Distribution Patterns and Determinants of Airbnb In New York City

Daniel Addai-Marnu

Abstract

This paper analyses Airbnb data in New York City in the form of listings attribute and their spatial distribution. To date, the growing number of homes/apartments and rooms rented via Airbnb and other peer-to-peer accommodation platforms has drawn concern about the effect of such practices on the accommodation market. Using the dataset of Airbnb listings, we analyse their distribution concentration in the city, spatial distribution of listings features by borough, the effect of subway network and tourist attraction on distribution and develop regression models to find the determinants of their distribution. We can conclude that Manhattan and Brooklyn has the majority of listings with higher concentrations throughout Manhattan and part of Brooklyn adjacent the East river and centre of the city. We also found that subway connectivity influences the distribution of listings, whereas tourist places has no significant influence on distribution. Finally, the distribution of Airbnb listings is mostly determined by location (boroughs).

Keywords: Airbnb; New York City; Feature Analysis; Spatial Distribution; Regression Analysis.

1 Problem Statement

Airbnb is a room service that allows people to rent their unused rooms or property directly through computer-based interactions to potential customers. Founded in 2008, Airbnb has experienced exponential growth over the last few years and now has more than 7 million accommodations worldwide.[1] Today, more people and investors have caught up in the trend and realised the prospect and success in Airbnb. As interest in Airbnb has escalated, it is worth conducting analysis that will inform investments strategies and maximise benefit.

In this study, we use visual analytics techniques to analyse Airbnb data with the aim of addressing the following questions:

- Which factors explain the distribution of Airbnb across the five boroughs of New York City? We investigate the distribution and concentration of listings, listings features distribution, the effect of subway network and places of interest on Airbnb distribution.
- Which factors determine the distribution of Airbnb listings in New York City? We investigate the possibility of using regression models to find the determinants of Airbnb distribution.

The primary data is Airbnb data of New York City made up 48377 listings and 16 attributes. We also use spatial data to aid in the spatial visualization of distribution patterns. The attributes of entries will aid in grouping listings by boroughs or neighbourhoods and analyse the effect of the features like price concentration and distribution concentration. The spatial data will help visualise the effects of features geographically, likewise the effect of spatial points like subway stations and places of interest (tourist attractions) on distribution.

2 State of the Art

Adamiak et al. use Airbnb dataset scrapped from Airbnb website and obtained hotel data from TripAdvisor. The data is aggregated in 79 tourist sites and 8124 municipalities in Spain to measure their spatial autocorrelation, concentration and find the determinants of Airbnb rental distributions using regression models.

First, they present a comprehensive spatial description of Airbnb listings, compares their distribution and also measure the spatial autocorrelation and concentration of the density to housing and hotel supply.

Secondly, they employ regression analysis to find determinants of territorial distribution of Airbnb to answer the following hypotheses:

- The number of flats and homes, used as primary dwellings and non-primary dwellings, including vacant flats and homes and second homes determine the location of Airbnb rentals.
- The service of Airbnb is concentrated in areas close to the coast, which are (with the exception of major cities) the key destinations for recreational tourism in Spain.
- Areas with established tourist sector and appealing to tourists have high Airbnb offer concentration.
- Locations where high tourist growth or high seasonal variations have limited the ability of existing accommodation to meet demand, Airbnb serves as additional accommodation supply.
- Since Airbnb offers a familiar experience, it is especially attractive to international tourists.

They employed the Hoover index to numerically describe the degree of spatial concentration as well as developed eight regression models in order to identify the factors that affect the distribution of Airbnb listings.

Gutiérrez et al. article compares the analyses of Airbnb spatial patterns in Barcelona to hotels and sightseeing spots. Their analysis is based on Airbnb geolocated data obtained from Inside Airbnb website.
Their goal was to conduct a comparative analysis of Airbnb and hotel accommodation spatial pattern along with the

factors explaining their distribution. The main hypotheses for their study was:

- Airbnb grows tourist pressure on the city centre by occupying central location that is not covered by hotels.
- With regards to the city's key sightseeing spots, Airbnb is more favourably located than hotels.
- The factors that explain the pattern of Airbnb location vary from those that explain hotels.

The analysis was carried out using exploratory spatial data analysis (ESDA). Multiple ordinary least square regression (OLS) models were used to identify the underlying mechanism driving the spatial patterns for Airbnb and hotels accommodation.

