Kevin Kiding

Housing Price Prediction

- Background:

  Predicting home prices is essential for real estate transactions, as it enables informed decisions, financial planning, and market analysis. Accurate forecasts enable buyers and vendors to confidently navigate complex markets, optimizing their decisions and strategies.

- Objective:

  Explore and analyze prepared data using EDA to uncover patterns, relationships, and insights. It aims to enhance decision-making regarding house prices and associated attributes. Through data visualization and analysis, it provides valuable insights into categorical and numerical variables, facilitating the subsequent stages of analysis.

Introduction

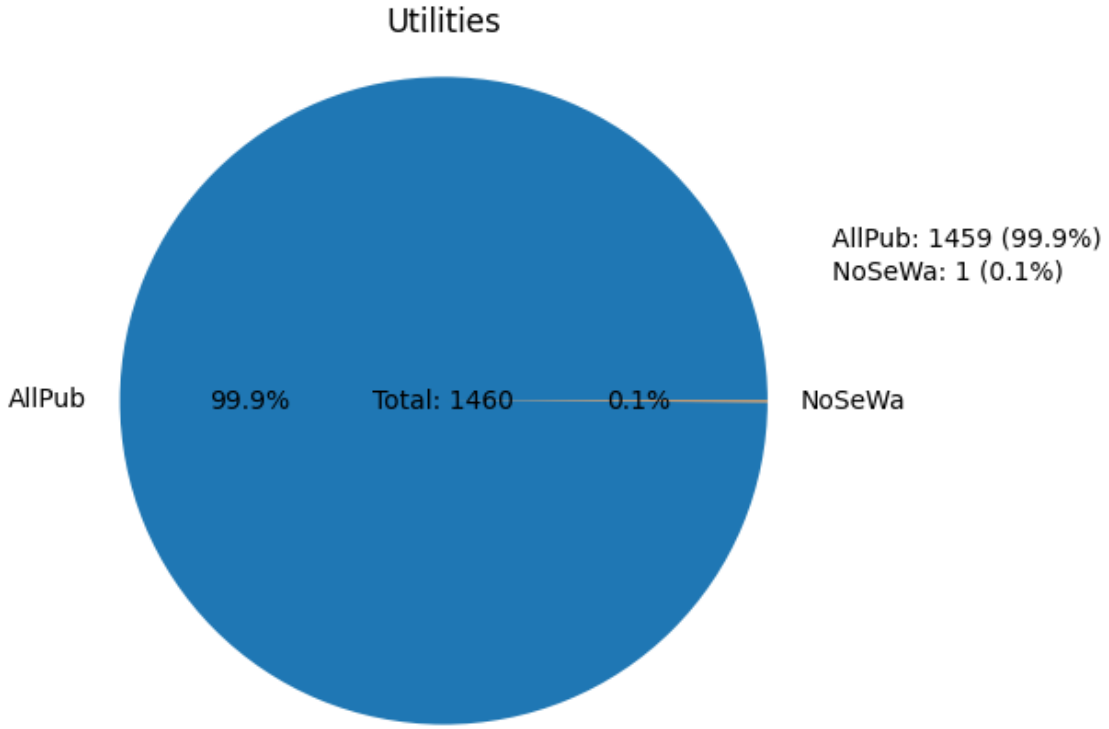
Here I would like to share a Google Colab link to access the data processing steps that I have done. You can click on this link to view and explore the steps involved in the data processing, including the code snippets and explanations used. This will allow you to gain insight into the data processing techniques applied and understand the overall workflow. Please access the links at your convenience.

https://drive.google.com/file/d/1kOWgWET2O_jJ3rewv10FHifSQBB2UOLY/view?usp=sharing

