

New York City Airbnb Analysis

By: Kriti Srivastava

Using Mongo DB and MongoDB Atlas



Content

- Why New York Airbnb ?
- Process Overview
- Data Source
- Data Description
- Data Cleansing
- Import to MongoDB and MongoDB Atlas
- Data Analysis and Visualizations

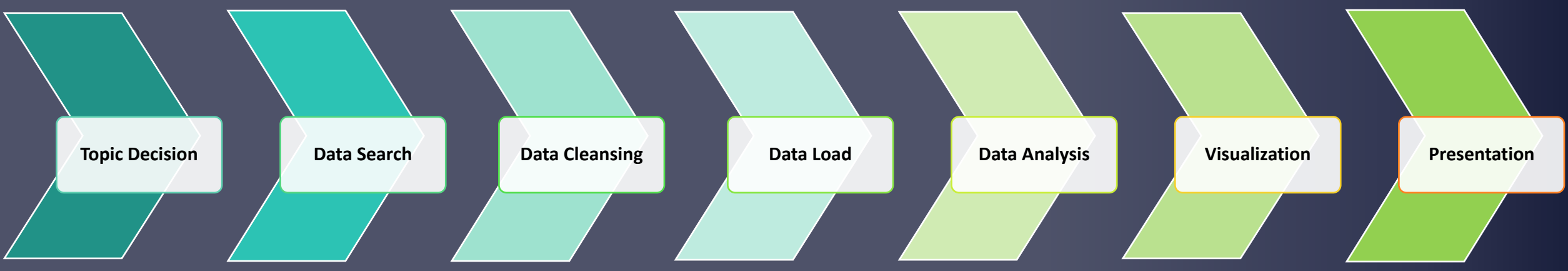


Why New York City Airbnb?

- Last summer me and my family visited New York City.
- At that time, we researched for the better Airbnb options with some questions in mind such as:
 - Which neighborhood has highest rating?
 - What are the pricing of the Airbnb in that neighborhood?
 - How are the availabilities?
 - The Airbnb options available near the attractions in the city etcetera.
- Therefore, I thought this is a great opportunity to use MongoDB to perform such kind of research and analysis and decided to work on New York City Airbnb for my project.



Process Overview



Topic Decision

Data Search

Data Cleansing

Data Load

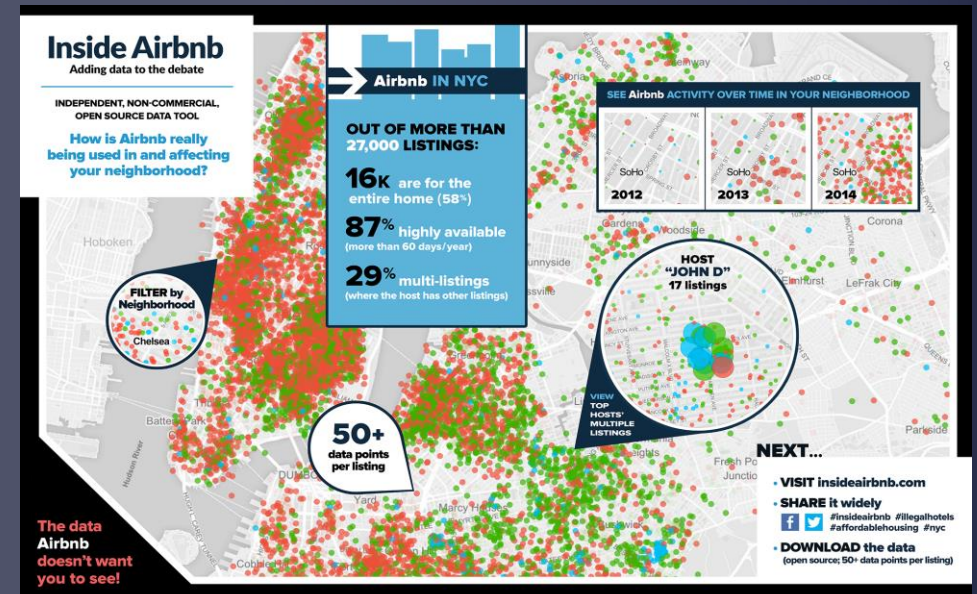
Data Analysis

Visualization

Presentation

Data Source

- The dataset describes the listing on Airbnb in NYC, NY for 2019.
- The direct source of the dataset is <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data> however the original source of the dataset is **Inside Airbnb**.
- The data is sourced from the **Inside Airbnb** (<http://insideairbnb.com/get-the-data.html>) which is sourced from publicly available information from the Airbnb site.
- According to the Inside Airbnb webpage *"Inside Airbnb is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world."*
- The data at Inside Airbnb has been already cleansed and aggregated where appropriate to facilitate public use.



Data Description

- This dataset describes the listing activity and metrics in NYC, NY for 2019.
- The dataset is consist of one table (collection)..
- Number of documents: **48,895**
- Although MongoDB do not have any concept of columns , however to explain the dataset, it has **16** columns

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_review	last_review	reviews_per_month	calculated_host_listings_count	availability_365
1	2595	Skylit Midtown	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home or apt	225	1	45	2019-05-21	0.38	2	355
2	2539	Clean & quiet	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
3	5099	Large Cozy	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Entire home or apt	200	3	74	2019-06-22	0.59	1	129
4	5022	Entire Apt: S...	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home or apt	80	10	9	2018-11-19	0.1	1	0
5	5121	BlissArtsSpa...	7356	Garon	Brooklyn	Bedford-Stu...	40.68688	-73.95596	Private room	60	45	49	2017-10-05	0.4	1	0

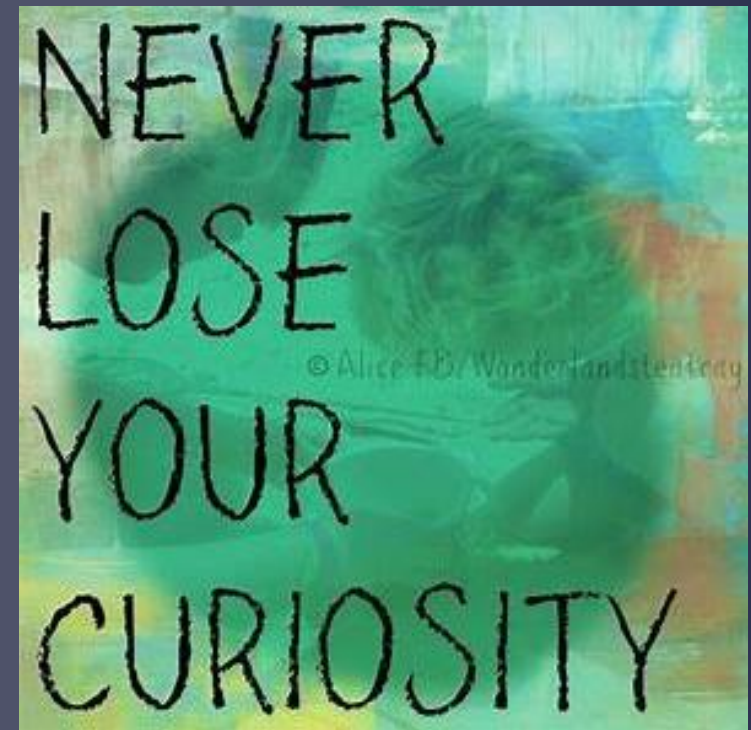
Data Columns Description

- **id** (Integer): Listing ID to uniquely identify each listed property.
- **name** (string) : Name of the listing.
- **host_id** (integer) : Host ID to uniquely identify host.
- **host_name** (string) : Name of the host of the listed property.
- **neighbourhood_group** (string) : Neighborhood group (Bronx, Brooklyn, Manhattan, Queens, Staten Island, etc)
- **neighbourhood** (string): Neighborhood of listed property.
- **latitude** (decimal) : Geospatial coordinate
- **longitude** (decimal) : Geospatial coordinate
- **roomtype** (string): Room type
- **price** (decimal): Price of the listing in dollars per night.
- **minimum_night** (integer): number of nights minimum required for booking.
- **number_of_review** (integer): number of reviews made on the listed property.
- **last_review** (data) : Date when last review made.
- **reviews_per_month** (integer): number of reviews made on the listed property per month.
- **calculated_host_listings_count** (integer) : count of the bookings that has been hosted by the property.
- **availability_365 (integer)**: number of days in the year the listing is available Airbnb booking.

Data Cleansing

Out of curiosity: I converted the .csv file to .json file by using free converting tools available online.

By converting the .csv file to .json file, it automatically solved the problem of null values in the collection.



