A Visualization Tool for Airbnbs in New York City

Berend van Beek[*] 1901354 Eline Ensinck[†] 1333941 Laurens Niesten[‡] 1265326 Lucas Boonman[§] 1463411

Index Terms: Visualization—Airbnbs—New York City—Entrepreneurship;

1 Abstract

This report concerns the data visualisation tool created for the course JBI100-Visualisation at Eindhoven University of Technology [1]. The tool can be found at `https://laurensn98.github.io/visualization/` and uses the Airbnb dataset about New York City from `http://insideairbnb.com/new-york-city/`. Using this dataset a (future) Airbnb entrepreneur can get interesting insights into the current Airbnb market. The tool integrates functions as simple as colour coding the different neighbourhood groups (to be able to see the distributions of Airbnbs across New York City), to more complex functions like showing the relation between different review aspects in a radar plot. These new functionalities as well as an intuitive design with a simple but efficient colour scheme results in a tool that is easy to use, understand and apply. By adding options to allow for selection and sorting of the data, the user is able to fine-tune their exploration on the map. In this way, the tool allows for flexible use and might be a good alternative to the current tools available, because of its different functionality and clear design and use. However, in the future still some improvements can be done, for instance by adding a SPLOM that shows the correlation between different attributes, or improving the functionality of the night and dark-mode.

2 Introduction

The visualization tool for this course will be designed for Airbnb entrepreneurs in New York City (NYC). They will either be interested in opening an Airbnb and thus want to explore the current market or they are already running one or multiple Airbnbs and want to improve their business. In either case the goal will be to analyze and thus explore a dataset, consisting of information about Airbnbs in NYC, to get an idea of the competition in the market and as such make appropriate choices when renting out an Airbnb. As we are dealing with multiple aspects that are of interest (e.g. price per night, location in the city, average review, etc.) a visualization tool will be suitable in order to cover all information in a comprehensible way. The user, who is a future or current Airbnb entrepreneur will be able to filter the data on aspects that he/she thinks are most relevant (e.g. a specific neighbourhood or room type), which is not easily achievable when presenting the data in a tabular format. As such, the tool will enable the user to identify patterns in the data. Ratings, for instance, are often important indicators for customers to determine which accommodation to book. In general these appear to be quite positive for Airbnbs [4], but overall ratings might therefore not give a clear idea. Our visualization tool will help to interpret these reviews in more detail by looking at the different aspects of a review.

~~2.1 Related Work~~

~~Current solutions generally do not provide the kind of insight that is needed for our user, an Airbnb entrepreneur, as they are limited to showing only price, distance from the city centre and a score for the location, for instance [3]. Also, the existing map available at the website of Airbnb itself [1] has some limits for our user as it is developed for people looking to stay at an Airbnb, rather than for people who would like to start a new Airbnb. The visualization tool available at the source website of the data provides more detailed statistics, but for instance does not distinguish between the different neighbourhood groups on the map (something that we will implement using colour to give a clear overview of the city). [5] Therefore, there seems to be a need for a more advanced visualization tool developed for in depth analyses suitable for Airbnb entrepreneurs. At the same time, quite a lot of different factors play a role in the large, changing business. As Airbnb reflects, people's interests and preferences have changed over time. Previously, the norm was to stay at hotels and occasionally a bed and breakfast, but this changed to an era where people are even able to rent out free bedrooms during the holiday season. As a result, aspects like the host and their behaviour, and type of room can make a big difference. [6] When trying to determine which investment would be suitable for an Airbnb location/property, it can be useful to know this information, because it is a very competititive field and it is important to stay ahead of competition to be successful [4]. Our visualization tool aims to provide this knowledge. Looking at existing visualisations of similar datasets, such as property data (of which Airbnbs might be seen as a subtype), gives inspiration on how to visualize the data, as they make use of principal component plots, for instance [9]. This is an example of a multivariate idiom and can give inspiration on how to visualize the data relevant for our user. The design decisions made for our tool will therefore be based on such examples, as well as the material presented in the course.~~

3 Data Analysis (What)

3.1 Domain Data Specification

3.1.1 Preprocessing

For the interim report, the dataset available at `https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata` was used. This dataset contains 102599 entries and 26 columns, where each cell represents a different Airbnb and each column gives information about an attribute of that Airbnb. However, to be able to perform more extensive analyses, the current tool is based on the dataset available at `http://insideairbnb.com/new-york-city/`, which includes some more detailed attributes, for instance about the reviews. However, using this dataset comes at the cost of losing some of the entries as this dataset consists of 'only' 41.533 entries. Just like the previous dataset, the data is presented in static tabular form, which means the entries will not change over time.

When loading the data it becomes clear that its naming is not consistent. To deal with this, attribute names are converted to make them all aligned, namely lowercase and using underscores.

[*]e-mail: b.c.b.v.beek@student.tue.nl
[†]e-mail: e.n.f.ensinck@student.tue.nl
[‡]e-mail: l.g.niesten@student.tue.nl
[§]e-mail: l.j.boonman@student.tue.nl

[1]To indicate the changes with respect to the interim report, we have coloured adjustments or new text in black, the old text has been switched to gray. No issues have been encountered in the group work.

