**California Housing Price Prediction**

Group 3: Raphael Luis Santos, Kelly Tran, Rick Liu, Huu Nhan Nguyen, Amy Ionescu

**ABSTRACT**

California housing prices have been a point of contention for the past several decades, with the state’s numbers consistently among the highest in the nation. In this paper, we tackle the factors that go into California’s housing prices, attempting to understand the median housing value based on a variety of factors and focusing on ocean proximity and medium income. Through using linear regression, k-Nearest Neighbor, and Random Forest regression, we noticed that while not linear, clear positive correlations developed between housing prices and income alongside housing prices and ocean proximity. As the housing market continues to rise above inflation, these results allow younger generations to maintain dreams of being a homeowner alive.

**Introduction**

California is among the most expensive housing markets in the country and is in the middle of a housing crisis. These circumstances can be attributed to various geographical, social, and economic factors–all of which intertwine and contribute to the rising pressures of living in California. As such, we hope to explore historical data from the U.S. Census to analyze the housing market and understand how various housing factors play a role in determining the stability of housing for California residents. To guide our research, we aim to answer three overarching questions. (1) How does ocean proximity influence the housing value? (2) In what capacity, does household income mitigate rising housing costs? (3) What implications can we draw to comment on current housing conditions? By utilizing techniques drawn from the data life cycle and machine learning, we will deepen the understanding of historical California housing prices and make implications for the current housing landscape.

**Problem Definition**

In this housing price prediction report, we seek to answer the questions of how different features impact the price of houses throughout the state of California using various data analysis techniques. Specifically, we want to know how ocean proximity influences the housing value, in what capacity, does household income mitigate rising housing costs, and what implications can we draw to comment on current housing conditions? We will attempt to answer these questions with techniques including Linear Regression, kNN Modeling, and Random Forest Analysis.

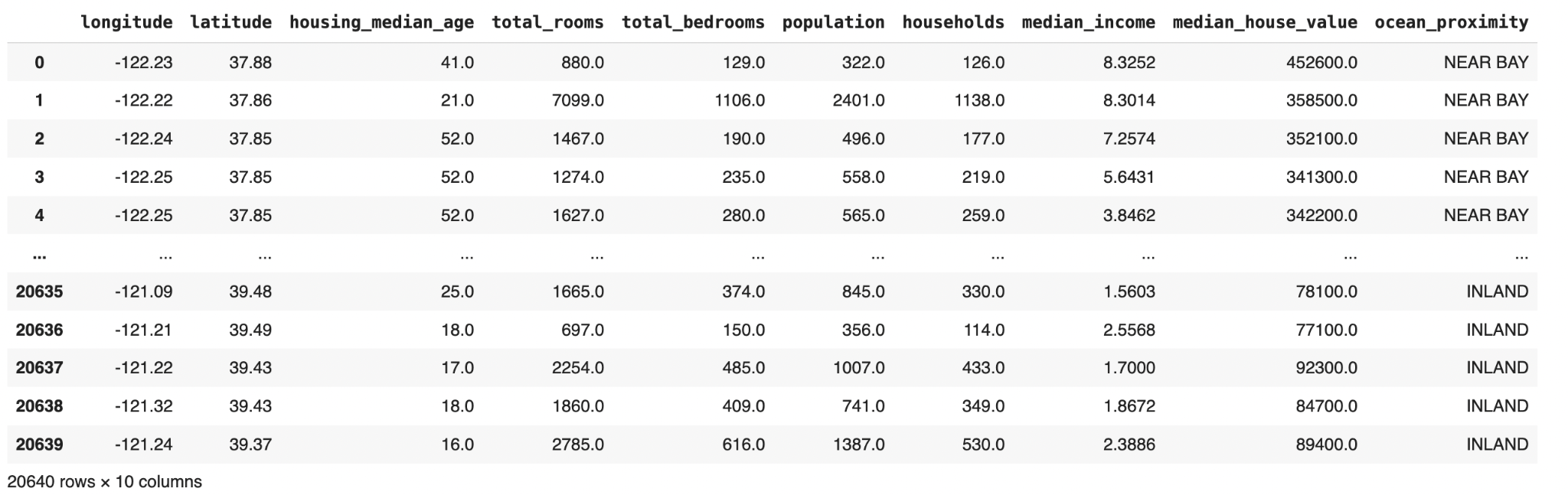
**Data Exploration and Data Preprocessing**

Overview: The California Housing Prices dataset provides a rich source of information on housing attributes and their relationship with housing prices across different California districts. The primary goal is to predict the 'median house value,' which is the target feature in this analysis.

