Capstone Project: Milestone Report 2
Jeffrey Ebert

New York City Property Sales

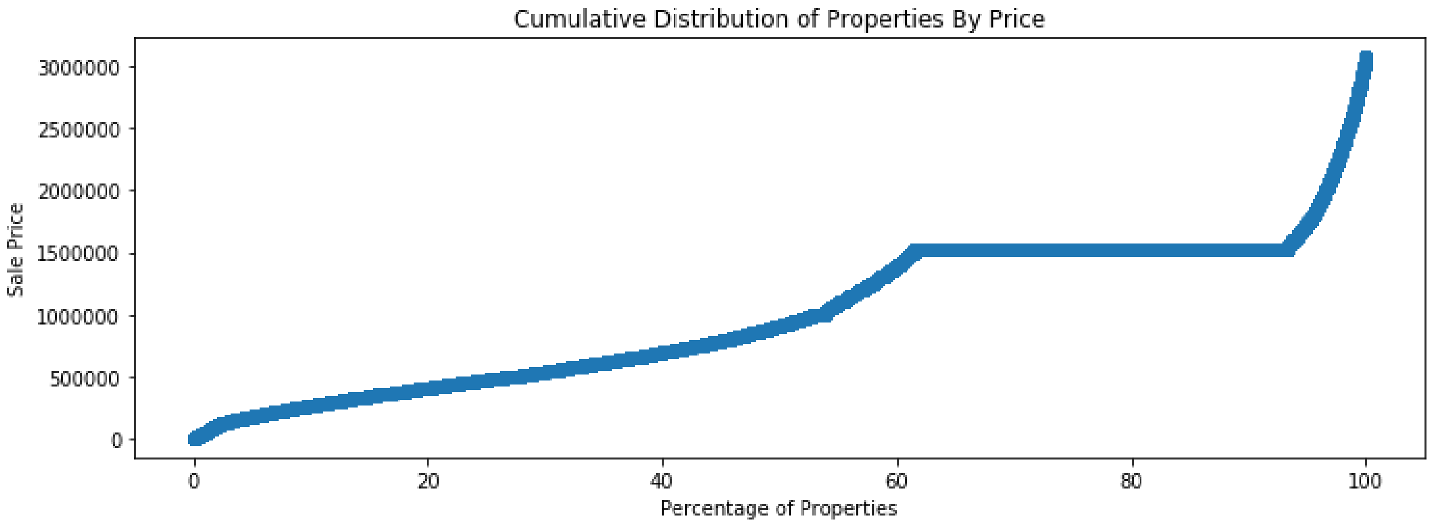
New York City, the most populated city in the United States, has 8.6 million inhabitants. As a result, it is a real estate hub. Multiple thousands of property transactions take place each year, varying from the buying and selling of different styles of homes, apartment buildings, offices, factories etc. As someone looking to buy or sell property in New York, it is important to know how to properly evaluate the price of a property. With various factors regarding a building's attributes such as its size, location and more, it can be difficult to ensure as a seller that you are providing a competitive price where you can get the most for the building's value without pricing too high and making the property overly difficult to sell. And as a potential buyer, it is equally important to make sure you are not overspending for a property based on its true value and to understand what are the most important determinants in the price of a specific property based off of similar New York City properties.

The data used for this project is comprised of property transactions in New York City from September 1st 2016 to August 31st 2017. The public dataset was uploaded from Kaggle.com. There are 84,548 original data entries, each denoting one property transaction and all of its available known features. The listed features are a property's location, including its borough, neighborhood, and address, the property's building class, tax class, size as represented by amount of commercial and residential units, land and gross square footage, the year the building was constructed, the date of the sale and the price that it was sold for.

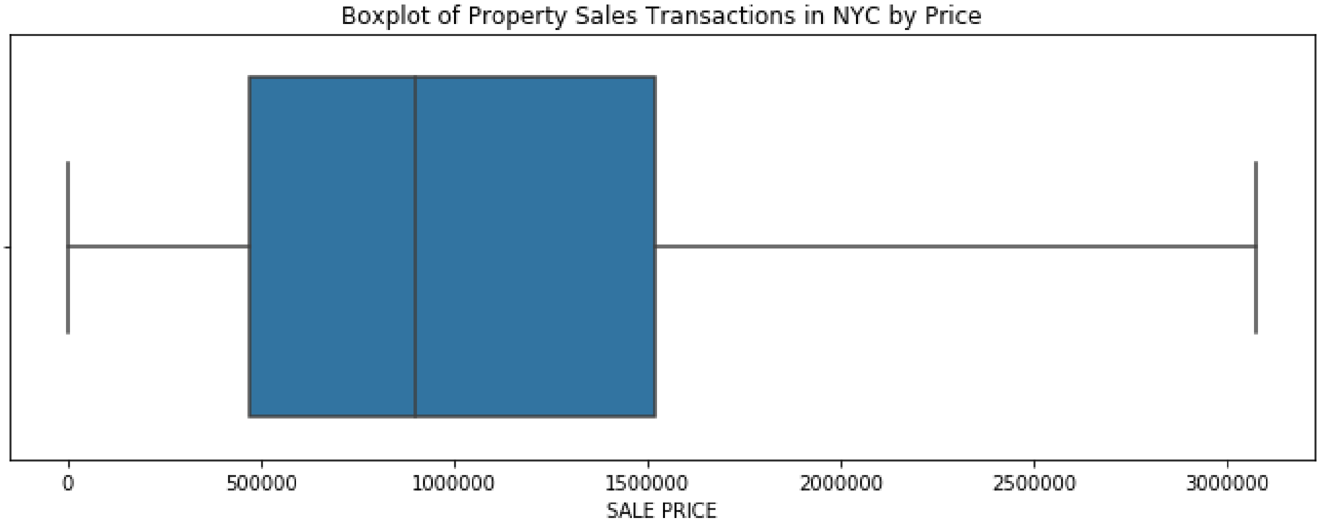
There were flaws in the data that required cleaning found by exploring the dataset. The dataset contained columns that provided no or useless information. These data columns, labelled EASEMENT, UNNAMED, and APARTMENT NUMBER were all removed. The data contained 765 duplicate values of property sales. These entries were removed from the dataset so each sale would not be counted an extra time. The columns of the square footage, building age, and price were converted to numerical data types, while the columns of tax classes, zip code, and lot were converted to categorical data types. The borough data column was listed as values from 1-5 designated the borough the property was in. To replace this with the proper name, the values were cross referenced by neighborhood to accurately label the borough titles.

There were data entries that included missing values and values of zero for a property's square footage, building year, and price. These missing values needed to be treated differently based on the variable. There were null or zero values for the square footage data columns for almost half of the dataset. Because there were so many values, it was best not to remove them. Instead, they were filled with the mean value for their type of square footage. There were also a small amount of high outliers for square footage that were removed when doing analysis based on square footage as to not significantly skew the data. There were over 6,000 data values that had a zero value for the year the building was built. These values were excluded when doing analysis involving the building age. The sale price data column also contained many null values and prices listed as zero or other very small values that could not have been possible for an actual property transaction in New York City. In order to protect the accuracy of the data available, all data values for price that were missing values or listed as $1,000 and lower were filled with the average price of all other property transactions. There were also over 4,000 upper bound outliers in the data by price that significantly altered the data mainly due to sales on Manhattan properties that were much greater than any other properties sold. These values were removed from the dataset to provide a more concise grouping of the data.

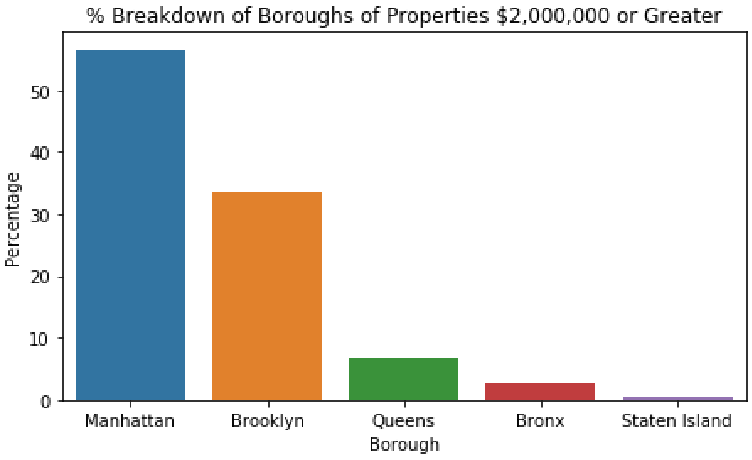
The first objective of the project is to understand how the data is separated by different features and evaluate the various types of transactions made with the distinctions between them that have an influence on the property's sale price. In doing so, we will first analyze the numerical features of a property. Below is a cumulative distribution graph of price with the outliers removed.

