House Prices

Advanced Regression Techniques EDA (Kaggle)

Ritesh Kumar

2024WI\_MS\_DSP\_422-DL\_SEC61: Practical Machine Learning

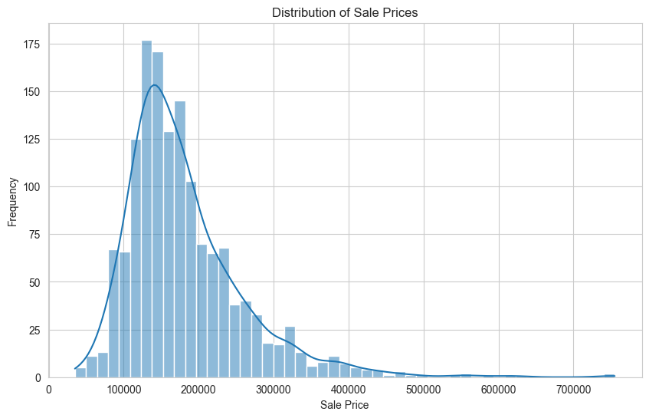
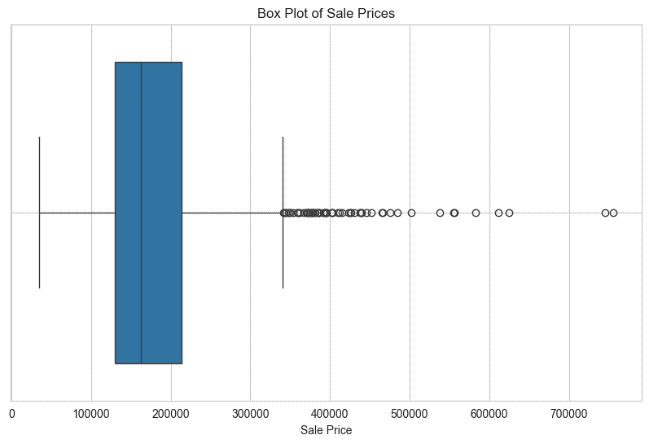
Module 1 Assignment

# Exploratory Data Analysis

# Donal Wedding and Narayana Darapaneni

# January 7, 2024

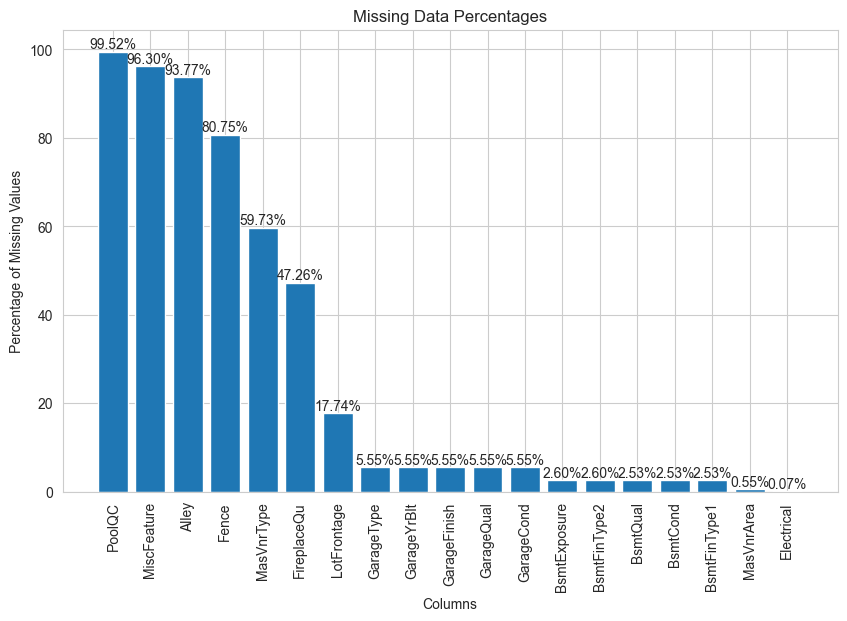
1. **Provide appropriate descriptive statistics and visualizations to help understand the marginal distribution of the dependent variable.**

The distribution of SalePrice is right-skewed with a concentration of values in the lower range and fewer high-value outliers, as shown in the histogram and boxplot. The median price is lower than the mean, which is elevated by these high-value sales. Descriptive statistics reveal a broad range in sale prices, with a mean of approximately $180,921 and a notable standard deviation of $79,442, indicating varied house prices within the dataset.

1. **Investigate missing data and outliers.**

Missing Values:



High Missing Values Approach (Above 40%): For features with very high missing values like PoolQC, MiscFeature, and Alley, it's effective to transform them into binary indicators that signal the presence or absence of the feature. However, for those with over 50% missing data, and where it's unclear if the data is missing or unrecorded, we have decided to remove these features from the analysis to avoid skewing results with unreliable data.  
Handling Moderate and Low Missing Values (Below 40%): For features with moderate to low missing data, we have used median and mode imputation to fill in the gaps in continuous and discrete features respectively, ensuring a straightforward and effective approach to maintain data integrity for our analysis.

Outliers:

The summary of outlier counts across various features shows that several columns have a significant number of outliers, exceeding 10% of the data. `RoofStyle` leads with 319 outliers, followed by `MSZoning` with 309 outliers. `SaleCondition` and `BldgType` also have a high number of outliers at 262 and 240, respectively. Other features with notable outliers include `BsmtExposure`, `Condition1`, `EnclosedPorch`, and `SaleType`, with over 190 outliers each. `LandContour`, `ExterCond`, `BsmtFinType2`, and `BsmtFinSF2` show fewer outliers but still exceed the 10% threshold. These outlier counts suggest significant variation in these attributes, which may require attention during data preprocessing for predictive modeling.

