ANALYSIS OF HOUSE PRICEPREDICTION IN CHENNAI CITY

Report submitted to the
SDM COLLEGE (Autonomous)

In partial fulfilment of the degree of
MASTER OF SCIENCE

IN
STATISTICS
By
NIDHISHA
Under the supervision of
Ms. SHWETHA KUMARI

Department of Postgraduate Studied
In Statistics

SRI DHARMASTHALA MANJUNATHESHWARA
COLLEGE (Autonomous)
UJIRE-574240

Contents:

1 Chapter 1

0.1 Introduction:

Real estate transactions are quite opaque sometimes and it may be difficult for a newbie to know the fair price of any given home. Thus, multiple real estate websites have the functionality to predict the prices of houses given different features regarding it. Such forecasting models will help buyers to identify a fair price for the home and also give insights to sellers as to how to build homes that fetch them more money. Chennai house sale price data is shared here and the participants are expected to build a sale price prediction model that will aid the customers to find a fair price for their homes and also help the sellers understand what factors are fetching more money for the houses.

Accurately estimating the value of real estate is an important problem for many stakeholders including house owners, house buyers, agents, creditors, and investors. It is also a difficult one. Though it is common knowledge that factors such as the size, number of rooms and location affect the price, there are many other things at play. Additionally, prices are sensitive to changes in market demand and the peculiarities of each situation, such as when a property needs to be urgently sold. The sales price of a property can be predicted in various ways, but is often based on regression techniques. All regression techniques essentially involve one or more predictor variables as input and a single target variable as output. In this paper, we compare different machine learning methods performance in predicting the selling price of houses based on a number of features such as the area, the number of bed- and bathrooms and the geographical position.

0.1.1 Objective and Scope of study:

- To compare how the sales prices of the house varies with different area in Chennai.

- To study the factor influencing the sales price of the house:
    - Depends on the distance from the main road to the house
    - Depends on the parking facility
    - Depends on the build type of the house.
    - Number of rooms, bathrooms, bedrooms.
    - Depends on utility available.
    - Depends on the square feet of the house
    - Depends on the sales condition
    - Depends on the types of street

- To predict the sales prices of the house for customers.

Scope of the study:

House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. There are three factors that influence the price of a house which include physical conditions, concept and location.

2 Chapter 2:

0.2 Materials and Methods:

About the Data:

House price prediction in Chennai city data is collected from "**https://www.kaggle.com/search".**

The data refers to the details of house and price in different area of the Chennai.

Row: Id of the house.

Columns:

- Area: different area in Chennai
- Int_sqt: Square feet of the house
- Dist_mainroad: Dsitance from the main road
- N_bedroom: Number of bedroom
- N_bathroom: Number of bathroom
- N_room: number of room
- Sale_condtion: Sale condition of the house
- Build_type: type of the building
- Utility_avail: available of the utility
- Park_facility: Availability of parking facilities
- Street: type of street
- Commis: commission

Location: Chennai.

Python Libraries:

Numpy: Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

Pandas: Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

Matplotlib: Matplotlib is a plotting library for the Python programming.

Plotly: Plotly's Python graphing library makes interactive, publication quality graphs. Example of how to make line plots, area, charts, bar chart, boxplots, Histogram.

Seaborn: Seaborn is Python data visualization library based on matplotlib.

Scikit learn: scikit learn is an open-source Python library that implements a range of machine learning, pre-processing, cross-validation, and visualization algorithms using a unified interface.

Statistical techniques:

- **t-Distribution:** The t-distribution, like the normal distribution, is bell-shaped and symmetric, but it has heavier tails, which means that it tends to produce values that fall far from its mean
- **One-way ANOVA**: A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- **Multiple linear regression**: Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.
- **Mann Whiteny U test**: The Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed

- **Bar Graph**: A bar graph can be defined as a graphical representation of data, quantities, or numbers using bars or strips. They are used to compare and contrast different types of data, frequencies, or other measures of distinct categories of data

- **Histogram:** A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a

data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins

- **Boxplot:** displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median.

