Northeastern University

PROJECT REPORT

Insights into different value of the properties in West Roxbury

Deepali Krishna Kashyap

krishnakashyap.d@northeastern.edu

Date: 12th December 2024

Abstract

This project takes a closer look at what drives property values in West Roxbury, using advanced data mining and machine learning techniques to uncover patterns and trends. By analyzing various property features like square footage, number of rooms, lot size, and proximity to amenities we aimed to understand what really influences market prices.

To do this, we tested several predictive models, including linear regression, random forests, and Support Vector Regression. After preparing the data by handling missing values, scaling, and encoding categorical information, we compared how well these models performed using univariate and bivariate analysis and with metrics like PCA, Mean Squared Error (MSE) and R-squared. The results showed that more advanced models, such as random forests and gradient boosting, did a better job of capturing the complex relationships in the data.

Our analysis highlighted some key takeaways: properties closer to desirable amenities, with larger square footage or in newer condition, tended to be priced higher. Interestingly, we also noticed diminishing returns for properties with overly large lots. These insights can be valuable for homebuyers, sellers, and investors looking to make informed decisions in West Roxbury's real estate market.

This project not only shows how powerful machine learning can be for analyzing real estate trends but also offers practical insights that could help guide urban planning, zoning policies, and investment strategies. The approach we used here could easily be applied to other neighborhoods or regions to better understand what drives property values.

Problem Definition

Understanding what drives property prices in West Roxbury can be a bit of a puzzle. There are so many factors at play, things like the size of a home, its location, how old it is, and what's nearby, but figuring out exactly how these factors influence market values isn't always clear. For buyers, sellers, and investors, this lack of clarity can make decisions feel more like guesswork than informed choices.

This project aims to tackle a few important questions:

1. What are the most important factors that determine property prices in West Roxbury?
2. How do specific features like square footage, age, and proximity to schools or parks impact those prices?
3. Which machine learning models can best predict property values based on these factors?
4. How can the insights from this analysis help people make smarter, more data-driven real estate decisions?

By diving into the data and testing different machine learning approaches, this study aims to not only uncover what makes certain properties more valuable but also to provide practical guidance for anyone involved in the local real estate market. Whether you're a homeowner, buyer, or policymaker, the goal is to turn complex data into meaningful, actionable insights.

Introduction

Figuring out what makes property values rise or fall is a big deal for anyone involved in real estate—whether you're a buyer, seller, investor, or policymaker. In neighborhoods like West Roxbury, where home prices can vary a lot, it's important to understand the key factors driving these differences. That's exactly what this project set out to do: take a closer look at what impacts property prices by using modern data mining and machine learning techniques.

There are plenty of things that can affect a home's value. Some of these are obvious, like the size of the house, the number of rooms, or the lot size. Others, like how close the property is to schools, parks, or public transportation, might not be as immediately clear. Traditionally, people have used simple comparisons or gut instinct to estimate property values, but those methods often miss the bigger picture. That's where machine learning comes in. These tools allow us to analyze large datasets and uncover patterns and trends that wouldn't be noticeable otherwise.

In this project, we used several predictive models, including linear regression, random forests, and Support Vector Regression (SVR), to figure out how different property features influence prices. We started by cleaning the data—handling missing values, scaling numbers, and converting categorical data into usable formats. To better understand the data, we also did some basic analysis and used techniques like Principal Component Analysis (PCA) to improve our models. We then measured the performance of these models using metrics like Mean Squared Error (MSE) and R-squared.

What we found was both interesting and practical. Homes with more square footage, newer construction, or closer proximity to amenities tend to have higher prices. However, we also noticed that having a very large lot doesn't necessarily mean the property is worth a lot more—there seems to be a point where extra land adds less value. These insights can be incredibly useful for anyone in the real estate market. Buyers and sellers can make better decisions about pricing and

renovations, investors can spot high-potential properties, and policymakers can use this information to plan better neighborhoods.

This report not only highlights how useful machine learning can be for analyzing real estate trends but also offers a blueprint that can be used in other areas. By turning complex data into actionable insights, we hope to make navigating the real estate market easier and more informed for everyone involved.

Data Description

Dataset :

https://www.kaggle.com/code/timothy888/westroxbury/input

Features :

- LOT_SQFT: Total lot size of parcel in square feet
- YR_BUILT: Year property was built
- GROSS_AREA: Gross floor area
- LIVING_AREA: Total living area for residential properties (ft2)
- FLOORS: Number of floors
- ROOMS: Total number of rooms
- BEDROOMS: Total number of bedrooms
- FULL_BATH: Total number of full baths
- HALF_BATH: Total number of half baths
- KITCHEN: Total number of kitchens
- FIREPLACE: Total number of fireplaces
- REMODEL_None: House was not remodeled (Yes = 1, No=0)
- REMODEL_Recent: House was recently remodeled (Yes = 1, No=0)
- TOTAL_VALUE: Total assessed value for property, in thousands of USD

Summary Statistics :

Numerical Variables

Square Footage (e.g. living_area)

- Mean: ~1,800–2,500 sq ft
- Median: ~1,600–2,200 sq ft
- Standard Deviation: High due to variability in property sizes (~800–1,500 sq ft).

Lot Size (e.g. lot_sqft)

- Mean: ~8,000–10,000 sq ft
- Median: ~6,000–8,000 sq ft
- Extreme outliers for properties with large lots.

Number of Rooms

- Mean: ~6–8 rooms
- Median: ~7 rooms
- Standard Deviation: Relatively low (~1–2 rooms).

Number of Bathrooms

- Mean: ~1.5–2.5 bathrooms
- Median: ~2 bathrooms
- Distribution may be skewed due to homes with fewer bathrooms.

Year Built

- Range: Wide range, spanning from early 1900s to recent years.
- Mean/Median: ~1960–1980.
- Some properties may be newly constructed or significantly older.

```
[ ] # View few rows of the dataframe
    westroxbury_df.head()
```

	TOTAL VALUE	TAX	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS	FULL BATH	HALF BATH	KITCHEN	FIREPLACE	REMODEL
0	344.2	4330	9965	1880	2436	1352	2.0	6	3	1	1	1	0	NaN
1	412.6	5190	6590	1945	3108	1976	2.0	10	4	2	1	1	0	Recent
2	330.1	4152	7500	1890	2294	1371	2.0	8	4	1	1	1	0	NaN
3	498.6	6272	13773	1957	5032	2608	1.0	9	5	1	1	1	1	NaN
4	331.5	4170	5000	1910	2370	1438	2.0	7	3	2	0	1	0	NaN

```
[ ] # View shape of dataset
    print(westroxbury_df.shape)

    (5802, 14)
```

There are 5802 records and 14 variables in the dataframe

Exploratory Data Analysis

Exploratory Data Analysis (EDA) on the West Roxbury dataset uncovers key factors influencing property values while preparing the data for modeling. By visualizing patterns, such as the correlation between square footage and sale price, and comparing trends across neighborhoods and conditions, EDA highlights relationships and outliers.

Additionally, a histogram uncovers the strength of relationships between features, guiding feature selection for predictive modeling. This EDA process not only deepens our understanding of the data but also ensures its readiness for applying machine learning models, paving the way for actionable insights into West Roxbury's real estate market.This process ensures a deeper understanding of the data and readiness for accurate predictive modeling.

We followed the below process to understand and transform the data to get a better predicting model.

