Furniture Sales Data

Group 06

Group Members :

Shashin Prabath	16375
Kavithi Wijesekara	16380
Dewmi Yasarathne	16382
Theekshana Yatawathura	16384

Contents

A. Introduction

Data Set Name : Furniture Sales Data.

Source of the data set :Furniture Sales Data (kaggle.com)

Target Variable (y) : Price

Auxiliary Variable (x) : Cost

No. of Observations : 2500

This dataset contains information on furniture sales with 2500 observations. It includes 15 variables (8 quantitative variables and 7 categorical variables). The dataset captures various aspects of furniture sales, such as pricing, cost, sales volume, discount percentage, inventory levels, delivery time, and different categorical attributes like furniture type, material, color, and store location.

In here, we use “price” as our response variable as it is a continuous variable. Since “cost” is also a continuous variable and it highly correlate with price, we choose “cost” as auxiliary variable in our study.

This dataset was analyzed using Simple Random Sampling, Stratified Sampling, and Two-stage Cluster Sampling separately and gain the best estimators for the parameters. In this study we mainly focus on the mean, total, proportion estimation. “category” variable will be used to define strata and “color” variable will be used to define clusters. All these methods are explained in detail in the next parts of the report.

B. Methodology

The sample size for a survey is manually determined using the following formulas.

$$n_0 = \left(Z_{\alpha/2} \times \frac{S}{e}\right)^2$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

n = Sample Size

N = Population Size

$Z_{\alpha/2}$ = Z value of the significance level

S = Data Variation

e = Margin of error

For all calculations and perform every sample and sample estimations we use R statistical software.

❖ Simple Random Sampling

- Simple Random Sampling is one of most simple and easy sampling method used to select a sample from a population in such a way that each individual has an equal chance of being selected.
- In Simple Random Sampling (SRS), it is assumed that each element in the population has an equal probability of being selected, and selections are made independently of one another.
- The population must be well-defined, finite, and represented by an accurate sampling frame. Typically, sampling is done without replacement, meaning no element is chosen more than once.
- There should be no systematic differences among population members that could introduce bias, and the sample size must be large enough to ensure representativeness and precision in estimates. If these assumptions hold, SRS provides a reliable, unbiased sample for analysis.
- First we need to calculate the sample size, the calculation is more easier with R software. Instead of calculate the sample size for SRS with replacement then using that calculate sample size for SRS here we use “rsampcalc” command.

```
> # Sample size calculation for demonstration (you can adjust based on your needs)
> a = rsampcalc(nrow(Furniture), e = 3, ci = 95, p = 0.5)
> a
[1] 748
```

Then the parameter estimation and population parameters are shown in the below table.

	Population	Sample 1		Sample 2	
	Mean	**Estimated Population Mean**	**Standard Error**	**Estimated Population Mean**	**Standard Error**
Price	274.50	275.02	4.7144	272.94	4.7834
Cost	191.93	191.96	3.5443	189.77	3.5997
	Total	**Estimated Population Total**	**Standard Error**	**Estimated Population Total**	**Standard Error**
Price	686238	687546.8	3526.4	682356.3	3578.0
Cost	479825	479893	2651.1	474435.2	2692.6
	Proportion	**Estimated Proportion**	**Standard Error**	**Estimated Proportion**	**Standard Error**
Store Type					
• Online	0.5228	0.5201	0.0183	0.5308	0.0183
• Retail	0.4772	0.4700	0.0183	0.4693	0.0183
Location					
• Rural	0.3588	0.3543	0.0175	0.3316	0.0172
• Suburban	0.3252	0.3329	0.0172	0.3516	0.0175
• Urban	0.3160	0.3128	0.0170	0.3168	0.0170

- Sample 1 has a higher mean price (275.02) than Sample 2 (272.94) and a lower standard error (4.7144 vs. 4.7834), indicating more precision in Sample 1's estimate. The estimated population total for Sample 1 (687546.8) is also higher than Sample 2 (682356.3).

- For cost, Sample 1's mean (191.96) is higher than Sample 2 (189.77), with a lower standard error (3.5443 vs. 3.5997), suggesting a more reliable estimate. The population total for Sample 1 (479893) is greater than Sample 2 (474435.2).

- In store types, Sample 1 has a slightly lower online proportion (0.5201 vs. 0.5308) but a higher retail proportion (0.4700 vs. 0.4693).

- For location, Sample 1 shows a higher rural proportion (0.3543 vs. 0.3316) and a lower suburban proportion (0.3329 vs. 0.3516). The urban proportions are similar (0.3128 vs. 0.3168).

- Overall, Sample 1 indicates higher prices and costs, more rural representation, while Sample 2 shows a stronger online presence and suburban focus. Sample 1 generally has more precise estimates.

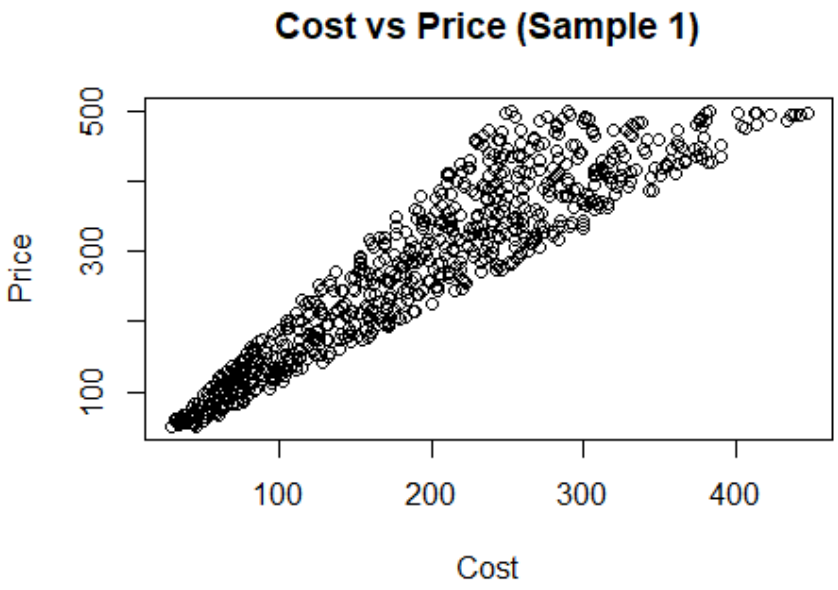
Regression Analysis

- The main idea is to use a regression model to estimate the relationship between the variable of interest (Price) and an auxiliary variable (cost).
- By combining the sample data with this auxiliary information, the regression estimate reduces the variance and improves the reliability of population estimates (such as totals or means) compared to basic SRS estimates.
- This method is particularly useful when there is a strong correlation between the auxiliary variable and the target variable, as it leads to smaller standard errors and more accurate population estimates.
- Since price and cost has a correlation coefficient greater than 0.9 we tried this method.

```
> cor_sample1 <- cor(srs_sample1$variables$price, srs_sample1$variables$cost, use = "complete.obs")
> cor_sample1
[1] 0.9314237
> cor_sample2 <- cor(srs_sample2$variables$price, srs_sample2$variables$cost, use = "complete.obs")
> cor_sample2
[1] 0.9318662
```

Sample 1

Since price and cost has a strong and linear relationship, we an show it using scatter plot.in below graph we can see there is a strong positive and linear relationship.



