DATA220. LAB-1

PAIR PROGRAMMING GROUP-37

NAME: SHRINIVAS BHUSANNAVAR

LAB1-Exploratory Data Analysis (EDA) Insights Documentation

Price Dataset Analysis

In this part, we shift our focus to the "Price.csv" dataset. Our objective is to conduct a comprehensive exploratory data analysis. We will delve into the dataset's characteristics, perform data preprocessing, and derive meaningful insights from the data. Python will be our tool of choice for this analysis.

The combined insights gathered from of this documentation aim to provide a comprehensive view of data analysis techniques and their applicability across different tools and datasets. Each part contributes to enhancing our understanding of data exploration, data cleaning, and the extraction of valuable information from datasets.

Exploratory Data Analysis (EDA) is a crucial initial step in understanding and extracting insights from a dataset, and this process is particularly valuable in the context of house pricing data. EDA helps us to uncover patterns, relationships, and outliers in the data, which can inform decision-making for real estate professionals, investors, and prospective homebuyers.

The dataset encompasses **4,600** houses and encompasses **18** features, with one of these features being the price of the houses. Consequently, we have a selection of 17 diverse features to investigate and analyze for their impact on house prices. Let's delve into an examination of these features., and all of them belong to the United States (USA).

Category: Numerical Variables

1. **Continuous**

- 'price': Represents continuous house prices in dollars.
- 'sqft_living': Represents continuous living area in square feet.
- 'sqft_lot': Represents continuous lot size in square feet.

2. **Discrete**
   - 'bedrooms': Represents the count of bedrooms in each house (integer values).
   - 'bathrooms': Represents the count of bathrooms in each house (integer values).
   - 'floors': Represents the count of floors in the houses (integer values).
   - 'yr_built': Represents the discrete year of construction for each house.
   - 'yr_renovated': Represents the discrete year of renovation (if any) for each house.

Category: Categorical Variables

1. **Binary (Nominal)**
   - 'waterfront': Represents whether a house is on the waterfront (1 for yes, 0 for no).
2. **Ordinal**
   - 'view': Represents the view rating on a scale from 0 to 4 (ordered categories).
   - 'condition': Represents the condition rating on a scale from 1 to 5 (ordered categories).
3. **Nominal**
   - 'street': Likely represents the nominal categorical variable of street addresses (non-ordered categories).
   - 'city': Represents the nominal categorical variable of the city where each house is located.
   - 'statezip': Likely represents the nominal categorical variable of state and ZIP codes.
   - 'country': Represents the nominal categorical variable of the country.

Here's a consolidated summary of these attributes and their corresponding statistics:

1. **Date**: This likely represents the date when information about each house was recorded.
2. **House Price**: Expressed in dollars, it indicates the cost of the house. The average price is around $551,963, with a broad range from $0 to $26,590,000.
3. **Number of Bedrooms:** Represents the count of bedrooms in each house. The average is approximately 3.4, with a range from 0 to 9.
4. **Number of Bathrooms:** Reflects the count of bathrooms in each house. The average is about 2.16, with a range from 0 to 8.
5. **Living Area:** Indicates the area of the living space in square feet. The average living area is approximately 2,139 square feet, with values ranging from 370 to 13,540 square feet.
6. **Plot Size:** Represents the size of the property's land in square feet. The average lot size is about 14,852 square feet, with a range from 638 to 1,074,218 square feet.
7**. Number of Floors:** Reflects the number of floors, with an average of approximately 1.51, ranging from 1 to 3.5.

