**Manhattan Sales Price Prediction Report**

**Sponsored By APT212 INC.**

**BUS 9430**

**Group 3**

**Yaxin Xu, Jessica Jiang, Eric Sedaghat, Zhirui Chen**

**INDEX**

1. **Executive Summary**
2. **Organization**
3. **Business Problem Statement**
4. **Analytical Problem Statement**
5. **Problem Solving Approach**
6. **Findings and Results**
7. **Risks and Improvements**
8. **Conclusion**
9. **Appendix**

**1. Executive Summary**

As one of the most significant real estate markets in the world, New York City has long been a target and promising market among investors. Being able to identify undervalued assets as well as being able to stay away from overpriced assets are some of the keys to being successful in real estate investment. Naturally, having the ability to keep up with the market and predict property prices for better investment decisions has been a part of every company's intrinsic goals in the real estate industry. Data analysis and machine learning algorithms have introduced new and more promising paths to achieve this goal - they offer important insights into the current trends and help investors determine the profitability of a property. More and more companies are starting to use or have been using data to improve their operations and adapt machine learning to investment functions. Data analysis allows for a fully objective and reproducible way of pricing assets without the risk of the bias of an appraiser inaccurately pricing an asset.

APT212, established in 2010, has been focusing on the Manhattan real estate market, providing renters, buyers, sellers a powerful marketplace and offering clients a streamlined, first-rate experience with innovation technology and human relationships. APT212 has an investment projection model built in Excel that factors in a property’s mortgage payment, operating cash flow, sales assumptions, ROI, IRR and net gain. Most of the outcomes are generated automatically with the default formulas, but users still need to input the projected rental and sales prices by themselves. There are a number of factors that impact real estate prices. Interest rates can impact the price and demand of real estate. Prices can also follow the cycles of the economy. Government policies and legislation can also boost or hinder the demand of real estate (Nguyen). Thus, it is hard to predict a property’s value without a certain real estate background. APT212 wants to deliver a model which is suitable for every user - no matter if you are an experienced investor or a first-time buyer who is new to the New York real estate market. The solution to this problem is to build a prediction model providing users a projected rental and sales price based on property’s information, for example, square feet, zip code, neighborhood, last sales price and help users to make better investment decisions.

Due to limited time, we focused on building the Manhattan sales price model. The dataset we used came from NYC Open Data, an annualized sales file displaying yearly sales information of properties sold in New York City. To get this dataset ready for modeling, we did data cleaning, filter based feature selection, and feature engineering. The final dataset we used had eight variables (neighborhood, zip code, year built, sale price, sale date, year, month and day) and 21,286 unique sales listings from January 1st, 2016 to December 31st, 2020. Then, we used xgboost, random forest, lasso regression, support vector regression and neural networks to run models. For each modeling algorithm, we used all numeric variables first then we did one hot encoding to the categorical columns. Having compared the mean squared error (MSE) for the training set and validation set, the xgboost model using categorical columns has the best performance, with 0.176 MSE for the training set and 0.297 for the validation set. Based on the result we have from the xgboost model, neighborhood plays a big role in the sales price change.

**2. Organization**

APT212 is a powerful, growing and disruptive force in New York City real estate, with services including sales, new developments, financing solutions, rentals and furnished short term rentals. APT212 provides personalized accommodation solutions for customer’s needs and budget, making the rental, sales process smooth and pressure-free. As one of the largest short-term rental platforms in New York, APT212 offers affordable and luxury rental services with an instant online booking system and connections to local booking agents. APT212 also provides a marketplace for owners to list their properties for rent or sales.

Our main focus while beginning the project was building a predictive model on rental values and sales price in Manhattan based on the market and property information to deliver a more user-friendly investment model. Due to the limited time, we decided to focus on the sales price prediction model. Our sales prediction model will help users who are not familiar with the real estate market to develop a better understanding of what type of returns on their investments they can expect. The model will also assist APT212 in seeing what type of properties yield the highest sales price as well as those with the lowest. It can also potentially show the demand in certain neighborhoods as some are more favorable than others. Additionally, it will help APT212 to see what buyers want out of a property. Key stakeholders in this project are the company, potential investors, property owners, and website visitors who are redirected to the website looking to use the investment model or learn more about the real estate industry.

**3. Business Problem Statement**

It’s no secret that the rich have been able to get richer by investing in real estate in New York City. Most people do not have the time nor the background to get involved in what could be an extremely lucrative investment opportunity regarding different properties in the city. APT212 aims to extend this opportunity to everyday people. APT212 wants to make investing in a property in the city more intuitive and available to common people, however as it stands the investment projection model requires users to input sales and rental prices in order to get accurate investment metrics. For someone who does not have a background in real estate and also does not have the time to research historical price charts and rent rolls, inputting these numbers with any confidence is nearly impossible. Driving away the business that APT212 is really targeting. By creating a predictive model that can take historical data and project sales value as well as rental values we are empowering the customer to confidently inform the investment projection model. As a result, someone who may not know much about real estate will now have the opportunity to fill out APT212’s investment projection model and look at the opportunity to purchase a unit the same as they would any other investment using metrics like ROI, IRR, and gains.

There are several different types of real estate analytics like descriptive, diagnostic, predictive and prescriptive. Descriptive helps derive useful information from the past. Diagnostics uses historical data to explain why something happened. Predictive forecasts into the future and prescriptive provides insight on what needs to be done to take advantage of an opportunity or avoid in the future. Prior to starting our project we thought it would be prudent to research similar projects conducted by others in the past to see what we can learn from them. First we looked at similar projects that were posted on Kaggle to see if there were any prevalent problems that we may need to look out for when conducting our project. We noted that a majority of the projects tended to follow relatively normal data cleaning methodologies without any significant or unique issues and conducted their feature selection according to what they were trying to achieve. Additionally, across Real Estate price prediction projects as a whole we noted that most found success incorporating models like XGBoost, Random Forest, and Lasso Regression in the model fitting stage which we took note of before moving on with our own project. Models can have overfitting or underfitting problems, which require the data science team to adjust the dataset dimensions, train more data, remove some features and even add hidden layers to the neural network. For real estate valuation, many use linear regression but the problems that come with that are autocorrelation and multicollinearity. Besides property valuation, regression analysis is also useful for pricing analysis for list prices and rental rates, demographics/ psychographic analysis of residential buyers and tenants, identifying targets for direct marketing, and ROI analysis for marketing campaigns (Barr).

