Understanding New York City Property Transactions

Jeffrey Ebert

Objective

- Understand how the thousands of property transactions throughout a one year period in New York City are categorized

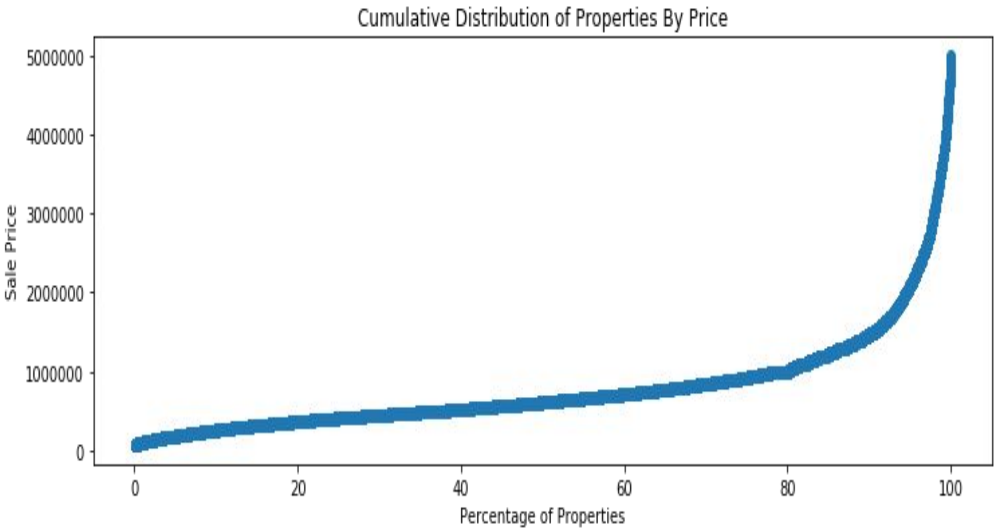
- Evaluate the varying types of transactions made with the distinctions between them and which have a significant influence on the sale price of a certain property

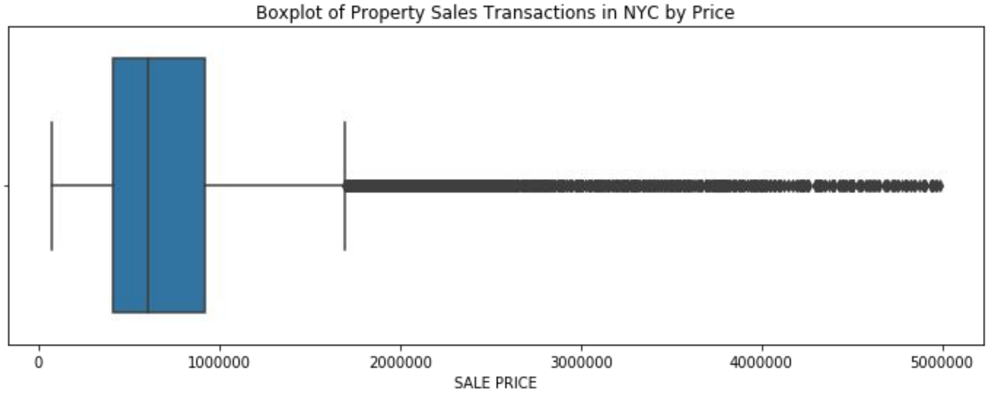
- Run machine learning models to predict values for the price of a property based on its features

