MAHARASHTRA EDUCATION SOCIETY'S

ABASAHEB GARWARE COLLEGE

Housing Prices In PUNE

SUBMITTED TO

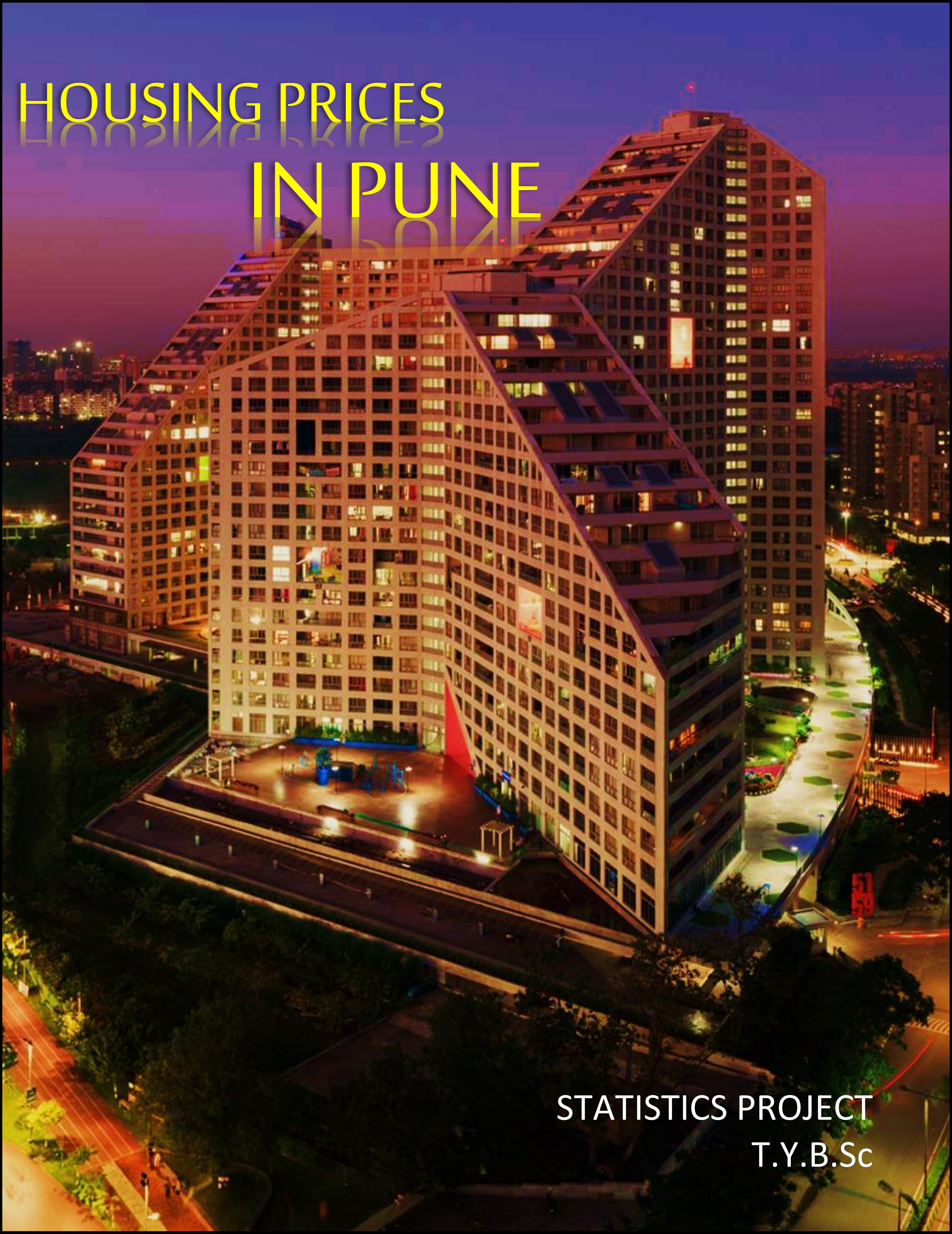
DEPARTMENT OF STATISTICS

IN THE FULLFILMENT OF

T.Y.B.Sc.

2020-2021

HOUSING PRICES IN PUNE



STATISTICS PROJECT
T.Y.B.Sc

ACKNOWLEDGEMENT

We have satisfaction upon completion of our project work **HOUSING PRICES IN PUNE** at the Department of Statistics of MES Abasaheb Garware College , Pune during the academic year 2020-21.

We express our gratitude towards **prof. Rutuja Joshi** for her valuable guidance throughout project work .

We are thankful to prof. **Sandesh Kurade** (Head, Department of Statistics) and the teaching staff of Department of Statistics for their valuable support and co-operation during completion of this project .

CERTIFICATE

This is to certify that the project report entitled HOUSING PRICES IN PUNE is being submitted by -

Gargi Sarjine (9756)

Aishwarya Bhujbal (9760)

Akriti Maurya (9764)

Lalitagauri Walimbe (9771)

Aakanksha Dixit (9774)

Tanvi Khamkar (9778)

as a partial fulfilment for the award of the degree of the Bachelor of Science (B.sc).This is a record of bonafide work carried out by them under supervision and guidance.

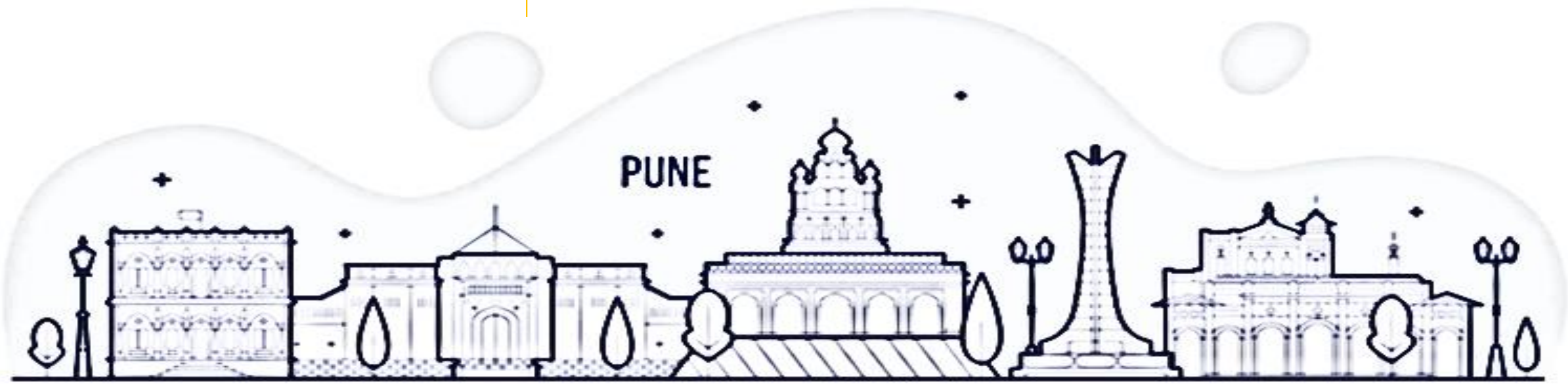
Prof .Rutuja Joshi
Project Guide

Prof.Sandesh Kurade
Head Dept .of Statistics

Place : Pune

Date: 30/4/2021

INDEX



OBJECTIVE

The objective of this project is to observe the effect of various factors like built up area, location, number of amenities etc. on house prices in Pune and to build a suitable regression model for predicting prices of other houses using multiple linear regression analysis.

NOTE – The data used for this project is publicly available on nobroker.com. It has not been misused in any case and its use is limited only for the project.

MOTIVATION

A friend of ours was interested in buying a flat in Pune. And we decided to help her out…

So we were looking for 2 bhks near the Koregaon park area and like most of the people we hired an agent and this is what we were told –

This one is for 1.6 crores and the other one is for 74 lakhs. Yes, both are 2 bhks.

We wondered why were the prices for a 2 bhk so different? What were the factors present in first house that made it more costlier? This sparked an interest in us to study what factors affect the price of a house. Why are some houses with smaller built up area costlier than others with a larger built up area? After some research we came to know that there are various factors like built up area, the location, presence or absence of certain amenities etc. all affect the price of a house.

It's quite possible there are people spending too much on a house with almost no facilities and people loosing too much by selling a house full of facilities at a very low price.

Hence we decided to study the various factors affecting the house prices at few major locations of Pune and to fit a regression model which helps to predict the proper price of a house.

ABSTRACT

The price of a house depends on various factors like its built up area, location, number of balconies, availability of parking security etc. We have considered various such factors to fit the best possible regression model.

We have collected a sample of 1200 observations. We have considered resale and newly constructed flats for this project. All of the observations have been collected from "*NoBroker.com*". We have chosen few major locations in Pune like Kothrud, Aundh, Hinjewadi, Koregaon park etc.

We have used proportional allocation to determine the number of observations to be taken from each location.

The distribution of cost is **non normal distribution** proved by Shapiro test performed in the statistical analysis. To obtain the most suitable model we have tried various transformations and also the Box Cox transformation.

KEYWORDS : Proportional allocation , Regression analysis, Multiple Linear regression, Shapiro Wilk test , Chi square test ,Kruskal Wallis test, correlation test, Box Cox transformation.

STATISTICAL DEFINITIONS

REGRESSION ANALYSIS

In statistical analysis the regression is the most powerful technique for estimating the relationship between independent variable often called regressors and dependent variables often called response variables. Using regression analysis one can fit regression model for the purpose of forecasting and prediction. The parameters of regression model can be estimated by ordinary least square method.

ASSUMPTIONS OF REGRESSION MODEL:-

- The relationship between response variables and regressor variables should be linear.
- The error term has zero mean.
- The error term has constant variance.
- The errors are uncorrelated.
- The errors are normally distributed.

LINEAR REGRESSION

Linear regression is the relationship between a scalar response and one or more explanatory variables also known as dependent and independent variables.

MULTIPLE LINEAR REGRESSION ANALYSIS

A regression model which involves more than one regressor variable is called multiple linear regression model. In general, the response Y may be related to K regressors then, model becomes,

$$Y = \beta_0 + \ \beta_1 X_1 + \ \beta_2 X_2 + \ \ldots\ldots + \beta_K X_K + \ \epsilon$$

where, β_j is the j^{th} regressors's coefficient.

<u>Assumptions:</u>

- **Linear Relationship**: There must be linear relationship between the outcome variable and the independent variables.
- **Multivariate Normality**: Multiple regression assumes that the residuals are normally distributed.
- **No Multicollinearity**: Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.
- **Homoscedasticity**: This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

STRATIFIED SAMPLING

Stratified sampling is a type of sampling method in which the total population is divided into smaller groups known as strata. Stratified sampling provides better coverage of the population since all the subgroups are represented in the sampling.

PROPORTIONAL ALLOCATION

Proportional allocation is a procedure for dividing a sample among the strata in a stratified sample survey.

Proportional allocation sets the sample size in each stratum equal to be proportional to the number of sampling units in that stratum.

i.e

$$\mathrm{ni} = \mathrm{Wi}{*}\mathrm{n}$$

where, Wi = Ni/N

N = population size

Ni = No. of units in the i^{th} stratum, i = 1,2,....,k

n = size of the stratified sample

ni = size of sample selected from i^{th} stratum, i = 1,2,....,k

Proportional allocation is useful if precise estimates are desired for the larger strata in the population, as large sample sizes are allocated to the large strata.

DUMMY VARIABLES

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in the study. Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups.

STEPWISE REGRESSION

Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, significance of previously added regressors is checked. If a variable is found insignificant, it is removed from the model. The model continues till an appropriate model is developed.

CHI SQUARE TEST

A chi square (χ^2) statistic is a test that measures how a model compares to actual observed data.

$$\chi^2 = \sum \frac{(\boldsymbol{Oi} - \boldsymbol{Ei})^2}{\boldsymbol{Ei}}$$

χ^2 = chi squared

Oi = observed value

Ei = expected value

CRAMER'S V TEST

Cramer's V is a measure of association between two nominal variables, giving a value between 0 and +1. It is based on Pearson’s chi-squared statistic.

Test statistic:

$$\boldsymbol{V} = \sqrt{\frac{\boldsymbol{\chi^2}}{\boldsymbol{n}*(\boldsymbol{q}-\boldsymbol{1})}}$$

Where q is min (row or column)

Testing criteria:

Value of V	**Interpretation**
V>0.25	Very strong
V>0.15	Strong
V>0.10	Moderate
V>0.05	Weak
V>0	No or very weak

CORRELATION TEST

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.

A high correlation means that two or more variables have a strong relationship with each other, while a weak correlation means that the variables are hardly related.

CORRELATION MATRIX

A simple measure of multicollinearity is inspection of off-diagonal elements r_{ij}; $|r_{ij}|$>0.9 indicates multicollinearity problem. It is helpful in detecting linear dependence between pairs of regressors.

EIGENSYSTEM ANALYSIS

Multicollinearity can also be detected from the eigenvalue of the correlation matrix of (X'X).

$$\boldsymbol{k = \frac{\lambda_{max}}{\lambda_{min}}}$$

where λ_k are the eigen values.

Value of k	**Multicollinearity**
k < 100	indicates no serious problem with multicollinearity
100 ≤ k ≤ 1000	Indicates moderate to strong multicollinearity
K > 1000	Indicates severe problem with multicollinearity

VARIANCE INFLATION FACTORS (VIF)

VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable.

In general, VIF ≥ 0.5 indicates multicollinearity.

BOXCOX METHOD

Box-cox method is used to transform response variable to correct non normality or non-constant variance which is one of the assumptions of linear regression model.

The useful class of transformation is called as power transformation y^{λ} where $\lambda$the parameter to be determined.

The best procedure to estimate y^{λ} is given below:

$$y^{\lambda} = \frac{y^{\lambda}-1}{\lambda}, \quad \text{if} \quad \lambda \neq 0$$

$$= \ln y \quad , \quad \text{if} \quad \lambda = 0$$

KRUSKAL WALLIS H TEST

In situations where the normality assumption of errors is not justified or failed the Kruskal Wallis H test is an alternative procedure to the F test used in analysis of variance.

The Kruskal Wallis H test is use for testing the equality of treatment means. The test procedure is given below

Ho = The treatments mean does not differ significantly.

vs

$H1$ = The treatments mean differ significantly.

Test statistic:

$$H=\frac{1}{s^2}\left[\sum_{i=1}^{t}\frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right]$$

Where ni is number of observations for ith treatment.
N = total number of observations

If there **are ties** in the observations

$$s^2=\frac{1}{(N-1)}\left[\sum_{i=1}^{t}\sum_{j=1}^{r_i} R_{ij}^2 - \frac{N(N+1)^2}{4}\right]$$

If there are **no ties** in the observations

$$s^2=\frac{N(N+1)}{12}$$

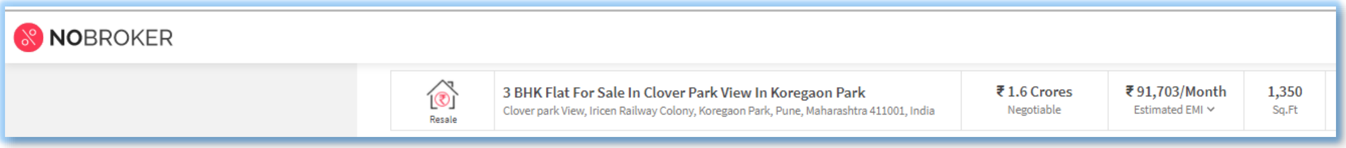
$$H=\frac{12}{N(N+1)}\left[\sum_{i=1}^{t}\frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right]$$

Decision criteria:

If $\geq S(t-1),\alpha$ then **reject H_0** accept otherwise. Where α is level of significance.

METHODOLOGY

Our project contains house sale prices and its features. This project uses the method of secondary data collection. This dataset consists of 31 house features and 1200 houses with their prices. We have collected the data from ***NoBroker.com*** site. NoBroker.com site gives us below information:



NOBROKER

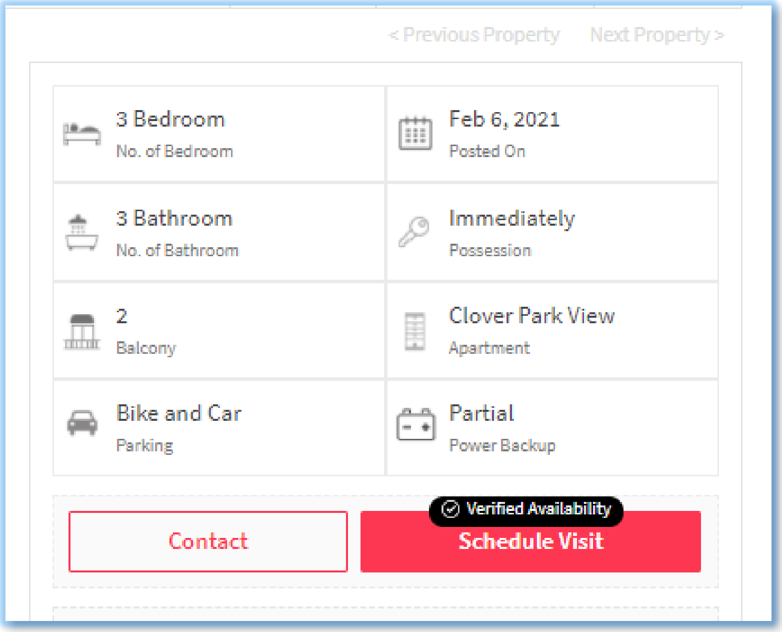
Resale

3 BHK Flat For Sale In Clover Park View In Koregaon Park
Clover park View, Iricen Railway Colony, Koregaon Park, Pune, Maharashtra 411001, India

₹ 1.6 Crores
Negotiable

₹ 91,703/Month
Estimated EMI

1,350
Sq.Ft



< Previous Property Next Property >

3 Bedroom No. of Bedroom	Feb 6, 2021 Posted On
3 Bathroom No. of Bathroom	Immediately Possession
2 Balcony	Clover Park View Apartment
Bike and Car Parking	Partial Power Backup

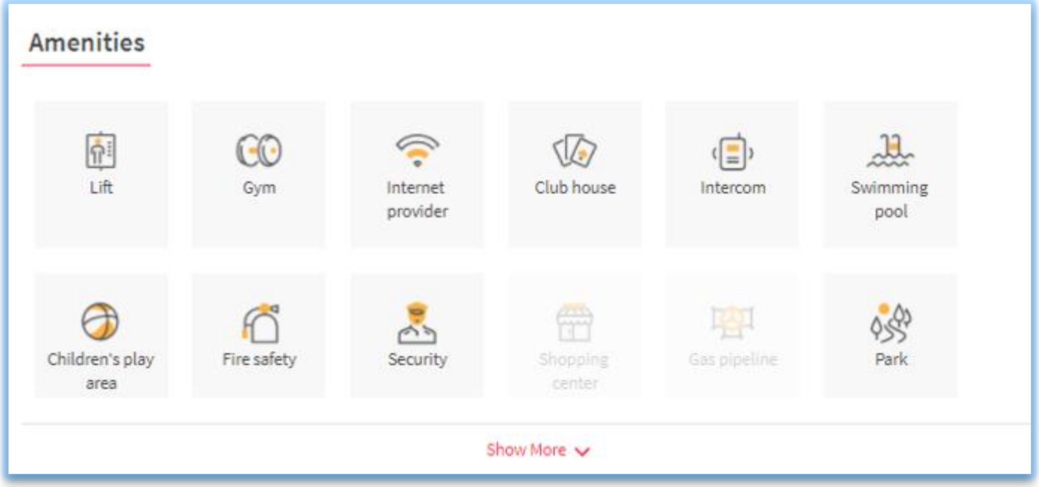
Contact

Verified Availability

Schedule Visit

These images give us the information about cost, building area, power backup, availability of parking, no. of bedrooms, bathrooms and balconies.

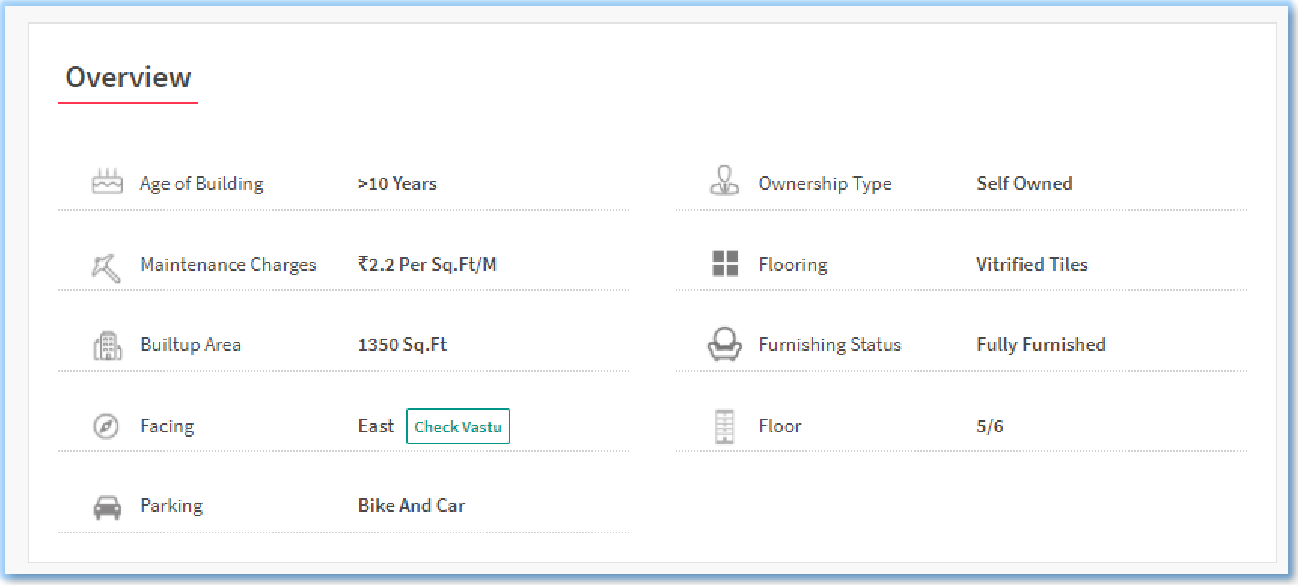
From this image we came to know about the presence/absence of various amenities.



Amenities

Lift
Gym
Internet provider
Club house
Intercom
Swimming pool
Children's play area
Fire safety
Security
Shopping center
Gas pipeline
Park

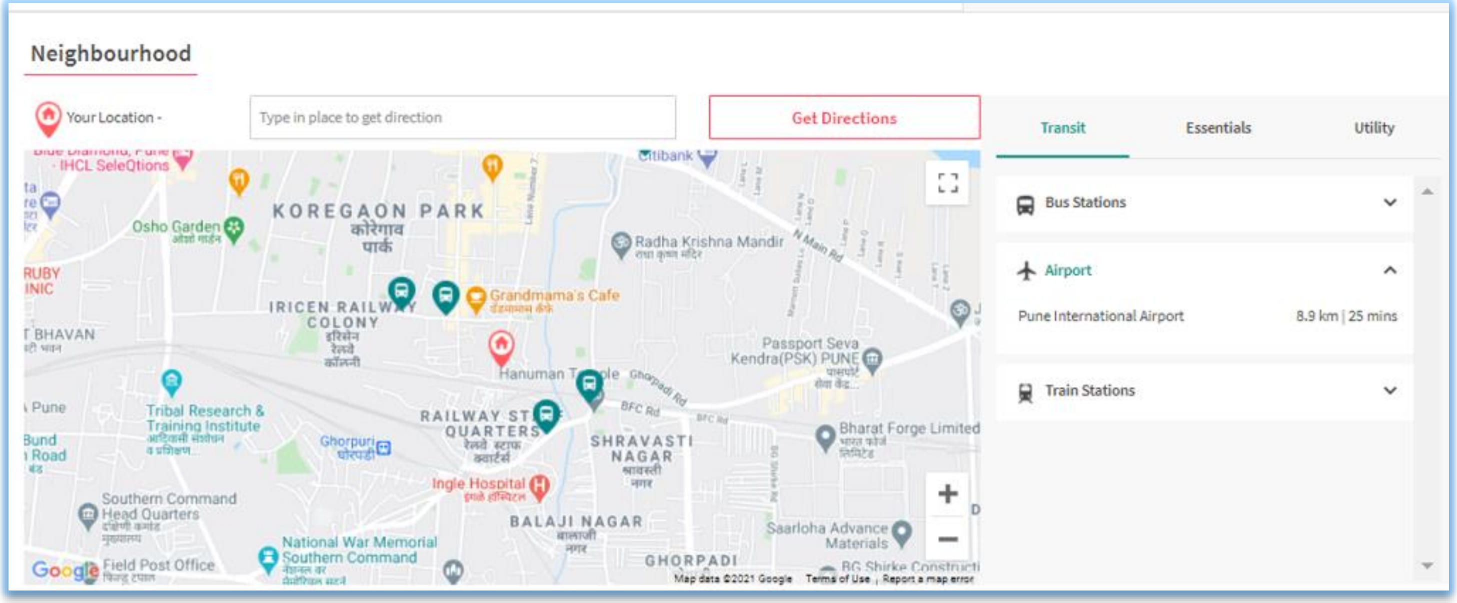
Show More



Overview

Age of Building	>10 Years	Ownership Type	Self Owned
Maintenance Charges	₹2.2 Per Sq.Ft/M	Flooring	Vitrified Tiles
Builtup Area	1350 Sq.Ft	Furnishing Status	Fully Furnished
Facing	East Check Vastu	Floor	5/6
Parking	Bike And Car		

We get information about age of the property, parking, floor, furnishing status and flooring from above image.



