Capstone Project: Milestone Report 1
Jeffrey Ebert

New York City Property Sales

With over 8.6 million people living in New York City, it is a national leader for real estate. There are thousands of various property transactions occurring every month, varying from the selling and buying of homes, apartment complexes, office buildings, warehouses, factories and more. As an individual looking to buy or sell property in New York, it is important to know the appropriate prices at which properties should be evaluated. With so many factors regarding a building's attributes such as its size, style, location and what the building contains, 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.

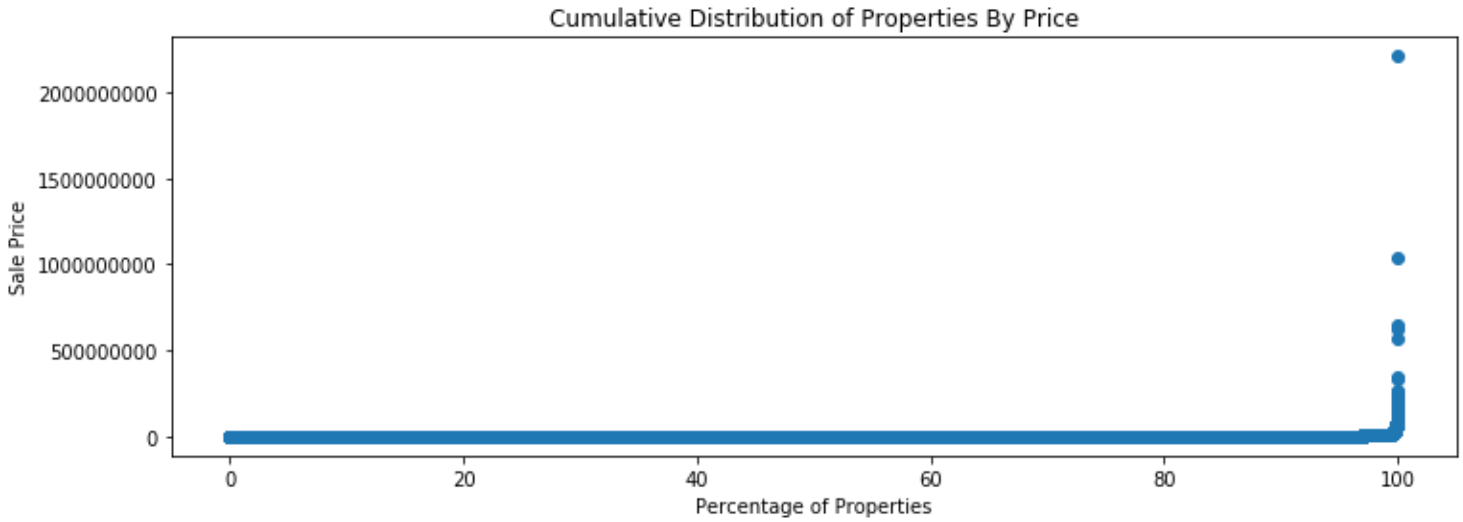
The data used for this project is comprised of all property transactions in New York City during a one year period from September 1st 2016 through August 31st 2017. Developed by the city of New York, it was uploaded from Kaggle.com as a public dataset. There are 84,548 data entries, each denoting one 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 its associated price.

For the purposes of formulating accurate and insightful analysis, the data needed to be cleaned and many values were ultimately removed. First, the dataset contained columns that did not provide any useful information. These data columns were all removed. Then I found that the data contained 765 duplicate values. These entries were removed from the dataset so each listing would only count once. Next, columns such as the square footage, building age, and price were converted to numerical data types, while columns such as tax class, zip code, and lot were converted to categorical data types. The borough data column was listed as values from 1-5

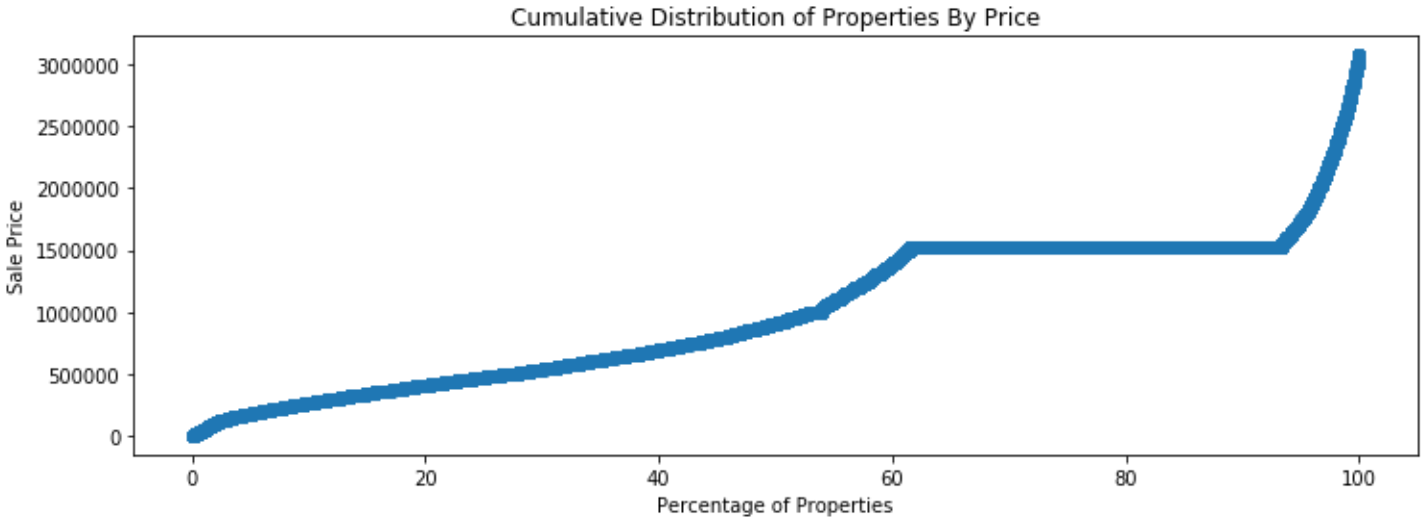
designated the borough the property was in. To make future analysis more clear it was important to note which value was assigned to which borough by checking its neighborhood and then changing the values to their appropriate titles.

There ended up being 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 to not skew the data. There were over 6000 data values that had a zero value for the year the building was built. These values will be 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 many upper bound outliers in the data by price that significantly altered the data. This was because of many sales on Manhattan homes that were much greater than any other properties. In order to provide a more balanced analysis these values were removed, but will be included again when evaluating various models based on price as the most extreme values could potentially help the models the most.

