An Empirical Analysis of California's 1990 Housing Market - 220138866

1. Introduction & Data Preparation:

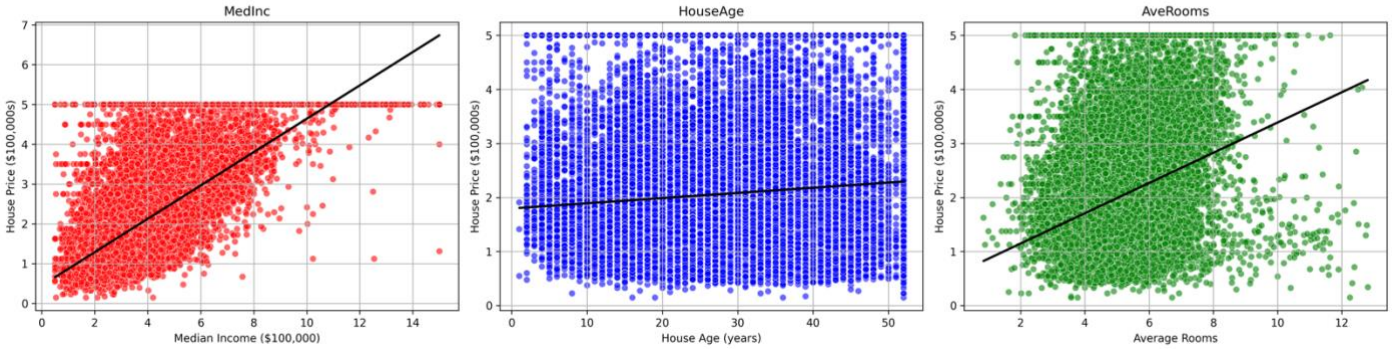
The housing market plays a crucial role in understanding regional economic inequality, household welfare and long-term demographic trends. California, in particular, has long been one of the most economically and geographically diverse states in the United States, making its housing market a valuable case for analysis. This report presents an analysis of the California Housing dataset, accessible through the Scikit-learn library, which contains median house values for districts across California based on the 1990 U.S. census data. The aim of this report is to analyse the dataset to understand the factors influencing median house values, clean the data by cleaning of unrealistic entries and outliers, engineer new features to enhance predictive capability, build and evaluate linear and regularised regression models, and identify the best-performing model and the most influential features in predicting housing prices.

To begin this report, it is important to state the features within the dataset – these are: MedInc (median income), HouseAge (median house age), AveRooms (average number of rooms per household), AveBedrms (average number of bedrooms per household), Population (block group population), AveOccup (average number of household members), Latitude, Longitude. The dataset comprises of 20,640 entries whereby descriptive analysis shows substantial variation across features such as income levels, population density and housing characteristics. The median house value, which serves as the target feature in this report, ranges from approximately $14,999 to $500,001 with an average of around $206,855. This wide range demonstrates the significant housing price disparities across California's districts in 1990. After examining the dataset, it was clear that the dataset did not have any missing values, however, it was very obvious that certain features did contain very unrealistic and unexplainable entries that were identified as potential outliers – they were: AveRooms, AveBedrms and AveOccup. To add to that, the distributions of these features are positively skewed with a few extremely high entries. Prior to any data cleaning, it is worth mentioning that three separate methods were considered to identify and clean of outliers such as using the 1.5*IQR rule (Tukey's fences), three sigma-rule or economic intuition.

Firstly, outliers in the AveRooms feature were identified and removed using the three sigma-rule as its distribution was approximately normally distributed. Entries beyond three standard deviations are extremely unlikely in this case and can disproportionately affect the model's performance, this led to 133 entries to be removed. As a consequence, the outliers within the AveBedrms feature were also removed as these outliers corresponded to the ones found in AveRooms. Moreover, outliers in the AveOccup feature were identified and removed using economic intuition rather than the three-sigma rule or Tukey's fences as both methods are heavily dependent on the mean of the dataset which was heavily skewed due to numerous outstanding outliers which would've eliminated around 700 entries – most of which would've been genuinely reasonable entries. While there is no strict threshold for household size, using economic intuition and contextual reasoning suggest that average household sizes greater than 15 are implausible. Although certain neighborhoods may experience higher levels of overcrowding or shared living arrangements, such extreme entries are unlikely and were removed – this removed only 19 entries.

