Storytelling Case Study: Airbnb, NYC

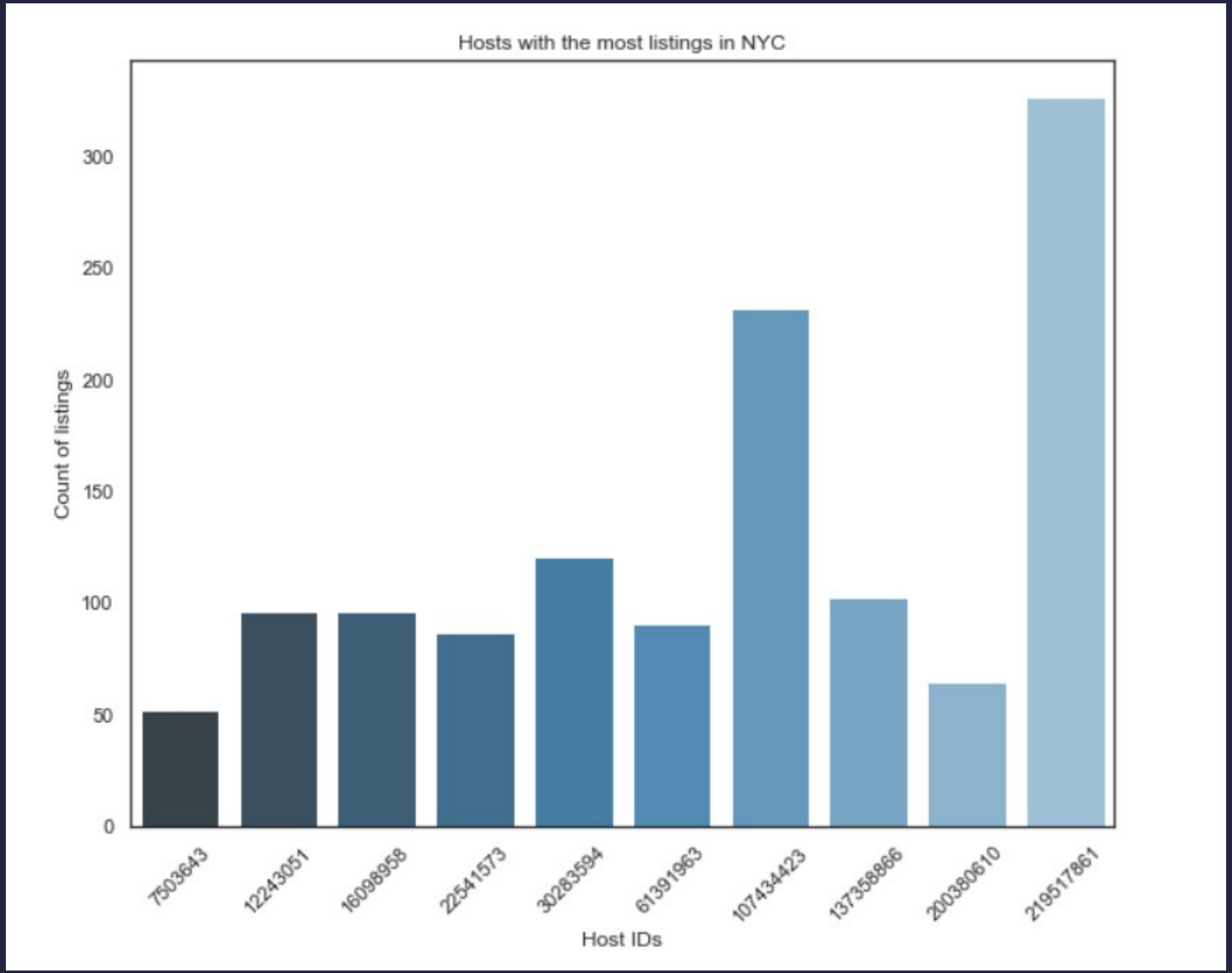
-SANDEEP SINGH, DEVIDUTTA PRADHAN

OVERVIEW OF DATA

- → This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.
- → After loading the dataset AB_2019_NYC we make the following observations. The 16 columns provide a very rich amount of information for deep data exploration we can do on this dataset. We do already see some missing values, which will require cleaning and handling of NaN values. Later, we may need to continue with mapping certain values to ones and zeros for predictive analytics.
- → After cleaning the data, we start visualizing the same.

