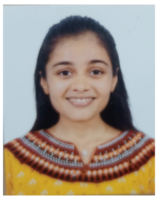
Exploratory Data Analysis on

Housing Price over the Years

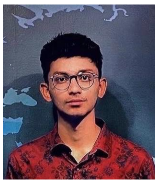
by

Group 6

Param Andharia
ID: 202203041
Course:
BTech(MNC)



Nakshi Shah
ID: 202411063
Course:
MTech(ICT-ML)



Rudra Patel
ID: 202201193
Course:
BTech(ICT)

Course Code: IT 462
Semester: Autumn 2024

Under the guidance of

Dr. Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

Acknowledgment

I am writing this letter to express my heartfelt gratitude for your guidance and support throughout the duration of our project and the course. Your invaluable assistance has played a pivotal role in shaping the successful completion of this endeavor.

I am extremely fortunate to have had the opportunity to work under your mentorship. Your expertise, encouragement, and willingness to share your knowledge have been instrumental in elevating the quality and scope of my project. Your constructive feedback and insightful suggestions have helped me overcome challenges and develop a deeper understanding of the subject matter.

Furthermore, I would like to extend my appreciation to the entire team at DA-IICT for fostering an environment of collaboration and innovation. The resources and facilities provided have been crucial in conducting comprehensive research and analysis.

I would also like to express my gratitude to my peers and colleagues who have been supportive throughout this journey. Their valuable input and camaraderie have been a constant source of motivation.

Completing this project has been a tremendous learning experience, and I am confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for my future endeavors.

Once again, thank you for your unwavering guidance and belief in my abilities. Your mentorship has been invaluable, and I am truly grateful for the opportunity to work with you.

Sincerely,
Param Andharia, 202203041
Nakshi Shah, 202411063
Rudra Patel, 202201193

DECLARATION

We, Param Andharia, Nakshi Shah, and Rudra Patel hereby declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

We acknowledge that the data used in this project is obtained from the site: `https://catalog.data.gov/dataset/real-estate-sales-2001-2018`. We also declare that we have adhered to the terms and conditions mentioned in the website for using the dataset. We confirm that the dataset used in this project is true and accurate to the best of our knowledge.

We acknowledge that we have received no external help or assistance in conducting this project, except for the guidance provided by our mentor Prof. Gopinath Panda. We declare that there is no conflict of interest in conducting this EDA project.

We hereby sign the declaration statement and confirm the submission of this report on 2nd July, 2023.

Param	Nmshah	
Param Andharia *ID:* 202203041 *Course:* BTech(MnC)	Nakshi Shah *ID:* 202411063 *Course:* MTech(ICT-ML)	Rudra Patel *ID:* 202201193 *Course:* BTech(ICT)

CERTIFICATE

This is to certify that Group 6 comprising Param Andharia, Nakshi SHah, and Rudra Patel has successfully completed an exploratory data analysis (EDA) project on the Housing Price Dataset which was obtained from `https://catalog.data.gov/dataset/real-estate-sales-2001-2018`.

The EDA project presented by Group 6 is their original work and has been completed under the guidance of the course instructor, Prof. Gopinath Panda, who has provided support and guidance throughout the project. The project is based on a thorough analysis of the Housing Price dataset, and the results presented in the report are based on the data obtained from the dataset.

This certificate is issued to recognize the successful completion of the EDA project on the Housing Price Dataset which demonstrates the analytical skills and knowledge of the students of Group 6 in the field of data analysis.

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

December 2, 2024

Contents

List of Figures

List of Tables

Abstract

This report analyzes a real estate(housing price) transaction dataset obtained from an open government platform of USA, comprising over 1 million records with columns such as List Year (varying from 2001 to 2022), Town, Address, Assessed Value, Property Type, Residential Type, and Sale Amount (our prediction variable). The study involves data preprocessing, exploratory analysis, and regression modeling to uncover trends and relationships. Our key findings include a moderate positive correlation between Assessed Values and Sale Amounts, along with other regional variations such as Address and Town.

Chapter 1. Introduction

1.1 Objective

This study aims to perform exploratory data analysis (EDA) on a housing dataset over the years 2001 to 2022 to uncover insights into property transactions in various Towns of USA, and using those to male predictions on Sales Amount (actual sale price of the property).

1.2 Background

This dataset was retrieved from the Government of USA's open data platform, which provides valuable insights into real estate transactions across various regions. It includes attributes like assessed values, sale amounts, and property types, spanning over ten lakh records.

