Final Report

1. **Problem Statement:** The broad idea behind this capstone project is to demonstrate the possibility of leveraging machine learning to predict the property prices and help investors, developers and, everyone who is seeking to invest in the property for both professional and personal purposes. Or to just simply answer the question "What's my House Worth"? More Specific business problem statement is

Can we use machine learning specifically Regression techniques to predict the sale price of the properties? Which factors are responsible for the maximum variation in the sale price? Can we make correct predictions with least amount of error?

The goals of this capstone project were to determine the factors that can help in the estimation of the property prices and use regression techniques to correctly estimate it with least amount of errors without any overfitting issues.

2. **Background:** While doing the research on this problem, I have come across the different methodologies by which these kind of problems were tackled, The common thing that was lacking in these methods were the lack of data cleaning and EDA and, categorical variables were not given importance. The regression techniques used were not regularised. These shortcomings in previous approaches motivated me to tackle this problem in best possible way.
3. **Details on the Data Source:**

Originally the dataset is from https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page which is in form of 5 different tables for each of the borough.

I have downloaded it from the Kaggle website- https://www.kaggle.com/new-york-city/nyc-property-sales

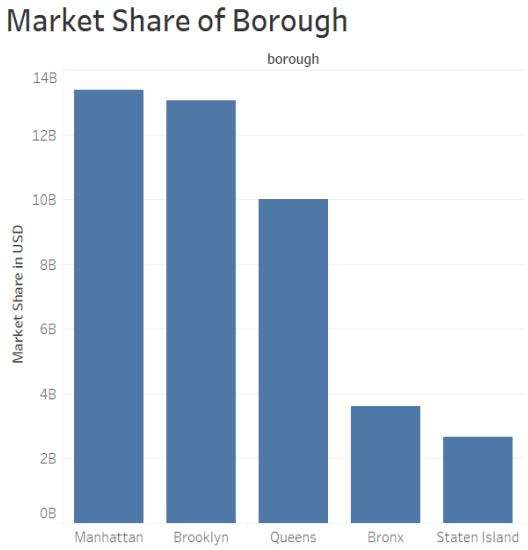
The data set is about the properties sold in New York City over a 12-month period from September 2016 to September 2017. The dataset is in form of a table that includes around 84548 rows and 22 columns originally. GLOSSARY of the columns is taken from the website below

https://www1.nyc.gov/assets/finance/downloads/pdf/07pdf/glossary_rsf071607.pdf

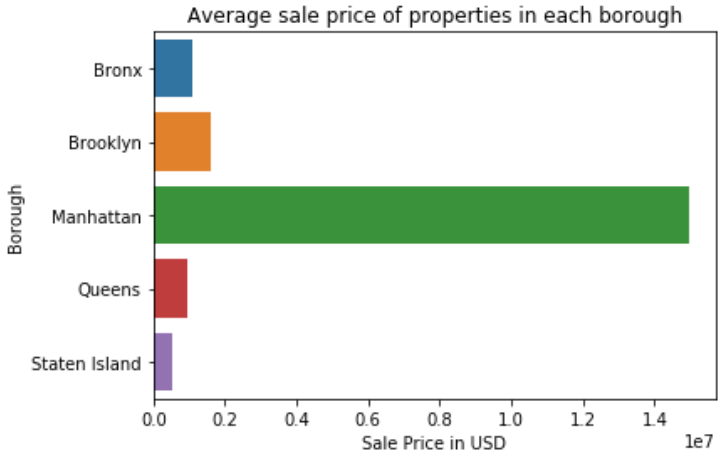
4. **Data preprocessing, feature engineering, data cleaning/transformation and EDA**
    - The original shape of the dataset is 84548 rows and 22 columns and final shape is 28040 rows and 31 rows.
    - This whole process includes conversion of the data types of columns into desired types, for example converting the **borough** into categorical type and also numerical values were converted to corresponding name of the Borough e.g. 1 into Manhattan, **sale date** into datetime, keeping **block** and **lot** into the numerical types as there are whole lot of unique values corresponding to both of them.
    - The columns **apartment**, **easement** and **unnamed: 0** were dropped. The reason behind deleting these columns is that the column unnamed: 0 is just analogous to index, easement has 100% of missing values and similar situation with apartment having almost 77% of missing values.
    - From the column **sale date, sale year and sale month** were extracted and then using **year built** the **age of building** was calculated as new column.
    - The columns **land square feet** and **gross square feet** are the ones with high correlation with **residential units**, so this variable is used to fill Nan values in **land square feet**, **gross square feet,** and **sale.** The mean value for each of these columns using group by function on the residential units' column is calculated and put in dictionary and then filled in corresponding columns using mean values. However, we could have filled these values with median as well, though not making much of the difference, only slightly changing standard deviations.
    - Target Variable - **sale price** In the dataset few properties have sale price = 0 USD, the reason of which was also explained in the Glossary that these are the properties with their ownership transferred from parents to children. Also, I have decided to keep the rows which have sale price greater than USD 50000 and less than USD 500,000,000. The reason behind keeping only these values is that sale price less than USD 50,000 either could be replaced or dropped. Similarly, there are only few values in sale price column which is greater than 500,000,000 USD e.g. The most

