Housing Price Analysis

By: Muhammad Fakhri Azhar

Data Analyst Project | 2025

Introduction

Hi! I’m Muhammad Fakhri Azhar, a physics graduate with a strong passion for data analysis. This project is part of my learning journey in turning data into insights.

Course License:

- Data Science Bootcamp @Kelas Work by Kelas.com
- Data Analyst Mini Course @RevoU
- Ms.Excel Short Class @MySkill
- Computer Training @FMIPA UNNES

Contact Info

Email : mfkriazh57@gmail.com
Phone : 0857-2454-9367
LinkedIn : Muhammad Fakhri Azhar
Portfolio : Click Here
GitHub : mfakhriazhar

Project Code Details on Github :
https://github.com/mfakhriazhar/housing-price-analysis/blob/main/Case_02_Final_Project.ipynb

Overview

Determining the price of a house also depends on various factors such as building area, exterior quality, and amenities. The Housing Price dataset provides information on properties for sale, and through Exploratory Data Analysis (EDA), patterns and key factors affecting house prices can be identified. The expected outcome is a better understanding of the main factors that determine house prices, as well as how data distribution patterns can help in making property price predictions or recommendations.

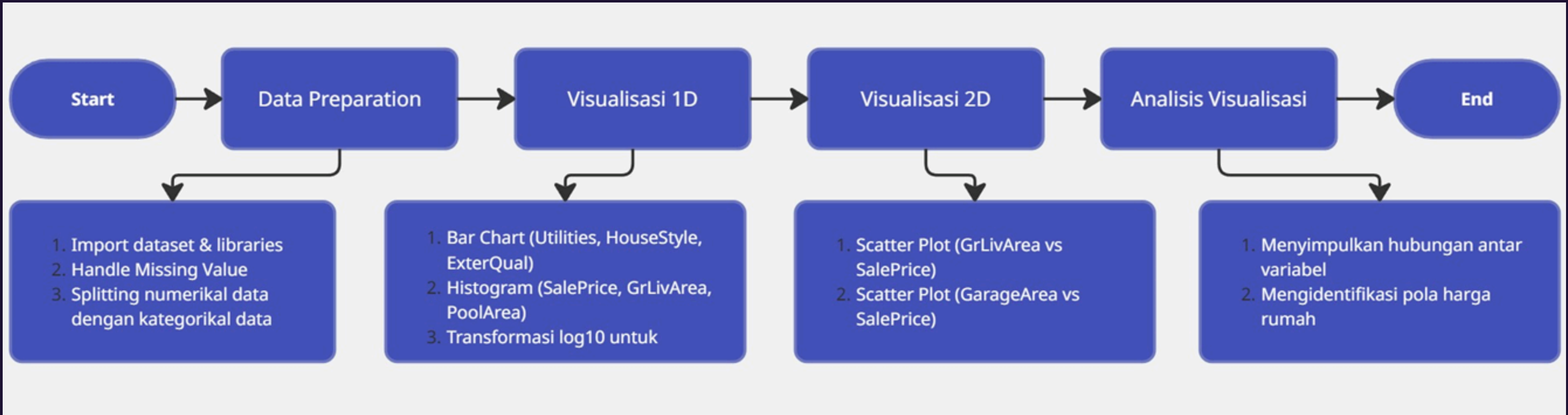
Dataset Link :

