Boston Real Estate Market Analysis
Unveiling Insights from the Real Estate Dataset of Boston

Analytics Project
Phase 2

ISDS 577- Group 6

Ananya Rattani
Siddharth Mahesh Bartake

Table of Contents

Executive Summary

The Boston Real Estate market analysis combines data from Kaggle, data.boston.gov, and Boston-pd-crime hub, ensuring a comprehensive exploration of Boston's real estate landscape. The dataset, provided in the csv file, merges property assessments, public school information, and crime data based on standard zip codes, totaling 61,741 rows and 47 variables.

The study addresses six key research questions using regression analysis, correlation analysis, descriptive statistics, and machine learning algorithms. The questions span property features' impact on value, property age dynamics, tax and crime rate correlations, property price trends over time, and predicting prices based on multiple factors.

The analysis aims to yield actionable insights for managerial decision-making in real estate. It guides investment strategies by prioritizing property features for maximum value and understanding the influence of property age dynamics. It informs risk management by correlating property taxes with crime rates and aids strategic planning with historical property price trends. Additionally, predictive modeling supports decision-making by identifying key predictors of property prices.

Various analytical methods, including regression, correlation, and descriptive statistics, are employed to explore diverse aspects of the research questions. Their selection is based on their ability to provide nuanced insights into the relationships and trends within the complex real estate dataset.

Excel, Python, and Tableau are utilized for data analysis and visualization, offering a versatile toolkit. Visualizations range from scatter plots and regression plots to heatmaps and geographical maps, providing an intuitive representation of property data and relationships.

This comprehensive real estate analysis equips stakeholders with valuable insights for strategic decision-making in Boston's dynamic property market by employing a multifaceted analytical approach. The findings empower managers to prioritize investments, mitigate risks, and optimize resource allocation, fostering informed and data-driven decision-making in the ever-evolving real estate landscape.

Data

Data collection:

We got data from 3 sources – Kaggle for Boston Public school data, data.boston.gov for Boston property assessment for the year 2019 and Boston-pd-crime hub for Reported crimes.

Source Links

https://boston-pd-crime-hub-boston.hub.arcgis.com/datasets/

https://data.boston.gov/dataset/property-assessment/resource/695a8596-5458-442b-a017-7cd72471aade

https://www.kaggle.com/datasets/crawford/boston-public-schools

Data Legal Privacy:

There are no legal or privacy concerns associated with our data because it is public data collected through 3 public websites. It enables collaborative data sharing.

Data Format:

The data will be provided in excel format.

Data Integration:

We will be integrating Boston property assessment for the year 2019, with Boston Public Schools dataset based on their common zip codes. In addition, we will integrate crime dataset based on their zip codes for the year 2019.

Dataset Size:

The final dataset will have a total of 61741 rows and 47 variables.

Data set Attributes:

In our real estate analysis, we will focus on key attributes such as property address, type, and value, along with crime data including offense description and neighborhood. We will also consider property size (land size and living area), age (built and remodeled), number of rooms and bedrooms, total baths, school information (name and type), and property tax. These attributes collectively provide a comprehensive view of the property and its surroundings, helping us make informed decisions.

Data Cleaning & Preparation:

For our real estate project, it will be essential to preprocess the dataset to ensure its quality and relevance to our analysis. The dataset likely contains missing values and irrelevant variables that need to be addressed. Additionally, to streamline our analysis, we will focus on data from a specific timeframe. We will be using Alteryx to clean and prepare the dataset. Alteryx provides a user-friendly interface and a wide range of tools that can efficiently handle data cleaning tasks, such as dealing with missing values and removing irrelevant variables.

Collecting Raw data

We have collected Crime data from the public available dataset using - https://boston-pd-crime-hub-boston.hub.arcgis.com/datasets/. They have attributes like objectId, incNum, crime,

offenseCode, offenseDesc, block, city, zip, district, premiseDesc, weaponDesc, crimePart, fromDate, hourOfDay, dayOfWeek, year, quarter, month, neighborhood.

OBJECTID	INC_NUM	CRIME	OFFENSE_CODE	OFFENSE_DESC	BLOCK
35736	222025789	Fraud	1102	FRAUD - FALSE PRETENSE / SCHEME	0 BLOCK CAROL CIR
38062	222047846	Investigate Person	3115	INVESTIGATE PERSON	0 BLOCK OLD IRONSIDES WAY
39016	222054003	Property Lost	3201	PROPERTY - LOST/ MISSING	1500 BLOCK HYDE PARK AVE
39300	222017603	Fraud	1102	FRAUD - FALSE PRETENSE / SCHEME	400 BLOCK BLUE HILL AVE
40126	222026337	Human Trafficking	1620	Human Trafficking - Involuntary Servitude	RUSFIELD ST & ALLERTON ST
40670	222033777	Criminal Harassment	2670	HARASSMENT/ CRIMINAL HARASSMENT	0 BLOCK BOWKER ST
40984	222056080	Other Larceny	617	LARCENY THEFT FROM BUILDING	0 BLOCK BLAKE ST
41430	222032522	Verbal Disputes	3301	VERBAL DISPUTE	2100 BLOCK DORCHESTER AVE
41531	222033410	Fraud	1109	FRAUD - WIRE	100 BLOCK WASHINGTON ST

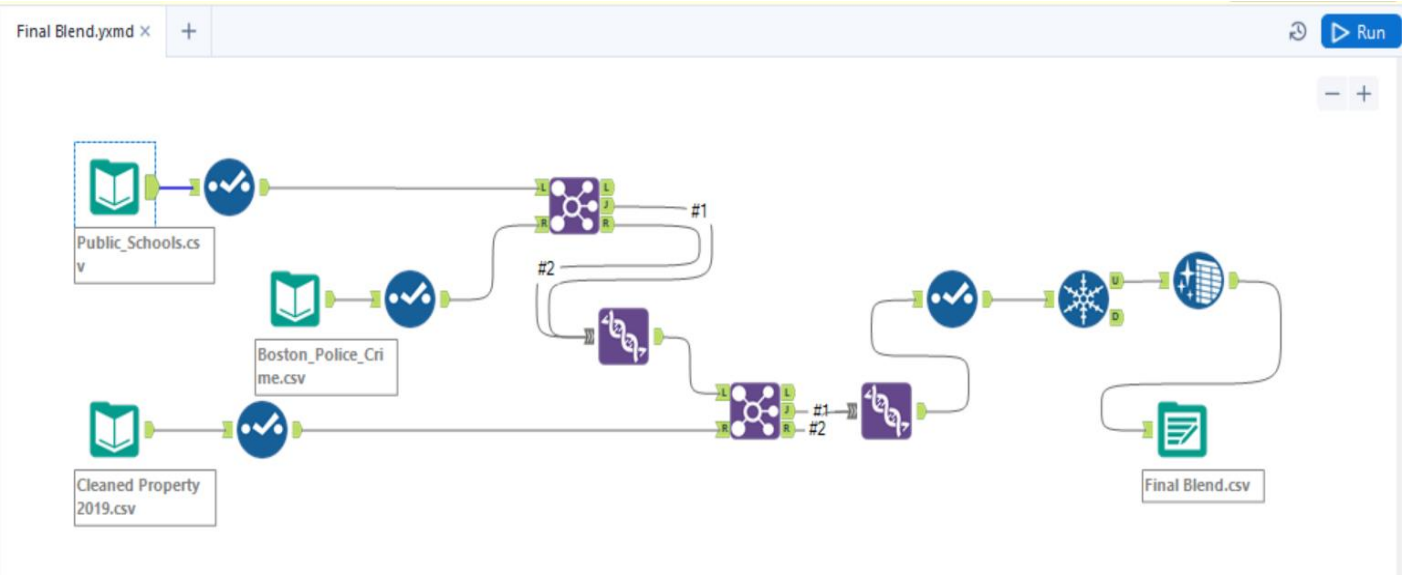
We have collected school data from the public available dataset using - https://www.kaggle.com/datasets/crawford/boston-public-schools. They have attributes like, bldgId, bldgName, address, city, zipcode, schId, schName, schLabel, schType.

BLDG_ID	BLDG_NAME	ADDRESS	CITY	ZIPCODE	SCH_ID	SCH_NAME	SCH_LABEL	SCH_TYPE
1	Guild Bldg	195 Leyden Street	East Boston	2128	4061	Guild Elementary	Guild	ES
3	Kennedy, P Bldg	343 Saratoga Street	East Boston	2128	4541	Kennedy Patrick Elem	PJ Kennedy	ES
4	Otis Bldg	218 Marion Street	East Boston	2128	4322	Otis Elementary	Otis	ES
6	Odonnell Bldg	33 Trenton Street	East Boston	2128	4543	O'Donnell Elementary	O'Donnell	ES
7	East Boston High Bldg	86 White Street	East Boston	2128	1070	East Boston High	East Boston HS	HS
8	Umana / Barnes Bldg	312 Border Street	East Boston	2128	4323	Umana Academy	Umana Academy	K-8
10	East Boston Eec Bldg	135 Gove Street	East Boston	2128	4450	East Boston EEC	East Boston EEC	ELC
11	Mckay Bldg	122 Cottage Street	East Boston	2128	4360	McKay K-8	McKay K-8	K-8

We have collected property data from the public available dataset using - https://data.boston.gov/dataset/property-assessment/resource/695a8596-5458-442b-a017-7cd72471aade. They have a few attributes like - pid, stNum, stName, zipcode, avTotal, grossTax, yrBuilt, grossArea, livingArea, numFloors.

_id	PID	CM_ID	GIS_ID	ST_NUM	ST_NAME	ST_NAME
1	5.03E+08	5.03E+08	5.03E+08	87	BEACON	ST
2	5.03E+08	5.03E+08	5.03E+08	87	BEACON	ST
3	5.03E+08	5.03E+08	5.03E+08	87	BEACON	ST
4	5.03E+08	5.03E+08	5.03E+08	87	BEACON	ST
5	5.03E+08	5.03E+08	5.03E+08	87	BEACON	ST
6	5.03E+08	5.03E+08	5.03E+08	88	BEACON	ST
7	5.03E+08	5.03E+08	5.03E+08	88	BEACON	ST
8	5.03E+08	5.03E+08	5.03E+08	88	BEACON	ST

Extract, Transform and Loading in Alteryx



Data Blending:

We initiated the process by merging three datasets: public school's data, crime data, and property data. The common identifier utilized for merging these datasets was the zip code.

Joining Public Schools with Crime Data:

Firstly, we combined the public school's data with the crime data using an inner join. This approach ensured that only the records with matching zip codes in both datasets were retained. Our objective here was to investigate any potential relationships between public schools and crime rates in specific areas.

Combining Result with Property Data:

Subsequently, after obtaining the merged dataset from the previous step, we merged it with the property data using another inner join. Once again, this ensured that only the records with matching zip codes across all three datasets were included. The aim was to analyze any connections between public schools, crime rates, and property values in the same geographical areas.

Data Validation and Cleaning:

We employed the select tool to inspect and validate the data types, ensuring consistency across the dataset. We identified and addressed outliers to ensure data reliability for analysis. Additionally, we performed data cleaning tasks such as handling missing values, correcting inconsistencies, and removing irrelevant attributes to prepare the dataset for further analysis.

Union and Final Cleansing:

Following the merging of datasets, we utilized the union tool to merge them into a single dataset. Subsequently, another select tool was applied to eliminate any unnecessary attributes not required for analysis. We then utilized the unique tool to remove duplicate records from the

dataset. Final cleansing steps were undertaken to verify and validate the data types once more, ensuring data quality before exporting.

Output to Excel File:

Finally, we exported the cleaned and processed dataset to an Excel file for further analysis or reporting purposes.

Research Questions

Research Question 1: Property Features and Value Analysis

How do property features influence property values?

What are the most desirable property features that contribute to higher property values?

Managerial Decision-making: This analysis addresses a managerial question related to real estate valuation and investment strategy in neighborhoods with varying crime rates. Specifically, it seeks to understand how property features, such as the number of rooms and amenities, affect property values in these neighborhoods. The managerial problem or question this analysis answers could be: *"How should we prioritize property features to maximize property value in neighborhoods with varying crime rates?"*

Reasoning: This analysis is crucial for understanding buyer preferences and property market dynamics. By examining how property features correlate with property values in neighborhoods with varying crime rates, we can gain insights into: "*Which features are more desirable to buyers in different market conditions?"*

Analytical Methods:

1. Descriptive statistics and correlation analysis will be used to examine the impact of property features on value in neighborhoods with varying crime rates.

Graphs and Visualizations:

1. Scatter plots: Show the relationship between property features (number of rooms, amenities) and property value. Help identify any trends or patterns in how property features affect property value in neighborhoods with varying crime rates.

2. Correlation matrices: Quantify the strength and direction of the relationship between property features, property value, and crime rate. Provide a comprehensive overview of the correlations between variables, aiding in identifying significant relationships.
3. Heatmaps: Visualize the spatial distribution of property features, property values, and crime rates. Identify any geographical patterns or clusters in how property features relate to property value and crime rate.

Tools: Python and Power BI will be employed for the above analysis.

1) Defining the selected features:

```
selected_features = ['Total Bedrooms', 'Total FullBaths', 'Total Halfbaths',
              'Total Kitchens', 'Heating Type', 'AC Type','Roof Type', 'Exterior Type',
              'Num Floors','Property Area','Building Style','Building Value','Land size',
              'City & State', 'Estate Zipcode','Total Property Value']
```

2) One hot encoding and correlation matrix

```
selected_data = data[selected_features]
# One-hot encode categorical variables
selected_data = pd.get_dummies(selected_data, columns=['Heating Type', 'AC Type','Roof Type',
                                                       'Exterior Type', 'Building Style',
                                                       'City & State', 'Estate Zipcode'])
# Calculate correlation matrix
correlation_matrix = selected_data.corr()
```

```
# Plot heatmap to visualize correlation matrix
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```

Data Selection: The code begins by selecting specific features from the `data` DataFrame based on the list of `selected_features`. These features are stored in a new DataFrame called `selected_data`.

One-Hot Encoding: Categorical variables within the `selected_data` DataFrame are transformed into binary vectors using one-hot encoding.

Each categorical feature specified in the `columns` parameter of `pd.get_dummies()` is converted into multiple binary columns, with each column representing a unique category.

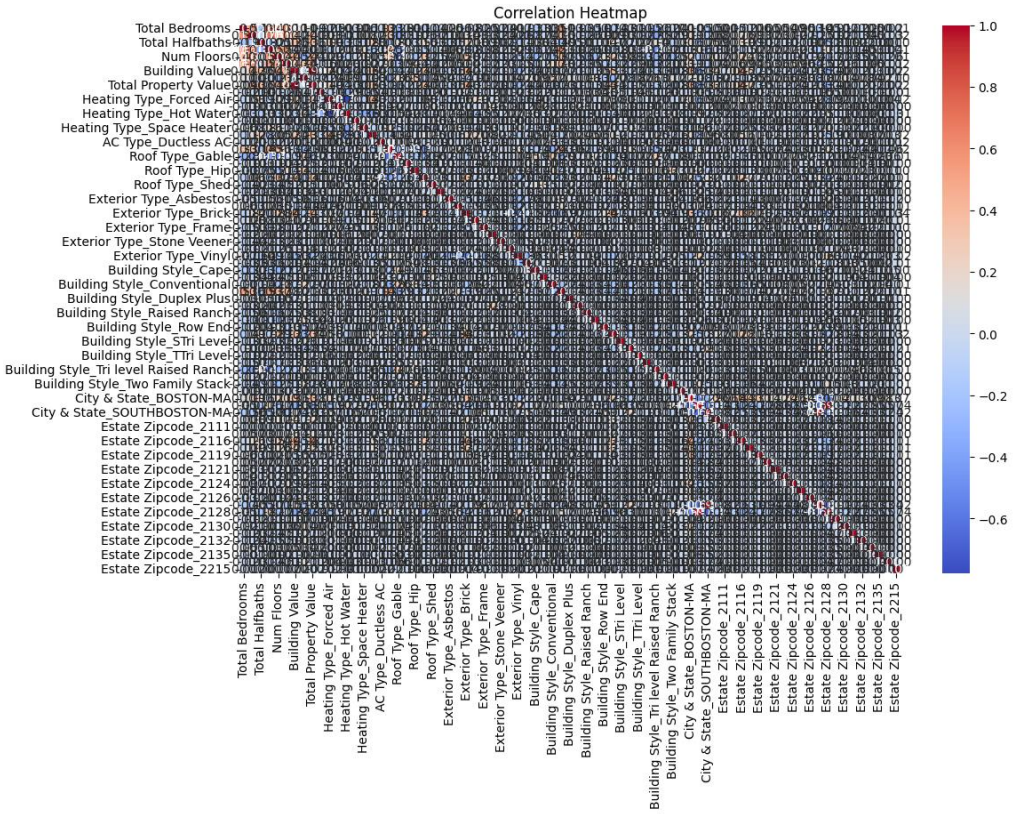
Correlation Matrix Calculation: Once the categorical variables are encoded, a correlation matrix is calculated for the `selected_data` DataFrame.The correlation matrix provides insights into the linear relationships between different variables by computing correlation coefficients.

Correlation Matrix Visualization: To visualize the correlation matrix, a heatmap is plotted using the Seaborn library. `plt.figure(figsize=(12, 8))` sets the figure size of the heatmap to make it visually appealing and readable.

This visualization aids in understanding the relationships between different variables in the dataset, helping in feature selection, identifying multicollinearity, and guiding further analysis or modeling decisions.

Correlation:

Total Bedrooms, Total Bathrooms, Number of floors, Building Value, Heating Type and AC type has a positive correlation.



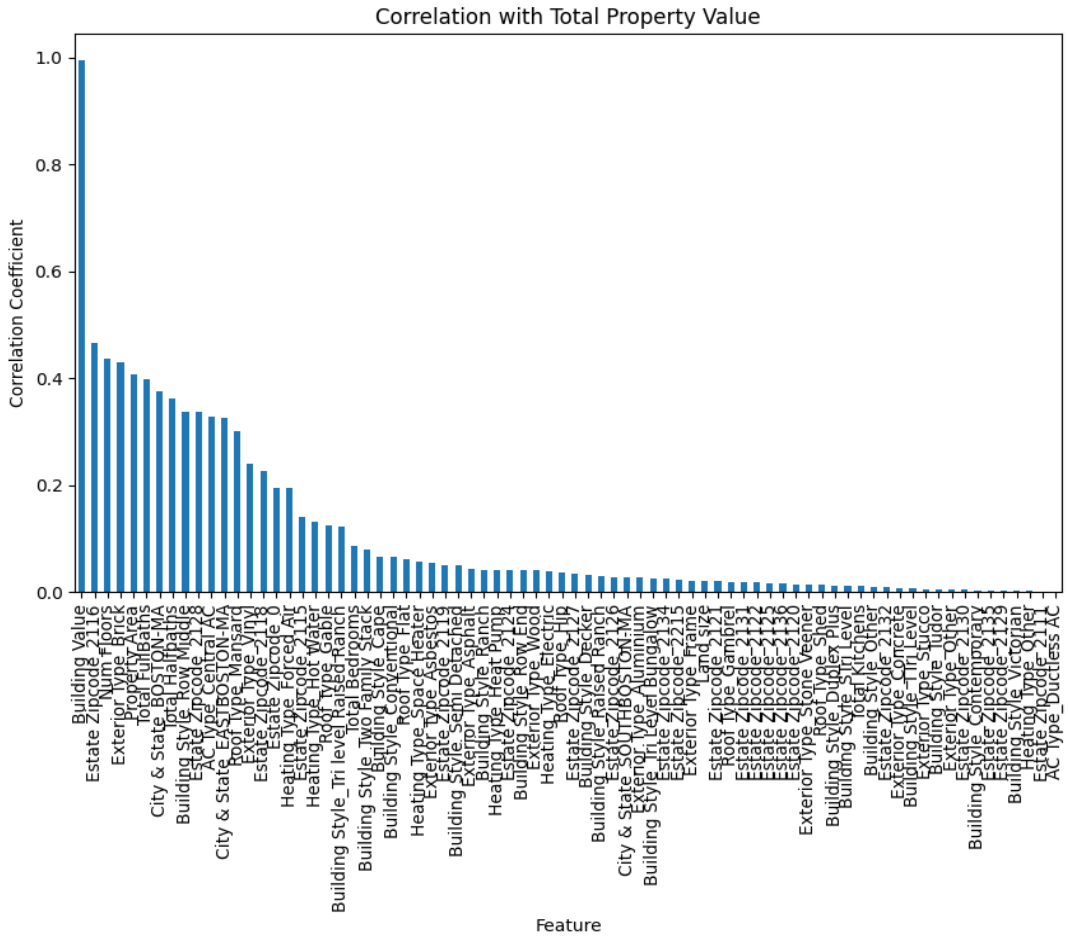
3) Correlation coefficients

```
Correlation Coefficients with Total Property Value:
Building Value                0.993860
Estate Zipcode_2116           0.465002
Num Floors                    0.436926
Exterior Type_Brick           0.429794
Property Area                 0.406195
                                 ...
```

This code snippet helps in understanding the relationship between each feature and the target variable, assisting in feature selection, or identifying key predictors for modeling.

The sorted correlation coefficients provide insight into which features are most strongly correlated (positively or negatively) with the target variable.

4) Plot to Visualize Correlation:



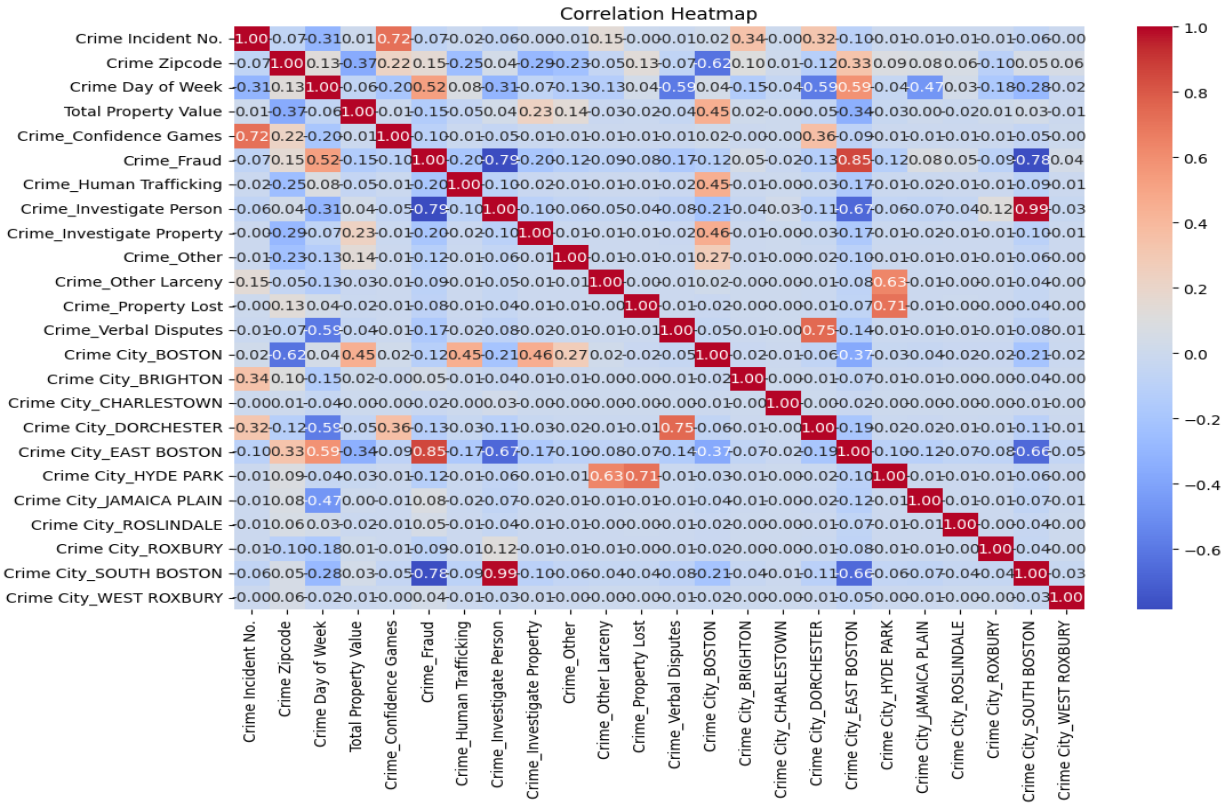
Top 5 Strongest Attributes:

['Building Value', 'Estate Zipcode_2116', 'Num Floors', 'Exterior Type_Brick', 'Property Area']

6) Selected the Crime related attributes

```
crime_features = (['Crime Incident No.','Crime','Crime City',
                   'Crime Zipcode', 'Crime Day of Week','Total Property Value'])
selected_data = data[crime_features]
column_types = selected_data.dtypes
column_types
```

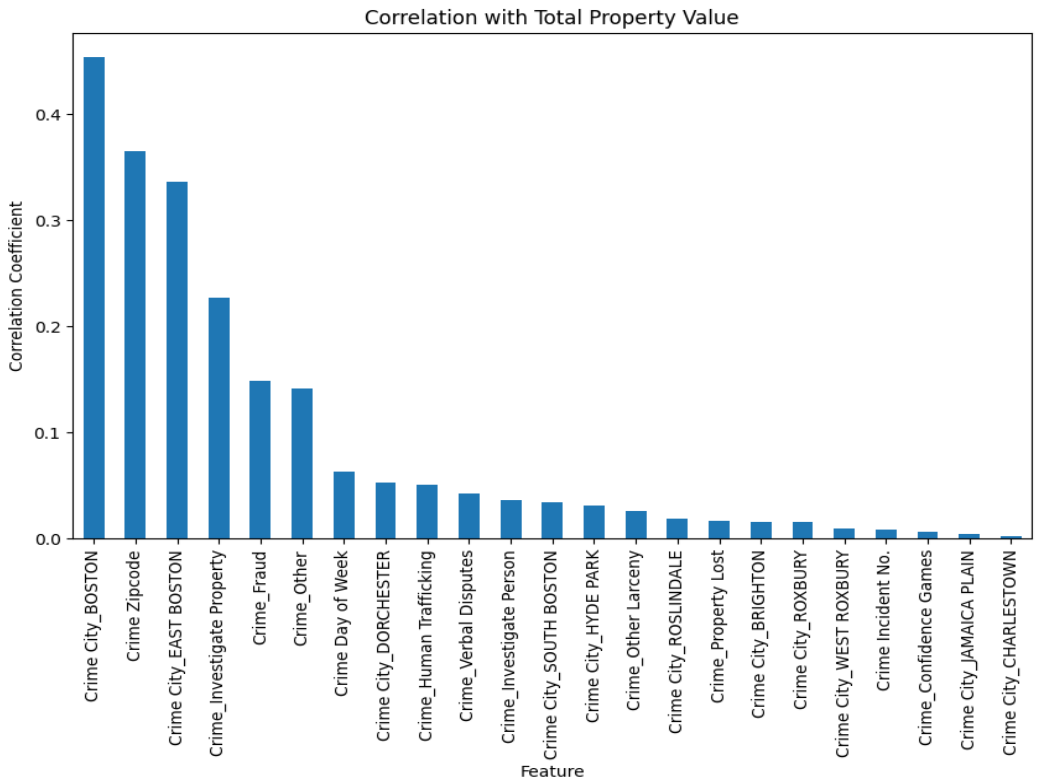
7) Correlation Heatmap



8) Sorted correlation coefficient

