ASSIGNMENT 2

FINAL REPORT

THE UNIVERSITY OF SYDNEY

Visual Analytics

COMP5048

Natalie Astalosh	(312067194)
Ming Sheng Choo	(490072615)
Julio Correa	(490086807)
Ian Johnston	(490228511)
Alan Poon	(480317036)
Jonathan Ross	(490204124)

Contents

1 Introduction

Airbnb is a people-to-people platform that connects hosts with guests for predominantly short-term accommodation rentals. Since launching in 2008, it has grown to feature 7 million places in more than 100,000 cities across the globe and is a major feature of modern travel. Tourists have a more unique and personalised way of travelling the world through Airbnb as they selectively craft the experience they want. Each party rates the other on a five-star scale after each stay, to provide prospective travellers and hosts, and Airbnb information about their experience [1]. Owners on Airbnb vary from private citizens with spare rooms to large corporate entities who make an entire business out of managing multiple properties.

1.1 Dataset and Tasks

The dataset we have chosen for our visualization implementation is the New York City Airbnb Open Data, available on Kaggle [2]. This set contains nearly 50,000 Airbnb properties in the city and includes the listing and owner name, geographic coordinates, the accommodation type, review history and price. Person-to-person platforms have grown in scale and profitability, and hence working with this dataset is familiar, challenging and topical work that proved of interest to the group.

We initially were drawn to creating an interactive dashboard that would be a visualisation of the whole dataset: the visualisation would be a manifestation of the data itself, based on the geographic overlay on the New York City map. We quickly realised this included low analysis of the data and could be achieved using only one platform, Tableau

We identified the property owners as a group who would be very interested in the dataset and strove to develop tasks that would provide deep insight to their investment on the Airbnb platform. These developed to be:

1. Design and implement an interactive dashboard for owners and investors to understand the market and competition in New York City
2. Assist owners in identifying the right price point in their borough, with respect to their desired return on investment
3. Make a recommendation on the best words to use in the listing title to make the property as appealing as possible

1.2 Aims and Contribution

We aimed to implement the above tasks in a timely fashion in a way that would be meaningful to the plethora of owners on Airbnb. We wanted to provide a simple and wholistic view of the competitive landscape so that owners need not do tedious research, but rather use our platform to cut the time required for making decisions. We set out to design a tool that could be used to improve their decision making in regards to Airbnb investments, to help maximise both the customer experience and profit.

There are 37,456 owners listing 48,896 properties on Airbnb in New York City, and it is therefore not feasible to address the thousands of owners on an individual basis. Therefore, we need to partition our owners into segments (profiles) that we can communicate with more efficiently.

Once we understand our owner profiles, we would like to create a visualization that is designed with a specific owner profile in mind. The visualisation would provide owners with information regarding their local market and competitors. Such a visualisation would communicate who the most successful owners in a market are, where they are operating, and what types of rooms they are listing. The analysis will therefore focus on the ownership, price, and room type distributions for the various markets (Neighbourhood Groups) in NYC.

We believe we have partially met these aims. While our platform is not completely comprehensive or inclusive of every piece of information an owner may like to know, it is a great visualisation of the available dataset. We have provided several real, actionable insights presented in a visual format that is easy to understand yet complex enough to provide meaning. Using our tool would allow an owner to understand what is happening in their neighbourhood at Airbnb and cut their decision making time. A deeper evaluation is provided in section 4.

2 Design

2.1 Analysis

2.1.1 Word Cloud

To get a sense for how owners describe their listings and to see whether there are any interesting patterns, we looked at the individual words used in listing titles. We were hoping to see if there are specific words associated with higher or lower value properties, or with unique prevalence in particular boroughs. We were also interested in the grouping or categories of words used, to see if there was a common theme amongst properties of similar type. The goal of this analysis was to aid potential new owners with naming their listings and help them choose descriptive words that will help their property stand out from competitors and attract more higher paying visitors.

2.1.2 Prolific Network Graph

We established owner profiles to understand the types of people who own Airbnb properties.