Thank you for your attention

Utilites

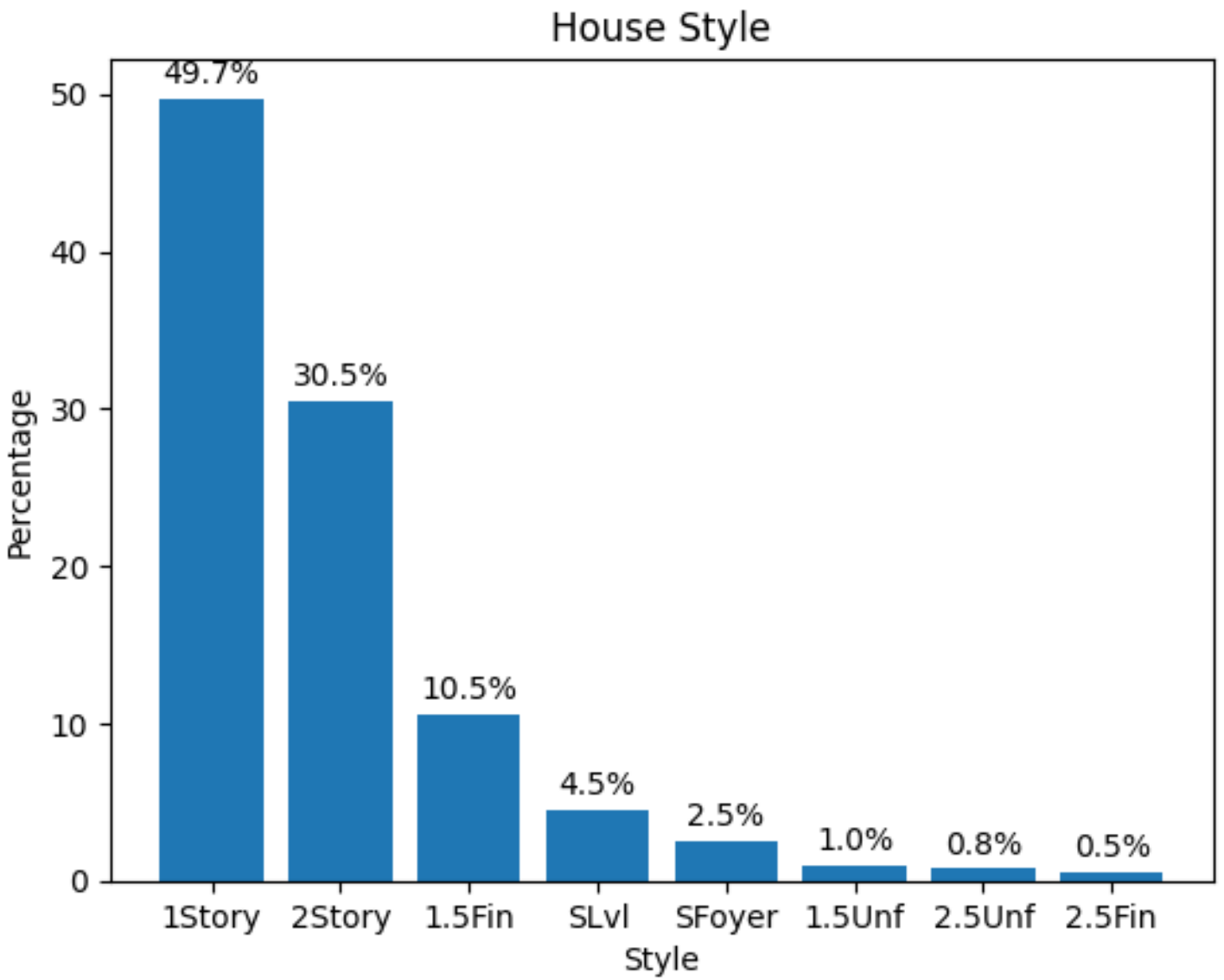
- Contains information about what facilities are available at the property.
  - AllPub: This indicates that the property has access to all common public utilities, including Electricity (E), Gas (G), Water (W), and Sanitation (S).
  - NoSeWa: This indicates that the property has only limited access to Electricity (E) and Gas (G). Properties with this value may not have access to public water supply and sanitation systems.



- The majority of properties (99.99%) have an 'AllPub' utility type, indicating full access to all public utilities.
- There is only one property (0.01%) with a utility type of 'NoSeWa', indicating limited access to utilities.
- This analysis provides an understanding of the distribution of utility types in an area and provides insights to homebuyers, property developers, and policymakers on the availability and accessibility of essential services.

House Style

HouseStyle is a column that describes the style or type of house that each property has. This data provides information on the number and percentage of houses with different styles, such as one-story, two-story, half-story, and split-level houses. Knowledge of these variations in house styles can provide insight into architectural preferences within the property market, thus aiding decision-making for businesspeople in the real estate industry.