Key Features:

* *‘longitude*’: Longitude coordinate
* *‘latitude’:* Latitude coordinate
* *‘housing\_median\_age’:* Median age of the housing units
* *‘total\_rooms’:* Total number of rooms in the house
* *‘total\_bedrooms’:* Total number of bedrooms in the house
* *‘population’:* Total population of the district
* *‘households’*: Number of households
* *‘median\_income’*: Median income of the district
* *‘median\_house\_value’:* Median house value (target feature)

**Data Exploration:**

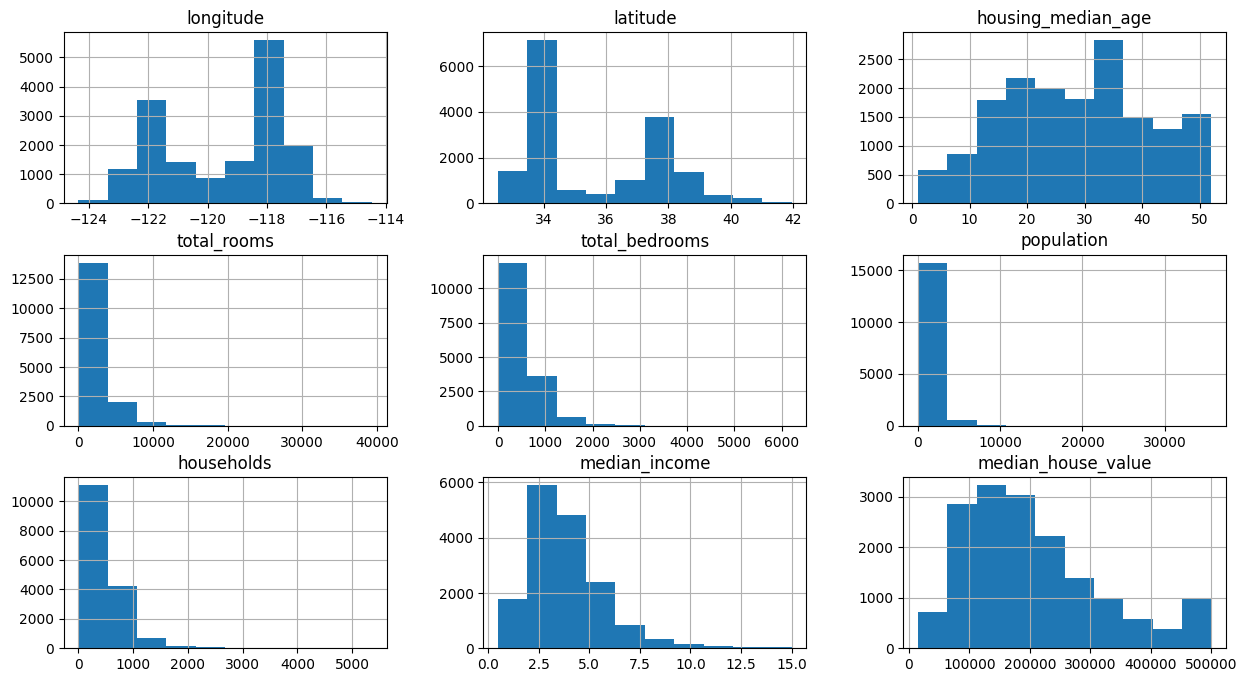


Dataset loaded from Kaggle: “California Housing Price”

Upon initial inspection, the dataset appears to be structured with rows representing districts and columns representing various features of those districts. Each feature provides a different aspect of housing and demographic information, with median\_house\_value being the target variable of interest.

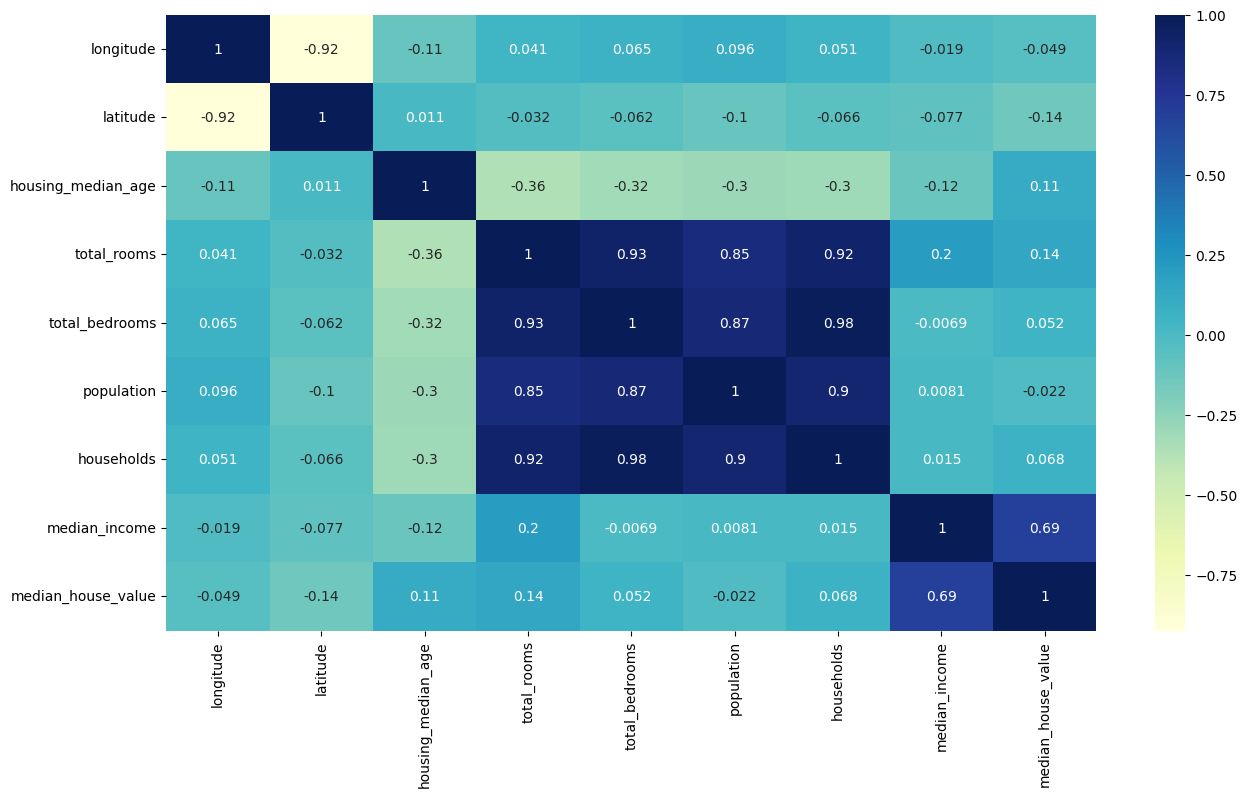
Visualization Insights:

* Histograms of features reveal the distribution and help identify any skewness or outliers. Box plots further illustrate the presence of outliers and the spread of the data.



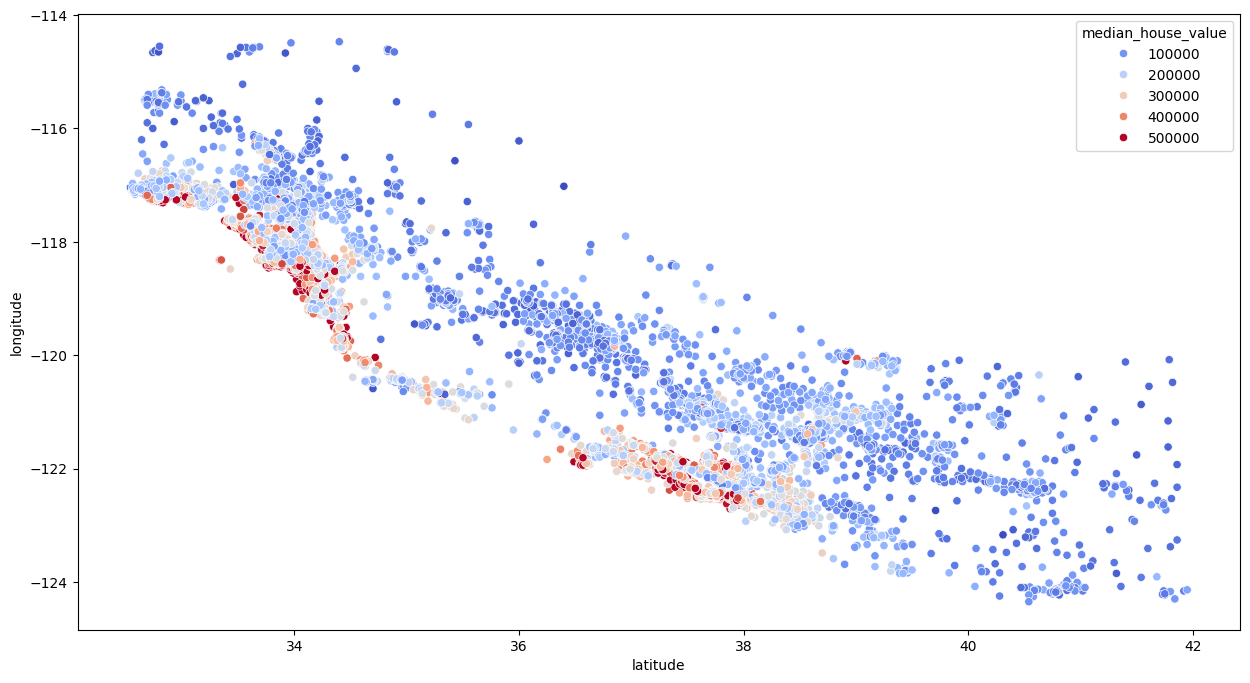
Median House Value Based on Different Features

* Correlation heatmaps provide a visual representation of feature relationships, highlighting how features are related to each other and to the target variable. This helps in understanding which features are most relevant for predicting house values.

