Delhi House Price Analysis and Prediction

Submitted by

Ritwik Saha, Biplab Mahato, Abhijit Chaki, Sarthak De, Shivangi Mehta

Roll No.:23MA60R02, 23MA60R04, 23MA60R06, 23MA60R14, 23MA60R18

M.Tech in CSDP

Department of Mathematics
IIT KHARAGPUR
Kharagpur, West Bengal
India, 721302

Acknowledgements

First of all, we owe our sincere gratitude to IIT KHARAGPUR for giving us the opportunity to make this project work. We would like to express our special thanks of gratitude to our project coordinator Dr. Buddhananda Banerjee, Professor, Department of Mathematics, IIT Kharagpur, Kharagpur-721302, West Bengal, India who gave us the golden opportunity to do this interesting project work on **"Delhi House Price Analysis and Prediction"**. I would also like to thank our friends and parents for their mental supports, which help us a lot to construct this project work within the limited time frame. Also we would like to give our heartfelt gratitude towards the contribution of https://www.youtube.com//, https://www.google.com/, https://www.kaggle.com/ and different sites which make the work a lot easier to all such aspirants.

Ritwik Saha, Biplab Mahato,
Abhijit Chaki, Sarthak De,
Shivangi Mehta
Department of Mathematics
IIT KHARAGPUR,
Kharagpur-721302,
West Bengal, India

Abstract

In this project, a predictive model has been built, analyzing the Housing Market of Delhi and describing how the House price depends on various factors. Thorough data analysis have been performed and Hypothesis Testing have been done gaining further insight into the data. At last a predictive model have been built using Linear Regression Model and the results have been discussed giving an understanding of the Housing Market of Delhi through Data.

Contents

1 Introduction

The aim of this project is to use different statistical methods like Descriptive Statistical Analysis, Inferential Statistics and Predictive Models and try to extract valuable information from a data-set like Delhi House Price Data-set and apply predictive models in it. The reason behind choosing such a numerical data-set is that, this particular data-set has good variability within it, which enables us to apply all the different techniques that have been learnt in the course. Choosing such a data-set of Delhi works in our favour as such data from a metropolitan region, like the capital, inherently has a lot of variability within it and has a lot of information to extract from it.

2 Data Description

The data-set contains the price records of houses in different regions of Delhi. The data-set has records of 1259 houses, each of which has 11 features including the price. Out of these features 6 are numeric type and rest are categorical type. [1]

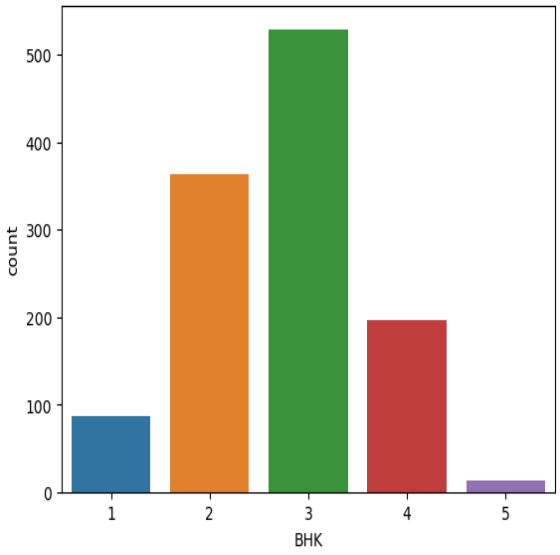
	Area	BHK	Bathroom	Furnishing	Locality	Parking	Price	Status	Transaction	Type	Per_Sqft
0	800.0	3	2.0	Semi-Furnished	Rohini Sector 25	1.0	6500000	Ready_to_move	New_Property	Builder_Floor	NaN
1	750.0	2	2.0	Semi-Furnished	J R Designers Floors, Rohini Sector 24	1.0	5000000	Ready_to_move	New_Property	Apartment	6667.0
2	950.0	2	2.0	Furnished	Citizen Apartment, Rohini Sector 13	1.0	15500000	Ready_to_move	Resale	Apartment	6667.0
3	600.0	2	2.0	Semi-Furnished	Rohini Sector 24	1.0	4200000	Ready_to_move	Resale	Builder_Floor	6667.0
4	650.0	2	2.0	Semi-Furnished	Rohini Sector 24 carpet area 650 sqft status R...	1.0	6200000	Ready_to_move	New_Property	Builder_Floor	6667.0
5	1300.0	4	3.0	Semi-Furnished	Rohini Sector 24	1.0	15500000	Ready_to_move	New_Property	Builder_Floor	6667.0
6	1350.0	4	3.0	Semi-Furnished	Rohini Sector 24	1.0	10000000	Ready_to_move	Resale	Builder_Floor	6667.0
7	650.0	2	2.0	Semi-Furnished	Delhi Homes, Rohini Sector 24	1.0	4000000	Ready_to_move	New_Property	Apartment	6154.0
8	985.0	3	3.0	Unfurnished	Rohini Sector 21	1.0	6800000	Almost_ready	New_Property	Builder_Floor	6154.0
9	1300.0	4	4.0	Semi-Furnished	Rohini Sector 22	1.0	15000000	Ready_to_move	New_Property	Builder_Floor	6154.0

