EDA and Insight Analysis Report

Colombo Apartments Pricing Data

Table of Contents

Introduction

The real estate property market is a market of significant growth and one with highly competitive prices. This study aims to analyze the rental landscape in Colombo by leveraging a comprehensive dataset of apartment listings from various neighborhoods.

The data set comprises various apartments from different neighborhoods across Colombo.

This report will explore how rental prices vary across neighborhoods, identify the most influential property attributes, and provide data-driven recommendations for stakeholders in the real estate industry. We wish to uncover how rental price trends vary across features and variables to uncover unique insights on the real estate market for Apartments in Colombo using the given dataset.

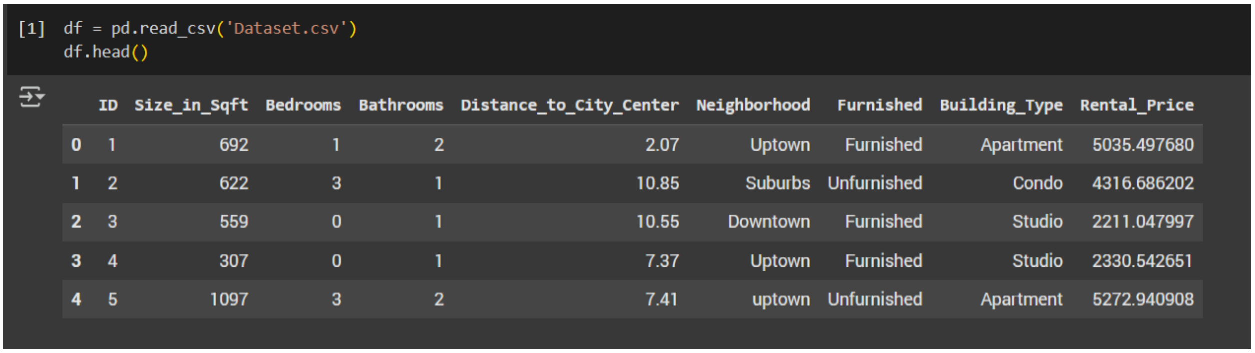
1.1 Dataset Overview

The first step in our analysis was to understand the dataset and its features. The case study document underlines the features and what they stand for.

Dataset Features

1. Apartment_ID: Unique identifier for each apartment listing
2. Neighborhood: Name of the neighbourhoods
3. Rental_Price: Monthly rental price
4. Size_in_Sqft: Size of the apartment in square feet
5. Distance_to_City_Center_KM: Distance to the Colombo Fort Station
6. Bedrooms: Number of bedrooms
7. Bathrooms: Number of bathrooms
8. Furnished: Whether the apartment is furnished
9. Building_Type: Type of the building

After loading the dataset to a Pandas Data frame and checking for the data shapes, we found the original dataset to contain 252 records across 9 features.

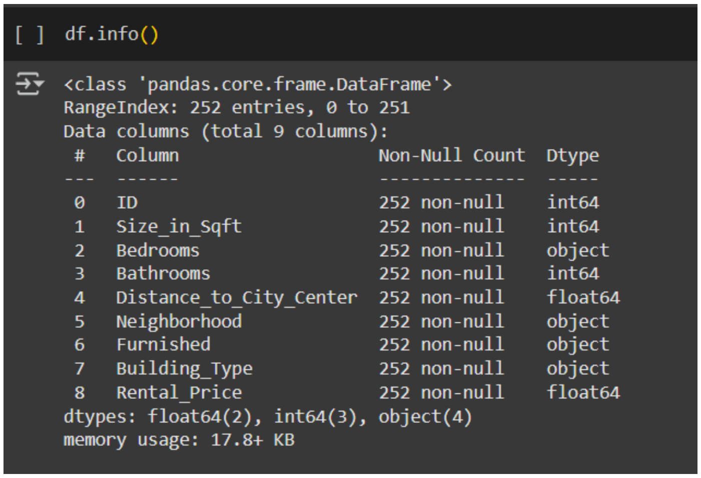


```
[1] df = pd.read_csv('Dataset.csv')
    df.head()
```

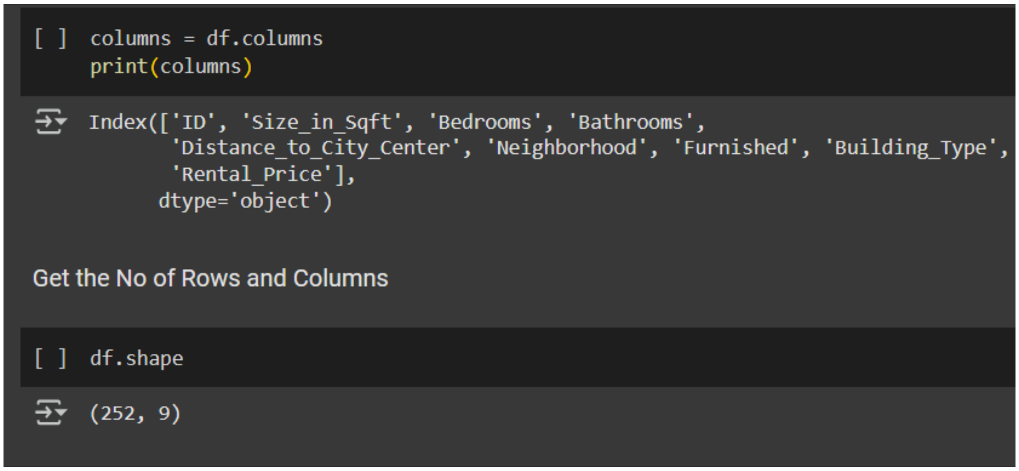
	ID	Size_in_Sqft	Bedrooms	Bathrooms	Distance_to_City_Center	Neighborhood	Furnished	Building_Type	Rental_Price
0	1	692	1	2	2.07	Uptown	Furnished	Apartment	5035.497680
1	2	622	3	1	10.85	Suburbs	Unfurnished	Condo	4316.686202
2	3	559	0	1	10.55	Downtown	Furnished	Studio	2211.047997
3	4	307	0	1	7.37	Uptown	Furnished	Studio	2330.542651
4	5	1097	3	2	7.41	uptown	Unfurnished	Apartment	5272.940908

Feature types for data columns include,

Categorical columns: Bedrooms, Neighborhood, Furnished, Building_Type
Numerical Columns: Distance_to_city_Cente, Rental_Price, Bathrooms, ID, Size_in_sqft



```
[ ] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252 entries, 0 to 251
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   ID                       252 non-null    int64
 1   Size_in_Sqft             252 non-null    int64
 2   Bedrooms                 252 non-null    object
 3   Bathrooms                252 non-null    int64
 4   Distance_to_City_Center  252 non-null    float64
 5   Neighborhood             252 non-null    object
 6   Furnished                252 non-null    object
 7   Building_Type            252 non-null    object
 8   Rental_Price             252 non-null    float64
dtypes: float64(2), int64(3), object(4)
memory usage: 17.8+ KB
```



```
[ ] columns = df.columns
    print(columns)

Index(['ID', 'Size_in_Sqft', 'Bedrooms', 'Bathrooms',
       'Distance_to_City_Center', 'Neighborhood', 'Furnished', 'Building_Type',
       'Rental_Price'],
      dtype='object')
```

Get the No of Rows and Columns

```
[ ] df.shape

(252, 9)
```

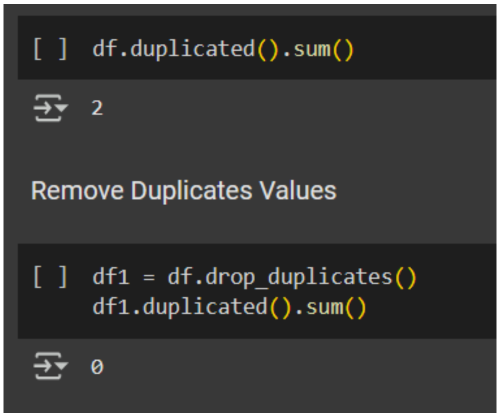
The most significant feature of this study is taken as the rental_price. The distribution of other fields against the rental price and their correlation is a significant focus of this analysis.

2. Methodology

This section goes into the preprocessing steps that were taken for analysis and the data analysis that were taken.

