PROJECT PROGRESS REPORT

PROJECT PROGESS OVERVIEW:

Aim of the Project:
The project is focused on building a system to predict house prices and analyzing various factors that impact property prices. This involves exploratory data analysis (EDA) to understand patterns in housing data and identifying features that significantly influence prices.

Completed Tasks :

1.Data Loading and Basic Exploration

- Successfully imported necessary libraries such as Pandas, NumPy, Matplotlib, and Seaborn.
- Loaded the dataset and performed an initial exploration using df.head(), df.tail(), df.shape, df.columns, df.describe(), and df.info().
- Checked for missing values in the dataset to understand data completeness.

2.Exploratory Data Analysis (EDA)

- Conducted various visual analyses to understand relationships between features and house prices:

Distribution of House Prices by Area:

- Created a bar plot of average prices by location to identify areas with significantly higher or lower prices.
- Used color palettes and appropriate labeling to enhance plot readability.

Correlation Analysis:

- Analyzed how the number of bedrooms, bathrooms, and floors correlate with house prices using a correlation matrix heatmap.
- Highlighted significant relationships between features and price to identify potential key variables.

House Condition vs. Price:

- Created a bar plot to show the average price based on the house condition (e.g., Excellent, Good, Fair, Poor).
- Provided insights into how house conditions affect selling prices.

Trend in House Prices Based on Year Built:

- Plotted a line graph to visualize the trend of house prices with respect to the year built.
- Investigated whether older houses tend to be priced differently compared to newer constructions.

Location Impact on Price:

- Generated another plot to examine which locations yield the highest average prices.

Garage Availability vs. Price:

- Created a bar plot comparing house prices for homes with and without garages, giving insights into the effect of garage availability on price.

3.Insights and Interpretation

- Each plot contributes to understanding how individual factors influence house prices, which will inform the predictive model later in the project.

CHALLENGES FACED :

1.Handling Missing or Incomplete Data

Issue: As noted in the code, missing values are checked, but if certain critical columns like Price, Location, or Bedrooms have substantial missing data, this could impact the analysis and predictions.

Solution: Missing data can be handled by either imputing with suitable values (e.g., median for numerical features, mode for categorical) or, if necessary, dropping rows or columns with excessive missing values. Careful exploration of the missing data is key to decide the best approach for each feature.

2.Feature Selection and Engineering

Issue: Some features (e.g., YearBuilt, Condition, Location) may impact prices differently depending on location or other factors. Deciding which features are most influential and whether additional derived features (like Property Age = Current Year - YearBuilt) are needed can be challenging.

Solution: Use techniques like correlation analysis, feature importance from initial models, or domain knowledge to guide feature selection. Testing the impact of derived features through trial models can also help determine their usefulness.

3.Categorical Variables and Encoding

Issue: Some columns, like Location and Condition, are categorical, and encoding them effectively is essential. Location might have many unique values, making it complex to encode, as traditional one-hot encoding could create a high-dimensional dataset.

Solution: For high-cardinality features like Location, consider approaches like target encoding (encoding based on average price for each location) or grouping less common values into broader categories. Testing different encoding strategies and evaluating their impact on model performance is recommended.

Unresolved Challenges and Proposed Approaches

1.Dealing with Changing Market Trends in Real Estate

Challenge: The housing market is dynamic, and a model trained on historical data might not adapt well to rapid changes.

Proposed Solution: Implement a model retraining strategy where new data is incorporated periodically to maintain relevance. Alternatively, exploring time-series models to capture trend changes over time could improve predictions.

COLLABORATION:

1. Meeting Frequency
Our team meets bi-weekly to discuss project progress, troubleshoot issues, and assign tasks. This schedule provides us with a balance of regular communication and enough time for independent work on assigned tasks. We also maintain an open communication channel (e.g., Slack, Teams) for quick questions or urgent updates between meetings.

2. Contributions from Group Members
Each team member has taken ownership of specific aspects of the project, and contributions are balanced to ensure meaningful input from everyone. For instance, one member focuses on data preprocessing, another on exploratory data analysis (EDA), and another on model selection and evaluation.

3. Addressing Potential Imbalances in Contribution
If we notice that contributions become uneven, we plan to address this through open discussion, where team members can share any challenges they're facing. We'll reassign or redistribute tasks as needed to ensure each person's workload is manageable and that all members feel fully involved. Additionally, we may adjust task complexity to better match each member's strengths and skill levels, ensuring that everyone's efforts are impactful.

NEXT STEPS:

1.Further Feature Engineering: Creating new features (e.g., age of property, price per square foot) to add predictive value.

2.Data Preprocessing: Handling missing values, encoding categorical variables, and standardizing or normalizing numerical features as necessary.

3.Modeling: Developing and training a machine learning model for house price prediction using algorithms such as linear regression, decision trees, or more advanced models like gradient boosting.

4.Model Evaluation: Evaluating the model performance using appropriate metrics (e.g., RMSE, MAE) and tuning for improved accuracy.