



# Housing Prices New York City

---

By  
Pizon Shetu



---

# Problem Statement

---

- How do we engage the New York City housing market?
- How should Capital Fortune price their homes?

# Predicting House Prices

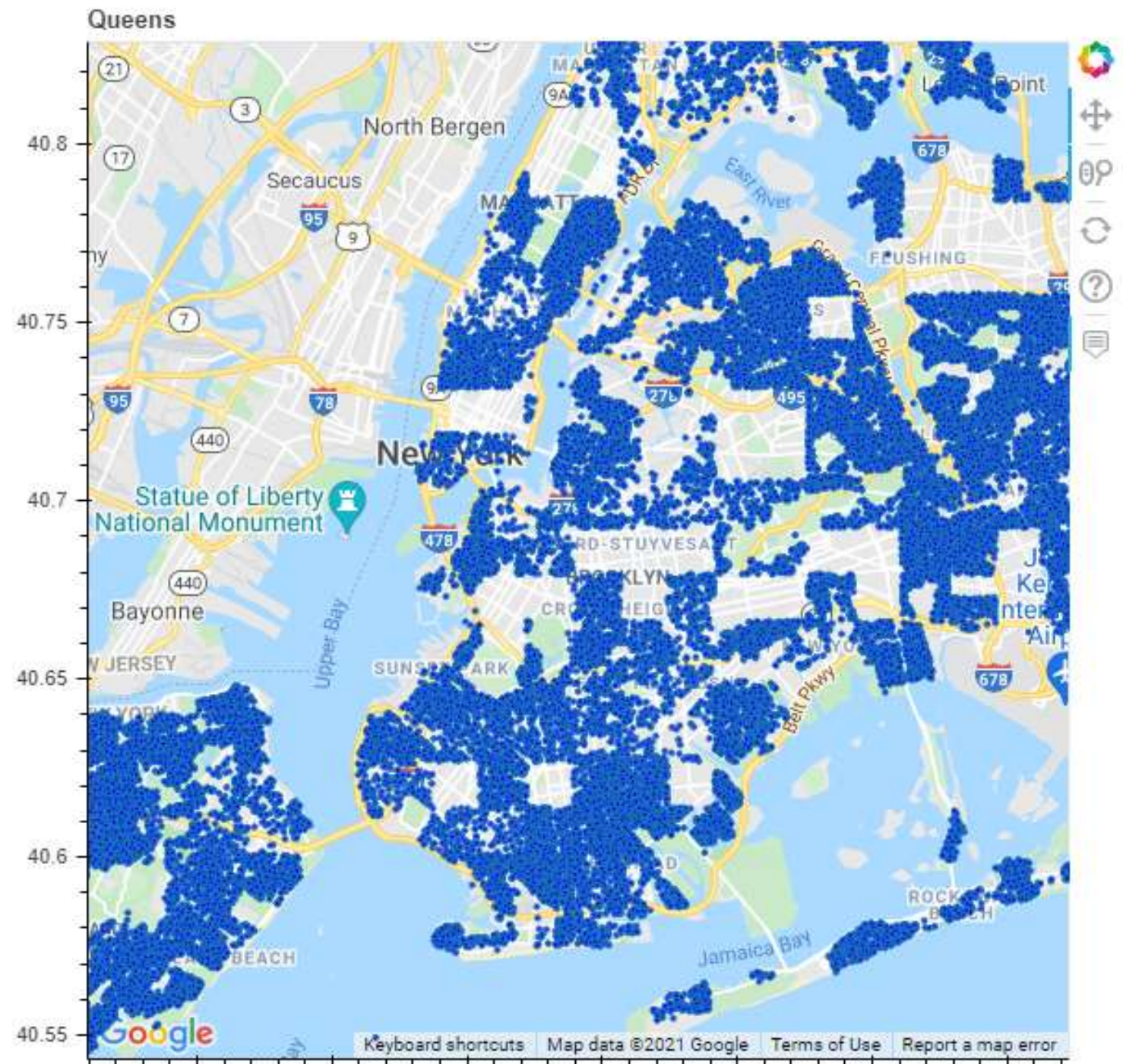
---

- How machine learning can help predict prices of homes in New York City
- Zillow's NYC Dataset