Import to MongoDB and MongoDB Atlas

- Uploaded the file from MongoDB shell locally to perform aggregations and analysis.
- Also uploaded file on MongoDB Atlas from my local shell for creating visualizations using MongoDB Charts.

Command to import file locally

```
C:\MongoDB>mongoimport -d air_bnb_ny -c project C:\IPD351\csvjson_edit.json --jsonArray
2020-05-07T08:30:53.811-0500 connected to: mongodb://localhost/
2020-05-07T08:30:56.199-0500 48895 document(s) imported successfully. 0 document(s) failed to import.

C:\MongoDB>
```

Command to import file on MongoDB Atlas

```
mongoimport --uri "mongodb://root:kriti1234@mongodb+srv://testcluster-4c0q9.mongodb.net/test:27017/finalProject?ssl=true&replicaSet=myAtlasRS&authSource=admin" --collection air_bnb_ny --drop --file C:\IPD351\csvjson_edit.json --jsonArray
```

Created Indexes

```
db.project.ensureIndex({neighbourhood_group: 1,  
neighbourhood: 1, price: 1})
```

```
db.project.ensureIndex({neighbourhood_group: 1,  
neighbourhood: 1, price: -1})
```

```
db.project.ensureIndex({neighbourhood_group: 1, price: -1})
```

```
db.project.ensureIndex({neighbourhood_group: 1, price: 1})
```

```
db.project.ensureIndex({neighbourhood_group: 1,  
neighbourhood: 1, room_type: 1, price: 1})
```

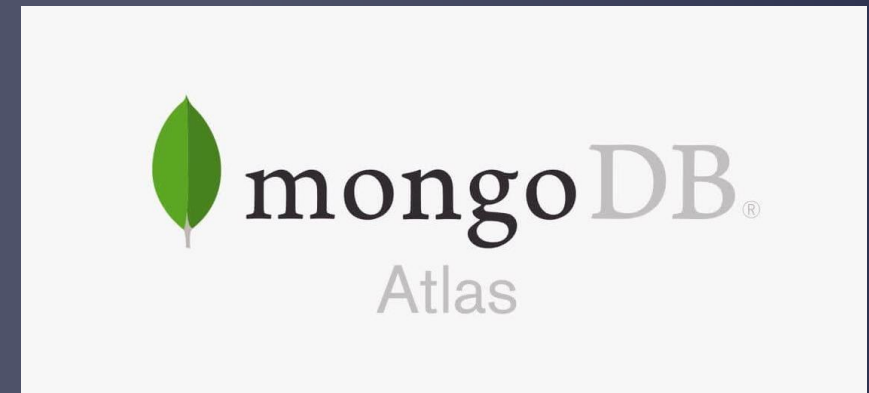
```
db.project.ensureIndex({neighbourhood_group: 1,  
room_type: 1, price: 1})
```

```
db.project.getIndexes()
```

```
/* 1 */  
[  
  {  
    "v" : 2,  
    "key" : {  
      "_id" : 1  
    },  
    "name" : "_id_",  
    "ns" : "air_bnb_ny.project"  
  },  
  {  
    "v" : 2,  
    "key" : {  
      "neighbourhood_group" : 1.0,  
      "neighbourhood" : 1.0,  
      "price" : 1.0  
    },  
    "name" : "neighbourhood_group_1_neighbourhood_1_price_1",  
    "ns" : "air_bnb_ny.project"  
  },  
  {  
    "v" : 2,  
    "key" : {  
      "neighbourhood_group" : 1.0,  
      "neighbourhood" : 1.0,  
      "price" : -1.0  
    },  
    "name" : "neighbourhood_group_1_neighbourhood_1_price_-1",  
    "ns" : "air_bnb_ny.project"  
  },  
  {  
    "v" : 2,  
    "key" : {  
      "neighbourhood_group" : 1.0,  
      "neighbourhood" : 1.0,  
      "room_type" : 1.0,  
      "price" : 1.0  
    },  
    "name" : "neighbourhood_group_1_neighbourhood_1_room_type_1_price_1",  
    "ns" : "air_bnb_ny.project"  
  },  
  {  
    "v" : 2,  
    "key" : {  
      "neighbourhood_group" : 1.0,  
      "room_type" : 1.0,  
      "price" : 1.0  
    },  
    "name" : "neighbourhood_group_1_room_type_1_price_1",  
    "ns" : "air_bnb_ny.project"  
  }  
]
```

Data Analysis and Visualizations

- Used MongoDB for aggregations and data analysis
- Used Mongo charts for creating charts and visualization.
- The Mongo chart is available for free in the MongoDB Atlas.



Rechecking for the missing values

No null values in the data set at this stage.

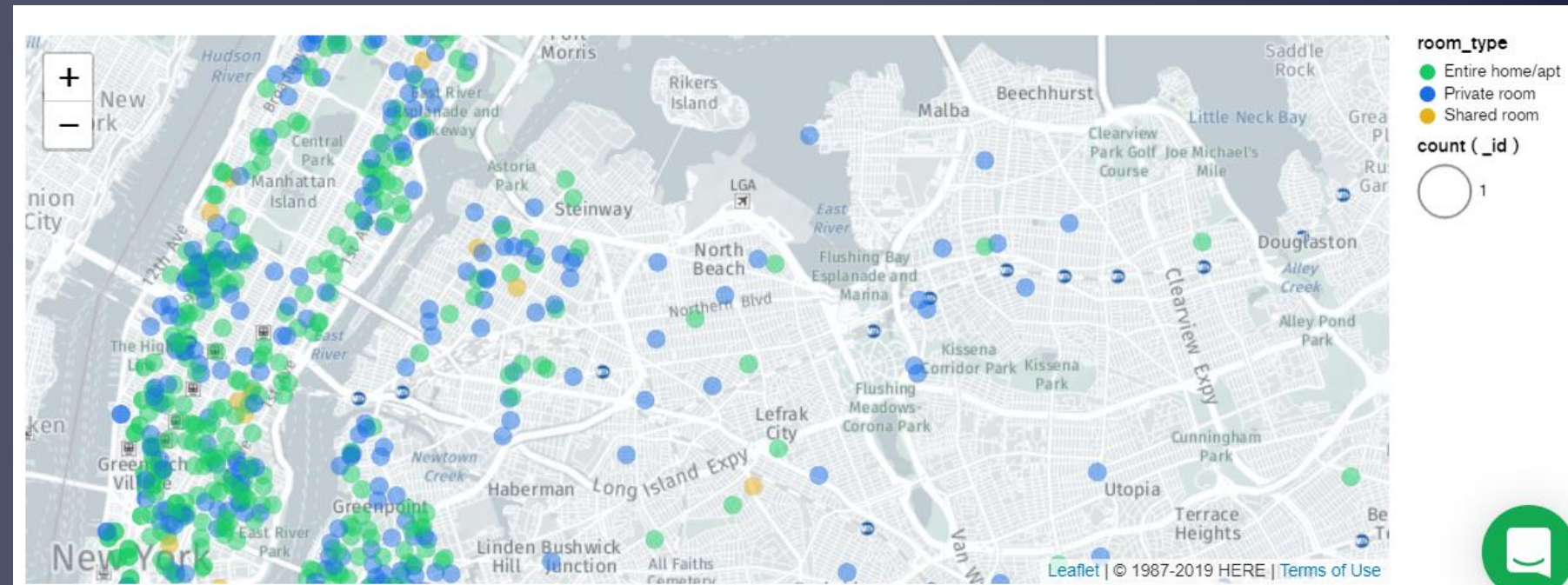
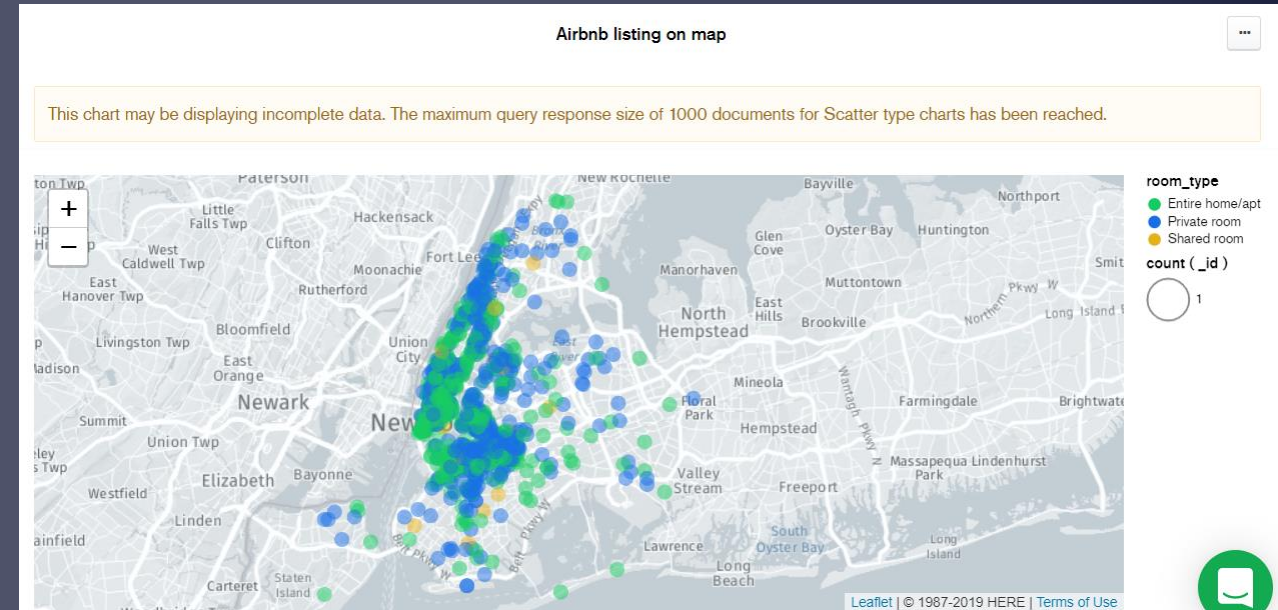
To me safe, I will also remove the nulls and missing values while writing the queries.

```
db.project.find({id : null}).count()
db.project.find({name :null}).count()
db.project.find({host_id : null}).count()
db.project.find({hostname : null}).count()
db.project.find({neighbourhood_group null }).count()
db.project.find({neighbourhood : null }).count()
db.project.find({latitude null}).count()
db.project.find({longitude : null}).count()
db.project.find({roomtype : null}).count()
db.project.find({price : null}).count()
db.project.find({minimum_nights : null}).count()
db.project.find({number_of_reviews : null}).count()
db.project.find({last_review : null}).count()
db.project.find({calculated_host_listings_count : null}).count()
db.project.find({availability_365 : null}).count()
```



