SIT743

Multivariate and Categorical Data Analysis

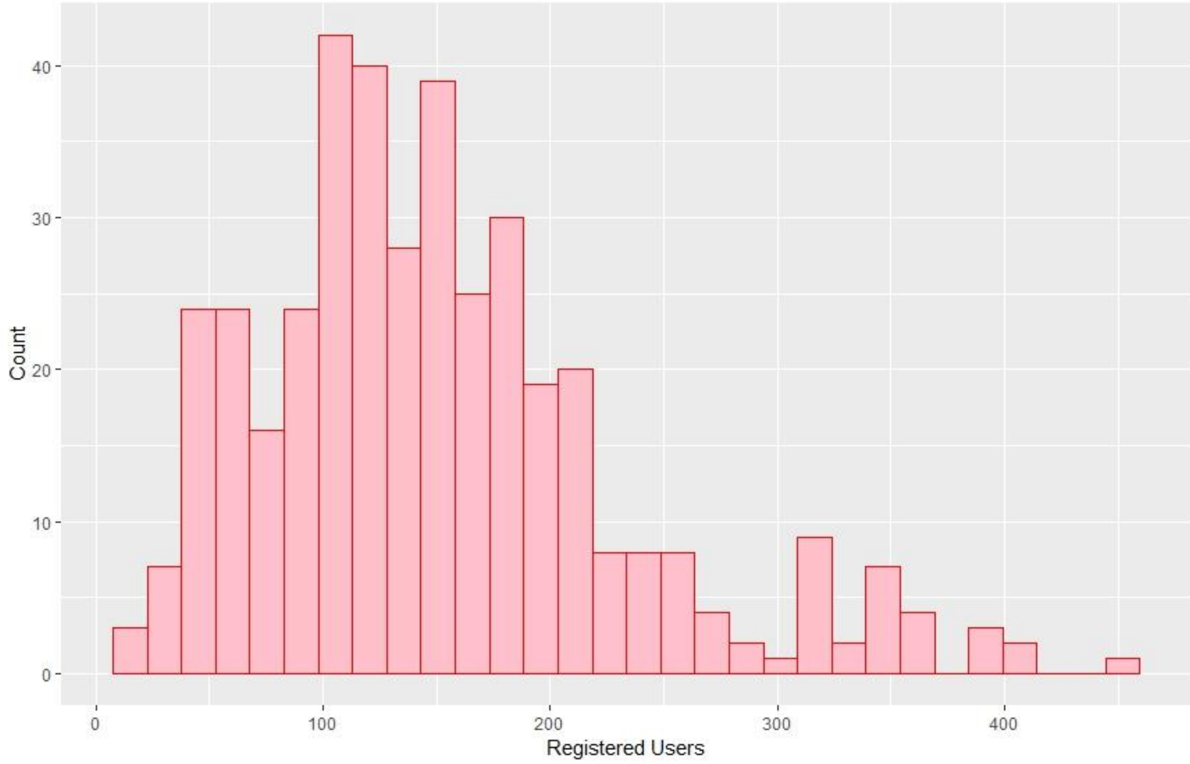
Assignment-1

Deakin University

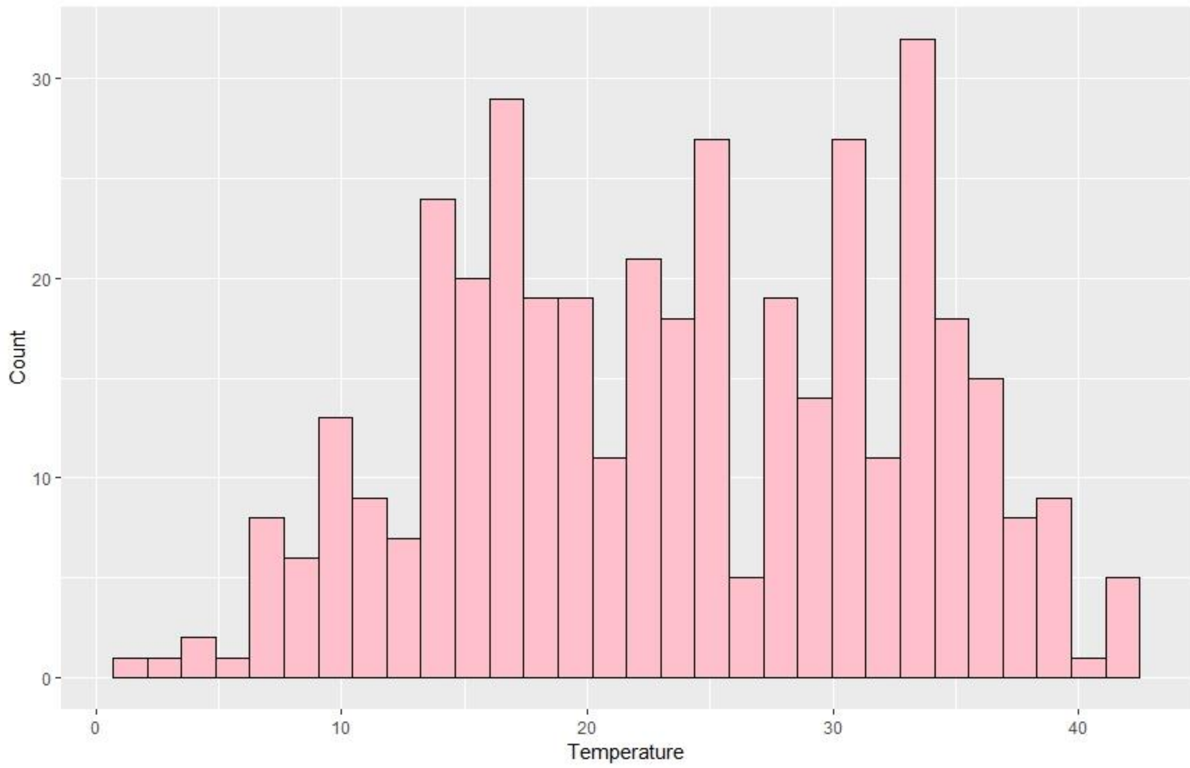
Submitted by: Shantanu Gupta

Student ID: 218200234

1.1)



The majority of the registered users that used a bike at that time is between 25 to 35.

