Northeastern University
College of Professional Studies

April 03, 2023

Airbnb New York Analysis (2011 - 2019) :

Submitted to:
Prof. Fatemeh Ahmadi Abkenari

Submitted by:
Abhilash Dikshit
Milan Prajapati
Shamim Sherafati
Smit Parmar

Airbnb New York Analysis (2011 - 2019)

Airbnb New York Analysis (2011 - 2019)
Airbnb has been facilitating unique and tailored travel experiences for both guests and hosts since 2008. The present dataset provides comprehensive information on the metrics and listing activity specific to New York City in 2019.

This dataset contains information about Airbnb host listings, neighbourhoods, price, minimum nights, review counts/rate and the availability throughout the year.

Limitations:
No data on customer stay duration, timeline, review score for each listing and no. of tourist attractions nearby.

Acknowledgements:
Dataset taken from Kaggle. This is a public dataset of Airbnb, and the original source can be found in their website.

Dataframe cleanup:

- Sorted as per price.
- Removed blank cells.
- Resized the data frame to rows= 4890 and attributes= 16 by removing the top 55 % and bottom 45%.

```
[1] "Uncleaned Dataframe Info:"
Number of Rows: 48895
Number of Columns: 16
Blank cells count: 10052

[1] "Removed Blank cells:"
Number of Rows after removing blank cells: 38843
Number of Columns after removing blank cells: 16
Blank cells count after removal: 0

[1] "Resized dataframe:"
Number of Rows: 4890
Number of Columns: 16
Blank cells count in cleaned dataframe: 0
```

Airbnb New York Analysis (2011 - 2019)

Dataset Output

	id <dbl>	name <chr>	host_id <dbl>	host_name <chr>	neighbourhood_group <chr>
1	13157147	Spacious 1 BR Apartment w/ private backyard	2708284	Agustina	Brooklyn
2	13149690	Private Cozy Bedroom - Central Park North	72008788	Chris	Manhattan
3	13135822	Gorgeous Bedroom in Manhattan Midtown West	54454582	Breath Beyond	Manhattan
4888	6809332	Cozy, Warm Home in the West Village	198010	Shenaz	Manhattan
4889	6657454	Spacious & sunny in Clinton Hill :)	6913285	Marc	Brooklyn
4890	6638377	Lightfilled studio in Brooklyn	31889552	Chris	Brooklyn

6 rows | 1-6 of 16 columns

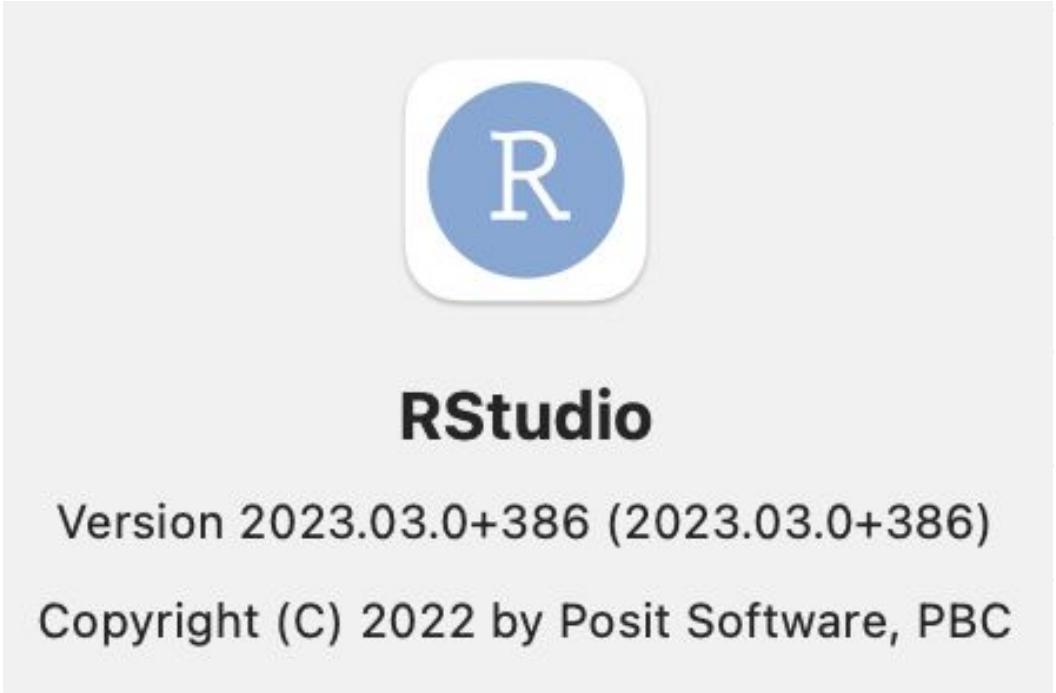
neighbourhood <chr>	latitude <dbl>	longitude <dbl>	room_type <chr>	price <dbl>	minimum_nights <dbl>	number_of_reviews <dbl>
Bedford-Stuyvesant	40.69	-73.95	Private room	75	6	37
Harlem	40.80	-73.95	Private room	75	3	107
Hell's Kitchen	40.76	-73.99	Private room	75	4	146
West Village	40.74	-74.01	Entire home/apt	95	5	4
Bedford-Stuyvesant	40.69	-73.96	Entire home/apt	95	5	4
Fort Greene	40.69	-73.97	Entire home/apt	95	3	1

6 rows | 7-13 of 16 columns

number_of_reviews <dbl>	last_review <chr>	reviews_per_month <dbl>	calculated_host_listings_count <dbl>	availability_365 <dbl>
37	2019-06-17	1.53	1	329
107	2019-06-13	2.86	1	279
146	2019-07-06	3.89	3	73
4	2016-02-21	0.09	1	0
4	2016-06-03	0.08	1	0
1	2015-07-27	0.02	1	0

6 rows | 13-17 of 16 columns

Approach 1: R Code



Bar chart showing the distribution of Room Types in New York City

Business Question:

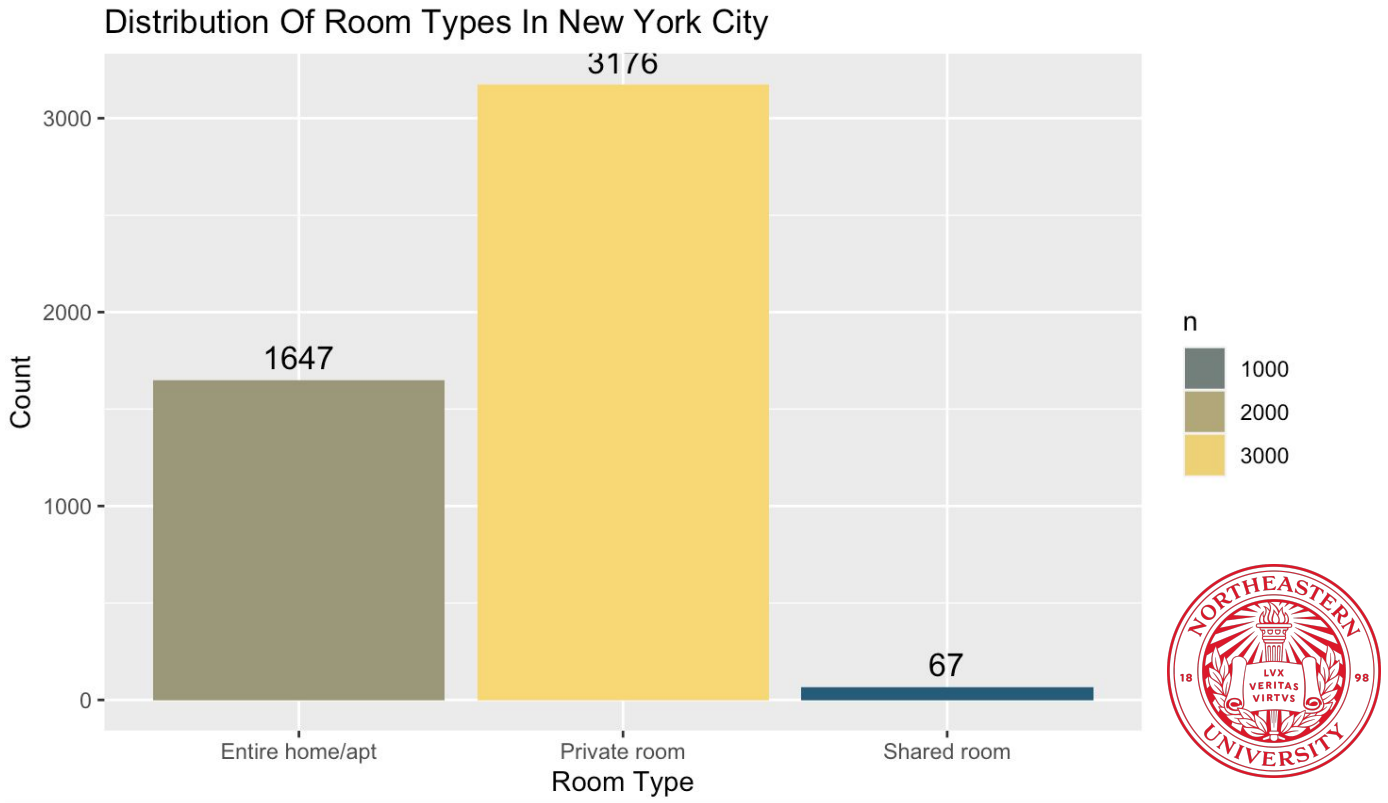
What is the distribution of the room types located in New York City?

