Final Report: NYC Property Sales Data Analysis

By Jimmy Cheng

Problem:
How can I create a model that is able to predict New York City property sales prices between 2022-2023 to within 20% margin of error based on intrinsic features of the property such neighborhood, building type, square footage?

Data:
The Department of Finance's Rolling Sales files lists properties that sold in during 2022-2023 in New York City for residential units, cooperative and condominiums, offices and warehouses. These files include features such as the neighborhood, building type, square footage, sales data and other categorical and numerical features.

Our dataset: https://www.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

Summary of results / findings:
I started by performing data wrangling, removing null values and incorrectly recorded data. I then performed Exploratory Data Analysis to look for trends and patterns, plot correlations between the columns and performed feature selection. The cleaned data is split into training and testing set and fitted through a pipeline which includes scaling the data and creating multiple models to compare results. The random forest regressor model is chosen since it produces the least RMSE and MAE with highest R squared value. I then examine the results closely and went back to our data set to create a subset without Manhattan and Brooklyn and re-ran the model. Finally, I performed hyperparameter tuning to create a model that produces the least RMSE and MAE values and selected it to be our prediction model.

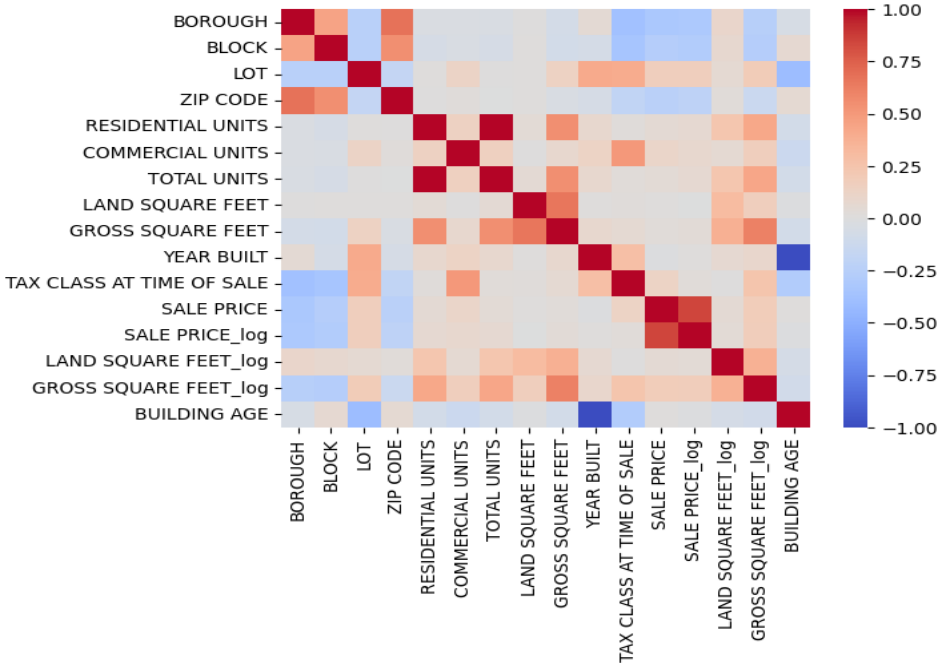
Data Wrangling:
Data is loaded and combined into a pd.DataFrame object. The complete dataset has 84391 sales observations, 21 columns. We first examine the observations with NULL values:

1) 'EASEMENT' Column has 0 Non-null columns, a closer inspection shows this as a NULL column, which can be dropped. According to the Glossary: An easement is a right, such as a right of way, which allows an entity to make limited use of another's real property.
2) Multiple BUILDING CLASS Columns were combined into a single column to reduce redundancy
3) Multiple TAX CLASS Columns were combined into a single column to reduce redundancy
4) 'APARTMENT NUMBER' and 'ADDRESS' contains information that could not be standardized, thus dropped
5) Many observations have SALE PRICE <$1000 because a low $ sale indicates that there was a transfer of ownership without a cash consideration. There can be a number of reasons for a $0 sale including transfers of ownership from parents to children. These observations are dropped.
6) The other remaining NA fields are replaced with Median observations grouped by Borough

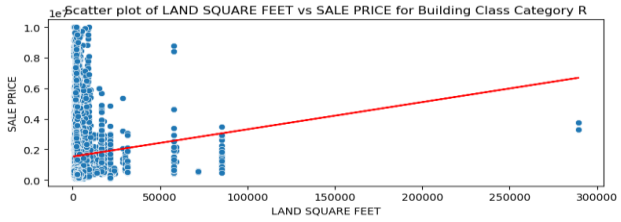
Through data wrangling, I managed to isolate a target of interest – SALE PRICE, presented in Log scale in this histogram below for observation purposes.