https://github.com/mfakhriazhar/housing-price-analysis/blob/main/train_house.csv

Project Goals

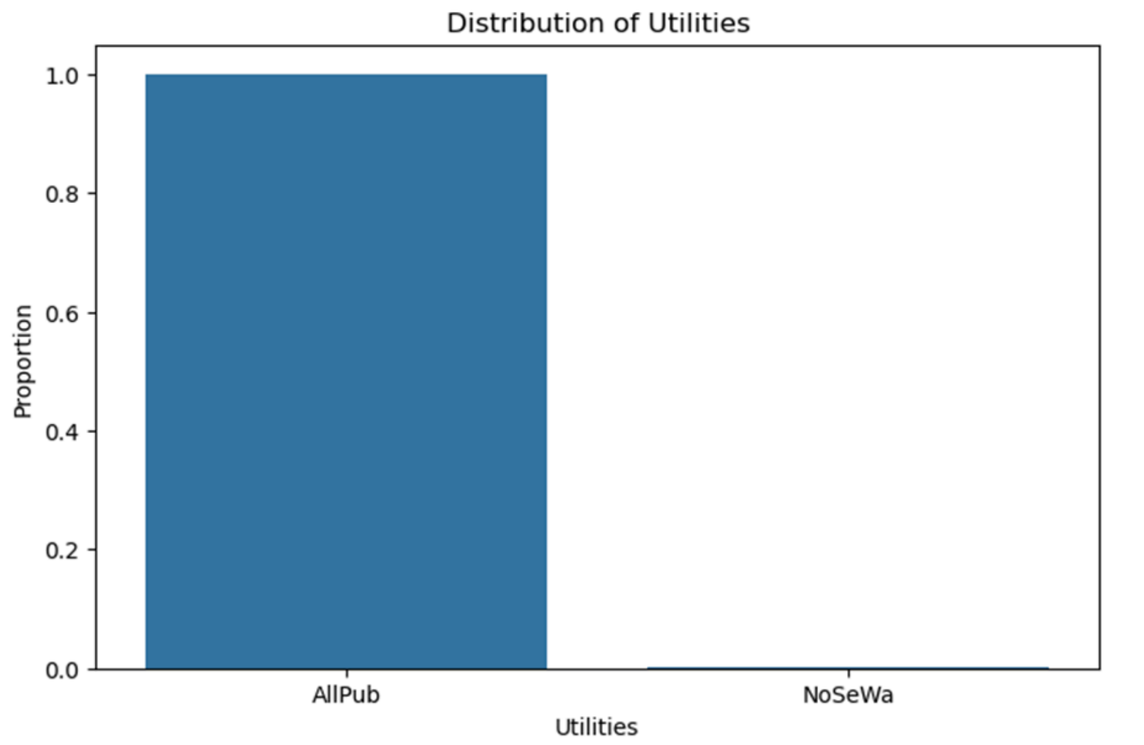
- Cleaning and preparing data, including handling missing values (NaN).
- Analyzed the distribution of numerical and categorical variables using 1D visualizations.
- Analyze relationships between variables using 2D visualizations to see how certain factors affect house prices.
- Gain insights that can help in decision-making related to property valuation.

Flowchart



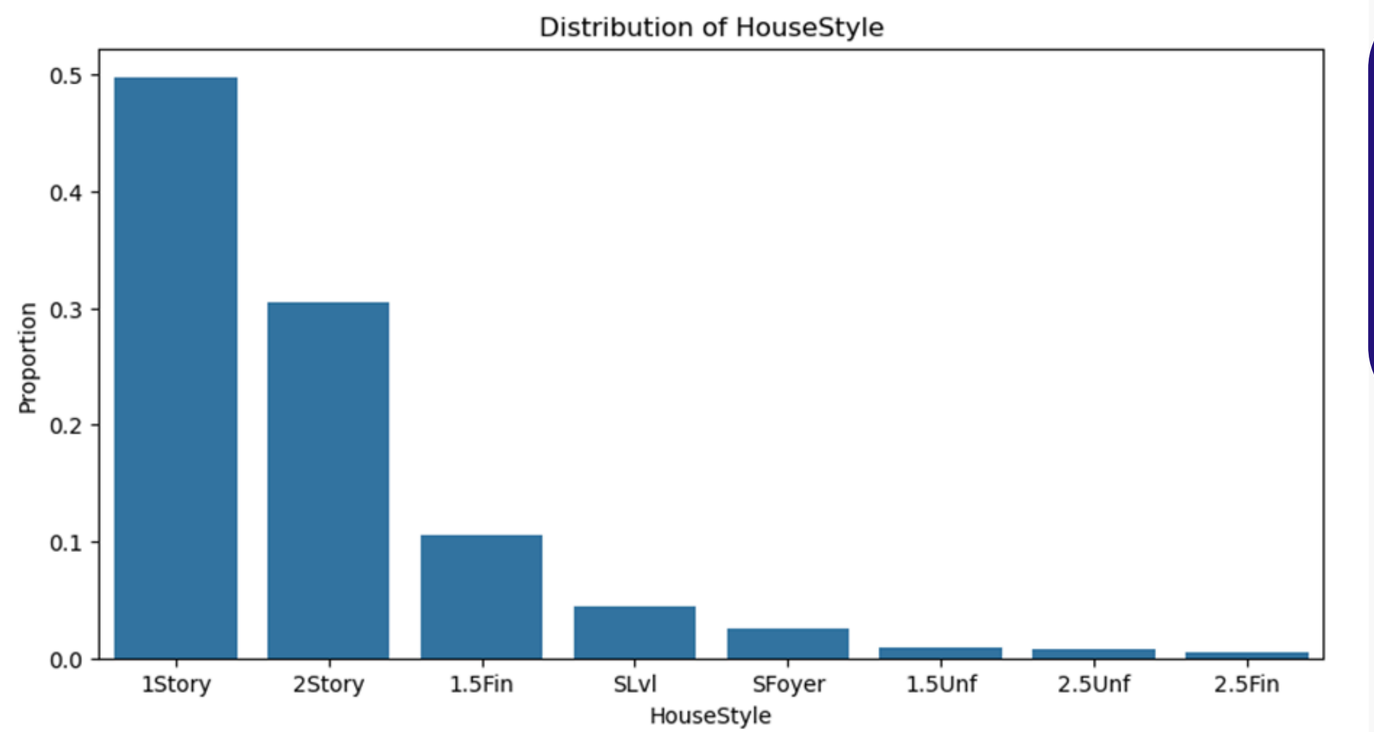
Stages of housing price data analysis from data preparation to visualization interpretation.

