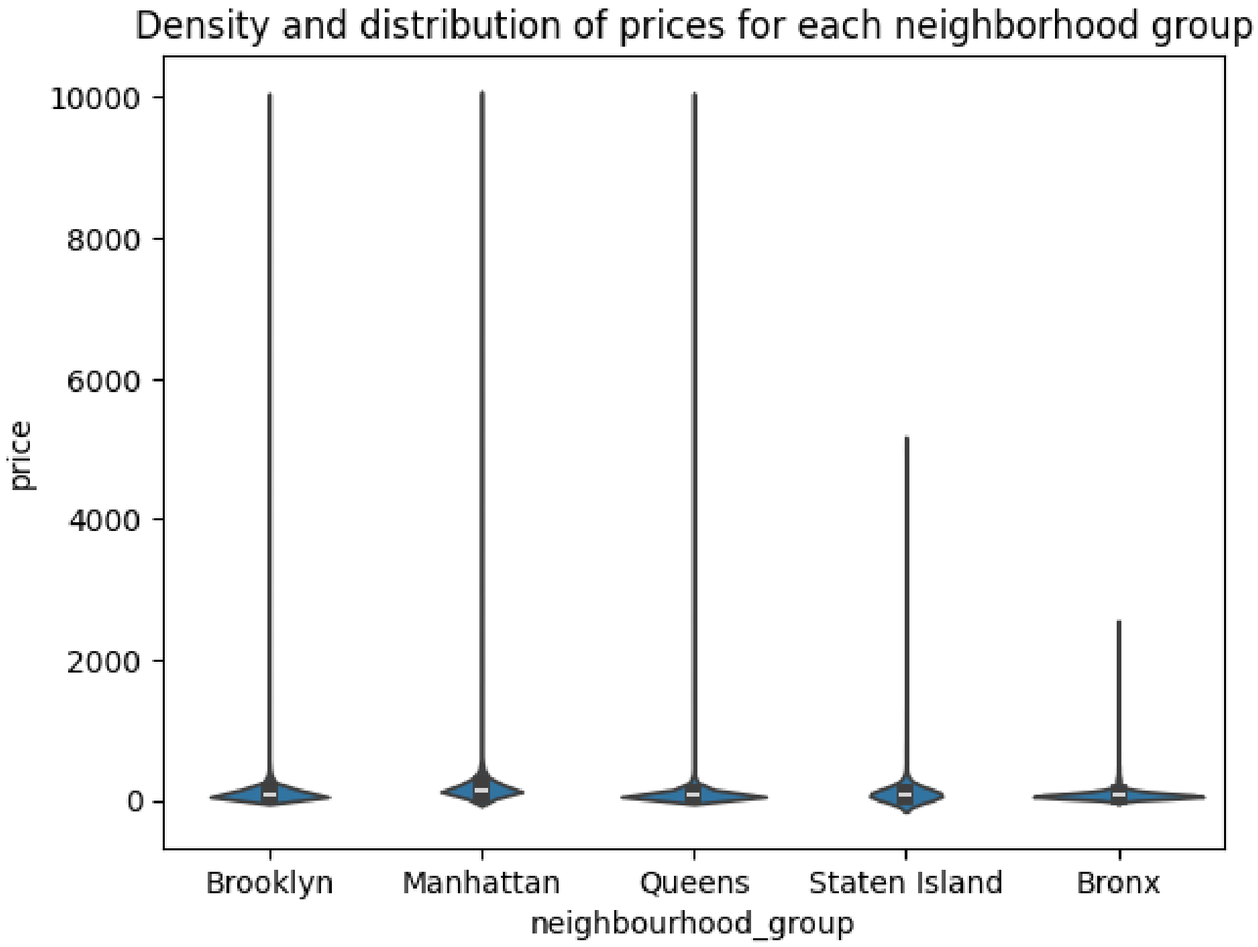