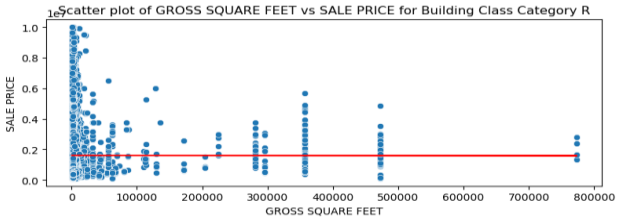
EDA:

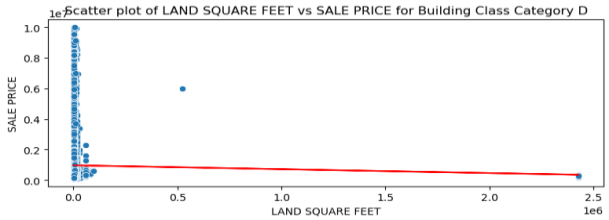
Looking at the correlation plots of my features, I found out Borough, Block, Zipcode, Gross Square Feet have higher correlation to Sale Price.

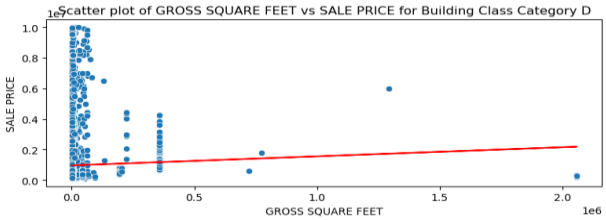


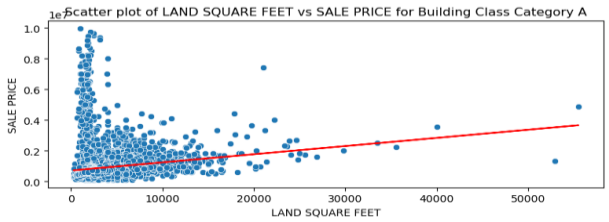
Looking at scatter plots below show that there may be some relationship between land square feet and gross square feet versus Sale Price for particular building types.