2.1 Checking for Duplicates and Handling them

The dataset was checked for duplicate entries, and any duplicates found were removed to ensure data integrity.



```
[ ]  df.duplicated().sum()
```

```
2
```

Remove Duplicates Values

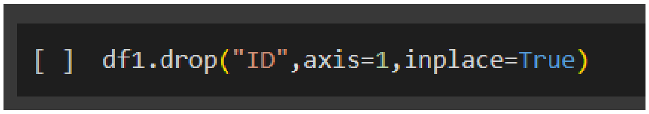
```
[ ]  df1 = df.drop_duplicates()
     df1.duplicated().sum()
```

```
0
```

2 duplicate entries were found and removed from the dataframe and our resultant dataset used for analysis from here on now has 250 entries.

2.2 Dropping the unwanted ID column

We also dropped the ID column from the dataset as we considered it as not required for the analysis.



```
[ ]  df1.drop("ID",axis=1,inplace=True)
```

2.3 Summary Statistics

2.3.1 Summary statistics for numerical columns

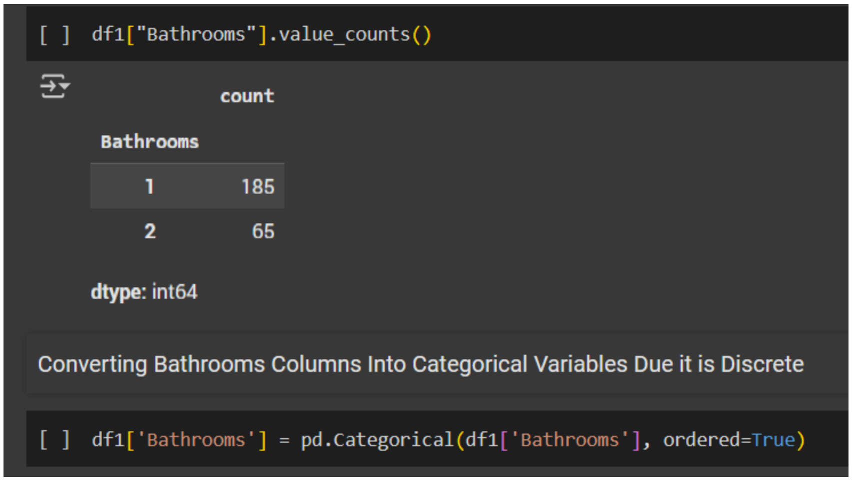
Summary statistics for the dataset, including measures such as mean, median, standard deviation, and quartiles, were computed for all numerical and features to understand their distributions and detect any anomalies.

Feature	Mean	Std Dev	Min	25%	50%	75%	Max
Size_in_Sqft	800.97	302.51	301.00	559.00	802.00	1023.50	1497.00
Bathrooms	1.26	0.44	1.00	1.00	1.00	2.00	2.00
Distance_to_City_Center	6.55	3.40	0.57	3.49	7.00	9.58	11.98
Rental_Price	4079.35	2052.21	804.81	2426.14	3938.17	5254.28	13179.28

2.3.2 Ordinal Transformation of the Bathrooms feature.

The bathrooms column was initially treated as a numerical variable. However, since it only had two values (1 or 2 bathrooms), it was better suited as an ordinal feature. The transformation was carried out accordingly.

If we had kept the bathroom feature as numeric, it would have resulted in incorrect results in our EDA steps when seeing the outputs for things like mean number of bathrooms etc as it results in decimal unrealistic values. The bedrooms column is already formatted as of datatype object so we can change the datatype of the bathrooms to the same.



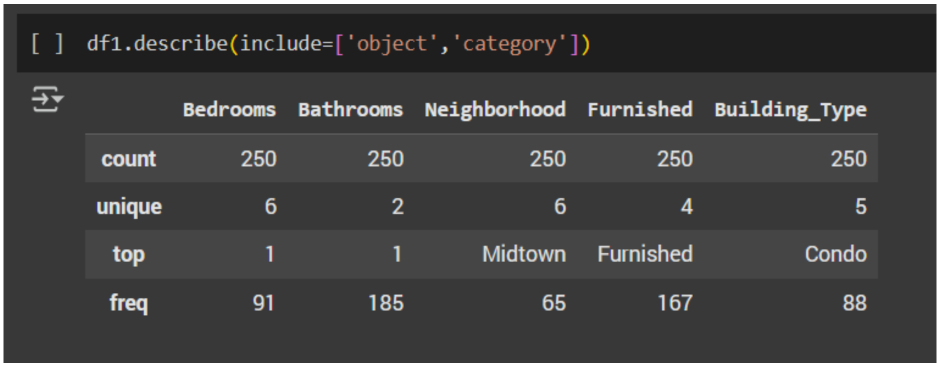
```
[ ] df1["Bathrooms"].value_counts()

            count
 Bathrooms
     1        185
     2         65

dtype: int64
```

Converting Bathrooms Columns Into Categorical Variables Due it is Discrete

```
[ ] df1['Bathrooms'] = pd.Categorical(df1['Bathrooms'], ordered=True)
```

2.3.3 Summary Statistics for Categorical Features

Summary statistics for categorical (string) features were computed, including unique value counts and frequency distributions, to assess the characteristics of these variables.



```
[ ] df1.describe(include=['object','category'])
```

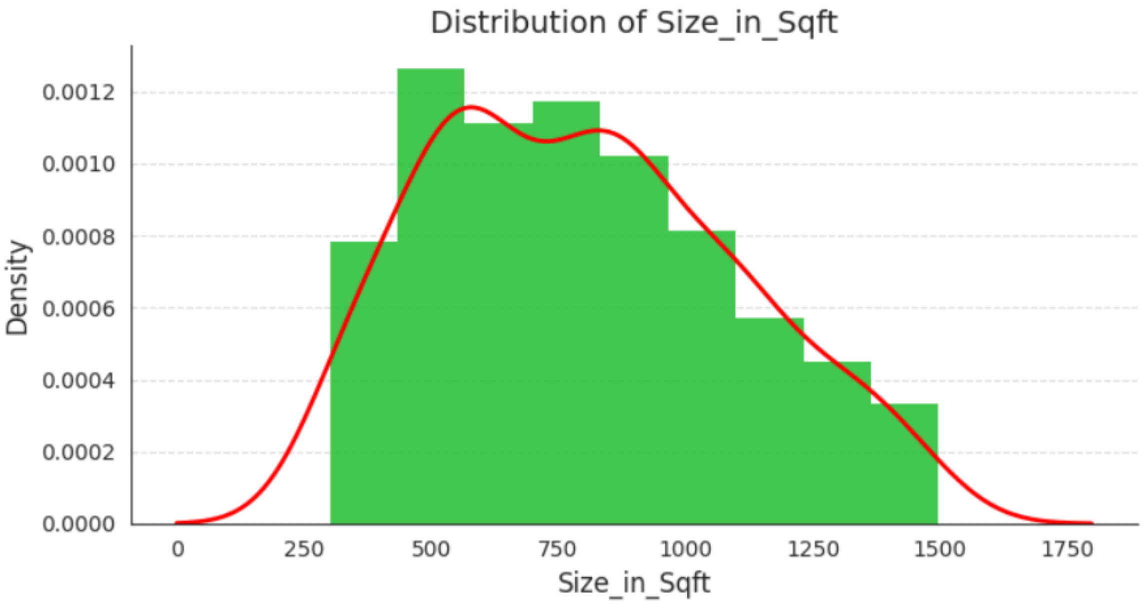
```
        Bedrooms  Bathrooms  Neighborhood  Furnished  Building_Type
count        250        250           250        250            250
unique         6          2             6          4              5
top            1          1       Midtown  Furnished          Condo
freq          91        185            65        167             88
```