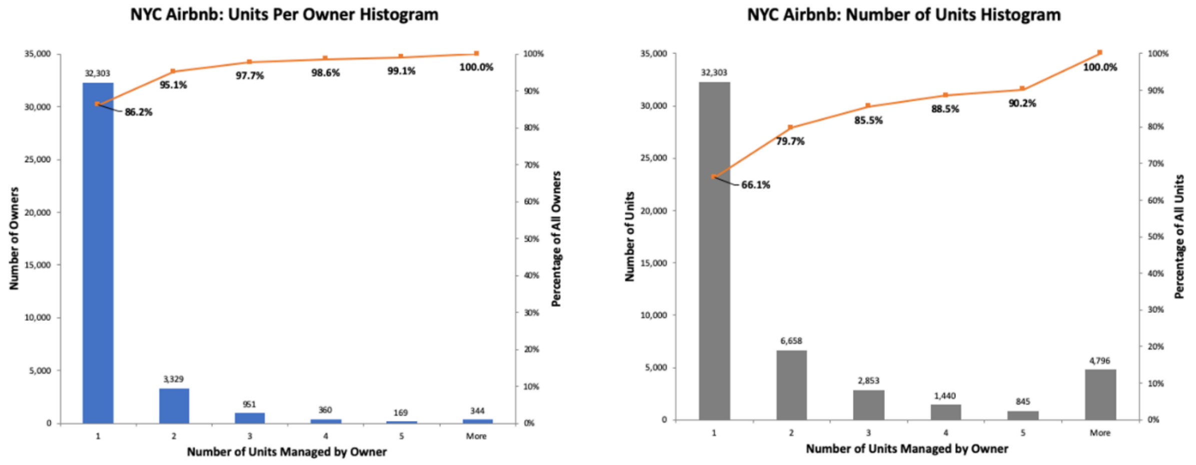


Figure 1: NYC Airbnb ownership overview: histogram of number of owners (left) and number of units (right), by units managed

The histogram in **Figure 1** (left) tells us that the number of units managed by owners is heavily skewed to the left. Over 86% of all Airbnb owners in NYC manage only a single unit. Only 0.9% of all owners list more than five properties. **Figure 1** (right) tells us that 86.2% of all owners manage only 66.1% of all properties. This indicates that there are a small number of owners managing a large number of properties: 10% of all properties in NYC are managed by only 344 (0.9%) of all owners.

We hence developed two owner profiles. The first is an owner with a small number of properties (less than or equal to five) and the second is an owner who manages five or more properties. We will now look at the product offering (room type) and price distribution for each of these owner profiles. We will first conduct an analysis of the market for owners with five or less properties.

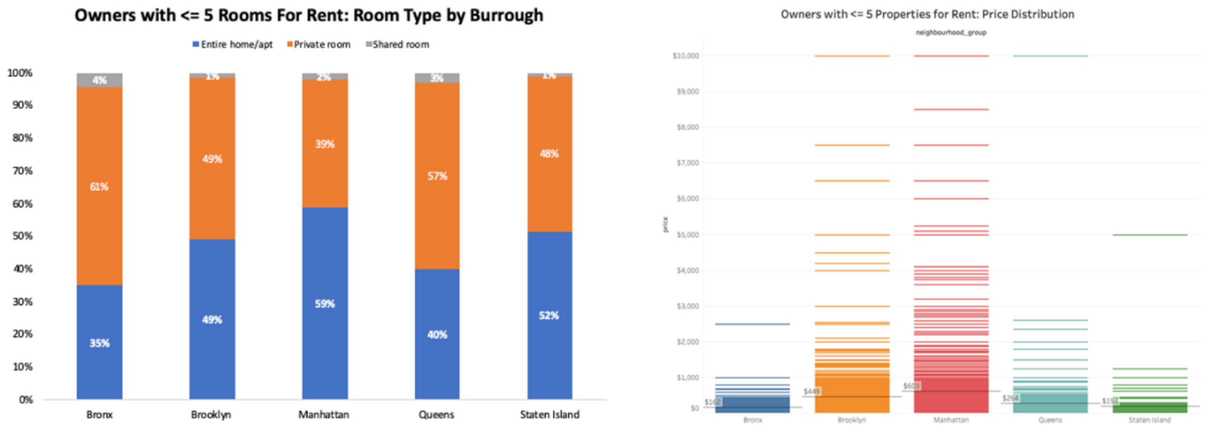


Figure 2: Profile of Owner with 5 or fewer rooms for rent. Room type (left) and price (right) distribution.

On average, half of the listing by owners with five or less properties are either entire homes or private rooms. The proportion of entire homes is largest and Manhattan (59%) and smallest in Bronx (35%), where the proportion of private rooms is the largest (61%). Shared rooms represent a small proportion (<5%) in all markets and represent the largest proportion in Bronx.

It is clear that Manhattan is the largest and most expensive market followed by Brooklyn, Queens, Bronx and Staten Island. In Manhattan and Brooklyn, there appears to be a market for exceptionally high-priced units. In Bronx, Queens, and Staten Island there is stronger clustering around the mean price and fewer higher priced units.

We will now conduct a similar analysis for an owner who manages more than five properties.

