Project Name

PREDICTIVE ANALYSIS AND MODELING OF CALIFORNIA HOUSING PRICES.

Student Name: *Rutvi Vipulkumar Kathiriya*

RIT Email ID: rk4739@rit.edu

Course: DSCI 633: Foundations of Data Science
Instructor: Dr. Nidhi Rastogi
Date: 15 December 2023

I. Data Science Problem

The primary objective of this project is to develop a predictive model for housing prices in California. The dataset utilized for this analysis is the California housing dataset, which comprises various features such as median income, house age, average rooms, and population. The overarching goal is to understand the relationships between these features and the target variable (median housing price) and construct a model that accurately predicts housing prices.

II. Data and Model Description

Dataset Overview

This Dataset has a lot of information from the 1990 U.S. census. It was carefully put together from rows that represent census block groups. These block groups, which are essential to the Census data, are the smallest physical units and usually have between 600 and 3,000 people living in them.

The California housing dataset covers a huge area. It has 20,640 cases, and each one is defined by a set of nine different features. These features include important things like the typical income, the average age of the homes, and their GPS coordinates. The target variable, which shows the typical home price, is the most important part of the dataset. The information basically shows a wide range of different housing-related factors spread out in different parts of the California state. This includes a lot of different kinds of information from many census block groups, which makes sure that we have a full and accurate picture of how housing works in California.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis involved computing key statistics for each feature, unveiling insights into the data's central tendencies and variability. The analysis aimed to identify potential outliers and anomalies through histograms and box plots, ensuring a thorough understanding of the dataset's characteristics.

The dataset was found to be free of missing values, and anomalies and outliers were visualized using histograms and box plots.

```
# Visualization for Anomaly Detection
print("\nHistograms for Feature Distribution:")
data.hist(bins=50, figsize=(20, 15))
plt.show()
```