Analysis:
From the graph above, we can observe that the number of Private rooms in New York are highest followed by Entire Apartments and then Shared Rooms.

Reason:
- High demand of hotel industry.
- Number of customer are high.
- High lifestyle

```
counts <- airbnb %>%
  count(room_type)

ggplot(data = counts) +
  geom_bar(mapping = aes(x = room_type, y = n, fill = n),
           stat = "identity", position = "dodge") +
  scale_fill_gradient(low = "#005E7D", high = "#FFD662", guide = "legend") +
  geom_text(aes(x = room_type, y = n, label = n),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 4.5) +
  labs(title = "Distribution Of Room Types In New York City",
       x = "Room Type", y = "Count")
```



Stacked bar chart showing the distribution of room types by neighbourhood group:

Business Question:

What is the distribution of the neighbourhood groups variable and percentage of Airbnb's located in each area?

Analysis:

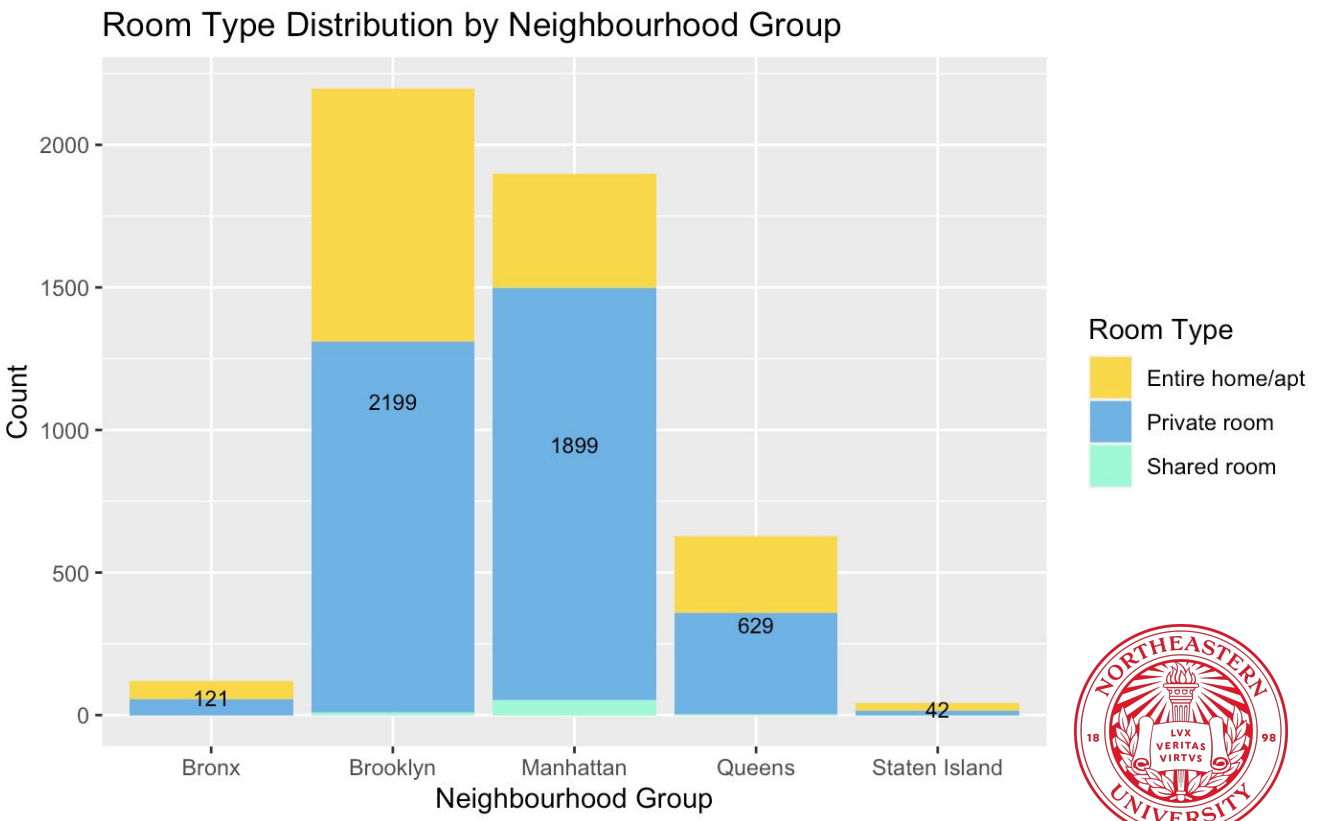
We have differentiated the neighbourhood in 5 groups as shows in the given table.

Out of Bronx, Brooklyn, Manhattan, Queens, and Staten Island, the percentage of Airbnb's located in Brooklyn is 45% which is the highest among all whereas Manhattan is at 39% followed by Queens which is at 13%, Bronx at 2% and Staten Island is below 1%. You can also see the segregation based on different colors.

Reason:

- Close by nature, parks and archeological sites.
- Good weather and food.
- Good commute and less traffic.

```
ggplot(data = airbnb) +
  geom_bar(mapping = aes(x = neighbourhood_group, fill = room_type),
           position = "stack") +
  stat_count(mapping = aes(x = neighbourhood_group, label = stat(count)),
             geom = "text", position = position_stack(vjust = 0.5), size = 3) +
  scale_fill_manual(values = c("#FFD700", "#56B4E9", "#7FFFD4")) +
  labs(title = "Room Type Distribution by Neighbourhood Group",
       x = "Neighbourhood Group", y = "Count") +
  guides(fill = guide_legend(title = "Room Type"))
```



Scatterplot showing the relationship between Price and Minimum nights:

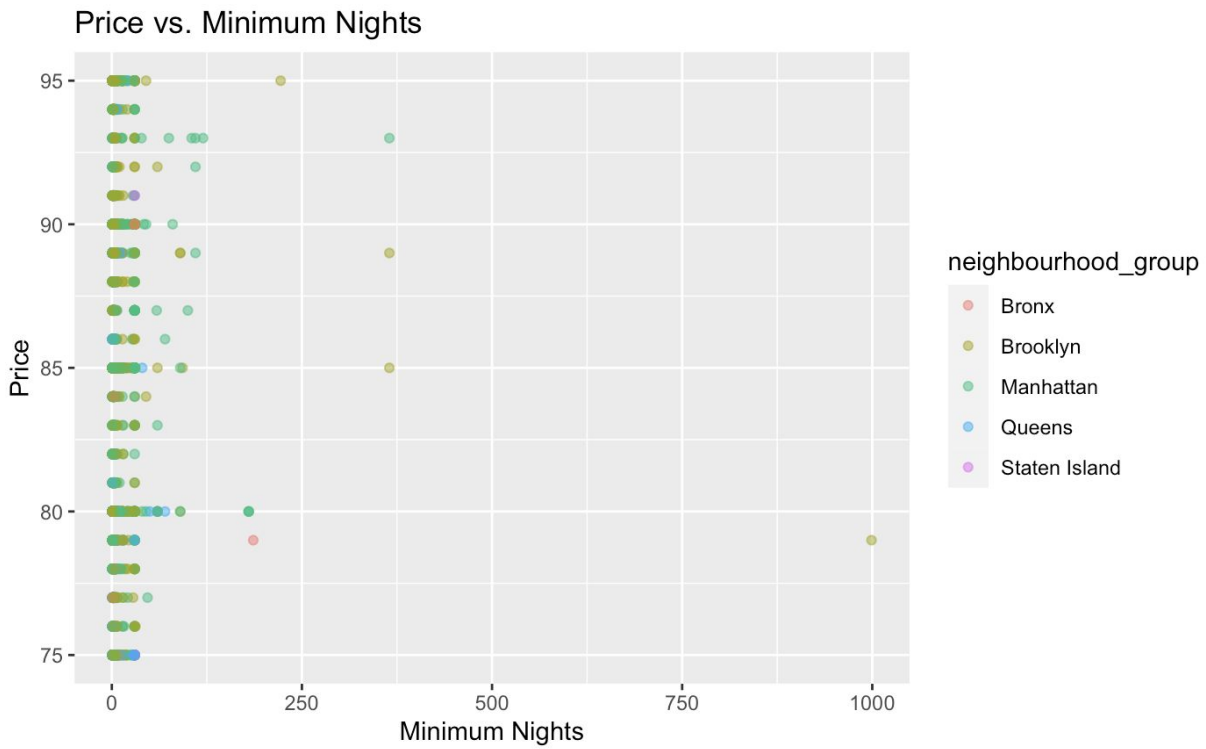
Business Question:

What is the Price of Airbnb's in different neighborhoods based on minimum nights?

