- Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation.

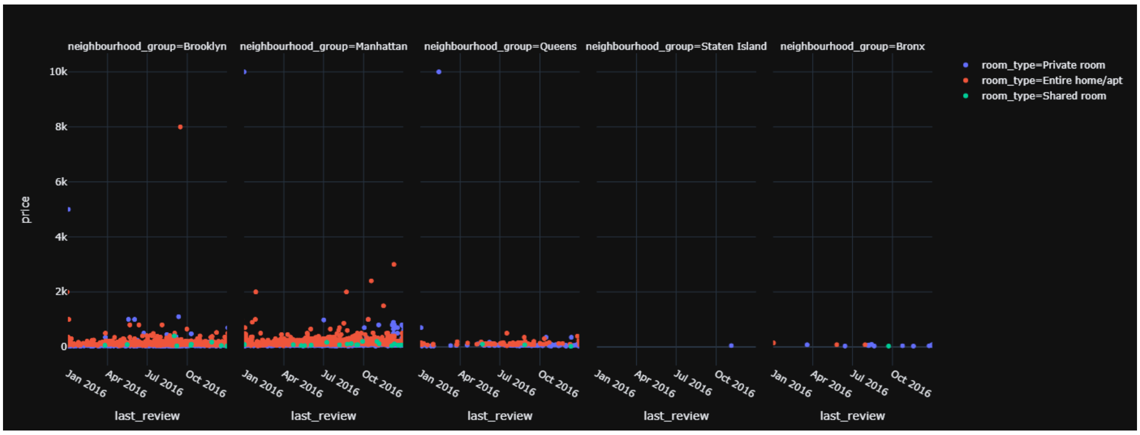
    - ➢ The main objective of this model is to predict the prices of apartments in New York city on Airbnb.

- Brief description of the data set you chose and a summary of its attributes.

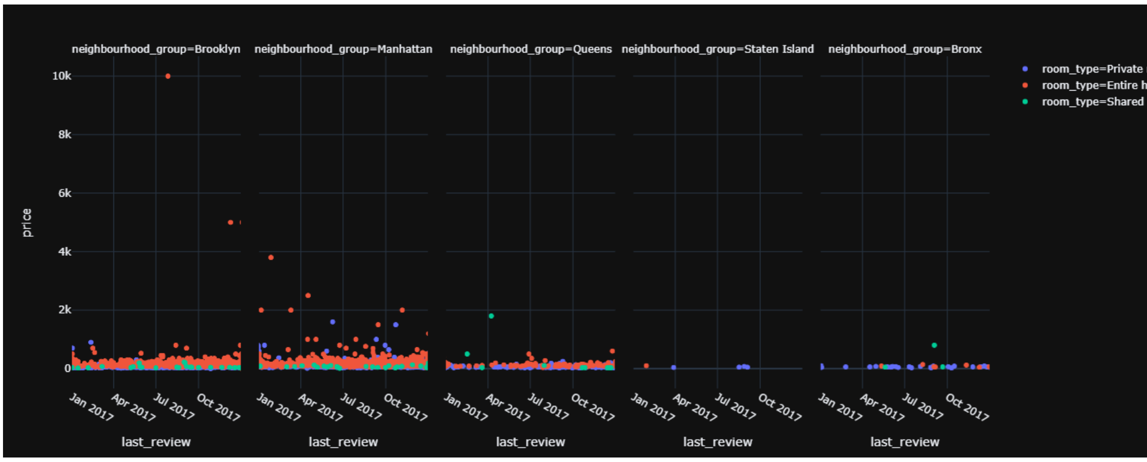
    - ➢ The dataset chosen for this report was Airbnb’s open dataset from 2019 house listing in New York city.
    - ➢ The dataset has 16 following attributes: -
        - ▪ Name: name of apartment
        - ▪ Id: serial number for the data
        - ▪ Host_name: The name of person or organisation to which the apartment belongs.
        - ▪ Host_id: unique number provided to the host
        - ▪ Latitude and longitude: for the position of apartments on map.
        - ▪ neighbourhood group: area to which the apartment is located.
        - ▪ Neighbourhood: Area in which apartment is located
        - ▪ room_type: Listed room types
        - ▪ price: Price of the room
        - ▪ minimum_nights: Amount of nights stayed
        - ▪ number_of_reviews
        - ▪ last_review: Latest review date
        - ▪ reviews_per_month: number of reviews per month
        - ▪ calculated_host_listings_count: amount of listing per host
        - ▪ availability_365: number of days when listing is available for booking

