***House Price Prediction (Linear Regression) - Result Sheet***

**1. Data Overview**

**Dataset**: California Housing Prices Dataset

**Target Variable**: Price

**Features Used**:

Median Income (MedInc)

Average Number of Rooms (AveRooms)

House Age (HouseAge)

Average Number of Bed Rooms ('AveBedrms')

Population (Population)

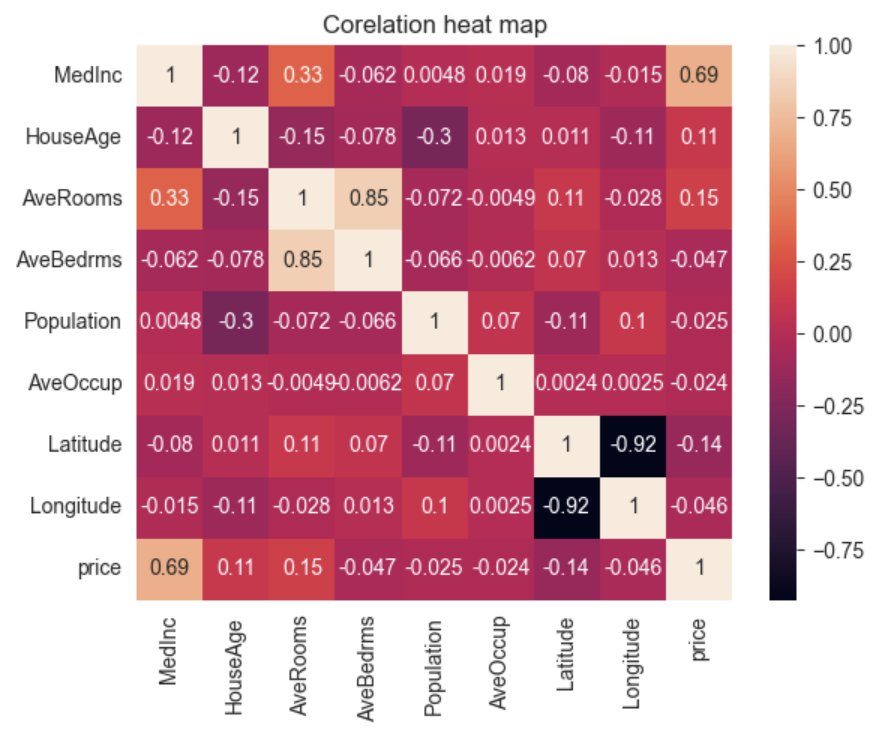
Average Number of (AveOccup)

Latitude (Latitude)

Longitude (Longitude)

**2. Preprocessing Steps**

* **Handling Missing Values**: No null values are present

**3. Exploratory Data Analysis**

** MedInc (Median Income):**

* **Median income is having strong positive correlation with price (0.69), meaning that higher median income is associated with higher house prices. Its positive correlation with AveRooms (0.33), suggesting higher income areas have more average rooms.**
* **HouseAge is having small positive correlation with price (0.11), indicating older houses may have slightly higher prices. Its negative correlation with Population (-0.30), suggesting that areas with older homes tend to have fewer people.**
* **AveRooms is having strong positive correlation with AveBedrms (0.85), as expected (more rooms tend to imply more bedrooms). It’s also having a moderate positive correlation with price (0.15).**
* **Strong negative correlation between Latitude and Longitude (-0.92), showing an expected spatial geographic relationship. Both have relatively weak correlations with house prices.**
* **Population is having very weak negative correlation with price (-0.025), suggesting population size has little impact on house prices.**
* **AveOccup isn’t having any significant correlations with other variables, indicating it has little impact on house prices or other factors.**

