**315 Project Timeline**

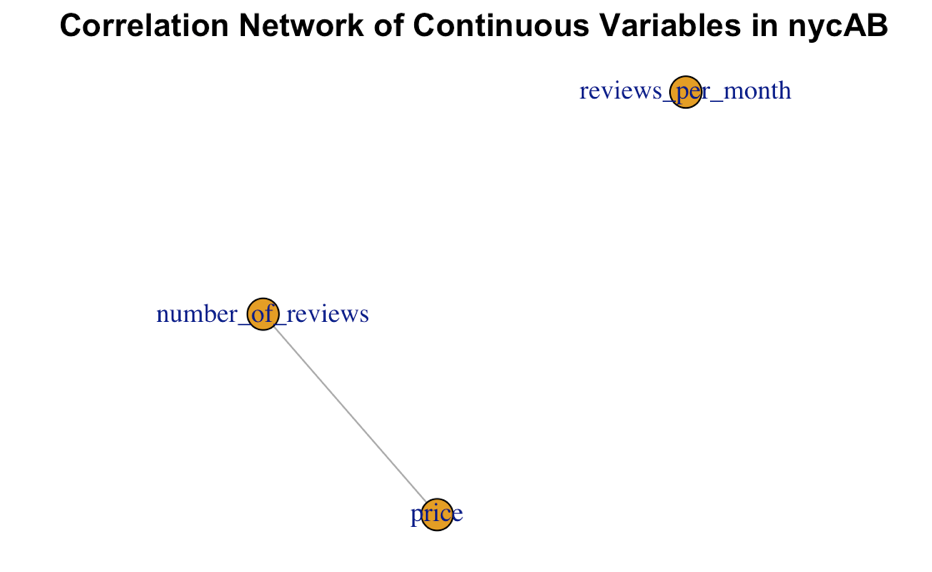
* **Dataset:** <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data#AB_NYC_2019.csv>
* **Variables:**
  + info about airbnb: `id`(nominal), `name`(nominal)
  + Host Info: `host\_id`(nominal), `host\_name`(nominal), `calculated\_host\_list`(numerical, amount of listing per host)
  + Geographic Info: `neighbourhood`(categorical), `latitude`(numerical), `longitude`(numerical)
  + Basic Airbnb Info: `room\_type`(categorical), `price`(numerical), `minimum\_nights`(numerical, amount of nights minimum to book), `availability\_365`(numerical, number of days when listing is available for booking)
  + Popularity Metric: `number\_of\_reviews`(numerical), `last\_review`(date-time-year, time of latest review), `reviews\_per\_month`(numerical, number of reviews per month)
* **Questions we like to explore?**
  + Which area and room type have more popular listings?
  + What variables affect prices
* **Part1: EDA [plot + interpretation] due Friday Nov 29**
  + Histogram of Price [Lucy]
    - Stacked Histogram/ Side-by-side Histogram
    - Color by neighborhood



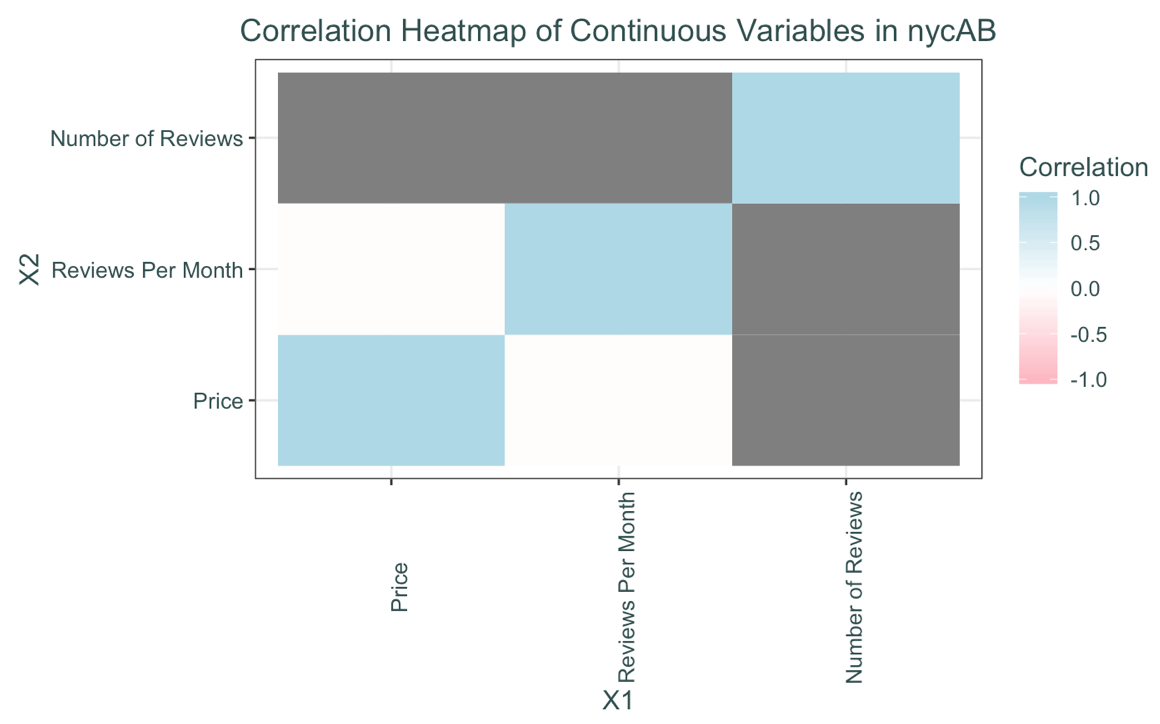
**Conclusion:**

The histogram for the price variable is unimodal, with a huge peak at the lower end. It is very skewed, with a long tail extending to the right hand side. The two center measurements that we typically use -- mean and median -- are very different in this case, with median 106 dollars, and mean 152 dollars. This is because the mean value is dragged toward the higher price end by the super expensive houses. It has multiple outliers, ranging from 2500 dollars to 10000 dollars.