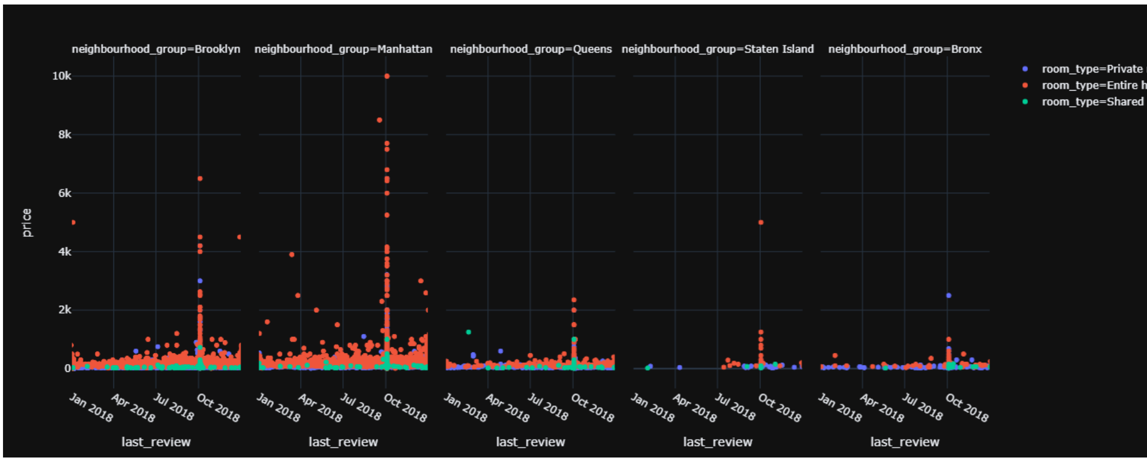
- Brief summary of data exploration and actions taken for data cleaning and feature engineering.

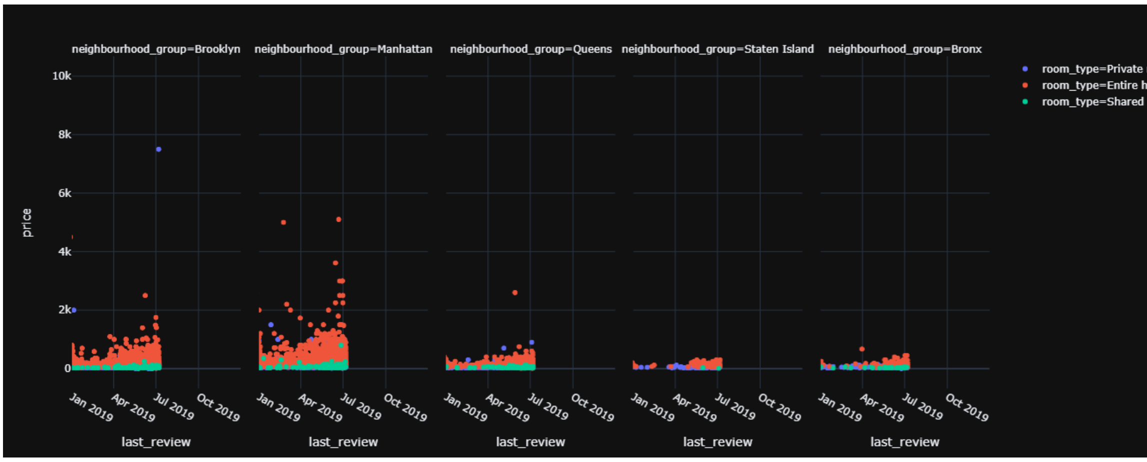
  - The null values were cleaned by apply the mean values to them
  - Datatype of last_review was chaged to datetime for more understanding