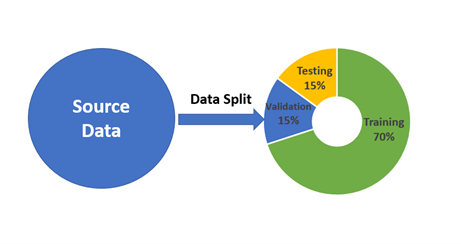


Correlation Matrix for A Better Visualization

* Scatter plots are essential for visualizing and understanding the relationship between geographic coordinates and house values. They help to uncover trends, patterns, and anomalies, providing valuable insights for further analysis and decision-making in real estate and urban planning contexts.



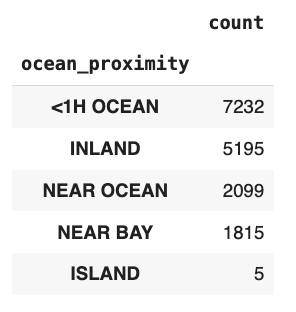
**Data Preprocessing**

****

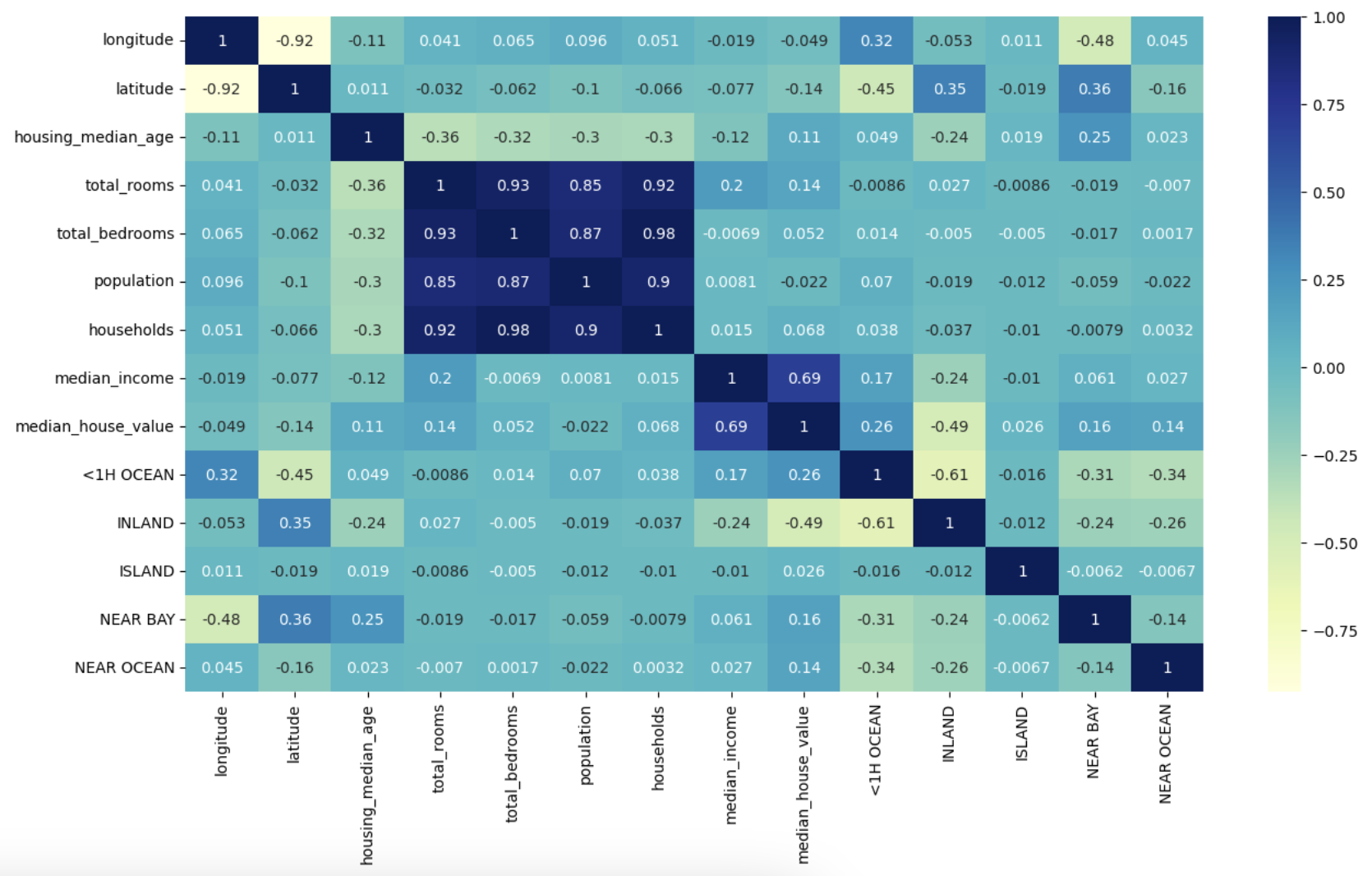
Before splitting data for training and testing purposes, making sure there are no issues in our dataset is extremely important.

