**A Statistical model to determine the most significant factors and understand their relation with medium house prices in California**

**Background:**

Real estate prices are influenced by a number of factors. In California, like in any other state, the old adage in real estate, "location, location, location," holds true. This pertains to proximity to urban centers, coastline, desirable neighborhoods, schools, parks as well as transportation hubs. Other factors include but are not limited to; economic indicators (robust job markets, and income levels), housing supply and demand, mortgage rates, natural calamities (e.g. earthquake and wild-fires) prone areas and population growth. All these factors influence real estate demand and pricing.

The aim of the study is to fit a statistical model to identify the most significant factors associated with the median house value in California.

**Data preparation**

**Source**: Dataset related to real estate prices in California was provided. N = 20640

**Variables:**

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 longitude 20640 non-null float64

1 latitude 20640 non-null float64

2 housing\_median\_age 20640 non-null float64

3 total\_rooms 20640 non-null float64

4 total\_bedrooms 20433 non-null float64

5 population 20640 non-null float64

6 households 20640 non-null float64

7 median\_income 20640 non-null float64

8 median\_house\_value 20640 non-null float64

9 ocean\_proximity 20640 non-null object

dtypes: float64(9), object(1)

**Cleaning**: No duplicated records. Missingness rate was 1% in the variable *total\_bedrooms*. Since model building is sensitive to missing data, these records were dropped.

**Transformations**:

**Derivations:** New variables created:

*avg\_rooms = total\_rooms/households*

*avg\_bedrooms = total\_bedrooms/households*

**Imputation**: Median was used for any NaN imputations.

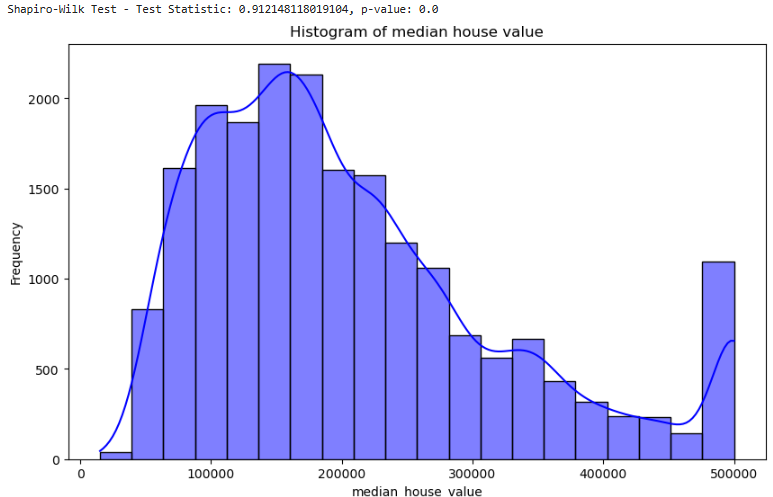
**Encoding:** OneHotEncoder was used to create dummy variables for categories of variable; *ocean\_proximity*. Labelencoder was used to create a numeric version of *ocean\_proximity* used in preliminary modelling and diagnostics checking.

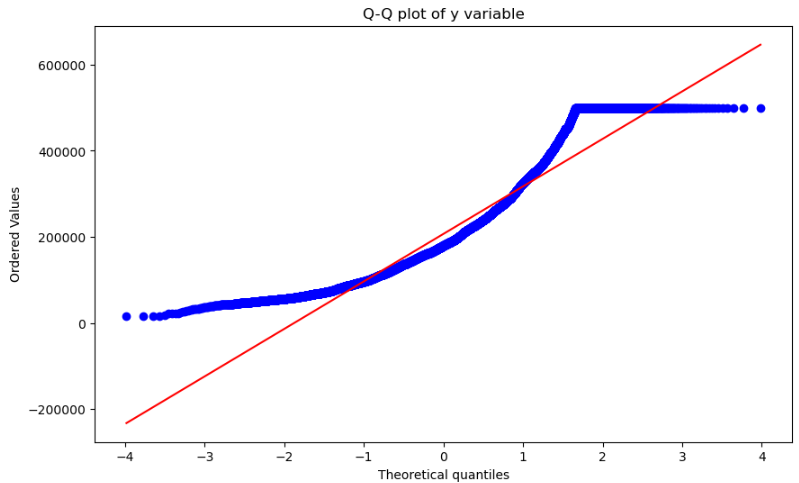
**Data splitting**: Dataset was split into a training (80%) and test (20%) set.

