Kevin Kiding

Housing Price Prediction

Background & Objective

- The background of the Exploratory Data Analysis (EDA) process in this case study is to understand and analyze the prepared data with the aim of discovering relevant patterns, relationships, and insights to support better decision-making in the context of house prices and their related attributes.

- The goal of the EDA process in this case study is to provide deeper insights and understanding of the factors that influence house prices, so that individuals can make more informed and appropriate decisions in buying or selling property.

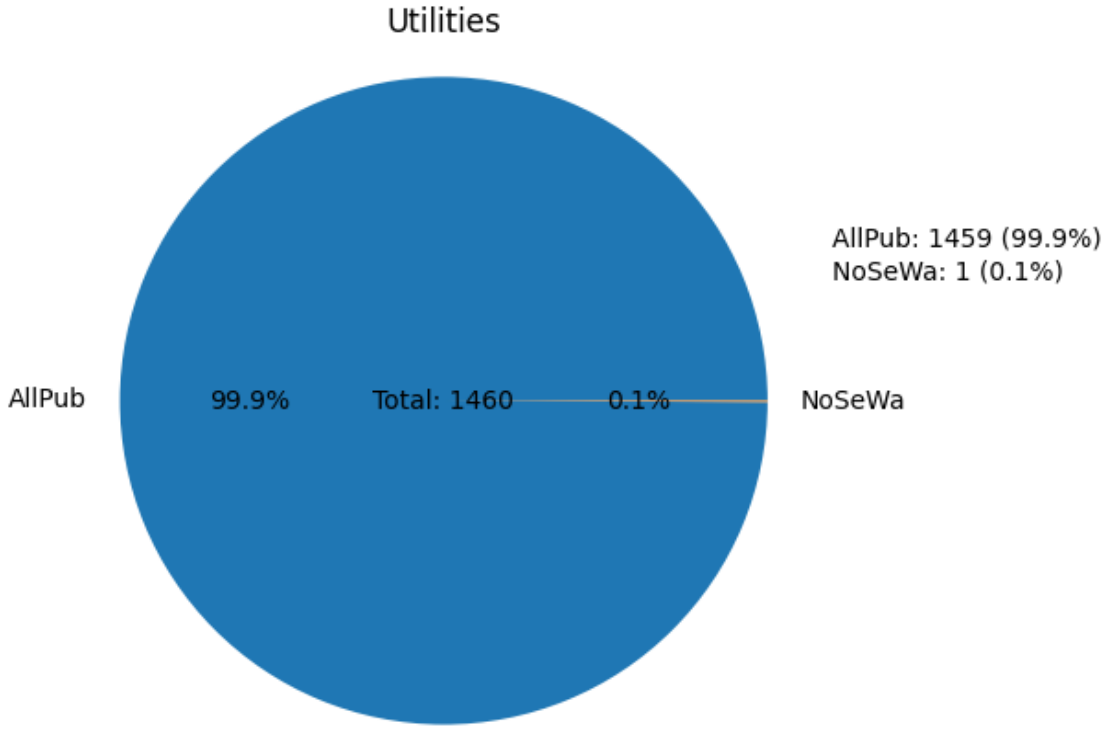
Introduction

I would like to share a link with you to access the data processing process that I have performed in Google Colab. By clicking on the provided link, you will be able to view and explore the steps involved in the data processing, including any code snippets or explanations that were utilized. This will enable you to gain insights into the data processing techniques applied and understand the overall workflow. Please feel free to access the link at your convenience.

