What Decides A House's Price in New York City

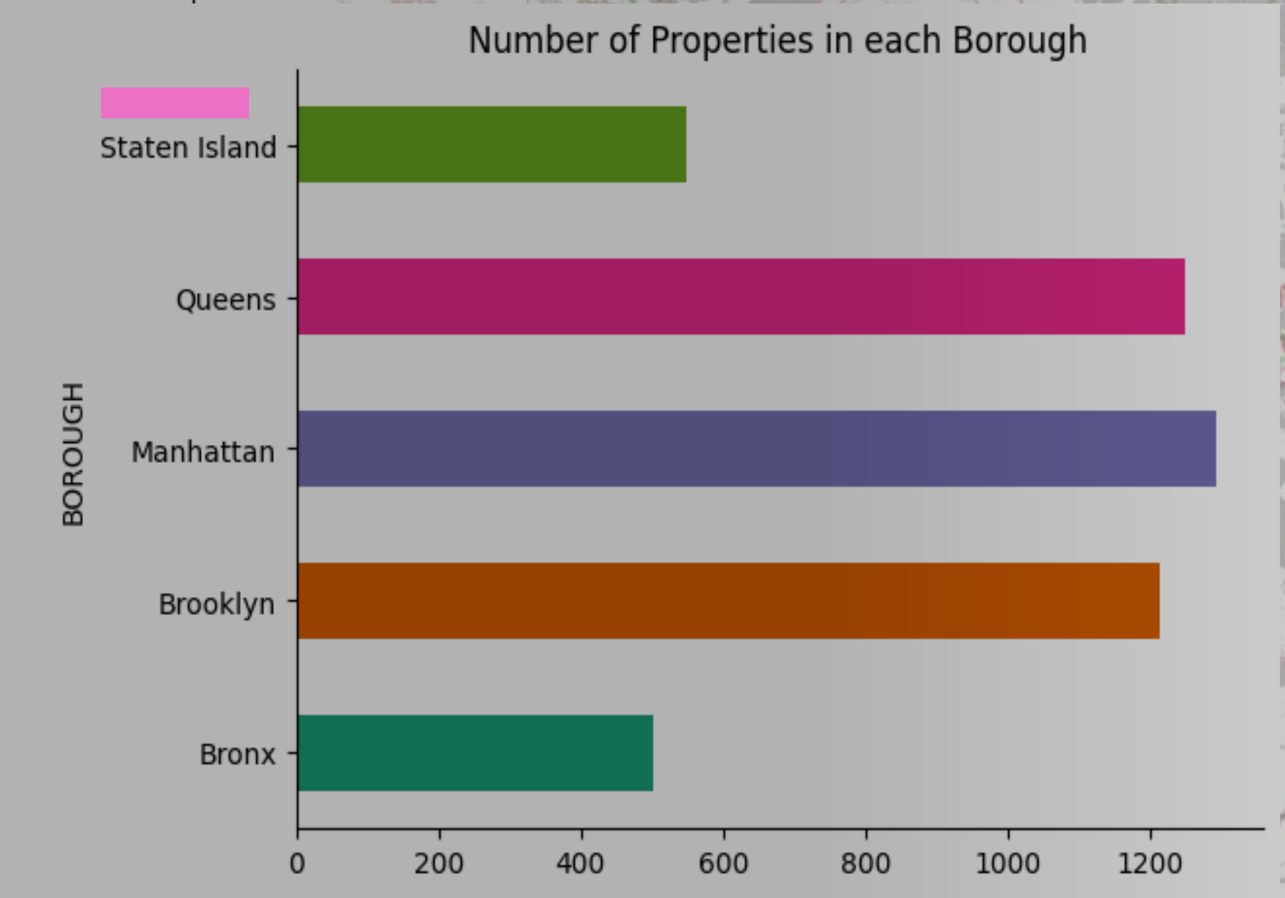
Team member: Wangsheng Wu (ww2674)

Summary of Project 1

- In project 1, I investigated the New York Housing Market dataset from **Kaggle** (https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market/data) on a very basic level which turned out to give me a very limited understanding of New York real estate market.
- I did not take good care of the dataset and cleaned it in a rude way (by removing lacking data and focusing on major group of data.)
- I did not analyze the main contributors of the price of a house but just predicting that price, failing to capture some interesting things.
- I did find the house price should be explained from multiple angles. But this single dataset lacks some features which could influence the price a lot.

Expectation of This Project

- In this project, I keep using the original dataset, but have it cleaned more elegantly and precisely.
- I also find new dataset(s) from **simplemaps.com** to add and explore more potential influencing features that may decide the price of a house.
- Also, I am going to apply more methodologies which I unfortunately did not even consider in the first project.
- Luckily, I get some interesting points about New York real estate market.