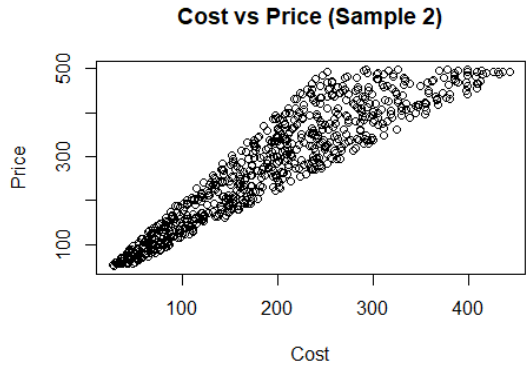

```
call:
lm(formula = price ~ cost, data = srs_sample1$variables)

Coefficients:
(Intercept)         cost
     37.199        1.239
```

- We can write least square regression line using above output.

Price = 37.199 + 1.239 * Cost
275. 04 = 37.199 + 1.239 *191.96

- When we estimate the Price using the mean of cost there is 83.11 deviation.

Sample 2



Like the graph in sample one this graph also has the almost same characteristics.

```
Call:
lm(formula = price ~ cost, data = srs_sample2$variables)

Coefficients:
(Intercept)         cost
     37.949        1.238
```

- We can write least square regression line using above output.

Price = 37.949 + 1.238 * Cost
275.60= 37.949 + 1.238 *191.96

- When we estimate the Price using the mean of cost there is 83.67 deviation.

Ratio Analysis

This method is particularly useful when there is a strong, consistent relationship between a target variable (Price) and an auxiliary variable (cost).

In ratio estimation, instead of estimating the total or mean directly from the sample, the ratio of the target variable to the auxiliary variable is calculated and used to project the population estimate.

Sample 1

```
> # Ratio Estimation for Sample 1
> mean_Y1 <- as.numeric(svymean(~price, srs_sample1))
> mean_X1 <- as.numeric(svymean(~cost, srs_sample1))
> ratio_estimate1 <- mean_Y1 / mean_X1
> ratio_estimate1
[1] 1.432711
```

- As above result we take a estimation for ratio estimation using sample 1 data.

For population,

T(hat)$_{yr}$= Ratio Estimation of sample 1 * total of cost in population

687450.6=1.432711 * 479825

Ratio estimated mean for price= 274.98

- Total of price in population is 686238.Estimation is slightly under estimated for the true parameter.

Sample 2

```
> # Ratio Estimation for Sample 2
> mean_Y2 <- as.numeric(svymean(~price, srs_sample2))
> mean_X2 <- as.numeric(svymean(~cost, srs_sample2))
> ratio_estimate2 <- mean_Y2 / mean_X2
> ratio_estimate2
[1] 1.438253
```

- As above result we take a estimation for ratio estimation using sample 2 data.

For population,

T(hat)yr= Ratio Estimation of sample 2 * total of cost in population

690109.7= 1.438253 * 479825

Ratio estimated mean for price= 276.04

- Total of price in population is 686238.Estimation is slightly over estimated for the true parameter.

❖ Stratified Random Sampling

This section will cover the stratified random sampling technique's methodology.

- First, within the target population, strata need to be created. This can be achieved by applying stratification to variables that significantly affect the outcome variable. The variable we used for stratification was category.

- In order to evaluate the Price in each stratum separately, we first divide the population into categories. We next take into account the fact that strata do not overlap and that within-stratum variance is minimal.

- R software package was used to conduct the analysis.

- Using the "rsampcalc" command, a random sample size with a 3 tolerance for error was produced. The result is 748.

```
> set.seed(1637)
> Samp_size=rsampcalc(nrow(Furniture),e=3,ci=95,0.5)
> Samp_size
[1] 748
```

- Then the samples from each stratum were then chosen using a simple method of random sampling, which can be done by using the "ssampcalc" command.

```
> strata_size=ssampcalc(Furniture,Samp_size,category)
> strata_size
```

- The sample sizes of each stratum are,

	category	Nh	wt[,1]	nh[,1]
	<chr>	<int>	<dbl>	<dbl>
1	Bed	481	0.192	144
2	Chair	497	0.199	149
3	Desk	501	0.200	150
4	Sofa	488	0.195	146
5	Table	533	0.213	159

```
>
```

Then, the parameter estimations and population parameters are shown in below chart.

	Population	Sample 1		Sample 2	
	Mean	**Estimated Population Mean**	**Standard Error**	**Estimated Population Mean**	**Standard Error**
Price	274.4952	283.49	4.8045	279.19	4.8001
Cost	191.9301	196.56	3.5405	193.63	3.601
	Total	**Estimated Population Total**	**Standard Error**	**Estimated Population Total**	**Standard Error**
Price	686238	708260	12003	697497	11992
Cost	479825.3	491079	8845.2	483749	8996.4
	Proportion	**Estimated Proportion**	**Standard Error**	**Estimated Proportion**	**Standard Error**
Store Type					
• Online	0.5228	0.5254	0.0183	0.53476	0.0182
• Retail	0.4772	0.4746	0.0183	0.46524	0.0182
Location					
• Rural	0.3160	0.31283	0.0170	0.32754	0.0171
• Suburban	0.3588	0.35027	0.0175	0.35695	0.0175
• Urban	0.3252	0.33690	0.0173	0.31551	0.0170

- The average cost of the population is 191.93, while the average costs of stratified Sample 1 and stratified Sample 2 are 196.56 and 193.63, respectively. While the

differences between the sample averages and the population average are relatively small.

- The total cost for the entire population is 479,825.3. In comparison, stratified sample 1 has a total cost of 491,079, which is slightly higher than the population total. Stratified sample 2, on the other hand, has a total cost of 483,749, which is slightly closer to the population total than sample 01.

- Overall, the proportions of store types in both samples are quite similar to the population proportions. This suggests that the samples are representative of the population in terms of store type distribution.