The study conducted by both papers is very similar to what we are undertaking in that they analyse spatial distribution and investigate the factors that determine these distributions. Gutiérrez et al. looks at the patterns of distribution between Airbnb and hotel accommodation using ESDS while Adamiak et al. analyse the spatial patterns of listings distribution using the Hoover index to measure their concentration. In finding the determinants of the distribution patterns, both articles use regression analysis. Using the lessons learned from these papers in our approach, we will use ESDA to analyse the spatial distribution of Airbnb listings and use regression analysis to identify distribution determinants of Airbnb distribution.

3 Properties of the Data

This paper is primarily based on the analysis of the Airbnb geolocated data of New York City obtained from the website of Inside Airbnb. The data uses public information collected from the website of Airbnb.[4] The data was compiled in September 2019, and the file listings.csv was downloaded. The data also comes with a png file (dimension: 1326 x 1291 and resolution: 72 x 72) which was utilised in the visualisation. The data consist of 48377 listings and 16 attributes:

- id (numeric) – listings unique id
- name (categorical) – name of listing
- host_id (numeric) – host's unique id
- host_name (categorical) – name of host
- neighbourhood_group (categorical) – listing's borough
- neighbourhood (categorical) – listing's neighbourhood
- latitude (numeric) – listing's latitude
- longitude (numeric) – listing's longitude
- room_type (categorical) – type listing eg entire home, private room or hotel room
- price (numeric) – price per night
- minimum_night (numeric) – the minimum stay period
- number_of_reviews (numeric) – number of reviews of listing
- last_review (datetime) – last review date
- reviews_per_month (numeric) – number of reviews per month
- calculated_host_listings_count (numeric) – number of listings belonging to the host
- availability_365 (numeric) – the listing's availability in the next 365 days

To aggregate Airbnb data over the boroughs and analyse distributions, we download spatial data made available on NYC OpenData website aimed at engaging New Yorkers in the information generated and used by the city government.[5] These geographic data were areas of interest points, subway station points and borough boundaries; their geojson files were downloaded.

Attribute	Borough				
	Manhattan	Brooklyn	Queens	Bronx	Staten Island
Number of listings	21183	19856	5853	1126	359
Room Type					
Entire Home/Apt (24898)	12828	9364	2135	392	179
Private Room (21852)	7559	9985	3468	671	169
Shared Room (1192)	467	441	210	63	11
Hotel Room (435)	329	66	40	0	0
Price Min($0), Avg($105), Max($10,000)	Min($): 0 Avg($): 150 Max($): 10,000	0 90 10,000	10 75 10,000	10 65 1,000	13 78 5,000
Minimum Nights Min(1), Avg(2), Max(1250)	Min: 1 Avg: 3 Max: 1250	1 2 500	1 2 500	1 2 365	1 2 90
Number of Reviews Min(0), Avg(5), Max(654)	Min: 0 Avg: 4 Max: 618	0 6 469	0 8 654	0 9 331	0 14 360
Reviews Per Month Min(0.0), Avg(0.38), Max(67.6)	Min: 0.0 Avg: 0.27 Max: 67.6	0.0 0.4 19.25	0.0 0.85 20.42	0.0 1.0 10.75	0.0 1.09 9.29
Availability 365 Min(0), Avg(47), Max(365)	Min: 0 Avg: 38 Max: 365	0 27 365	0 110 365	0 142 365	0 244 365
Calculated Host Listings Count Min(1), Avg(1), Max(387)	Min: 1 Avg: 1 Max: 387	1 1 240	1 1 114	1 1 38	1 1 7

Table 1: Basic Statistics of Airbnb Data

The areas of interest are geodata points representing areas of interest like Yankee stadium, JFK airport. These points were plotted on the spatial distribution to explore their effect on the distribution. Similarly, the subway station geodata points were plotted on the spatial distribution of listings to investigate their influence. The borough boundaries geodata was used in visualising the distribution and concentration of listings in boroughs. The "neighbourhood_group" (boroughs) in the Airbnb data was linked with "boroname" (boroughs) in the borough boundaries geodata to distribute the listings into the various boroughs.