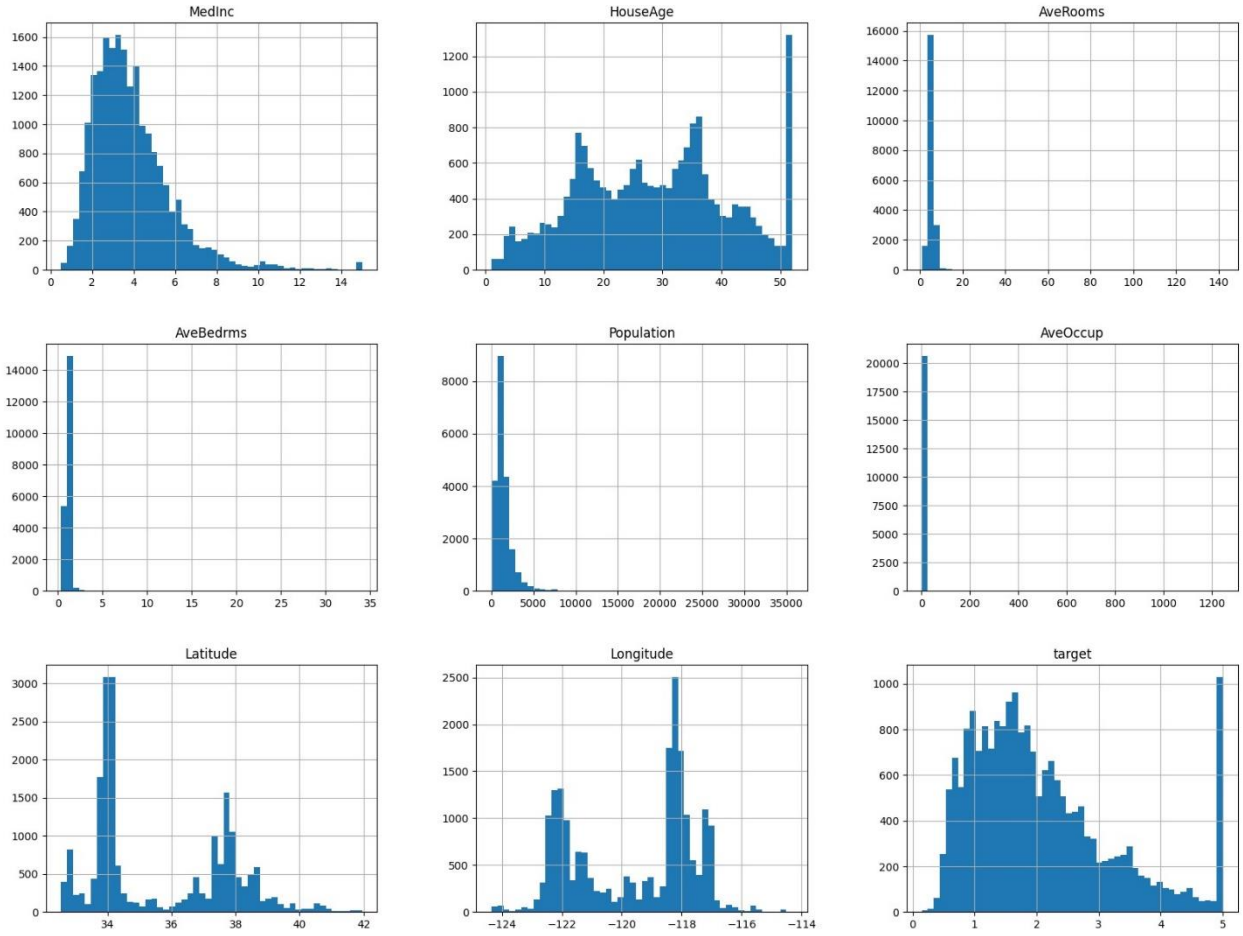

Figure 1: Histograms for Feature Distribution

The picture shows histograms for different traits in the dataset. Each histogram shows how often data points fall within a certain set of values, or "bins."

Key Observations:

- AveRooms: The distribution looks even, with only a small rightward skew. This means that most houses have a normal number of rooms, but some may have a lot more.
- AveRooms: the distribution of AveBedrms has a small right skew, which means that houses with more bedrooms are more likely to be found.
- Population: The data is skewed to the right, and the numbers that are higher in the population have a longer tail. This means that even though most places have average numbers, some may have a lot more people living in them.
- Medium: The distribution is skewed to the left, with a longer tail going down to lower median income levels. This means that a bigger part of the dataset is made up of people with lower incomes.
- Latitude and Longitude: The two sets of numbers look pretty much the same, which means that the data points are spread out evenly across the world.

Potential Anomalies:

- Outliers at the tails of some distributions: There may be problems that need to be investigated if there are data points at the very ends of some histograms, like a very high number of rooms in AveRooms.
- Not even spacing between bins: Sometimes, the bins might not be spaced out evenly, which could hide trends or change how the data is shown.

Box Plots for Outlier Detection:

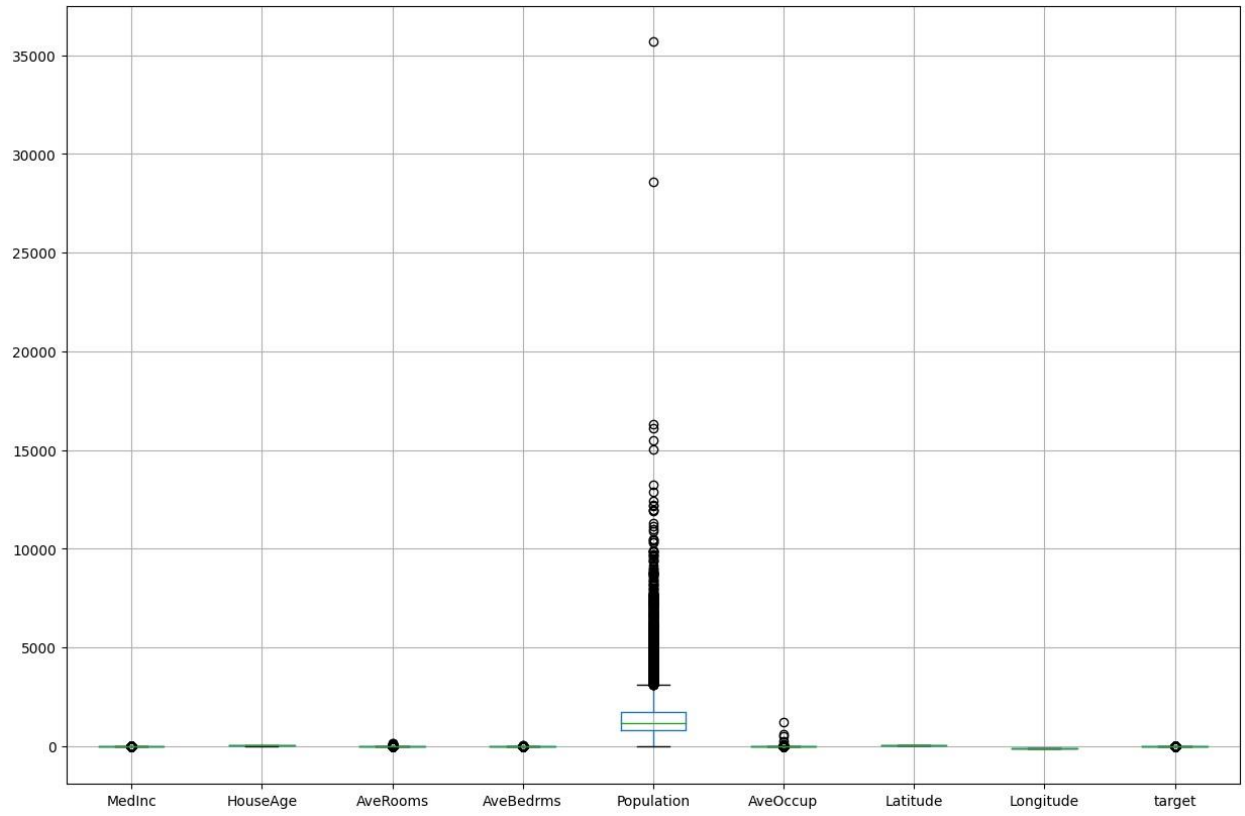

Figure 2 : Box Plots for Outlier Detection

Model Description

In this part, we provide an in-depth look at the regression models used to forecast house prices in the California dataset. Each model is designed to capture the complex interactions between many characteristics and the goal variable, median house price.

1. Regression Linear

A fundamental technique in regression analysis, linear regression, presupposes a linear connection between the independent variables and the target variable. The goal of the model is to identify the best linear equation that minimizes the sum of squared discrepancies between anticipated and actual values. Linear Regression is used in this research to estimate the median dwelling price based on characteristics such as median income, house age, and geographical coordinates.

2. Ridge Regression

Ridge Regression adds regularization to the linear regression cost function, which helps to prevent overfitting by penalizing large coefficients. The regularization term, which is controlled by a hyperparameter (alpha), is especially useful in dealing with multicollinearity difficulties. Ridge Regression improves model stability by restricting the magnitude of coefficients.

3. Lasso Regression

Similar to Ridge Regression, Lasso Regression uses regularization, but with L1 regularization, which encourages sparsity in the coefficient vector. Some coefficients are set to exactly zero, resulting in feature selection. When dealing with datasets with numerous features, Lasso Regression improves model interpretability.

4. Random Forest Regression

Random Forest Regression is a form of ensemble learning that builds several decision trees during training. The final prediction is the average of all tree predictions, providing robust performance and resistance to overfitting. To improve forecast accuracy, model parameters such as the number of trees and tree depth can be tweaked.

5. PCA (Principal Component Analysis)

Principal Component Analysis (PCA) is a technique for reducing dimensionality that is used in conjunction with other models. It converts the original characteristics into principle components, capturing the data's most important variance. By focusing on the most informative characteristics, PCA can help reduce computational complexity and potentially improve model performance.

These regression models constitute a diversified toolkit for handling the challenge of predicting house prices. Models are chosen based on their distinct characteristics, such as linear relationships, regularization capabilities, ensemble learning, and dimensionality reduction. In the following sections, we assess the performance of each model to determine the most effective method for predicting home prices in California.

III. Analysis Strategy

Feature Engineering

Skewed features, such as 'AveRooms' and 'AveBedrms,' underwent log transformation to mitigate skewness and enhance model performance. The impact of log transformation on feature distribution was visualized to assess its effectiveness.

