Nil Distefano u172948
Pol Ayala u173483

Visual Analytics

Final Project Report

New York AirBnB analysis

1. Project description

New York is a very transitory city and hence short term homestays and rentals are very common. This service is provided mainly by AirBnB and in this project we will analyse one of its datasets, Airbnb Open Data provided by Arian Azmoudeh. This dataset contains information such as the host name, the type of the rental (house, room…), information on the reviews and price. In addition, it also provides geolocation data (coordinates) for every listing.

The aim is to retrieve as much information as we can from the source of data that we have in hopes of being able to predict the price of an AirBnB given its characteristics.

1. Which borough[1] has more AirBnBs? Is the average price related to each of the different boroughs?
2. Which is the most common room type? Is it related to lower prices?
3. Are listings with high review rates more expensive than others?
4. Is the distance between the AirBnB and the nearest subway station an influential factor to determine the price?

2. Process and methodology

2.1. Data Analysis and visualisation

Following the usual machine learning workflow we started by taking an overall look at our data and making some cleaning. Our dataset has a total of 17 columns, some of which are worth mentioning in the report:

- Categorical:
  - `neighbourhood_group`: Essentially the New York Borough of the entry. This is a categorical feature with five unique values and no missing entries.
  - `subway_proximity`: Discussed in 2.2.
  - `room_type`: Determines the type of housing, i.e. House, private room or shared.
- Numerical:
  - `latitude` and `longitude`: Floating point values that geographically locate each airbnb.
  - `price`: Monthly cost of the airbnb that we will use as the main target for our study.

After removing missing values we extract some insights and relationships. See Results section.

2.2. Location intelligence

The goal of this section is to classify each listing based on their distance to the nearest subway station. To do so, we will use service areas based on the streets of New York and the position of the subway stations. The shapefiles for this information are available in NYC Open Data.

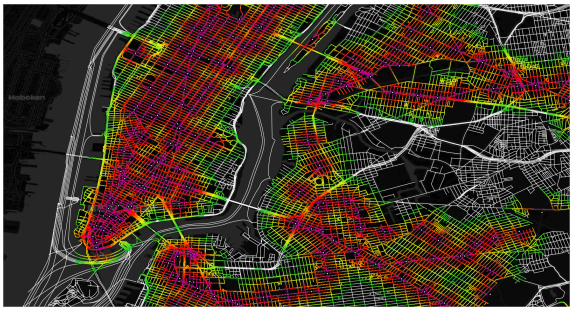
New York Streets

New York Subway Stations

With this information we can determine multiple service areas on our railway layer based on the subway stations layer. For this particular case we will generate 10 service areas with maximum distances: **100**, **200**, **300**, **400**, **500**, **600**, **700**, **800**, **900** and **1000** metres.

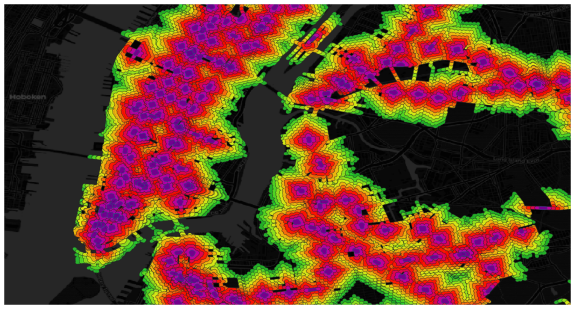
[1] *Borough: a town or district which is an administrative unit. In New York there are five boroughs: Bronx, Queens, Manhattan, Staten Island and Brooklyn.*

Nil Distefano u172948
Pol Ayala u173483

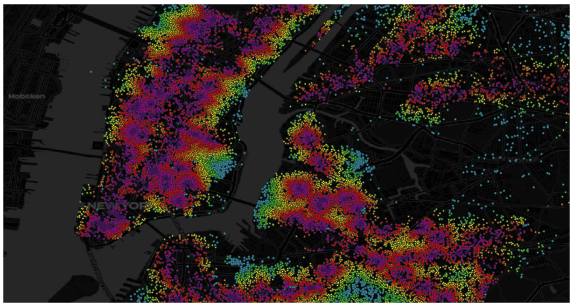


Service areas overlayed

Next we import our dataset containing each AirBnB. We will consider that an AirBnB is within a service area if it falls within 100m of any point in that area.



Service areas with 100m influence



AirBnBs in each service area

Finally we export eleven datasets, each containing the listings in each service plus an additional one containing the ones not included in any of them, i.e. those at distance >1000m. These datasets will be combined with the original dataset to obtain a new feature `subway_proximity`. Comparison with the average prices of the airbnbs at different ranges will show the effect of the proximity. See Results section.

2.3. Prediction and model explainability

For this last section we will train a simple model and analyse its behaviour with the SHAP library. Before doing so, we need to prepare the data as some features still contain categorical variables that need to be encoded and other features that can be dropped.