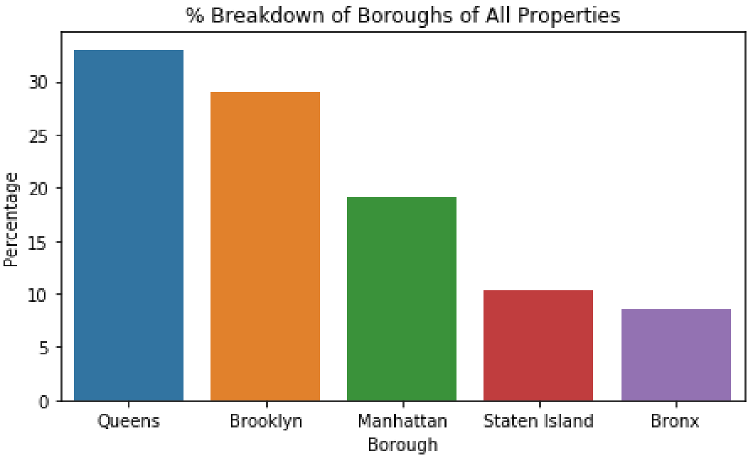


With a more narrow scope for price, we see a more gradual increase in the percentage of properties by price in the cumulative distribution graph. There is a gradual increase in the percentage of properties up to approximately $100,000, with a steeper increase up to approximately $1,500,000. Due to forward filling the null and smallest values with the average value for all other prices, the curve is steady for over 30% of the data. For the remainder of the distribution, there is the steepest incline indicating a diminishing amount of property sales with the highest values in price. Below is a boxplot of the distribution of price.

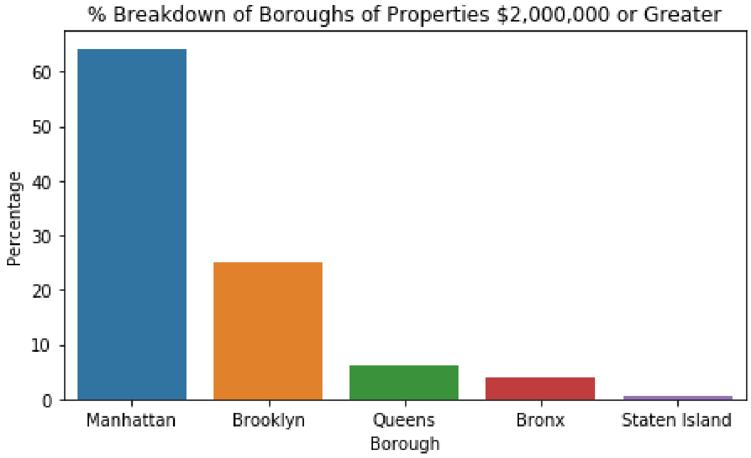


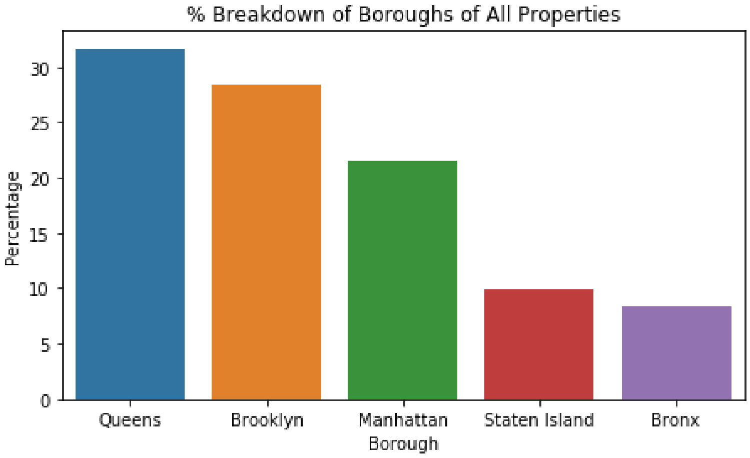
The data is most concentrated in the price range from $500,000 to $1,500,000, extending up to over $3,000,000 with the outliers removed. The upper echelon of properties in this data have a price range from 2 million to just over 3 million dollars. The next analysis will take a look at what are the most common attributes of the most expensive properties.



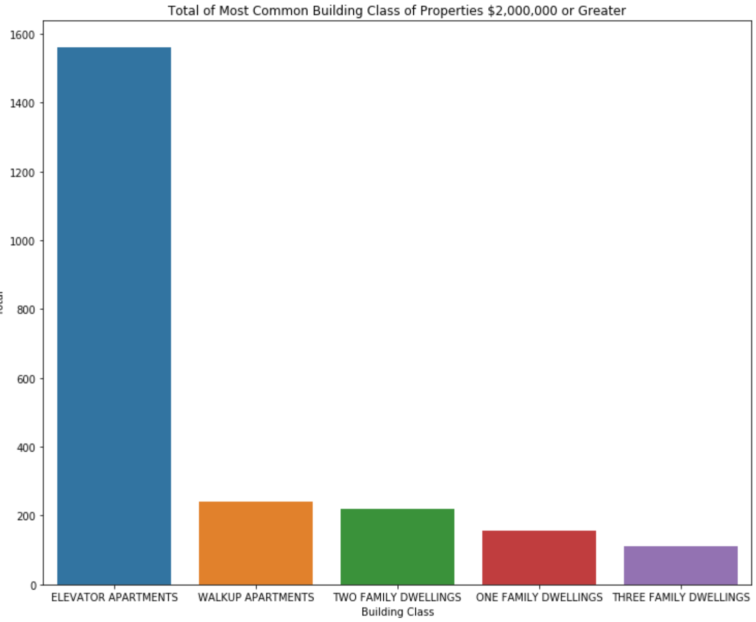


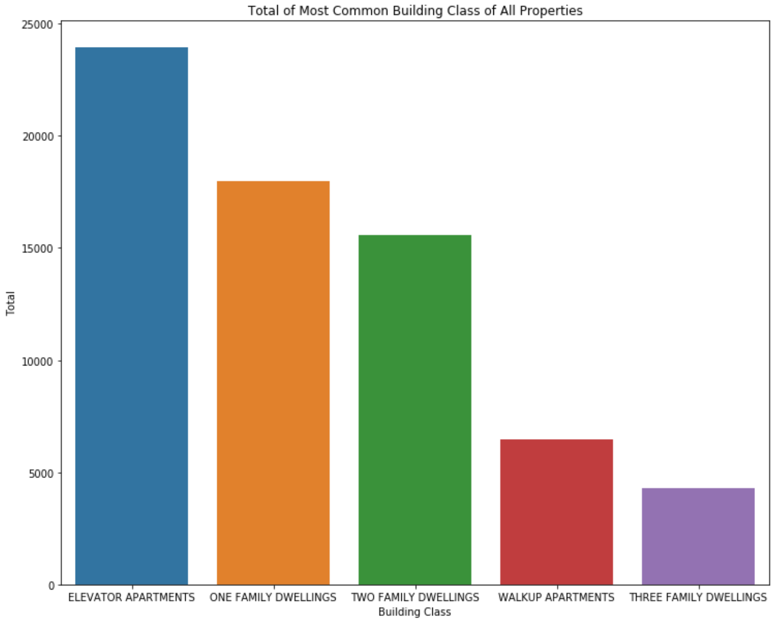
The left graph shows the percentage of properties by borough of transactions over 2 million dollars. Manhattan has over 50% of them and the next highest share is Brooklyn with more than 30%. Staten Island has the least amount despite having more total transactions than the Bronx, as represented by the graph on the right. Although Queens has the highest percentage of properties sold overall, they have less than 10% of the properties sold at over $2,000,000. It is important to note that these visuals do not include the upper bound outliers that were removed. Using those data points, these graphs are as follows.