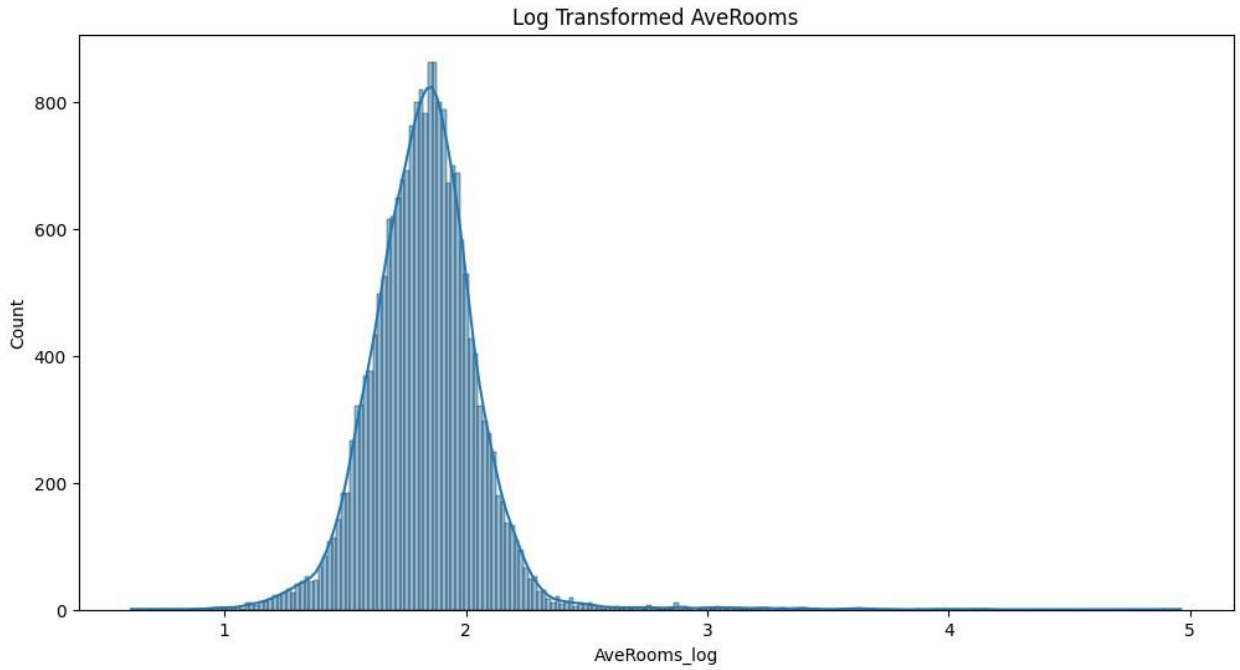

Figure 3: Visualization of Log Transformation:

```
# Applying Log Transformation to skewed features
skewed_features = ['AveRooms', 'AveBedrms']  # Replace with actual skewed features
for feature in skewed_features:
    data[feature + '_log'] = np.log1p(data[feature])

# Visualize the effect of Log Transformation
print("\nVisualization of Log Transformation:")
plt.figure(figsize=(12, 6))
sns.histplot(data['AveRooms_log'], kde=True)
plt.title('Log Transformed AveRooms')
plt.show()
```

The effect of log transformation on the distribution of AveRooms in your California housing dataset is demonstrated in this visualization. We can acquire useful insights into the possible benefits of this data preparation strategy by comparing the original and changed distributions.

Distribution using Log Transform:

The log transformation has a significant impact on the distribution (insert image of log-transformed distribution). The converted data now has a more symmetrical and bell-shaped curve, indicating that it is closer to normalcy.

This transition has various benefits:

- Reduced skewness: The log transformation effectively mitigates the right skew, allowing each data point to be given a more equal weight in the analysis.

- Model performance has been improved: Because the assumptions of normalcy are better met, normalizing the distribution can lead to more accurate and robust regression models.
- Log transformation can also serve to stabilize the variance of the data, making it less prone to outliers and enhancing the model's overall stability.

Overall:

The effect of log transformation on AveRooms is visualized to highlight its potential to increase data quality and model performance. Log transformation, by reducing skewness and normalizing the distribution, can provide more precise and trustworthy insights into the correlations between characteristics and the target variable.

Correlation Analysis

A correlation matrix was constructed and visualized using a heatmap to understand feature relationships. This analysis guided feature selection and contributed to the development of robust predictive models.