1.3 Significance

Analyzing this dataset can assist researchers and analysts in understanding market trends, identifying anomalies, and making informed decisions in housing schemes.

Chapter 2. Dataset Description

2.1 Source Details

The dataset was sourced from the Government Open Data Platform (Click Here, a repository for publicly accessible data.

2.2 Features

Before any pre-processing, key attributes in the dataset included:

Column Name	**Description**	**Data Type**
Serial Number	Unique transaction identifier	Integer
List Year	Year the property was listed	Integer
Date Recorded	Date the transaction was recorded	Date
Town	Name of the town where the property is located	String
Address	Full address of the property	String
Assessed Value	Government-assessed property value	Float
Sale Amount	Actual sale price of the property	Float
Sales Ratio	Ratio of sale amount to assessed value	Float
Property Type	General type of property (e.g., residential)	String
Residential Type	Specific type of residential property	String
Non Use Code	Code indicating reasons for non-use of property	String
Assessor Remarks	Additional notes from the property assessor	String
OPM Remarks	Additional notes from Office of Property Mgmt	String
Location	Geographic coordinates or descriptive location	String

Table 2.1: Dataset Features and Descriptions

2.3 Summary Statistics for Numerical Columns

Table 2.2: Summary Statistics

Statistic	Serial Number	List Year	Assessed Value	Sale Amount	Sales Ratio
Count	1,048,575	1,048,575	1,048,575	1,048,575	1,048,575
Mean	511,561.4	2010.857	280,725.3	404,686.1	9.999754
Std	7,454,481.0	6.716	1,661,631.0	5,245,596.0	1,843.322
Min	0.0	2001.0	0.0	0.0	0.0
25%	30,565.0	2004.0	88,340.0	144,000.0	0.4738
50%	80,068.0	2011.0	139,730.0	232,000.0	0.6075
75%	160,668.0	2017.0	226,800.0	375,000.0	0.7716
Max	2,000,500,000.0	2022.0	881,510,000.0	5,000,000,000.0	1,226,420.0

Chapter 3. Methodology

3.1 Data Loading and Preprocessing

3.1.1 Steps

1. **Data Import**: The dataset was loaded from an online source in Excel format. The initial data import process included inspecting the first few rows (by df.head) to ensure correct structure.

Table 3.1: First Five Rows (transposed)

	Row 1	Row 2	Row 3	Row 4	Row 5
Serial Number	220008	2020348	20002	210317	200212
List Year	2022	2020	2020	2021	2020
Date Recorded	01/30/2023	09/13/2021	2020-02-10	2022-05-07	2021-09-03
Town	Andover	Ansonia	Ashford	Avon	Avon
Address	618 ROUTE 6	230 WAKELEE AVE	390 TURNPIKE RD	53 COTSWOLD WAY	5 CHESTNUT DRIVE
Assessed Value	139,020.0	150,500.0	253,000.0	329,730.0	130,400.0
Sale Amount	232,000.0	325,000.0	430,000.0	805,000.0	179,900.0
Sales Ratio	0.5992	0.4630	0.5883	0.4096	0.7248
Property Type	Residential	Commercial	Residential	Residential	Residential
Residential Type	Single Family	NaN	Single Family	Single Family	Condo
Non Use Code	NaN	NaN	NaN	NaN	NaN
Assessor Remarks	NaN	NaN	NaN	NaN	NaN
OPM Remarks	NaN	NaN	NaN	NaN	NaN
Location	POINT (-72.343629 41.728432)	NaN	NaN	POINT (-72.846366 41.781677)	NaN

2. **Handling Missing Data**:

Table 3.2: Missing Values in each Column

Serial Number	0
List Year	0
Date Recorded	2
Town	0
Address	51
Assessed Value	0
Sale Amount	0
Sales Ratio	0
Property Type	378726
Residential Type	394669
Non Use Code	748898
Assessor Remarks	888251
OPM remarks	1037575
Location	758031

A detailed check for missing values was performed to identify null values in the dataset. Initially, irrelevant columns with very large number of missing values, such as ”OPM remarks”, ” Assessor Remarks”, ”Loation” and ”Non Use Code,” were removed to streamline the dataset. For critical fields like ”Property Type” and ”Residential Type”, the cleaning is described in the Cleaning Section below. For ”Address” column, there were only a few missing values, so we decided to delete those rows.

3. **Grouping and Cleaning**: Property types were grouped into two main categories—”Residential” and ”Non-Residential”—to simplify the analysis. This was because the ”Non-Residential” property types will have null values in the column ”Residential Type” column due to obcious reasons. So choosing to delete those null values will result in deleting all ”Non - Residential” type data. So, we decided to partition the data set into two datasets: one with all houses being residential (df_residential), and the other being non-residential properties(df_non_residential). Additionally, the ”Address” column had mostly all unique values. To get more values of similar type, we decided to delete the initial numbers from the text in address, that is 618 Route Ave becomes Route Ave, which fascilitates in grouping data later on.

4. **Categorical Data Encoding**: Certain categorical variables, such as ”Property Type,” required transformation for modeling. Label Encoding was applied to convert these categorical variables into numerical values, facilitating their use in machine learning models.

5. **Data Standardization**: To ensure that numerical variables were on the same scale, standardization was applied. This involved scaling the features so that they had a mean of 0 and a standard deviation of 1. Standardization was crucial for models sensitive to the magnitude of input features, ensuring that all variables contributed equally to the analysis.

Outcome: The preprocessing phase resulted in a clean, consistent dataset, free of missing values and irrelevant data. Numerical and categorical features were standardized and encoded, ensuring compatibility with machine learning models and improving the quality of subsequent analysis.

3.2 Exploratory Data Analysis (EDA)

3.2.1 Objective

Gain a deeper understanding of the dataset’s structure, distributions, and relationships. This stage focuses on identifying trends, correlations, and outliers that may impact model performance or provide insights into the data.

3.2.2 Steps

1. **Statistical Summary**: Basic descriptive statistics were generated for numerical and categorical variables, including measures such as mean, median, mode, standard deviation, and quantiles. This helped identify general patterns and central tendencies in the data.

2. **Visual Analysis**: Various visualization tools were employed to explore data trends and detect outliers. Box plots were utilized to examine the spread of data, identifying potential outliers across numerical features. Histograms provided insights into the distribution of key variables, such as property values, highlighting skewness or uniformity. In addition to this, bar plots were generated to compare categorical variables and their corresponding numerical values, offering a clear view of trends within different categories. Density plots were used to visualize the smooth distribution of numerical features, especially to understand the spread of the sale amount and property values. Finally, pair plots were created to examine the relationships between multiple numerical features, allowing us to visualize correlations and identify any clustering or patterns that may exist between them.

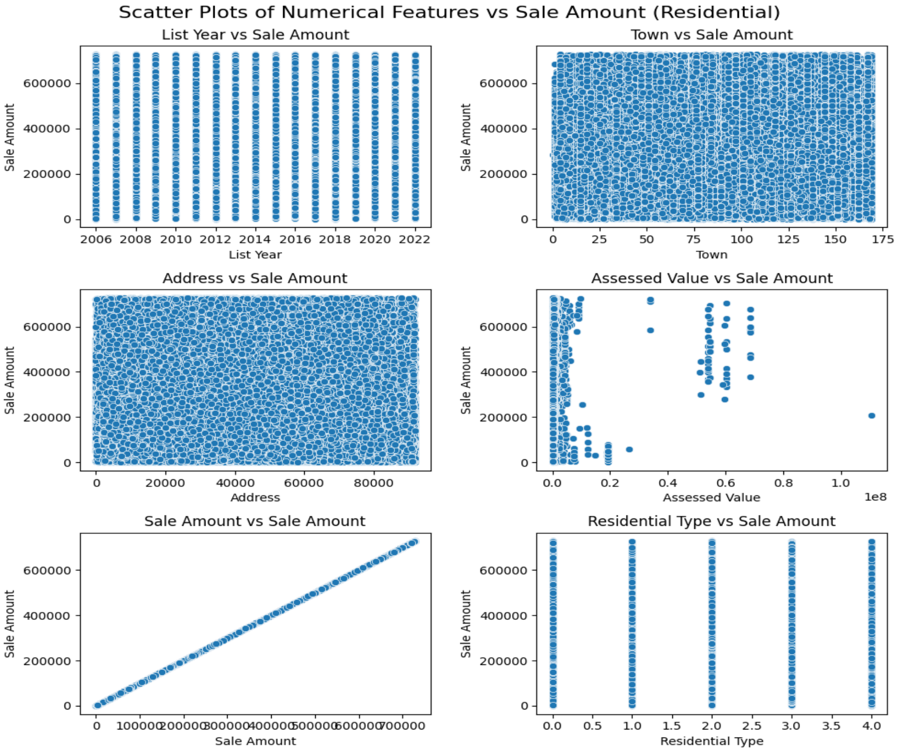


Figure 3.1: Scatter Plots of Numerical Features vs Sale Amount (Residential). Key Observations:
- List Year vs. Sale Amount: Slight upward trend, weak correlation.
- Town/Address vs. Sale Amount: No clear pattern.
- Assessed Value vs. Sale Amount: Positive correlation.
- Residential Type vs. Sale Amount: Distinct clusters suggest variations.

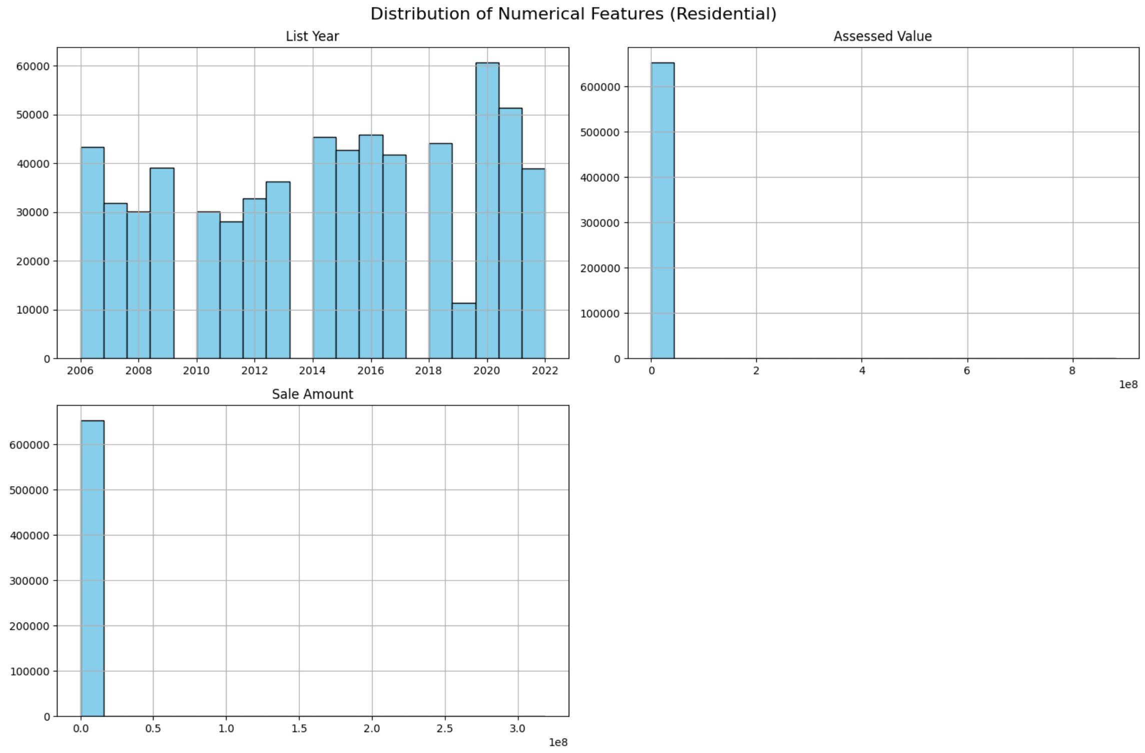


Figure 3.2: Distribution of Numerical Features (Residential). Key Observations:
- List Year: Uniform distribution.
- Assessed Value/Sale Amount: Right-skewed, most properties priced lower.

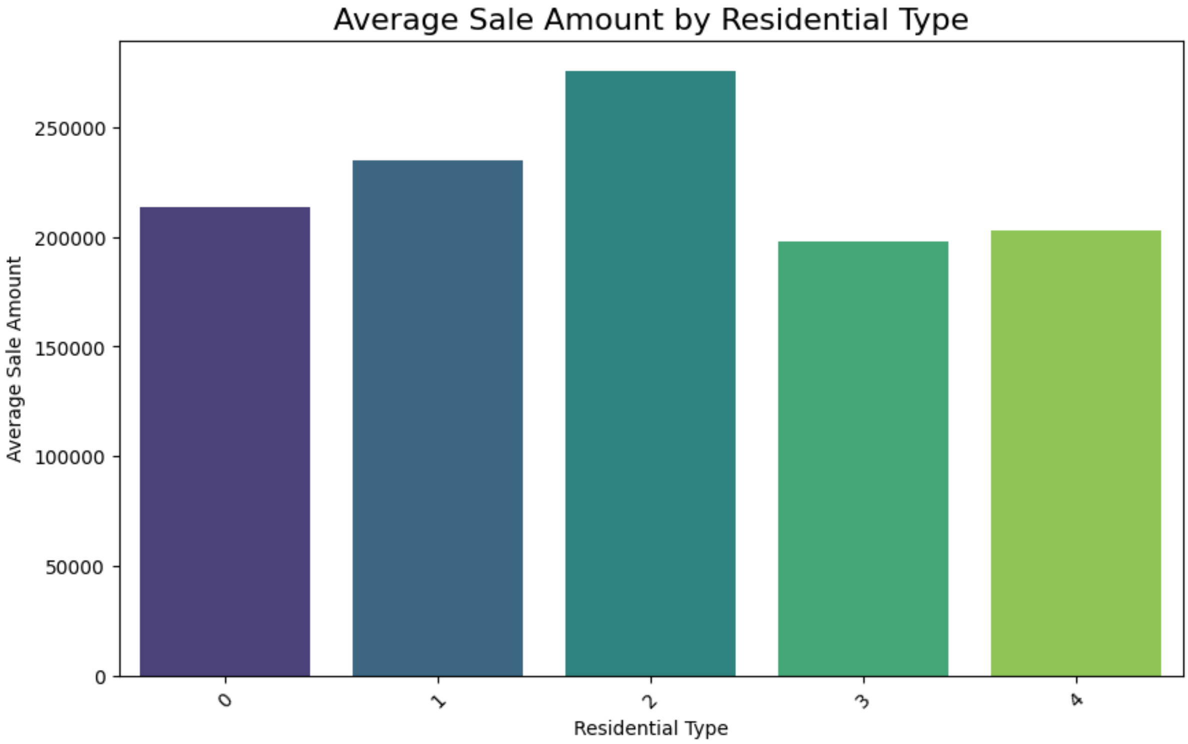


Figure 3.3: Average Sale Amount by Residential Type. Key Observations:
- Type 2: Highest average sale amount.
- Types 3/4: Lowest average sale amounts, Type 4 slightly lower.

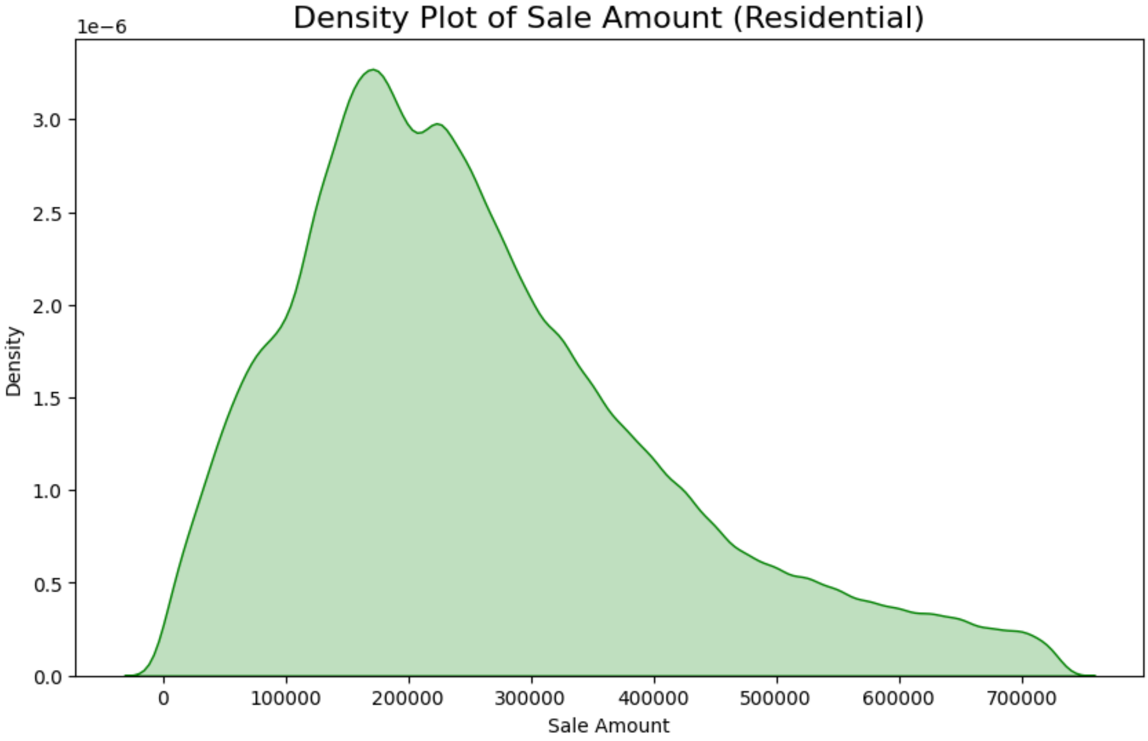


Figure 3.4: Density Plot of Sale Amount (Residential). Key Observation:
- Right-skewed distribution with a peak around 100,000 to 200,000.

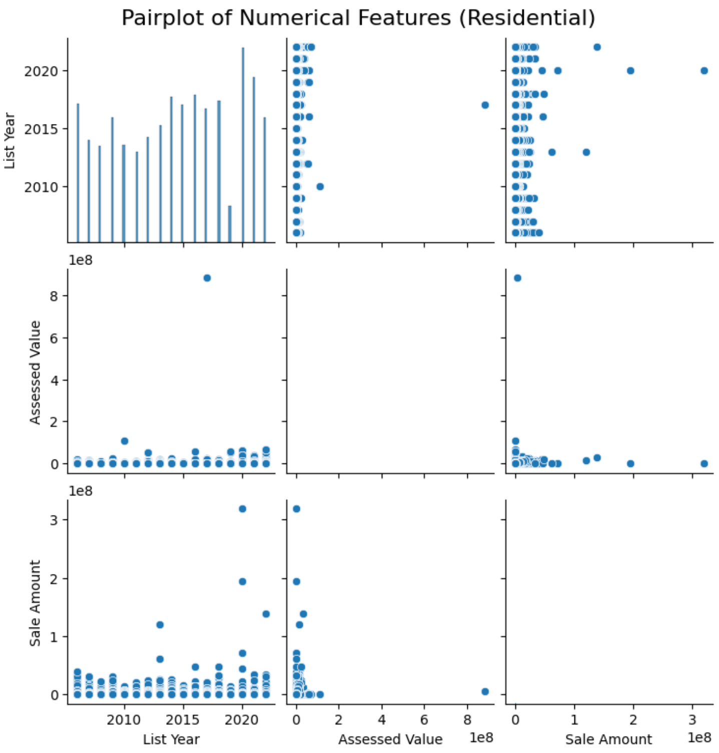


Figure 3.5: Pairplot of Numerical Features (Residential). Key Observations:
- Assessed Value vs. Sale Amount: Positive correlation.
- Other pairs: Weak/no significant relationships.

3. **Correlation Analysis** : To understand the relationships between numerical variables, correlation matrices were constructed. These matrices visually represented the degree of association between different features, helping identify which variables had the strongest relationships with the target variables.

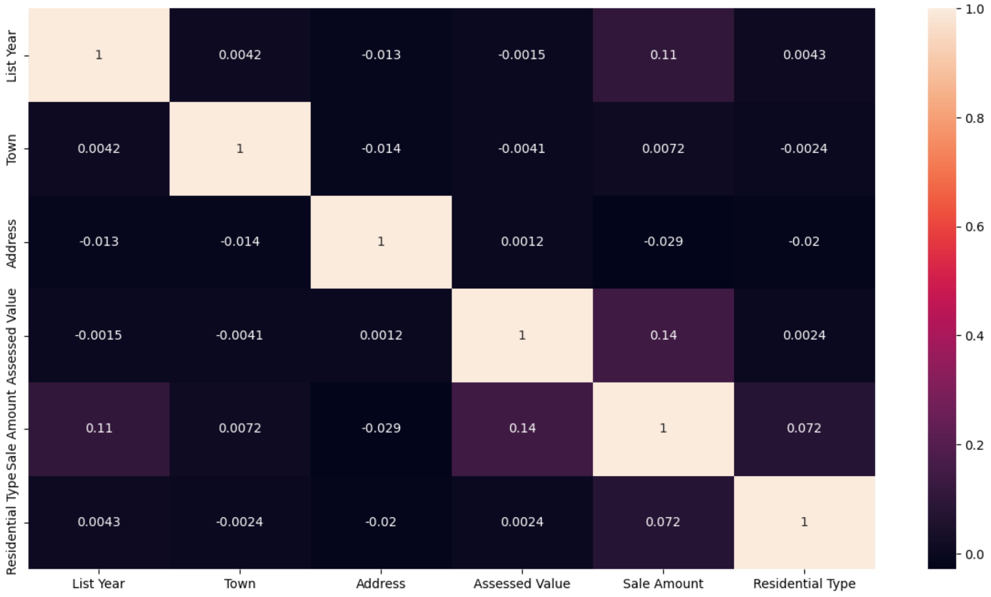


Figure 3.6: Correlation plot

4. **Outlier Detection and Handling**: Outliers were detected using the Interquartile Range (IQR) method. Data points falling significantly outside the typical range were flagged as potential anomalies. Outliers that had a high likelihood of skewing results were removed to maintain the validity of the data for predictive modeling.

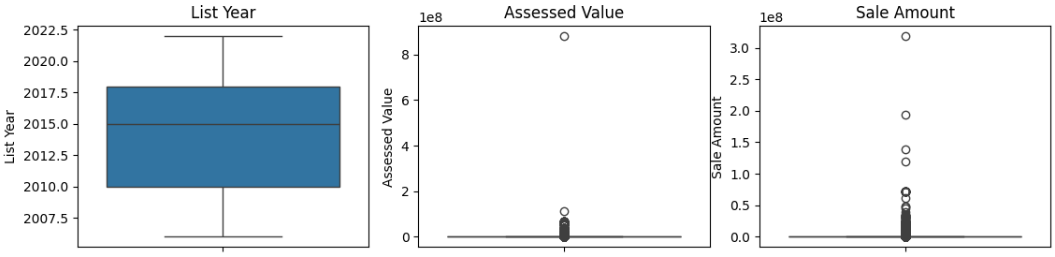


Figure 3.7: Box lot

Outcome: EDA provided a comprehensive overview of data behavior, trends, and relationships. The identification and handling of outliers ensured that the dataset was reliable and free from extreme anomalies that could distort analysis results. Key variables were identified for focus in the modeling phase.

3.3 Model Fitting and Performance

3.3.1 Feature Scaling

Before fitting any predictive model, the dataset underwent preprocessing to ensure that the features were standardized, as many machine learning models are sensitive to the scale of data. Below are the preprocessing steps:

1. **Splitting Features and Target Variable**: The dataset was split into X (features) and y (target variable). X consisted of all predictor variables, while y was the target variable—Sale Amount.

2. **Feature Scaling with StandardScaler**: A StandardScaler was applied to X to standardize the features, transforming them to have a mean of 0 and a standard deviation of 1. This scaling ensures that all features contribute equally to the model training, preventing dominance by any feature due to differing scales.

3.3.2 Data Splitting

To validate the performance of the models, the data was split into training and testing sets using two configurations:

- 80-20 Split: 80% training and 20% testing.

- 70-30 Split: 70% training and 30% testing.

The purpose of using two splits was to observe the consistency and robustness of each model's performance under different training/testing distributions.

3.3.3 Model Descriptions and Performance

Linear Regression

Purpose: Linear Regression was employed as a baseline model to capture a simple linear relationship between the predictors and the target variable.

Performance:

- 80-20 Split: MSE: 21,737,995,103.50
  R-squared: 0.0381

- 70-30 Split: MSE: 21,622,571,538.02
  R-squared: 0.0405

Analysis: The low R-squared values indicated that Linear Regression could not effectively capture the data's complexity.

Gradient Boosting Regressor

Purpose: Gradient Boosting Regressor, a powerful ensemble technique, was utilized to improve predictive accuracy by correcting errors iteratively using sequential decision trees.

Performance:

- 80-20 Split: MSE: 6,080,061,233.61
  R-squared: 0.7310
- 70-30 Split: MSE: 6,078,494,772.18
  R-squared: 0.7303

Analysis: This model outperformed Linear Regression significantly, showcasing its strength in capturing complex patterns.

XGBoost Regressor

Purpose: XGBoost, known for its efficiency and accuracy, was chosen for its optimized gradient boosting technique that enhances prediction using a series of decision trees.

Performance:

- 80-20 Split: MSE: 5,282,967,511.96
  R-squared: 0.7662
- 70-30 Split: MSE: 5,304,788,282.50
  R-squared: 0.7646

Analysis: XGBoost outperformed Gradient Boosting, indicating a better fit to the data's intricacies.

Random Forest Regressor

Purpose: Random Forest, another ensemble technique, was utilized to minimize overfitting by averaging multiple decision trees.

Performance:

- 80-20 Split: MSE: 5,231,274,192.41
  R-squared: 0.7685
- 70-30 Split: MSE: 5,306,053,404.58
  R-squared: 0.7646

3.3.4 Comparison of Model Performance

The table below provides a summary of the models' performance metrics for both the 80-20 and 70-30 splits:

3.4 Comparison of Model Performance

In this section, we compare the performance of the different models based on their evaluation metrics. We utilized both the 80-20 split and the 70-30 split to assess how well the models performed with different training and testing distributions. The models compared include Linear Regression, Gradient Boosting, XGBoost, and Random Forest.

The performance metrics used for comparison are:

- **Mean Squared Error (MSE):** A lower MSE indicates a better fit, as it reflects the average squared difference between the predicted and actual values.

- **R-squared:** A higher R-squared value indicates a better fit, showing the proportion of the variance in the dependent variable that is predictable from the independent variables. The R-squared value is not too high because of very low correlation between the columns of the dataset.

The table below presents a summary of the performance of these models based on the MSE and R-squared for both the 80-20 and 70-30 splits:

Model	80-20 Split (MSE)	80-20 Split (R-squared)	70-30 Split (MSE)	70-30 Split (R-squared)
Linear Regression	21,737,995,103.50	0.0381	21,622,571,538.02	0.0405
Gradient Boosting	6,080,061,233.61	0.7310	6,078,494,772.18	0.7303
XGBoost	5,282,967,511.96	0.7662	5,304,788,282.50	0.7646
Random Forest	5,231,274,192.41	0.7685	5,306,053,404.58	0.7646

Table 3.3: Comparison of Model Performance on Residential Data

Table 3.4: Comparison of Model Performance on Non-Residential Data on 80-20 Split

Model	MSE	R-squared
Linear Regression	91867851349.7632	0.015474283286558865
GradientBoostRegressor	43370078032.20582	0.5352132815645159
XGBRegressor	40401476250.68998	0.5670270744598822

Table 3.5: Performance Comparison of Models

As observed from the table:

- **Linear Regression** shows poor performance with a low R-squared value, indicating it could not capture the complexity of the data.
- **Gradient Boosting** and **XGBoost** demonstrate strong performance, with R-squared values over 0.73 for both splits, indicating good predictive power.
- **Random Forest** performed the best with an R-squared value of 0.7685 for the 80-20 split, suggesting it captured the most variance in the target variable.

The ensemble models—Gradient Boosting, XGBoost, and Random Forest—consistently outperformed Linear Regression, highlighting their ability to handle complex patterns in the data.

Chapter 4. Results and Discussion

4.1 Key Findings

- **Property Type Impact:** The type of property significantly influenced sale prices. Single-family homes generally showed higher median sale values compared to multi-family and commercial properties.
- **Year Built and Sale Amount:** Properties built more recently tend to have higher sale prices, indicating that newer constructions hold greater market value.
- **Geographical Variation:** Sale prices varied significantly by location. Urban regions showed higher average sale amounts, while rural areas had lower averages.

4.2 Regression Analysis

The regression models confirmed that several factors were statistically significant in predicting sale prices:

- Assessed Value was a strong predictor, explaining a large portion of the variance in sale prices.
- Square Footage also showed a positive correlation with sale prices, indicating that larger properties tend to sell for more.
- Year of Sale revealed slight inflationary trends in the market, with recent years showing a steady increase in property values.

4.3 Model Performance and Limitations

The random forest model achieved a reasonable accuracy in predicting property sale prices, with an R^2 value indicating that a significant portion of the variability was captured by the model. However, certain limitations were observed:

- Outliers impacted model performance.
- The model was less effective in predicting high-end luxury properties.
- The R^2 value was overall low because of very low correlation between the columns, which is due to the chosen dataset.

4.4 Discussion of Results

The findings suggest that traditional factors like property type, assessed value, and square footage remain strong indicators of market value. However, regional variations indicate that location-specific factors are equally crucial in understanding real estate dynamics. The presence of outliers and specific anomalies highlights the complexity of the real estate market and the need for localized models or more granular data in some cases.

Chapter 5. Conclusion

The study aimed to analyze real estate transactions to identify key factors influencing property sale prices. Through comprehensive data preprocessing, exploratory analysis, and regression modeling, several insights were gathered.

5.1 Future Work

- Utilizing more sophisticated machine learning models.
- Incorporating additional features like property renovation status, amenities, or proximity to key facilities.
- More granular regional analysis focusing on micro-locations within cities.

In conclusion, this analysis provides a solid foundation for understanding real estate market trends and highlights the importance of robust data handling and model selection in making accurate predictions.

Chapter 6. References

- **Dataset**: https://data.ct.gov/api/views/5mzw-sjtu/rows.csv?accessType=DOWNLOAD
- https://www.data.gov.in/catalog/housing-price-index-india
- https://www.housingpriceindex.in/