Figure 1: Dataset

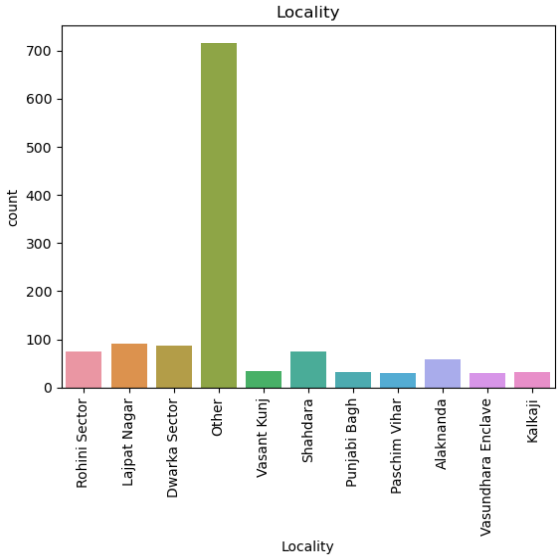
3 Exploratory Data Analysis

In this data exploration part, we have shown distributions of different features within the data-set. We also tried to find various relationships and further insights within the data-set.

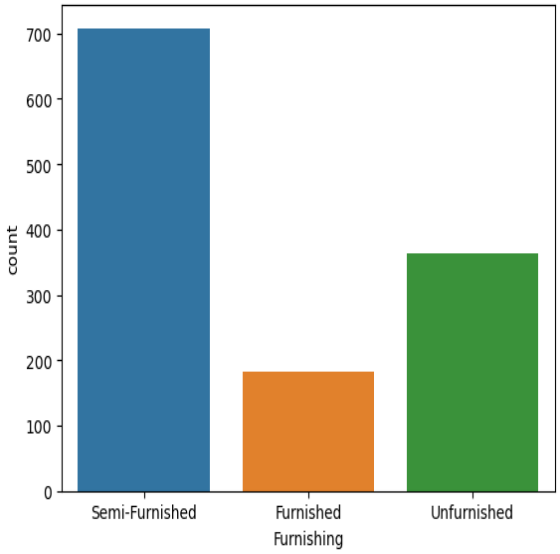
[1] https://www.kaggle.com/code/neelkamal692/starter-delhi-house-price-prediction/input



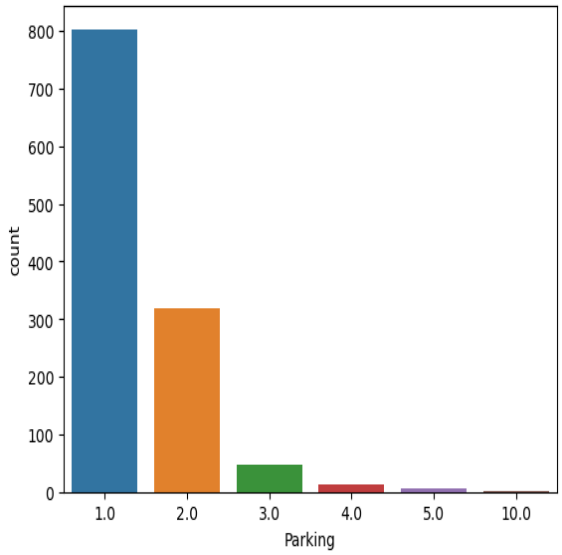
BHK (Bedroom Hall Kitchen): According to the data-set, most of the houses in Delhi are 3-BHK, followed by 2,4,1 and 5-BHK. This gives an idea of the most preferable property sizes to the buyers.