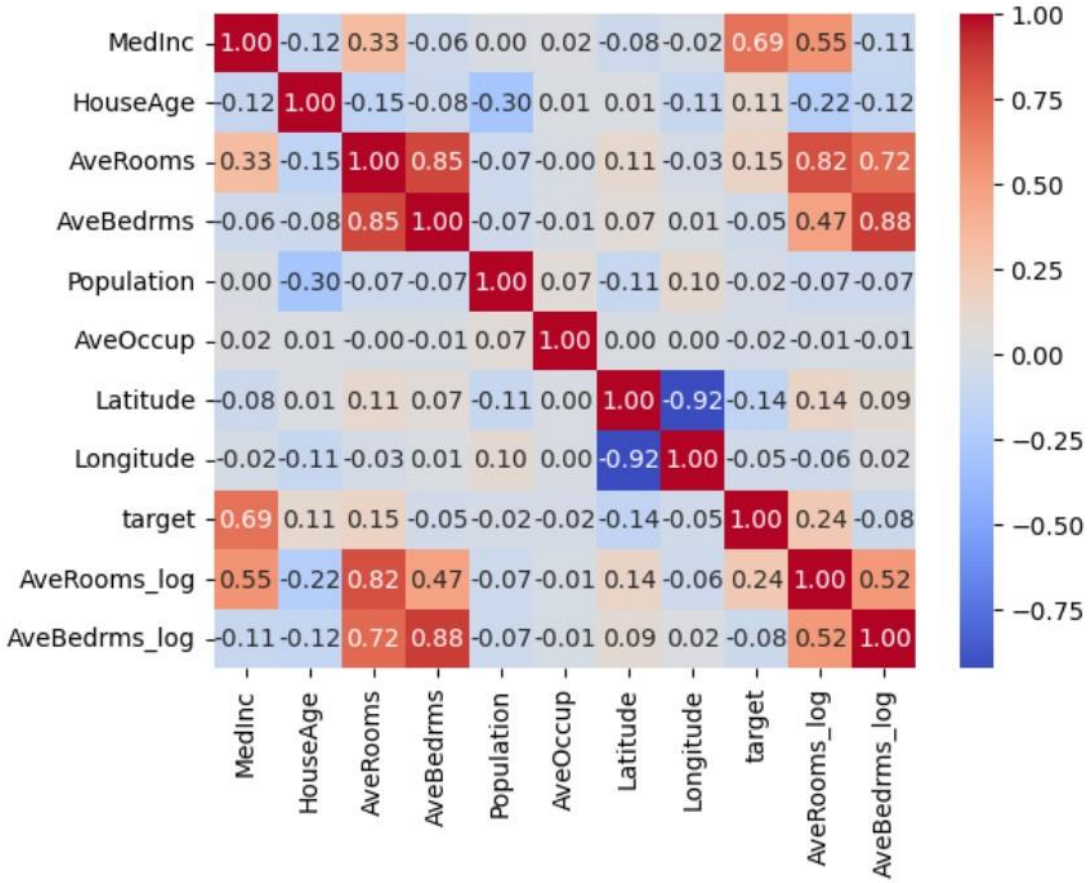

Figure 4: Correlation Matrix

- This heatmap reveals the relationships between various features in the California housing dataset, with color intensity indicating the strength of the correlation.
- Strong positive correlations, like the one between AveRooms and AveBedrms, suggest complementary trends, while negative correlations, like Median Income and Longitude, hint at potential trade-offs.

- This visual analysis can guide feature selection for modeling, helping to identify potentially redundant or irrelevant features.
- Overall, the heatmap offers valuable insights into the complex relationships within the California housing data, informing further analysis and model building.

Geographic Visualization

Scatter plots on the California map were created to visualize the spatial distribution of housing prices. This geographical representation provided insights into regional variations, facilitating a comprehensive understanding of the dataset.