# Homes in dataset



# Data Wrangling

---

75,000 observations with over 1500 features

---

Shrunk to 65,000 homes and 24 features

---

Main variable – Price

---

Dropping of redundant, useless features and rows with missing values

---

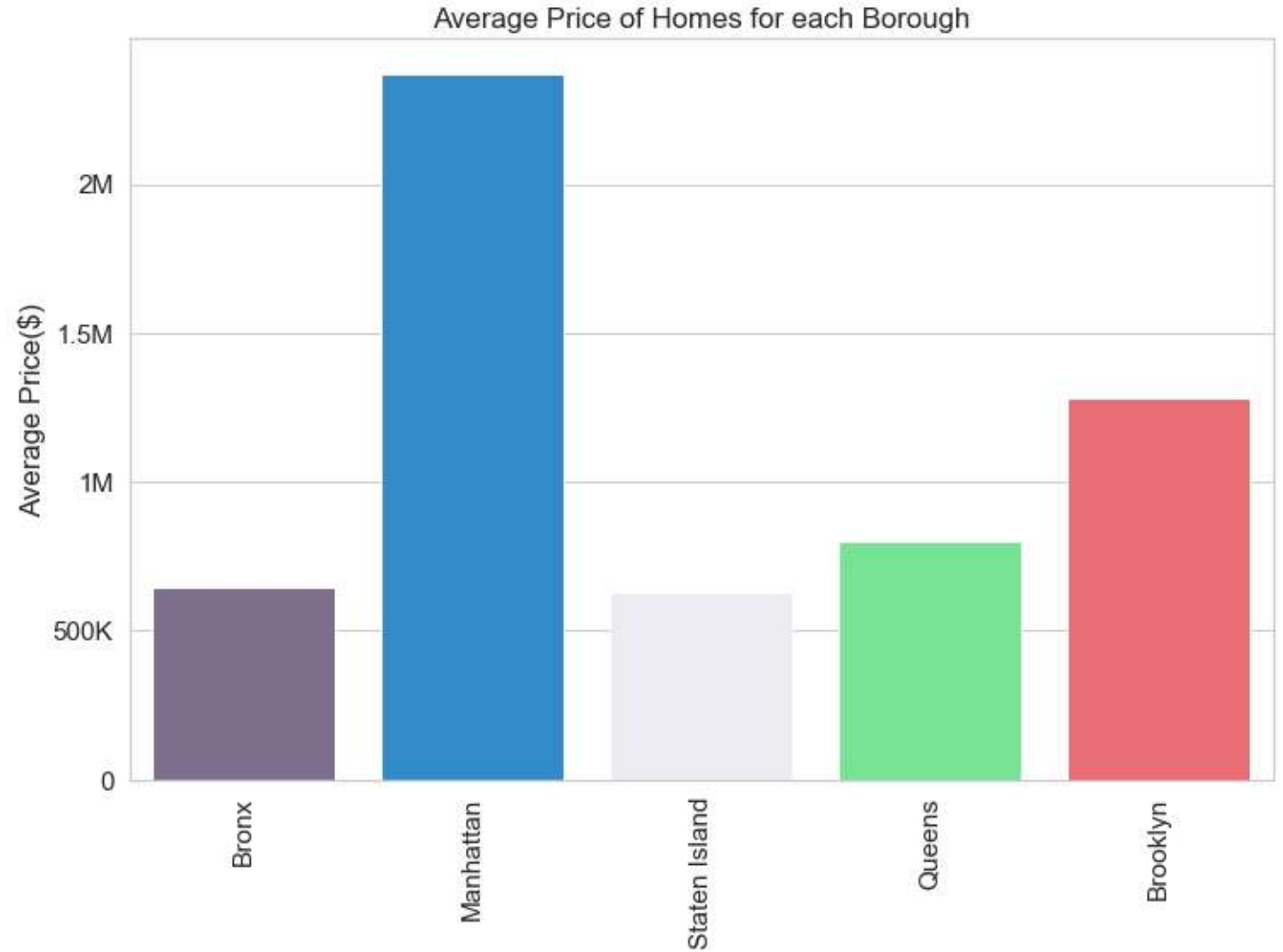
# Imputations - MICE

---

- Imputation method utilizing Gradient Regression on missing values
- Estimates missing value of one feature by learning from other features and looking at past values given data from other features.
- 6 Step process
- Greater results than mean or median imputation.