In pre-processing the data, we found an issue with the reviews per month and the last review date in that when the number of reviews is zero, the reviews per month and the last review date have missing values. This was expected

because with no reviews the reviews per month and date of last review will not exist. We filled the missing values of reviews per month with zero and that of the last review date with the oldest date of the last review. In dealing with outliers, we did not want to delete entries as we wanted the analysis to reflect the actual behaviour of the total population. Therefore, we standardised the data to reduce the effect of outliers present. Most of the features were skewed; thus, they were log-transformed to improve their distribution to boost the accuracy of regression models. We used Kendall correlation the see the dependency of attributes on each other.

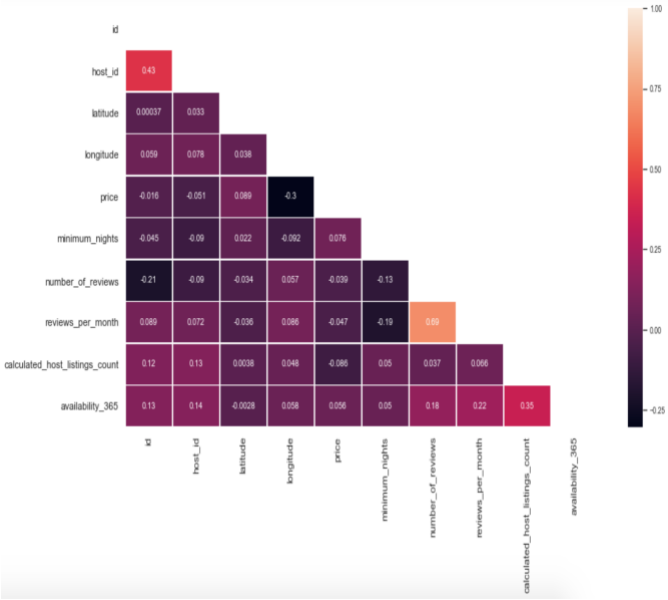

Fig 1: Kendall Correlation Between Features

4 ANALYSIS

4.1 Approach

Our aim in this analysis is to leverage the data within a visual analytics approach to deliver appropriate answers to our research questions. In light of that, we have identified these tasks to aid guide us.

The first step is data preprocessing and transformation (this is not necessarily an analytical task but essential in the process). There are two stages in this respect; the initial preprocessing for the exploratory spatial data analysis followed by further preprocessing and transformation for the regression analysis. The initial preprocessing involved deleting duplicate entries and filling in missing values to make the data tidy for the exploratory spatial data analysis. The second stage was to engineer a new to feature from the last review date in order to be represented in the regression analysis. The dates were transformed into days so that instead of having the last review date, it became the number of days after the last review. Additionally, all the categorical data were encoded; one-hot encoding was used because none was ordinal.

In general, the distributions of most the attributes were skewed, so we applied natural logarithm to improve the distribution. Outliers were not removed because we wanted the analysis to be reflective of the actual population of listings, but rather their effect was minimized by standardization of the features.

Descriptive statistics and density maps are used in the exploratory spatial analysis to determine the intensity and concentration of listings. Aggregated data at borough level and density mapping are the initial visual approach to distribution of listings. The descriptive statistics enable measurements to be taken to determine the dispersion or degree of concentration of listings and listing types. Based on the aggregated data, the location patterns were analysed and considered against geodata points to analyse their spatial interaction.

Multiple regression models (xgboost regressors) were used to identify the driving factors underlying the distribution of listings in different boroughs. The aim was to use the built models' explanatory variables to explain these factors. The models utilised normalised (standardised) features of the independent predictors. Pairwise scatter plots, correlation matrix and plots of feature importance helped determine essential features and their relationship with distribution. In conclusion, findings in visualisation steered the model iterative building process.

