Presentation and Visualization

Delivery 1: Dashboard

24th November 2021 - Rodolfo Castro, Xavier de Juan Pulido, Lorenzo Andrés Vigo del Rosso

First of all, it should be noted that we have chosen the **“Airbnb Listings in New York City”** dataset available in Tableau’s Public Resources as the dataset we will use during this course. This decision was taken due to the disponibility of temporal and geographical information in the set’s entries: we consider this fact will enable us to portray information in various visually appealing formats and also to draw conclusions based on the where the venues are located, which venues are close to each other, when the host posted them, the reported characteristics of the apartments and the ratings they were given.

Part 1: Data (bottom-up): Study and describe the data.

Our first task in this project is to describe the data included in the chosen dataset. To do so, we will reply to the proposed questions in Task 1:

1. **What type of data do you have? Has it a temporal nature? Has it a geographical nature?**

Our data has both temporal and geographical nature. For temporal, it has a “Host Since” feature which can give us a nice view of trends in airbnb rentals in NYC. For geographical, it has a couple of features, “Zipcode” and “Neighbourhood”, which can give us a different perspective when compared with prices and numbers of beds, for example.

2. **What is the range of values? Which units are used? What precision is required?**

Most values are categorical but we can still find numerical values in the dataset. One of the most important ones has a continuous nature: “Host Since", which ranges from 2008 (creation of AirBnb) until 2015. Exact day of “Host Since” is included in the feature but for a long term analysis and visualizations is probably not necessary, unless we find some correlations between seasonality and first time hosting.

Among the categorical properties, we might find the columns “Host Id” (ordinal), Name (nominal), Neighbourhood (nominal), Property Type (nominal), Room Type (nominal) and Zip Code (ordinal).

The rest of the data is numerical. Both rating columns include values from 20 to 100 (points), but we imagine the accepted range is from 0 to 100. The number of beds lies between 0 and 16 (beds). The number of records is always 1. The number of reviews presents values from 0 to 257 (reviews). Last, the prices range from 10 to 10000 (dollars, as the data is extracted from the USA). All range limits are inclusive.

In terms of precision, most values are completely concrete. However, price does not include cents and the review rating mean is rounded to the closest integer.

3. **What is it’s life span? (how often shall it be updated)**

Data should be updated at least every month, to get a good view of first time hosting being compared to other features. This data finishes in 2015, so for an optimal analysis we assume we have all the data from 2008 until 2015.

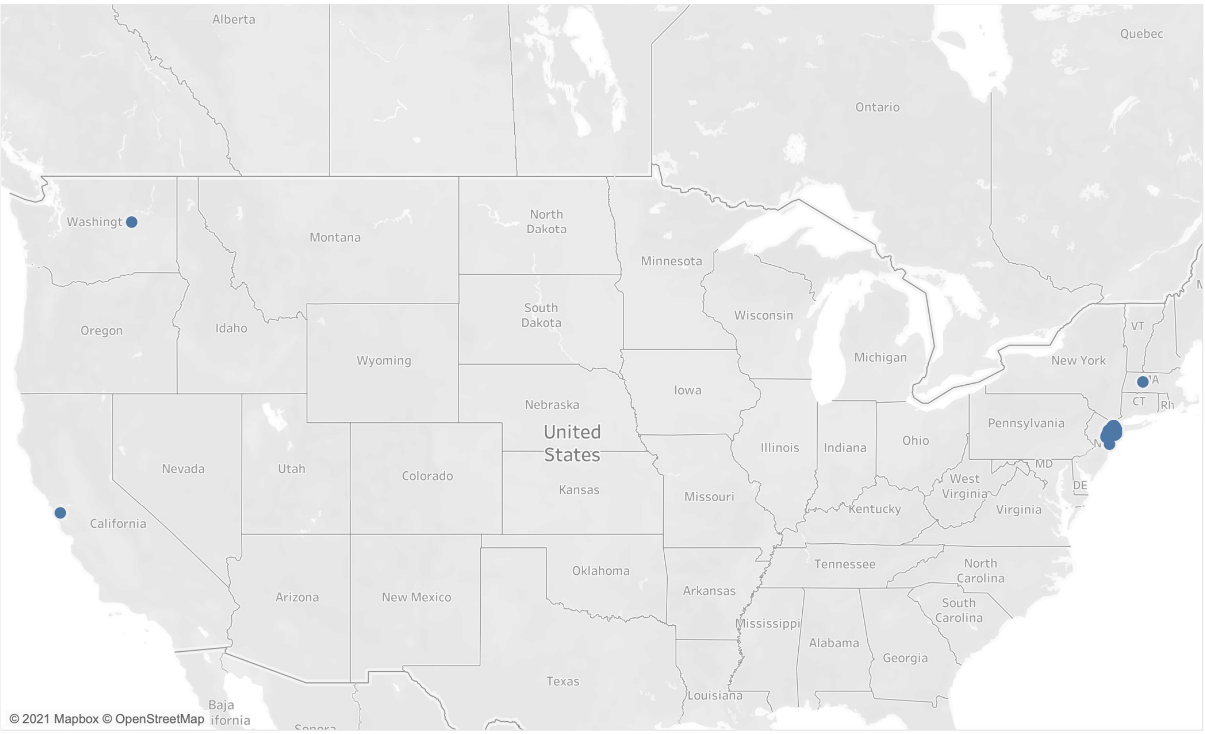
4. **Are there outliers? Identify relations and groupings.**

   Yes, there are several outliers.

   - Zip code: few apartments outside of NY: Washington State, California, Massachusetts and NJ.
   - Price: not that many cheap rentals but makes sense because it is NYC. All rentals available under 50$ and all those over 2000$ should definitely be considered as outliers.
   - Neighbourhoods: Manhattan and Brooklyn account for more than 80% of airbnb properties, mostly because they are the most popular neighbourhoods for tourists
   - Property Type: 90% are apartments. Everything else seems like an outlier, especially the one labeled as “Castle”, for example.
   - Room Type: 90% are “Entire home/apt” or “Private room”. For example, less than 2% of available rentals are labeled as “Shared room”, which could be considered as an outlier.. Also, it is pretty clear that entire homes or apartments are more expensive than private rooms.

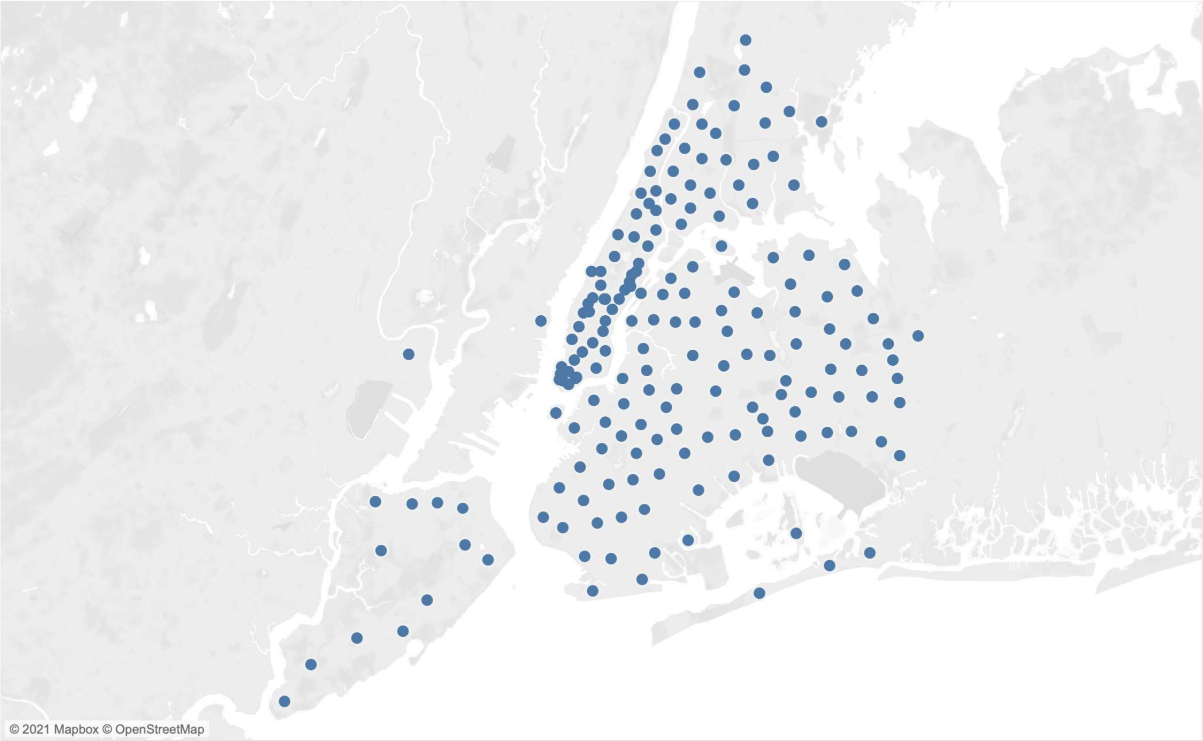
5. **Create some graphs to explore distribution and relations. Explain what you observe.**

Zip Code Distribution with Outliers



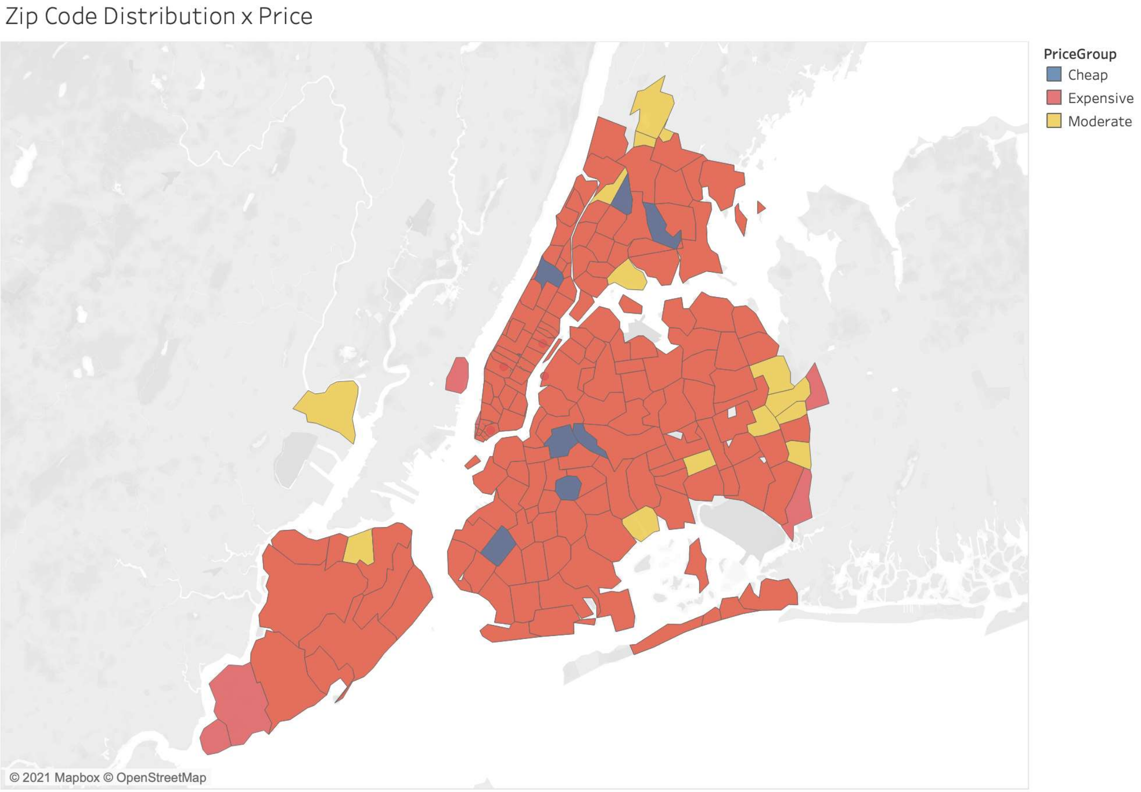
Map based on Longitude (generated) and Latitude (generated). Details are shown for Zipcode. The view is filtered on Latitude (generated) and Longitude (generated). The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