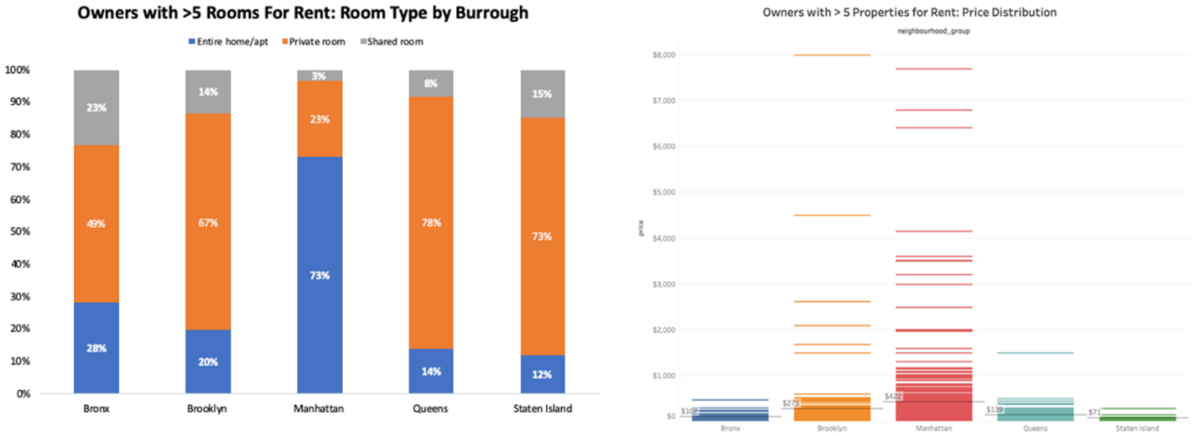
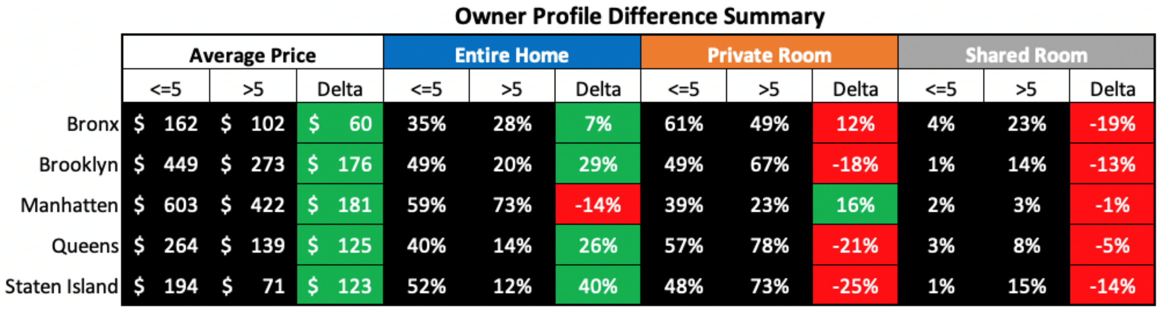


Figure 3: Profile of Owner with more than 5 rooms for rent. Room type (left) and price(right) distribution.

The differences between **Figure 2** and **Figure 3** illustrate the variations between our two owner profiles in terms of their product and pricing strategies. Owners who list more than five properties are offering more private and shared rooms. Not surprisingly, the prices for these units are also, on average, much lower than those offered by owners listing five or less units. We can also see that owners with five or less properties are offering a greater number of luxury units as indicated by the wider distribution of prices.

Table 1 below provides a summary of the pricing and product differences between our two owner profiles.



Owner Profile Difference Summary

	Average Price			Entire Home			Private Room			Shared Room		
	<=5	>5	Delta	<=5	>5	Delta	<=5	>5	Delta	<=5	>5	Delta
Bronx	$ 162	$ 102	$ 60	35%	28%	7%	61%	49%	12%	4%	23%	-19%
Brooklyn	$ 449	$ 273	$ 176	49%	20%	29%	49%	67%	-18%	1%	14%	-13%
Manhatten	$ 603	$ 422	$ 181	59%	73%	-14%	39%	23%	16%	2%	3%	-1%
Queens	$ 264	$ 139	$ 125	40%	14%	26%	57%	78%	-21%	3%	8%	-5%
Staten Island	$ 194	$ 71	$ 123	52%	12%	40%	48%	73%	-25%	1%	15%	-14%

Table 1: Summary of pricing and product differences between owners with five or less properties and owners with more than five properties.

2.1.3 Profitability Dashboard

A well-designed dashboard can show not only the generic information but also hidden information in the dataset. We developed a model to estimate the expected monthly occupancy of each home, then calculated the expected profitability by comparing leasing on Airbnb and the average price of a mortgage credit debt in New York. We then developed a pricing and income model to identify the average revenue.

The occupancy rate is generated based on the number of reviews per month, the number of nights, the number of available days in a year and a constant factor. To find the constant, we used the hotel occupancy rate of North America (65%) [3] and of New York City in 2015 (70%) [4]. We therefore used a factor of 1.5 in our calculations **(Appendix B. Code)**. We then multiplied the number of reviews per month and minimum nights and use it as a proxy in order to estimate the number of nights an average stay for each listing.

