NYC Sales Capstone Project
Jeffrey Ebert

New York City Property Sales

With over 8.6 million people currently living in New York City, the area is continually a national leader for real estate. 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.

The data used for this project is comprised of property transactions in New York City from September 1st 2016 to August 31st 2017. The dataset was uploaded from a public source on 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.

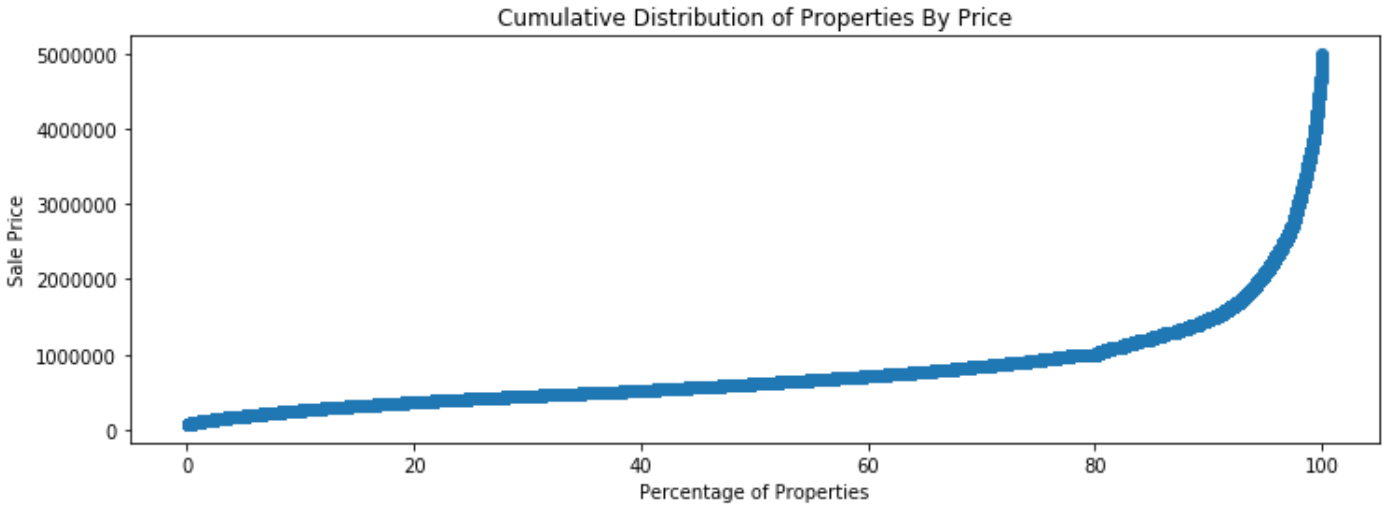
There were inaccuracies in the data found by exploring the dataset, labelled df, that required cleaning. The dataset contained columns that provided no or useless information. These data columns, labelled EASEMENT, UNNAMED, and APARTMENT NUMBER were all removed. The data contained 765 duplicate values 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.

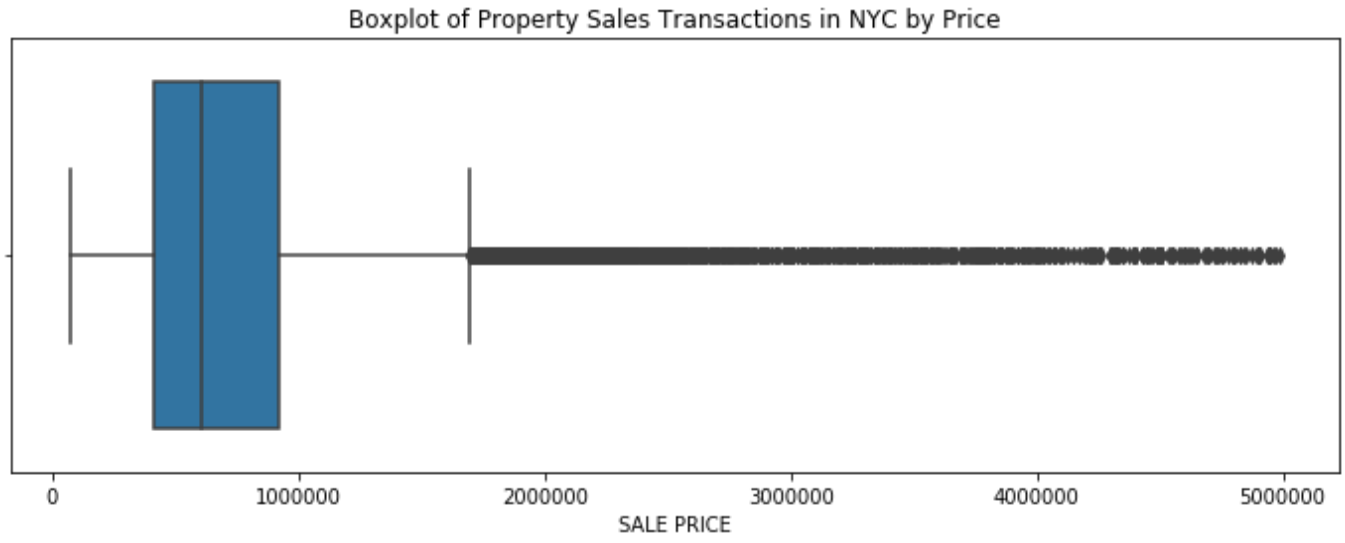
There were data entries that included missing values and values of zero for a property's square footage, building year, and price. These missing values needed to be treated differently based on the variable. There were null or zero values for the square footage data columns for almost half of the dataset. Because there were so many values, it may not be best not to remove them. Instead, they were filled with the mean value for their type of square footage. I also created another iteration of the dataset df (df_o) where the data entries with null values for square footage were deleted but the zero values were kept. There were also a small amount of high outliers for square footage that were removed when analyzing price based on square footage to not 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 observing the relationship between price and building age, but kept in the dataset for future analysis. There were also data entries found where the amount of commercial and residential units did not equal the amount of total units, which could not have been possible. When observing the relationship between price and the amount of units these values were removed but kept for other analysis.

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, for the dataframe df, 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 4,074 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. An iteration of dataframe df was created (df_m) where these outliers were removed from the dataset to provide a more concise grouping of the data. This dataset contains 79,709 rows. On the dataframe df_o, without any null or zero values filled by a column's average values, close to 11% of the data contained zero values for price and 15% of the total data included prices below $75,000. In this dataframe, there were similarly a few very high values for price that altered the data. These values were removed here in order to keep a more consistent grouping of prices. The new minimum and maximum for price in this dataframe ranges from $75,000 to $5,000,000 and it contains 35,007 rows. Another dataframe was created (df_n) from the original data where all null and zero values were removed. As a result this is the shortest dataframe with 29,162 rows. All of these different variations of the data will be tested in the machine learning models in order to see which iterations of the data and its features can predict price the best.

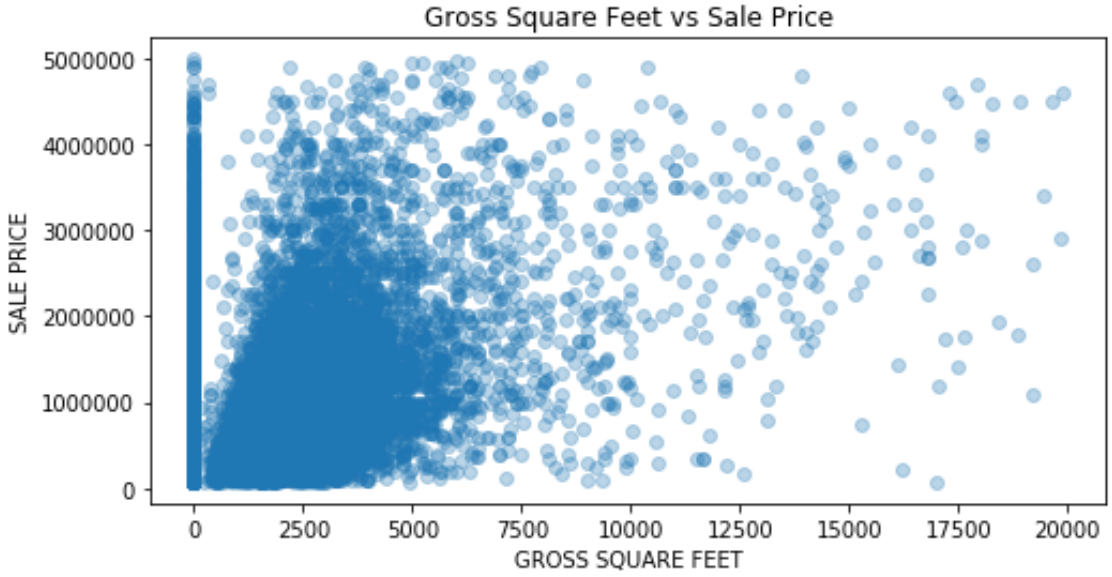
The first objective of the project is to understand how the data is broken down and to evaluate the various types of transactions made with the distinctions between them that have an influence on the property's sale price. In doing so, The numerical features of a property will be observed first. All of the following data being used comes from the dataframe df_o described above, where null values and outliers of price have been removed. Below is a cumulative distribution graph of price.