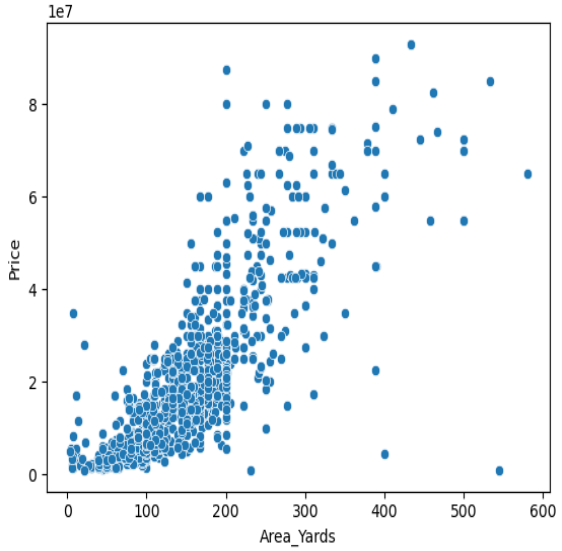
Since there are so many localities and address in the dataset, we group nearly half of them in the top ten localities (count wise), and the remaining localities are grouped as 'Others'. Upon visualizing the locality on the graph, we can see that after the 'Other' category, the Dwarka Sector has highest number of houses followed by Lajpat Naagr and Rohini Sector. From this information, we assume that these localities are good to settle in Delhi.



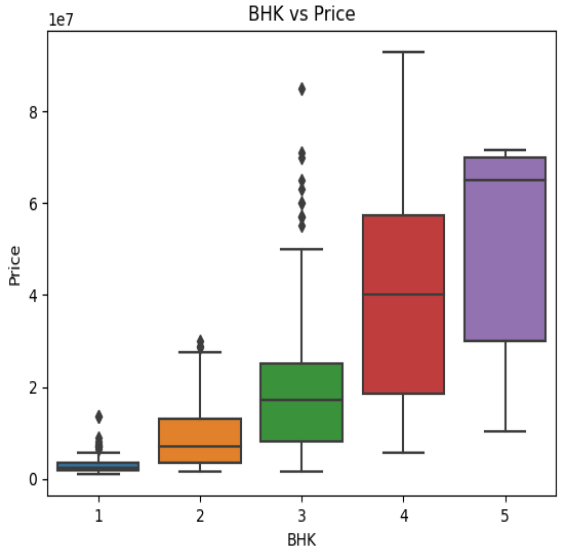
Delhi is the capital region of India. Thus a lot of people from all over the country come to Delhi and naturally look for accommodations. For locals, unfurnished house are preferred, while people from neighbouring regions might prefer Semi-furnished. But, people migrating from distant states can't afford to move their furniture, thus preferring Fully-furnished houses.



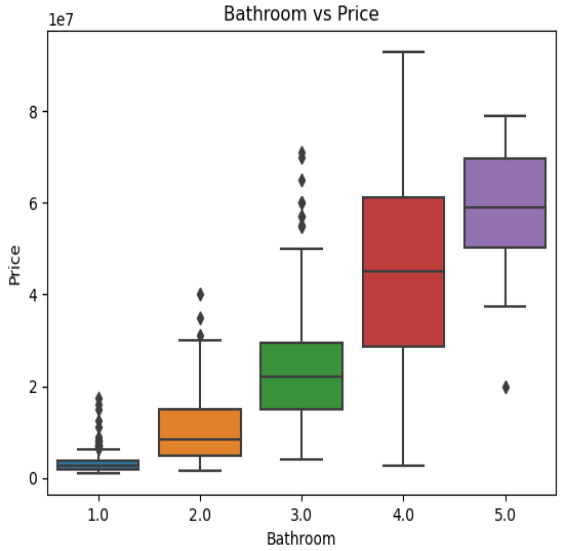
Majority of the houses in Delhi has one car parking which is quite common. Few of the houses have enough space for two car parking and very few houses have more than two car parking space.