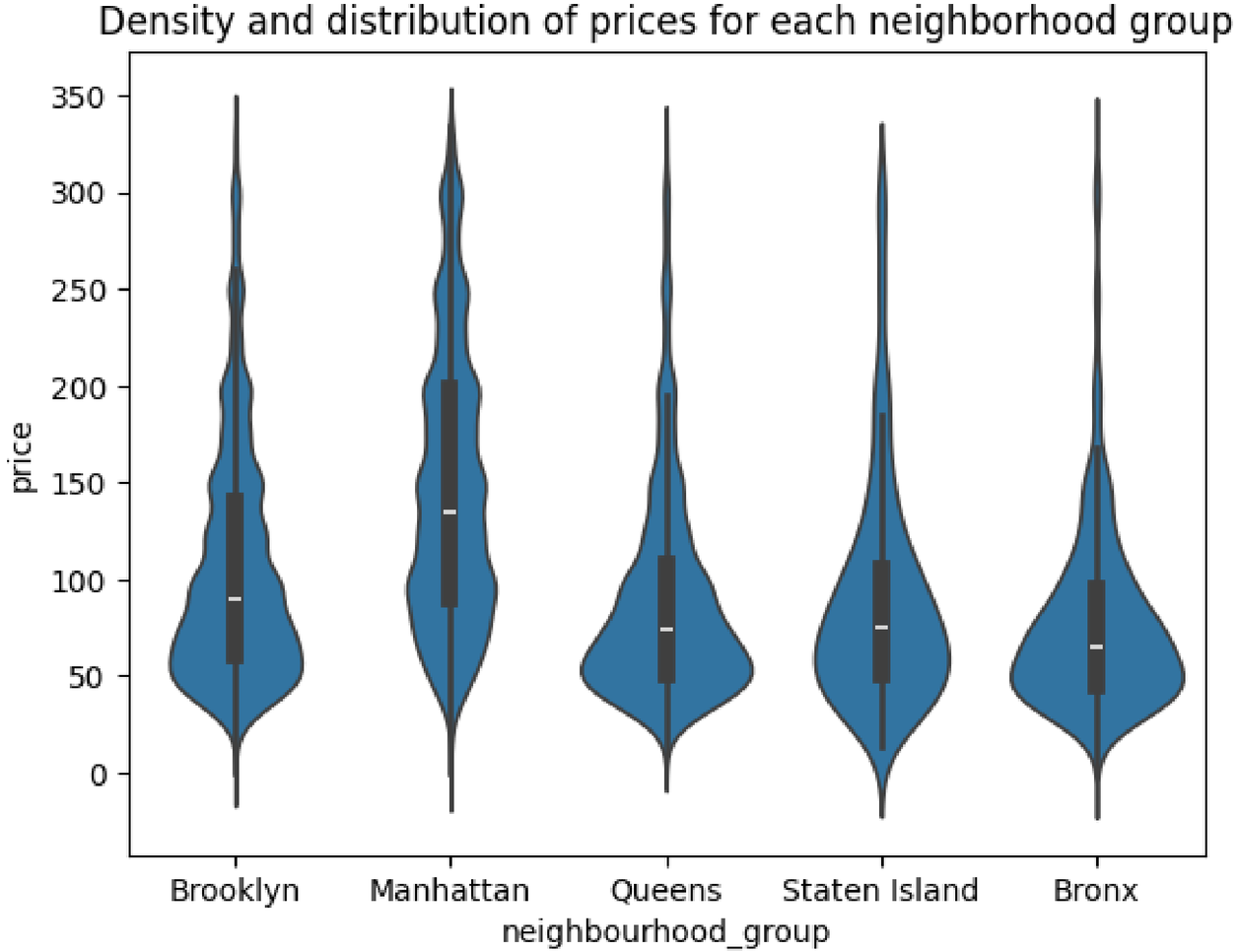
# NEW YORK AIRBNB DATA