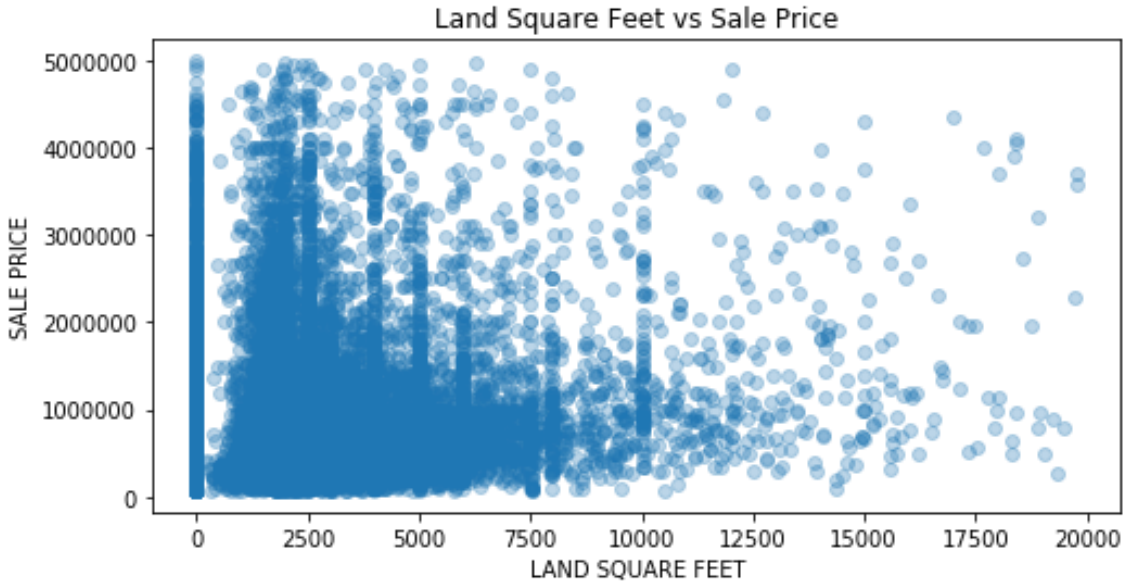


This cumulative distribution graph reflects the percentage share of the properties of the data with a price range from $75,000 to $5,000,000. There is a perpetual increase from 0% up to approximately 80% to a price of around $1,000,000. At this point, the curve steeply inclines, representing a diminishing amount of properties with continually higher prices. Viewing this data in the form of a boxplot looks like this:



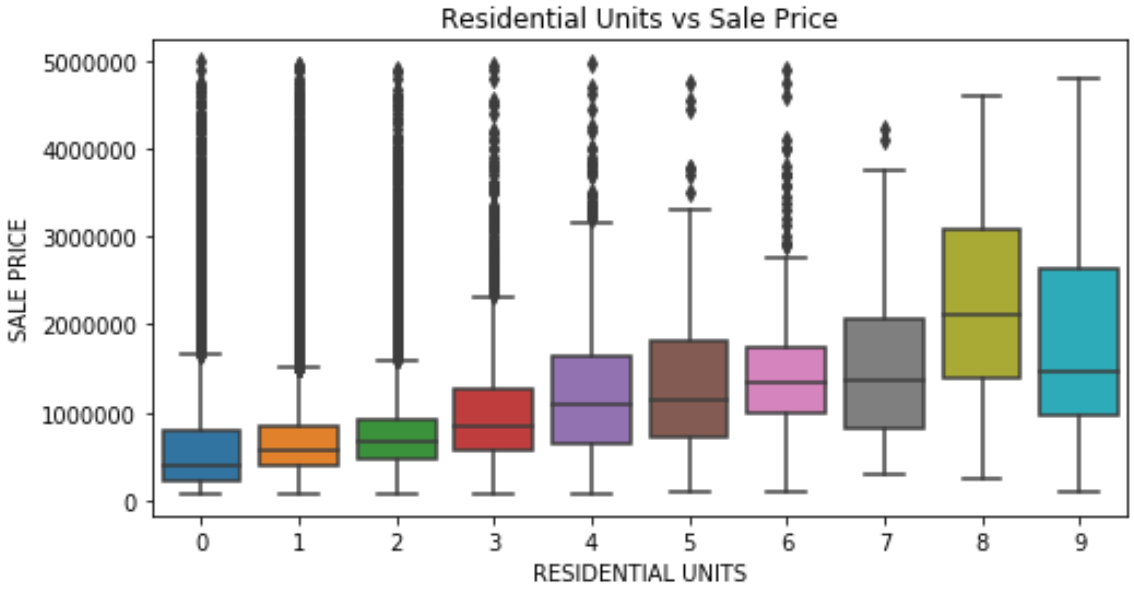
The data is most concentrated in the price range from $500,000 to $1,000,000, extending up to over $5,000,000. The upper echelon of properties in this data have a price range of 2 to 5 million dollars. The next analysis will take a look at the relationship for values of a property's square footage and price.