**4. Analytical Problem Statement**

Our dataset came from NYC Open Data, an annualized sales file displaying yearly sales information of properties sold in New York City. This sales dataset was updated June 15, 2021 and provided by the NYC Department of Finance, which is an up-to-date and reliable source. The original dataset has over 414,000 rows and 29 columns and included properties sold between January 1st, 2016 and December 31st, 2020. Initially our dataset contained sales of multi unit assets both residential and commercial. Each row logged an individual unit or a whole building. From here we faced a myriad of tasks that needed to be completed in order to achieve our business objective. To start, we would clean the dataset as an enormous dataset would negatively affect our model. Feature selection is important to avoid the overfitting. Here we could use forward selection or backward selection so we wouldn’t miss any variables. We needed to identify which columns we would be able to use in order to fit our model; we believed some variables might not contribute any predictive abilities and others ran the risk of overlap of information. Additionally, we needed to filter the values so the dataset would not include irrelevant listings, since we wanted to focus on single unit properties in Manhattan only.

In terms of sales value predictions, before model selection, the data has to be transformed to normalize the features and the response variable. Random forest and gradient boosting appear to be some of the models used for prediction. In terms of choosing the best model, one has to assign the appropriate values to NAs, normalize variables, and optimize the hyperparameters for those models (Jermain). We trained and validated the data with the following models (XGBoost, Random Forest, Lasso Regression, Support Vector Regression, Neural Networks) to see which ones would produce the best performance based on the MSE value. MSE is the average of the square errors. The lower the MSE value the better and 0 refers to a perfect model. Therefore, when we look at the training and validation MSEs for each model, we want them to be as close to 0 as possible. Also, if the MSE ratio for training and validation set difference is lower than 10%, the prediction model is fairly precise.

**5. Problem Solving Approach**

After the initial exploration of this dataset, we decided to drop some features that were not useful for our prediction purposes - tax block, tax lot, easement, BIN (Bank Identification Number), BBL (parcel number). After looking closer at the remaining variables, we excluded some categorical columns including tax class as of final roll, tax class at time of sale, building class category, building class of final roll, address, and NTA, and numeric columns including apartment number, community board, council district, and census tract. Those columns included too many unique values and would expand the dimensions of the dataset largely if we did one-hot-encoding.

In the next stage of the feature selection, we filtered variables and examined the data type for each column. Since APT212 specializes in Manhattan listings, we filtered the “borough” variable to only Manhattan. We also filtered “total units” and “residential units”columns to keep single and residential property. We needed time series data to generate a well-performed model that would accurately make price predictions, so we changed the “sale date” column from object to datetime. Additionally, we did feature engineering to extract the year, month and date from the “sale date” column and added three new variables. In the “sales price” variable, we removed any sales prices less than $10,000 since those listing prices were not realistic. The variables we ended up keeping which we used for modeling are neighborhood, zip code, year built, sale price, sale date, year, month and day.

There are some limitations in the data. Firstly, it includes the sales price and sales date but it is missing information for purchase price and purchase date. We could have used those 4 variables to do some comparison, feature engineering, and data visualization to see the change of property market prices over time. The second limitation is the large amount of missing values. The approach we took to resolve this issue was to drop all the columns that include missing value. As a result, we ended up having concise and consistent data files that were ready for further cleaning. Last but not the least, the land square feet and gross square feet had too many “0” values. Square feet is a key factor to determine a property’s sales price - the larger square feet, the higher sales price generally. Different to the approach to replace the missing values by the mean or median, we couldn’t do the same approach here because the square feet had huge differences based on the property layouts and we had no information in the dataset in apartment bedroom number and bathroom number. Thus, we had to exclude those two key variables to keep the dataset information accurate. Originally, APT212 wanted users to input the property’s square feet and our prediction model would predict the sales price based on price per square feet. In this case, we would have to figure out other predictors, like neighborhood, zip code, year build, sale date.

**6. Findings and Results**

Our exploration revealed some expected patterns and behaviors in the data as well as details that would’ve never been uncovered at a glance. In Figure 1, we did a box plot against sales price distribution. From the above box plot, we could barely see the box. Thus, we dropped listing prices that were below the first percentile and above the 99th percentile. Having cleaned up the outliers, we made a second box plot and it was much more acceptable for modeling. In Figure 2, it displays the correlation matrix heatmap amongst the numeric variables: Zip Code, Year Built, Year, Day and Month. Overall there was very little correlation amongst the five variables with each other. In Figure 3, it displays the number of records per neighborhood in Manhattan. This visual shows that the Upper East Side (79th Street - 96th Street) has the most property sales records followed by the Upper East Side (59th Street - 79th Street). Inwood and Morningside Heights have the least number of sales records. Overall, we can conclude that affluent neighborhoods draw in more sales, especially those near Central Park. In Figure 4, the graph shows the mean sales price for each neighborhood. Greenwich Village West sits comfortably at the top being the most expensive neighborhood on average sales price. Washington Heights Lower, Washington Heights Upper and Inwood are the least expensive neighborhoods on average sales price. In Figure 5, it displays the mean sales price based on the building year built. Properties built in 1938 have the highest average sales price while those built in 1954 have the lowest average sales price. However, we should note that it can also depend on the number of properties that were on sale for that building year. For example, there could’ve been 200 properties that were built in 1938 on sale, while only 10 properties that were built in 1938 were on sale. In Figure 6, we see the comparisons of the mean sales price across the years and across the top 5 mean sales price neighborhoods. The highest average sales price amongst this time span was in early 2019 in Midtown West. Oddly enough, prior to the spike, its average sales price was one of the lowest points. We can also see the exact point in which the pandemic affected the property sales. However, it seems to bounce back to normal levels shortly thereafter.