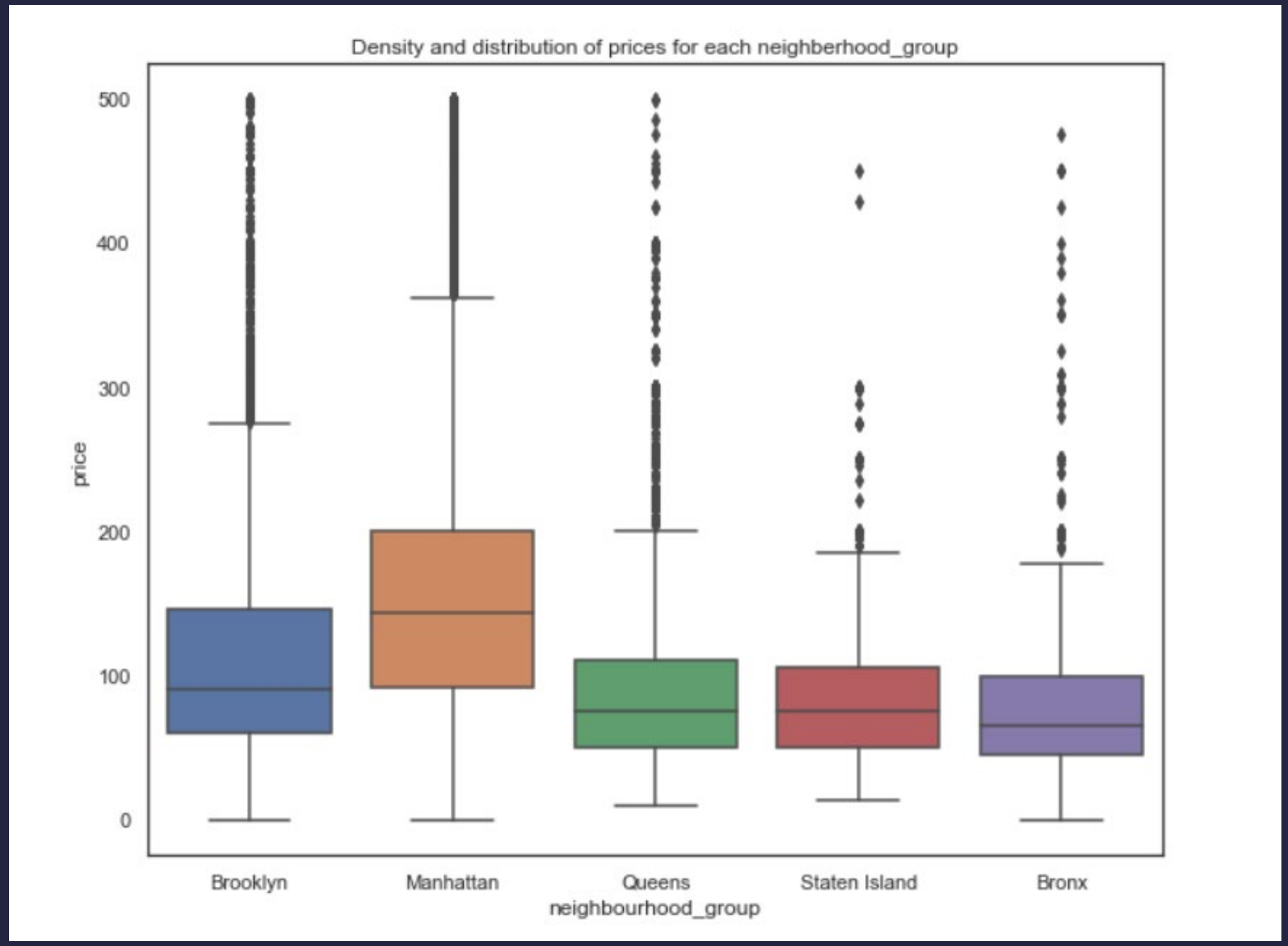
HOSTS WITH MOST LISTINGS

- → From here we can observe there are 300+ listings for host with 219517861 ID.
- → Whereas host with 7503643 ID has a minimum with 50 listings.