% Breakdown of Boroughs of Properties $2,000,000 or Greater


% Breakdown of Boroughs of All Properties

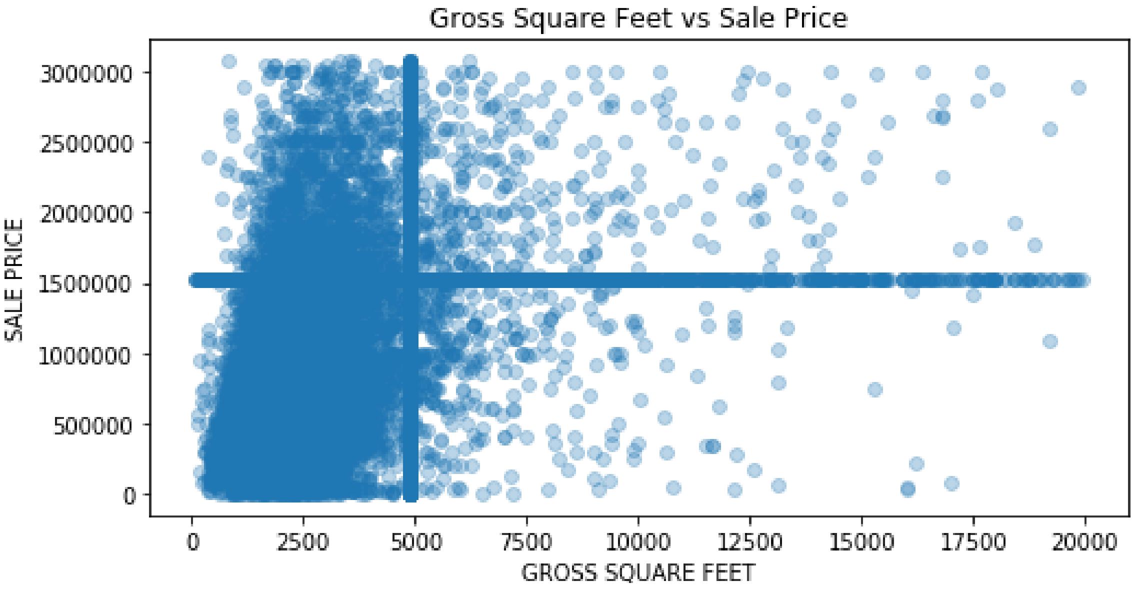
These two graphs are similar to above, with a noticeable difference in the increase in percentage for Manhattan. This is consistent with the observations from data wrangling where most of the upper bound outliers removed were properties in Manhattan. More of the extremely high prices are in Manhattan than the other four boroughs combined. Below, visuals of the most common building categories in the upper tier of price are compared with the overall frequency share of the most common building types for all of the properties regardless of price.

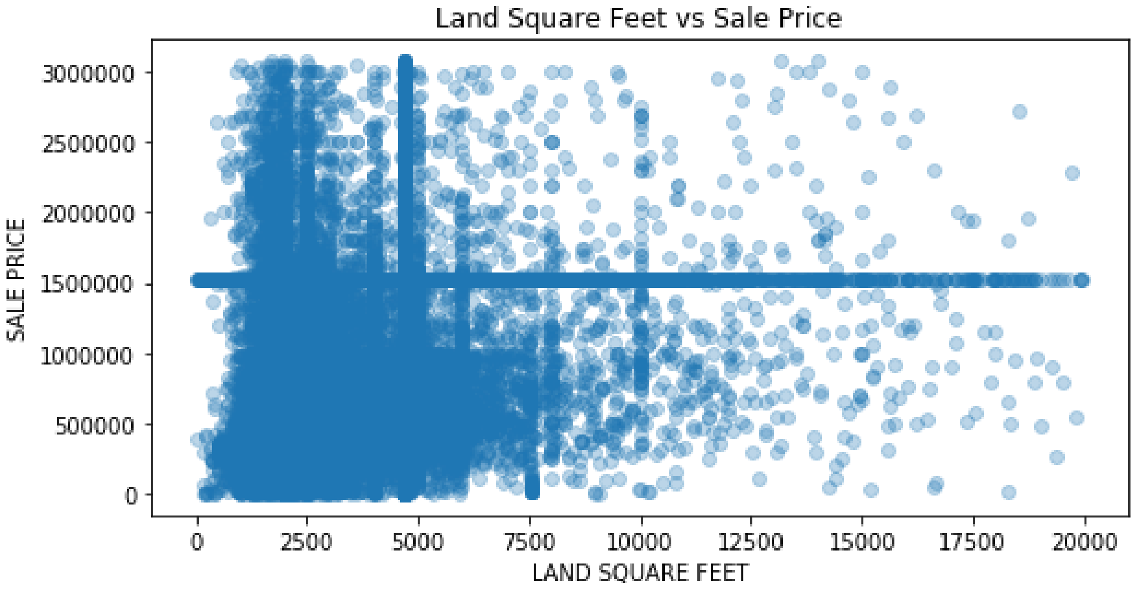

Total of Most Common Building Class of Properties $2,000,000 or Greater


Total of Most Common Building Class of All Properties

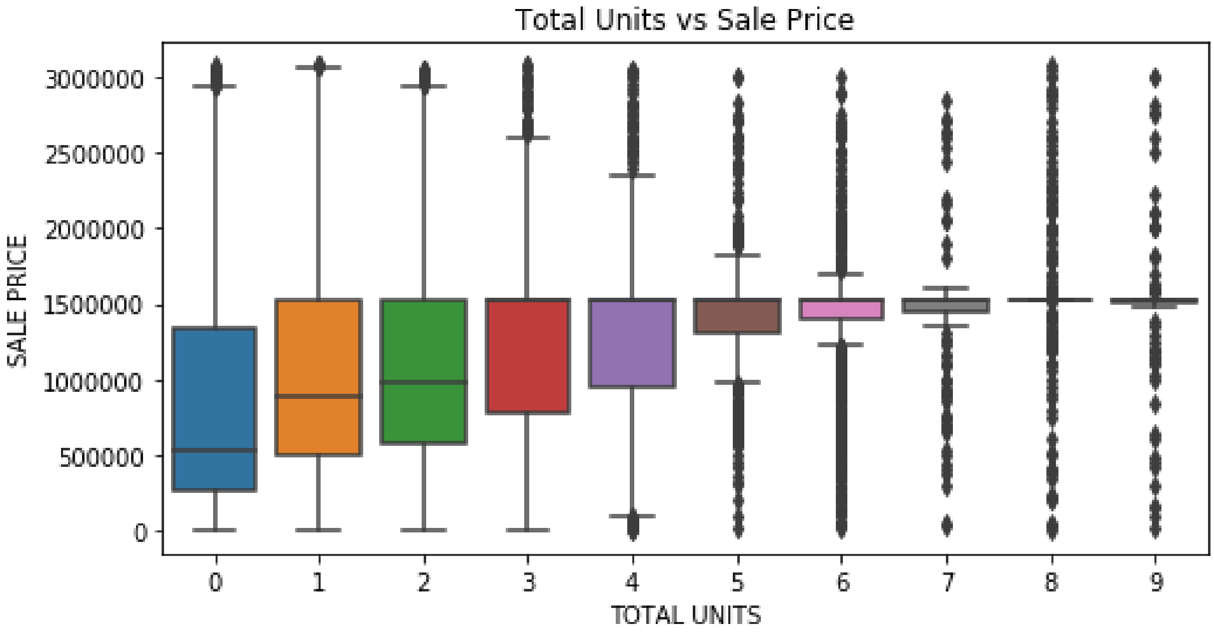
Of the most expensive properties, the majority of them are elevator apartments, which is the most frequent building class of all the properties. The most common types of properties comprise of the most frequent types of the most expensive properties as well.

Other factors that could determine a property's price are its size, represented by square footage and the amount of commercial and residential units it contains. The following are distributions of price by their gross and land square footage values, with extreme outliers in square footage removed to improve readability.