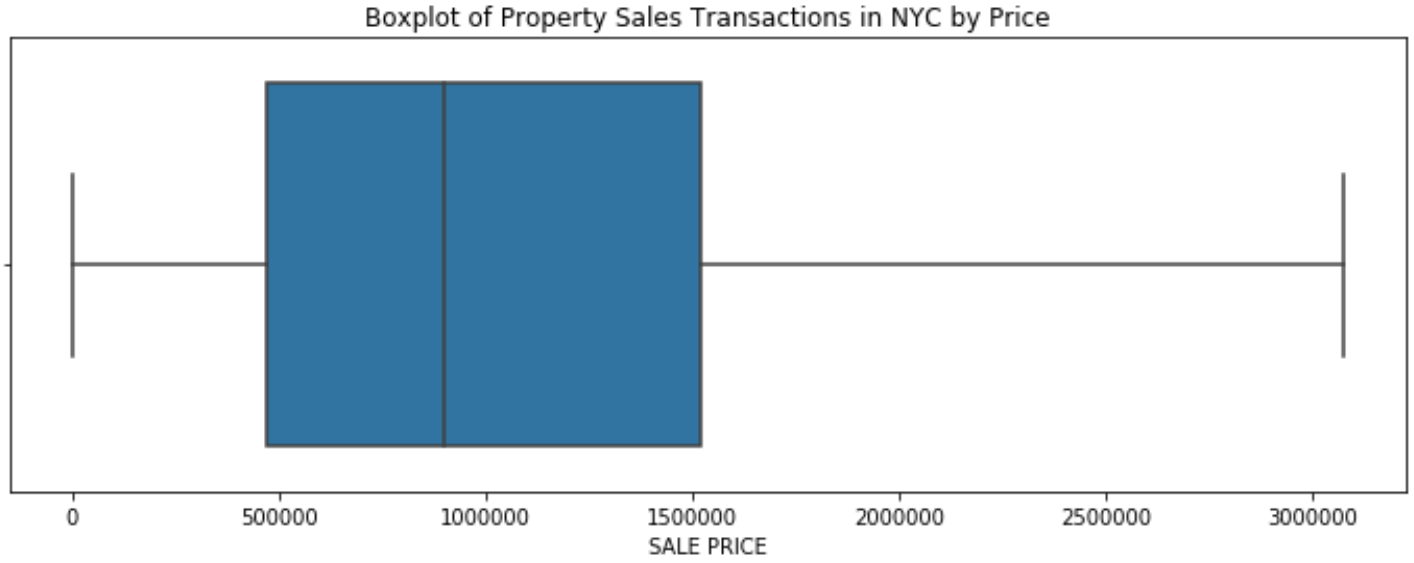
The first objective of the project is to understand the breakdown of the data available and evaluate the different types of transactions made with the distinctions between them that have an influence on the property's sale price. In doing so, we will first analyze the numerical features of a property.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



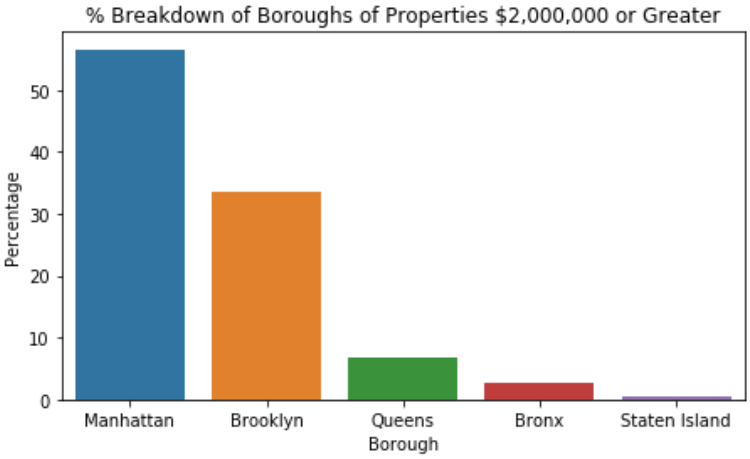
This cumulative distribution chart shows that there are a few outliers that highly skew the data. This makes sense as many of the most highly priced properties in the world are in Manhattan. After removing these outliers our graph looks like this.

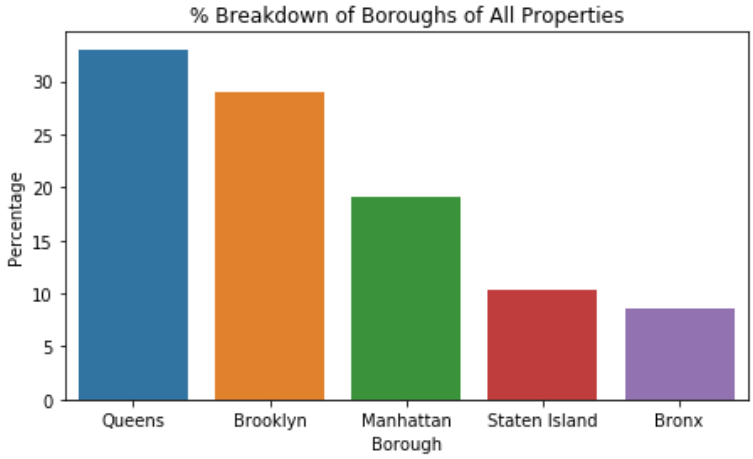


With a more narrow scope for price, we see a more gradual increase in the percentage of properties by price in the cumulative distribution graph. There is a gradual increase in the percentage of properties up to approximately 1$00,000, with a steeper increase up to approximately $1,500,000. Then, due to filling the null and smallest values with the average value for price the curve is steady for over 30% of the data. For the remainder of the distribution, the steepest incline exists indicating a diminishing amount of property sales with the highest values in price. This is similar to what we saw in the previous graph, but it is more pronounced now without the extreme outliers in the data present. Let's take a look at this data in a different visual context.