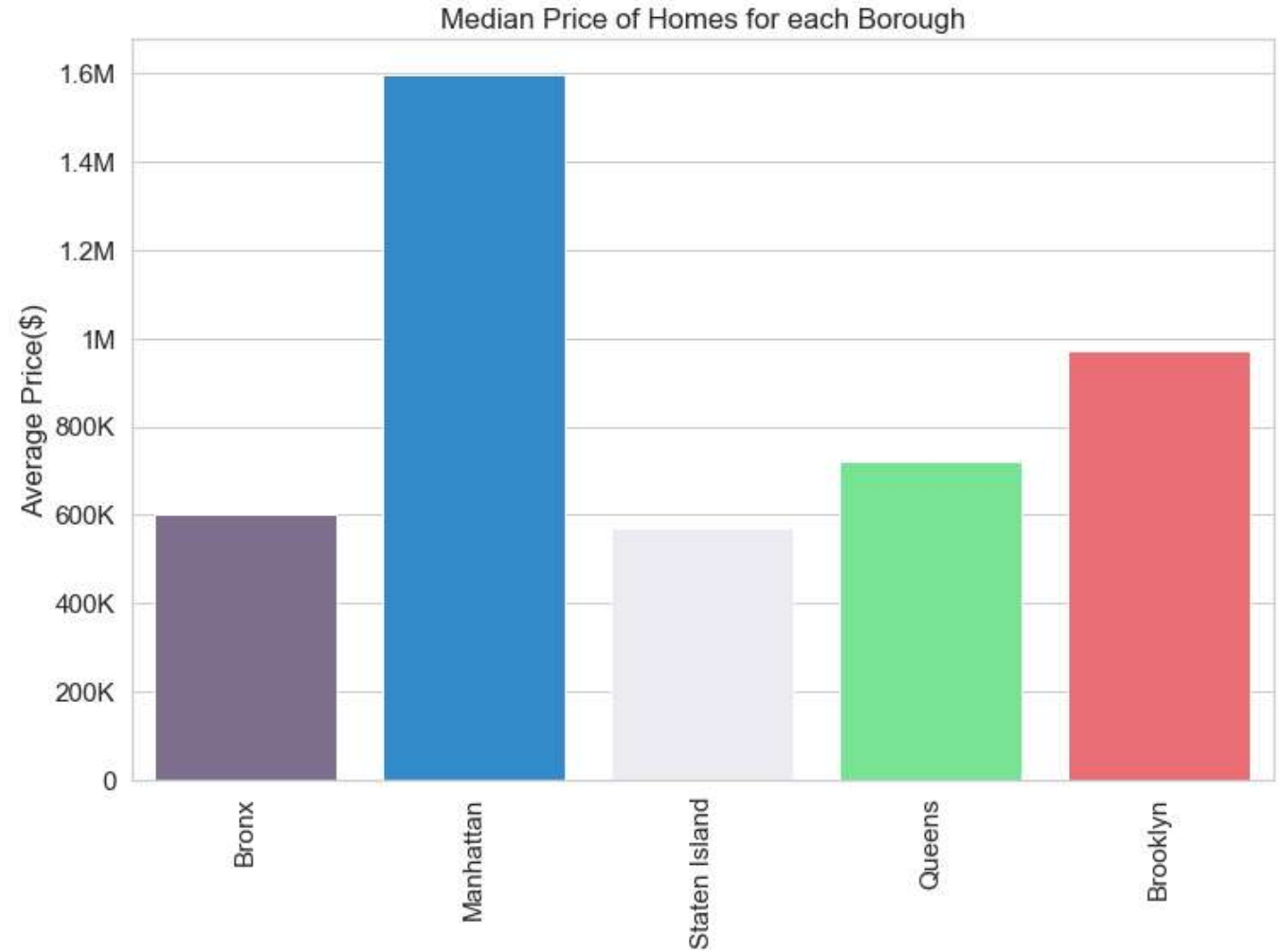
# Location Impact

- Manhattan has the highest average prices for homes



# Median Price of Homes