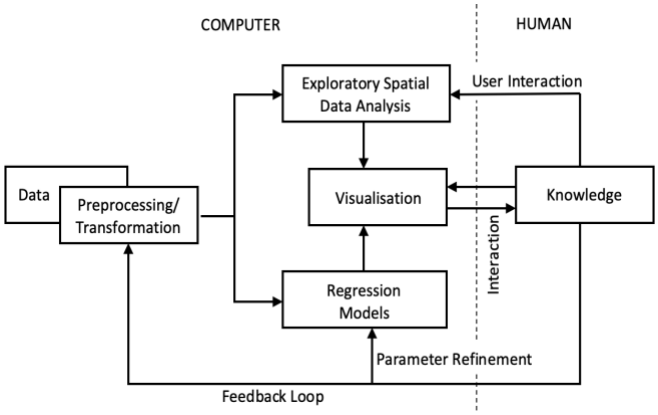

Fig 2: Visual Analytics Process

4.2 Process

In this section, we seek to analyse listing distribution and study the driving mechanism behind the distribution. We started by visualising the distribution of listings count by borough and further refine it by listing type count per borough. Manhattan has the most number of listings with most being entire home/apartment, followed by private room, shared room hotel room respectively. Brooklyn, Queens, Bronx and Staten Island follow respectively in listings count but with the order of listing type magnitude differing from Manhattan. They have most of their listings being private room, then entire home/apartment, shared

room and hotel room respectively. Bronx and Staten Island have no hotel room listings.

To find the exact spatial concentration of listings distribution, we visualised the heatmap of listings distribution to investigate the exact locations where listings are concentrated. There are high concentration of listings throughout Manhattan and centrally; Brooklyn, adjacent to the East River which separates Manhattan and Brooklyn.

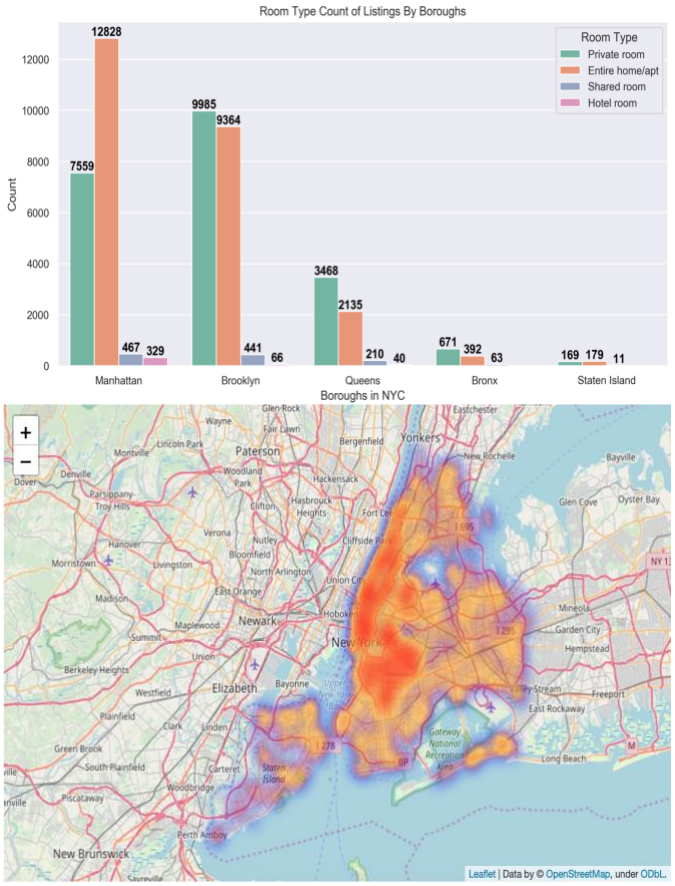

Fig 3: a) Listings Type Count By Borough b) Heatmap Of Listings Distribution

Next, we visualised listings distribution by price; prices were bin iteratively to arrive at four bins for visualisation. The bins were $0 - $500, $501 - $1000, $1001 - $2000 and above $2000 to aid in visualising the pattern in price distribution of listings. From the plot of $0 - $500, we notice that majority of the listings fall into this category with the highest prices being concentrated in Manhattan. Visualising the $501 - $1000 range shows a reduction in number of listings with listings converging towards Manhattan and Brooklyn. The $1001 - $2000 range show a further reduction in numbers, concentrated in areas with very high concentration of listings visualised previously from fig1. Finally, visualising prices above $2000 shows listings concentrated south-east Manhattan bordered by the East and Hudson Rivers with a few scattered in Brooklyn.

