Capstone Project – 1

Airbnb Dataset Analysis

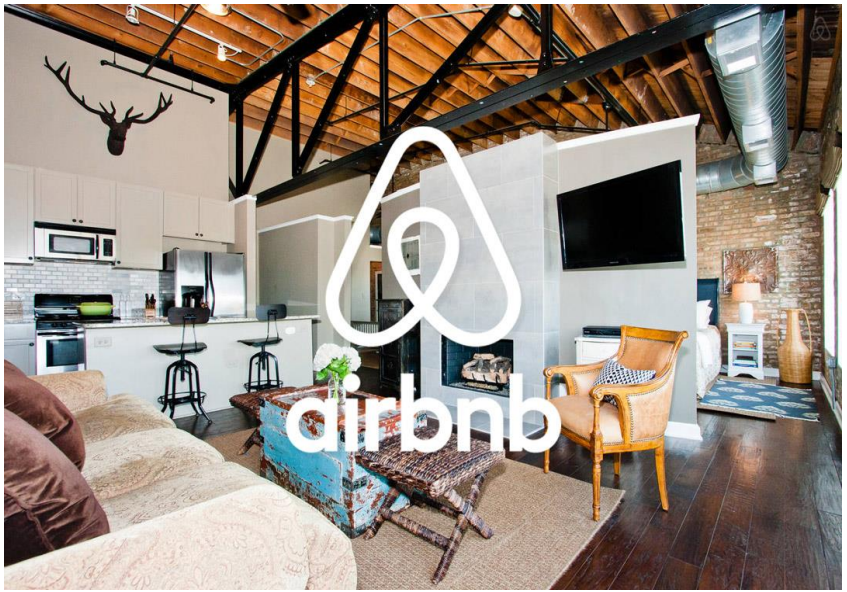
By – Shri Prakash Yadav

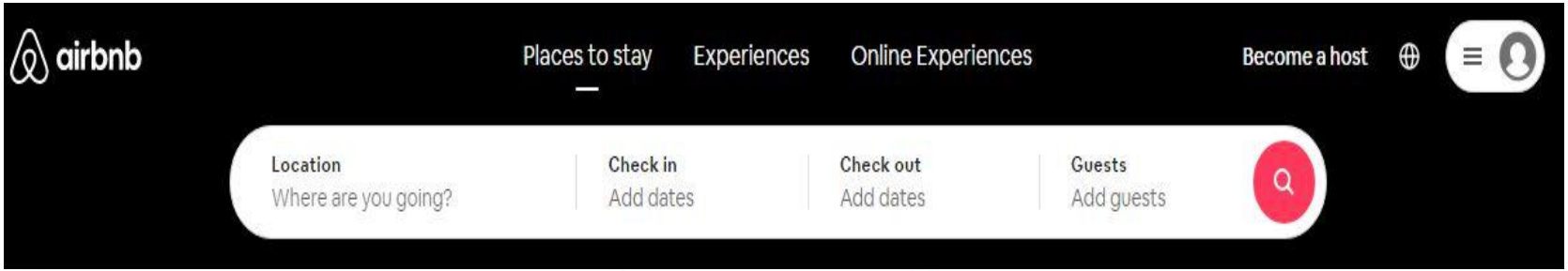
Points of Discussion

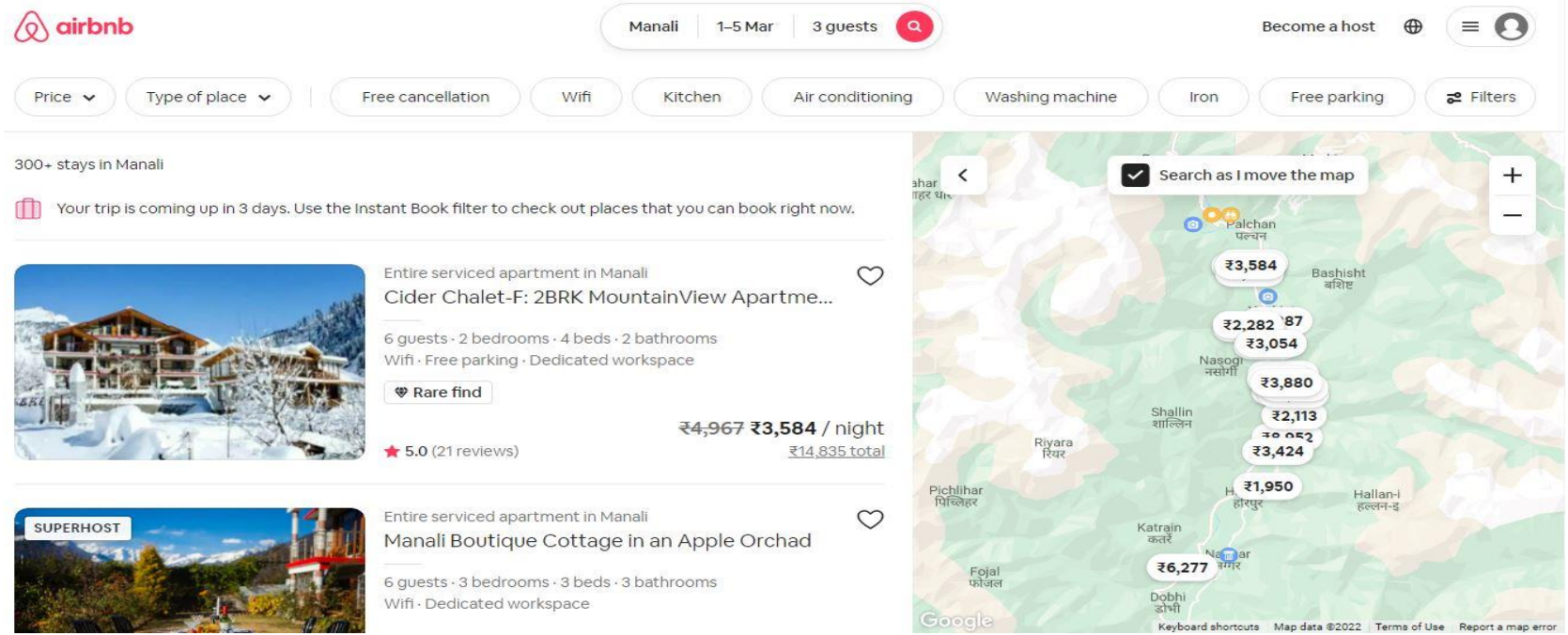
- What is airbnb
- Data Summary
- Visits per neighbourhood_groups
- Top and bottom neighbourhoods
- Room_types
- Correlation
- Price Distribution
- Outliers Removal
- Conclusion
- Challenges

What is Airbnb?

- Airbnb is an online platform which basically connects guests and hosts. Guests is someone who wants to book home stays or rent apartments in any area or neighbourhood mainly for vacation and tourism purpose. Host is someone who wants to rent out his apartment, room or house. Airbnb platform has become very popular service and is being used by a lot of users throughout the world.




airbnb
Places to stay
Experiences
Online Experiences
Become a host
Location
Where are you going?
Check in
Add dates
Check out
Add dates
Guests
Add guests


