Jacob Burns

Andrea Gonzales

Raj Patel

Cole Smith

Derrick Subnaik

**Numerical Methods Project**

# **Problem Statement**

Our group’s goal is to develop a predictive model for California housing prices by estimating the median house value of each census tract using both traditional socio-economic indicators and novel geographic features. Specifically, we will compare two dataset versions: the classic California Housing dataset, which includes attributes such as longitude, latitude, housing age, room and bedroom counts, population, households, median income, and a categorical ocean-proximity indicator; and an enriched version that replaces the proximity category with continuous distance measures to the coast and major urban centers (Los Angeles, San Diego, San Jose, and San Francisco). By applying ordinary least squares regression to each version, evaluating and reporting prediction errors, and analyzing the resulting coefficient magnitudes, we aim to determine how engineered distance features affect model accuracy and to identify which combination of variables most strongly predicts regional housing values.

**Description of Datasets**

For our analysis, we utilize the two California Housing Datasets which we pulled from Kaggle, we have chosen to use the 1990 U.S. census. These datasets contain information on housing blocks from various districts in California and serve as a valuable resource for understanding housing price determinants in the state. The datasets consist of 20,640 observations, with each observation representing a block group (the smallest geographical unit for which the U.S. Census Bureau publishes sample data).  
First and Second Dataset Features:

* **longitude**: Geographic coordinate (negative values, float)
* **latitude**: Geographic coordinate (float)
* **housing\_median\_age**: Median age of houses in the block (numeric)
* **total\_rooms**: Total number of rooms in the block (numeric)
* **total\_bedrooms**: Total number of bedrooms in the block (numeric)
* **population**: Total population in the block (numeric)
* **households**: Number of households in the block (numeric)
* **median\_income**: Median income of households in the block (numeric, scaled)
* **median\_house\_value**: Median house value for houses in the block (numeric) - our target variable
* **ocean\_proximity/Distance\_to\_coast**: Categorical variable indicating proximity to the ocean (categorical: 'NEAR BAY', 'NEAR OCEAN', '<1H OCEAN', 'INLAND', 'ISLAND'). While the Distance\_to\_coast variable is a float that measures distance in meters.

Exclusive to Second Dataset:

* **Distance\_to\_LA:** The distance from the property to LA in meters (float).
* **Distance\_to\_SanDiego:** The distance from the property to San Diego in meters (float).
* **Distance\_to\_SanJose:** The distance from the property to San Jos in meters (float).
* **Distance\_to\_SanFrancisco:** The distance from the property to San Francisco in meters(float).

The datasets provide a comprehensive snapshot of housing characteristics across different regions of California, including urban centers, suburban areas, and coastal communities. This diversity allows us to examine how various factors, including geographical location, housing attributes, population density, and economic indicators, influence housing prices throughout the state.

Some notable characteristics of the datasets include:

* No missing values except in the total\_bedrooms column.
* Natural geographical clustering of data points based on population centers.
* Significant variations in housing prices across different regions.
* Strong correlation between certain features such as total\_rooms and households.

These datasets enables’ us to develop and evaluate our predictive models for California housing prices while gaining insights into the factors that drive real estate valuation in one of America's most diverse housing markets.

# **Five Data Manipulations**

Data Manipulation 1: What we did for our first data manipulation was feature scaling which standardizes all features to have mean = 0 and standard deviation = 1.

Data Manipulation 2: What we did for our second data manipulation was feature ratios standardize all features to have mean = 0 and standard deviation = 1.

Data Manipulation 3: What we did for our third data manipulation was location-based features, which uses rooms per household, bedrooms per household, and population per household.

Data Manipulation 4: What we did for our fourth data manipulation was binning median age.

Data Manipulation 5: What we did for our fifth data manipulation was Polynomial Features for Income.

Data Manipulation 6: What we did for our sixth data manipulation was Geographic Clustering

Data Manipulation 7: The last manipulation was combing all of the features All Features Combined

