Annie Jiang
CIS 3120
May 22, 2024

NYC Housing Regression Report

Housing prices in New York City (NYC) are influenced by several factors, ranging from the property's physical attributes to its location within the city's diverse boroughs. Understanding the roles of these features in determining price is crucial for stakeholders, including homebuyers, real estate investors, and policymakers.

In this project, I created a linear regression model to estimate the price of a house located in one of NYC's five boroughs based on information such as type of property, number of bedrooms and bathrooms, square footage, and borough. The dataset, from Kaggle, used in this project includes 4801 observations of property listed in New York State with 17 columns.

The first step in this project was to preprocess the data. I looked at the descriptive statistics and data types for each column to become more familiar with the data. I checked for null values, and there were none. I noticed that there wasn't a dedicated city column, only columns for locality and sublocality, which I noticed were a mix of counties and cities, and a state column that was formatted as city, state, zip code. Therefore, I decided to extract the city from the state column and store it in a new column called "City." Next, I dropped rows where the City value wasn't in the five boroughs, Manhattan, Brooklyn, Queens, Bronx, and Staten Island. After that, I decided to exclude several types of housing from the analysis, such as condops, properties in foreclosure, mobile homes, etc. for the sake of simplicity. When I was looking at the descriptive statistics, I noticed that the price of 2 E 55th St Unit 803 was listed at over $2 billion. Therefore, I investigated the property and found that it actually sold for $350k and was never worth $2 billion as the data suggested, so I fixed the value. This would help the linear regression model as a property with a price of $2 billion would skew the results greatly. In addition to that, I removed outliers in the price column using the interquartile range method to ensure more robust regression results. Finally, I dropped the columns that would be irrelevant to linear regression, such as broker title, administrative area level 2, address, etc. I also dropped the longitude and latitude columns despite them being numerical since they were not directly interpretable as meaningful features for predicting housing prices, and since we have the city column to encapsulate the location of the housing.

After preprocessing the data, I created data visualizations to attempt to uncover trends in the data. Firstly, a correlation matrix (Plot 1) shows the relationships between various numerical features in the dataset, including price, number of bedrooms, number of bathrooms, and property square footage. From the heatmap, we can see that there is a moderate positive correlation between price and square footage. This suggests that larger properties tend to have higher prices, which is a logical and expected trend in real estate. We also see a strong positive correlation between number of bedrooms or number of bathrooms and square footage. This also makes sense as larger properties tend to have more bedrooms and bathrooms. For Plot 2, I created a pair plot with borough as the hue to further visualize the relationships between the key numerical features, segmented by borough. In the price vs. square footage plot, there is a subtle positive relationship between the two variables. Manhattan and Brooklyn seem to have higher-priced properties for similar square footage compared to other boroughs. Plot 3, which are multiple box plots showing the price distribution by borough, provides a clear comparison of housing prices across the five boroughs of NYC. Looking at the Manhattan box plot, the median price is significantly higher compared to the other boroughs, and there is a wider range of prices. In contrast, Staten Island has the lowest median price and the most narrow range of prices out of all the boroughs. Plot 4 shows box plots of property square footage by borough, where we can see that Staten Island, Brooklyn, the Bronx, and Queens have relatively similar median property sizes, indicating more uniform housing sizes in these boroughs compared to Manhattan. The presence of outliers in all boroughs also suggests that there are exceptions with significantly larger properties.

After finding these insights in the data, I began to prepare it for a linear regression model. I encoded the categorical columns, city and type, using OneHotEncoder from the sklearn.preprocessing module. This converts categorical variables into a format that can be provided to machine learning algorithms to do a better job in prediction. Next, I standardized the numerical columns, beds, bath, and property square feet, using StandardScaler from the same module. StandardScaler standardizes features by removing the mean and scaling to unit variance. Next, I split the data into training and test sets in a 70:30 ratio. This ensures that the model is trained on a majority of the data and tested on a separate subset to evaluate its performance. Afterwards, I fitted the model on the training data using LinearRegression from the sklearn.linear_model module. Moreover, I calculated the model's coefficients and intercept to understand the impact of each feature on the price.