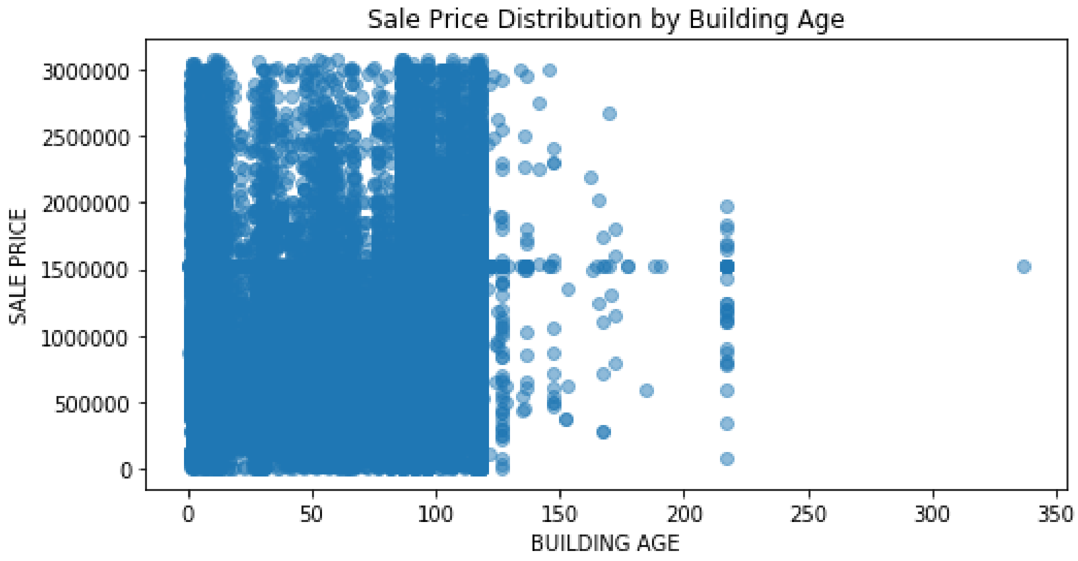




These graphs show the relationship between gross square footage and land square footage with price. It does not appear that land square feet, which represents the total land area of the property, and price have a significant relationship. There are many data points that have larger amounts of land square footage but are still on the bottom spectrum in terms of its price. This graph is much more dispersed than the one represent gross square footage. It seems that gross square footage is much more positively associated with price than land square feet. As the gross square footage increases, there become more data points above the average than below. Gross square footage here is referring to the total land area of all floors of a building. These charts can show that while gross square feet could be indicative of price there are many variables to consider when pricing a property, such as its location and other features besides its size. The following are boxplots of a property's total units by its price.

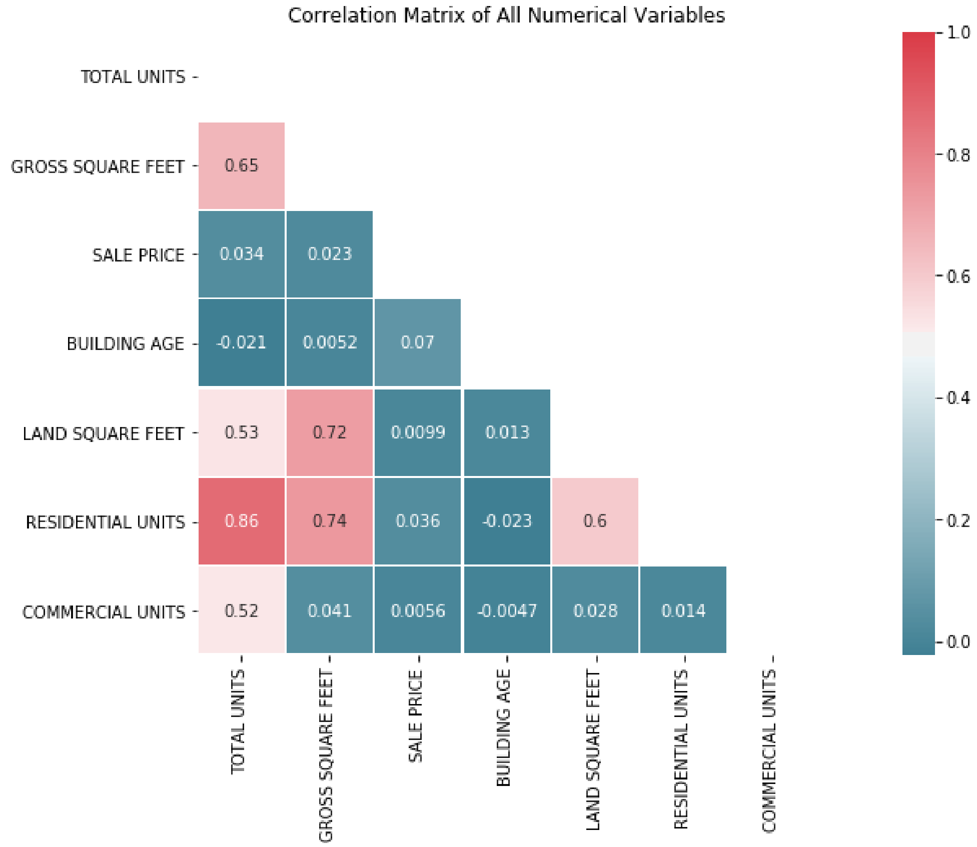


These boxplots show a general trend of the median price increasing as the amount of units increase when the value for total units is lower. As the values increase the median price levels off. This makes sense logistically because when more space is added the price generally increases, but there becomes a point where the addition of more units does not influence price as much. The following is a distribution between how many years ago the property was built and its sale price.

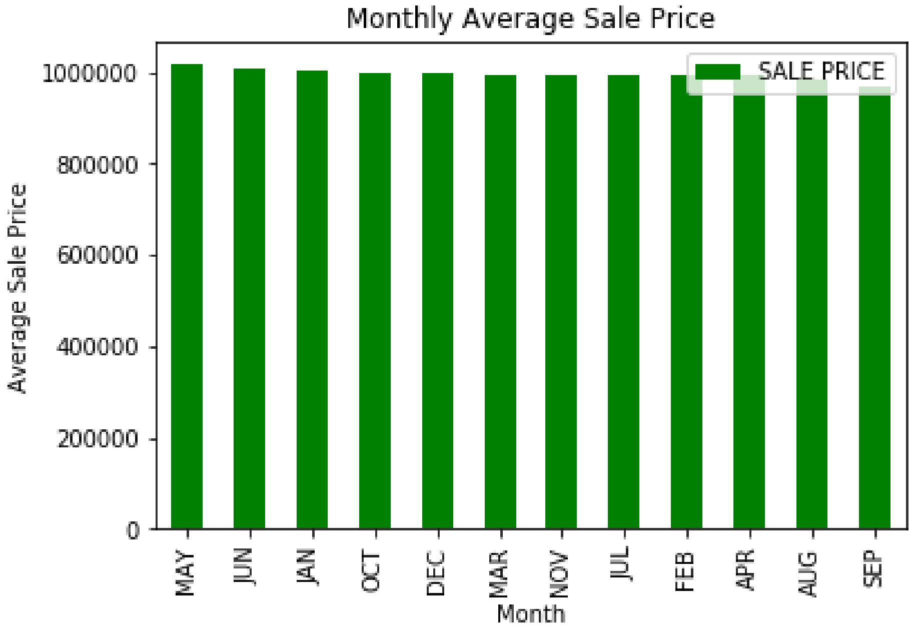


The age of the building also does not appear to tell us much in terms of its price. Most of the building ages are around 120 years old or less and are priced consistently. The oldest buildings, existing for about 125 years and more, are easier to view on the graph because there are considerably less of them. Many of these points are some of the lowest priced while others are among the highest. The age of a building is well represented across different values of price.