- **Scatter plot**: Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.**:**

Literature Review:

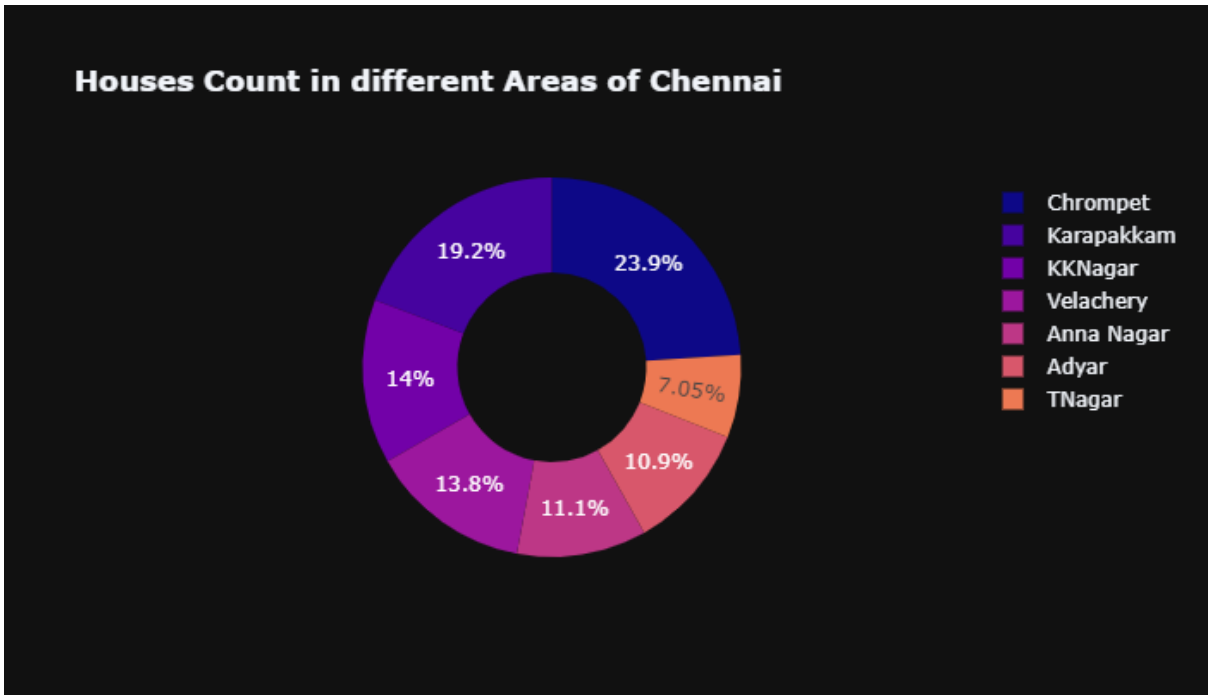
Sifei Lu, proposed a hybrid regression technique for house price prediction. With limited dataset and data features, creative feature engineering method is examined in this paper. The proposed approach has recently has been deployed as the key kernel for Kaggle Challenge "House Price: Advance Regression Techniques". The goal of the paper is to predict reasonable price for customers with respect to their budgets and priorities.

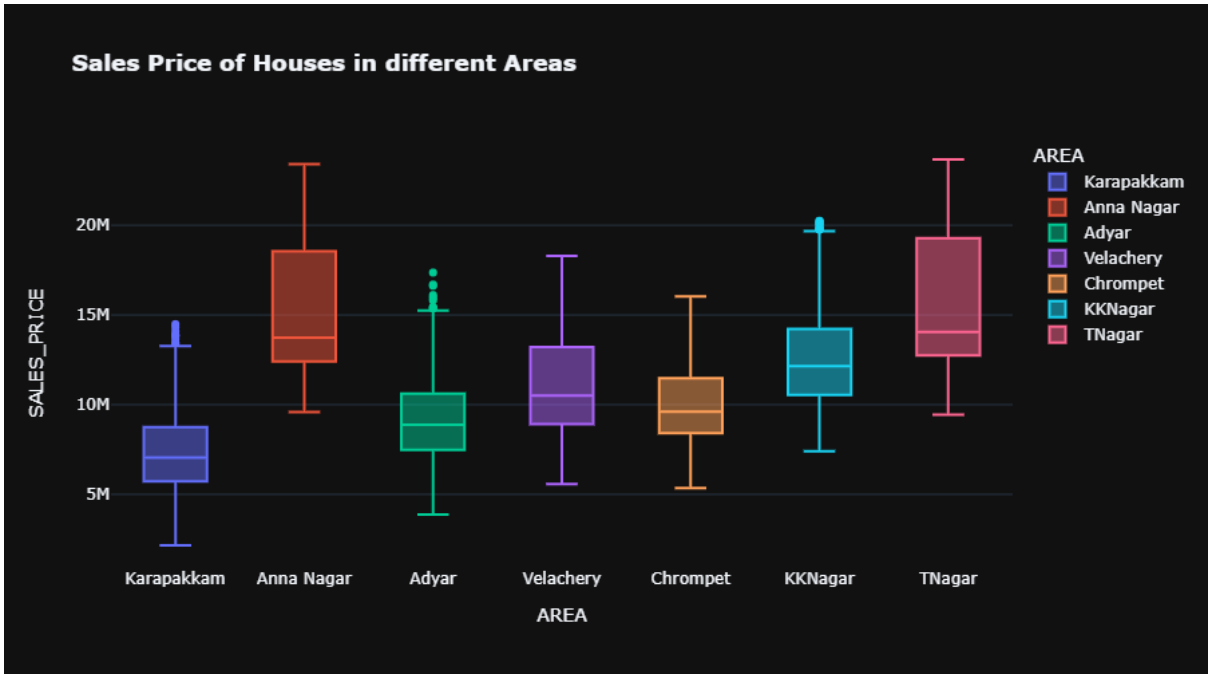
In 2020, Shuzlina Abdul Rahman. conducted a survey on" House Price Prediction using a Machine Learning Mode". Objectives of this survey was to determine the house price prediction. The survey was concluded as follows: Predicting the fair price of the house for customer in that current scenario. And authors have aimed to developed different methods in order to estimate the real market price.