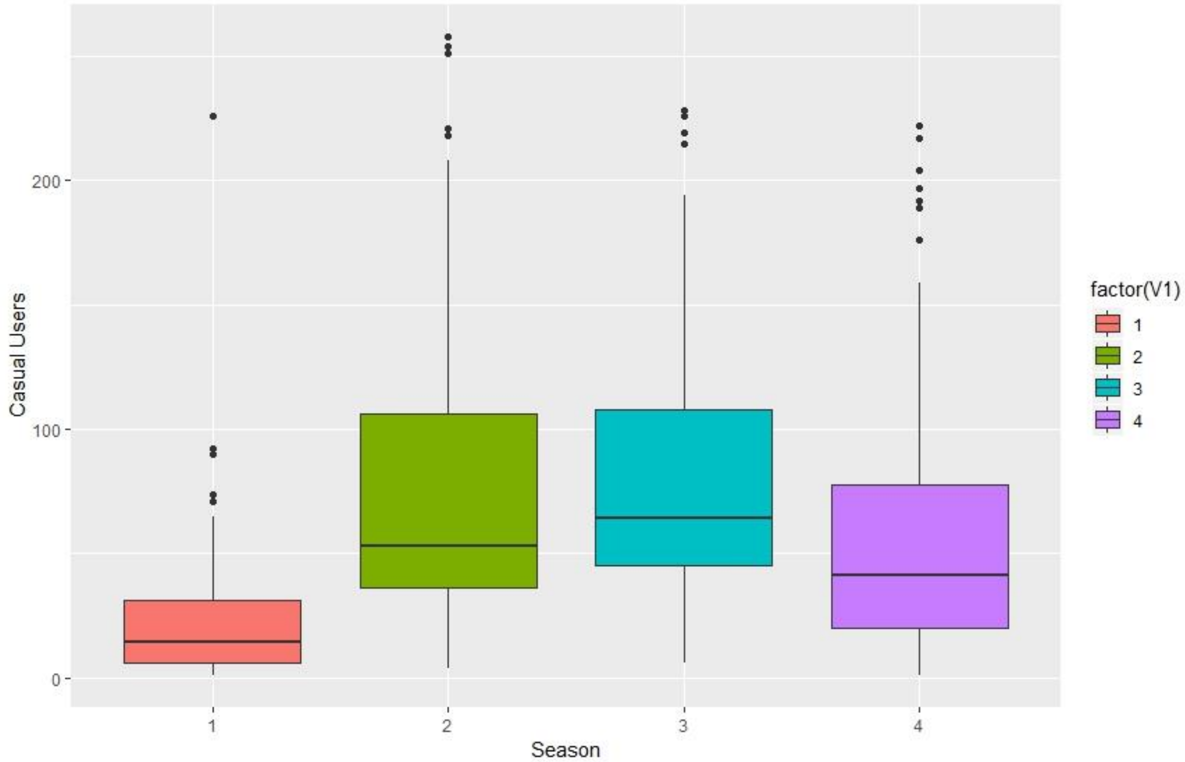


The range of temperature between 11am-12pm in U.S. cities is usually between 15 to 35 in Celsius.

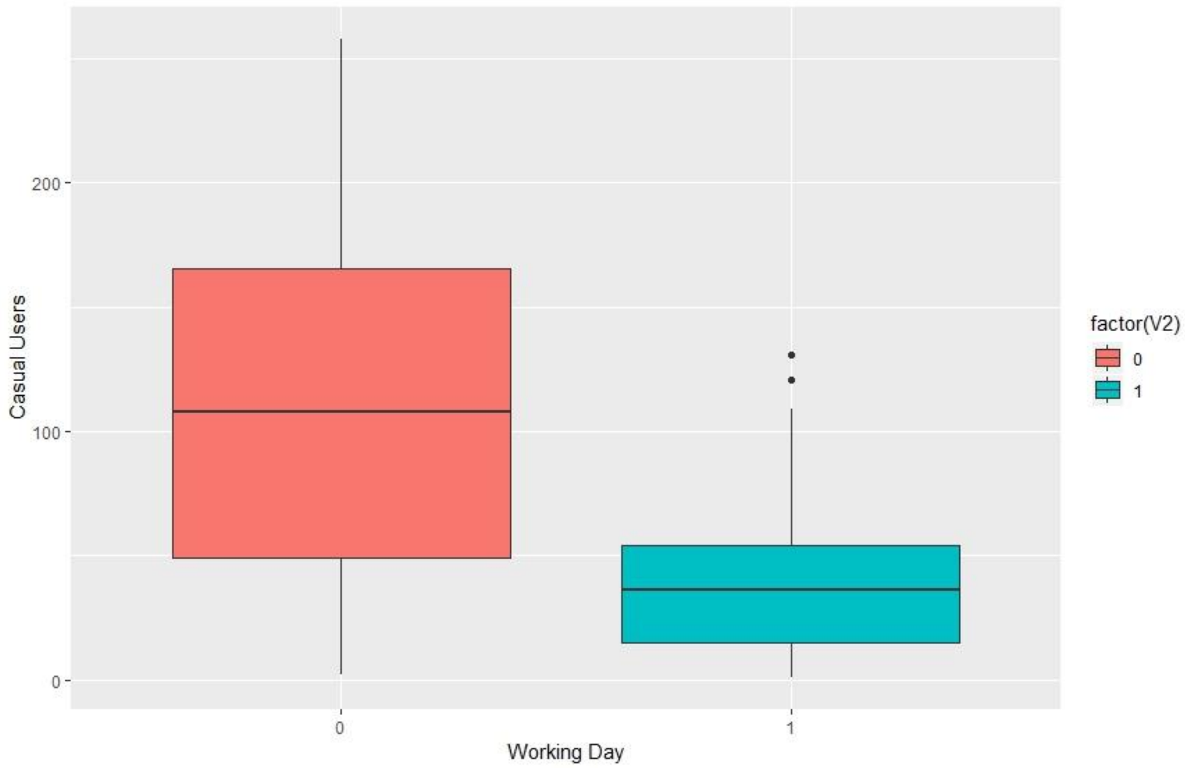
1.2)

```
> #1.2)
> #Five Number Summary for Casual Users
> df=as.data.frame(my.data)
> min(df$V8)
[1] 1
> max(df$V8)
[1] 258
> median(df$V8)
[1] 45
> quantile(df$V8)
  0%  25%  50%  75% 100% 
   1   19   45   76  258 
> fivenum(df$V8)
[1]   1  19  45  76 258
> summary(df$V8)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   19.00   45.00   60.53   76.00  258.00 
> mean(df$V8)
[1] 60.53
> #Five Number Summary for Registered Users
> min(df$V9)
[1] 9
> max(df$V9)
[1] 446
> median(df$V9)
[1] 138
> quantile(df$V9)
  0%  25%  50%  75% 100% 
   9   99  138  187  446 
> fivenum(df$V9)
[1]   9  99 138 187 446
> summary(df$V9)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    9.0    99.0   138.0   150.6   187.0   446.0 
> mean(df$V9)
[1] 150.5525
```