Error Handling:

* Detection of Missing Data: Non-null value for all dataset in each feature: ’df.dropna(inplace=True)’to make sure that they are compatible with each other because it is crucial for accurate analysis and modeling.
* Navigate the feature which needs a ‘special’ handling (e.g. ocean proximity) ‘train\_df.ocean\_proximity.value\_counts()’due to its incompatibility with other values.



* Drop the feature that is not compatible with each other or plan to use it for a special purpose: ‘train\_df.select\_dtypes(include=[np.number])’. For instance, we drop the ‘ocean proximity’ because its value is non-numeric and can’t serve the training purpose.
* One-hot Encoded: it is hard to exclude the important feature such as ‘ocean proximity’ out of the determination for the housing price since location plays an important role. Therefore, one-hot encoding for non-numeric values into binary values would solve this problem. ‘train\_df.join(pd.get\_dummies(train\_df.ocean\_proximity)).drop(['ocean\_proximity'], axis = 1)’.This not only solve the incompatibility issue but also make conclusions more reliable.



Correlation Matrix within ‘ocean proximity’ one-hot encoded

Adding New Features:

Adding new features like *‘bedroom\_ratio’* (total bedrooms divided by total rooms) and *‘household\_rooms’* (total rooms divided by total households) can enhance data preprocessing by providing more granular insights into housing conditions.

* Bedroom\_ratio helps quantify the proportion of bedrooms relative to total rooms, which can indicate housing density and potential overcrowding.
* Household\_rooms reflects the average number of rooms per household, offering insight into living space distribution.

These features can capture more nuanced relationships with median\_house\_value, potentially improving model accuracy by incorporating relevant aspects of housing quality and usage that might not be apparent from raw counts alone.

1**Linear Regression**

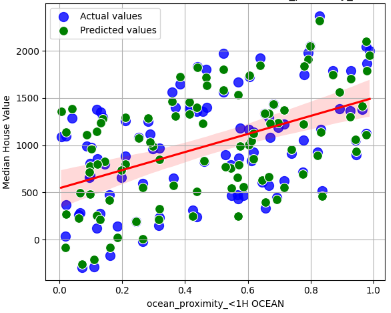
Linear Regression was the first model we used to predict housing prices. This method uses the relationship between dependent variables and independent variables to make the best prediction using a linear equation based on the independent variables. The use of Linear Regression comes with some assumptions, such as the relationship between the variables maintaining linearity. Additionally, it assumes observations are independent of each other, and that the variance of errors are constant and consistent across all levels, and that there is normal distribution.

1.1**Experiment Design and Evaluation**

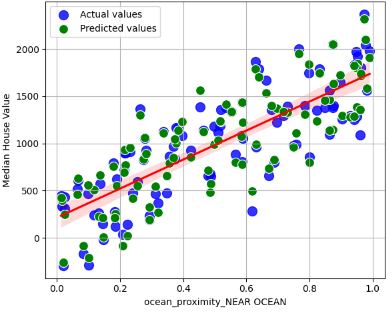
Utilizing our preprocessed data, we used sklearn’s Linear Regression to train our model. We had adjusted the test data to match the form of the training data set, which meant going through the same one-hot encoding for ocean-proximity. Once the test and training data were within the same format, we evaluated the model using R-squared. The model was then also evaluated after being scaled using sklearn’s standard scaler, to see if the score would improve in any way.

The best R-squared score for both the scaled and unscaled versions of the Linear Regression model revealed that it did not fit the dataset very well. They both landed at around -0.623, while being only minor decimal points apart. This is likely due to the variability in the models because of the way in which the vast amount of variables interacted with the resulting median house value. This would indicate the relationship between them is likely not linear.

However, the model was still able to map some relationships between individual variables that appear to have linearity. This was exemplified with the Actual and Predicted Median House Value and ocean\_proximity.



Actual Vs. Predicted Median House Value vs. ocean\_proximity <1H OCEAN



Actual vs. Predicted Median House Value vs. ocean\_proximity\_NEAR OCEAN

As seen in the graphs above, it would appear that the closer the house is to the ocean, the higher its median house value would be. But, the spread surrounding the linear regression line still indicates that large amount of variability in the dataset. With this model and prediction, we concluded that linear regression did not fit well and produced the least reliable results.

