FINAL PROJECT

DATA VISUALIZATION - 6380



GROUP 4

Guruprasad Annadurai

Alice Bergeret

Gogul Vengatesh Krishnamoorthi

Shyam Pandalanghat Balasubramanian

Siddique Ansaaruddin Syed

**TABLE OF CONTENTS**

1. Introduction                                    Page2
2. Dataset choice: why?                                Page2
3. Finding and insights                                Page4
4. Methods used                                    Page7
5. Conclusion                                    Page8
6. References Page9

**1.Introduction**

We are not just a company who intend to help people analyze the future of their flat in Airbnb or open a hotel, but also for people who want to find a great place to stay in NYC.

On one hand, people can run a real business by renting their properties. Yes! It allows you to have an additional income. Yet, it still requires you to take the right decisions before register your flat on this application. Is it really worthy to rent via Airbnb, or no?

Owners of real estate have multiple choices, particularly renting it on their own without using Airbnb. That’s what our company does: make it clear for you to take a decision.

On the other hand, tourists who don’t know NYC accommodation really well need information to make the right choice: Best Place, Right Price and Trustworthy.

**2.Dataset choice: Why?**

The reasons we think our dataset is reliable are:

* **Integrity** : This public dataset is part of Airbnb, and the original source can be found on its website (<http://insideairbnb.com/new-york-city/>)
* **Content**: This data file includes all needed information to find out more about price, popularity of place, type of renting you can find
* **Context**: This dataset is recent, it has been released in the middle of 2019
* **Company we choose**: A‌i‌r‌b‌n‌b‌,‌ ‌I‌n‌c‌.‌ is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking. It is very popular for short renting.

In terms of data acquisition, we have one primary dataset: it’s enough for our purpose.

(Source of our dataset : [https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/version/3#](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/version/3))

Our dataset contains records from flats which are registered on Airbnb, in New York City. There are many neighborhoods, and the dataset grouped them by 4 districts to make possible the manipulation of granularity. The dataset includes latitudes and longitudes, that help pinpoint the exact geo location. In our opinion, the aim metrics are here: Price, Type of rooms (3 types) and popularity we estimate with number of reviews.

The group work was about analysis, and we knew that we can’t do analytics on non-numeric data. That’s why we didn’t look for Airbnb dataset with text reviews.