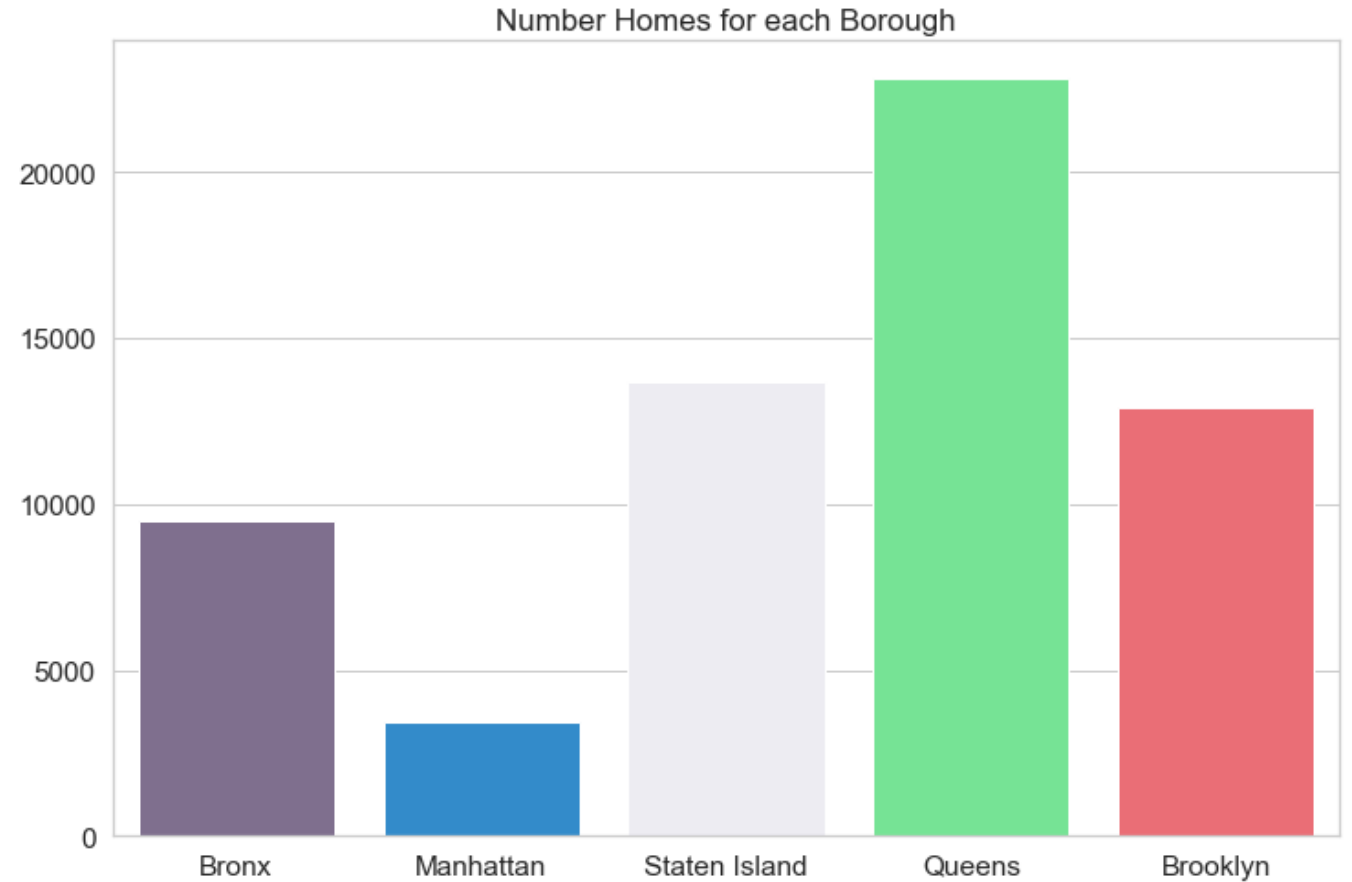
Average Price of homes can be skewed by outliers, Median shows how much most houses are sold for.