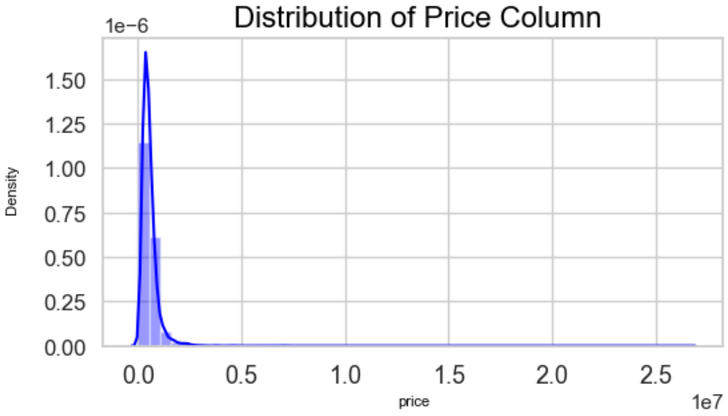
8. **Location by the Sea:** A binary feature, indicating whether the house is situated by the sea or not. The majority of houses are not by the sea.
9. **View:** Rated on a scale from 0 to 4, describing the view from the property. The average view rating is around 0.24.
10. **House Condition:** Rated on a scale from 1 to 5, assessing the condition of the house. There are 5 unique condition values, with an average condition rating of 3.45.
11**. Ground Area of the House**: Represents the area the house covers in square feet. The average above-ground area is about 1,827 square feet, with a range from 370 to 9,410 square feet.
12. **Basement Area**: If available, it indicates the size of the basement in square feet. The average basement area, if present, is approximately 312 square feet, ranging from 0 to 4,820 square feet.
13**. Year of Construction**: Reflects the year the house was originally built. The average year of construction is around 1970, with a range from 1900 to 2014.
14. **Year of Renovation**: If applicable, it shows the year when the house underwent renovation. The average year of renovation is about 808, with a range from 0 to 2014.
15. **Address**: Specifies the specific address of the property.
16. **City:** Indicates the city where the house is located.
17. **Postal Code**: Represents the postal code associated with the property's location.

These summary statistics provide insights into the central tendencies and variability of each attribute, enabling a better understanding of the dataset's characteristics. Moreover, an analysis of unique values within each column reveals the diversity and distinctiveness of the information recorded for these attributes, which is essential for data exploration and comprehension.

Data Cleaning

We identified instances in the dataset where the 'price' column had a value of $0. This $0 price is likely an invalid or outlier value, as it doesn't align with typical real estate pricing. As part of our exploratory data analysis (EDA) process, we have decided to remove all rows with a 'price' of $0. This data cleaning step ensures that our analysis is based on accurate and meaningful information, as including such outliers could skew our results and conclusions."

The distribution of price column:



The distribution depicted in the histogram appears to be **positively skewed distribution**. This means that the majority of data points are concentrated on the left side of the distribution, and the tail of the distribution extends to the right. In other words, there are relatively few data points with very high values, which cause the rightward skew.

The distribution, with its mean (557,905), median (465,000), and mode (300,000), does not exhibit the characteristics of a normal distribution, as it lacks symmetry and does not meet the criteria of a typical bell-shaped curve.as it does not have a single peak in the middle, and the mean, median, and mode are not equal.

While the histogram does not resemble a normal distribution.

Shapiro-Wilk Test:

The Shapiro-Wilk test is a statistical test used to determine whether a given dataset follows a normal distribution. A normal distribution, also known as a Gaussian distribution, is a symmetrical probability distribution characterized by a bell-shaped curve. Many statistical analyses and techniques assume that data is normally distributed, so it's important to check if your data meets this assumption before using those methods.

The Shapiro-Wilk test resulted in a p-value of 0.0, suggesting that the price data is highly unlikely to have a normal distribution. A p-value of 0 typically means strong evidence against the idea that the data follows a normal distribution. Given the extremely small p-value, we can confidently state that the price data does not conform to a normal distribution.

The Hypothesis "The year built has a significant impact on sale price.”. Do a hypothesis test using a t-test: split into two groups: properties built before 1990 and those built-in or after 1990.

Analyzing Standard Deviations and Choosing the Test:

To compare the two groups, one comprising property built before 1990 and the other properties built in or after 1990, we first assess the standard deviations of their 'price' data.

For our data, the results show different standard deviations:

- **Standard deviation for properties built before 1990 : 427923.98**
- **Standard deviation for properties built in or after 1990 : 768153.62**

The notable disparity in standard deviations between the two groups indicates that the variances are not equal. This discrepancy goes against the assumption of equal variance, which is necessary for the student's t-test.

To address this, we opt for Welch's t-test, which doesn't assume equal variance. Welch's t-test adjusts the degrees of freedom, accommodating the unequal variance between the groups, making it a suitable choice for our analysis.

After performing the hypothesis testing:

1. **Null Hypothesis (H0):** It assumes that there is no significant impact of the year built on sale price. In other words, the average sale prices for properties built before 1990 and those built in or after 1990 are not significantly different.

2. **Alternative Hypothesis (Ha):** It suggests that there is a significant impact of the year built on sale price. This means that the average sale prices for the two groups are indeed significantly different.