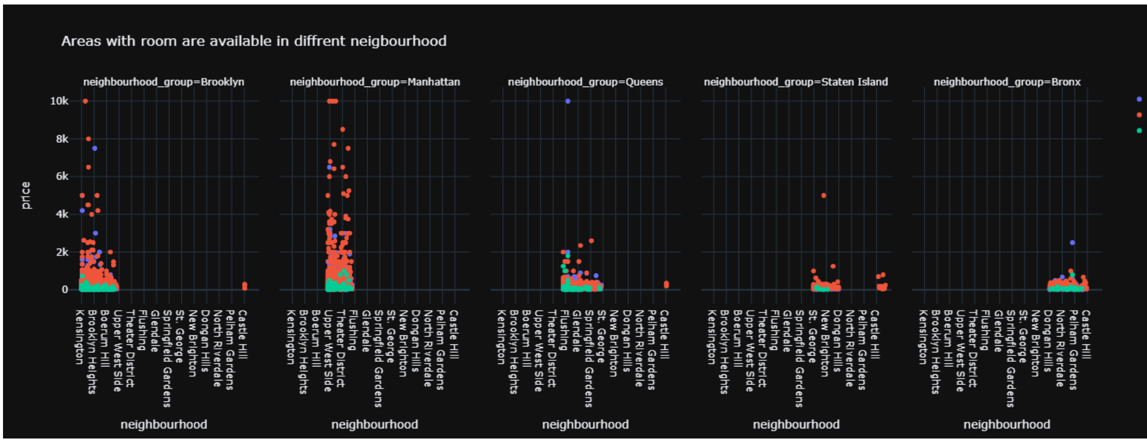


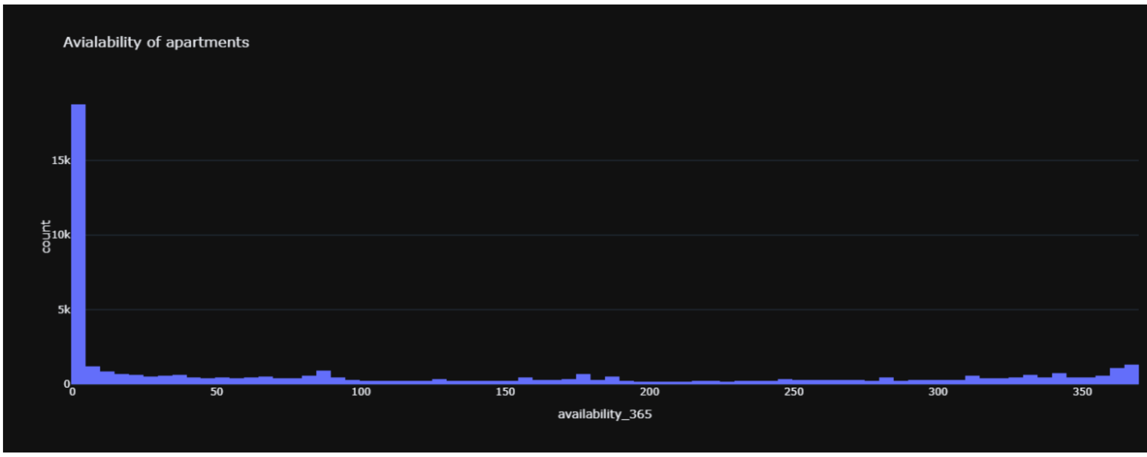
  -

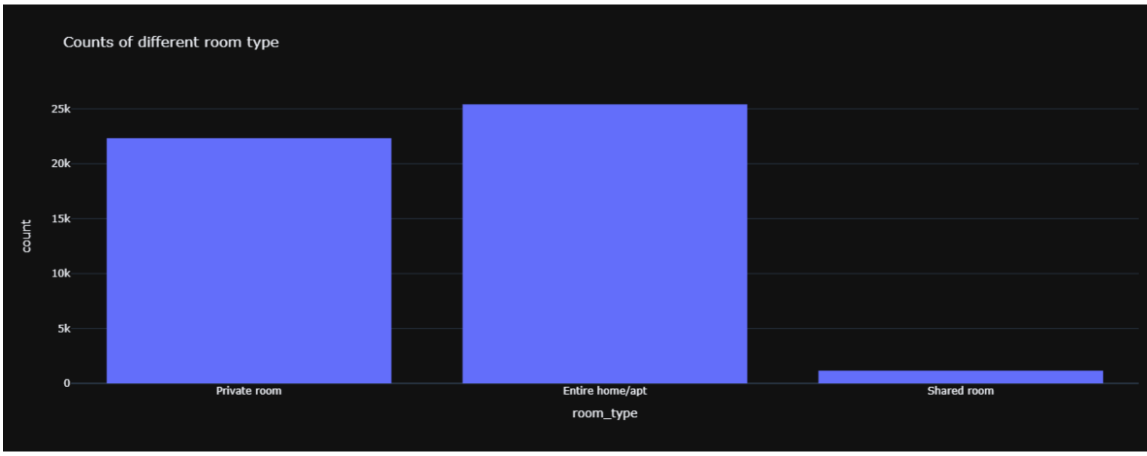


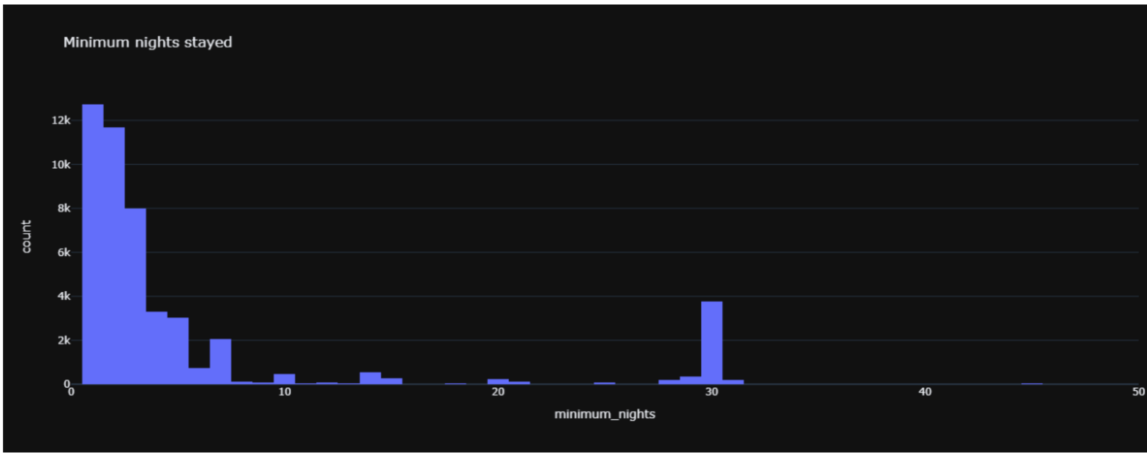


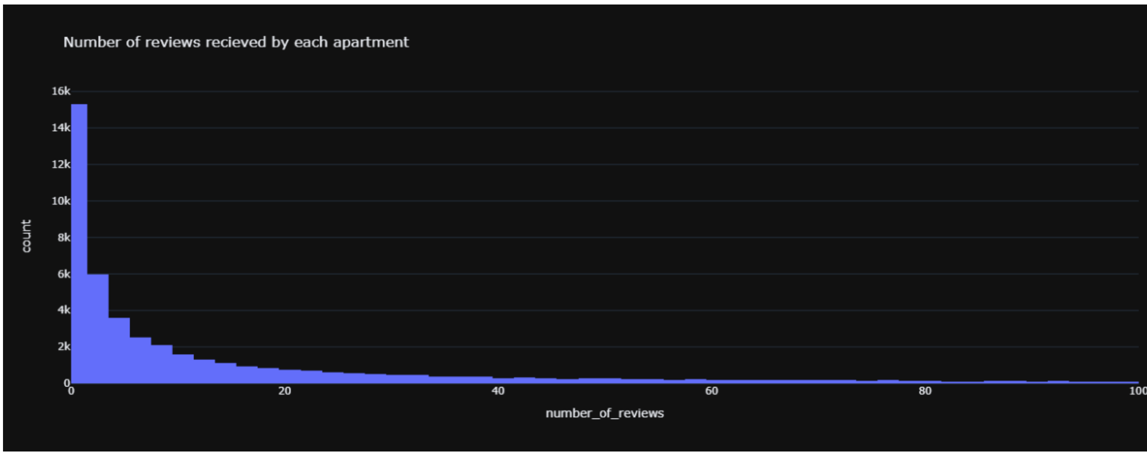

neighbourhood_group=Brooklyn
neighbourhood_group=Manhattan
neighbourhood_group=Queens
neighbourhood_group=Staten Island
neighbourhood_group=Bronx
room_type=Private
room_type=Entire h
room_type=Shared
price
10k
8k
6k
4k
2k
0
Jan 2019
Apr 2019
Jul 2019
Oct 2019
last_review