**Data Modelling**

**Diagnostics:**

**Normality:** Q-Q plot and Shapiro-wilk Test revealed non-normality of median house value. Best transformation method was square-root.



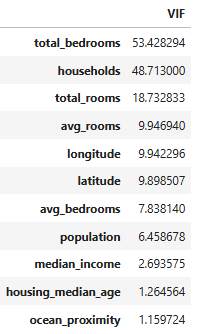


**Shapiro-Wilk test**: H0: The data are normally distributed. H1: The data is not normally distributed.

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is sufficient evidence that the data is not from a normally distributed population.

**Outlying records:** Records with *avg\_rooms* >80 were dropped.

**Multicollinearity:** Large pairwise correlations (r>0.8) between predictors results into inflated beta coefficients hence misleading conclusions. Variance Inflation Factor (VIF) was used to assess multicollinearity.



Variables; *total\_bedrooms, households* and *total\_rooms* with a VIF value > 10 were excluded from the analysis.

**Dimension reduction/Feature importance:** Permutation feature importance, a model inspection technique that measures the contribution of each feature to a fitted model’s statistical performance, was used.

('longitude', 1.1528406659786476)

('latitude', 1.2244316537767588)

('housing\_median\_age', 0.019378949198610607)

('avg\_rooms', 0.061314080264007045)

('avg\_bedrooms', 0.07209263899646859)

('population', -3.420039031976074e-05)

('median\_income', 1.0994103406765414)

('ocean\_proximity', 0.0007842031488236145)

Longitude and latitude (“location location location”) followed by median income were the most important variables while population was the least important one.

**Remedies (re-fitting model and diagnostics re-check)**

**Dimension reduction/Feature selection:** Backward elimination with a significance level of 5% (P-value=0.05) was used to fit the best model from all potential predictors.

**Final model:**

R-squared: 0.6153819336905948 Adjusted R-squared: 0.6150678828485034

F-value: 1959.4977984850889 F-test p-value: 0.0

**Equation**: y = -2282178.1816 + (-26825.6701 \* longitude) + (-24814.3436 \* latitude) + (865.7194 \* housing\_median\_age) + (-10486.5780 \* avg\_rooms) + (68769.3219 \* avg\_bedrooms) + (43023.6992 \* median\_income) + (-33062.8238 \* ocean\_proximity\_<1H OCEAN) + (-71131.9366 \* ocean\_proximity\_INLAND) + (160319.1778 \* ocean\_proximity\_ISLAND) + (-31128.2978 \* ocean\_proximity\_NEAR BAY) + (-24996.1195 \* ocean\_proximity\_NEAR OCEAN)

MSE on Validation Set: 5190909557.2384 Test MSE: 4903442897.667893

**Interpretation:**

Model is significant (p-value for the F test <0.0001)

Maximum VIF < 10

Adjusted R-squared= 0.62 meaning that 62% of the variance in the medium house prices in California is explained by the variables in the model.

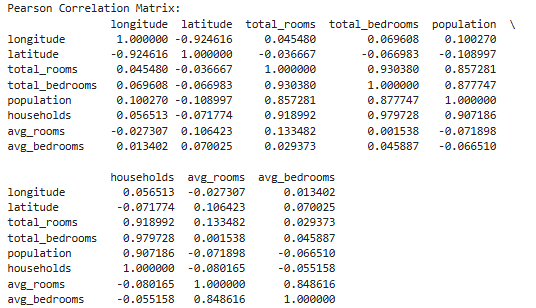
The location of a property has a significant impact on its value. Every unit increase in longitude results into $26,825 decrease in the medium house price while for every unit increase in latitude results into $24,814 decrease in the medium house price.