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

Utilites

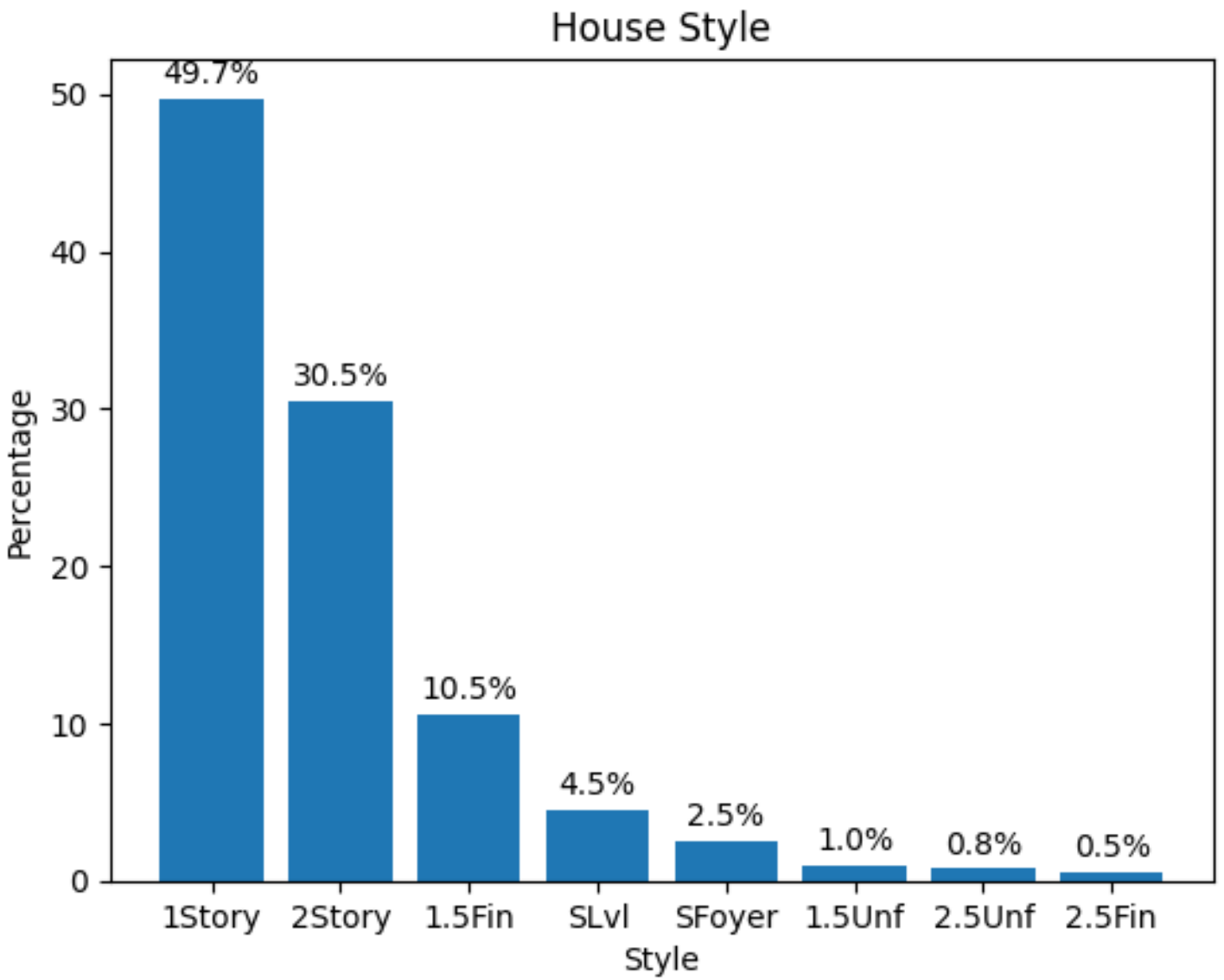
- Contains information about what facilities are available at the property.
  - AllPub: This indicates that the property has access to all common public utilities, including Electricity (E), Gas (G), Water (W), and Sanitation (S).
  - NoSeWa: This indicates that the property has only limited access to 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 the utility type 'AllPub', indicating full access to all public utilities.
- Only a single property (0.01%) has the utility type 'NoSeWa', indicating limited access to utilities.
- Understanding the distribution of utility types provides insights into the availability and accessibility of essential services, benefiting stakeholders such as homebuyers, developers, and policymakers.

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.

