PPT (Pinagtagpo Pero di Tinadhana)

Aaron Cloyd Villarta

Mike Lawrence Alpas

Godwin Ryan Sanjorjo

Daniel Gilbert Dela Pena

Franz Jason Dolores

REPORT ON DATA EXPLORATION TECHNIQUES (USING HOUSING PRICE DATASET)

Introduction

The dataset used in this analysis is the Housing Price Dataset, which contains 545 entries. It includes both numerical and categorical variables, such as:

- Numerical variables: price, area, bedrooms, bathrooms, stories, and parking.
- Categorical variables: mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea, and furnishingstatus.

The dataset helps analyze the relationship between various house characteristics and their respective prices. Missing values were handled by filling numerical columns with their mean and categorical columns with their mode.

Key Statistics

Key descriptive statistics for the numerical variables are as follows:

- **Mean Price**: 4,766,729
- **Median Price**: 4,340,000
- **Mode Bedrooms**: 3
- **Standard Deviation of Price**: 1,870,440
- **Range of Area**: 14,550 square feet (Min: 1,650, Max: 16,200)

The central tendency (mean, median, mode) highlights the typical property characteristics, while the variability (standard deviation, range) reveals a significant spread in prices and house sizes.

Insights from Descriptive Statistics

- **Skewness**: The data for price and area appears right-skewed, as indicated by a mean higher than the median. This means there are some very expensive and large properties pulling the distribution to the right.
- **Outliers**: The box plots reveal several outliers, especially for price and area. These outliers represent the properties that are much more expensive and larger than most homes in the dataset.
- **Spread**: The range of prices is quite large, indicating that the dataset includes properties from relatively affordable homes to very high-end properties. The spread in parking values is much smaller, showing that most properties have between 0 and 3 parking spaces.