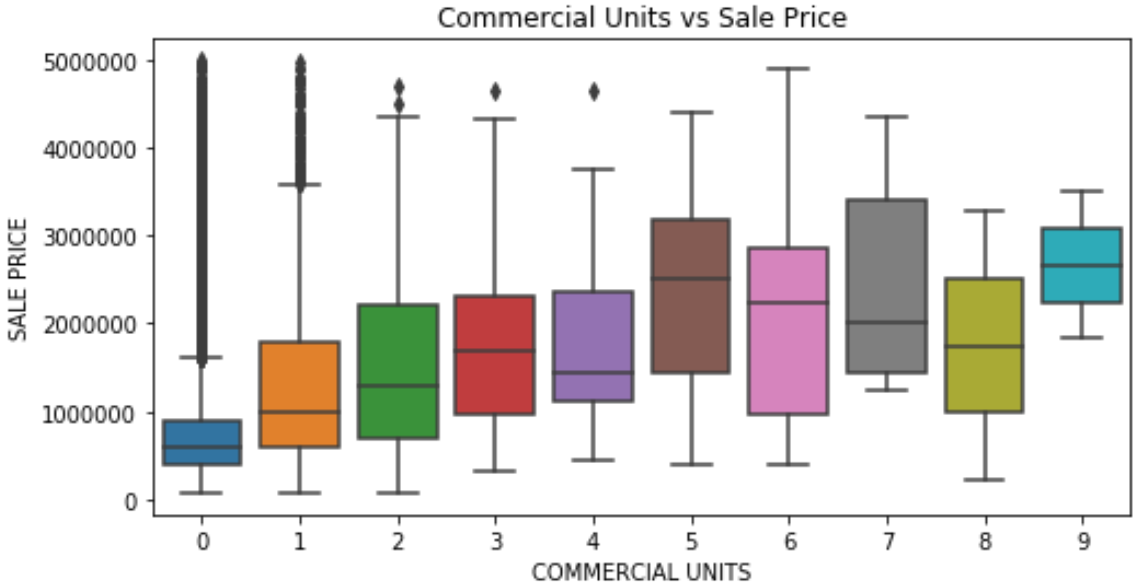




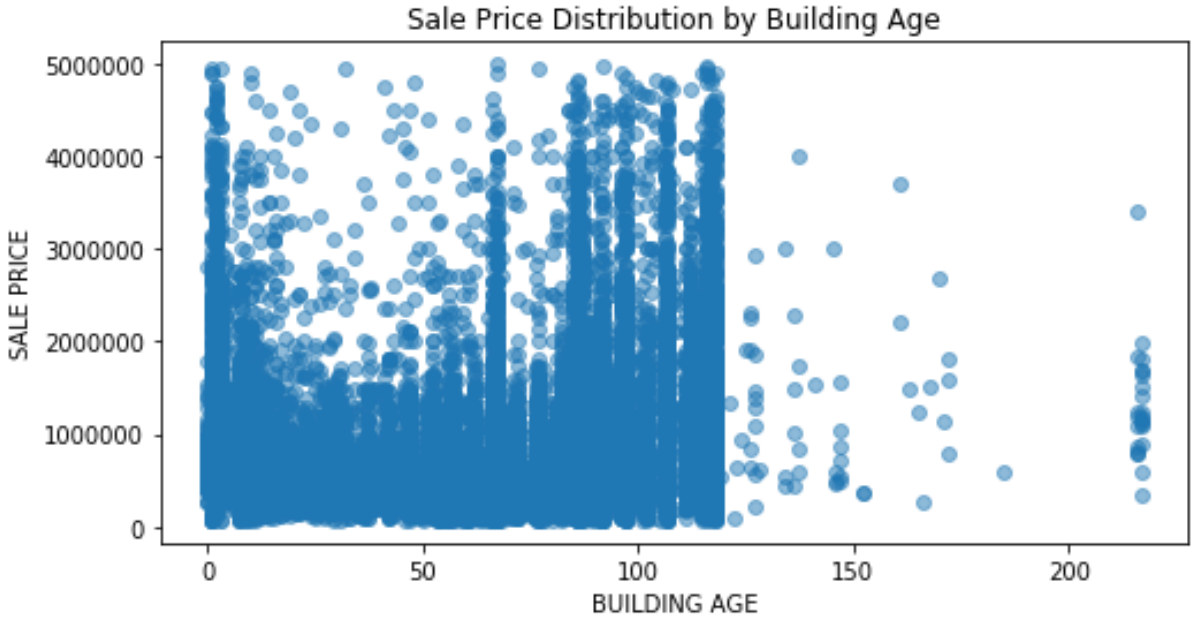
These graphs show the relationship between gross square footage and land square footage with price for properties in New York City. There does not appear to be a significant relationship for land square feet, which represents the total land area of the property. There are many data points that have larger amounts of land square footage

but are still on the bottom of the spectrum in terms of price. This graph is much more dispersed. Gross square footage is much more positively associated with price than land square feet. As the gross square footage values increase, there become more data points toward the top of the graph. Gross square feet here is referring to the total land area of all floors of a building. This can tell us that while gross square feet could be indicative of price there are many variables to consider when pricing a property other than its square feet. Another variable to measure the size of a property are the amount of commercial and residential units it contains.



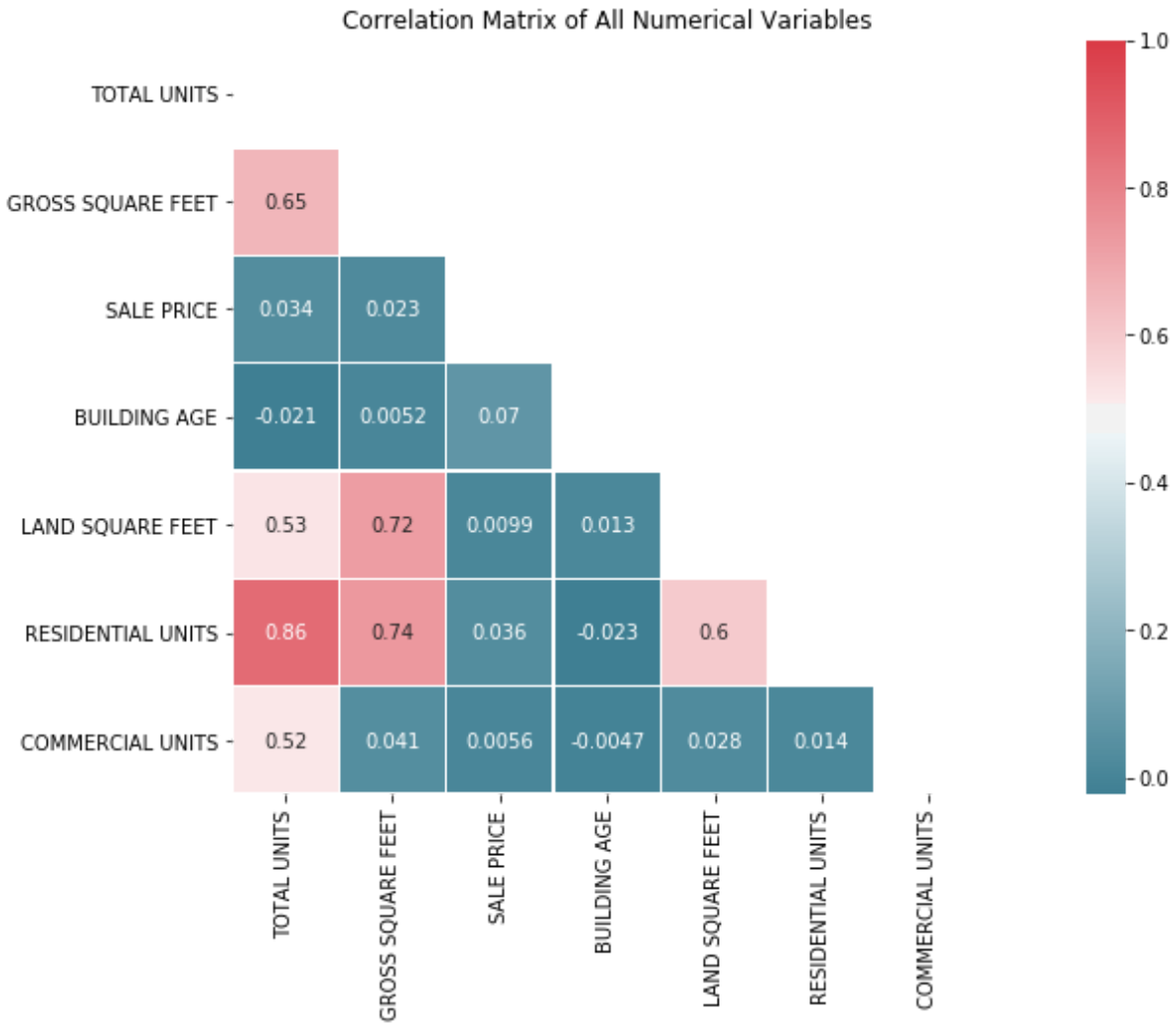


These boxplots reflect a trend of the median price and range of prices mainly increasing as the amount of residential and commercial units increase for a property. This makes sense logistically because when more space is added the price increases. The only mark where the median price of residential units decreases is at 9, however the 75th percentile price is the highest of all. For commercial units, there are slight dips at 4, 7, and 8 units but the median price for 9 commercial units is the highest for all properties with less than 10 units. The following is a visualization regarding how many years ago the property was built and its sale price.

