

California Dreamin' or Nightmarin'?

Revealing the Hidden Factors Behind 1990's Housing Market

Capstone Project #3 | Inggar Gumintang | JCDSOL - 014 - 2

Conclusion on Model Performance

Business Recommendations

Profit Margin Estimation

Review on Selected Model

Modeling

Data Understanding and Preparation

Description of the Dataset

Data Cleaning

EDA

Data Manipulation

Data Preprocessing

Business Problem Understanding


Context

Problem Statement

Goal

Business Problem Understanding

Context: The 1990 U.S. economic recession significantly affected housing prices, with national home prices rising by only 0.8% and California experiencing even slower growth.



Problem Statement: The 1990 recession slowed California's housing market, with price growth lagging behind the national average. This study identifies factors influencing housing prices in California, focusing on property attributes, buyer income, and location. Using multiple linear regression, we analyze the key determinants of housing price variations in 1990.

Goal: The aim of this study is to create a predictive model for housing prices in California in 1990.

What We Need?

We need to analyze the dataset to uncover patterns and relationships between property characteristics, median income, and geographic factors affecting housing prices. The next step is to build a regression model to predict housing prices, helping users understand these influencing factors and make informed pricing decisions.

Data Understanding and Preparation - Description of the Dataset

Representing its position along the east-west axis on the Earth's surface

longitude

Indicating its position along the north-south axis

latitude

Age of housing structures

Housing median age

Total number of rooms in the property

total_rooms

total_bedrooms

Total number of bedrooms in the property

population

Total population residing in the neighborhood

households

Total number of households in the neighborhood

median_income

Median income of households in the neighborhood

median house value

Median value of houses in the neighborhood

ocean_proximity

Proximity of the property to the ocean

Data Understanding and Preparation - Data Cleaning

1

Check Data Types

	data_features	data_type	null	null_percentage	unique	unique_sample	filled_count
0	longitude	float64	0	0.00	806	[-124.07, -118.12, -119.33, -123.17, -116.24]	14448
1	latitude	float64	0	0.00	836	[38.47, 38.41, 37.8, 39.34, 38.5]	14448
2	housing_median_age	float64	0	0.00	52	[22.0, 36.0, 15.0, 49.0, 5.0]	14448
3	total_rooms	float64	0	0.00	5227	[2380.0, 2064.0, 13814.0, 2631.0, 860.0]	14448
4	total_bedrooms	float64	137	0.95	1748	[545.0, 42.0, 1284.0, 2289.0, 3479.0]	14311
5	population	float64	0	0.00	3498	[1556.0, 878.0, 3057.0, 887.0, 1756.0]	14448
6	households	float64	0	0.00	1649	[181.0, 977.0, 1336.0, 2338.0, 1594.0]	14448
7	median_income	float64	0	0.00	9797	[3.962, 1.6172, 6.6712, 3.0432, 2.3011]	14448
8	ocean_proximity	object	0	0.00	5	[<1H OCEAN, NEAR OCEAN, INLAND, ISLAND, NEAR BAY]	14448
9	median_house_value	float64	0	0.00	3548	[173700.0, 159000.0, 64300.0, 90700.0, 159400.0]	14448

Should be encoded

2

Check Duplicate

Duplicate rows:

Empty DataFrame

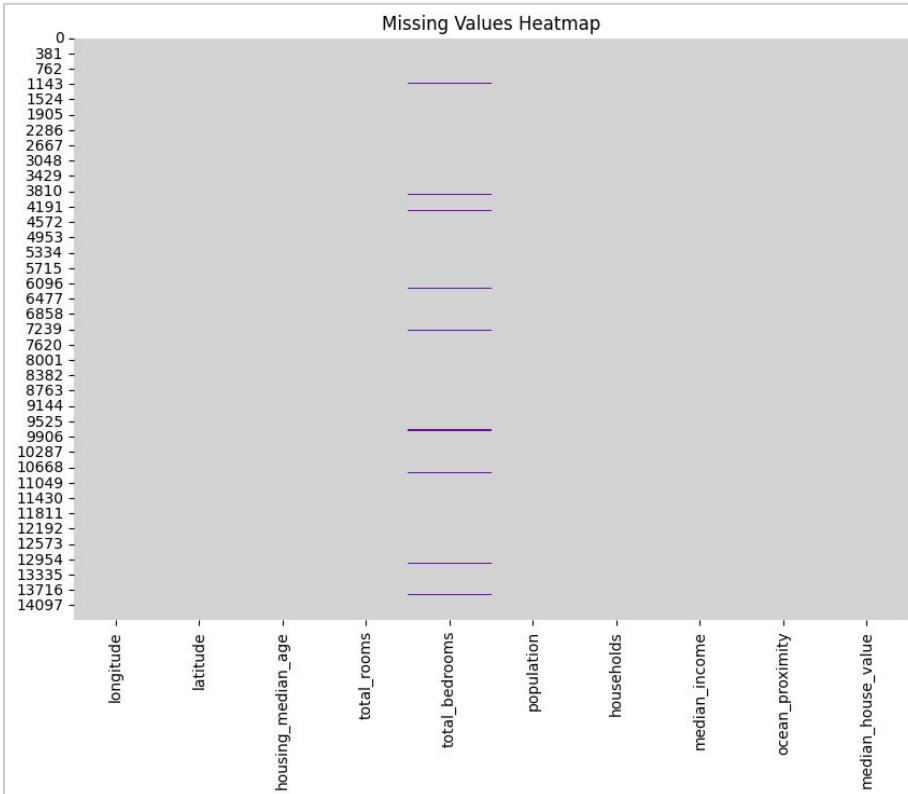
Columns: [longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, ocean_proximity, median_house_value]