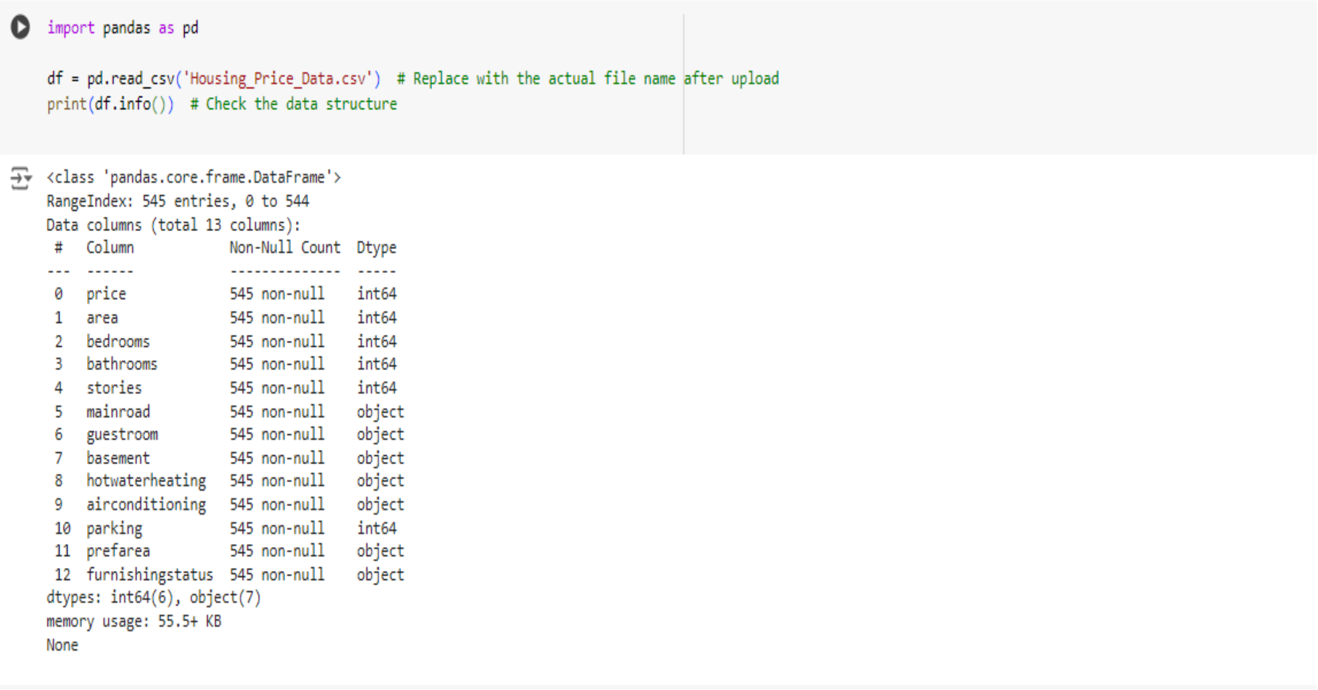
Visualizations

- **Histograms**: The histogram visualizations for numerical variables show a skewed distribution for both price and area. A larger proportion of homes are on the smaller and less expensive end, with fewer large, high-priced homes.
- **Box Plots**: Box plots confirm the presence of several outliers in price, area, and parking, which reflect properties that significantly deviate from the typical house in the dataset.
- **Correlation Heatmap**: A heatmap of the correlation matrix shows a strong positive correlation between:
  - price and area (larger properties tend to be more expensive).
  - price and the number of bedrooms, bathrooms, and stories, reinforcing the idea that more amenities and space are tied to higher prices.

Conclusion

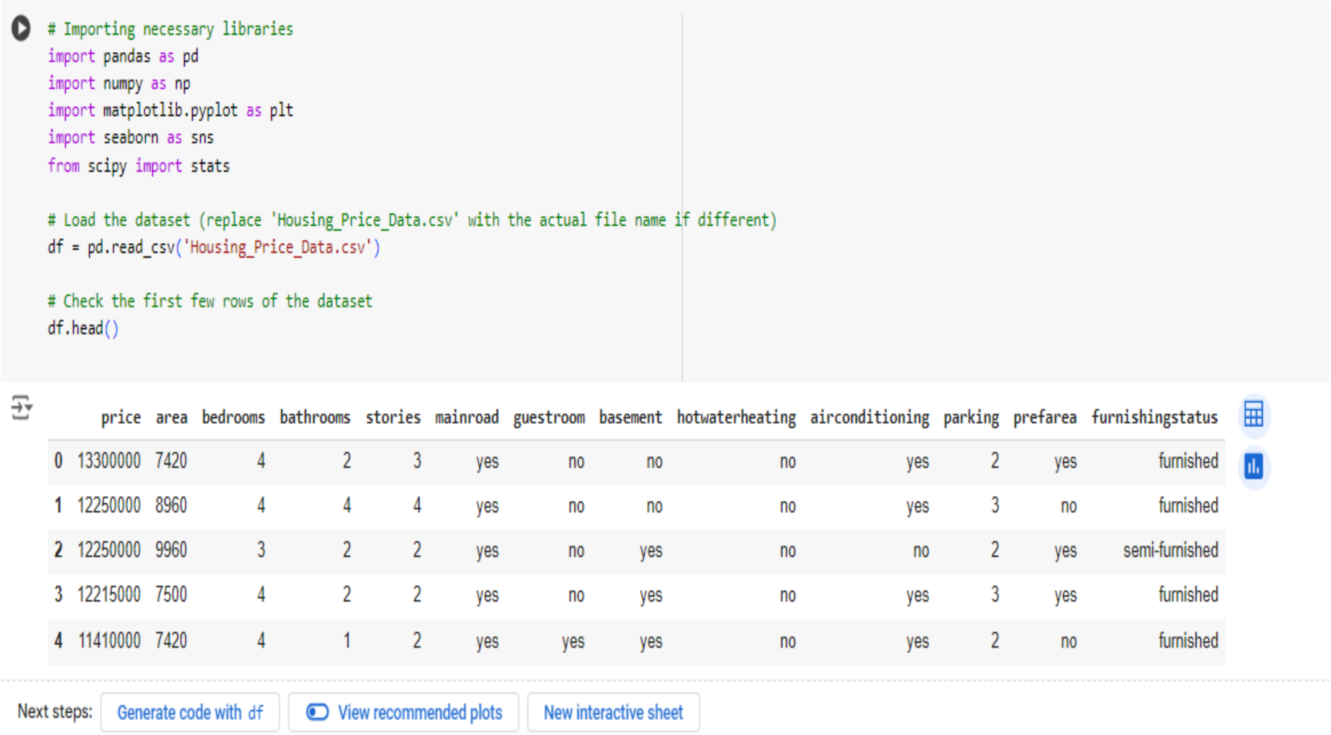
This dataset shows significant variability in housing prices and features. The analysis reveals that:

- Larger homes with more bedrooms, bathrooms, and stories are generally more expensive.
- There are some highly priced outliers, contributing to the right-skewed nature of the price and area distributions.