Zip Code Distribution



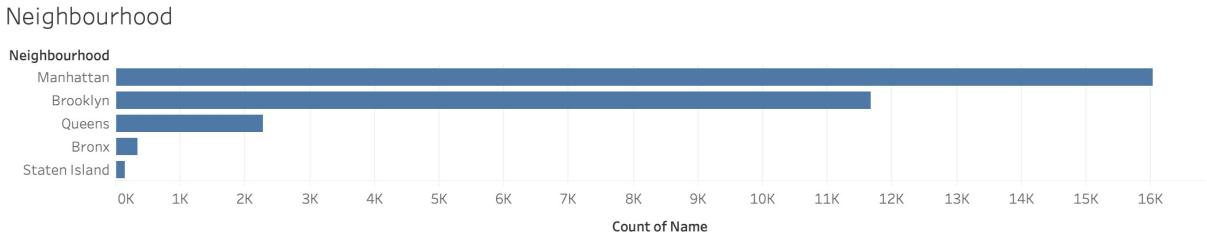
Map based on Longitude (generated) and Latitude (generated). Details are shown for Zipcode. The view is filtered on Zipcode, Latitude (generated) and Longitude (generated). The Zipcode filter excludes 01003, 07712, 94103 and 99135. The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

These two images show the zip code geographical distribution from each property. The first one shows a cluster of hosts in New York city and also some isolated properties outside this area, the ones that will be considered as outliers. The second one zooms in the New York cluster showing the properties distribution from each neighbourhood.



Map based on Longitude (generated) and Latitude (generated). Color shows details about PriceGroup. Details are shown for Zipcode. The view is filtered on Zipcode, Latitude (generated) and Longitude (generated). The Zipcode filter excludes 01003, 07712, 94103 and 99135. The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

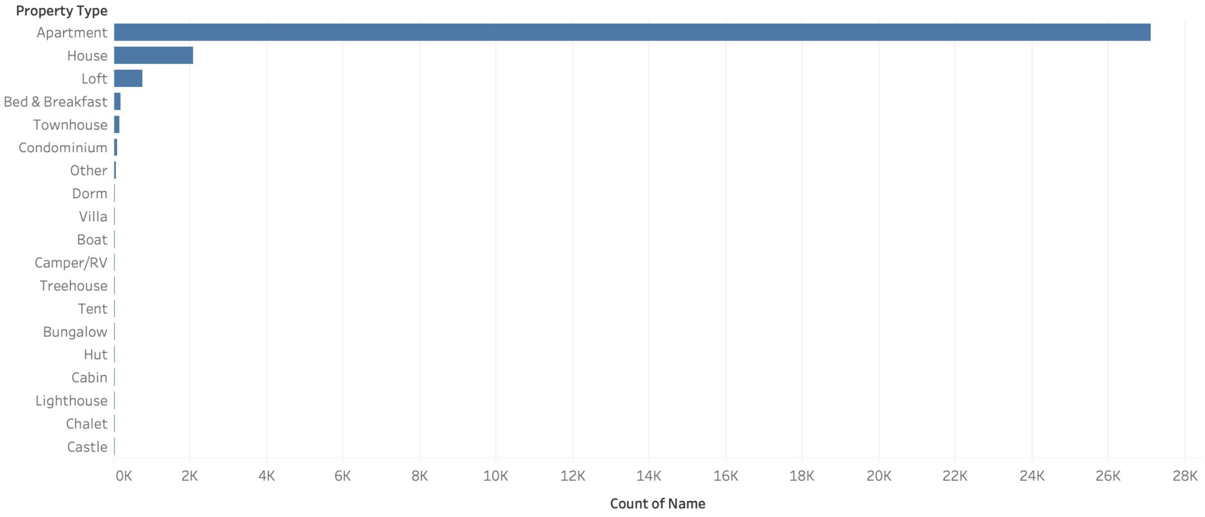
Again, the map shows the properties distribution in New York but adding a price feature. Each color groups price ranges into a categorical variable with three possible values: cheap (blue), moderate (yellow) and expensive (red).



Count of Name for each Neighbourhood.

The above bar plot represents the number of properties in each New York neighbourhood. It is clearly shown that the most number of properties is located in Manhattan and Brooklyn.

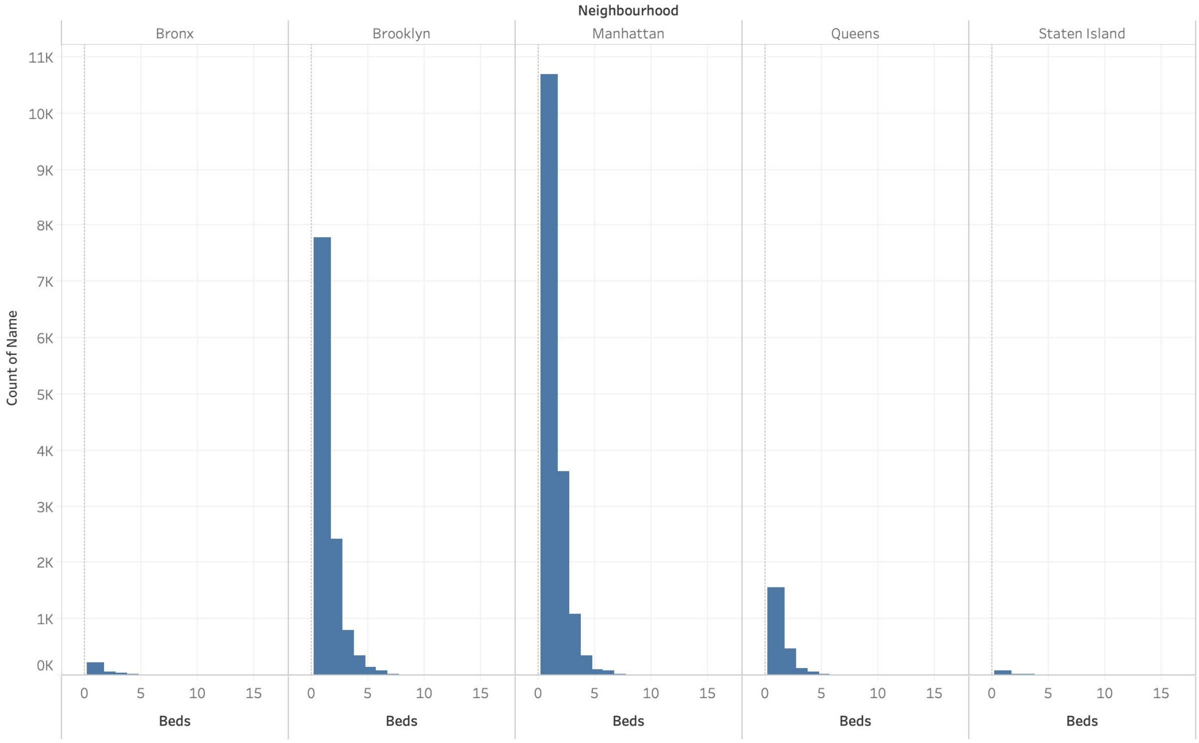
Property Type



Count of Name for each Property Type. The view is filtered on Property Type, which excludes Null.

For each host, it is specified the property type being rented. The previous bar plot represents the number of properties grouped by its type. It is observable that most properties are apartments.