* + Correlation Network [Lynn]
    - Igraph package
    - Refer to HW9 1a
    - See which variables are connected to price, try different significant level

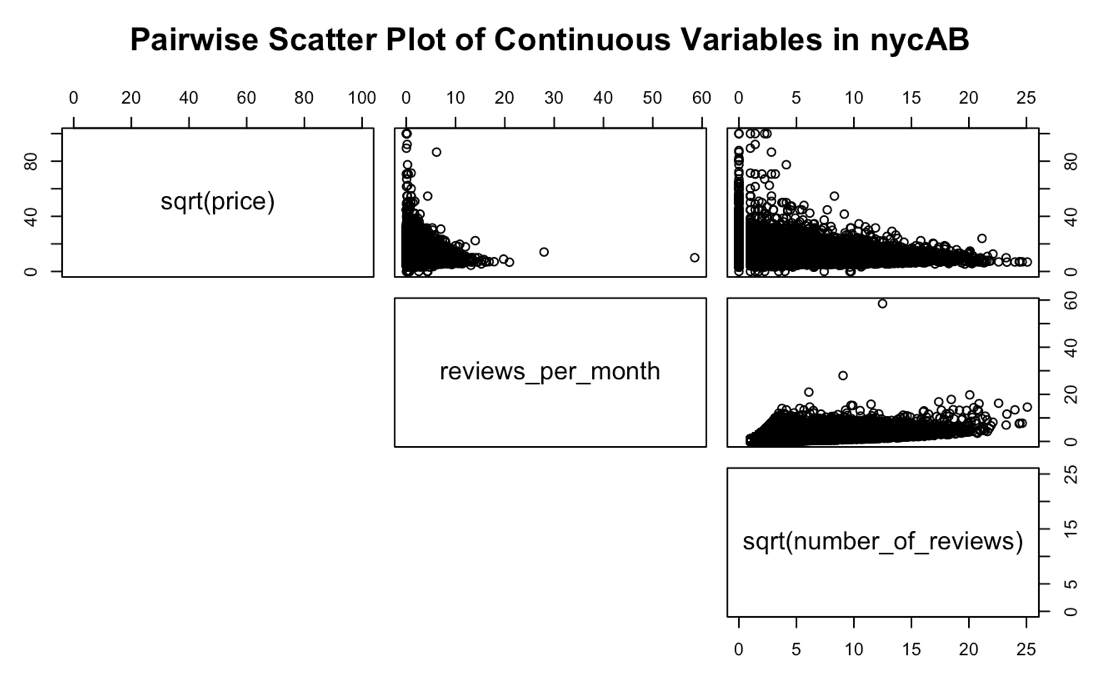
**Conclusion:**

From the graph we can see that number\_of\_reviews is related to price.

* + Correlation Heatmap [Lynn]

**Conclusion:**

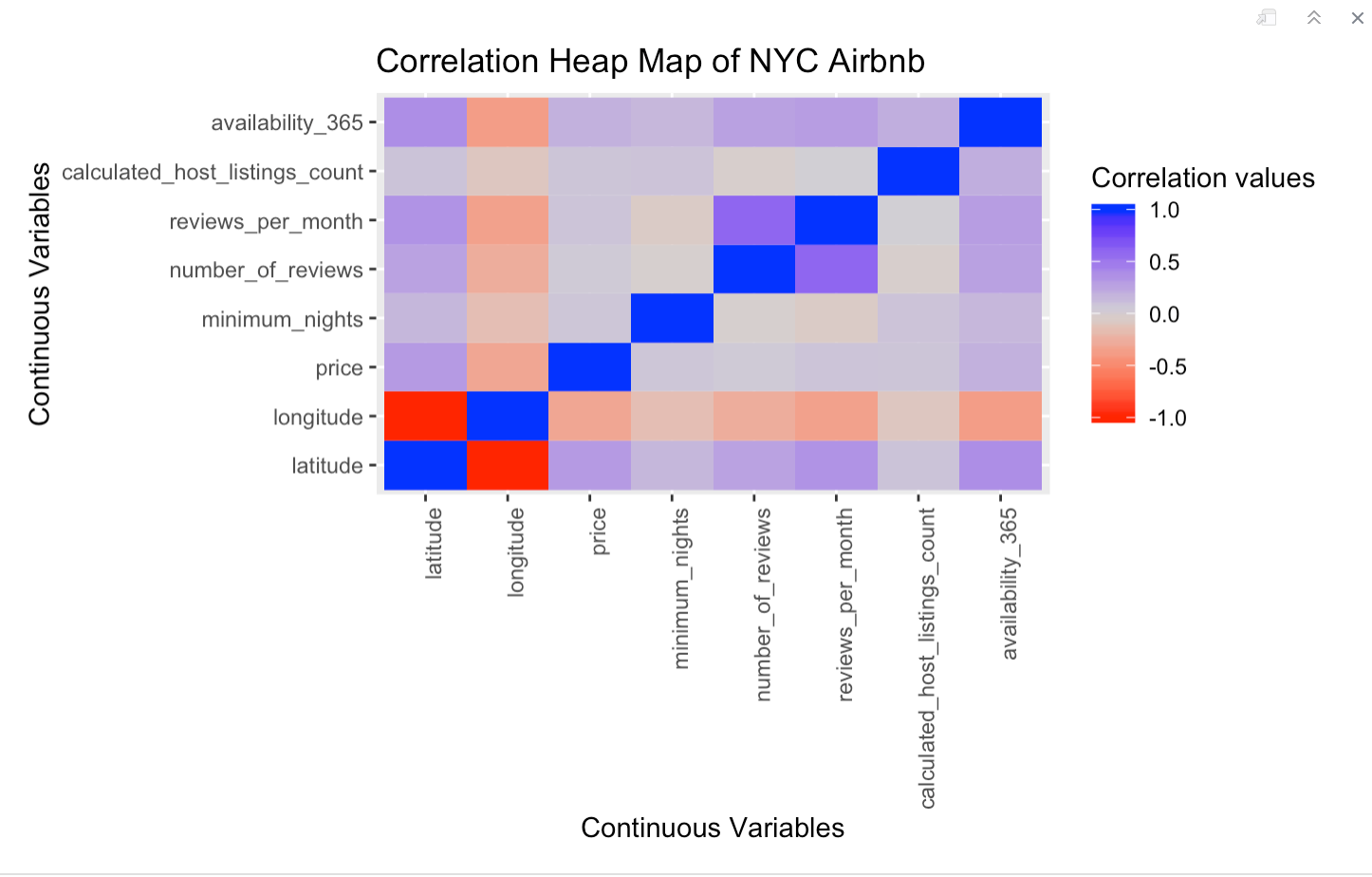
From the graph we can see that reviews\_per\_month has positive linear relationship to price.