For the XGBoost model, we first excluded the categorical data to run XGBoost to see the performance. Without using categorical data, the MSE score is 0.1473 for the training set and 0.3188 for the validation set. It is not a good performance as the MSE value for validation set is too high and there is an overfitting problem. The approach we used to improve the model performance was to use one hot encoding to encode both ‘Neighborhood’ and ‘Month’ columns. It kept showing an error saying the training set was missing "NEIGHBORHOOD\_MORNINGSIDE HEIGHTS". Having taken a look at the unique value for this variable, we realized that there was only one unique value, which made sense now because the only value had been sitting on the validation set. We needed to exclude this column. After exclusion, we got an MSE of 0.1764 for the training set and 0.2967 for the validation set. It is better to keep the categorical columns. After we added encoding to categorical data, the MSE for the validation set became smaller and got closer to the MSE of the training set even though the MSE of the training set got slightly higher. The difference of MSE here is around 12%, slightly higher than 10%, meaning there was still a small overfitting problem. We also tried several times to adjust parameters in order to see how the MSE scores changed. However, due to the limited variables of the data, 0.17 and 0.29 was the best score we could get. Also, the XGBoost indicates that neighborhood is the most important feature for predicting price.

For the other four models, we repeated the steps above - ran all numeric variables first and then used categorical variables with the one-hot-encoding. We realized all models had better performance (lower MSE scores) with the categorical columns. The Figure 7 and 8 show the MSE scores for all the five models we used. Overall, the XGBoost model performed the best as it had the lowest MSE scores and acceptable ratio difference (12%) between training set and validation set. Random Forest, Lasso Regression and SVR did not have overfitting problems but the MSE scores were over 0.50. Neural networks had a big overfitting problem, as there was a 23% difference between the training and validation MSEs.

**7. Risks and Improvements**

Some risks that occur in this project are quality, market, and resources. The real estate market is tricky. By doing analysis, investors and companies can understand the housing market. By forecasting, it provides information about the future potential of a particular market. Trends in the market can potentially be risky because it could end up being a good investment then later on end up not being the best property due to a variety of factors. There’s competitive and reputation risks, making sure the company is staying ahead of its competitors and gaining an edge above the others. This also factors in the quality risk. As much as a company wants the latest data to edge out the competition, they should also make sure the models and the data that they use is the best available. The quality of the model should not suffer because of the lack of time. Another risk is financial. The model has to be accurate enough that users feel comfortable putting that property up for that sales price, or users believe that a certain property would be sold for a certain amount. The company is also investing time and money into this model in hopes that this will bring them new customers.

While our project used standard variables to make the prediction model, the real estate industry has evolved by also including new variables that can help determine a property’s sales price. It is a mix of community features. For example, the number of grocery stores can increase the price but having too many can also lower the price. Macroeconomic and demographic indicators like crime rate and median age, can also impact pricing. This would involve the team analyzing and figuring out the weight of each of these nontraditional variables. How much does each of these indicators affect the sales price? Proximity to highly rated restaurants or changes in the number of nearby apparel stores explain 60% of changes in rent (Asaftei). By expanding and improving our current model with these variables, realtors, investors and home buyers have access to information that will help them make smarter investment decisions taking into account risks and market trends.

An improvement here is to keep more categorical variables before we exclude them. For example, given more time, we would add “Building class category” as a variable for the modeling. There were unique values like “Rentals-Walk Up Building”, “Rentals-Elevator Building”，“Condos Building”, and “Coop Building”. We could add one-hot-encoding here because the “Building class category” implies additional property information and they can differ the sales price. Even though the reason why we dropped this variable was the big number of its unique variables, we could do some text mining here. We can transfer all unique variables containing the key words “Walkup”, “Elevator”, “Coop”, “Condo” into this single word. In this case, we can minimize the unique values and it may help with the overfitting problem. Another improvement is to add a price predicting calculator to the model. We failed to do so due to the missing variables of the purchase date and purchase price. The price calculator could generate current market values by calculating the distance between the property’s last and sold price and dates converted to standard units. For our future project on price prediction, we can apply this price calculator if the dataset included both purchase and sales details. An example of a possible outcome is shown in Figure 9.

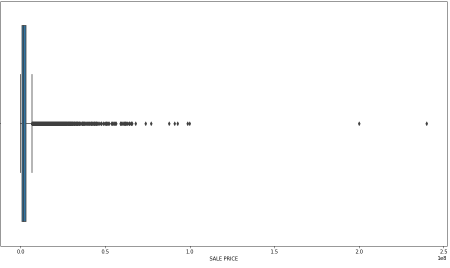
**8. Conclusion**

Xgboost model has the best performance and based on the result, neighborhood is a big factor in determining sales price, which aligns with our original hypothesis. An affluent neighborhood is more likely to enlist higher pricing for their property therefore excluding certain groups of people from living there. Not only does real estate have economic impact but also social and societal impact. We can look into gentrification and how property sales prices rise after a certain economic group enters the neighborhood. We built a base sales prediction model that helps users develop a better understanding of their returns on investment and what properties yield the highest sales price. However, as mentioned before, there are many more other important factors than the ones that we’ve included in our model that result in sales pricing. It’s important to note that the model should continuously improve and adjust as the real estate market goes through its cycles.