3 Chapter 3

0.3 Results and Discussion:

Sales prices of the house varies with different area in Chennai.





Conclusion:

From the graph we observed that number of houses in crompet is more and number of houses in Tnager is less. From the above box plot we observed that sales prices of the house in Anna Nagar and T Nagar is more. So our conclusion is sales prices of the house varies with different area in Chennai.

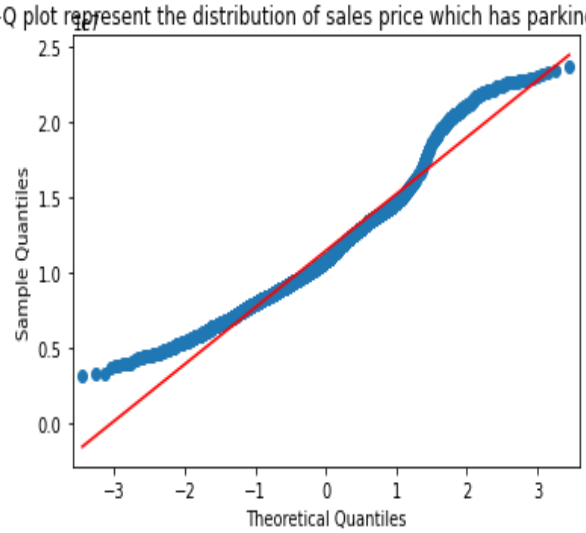
The factor influencing the sales price of the house:

- **The factor parking facility is influencing the sales price of the house:**
  <u>t- distribution:</u>

  Checking the assumption:

  Checking the normality:

Q-Q plot represent the distribution of sales price which has parking faciliies.



Both graph are not normally distributed and it is positively skewed. Hence we are using the non-parametric Mann Whitney U test.

Hypothesis:

H_0: There is no significant difference of the prices of the house based on the parking facilities.

H_1: There is a significant difference of the prices of the house based on the parking facilities.

Using Mann-Whitney U test

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

P_value= `5.39632214e-38`

We reject the null hypothesis if p_value<0.05 Hence we reject the null hypothesis.

Conclusion:

There is a significant difference of the prices of the house Based on the parking facilities.

Hypothesis:

H_0: There is no significant mean difference between the sale prices of the house based on the the parking facilities

H_1: There is a significant mean difference between the sale prices of the house based on the the parking facilities.

Using t-distribution,

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t= 10.95155479
p_value= 3.58167338e-07

We reject the null hypothesis if p_value<0.05 Hence we reject the null hypothesis.

Conclusion:

There is a significant mean difference between the sale prices of the house based on the the parking facilities.

- **The factor build type is influencing the sales price of the house:**

Hypothesis:

H_0: There is no significant mean difference between the sale prices of the house based on the different build type.

H_1: There is significant mean difference between the sale prices of the house based on the different build type.

