1. Introduction

This report summarizes a regression analysis on the **California Housing** dataset. The goal was to predict the **median house value** (`MedHouseValue`) using various features such as **median income**, **house age**, **average rooms**, **latitude**, **longitude**, etc. We employed **XGBoost** with **polynomial features** and **hyperparameter tuning** to improve predictive performance.

2. Data Exploration & Correlation Heatmap

1. **Correlation Heatmap**
    - The heatmap indicates **`MedInc` (Median Income)** is most strongly correlated with `MedHouseValue`.
    - Other features like `HouseAge`, `AveRooms`, and `AveBedrms` show moderate positive correlations.
    - **Latitude** and **Longitude** exhibit negative or moderate correlations, indicating location significantly influences housing prices (coastal areas often have higher values).

2. **Key Observations**
    - **Median Income** stands out as the single most important predictor.
    - **HouseAge** has a smaller positive relationship with house value.
    - **Population** and **AveOccup** are less correlated, though they can still be relevant in a non-linear model.

3. Modeling Approach

1. **Feature Engineering**
    - We introduced **polynomial features** (degree=2) for high-impact variables (e.g., `MedInc` and `AveRooms`) to capture non-linear relationships.
    - This added new terms like `MedInc^2` and `MedInc * AveRooms`.
2. **Model & Hyperparameters**
    - We used an **XGBoost** regressor with the objective set to `"reg:squarederror"`.
    - **RandomizedSearchCV** (with a small search space) optimized hyperparameters such as `n_estimators`, `learning_rate`, `max_depth`, etc.
3. **Train/Test Split**
    - An **80/20 split** was used, ensuring the model was trained on the majority of data and tested on unseen data for performance metrics.

4. Results & Performance

From the console output:

- **MAE (Mean Absolute Error)**: ~0.31
    - On average, predictions are off by about $31k, since the target is in hundreds of thousands of dollars.
- **MSE (Mean Squared Error)**: ~0.22
    - Squared errors penalize large deviations more heavily; ~0.22 in the scale of "hundreds of thousands" can still be reasonable.
- **R^2**: ~0.8319
    - The model explains about **83%** of the variance in median house values, which is a strong improvement over simpler baselines.

These results indicate that polynomial features plus XGBoost capture much of the complexity in California's housing market.

5. Feature Importance

A bar chart of the **top 10 features** in XGBoost shows:

1. **`MedInc^2`** (polynomial term) ranks highest, emphasizing the strong non-linear effect of income on house value.
2. **`MedInc`** itself is also near the top, confirming income is critical.
3. **Cross-term** (`MedInc * AveRooms`) often appears in the top features, suggesting interactions between household income and room count matter.
4. **Location** variables (`Longitude`, `Latitude`) appear mid-range, highlighting how coastal proximity or specific geographic zones influence pricing.
5. **`AveRooms`**, `AveBedrms`, and `HouseAge` also appear, though with lower relative importance compared to income-related features.

6. Conclusions & Recommendations

1. **Income Dominates**
    - Median Income (`MedInc`) and its polynomial transformations are the strongest predictors of house value in this dataset.
2. **Non-Linear Patterns**
    - Polynomial features significantly boost performance, indicating a **non-linear** relationship between income, rooms, and housing prices.
3. **Location Matters**
    - Geographic coordinates (`Latitude` and `Longitude`) also help differentiate higher- vs. lower-value areas.
4. **Model Performance**
    - An **R^2 of ~0.8319** is quite good. Further tuning (e.g., more folds in cross-validation, expanded hyperparameter ranges, additional polynomial features) could improve results.
5. **Future Work**

- **Regularization** (e.g., L1, L2) might reduce overfitting when many polynomial terms are introduced.
- **Feature Engineering** for other interactions (e.g., `HouseAge * MedInc`) may uncover additional patterns.
- **Ensemble Approaches** (averaging multiple models) can further refine predictions.

Overall, the analysis confirms that **income, room count, and location** are primary drivers of house value in California, and that **non-linear modeling** (via XGBoost + polynomial features) significantly enhances predictive accuracy.