**Group 10**

# DATA PREPROCESSING

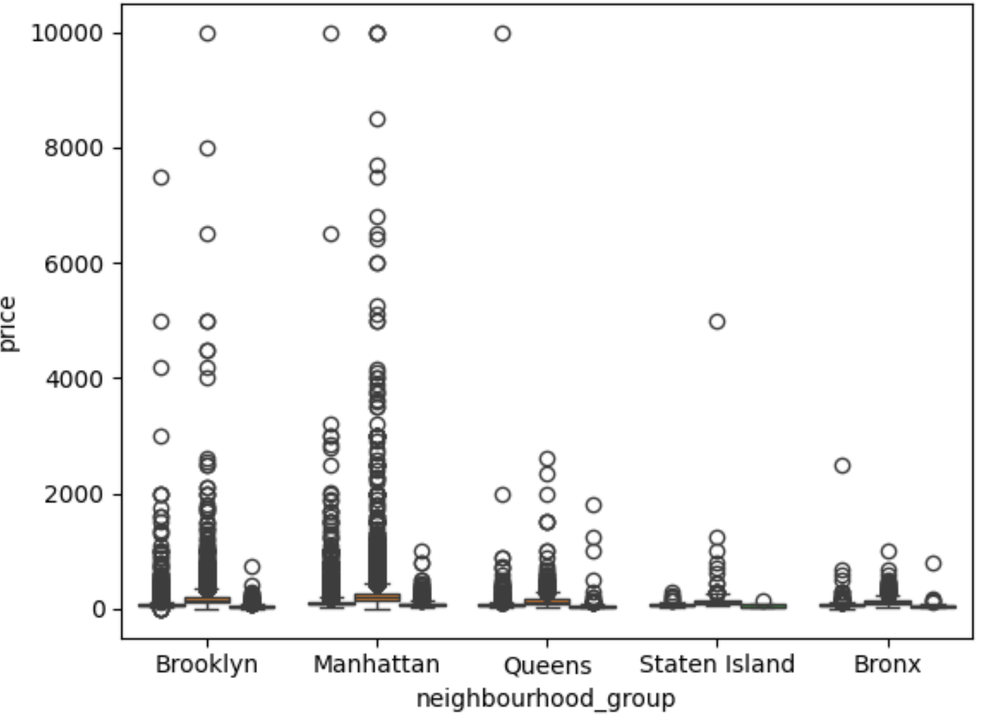


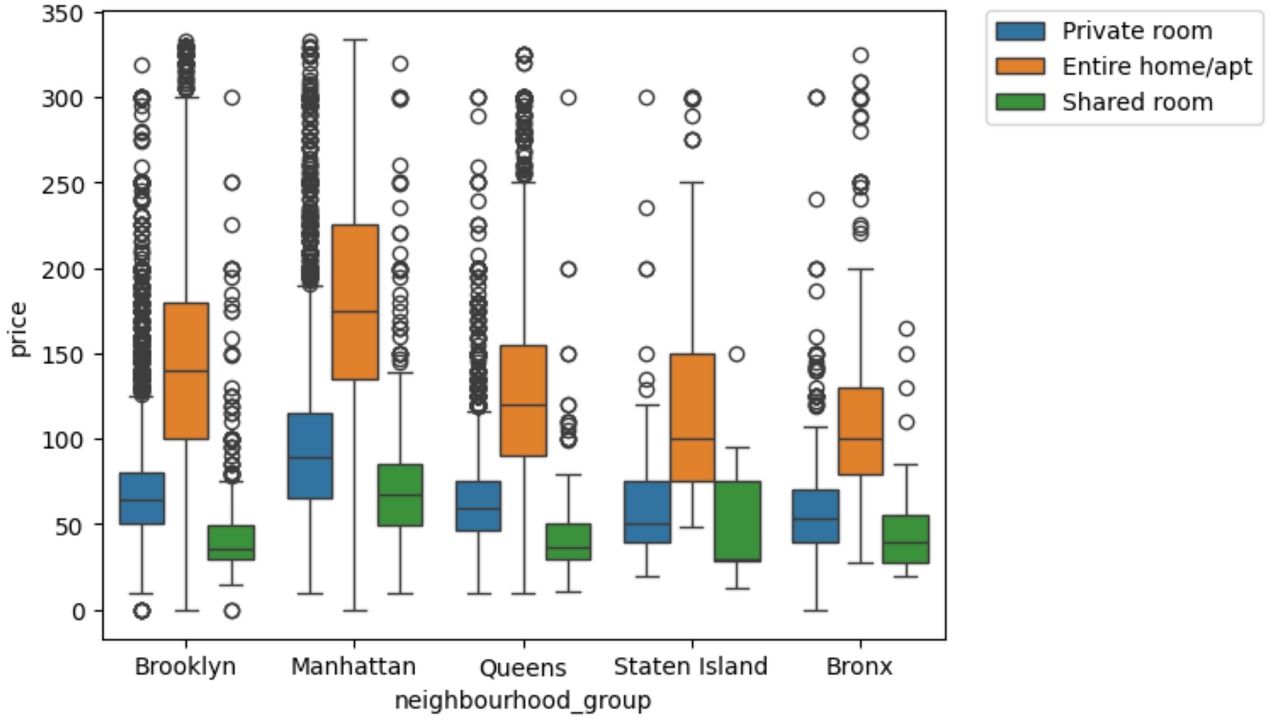



# DATA PREPROCESSING

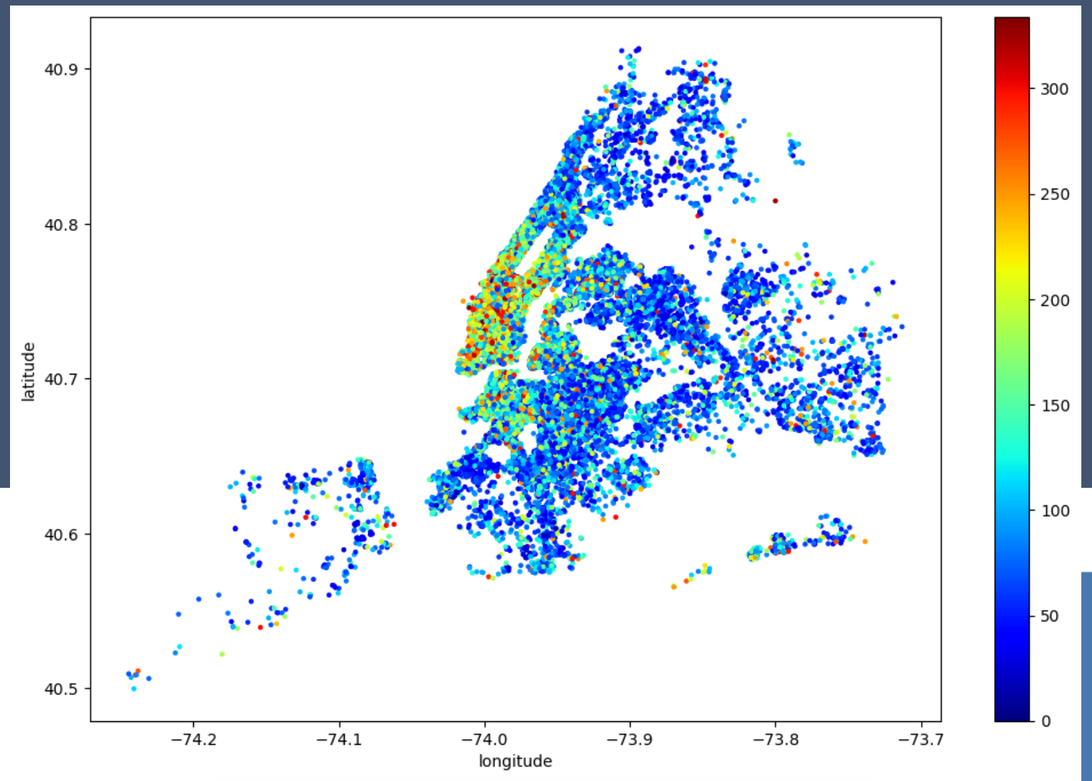