Index: []

Data Understanding and Preparation - Data Cleaning

3

Handling Missing Value



139 Missing Value (per 14448) or **0.95%**
of the total entries in the 'total_bedrooms' column

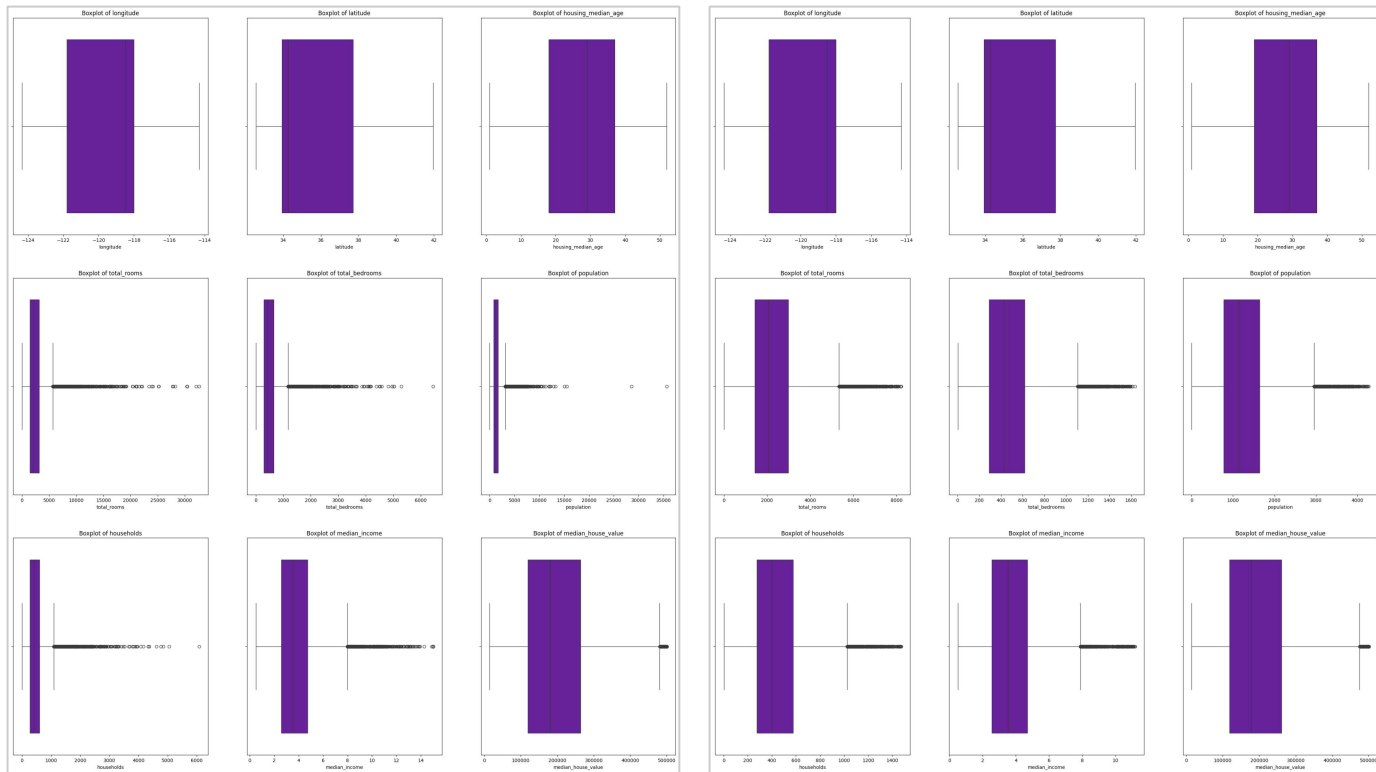
KNN Imputer

```
longitude      0
latitude       0
housing_median_age  0
total_rooms    0
total_bedrooms  0
population     0
households     0
median_income  0
median_house_value  0
ocean_proximity  0
dtype: int64
```

Data Understanding and Preparation - Data Cleaning

4

Handling Outliers



Before

After

Interquartile Range (IQR)