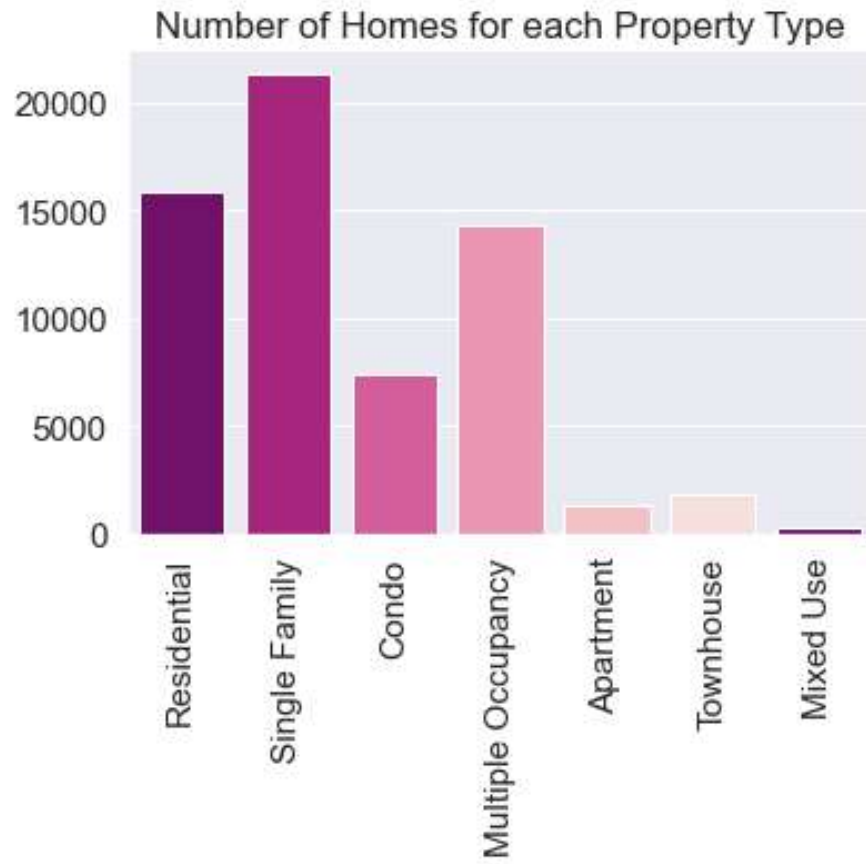
# Location count

- Manhattan barely having many homes for sales compared to other boroughs
- Queens leading in home listings



