Linear Regression Model Performance Report- Team 1

1. Kaggle House Prices Dataset- Ames, Iowa,USA

Overview :

This dataset gave us a rich collection of housing features to analyze, including things like overall quality ratings, living area square footage, garage capacity, and the year each home was built. Our main goal was to predict the sale price. We spent time cleaning up the data by filling in missing values, converting categorical information into numbers the model could understand, and scaling everything to a common range. We also identified and handled outliers that might have thrown off our predictions.

Correlation Insights:

When we looked at which features had the strongest relationship with sale price, we found ten clear winners: Overall Quality came out on top with a correlation of 0.80, followed by Living Area at 0.70, and features like Garage Cars, Garage Area, and Total Basement Square Footage all clustering around 0.63-0.64. Rounding out the top ten were First Floor Square Footage, Year Built, Number of Full Bathrooms, Year of Remodel, and Masonry Veneer Area.

Model Performance:

Metric	**Training**	**Testing**
MSE	0.010	0.02
RMSE	$0.11	$0.14
R^2	0.9286	0.8953

Interpretation:

1. Our model did remarkably well, capturing about **90% of what drives house price variation**. After applying a logarithmic transformation to handle the wide range of prices, we ended up with very small prediction errors.

2. The difference between training and testing performance (only about 3%) tells us the model isn't just memorizing the training data—it's actually learning patterns that hold up with new homes.
3. The features pulling the most weight were Living Area (pushing prices up significantly), Kitchen Above Grade (surprisingly pulling prices down), and Street type. For a straightforward linear approach, these results are honestly quite impressive. The model is easy to interpret and doesn't show signs of overfitting.

2. California Housing Dataset

Overview :

This dataset gave us a different perspective, focusing on district-level housing characteristics across California. We had information about median income in each area, average number of rooms and bedrooms, population density, how old the houses were, and geographic coordinates. The nice thing here was that the data was already pretty clean—no missing values to worry about—so we could jump straight into scaling features and building the model.

Model Performance

Metric	Training	Testing
MSE	0.5179	0.5559
RMSE	$0.72	$0.75
MAE	$0.53	$0.53
R²	0.6126	0.5758

Interpretation:

This model captured about **58% of the variation** in house prices—not as stellar as the Kaggle dataset, but still respectable. Our typical prediction error came in around $75,000, which is moderate given California's housing market. The silver lining is that training and testing scores stayed pretty close together, meaning the model generalizes reasonably well to new districts.

Median income turned out to be the strongest predictor (not surprising), while average bedrooms and house age also played important roles. The moderate performance suggests there are probably some non-linear patterns in the data that a simple linear model can't quite capture. This points toward trying more sophisticated approaches like Random Forests or Gradient Boosting in the future.

3. Overall analysis

1. Looking at both projects together, the Kaggle model clearly outperformed with its 90% R^2 score, largely thanks to having more detailed features and some solid preprocessing work. The California model, while more modest at 58% R^2, still did a decent job and proved it could handle new data without falling apart.
2. Both analyses confirmed what real estate agents have known forever: income levels, living space, and build quality are huge drivers of home prices.