Layout of the Houses in the Dataset on a Map

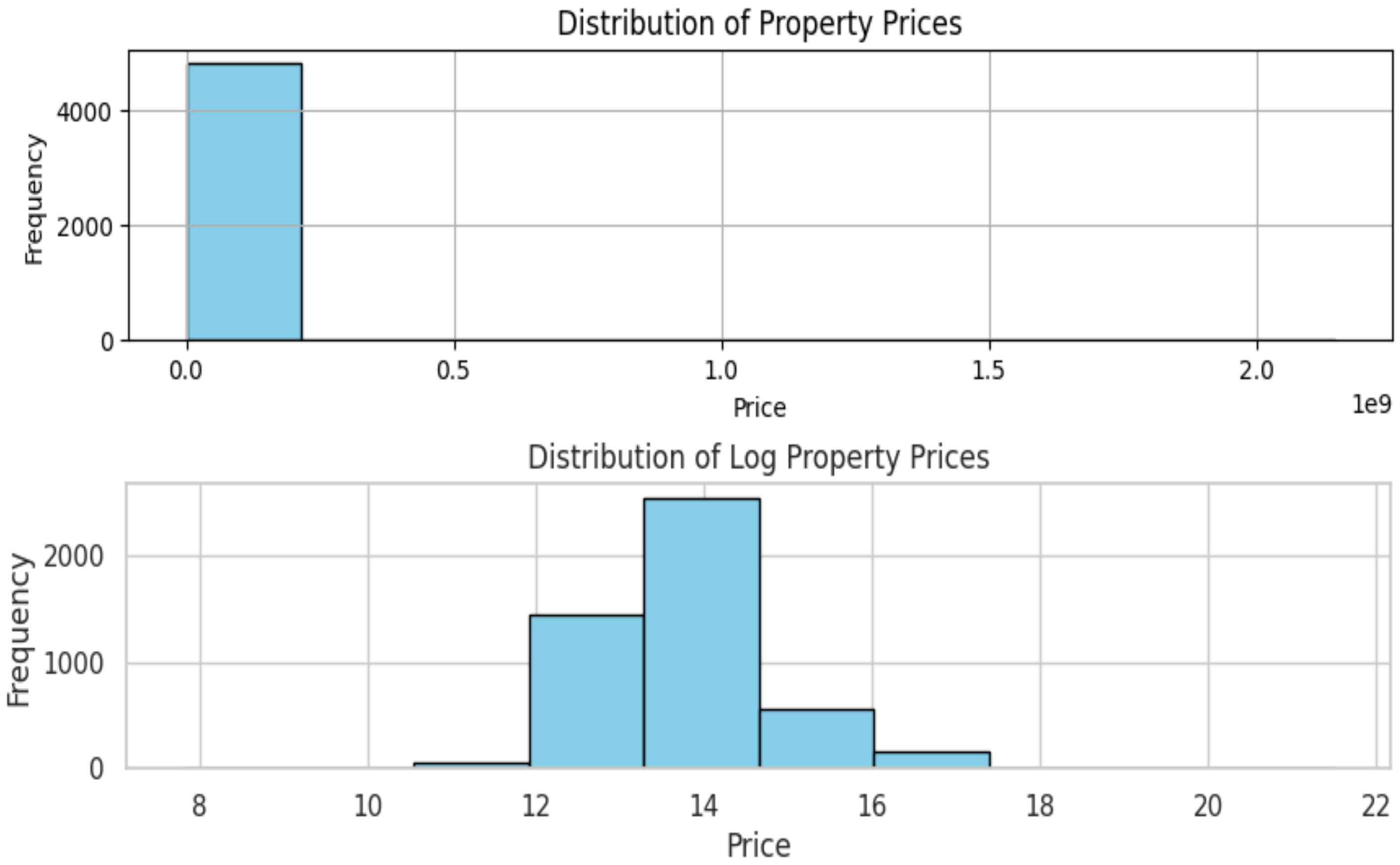
- Data points are “evenly” scattered in each borough of New York
- Each borough has comparable number of properties.

Introduction of New Datasets

- I retrieved house median income, education rate (percentage of people have a degree of college or higher), and median home value in each zip region of New York City.
- Merged these data with the original dataset and hence introducing 3 new features for exploration in this project.

