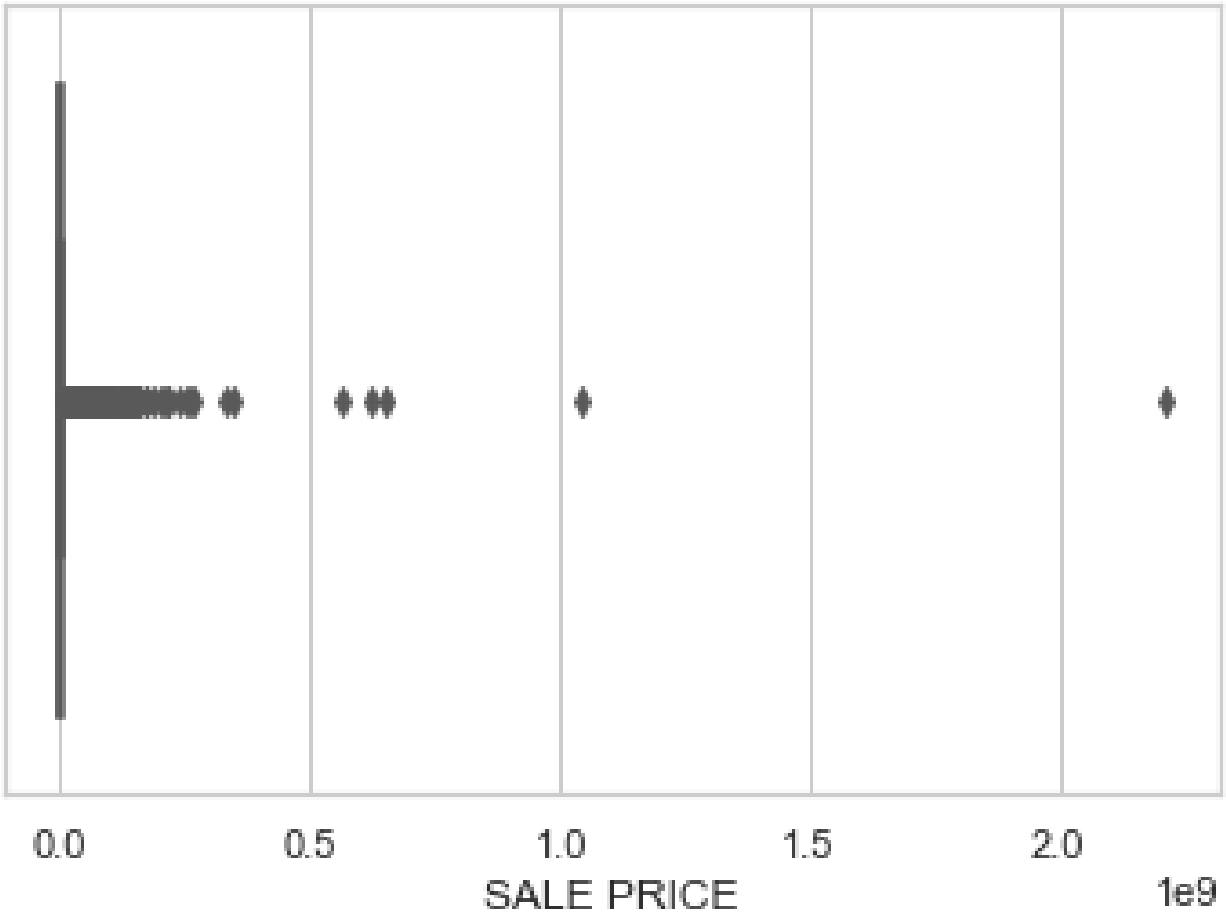
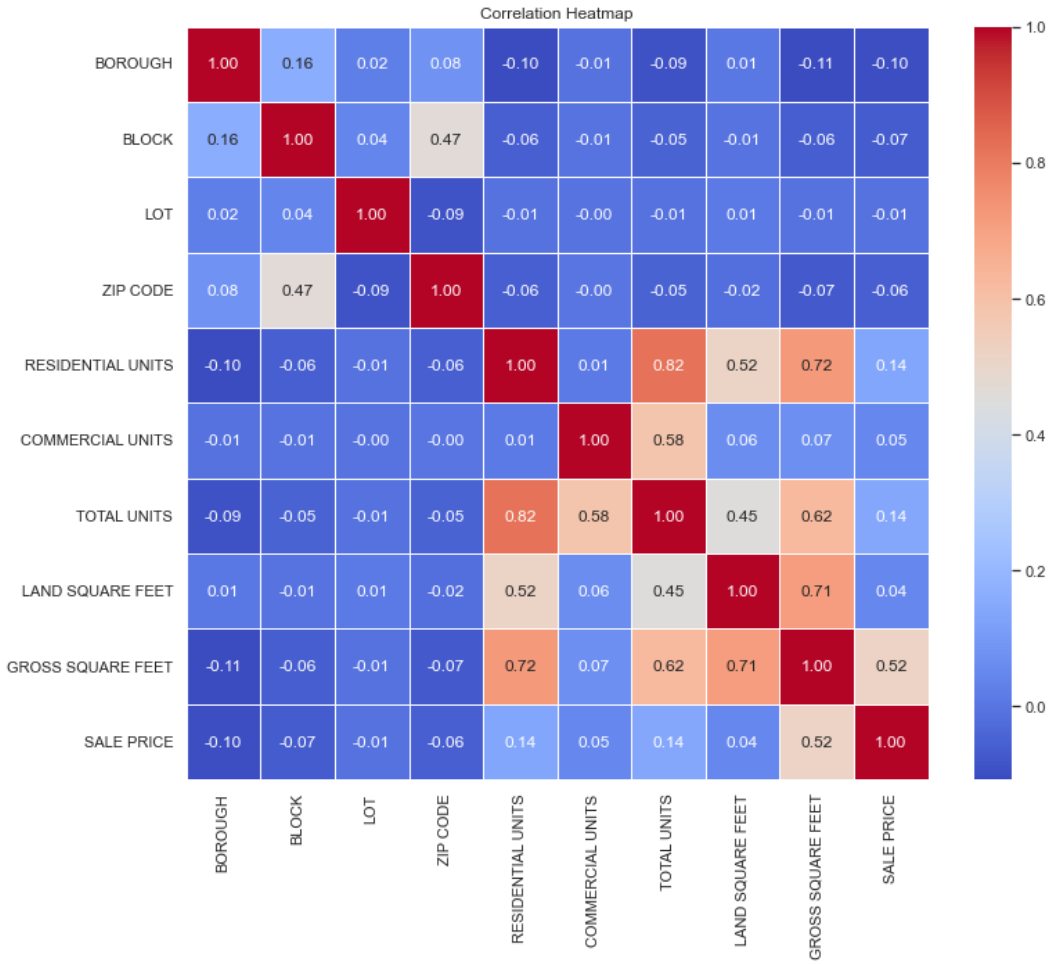
NYC Property Sales Value