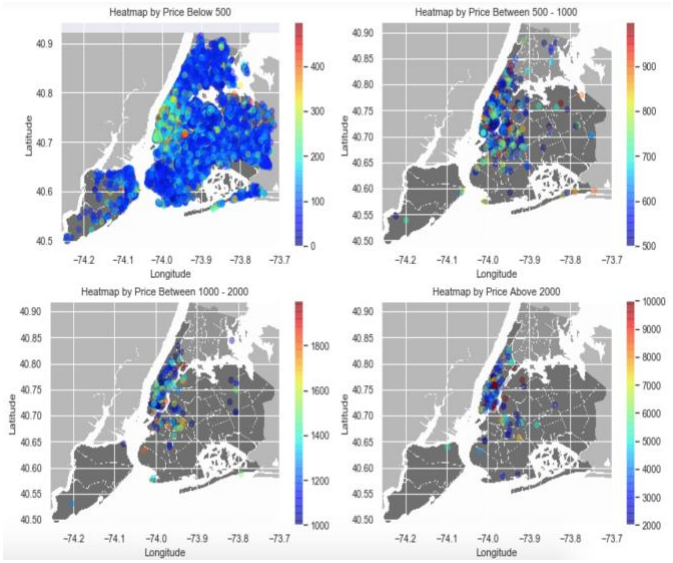

Fig 4: Price Distribution Pattern Heatmaps

We also visualised distribution of listing using the number of listings belonging to a host, that is, where hosts with many listings have their listings located. As done previously, we iteratively bin them to find their distribution pattern through exploratory spatial data analysis. We arrived at bins of hosts listings count of 1 – 10, 11 – 50, 51 – 100 and above 100. Most of the hosts have less than ten listings. Visualising hosts with listings between 11-50 show their listings converging to Manhattan and Brooklyn, that is, these hosts have most of their listings in these two boroughs. As hosts listing count moves to between 51-100, most of them have all their listings in Manhattan except for one with some listings in Brooklyn. Host with listing count above 100 have a similar distribution, with the majority in Manhattan and one host with some listings in Brooklyn.

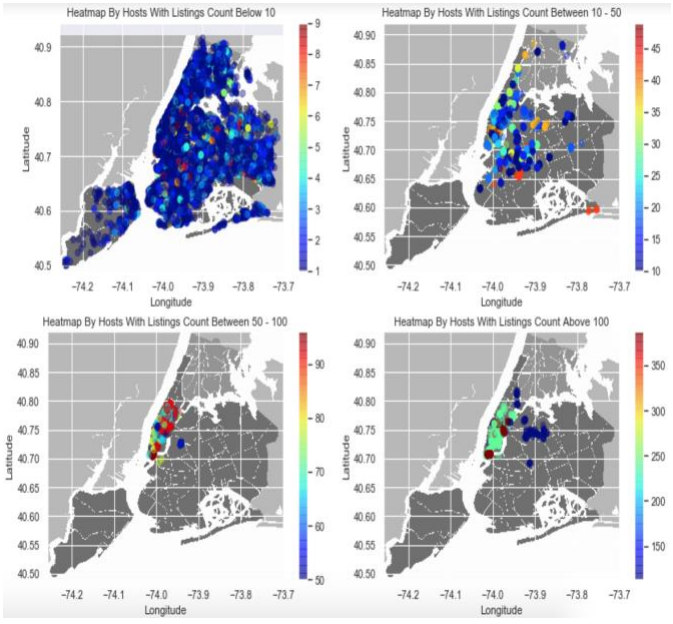

Fig 5: Calculated Host Listings Count Distribution

We spatially also explored the minimum night or minimum stay of listings effect on distribution. Similarly, to elaborate on the distribution with iteratively explored different binning and arrived at four bins; below 10 nights, 11 – 100 nights, 101 – 200 nights and above 200 nights. Most of the listings have minimum nights below 10 nights, and they spread throughout all the boroughs. As the minimum nights increase to between 11 – 100 nights the number of listings reduce and their distribution starts converging to Manhattan and Brooklyn. As we progress to listings with minimum nights between 101 – 200, the number of listings further goes down, and the distribution converges to Manhattan and Brooklyn. With listings with minimum nights above 200, the number of listing reduces further, and listings distribution also converges further to areas of high listing concentration in Manhattan and Brooklyn.