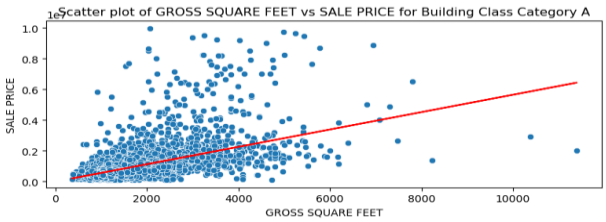


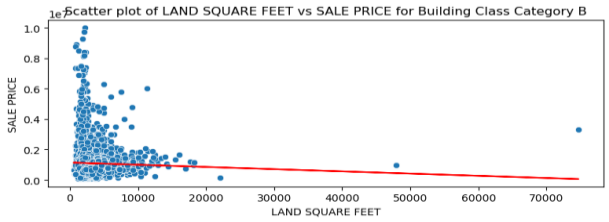


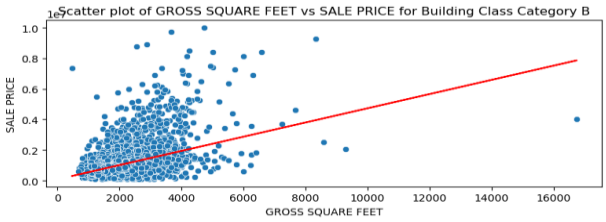


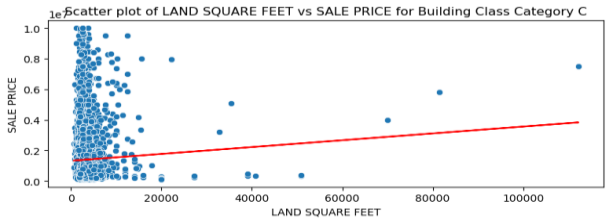


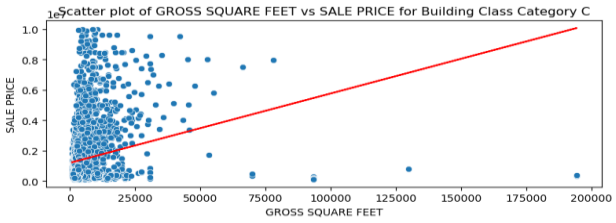










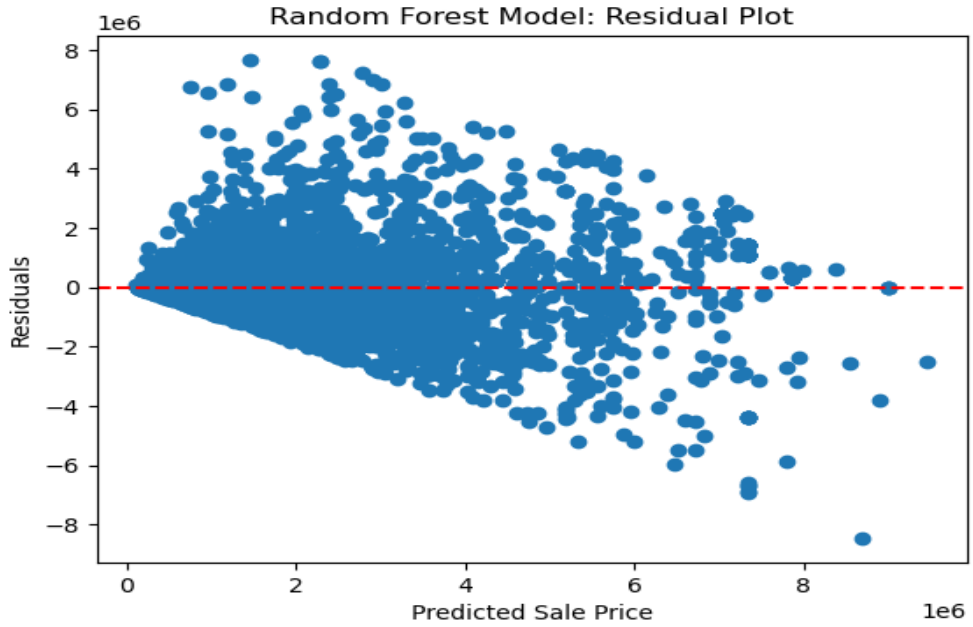


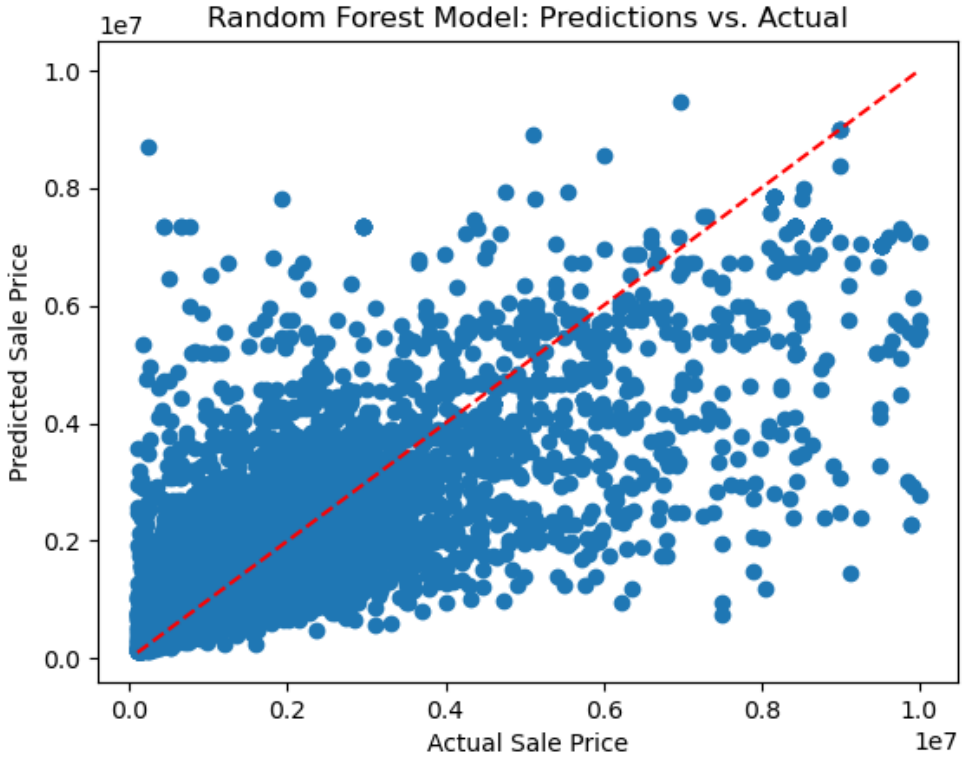
EDA allowed us to identify some hypothesis that ZIP CODE, GROSS SQUARE FEET and LAND SQUARE FEET may help us build a model to effectively predict SALE PRICE.