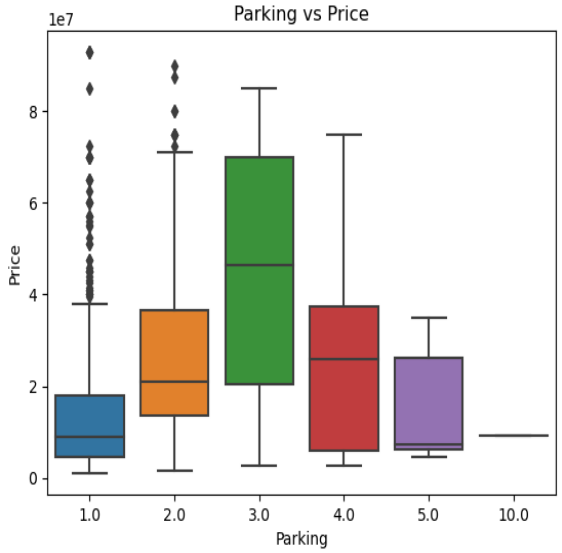
In the Scatter Plot between the target variable Price and feature Area Yards we can observe, there is a general increment in the property prices with an increase in the area, a pattern commonly observed in real estate dynamics. However, an intriguing observation surfaces within this trend. Certain houses exhibit a lower price relative to others with similar areas, indicating that additional factors beyond mere size play a pivotal role in determining house prices.



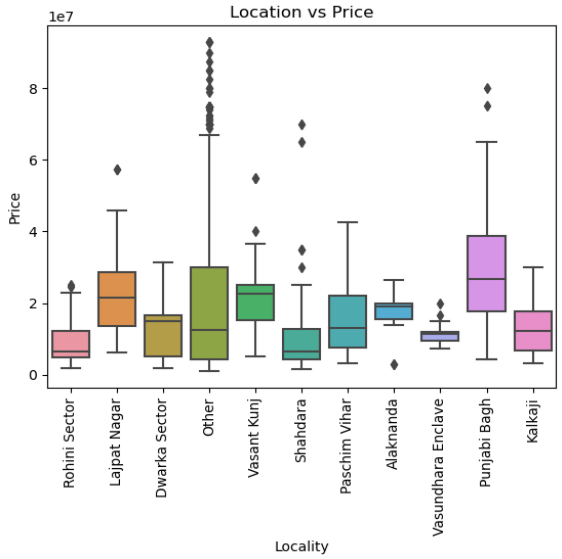
The Box Plot analysis indicates a clear relationship between house prices and the number of bedrooms ('BHK') in the Delhi real estate market. Evidently, there is a positive correlation, showcasing that as the 'BHK' count increases, so does the median house price. Notably, 5 BHK houses claim the highest median price, nearing 7,00,00,000 INR, closely followed by 4 BHK houses with a median price of approximately 4,00,00,000 INR.



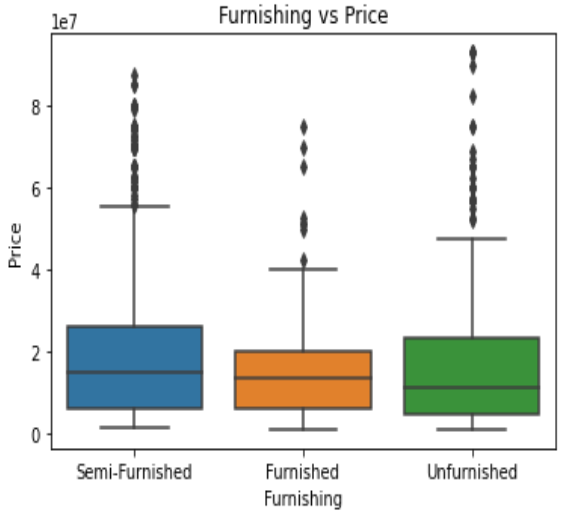
The observed trend in the graph, illustrating the association between house prices and the number of bathrooms, mirrors the patterns discerned in the earlier analysis of BHK and Price. Notably, there is a noticeable escalation in house prices corresponding to an increase in the number of bathrooms. Furthermore, the median prices for each bathroom count closely resemble the trends observed in the previous graph depicting BHK and Price. This consistent alignment strongly suggests a robust correlation between the number of bathrooms and the BHK configuration of the house.



