Interpretation Report: House Rent Prediction

Objective:

To build a machine learning model that predicts monthly house rent prices using key property features such as location, size, furnishing status, and tenant preference. The goal is to identify the influential factors and provide accurate rent estimations to aid in real estate decision-making.

Step 1: Data Description

The dataset contains details about house rental listings across major Indian cities. Key attributes include:

- **Size** (sq.ft)
- **BHK**: Number of bedrooms
- **Bathroom**: Number of bathrooms
- **Floor**: Current and total floor numbers
- **Area Type**: Layout type (e.g., Super Area)
- **Area Locality**: Local neighborhood name
- **City**: Location of the house
- **Furnishing Status**: Furnished, Semi-Furnished, or Unfurnished
- **Tenant Preferred**: Target customer segment (Family, Bachelors, Company)
- **Point of Contact**: Broker, Owner, etc.
- **Rent**: Monthly rent (target variable)

Step 2: Data Preprocessing

- **Feature Engineering**:
    - The "Floor" feature was split into **Current Floor** and **Total Floors**.
    - "Ground" was replaced with 0 for numerical consistency.

```python
# Split the "Floor" column into "Current Floor" and "Total Floors"

df[['Current Floor', 'Total Floors']] = df['Floor'].str.split(' out of ', expand=True)
df.head(2)
```

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact	Current Floor	Total Floors
0	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner	Ground	2
1	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner	1	3

```python
# Replace "Ground" with 0 in the "Current Floor" column

df['Current Floor'] = df['Current Floor'].replace({'Ground': 0})
df.head(2)
```

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact	Current Floor	Total Floors
0	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner	0	2
1	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner	1	3

- **Encoding Categorical Variables**:
  - All categorical fields were encoded using **Label Encoder**:
    - Area Type, Area Locality, City, Furnishing Status, Tenant Preferred, and Point of Contact.

```python
from sklearn.preprocessing import LabelEncoder

# Initialize LabelEncoder
encoder = LabelEncoder()
df['Area_Type_encoded'] = encoder.fit_transform(df['Area Type'])


# Initialize LabelEncoder
encoder = LabelEncoder()
df['Area_Locality_encoded'] = encoder.fit_transform(df['Area Locality'])

# Initialize LabelEncoder
encoder = LabelEncoder()
df['City_encoded'] = encoder.fit_transform(df['City'])

# Initialize LabelEncoder
encoder = LabelEncoder()
df['Furnishing_Status_encoded'] = encoder.fit_transform(df['Furnishing Status'])

# Initialize LabelEncoder
encoder = LabelEncoder()
df['Tenant_Preferred_encoded'] = encoder.fit_transform(df['Tenant Preferred'])

# Initialize LabelEncoder
encoder = LabelEncoder()
df['Point_of_Contact_encoded'] = encoder.fit_transform(df['Point of Contact'])

df.head(2)
```

- **Scaling**:
    - Numerical features were standardized using **StandardScaler** for consistent scaling across features.