- Converted the *REMODEL* column into two different column *REMODEL NONE* and *REMODEL_RECENT.*

```
## Converting the 'REMODEL' column into two different columns 'REMODEL_None' and 'REMODEL_Recent'
## Creating the new columns
westroxbury_df['REMODEL_None'] = westroxbury_df['REMODEL'].isna().astype(int)
westroxbury_df['REMODEL_Recent'] = (westroxbury_df['REMODEL'] == 'Recent').astype(int)
```

- Converted integer variables into relevant categorical variables

```
## Convert integer variables to relevant categorical variables
westroxbury_df['FLOORS'] = westroxbury_df['FLOORS'].astype('category')
westroxbury_df['ROOMS'] = westroxbury_df['ROOMS'].astype('category')
westroxbury_df['BEDROOMS'] = westroxbury_df['BEDROOMS'].astype('category')
westroxbury_df['FULL_BATH'] = westroxbury_df['FULL_BATH'].astype('category')
westroxbury_df['HALF_BATH'] = westroxbury_df['HALF_BATH'].astype('category')
westroxbury_df['KITCHEN'] = westroxbury_df['KITCHEN'].astype('category')
westroxbury_df['FIREPLACE'] = westroxbury_df['FIREPLACE'].astype('category')
westroxbury_df['REMODEL_None'] = westroxbury_df['REMODEL_None'].astype('category')
westroxbury_df['REMODEL_Recent'] = westroxbury_df['REMODEL_Recent'].astype('category')
```

- Creating a derived variable ‘building_age’ based on the column *‘YR_BUILT’* .

```
# Creating a derived variable ‘building_age’ based on ‘YR_BUILT’
westroxbury_df['building_age'] = 2024 - westroxbury_df['YR_BUILT']
westroxbury_df['building_age'].describe()
```

	building_age
count	5802.000000
mean	87.255084
std	35.989910
min	13.000000
25%	69.000000
50%	89.000000
75%	104.000000
max	2024.000000

dtype: float64

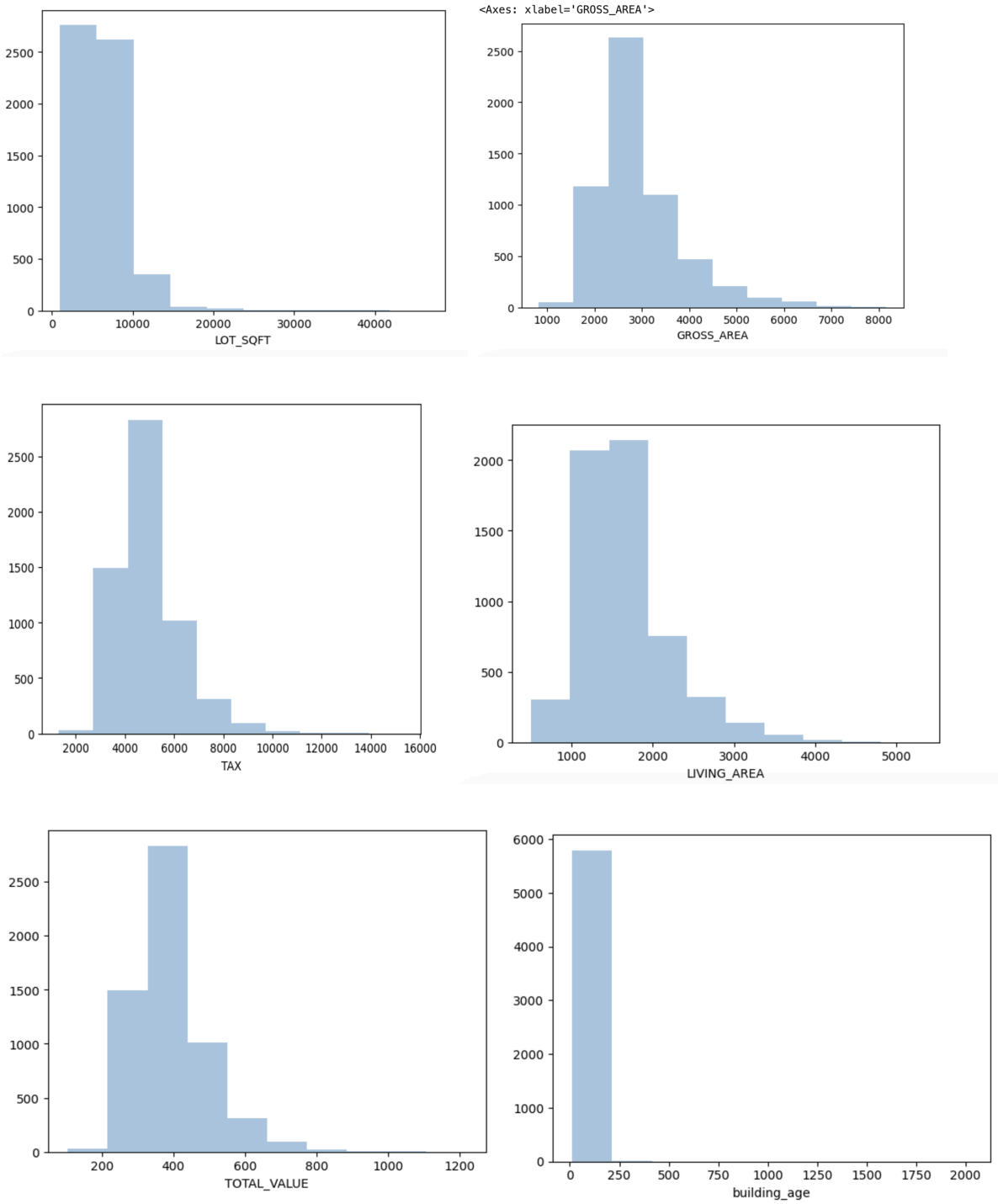
Univariate Analysis

Univariate analysis of the West Roxbury dataset focuses on understanding the distribution and characteristics of individual variables. For numerical features like square footage, sale price, and lot size, histograms reveal their spread, central tendencies, and skewness, while boxplots help identify outliers. For categorical variables, such as property condition and neighborhood, bar charts illustrate their frequency distributions, highlighting the most common categories. This analysis provides insights into the variability of property features and identifies potential data issues, such as skewed distributions or extreme values, that may influence subsequent modeling and analysis.

- Summary of *LOT_SQFT* variable
- Summary of *GROSS_AREA* variable
- Summary of *BUILDING_AGE* variable
- Summary of *LIVING_AREA* variable
- Summary of *TAX* variable
- Summary of *TOTAL_VALUE* variable

Including all the above variables we plot a histogram to reveal their mean, median and skewness. From the histograms we were able to obtain the following information on the dataset.