Neighbourhood

Your Location -

Type in place to get direction

Get Directions

Transit Essentials Utility

Bus Stations

Airport

Pune International Airport	8.9 km | 25 mins

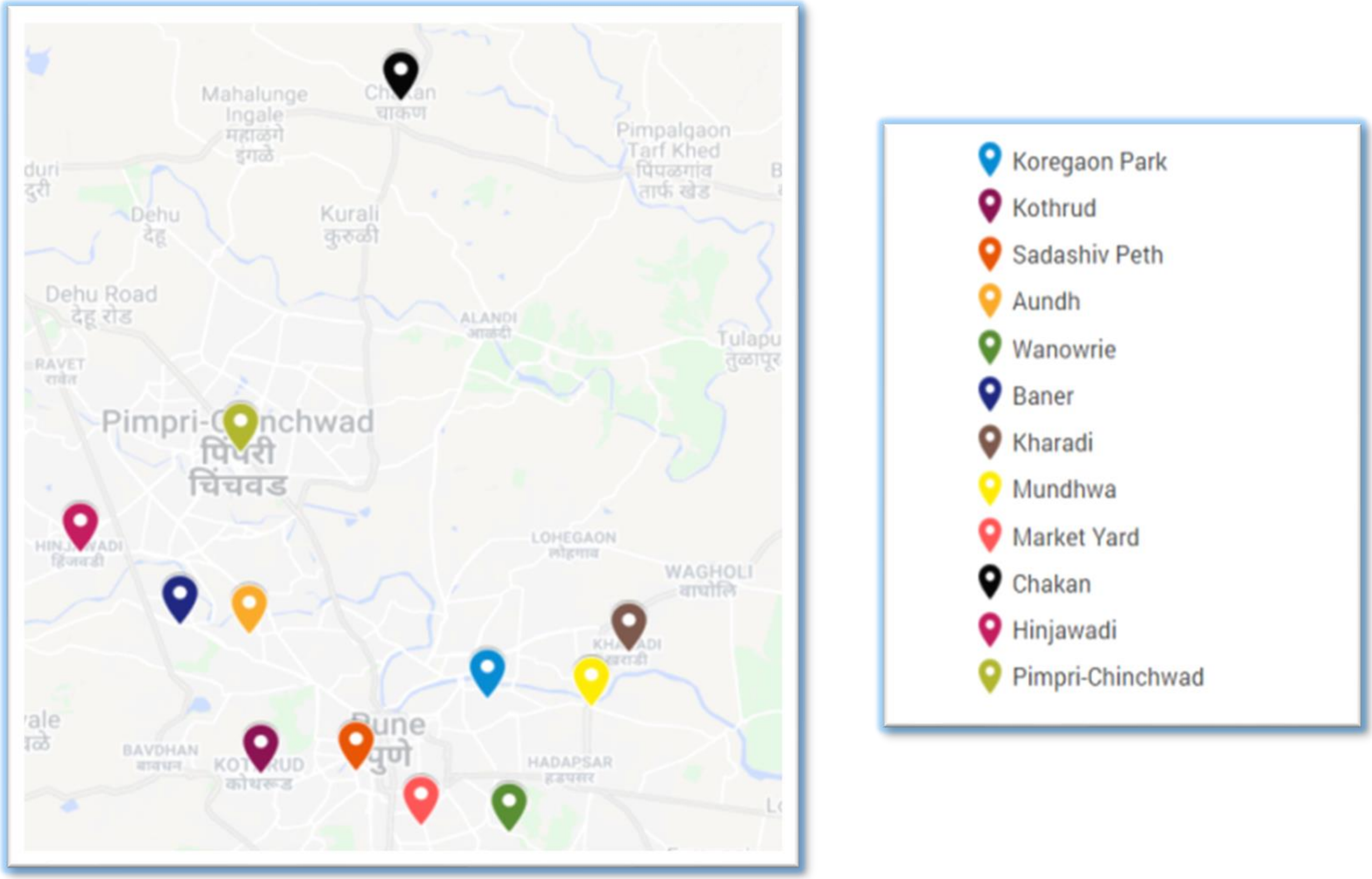
Train Stations

From this we could find distance of any desired location from the property.

Note : The data was collected between 27 – 31st march 2021.

For this project we have taken some residential, commercial and industrial areas in Pune. These areas are as follows:

Residential	Commercial	Industrial
Kothrud	Baner	Chakan
Sadashiv Peth	Kharadi	Hinjewadi P1
Aundh	Mundhwa	Pimpri Chinchwad Municipal Corporation(PCMC(MIDC))
Wanowrie	Market yard	
Koregaon park(Kp)		



DATA INFORMATION

Some factors and their explanations:

Factors	Explanation
Cost	Total cost of flats (in lakhs).
Locations	Koregaon Park (Kp), Aundh, Kharadi, Wanowrie, Chakan, Sadashiv Peth, Pimpri Chinchwad Municipal Corporation (PCMC), Mundhwa, Market yard, Kothrud, Baner, Hinjewadi Phase1.
Build Up Area	Total area measure on outer line of flat.
Age of Property	Effective age refers to state of the property.
Flooring	Vitrified, Mosaic, cement, Wooden, Marble/Granite.
Airport Distance	Distance of Pune International Airport from property.
Railway Station Distance	Distance of Pune Railway Station from property.
Bus Stand Distance	Distance of Swargate Bus Stand from property.
Amenities	Parking, lift, fire safe, gas pipeline, club house, pool, gym, power backup, children’s play area, park, sewage treatment plant, intercom, internet provider, shopping center, security, visitor parking.

We coded certain variables. The codes are listed below:

Factors	Codes
Furnishing	0 – Unfurnished, 1 – Fully furnished, 2 – Semi furnished.
Amenities	0 – Absence of amenity, 1- Presence of amenity.
Power Backup	0 – No backup, 1 – Full backup, 2 – Partial backup.
Parking	0 – No parking, 1 – Parking for bike, 2 – Parking for car, 3 – Parking for both bike and car.

DATA

cost(in lakhs)	Location	Build area (sq ft)	age of property	furnishing(yes/no/semi)	.	.	.	.	visitor parking
160	Kp	1340	>10	2	.	.	.	.	0
90	Kp	500	>10	1	.	.	.	.	0
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
67	Kharadi	878	5-10	2	.	.	.	.	1
48	Kharadi	600	5-10	2	.	.	.	.	1
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
110	Wanowrie	1385	5-10	2	.	.	.	.	1
110	Wanowrie	1680	5-10	1	.	.	.	.	0
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
35	Chakan	800	0	0	.	.	.	.	1
22	Chakan	475	3-5	0	.	.	.	.	1
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
45	Hinjewadi P1	712	>10	0	.	.	.	.	1
95	Hinjewadi P1	1648	5-10	0	.	.	.	.	1

EXPLORATORY ANALYSIS

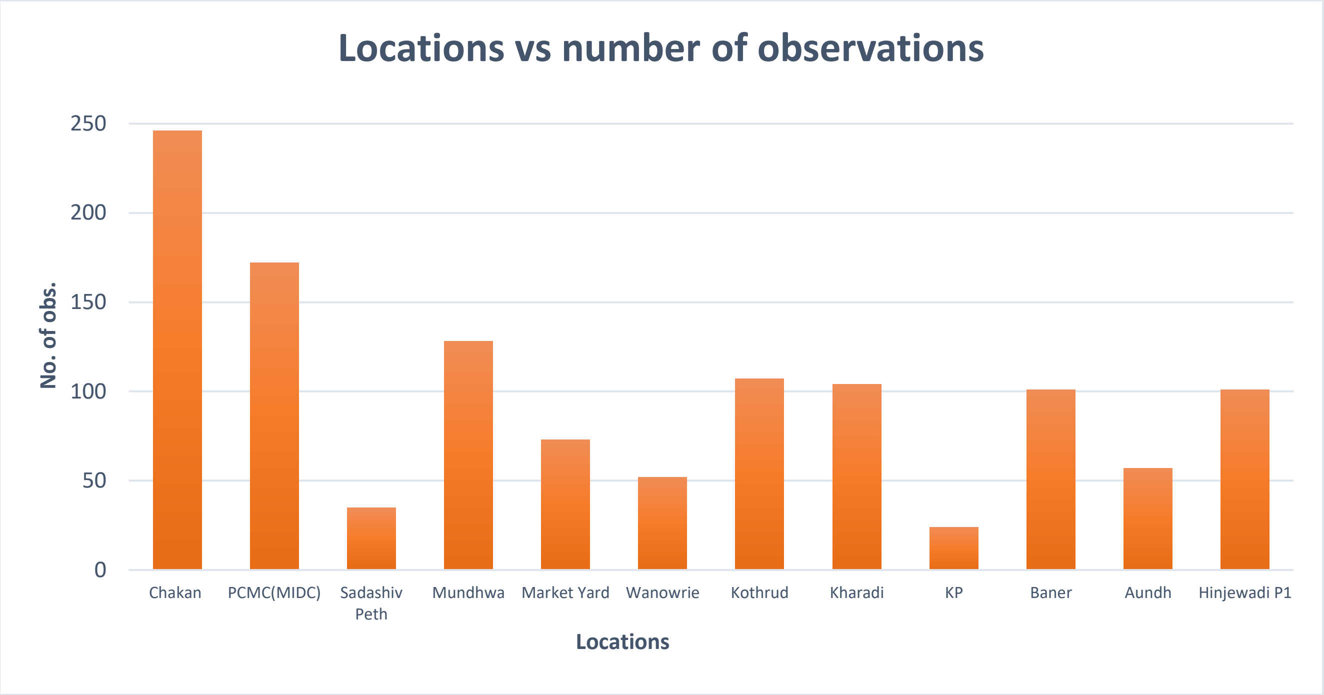
DESCRIPTIVE STATISTICS

	Cost	Build area(sq ft)	Dist airport	Dist railway	Dist swargate
Mean	70.63808	958.2925	18.1454167	15.996	17.14741667
Standard Error	1.5729	14.22026702	0.20783642	0.289477592	0.30159364
Median	58	879	18.4	12.2	14.2
Mode	55	1000	28.8	34.6	36
Standard Deviation	54.48684	492.6044995	7.19966492	10.02779793	10.44751017
Sample Variance	2968.816	242659.1929	51.835175	100.5567313	109.1504687
Kurtosis	28.0359	12.47102312	-0.9248221	-0.492942864	-0.53060713
Skewness	3.44571	2.592698779	0.10966526	0.912528323	0.71165069
Range	772	4833	33.8	41.2	44
Minimum	8	167	3.9	2.3	0.9
Maximum	780	5000	37.7	43.5	44.9
Sum	84765.69	1149951	21774.5	19195.2	20576.9
Count	1200	1200	1200	1200	1200
First quartile	34.625	612	12.6	8.8	9.7
Third quartile	88	1144.25	24.25	19.6	20.4

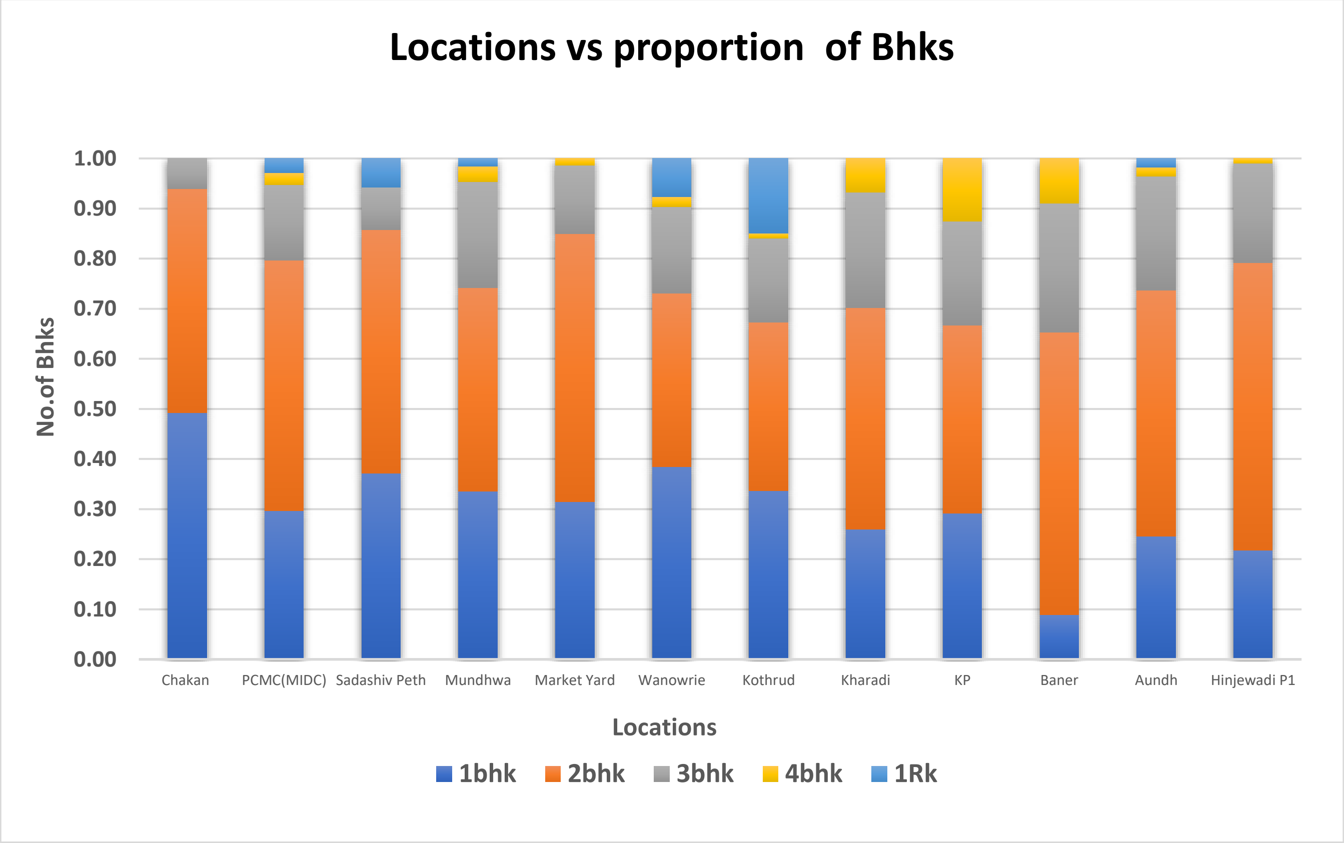
Interpretation:

The average cost of flats in our sample is approximately 71 lakhs having average Build area 958 sq ft. The average distance of Pune Airport, Pune railway station and Swargate bus stand from the flats in the sample is approximately 18 km, 16 km and 17 km respectively. The distribution of Cost prices and build area of flats in the sample is leptokurtic while that of distances from Pune Airport, Pune railway station and Swargate bus stand is platykurtic.

BAR AND PIE CHARTS

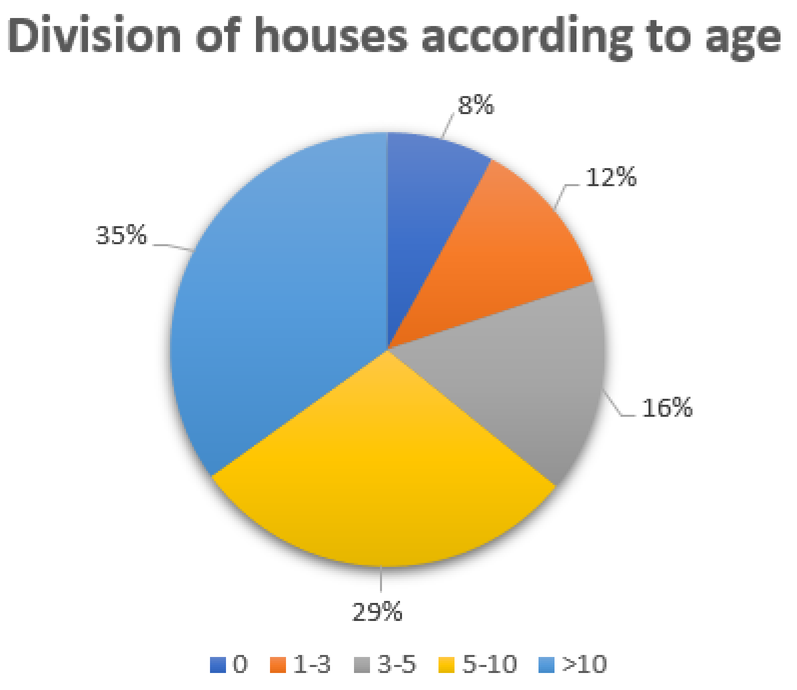


Interpretation – In our sample most observations are from Chakan and least observations are from KP i.e. Koregaon Park. Equal number of observations have been taken from Baner and Hinjewadi P1.



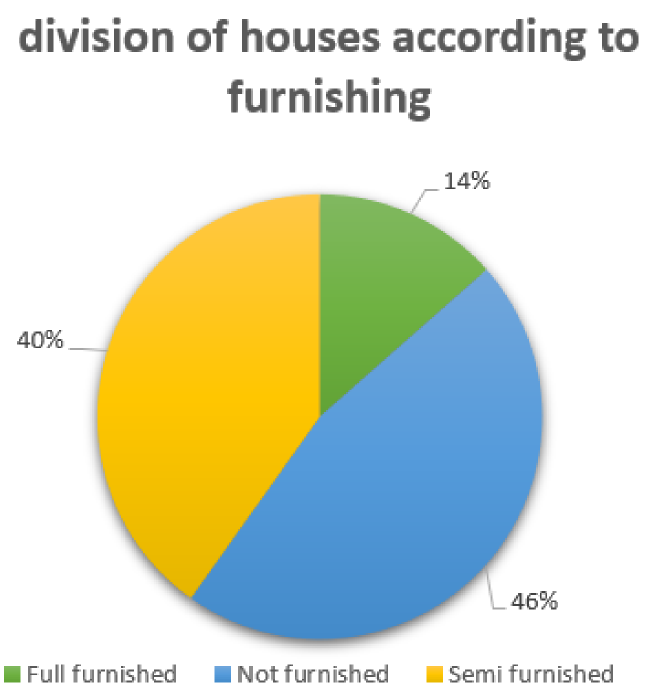
Interpretation – In our sample -

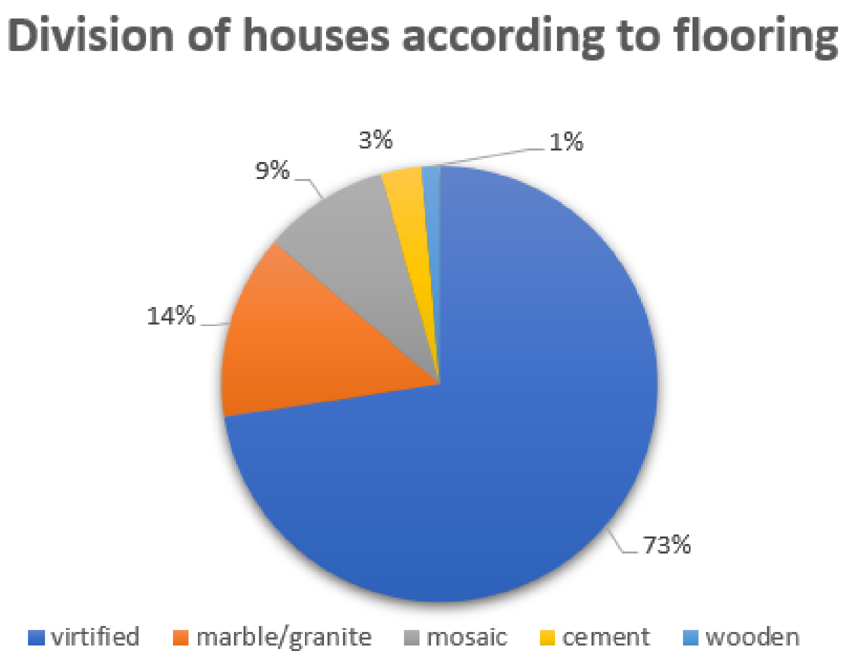
1) There are no observations in KP , Baner , Kharadi , Hinjewadi P1 , Market Yard and Chakan for 1RK .
2) Baner has the least number of 1 BHKs and Chakan has highest.
3) There are no 4Bhk flats in Sadashiv Peth and Chakan.



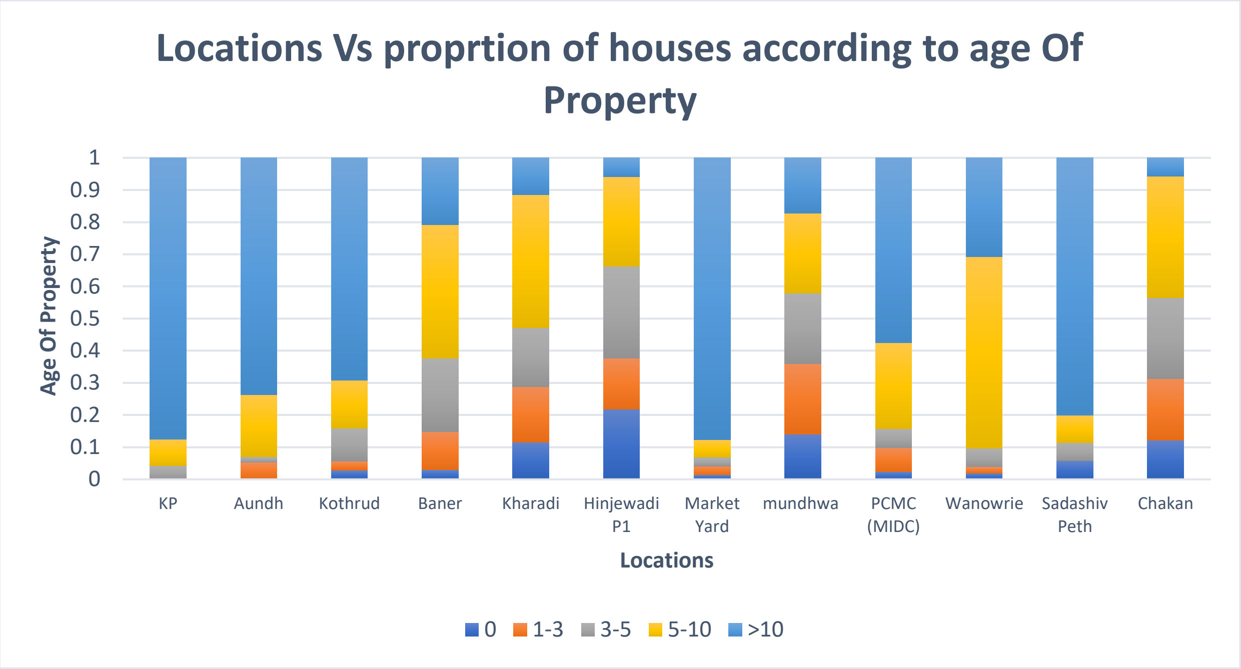
Interpretation – In our sample most of the houses (35%) have age more than 10 years and only a few (8%) are newly constructed.

Interpretation – In our sample 46% of the houses are not furnished and only some (14%) are fully furnished. There are 40% houses that are semi furnished.



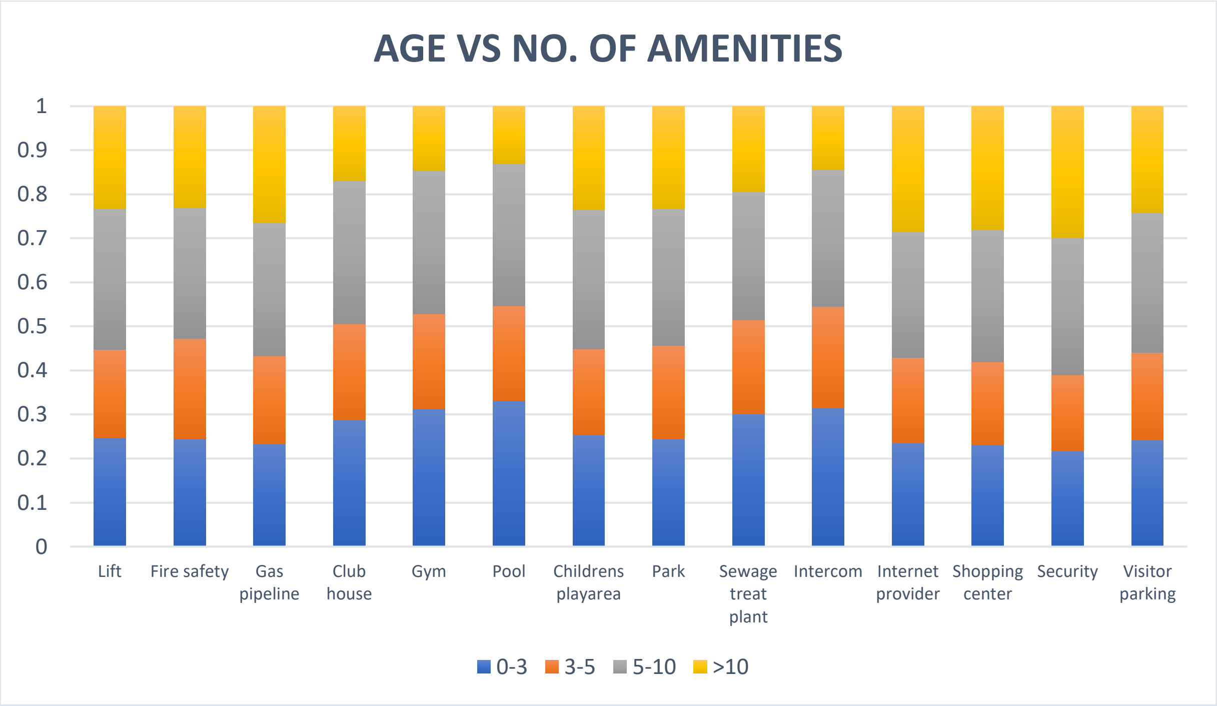


Interpretation – In our sample most of the houses (73%) in the sample have vitrified tiles and very few have wooden flooring.