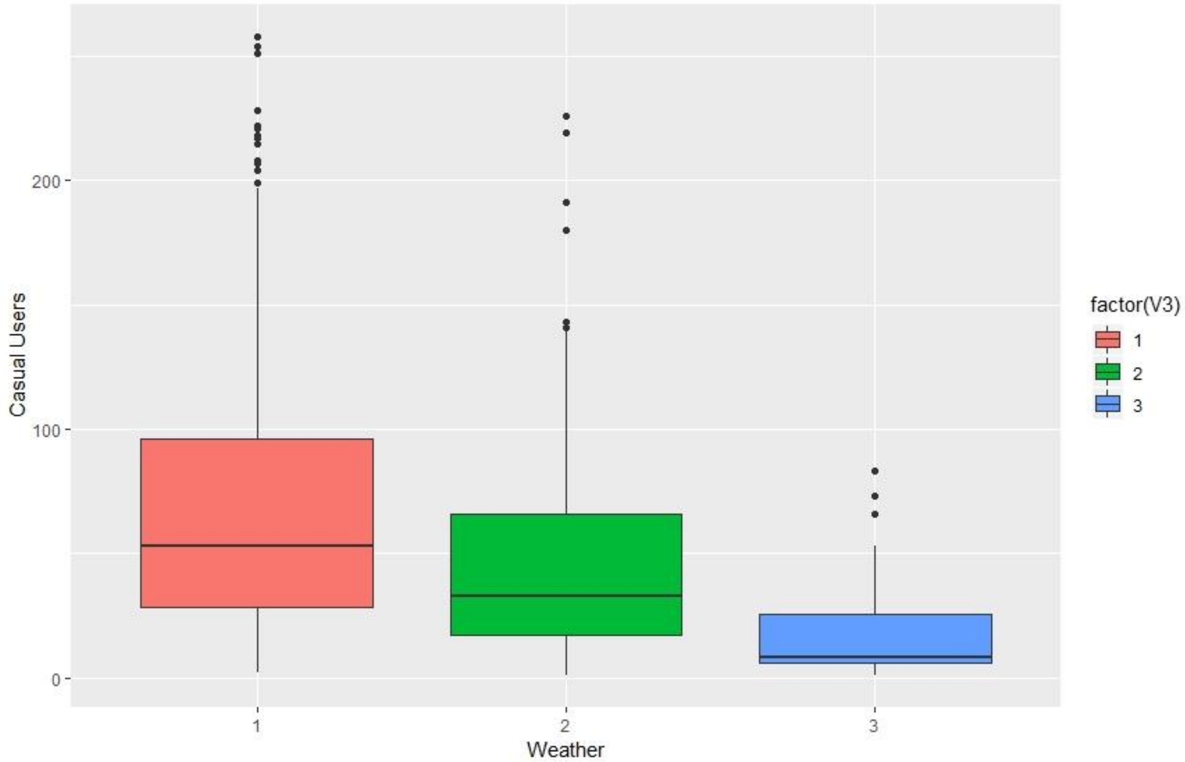
1.3)



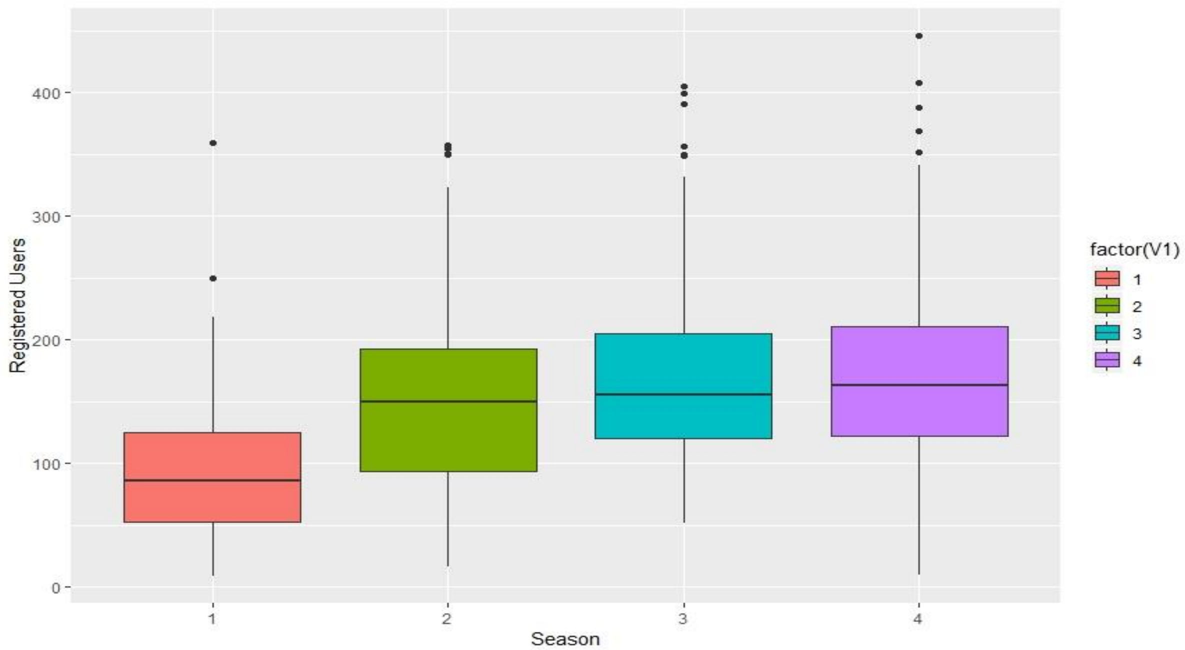
Most of the casual users will rent the bike in summer and autumn season than comparison to spring and winter.