Analysis:

- Based on this scatterplot, we can observe that the price of Airbnb's start from $ 75.
- The different neighborhoods are shown through different colors.
- We can also observe that we have an outlier that is in Brooklyn and allows a minimum of 1000 nights.

```
ggplot(data = airbnb) +
  geom_point(mapping = aes(x = minimum_nights, y = price, color = neighbourhood_group), alpha = 0.5) +
  labs(title = "Price vs. Minimum Nights", x = "Minimum Nights", y = "Price")
```



Boxplot showing the distribution of prices by room type:

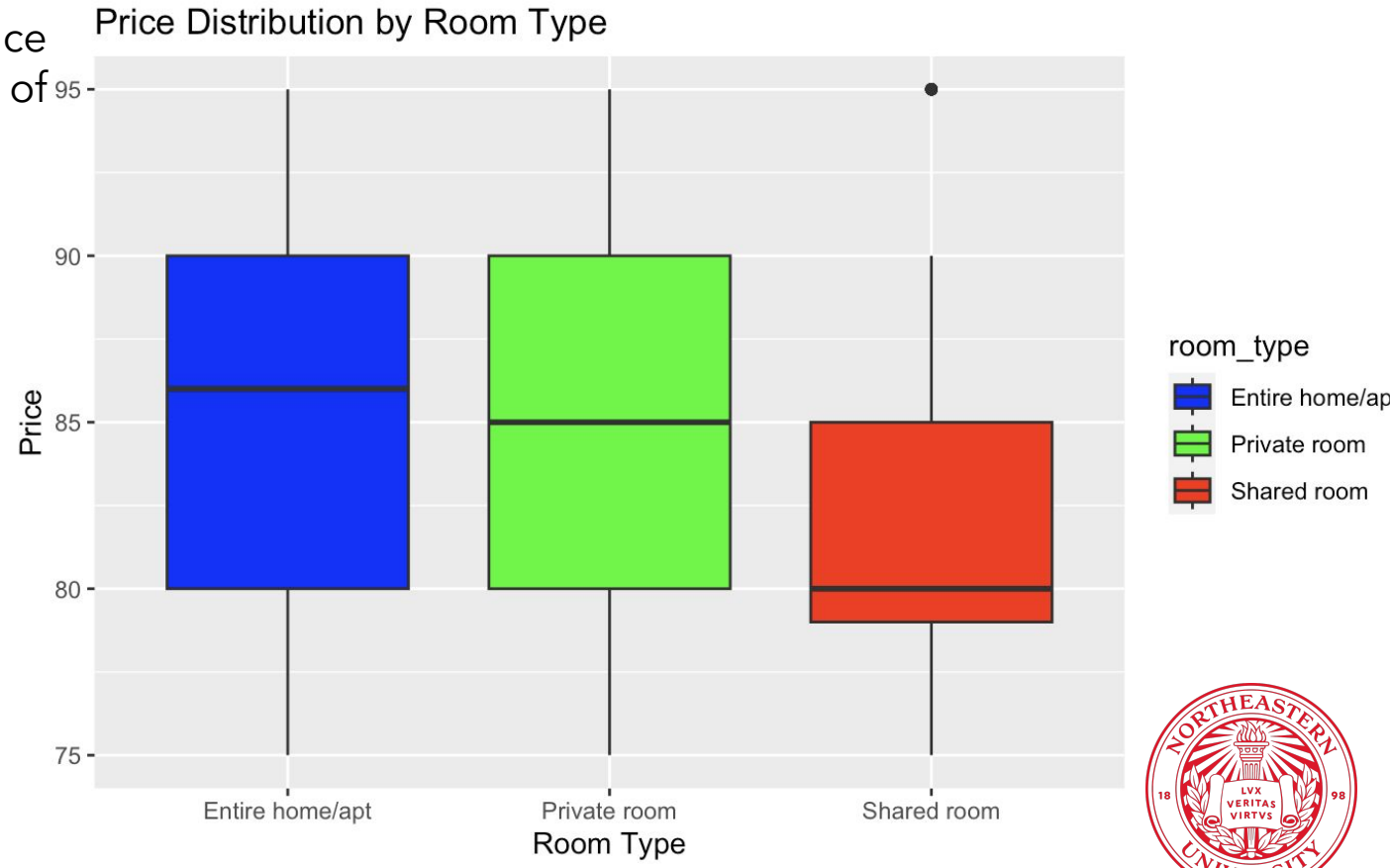
Business Question:
Is there a pricing difference between the various room types?

```
ggplot(data = airbnb) +
  geom_boxplot(mapping = aes(x = room_type, y = price, fill = room_type)) +
  scale_fill_manual(values = c("blue", "green", "red")) +
  labs(title = "Price Distribution by Room Type", x = "Room Type", y = "Price")
```

Analysis:
Entire home or apartment has the highest median price out of all of them i.e., 86. It also has the least number of outliers whereas Private room has 85 median price.
Shared room has the lowest median price i.e., 80.

Reason:

- Close by nature and parks
- Good weather and food.
- Good commute and less traffic.
- Providing more privacy



Map view for specific room types as per price

Business Question:

How much is the price density as per room types across a region?

```
qmplot(longitude, latitude, data = airbnb, geom = "point", color = price, main = "Density of Latitude and Longitude") +
facet_wrap(~ room_type)
```

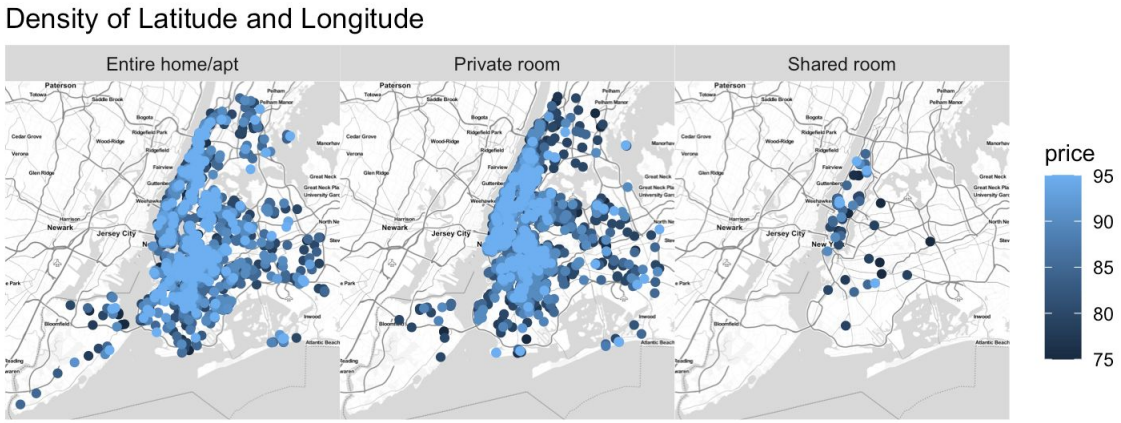
```
mapview(airbnb_unclean, xcol = "longitude", ycol = "latitude", crs = 4326, grid = FALSE) # coordinate reference system (crs)
```

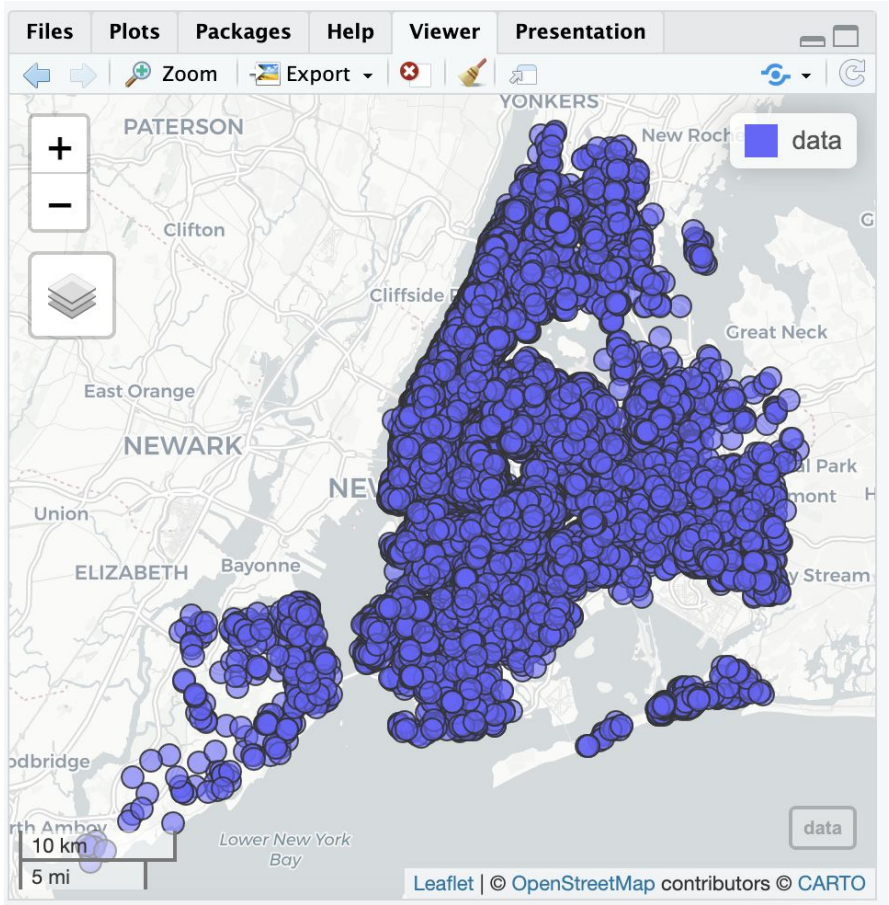
Analysis:

- We can clearly analyse the density of prices across all 3 room types and Private rooms has the highest denser plot in the region.