# **Error Report**

Manip=Manipulation

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset 1 | Manip.1 | Manip.2 | Manip.3 | Manip.4 | Manip.5 | Manip.6 | Manip.7 |
| Training MSE | 4680954167.59 | 4560438231.38 | 4474899208.12 | 4663394918.68 | 4453844618.77 | 4536190444.78 | 4676276461.99 |
| Training RMSE | 68417.50 | 67531.02 | 66894.69 | 68289.05 | 66737.13 | 67351.25 | 68383.31 |
| Training R^2 | 0.6498 | 0.6588 | 0.6652 | 0.6511 | 0.6668 | 0.6607 | 0.6502 |
| Testing MSE | 4809839606.90 | 4768269792.55 | 4692059965.52 | 4801222542.96 | 4664873321.17 | 4662883069.00 | 4810897773.48 |
| Testing RMSE | 69353.01 | 69052.66 | 68498.61 | 69290.85 | 68299.88 | 68285.31 | 69360.64 |
| Testing R^2 | 0.6330 | 0.6361 | 0.6419 | 0.6336 | 0.6440 | 0.6442 | 0.6329 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset 2 | Manip.1 | Manip.2 | Manip.3 | Manip.4 | Manip.5 | Manip.6 | Manip.7 |
| Training MSE | 4683203783.50 | 4568841694.43 | 4638333374.45 | 4668001368.03 | 4460202162.73 | 4532855082.31 | 4193515667.58 |
| Training RMSE | 68433.93 | 67593.21 | 68105.31 | 68322.77 | 66784.74 | 67326.48 | 64757.36 |
| Training R^2 | 0.6497 | 0.6582 | 0.6530 | 0.6508 | 0.6663 | 0.6609 | 0.6863 |
| Testing MSE | 4908476721.16 | 5280716470.09 | 4831461570.55 | 4898453179.75 | 4812589380.30 | 4699077508.39 | 4821986474.01 |
| Testing RMSE | 70060.52 | 72668.54 | 69508.72 | 69988.95 | 69372.83 | 68549.82 | 69440.52 |
| Testing R^2 | 0.6254 | 0.5970 | 0.6313 | 0.6262 | 0.6327 | 0.6414 | 0.6320 |

# **Comparison of Performance**

Dataset 1:

A graph with blue lines

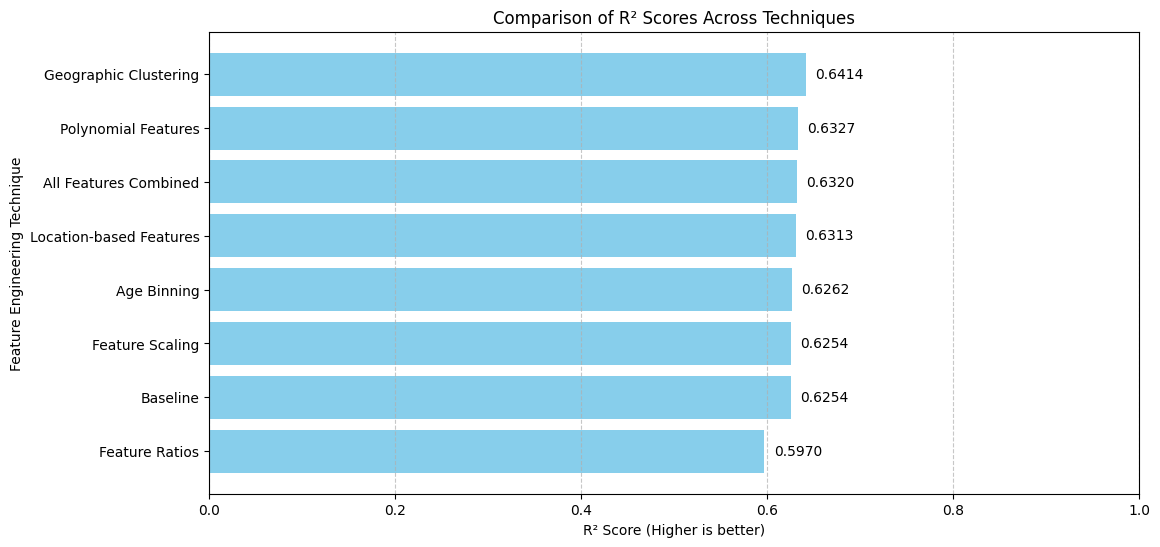
AI-generated content may be incorrect.

The R^2 value or the coefficient of determination represents the proportion of the variation in the dependent variable that is predicted from the independent variable. In this case for all of our data manipulation methods, it tells us how well the linear regression model (least squares method) explains the variability of the target variable.  
  