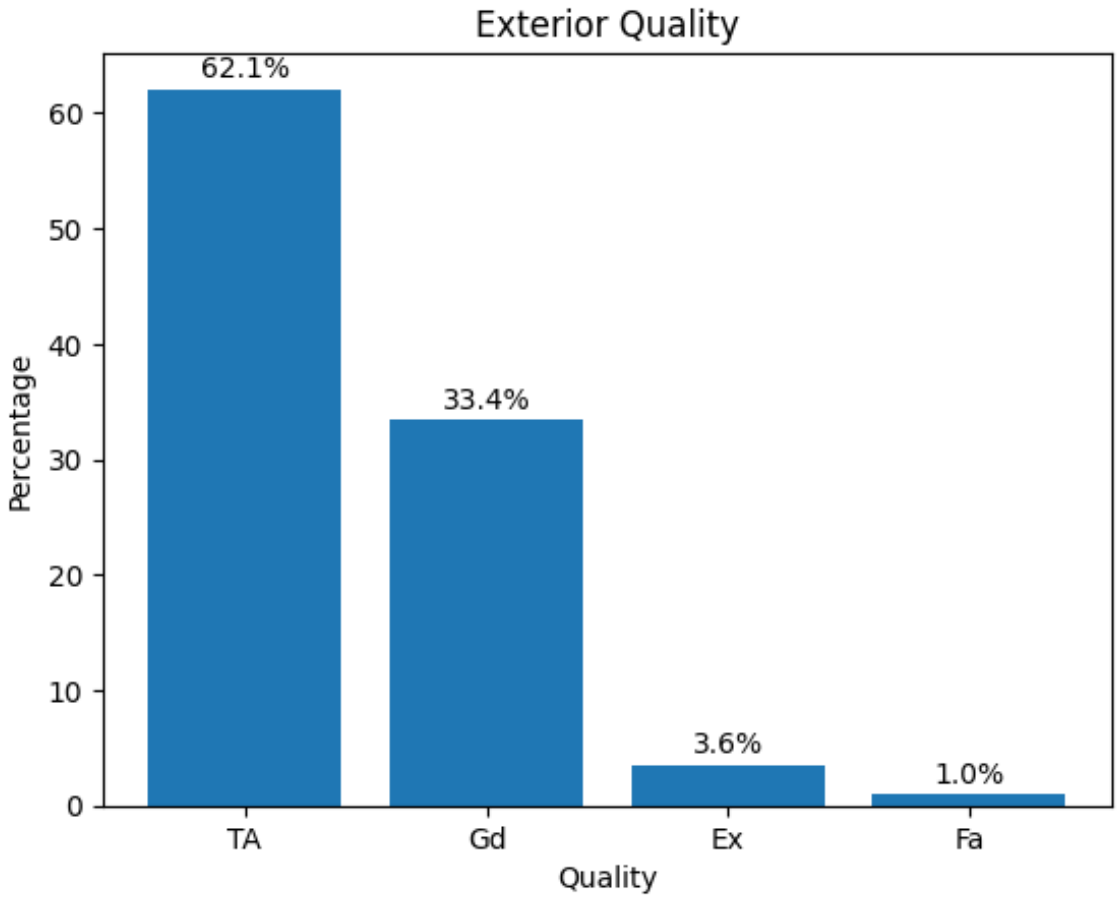

House Style
49.7%
30.5%
10.5%
4.5%
2.5%
1.0%
0.8%
0.5%
50
40
30
20
10
0
Percentage
1Story
2Story
1.5Fin
SLvl
SFoyer
1.5Unf
2.5Unf
2.5Fin
Style

Explanation

- In this dataset, there are various house styles that reflect architectural preferences and market trends. The 1-Story house style is the most common, representing 49.7% of properties. This single-storey design offers convenience and practical accessibility. The 2-Story home style covers 30.5% of properties, with a spacious two-story design suitable for families.
- The 1.5-Fin home style (10.5%) offers a unique layout with one and a half floors and a finished second floor. The S-Lvl home style (4.5%) is a split-level home with visually separated living areas. The S-Foyer home style (2.5%) is a home with separate entrances leading to the upper and lower levels.
- The 1.5-Unf house style (1%) is a one-and-a-half-story house with an unfinished second floor, allowing for personalized customization. The 2.5-Unf house style (0.8%) is a two-and-a-half-story house with an unfinished second floor. Meanwhile, the 2.5-Fin (0.5%) house style is a two-and-a-half-storey house with a finished second floor.
- Understanding the distribution of these house styles provides insights into architectural preferences and market trends, which can aid in decision-making in the housing industry. Each house style has its own uniqueness and characteristics that can attract buyers and determine the selling value of the property.

Exterior Quality (ExterQual)