Reason:

- Good hospitality service and corporate discounts, seasonal price change
- Better options on room availability.





Density plot showing the distribution of prices:

Business Question:

Density plot showing the distribution of prices:

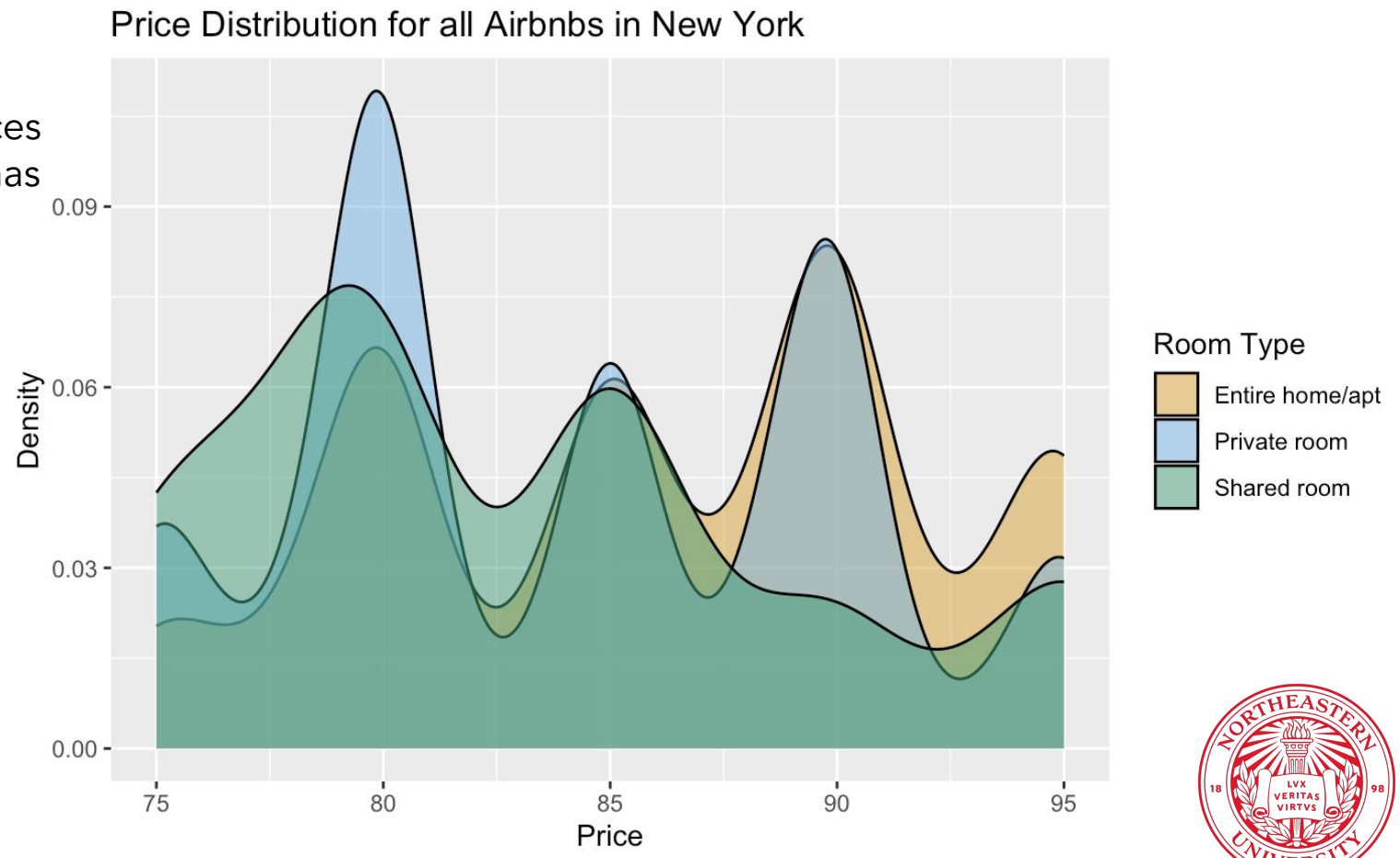
```
ggplot(data = airbnb, aes(x = price, fill = room_type)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("#E69F00", "#56B4E9", "#009E73")) +
  labs(title = "Price Distribution for all Airbnbs in New York", x = "Price", y = "Density", fill = "Room Type")
```

Analysis:

- We can clearly analyse the density of prices across all 3 room types and Private rooms has the highest denser plot in the region.

Reason:

- Good hospitality service and corporate discounts.
- Better options on room availability.



Approach 2: Shiny R

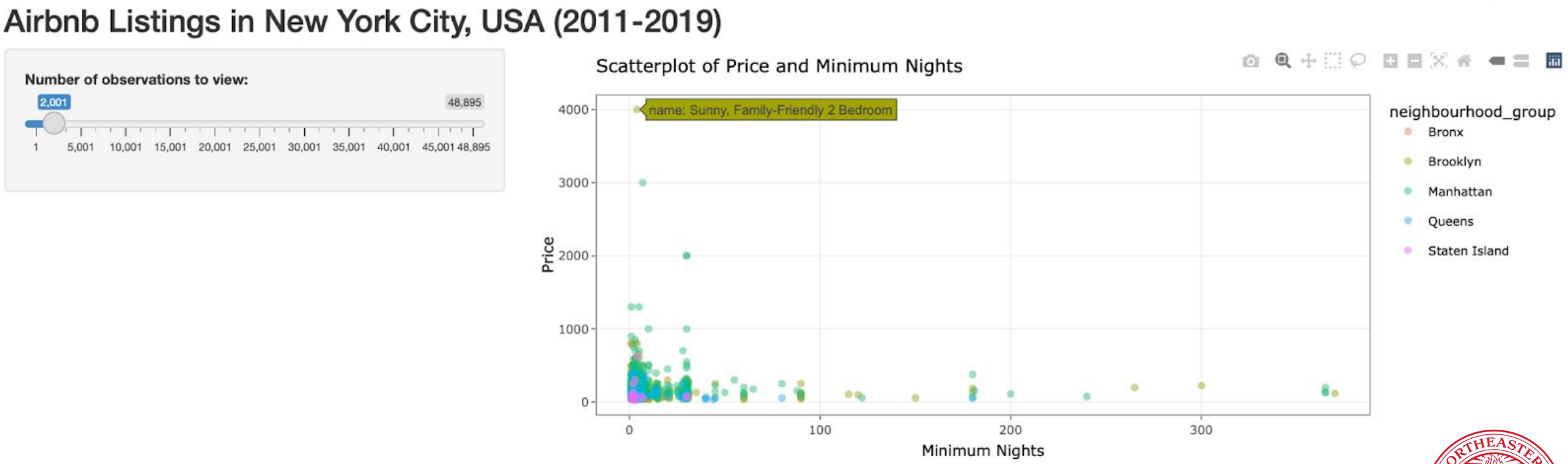
Scatter Plot showing distribution of prices based on neighborhood

Business Question:

Density plot showing the distribution of prices:

Analysis:

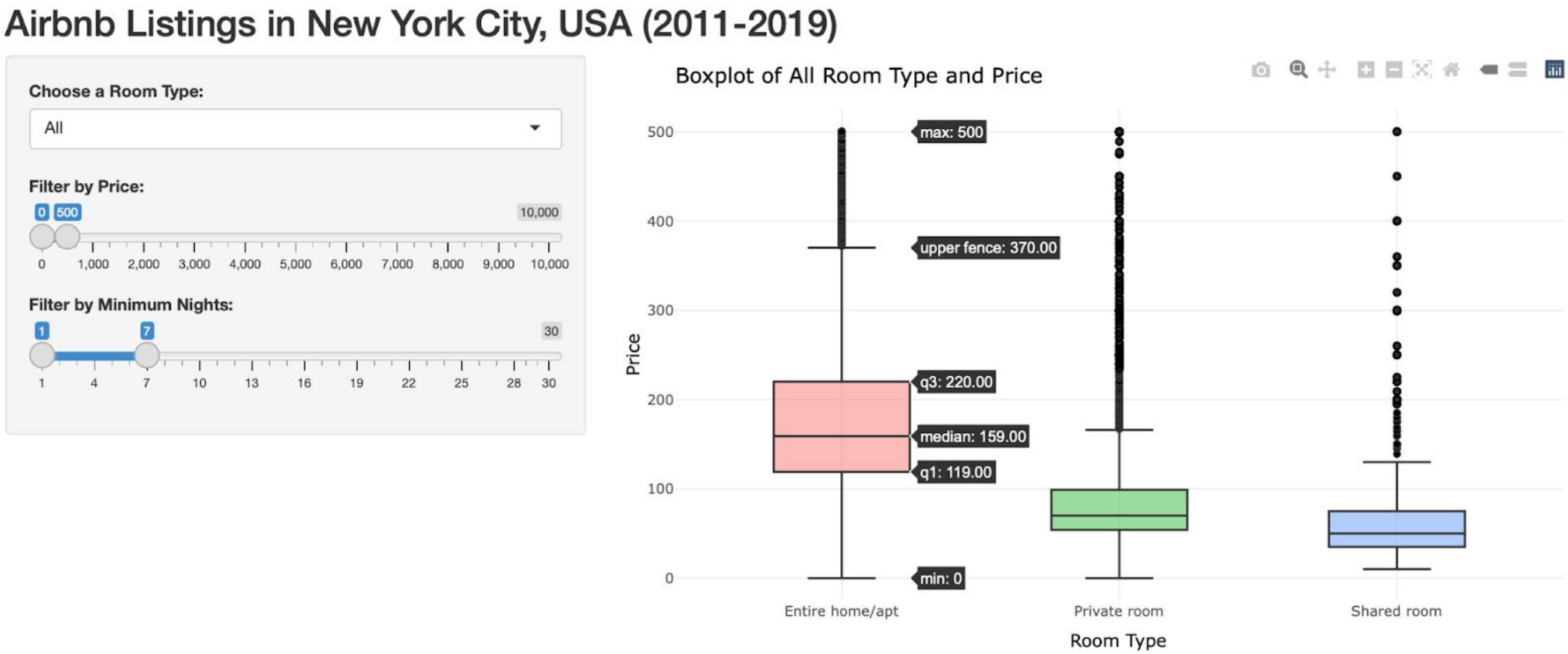
- We can clearly analyse the density of prices across various location base on the price



Box plot showing the Room Type and its Price

Analysis:

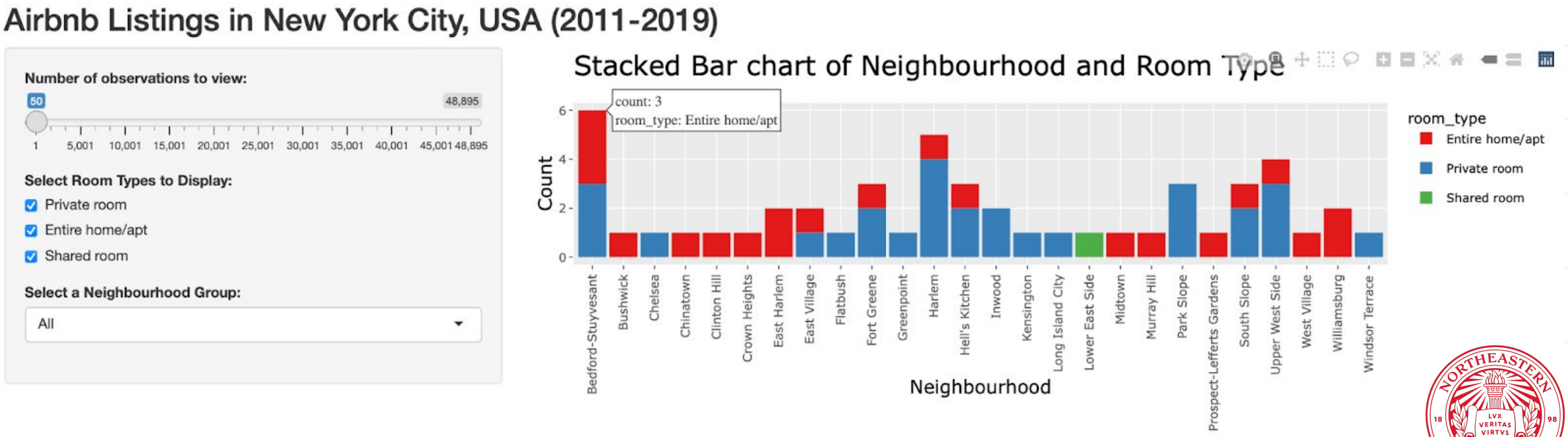
- We can clearly analyse the prices of these three types of the room with this box plot.



Stacked Bar Chart showing Room Types and Neighbourhood

Analysis:

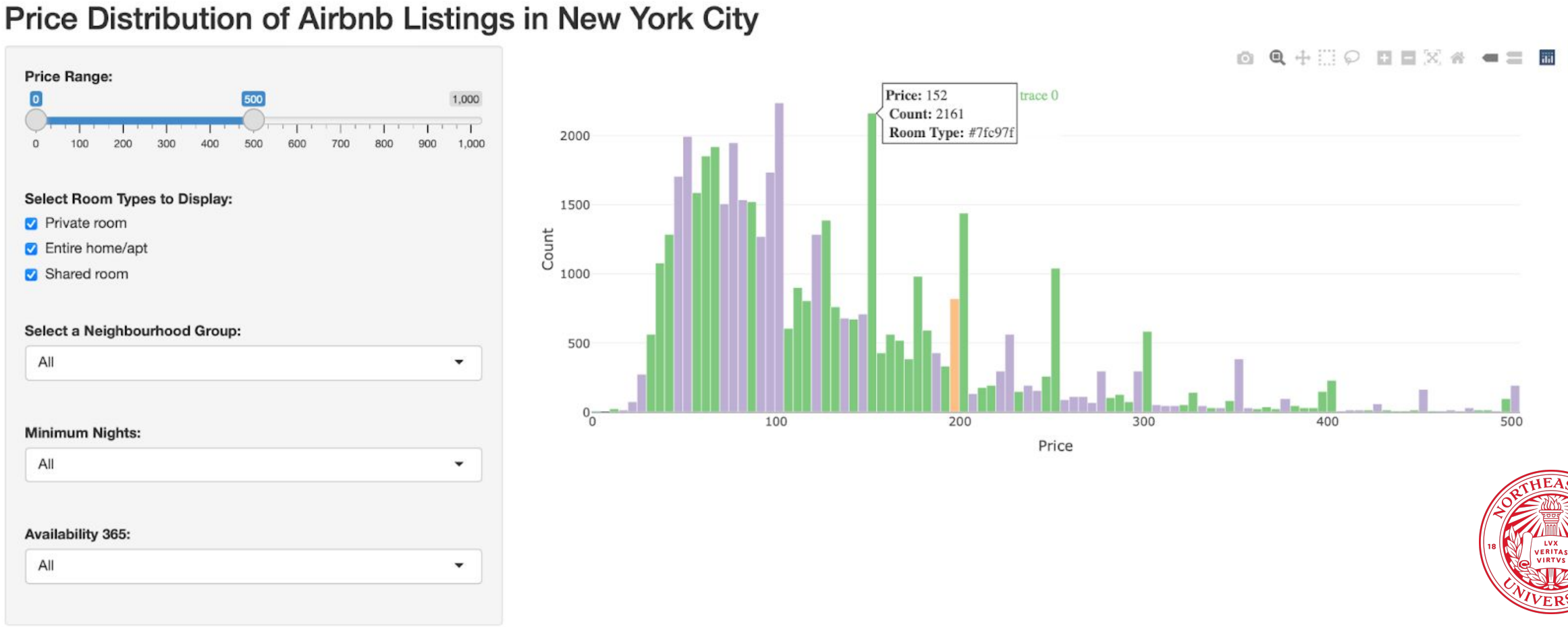
- We can clearly analyse the density of prices across all 3 room types and Private rooms has the highest denser plot in the region.



