Carlos Hernandez Luis D Jimenez Guerra Trevor Werner

**Class Project – What Stories do Data Tell?**

**Analyzing and Interpreting New York Housing Prices**

**Brief Introduction**

Our team selected a dataset from Kaggle that includes housing prices in New York, providing valuable insights into the real estate market. It contains 4802 rows and the following columns:

- TYPE
  - This column contains information on the type of house e.g. house, condo, multi family, etc.
- BEDS
  - This column contains information on the number of bedrooms on the house sold.
- BATH
  - This column contains information on the number of bathrooms on the house sold.
- PROPERTYSQFT
  - This column contains information on the area (square footage) of the home sold.
- LATITUDE :
  - This column contains the latitude for the location of where the house is located
- LONGITUDE :
  - This column contains the longitude for the location of where the house is located
- PRICE:
  - This column contains the selling price from the house.

**Data Preprocessing**

In order to analyze and model the data correctly, we removed outliers from the 'PRICE' column using the IQR(Inter Quartile Range) method. In addition, we kept the main categories in the 'TYPE' and grouped the categories with low sample size as 'Others' in order to balance the design of the test and the training of ML models.

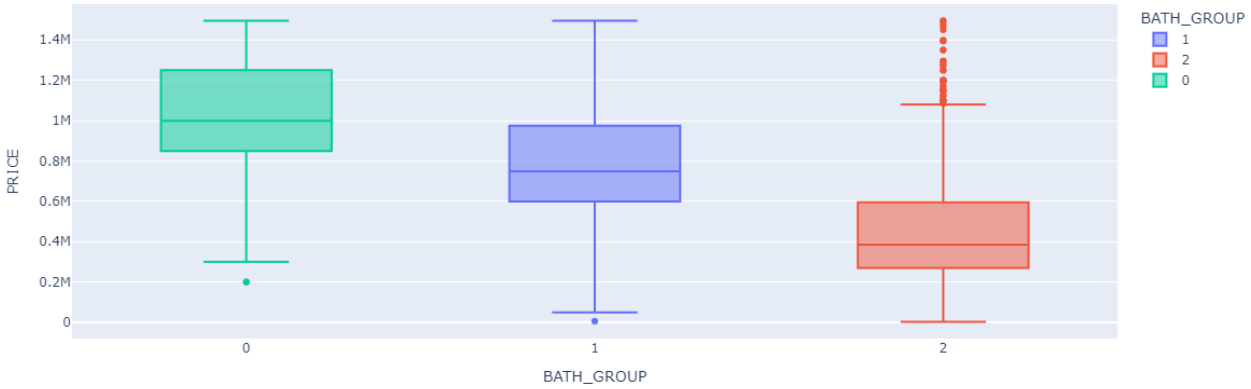
Furthermore, our team used KMeans to group the bathroom and bedroom categories into 'Low' , 'Medium' and 'High' as a way to balance the features. Additionally, due to the limited

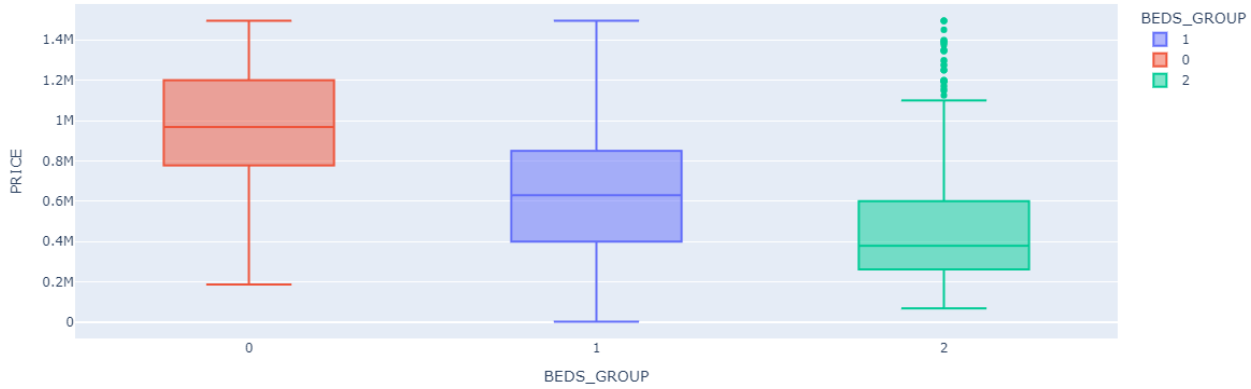
amount of predictive features in the dataset, we performed feature engineering by taking the square root and the log transformation, as well as the square and cube of the features: 'PROPERTYSQFT' , 'LATITUDE' , 'LONGITUDE' (although for the 'LATITUDE' and 'LONGITUDE' we did not use square root and log transformation). Lastly, each numerical feature scaled using a MinMax scaler instead of the Standard Scaler due to the skewness of the data.