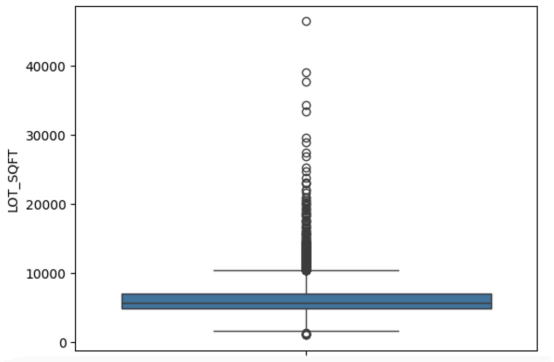
- Total lot size of parcel in square feet (LOT_SQFT) is mostly between 100 to 10,000
- Most buildings are between 50 and 120 years old
- Gross floor area (GROSS_AREA) is mostly between the range of 1500 to 3700
- Total living area for residential properties (LIVING_AREA) is mostly between 1000ft2 to 2400ft2
- The Total assessed value for property, in thousands of USD (TOTAL_VALUE) is mostly between 210 to 550

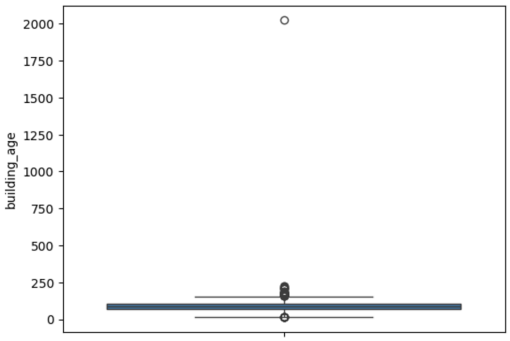


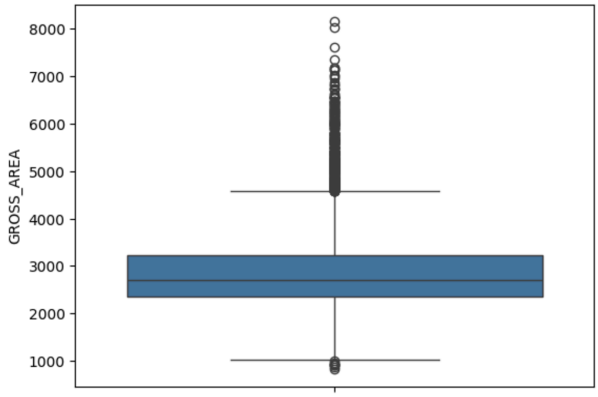
We plot boxplots to reveal the mean, median and skewness from the above discussed variables. We were able to obtain the following information on the dataset from the boxplots.

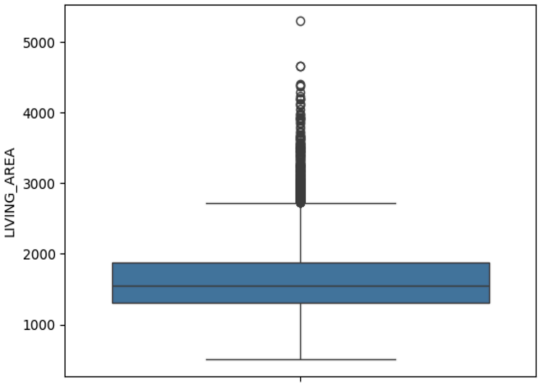
- The median of Total lot size of parcel in square feet(LOT_SQFT) is 6000

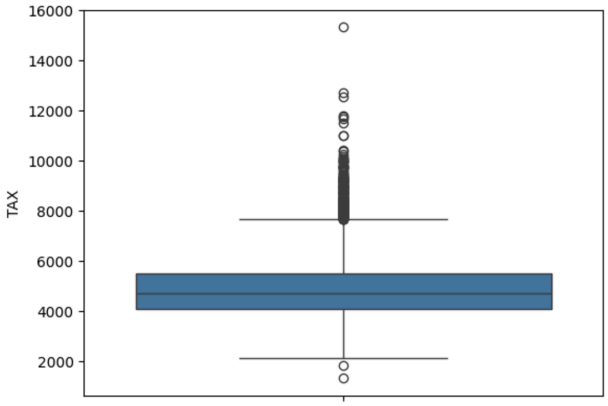
- The median building age is around 85 years
- The median of Gross floor area(GROSS_AREA) is 2700
- The median of Total living area for residential properties (ft2)(LIVING_AREA) is 1500ft2
- The median of Total assessed value for property, in thousands of USD(TOTAL_VALUE) is 380

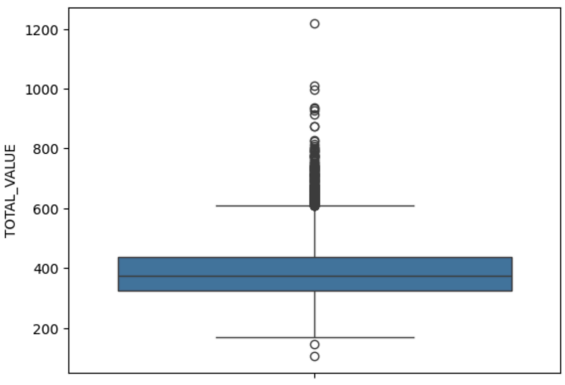






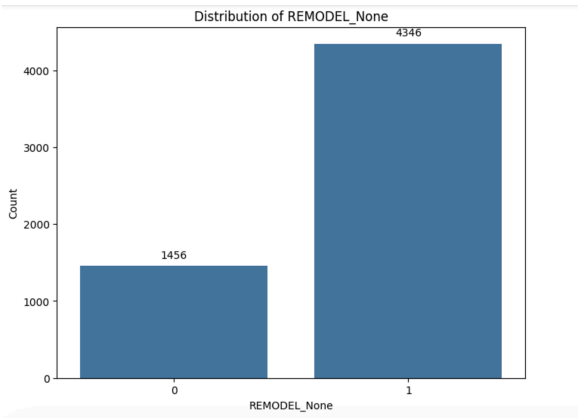


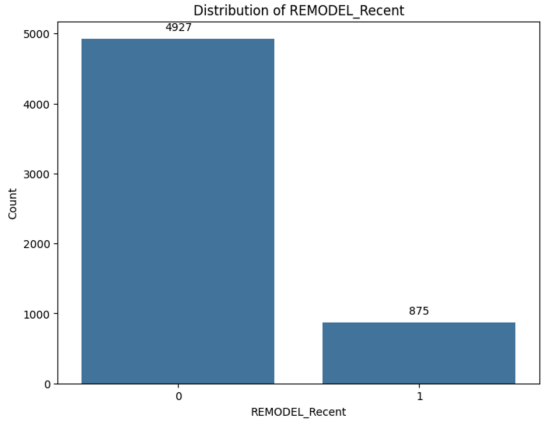


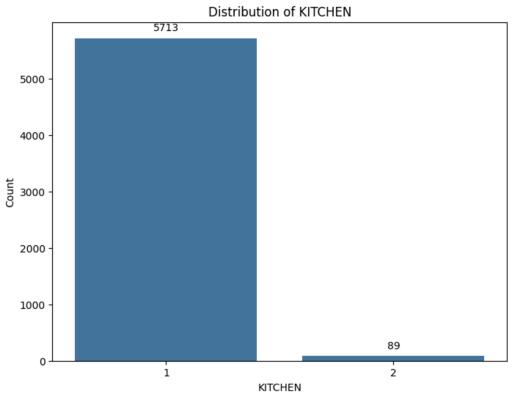


We plot count plots to explore the frequency distribution of categorical variables like property condition, neighborhood, and zoning type. Count Plots are especially useful for spotting imbalances in the data, like underrepresented categories, which may need special consideration during modeling. The below information is extracted from the plots.