3.1.2 Missing Values

Some of the listings have missing values. If the missing value is for an attribute that has only one possible value, such as country, this will be easily fixed (namely by relabeling it to United States). The same holds for missing neighbourhood groups; once a neighbourhood is known, the neighbourhood group can also be determined. However, if the missing value is for one of the other attributes, it is much harder to determine what it should be. These listings will be left in the data, but the missing values will not account for the relevant filters. For instance, if a price is undetermined for an Airbnb, the datapoint will not be shown if the price is necessary for the visualization, but it will remain useful for showing visualizations involving the other attributes. Note that the visualization tool thus makes use of a reduced dataset because of these missing values. Each plot will include statistics, so that it becomes clear whether, for instance, the plot is based on only 10 or 10.000 datapoints.

The most important attributes are longitude and latitude, since these values are used for giving the Airbnbs a position on the map. Without one of these two values, the location on the map cannot be determined, and as such the datapoint cannot be presented. Hence, Airbnbs without at least one of these values will need to be removed. Fortunately, after analysis, it became clear that this only applied to 8 cases in the old dataset, and to none of the Airbnbs in the new dataset.

3.1.3 Interpretation of Attributes

The semantics of all the attributes can be easily obtained through the attribute names. To clarify, the neighbourhood and neighbourhood group show the locations of the Airbnbs. A neighbourhood group is a collection of smaller neighbourhoods. As mentioned before, the attributes with the exact longitude and latitude coordinates describe the location of the Airbnb. The price is how much the room costs per night, the number of reviews show the total number of reviews a particular room has, and the review rate number is the average rating (on a scale from 1 to 5). More specifically, in the extended dataset the accuracy, cleanliness, checkin, communication, location, and value scores per review are available as well. Besides these, also some other attributes have been taken into account with respect to the interim report. Some focus on specific properties of the Airbnb, such as availability 365 (which shows the number of days the Airbnb is available each year), bathrooms, bedrooms, and beds. Others focus on the host, such as host acceptance rate (how many reservations are actually being accepted) and the host ID.

3.2 Data Abstraction

An overview of the data abstraction with all attributes that seem relevant for our user, the Airbnb entrepreneur, can be seen in Table 1.

Attribute	Categorical / Ordered	Ordering Type
Availability 365	Ordered	Quantitative: sequential
Bathrooms	Ordered	Quantitative: sequential
Bedrooms	Ordered	Quantitative: sequential
Beds	Ordered	Quantitative: sequential
Host Acceptance Rate	Ordered	Quantitative: sequential
Host ID	Categorical	Unordered
ID	Categorical	Unordered
Latitude	Ordered	Quantitative: Diverging
Longitude	Ordered	Quantitative: Diverging
Name	Categorical	Unordered
Neighbourhood	Categorical	Unordered
Neighbourhood group	Categorical	Unordered
Number of reviews	Ordered	Quantitative: sequential
Price	Ordered	Quantitative: sequential
Property type	Categorical	Unordered
Review: accuracy	Ordered	Quantitative: sequential
Review: checkin	Ordered	Quantitative: sequential
Review: cleanliness	Ordered	Quantitative: sequential
Review: communication	Ordered	Quantitative: sequential
Review: location	Ordered	Quantitative: sequential
Review: rating	Ordered	Quantitative: sequential
Reviews per month	Ordered	Quantitative: sequential
Room type	Categorical	Unordered

Table 1: Classification of Attributes

4 TASK ANALYSIS (WHY)

4.1 Domain Specific Tasks

There are several things the user may want to get out of the dataset. As a future Airbnb entrepreneur the user may first want to broadly **explore** the current offer of Airbnbs in NYC. He/she can do this by **investigating the total number** of Airbnbs in a specific area or neighbourhood to begin with. In this way you get an idea of the current offer, but it says nothing yet about the quality, type of accommodation, availability, etc. More **detailed analyses** will be necessary for this. Focusing on a region, one might for instance inspect the type of Airbnbs present; are they small/large, cheap/expensive, etc.

Based on the available budget of the user, it could be interesting to **see where the expensive Airbnbs are**, versus the somewhat more affordable but potentially also lower rated ones. If there are Airbnbs close to a region where someone wants to open an Airbnb and all of them have a very high rating, the current accommodations may take customers away from a newly opened Airbnb without any (positive) ratings yet. On the other hand, if there are only cheap Airbnbs closeby, it is wise to adjust your price accordingly. At the same time, opening a cheap Airbnb in a region with mostly expensive Airbnbs available, might attract more attention from future customers. Hence, only by inspecting the price, already quite some useful information can be obtained. However, it may also be important to **see the correlation between price and rating of the competitors**, as value for money is of course relevant for a customer to decide which Airbnb to book and what to aim for. Focusing more on the available ratings, it might be interesting to see not only the overall rating of a specific Airbnb, but also how these ratings are distributed over multiple review aspects (namely accuracy, cleanliness, checkin, communication, location, and value). In this way, the user acquires knowledge about which aspects are lacking in the current market, or what the minimal level of e.g. cleanliness is that is expected.

At the same time, a user that is already renting Airbnbs might also get profit from the visualization tool. Useful **insights can be obtained about what to offer and for which prices** to improve their business. As an example, if all other Airbnbs nearby have higher prices, then it seems reasonable to also raise this.