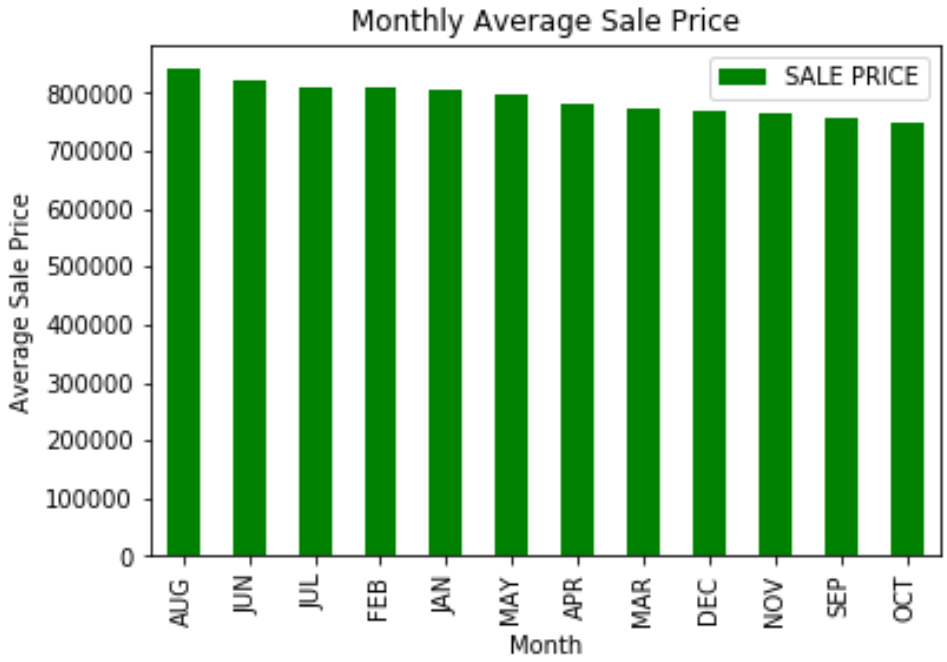


The age of the building also does not appear to tell us much in terms of its price. Most of the building ages are around 120 years old or less and are priced consistently. The oldest buildings, aged around 125 years or more, are easier to view on the graph because there are considerably less of them. Some of these are some of the lowest priced while others are among the highest, but they are represented in all areas of the price scale.

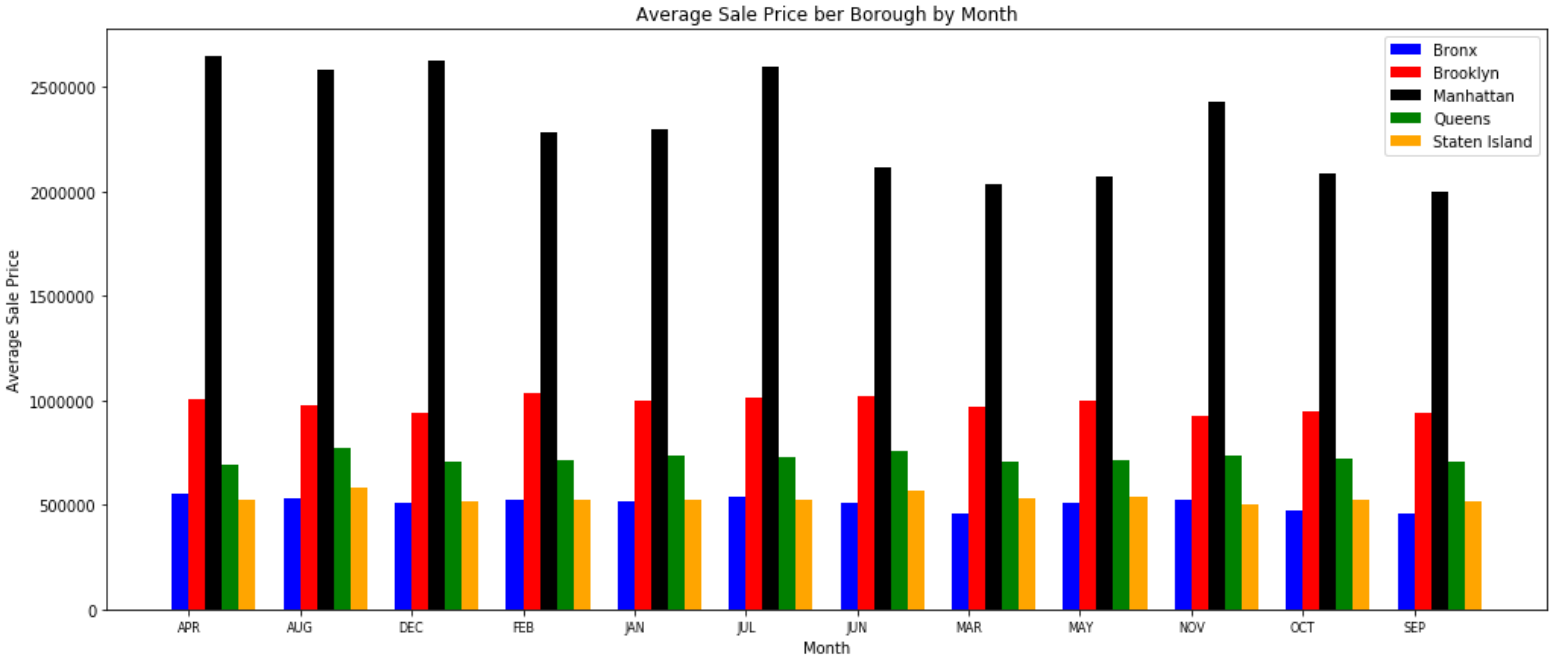
After viewing the visual relationships between price and a property's numerical aspects we can see if their correlations are consistent with those results.



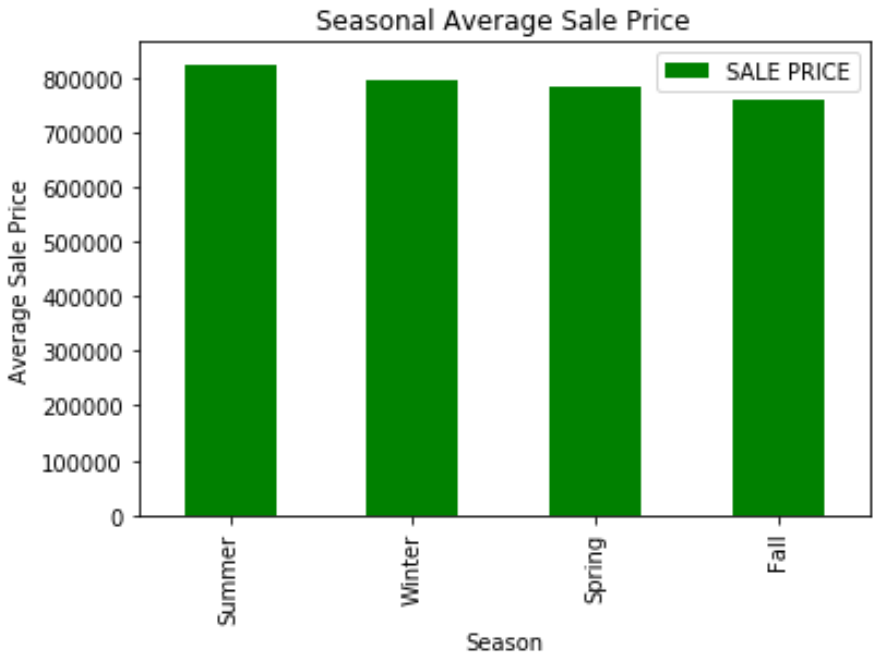
Based on the matrix above, the relationships with the highest correlations are between residential units with total units and gross square feet with residential units, which make sense intuitively. The relationships most significant for this project's purposes are associated with price. Total units and 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.



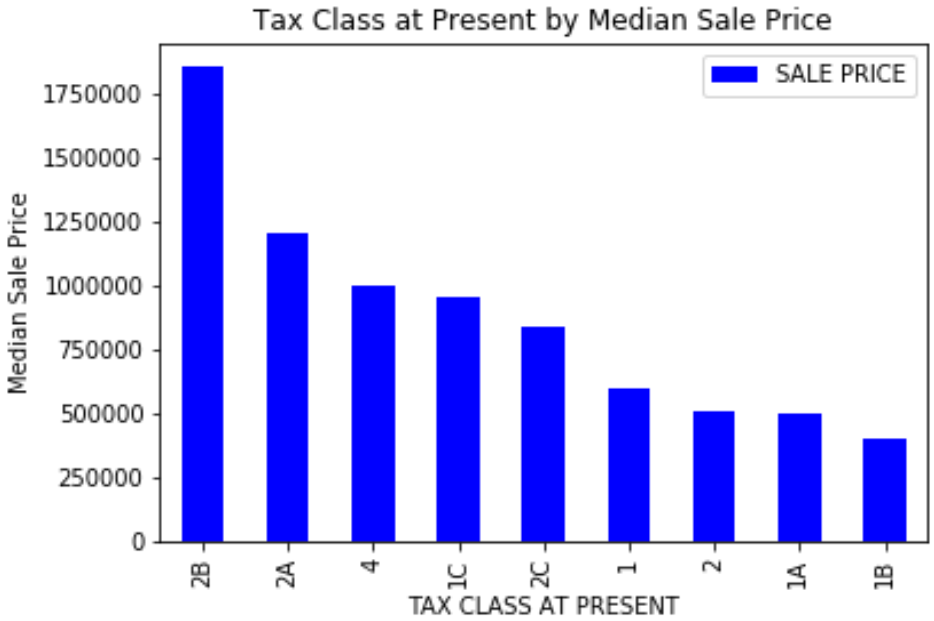
Based on the graph above, the month of August has the largest average sale price with June and July following. These months having the greatest values for average price could be related to the increased demand for new properties during the summer compared to winter or other colder months in New York. If one is looking for the most economical property transaction it could be smart to complete it in the fall, which had the lowest average monthly prices of November, September and October. A more in depth look at this will evaluate average monthly price by the property's borough.