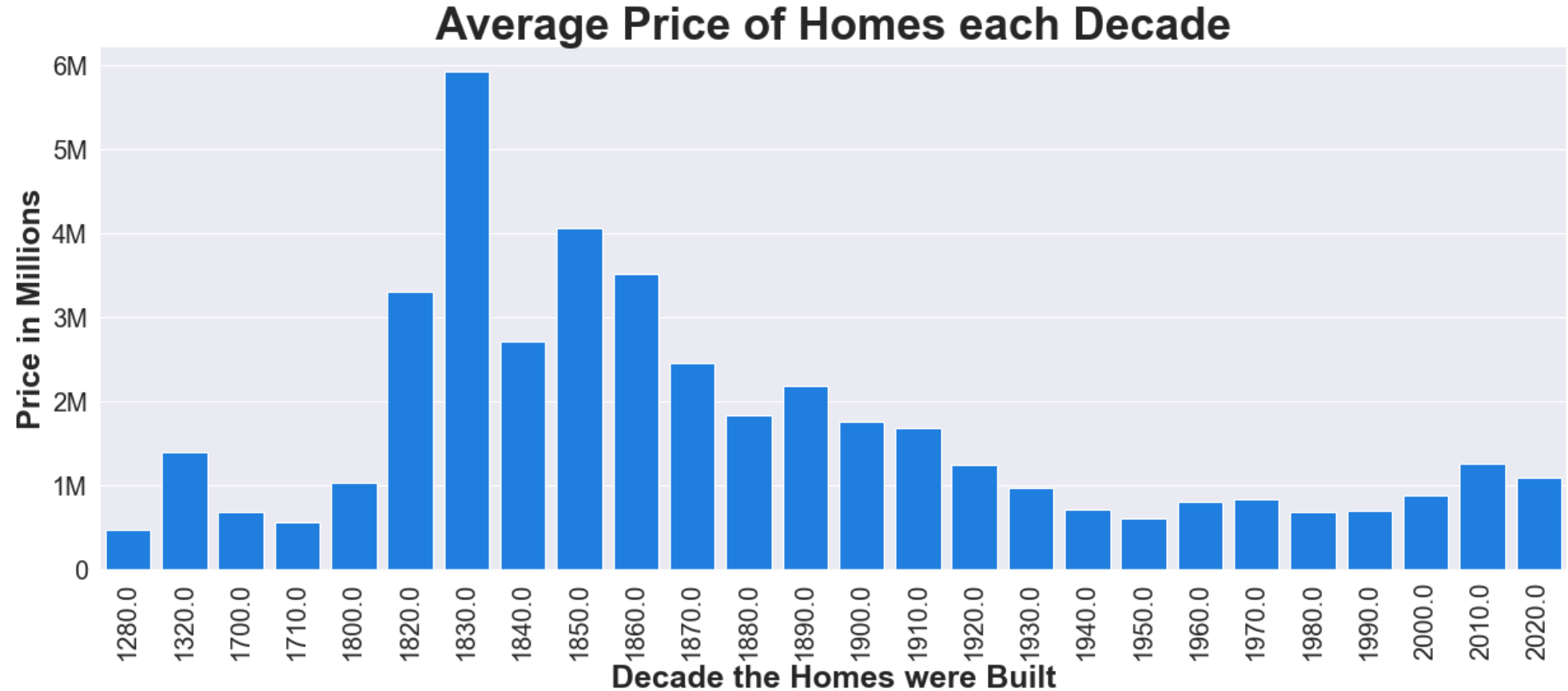
# Property Type

- Due to low sample size data for Townhouse and Mixed Use properties, we can have skewed data, as few homes might not give us the whole picture for prices for this property type.

