CASE STUDY: CALIFORNIA HOUSE- PRICE DRIVERS

REQUIREMENTS

- Problem definition
- Independent variables
- Dependent variables
- Required python libraries
- Data loading
- Data preprocessing
- Missing data removal
- Outlier detection
- Data Visualization
- Histogram/Boxplots/Scatter plots
- Inter-quantile-range [IQR]
- Percentiles/ Quantiles
- Correlation Analysis
- Correlation Heatmaps
- String Variable Handling
- Dummy Variables

CAUSAL ANALYSIS

From the statistical analysis

OLS Regression Results

==========

Dep. Variable: median_house_value R-squared: 0.588

Model: OLS Adj. R-squared: 0.588

Method: Least Squares F-statistic: 1973.

Date: Thu, 01 Aug 2024 Prob (F-statistic): 0.00

Time: 12:20:12 Log-Likelihood: -1.8879e+05

No. Observations: 15220 AIC: 3.776e+05

Df Residuals: 15208 BIC: 3.777e+05

Df Model: 11

Covariance Type: nonrobust

==========

==

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.747e+06	9.16e+04	-19.073	0.000	-1.93e+06	-1.57e+06
longitude	-2.268e+04	997.843	-22.728	0.000	-2.46e+04	-2.07e+04
latitude	-2.109e+04	981.562	-21.490	0.000	-2.3e+04	-1.92e+04
housing_median_age	846.4636	44.561	18.996	0.000	759.118	933.809
total_rooms	-2.6698	0.716	-3.728	0.000	-4.073	-1.266
population	-33.1937	1.050	-31.620	0.000	-35.251	-31.136
households	124.3609	4.388	28.343	0.000	115.761	132.961
median_income	3.562e+04	424.453	83.928	0.000	3.48e+04	3.65e+04
ocean_proximity_<1H OCEAN	-1.7e+05	2.95e+04	-5.755	0.000	-2.28e+05	-1.12e+05
ocean_proximity_INLAND	-2.108e+05	2.96e+04	-7.123	0.000	-2.69e+05	-1.53e+05
ocean_proximity_NEAR BAY	-1.779e+05	2.96e+04	-6.015	0.000	-2.36e+05	-1.2e+05
ocean_proximity_NEAR OCEAN	-1.689e+05	2.95e+04	-5.716	0.000	-2.27e+05	-1.11e+05

==========

Omnibus: 3434.427 Durbin-Watson: 2.004

Prob(Omnibus): 0.000 Jarque-Bera (JB): 10645.381

Skew: 1.157 Prob(JB): 0.00

Kurtosis: 6.381 Cond. No. 7.61e+05

The output provided is from an Ordinary Least Squares (OLS) regression analysis. This statistical method is used to model the relationship between a dependent variable (in this case, median_house_value) and one or more independent variables.

Here is a detailed interpretation of the results:

Model Summary

- **Dependent Variable**: median_house_value (the variable we are trying to predict)
- **R-squared**: 0.588
    - This value indicates that 58.8% of the variance in the median_house_value is explained by the model. This is a relatively good fit, suggesting the independent variables collectively explain a substantial portion of the variation in house prices.
- **Adj. R-squared**: 0.588
    - The adjusted R-squared is almost the same as the R-squared value, indicating that the model is well-specified and additional variables do not significantly increase the explanatory power.

ANOVA Table

- **F-statistic**: 1973
    - This tests whether at least one of the regression coefficients is different from zero. Given the high F-statistic value and the p-value of 0.000, the model is statistically significant.
- **Prob (F-statistic)**: 0.00
    - The p-value associated with the F-statistic. A value of 0.00 indicates that the model is highly significant.

Coefficients Table

This table shows the estimated coefficients for each independent variable, their standard errors, t-values, and p-values.

- **const** (Intercept): -1.747e+06
    - This is the estimated value of median_house_value when all independent variables are zero. A very high negative value, indicating that the baseline without other factors is very low (but note that interpretation of the intercept in regression models with categorical variables needs caution).
- **longitude**: -2.268e+04
    - Each unit increase in longitude decreases median_house_value by 22,680 on average, holding other factors constant. This is statistically significant ($p < 0.05$).
- **latitude**: -2.109e+04
    - Each unit increase in latitude decreases median_house_value by 21,090 on average, holding other factors constant. This is also statistically significant.
- **housing_median_age**: 846.4636
    - Each unit increase in median age of houses increases median_house_value by approximately 846, holding other factors constant.
- **total_rooms**: -2.6698
    - Each additional room decreases median_house_value by approximately 2.67, holding other factors constant. This could suggest that more rooms might be associated with less desirable houses or larger homes in less expensive areas.
- **population**: -33.1937

    - Each additional person in the population decreases median_house_value by about 33.19, which might reflect overcrowding reducing house prices.

- **households**: 124.3609
    - Each additional household increases median_house_value by about 124.36.
- **median_income**: 3.562e+04
    - Each unit increase in median income increases median_house_value by 35,620, holding other factors constant. This has a very large and significant positive effect, as expected.
- **ocean_proximity_<1H OCEAN**: -1.7e+05
    - Houses within 1 hour of the ocean decrease in value by 170,000 compared to the baseline category.
- **ocean_proximity_INLAND**: -2.108e+05
    - Inland houses are worth 210,800 less than the baseline category.
- **ocean_proximity_NEAR BAY**: -1.779e+05
    - Houses near a bay are worth 177,900 less than the baseline category.
- **ocean_proximity_NEAR OCEAN**: -1.689e+05
    - Houses near the ocean are worth 168,900 less than the baseline category.

Diagnostics

- **Omnibus**: 3434.427
    - This tests the skewness and kurtosis of the residuals. A significant value suggests the residuals are not normally distributed.
- **Prob(Omnibus)**: 0.000
    - Indicates that the residuals are not normally distributed (as $p < 0.05$).
- **Jarque-Bera (JB)**: 10645.381
    - Another test for normality of residuals. A high value indicates non-normality.
- **Skew**: 1.157
    - The distribution of the residuals is positively skewed.
- **Kurtosis**: 6.381
    - Indicates the residuals have heavier tails than a normal distribution.
- **Durbin-Watson**: 2.004
    - Tests for autocorrelation in the residuals. A value close to 2 suggests no autocorrelation.

Conclusion

The model explains a significant portion of the variability in median_house_value (58.8%). Most of the independent variables are statistically significant, with median income having the most substantial positive effect on house prices. The diagnostic tests suggest some issues with the normality of residuals, which could impact the validity of statistical tests and confidence intervals.

4o