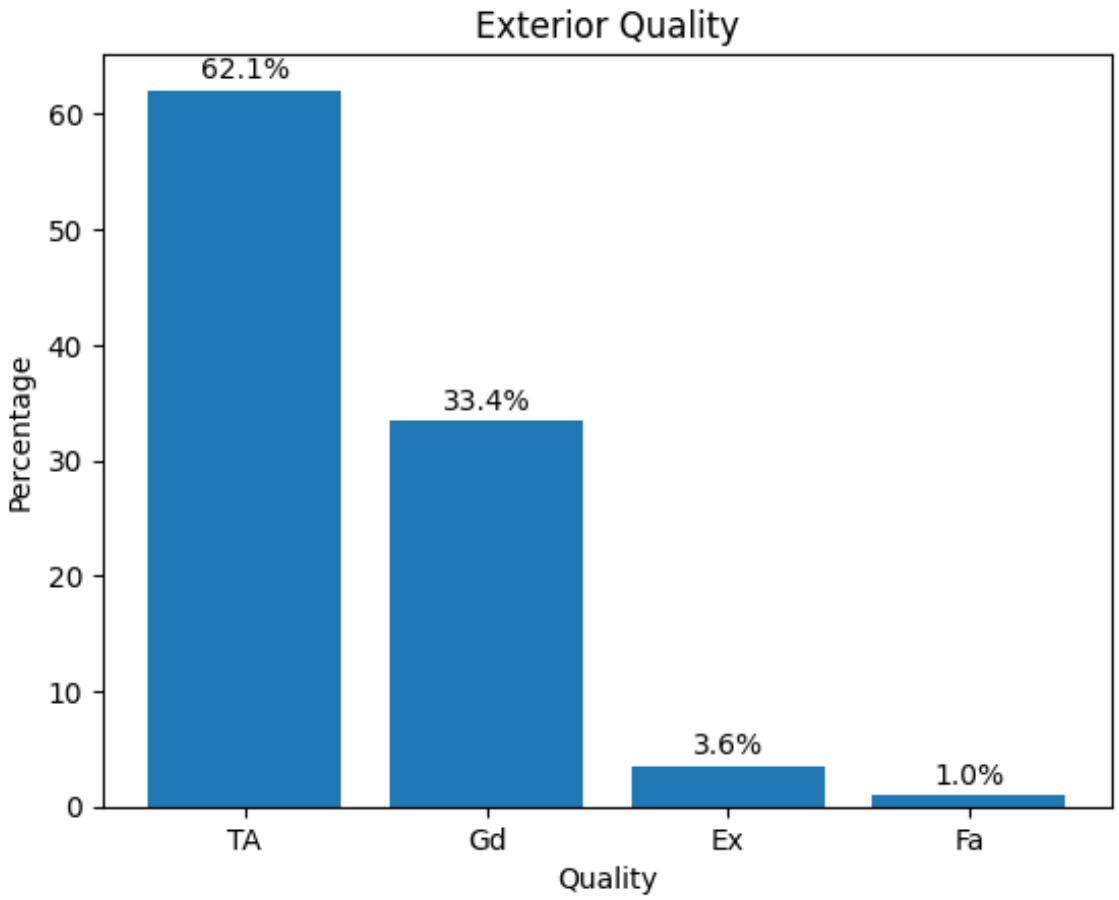

House Style
49.7%
30.5%
10.5%
4.5%
2.5%
1.0%
0.8%
0.5%
50
40
30
20
10
0
Percentage
1Story
2Story
1.5Fin
SLvl
SFoyer
1.5Unf
2.5Unf
2.5Fin
Style

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.