Preprocessing and Training Data Development:

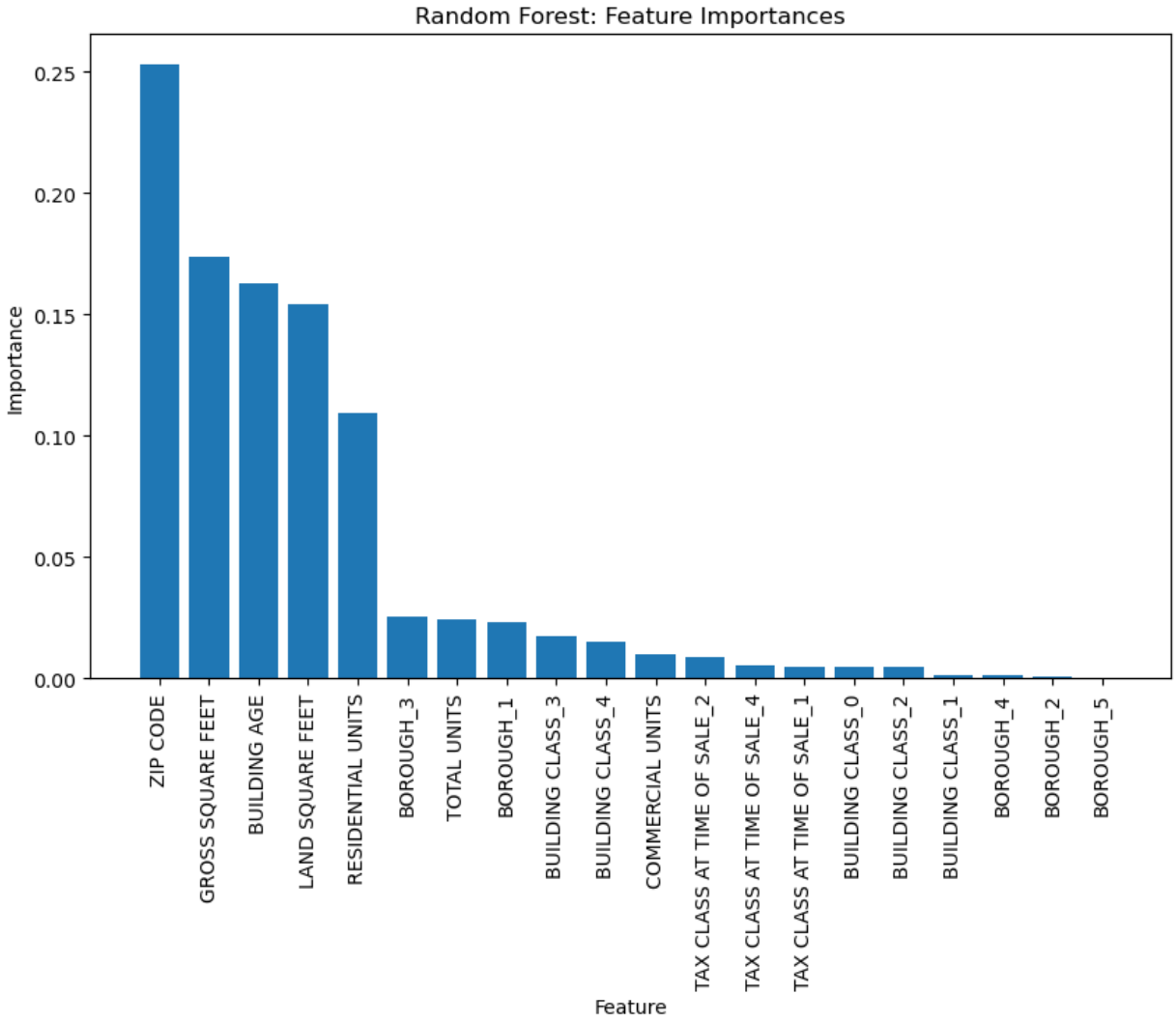
1) I performed label encoding and one hot encoding to turn the categorical columns into numerical columns
2) Data is split into a 70/30 Train/test split, K-fold cross validation (K=5)
3) Model Evaluation metrics being used are the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-squared coefficient. These metrics are selected because predicting sale price is a regression problem and these RMSE and MAE will tell how much the model predictions will differ from actual results, lower being better. R-squared coefficient will tell me how much variance is explained by the model, higher being better.

