EXECUTIVE SUMMARY

1. INTRODUCTION

The goal is to understand the relationship between house features and how these variables affect house prices. Using more than one model, predict the price of the house using the given dataset and compare the accuracy of the models along with the drawbacks of each technique's assumptions, also recommending the final prediction model.

2. DATA DESCRIPTION

Field Name	Data Type
Transaction date	Numeric
House Age	Numeric
Distance from nearest Metro station (km)	Numeric
Number of convenience stores	Categorical (Ordinal)
latitude	Numeric
longitude	Numeric
Number of bedrooms	Categorical (Ordinal)
House size (sqft)	Numeric
House price of unit area	Numeric

3. EXPLORATORY DATA ANALYSIS

The objectives of EDA can be summarized as follow:

- Maximize insight into the data/understand the data structure.
- EDA is an approach to analyzing data using non-visual and visual techniques.
- EDA involves a thorough analysis of data to understand the current business situation.
- EDA's objective is to extract ''Gold' from the ''Data mine'' based on domain understanding.

As a first step, importing all the necessary libraries, we think that will be required to perform the EDA. Loading the data set – Loading the 'DS - Assignment Part 1 data set.xlsx 'file using pandas. For this, we will be using a read excel file.

Head of the dataset: After reading the xlsx file, the head () command gives the below output –

	Transaction date	House Age	Distance from nearest Metro station (km)	Number of convenience stores	latitude	longitude	Number of bedrooms	House size (sqft)	House price of unit area
0	2012.916667	32.0	84.87882	10	24.98298	121.54024	1	575	37.9
1	2012.916667	19.5	306.59470	9	24.98034	121.53951	2	1240	42.2
2	2013.583333	13.3	561.98450	5	24.98746	121.54391	3	1060	47.3
3	2013.500000	13.3	561.98450	5	24.98746	121.54391	2	875	54.8
4	2012.833333	5.0	390.56840	5	24.97937	121.54245	1	491	43.1

From the head table, we infer that,

- The dataset contains 9 variables such as 'House Age', 'Distance from nearest Metro station (km)', 'Number of convenience stores, 'latitude', 'longitude', 'Number of bedrooms', 'House size (sqft)', 'House price of unit area'. The variable 'Transaction date' is not useful for our analysis and it will be dropped in the future.
- There are 8 Independent variables and 1 dependent variable as 'House price of unit area'

The shape of the dataset: Output from the shape () command is –

```
The dataset has 414 rows and 8 columns.
```

info () is used to check the Information about the data and the datatypes of each respective attribute: Output from the Info () command is –