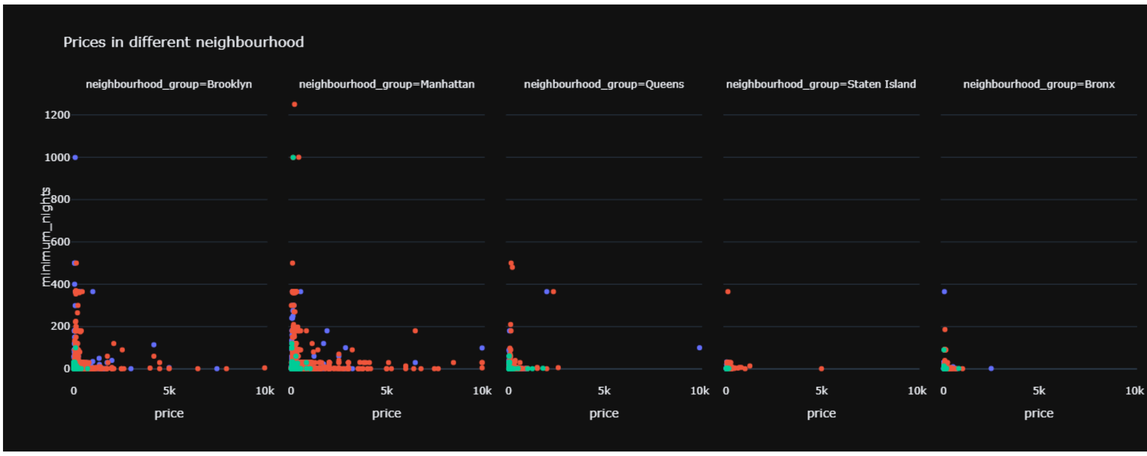

Areas with room are available in diffrent neigbourhood
neighbourhood_group=Brooklyn
neighbourhood_group=Manhattan
neighbourhood_group=Queens
neighbourhood_group=Staten Island
neighbourhood_group=Bronx
price
10k
8k
6k
4k
2k
0
Kensington
Brooklyn Heights
Boerum Hill
Upper West Side
Theater District
Flushing
Glendale
Springfield Gardens
St. George
New Brighton
Dongan Hills
North Riverdale
Pelham Gardens
Castle Hill
neighbourhood


Avialability of apartments
count
15k
10k
5k
0
0
50
100
150
200
250
300
350
availability_365

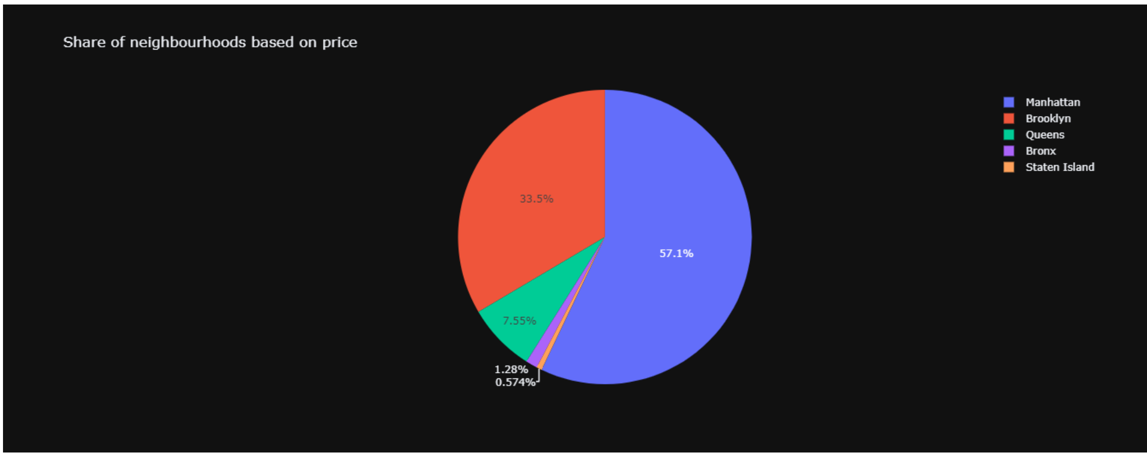

Counts of different room type
count
25k
20k
15k
10k
5k
0
Private room
Entire home/apt
Shared room
room_type