Finally, I evaluated the model's performance on the test data using metrics such as R-squared, mean squared error, root mean squared error, and mean absolute error. The R-squared value was 0.448, and this indicates how well the independent variables explain the variability of the dependent variable. The mean squared error (MSE), which was 3.76×10^{11}, provides a measure of the average squared error between predicted and actual values. The root mean squared error (RMSE) is the square root of the MSE, which offers a more interpretable metric of error in the same units as the dependent variable, and in this case it was 6.12×10^{5}. Finally, the mean absolute error (MAE) was 4.22×10^{5}. This measures the average absolute error between predictions and actual values.

Once I created my linear regression model, I began to deploy it to a web interface through PythonAnywhere where non-technical users could make housing price predictions with the model. I saved the linear regression model using Python's pickle library. I saved the StandardScaler using pickle as well, since I standardized the numerical variables with it. This is crucial for ensuring that any new input data is scaled in the same way as the training data, maintaining the consistency of the predictions. After that, I created a form where users can input data for every feature, such as borough, number of bedrooms, number of bathrooms, property square footage, and type of property. When the user submits a complete form, the model makes a prediction and the website displays the estimated price of the housing.

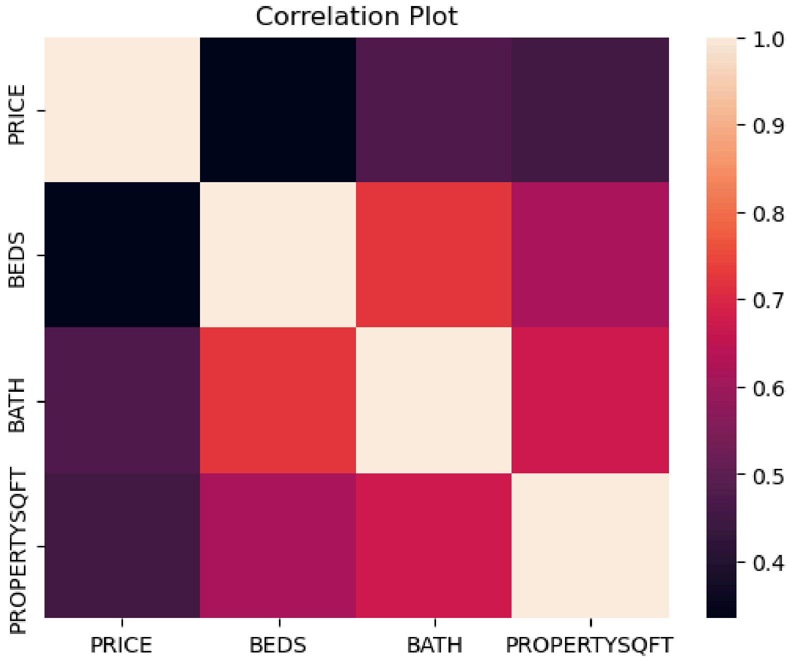
This predictive model can solve problems for homebuyers, real estate investors, and real estate agents. Prospective homebuyers can use the website to estimate the value of properties they are interested in, which helps them make informed decisions and negotiate better deals. Real estate investors can quickly assess the potential value of properties, enabling them to identify undervalued properties and make strategic investment decisions. Furthermore, agents can use the tool to provide clients with data-driven price estimates.

However, one problem I ran into when deploying the model was that when I would submit a form, I would receive predictions with astronomically high prices. For example, when I input a 2 bedroom, 1 bathroom condo in Brooklyn with square footage of 800, the predicted price would be $\$2.29 \times 10^{18}$, which cannot be a real price for property. I am unsure where the process went wrong that the model returns this high of a number. Regardless, this program still has potential to be helpful if the issue with the predictions can be resolved.