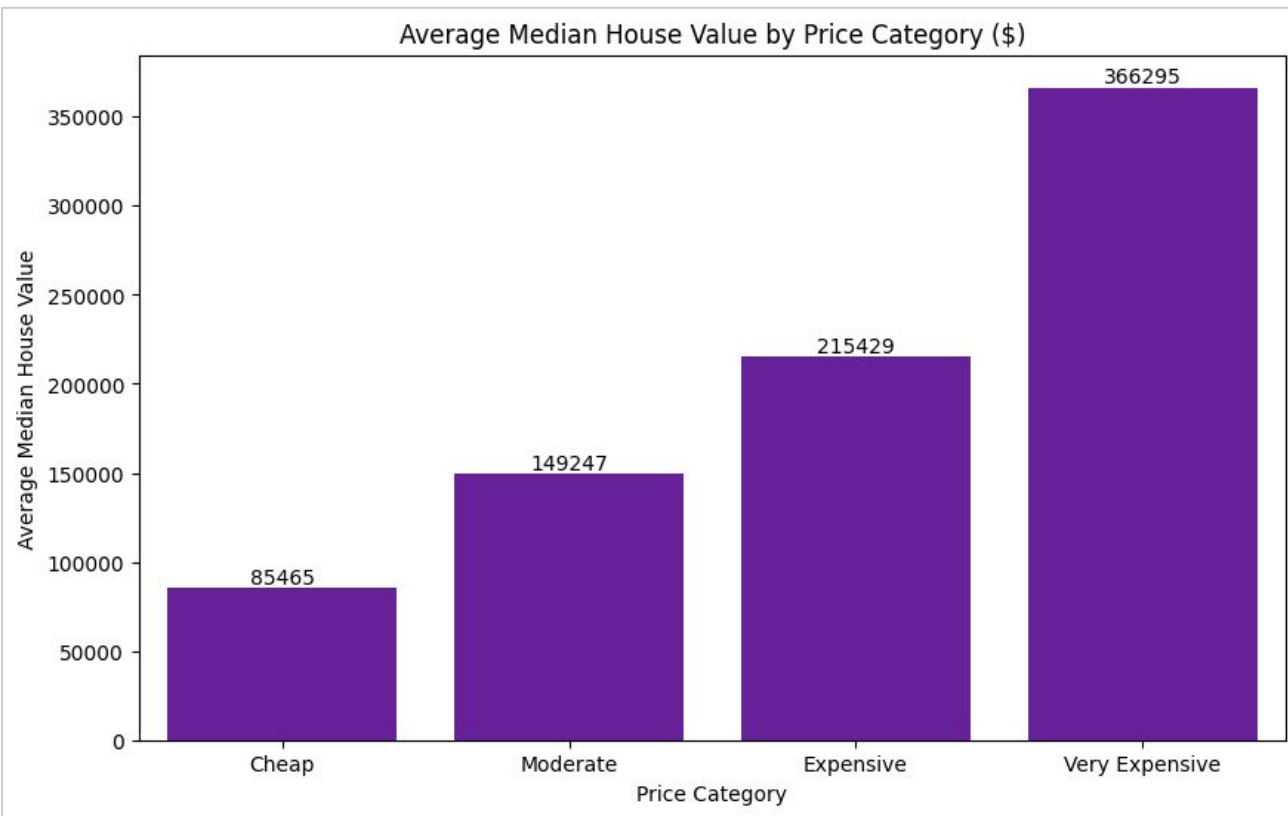
Number of entries removed: **659** (per 14448)
Percentage of entries removed: **4.56%**

```
# Column Non-Null Count Dtype
---
0 longitude 13789 non-null float64
1 latitude 13789 non-null float64
2 housing_median_age 13789 non-null float64
3 total_rooms 13789 non-null float64
4 total_bedrooms 13789 non-null float64
5 population 13789 non-null float64
6 households 13789 non-null float64
7 median_income 13789 non-null float64
8 median_house_value 13789 non-null float64
9 ocean_proximity 13789 non-null object
dtypes: float64(9), object(1)
memory usage: 1.2+ MB
```