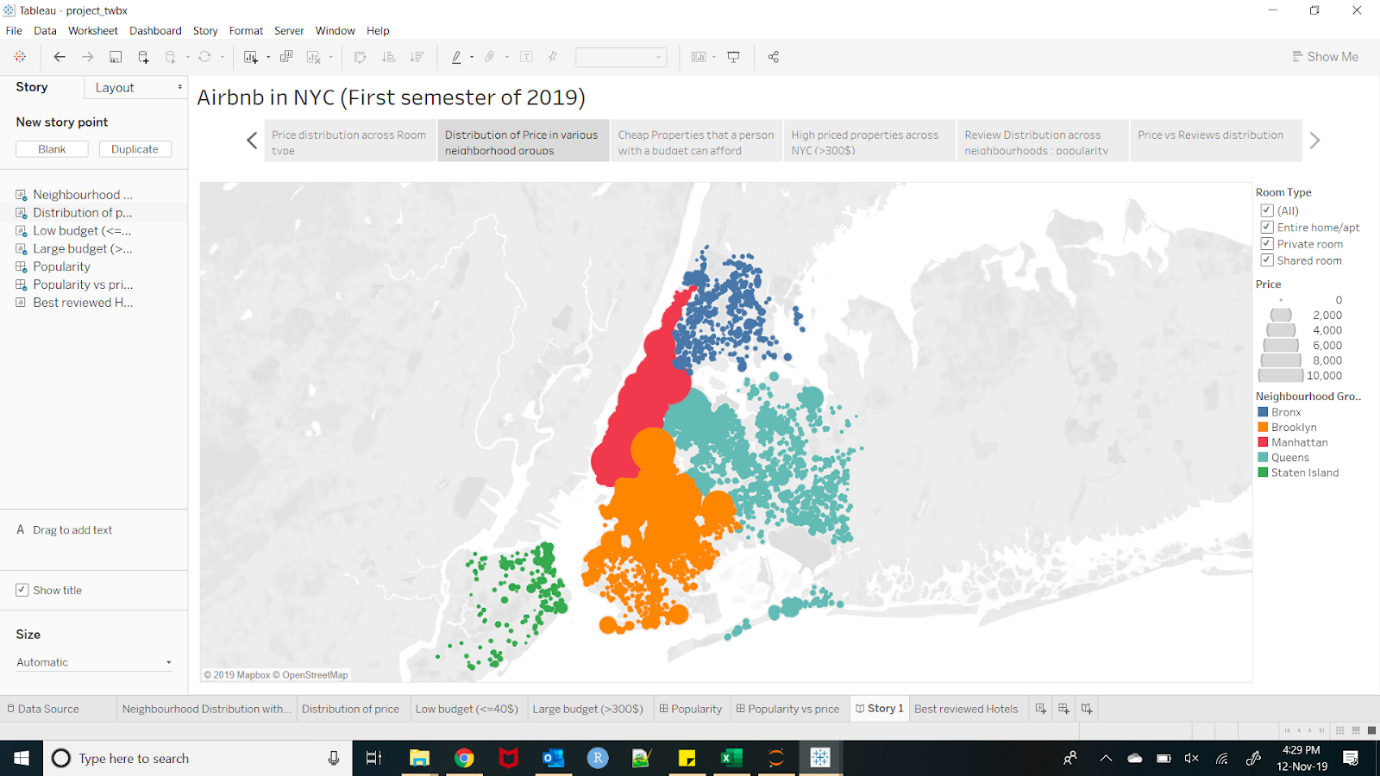
*Map of NYC districts with its two main airports*

**3.Finding and insights:**

By manipulating our dataset with Tableau, we found some insights:

**Most expensive place**: We used a map with price distribution to figure out which district is the most expensive. The fact we could see the data with their geographical location allows us to find this insight.

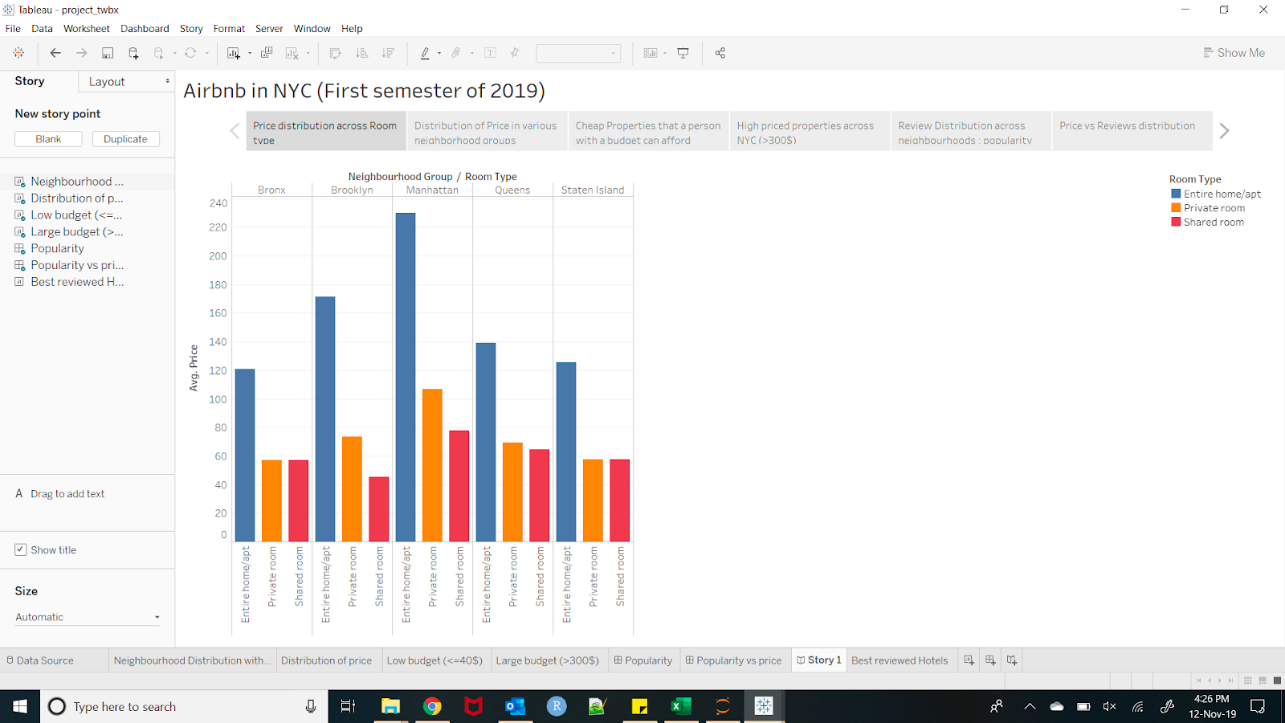
* Insight: Manhattan, undoubtedly, the more attractive place, is the most expensive place to stay. We can see a large difference, the more we move away from Manhattan, the more the prices begin to drop. This is shown by lower prices in districts like Staten Island or the Bronx.