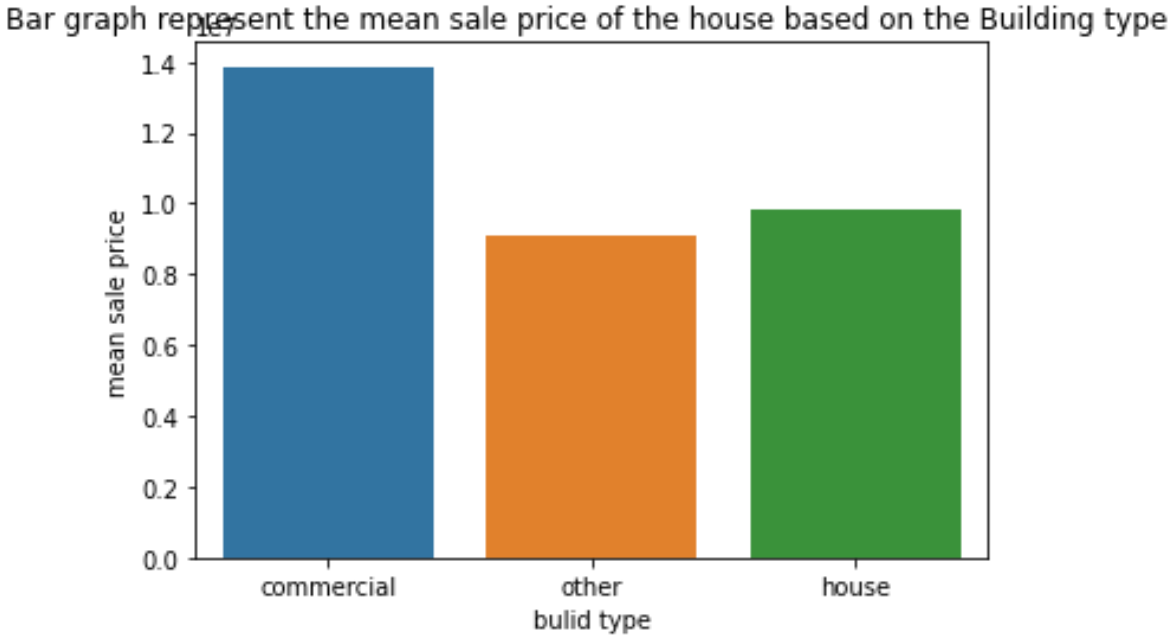
Using the ANOVA,

p_value= 000000.12

We reject the null hypothesis if p_value<0.05 Hence we reject the null hypothesis.

Conclusion:

There is a significant mean difference between the prices of the house between the build type.



Conclusion:

From the above bar graph we observed that mean sale price of the commercial land is more.

- **The factor available of utility is influencing the sales of the house:**

Hypothesis:

H_0: There is no significant mean difference between the prices of the house based on the available of utility.

H_1: There is a significant mean difference between the prices of the house based on the available of utility.

Using the ANOVA,

p_value= `3.58167338e-07`

We reject the null hypothesis if p_value<0.05 Hence we reject the null hypothesis.

Conclusion:

There is a significant mean difference between the prices of the house based on the available of utility.

- **The factor sales condition is influencing the sales of the house:**

Hypothesis:

H_0: There is no significant mean difference between the prices of the house based on the sales condition.

H_1: There is a significant mean difference between the prices of the house based on the sales condition.

Using the ANOVA,

p_value= `3.58167338e-07`

We reject the null hypothesis if p_value<0.05 Hence we reject the null hypothesis.

Conclusion:

There is a significant mean difference between the prices of the house based on the sales condition.

- **The factor types of street is influencing the sales of the house:**

Hypothesis:

H_0: There is no significant mean difference between the prices of the house based on the types of street.

H_1: There is a significant mean difference between the prices of the house based on the types of street.

Using the ANOVA,

p_value= `1.62790824e-38`

We reject the null hypothesis if p_value<0.05 Hence we reject the null hypothesis.

