EDA AND INITIAL CLEANING OF HOUSE DATASET

PRESENTED BY: SONAM MEHTA

SUMMARY STATISTICS

- **Overview of Raw Data**
- **Data Cleaning**
- **Data Type Adjustment**
- **Observations**

OVERVIEW OF RAW DATA

- The dataset provided contains information about houses, including features such as sold price, location (zip code), dimensions, and various amenities. The goal is to preprocess this data to prepare it for modeling.
- After reading this raw dataset noticed it has 5000 observations and 16 features.
- After Summary statistics were generated using the describe() method to understand the distribution of numerical features.
- Dataset was checked for missing values using is null().sum(). Notably: Columns like kitchen features, garage, and others contained missing values.
- The dataset was examined for duplicate entries using duplicated().sum(). It was found that there were no duplicate rows.

DATA CLEANING STEPS

- **Handling Missing Values**

  - Missing values in the dataset were handled as follows: For categorical columns (e.g., fireplaces), missing entries were replaced with "Unknown".

  - For numerical columns, rows with missing values in critical fields like sold price, bedrooms, or bathrooms were dropped to maintain data integrity.

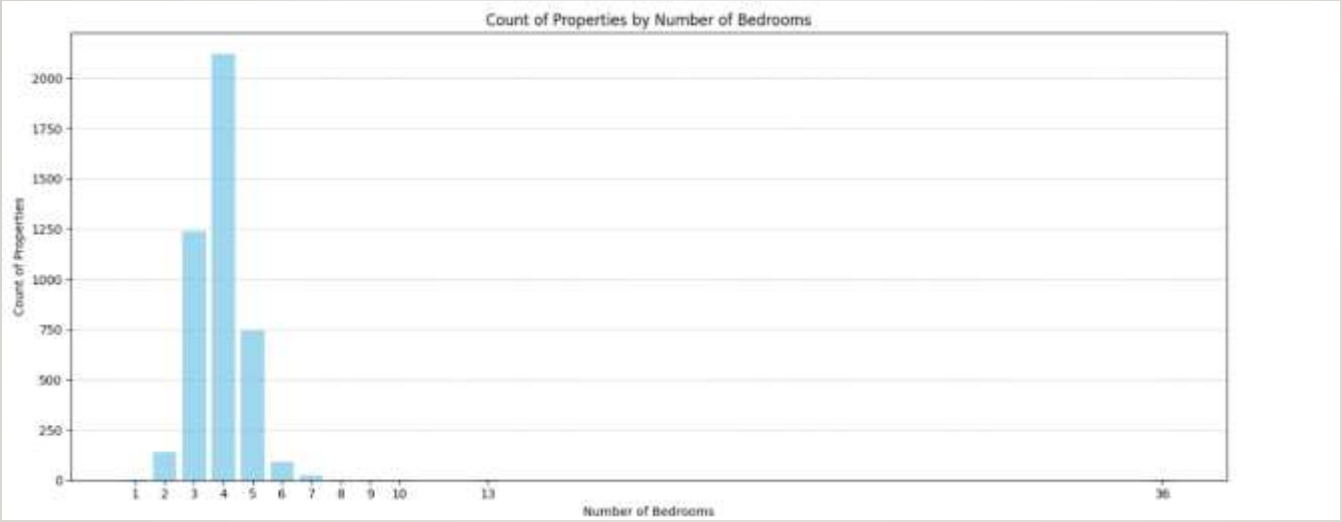
DATA TYPE ADJUSTMENT

- The cleaned dataset now contains 4370 entries with the following key characteristics: All missing values have been addressed.
- Certain features were of particular dataset. Updated them to int data type for efficiency for further requests.

- **Outlier Detection**
  - Outliers in the sold price and other numerical features were identified using scatter plots and box plots. Extreme outliers can be evaluated and removed if deemed necessary based on domain knowledge.

OBSERVATIONS

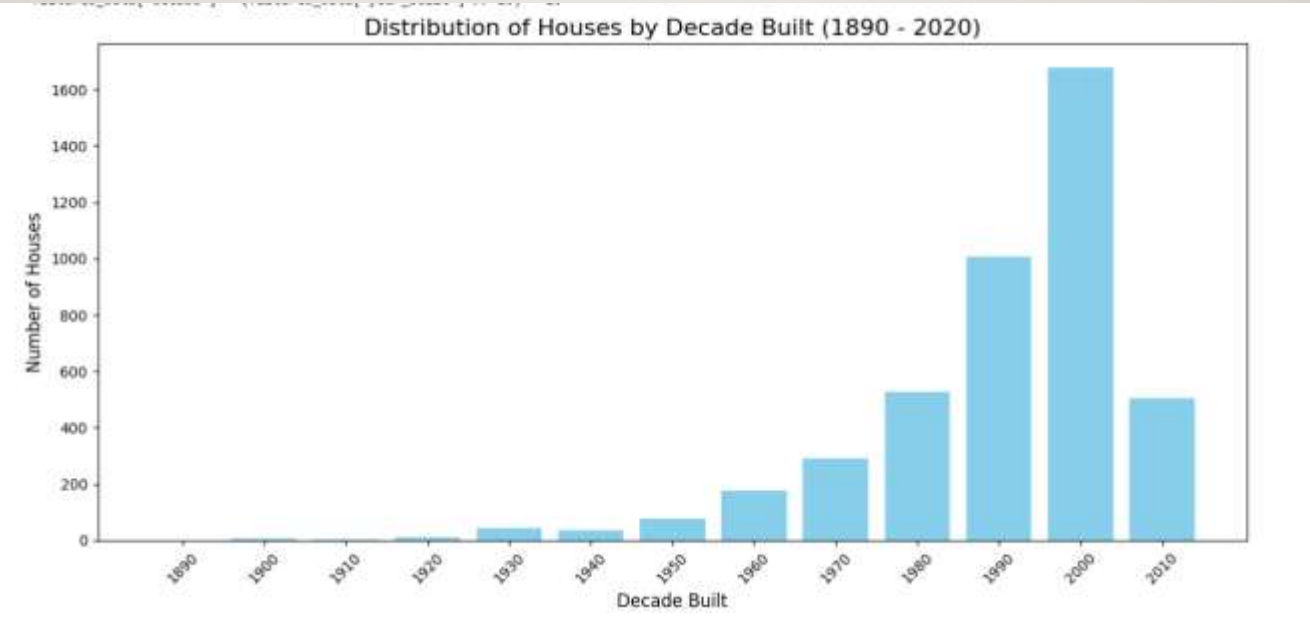
- Count of Properties by Number of Bedrooms.



- For Example: Properties with 4 bedrooms are the most common, with a count of 2,388.
- The property count decreases significantly for homes with more than 5 bedrooms.
- Outliers include properties with unusual bedroom counts, such as 36 bedrooms (3 properties) and 19 or 18 bedrooms (1 property each).

OBSERVATIONS

- Bar chart estimating the distribution of houses built by decade, providing insights into property age and investment potential. The graph illustrates the pattern of house construction over decades, helping identify whether properties are older or newer for informed decision-making.



OBSERVATIONS

- Scatter plot chart showing the maximum sold price grouped by zip code and number of bedrooms, enabling analysis of areas with the most properties and their corresponding bedroom counts relative to sold prices.


THANK
YOU