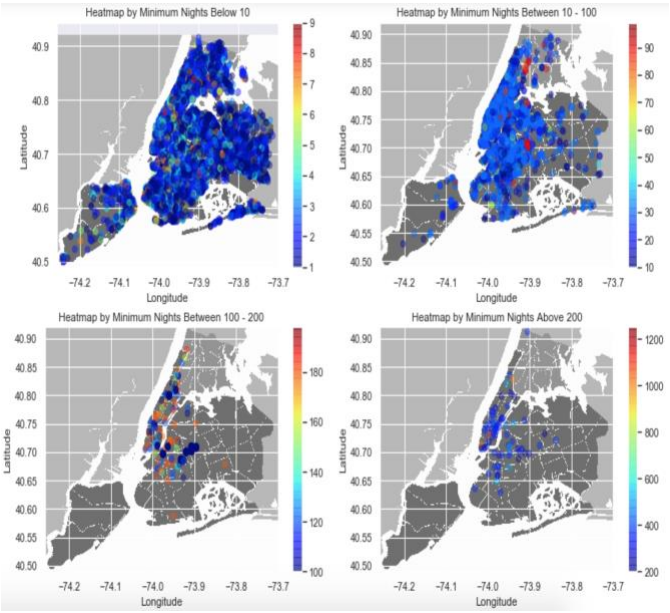

Fig 6: Minimum Night Distribution of Listings

Next on the exploratory spatial data analysis was to visualise the effect of subway connectivity using subway stations and places or points of interest on listings distribution. Visualising the subway station, we plotted the concentration of listing distribution of boroughs and then plotted the subway station point on the spatial distribution of the boroughs to investigate the relationship. From fig 7, We can see that Manhattan and Brooklyn have a high concentration of listings. From the subway station, we can see that these two boroughs have high concentration of subway stations as well, thus, have good subway network. This shows that boroughs with good subway network tend to have a high number of Airbnb listings.

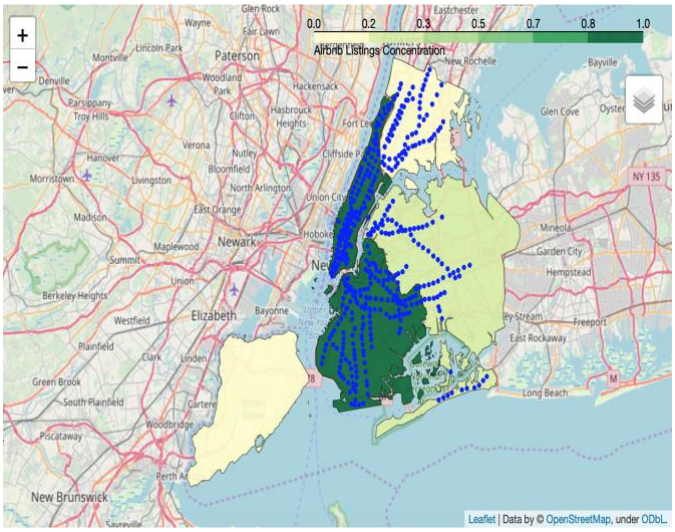

Fig 7: Listings Distribution Concentration With Subway Stations (Blue Points)

The last of the spatial exploration was to visualise the effect of places of interest (tourist attractions) on listings distribution. Here as well, borough concentration of listings was plotted, and the points of interest were plotted over it to investigate their effect. From fig 8, it can be seen that places of interest are scattered throughout the city and in all boroughs and the location of places of interest, in general, does not have any bearing on the distribution of listings. Although there was no correlation between the distribution of listings and the places of interest, this is not conclusive. Further investigation of specific places of interest may help find how they affect the distribution of Airbnb listings.

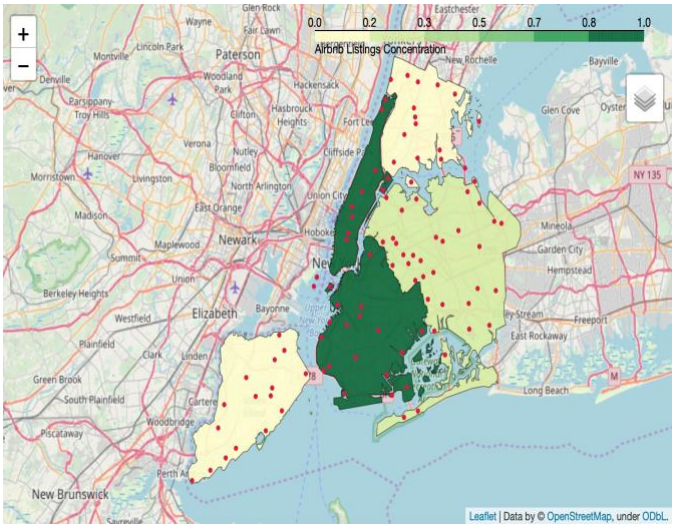

Fig 8: Listing Distribution Concentration With Places Of Interest (Red Points)