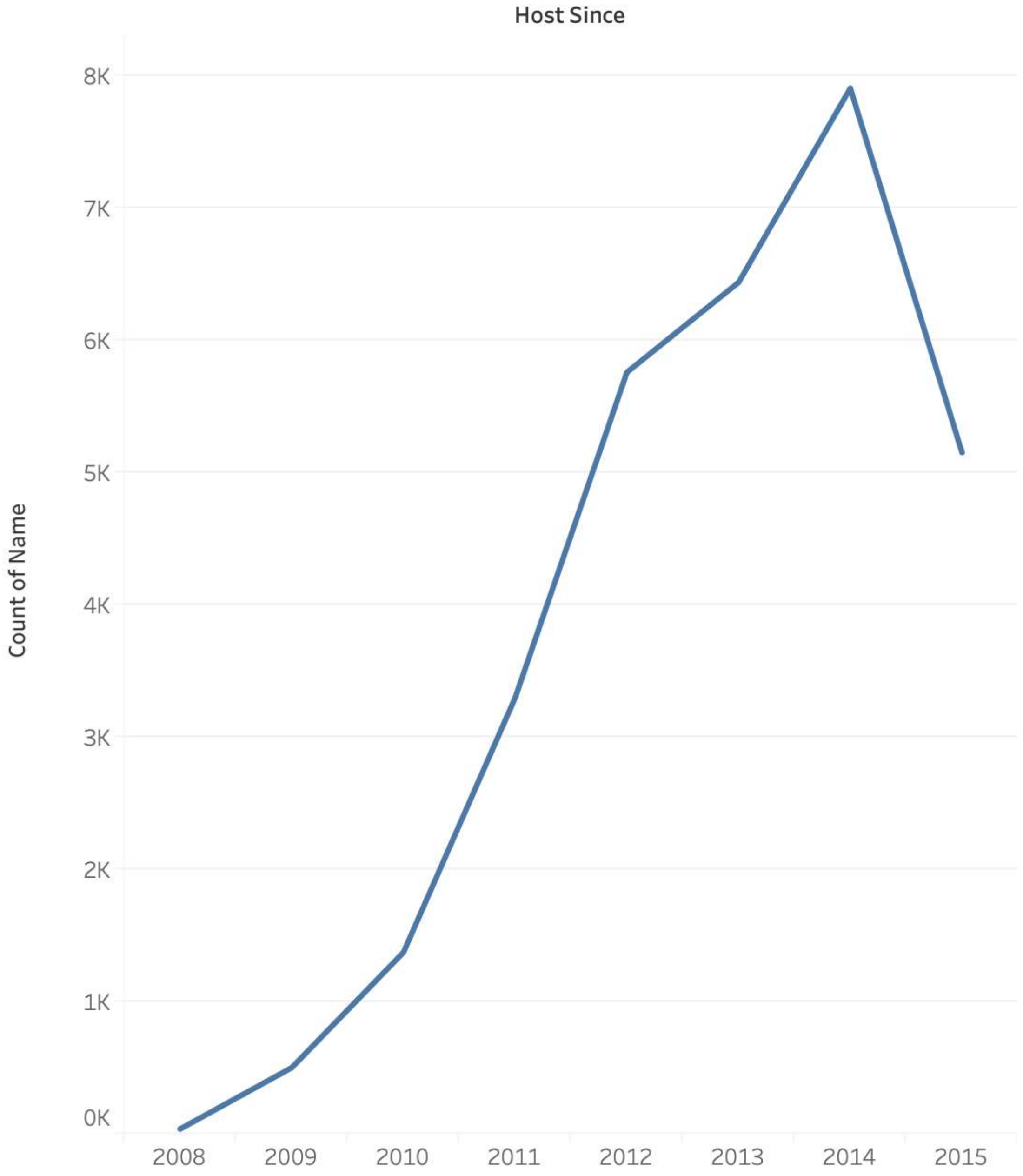
Neighbourhood x Bed Count



The plot of count of Name for Beds broken down by Neighbourhood. The view is filtered on Beds, which keeps non-Null values only.

For each property, we have the number of beds. The previous chart shows, for each neighbourhood, the number of properties with a specific number of beds: 0, 1, 2…

First Time Hosting Year



The trend of count of Name for Host Since Year. The view is filtered on Host Since Year, which excludes Null.

The last graph represents the amount of properties in renting from 2008 to 2015. The value from each year represents the number of properties that started to be in the airbnb market.

Part 2: Audience (top-bottom): Who is your audience?

Our second task in this project is to describe the audience. Before we do anything else, we list who is the potential audience to then be able to describe them:

- USA / NY Government (public sector)
- AirBnb hosts / real estate investors (private sector)

Now that we have a view of who the audience is, we can start asking more questions. To do so, we will reply to the proposed questions in Task 2:

1. **Are they experts in the use of charts? Which platform will they use to see your charts? How often will they consult the visualization?**

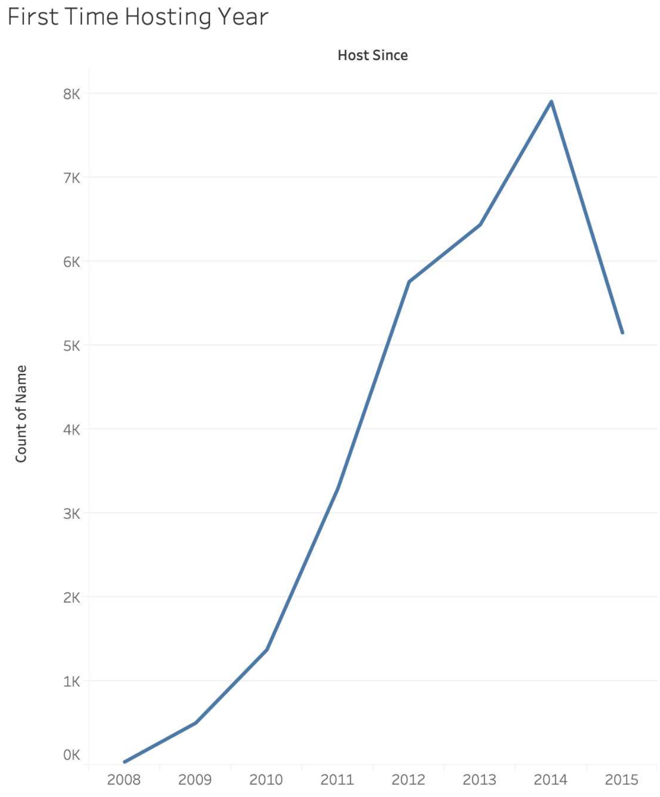
Our audience are definitely experts in the use of charts. When talking about real estate development and investing, either private or public, using charts and graphs are essential. Obviously it depends which type of chart but at least some geographical chart about prices, for example, is something all our audience should be an expert of.

As far as platforms, the audience will prefer to see the charts through a web app view or interactive dashboard. Both platforms used in class Tableau and Altair would be useful but only in the development of the charts; the audience will not be proficient in these tools, but they will be able to understand the charts.

Consultation of the visualization depends on the involvement of projects in the area. For example, someone in the private sector, let's say an AirBnb investor looking to open an Airbnb property in NY, will consult the visualization regularly most likely for the entire life of their Airbnb journey. They will use these charts and visualizations to compare their listing with others and see how they can fine tune some features to be able to beat the competition, for example lowering the price or buying a sofa bed for extra people.

2. **Describe a persona, their goal (as questions they may want to answer) and a scenario of use of your visualization**

A USA or NY Government official, from the Mayor's office, senators or representatives, might want to get a better look at the expected economic growth of their state (in this case NY). A few indicators of economic growth they might want to look at are: tourism, new housing development and prices of houses. From our data and visualization, we can give a glimpse of many of these indicators. We can use the chart below to show that first time airbnb listings are declining. The government official can then dig deeper on why this is happening. Are houses too expensive? Are tourists not attracted to NYC anymore? And so on.



The trend of count of Name for Host Since Year. The view is filtered on Host Since Year, which excludes Null.

Another persona from our audience could be a host from Airbnb. Someone registered on the platform as a host having a property in New York that could or not be in the platform to be rented.
The hosts might want to consult nearby properties to set a price for the rental. From our data we could give the host the average price, the number of guests, the average reviews and the property types being rented near the host property. With this information, hosts will be able to fine tune features from its property and beat the competitors in its area.

3. **What message do you want to convey with your data? What questions do you want to answer (not only for end-users; think of internal customers as well)?**

The message we want to portray towards the public sector is that the growth of the Airbnb business leads to economic growth of the city, as it transforms the touristic demand into a new use for properties and a new source of income to property owners and direct and indirect employees. The questions that may be answered are:

- Is tourism to the city (NY) growing or slowing? (Represented by the increase of available Airbnb versus trying to supply a higher touristic demand).
- Are tourists having a good experience in the city taking into account the ratings they are giving to the properties they stay in?
- How much are tourists willing to spend in bookings when they come to the city?
- Is Airbnb slowly taking up too many properties in the city and leaving New Yorkers with no place to live? Is the ratio of available properties for citizens similar throughout the neighbourhoods?

From a host perspective, the main message could be a description of the area nearby to a host property in NY. The description includes a market comparison analysis based on the nearby property features present in our dataset. The questions we want to answer are:

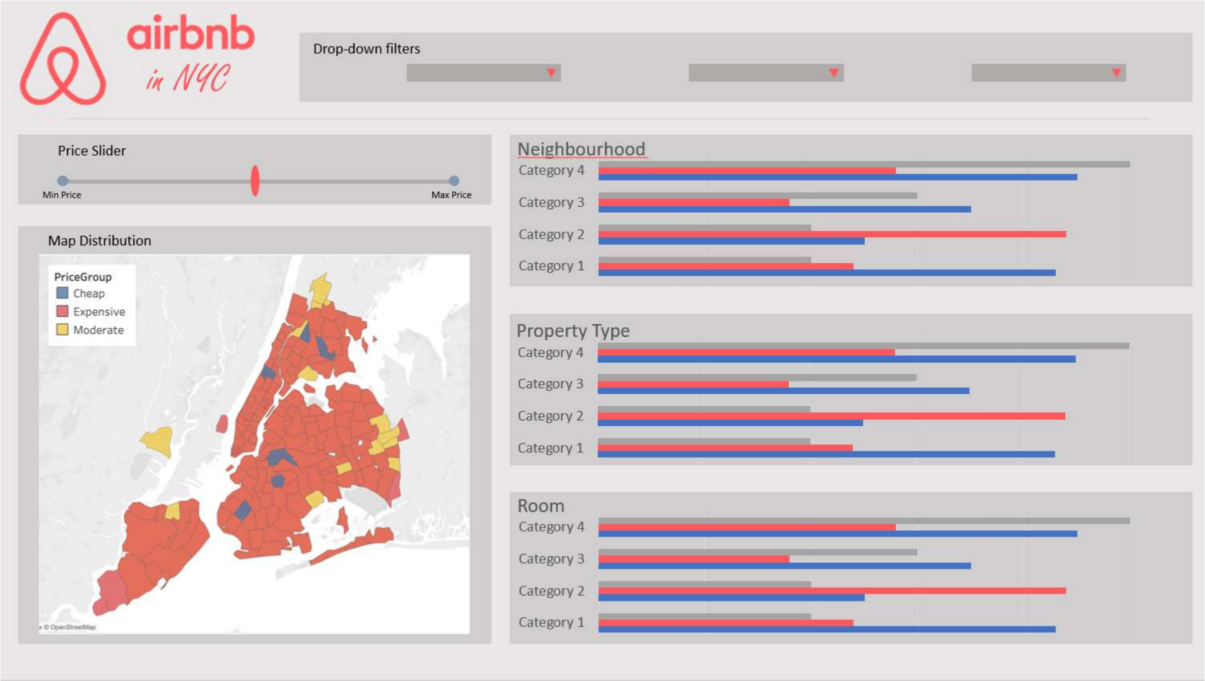
- How many properties are being rented in a specific area (in NY)?
- Which is the average price in a specific area (in NY)?
- Which is the average rating in a specific area (in NY)?
- Which is the most common property or room type in a specific area (in NY)?
- How can I improve my property to beat the competitors?
- From a predicting perspective, which could be the price for my property?

In our project, we will work on portraying the message destined to the hosts as it is more interesting to us to work with users of our platform. Also, the questions related to the hosts imply using more of the data available in the dataset.