All these considerations together will give an idea of whether it is feasible to start an Airbnb within a certain region, and which type (e.g. cheap/expensive, small/large) would be good to consider opening , but also which aspects might improve a current Airbnb business. To summarize, the following tasks can be formulated:

- Analyze the type of Airbnbs available in NYC (number of beds, room type, etc.);
- Determine a suitable region for opening an Airbnb;
- Investigate the average price for an Airbnb in NYC, taking into account both rating and type of accommodation.
- Inspect the scores of different review aspects (accuracy, cleanliness, checkin, communication, location, and value);
- Analyze how to improve existing Airbnbs.

4.2 Task Abstraction

Based on this task analysis, the following task abstraction, presented in table 2, is derived. The first task which analyses the current Airbnbs available, tries to provide an overview of (and thus summarize) the general features in the data. The second task, tries to determine a suitable region, which is an example where the target is known (for instance density of Airbnbs), but the location is yet unkown (hence, search and locate). The third task analyzes the distribution of the price attribute, potentially filtered on multiple aspects. Going more into detail, the fourth task tries to discover the correlation between the scores for different review aspects. Finally, the task of analyzing possible improvement points, is about exploring dependencies in the data, such as higher ratings mean higher prices.

Task	Action	Target
Analyze Airbnbs available	Summarize	Features
Determine suitable region	Search	Locate
Investigate price	Summarize	Distribution
Inspect reviews	Discover	Correlation
Analyze improvement	Explore	Dependency

Table 2: Action-target pairs

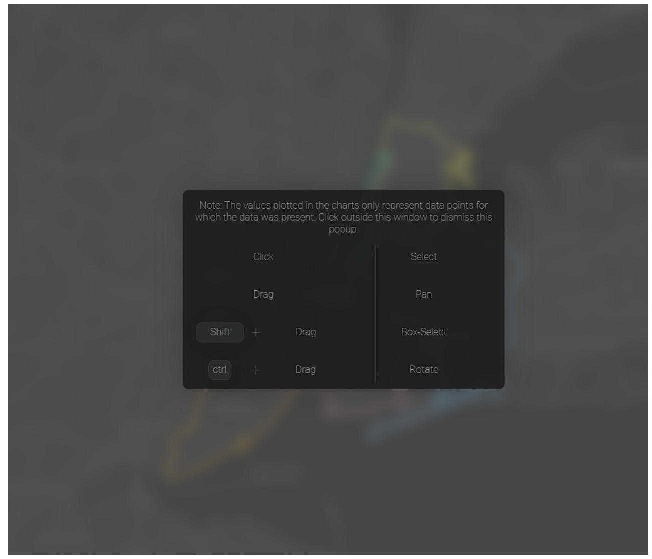

Figure 2: Initial pop-up with instructions

5 Current Solution

To provide the tools that entrepreneurs need, there are a number of options and tools that will be needed within the visualisation. Firstly the user would need to be able to compare different areas within NYC and the Airbnbs in them. Some additional information about the existing Airbnbs would of course also be needed to make the tool more informative and useful. In our solution we aim to provide information on (among others) prices, types of units, number of accommodations and number of bedrooms in the Airbnbs, as these have been shown to be important factors in customers' choice to rent an AirBnB. [2] In order to show all these aspects the first step was to visualise the spatial data. To do this we use an interactive map of NYC onto which we have mapped the locations of all the Airbnbs. An example sketch can be seen in Fig. 1. This sketch implemented in our visualization tool can be seen in Fig. 4.

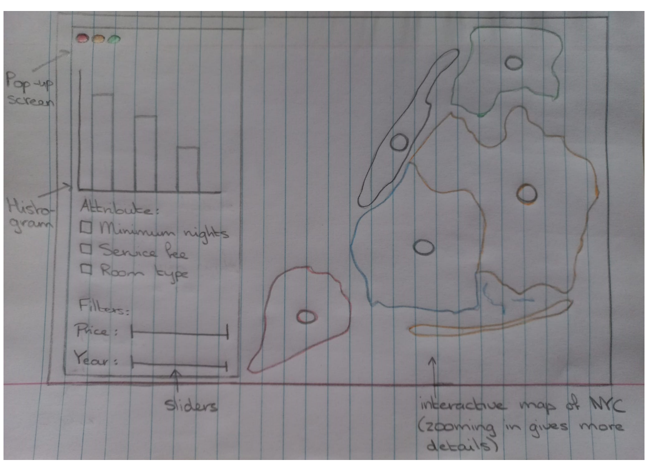

Figure 1: Sketch of the visualization tool

When opening the visualization via the web browser, a pop-up is shown that gives information about how to use the tool, as shown in Fig. 2. This pop-up makes use of the Gestalt principle of Figure Ground, as the foreground is separated from the background by making the latter blurred.

As soon as the actual map is being displayed, it becomes clear that the user has multiple ways of interacting with it. First of all, he/she can decide to switch the interaction mode to ”automatic”, ”Airbnbs”, or ”neighbourhood groups”, as shown in Fig. 3. When automatic is selected, individual Airbnbs will be shown by clicking on one of the neighbourhood groups. When zooming out the lay-out of the Airbnbs will change (switching back from individual Airbnbs to the neighbourhood groups), which is an example of **semantic zooming** as the visual encoding changes. However, the user can also decide to always view just the neighbourhood groups or Airbnbs by selecting one of these options.