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 414 entries, 0 to 413
Data columns (total 8 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   House Age                               414 non-null    float64
 1   Distance from nearest Metro station (km)  414 non-null    float64
 2   Number of convenience stores            414 non-null    int64
 3   latitude                                414 non-null    float64
 4   longitude                               414 non-null    float64
 5   Number of bedrooms                      414 non-null    int64
 6   House size (sqft)                       414 non-null    float64
 7   House price of unit area                414 non-null    float64
dtypes: float64(6), int64(2)
memory usage: 26.0 KB
```

We infer from Info () function that the dataset has 414 instances with 8 attributes. There are 6 float types, and 2 int types, but the variables 'Number of convenience stores '& 'Number of bedrooms ' are categorical (Ordinal) in nature

Duplication check: Output from duplicated () with sum command is –

The dataset has 0 duplications. Though there is no customer ID or any unique identifier to determine whether it is true duplication or not, here we are assuming the dataset does not contain any duplicate values

Descriptive Analytics: Describe method will help us see how data is spread for the numerical values, also we can see the minimum value, mean values, different percentile values, and maximum values.

Output from Describe with Transpose option is –

	count	mean	std	min	25%	50%	75%	max
House Age	414.0	17.71	11.39	0.00	9.02	16.10	28.15	43.80
Distance from nearest Metro station (km)	414.0	974.55	968.92	23.38	289.32	492.23	1454.28	3201.71
Number of convenience stores	414.0	4.09	2.95	0.00	1.00	4.00	6.00	10.00
latitude	414.0	24.97	0.01	24.94	24.96	24.97	24.98	25.00
longitude	414.0	121.53	0.01	121.51	121.53	121.54	121.54	121.57
Number of bedrooms	414.0	1.99	0.82	1.00	1.00	2.00	3.00	3.00
House size (sqft)	414.0	931.48	348.91	402.00	548.00	975.00	1234.75	1500.00
House price of unit area	414.0	37.86	13.11	7.60	27.70	38.45	46.60	74.95

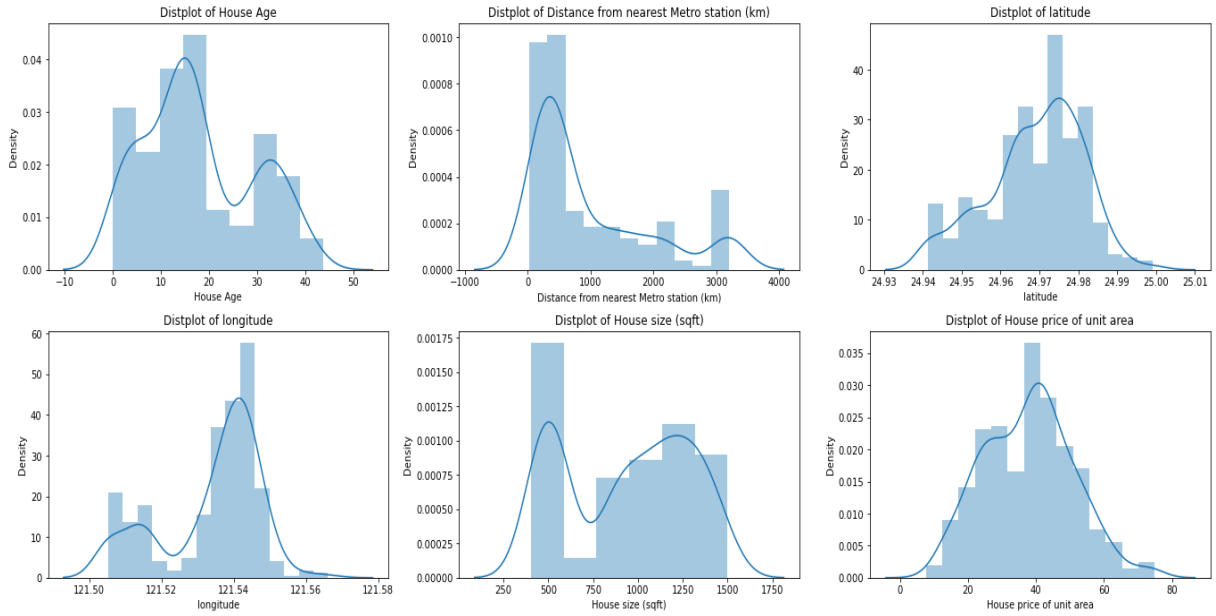
Key Observation –

- The mean and median values are not far apart, this indicates less skewness