**Room type available and their price**: Three types of rooms are available in Airbnb: Entire flats, Private rooms and Shared room. The idea was about discover which type of rooms are more available in which districts, and for what price. It provides information about what you can rent, and where. The price is interesting here because it shows the value of different types of room according to the district. An entire room in Manhattan doesn’t have the same value as one in Brooklyn.

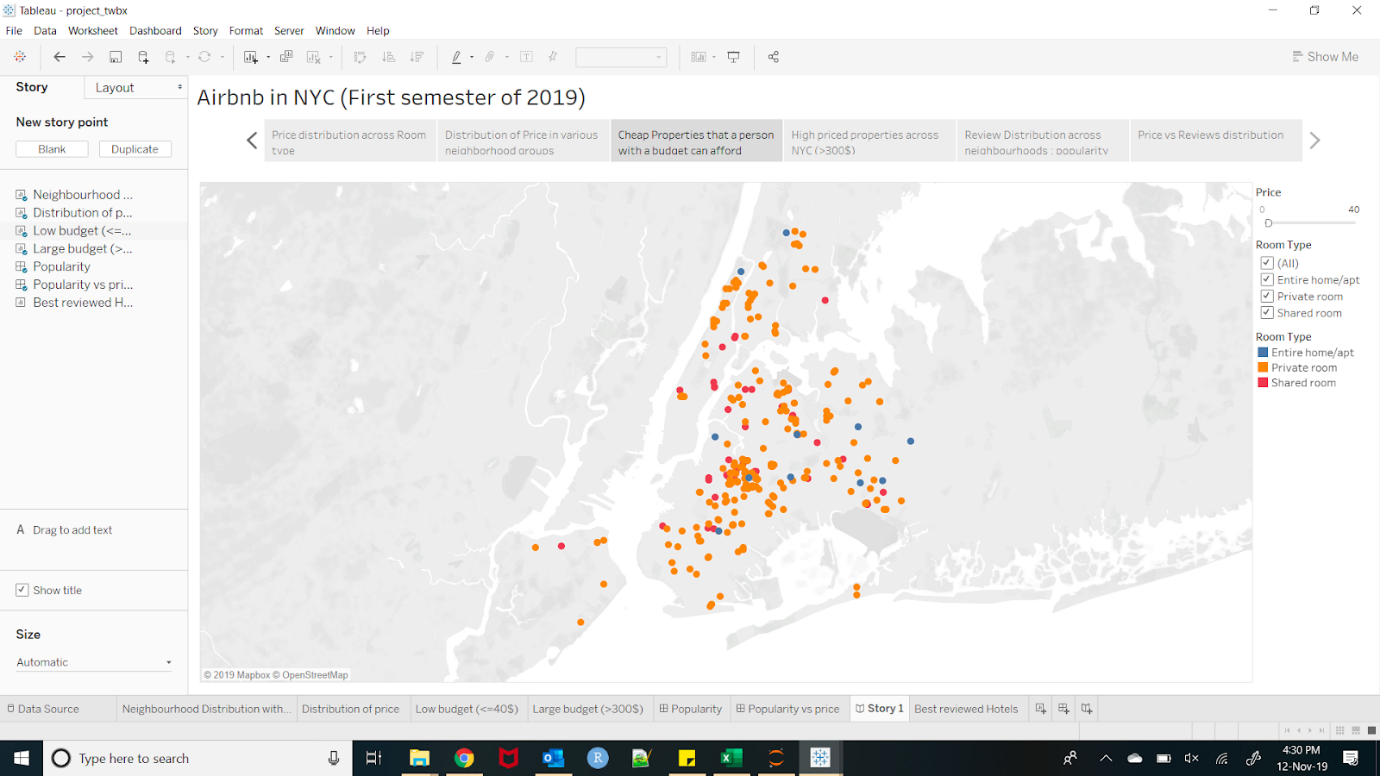
* Insight: Entire flat represents the majority of the available rental in all the districts. Yet, the average price for Manhattan is high, about 230$ for one night. Brooklyn is the second most expensive place, with an average cost of 170$.