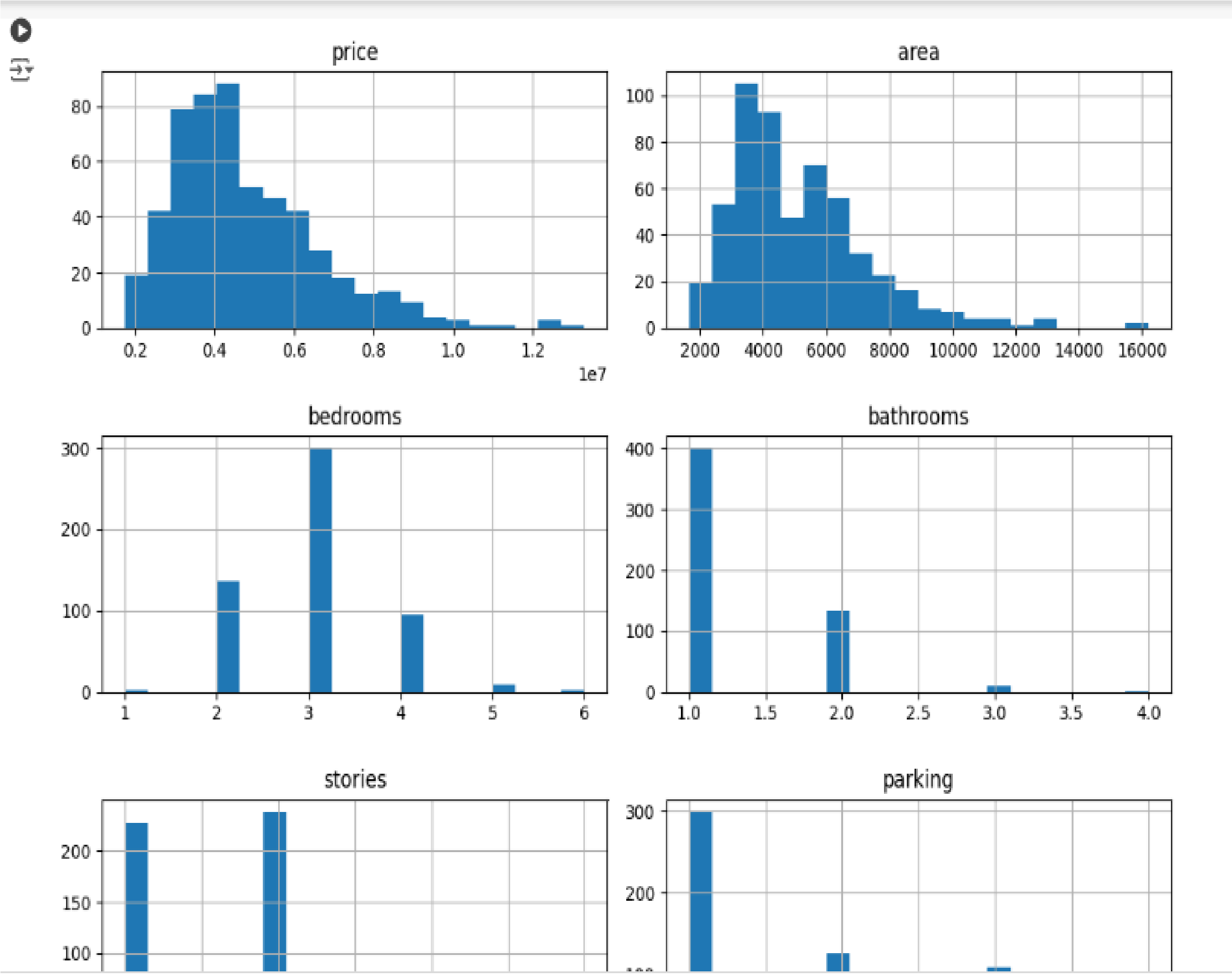
```python
import pandas as pd

df = pd.read_csv('Housing_Price_Data.csv')  # Replace with the actual file name after upload
print(df.info())  # Check the data structure
```