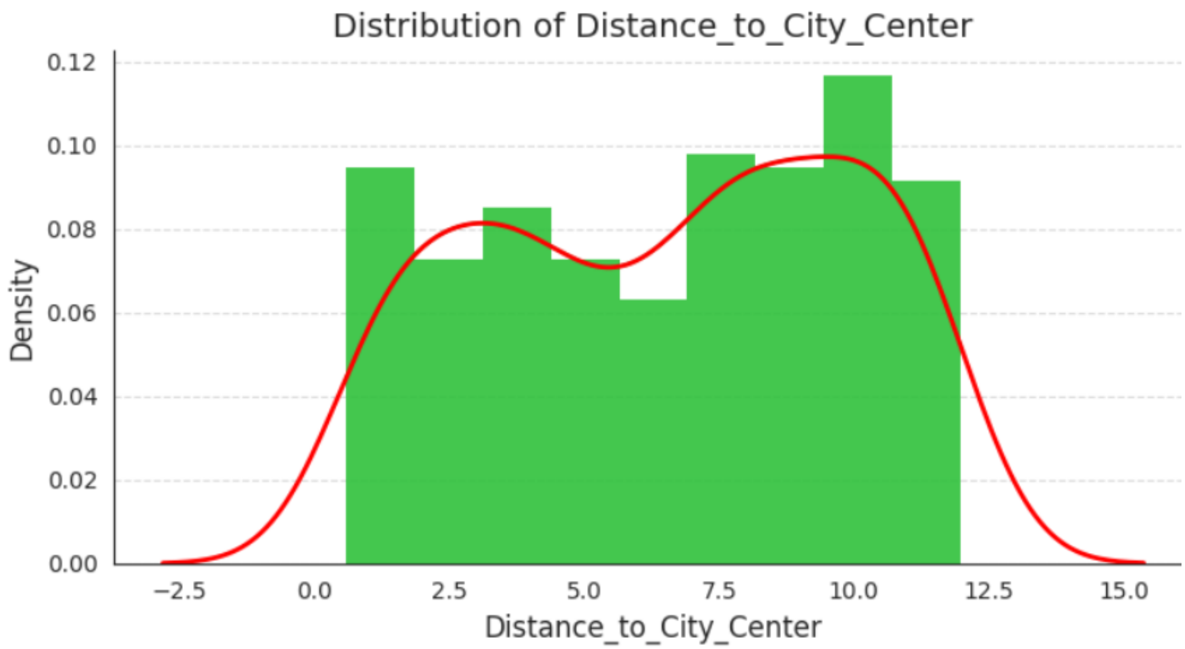
The number of unique values in each categorical column and the modes in each category were found this way.

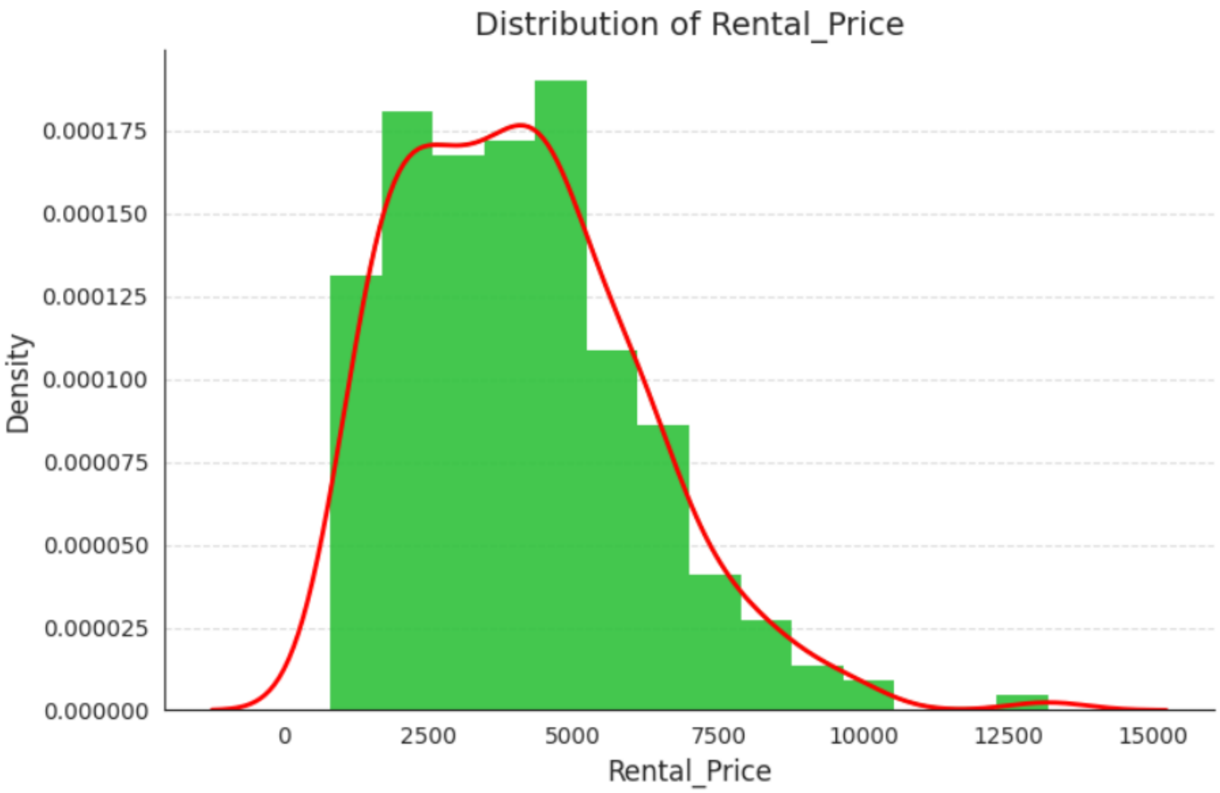
2.4 Numerical feature analysis

2.4.1 Plotting numerical column distributions

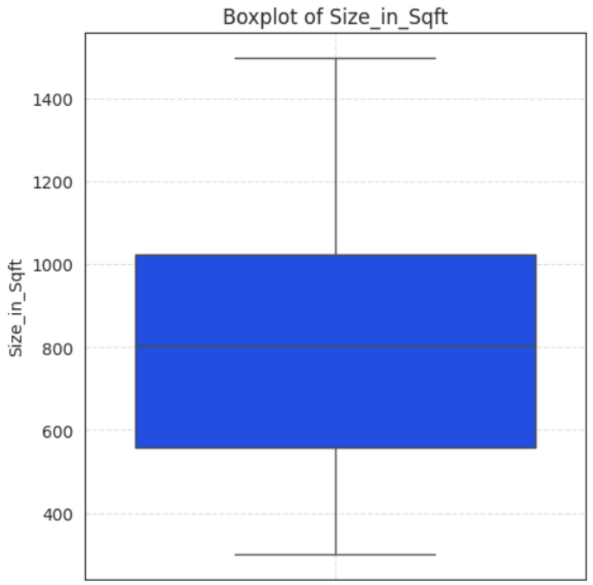
The distribution trend of each of the numerical columns were plotted using histograms alongside a KDE plot to see the general data distribution in each column when taken separately.

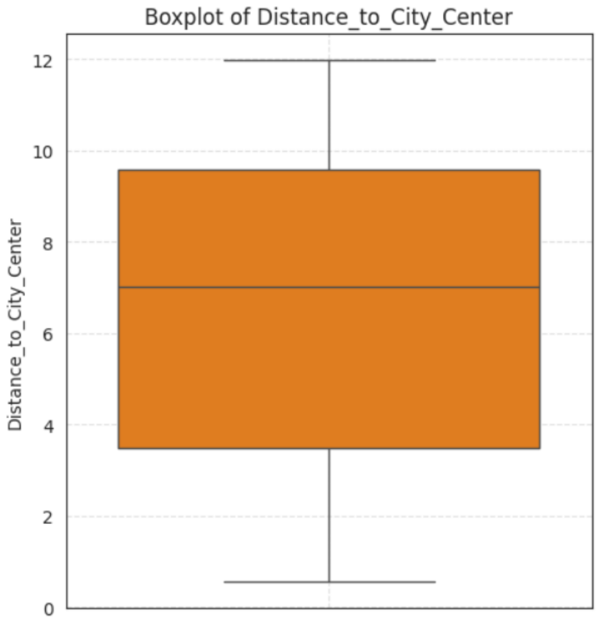


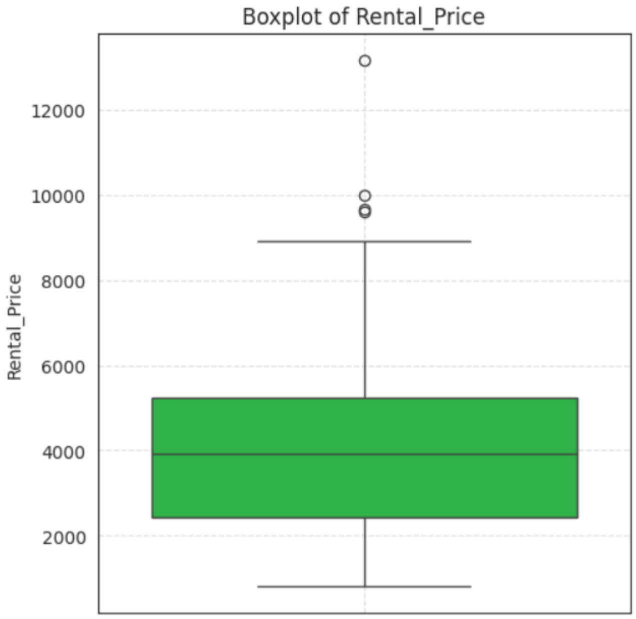

Distribution of Distance_to_City_Center
Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
Distance_to_City_Center


Distribution of Rental_Price
Density
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
0
2500
5000
7500
10000
12500
15000
Rental_Price