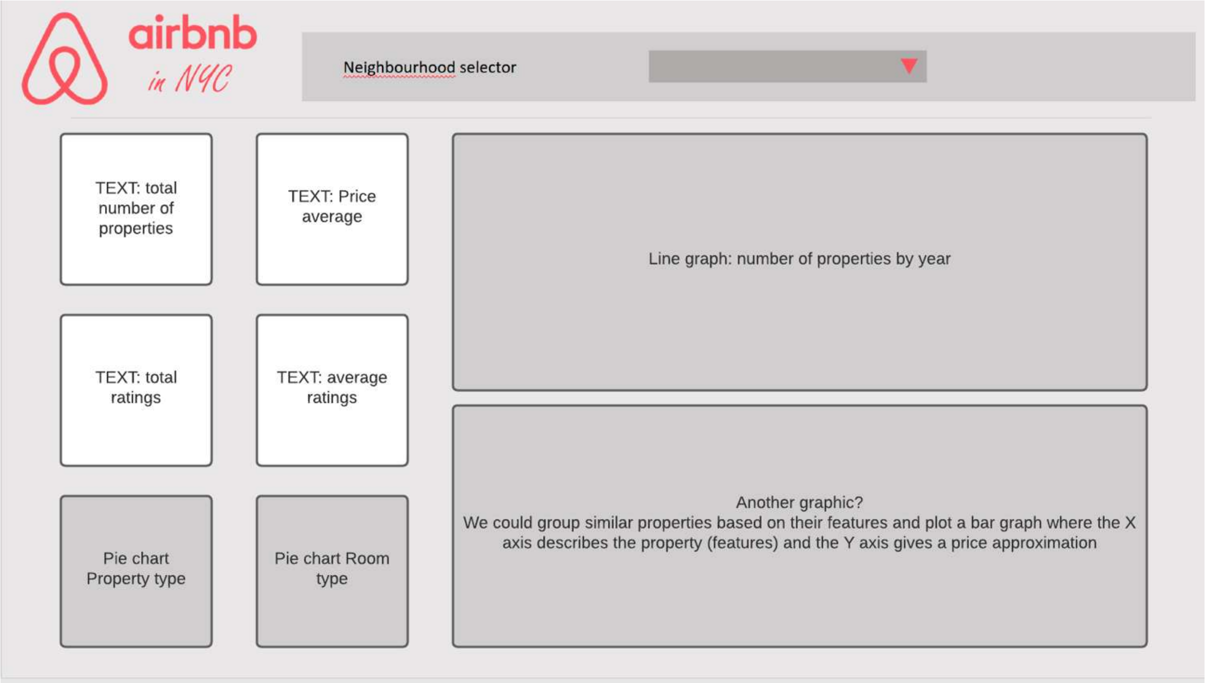
4. **Create a "napkin" design, low-fi, and explain how this design will answer the proposed questions. Include any contextual information that is relevant to your message**

In order to start the design of our dashboard, we were asked to present a first napkin of it. This napkin will be an initial iteration of our design consisting in a first sketch or a simple approach to the final result.

We are expecting our dashboard to include at least two screens:



We want to include an interactive map which data could be restricted by the viewer through the map, the list selectors at the top and through a price slider. We expect to show information on the density of available venues per geographical zone or neighbourhood and the distribution of venues in terms of property type and number of rooms. The following filtering criteria have been already considered: price and different categories.


airbnb
in NYC
Neighbourhood selector
TEXT: total number of properties
TEXT: Price average
Line graph: number of properties by year
TEXT: total ratings
TEXT: average ratings
Another graphic?
We could group similar properties based on their features and plot a bar graph where the X axis describes the property (features) and the Y axis gives a price approximation
Pie chart Property type
Pie chart Room type

In the next screen, we thought about adding first glance statistics at the left. This way, viewers will receive concise and very relevant information almost immediately. We also considered the idea of adding graphs over time on this screen that show the evolution of the business.

There are some ideas we want to take into account in the soon future. We may decide to change the order of the screens and choose a different color schema and fonts. Also, there are some gaps yet to be filled in the second screen, where we may add graphs analysing the relationship between variables. Lastly, we have been thinking about different graphs for the left bottom corner in the last image such as a graph showing the price distribution with respect to other properties.

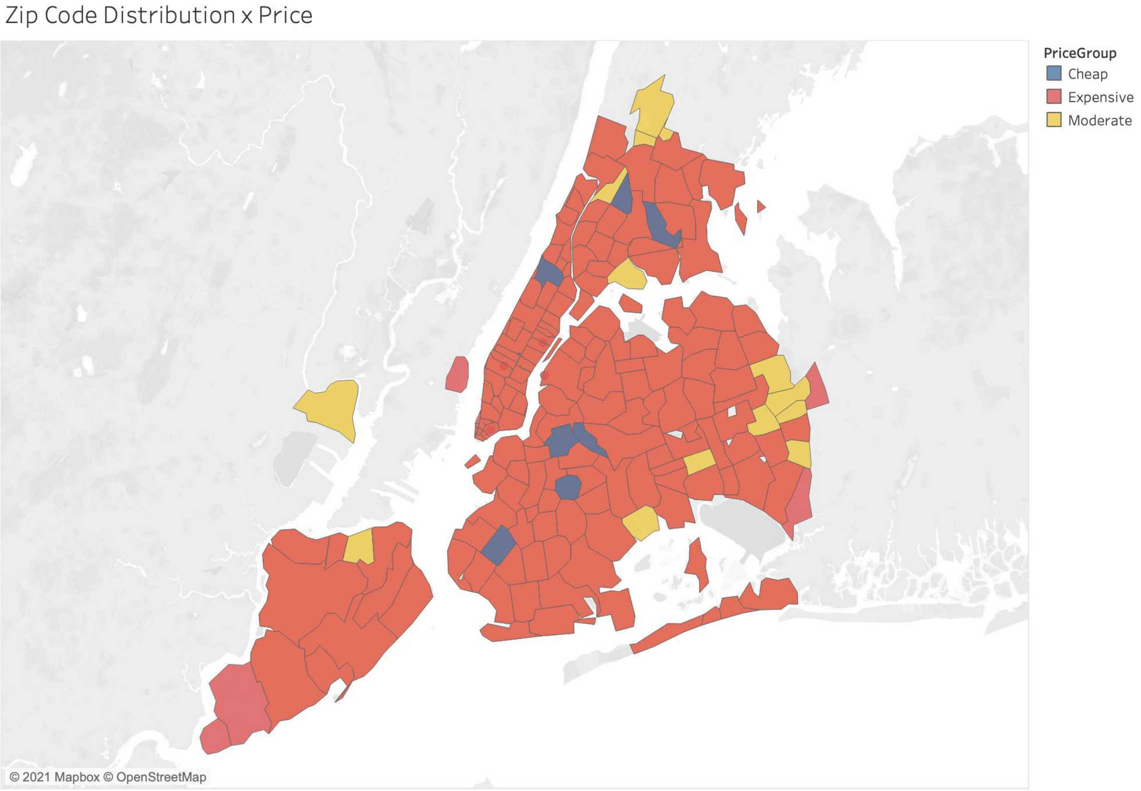
As to how these graphs would answer the user's questions:

- How many properties are being rented in a specific area (in NY)? We are showing a graph with the distribution of properties through all the neighbourhoods.
- Which is the average price in a specific area (in NY)? By selecting the neighbourhood in the second slide, the average price will be restricted to that area.
- Which is the average rating in a specific area (in NY)? Following the same criteria as the previous question.
- Which is the most common property or room type in a specific area (in NY)? We are showing a graph with the distribution of the property types for the whole NY, but it may be restricted to an area selecting a specific neighbourhood in the graph above.
- How can I improve my property to beat the competitors? Through all the information shown in the graphs and by filtering it as already explained, you may get an insight of the kind of properties that are offered in every neighbourhood along to the price they are rented for.
- From a predicting perspective, which could be the price for my property? You may restrict the data to properties similar and close to yours, and then analyze the average price.

Part 3: Selection of chart and encoding

1. **Decide the type of charts and the encoding of the data taking into account the objectives. Describe decisions on encoding and used charts. Justify decisions taken based on perception properties, principles and best practices. Prepare the layout taking into account Gestalt principles of grouping and relating.**

In order to achieve the objectives we stated in previous deliveries, we need to select which charts we want to use for our dashboard, select the encoding of these ones and justify these selections. Let’s discuss the decisions taken for each chart of our dashboard:



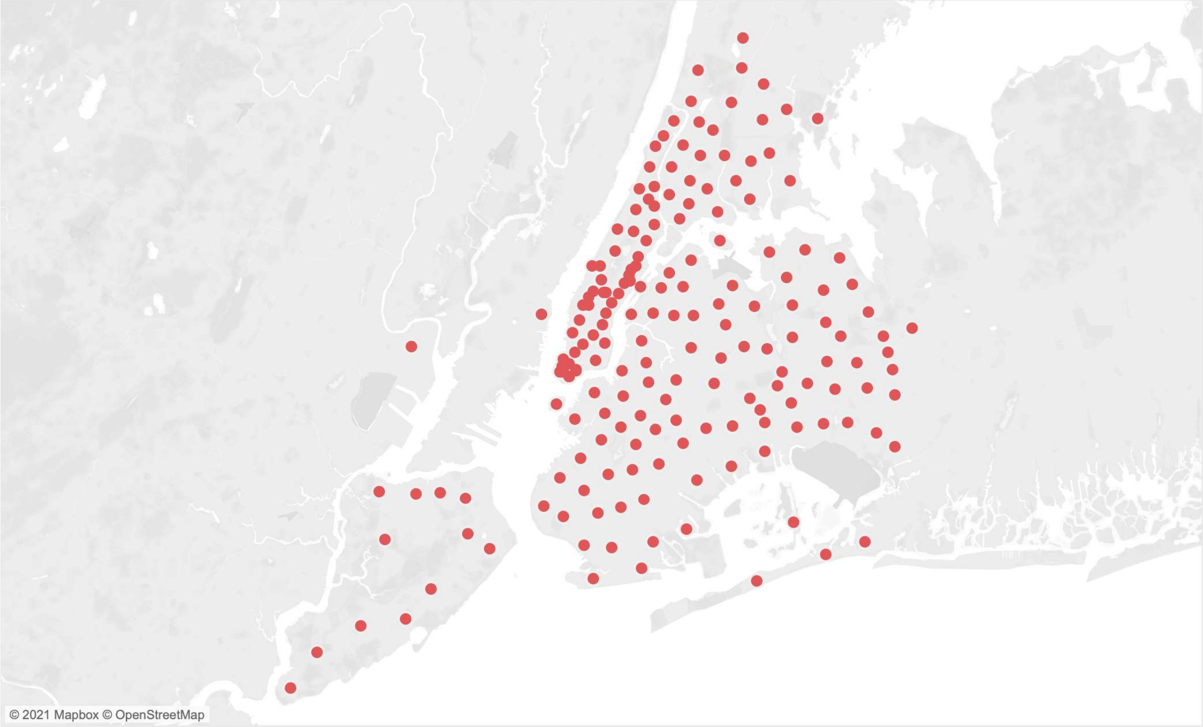
Map based on Longitude (generated) and Latitude (generated). Color shows details about PriceGroup. Details are shown for Zipcode. The view is filtered on Zipcode, Latitude (generated) and Longitude (generated). The Zipcode filter excludes 01003, 07712, 94103 and 99135. The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

Type: Basic Choropleth

Encoding: Price (grouped in 3 classes) and Zip Code (grouped according to their districts). Each one of the price ranges is encoded by a different color and there is a legend to guide the user. As most of the districts end up classified as ‘Expensive’, we may add in the future more ranges or modify the thresholds among the ones that already exist. Moreover, changing the legend labels to numerical values and modifying the color notation may be good ideas to apply in the following iterations. Chart ranges will depend on the filters included in the dashboard such as sliders (for price) or dropdown to show only specific neighborhoods.