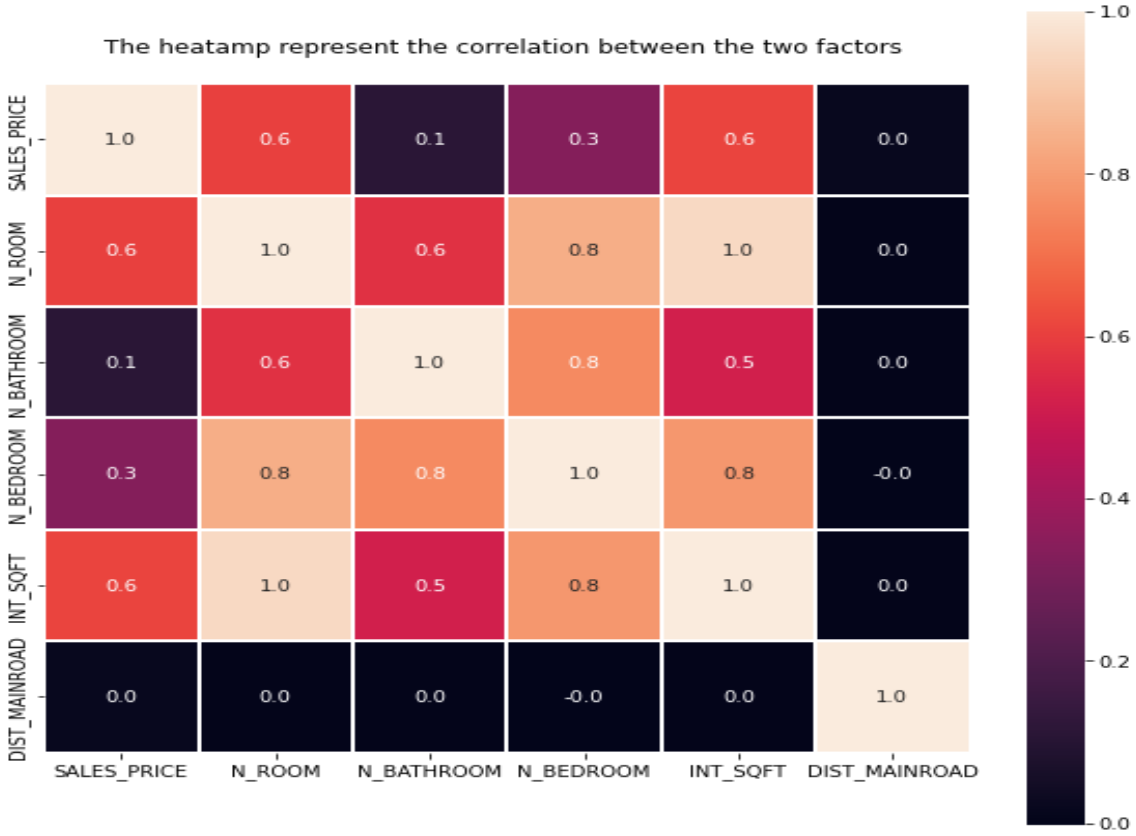
Conclusion:

There is a significant mean difference between the prices of the house based on the types of street.

Overall Conclusion:

The parking facilities, build type, utility availables, sales condition, street these are factors influenced to the sale prices of the house.

- **The factors distance from the main road, number of rooms, bathrooms, bedrooms, square feet is influencing the sales price of the house:**



Conclusion:

From the above graph represent the number of rooms and square feet is correlated with sales price As number of rooms and square feet increases the prices of the house also increases. Number of rooms is correlated with bedrooms and bathrooms, so the factors number of bedroom and bathrooms influence the prices of the house. But the factor distance from the main road is not correlated with prices of the house. So price of the house is not effect from the distance of the house from main road.

To predict the sales prices of the house for customers:

By using multiple linear regression,

$Y=b_o+b_1X_1+b_2X_2\ldots+b_nX_n$
Y=Sales of the house prices
X=All the regressor.

(i) Here we took all the factors for the multiple linear regression.

Mean square Error (MSE): 1375097601245.89
Mean root square Error (MRSE): 1172645.5565284379
R2_score: 0.89533

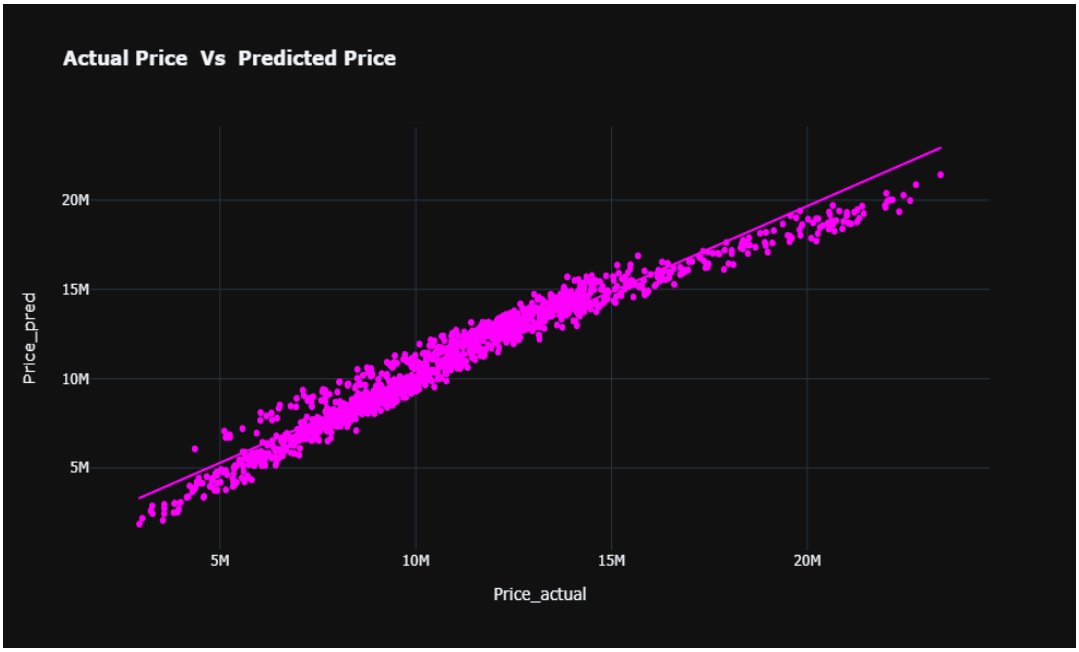
Actual value v|s predicted value:

	Price_actual	Price_pred
1543	13282450	1.281938e+07
3497	22022260	2.038723e+07
5435	15053510	1.550778e+07
3539	5624750	5.897580e+06
6470	5076885	4.767712e+06
4231	14873340	1.452878e+07
2098	10933000	1.254410e+07
2318	7947370	7.865278e+06
4903	14083210	1.421266e+07
1426	8528890	8.808100e+06

(ii) Here we took only influencing the factors and built the multiple linear regression.

Mean square Error (MSE): 602389011094.275
Mean root square Error (MRSE): 776137.2372810591
R2_score: 0.954147

	Price_actual	Price_pred
6827	11469920	1.229784e+07
4728	10313310	1.144022e+07
3263	15043490	1.572328e+07
4710	9664460	9.786042e+06
4046	12632900	1.217865e+07
3989	9808360	9.714722e+06
7050	9406680	9.088067e+06
1009	7809500	7.432199e+06
577	10769525	1.029253e+07
3664	12149050	1.302103e+07
3955	9045605	8.589710e+06
346	7794500	9.133068e+06

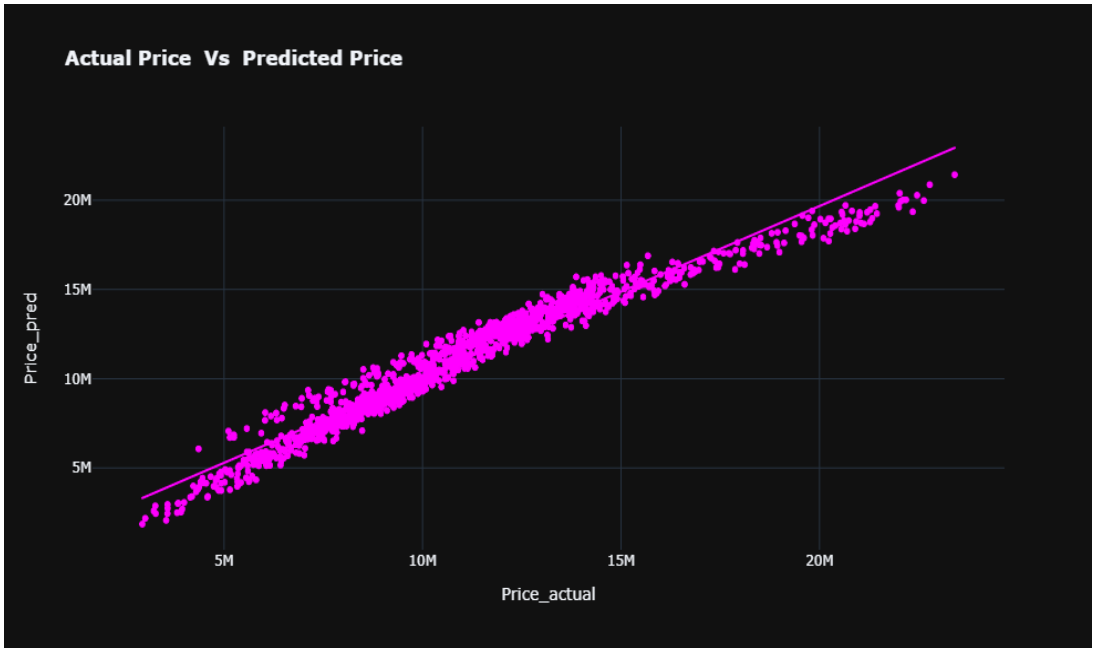


Conclusion: From the above result we observed that in graph (i) variables are spreaded much comparatively graph (ii). Then the mean square error is more in graph (i) comparatively graph (ii) and R score is more in graph (ii). So the second model is more accurate than the first model.

