1. Brief description of the data set and a summary of its attributes

The house prices: advanced regression techniques from kaggle

Housing prices

Msssubclass: identifies the type of dwelling involved in the sale. From size, number of floors, number of rooms, etc.

1. Initial plan for data exploration

Looking through the data and identify variables and looking for null values, sizes, and number of non-numeric values, identifying outliers. Using: .describe(), .info(), .head(), dtypes()

1. Actions taken for data cleaning and feature engineering

Checking for skewness. The predictions generated by the final model will be log-transformed, so I'll convert these predictions back their original form later on. I will start with making a histogram with plt.hist(). Then I will transform train.salesprice with np.log() and calculate the skewness and plot the data again. Thenn, I will check for correlations between the features and the target with .corr()followed by using .unique() to get the unique values. Then to further look at the relationship between ground living area and sales price with scatter plots as well as with garage area.

1. Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner

There is a positive correlation between sales price and overall quality, ground living area, garage cars, and garage area. Using a pivot table as well as a scatter plot, I visualized the correlation between sales price and overall quality.

Using feature engineering and plotting a bar plot of the correlation between street features and sales price. This shows that partial has a significantly higher median sales price.

1. Formulating at least 3 hypotheses about this data

- my hypotheses are the larger the house the higher the sales price

- the better the overall quality the higher the sales price

- the larger the lot area ground living area the higher the sales price

1. Conducting a formal significance test for one of the hypotheses and discuss the results

Testing hypothesis of the influence of lot area on sales price

H0: the earlier built date of the house result in the lower house price

H1: the earlier built date of the house does not result in the lower house price

Alpha to be set to 0.05, which is a 95% confidence level.

I'll calculate a Pearson correlation coefficient and the p-value for testing correlation from scipy.

The result of the p-value suggest we should reject the null hypothesis.

1. Suggestions for next steps in analyzing this data

Building a linear model, that will separate the features and the target variable for modeling.

1. A paragraph that summarizes the quality of this data set and a request for additional data if needed Because this is a data set from Kaggle intended for beginners, I found the data to be well organized and complete.

My code can be found at: <https://www.kaggle.com/code/ryanjt/house-prices>