The significance level (alpha) is set at 0.05, which is a commonly used threshold. If the calculated p-value is less than alpha, it provides evidence against the null hypothesis, indicating that the data supports the alternative hypothesis. **In this case, the code's output concludes that "The year built has a significant impact on sale price."**

On the other hand, if the p-value is greater than or equal to alpha, it fails to provide sufficient evidence to reject the null hypothesis. This implies that the data does not support the idea of a significant impact of the year built on sale price, and the code's output states, "There is no significant impact of year built on sale price."

The specific result, "The year built has a significant impact on sale price (p-value =", indicates that the data suggests a significant relationship between the year built and sale price based on the chosen significance level.

Consider the Hypothesis "The year built has a significant impact on sale price.Do a Hypothesis test using ANOVA: Assume that you have 3 groups: groupA has all houses built in 1990, groupB has all houses that were built in 2000 and groupC has all houses built in 2010 (alpha/confidence interval = 95%)

F-statistic: The F-statistic is a measure of the ratio of the variance between the groups (variation in sale prices between the groups) to the variance within the groups (variation in sale prices within each group). In this case, the F-statistic is approximately 3.54.

p-value: The p-value is a probability that helps you assess whether the observed differences in sale prices between the three groups could have occurred by random chance. A smaller p-value indicates stronger evidence against the null hypothesis. In this case, the p-value is approximately 0.0315.

Interpreting the results:

Reject the null hypothesis: In hypothesis testing, when the p-value is less than the chosen significance level (alpha), which is typically set at 0.05, we reject the null hypothesis. In this case, with a p-value of 0.0315, which is less than 0.05, you reject the null hypothesis.

There is a significant difference in sale price between the three groups: Since we have rejected the null hypothesis, this means that there is statistical evidence to support the alternative hypothesis. In practical terms**, it indicates that there is a significant difference in the sale prices of properties built in 1990, 2000, and 2010.**

So, in summary, the ANOVA test results suggest that the year a property was built has a statistically significant impact on its sale price. There are differences in sale prices between the three groups, and these differences are unlikely to have occurred by random chance alone.

Does The ANOVA Conclusion Change from The T-Test Or Is It The Same?

The conclusions of both these tests are the **same**: The year built significantly impacts the house sale price. However, the ANOVA test provides a more granular analysis by comparing three different years rather than just a split at 1990. This might reveal more subtle differences in property prices depending on the exact year of construction.

Since you have rejected the null hypothesis, this means that there is a significant difference between the sale prices of the two groups. Simply put, the year a property was built has a statistically significant effect on the sale price.

Since you have rejected the null hypothesis in this scenario the same as the t-test, it suggests that there is a significant difference in the average sale prices between these three groups. All in all, the house construction year affects the house sale price.

Calculate The Covariance Matrix of The Numerical Features Present In The Dataset

First, we create a list of numerical columns by picking those that have numeric data types.

Columns with numerical data:
['Price', 'Bedrooms', 'Bathrooms', 'Sqft_Living', 'Sqft_Lot', 'Floors', 'Waterfront', 'View', 'Condition', 'Sqft_Above', 'Sqft_Basement', 'Yr_Built', 'Yr_Renovated']

dataframe.cov(): Method is used to compute the covariance matrix for the numerical columns in a DataFrame.

The diagonal contains the variances Off-diagonals contain the covariances Symmetric matrix Positive covariance means positive relationship Negative covariance means negative or inverse relationship Reviewing the covariance matrix can identify relationships between features that could be further analyzed through correlation plots or regression models.

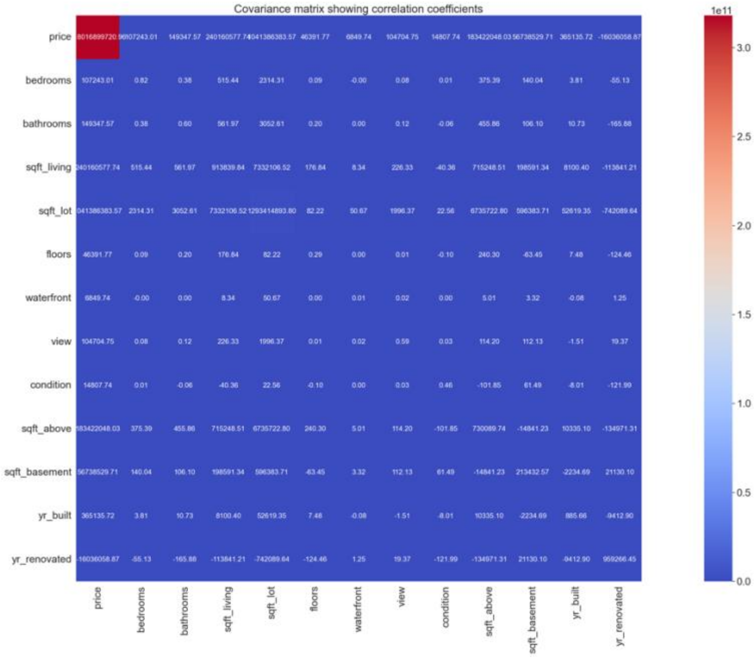
Create a heatmap of the covariance matrix. What do the colors in the heatmap represent?