To identify which factors most strongly influence housing prices, pairwise correlations between three features and the median house value were examined – these were: MedInc, HouseAge, AveRooms. The regression results revealed that median income had fit the data best and had a significantly higher correlation with housing prices than the other two features, indicating that higher income areas generally have higher property values. Visualising the relationship between HouseAge and AveRooms with housing prices shows no clear trend and the analysis corroborates this with negligible correlation coefficients.

Figure 1: Relationship between Features and House Price



2. Feature Engineering & Models:

Feature engineering is a process that transforms original features within a dataset into new meaningful and machine-readable features to enhance the performance of machine learning models. These transformations aim to improve the predictive capabilities to better represent underlying economic, socioeconomic and demographic factors influencing housing prices. Intuitively, housing prices especially in a relatively affluent state such as California would see a pattern where properties in more habitable, lavish and comfortable metropolitan cities to be more expensive such as Los Angeles or San Francisco. Furthermore, it is also expected that properties along hot-and-coming coastal regions within these big cities are to be valued more than properties situated in the outskirts as visualised in Image 1 below. Hence, it can be concluded that geographical location plays a crucial role in determining property values. To capture these spatial patterns, three features were engineered: $Latitude^2$, $Longitude^2$ and the interaction between them $Latitude \times Longitude$. Spatial patterns describe how housing prices vary across geographic location reflecting differences in property valuation between coastal and city properties. These new features allow the model to study and account for any nonlinear and interactive geographical effects on housing prices.

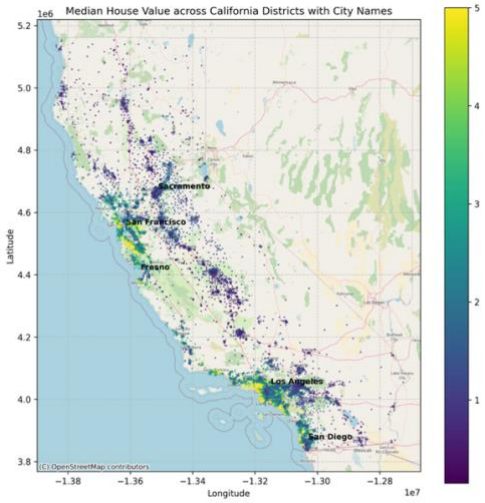


Image 1: Geographic distribution of median house values in California

Prior to running any regression analyses with these new features, the dataset had been randomly allocated to one of two groups: 80% of the data was allocated to the training group to fit and learn the existing data, the remaining 20% was allocated to the testing group to test its predictive accuracy on unseen data. This split ensures that the model is trained on a sufficient pool of the data while reserving a subset for evaluating its generalisation performance.

The regression analyses begin by evaluating all original features within the dataset against the target feature, median house values. The fitted model explains approximately 65.2% ($R^2 = 0.652$) of the variance within the training group and 66.9% ($R^2 = 0.669$) of the variance within the testing group with corresponding RMSE values of 0.681 and 0.667 in hundreds of thousands of dollars. This implies the predictions deviated by roughly $66,700 to &68,100 compared to the actual values within the two groups respectively. This indicates a reasonably good fit without the use of the engineered features thus far. This model will be called “Model 1” for simplification purposes when distinguishing between different models.