Proposal

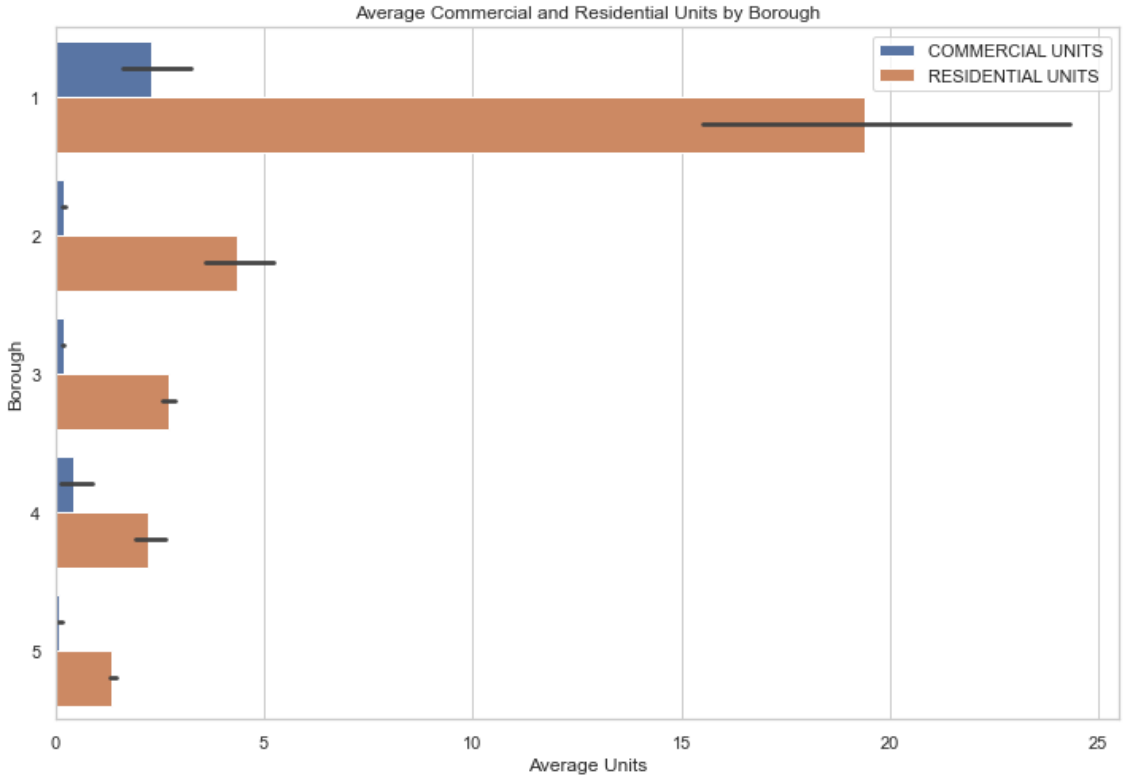
- **Abstract-** This white paper explores the drivers of real estate property values. Buyers,sellers, and investors in real estate all thirst for accurate real estate property assessments. Robust forecasting will drive the efficiency of the real estate market at-large. I will use Machine Learning Regression models to choose the best features and then to accurately predict property values based on these features. These models promise to give insight into what interactions of features drive price.
- **Business Problem –** How can we better inform potential buyers and sellers of buildings in NYC what the true value of a property is? This information is invaluable in making business decisions. If the asking price of a property is well above the predicted value, one should be wary of investing in it until he can account for the price hike.
- **Background/History-** There are many factors which affect real estate value. These include location, square footage, real estate comparables or comparable properties' value, age of property, condition of property, strength of economy, and interest rates (Sofi, 2023). One approach to evaluating a property is called the "income approach" (The three approaches to value - province of Manitoba). This approach asserts that the value of a property is driven by the income that it generates. Predicting real estate values more accurately will only make the market more robust. Lenders need good evaluations of real estate values in order to calculate how much they can loan and at what interest rate. Market transparency leads to healthier real estate transactions. Clarity vis-a-vis pricing perforce builds trust amongst all principals.
- **Data Explanation-** The data I will use comes from "Rolling Sales Data" from the City of New York (nyc.gov). I accessed the data through Kaggle (City of New York, 2017). Kaggle merged several datasets to include sales records from all five boroughs from September 2016 to September 2017. This dataset includes all property sales from this 12-month time period. The Department of Finance of The City of New York provides a Glossary of Terms for Property Sales Files which serves as a compendious data dictionary (Glossary of Terms). SALE PRICE is the target variable.