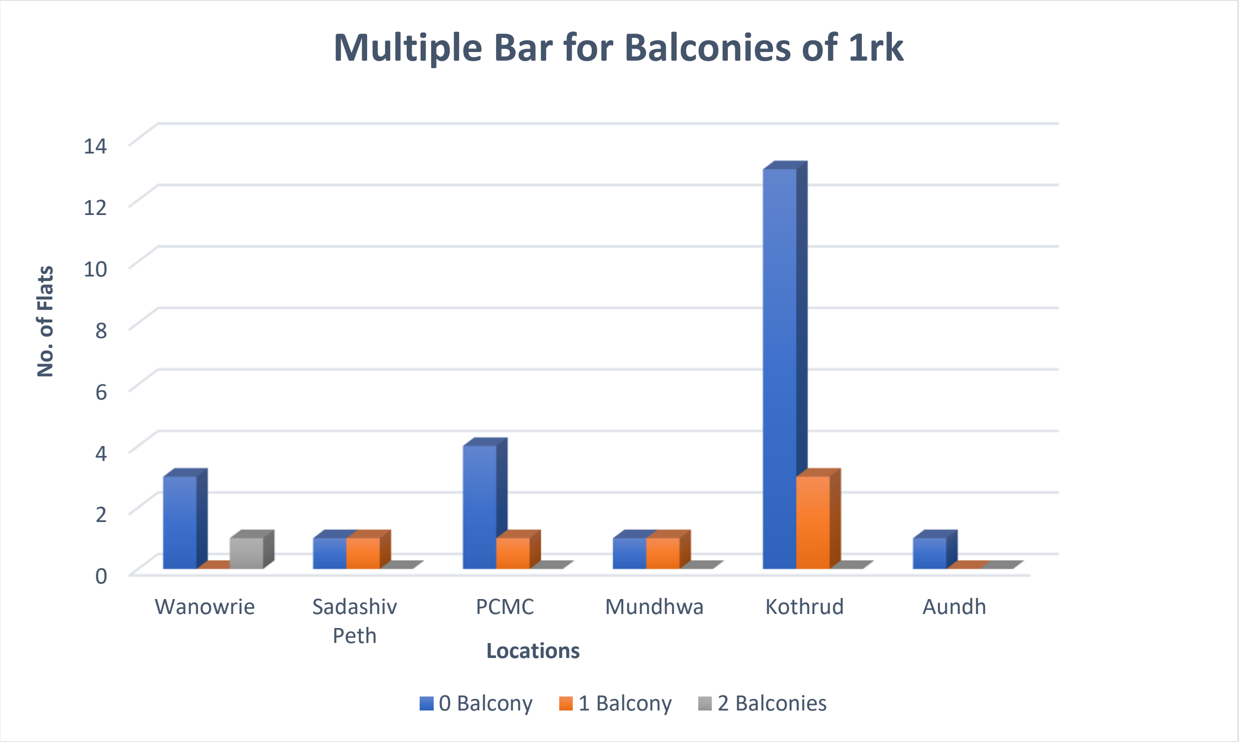
Interpretation – In our sample –

1) In most of the locations we observe high proportion of houses aged more than 5 years.
2) Koregaon Park and Market Yard have the highest proportion of houses aged more than 10 years followed by Sadashiv Peth.
3) Hinjewadi P1 and Mundhwa have the highest proportion of newly constructed flats (aged 0 – 3 years).
4) Wanowrie has the highest proportion of houses having age between 5 – 10 years and Market yard .



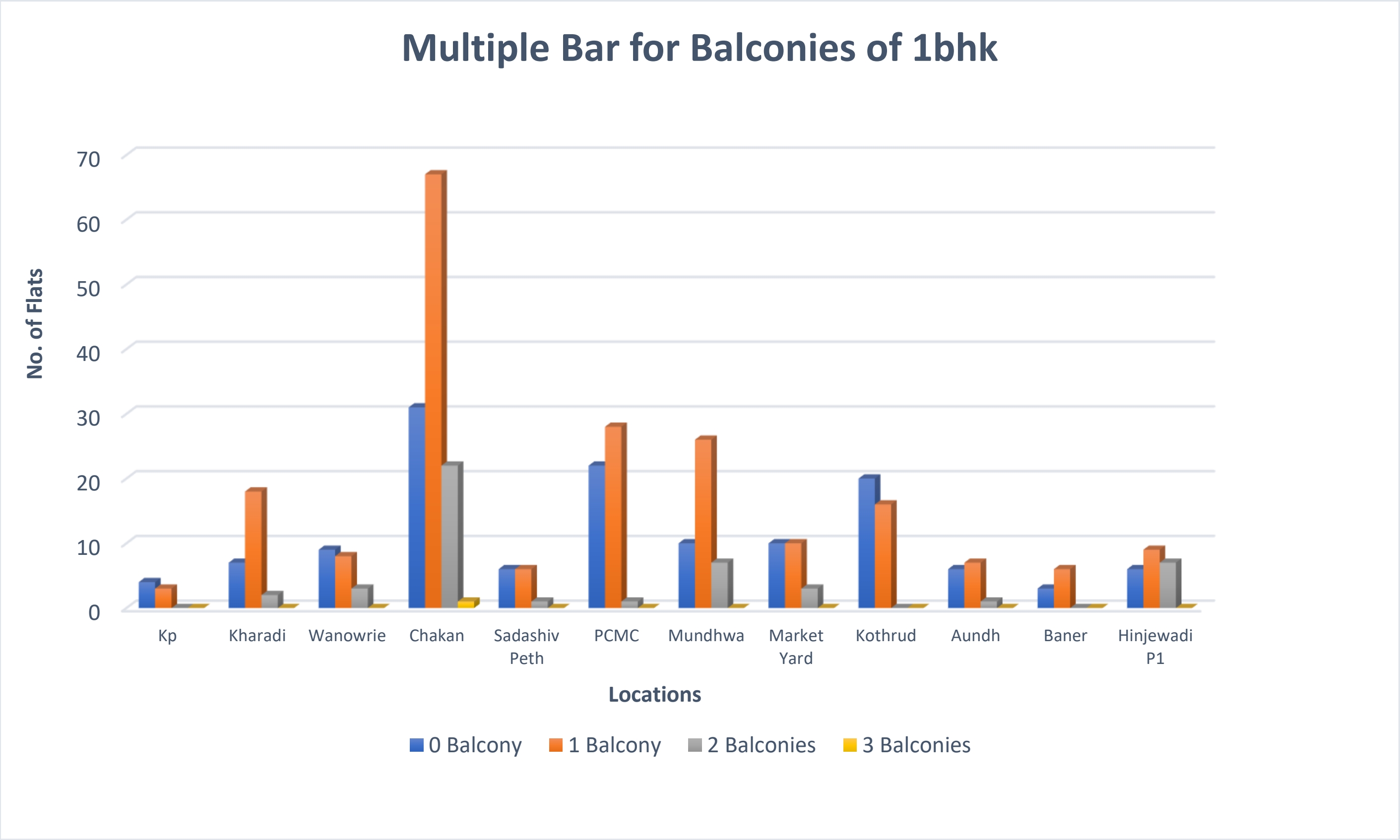
Interpretation – In our sample –

1) High proportion of houses between 5-10 years and very low of houses aged >10 years have new amenities like gas pipeline, clubhouse, gym, swimming pool, intercom, internet provider, park, sewage treat plant, visitor parking and shopping center.
2) The presence of common amenities like lift, fire safety and security is slightly more in proportion at houses between 5-10 years.



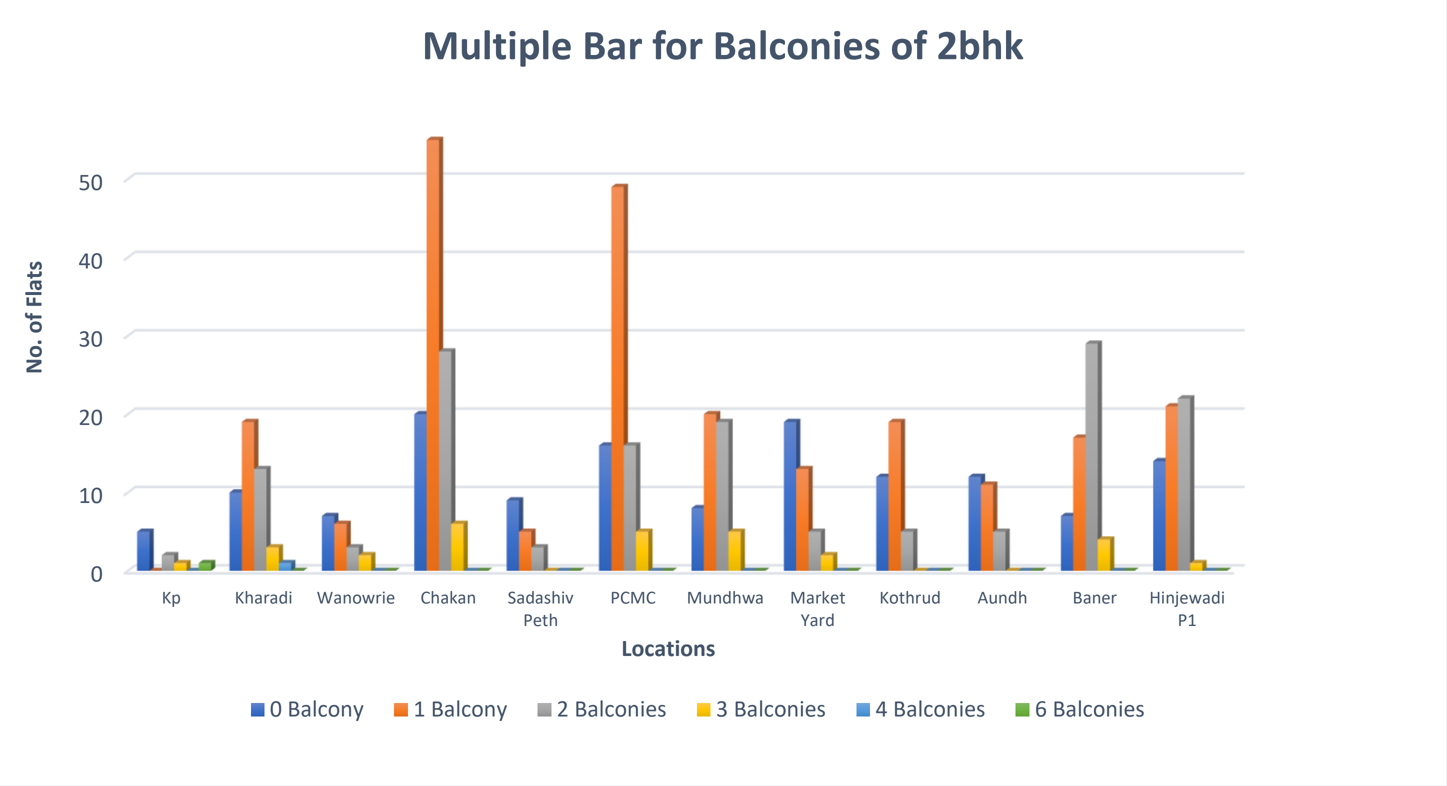
Interpretation – In our sample –

From above graph we conclude that most of the 1rk flats have 0 balconies and only flats in Wanowrie have 2 balconies.



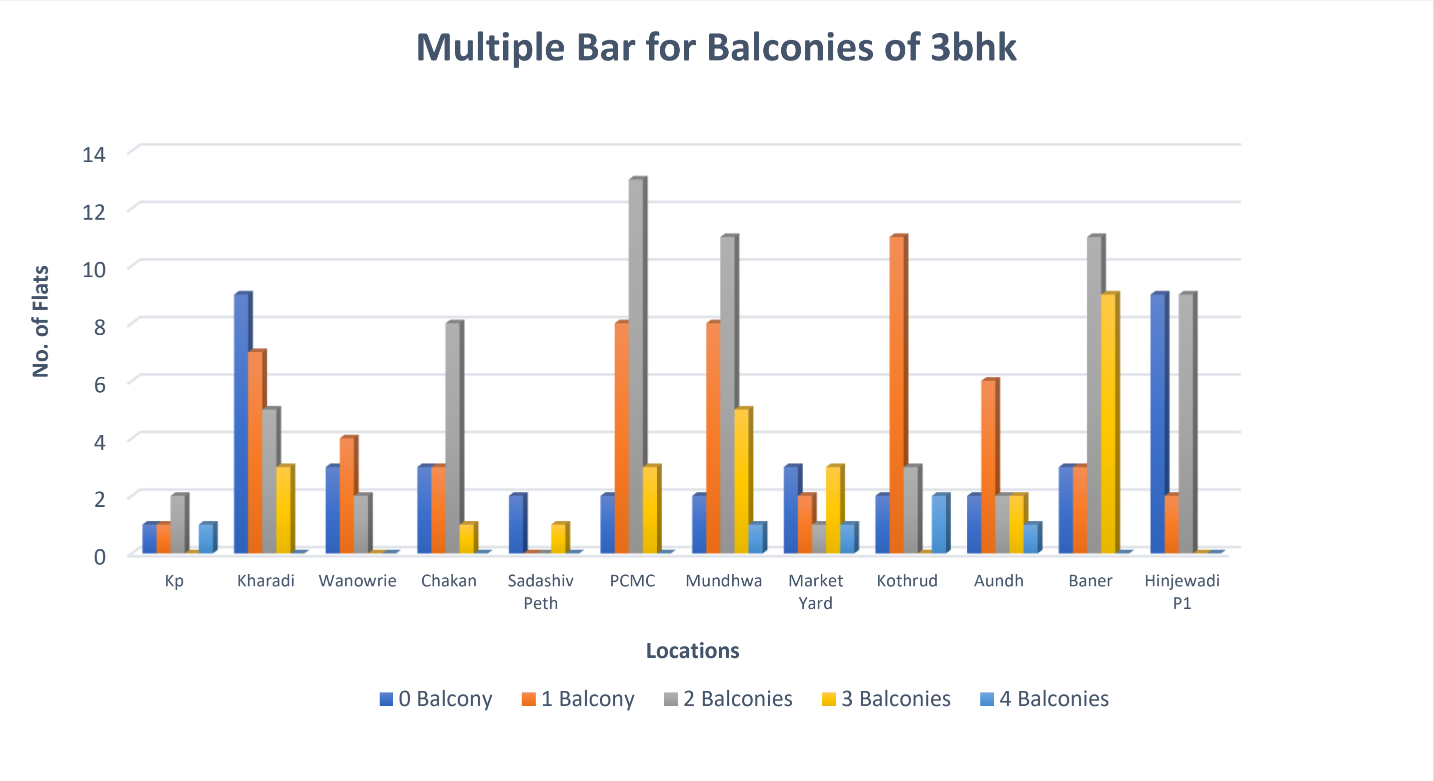
Interpretation – In our sample –

From the above graph we conclude that most of the 1bhk flats have one balcony .Chakan has highest number of one BHKs with 1 balcony.



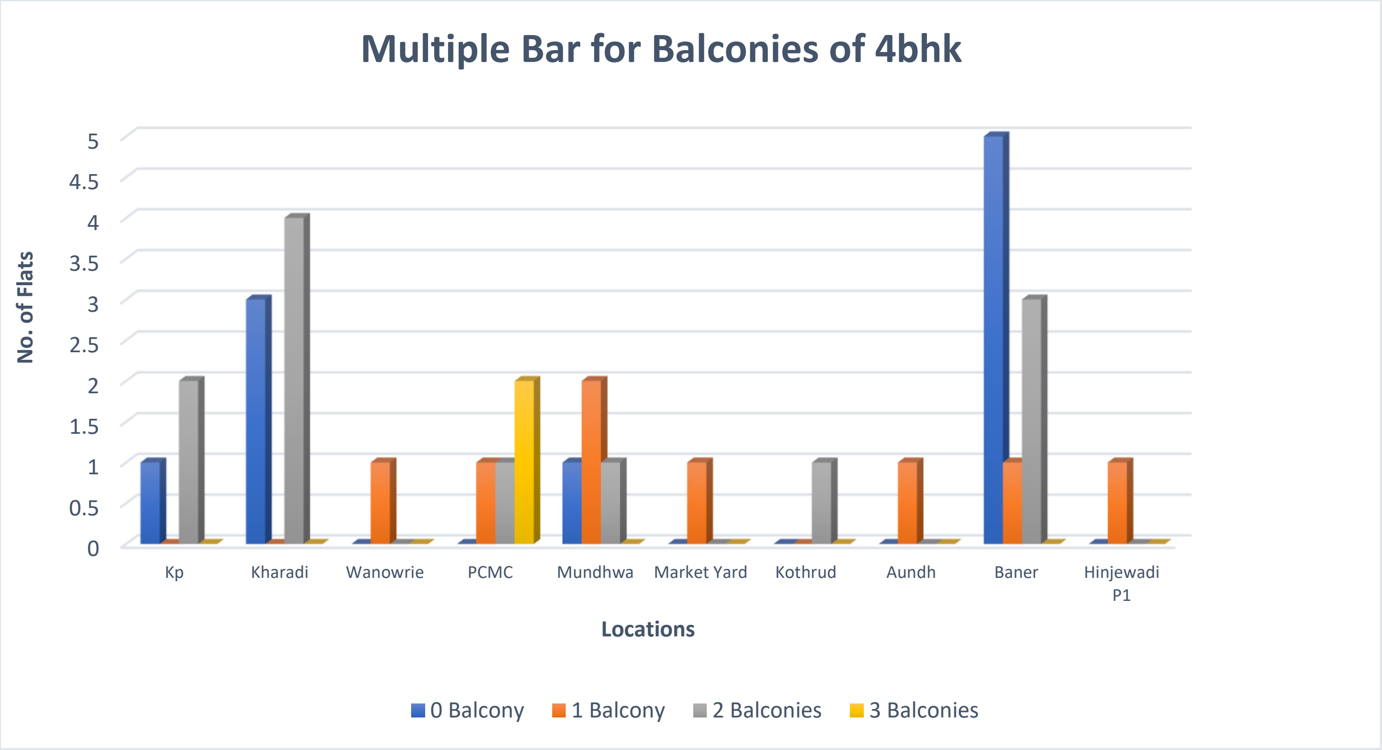
Interpretation – In our sample –

From the above graph we conclude that at most of the locations 2bhk flats have 1 or 2 balconies .Chakan and PCMC have large number of 2 BHKS with 1 balcony. Baner has highest number has 2bhks with two balconies.



Interpretation – In our sample –

From the above graph we conclude that at most of the locations 3bhk flats have 1, 2 balconies. At Baner we can observe large number of flats with 3 balconies.

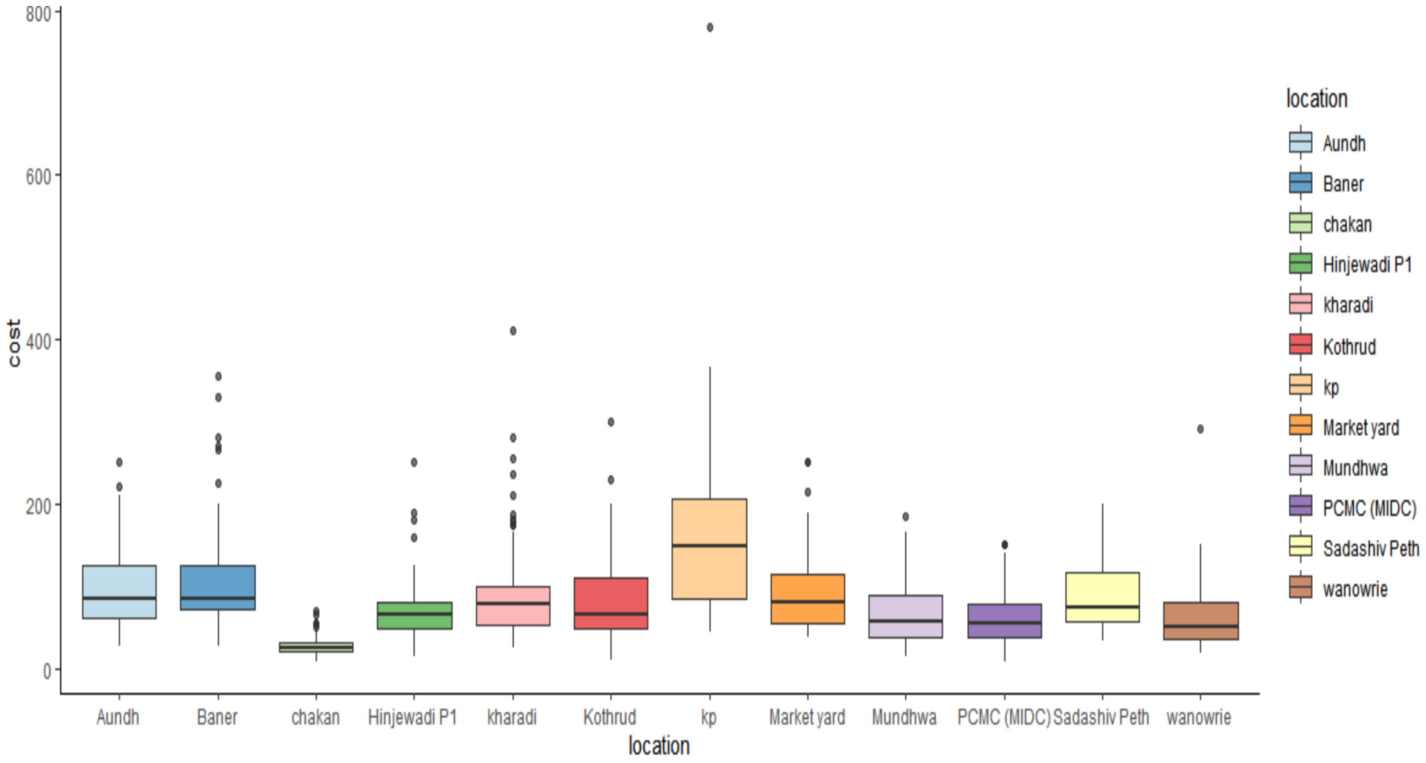


Interpretation – In our sample –

From the above graph observed that at most of the locations 4bhk flats have 0, 1 or 2 balconies. PCMC is the only location with 4 BHKS having 3 balconies.

BOX PLOTS

1) Cost (location wise)



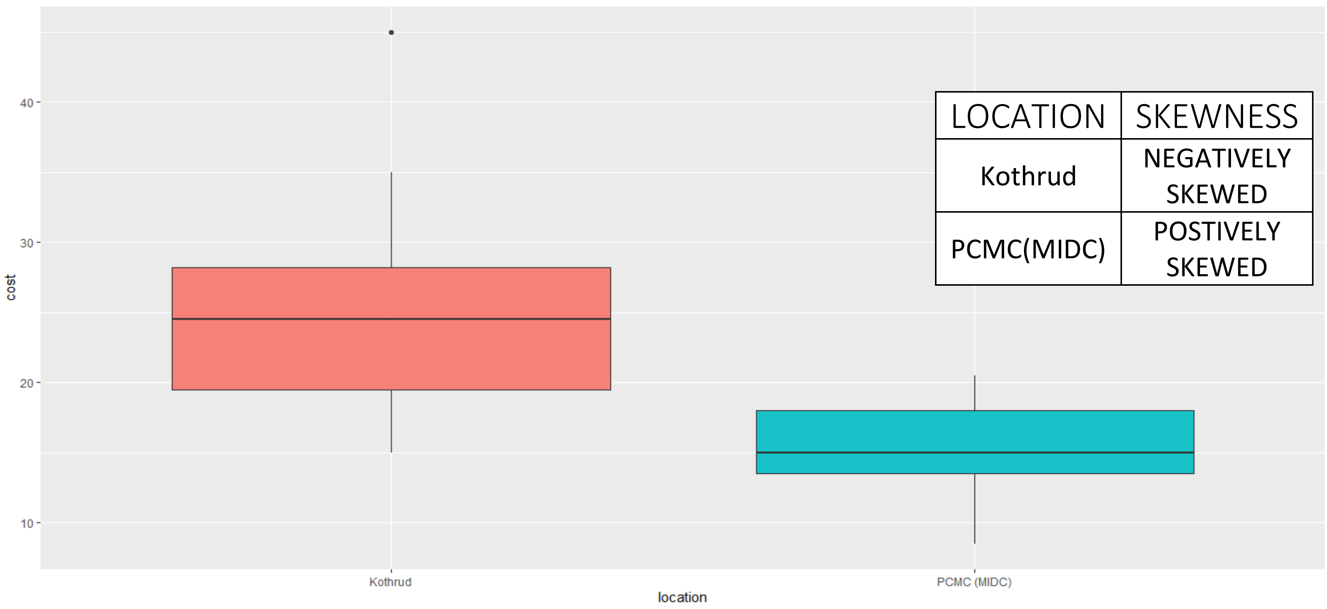
LOCATION	SKEWNESS	LOCATION	SKEWNESS
Aundh	Positive	Koregaon Park	Negative
Baner	Positive	Market yard	Positive
Chakan	Symmetric	Mundhwa	Positive
Hinjewadi P1	Negative	PCMC (MIDC)	Symmetric
Kharadi	Negative	Sadashiv Peth	Positive
Kothrud	Positive	Wanowrie	Positive

REMARKS – 1) There are few homes in all categories that have very high prices than the rest of the houses their respective groups.

2) The houses in Koregaon park (kp) are mostly costlier than the houses in other areas and Chakan has the cheapest houses out of all considered locations.

