RIYADH REAL ESTATE

Tuwaiq Academy | AI Model Building & Development Bootcamp

Data Science Project

Prepared by
Nujud Almaleki

Dashboard Link:

GitHub Link:

Table of Contents

1. Introduction

1.1 Dataset

Title: Riyadh Real Estate | عقارات الرياض

Kaggle Link: https://www.kaggle.com/datasets/mohammedalsubaie/riyadh-real-estate?resource=download

Descriptions:

The data contains the features of properties offered in Riyadh (type, area, price, number of rooms, etc.). The aim is to analyze prices and market characteristics to understand trends and identify the factors influencing prices.

Why did I select this dataset?

Because it touches on the real estate market in Riyadh and allows for the extraction of strong insights that help in understanding market trends.

1.2 Problem Definition

My main goal is to explore how the housing market behaves and what factors impact property prices in this growing metropolis.

The main questions to be answered:

1. What are the most common property types listed in Riyadh?
2. How are **prices** distributed? Are there outliers or price clusters?
3. Does **area (m²)** have a strong relationship with **price**?
4. What is the impact of **Property Type** and **District** on the average **price**?
5. Which **property types** show the highest and lowest prices?
6. Which **locations** are expensive or affordable?
7. Do **rooms** and **bathrooms** impact **price**?
8. Are there correlations between **features**?

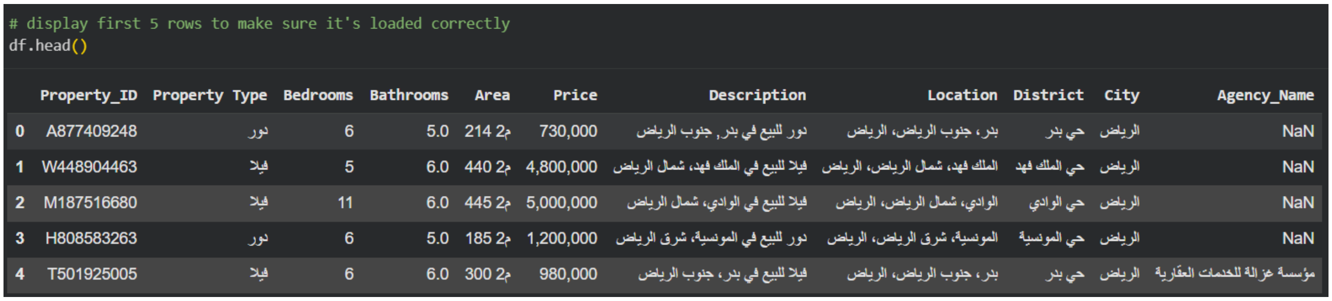
By the end of the analysis, I expect to identify:

- The dominant property type in the market
- Price patterns and typical ranges
- High-value and affordable neighborhoods
- The effect of property size and features on pricing
- Key factors that influence the market overall

2. Data Exploration

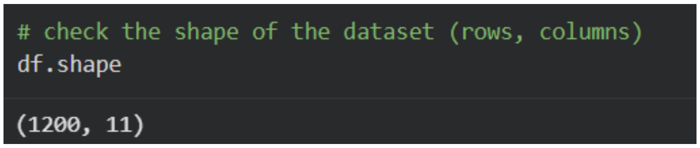
In this section, I explored the structure of the dataset using basic Pandas functions.

Images below show data structure, features, data types, and missing values.

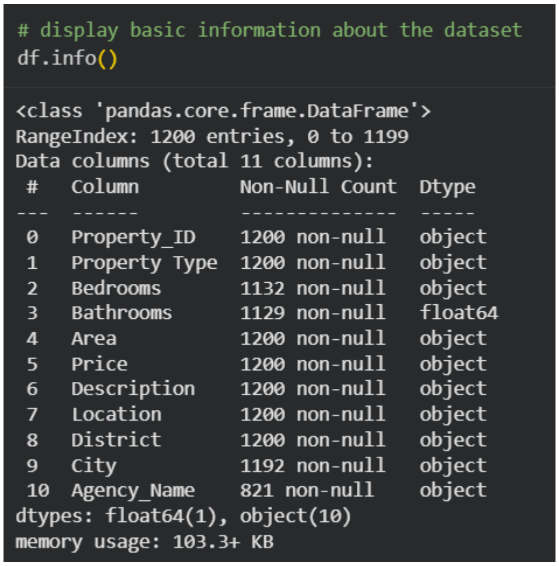


```
# display first 5 rows to make sure it's loaded correctly
df.head()
```

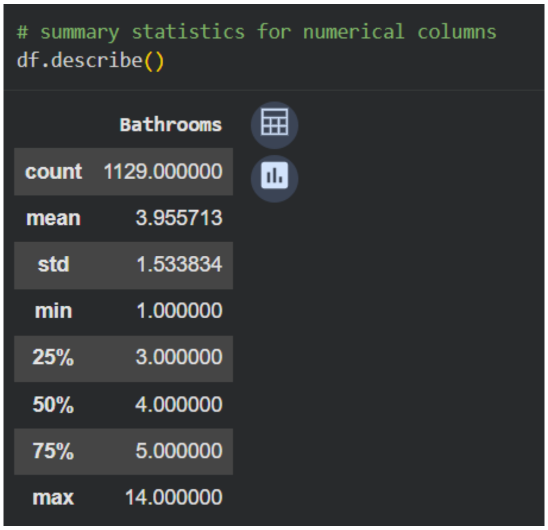
	Property_ID	Property Type	Bedrooms	Bathrooms	Area	Price	Description	Location	District	City	Agency_Name
0	A877409248	دور	6	5.0	214 م2	730,000	دور للبيع في بدر, جنوب الرياض	بدر، جنوب الرياض، الرياض	حي بدر	الرياض	NaN
1	W448904463	فيلا	5	6.0	440 م2	4,800,000	فيلا للبيع في الملك فهد، شمال الرياض	الملك فهد، شمال الرياض، الرياض	حي الملك فهد	الرياض	NaN
2	M187516680	فيلا	11	6.0	445 م2	5,000,000	فيلا للبيع في الوادي، شمال الرياض	الوادي، شمال الرياض، الرياض	حي الوادي	الرياض	NaN
3	H808583263	دور	6	5.0	185 م2	1,200,000	دور للبيع في المونسية، شرق الرياض	المونسية، شرق الرياض، الرياض	حي المونسية	الرياض	NaN
4	T501925005	فيلا	6	6.0	300 م2	980,000	فيلا للبيع في بدر، جنوب الرياض	بدر، جنوب الرياض، الرياض	حي بدر	الرياض	مؤسسة غزالة للخدمات العقارية



```
# check the shape of the dataset (rows, columns)
df.shape

(1200, 11)
```

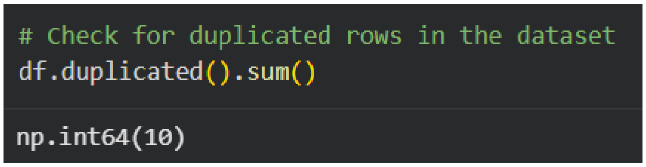


```
# display basic information about the dataset
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 11 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Property_ID    1200 non-null   object 
 1   Property Type  1200 non-null   object 
 2   Bedrooms       1132 non-null   object 
 3   Bathrooms      1129 non-null   float64
 4   Area           1200 non-null   object 
 5   Price          1200 non-null   object 
 6   Description    1200 non-null   object 
 7   Location       1200 non-null   object 
 8   District       1200 non-null   object 
 9   City           1192 non-null   object 
 10  Agency_Name    821 non-null    object 
dtypes: float64(1), object(10)
memory usage: 103.3+ KB
```



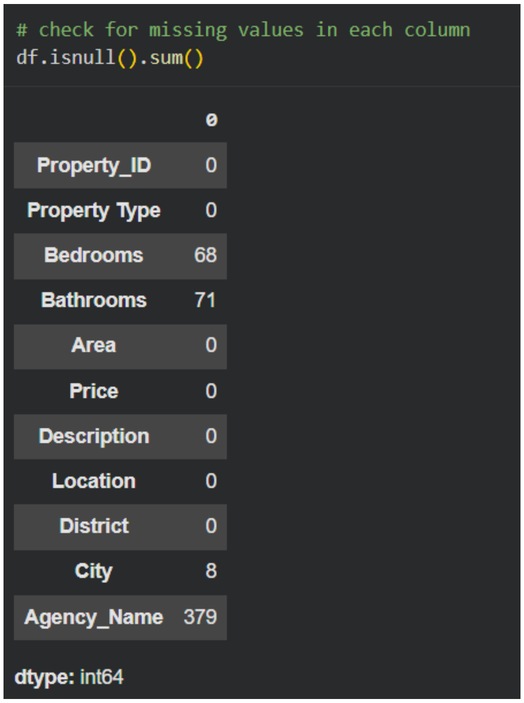
```
# summary statistics for numerical columns
df.describe()
```