  
A graph of a comparison of rse across techniques

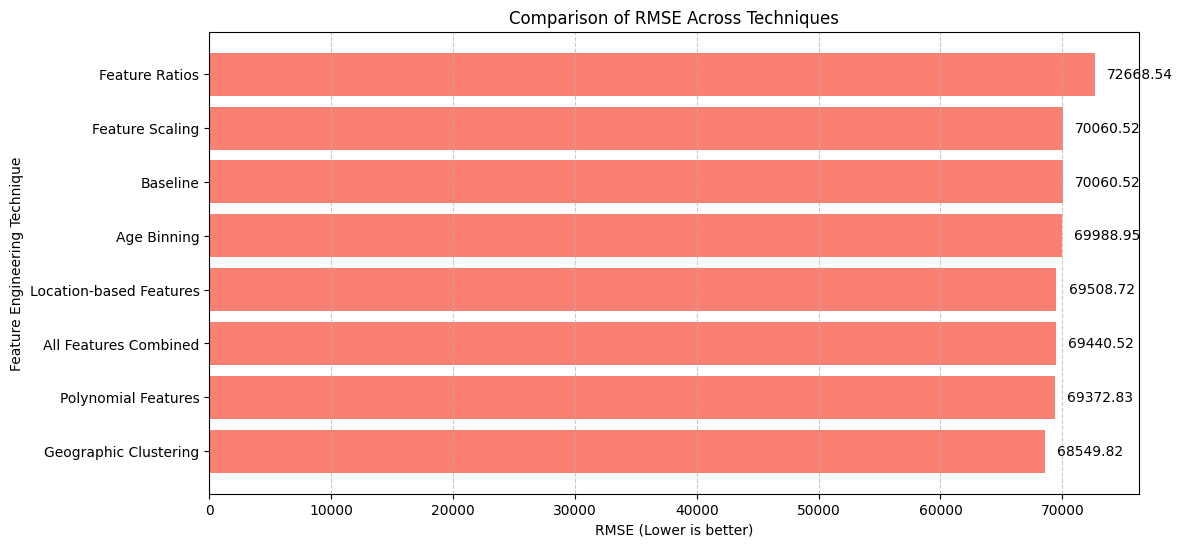
AI-generated content may be incorrect.

The RMSE value or the root mean squared error represents the average magnitude of the prediction errors, or how far off our models’ predictions are from the actual target values. The lower the RMSE the closer our models’ predictions are to the actual values, while the higher they are means that there is more deviation making it have worse performance.  
  
In this case we can see that from the bar graph above that see that the best performing RMSE would be from the geographic clustering manipulation, which means that this manipulations prediction is the closest to the actual values and has the best performance, while the worse performance manipulation model would be all of the features combined. But if we were targeting a specific manipulation the worst would be feature scaling.

Dataset 2:



When analyzing the above graph for the second data set that we ran the data on, the R^2 gives us a good idea of how the Linear Regression (least squares method). When analyzing this data, the higher is better which leads us to Geographic Clustering method was the most impactful in predicting the values. On the flip side Feature ratios for this data set is the one that worked the worst. All of the data manipulation resulted in close results.



When analyzing the RMSE for our second data set, as stated before, we look at the lower number for better results. In our analysis of our second data set we had our geographic clustering as the best result of our data set or the prediction that was closest to the actual number. In our second data set the Feature Ratios was the manipulation that worked the worst in terms of RMSE. This lines up well with the R^2 data and the numbers that we also got there.

# **Analysis of Model Weights**

Dataset 1:

When analyzing the model weights, we can use the R^2 variable for each manipulation model used which will show us how well the linear regression model explains variability. The best performing manipulation given the R^2 values is geographic clustering. This means that out of all manipulation techniques geographic clustering helped the most in following the linear regression model that we used. At the same time, we can see that our last manipulation method of all the features combined was the least successful with an R^2 value of 0.6329. If we were targeting a specific model the worst would be feature scaling with an R^2 of 0.6330.

Dataset 2:

When analyzing the weights of the second data set, we looked at the R^2 value. As stated, when we talked about the first data set the linear regression model explains this variability. In our second data this application gave us that the Geographic Clustering method of data manipulation was higher, allowing the weight for this manipulation to be the most impactful. When we applied our feature ratios, we saw that this one R^2 value performed the worst and was the least impactful on model weights.

# **Additional Points**

We have tested Linear regression model on the California housing dataset, and the approximate R2 value we got is 0.62 which is pretty low. This is because of couple reasons, first linear Regression (LR) assumes linear relationship between features and target. Second, LR treats features independently unless interaction terms are manually created. Third, LR struggles with complex spatial patterns in California housing. Lastly, LR is highly sensitive to extreme property values.

Due to this reason we have also include a second machine learning model using XGBoost and Tune XGBoost for prediction prices accurately. This two model gives R2 values of approximate 0.85 (for default XGBoost model) and 0.90 (for Tuned XGBoost model). XGBoost is better choice for this type of dataset because of following reasons, first XGBoost automatically models complex non-linear patters though decision trees. Second, it naturally discover and models interactions between features. Third, it can partition data into region with similar characteristics. Lastly it is more robust due to its ensemble natures.