Data Xplore 2025

# EDA & INSIGHT ANALYSIS REPORT

Prepared by:

**Team DXP10**

# Introduction

This report presents an in-depth analysis of the rental landscape in Colombo, using data-driven insights to uncover trends and pricing patterns across various neighborhoods.


To achieve this, we conducted an **Exploratory Data Analysis (EDA)** using **Python** in a **Kaggle Jupyter Notebook**. Leveraging powerful libraries such as **Pandas, Matplotlib, and Seaborn**, we cleaned, processed, and visualized the dataset to extract meaningful insights.

The dataset includes apartment listings with the following key features:

- **Apartment_ID**: Unique identifier for each apartment listing.
- **Neighborhood**: Name of the neighborhood where the apartment is located.
- **Rental_Price**: Monthly rental price of the apartment.
- **Size_in_Sqft**: Total size of the apartment in square feet.
- **Distance_to_City_Center:** Distance from the apartment to Colombo Fort Station.
- **Bedrooms**: Number of bedrooms in the apartment.
- **Bathrooms**: Number of bathrooms in the apartment.
- **Furnished**: Indicates whether the apartment is furnished or not.
- **Building_Type**: Type of building the apartment is located in.

By analyzing these variables, we aim to identify key drivers of rental prices and their impact on different types of apartments. Our findings provide valuable insights into how factors like location, apartment size, and amenities influence rental costs, enabling stakeholders to make informed decisions in the dynamic Colombo real estate market.

# Methodology

To extract meaningful insights from the dataset, we implemented a structured approach consisting of data preprocessing and data analysis. These steps ensured data accuracy, consistency, and relevance for uncovering key rental market trends in Colombo.

**Data Cleaning and Preprocessing**

**Statistical and Correlation Analysis**

**Data Visualization**

# Data Preprocessing

The initial data types and memory usage of each column were as follows:

```
<class 'pandas.core.frame.DataFrame'>
Index: 252 entries, 1 to 250
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Size_in_Sqft            252 non-null    int64
 1   Bedrooms                252 non-null    object
 2   Bathrooms               252 non-null    int64
 3   Distance_to_City_Center 252 non-null    float64
 4   Neighborhood            252 non-null    object
 5   Furnished               252 non-null    object
 6   Building_Type           252 non-null    object
 7   Rental_Price            252 non-null    float64
dtypes: float64(2), int64(2), object(4)
memory usage: 71.9 KB
```

The Dtype being 'int64' means that the data is in integer format and 'object' means that the data is either text or categorical.

**We did the following to clean our dataset:**

- Identified and removed duplicates.

```
dataset[dataset.duplicated()]
```

| | ID | Size_in_Sqft | Bedrooms | Bathrooms | Distance_to_City_Center | Neighborhood | Furnished | Building_Type | Rental_Price |
|---|---|---|---|---|---|---|---|---|---|
| 67 | 67 | 301 | 0 | 1 | 11.36 | Downtown | Furnished | Studio | 1569.109771 |
| 188 | 187 | 631 | 1 | 1 | 1.43 | Uptown | Unfurnished | Apartment | 5191.320454 |

```
filtered_rows = dataset[dataset['ID'] == 67]
print(filtered_rows)
```

```
    ID  Size_in_Sqft Bedrooms  Bathrooms  Distance_to_City_Center  \
66  67           301        0          1                    11.36
67  67           301        0          1                    11.36

   Neighborhood  Furnished Building_Type  Rental_Price
66     Downtown  Furnished        Studio   1569.109771
67     Downtown  Furnished        Studio   1569.109771
```

```
filtered_rows = dataset[dataset['ID'] == 187]
print(filtered_rows)
```

```
      ID  Size_in_Sqft Bedrooms  Bathrooms  Distance_to_City_Center  \
187  187           631        1          1                     1.43
188  187           631        1          1                     1.43

    Neighborhood    Furnished Building_Type  Rental_Price
187       Uptown  Unfurnished     Apartment   5191.320454
188       Uptown  Unfurnished     Apartment   5191.320454
```