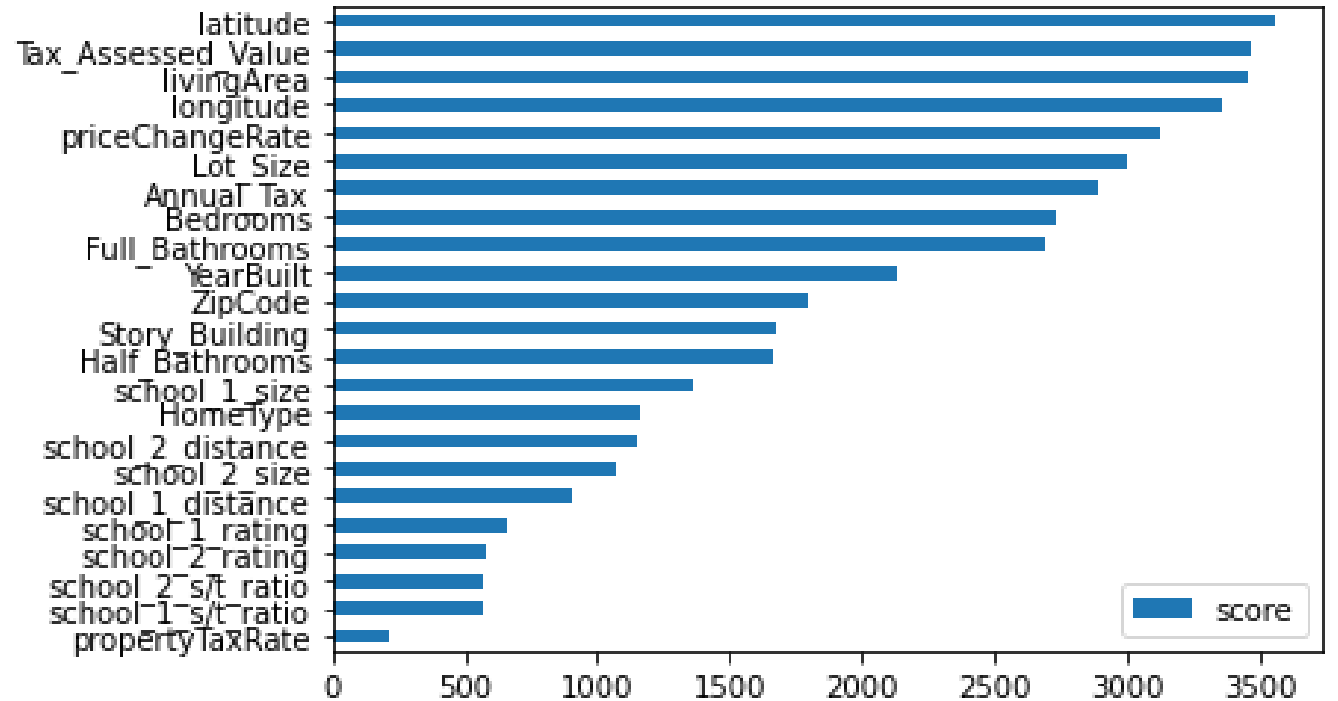


# Impact of Built Year

- 70 % of homes in early 1800's to 1900's were in Manhattan or Brooklyn
- Recent Homes see an uptick with over 50% of homes in lower priced boroughs