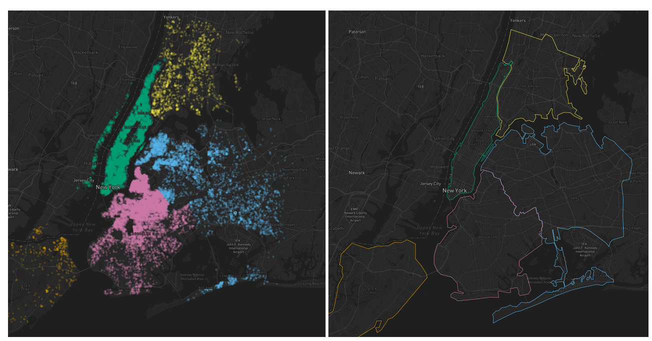

Figure 3: Interactive mode on ”Airbnb” mode (left) and ”Neighbourhood groups” mode (right)

Since the location data was considered the most important attribute for Airbnbs [7], using the interactive map shows use of the **effectiveness principle** as it is the most prominently displayed attribute. In addition to this, the neighbourhoods are shown as outlines on the map and can be selected to show data on the neighbourhood in its entirety. In the interim report we mentioned the design of both a dark and light mode. However, because of the complexity we decided to only focus on the implementation of the dark mode for the final report. An example of how the light mode would look like can be seen in Fig. 5.

More informative details about the Airbnbs will be shown in a pop-up menu, visible in Fig. 4 that opens when a neighbourhood, a specific Airbnb or a self-determined region on the map is selected.

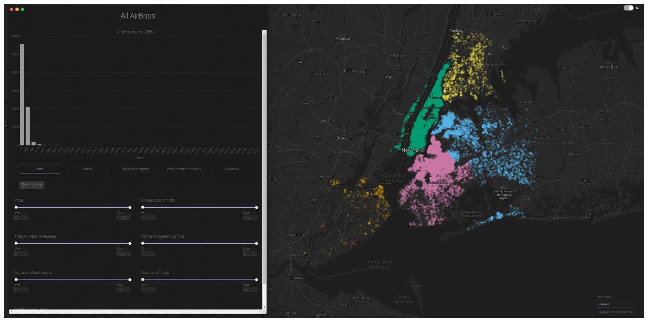

Figure 4: Entire map with pop-out menu in dark mode

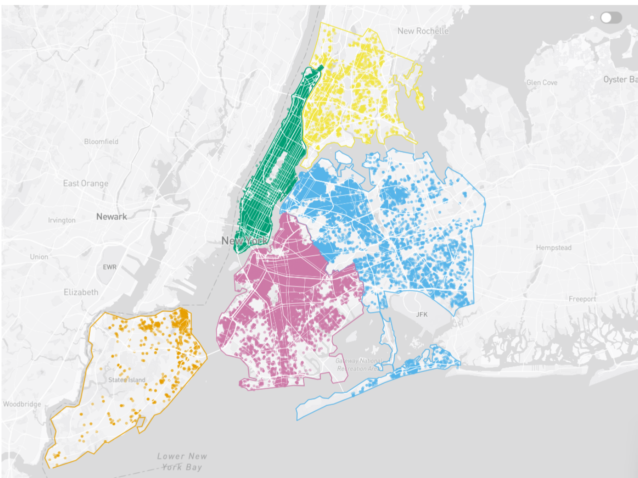

Figure 5: Entire map in light mode

Selecting a rectangular region on the map can be done by the combination of shift+drag and is a type of **data reduction**, which can be obtained through interaction with the tool. Use of this feature is shown in Fig. 6.

In the prototype implemented for the interim report, only a price distribution of the Airbnbs was visible. Now, for the final version, this window has been improved by adding extensive interaction possibilities of which a first glance can be seen in Fig. 4. The user is now able to choose the attribute of which he/she would like to see the distribution. The selection consists of price, rating, reviews per month, total number of reviews, and availability, giving a diverse range of options. Additionally, in this screen the user has the ability to **filter the data**, as it includes menus to select the desired price range, number of beds, rating score, etc. A detailed look at these filters is given in Fig. 7. Once a selection has been done using the filters, the map gets updated as well. Thus, both the histogram and the map are linked with each other, providing **direct feedback** (e.g. visually showing where those Airbnbs with prices in range 100-200 dollars are, or those Airbnbs with availability of more than 300 days a year). In this way, more detailed analyses are possible, fine-tuned exactly to the wishes of the user. So-called linking and brushing is also implemented by enabling the user to select a bar in the histogram and then viewing only these datapoints on the map. An example of this is shown in Fig. 8. If the user wants to remove the selection made, this can be done using the ”reset filters” button, which is also visible in Fig. 7.

As can be seen in Fig. 9, when hovering over a neighbourhood group, the mouse cursor changes into a pointer and the outlines of the neighbourhood group will change colour and turn white. This