airbnb
Manali
1–5 Mar
3 guests
Become a host
Price
Type of place
Free cancellation
Wifi
Kitchen
Air conditioning
Washing machine
Iron
Free parking
Filters
300+ stays in Manali
Your trip is coming up in 3 days. Use the Instant Book filter to check out places that you can book right now.
Entire serviced apartment in Manali
Cider Chalet-F: 2BRK MountainView Apartme...
6 guests · 2 bedrooms · 4 beds · 2 bathrooms
Wifi · Free parking · Dedicated workspace
Rare find
₹4,967 ₹3,584 / night
5.0 (21 reviews)
₹14,835 total
SUPERHOST
Entire serviced apartment in Manali
Manali Boutique Cottage in an Apple Orchad
6 guests · 3 bedrooms · 3 beds · 3 bathrooms
Wifi · Dedicated workspace
Search as I move the map
Palchan
₹3,584
Bashisht
₹2,282
₹3,054
₹3,880
₹2,113
₹3,424
₹1,950
Shallin
Riyara
Pichlihar
Katrain
₹6,277
Fojal
Dobhi
Google
Keyboard shortcuts
Map data ©2022
Terms of Use
Report a map error

Data Summary

This data set consists of 48895 rows and16 columns which are listed below:

Numeric:-

- **ID** – Each listing has a unique id.
- **HOST_ID** – Each host is assigned with a unique host_id.
- **Latitude** – Each host's angular distance from north or south of the earth's equator.
- **Longitude** – Position of a host with respect to the north-south running lines on earth.
- **Price** – Price of any rented space.
- **Minimum_nights** – minimum number of nights for which booking is open.
- **number_of_reviews** – Number of reviews a host has received.

Data Summary

- **reviews_per_month** – how many reviews a host has received per month.
- **calculated_host_listing_count** – It is basically the number of time a particular host has used airbnb in that dataset.
- **availability_365** – number of days a host is available in 365 days.

- **DateTime object:-**
- **last_review** – date of the last review received.

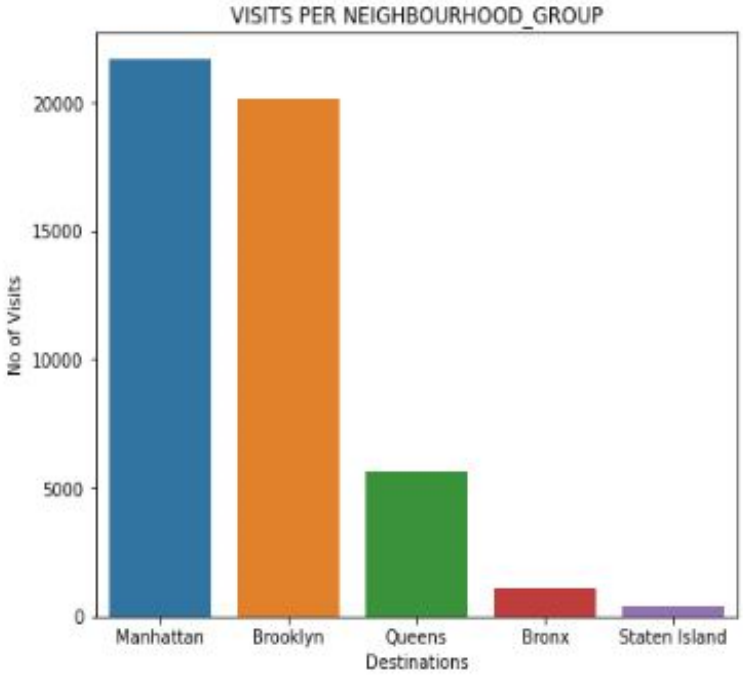
- **String objects:-**
- **name** – name of the rented room or appartment.
- **host_name** – name of the host

Data Summary

- **neighbourhood_group** – neighbourhood_group is name by which the group of neighbourhoods surrounding any host is known.
- **neighbourhood** – neighbourhood is the neighbourhood of any host.
- **room_type** – type of room a host is renting.