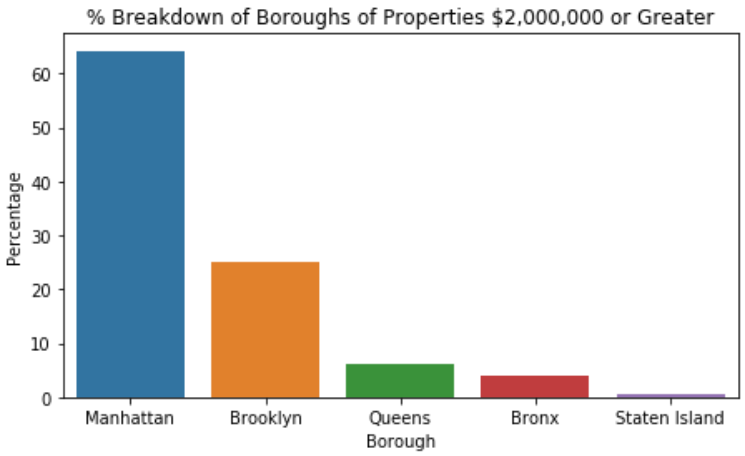


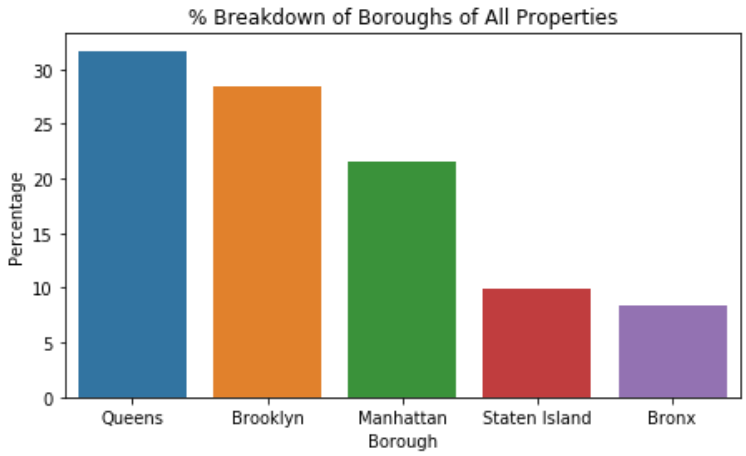
Here we can see the data is most concentrated in the price range from $500,000 to $1,500,000, extending up to over $3,000,000. The upper echelon of properties in this data have a price range of 2 to 3 million dollars. The next analysis will take a look at what are the most common attributes of the most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



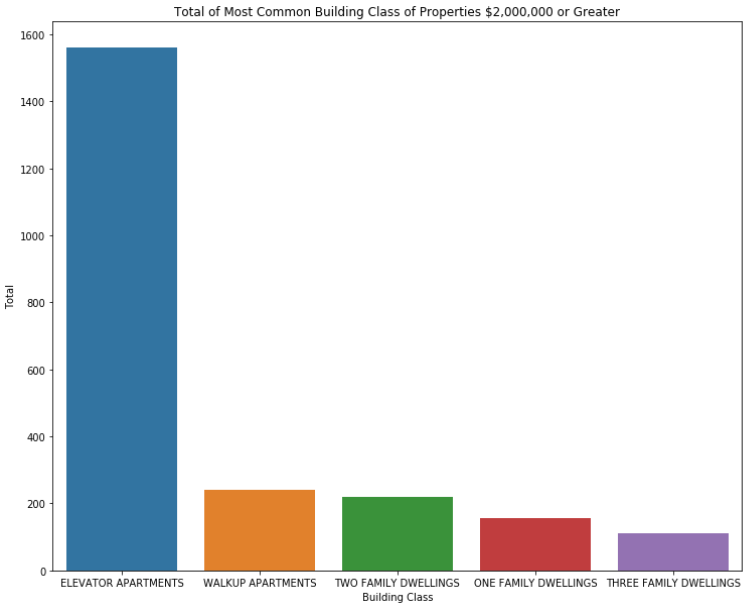


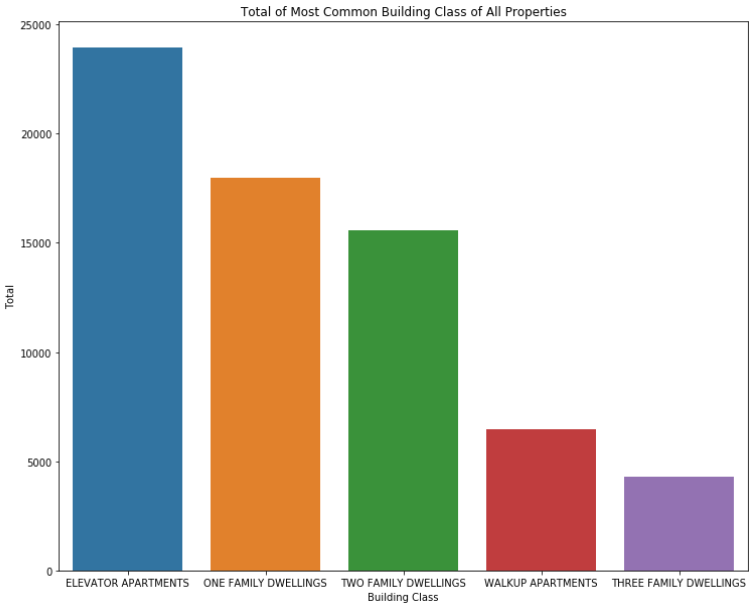
The graph on the left shows the percentage of properties by borough of transactions over 2 million dollars. Manhattan has more than half of the total share and next highest is Brooklyn with just over 30%. Staten Island has the least amount despite being ahead of the Bronx in having more total transactions. Even though Queens has the highest percentage of properties overall, they have less than 10% of the properties sold at over 2 million dollars. It is important to note that these visuals do not include the upper bound outliers we removed. Using those data points, these graphs would look like this:



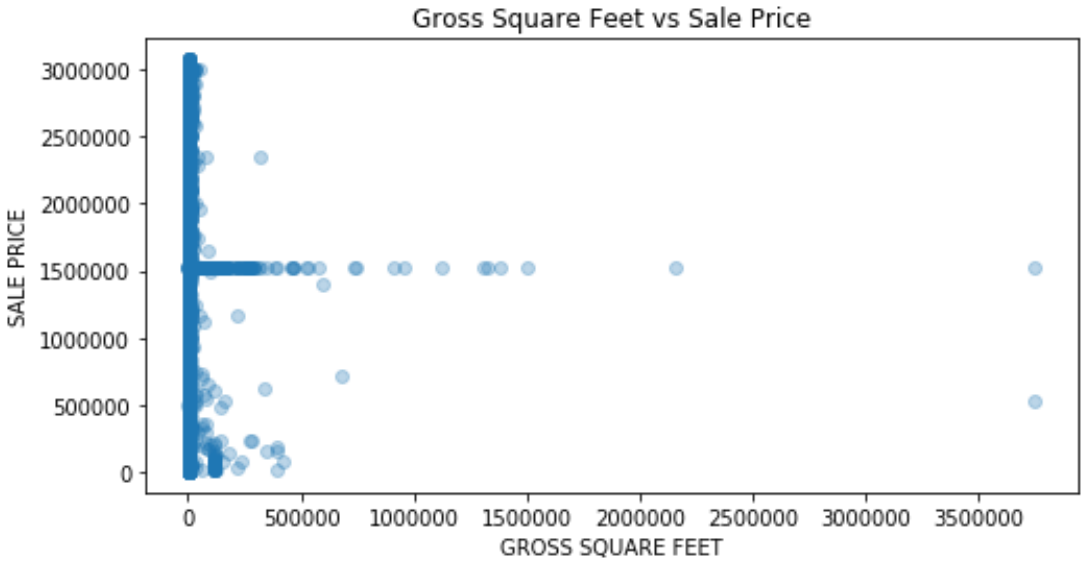


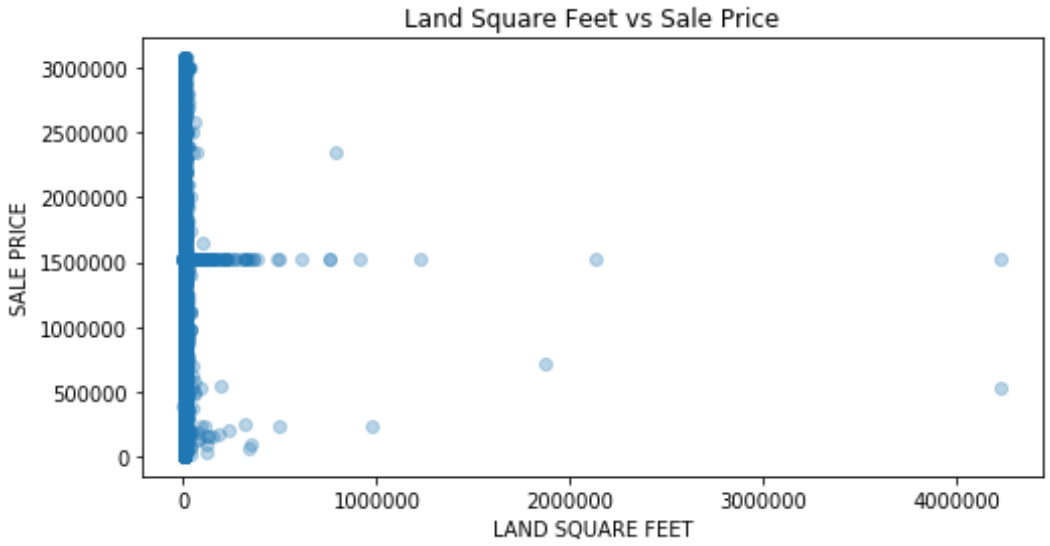
These two graphs comparisons are very similar to what we have just observed, with a noticeable difference in the increase in percentage share of Manhattan in having the most expensive properties. It is clear there are more very high prices in Manhattan than the other four boroughs combined. Next we will look at the most common building categories of the most expensive properties.



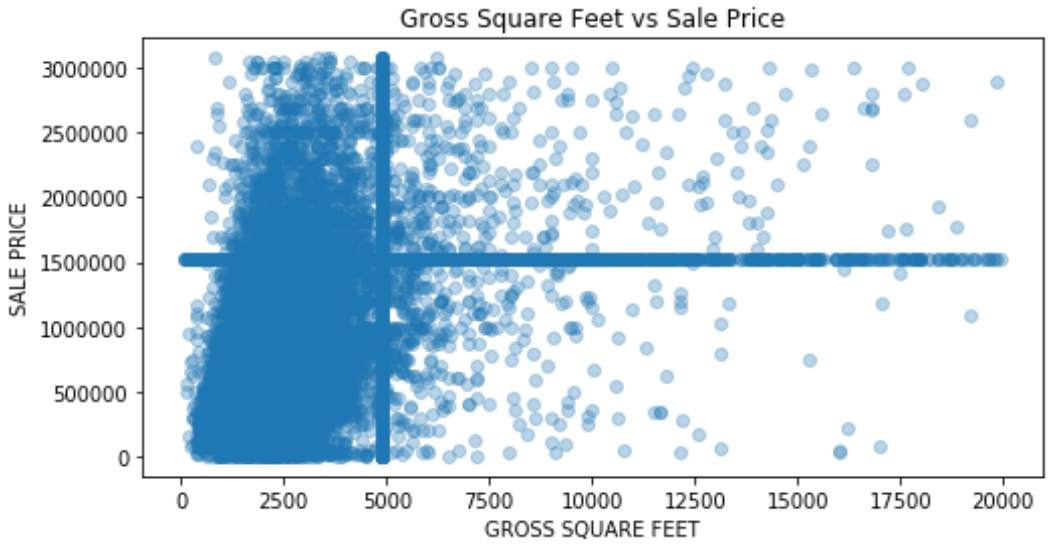


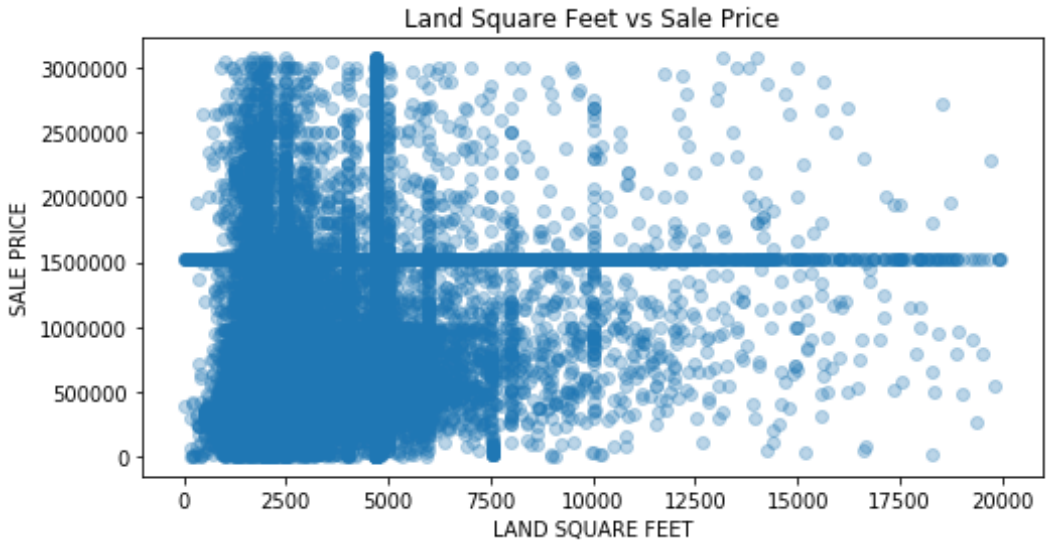
Of the most expensive properties, the majority of them are overwhelmingly elevator apartments, which is the most frequent building class of all the properties. It is interesting that the most common properties also make up the most frequent of the most expensive properties as well. Other variables that influence a property's price are its size, represented by square footage and the amount of units it contains. The following is a distribution of prices by their gross square footage.



