is the average price different between neighborhoods?**

Using the **New York’s AirBNBs dataframe** the mean and standard deviation of the price will be described across different neighborhoods and differences will be tested.

* 1. Is there an association between the average price and the neighborhood?

A simple linear regression will test the association between price (outcome/dependent variable) and the neighbourhood (categorical independent variable).

* 1. Is the average price different between whole apartments and rooms?

Using the **New York’s AirBNBs dataframe** the mean and standard deviation of the price will be described across different types of accommodation (eg; whole apartment vs room only) and differences will be tested.

* 1. Is the average price different according to the number of venues nearby?

Using the **New York’s AirBNBs and adding the local venues from the API**, the mean and standard deviation of the price will be described across different neighborhoods and differences will be tested.

* 1. Is there an association between the average price and the number of venues nearby?

Using the **New York’s AirBNBs and adding the local venues from the API**, a simple linear regression will test the association between price (outcome/dependent variable) and the number of nearby venues (continuous dependent variable).

* 1. Considering the simultaneous effect of all candidate factors which may be associated with the price of a rental property?

Using the **New York’s AirBNBs and adding the local venues from the API**, a multiple linear regression will test the association between price (outcome/dependent variable) and a set of dependent variables (neighborhood and number of nearby venues).

In order to test differences of price in USD (continuous variable) between groups the following statistical tests may be used;

* Student’s t-test; comparing two independent groups if prices have a normal distribution;
* Wilcoxon-Mann Whitney test; comparing two independent groups if prices have a non-normal distribution;
* Analysis of covariance (ANOVA); when comparing prices across more than 2 independent groups if prices have a normal distribution;
* Kruskal Wallis; when comparing prices across more than 2 independent groups if prices have a non-normal distribution.