Data Understanding and Preparation - EDA

1

Price Category



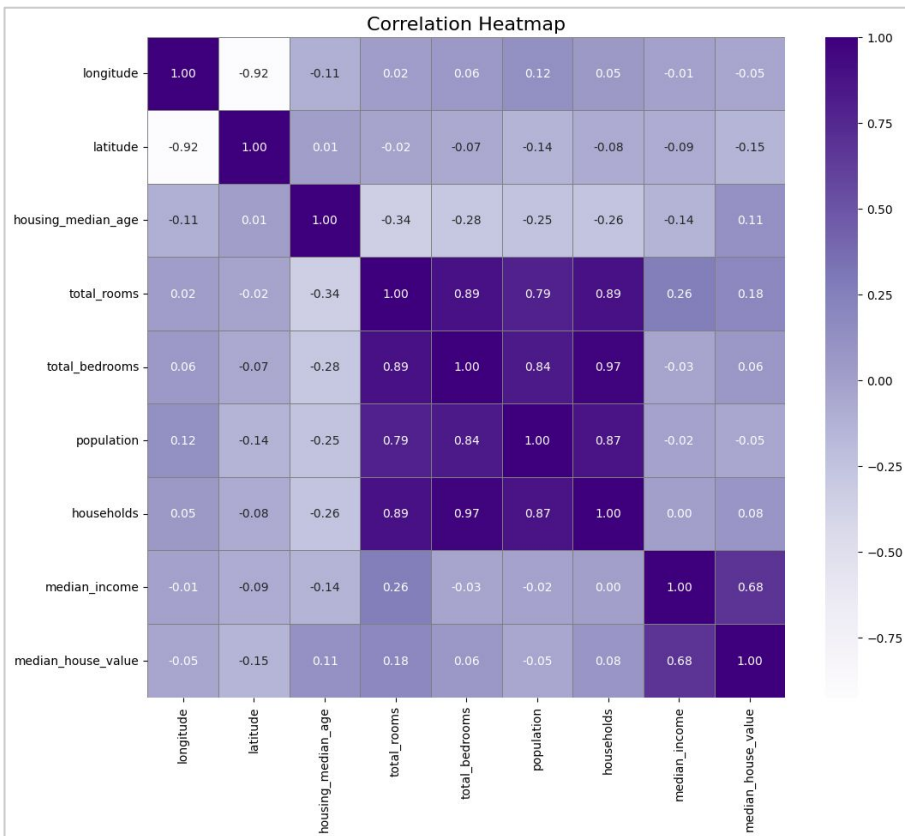
Quartile Categories and Calculation

```
Total houses per price category:  
price_category  
Cheap          3452  
Moderate        3449  
Very Expensive  3444  
Expensive       3444  
Name: count, dtype: int64
```


Data Understanding and Preparation - EDA

2

Correlation



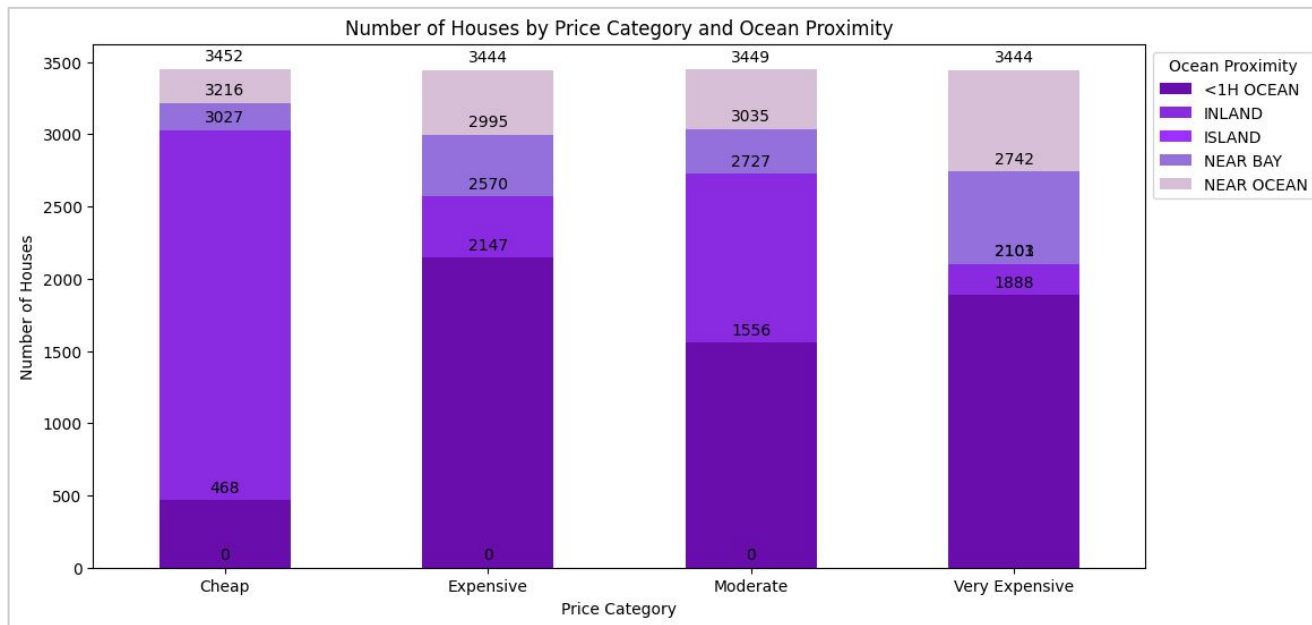
Key Insights:

- **Income and House Value:** The correlation of 0.68 between median income and median house value suggests a strong relationship. Generally, customers seek homes within their budget, and higher income typically enables higher spending on more expensive properties.
- **Household Size and Property Features:** High correlations between households, population, total bedrooms, and total rooms indicate that larger households and populations are associated with more spacious homes. This trend suggests that customers with larger families or more significant household sizes might prefer properties with more rooms and amenities.