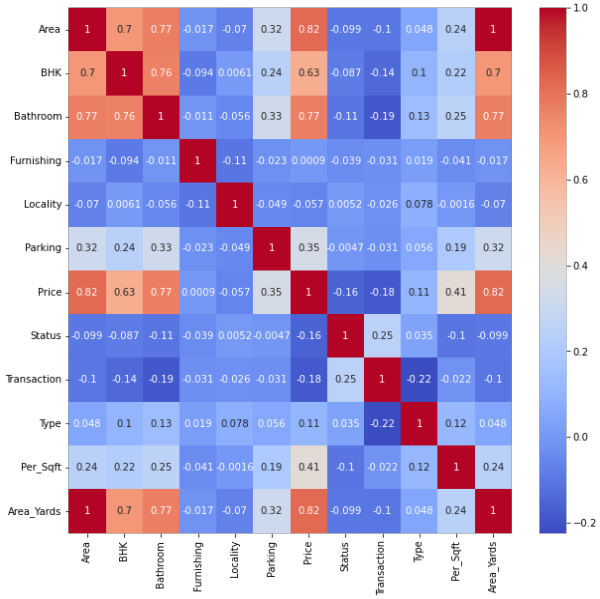
This boxplot visually captures the correlation between the availability of parking spaces and the corresponding house prices. Notably, residences offering 3 parking spaces have the highest median price, which is 4,50,00,000 INR, followed by houses with 2 and 1 parking spaces in descending order. The graph suggests a trend where the number of parking spaces positively influences house prices up to a certain point. Interestingly, houses with 4 or more parking spaces exhibit a lower median price, indicating a potential threshold where excessive parking space may not significantly impact the property's value.



This graphical representation provides insights into the correlation between different localities and house prices. Punjabi Bagh emerges as the locale with the highest median price. Following closely are Lagpat Nagar and Vasant Kunj, both of which can be categorized as posh localities.



A very little difference is observed in the median house prices based on the furnishing status. Notably, an intriguing trend surfaces, as furnished houses exhibit a lower median price compared to semi-furnished residences. Unfurnished houses, on the other hand, have the lowest median price.



Examining the heatmap of the correlation matrix, it becomes evident that the house price exhibits a strong positive correlation with land area, BHK, and bathroom count. This observation validates our earlier findings regarding the relationships among these variables. The heatmap visually reinforces the robust positive correlations, underscoring the interdependencies between the price of the house and key features such as land area, BHK count, and the number of bathrooms.

4 Methodology

We have performed Hypothesis Testing and built a Predictive Model that predicts the possible House Price.

4.1 Hypothesis Testing

We have performed a hypothesis testing with 2 samples to test the mean price of Unfurnished and Semi-Furnished Houses. The Hypothesis are:

$$H_0 : \mu_1 - \mu_2 = d$$

$$H_1 : \mu_1 - \mu_2 \neq d$$

Where, μ_1 is the population mean of prices of Unfurnished Houses and mu_2 is the population mean of prices of Semi-Furnished Houses.
Considering the population variance to be known and sample size of 50, we did Two-Tail test using 2-sample Z-test. The test statistic is:

$$Z = \frac{\bar{X} - \bar{Y} + (\mu_2 - \mu_1)}{\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{m})}}$$

where X is the sample of "Unfurnished" houses and Y is the sample of "Semi-Furnished" houses.