Utilities Distribution



From the Distribution of Utilities bar chart, it can be seen that almost all houses have All public Utilities (AllPub) facilities, while houses with Electricity and Gas Only (NoSeWa) facilities are very rare. Due to the unbalanced distribution, this variable is less informative and could be considered for deletion or reduction to a binary variable.

House Style Distribution

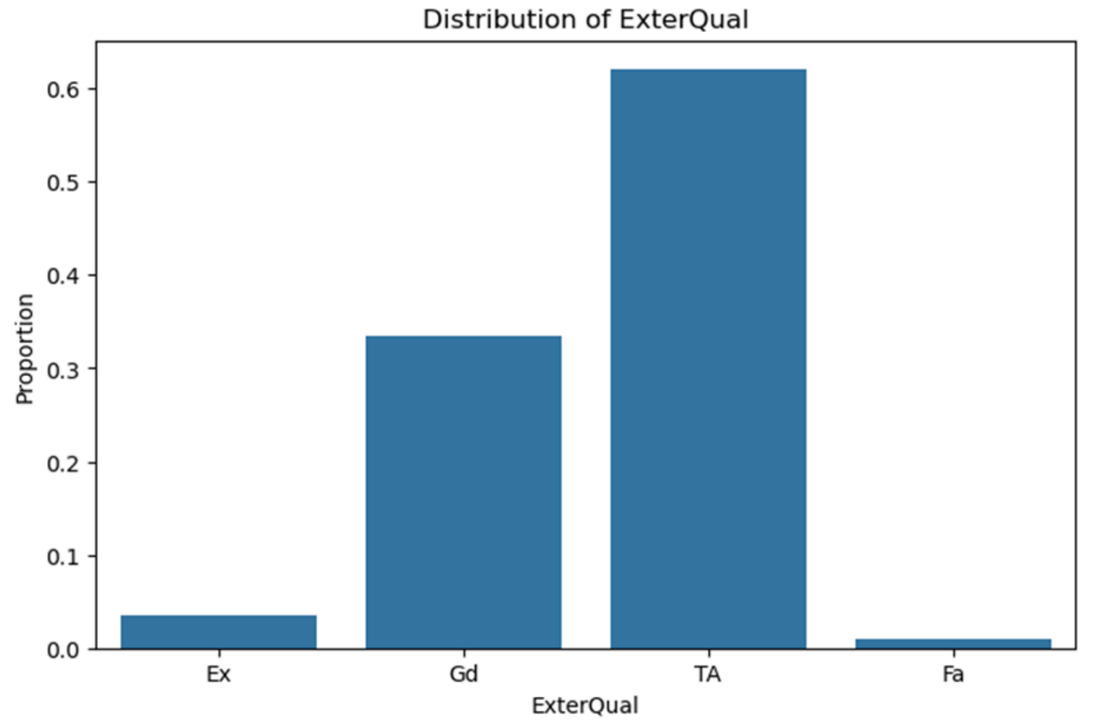


HouseStyle: Style of property (e.g. 1-story, 2-story, etc.)

- 1Story : One story
- 1.5Fin : One and one-half story: 2nd level finished
- 1.5Unf : One and one-half story: 2nd level unfinished
- 2Story : Two story
- 2.5Fin : Two and one-half story: 2nd level finished
- 2.5Unf : Two and one-half story: 2nd level unfinished
- SFoyer : Split Foyer
- SLvl : Split Level

For the Distribution of House Style bar chart, it is noted that, the majority of houses are of the “1Story” and “2Story” types, covering more than 80% of the dataset. Other house styles such as “1.5Unf”, “2.5Fin”, and “2.5Unf” are very rare, indicating that split-level houses are less common.

