RAMAKRISHNA MISSION RESIDENTIAL COLLEGE (AUTONOMOUS)

NARENDRAPUR

STATISTICS PROJECT

HOUSING PRICES OF KOLKATA METROPOLITAN AREA

NAME: ARGHADEEP BASU

COLLEGE ROLL NO.: STUG/175/18

SEMESTER: 06

YEAR: 2020-21

SUBJECT: STATISTICS HONOURS

Acknowledgement

Many people have contributed directly and indirectly to my project work. I want to express my deepest gratitude to them.

I would like to express my sincere thanks and gratitude to my professors of the Department of Statistics, Ramakrishna Mission Residential College, Narendrapur, Dr Dilip Kumar Sahoo, Dr Parthasarathi Chakrabarti, Sri Tulsidas Mukherjee, Sri Palas Pal, Sri Subhadeep Banerjee for encouraging me to explore different topics for the project work and helping me whenever I faced any difficulty.

I would also like to thank my fellow classmates who helped me whenever I needed their help to complete my project.

Arghadeep Basu

Table of Contents:

INTRODUCTION

As we all know, due to rapid urban development in India, more and more people are flocking from the hinterlands to the various cities and urban conglomerates across the country in search for a better living condition. The metropolises and megapolises, which are already booming with people are no exception, either. Hence the real estate industry has flourished. All cities are expanding their city limits and new urban developments are being formed. Once sparsely populated suburbs are now gasping for space. The city of joy, Kolkata is not an exception, either. The Kolkata Metropolitan area, which has a population of more than 15 million, is continuously expanding its city limits.

Kolkata Metropolitan Area, also known as **Greater Kolkata**, is the urban agglomeration of the city of Kolkata in the Indian state of West Bengal. It is the third most populous metropolitan area in India after Delhi and Mumbai. The area is administered by the Kolkata Metropolitan Development Authority (KMDA). The area covers four municipal corporations along with 37 municipalities. Kolkata metropolitan district was legally defined in the schedule of the *Calcutta Metropolitan Planning Area (Use and Development of Land) Control Act, 1965* (West Bengal Act XIV of 1965), and, after repeal of that Act, redefined as Kolkata metropolitan area in the first schedule of *West Bengal Town and Country (Planning and Development) Act, 1979* (West Bengal Act XIII of 1979).

Jurisdiction

Municipal Corporations	Kolkata, Bidhannagar, Howrah, Chadannagar
Municipalities	1. North 24 Parganas district : Baranagar, Barasat, Barrackpore, Bhatpara, Dum Dum, Garulia, Halisahar, Kamarhati, Kanchrapara, Khardah, Madhyamgram, Naihati, New Barrackpore, North Barrackpur, North Dumdum, Panihati, South Dumdum, Titagarh 2. South 24 Parganas district : Baruipur, Budge Budge, Jaynagar Majilpur, Maheshtala, Pujali, Rajpur Sonarpur 3. Howrah district : Uluberia 4. Nadia district : Gayespur, Kalyani 5. Hooghly district : Baidyabati, Bhadreswar, Bansberia, Champdani, Dankuni, Hooghly-Chinsurah, Konnagar, Rishra, Serampore, Uttarpara Kotrung

According to the 2011 census data, the total population of the Kolkata metropolitan area was 14,112,536. KMDA report states the total area is 1,886.67 km², making the population density 7,480 per km².

OBJECTIVE

In this project I would like to eshtablish whether various factors which seem to predict the price of an apartment in and around the Kolkata Metropoliton Region can actually predict the price or not. In other words I would like to see how price of apartments actually depend on the different factors like area, location, whether the apartment is new or is it being resold, etc.

DATA AND PROBLEM DEFINITION

I have collected a dataset on housing prices of 6104 houses in and around the Kolkata Metropolitan Area from a website named Kaggle. The link for the dataset in given here →https://www.kaggle.com/ruchi798/housing-prices-in-metropolitan-areas-of-india?select=Kolkata.csv

The dataset comprises of columns like price of the apartment, area, location, number of bedrooms, whether the apartment is new or is it being resold, whether car parking ,24X7 security and lift are available or not available or there is nothing mentioned about their availability.

Variable to be explained: Price of the apartment (in Rs)

Explanatory factors: 1. Area of the apartment (in sq foot)

2. Number of Bedrooms (1,2,3,4)

3. Location (Name of place)

4. Resale/New apartment (1 denotes resale and 0 denotes new)

5. 24X7 security (1 denotes present, 0 denotes absent,9 denotes not mentioned)

6. Car parking (1 denotes present, 0 denotes absent,9 denotes not mentioned)

7. Lift available (1 denotes present, 0 denotes absent,9 denotes not mentioned)

Note:

1. I created 3 new factors from the given data

(A). **price per square foot** (dividing the area by price)

(B). **distance from city centre**(there were 300 different locations within the city. I took the city centre to be esplanade. With the help of google maps, I found out the distance of each location from esplanade and that distance is the distance from city centre)

(C). **Zone** (I divided the different locations into 9 zones, namely **north Kolkata, south Kolkata, north suburbs, south suburbs, south west suburbs, east Kolkata, Howrah, new urban developments and far away municipalities.** Then I categorised the locations to their correspondent zones.

CONCEPT:

1. Testing for statistical independence

In this case, an "observation" consists of the values of two outcomes and the null hypothesis is that the occurrence of these outcomes is statistically independent. Each observation is allocated to one cell of a two-dimensional array of cells (called a contingency table) according to the values of the two outcomes. If there are *r* rows and *c* columns in the table, the "theoretical frequency" for a cell, given the hypothesis of independence, is

$$E_{i,j} = N p_{i\cdot} p_{\cdot j}$$

where *N* is the total sample size (the sum of all cells in the table), and

$$p_{i\cdot} = \frac{1}{N} O_{i\cdot} N = \sum_{j=1}^{c} \frac{1}{N} O_{i,j} N$$

is the fraction of observations of type *i* ignoring the column attribute (fraction of row totals), and

$$p_{\cdot j} = O_{\cdot j} = \sum_{i=1} \frac{O_{ij}}{N}$$

is the fraction of observations of type *j* ignoring the row attribute (fraction of column totals), The term "frequencies" refers to absolute numbers rather than already normalized values.