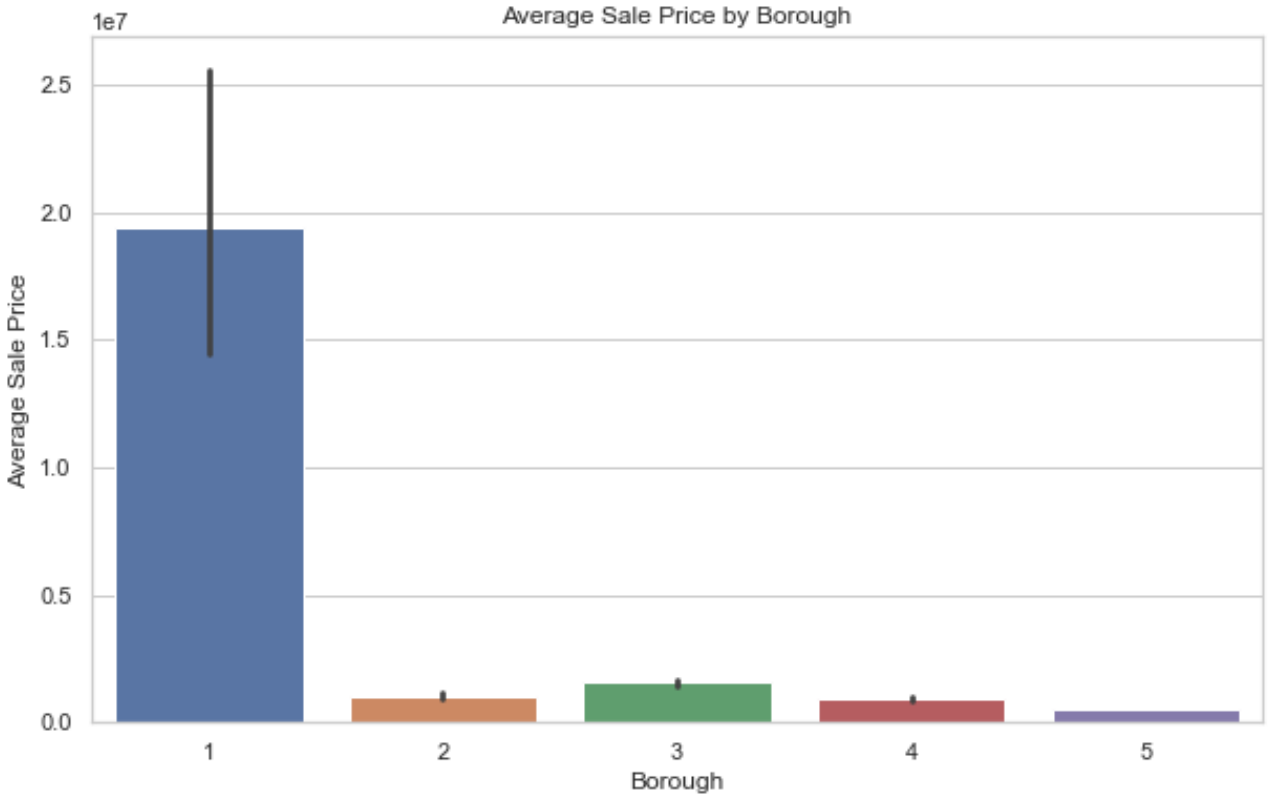
-
- Note the preponderance of data points at 0, or with no consideration. These are obviously not sales.
- It describes the fields included in the dataset. These fields include borough, neighborhood, building class category, tax class, tax block, tax lot, whether this property has an easement, building class at present, address, zip code, residential units, commercial units, total units, land square feet, gross square feet, year built, building class at time of sale, sales price, and sales date. Borough 1 is Manhattan, 2 is The Bronx, 3 is Brooklyn, 4 is Queens, and 5 is Staten Island.



- Note the significant (0.52) correlation between GROSS SQUARE FEET and SALE PRICE. LAND SQUARE FEET, TOTAL UNITS, COMMERCIAL UNITS, and RESIDENTIAL UNITS are each poorly correlated with SALE PRICE. Both COMMERCIAL UNITS and RESIDENTIAL UNITS make up TOTAL UNITS and therefore are correlated.



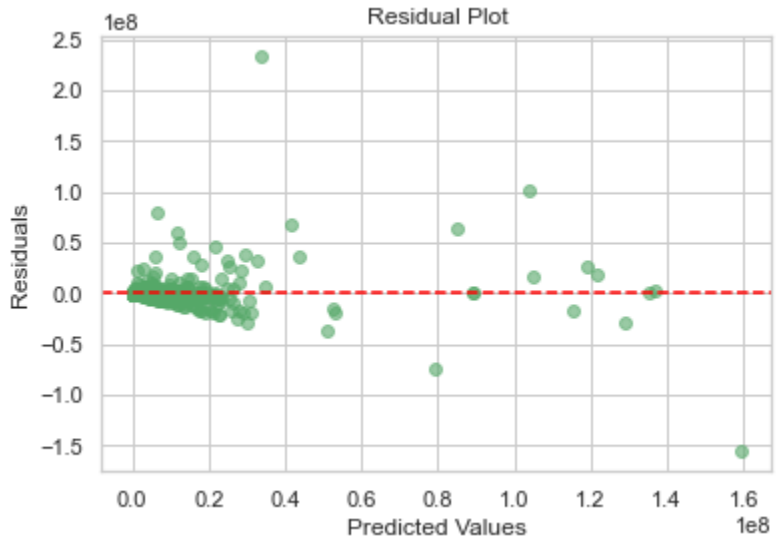
- Manhattan far outclassed all the other boroughs in both average commercial units and average residential units. Queens displayed more commercial units than the other three, but is a distant second to Manhattan.
- **Methods-** I will have to wrangle this data to make it usable. While the dataset includes 84,547 entries, most of the rows are unusable, whether due to null values, or other data issues. This is a regression problem with many categorical features. I used one-hot encoding on the BOROUGH column. This feature has great impact on predictions. I did a Principal Component Analysis (PCA) on the features and then created a myriad of ML models using principal components, but these models did not prove to be robust.