expensive property in the original dataset is around 2 Billion USD. These few values will act as outliers and can affect the models employed on the data and could alter the results i.e. predictions made on the target variable.

- Only those properties are kept which have total units = residential units + commercial units.
- Zero values in **sale price, land square feet** and **gross square feet** are dropped and only those properties that were built after 1650 are kept.
- The columns were made to go under log transformations to make their distribution normal, which was not the case without the transformation.
- **Manhattan** and **Brooklyn** boroughs make most of the market share

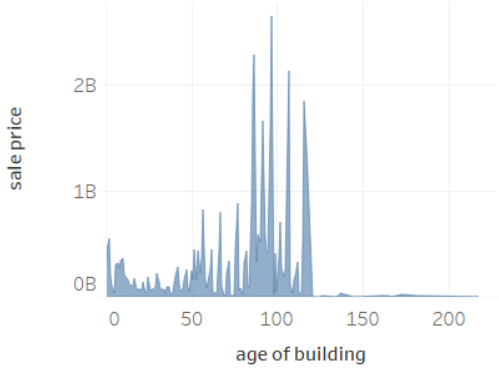


- The average price of a property is highest in Manhattan making it most expensive place to live in New York

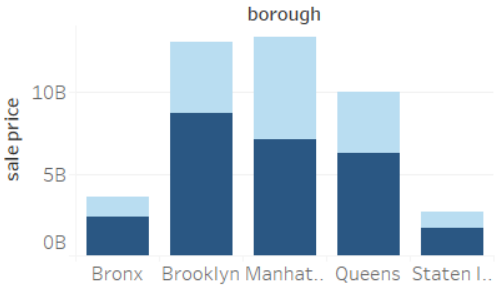


- Other Visualisations

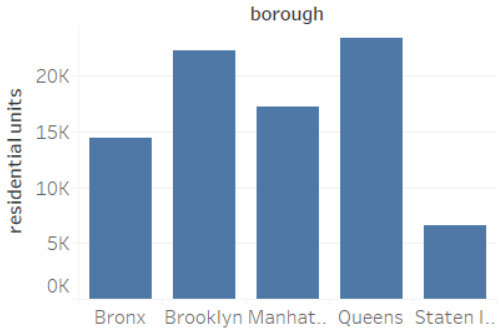
Effect of Age of the building on Sale price of the properties



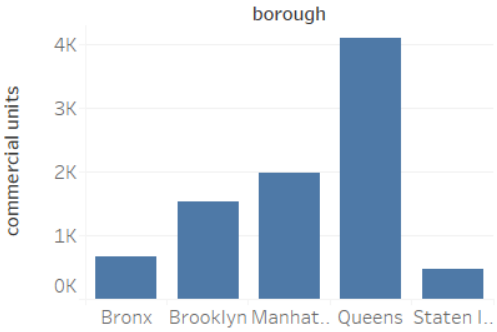
Comparison of Sales in 2016 and 2017 in each Borough(Brooklyn has highest property sales in 2017)



Residential Units in Boroughs(Queens has maximum residential units)



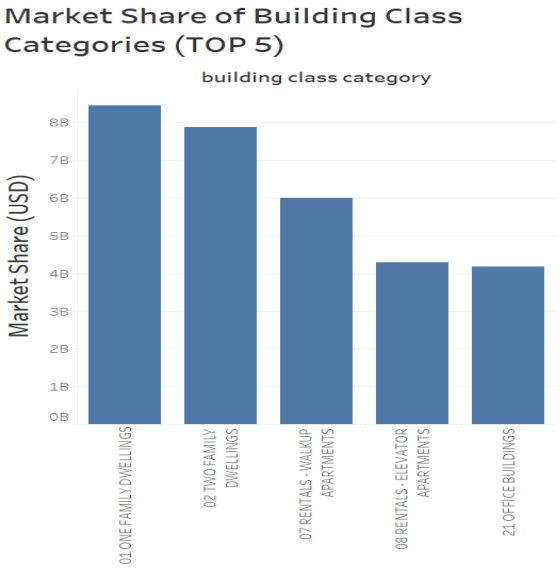
Commerical Units in Borough(Queens has highest number of commercial units)