- Checked for missing values. None were found.

```
ID                         0
Size_in_Sqft               0
Bedrooms                   0
Bathrooms                  0
Distance_to_City_Center    0
Neighborhood               0
Furnished                  0
Building_Type              0
Rental_Price               0
dtype: int64
```

- Identified and corrected the data inconsistencies.
  - In ‘Bedrooms’ column , there were two entries with ‘O’ , we changed it to ‘-1'

```
Bedrooms
1    92
0    71
2    48
3    34
4     5
O     2
Name: count, dtype: int64
```

- In “Furnished” , “Neighborhood” , “Building_type” columns, there were non unique features with different names.
  We changed them as follows:

**Before**

```
Furnished
Furnished      168
Unfurnished     79
unfurnished      4
furnished        1
Name: count, dtype: int64
```

```
Neighborhood
Midtown      65
Uptown       62
Suburbs      58
Downtown     56
uptown        5
suburbs       4
Name: count, dtype: int64
```

```
Building_Type
Condo         88
Apartment     82
Studio        76
apartment      2
condo          2
Name: count, dtype: int64
```

**After**

```
Furnished
Furnished      168
Unfurnished     82
Name: count, dtype: int64
```

```
Neighborhood
Uptown       67
Midtown      65
Suburbs      62
Downtown     56
Name: count, dtype: int64
```

```
Building_Type
Condo         90
Apartment     84
Studio        76
Name: count, dtype: int64
```

- One hot encoded features ‘Building_Type’ , ‘ Neighborhood’ , ‘Furnished’ , and renamed the columns as required.

- Through this, we further managed to reduce the memory usage of our dataset which in turn will reduce the computational power
  for data processing and visualization.

**Before:** memory_usage = 71.9 KB

```
<class 'pandas.core.frame.DataFrame'>
Index: 252 entries, 1 to 250
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Size_in_Sqft            252 non-null    int64
 1   Bedrooms                252 non-null    object
 2   Bathrooms               252 non-null    int64
 3   Distance_to_City_Center 252 non-null    float64
 4   Neighborhood            252 non-null    object
 5   Furnished               252 non-null    object
 6   Building_Type           252 non-null    object
 7   Rental_Price            252 non-null    float64
dtypes: float64(2), int64(2), object(4)
memory usage: 71.9 KB
```

**After:** memory_usage = 27.3 KB

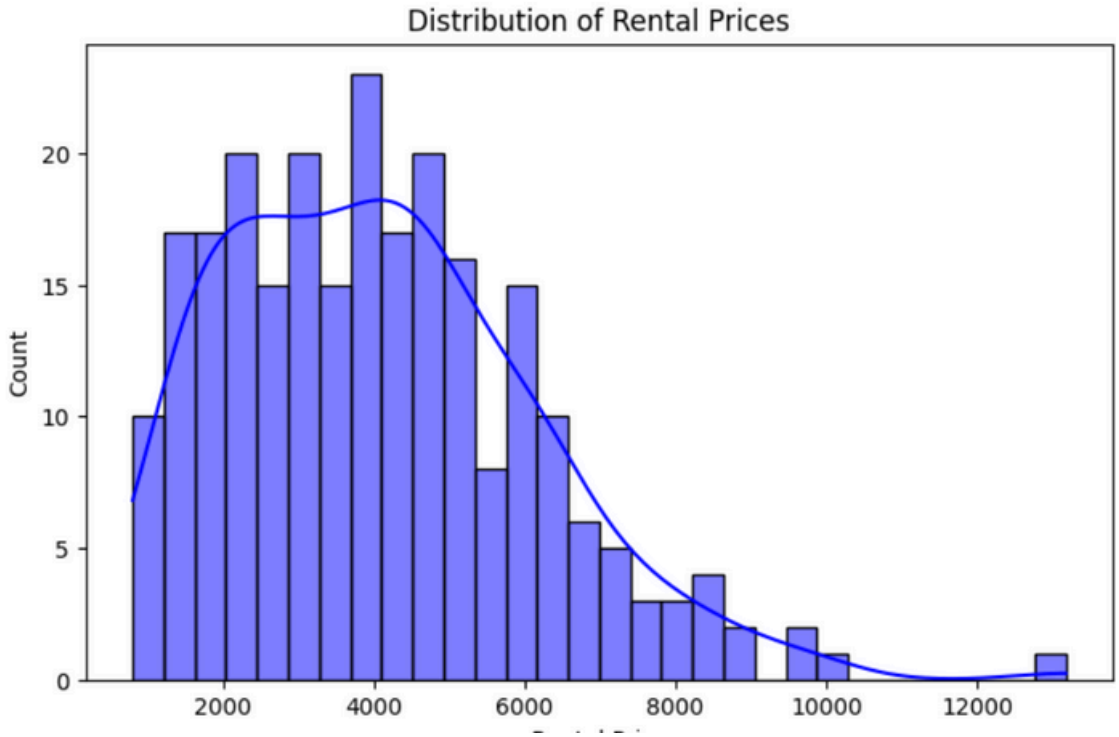
```
<class 'pandas.core.frame.DataFrame'>
Index: 250 entries, 1 to 250
Data columns (total 13 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Size_in_Sqft            250 non-null    int64
 1   Bedrooms                250 non-null    int64
 2   Bathrooms               250 non-null    int64
 3   Distance_to_City_Center 250 non-null    float64
 4   Rental_Price            250 non-null    float64
 5   Neighborhood_Downtown   250 non-null    float64
 6   Neighborhood_Midtown    250 non-null    float64
 7   Neighborhood_Suburbs    250 non-null    float64
 8   Neighborhood_Uptown     250 non-null    float64
 9   Furnished               250 non-null    float64
 10  Building_Type_Apartment 250 non-null    float64
 11  Building_Type_Condo     250 non-null    float64
 12  Building_Type_Studio    250 non-null    float64
dtypes: float64(10), int64(3)
memory usage: 27.3 KB
```