2.4.2 Plotting Boxplots to further understand the distribution and outlier detection







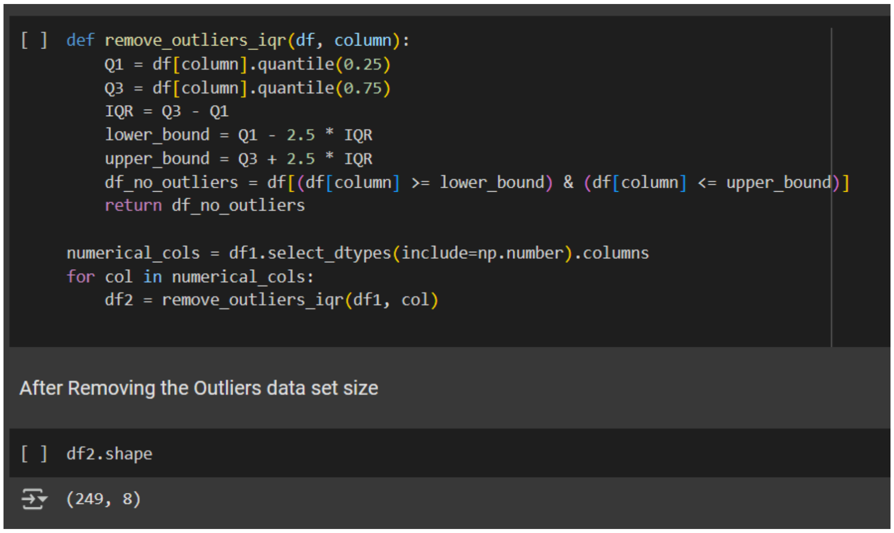
We can thus identify 3 extreme values in the rental price column using the boxplot analysis.

2.4.3 Removing Outliers

The Interquartile Range (IQR) method was used to detect and remove outliers.

When looking at the plots of features such as rental price, the maximum value of the column 13,173 while the median value is 3938. Large values such as this skews our analysis if we look at features as the mean.

Also we found a couple of values that were significantly higher but we've found these values to be possibly important for the analysis as some high values are possible in a market such as real estate. Thus we only wanted to capture significantly larger values so we opted to use 2.5 * IQR to remove the outliers instead of the usual 1.5 * IQR used.



```
[ ] def remove_outliers_iqr(df, column):
        Q1 = df[column].quantile(0.25)
        Q3 = df[column].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 2.5 * IQR
        upper_bound = Q3 + 2.5 * IQR
        df_no_outliers = df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]
        return df_no_outliers

    numerical_cols = df1.select_dtypes(include=np.number).columns
    for col in numerical_cols:
        df2 = remove_outliers_iqr(df1, col)
```

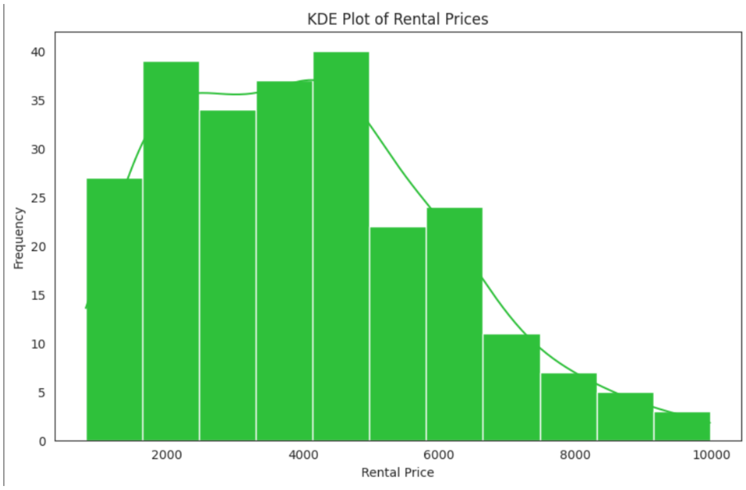
After Removing the Outliers data set size

```
[ ] df2.shape
```

```
(249, 8)
```

After removing the outliers, our resultant dataframe now has 249 entries when we remove that single entry with that large value in the Rental Price Column.

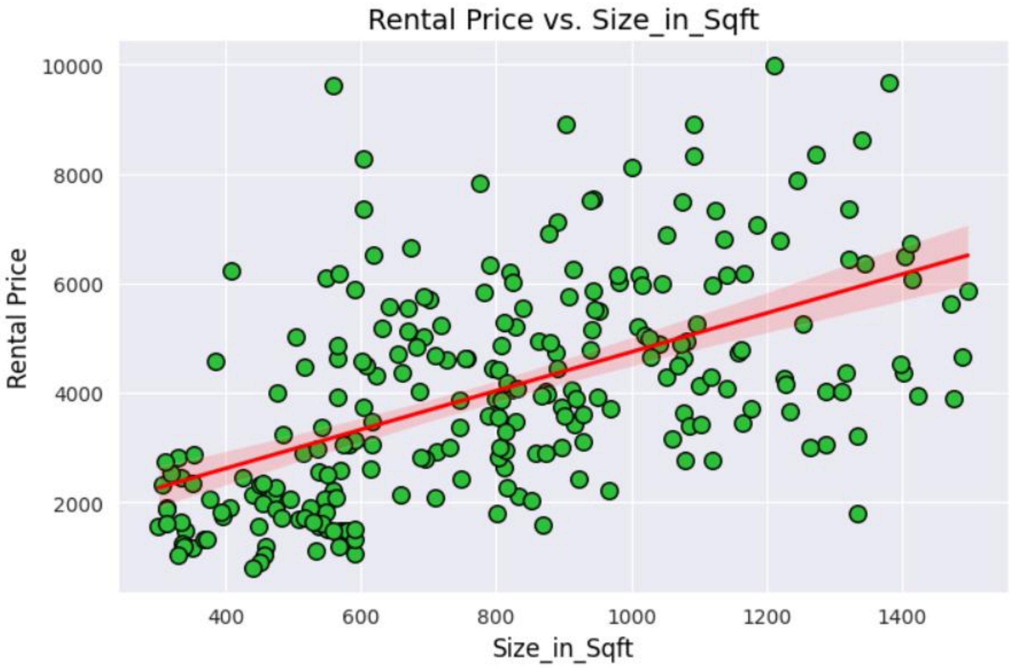
The histogram distribution of the rental value feature now takes a following shape.



2.4.4 Bivariate analysis for Numerical Columns

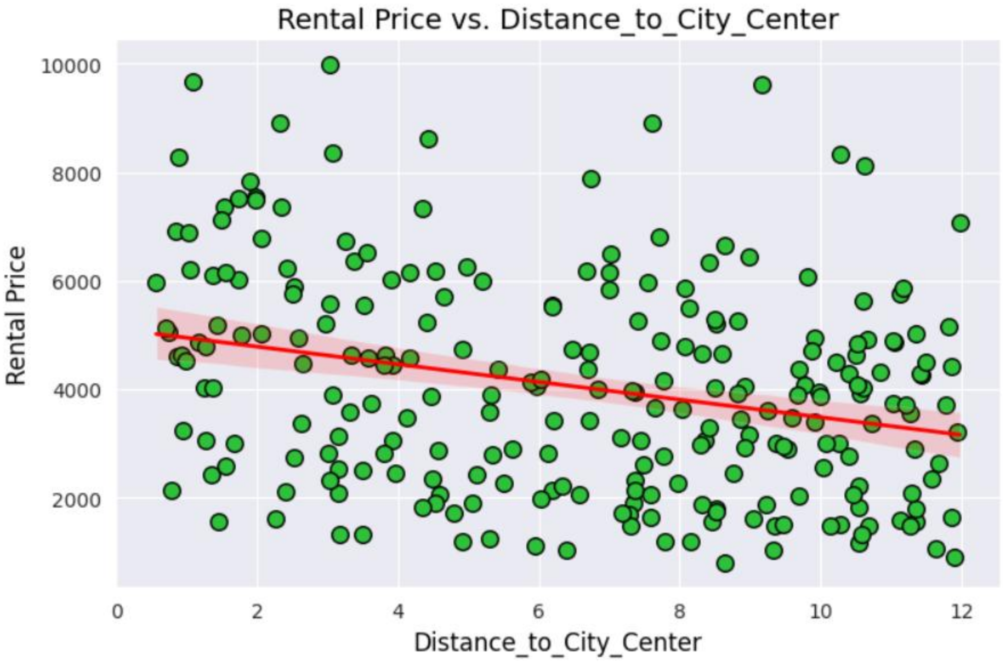
As our goal is to understand the variation of rental pricing based on the features on the dataset, the next step was to plot each of the numerical features against the rental property to understand a relationship or a general trend.

