Predict House Prices in California

Executive Summary

housing is a unique asset. Both an investment and a consumption good, it is traded in markets that are subject to significant search frictions and information asymmetries. In addition, housing accounts for a large share of wealth in the economy. As a result, changes in house prices can have large effects on aggregate economic activity. this makes housing an ideal asset for the study of a range of questions of broader economic interest (What Can Housing Markets Teach Us about Economics?, n.d.). The aim of this study is to estimate house prices in California to help real estate developers appraise their property. The hypothesis derived from the goal of the project is summarized as follows: A Linear Regression Model can be constructed on the research data set with an accuracy greater than 70%.

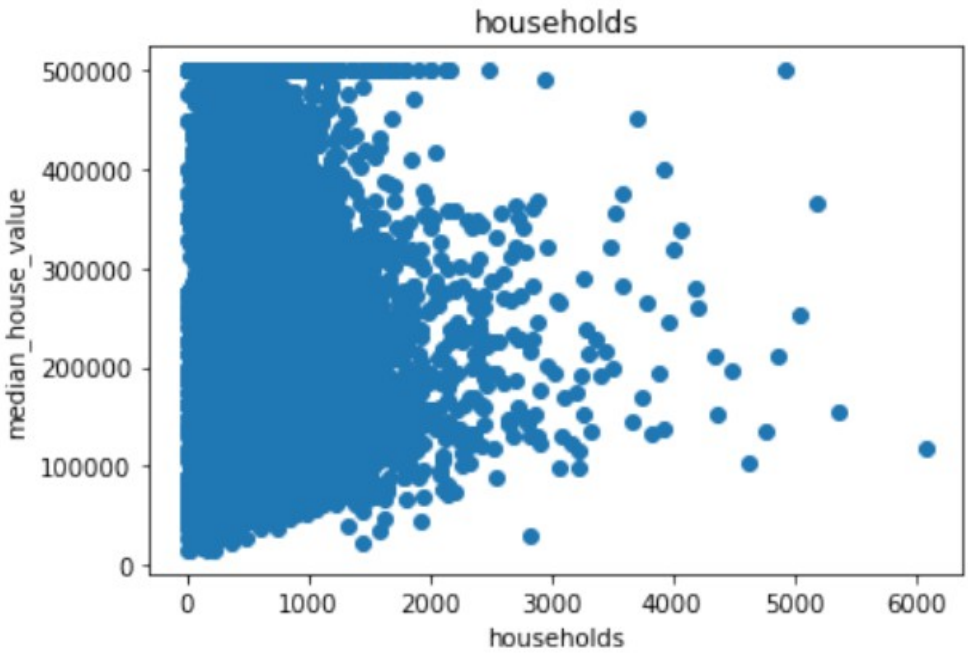
The data collected for this study is publicly available information provided by the U.S. census bureau (California Housing Prices, n.d.). The data contains information from the 1990 California census. The data set contains 20,640 rows. The data set contains the following variables of longitude, latitude, housing_median_age, total_rooms,

total_bedrooms, population, households, median_income, median_house_value, ocean_proximity. The target variable is the continuous variable called median house value.

The analysis steps included:

- Forward fill imputation for the continuous variables with missing values
- Transforming the categorical variables with one-hot encoding
- Detecting missing values
- Detect duplicate rows
- Visualization of all variables with box plots, scatter plots, and Q-Q plots.
- Construction of MLR model on training data set.
- Test accuracy of the MLR model.
- Use stepwise backwards elimination of predictor variables if necessary.

Exploratory data analysis of predictor variables with scatter plots informed that the predictor variables did not have a linear relationship with the response variable. Scatter plots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the predictor variables and the response variables, and the strength of the relationship (*3.2: Scatter Plots*, 2021). The households variable is shown in the scatter plot below.



Multiple Linear Regression was used to identify the statistically significant independent variables and interaction effects, build the model, and test the model for accuracy. Once the model was built the summary output showed that the adjusted R-squared value was 0.649. This was less than the required 0.70 needed to reject the null hypothesis. The model was not reduced due to the fact that no coefficient had a P value greater than 0.05.

our ago
sterday
ites ago
ars ago
sterday
ays ago
ays ago

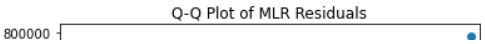
OLS Regression Results

Dep. Variable:	median_house_value	R-squared:	0.650
Model:	OLS	Adj. R-squared:	0.649
Method:	Least Squares	F-statistic:	2550.
Date:	Fri, 24 May 2024	Prob (F-statistic):	0.00
Time:	16:02:46	Log-Likelihood:	-2.0727e+05
No. Observations:	16512	AIC:	4.146e+05
Df Residuals:	16499	BIC:	4.147e+05
Df Model:	12		
Covariance Type:	nonrobust		