The best option in terms of saving money is to rent a shared room in Brooklyn, for an average price of 40$.



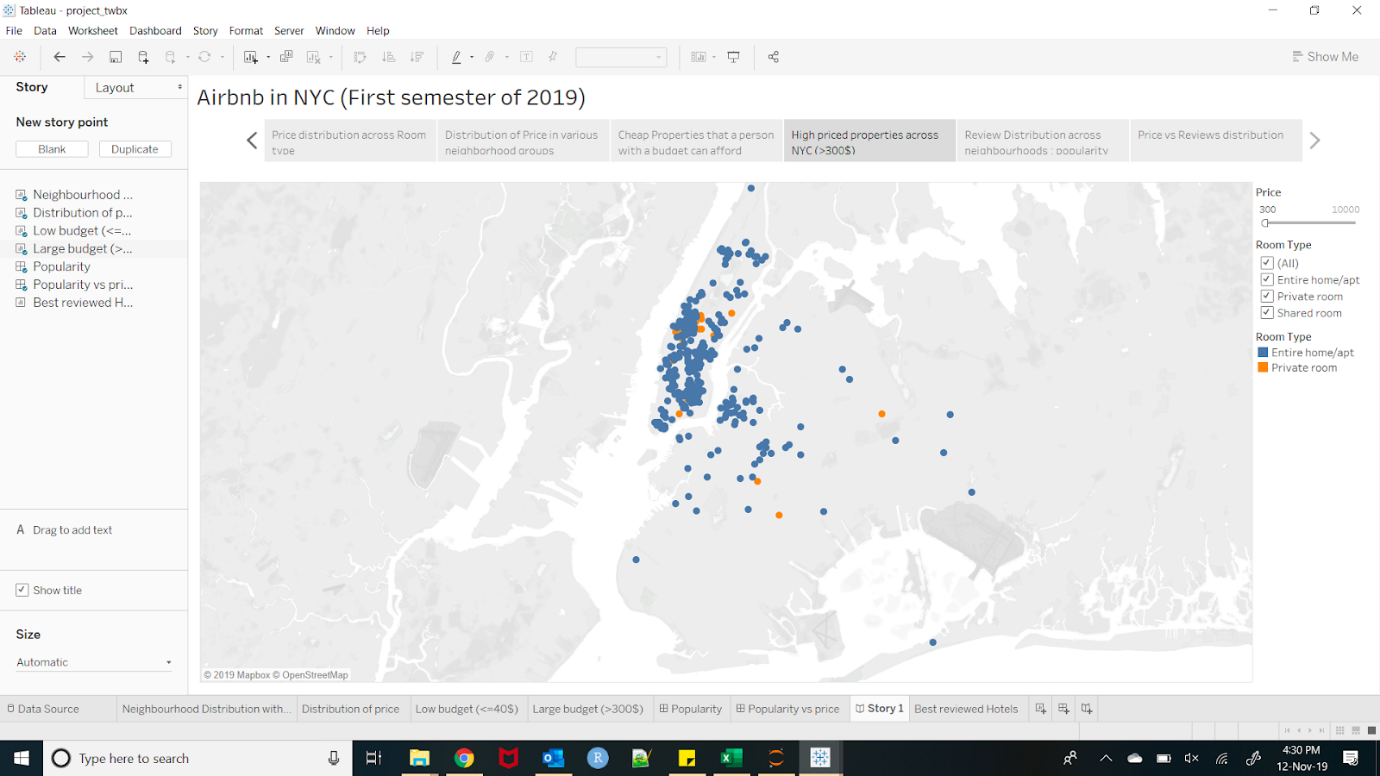
**The cheapest place to stay**: It is possible to stay over at NYC at $40 per night, thereby proving our hypothesis.. We apply a filter on the map to figure out where we can find these cheap properties.