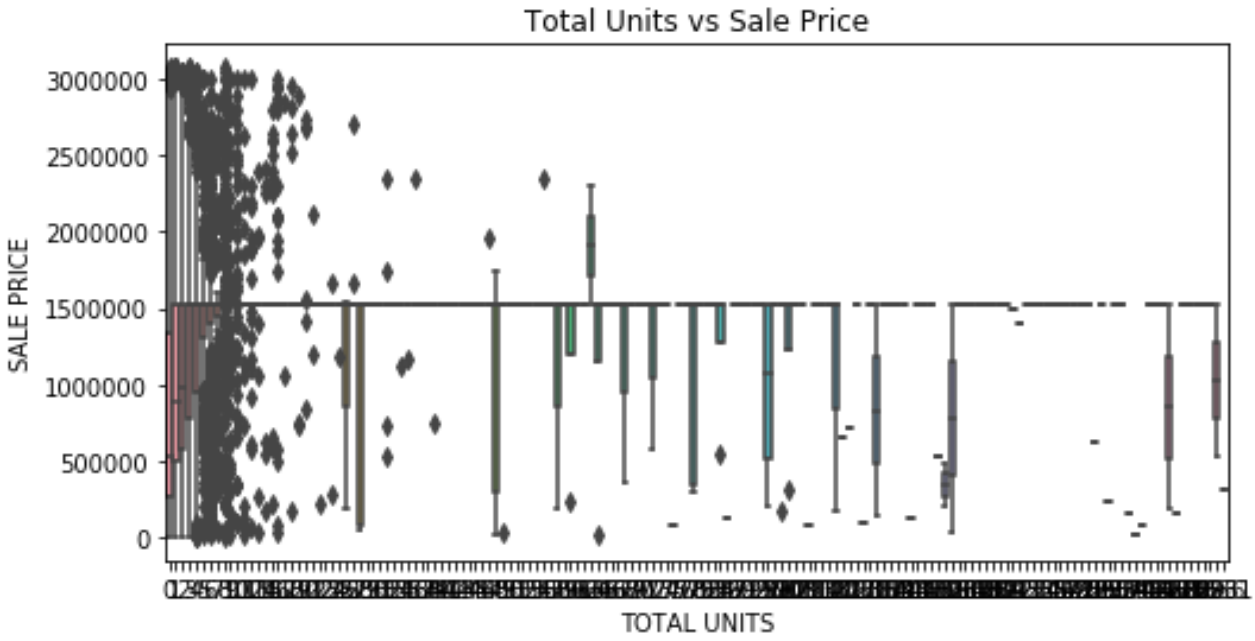


There are a few outliers skewing the data in these graphs. Let's see what the distribution between price and square footage looks like with these values removed.