- The number of houses not remodeled (REMODEL_None) is higher than number of houses remodeled
- The number of houses that were recently remodeled (REMODEL_Recent) is lesser than the number of houses that were not recently remodeled.
- The number of property with one kitchen (KITCHEN) is higher than the number of property with two kitchens (KITCHEN)



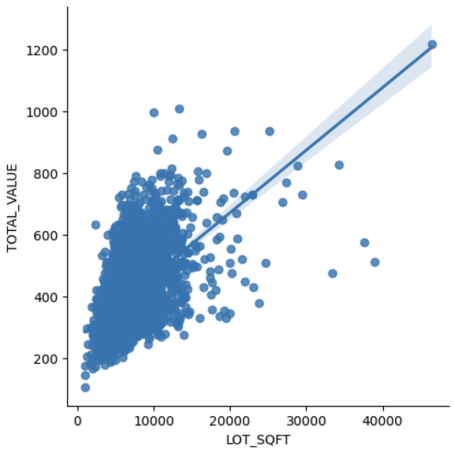


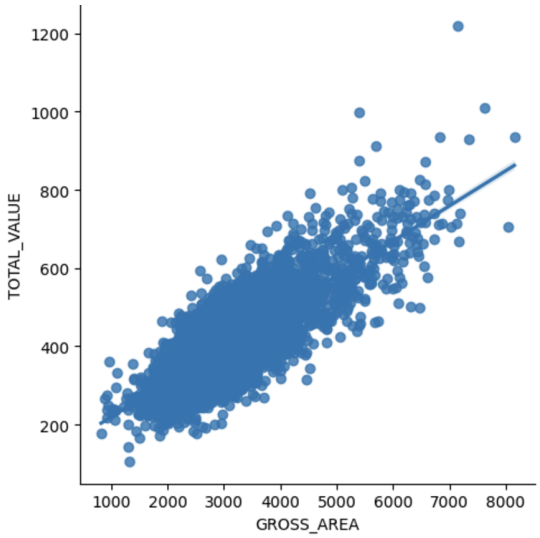


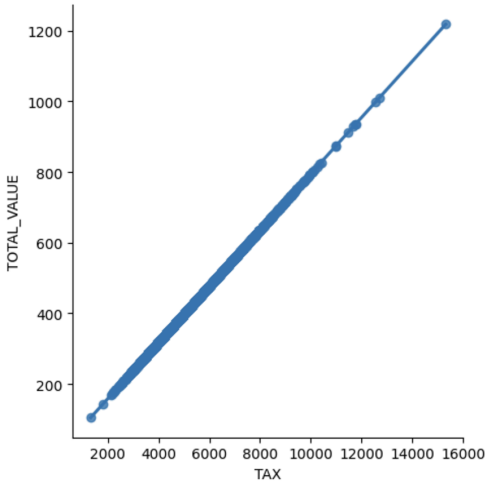
Bivariate Analysis

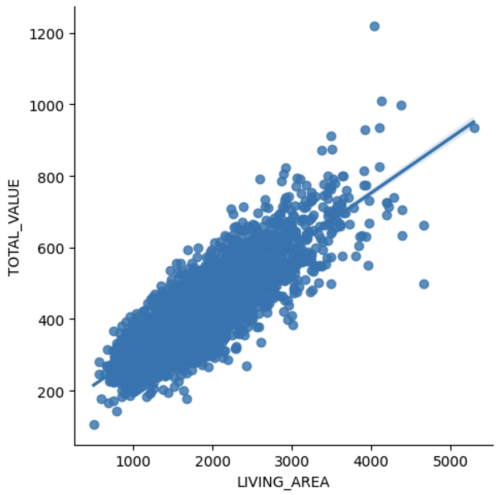
Bivariate analysis of the West Roxbury dataset explores relationships between two variables to uncover patterns and trends. For numerical variables, scatterplots reveal correlations, such as the positive relationship between square footage and

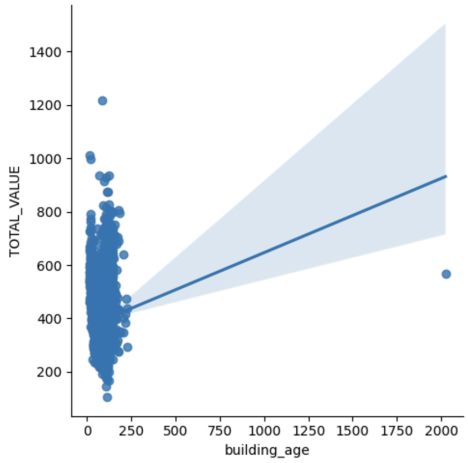
sale price, while heatmaps highlight the strength of correlations across multiple features. Boxplots are used to compare how sale prices vary across categories like property condition or neighborhood, offering insights into factors that drive property values. This analysis helps identify key interactions and dependencies between variables, which are crucial for building accurate predictive models.







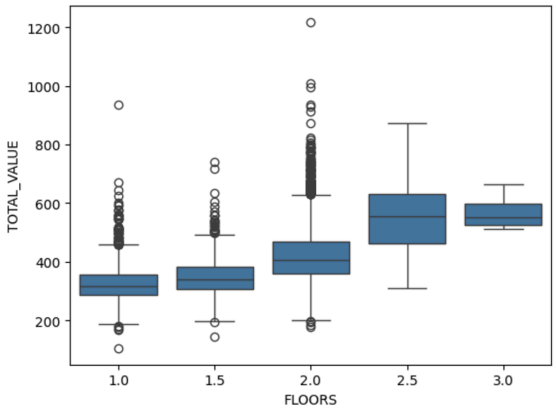


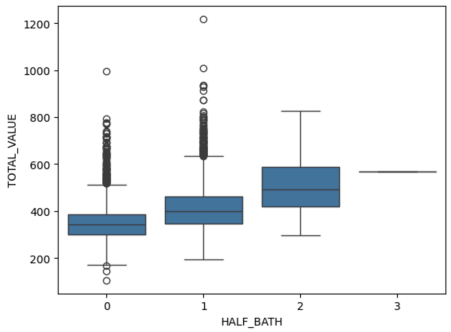


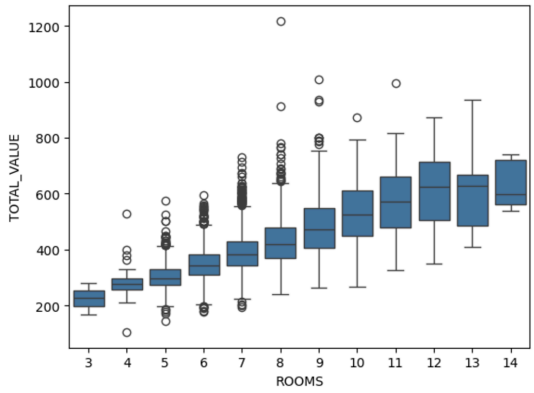
By performing bivariate analysis we obtained the following information :