Data

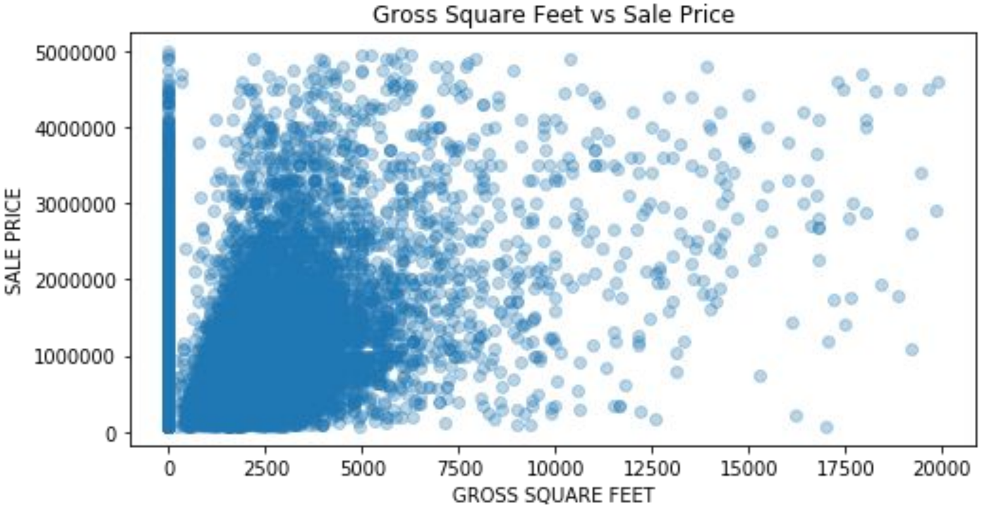
- The data used for this project consists of over 80,000 property transactions in New York City from September 2016 to August 2017.
- Each data entry lists the attributes of the property. This includes its borough, neighborhood, land square footage, gross square footage, amount of residential and commercial units, year the building was built, date of the sale and its associated price
- The dataset required extensive cleaning and removal of missing and inaccurate data values to ensure the accuracy of the values and analysis made
- The dataset was uploaded via a public dataset on Kaggle.com