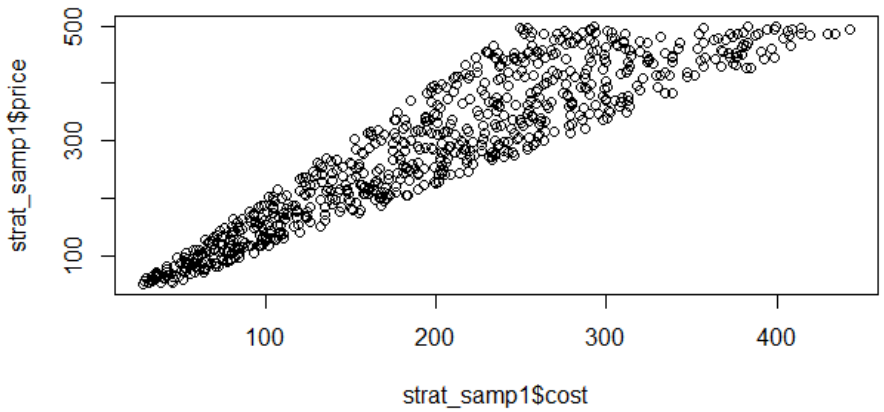
- In this case, Stratified Samples 01 and 02 have roughly the same estimated values for the variable's **location** and **season** with lesser standard errors.

Regression Analysis

To find the **price** mean, regression estimation was carried out.

Sample 01

Important conclusions are,



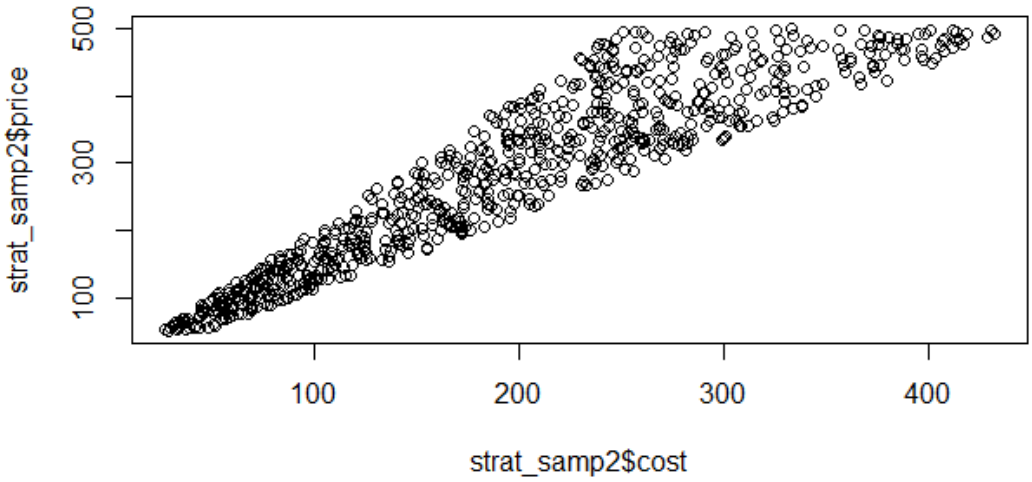
Coefficients:

(Intercept)	strat_samp1$cost
36.304	1.258

- Mean for **cost** in population was obtained as 191.9301
- Then the calculated expected mean for **price** using regression model is 277.6681.

Sample 02

Important conclusions are,



Coefficients:

(Intercept)	strat_samp2$cost
38.794	1.242

- Mean for **cost** in population was obtained as 191.9301
- Then the calculated expected mean for **price** using regression model is 282.8279.

We can conclude that the Regression model results for samples 1 and 2 show a slightly difference expected mean price, which is not a significant value.

Ratio analysis

Sample 01

After that, a ratio estimation approach is used to estimate the price mean.

```
$total
                    strat_samp1$cost
strat_samp1$price         692028.3

$se
                    strat_samp1$cost
strat_samp1$price         4680.451

 > Estimated_mean = 692028.3/2500
 > Estimated_mean
 [1] 276.8113
```

Sample 02

A ratio estimation approach is used to estimate the price mean.

```
$total
                    strat_samp2$cost
strat_samp2$price         691839.5

$se
                    strat_samp2$cost
strat_samp2$price         4685.425

  > Estimated_mean = 691839.5/2500
  > Estimated_mean
  [1] 276.7358
```

The price estimates that we were able to obtain by ratio estimation are extremely similar.

❖ Two – Stage Cluster Sampling

- Clustering Variable : Color of the Furniture
- Number of clusters in the Population : 6
- Number of observations in each Cluster :

```
> color=table(color)
> color
color
Black  Blue Brown Green   Red White
  448   401   406   399   419   427
```

- Number of arbitrary selected Clusters : 4
- Selected Clusters for Sample 1 :

```
> #we arbitrary select 4 clusters
> n=4
> set.seed(2109)
> clusters1 = sample(x = unique(Furniture$color),size = n,replace = F)
> clusters1
[1] Red   Blue  White Brown
```

	color	Population Size	Sample Size
1	Red	419	301
2	Blue	401	292
3	White	427	305
4	Brown	406	295

Sample size = 1193

- Selected Clusters for Sample 2 :

```
> set.seed(555)
> clusters1 = sample(x = unique(Furniture$color),size = n,replace = F)
> clusters1
[1] Blue  Red   Green Brown
Levels: Black Blue Brown Green Red White
```

	color	Population Size	Sample Size
1	Blue	401	292
2	Red	419	301
3	Green	399	291
4	Brown	406	295