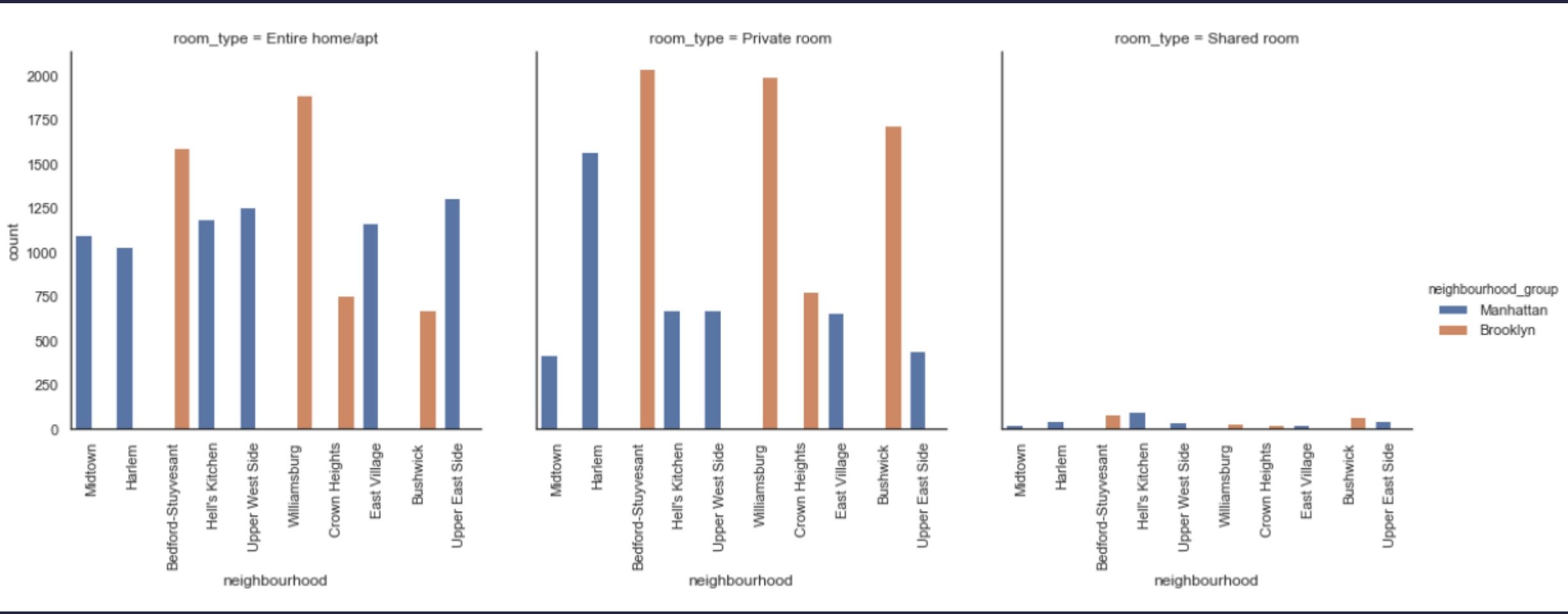
PRICE DISTRIBUTION OVER NEIGHBOURHOOD

- → we can state that Manhattan has the highest range of prices for the listings with $150 price as average observation, followed by Brooklyn with $90 per night.
- → Bronx is the cheapest of them all.