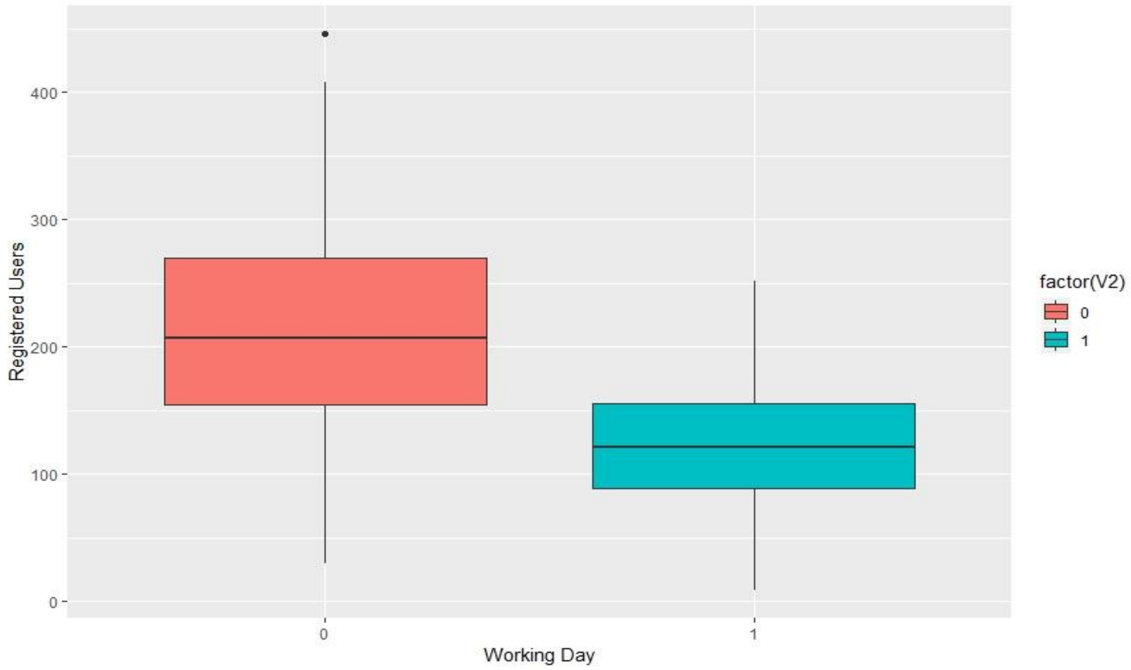
Most Casual users prefer bike in weekend than workday.



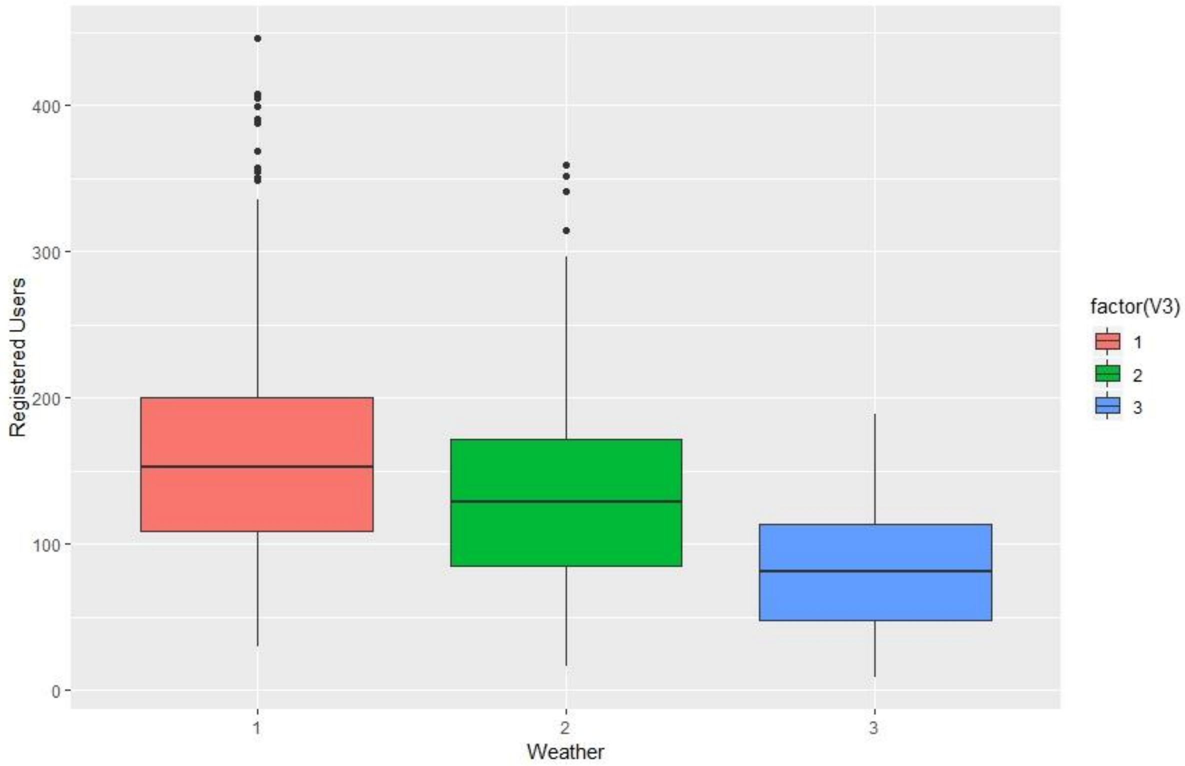
Most of the people to hire bike when there is a clear, few clouds season. No one is renting the bike when there is a heavy storm.