	Bathrooms
count	1129.000000
mean	3.955713
std	1.533834
min	1.000000
25%	3.000000
50%	4.000000
75%	5.000000
max	14.000000



```
# Check for duplicated rows in the dataset
df.duplicated().sum()
```

```
np.int64(10)
```



```
# check for missing values in each column
df.isnull().sum()
```

```
                 0
  Property_ID    0
Property Type    0
     Bedrooms   68
    Bathrooms   71
         Area    0
        Price    0
  Description    0
     Location    0
     District    0
         City    8
  Agency_Name  379

dtype: int64
```

Main Observations:

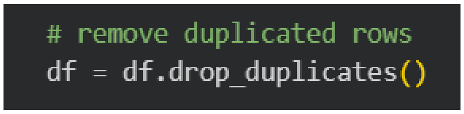
After loading the Riyadh Real Estate dataset, I explored its structure to understand the data before applying any cleaning or preprocessing steps, and I find that:

- The dataset contains 1200 rows and 11 columns.
- Most columns are categorical, while only one (Bathrooms) is numerical.
- The dataset includes missing values:
    - Bedrooms: 68 missing
    - Bathrooms: 71 missing
    - City: 8 missing
    - Agency_Name: 379 missing (very high)
- There are 10 duplicated rows that need to be removed.
- Column names are generally readable, but some values contain extra symbols (e.g., "2م" in Area) which need cleaning.

3. Data Cleaning

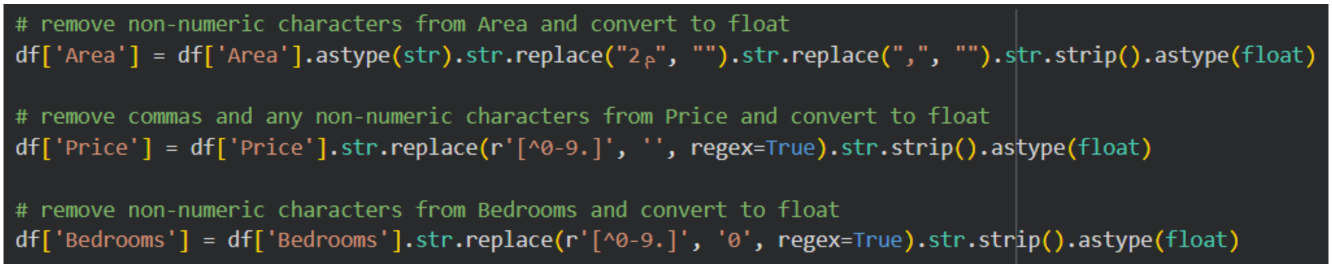
Several cleaning steps were applied to improve data quality and prepare for analysis:

1. Removed duplicates



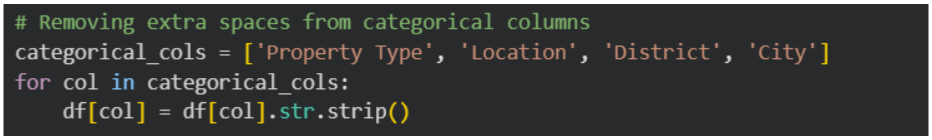
```
# remove duplicated rows
df = df.drop_duplicates()
```

2. Converted data types



```
# remove non-numeric characters from Area and convert to float
df['Area'] = df['Area'].astype(str).str.replace("2م", "").str.replace(",", "").str.strip().astype(float)

# remove commas and any non-numeric characters from Price and convert to float
df['Price'] = df['Price'].str.replace(r'[^0-9.]', '', regex=True).str.strip().astype(float)

# remove non-numeric characters from Bedrooms and convert to float
df['Bedrooms'] = df['Bedrooms'].str.replace(r'[^0-9.]', '0', regex=True).str.strip().astype(float)
```

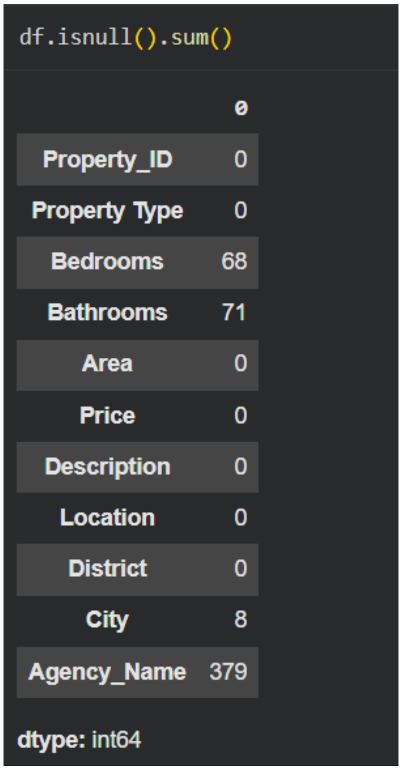
3. Handling Categorical Data



```
# Removing extra spaces from categorical columns
categorical_cols = ['Property Type', 'Location', 'District', 'City']
for col in categorical_cols:
    df[col] = df[col].str.strip()
```

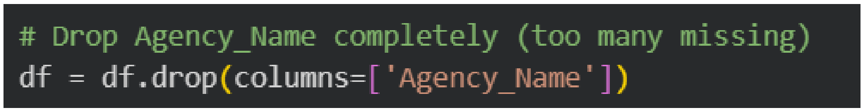
4. Handling Missing Values

First, show all the missing values, then handle each of them in the appropriate way.



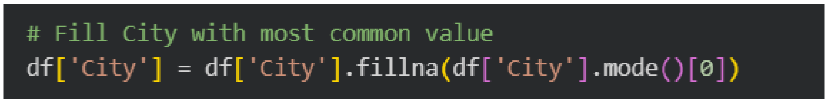
```
df.isnull().sum()

                 0
  Property_ID    0
Property Type    0
     Bedrooms   68
    Bathrooms   71
         Area    0
        Price    0
  Description    0
     Location    0
     District    0
         City    8
  Agency_Name  379

dtype: int64
```

a) 379 missing values in Agency_Name:
This column in the dataset is optional and has many missing values; therefore, the best approach here is to drop the entire column.



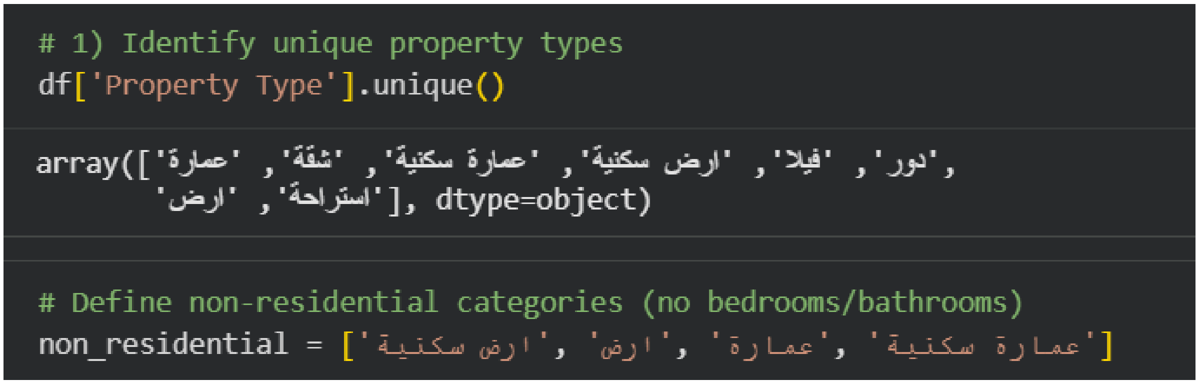
```
# Drop Agency_Name completely (too many missing)
df = df.drop(columns=['Agency_Name'])
```

b) 8 missing in City:
This data set is limited to properties in Riyadh, so logically, all values in the City column are Riyadh. Therefore, the appropriate method here is to fill in the missing values with the most frequently occurring value in the column.



```
# Fill City with most common value
df['City'] = df['City'].fillna(df['City'].mode()[0])
```

c) Bedrooms & Bathrooms Missing Values:
First, identify property types that do not have bedrooms or bathrooms (e.g., land or rest houses). They were classified as non-residential to avoid incorrect analysis and ensure the accuracy of the results.