## Filtered outliers using IQR

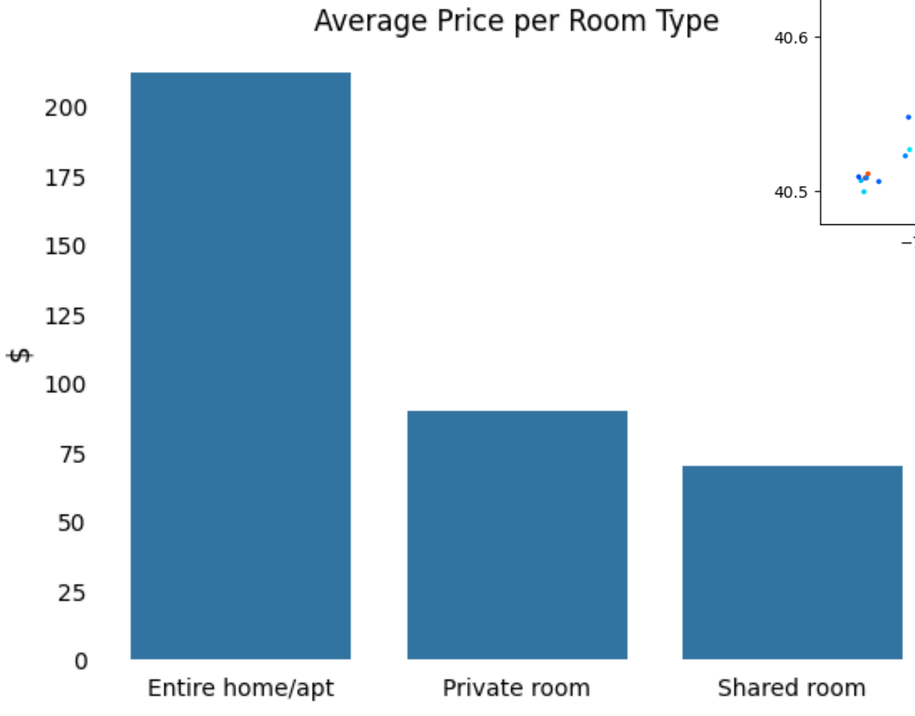
- First, dropped rows in which rental price = 0
- Then, applied IQR formula to remove outliers in the dataset