Most of the registered users will take bike equally in summer, autumn and fall season. There are less users in the Spring Season as compared to other three seasons.

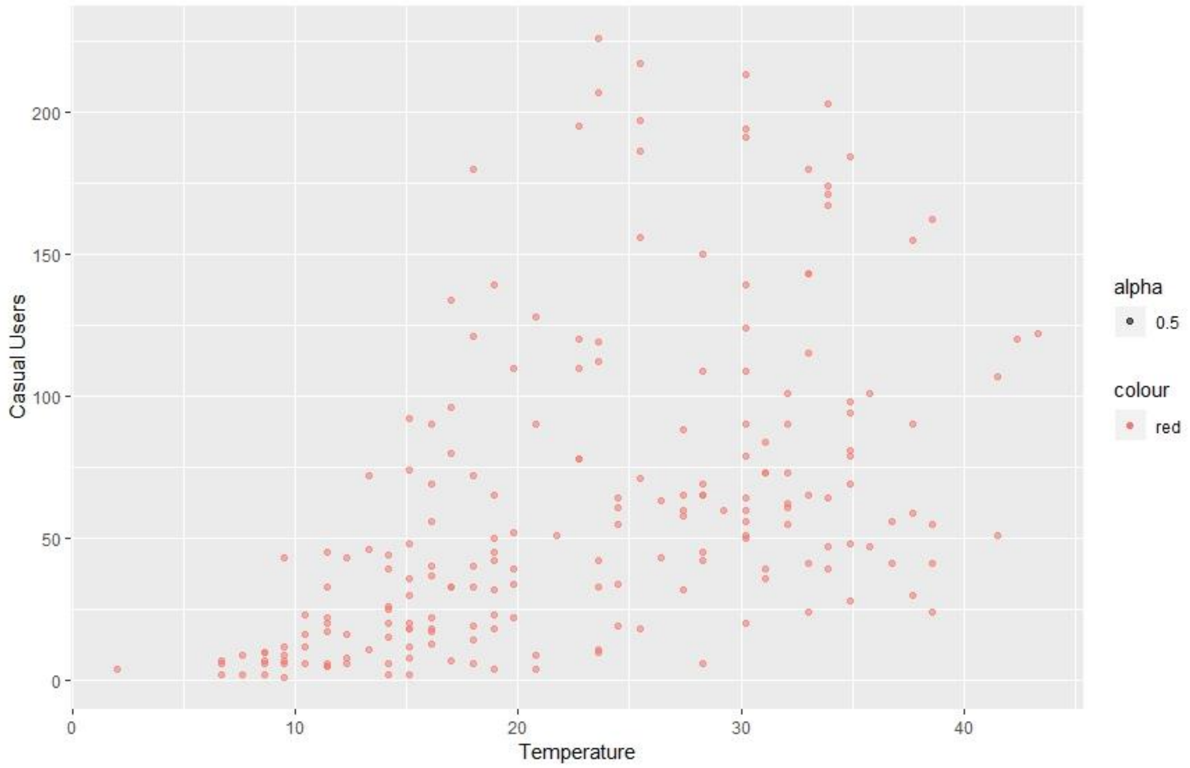


Most registered users prefer bike in weekend than workday. Probably because of it’s a holiday (No-office duties) and people wanted to go out and explore the cities.



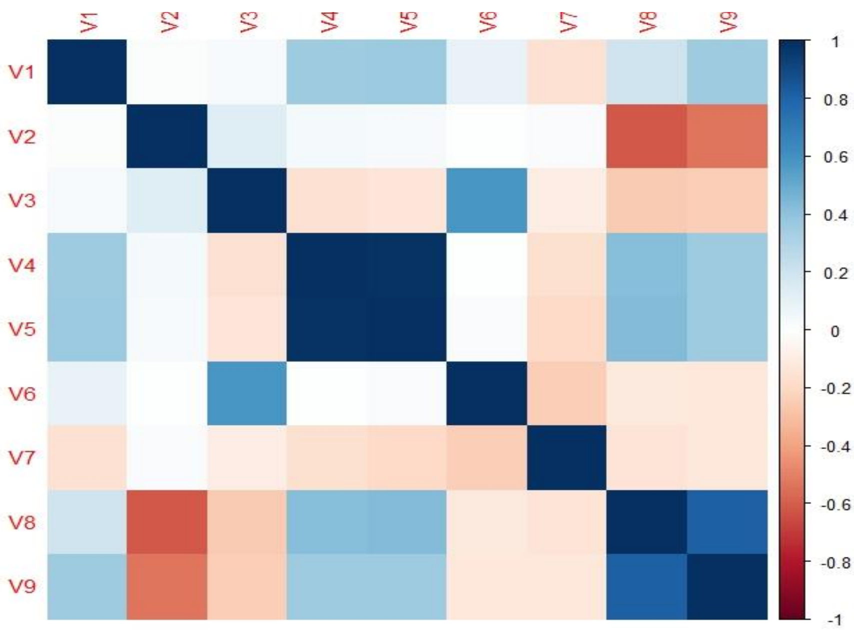
Most of the people to hire bike when there is a clear, few clouds season. No one is renting the bike when there is a heavy storm.