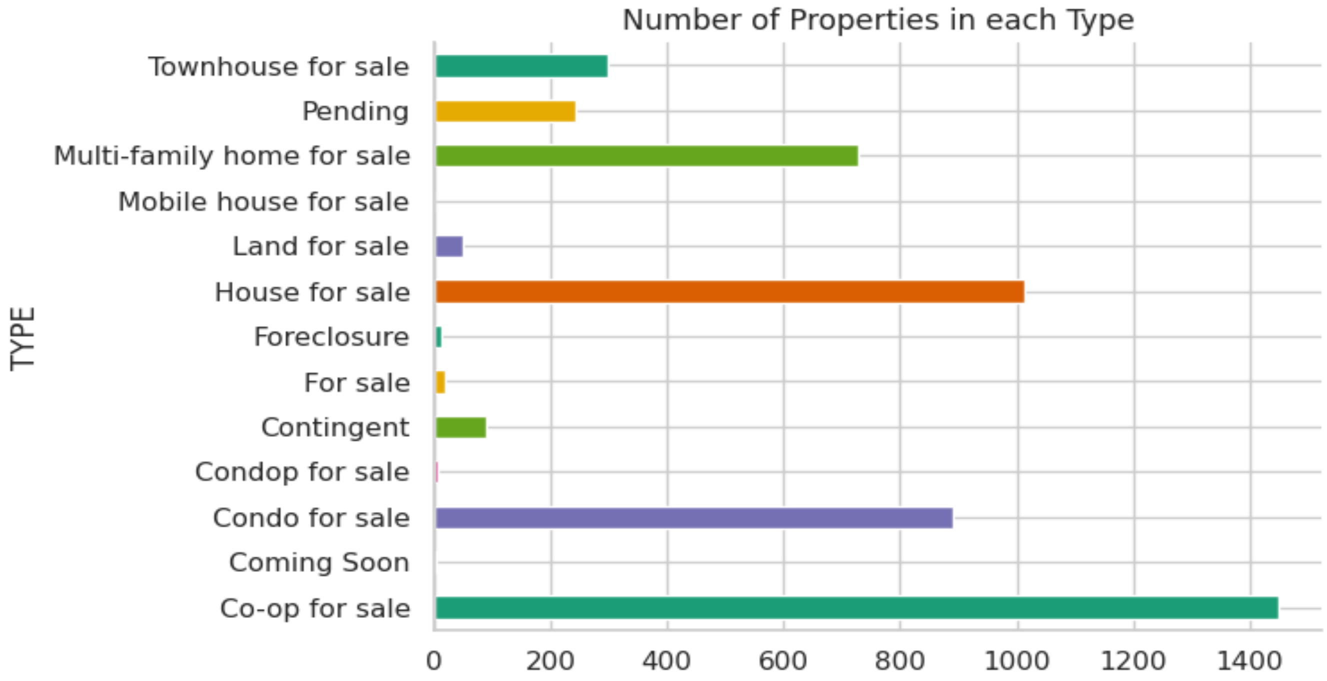
Feature Engineering

- Created new feature “log_price” from “PRICE” & “log_sqft” from “PROPERTYSQFT” since price and size data have a strong right-skewed distribution. It is a normal practice to consider doing log transformation to normalize their distribution, reducing skewness, and increase interpretability.