# Most Impactful Features





---

MAE – Mean Absolute Error

## **Metrics for evaluation**

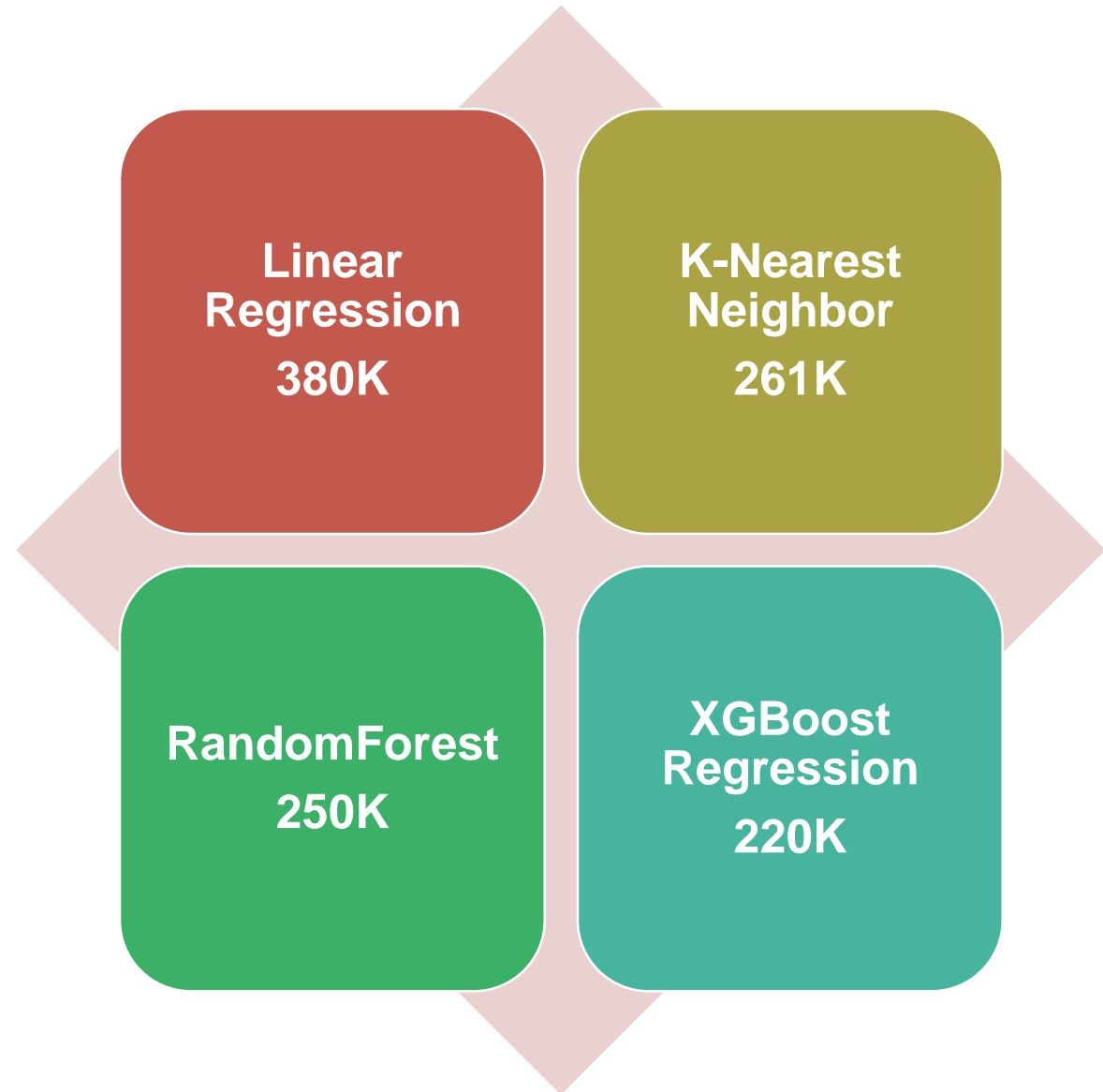
---

On average how far is the prediction from the actual

---

A decorative horizontal line at the bottom of the slide, consisting of a short brown segment on the left and a longer light brown segment extending to the right.

# Model Selection



# Hyper-Parameter Tuning

---



MAX\_DEPTH,  
MIN\_CHILD\_WEIGHT, ETA,  
SUBSAMPLE, AND COL\_SAMPLE



REDUCTION OF NEARLY 30K IN  
MAE ON TESTING DATA AND 65.7K  
IN TRAINING DATA.

# Takeaways

---

- **Best results achieved via XGBoost**
- **Location is the largest driving force of price**
- **Multi-family homes and Condominium have greater returns**
- **Townhouses yield a high price but due to lack of data it is inconclusive**
- **Newly built homes garner larger premium over old and remodeled homes**