Airbnb Listings by Room Type

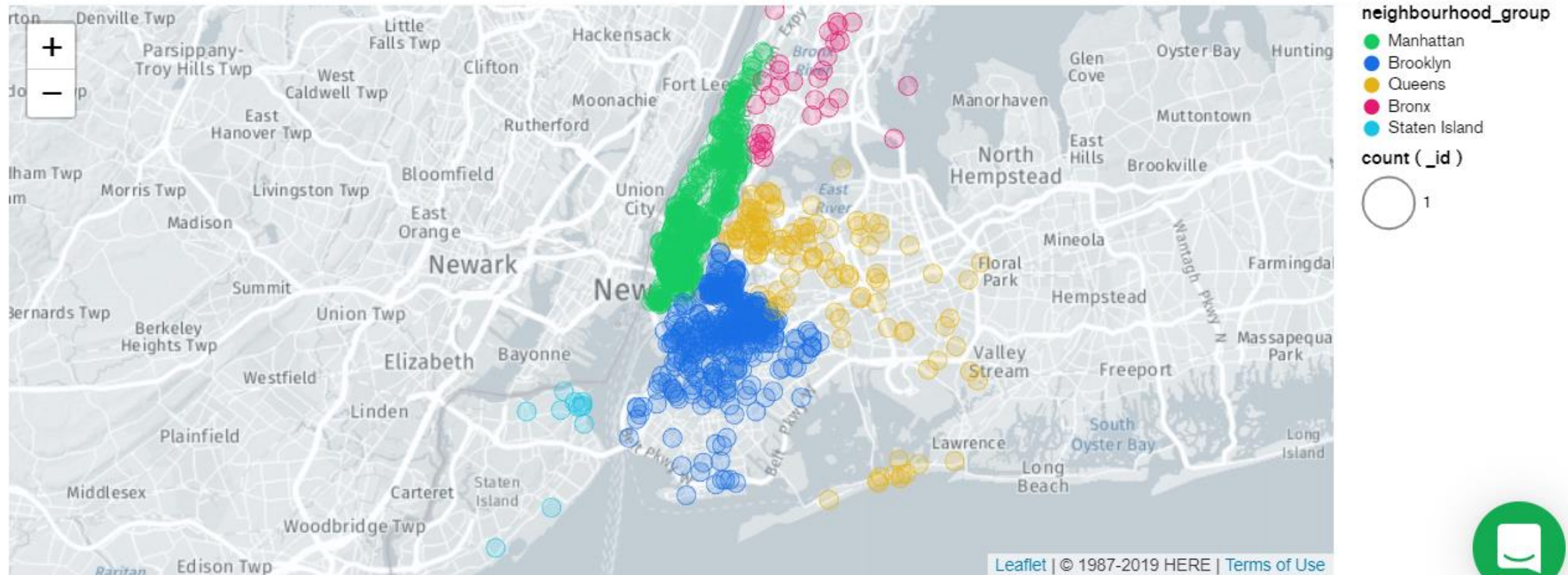
- Every point on a map shows an Airbnb listing.
- The points are colored as per the room type.
- The map is generated using Geospatial feature in the MongoDB chart that is available on MongoDB Atlas.
- The map initially shows up to 1000 documents, however as we zoom in, the map shows more documents in the selected area.



Airbnb listings on map by Neighborhood Group

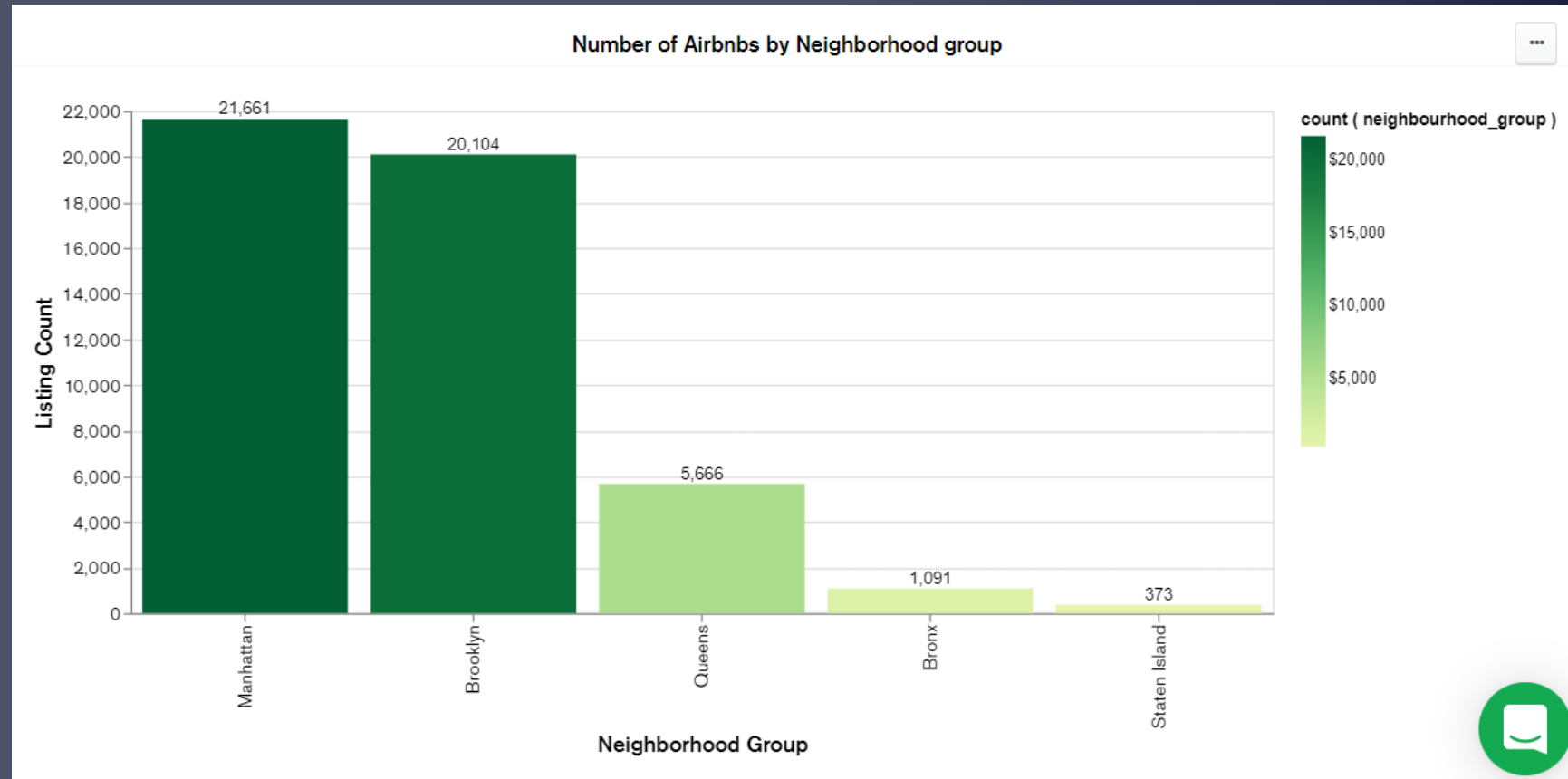
Airbnb listing on map

This chart may be displaying incomplete data. The maximum query response size of 1000 documents for Scatter type charts has been reached.