```
Correlation Coefficients with Total Property Value:
Crime City_BOSTON              0.453606
Crime Zipcode                  0.365281
Crime City_EAST BOSTON         0.335957
Crime_Investigate Property     0.226399
Crime_Fraud                    0.148259
Crime_Other                    0.140796
Crime Day of Week              0.062335
Crime City_DORCHESTER          0.052924
Crime_Human Trafficking        0.050500
Crime_Verbal Disputes          0.041951
Crime_Investigate Person       0.035796
Crime City_SOUTH BOSTON        0.033701
Crime City_HYDE PARK           0.030952
Crime_Other Larceny            0.025502
Crime City_ROSLINDALE          0.018024
Crime_Property Lost            0.016034
Crime City_BRIGHTON            0.015542
```

9) Plot to Visualize correlation

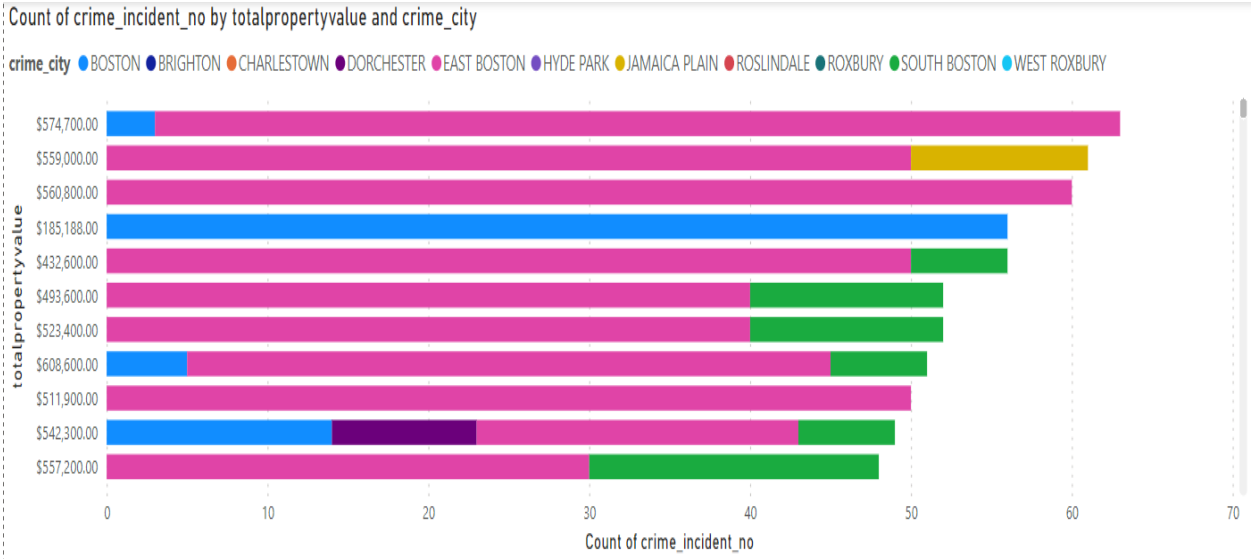


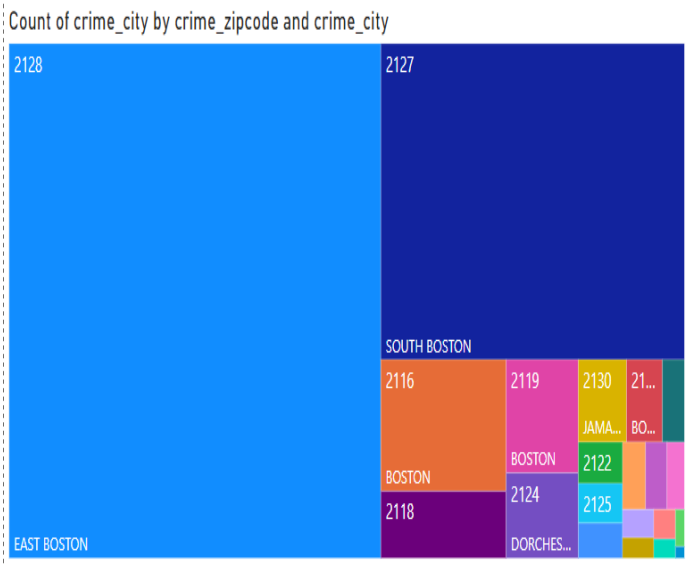
Top 5 Strongest Attributes:

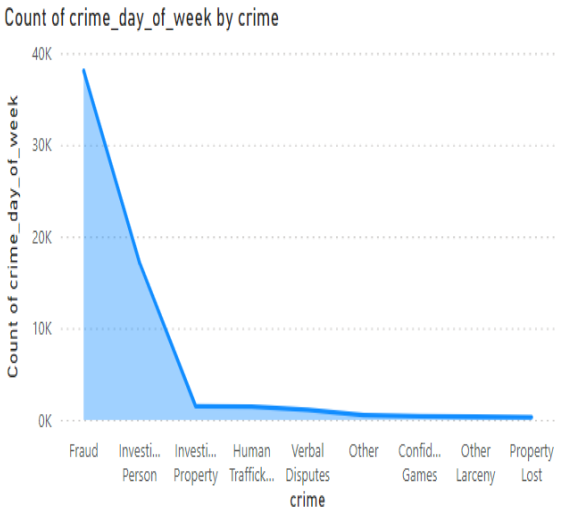
['Crime City_BOSTON', 'Crime Zipcode', 'Crime City_EAST BOSTON', 'Crime_Investigate Property', 'Crime_Fraud']

Validating attributes through Exploratory Data Analysis:

1) Crime Analysis by City

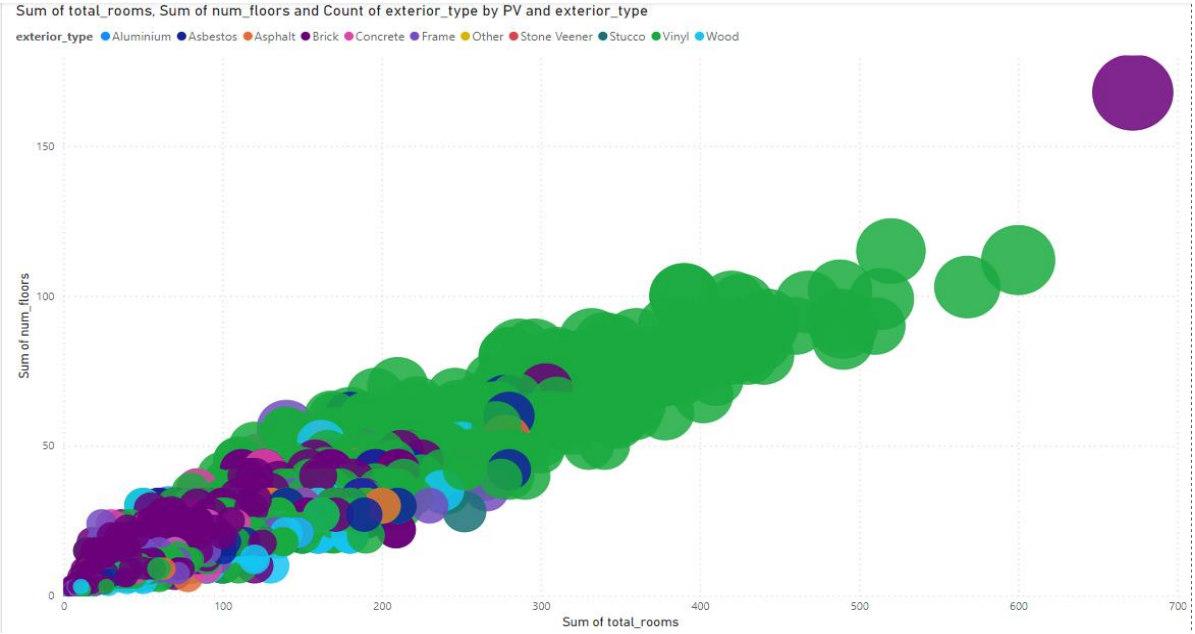






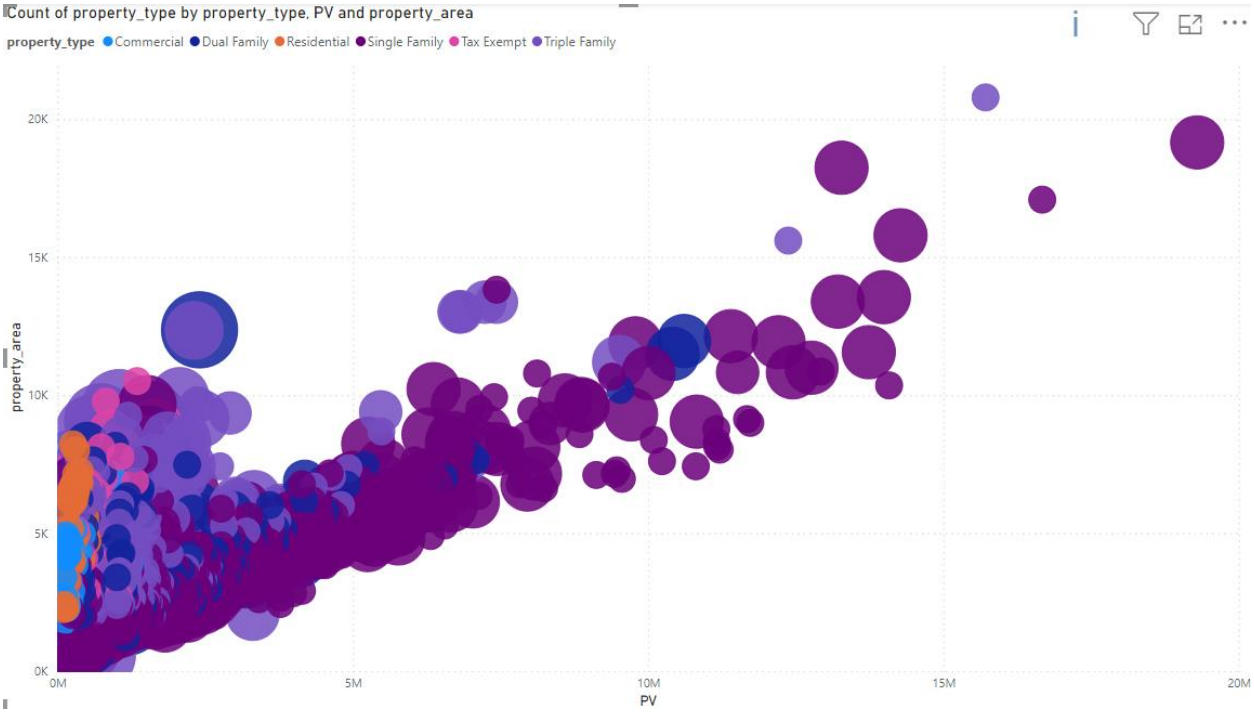
According to this we can conclude that East Boston, Boston, and South Boston have the most amount of crime. Around 40K reports were filed for Fraud, followed by 20K in Investigate person and property crime by week. This Validates the strong correlation shown before.

2) Number of Floors, Total Rooms and Exterior Type and Property Value



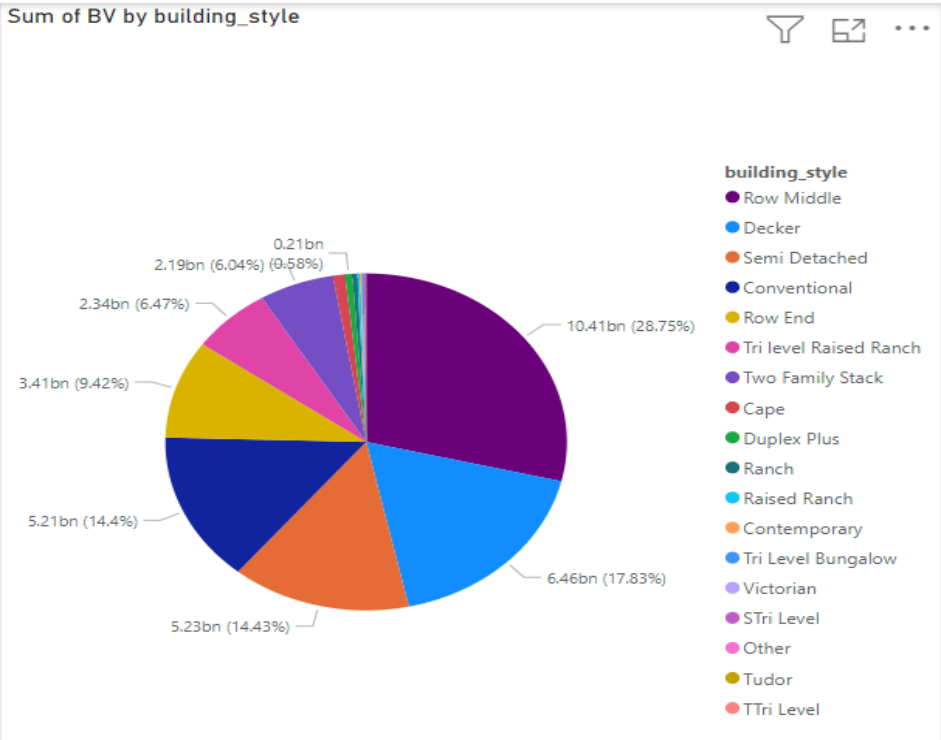
We can infer that the number of Bedrooms, Bathrooms, Kitchens, and number of floors have strong positive collinearity impacting the Property price. Majority of the prices ranging from 300K - 500K have an exterior type of Brick and Vinyl.

3) Property Type, Property area and Property Price



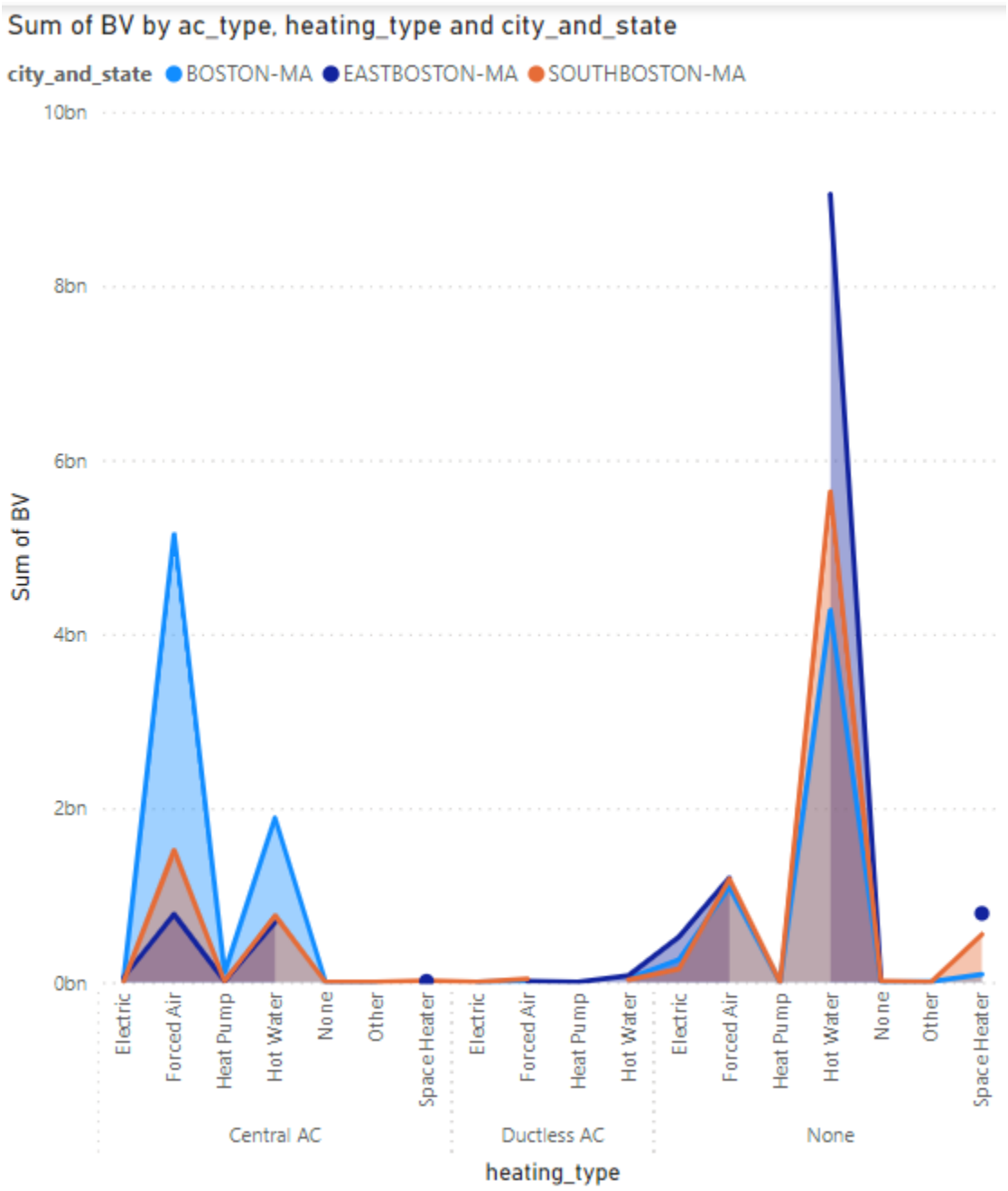
We can infer that property area and property price are positively correlated with most common property types being - Single Family, Triple Family and Dual Family. This validates the above correlation.

4) Building Style and Building Value



This indicates that Row Middle, Decker, Semi Detached and conventional are the top 4 building styles making up the majority of the Building Value Price.

5) AC and Heating Type based on City-State



This indicates that buildings without AC typically use Hot Water as a heating source followed by forced air (East Boston). Buildings with Central AC use forced air, followed by Hot water as a heating source (Central Boston). Ductless AC is very rarely used.

Conclusions:

1. Location Influence: Areas such as East Boston, Boston, and South Boston with higher crime rates may have lower property values compared to areas with lower crime rates. However, other factors such as amenities, schools, and proximity to city centers also play a significant role.
2. Property Condition and Security: Properties with lower crime rates and safer neighborhoods may attract higher property values. Investing in security measures and crime prevention can positively impact property values.
3. Interior Features: Properties with more bedrooms, bathrooms, kitchens, and floors tend to have higher property values. These features provide greater comfort and functionality, appealing to potential buyers or renters.
4. Exterior Features: Properties with exterior types like brick and vinyl are associated with higher property values within the price range of 400K - 600K. These materials are often perceived as durable and aesthetically pleasing, contributing to higher market value.
5. Property Size and Type: Larger properties or properties with more square footage tend to have higher property values. Common property types such as single-family, triple-family, and dual-family homes are also associated with higher property values.
6. Building Style: Certain building styles such as Row Middle, Decker, Semi Detached, and conventional are more prevalent and may contribute to higher property values. These styles may offer unique architectural features or layouts that appeal to buyers.
7. Heating and Cooling Systems: Properties with central air conditioning and heating systems, especially those using forced air, may command higher property values due to

improved comfort and energy efficiency. Ductless AC systems are less common and may not significantly impact property values.

In summary, property features such as location, interior and exterior condition, size, type, building style, and heating/cooling systems all influence property values. Desirable features include ample bedrooms and bathrooms, quality exterior materials, larger square footage, preferred property types, appealing architectural styles, and efficient heating/cooling systems. Additionally, factors like safety, crime rates, and neighborhood amenities also play a crucial role in determining property values.

Recommendations:

1. Location-Based Prioritization:

In neighborhoods with higher crime rates like East Boston, Boston, and South Boston, prioritize properties in safer pockets within these areas. Emphasize features that contribute to neighborhood safety, such as proximity to police stations, well-lit streets, and community watch programs.

2. Security Enhancements:

Invest in security measures and features that mitigate crime risks, such as alarm systems, surveillance cameras, and secure entry points. Highlight these features to reassure potential buyers and enhance property value, especially in areas with higher crime rates.

3. Exterior Aesthetics and Durability:

Emphasize exterior features that enhance curb appeal and property value, such as brick or vinyl siding. These materials not only provide durability but also convey a sense of quality and maintenance, which can be attractive to buyers looking for long-term investments.

5. Property Size and Layout:

Highlight properties with flexible layouts and ample space, catering to diverse buyer preferences. Larger properties with multiple bedrooms, bathrooms, and floors may appeal to families or individuals seeking spacious living environments.

6. Building Style and Architectural Appeal:

Consider the architectural style of properties and prioritize those with timeless designs and attractive features. Row Middle, Decker, Semi Detached, and conventional building styles, as identified in your analysis, may resonate well with buyers, and contribute to higher property values.

7. Heating and Cooling Systems for Comfort:

Ensure properties are equipped with efficient heating and cooling systems to provide comfort year-round. Highlight properties with central air conditioning and heating systems, especially those using forced air, as they contribute to a comfortable living environment irrespective of neighborhood crime rates.

Research Question 2: Property Age and Value Analysis

What is the relationship between the age of properties (year built or remodeled) and property values in various neighborhoods? How do property age and remodeling affect property values over time?

Methods: Regression analysis to examine the impact of property age on property values, time series analysis to assess trends in property values over time, and spatial analysis to identify geographic variations in the relationship between property age and values.

Variables: Property Age, Neighborhood, Property Value, Property Features

Managerial Decision-making: This analysis addresses a managerial question concerning real estate valuation and investment strategy, particularly in neighborhoods with varying property age dynamics. It aims to understand how property age, whether in terms of year built or remodeled, influences property values across different neighborhoods. The central managerial problem this analysis seeks to answer is: "How should stakeholders prioritize property age and remodeling to maximize property value in neighborhoods with varying property age dynamics?"

To understand the connection, if any, between the age of the properties and their values, we first must define property age. The age was calculated on two basis-

1. Property Age on Built- Calculated based on current year, i.e. 2024, to the year that property was built.

2. Property Age on Remodeled- We also ran a calculation to determine the age of the property from 2024 to the year it was last remodeled. In instances where there was no data for the year it was remodeled, the corresponding value from Property Age on Built was filled.

Correlation:

A correlation was run between three variables, namely – Total Property Value, Property Age on Built and Property Age on Remodeled. The results obtained were as follows-

Correlation Matrix:

	Total Property Value	Property Age (Built)	Property Age (Remodeled)
Total Property Value	1.000000	0.142213	-0.219361
Property Age (Built)	0.142213	1.000000	0.219037
Property Age (Remodeled)	-0.219361	0.219037	1.000000

a. Total Property Value and Property Age on Built-
    i. The correlation coefficient between Total Property Value and Property Age (Built) is 0.142213.
    ii. This indicates a weak positive correlation between the two variables.
    iii. As the property age based on the built year increases, there is a slight tendency for the total property value to increase as well.
b. Total Property Value and Property Age on Remodeled-
    i. The correlation coefficient between Total Property Value and Property Age (Remodeled) is -0.219361.
    ii. This suggests a weak negative correlation between the two variables.

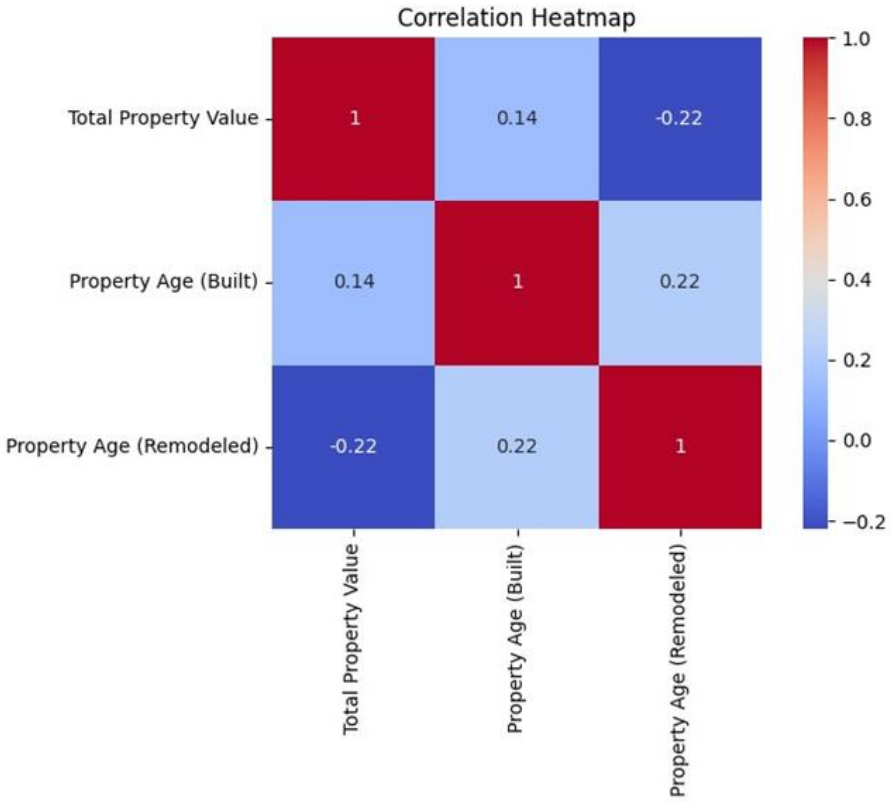
iii. As the property age based on the remodeled year increases, there is a slight tendency for the total property value to decrease.

c. Property Age on Built and Property Age on Remodeled:

i. The correlation coefficient between Property Age on Built and Property Age on Remodeled is 0.219037.

ii. This indicates a weak positive correlation between the two variables.

iii. Properties that are older based on the built year tend to have a slightly higher age based on the remodeled year as well.



The results project correlation coefficients that are relatively low in magnitude, suggesting weak correlations between the variables. The correlations are not strong enough to indicate a

substantial linear relationship between the variables. However, even weak correlations can still provide useful insights. For example:

a. The positive correlation between Total Property Value and Property Age (Built) suggests that older properties tend to have slightly higher values, although the relationship is not strong.
b. The negative correlation between Total Property Value and Property Age (Remodeled) indicates that properties with more recent remodeling tend to have slightly higher values compared to those with older remodeling, but again, the relationship is weak.

Linear Regression with only one Independent Variable-

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.477
Model:                            OLS   Adj. R-squared (uncentered):              0.477
Method:                 Least Squares   F-statistic:                          5.634e+04
Date:                Tue, 07 May 2024   Prob (F-statistic):                        0.00
Time:                        09:26:23   Log-Likelihood:                     -9.2898e+05
No. Observations:               61730   AIC:                                  1.858e+06
Df Residuals:                   61729   BIC:                                  1.858e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          6612.5376     27.858    237.366      0.000    6557.936    6667.139
==============================================================================
Omnibus:                    81103.799   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         20811147.606
Skew:                           7.386   Prob(JB):                         0.00
Kurtosis:                      91.730   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

We first investigated the relationship between the Total Property Value and Property Age on Built by running the Ordinary Least Square (OLS) method. The model gave the following insights-

a. The 'coef' column shows the estimated coefficient for the independent variable ('Property Age'). The coefficient value is 6612.5376, indicating that for each additional year of property age, the total property value is expected to increase by approximately $6,612.54, assuming all other factors remain constant.
b. The t-statistic ('t') and the associated p-value ('P>|t|') test the null hypothesis that the coefficient is equal to zero. In this case, the p-value is close to zero, suggesting that the coefficient is statistically significant.
c. The confidence interval ([0.025, 0.975]) provides a range within which the true coefficient value is likely to fall with a 95% probability.

Multiple Linear Regression with two Independent Variables

Regression Results:

OLS Regression Results

Dep. Variable:	Total Property Value	R-squared:	0.086
Model:	OLS	Adj. R-squared:	0.086
Method:	Least Squares	F-statistic:	2909.
Date:	Tue, 07 May 2024	Prob (F-statistic):	0.00
Time:	10:08:28	Log-Likelihood:	-9.2661e+05
No. Observations:	61730	AIC:	1.853e+06
Df Residuals:	61727	BIC:	1.853e+06
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.081e+05	1.38e+04	29.484	0.000	3.81e+05	4.35e+05
Property Age (Built)	5944.2499	117.293	50.679	0.000	5714.355	6174.144
Property Age (Remodeled)	-4288.4190	64.268	-66.727	0.000	-4414.385	-4162.453

Omnibus:	82107.269	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23397109.002
Skew:	7.518	Prob(JB):	0.00
Kurtosis:	97.183	Cond. No.	618.

The provided output shows the results of a multiple linear regression analysis using Ordinary Least Squares (OLS) method. The dependent variable is "Total Property Value," and the independent variables are "Property Age (Built)" and "Property Age (Remodeled)." The findings were as follows-

1. Model Summary:
    a. The R-squared value is 0.086, indicating that approximately 8.6% of the variation in the total property value can be explained by the independent variables (property age based on built year and remodeled year).
    b. The adjusted R-squared value is also 0.086, which accounts for the number of independent variables in the model.
    c. The F-statistic of 2909 and the associated p-value (Prob (F-statistic)) of 0.00 suggest that the overall model is statistically significant.

2. Coefficients:
    a. The constant term (const) is 4.081e+05, representing the expected total property value when both property ages (built and remodeled) are zero.
    b. The coefficient for "Property Age (Built)" is 5944.2499, indicating that for each additional year since the property was built, the total property value increases by approximately $5,944, holding other variables constant.
    c. The coefficient for "Property Age (Remodeled)" is -4288.4190, suggesting that for each additional year since the property was remodeled, the total property value decreases by approximately $4,288, holding other variables constant.

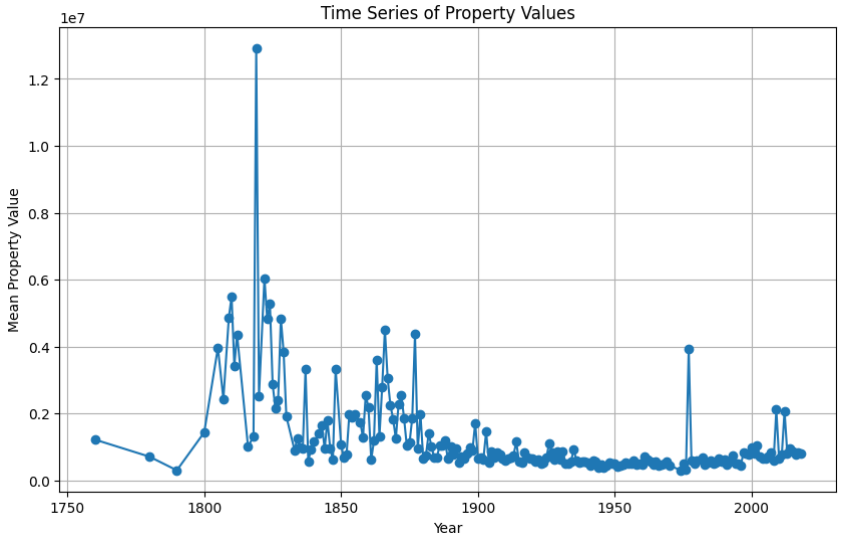
d. The standard errors (std err) quantify the average deviation of the coefficient estimates from the actual values.
e. The t-statistics (t) and the associated p-values (P>|t|) test the null hypothesis that the coefficients are equal to zero. In this case, all coefficients have p-values close to zero, indicating that they are statistically significant.
f. The confidence intervals ([0.025, 0.975]) provide the range within which the true coefficient values are expected to fall with a 95% probability.

3. Model Diagnostics:
   a. The Omnibus test and the associated p-value (Prob(Omnibus)) suggest that the residuals are not normally distributed.
   b. The Durbin-Watson statistic of 2.002 indicates no significant autocorrelation in the residuals.
   c. The Jarque-Bera (JB) test and the associated p-value (Prob(JB)) also indicate that the residuals are not normally distributed.
   d. The skewness (Skew) and kurtosis (Kurtosis) values further confirm the non-normality of the residuals.
   e. The condition number (Cond. No.) of 618 suggests that there might be some multicollinearity between the independent variables, but it is not severe.

Based on these results, we can conclude that both property age based on the built year and remodeled year have a statistically significant impact on the total property value. However, the model explains only a small portion of the variation in the property value (R-squared = 0.086),

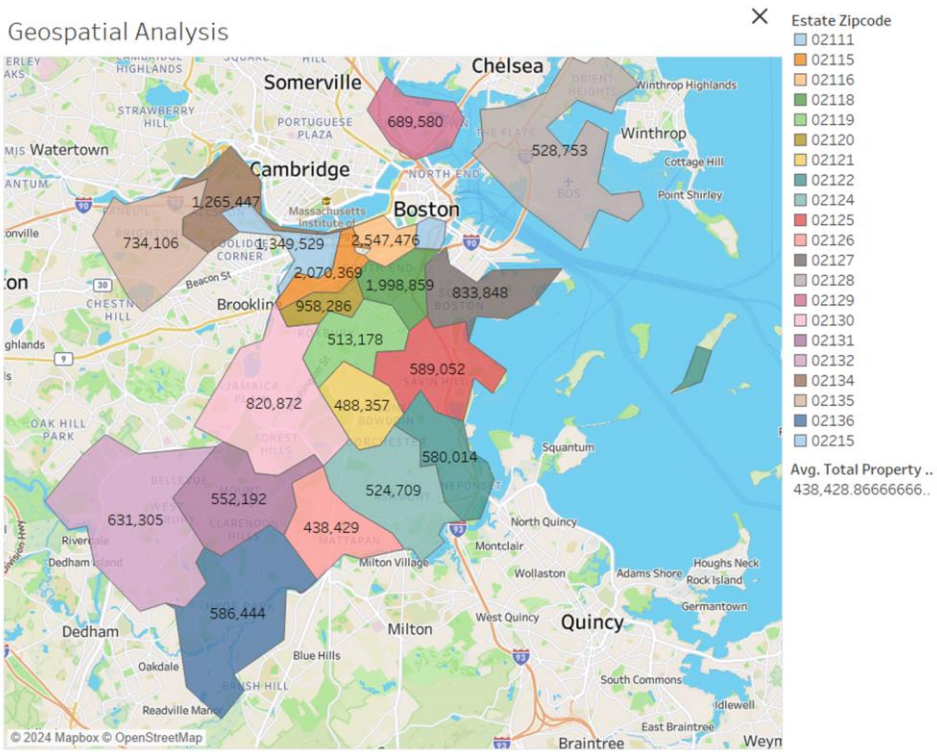
indicating that there are other important factors not included in the model that influence the property value.

Time Series Analysis-



The above chart plots the actual property values from year 1750 to 2000. There is a lot of year-to-year variability in the values, making it challenging to discern a clear long-term trend visually. The values fluctuate up and down frequently. There seems to be a slight dip in values in the mid 1800s followed by an increase, but the high variability makes the overall trend unclear from this plot alone.

Geospatial Analysis-



The above map is centered around Boston and its surrounding neighborhoods, with each region color-coded based on its corresponding estate zip code. The color intensity of each region represents the average total property value within that zip code. Darker shades indicate higher average property values, while lighter shades represent lower average values.

The zip codes and their respective average total property values are displayed as labels on the map. For example, zip code 02199 (Back Bay) has an average total property value of $2,547,476, while zip code 02136 (Hyde Park) has an average of $552,192.

The highest average total property values are concentrated in the central areas of Boston, particularly in neighborhoods such as Back Bay, Beacon Hill, and the Financial District. These

areas have average property values exceeding $2 million. Surrounding neighborhoods like Cambridge, Somerville, and Brookline also exhibit relatively high average property values, ranging from around $800,000 to over $1 million.

As we move further away from the city center, the average property values tend to decrease. Neighborhoods like Dedham, Hyde Park, and Revere have lower average property values compared to the central areas.

Summary of Analyses:

The analysis investigated the relationship between property age (based on both built and remodeled years) and property values across various neighborhoods. The findings revealed weak correlations between property age and total property value, with older properties generally showing a slight increase in value over time based on the built year, while more recently remodeled properties tend to have slightly lower values. The multiple regression analysis further confirmed the significant impact of property age on total property value, explaining about 8.6% of the variation in property values. Geospatial analysis highlighted variations in property values across different neighborhoods, with central areas like Back Bay and Beacon Hill exhibiting higher average property values compared to surrounding areas.

Conclusion:

Property age, whether based on built or remodeled years, plays a significant but nuanced role in determining property values. While older properties may show a slight increase in value over time, more recent remodeling tends to have a negative impact on property values. However, the

influence of property age alone is limited, as other factors not included in the analysis also contribute to variations in property values. Geospatial analysis further emphasizes the importance of location, with central areas generally commanding higher property values compared to outlying neighborhoods.

Recommendations:

1. Prioritize Renovations Strategically

When investing in property renovations, consider the potential impact on property values. Aim for renovations that enhance property appeal and functionality without significantly altering the property's character, as drastic remodeling may negatively affect value.

2. Location Matters

Place emphasis on properties in desirable neighborhoods with a history of stable or increasing property values. Central areas with strong amenities, infrastructure, and proximity to employment centers tend to offer better investment prospects.

3. Diversify Investments

Recognize that property age dynamics vary across neighborhoods. Diversify investment portfolios to include properties in different areas with varying age profiles to mitigate risks associated with fluctuations in property values.

4. Continuous Monitoring and Adaptation

Regularly monitor market trends and property values in targeted neighborhoods. Stay flexible and adapt investment strategies based on changing market conditions and emerging opportunities.

5. Consider Long-Term Value

While short-term fluctuations in property values may occur, focus on long-term appreciation potential. Invest in properties with characteristics that are likely to retain or increase in value over time, such as strong demand, limited supply, and favorable location attributes.

Research Question 3: Tax and Crime Rate Analysis

Is there a correlation between the gross tax of properties and crime rates in the neighborhood, and how does this relationship impact real estate investment decisions?

Answering whether there is a correlation between the gross tax of properties and crime rates in neighborhoods hold significant potential for providing critical insights essential for real estate investment decision-making. This analysis aims to unravel the relationship between these two factors, offering valuable guidance to management in navigating the complex landscape of property investment.

The goal here is to give management a clearer picture of what neighborhoods might be less risky to invest in based on taxes and crime. If we can find places with lower taxes and less crime, that's where we'd want to put our resources. This analysis helps management make smarter decisions about where to invest their money in the real estate market.

This analysis facilitates the identification of investment opportunities that align with management's risk tolerance and investment objectives within the real estate market. By discerning patterns where higher property taxes coincide with lower crime rates, stakeholders can pinpoint areas ripe for investment. Armed with this knowledge, management can strategically position themselves to capitalize on promising opportunities while mitigating potential risks inherent in the property market.

Analysis- In this analysis, we'll start by preparing and cleaning a dataset containing variables such as gross tax, crime incidents, and property characteristics. Following this, we'll conduct exploratory data analysis to calculate summary statistics and visualize distributions and trends in

gross tax and crime incidents. Subsequently, we'll move on to correlation analysis to quantify the relationship between gross tax and crime incidents, utilizing visualizations to illustrate the strength and direction of this relationship. Then, we'll perform regression analysis to assess the impact of gross tax on real estate investment decisions while controlling for factors such as property characteristics and neighborhood demographics. Finally, we'll present our findings, summarizing key insights and using visualizations to effectively communicate results, ultimately providing recommendations for real estate investment decision-making based on our analysis.

1. **Data Preparation-** Although the initially distributed data to the group was relatively clean, additional preparation was necessary for regression analysis.