- I did a Linear Regression on all features as a baseline and it produced a good model. I explored other Linear Regression models with different combinations of features. I built a Decision Tree, a Random Forest, a Support Vector Regression (SVR), a kNN Regressor, a Lasso Regression, a Ridge Regression, an Elastic Net Regression, tuned models of Random Forest and Ridge Regression, XGBoost Regression, and a bagged regression with Random Forest.
- I created an 80/20 train/test split across the board for the myriad of models. This means that the model has never seen the test data befoe validating on it. It prevents overfitting because one would see a degradation of performance from the train to the test. I displayed the folowing standard metrics for a regression problem for each model: Mean Squared Error, Mean Absolute Error, R-Squared, Explained Variance Score, Maximum Residual Error, and Median Absolute Error.
- **Analysis –** Linear Regression probably performed so well because the features satisfied the assumptions of linearity, homoscedasticity, and normality (Simple Linear Regression). The most robust model for this data is RandomForest with R^2 = 0.5874, even better than the tuned RandomForest which produced an R^2 = 0.5406. R^2, or Coefficient of Determination, give us a measure of the goodness of fit. It is a statistical measure of how well the regression line fits the actual data (Numeracy, Math, and Statistics). The Random Forest algorithm is an ensemble method which leverages both bagging and feature randomness to create an uncorrelated forest of decision trees (What is Random Forest? IBM). It reduces overfitting and deals well with missing values. See Appendix 1 for best models' metrics.
- **Conclusion-** Any one of the top six models (Appendix 1) will give a stong prediction of property value based on the available features. Use the Random Forest model because it is the best. Borough is such a factor in property value that one must first choose whether he wishes to buy and sell in Manhattan where values are incomparably higher than the other four boroughs.