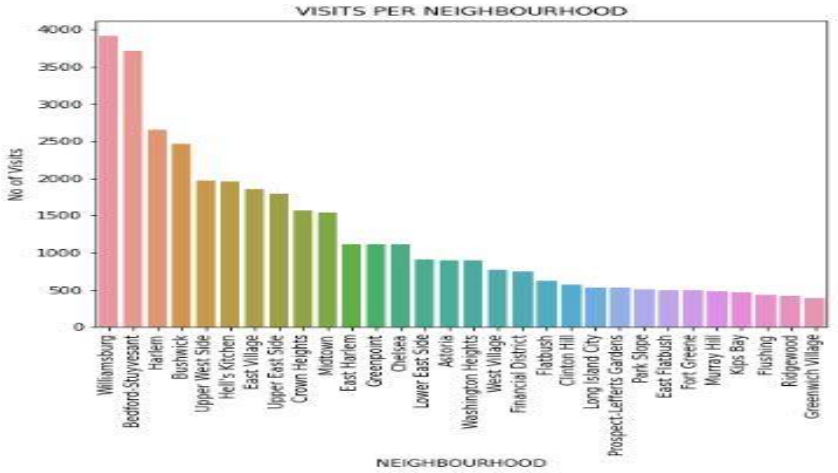
Visits per neighbourhood_groups

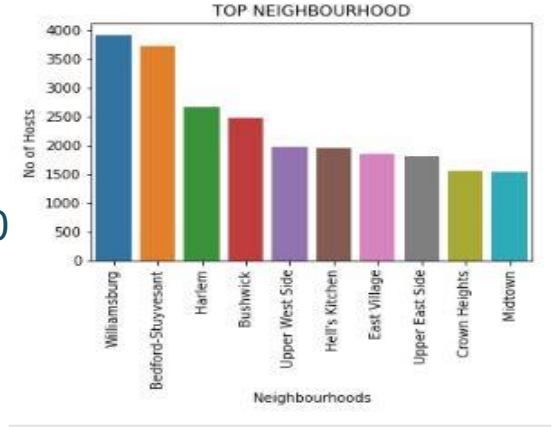
- We have 5 unique neighbourhood_groups Brooklyn, Manhattan, Queens, Bronx and Staten Island.
- Manhattan is the most visited neighbourhood_group with total visits of 21661 followed by Brooklyn 20104. Staten Island is the least visited Neighbourhood_group with total visit of 373.

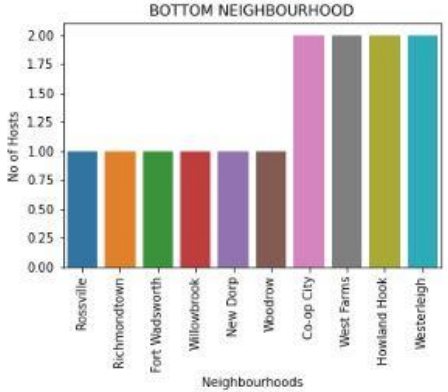


Top and Bottom Neighbourhoods

- We have 221 unique neighbourhoods.
- Williamsburg, Bedford-Stuyvesant, Harlem are the top visited neighbourhoods whereas Rossville, Richmondtown and Fort-Wadsworth are least visited neighbourhoods.
- Total visit to williamsburg is 3920 whereas Rossville has single visit.