The most expensive neighbourhood in Borough

Borough	**Neighbourhood**
Bronx	Melrouse/Concourse
Brooklyn	Bedford Stuyvesant
Manhattan	Financial
Queens	Flushing-North
Staten Island	Great Kills

- In terms of **building class category, ONE FAMILY DWELLINGS** make most of the market share



- One Hot Encoding was done on **borough** and **tax class at present** to include them into the modelling.

5. Firstly, feature importance and correlation matrix were used to find the important features in the dataset and **land square feet** and **gross square feet** came out to be most predictive of the target variable. This was before the inclusion of categorical features.

Then, Ordinary Least Square Method was used with the backward elimination method to remove the variables that has p-value has than 0.05 and see the Beta values and the variations they bring to the target variable. The model was evaluated by plotting between the residuals and fitted values and no pattern was found in the plot. The data was split into the training(75%) and test set(25%) and training set was scaled .After this regression techniques like Linear Regression, Random forest Regression, Decision Tree, ridge and lasso were fitted on the training set and r-square , mean absolute error, mean squared error, mean percentage error were calculated and compared and models are ranked as below

Case 1: With only Numerical Variables

Rank	Model	R-Square	Mean Absolute Error	Mean Squared Error	Mean Percentage Error
1	Random Forest Regressor	0.4946	0.43	0.3876	3.1963
2	Ridge Regression (α = 120)	0.4553	0.4544	0.4175	3.3694
3	Linear Regression	0.4553	0.4547	0.4177	3.372
4	Ridge Regression (α = 1)	0.4553	0.4547	0.4177	3.372
5	Lasso Regression (α = 0)	0.4553	0.4547	0.4177	3.372
6	Decision Tree Regression	0.071	0.58	0.7124	4.2947
7	Lasso Regression (α = 1)	-9.11E-05	0.6036	0.7669	4.3861

After these Categorical variables were included using one hot encoding, the beta values corresponding to the variables are

```
const                    5.403370
land square feet         0.130391
gross square feet        0.595826
total units             -0.142175
residential units        0.101370
commercial units         0.183307
age of building         -0.005821
tax class at present4    1.921065
tax class at present1    1.645775
tax class at present2    1.836530
boroughBronx             0.559607
boroughBrooklyn          1.149779
boroughManhattan         2.102531
boroughQueens            0.920887
boroughStaten Island     0.670567
```

Case 2: Numerical Variables with Categorical Variables like borough, tax class at present

Rank	Model	R-Square	Mean Absolute Error	Mean Squared Error	Mean Percentage Error
1	Random Forest Regressor	0.5721	0.3869	0.3281	2.8867
2	Ridge Regression (α = 80)	-9.11E-05	0.4039	0.3484	3.0123
3	Linear Regression	0.5451	0.4040	0.3489	3.0135
3	Lasso Regression (α = 0)	0.5451	0.4040	0.3489	3.0135
4	Random forest Regressor (n_estimators = 201, max_depth = 1)	0.3516	0.51190	0.4972	3.7737
5	Decision Tree Regression	0.2148	0.5147	0.6022	3.2889

6. **Findings and Conclusions:** In both the cases discussed above, Random forest Regressor works out to be the best model in the prediction of the sale price with only 2.89 % in Case 2. With the inclusion of more information about the dataset, we can see that the predictions have become better. When the Categorical variables like borough was added, the Beta value for the Borough Manhattan was maximum which was maximum for gross square feet before without categorical features. Thus, we can conclude that Borough in which the property is located is an important factor in the prediction of the sale price. Also, tax class of the property is an important factor as well, out shadowing other factors as beta values corresponding to them were also higher explaining that they are responsible for more variation in the sale price.

   These results are in line with the initial goals of the capstone project which was to correctly predict the sale price of properties with least amount of error and

find the factors responsible for the maximum variation in the target variable i.e. sale price.

7. **Summary and Future Directions:** Summing up, Inclusion of categorical variables have improved the model score and decreased the errors in predictions made for the sale price, In terms of modelling for Random Forest model, the regularisation has increased the error. Linear regression is working better with categorical variables, so we can include more categorical features like **building class at present** and perform the iterations. Also, neural networks can be employed, and code can be made more efficient. The way of calculation for errors can be improved.