**Key Insights and EDA**

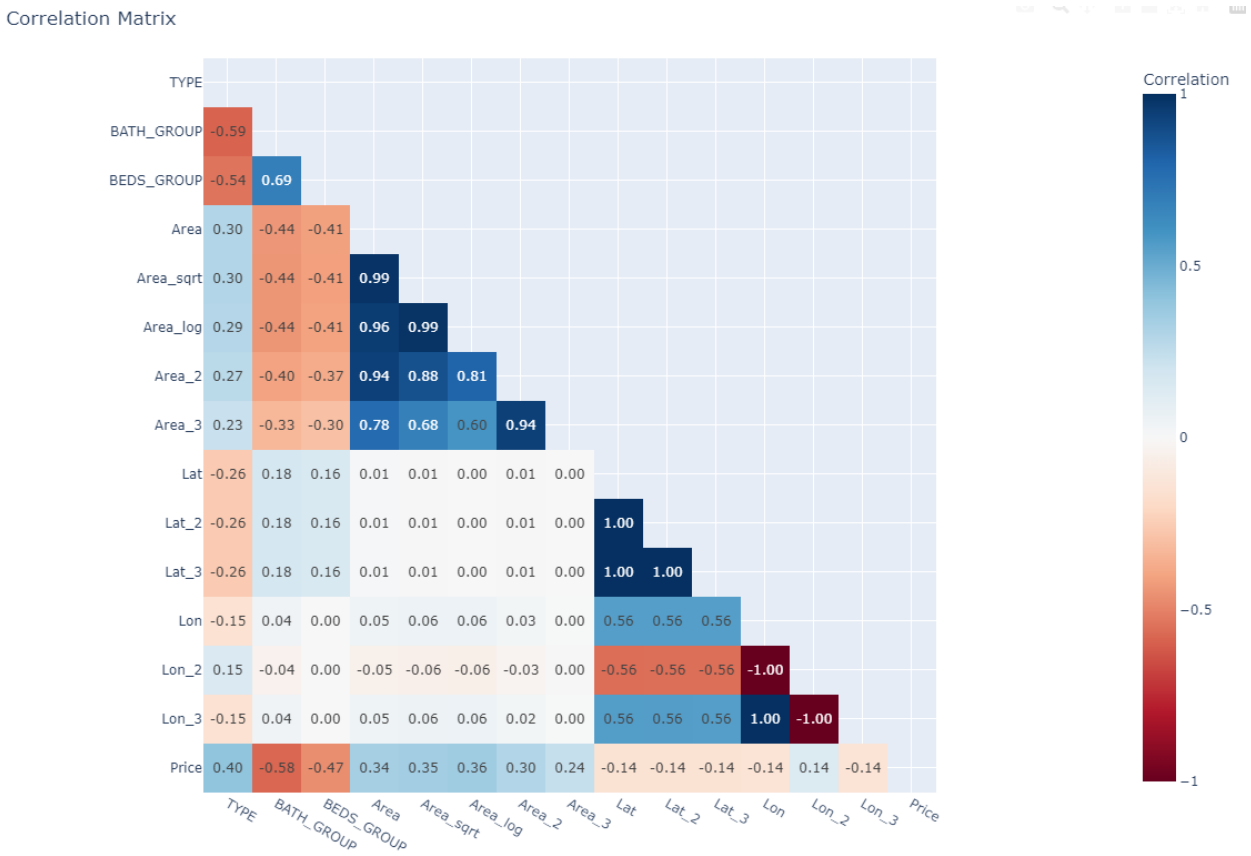
- From the figure below, we can see how the prices for some of the housing types is skewed and in some cases with heavy tails.