ExterQual Distribution

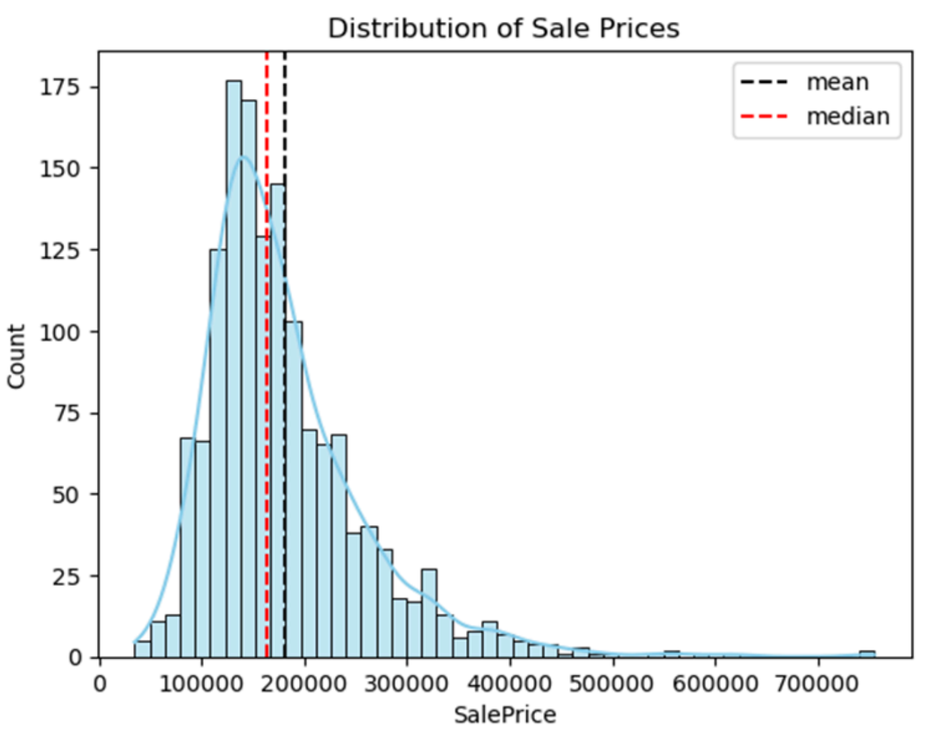


ExterQual: exterior material quality

- Ex Excellent
- Gd Good
- TA Average/Typical
- Fa Fair

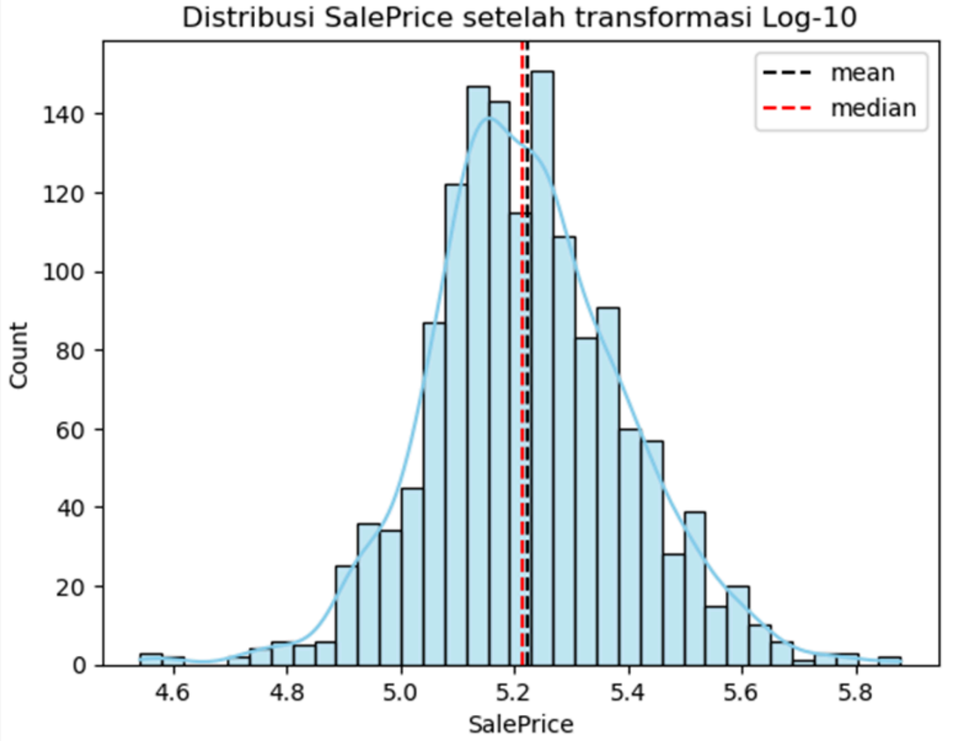
As for the Distribution of ExterQual bar chart, most of the houses have exterior qualities of “TA” (Typical/Average) and “Gd” (Good), while ‘Ex’ (Excellent) and “Fa” (Fair) are very rare. This shows that most houses have standard exterior materials, with few being excellent or poor.