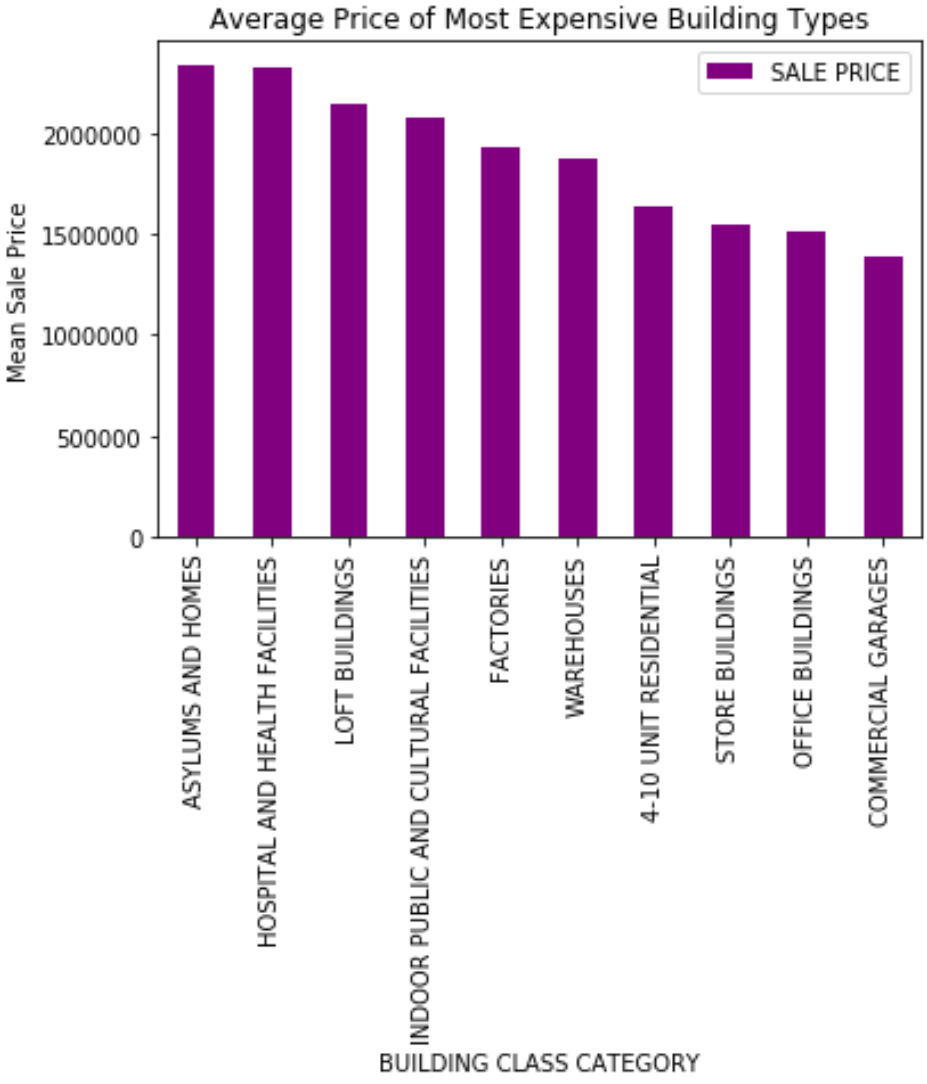
This graph represents each borough's average sale price by month, listed alphabetically. From this we are able to notice the differences in price by borough comparing month to month. In Manhattan, these differences are most dramatic due to some higher property values, however most months by borough appear mainly consistent. August, the highest priced month overall, has the highest average by month for Queens and Staten Island. Similarly, here is a breakdown of price by the season in which a property is sold.

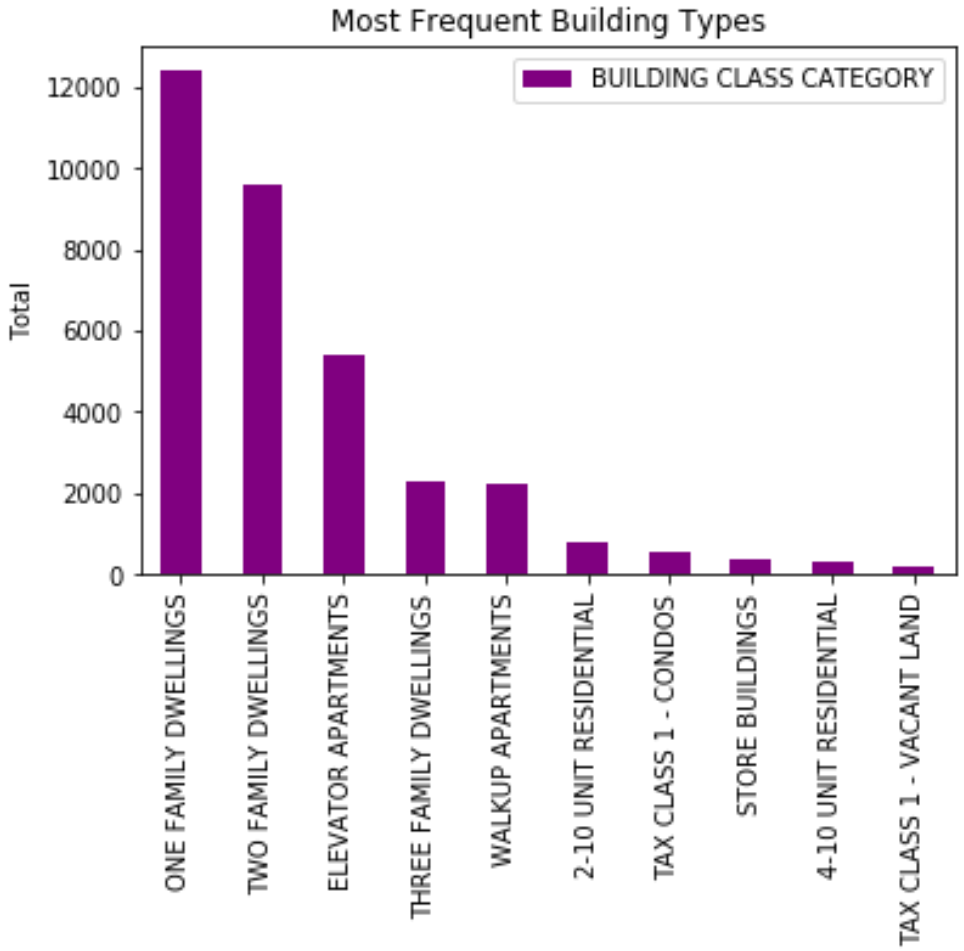


Based on the chart, these values are consistent with the monthly data where summer has the highest average price and fall has the lowest. There is a difference of $66,991 between these two seasons. Varying from the date, another categorical variable from the dataset is a property's tax class.

