CSST 102

Renalyn N. Pino
BSCS - 3A

SUMMARY REPORT

● Data Exploration and Visualization

1. Loading and Exploring the Data

- The dataset was loaded using `pandas`, and initial exploration was performed using `df.info()`, `df.describe()`, and checking for missing values with `df.isnull().sum()`.

- The dataset had 1000 entries and 5 columns: 'Size (sqft)', 'Bedrooms', 'Age', 'Proximity to Downtown (miles)', and 'Price'. There were no missing values.

2. Visualizations

Line Plot (Size vs. Price)

- Larger house sizes tend to correlate with higher prices.

Bar Plot (Bedrooms vs. Price)

- More bedrooms slightly increase house prices.

Histogram (Age vs. Price)

- Older houses tend to be cheaper.

Line Plot (Proximity to Downtown vs. Price)

- Houses closer to downtown are more expensive.

Price Distribution

- House prices are right-skewed, with more expensive houses being less common.

Correlation Matrix

- Strong correlation between house size and price was observed, and weaker negative correlations were found between proximity to downtown and house price.

● Data Preprocessing

1. Handling Missing Data

- Although there were no missing values, I used `df.fillna(df.mean(), inplace=True)` to ensure any future missing values are replaced with column means.

2. Scaling the Data

- Features ('Size (sqft)', 'Bedrooms', 'Age', and 'Proximity to Downtown (miles)') were standardized using `StandardScaler()` for better performance in the regression model.

● Model Development

CSST 102

1. Splitting the Data
 - The dataset was split into training and testing sets (70% train, 30% test) using `train_test_split()`.
 - Independent variables: 'Size', 'Bedrooms', 'Age', 'Proximity to Downtown'.
 - Dependent variable: 'Price'.

2. Model Training
 - A **Linear Regression** model was trained on the training set using `LinearRegression().fit()`.

3. Model Coefficients
 - The coefficients showed that larger house sizes significantly increased prices, while more distance from downtown and house age slightly reduced prices.

● Model Evaluation

1. Prediction & Error Metrics
- The model's performance was evaluated using `mean_squared_error()` and `r2_score()`.

 - The **Mean Squared Error (MSE)** was **100,214,724.63**, indicating how much on average the model's predictions deviate from actual prices.
 - The R-squared value was 1.00, indicating a perfect fit.

2. Visualization of Predictions
 - A scatter plot comparing actual vs. predicted prices showed strong alignment along the "perfect prediction" line, reinforcing the model's accuracy.

● Challenges Faced

1. Handling Multicollinearity
 - There was a strong correlation between 'Size' and 'Price', which might lead to multicollinearity. Addressing this through more complex feature selection could further improve the model.

2. Visualizing Results
 - A challenge was ensuring the plots effectively communicated insights. I adjusted plot sizes and labels for better clarity.

● Conclusion

Real-World Applicability
The model is useful for predicting house prices based on a few key features. It can help real estate agencies and home buyers estimate prices based on size, number of bedrooms, proximity to downtown, and age.

Potential Limitations in real-world scenarios

Future improvements could include more data features like neighborhood quality or modern amenities to refine predictions and handle possible multicollinearity.