Histogram plot showing different Neighbourhood and Prices

Analysis:

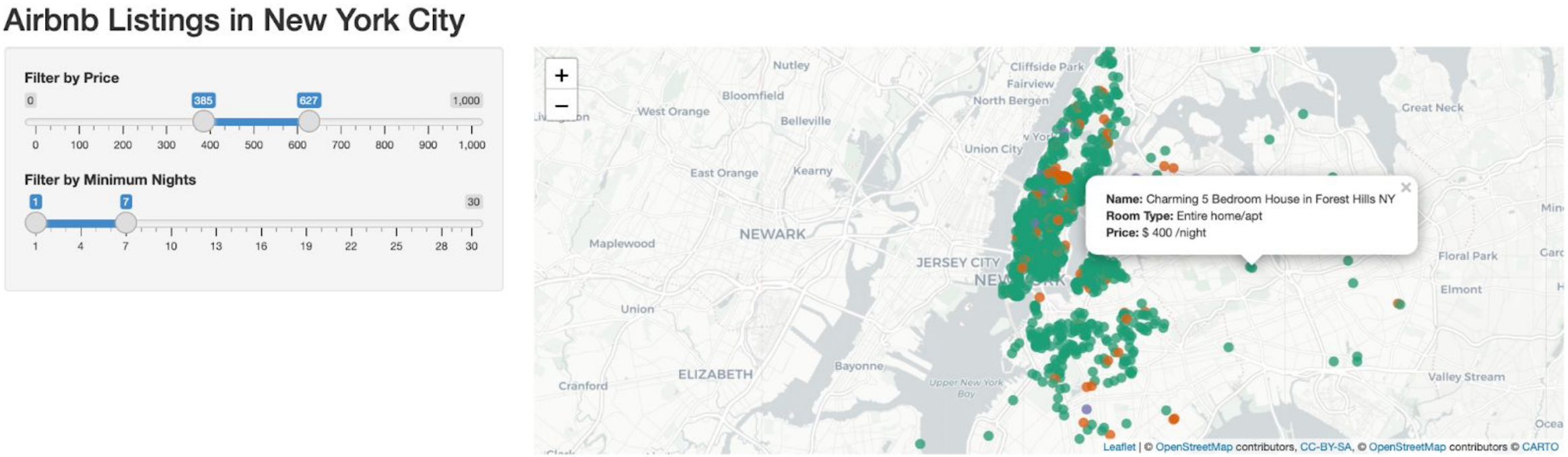
- We can clearly analyse the density of prices across all 3 room types in different neighbourhoods.



Map view showing the Prices and Minimum Stays

Analysis:

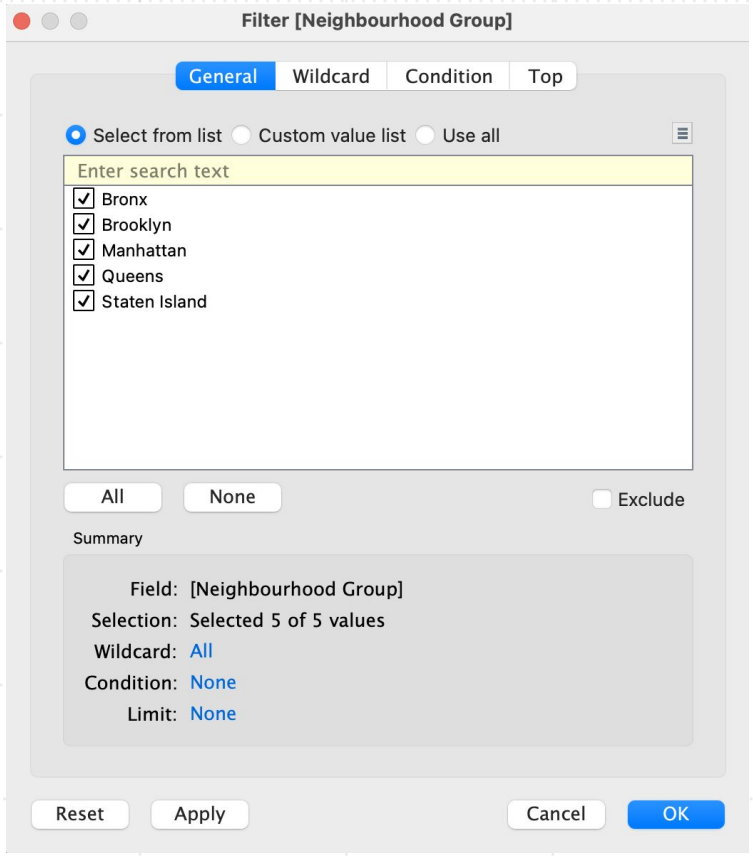
- We can clearly analyse the density of prices across all 3 room types and Private rooms has the highest denser plot in the region.

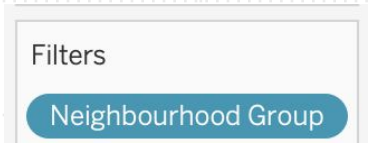


Approach 3: Tableau

Filter, Parameter and Dynamic measure

Filter: We created a filter to display only the data for a specific neighborhood. To do this, we added a "neighbourhood_group" field to the Filters shelf, selected the neighborhoods to display. We then used the filter to dynamically change the data displayed in the visualization.

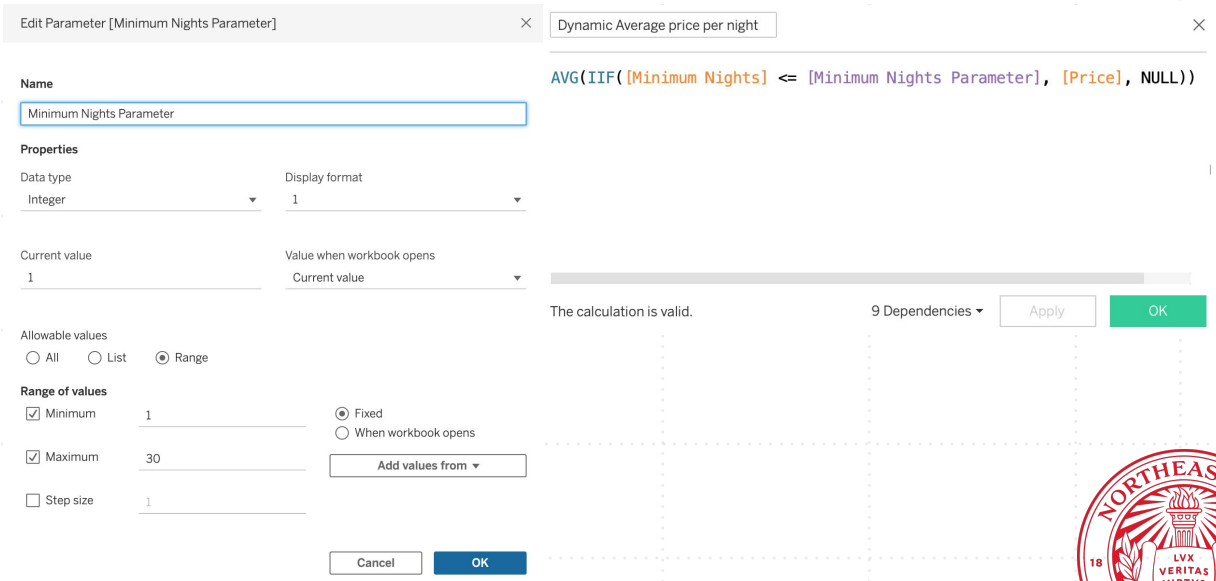




Parameter: We created a parameter on the "minimum_nights" field to allow the user to change the minimum number of nights for a stay.

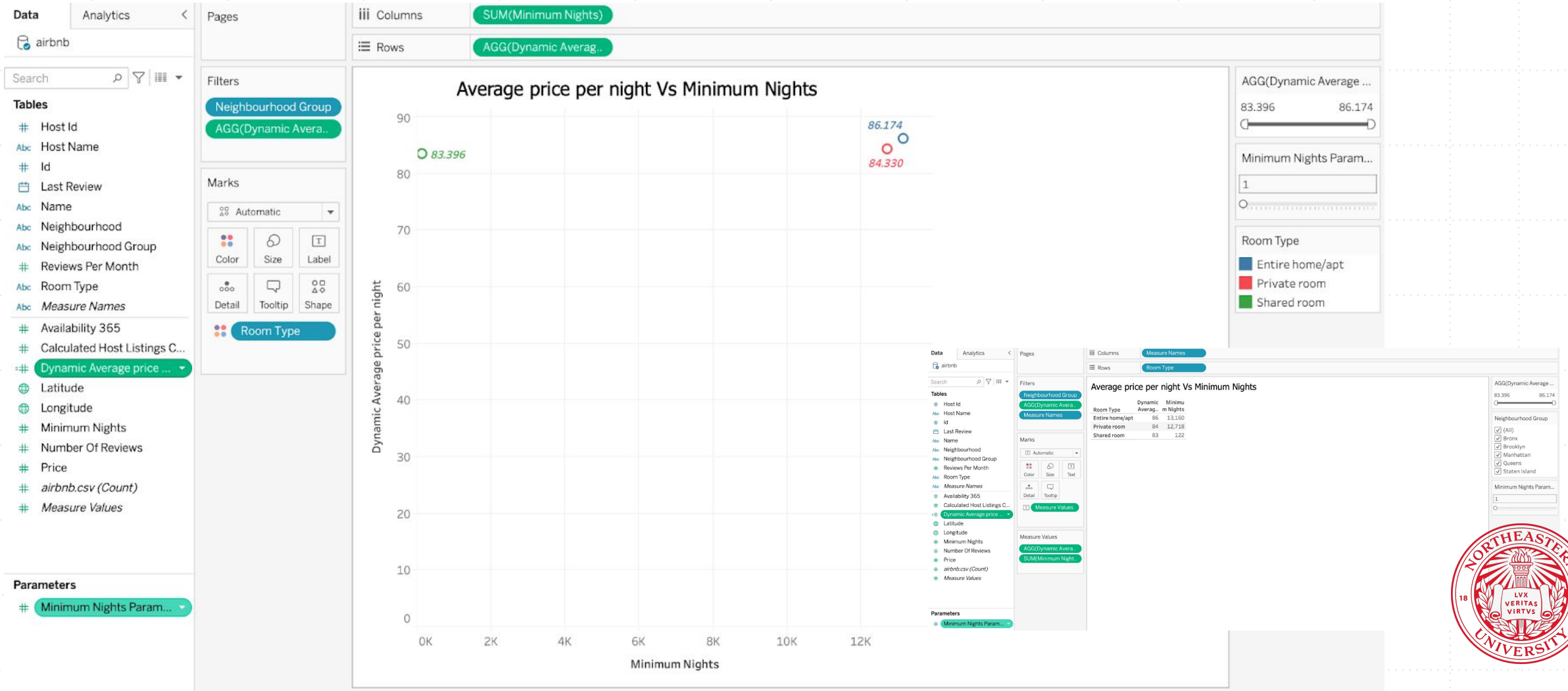
We can now use the parameter to dynamically change the data displayed in the visualization.