Model 2 represents the regression with the original features within the dataset plus the engineered features: $Latitude^2$, $Longitude^2$ and the interaction between them $Latitude \times Longitude$. The purpose of this regression is to observe whether the engineered features can better fit and predict the data better with the target feature. This fitted model explains approximately 66.0% ($R^2 = 0.660$) of the variance within the training group and 67.6% ($R^2 = 0.676$) of the variance within the testing group with corresponding RSME values of 0.673 and 0.659 (in hundreds of thousands of dollars). This implies that the predictions deviated by roughly $67,300 to &65,900 compared to the actual values within the two groups respectively. Therefore, given the results, it can be deduced that this model with the engineered features had fit the data slightly better with a higher training and testing R^2. Surprisingly, the training RMSE is higher in model 2 than in model 1 while the testing RMSE is lower in model 2 than in model. It was expected that given the addition of the engineered features for R^2 to rise (which has been realised) and RMSE within both the training and testing sets to fall (partially realised), this pattern may not continue as we introduce more models in our analyses.

Ridge regression is a statistical method to prevent overfitting and multicollinearity in a linear regression by adding a penalty term. This method is particularly useful as it reduces the predictor features toward zero, reducing variances and leading to more stable and reliable predictions. Prior to analysing our linear models using Ridge regression and later Lasso regression, the training and testing groups must be scaled or standardized. This means that the features are all in the same numerical scale with $\mu = 0 \; and \; \sigma^2 = 1$. This ensures fair regularisation across these features. If these features are not scaled and one or multiple features is measured in much larger numbers than the others then the penalty term becomes unfairly unbiased toward features with a smaller magnitude.

Two Ridge models were estimated these were: Model 3 (original features) and Model 4 (original +engineered features). Both models were evaluated across multiple α levels [0.1, 1, 10, 100, 1,000] to examine how different levels of regularisation affected performance. For both models, α = 0.1, yielded the best-balanced performance achieving the lowest RSME and highest R^2 value. The Ridge model using only the original features (Model 3) explained 65.2% of the variance and 66.9% of the variance within the training and testing groups respectively with a corresponding RMSE values of 0.681 and 0.667 (in hundreds of thousands of dollars). When the engineered spatial features were introduced (Model 4) the model achieved R^2 scores of 65.7% and 67.2%, with RMSE values of 0.676 and 0.664 within the training and testing groups respectively. This demonstrates that the inclusion of the engineered features marginally improved model predictive performance. It is worth mentioning that as α increases, there is a persistent downward R^2 and upward RMSE trend. Moreover, both models were evaluated across multiple α levels [0.1, 1, 10, 100, 1,000] using 5-fold cross-validation to identify the optimal level of regularisation. The analyses showed that α = 0.1 produced the highest R^2 score and was therefore selected as the best parameter for the final models.

Lasso regression is another method used for regularisation and feature selection. It adds a penalty to the model based on the absolute value of the coefficients which shrinks lesser important features to zero, thereby performing feature selection. This makes it very useful to identify the most influential features that affect housing prices.

Similar to the previous analysis, two Lasso models were evaluated: Model 5 (original features) and Model 6 (original + engineered features). Model 5 explained 53.3% of the variance and 54.8% of the variance within the training and testing groups respectively with corresponding RMSE values of 0.788 and 0.780 (in hundreds of thousands of dollars). When the engineered spatial features were introduced (Model 6) the model achieved R^2 scores of 53.4% and 54.8%, with RMSE values of 0.788 and 0.779 within the training and testing groups respectively. Similar to Ridge regression, both models were evaluated across multiple α levels [0.1, 1, 10, 100, 1,000] and found α = 0.1 to be the best parameter for the final models. This suggests that Lasso regression actually did worse than Linear and Ridge regression as it had eliminated several features that had affected regression performance and results.

Model	Training RMSE ($100k)	Testing RMSE ($100k)	Training R-squared	Testing R-squared
1	0.6811	0.6672	0.6516	0.6687
2	0.6726	0.6592	0.6603	0.6766
3	0.6811	0.6672	0.6516	0.6687
4	0.6762	0.6637	0.6566	0.6722
5	0.7882	0.7797	0.5334	0.5475
6	0.7878	0.7794	0.5339	0.5479

Table 1: Comparison of all models against respective RSME and R^2 scores.