The heatmap of the covariance matrix represents the direction of the covariance between each pair of variables.

Covariance indicates the direction of the linear relationship between numerical variables. If the value is positive, then the direction is upward. If the value is negative, then the direction is downward.

The covariance value won't help to compare the strength of the relationship between different variables, as the units of the variables may be the same or different. This makes the covariance difficult to use to interpret the strength of direction.

For this reason, correlation is used, as it is a normalized version of covariance. And, the range of value is between - 1 to +1, which indicates both direction and strength of the linear relationship between 2 numerical variables.



- **Red:** Strong positive correlation
- **Dark blue:** Strong negative correlation
- **Light blue:** Weak negative correlation
- **Light red:** Weak positive correlation
- **Gray:** No correlation
- The darker the color, the stronger the correlation

Compute the eigenvalue, eigenvector, and Rank of the covariance matrix.

In code, we're performing calculations on the covariance matrix of a dataset, and here's a simple explanation of what's happening along with the output for documentation:

We are working with a covariance matrix that summarizes the relationships between different variables in our dataset. The code below performs three key calculations:

1. **Eigenvalues and Eigenvectors**: We compute the eigenvalues and eigenvectors of the covariance matrix. Eigenvalues represent the magnitude of variation in the data, while eigenvectors provide the direction of that variation. These values are important for understanding the data's principal components.

2. **Rank Calculation:** We determine the rank of the covariance matrix. The rank represents the number of linearly independent variables in the data. A rank of 12, in this case, indicates that there are 12 linearly independent variables among the features.

Outcome Results:

```
Eigenvalues:
[3.94220649e+00 2.07167801e+00 1.24522399e+00 1.13534974e+00
 1.00257135e+00 8.44979485e-01 6.87318346e-01 6.18825531e-01
 2.25679557e-01 3.75567539e-01 4.10327308e-01 4.43129801e-01
 6.12547932e-16]
```

Eigenvalues: These are the eigenvalues of the covariance matrix, representing the amount of variation in the data associated with each eigenvector. The values are presented in descending order, with the largest eigenvalue first. The last value, which is very close to zero, often indicates negligible variation.

Rank of the Covariance Matrix: The rank of the covariance matrix is 12, indicating that there are 12 linearly independent variables among the features.

This information is valuable for various data analysis and dimensionality reduction techniques, such as PCA, as it helps in understanding the underlying structure and relationships in the dataset.

Interpret the Eigenvectors in the context of the dataset. What do they represent and their significance?

Eigenvectors represent the direction of maximum variance in the dataset. In the context of our housing price dataset, these eigenvectors can be seen as patterns or relationships among the various features. Each eigenvector corresponds to a combination of features, and their values indicate how much each feature contributes to that pattern.

Based on eigen vectors data output :

When eigenvector has higher values for 'bedrooms,' 'bathrooms,' and 'living area,' it suggests that these features are correlated in a specific way within the data. This could mean that houses with more bedrooms, bathrooms, and larger living areas tend to have higher prices. On the other hand, an eigenvector with significant contributions from 'condition' and 'year built' will represent a different pattern, indicating that the condition and age of a house play a crucial role in determining its price.

Significance:

1. **Feature Importance:** The eigenvectors help us identify which features have the most impact on the dataset's variability. Features associated with the largest eigenvalues have the most influence, making them crucial for understanding price fluctuations.
2. **Dimension Reduction:** Eigenvectors are used in techniques like Principal Component Analysis (PCA) to reduce the dataset's dimensionality while retaining as much information as possible. This can simplify the dataset while preserving its essential characteristics.
3. **Feature Relationships:** Eigenvectors show how features are related. Positive or negative values within eigenvectors indicate whether features tend to increase or decrease together, revealing underlying patterns in the dataset.

In summary, eigenvectors in our price dataset help us identify influential features, understand feature relationships, and can be used for dimension reduction to simplify analysis. They are a valuable tool in exploratory data analysis for uncovering the dataset's underlying structure.