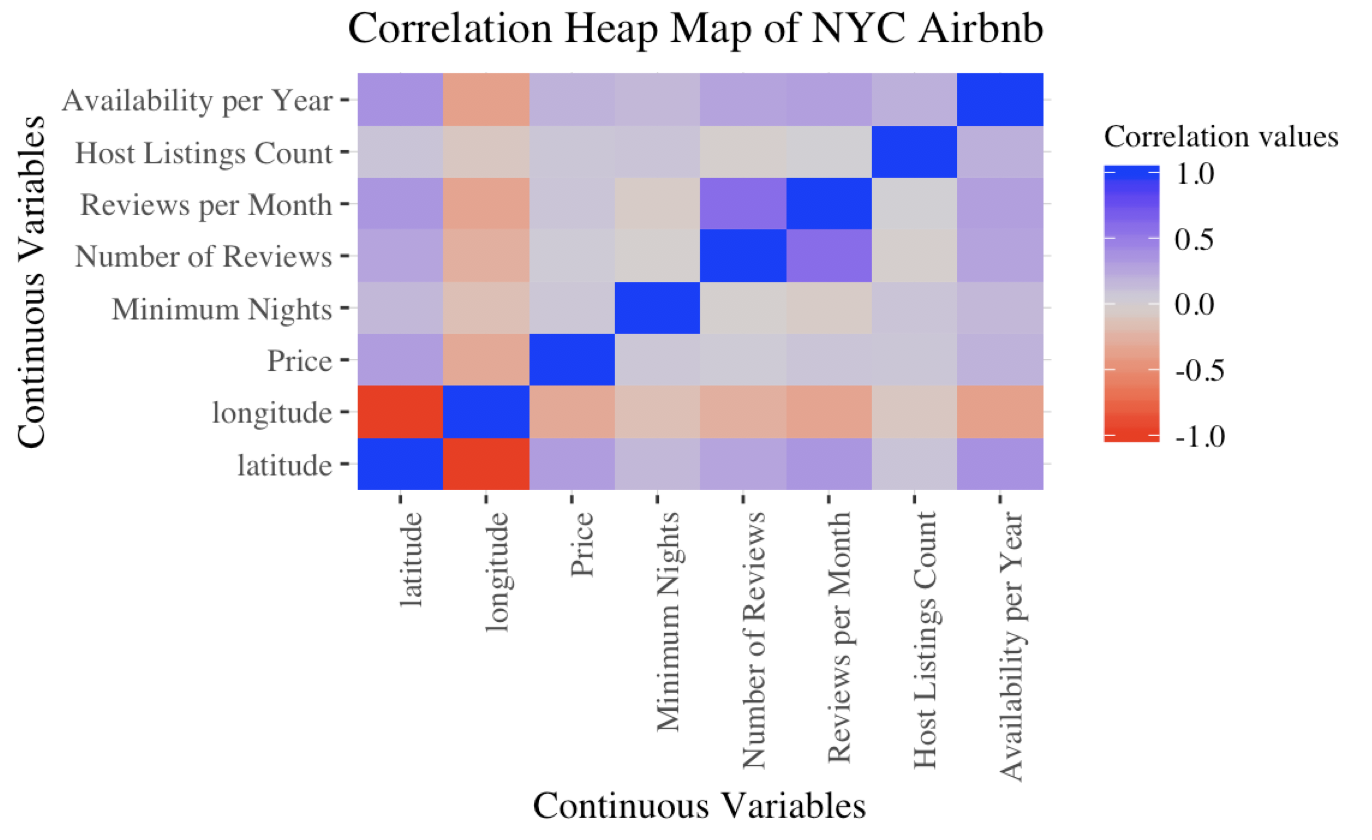
* + Pairwise Scatter Plot [ Lynn ]

**Conclusion:**

In this scatter plot, we apply square root to price variable and num\_of\_reviews variable to better display relationship between these continuous variables.

* + From the plot, we can see that when sqrt(price) > 15, sqrt(price) and reviews\_per\_month are negatively correlated. When sqrt(price) <= 15, sqrt(price) and reviews\_per\_month are positively correlated.
  + The linear relationship has the same pattern between sqrt(price) and sqrt(number\_of\_reviews).
  + Reviews\_per\_month and sqrt(number\_of\_reviews) has a positive linear relationship. Which makes sense
  + Correlation heatmap: [Lucy]





**Conclusion:**

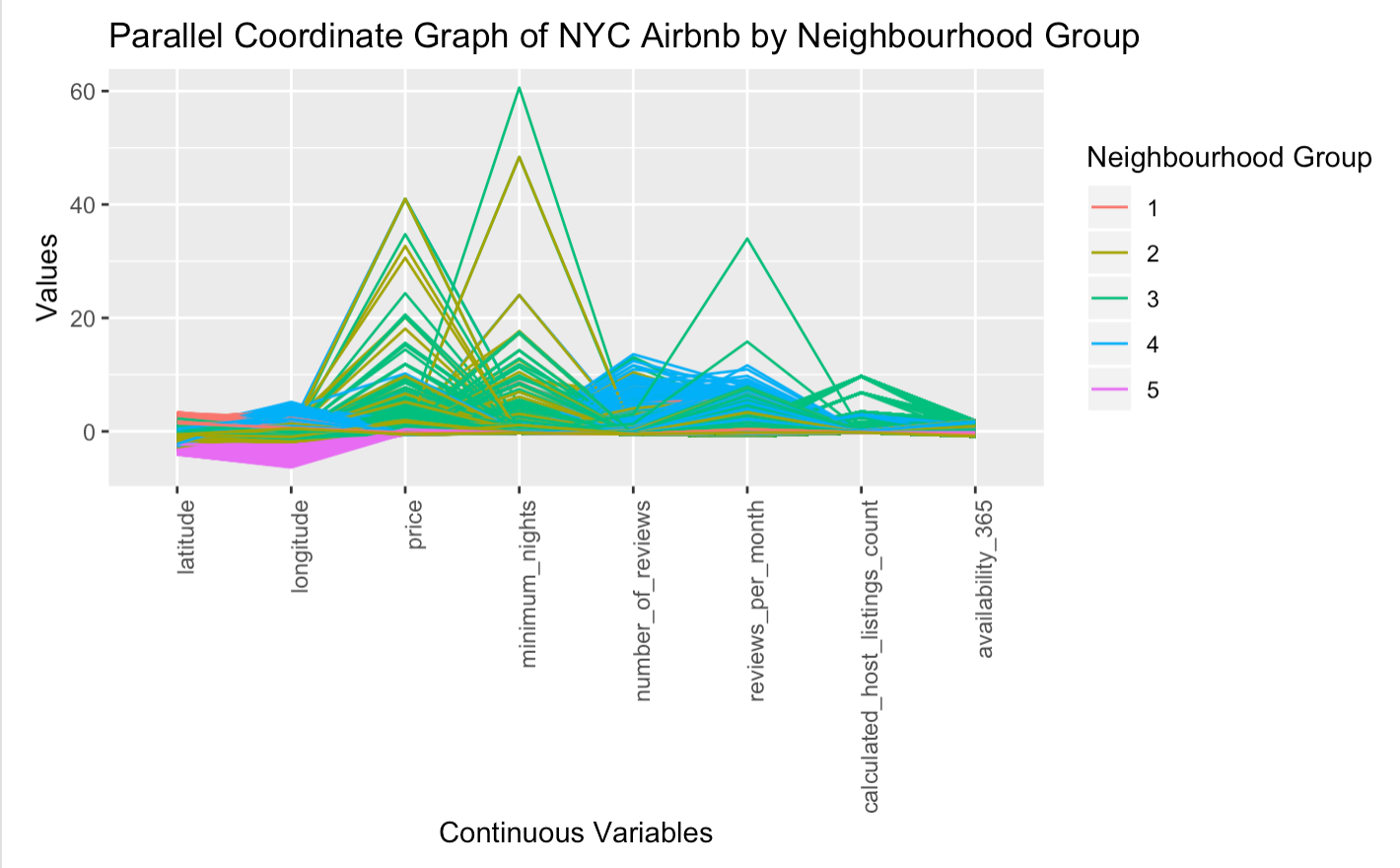
The correlation Heatmap shows us correlation between each continuous variables for the NYC Airbnb dataset (id and host\_id are not included).