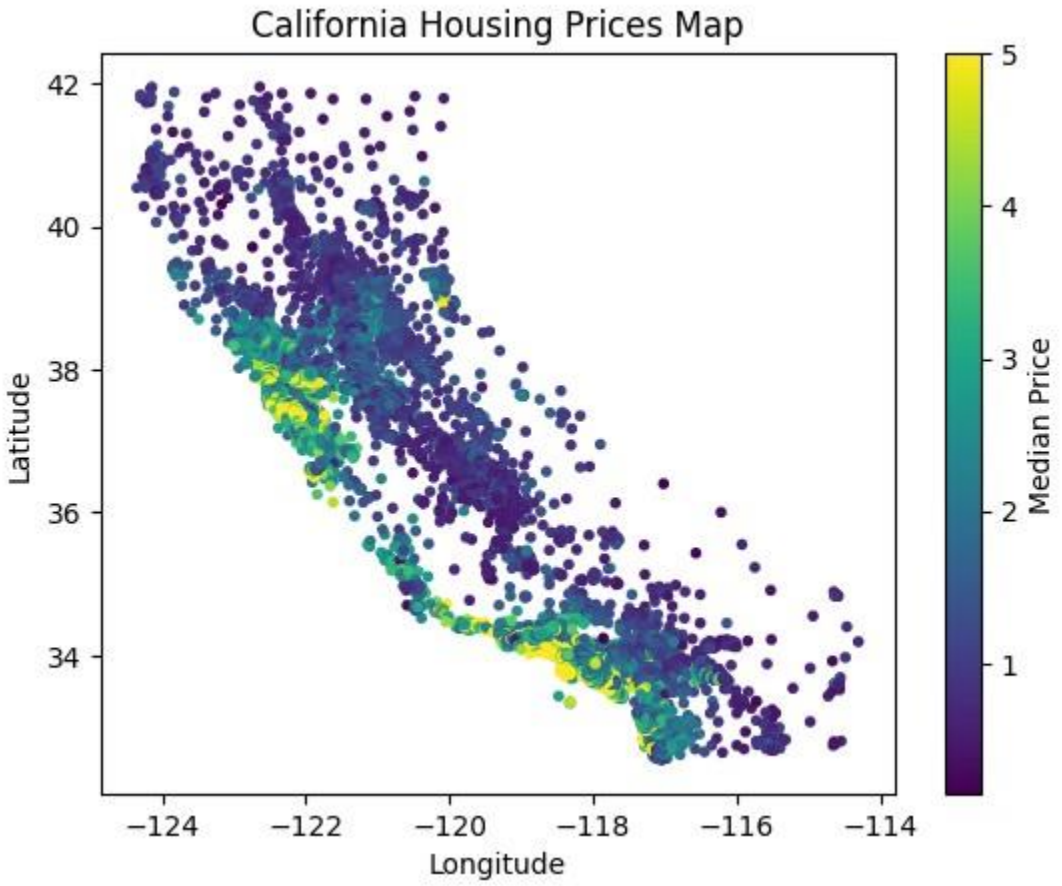

Figure 5: California Housing Prices Map

- Regional variances: The graphic clearly shows various regional variances in California house prices. Higher median prices are reflected by darker green and yellow hues in coastal cities like San Francisco and San Diego, as well as the Central Valley. In contrast, the inland regions and northernmost parts of the state have lower median prices, as illustrated by the brighter blues and greens.

- Clustering and Trends: The scatter plot indicates potential pricing groupings within specific locations. The San Francisco Bay Area, for example, shows as a dense cluster of darker greens, indicating a narrow range of high median values within that area. Similarly, the Central Valley has a concentrated zone of yellows and greens, indicating a cluster of median prices ranging from moderate to higher.

- Visualizing Price Distribution: This map works well in conjunction with other visualizations such as histograms or scatter plots of attributes. It allows you to see how price trends differ geographically and uncover potential linkages between geography and other factors impacting home costs by providing a visual context.

IV. Analysis Code

Model evaluation was performed using key metrics: Mean Squared Error (MSE) and R-squared (R2) Score. The models were assessed both with and without Principal Component Analysis (PCA) to evaluate the impact of dimensionality reduction.

Model Performance:

```
# Model Selection and Training
linear_model = LinearRegression()
linear_model.fit(X_train, y_train)

ridge_model = Ridge(alpha=1.0)
ridge_model.fit(X_train, y_train)

lasso_model = Lasso(alpha=0.1)
lasso_model.fit(X_train, y_train)

rf_model = RandomForestRegressor(random_state=42)
param_grid_rf = {
    'n_estimators': [100, 200],
    'max_depth': [10, 20, 30],
    'min_samples_split': [2, 5]
}
```

```
[ ] grid_search_rf = GridSearchCV(rf_model, param_grid_rf, cv=3, scoring='neg_mean_squared_error')
    grid_search_rf.fit(X_train, y_train)
    best_rf_model = grid_search_rf.best_estimator_
```