To finalize, we set an upper bound limit of 70% to the monthly occupancy rate of Airbnb based on the information obtained from Statista as mentioned above. Thus, as in the first case described, we penalized the maximum values of Airbnb and put them below our reference (hotel industry).

To calculate the expected revenue, we found the cost of an average lease, by modifying the data available at Rent Jungle [5].

Type of room	Monthly rent (USD)
Entire home/apt	3,519.00
Shared room	824.60
Private room	1,178.00

Table 1: Average rent by room type

With all the calculations above, we then compared the expected monthly income to rank the boroughs by profitability. The most profitable listings are in Manhattan and Brooklyn, the two more affluent and tourist-frequented areas of New York.

2.2 Visualisation

2.2.1 Word cloud

For the visualisation we wanted to produce a word map from the listing titles. Separate word maps would be produced for each borough and each accommodation type (entire apartment, private room, shared room). The value, or importance of each word displayed would be our primary consideration, but we also wanted to convey the prevalence of the words as well. Finally, we wanted the borough and accommodation type to be quickly and clearly identifiable. Thusly, we decided associated value would determine the size of each word, whilst using colour intensity to show how frequently a word was used. To differentiate between accommodation, we would use a distinctive colour palette for each type, and to convey location each word map would be projected in the shape of the relevant borough.

2.2.2 Prolific Network Graph

It is clear from the analysis that unique owner groups exist. Accordingly, our visualisation must be designed with a specific owner profile in mind. If we attempt to communicate with all owners using only one visual, we run the risk of providing irrelevant information that is not useful to any of the owners.

Our visualisation will provide information about the market conditions for a specific owner profile. Specifically, it will communicate who the most successful owners are, what types of units they are offering, the pricing strategy for these units, and which markets these units are provided. The visualisation should communicate the profile of a successful owner within a specific market segment. Accordingly, the visualization should be focused on the owner and make clear the relationship with their properties. Relationships between owners are not necessarily important but we do need the ability to compare profiles of different owners. Finally, because we are working with thousands of listings and owners, our visualization must have the ability to use as input large numbers of data points.

We expect that a network graph will meet the needs outlined above for our visualisation. Specifically, it will be capable of communicating information regarding centrality and connectedness for a large number of inputs.

2.2.3 Profitability Dashboard

As for the visualization, we wanted to have infographic dashboard which will bring everything together and allow for dynamic interaction with user. With the expected revenue calculated as outlined above, we combined summary bar graphs with the NYC map to highlight the most profitable neighbourhoods.

3 Implementation

3.1 Word Cloud

The first step in creating our visualisation was to extract the Airbnb listing titles, borough locations, room types, and prices from the dataset. To achieve this, we wrote a Python script using the Pandas module that produced a new csv file with these details for every listing. Once we had our new csv file, we wrote a second Python script, again using Pandas, to separate listings by borough and accommodation type, then take each listing title and split it into individual words. Each word was saved with its borough, accommodation type, a count of the number of times it occurred, and the mean average of the prices of all listings it had appeared in. This list of individual words was published in a separate csv file for each borough and accommodation type, giving us fifteen files. As a threshold for relevance, a word had to appear in at least five listings for it to make it into the final list.

As we were unable to produce a word cloud in a specific shape in Tableau, we instead decided to use the WordArt.com to project the wordlist onto the maps of the boroughs. For the borough maps we used GIMP to cut and crop each borough from the map of New York provided with the dataset. The WordArt.com word cloud art creator allows data to be imported from a csv or excel file. Words can also be associated with a number indicating relative size and a hex colour. Since the size of each word was to represent the price or value of the words, we used the average listing price. However, for the word colour we needed to find a way to convert the number of times the word appeared into a hex colour. This was achieved in two steps: the first was to write a Java script using D3 to produce a series of hex colour gradients from a specified palette, and the second step was to edit our Python script to apply the relevant intensity of colour to each word based on prevalence. With this done we imported our word lists with values and hex colours and borough maps into WordArt.com to produce our word maps. Full code can be found in **Appendix B.** Code.

3.2 Prolific Network Graph

We used Excel and Tableau for our data analysis and for preparing the data for input into Gephi and Tulip. Data preparation included segmenting the owners into our two owner profiles (greater or less than five properties) and the identification of nodes (owners and units) and edges (owners to property). Gephi was used import the nodes and edges from Excel. A graphml file was then exported to Tulip for creation of the graph. Tulip was preferred over Gephi because of its wider offering of network graph algorithms

Our initial intention was to provide a network graph for both of our owner profiles. However, due to hardware and software limitations we were forced to focus on owners with more than five properties. A network analysis of owners with five or less properties required the ability to handle over 100,000 combined nodes and edges. Through trial and error, we found that a maximum of 30,000 combined nodes and edges was within our software, hardware, and run-time constraints. Fortunately, the analysis of owners with more than 5 listings only required the capability to process algorithms for approximately 10,000 nodes and edges.