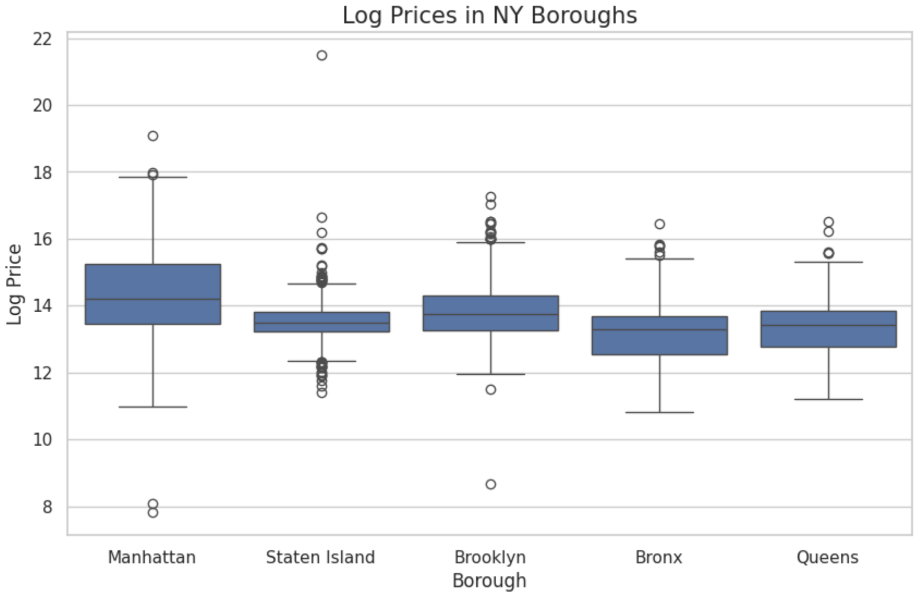


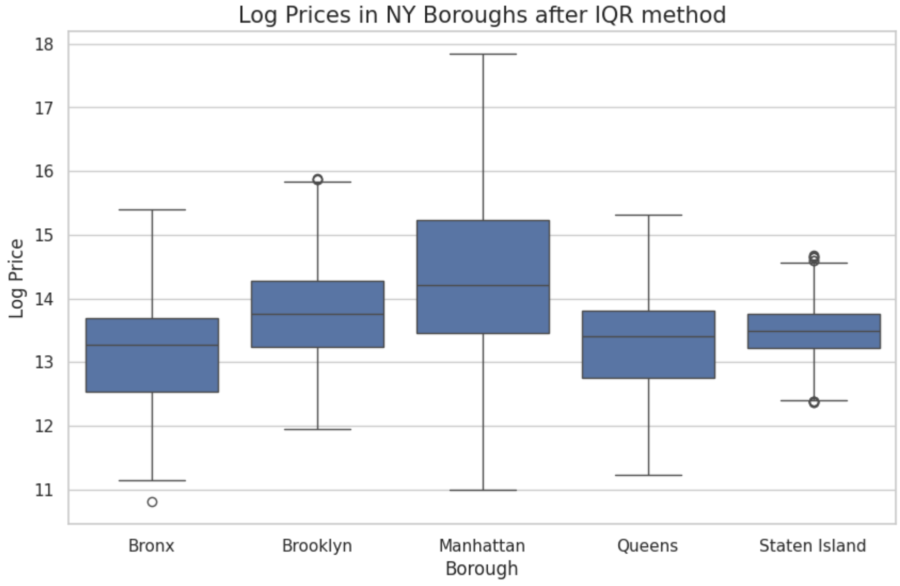
Feature Engineering

- Cleaned "TYPE" column by grouping rare types into a group of "other."



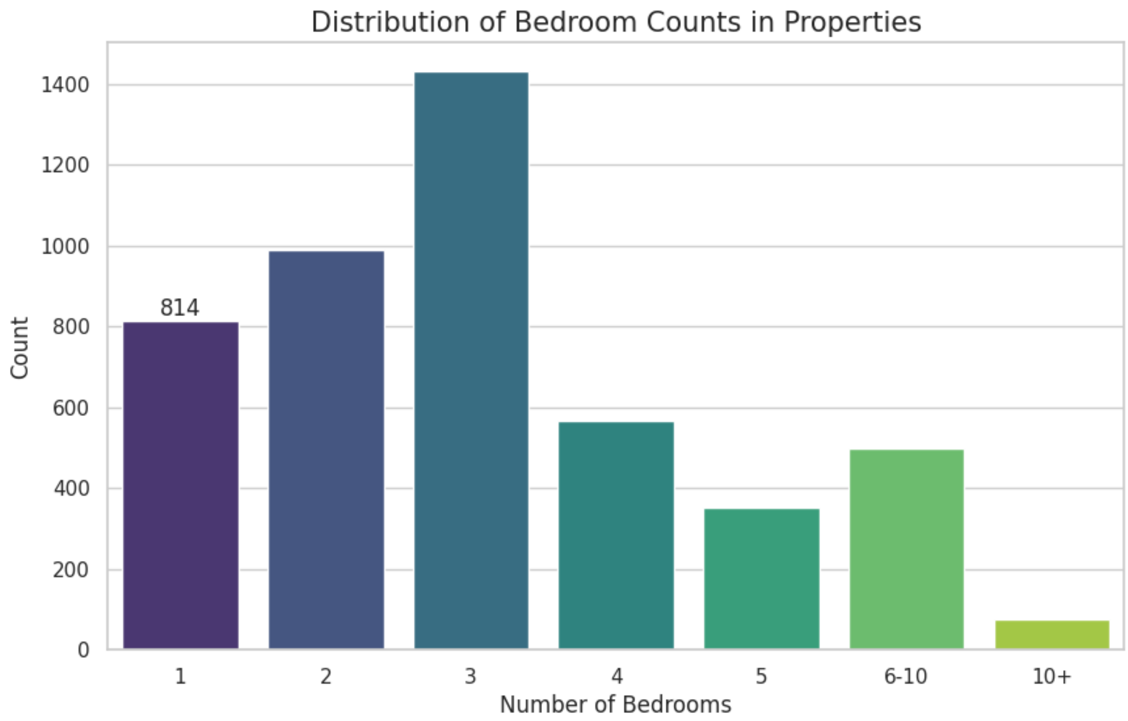
- Used IQR to remove outliers for columns like "log_price" and "PROPERTYSQFT" in groups like "BORROUGH" and "TYPE"