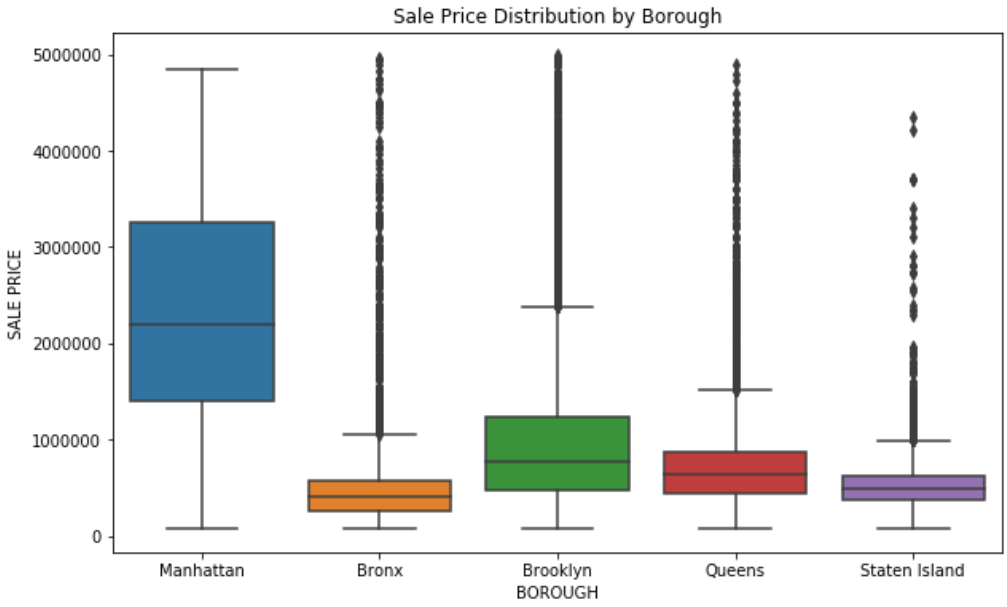


Every property in New York City is assigned to a tax class of 1, 2, 3, or 4 based on the use of the property. Class 1 includes most residential properties of up to 3 units. Class 2 includes all other properties that are mainly residential including cooperatives and condominiums. Class 3 includes property 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 2. These are comprised of the sales of mainly apartment building complexes. Class 1 has the lowest median prices of the tax groups. Next we can look at which building classes are most expensive compared to the total counts of the most frequent building classes.

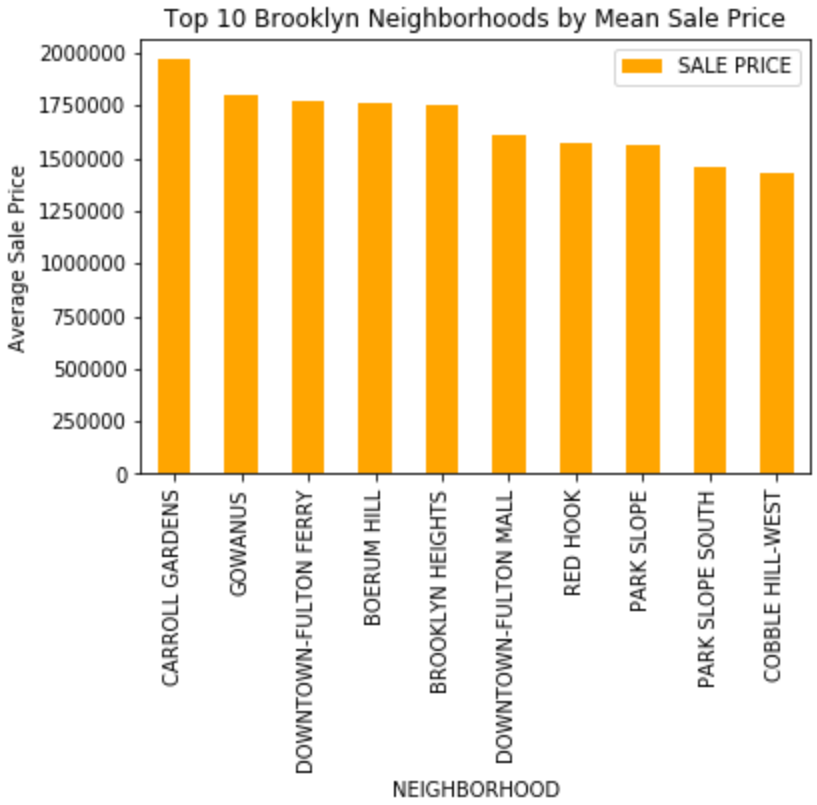


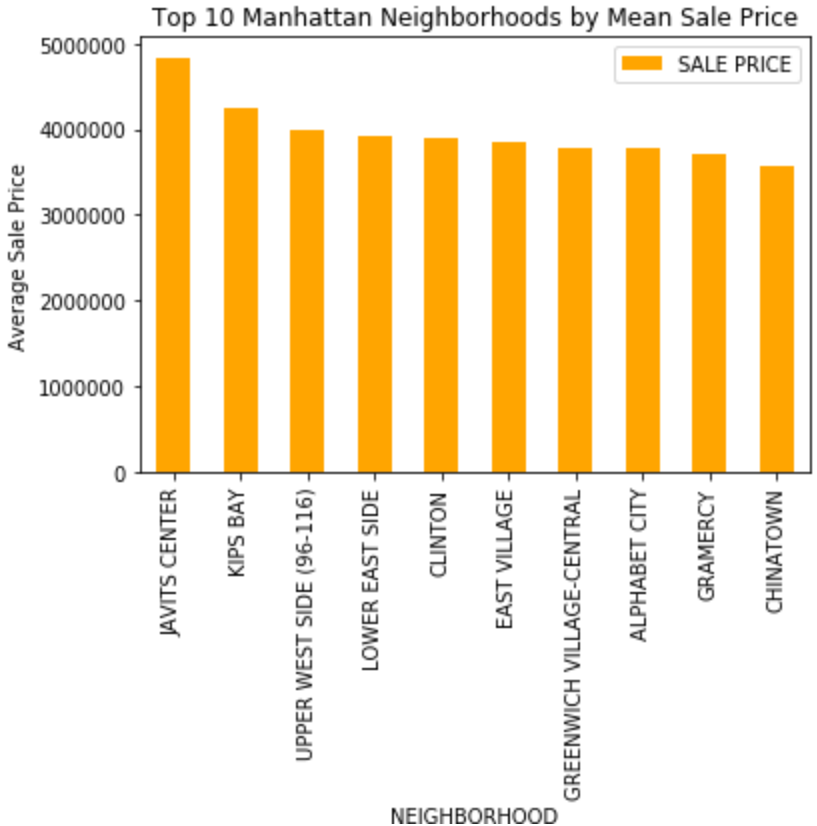


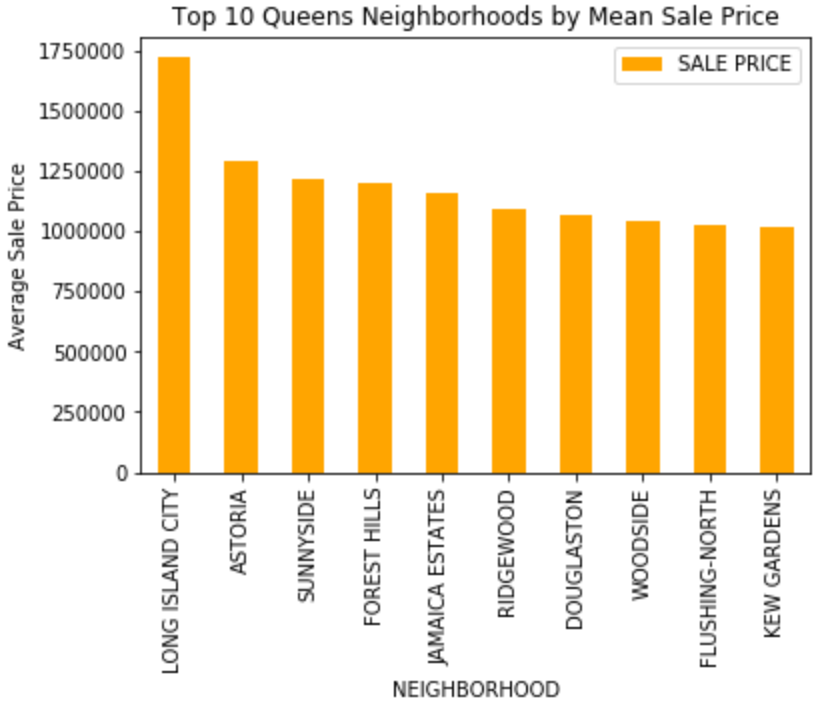
The most frequent property types sold in New York City were homes and apartment buildings. This differs from the most expensive property types, where facilities such as asylums and hospitals were the most expensive on average. Below is a distribution of sale prices by borough represented in boxplots.