Justification: Geographical data is easier to understand when located in a map. This representation is a good summary of the price distributions per area and allows the user to identify patterns depending on location at first glance. In addition, viewers will be able to compare data respect to location immediately, especially between neighbouring or close districts.

Zip Code Distribution

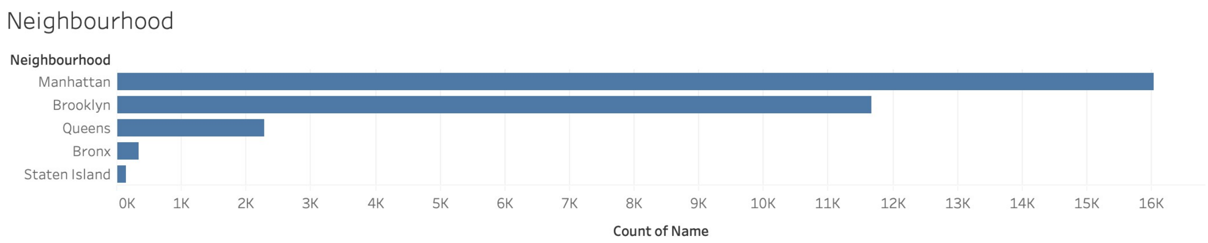


Map based on Longitude (generated) and Latitude (generated). Details are shown for Zipcode. The view is filtered on Zipcode, Latitude (generated) and Longitude (generated). The Zipcode filter excludes 01003, 07712, 94103 and 99135. The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

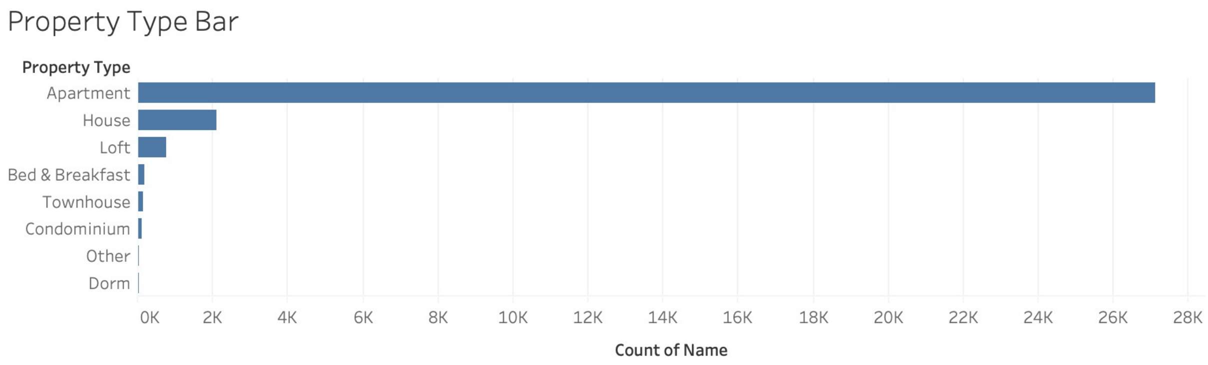
Type: Proportional Symbol

Encoding: Size of the symbol: there is a circle per district representing the total amount of Airbnb related properties that can be found there.

Justification: We were aiming to show the absolute value of the number of properties per district, rather than a ratio. Sadly, there are not great differences in values so this graph type falls a little short: it only shows that there is an even distribution. We should review the data and the chosen graph type.



Count of Name for each Neighbourhood.



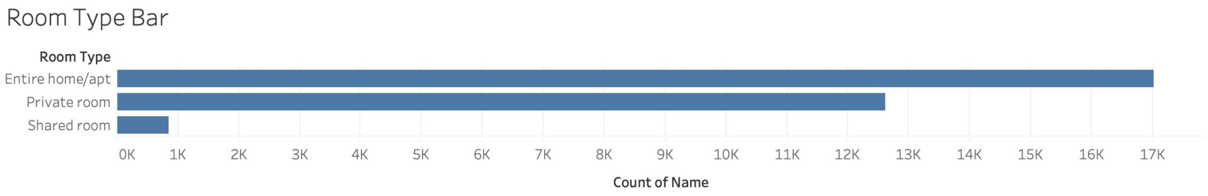
Count of Name for each Property Type. The view is filtered on Property Type and count of Name. The Property Type filter excludes Null. The count of Name filter ranges from 31 to 27,102.

Type: Ordered Horizontal Bar Chart

Encoding: Categorical variables in the vertical axis, amount in the horizontal axis. There is only one series, so there is no specific color encoding.

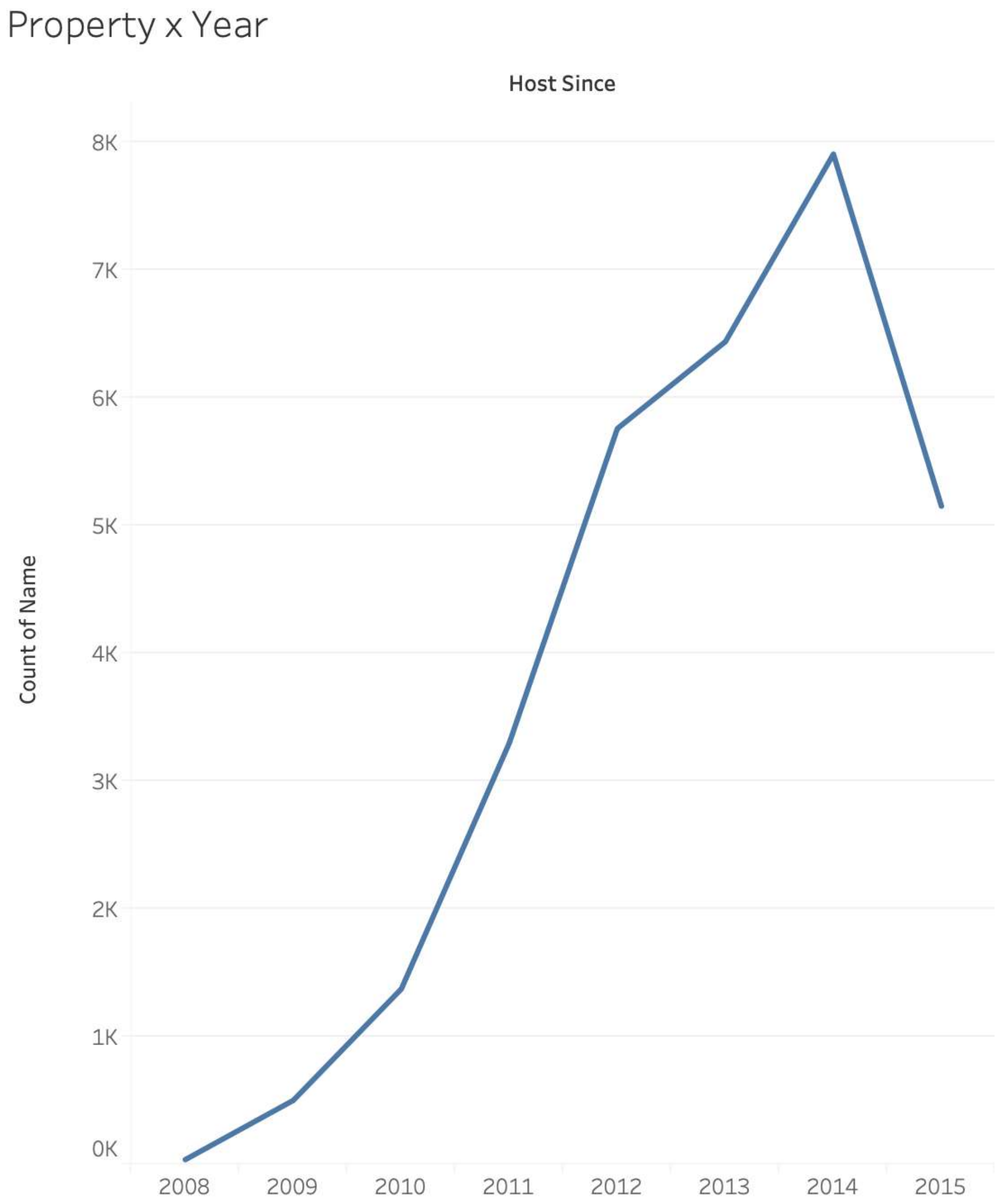
Justification: We wanted to show the values of these two categorical variables (Neighbourhood of the venues and property type) and a bar chart is the simplest and most efficient tool to do so. We could have chosen a pie chart, but we consider that the ratio is not as relevant as the absolute value in these two cases.

We choose the horizontal version as it is more adept for categorical variables (vertical bars are more common when the independent variable is related to time, for example). Ordering the value gives the viewer a sense of ranking: this may be useful in order to convey our message to the hosts or potential hosts to invest in Airbnb and where to invest (where everybody does and is proven to work or in less usual places and try to stand out).



Count of Name for each Room Type.

In the case of Room Type, we are also considering conveying the information in the shape of a pie chart. Given that there are less possible values for the categorical variable and that the absolute value may be inferred from the amount of properties, showing just the ratio may do the job. We will show the result in the following pages.



The trend of count of Name for Host Since Year. The view is filtered on Host Since Year, which excludes Null.