By using the ridge regression:

```
Mean root square error(MRSE):776096.4457551298.
R2_score: 0.954152
```

	Price_actual	Price_pred
339	11045960	1.146317e+07
6693	8416200	8.286463e+06
5259	11544130	1.232166e+07
6169	14299300	1.374750e+07
2977	13172210	1.290670e+07
3950	8957390	8.865209e+06
6909	14344300	1.402479e+07
2398	5783750	5.737768e+06
5995	6972000	7.157933e+06
3946	10065600	9.669355e+06



Conclusion: From the above graph we observed that the data is not spreded much and accuracy of the model is 0.954152.

4 Chapter 4

0.4 Conclusion

- The sales price of the house is varies in different area. Sales price of the house is very high in Anna nagar and TN nagar.
- Sales price of the house is varies based on the parking facilities. If there is a parking facilities then price of the house is high.
- Based on the build type, the price of the house is varies. Commercial building type as high price.
- Sales price of the house is high based on the available of the utility.
- Based on the type of the street and sales condition of the house, the price varies.
- Based on the number of the rooms, bathroom, bedroom and square feet of the house the price of the house is varies but the sales price of the house is not vary based on the distance of the main road from house.
- Our model (ii) and (iii) predicted value is almost similar to the actual value. Hence we can conclude that model (ii) and (iii) has more accuracy.

0.4.1

Overall Conclusion:

House price prediction can helps to develops determine the selling price of a house the customer to arrange the time to purchase a house. It helps the real estate owner to define the predicted amount of the house using various influencing factors. Real estate is the least transparent industry in our ecosystem. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main objectives of this project. Here we evaluated based on every basic parameter that is considered while determining the price of the house. We observed which and all factors influencing the price of the house. Our project also includes estimating the price of houses without any expectations of market prices and cost increment. The main objective of this project is prediction of residential prices for the customers considering their financial plans and needs. It will helps the client based on the requirements.

5 References:

https://www.sciencedirect.com/science/article/pii/S1877050920316318?ref

https://www.geeksforgeeks.org/

https://www.diva-portal.org/smash/get/diva2:1456610/FULLTEXT01.pdf

https://www.researchgate.net/publication/349477129_House_Price_Prediction

Appendix:

checking the normalities:

```
sm.qqplot(df["SALES_PRICE"],line="s")
plt.title(" Q-Q plot represent the distribution of sales price")
```

Influencing the sales price of the house based on parking facilities

```
df["PARK_FACIL"].unique()park_no,park_yes=df.groupby("PARK_FACIL")["SALES
_PRICE"]
park_no=pd.DataFrame({"sales_park_no":park_no[1]})
park_yes=pd.DataFrame({"sales_park_yes":park_yes[1]})
sm.qqplot(park_no["sales_park_no"],line="s")
plt.title(" Q-Q plot represent the distribution of sales price which has no parking
facilities")
from scipy.stats import mannwhitneyu
 from scipy.stats import ttest_ind
stat, p_value = mannwhitneyu(park_no, park_yes)
p_value
if(p_value<0.05):
      print("We are rejecting the null hypothesis")
      print()
else:
      print("We are not rejecting the null hypotesis")
      print()
```

Influencing the sales price of the house based on type of the buildings:

```
df["BUILDTYPE"].unique()

from scipy.stats import f_oneway

commercial,other,house=df.groupby("BUILDTYPE")["SALES_PRICE"]
commercial=pd.DataFrame({"commercial":commercial[1]})
other=pd.DataFrame({"other":other[1]})
house=pd.DataFrame({"house":house[1]})

statistic,p_value=f_oneway(commercial,other,house)
p_value

if(p_value<0.05):
   print("We are rejecting the null hypothesis")
   print()
else:
   print("We are not rejecting the null hypotesis")
   print()

build=["commercial","other","house"]
a=[np.mean(commercial["commercial"]),np.mean(other["other"]),np.mean(house["h
ouse"])]

a=pd.DataFrame({"mean":a,"type":build})
sns.barplot(x=a["type"],y=a["mean"])
plt.xlabel("bulid type")
plt.ylabel("mean sale price")
plt.title("Bar graph represent the mean sale price of the house based on the Building
type ")
```

Influencing the sales price of the house based on the sales condition

```
df["SALE_COND"].unique()

from scipy.stats import f_oneway

abnormal,family,partial,adjland,normal_sale=df.groupby("SALE_COND")["SALES
_PRICE"]
abnormal=pd.DataFrame({"AbNormal":abnormal[1]})
family=pd.DataFrame({"Family":family[1]})
partial=pd.DataFrame({"Partial":partial[1]})
adjland=pd.DataFrame({"AdjLand":adjland[1]})
normal_sale=pd.DataFrame({"Normal Sale":normal_sale[1]})
```

```
statistic,p_value=f_oneway(abnormal,family,partial,adjland,normal_sale)
p_value

if(p_value<0.05):
    print("We are rejecting the null hypothesis")
    print()
else:
    print("We are not rejecting the null hypotesis")
    print()
```