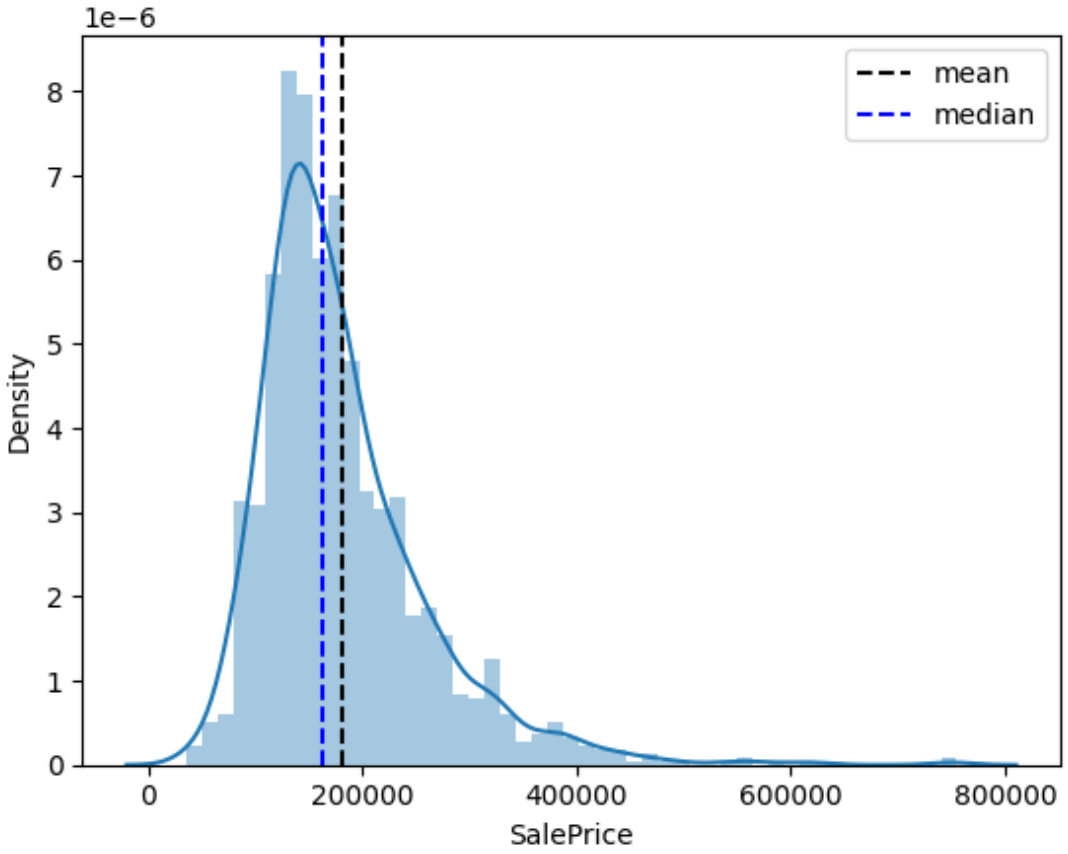
- The 'ExterQual' column in the dataset represents the material quality of the property's exterior. This column assesses the overall condition and attractiveness of the external features of the house. The values in this column are categorical and indicate different levels of quality, ranging from excellent to poor.



- The majority of properties (62.1%) have a fair or average exterior quality, with good quality (33.4%) being the second most common.
- Understanding ExterQual data is important in property transactions as it can provide insight into the exterior quality of a property.
- The exterior quality of a property plays an important role in determining its value and appeal to potential buyers, with properties that have good quality tending to have a higher value and attract buyers.

SalePrice

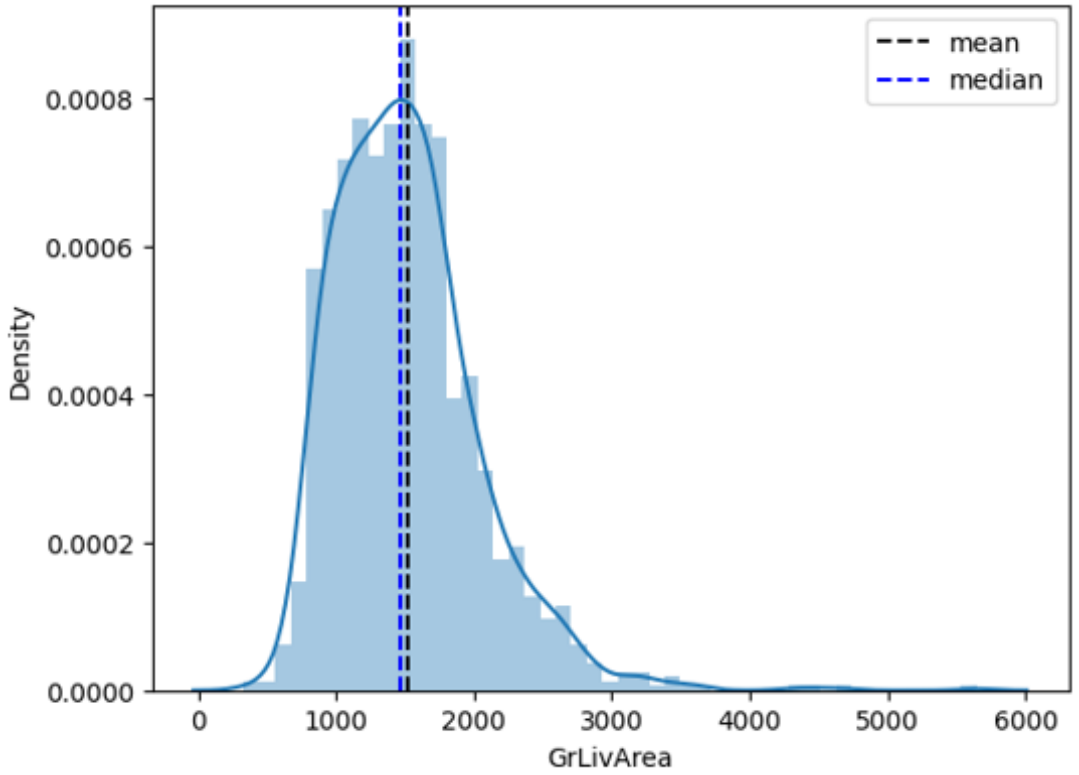
- The "SalePrice" column provides in-depth insights and an understanding of the factors that influence house prices. The column's analysis enables informed decision-making in buying or selling a property, gaining optimal value, and optimizing profits. The information in the "SalePrice" column helps individuals recognize market trends, adjust budgets, and take advantage of the opportunity in the competitive real estate industry.
- With the knowledge of house price factors through "SalePrice", individuals can make smarter and more effective decisions in their property buying and selling activities.