Rental Price vs Size in Square Feet



There is a positive relationship between the rental price and the size in square feet. This trend is understandable as the rental price increases when the size of the property increases.

Rental price and the distance from Petta to the property

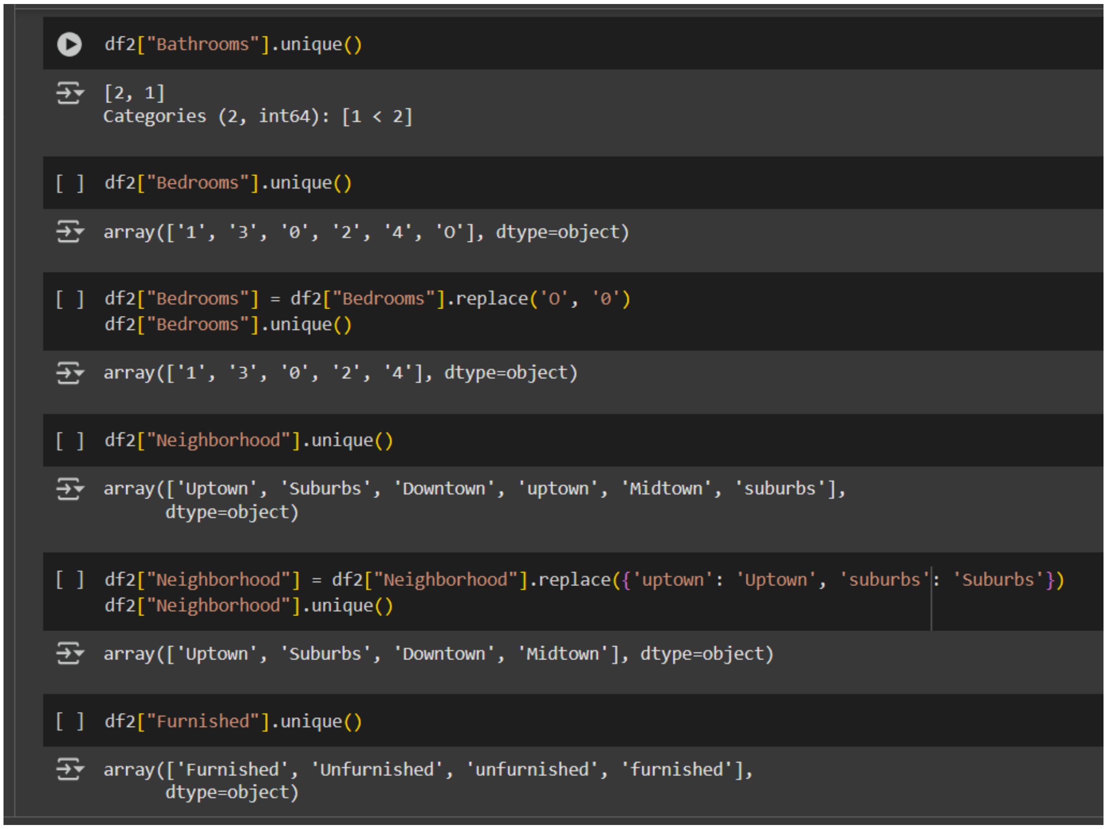


Bivariate analysis between rental price and the distance from Colombo Fort to the property doesn't result in a strong correlation and there is a small negative correlation between distance and rental price. Though now a distinct increase, this conveys that more distant the property is from the fort (Colombo), there is a smaller trend of a decrease in rental price.

2.5 Categorical Feature Analysis

2.5.1 Data Cleaning in Categorical features

After checking the unique value counts in each of the categorical columns, some values are differently formatted.



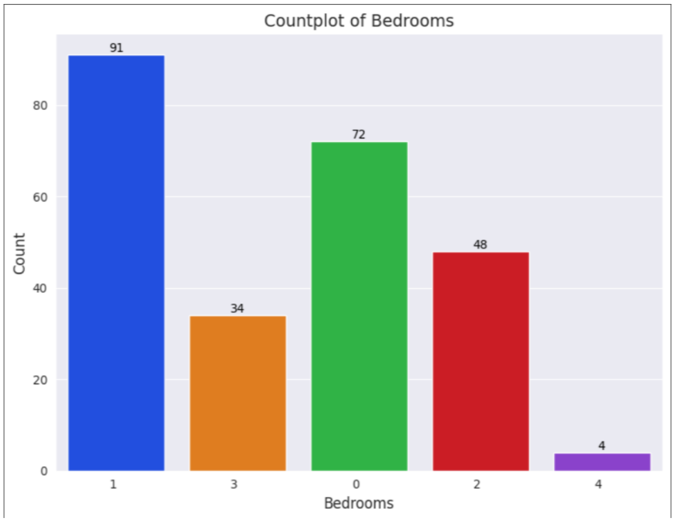
```
df2["Bathrooms"].unique()

[2, 1]
Categories (2, int64): [1 < 2]

[ ] df2["Bedrooms"].unique()

array(['1', '3', '0', '2', '4', 'O'], dtype=object)

[ ] df2["Bedrooms"] = df2["Bedrooms"].replace('O', '0')
    df2["Bedrooms"].unique()

array(['1', '3', '0', '2', '4'], dtype=object)

[ ] df2["Neighborhood"].unique()

array(['Uptown', 'Suburbs', 'Downtown', 'uptown', 'Midtown', 'suburbs'],
      dtype=object)

[ ] df2["Neighborhood"] = df2["Neighborhood"].replace({'uptown': 'Uptown', 'suburbs': 'Suburbs'})
    df2["Neighborhood"].unique()

array(['Uptown', 'Suburbs', 'Downtown', 'Midtown'], dtype=object)

[ ] df2["Furnished"].unique()

array(['Furnished', 'Unfurnished', 'unfurnished', 'furnished'],
      dtype=object)
```

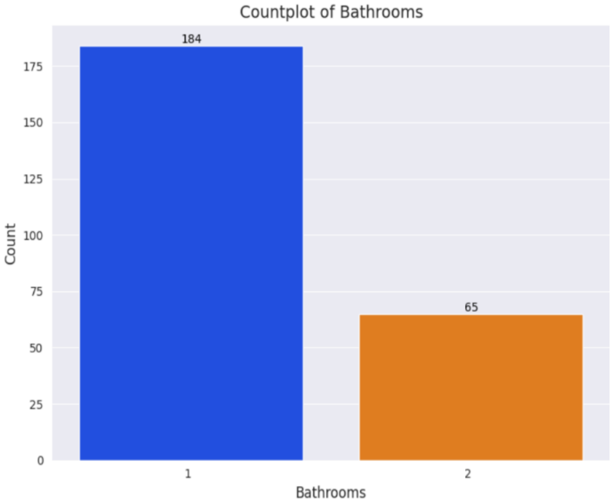
As some feature values were named inconsistently with different latter casing, we were to replace all the feature values with a consistent naming. Thus these features are consistently named to be calculated together instead of being treated as separate.

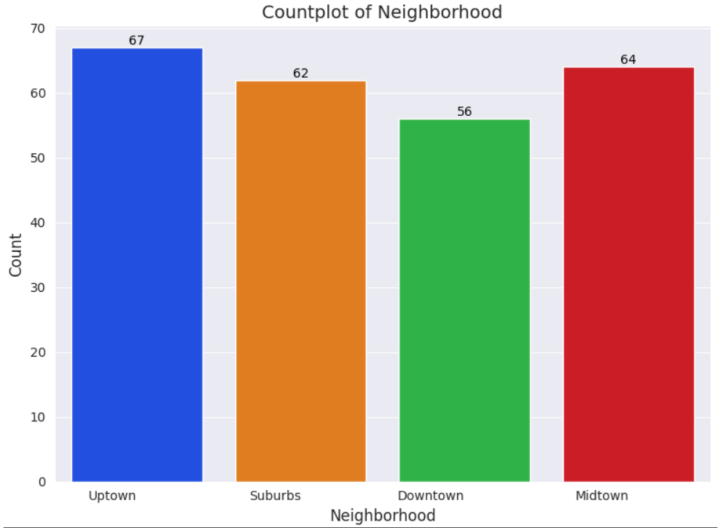
The bedrooms column has a value named ‘O’, we replaced the value of ‘O’ with 0.