The value of the test-statistic is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$= N \sum_{ij} p_{i.} p_{.j} \left(\frac{\left(\left(O_{ij}/N\right) - p_{i.} p_{.}\right)^2}{p_{i.} p_{.j}} \right)$$

Note that χ^2 is 0 if and only if Oi, j = Ei, j ∀i, j i.e., only if the expected and true number of observations are equal in all cells.

Fitting the model of "independence" reduces the number of degrees of freedom by *p* = *r* + *c* − 1. The number of degrees of freedom is equal to the number of cells *rc*, minus the reduction in degrees of freedom, *p*, which reduces to (*r* − 1)(*c* − 1).

For the test of independence, also known as the test of homogeneity, a chi-squared probability of less than or equal to 0.05 (or the chi-squared statistic being at or larger than the 0.05 critical point) is commonly interpreted by applied workers as justification for rejecting the null hypothesis that the row variable is independent of the column variable. The alternative hypothesis corresponds to the variables having an association or relationship where the structure of this relationship is not specified.

2. Test For Existence of Regression Between Two Variables

Suppose the sample values of two variables x and y are arranged in arrays of y according to fixed values x as given below: -

x_1	…	x_k
y_{11}	$\cdots$	y_{k1}
$\vdots$	$\ddots$	$\vdots$
y_{1n_1}	$\cdots$	y_{kn_k}

Define: 1. $\bar{y}_{i0} = \frac{1}{n_i}\sum_{j=1}^{ni} y_{ij}$, $i = 1(|)k$

2. $\bar{y}_{00} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_{i0}$, $n = \sum_{1}^{k} n_i$

3. $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$

Here,

$$e_{yx}^2 = \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i0}-\bar{y}_{00})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{00})^2} \quad , \quad e_{yx} = +\sqrt{e_{yx}^2} = sample\ correlation\ ratio$$

$$r = \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{00})(x_i-\bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{00})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{k}n_i(x_i-\bar{x})^2}}$$ = sample correlation coefficient

We assume, $\{y_{ij}|x = x_i\} \sim N(\mu_i, \sigma^2); i = 1(|)k$

⇨ $E(y_{ij}|x = x_i) = \mu_i$

H_0 : There exists a true regression of y on x

⇔ $H_0 : \mu_1 = \mu_2 \ldots\ldots = \mu_k$

Validity of H_0 means the absence of regression of y on x

Define $\eta_{yx}^2 = \frac{V(E(y|X))}{V(y)}$ where $E(y|X=x_i) = \mu_i$, for i=1(|)k

$\eta_{yx} = +\sqrt{\eta_{yx}^2}$ = population correlation ratio

Now,

$H_0 : \mu_1 = \mu_2 \ldots\ldots\ldots = \mu_k$
⇔ H_0 : E(y|x) = constant

⇔ $H_0 : \eta_{yx}^2 = 0$ against $H_1 : \eta_{yx}^2 > 0$

We note that, under H$_0$,

$e_{yx}^2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2 = \sum_{i=1}^{k} n_i (\bar{y}_{i0} - \bar{y}_{00})^2 \sim \sigma^2 \chi_{k-1}^2$ ....(*)

Again, $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i0})^2 + \sum_{i=1}^{k} n_i (\bar{y}_{i0} - \bar{y}_{00})^2$

Therefore, $(1-e_{yx}^2) \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2 - e_{yx}^2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2$

$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2 - \sum_{i=1}^{k} n_i (\bar{y}_{i0} - \bar{y}_{00})^2$

$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i0})^2 \sim \sigma^2 \chi_{n-k}^2$ ....(**)

(*) and (**) are independent

Under H$_0$, F = $\frac{e_{yx}^2/k-1}{1-e_{yx}^2/n-k} \sim F_{k-1,n-k}$ Here F is our test statistic.

For alternative hypothesis, H$_1$: $\eta_{yx}^2 > 0$, a large value of F supports H$_1$ and we reject H$_0$ if F$_0$>$F_{\alpha;k-1,n-k}$

Here, α∈ (0,1) = level of significance of the test,$F_{\alpha;k-1,n-k} = upper\ 100\alpha\%\ point\ of\ F_{k-1,n-k}$

F$_0$ = observed value of F for the given sample.

Note: (a) $F_{\alpha;k-1,n-k}$ is the critical value of the test.

(b) P=P[F≥ F$_0$] where F is a random variable $\sim F_{k-1,n-k}$ = The p-value of the test.

If p≤α , then we reject H$_0$ at the level of significance α ∈ (0,1)

Test for linearity:

If H$_0$ from the previous test is rejected i.e. if the regression of y on x is established then we may like to test whether the regression is linear i.e. we want to test H$_0$: $\mu_i = \alpha + \beta x_i, i = 1(|)k$ vs H$_1$: H$_0$ is false.

We note that,

$$r^2 \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2 = \frac{\{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})(x_i - \bar{x})\}^2}{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$$

And, $e_{yx}^2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2 = \sum_{i=1}^{k} n_i (\bar{y}_{i0} - \bar{y}_{00})^2$

$$\Rightarrow (e_{yx}^2 - r^2) \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2 = \sum_{i=1}^{k} n_i (\bar{y}_{i0} - \bar{y}_{00})^2 - \frac{\{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})(x_i - \bar{x})\}^2}{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$$

$$= \sum_{i=1}^{k} n_i (\bar{y}_{i0} - \bar{y}_{00})^2 - \hat{\beta}^2 \sum_{i=1}^{k} n_i (x_i - \bar{x})^2$$

$\sim \sigma^2 \chi_{k-2}^2$, under H$_0$..(*)

Where , $\hat{\beta} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})(x_i - \bar{x})}{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2} = least\ square\ estimate\ of\ \beta$

Also, $(1-e_{yx}^2) \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2 \sim \sigma^2 \chi_{n-k}^2$ ....(**)

(*) and (**) are independent

Under H$_0$, F = $\frac{(e_{yx}^2 - r^2)/k-2}{1-e_{yx}^2/n-k} \sim F_{k-2,n-k}$ Here F is our test statistic.