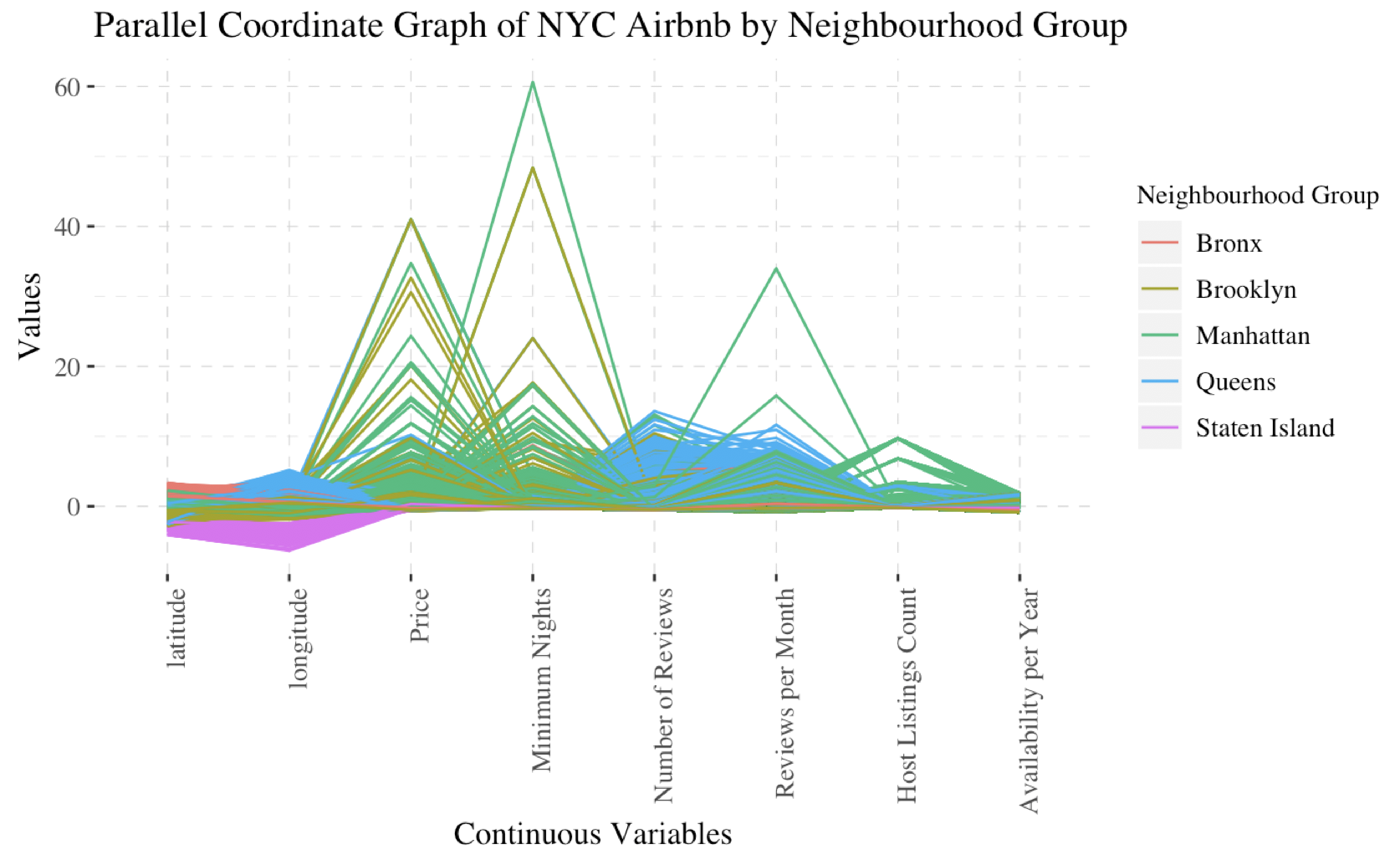
- Positively correlated variables are: review\_per\_month and number\_of\_reviews, review\_per\_month and avalibility\_365, latitude and avalibility\_365, latitude and review\_per\_month, latitude and price.

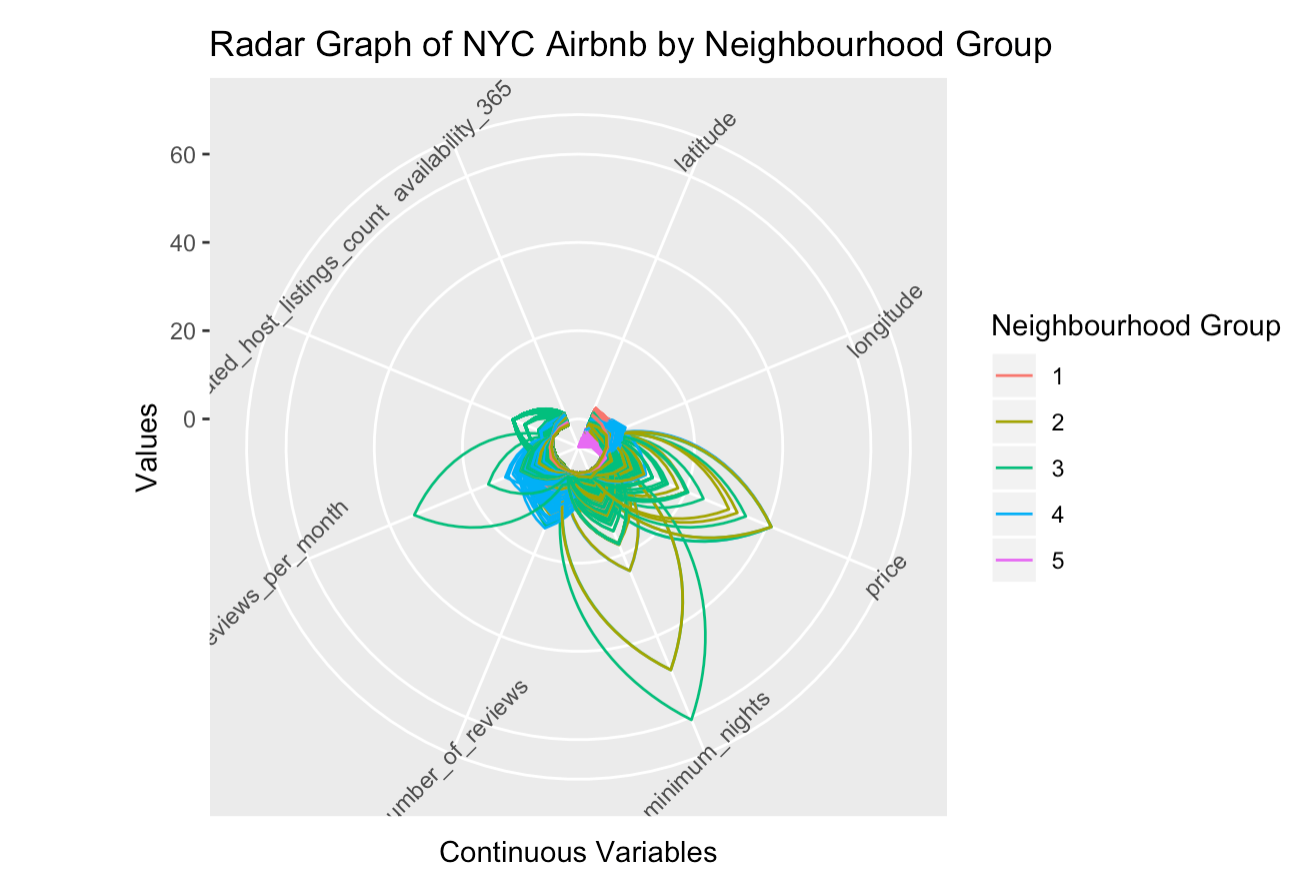
- Negatively correlated variables are: latitude and longitude, longitude and avalibility\_365, longitude and reviews\_per\_month, longitude and price.