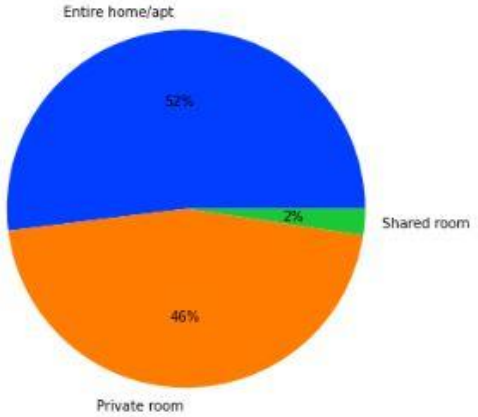


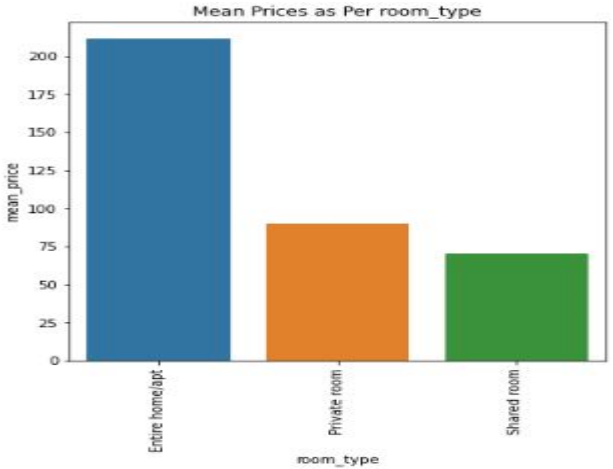




Room_types

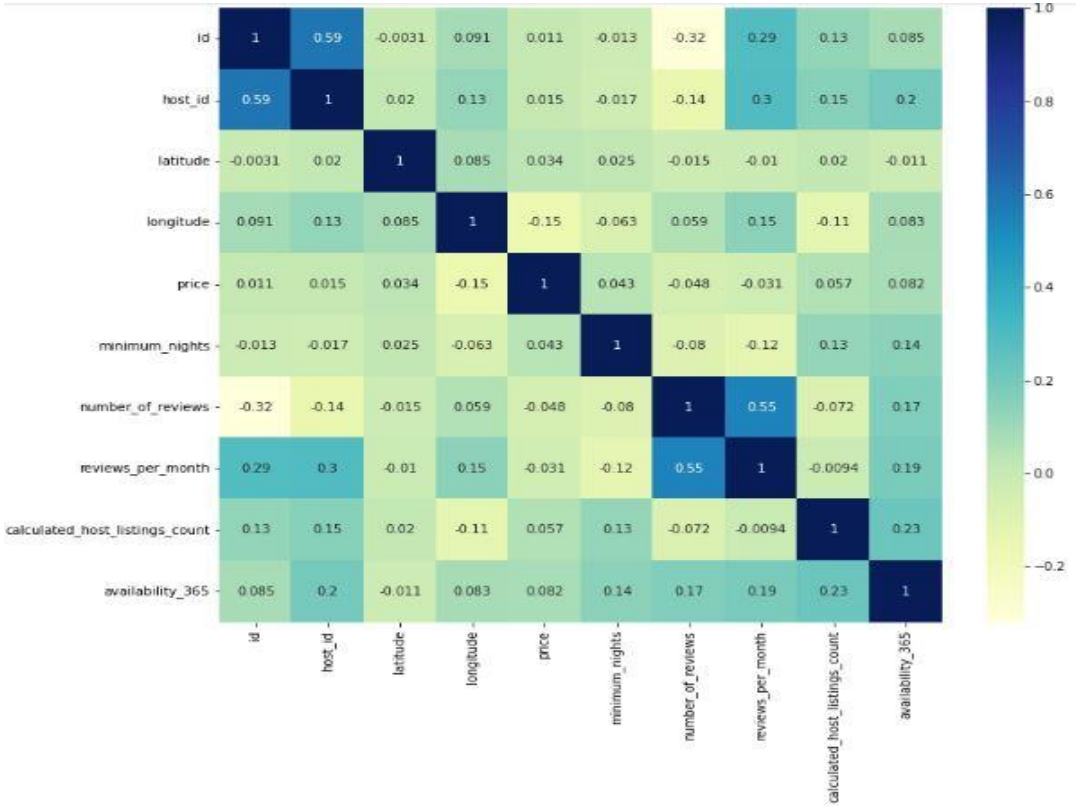
- We have three room types entire_home/apt, shared room and Private room with constitution of 52% ,2% and 46% respectively.
- Entire home/apt is the costliest room_type with a mean price of 211 whereas Shared room is the cheapest with a mean price of 70.Mean price for private room type is 89.





