

Housing Dataset Cleaning & Visualization

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [2]: # Load dataset
df = pd.read_csv("Housing.csv")
df.head()
```

```
Out[2]:
```

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotw
0	13300000	7420	4	2	3	yes	no	no	
1	12250000	8960	4	4	4	yes	no	no	
2	12250000	9960	3	2	2	yes	no	yes	
3	12215000	7500	4	2	2	yes	no	yes	
4	11410000	7420	4	1	2	yes	yes	yes	

◀ ————— ▶

Handle Missing Values (Simulated)

```
In [3]: # Simulate missing values
np.random.seed(123)
missing_indices = np.random.choice(df.index, size=20, replace=False)
df.loc[missing_indices[:10], 'bedrooms'] = np.nan
df.loc[missing_indices[10:], 'bathrooms'] = np.nan

# Fill missing values with median
df['bedrooms'] = df['bedrooms'].fillna(df['bedrooms'].median())
df['bathrooms'] = df['bathrooms'].fillna(df['bathrooms'].median())
```

Handle Outliers using IQR

```
In [4]: def cap_outliers(series):
    Q1 = series.quantile(0.25)
    Q3 = series.quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    return np.clip(series, lower, upper)
```

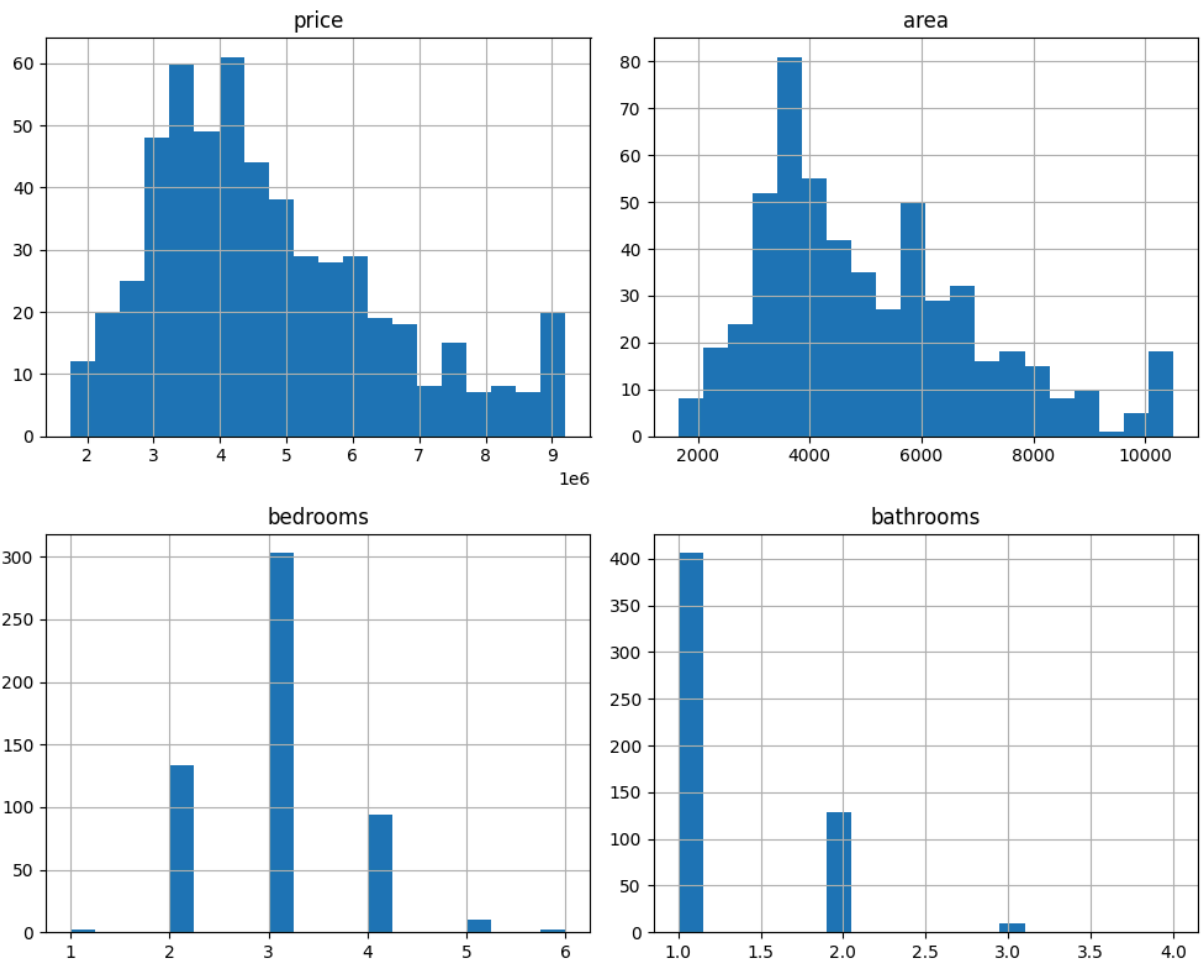
```
df['price'] = cap_outliers(df['price'])
df['area'] = cap_outliers(df['area'])
```

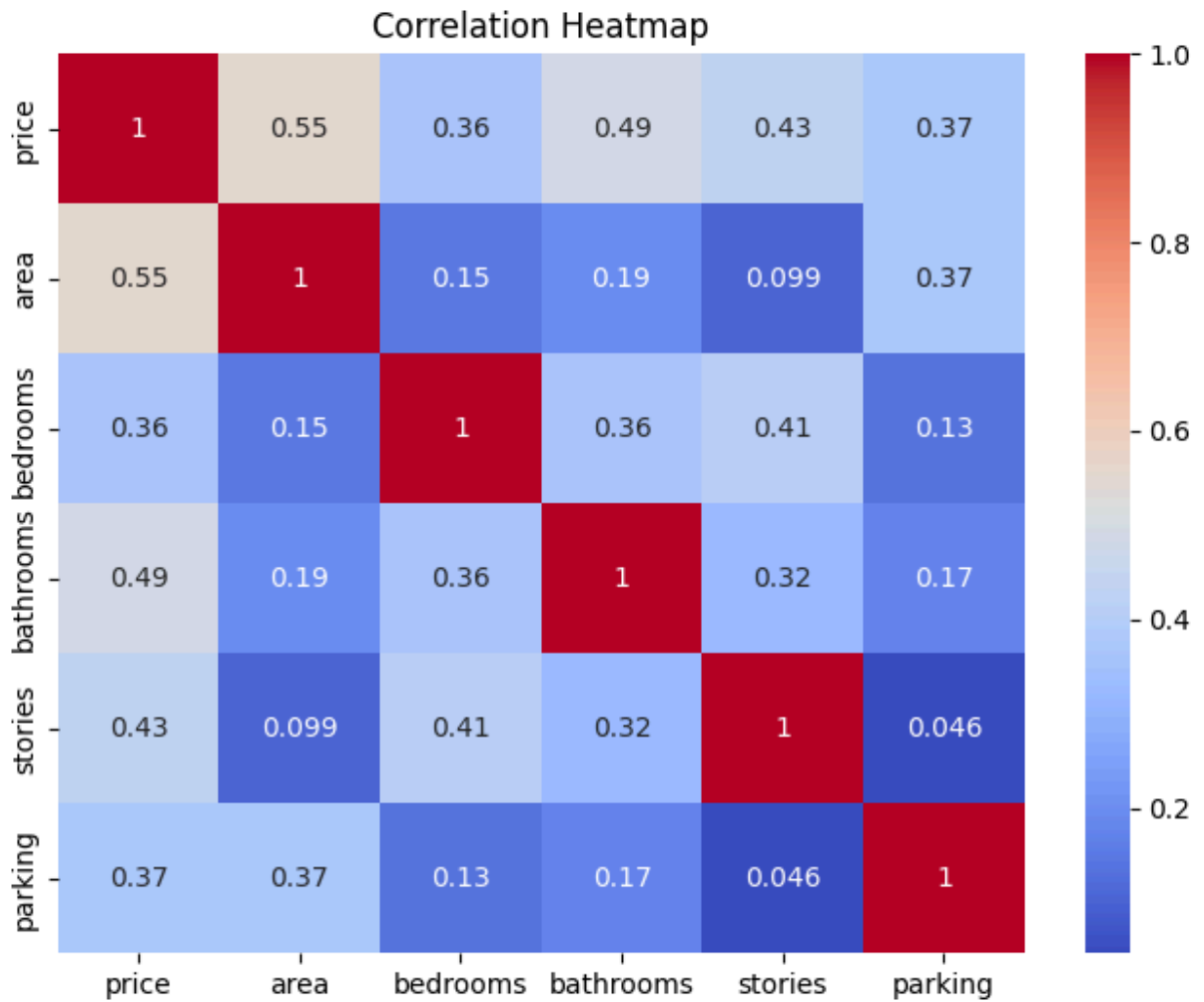


Data Visualization

```
In [5]: # Histograms after cleaning . These plots show the distribution of numeric features
df[['price', 'area', 'bedrooms', 'bathrooms']].hist(bins=20, figsize=(10, 8))
plt.tight_layout()
plt.show()

# Correlation heatmap
plt.figure(figsize=(8, 6))
numeric_df = df.select_dtypes(include='number') # keep only numeric columns
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```





Save Cleaned Dataset

```
In [6]: df.to_csv("Housing_Cleaned.csv", index=False)
print("Cleaned dataset saved as Housing_Cleaned.csv")
```

Cleaned dataset saved as Housing_Cleaned.csv