- By looking at the dataset, it appears that there are outliers in the variables. The same is visible from the distribution of 5 values (min, 25 percentile, 50 percentile, 75 percentile, and maximum)
- There seem to be no bad /anomalies present in the data

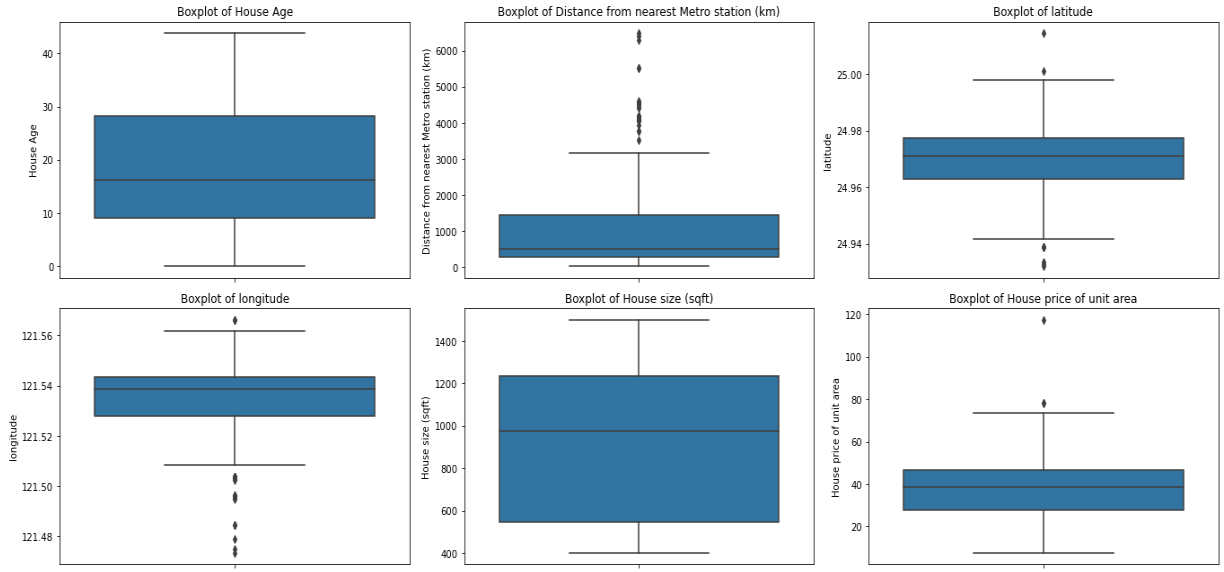
4. UNIVARIANT ANALYSIS

The objective of the univariant analysis is to derive the data, define, analyze and summarize the pattern present in it. A dataset explores each variable separately, such as the numerical and categorical variables. Some of the patterns that can be easily identified with univariant analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation. Univariant analysis can be described and visualized with the help of the most used plots of Histogram/Distplot and Boxplot.

Continuous Variable – Histogram



Continuous Variable – Boxplot

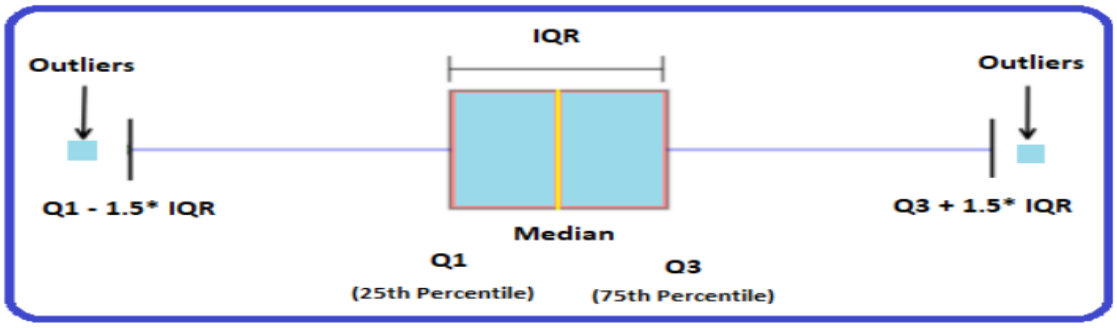


Checking the Skewness, Kurtosis & Outlier

	Skewness		kurtosis		outlier
Distance from nearest Metro station (km)	1.22	**Distance from nearest Metro station (km)**	0.22	**Distance from nearest Metro station (km)**	37
House Age	0.38	**latitude**	-0.25	**longitude**	35
House price of unit area	0.18	**House price of unit area**	-0.28	**latitude**	8
House size (sqft)	-0.11	**longitude**	-0.34	**House price of unit area**	3
latitude	-0.45	**House Age**	-0.88	**House Age**	0
longitude	-0.82	**House size (sqft)**	-1.42	**House size (sqft)**	0

Key Observation –

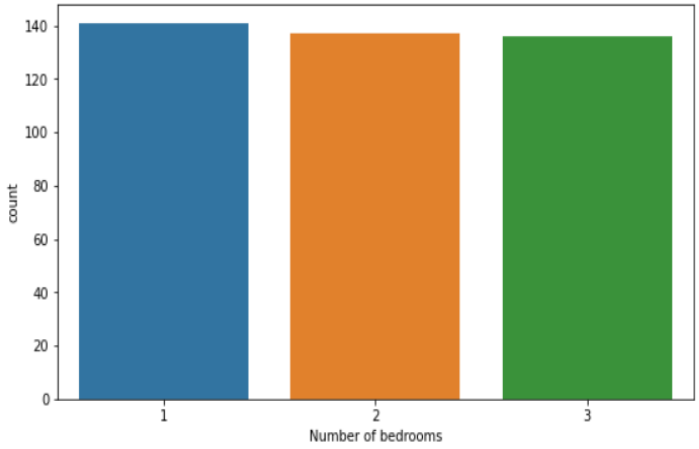
- Except for variable 'House Age' & House size', all the continuous variables have outliers. This means all variables have values that are out of the range of (Q1 – 1.5* IQR) to (Q3 +1.5 * IQR) as shown below. However, as there is no value that+3 seems erroneous, we will not remove these values



- All the continuous variables have outliers. This means all variables have values that are out of the range of (Q1 – 1.5* IQR) to (Q3 +1.5 * IQR) as shown below. However, as no value seems erroneous, we will not remove these values
- Variable 'Distance from nearest metro station (km)' is slightly right skewed and all other variables almost follow Normal distribution

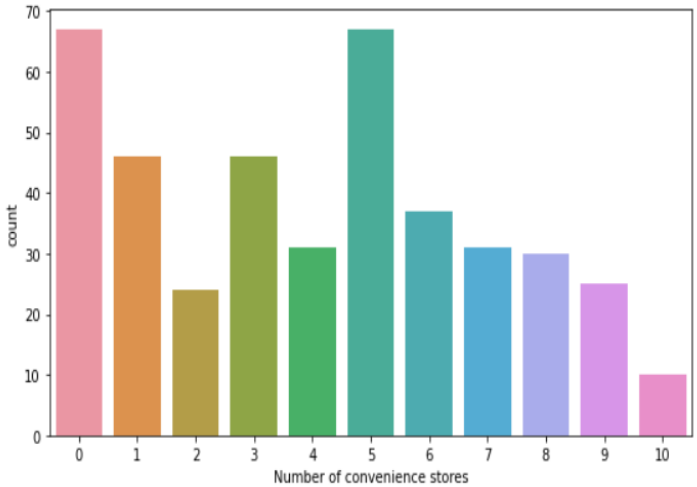
Categorical Variable – Countplot

Variable - Number of bedrooms