For an alternative hypothesis H_1: 'regression is not linear', a large value of F supports H_1 and we reject H_0 if $F_0 > F_{\alpha;k-2,n-k}$

Here, αϵ (0,1) = level of significance of the test, $F_{\alpha;k-2,n-k} = upper\ 100\alpha\%\ point\ of\ F_{k-2,n-k}$

F_0 = observed value of F for the given sample.

Note: (a) $F_{\alpha;k-2,n-k}$ is the critical value of the test.

(b) P=P[F≥ F_0] where F is a random variable $\sim F_{k-2,n-k}$ = The p-value of the test.

If p≤α , then we reject H_0 at the level of significance α ϵ (0,1)

ANALYSIS:

1. With the help of Ms-Excel , I created the frequency distribution table for observed and expected prices and area as given below:

Table for observed frequencies

	Area of the house (in square feet)										
Price (in rupees)	**350-1349**	**1350-2349**	**2350-3349**	**3350-4349**	**4350-5349**	**5350-6349**	**6350-7349**	**7350-8349**	**8350-9349**	**9350-10349**	**Grand Total**
2000000-3499999	1340	386	34	9	19	6	3	0	1	1	**1799**
3500000-4999999	1065	346	27	12	6	14	0	1	2	0	**1473**
5000000-6499999	596	247	19	8	10	6	0	0	0	0	**886**
6500000-7999999	449	225	20	7	2	5	1	1	0	0	**710**
8000000-9499999	267	166	16	4	4	8	0	0	0	0	**465**
9500000-10999999	167	88	10	5	2	1	2	0	0	0	**275**
11000000-12499999	116	59	5	0	2	1	0	0	0	0	**183**
12500000-13999999	69	46	5	1	1	0	0	0	0	0	**122**
14000000-15499999	40	32	3	1	1	0	0	0	0	0	**77**
15500000-16999999	26	22	2	1	0	0	0	0	0	0	**51**
17000000-18499999	23	17	4	2	0	0	0	0	0	0	**46**
18500000-19999999	13	2	1	1	0	0	0	0	0	0	**17**
Grand Total	**4171**	**1636**	**146**	**51**	**47**	**41**	**6**	**2**	**3**	**1**	**6104**

Table for expected frequencies

	Area of the house (in square feet)										
Price (in rupees)	**350-1349**	**1350-2349**	**2350-3349**	**3350-4349**	**4350-5349**	**5350-6349**	**6350-7349**	**7350-8349**	**8350-9349**	**9350-10349**	**Grand Total**
2000000-3499999	1229.297	482.1697	43.029817	15.030963	13.852064	12.08372	1.768349	0.58945	0.884174	0.2947248	**1799**
3500000-4999999	1006.534	394.7949	35.232307	12.307176	11.341907	9.894004	1.447903	0.482634	0.723952	0.2413172	**1473**
5000000-6499999	605.4237	237.4666	21.192005	7.4026868	6.8220839	5.95118	0.870904	0.290301	0.435452	0.1451507	**886**
6500000-7999999	485.1589	190.2949	16.982307	5.9321756	5.4669069	4.769004	0.697903	0.232634	0.348952	0.1163172	**710**
8000000-9499999	317.7449	124.6298	11.122215	3.8851573	3.5804391	3.123362	0.457077	0.152359	0.228539	0.0761796	**465**
9500000-10999999	187.9137	73.70577	6.577654	2.2976737	2.117464	1.847149	0.270315	0.090105	0.135157	0.0450524	**275**
11000000-12499999	125.048	49.04784	4.3771298	1.5289974	1.409076	1.229194	0.179882	0.059961	0.089941	0.0299803	**183**
12500000-13999999	83.36533	32.69856	2.9180865	1.0193316	0.939384	0.819463	0.119921	0.039974	0.059961	0.0199869	**122**
14000000-15499999	52.61583	20.63761	1.8417431	0.6433486	0.5928899	0.517202	0.075688	0.025229	0.037844	0.0126147	**77**
15500000-16999999	34.84944	13.66907	1.2198558	0.426114	0.3926933	0.342562	0.050131	0.01671	0.025066	0.0083552	**51**
17000000-18499999	31.43283	12.32896	1.1002621	0.3843381	0.354194	0.308978	0.045216	0.015072	0.022608	0.007536	**46**
18500000-19999999	11.61648	4.556356	0.4066186	0.142038	0.1308978	0.114187	0.01671	0.00557	0.008355	0.0027851	**17**
Grand Total	**4171**	**1636**	**146**	**51**	**47**	**41**	**6**	**2**	**3**	**1**	**6104**

H_0 : Price and area are independent

H_1 : Price and area are dependent

The value of the χ^2 *is obtained as* 198.23775 using the formula

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$$

$$= N\sum_{ij} p_{i.}p_{.j}\left(\frac{((O_{ij}/_{N})-p_{i.}p_{.j})}{p_{i.}p_{.j}}\right)^2$$

Degree of freedom= 11 X 9 = 99

The critical value is $\chi^2_{0.05;99}$ = 123.22522

Since χ^2_{observed} =198 > $\chi^2_{0.05;99}$ = 123.22522

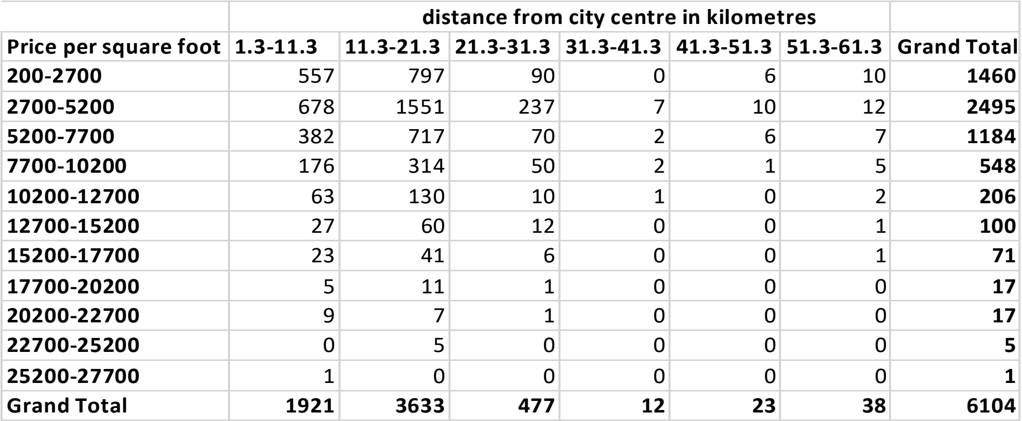
The test is rejected at 5% level

Also p = p-value of the test = P[$\chi^2 \geq \chi^2_{\text{observed}}$] =1.3 X 10^{-8} << 0.05

Hence , the test is rejected at 5% level.