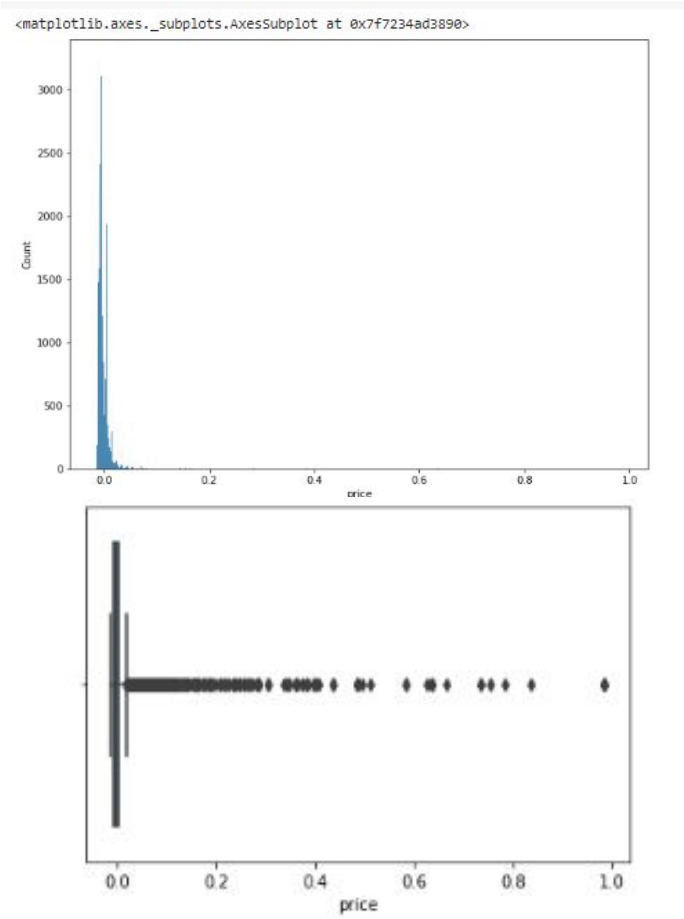
Correlation

- Variables does not show any correlation at all.
  We can spot certain darker regions like reviews_per_month with number_of_reviews and id with host_id but these are not anyway useful.
- We can see that none of the numerical feature affects the price.



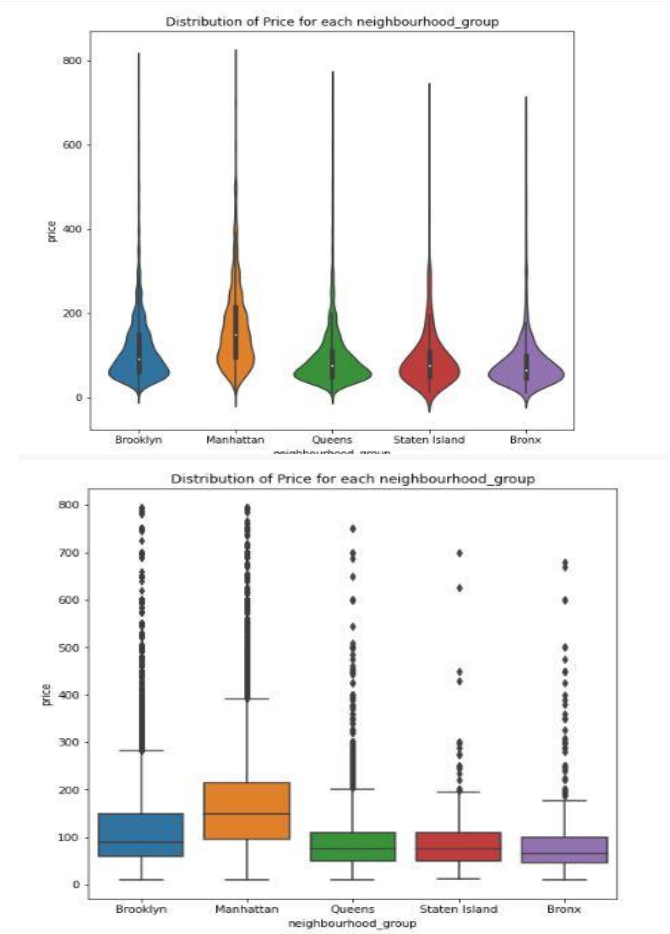
Price Distribution

- In the PDF plot price we can see that it is positively skewed.
- Min, mean and max prices are 0,152 and 10000 respectively.
- The 99th percentile of price is 799.