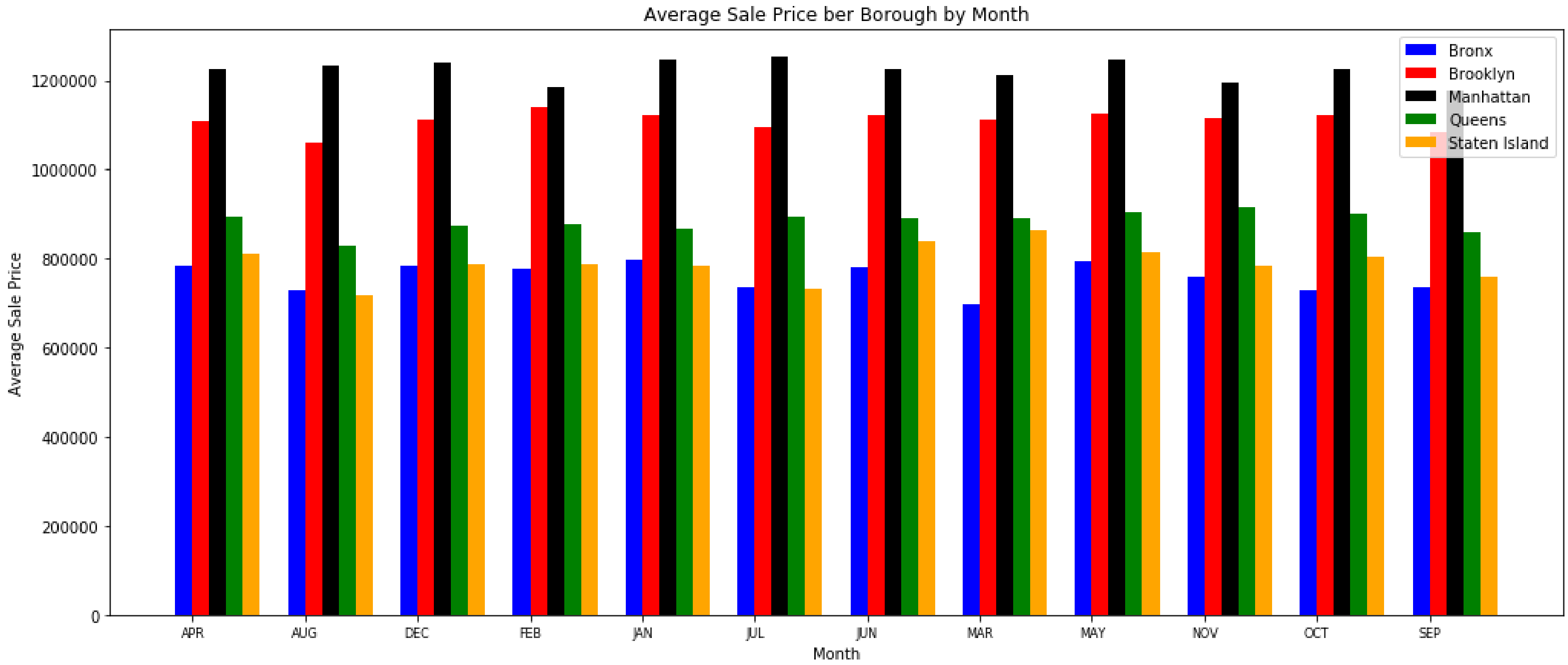
After viewing the visual relationships between price and a property's numerical aspects we can see if their correlations are consistent with these results.



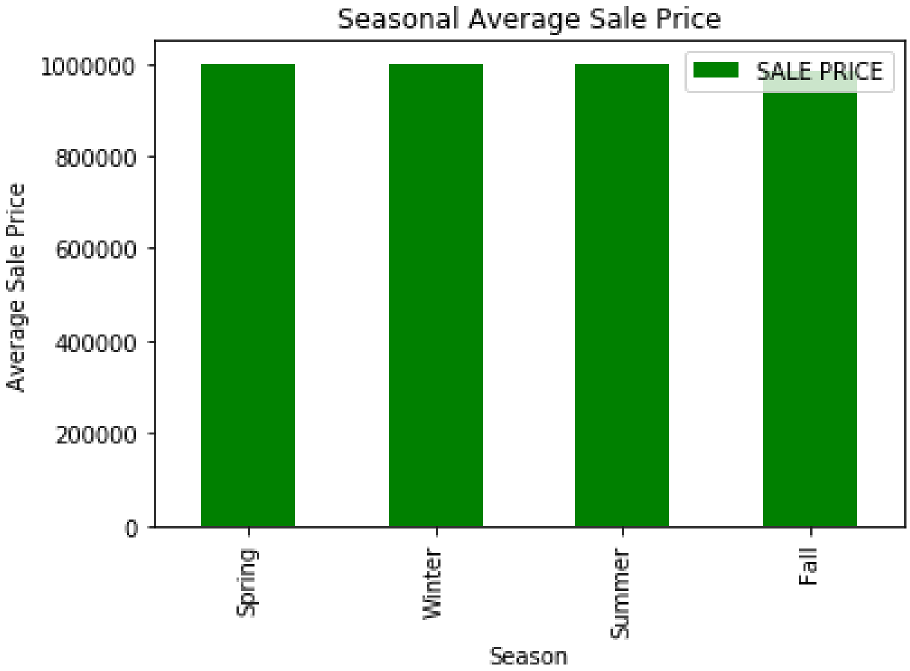
Based on the matrix above, the relationships with the highest correlations are between residential units with total units and gross square feet with residential units, which make sense intuitively. The relationships most significant for this project's purposes are associated with price. Total units and gross square footage have the highest correlations with price, which is largely consistent with the previous visualizations of these variables' distributions. Because these values are all numerical, the categorical values of the properties are not being included. In doing this, first we will see how price changes by the time of the year a property is sold.



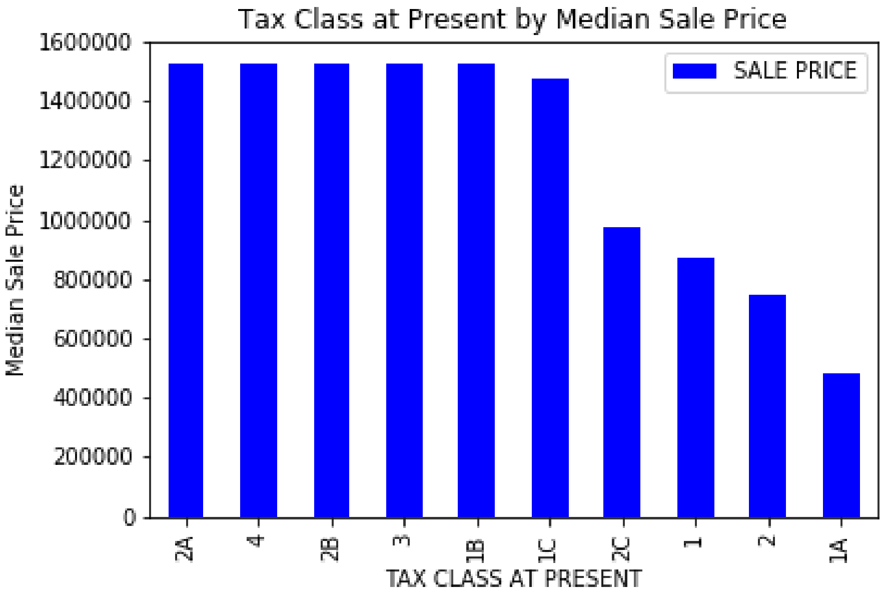
Based on this chart, the month of May has the largest average sale price with June and January slightly below. May and June having the greatest values for average price could be related to more demand for new properties just before the summer starts, compared to winter or other colder months in New York. If someone is looking for the most economical property transaction it could be wise to do so in the fall, or just after the peak of summertime is over as August and September had the lowest monthly sale prices. A more in depth look at this will evaluate average monthly price by the borough of a property.



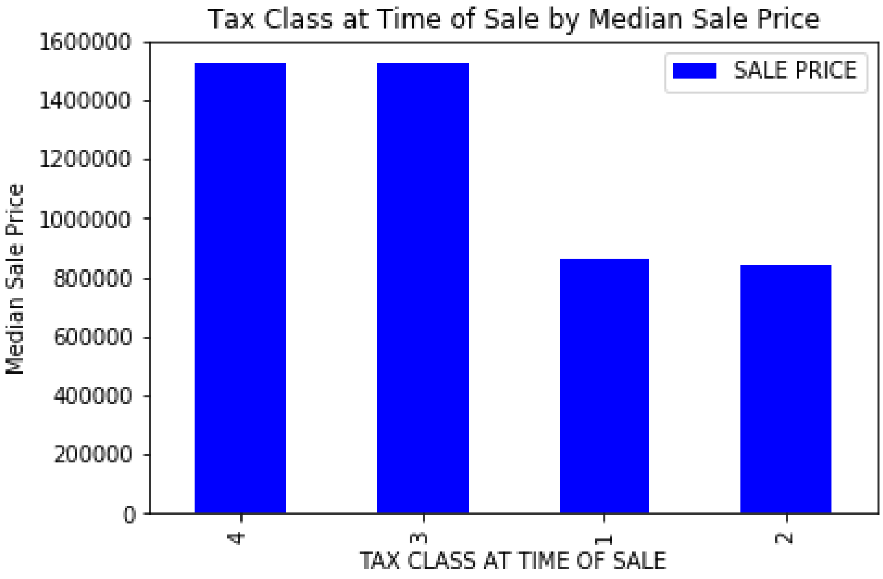
This graph represents each borough's average sale price by the month, listed alphabetically. From this we are able to notice the differences in price by borough comparing month to month. In Manhattan, these differences are most dramatic due to some higher property values, however the values appear mainly consistent across each borough. While May has the highest average overall, it is only the highest average priced month in Brooklyn, but it is among the top few in each borough. Similarly, here is a breakdown of price by the season in which a property is sold.