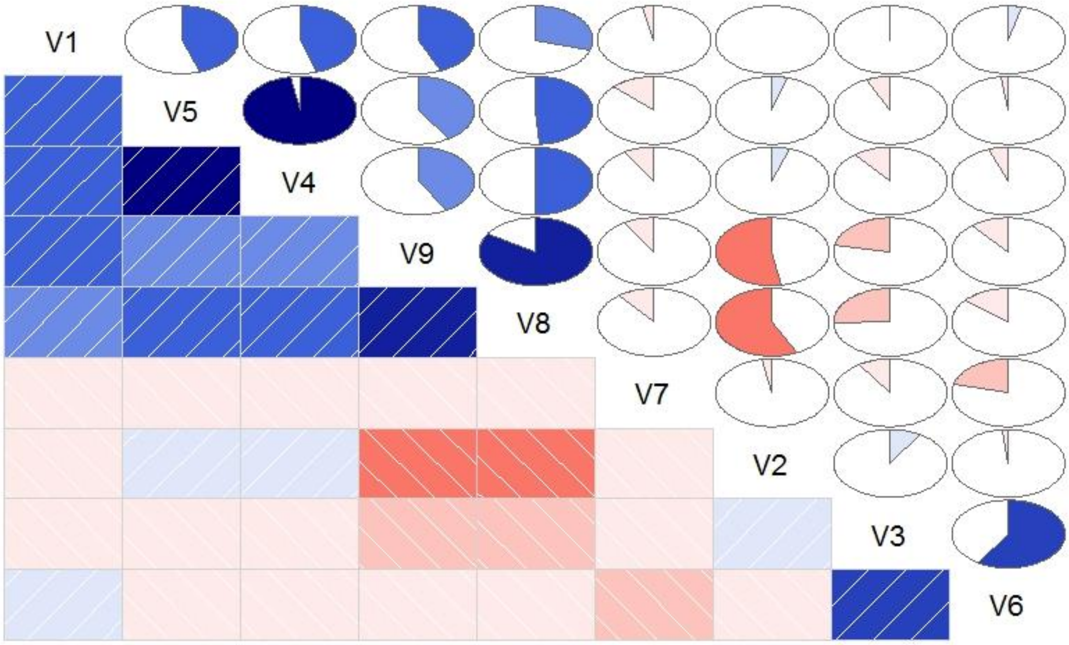
1.4)



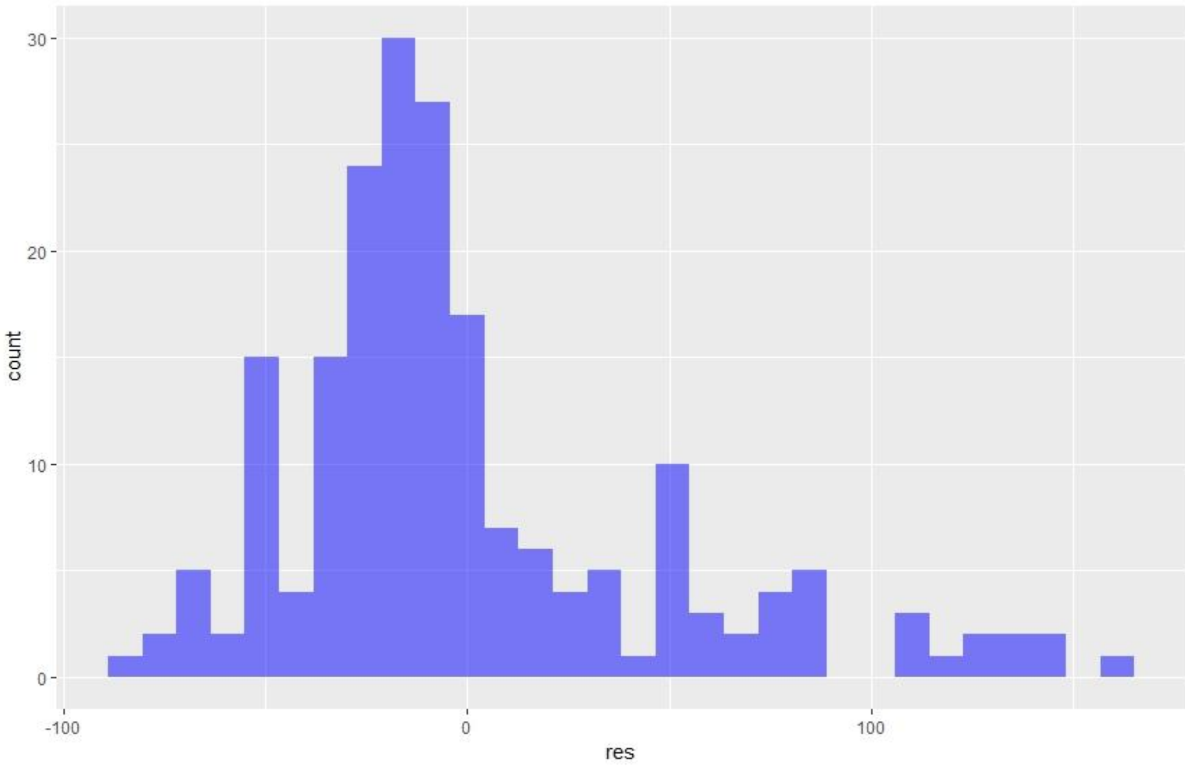
There is a positive correlation between these two variables. More the temperature, more is the casual users.