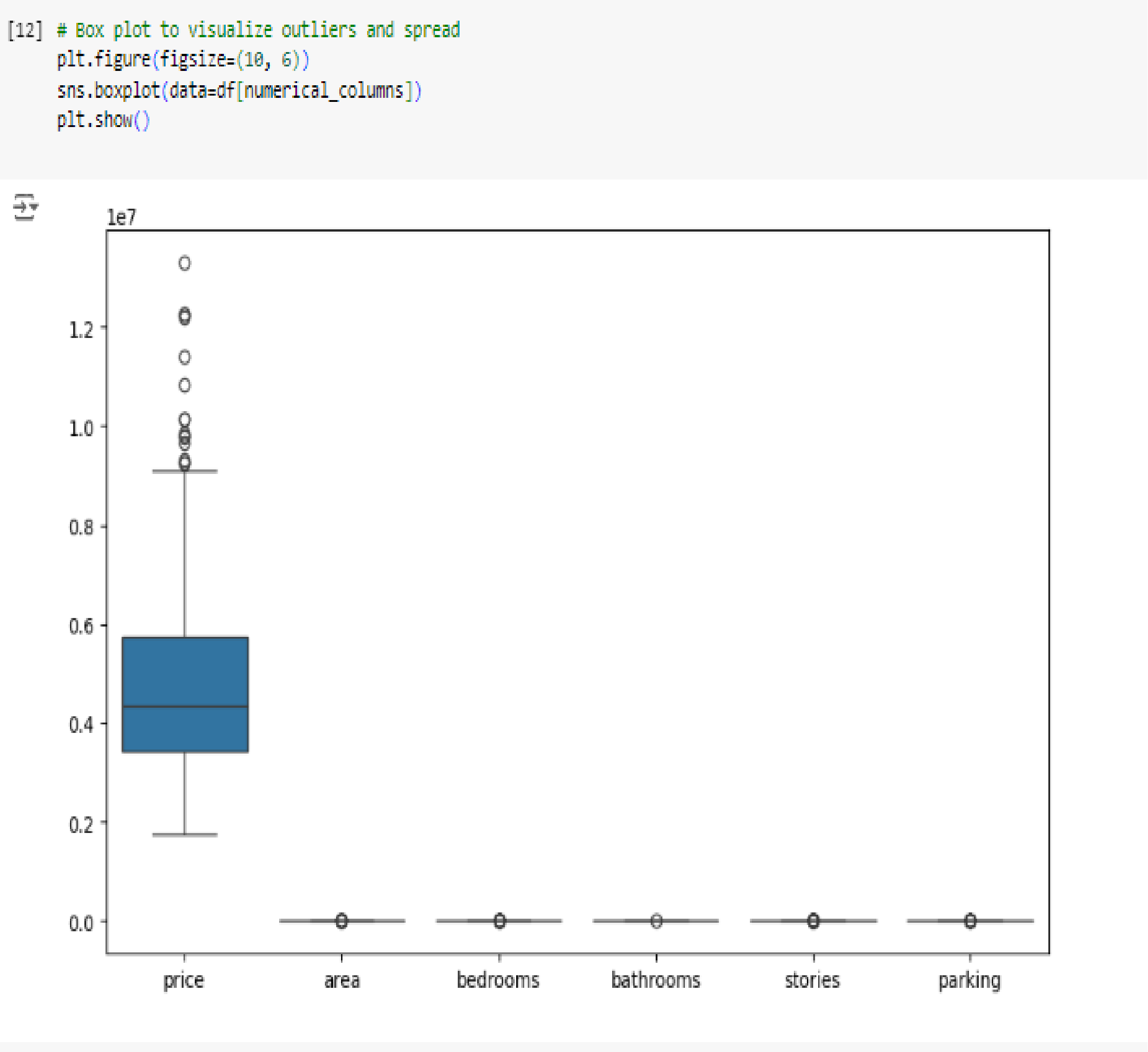
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 545 entries, 0 to 544
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   price             545 non-null    int64
 1   area              545 non-null    int64
 2   bedrooms          545 non-null    int64
 3   bathrooms         545 non-null    int64
 4   stories           545 non-null    int64
 5   mainroad          545 non-null    object
 6   guestroom         545 non-null    object
 7   basement          545 non-null    object
 8   hotwaterheating   545 non-null    object
 9   airconditioning   545 non-null    object
 10  parking           545 non-null    int64
 11  prefarea          545 non-null    object
 12  furnishingstatus  545 non-null    object
dtypes: int64(6), object(7)
memory usage: 55.5+ KB
None
```



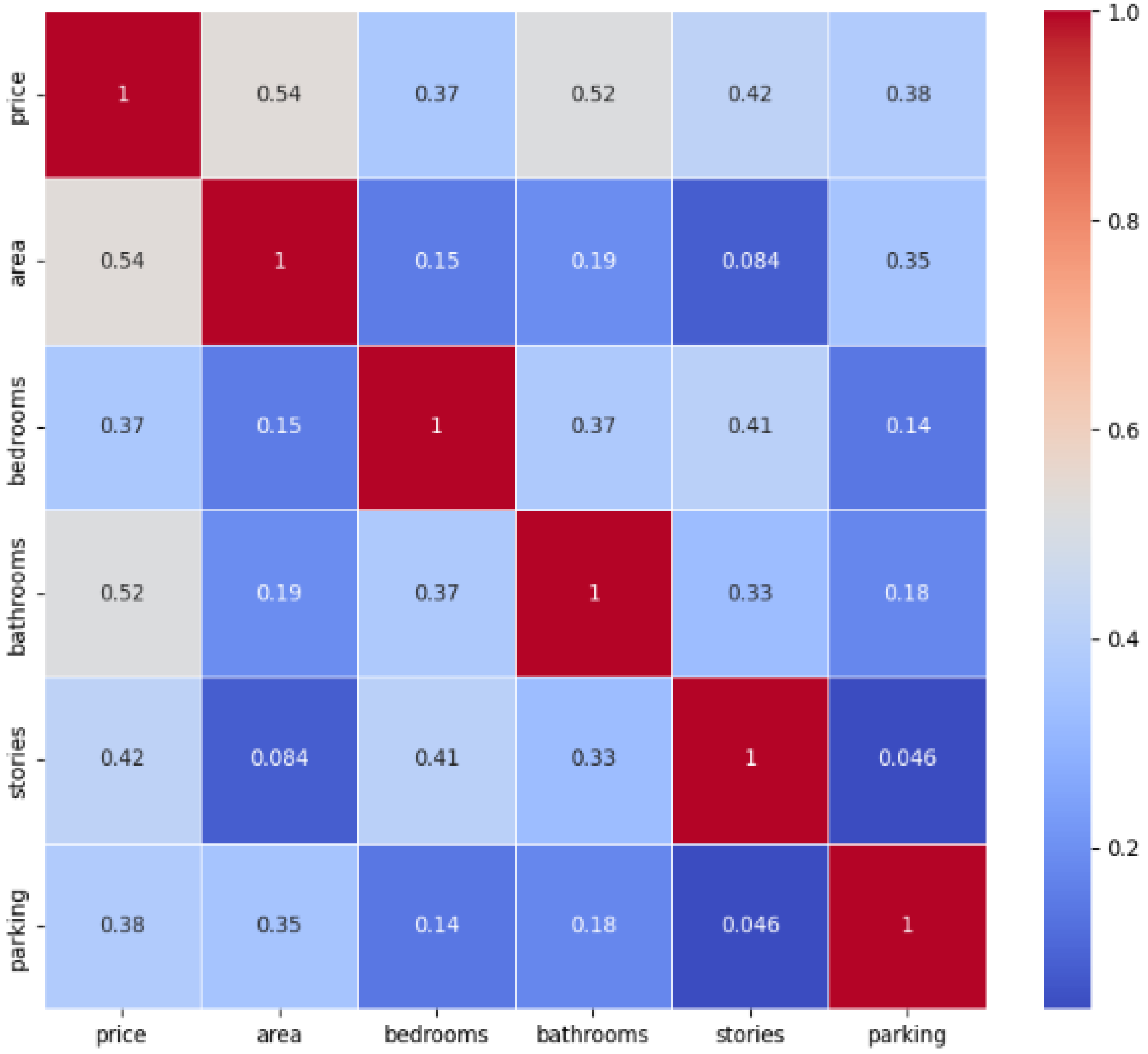
```python
# Importing necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