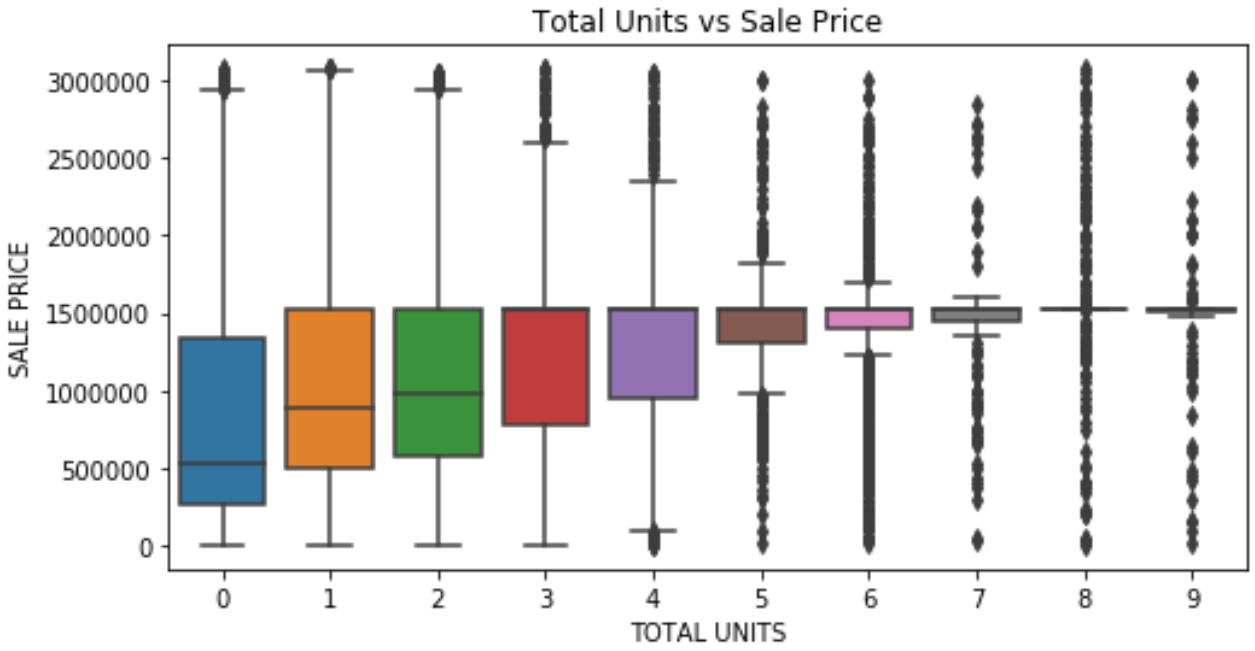




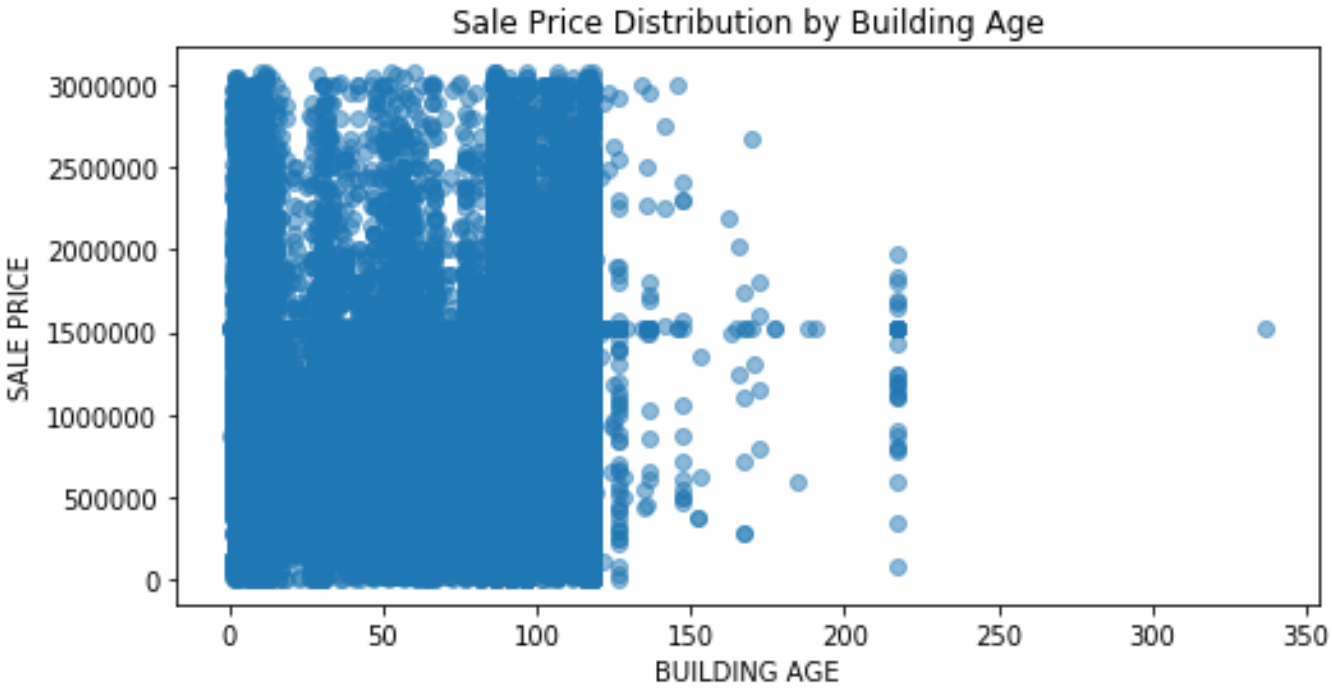
These graphs show the relationship between gross square footage and land square footage with price for properties in New York City. There does not appear to be a significant relationship for land square feet, which represents the total land area of the property. There are many data points that have larger 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. There are many fewer of the greater square footage data points below the average than there are above the average. As the gross square footage increases, 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, such as its location and other features besides its size. Next will be observing how units affect price.



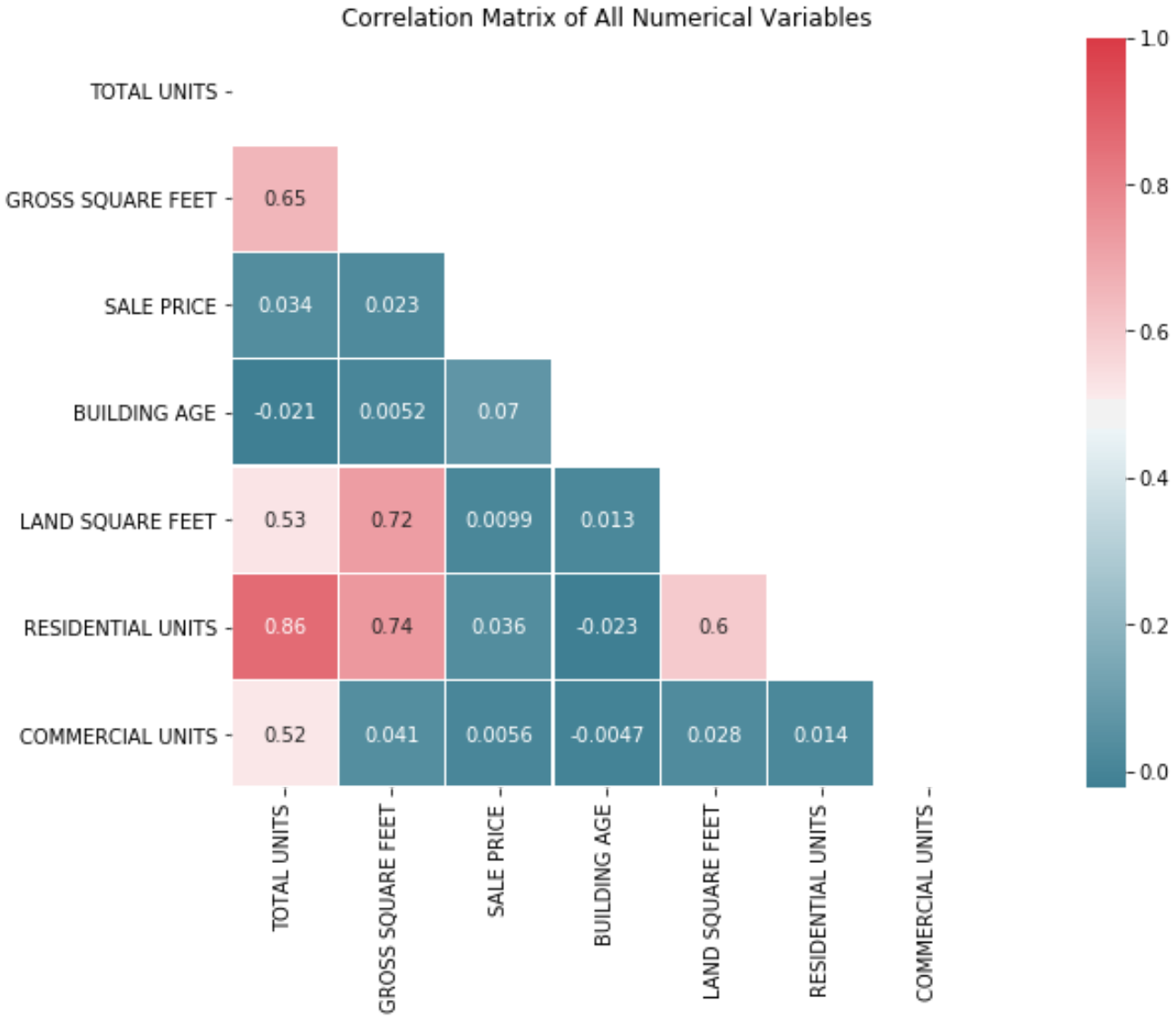
This graph does not tell us much due to the high variance in the amount of total units in the dataset. As a result, we can exclude the upper value for units. The majority of the data is comprised of properties with less than ten units.