1.5) Exploratory Data Analysis of Data

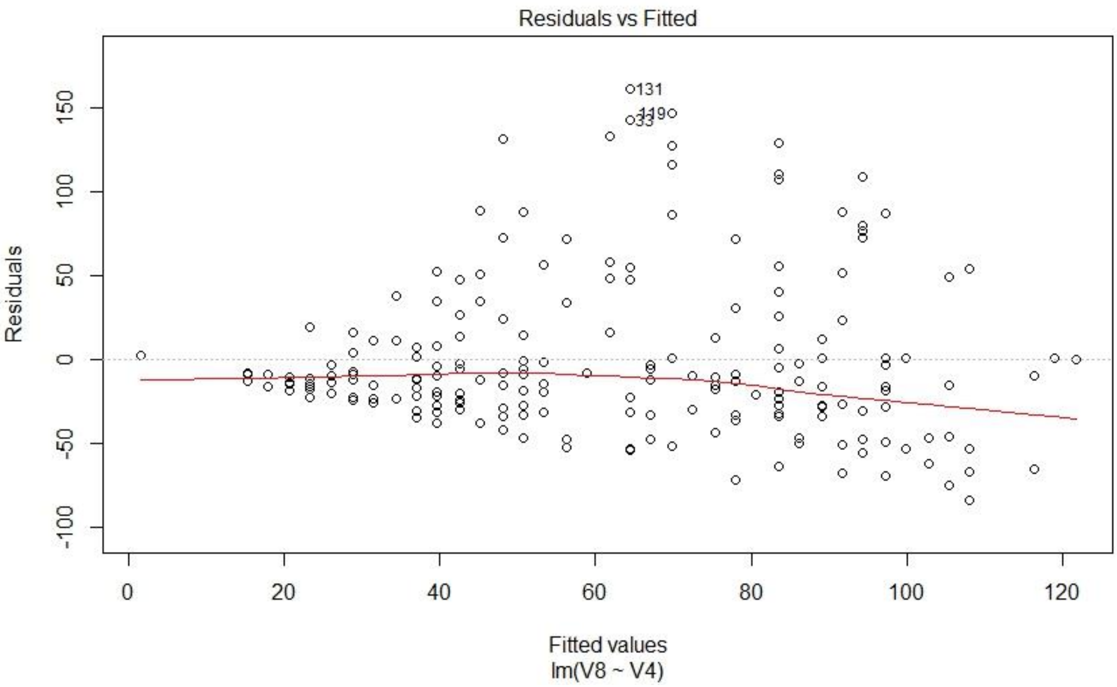


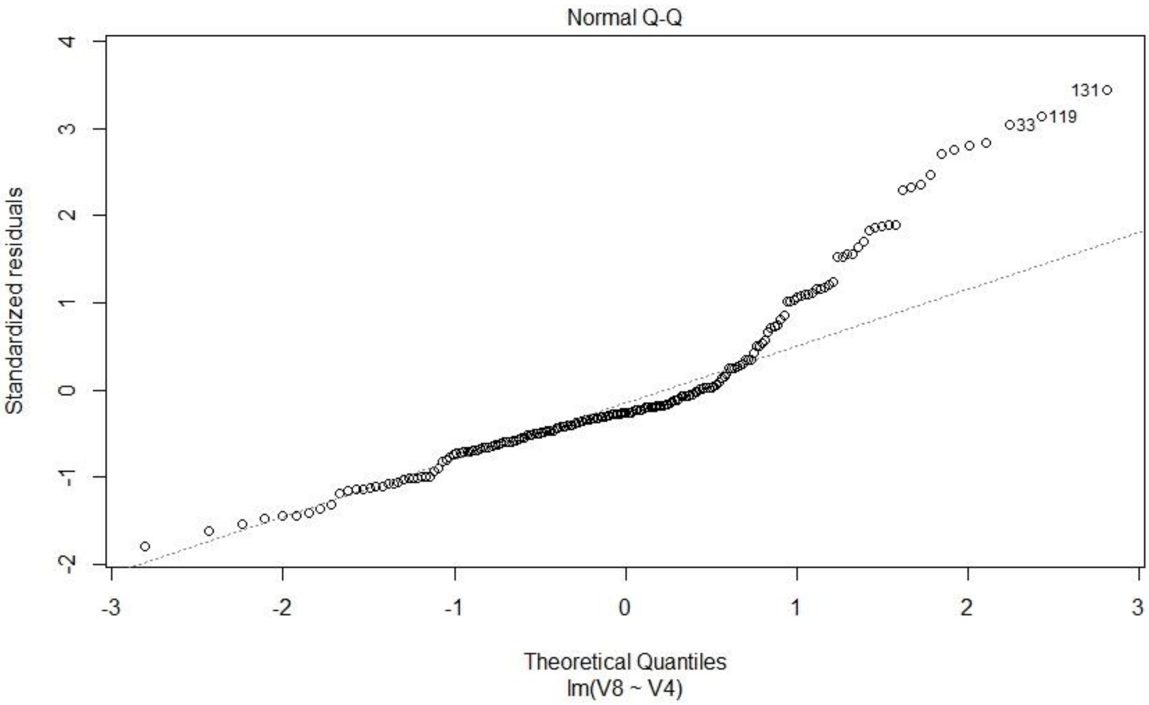


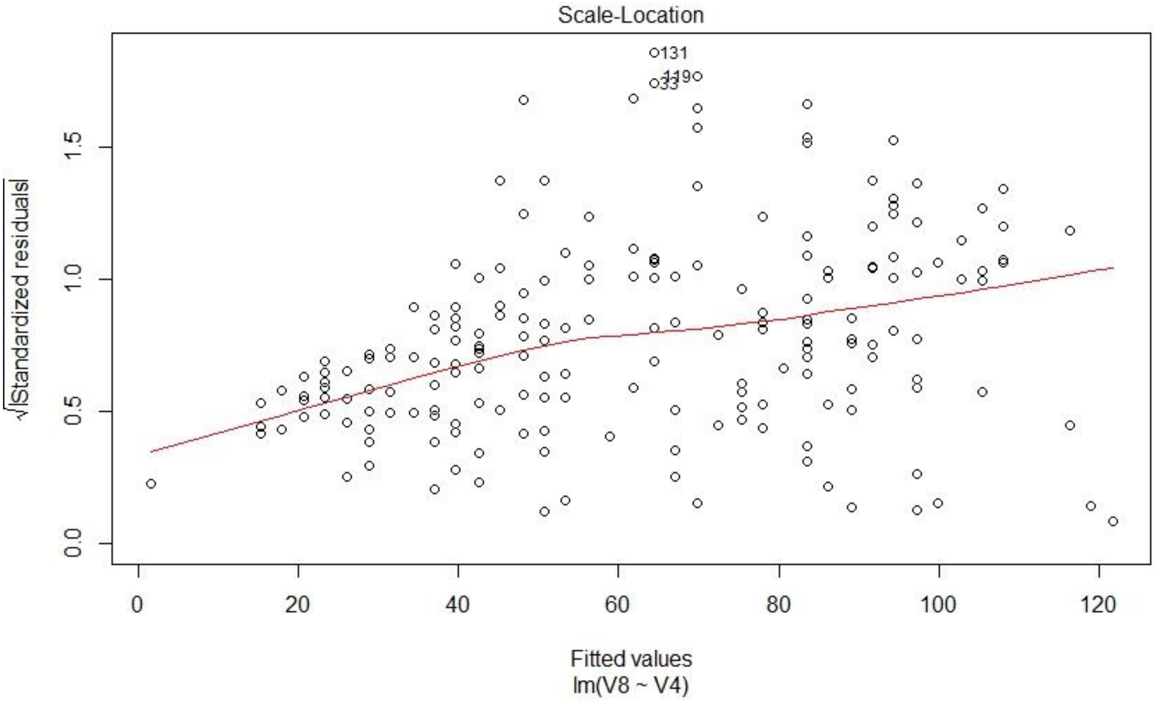

```
> print(cor.data)
            V1          V2          V3           V4          V5           V6          V7         V8         V9
V1  1.00000000  0.01783956  0.03609458  0.350111367  0.36042505  0.097023814 -0.15027376  0.2051622  0.3545110
V2  0.01783956  1.00000000  0.13032494  0.045072521  0.03305743  0.007043480  0.02107149 -0.6116061 -0.5320887
V3  0.03609458  0.13032494  1.00000000 -0.152149089 -0.13569116  0.581610599 -0.09868868 -0.2577931 -0.2454199
V4  0.35011137  0.04507252 -0.15214909  1.000000000  0.98432940  0.003430747 -0.16932824  0.4292105  0.3507902
V5  0.36042505  0.03305743 -0.13569116  0.984329398  1.00000000  0.029110143 -0.19832236  0.4363685  0.3547911
V6  0.09702381  0.00704348  0.58161060  0.003430747  0.02911014  1.000000000 -0.24531030 -0.1180926 -0.1263170
V7 -0.15027376  0.02107149 -0.09868868 -0.169328236 -0.19832236 -0.245310300  1.00000000 -0.1491998 -0.1202972
V8  0.20516222 -0.61160606 -0.25779313  0.429210461  0.43636855 -0.118092649 -0.14919983  1.0000000  0.8125283
V9  0.35451101 -0.53208872 -0.24541992  0.350790220  0.35479113 -0.126316959 -0.12029724  0.8125283  1.0000000
```