Airbnb listings by Neighborhood Group

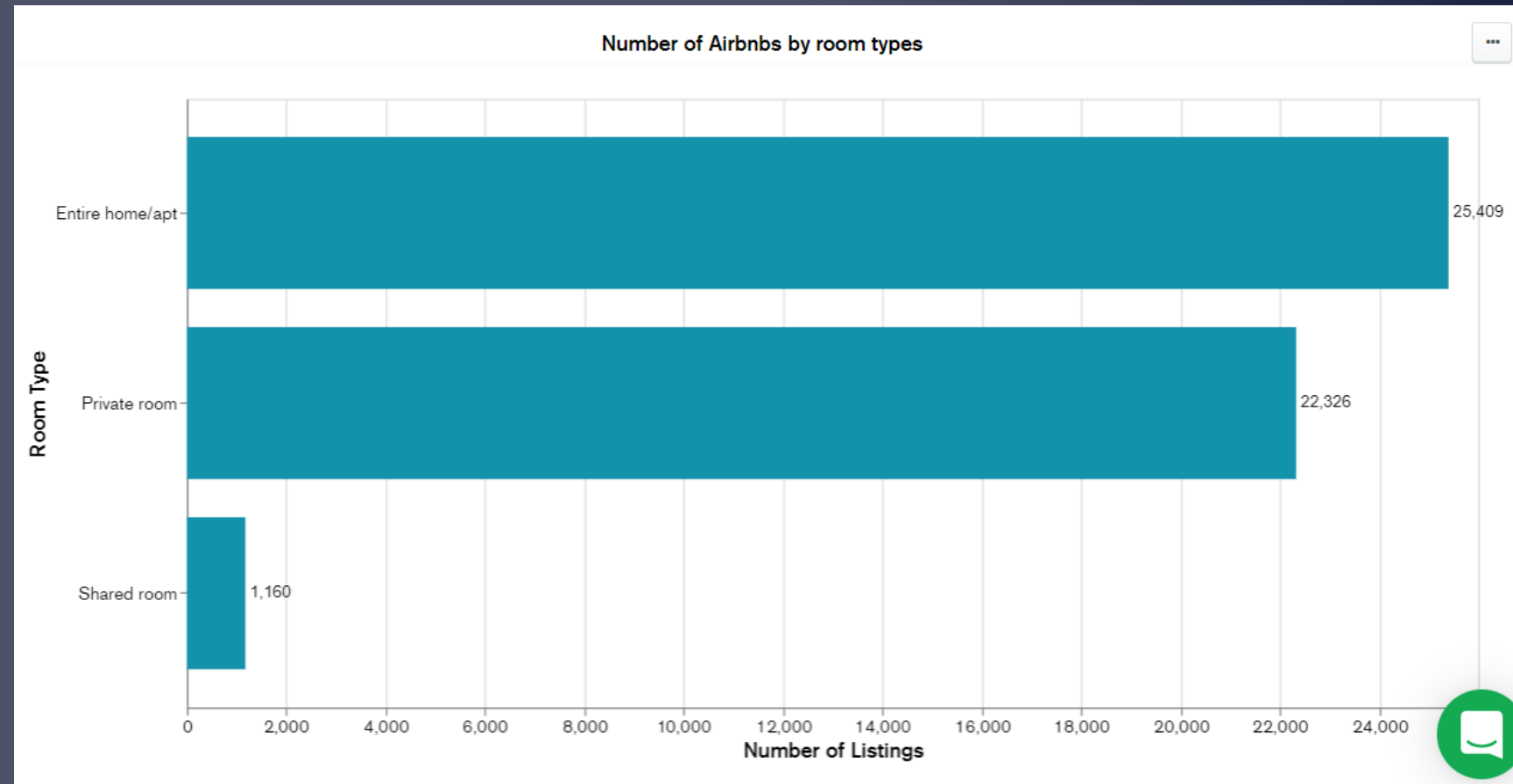
- Manhattan has largest number of Airbnb listing.
- Manhattan had 21,661 listings on Airbnb.
- The Brooklyn came next close to Manhattan with total count of 20,104 listings
- Queens, Bronx and Staten Island is way behind from Manhattan and Brooklyn.
- Then comes Queens with 5,666 listings, Bronx with 1,091.
- Staten Island has least number of listings with just 373.
- **Manhattan and Brooklyn has a lot of options to choose from.**



```
db.project.aggregate ([
  {$group: {"_id": "$neighbourhood_group", "Count" :{ $sum: 1}}}
])
```

Airbnb listings by Room Type

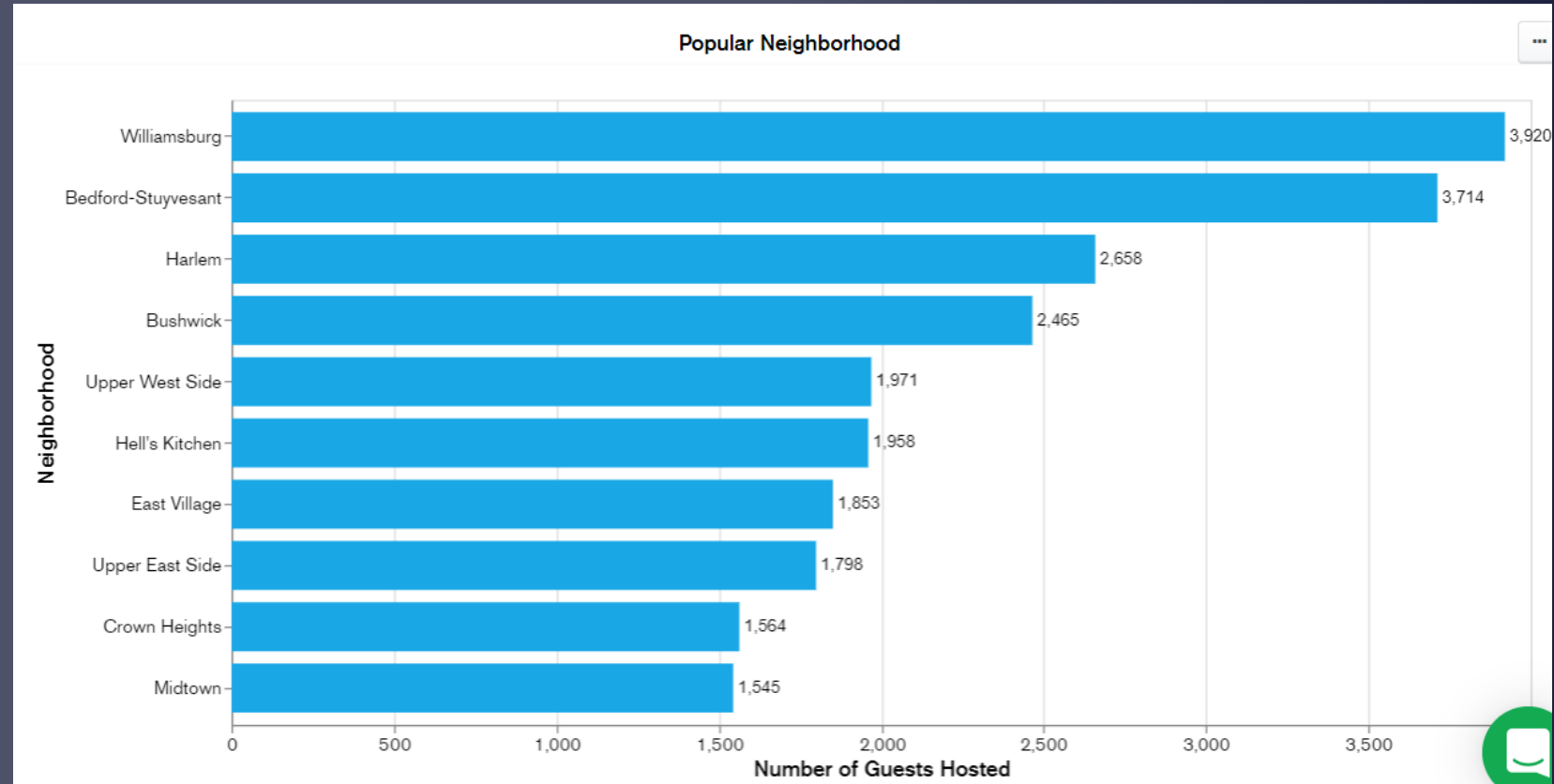
- Airbnb New York in 2019 offered 25,409 listings at the Entire home or Apartment.
- 22,326 private rooms.
- 1,160 shared room. Which is very low as compare to that of private rooms and entire home or apartments.



```
db.project.aggregate ([
  {$group: {"_id": "$neighbourhood_group", "Count" :{ $sum: 1}}}
])
```