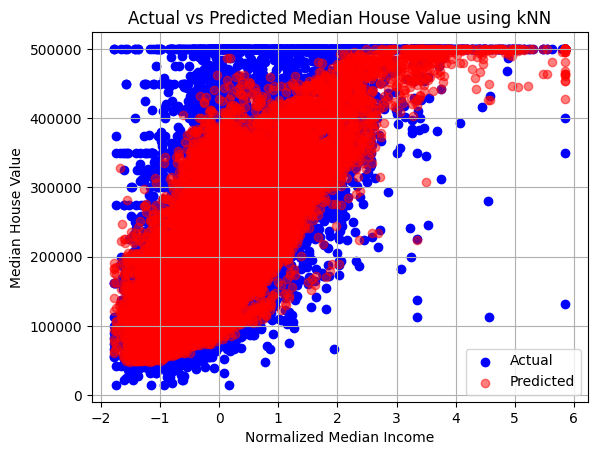
2**kNN**

k-Nearest Neighbor (kNN) is a technique that uses the proximity of points to make classifications or predictions about the grouping of an individual data point. In our implementation, we leverage the following steps to conduct kNN on our cleaned data. As an effort to prevent overfitting and improve the performance of our model, we began by running a k-fold cross validation and plotting the mean squared error (MSE) across a range of k values to select the most optimal hyperparameter. Then, using the most optimal k value, we ran the k-NN algorithm and produced the R-squared value to compare performance. For various features, these steps were repeated to assess how different housing factors could influence the median housing value across California.

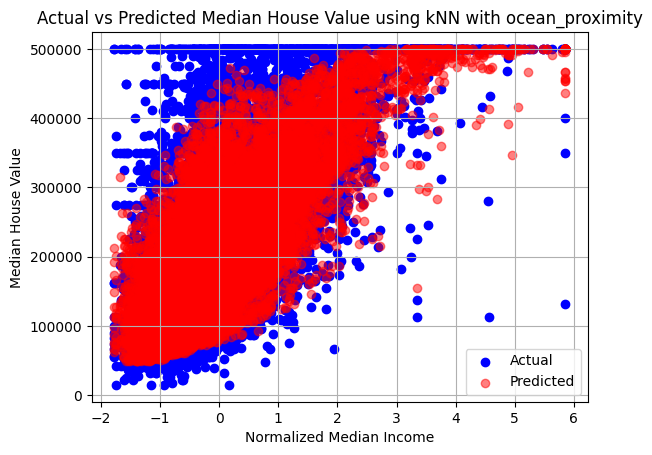
2.1**Experiment Design and Evaluation**

By using k-Nearest Neighbors regression we were able to predict housing prices (‘median\_house\_value’) using various features from the housing dataset.

We decided to test a kNN model without one-hot encoding the ‘ocean\_proximity; feature first. After cleaning the dataset and imputing a column’s mean for missing values in certain features, we then normalized the features. We tested different numbers of neighbors (k) from 1 to 15 using mean squared error (MSE) for each k to identify the optimal value, where we found that a k value of 10 yielded the lowest MSE of 3709167220.827241. This k-value was used to build a kNN model to assess the proportion of variance with the R-squared metric of 0.77697, which suggests that kNN is a decent fit for the dataset.



We then attempted to use a kNN model but this time including ‘ocean\_proximity’ by one-hot encoding the categorical values into quantitative ones. We followed the same steps as the model without the one-hot encoding and tested different numbers of k from 1 to 15 to identify the best k. We found this to be k = 12, with an associated MSE of 3561524474.537329. This k-value was then used to build another kNN model to assess how well the model fit the data, where we received an R-squared value of 0.77613, still suggesting that kNN is a good fit for the dataset.