* Insight found: Most of these places are located in Brooklyn.



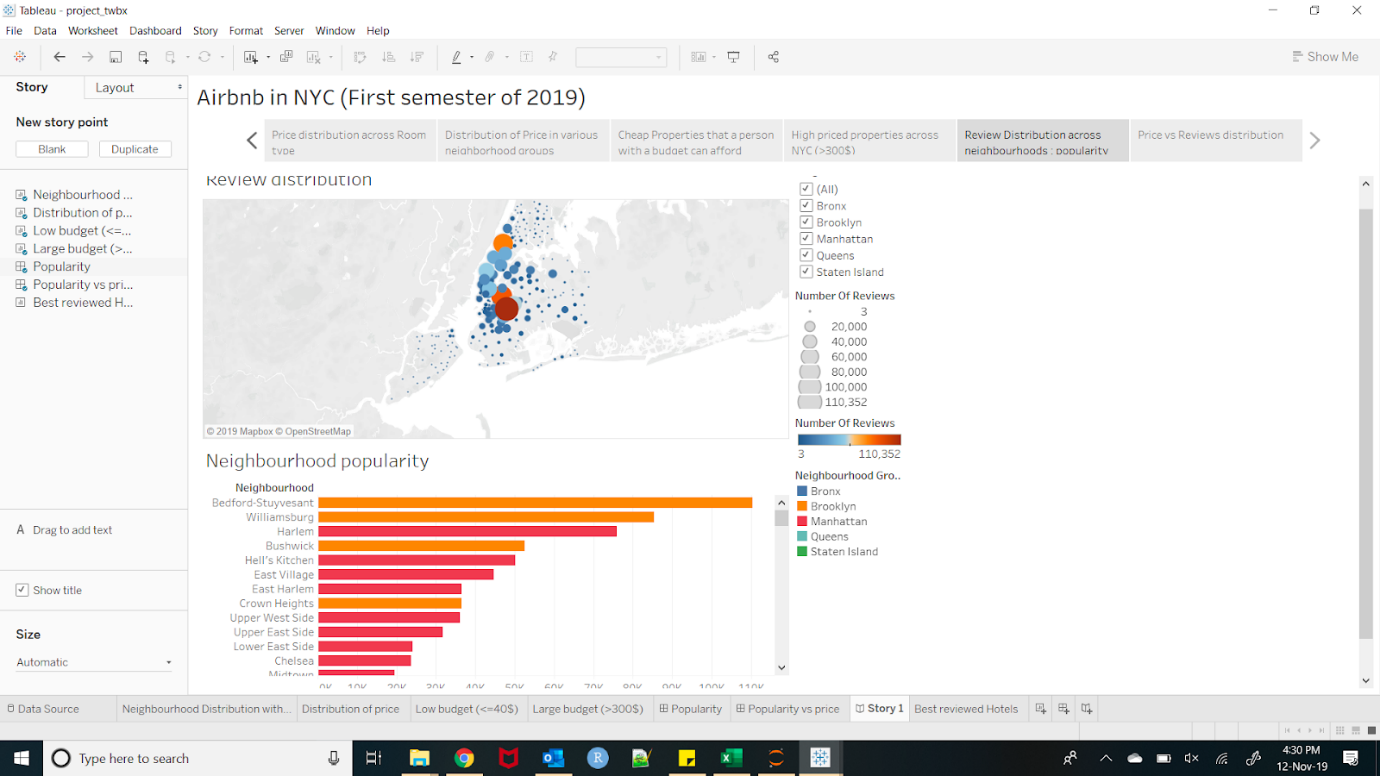
**Large budget**: It’s well known that NYC is an expensive destination. That’s why during the last insight we found on Tableau, focused on it. We applied a filter to show only the renting which costs almost 300$ for a night.