Most Popular Neighborhood

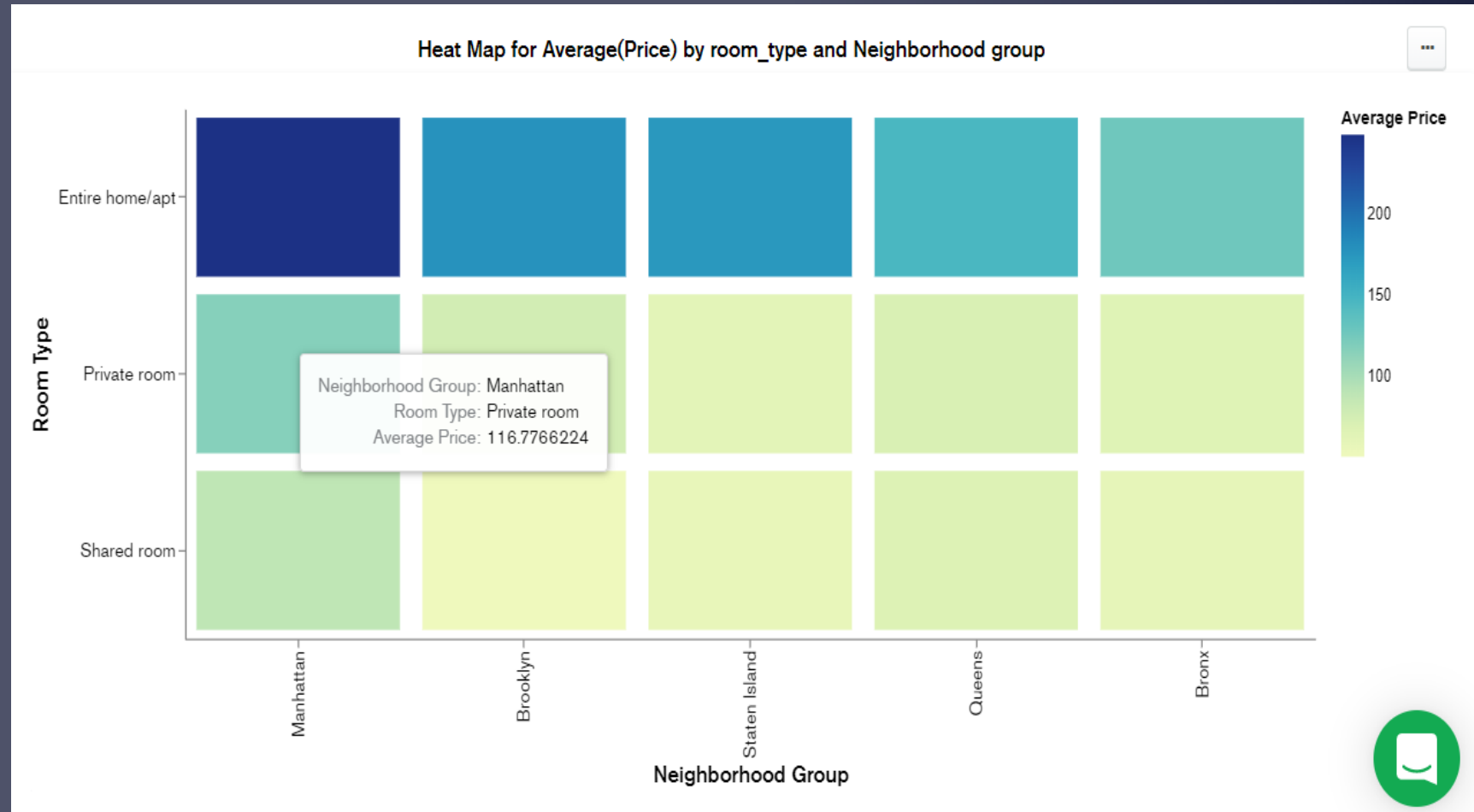
- These are the top 10 most popular neighborhoods.
- Williamsburg from Brooklyn is the most popular neighborhood.
- The popularity has been calculated by the count of the number of bookings hosted by the listed property.



```
db.project.aggregate ([
  {$group:{"_id":{"neighbourhood":"$neighbourhood"},"Num_hosting":{"$sum":{"$cond":[{"$ne":[{"$type":"$calculated_host_listings_count"},"missing"]}],1,0}}}},
  {$sort:{Num_hosting : -1}},
  {$project:{"_id": 0,neighbourhood: "$_id.neighbourhood",Num_hosting: 1}}])
```

Average Room Types Prices in all Neighborhood Group

- Obviously, entire home or apartment in Manhattan is most expensive.
- The average price of the entire room or apartment in Manhattan is almost \$250 per night.
- Shared rooms in Brooklyn are least expensive.
- By looking at the heat, we can see that average price of the private rooms per night is under \$150.
- Average pricing of shared rooms are under \$100 per night.
- Average pricing of Entire home/apt in Queens and Bronx are under \$150 per night.



```
db.project.aggregate([
  {$match: {"price": {$gt: 0 }, "neighbourhood_group": { $nin: [ null, "" ]}, "room_type": { $nin: [ null, "" ]}} ,
  {$group: {"_id" :{ neighbourhood_group:"$neighbourhood_group", room_type :"$room_type" }, "average_price": {$avg : "$price" } }},
  {$sort :{average_price : -1}},
  {$project :{ "_id": 0, neighbourhood_group: "$_id.neighbourhood_group",room_type : "$_id.room_type" , average_price : 1}} ])
```

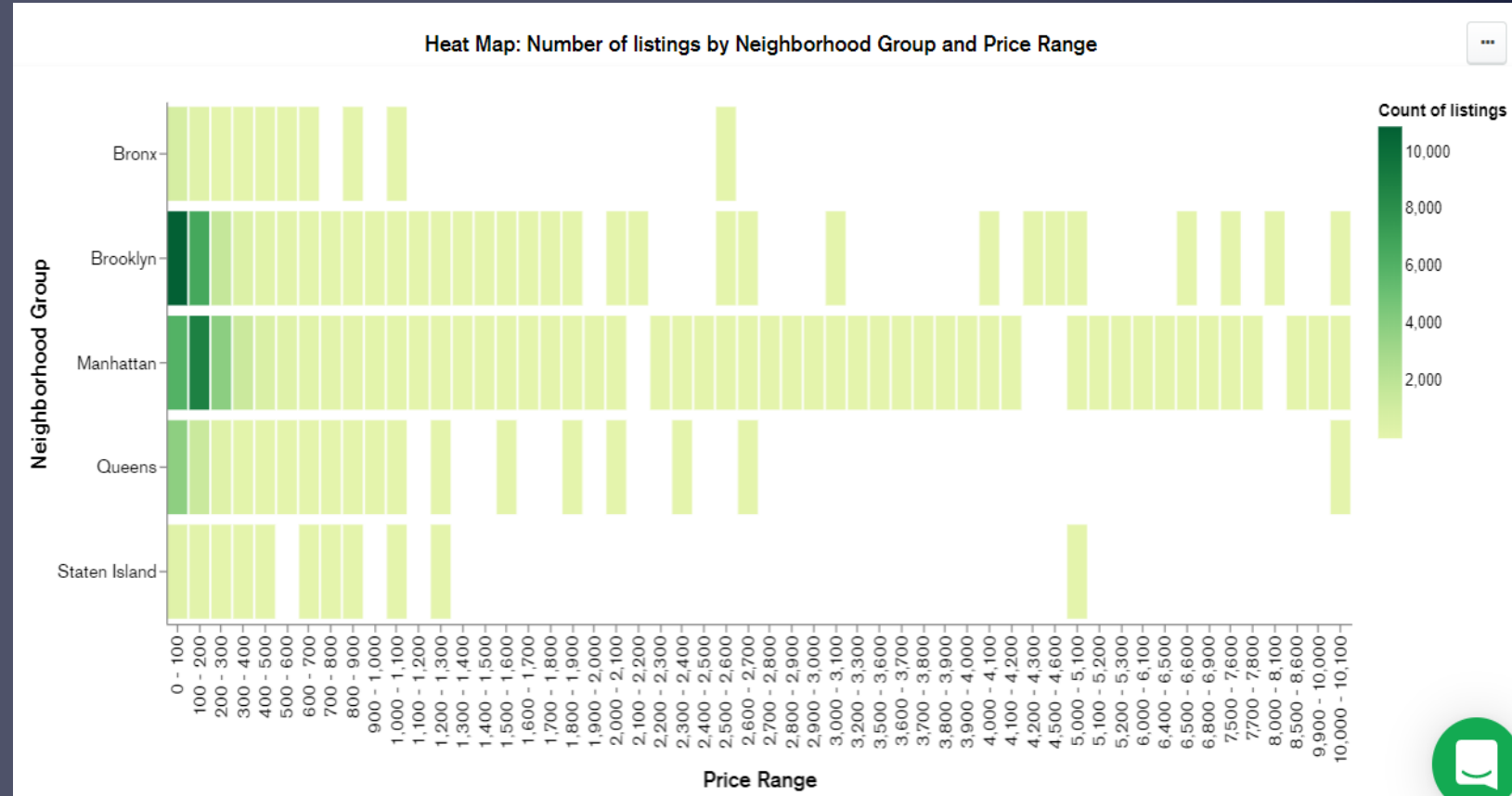
Average Room Types Prices in all Neighborhood Group