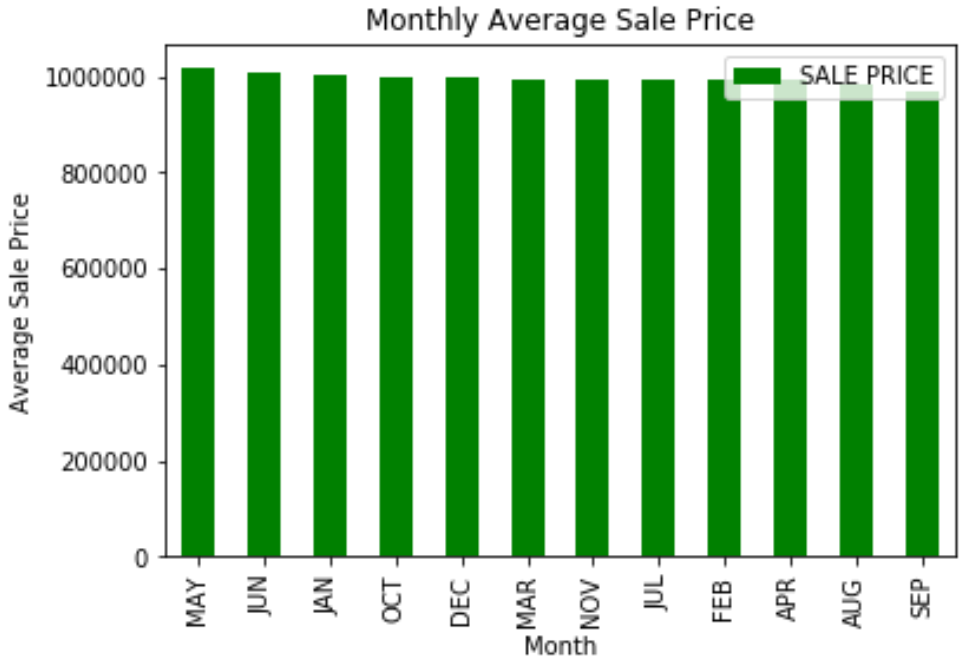
These boxplots show a general trend of the median price increasing as the amount of units increase when the value for total units is low. As the values increase the median price levels off. This makes sense logistically because when more space is added the price increases, but there becomes a point where the addition of more units does not influence price as significantly. Next we can observe how the age of a building impacts price.



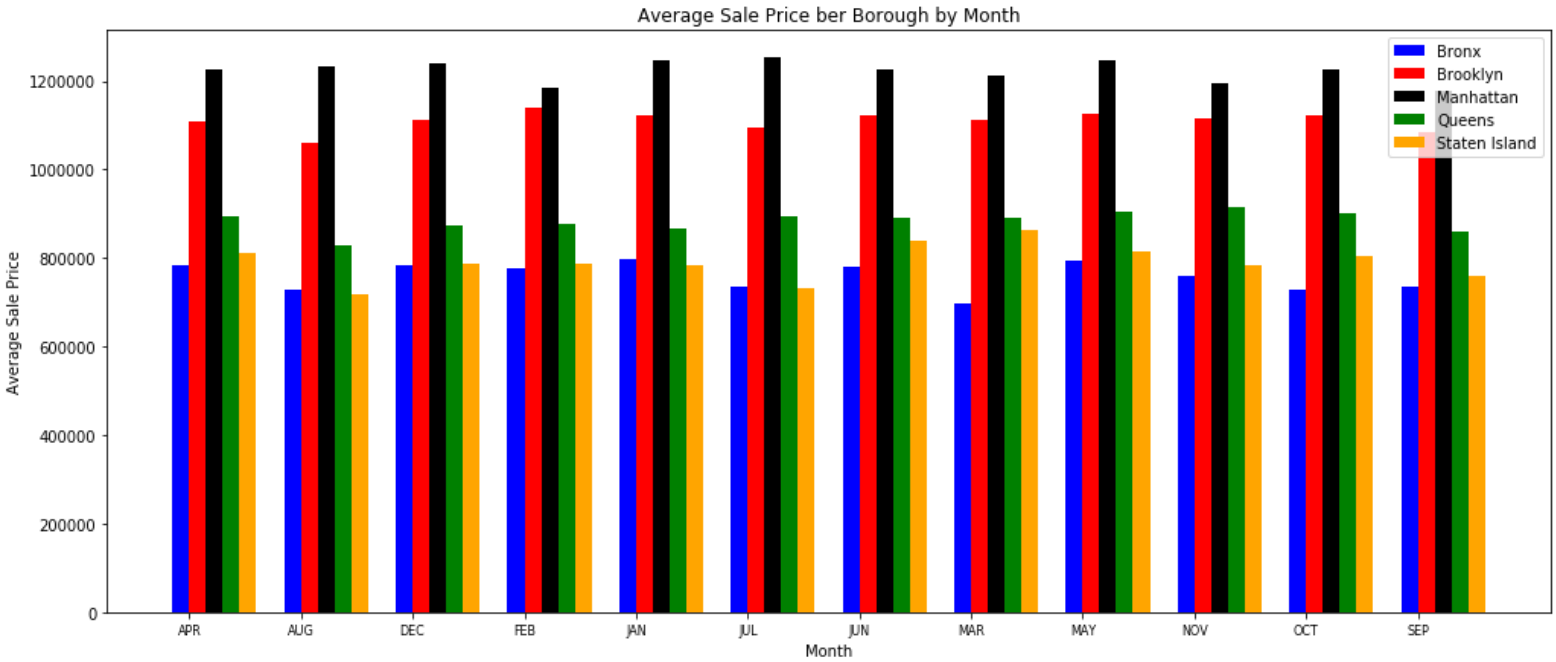
The age of the building does not appear to tell us much in terms of its price. Most of the building ages are around 120 years old or less and are priced consistently. The oldest buildings, aged around 125 years or more, are easier to view on the graph because there are considerably less of them. Some 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. Now that we have viewed the relationships between price and a property's numerical aspects we can view their true correlations.