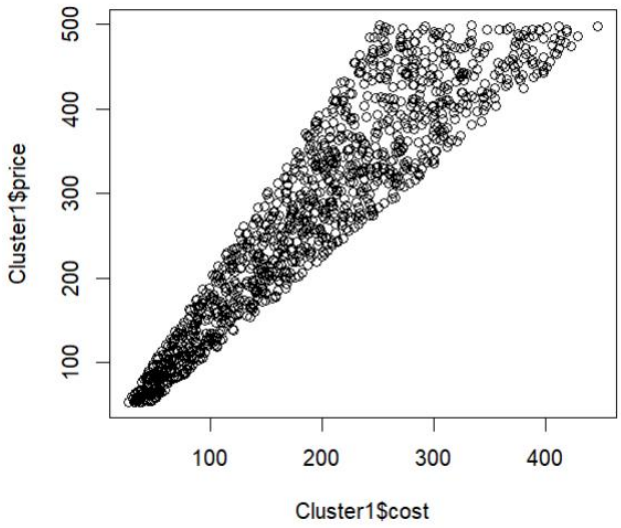
Sample size = 1179

The parameter estimations and population parameters are shown in below chart,

	Population	Sample 1		Sample 2	
	Mean	**Estimated Population Mean**	**Standard Error**	**Estimated Population Mean**	**Standard Error**
Price	274.4952	275.03	2.9258	270.62	3.756
Cost	191.9301	190.62	2.4964	189.14	2.3464
	Total	**Estimated Population Total**	**Standard Error**	**Estimated Population Total**	**Standard Error**
Price	686238	681941	16481	659644	16395
Cost	479825.3	472653	11967	461029	10784
	Proportion	**Estimated Proportion**	**Standard Error**	**Estimated Proportion**	**Standard Error**
Store Type					
• Online	0.5228	0.51724	0.007	0.531	0.0071
• Retail	0.4772	0.48276	0.007	0.469	0.0071
Location					
• Rural	0.3588	0.36886	0.0054	0.36468	0.0119
• Suburban	0.3252	0.30594	0.0082	0.30283	0.0081
• Urban	0.3160	0.32520	0.0052	0.33249	0.0073

Regression analysis

Sample 01



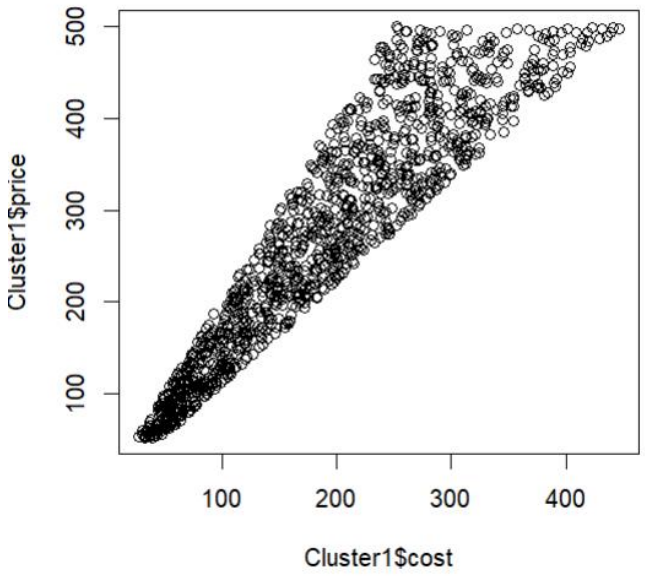
Valuable insights are,

Coefficients:

Intercept	Cost
34.92207	1.25957

- Mean of Price in population = 274.4952
- The calculated expected Mean of Price using the regression model is, 34.92207 + 1.25957*191.9301 = 276.6715

Sample 02



Valuable insights are,

Coefficients:

Intercept	Cost
35.29215	1.24417

- Mean of Price in population = 274.4952
- The calculated expected Mean of Price using the regression model is, 34.92207 + 1.24417 *191.9301 = 274.0858

Ratio Analysis

The ratio estimation approach is used to estimate the price mean.

Sample 01

```
$total
                Cluster1$cost
Cluster1$price         692290

$se
                Cluster1$cost
Cluster1$price       3255.016
```

Estimated mean = 692290 / 2500 = 276.916

Sample 02

```
$total
                Cluster1$cost
Cluster1$price       686536.8

$se
                Cluster1$cost
Cluster1$price       2786.968
```

Estimated mean = 686536.8 / 2500 = 274.61472

C. Results of the Study

❖ Simple Random Sampling

	Population	Normal Estimation
	Mean	**Estimated Mean (Standard Error)**
Cost	191.93	191.96(3.54) 189.77(3.60)
Price	274.5	275.02(4.711) 272.94(4.78)
	Totals	**Estimated Totals**
Cost	479825	479893(2651.10) 474435.2(2692.60)
Price	686238	687546.8(3526.40) 682356.3(3578.0)
	Proportions	**Estimated Proportions**
Stock_type		
• Online	0.5228	0.5201 0.5308
• Retail	0.4772	0.4700 0.4693
Location		
• Urban	0.3588	0.3545 0.3316
• Rural	0.3252	0.3329 0.3516
• Suburban	0.3160	0.3128 0.3128

Regression Estimation	Ratio Estimation
Estimated Mean (Price)	Estimated Mean (Price)
275.04 276.60	274.98 276.04

❖Stratified Random Sampling

	Population	**Normal Estimation**
	Mean	**Estimated Mean (Standard Error)**
Cost	191.9301	196.56 (3.5405) 193.63 (3.601)
Price	274.4952	283.49 (4.8045) 279.19 (4.8001)
	Totals	**Estimated Totals**
Cost	117221	491079 (8845.2) 483749 (8996.4)
Price	143129	708260 (12003) 697497 (11992)
	Proportions	**Estimated Proportions**
Stock_type		
• Online	0.5228	0.5254 (0.0183) 0.53476 (0.0182)
• Retail	0.4772	0.4746 (0.0183) 0.46524 (0.0182)
Location		

• Urban	0.3160	0.31283 (0.0170) 0.32754 (0.0171)
• Rural	0.3588	0.35027 (0.0175) 0.35695 (0.0175)
• Suburban	0.3252	0.33690 (0.0173) 0.31551 (0.0170)

Regression Estimation	Ratio Estimation
Estimated Mean (Price)	Estimated Mean (Price)
277.6681, 282.8279	276.8113, 276.7358

❖ Two – Stage Cluster Sampling

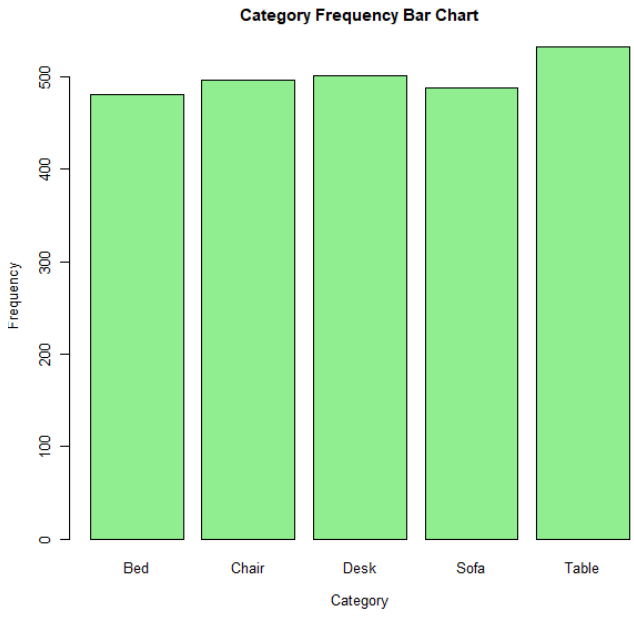
	Population	**Normal Estimation**
	Mean	**Estimated Mean (Standard Error)**
Cost	191.9301	160.62 (2.4964) 189.14 (2.3464)
Price	274.4952	275.03 (2.9258) 270.62 (3.756)
	Totals	**Estimated Totals**
Cost	479825.3	472653 (11967) 461029 (10784)
Price	686238	681941 (16481) 472653 (11967)
	Proportions	**Estimated Proportions**
Stock_type		
• Online	0.5228	0.51724 (0.007) 0.531 (0.0071)
• Retail	0.4772	0.48276 (0.007)