2) Cost for various locations (BHK wise)

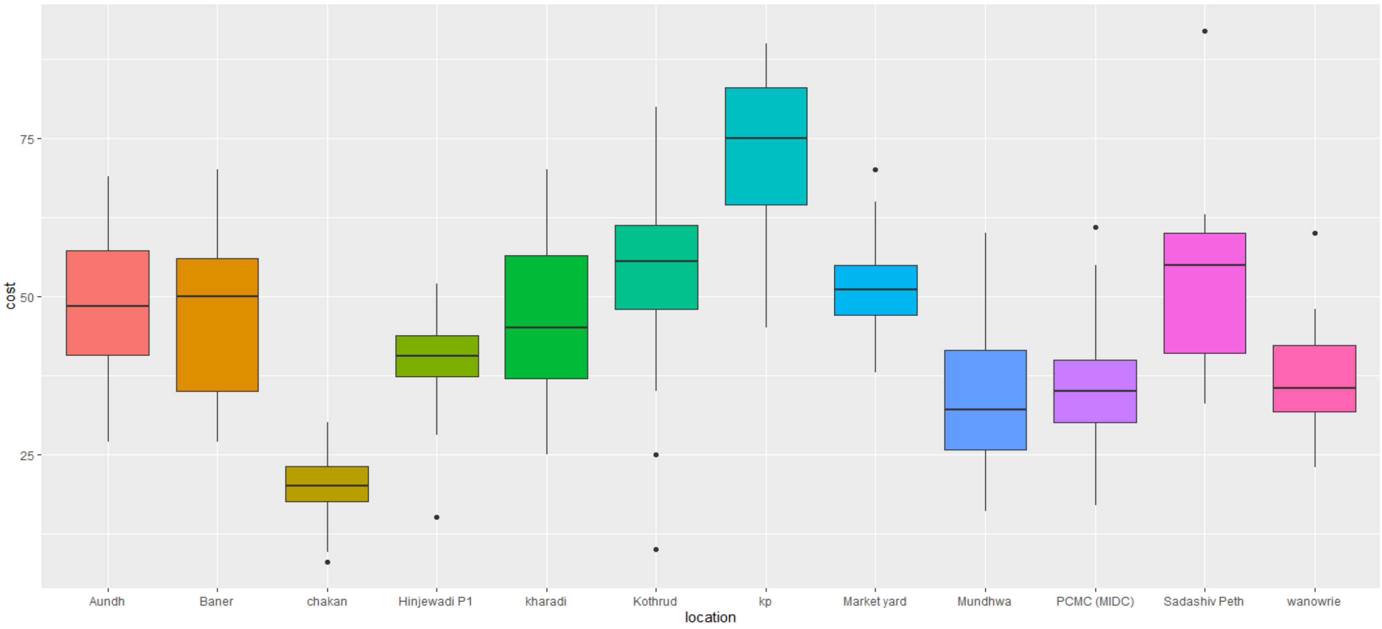
I. 1rk



LOCATION	SKEWNESS
Kothrud	NEGATIVELY SKEWED
PCMC(MIDC)	POSTIVELY SKEWED

INTERPRETATION- From boxplot of kothrud is it clear that one flat has cost price higher than all the others in kothrud area. Also almost 75% of cost prices of flats in kothrud are higher than that of all of those in PCMC (MIDC).

II. 1 BHK

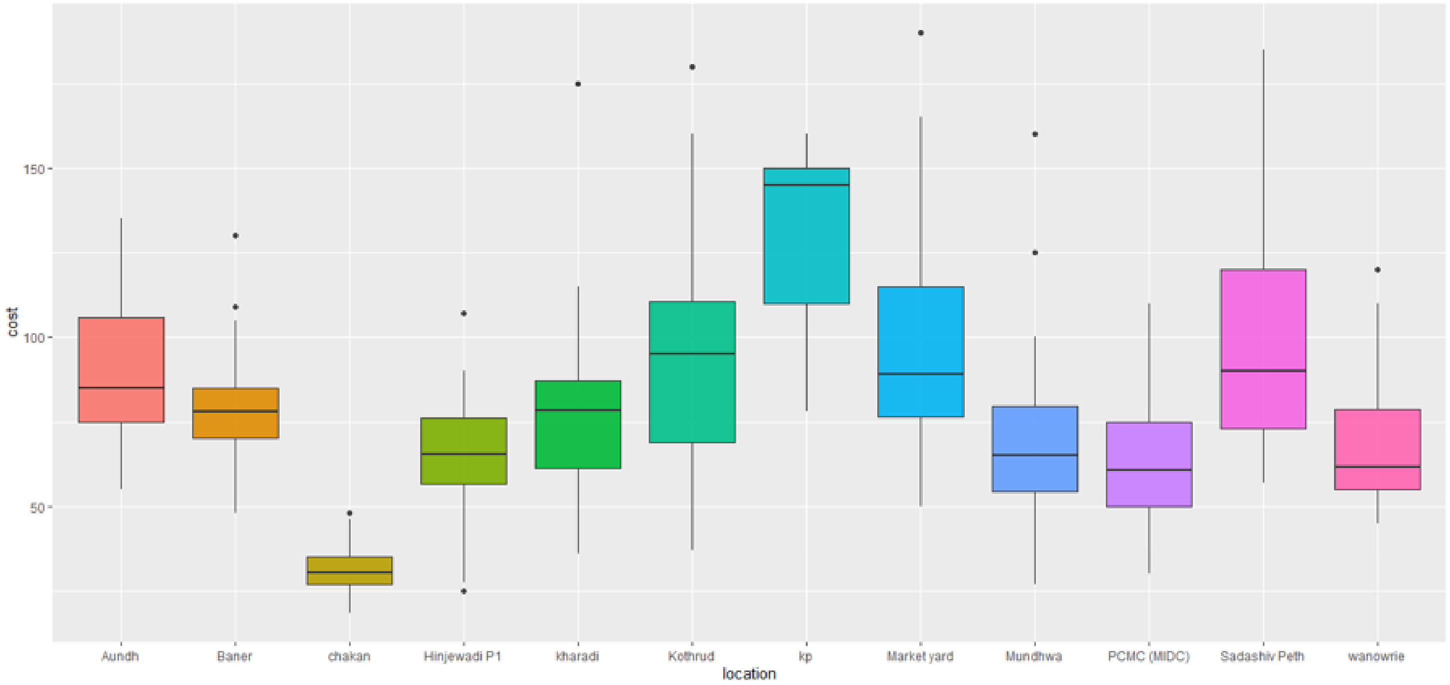


LOCATIONS	SKEWNESS
Aundh	Positively skewed
Baner	symmetric
Chakan	symmetric
Hinjewadi p1	symmetric
Kharadi	Positively skewed
Kothrud	Negatively skewed
Koregaon park(kp)	Negatively skewed
Market yard	symmetric
Mundhwa	Positively skewed
PCMC(MIDC)	symmetric
Sadashiv peth	Negatively skewed
Wanowrie	Positively skewed

INTERPRETATION-

From above boxplot is it clear that Chakan and Hinjewadi p1 have one flat each whose Cost Price is lower than all other flats in its area and similarly Kothrud has such two flats. Also market yard, PCMC (MIDC), Sadashiv Peth and Wanowrie has one flat each who's Cost Price is higher than all others in its area respectively. We also observe that cost prices of KP is mostly higher and that of Chakan is lower than all other areas.

III. 2 BHK

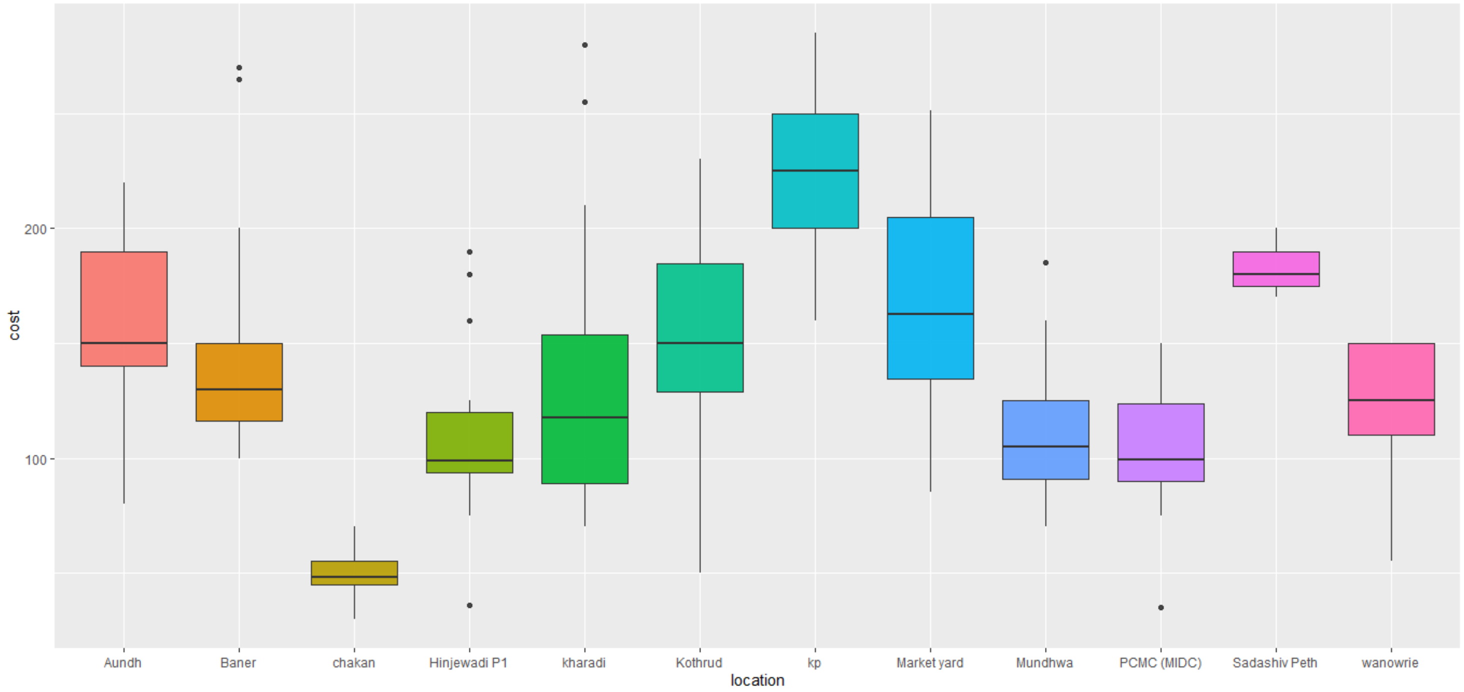


LOCATIONS	SKEWNESS
Aundh	Positively skewed
Baner	negatively skewed
Chakan	Positively skewed
Hinjewadi p1	Positively skewed
Kharadi	negatively skewed
Kothrud	Negatively skewed
Koregaon park(kp)	Negatively skewed
Market yard	Positively skewed
Mundhwa	Positively skewed
PCMC(MIDC)	Positively skewed
Sadashiv peth	Positively skewed
Wanowrie	Positively skewed

INTERPRETATION-

From boxplot of Chakan, Hinjewadi p1, Kharadi, Kothrud, market yard and Wanowrie is it clear that they have one flat each who’s Cost Price is higher than all other flats in its area and similarly Baner and Mundhwa has such two flats. Also Hinjewadi p1 also has one flat each whose Cost Price is lower than all others in its area. We also observe that cost prices of KP is mostly higher and that of Chakan is lower than all other areas just as in the case of 1bhk.

IV. 3 BHK

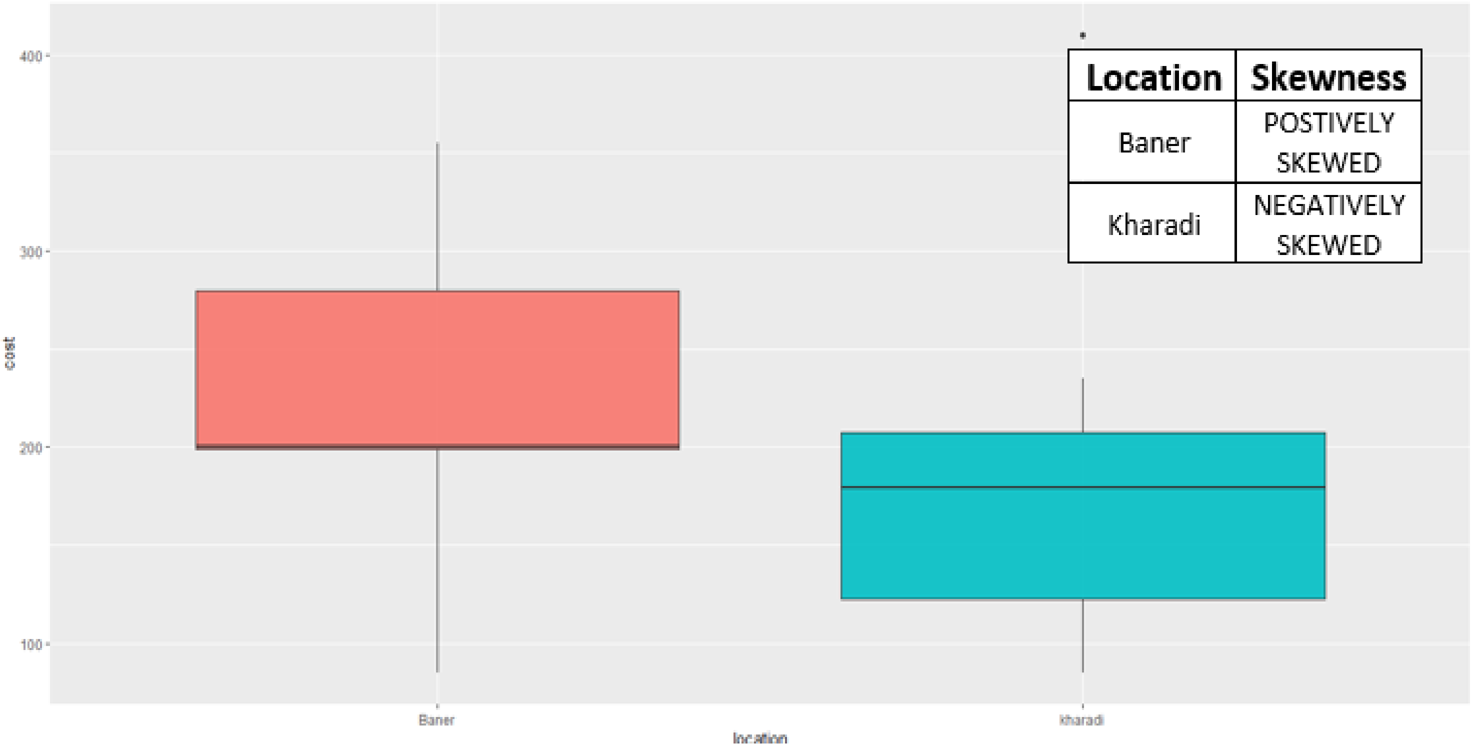


LOCATIONS	SKEWNESS
Aundh	Positively skewed
Baner	Positively skewed
Chakan	Positively skewed
Hinjewadi p1	Positively skewed
Kharadi	Positively skewed
Kothrud	Positively skewed
Koregaon park(kp)	Symmetric
Market yard	Positively skewed
Mundhwa	Positively skewed
PCMC(MIDC)	Positively skewed
Sadashiv peth	Positively skewed
Wanowrie	Positively skewed

INTERPRETATION-

From boxplot of Baner and Kharadi is it clear that they have two flats each whose Cost Price is higher than all other flats in its area Also Hinjewadi p1 also has one flat whose Cost Price is lower while three whose Cost price is higher, In case of Mundhwa, there' one whose price is higher while for PCMC (MIDC), there's one whose price is lower than all the others in its area for each. We also observe that cost prices of KP is mostly higher and that of Chakan is lower than all other areas just as in the case of 1bhk and 2 bhk.

V. 4 BHK

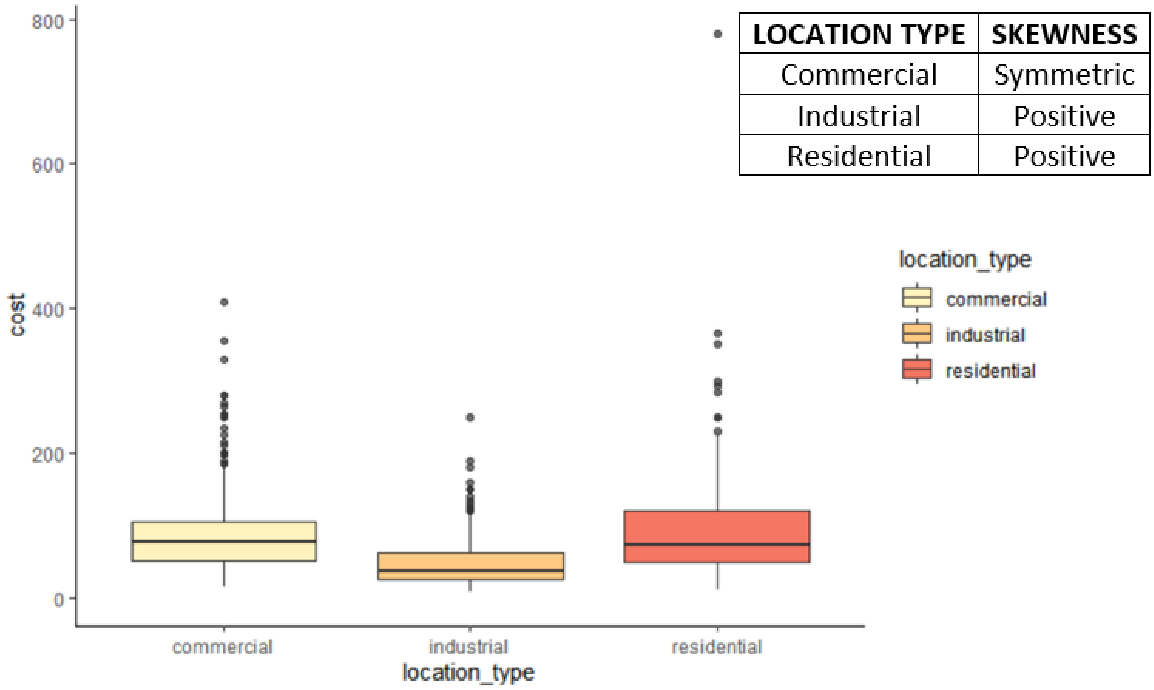


Location	Skewness
Baner	POSTIVELY SKEWED
Kharadi	NEGATIVELY SKEWED

INTERPRETATION-

From boxplot of Kharadi is it clear that one flat has cost price higher than all the others in Kharadi area. Also almost 75% of cost prices of flats in Baner are higher than that of all of those in Kharadi

3) Cost (location type wise)



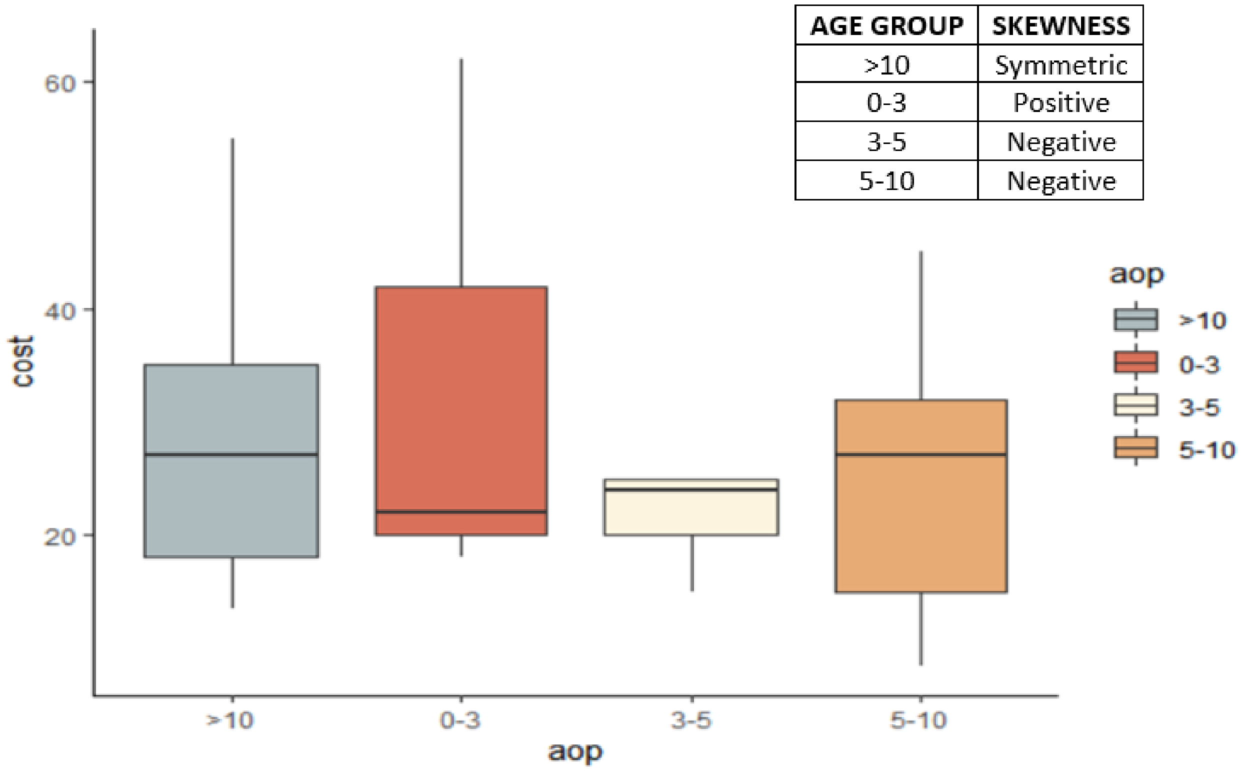
LOCATION TYPE	SKEWNESS
Commercial	Symmetric
Industrial	Positive
Residential	Positive

REMARKS – 1) There are few homes in all categories that have very high prices than the rest of the houses their respective groups.

2) The houses in residential areas are mostly more costly than the houses in other areas.

4) Cost vs age of property (BHK wise)

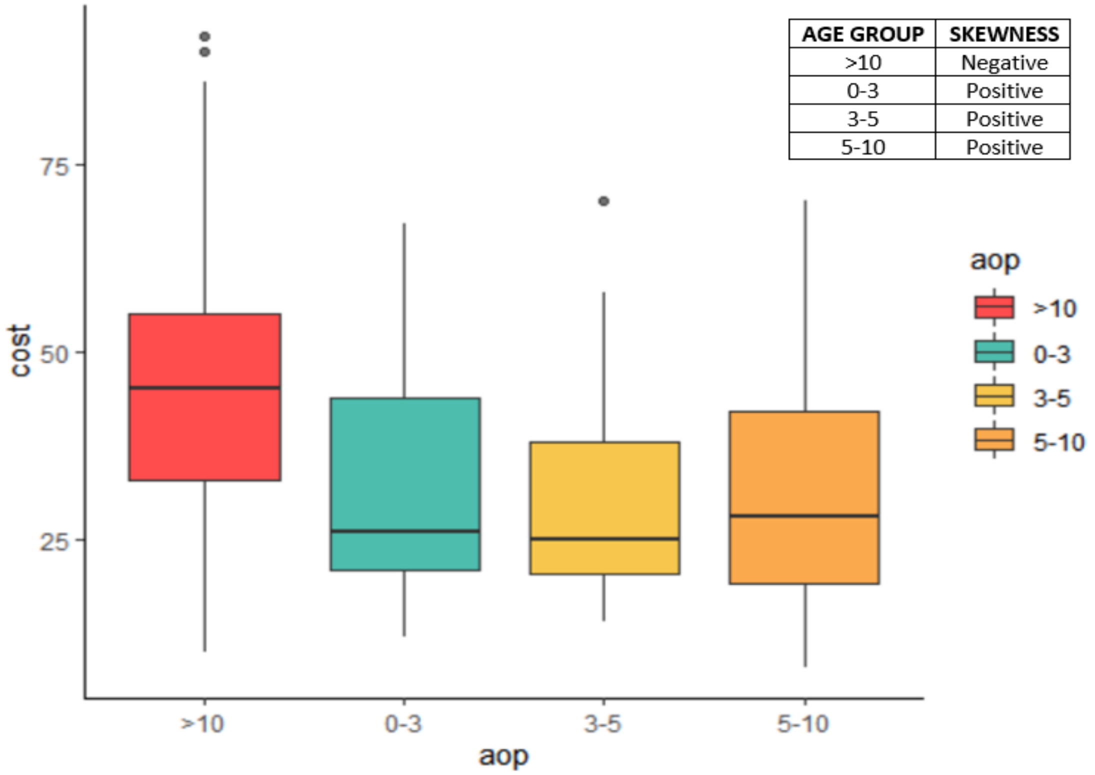
I. 1 rk



AGE GROUP	SKEWNESS
>10	Symmetric
0-3	Positive
3-5	Negative
5-10	Negative

REMARK – Newly constructed houses (0-3) are much costlier than the others.

II. 1 BHK

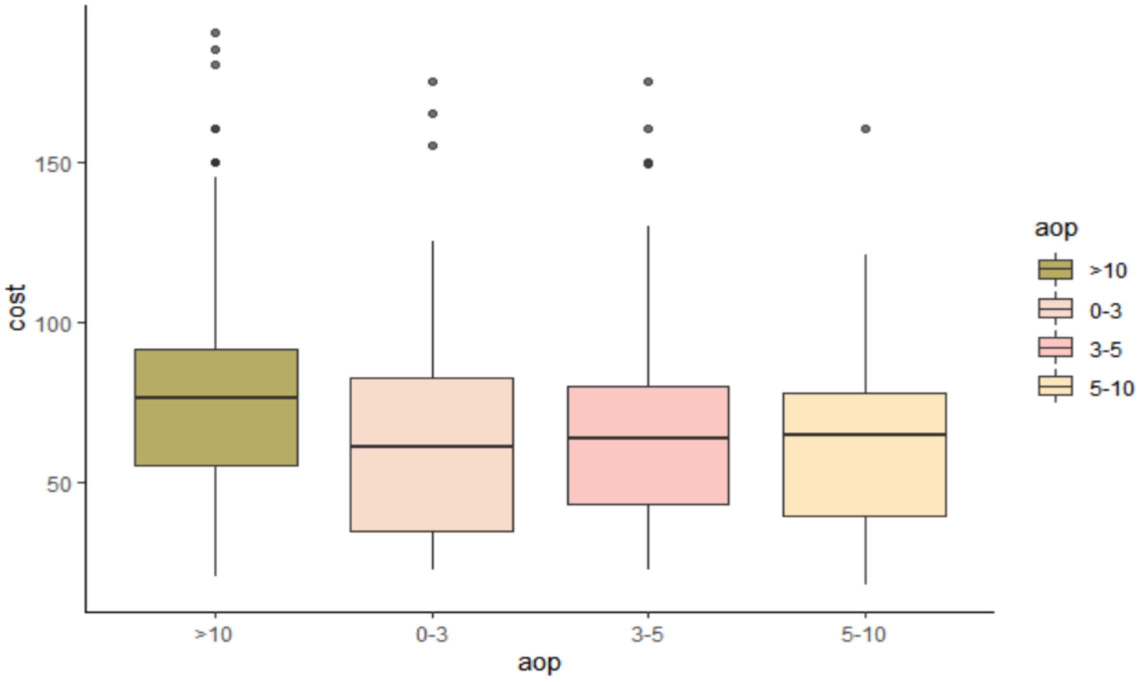


AGE GROUP	SKEWNESS
>10	Negative
0-3	Positive
3-5	Positive
5-10	Positive

REMARKS – 1) There are outliers in >10 and 3-5 age groups.