# Load the dataset (replace 'Housing_Price_Data.csv' with the actual file name if different)
df = pd.read_csv('Housing_Price_Data.csv')

# Check the first few rows of the dataset
df.head()
```

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished


price
0
20
40
60
80
0.2
0.4
0.6
0.8
1.0
1.2
1e7
area
0
20
40
60
80
100
2000
4000
6000
8000
10000
12000
14000
16000
bedrooms
0
100
200
300
1
2
3
4
5
6
bathrooms
0
100
200
300
400
1.0
1.5
2.0
2.5
3.0
3.5
4.0
stories
100
150
200
parking
200
300



```
[12] # Box plot to visualize outliers and spread
     plt.figure(figsize=(10, 6))
     sns.boxplot(data=df[numerical_columns])
     plt.show()
```


price
area
bedrooms
bathrooms
stories
parking
1
0.54
0.37
0.52
0.42
0.38
0.54
1
0.15
0.19
0.084
0.35
0.37
0.15
1
0.37
0.41
0.14
0.52
0.19
0.37
1
0.33
0.18
0.42
0.084
0.41
0.33
1
0.046
0.38
0.35
0.14
0.18
0.046
1
price
area
bedrooms
bathrooms
stories
parking
1.0
0.8
0.6
0.4
0.2