Minimum nights stayed
count
12k
10k
8k
6k
4k
2k
0
0
10
20
30
40
50
minimum_nights

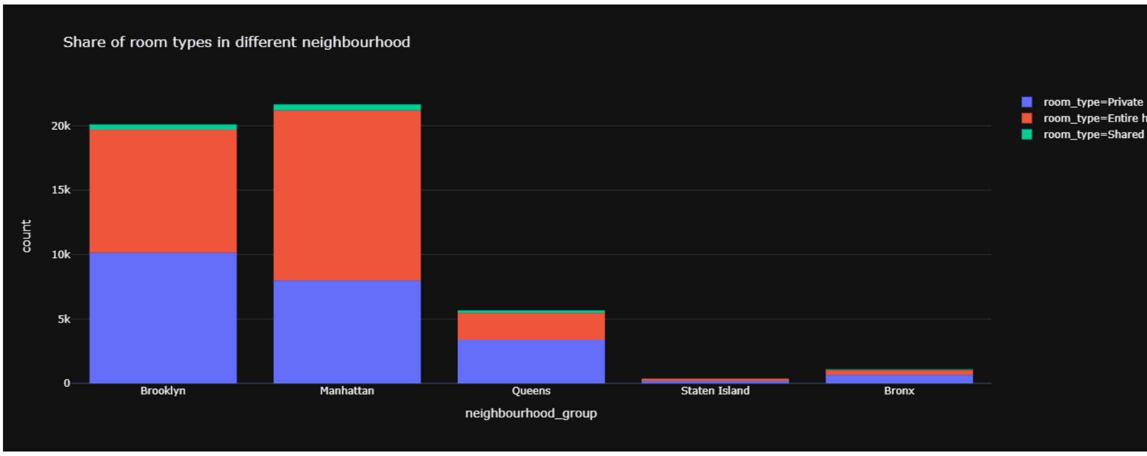

Number of reviews recieved by each apartment
count
16k
14k
12k
10k
8k
6k
4k
2k
0
0
20
40
60
80
100
number_of_reviews