Dynamic measure: We created a dynamic measure to show the average price per night on the "price" field and enter the below formula:



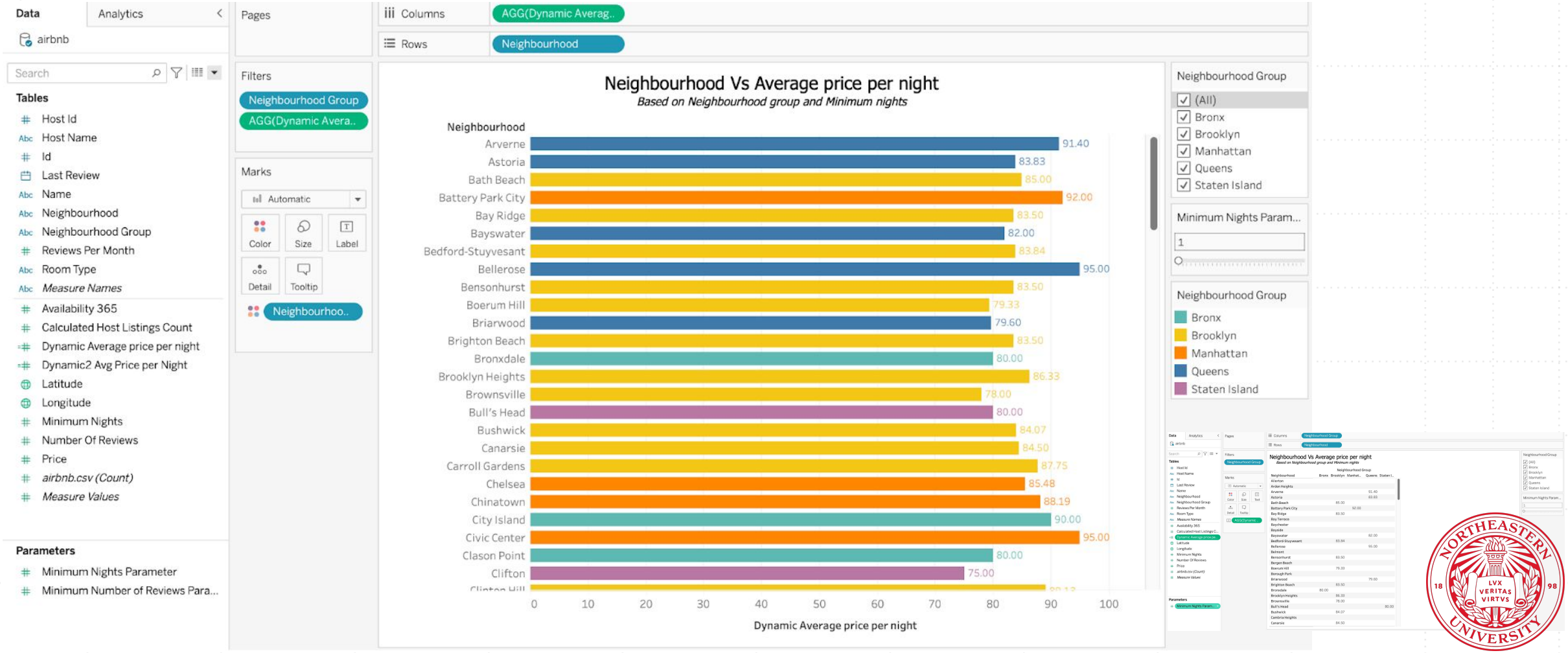
Scatter Plot:

Drag the "minimum_nights" field to the Columns shelf, and the dynamic measure created in the previous step (e.g., "Average Price per Night") to the Rows shelf. We added the "room_type" field to the Color shelf to differentiate between the different types of rooms.



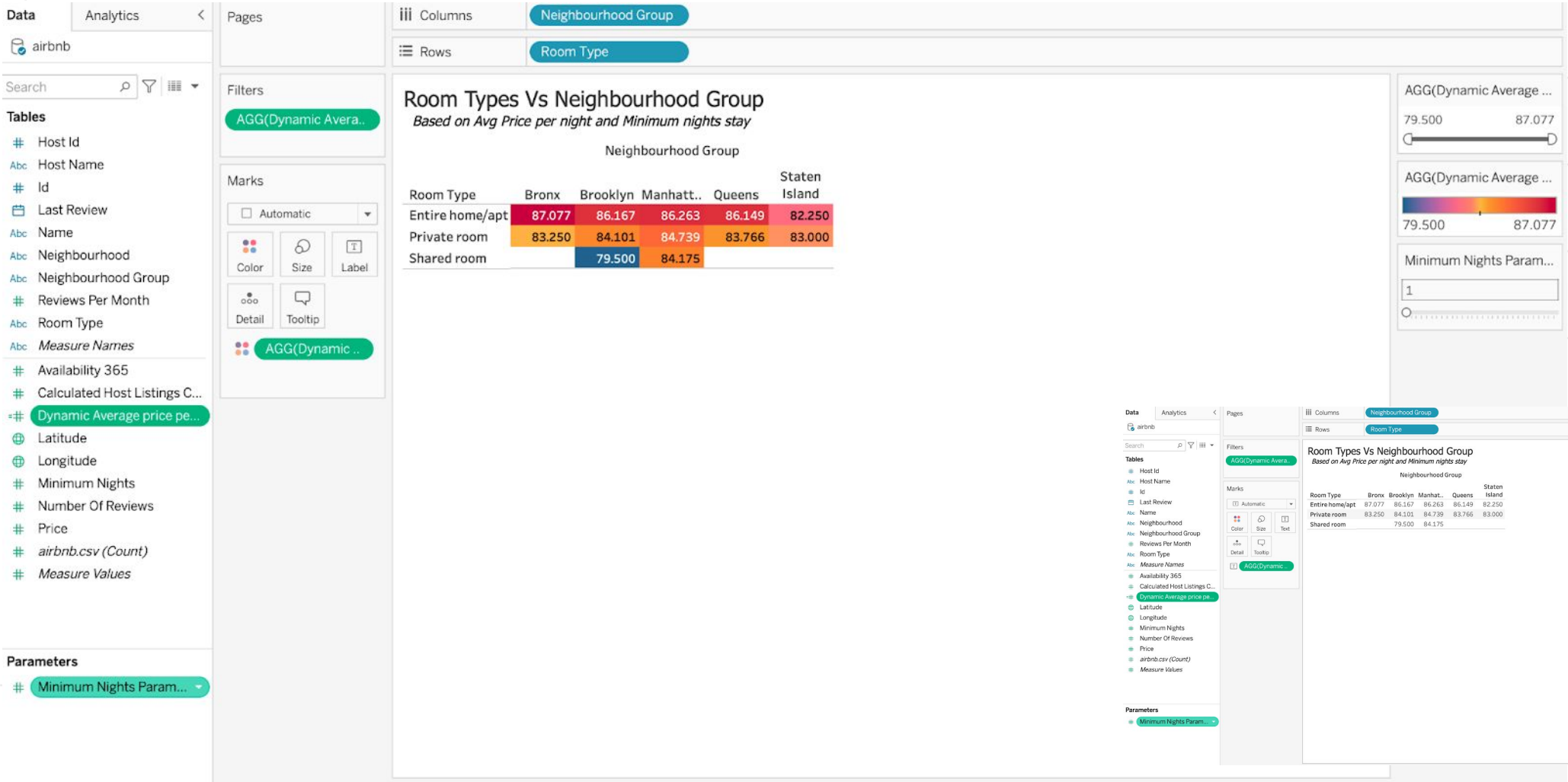
Bar Chart:

Drag the "neighbourhood" field to the Rows shelf, the dynamic measure created in the previous step to the Columns shelf, and the filter created in the first step (e.g., "Neighbourhood Group") to the Filters shelf. This will create a bar chart showing the average price per night for each neighbourhood.