Based on this chart these values are consistent with the monthly data where spring has the highest average price and fall has the lowest. There is a greater than $15,000 difference between these two seasons. Varying from the date, another variable that represents what type of building a property is its tax class.

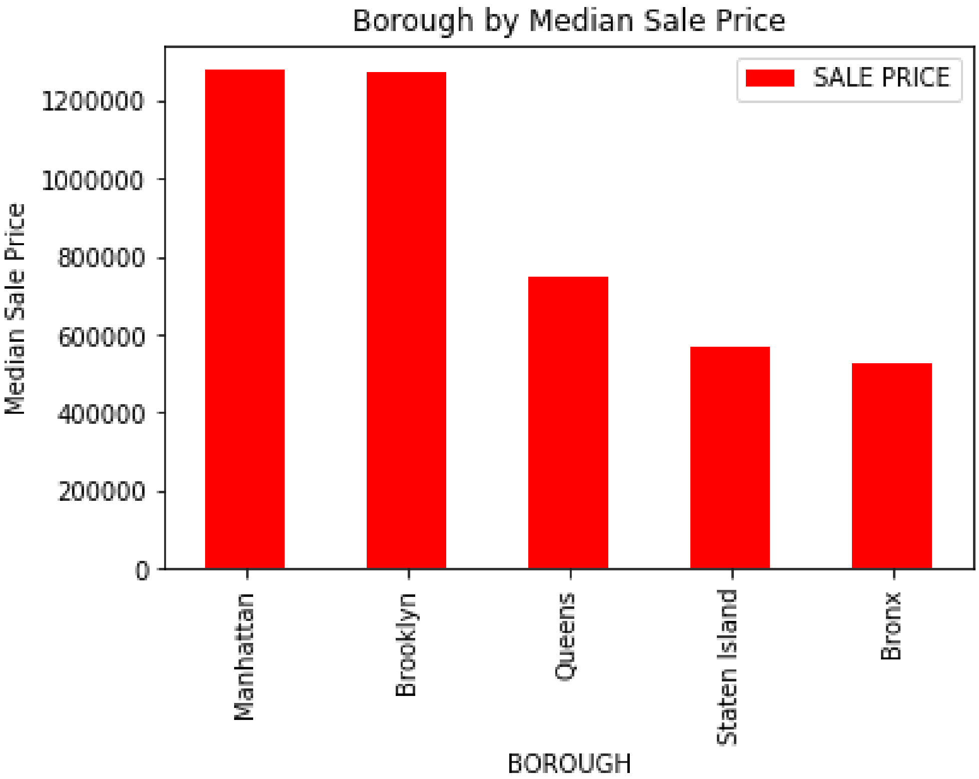


Every property in New York City is assigned to a tax class of 1, 2, 3, or 4 based on the usage of the property. Class 1 includes most residential properties of up to 3 units. Class 2 includes all other properties that are mainly residential including cooperatives and condominiums. Class 3 includes property with equipment owned by a gas, electrical, or telephone company. Class 4 includes all other types of properties such as warehouses, offices, and factories. The most expensive property sales transactions come from tax class 2A. These are comprised of the sales of mainly apartment building complexes. Class 1 has the lowest median prices of the tax groups. This next graphic represents the tax class of a building when it is sold.

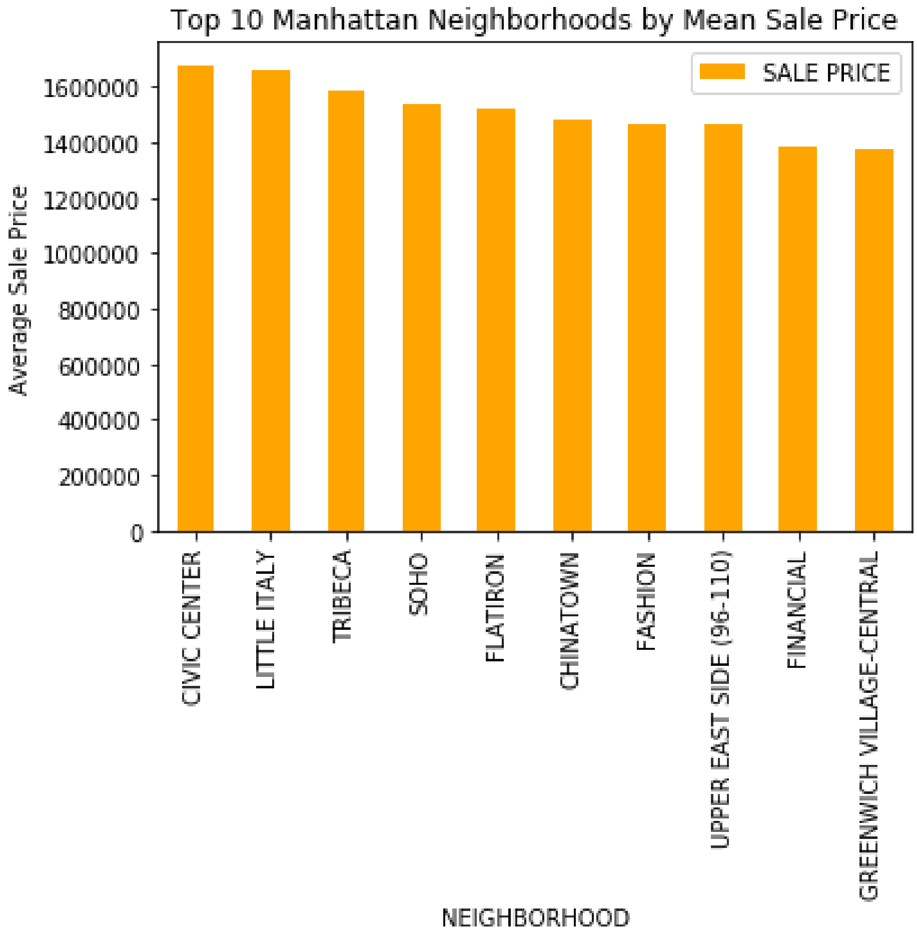


At the time of sale, the values of properties in tax class 4 are the most expensive, followed by tax class 3, 2, then 1. This is not completely consistent to what we viewed in the last graph representing price of the current tax class. Here, class 2 has the lowest median prices while class 4 has become the highest priced.