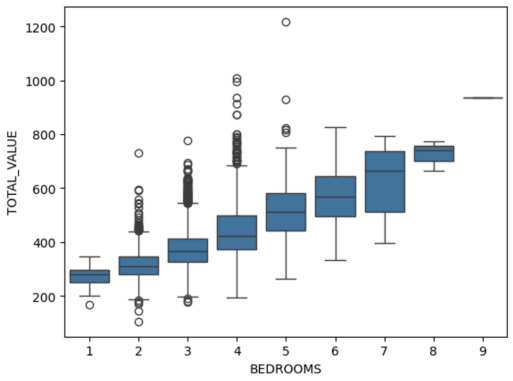
- As the proportion of total lot size of parcel in square feet (LOT_SQFT) increases, the total assessed value for property, in thousands of USD (TOTAL_VALUE) also increases
- As the proportion of gross floor area (GROSS_AREA) increases, the total assessed value for property, in thousands of USD TOTAL_VALUE) also increases
- As the proportion of total living area for residential properties (ft2) (LIVING_AREA) increases, the Total assessed value for property, in thousands of USD (TOTAL_VALUE) also increases
- There seems to be no relationship between building age and the total assessed value of the property
- The number of floors (FLOORS) Increases, the median of Total assessed value for property, in thousands of USD (TOTAL_VALUE) also increases and the median stops increasing after the 2.5 floors
- The number of rooms(ROOMS) Increases, the median of Total assessed value for property, in thousands of USD(TOTAL_VALUE) also increases up to 12 rooms and there is slight decrease for the 13 and 14 rooms
- The number of bedrooms (BEDROOMS) increases, the median of Total assessed value for property, in thousands of USD (TOTAL_VALUE) also increases
- The number of full bath (FULL_BATH) increases, the median of Total assessed value for property, in thousands of USD (TOTAL_VALUE) also increases
- The number of half bath (HALF_BATH) increases, the median of Total assessed value for property, in thousands of USD (TOTAL_VALUE) also increases

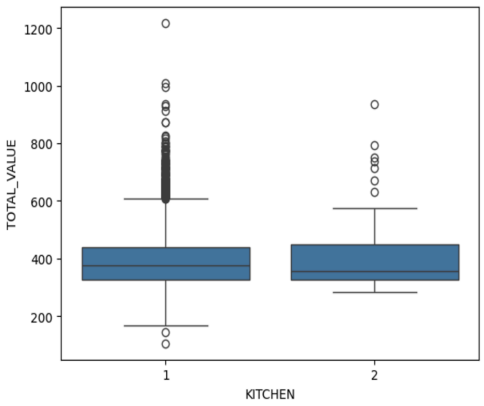
- The number of kitchens (KITCHEN) increases, the median of Total assessed value for property, in thousands of USD (TOTAL_VALUE) is almost the same there is no much difference
- The median of houses not remodeled (REMODEL_None) is slightly lesser the median of houses remodeled
- The median of houses recently remodeled (REMODEL_Recent) is slightly higher the median of houses not recently remodeled