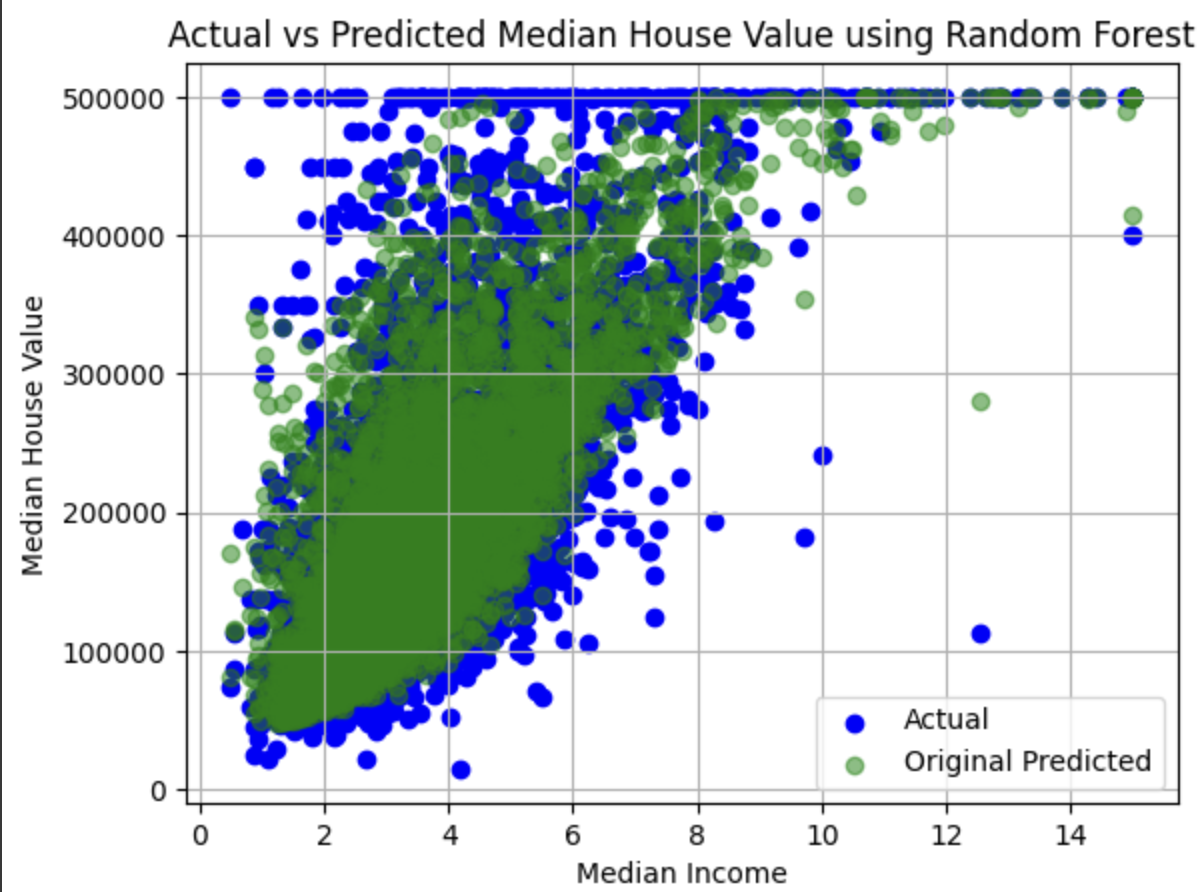
The results from the kNN models indicate that housing prices can be effectively predicted using features like ‘median\_income’ and ‘ocean\_proximity.’ The models suggest that these features significantly influence housing prices and can aid in reliable predictions.

3**Random Forest**

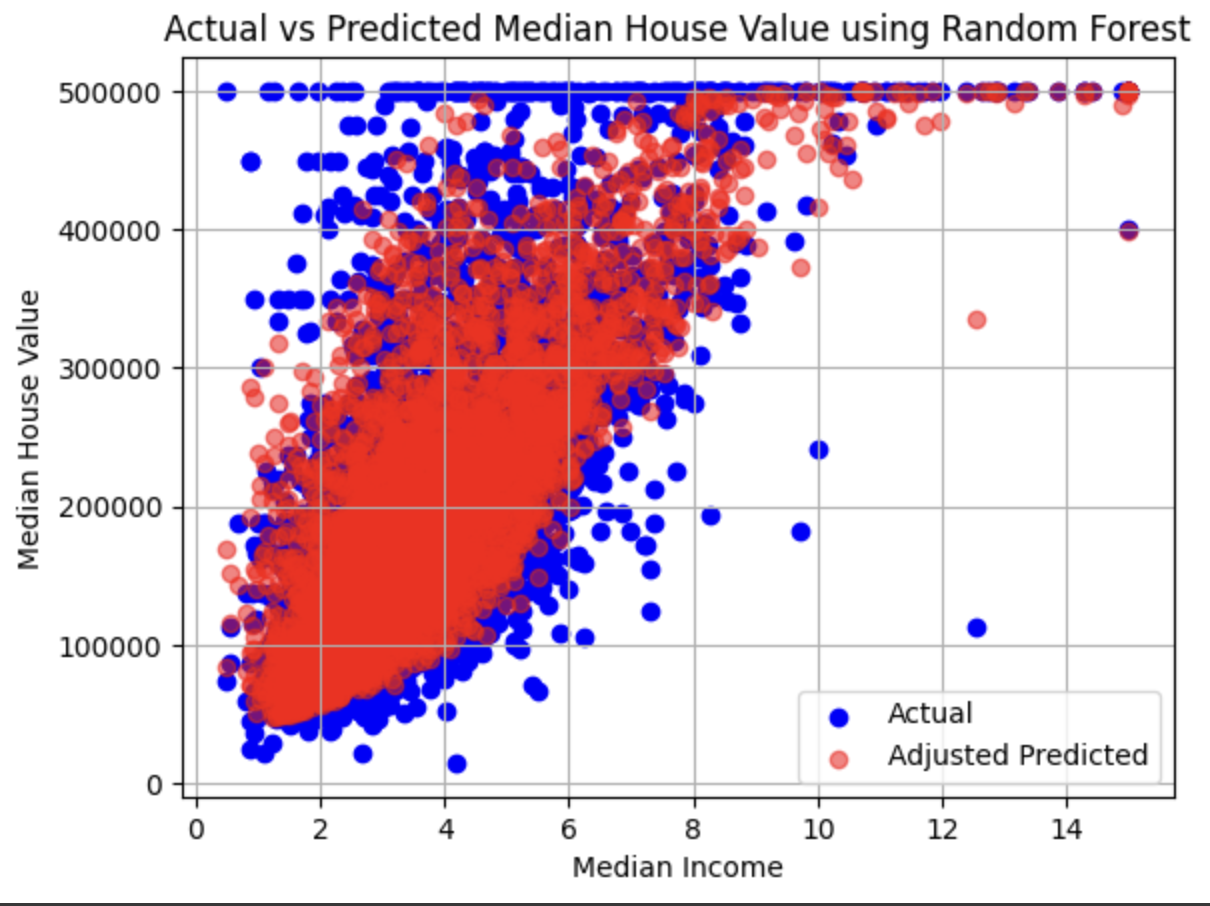
Random Forest is a machine learning algorithm that uses a series of decision trees to produce a final result. Decision trees are classification problems which give you a choice between several possible class labels. They are common supervised learning algorithms but are prone to overfitting and bias. There, random forest comes in, and allows the multiple decision trees to predict more accurate results by selecting only a few features.