Data Understanding and Preparation - EDA

3

Business Opportunity Based on Ocean Proximity



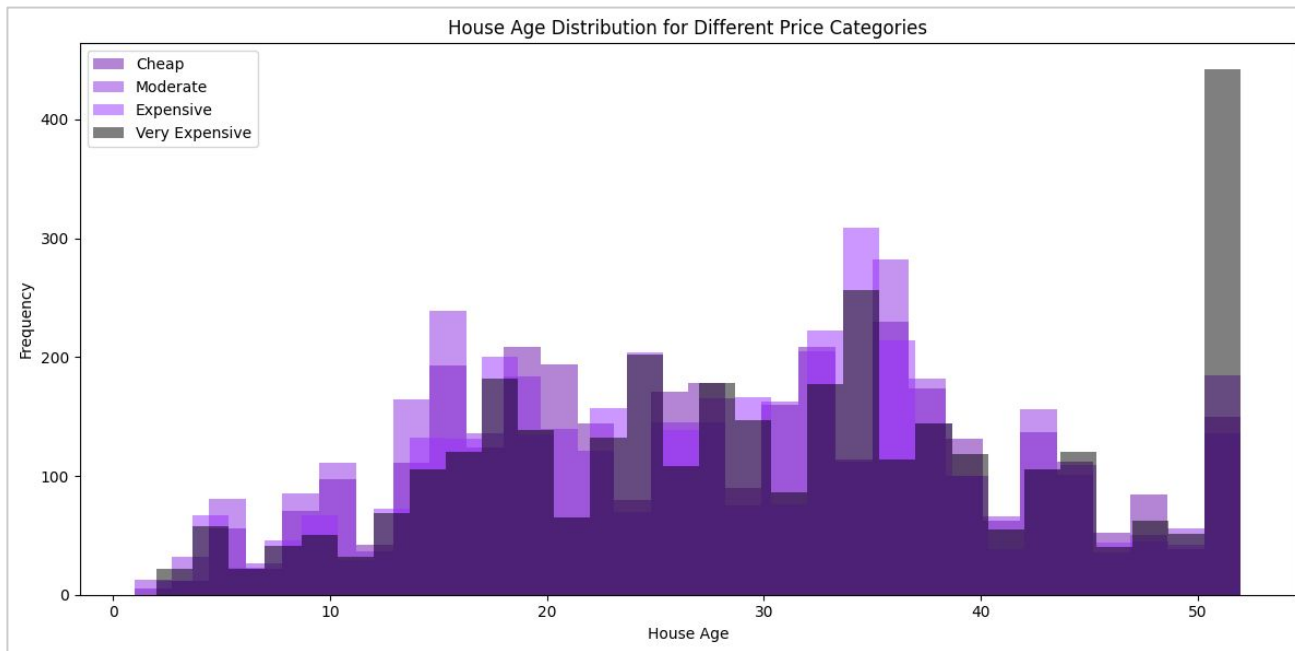
When considering the needs for family housing and the economic recession during that period, it turns out there were many affordable houses available, as indicated by the 'Cheap' or affordable category on the graph. While these houses are not near the ocean, which could be an additional value, this doesn't mean we should only focus on analyzing cheap houses.

During the recession, some investors were willing to enter the market for long-term investments, such as developing properties into vacation villas. It is evident that in the 'Very Expensive' category, most properties are near the ocean. This proximity to the ocean becomes a highly promising value, especially if investments were made in 1990. The trend of vacationing has seen significant growth in the 2020s, reinforcing the long-term value of such investments.

Data Understanding and Preparation - EDA

4

Cost Opportunity Based on House Age



Despite the colors being highly overlapped, it is evident that 'Very Expensive' houses show a significant spike in frequency for homes older than 50 years. This observation is not a problem if we are considering long-term investment, as it relates closely to revenue substitution for business development.

On the other hand, for 'Cheap' houses, there are no major concerns. These houses tend to have a more varied range of ages and conditions. They can be older but well-maintained, or newer but less well-kept.

Data Understanding and Preparation - EDA

4

Business Opportunity Based on Room Types



From the data, no unique patterns are evident as the average number of bedrooms is consistent across each price category, and the differences in the number of rooms are not significant. This might be due to inaccuracies in data collection or classification, such as distinguishing between specific types of rooms like dining rooms, kitchens, and others. This observation suggests that feature classification may need reevaluation.

Additionally, it would be more insightful to analyze the average house size and outdoor space for each price category. This could provide a clearer understanding of how these factors correlate with the different price categories and enhance the evaluation of property features.

Data Understanding and Preparation - Data Manipulation

1

Add Column

Since our dataset values are aggregated by block (with multiple houses per block), we need to calculate room count, bedroom count, and population per house.