- Almost not correlated variables are: price and minimum\_nights, price and number\_of\_reviews, latitude and calculated\_host\_listing\_count.

* + Parallel Coordinate Graph & Radar Graph [Lucy]
    - (on Numerical Variables)
    - Color by neighbourhood
    - Refer to HW8 1a



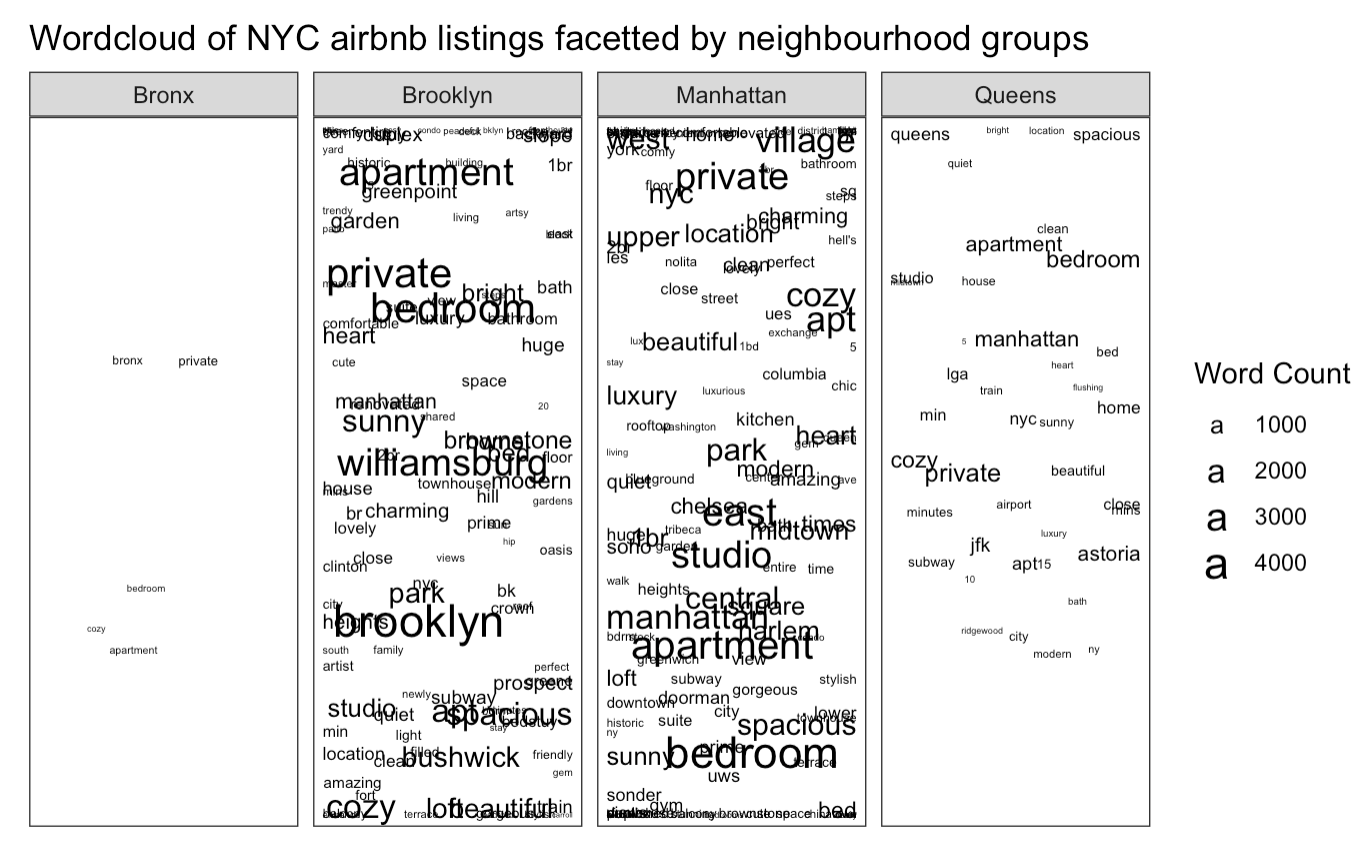




**Conclusion:**

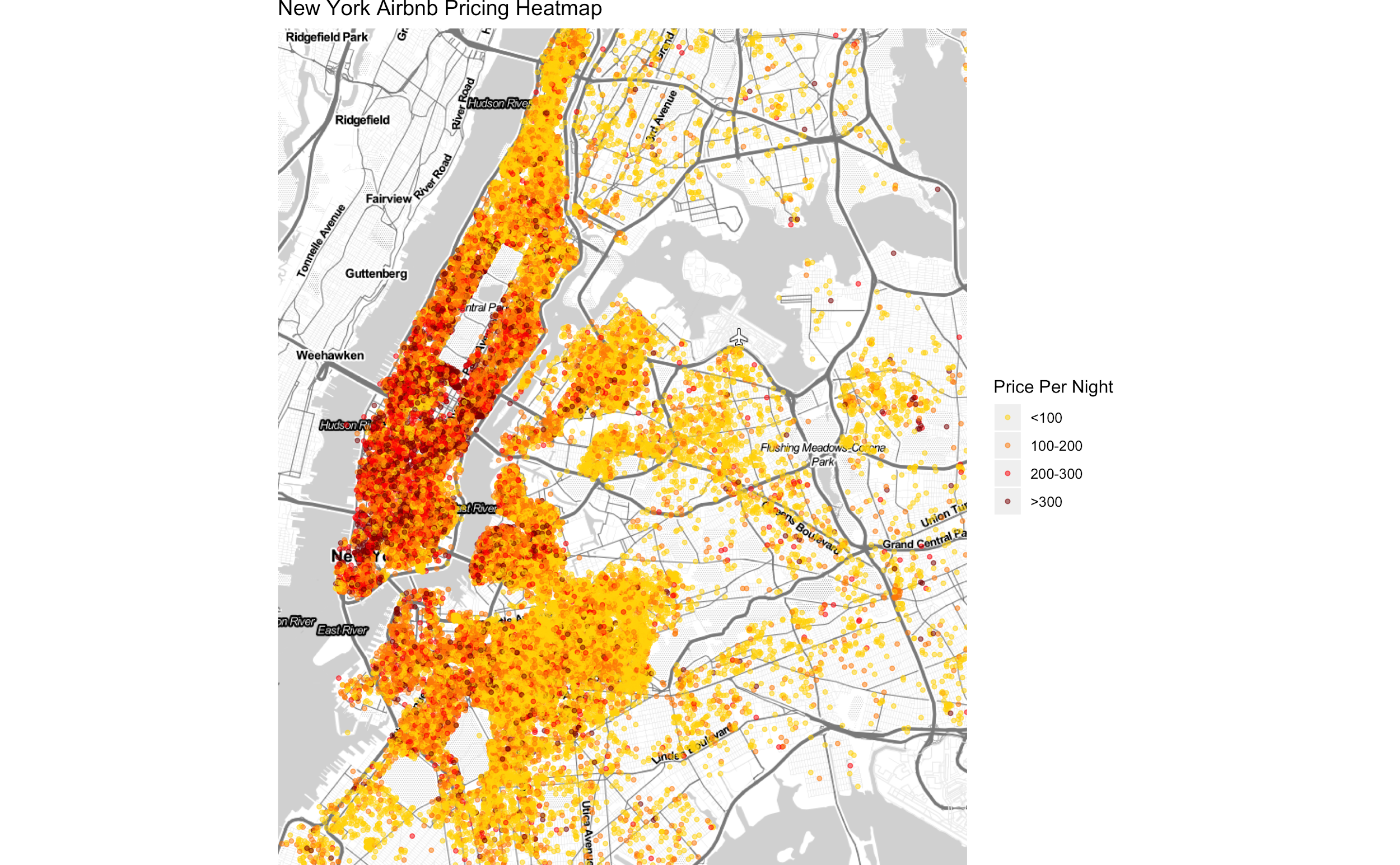
Parallel Coordinate plot allows us to see all the values for each variable, for different neighbourhood groups. Neighbourhood 2 and 3 has the highest price, and highest requirement of minimum nights. Neighbourhood 4 has the most number\_of\_reviews. Neighbourhood 5 has the lowest longitude. Also we say that price and minimum\_nights are negatively related because in the parallel coordinate plot, the lines between these variables are crossing, forming an “X” shape, which indicates a negative correlation.