Based on the results presented in Table 1, the inclusion of the engineered spatial features did help as it produced lower RMSE values as well as higher R^2 scores across Linear, Ridge and Lasso regressions – Models 2, 4, 6 – compared to the models consisting of just the original features. Although the improvement in output is negligible, in some cases the difference is less than a percent, it still fitted the data better by explaining the variation more and the predictions were closer to the actual values.

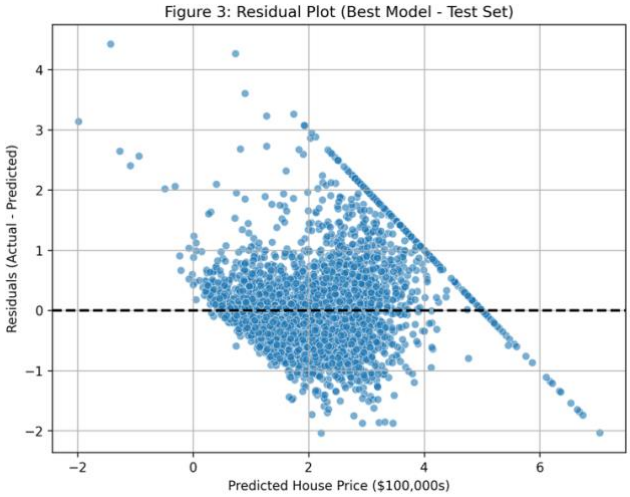
While the engineered spatial features improved predictive accuracy in Linear and Ridge models, Lasso eliminated them suggesting that their contribution was not substantial enough once the weaker predictors were penalised. This suggests that the engineered features are largely rooted in other features such as median income.

3. Model Analysis

This section of the report evaluates the interpretability and performance of the models developed in the previous section. The analysis focuses on the best-performing model, comparing the predicted versus actual housing prices and examining residual patterns to evaluate model fit. This discussion also identifies the most influential predictors based on coefficient magnitude and evaluates the effects of regularisation methods, economic relevance and model limitations.

The scatter plot in Figure 2 shows a clear positive trend, indicating that the model successfully captures the overall relationship between predicted and observed prices. However, some dispersion is visible at the higher end of the price range where the model tends to slightly underestimate expensive properties, and at the lower end, where it marginally overestimates cheaper houses. This pattern is typical of linear models and reflects moderate predictive error but an overall consistent trend.

Figure 3 visualises the difference between the model's predicted and actual values (i.e. the residuals) against the predicted values themselves. It is used to assess whether the model has homoskedasticity: constant variance of errors. Ideally, the residuals should be randomly scattered around zero without a clear pattern, indicating that the model captures the systematic structure in the data and that the errors are evenly distributed across the predicted house values. However, as shown in the plot, the residuals are not scattered evenly around zero, hence, indicating that the model does not fully satisfy the assumption of homoskedasticity.



Heteroskedasticity is an issue because it violates a key assumption in linear regression that the variance of the error terms remains constant. If violated, the model's standard errors become unreliable, leading to biased statistical inference such as in hypothesis testing and calculating confidence intervals. To corroborate this result, the Breusch-Pagan test was conducted to see whether this model is statistically proven to be heteroskedastic, and it found sufficient evidence to suggest presence of heteroskedasticity in the residuals. A potential solution to this issue would be to take a transformation of the model such as the logarithm of the model, however, the Breusch-Pagan test found presence of heteroskedasticity within this model as well. This may be worth investigating further but this is beyond the scope of this report.