Here we can see the correlations between all of the numerical values in the dataset. The relationships with the highest correlations are residential units with total units and gross square feet with residential units, which make sense intuitively. The relationships most significant for our analysis are associated with price. Total units and gross square footage have the highest correlations with price, which is largely what we saw based off of the previous visualizations. One variable not being evaluated here is the date. Now we will shift to observing the categorical variables in the dataset and how they are related to price. The next visuals represents how price changes by the time of year the property is sold.

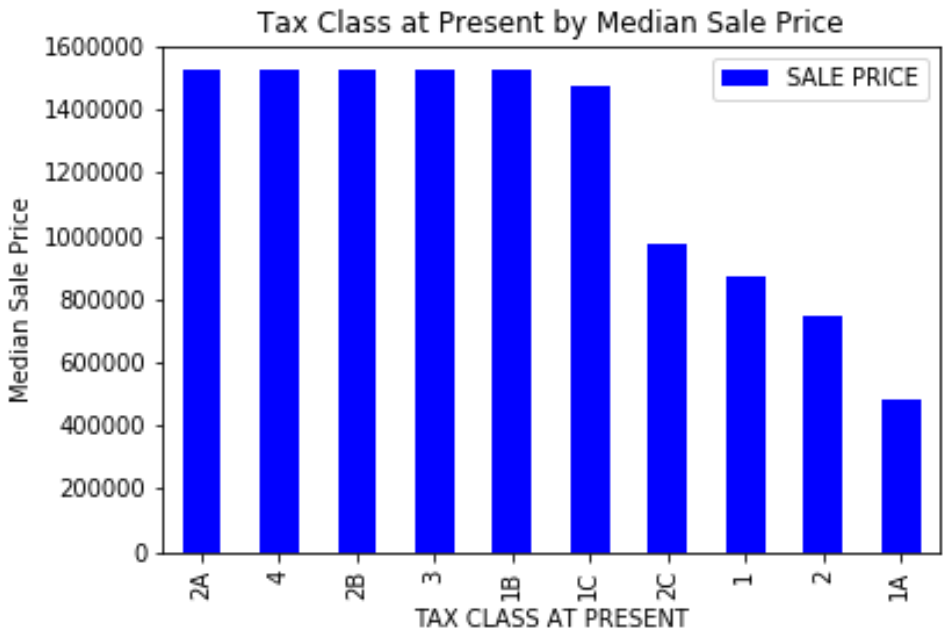


Based on the graph, the month of May has the largest average sale price with June and January just behind it. May and June having the greatest values for average price could be related to more demand for new properties just before the summer compared to winter or other colder months in New York. If one is looking for the most economical property transaction it could be wise to complete it in the fall, just after summer is over, which had the lowest average monthly prices of August and September. A more in depth look at this will evaluate average monthly price by the property's borough.



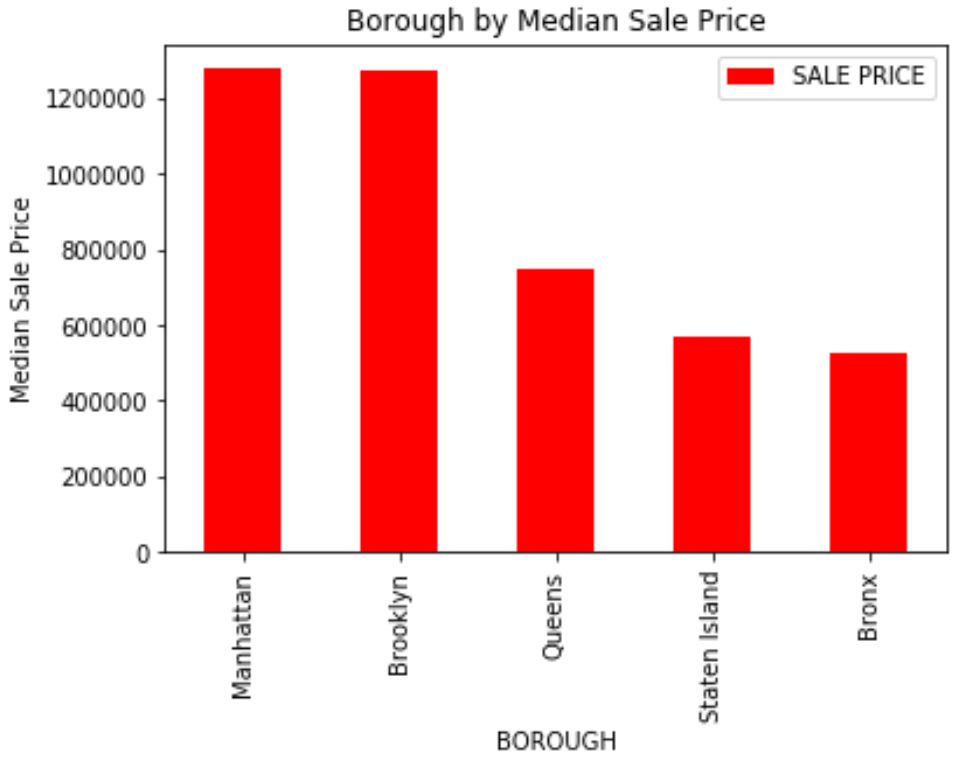
This graph represents each borough's average sale price by month, listed alphabetically. From this we are able to notice the differences in price by borough comparing month to month. In Manhattan, these differences are most dramatic due to some higher property values, however the values appear mainly consistent. While May has the highest average overall, it is only the highest averaged price month in Brooklyn, but it is among the top few in each borough. Let's see if the monthly data is consistent with the results by season.

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 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 2A. These are comprised of the sales of mainly apartment building complexes. Class 1 has the lowest median prices of the tax groups.

We have previously seen the breakdown of the five boroughs by their most expensive properties, but we can see if that is consistent with their overall median prices.



This is similar to what we saw earlier in the analysis, where Manhattan is the leader among expensive properties, but 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.

This analysis has shown the differences in the dataset's properties sold in New York City from September 2016 to August 2017. There is much to consider when evaluating the price of a property, such as its type, where it is located, its features and its size. While we have explored many different variables related to a building's price there are many others that this data does not mention that could be an important factor in how people evaluate the properties. Information regarding the condition of a property, its amenities, proximity to subways and available parking are not referenced and could all have an influence on how a property is priced.