```
1    141
2    137
3    136
Name: Number of bedrooms, dtype: int64
```

Variable - Number of convenience stores



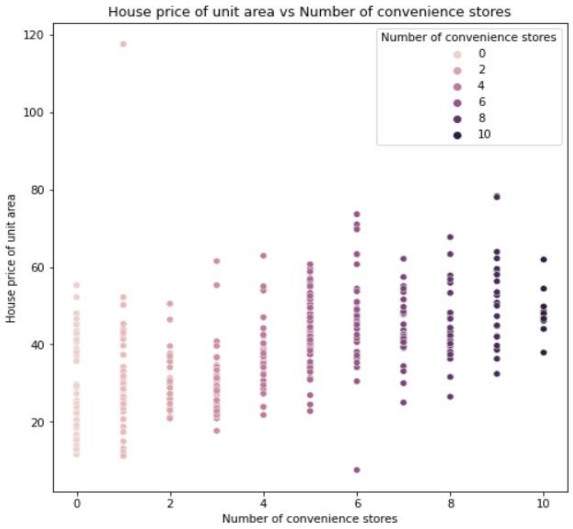

```
5     67
0     67
3     46
1     46
6     37
7     31
4     31
8     30
9     25
2     24
10    10
Name: Number of convenience stores, dtype: int64
```

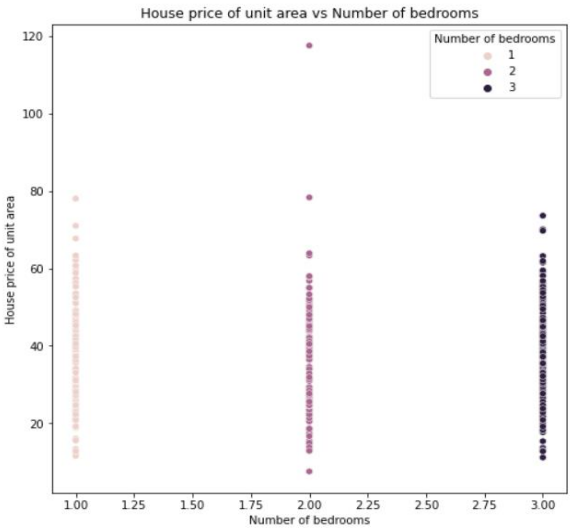
Key Observation –

- **Number of bedrooms:** There is no drastic difference found between rooms 1, 2 & 3
- **Number of convenience stores:** 67% of houses do not have nearby convenience stores

5. BIVARIANT ANALYSIS

Strip plot - Target (House price of unit area) and Categorical variables

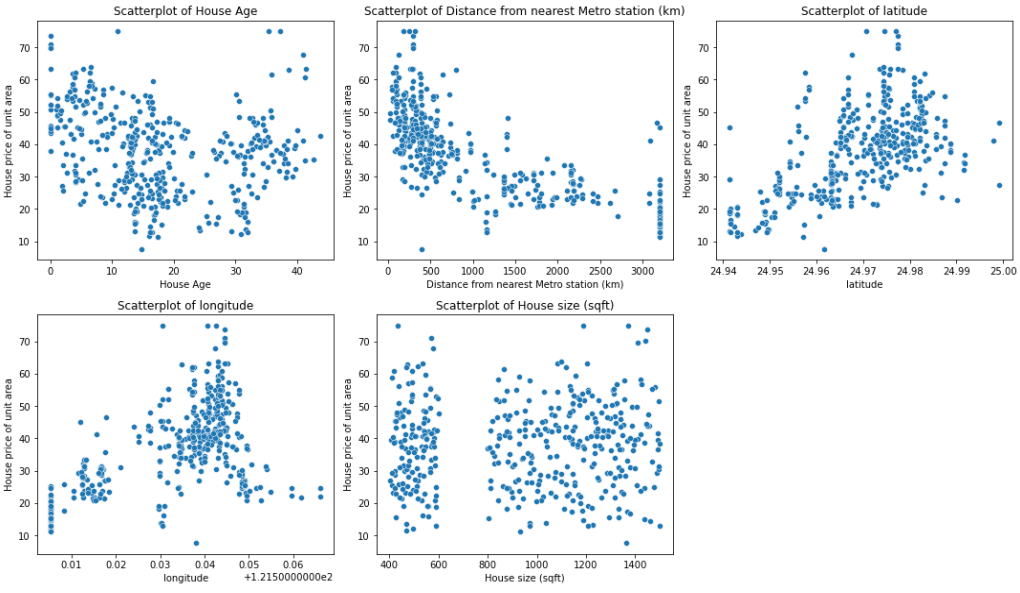




Key Observation –

- Variable Number of convenience stores vs House price of the unit area shows positive relation when the Number of convenience stores increases the House price of the unit area also increased
- Variable Number of bedrooms vs House price of unit area shows does not show any relationship
- Among the categorical variable, the number of convenience stores is showing a strong predictor as compared to others and the strip plot also suggests the same

Strip plot - Target (House price of unit area) and continuous variables

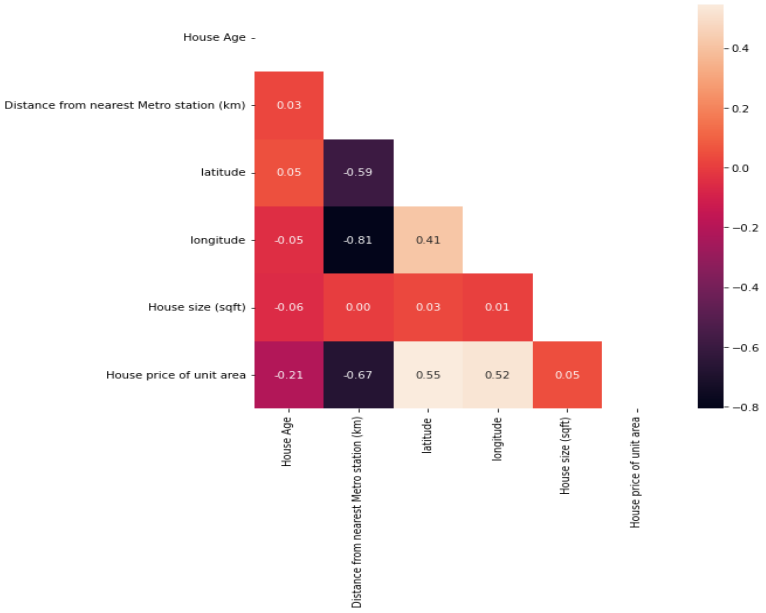


Key Observation –

- Variable Distance from nearest Metro station (km) vs House price of the unit area show some decreasing pattern, this indicates a negative correlation
- Variable latitude & longitude vs House price of a unit area shows some positive correlation when latitude increases the House price of a unit area slightly gets increased
- Variable House Age & House size (sqft) longitude vs House price seems no relationship and all the data points are scattered as cloud

- Among the categorical variable, Distance from the nearest Metro station (km), Latitude & Longitude is showing slightly good predictors as compared to others and the strip plot also suggests the same

HeatMap- Target (House price of unit area) and continuous variables



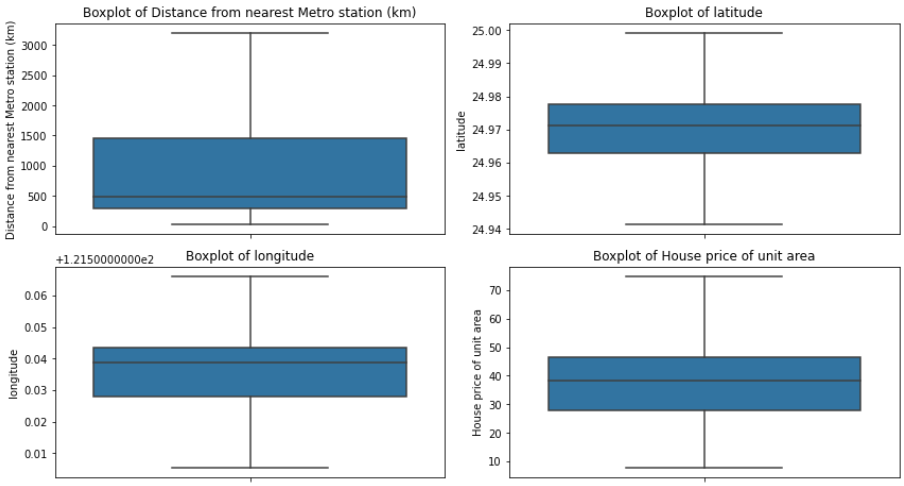
Key Observation:

Here we see high multicollinearity between the independent variable and only variable 'Latitude', 'Longitude' & 'House_size' is correlated to the dependent variable

6. OUTLIER TREATMENT

The presence of outliers in the dataset may affect the output during Clustering. That is because each centroid is a mean, which is a measure of central tendency whose value is affected by extreme values.

- As per univariant analysis (Boxplot), we deducted the presence of outliers in all the variables
- Using IQR Capping method, Imputing the Outlier values by replacing the observations outside the lower limit with the value of the 25th percentile; and the observations that lie above the upper limit, with the value of the 75th percentile of the same dataset



The below boxplot shows that after IQR imputation no outliers got deducted across the variables

6. MODEL DEVELOPMENT

Linear Regression Model: -