The Tulip Bubble Tree Pack layout was used to create the graph. We considered the most successful owners to be those with the most properties under management, and we used the top twenty owners for the graph. Each owner and their properties are nodes and are connected by edges that indicate which neighbourhood group the property belongs to. The property nodes are sized by price and coloured by property type.

3.3 Profitability Dashboard

We compared the expected income of Airbnb against its direct alternative cost, the monthly lease. We chose Tableau as the main tool to display our results. Starting from the overview map as shown in the figure below, it is drawn by using geographic information data that can be obtained from NYC OpenData [6], we then make use of Tableau to convert GeojSON file into table format so that we can inner join the table with the main Airbnb dataset to create the five borough outlines. The graphs are dynamically interactive when the cursor hovers over each borough (see Appendix for code).

The dashboard not only contains the profitability comparisons, but also the word clouds and network graph of owners. In this way, each visualisation is consolidated in one place for easy access.

4 Evaluation

4.1 Results

4.1.1 Word cloud

Shown here are two of the word maps we produced; the full set can be found in the Appendix. The visualisations shown here represent entire apartments in Queens, and private rooms in the Bronx. Word map of all boroughs is displayed in **Appendix C**. Images.

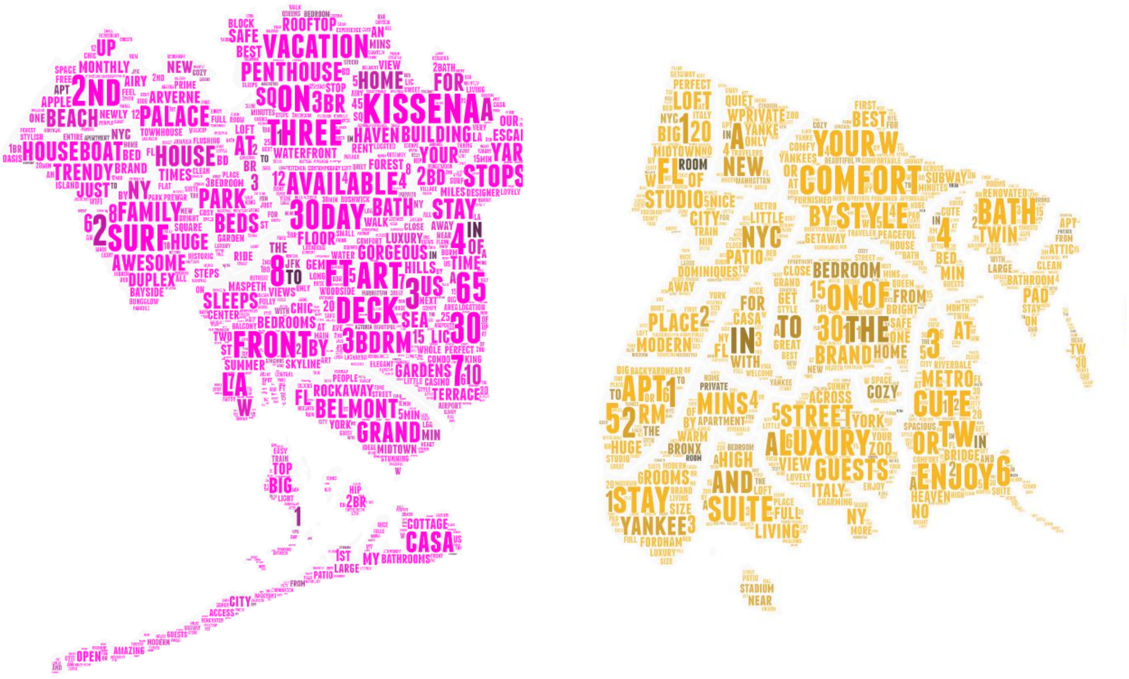


Figure 4: Word cloud of entire apartments in Queens(left) and private rooms in the Bronx(right), where word size indicates average price of listing containing word, and darker colour indicates word frequency.