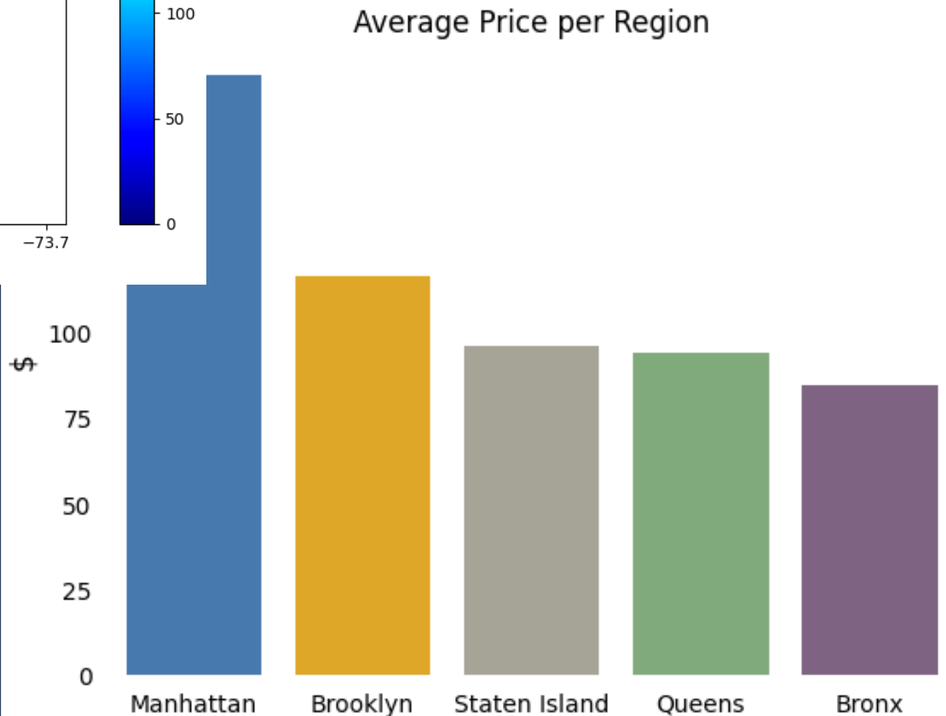


# EXPLORATORY DATA ANALYSIS

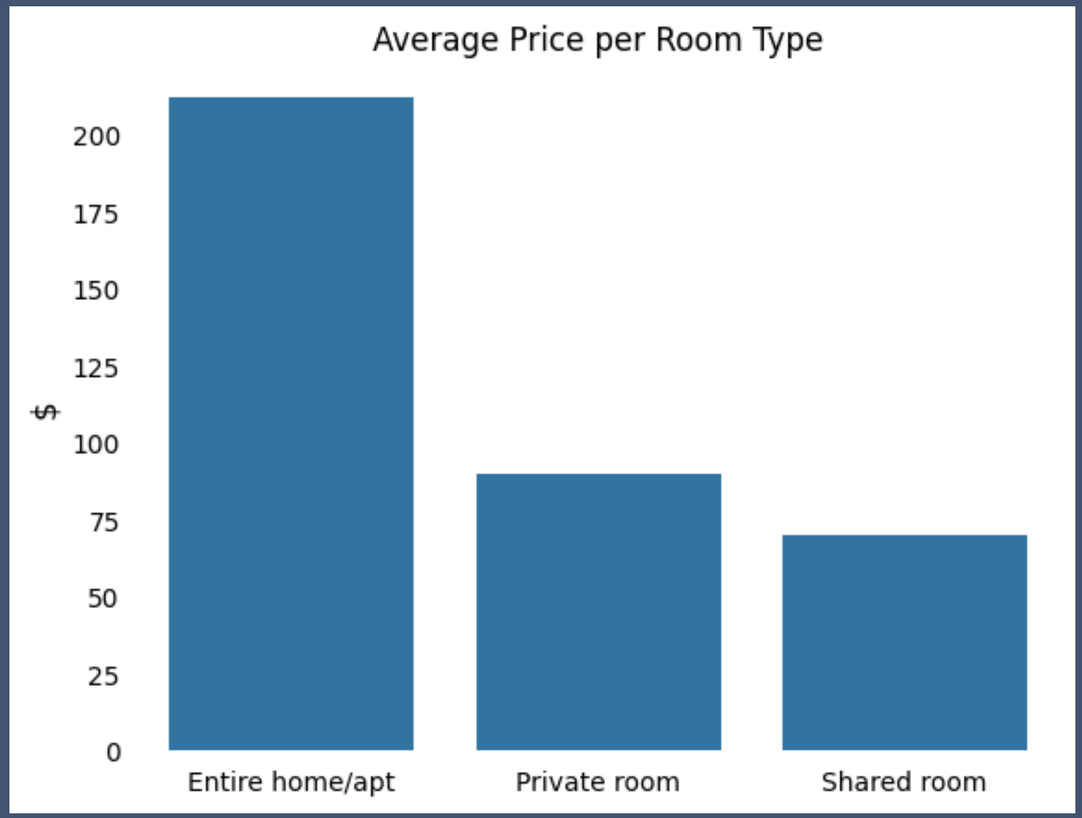


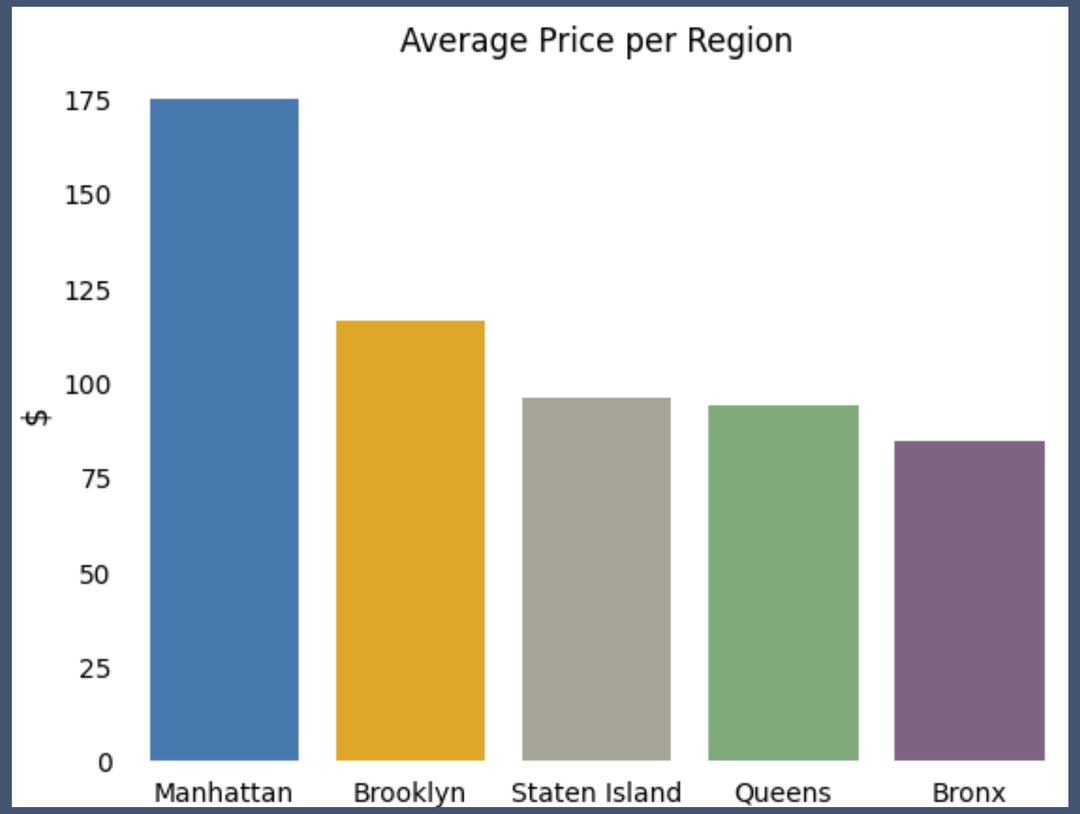
# PRICE BASED ON ROOM TYPE

The first variable that we identified as having a correlation with price is the room type.



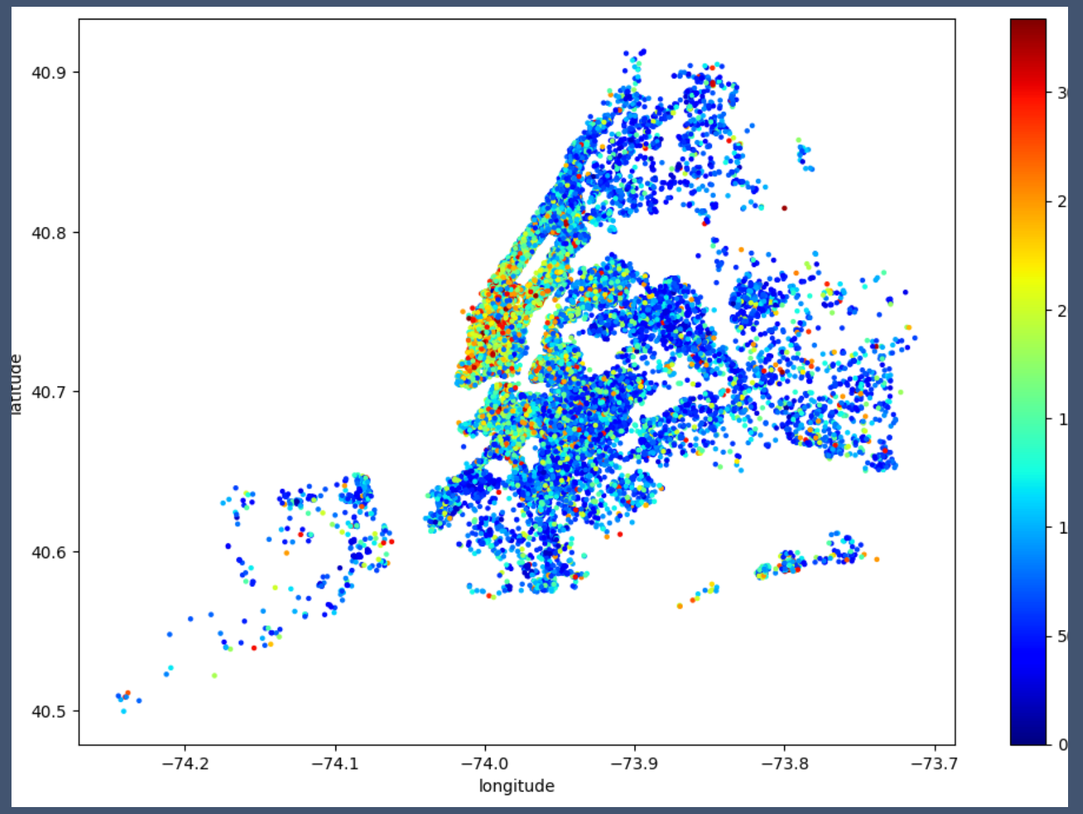
# PRICE BASED ON REGION

The next variable we identified as having a correlation with price is the ‘neighbourhood_group’ variable, aka region.



# PRICE BASED ON LATITUDE & LONGITUDE

Similarly to region, we also identified that the latitude and longitude of the listing played a role in influencing the listing price.



