Dat Dao

Final Project: Data Exploration and Analysis

This dataset contains information about residential properties in Ames, Iowa, including various characteristics of the homes such as living area, number of bedrooms and bathrooms, garage size, overall quality and condition, year built and remodeled, and sale price.

The dataset contains a total of 16 columns, including an ID column, 13 numerical columns, and 2 categorical columns (Neighborhood, Condition1, BldgType, and HouseStyle). The target variable is SalePrice, which represents the property’s final sale price in US dollars.

Some of the key features in this dataset that are often considered essential factors in determining the value of a residential property include living area (GrLivArea), overall quality and condition (OverallQual and OverallCond), number of bedrooms and bathrooms (BedroomAbvGr, FullBath, HalfBath, and BsmtFullBath), garage size (GarageCars), and year built and remodeled (YearBuilt and YearRemodAdd).

This dataset could be used to build a regression model to predict the sale price of a property based on its various characteristics. Additionally, it could be used to explore the relationships between different features and the sale price of a property and to gain insights into the factors that affect property values in the Ames, Iowa, area.

There is no difference between the price for conditions 6 and 7. The price for condition 5 is spread out between 100k-200k. Since the price for 6 and 7 conditions is concentrated between 100k-175k. The trend is that the condition overall higher seems to be sold at a higher price as the chart shifted to the right more.

Based on the scatter plots, we observe that many houses were sold for under 300,000 and had an area of under 3,000 sqft. Although there appears to be a trend that larger houses are sold at higher prices, this relationship is not entirely clear since several other factors affect house prices. However, we can deduce that there is a positive correlation between the quality of a house and its sale price. Houses with a higher quality rating are generally sold at higher prices. Specifically, the sale prices for houses with quality ratings of 5, 6, and 7 range from around 80,000 to 350,000 and are relatively consistent.

The Pearson correlation coefficient is a measure of the strength and direction of the linear relationship between two variables. Pearson's correlation coefficient between GrLivArea and SalePrice of 0.5893156056586915 indicates a moderately strong positive linear relationship between the two variables. A positive correlation means that as one variable increases, the other variable also tends to increase. In this case, as a house's living area (GrLivArea) increases, the house's sale price (SalePrice) also tends to increase. The magnitude of the correlation coefficient, which ranges from -1 to 1, indicates the strength of the relationship. A coefficient of 0 indicates no linear relationship, a coefficient of 1 indicates a perfect positive linear relationship, and a coefficient of -1 indicates a perfect negative linear relationship. A coefficient of 0.5893156056586915 indicates a moderately strong positive linear relationship, which means that GrLivArea and SalePrice are positively correlated. However, there may still be some unexplained variability in the data that is not captured by the linear relationship.

A value of 0.6258949464633632 indicates a positive monotonic relationship between the two variables, where higher values of GrLivArea tend to be associated with higher values of SalePrice. The second value returned by the spearman function (p\_value) is the associated p-value, which is a measure of the statistical significance of the correlation coefficient. In this case, the p-value is 2.260393e-317, which is very small (close to zero), indicating that the correlation coefficient is statistically significant and unlikely to have occurred by chance. This suggests that there is a genuine and significant relationship between the GrLivArea and SalePrice variables based on their ranks.

R-squared: This is the coefficient of determination, which measures the proportion of variance in the response variable explained by the predictor variables. In this case, the R-squared value is 0.452, which means that the predictor variables explain 45.2% of the variance in the sale price.

Adj. R-squared: This is the adjusted R-squared value, which considers the number of predictor variables in the model. In this case, the adjusted R-squared value is 0.449.

F-statistic: This is the F-statistic for the overall significance of the model. It tests whether the predictor variables as a group significantly affect the response variable. The F-statistic value is 171.2, and the p-value for the F-test is 0.00, which means that the model is statistically significant.

Coefficients: These are the estimated regression coefficients for each predictor variable. The coefficient represents the change in the response variable (sale price) associated with a one-unit increase in the predictor variable, holding all other variables constant. The intercept (const) coefficient is -21010, which means that the predicted sale price when all predictor variables are equal to zero is $21,010. The other coefficients represent the change in the sale price associated with a one-unit increase in each predictor variable.

Standard errors: These are the standard errors of the regression coefficients, which measure the variability of the estimated coefficients. The const coefficient has a t-value of -0.488 and a p-value of 0.625, indicating that it is not statistically significant at the conventional alpha level of 0.05. This means that the intercept term is not significantly different from zero, and the regression line is not significantly different from intersecting the y-axis at zero.