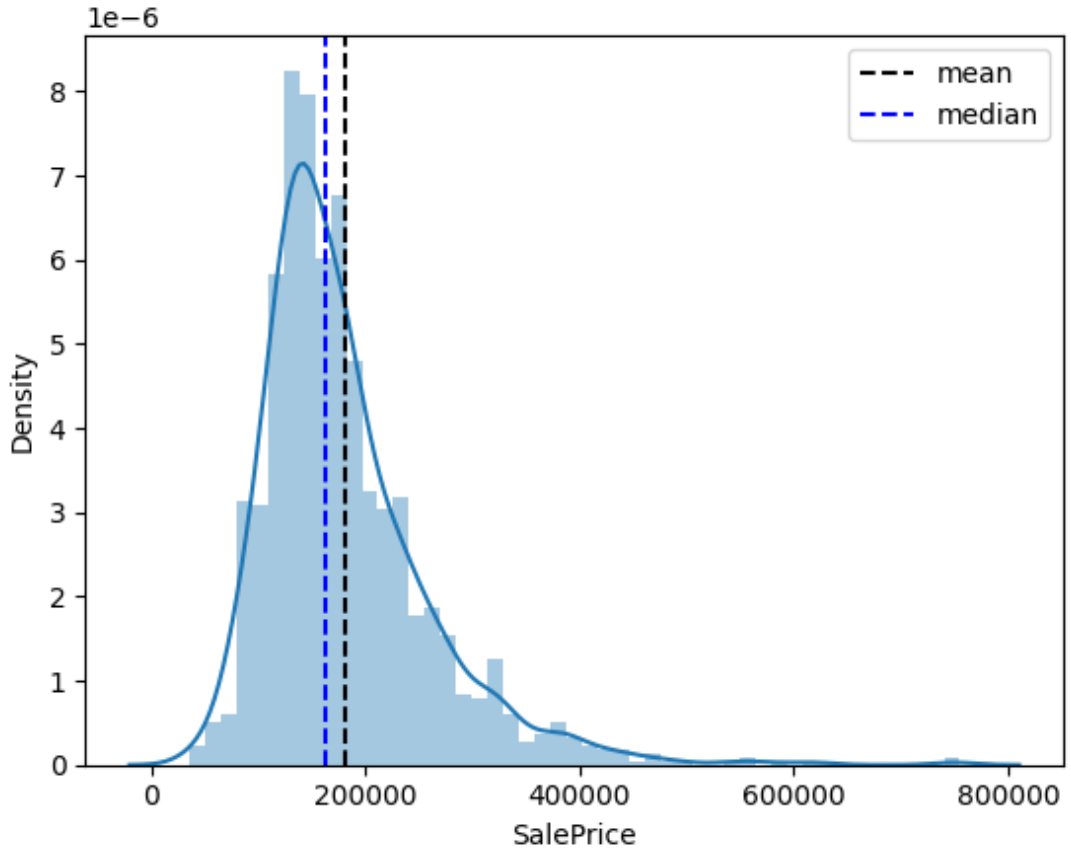


- The ExterQual data provides insight into the exterior quality of the properties in this dataset. The majority of properties have average or common exterior quality (62.1%), with good quality (33.4%) being the second most common. Properties with excellent (3.6%) and fair (1.0%) exterior quality have lower proportions. Understanding this factor is important in the purchase or sale of a property. By knowing the exterior quality of the property, a more informed decision can be made in determining the sale price or purchase offer. This factor also affects the value and attractiveness of the property to potential buyers.

- My point of view is that the exterior quality of a property plays an important role in determining its value and appeal. Properties with good exterior quality tend to have a higher value and attract potential buyers. On the other hand, properties with fair or average exterior quality may require renovations or repairs to increase their value and appeal. Therefore, understanding ExterQual data provides a competitive advantage in property transactions and helps in making smarter and more profitable decisions.

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.

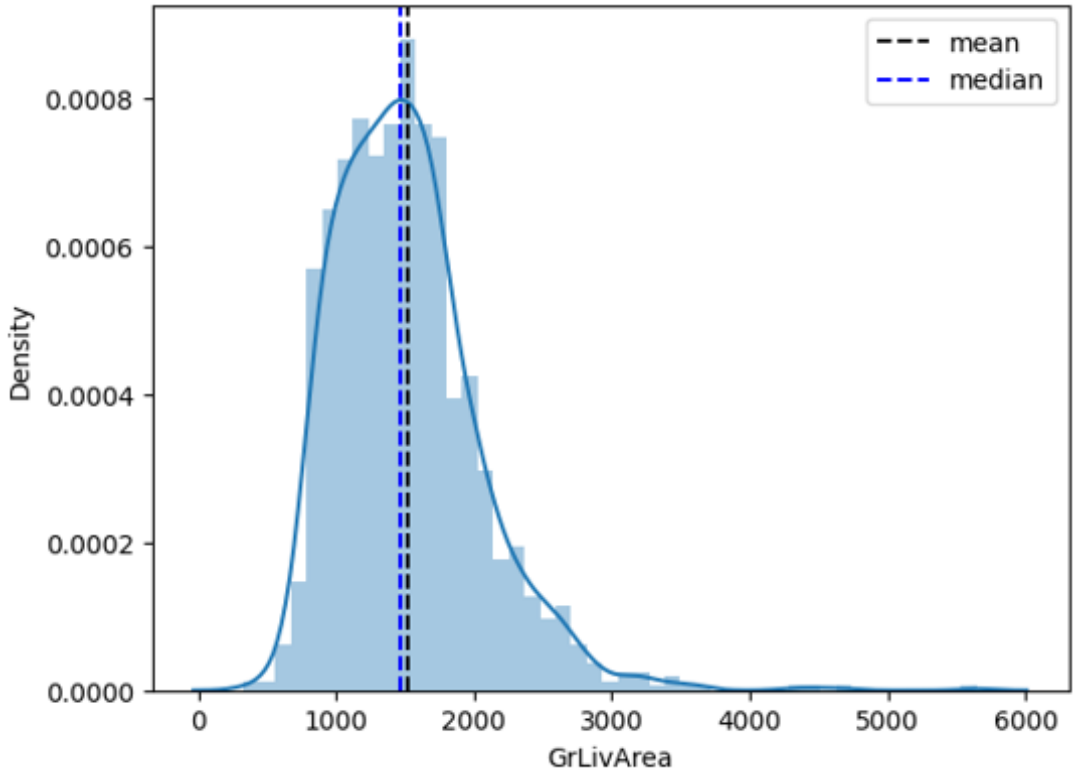


- This analysis depicts the distribution of home sale prices through histograms, providing a deeper understanding of the price variations in the dataset. From the visualization, it can be seen that the median house price is $163,000 with an average of $180,921,196, indicating significant variation. The right-skewed distribution indicates the presence of a number of properties with much higher prices. Nonetheless, most home prices fall within the $100,000 to $300,000 range, which is a common and desirable option for many buyers.