Sale Price Distribution



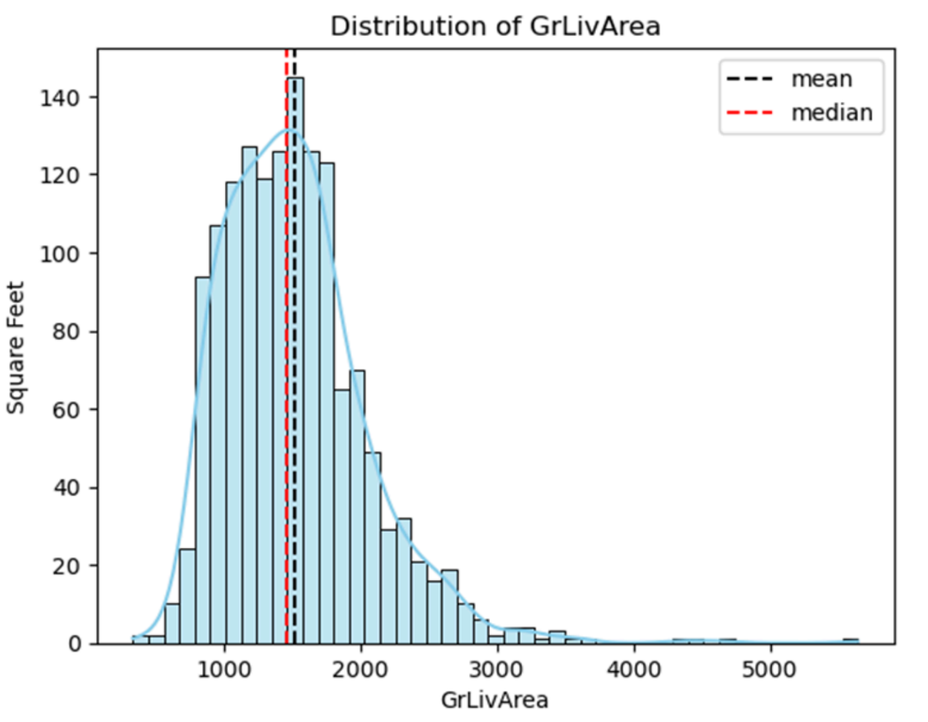
From the Distribution of Sale Prices histogram, it can be seen that the distribution of house prices is skewed to the right, with a mean value (180,921) greater than the median (163,000), indicating outliers at high prices. Most homes are in the 100,000-200,000 range, with a standard deviation of 79,442, indicating a large variation in prices. As the median is more representative, further analysis may consider data transformation to deal with skewness.

Sale Price Distribution



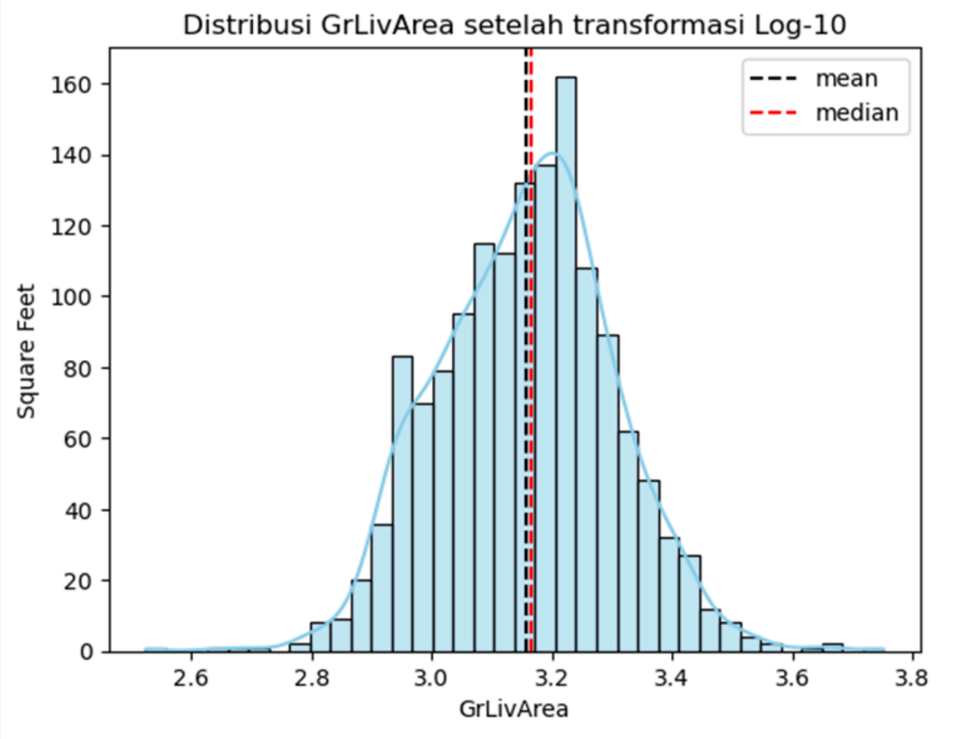
After log10 transformation of SalePrice, the mean (5.22) and median (5.21) values almost coincide, indicating a more symmetrical distribution than before transformation. The smaller standard deviation (0.17) also indicates that the data variation is better controlled.