	data_features	data_type	null	null_percentage	unique	unique_sample	filled_count
0	longitude	float64	0	0.0	798	[-119.23, -118.33, -124.19, -118.98, -121.69]	13789
1	latitude	float64	0	0.0	831	[33.0, 37.09, 36.75, 36.84, 37.95]	13789
2	housing_median_age	float64	0	0.0	52	[40.0, 49.0, 48.0, 9.0, 23.0]	13789
3	total_rooms	float64	0	0.0	4723	[794.0, 2242.0, 408.0, 193.0, 195.0]	13789
4	total_bedrooms	float64	0	0.0	1465	[683.0, 1073.0, 314.0, 1079.0, 416.0]	13789
5	population	float64	0	0.0	3038	[593.0, 1454.0, 1296.0, 2284.0, 3298.0]	13789
6	households	float64	0	0.0	1279	[31.0, 204.0, 1147.0, 386.0, 454.0]	13789
7	median_income	float64	0	0.0	9290	[4.3472, 4.9676, 7.1273, 4.2153, 5.1041]	13789
8	median_house_value	float64	0	0.0	3502	[68200.0, 141000.0, 52400.0, 279000.0, 51300.0]	13789
9	ocean_proximity	object	0	0.0	5	[NEAR BAY, INLAND, ISLAND, <1H OCEAN, NEAR OCEAN]	13789
10	price_category	object	0	0.0	4	[Expensive, Very Expensive, Moderate, Cheap]	13789
11	avg_bedrooms_per_household	float64	0	0.0	10255	[1.0772727272727274, 1.084070796460177, 1.0094...	13789
12	avg_rooms_per_household	float64	0	0.0	13146	[5.283908045977012, 6.495180722891567, 4.91129...	13789
13	rooms_per_household	float64	0	0.0	39	[26.0, 13.0, 10.0, 7.0, 3.0]	13789
14	bedrooms_per_household	float64	0	0.0	13	[4.0, 10.0, 34.0, 15.0, 3.0]	13789
15	population_per_household	float64	0	0.0	23	[600.0, 64.0, 17.0, 6.0, 18.0]	13789

Data Understanding and Preparation - Data Preprocessing

1

One-Hot Encoding 'Ocean Proximity'

One-hot encoding is used for the variable 'ocean_proximity' to convert categorical values into a numerical format suitable for machine learning algorithms. Since many algorithms require numerical input, one-hot encoding transforms each category into a binary column, indicating the presence or absence of each category. This allows the model to interpret categorical data without implying any ordinal relationship between categories, ensuring accurate analysis and predictions.



```
Data columns (total 19 columns):
```

#	Column	Non-Null Count	Dtype
0	longitude	13789 non-null	float64
1	latitude	13789 non-null	float64
2	housing_median_age	13789 non-null	float64
3	total_rooms	13789 non-null	float64
4	total_bedrooms	13789 non-null	float64
5	population	13789 non-null	float64
6	households	13789 non-null	float64
7	median_income	13789 non-null	float64
8	median_house_value	13789 non-null	float64
9	price_category	13789 non-null	object
10	avg_bedrooms_per_household	13789 non-null	float64
11	avg_rooms_per_household	13789 non-null	float64
12	rooms_per_household	13789 non-null	float64
13	bedrooms_per_household	13789 non-null	float64
14	population_per_household	13789 non-null	float64
15	ocean_proximity_INLAND	13789 non-null	float64
16	ocean_proximity_ISLAND	13789 non-null	float64
17	ocean_proximity_NEAR BAY	13789 non-null	float64
18	ocean_proximity_NEAR OCEAN	13789 non-null	float64

dtypes: float64(18), object(1)
memory usage: 2.0+ MB

Data Understanding and Preparation - Data Preprocessing

2

Feature Selection

Backward regression, or backward elimination, is a feature selection method used to improve model performance by systematically removing less significant variables.

Check for multicollinearity using **VIF**

	Feature	VIF	Multicollinearity
0	longitude	831.154830	Yes
1	latitude	798.614459	Yes
2	housing_median_age	8.430218	No
3	rooms_per_household	41.584760	Yes
4	bedrooms_per_household	32.209470	Yes
5	population_per_household	1.333357	No
6	households	4.811540	No
7	median_income	12.549057	Yes
8	ocean_proximity_INLAND	2.782084	No
9	ocean_proximity_ISLAND	1.000672	No
10	ocean_proximity_NEAR BAY	1.678039	No
11	ocean proximity NEAR OCEAN	1.304819	No

Variables with multicollinearity issues are often closely related, causing the model to struggle in isolating the contribution of each feature to the target variable.

Solutions and Approaches

- Removing Variables: Consider removing one of the highly correlated variables, such as removing rooms_per_household or bedrooms_per_household, to reduce multicollinearity.

Next Steps

- Feature Selection: Based on the above analysis, select the most relevant features and reduce features with high multicollinearity.

Data Understanding and Preparation - Data Preprocessing

2

Feature Selection

**B
E
F
O
R
E**

	Feature	VIF	Multicollinearity
0	longitude	831.154830	Yes
1	latitude	798.614459	Yes
2	housing_median_age	8.430218	No
3	rooms_per_household	41.584760	Yes
4	bedrooms_per_household	32.209470	Yes
5	population_per_household	1.333357	No
6	households	4.811540	No
7	median_income	12.549057	Yes
8	ocean_proximity_INLAND	2.782084	No
9	ocean_proximity_ISLAND	1.000672	No
10	ocean_proximity_NEAR BAY	1.678039	No
11	ocean proximity NEAR OCEAN	1.304819	No

**A
F
T
E
R**

	Feature	VIF	Multicollinearity
0	housing_median_age	4.425413	No
1	bedrooms_per_household	5.067358	No
2	median_income	4.438443	No
3	population_per_household	1.314179	No
4	households	3.268195	No
5	ocean_proximity_INLAND	1.646151	No
6	ocean_proximity_ISLAND	1.000497	No
7	ocean_proximity_NEAR BAY	1.320911	No
8	ocean_proximity_NEAR OCEAN	1.273579	No

The VIF analysis indicates that there are no significant multicollinearity issues among the features, allowing for the continued development of the model.