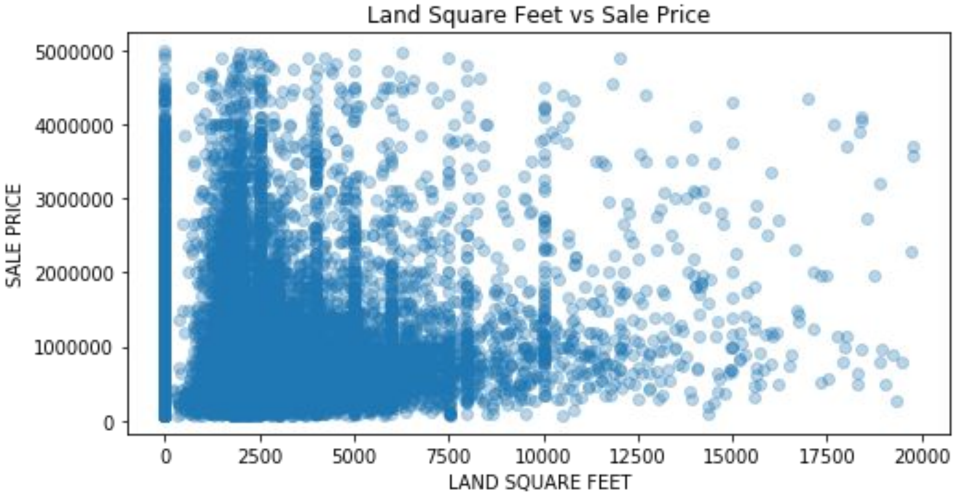
Exploratory Data Analysis



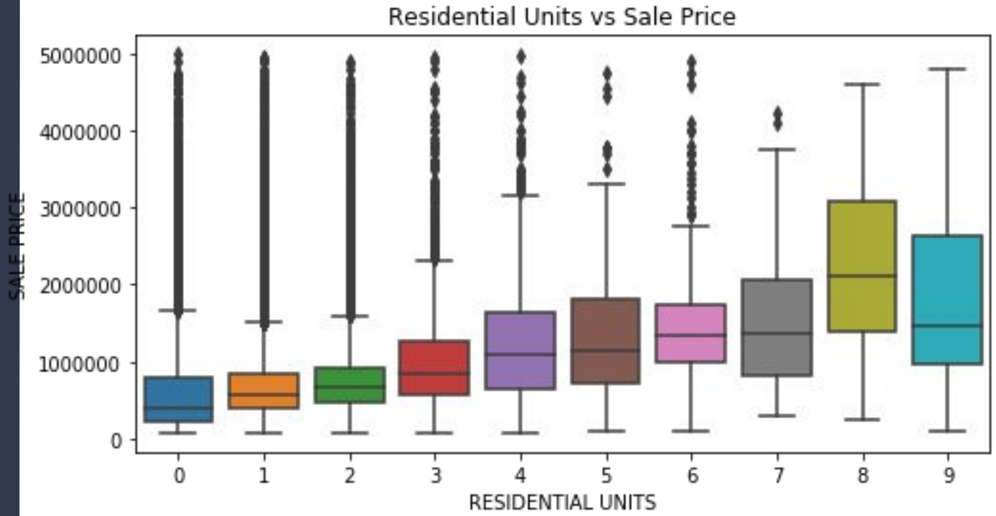


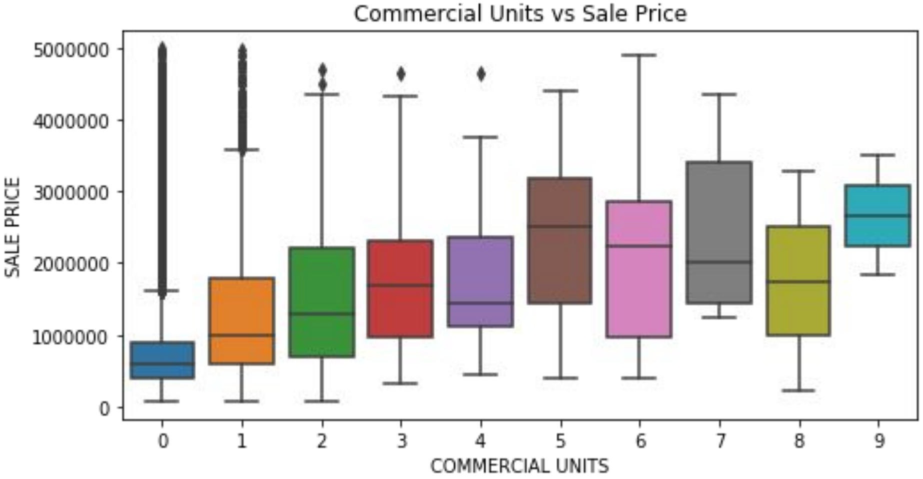
Exploratory Data Analysis



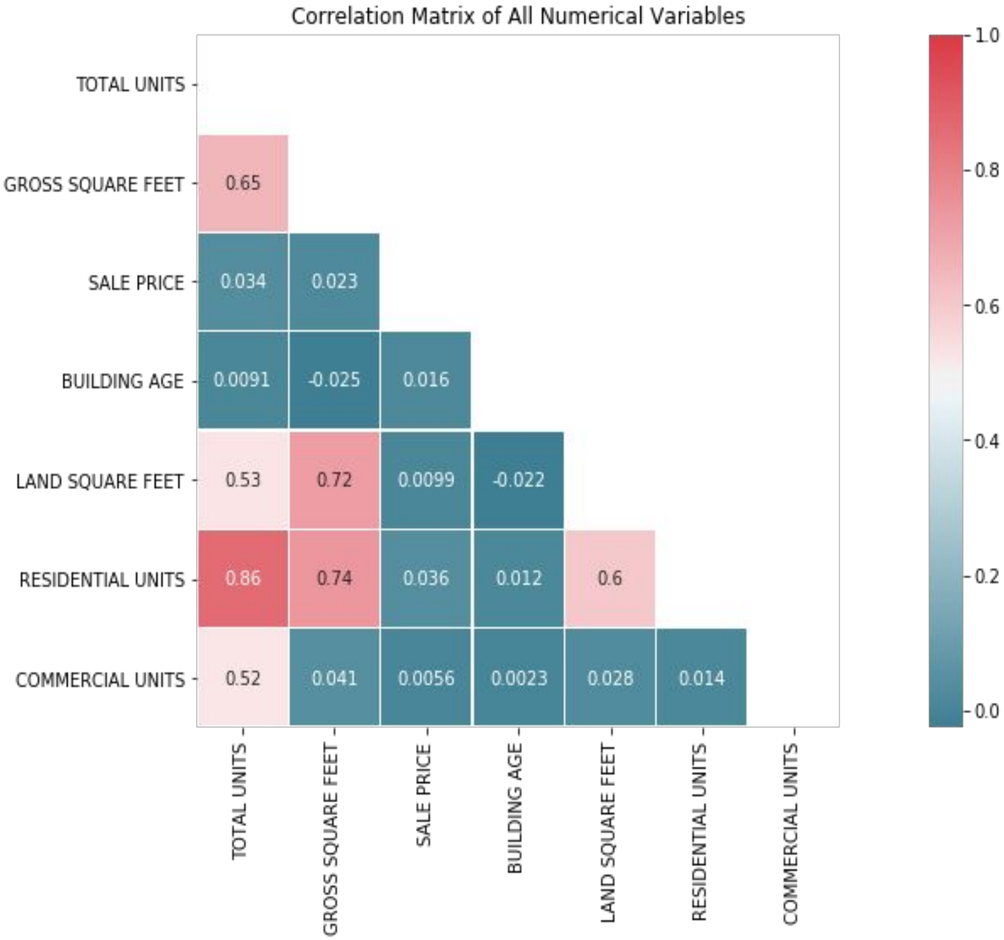


Exploratory Data Analysis





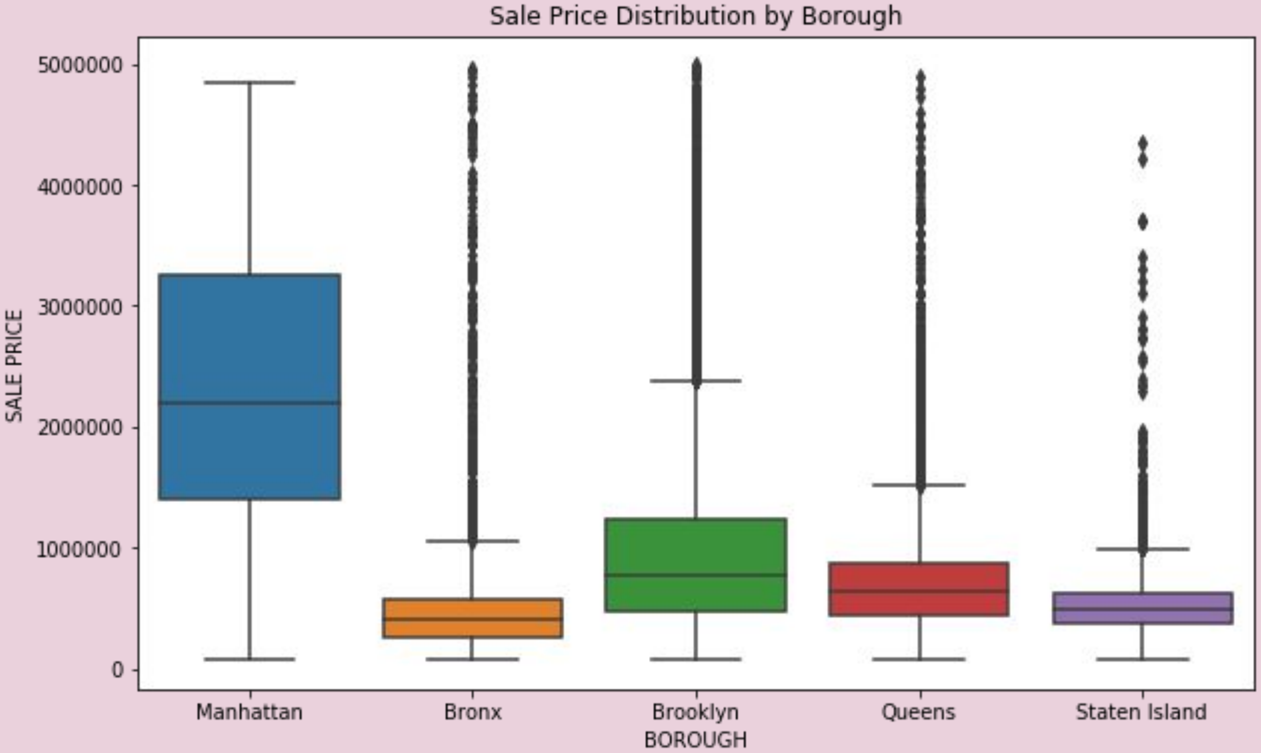
- This correlation matrix represents the relationship all numerical variables have with one another, including price.
- Based on the matrix and the previous visuals the most significant variables in determining the price of a property are its amount of residential units and gross square feet