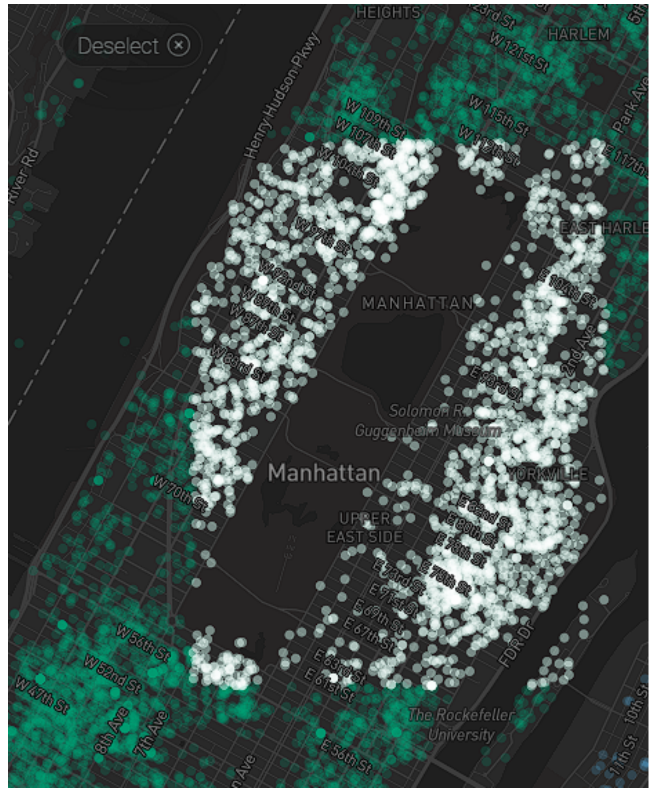

Figure 6: Box selection on the interactive map with the deselect option in the top left

is to show that the neighbourhood can be selected. This change in colour will signal this to the users pre-attentive vision. This theory is also applied to specific Airbnbs on the map, turning the dot red, increasing its size, and adding a white outline when it is selected, drawing the attention of the user. As we have seen during the course, only changing one attribute at a time, makes it easy for the user to notice it, as it then ”pop outs” of the visualization. In addition to this, **transparency** is added to the colouring of the dots to deal with the high density of Airbnbs on the map taking into account **scalability** and clearly visualising the density of datapoints on the map.

New with respect to the interim report is the implementation of a **radar plot** when selecting an individual Airbnb. This multivariate idiom shows the different review scores, and can visualize the underlying relations between them. Interesting to the user might be to see patterns in these star shapes. For example, all stars in region X that score high on cleanliness score low on communication, hence you should make sure to also score high on communication to stand out within this region. An example of a radar plot is shown in Fig. 11.

Fig.10 shows that the dots on the map representing the individual Airbnbs are coloured to be the same as the outline of the neighbourhood the Airbnb is in. This shows the next layer of data visualized in the map by showing the Airbnbs that are in the same neighbourhood in the same colour. To make sure the visualization tool is also suitable for (red-green) colourblind people, the following colour scheme has been chosen: orange, sky blue, bluish green, reddish purple and yellow [8]. These match the categorical data, as they are unordered, which is in line with the **expressiveness principle** that says that the visual encoding should match the data characteristics.