Linear regression is a supervised Machine Learning model and a way to identify a relationship between two or more variables. We use this relationship to predict the values for one variable for a given set of value(s) of the other variable(s). The variable, which is used in prediction is termed as independent/explanatory/regress or variable whereas the predicted variable is termed as dependent/target/response/regress and variable. Linear regression assumes that the dependent variable is linearly related to the estimated parameter(s).

The main goal of the Linear Regression model is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized

Creating multiple models and checking their performance of Predictions on Train and Test sets using R-square, RMSE & Adj Square: -

Here we build a Multiple Linear Regression model and check their model performance metrics, at the end we will compare the created models and select the best fit model to create a final linear equation.

In machine learning and Multiple Linear Regression literature the above equation is used in the form: -

Y= $m1x1$ + $m2x2$+ $m3x3$+ ……+ $mnxn$+C+e
y - Dependent/target/predicted variable
xi – Independent/Predictor variable
mi- Coefficients for the ith independent/Predictor variable
C – Constant/intercept/bias
e – Residual error/unexplained variance/difference between actual and prediction

y (House price of unit area) = C+ m1 * House Age + m2 * Distance from nearest Metro station (km) + m3 * Number of convenience stores+ m4 * latitude+ m5 * longitude+ m6 * Number of bedrooms+ m7 * House size (sqft)

Model 1: The first model is made using all the variables

The intercept for our model is -5277.181481260923
The coefficient for House_Age is -0.25245919920020327
The coefficient for Distance_from_nearest_Metro_station is -0.006258395337943501
The coefficient for Number_of_convenience_stores is 0.9201665828946162
The coefficient for latitude is 241.00795476786396
The coefficient for longitude is -5.725842054574471
The coefficient for Number_of_bedrooms is 2.165750016352232
The coefficient for House_size is -0.004674419443318952

Key Observations:

The above coefficients of determinants output describe among 7 independent variables,
variable 'latitude' has the most weightage and acts as a good predictor for the target variable 'price'

When 'latitude' increases by 1 unit, the 'House price of unit area' increases by 241.00 units, keeping all other predictors constant

- Variable House_Age, Distance_from_nearest_metro_station, longitude & House_size has a negative coefficient, and it acts as a weak predictor for the target variable 'House price of unit area'
- Variable 'Number_of_bedrooms' & 'Number_of_convenience_stores has some moderate weightage in predicting target variable 'House price of unit area'

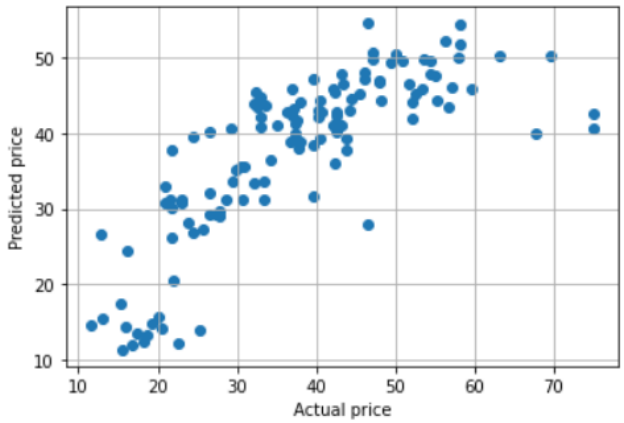
Model Performance

	Model 1
Intercept	-5277.18
R.Sq Train	0.64
R.Sq Test	0.62
RMSE Train	7.72
RMSE Test	8.43

Key Observations:

R squared value for train & test is close to 60%, thus the model has become a quite unstable post dropping the variables with high multicollinearity. but this is not a reliable metrics since it might contain some statistical fluke. To arrest this statistical fluke, we will consider the adjusted R squared