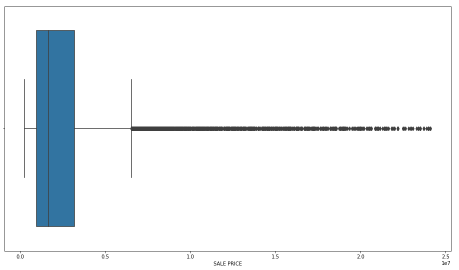
As the pandemic subsides bringing more people back to Manhattan, rental inventory in Manhattan is above its five and ten year averages. The market is still below where it was pre-pandemic so there are still good deals available. Not only have rental prices gone below pre-pandemic levels, but landlords are offering a couple months of free rent as concessions. The median selling price in Manhattan is $1.13 million in the second quarter of 2021. Brokers closed almost 3,000 sales in Manhattan the second quarter (Paynter). There is an especially high demand for luxury real estate. In Q2 of 2021, NYC recorded 74 transactions of 118 properties for a total of $1.24 billion, with multifamily properties comprising more than $20 million (Shkury). The real estate market is booming right now with many people looking at properties and not enough properties being put into the market. Without a doubt, APT212 will do their best to provide the best for their clientele and those new to the real estate market looking to learn more.

**9. Appendix**

**Figure 1: Box Plot Against Sale Price**

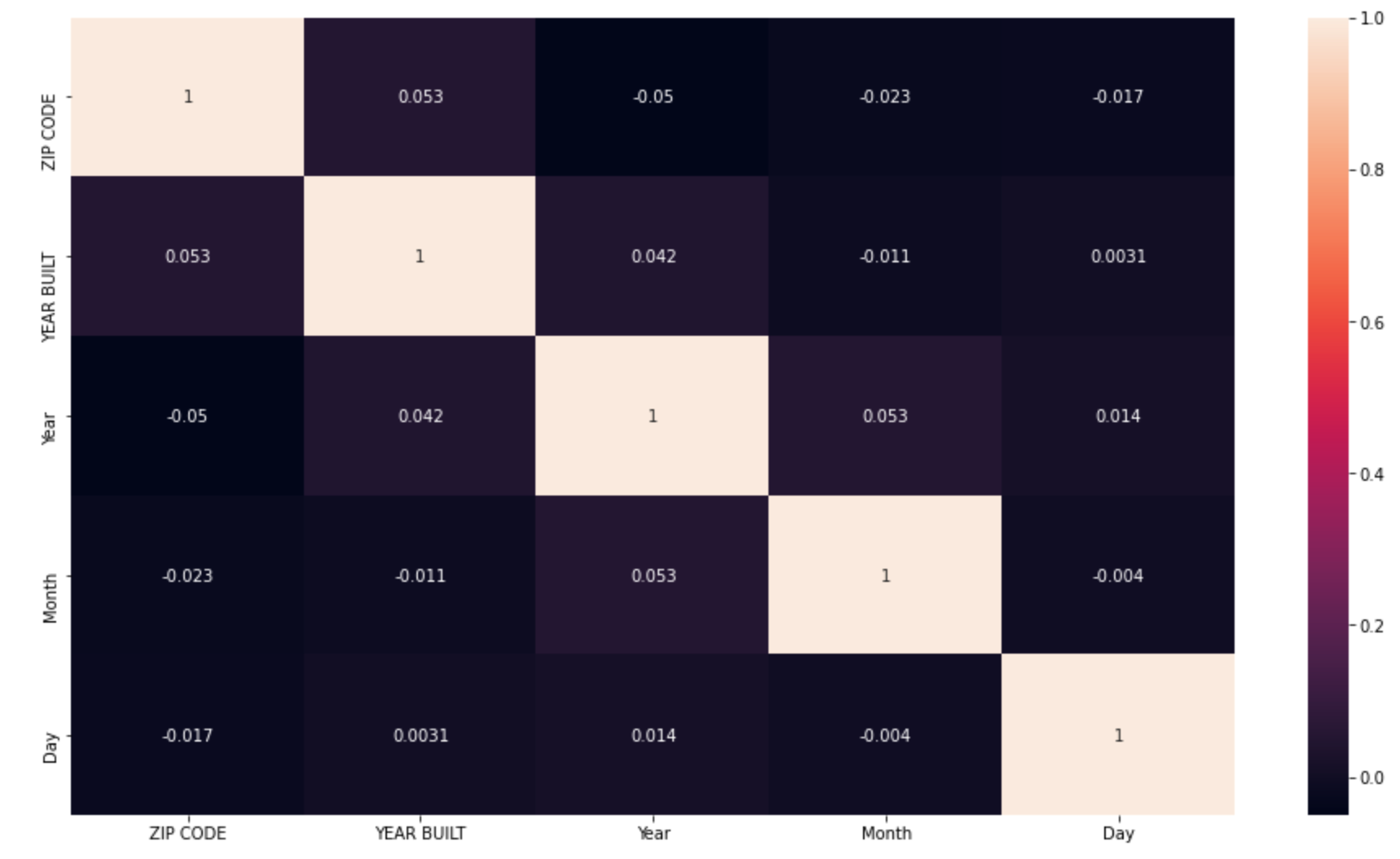
****

**(Before)**

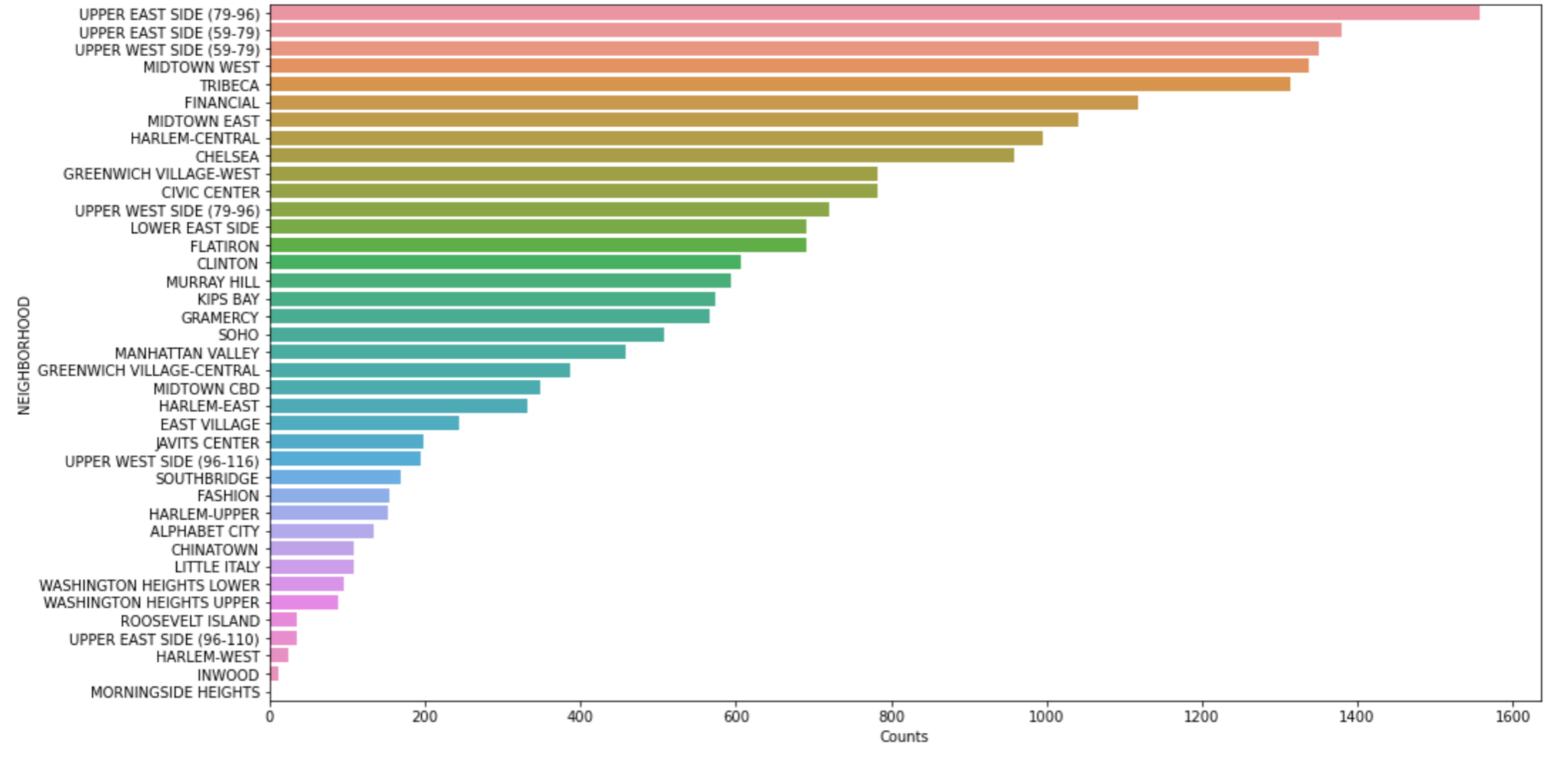
****

**(After)**

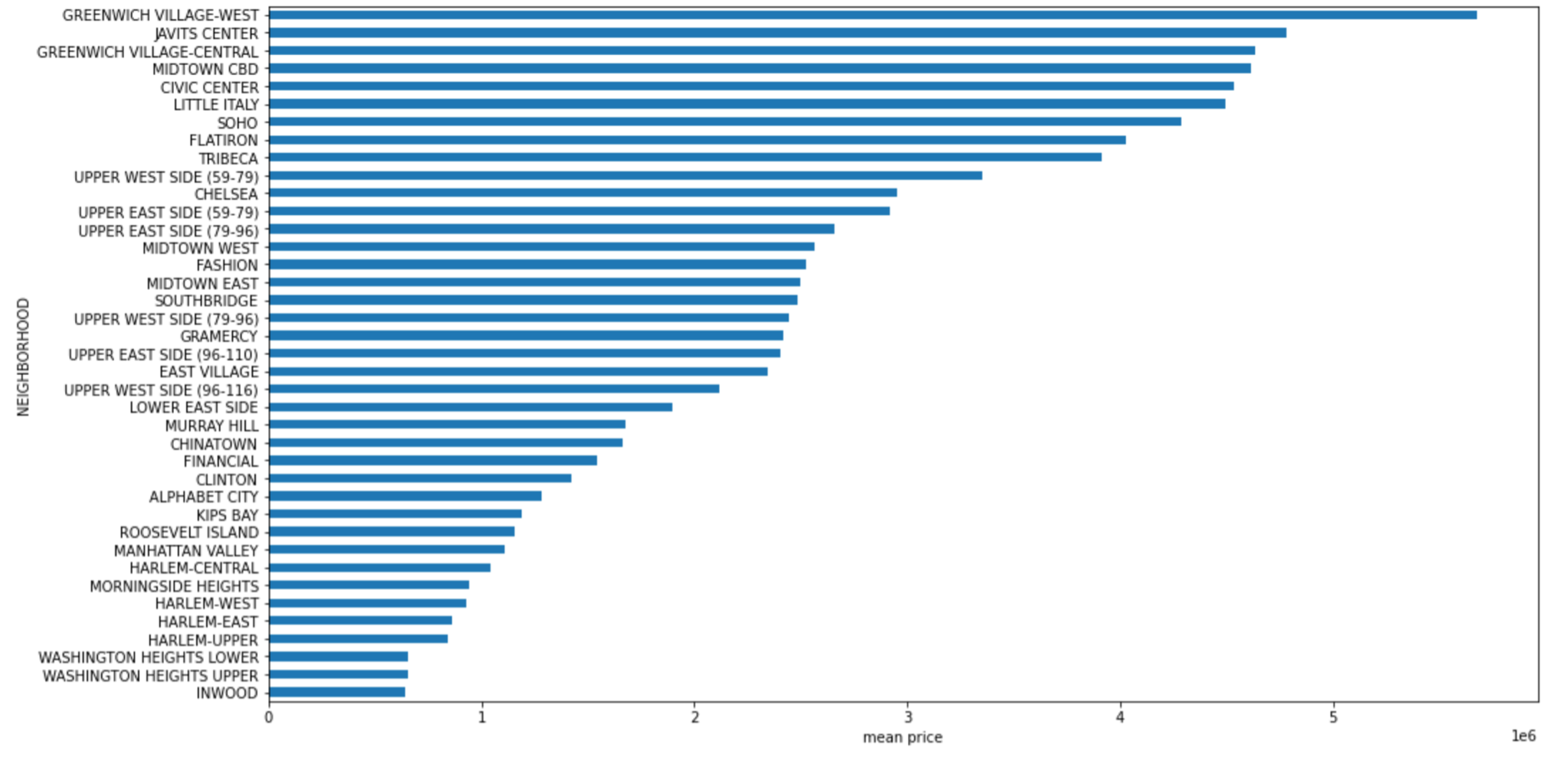
**Figure 2: Correlation Matrix Heatmap**