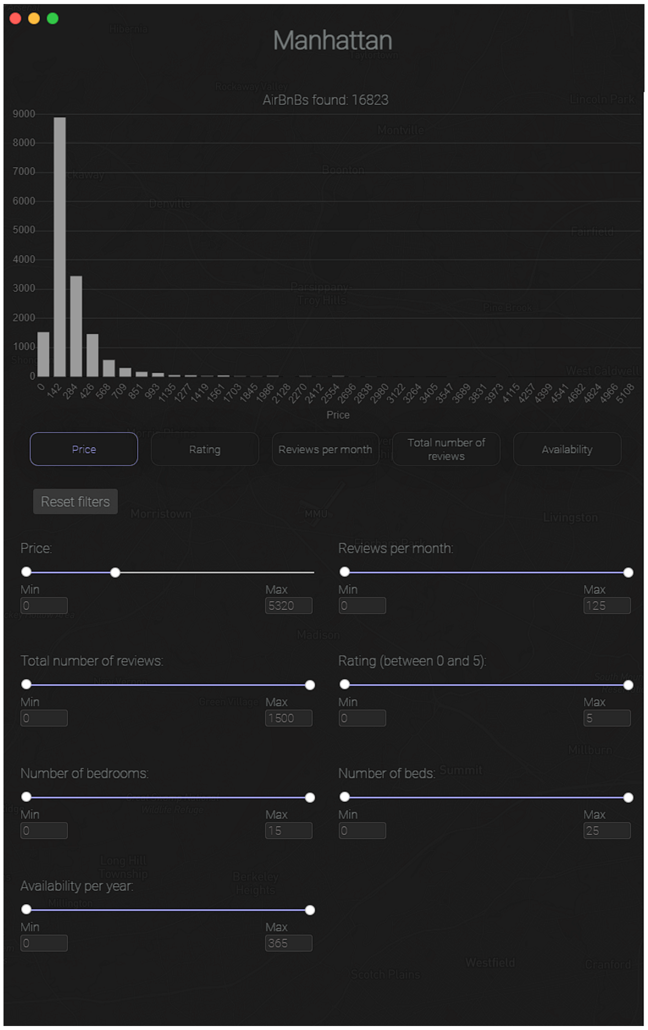

Figure 7: A closer look at the filter menu

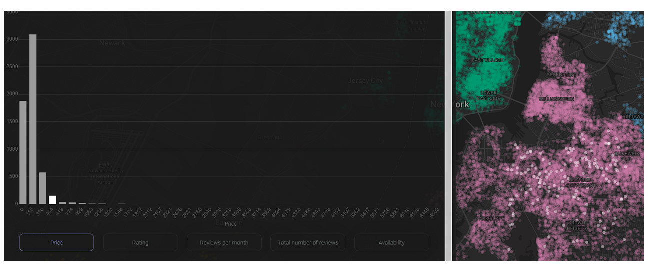

Figure 8: Viewing a set of Airbnbs by selecting a bar in the graph

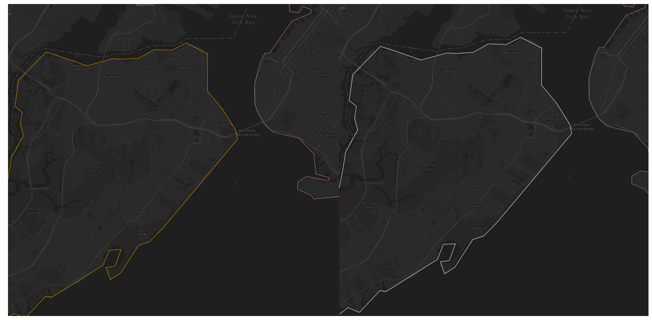

Figure 9: Staten Island selected vs. unselected

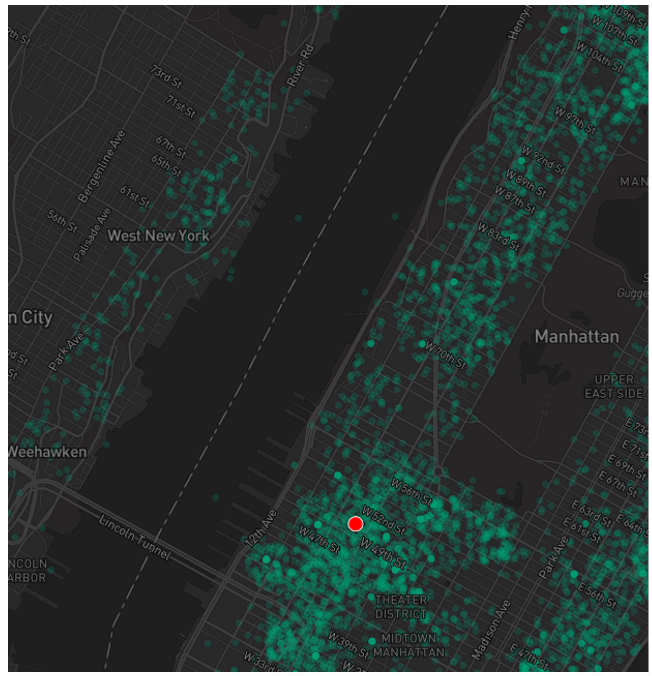

Figure 10: Single Airbnb selection on the interactive map

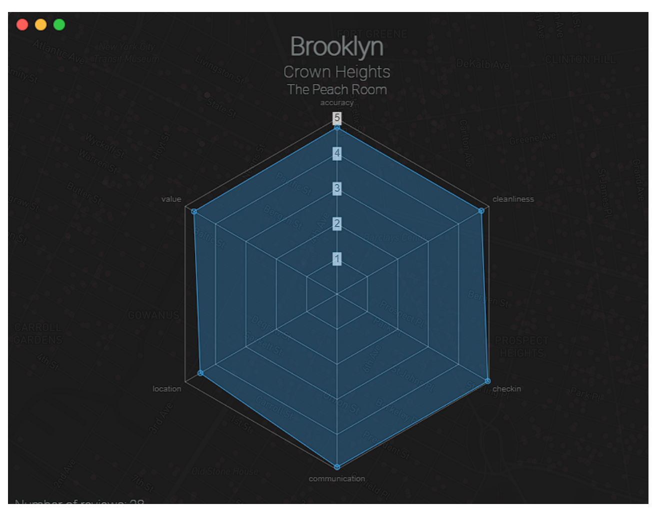

Figure 11: Radar plot showing data for a specific Airbnb

6 Implementation

Originally, we had decided on using dash to implement our visualization. However, we soon realized that this had some major limitations. For instance, event listeners on mouse-wheel scroll are not possible in dash as the coupled JavaScript can not access the DOM. This functionality was a detrimental part to some of the ideas we had. Therefore, unable to implement these ideas in dash, we made the timely decision to switch over to a combination of JavaScript, HTML, Css, and Python. Python is only used for preprocessing the data. We have also made use of the libraries Chartjs and Mapbox to implement our visualization. It can be viewed at:

https://laurensn98.github.io/visualization/

7 USE CASES

7.1 Use case 1: Starting a new Airbnb

The first user outlined at the start of this report was someone interested in starting an Airbnb within New York City. To this end, the tool can be used to both get a general overview of the city as well as allow for more in depth information about an area of interest. In this case, one could start with looking at the general overview of the map and using the setting in the bottom right to view all Airbnb locations currently available, as is visible in Fig. 4. From this view, it becomes clear that most of the Airbnb locations are within Manhattan or the north of Brooklyn and other areas are far less saturated with existing Airbnbs. As such one could conclude, for example, that it might be interesting to look into locations outside of Manhattan and Brooklyn since there is less competition in these areas. Hence, the user might want to take a look at the area of the Bronx bordering on Manhattan, shown in Fig. 12.

