Introduction:

Housing prices in New York City are among the most expensive in the entire country, with many residents being forced to prioritize affordability over desirability. Some features that are determined to increase the desirability of homes include square footage, the number of essential features such as bedrooms and bathrooms, and the overall cleanliness of the area.

The goal of my project is to use these desirability features to predict the price at which NYC homes can be sold. I have developed several models using square footage, number of bedrooms, number of bathrooms, and the number of rat sightings in the general area to predict price.

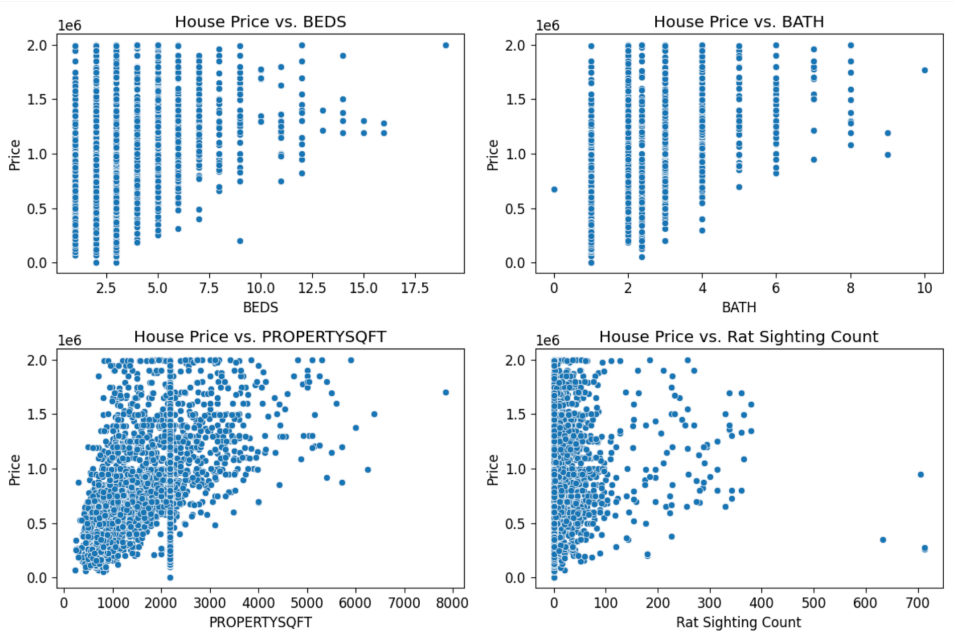
Data and Methodology:

Housing Data: https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market
Rat Data: https://www.kaggle.com/datasets/new-york-city/nyc-rat-sightings

Both of my datasets were publicly available via Kaggle. My first dataset consisted of housing data in New York City, with each entry being a home in New York that has been sold or is currently for sale. From this dataset, I gathered important features such as sale price, number of bedrooms and bathrooms, and the square footage.

To add greater depth to my models, I sought a feature that could accurately portray the cleanliness of the areas associated with each housing option. This led me to another public dataset on Kaggle, which records rodent complaints to 3-1-1 in NYC. I chose to group the number of rat sightings by street and then merge my datasets on the name of the street. I felt that with the data I was using, this would be the best way to portray cleanliness of a property. Better features can likely be found or developed, but due to time constraints and lack of common information to merge tables, I felt this would be appropriate.

In exploratory data analysis, we found that a fairly linear, positive relationship exists between sale price and square footage, with the other features not showing as strong of a linear relationship. This is demonstrated in the following scatterplots:



Mostly positive relationships can be foundamong the other variables as well, though there seems to be a lot more noise. This makes sense as number of bedrooms and bathrooms will often increase the property price, though this can be situational to fit some needs better. However, a positive relationship between rat sightings and house price is unexpected, and could possibly signal that this statistic is more about urban population than hygiene concerns. If future linear models were to be explored, this variable could possibly be transformed, though I was hesitant to perform a transformation in fear of negatively affecting other models' performances.

The models I selected to predict the sale price of each house were Multi-Linear Regression, K Nearest Neighbor Regression, and Random Forest Regression. Multi-Linear Regression tries to find the relationships between predictor variables and the target variable, allowing for accurate and interpretable predictions. In this project, K Nearest Neighbors finds the five closest neighbors to each point, averages their sales price, then predicts the sales price of new points. Random Forest Regression takes the average of many different decision trees to predict the sales price of a property.

Results and Analysis:

The following are RMSE (root mean squared error) and R-squared of the different models used:

Model	RMSE	R-squared
Multi-Linear Regression	353910.98	0.358804
KNN Regression	346268.18	0.386199
Random Forest Regression	350718.31	0.370320

Of the three models, KNN Regression performed the best, with Random Forest Regression performing slightly worse based on both RMSE and R-squared, though all three models perform fairly similarly. In all three models, the most important feature in predicting appears to be square footage. In multi-linear regression, it had a coefficient of 85.91, meaning that holding all else constat, each square foot increase in area increased the sale price by $85.91. Next most important was number of bathrooms, number of beds, and rat sightings (this surprisingly had a positive effect on price, which will be discussed in full in conclusion.)

In KNN Regression, using permutation importance, square footage was once again found to be the most important feature, being followed by bathrooms, bedrooms, then rat sightings. Permutation importance shuffles each feature individually, seeing which one causes the greatest effect on model performance. In Random Forest Regression, number of bathrooms was found to be the most important feature, followed by square footage, rat sightings, and then number of bedrooms.

Model Comparison and Discussion:

Of the models used, the decision should be between all three models as they all performed fairly similarly in predictive power, though both nonparametric models outperformed Multi-Linear Regression. I believe Multi-Linear Regression is the best option as it is much more interpretable than Random Forest Regression and KNN Regression. Making this decision will cause a slight decline in the predictive power of our model, but I believe the increase in interpretability is worth the sacrifice.

Multi-Linear Regression is considered more interpretable in large part due to the coefficient results which allow you to see the direct effect each feature has on the property price. Random Forest Regression and KNN Regression do not offer coefficients or an equation of any kind, making the results much less transparent.

Conclusion and Future Steps:

In conclusion, the model that was selected for price prediction due to predictive performance and interpretability is Multi-Linear Regression. The feature of greatest importance across most models used was square footage, which makes sense given how great of a commodity larger apartments can be in a packed city such as New York City.

My recommendation for future steps would be to find a better way to measure hygiene of the properties than the number of rats spotted in the area. In our models, an increase in rat sightings led to an increase in sale price. This seems very counterintuitive, as it is safe to assume that most people do not prefer rats to be in their home. However, apartments found in the center of the city or near popular tourist sights are going to draw a higher price, while increased population leads to increased waste, which will bring in more rats. A better metric can lightly be found that will better represent hygiene concerns in the city, though using rat sightings data can also be useful and amusing.