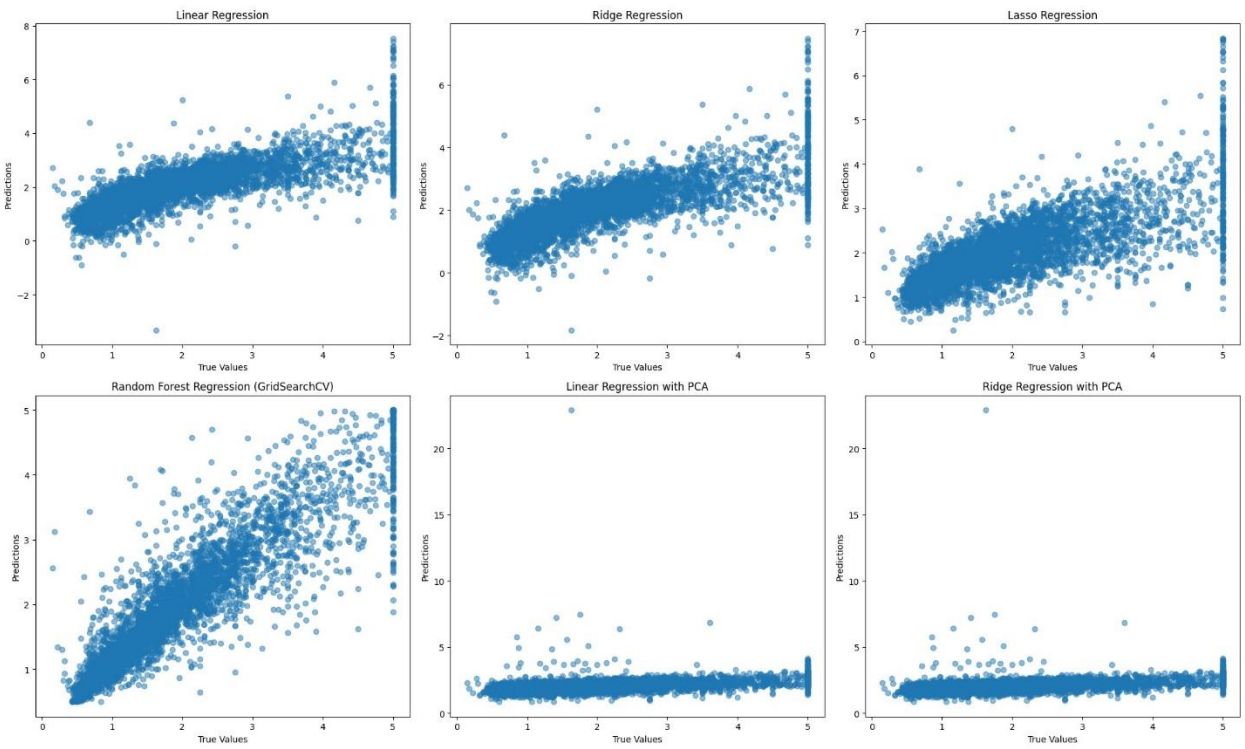

Figure 6: Visualization of model prediction

show how well each model describes the connection between what was expected and what happened with home prices. Even though all models follow the general trend, there is a big difference in how close they are to the diagonal line that shows perfect forecast.

Random Forest with GridSearchCV does a great job here. Its scatter points are very close to the vertical, which means it is more accurate than the other models. This benefit comes from several sources:

- Non-linearity: Unlike linear models like Linear Regression or Ridge Regression, Random Forests can deal with relationships between features and the goal variable that are not simple or linear. This is very important for finding small but important changes in the California housing data, where things like location, perks, and market trends all work together in complex ways.
- Feature interactions: Random Forests look at how features affect each other, which gives us a more complete picture of how various things affect prices. This can be very important for figuring out how a bunch of seemingly unimportant traits work together to make a prediction.
- Hyperparameter optimization: GridSearchCV tries out different combos of hyperparameters (tuning knobs) for the Random Forest model to make it work better with this dataset. With this fine-tuning, the model can make the most of its strengths.