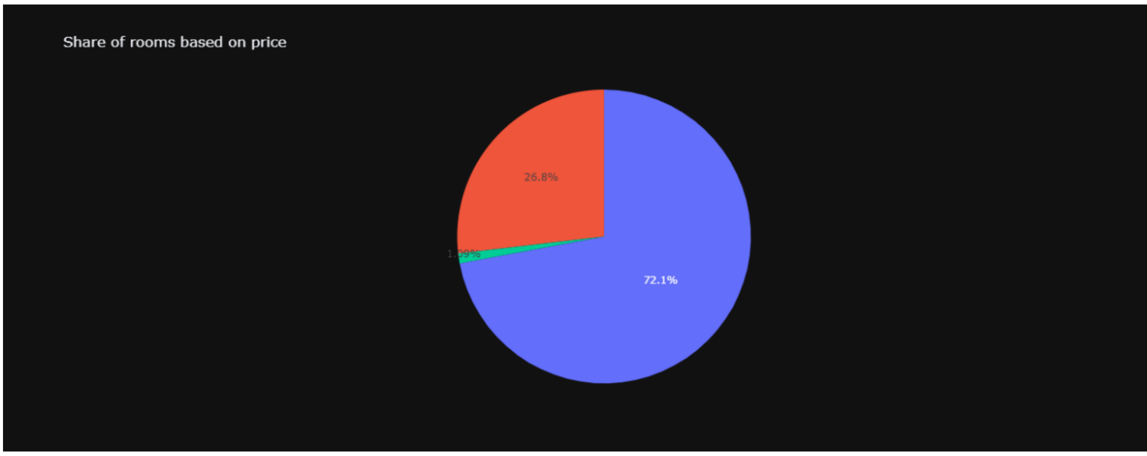
Prices in different neighbourhood

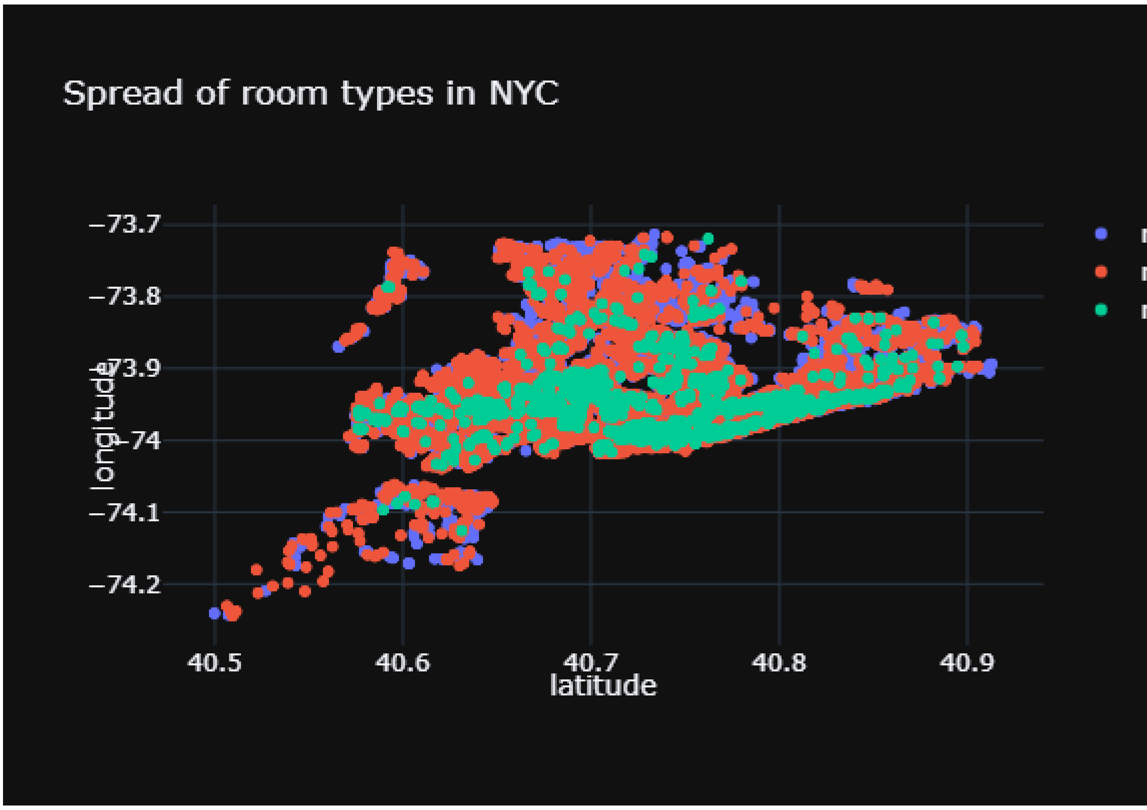


Share of neighbourhoods based on price



Share of room types in different neighbourhood





- Summary of training at least three linear regression models which should be variations that cover using a simple linear regression as a baseline, adding polynomial effects, and using a regularization regression. Preferably, all use the same training and test splits, or the same cross-validation method.

  ➢ The dataset was split into 70:30 ie, 70% for training and 30% for testing

  ➢ Linear regression:

- Mean squared error:39962614951181760.00000
- Coefficient of determination:-664107546751.53162

- Ridge regression:
  - 53649.20341
  - Coefficient of determination:0.10845
- Lasso regression:
  - Mean squared error: 56027.12505
  - Coefficient of determination: 0.06893
- Cross validation:

  Cross validation score of linear regression =-492005747090.12537

  Cross validation score of ridge regression = 0.10921

  Cross validation score of lasso regression = 0.07121

- A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability.

  - The model chosen for this dataset is ridge regressor due to the reason that it takes more features into account and has the better overall performance compared to the other two models

- Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your linear regression model.

  - The key findings that I saw from my linear regression model was even though it highly simple most of the time it is inefficient this due to the fact that in this dataset all the point where scatter and the dimensionality was high due to this reason fitting a simple straight line was not possible.
  - The mean squared error value is very high indicating that the straight line is pass through very less points.

    - ➢ The negative cross validation score shows that the model is struggling to find a relationship to find between the features and produce a straight line

- Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model adding specific data features to achieve a better explanation or a better prediction.

    - ➢ The next step of improving this model would be to consider other models like decision tree regressor, random forest regressor etc.
    - ➢ The model could also be improved by hyperparameter tuning.
    - ➢ Finding the important feature using feature importance can also improve the model a lot.