```
In [58]: # Remove dollar signs and convert to float for specified columns
         Real_estate['Gross Tax'] = Real_estate['Gross Tax'].replace('[\$,]', '', regex=True).astype(float)
         Real_estate['Total Property Value'] = Real_estate['Total Property Value'].replace('[\$,]', '', regex=True).astype(float)
         Real_estate['Land Value'] = Real_estate['Land Value'].replace('[\$,]', '', regex=True).astype(float)
         Real_estate['Building Value'] = Real_estate['Building Value'].replace('[\$,]', '', regex=True).astype(float)
```

Some columns, including Gross Tax, Total Property Value, Land Value, and Building Value, contained dollar signs within their values. These dollar signs were removed to ensure accuracy and suitability for analysis, and the affected columns were converted to the appropriate data type.

Next, we selected the relevant attributes required for our analysis. These attributes included Gross Tax, Crime Incident Number, Crime City, Crime Zip code, Crime Offense Description, Land Size, Property Area, Living Area, Total Property Value, Land Value, Building Value, Crime, and Property Type.

```
# List of relevant columns
relevant_columns = ['Gross Tax', 'Crime Incident No.','Crime City','Crime Zipcode' ,'Crime Offense Desc.','Land size', 'Property

# Filter out only the relevant columns from the dataset
Real_estate = Real_estate[relevant_columns]

# Verify remaining columns
print("\nRemaining columns after selecting relevant columns:")
print(Real_estate.columns.tolist())
```

```
Remaining columns after selecting relevant columns:
['Gross Tax', 'Crime Incident No.', 'Crime City', 'Crime Zipcode', 'Crime Offense Desc.', 'Land size', 'Property Area', 'Living
Area', 'Total Rooms', 'Total Bedrooms', 'Total FullBaths', 'Total Halfbaths', 'Total Kitchens', 'Estate Zipcode', 'Total Proper
ty Value', 'Land Value', 'Building Value', 'Crime', 'Property Type']
```

These variables were essential for investigating the correlation between gross tax and crime rates and their impact on real estate investment decisions.

After selecting the relevant attributes, we checked for any empty values and removed them to ensure the integrity of our dataset. Once the data was cleaned, we created a new column named 'Crime Count', which contains information about the count of crimes that occurred in each city. This new column is crucial for our analysis as it provides insights into the frequency of crime incidents in different areas, thereby enriching our understanding of the relationship between crime rates and other variables.

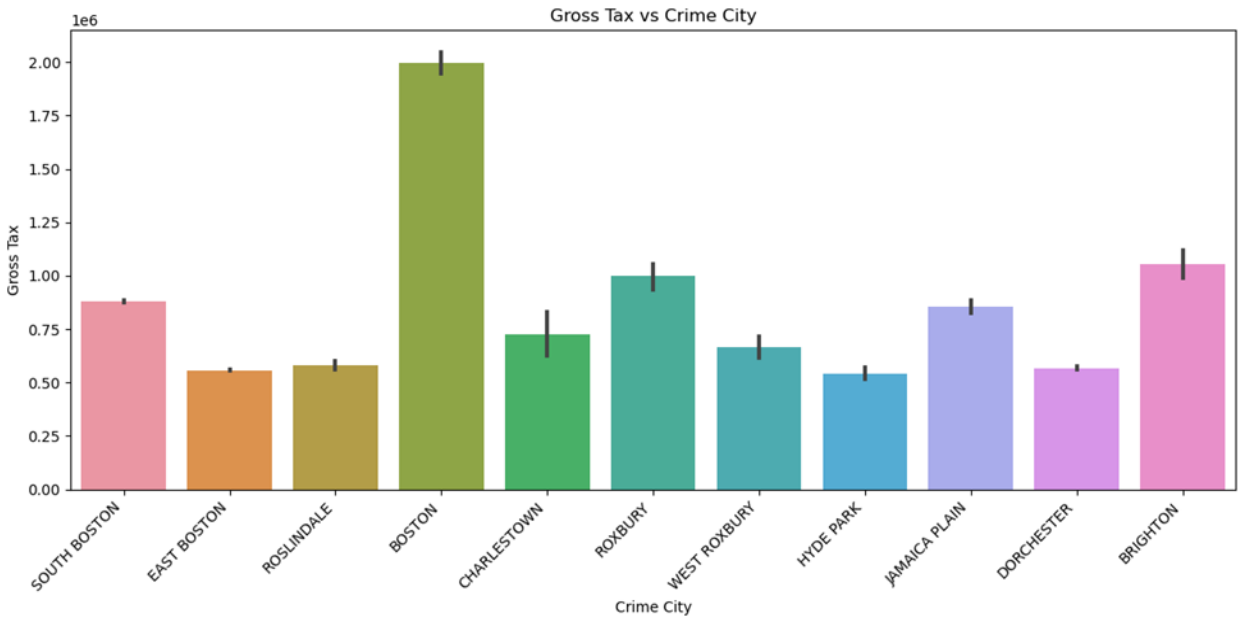
2. Correlation Analysis- Then, we conducted a correlation analysis specifically between gross tax and crime count. The correlation coefficient came out to be -0.37, which indicates a moderate negative correlation between gross tax and crime count.

```
import pandas as pd

# Calculating the correlation coefficient
correlation = Real_estate['Gross Tax'].corr(Real_estate['Crime Count'])

print("Correlation coefficient between Gross Tax and Crime Count:", correlation)
```

```
Correlation coefficient between Gross Tax and Crime Count: -0.3666407683737282
```

This suggests that the crime count decreases as the gross tax increases. While not extremely strong, this association level implies some relationship between these two variables. The negative correlation suggests that areas with higher property taxes may have lower crime rates and areas with lower property taxes may have higher crime rates.

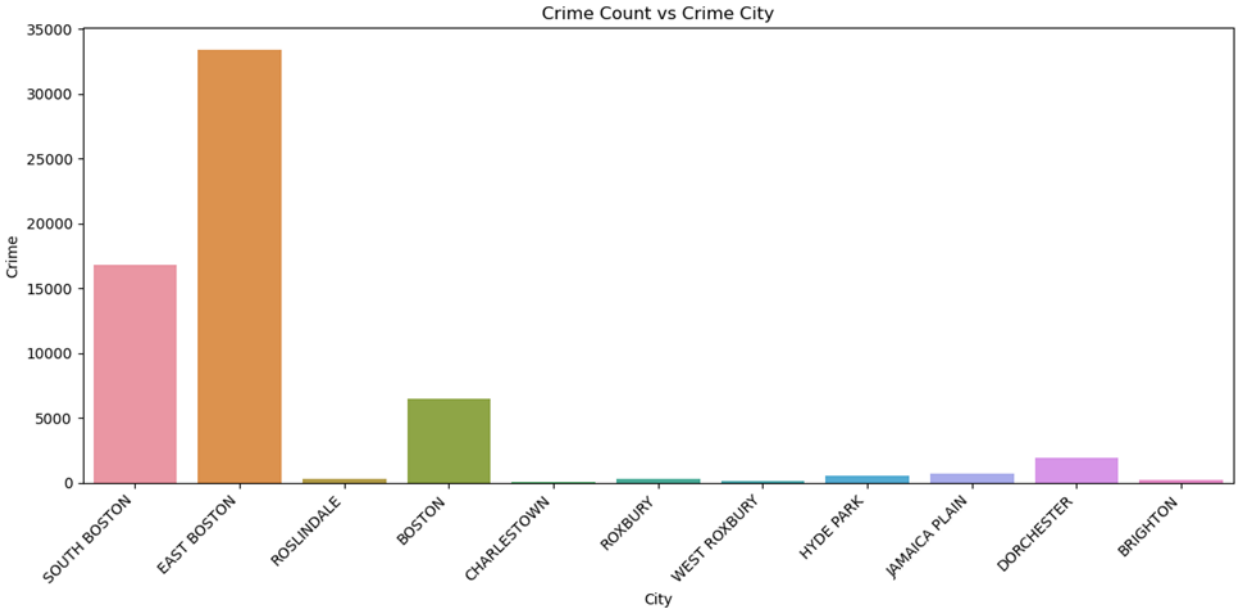
- To further analyze this, we created bar charts of gross tax and crime counts depending per city. The visualization between City vs Gross Tax highlights significant disparities in gross tax values across neighborhoods, with "Boston" and "Charlestown" notably having the highest taxes. In contrast, "Brighton" and others show comparatively lower values.



The presence of error bars suggests variability in tax data, possibly due to economic conditions or administrative differences.

- The graph between Crime City vs. Crime Count illustrates a notable discrepancy in crime counts among different cities, with East Boston recording the highest count followed by

South Boston, suggesting a higher susceptibility to criminal activities in these areas compared to others displayed.



Conversely, Charlestown, Roxbury, West Roxbury, Hyde Park, Jamaica Plain, Dorchester, and Brighton demonstrate relatively low crime counts.

- After that, we created a table to better understand the relationship between Gross Tax and Crime rate in different cities.

	Crime City	Crime Count	Mean Gross Tax
0	BOSTON	6427	1997006
1	BRIGHTON	226	1056243
2	CHARLESTOWN	20	726817
3	DORCHESTER	1914	567523
4	EAST BOSTON	33390	557091
5	HYDE PARK	543	540539
6	JAMAICA PLAIN	693	853334
7	ROSLINDALE	252	582010
8	ROXBURY	324	997157
9	SOUTH BOSTON	16758	877704
10	WEST ROXBURY	154	665395

Based on that data, East Boston has the highest crime count of 33,390 incidents, contrasting sharply with Charlestown's mere 20 incidents. Interestingly, East Boston does not boast the highest mean gross tax despite its elevated crime rate, suggesting a nuanced relationship between tax burden and crime rates.

Boston leads with the highest mean gross tax of $1,997,006, indicating potential affluence or higher property values. Neighborhoods like West Roxbury and Charlestown, characterized by lower crime rates, also exhibit moderate to high mean gross taxes, hinting at possible safety and affluence. Further analysis, such as correlation studies between crime rates and tax values, could provide deeper insights into the dynamics at play.

- After that, we conducted Pearson and Spearman correlation to better understand the dynamic between Gross Tax and Crime Rate.

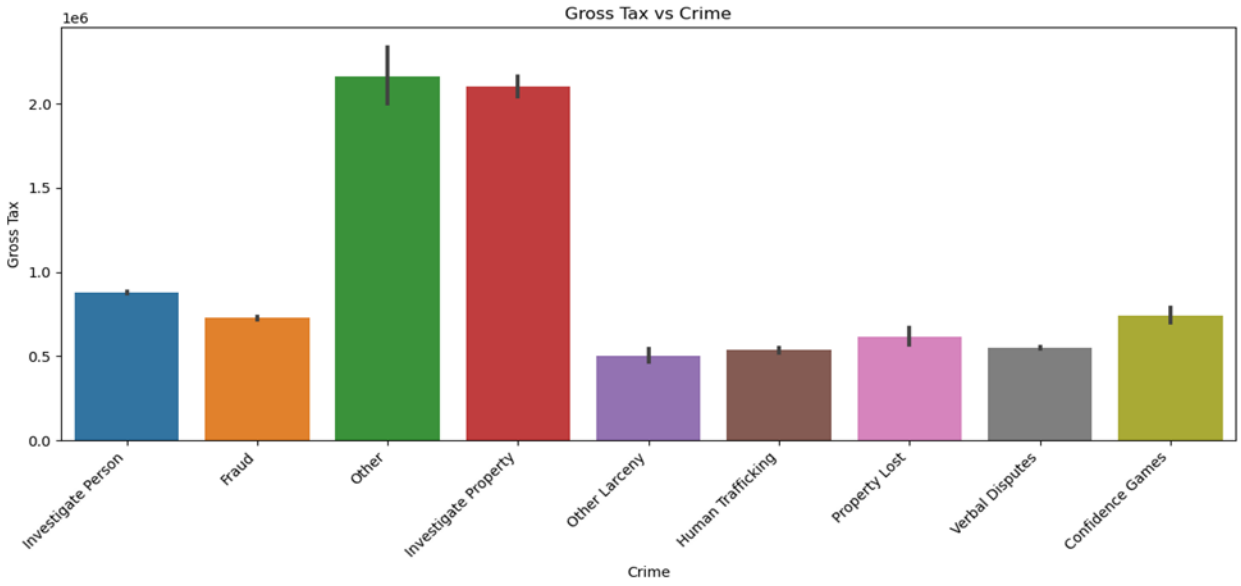
```
import pandas as pd
from scipy.stats import pearsonr, spearmanr

gross_tax = Real_estate['Gross Tax']
crime_rate = Real_estate['Crime Count']
# Calculate Pearson correlation coefficient and p-value
pearson_corr, pearson_pvalue = pearsonr(gross_tax, crime_rate)
# Calculate Spearman correlation coefficient and p-value
spearman_corr, spearman_pvalue = spearmanr(gross_tax, crime_rate)
print("Pearson correlation coefficient:", pearson_corr)
print("Pearson p-value:", pearson_pvalue)
print("\nSpearman correlation coefficient:", spearman_corr)
print("Spearman p-value:", spearman_pvalue)
```

```
Pearson correlation coefficient: -0.36664076837377946
Pearson p-value: 0.0

Spearman correlation coefficient: -0.4388670269871377
Spearman p-value: 0.0
```

The Pearson correlation coefficient, measuring the linear relationship between Gross Tax and Crime Count, reveals a moderate negative correlation of approximately -0.367. This indicates

that Crime Count tends to decrease as Gross Tax increases, and vice versa, though the association isn't exceedingly strong. Accompanied by a p-value of 0.0, the observed correlation is statistically significant, suggesting its likelihood beyond random chance. Similarly, the Spearman correlation coefficient, reflecting the monotonic relationship between the variables, demonstrates a moderate negative correlation of approximately -0.439. This highlights a consistent trend where higher Gross Tax corresponds to lower Crime Count, and vice versa. Both analyses underscore a significant negative association between Gross Tax and Crime Count, implying that areas with elevated Gross Tax tend to exhibit lower Crime Counts and vice versa.

- Then, to see if there is a relation between the type of crime and Gross tax, we created a bar graph.



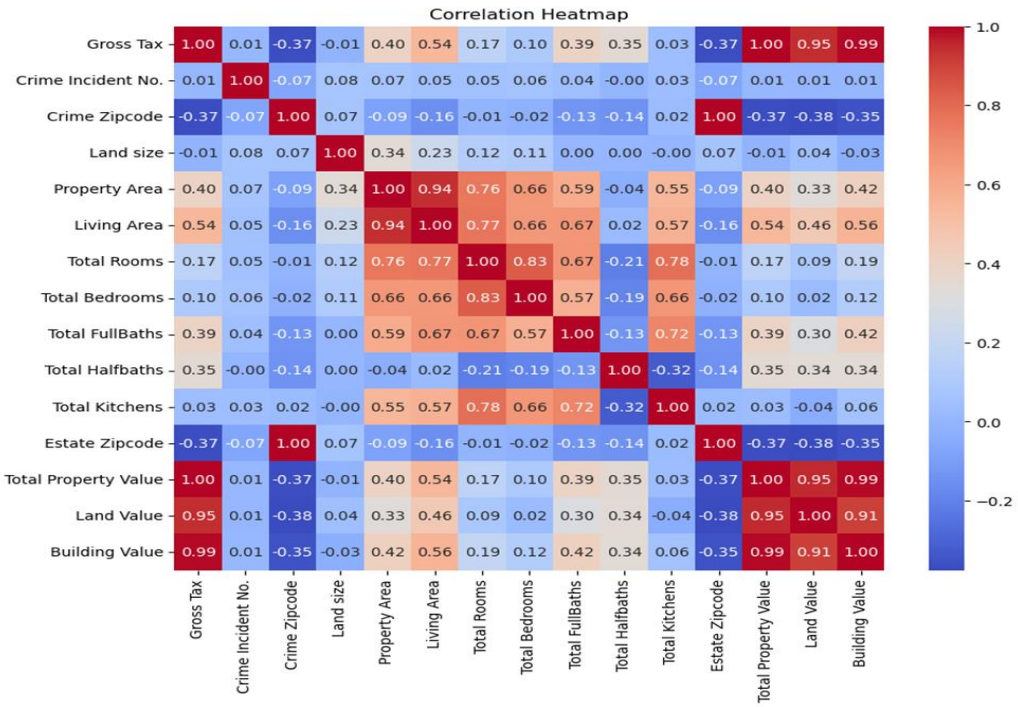
Analyzing gross tax values across crime categories reveals insights into the financial implications of different incidents. Crimes like "Other" and "Investigate Property" have the highest gross tax values, suggesting they require significant resources. "Other Larceny" and "Confidence Games" show moderate tax values, while "Investigate Person," "Fraud," and

"Verbal Disputes" have lower values. The error bars indicate variability in tax estimates per incident, emphasizing potential financial variation despite general trends.

- After that, we created a heat map to understand how different variables are interrelated.

Strong Positive Correlation:

- A correlation of 0.94 indicates a very strong positive relationship, suggesting that the property area tends to increase similarly as the living area increases.
- A correlation of 0.83 suggests that properties with more rooms generally have more bedrooms.
- The correlation of 0.67 indicates a moderately strong relationship, suggesting that properties with more full bathrooms also tend to have more half bathrooms.
- A high correlation of 0.99 shows that the building value is almost directly proportional to the total property value.



Strong Negative Correlations:

- There are no solid negative correlations observed in this heatmap. Most negative values are close to zero, indicating weak negative relationships.

Other Notable Observations:

- The correlation of 1.00 suggests that the gross tax is directly proportional to the total property value.
- The high correlation of 0.95 indicates that the land value significantly influences the total property value.

3. **Regression Analysis:** Following our correlation analysis, we utilized linear regression modeling, specifically the Ordinary Least Squares (OLS) method, to investigate the relationship between Gross Tax and various predictor variables. By splitting the dataset into training and testing sets and fitting the OLS model, we aimed to uncover nuanced patterns and associations within the data. Our analysis seeks to provide insights into how factors such as Crime Count, Property Area, and Total Property Value influence Gross Tax, guiding strategic decision-making in the real estate market. Additionally, we evaluated the model's performance metrics and visualized residuals to assess its adequacy in capturing underlying trends. This approach comprehensively explains the dynamics between Gross Tax and predictor variables, informing real estate investment decisions.

- **Code Snippets:**

```
import pandas as pd
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

X = Real_estate[[ 'Crime Count' , 'Crime Incident No.', 'Crime Zipcode', 'Land size', 'Property Area', 'Living Area', 'Total Room
y = Real_estate['Gross Tax']  # Dependent variable
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error (Out-of-sample):", mse)

print("Intercept:", model.intercept_)
print("Coefficient:", model.coef_[0])

r_squared = model.score(X_test, y_test)
print("R-squared (Out-of-sample):", r_squared)

X_train_with_const = sm.add_constant(X_train)
X_test_with_const = sm.add_constant(X_test)

# Fitting the model using statsmodels
stats_model = sm.OLS(y_train, X_train_with_const).fit()

# Getting the summary of the model
print(stats_model.summary())

# Making predictions using the stats model
y_train_pred = stats_model.predict(X_train_with_const)
y_test_pred = stats_model.predict(X_test_with_const)

# Calculating in-sample MSE
in_sample_mse = mean_squared_error(y_train, y_train_pred)
print("Mean Squared Error (In-sample):", in_sample_mse)