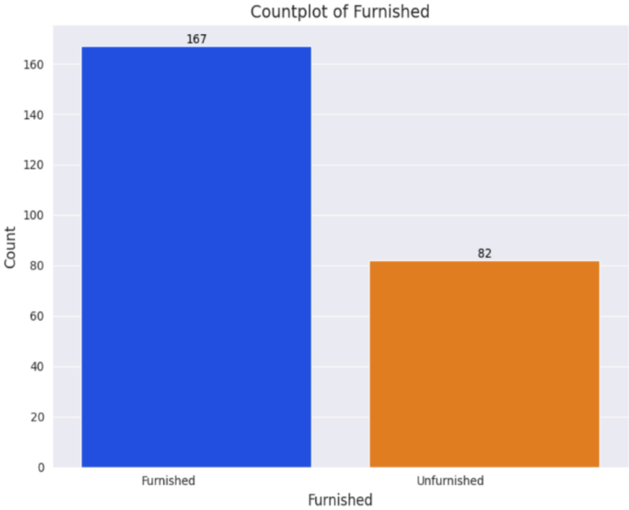
2.5.2 Countplots for each of the categorical features to see value counts for each feature

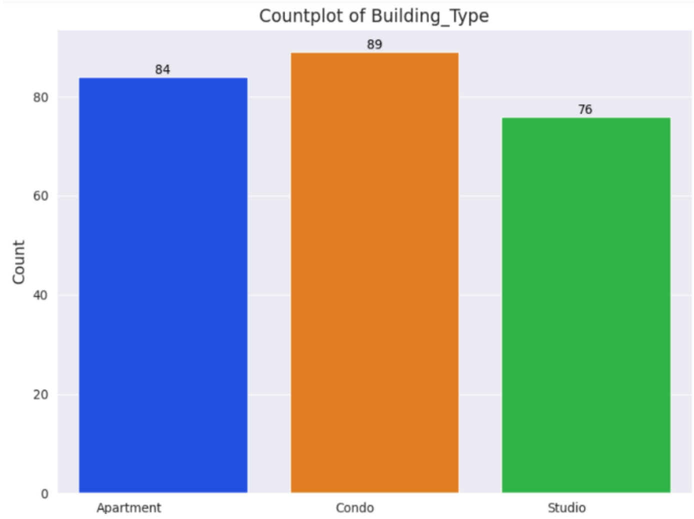
For each categorical feature, we created countplots to visualize the frequency distribution of its values. Countplots are particularly useful for understanding the balance of categories within a feature and providing a clear representation of the number of occurrences for each category.



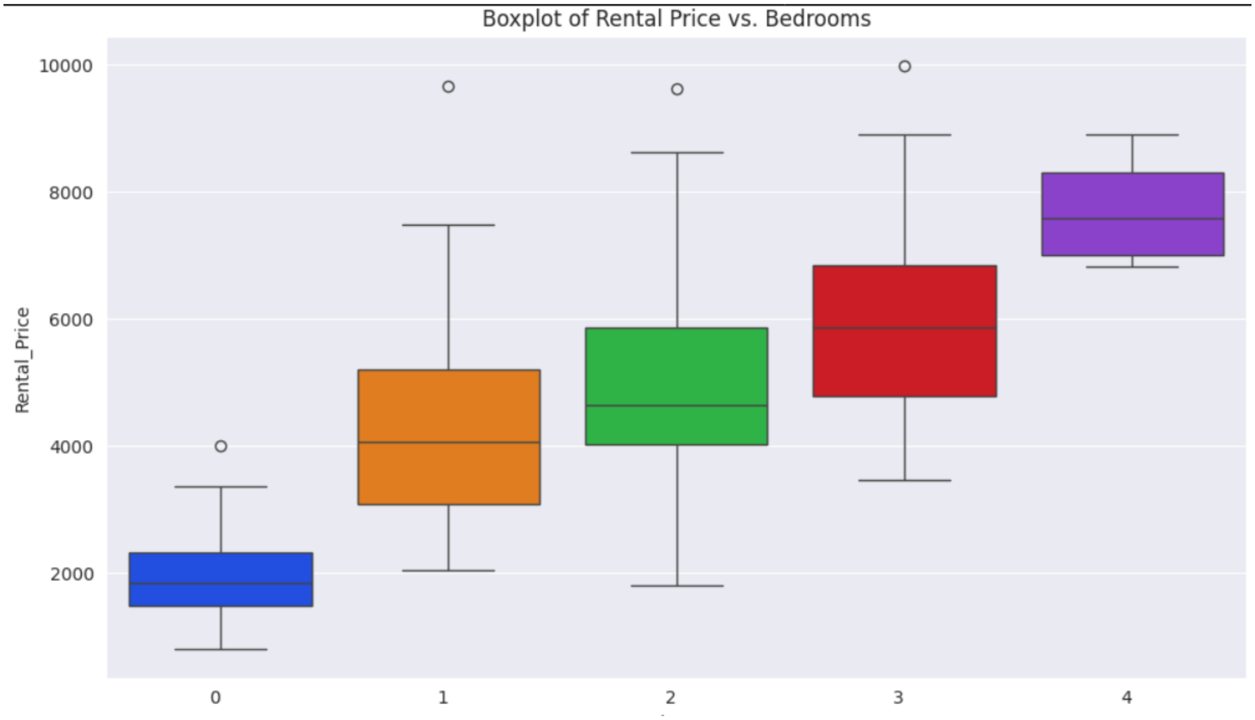


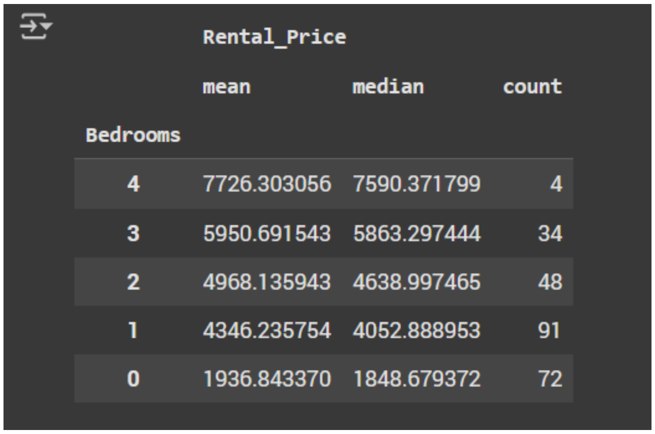






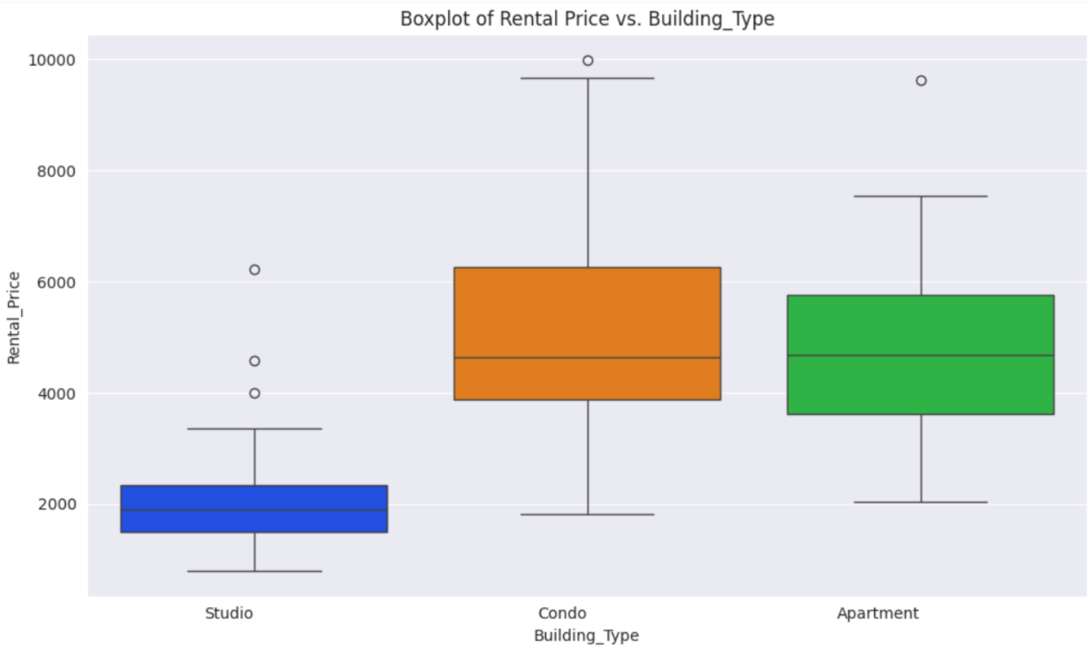
2.5.3 Getting the boxplot distributions in rental values against each of the categorical columns and additional groupings