Square footage, especially gross will drive property value. Invest in footage.

- **Assumptions-** This project assumes independence. The Assumption of Independence means that no two observations in the dataset are related or affected by one another (Zach, 2021). My dataset includes all property sales in NYC during that year, which is juxtaposed to simple random sampling, which is the best way to avoid dependence. The dataset does not give information about the buyers or the sellers. Did any of the principals participate in more than one sale? It does not tell us if one sale spurred another sale, or prevented another sale. If the assumption of independence is violated, the models' efficacy is called into question. I asssume that the dataset is reliable because it is a government one. It appears to be a big assumption, especially considering how I had to discard most of the data because it was unusable. I assume that the features I modeled were the best drivers of property value.
- **Limitations-** The data is limited to a one-year period. One need be leery of extrapolating to other years where market conditions vary. Property values changed so much over the past year because of interest rate movements. These same conditions did not exist during the 2016-2017 time period. This project is geographically limited. Because there is no place in the world like NYC, one cannot simply extrapolate to other locales. One must adjust to the local variations in real estate conditions. Just outside of NYC, for instance, square footage is not the end-all. Above all, the real estate market is dynamic and therefore one must be wary and adjust for any data which is not real-time.
- **Challenges-** Data quality is an ongoing challenge with this project. I was able to use only a third of the year's property sales because of the poor recording of sales. This is surprising because the data source is nyc.gov. Outliers must always be investigated, both to ensure that it is a correct input,

and even after confirming its veracity, and to explore whether these outliers unduly affect the models. While the Random Forest model is most robust, it is a black-box model. One must endeavor to explain these models to the stakeholders to maintain transparency. One must carefully explain how the ensemble model comes together in order to get buy-in.

- **Future Uses/Additional Applications-** This project may very well impugn the "income approach" mentioned above. I performed predictions by borough using the mean for each feature. I made four predictions for each borough: one by adding a residential unit to the mean, the second by subtracting a residential unit from the mean, a third prediction by adding a commercial unit to the mean, and a fourth prediction by subtracting one commercial unit from the mean.I made each prediction assuming the mean in the other features. I found that subtracting one residential unit in Manhattan led to almost a one million dollar appreciation. Subtracting one commercial unit from the mean led to an appreciation of more than one-and-a-half million dollars over adding one commercial unit. The Bronx and Staten Island showed the same trend. The "income approach" demands that adding units adds rents and therefore adds value.
- **Recommendations-** Use the Random Forest model with confidence.
- Note the random distribution of points and the congregation around 0, signs of a robust model.
- 