# Data Visualization

- We used matplotlib and seaborn python libraries for visualising.

```
import pandas as pd
pd.plotting.register_matplotlib_converters()
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

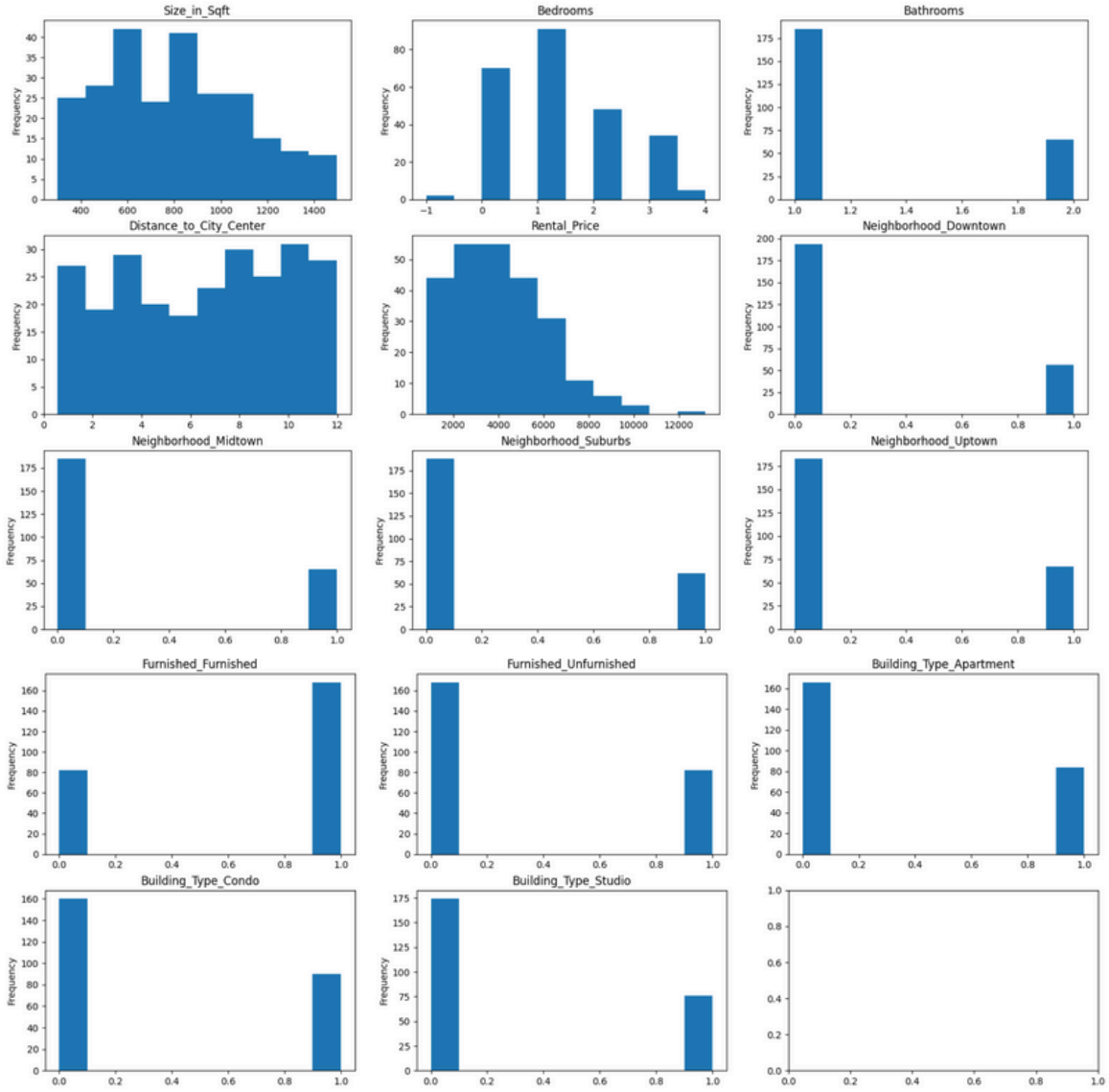
- Below is the distribution of Rental Prices with frequency, as seen it is moderately positively skewed.



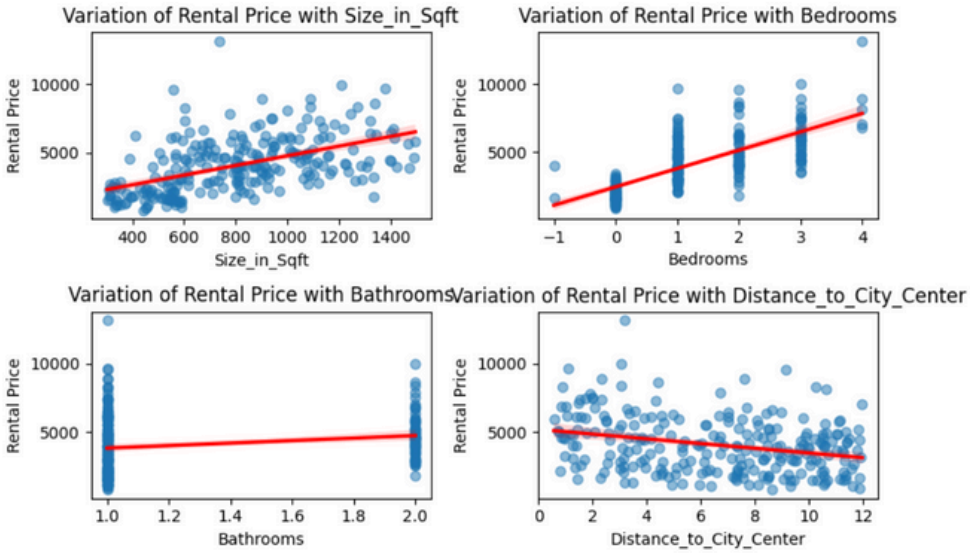

- We used matplotlib and seaborn python libraries for visualising.

```
import pandas as pd
pd.plotting.register_matplotlib_converters()
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

- Following is the variation of each feature with it's respective frequency.