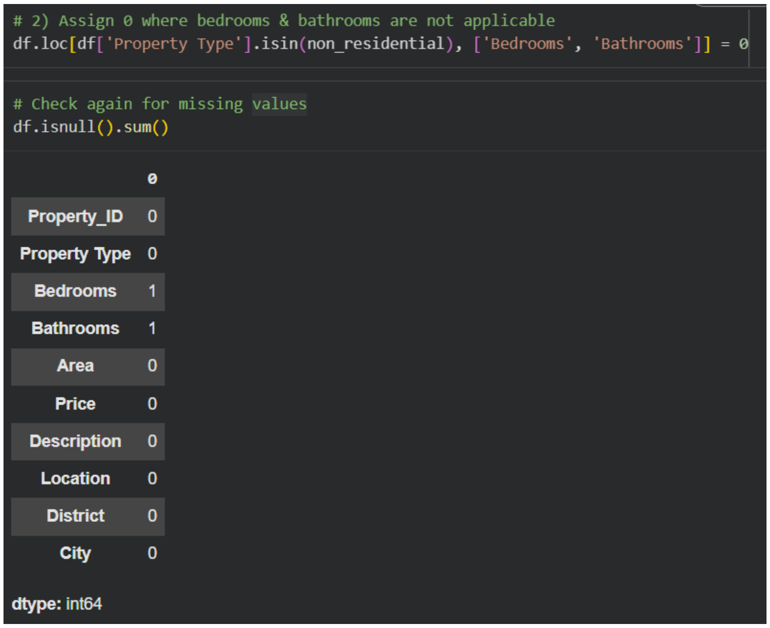


```
# 1) Identify unique property types
df['Property Type'].unique()
```

```
array(['عمارة', 'شقة', 'عمارة سكنية', 'ارض سكنية', 'فيلا', 'دور',
       'استراحة', 'ارض'], dtype=object)
```

```
# Define non-residential categories (no bedrooms/bathrooms)
non_residential = ['ارض سكنية', 'ارض', 'عمارة', 'عمارة سكنية']
```

For non-residential properties, bedrooms and bathrooms do not apply, so I assigned them a value of 0 to remove missing values and keep the dataset consistent for analysis.

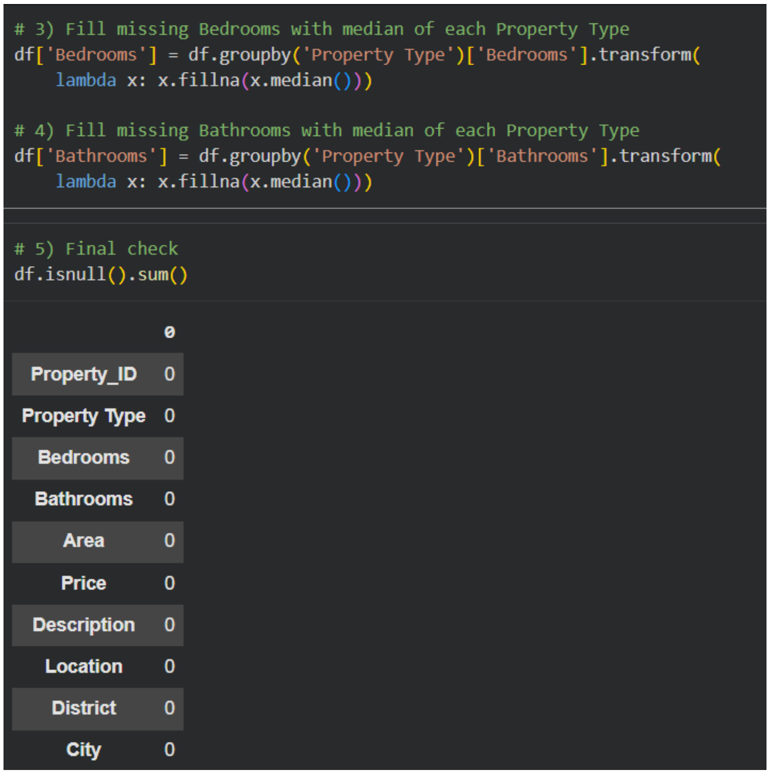


```
# 2) Assign 0 where bedrooms & bathrooms are not applicable
df.loc[df['Property Type'].isin(non_residential), ['Bedrooms', 'Bathrooms']] = 0
```

```
# Check again for missing values
df.isnull().sum()
```

```
               0
  Property_ID  0
Property Type  0
     Bedrooms  1
    Bathrooms  1
         Area  0
        Price  0
  Description  0
     Location  0
     District  0
         City  0

dtype: int64
```

The remaining missing bedroom and bathroom values were filled using the median of each property type. This keeps the data realistic and prevents skewing the analysis with incorrect or random values.

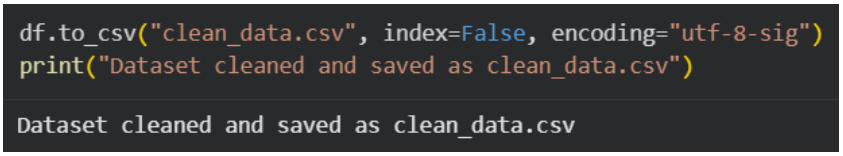


```
# 3) Fill missing Bedrooms with median of each Property Type
df['Bedrooms'] = df.groupby('Property Type')['Bedrooms'].transform(
    lambda x: x.fillna(x.median()))

# 4) Fill missing Bathrooms with median of each Property Type
df['Bathrooms'] = df.groupby('Property Type')['Bathrooms'].transform(
    lambda x: x.fillna(x.median()))
```

```
# 5) Final check
df.isnull().sum()
```

```
               0
  Property_ID  0
Property Type  0
     Bedrooms  0
    Bathrooms  0
         Area  0
        Price  0
  Description  0
     Location  0
     District  0
         City  0
```

5. Saved final clean dataset as clean_data.csv

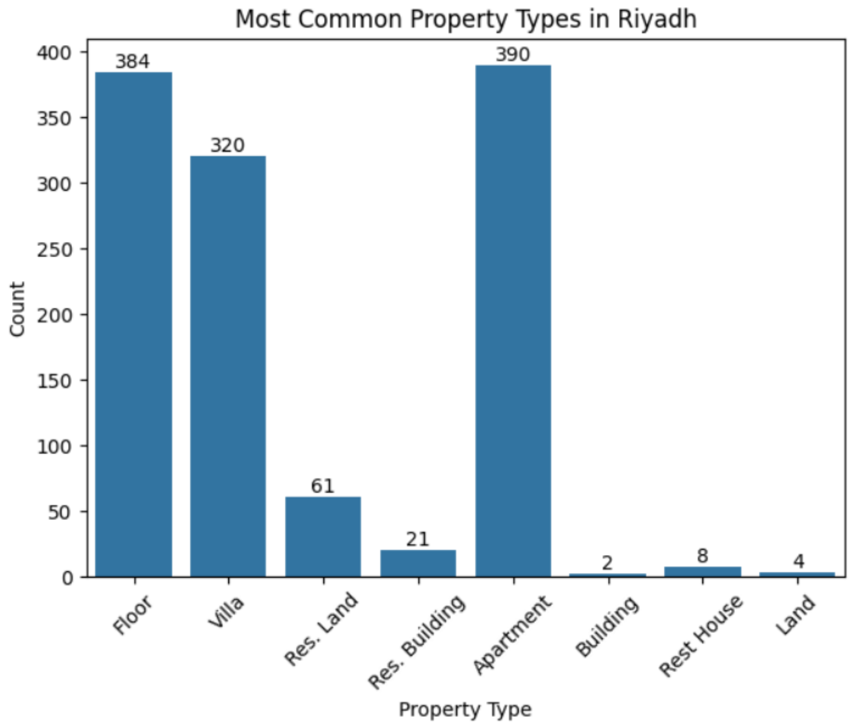


```
df.to_csv("clean_data.csv", index=False, encoding="utf-8-sig")
print("Dataset cleaned and saved as clean_data.csv")
```

```
Dataset cleaned and saved as clean_data.csv
```