- This analysis uses histograms to illustrate the distribution of home sale prices, focusing on the median price of $163,000 and the average of $180,921,196. Significant price variations are seen in this dataset.
- While there are some properties with much higher prices, most home prices fall within the $100,000 to $300,000 range, which is a common and desirable option for many buyers.
- Knowing the distribution of home prices and these dominant price ranges provides important insights into property-related decision-making, allowing individuals to make smarter decisions based on personal needs and preferences, as well as factors that influence prices within these ranges.

GrLivArea

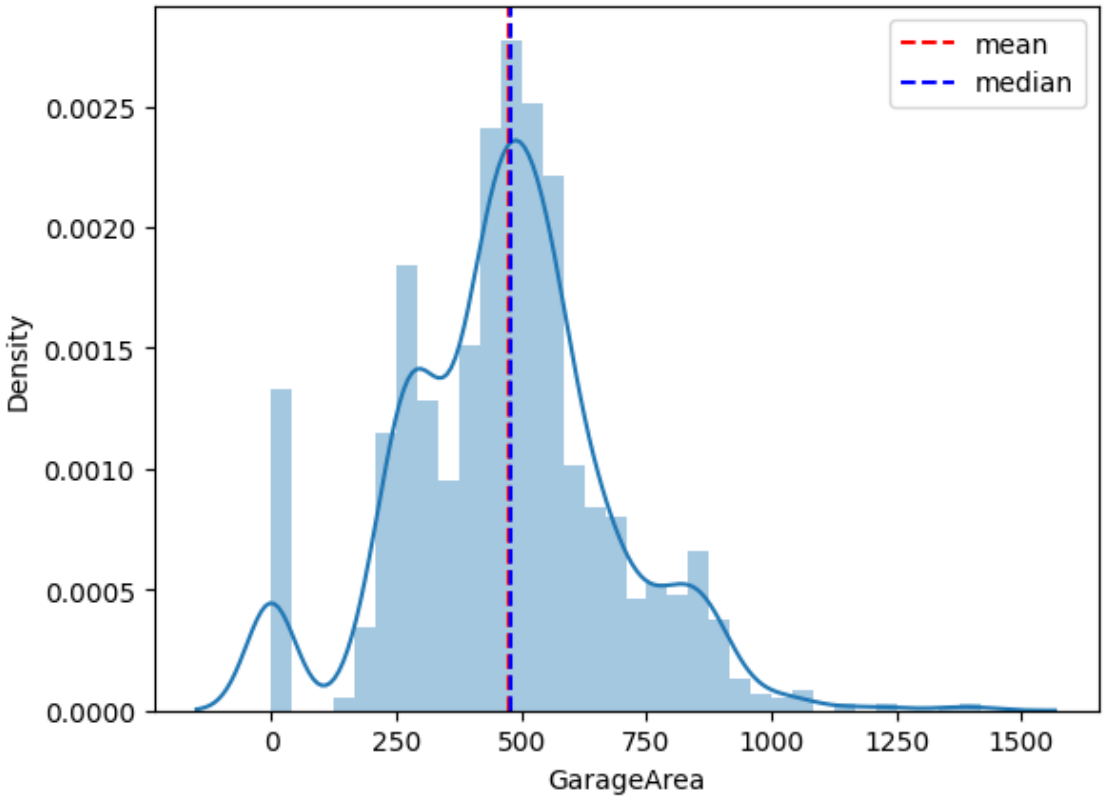
- The GrLivArea (Ground Living Area) column presents information about the living area of each property in the dataset. This data describes the size or area of the living area available in the house. The living area is important because it can affect the comfort and the functionality of the space for the occupants of the house.



- GrLivArea is a column in the dataset that reflects the living area of each property, with the right-skewed distribution indicating that most homes in the United States have a living area between 1000 and 2000 square feet.
- Understanding the size of a property's living area is important for buyers and sellers to make informed decisions.
- In my perspective, square footage plays a significant role in determining the value and appeal of a property, where larger square footage tends to command a higher price and attract more potential buyers. However, other factors also need to be considered in determining the price and purchase decision of a property.

GarageArea

- The ‘GarageArea’ column in this dataset reflects the garage area of each property. Garage area can provide insight into the capacity and ability of a garage to accommodate vehicles and provide additional storage space.