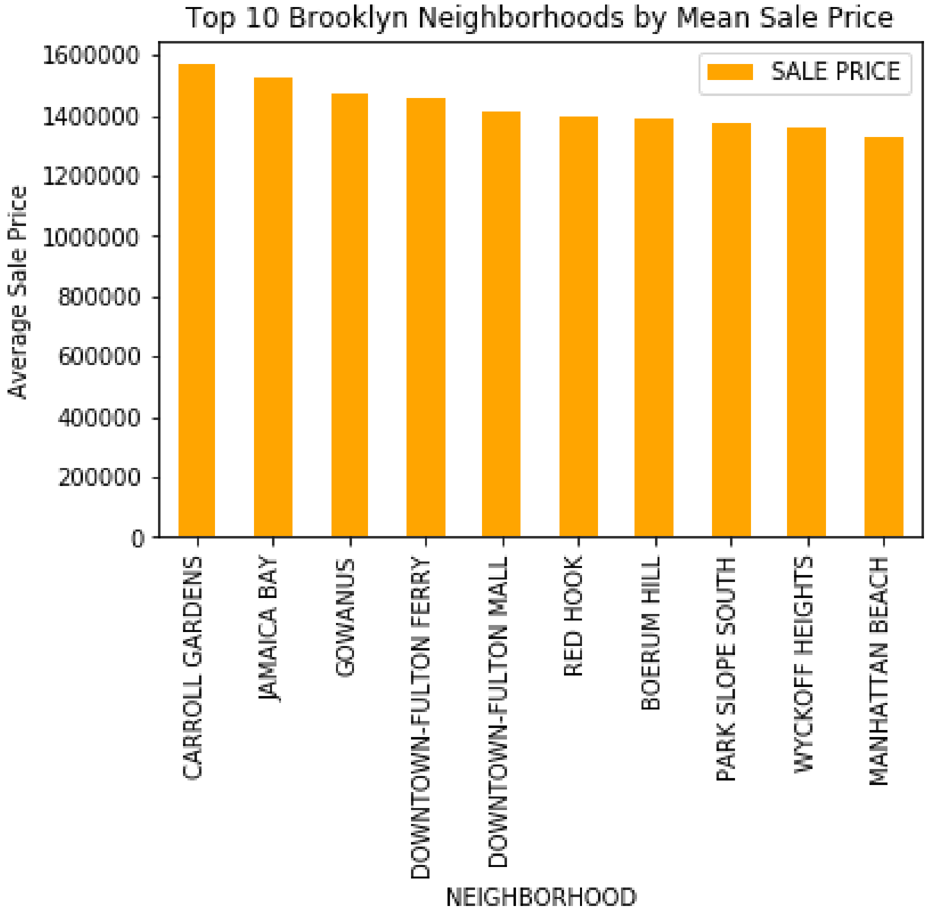
We have previously seen the five boroughs' percentage share of the most expensive properties, but will that be consistent with their overall median prices?



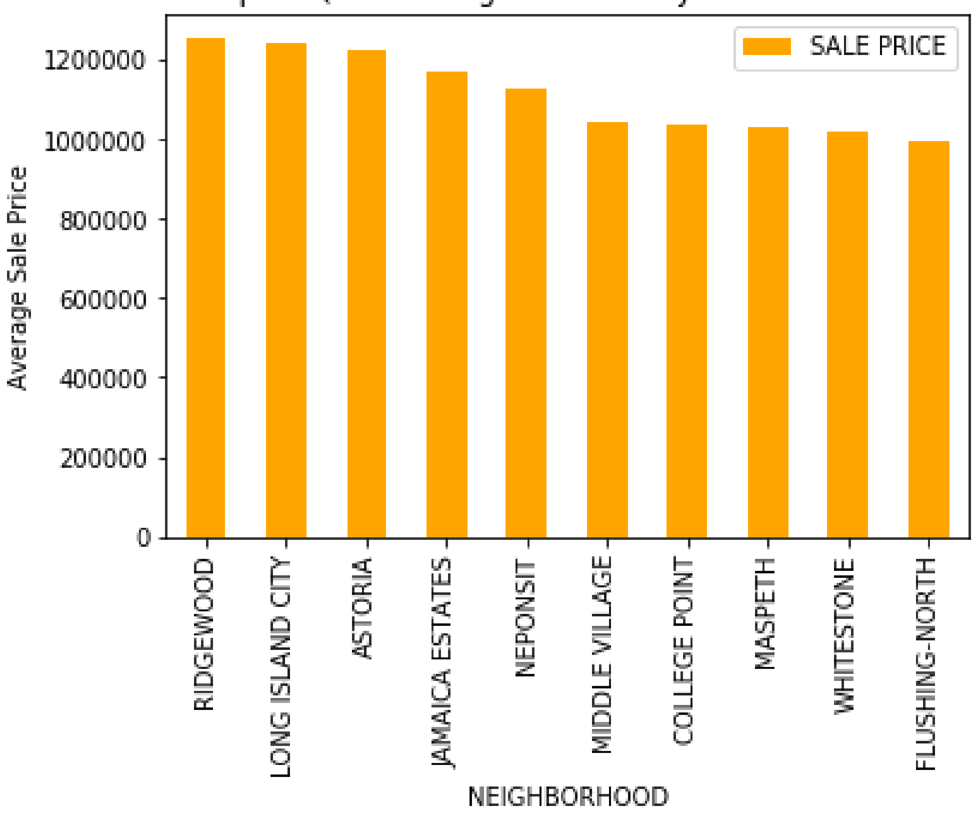
This is similar to our previous results, where Manhattan is the leader among expensive properties. However, its median price is not much greater than that of Brooklyn's. Staten Island and the Bronx have been very similar with one another in terms of prices throughout and again are the lowest priced as represented by this graph. A deeper analysis by location would include viewing which neighborhoods in each borough sold for the highest average price.



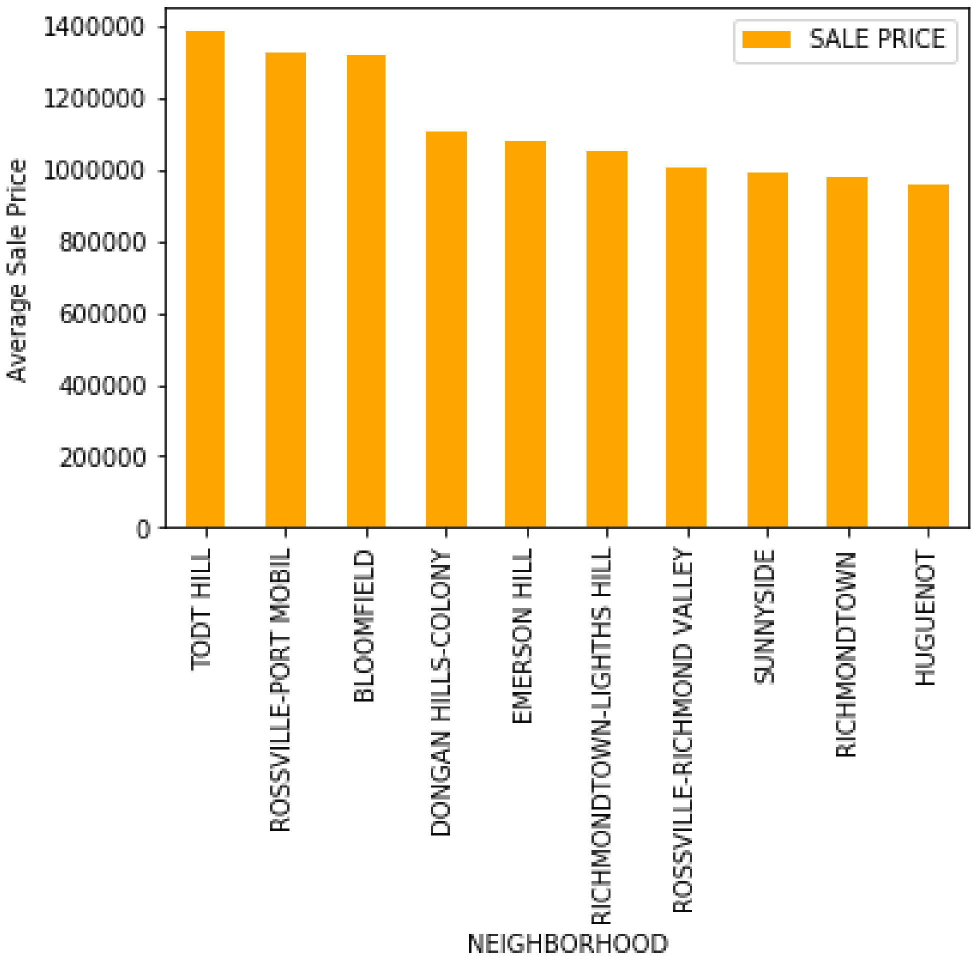
Top 10 Manhattan Neighborhoods by Mean Sale Price