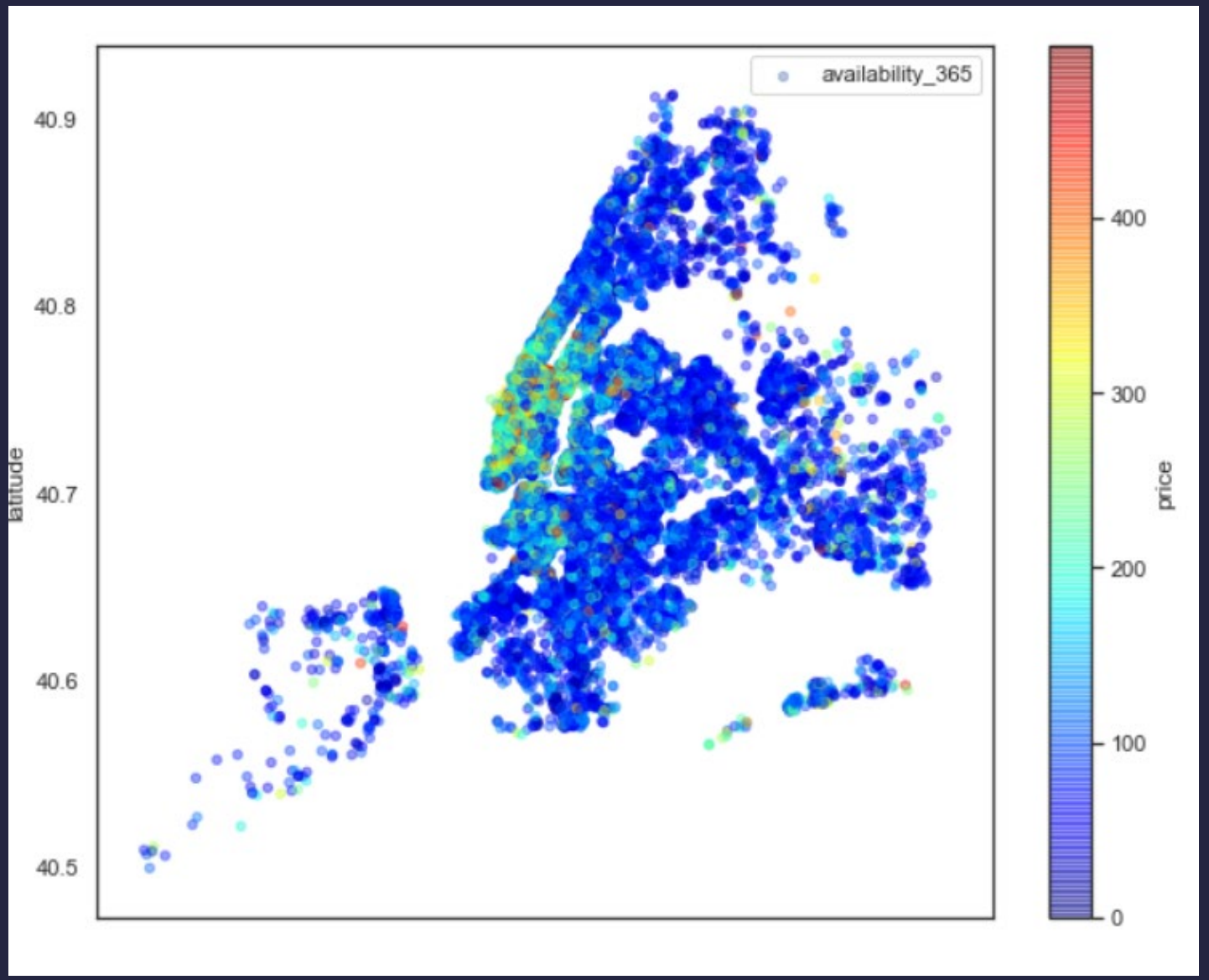
ROOM TYPES OVER NEIGHBOURHOOD

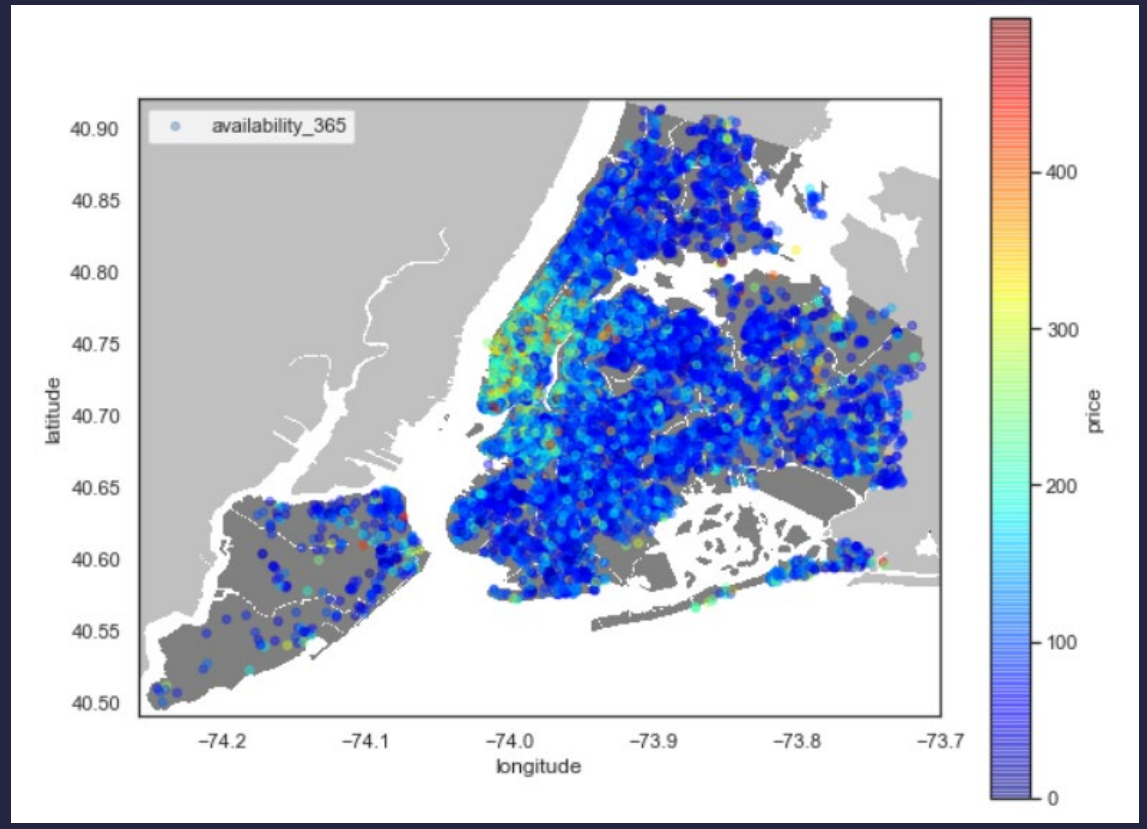
→ Here we can see the 3 types of room available in Manhattan and Brooklyn.



SCATTER PLOT OF AVAILABLITY

- → The following scatter plot is to depict the longitude and latitude of the areas with available rooms.
- → But it is hard to interpret without a map.
- → We download the map from https://upload.wikimedia.org/wikipedia/commons/e/ec/Neighbourhoods_New_York_City_Map.PNG