Ground Live Area Distribution



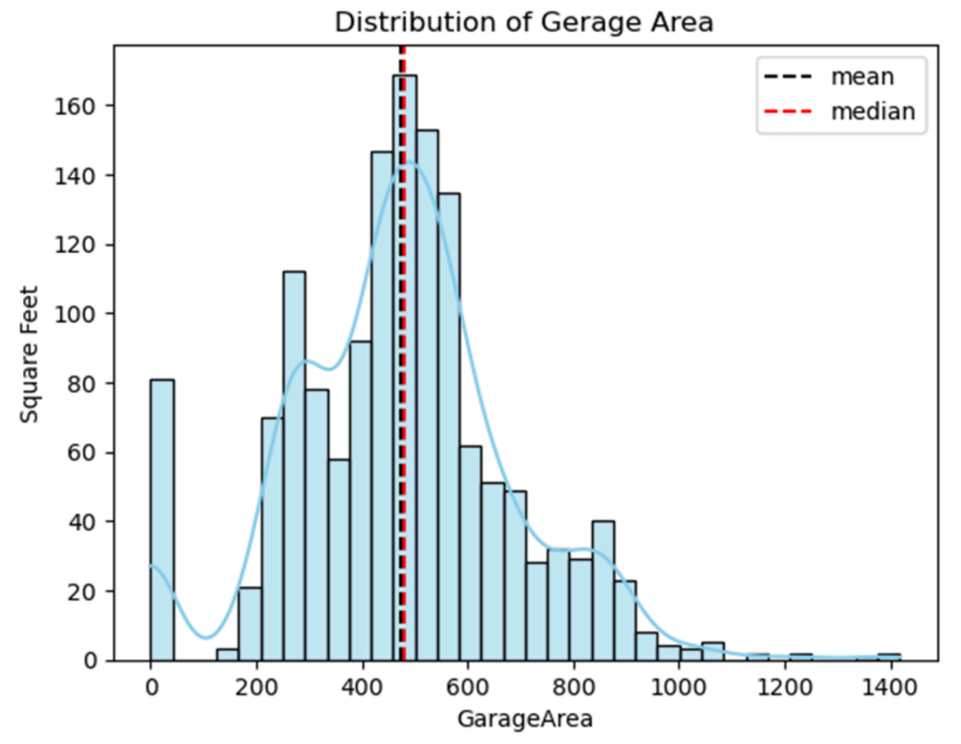
Histogram Distribution of GrLivArea shows the distribution of living area above ground in square feet. The distribution is right-skewed, with the mean value greater than the median. This shows that there are a few homes with much larger areas than the majority, which may be outliers. Most of the houses have an area of around 1,000 - 2,000 square feet, as seen from the peak of the distribution.

Ground Live Area Distribution



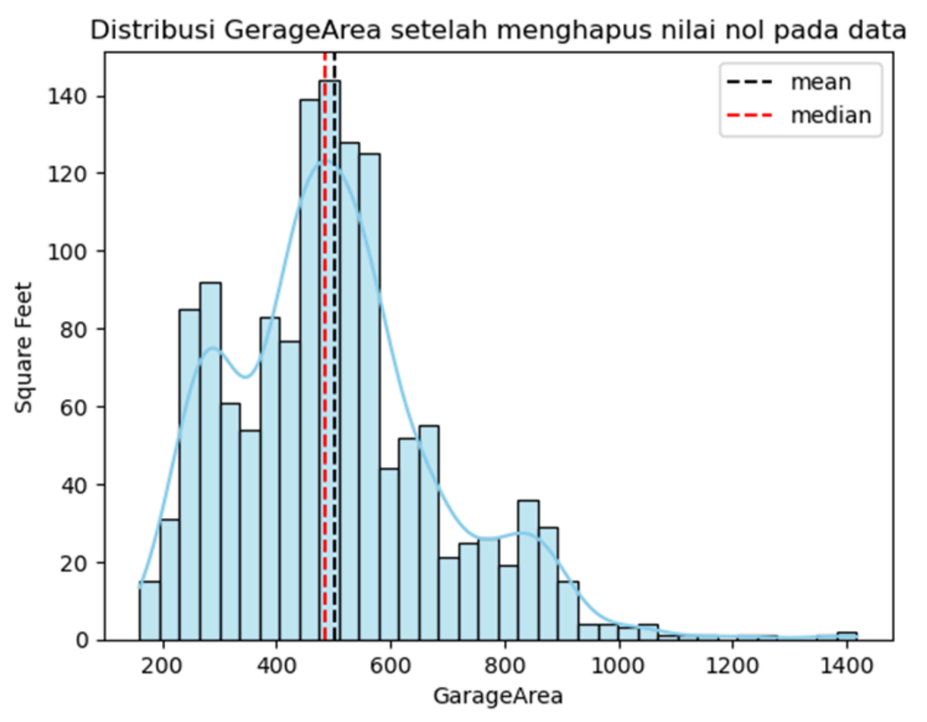
Just like SalePrice, the distribution of GrLivArea after log10 transformation is also more symmetrical with a small difference between the mean (3.16) and median (3.17). This shows that the transformation successfully reduces skewness and normalizes the data.