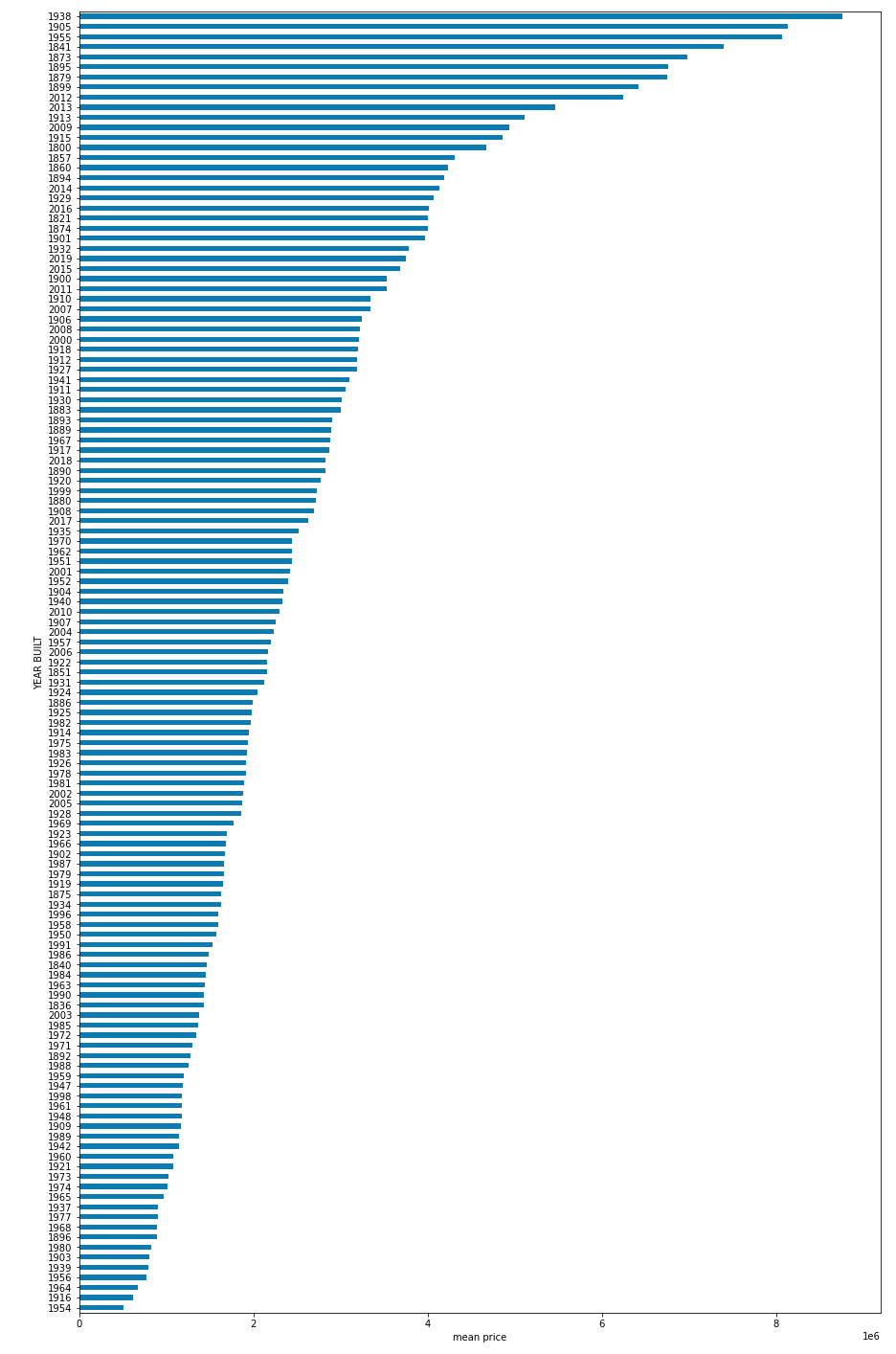


**Figure 3: Number of Sales Per Neighborhood**

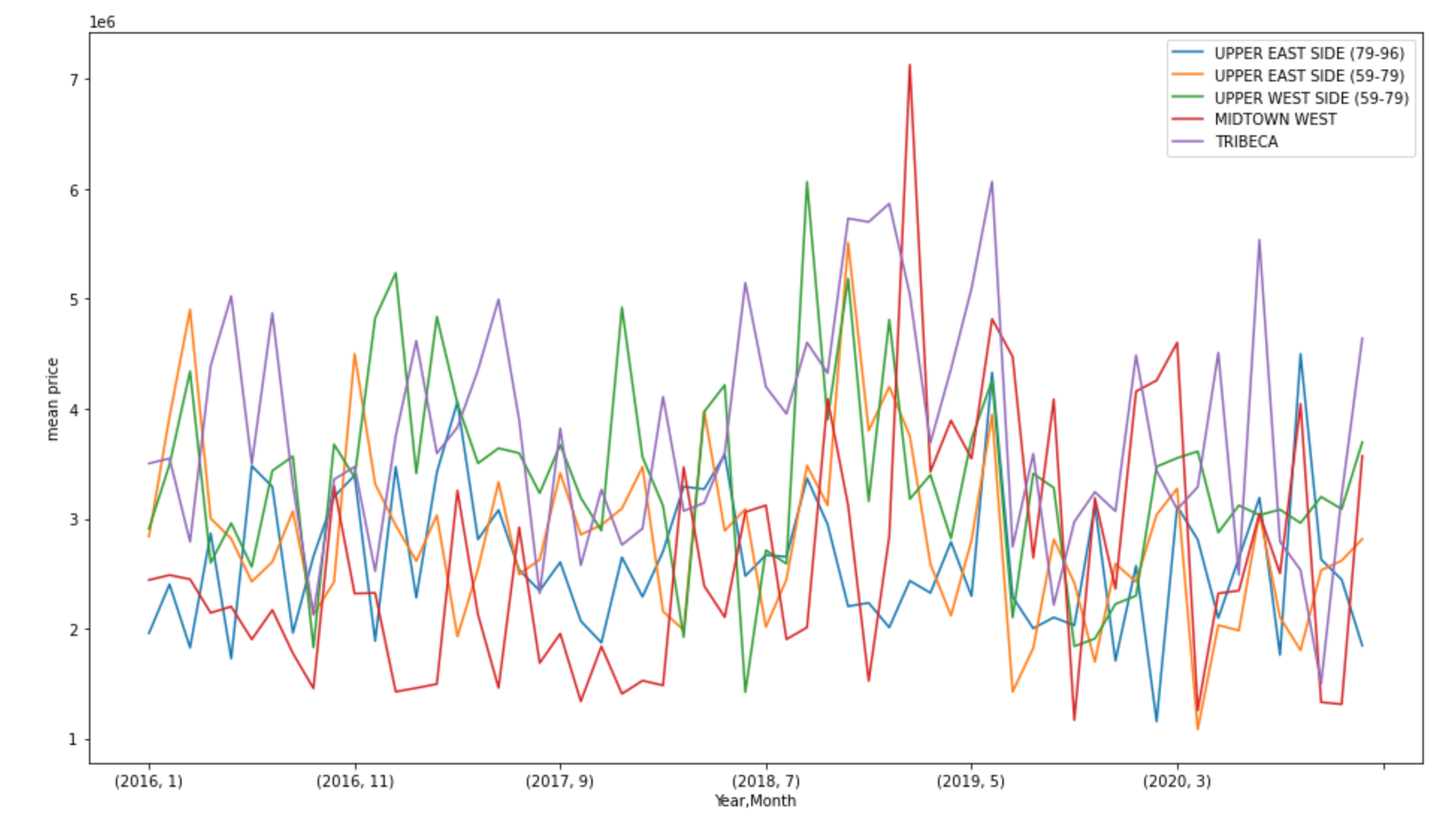


**Figure 4: Mean Sales Price for Each Neighborhood**

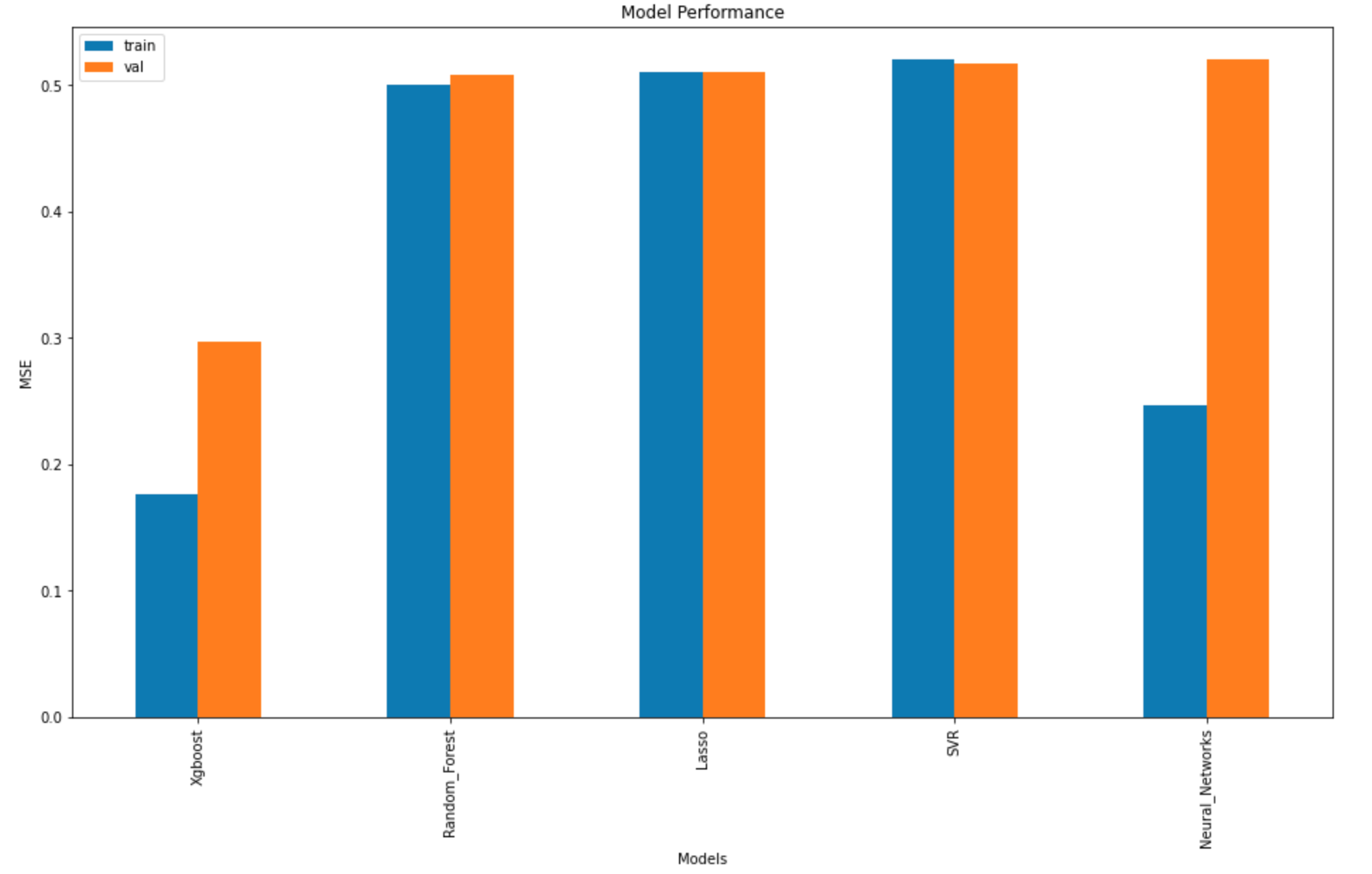


**Figure 5: Mean Sales Price** **Based on Building Year Built**

**Figure 6: Top 5 Neighborhood’s Mean Sales Price**



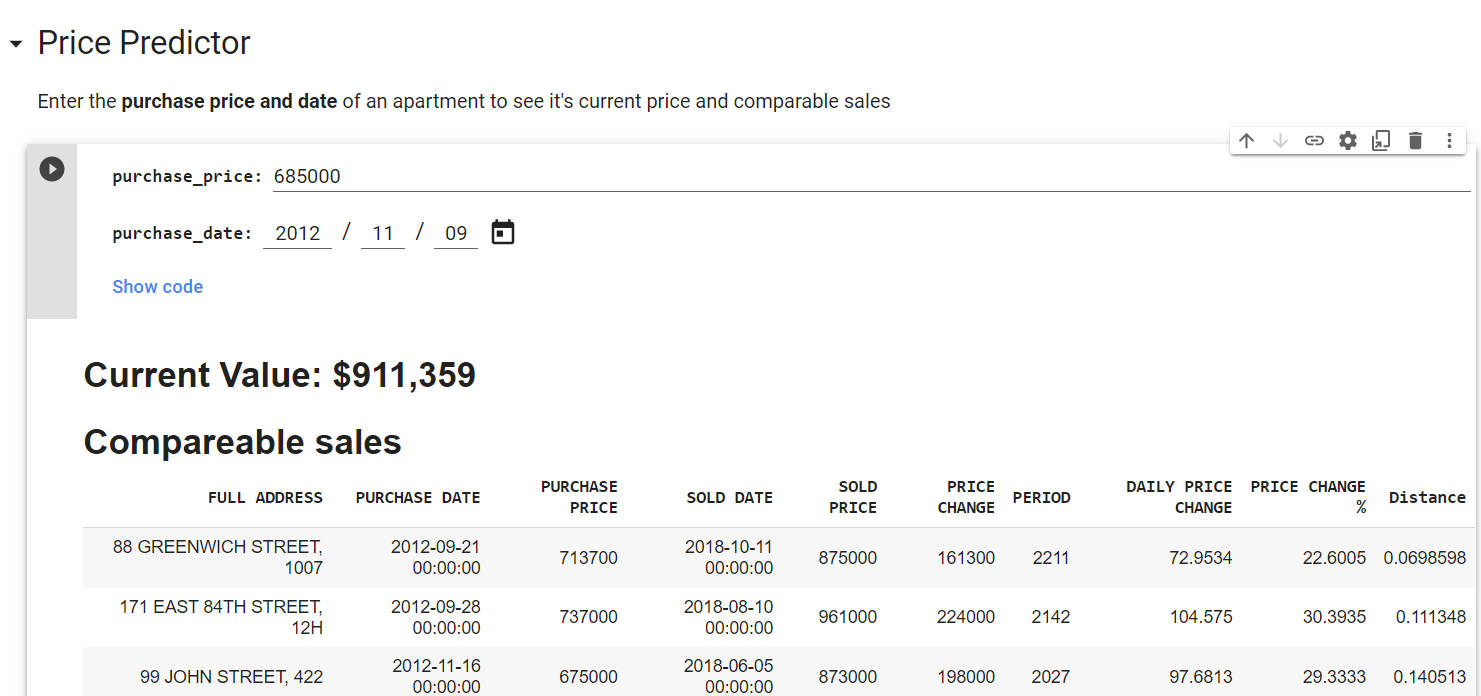
**Figure 7: Model Performance Based on Training & Validation MSE**



**Figure 8: MSE Chart**

| Mean Squared Error | **Training** | **Validation** |
| --- | --- | --- |
| **XGBoost** | 0.176455 | 0.296757 |
| **Random Forest** | 0.500639 | 0.508324 |
| **Lasso** | 0.509807 | 0.510199 |