		0.469 (0.0071)
Location		
• Urban	0.3588	0.36886 (0.0054) 0.36468 (0.0119)
• Rural	0.3252	0.30594 (0.0082) 0.30283 (0.0081)
• Suburban	0.3160	0.32520 (0.0052) 0.33249 (0.0073)

Regression Estimation	Ratio Estimation
Estimated Mean (Price)	Estimated Mean (Price)
276.6715 , 274.0858	276.916, 274.61472

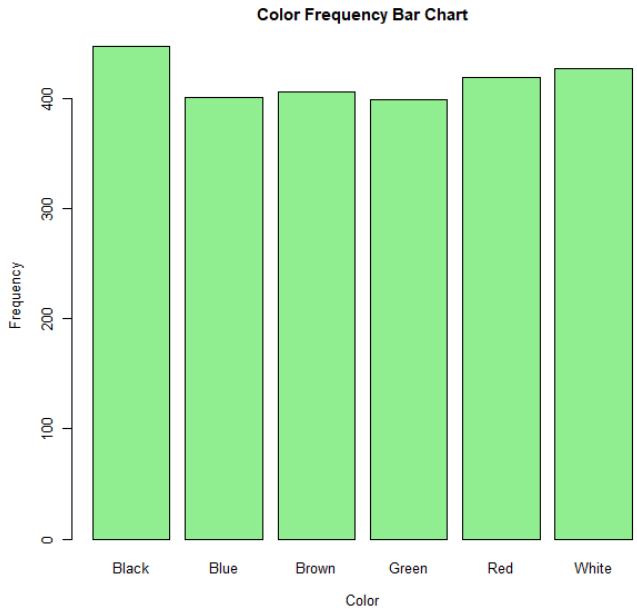
D. Graphical Analysis

❖ Graphical Analysis about Population



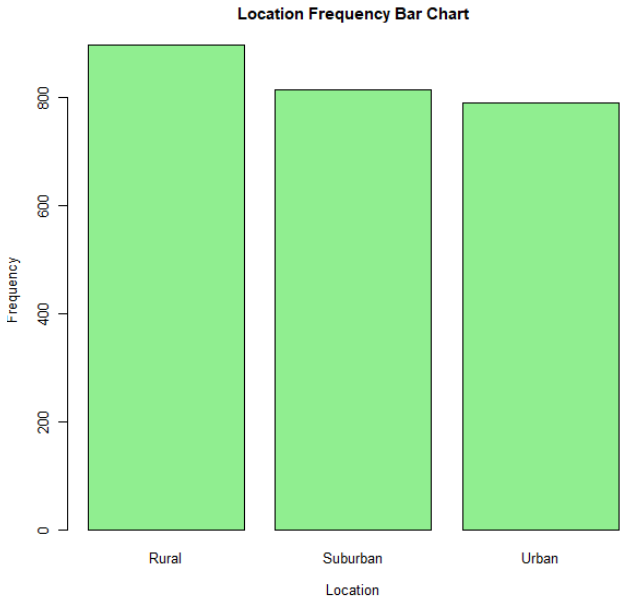

```
> category

  Bed Chair  Desk  Sofa Table
  481   497   501   488   533
```

There are 481 records from bed, 497 from chair, 501 from desk, 488 from sofa, and 533 from table categories.



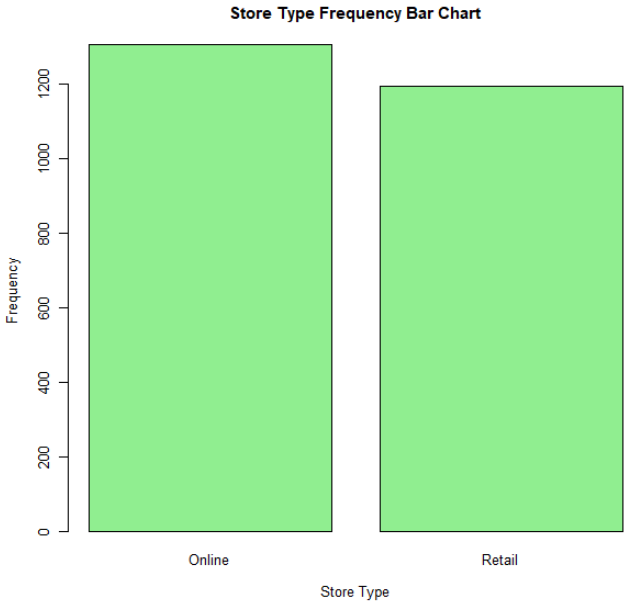

```
> color

Black  Blue Brown Green   Red White
  448   401   406   399   419   427
```

There are 448 records from black, 401 from blue, 406 from brown, 399 from green, 419 from red, and 427 from white colors.



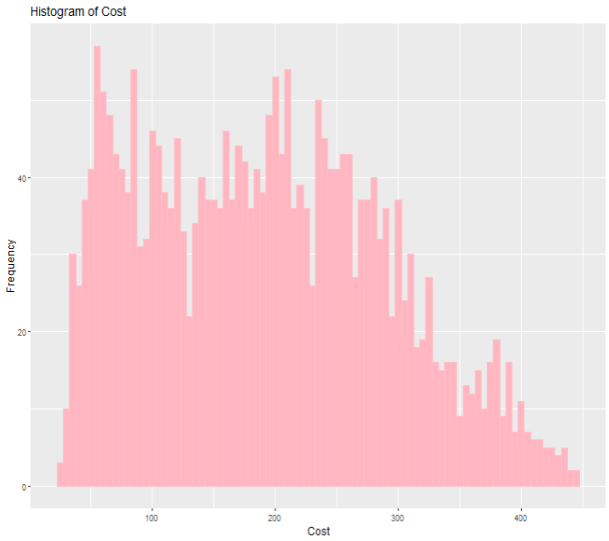

```
> location

   Rural Suburban    Urban
     897      813      790
```

There are 897 records from rural areas, 813 from suburban areas, and 790 from urban areas.