Gerage Area Distribution



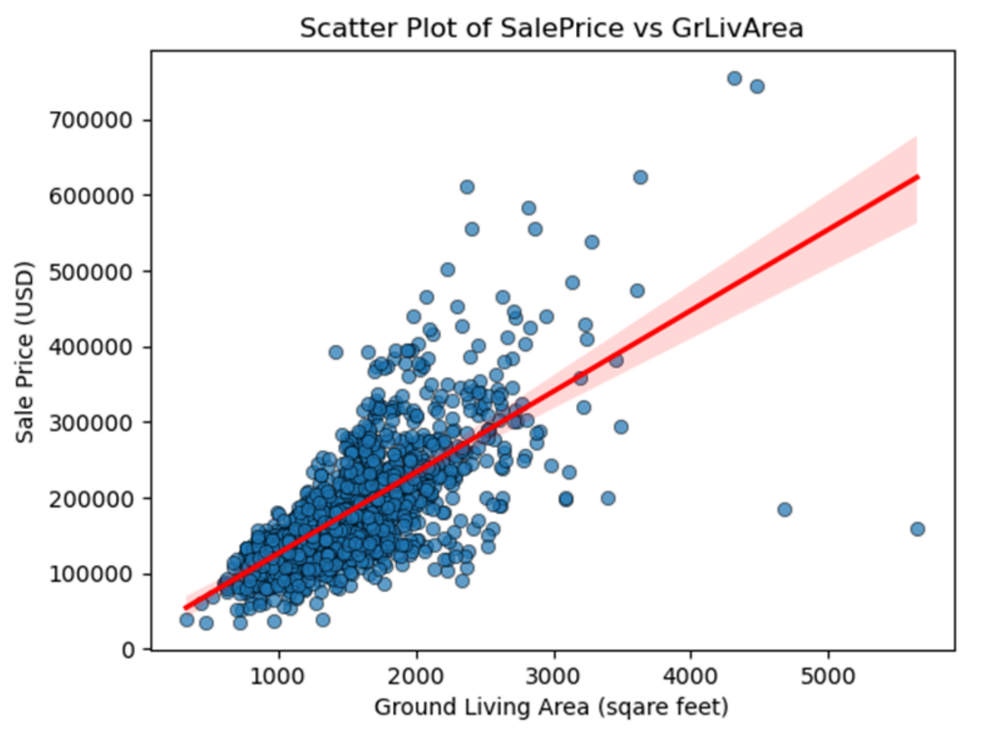
The GarageArea histogram shows the distribution of garage area in square feet. The distribution is skewed to the right, as the mean is larger than the median although the values are only slightly different but has a long tail to the right, indicating the presence of some very large garages as outliers. Most of the houses have garages of around 400 - 600 square feet, which reflects the general capacity for 1-2 cars. In addition, there are 81 homes that do not have garages out of all the homes in the dataset.