- Understanding the distribution of house prices and this dominant price range provides insights into property-related decision-making. Individuals can make smarter and more informed decisions by considering personal needs and preferences, as well as the factors that influence prices within these ranges. In-depth analysis and understanding of market trends can also assist individuals in making property transactions as expected.

GrLivArea

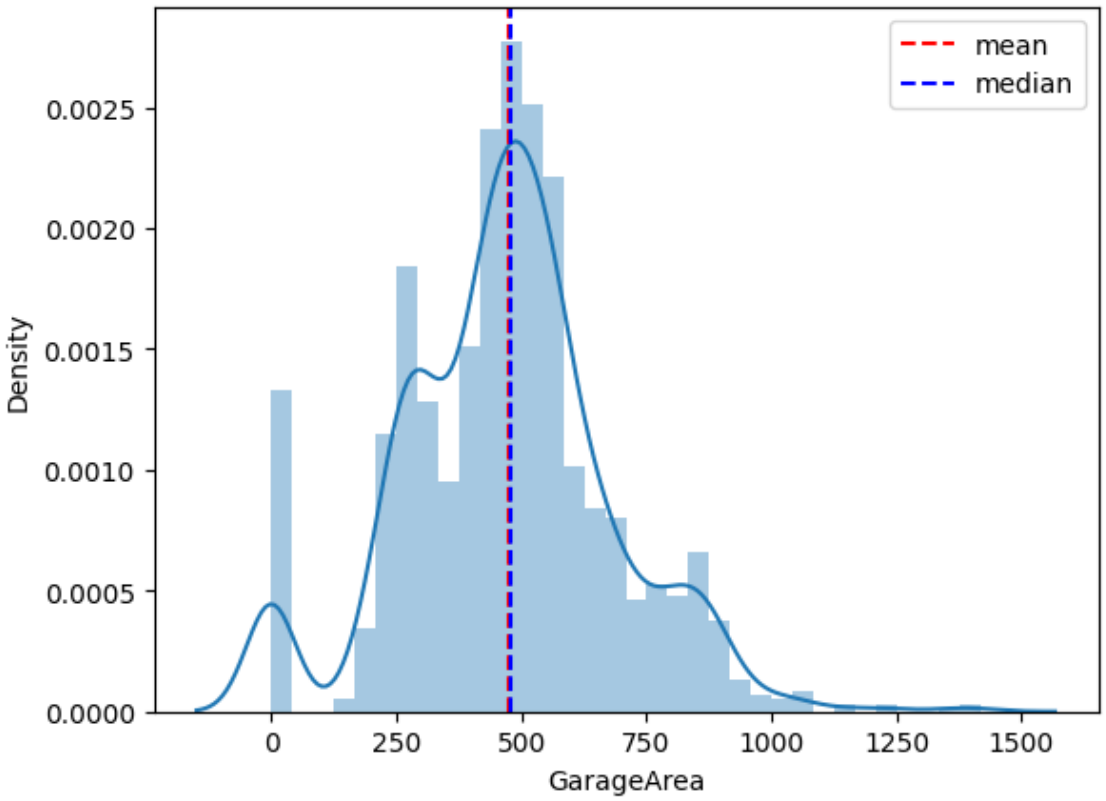
- The GrLivArea (Ground Living Area) column presents information about the living area of each property in the dataset. This data describes the size or area of the living area available in the house. The living area is important because it can affect the comfort and the functionality of the space for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



- The GrLivArea column in this dataset represents the ground living area of each property. The right-skewed distribution indicates that the majority of homes in the United States have a living area between approximately 1000 and 2000 square feet. The median living area is 163,000, while the mean is 180,921.196, showing some variation in the dataset. Understanding the size of the living area is crucial for buyers and sellers to make informed decisions.
- In my opinion, the living area plays a significant role in determining the value and appeal of a property, with larger living areas generally commanding higher prices and attracting more interest from potential buyers. However, it's important to consider other factors that may also impact the property's price and purchase decision.

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.