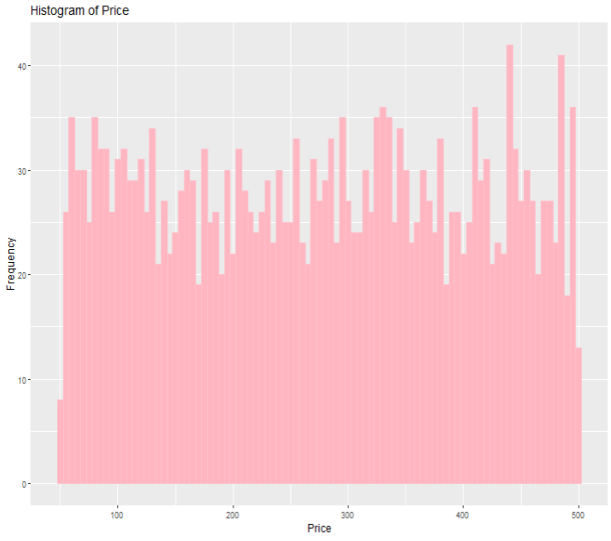

```
> store_type

Online Retail
  1307   1193
```

There are 1307 records for online store and 1193 records for retail store.



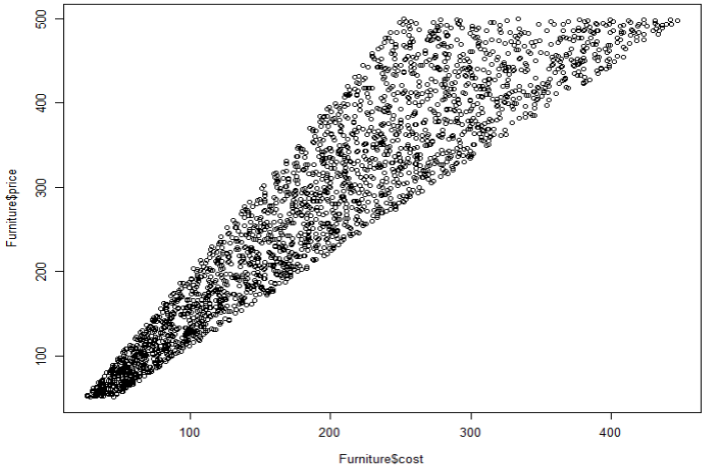

```
> summary(Furniture$cost)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.51  106.40  189.34  191.93  263.19  447.02
```

- We can observe that cost variable is roughly skewed to the right.
- Most of the observations have costs between 52.5 – 57.5 range.
- The smallest observed value of the cost variable is 26.51 while the highest observed value is 447.02.
- Cost variable has a mean of 191.93 and median of 189.34.




```
> summary(Furniture$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   50.7   159.1   277.6   274.5   387.4   499.9
```

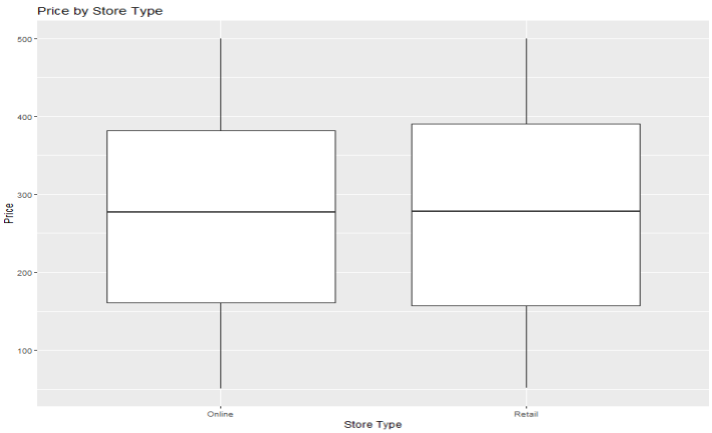
- We can't observe any particular pattern in price variable according to this histogram.
- Most of the observations have prices between 437.5 – 442.5 range.
- The smallest observed value of the price variable is 50.7 while the highest observed value is 499.9.
- Price variable has a mean of 274.5 and median of 277.6.



R = 0.9327871

correlation coefficient of 0.9328 is very close to 1, which implies a very strong positive linear relationship between price and cost variables, indicating that higher prices are typically associated with higher costs.

Therefore, we can use this cost variable as our auxiliary variable.



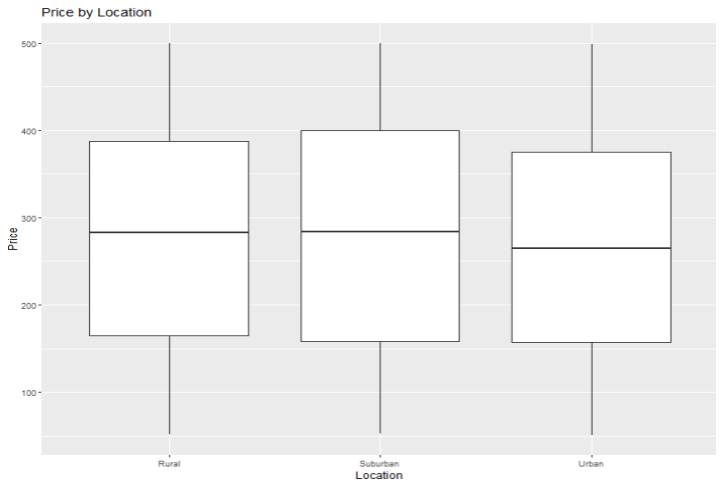
These box plots show how mean values of price vary with the store type.

Both have an approximately normal distribution. There are no any outliers.

We can observe both mean values are approximately equal, which implies that store type has less impact of the price of the furniture.