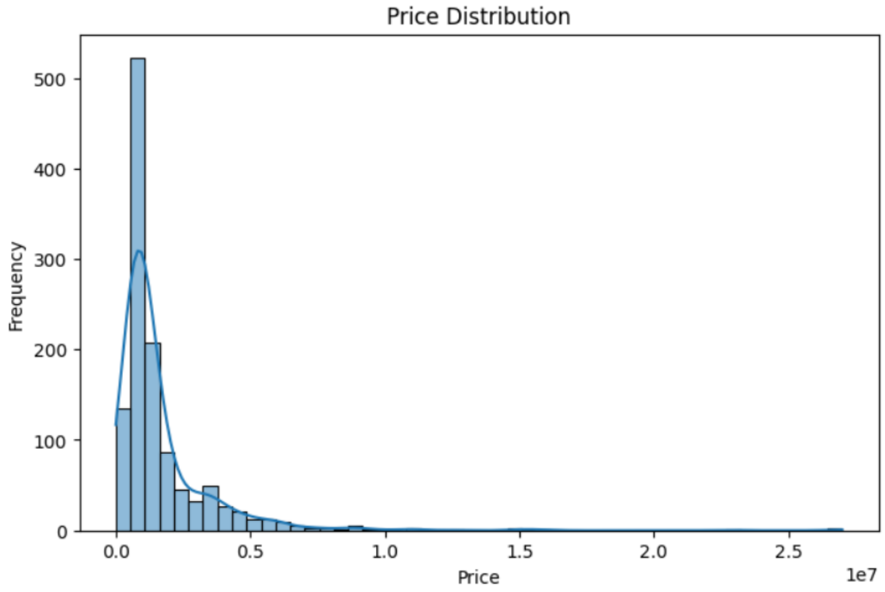
4. Exploratory Data Analysis (EDA)

1) What are the most common property types listed in Riyadh?

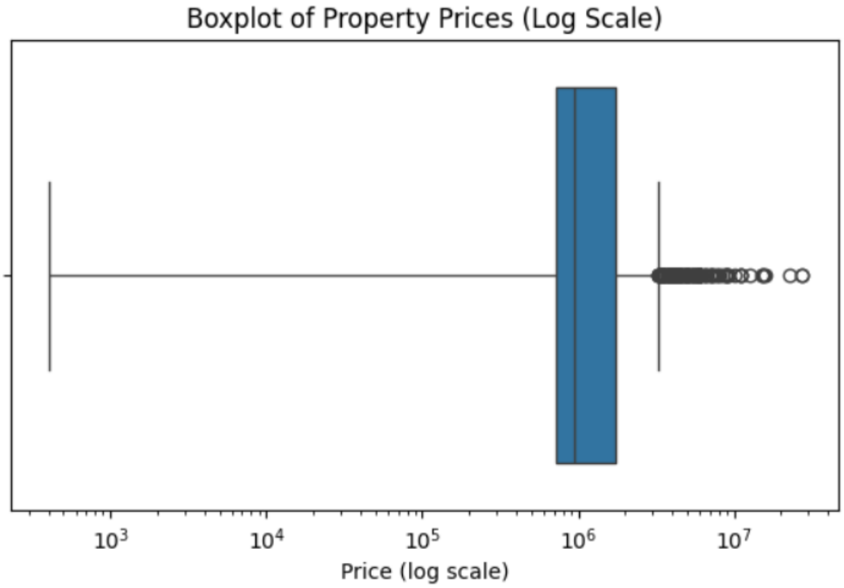


From the bar chart, it is clear that the most common property types in Riyadh are Apartment, Floor, and Villa, with each having over 300 listings. Other types such as Residential Land appear less frequently, while Building, Land, and Rest House occur only in a few records — indicating that they are rare within the dataset.

2) How are prices distributed? Are there outliers or price clusters?

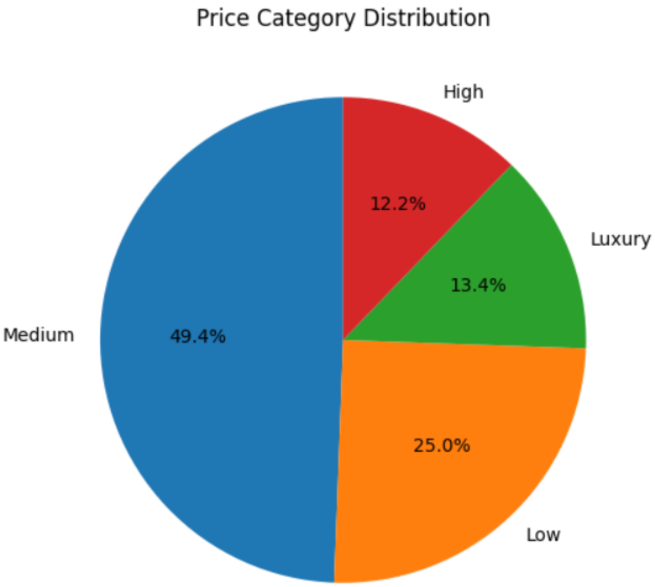


The histogram shows that most properties are priced in the lower range, indicating that affordable and mid-range properties make up the majority of the market. as prices increase, the number of properties decreases sharply, which means expensive units are not common.



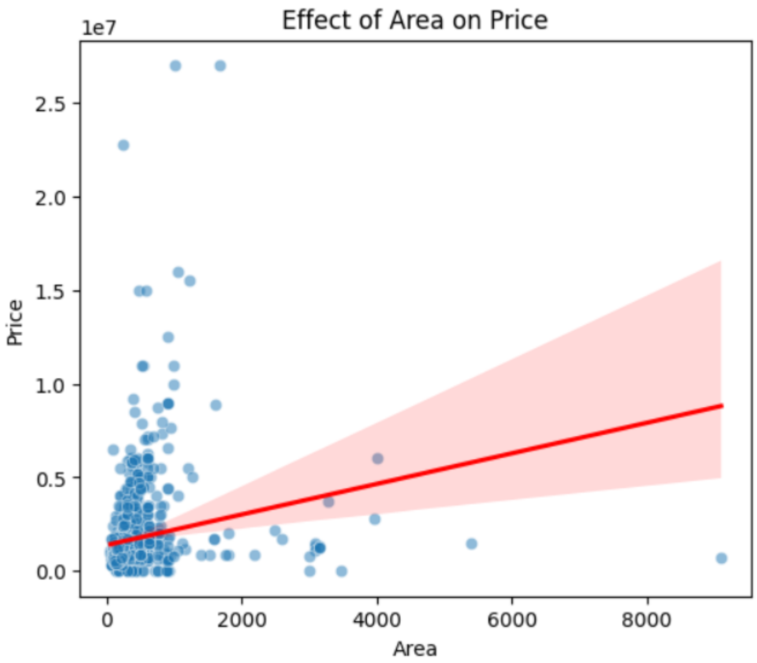
the boxplot confirms that a large number of properties are priced normally within a specific range, while several properties lie far outside this range, appearing as outliers. these outliers represent high-end or unique properties with significantly higher prices.

both plots together show that Riyadh's real estate market is mostly concentrated around moderate prices, with a smaller luxury segment that appears as outliers — this indicates a divided market with typical pricing and a premium category.



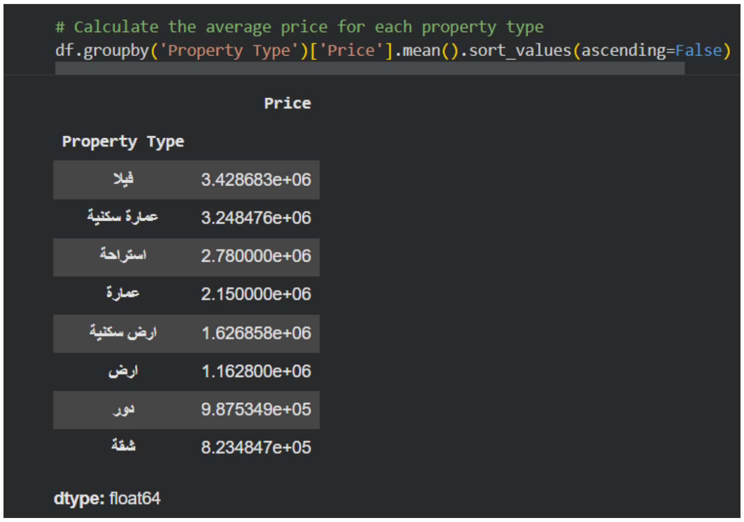
To better understand the price distribution, properties were grouped into four categories using the IQR method. nearly half of the listings fall into the medium-price range, while only a small portion belong to the high and luxury categories. this supports the boxplot findings, showing that the market is mainly concentrated around typical prices with fewer premium properties.