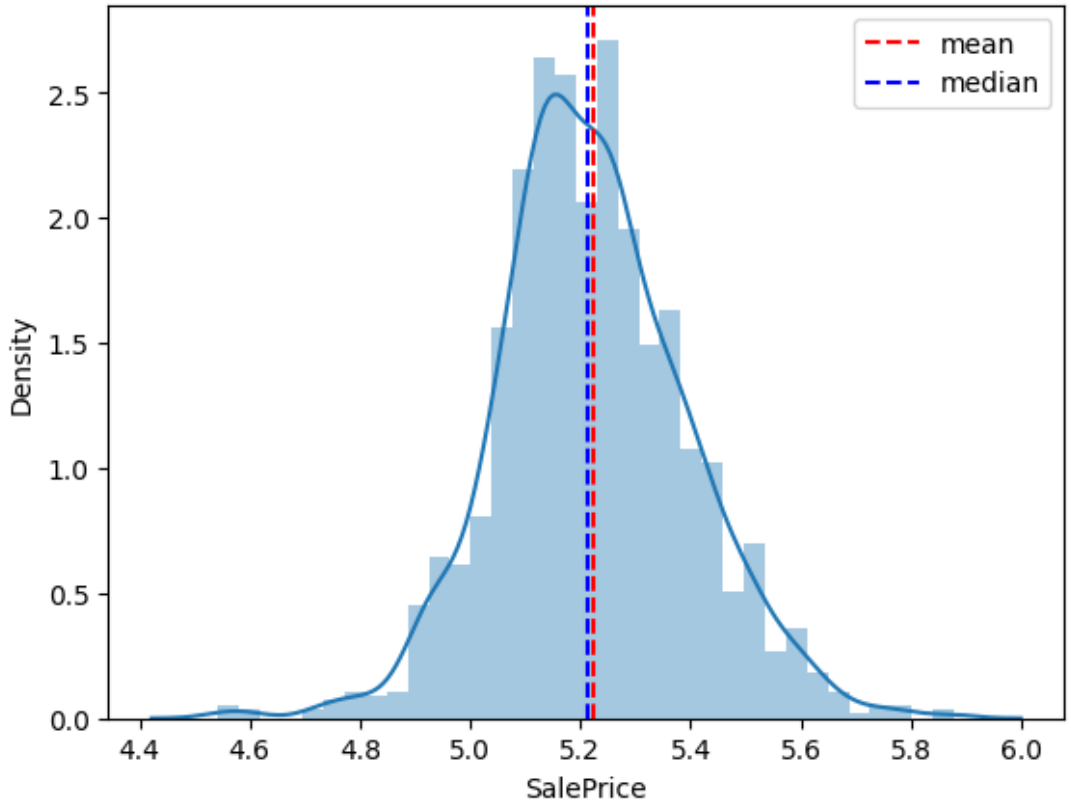
- The GarageArea column reflects the garage area of each property in the dataset, with the distribution of garage data skewed to the left.
- In the histogram visualization, most properties have a garage area between 250 and 875 square feet, but there are variations in garage area outside of this range.
- Understanding garage square footage is important in evaluating garage functions and needs, both for property buyers and sellers. Garage square footage also affects property value and convenience, although individual preferences and needs may vary, so the ideal garage size may be different for each person.

Processing Data Numerical varaibel

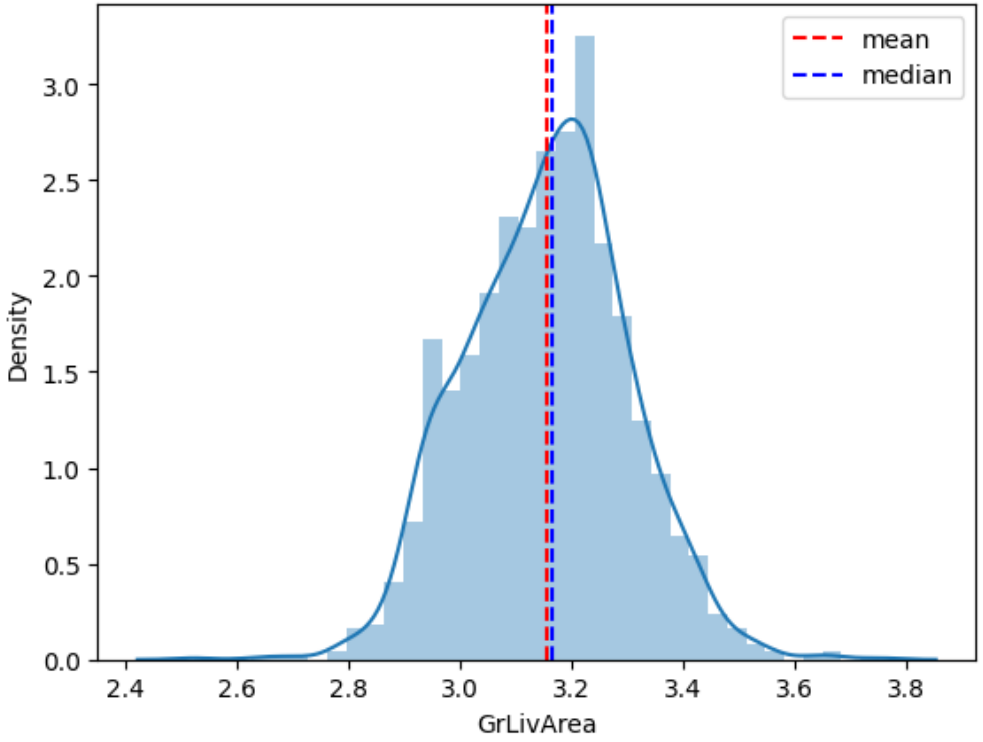
- In this analysis, the SalePrice and GarageArea columns use the log-10 technique to process the data. The aim is to provide a deeper understanding of the factors that influence house prices, so that individuals can make more informed decisions in property transactions.
- Data processing with the log-10 technique in the SalePrice column helps to equalize the scale of home price values, making it easier to compare and analyze factors such as location and property size.
- This method is important as it provides deeper insights and allows individuals to make informed decisions in the purchase or sale of property. By understanding the factors that influence house prices, individuals can conduct property transactions with confidence and maximize the success of their transactions.