Model Selection:
I fitted the training data into linear regression, polynomial regression, random forest regressor, XGBoost Regressor, KNN Regressor and PCA + Linear Regression. Using my defined model evaluation metrics, I identified the random forest regressor produces the best results.





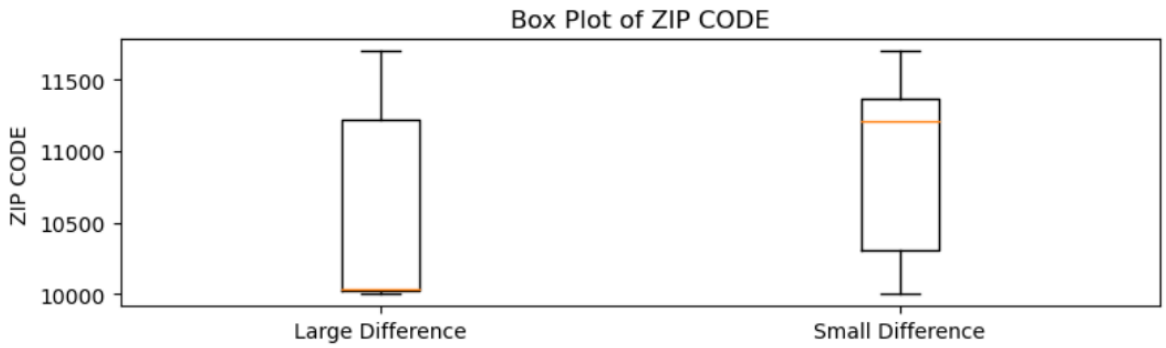

```
Root Mean Squared Error: 771417.1717335558
Mean Absolute Error: 390155.74002615956
R-squared: 0.6605585383252054
```

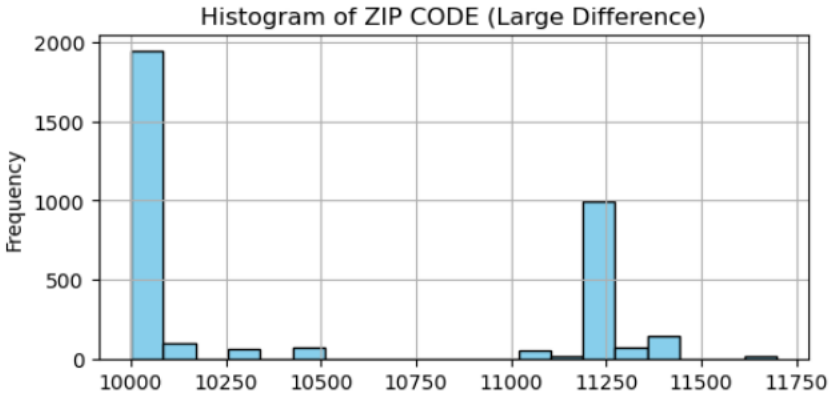


The random forest regressor model also determined that Zip Code, Gross Square feet, building age to be the top 3 features explaining the most variance. This matches my hypothesis from EDA! However, there was still too much unexplained variance and for average sale price of $1,220,000, a mean absolute error of $390,155 is still too much! I need to better understand where the error is coming from in order to fine-tune this model.