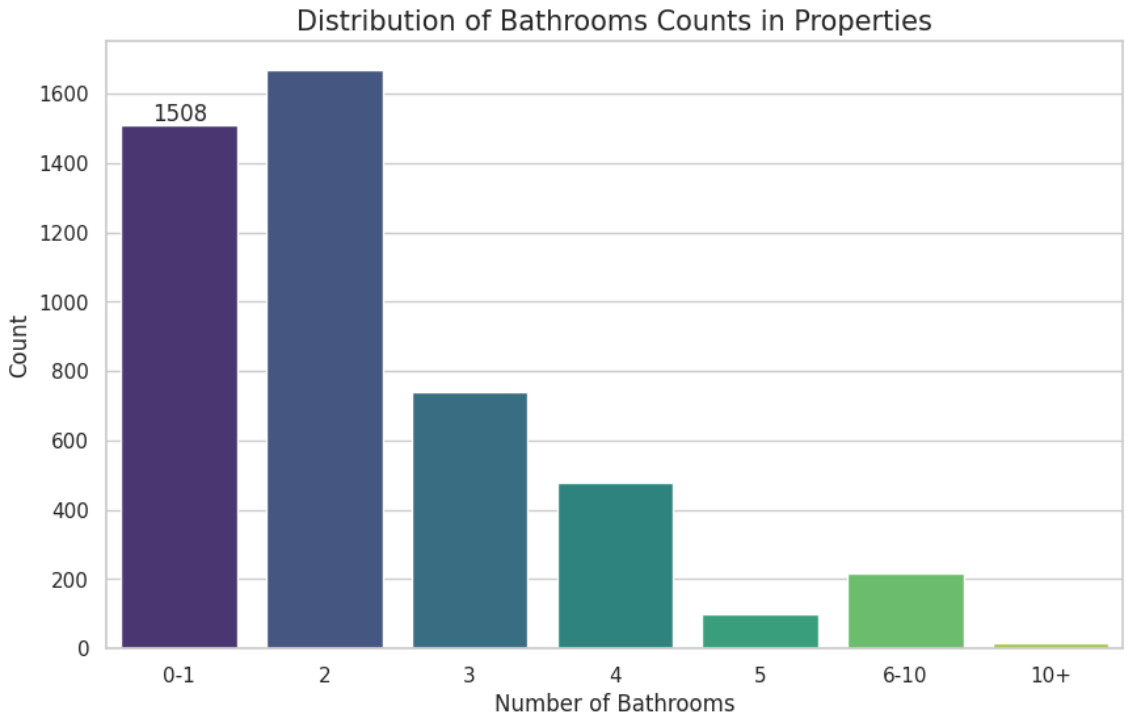




Feature Engineering

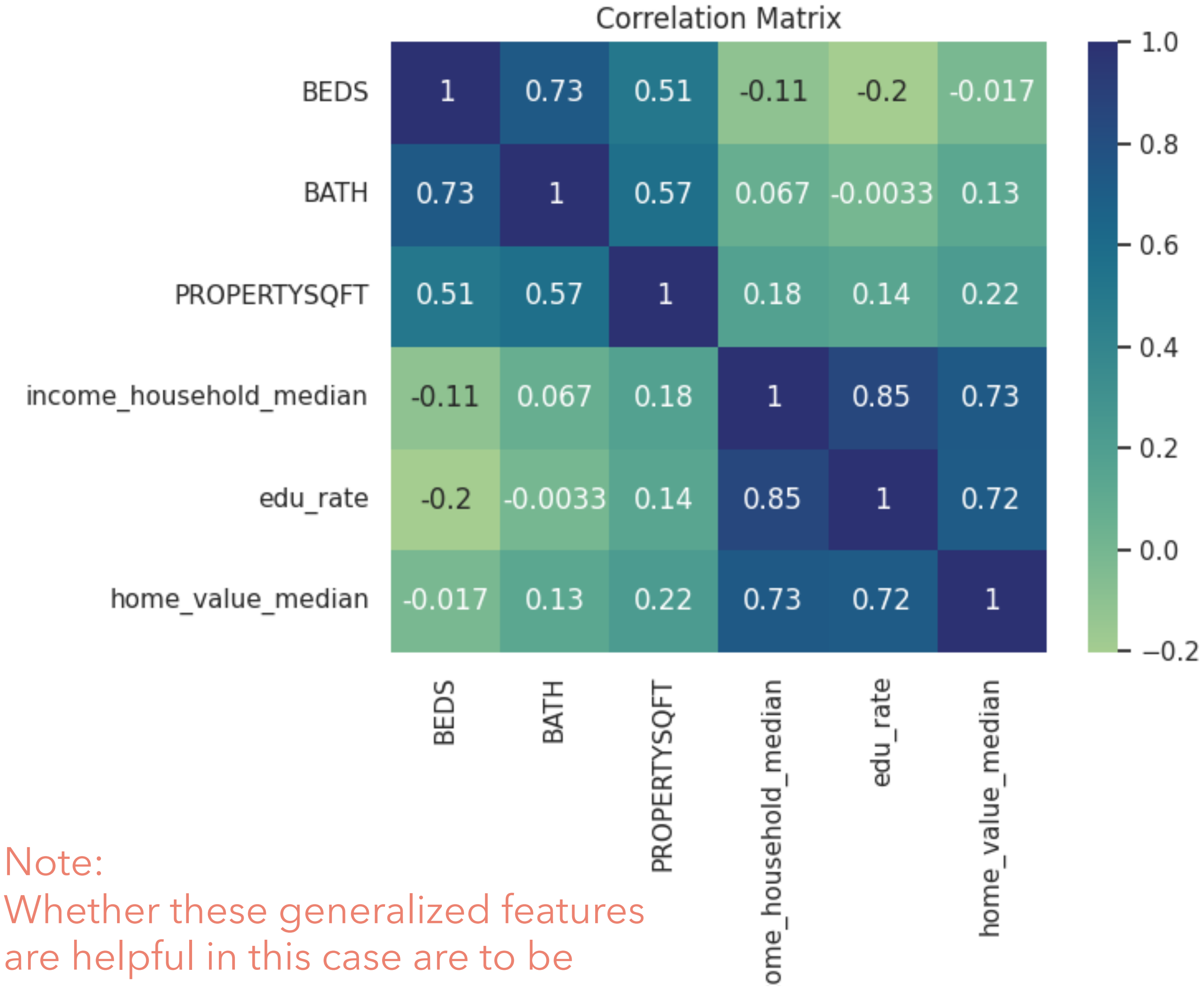
- Created categorical columns for "BEDS" and "BATH" using bin technique.