| **SVR** | 0.520197 | 0.516380 |
| **Neural Networks** | 0.246855 | 0.519644 |

**Figure 9: Price Predictor Example**

****

**Citations**

Asaftei, Gabriel Morgan, et al. “Getting Ahead of the Market: How Big Data Is Transforming Real Estate.” *McKinsey & Company*, McKinsey & Company, 30 Mar. 2021, www.mckinsey.com/industries/real-estate/our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate.

Barr, Daniel. “Real Estate Valuation Using Regression Analysis – a Tutorial.” *Toptal Finance Blog*, Toptal, 28 June 2018, www.toptal.com/finance/real-estate/real-estate-valuation.

Jermain, Nate. “Home Value Prediction.” *Medium*, Towards Data Science, 5 Apr. 2019, towardsdatascience.com/home-value-prediction-2de1c293853c.

Nguyen, Joseph. “4 Key Factors That Drive the Real Estate Market.” *Investopedia*, Investopedia, 28 July 2021, www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp.

Paynter, Sarah. “Here's Where NYC's Real Estate Market Stands Right Now.” *New York Post*, New York Post, 20 July 2021, nypost.com/article/nyc-real-estate-market-housing-prices/.

Shkury, Shimon. “Q2 2021 Numbers Affirm New York City's MULTIFAMILY COMEBACK.” *Forbes*, Forbes Magazine, 6 Aug. 2021, www.forbes.com/sites/shimonshkury/2021/08/06/q2-2021-numbers-affirm-new-york-citys-multifamily-comeback/?sh=4ff333d52144.