4.2 Prediction Model

4.2.1 Checking Assumption

Linear Regression relies on several assumptions to ensure the validity and reliability of the estimates and inferences. Here we check two assumptions:

- Linearity
- Normality of Residue

4.2.2 Building the Model

	Area	Bathroom	BHK	Furnishing	Price
0	800.0	2.0	3	1	6500000
1	750.0	2.0	2	1	5000000
2	950.0	2.0	2	0	15500000
3	600.0	2.0	2	1	4200000
4	650.0	2.0	2	1	6200000
5	1300.0	3.0	4	1	15500000
6	1350.0	3.0	4	1	10000000

For the prediction of House Price, we have built a prediction model using Multiple Linear Regression model, where our independent features are: "Area", "Bathroom", "BHK", "Furnishing". The dependent feature, that we are going to predict is: "Price".
The model is represented by:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}$$

5 Results and Discussion

5.1 Result of Hypothesis Testing

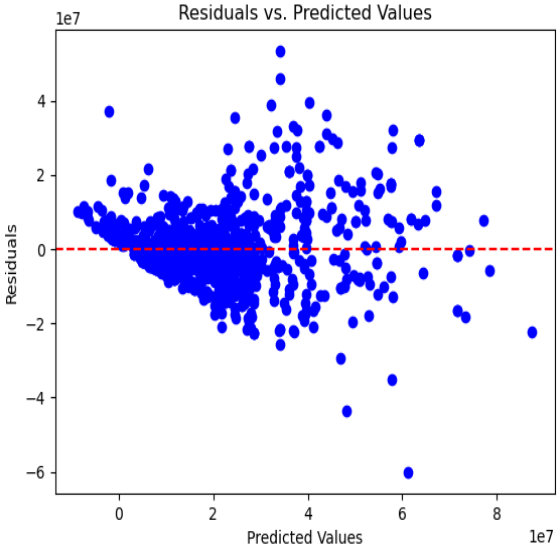
With 95% Confidence ($\alpha = 0.05$) we have failed to reject the null hypothesis H_0, as the observed value of the Z-statistic is about $9.37X10^{-9}$ i.e. within the range -1.96 to 1.96, which are the critical points of the Z-statistic obtained from Z-Table.

5.2 Price Prediction Results

Here we check the validity of the assumptions:

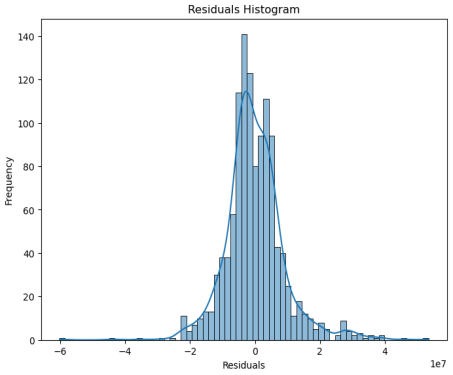
- Linearity
- Normality of Residue

5.2.1 Linearity

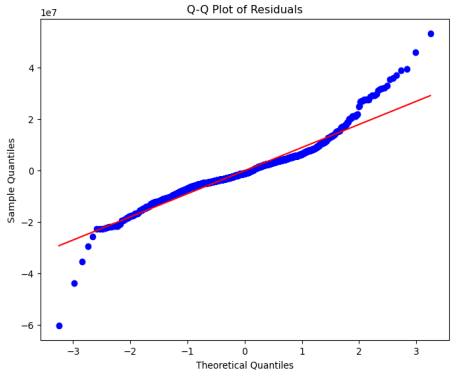


Here we have checked the linear relationship between dependent and independent variables. In the figure we clearly see that the residues are randomly scattered around 0 with no pattern. So, the data is linear.

5.2.2 Normality of Residue



(a) Residuals Histogram



(b) QQplot

Here we check whether residues are following normal distribution or not. From the given two figures we can conclude that our residues follow normal distribution.

Since, both the assumptions are valid, we proceed for the model building.

5.2.3 Model Building

	Area	Bathroom	BHK	Furnishing	Price
0	800.0	2.0	3	1	6500000
1	750.0	2.0	2	1	5000000
2	950.0	2.0	2	0	15500000
3	600.0	2.0	2	1	4200000
4	650.0	2.0	2	1	6200000
5	1300.0	3.0	4	1	15500000
6	1350.0	3.0	4	1	10000000

Here multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

where β_0 is intercept, $\beta_1, \beta_2, \beta_3, \beta_4$ are coefficients and X_i's are independent columns.