- Result of the same query in table format visual.
- Now we can see exactly what is the average price of the listings.
- Here we can see the Private room in Manhattan has the average price of \$116.78 per night, as shown in heat map.
- Entire home/ apartment in Manhattan is most expensive with the average price of \$ 249.58 per night.
- Shared room is Brooklyn is least expensive with the average price of \$ 50.77per night.

neighbourhood_group	room_type	average_price ↑
Manhattan	Entire home/apt	249.25799363539932
Brooklyn	Entire home/apt	178.34620213433772
Staten Island	Entire home/apt	173.8465909090909
Queens	Entire home/apt	147.05057251908397
Bronx	Entire home/apt	127.5065963060686
Manhattan	Private room	116.7766224004009
Manhattan	Shared room	88.97708333333334
Brooklyn	Private room	76.54542761208769
Queens	Private room	71.76245551601423
Queens	Shared room	69.02020202020202
Bronx	Private room	66.89093701996927
Staten Island	Private room	62.29255319148936
Bronx	Shared room	59.8
Staten Island	Shared room	57.44444444444444
Brooklyn	Shared room	50.77372262773723

Listings in Neighborhood Group by Price Bucket

- This chart shows many information about the price range of listings in New York City.
- Manhattan and Brooklyn have listings up to \$10,100 per night.
- However, Brooklyn alone has more that 10,000 listings that are under \$100 per night. Near about 7,000 listing that are between \$100 to \$200 per night. Brooklyn properties are least expensive.
- Manhattan has nearly 8,000 listings between \$100 to \$200 per night. Near about 6,000 listings under \$100 per night.
- Rooms in Bronx & Staten Island typically range up to \$1300/night barring a few in higher range.



```
db.project.aggregate([
  {$match: {price: {$gte: 1 }, neighbourhood_group: {$nin: [null, "" ]}}},
  {$addFields: {price: {$cond: {if: {$in: [{ $type: "$price" },["double", "int","long","decimal"]]}, then: "$price", "else": null }}}},
  {$addFields: {price_limit1: {$multiply: [{ $trunc: { $divide: ["$price", 100]}}, 100]}}},
  {$addFields: {price_limit2string: {$convert: { input: {$add: ["$price_limit1", 100]}, to: "string"}}}},
  {$addFields: {price_limit1string: {$convert: { input: "$price_limit1", to: "string"}} } },
  {$addFields: {price_range: {$concat: ["$price_limit1string","-","$price_limit2string"]}}},
  {$group: { "_id": { price_range: "$price_range", neighbourhood_group: "$neighbourhood_group"}, "count": {$sum: {$cond: [{ $ne: [{ $type: "$_id" }, "missing"]}, 1, 0 ]}}}},
  {$sort: { _id : 1}}, {$project: { "_id": 0, price_range: "$_id.price_range", neighbourhood_group: "$_id.neighbourhood_group", "count": 1}} ])
```

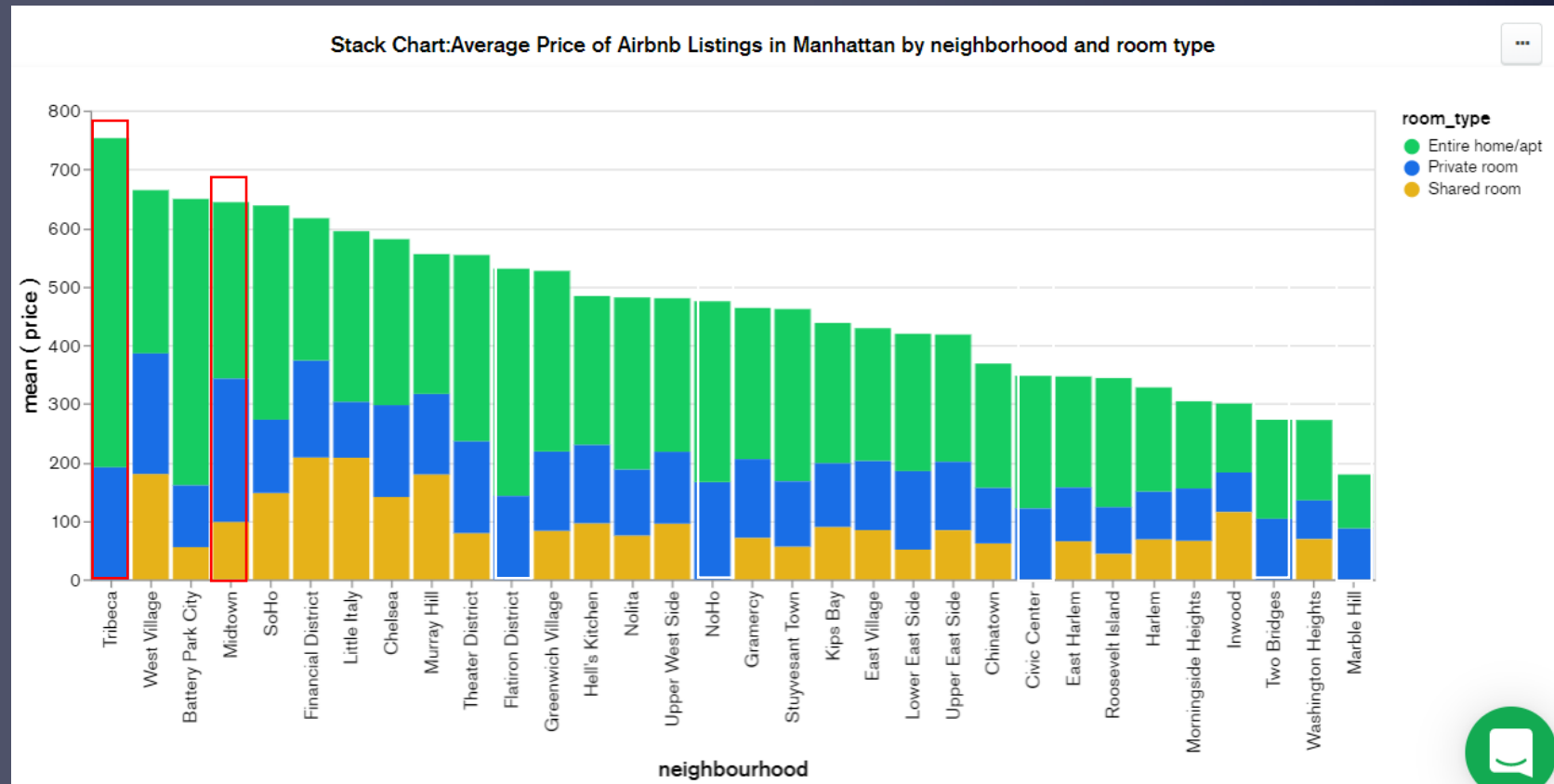

Listings in Neighborhood Group by Price Bucket

- Result of the same query in tabular format giving the count of rooms in different price ranges.
- This the snapshot of the top few rows of the table.
- Now we can see exactly what is the average price of the listings.