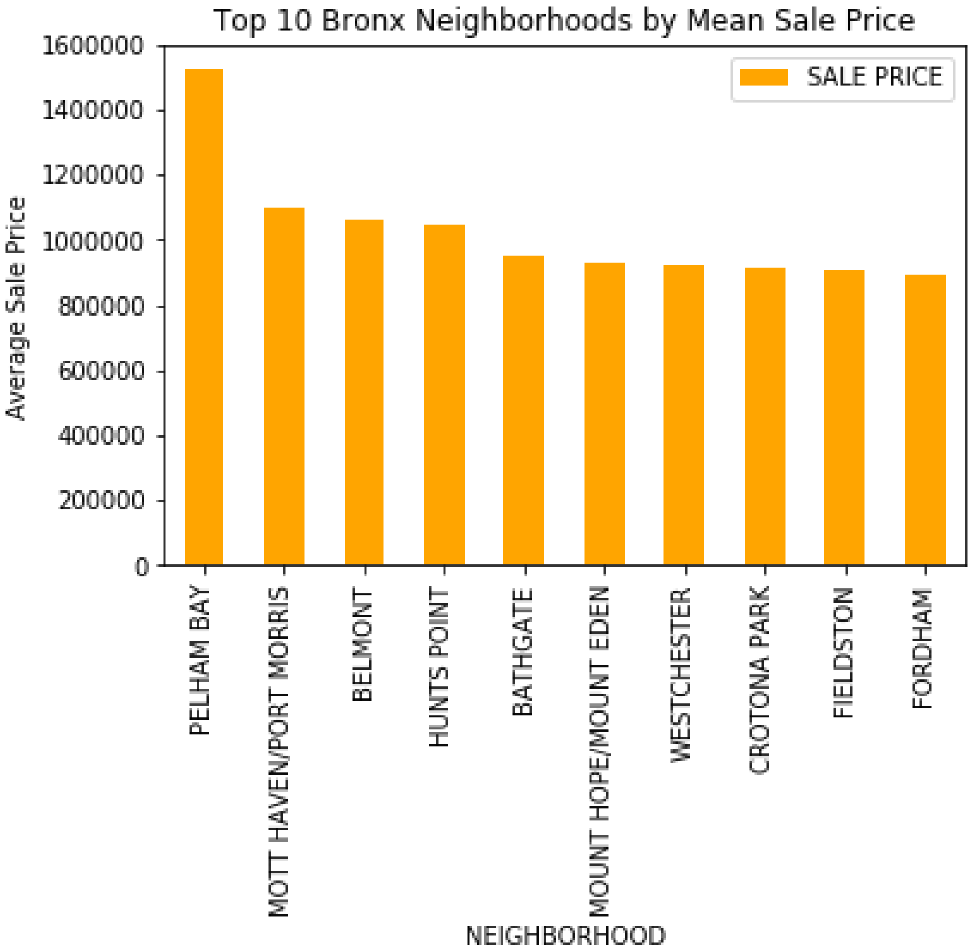
Top 10 Brooklyn Neighborhoods by Mean Sale Price



Top 10 Queens Neighborhoods by Mean Sale Price



Top 10 Staten Island Neighborhoods by Mean Sale Price



These graphs are mainly consistent from what we have learned about the average prices of properties by borough. Despite the Bronx having the lowest median price per borough, their top priced neighborhood, Pelham Bay, has a higher average sale price than of any neighborhood in Queens or Staten Island. This area shows the most disparity between any borough's adjacently ranked top 10 properties with a greater than $420,000 difference between the next highest priced neighborhood. Civic Center in Manhattan is the highest overall priced neighborhood on average while Manhattan holds the top 3 overall priced neighborhoods and 6 of the top 10, while Brooklyn has 3 of the top ten highest averaged priced neighborhoods. This is also consistent with what we have seen where Brooklyn is the next priciest borough behind Manhattan.

The next section of the project evaluates various machine learning models, with the objective of creating a predictive model that can accurately predict the price of a property based on its features and reduce the amount of error in these predictions. Different types of machine learning algorithms will be applied to see which performs best based on the information available to it.

The variables being included into these models to predict price will consist of a property's borough, building class category, the age of the building, the amount of commercial units, residential units, gross square feet, land square feet, month of the sale and that season. For the categorical data columns of borough, building category, month and season, dummy variables were created so the models can indicate which of these corresponding values of these features the property is associated with. From our dataset of 79,709 properties, 80% of the data was taken to use as training data for each model, while the remaining 20% will be the data that is tested on. That gives 63,737 values to be used for training and 15,942 values that are predicted on in any iteration of a model.

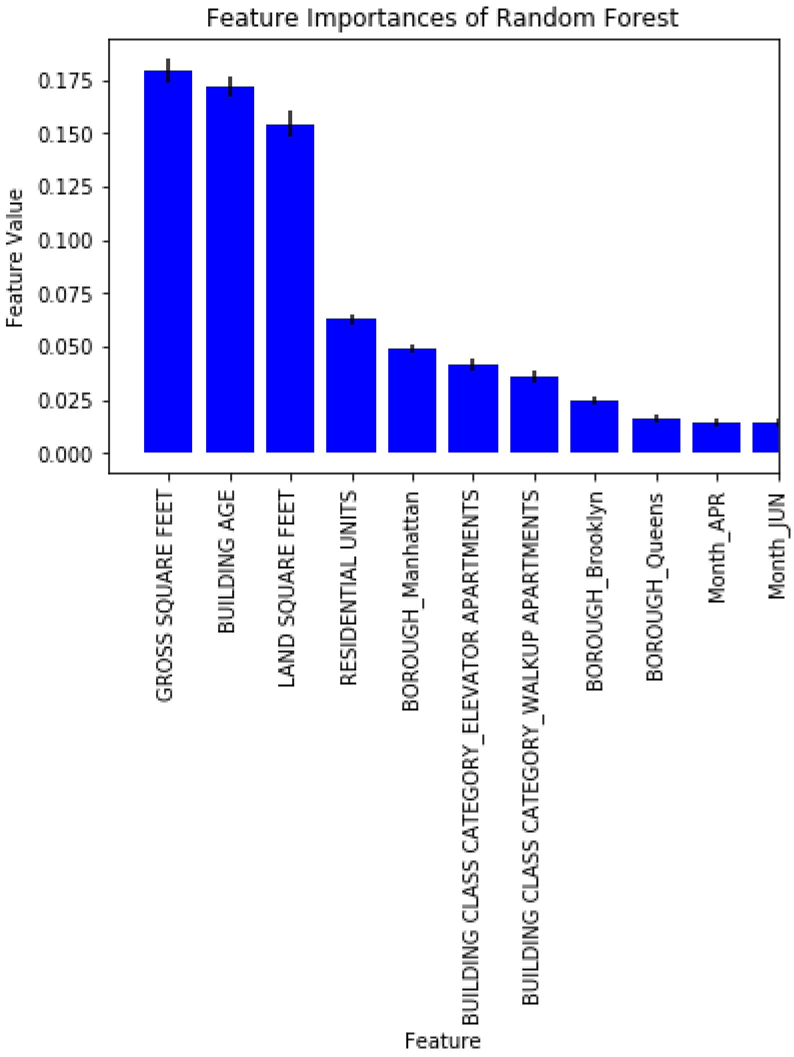
Below is the report of the values measuring the performance of the models.

Linear Regression:
R^2: 0.185
Root Mean Squared Error: 0.899
Average 5-Fold CV Score:
-3.372621123002836e+17

Random Forest Regression:
R^2: 0.181
Root Mean Squared Error: 0.901
Average 5-Fold CV Score: 0.16945

Ridge Regression:
R^2: 0.185
Root Mean Squared Error: 0.899
Average 5-Fold CV Score: 0.19046

Elastic Net Regression:
Tuned Elastic Net l1 ratio: {'l1_ratio': 0.0}
Tuned Elastic Net R squared: 0.0655
Tuned Elastic Net MSE: 0.9265



The models used were based off of the dataframe that eliminated the upper outliers for price. None of these models performed particularly well at reducing the amount of error in the predictions, but the Ridge Regression performed the best with the same values as the Linear Regression for R-Squared and Root Mean Squared Error of .185 and .899 respectively. The Ridge Regression featured an average five fold cross validation score of .19 while the Linear Regression value was an extremely high negative value. The Random Forest model also performed similarly to these values. Based on the above graph of the Random Forest feature importance values, the most important features in the model were a property's value for gross square feet, building age, land square feet, then amount of residential units.

Conclusions:

There are many different types of property transactions across New York City, and this is easy to see based off of this limited data alone. Without even viewing a property we are able to gain insight on the intricacies of NYC real estate and understand the distinctions and determinants of what goes into the sale price of a property. While the information we had available was valuable, it likely does not tell the complete story. In actuality, there are many additional factors that can determine a property's price, based on what type of property it is. Some other variables may include the proximity of the property to subways or major NYC landmarks, its amenities and what is included with the property, how nice it is, the crime rate in the area, the property's accessibility to available parking and more. While the models created did not give way to an accurate predictive source for price modelling, they demonstrated a course of action to take in the case of more available and uniform data that could yield better results.