Attempt of Dimensionality Reduction Using PCA

- Since "BEDS", "BATH", and "PROPERTYSQFT" have strong correlation.
- So do "income_house_median", "edu_rate", and "home_value_median."
- Is it a good idea to generalize them into a single feature to better our model performance?
- Created new features "property_feature_pt" and "environment_feature_pt" using Principal Component Analysis.



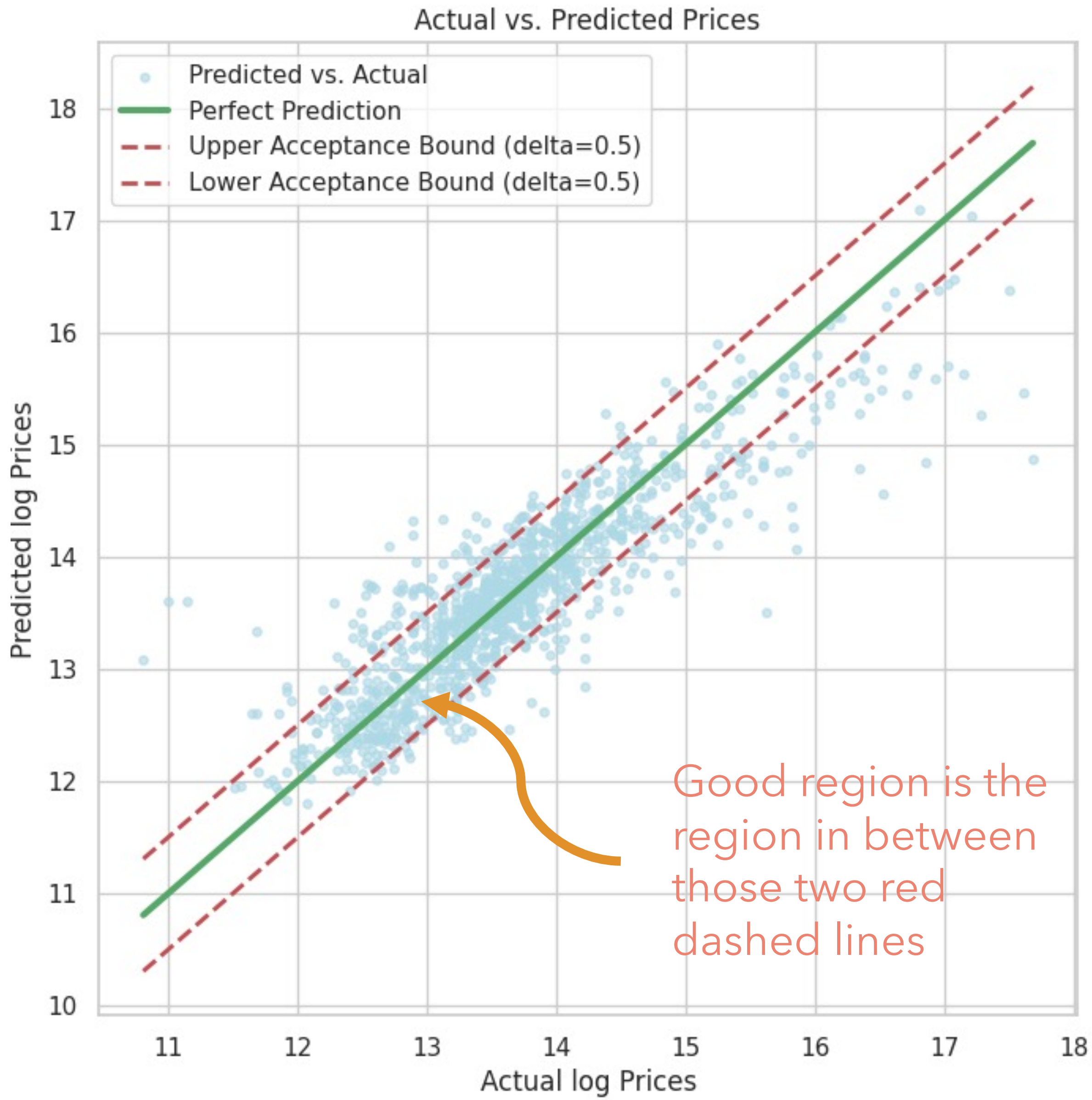
Note:
Whether these generalized features are helpful in this case are to be examined in model training later.