In finding the determinants of listings distribution, five xgboost regression models were used in predicting the location of a listing and finding feature importance. Three additional features were engineered before modelling; the estimated total annual income of the listing the minimum

night spend (which is the amount spent per minimum stay), and the number of days after the last review. We also encoded the categorical features so it can be utilised in the regression analysis. Moreover, we investigated the distribution of the features and realised most of them were skewed; therefore, to reduce the skewness and improve our models, we applied natural logarithm to transform them.

Five regression models were built each to investigate the determining factors (determinants) of predicting the location of a listing to be in Manhattan, Brooklyn, Queens, Bronx and Staten Island respectively. The models' hyperparameters were tuned using grid search to find the optimum accuracy. The R-squared of the fives regression models, a measure of how close the data is to the fitted regression line, were 0.9783, 0.9766, 0.9295, 0.9733 and 0.9912. All the models performed well. The determinants of distribution was arrived at by finding the feature importance of the regression models. The prominent features from the regression are latitude and longitude, which indicate that location in the main factor that influence the distribution of listings. Other small influences were noticed but as compared to that of location makes their effect less relevant. Fig 9 shows a plot of the combined features of importance of all the models.

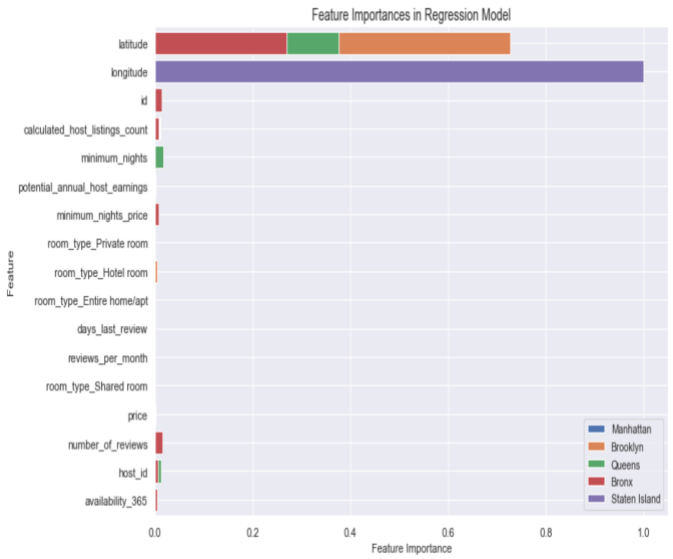

Fig 9: Combined Features Importance Of Regression Models

4.3 Results

From the exploratory analysis, the Airbnb listings followed a general pattern distribution. Listings concentration is predominant in Manhattan and Brooklyn (centrally), near the East River, which divides Manhattan and Brooklyn. Majority of the listings are entire home/apartment and private room and these dominate in all boroughs. Bronx and Staten Island have no hotel room listings. Price distribution of listings shows a decrease in the number of listings as price increase and listings distribution tend to converge at Manhattan and Brooklyn when listings are highly concentrated. The number of listings a host have also followed a similar pattern and the hosts with many listings have most of their listings located in areas with very high concentration of listings.

Distribution listings regarding minimum nights also converges towards Manhattan and Brooklyn as listings minimum nights increases. That is, listings with higher minimum stay period tend to be in these locations. From the analysis on subway connectivity, we noticed that areas with a high number of listings (Manhattan and Brooklyn) have good subway network evident in the higher number of subway stations located in these boroughs. However, we found no correlation between listing distribution and place of interest in general.