* + Word cloud of hosting name [Chen Xi]
    - (determine most popular words for house listing)
    - Facet by neighborhood\_group

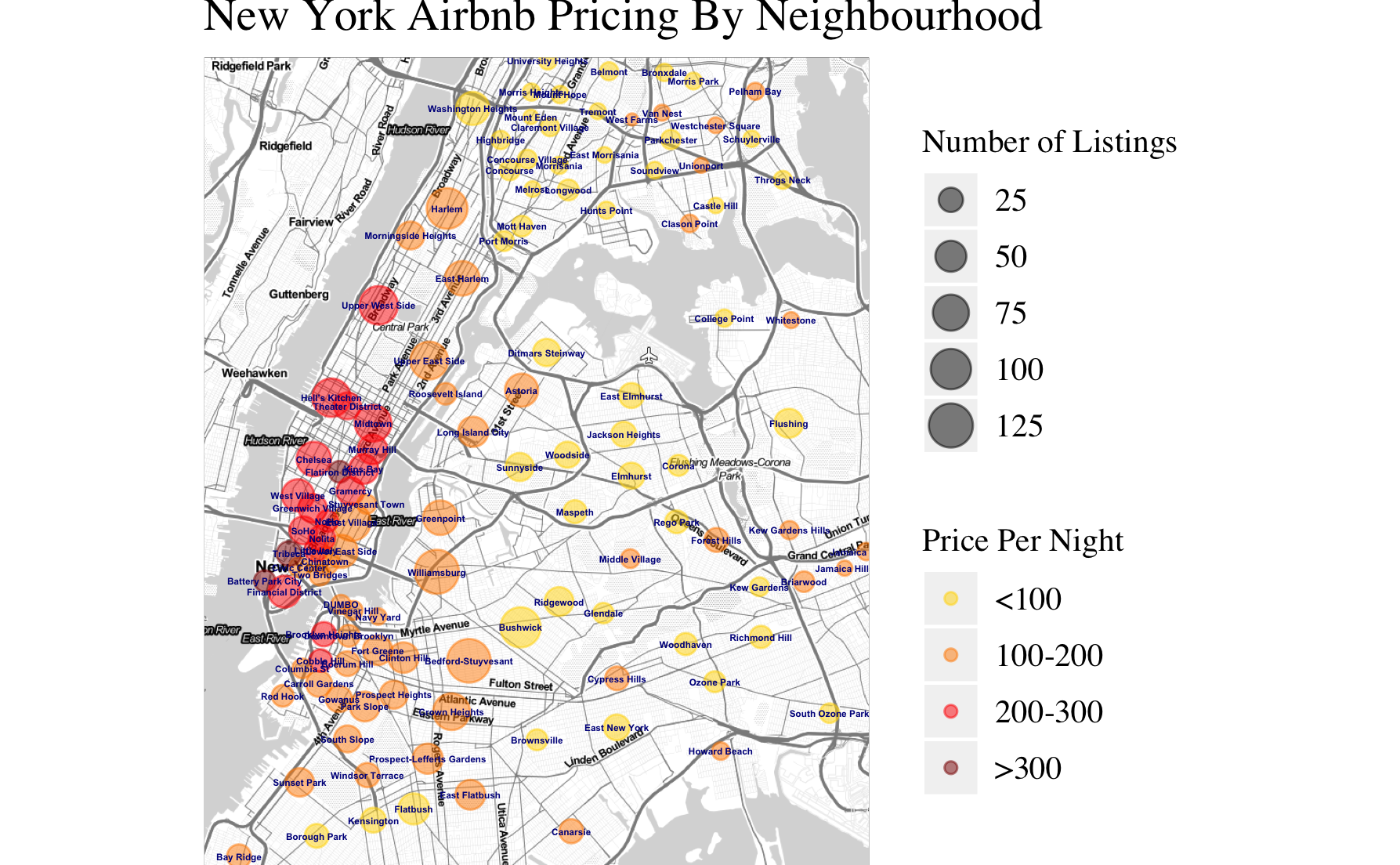


**Conclusion:** From the word cloud, we can see that listings, regardless of which neighbourhood it belongs to, all use the word "apartment", "private" and "bedroom" in their descriptions frequently, indicating that the majority of the listings are private bedrooms of apartments. Also, we can see that most of the listings of NYC airbnb come from Brooklyn, Manhattan and Queens neighbourhoods. In addition, "cozy" is a word that is used frequently in the descriptions for listings from all neighbourhoods and the descriptions of the listings all tend to include the neighbourhoods' names.