3) Does area (m²) have a strong relationship with price?



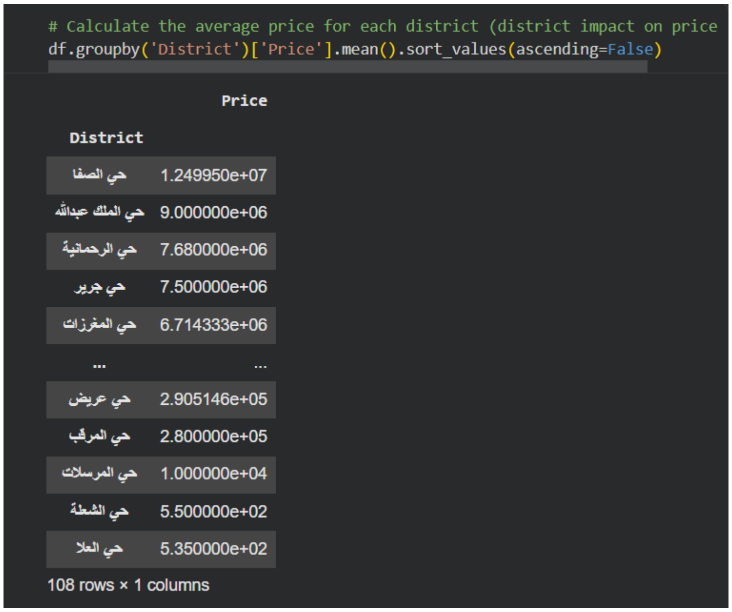
The scatter plot with regression line shows a weak positive relationship between area and price. Most properties are clustered within small areas (under 1,000 m²), but their prices vary significantly. Some large-area properties do not have high prices, indicating that area alone is not a strong predictor of price. The presence of outliers suggests that other features may influence price more than area.

4) What is the impact of Property Type and District on the average price?



```
# Calculate the average price for each property type
df.groupby('Property Type')['Price'].mean().sort_values(ascending=False)
```

```
                      Price
Property Type
         فيلا    3.428683e+06
   عمارة سكنية    3.248476e+06
       استراحة    2.780000e+06
         عمارة    2.150000e+06
     ارض سكنية    1.626858e+06
           ارض    1.162800e+06
           دور    9.875349e+05
           شقة    8.234847e+05

dtype: float64
```



```
# Calculate the average price for each district (district impact on price
df.groupby('District')['Price'].mean().sort_values(ascending=False)
```

```
                    Price
District
حي الصفا            1.249950e+07
حي الملك عبدالله    9.000000e+06
حي الرحمانية        7.680000e+06
حي جرير             7.500000e+06
حي المغرزات         6.714333e+06
...                 ...
حي عريض             2.905146e+05
حي المرقب           2.800000e+05
حي المرسلات         1.000000e+04
حي الشعلة           5.500000e+02
حي العلا            5.350000e+02
108 rows × 1 columns
```

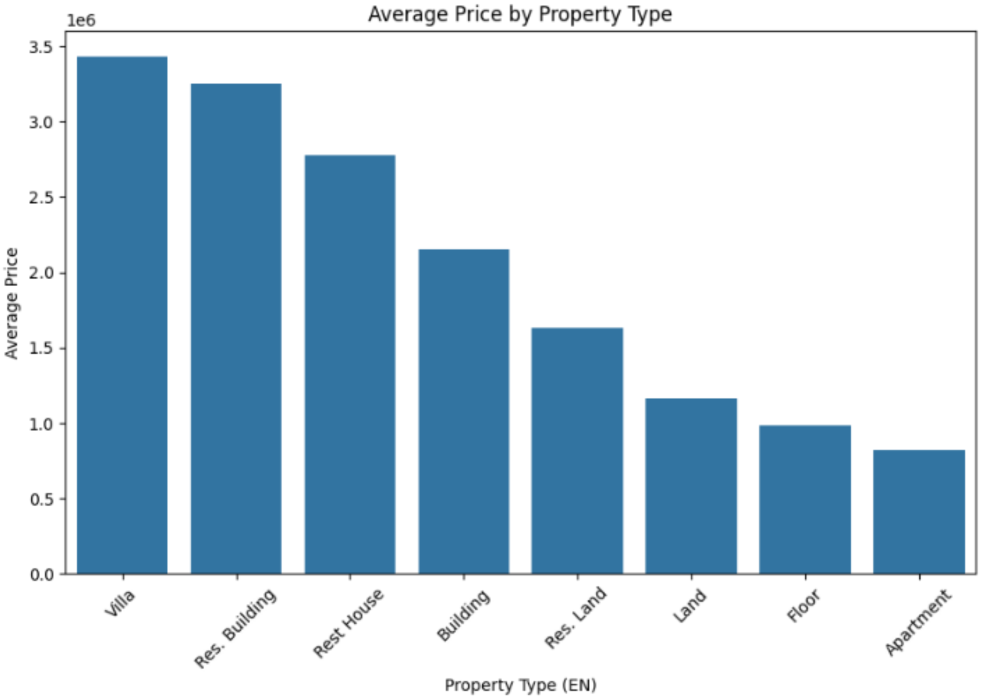
The results show that both Property Type and District have a strong impact on pricing.

Villas and residential buildings recorded the highest average prices, while other property types such as apartments showed lower averages.

Similarly, there is a significant variation in prices between districts, where neighborhoods such as حي الصفا and حي الملك عبدالله have much higher average prices compared to other areas.

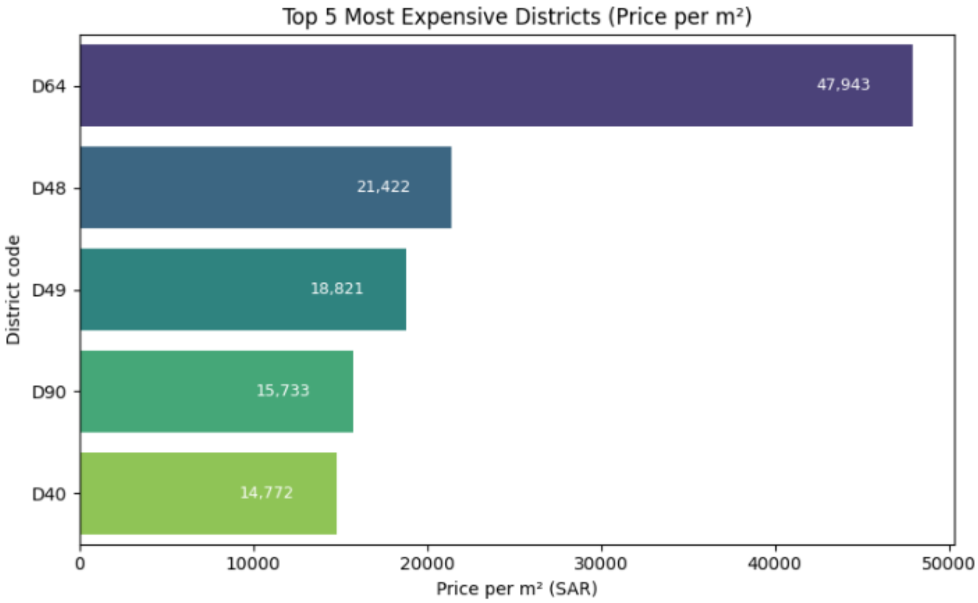
This indicates that location and property type are key factors affecting real estate value, making them strong candidates as important features for further predictive modeling.

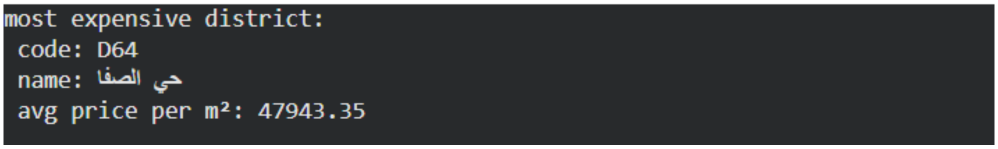
5) Which property types show the highest and lowest prices?