Inverse of the Covariance Matrix

Calculating the inverse of the covariance matrix is an essential step in various statistical and machine learning techniques. Here's an explanation of what this code does and its output for your documentation:

Explanation:

In this code, we compute the inverse of the covariance matrix. The covariance matrix contains information about how each pair of variables in the dataset relates to one another. Calculating the inverse allows us to analyze the relationships from a different perspective, which can be valuable in various data analysis and modeling tasks.

Discuss the impact of the matrix rank on the feasibility of solving a linear regression problem using these features.

The rank of a matrix indicates how many independent pieces of information it contains. For a linear regression problem, the matrix contains the features we want to use to predict the output.

In our case, we have 13 features, but the matrix rank is only 12. This means that out of the 13 features, there are only 12 independent pieces of information. The 13th feature is redundant and can be calculated from the other 12 features. Having redundant features can cause problems when trying to solve the linear regression. With 13 features but only 12 independent pieces of information, the matrix becomes non-invertible. This means we cannot uniquely solve for the regression coefficients.

In technical terms, this is called a "**rank deficient**" matrix. The matrix does not have full rank.

The impact of this rank deficiency on solving a linear regression problem is as follows:

1. **Overfitting**: Excess features compared to linearly independent observations may lead to overfitting, compromising the model's generalization.

2. **Ill-Conditioned Matrices**: Multicollinearity can create ill-conditioned design matrices, causing numerical instability in regression solution.

3. **Reduced Model Interpretability**: High feature correlation hampers isolating individual feature impacts, diminishing model interpretability and feature importance identification.

In summary, when the matrix rank is lower than the number of features in a linear regression problem, it's important to address multicollinearity to ensure a well-posed and interpretable model and to prevent issues like overfitting and numerical instability.

- Sqft_living vs. Bathrooms: Strong positive correlation (0.76). More bathrooms tend to lead to a more sqft_area.
- Sqft_living vs. Bedrooms: Moderate positive correlation (0.6). More Bedrooms tend to lead to a more sqft_area.
- Sqft_above vs. Sqft_living: Strong positive correlation (0.88). Larger properties tend to have higher Sqft_above.
- Condition vs. Year Built: Moderate negative correlation (-0.40). Older properties typically will not have well condition.

How does the rank relate to multicollinearity among the independent variables? Provide examples from the dataset.

The rank of a matrix is the number of linearly independent rows or columns in the matrix. Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated.

There is a close relationship between the rank of a matrix and multicollinearity. If the rank of a matrix is less than the number of columns in the matrix, then the matrix is ill-conditioned, and multicollinearity is present.

Looking at the correlation matrix, we can see evidence of potential multicollinearity:

- **The correlation between Sqft_living and sqrft_above is 0.88**. This is a very strong correlation, so there is a high chance of multicollinearity between these two features.

- **The correlation between Sqft_living vs. Bathrooms is 0.76**. This is a very strong correlation, so there is a high chance of multicollinearity between these two features.

If we want to develop a model to predict the price of a property, its recommended removing one or more of the correlated features from model. For example, we can remove the sqrft_above feature, since it is highly correlated with the sqrft_living feature. We can also combine the bedrooms and bathrooms features into a single feature, such as "total number of bedrooms and bathrooms."

If we try to fit a linear regression model with all 13 features, the model will be overfitting to the data. This is because the model will not be able to determine which features are most important for predicting the target variable, since few of the features are highly correlated.

Create a matrix X with the selected_features = ['bedrooms', 'sqft_lot', 'floors', 'yr_built'] and Y with the target feature and Print matrix X and Y

Creating matrices X and Y based on selected features is a fundamental step in preparing data for machine learning.

In this code, we are selecting specific features from the dataset to

create a matrix X and a vector Y

- **Matrix X (X)**: It is a matrix that contains the selected features (in this case, 'bedrooms,' 'sqft_lot,' 'floors,' and 'yr_built'). Each row represents a data point, and each column corresponds to a selected feature. X is used as the input data for machine learning algorithms.
- **Vector Y (Y)**: It is a vector that contains the target feature, 'price.' Y represents the values we want to predict using machine learning models. Each element in Y corresponds to the price of a property in the dataset.

These matrices are crucial for training and evaluating machine learning models to predict house prices based on the selected features.

Calculating the transpose of matrix X

In this code, we compute the transpose of matrix X (denoted as X_transpose). The transpose operation essentially flips the rows and columns of the matrix, so rows become columns and columns become rows.