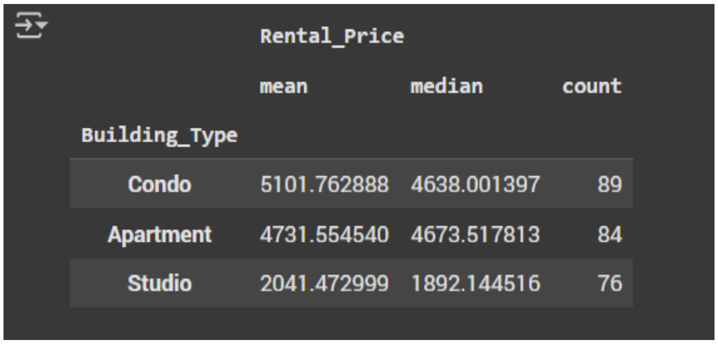




	Rental_Price		
	mean	median	count
Bedrooms			
4	7726.303056	7590.371799	4
3	5950.691543	5863.297444	34
2	4968.135943	4638.997465	48
1	4346.235754	4052.888953	91
0	1936.843370	1848.679372	72

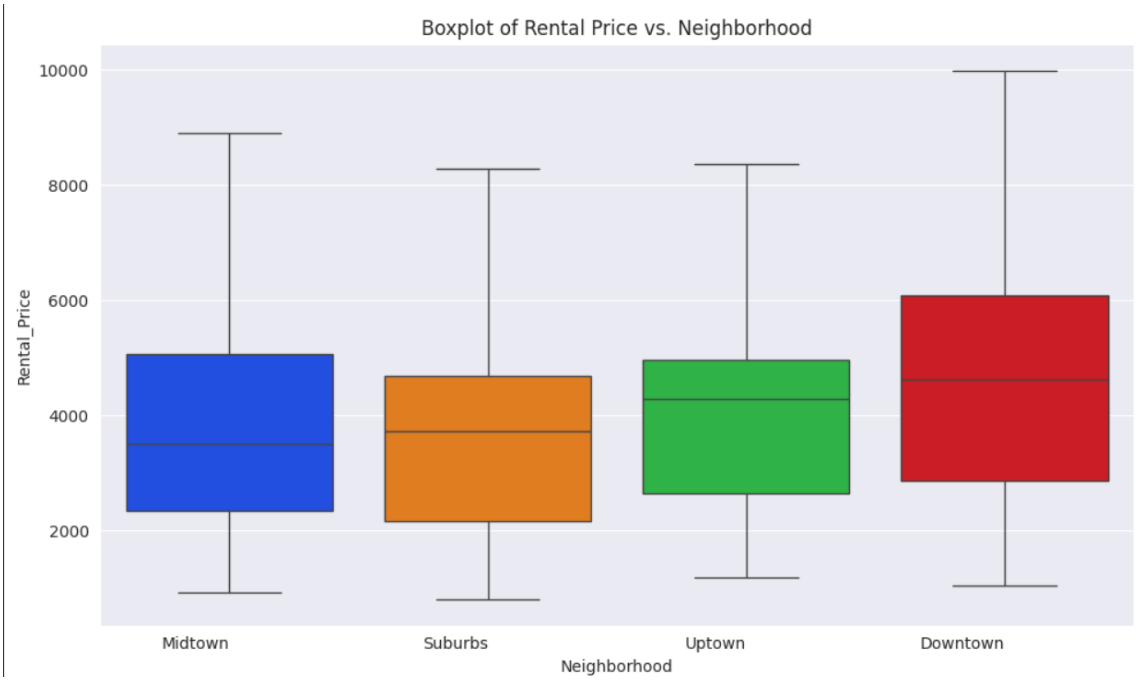
We can thus see that more bedrooms mean a higher rental price.

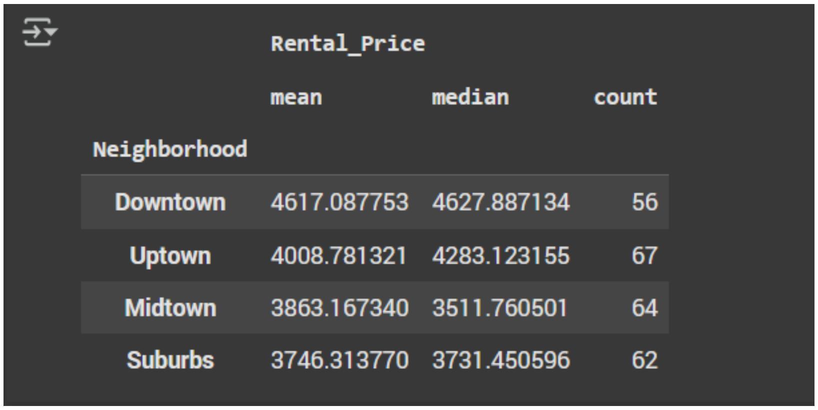




	Rental_Price		
	mean	median	count
Building_Type			
Condo	5101.762888	4638.001397	89
Apartment	4731.554540	4673.517813	84
Studio	2041.472999	1892.144516	76

Condo prices have been distributed at a larger variance than apartments and studio prices.



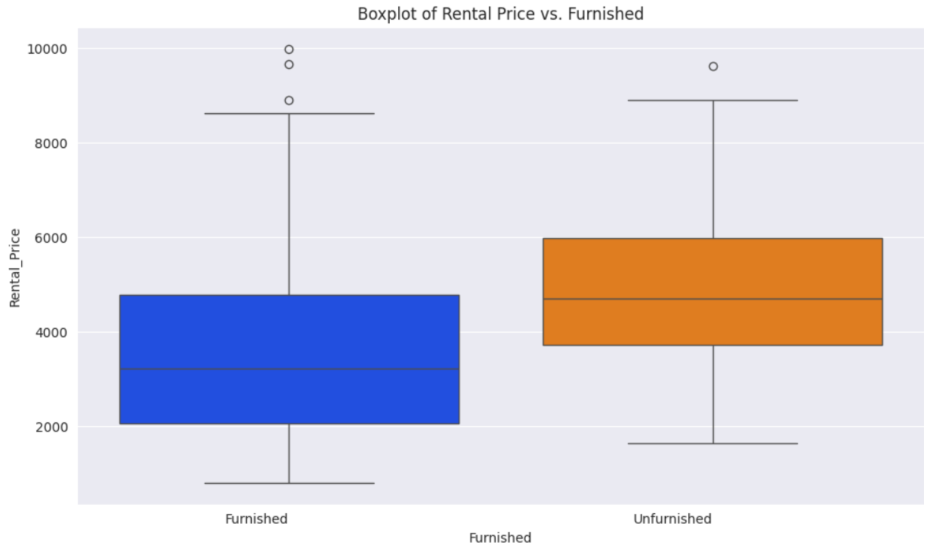


	Rental_Price		
	mean	median	count
Neighborhood			
Downtown	4617.087753	4627.887134	56
Uptown	4008.781321	4283.123155	67
Midtown	3863.167340	3511.760501	64
Suburbs	3746.313770	3731.450596	62

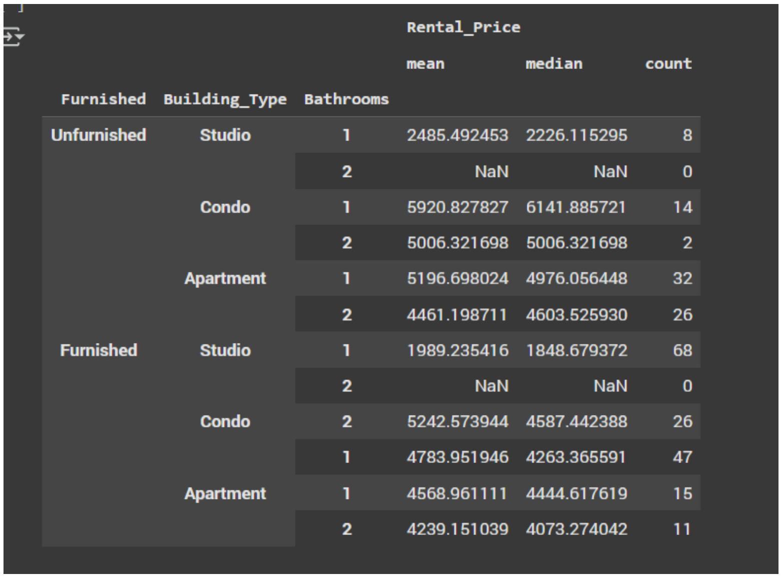
The rental price data by neighborhood shows Downtown with the highest average rent at 4617.09, indicating a relatively balanced market with fewer properties (56). Uptown follows with an average of 4008.78, featuring a larger market size (67) and some higher-priced outliers. Midtown's average is 3863.17, reflecting a market with more variation and some lower-priced properties.