3.1**Experiment Design and Evaluation**

By using Random Forest regression, we were able to predict the housing values using the features from the housing dataset. Like with kNN and linear regression, we began by cleaning the dataset and imputing a column’s mean for missing values in certain features. We then used the builtin Random Forest regressor from sklearn to build a model with an R-squared metric of 0.8137121447334693, which suggests that random forest is already a decent predictor of housing values.



We then began to alter some of the default parameters from the Random Forest regressor, focusing on the six most important – bootstrap, n\_estimators, max\_features, max\_depth, min\_samples\_leaf and max\_leaf\_nodes. Out of these, only three proved to improve the model with alterations, n\_estimators which is the number of decision trees in the forest, max\_features which is the maximum number of features to consider when looking for the best split, and min\_samples\_leaf which is the minimum number of samples required to be at a leaf node . These changes brought the R-squared metric up by roughly 0.01 to 0.8257831303, once again suggesting that random forest is a solid predictor of housing values.



These results indicate that housing values can be effectively predicted using median income, with it significantly influencing housing prices. Furthermore, they show that changing the default parameters on the random forest regression from sklearn only slightly impacted our results.

**Conclusion**

The analysis of the California Housing Prices dataset reveals a significant positive relationship between proximity to the ocean and house values. Districts closer to the coast tend to have higher median house values, underscoring the premium placed on coastal properties.

However, a comparison of housing conditions from 1990 to 2024 highlights a concerning trend: while median household income has stagnated or declined relative to the growing cost of homes, making it increasingly difficult to qualify for mortgages on even lower-tier homes. According to recent affordability metrics “Legislative Analyst’s Office (LAO)”, the gap between income and housing costs has widened, reflecting a broader issue of housing affordability across California. This growing disparity suggests that the market has become less accessible for many residents, with the cost of homeownership outpacing income growth. These insights stress the need for targeted housing policies and affordability initiatives to address the escalating challenges faced by prospective homeowners.

**REFERENCES**

[1] Cameron Nugent. 2023. California Housing Prices. Kaggle. Available at: <https://www.kaggle.com/datasets/camnugent/california-housing-prices>.

[2] U.S. Census Bureau. 1990. General Housing Characteristics, California. U.S. Department of Commerce. Available at:

<https://www2.census.gov/library/publications/decennial/1990/ch-1/ch-1-6.pdf>.

[3] Legislative Analyst’s Office. 2023. California’s High Housing Costs: Causes and Consequences. Available at:

<https://lao.ca.gov/LAOEconTax/Article/Detail/793#:~:text=Affordability%20depends%20on%20both%20the,income%20in%202022%20(%2485%2C300)>.

[4] IBM. 2023. K-Nearest Neighbors (KNN). Available at:

<https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20>.

Task Distribution

| Task | People |
| --- | --- |
| Collecting and Preprocessing Data | Huu Nhan Nguyen |
| Implementing & Evaluating Linear Regression | Raphael Luis Santos |
| Implementing & Evaluating KNN | Kelly Tran, Rick Liu |
| Implementing & Evaluating Regression Trees/Random Forest | Amy Ionescu |
| Algorithm Comparisons | All |
| Writing Report | All |

Submission Folder:

<https://drive.google.com/drive/folders/1d2HzqO9fKB5uTFTcvDhWQKH6MerqTmVz?usp=sharing>

Specific Links:

Dataset (In reference as well) :

<https://www.kaggle.com/datasets/camnugent/california-housing-prices>

Preprocessing and Linear Regression Code:

<https://colab.research.google.com/drive/1XAwoCqX-fDbs6Fv3m8cSb0KLDu485Yg1?usp=sharing>

KNN Code:

<https://colab.research.google.com/drive/1LSfKO-XFw9mzGHQMLMuvfY2Rk6hOmmDl?usp=sharing>

Regression Trees/Random Forest Code:

<https://colab.research.google.com/drive/1Xz9SplK2dr2RcuBRzfr9A3UMSmOqZVRY?usp=sharing>