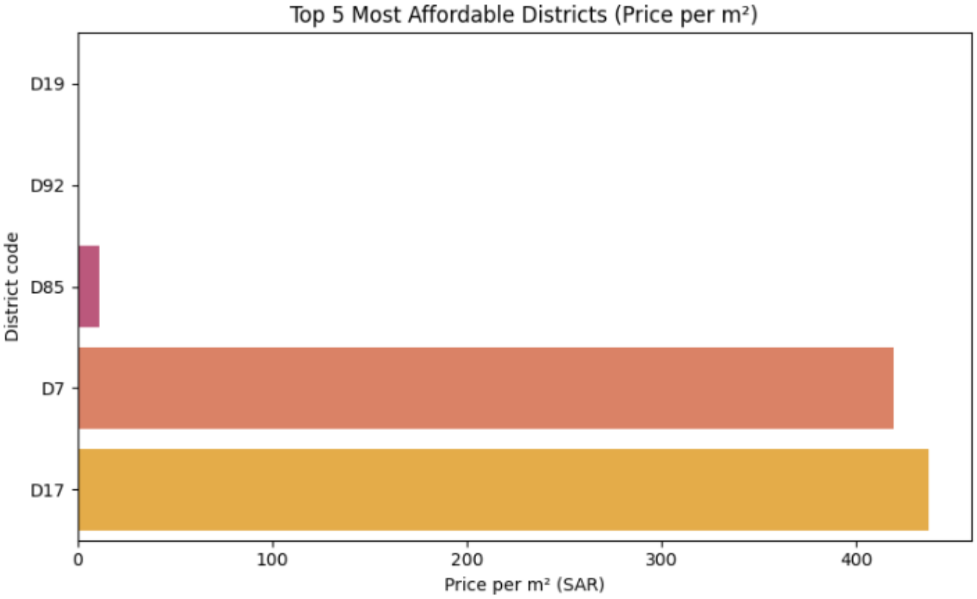
The bar chart shows the average price for each property type. Villas and residential buildings have the highest average prices, while apartments and single-floor properties show the lowest prices. This indicates that property type has a strong influence on real estate pricing, making it a key factor to consider in further analysis and prediction models.

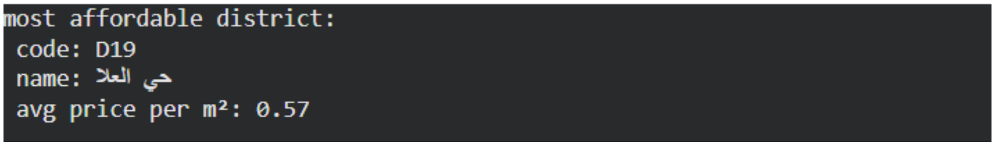
6) Which locations are expensive or affordable?





```
most expensive district:
 code: D64
 name: حي الصفا
 avg price per m²: 47943.35
```





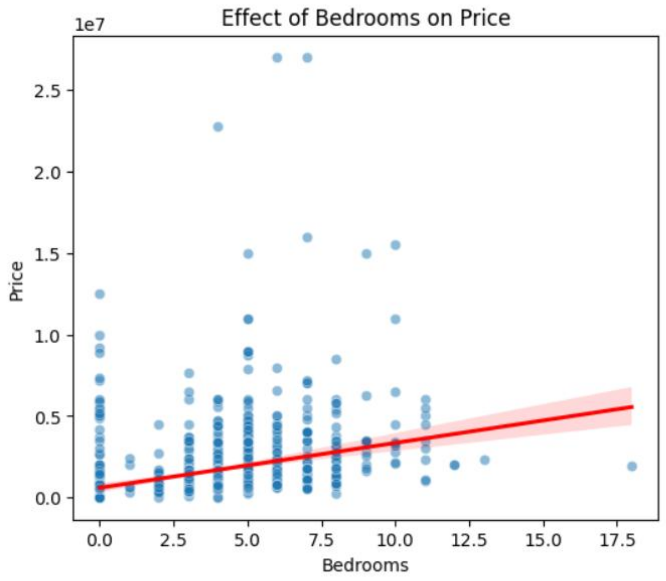
```
most affordable district:
 code: D19
 name: حي العلا
 avg price per m²: 0.57
```

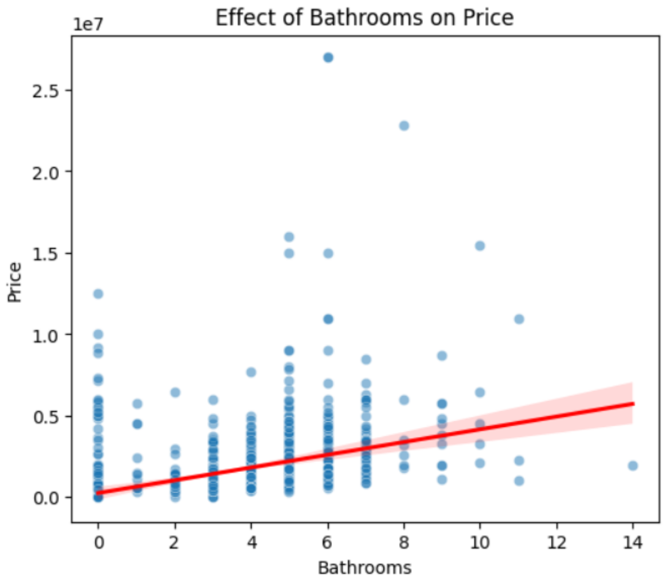
The visualizations clearly show a significant difference between districts in terms of average price per square meter.

- From the first chart, district D64 (حي الصفا) stands out as the most expensive area with a very high average price per m² (around 47,943 SAR), which is much higher than the other top districts. This indicates that D64 is likely a premium location, possibly offering better services, facilities, or a strategic location.
- On the other hand, the second chart highlights the most affordable districts, where D19 (حي العلا) has the lowest average price per m² (around 0.57 SAR), which may indicate that this area is still under development or lacks infrastructure.

the visualizations show a strong relationship between district and price. Some districts have significantly higher average prices per m², while others are much more affordable. This large variation indicates that district is one of the most influential factors in determining property value.

7) Do rooms and bathrooms impact price?