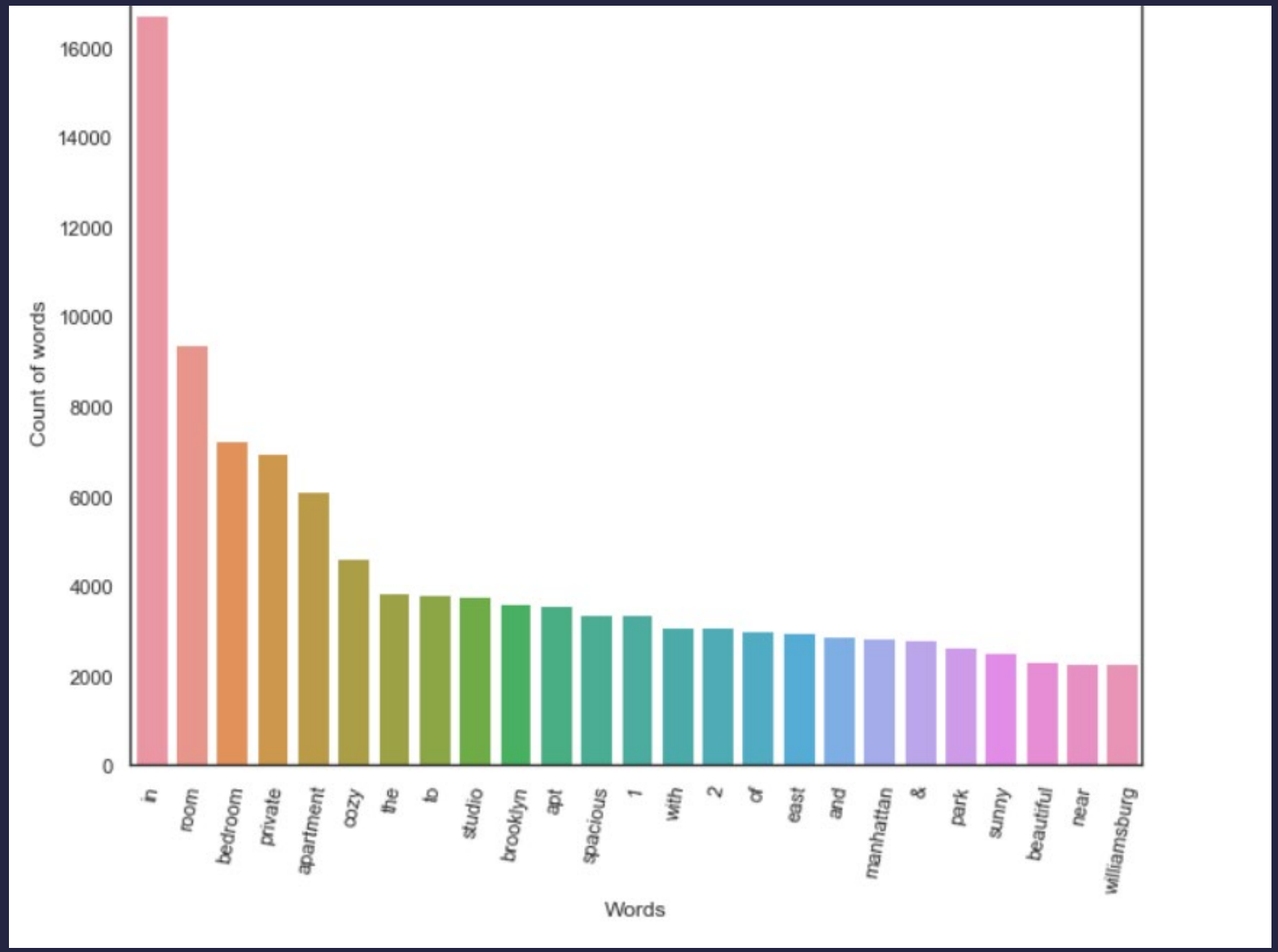


→ Now the scatter plot is easier to interpret than before

→ We can observe price range of 200-300, at the (-74,40.75) geographic location

NAMES IN LISTINGS

- → The bar chart shows that hosts are simply describing their listing in a short form with very specific terms for easier search by a potential traveler.
- → This shows that there are no catchphrases or 'popular/trending' terms that are used for names.



CONCLUSION

- → This dataset for the 2019 year appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented.
- → We have found hosts that take good advantage of the Airbnb platform and provide the most listings, and that our top host has 327 listings.
- → We proceeded with analyzing boroughs and neighborhood listing densities and what areas were more popular than another, and used latitude and longitude columns to create a geographical heatmap color-coded by the price of listings.
- → Further, we came back to the first column with name strings and had to do a bit more coding to parse each title and analyze existing trends on how listings are named as well as what was the count for the most used words by hosts