* Insight: We clearly see than Manhattan propose that type of rental, besides it’s the district with the highest attractions. We recommend people with money to spend, who want to experience the intricacies of NYC to book in this area.



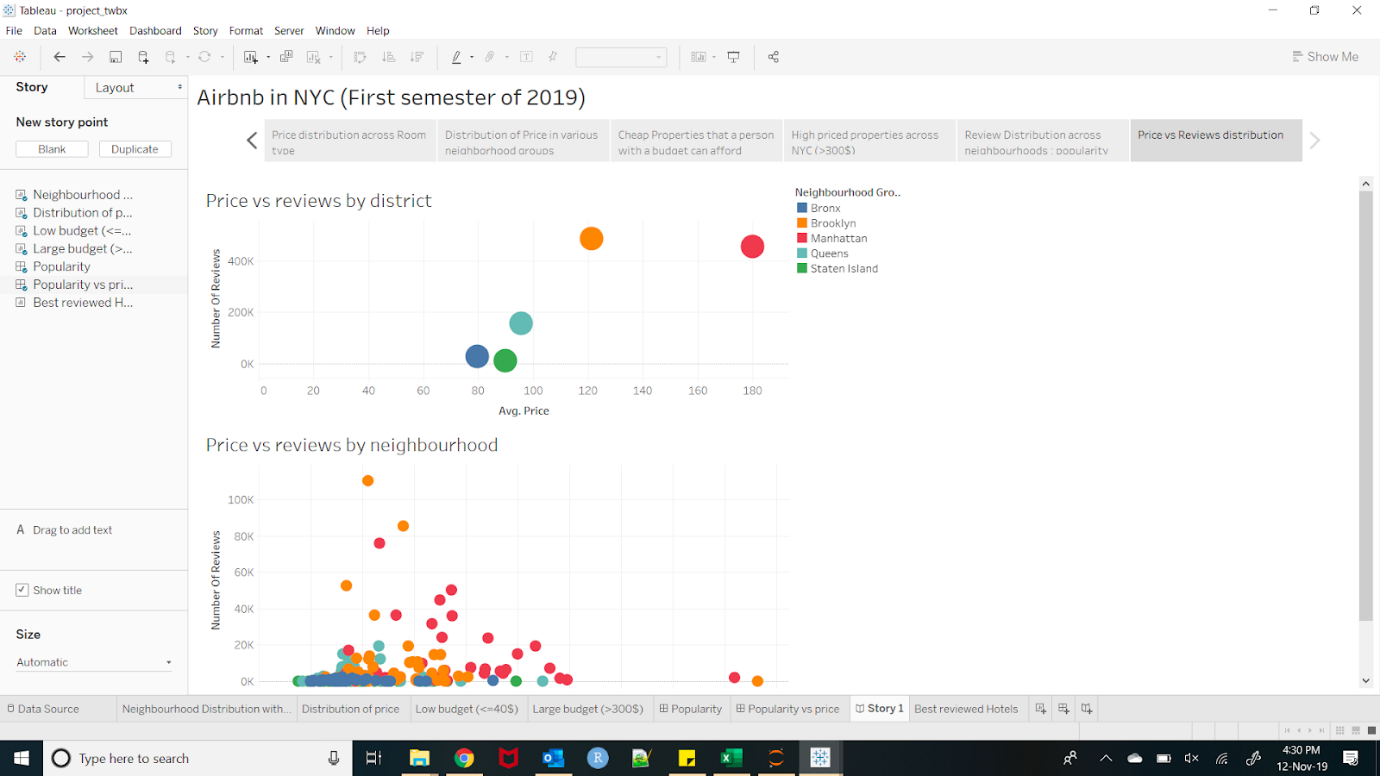
**Most popular place**: We know that the number of reviews shows the popularity of a place. The more reviews a place has, the more popular it is. We decided to use two different visualizations to precise our insight.