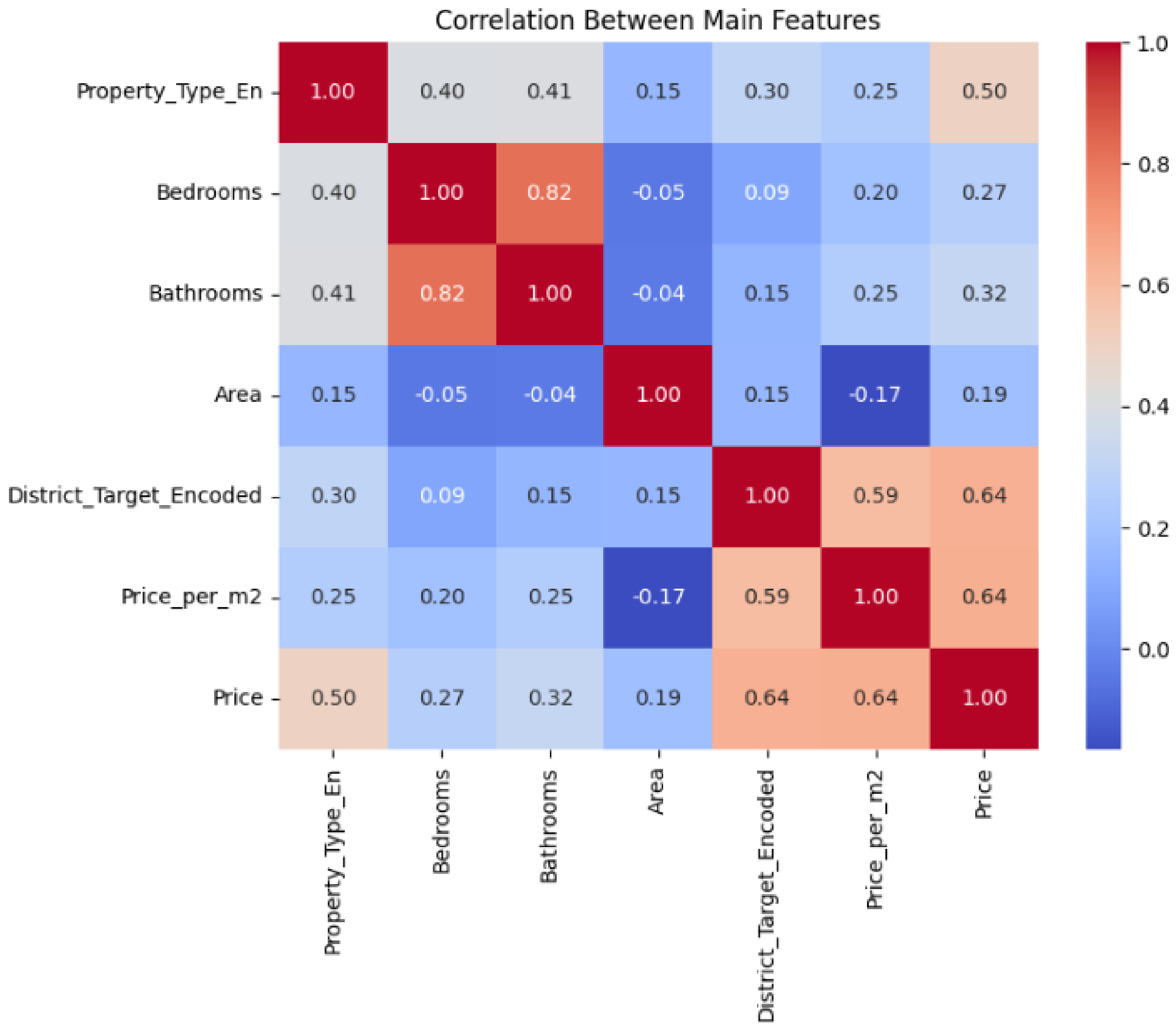




The scatter plot shows that adding more bedrooms generally increases the price, but the relationship is not very strong. The regression line has a slight positive slope, meaning that properties with more bedrooms tend to be more expensive. However, the data points are widely scattered, which indicates that the price is also affected by other factors besides bedrooms (such as location, land size, or overall house quality).

Similar to bedrooms, Bathrooms also have a weak positive impact on price. They contribute to price increase, but they alone cannot predict property price accurately.

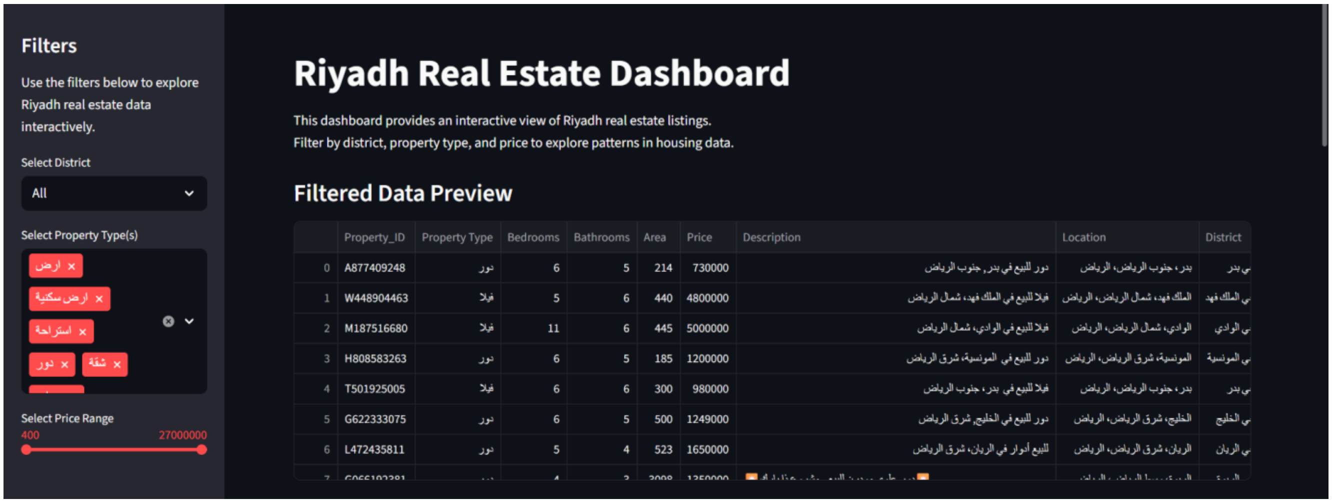
8) Are there correlations between features?



The heatmap shows the correlations between the main features and the property price. Categorical variables such as property type and district were encoded to numerical form to make them compatible with correlation analysis. After encoding, the results revealed that price per square meter and district are the most influential features, followed by property type, bathrooms, and bedrooms. This indicates that both the location and property characteristics have a strong impact on real estate prices, making them important predictors for future modeling.

5. Dashboard Overview

Thث dashboard offers an interactive exploration of Riyadh's real estate market. Users can apply filters such as district, property type, and price range to instantly update the dataset preview and uncover patterns in the housing data. The dashboard also provides real-time summary statistics, including total property count, average property price, and average price per square meter. Visualizations, such as the 'Average Price by District' plot, help highlight differences across neighborhoods, making it easier to identify high-value areas and trends. Overall, the dashboard transforms static analysis into an engaging, user-friendly tool for deeper market understanding.



Filters

Use the filters below to explore Riyadh real estate data interactively.

Select District: All

Select Property Type(s): ارض ×, ارض سكنية ×, استراحة ×, دور ×, شقة ×

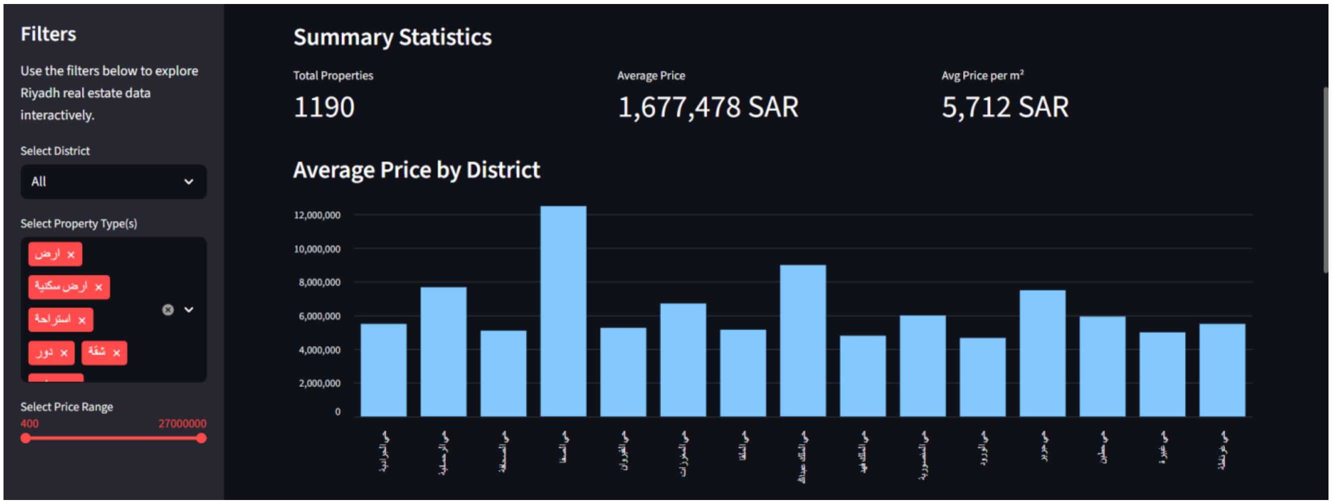
Select Price Range: 400 – 27000000

Riyadh Real Estate Dashboard

This dashboard provides an interactive view of Riyadh real estate listings.
Filter by district, property type, and price to explore patterns in housing data.

Filtered Data Preview

	Property_ID	Property Type	Bedrooms	Bathrooms	Area	Price	Description	Location	District
0	A877409248	دور	6	5	214	730000	دور للبيع في بدر, جنوب الرياض	بدر ، جنوب الرياض، الرياض	ي بدر
1	W448904463	فيلا	5	6	440	4800000	فيلا للبيع في الملك فهد، شمال الرياض	الملك فهد، شمال الرياض، الرياض	ي الملك فهد
2	M187516680	فيلا	11	6	445	5000000	فيلا للبيع في الوادي، شمال الرياض	الوادي، شمال الرياض، الرياض	ي الوادي
3	H808583263	دور	6	5	185	1200000	دور للبيع في المونسية، شرق الرياض	المونسية، شرق الرياض، الرياض	ي المونسية
4	T501925005	فيلا	6	6	300	980000	فيلا للبيع في بدر ، جنوب الرياض	بدر ، جنوب الرياض، الرياض	ي بدر
5	G622333075	دور	6	5	500	1249000	دور للبيع في الخليج, شرق الرياض	الخليج، شرق الرياض، الرياض	ي الخليج
6	L472435811	دور	5	4	523	1650000	للبيع أدوار في الريان، شرق الرياض	الريان، شرق الرياض، الرياض	ي الريان



Filters

Use the filters below to explore Riyadh real estate data interactively.