2) The houses aged more than 10 are costlier despite being older than the others.

III. 2 BHK



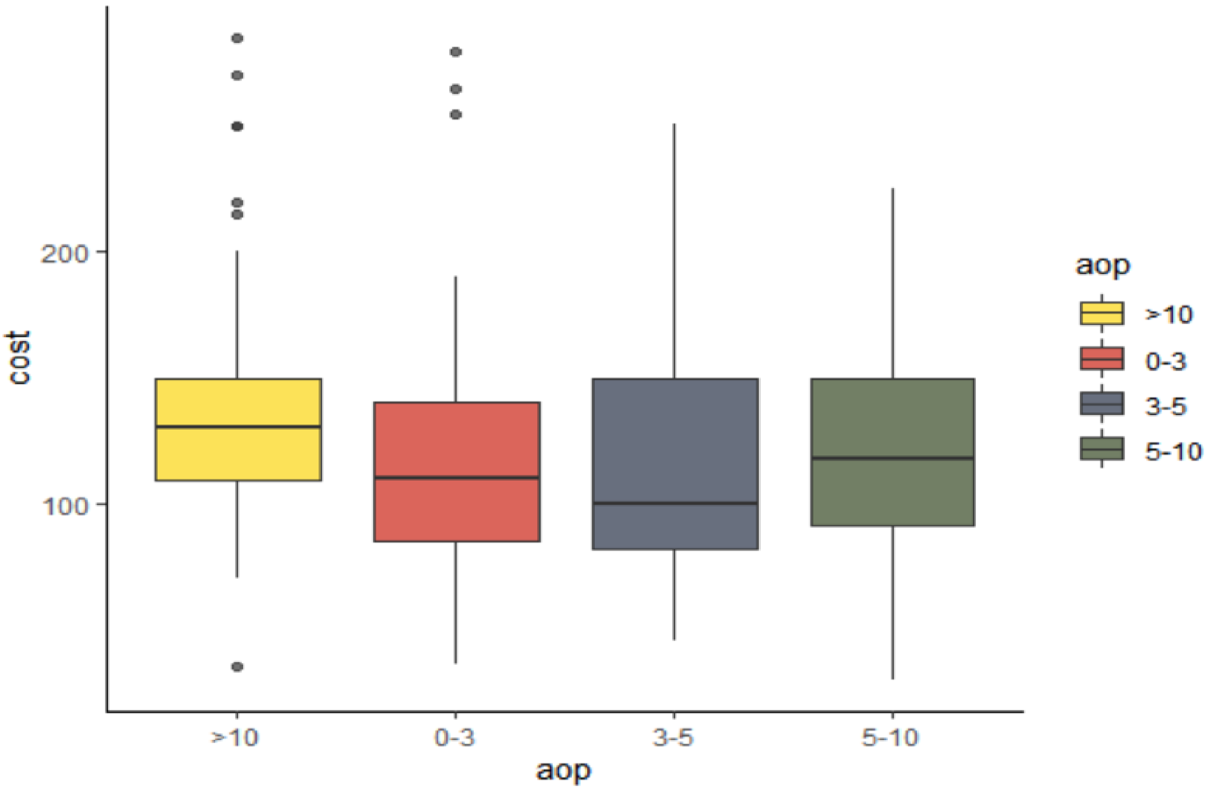
AGE GROUP	SKEWNESS
>10	Negative
0-3	Negative
3-5	Negative
5-10	Negative

REMARKS –

1) There are outliers in all age groups.

2) The houses aged more than 10 are costlier despite being older than the others.

IV. 3 BHK



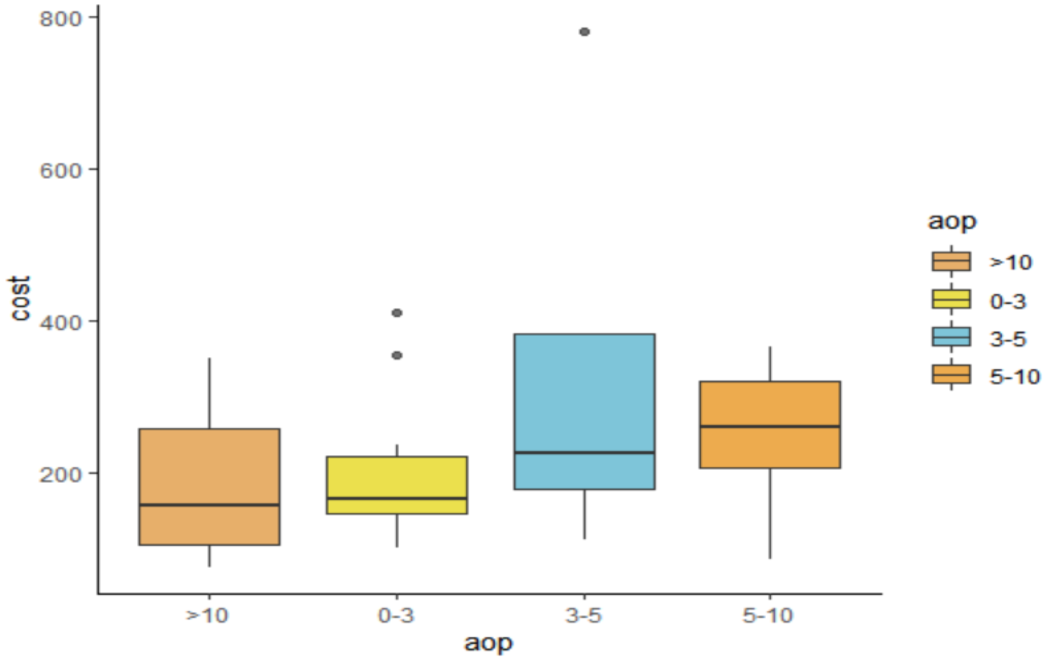
AGE GROUP	SKEWNESS
>10	Negative
0-3	Positive
3-5	Positive
5-10	Positive

REMARKS –

1) There are outliers in >10 and 0-3 age groups.

2) The houses aged more than 10 are slightly more costlier despite being older than the others.

V. 4 BHK



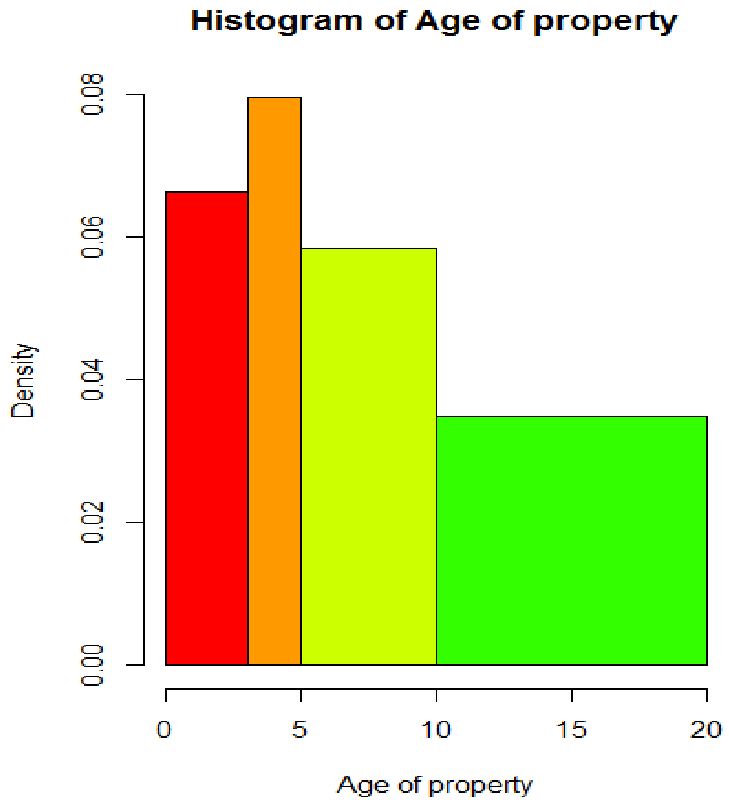
AGE GROUP	SKEWNESS
>10	Positive
0-3	Positive
3-5	Positive
5-10	Symmetric

REMARKS –

1) There are outliers in 0-3 and 3-5 age groups.

2) The houses aged 3-5 are costlier than the others.

HISTOGRAM

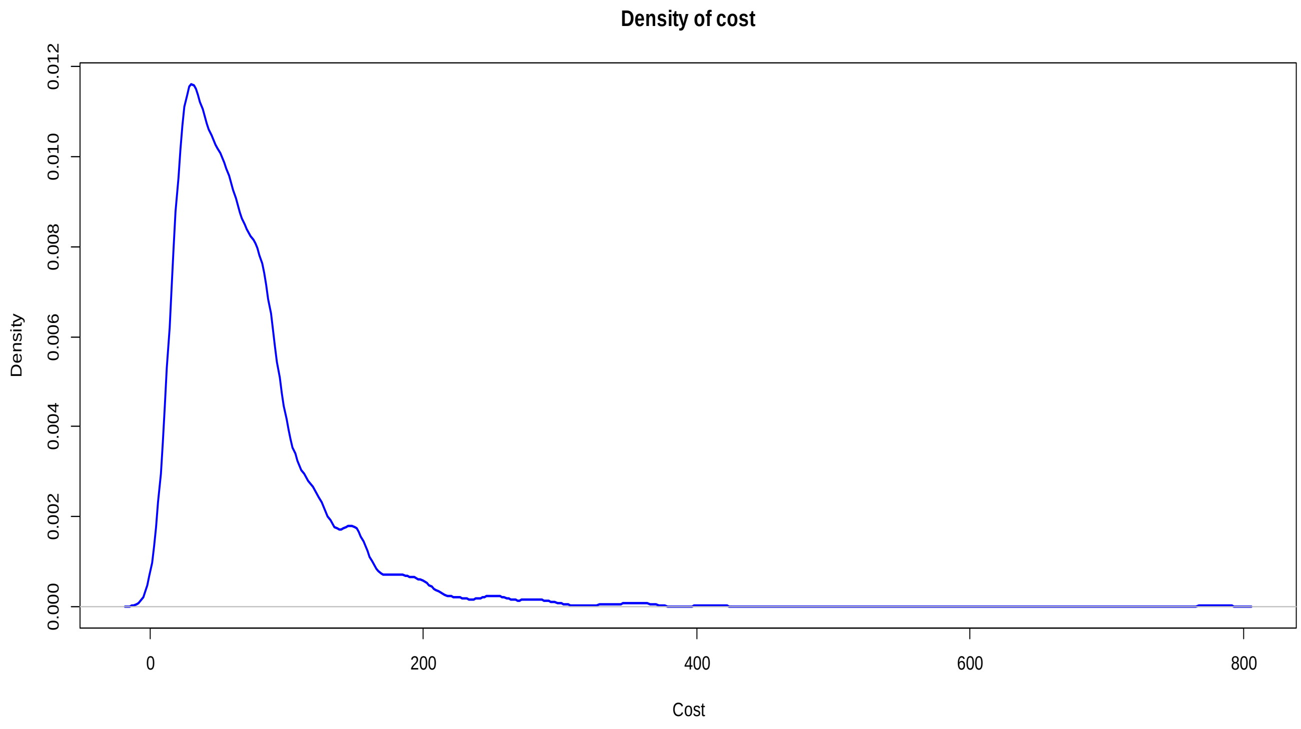


REMARKS – In our sample -

1) The maximum number of houses belong to >10 age group followed by 5- 10 age group.

2) The minimum number of houses belong to 3-5 age group.

DENSITY PLOT



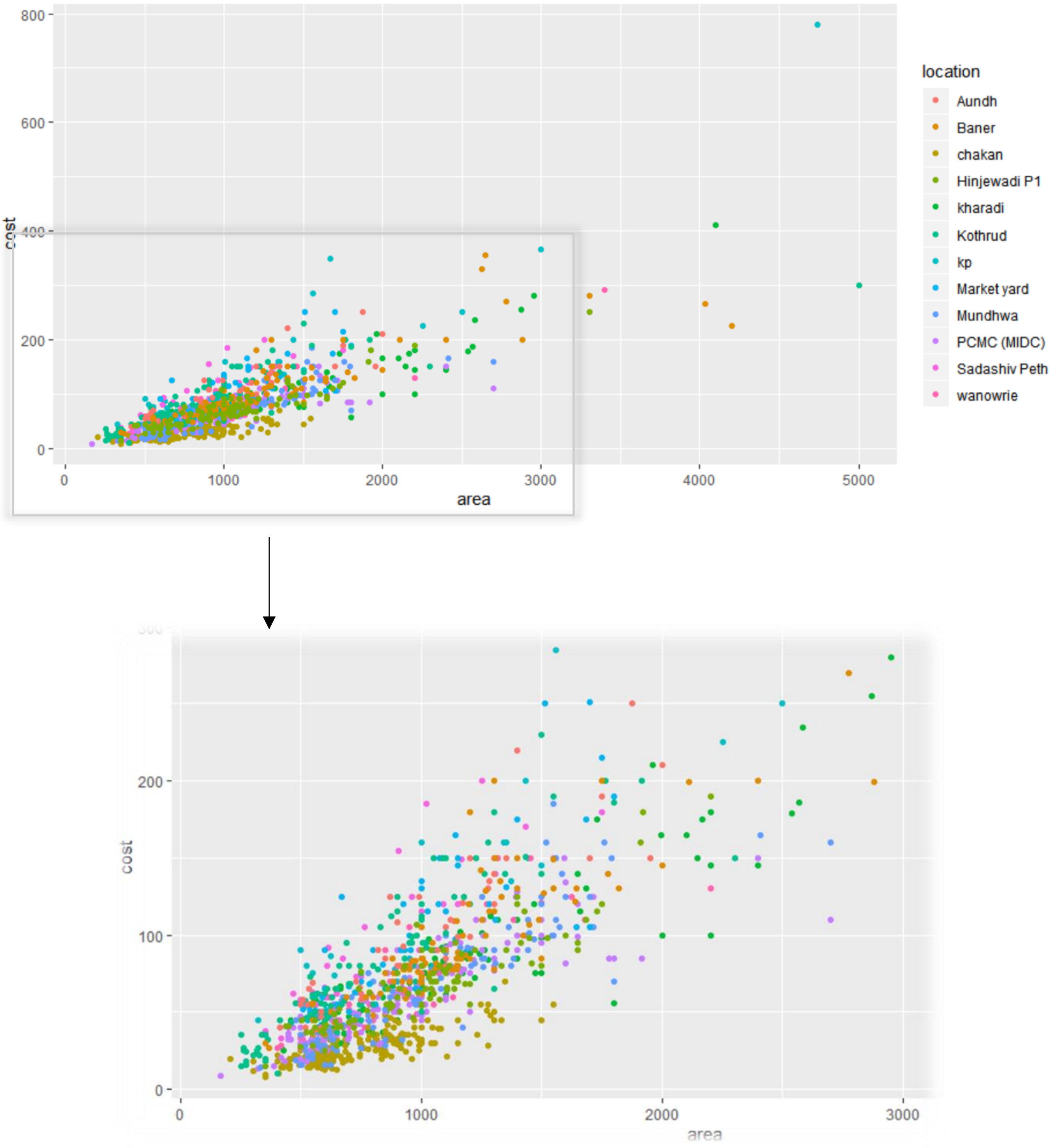
REMARKS –

1) Cost is not normally distributed.

2) The distribution is positively skewed.

SCATTER PLOTS

1) Scattter plot of cost Vs area plotted location wise



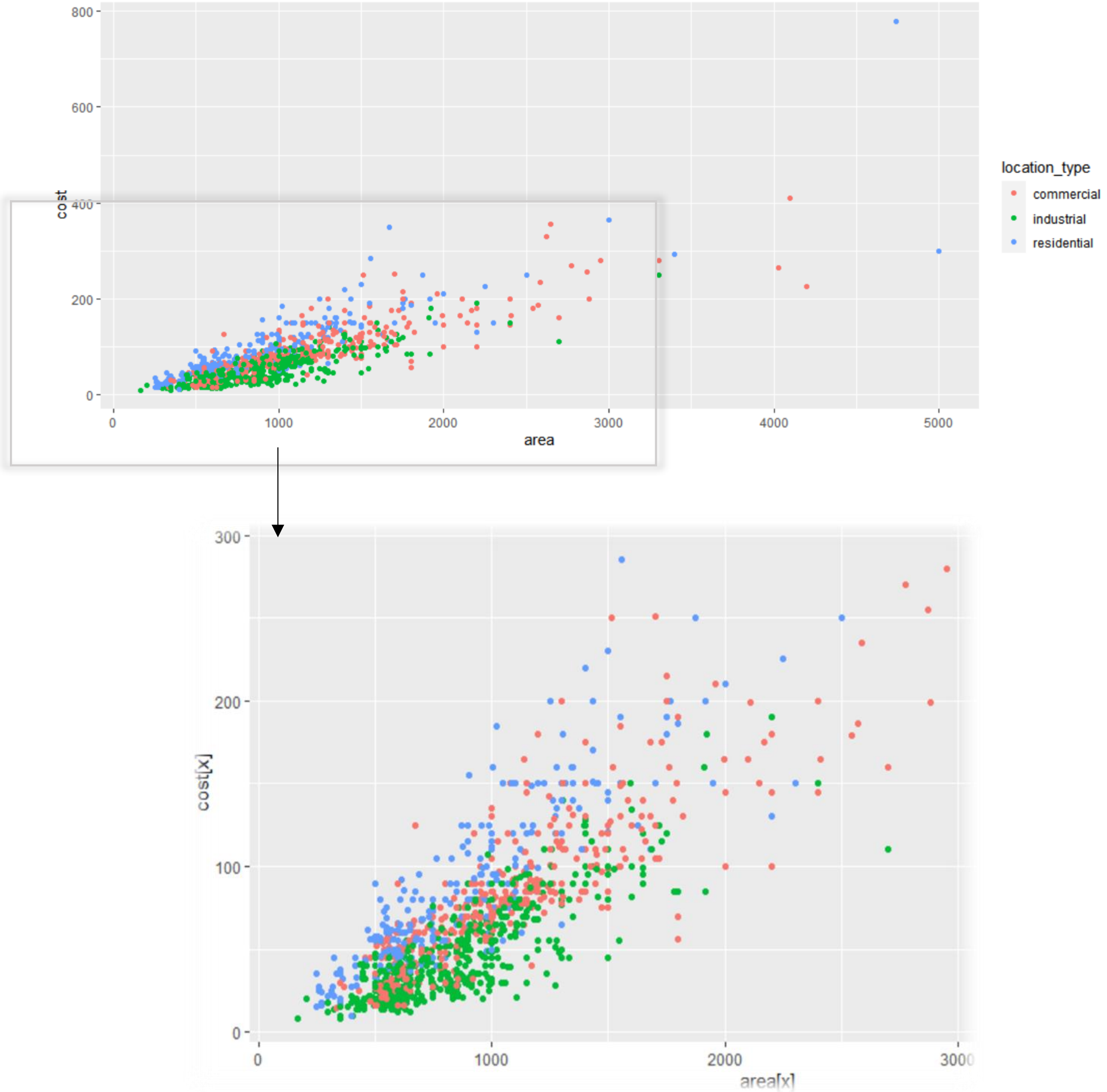
REMARKS –

1) As the area (in sq. ft) increases the cost also increases for all locations.

2) Prices of houses at Chakan stay at the bottom part of the plot i.e even though the area increases prices don't increase rapidly.

3) For certain locations prices increase very rapidly than others. E.g. – Kothrud, Koregaon Park.

2) Scattter plot of cost Vs area plotted location type wise



REMARKS –

1) As the area (in sqft) increases the cost also increases for all locations types.

2) Prices at the residential area are always at the top in plot. Hence houses are costlier in residential areas.

3) Prices at the industrial areas seem to be quite low.

KRUSKAL WALLIS TESTS

- **For Bhk types**

H0: Average cost is same for all bhk type.

Vs

H1: Average cost differ significantly for all bhk type.

Kruskal-Wallis Rank Test Sum

Kruskal-Wallis chi-squared = 631.9, df = 4, p-value < 2.2e-16

Conclusion: Here p value is less than level of significance α =0.05, hence we reject null hypothesis. Therefore, average cost differs significantly for all bhks

- **For Furnishing types**

H_0: Average cost of houses is same for all furnishing

Vs

H_1: Average cost of houses differ significantly for all furnishing

Kruskal-Wallis rank sum test

Kruskal-Wallis chi-squared = 155.83, DF = 2, p-value < 2.2e-16

Conclusion: As p-value is less than level of significance α = 0.05, H_0 is rejected. Hence cost differs significantly for all furnishing types.

- **For Age of property**

H_O: Average cost of houses is same for all age groups.

Vs

H_1: Average cost of houses differ significantly for all age groups.

Kruskal-Wallis rank sum test

Kruskal-Wallis chi-squared = 30.022, df = 3, p-value = 1.366e-06

Here p-value is less than 0.05 level of significance, so we reject the null hypothesis.

Hence, Average cost of houses differ significantly for all age groups.

- **For location types**

Ho : Average cost of houses is same for all location types.

Vs

H1 : Average cost of houses differ significantly for all location types.

Kruskal-Wallis rank sum test

Kruskal-Wallis chi-squared = 269.29, df = 2, p-value < 2.2e-16

Conclusion: As p value is less than level of significance α = 0.05 Ho is rejected. Hence average cost differs significantly for all location types.

- **For locations.**

H_O: Average cost of houses is same for all locations.

Vs

H_1: Average cost of houses differ significantly for all locations.

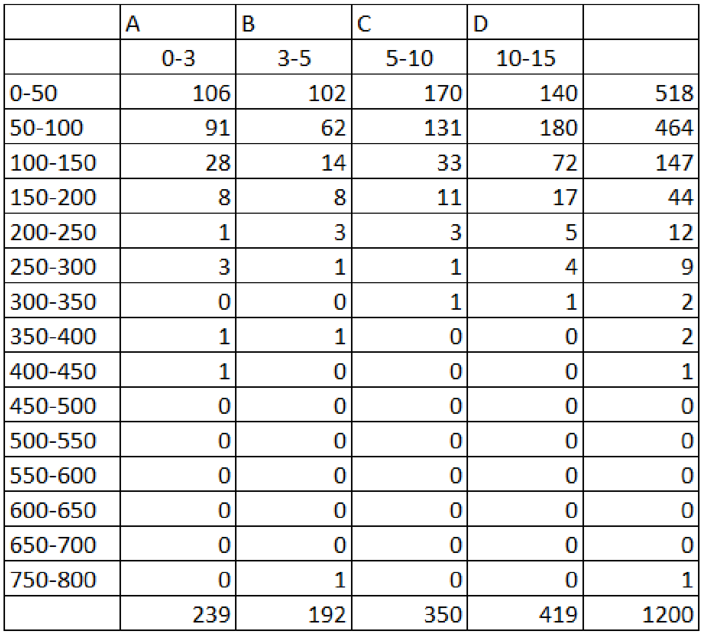
Kruskal-Wallis rank sum test

Kruskal-Wallis chi-squared = 424.46, df = 11, p-value < 2.2e-16

Here p-value is less than level of significance α = 0.05, so we reject null hypothesis. Hence, Average cost of houses differ significantly for all locations.

TESTS FOR INDEPENDENCE

1) TEST FOR CHECKING ASSOCIATION BETWEEN AGE AND COST PRICES OF FLATS



	A	B	C	D	
	0-3	3-5	5-10	10-15	
0-50	106	102	170	140	518
50-100	91	62	131	180	464
100-150	28	14	33	72	147
150-200	8	8	11	17	44
200-250	1	3	3	5	12
250-300	3	1	1	4	9
300-350	0	0	1	1	2
350-400	1	1	0	0	2
400-450	1	0	0	0	1
450-500	0	0	0	0	0
500-550	0	0	0	0	0
550-600	0	0	0	0	0
600-650	0	0	0	0	0
650-700	0	0	0	0	0
750-800	0	1	0	0	1
	239	192	350	419	1200

H o : There is no association between age and cost prices of flats.

H 1 : There is association between age and cost prices of flats.

Pearson's Chi-squared test

X-squared = 69.116, df = 36, p-value =0.000742

Here p value is less than 0.05 level of significance, therefore we reject null hypothesis.

Hence age and cost are associated.

- **TO CHECK EXACT AMOUNT OF ASSOCIATION BETWEEN AGE AND COST PRICES OF FLATS**

Cramer's V test:
Test statistic:

$$V=\sqrt{\frac{\chi^2}{n*(q-1)}}$$

Where q is min (row and column)

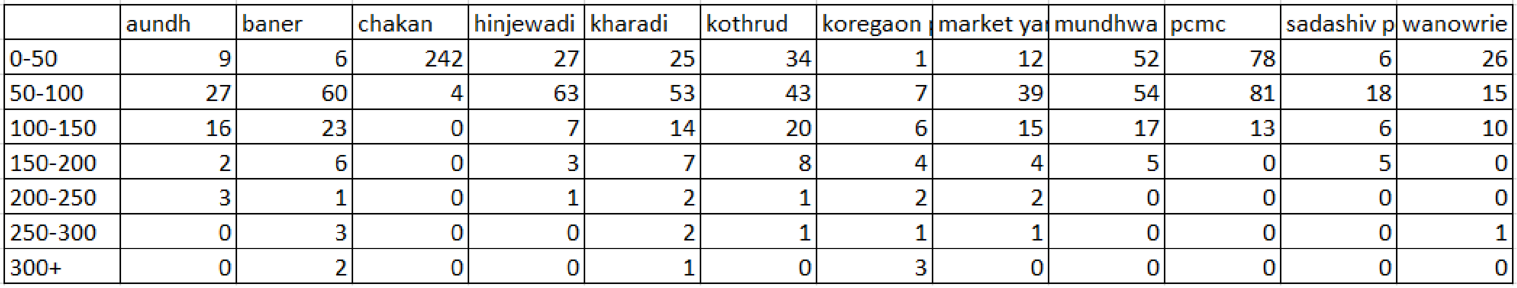
X-squared = 69.116
n= 1200

q=4

Test statistic: 0.1385601

Conclusion: Since, **test statistic value is greater than 0.10** and hence there is a **moderate** association between age and cost.