* + Map of NYC [Serena]
    - Each neighborhood\_group is colored by mean price range in that area
    - Add dot to each neighborhood\_group, size by number of listings in this neighborhood

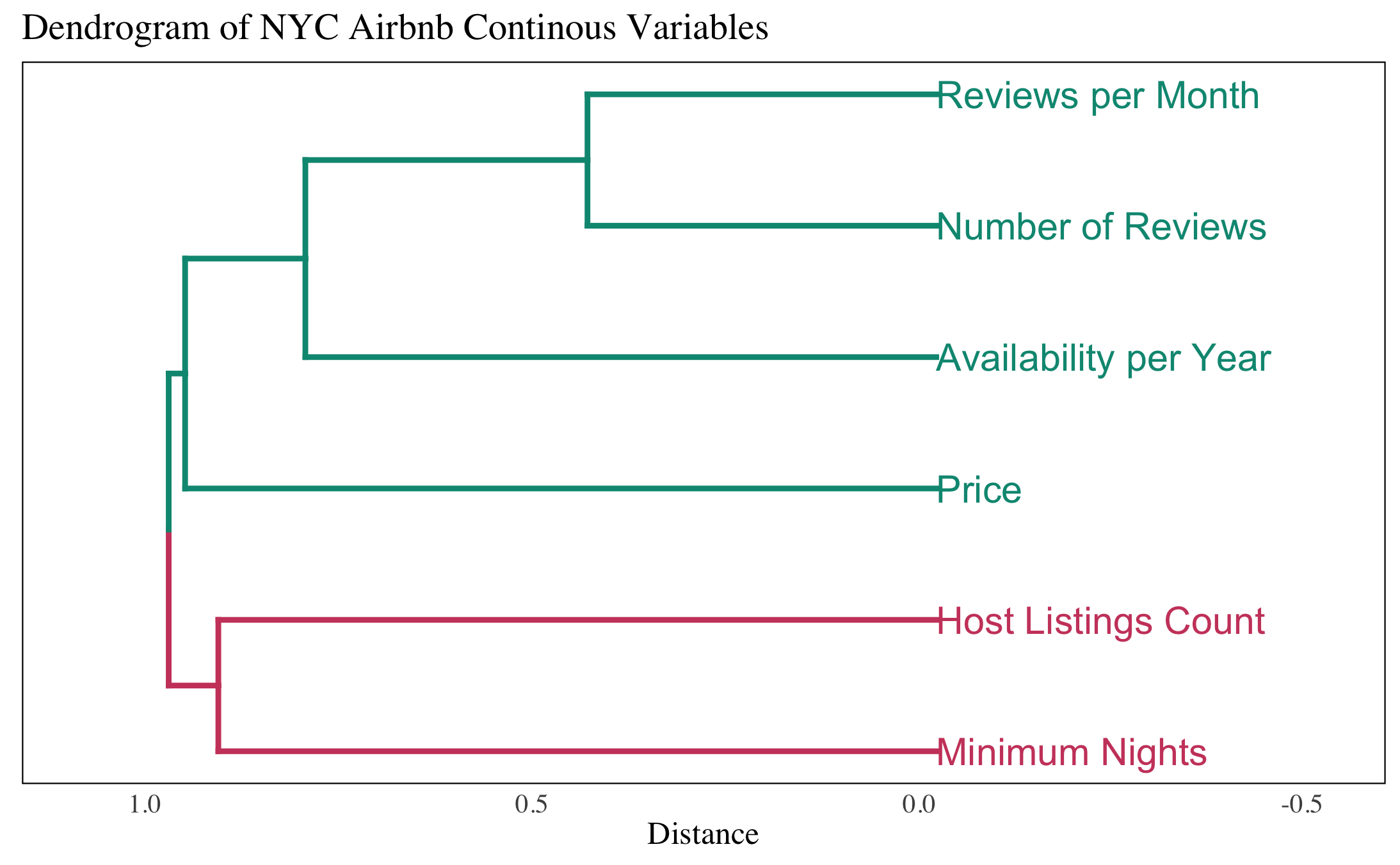


**Conclusion:** This heatmap shows that listings on mid to lower manhattan and river side brooklyn area are the most expensive.



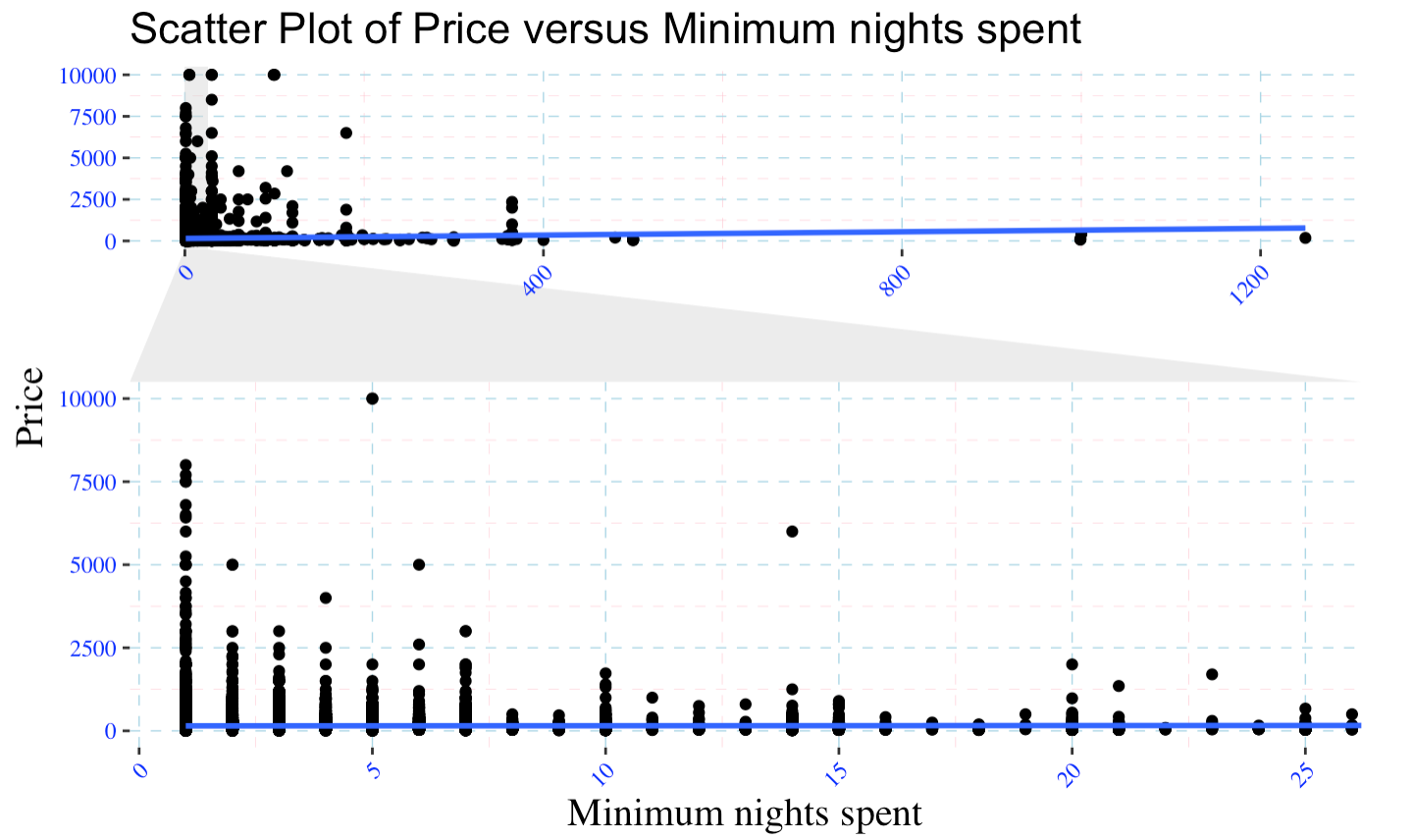
**Conclusion:** Most of the airbnb listings in NYC are located on manhattan and brooklyn. Comparing the dot sizes for each neighbourhood,Harlem, Upper West Size, Upper East Side, Hell’s Kitchen, Bedford Stuyvesant and Williamsburg are the neighbourhoods that have most listings. Price per night can be observed by comparing the color. Tribeca, Battery Park City and Flatiron District are the most expensive neighbourhoods.