1. **Investigate at least three potential predictors of the dependent variable and provide appropriate graphs / statistics to demonstrate the relationships.**

The correlation heatmap indicates that `GrLivArea` has a strong positive correlation with `SalePrice`, suggesting significant influence on sale prices, while `GarageArea` and `YearBuilt` show moderate positive correlations, implying that larger garages and newer homes tend to sell for higher prices. `MasVnrArea` has a weaker positive relationship with `SalePrice`, and the correlations among the features themselves are relatively low, suggesting that they contribute independently to predicting `SalePrice`. These insights can be instrumental in developing a predictive model for housing prices.

A screenshot of a graph

Description automatically generatedA graph of a graph

Description automatically generated with medium confidenceA graph with a blue rectangular bar

Description automatically generated

A graph of a graph of a person

Description automatically generated with medium confidenceA graph with a blue rectangular bar

Description automatically generated

A graph of a graph of a person

Description automatically generated with medium confidenceA graph with a blue rectangle

Description automatically generated

A graph with blue bars

Description automatically generatedA graph with a line graph

Description automatically generated with medium confidence

1. Histograms and Boxplots of `GrLivArea` and `GarageArea` indicate that most of the houses have a living area and garage area within a moderate range, but there are a few houses with exceptionally large areas.
2. Histogram and Boxplot of `YearBuilt` suggest that the dataset includes many newer homes, with fewer older homes.
3. Histogram and Boxplot of `MasVnrArea` suggest that the typical house has a modest amount of masonry veneer, with a few houses having significantly more.

A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated

A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated

1. SalePrice vs GrLivArea: There's a positive correlation between `GrLivArea` (above-grade living area square feet) and `SalePrice`. As the living area increases, the sale price tends to increase as well.
2. SalePrice vs YearBuilt: There's a general trend indicating that newer houses tend to sell for higher prices, but the relationship is not as strong or linear as with `GrLivArea`.
3. SalePrice vs GarageArea: A moderate positive correlation exists between `GarageArea` and `SalePrice`, with some spread in the data points.
4. SalePrice vs MasVnrArea: suggests a slight positive trend, the relationship is not as strong, and there is a lot of spread in the sale prices for houses with small to moderate masonry veneer areas.
5. **Engage in feature creation by splitting, merging, or otherwise generating a new predictor.**

To consolidate the house size-related features into a single variable, we sum the following attributes: ‘GrLivArea’ (above-grade living area square feet), ‘1stFlrSF’ (first floor square feet), ‘2ndFlrSF’ (second floor square feet), ‘BsmtFinSF1: Type 1 finished square feet’, and ‘LowQualFinSF’ (low-quality finished square feet across all floors). This new aggregated variable represents the total living area of the house in square feet, capturing the comprehensive size across all floors and quality levels.

A screenshot of a graph

Description automatically generated

The correlation of TotalSF is 0.75 with SalePrice, whereas that of GrLivArea was 0.71. When comparing to the other three variables, the corrleation of TotalSF has slightly increased vis-a-vis GrLivArea.

A graph of blue bars

Description automatically generated A graph with a bar and a line

Description automatically generated with medium confidence

The `TotalSF` variable, representing the total square footage of houses, exhibits a right-skewed distribution with most houses clustered in the moderate square footage range but with a long tail of outliers indicating some houses have significantly larger areas. The median square footage is lower than the mean, which is skewed upwards by these larger properties, and the considerable standard deviation reflects a wide variance in house sizes. The presence of outliers suggests that while `TotalSF` could be a strong predictor for `SalePrice`.

A graph with blue dots

Description automatically generated

The scatter plot depicting `SalePrice` versus `TotalSF` indicates a positive correlation between the total square footage of houses and their sale prices. As the total square footage increases, there tends to be an increase in sale price, suggesting that larger homes generally sell for more. The plot also shows a concentration of data points in the lower to mid-range of total square footage, with fewer homes on the larger end, reflecting the right-skewed distribution observed in the histogram. There are outliers with both unusually high square footage and sale prices, which could influence any predictive modeling efforts and might need to be addressed. This trend is consistent with real estate market expectations where the size of a property is a key determinant of its value.

1. **Using the dependent variable, perform both min-max and standard scaling in Python.**

Before min-max scaling, the features in the dataset display a wide range of values with varying degrees of spread, as indicated by their standard deviations, particularly `TotalSF` and `SalePrice` showing large spreads. After scaling, all features are uniformly transformed to a [0, 1] scale, which normalizes their distributions for comparability and suitability for scale-sensitive algorithms, while preserving their inherent distribution characteristics. The rescaling is evident in the normalized mean values and standard deviations, which are now proportionally smaller, reflecting the unified scale across all features.

A graph of a graph with lines and numbers

Description automatically generated with medium confidence

In the right plot (after scaling), all features are now on the same [0, 1] scale, and their distributions can be easily compared. The medians, quartiles, and outliers are now visible on a consistent scale, which is especially important for algorithms that depend on the distance between data points. Outliers remain visible post-scaling, but they no longer dominate the scale of the plot. This normalization enables a more meaningful comparison across features and is necessary for many machine learning models to perform optimally.

**Code and Output**