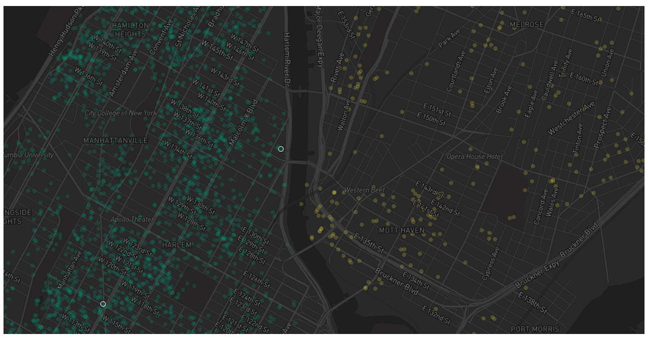

Figure 12: The area where Manhattan borders The Bronx

This area is close to the city centre in Manhattan and a similar distance from the airport. However when using the box-select tool to compare prices, the user might notice that while the average price of an Airbnb in Manhattan is around 250 dollars, this area in the Bronx averages at only around 100 dollars, giving an indication for the price. Finally once the location has been identified the tool can be used again to inspect different properties of nearby Airbnbs to see what they may be missing or what they excel at. Looking at surrounding Airbnbs it becomes noticeable that many of these locations get a lower rating on communication, as shown in Fig. 13. It may therefore be a consideration to dedicate some extra attention and/or resources to this aspect. The radar plots on individual Airbnbs allow for a quick and accurate insight into what properties surrounding units have such that the new unit can either conform to these or possibly improve on them to stand out from the rest.

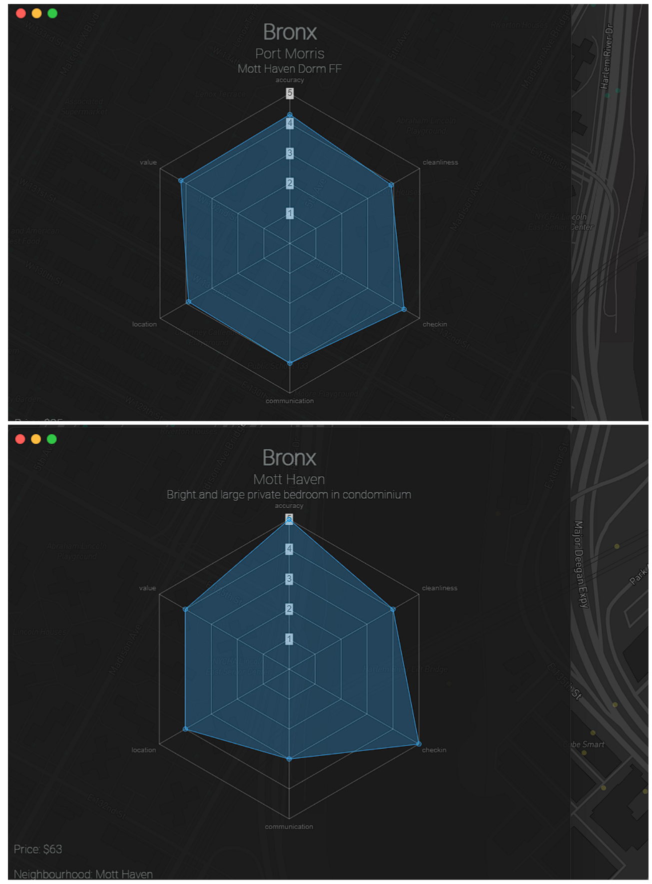

Figure 13: Two examples of the radar plots of Airbnbs in The Bronx

7.2 Use case 2: Improving an existing Airbnb

Secondly, someone who already has an Airbnb in NYC might be looking to improve their current business. Firstly they would use the interactive map to look up the area nearby their Airbnb (and assuming the data is up-to-date they would also be able to find the details of their own accommodation).

Assume the user wants to improve popularity of their Airbnb in Brooklyn. Looking around the area with our tool, it becomes clear that there is quite a high density of Airbnb locations here, especially as we get closer to Manhattan. The Airbnb would look like Fig. 14 in case the review data cannot be found. Of course, other Airbnbs do have this information available in the dataset. Next, we can take a look at some radar plots of other Airbnbs and check the bar graphs with the box select tool, as shown in Fig. 15.

Using these tools you can see that while there are many Airbnbs with good ratings in the same area, the availability is generally quite low. It may be a good idea to advertise high availability and show this on the listing. Other information you can find is that the price set for the Airbnb is lower than most other locations around, meaning you can use this information to either raise the price or advertise that the Airbnb is a cheaper option than others in the area. From using this tool the user can thus gain valuable insight in how to improve our Airbnb.