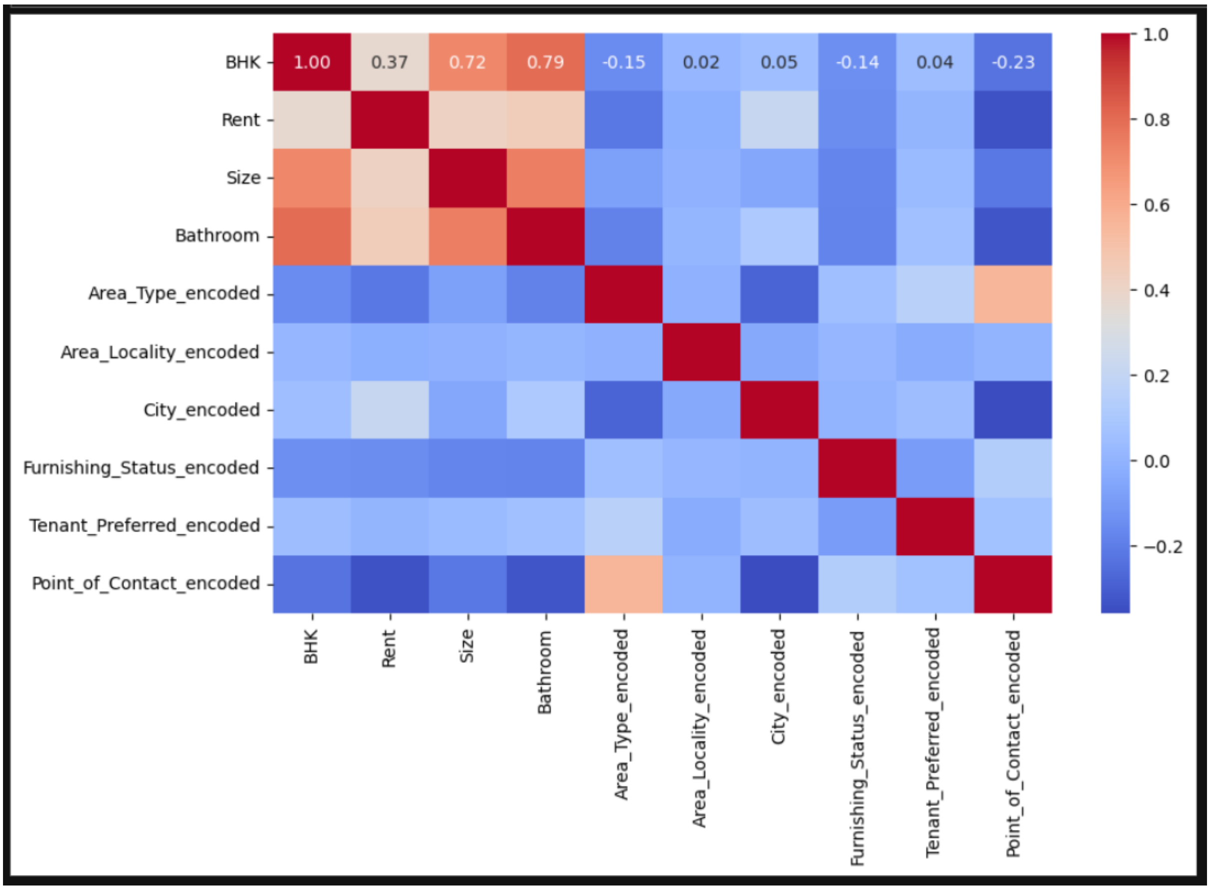
```
# Initialize StandardScaler
scaler = StandardScaler()

# Fit and transform the data
standardized_values = scaler.fit_transform(numeric_cols)

# Convert back to DataFrame
df_standardized = pd.DataFrame(standardized_values, columns=numeric_cols.columns)

df_standardized.head(2)
```

Step 3: Correlation Analysis

- A heatmap was generated using **Seaborn** to examine relationships between numerical features.
- **Size**, **BHK**, and **Bathroom** showed strong positive correlation with **Rent**.
- **City** and **Furnishing Status** also had notable influence based on encoded values.

```
# Compute correlation matrix
correlation_matrix = df_standardized.corr()

# Display correlation values
correlation_matrix
```

```
import seaborn as sns
import matplotlib.pyplot as plt

# Set figure size
plt.figure(figsize=(10,6))

# Create heatmap
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f")

# Show plot
plt.show()
```



Step 4: Model Building

- A **Linear Regression model** was created using **statsmodels**.
- The model formula included key independent variables like:
  - Size, BHK, bathroom, Point of contact.
- Splitting data into different dataset as training dataset and testing dataset.

```
# independent variables
x = df_standardized[['BHK', 'Size', 'Bathroom', 'Point_of_Contact_encoded']]
y = df_standardized['Rent']  # Dependent variable
```

```
print(x.shape, y.shape)
```

```
(4746, 4) (4746,)
```

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression


# Split dataset into training (80%) and testing (20%)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=21)
x_train.head(2)
```

Step 5: Model Interpretation

- Durbin-Watson statistic was found to be 1.68, indicating mild positive autocorrelation among the residuals.
  - This suggests that prediction errors may not be completely independent.
  - It slightly violates the regression assumption of independent errors and may affect statistical inference.
  - A more robust model or diagnostics may be needed for refinement.
- Key Influencers: Size, Bathroom, and BHK significantly impacted rent prices.

Results

1. **Significant Predictors**:
   - **Size of the property**, **number of bathrooms**, and **BHK** count.
   - **City** and **Furnishing Status** also influenced rent values due to market segmentation.
2. **Model Accuracy**:
   - The regression model showed moderate explanatory power.
   - May improve further with advanced models like Random Forest or XGBoost.
3. **Business Application**:
   - Useful for brokers, property owners, and real-estate portals to estimate rent values.
   - Can assist tenants in understanding fair pricing.

Conclusion

The House Rent Prediction project successfully modeled rent prices based on property features using linear regression. With standard preprocessing and careful feature engineering, meaningful insights were derived. For better precision, non-linear models or ensemble techniques could be tested in future iterations.