Data Methodology

Step 1: Storyboarding

- Went through the data to get familiarized with it and noted down important fields
- Made a mind map of the various slides of the presentation
- Made a rough template based on this mind map

Step 2: Data Wrangling

- Deleting the columns which have all null values

I have identified the columns which have all null values in it and deleted the columns. Before that I have copied my listings dataset into listing_duplicate. All data manipulation will be done on this dataframe

```
listings_duplicate <- listings
a<- sapply(listings_duplicate, function(y) sum(length(which(i
s.na(y)))))
listings_duplicate <- listings[,colSums(is.na(listings)) < nr
ow(listings)]
length(listings_duplicate)
```

. This removed 6 columns which have entire nulls

- Removing columns which have url's in it

In our dataset we have few columns which have url data of the Airbnb houses, host profile, Airbnb pictures etc., As we don't have anything to do with the url data we can identify these columns and remove them

```
listings_duplicate[,names(listings_duplicate)[grep("url",name
s(listings_duplicate))]] <- NULL
length(listings_duplicate)
```

`

- Deleting column which have same value in all the row

In our dataset we have few columns which have same value in all rows, columns are scrape_id, state, country, country_code. As we are doing analysis on Melbourne city, we don't need columns such as state, country and country_code.

```
unique(listings$scrape_id)
unique(listings$state)
unique(listings$country_code)
unique(listings$country)
listings_duplicate[,c( "scrape_id", "state", "country_code",
"country")] <- NULL
```

- Deleting columns which have more than 70% null values:

Though we have deleted columns which have all null values, we still have columns which have more then 70% of the null data in it. As these nulls will affect our analysis we have two options to overcome with these.

i. We can replace all nulls with the median, mean or mode if those columns are necessary.

ii. We can simply delete these columns if they are not necessary.

In my dataset exploration, these columns will not affect the analysis. So here am choosing the second option to delete these columns.

```
total <- nrow(listings_duplicate)
b <-  round(total * 0.7)
#b
listings_duplicate <- listings_duplicate[,colSums(is.na(listi
ngs_duplicate)) < b]
length(listings_duplicate)
```

- Columns which are not required:

There are few columns which are not required for exploring, so it's a good option to delete those columns and reduce the space.
c("security_deposit","weekly_price","monthly_price","first_review","jurisdiction_names","zipcode","street", "market","cleaning_fee","name","interaction","access","space","notes","summary","description","host_name","host_has_profile_pic","host_verifications","host_neighborhood","require_guest_profile_picture","require_guest_phone_verification","calculated_host_listings_count","host_location","transit","neighborhood_overview","house_rules","host_about","license","requires_license","host_neighbourhod")

In [8]:

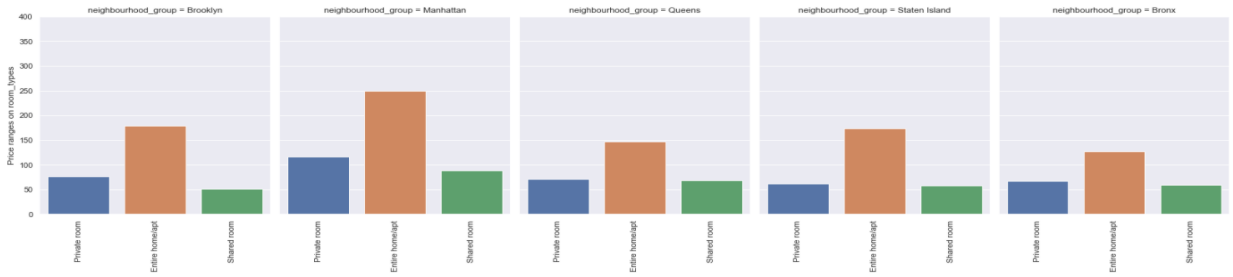
- Redundant Columns

There might be columns which will provide same information. As this leads to data redundancy we have to delete these columns

```
listings_duplicate[,names(listings_duplicate[(duplicated(t(li
stings_duplicate)))])] <- NULL
```

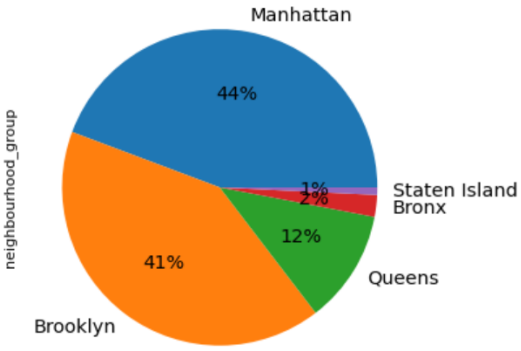
Step 3: Data Analysis

- Checked neighborhood grouped wise distribution of price and room type
- Manhattan is costliest overall and the cheapest are:
    a. Entire apt: Bronx
    b. Private room: Staten Islands
    c. Shared room: Brooklyn

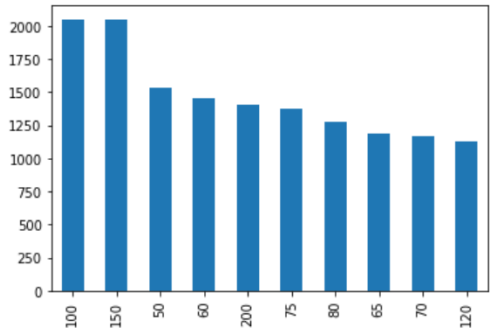


- By the two scatterplots of latitude vs longitude we can infer there's is very less shared room throughout NYC as compared to private and Entire home/apt.
- 95% of the listings on Airbnb are either Private room or Entire/home apt. Very few guests had opted for shared rooms on Airbnb.
- Also, guests mostly prefer this room types when they are looking for a rent on Airbnb as we found out previously in our analysis.

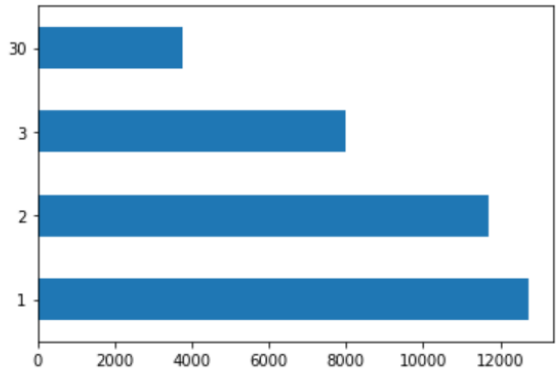
- Manhattan has the highest neighborhood group where as Staten Island has the least.
- Brooklyn has 41% of the entire neighborhood group



- Almost 2k+ airbnb's has a price of 100 dollars and 150 dollars each respectively.
- 1.5k airbnb's have around 50 dollars price.



- We can observe that most of almost 12k people used 1 night stay in airbnb.
- 11k people choose 2 night stay while 7k choose 3 night stay.
- Almost 3.7k stayed upto a month.



Step 4: Presentation

- Made the presentation adhering to best practices and pyramid principle