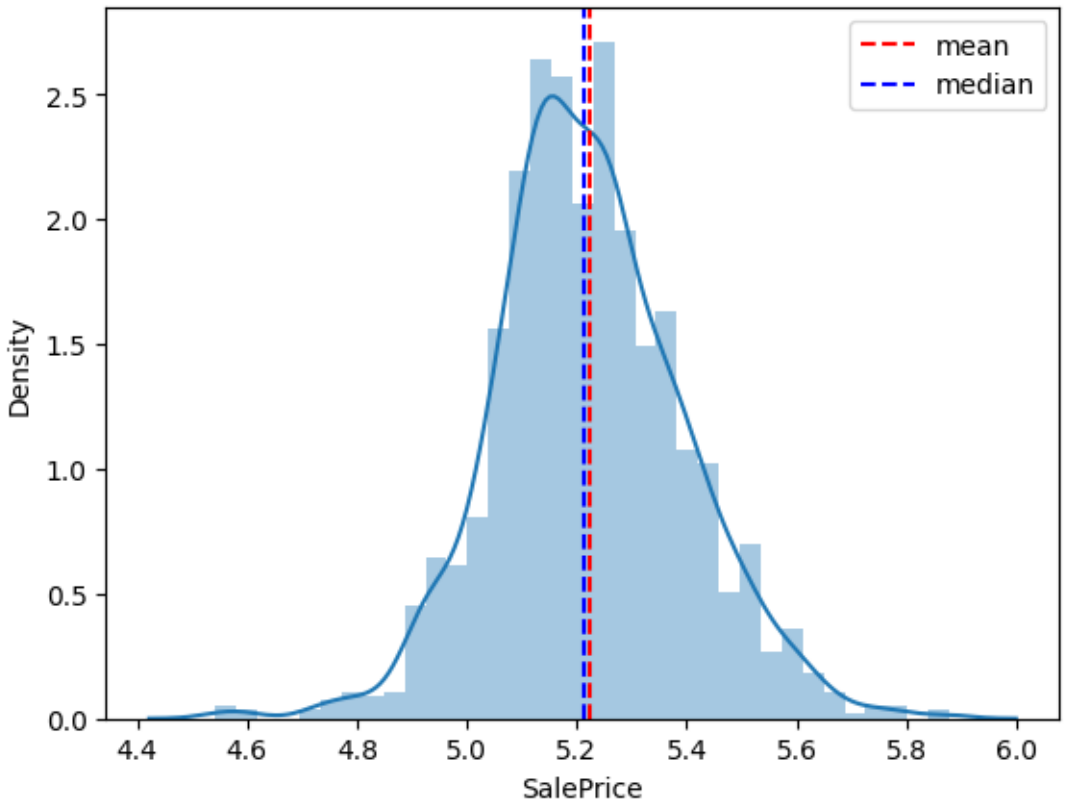


- The GarageArea column reflects the garage area of each property in the dataset. In the histogram visualization, the distribution of the garage area data is skewed to the left. The median garage area is 163,000, while the mean is 180,921,196. The standard deviation of 79,442,503 shows the variation of garage area in this dataset. Most properties have garage areas between 250 and 875 square feet, but there are also properties with garage areas that are smaller or larger than this range.
- Understanding garage square footage is important in evaluating garage functionality and needs, both for property buyers and sellers. Garage square footage also affects the value and convenience of the property, but it should be noted that 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. This is done 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 house price values, making it easier to compare and analyze factors such as location and property size.
- This approach is important as it provides deeper insights and allows individuals to make informed and informed decisions in buying or selling property. By understanding the factors that influence house prices, individuals can conduct property transactions with confidence and maximize the success of their transactions.

