Ames Housing Dataset: An Exploration Journey

Background

As a home buyer, there are many factors that influence the decision-making process; it could be the number of bedrooms, the height of the basement ceiling, or even the proximity to an east-west railroad. In this assignment, we are diving deep into the Ames Housing dataset, which contains information on residential homes in Ames, Iowa. The dataset features a rich set of 79 explanatory variables describing almost every aspect of a residential home.

Your task is to explore this dataset, uncover patterns, and draw insights using your knowledge of Python and the Pandas library. Feel free to consult the accompanying data description file to understand the fields better. Below, you will find a brief description of some of the fields in the dataset:

- `SalePrice`: The property's sale price in dollars (the target variable).
- `MSSubClass`: The building class.
- `MSZoning`: The general zoning classification.
- `LotFrontage`: Linear feet of street connected to the property.
- `LotArea`: Lot size in square feet.
- `Street`: Type of road access.
- (Refer to the data description file for definitions of all other fields)

Problem Set

Problem 1: Dataset Overview

1.1 Load the Ames Housing dataset into a Pandas DataFrame. Display the first 5 rows to understand the initial structure of your data.
1.2 How many columns are there in the dataset? What kind of information does the dataset contain?
1.3 Are there any missing values in the dataset? How can you identify and count the missing values in each column?
1.4 What are the top 5 neighborhoods with the highest average `SalePrice`?
1.5 Determine the proportion of houses that have a swimming pool (`PoolArea` > 0).
1.6 Can you find out if there is a relationship between the overall quality (`OverallQual`) of a house and its sale price (`SalePrice`)?

Problem 2: Data Access and Manipulation

2.1 Can you extract the information on the `OverallQual` and `OverallCond` of houses built in the year 2000 and after?
2.2 What is the average lot area (`LotArea`) and sale price (`SalePrice`) of houses in different neighborhoods (`Neighborhood`)?

2.3 Can you find and display the details of the house with the highest sale price?
2.4 Identify the houses that have more than 2 fireplaces. How does the number of fireplaces affect the sale price?
2.5 How many houses have a garage area (`GarageArea`) larger than 800 square feet? What is their average sale price?

Problem 3: Data Cleaning

3.1 The `Alley` column has many missing values. What strategies can you employ to handle these missing values? Would you remove the column or replace missing values? Justify your choice.
3.2 Identify columns with a high percentage of missing values (e.g., more than 80%). What would you suggest doing with these columns, and why?
3.3 Identify columns that have a single unique value and remove them from the DataFrame. What is the rationale behind removing such columns?

Problem 4: Aggregation and Grouping

4.1 What is the median sale price of houses with different building types (`BldgType`)? How does the building type affect the sale price?
4.2 Can you find out which year had the highest number of houses sold? How many houses were sold that year?
4.3 Can you group the houses by neighborhood and building type and then find the average sale price for each group?
4.4 Find out the year with the highest and lowest number of sales. Can you explain any trends or patterns that you observe?

Problem 5: Deep Dive with Pivot Tables

5.1 Create a pivot table to analyze the relationship between the overall quality (`OverallQual`) and the sale condition (`SaleCondition`) with regards to the average sale price. What insights can you gather from this table?
5.2 Can you create a pivot table that shows the median sale price for each combination of the overall condition (`OverallCond`) and the year the house was built (`YearBuilt`)?
5.3 Create a pivot table that shows the median `LotArea` for each combination of neighborhood and lot configuration (`LotConfig`). What insights can this table provide?
5.4 Using a pivot table, analyze the sale price trend over the years. What observations can you make about the sale price trends?

Problem 6: Visualization and Styling

6.1 Can you create a correlation matrix of the numerical columns and then visualize this matrix using a heatmap? What do you observe from this heatmap regarding the relationships between different variables?
6.2 Visualize the trend of average sale price over the years. Can you identify any patterns or trends in the data?
6.3 Create a scatter plot to visualize the relationship between the ground living area (`GrLivArea`) and the sale price. What can you infer from this scatter plot?
6.4 Can you create a boxplot to visualize the distribution of sale prices in different neighborhoods?