# DATA PREPROCESSING

| | id | host_id | latitude | longitude | price | minimum_nights | number_of_reviews | reviews_per_month | calculated_host_listings_count | availability_365 | _Bronx | _Brooklyn | _Manhattan | _Queens | _Staten Island | _Entire home/apt | _Private room | _Shared room |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **id** | 1.000000 | 0.588052 | -0.003031 | 0.091243 | 0.023676 | -0.014451 | -0.320804 | 0.291545 | 0.133959 | 0.084466 | 0.051214 | -0.058130 | -0.019690 | 0.090396 | 0.020771 | -0.054689 | 0.037347 | 0.057200 |
| **host_id** | 0.588052 | 1.000000 | 0.020182 | 0.128344 | 0.033510 | -0.018608 | -0.140451 | 0.296036 | 0.155661 | 0.204641 | 0.073342 | -0.116284 | 0.001861 | 0.132324 | 0.034578 | -0.077865 | 0.056625 | 0.070159 |
| **latitude** | -0.003031 | 0.020182 | 1.000000 | 0.085424 | 0.065374 | 0.025170 | -0.015185 | -0.009881 | 0.019518 | -0.011236 | 0.331174 | -0.673293 | 0.590831 | 0.017652 | -0.190524 | -0.006179 | 0.004790 | 0.004599 |
| **longitude** | 0.091243 | 0.128344 | 0.085424 | 1.000000 | -0.271110 | -0.064274 | 0.057869 | 0.145470 | -0.115696 | 0.085503 | 0.221146 | 0.015265 | -0.432327 | 0.622690 | -0.291903 | -0.190642 | 0.182360 | 0.028835 |
| **price** | 0.023676 | 0.033510 | 0.065374 | -0.271110 | 1.000000 | 0.023368 | -0.056026 | -0.029905 | 0.144355 | 0.113578 | -0.076426 | -0.173023 | 0.297928 | -0.150839 | -0.034755 | 0.503432 | -0.472587 | -0.105483 |

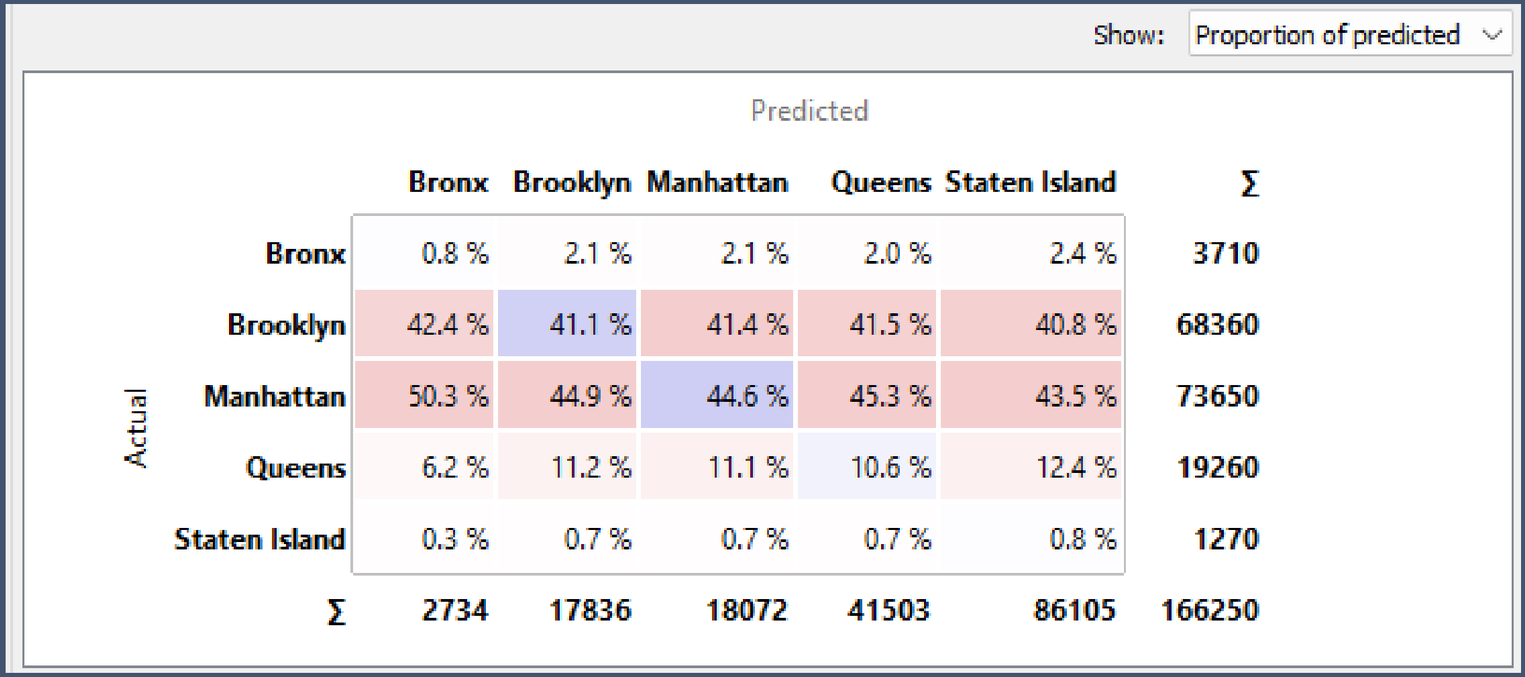
# CONVERTING CATEGORICAL DATA

- In order to train different models on the dataset, the desired categorical data for training must be converted into ‘dummy variables’
  - pandas.get_dummies:
    - room_type & neighbourhood_group