- Educate the stakeholders. Explain that a prediction is likely, or very likely, to come true and to weigh business decisions based on that. Anyone who has access to these predictions will perforce have the upper hand in negotiations because they are privy to the most "real", accurate price. Always keep an eye out for new possible features which potentially capture the data even better.
- **Implementation Plan-** Get the model up and running. Feed it the name of the borough, the square footage, and the units and leverage the robust predictions to have a good idea of the property value. One must be cognisent of any other contributing factors like a neighborhood becoming trendy, i.e. the place to be. Negotiate from a position of strength. The price you are quoting is the real value of the property.
- **Ethical Concerns-** There must be clear documentation of all decisions of how to make the data usable, from imputation for null values to excluding rows. One must ensure that this

descriptive personal data is allowed to be used, both legally and ethically. One must clearly describe the algorithms used to predict values so the consumer will understand what the models mean. Beware of historical discriminatory practices perpetrated by the real estate community. From the practice of redlining developed by the Home Owners' Loan Corporation which "redlined" black households as too risky to lend to, to predatory mortgages and loans which disproportionately injured people of color, the real estate business has been anything but inclusive (Historic Housing Discrimination). One must not perpetuate this. One must look at the granular data areas and prevent gentrification, even at the expense of profit. Increased real estate values may translate into unbearable costs for the economically disadvantaged. One must prevent models from hurting the disadvantaged. Open up lines of communication with the local community and its residents to inform them in real time how the real estate industry will mold the community's property pricing. Be responsive to stated concerns.

- Because of the real estate industry's woeful ethical history, it is imperative to conduct ethical audits. There are multiple parts to an ethics audit (How to conduct an ethics audit). The process of the audit is to compare actual company behavior with the written ethical code of the company. The company must strive to create ethics metrics to measure its adherence to its ethical standards. Exercise vigilance in looking out for more areas to improve in ethics. Use ethics violations as a rubric for continuous education.

References

Glossary of terms for property sales files - nyc.gov. (n.d.). https://www.nyc.gov/assets/finance/downloads/pdf/07pdf/glossary_rsf071607.pdf

C. of N. York (2017, September 22). *NYC property sales*. Kaggle. https://www.kaggle.com/datasets/new-york-city/nyc-property-sales

Historic housing discrimination in the U.S. Habitat for Humanity. (n.d.). https://www.habitat.org/stories/historic-housing-discrimination-us#:~:text=Black%20families%20were%20prevented%20from,redlined%E2%80%9D%20areas%20unsafe%20for%20lending.

Numeracy, Maths and statistics - academic skills kit. (n.d.). https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html

Real estate contract basics. Wise. (2017, June 7). https://wisepropertymanagement.com/real-estate-contract-basics/#:~:text=Consideration%20is%20anything%20of%20legal,contract%20is%20not%20legally%20enforceable.

Rolling Sales Data. (n.d.). https://www.nyc.gov/site/finance/property/property-rolling-sales-data.page

Simple linear regression. (n.d.). https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html#:~:text=There%20are%20four%20assumptions%20associated,are%20independent%20of%20each%20other.

SoFi. (2023, December 31). *7 important factors that affect property value*. https://www.sofi.com/learn/content/factors-that-affect-property-value/

The three approaches to value - province of Manitoba. (n.d.-b). https://www.gov.mb.ca/mao/public/fact_sheets/approaches_to_value.pdf

Wall, S. (2023, June 26). *Brokers criticize LLC transparency act*. The Real Deal. https://therealdeal.com/new-york/2023/06/26/brokers-cry-privacy-in-pushback-on-bill-exposing-llcs/

What is Random Forest?. IBM. (n.d.). https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

Zach. (2021, April 12). *What is the Assumption of Independence in Statistics?*. Statology. https://www.statology.org/assumption-of-independence/

Appendix A

Model Metrics Comparison

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R-squared (R
Tuned RandomForest	24172237304663.914	789729.41	0.5406
Ridge Regression	50297199319095.74	1321224.69	0.5286
RandomForest	26042280118408.414	775307.53	0.5874
Linear (Gross & Total)	139070168737840.92	1535408.61	0.54
Linear (Gross)	149623544146030.88	1382152.65	0.51
Linear (All)	155075890534231.84	1726621.12	0.49

Appendix B

Unit Predictions

	1 Residential Unit More Than the Mean	One Residential Unit Less Than the Mean	One Commercial Unit More Than the Mean	One Commercial Unit Less Than the Mean
Manhattan	$12,693,147.12	$13,619,417.14	$13,185,238.05	$14,735,632.62
Bronx	$583,473.49	$669,737.74	$643,224.76	$643,347.62
Brooklyn	$1,663,263.79	$1,095,173.56	$1,647,062.06	$1,134,386.28
Queens	$1,260,107.65	$1,182,891.90	$1,339,586.33	$1,041,043.13
Staten Island	$614,241.55	$857,385.06	$646,237.26	$706,012.26