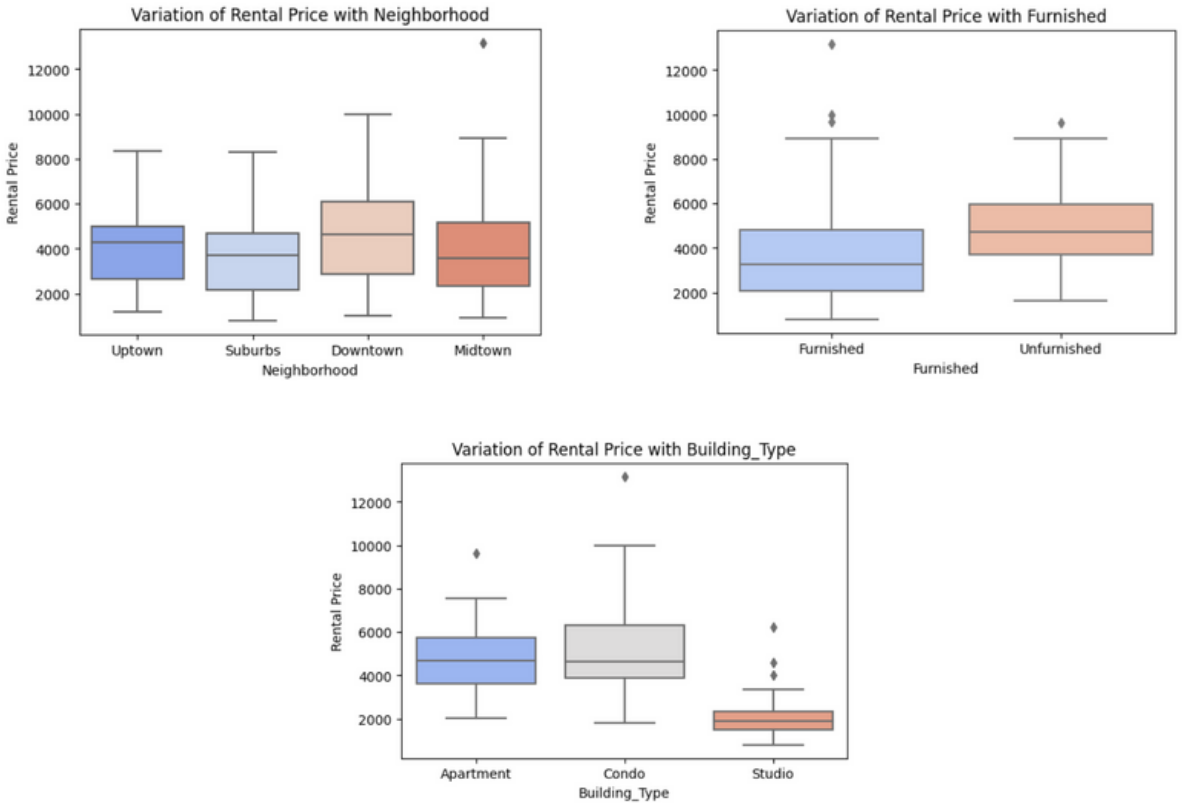



- Following is the variation of numerical features vs the Rental_Price



- With the number of bathrooms, and bedrooms the rental price has increased linearly , and the effect is more significant with bedroom count.
- According to these scatterplots, there is a linear relationship between Rental Price and Size_in_Sqft as well as Rental Price and Bedrooms.
- There is only a slight increase in the rental price with the number of bathrooms and a slight decrease in the rental price with the Distance_to_City_Center.