2. With the help of Ms-Excel , I created the frequency distribution table for observed and expected prices per square foot and distance from city centre as given below. In this case I have taken price per square foot instead of price as my explained variable since prices at different locations vary due to their respective areas. Since price per square foot also includes area , it is a better measure than price alone.

Table for observed frequencies



	distance from city centre in kilometres						
Price per square foot	**1.3-11.3**	**11.3-21.3**	**21.3-31.3**	**31.3-41.3**	**41.3-51.3**	**51.3-61.3**	**Grand Total**
200-2700	557	797	90	0	6	10	**1460**
2700-5200	678	1551	237	7	10	12	**2495**
5200-7700	382	717	70	2	6	7	**1184**
7700-10200	176	314	50	2	1	5	**548**
10200-12700	63	130	10	1	0	2	**206**
12700-15200	27	60	12	0	0	1	**100**
15200-17700	23	41	6	0	0	1	**71**
17700-20200	5	11	1	0	0	0	**17**
20200-22700	9	7	1	0	0	0	**17**
22700-25200	0	5	0	0	0	0	**5**
25200-27700	1	0	0	0	0	0	**1**
Grand Total	**1921**	**3633**	**477**	**12**	**23**	**38**	**6104**

Table for expected frequencies

	distance from city centre in kilometres						
Price per square foot	**1.3-11.3**	**11.3-21.3**	**21.3-31.3**	**31.3-41.3**	**41.3-51.3**	**51.3-61.3**	**Grand Total**
200-2700	459.479	868.9679	114.0924	2.870249	5.501311	9.089122	**1460**
2700-5200	785.2056	1484.983	194.973	4.90498	9.401212	15.53244	**2495**
5200-7700	372.6186	704.6972	92.52425	2.327654	4.461337	7.370904	**1184**
7700-10200	172.462	326.1606	42.82372	1.077326	2.064875	3.411533	**548**
10200-12700	64.8306	122.6078	16.09797	0.40498	0.776212	1.282438	**206**
12700-15200	31.47117	59.51835	7.814548	0.196592	0.376802	0.622543	**100**
15200-17700	22.34453	42.25803	5.548329	0.139581	0.267529	0.442005	**71**
17700-20200	5.350098	10.11812	1.328473	0.033421	0.064056	0.105832	**17**
20200-22700	5.350098	10.11812	1.328473	0.033421	0.064056	0.105832	**17**
22700-25200	1.573558	2.975917	0.390727	0.00983	0.01884	0.031127	**5**
25200-27700	0.314712	0.595183	0.078145	0.001966	0.003768	0.006225	**1**
Grand Total	**1921**	**3633**	**477**	**12**	**23**	**38**	**6104**

H_0 : Price per square foot and distance from city centre are independent

H_1 : Price per square foot and distance from city centre are dependent

The value of the χ^2 *is obtained as* 92.90265 using the formula

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$= N \sum_{ij} p_{i.}p_{.j} \left(\frac{\left(\left(O_{ij}/_N\right) - p_{i.}p_{.j}\right)^2}{p_{i.}p_{.j}}\right)$$

Degree of freedom= 10 X 5 = 50

The critical value is $\chi^2_{0.05;50}$ = 67.504807

Since χ^2_{observed} = 92.90265 > $\chi^2_{0.05;50}$ = 67.504807

The test is rejected at 5% level

Also p = p-value of the test = P[$\chi^2 \geq \chi^2_{\text{observed}}$] = 0.0002188 < 0.05

<u>**Hence, test is rejected at 5% level.**</u>

3. With the help of Ms-Excel , I created the frequency distribution table for observed and expected price per square foot and whether the apartment is being resold or is new as given below:

Table for observed frequencies

	resale/new		
price per sq foot	**0**	**1**	**Grand Total**
200-2700	958	502	**1460**
2700-5200	1769	726	**2495**
5200-7700	808	376	**1184**
7700-10200	373	175	**548**
10200-12700	144	62	**206**
12700-15200	77	23	**100**
15200-17700	49	22	**71**
17700-20200	13	4	**17**
20200-22700	11	6	**17**
22700-25200	3	2	**5**
25200-27700	1		**1**
Grand Total	**4206**	**1898**	**6104**

Table for expected frequencies

	resale/new		
price per sq foot	**0**	**1**	**Grand Total**
200-2700	1006.022	453.9777	**1460**
2700-5200	1719.196	775.8044	**2495**
5200-7700	815.8427	368.1573	**1184**
7700-10200	377.6029	170.3971	**548**
10200-12700	141.9456	64.05439	**206**
12700-15200	68.90564	31.09436	**100**
15200-17700	48.923	22.077	**71**
17700-20200	11.71396	5.286042	**17**
20200-22700	11.71396	5.286042	**17**
22700-25200	3.445282	1.554718	**5**
25200-27700	0.689056	0.310944	**1**
Grand Total	**4206**	**1898**	**6104**

H_0 : Price per square foot and resold/new are independent

H_1 : Price per square foot and resold/new are dependent

The value of the χ^2 *is obtained as* 16.819517 using the formula

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$= N\sum_{ij} p_{i.}p_{.j}\left(\frac{\left(\left(^{O_{ij}}/_{N}\right) - p_{i.}p_{.j}\right)^2}{p_{i.}p_{.j}}\right)$$

Degree of freedom= 10 X 1= 10

The critical value is $\chi^2_{0.05;10} = 18.30704$

Since $\chi^2_{observed} = 16.81951 < \chi^2_{0.05;10} = 18.30704$

The test is accepted at 5% level

Also p = p-value of the test = $P[\chi^2 \geq \chi^2_{observed}] = 0.08597 > 0.05$