Summary of multiple regression:

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.724
Model:	OLS	Adj. R-squared:	0.723
Method:	Least Squares	F-statistic:	776.7
Date:	Mon, 13 Nov 2023	Prob (F-statistic):	0.00
Time:	14:48:12	Log-Likelihood:	-20766.
No. Observations:	1189	AIC:	4.154e+04
Df Residuals:	1184	BIC:	4.157e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.307e+07	1.1e+06	-11.862	0.000	-1.52e+07	-1.09e+07
Area	1.393e+04	587.646	23.700	0.000	1.28e+04	1.51e+04
Bathroom	6.7e+06	5.12e+05	13.091	0.000	5.7e+06	7.7e+06
BHK	-1.237e+06	5.13e+05	-2.409	0.016	-2.24e+06	-2.3e+05
Furnishing	2.476e+05	4.26e+05	0.581	0.561	-5.88e+05	1.08e+06

Omnibus:	192.491	Durbin-Watson:	1.632
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1246.676
Skew:	0.570	Prob(JB):	1.94e-271
Kurtosis:	7.885	Cond. No.	6.46e+03

Figure 2: Regression summary

6 Conclusion

In summary, this project has not only unveiled crucial insights into the Housing Market of Delhi and it's dynamics, but it has also demonstrated the potential for applying such models to predict the housing prices which helps both buyers and real-estate agents. The models used are multi linear regression. We have been able to apply various statistical methods to analyze and build a predictive model for the housing market.

This model will predict the house prices correctly so that the real estate company does not suffer from any losses. The buyers will get the idea of the house price so that they can decide whether they can buy it or not.

7 References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "Introduction to Statistical Learning." Springer.

2. John A Rice. Mathematical statistics and data analysis. Cengage Learning, 2006.

3. https://www.kaggle.com/code/neelkamal692/starter-delhi-house-price-prediction/input