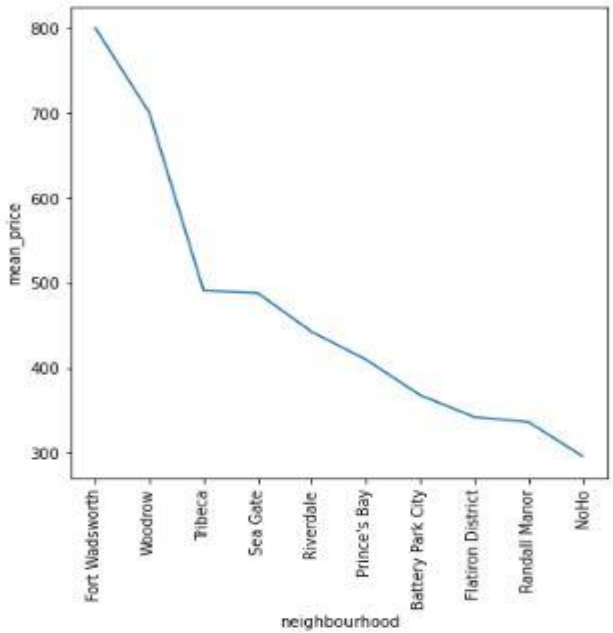
Price Distribution

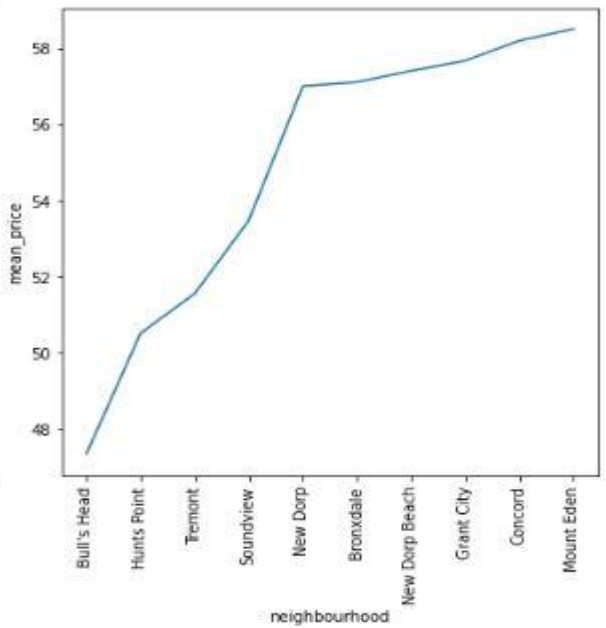
- This is the price distribution we can see that Manhattan has more spread and whereas other neighbourhood_groups have more peakedness.
- Manhattan has the highest mean price 196 and Bronx has the lowest mean price 99.
- For each neighbourhood when we reach maximum prices, data points are rare specially In Queens , Staten Island and Bronx. These could be outliers.