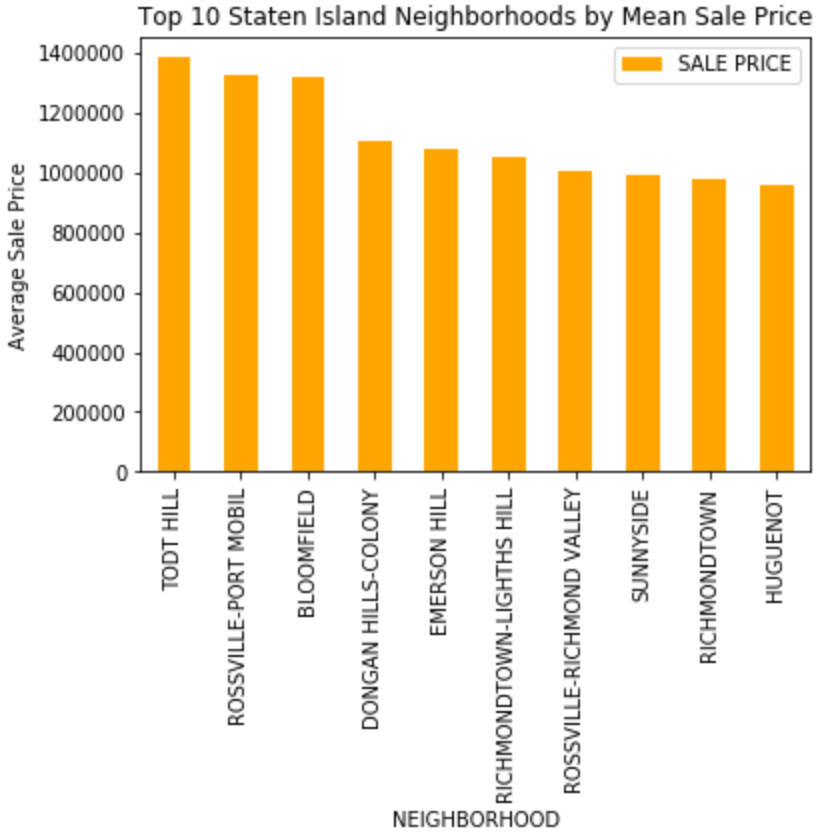


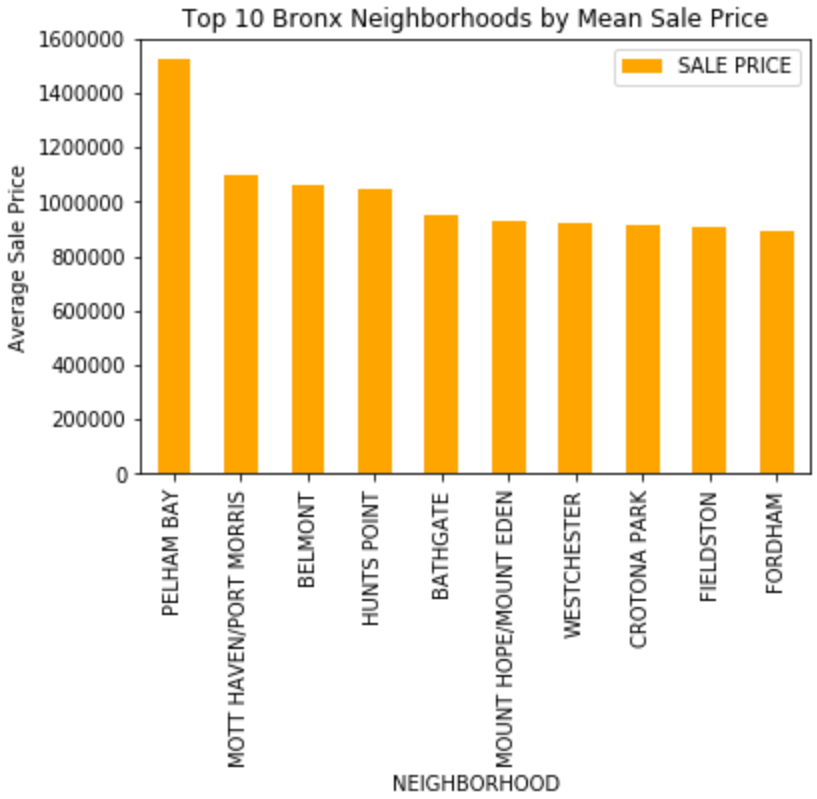
This is consistent with our the earlier results, where Manhattan has the most expensive properties by a significant margin. Its 25th percentile for price is greater than the 75th percentile price of all other boroughs. Staten Island and the Bronx have been very similar with one another in terms of prices throughout and again are the lowest priced as represented by this graph. We can go deeper down into location data by viewing which neighborhoods in each borough sold for the highest average price.






Top 10 Queens Neighborhoods by Mean Sale Price
SALE PRICE
Average Sale Price
1750000
1500000
1250000
1000000
750000
500000
250000
0
LONG ISLAND CITY
ASTORIA
SUNNYSIDE
FOREST HILLS
JAMAICA ESTATES
RIDGEWOOD
DOUGLASTON
WOODSIDE
FLUSHING-NORTH
KEW GARDENS
NEIGHBORHOOD


Top 10 Staten Island Neighborhoods by Mean Sale Price
SALE PRICE
Average Sale Price
1400000
1200000
1000000
800000
600000
400000
200000
0
TODT HILL
ROSSVILLE-PORT MOBIL
BLOOMFIELD
DONGAN HILLS-COLONY
EMERSON HILL
RICHMONDTOWN-LIGHTHS HILL
ROSSVILLE-RICHMOND VALLEY
SUNNYSIDE
RICHMONDTOWN
HUGUENOT
NEIGHBORHOOD



These graphs are mainly consistent from what we have seen 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 any neighborhood in Queens or Staten Island other than Long Island City. Javits Center in Manhattan is the highest overall priced neighborhood on average while Manhattan holds the top 10 overall priced neighborhoods. There 10th most expensive neighborhood is almost double Brooklyn's highest, the next borough with the top priced neighborhoods.

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 on the different iterations of the dataset available to see which performs best based on the information available.

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. The first dataset tested on is the main one used in the analysis above, where all null values are

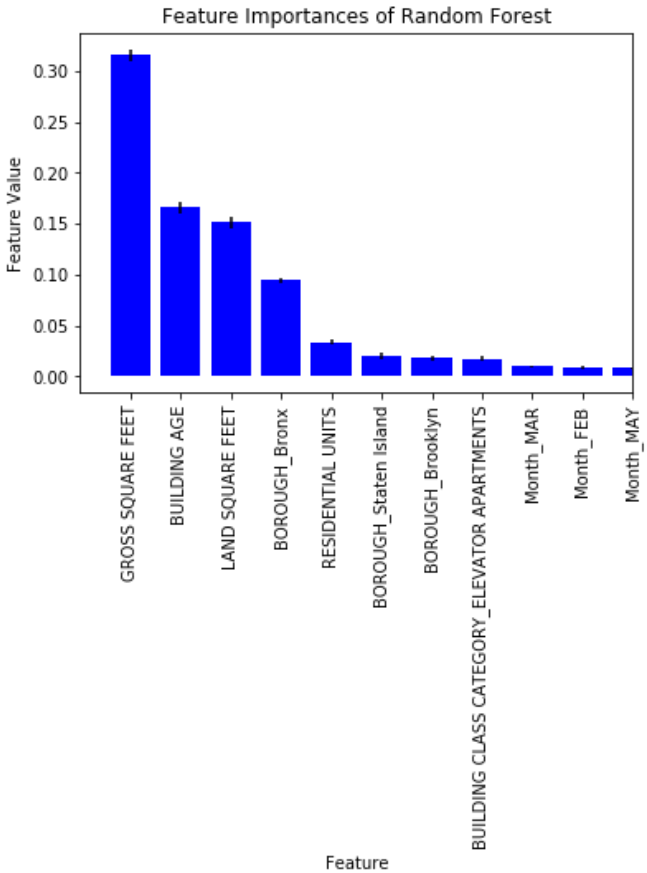
deleted but values of zero were left untouched. All of the dataframes used in modelling were broken into training and test data splits of 80% training data and 20% test data. For the datafame labelled df_o of size 35,007, 28,005 of those go to the training data and 7,002 go towards the test data. The next dataframe from our dataset is df_m, with forward filled values and no outliers, consisting of 79,709 properties, That gives 63,737 values to be used for training and 15,942 values that are predicted upon. The most similar dataframe to df_m is df, which includes all of df_m's values along with the upper bound outliers for price. This dataset of 83,783 has 67,026 values in the training set and 16,757 values in the test set. The last and smallest dataset is df_n where all null and zero values were removed. This dataframe consists of 29,162 rows, where 23,329 values are included in the training set and 5,833 in the test set. All of these models have a considerable amount of data to them to try and gather an accurate predictor of price.