Type: Line Chart

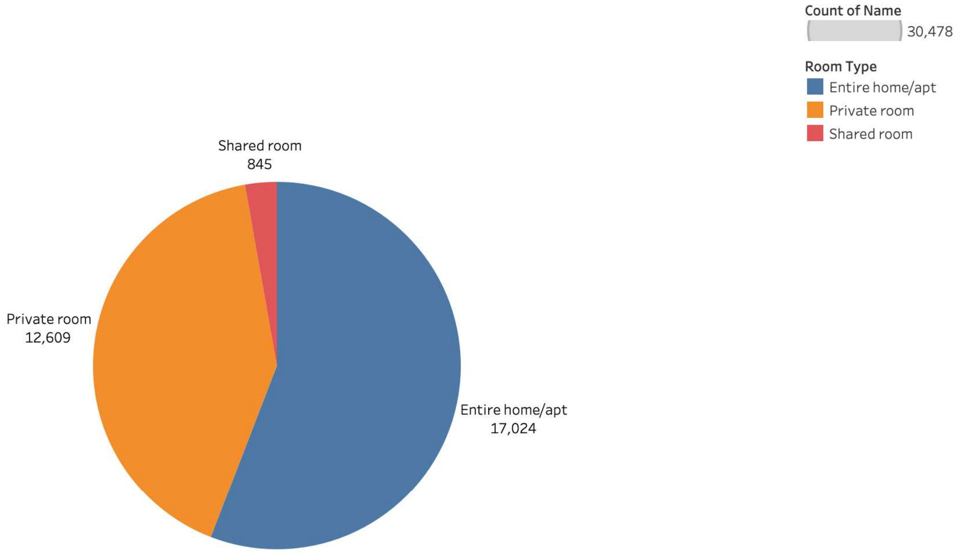
Encoding: Time in the horizontal axis, amount of new hosts per year in the y-axis.

Justification: As a part of our message we want to convey that Airbnb is a strong and growing business, in which potential hosts should start investing and current hosts should consider investing more. For that reason, we wanted to show the evolution over the time of the number of hosts in NYC Airbnb: to make it clear that more and more people are taking part in the business every year.

The best graph type to show an evolution over time is the line chart.

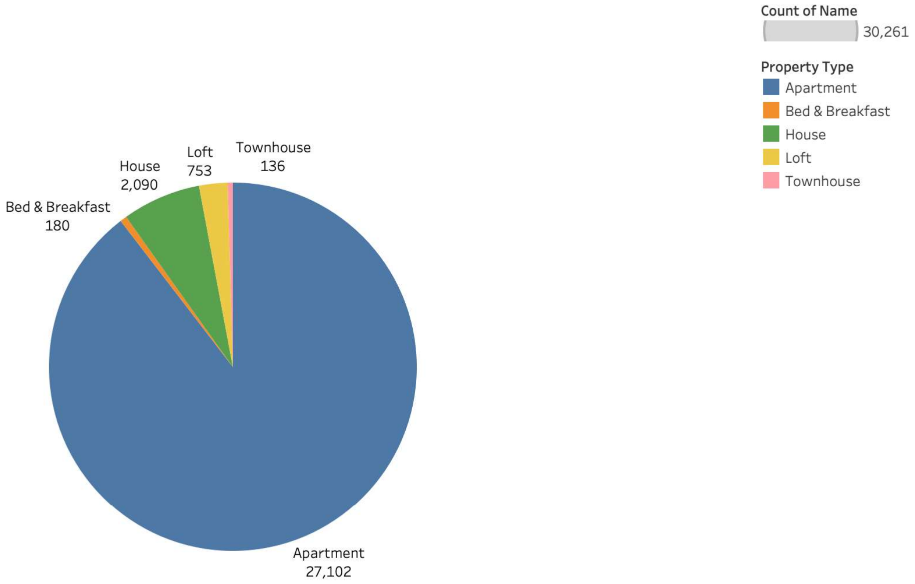
The drop in the last year may be due to the dataset not reaching the end of 2015. Also, we would have loved to show an accumulative evolution over time, but we have no information on hosts that left the business each year.

Room Type Pie



Room Type and count of Name. Color shows details about Room Type. Size shows count of Name. The marks are labeled by Room Type and count of Name.

Property Type Pie



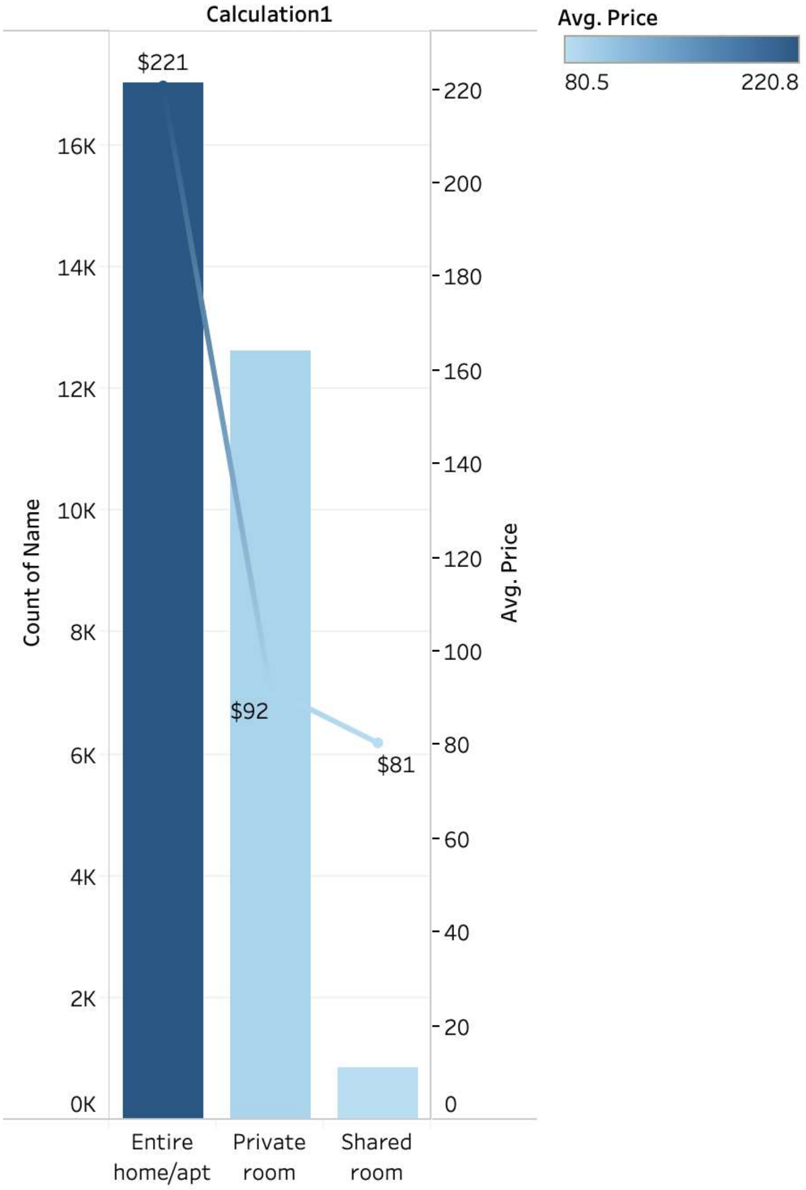
Property Type and count of Name. Color shows details about Property Type. Size shows count of Name. The marks are labeled by Property Type and count of Name. The view is filtered on Property Type and count of Name. The Property Type filter excludes Null. The count of Name filter ranges from 100 to 27,102.

Type: Pie chart

Encoding: The chart encodes the components, and its size, of a categorical variable. The first one encodes the different room types and the second one the different property types in our dataset.

Justification: We want to show the components and size distribution of these two categorical variables. The pie chart allows us to determine the different values for that variable and the distribution over our dataset. By using the sizes of each category in the chart, we can compare the number of properties from each category and it’s easy to understand the proportion of each one inside the dataset.

Dynamic Table



The trends of count of Name and average of Price for Calculation1. Color shows average of Price. The view is filtered on Calculation1, which keeps Entire home/apt, Private room and Shared room.

This is a dynamic chart where the user can choose the variable in the x-axis showing different representations based on our data. However, the type of chart will be the same.

Type: Column + line average

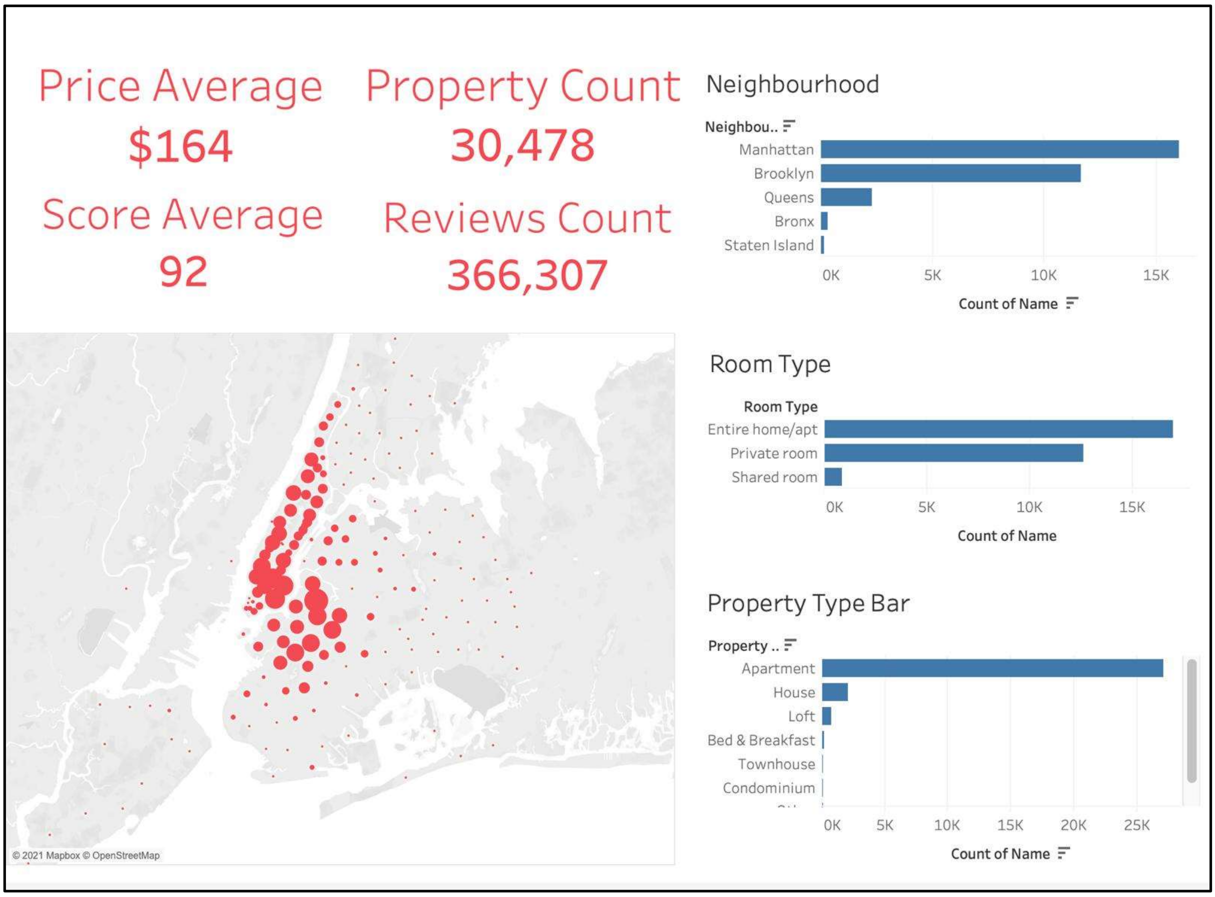
Encoding: categorical data on the x-axis (depending on the variable) and the count of that category in the y-axis. There will also be a line joining the price average for each category.