* Insight: We can see that the most popular places are located in Manhattan.



**Popularity vs price**: Here is an insight we found relevant. What we can find was if the price was a difficulty for the visitors. If the number of reviews is low and the price is high, it proves that visitors don’t want to spend too much money for their renting. Inversely, a high number of reviews with a high price proves that the price isn’t an important factor.

* Insight: We notice is that Manhattan has the highest price but compensates with a good popularity. Behind it is Brooklyn which registers a good popularity with lower prices. We definitely can say that the other three districts, Bronx, Staten Island and Queens are not much appreciated despite their small prices.



To complete these insights, we have decided to create a program to make *predictions*. They are more for business interest. Indeed, our data set didn’t have a time dimension. Here are the insights we found out:

**Influence of Airports**: When you travel somewhere, one of the most used means of transports is the aeroplane. So, we have taken the number of reviews for each hotel which is more than 15 (to filter the most popular) and compute their distance with two airports.

* Insight: 80% of the hotels are near to the Airports. This indicates that the number of people booking near the airport is high. So, if a new hotel is being opened, proximity to the airport would be an asset.

**Prediction of prices**: We have made a prediction of prices for a room based on the different Regions in New York City. In the case of opening a new Hotel in the city, one can give out the details of the hotel and find out the optimum prices for the room. This is done with the help of a Randon forest model which takes the current City data and predicts the future prices of the hotel rooms for a particular Neighborhood.

The price prediction can be a major factor in market analysis. It can be used by a new hotel owner as well as an existing owner to predict the market price in their geographical location.

**4.Overview**

**Cleaning of the dataset**

We noticed that less than NULL values represent less than 2%, that’s why we decided to delete it. We used Excel. To clean the dataset, we filter the rows with null values, and delete it. We then used python to impute a few missing fields with median value of the attribute. The dataset was quite clean, that’s why we didn’t have to use many techniques of cleaning.

**Visualization: Choice of Tools**

We decided to use Tableau, because this software is intuitive. You can use Excel files with it, and it is the format of our dataset. Besides, we created reports using the dashboard function, and it helped us to define the conduct of the video.

**Finding the insights**

On Tableau: We decided to start manipulating the data with a map. Indeed, the data set provides us locations (latitude and longitude) data, so it was easy to display it. It allows us to start finding insight by adding some measures like prices, number of reviews or type of rooms. We thought that it was a good way to display our visualizations in the video too, because it was linked to our project.

**Programming:**  
We predict the price of hotels using linear regression and Random Forest Regressor in python. The latter seemed to provide more accuracy when tested on a validation set. Hence we have set the same as our final model.

**5.Conclusion**

For all our visualizations and analysis of the data, we used various types of graphs: maps with plot distribution, bar charts and scatterplot.

It allows us to produce a complete study to provide different types of users with information about the Airbnb application.

On one hand, tourists are able to know that the best deal is Brooklyn: cheaper and popular. If they prefer to have more choice of entire flat, but with the highest price, Manhattan would be a better choice.

On the other hand, for people who wanted to register their flat on Airbnb, they can figure out whether they put their flat into the perfect price range. They can know that the distance from airports is a critical factor for the price.

**6.References:**

1. Kaggle – website to obtain the datasets
2. Youtube – Clips used in video presentation
3. Shuttershot – Images and gif used in the video presentation
4. Moovly.com – Clips used in video presentation