Modelling

1

Model Selection & Cross Validation


	Model	Mean MSE	Std MSE	Mean RMSE	Std RMSE	Mean MAPE	Std MAPE	Mean R2	Std R2
11	Gradient Boosting	4.031740e+09	3.171170e+08	63447.379826	2483.920608	0.258546	0.006953	0.684748	0.016003
7	Stacking - Linear Regression	4.172266e+09	2.666318e+08	64560.441675	2053.130003	0.260214	0.007484	0.673631	0.011662
9	Random Forest	4.292375e+09	3.427641e+08	65464.738663	2596.672629	0.271216	0.007494	0.664379	0.017366
0	KNN	4.441588e+09	2.611836e+08	66616.798544	1946.898081	0.263178	0.008691	0.652380	0.014250
4	Voting Regressor	4.479980e+09	2.953263e+08	66896.742290	2192.319803	0.267866	0.008219	0.649531	0.014387
12	XGBoost	4.734209e+09	2.397790e+08	68783.614443	1738.718184	0.281251	0.007123	0.629464	0.011199
5	Stacking - KNN	4.894675e+09	3.203323e+08	69924.620843	2285.217518	0.283884	0.006798	0.617125	0.014811
8	Bagging Regressor	4.973491e+09	3.800228e+08	70471.004508	2707.100009	0.292990	0.003766	0.611201	0.017039
2	Linear Regression	4.988316e+09	3.807265e+08	70575.966622	2710.844254	0.293190	0.004152	0.610050	0.016919
10	AdaBoost	6.231703e+09	5.010388e+08	78875.548439	3217.296767	0.430870	0.021155	0.512548	0.030374
1	Decision Tree	7.949142e+09	4.954557e+08	89113.815245	2805.398621	0.344926	0.013833	0.377455	0.037135
6	Stacking - DT	8.376965e+09	2.602173e+08	91514.663772	1425.183694	0.363956	0.004484	0.344111	0.005955
3	Support Vector Regressor	1.340753e+10	6.507206e+08	115756.142944	2835.632460	0.518273	0.009043	-0.049034	0.012033

Modelling


1

Model Selection & Cross Validation

Cross-Validation provides a more reliable estimate of a model's performance compared to using a single train-test split.



	Model	Test RMSE	Test MAPE	Test R2
0	Gradient Boosting	65084.330297	0.256914	0.676900
1	Stacking - Linear Regression	82293.848227	0.270221	0.483443



Based on the evaluation results, Gradient Boosting is the best model with the lowest RMSE and MAPE, as well as the highest R2, indicating superior accuracy and explanatory power compared to Stacking - Linear Regression, which performs worse across all metrics. Other models like Random Forest and XGBoost also perform better than Stacking - Linear Regression, but still fall short compared to Gradient Boosting.

Modelling

2

Hyperparameter Tuning

Results Before Hyperparameter Tuning:

	Model	Test RMSE	Test MAPE	Test R2
0	Gradient Boosting	64573.097024	0.256774	0.681956

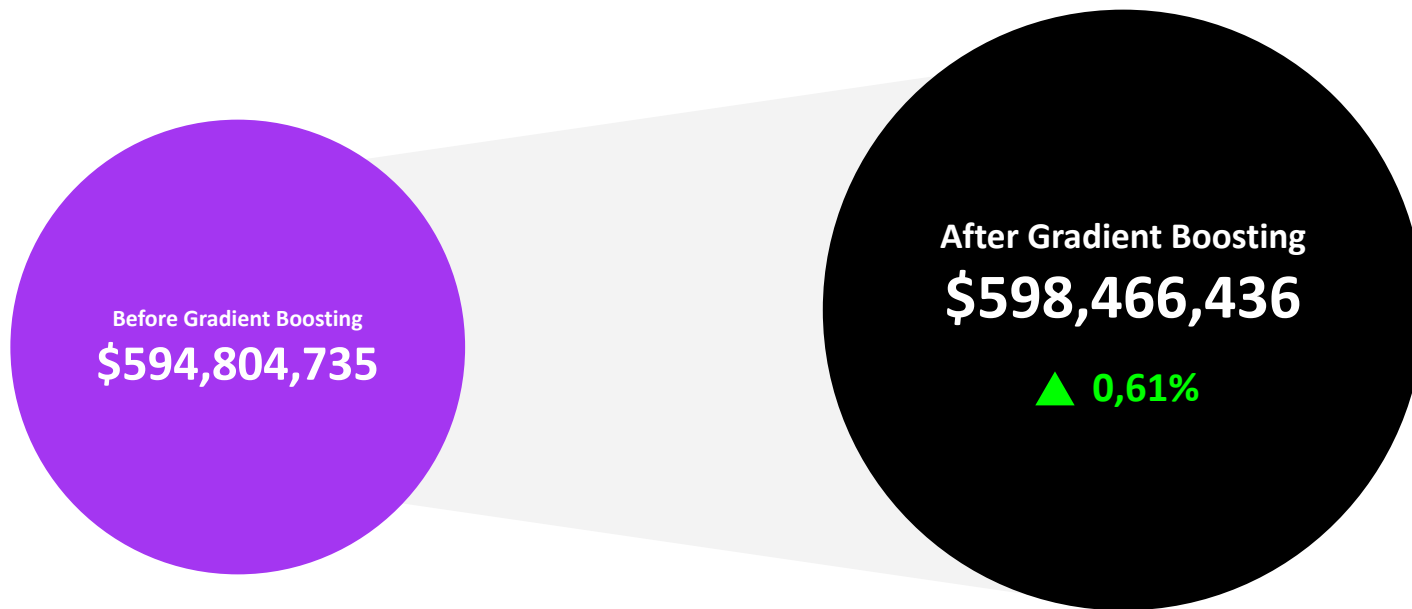
Results After Hyperparameter Tuning:

	Model	Test RMSE	Test MAPE	Test R2
0	Gradient Boosting	64573.097024	0.256774	0.681956



The hyperparameter tuning for Gradient Boosting did not lead to significant improvements in test metrics. The test RMSE, MAPE, and R^2 values remained the same before and after tuning, indicating that the selected hyperparameters were already well-suited for this model or that further tuning may be needed to observe more notable changes.

Conclusion on Model Performance - Profit Margin Estimation



Conclusion on Model Performance - Review on Selected Model

The analysis identifies **Gradient Boosting as the most effective** model for predicting housing prices in California during 1990. It achieved the lowest RMSE and MAPE, and the highest R-squared value, demonstrating its superior accuracy compared to other models like Stacking - Linear Regression, Random Forest, and XGBoost. However, **hyperparameter tuning did not lead to significant improvements**, suggesting that the current hyperparameters are well-suited or further fine-tuning might be required.

Model Limitations

Historical Data

- **Detail:** Data is specific to 1990, potentially making findings less relevant to current market conditions.
- **Impact:** Insights may not fully reflect today's housing market.

Limited Features

- **Detail:** Excludes important details like specific room types and lot size.
- **Impact:** Model accuracy may be constrained by the lack of granular data.

Economic Changes

- **Detail:** Focuses on the 1990 recession, which may not apply to other economic conditions.
- **Impact:** Results might not be applicable to different economic scenarios.

Data Quality

- **Detail:** Potential data errors or missing values.
- **Impact:** Can affect the overall accuracy of the model.

Hyperparameter Tuning

- **Detail:** Did not significantly improve performance.
- **Impact:** Model's potential might be limited by current settings.

External Factors

- **Detail:** Excludes factors like interest rates or tax incentives.
- **Impact:** Model might not provide a complete picture of housing prices.

Conclusion on Model Performance - Business Recommendations

Feature Expansion: Improve the dataset with detailed property features and lot size to enhance model accuracy and understanding of pricing factors.

Target Market Analysis: Focus on high-value properties in desirable locations like "Near Ocean" or "Island" for long-term investments. Tailor offerings to affluent buyers or those interested in vacation properties.

Price Segmentation: Use insights to better segment the market. Develop marketing strategies that highlight key property attributes in different price categories.

References

Influencing Factors of California Housing Prices in 1990: a Multiple Linear Regression Analysis

Yitan Hao^{1,*,*†}, Luhan Zhuang^{2,*,*†}, Zitao Ying^{3,*,*††}, Juncheng Zhai^{4,*,*††}

* Corresponding author: 118010088@link.cuhk.edu.cn¹, *luhan.zhuang@mail.utoronto.ca²,
*3180103695@zju.edu.cn³, *1902020040@mail.bnuz.edu.cn⁴

School of Data Science, The Chinese University of Hong Kong, Shenzhen, China¹

Faculty of Arts and Science, University of Toronto, Toronto, Canada²

Electronic Engineering, Zhejiang University, Hangzhou, China³

School of Management, Beijing Normal University Zhuhai, Zhuhai, China⁴

[†]These two authors contributed equally.

^{††}These two authors contributed equally.

[Link](#)

**JUST THE
FACTS**

THE CALIFORNIA ECONOMY:
CRISIS IN THE HOUSING MARKET

MARCH 2008

- **IN 2007, CALIFORNIA HOME PRICES SUFFERED THE FASTEST AND STEEPEST DECLINE IN 25 YEARS.** California home prices fell 6.6% between the fourth quarter of 2006 and the fourth quarter of 2007. (Just two years ago, home prices rose 21% in California.) Nationally, home prices rose 0.8%, well ahead of California but the slowest national growth since 1990.

[Link](#)

Thank **You !**

Capstone Project #3 | Inggar Gumintang | JCDSOL - 014 - 2