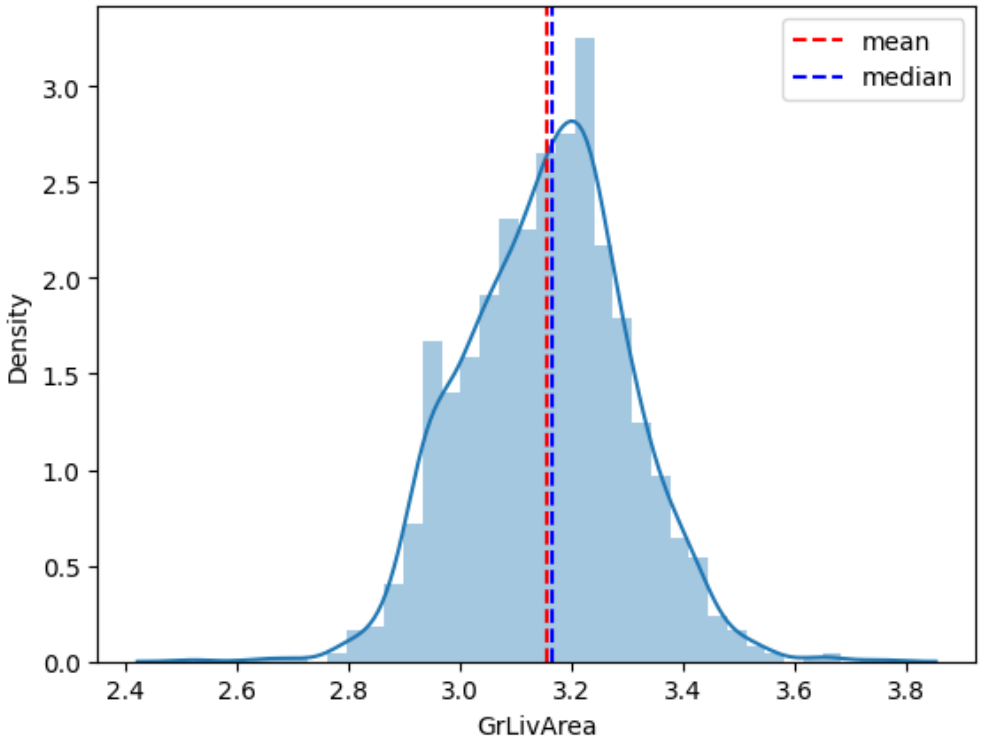
SalePrice



The 'SalePrice' column in this dataset is the sales price of the house. To facilitate understanding of the variation in house prices, a log-10 transformation was performed on this column. After the transformation, we see that the median value becomes 5,212 and the mean value becomes 5,222, with a standard deviation of 0.173. The log-10 transformation changes the scale of house price values to be more proportional, allowing for a better comparison between larger and smaller house prices. This provides a better insight into the factors affecting house prices and helps in making more informed decisions in property transactions.

Personally, the log-10 transformation on the SalePrice column is very useful in overcoming the unsymmetrical distribution and clarifying the pattern of house price variability. By using a comparable scale of 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 showed a median of 3.166, an average of 3.156, and a standard deviation of 0.145.
- I personally find that the log-10 transformation of the ‘GarageArea’ column provides significant benefits. With this transformation, the data distribution becomes more balanced and allows for a better understanding of the variation in garage area. Garage area has an important influence in determining the value and attractiveness of a property. The transformation provides a more comprehensive and in-depth understanding in analyzing garage area data and optimizing 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))
```

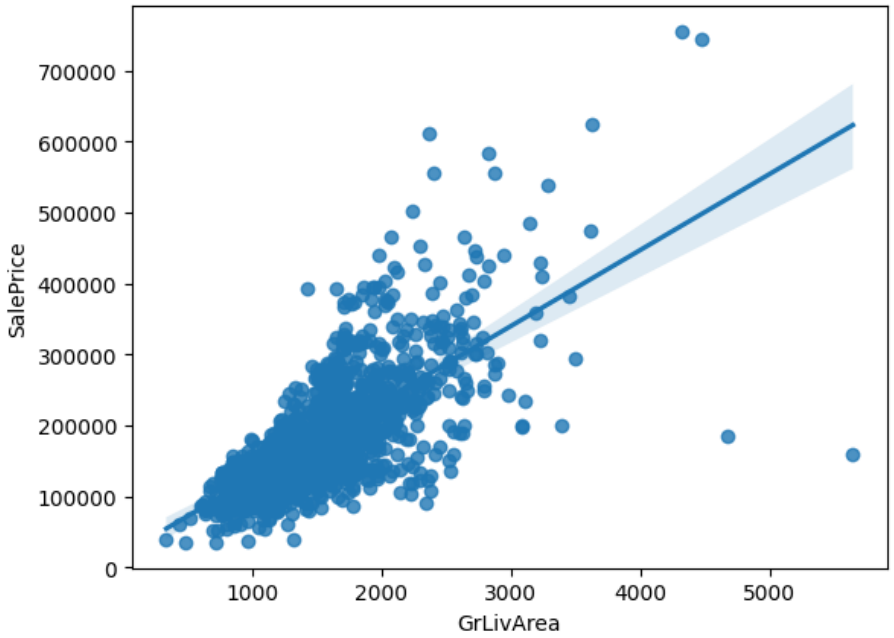
```
Number of non-zero values:  1379
Number of zero values:  81
```