<u>Hence , the test is accepted at 5% level</u>

4. With the help of Ms-Excel , I created the frequency distribution table for observed and expected price per square foot and availability of 24 X 7 security as given below(1 represents available, 0 represents not available,9 represents not mentioned):

<u>Table for observed frequencies</u>

	24 X 7 security			
Price per square foot	0	1	9	**Grand Total**
200-2700	2	6	1452	**1460**
2700-5200	15	29	2451	**2495**
5200-7700	6	7	1171	**1184**
7700-10200	2	5	541	**548**
10200-12700	0	0	206	**206**
12700-15200	0	0	100	**100**
15200-17700	0	0	71	**71**
17700-20200	0	0	17	**17**
20200-22700	0	0	17	**17**
22700-25200	0	0	5	**5**
25200-27700	0	0	1	**1**
Grand Total	**25**	**47**	**6032**	**6104**

<u>Table for expected frequencies</u>

	24 X 7 security			
Price per square foot	0	1	9	**Grand Total**
200-2700	5.979685	11.24181	1442.779	**1460**
2700-5200	10.21871	19.21117	2465.57	**2495**
5200-7700	4.849279	9.116645	1170.034	**1184**
7700-10200	2.24443	4.219528	541.536	**548**
10200-12700	0.843709	1.586173	203.5701	**206**
12700-15200	0.409567	0.769987	98.82045	**100**
15200-17700	0.290793	0.546691	70.16252	**71**
17700-20200	0.069626	0.130898	16.79948	**17**
20200-22700	0.069626	0.130898	16.79948	**17**
22700-25200	0.020478	0.038499	4.941022	**5**
25200-27700	0.004096	0.0077	0.988204	**1**
Grand Total	**25**	**47**	**6032**	**6104**

H_0 : Price per square foot and 24 x 7 security availability are independent

H_1 : Price per square foot and 24 x 7 security availability are dependent

The value of the χ^2 *is obtained as* 18.37698 using the formula

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$= N\sum_{ij} p_{i.}p_{.j}\left(\frac{((O_{ij}/_{N}) - p_{i.}p_{.j})}{p_{i.}p_{.j}}\right)^2$$

Degree of freedom= 10 X 2= 20

The critical value is $\chi^2_{0.05;20}$ = 31.410433

Since χ^2_{observed} = 18.37698 < $\chi^2_{0.05;20}$ = 31.410433

The test is accepted at 5% level

Also p = p-value of the test = P[$\chi^2 \geq \chi^2_{\text{observed}}$] = 0.5625898 > 0.05

Hence , the test is accepted at 5% level.

5. With the help of Ms-Excel , I created the frequency distribution table for observed and expected price per square foot and availability of car parking
as given below(1 represents available, 0 represents not available,9 represents not mentioned):

Table for observed frequencies

		Car parking		
Price per square foot	0	1	9	**Grand Total**
200-2700	6	2	1452	**1460**
2700-5200	23	21	2451	**2495**
5200-7700	7	6	1171	**1184**
7700-10200	2	5	541	**548**
10200-12700	0	0	206	**206**
12700-15200	0	0	100	**100**
15200-17700	0	0	71	**71**
17700-20200	0	0	17	**17**
20200-22700	0	0	17	**17**
22700-25200	0	0	5	**5**
25200-27700	0	0	1	**1**
Grand Total	**38**	**34**	**6032**	**6104**

Table for expected frequencies

		Car parking		
Price per square foot	0	1	9	**Grand Total**
200-2700	9.089122	8.132372	1442.779	**1460**
2700-5200	15.53244	13.89744	2465.57	**2495**
5200-7700	7.370904	6.59502	1170.034	**1184**
7700-10200	3.411533	3.052425	541.536	**548**
10200-12700	1.282438	1.147444	203.5701	**206**
12700-15200	0.622543	0.557012	98.82045	**100**
15200-17700	0.442005	0.395478	70.16252	**71**
17700-20200	0.105832	0.094692	16.79948	**17**
20200-22700	0.105832	0.094692	16.79948	**17**
22700-25200	0.031127	0.027851	4.941022	**5**
25200-27700	0.006225	0.00557	0.988204	**1**
Grand Total	**38**	**34**	**6032**	**6104**

H_0 : Price per square foot and car parking availability are independent

H_1 : Price per square foot and car parking availability are dependent

The value of the χ^2 *is obtained as* 19.917059 using the formula

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$= N\sum_{ij} p_{i.}p_{.j}\left(\frac{\left(\left(O_{ij}/N\right) - p_{i.}p_{.j}\right)^2}{p_{i.}p_{.j}}\right)$$

Degree of freedom= 10 X 2= 20

The critical value is $\chi^2_{0.05;20}$ = 31.410433

Since χ^2_{observed} = 19.917059 < $\chi^2_{0.05;20}$ = 31.410433

The test is accepted at 5% level

Also p = p-value of the test = P[$\chi^2 \geq \chi^2_{\text{observed}}$] = 0.4631288 > 0.05

Hence , the test is accepted at 5% level.

6. With the help of Ms-Excel , I created the frequency distribution table for observed and expected price per square foot and availability of lift service
as given below(1 represents available, 0 represents not available,9 represents not mentioned):

Table for observed frequencies

		Lift		
Price per square foot	**0**	**1**	**9**	**Grand Total**
200-2700	3	5	1452	**1460**
2700-5200	11	33	2451	**2495**
5200-7700	3	10	1171	**1184**
7700-10200	0	7	541	**548**
10200-12700	0	0	206	**206**
12700-15200	0	0	100	**100**
15200-17700	0	0	71	**71**
17700-20200	0	0	17	**17**
20200-22700	0	0	17	**17**
22700-25200	0	0	5	**5**
25200-27700	0	0	1	**1**
Grand Total	**17**	**55**	**6032**	**6104**