Count of listings					
Price Range	Bronx	Brooklyn	Manhattan	Queens	Staten Island
0 - 100	822	10,895	6,035	3,865	249
100 - 200	214	6,667	8,869	1,390	93
200 - 300	26	1,620	3,921	280	20
300 - 400	14	462	1,383	68	3
400 - 500	5	172	547	27	2
500 - 600	2	101	264	9	
600 - 700	4	44	167	7	1
700 - 800		31	126	4	1

Average price by Neighborhoods in Manhattan

- I focused my research to Manhattan.
- I wanted to see the average pricing of listings in different neighborhoods of Manhattan.
- The stack chart that shows the average pricing of listings of all the three room types by neighborhoods.
- Tribeca, NoHo, Flatiron District, Civic Center, Two Bridges, and Marble Hill don't even have any listings with Shared room. Interesting!!
- Tribeca has most expensive Entire home/apartment listings (\$561.8) however shared rooms have decent pricing (\$191.2).
- Whereas, Midtown has expensive shared room listings (\$244.4).



```
db.project.aggregate([
  {$match: {"neighbourhood_group": {"$in": ["Manhattan"]}, "price": {"$gte": 1 }}},
  {$group: {"_id": {"neighbourhood": "$neighbourhood", "room_type": "$room_type"},
    "average_price": {$avg: "$price"}}},
  {$sort: {"average_price": -1}},
  {$project: {"_id": 0, neighbourhood: "$_id.neighbourhood", room_type:
    "$_id.room_type", average_price: 1 } } ])
```

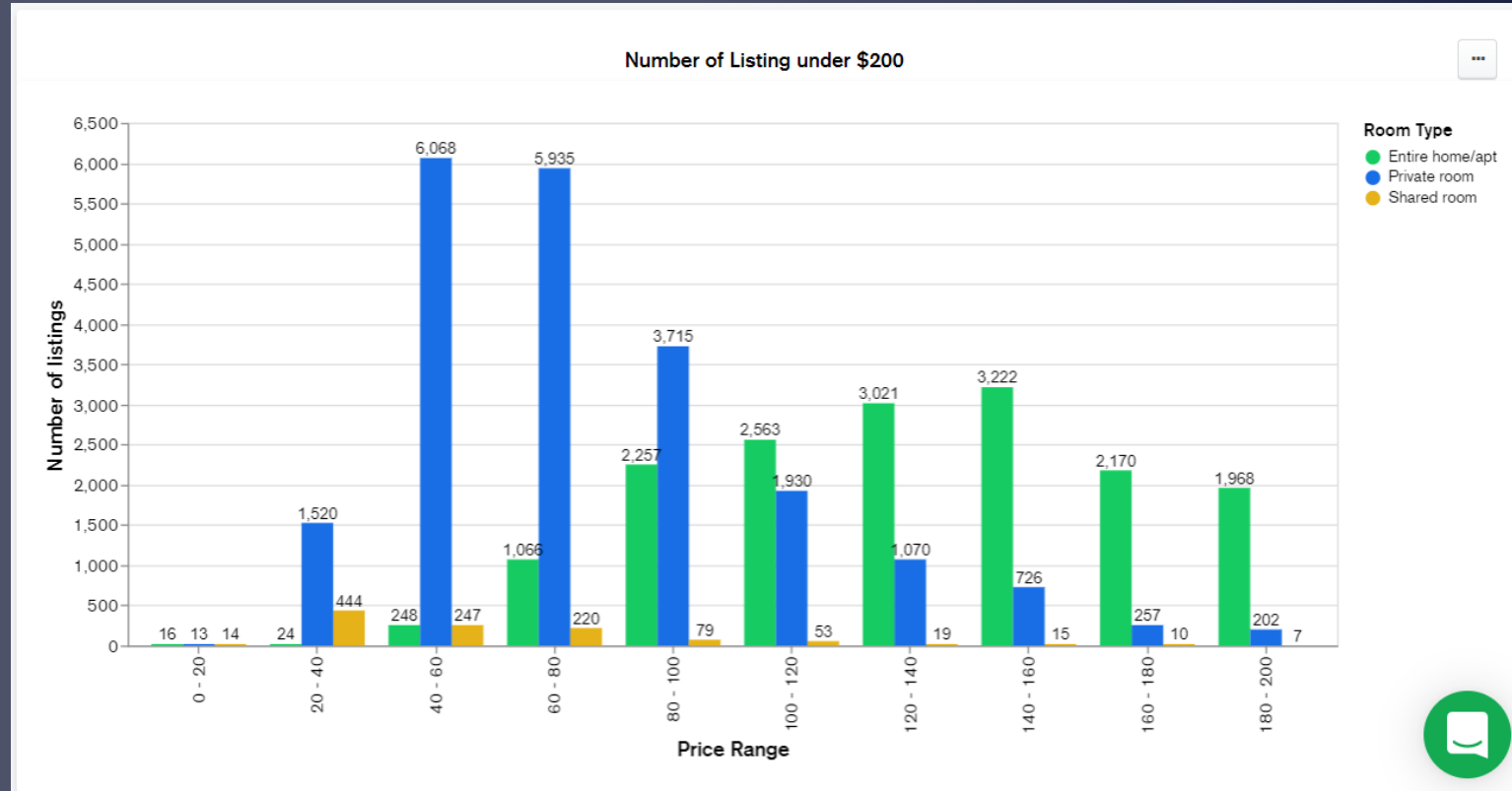
For Mid-Budget Travelers

(Private rooms under \$200 per night)



Listings under \$200

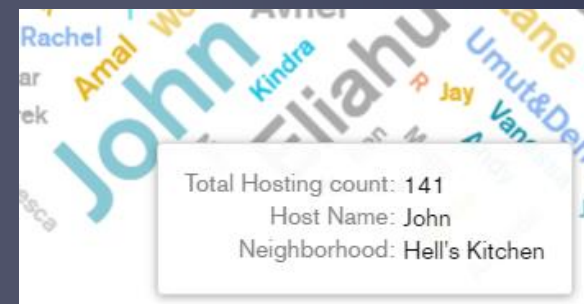
- There are few rooms in the 0-\$20 range which is surprising.
- The high density price range is between \$40-\$100 private rooms.
- There are very few shared room available which again is surprising.
- Shared rooms are for highly budget conscious travelers and they have few rooms between \$20-\$80 range.



```
db.project.aggregate ([
  {$match: {room_type: {$nin: [null,""]}, price: {$gte: 1, $lt: 200 }}},
  {$addFields: {price: {$cond: {if: {$in: [{$type: "$price"},["double","int","long","decimal"]]},then: "$price","else": null}}}},
  {$addFields: {price_limit1: {$multiply: [{$trunc: {$divide: ["$price",20]}},20]}},
  {$addFields: {price_limit2string: {$convert: { input: {$add: ["$price_limit1", 20]}, to: "string"}}}},
  {$addFields: {price_limit1string: {$convert: { input: "$price_limit1", to: "string"}}}},
  {$addFields: {price_limit: {$concat: ["$price_limit1string","-","$price_limit2string"]}},
  {$group: {"_id": {price_range: "$price_limit",room_type: "$room_type"},total_listing: {$sum: {$cond: [{$ne: [{$type: "$id"},"missing"]}, 1,0 ]}}}},
  {$sort:{_id : 1}},
  {$project: {"_id": 0,price_range: "$_id.price_range",room_type: "$_id.room_type",total_listing: 1 }} ])
```

Most Popular Hosts in Manhattan

- These are the 100 most popular hosts. The popularity is calculated by the number of guests they have hosted in 2019.
- The host's name is colored as per their neighborhood which can be viewed on mouse-over action.
- These are the hosts for the properties with private rooms that are under \$200 per night.
- The same word cloud can be created for other locations such as Brooklyn.



```
db.project.aggregate([
  {$match: {neighbourhood_group: {$in: ["Manhattan"]},availability_365: 365, room_type: {$in: ["Private room" ]}, price: {$gte: 10, $lte: 200 }},
  {$group: {"_id": {host_name: "$host_name", neighbourhood: "$neighbourhood"}, total_hosting_count: {$sum: "$calculated_host_listings_count"}}} ,
  {$sort:{total_hosting_count : -1}},
  {$project: {"_id": 0, host_name: "$_id.host_name", neighbourhood: "$_id.neighbourhood", total_hosting_count: 1}},
  {$limit: 100})
```