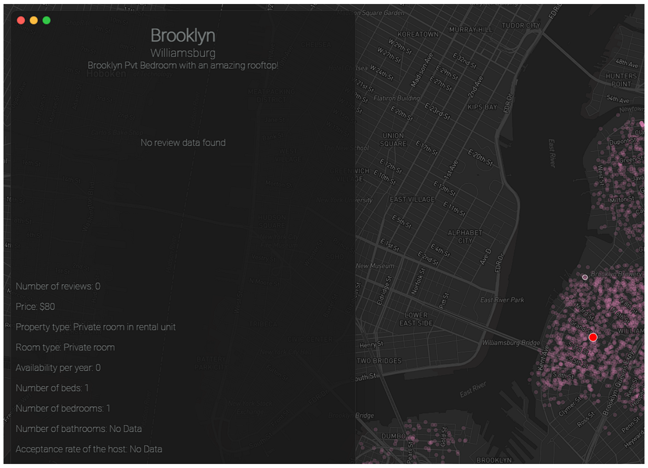

Figure 14: Our Airbnb location in Brooklyn

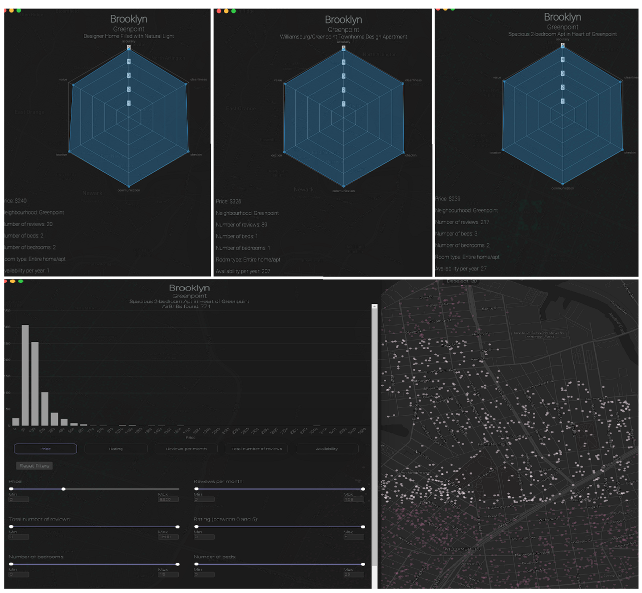

Figure 15: Information on other Airbnbs in Brooklyn

8 Conclusion

The presented visualization tool for Airbnbs in NYC has been developed over the course of 10 weeks. Although we initially started by using the Kaggle dataset, we switched to an extended version of this dataset available at Airbnb Inside [5]. An advantage of doing so was that we were now able to use some more detailed attributes, mainly about the review scores, which were visualized in the radar plots. However, the size of this new dataset was smaller, which was a compromise we had to make. As became clear from the user cases, **new insights** can be done with our tool by combining its different views and many interaction possibilities such as zooming, selection, and extensive filtering. This can be relevant for both a new Airbnb entrepreneur who is, for instance, interested in finding the best region for opening an Airbnb, and who acquires knowledge about this by checking the current offerings and ratings in different regions to see how much competition there is in a certain region. But at the same time, a current Airbnb entrepreneur can use the tool to compare his/her own Airbnb to the other options available on the market and to find certain trends in the data, e.g. by looking at the shapes of the radar plot, to see what should be improved, for instance.

Besides this, one of the advantages of our tool is its **user-friendliness**, which is immediately noticeable when opening the tool and viewing the pop-up window giving explanation about its functionality. Moreover, we have made sure to implement a **colour scheme** that is easily distinguishable for colourblind people, and matched the data characteristics to the visual encoding (e.g. using different hues for categorical data). However, due to **time constraints**, still some bugs will probably be present in the tool, which might limit the user experience at this point. Fortunately, these issues can be solved when time allows, and have nothing to do with the design choices made. Another disadvantage is the fact that the tool currently does not focus on finding correlations between sets of different attributes (e.g. how the number of beds relates to the price of an Airbnb, although one could filter on both aspects). To make this easier, a suggestion for **future work** would be to implement a scatter plot matrix (SPLOM) showing the relation between different attributes in the dataset. Building on top of this, the implementation of the light mode (next to the dark mode) would be interesting to realize as well, to make the tool even more suitable and customizable to the wishes of the user.

To conclude, our tool has been able to help Airbnb entrepreneurs to get an idea of the current market and what would be good investments to make. However, still some improvements could be done to enable different analyses or improve its usability further.

References

[1] Airbnb. Airbnb map view. `https://www.airbnb.com/?drawer_open=true`, 2023. Accessed: 2023-01-29.

[2] S. D.Guttentag and M.Havitz. Why tourists choose airbnb: A motivation-based segmentation study. *Sage Journals*, 57(3), Apr. 2017. doi: 10.1177/0047287517696980

[3] T. G.DUDÁS, L.BOROS and B.KOVALCSIK. The visualization of the spatiality of airbnb in budapest using 3-band raster representation. *Geographia Technica*, 12(1):23–30, May 2017. doi: 10.21163/GT_2017.121.03

[4] D. G.Zervas and J.W.Byers. A first look at online reputation on airbnb, where every stay is above average. *Marketing Letters*, 32:1–16, Nov. 2021. doi: 10.1007/s11002-020-09546-4

[5] Inside Airbnb. Source of the data used. `http://insideairbnb.com/new-york-city/?`, 2023. Accessed: 2023-02-03.

[6] J.Oskam and A.Boswijk. Airbnb: the future of networked hospitality businesses. *Journal of Tourism Futures*, 2(1), Mar. 2016. doi: 10.1108/JTF-11-2015-0048

[7] P. V.R.Perez-Sanchez, L.Serrano-Estrada and R.T.Mora-Garcia. The what, where, and why of airbnb price determinants. *Sustainability*, 10(12), Dec. 2018. doi: 10.3390/su10124596

[8] B. Wong. Color blindness. *nature methods*, 8(6):441, 2011.

[9] J. X.J.Ge. Analyse property data through visualisation. In *Proceedings from the PRRES Conference*, 25th Annual Pacific-RIM Real Estate Society Conference. PRRES, Melbourne, Australia, 2019.