The Suburbs have the lowest average rental price at 3746.31, with a more consistent and uniform pricing structure across 62 properties. Overall, Downtown has the highest rents, while the Suburbs have the lowest, with Uptown and Midtown in between.

Furnished vs Rental Price



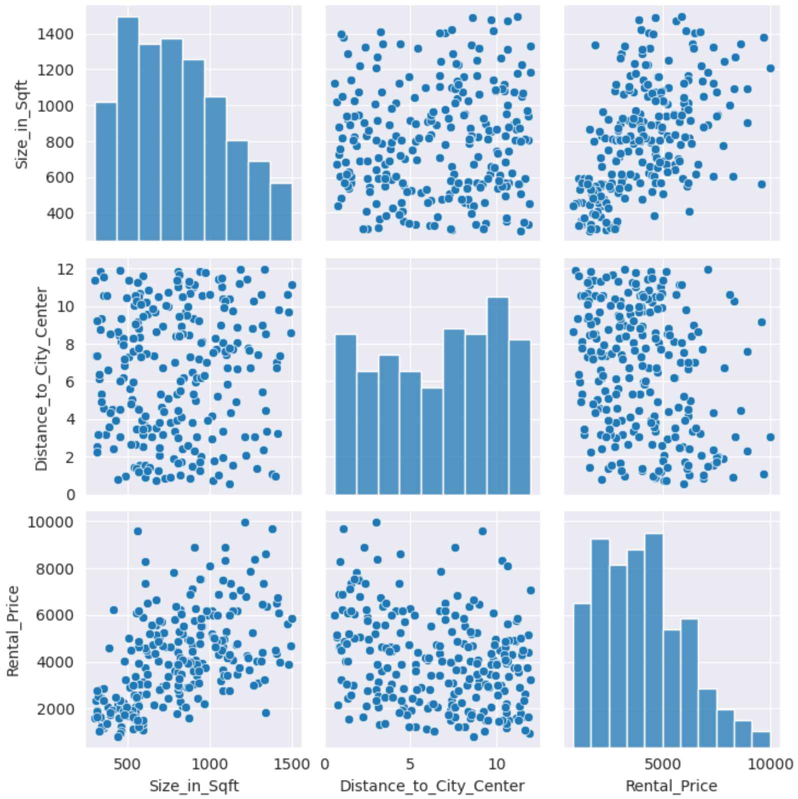
We are seeing a rather unexpected pattern here. Our initial expectations were to see higher prices for furnished apartments. As we found this difference an interesting finding, further grouping of data revealed that this was a consistent trend across Building categories



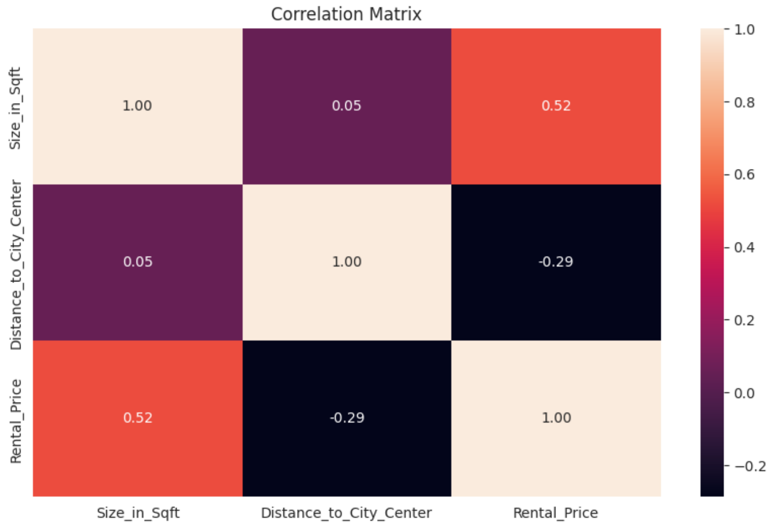
			Rental_Price		
			mean	median	count
Furnished	Building_Type	Bathrooms			
Unfurnished	Studio	1	2485.492453	2226.115295	8
		2	NaN	NaN	0
	Condo	1	5920.827827	6141.885721	14
		2	5006.321698	5006.321698	2
	Apartment	1	5196.698024	4976.056448	32
		2	4461.198711	4603.525930	26
Furnished	Studio	1	1989.235416	1848.679372	68
		2	NaN	NaN	0
	Condo	2	5242.573944	4587.442388	26
		1	4783.951946	4263.365591	47
	Apartment	1	4568.961111	4444.617619	15
		2	4239.151039	4073.274042	11

2.6 Correlation analysis of the numerical features

Pair plots for plotting distribution



Correlation matrix



The correlation matrix offers a deeper understanding of the relationships between key numerical features. The correlation between *Size_in_Sqft* and *Rental_Price* is notably strong, with a coefficient of 0.519, indicating a moderate positive relationship. This suggests that larger properties tend to have higher rental prices, which aligns with expectations in housing markets. On the other hand, the correlation between *Size_in_Sqft* and *Distance_to_City_Center* is very low at 0.0507, suggesting that the size of a property is largely independent of its distance from the city center.

The correlation between *Rental_Price* and *Distance_to_City_Center* is negative at -0.2867, indicating a weak but inverse relationship. This suggests that properties farther from the city center tend to have lower rental prices, which is a common trend in urban housing markets where proximity to the city center generally drives up rental prices.

3. Discussion

The analysis of rental prices in Sri Lanka has revealed several important relationships between various property characteristics and rental costs. The findings suggest some expected trends, while also uncovering some unexpected results that warrant further investigation.

3.1. Size (in Square Feet) and Rental Price

One of the most significant findings was the positive correlation between the size of the property and rental price. With a correlation coefficient of 0.52, this suggests that larger properties tend to have higher rental prices, which is in line with common expectations in real estate markets worldwide. Larger properties offer more space, which typically drives up rental demand and, consequently, the rental price.

3.2 Distance to City Center and Rental Price

Another expected finding was the negative relationship between distance from the city center and rental price, with a correlation of -0.29. Properties located farther away from the city center generally tend to have lower rental prices. This is a common trend observed in urban markets, where proximity to the city center usually correlates with higher rental demand due to accessibility to amenities, work opportunities, and public transport.

3.3. Neighborhood and Rental Prices

The analysis of rental prices by the neighborhood revealed interesting trends:

- Downtown emerges as the area with the highest average rental prices, which is understandable given its central location, proximity to business hubs, and access to various amenities and services.
- On the other hand, Suburbs represents the neighborhood with the lowest average rental prices, likely due to being farther from the city center and the limited availability of high-demand features or facilities.
- Midtown and Suburbs display similar average rental prices, suggesting that while Midtown may have some central benefits, Suburbs provide enough value in terms of space and affordability to match Midtown's prices.

3.4. Building Type and Rental Prices

The analysis of rental prices by building type further highlights the influence of property characteristics on rental costs. Condos were found to command the highest rental prices. Condominiums are often associated with modern amenities, prime locations, and desirable facilities, contributing to their higher rental costs. Conversely, *Studios* had the lowest rental prices, which is expected as these smaller, more compact spaces typically appeal to single individuals or students who prioritize cost over space.

3.5. Bedrooms, Bathrooms, and Rental Price

Both the number of bedroom and bathrooms showed positive relationships with rental price, which is intuitive. More bedrooms and bathrooms typically translate to larger living spaces and greater comfort, thereby increasing the rental price.

This relationship also suggests that renters are willing to pay a premium for properties that offer more space and amenities for family or group living.

3.6. Unexpected Findings with Furniture and Rental Price

A surprising result emerged when analyzing the relationship between furniture and rental prices. Contrary to expectations, furnished properties had a lower average rental price compared to non-furnished properties. This finding was unexpected as one would assume that furnished properties, offering convenience for tenants, would typically command higher prices.

Upon deeper exploration during our exploratory data analysis (EDA), we hypothesize that this could be due to a labeling mistake, where the classification of some properties as "furnished" might not have been accurate. This anomaly requires further verification and could be an area for additional data cleaning.

4. Conclusion

The analysis highlights some clear patterns in the rental market of Sri Lanka, particularly the positive correlation between size, bedrooms, and bathrooms with rental price, as well as the expected price differences between neighborhoods and building types.

We have also found expected trends of decrease in rental price with distance from the fort though as not as significant.

However, the unexpected result regarding furnished versus non-furnished properties emphasizes the importance of data accuracy and validation.