Feature	Coefficient	Coefficient Magnitude
Latitude	14.34	14.34
Longitude	11.64	11.64
AveBedrms	0.79	0.79
MedInc	0.42	0.42
AveOccup	-0.3	0.3

Table 2: Coefficient magnitude

Table 2 presents the top five features ranked by coefficient magnitude in the best-performing model. Latitude and Longitude exhibit the largest coefficient indicating that geographic location remains the largest predictor of housing prices. This is followed by AveBedrms and MedInc suggesting that higher household income levels as well as more bedroom count are associated with more expensive properties. On the other hand, AveOccup , with a negative coefficient, implies that houses with more occupants tend to be located in cheaper neighborhoods.

Comparing the performance of Ridge and Lasso regression, Ridge achieved slightly higher R^2 scores and lower RMSE across both training and testing sets as shown in Table 1. Model 4 – original and engineered spatial features – performed the best amongst all the models under regularisation methods. This model was regressed under Ridge with an α = 0.1 suggesting that the Linear model of these features needed minimal penalisation to obtain optimal performance. This outcome indicates that most predictor features already contributed meaningful explanatory power and the model benefitted from Ridge through reducing overfitting rather than drastically altering feature weights. In contrast, the Lasso models applied a stronger penalty that eliminated several relevant predictor features which weakened its explanatory ability and lowered predictive accuracy.

Comparing feature magnitude in Table 2, the geographical features Latitude and Longitude as well as the average number of bedrooms per household were the biggest factors affecting housing prices. This supports economic and real estate market intuition where it is expected that more wealthy neighborhoods in big metropolitan cities such as Long Beach in Los Angeles possess more expensive properties. Additionally, the average number of bedrooms per household reflects both the size and quality of the property, which are strong determinants of value. Larger homes with more bedrooms typically command higher prices, as they indicate greater living space and are often located in higher-income areas.

In the previous section, heteroskedasticity was identified as a key limitation of this model. As it has been forementioned, this will be a brief summary. If it is observed that our error terms do not have constant variance across all levels of predicted values, the standard errors become unreliable and this sabotages our results in hypotheses testing as the test statistics calculated would be completely redundant and harmful for statistical inference as they would increase the chances of a Type I or II error leading to misleading statistical interpretation.

As shown in Figure 2, the model performs better on affordable and mid-range priced houses than on expensive properties. The residuals grew larger and became more dispersed for higher-priced houses, indicating that the model tends to underestimate their value. This may be due to the factor that houses in this bracket may be influenced by other factors not captured in the features available or need new features to capture the effects of proximity to amenities, tourist attractions, public transport etc. Consequently, while the model generalises well for the majority of properties, it struggles to account for the variability and price extremes of luxury housing markets.

4. Conclusion

To sum up the report, the model which best represented the dataset and had the best predictive performance on unseen data was Model 2 – the Linear regression with our original features plus the engineered spatial features. It explained 66% and 67.7% of the variance within the training and testing groups respectively, whereby, this model achieved the highest R^2 scores among all models. Furthermore, Model 2 also achieved the lowest RMSE values across all models with 0.673 and 0.659 (in hundreds of thousands of dollars) within the training and testing groups respectively. This signifies that on average our model's predictions on housing prices deviated by $67,300 and $65,900 from the actual housing prices across the two groups.

Based on model outputs, the three most important and influential features for predicting housing prices are: Latitude, Longitude and AveBedrms. These features keep appearing as the strongest predictors across all model specifications, confirming their importance in explaining price variation. Their influence has been discussed in the previous section and their consistent performance across models highlights their reliability as key factors of housing value.

The main limitation for the linear model is the presence of heteroskedasticity, where the variance of the residuals increases with higher predicted house prices. As previously discussed, this violates one of the key assumptions of linear regression and results in unreliable standard errors, reducing the validity of statistical inference. To combat this, a logarithmic transformation was performed, however, this did not remove the trend of heteroskedasticity. Therefore, more transformations or solutions should be presented to eliminate such an issue, however, this is not an objective of this report and goes beyond the scope of the report.