- Following is the variation of categorical features vs the Rental_Price

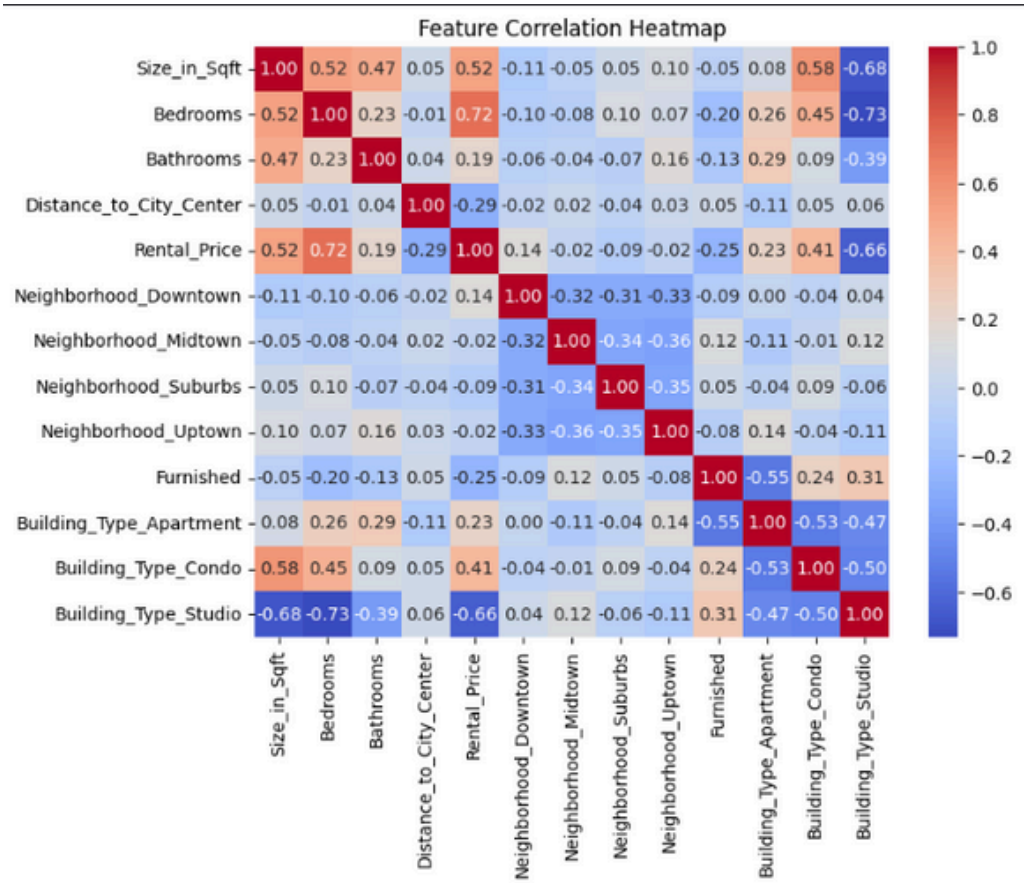



- According to the boxplot, the Neighborhood does not appear to have a significant impact on Rental_Price.
- However, there is a slight difference in rental prices between furnished and unfurnished properties.
- For Building Type, rental prices for Apartments and Condos are quite similar, while Studios show a noticeable difference compared to the other two categories.

# Data Analysis

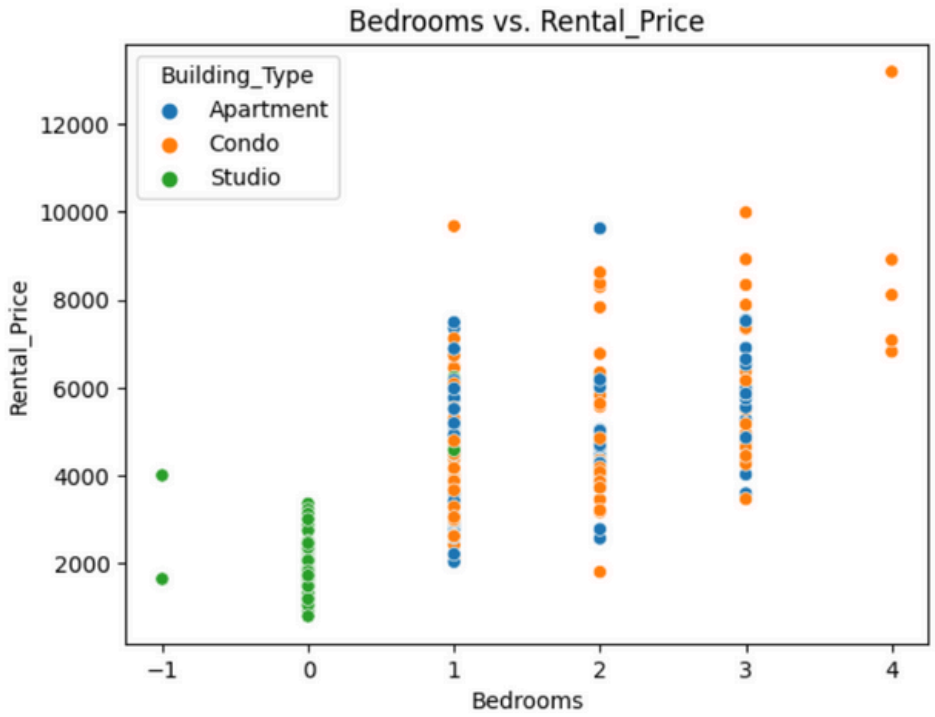
- Below are the mutual information scores calculated for the given features given in sorted order.

```
Selected Features and Their Mutual Information Scores:
                          MI Score
Bedrooms                  0.499703
Building_Type_Studio      0.411192
Size_in_Sqft              0.269920
Building_Type_Condo       0.138937
Building_Type_Apartment   0.132483
Neighborhood_Downtown     0.052551
Bathrooms                 0.052086
Neighborhood_Uptown       0.021079
Neighborhood_Midtown      0.016173
Furnished                 0.001800
Distance_to_City_Center   0.000000
Neighborhood_Suburbs      0.000000
```