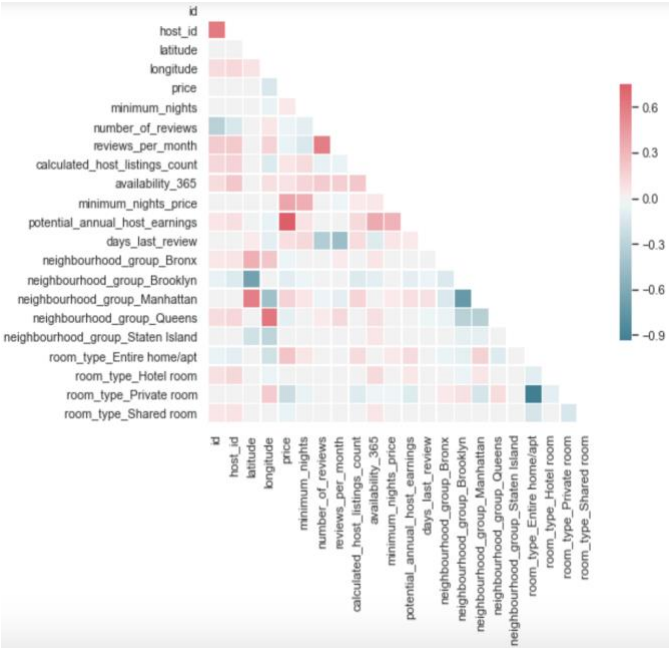

Fig 10: Correlation Matrix To Show Inter-dependency Of Listings Features

5 Critical Reflection

The eruption of Airbnb in the accommodation industry, especially in major cities, has seen a tremendous increase and has drawn and now receiving attention. This study aims to bring to bear patterns in the distribution of Airbnb and the driving mechanism behind it.

The results of the analysis show a general pattern in the distribution of the Airbnb listings offered in New York City; in Manhattan and centrally in Brooklyn. Exploratory spatial data analysis enabled us to visualise the distribution of Airbnb more regular and in a straightforward manner, putting listings in samples or bins and investigating distribution trends. These bins tend to reduce the effect of outliers as an outlier cannot be in all bin samples. Additionally, with the ability of manually iteratively visualise through exploratory spatial data analysis, we

intuitively used visualisation to tinker and investigate patterns we sought to see.

The regression analysis on determinants of Airbnb distribution confirms a strong relationship with location with the spatial analysis also reaffirming it with strong distributions in areas in the financial district (Manhattan) and central locations (Brooklyn) with proximity to the East River.

The analysis produced no correlation between places of interest and the distribution of Airbnb listings, but this is not conclusive. Possible analysis of individual tourist attractions may aid find how each influences the distribution of listings. However, from the analysis, subway accessibility affects the distribution of Airbnb listings, with large numbers of Airbnb listings in boroughs (Manhattan and Brooklyn) with good subway network.

Our approach has some limitations; first, regression analyses cannot discover causal relationships. Even though we can justifiably argue the driving factors of Airbnb distribution can be explained, a causal relationship cannot be established. In future work, investigation with other predictive models produce interesting relationships.

Secondly, even though our exploratory spatial analysis produced patterns in distribution. If we aim at capturing a true variety of distribution, socio-economic conditions, demographics, neighbourhood characteristics and service provision of listings will shed light on distribution patterns.

Our findings have shown a striking consistency of distribution patterns across New York City. This consistency suggests that, to an extent, there exist patterns that govern the distribution of Airbnb in this city. This means our approach can be applied to a city that has not been previously investigated to identify patterns in Airbnb distributions.

Table of word counts

Problem statement	252
State of the art	484
Properties of the data	489
Analysis: Approach	374
Analysis: Process	1124
Analysis: Results	197
Critical reflection	388

REFERENCES

[1] Airbnb About-Us, accessed 08 December 2019, https://news.airbnb.com/about-us/

[2] Adamiak, C., Szyda, B., Dubownik, A. and García-Álvarez, D. (2019) 'Airbnb Offer in Spain—Spatial Analysis of the Pattern and Determinants of Its Distribution', *ISPRS International Journal of Geo-Information*. MDPI AG, 8(3), p. 155. doi: 10.3390/ijgi8030155.

[3] Gutiérrez, J., García-Palomares, J., Romanillos, G., Salas-Olmedo, M H. (2017). The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona. *Tourism Management*. 62. 278–291. 10.1016/j.tourman.2017.05.003.

[4] Inside Airbnb, accessed 29 November 2019, http://insideairbnb.com/get-the-data.html

[5] NYC OpenData, accessed 02 December 2019, https://opendata.cityofnewyork.us/data/