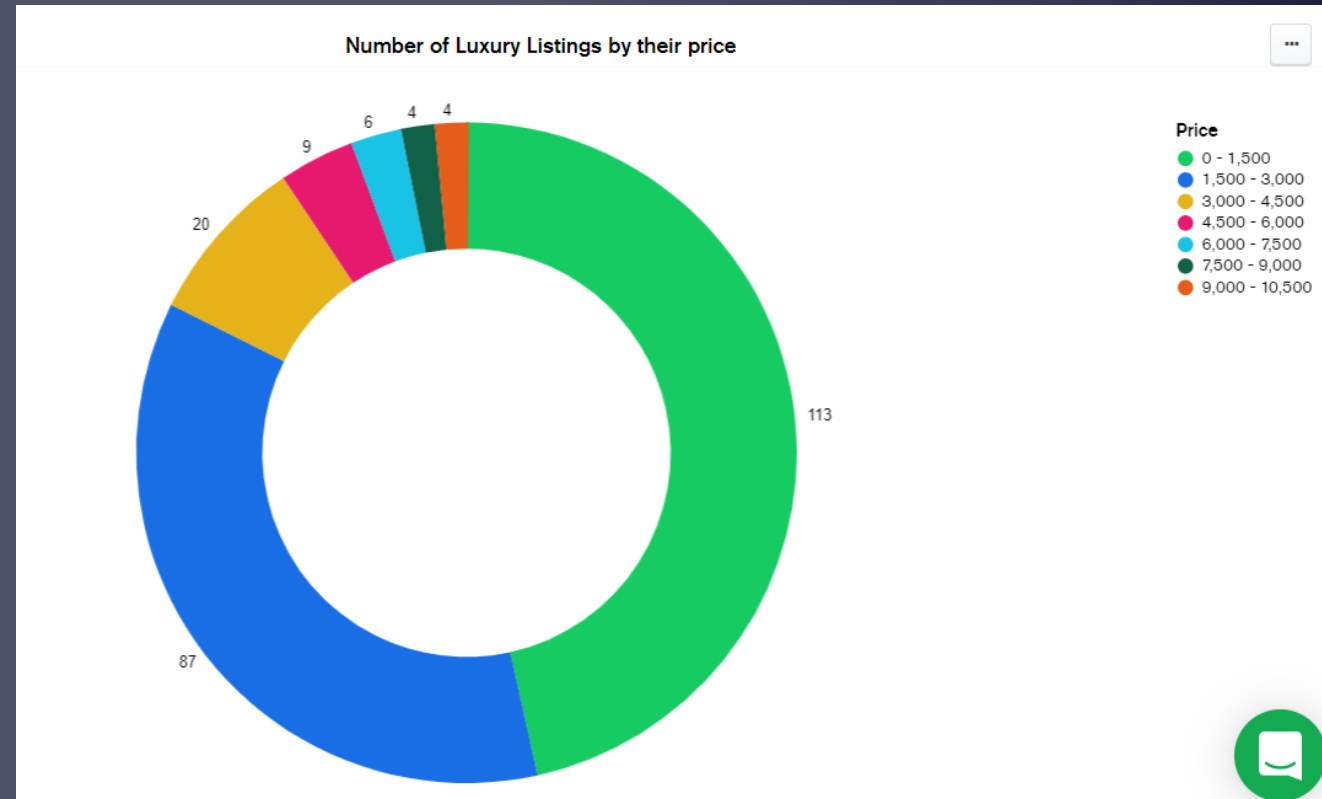

For Luxurious Segment

(Entire home/ Apartments with \$1000 and above per night)



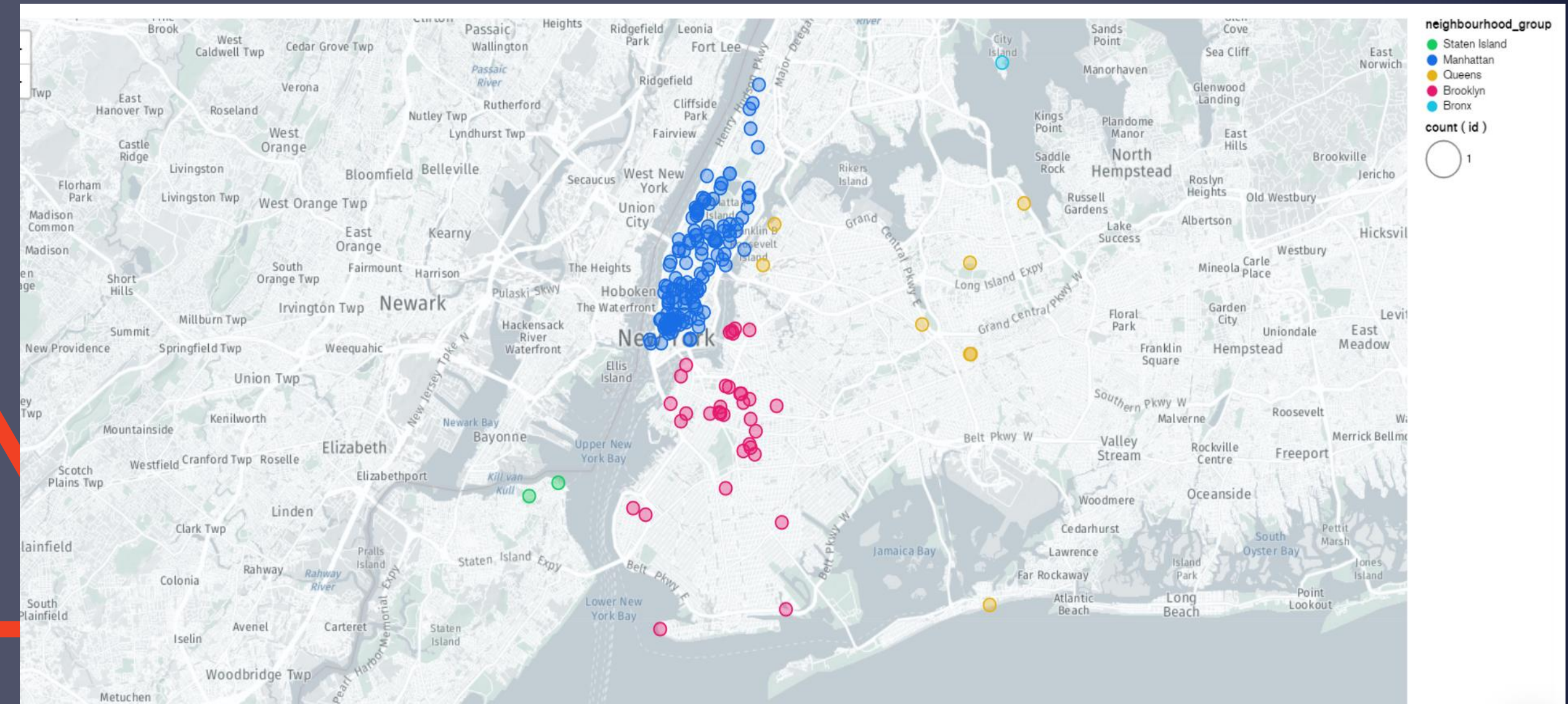
Number of Luxurious Listings by Price

- The donut chart shows the listing that are available in a given price range.
- The top end segment has 14 properties if we look at prices over \$6000/night.
- Travelers have abundant options between \$1000-\$3000 range.



```
db.project.aggregate ([
  {$match: {room_type: {$in: ["Entire home or apt"]},price: {"$gte": 1000 }}} ,
  {$addFields: {price: {$cond : {if: {$in: [ {$type: "$price"},["double","int","long","decimal"]]},then: "$price","else": null}}}} ,
  {$addFields: {price_limit1: {$multiply: [{$trunc: {$divide: ["$price",1500]}},1500]}},
  {$addFields: {price_limit2string: {$convert: { input: {$add: ["$price_limit1", 1500]}, to: "string"}}}},
  {$addFields: {price_limit1string: {$convert: { input: "$price_limit1", to: "string"}} } ,
  {$addFields: {price_limit: {$concat: ["$price_limit1string","-","$price_limit2string"]}},
  {$group: {"_id": {price_limit: "$price_limit"},total_listings: {$sum: {$cond: [{"$ne": [{"$type": "$_id"},"missing" ]},1,0 ]}}},
  {$sort: {_id : 1}},{$project: {"_id": 0, price_limit: "$_id.price_limit", total_listings: 1}} ] )
```


Location of Luxurious Listings



Conclusion

1. I prefer to book the complete apartment as I plan my travel with family or group of friends. Manhattan will be my choice of the location as it provides good options within and over \$1000/night range.
2. I will also look at top hosts to decide which apartment I will be booking as the overall experience matters for me the most.
3. The data also gives me insights about the availability of some of the property so I will try to book it in advance.

Below points are form AirBnB perspective -

1. AirBnB can add a new 'luxury' segment to their offering to attract high worth clients.
2. AirBnB could also do additional advertising for the hosts so that new hosts can gain more clients and existing hosts can promote their experiences and good reviews.



Thank you