- Based on the MI scores and the correlation matrix, we can say that the number of bedrooms, building type being studio, size in squarefeet, building type being condo, and apartment have the most significant effect on the rental price.

- Building type being studio is a significant factor for rental_price, while the type being condo or apartment is not very significant.(Based on MI scores).
- There's a considerable collinearity between the number of bedrooms and the building type being a studio.
- Based on the below plot, we can say that most of the studios have 0 bedrooms.Therefore the two data entries where bedroom count was entered as 'O' could be ,zero count mistakenly entered as 'O'.




- There is an outlier in rental price.

| ID | Size_in_Sqft | Bedrooms | Bathrooms | Distance_to_City_Center | Neighborhood | Furnished | Building_Type | Rental_Price |
|---|---|---|---|---|---|---|---|---|
| 126 | 736 | 4 | 1 | 3.19 | Midtown | Furnished | Condo | 13179.27658 |

- Even compared to other houses with 4 bedrooms, this is a significant outlier.(Since we identified as number of bedrooms being a significant factor affecting rental price.

| ID | Size_in_Sqft | Bedrooms | Bathrooms | Distance_to_City_Center | Neighborhood | Furnished | Building_Type | Rental_Price |
|---|---|---|---|---|---|---|---|---|
| 98 | 1137 | 4 | 1 | 7.70 | suburbs | unfurnished | Condo | 6817.038486 |
| 109 | 1185 | 4 | 2 | 11.98 | Midtown | Furnished | Condo | 7074.471600 |
| 126 | 736 | 4 | 1 | 3.19 | Midtown | Furnished | Condo | 13179.276580 |
| 209 | 1091 | 4 | 1 | 7.61 | Midtown | Furnished | Condo | 8907.430141 |
| 212 | 1001 | 4 | 1 | 10.63 | Downtown | Unfurnished | Condo | 8106.271997 |

# Conclusion

- Based on our analysis, the features which affect the rental price with decreasing order of significance are as follows:

| | MI Score |
|---|---|
| Bedrooms | 0.499703 |
| Building_Type_Studio | 0.411192 |
| Size_in_Sqft | 0.269920 |
| Building_Type_Condo | 0.138937 |
| Building_Type_Apartment | 0.132483 |
| Neighborhood_Downtown | 0.052551 |
| Bathrooms | 0.052086 |
| Neighborhood_Uptown | 0.021079 |
| Neighborhood_Midtown | 0.016173 |
| Furnished | 0.001800 |
| Distance_to_City_Center | 0.000000 |
| Neighborhood_Suburbs | 0.000000 |

- Based on our analysis, the apartment features that affect the rental price the most are the number of bedrooms.
- The features, number of bathrooms, and being furnished or not don't have a significant impact on the rental price.
- The rental price trend that we have identified is that the size of the apartment has an approximately positive linear relationship with the rental price. But the neighborhood doesn't seem to have a strong relationship with the rental price.
- Distance to the city center also doesn't have much influence on the rental price but we can see a small decrease in the rental price with the increase in distance.
- The building-type studio makes the rental price relatively smaller than that being a condo or an apartment( which doesn't have much significance  to the rental price)

# Thank you

Prepared by:

**Team DXP10**