2.3.1. Dropping irrelevant features and encoding categorical variables

Here are the correlations of every numerical feature with our target variable:

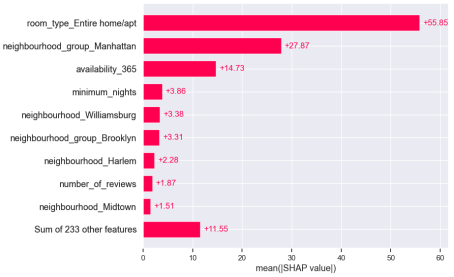
	longitude	number_of_reviews	reviews_per_month	id	host_id	minimum_nights	latitude	calculated_host_listings_count	availability_365
price	-0.155298	-0.035924	-0.030623	-0.006696	0.006263	0.025501	0.031344	0.052895	0.078276

We decided to drop those that had no direct relation to the price or had very small correlation, for example, `host_id` or the geographical components. Then we applied a one-hot encoding to every categorical variable like `neighbourhood`, including also our generated feature `subway_proximity`.

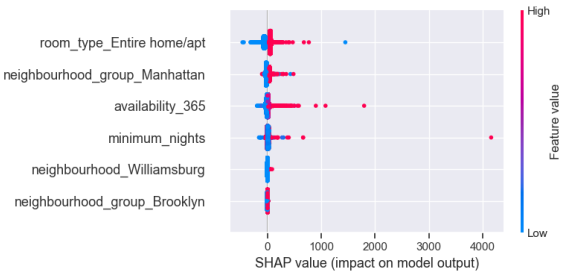
2.3.2. Model training

We trained three different models: Linear Regression, Random Forest Regression and Gradient Boosting Regression. Nonetheless, we will stick with Random Forest as it is the one that yielded the smallest mean squared error for the test set.

2.3.3. SHAP visualisation



SHAP bar plot



SHAP waterfall plot

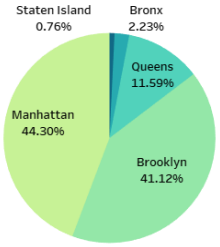
Nil Distefano u172948
Pol Ayala u173483

Our model is influenced by three features. These being weather room_type is Entire home/apt, if the airbnb is in Manhattan and its availability. The first two features were kind of expected given our initial analysis, nonetheless, the appearance of availability seems rare. A deeper look in the waterfall plot shows that in some cases high availability comes at a higher price. We will further discuss this in the next section.

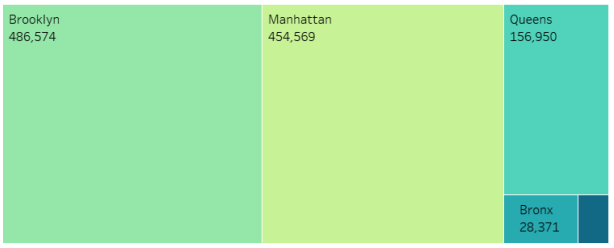
3. Results

With all the information gathered from the above we can go back to our initial questions in hopes of finding some useful insights.

1. **Which borough has more AirBnBs? Is the average price related to each of the different boroughs?**
   Manhattan and Brooklyn, the most wealthy boroughs in New York, is where most of the AirBnBs are situated and where prices are higher as well on average ($180 and $121 respectively). The prices are around twice the average price in the other boroughs. Bronx comes last where the average price is only $79. Consequently, Manhattan and Brooklyn are also the boroughs with higher reviews.



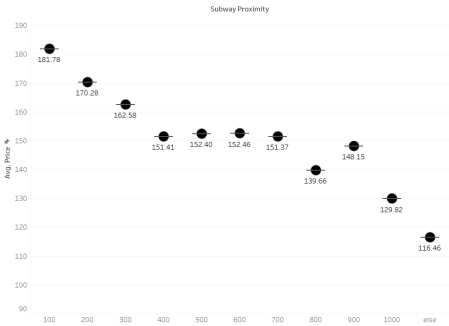
Listings distribution by borough



Reviews by borough

2. **Which is the most common room type? Is it related to lower prices?**
   The room type options and prices are distributed as follows:
   - 52% offer an Entire home/apartment at an average price of $196.
   - 45% offer a Private room at an average price of $83.
   - 2% offer Shared rooms at a cost of $63.

   Even if private rooms are around 20 times more common than a shared room, their price is barely different. This trend is also present when taking into account the borough in which these listings are located. Manhattan shows prices over 50% of the mean price for an Entire home/apartment.
3. **Are listings with high review rates more expensive than others?**
   Cheap AirBnBs, below the total average price ($140) benefit from high review rates. Especially Private rooms which get more than 10 reviews per month in some cases.
4. **Is the distance between the AirBnB and the nearest subway station an influential factor to determine the price?**
   When looking at the price distribution at each distance interval we observe a similar distribution. Nonetheless, as we get away from the subway stations, the distribution mean decreases and the percentage of lodgings with low prices increases. Furthermore, when taking a look at the absolute mean price in each distance interval this tendency is even more clear. While under 600m of any subway station prices are over the mean ($140), as we get further from these, prices reach minimum prices as up to less than $110 at more than 1000m away. The difference between



Average price per proximity interval