Gerage Area Distribution



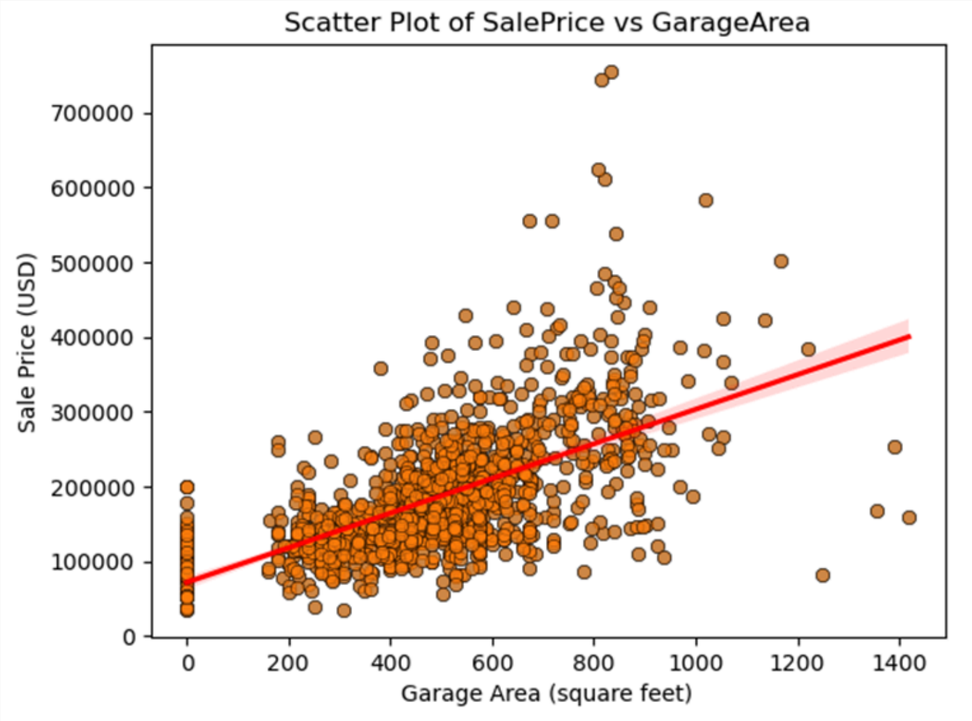
For the GarageArea distribution, after removing the zero values, the distribution remains slightly skewed to the right, with the mean (500.76) being greater than the median (484). However, compared to before, the distribution is more centered and has more controllable variation (standard deviation 185.69). Overall, the log10 transformation is effective in dealing with skewness in SalePrice and GrLivArea, while for GarageArea, additional steps such as transformation or outlier handling may be needed to normalize the distribution.

Sale Price vs Ground Live Area



The scatter plot above shows a positive relationship between Ground Living Area (GrLivArea) and Sale Price (USD), where the larger the living area, the higher the sale price of the house. The red regression line indicates a linear trend, with the pink band being the level of uncertainty. Some outliers are visible, such as large homes with low prices or vice versa, which may be influenced by other factors such as location and building quality.

Sale Price vs Gerage Area

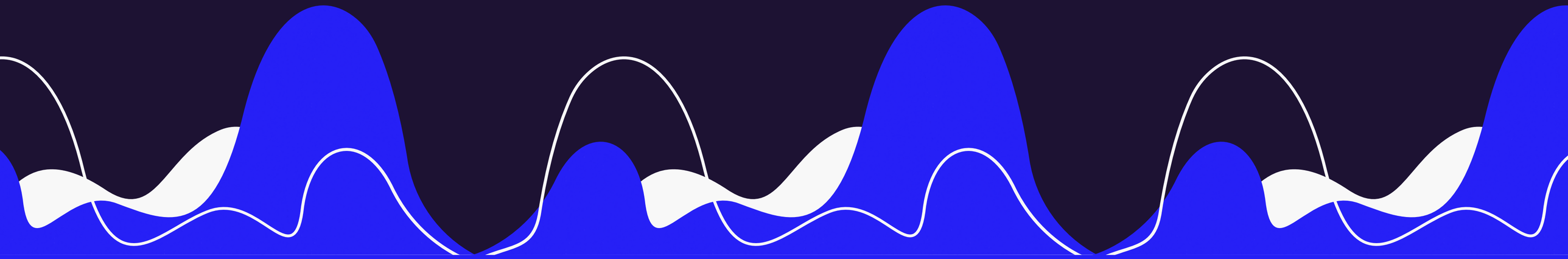


While this scatter plot shows a positive relationship between Garage Area and Sale Price (USD), where the larger the garage, the higher the house price. The red regression line shows the trend in the scatter plot with pink uncertainty bands. However, the correlation appears weaker compared to GrLivArea vs SalePrice, as the data points are more spread out. This is understandable as the main space of the house is a major factor in determining the comfort and value of the property, while the garage, although important, is not a major factor in determining the sale price of the house.

Conclusion

- Analysis of the Housing Price dataset shows that Ground Living Area (GrLivArea) has a strong positive correlation with house price (SalePrice), making it a major factor in determining property value.
- Garage Area is also influential, although the correlation is weaker. The distribution of house prices tends to be right-skewed, with some outliers in the form of houses that are priced very high or low relative to their size and features.
- Other factors such as exterior quality (ExterQual) also contribute significantly to price, while Utilities have a smaller influence.
- The log-10 transformation helps normalize the price distribution for more accurate analysis.
- Overall, house prices are influenced by a combination of square footage, amenities, and other external factors, so a holistic approach is required in property analysis.

Thank You!



Muhammad Fakhri Azhar

Data Analyst Project | 2025