4.1.2 Prolific Network Graph

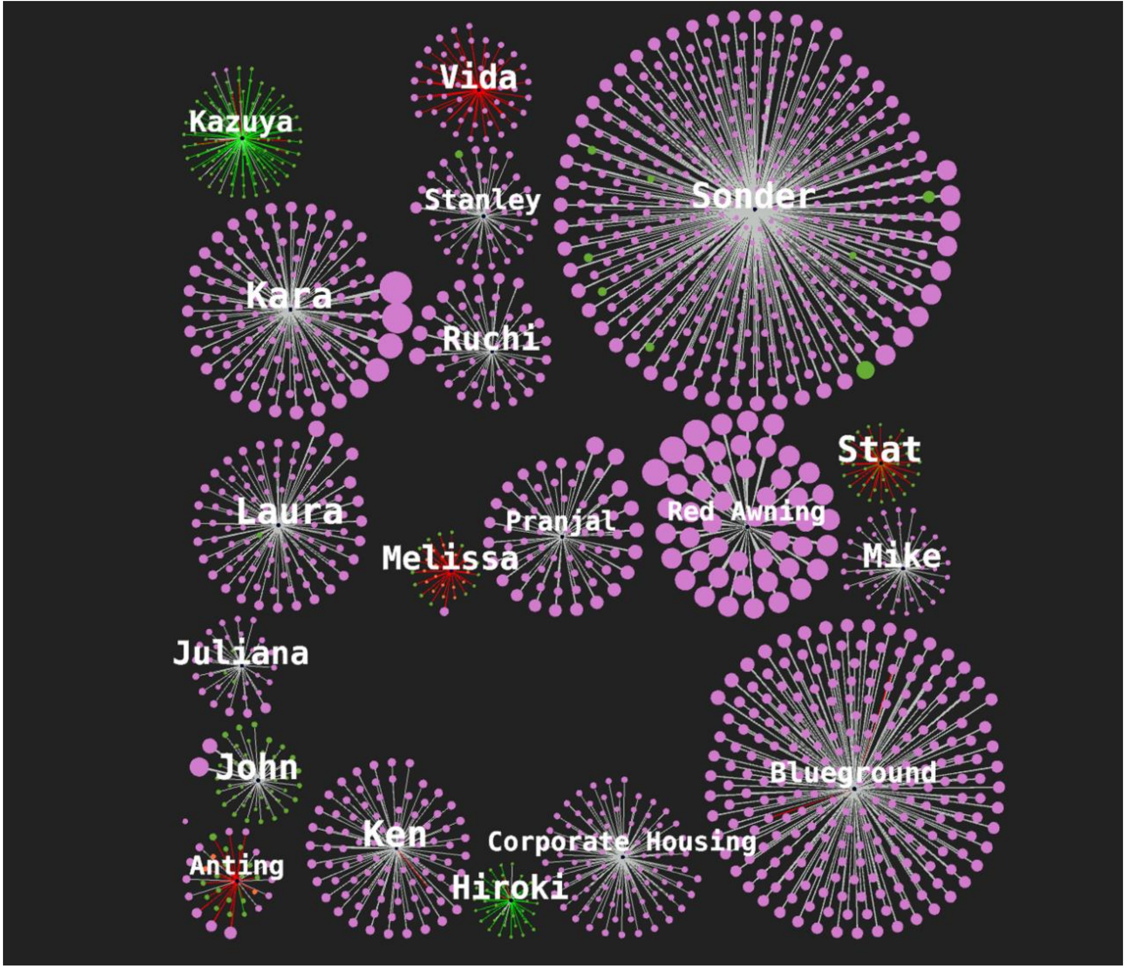


Figure 5: Network graph of most successful owners

4.1.3 Dashboard

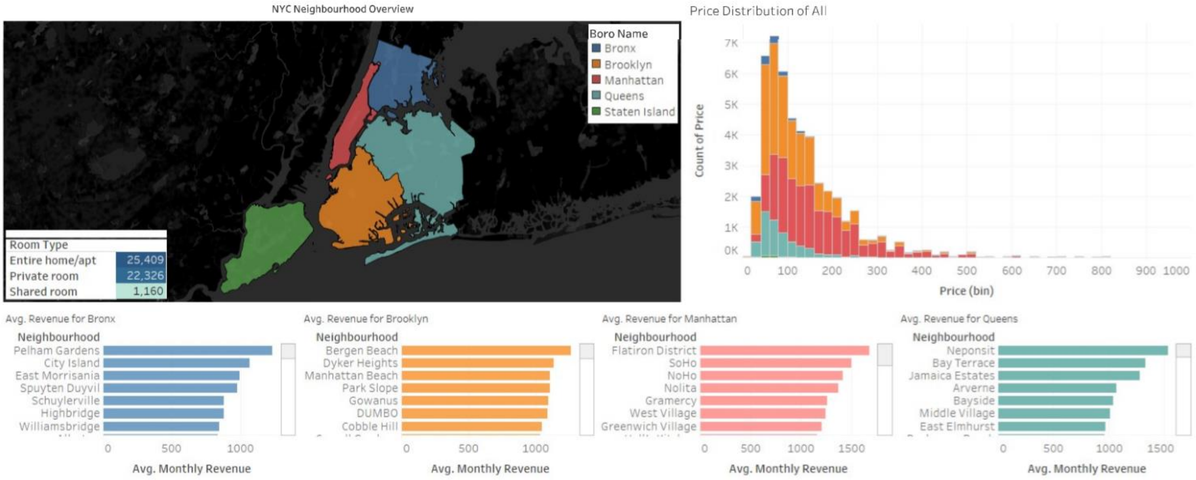


Figure 6: Dashboard 1 with general information of each borough and ranking based on average revenue