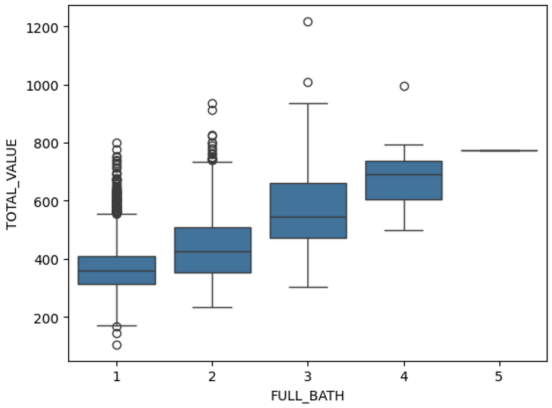


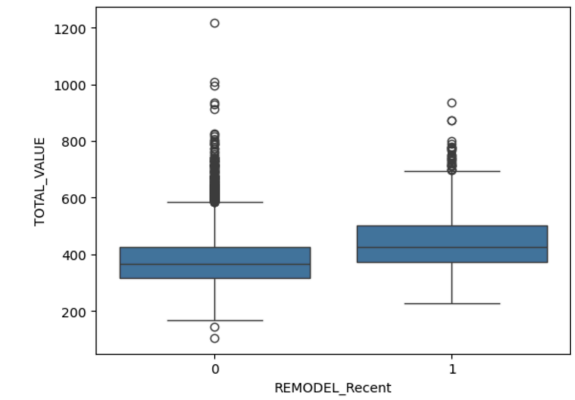


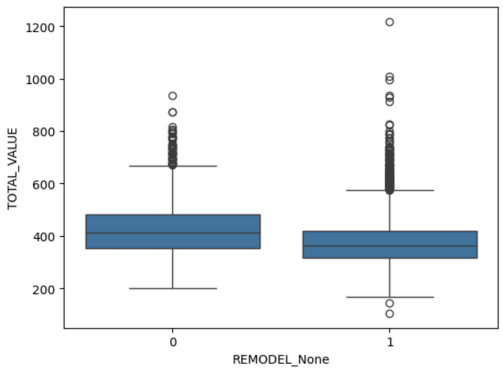








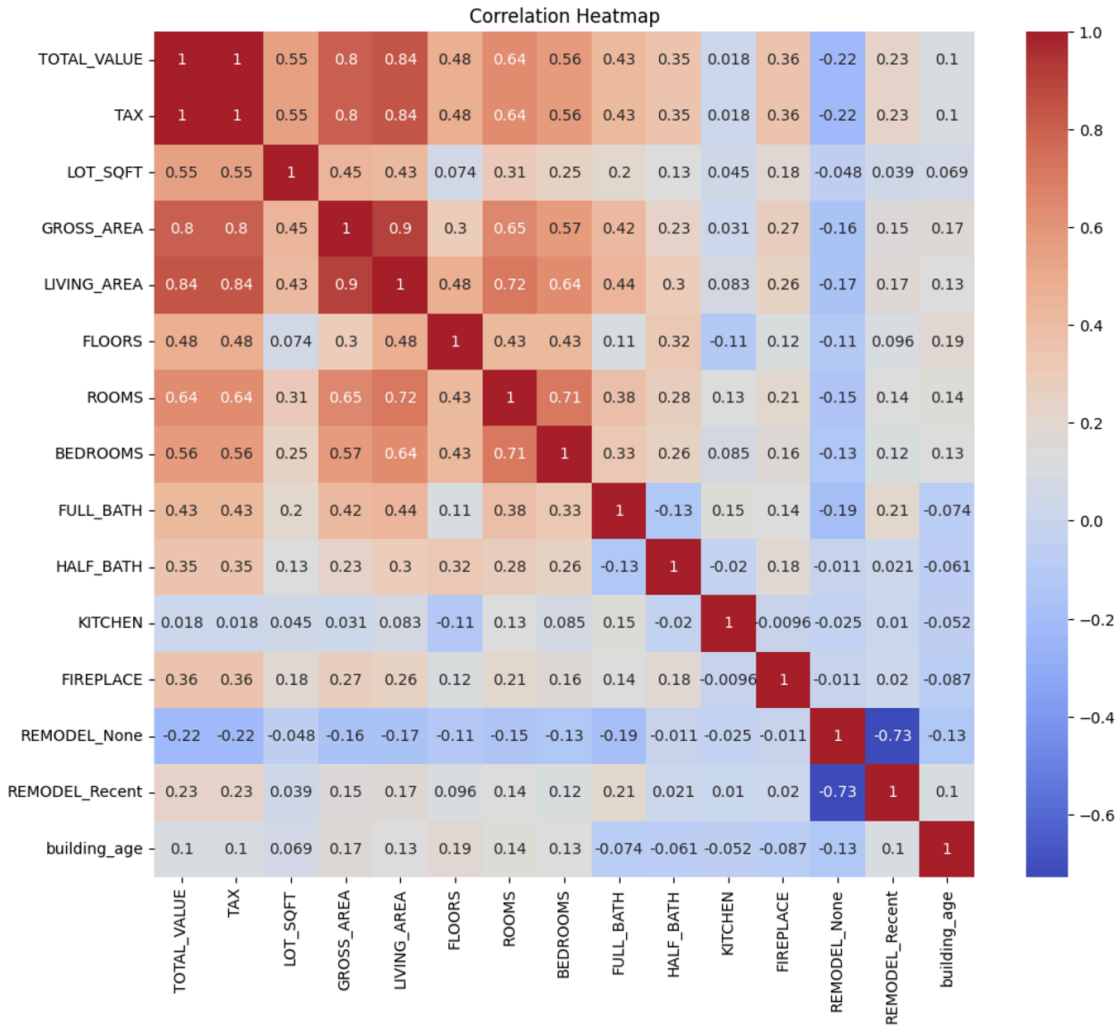


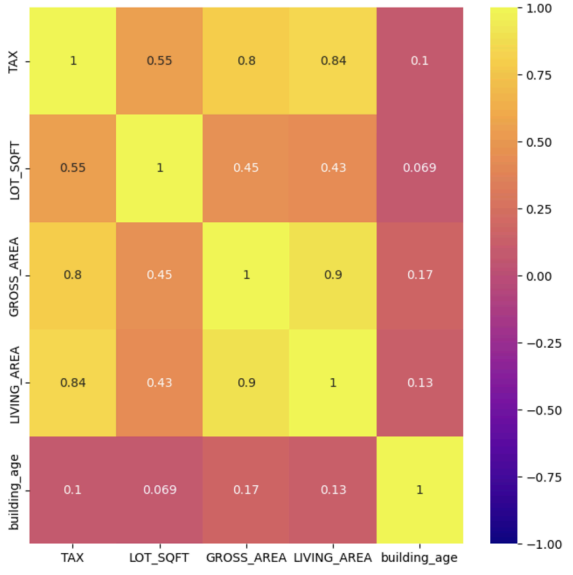


A correlation matrix shows how closely related the numerical features in the dataset are to each other. For example, it can reveal if total value and the number of rooms or other variables are strongly connected, which might mean they overlap in the information they provide. This is helpful for spotting patterns or potential issues, like features that are too similar and could affect the performance of predictive models. A heatmap of the matrix makes it easy to see these relationships, helping us decide which features to keep or adjust for better analysis.

The following correlation matrix shows a strong relationship between GROSS_AREA and LIVING_AREA (0.8998), indicating these variables are closely linked. TAX is moderately correlated with LIVING_AREA (0.8371) and GROSS_AREA (0.8005), reflecting their influence on taxes. In contrast, LOT_SQFT and building_age have weaker correlations with all variables, with building_age showing minimal association.

	TAX	LOT_SQFT	GROSS_AREA	LIVING_AREA	building_age
TAX	1.000000	0.546120	0.800518	0.837122	0.100918
LOT_SQFT	0.546120	1.000000	0.448880	0.426045	0.068908
GROSS_AREA	0.800518	0.448880	1.000000	0.899775	0.167928
LIVING_AREA	0.837122	0.426045	0.899775	1.000000	0.131274
building_age	0.100918	0.068908	0.167928	0.131274	1.000000


Correlation Heatmap
TOTAL_VALUE 1 1 0.55 0.8 0.84 0.48 0.64 0.56 0.43 0.35 0.018 0.36 -0.22 0.23 0.1
TAX 1 1 0.55 0.8 0.84 0.48 0.64 0.56 0.43 0.35 0.018 0.36 -0.22 0.23 0.1
LOT_SQFT 0.55 0.55 1 0.45 0.43 0.074 0.31 0.25 0.2 0.13 0.045 0.18 -0.048 0.039 0.069
GROSS_AREA 0.8 0.8 0.45 1 0.9 0.3 0.65 0.57 0.42 0.23 0.031 0.27 -0.16 0.15 0.17
LIVING_AREA 0.84 0.84 0.43 0.9 1 0.48 0.72 0.64 0.44 0.3 0.083 0.26 -0.17 0.17 0.13
FLOORS 0.48 0.48 0.074 0.3 0.48 1 0.43 0.43 0.11 0.32 -0.11 0.12 -0.11 0.096 0.19
ROOMS 0.64 0.64 0.31 0.65 0.72 0.43 1 0.71 0.38 0.28 0.13 0.21 -0.15 0.14 0.14
BEDROOMS 0.56 0.56 0.25 0.57 0.64 0.43 0.71 1 0.33 0.26 0.085 0.16 -0.13 0.12 0.13
FULL_BATH 0.43 0.43 0.2 0.42 0.44 0.11 0.38 0.33 1 -0.13 0.15 0.14 -0.19 0.21 -0.074
HALF_BATH 0.35 0.35 0.13 0.23 0.3 0.32 0.28 0.26 -0.13 1 -0.02 0.18 -0.011 0.021 -0.061
KITCHEN 0.018 0.018 0.045 0.031 0.083 -0.11 0.13 0.085 0.15 -0.02 1 -0.0096 -0.025 0.01 -0.052
FIREPLACE 0.36 0.36 0.18 0.27 0.26 0.12 0.21 0.16 0.14 0.18 -0.0096 1 -0.011 0.02 -0.087
REMODEL_None -0.22 -0.22 -0.048 -0.16 -0.17 -0.11 -0.15 -0.13 -0.19 -0.011 -0.025 -0.011 1 -0.73 -0.13
REMODEL_Recent 0.23 0.23 0.039 0.15 0.17 0.096 0.14 0.12 0.21 0.021 0.01 0.02 -0.73 1 0.1
building_age 0.1 0.1 0.069 0.17 0.13 0.19 0.14 0.13 -0.074 -0.061 -0.052 -0.087 -0.13 0.1 1
TOTAL_VALUE TAX LOT_SQFT GROSS_AREA LIVING_AREA FLOORS ROOMS BEDROOMS FULL_BATH HALF_BATH KITCHEN FIREPLACE REMODEL_None REMODEL_Recent building_age
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6


TAX 1 0.55 0.8 0.84 0.1
LOT_SQFT 0.55 1 0.45 0.43 0.069
GROSS_AREA 0.8 0.45 1 0.9 0.17
LIVING_AREA 0.84 0.43 0.9 1 0.13
building_age 0.1 0.069 0.17 0.13 1
TAX LOT_SQFT GROSS_AREA LIVING_AREA building_age
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00

Model Exploration, Performance Evaluation and Comparison

1. Linear Regression Model

A linear regression model offers a simple way to predict property prices by looking at how features like square footage, lot size, and number of rooms influence sale price. It works by assuming a straight-line relationship between these features and the price, which makes it easy to understand how each one affects the prediction. For example, the model can show how much adding extra square footage might increase a home's value. While it's straightforward and quick, linear regression might struggle with more complex patterns or when features are too similar to each other.

To evaluate the accuracy of predictions, we calculate metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). RMSE measures the average magnitude of errors, emphasizing larger errors by squaring them, making it useful for spotting big deviations in predictions. MAE, on the other hand, provides the average error in absolute terms, offering a straightforward interpretation of how far predictions deviate from actual values. MAPE expresses errors as a percentage, which helps understand model performance relative to the scale of the target variable, such as sale price. Together, these metrics give a good view of how well a model is performing and highlight areas for improvement.