Select District: All

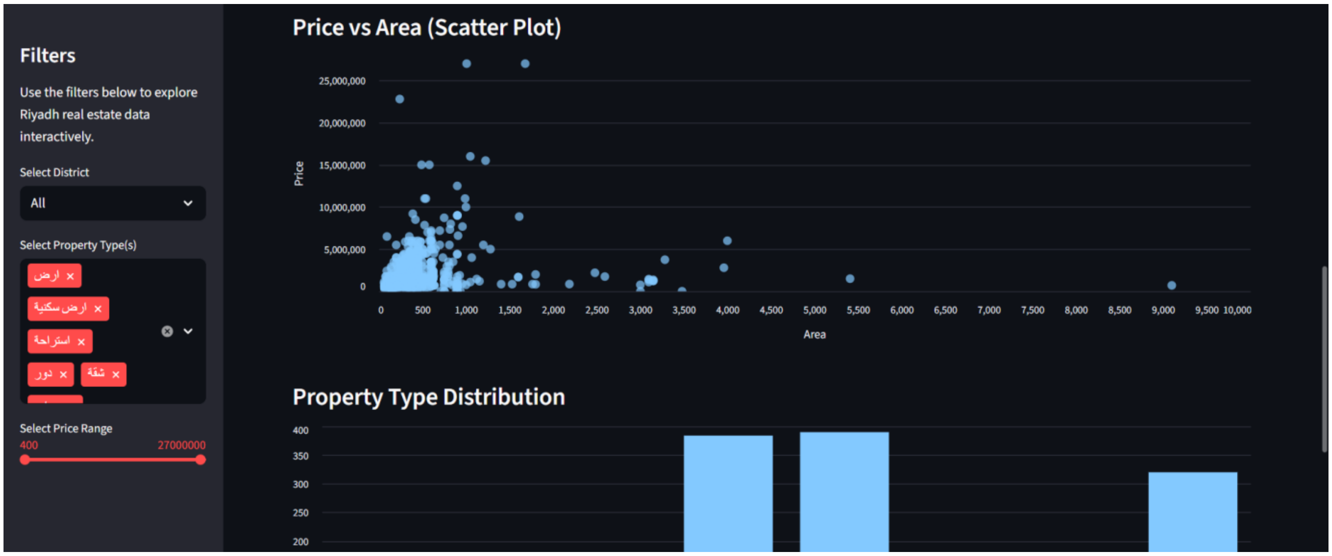
Select Property Type(s): ارض ×, ارض سكنية ×, استراحة ×, دور ×, شقة ×

Select Price Range: 400 – 27000000

Summary Statistics

Total Properties	Average Price	Avg Price per m²
1190	1,677,478 SAR	5,712 SAR

Average Price by District



Filters

Use the filters below to explore Riyadh real estate data interactively.

Select District

All

Select Property Type(s)

ارض ×

ارض سكنية ×

استراحة ×

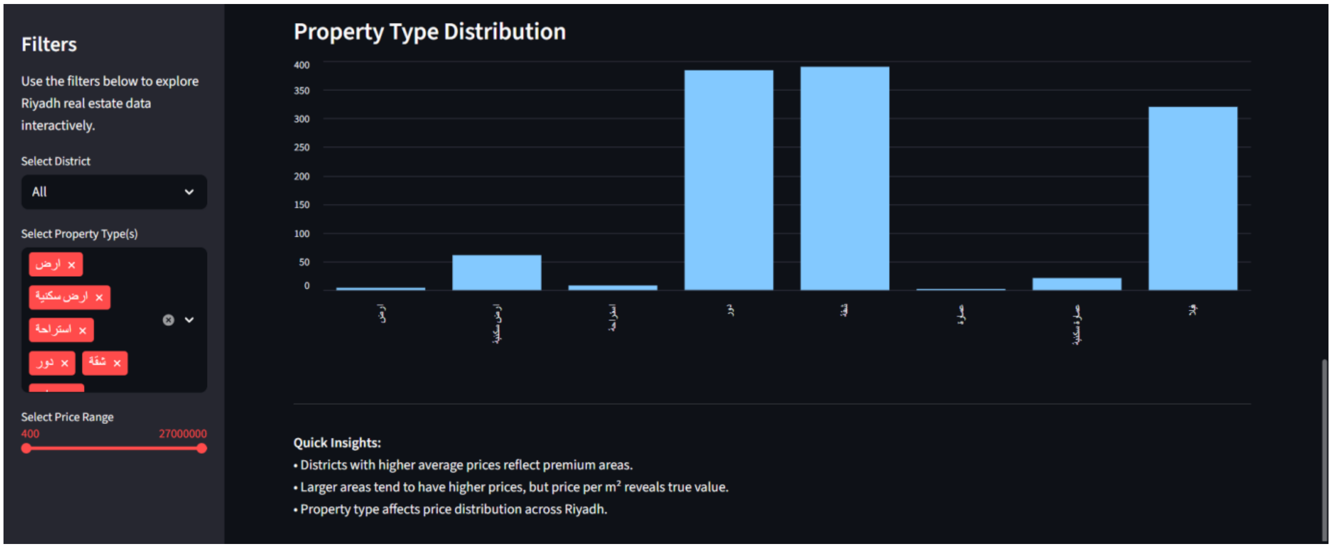
دور × شقة ×

Select Price Range

400 27000000

Price vs Area (Scatter Plot)

Property Type Distribution



Filters

Use the filters below to explore Riyadh real estate data interactively.

Select District

All

Select Property Type(s)

ارض ×

ارض سكنية ×

استراحة ×

دور × شقة ×

Select Price Range

400 27000000

Property Type Distribution

Quick Insights:

- Districts with higher average prices reflect premium areas.
- Larger areas tend to have higher prices, but price per m² reveals true value.
- Property type affects price distribution across Riyadh.

6. Insights & Conclusion

After exploring the Real Estate dataset in Riyadh and answering the main analytical questions, here are the key insights:

---

1. Relationship Between Price and District (Neighborhood)

Insight: There is a clear variation in average property prices across different districts. Some districts consistently show significantly higher prices per square meter, indicating that location is a strong factor influencing real estate value. This highlights the importance of district selection when evaluating property investments.

---

2. Impact of Property Type on Price

Insight: Property type has a noticeable effect on the final price. Certain types of real estate (e.g., villas or residential buildings) tend to be more expensive than apartments or land. This suggests that property type plays a major role in determining market value and investment potential.

---

3. Area and Price Relationship

Insight: There is a positive relationship between area and price, meaning larger properties generally tend to cost more. However, some outliers show properties with small areas but high prices — this may indicate premium locations or renovated properties with higher value.

---

4. Price per Square Meter as a Better Indicator

Insight: By creating the new feature Price_per_m2, a clearer pattern emerged. This metric helped compare properties more fairly and showed which districts provide better value for money. It also revealed districts with overpriced properties compared to their actual area.

---

5. District and Price_per_m2 – Strongest Relationship

Insight: The district showed the strongest correlation with Price_per_m2. This confirms that location is the most influential factor in Riyadh's real estate market, more than rooms, bathrooms, or area alone.

Overall Conclusion:

- Location (district) is the most important factor affecting real estate prices.
- Property type also plays a noticeable role in determining price.
- Area correlates with price, but Price_per_m2 gives a more accurate understanding of real value.
- Outliers suggest the presence of luxury properties or unique investment opportunities.

These insights provide a clearer understanding of the Riyadh real estate market and can support smarter decision-making for buyers, investors, and urban planning.

What I Learned:

- How to clean and analyze real-world data using Pandas.
- How visualization helps reveal valuable market patterns.

- How to build a simple interactive dashboard with Streamlit.
- How to extract insights that support decision-making.

Limitations of the Dataset:

- Some property types lacked complete information (rooms/bathrooms).
- District naming was inconsistent and required encoding.
- No geographic coordinates, which limited location-based visual analysis.
- No time dimension, making it difficult to study price trends over years.

Recommendations for Future Analysis

- Include historical data to analyze price trends over time.
- Add geographic coordinates to perform mapping and spatial analysis.
- Collect more property features (age, condition, amenities).
- Apply predictive modeling to estimate property prices more accurately.