Evaluation of plot between the actual and predicted price for linear regression



From the above scatter plot, we infer that data points got slightly scattered but do not look cloudy. This indicates that our model did a reasonable prediction and has a slightly positive linear line

OLS Output

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     House_price_of_unit_area   R-squared:                       0.637
Model:                                 OLS   Adj. R-squared:                  0.628
Method:                      Least Squares   F-statistic:                     70.41
Date:                     Thu, 01 Dec 2022   Prob (F-statistic):           3.64e-58
Time:                             17:56:19   Log-Likelihood:                -1000.6
No. Observations:                      289   AIC:                             2017.
Df Residuals:                          281   BIC:                             2047.
Df Model:                                7
Covariance Type:                 nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
Intercept                          -5277.1815   6366.333     -0.829      0.408   -1.78e+04    7254.576
House_Age                             -0.2525      0.040     -6.269      0.000      -0.332      -0.173
Distance_from_nearest_Metro_station   -0.0063      0.001     -6.635      0.000      -0.008      -0.004
Number_of_convenience_stores           0.9202      0.217      4.250      0.000       0.494       1.346
latitude                             241.0080     46.257      5.210      0.000     149.955     332.061
longitude                             -5.7258     50.438     -0.114      0.910    -105.010      93.558
Number_of_bedrooms                     2.1658      0.942      2.300      0.022       0.312       4.019
House_size                            -0.0047      0.002     -2.179      0.030      -0.009      -0.000
==============================================================================
Omnibus:                       52.168   Durbin-Watson:                   2.179
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              148.270
Skew:                           0.800   Prob(JB):                     6.36e-33
Kurtosis:                       6.123   Cond. No.                     2.12e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.12e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

From the above OLS summary, we infer that -

- R-squared and adjusted R-squared values are the same which is equal to 0.63
- The overall p-value for the model is 0.00 which is less than 0.05, this indicates that there is some relation between dependent and independent variables.
- The P-Value for the feature Longitude is > 0.05 this indicates that there's strong multicollinearity among some of the independent variable

In addition, the output from the sklearn's Linear Regression & statsmodel's OLS is similar. Therefore, we will continue using sklearn's Linear Regression model for further analysis

Model 2: (Model using House_Age, Distance_from_nearest_Metro_station, Number_of_convenience_stores and House price of unit area) – Regression, Ridge & Lasso

Regression Model:

```
The intercept for our model is 45.13681784564828
The coefficient for House_Age is -0.22965580678642025
The coefficient for Distance_from_nearest_Metro_station is -0.007734221979948951
The coefficient for Number_of_convenience_stores is 1.0393791462820763
```

Ridge Model:

```
The intercept for our model is 45.13681784564828
The coefficient for House_Age is -0.22965457691536667
The coefficient for Distance_from_nearest_Metro_station is -0.007734694299160217
The coefficient for Number_of_convenience_stores is 1.03915312029989
```

Lasso Model:

```
The intercept for our model is 45.2523763503691
The coefficient for House_Age is -0.22895586
The coefficient for Distance_from_nearest_Metro_station is -0.00777804
The coefficient for Number_of_convenience_stores is 1.0184555
```

Key Observations:

The above coefficients of determinants of 3 outputs describe 5 independent variables,
In all three models, variable 'Number_of_convenience_stores' has slight weightage and acts as a good predictor for the target variable 'price'

Model Performance

	Model 2	Model 3	Model 4
Intercept	45.14	45.14	45.25
R.Sq Train	0.59	0.59	0.59
R.Sq Test	0.64	0.64	0.63
RMSE Train	8.18	8.18	8.18
RMSE Test	8.30	8.30	8.30

- More or less all three models give similar results but with fewer complex models. Complexity is a function of variables and coefficients
- As a result, we can say that compared to Model 1, Models 2,3 & 4 predicted well in terms of R.sq.

Model 3: The model is made using all the variables and fitted into Decision Tree Regressor, Random Forest Regressor, ANN Regressor

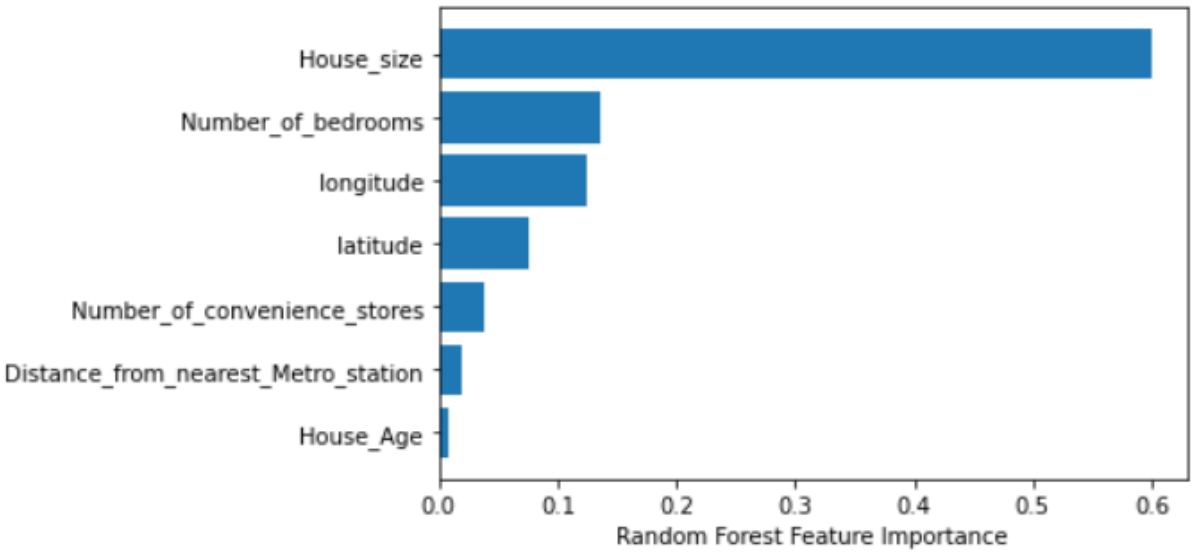
1. Fitted the model on scaled data using StandarScalar

2. Model Performance

	Train RMSE	Test RMSE	R.Sq_train	R.Sq_test
Decision Tree Regressor	0.000000	9.984280	1.000000	0.472095
Random Forest Regressor	2.592766	7.068913	0.959011	0.735377
ANN Regressor	7.247218	7.767452	0.679751	0.680494

Here we infer that the Random Forest regressor gives seems to predict high R.sq but we see a big difference in train & test this is an indication of overfitting Model

3. Feature Importance



From the above feature Importance, we can see 'House_size' is the most important feature and it has the most effective over the price house, whereas 'House' Age is the least important feature and thus has the least effect on the price of the House

Model 4: The model is made using all the variables and fitted into Tunned Decision Tree Regressor, Random Forest Regressor, ANN Regressor

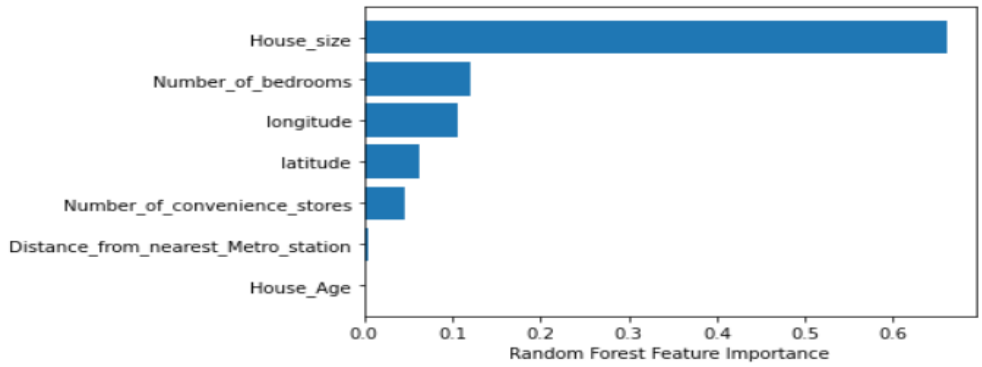
1. Fitted the Tunned Model on scaled data using StandarScalar

2. Model Performance

	Train RMSE	Test RMSE	Training Score	Test Score
Decision Tree Regressor	6.292982	7.201592	0.758533	0.725350
Random Forest Regressor	5.654434	6.879425	0.805050	0.749374
ANN Regressor	7.477800	7.689213	0.659049	0.686898

Here we infer that the Tunned Random Forest regressor gives seems to predict high R.sq. This model seems quite stable. Can't see a big difference in train and test results as compared to the model