Model Training & Evaluation

- Model Choice: Linear Regression, Decision Tree, and Random Forest.
- Evaluation methods:
  - Normal performance metrics: MSE, MAE, and R^2.
  - Also created customized calculation for MSE, MAE, and R^2 in this case.
  - Feature Importance (Help us understand what are deciding features of the house price)

Explanation of the Customized Metrics

- Idea: House price is not stable. It fluctuates!
- In a range of 10 years, it is normal to expect a fluctuation rate (more likely an increase rate) of 50% by examining the paper *Trends in New York City Housing Price Appreciation* by NYU Furman Center.
- I set a good region of ± 50% for price prediction. If a price prediction falls into that region, error is not counted. Namely, MSE, and MAE are calculated by (actual, prediction) points outside that region. Customed R^2 is thus a weighted average of those two prediction results.



Linear Regression

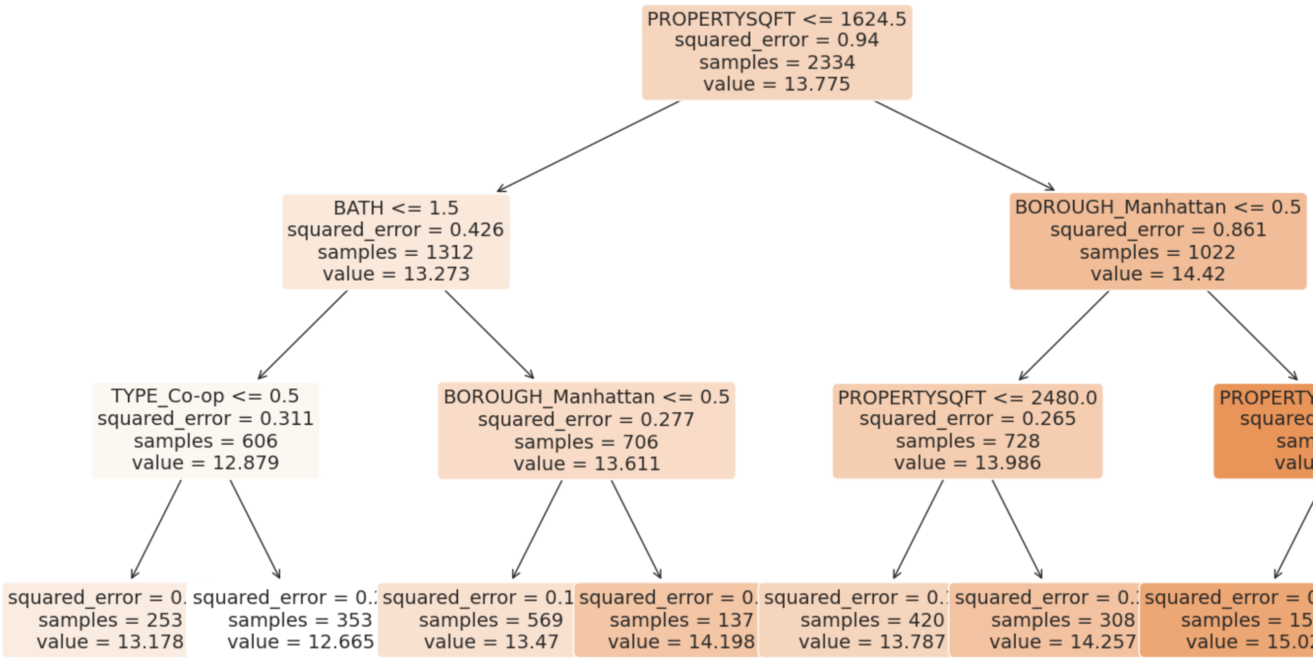
- `select_feature = ['BEDS', 'BATH', 'PROPERTYSQFT', 'BEDS_category', 'BATH_category', 'TYPE', 'BOROUGH']`
- 97.1% of the data can be explained using the customized metrics.
- 78.36% pairs of data fall into the good region we set.
- Linear Regression might be a good model in predicting house price?
- Anyway, Linear Regression Model with these selected features can be a good basic control model for testing other models.

```
Summary of Model
----------------------------------------
----------------------------------------
Mean Squared Error: 0.1742
Mean Absolute Error: 0.3178
True Mean Absolute Error: 680806.7402
R-squared: 0.8179
----------------------------------------
Custom Mean Squared Error: 0.1280
Custom Mean Absolute Error: 0.1595
Custom R-squared: 0.9710
----------------------------------------
Number of all (actual, pred) points: 730
Number of points within the region: 572
Percentage of points within the region: 78.36%
----------------------------------------
```