Output:

- **Transpose of matrix X (X_transpose)**: The output is the transposed matrix X. It contains the same data as the original matrix X, but with rows and columns interchanged. Each row in X_transpose corresponds to a specific feature ('bedrooms,' 'sqft_lot,' 'floors,' 'yr_built'), and each column represents a data point (a property).

The transposed matrix is useful for various operations, including matrix multiplication and solving linear equations. It can be valuable for certain data analysis and modeling tasks.

Solving a linear system of equations : it is a crucial step in linear regression to find the coefficients.

In this code, we solve the linear system of equations 'X * a = Y' to find the coefficients 'a' that best fit the data. This is essential in linear regression, where we aim to model the relationship between selected features (matrix X) and the target variable (vector Y).

Outcome:

Coefficients (a): The output is the vector 'a,' which contains the coefficients for each selected feature. These coefficients represent the weights assigned to each feature in the linear regression model. In the context of our dataset, these coefficients (rounded for simplicity) are as follows:

- Coefficient for 'bedrooms': approximately 118,263

- Coefficient for 'sqft_lot': approximately 0.59

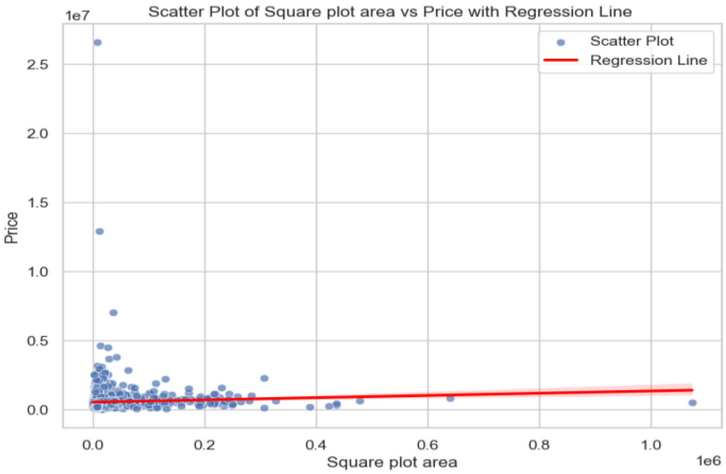
- Coefficient for 'floors': approximately 128,190

- Coefficient for 'yr_built': approximately -23.71

These coefficients are used to construct the linear regression equation that predicts house prices based on the selected features.

The code also calculates other statistical information such as residuals (the differences between predicted and actual values), rank, and singular values. These details can be valuable for assessing the quality and reliability of the linear regression model.

The scatter plot of square plot area vs price with regression line shows a positive correlation between the two variables, meaning that as the square plot area increases, the price also tends to increase. The regression line is a best-fit line that shows the overall trend of the data.



The scatter plot shows that there is a lot of variation in the data, with some plots falling well above or below the regression line. This suggests that there are other factors that also affect the price of a square plot of land, such as yr_built, amenities, and overall market conditions.

1. Linear Relationship: The scatter plot shows a linear trend where, in general, as the square plot area increases, house prices tend to rise. This suggests a positive correlation between the two variables. It's essential to note that while this correlation is evident, there is still some level of variability in house prices for a given square plot area.

2. Outliers: There are a few data points that deviate significantly from the linear trend. These points, which lie far from the regression line, may be considered as potential outliers. Investigating these outliers could reveal unique properties or unusual circumstances that affect pricing.

3. Price Range: The plot demonstrates that houses with smaller square plot areas are available in a wide price range, but as the square plot area increases, the price range narrows. This suggests that square plot area becomes a more critical factor in determining prices for larger properties.

4. Regression Line: The red regression line on the plot represents the linear relationship between square plot area and price. This line can be used for price predictions based on square plot area. The slope of the line indicates how much price increases, on average, for each additional unit of square plot area.

5. Decision Support: For individuals involved in real estate, understanding this relationship is vital for pricing strategies. It helps to estimate a house's price based on its square plot area, which is useful for both buyers and sellers. However, other factors not considered in this analysis may also influence house prices.

Overall, this scatter plot is a useful visualization for understanding how the square plot area impacts house prices. It provides a starting point for pricing decisions and further analysis of real estate data. Further investigation and modeling can incorporate additional variables to enhance the accuracy of house price predictions.

However, it is important to be aware of the limitations of the scatter plot and to interpret it carefully.