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

df_o (No nulls, zeros):

Linear Regression:
R^2: -16092131431196237824.000
Root Mean Squared Error: 4020494475.945
Average 5-Fold CV Score: -3.7446265252874486e+19
[-1.72903580e+20 -1.29149249e+19 -1.51537935e+17 2.93069491e-01
 -1.26128389e+18]

Random Forest:
R^2: 0.447
Root Mean Squared Error: 0.745
Average 5-Fold CV Score: 0.42970
[0.4287971 0.44980749 0.43517681 0.40149778
0.43321305]



Ridge Regression:
R^2: 0.303
Root Mean Squared Error: 0.837
Average 5-Fold CV Score: 0.30289
[0.30552408 0.31068041 0.30578442 0.29296507
0.29947233]

Elastic Net Regression:
Tuned ElasticNet l1 ratio: {'l1_ratio': 0.0}
Tuned ElasticNet R^2: 0.1473
Tuned ElasticNet MSE: 0.8565

The best model used on this set of data was the Random Forest with a root mean square error of .745. Based on the feature importance chart for the Random Forest model, gross square footage is the most important feature in the model, followed by building age, and land square footage. The most important feature for borough is the Bronx, which is not surprising given it has consistently been the lowest priced borough.

df_m (No nulls, or zeros, forward fill average values, without high outlier values):

Linear Regression:

R^2: 0.189
Root Mean Squared Error: 0.897
Average 5-Fold CV Score: -4.60504188652702e+18
[-1.48385143e+19 -7.63589634e+16 1.91011320e-01 2.01206119e-01
 -8.11033617e+18]

Random Forest:

R^2: 0.174
Root Mean Squared Error: 0.905
Average 5-Fold CV Score: 0.16598
[0.16129846 0.18016286 0.17912653 0.14208922 0.16721909]

Ridge Regression:

R^2: 0.189
Root Mean Squared Error: 0.897
Average 5-Fold CV Score: 0.1949
[0.19723403 0.19144762 0.19124274 0.20100839 0.19355558]

Elastic Net Regression:

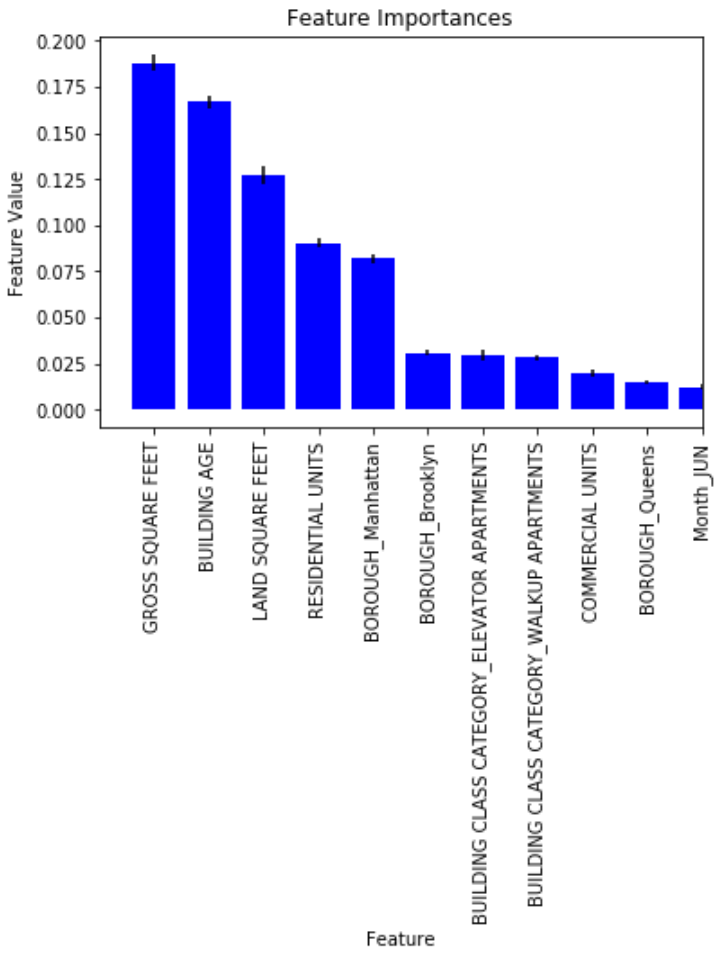
Tuned ElasticNet l1 ratio: {'l1_ratio': 0.0}
Tuned ElasticNet R^2: 0.0670
Tuned ElasticNet MSE: 0.9251

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 .189 and .897 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.

df (No nulls, or zeros, forward fill average values, with high outlier values):

Linear Regression:
R^2: 0.278
Root Mean Squared Error: 0.861
Average 5-Fold CV Score: -1.7892384042431547e+18
[2.63405483e-01 2.60810577e-01 2.77369847e-01 -8.94619202e+18
  2.55172872e-01]

Random Forest:
R^2: 0.271
Root Mean Squared Error: 0.865
Average 5-Fold CV Score: 0.24923
[0.26105479 0.2189147 0.26460067 0.26100433
0.24057206]



Ridge Regression:
R^2: 0.278
Root Mean Squared Error: 0.861
Average 5-Fold CV Score: 0.26461
[0.26556808 0.26257665 0.27898279 0.26078695
0.2551135]

Elastic Net Regression:
Tuned ElasticNet l1 ratio: {'l1_ratio': 0.0}
Tuned ElasticNet R squared: 0.1156
Tuned ElasticNet MSE: 0.9075

These models all had better scores than that of the previous data without the high outliers. This shows that the most extreme data values can provide valuable insight in terms of accurately estimating price based on the property's features. Based on our Random Forest feature importance graph, gross square footage is the most important feature in the model, followed by building age, land square footage, then the amount of residential units. The most important feature for borough is Manhattan, which makes sense given it has consistently been the top priced borough by wide margins.

df_n (Nulls and zero values deleted):

Linear Regression:
R^2: 0.074
Root Mean Squared Error: 0.978
Average 5-Fold CV Score: 0.0780536013513172
[0.08932623 0.06339105 0.07006086 0.06966762 0.09782226]

Random Forest:
R^2: -0.066
Root Mean Squared Error: 1.049
Average 5-Fold CV Score: -0.06655
[-0.03245468 -0.08069875 -0.10194674 -0.07909416 -0.03853894]

Ridge Regression:
R^2: 0.0737
Root Mean Squared Error: 0.978
Average 5-Fold CV Score: 0.078234
[0.08909331 0.0636173 0.07037561 0.07043282 0.09764889]

Elastic Net Regression:
Tuned ElasticNet l1 ratio: {'l1_ratio': 0.0}
Tuned ElasticNet R squared: 0.0569
Tuned ElasticNet MSE: 0.9742

These models performed consistently the worst of all the datasets, possibly because there was the least amount of available data for the model to test. While none of the models on any of the four datasets provided something that one could use to accurately predict the price of a property based on its features most of the time, they could possibly be a good building block if there was more available data to work with and more characteristics of a property listed.

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. 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 aesthetically pleasing 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.