Price Distribution

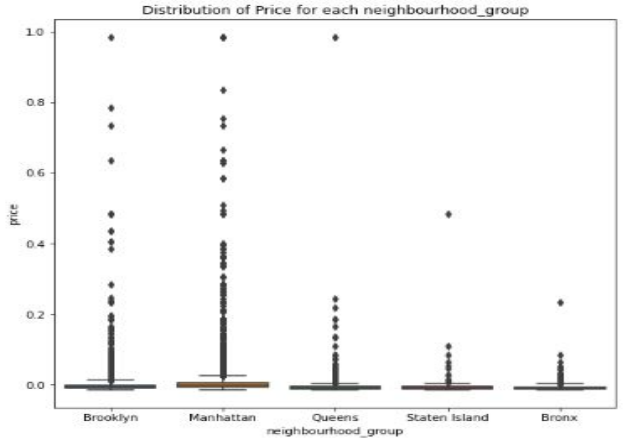
- Fort-Wadsworth neighbourhood has the maximum mean price of 800 whereas Bull's Head has the minimum mean price of 47.

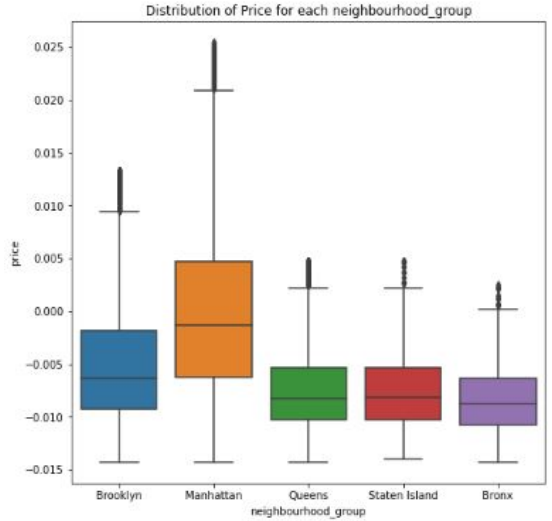




Outliers Removal

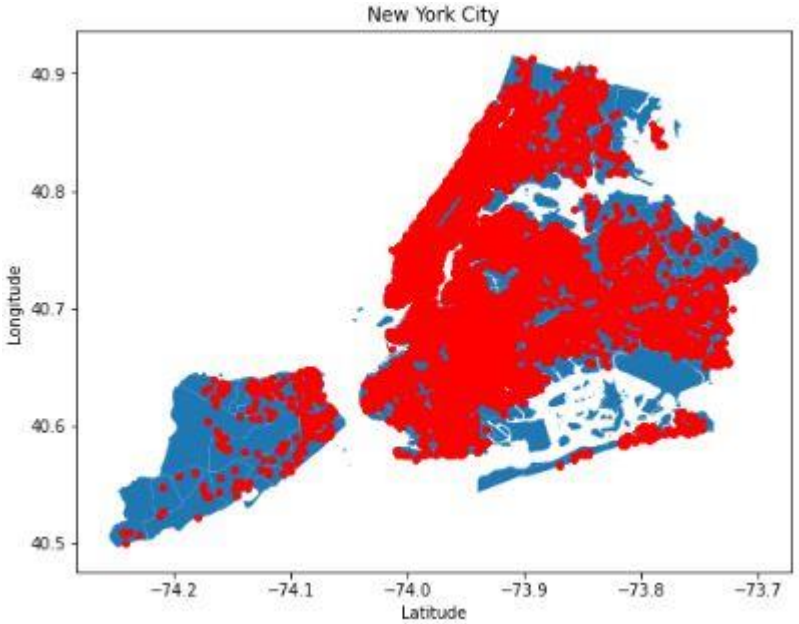
- First image is after standardisation.
- We can clearly see outliers.
- After removing outliers distribution looks like below plot.
- After removing outliers we are left with 46154 data points.





Outliers removal

- I have plotted the longitudes
- and latitudes on New York City
- map as we can see that
- we don’t have any point
- extreme points.



Conclusion

- I was able to find certain trends but was not able to find strong relationships. This could be possible because the dataset was large but most of the features were not relevant.
- Features were not having any strong dependencies. Still I tried to pick certain insights from data like price mostly depends on categorical features not on any numerical feature.

Challenges

- Finding relationship between features was a challenge as most of the features were irrelevant.

THANK YOU