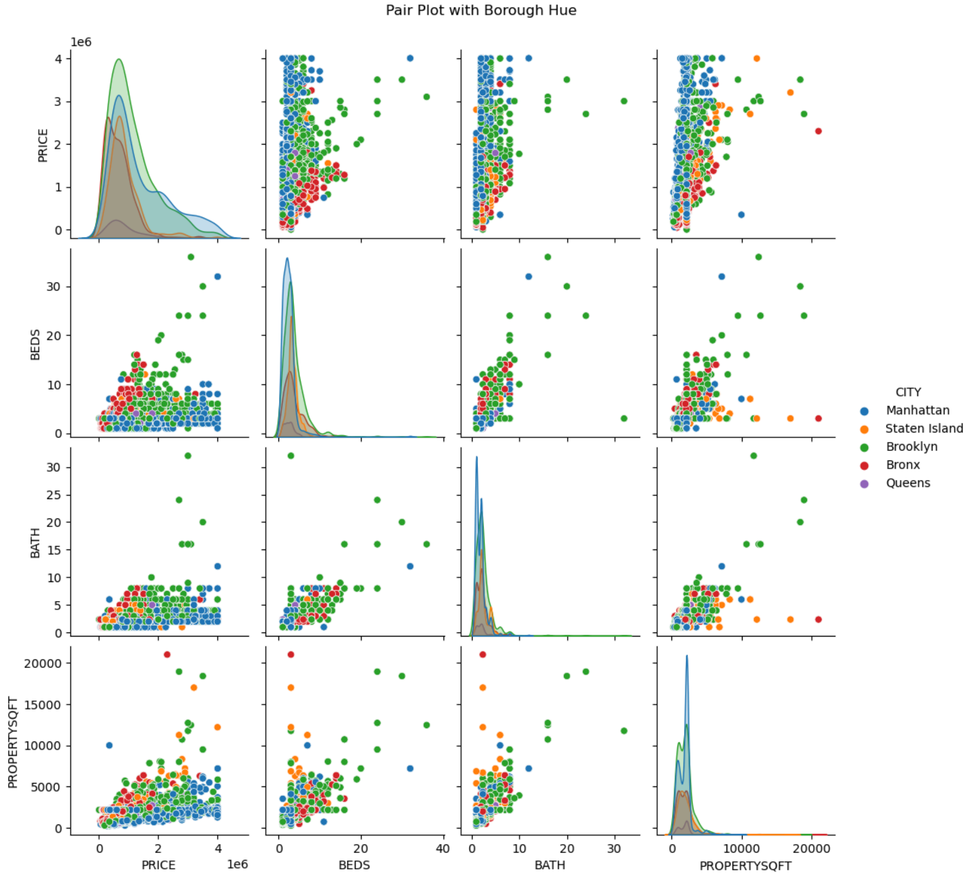
In summary, I created a linear regression model to predict housing prices in NYC based on various features. I deployed the model on the website so that non-technical users can upload data to it and make predictions themselves. This model can solve problems for prospective homebuyers, real estate investors, and real estate agents if its current issue with predicting exorbitantly high prices is fixed.