4. Mueller, J. P., & Massaron, L. (2016). "Python for Data Science For Dummies." Wiley.

8 Appendix

```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

from mpl_toolkits.mplot3d import Axes3D
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt # plotting
import numpy as np # linear algebra
import os # accessing directory structure
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

df=pd.read_csv("MagicBricks.csv")

df.isnull().sum()

# Replacing missing value in Per_Sqft
df['Per_Sqft'] = df['Per_Sqft'].fillna(df['Price']/df['Area'])

# Replacing missing values in Parking, Bathroom, Furnishing and Type
df['Parking'].fillna(df['Parking'].mode()[0], inplace=True)
df['Bathroom'].fillna(df['Bathroom'].mode()[0], inplace=True)
df['Furnishing'].fillna(df['Furnishing'].mode()[0], inplace=True)
df['Type'].fillna(df['Type'].mode()[0], inplace=True)


# Type casting
df[['Parking', 'Bathroom']].astype('int64')

def grp_local(locality):
   locality = locality.lower() # avoid case sensitive
   if 'rohini' in locality:
    return 'Rohini Sector'
   elif 'dwarka' in locality:
    return 'Dwarka Sector'
   elif 'shahdara' in locality:
    return 'Shahdara'
   elif 'vasant' in locality:
    return 'Vasant Kunj'
   elif 'paschim' in locality:
    return 'Paschim Vihar'
   elif 'alaknanda' in locality:
    return 'Alaknanda'
   elif 'vasundhar' in locality:
    return 'Vasundhara Enclave'
```

```
    elif 'punjabi' in locality:
     return 'Punjabi Bagh'
    elif 'kalkaji' in locality:
     return 'Kalkaji'
    elif 'lajpat' in locality:
     return 'Lajpat Nagar'
    else:
     return 'Other'
df['Locality'] = df['Locality'].apply(grp_local)


# Using Z - score to remove outliers
from scipy import stats
# Z score
z = np.abs(stats.zscore(df[df.dtypes[df.dtypes != 'object'].index]))
# Removing outliers
df = df[(z < 3).all(axis=1)]

sns.histplot(x = df['Area_Yards'], kde = True, bins =
50).set_title('Area_Yards')

sns.countplot(x = 'BHK', data = df).set_title('BHK')

sns.countplot(x = 'Bathroom', data = df).set_title('Bathroom')

sns.countplot(x='Furnishing',data=df).set_title('')

sns.countplot(x = 'Parking', data = df).set_title('')

sns.scatterplot(x = 'Area_Yards', y = 'Price', data = df)

sns.boxplot(x = 'BHK', y = 'Price',
data = df).set_title('BHK vs Price')

sns.boxplot(x = 'Bathroom', y = 'Price', data = df).set_title
('Bathroom vs Price')

sns.boxplot(x = 'Parking', y = 'Price',
data = df).set_title('Parking vs Price')

sns.boxplot(x='Locality', y='Price', data=df).set_title('Location
vs Price')plt.xticks(rotation=90)

#Hypothesis testing
X=df[df["Furnishing"]=="Unfurnished"]
Y=df[df["Furnishing"]=="Semi-Furnished"]

mu_1=X["Price"].mean()
```

```
mu_2=Y["Price"].mean()

arr_X=np.random.randint(1,183,50)
arr_Y=np.random.randint(1,183,50)

X_sample_df=X.iloc[arr_X]
Y_sample_df=Y.iloc[arr_Y]

X_bar=X_sample_df["Price"].mean()
Y_bar=Y_sample_df["Price"].mean()
print(X_bar)
print(Y_bar)

var_1=X["Price"].var()
var_2=Y["Price"].var()

#Hypothesis testing for "Furnished"
z_obs=((X_bar-Y_bar)-(mu_1-mu_2))/((var_1*(1/50 + 1/50)**0.5))
print(z_obs)

if(z_obs>1.96 or z_obs<-1.96):
    print("Null Hypothesis is Rejected")
else:
    print("Null Hypothesis is Accepted")


#Data processing
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
# Columns for label encoding
cols = ['Furnishing', 'Locality', 'Status', 'Transaction', 'Type']
for i in cols:
   le.fit(df[i])
   df[i] = le.transform(df[i])
   print(i, df[i].unique())

corr_df = df.corr()

strong_relation_features = pd.Series(corr_df['Price']).
nlargest(n=7).iloc[1:]

new2=df[["Area", "Bathroom", "BHK","Furnishing","Price"]]

fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 5))

# Plot x1 and y
ax1.scatter(new2["Area"], df['Price'], color='blue')
ax1.set_title('Scatter plot of Arae and Price')
```

```
ax1.set_xlabel('Area')
ax1.set_ylabel('Price')

# Plot x2 and y
ax2.scatter(df['Bathroom'], df['Price'], color='red')
ax2.set_title('Scatter plot of Bathroom and Price')
ax2.set_xlabel('Bathroom')
ax2.set_ylabel('Price')

plt.tight_layout()
plt.show()

# Fit a linear regression model
X = df[["Area",        "Bathroom",        "BHK",        "Furnishing"]]
y = df['Price']
from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(X, y)

# Calculate predicted values and residuals
y_pred = model.predict(X)
residuals = y - y_pred

# Plot residuals against predicted values
plt.scatter(y_pred, residuals, color='blue')
plt.axhline(y=0, color='red', linestyle='--')
plt.title('Residuals vs. Predicted Values')
plt.xlabel('Predicted Values')
plt.ylabel('Residuals')

plt.show()

# Histogram
plt.figure(figsize=(8, 6))
sns.histplot(residuals, kde=True)
plt.title('Residuals Histogram')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.show()


from scipy import stats
plt.figure(figsize=(8, 6))
stats.probplot(residuals, plot=plt)
plt.title('Q-Q Plot of Residuals')
plt.xlabel('Theoretical Quantiles')
plt.ylabel('Sample Quantiles')
```

```
plt.show()


#Model Fitting
import pandas as pd
import statsmodels.api as sm
# Load the dataset
# Define the independent variables (add a constant for the intercept)
X=df[["Area",        "Bathroom",        "BHK",        "Furnishing"]]
X = sm.add_constant(X)
# Define the dependent variable
y = df['Price']
# Fit the model using the independent and dependent variables
model = sm.OLS(y, X).fit()
# Print the summary of the model
print(model.summary())
```