Residual Analysis:

I defined a residual of $500,000 to be too large created a subset of observations with large residuals. Looking at the model's leading feature: Zip Code across observations that create large residuals and observations that have small residuals, I notice most large residuals are in the Zip Code of 10000, which corresponds to Manhattan – Borough 0. In fact, 46% of the model's predictions in Borough 0 create large residuals and 23% of the model's predictions in Borough 3 (Brooklyn) create large residuals.

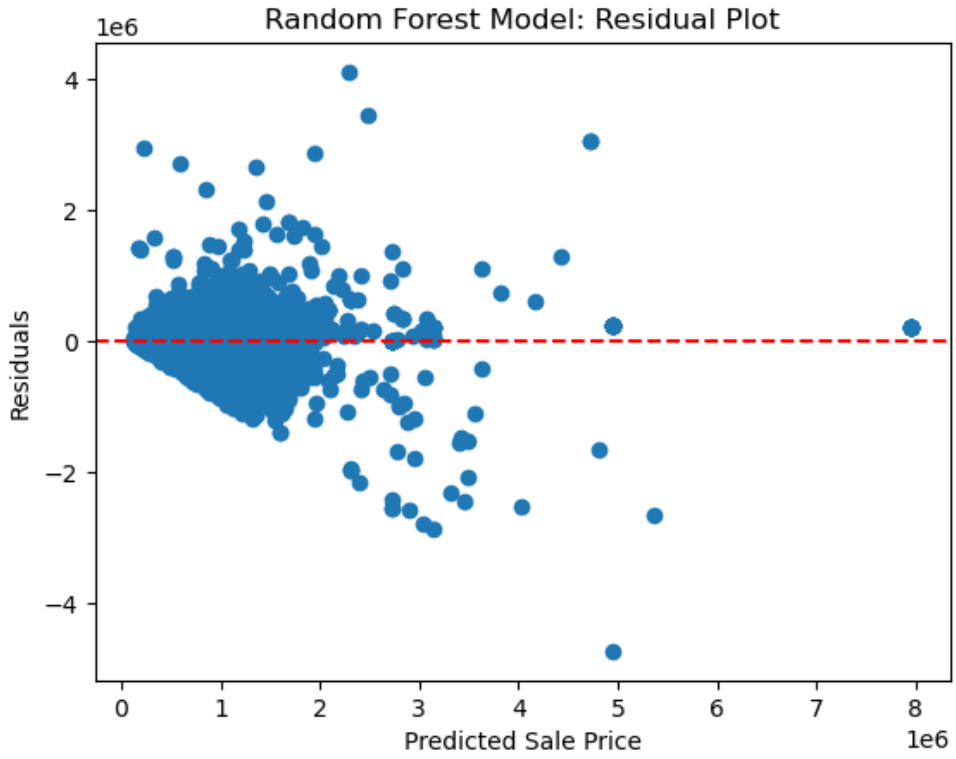


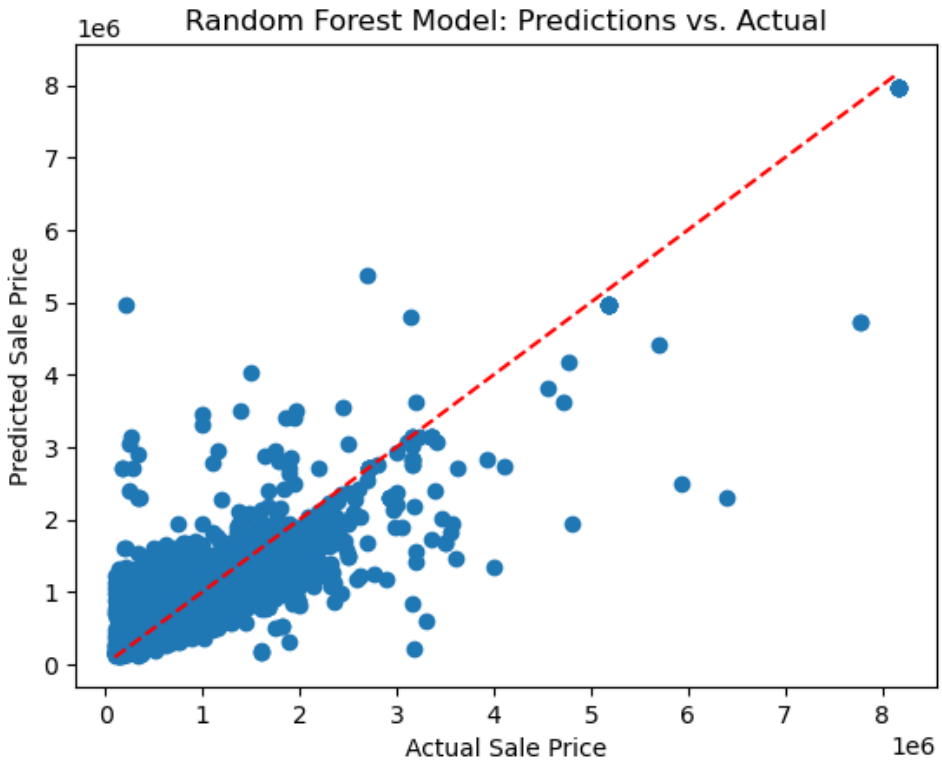



```
% of units per Borough predicted incorrectly
0.4592674805771365
0.056588520614389654
0.2334355104928083
0.0636211931581143
0.024475524475524476
```

I believe the reason why my model is unable to correctly predict sale prices from those regions is because properties in those regions often have extrinsic premiums attached and not reflected by my dataset, which only contains intrinsic data about the property! These extrinsic premiums may in fact be non-linear or varies with time. With the lack of further data, I have decided to eliminate observations from Manhattan and Brooklyn and limit my prediction problem to the remaining three boroughs of NYC.

Final Model and HyperParameter Tuning:
My final model is a random forest regressor model that predicts the Sale Price of properties in Queens, The Bronx and Staten Island of NYC.






```
Root Mean Squared Error: 289961.4004652916
Mean Absolute Error: 158457.16708724495
R-squared: 0.7437885449430531
```

The model now has an RMSE of 289,961 and MAE of 158,457 with R squared value of 0.743.