Table for expected frequencies

		Lift		
Price per square foot	**0**	**1**	**9**	**Grand Total**
200-2700	4.066186	13.15531	1442.779	**1460**
2700-5200	6.948722	22.48116	2465.57	**2495**
5200-7700	3.29751	10.66841	1170.034	**1184**
7700-10200	1.526212	4.937746	541.536	**548**
10200-12700	0.573722	1.85616	203.5701	**206**
12700-15200	0.278506	0.901048	98.82045	**100**
15200-17700	0.197739	0.639744	70.16252	**71**
17700-20200	0.047346	0.153178	16.79948	**17**
20200-22700	0.047346	0.153178	16.79948	**17**
22700-25200	0.013925	0.045052	4.941022	**5**
25200-27700	0.002785	0.00901	0.988204	**1**
Grand Total	**17**	**55**	**6032**	**6104**

H_0 : Price per square foot and lift availability are independent

H_1 : Price per square foot and lift availability are dependent

The value of the χ^2 *is obtained as* 20.199018 using the formula

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$= N\sum_{ij} p_{i.}p_{.j}\left(\frac{\left(\frac{O_{ij}}{N}\right)-p_{i.}p_{.j}}{p_{i.}p_{.j}}\right)^2$$

Degree of freedom= 10 X 2= 20

The critical value is $\chi^2_{0.05;20}$ = 31.410433

Since χ^2_{observed} = 20.199018 < $\chi^2_{0.05;20}$ = 31.410433

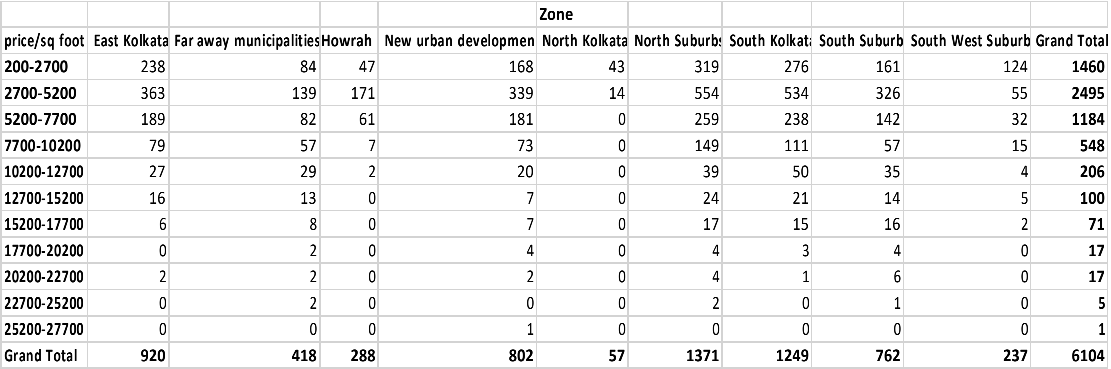
The test is accepted at 5% level

Also p = p-value of the test = P[$\chi^2 \geq \chi^2_{\text{observed}}$] = 0.44554371 > 0.05

Hence , the test is accepted at 5% level.

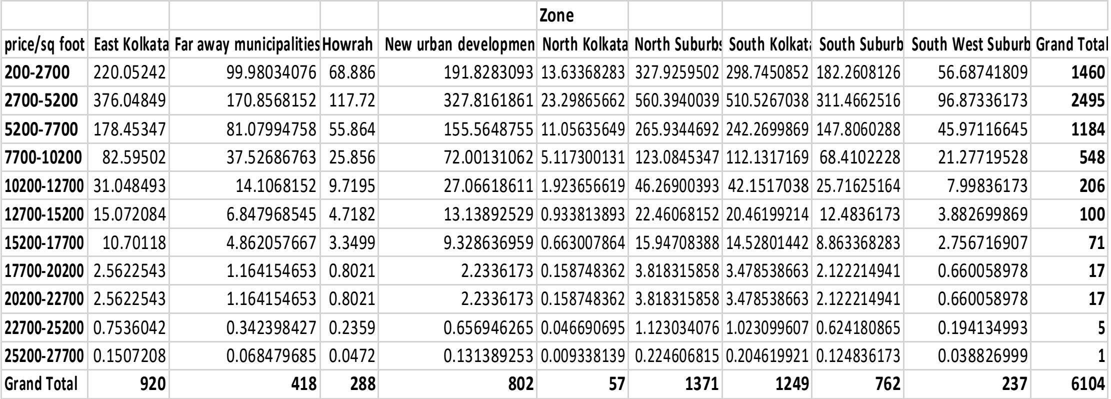
7. With the help of Ms-Excel , I created the frequency distribution table for observed and expected prices per square foot and the zone in which the house lies as given below.

Table for observed frequencies



					Zone					
price/sq foot	East Kolkata	Far away municipalities	Howrah	New urban developmen	North Kolkata	North Suburbs	South Kolkata	South Suburb	South West Suburb	Grand Total
200-2700	238	84	47	168	43	319	276	161	124	1460
2700-5200	363	139	171	339	14	554	534	326	55	2495
5200-7700	189	82	61	181	0	259	238	142	32	1184
7700-10200	79	57	7	73	0	149	111	57	15	548
10200-12700	27	29	2	20	0	39	50	35	4	206
12700-15200	16	13	0	7	0	24	21	14	5	100
15200-17700	6	8	0	7	0	17	15	16	2	71
17700-20200	0	2	0	4	0	4	3	4	0	17
20200-22700	2	2	0	2	0	4	1	6	0	17
22700-25200	0	2	0	0	0	2	0	1	0	5
25200-27700	0	0	0	1	0	0	0	0	0	1
Grand Total	920	418	288	802	57	1371	1249	762	237	6104