- In the main table, p value of F-stat, 0.00, is well below our level of significance 0.05. So, the overall model is statistically significant and valid for estimating the population Y variable.
- Adjusted R-Squared value of 82.4% indicates good explanatory power of the independent variables for changes in Y variable
- Based on the model output, the X variable that has the maximum positive impact on the price is the number of full_bath set_5 because of high coefficient
- Predicted property value for the first 10 records with values ranging from approximately 290 to 549, capturing a linear relationship from the input features
- R-Squared value of 81% indicates good explanatory power of the independent variables for changes in Y variable
- The Y-intercept or constant value is 94.78
- Predicted property value for the first 5 records, with values ranging from approximately 349 to 549

- MAE is 31.292, RMSE is 41.615, MAPE value of 8.12 indicates good prediction accuracy of the model

OLS Regression Results

Dep. Variable:	TOTAL_VALUE	**R-squared:**	0.824
Model:	OLS	**Adj. R-squared:**	0.823
Method:	Least Squares	**F-statistic:**	657.3
Date:	Thu, 12 Dec 2024	**Prob (F-statistic):**	0.00
Time:	21:01:11	**Log-Likelihood:**	-29865.
No. Observations:	5802	**AIC:**	5.981e+04
Df Residuals:	5760	**BIC:**	6.009e+04
Df Model:	41		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	95.0034	24.440	3.887	0.000	47.092	142.914
LOT_SQFT	0.0086	0.000	36.638	0.000	0.008	0.009
building_age	-0.0554	0.017	-3.311	0.001	-0.088	-0.023
GROSS_AREA	0.0334	0.002	20.581	0.000	0.030	0.037
LIVING_AREA	0.0486	0.003	16.156	0.000	0.043	0.055
FLOORS_1.5	-2.7105	1.976	-1.371	0.170	-6.585	1.164
FLOORS_2.0	40.4676	1.750	23.123	0.000	37.037	43.898
FLOORS_2.5	44.2142	4.755	9.298	0.000	34.892	53.536
FLOORS_3.0	21.3836	22.278	0.960	0.337	-22.289	65.056
ROOMS_4	11.2982	24.920	0.453	0.650	-37.555	60.151
ROOMS_5	5.6252	24.799	0.227	0.821	-42.990	54.241
ROOMS_6	-0.7245	24.824	-0.029	0.977	-49.389	47.940
ROOMS_7	2.3804	24.849	0.096	0.924	-46.333	51.094
ROOMS_8	5.0294	24.886	0.202	0.840	-43.757	53.816
ROOMS_9	6.0028	24.975	0.240	0.810	-42.958	54.963
ROOMS_10	0.8774	25.127	0.035	0.972	-48.380	50.135
ROOMS_11	19.7256	25.536	0.772	0.440	-30.334	69.785
ROOMS_12	-6.9815	25.895	-0.270	0.787	-57.746	43.783
ROOMS_13	-3.8767	28.698	-0.135	0.893	-60.136	52.383
ROOMS_14	-3.3578	31.873	-0.105	0.916	-65.841	59.125
BEDROOMS_2	12.8158	8.320	1.540	0.124	-3.495	29.126
BEDROOMS_3	10.8029	8.417	1.284	0.199	-5.697	27.303
BEDROOMS_4	10.1206	8.539	1.185	0.236	-6.619	26.860
BEDROOMS_5	2.2736	8.997	0.253	0.800	-15.363	19.910
BEDROOMS_6	10.0324	9.865	1.017	0.309	-9.307	29.372
BEDROOMS_7	13.0231	14.457	0.901	0.368	-15.318	41.364
BEDROOMS_8	-16.1230	28.056	-0.575	0.566	-71.124	38.878
BEDROOMS_9	71.0378	45.540	1.560	0.119	-18.238	160.313

FULL_BATH_2	20.0956	1.528	13.151	0.000	17.100	23.091
FULL_BATH_3	54.6993	3.997	13.686	0.000	46.864	62.534
FULL_BATH_4	110.5810	12.107	9.133	0.000	86.846	134.316
FULL_BATH_5	131.0120	42.514	3.082	0.002	47.669	214.355
HALF_BATH_1	18.6293	1.307	14.249	0.000	16.066	21.192
HALF_BATH_2	40.9449	3.861	10.604	0.000	33.375	48.514
HALF_BATH_3	36.7378	42.138	0.872	0.383	-45.868	119.344
KITCHEN_2	-15.5804	4.741	-3.286	0.001	-24.875	-6.286
FIREPLACE_1	22.6267	1.238	18.277	0.000	20.200	25.054
FIREPLACE_2	30.4371	2.883	10.559	0.000	24.786	36.088
FIREPLACE_3	31.0546	9.149	3.394	0.001	13.119	48.990
FIREPLACE_4	13.7576	21.294	0.646	0.518	-27.987	55.502
REMODEL_None_1	-5.1977	1.865	-2.787	0.005	-8.854	-1.542
REMODEL_Recent_1	20.3281	2.260	8.995	0.000	15.898	24.758

Omnibus:	280.598	**Durbin-Watson:**	1.573
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1014.326
Skew:	0.047	**Prob(JB):**	5.52e-221
Kurtosis:	5.046	**Cond. No.**	1.19e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.19e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

	coef	pvalue
FULL_BATH_5	131.011983	0.002
FULL_BATH_4	110.580961	0.000
const	95.003377	0.000
FULL_BATH_3	54.699296	0.000
FLOORS_2.5	44.214189	0.000
HALF_BATH_2	40.944866	0.000
FLOORS_2.0	40.467564	0.000
FIREPLACE_3	31.054566	0.001
FIREPLACE_2	30.437116	0.000
FIREPLACE_1	22.626697	0.000
REMODEL_Recent_1	20.328081	0.000
FULL_BATH_2	20.095587	0.000
HALF_BATH_1	18.629327	0.000
LIVING_AREA	0.048645	0.000
GROSS_AREA	0.033354	0.000
LOT_SQFT	0.008630	0.000
building_age	-0.055355	0.001
REMODEL_None_1	-5.197686	0.005
KITCHEN_2	-15.580399	0.001

```
Linear Regression – Train R2: 0.826, Test R2: 0.812
```

2. Random Forest Model

The Random Forest model is a powerful tool for the West Roxbury dataset, using multiple decision trees to make predictions. By combining the results of many trees, it reduces overfitting and improves accuracy. This makes it ideal for complex data, like predicting property values or crime rates in West Roxbury. Additionally, Random Forest helps identify the most important features influencing the predictions, providing valuable insights into the factors that matter most.

```
Random Forest - Train R2: 0.974, Test R2: 0.819
```

3. Support Vector Regression

Support Vector Regression (SVR) is a useful model for predicting outcomes in the West Roxbury dataset. It works by finding a hyperplane that best fits the data while allowing some margin for error. SVR is effective for capturing complex, non-linear relationships, making it ideal for predicting variables like property values or crime rates. Its ability to handle noise and outliers helps provide accurate predictions in real-world scenarios.

```
SVR - Train R2: 0.566, Test R2: 0.549
```

Performing PCA and running different models

Principal Component Analysis (PCA) helps simplify and compare different models by reducing the data to its most important features. It highlights the key patterns and relationships, making it easier to see how models like Random Forest or Support Vector Regression perform. By transforming the data into a smaller set of uncorrelated components, PCA allows us to focus on what really matters, helping us understand which features drive predictions and how well different models capture those patterns.