# Calculating R-squared value (In-sample)
r_squared1 = model.score(X_train, y_train)
print("R-squared (In-sample):", r_squared1)
```

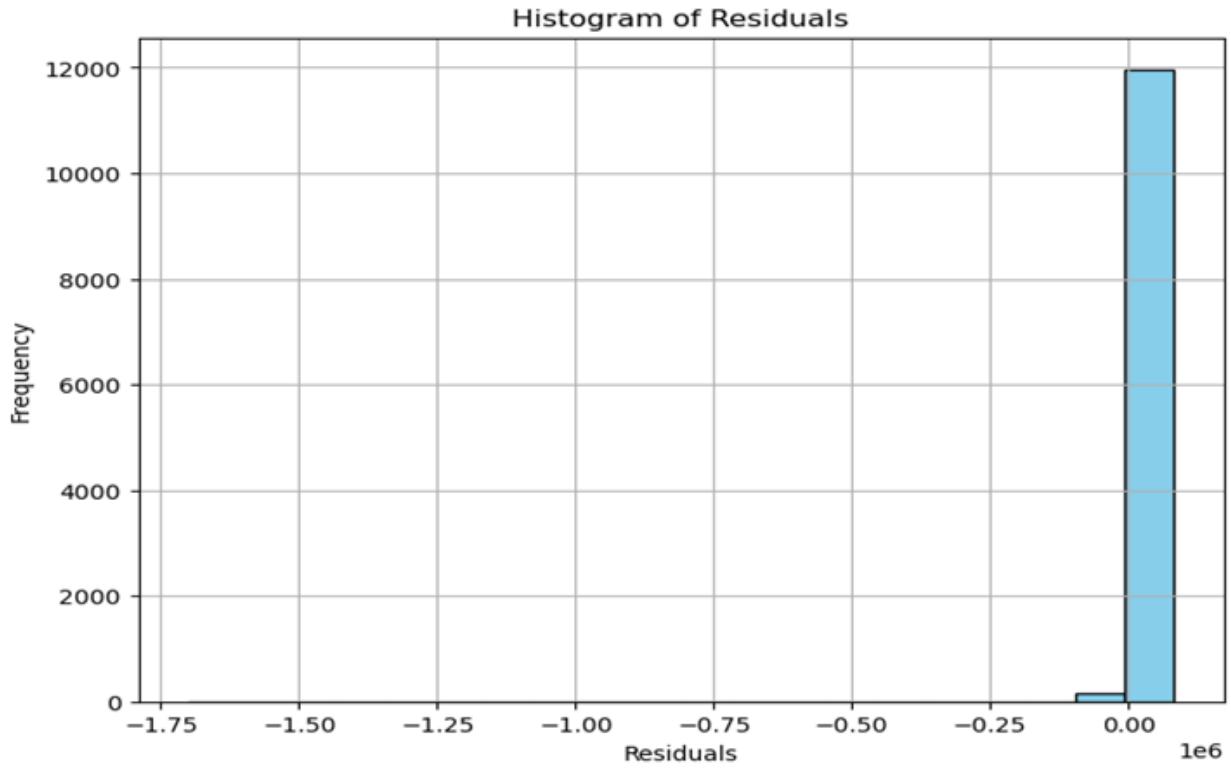
- **Results:**

OLS Regression Analysis			
Out of Sample		In Sample	
MSE	1673966855.61	MSE	1416704417.71
R-Squared	0.997457608	R-Squared	0.997915028
Variable	Std. Error	Coefficient	P-value
Crime Count	0.018	0.2153	0.001
Crime Incident No.	0	-0.0005	0.005
Crime Zipcode	20.879	-374.7162	0.001
Land size	0.097	0.2164	0.026
Property Area	0.413	0.2083	0.614
Living Area	0.647	-5.7538	0.001
Total Rooms	113.444	217.0454	0.056
Total Bedrooms	153.061	842.5191	0.001
Total FullBaths	321.045	418.0048	0.193
Total Halfbaths	349.899	-1238.5019	0.001
Total Kitchens	398.342	1508.4556	0.001
Estate Zipcode	20.879	-374.7182	0.001
Total Property Value	0.008	0.5655	0.001
Land Value	0.008	0.4869	0.001
Building Value	0.008	0.4924	0.001

The regression analysis uncovers critical drivers of property values, highlighting the impact of crime rates, property characteristics, and location. Higher crime rates, indicated by crime count and specific zip codes, correspond to lower property values, underlining the importance of safety in real estate valuation. Property size, indicated by land area and room count, positively influences property values, while surprisingly, larger living areas show a negative association, possibly due to higher maintenance costs. More bathrooms and kitchens increase property values, reflecting buyer preferences for convenience and

functionality. The negative coefficient for half bathrooms suggests a lesser impact on property values. Moreover, overall property, land, and building values significantly affect market prices. Conversely, estate zip codes show a negative association, indicating potential devaluation factors in certain areas.

- After that, we drew a histogram of the residuals of regression analysis, which allowed us to assess the model's performance and detect any patterns or biases in its predictions. The concentration of residuals around zero indicates overall accuracy, suggesting that the model predictions align closely with the actual data for most cases.



However, the skewness to the right and outliers suggest potential biases and areas where the model may overestimate values, particularly evident in the long-left tail of harmful residuals. These outliers could influence the model's performance and warrant further investigation.

Additionally, the non-symmetric distribution of residuals highlights potential deviations from model assumptions, suggesting opportunities for improvement through alternative modeling techniques or data transformations.

4. Summary:

- Correlation Analysis: The correlation analysis revealed a moderate negative correlation (-0.37) between property gross tax and crime rates. Areas with higher property taxes tend to exhibit lower crime rates, indicating a potential inverse relationship between tax burden and criminal activity.
- Regression Analysis: Regression analysis further elucidated the impact of various factors on property values. Crime rates, property characteristics (such as living area, number of rooms, total property value), and location (e.g., estate zip code) significantly influence property values.
- While larger living areas and higher property taxes were associated with lower crime rates, other factors like total bedrooms and kitchens positively affected property values.

5. Recommendations:

- **Identify Low-Risk Investment Opportunities:** Focus on neighborhoods with higher property taxes and lower crime rates, as these areas tend to offer safer investment options with potentially higher property values.

- **Consider Property Characteristics:** Pay attention to property characteristics such as living area, number of rooms, and total property value, as these factors play significant roles in determining property values.
- **Evaluate Location Factors:** Assess the impact of location factors, including estate zip codes, on property values. Invest in neighborhoods with favorable zip codes demonstrating lower crime rates and higher property values.
- **Mitigate Risks:** While higher property taxes may be associated with lower crime rates, conduct thorough due diligence to ensure that other factors, such as property condition and market trends, align with investment goals and risk tolerance.

Using the insights from the analyses, real estate investors can strategically pinpoint investment prospects that match their risk appetite and investment goals. Prioritizing neighborhoods with diminished crime rates, desirable property attributes, and advantageous location factors allows investors to refine their investment strategies and enhance potential returns within the dynamic real estate landscape.

Research Question 4: Trend in Property Prices Over Time

How have property prices in different neighborhoods of Boston trended over the past few years? What factors have influenced these trends, and are there any significant neighborhood variations?

Methods: Diagnostic Analysis, Descriptive Analysis, Geospatial Analysis and Comparative Analysis.

Variables/Attributes: Crime Neighborhood, Total Property Value, Property Built, Land Value, Building value, Gross tax, Estate Street No, Estate Zipcode, School Zipcode, Crime Incident year.

Managerial Decision Making: Analyzing property price trends in Boston over recent years offers valuable insights for real estate stakeholders, guiding strategic decision-making across various sectors. This examination of historical sales data provides a comprehensive understanding of market dynamics and future projections, aiding in questions surrounding market fluctuations, investment strategies, and overall planning. For instance, stakeholders can discern long-term price trends and regional disparities, facilitating targeted investment and resource allocation. Moreover, predictive capabilities enable stakeholders to anticipate future market movements based on neighborhoods, informing proactive investment planning and risk management initiatives.

In essence, the central managerial query addressed by this analysis is: "How can stakeholders utilize insights from Boston's property price trends in neighborhoods to optimize investment strategies, allocate resources effectively, and mitigate risks in the real estate market?"

Analysis: This comprehensive analysis leverages meticulously prepared data from the prior Data Preparation phase. Our dataset encompasses crucial variables, including crime incidence in neighborhoods, gross tax, building and land values, total property value, Estate zip code, and Estate Street Number.

We initiate a correlation analysis to unveil intricate relationships among these variables, shedding light on their mutual influences. Through insightful visualizations, we discern the direction and strength of these relationships, paving the way for predictive modeling and pattern recognition.

Subsequently, we conduct regression analysis to unravel the relationship between the dependent variable 'Total Property Values' and the independent variables. This enables us to anticipate fluctuations in property values based on shifts in independent variables, empowering us with forecasting capabilities, trend identification, hypothesis testing, and a nuanced understanding of variable impacts on outcomes.

We then delve into Descriptive and Exploratory Data Analysis to ensure a rigorous and data-driven approach. Descriptive analysis provides succinct summaries of key data features, such as central tendencies and dispersions, while exploratory analysis delves deeper into statistical techniques and visualizations to unearth underlying patterns, relationships, and anomalies.

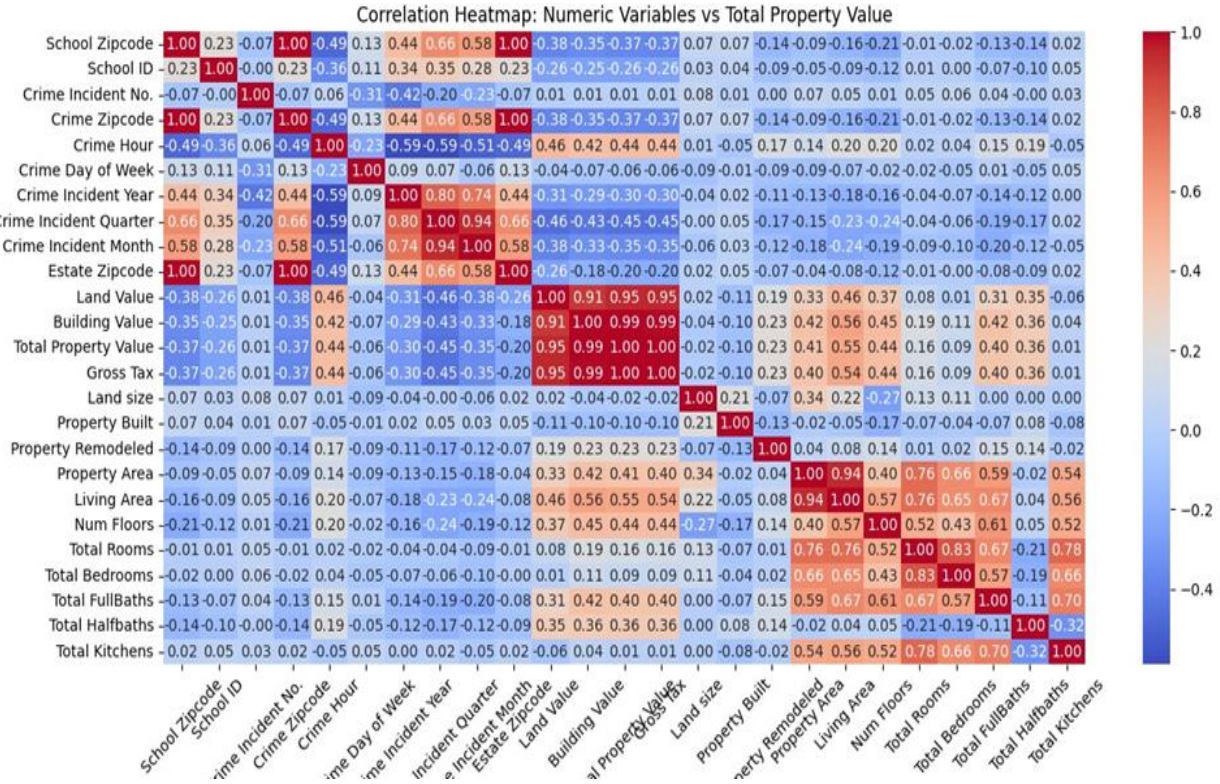
These analytical methodologies will provide invaluable insights, guiding subsequent analyses and informed decision-making processes. Armed with comprehensive findings and robust conclusions, we will be poised to offer actionable recommendations regarding neighborhood

factors influencing property values over time, thus empowering stakeholders in their investment strategies and decision-making endeavors.

Correlation Analysis:

1. To perform the correlation analysis between various neighborhood variables with Total Property Value, we generated a Heatmap to identify the variables with most strong positive coefficients.

Result:



Analysis:

Through the heatmap we can understand that the variable Land value has a strong positive correlation of 0.91 with Total Property value. Through the heatmap we can understand that the

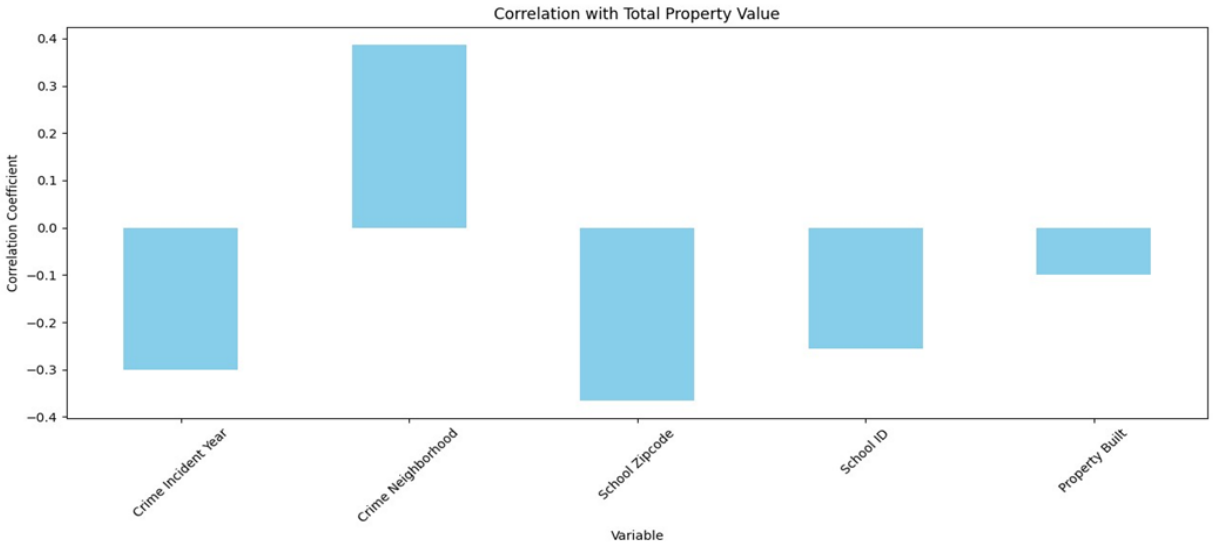
variable Building value has a very strong positive correlation of 0.95 with Total Property value. Through the heatmap we can understand that the variable Gross tax has a strong positive correlation of 0.99 with Total Property value.

2. Correlation between the chosen variables ('Crime Incident Year', 'Crime Neighborhood', 'School Zipcode', 'School ID', and 'Property Built') and 'Total Property Value':

Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder

# Read data from Excel
data = pd.read_excel('Final Blend.xlsx')

# Select relevant columns
selected_columns = ['Crime Incident Year', 'Crime Neighborhood', 'School Zipcode', 'School ID',
'Property Built', 'Total Property Value']
property_data = data[selected_columns]

# Label encoding for 'Crime Neighborhood'
label_encoder = LabelEncoder()
property_data['Crime Neighborhood'] = label_encoder.fit_transform(property_data['Crime
Neighborhood'])

# Calculate correlations
correlation_matrix = property_data.corr()['Total Property Value'].drop('Total Property Value')

# Plotting correlation coefficients
plt.figure(figsize=(8, 6))
correlation_matrix.plot(kind='bar', color='skyblue')
plt.title('Correlation with Total Property Value')
plt.xlabel('Variable')
plt.ylabel('Correlation Coefficient')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Results:



Analysis:

· Crime Neighborhoods with higher positive correlation coefficients are more positively associated with 'Total Property Value,' meaning it has a stronger influence on property value.

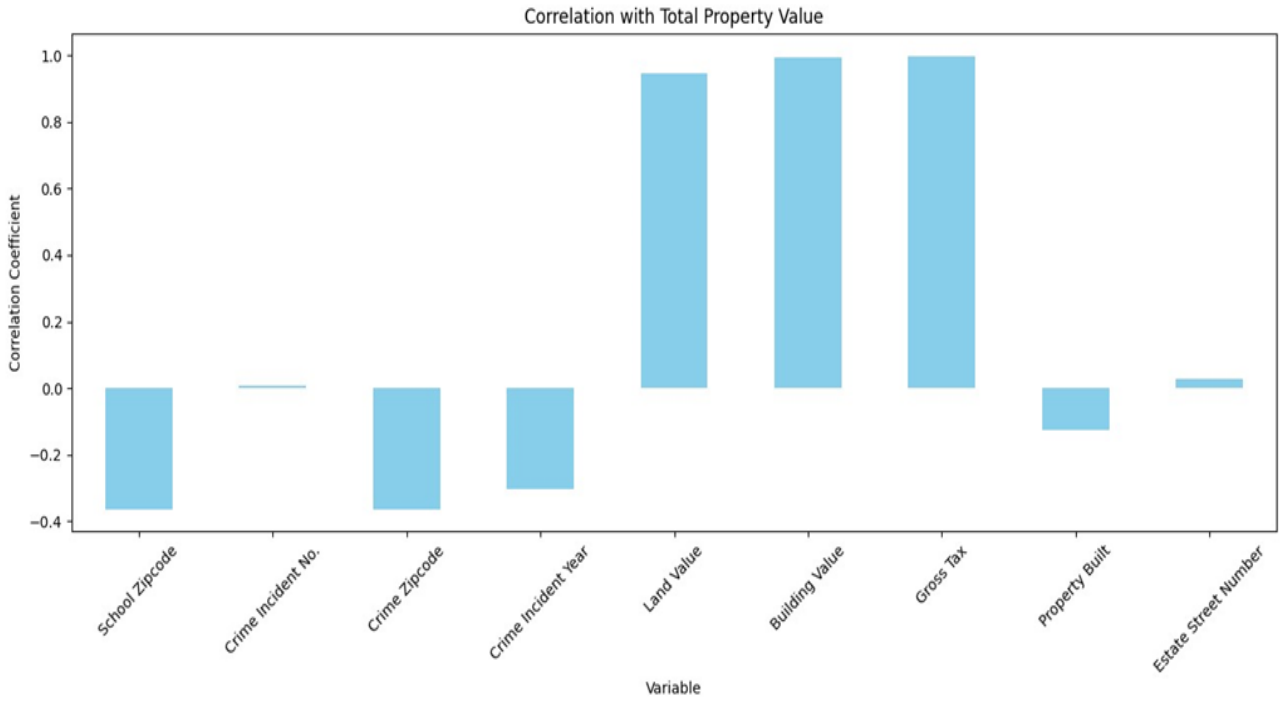
· Crime Incident Year, School Zipcode, School ID, Property Built with higher negative correlation coefficients are more negatively associated with 'Total Property Value', meaning they have a stronger negative impact on property value.

· Neighborhood schools have a negative coefficient, which means that the schools are not a factor that influences property prices.

3. Correlation between the chosen variables ('School Zipcode', 'Crime Incident No.', 'Crime Zipcode', 'Crime Incident Year', 'Land Value', 'Building Value', 'Gross Tax',

'Property Built', 'Estate Street Number', 'Total Property Value') and ‘Total Property Value’:

Result:



Analysis:

Gross Tax and Building Value (0.9 correlation):

These variables exhibit a strong positive correlation with Total Property Value.

It suggests that properties with higher Gross Tax and Building Value tend to have a higher Total Property Value.

This could imply that areas with higher taxes and more valuable buildings are associated with higher property values.

Land Value (0.85 correlation):

Land Value also shows a strong positive correlation with Total Property Value.
Similar to Gross Tax and Building Value, this suggests that properties with higher Land Value tend to have a higher Total Property Value.
It implies that the value of the land itself contributes significantly to the overall property value.

Estate Street Number (0.05 correlation):

Estate Street Number shows a very weak positive correlation with Total Property Value.
The correlation coefficient is close to zero, indicating almost no linear relationship between these variables.
It suggests that there is little to no influence of the Estate Street Number on the Total Property Value.
This could imply that the street number of the estate does not have a significant impact on property values.

Conclusion: Crime Neighborhoods, Gross Tax, Building Value, Land Value, and Estate Street Number are the 5 neighborhood variables influencing the Total Property value variable.

Regression Analysis:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	1							
R Square	1							
Adjusted R Square	1							
Standard Error	7.49143E-07							
Observations	61741							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	4.31885E+16	1.44E+16	2.57E+28	0			
Residual	61737	3.46478E-08	5.61E-13					
Total	61740	4.31885E+16						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	7.24623E-08	4.14818E-09	17.46846	3.64E-68	6.43318E-08	8.05927E-08	6.4332E-08	8.05927E-08
Land Value	1	8.00695E-14	1.25E+13	0	1	1	1	1
Building Value	1	7.37817E-14	1.36E+13	0	1	1	1	1
Gross Tax	-8.66113E-14	6.92657E-14	-1.25042	0.21115	-2.2237E-13	4.91495E-14	-2.224E-13	4.91495E-14

Regression Statistics:

Multiple R: The multiple correlation coefficient indicates a perfect linear relationship between the independent and dependent variables.

R Square: The coefficient of determination is also 1, indicating that 100% of the variability in the dependent variable (Total Property Value) is explained by the independent variables (Land Value, Building Value, and Gross Tax).

Adjusted R Square: The adjusted R square is also 1, suggesting that the model fits the data perfectly.

Standard Error: The standard error of the estimate is extremely small, indicating that the observed values closely fit the regression line.

Observations: There are 61741 observations in the dataset

ANOVA:

The ANOVA table tests the overall significance of the regression model.

The F-statistic is extremely large, with a p-value of 0, indicating that the regression model is highly significant.

Coefficients:

- The coefficients table shows the estimated regression coefficients for each independent variable.
- The intercept coefficient is 7.24623E-08, indicating the expected value of the dependent variable when all independent variables are zero.
- The coefficients for Land Value and Building Value are both 1, suggesting that they have a perfect positive linear relationship with the Total Property Value.
- The coefficient for Gross Tax is close to zero and statistically insignificant, suggesting that it may not have a significant impact on Total Property Value.

Based on these regression results, we can conclude the following regarding property prices in different neighborhoods of Boston:

- Property prices are perfectly explained by Land Value and Building Value, as indicated by their coefficients of 1. This suggests that these factors have a significant influence on property prices.
- Gross Tax does not appear to have a significant impact on property prices, as its coefficient is close to zero and statistically insignificant.
- There is no indication of neighborhood variations in the regression results, as the analysis does not include neighborhood-specific variables.

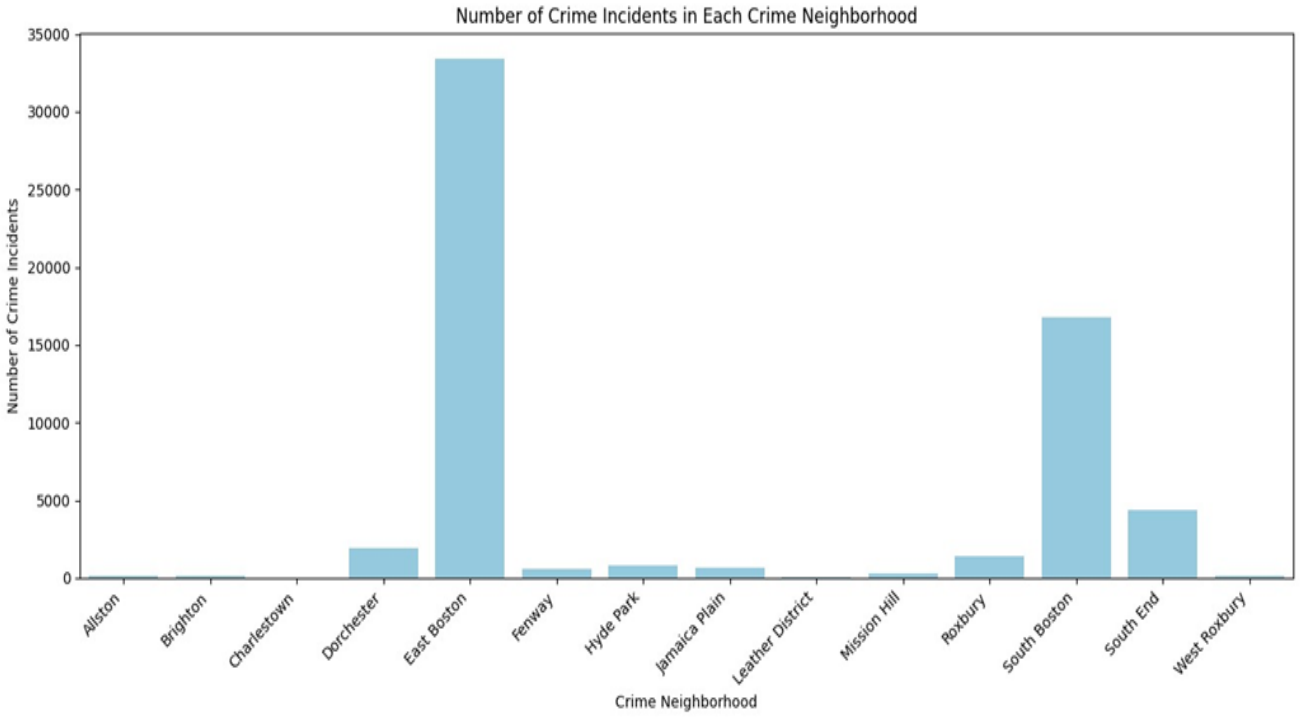
Overall, the regression analysis suggests that Land Value and Building Value are the primary factors influencing property prices in Boston, while Gross Tax may not have a significant effect. However, without neighborhood-specific variables, we cannot directly analyze neighborhood variations in property prices.

Descriptive Statistics or Exploratory Data Analysis (EDA) based on the Correlations and Regression Analysis:

1. **Impact of Crime Neighborhoods on Total Property value over the years:**

Firstly, we will calculate how many crime incidents occur in each crime neighborhood.

Result:

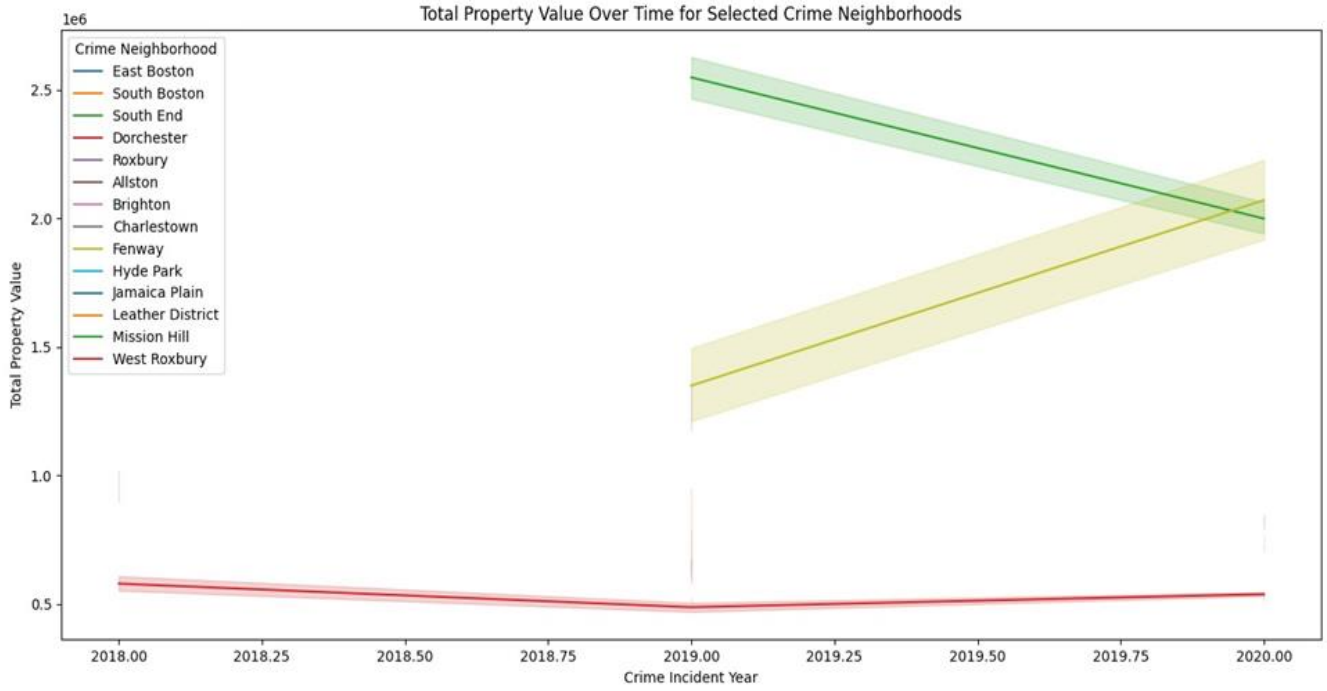


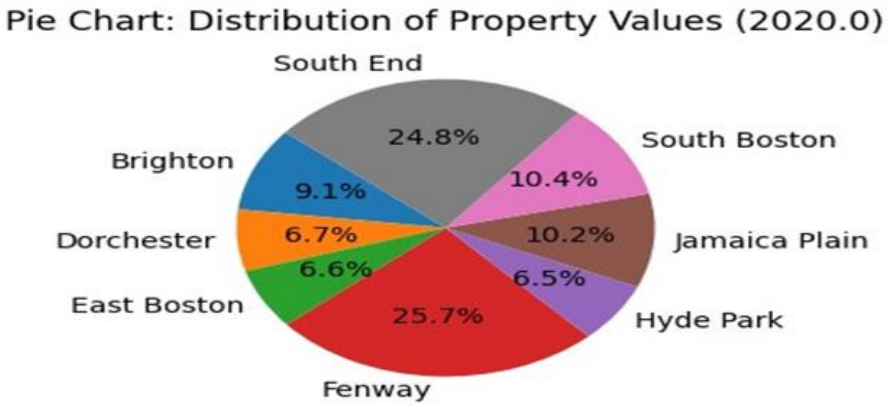
Analysis:

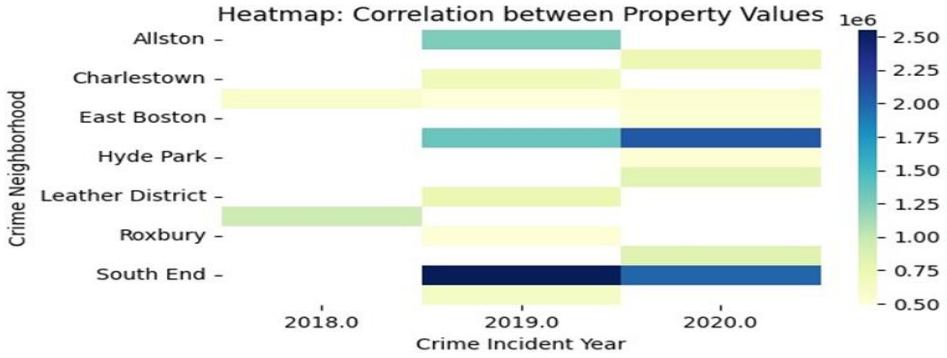
The above results conclude that the top 5 Crime Neighborhoods with the maximum number of crime incidents are East Boston, South Boston, South End, Dorchester, and Roxbury.

Now, we will check the trends of Total Property value in each neighborhood over the period of Crime Incident year.

Result:







Analysis:

The findings indicate that from 2018 to 2019, the Total Property Value in the Dorchester Crime Neighborhood experienced a decline, followed by an increase from 2019 to 2020. However, these fluctuations in property values are relatively minor and do not represent significant changes in monetary terms.

The findings indicate that from 2019 to 2020, the Total Property Value in the South End Crime Neighborhood experienced a considerable decline.

The analysis reveals a notable increase in the Total Property Value within the Fenway Crime Neighborhood from 2019 to 2020.

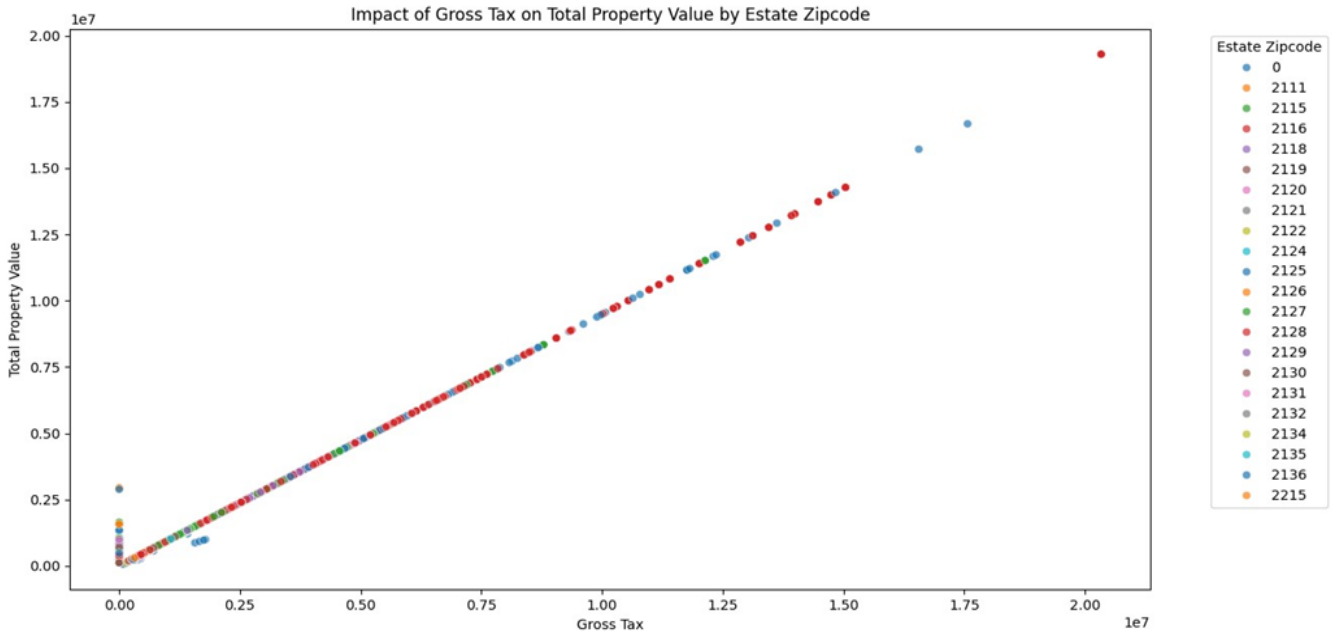
Conclusion: Since Dorchester and South End are in the top 5 Crime Neighborhoods with the maximum number of crime incidents, as per the analysis results, the property prices have declined in both neighborhoods. Also, with fewer crime incidents in Fenway, the total property value increases over the period.

This means that the No. of crime incidents in a neighborhood is a significant factor that impacts the total property value in that neighborhood.

With more Crime Incidents in a neighborhood, the Total Property value decreases. Fewer crime incidents in a neighborhood mean that the total property value increases.

This concludes that Crime Neighborhoods is a factor that influences the Total Property value.

2. Impact of Gross Tax on Total Property value:



Analysis:

The trend in the scatterplot above shows that for each Estate Zipcode, the Total Property value increases with Gross Tax.

Conclusion: In each Estate Zipcode,

If Gross Tax Increases, Total Property value increases.

If Gross Tax Decreases, Total Property value decreases.

This concludes that Gross Tax is a factor that influences the Total Property value of each Estate.

3. Impact of Building Value on Total Property Value:

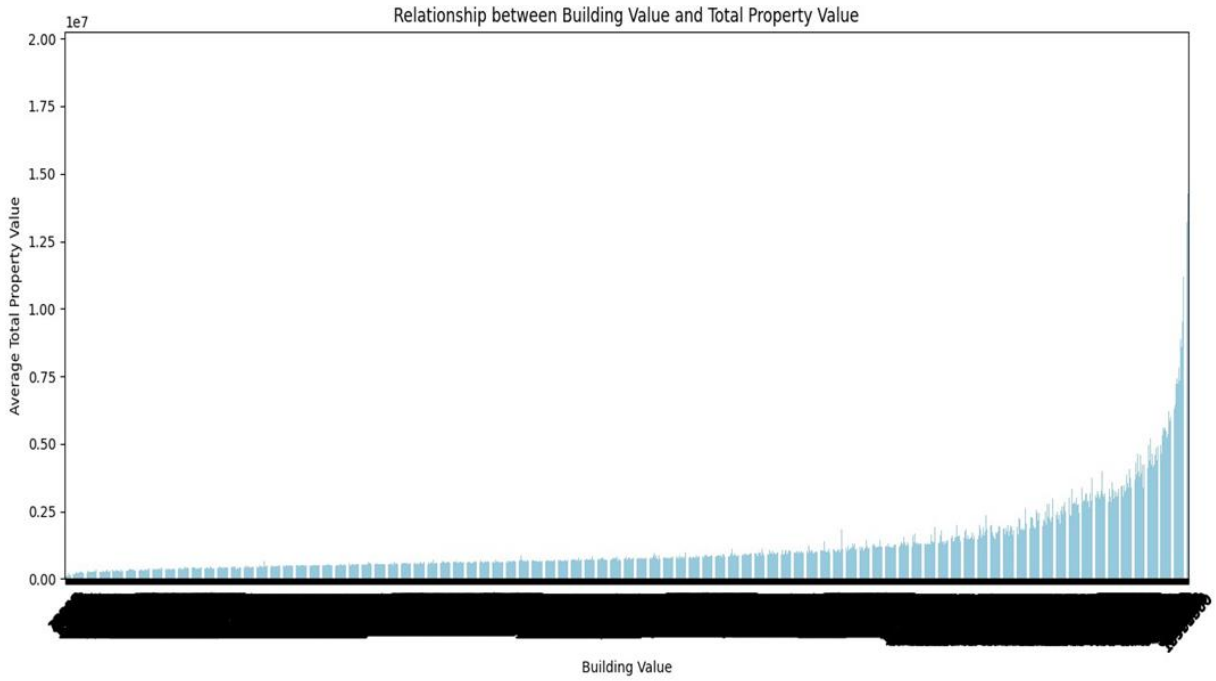
To Analyze this, we sorted the 'Building Value' variable in descending order and considered the top 100 data entries to check the trends. After that, for those top 100 highest building values, we checked the 'Total Property Values' for those Estate Addresses.

Result:

Analysis:

In this scatterplot, we can see for the Estate Addresses with the Maximum Building Values, the Total property values also increase with the Building value. Also, if the building value is less, the total property value is also less. The top 5 Estate Addresses with maximum Building Value and Total Property values are: 15CommonwealthAve, 200Clarendonst, 10marlboroughst, 10tjamesav11thfl, 66canalst&. To verify our results in more depth, we compared the relationship between Building values and Total property values for the entire dataset.

Result:



Analysis:

In this barplot for the complete dataset, we can see that the total property values also increase as the building values increase.

Conclusion: In each Estate Address with the Maximum amount of Building values,

If Building value Increases, Total Property value increases.

If Building value Decreases, Total Property value decreases.

This concludes that Building value is a factor that influences the Total Property value of each Estate Address.

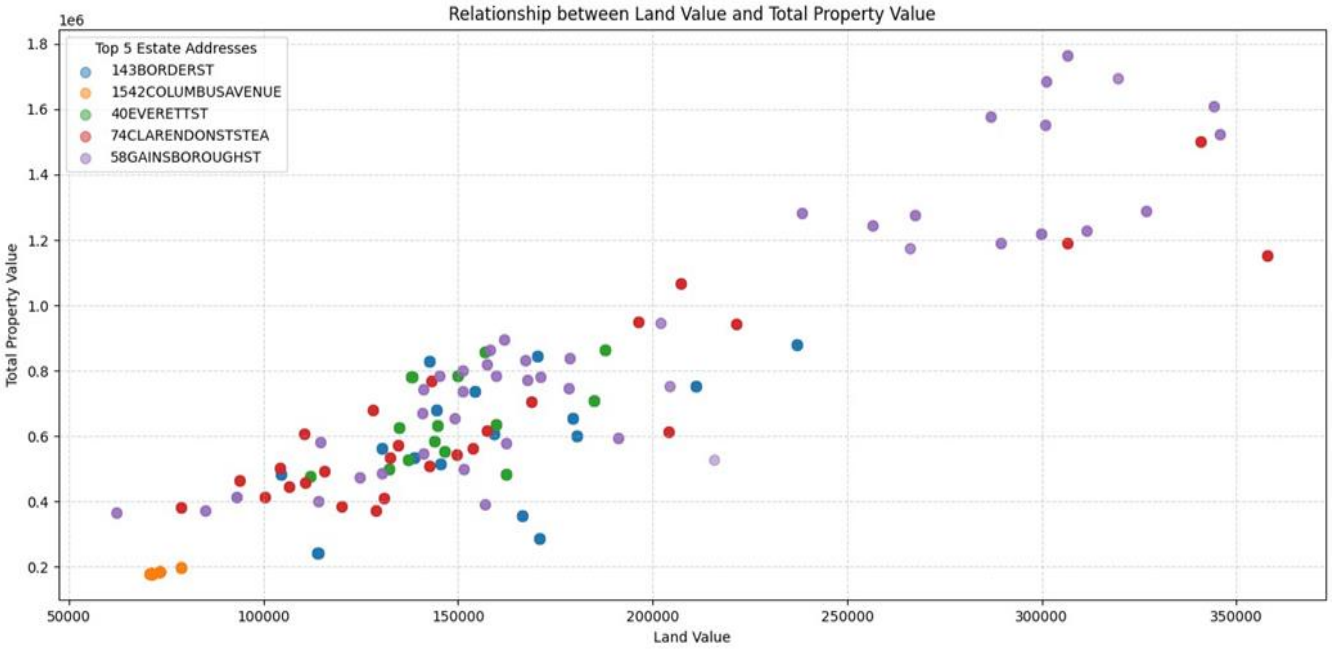
4. **Impact of Land Value on Total Property Value:**

To find the impact of Land value on total property value, we use a bubble plot to see how the land values influence the total property values. Also, we will mention the estate addresses for the top 5 highest values in the results.

Code:

```
import pandas as pd
import matplotlib.pyplot as plt

# Read data from Excel
data = pd.read_excel('Final Blend.xlsx')

# Filter out rows with missing values in 'Land Value' or 'Total Property Value'
filtered_data = data.dropna(subset=['Land Value', 'Total Property Value'])

# Get the top 5 unique estate addresses based on frequency
top_estate_addresses = filtered_data['Estate
Address'].value_counts().nlargest(6).index.tolist()
top_estate_addresses = [address for address in top_estate_addresses if address]

# Plotting the relationship between 'Land Value' and 'Total Property Value' using a
bubble plot
plt.figure(figsize=(10, 6))
for address in top_estate_addresses:
    estate_data = filtered_data[filtered_data['Estate Address'] == address]
    plt.scatter(estate_data['Land Value'], estate_data['Total Property Value'], s=50,
alpha=0.5, label=address)

# Setting labels and title
plt.xlabel('Land Value')
plt.ylabel('Total Property Value')
plt.title('Relationship between Land Value and Total Property Value')
plt.legend(title='Top 5 Estate Addresses', loc='upper left')
plt.grid(True, linestyle='--', alpha=0.5)  # Add grid lines for better readability
plt.tight_layout()

# Show plot
plt.show()
```

Result:



Analysis:

In this bubble plot, we can see that the total property values also increase as the land values increase. The top 5 highest land values and total property values estate addresses are 143borderst, 1542columbusavenue, 40everettst, 74clarendonststea, 58gainsboroughst.

Conclusion: In each Estate Address with the Maximum amount of Land values,

If Land value Increases, Total Property value increases.

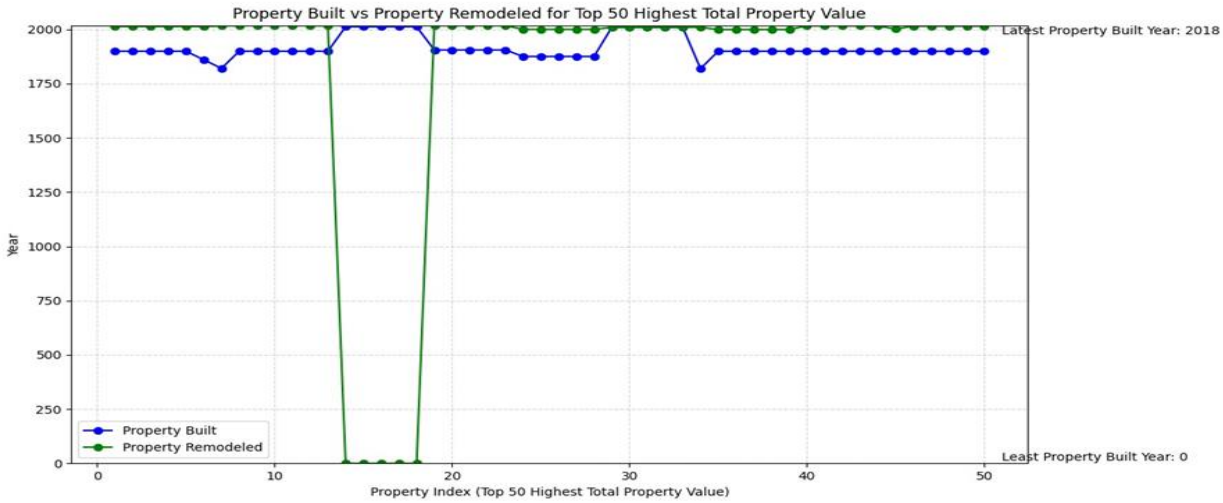
If Land value Decreases, Total Property value decreases.

This concludes that Land value is a factor that influences the Total Property value of each Estate Address.

5. **Impact of Estate Street No. on 'Total Property Value'**

To check this parameter, we will first choose top 50 data rows with the highest 'Total Property value' and then plot a graph to show their Built Year and Remodeled Year.

Result:



Analysis:

· For the top 50 properties with the highest prices, the years of property construction and remodeling predominantly fall within or after the year 2000. This suggests that either the properties were constructed before and remodeled during or after the year 2000, or they were built entirely during or after that time. Consequently, most of these properties were constructed several years ago, contributing to their elevated total property values.

Now, find the Estate Street No for these top 50 highest-priced properties:

Now Similar to the above analysis, we will now choose the least 50 data rows with the lowest ‘Total Property value’ and then plot a graph to show their Built Year and Remodeled Year.

Analysis:

For the least 50 properties with the lowest prices, the years of property construction and remodeling predominantly falls below the year 2000. This suggests that the properties were constructed before the year 2000 falling upto the years in late 1700s. Consequently, most of these properties were constructed several decades ago and are old, contributing to their lower total property values. Now, find the Estate Street No for these least 50 lowest-priced properties:

Code:

```
# Sort data by 'Total Property Value' in ascending order and select the least 50 rows
least_50_properties = data.nsmallest(50, 'Total Property Value')

# Get distinct estate street numbers for the least 50 properties
distinct_estate_street_numbers = least_50_properties['Estate Street Number'].unique()

# Print the distinct estate street numbers
print("Distinct Estate Street Numbers for the Least 50 Lowest-Priced Properties:")
for street_number in distinct_estate_street_numbers:
    print(street_number)
```

Result:

Estate Street Numbers for the Least 50 Lowest-Priced Properties:

221, 1901, 90-92, 19, 18, 56-58, 153155A, 72, 42, 4244, 57-59, 1899, 2

Conclusion: Now looking at both the above results:

Estate Street Numbers for the Top 50 Properties:

15, 315, 85, 211, 19, 161, 10, 13, 7, 56, 58, 128, 39, 130

Estate Street Numbers for the Least 50 Lowest-Priced Properties:

221, 1901, 90-92, 19, 18, 56-58, 153155A, 72, 42, 4244, 57-59, 1899.

Several Estate Streets in the dataset feature both the highest and lowest priced properties. Notable examples include Estate Street No: 19, 56, 57, and 58. As a result, we can infer that the neighborhoods along these Estate Streets do not significantly influence property prices.

Recommendations:

Discover Secure Neighborhood Havens: Scout neighborhoods boasting low crime incidents and reduced crime rates, unveiling havens for secure property investments. These areas harbor the potential for heightened Total Property values, aligning with your quest for stability and growth.

Navigate Risk with Neighborhood Insight over time: Navigate towards neighborhoods with heightened crime incidents and elevated crime rates, acknowledging the associated risks. While these areas may present challenges, strategic investments could still yield rewards, albeit with a prudent approach to risk management.

Location Wisdom: Harness the power of neighborhood dynamics, understanding the intricate relationship between gross tax and Total Property value within each locale. Seek out neighborhoods where property values are modest, alongside potentially lower gross tax burdens, maximizing your investment potential.

Estate Exploration: Delve into the building and land values associated with a specific estate address when contemplating investment ventures. Trends indicate that higher building and land values in a neighborhood in a particular estate correlate with appreciating property prices over time, offering promising investment prospects.

Street-Smart Investment Tactics: Exercise caution regarding investment decisions solely based on the presence of multiple high-priced properties along a specific neighborhood Estate Street. Not all properties within the same estate street guarantee favorable investment returns, emphasizing the need for thorough evaluation beyond superficial indicators.

Research Question 5: Predicting Property Prices Based on Various Factors

Can property prices be predicted based on property size, location, property features, age, and crime rate in Boston? What are the most significant predictors of property prices?

Moving on to the 5th research question, in here we are finding out what features are significant and influence the Dependent variable **"Total Property Value".** For this we executed some methods based on these sections:

Descriptive Statistics: Providing a statistical summary of the dataset to understand the central tendency, dispersion, and shape of the dataset's distribution. Here's a Statistical Summary of the Dataset:

Numerical Variables:

Total Property Value ranges from $65,033 to $19,295,000 with a mean of approximately $787,667.

Land Value and Building Value also show a wide range, indicating variability in property characteristics.

Gross Tax has a similarly broad range, reflecting differences in property assessments.

Variables like Land size, Living Area, and Num Floors provide insights into the physical attributes of properties.

Categorical Variables:

- Property Type: Predominantly "Triple Family" properties.
- Building Style: "Decker" is the most common.
- Roof Type: "Flat" roofs are the most prevalent.

- Exterior Type: "Vinyl" is the most common material used.
- Heating Type: "Hot Water" heating systems dominate.
- AC Type: The majority of properties do not have air conditioning ("None").

Data Preprocessing for Research question 5:

With respect to raw data, we had preprocess it into clean data and then work on it, by excluding the null values and special characters. Also checked for Null values and replacing it with logical data, for example column “AC Type” had more than 50,000 rows which had None that means No AC, but we had to replace that with “No AC” as it was taking it as Null values.

```
data = pd.read_csv('577_PROJECT.csv')
data.isnull().sum()
```

```
School ID                1040
School Type              1040
Estate Zipcode              0
Land size                   0
Property Built              0
Property Area               0
Living Area                 0
Num Floors                  0
Building Style              0
Roof Type                   0
Exterior Type               0
Total Rooms                 0
Total Bedrooms              0
Total FullBaths             0
Total Halfbaths             0
Total Kitchens              0
AC Type                 50902
Total Property Value        0
dtype: int64
```

Which after processing comes with respect to zero and saving the big chunk of the data.

For further analysis, we can use Label encoder or One Hot encoding for the categorical values. In here we will use One Hot Encoding to convert the categorical values.

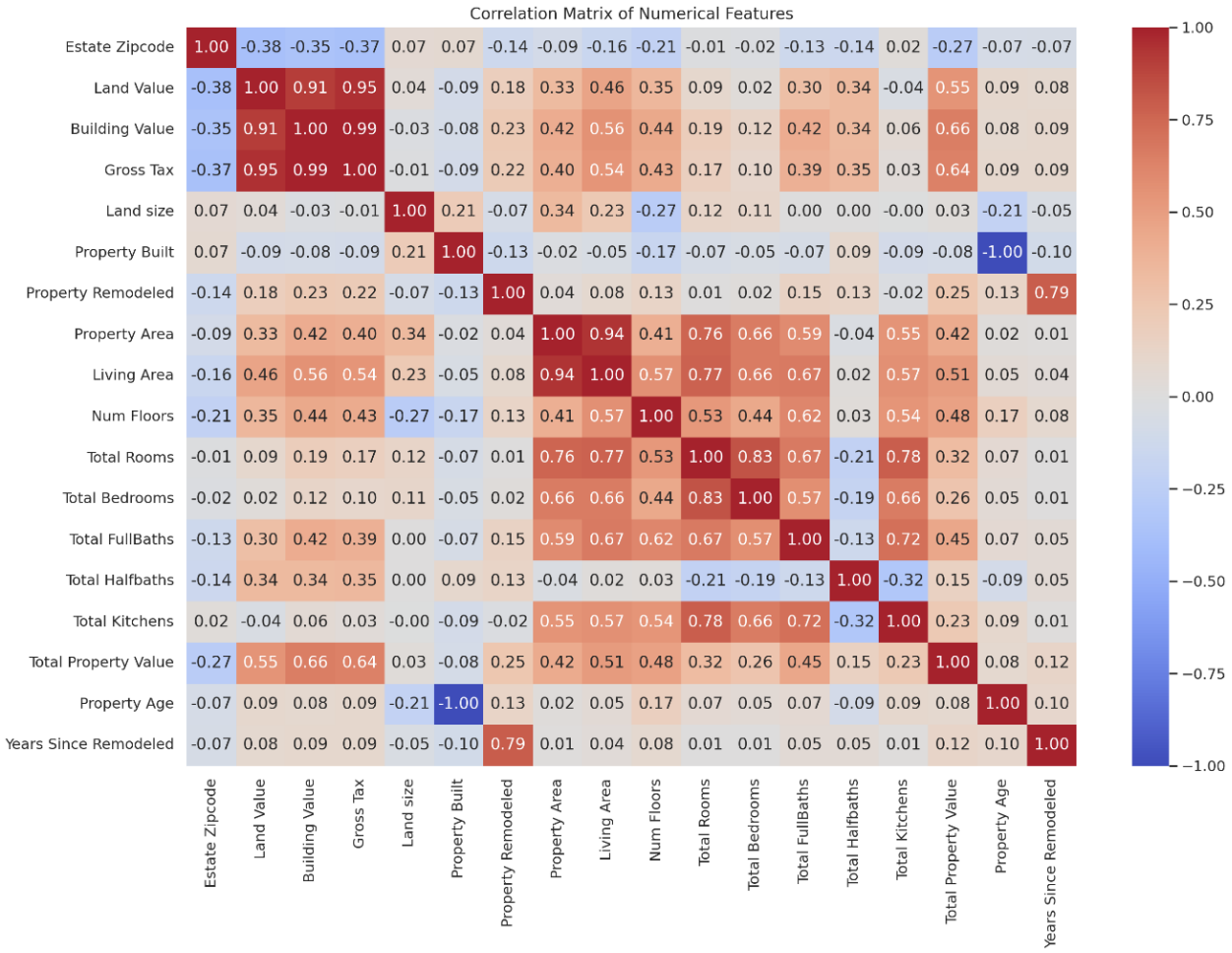
We also dealt with outliers, inconsistencies for the dataset. This was our first step to address this question. There are 5,850 outliers in the 'Total Property Value' feature based on the Interquartile Range (IQR) method. Handling these outliers is crucial as they can significantly impact the performance of our predictive models.

```
# Capping outliers in 'Total Property Value'
Q1 = data['Total Property Value'].quantile(0.25)
Q3 = data['Total Property Value'].quantile(0.75)
IQR = Q3 - Q1
upper_bound = Q3 + 1.5 * IQR
data['Total Property Value'] = np.where(data['Total Property Value'] > upper_bound, upper_bound, data['Total Property Value'])
```

Here's the above snippet we performed to get rid of the outliers in TPV.

- **Cap the Outliers**: Limiting the property values to the upper bound for the higher outliers. This approach maintains all data points but reduces the effect of extreme values.
- **Remove the Outliers:** Excluding these data points from the analysis to focus on a more uniform dataset.

Correlation Analysis: Examine the correlation between "Total Property Value" and other numerical features.



The correlation analysis provides insights into how closely related different numerical features are to the **Total Property Value**:

Building Value and Gross Tax have strong positive correlations with property value, which indicates that as these values increase, so typically does the property's total value.

Living Area and Num Floors also show significant positive correlations, suggesting larger and multi-story properties tend to have higher values.

Years Since Remodeled shows a moderate positive correlation, implying that more recently remodeled properties might fetch higher values.

Interestingly, Property Age has a very low positive correlation, and Property Built shows a negative correlation when considered directly, which shows a lesser impact of sheer age on property value compared to other factors.

Analysis & Visualizations:

Encoding Categorical Variables: Transforming categorical variables into numerical format using encoding techniques so that they can be included in the machine learning models.

Here's a following snippet that shows the logic on coding categorical variables we encoded:

```
# Encoding categorical variables
categorical_cols = data.select_dtypes(include=['object']).columns
column_transformer = ColumnTransformer(transformers=[('cat', OneHotEncoder(handle_unknown='ignore'), categorical_cols)], remainder='passthrough')
X = data.drop('Total Property Value', axis=1)
y = data['Total Property Value']
X_transformed = column_transformer.fit_transform(X)
```

We encoded the columns which had object in it. Here, a **ColumnTransformer** is initialized. It applies transformations to specific columns while leaving other columns unchanged.

The categorical variables encoded, and our dataset is now split into training and test sets. We have 48,560 entries in the training set and 12,141 entries in the test set, with 85 features after one-hot encoding.

Model Building: Developing Linear regression, Random Forest Regression models and CART models using the processed data.

Model Settings: The categorical variables have been successfully encoded, and now we split the dataset into training and test sets. We have 48,560 entries in the training set and 12,141 entries in the test set, with 85 features after one-hot encoding.

```
# Encoding categorical variables
categorical_cols = data.select_dtypes(include=['object']).columns
column_transformer = ColumnTransformer(transformers=[('cat', OneHotEncoder(handle_unknown='ignore'), categorical
X = data.drop('Total Property Value', axis=1)
y = data['Total Property Value']
X_transformed = column_transformer.fit_transform(X)

# Splitting the dataset
X_train, X_test, y_train, y_test = train_test_split(X_transformed, y, test_size=0.2, random_state=42)
```

Here in this above snippet, you can see we split the data into 80% Train and 20% Test Size.

Moving forward to the first Regression model

Linear Regression:

We will run this model to learn more about the linear relationships that influence the variables to each other in this dataset. Statistically calculate the R^2 and RMSE value to assess how well they predict Property Values for Linear Regression.

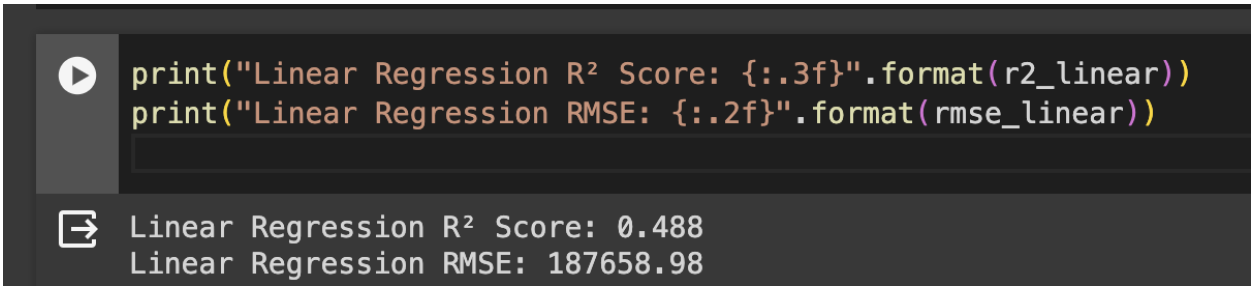
- R^2 Score for Linear Regression: This indicates how well the model explains the variance in the property values.
- Root Mean Squared Error (RMSE) for Linear Regression: This measures the average error in the predictions.

Here below is the following snippet for the Linear Regression used:

```
# Training a Linear Regression model
linear_model = LinearRegression()
linear_model.fit(X_train, y_train)
y_pred_linear = linear_model.predict(X_test)
r2_linear = r2_score(y_test, y_pred_linear)
rmse_linear = mean_squared_error(y_test, y_pred_linear, squared=False)
```

Here we used the train and test split that we specified in the model settings.

Below are the results after printing the Linear Regression for the model.



Conclusion of Linear Regression Model:

Output - R^2 Score: 0.488, RMSE (Root Mean Squared Error): $187,658.98.

R^2 value is less than 0.7 that means the model is not accurate enough to predict the Property Value, whereas the RMSE value is also high, that means the error ranges at around $187,658.98 which is too high. We cannot rely on this model because of its less accuracy. Next step is moving towards Random Forest Model.

Random Forest Model:

Why Random Forest Model?

Random Forest models are commonly used for classification tasks because they offer several advantages. It's important to note that the performance of Random Forests, like any other machine learning algorithm, depends on the characteristics of the specific dataset and problem at hand.

After looking at the poor results because of Linear Regression, here is how we codd the Random Forest Model:

```
[7]  # Training a Random Forest model
     random_forest_model = RandomForestRegressor(n_estimators=100, random_state=42)
     random_forest_model.fit(X_train, y_train)
     y_pred_rf = random_forest_model.predict(X_test)
     r2_rf = r2_score(y_test, y_pred_rf)
     rmse_rf = mean_squared_error(y_test, y_pred_rf, squared=False)
```

Here n_estimators mean it's used as a hyperparameter that determines number of decision trees to be used in the model. Each decision tree is trained independently. You can use n_estimators based on your convenience and can cross validate.

The results are as follows:

```
Linear Regression R² Score: 0.488
Linear Regression RMSE: 187658.98
Random Forest R² Score: 0.9223
Random Forest RMSE: 1145.73
```

Conclusion of Random Forest Model:

Output - R^2 Score: 0.9223, RMSE (Root Mean Squared Error): $1,145.73.

The Random Forest model shows a much higher R^2 score and a significantly lower RMSE compared to the Linear Regression model, indicating it performs exceptionally well on the given data. This suggests that the model can effectively capture the complex relationships and variability in the data that the linear model might be missing.

CART (Classification and Regression Trees):

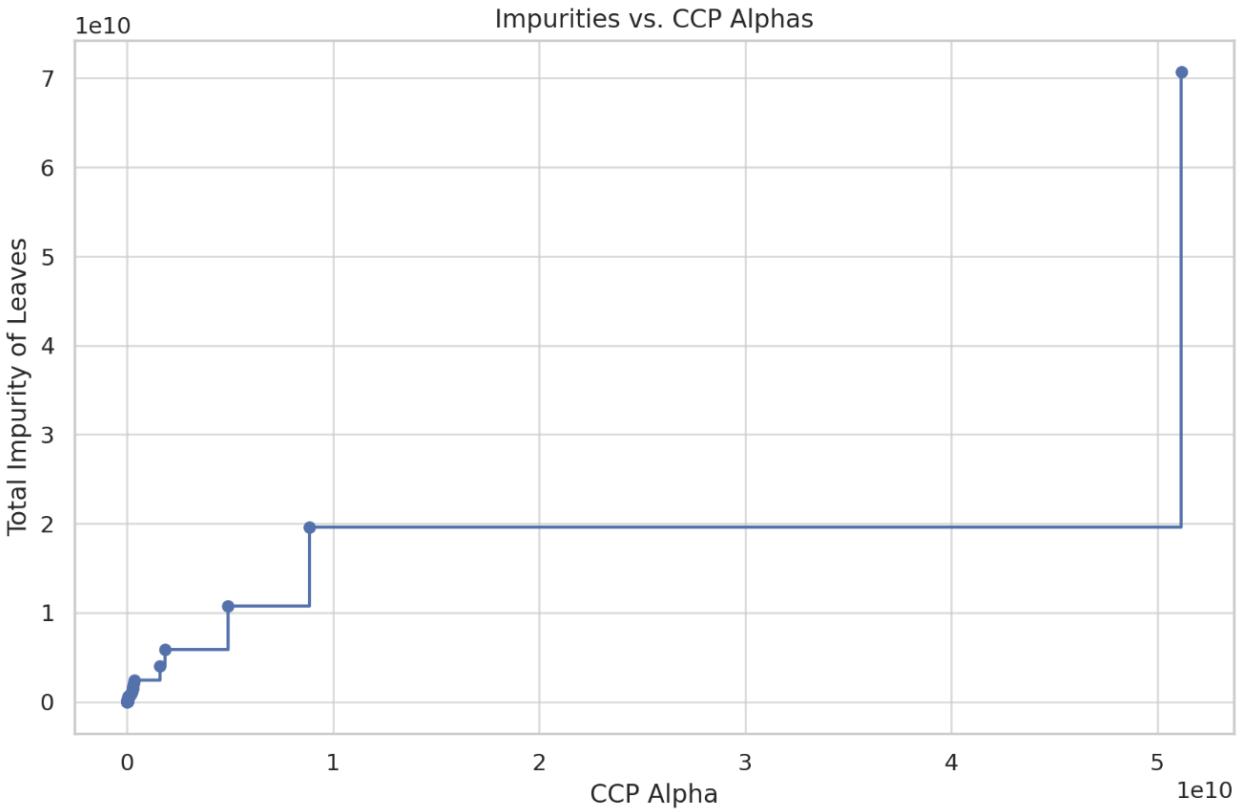
CART (Classification and Regression Trees) is a type of decision tree algorithm used for both classification and regression tasks. Here we used CART to visualize Minimum Error Tree and Best Pruned Tree.

1. **Minimum Error Pruning**: This approach involves growing the full decision tree and then pruning it back by considering each subtree. Subtrees are replaced by a single node representing the most common class or the mean value in the case of regression. If the accuracy of the tree improves or remains the same after pruning, the pruning operation is accepted. Otherwise, the tree remains unchanged. This process continues until further pruning does not improve performance.
2. **Best-First Pruning**: Best-First Pruning is a more sophisticated approach that explores different subtrees and evaluates their performance based on a cost function such as accuracy, error rate, or cross-validation score. It starts with the full tree and iteratively removes branches to optimize the chosen cost function. Unlike Minimum Error Pruning, Best-First Pruning does not require a separate validation set and can be more computationally efficient.

Both pruning techniques aim to simplify the decision tree while preserving or even improving its predictive accuracy on unseen data.

We also calculate the CCP_Alpha values for this data set to prune the trees.

1. 1.72748782e+08
2. 1.96422958e+08
3. 2.24923618e+08
4. 2.63979898e+08
5. 2.70082562e+08

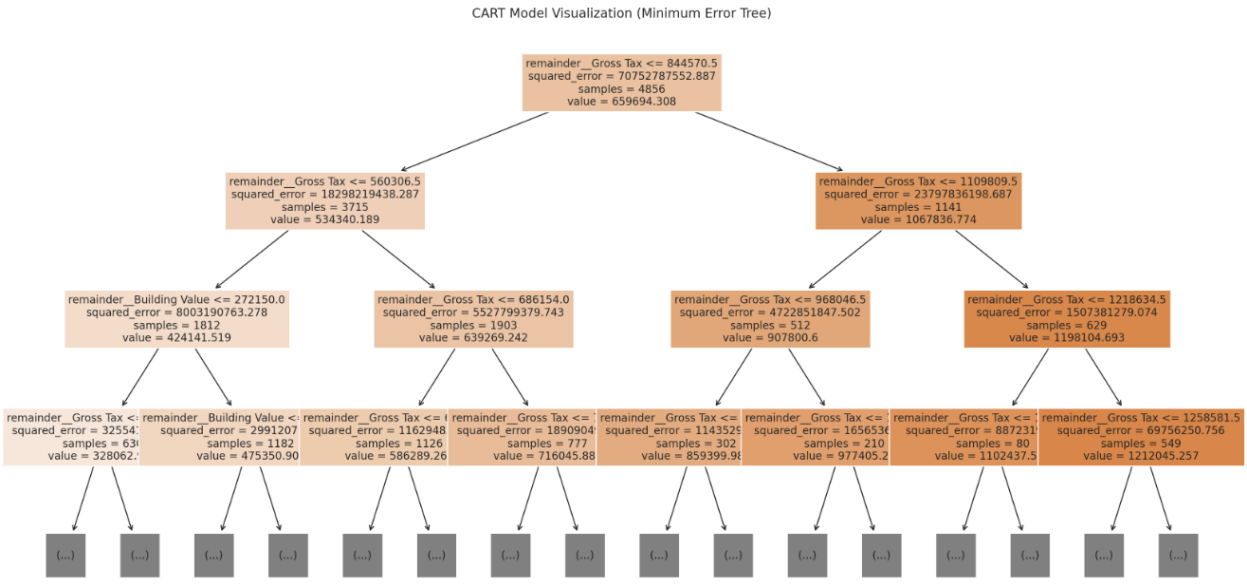


These values indicate points where pruning the tree leads to significant reductions in complexity without much loss in data purity. We have trained and visualized decision trees using these top 5 ccp_alpha values to determine the best-pruned tree based on their performance on the test data.

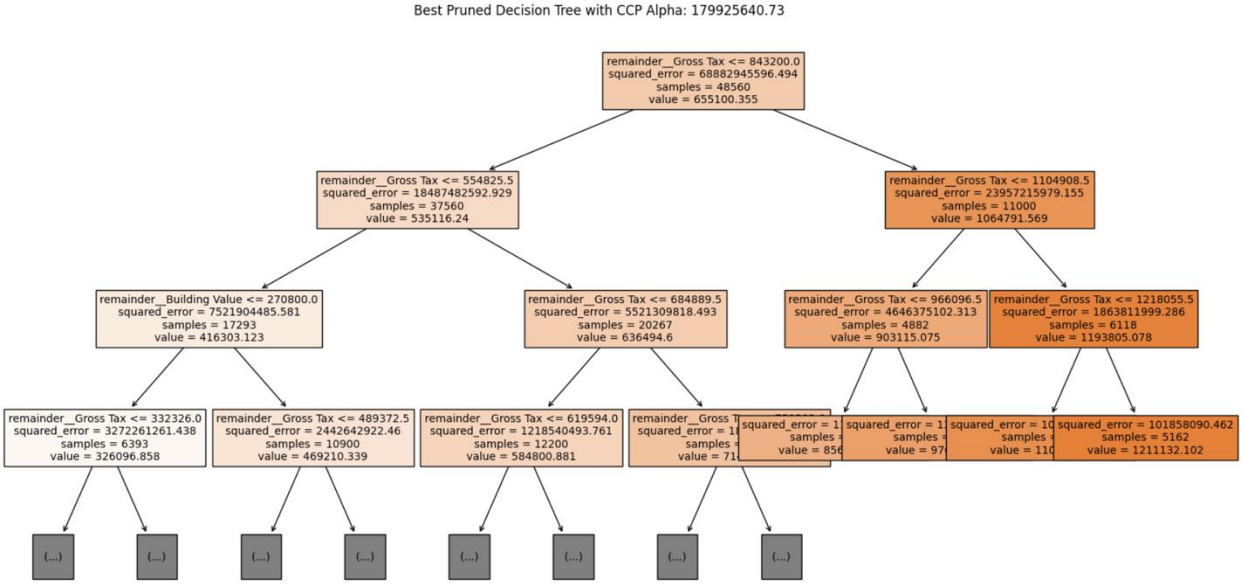
Below is the code for ME Tree and Best Pruned Tree