2) TEST FOR CHECKING ASSOCIATION BETWEEN LOCATION AND COST PRICES OF FLATS



	aundh	baner	chakan	hinjewadi	kharadi	kothrud	koregaon	market ya	mundhwa	pcmc	sadashiv p	wanowrie
0-50	9	6	242	27	25	34	1	12	52	78	6	26
50-100	27	60	4	63	53	43	7	39	54	81	18	15
100-150	16	23	0	7	14	20	6	15	17	13	6	10
150-200	2	6	0	3	7	8	4	4	5	0	5	0
200-250	3	1	0	1	2	1	2	2	0	0	0	0
250-300	0	3	0	0	2	1	1	1	0	0	0	1
300+	0	2	0	0	1	0	3	0	0	0	0	0

H o : There is no association between location and cost prices of flats.

H 1 : There is association between location and cost prices of flats.

Pearson's Chi-squared test

X-squared = 639.97 , df = 66 , p-value < 2.2e-16

Here p value is less than 0.05 level of significance, therefore we reject null hypothesis.

Hence location and cost are associated.

- **TO CHECK EXACT AMOUNT OF ASSOCIATION BETWEEN LOCATION AND COST PRICES OF FLATS**

Cramer's V test:

Test statistic:

$$V=\sqrt{\frac{\chi^2}{n*(q-1)}}$$

Where q is min (row and column)

X-squared = 639.97

N = 1200

q = 7

Test statistic: 0.298135

Conclusion: Since, **test statistic value is greater than 0.25** and hence there is a **strong** association between location and cost.

3) TEST FOR CHECKING ASSOCIATION OF COST AND LOCATION TYPE

	Location Type			
Cost	**Commercial**	**Residential**	**Industrial**	**Total**
0-100	301	186	495	982
100-200	91	77	23	191
200-300	11	9	1	21
>300	3	3	0	6
Total	406	275	519	1200

H_o : There is no association between location type and cost prices of flats.

H_1 : There is association between location type and cost prices of flats.

Pearson's Chi-squared test

X-squared = 118.3 , df = 6 , p-value < 2.2e-16

Here p value is less than 0.05 level of significance, therefore we reject null hypothesis.

There is association between location type and cost prices of flats.

- **TO CHECK EXACT AMOUNT OF ASSOCIATION BETWEEN LOCATION TYPE AND COST PRICES OF FLATS**

Cramer's V test:

Test statistic:

$$V=\sqrt{\frac{\chi^2}{n*(q-1)}}$$

Where q is min (row and column)

χ^2=118.3

n=1200

q=3

Test statistic: 0.2220173

Conclusion: Since, **test statistic value is greater than 0.15** and hence there is a **strong association** between **location type and cost prices of flats.**

4) TEST FOR CHECKING ASSOCIATION OF COST AND FURNISHING TYPE

	Furnishing Type			
	Furnished	**Semi**	**Unfurnished**	**Total**
0-100	101	377	504	982
100-200	53	91	47	191
200-300	6	11	4	21
>300	2	3	1	6
Total	162	482	556	1200

H_o : There is no association between furnishing type and cost prices of flats.

H_1 : There is association between furnishing type and cost prices of flats.

Pearson's Chi-squared test

X-squared = 75.129 , df = 6 , p-value = 3.611e-14

Here p value is less than 0.05 level of significance, therefore we reject null hypothesis.

There is association between furnishing type and cost prices of flats.

- **TO CHECK EXACT AMOUNT OF ASSOCIATION BETWEEN FURNISHING TYPE AND COST PRICES OF FLATS**

Cramer's V test:

Test statistic:

$$V= \sqrt{\frac{\chi^2}{n*(q-1)}}$$

Where q is min (row and column)

χ^2=75.129

n=1200

q=3

Test statistic: 0.1769287

Conclusion: Since, **test statistic value is greater than 0.15** and hence there is a **strong association** between **furnishing type and cost prices of flats.**

4) TEST FOR CHECKING ASSOCIATION OF COST AND POWER BACKUP TYPE

	Power Backup			
	Yes	**No**	**Partial**	**Total**
0-100	128	702	152	982
100-200	115	38	38	191
200-300	6	13	2	21
>300	1	5	0	6
Total	250	758	192	1200

H_o : There is no association between power backup type and cost prices of flats.

H_1 : There is association between power backup type and cost prices of flats.

Pearson's Chi-squared test

X-squared = 242.73 , df = 6 , p-value < 2.2e-16.

Here p value is less than 0.05 level of significance, therefore we reject null hypothesis.

There is association between power backup type and cost prices of flats.

- **TO CHECK EXACT AMOUNT OF ASSOCIATION BETWEEN POWER BACKUP AND COST PRICES OF FLATS**

Cramer’s V test:

Test statistic:

$$V= \sqrt{\frac{\chi^2}{n*(q-1)}}$$

Where q is min (row and column)

χ^2=242.73

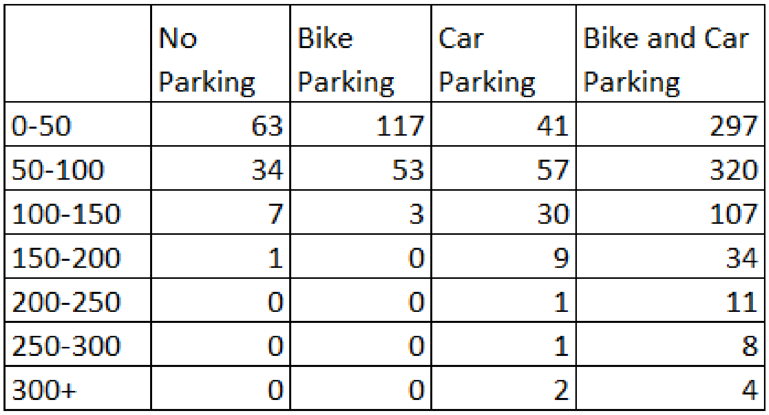
n=1200

q=3

Test statistic: 0.3180212

Conclusion: Since, **test statistic value is greater than 0.15** and hence there is a **very strong association** between **power backup type and cost prices of flats.**

5) TEST FOR CHECKING ASSOCIATION OF COST AND PARKING TYPE



	No Parking	Bike Parking	Car Parking	Bike and Car Parking
0-50	63	117	41	297
50-100	34	53	57	320
100-150	7	3	30	107
150-200	1	0	9	34
200-250	0	0	1	11
250-300	0	0	1	8
300+	0	0	2	4

H o : There is no association between availability of parking and cost prices of flats.

H 1 : There is association between location availability of parking and cost prices of flats.

Pearson's Chi-squared test

X-squared = 99.432 , df = 18 , p-value = 2.814e-13

Here p value is less than 0.05 level of significance, therefore we reject null hypothesis.

There is association between parking type and cost prices of flats.

- **TO CHECK EXACT AMOUNT OF ASSOCIATION BETWEEN LOCATION TYPE AND COST PRICES OF FLATS**

χ^2= 99.432

N =1200

q = 4

Test statistic: 0.1661927

Conclusion: Since, **test statistic value is greater than 0.15** and hence there is a **strong association** between **parking type and cost prices of flats.**

Statistical Analysis

To build a regression model, let-

Response variable – Cost of houses (in lakhs)

Regressors –

Numeric regressors	Categorical regressors
1. Build Area (sq. ft.)	
2. Number of Bedrooms	1. Location
3. Number of Bathrooms	2. Location Type
4. Number of Balconies	3. Furnishing
5. Total of Amenities	4. Parking,
6. Distance from Airport	5. Power Backup
7. Distance from Railway Station	6. Age of Property
8. Distance from Bus Stop.	

Assumptions check for a Linear Regression Model-

1. NORMALITY TEST:

Shapiro-Wilk normality test of **Cost prices**

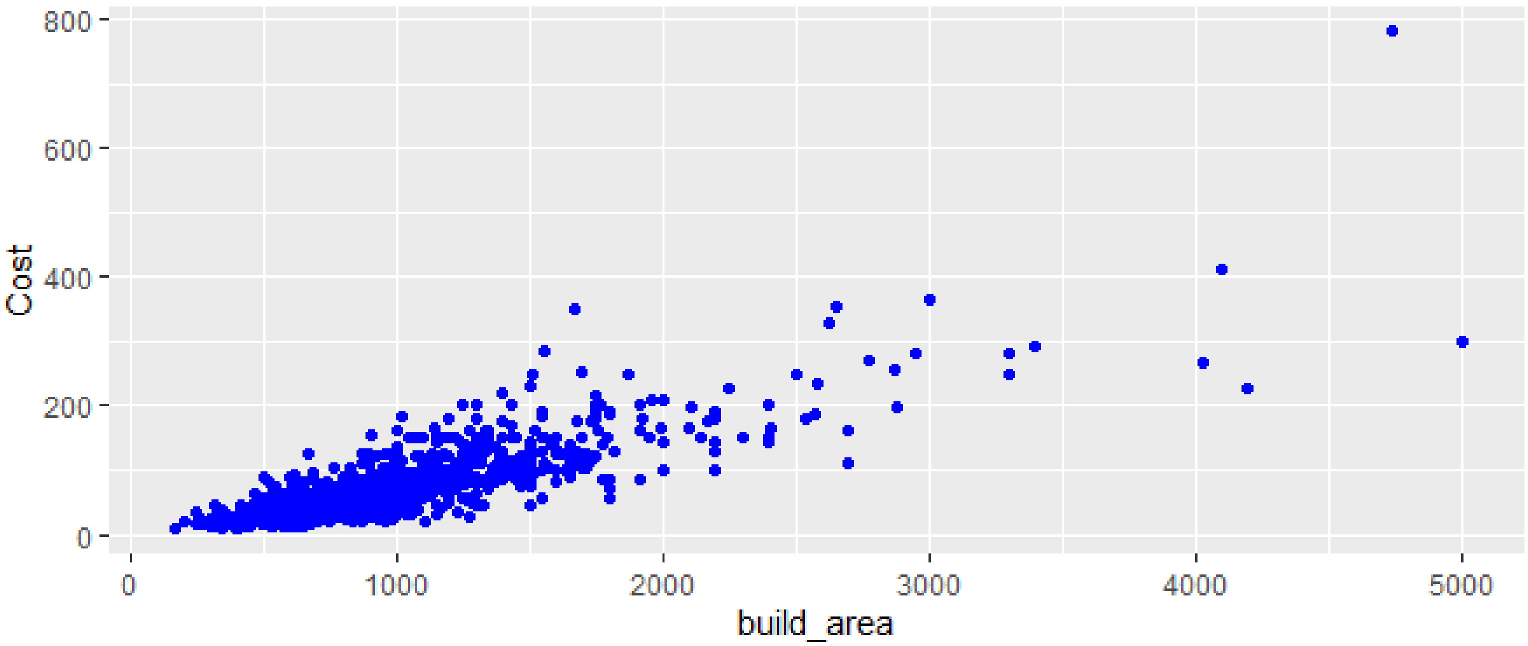
```
data:  Cost (in lakhs)

W = 0.75418, p-value < 2.2e-16
```

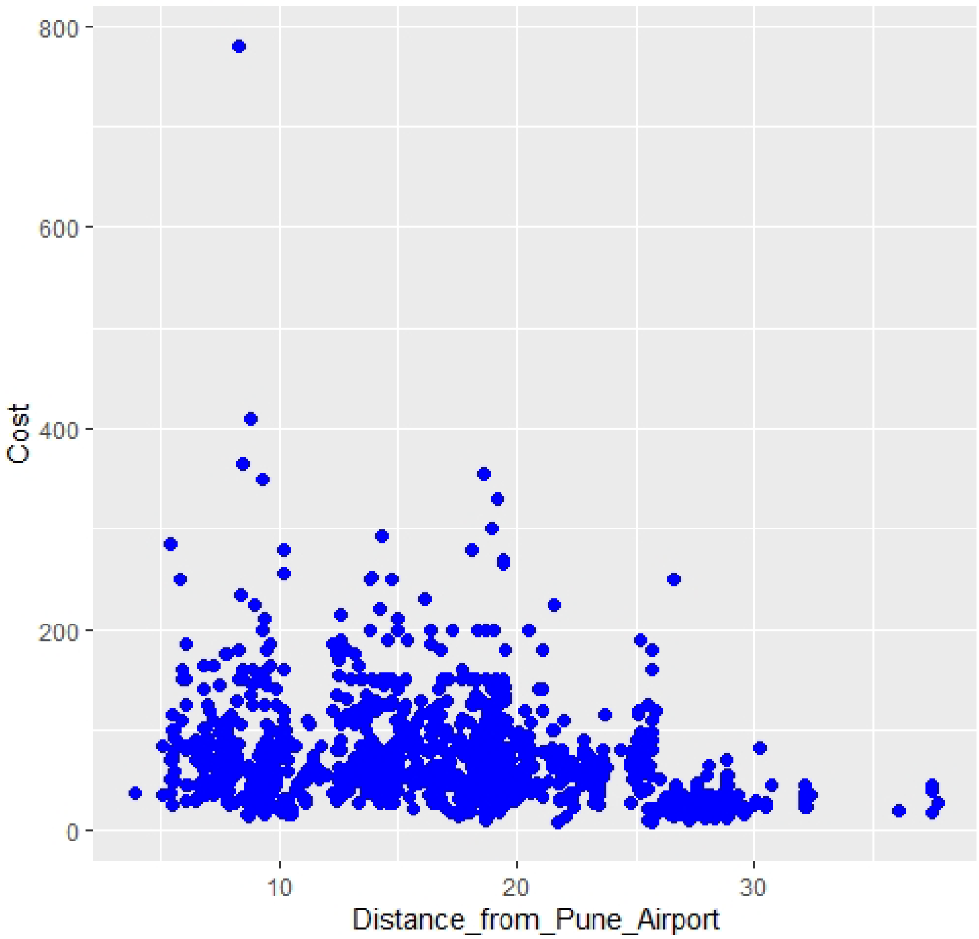
The response variable i.e. Cost prices of houses is not normally distributed.

2. To check linearity assumption between regressors and response variable

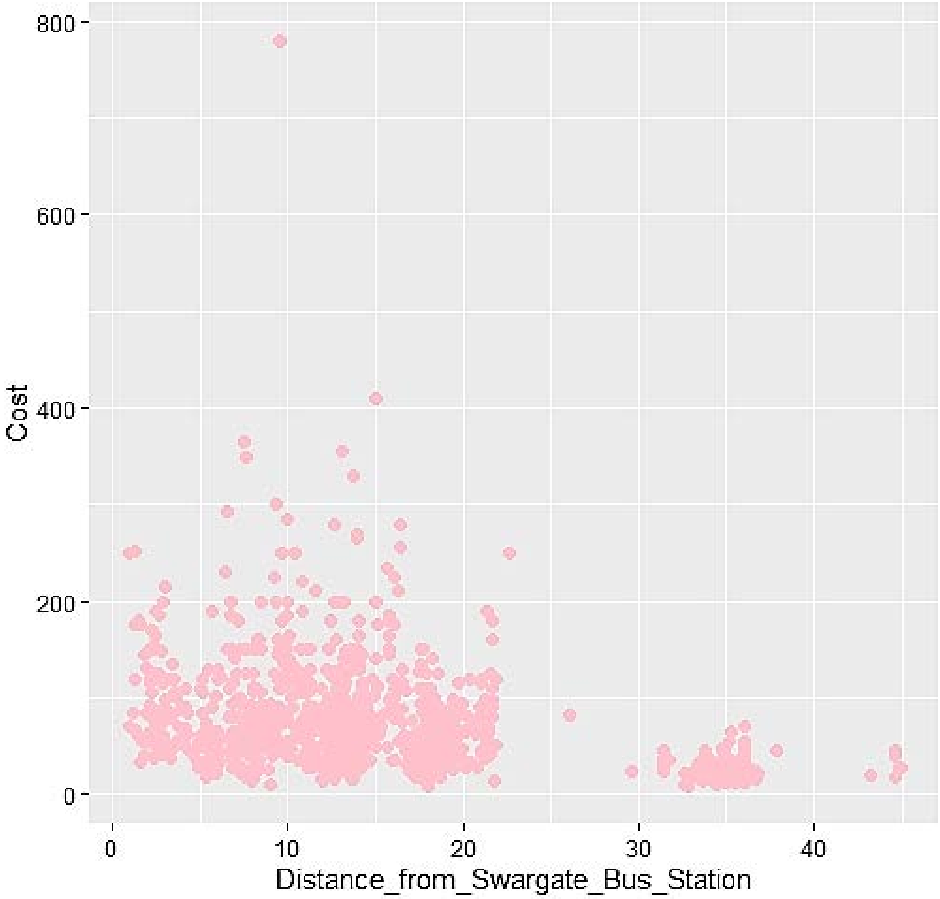
- Scatter plots



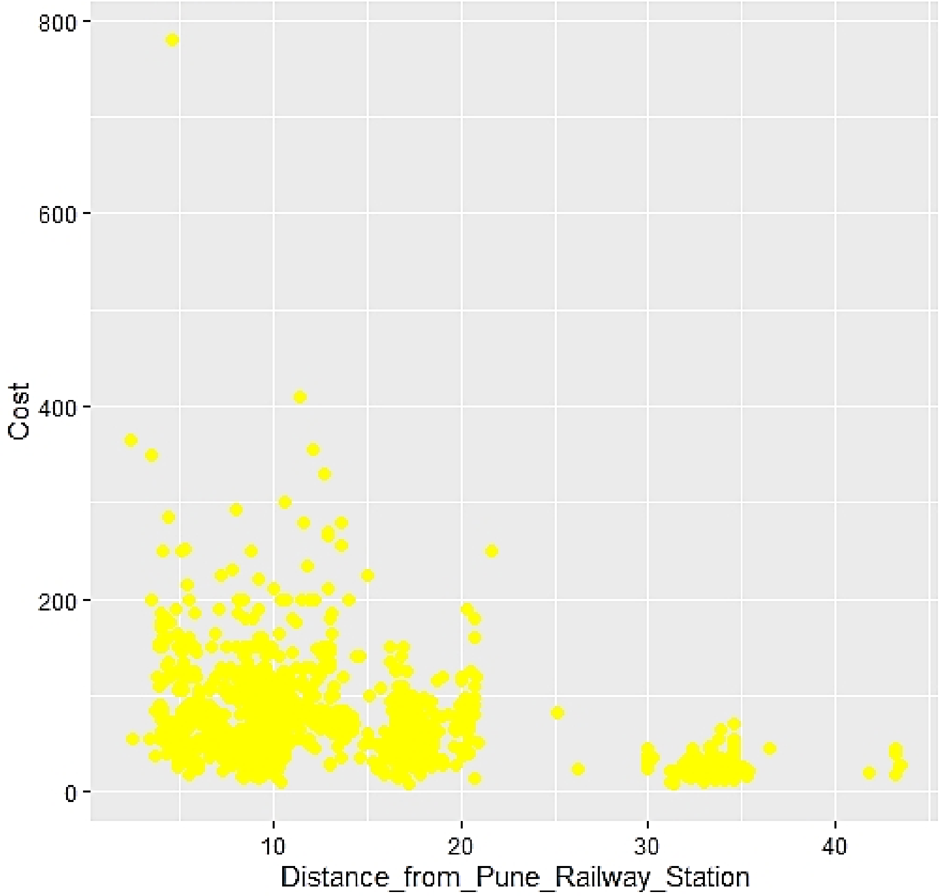
There is strong correlation between cost and build area of houses.



There is weak and negative correlation between the distance from Pune Airport and cost of the house



There is weak and negative correlation between the distance from Swargate Bus Station and cost of the house.



There is weak and negative correlation between the distance from Pune Railway Station and the cost of the house.

CORRELATION TESTS

COST	BUILD AREA	NO. OF BEDROOMS	NO. OF BATHROOMS	NO. OF BALCONIES	DISTANCE FROM AIR PORT	DISTANCE FROM RAILWAY STATION	DISTANCE FROM BUS STATION
r	0.8314197	0.6790053	0.6853033	0.1714248	-0.3504131	-0.4268997	-0.4110883
t STATISTIC	51.791	32.013	32.557	6.0225	-12.95	-16.34	-15.609
DF	1198	1198	1197	1198	1198	1198	1198
P VALUE	< 2.2e-16	< 2.2e-16	< 2.2e-16	2.281e-09	< 2.2e-16	< 2.2e-16	< 2.2e-16
DECISION	positive correlation	positive correlation	positive correlation	positive correlation	negative correlation	negative correlation	negative correlation

Since all the p values are less than 0.05 l.o.s , on the basis of correlation t test we conclude that built up area , no. of bedrooms, bathrooms, balconies ,distance from the airport ,railway station, bus station are linearly correlated with cost of houses.

CHI SQ TESTS TO CHECK ASSOCIATION BETWEEN COST AND CATEGORICAL VARIABLES

Variable ->	**Location**	**Location Type**	**Age Of Property**	**Furnishing**	**Power Backup**	**Parking**
χ^2	639.97	118.3	69.116	75.129	242.73	99.432
Degrees of freedom	66-	6	36	6	6	18
p-value	< 2.2e-16	< 2.2e-16	0.000742	3.611e-14	< 2.2e-16.	2.814e-13
cramer's V test statistic	0.298135	0.2220173	0.1385601	0.1769287	0.3180212	0.1661927
Null hypothesis rejected/accepted	rejected	rejected	rejected	rejected	rejected	rejected
Associated/not associated	**Associated**	**Associated**	**Associated**	**Associated**	**Associated**	**Associated**

From the above test it is clear that cost and location,location type , age of property, furnishing, power backup and parking is associated with each other.

EIGENVALUE ANALYSIS

Lambda	Eigen Values	Condition Indices	Values
λ_1	3.87945	K_1	1.000
λ_2	3.21819	K_2	1.205
λ_3	1.33939	K_3	2.896
λ_4	1.11541	K_4	3.478
λ_5	0.91570-	K_5	4.237
λ_6	0.87508	K_6	4.433
λ_7	0.72453	K_7	5.354
λ_8	0.64840	K_8	5.983
λ_8	0.48754	K_9	7.957
λ_{10}	0.35153	K_{10}	11.036
λ_{11}	0.21419	K_{11}	18.112
λ_{12}	0.14406	K_{12}	26.930
λ_{13}	0.08224	K_{13}	47.174
λ_{14}	0.00429	K_{14}	903.885

CONDITION NUMBER = 903.885

From this it is clear that condition number is **greater** than 100 which indicates **strong** multicollinearity in the data set.

Hence, Regressors are dependent.