- Housing price for the bath and bed categories is also separated as a result of the KMeans as seen below:




- Correlation matrix between price and the features engineered as well as the features available originally




**Hypothesis Testing**

We performed a t-test to assess whether the average price for the housing type is statistically different. For this, we used multiple t-test by finding all the paired combinations of the housing categories, and adding a Bonferroni correction, adjusting the alpha level by the number of comparisons being made. This decreases our chances of a Type I error. The null and alternative hypothesis in was set up as following:

$$H_0: \overline{price}_a = \overline{price}_b$$

$$H_a: \overline{price}_a \neq \overline{price}_b$$

Where *a* and *b* are the housing categories being tested.

The results showed a rejection of the null hypothesis on all tests except for the comparison between Condo and 'Other' as seen below:

| comparison | t-stat | p-value | p_value_corrected | rejected |
|---|---|---|---|---|
| Condo vs House | -6.178246 | 8.902275e-10 | 8.902275e-09 | True |
| Condo vs Co-op | 15.887604 | 3.841818e-51 | 3.841818e-50 | True |
| Condo vs Other | -0.157049 | 8.752409e-01 | 1.000000e+00 | False |
| Condo vs Multi-family home | -16.303075 | 1.201109e-53 | 1.201109e-52 | True |
| House vs Co-op | 27.210122 | 1.716747e-136 | 1.716747e-135 | True |
| House vs Other | 5.218287 | 2.331359e-07 | 2.331359e-06 | True |
| House vs Multi-family home | -12.081873 | 1.020172e-31 | 1.020172e-30 | True |
| Co-op vs Other | -13.829338 | 2.129876e-38 | 2.129876e-37 | True |
| Co-op vs Multi-family home | -37.994504 | 2.196711e-198 | 2.196711e-197 | True |
| Other vs Multi-family home | -14.236903 | 3.852831e-41 | 3.852831e-40 | True |

**Machine Learning Models**

4 models were built, a Linear Regression , K Nearest Neighbors with 5 neighbors , Decision Tree with a *squared error* and the criterion and a Random Forest. The Random Forest was trained using 100 estimators, *Friedman MSE* as the criterion and then the model was used to select the best predictive features in the problem based on the feature importance. Then the selected features were used to fit the remaining models. The most important features turned out to be *BATH_GROUP* , $Longitude^2$ , *TYPE* , and $Latitude^2$ .

**Machine Learning Results**

The Linear regression model residuals showed signs of underfitting, as the residuals showed a trend where the higher the predicted price, the higher the residuals were. The decision Tree and K Nearest Neighbors showed normal residuals centered around 0.

Below as the results for the models in the testing set:

| Model | RMSE | MAPE |
|---|---|---|
| Decision Tree | 12.45 | 14.71 |
| Random Forest | 12.44 | 14.70 |
| KNN | 12.44 | 14.70 |

Linear Regression model was not selected for testing due to not meeting the residual assumptions.

**Final Conclusions**

Finally, feature engineering revealed key predictors such as the number of bathrooms, geographical location (latitude and longitude), and property type, which were essential in driving price variations. Our hypothesis testing confirmed significant price differences across housing types, with all pairwise comparisons rejecting the null hypothesis, except for the comparison between condos and "other" types.

**Dataset link:**

https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market