Appendix

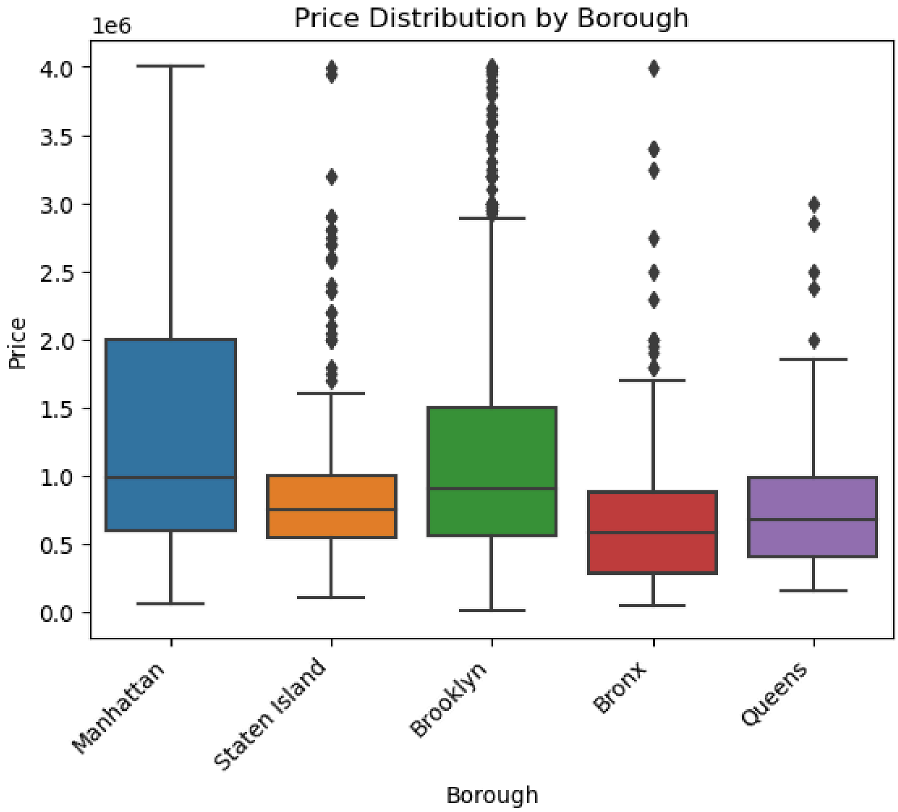
Plot 1 - Correlation matrix



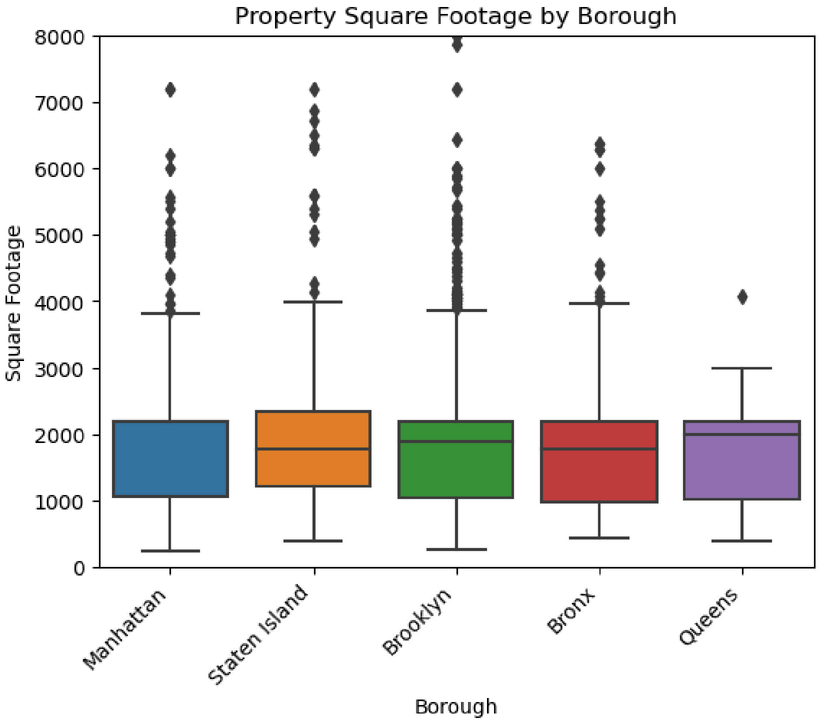
Plot 2 - Pairplot



Plot 3 - Price Distribution by Borough



Plot 4 - Property Square Footage by Borough


Property Square Footage by Borough
Square Footage
Manhattan
Staten Island
Brooklyn
Bronx
Queens
Borough