Logistic Regression

- `select_feature = ['BEDS', 'BATH', 'PROPERTYSQFT', 'BEDS_category', 'BATH_category', 'TYPE', 'BOROUGH']`
- 97.4% of the data can be explained using the customized metrics.
- 80.99% pairs of data fall into the good region we set.
- Logistic Regression generally performs worse than Linear Regression.
- Max_depth parameter can affect the performance of logistic regression model. Best choice is 6 for the feature selection above.

```
Mean Squared Error: 0.1843
Mean Absolute Error: 0.3172
True Mean Absolute Error: 650194.1903
R-squared: 0.8138
---------------------------------------------
Custom Mean Squared Error: 0.1354
Custom Mean Absolute Error: 0.1517
Custom R-squared: 0.9740
---------------------------------------------
Number of all (actual, pred) points: 584
Number of points within the region: 473
Percentage of points within the region: 80.99%
---------------------------------------------
```



Random Forest (Ideal Model)

- `select_feature = ['BEDS', 'BEDS_category', 'BATH', 'BATH_category', 'log_sqft', 'property_feature_pt', 'TYPE', 'BOROUGH', 'environment_feature_pt', 'edu_rate', 'income_household_median', 'log_home_value_median']`
- With all those features, we find the top deciding features are: our generalized property_feature_pt, edu_rate, BATH which is a component of property_feature_pt, BATH_category of 0-1, our generalized environment_feature_pt, and log_home_value_median which is component of environment_feature_pt.
- Up to 98.7% of these test data are explained by the Random Forest Model.
- 86.59% pairs of data fall into the good region we set.

```
Summary of Model
----------------------------------------------
----------------------------------------------
Mean Squared Error: 0.1255
Mean Absolute Error: 0.2565
True Mean Absolute Error: 529900.0611
R-squared: 0.8538
----------------------------------------------
Custom Mean Squared Error: 0.0832
Custom Mean Absolute Error: 0.1005
Custom R-squared: 0.9870
----------------------------------------------
Number of all (actual, pred) points: 1126
Number of points within the region: 975
Percentage of points within the region: 86.59%
----------------------------------------------
Ranked Feature Importance:
1. property_feature_pt: 20.05%
2. edu_rate: 16.36%
3. BATH: 14.96%
4. BATH_category_0-1: 12.13%
5. environment_feature_pt: 11.22%
6. log_home_value_median: 5.10%
----------------------------------------------
```

Feature Selection & Model Tuning

- Applied cross validation and grid search to pick the best hyperparameters for the Random Forest Model. But training a best model is not a major consideration in this project since we want to find the general deciding factors that influence the house price in New York.
- Use all those features we have might be a good idea in such a case considering our limited size of dataset and feature columns.
- Using generalized feature like “property_feature_pt” in model training performs a little bit worse than using these component features directly in all the three models as I investigated.
  PCA does not works as excellently as in extremely wide dataframes (i.e., scenarios of dataframes having many many features).

Extra Exploration

- Split the dataset into two sub-dataset by their "PROPERTYSQFT" value since I find there are a great portion of properties are of a size value ~2200 SqFt.
- It might be a good idea to investigate houses according to that size value.
- Models are better fitted for regular houses (not ~2200 SqFt)
- Popular houses (~2200 SqFt) are a little bit worse fitted by these models which might due to the reduction of one feature ("PROPERTYSQFT").

Interesting Findings

- Bathroom number ("BATH" & "BATH_category") can be a deciding feature in deciding/predicting the price of a house in New York since BATH contribute greatly to the price prediction in these models. This makes sense since in renting market, a room with a bath can differentiate it greatly in price away from its competitors which might even have larger area.
- Environment features like median household income, education rate, median home value of the region the house is in are all good representatives of the value of that house.

Limitation of the Project & Conclusion

- The size of the dataset is a flaw. Only 4801 rows of raw data, and about 4501 rows of data in cleaned dataset used for model training.
- House price is a complex combination of features from various levels and angles. More features we have, better model we can make.
- Even though being limited by the dataset, we can still get some idea of New York housing market and find interesting points in decisive factors of house price in New York.

Thank You!