* + - Refer to HW10
  + Dendrogram on all numerical variables: [Serena]
    - Refer to HW8 3b
    - See which variables are more related

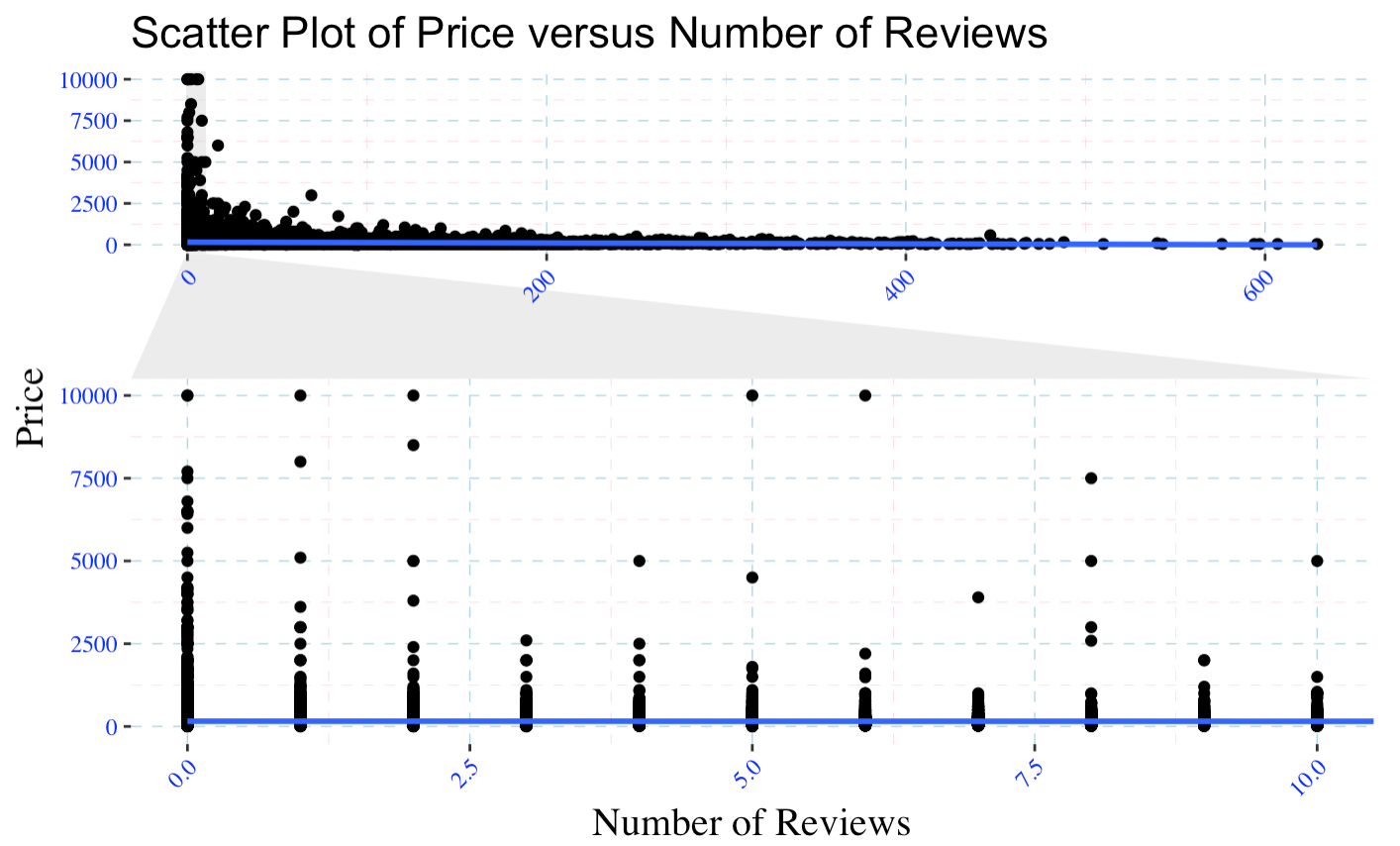


**Conclusion:** The dendrogram of continuous variables in the dataset shows 2 clusters of the variables. Reviews per Month, Number of Reviews and Availability per Year are more correlated with Price. On the other hand, Hosting Listings Count and Minimum Nights are more correlated.

* **Explore how each variable is related to prices (decide after EDA) Due Sunday Dec1**
  + Zoom-in Scatter Plot: [Chen Xi]
    - Relationship between two variables (color/ shape), then zooming in
    - Refer to HW8 7



**Conclusion:** from this plot, we can see that for minimum nights stayed from 0 to 25, the price generally decreases as minimum nights required increases and this is reasonable. For minimum nights required from 25 to around 400, price remains relatively stable and low. While it is not usual for airbnb listings to require minimum nights stayed longer than 400 days, There are several outliers with prices at low values. In addition, a linear regression does not capture the relationship between these two variables very well.



**Conclusion:** from this plot, we can see that there are many more listings with few reviews (around 0 to 10) than those with a large number of reviews (> 200). From the zoom-in scatter plot, we can see that number of reviews roughly decreases as price increases, showing a negative relationship, for 0-10 number of reviews. For number of reviews more than 10, price remains relatively stable. In addition, a linear regression does not capture the relationship between these two variables very well.

* + Linear Regression: [Chen Xi]
    - Geom\_smooth
    - Refer to HW6 problem 4
* **Make Poster + Submit Due Monday (Remote)**

Additional:

