Housing Prediction: An Exploratory Data Analysis

Author Name: Naga Avinash Matta

Mail-id: nm55284n@pace.edu

school of Seidenberg, pace university

Course : Practical Data science

Program Name: Ms in data science

Date: 12-11-2024

Agenda

- Project Overview
- Data
- Exploratory Data Analysis (EDA)
- Modeling Methods
- Findings
- Recommendations & Technical Next Steps
- Q&A

Executive summary slide:

Business Problem:

In the real estate market, accurately predicting house prices is essential for buyers, sellers, real estate agents, and investors. Pricing too high or too low can lead to significant financial losses or missed opportunities.

The objective of this project is to build a predictive model that estimates house prices based on various attributes, such as location, size, number of bedrooms and bathrooms, and other relevant features. By accurately forecasting house prices, stakeholders can make informed decisions, optimize investments, set competitive prices, and enhance customer satisfaction.

Solution summary:

- Conducted EDA to identify influential features.
- Built a predictive model focusing on major factors impacting housing prices.
- Identified key insights into feature-price relationships (e.g., area, bedrooms).
- Developed actionable recommendations based on insights, supporting business strategies.

Project plan Recap slide:

Overview of Phases:

Each topic represents a major project phase: *Data Collection*, *EDA*, *Model Development*, *Findings & Interpretation*, and *Recommendations*. This gives a quick overview of the workflow.

Duration and Timeline:

The chart's x-axis displays hypothetical project weeks, making it clear how long each phase took or is expected to take. It provides transparency on time allocation, helping stakeholders understand the project timeline.

Purpose of the Slide:

This slide summarizes the project's progress, helping stakeholders quickly identify which phases are completed and what remains. This recap aids in managing expectations and setting realistic next steps for the remaining tasks.

Project plan Recap slide:

DEVIVERABLE	DETAILS	DUE DATES	STATUS
Data & EDA	Identify the trends and patterns	11-05-2024	Completed
Methods, Findings and recommendations	Finding out the methods	11-12-2024	Pending
Final Presentation	Final Completion Deck	12-03-2024	Pending


DATA

Data: Housing Dataset

Data Source: Housing.csv

The housing dataset was collected from a simulated real estate database. It includes various property attributes and sale prices, simulating typical housing data in the market.

Sample Size: 545 records

The dataset contains few observations and few variables. Each row represents a unique property listing, while columns represent characteristics (features) like area, bedrooms, bathrooms, etc.

Time Period:

Since this is a simulated dataset for analysis, it does not cover a specific time period. However, it is structured to resemble real-world housing data trends.

Inclusions and Exclusions:

Included:

Key property features relevant to pricing, such as size, rooms, amenities (e.g., guest room, basement, air conditioning), furnishing status, and access to main roads.

Excluded:

Data on property age, neighborhood quality, and local market conditions were not included, which may limit the depth of market insights but keeps the dataset focused on generalizable property attributes.

Data: Housing Dataset – Classifications & Assumptions

Classification:

Categorical variables, like "furnishing status" and "main road access," are encoded as binary or categorical values to ensure consistency in analysis.

Assumptions:

Property Licensing:

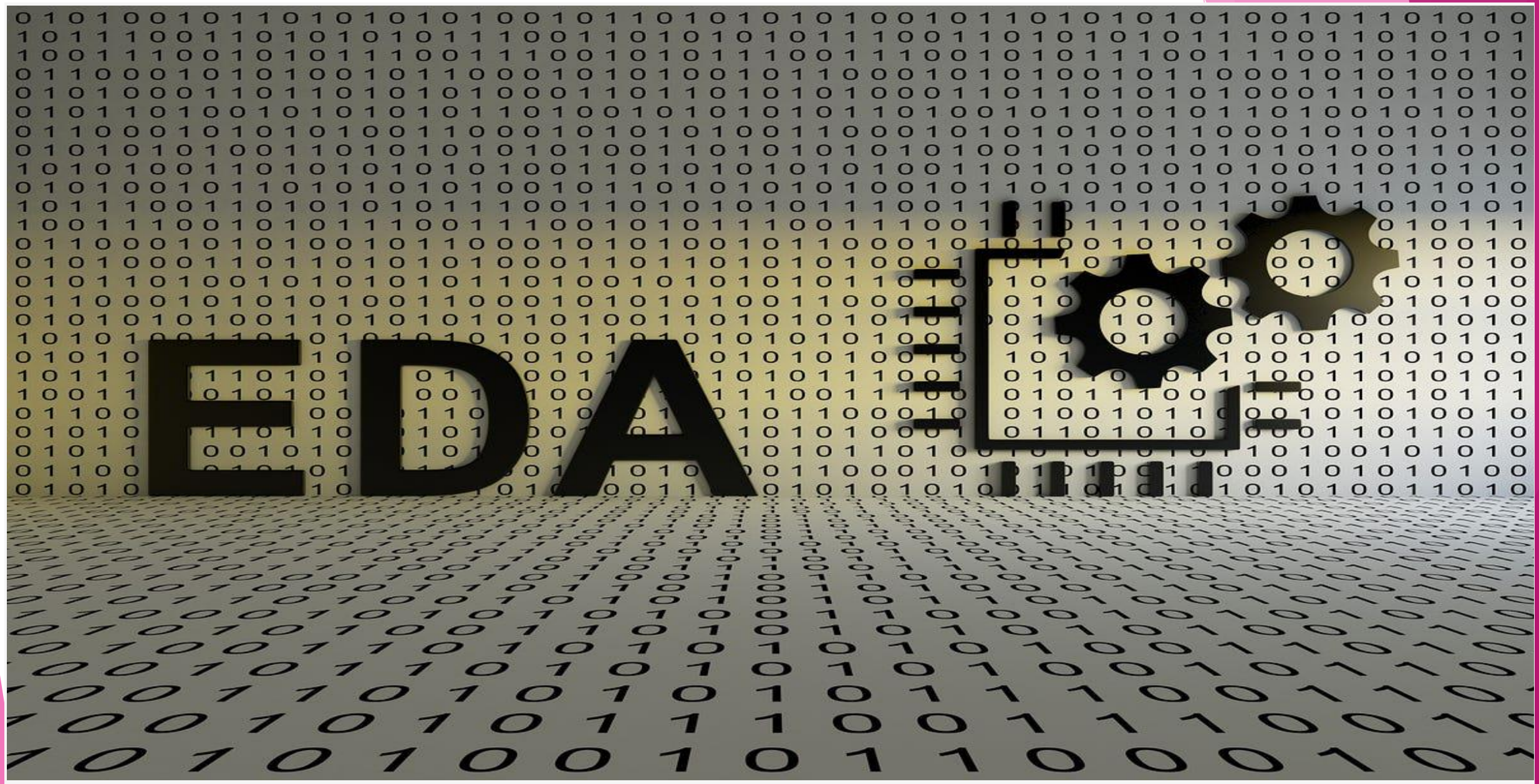
We assume that all properties meet local zoning and licensing requirements, as external market data typically verifies property validity and legality.

Housing Demand and market conditions:

It’s assumed that market trends influencing housing prices, such as seasonal demand fluctuations, remain consistent across the sample.

Uniform Data Quality:

Assumes the dataset accurately captures real estate attributes relevant to price. Any missing data has been treated or imputed to maintain completeness for modeling.


EDA

EDA: Housing Dataset Visualizations

The visualizations builds on the relationship between housing features and price, establishing a coherent narrative for modeling house prices based on size, amenities, and structure. This setup aids in identifying and selecting features critical to accurate predictions, tying back to the business goal of understanding and forecasting housing prices effectively.

Visualization 1: Distribution of House Prices

Visualization 2: House Price vs. Area (sq ft)

Visualization 3: House Prices by Number of Bedrooms

Visualization 4: Correlation Heatmap of Housing Features

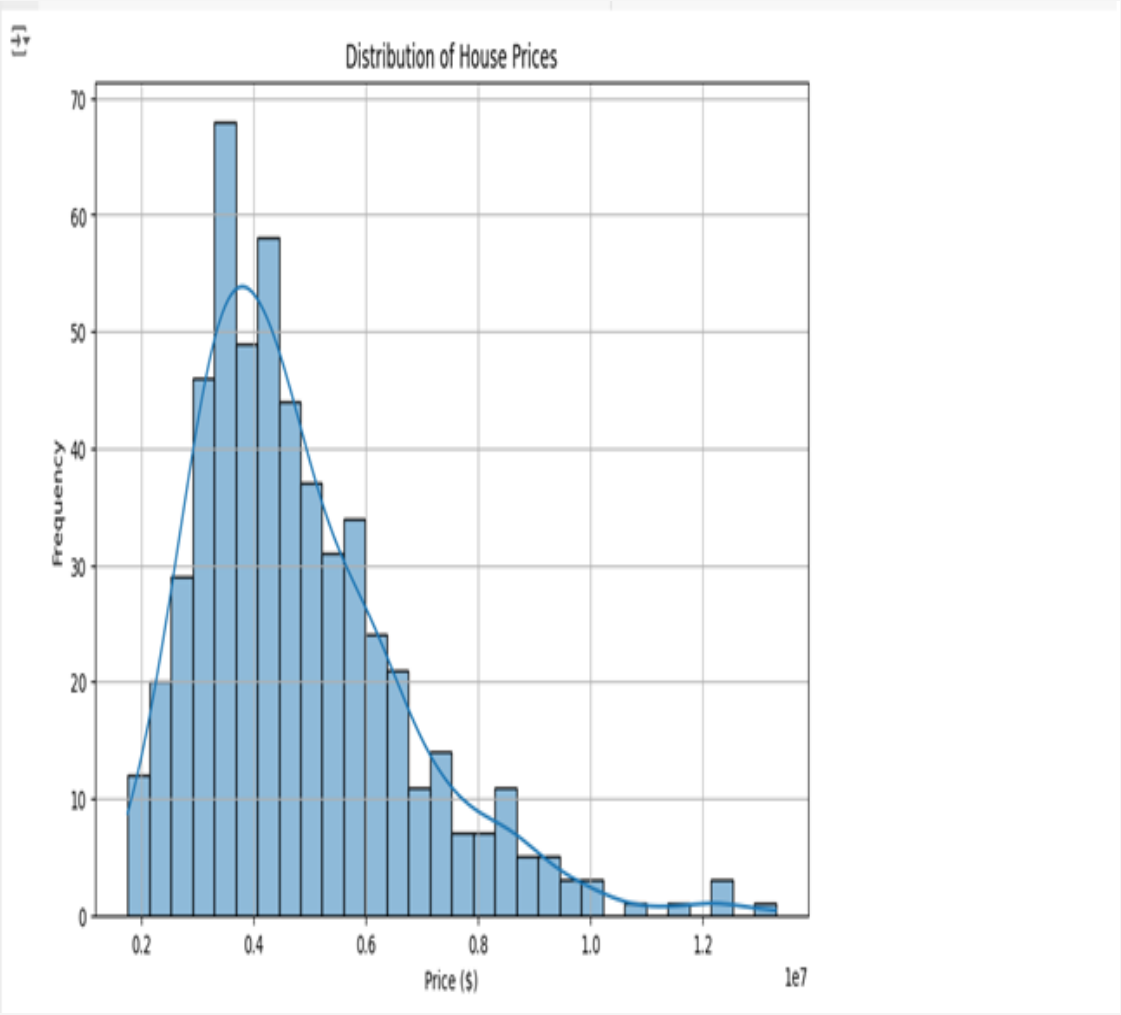
Visualization 5: Average House Price by Number of Stories

EDA: Visualization 1

Title: Distribution of House Prices in the Dataset

Summary:

This histogram displays the distribution of house prices across the dataset. The visualization helps us understand the range and central tendency of house prices, which is essential in predicting house values. A concentration in certain price ranges may suggest areas of the market where demand is highest or where our model could face challenges in prediction due to limited price variance.

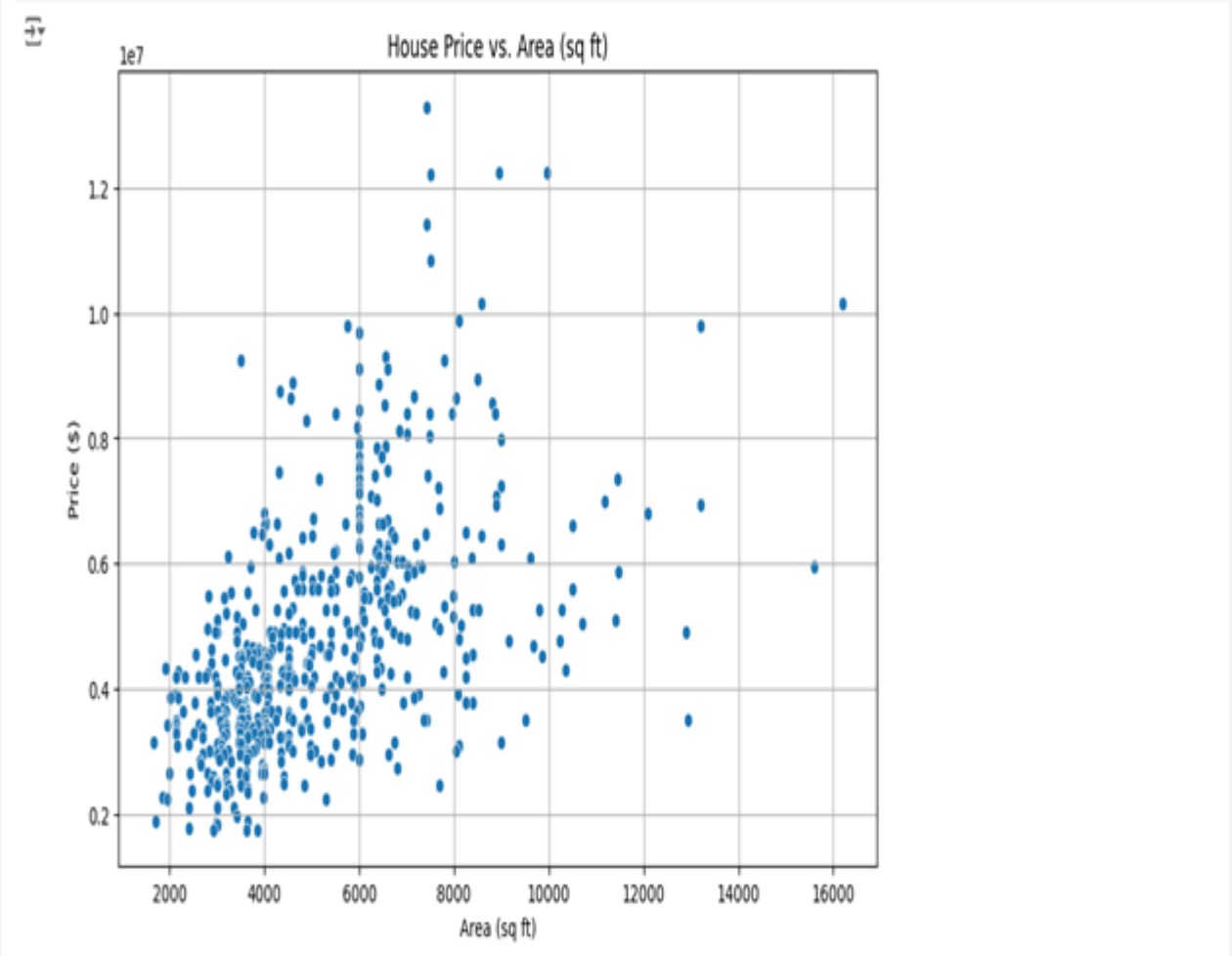


EDA: Visualization-2

Title: Relationship Between House Price and Area

Summary:

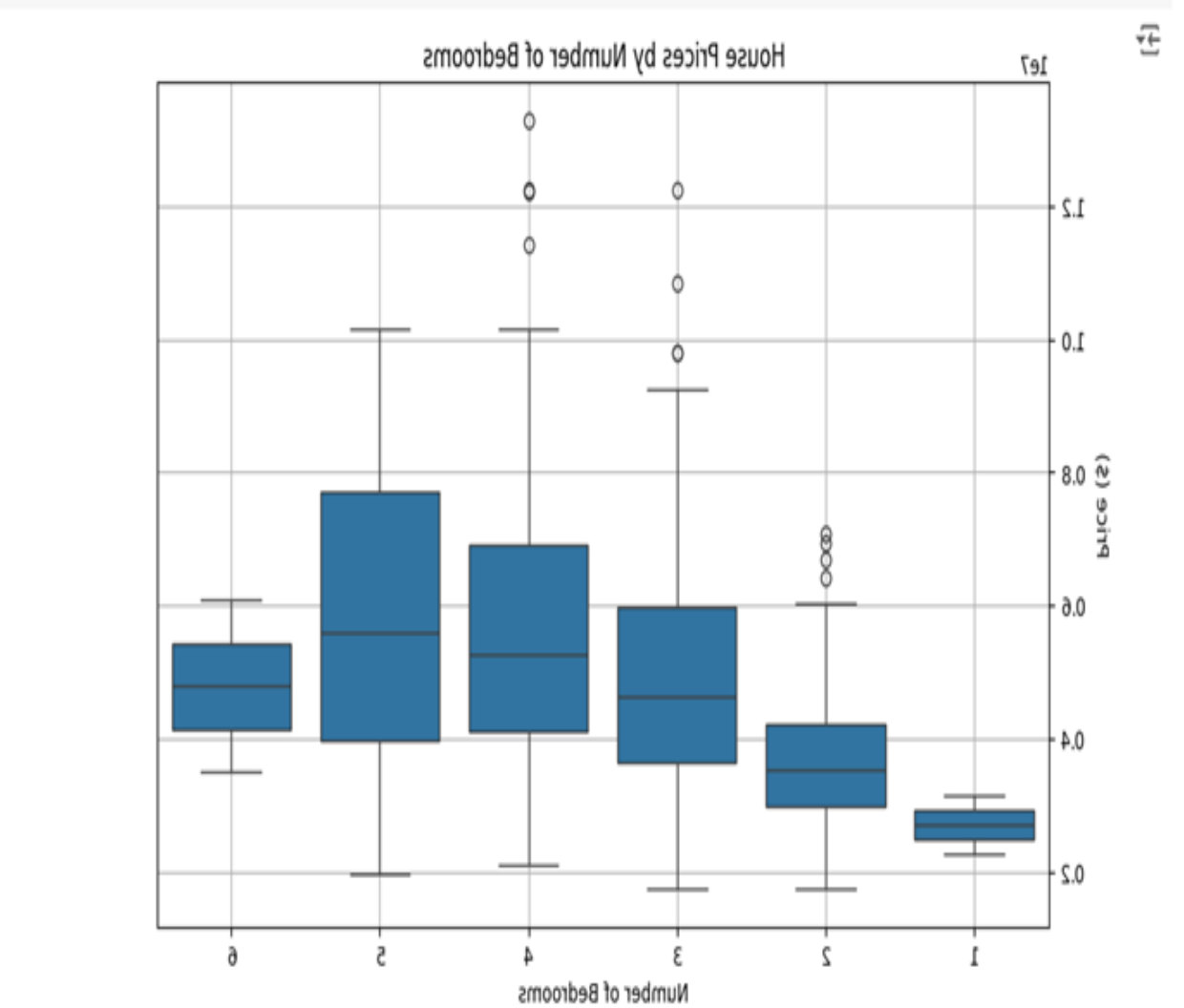
This scatter plot illustrates a positive relationship between the area of a house and its price. Larger homes tend to have higher prices, reinforcing that square footage is an essential predictor. This relationship suggests that increasing area may drive up property values, which is key to predicting house prices.



EDA: Visualization 3

Title: House Prices by Number of Bedrooms

Summary: This box plot shows the distribution of house prices based on the number of bedrooms. Houses with more bedrooms generally command higher prices. This insight underlines the importance of bedrooms as a feature in the model since families or buyers seeking larger homes may pay a premium.

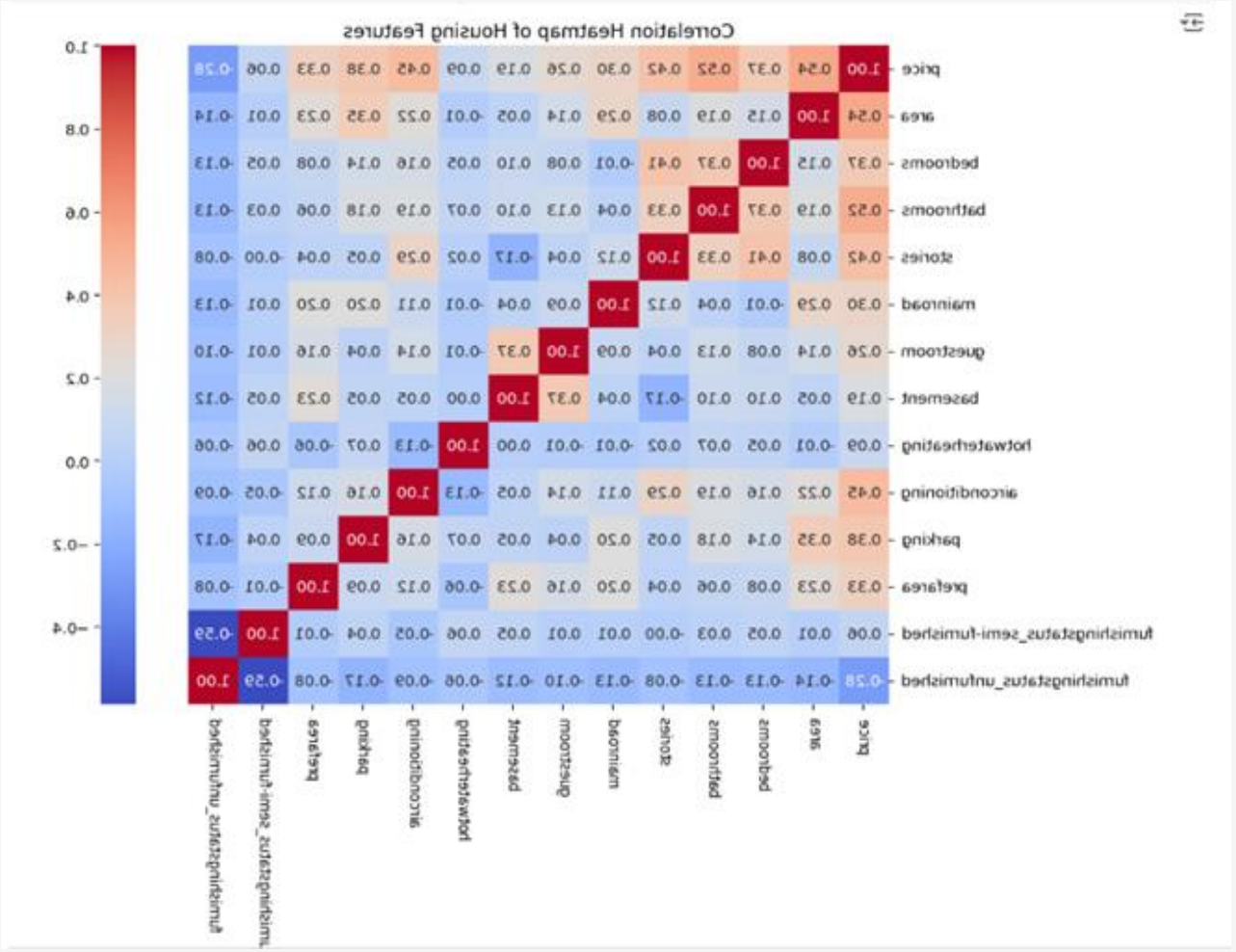


EDA: Visualization 4

Title: Correlation Heatmap of Key Housing Features

Summary:

This heatmap visualizes the correlation between features such as area, number of bedrooms, air conditioning, and price. Strong positive correlations between price and factors like area, number of bedrooms, and stories highlight these as significant predictors. This heatmap aids in feature selection, helping focus on high-impact variables for modeling.

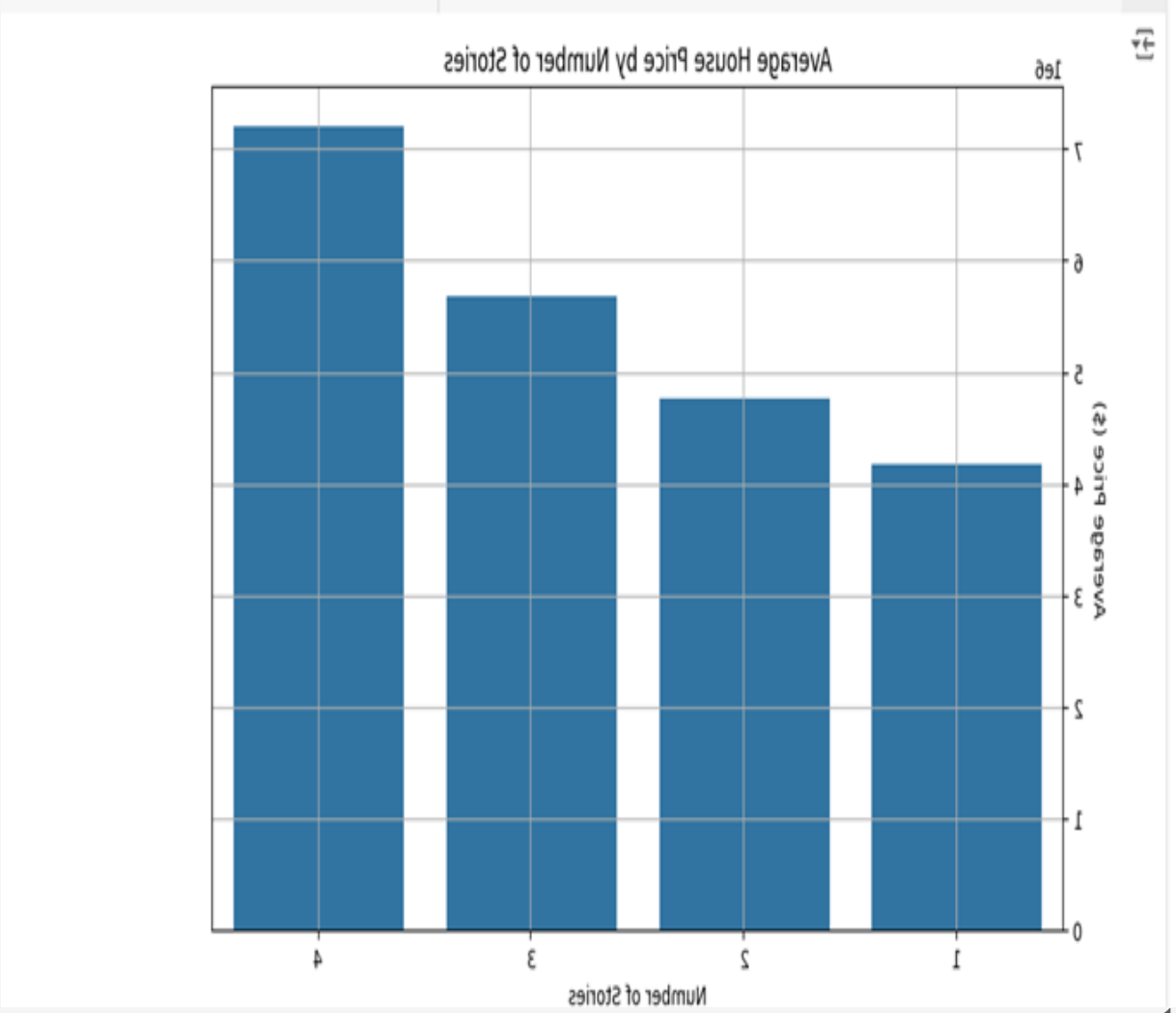


EDA: Visualization 5

Title: Average House Price by Number of Stories

Summary:

This bar chart shows the average price for houses based on the number of stories. Multi-story houses tend to have higher average prices, suggesting that additional stories are valued in this market. Including the number of stories in the model can capture the added value of multi-story properties, refining predictions.



Modeling methods

Modeling Methods:

1. **Outcome Variable:**

Outcome Variable (Target):

The outcome variable is the specific value or attribute we are trying to predict, such as housing prices in a real estate project.

Our target variable, **price**, represents the predicted housing price. The goal is to accurately forecast housing prices based on various features of each property, supporting insights into factors that influence house valuations.
All selected features connect back to this “so what” variable, with the goal of guiding decision-making related to pricing strategies, market assessment, and investment opportunities.

In Summary, the outcome variable provides direction for the project, shaping the choice of features and models and ensuring that all analysis remains centered on the primary business objective.

Modeling Methods: Features

2. Features:

Feature Selection:

The features used here were chosen based on their relevance to housing price predictions. Each feature potentially contributes to price prediction, and our code excludes **price** from the feature set X as it's the target. These features can be grouped as follows:

Physical Attributes:

Includes features like area, number of bedrooms, and bathrooms, as these are directly associated with a property's structure and usability.

Amenities and Extras:

Binary variables such as **mainroad**, **guestroom**, **basement**, **hotwaterheating**, **airconditioning**, **parking**, and **prefarea** indicate desirable amenities and access points, which often increase property value.

Furnishing Status:

Encoded into dummy variables, **furnishingstatus** reflects whether a property is fully, semi, or unfurnished. Homes with different furnishing levels appeal to varying buyer needs and budgets.

Modeling Methods: Model Type and Rationale

3. Model Type and Rationale:

Model Choice:

We opted for **Linear Regression** due to its interpretability and ease of application for predicting a continuous variable like house price. Linear regression is useful in this context as it provides an estimation of how each feature impacts the target variable (price) linearly, making it easy to interpret for business decisions.

In simple terms, **linear regression** is a model that looks at how each characteristic of a home (e.g., number of bedrooms, area, amenities) affects its price. Think of it as a rule where each factor adds or subtracts value from the price, similar to a recipe where each ingredient contributes to the final taste. For example, adding square footage generally raises the price, while the presence of certain amenities like air conditioning or parking may also increase value. This straightforward relationship allows us to see what drives prices up or down, guiding property-related decisions.

Modeling Methods: Model Training and Predictions

Model Training and Prediction Code :

After defining the features X and target variable Y, we split the data into training and testing sets. The model is then trained on x_train and y_train, learning the relationship between features and price. Once trained, we use x_test data to make predictions and evaluate model performance. This split ensures our model can generalize well to new data, important for accuracy in real-world applications.

Feature Coefficients Interpretation:

Printing out the model's coefficients provides insights into each feature's impact on price. Positive coefficients indicate features that tend to increase price, while negative coefficients show those that decrease it. For example, area may have a high positive coefficient, suggesting larger homes correlate strongly with higher prices.


FINDINGS

Findings: “Key Drivers of Housing Prices”

Main Finding:

Present a bar chart or table of the model**’s** coefficients to identify the most influential factors on housing prices.

Interpretation:

Features like **area (sq ft)**, **number of stories**, and **presence of amenities (e.g., air conditioning, guest rooms)** may display significant positive coefficients, meaning they contribute meaningfully to higher property prices.

Business Insight:

These results suggest that investment in increasing property size or enhancing amenities could add substantial value, guiding development and renovation decisions.

Findings: “Impact of Square Footage on Housing Price”

Main Finding:

Display a scatter plot showing the positive correlation between **area** (sq ft) and price.

Interpretation:

Square footage consistently increases housing prices, indicating that buyers prioritize larger living spaces.

Business Insight:

Focus on promoting larger properties or additions to existing homes as a reliable way to enhance market value. Additionally, this could suggest adjusting pricing models to account more directly for square footage.

Findings: “Role of Amenities on House Value”

Main Finding:

Coefficients for features like **air conditioning, guest room availability, and basement** reveal how specific amenities impact price.

Interpretation:

Significant coefficients for these amenities indicate that houses with these features are valued higher. For example, air conditioning could increase comfort and attractiveness in warmer regions, adding to property appeal.

Business Insight:

Stakeholders could focus on upgrading properties with popular amenities to capture higher valuations. Marketing strategies can also highlight these features to attract a premium buyer segment.

Findings: “Unexpected Findings: Features with Low Impact”

Main Finding:

Some features, like **proximity to the main road** or **furnishing status**, may show little to no statistical significance in affecting prices.

Interpretation:

These results are counterintuitive if previous assumptions suggested these features would impact valuation. It may indicate that in this market, buyers prioritize other features more.

Business Insight:

This slide could encourage further investigation. It might involve gathering additional data on location or re-evaluating assumptions about consumer preferences, which could reveal shifts in market demands or dataset limitations.

Recommendations

And

Next Steps

Recommendations & Technical Next Steps

Key Recommendations:

Invest in Expanding Square Footage:

Finding: Square footage has a significant positive correlation with housing price, meaning larger homes are consistently valued higher.
Actionable Recommendation: For properties under development or renovation, prioritize expansion projects that increase living space, particularly in areas where buyers value larger homes. Marketing campaigns should emphasize square footage as a major selling point.

Enhance Key Amenities:

Finding: Amenities such as air conditioning, guest rooms, and basements add notable value to a property's price.
Actionable Recommendation: Focus on upgrading or adding popular amenities to properties where feasible. These upgrades should also be featured in property listings and marketing materials to attract buyers willing to pay a premium for these conveniences.

Recommendations & Technical Next Steps

Technical Next Steps:
Build a More Advanced Model:

Rationale: A linear regression model offers basic insights, but more advanced techniques like Random Forest or Gradient Boosting could capture non-linear relationships and interactions between features, potentially enhancing prediction accuracy.
Implementation: Evaluate additional models to compare against the current linear regression approach and assess if they yield better predictions. This could involve hyperparameter tuning and cross-validation to optimize performance.

Expand Data Collection:

Rationale: Additional data, such as neighborhood crime rates, proximity to schools, or environmental factors, could provide more context and improve the model's prediction capability by including factors buyers might value.
Implementation: Collaborate with data collection teams to incorporate external datasets that expand the current feature set. Future analyses can investigate how these factors impact housing prices, addressing business questions about emerging market trends.


APPENDIX

Appendix:

Project Code Repository:
Git hub Link:

Technical Model Overview: Linear Regression
Model Type and Choice:

Linear regression was chosen for its interpretability and because it directly connects features (like area, number of bedrooms, etc.) to the outcome variable (house price).

Model Details:

- The model was fit on a training dataset split from the original data, with an 80-20 training-to-test ratio.
- The model coefficients represent the effect each feature has on the price. For instance, the area coefficient shows how each additional square foot impacts the price prediction.

Interpretation of Coefficients:

Coefficients can provide insights, such as whether area is a strong predictor of price. If statistically significant, it means the relationship is meaningful in predicting the price.

Appendix:

Supplemental Data Visualizations:

Additional charts, like:

- Scatter plots to show relationships not initially highlighted.
- Distribution plots for variables that may influence the housing market in subtle ways.
- A bar plot of average price by neighborhood or other categorical data.

THANK YOU!