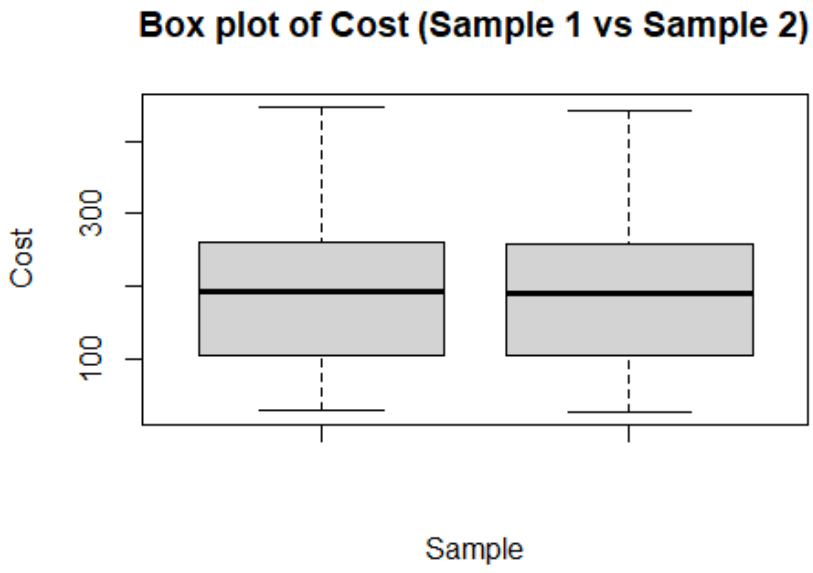
These box plots show how mean values of price vary with the location.

All three types of location have an approximately normal distribution.


Price by Location
500
400
300
200
100
Price
Rural
Suburban
Urban
Location

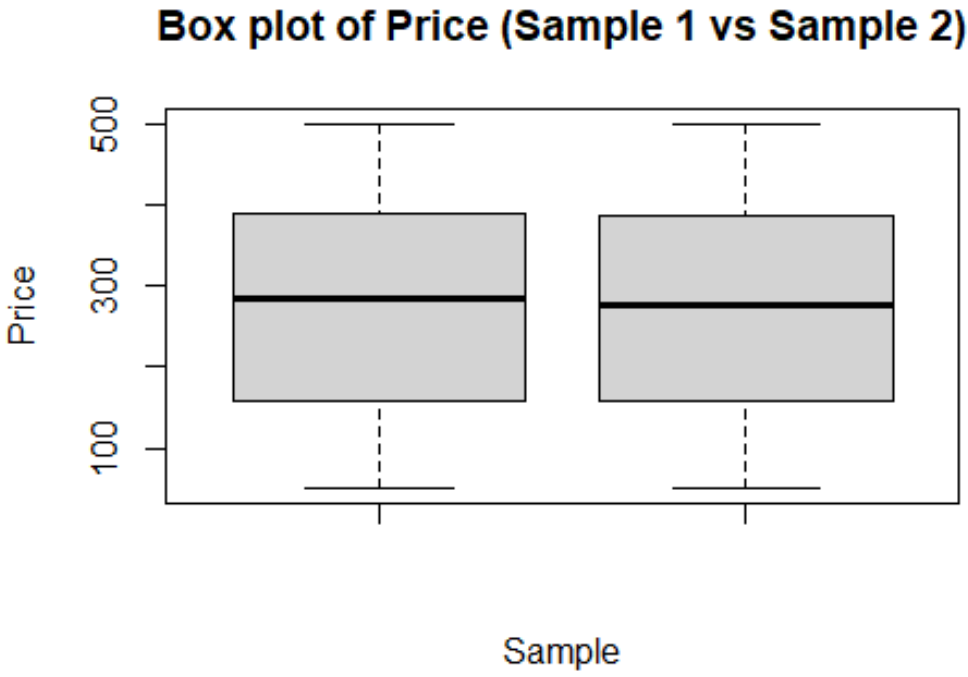
❖ Graphical Analysis about Samples

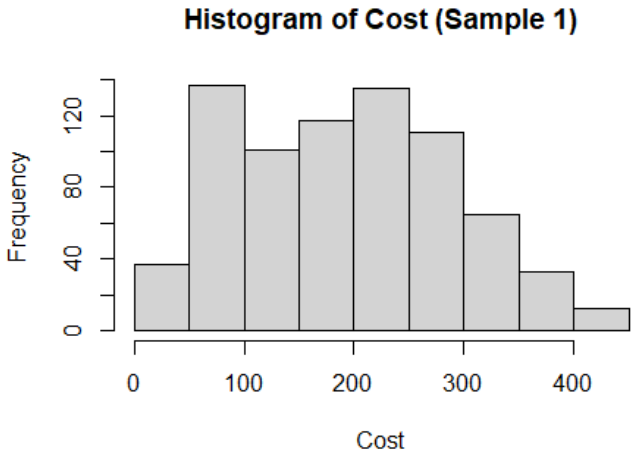
1. Simple Random Sample

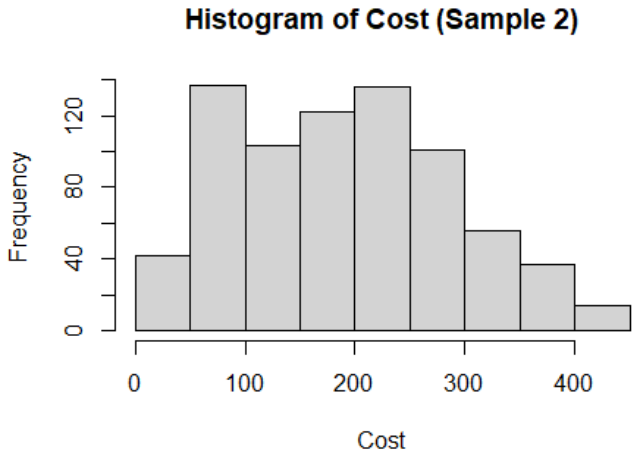


Range for sample 2 is lesser than sample one. For sample one and two medians are almost same. There is no outlies for both plots.

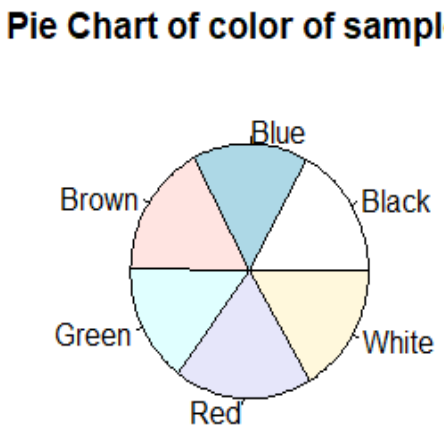
In this graph range for sample 2 is less than range for sample 1.meadin of sample one is slightly greater than sample 2 median.