With room type as a target variable, The number of instances that were incorrectly classified as negative by the model when they were actually positive is over 50% (these are our false negatives.) The only factors below 50% are actual negatives, false positives, and true negatives.

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | **Entire home/apt** | **Private room** | **Shared room** | **Σ** |
| Actual | **Entire home/apt** | 51.9 % | 52.1 % | 51.9 % | **86400** |
| | **Private room** | 45.7 % | 45.7 % | 45.6 % | **75910** |
| | **Shared room** | 2.3 % | 2.2 % | 2.4 % | **3940** |
| | **Σ** | **16446** | **41273** | **108531** | **166250** |

# Confusion matrix based on neighborhood

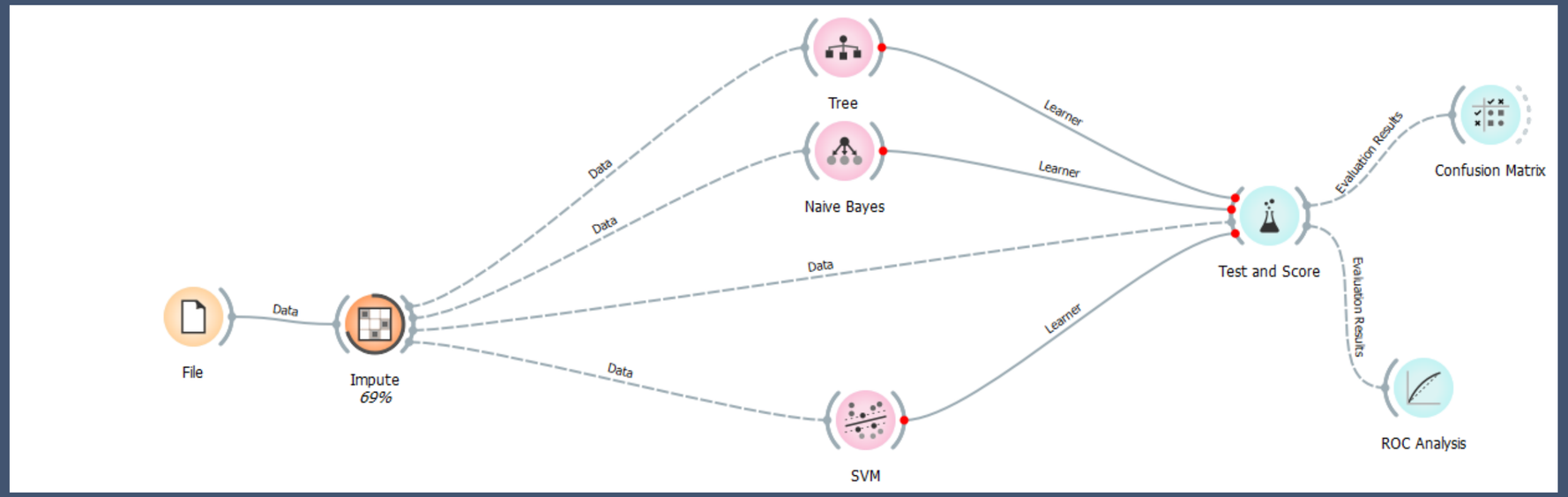


Show: Proportion of predicted

Predicted

| Actual | Bronx | Brooklyn | Manhattan | Queens | Staten Island | Σ |
|---|---|---|---|---|---|---|
| Bronx | 0.8 % | 2.1 % | 2.1 % | 2.0 % | 2.4 % | 3710 |
| Brooklyn | 42.4 % | 41.1 % | 41.4 % | 41.5 % | 40.8 % | 68360 |
| Manhattan | 50.3 % | 44.9 % | 44.6 % | 45.3 % | 43.5 % | 73650 |
| Queens | 6.2 % | 11.2 % | 11.1 % | 10.6 % | 12.4 % | 19260 |
| Staten Island | 0.3 % | 0.7 % | 0.7 % | 0.7 % | 0.8 % | 1270 |
| Σ | 2734 | 17836 | 18072 | 41503 | 86105 | 166250 |

# MODEL CONSTRUCTION IN ORANGE



# CONCLUSION

- In data preprocessing, we were able to clean the dataset and apply the IQR formula to remove outliers which skewed the data
- After preprocessing, we identified the variables that correlated the most with the price per listing
  - Room type
  - Region
  - Longitude & Latitude