SalePrice

- In this dataset, the 'SalePrice' column is the sale price of the house. To make it easier to understand the variation in house prices, a log-10 transformation was performed on this column.
- After the transformation, the median value becomes 5,212 and the mean becomes 5,222, with a standard deviation of 0.173. The log-10 transformation rescales the house price values to be more proportional, allowing for a better comparison between larger and smaller house prices.
- The log-10 transformation of the SalePrice column is particularly useful in overcoming unsymmetrical distributions and clarifying patterns of house price variation. By using a scale of comparable values, individuals can more easily compare house prices between different properties and gain a more accurate understanding of house price variations in this dataset.



GarageArea



- The 'GarageArea' column reflects the garage area of each property in the dataset. To gain a deeper understanding of the variation in garage area, a log-10 transformation was performed on this column.
- The transformation results show a median of 3.166, a mean of 3.156, and a standard deviation of 0.145.
- The log-10 transformation of the 'GarageArea' column provides significant benefits by balancing the data distribution and offering a clearer understanding of the variation in garage area. This variable plays a crucial role in determining the value and appeal of a property. The transformation enhances the analysis of garage area data, enabling more informed property-related decision-making.

Delete zero values in the 'GarageArea' column

```
[ ] # see how many zero values there are in GarageArea
    print("Number of non-zero values: ", np.sum(house_numeric["GarageArea"] != 0))
    print("Number of zero values: ", np.sum(house_numeric["GarageArea"] == 0))
```

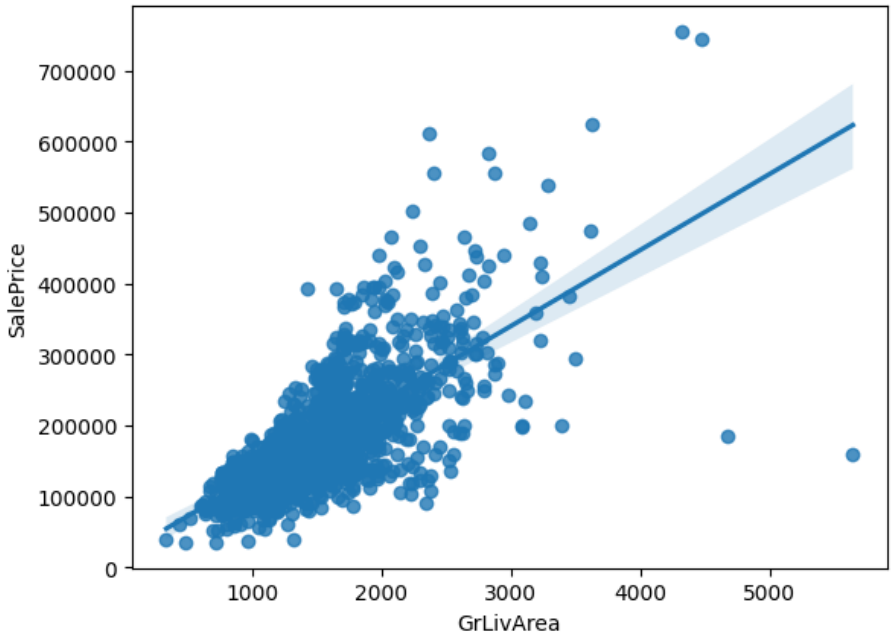
```
Number of non-zero values:  1379
Number of zero values:  81
```

```
# Removing zero values from GarageArea
x = house_numeric["GarageArea"][house_numeric["GarageArea"] != 0]
sns.distplot(x, axlabel=x.name)
line1 = plt.axvline(x.mean(), color='r', linestyle='--', label='mean')
line2 = plt.axvline(x.median(), color='b', linestyle='--', label='median')
first_legend = plt.legend(handles=[line1, line2], loc=1)

print('Median value after removing zero values from GarageArea: {:.2f}'.format(x.median()))
print('Mean value after removing zero values from GarageArea: {:.2f}'.format(x.mean()))
print('Standard deviation value after removing zero values from GarageArea: {:.2f}'.format(x.std()))

plt.show()
```

```
<ipython-input-33-9d2ceeeb3fb5>:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(x, axlabel=x.name)
Median value after removing zero values from GarageArea: 484.00
Mean value after removing zero values from GarageArea: 500.76
Standard deviation value after removing zero values from GarageArea: 185.68
```