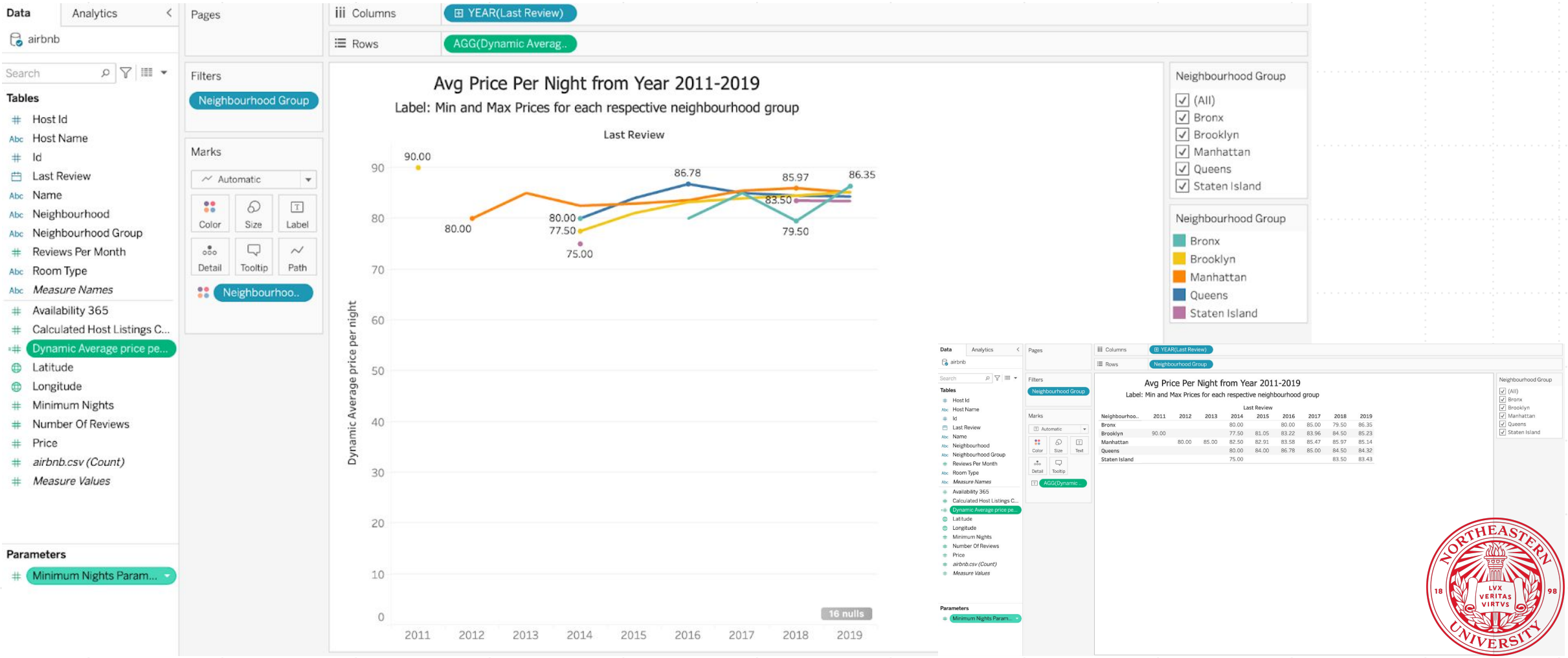
Heat Map:

Drag the "neighbourhood_group" field to the Columns shelf, the "room_type" field to the Rows shelf, and the dynamic measure created in the previous step to the Color shelf. We also added the "minimum_nights" parameter to the Filters shelf, so the user can adjust the minimum number of nights to see how the price per night changes.



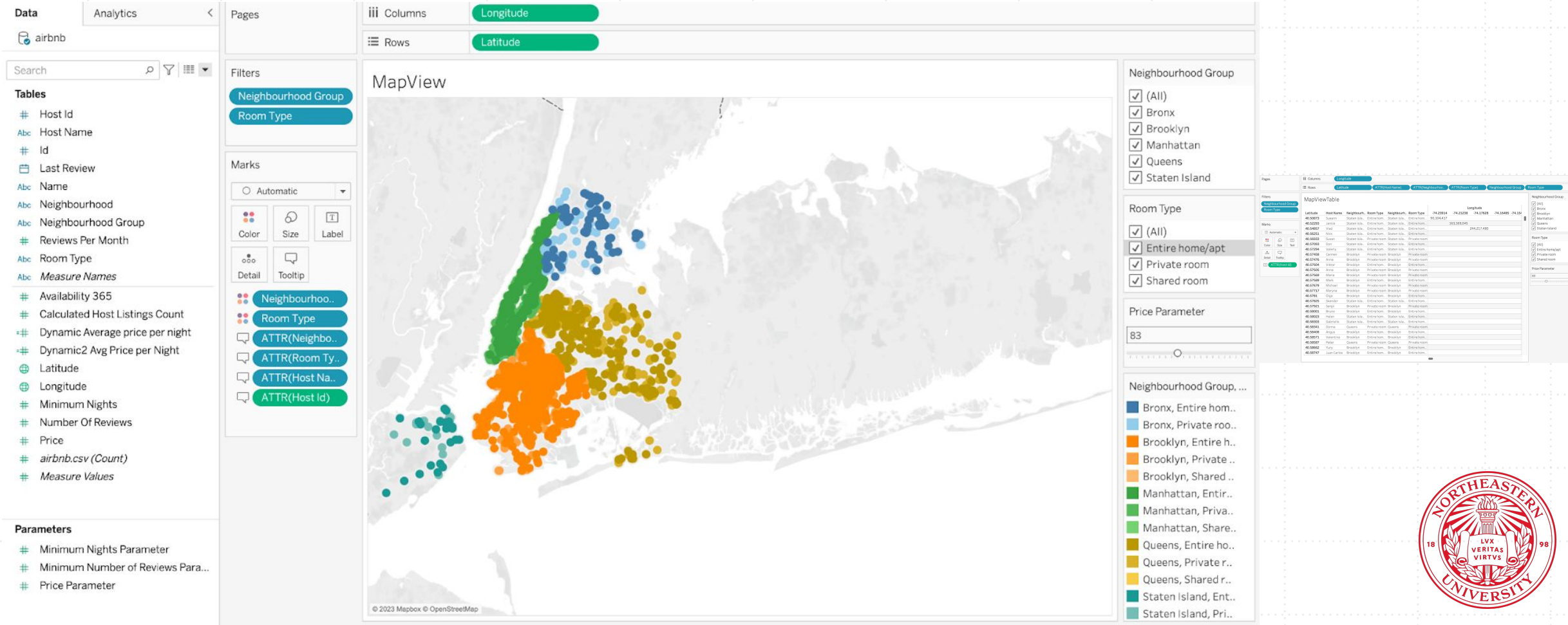
Line Chart:

Drag the "last_review" field to the Columns shelf, the dynamic measure created in the previous step to the Rows shelf, and the "neighbourhood_group" field to the Color shelf. We also added a filter to show data for a specific neighbourhood.



Map View:

Drag the "host_id" field to the Columns shelf, the "minimum_nights" parameter to the Filters shelf, and the dynamic measure created in the previous step to the Rows shelf. We also added the "room_type" field to the Color shelf to differentiate between the different types of rooms. This chart will show the average price per night for each host, stacked by the type of room.



Conclusion

The choice of data visualization tool depends on the specific needs and goals of the project. In the case of the dataset provided, R code, Shiny R, and Tableau can all be effective tools for communicating and visualizing the data.

R code is a popular choice for data analysis and visualization because it provides a wide range of customizable graphs and charts. With R, users can create complex visualizations that display multiple data variables, and can use advanced statistical techniques to explore relationships between variables. R code also allows for reproducibility and sharing of analysis and visualizations with others.

Shiny R is a web application framework for R that allows users to build interactive web applications that display data visualizations. Shiny R can be a great choice for projects where stakeholders need to interact with the data and make decisions based on the analysis. Shiny R applications can also be easily customized and updated as needed, allowing for more flexibility in data exploration.

Tableau is a popular data visualization tool that allows users to create interactive dashboards and visualizations. Tableau provides a user-friendly interface that allows users to quickly create charts and graphs with drag-and-drop functionality. Tableau also provides advanced analytics and mapping capabilities, making it a good choice for projects that require more complex visualizations.

In conclusion, all three tools can be effective for visualizing the provided dataset. R code provides more customization and advanced statistical techniques, while Shiny R allows for more interactivity and flexibility in data exploration. Tableau provides a user-friendly interface and advanced analytics capabilities. Ultimately, the choice of tool will depend on the specific needs and goals of the project, as well as the user's level of experience and comfort with the tool.

References

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). shiny: Web Application Framework for R. R package version 1.6.0. Retrieved from https://CRAN.R-project.org/package=shiny

RStudio. (n.d.). Leaflet for R. Retrieved March 19, 2023, from https://rstudio.github.io/leaflet/

Kaggle. (2019). Airbnb NYC 2019. Retrieved from https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

R Core Team. (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

```
1  def gratitude():
2      print("Thank you.")
3
```

Any Questions?