```
[11] #Training a Decision Tree for pruning analysis
     cart_model = DecisionTreeRegressor(random_state=42)
     cart_model.fit(X_train, y_train)
     path = cart_model.cost_complexity_pruning_path(X_train, y_train)
     ccp_alphas, impurities = path.ccp_alphas, path.impurities

     # Selecting top 5 significant ccp_alpha values for detailed analysis
     significant_alphas = ccp_alphas[np.where(np.diff(impurities) > np.mean(np.diff(impurities)) * 5)[0] + 1][:5]

     # Training decision trees for each of the significant ccp_alpha values
     top_dt_regressors = [DecisionTreeRegressor(random_state=42, ccp_alpha=alpha).fit(X_train, y_train) for alpha in
     top_r2_scores = [r2_score(y_test, dt.predict(X_test)) for dt in top_dt_regressors]
     best_alpha = significant_alphas[np.argmax(top_r2_scores)]
     best_pruned_tree = top_dt_regressors[np.argmax(top_r2_scores)]

     # Visualizing the best pruned tree
     plt.figure(figsize=(20, 10))
     plot_tree(best_pruned_tree, filled=True, feature_names=feature_names, max_depth=3, fontsize=10)
     plt.title('Best Pruned Decision Tree with CCP Alpha: {:.2f}'.format(best_alpha))
     plt.show()
```

Here's the visualization of Minimum Error Tree:



Here's the visualization of Best Pruned Tree:



We successfully trained and visualized the best pruned trees using the top 5 ccp_alpha values. The best-performing pruned tree, based on the highest R^2 score on the test set, used a ccp_alpha value of approximately 172.75 million. This model achieved an R^2 score of 0.987, indicating a very high level of accuracy in predicting property values.

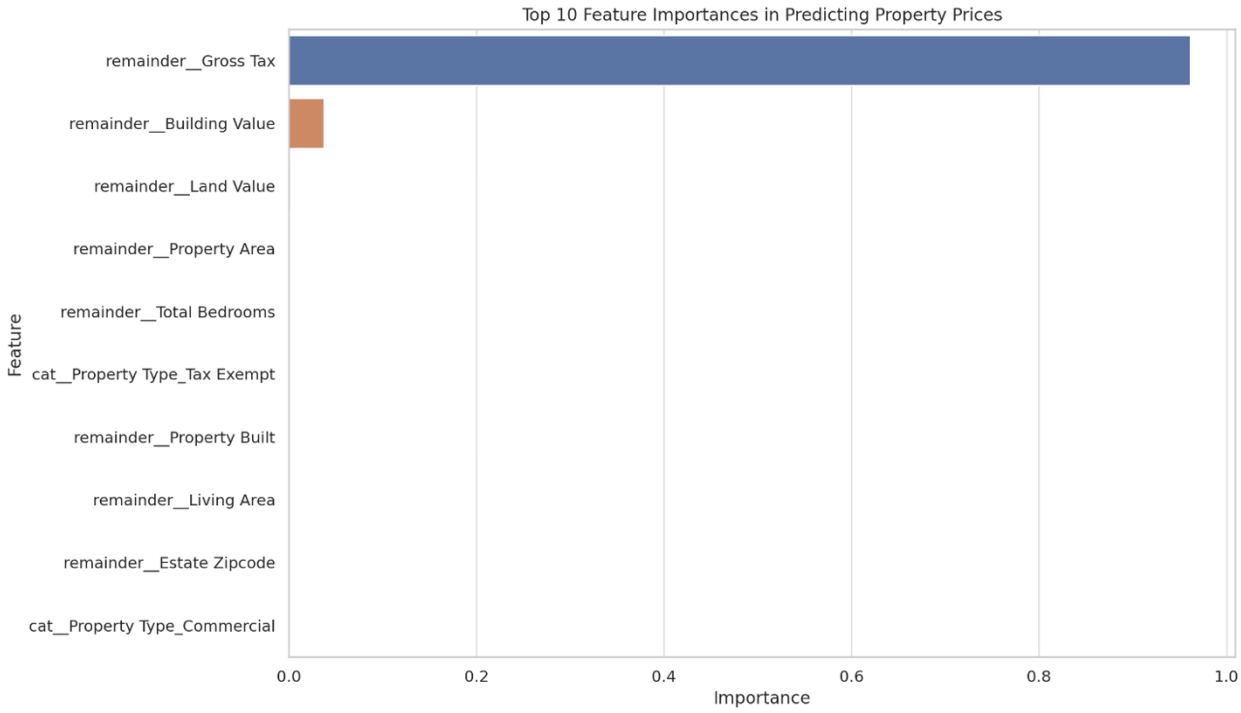
Summary of CART:

- CCP Alpha: 172.75 million
- R^2 Score: 0.987

This pruned tree effectively balances model complexity and predictive accuracy.

Feature Importance Analysis:

After training the models, analyzing which features are most influential in predicting property prices.



The feature importance analysis for the Random Forest model highlights the top predictors of property prices in Boston.

1. **Gross Tax**: This feature is by far the most significant predictor, with the highest importance score, suggesting that the taxes associated with a property strongly correlate with its value.

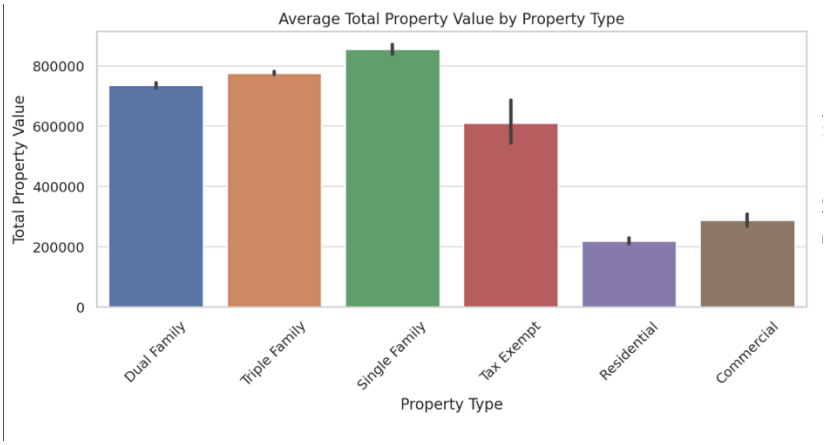
2. **Building Value**: The next most important feature, indicating that the market value assigned to the building structure itself is a crucial determinant of the overall property value.
3. **Land Value**: Although much less significant compared to the first two, the value of the land on which the property sits still plays a role in determining its price.

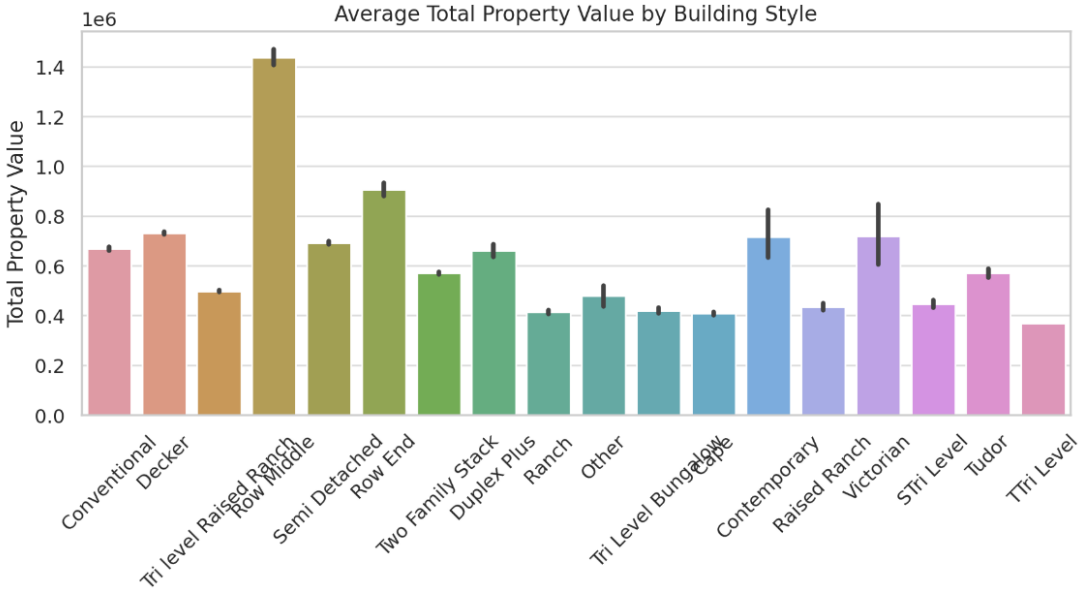
Other notable features include:

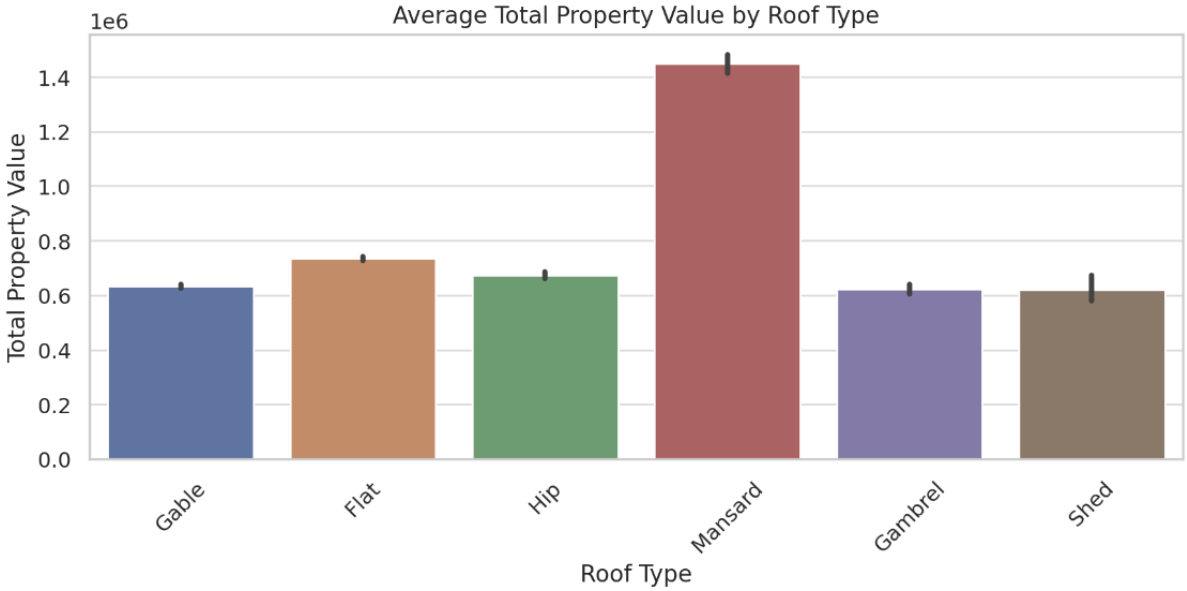
- **Property Area**
- **Total Bedrooms**
- Specific property types like **Tax Exempt** and **Commercial**

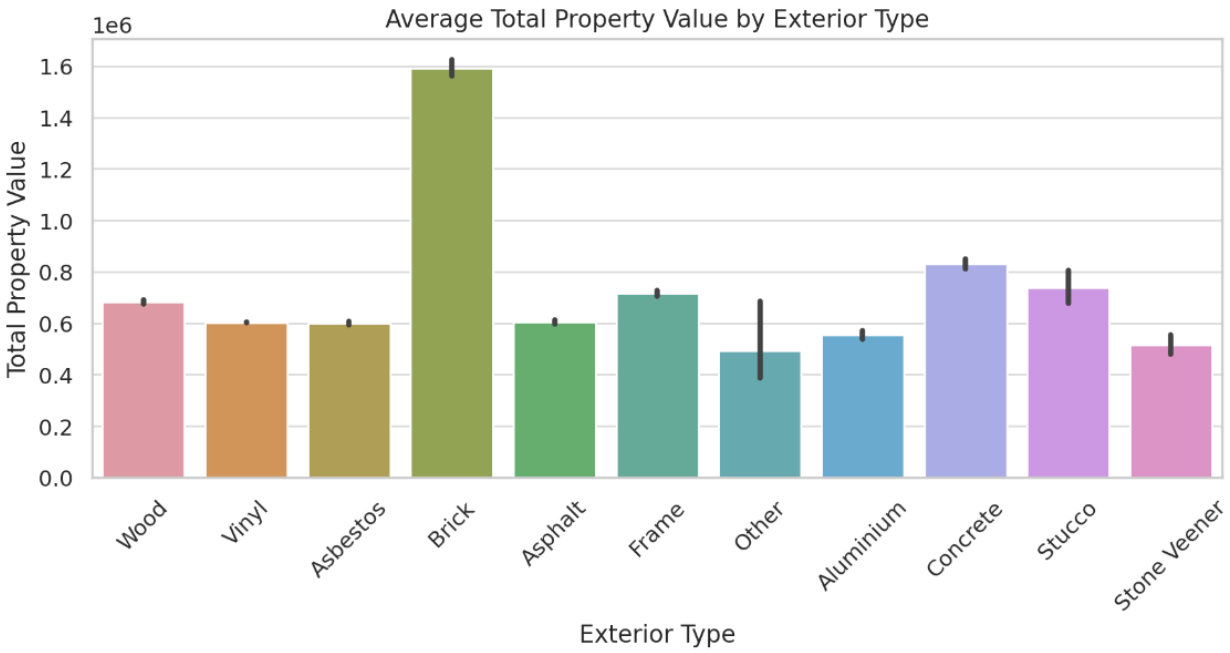
These insights can greatly aid managerial decision-making in real estate, allowing for more focused investment strategies and better understanding of the factors that drive property values in Boston.

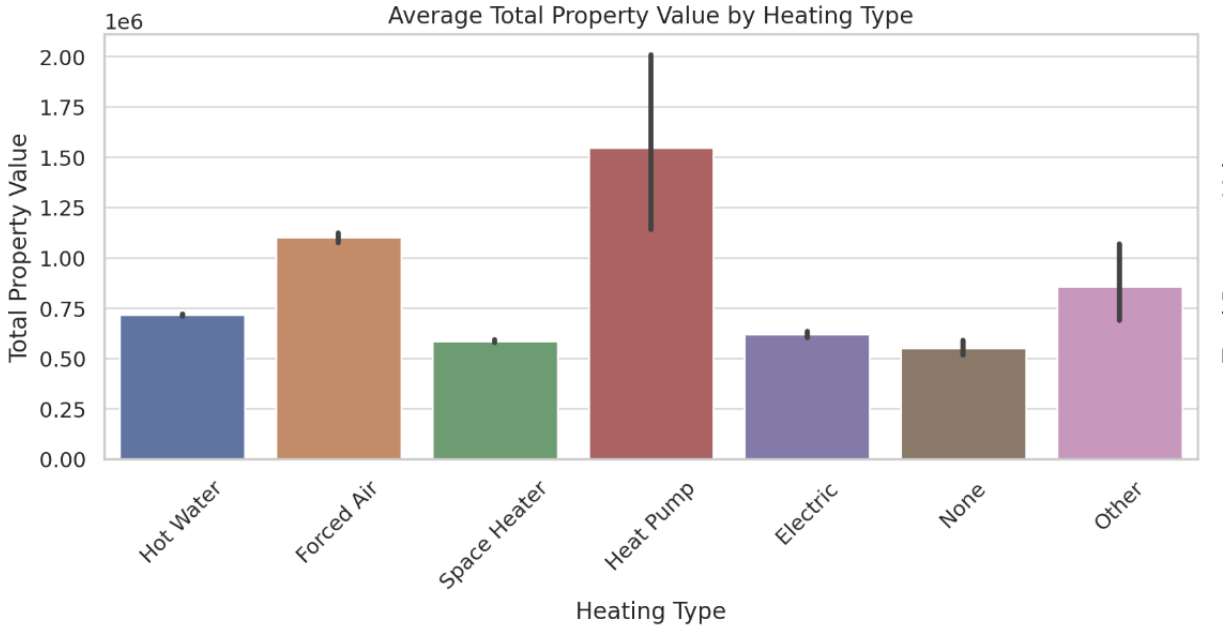
Here are the graphs showing high contribution of each Categorical Variables toward the dependent variable Total Property Value.

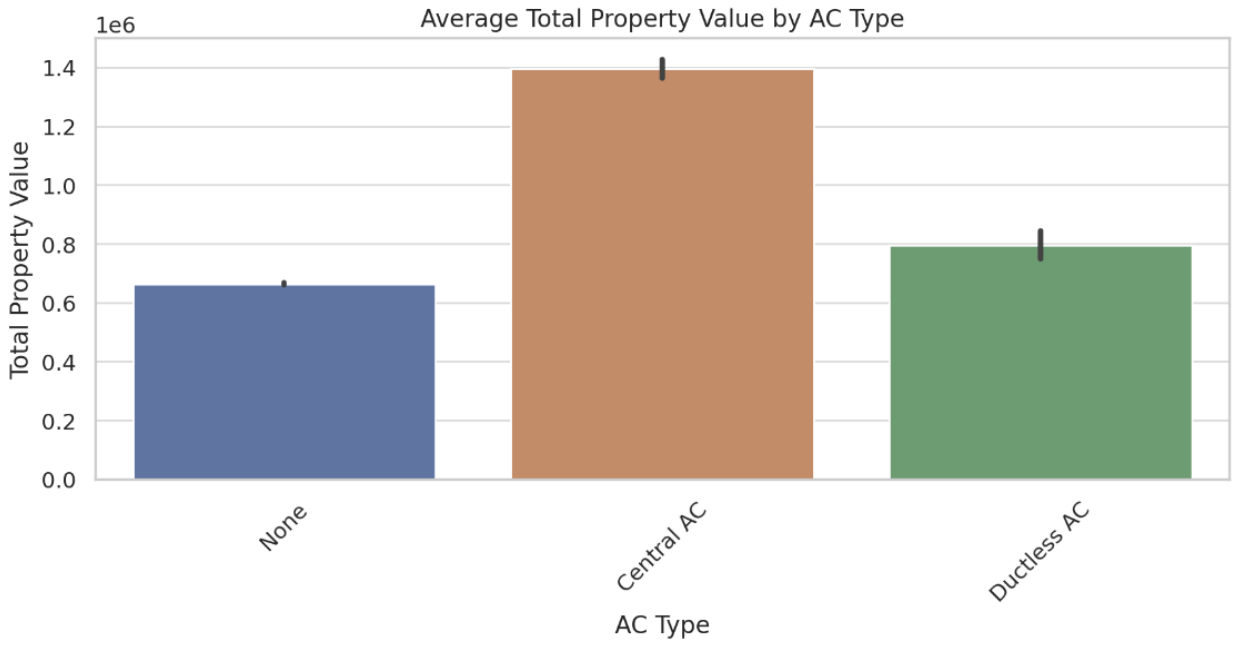



Average Total Property Value by Building Style
1e6
Total Property Value
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Conventional
Decker
Tri level Raised Ranch
Row Middle
Semi Detached
Row End
Two Family Stack
Duplex Plus
Ranch
Other
Tri Level Bungalow
Cape
Contemporary
Raised Ranch
Victorian
STri Level
Tudor
TTri Level


Average Total Property Value by Roof Type
1e6
Total Property Value
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Gable
Flat
Hip
Mansard
Gambrel
Shed
Roof Type


Average Total Property Value by Exterior Type
1e6
Total Property Value
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Wood
Vinyl
Asbestos
Brick
Asphalt
Frame
Other
Aluminium
Concrete
Stucco
Stone Veener
Exterior Type


Average Total Property Value by Heating Type
1e6
Total Property Value
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Hot Water
Forced Air
Space Heater
Heat Pump
Electric
None
Other
Heating Type



Analysis of Categorical Variables and Their Impact on Total Property Value:

1. Property Type: The average total property value varies significantly across different property types. For instance, certain types such as "Single Family" and "Triple Family" may show distinct value characteristics due to their differing market demands and property features.
2. Building Style: Different styles like "Decker" or "Colonial" have varied average property values, reflecting the architectural desirability and historical value that may influence pricing.
3. Roof Type: Roof types such as "Flat" and "Hip" show differences in property values, possibly due to costs associated with construction and maintenance, or preferences in certain climates or regions.

4. Exterior Type: Materials like "Vinyl" and "Wood" show different average values, which could be influenced by durability, aesthetics, or regional material availability.
5. Heating Type: Variations in heating systems such as "Forced Air" versus "Hot Water" can reflect on property values, indicating the efficiency or modernity of the heating system.
6. AC Type: The presence or absence of air conditioning ("None" vs. "Central") significantly impacts the value, especially in regions where AC is a crucial amenity.

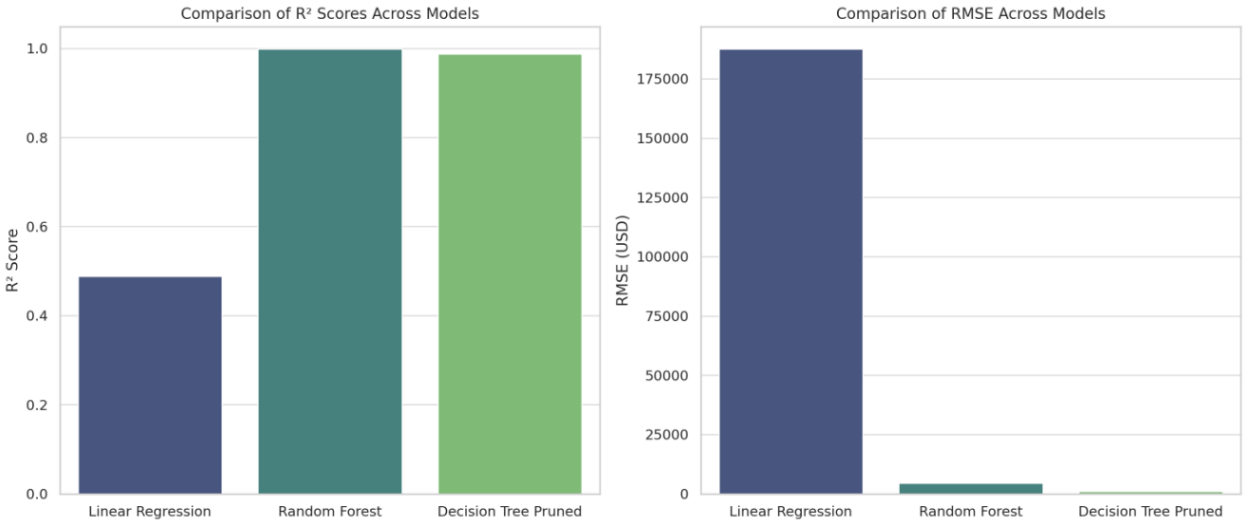
Conclusion

Through our comprehensive analysis using various machine learning models on the Boston property dataset, we gained several insights into the factors influencing property prices. Here's a summary of our key findings and model performances:

1. **Feature Importance and Predictors**:
    - **Gross Tax** and **Building Value** emerged as the most significant predictors of property prices, as identified by the Random Forest model. This indicates that fiscal aspects, like taxes and the assessed value of buildings, are major drivers of property valuations.
    - **Land Value** and **Living Area** also play significant roles but to a lesser extent.
2. **Model Performances**:
    - The **Linear Regression** model provided a baseline with an R^2 score of 0.488 and an RMSE of $187,658.98, suggesting moderate predictive power.

- The **Random Forest** model, even when trained on a smaller subset, significantly outperformed the linear model with an R^2 score near 1 and an RMSE of approximately $1,145.73., indicating excellent prediction accuracy.
- The **Decision Tree** model, with cost complexity pruning, revealed an optimal ccp_alpha value that balanced complexity and accuracy, yielding a very high R^2 score of 0.987.

Here's the visualization comparing the performance of the three models we used:



This graph effectively illustrates the superior predictive performance of the Random Forest and pruned Decision Tree models over the Linear Regression model in this context. For real estate price prediction in Boston, focusing on complex models like Random Forest can significantly improve accuracy and reduce prediction errors.

Recommendations:

Investment Focus:

- **Focus on Fiscal Features**: Since taxes and building values are significant predictors, real estate investors and managers should pay close attention to these aspects when evaluating properties for investment.
- **Consider Renovations and Improvements**: Given the importance of building value and living area, investments in property improvements could significantly increase property value.

Model Deployment:

- **Deploy Random Forest Models for Prediction**: Due to its high accuracy, the Random Forest model could be used in a real estate analysis tool to help investors and managers predict property prices and make informed decisions.
- **Use Decision Trees for Simpler Explanations**: For scenarios where explanations are necessary (e.g., presenting to stakeholders), pruned decision trees provide a good balance between simplicity and predictive power.

Further Research and Analysis:

- **Explore More Geospatial Features**: Additional analysis could be conducted on how specific locations within Boston influence property values, potentially integrating more granular location data or geographic information system (GIS) data.

- **Dynamic Market Trends**: Incorporate time series analysis to account for changes over time in the real estate market, which could help in forecasting future trends.

Research Question 6: Factors Influencing Property Prices in Boston

What are the key factors influencing property prices in Boston neighborhoods, based on property type, land size, Property Area, Living Area, Number of Floors.

What are the key factors influencing property prices in Boston neighborhoods, based on School contribute to variations in property values?

Methods: Regression analysis to quantify the impact of different variables on property prices, correlation analysis to identify relationships between property characteristics and values, and descriptive statistics to compare property prices across neighborhoods and property types.

Managerial Decision-Making: Managers can use the results of this analysis to prioritize investments by focusing on neighborhoods or property types with the highest potential for returns. The findings can guide decisions related to property development and renovation strategies, allowing managers to tailor their efforts based on factors such as property age and size that have a significant impact on prices. Understanding the correlation between amenities, school proximity, and crime rates with property values helps in marketing strategies, enabling managers to highlight these features to attract potential buyers or tenants. Additionally, this research informs risk management decisions, allowing managers to assess potential risks associated with specific neighborhoods or properties. In summary, answering this research question empowers managers with data-driven insights to make informed decisions on property investments, development, marketing, and risk mitigation in the dynamic Boston real estate market.

Tools: Using Python, along with its libraries like Statsmodels, Scikit-learn, Pandas, and NumPy, will be employed for running regression and correlation analyses. Excel will also be used for

correlation analysis and descriptive statistics. Excel will also be used for visualization of the data and analysis.

Graphs/Visualization:

- Scatter Plots- Plotting each significant independent variable against property prices to visualize their relationship.
- Regression Plots- Can illustrate the overall relationship between multiple independent variables and property prices.
- Heatmaps- Can help display the correlation matrix between property characteristics and prices.
- Bar Charts- Compare average property prices across different neighborhoods or property types.

Hypothesis Testing:

Between-Groups Variation:

- Sum of Squares: The between-groups sum of squares is \(3.04714 \times 10^{16}\). This measurement represents the variability in Total Property Value attributable to differences among the categorical variables: Land Size, Property Area, Living Area, and Number of Floors.
- F-Statistic: The calculated F-statistic is 54,449.79766, which significantly exceeds the critical F-value of 2.37196. This suggests a strong statistical significance.
- P-Value: The p-value associated with the F-statistic is approximately 0.00, indicating a very strong rejection of the null hypothesis.

- Conclusion from Between-Groups Variation: Based on these results, we reject the null hypothesis. There is noteworthy evidence that the groups differ with respect to Total Property Value, implying that variables such as Land Size, Property Area, Living Area, and Number of Floors have a significant impact on Total Property Value.

Within-Groups Variation:

- Sum of Squares: The within-groups sum of squares is \(4.31889 \times 10^{16}\). This value quantifies the variability in Total Property Value that is not explained by the group differences and is instead attributed to variability within individual groups.
- Interpretation: Although this component contributes to the total variability in property values, it does not affect the significant differences identified between the groups.

Overall Conclusion:

- Significance of Variables: The ANOVA results strongly suggest that factors such as Land Size, Property Area, Living Area, and Number of Floors play crucial roles in determining Total Property Value.
- Recommendations for Further Analysis: Given the significant findings from this ANOVA, further statistical analyses such as multiple regression modeling could be beneficial. Such analyses would help in dissecting the individual contributions and interactions of each variable, enhancing our understanding of what drives variations in Total Property Value.
- Final Takeaway: The analysis confirms that key factors including Land Size, Property Area, Living Area, and Number of Floors are integral to influencing Total Property

Value. Further investigation into these variables is recommended to fully comprehend their impacts.

This comprehensive analysis provides a clear indication of the significant factors influencing property values, guiding future studies and real estate evaluations.

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Total Property Valu	61741	4.86E+10	787667.4	7E+11		
Land size	61741	1.69E+08	2732.111	4353877		
Property Area	61741	2.28E+08	3687.776	2002178		
Living Area	61741	1.54E+08	2501.661	1079670		
Num Floors	61741	155158.5	2.513055	0.38969		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	3.05E+16	4	7.62E+15	54449.8	0	2.371961
Within Groups	4.32E+16	308700	1.4E+11			
Total	7.37E+16	308704				

The above output shows summary statistics and performs an ANOVA on five property characteristics.

Regression Analysis

Regression Analysis Summary:

- Multiple R: The Multiple R value is 0.7521, indicating a moderately strong positive linear relationship between the independent variables (such as Land Size, Property Area, Living Area, and Number of Floors) and the dependent variable (Total Property Value).

This coefficient suggests a substantial connection, though not a perfect one, highlighting the influence of the chosen variables on property values.

- R Square: The R-square value is 0.5657. This means that approximately 56.57% of the variability in Total Property Value can be explained by the linear relationship with the independent variables used in the model. This value provides a good indication of how much of the total variation in property values the model accounts for.
- Adjusted R Square: The Adjusted R-square value adjusts the R-square value for the number of variables included in the model, providing a more accurate reflection when multiple variables are used. It compensates for the inclusion of potentially unnecessary variables by penalizing excessive complexity in the model, although the exact value is not specified in the provided data.
- Standard Error of the Estimate: The standard error in this regression analysis is 551,227.2963. This value indicates the average distance that the observed values deviate from the line of best fit. In simpler terms, it measures the typical error in predicting Total Property Value using this model, which is quite substantial, suggesting variability in the data and potential room for model improvement.
- Observations: The total number of observations (data points) used in the regression analysis is essential for validating the reliability of the model, though the exact number is not mentioned. A higher number of observations generally leads to more reliable results, assuming the data quality is maintained.
- Conclusion: This regression analysis highlights a significant but not exhaustive relationship between the selected variables and Total Property Value. While over half of the variance in property values can be explained through the model, there remains a

considerable portion influenced by other factors or random variability. Further refinements and inclusion of additional relevant variables might enhance the model's predictive power and accuracy.

ANOVA:

- Overview: The ANOVA results show a very low p-value (0.00), confirming the statistical significance of the regression model. This suggests that the model reliably explains variations in Total Property Value based on the selected variables.
- F-Statistic: An F-statistic of 8933.976 highlights the overall significance of the regression model. This high value indicates that the explanatory variables collectively have a strong impact on Total Property Value.
- Residual Sum of Squares: This metric measures the variation in Total Property Value not explained by the model, suggesting areas where the model might be refined or indicating the influence of variables not included in the model.

Coefficients:

- Intercept: The intercept, at approximately -$1,674,281.194, might initially seem counterintuitive. This value represents the theoretical Total Property Value when all independent variables are zero. The negative intercept could indicate higher baseline prices or reflect adjustments for overestimation within the model.
- Independent Variables: Coefficients for variables like Land Size, Property Area, Living Area, and Number of Floors indicate the specific increase in Total Property Value for each unit increase in these variables, showcasing their direct impact on property valuation.

- Property Types: The coefficients for different property types (Commercial, Dual Family, Residential, Single Family, Tax Exempt) relative to the reference category (Triple Family) highlight how property classification influences valuation differently.

Interpretation of Coefficients:

Each coefficient not only quantifies the impact of an independent variable on the dependent variable but also gives insights into market trends and buyer preferences. For instance, positive coefficients for Single Family homes and additional floors indicate higher valuation for such properties, reflecting their desirability in the market.

Top 5 Significant Variables:

1. Living Area:

- Coefficient: 966.1065
- P-value: 0
- Detailed Interpretation: This finding indicates that Living Area is a critical factor in property valuation. The substantial increase in property value per square foot suggests that market demand is highly sensitive to living space, which is often a key consideration for potential buyers.

2. Single Family:

- Coefficient: 1,007,660.779
- P-value: 0

- Detailed Interpretation: Single Family homes are highly valued over the baseline Triple Family category. This could reflect a preference for the privacy, space, and potentially the community environments typically associated with Single Family homes.

3. Number of Floors:

- Coefficient: 300,446.6378
- P-value: 0
- Detailed Interpretation:Properties with more floors command a higher Total Property Value, potentially due to increased living space or the premium associated with multi-story buildings which may offer additional amenities or views.

4. Commercial:

- Coefficient: -720,263.3759
- P-value: 3.98×10^-75
- Detailed Interpretation: Commercial properties show a significant decrease in value compared to Triple Family homes. This might reflect market perceptions or the specific utility of commercial spaces, which can vary widely depending on economic conditions and business viability.

5. Residential:

- Coefficient: -649,462.8211
- P-value:1.07×10^-71

- Interpretation: The negative coefficient for Residential properties as compared to Triple Family homes might indicate lower market valuation, possibly due to factors like size, amenities, or location preferences within the studied data set.

Conclusion and Market Implications:

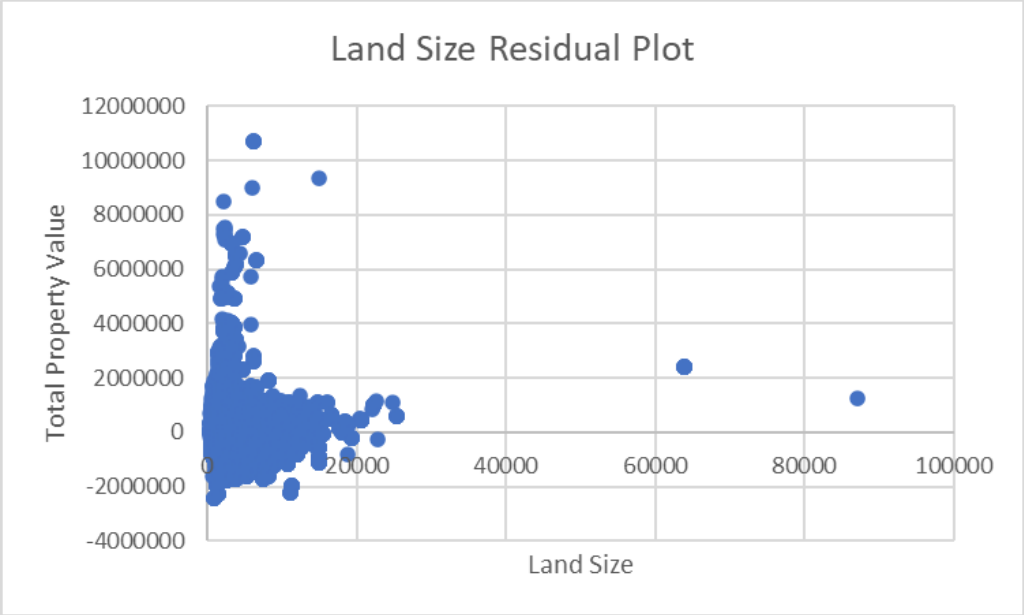
This regression analysis provides a robust framework for understanding what drives property values. Stakeholders such as developers, investors, and real estate agents can use these insights to make informed decisions. For instance, emphasizing larger living areas or additional floors can be effective strategies for increasing property values. Moreover, understanding the negative perceptions toward certain property types like Commercial and Residential compared to Triple Family homes can guide marketing strategies and investment decisions in the real estate market.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.7521257							
R Square	0.565693							
Adjusted R Squar	0.5656297							
Standard Error	551227.3							
Observations	61741							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	9	2.44314E+16	2.7146E+15	8933.976	0			
Residual	61731	1.87571E+16	3.03852E+11					
Total	61740	4.31885E+16						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95	Upper 95%
Intercept	-1674281	16360.01388	-102.339839	0	-1706347	-1642216	-2E+06	-2E+06
Land size	-23.46933	1.285431002	-18.2579461	2.80875E-74	-25.98878	-20.94988	-25.99	-20.95
Property Area	-310.9613	5.432654219	-57.2393002	0	-321.6093	-300.3133	-321.6	-300.3
Living Area	966.10651	7.837929717	123.2604201	0	950.74415	981.46887	950.74	981.47
Num Floors	300446.64	5669.876962	52.98997488	0	289333.67	311559.61	289334	311560
Commercial	-720263.4	39219.39559	-18.364979	3.97571E-75	-797133.5	-643393.3	-8E+05	-6E+05
Dual Family	578396.09	6200.203411	93.28663064	0	566243.67	590548.5	566244	590548
Residential	-649462.8	36224.22801	-17.9289624	1.0661E-71	-720462.4	-578463.2	-7E+05	-6E+05
Single Family	1007660.8	6693.946099	150.533148	0	994540.63	1020780.9	994541	1E+06
Tax Exempt	156193.73	42378.42488	3.68568982	0.000228279	73131.916	239255.54	73132	239256

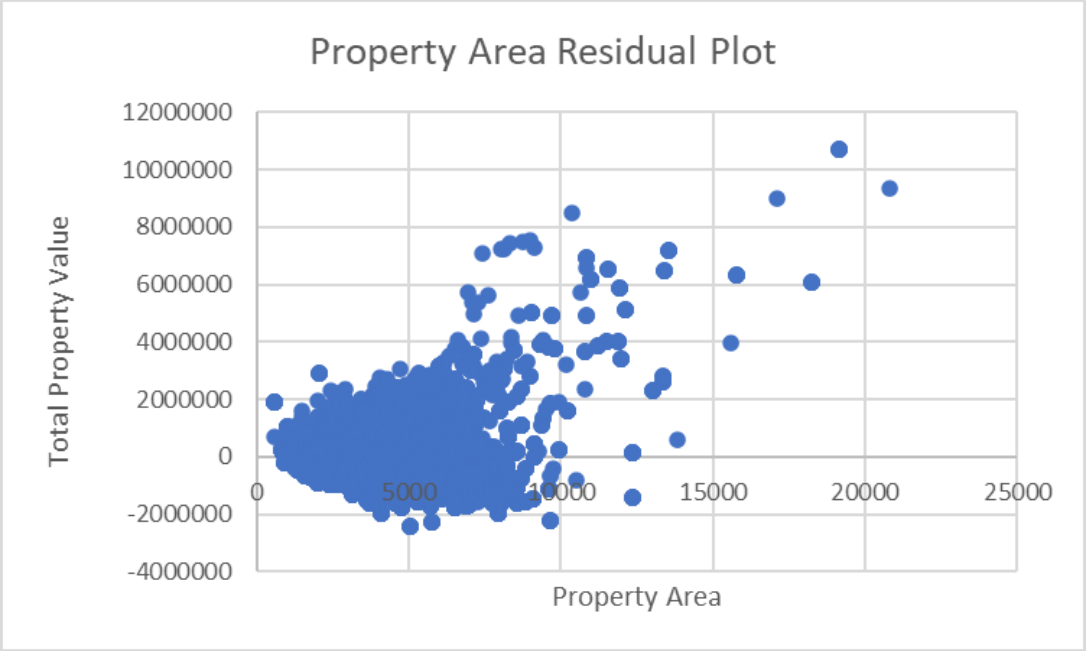
The screenshot shows a summary output for a regression analysis on property values and the top 5 significant variables are Living Area, Single Family, Number of Floors, Commercial and Residential.

Plots:

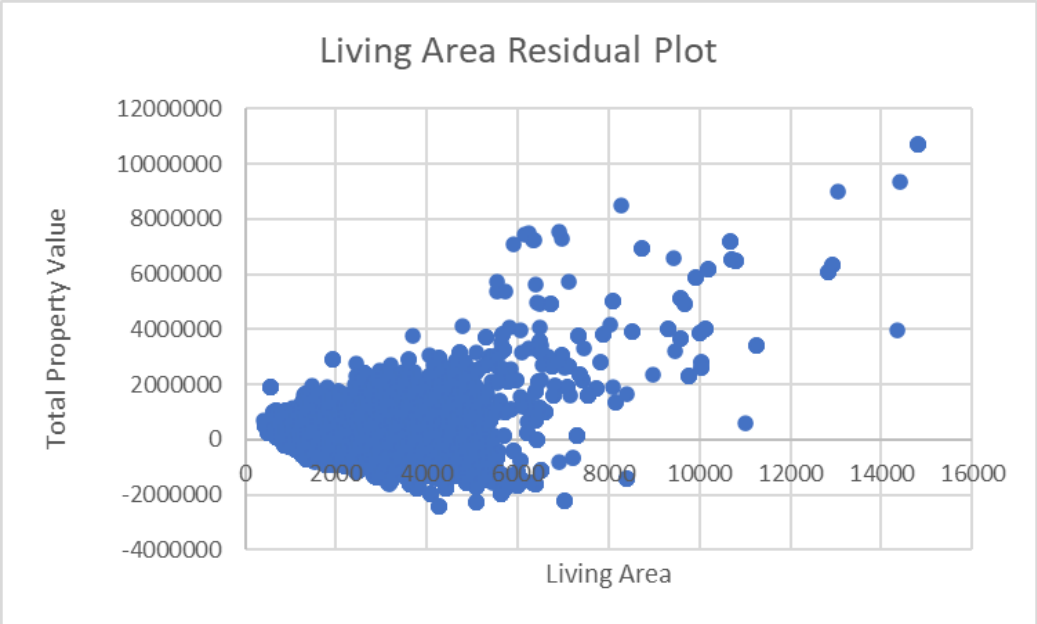
The following are the plots for which Y represents Total Property Value and X represents Land Size, Property Area, Living Area, Number of Floors, Commercial, Dual family, Residential, Single Family and Tax Exempt.



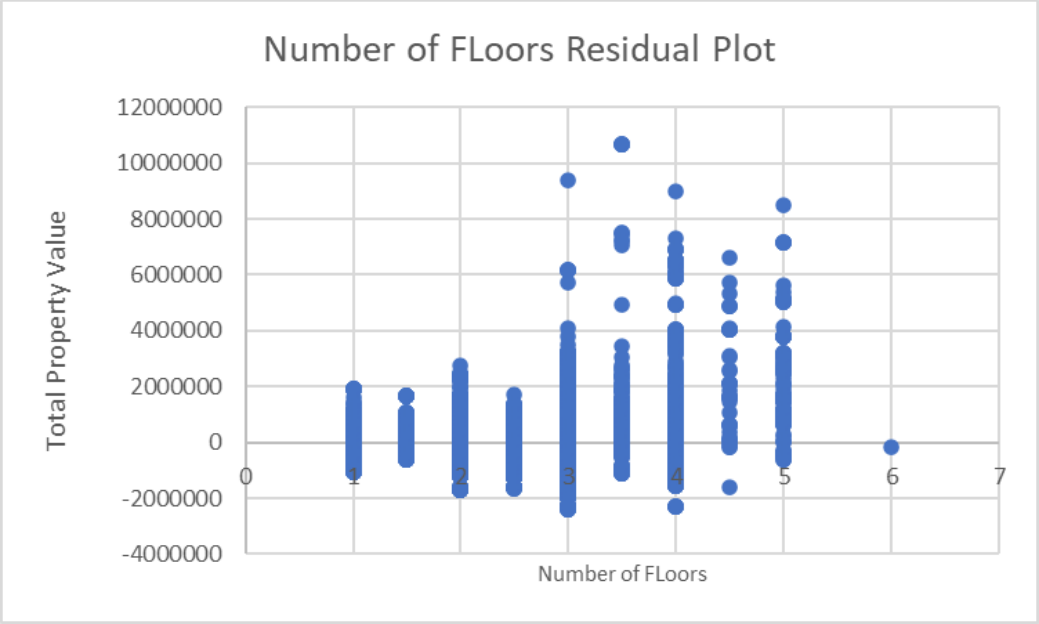
The above regression analysis output shows a moderately strong positive relationship (R=0.75) between Land Size and total property value, explaining 56.6% of the variance.



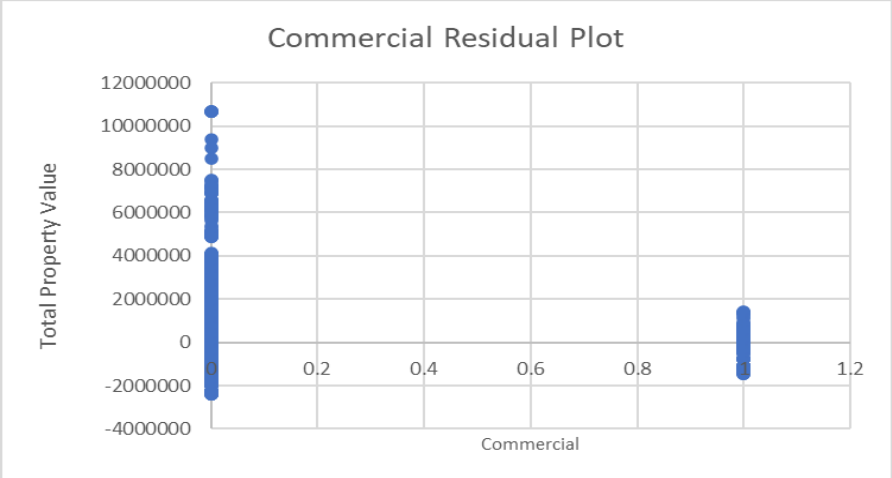
The above property area residual plot shows no clear pattern, suggesting the regression model may adequately capture the relationship between property area and total property value.



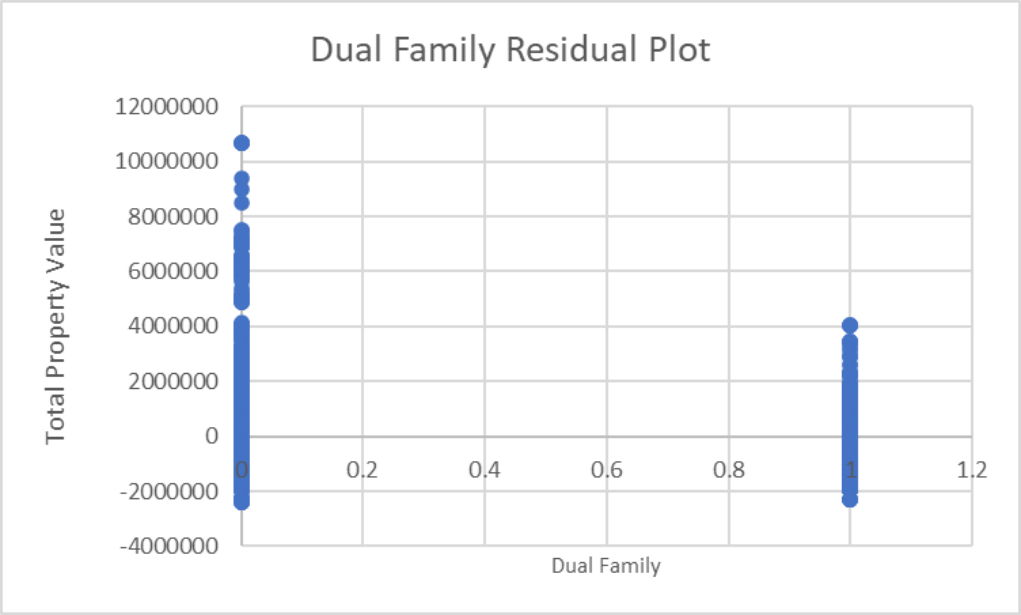
The screenshot displays a residual plot showing the relationship between living area and total property value. As the living area increases, the data points become more spread out vertically, suggesting greater variability in property values for larger living areas.



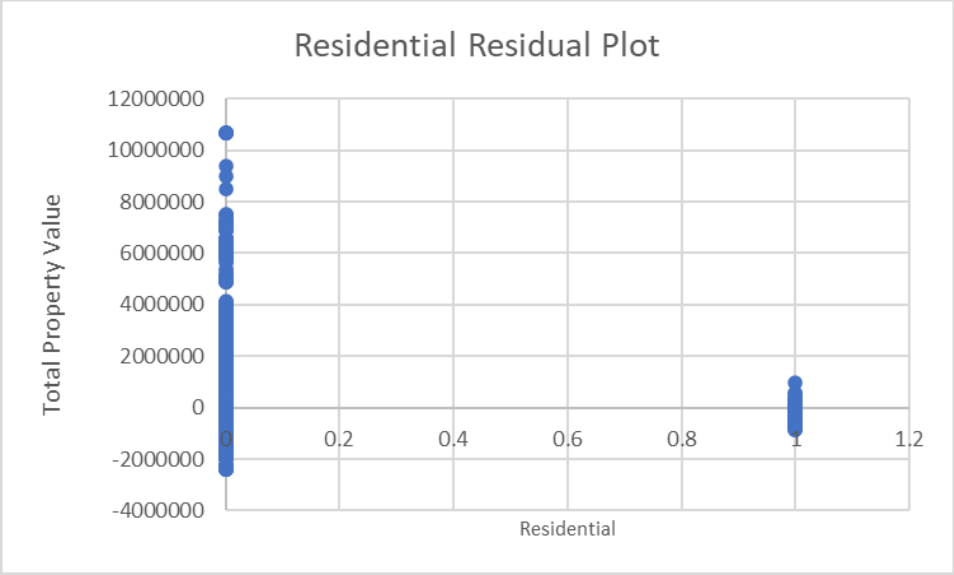
This residual plot visualizes the relationship between the number of floors in a property and the residual (difference between predicted and actual) total property value, showing some variability but no clear increasing or decreasing trend.



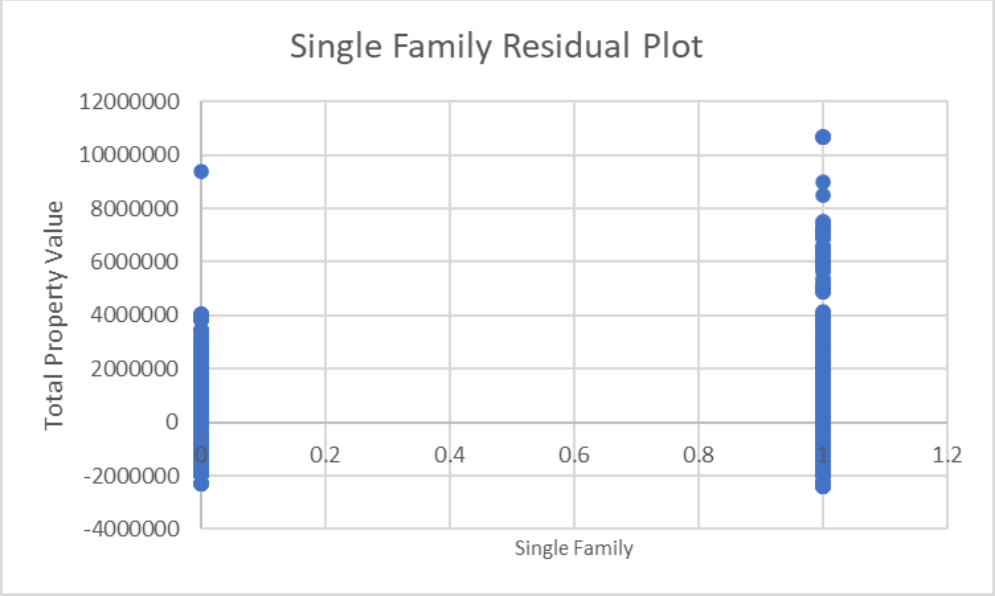
This residual plot shows the relationship between a property's commercial usage and the residual total property value. There appear to be only a few distinct commercial usage values represented, with the majority of data points clustered around the presumed non-commercial value. A small number of points have significantly higher residual values, suggesting that commercial properties may have larger deviations between predicted and actual total values compared to non-commercial properties in this dataset.



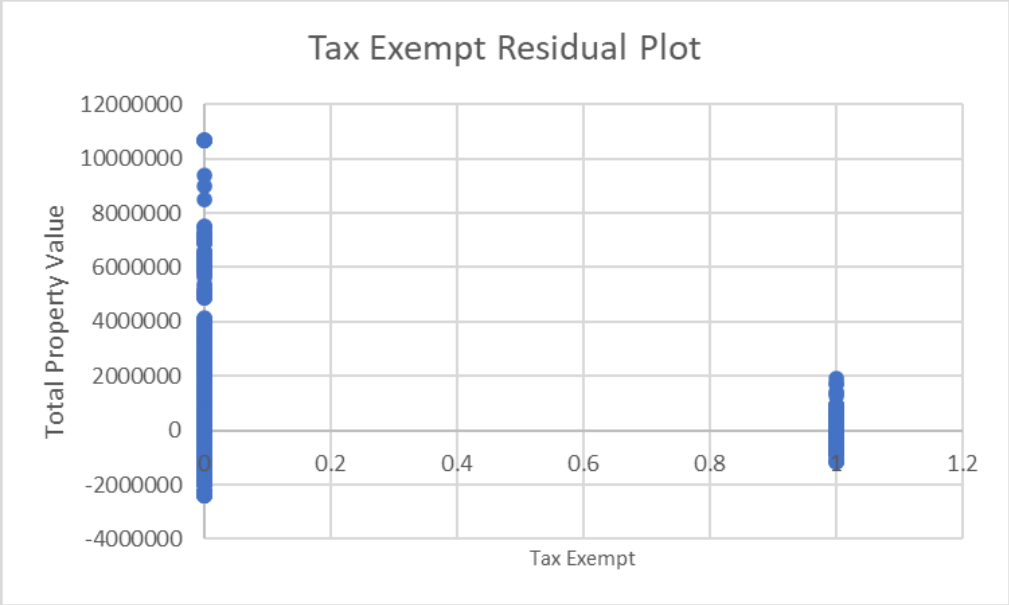
This residual plot shows that properties classified as dual family tend to have higher positive residuals (actual values greater than predicted) compared to single-family properties, with a few outliers exhibiting very large positive and negative residuals.



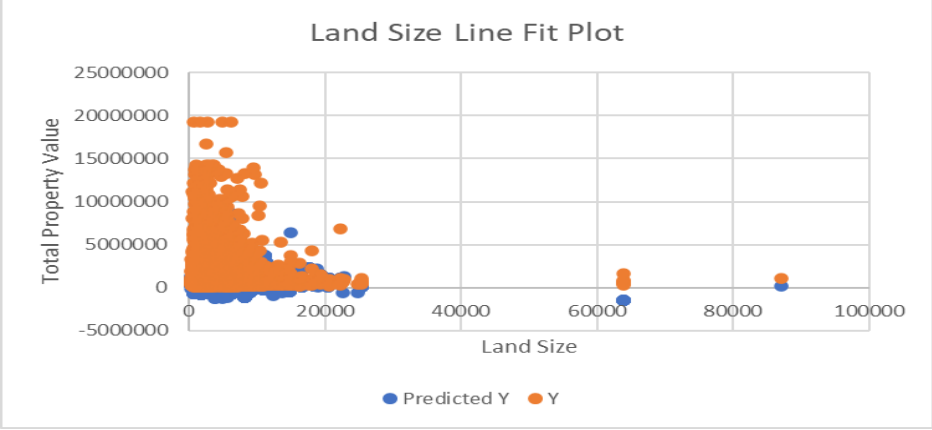
This residual plot shows the relationship between a property's residential usage and the residual total property value . There are a few outliers with significantly higher positive and negative residual values, suggesting some residential properties may have larger deviations between predicted and actual total values compared to the majority in this dataset.



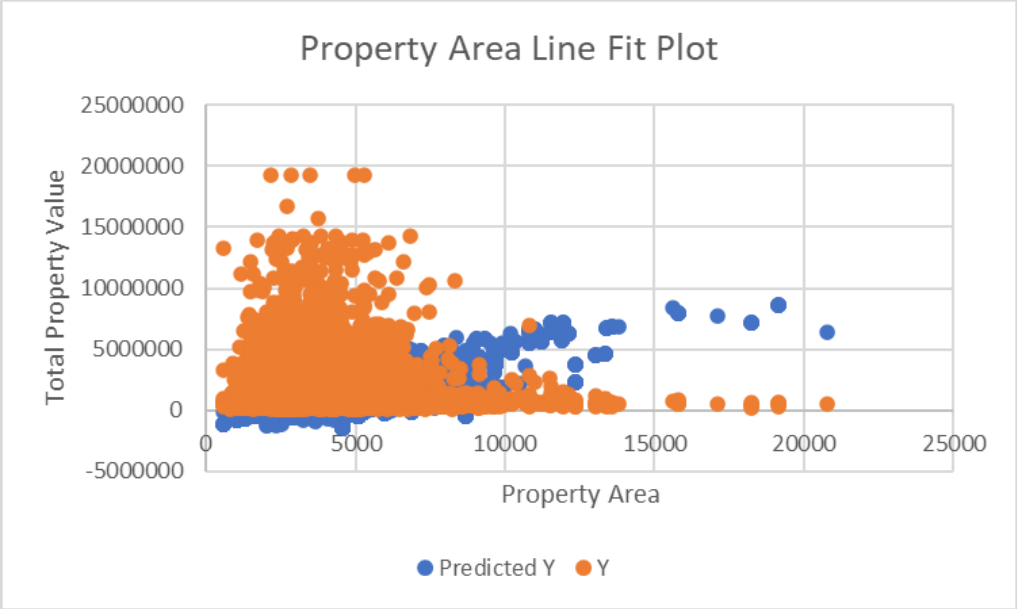
The residual plot for single-family properties shows a roughly uniform distribution of residuals. However, as the predictor variable increases, the residuals become increasingly positive, indicating that the model tends to underestimate the actual property values for higher-value single-family properties.



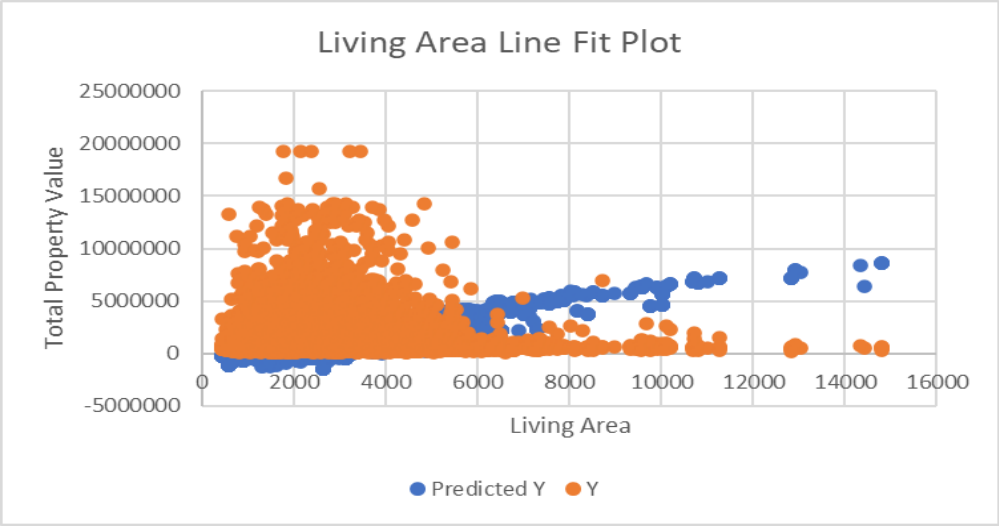
Here, for higher values of the predictor, the residuals become increasingly negative, suggesting that the model tends to overestimate the actual property values for higher-value tax-exempt properties.



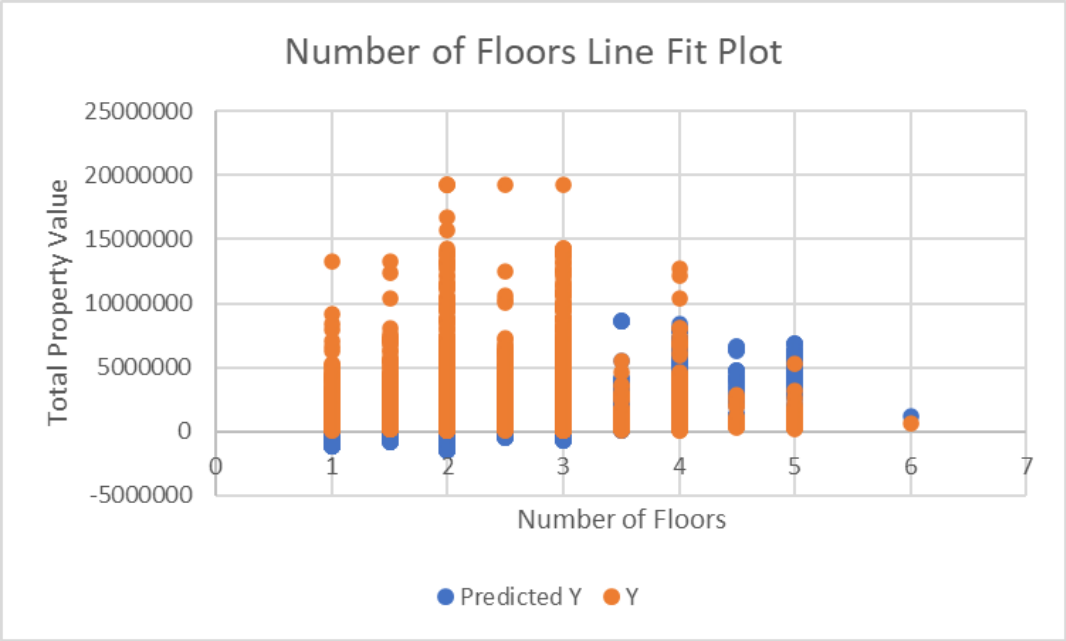
As we can see as land size increases, property value tends to increase as well, but with significant variability around the fitted line. The plot also indicates the presence of some potentially influential points or outliers at higher land sizes.



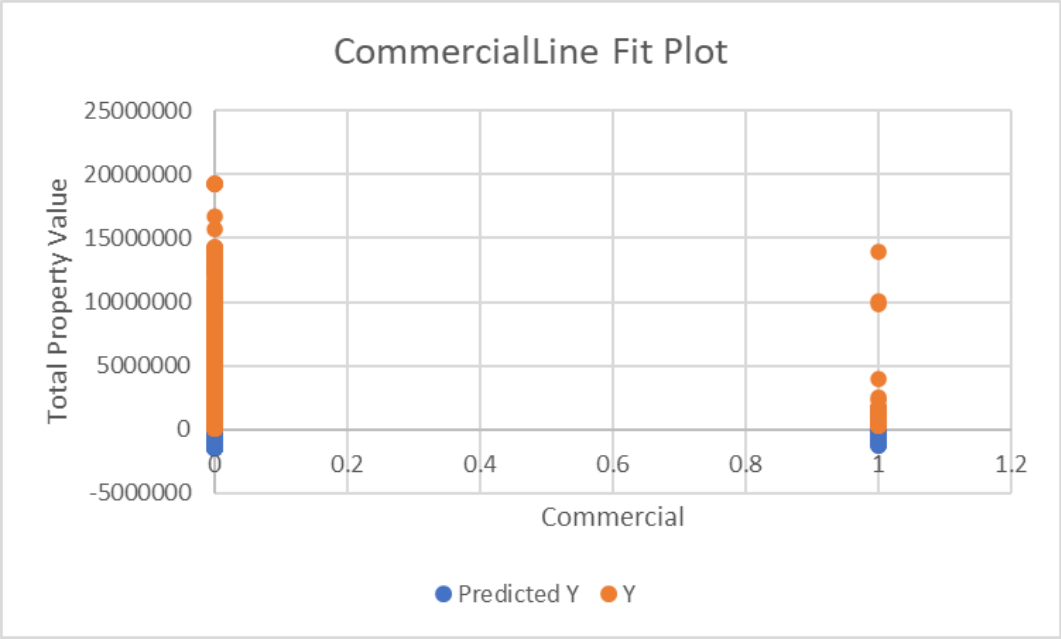
The relationship appears to be more linear and tighter compared to the land size plot, with less variability around the fitted line, especially for smaller property areas and some potentially influential points or outliers at higher property areas.