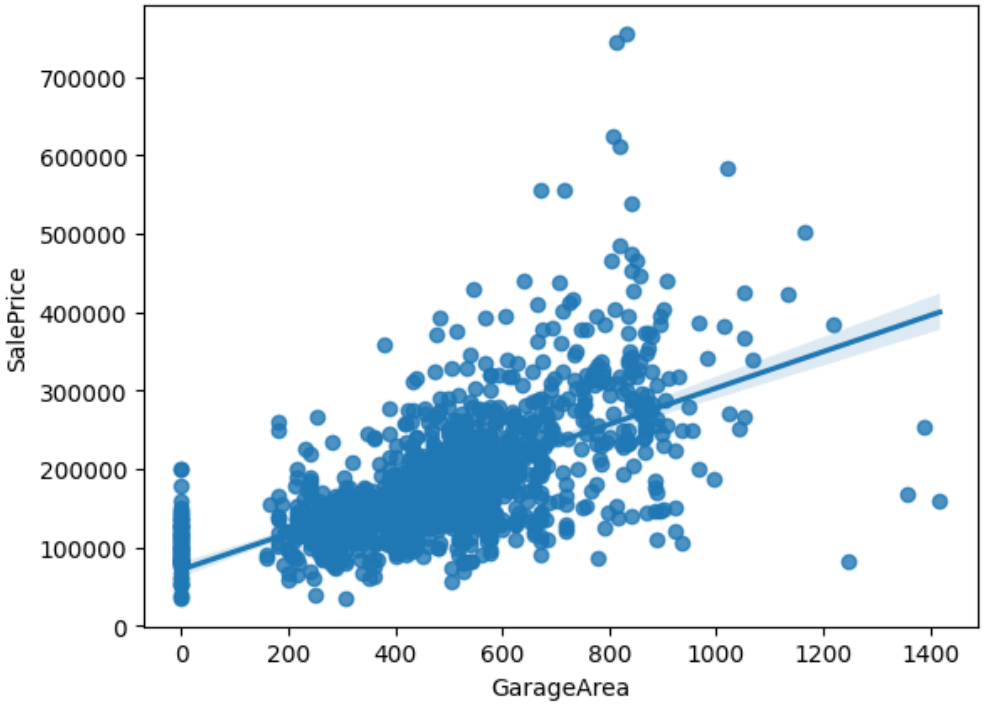
- In the 'GarageArea' column, zero values indicating the absence of garages on some properties were removed from the analysis to gain a more accurate understanding of the garage area on properties that actually have garages.
- After removing the null values, the analysis shows that the median garage area is 484.00, the mean is 500.76, and the standard deviation is 185.68. This provides a more accurate understanding of the variation in garage area on properties with garages, which impacts the value and attractiveness of the property.
- By eliminating zero-values, focus is given to properties that actually have garages, allowing for more informed decision-making in the purchase or sale of a property. Considering the garage area of a property that has a garage effectively helps in optimizing the outcome of a property transaction and understanding the factors that influence home prices.

Relationship between ‘GrLivArea’ and ‘SalePrice’



- In the regression graph between the columns 'GrLivArea' and 'SalePrice’. There is a positive relationship between the two. That is, the larger the residential area of a property, the higher the sale price. This shows that residential area is one of the factors that affect property prices.
- The impact in the property business is the importance of considering the size of the residential area in determining the selling price. This information provides insight for sellers to recognize the added value of a larger living area, which can reflect the level of comfort and appeal to consumers.
- In addition, understanding the positive relationship between square footage and selling price also helps buyers evaluate property values and make more informed purchasing decisions. By paying attention to this relationship, property businesses can optimize pricing strategies and improve customer satisfaction.

Relationship between ‘GarageArea’ and ‘SalePrice’



- The regression plot between the columns 'GarageArea' (garage area) and 'SalePrice' (sale price) shows a positive relationship between the two variables. The larger the garage area, the higher the selling price of the property.
- This information can be used to determine the selling price or value of the property to be purchased. Property developers can also consider potential buyers' preferences regarding garage area in planning and designing new properties.
- In fact, aside from the positive relationship between garage space and property selling price, other factors such as location, property condition, and other features also affect the overall price.

Conclusion

Through data exploration and analysis using EDA, the goal is to uncover patterns, relationships, and insights that can improve decision-making regarding house prices and related attributes. Through data visualization and analysis, EDA provides a deeper understanding of categorical and numerical variables associated with property prices. Thus, this analysis can help users to recognize important factors affecting house prices and make smarter decisions in property transactions.

Thank You