This is a much accurate model and is now able to predict Sale Prices to within 20% margin of error.

I performed a RandomizedSearchCV and obtained the results

```
Root Mean Squared Error: 288290.6234495329

Mean Absolute Error: 157226.32305117117
R-squared: 0.7467326534983358
```

Best Params:

```
Best Hyperparameters:
n_estimators: 900
min_samples_split: 5
min_samples_leaf: 1
max_depth: 60
bootstrap: True
```

Conclusion:
Out of the box – Random Forest Regressor was the best model predicting Sale Price within NYC, outperforming Linear model, K Nearest Neighbor Regressor and XGBoost Regressor. However, there was still too much variance not explained by the model. The model also produced an error that is greater than 20% of the prediction's value.

By reducing the scope of the problem down to only three boroughs of NYC, namely Queens, The Bronx and Staten Island, I am able to create a model that can predict sale price of units of 2022 – 2023 to within a margin of error of 20%. The most important features are zip code and gross/land square feet, which makes sense that property values are largely determined by location and square footage.

Future research topics would include obtaining more data to improve the predictability of our random forest model. For example, what makes Manhattan and Brooklyn property prices unpredictable? Is there an extrinsic premium that can be predicted through other data, like mortgage rates, interest rates, employment data, demographics? Due to time constraint, I am unable to source reliable data to further explore these relationships. Having a latitude/longitude coordinate system for address would probably be helpful and allow our model to map sale prices to geography easily.