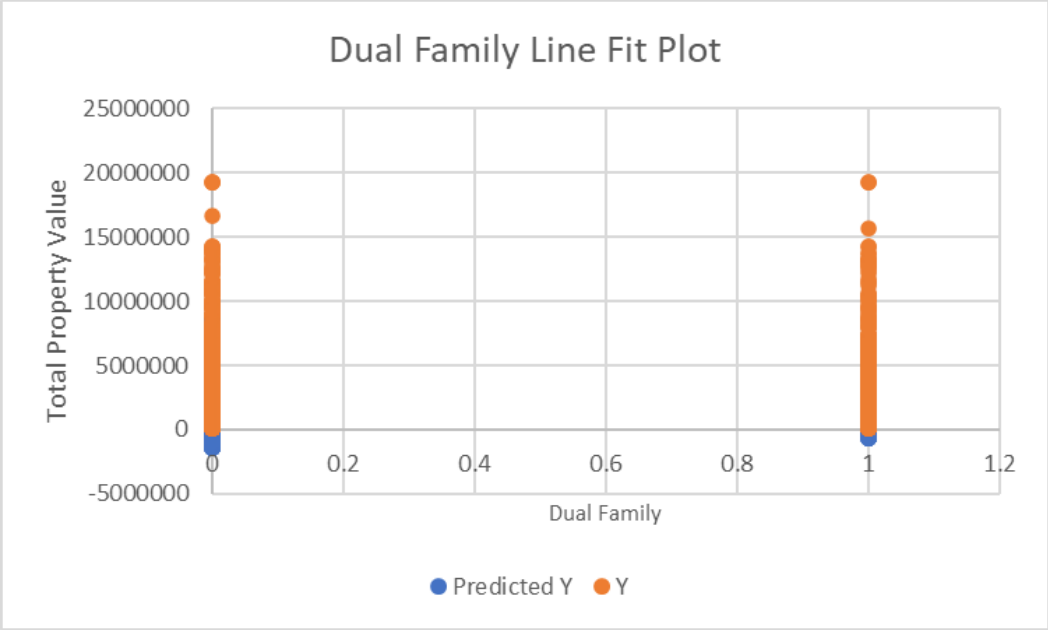
This line fit plot illustrates the positive relationship between living area and property value, with some variability and potential outliers at higher living area values.



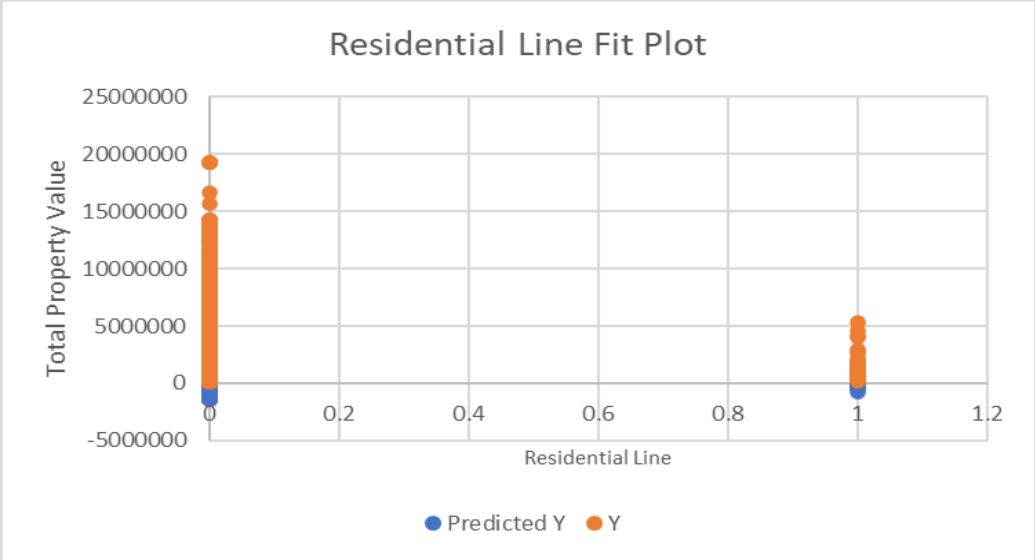
In the above image as the number of floors increases, the total property value tends to increase as well, with some fluctuations.



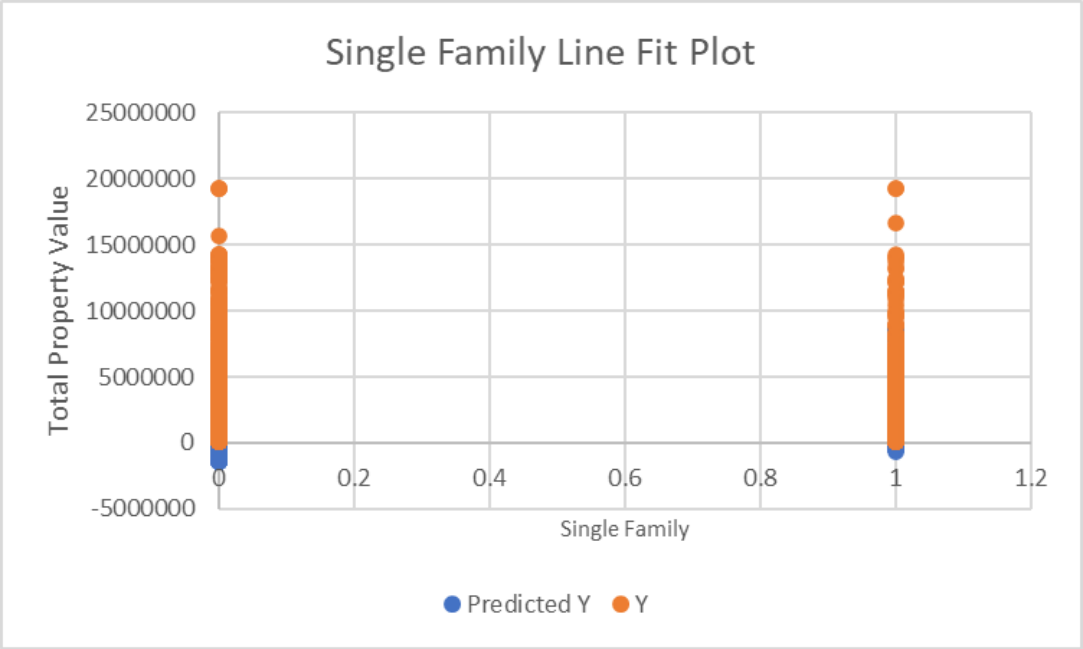
The above data points suggest that as the commercial property value increases, the total property value also increases. However, there are fewer data points for higher commercial property values.



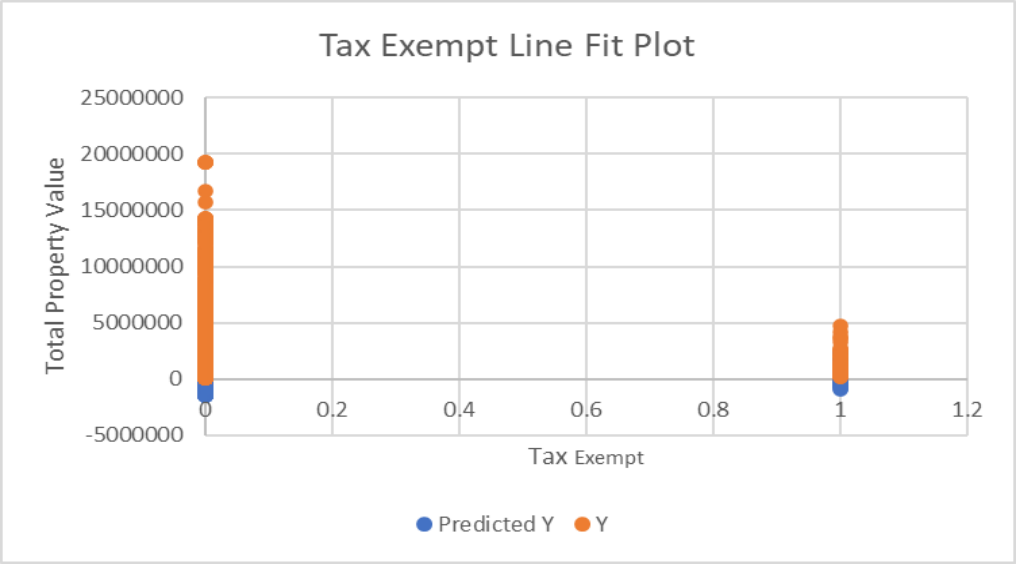
The above data points indicate that as the dual family property value increases, the total property value tends to increase as well, following a roughly linear trend.



The above data points suggest that as the residential property value increases, the total property value also increases, although there are fewer data points for higher residential property values.



The above data points indicate that as the single-family property value increases, the total property value tends to increase as well, following a roughly linear trend similar to the dual family property type.



The above data points suggest that as the tax-exempt value increases, the total property value also increases, although there are fewer data points for higher tax exempt property values.

Here is the snippet of regression output in python:

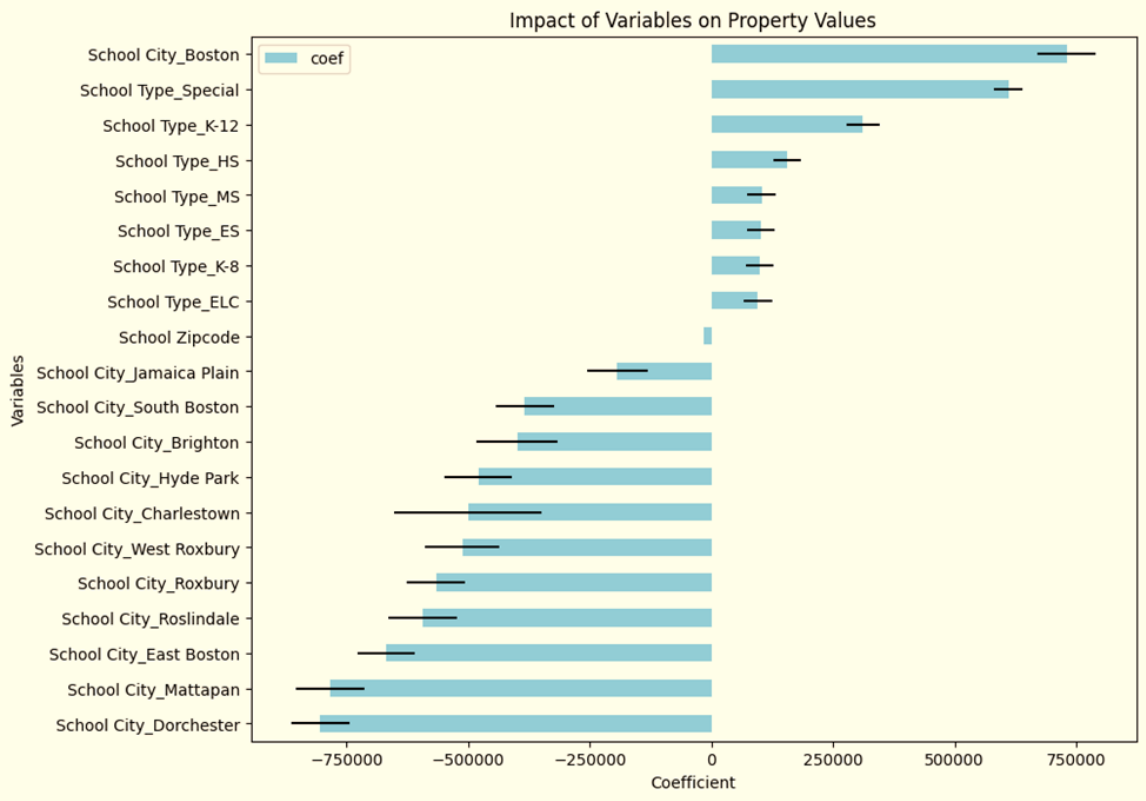
OLS Regression Results

Dep. Variable:	Total Property Value	R-squared:	0.366
Model:	OLS	Adj. R-squared:	0.366
Method:	Least Squares	F-statistic:	1753.
Date:	Sun, 05 May 2024	Prob (F-statistic):	0.00
Time:	15:19:41	Log-Likelihood:	-8.9576e+05
No. Observations:	60701	AIC:	1.792e+06
Df Residuals:	60680	BIC:	1.792e+06
Df Model:	20		
Covariance Type:	nonrobust		

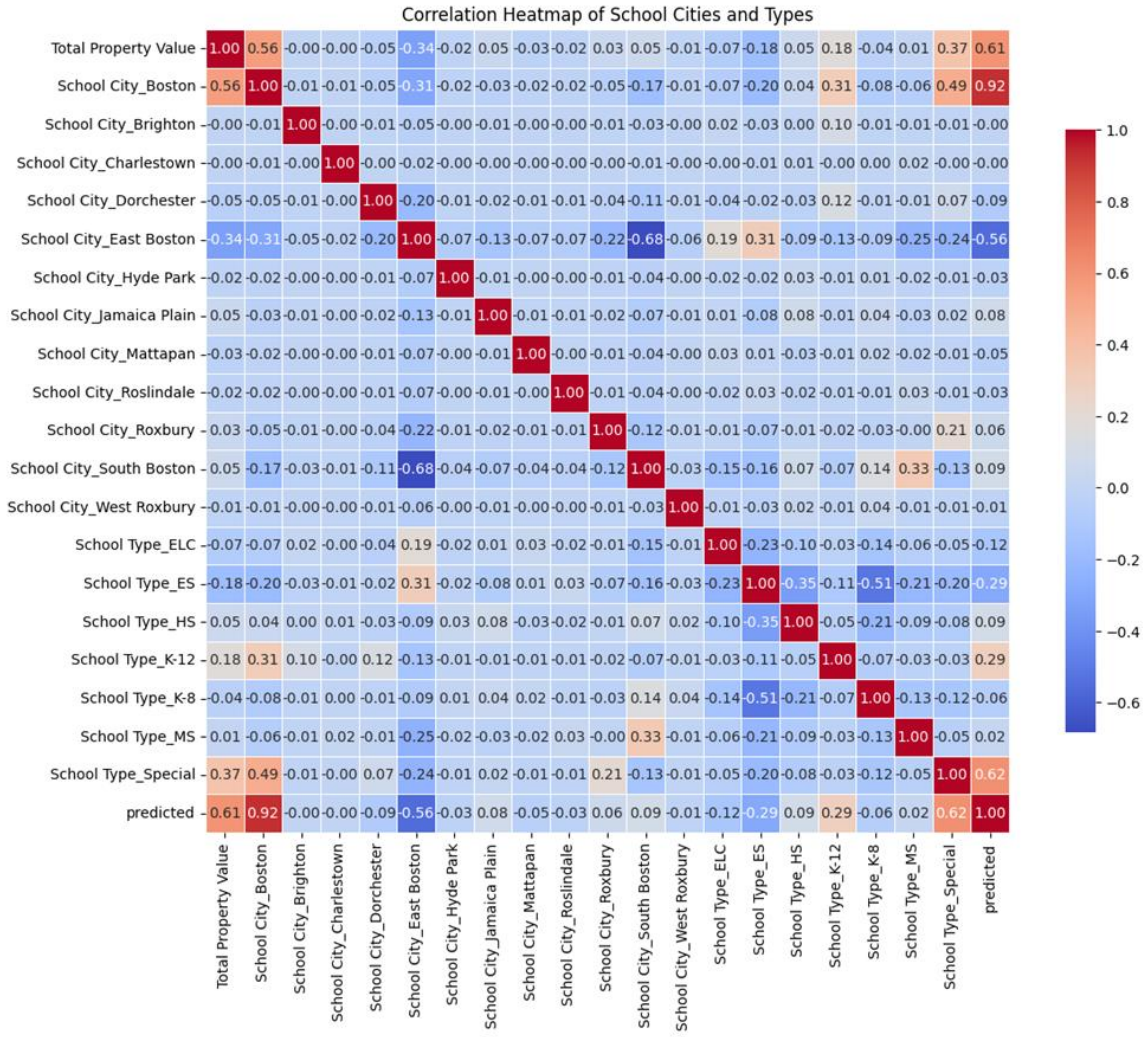
	coef	std err	t	P>|t|	[0.025	0.975]
const	3.493e+07	1.54e+06	22.663	0.000	3.19e+07	3.8e+07
School Zipcode	-1.59e+04	721.916	-22.027	0.000	-1.73e+04	-1.45e+04
School City_Boston	7.305e+05	6.02e+04	12.135	0.000	6.12e+05	8.48e+05
School City_Brighton	-3.997e+05	8.3e+04	-4.813	0.000	-5.63e+05	-2.37e+05
School City_Charlestown	-5.009e+05	1.51e+05	-3.326	0.001	-7.96e+05	-2.06e+05
School City_Dorchester	-8.039e+05	6.05e+04	-13.299	0.000	-9.22e+05	-6.85e+05
School City_East Boston	-6.684e+05	5.87e+04	-11.389	0.000	-7.83e+05	-5.53e+05
School City_Hyde Park	-4.797e+05	6.95e+04	-6.899	0.000	-6.16e+05	-3.43e+05
School City_Jamaica Plain	-1.938e+05	6.22e+04	-3.117	0.002	-3.16e+05	-7.19e+04
School City_Mattapan	-7.843e+05	6.98e+04	-11.239	0.000	-9.21e+05	-6.47e+05
School City_Roslindale	-5.926e+05	7.04e+04	-8.421	0.000	-7.31e+05	-4.55e+05
School City_Roxbury	-5.666e+05	6.06e+04	-9.343	0.000	-6.85e+05	-4.48e+05
School City_South Boston	-3.834e+05	5.88e+04	-6.522	0.000	-4.99e+05	-2.68e+05
School City_West Roxbury	-5.122e+05	7.69e+04	-6.664	0.000	-6.63e+05	-3.62e+05
School Type_ELC	9.549e+04	2.97e+04	3.212	0.001	3.72e+04	1.54e+05
School Type_ES	1.009e+05	2.79e+04	3.614	0.000	4.62e+04	1.56e+05
School Type_HS	1.563e+05	2.83e+04	5.525	0.000	1.01e+05	2.12e+05
School Type_K-12	3.116e+05	3.45e+04	9.033	0.000	2.44e+05	3.79e+05
School Type_K-8	9.862e+04	2.81e+04	3.513	0.000	4.36e+04	1.54e+05
School Type_MS	1.029e+05	3.02e+04	3.412	0.001	4.38e+04	1.62e+05
School Type_Special	6.104e+05	2.93e+04	20.860	0.000	5.53e+05	6.68e+05

Omnibus:	85015.136	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	48311071.484
Skew:	8.012	Prob(JB):	0.00
Kurtosis:	140.275	Cond. No.	1.30e+06

The OLS Regression Results provide a detailed statistical analysis concerning how different school-related factors, particularly geographic location and school type, affect the Total Property Value in a dataset comprising 60,701 observations. The model explains approximately 36.6% of the variance in property values, as indicated by an R-squared value of 0.366. This is a moderate level of explanatory power, suggesting that while school-related factors are important, other variables not included in this model also significantly impact property values.



The regression model has effectively identified key factors influencing property values linked to school characteristics and locations. The visualized coefficients with their error bars provide a clear and actionable understanding, which can facilitate strategic decision-making in real estate and urban educational policy planning.



The correlation heatmap provides a visual representation of how 'School City' and 'School Type' variables in your dataset are interrelated, using color gradients to indicate the strength and direction of correlations.

Key Insights from the Heatmap:

- Color Coding: Red indicates positive correlations, blue shows negative correlations, and the color intensity signifies the strength of these relationships.

- Correlation Coefficients: Values range from -1 (strong negative correlation) to +1 (strong positive correlation), with 0 indicating no correlation.
- Urban Planning: Insights from the heatmap can guide decisions on where to place new schools, amenities, or housing developments based on existing school types and locations.
- Resource Allocation: Policymakers can use these correlations to strategically allocate educational resources and infrastructure investments.
- Investment Opportunities: Real estate investors might target areas with certain school types that positively correlate with property values, suggesting potential for growth.

This concise analysis enables stakeholders to make informed decisions about educational placements, resource distribution, and investment strategies based on the relationships highlighted in the heatmap.

Conclusion

The OLS regression analysis and visualizations have provided substantial insights into the factors influencing property prices in Boston, with a specific focus on the role of schools. The analysis highlighted that school type and geographic location, represented by school city and zip code, significantly impact property values. Different school types and locations have varying degrees of influence on property prices, with properties near K-12 and Special schools showing higher values. The correlation heatmap revealed significant relationships between different school types and locations, offering a detailed perspective on how these factors interconnect and influence property valuations.

Recommendations:

1. Strategic Investments:

Real estate investors and developers should focus on properties near schools with positive impacts on property values, particularly K-12 and Special schools, as these areas offer the highest potential for returns.

Consider the geographic trends identified through the analysis to target investments in zip codes and neighborhoods that consistently demonstrate higher property values.

2. Development and Renovation:

Property developers should tailor their projects to meet the demands of families and individuals likely to be attracted to areas near high-value schools. This includes developing properties with features that appeal to this demographic, such as enhanced safety measures, family-friendly amenities, and community spaces.

Renovation strategies should also consider the local school types to increase property appeal and marketability, potentially increasing property values further.

3. Marketing Strategies:

Marketers and real estate agents should highlight proximity to desirable schools and advantageous school districts in their promotional materials, using these attributes as key selling points.

Tailor marketing campaigns to emphasize the benefits of living near schools that enhance property values, aligning promotional strategies with the insights derived from the data analysis.

4. Policy Development:

Policymakers should consider the insights from this analysis to guide decisions regarding urban planning and educational funding. Enhancing school quality in lower-value neighborhoods could be a strategy to elevate property values and improve overall community welfare.

Encourage developments around high-value school types to ensure balanced growth and accessibility for all residents.

5. Further Statistical Analysis:

Expand the current model to include more variables such as public transportation access, crime rates, and demographic changes over time, which could provide a deeper understanding of their collective impact on property values.

Employ advanced modeling techniques to capture more complex relationships and interactions between factors affecting property prices.

By implementing these recommendations, stakeholders can optimize their approaches to real estate investment, development, and management, leveraging educational infrastructure as a critical element in driving property value growth in Boston's dynamic market.

End of all research questions.

Conclusion of the Report

This comprehensive study aimed to understand the key factors influencing property values in Boston, with a particular emphasis on the impact of crime rates, property characteristics, location, and school quality. Through extensive data analysis and modeling, several significant findings emerged:

1. **Location and Neighborhood Factors**: The analysis revealed that property values are strongly influenced by location and neighborhood characteristics. Areas with higher crime rates, such as East Boston, Boston, and South Boston, generally exhibited lower property values compared to safer neighborhoods. However, proximity to city centers, amenities, and quality schools also played a crucial role in determining property values.
2. **Property Characteristics**: Interior features like the number of bedrooms, bathrooms, kitchens, and floors were positively associated with higher property values. Exterior factors, such as the use of durable materials like brick and vinyl, also contributed to increased property valuations, particularly in the $400K - $600K price range. Larger properties with more square footage and preferred architectural styles (e.g., Row Middle, Decker, Semi-Detached, and conventional) tended to command higher prices.
3. **Heating and Cooling Systems**: Properties with central air conditioning and heating systems, especially those using forced air, were associated with higher property values due to improved comfort and energy efficiency. However, the presence of ductless AC systems did not significantly impact property values.
4. **Property Age and Remodeling**: The analysis revealed a nuanced relationship between property age and value. While older properties generally showed a slight increase in

value over time, more recently remodeled properties tended to have slightly lower values. However, the influence of property age alone was limited, as other factors not included in the analysis also contributed to variations in property values.

5. **Crime Rates and Property Taxes**: The study found a moderate negative correlation between property gross tax and crime rates, suggesting that areas with higher property taxes tend to exhibit lower crime rates. Regression analysis further confirmed that crime rates, property characteristics, and location significantly influence property values.
6. **School Quality and Location**: The analysis highlighted the significant impact of school type and geographic location on property values. Properties near K-12 and Special schools generally exhibited higher valuations, indicating a strong influence of school quality and location on property prices.
7. **Model Performances**: The study employed various machine learning models, with the Random Forest model and the Decision Tree model with cost complexity pruning outperforming the baseline Linear Regression model, demonstrating excellent prediction accuracy for property values.

In conclusion, this comprehensive analysis has provided valuable insights into the complex interplay of factors influencing property values in Boston. By considering crime rates, property characteristics, location, school quality, and leveraging advanced machine learning techniques, this study offers a robust framework for understanding and predicting property valuations, enabling informed decision-making for stakeholders in the real estate market.

References:

1. *"Applications." Boston PD Crime Hub. Accessed March 10, 2024. https://boston-pd-crime-hub-boston.hub.arcgis.com/pages/apps*

2. *"Analyze Boston." PROPERTY ASSESSMENT FY2019, 2019. https://data.boston.gov/dataset/property-assessment/resource/695a8596-5458-442b-a017-7cd72471aade*

3. *Crawford, Chris. "Boston Public Schools." Kaggle, September 18, 2018. https://www.kaggle.com/datasets/crawford/boston-public-schools*