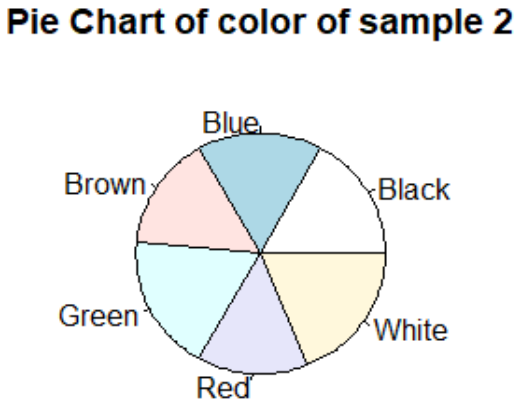






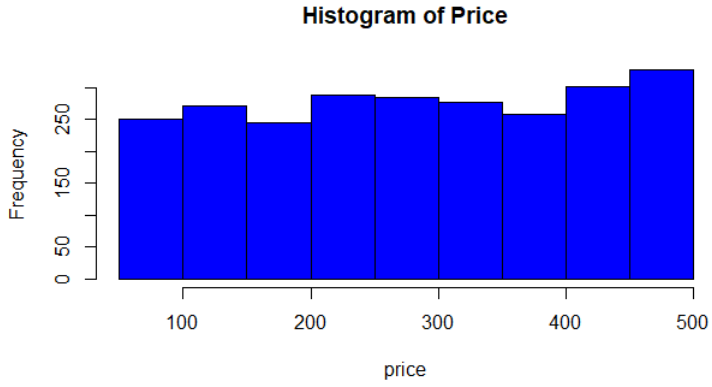
In these two histograms we an see both samples cost variable follow a normal distribution and both plots has same shape almost.



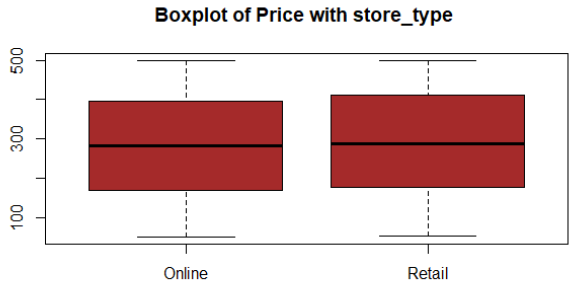


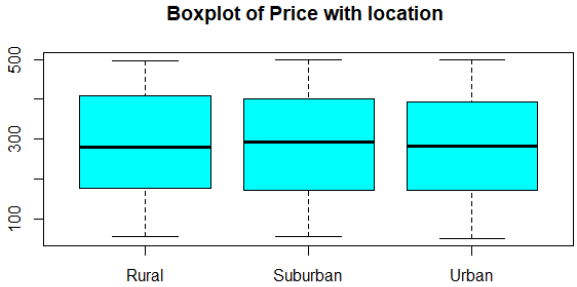
Both sample contain the furniture of all the 6 colors.in sample one they have more furniture in red than the other olours.in sample two they have more furniture in green than the other colors.

2. Stratified Random Sample

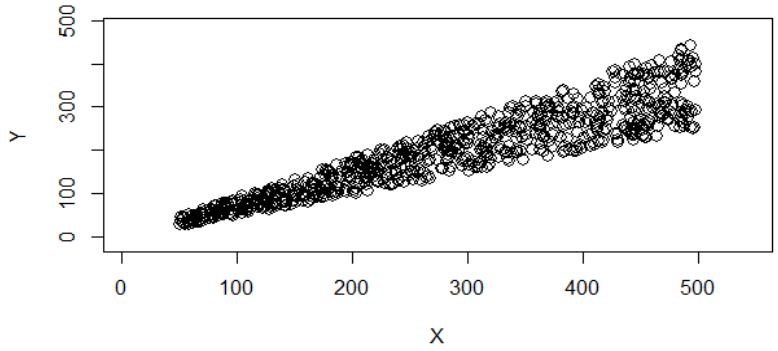


- We can see the Price of the Furniture has roughly uniformly distributed.





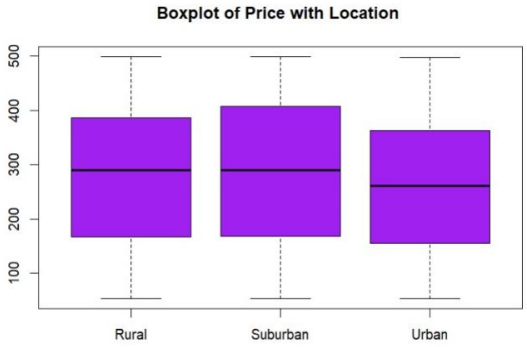
- Box plots are the same as in the Simple Random Sample plots with no outliers to be observed.



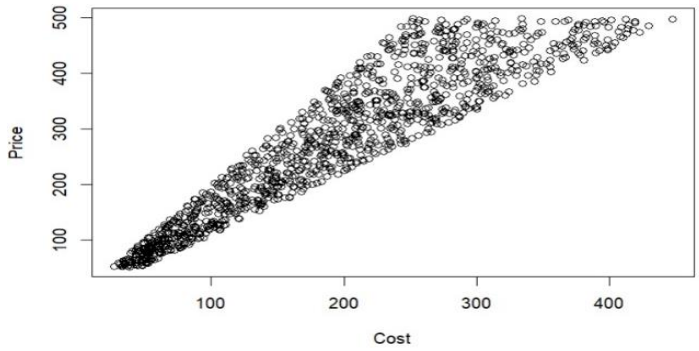
- Price (y) and cost (x) have a strong positive correlation. (since r= 0.9327871)

3. Two - Stage Cluster Sample

- We can see the Price of the Furniture has roughly uniformly distributed.



- Box plots are the same as in the Simple Random Sample and Stratified Random Sample plots with no outliers to be observed.



- Price (y) and Cost (x) have a strong positive correlation which is r = 0.9303849

E. Conclusion of the Analysis

- The results of this study that regards the sampling designs; simple random sampling, stratified random sampling and the two-stage cluster sampling for the Furniture dataset is discussed above.
- Each of the three sampling designs are built twice and compared with each other and with the actual population values.
- When compare the Results of the analysis we can observe population estimation do not differ significantly with the 3 methods of sampling.
- But the standard error of estimations in the Two-staged cluster sampling is lower when compared to the other two sampling methods. Therefore, Two-staged cluster sampling method is more suitable for this dataset.
- In every sampling method regression estimation method gives more precise estimation for the population parameters.
- In graphical analysis, Cost variable follows a positively skewed distribution while Price variable does not follow any particular distribution.
- There is a strong positive linear relationship between Price and Cost variables. This linear relationship helps us to do the ratio and regression analysis.
- We can see furniture prices of both online and retail stores are approximately same. It implies that furniture prices are not dependent on store type.
- Also, we can see furniture prices for both rural and suburban areas are approximately equal while furniture prices for urban areas are seemed to be a little bit lower than rural and suburban areas.