Justification: this type of chart allows us to show the correlation between (each category of) a variable and the property price average. With this chart we want to allow the user to select a variable on the x-axis and see the impact in the property price average.

2. Create a 2nd “napkin” design, hi-fi.

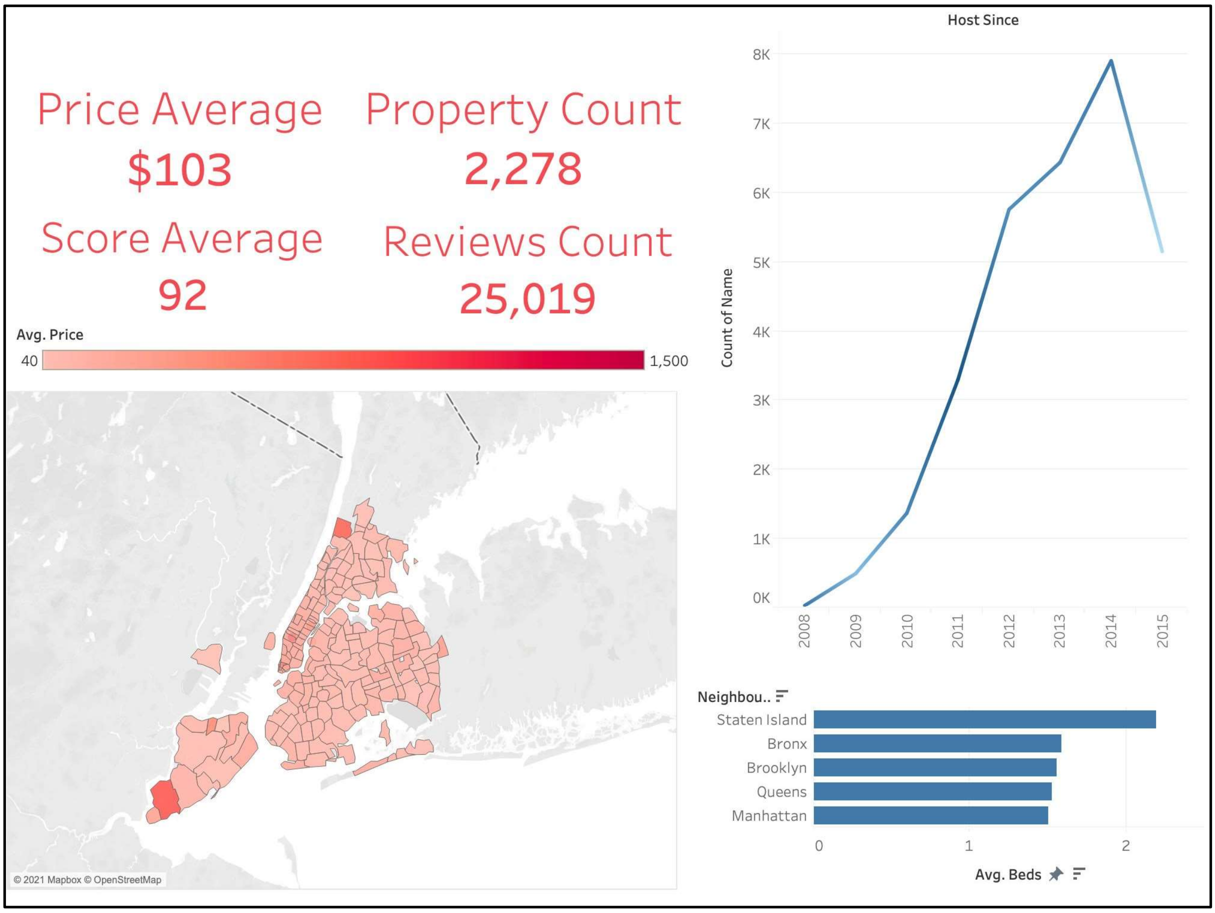
In order to deliver a hi-fi design, we decided to refine our charts and start to think about our final delivery. We needed to start taking into account perception, design aesthetics such as scales and labels. For this purpose we’ve prepared the following design.

As a main page:



First of all, we can see the aggregations at the top left. This first glance information encodes interesting data for the user so we had to redistribute the space. We’ve also modified the map representations changing the sizes to emphasize differences on the number of properties on each area. And we decided to use the bar charts to show generic information about our data because we thought the absolute value is more relevant for the categorical variables being represented.

And as a second page:



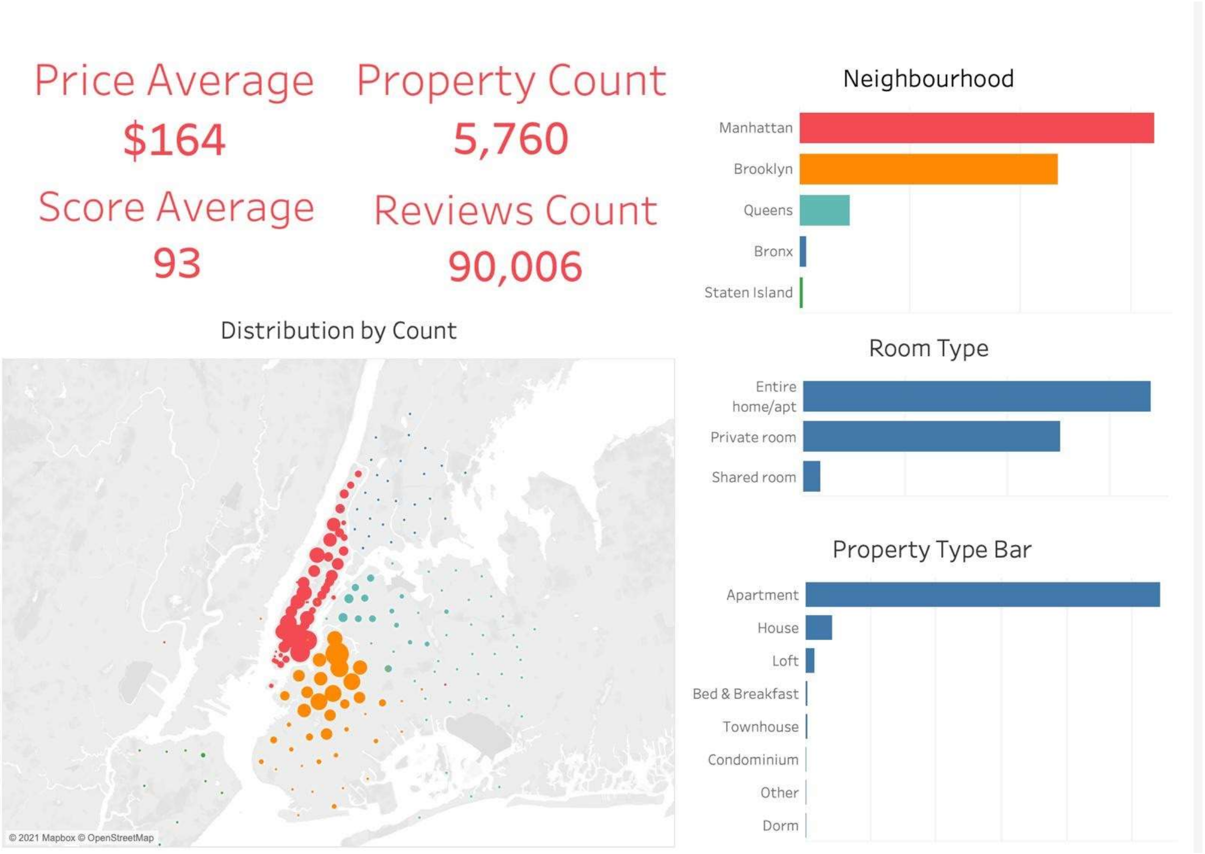
On this page we can also observe the first glance information stated before. The map on this screen has also been modified to include a color scale for the price average. At the right side of the page, we have the evolution of host incorporation in the platform. And, at the bottom right, there's an interactive bar plot that can encode multiple variables. The idea is to let the user choose a variable and represent it on the plot.

As a final comment, we think this is not a final design and we present a possible distribution of our charts. In future deliveries we want to unify the color palette, manage whitespaces and margins and give the dashboard a corporate aesthetic.

Part 4: Implementation

1. **Prepare the final dashboard including all your charts in an HTML page. You may ask support from the professor for this final task, as we have not dedicated a lot of time to HTML.**

As we approached the final delivery, we once again concluded that a refinement of our dashboard had to be performed. We tried to improve the titles, labels, and color used to give some context on the map visualizations.
The dashboard for this week is the following:



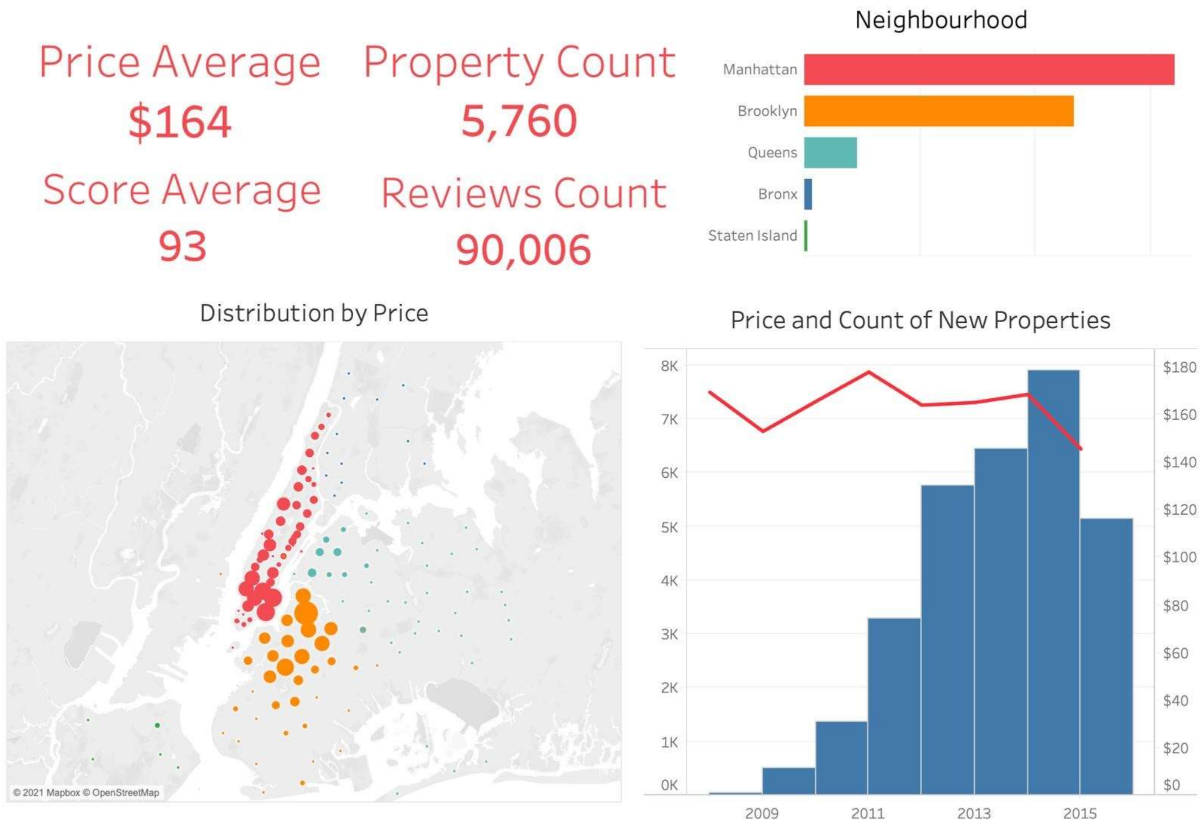
Price Average
$164

Property Count
5,760

Score Average
93

Reviews Count
90,006

The first dashboard page is similar to the previous delivery, we focused on giving clarity to the map and preserving consistency with the neighbourhood bar chart. We have given different colors to each neighbourhood and then the map has been painted preserving these colors in order to clarify which properties belong to each neighbourhood.



We have remodeled the second dashboard page in order to give different information. We also applied the colors strategy on the bar chart being consistent with the map. The map now shows the distribution of price and on the right side we plot the evolution of the properties during last years: the price evolution and also the new properties evolution.

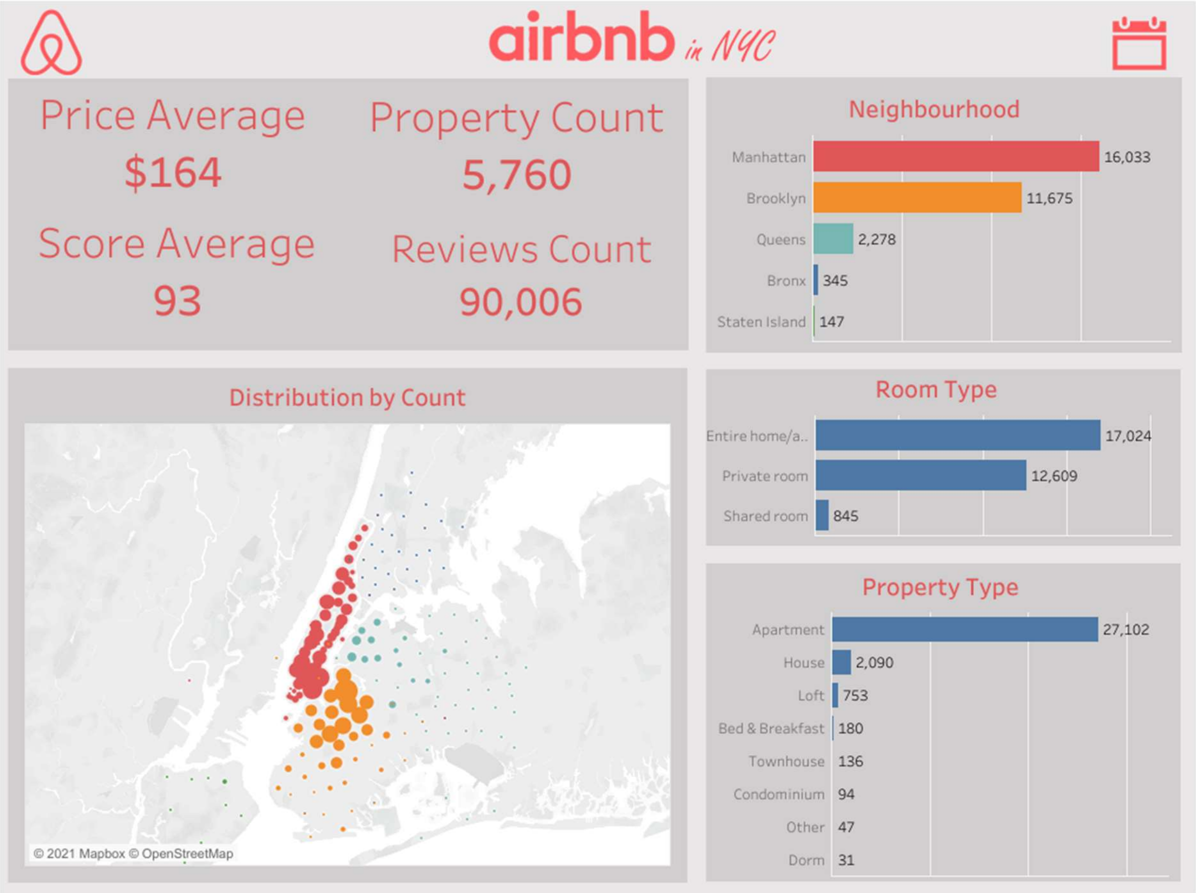
In the final delivery we will be looking forward to giving our dashboard a final corporative appearance by adding a header with the Airbnb logo and maybe a slogan to help us convey our message: that Airbnb is a solid and growing business that has spread all over NYC and that represents a great opportunity to invest in.

Part 5: Final dashboard

Finally, we are in the final delivery. As we said in the previous section, this delivery focuses on giving a corporative appearance to the dashboard. Also, based on the feedback from the last delivery, we tried to improve our dashboard by applying some modifications:

- Include the logo, a title and a navigation button in the header of the dashboard.
- As we used some charts that don’t include axis, we labeled the bar with the value count for each of them. This also allows to compare the charts even if they are not in the same scale.
- We also decided to remove the dual chart and replace it with two line charts showing the evolution and the information clearly.

The final dashboard is the following:



airbnb in NYC

Price Average
$164

Property Count
5,760

Score Average
93

Reviews Count
90,006

Neighbourhood

Manhattan	16,033
Brooklyn	11,675
Queens	2,278
Bronx	345
Staten Island	147

Distribution by Count

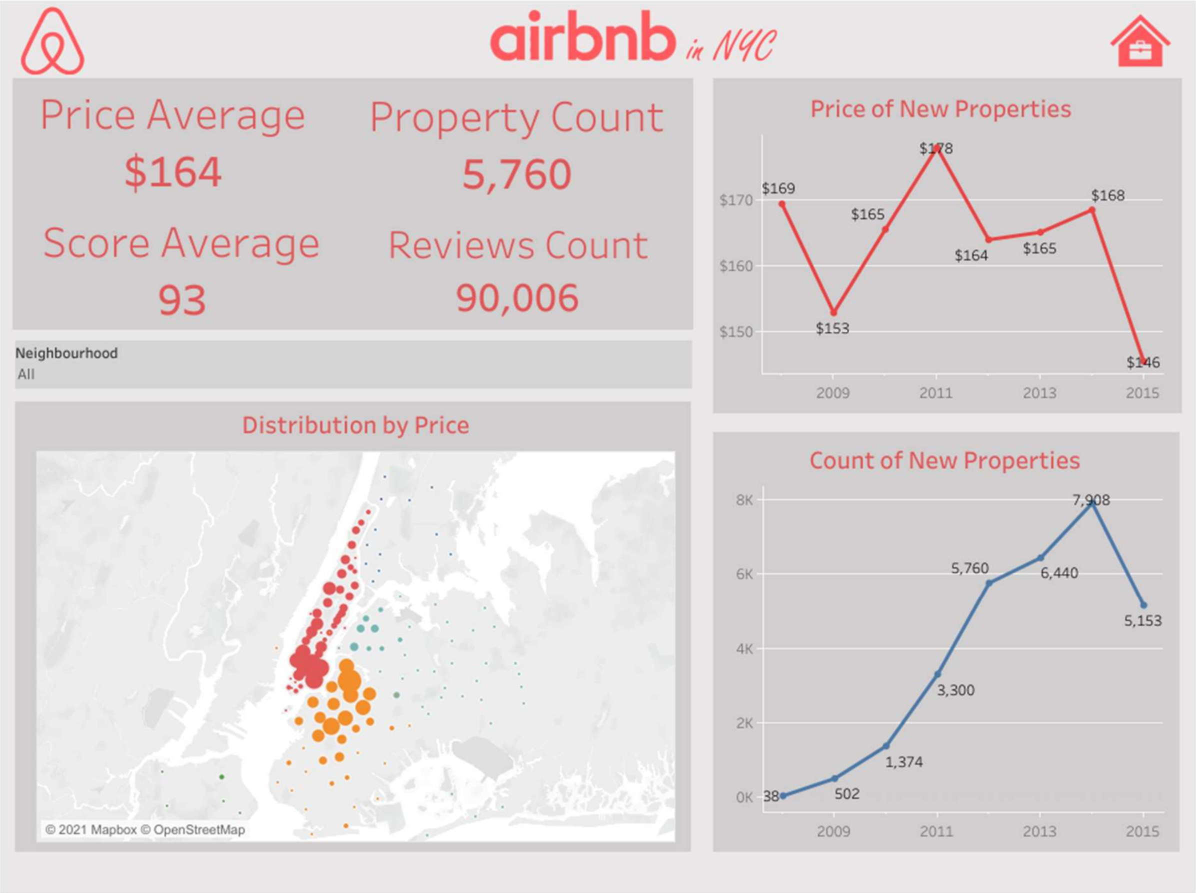
© 2021 Mapbox © OpenStreetMap

Room Type

Entire home/a..	17,024
Private room	12,609
Shared room	845

Property Type

Apartment	27,102
House	2,090
Loft	753
Bed & Breakfast	180
Townhouse	136
Condominium	94
Other	47
Dorm	31



airbnb in NYC

Price Average
$164

Property Count
5,760

Score Average
93

Reviews Count
90,006

Neighbourhood
All

Distribution by Price

© 2021 Mapbox © OpenStreetMap

Price of New Properties

Count of New Properties

We can appreciate that the dashboard is very similar to the delivery from the previous week. We tried to give a structure playing with the color contrast and a corporate appearance. The main page is almost the same but in the second page the charts have been changed in order to avoid the dual chart and repeat some KPIs. The idea is to keep the most important KPIs fixed in both dashboards, so after applying filters they will change accordingly and the user doesn't have to go back to the previous dashboard to check the KPIs.