Impact of PCA:

The scatter plots for PCA-transformed models show a different picture. Linear Regression and Ridge Regression with PCA both do pretty good work, but they are not as accurate as the Random Forest models. This means that PCA might not be the best method for this dataset, even though it can be useful in some cases.

This is why:

- Loss of information: PCA lowers the number of dimensions in the data by getting rid of some traits or variations on them. This can make things run more smoothly and cut down on overfitting, but it could also get rid of useful data that helps make correct predictions. In the California property data, some small details may not seem important on their own, but when put together with other details, they may reveal important information.
- Limited applicability for non-linear models: PCA is best for relationships that are linear, while Random Forests is best for interactions that are not linear. If you use PCA on the data before the Random Forest model, you might get rid of the information it needs to work well.

All the models can show the general direction of housing prices, but the Random Forest model with GridSearchCV is the most accurate because it can handle complex feature interactions and relationships that aren't linear. PCA, on the other hand, can help reduce the number of dimensions, but it might not be the best choice for datasets with complex relationships because it could miss important data. To do good data analysis and modeling, you need to understand these details and make smart decisions about which models to use and how to prepare the data.

Results:

```
[ ]  # Function to Evaluate Model Performance
     def evaluate_model(name, model, X_test, y_test):
         y_pred = model.predict(X_test)
         mse = mean_squared_error(y_test, y_pred)
         r2 = r2_score(y_test, y_pred)
         print(f"{name}:\nMean Squared Error: {mse}\nR2 Score: {r2}\n")
```

Without PCA:

Linear Regression:
- MSE: 0.5266
- R2 Score: 0.5981

The Linear Regression model had a Mean Squared Error (MSE) of 0.5266 and an R2 Score of 0.5981, indicating that it was effective at capturing variance in the target variable.

Ridge Regression:
- MSE: 0.5239
- R2 Score: 0.6002

Ridge Regression performed somewhat better, with an MSE of 0.5239 and an R2 Score of 0.6002, demonstrating its capacity to reduce overfitting and improve predictive accuracy.

Lasso Regression:
- MSE: 0.6135
- R2 Score: 0.5318

While delivering adequate predictions, Lasso Regression had a higher MSE of 0.6135 and an R2 Score of 0.5318 when compared to linear models.

Random Forest Regression (GridSearchCV):

- MSE: 0.2545
- R2 Score: 0.8058

The Random Forest Regression model greatly outperformed other models, with a low MSE of 0.2545 and a high R2 Score of 0.8058, highlighting its robust prediction capabilities.

- The Random Forest Regression model demonstrated superior predictive performance.

With PCA:

Linear Regression with PCA:

- MSE: 1.1461
- R2 Score: 0.1254

Adding PCA to Linear Regression resulted in a significant increase in MSE (1.1461) and a decrease in R2 Score (0.1254), indicating a significant reduction in predictive accuracy.

Ridge Regression with PCA:

- MSE: 1.1461
- R2 Score: 0.1254

With an MSE of 1.1461 and an R2 Score of 0.1254, Ridge Regression with PCA performed similarly to Linear Regression with PCA.

Lasso Regression with PCA:

- MSE: 1.2512
- R2 Score: 0.0452

When compared to its non-PCA counterpart, Lasso Regression with PCA had a higher MSE of 1.2512 and a lower R2 Score of 0.0452, suggesting inferior predictive accuracy.

Random Forest Regression with PCA:

- MSE: 0.5941
- R2 Score: 0.5467

Random Forest Regression with PCA demonstrated robustness to dimensionality reduction, with an MSE of 0.5941 and an R2 Score of 0.5467.

- While Random Forest Regression with PCA maintained competitive performance, linear models faced a considerable loss in accuracy, indicating a subtle impact of PCA on model predictive capabilities.

V. Conclusion

In conclusion, this project delved into the intricate landscape of the California housing market, employing a diverse set of regression models to predict housing prices. The Random Forest Regression model proved to be the most reliable predictor, with greater accuracy and performance. Geographical representations revealed regional variances, allowing for a more detailed understanding of spatial implications on price. The use of feature engineering techniques like log transformation proved useful in improving model performance. While linear models performed well in terms of accuracy, the effect of Principal Component Analysis (PCA) on dimensionality reduction varied across models. This detailed analysis provides stakeholders with essential tools for predicting and decision-making in California's ever-changing housing market.