3. Feature Importance



From the above feature Importance, we can see 'House_size' is the most important feature and it has the most effective over the price house, whereas 'House' Age is the least important feature and thus has the least effect on the price of the House

7. MODEL COMPARISON

Default Parameter

	Train RMSE	Test RMSE	R.Sq_train	R.Sq_test
Decision Tree Regressor	0.000000	9.984280	1.000000	0.472095
Random Forest Regressor	2.592766	7.068913	0.959011	0.735377
ANN Regressor	7.247218	7.767452	0.679751	0.680494

Tunned Parameter

	Train RMSE	Test RMSE	Training Score	Test Score
Decision Tree Regressor	6.292982	7.201592	0.758533	0.725350
Random Forest Regressor	5.654434	6.879425	0.805050	0.749374
ANN Regressor	7.477800	7.689213	0.659049	0.686898

Model 1: Linear Regression (All the variables)
Model 2: Linear Regression (House_Age, Distance_from_nearest_Metro_station, Number_of_convenience_stores and House price of unit area)
Model 3: Ridge (House_Age, Distance_from_nearest_Metro_station, Number_of_convenience_stores and House price of unit area)
Model 4: Lasso (House_Age, Distance_from_nearest_Metro_station, Number_of_convenience_stores and House price of unit area)

	Model 1	Model 2	Model 3	Model 4
Intercept	-5277.18	45.14	45.14	45.25
R.Sq Train	0.64	0.59	0.59	0.59
R.Sq Test	0.62	0.64	0.64	0.63
RMSE Train	7.72	8.18	8.18	8.18
RMSE Test	8.43	8.30	8.30	8.30

Key Inferences:

- Modelsel 1, 2, 3, and 4 give more or less the same R.Square and it slightly predicted poor in both training and test and leads to underfitting
- Model default parameter Decision Tree Regressor, Random Forest Regressor predicted well in train data but poor in test data in terms of RMSE & R.Square and this leads to overfitting and Model Ann Regressor predicted poor in both train and test data in terms of RMSE & R.Square leads to underfitting

- Model Tunned Parameter Ann Regressor & Decision Tree Regressor gives more or less the same in terms of RMSE & R.Square and it slightly predicted poor in both training and test and leads to underfitting
- Model Tunned Random Forest regressor gives seems to predict high R.sq. This model seems quite stable. Can't see a big difference in train and test results as compared to the model
- Here we conclude that our model doesn't perform more effectively, this is due to high multicollinearity between independent variables and useless variables, and as per the size of the rows, most of the features seem to be useless, due to the lack of feature model didn't perform great extend to predict the price. Still, we violated some of the linear model assumptions and created 10 models. As per the comparison of 10 models, we chose Tunned Random Forest model, moderately to be our best fit model since this model gives the best RMSE score & R.sq value.

Compared to 10 models, the model Tunned Random Forest regressor gives a reasonable RMSE & R.sq value

The top 5 most important features are:

['House_size', 'Number_of_bedrooms', 'longitude', 'latitude', 'Number_of_convenience_stores']