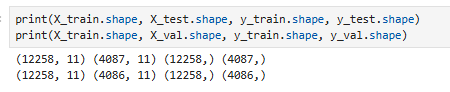
For every unit increase in the number of the house’s bedrooms, the medium house price increased by $68,769 while if the overall size of the house increased, its median price would decrease by $10,486.

For every unit increase in the house’s median age results into $865 increase in the medium house price.

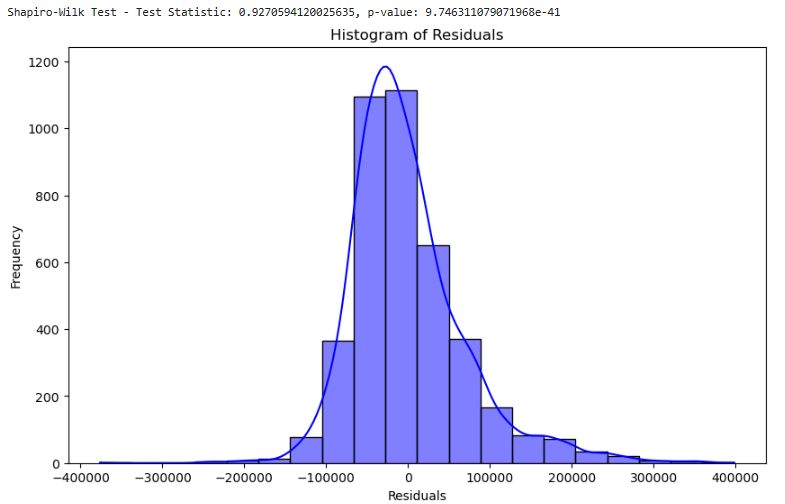
Island properties and those near the ocean had a higher median value compared to in-land properties.

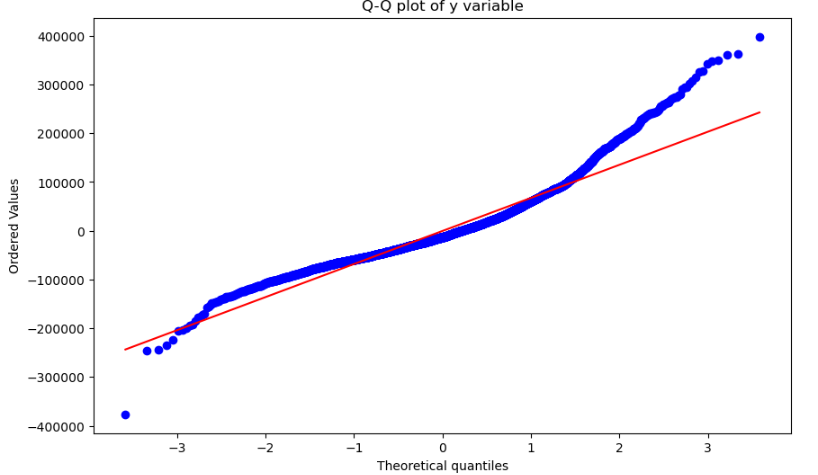
**Appendix:**

****



**Diagnostics:**

****

****

**Limitations:**

Data is aggregate, not at household level.

Not enough time to explore interactions and other model diagnostics.

**References:**

[OneHotEncoder — scikit-learn 1.5.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html)

[4.2. Permutation feature importance — scikit-learn 1.5.0 documentation](https://scikit-learn.org/stable/modules/permutation_importance.html#id2)

[Backward Elimination for Feature Selection in Machine Learning | by Sunny Srinidhi | Towards Data Science](https://towardsdatascience.com/backward-elimination-for-feature-selection-in-machine-learning-c6a3a8f8cef4)