MULTIPLE REGRESSION MODEL

1. ALL factors (complete data: n=1000, with INTERCEPT)

Regression Equation

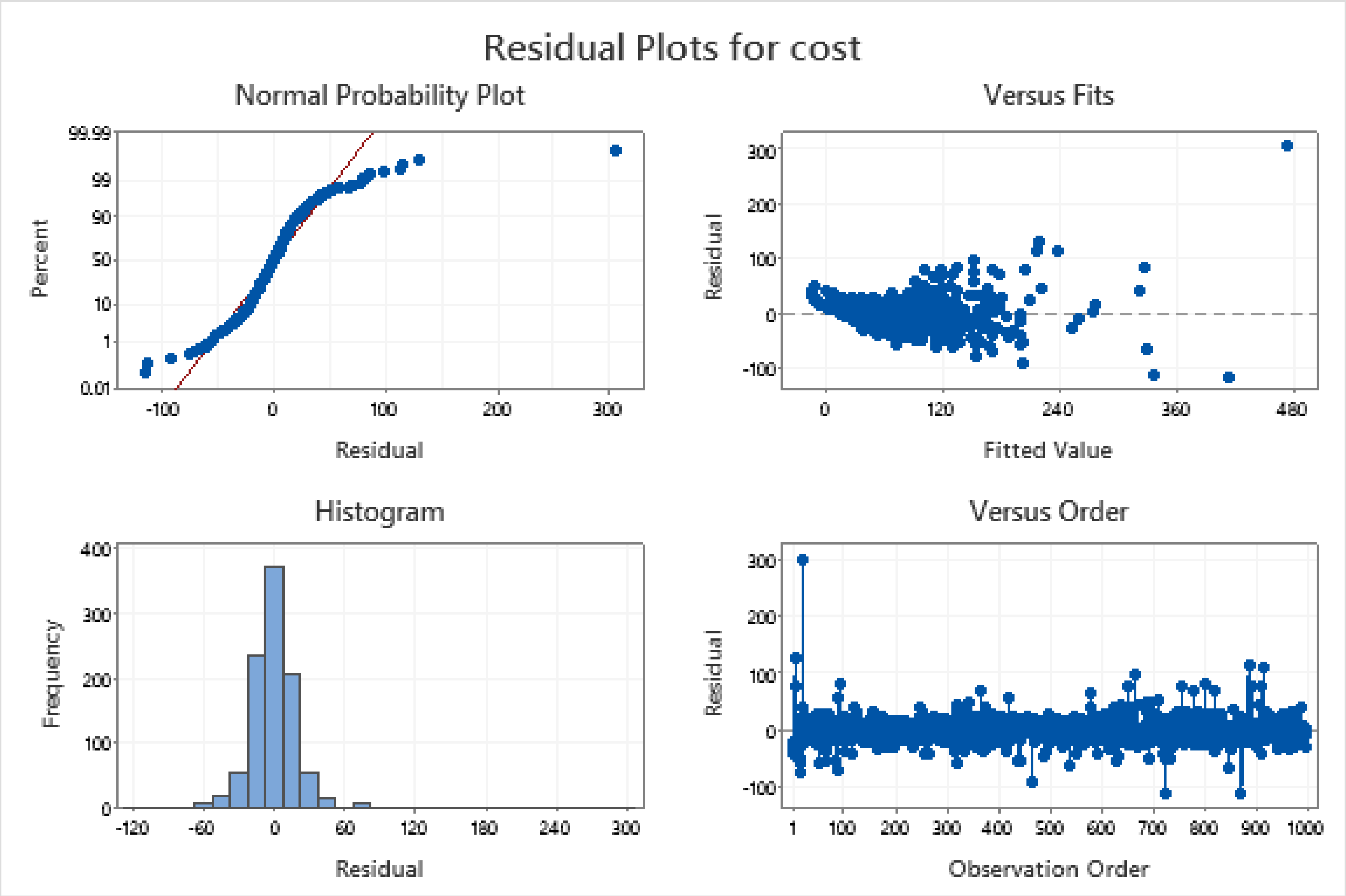
cost(in lakhs) = 89.34 + 0.07987 Build area (sq ft) - 2.715 no. of balconies + 1.030 total
+ 3.96 No. of bedrooms - 2.467 dis bus stop + 0.0 locode_0 - 75.10 locode_1
- 87.12 locode_2 - 44.3 locode_3 - 63.16 locode_4 - 63.74 locode_5
- 84.80 locode_6 - 68.40 locode_7 - 60.67 locode_8 - 52.20 locode_9
- 63.02 locode_10 - 67.76 locode_11 + 0.0 aopcode_0 + 1.57 aopcode_1
- 7.17 aopcode_2 - 6.85 aopcode_3 + 0.0 parking_0 - 0.43 parking_1
+ 8.26 parking_2 + 4.30 parking_3

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	89.34	8.80	10.16	0.000	
Build area (sq ft)	0.07987	0.00261	30.65	0.000	2.94
no. of balconies	-2.715	0.969	-2.80	0.005	1.30
total	1.030	0.229	4.50	0.000	1.55
No. of bedrooms	3.96	1.60	2.47	0.014	2.98
dis bus stop	-2.467	0.578	-4.27	0.000	63.53
locode					
1	-75.10	6.97	-10.77	0.000	6.65
2	-87.12	7.03	-12.39	0.000	3.69
3	-44.3	16.1	-2.74	0.006	72.05
4	-63.16	8.23	-7.67	0.000	3.54
5	-63.74	7.99	-7.98	0.000	12.78
6	-84.80	6.43	-13.20	0.000	6.91
7	-68.40	7.24	-9.45	0.000	5.53
8	-60.67	6.20	-9.78	0.000	5.86
9	-52.20	6.62	-7.88	0.000	3.69
10	-63.02	6.72	-9.38	0.000	5.91
11	-67.76	8.86	-7.65	0.000	10.50
aopcode					
1	1.57	2.57	0.61	0.541	1.55
2	-7.17	2.30	-3.11	0.002	1.88
3	-6.85	2.59	-2.64	0.008	2.68
parking					
1	-0.43	3.44	-0.13	0.900	2.57
2	8.26	3.61	2.29	0.022	2.39
3	4.30	2.97	1.45	0.148	3.48

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
23.9840	81.86%	81.45%	79.53%



We observe that in stepwise selection procedure number of bathrooms, distance from railway station and distance from airport factors were eliminated because of their insignificance at 0.05 level of significance. VIF of distance from bus stop is high. Also normality and constant variance of errors assumption is violated as seen from the above plots.

Model(without INTERCEPT)

Regression Equation

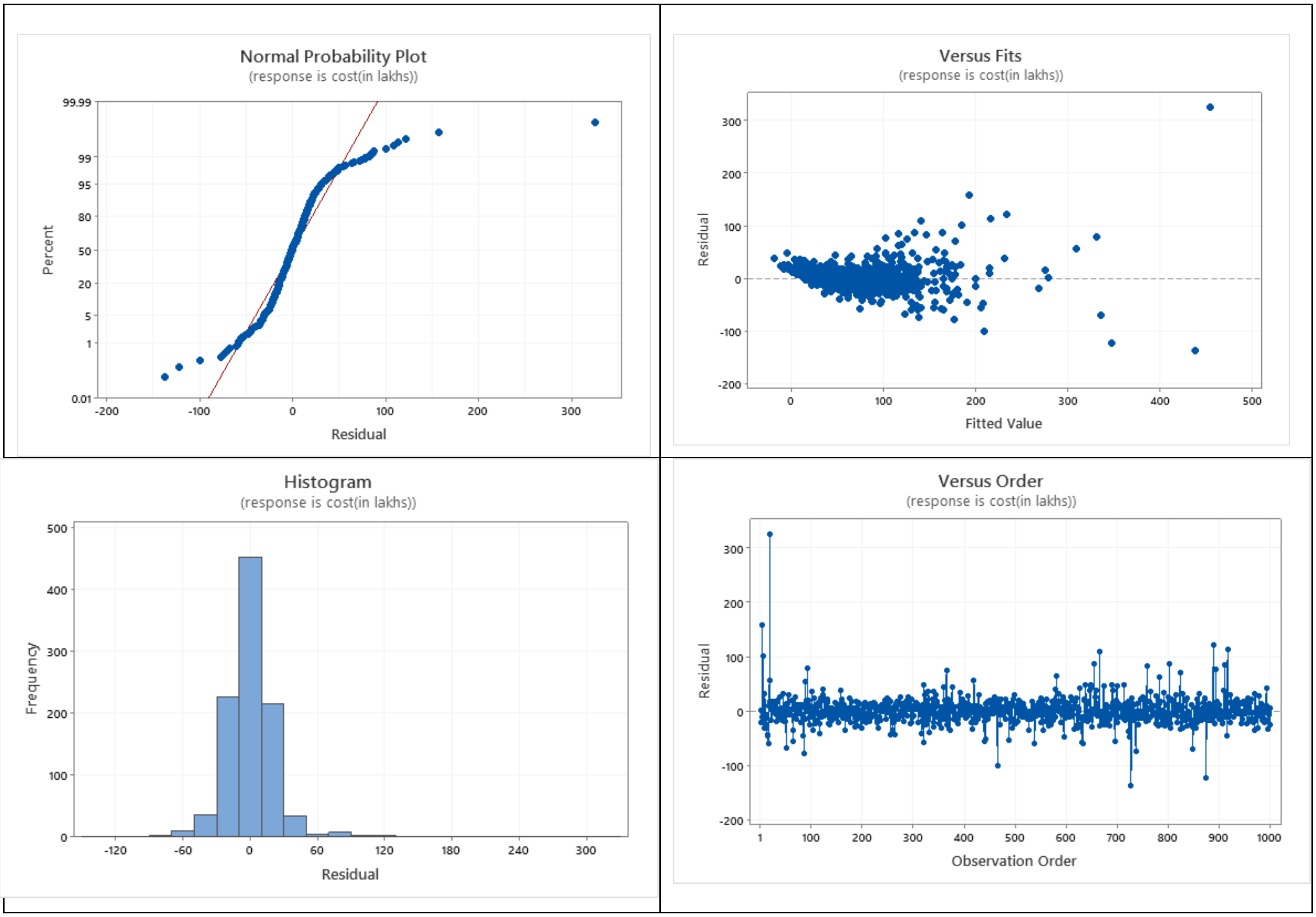
cost(in lakhs) = 0.08212 Build area (sq ft) + 0.785 total + 3.51 no. of bathrooms
+ 2.211 dis airport - 7.76 dis railway st + 5.32 dis bus stop + 0.0 locode_0
- 43.55 locode_1 - 30.60 locode_2 - 37.5 locode_3 + 4.6 locode_4
- 38.66 locode_5 - 52.64 locode_6 + 0.5 locode_7 - 11.44 locode_8
- 19.35 locode_9 - 36.00 locode_10 - 48.7 locode_11 + 0.0 aopcode_0
+ 3.10 aopcode_1 - 4.00 aopcode_2 - 0.80 aopcode_3 + 0.0 parking_0
+ 3.29 parking_1 + 12.10 parking_2 + 8.52 parking_3

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Build area (sq ft)	0.08212	0.00270	30.44	0.000	13.91
total	0.785	0.226	3.47	0.001	3.90
no. of bathrooms	3.51	1.72	2.05	0.041	18.99
dis airport	2.211	0.744	2.97	0.003	337.99
dis railway st	-7.76	1.30	-5.99	0.000	948.51
dis bus stop	5.32	1.10	4.84	0.000	772.65
locode					
1	-43.55	7.73	-5.64	0.000	8.35
2	-30.60	7.60	-4.03	0.000	4.23
3	-37.5	18.6	-2.02	0.044	112.10
4	4.6	10.2	0.45	0.654	5.24
5	-38.66	9.84	-3.93	0.000	20.95
6	-52.64	5.93	-8.88	0.000	6.17
7	0.5	10.3	0.05	0.958	11.25
8	-11.44	9.90	-1.16	0.248	15.45
9	-19.35	7.51	-2.58	0.010	4.68
10	-36.00	9.11	-3.95	0.000	11.07
11	-48.7	11.6	-4.21	0.000	18.32
aopcode					
1	3.10	2.64	1.18	0.240	1.82
2	-4.00	2.35	-1.70	0.090	2.56
3	-0.80	2.58	-0.31	0.756	3.84
parking					
1	3.29	3.53	0.93	0.352	2.95
2	12.10	3.71	3.26	0.001	2.69
3	8.52	3.04	2.80	0.005	9.85

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
24.8026	92.70%	92.53%	91.84%



We observe that R^2 is greater for no intercept model. Also if the values of all factors are zero it is expected that the value of cost be zero too. Hence, we consider NO intercept model for further analysis. Stepwise regression included some factors that were not included in the with intercept model but VIF's of few factors are high indicating distances factors might be highly correlated to each other while build area and number of bathrooms may be correlated with each other. Also we observed that the normality assumption of errors is not satisfied as clearly seen in residual plot. This indicates that the underlying distribution is light-tailed. The above figure has flattening at the extremes for the curves. The assumption of constant variance of errors isn't satisfied either. A nonlinear trend as well as outward opening funnel pattern is seen in the residuals versus fitted values plot. We try removing few outliers to correct non normality and the required factors to remove multicollinearity.

Model after removing **Outliers and factors with high VIF**

Regression Equation

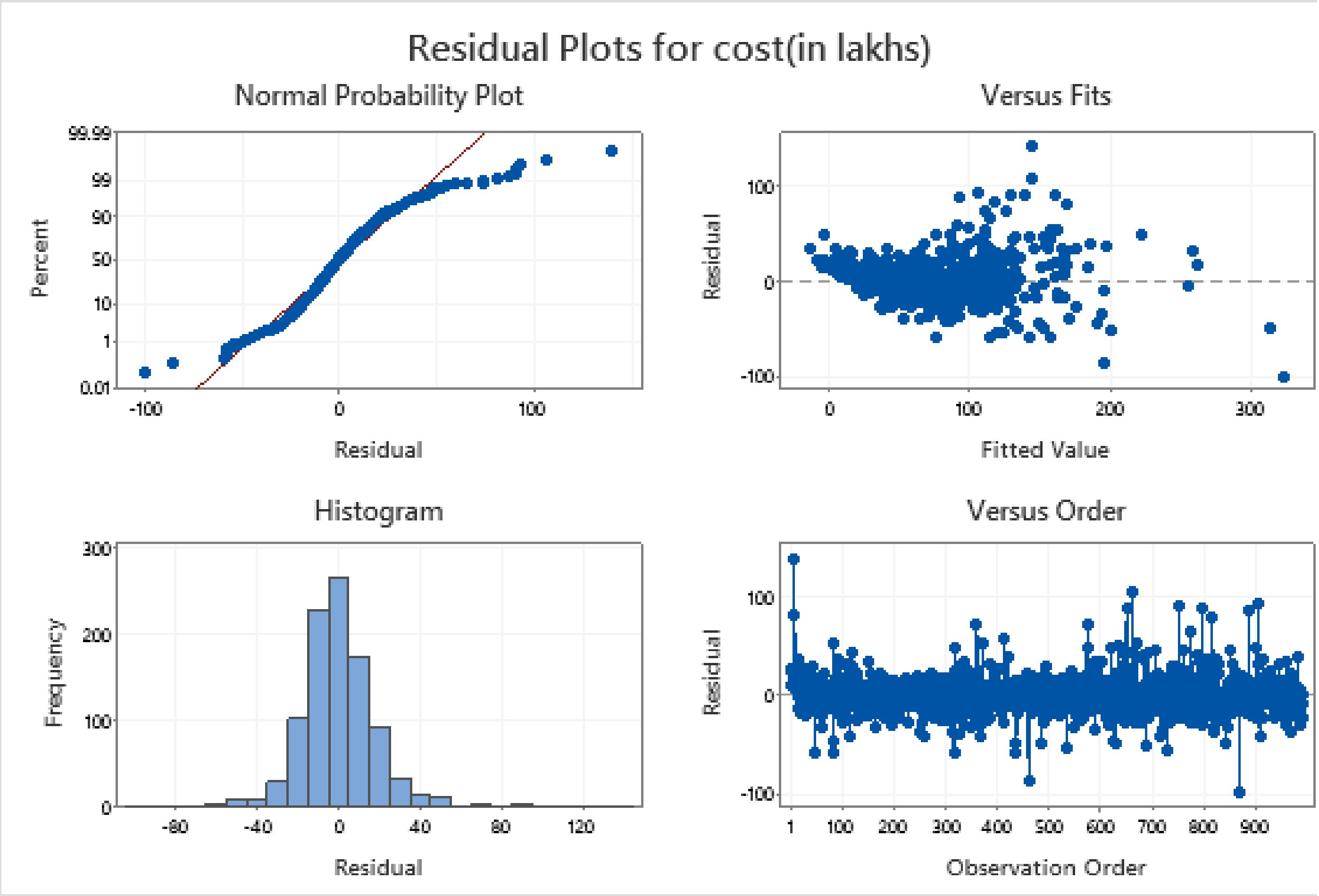
cost(in lakhs) = 0.07716 Build area (sq ft) + 0.936 total + 0.0 locode_0 - 24.02 locode_1 - 17.07 locode_2 - 43.66 locode_3 + 12.95 locode_4 - 25.63 locode_5 - 27.62 locode_6 + 4.91 locode_7 + 3.67 locode_8 + 2.89 locode_9 - 12.82 locode_10 - 28.38 locode_11 + 0.0 aopcode_0 + 1.73 aopcode_1 - 3.12 aopcode_2 + 2.44 aopcode_3 + 0.0 furnishing(yes/no/semi)_0 + 6.88 furnishing(yes/no/semi)_1 + 3.32 furnishing(yes/no/semi)_2 + 0.0 parking_0 + 2.25 parking_1 + 12.79 parking_2 + 10.13 parking_3

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Build area (sq ft)	0.07716	0.00160	48.22	0.000	6.79
total	0.936	0.184	5.10	0.000	3.83
locode					
1	-24.02	3.63	-6.61	0.000	2.74
2	-17.07	4.03	-4.24	0.000	1.78
3	-43.66	2.93	-14.91	0.000	4.17
4	12.95	4.72	2.74	0.006	1.69
5	-25.63	3.30	-7.76	0.000	3.55
6	-27.62	3.39	-8.16	0.000	3.02
7	4.91	3.75	1.31	0.191	2.24
8	3.67	3.42	1.07	0.284	2.74
9	2.89	4.08	0.71	0.479	2.07
10	-12.82	3.74	-3.43	0.001	2.74
11	-28.38	3.63	-7.81	0.000	2.71
aopcode					
1	1.73	2.13	0.81	0.417	1.77
2	-3.12	1.90	-1.64	0.101	2.49
3	2.44	2.03	1.20	0.231	3.57
furnishing(yes/no/semi)					
1	6.88	2.21	3.11	0.002	1.50
2	3.32	1.50	2.21	0.027	2.18
parking					
1	2.25	2.83	0.80	0.426	2.85
2	12.79	2.98	4.29	0.000	2.56
3	10.13	2.42	4.19	0.000	9.26

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
20.2393	94.21%	94.08%	93.89%



- We observe that the non-normality is still seen in the residual plot.
- The residuals versus fitted values show same trend as earlier models .i.e. outward opening funnel pattern and nonlinear trend.
- We also see that the VIF of factors in not as high as before. Hence no serious multicollinearity problem in the data remains.
- We further apply transformations on the dependent variable i.e. cost prices to deal with non-normality and non-constant variance of errors.

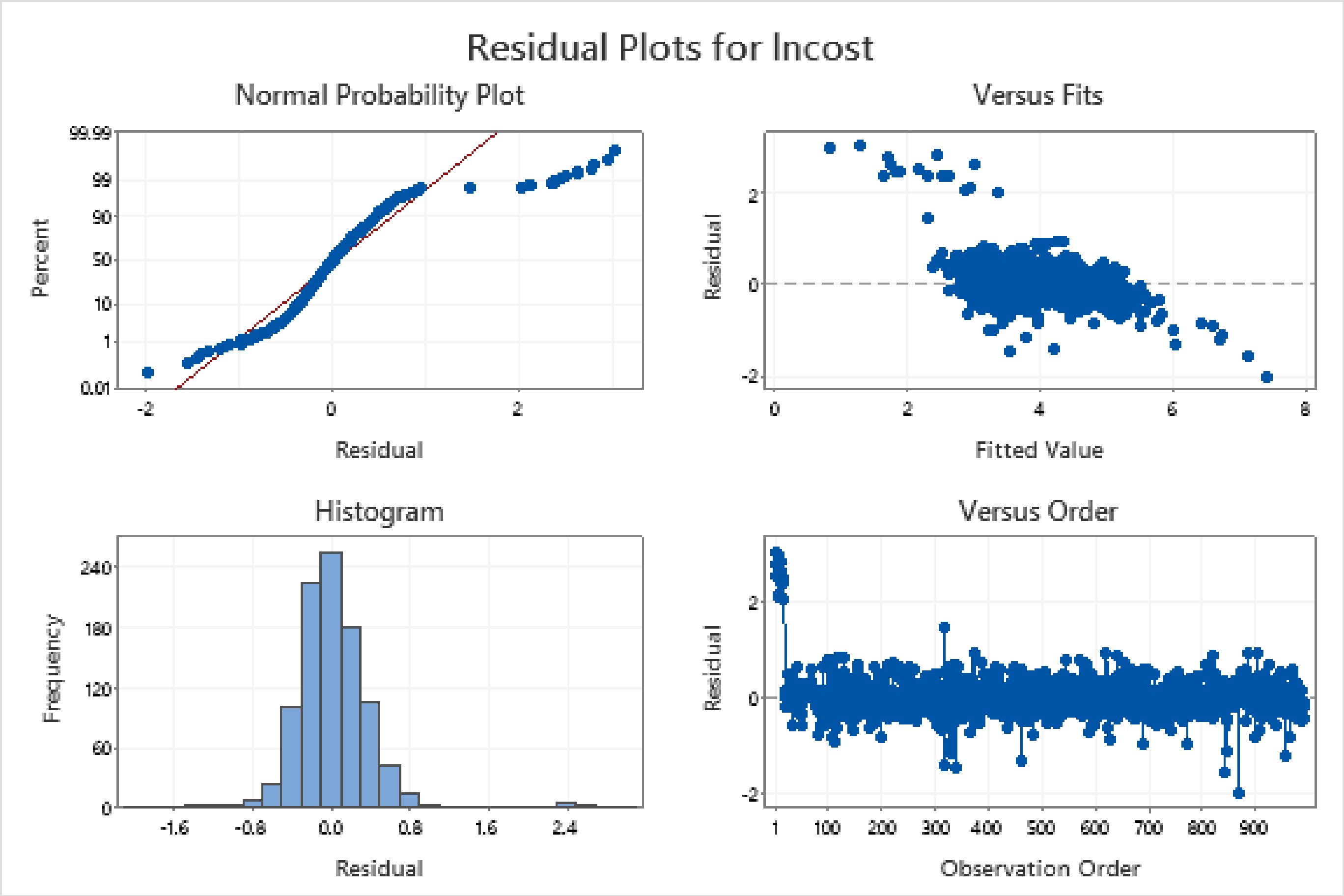
Ln transformation

Regression Equation

Incost = 0.001040 Build area (sq ft) + 0.0792 no. of balconies + 0.02414 total + 0.0 locode_0
+ 2.1490 locode_1 + 2.1662 locode_2 + 1.6673 locode_3 + 2.553 locode_4 + 2.0245 l
+ 2.0393 locode_6 + 2.3942 locode_7 + 2.4021 locode_8 + 2.3058 locode_9
+ 2.1652 locode_10 + 2.1324 locode_11 + 0.0 aopcode_0 + 0.1581 aopcode_1
+ 0.1780 aopcode_2 + 0.4657 aopcode_3 + 0.0 furnishing(yes/no/semi)_0
+ 0.1400 furnishing(yes/no/semi)_1 + 0.1428 furnishing(yes/no/semi)_2 + 0.0 parking
+ 0.2969 parking_1+ 0.5071 parking_2 + 0.4868 parking_3

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.472114	98.69%	98.66%	98.61%

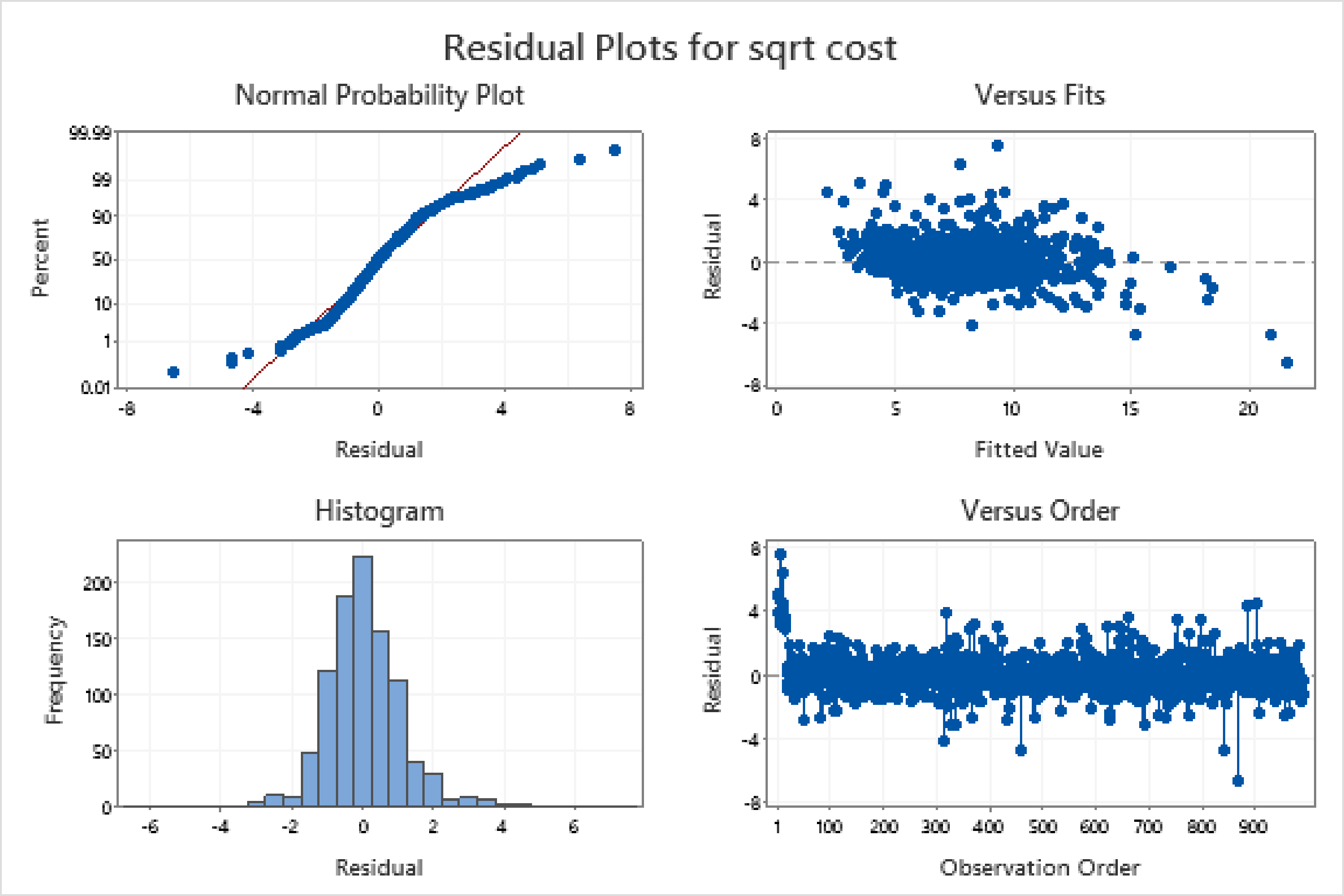


We observe that the non-normality is still seen in the residual plot. It is still heavy tailed. The residuals versus fitted values show a double bow pattern in the middle indicating non-constant variance.