Influencing the sales price of the house based on the distance from the house, number of rooms, bedrooms, bathrooms and square feet of the house

```
df_corr=df[["SALES_PRICE","N_ROOM","N_BATHROOM","N_BEDROOM",
"INT_SQFT","DIST_MAINROAD"]].corr()
df_corr

f,ax=plt.subplots(figsize=(10,10))
sns.heatmap(df_corr,annot=True,fmt="0.1f",linewidth="0.4",ax=ax,square=True)
plt.title("The heatamp represent the correlation between the two factors \n")
```

Compare the sales price of the house in different area:

```
area_sales=df["AREA"].value_counts()
area_sales
area_sales=pd.DataFrame(area_sales)
area_sales.reset_index(inplace=True)
area_sales.rename(columns={"index":"Area","AREA":"No_of_house"},inplace=Tru
e)
area_sales

px.pie(area_sales,values="No_of_house",template='plotly_dark',names="Area",colo
r_discrete_sequence=px.colors.sequential.Plasma,hole=0.5,width=700,height=400,ti
tle='<b> Houses Count in different Areas of Chennai</b>')

px.box(df,x='AREA',y='SALES_PRICE',color='AREA',template='plotly_dark',widt
h=900,height=500,title='<b> Sales Price of Houses in different Areas')
```

Build the regression¶

```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
```

```
# Coverting the labels into a numeric form using Label Encoder
for col in df.columns:
    if(df[col].dtypes=="object"):
        df[col]=le.fit_transform(df[col])

x=df.drop('SALES_PRICE',axis=1)
y=df['SALES_PRICE']

from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
x_scaled=scaler.fit_transform(x)

from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=42)
x_train.shape,x_test.shape,y_train.shape,y_test.shape

from sklearn import linear_model
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score

reg=linear_model.LinearRegression()
reg.fit(x_train,y_train)

y_pred=reg.predict(x_train)

print('Mean square Error (MsE):', round(mean_squared_error(y_train, y_pred),3))
print('Mean square Error (MRSE):', math.sqrt(round(mean_squared_error(y_train,
y_pred),3)))
print('R2_score:', round(r2_score(y_train, y_pred),6))

y_pred=reg.predict(x_test)


print('Mean square Error (MsE):', round(mean_squared_error(y_test, y_pred),3))
print('Mean square Error (MRSE):', math.sqrt(round(mean_squared_error(y_test,
y_pred),3)))
print('R2_score:', round(r2_score(y_test, y_pred),6))


out=pd.DataFrame({'Price_actual':y_test,'Price_pred':y_pred})
result=df.merge(out,left_index=True,right_index=True)
result[['Price_actual','Price_pred']].sample(20)


px.scatter(result,x='Price_actual',y='Price_pred',trendline='ols',color_discrete_seque
nce=['magenta'],template='plotly_dark',title='<b> Actual Price  Vs  Predicted Price ')
```

Ridge regression

```
df=pd.read_csv("df.csv")
df=df.drop(columns=["PRT_ID","Unnamed: 0"])

df=pd.get_dummies(df,columns=["AREA","SALE_COND","SALE_COND","PAR
K_FACIL","BUILDTYPE","UTILITY_AVAIL","STREET","MZZONE"])
df.head(2)

x=df.drop(['SALES_PRICE','DIST_MAINROAD'],axis=1)
y=df['SALES_PRICE']

from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
x_scaled=scaler.fit_transform(x)

ridge=Ridge(normalize=True)
search=GridSearchCV(estimator=ridge,param_grid={"alpha":np.logspace(-5,2,8)},
            scoring="neg_mean_squared_error",n_jobs=1,refit=True,cv=10)

search.fit(x_train,y_train)
search.best_params_



ridge=Ridge(normalize=True,alpha=0.0001)
ridge.fit(x_train,y_train

y_pred=ridge.predict(x_train)

print('R2_score:', round(r2_score(y_train, y_pred),6))

y_pred=ridge.predict(x_test)
m=mean_squared_error(y_test,y_pred)
print("Mean root square error(MRSE)",math.sqrt(m))

print('R2_score:',round(r2_score(y_test,y_pred=ridge.predict(x_test)),6))

out=pd.DataFrame({'Price_actual':y_test,'Price_pred':y_pred})
result=df.merge(out,left_index=True,right_index=True)
result[['Price_actual','Price_pred']].sample(10)
```