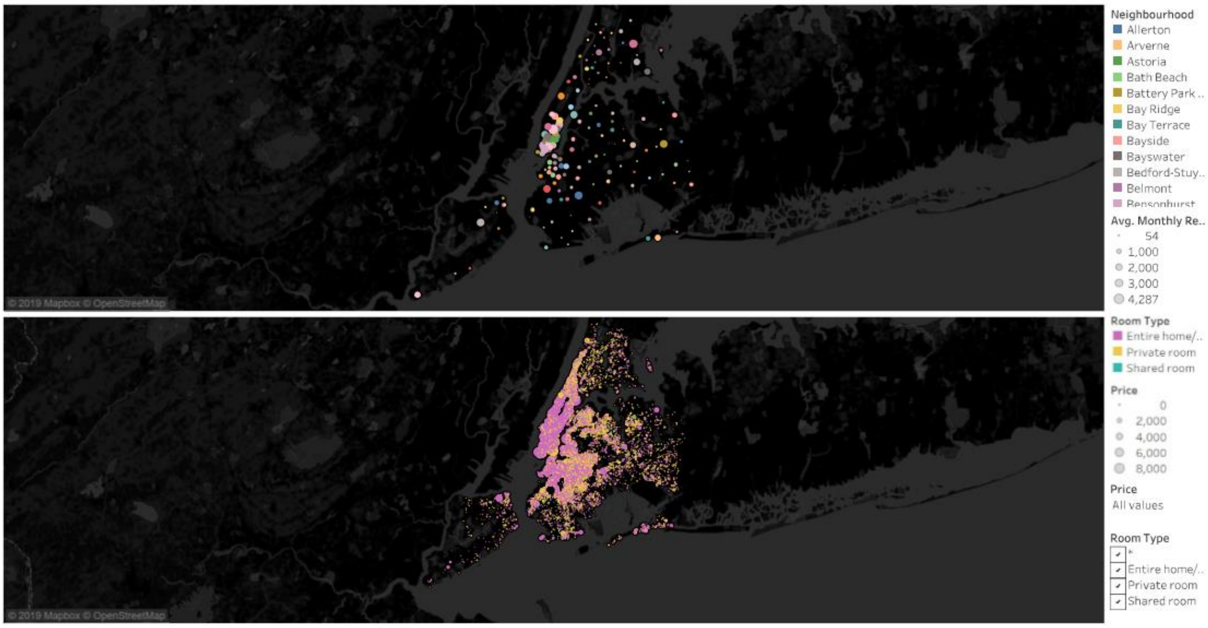


Figure 7: Dashboard 2of most profitable borough and current listing map with filtering

4.2 Discussion

4.2.1 Word Cloud

The two visualisations above show the success of our approach in creating these word maps. On initial impression the colour palettes are clearly distinguishable and match the colours chosen to represent the same accommodation type in our dashboard. The boroughs can also be easily referenced to the overall map of the city and are likely very recognisable to anyone familiar with New York. Both word size and colour gradient are clear to the viewer and clearly convey importance, although a legend or description is necessary to explain their specific relevance.

Regarding our initial aim, there are a lot of similarities between the words in each plot, especially those concerning facilities. Clearly property owners in New York find it preferable or necessary to specify the facilities and size of their properties in the listing titles; a tactic future property owners may wish to emulate. However, there are also words clearly referencing the character of the boroughs; "Yankee" appears in the Bronx, in reference to Yankee Baseball Stadium, and "Surf" and "Houseboat" both appear in Queens in reference to the southern waterfront. Clearly there is also value in highlighting the unique features of a property's borough in the listing title too.

Whilst our visualisations achieve their intended aim, they have the potential to be improved further. The nature of the WordArt.com word cloud art creator meant that some of our words were repeated at a smaller scale as the software tool attempted to completely fill the image space. Repeated words are of a sufficiently small size as to not impeded the clarity of the visualisation; however, a more rigorous revision would remove this problem.

4.2.2 Prolific Network Graph

We will evaluate the network graph based on both aesthetic and faithfulness metrics. A summary of the evaluation criteria and results are found below in **Table 2**.