**4. Model Performance**

**Model**: Linear Regression

**Train-Test Split**:

* + Training set size: 56%
  + Validation set size: 14%
  + Test set size: 30%

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

**Evaluation Metrics:**

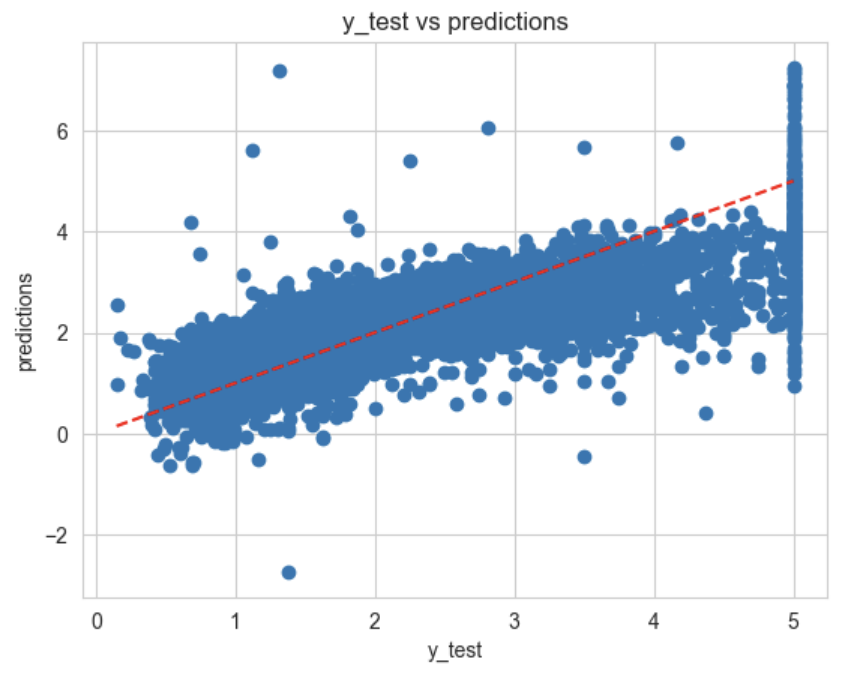
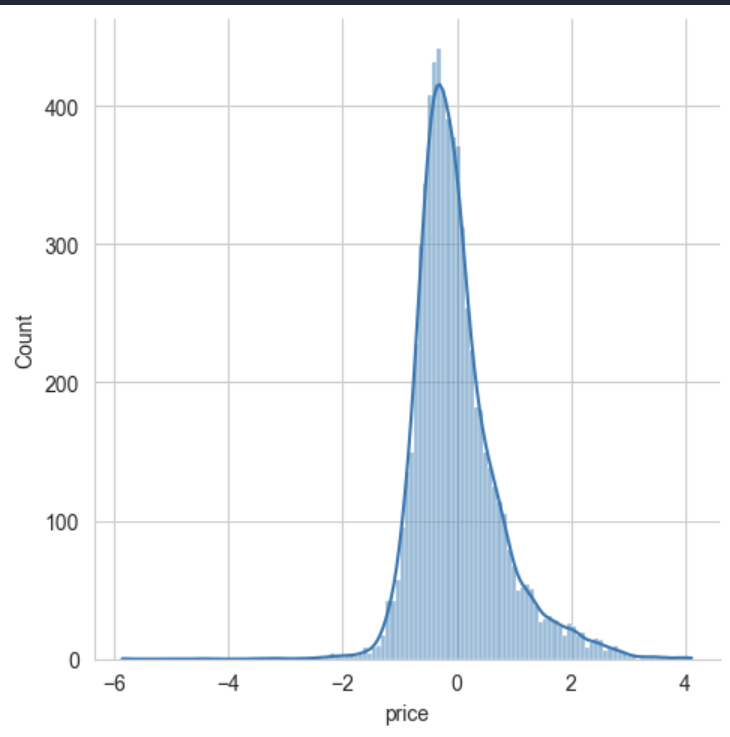
|  |  |  |
| --- | --- | --- |
| **Metrics** | **Train Set** | **Test Set** |
| **Mean Absolute Error** | 0.5299412801865069 | 0.5374839109617068 |
| **Mean Squared Error** | 0.5245717015872389 | 0.5468505447982505 |
| **Root Men Squared Error** | 0.7242732230223887 | 0.7394934379683503 |

**Metric Interpretation:**

* MAE: Calculates the average of the absolute differences between predicted and actual values.
* **MSE**: Measures the average squared difference between actual and predicted prices. Lower values indicate better performance.
* **RMSE**: The square root of MSE, providing error in the same unit as the target (i.e., house prices).

**5. Residual Analysis**

* **Residual Plot**: Visualizing the residuals (difference between actual and predicted values) to assess model fit.



* While many data points are very close to the regression line, there are some deviations. Some predictions appear to be quite far from the actual values, especially at the upper end of y\_test, where the variance increases. This suggests the model overestimates or underestimates at the extremes.
* In the histplot for (y\_test-predictions) the mean is almost zero which indicates the y\_test is almost equal to predictions and model is working well

**6. Feature Importance**

* **Coefficients**: Interpreting the coefficients from the linear regression model.

|  |  |  |
| --- | --- | --- |
| **Columns** | **Coefficients** | **Interpretation** |
| MedInc | 0.443325 | For every one-unit increase in median income, the house price increases by approximately 0.4433 units, holding all other variables constant. |
| HouseAge | 0.009035 | For every one-unit increase in house age, the house price increases by approximately 0.009035units, holding all other variables constant. |
| AveRooms | -0.116506 | For every one-unit increase in average no of rooms, the house price decreases by approximately 0.116506units, holding all other variables constant. |
| AveBedrms | 0.612772 | For every one-unit increase in average no of bedrooms, the house price increases by approximately 0.612772units, holding all other variables constant. |
| Population | -0.000009 | For every one-unit increase in Population, the house price decreases by approximately 0.000009 units, holding all other variables constant. |
| AveOccup | -0.006233 | For every one-unit increase in Average Occupancy, the house price decreases by approximately 0.006233 units, holding all other variables constant. |
| Latitude | -0.404732 | For every one-unit increase in Latitude,the house price decreases by approximately 0.006233 units, holding all other variables constant. |
| Longitude | -0.417610 | For every one-unit increase in Longitude, the house price decreases by approximately 0.41 units, holding all other variables constant. |

**7.Conclusion**

* **Summary of Findings**:
  + **Key Influencers**: Median Income and Average no of bedrooms are the strongest predictors of house prices.
  + **Model Performance**: The model performed well with an RMS score of 0.7394934379683503 on the test set, indicating that the features explain a significant portion of the variance in house prices.
  + **Improvement Areas**: Further enhancement might include adding more relevant features, handling outliers better, or experimenting with more advanced models.