Correlation Matrix of All Numerical Variables

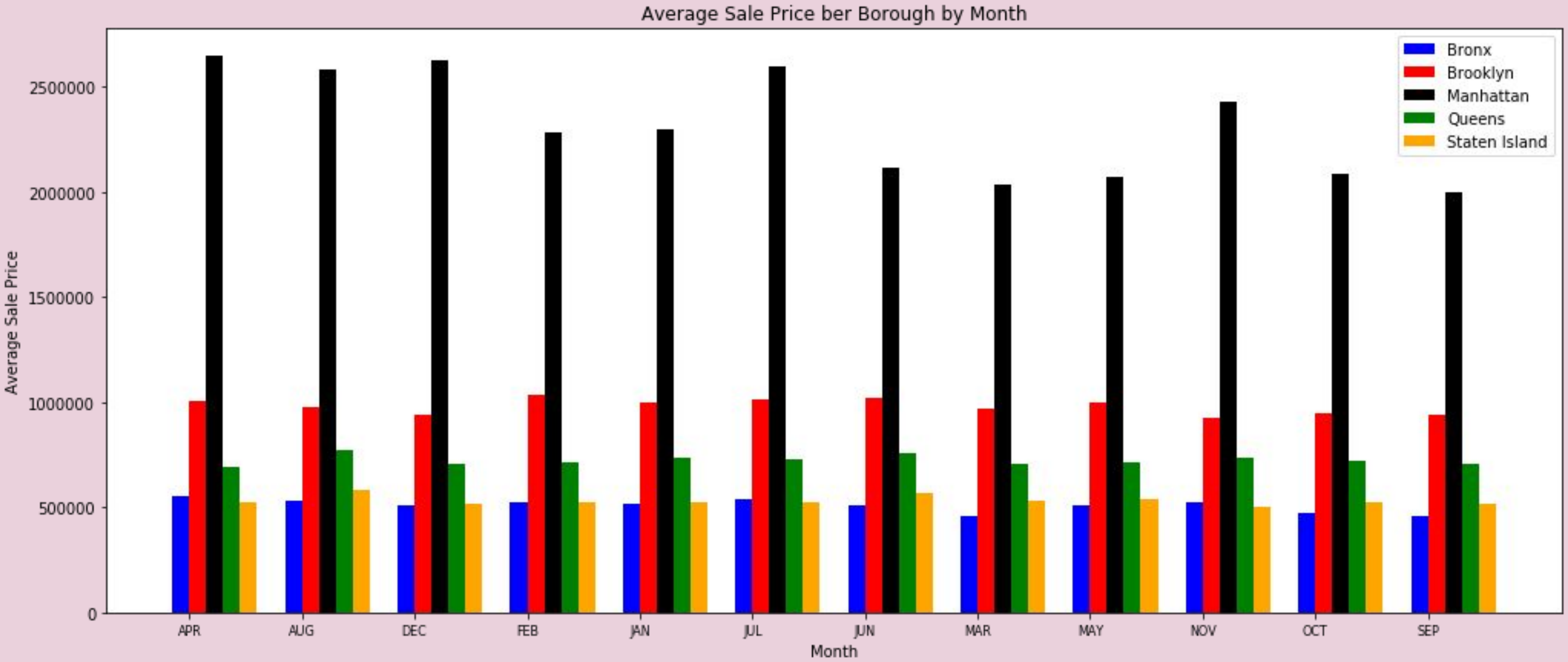
	TOTAL UNITS	GROSS SQUARE FEET	SALE PRICE	BUILDING AGE	LAND SQUARE FEET	RESIDENTIAL UNITS	COMMERCIAL UNITS
TOTAL UNITS							
GROSS SQUARE FEET	0.65						
SALE PRICE	0.034	0.023					
BUILDING AGE	0.0091	-0.025	0.016				
LAND SQUARE FEET	0.53	0.72	0.0099	-0.022			
RESIDENTIAL UNITS	0.86	0.74	0.036	0.012	0.6		
COMMERCIAL UNITS	0.52	0.041	0.0056	0.0023	0.028	0.014	

Median Price by Borough



- Based on these plots, Manhattan is the most expensive property, followed by Brooklyn, Queens, Staten Island and the Bronx

Monthly Sale Price by Borough



Models and Conclusion

- With multiple numerical and categorical variables of properties to evaluate price, various machine learning models on the data were used to predict the price of a certain property.

- Out of the linear regression, random forest, ridge regression and elastic net regression models, the random forest performed the best at predicting price with a root mean squared error of .745 and an R squared value of .447.

- While this and the other models created did not do a good job at eliminating error in evaluating the sale price of a property, it can likely be attributed due to a high variance of data and not having enough information pertinent to factors associated with a property’s true value.