```
# 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
```

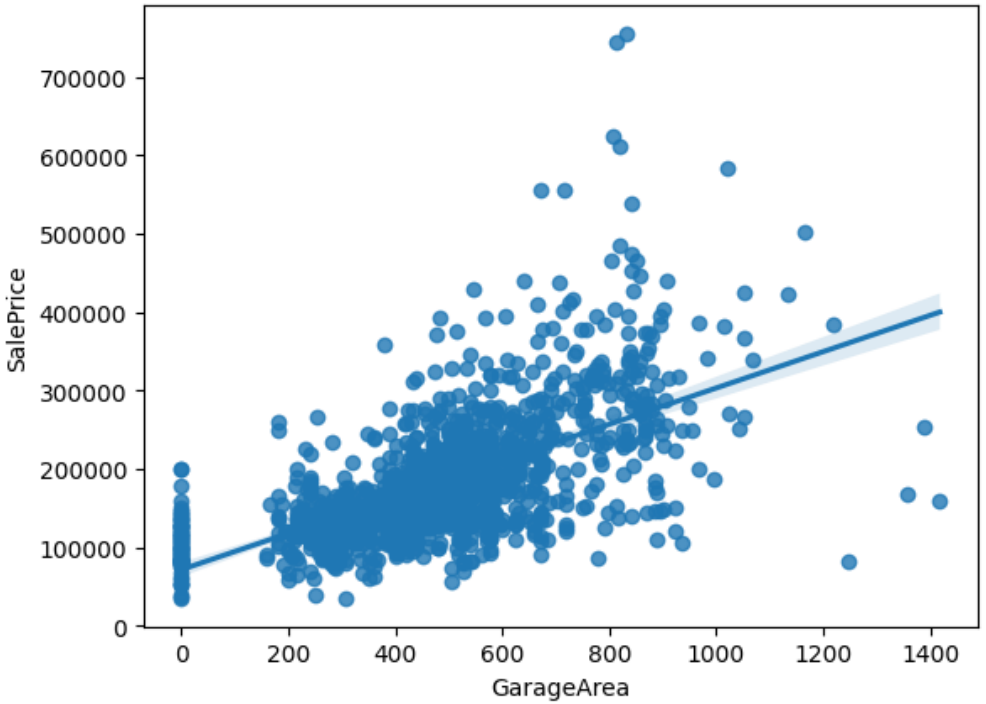
- In the GarageArea column, there are some properties in the dataset that do not have garages, indicated by zero values. To get a more accurate understanding of the garage area of properties that actually have garages, these zero values were removed from the analysis. By removing the zero values, focus can be put on those properties that actually have garages, which are likely to have a more significant influence on the value and attractiveness of the property. The results show that after removing the null values, 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, allowing for more informed decision-making in buying or selling a property.
- By removing the zero values, focus can be given to those properties that actually have garages, thus providing a more accurate understanding of the factors that influence house prices. By considering the garage area of properties that actually have a garage, we can make a more informed decision in buying or selling a property and maximize the outcome of the property transaction.

Relationship between ‘GrLivArea’ and ‘SalePrice’



- In the regression graph between the 'GrLivArea' (living area) and 'SalePrice' columns, there is a positive relationship between the two. That means, the larger the living area of a property, the higher the sales price. This shows that living area is one of the factors that affects property prices.
- The impact on the property business is the importance of considering the living area in setting the selling price. This information provides insight for property sellers to recognize the added value of a larger living area, which can reflect the comfort level and desirability of consumers.
- In addition, understanding the positive relationship between living area 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 'GarageArea' and 'SalePrice' columns shows a positive relationship between the two variables. A larger Garage Area tends to be associated with a higher Sale Price. This information can be used by property owners and potential buyers to determine the selling price or value of the property to be purchased. Property developers can also consider prospective buyers' preferences regarding garage area in planning and designing new properties.
- However, keep in mind that this relationship is not absolute, and other factors such as location and property condition also affect the overall price. A more comprehensive analysis and holistic consideration is required in determining the overall property price.

Conclusion

This analysis provides an in-depth understanding of the factors that influence home prices, such as utility type, home style, exterior quality, selling price, living area, and garage area. This knowledge helps individuals make more informed decisions in property transactions, maximize property value, and increase profits. An understanding of the relationship between square footage and selling price, market preferences for certain features, and market trends provides a competitive edge in the competitive property industry. In making property purchase or sale decisions, a deep understanding of these factors is critical to achieving a successful property transaction.

Thank You