Table for expected frequencies



					Zone					
price/sq foot	East Kolkata	Far away municipalities	Howrah	New urban developmen	North Kolkata	North Suburbs	South Kolkata	South Suburb	South West Suburb	Grand Total
200-2700	220.05242	99.98034076	68.886	191.8283093	13.63368283	327.9259502	298.7450852	182.2608126	56.68741809	1460
2700-5200	376.04849	170.8568152	117.72	327.8161861	23.29865662	560.3940039	510.5267038	311.4662516	96.87336173	2495
5200-7700	178.45347	81.07994758	55.864	155.5648755	11.05635649	265.9344692	242.2699869	147.8060288	45.97116645	1184
7700-10200	82.59502	37.52686763	25.856	72.00131062	5.117300131	123.0845347	112.1317169	68.4102228	21.27719528	548
10200-12700	31.048493	14.1068152	9.7195	27.06618611	1.923656619	46.26900393	42.1517038	25.71625164	7.99836173	206
12700-15200	15.072084	6.847968545	4.7182	13.13892529	0.933813893	22.46068152	20.46199214	12.4836173	3.882699869	100
15200-17700	10.70118	4.862057667	3.3499	9.328636959	0.663007864	15.94708388	14.52801442	8.863368283	2.756716907	71
17700-20200	2.5622543	1.164154653	0.8021	2.2336173	0.158748362	3.818315858	3.478538663	2.122214941	0.660058978	17
20200-22700	2.5622543	1.164154653	0.8021	2.2336173	0.158748362	3.818315858	3.478538663	2.122214941	0.660058978	17
22700-25200	0.7536042	0.342398427	0.2359	0.656946265	0.046690695	1.123034076	1.023099607	0.624180865	0.194134993	5
25200-27700	0.1507208	0.068479685	0.0472	0.131389253	0.009338139	0.224606815	0.204619921	0.124836173	0.038826999	1
Grand Total	920	418	288	802	57	1371	1249	762	237	6104

H_0 : Price per square foot and zone are independent

H_1 : Price per square foot and zone are dependent

The value of the χ^2 *is obtained as* 376.67699 using the formula

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$= N \sum_{ij} p_{i.}p_{.j} \left(\frac{\left(\left(O_{ij}/_N\right) - p_{i.}p_{.j}\right)^2}{p_{i.}p_{.j}}\right)$$

Degree of freedom= 10 X 8 = 80

The critical value is $\chi^2_{0.05;80}$ = 101.879474

Since χ^2_{observed} = 376.67699> $\chi^2_{0.05;\,80}$ = 101.879474

The test is rejected at 5% level

Also p = p-value of the test = P[$\chi^2 \geq \chi^2_{\text{observed}}$] = 5.23 X 10^{-40} <<< 0.05

<u>**Hence, test is rejected at 5% level.**</u>

8. We now test whether any regression exists between price and area

$H_0 : \mu_1 = \mu_2 \ldots\ldots\ldots = \mu_k$

⇔ H_0 : E(y|x) = constant

⇔ $H_0 : \eta^2_{yx} = 0$ against $H_1 : \eta^2_{yx} > 0$

Where y denotes price and x denotes area.

Under H_0, F = $\frac{e^2_{yx}/k-1}{1-e^2_{yx}/n-k} \sim F_{k-1,n-k}$. Here F is our test statistic.

And, $e^2_{yx} = \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i0}-\bar{y}_{00})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{00})^2}$

For the given problem,

e^2_{yx}= 0.017864494, k = 10, n = 6104

F = 12.316271 , critical value at 5% level $= F_{0.05;k-1,n-k} =$ 1.88141676

Since, $F_{\text{observed}} > F_{\text{critical}}$, the test is rejected at 5 % level

p-value of the test = p = P[F>F_{observed}] = 1.5313 X 10^{-19}

since p << 0.05 , the test is rejected at 5% level

Therefore, there exists a significant regression of y on x at 5% level

Now we shall try to test whether there exists linearity in the regression of y on x.

H$_0$: $\mu_i = \alpha + \beta x_i, i = 1(|)k$ v

H$_1$: H$_0$ is false.

Where E $(y_{ij}|x = x_i) = \mu_i$

For this problem, Under H$_0$, F = $\frac{(e_{yx}^2 - r^2)/k-2}{1-e_{yx}^2/n-k} \sim F_{k-2,n-k}$. Here F is our test statistic

Where $e_{yx}^2 = \frac{\sum_{i=1}^{k} n_i (\bar{y}_{i0} - \bar{y}_{00})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{00})^2}$, r^2 = $\left(\frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{00})(x_i - \bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{00})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}}\right)^2$

For this problem, e_{yx}^2= 0.017864494, k = 10, n = 6104 , r^2 =(0.0840236)^2 = 0.00706

F$_{\text{observed}}$ = 8.232975145 , , critical value at 5% level = $F_{0.05;k-2,n-k}$ = 1.939924685

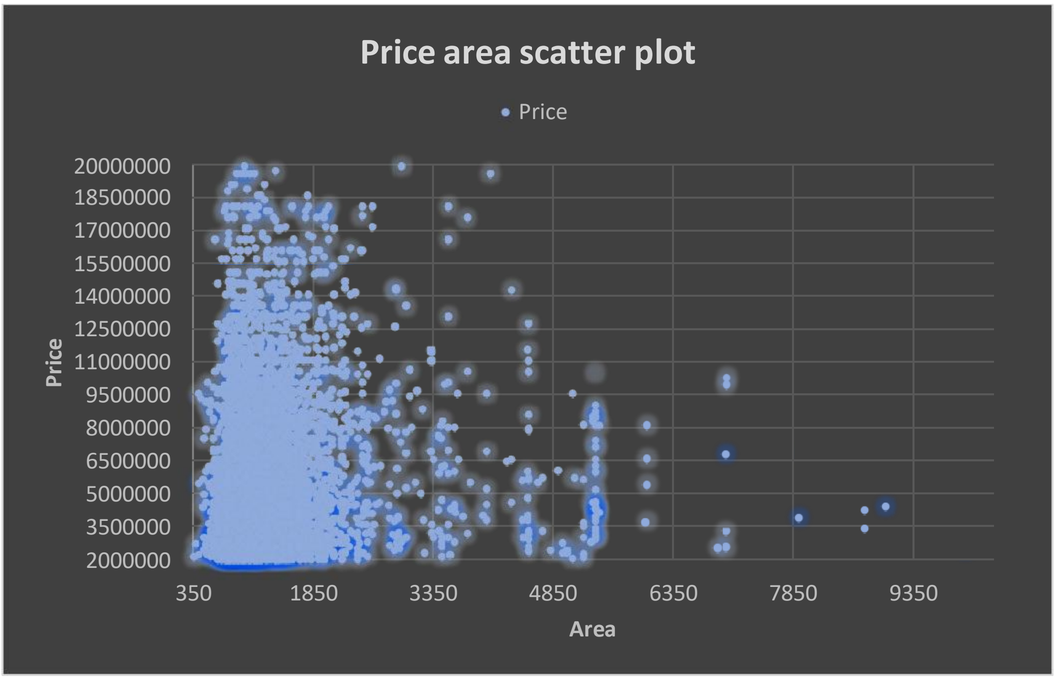
Since, F$_{\text{observed}}$ > F$_{\text{critical}}$, the test is rejected at 5 % level

p-value of the test = p = P[F>F$_{\text{observed}}$] = 3.76852 X 10^{-11}

since p << 0.05 , the test is rejected at 5% level

Hence, linearity is not significant at 5% level.