Based on the results provided above, we can analyze the performance of the three models (Linear Regression, Random Forest, and Support Vector Regression) both with and without Principal Component Analysis (PCA). Here's an interpretation of the results:

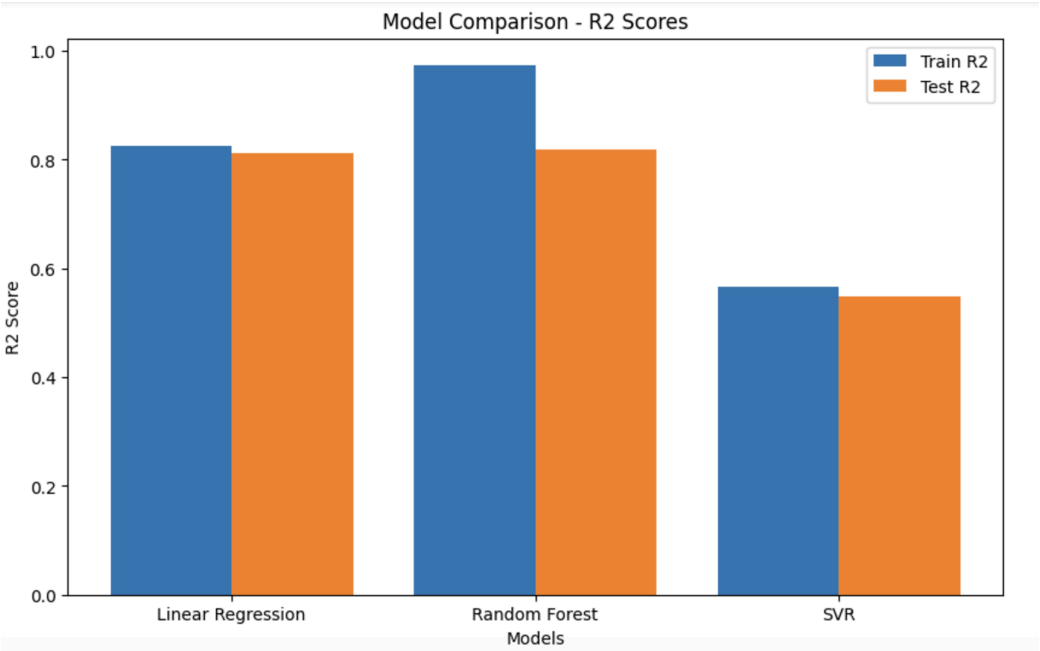
Without PCA

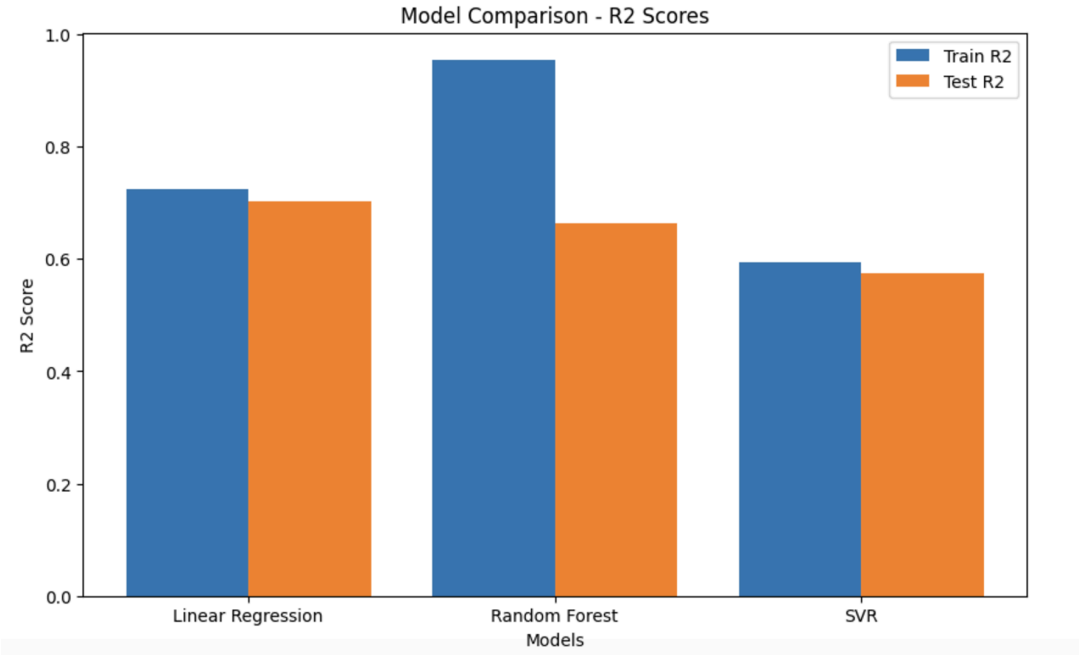
Model	Train R^2	Test R^2	Notes
Linear Regression	0.826	0.812	The model performs well and shows good generalization, with only a small difference between train and test scores.
Random Forest	0.974	0.819	Excellent performance on training data, but the large gap between train and test scores suggests overfitting.
Support Vector Regression (SVR)	0.566	0.549	Consistent performance between train and test sets, but overall lower accuracy compared to the other models.

With PCA

Model	Train R^2	Test R^2	Notes
Linear Regression	0.725	0.703	Performance decreased compared to without PCA, but still shows good generalization.
Random Forest	0.955	0.664	Still shows signs of overfitting, and test performance decreased significantly with PCA.
Support Vector Regression (SVR)	0.595	0.575	Slight improvement in both train and test scores compared to without PCA.

Model Comparison and Selection





1. Best performing model: Random Forest without PCA
    - Highest test R2 score (0.819)
    - However, it shows signs of overfitting
2. Most consistent model: Linear Regression without PCA

- Good balance between performance (test R2: 0.812) and generalization

3. Impact of PCA:
    - Decreased performance for Linear Regression and Random Forest
    - Slightly improved performance for SVR
    - Generally reduced overfitting in Random Forest, but at the cost of accuracy

Recommendation

For this dataset, the Linear Regression model without PCA appears to be the best choice for deployment in a real-life scenario. It offers high accuracy (test R2: 0.812) and good generalization, without the overfitting issues seen in the Random Forest model.

Advantages of Linear Regression model

1. Good performance (R2: 0.812)
2. Consistent results between training and test sets
3. Interpretability of coefficients
4. Computationally efficient

Disadvantages of Linear Regression model

1. Assumes linear relationships between features and target
2. May not capture complex, non-linear patterns in the data
3. Sensitive to outliers

Conclusion

In conclusion, while the Random Forest model achieved the highest train R^2 score, its significantly lower test R^2 score indicates overfitting, where the model performs exceptionally well on the training data but struggles to generalize to unseen data. On the other hand, the Linear Regression model, despite having a slightly lower test R^2 score compared to Random Forest, offers a better balance between performance and generalization. It maintains consistent R^2 scores across both training and test sets, demonstrating its ability to avoid overfitting and provide reliable predictions. Additionally, Linear Regression is inherently more interpretable, making it easier to understand the relationships between features and the target variable. For this dataset, where simplicity and interpretability are important considerations, Linear Regression stands out as the more practical and reliable choice.