SQRT TRANSFORMATION

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.20953	97.93%	97.89%	97.81%

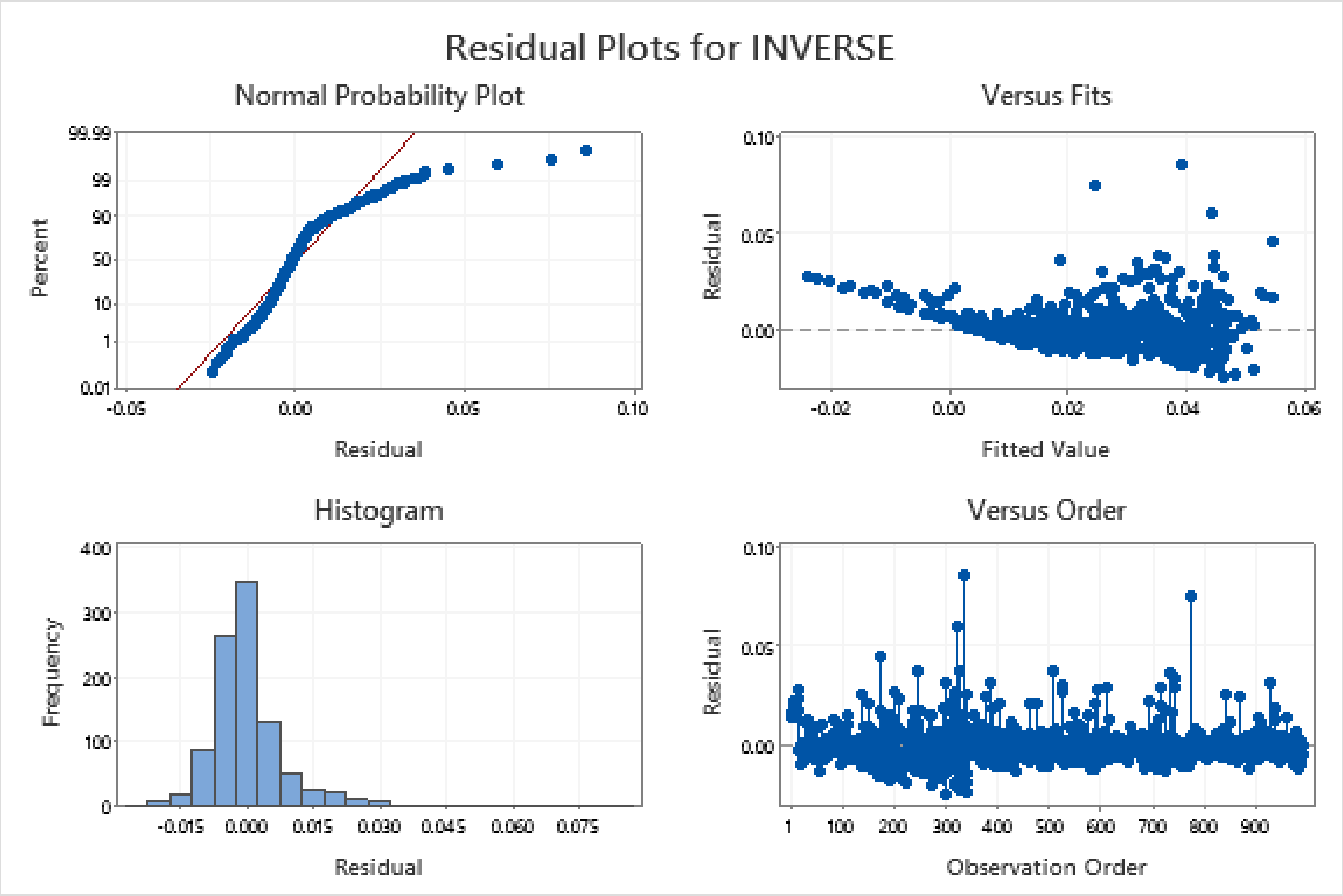


We observe that the non-normality is still seen in the residual plot. It is still heavy tailed. The residuals versus fitted values show a double bow pattern in the middle indicating non-constant variance.

INVERSE TRANSFORMATION

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0095415	87.74%	87.47%	87.23%



We observe that the non-normality is still seen in the residual plot. It is still heavy tailed. The residuals versus fitted values show a outward funnel pattern indicating non-constant variance. Also the histogram of residuals show a positively skewed distribution.

BOXCOX TRANSFORMATION

To find value of λ so that cost$^{\lambda}$ can be calculated and can be used as new response. Here value of **λ is** 0.571867(* this value is automatically calculated by minitab and is the optimal value of λ.)

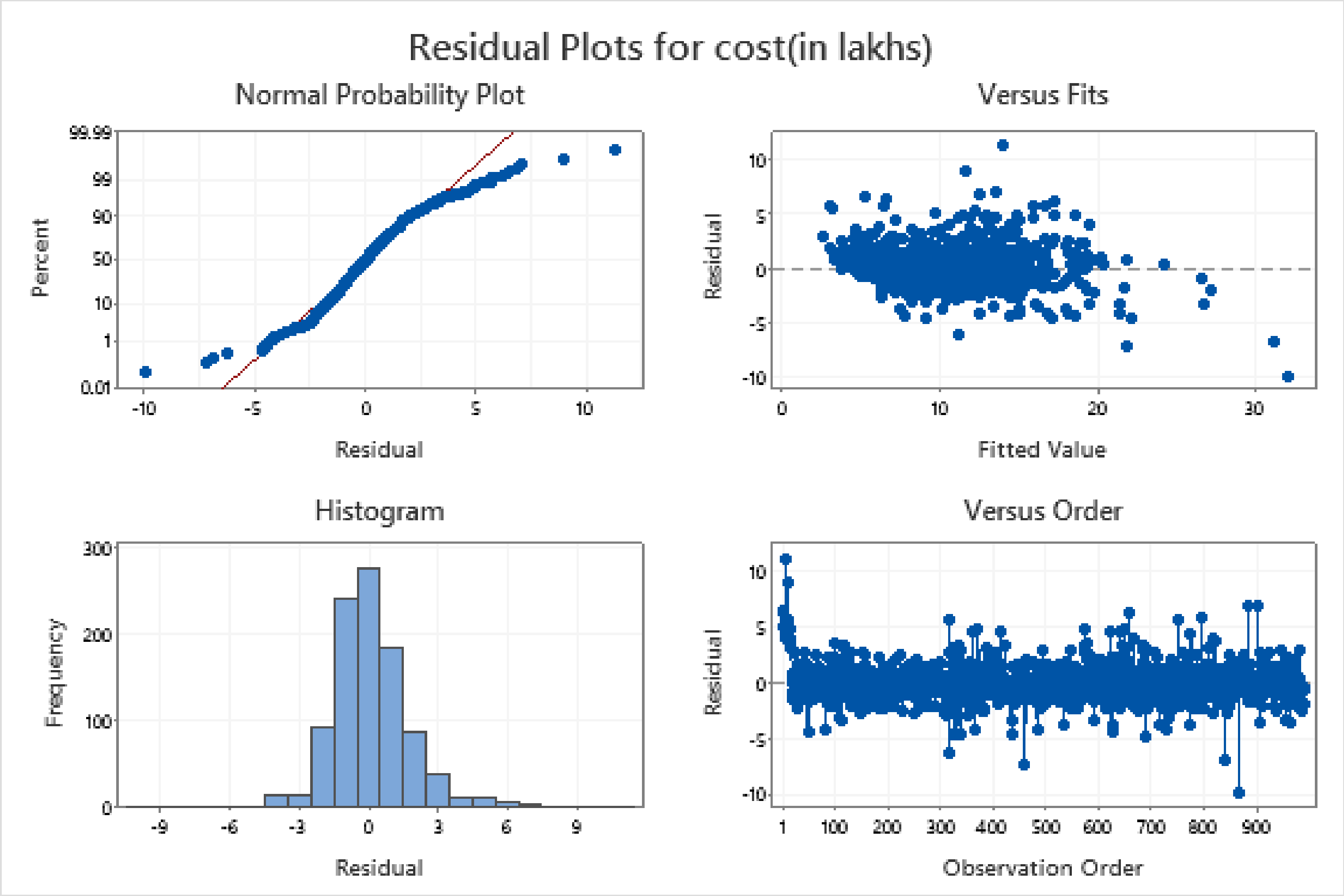
Rounded λ	0.571867
Estimated λ	0.571867
95% CI for λ	(0.530367, 0.614367)

Regression Equation

cost(in lakhs)^0.571867 = 0.006464 Build area (sq ft) + 0.2014 no. of balconies+ 0.1080 total
+ 0.0 locode_0 + 1.908 locode_1 + 2.239 locode_2- 0.532 locode_3
+ 4.982 locode_4 + 1.557 locode_5 + 1.390 locode_6+ 4.149 locode_7
+ 4.054 locode_8+ 3.850 locode_9+ 2.601 locode_10+ 1.563 locode_11
+ 0.0 aopcode_0+ 0.275 aopcode_1+ 0.063 aopcode_2
+ 0.877 aopcode_3+ 0.0 furnishing(yes/no/semi)_0
+ 0.692 furnishing(yes/no/semi)_1+ 0.476 furnishing(yes/no/semi)_2
+ 0.0 parking_0 + 0.515 parking_1+ 1.644 parking_2 + 1.427 parking_3
+ 0.0 power backup_0+ 0.023 power backup_1 - 0.033 power backup_2

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
1.80415	97.58%	97.52%	97.43%



We observe that the non-normality is still seen in the residual plot but it is better than the other above plots. The residuals versus fitted values still shows non-constant variance. This might be due to the non-linearity observed in the data set. So the best regression equation is given by-

cost(in lakhs)^0.571867 = 0.006464 Build area (sq ft) + 0.2014 no. of balconies+ 0.1080 total
+ 0.0 locode_0 + 1.908 locode_1 + 2.239 locode_2- 0.532 locode_3
+ 4.982 locode_4 + 1.557 locode_5 + 1.390 locode_6+ 4.149 locode_7
+ 4.054 locode_8+ 3.850 locode_9+ 2.601 locode_10+ 1.563 locode_11
+ 0.0 aopcode_0+ 0.275 aopcode_1+ 0.063 aopcode_2
+ 0.877 aopcode_3+ 0.0 furnishing(yes/no/semi)_0
+ 0.692 furnishing(yes/no/semi)_1+ 0.476 furnishing(yes/no/semi)_2
+ 0.0 parking_0 + 0.515 parking_1+ 1.644 parking_2 + 1.427 parking_3
+ 0.0 power backup_0+ 0.023 power backup_1 - 0.033 power backup_2

CONCLUSIONS

1) The distribution of cost is **not normal** and is positively skewed.
2) The cost and build up area are **strongly correlated (positive)** but we observe slight non linear nature from the scatter plots.
3) There is **positive correlation** between cost and number of bedrooms, bathrooms and balconies.
4) There is **negative correlation** between cost and the distance of the house from airport ,railway station (Pune junction) and bus station (Swargate).
5) Cost and location are **strongly associated**. There is significant difference in the cost of homes of similar built up areas at different locations.
6) Koregaon Park has the costliest homes and Chakan cheapest. We observe high variability in cost of homes at Koregaon Park but the median price is much higher than prices at all other locations.
7) Cost and location type are also **strongly associated**. Residential areas have higher prices than industrial and commercial areas. Also average cost of house **differs significantly** for all location types.
8) Cost and Age are **moderately associated**. Cost reduces as age increases though location still remains a prominent factor affecting the prices of a home. Also average cost of house differs significantly for all age groups.
9) The average price of a house of area about 958.29 sq ft is 70.27673 lakhs (according to our sample).
10) The average **cost differs significantly** for all furnishing types.
11) The most suitable model we could build is –

cost(in lakhs)^0.568912 = 0.006324 Build area (sq ft) + 0.1962 no. of balconies
+ 0.1023 total + 0.0 locode_0 + 2.142 locode_1 + 2.464 locode_2
- 0.269 locode_3 + 5.183 locode_4 + 1.805 locode_5 + 1.637 locode_6
+ 4.350 locode_7 + 4.252 locode_8 + 4.069 locode_9
+ 2.865 locode_10 + 1.804 locode_11 + 0.0 aopcode_0
+ 0.228 aopcode_1 + 0.033 aopcode_2 + 0.779 aopcode_3
+ 0.0 furnishing(yes/no/semi)_0 + 0.697 furnishing(yes/no/semi)_1
+ 0.427 furnishing(yes/no/semi)_2 + 0.0 parking_0 + 0.389 parking_1
+ 1.523 parking_2 + 1.296 parking_3 + 0.0 power backup_0
- 0.030 power backup_1 - 0.031 power backup_2

As obtained by the Box Cox method.

TESTING THE MODEL

COST	PREDICTED VALUE	C.I. LL	C.I. UL		COST	PREDICTED VALUE	C.I. LL	C.I. UL
160.00	74.594	67.989	81.456		30.00	21.319	17.901	24.987
250.00	179.828	167.288	192.750		14.00	16.291	13.065	19.812
150.00	77.371	70.811	84.175		23.00	21.225	18.503	24.104
86.00	30.651	26.517	35.035		25.00	25.361	22.338	28.545
160.00	54.311	47.569	61.427		24.00	16.947	13.890	20.258
67.00	67.417	61.906	73.125		22.75	38.890	35.030	42.919
102.00	92.484	85.344	99.866		32.00	28.016	24.899	31.286
58.00	55.172	50.212	60.328		13.50	7.976	6.225	9.907
82.00	83.361	75.983	91.027		25.00	38.846	36.042	41.738
30.00	42.520	37.081	48.271		70.00	68.779	62.621	75.179
42.00	41.956	36.630	47.584		24.00	24.061	21.679	26.548
175.00	136.976	128.205	145.990		24.00	17.717	13.996	21.802
62.00	63.182	57.497	69.092		55.00	81.026	74.660	87.611
25.00	21.405	17.260	25.921		30.00	18.558	15.634	21.692
255.00	251.988	238.838	265.435		27.00	26.939	23.653	30.404
77.00	69.612	64.162	75.249		15.00	18.541	16.443	20.745
35.00	29.053	25.415	32.894		15.50	15.233	12.697	17.962
61.00	79.215	73.982	84.600		29.00	28.925	25.187	32.880
85.00	61.277	56.458	66.261		18.50	13.257	10.636	16.118
60.00	60.725	55.324	66.338		35.00	48.660	44.259	53.237
280.00	269.356	256.053	282.943		15.00	10.879	8.613	13.366
179.00	206.274	194.069	218.792		21.50	23.527	21.073	26.095
120.00	114.624	108.137	121.271		25.00	24.058	21.153	27.119
121.00	113.600	104.221	123.319		30.00	15.570	12.401	19.040
105.00	61.393	54.832	68.266		30.00	14.713	12.286	17.322
34.00	44.600	38.963	50.556		35.00	40.378	36.626	44.283
52.00	46.679	40.632	53.078		30.00	40.098	36.359	43.991
55.00	57.974	50.704	65.651		45.00	8.329	6.536	10.303
85.00	126.109	115.980	136.594		185.00	108.525	98.809	118.624
32.00	43.070	37.686	48.755		58.00	55.482	48.497	62.862
28.50	57.259	52.981	61.677		50.00	65.434	58.621	72.562
21.00	40.348	37.027	43.788		170.00	153.029	142.129	164.269

43.00	50.836	45.730	56.169		35.00	37.639	34.265	41.145
26.00	32.971	29.178	36.959		62.00	48.821	44.864	52.918
12.50	22.358	19.842	24.999		65.00	61.934	56.833	67.220
18.00	9.948	7.471	12.717		48.00	76.397	71.391	81.546
19.00	18.332	16.287	20.480		99.00	122.328	115.541	129.277
35.00	28.994	26.499	31.583		94.00	78.458	72.116	85.025
15.00	14.357	11.240	17.789		77.00	73.762	69.197	78.451
15.00	16.748	13.992	19.710		84.00	70.356	65.432	75.431
39.00	52.951	48.679	57.375		55.00	33.394	29.017	38.029
9.50	5.219	3.155	7.706		37.00	30.467	27.125	33.972
45.00	43.441	40.005	46.997		37.00	43.098	38.185	48.261
32.00	29.579	26.912	32.352		34.00	50.795	46.420	55.336
21.00	19.379	16.650	22.281		75.00	68.580	63.742	73.567
18.00	5.551	3.744	7.650		62.00	57.593	52.158	63.254
8.00	6.680	4.823	8.788		50.00	80.031	75.423	84.754
45.00	70.899	65.071	76.937		47.00	50.807	45.911	55.912
22.00	20.617	18.201	23.159		30.00	40.871	36.738	45.189
32.00	40.753	37.280	44.355		150.00	91.593	84.849	98.553
75.00	67.270	62.381	72.313		45.00	63.965	59.454	68.615
55.00	57.451	52.183	62.930		150.00	125.142	117.766	132.707
40.00	27.385	23.659	31.339		35.00	49.230	45.117	53.494
37.00	41.684	37.315	46.255		80.00	76.705	71.914	81.626
65.00	71.949	65.529	78.621		58.00	49.052	44.793	53.475
40.00	31.448	28.239	34.803		50.00	48.783	43.792	53.999
29.00	45.560	41.605	49.666		60.00	46.852	42.801	51.056
49.00	48.073	43.530	52.806		200.00	155.413	147.644	163.349
17.00	25.502	22.493	28.670		125.00	96.472	88.447	104.790
60.00	61.000	57.071	65.039		62.00	68.519	62.606	74.656
60.00	47.376	42.557	52.411		58.00	53.743	48.350	59.376
39.50	34.849	31.036	38.847		80.00	91.533	84.411	98.899
27.00	24.467	20.950	28.212		75.00	77.666	70.929	84.660
50.00	58.077	54.138	62.132		93.00	88.339	82.145	94.722
67.00	63.962	58.084	70.077		89.00	98.502	90.823	106.443
50.00	58.004	53.304	62.871		69.00	71.702	66.162	77.429

57.00	46.492	41.848	51.342		82.00	81.546	75.930	87.332
100.00	108.305	102.204	114.556		85.00	100.831	94.029	107.833
20.00	19.356	16.429	22.484		70.00	70.596	64.950	76.441
8.50	7.822	5.815	10.076		100.00	113.600	107.158	120.201
37.00	37.113	33.565	40.811		130.00	139.901	132.491	147.481
55.00	64.555	60.130	69.112		100.00	115.701	108.070	123.552
55.00	44.442	39.980	49.102		85.00	90.117	83.545	96.898
25.00	32.399	27.985	37.083		125.00	100.440	93.500	107.589
91.00	95.931	89.335	102.726		59.25	68.323	62.336	74.540
90.00	91.187	84.817	97.752		145.00	161.454	151.545	171.627
52.00	43.829	39.152	48.728		85.00	90.176	84.350	96.166
35.00	43.686	38.672	48.955		70.00	80.082	72.943	87.502
90.00	88.960	83.008	95.085		150.00	107.782	100.061	115.744
60.00	27.149	23.632	30.870		76.00	74.347	68.857	80.015
72.00	62.773	58.346	67.337		70.00	82.196	75.100	89.560
32.00	36.557	32.832	40.450		149.00	130.168	122.841	137.673
101.00	98.215	91.878	104.730		199.00	271.800	258.747	285.125
45.00	43.824	39.360	48.488		98.00	118.370	111.179	125.750
42.00	43.488	38.630	48.587		65.00	62.322	57.143	67.690
100.00	80.791	75.062	86.697		41.00	21.804	17.756	26.199
28.00	16.280	12.987	19.883		38.00	29.620	25.362	34.155
45.00	40.637	36.019	45.489		36.00	33.330	29.370	37.499
16.00	19.354	16.232	22.704		53.00	64.205	58.936	69.662
42.00	33.004	28.228	38.092		59.00	66.455	61.295	71.790
165.00	199.694	188.993	210.642		75.00	65.581	60.199	71.156
150.00	115.212	107.136	123.534		51.00	52.995	48.010	58.187
38.00	50.522	45.364	55.914		40.00	53.112	47.903	58.547
85.00	110.880	104.110	117.829		120.00	88.317	82.147	94.675
64.00	88.765	81.760	96.012		160.00	149.001	140.375	157.843
90.00	81.952	75.386	88.749		80.00	41.118	35.967	46.558
55.00	80.507	74.523	86.685		65.00	50.790	45.170	56.686
53.50	48.394	43.288	53.740		70.00	79.966	73.790	86.351
65.00	87.852	80.866	95.082		45.00	44.293	39.652	49.149
95.00	66.683	61.686	71.844		80.00	76.428	70.691	82.354

SCOPE

Throughout the project we have tried to build a regression model which will predict the price of a house considering various factors that actually affect the cost of house. Such a model can be developed into a software / application and can be made available at various online property sites like no broker so that all sellers and buyers can use it to get an idea of the cost at which they can sell or buy a house. A model like this can be used by potential buyers to get an idea of the approximate cost they would need to pay for houses at certain areas. House buyers can also understand how the presence of certain amenities like pool, club house and also the age of the property affects the price of a house. Having a model like this will help people easily find their dream home.

LIMITATIONS

1. This data as earlier stated is collected from a website open for sellers to put their property on sale on the website and hence has a possibility of data manipulation to attract buyers consequently affecting our validity of true data.
2. There is a possibility of Non-sampling error during data collection from the given website as all 1200 observation was noted manually in excel.
3. Since numbers of observations to be taken from each location was decided using Proportional Allocation, it is assumed that the number of properties for sale in each location on the given website was constant during the data collection.
4. The locations considered in this project are among the top locations of each category namely Residential, Commercial and Industrial areas in pune, according to information that was available on the internet and as per best of our knowledge hence may be inaccurate.
5. No data was available for some observations of flooring type factor and so couldn't be included in the regression model. Analysis of flooring type is done on the available data.

REFERENCES

1. Introduction to Linear Regression analysis (Fifth Edition) by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining
2. https://www.nobroker.in (for data collection)
3. https://blog.kohinoorpune.com/top-commercial-real-estate-locations-in-pune (top commercial real estate locations)
4. https://zolostays.com/blog/best-places-in-pune-to-live-a-comfortable-life/ (top residential real estate locations)
5. https://www.khedcity.com/pune-industrial-area/ (top industrial real estate locations)

APPENDIX

B

1) Box plot

```
d <- projectdatafinal
str(d)

d$cost <- as.numeric(d$cost)

library(ggplot2)

ggplot(d,aes(x = d$location , y = d$cost , fill = d$location)) + geom_boxplot(alpha = 0.3) + theme(legend.position = "none") + scale_fill_brewer(palette = "Blue") + theme_classic()
```

2) Box Cox transformation (using Minitab)

Stat -> Regression -> regression -> Fit regression model -> enter the response variable and the regressors -> click Options -> select the lambda for Box-Cox transformation -> click ok -> click OK.

C

1) Chi sq test

for age groups

```
> x = cbind(A1,A2,A3,A4)

> chisq.test(x)


     Pearson's Chi-squared test


data:  A1 and A2

X-squared = 69.116, df = 36, p-value =

0.000742
```

2) Cramer’s V test

```
> sum(x)

[1] 1200

> q = min(nrow(x),ncol(x))

> q

[1] 4

> y = sqrt(69.116/(sum(x)*(q-1)))

> y

[1] 0.1385601
```

3) Correlation test

```
cor.test(cost,ba)

        Pearson's product-moment correlation

data:  cost and ba
t = 51.791, df = 1198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8130855 0.8481061
sample estimates:
      cor
0.8314197
```

D

1) Density Plot

```
>cost=scan(‘clipboard’)

>cost

>plot(density(cost),main="Density cost",xlab="Cost",lwd=2,col="blue")
```

E

1) Eigen system analysis (using Minitab)

Stat -> Multivariate -> Factor analysis -> click Storage-> Enter the column to store the eigen values and select the correlation matrix -> click ok -> click OK.

Note – It is essential to find the correlation matrix and store it before performing the eigen system analysis. Stat -> Basic statistics -> Correlation -> Select the variables -> Options -> select the Method and check Store correlation matrix -> ok -> O

H

1) Histogram

```
v1=c(1.5,4,7.5,15)

v1

v=c(239,191,351,419)

v2=rep(v1,v)

v2

hist(v2,breaks=c(0,3,5,10,20),col=rainbow(10),main="Histogram of Age of property",xlab="Age of property")
```

2) Heat map (Correlation plot)

```
d = projectdatafinal

str(d)
d$cost <- as.numeric(d$cost)
d$ba <- as.numeric(d$ba)
d$bath <- as.numeric(d$bath)
d$balc <- as.numeric(d$balc)

install.packages(corrplot)
library(corrplot)

str(d1)

d2 = cor(d1)
d2

corrplot(d2,method ='color')
title("Heat map")
```

K

1) Kruskal Wallis test

```
res = scan('clipboard')
ind = scan('clipboard')
com = scan('clipboard')

l = list(res,ind,com)
kruskal.test(l)
```

L

1) Location type wise scatter plot

I. With all observations

```
cost <- scan('clipboard')
area <- scan('clipboard')
plot(cost,area)
d7 = data.frame(cost,area,location_type)

install.packages("ggplot2")
library(ggplot2)
ggplot(d7,aes(x = area,y = cost, color = location_type)) + geom_point()
```

II. With few highest observations removed

```
cost <- scan('clipboard')
area <- scan('clipboard')
plot(cost,area)
x <- which(cost<300)
d8 = data.frame(cost[x],area[x],location_type[x])
install.packages("ggplot2")
library(ggplot2)
ggplot(d8,aes(x = area[x],y = cost[x], color = location_type[x])) + geom_point()
```

M

1) Multiple linear regression model(using Minitab)

Stat -> Regression -> regression -> Fit regression model -> Select the response variable and the regressors -> click OK.

R

1) Residual plots for the fitted model (using Minitab)

Stat -> Regression -> regression -> Fit regression model -> Select the response variable and the regressors -> click Graphs -> select the required charts -> click ok -> click OK.

S

1) Scatter plot

```
library('ggplot2')

x<-scan("clipboard")

x

y<-scan("clipboard")

y

d1<-data.frame(x,y)

colnames(d1)<-c("Distance_from_Pune_Airport","Cost")

d1

ggplot(d1,aes(x=Distance_from_Pune_Airport,y=Cost))+geom_point(col="blue",size=1.5)+ggtitle("Scatter Plot")
```

2) Stepwise multiple linear regression

Stat -> Regression -> regression -> Fit regression model ->Select the response variable and the regressors -> click Stepwise -> select the required method-> click ok -> click OK.

3) Shapiro Wilk test

```
> y <- scan( ‘clipboard’ )

> shapiro.test(y)
```

T

1) Transforming the response variable then fitting the model

Stat -> Regression -> regression -> Fit regression model -> Select the transformed response variable and the regressors -> click Stepwise -> select the required method-> click ok -> click OK.

THE END