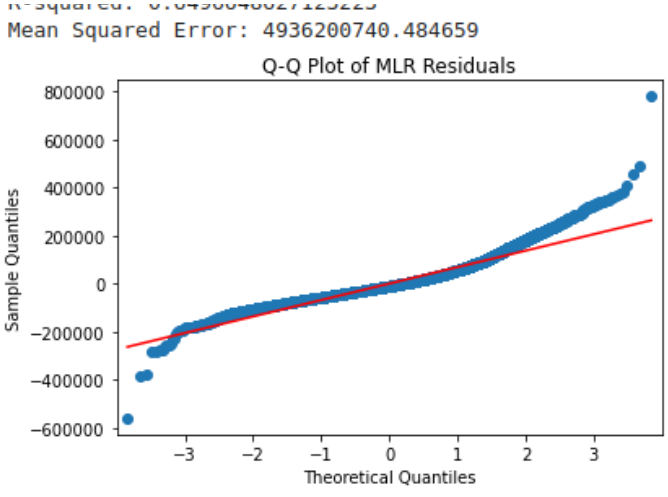
	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.276e+06	9.73e+04	-23.394	0.000	-2.47e+06	-2.08e+06
longitude	-2.684e+04	1127.047	-23.813	0.000	-2.9e+04	-2.46e+04
latitude	-2.547e+04	1111.486	-22.914	0.000	-2.76e+04	-2.33e+04
housing_median_age	1102.1851	48.605	22.676	0.000	1006.914	1197.456
total_rooms	-6.0215	0.886	-6.796	0.000	-7.758	-4.285
total_bedrooms	102.7894	7.697	13.355	0.000	87.703	117.876
population	-38.1729	1.188	-32.129	0.000	-40.502	-35.844
households	48.2528	8.375	5.761	0.000	31.836	64.669
median_income	3.947e+04	375.091	105.238	0.000	3.87e+04	4.02e+04
ocean_proximity_INLAND	-3.979e+04	1933.681	-20.576	0.000	-4.36e+04	-3.6e+04
ocean_proximity_ISLAND	1.361e+05	3.43e+04	3.972	0.000	6.89e+04	2.03e+05
ocean_proximity_NEAR BAY	-5136.6422	2111.676	-2.432	0.015	-9275.756	-997.529
ocean_proximity_NEAR OCEAN	3431.1401	1751.612	1.959	0.050	-2.208	6864.488

Omnibus:	4119.707	Durbin-Watson:	1.967
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16516.873
Skew:	1.189	Prob(JB):	0.00
Kurtosis:	7.284	Cond. No.	7.21e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.21e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
R-squared: 0.6496648627123223
Mean Squared Error: 4936200740.484659



The Q-Q plot of the residuals has a non normal distribution which indicates a poor fitting model and aligns with the poor accuracy score.



E.

The limitations of this study are mostly related to the available dataset. There was not a linear relationship between the predictor variables and the target variable. Not having domain specific knowledge to aid in identification of outliers was also a notable limitation.

Some proposed actions based on the findings are to not use this model for predicting house prices. Further study should focus on increasing the accuracy to 70% or greater before it can be used effectively. Some of the suggestions below may increase the accuracy to the desired threshold.

Some suggestions for future research include but are not limited to the following:

- Working with a domain expert to help identify outliers
- apply logarithmic transformations to the skewed predictors to normalize them. This may increase the accuracy of the model to greater than 70%. Logarithmic transformation is a convenient means of transforming a highly skewed variable into a more normalized dataset. When modeling variables with non-linear relationships, the chances of producing errors may also be skewed negatively (DEV Community, 2019).

The most important benefit of this analysis is to more accurately appraise houses. Houses hold a large portion of wealth in the economy and they have an effect on the aggregate economy so it is important that developers and investors have statistically significant data to help appraise their houses. Another benefit of this study is to understand how the coefficients of the MLR model correlate with the house value. This can be of use when deciding on ways to increase the value of a house.

References:

California Housing Prices. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/camnugent/california-housing-prices

DEV Community. (2019, April 19). Logarithmic Transformation in Linear Regression Models: Why & When. The DEV Community; dev.to.

https://dev.to/rokaandy/logarithmic-transformation-in-linear-regression-models-why-when-3a7c

3.2: Scatter Plots. (2021, September 8). Statistics LibreTexts. https://stats.libretexts.org/Courses/City_University_of_New_York/Introductory_Statistics_with_Probability_(CUNY)/03%3A_Introduction_to_Linear_Regression_and_Correlation/3.02%3A_Scatter_Plots

What Can Housing Markets Teach Us about Economics? (n.d.). NBER. https://www.nber.org/reporter/2016number4/what-can-housing-markets-teach-us-about-economics