The network graph succeeds in achieving many positive aesthetic qualities. There are no edge crossings or bends which is primarily a result of the graph displaying unconnected owners but is also achieved through the bubble layout that efficiently uses area. Bubble positions were established by the algorithm and determined the aspect ratio. By packing the property nodes tightly around the owner, we were able to minimize the edge length and achieve optimal angular resolution. Alternative layouts, such as a circular layout, were not as effective in minimizing edge length.

Because the nodes do not map to the geography, we do not consider the graph a success in achieving symmetry or orthogonality. Although there is symmetry for each owner node, we would have preferred to achieve a symmetry based on geographic location.

Criteria	***Evaluation***	***Criteria***	***Evaluation***
Crossings	✓	*Bends*	✓
Area	✓	*Angular Resolution*	✓
Aspect Ratio	✓	*Symmetry*	✘
Edge Length	✓	*Orthogonality*	✘

Table 2: Summary of network graph evaluation metrics

We also evaluate the network graph based on faithfulness metrics. A summary of this evaluation can be found in **Table 3**. We find that the graph layout, use of scaling, and use of colour to be contributing factors to making the graph very readable. In some cases, colour choices were optimized to improve readability at the expense of reduced consistency with the other visualisations in our overall dashboard.

In choosing to place focus on the individual owners, we were not able to remain faithful to the underlying geographic pattern of the data. Although this was a conscious choice that we felt supported our project aims, we would have preferred to find a way to both represent the geography and maintain a focus on the owner.

Measure	**Factors**	**Evaluation**
Faithfulness	Centrality of owner (+) Underlying geography not visible (-) Proportion of total network shown (-)	✘
Readability	Use of node scaling (+) Use of node and edge colour (+) Colour discrepancies with owner dashboard (-)	✓

Table 3: Summary of Network Graph Faithfulness Metrics

4.2.3 Dashboard

This evaluation was completed using the methodology outlined by Munzer [7].

1. **Performance:** Visualization is easy to generate, and computational time is low (polynomial complexity), since all the calculations to get the results are only linear combinations based on the code presented in section 2.1.3.
2. **Utility:** The presented graphs in the dashboard is user-friendly, user can understand the purpose of the visuals and quickly determine where the most profitable suburbs are located, and it helps the owner of a listing to make a better decision.
3. **User interface & interaction:** One shortcoming of the visualization is that the distribution of colours on the map does not have an a priori definition, and we use the default colour template that tableau offers. We did not meet these criteria, as we failed to harmonise the colours of the suburbs. Had we done so, we would have followed the rule raised by Munzer, Tamara, in Chapter 3 of his book, which is "Get It Right in Black and White."
4. **Information content:** It is important to note that the visualization contains a good density of information, characterized by the balance between categorical and numerical data. Thus, categories are shown on the map, which correspond to the neighbourhoods where the most profitable properties exist. On the other hand, quantitative/ordinal data is displayed at the bottom, in the graphs that rank these neighbourhoods according to their aggregate profitability.

5 Conclusion

We designed a user-friendly dashboard that provides valuable information to owners of Airbnb properties in New York City. We examined the words most commonly used in listings, and displayed the most expensive and frequent words by room type and borough in a word cloud optimised for readability. We conducted analysis of patterns of ownership, and found two distinct groups: those who owned many properties, and those who owned few. We developed a network graph of the owners who possessed many properties that shows where these are located and their room types, paying close attention to edge length and minimising crossings by disconnecting owner nodes. We developed a model to calculate the expected profitability of listings from the available data, and overlaid this information geographically onto a map of the city. Finally, we tied each visualisation together in one functional, cohesive dashboard that empowers owners with the information they need to make key decisions on their investment.

6 References

The Introduction uses text submitted in the *Initial Report* by these authors on 26 September 2019.

[1] Airbnb, (2019) *About Us.* Airbnb. Viewed 23 September 2019 <https://press.airbnb.com/about-us/>

[2] Dgomonov, (2019). *New York City Airbnb Open Data.* Kaggle. Viewed 23 September 2019 <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>

[3] S. Lock, (2019). Statista. Viewed 22 October 2019 <https://www.statista.com/statistics/587908/occupancy-rate-manhattan-by-quarter/>

[4] Statista research Department, (2016). Statista. Viewed on 22 October 2019 <https://www.statista.com/statistics/483846/new-york-city-airbnb-monthly-occpancy-rates/>

[5] RentJungle, (2019). Rent trend data in New York. Viewed on 24 October 2019 <https://www.rentjungle.com/average-rent-in-new-york-rent-trends/>

[6] NYC OpenData, (2019). Borough Boundaries. NYC OpenData. Viewed 20 October 2019 <https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm>

[7] Munzner, T. Visualization Design and Analysis: Abstractions, Principles, and Methods. Department of Computer Science, University of British Columbia. P. 10-14.