This finding is quite justified by looking at the scatter plot diagram given below:



From the scatter plot, it is quite evident that there cannot be any linear trend that can explain the price when area is given.

9. We now test whether any regression exists between price per square foot and distance from city centre.

$H_0 : \mu_1 = \mu_2 \ldots\ldots\ldots = \mu_k$

⇔ $H_0 : E(y|x) = \text{constant}$

⇔ $H_0 : \eta^2_{yx} = 0$ against $H_1 : \eta^2_{yx} > 0$

Where y denotes price per square foot and x denotes distance from city centre .

Under H_0, F = $\frac{e^2_{yx}/k-1}{1-e^2_{yx}/n-k} \sim F_{k-1,n-k}$. Here F is our test statistic.

And, $e^2_{yx} = \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i0}-\bar{y}_{00})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{00})^2}$

For the given problem,

e^2_{yx}=0.00093264, k = 6, n = 6104

F = 1.13850817 , critical value at 5% level $= F_{0.05;k-1,n-k}$ = 2.21556518

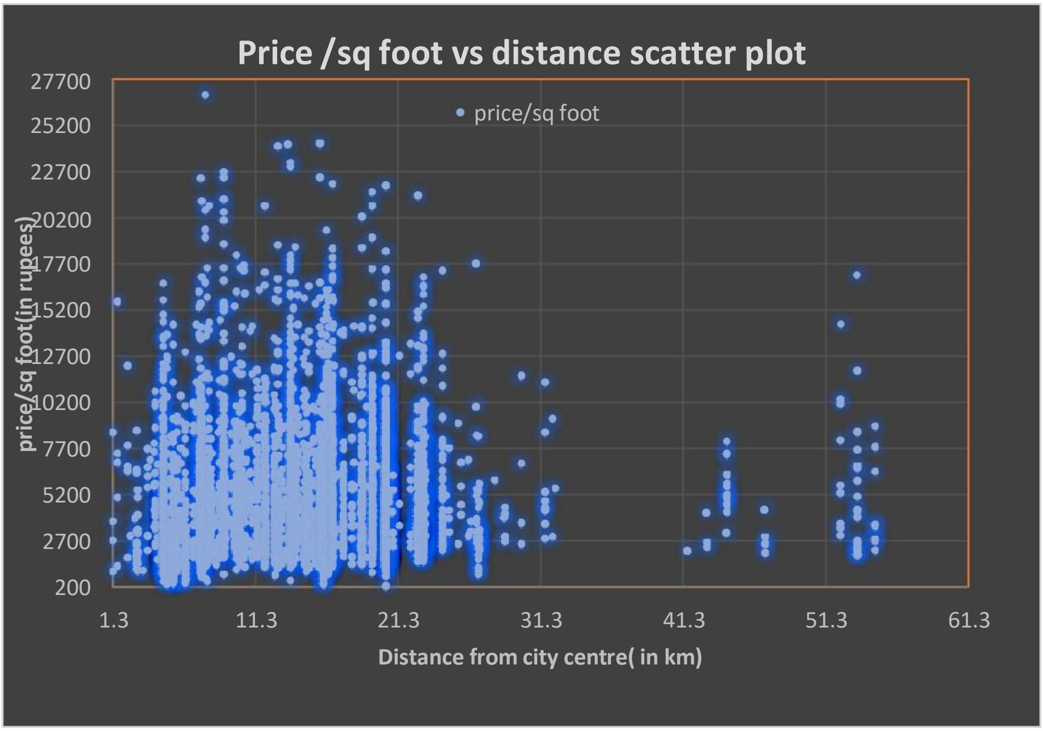
Since, $F_{\text{observed}} < F_{\text{critical}}$, the test is accepted at 5 % level

p-value of the test = p = $P[F>F_{\text{observed}}]$ = 0.33742623

since p > 0.05 , the test is accepted at 5% level

Therefore, there does not exist any regression of y on x at 5% level of significance, which is c

learly evident from the scatter plot given below:



RESULT:

1. **Price and area are dependent at 5% level. In fact, there exists a true regression of price on area at 5 % level**

2. **Price per square foot and distance from city centre, although dependent, there does not exist any regression of price per square foot on distance at 5% level.**

3. **Price per square foot and the criteria resold/new are independent at 5% level.**

4. **Price per square foot and 24X7 security availability are independent at 5% level.**

5. **Price per square foot and car parking facilities are independent at 5% level.**

6. **Price per square foot and lift availability are independent at 5% level.**

7. **Price per square foot and zone are dependent at 5% level.**

CONCLUSION:

Hence we can conclude that while price is significantly dependent on area, the variability explained by area is not quite high. This may be due to the reason that the presence of various local and random factors also come into play to explain the price of an apartment.

On the other hand , however, price per square foot is significantly dependent on the zone of the city in which the apartment is present and also, to some extent ,can be explained by the distance of the apartment from the city centre, although the dependence is still, quite low. The reason again may be some local and random factors like income of people in that area, target customer base, etc.

Surprisingly, the prices per square foot are independent of the other factors like lift availability, car parking ,24 X 7 security and whether the apartment is new/resold. The reason may be a biasness in the available data, for the frequency of 'not mentioned' category is seen to be overwhelmingly large.

BIBLIOGRAPHY

1. Fundamentals of Statistics (Vol 1&2), by Gun, Gupta, Dasgupta
2. Introduction to linear regression analysis, by Montgomery
3. An Introduction to Theory of Statistics, by Yule, Kendall
4. www.kaggle.com
5. www.wikipedia.com

BOOKS

WEBSITES