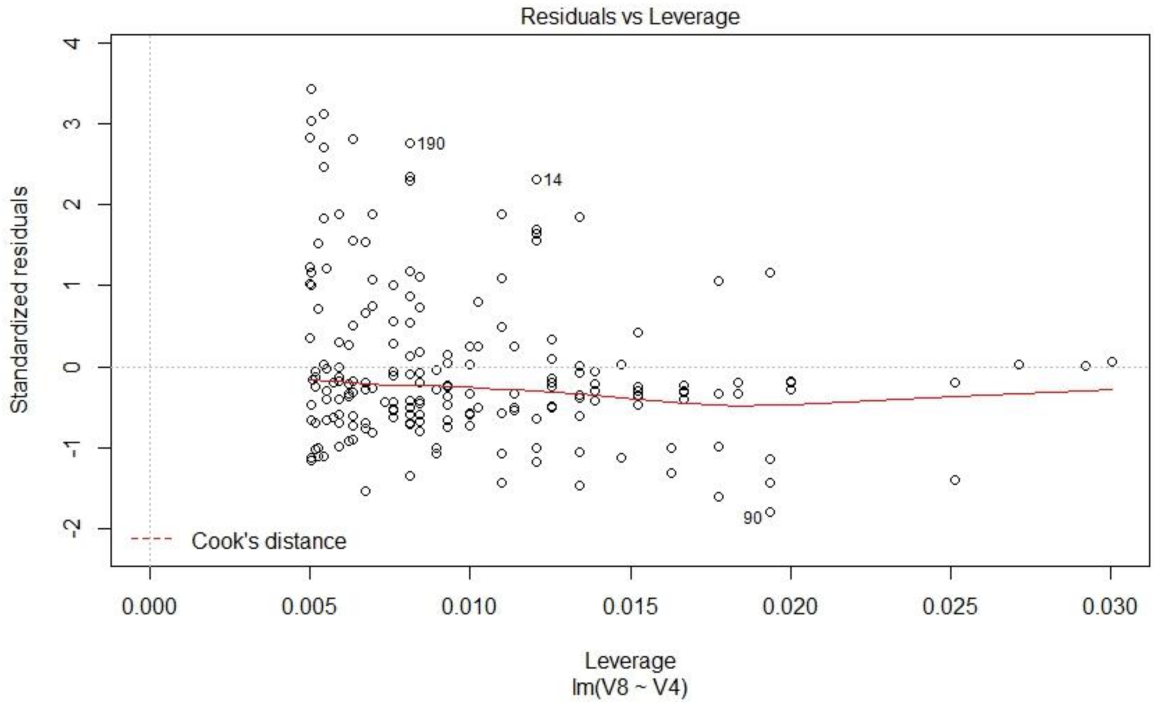


```
> head(res)
         res
1 -13.0361915
2   0.9542602
3 -15.8306845
4 -47.2226020
5  25.4376991
6  34.3557260
```






Scale-Location
√|Standardized residuals|
Fitted values
lm(V8 ~ V4)
131
119
33


Residuals vs Leverage
Standardized residuals
Leverage
lm(V8 ~ V4)
Cook's distance
190
14
90

```
> #Linear Regression Model
> model<-lm(V8~V4,change2)
> summary(model)

Call:
lm(formula = V8 ~ V4, data = change2)

Residuals:
   Min     1Q Median     3Q    Max 
-83.99 -27.66 -12.10  13.66 161.63 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -4.2738     8.8191  -0.485    0.628    
V4            2.9085     0.3578   8.129 4.64e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.1 on 198 degrees of freedom
Multiple R-squared:  0.2502,	Adjusted R-squared:  0.2464 
F-statistic: 66.08 on 1 and 198 DF,  p-value: 4.637e-14
```

```
> #Correlation Coefficient
> cor.test(~V8+V4,data=change2,method="pearson")

	Pearson's product-moment correlation

data:  V8 and V4
t = 8.1287, df = 198, p-value = 4.637e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3884332 0.5974915
sample estimates:
      cor 
0.5002168 
```

#Linear Regression Equation

Casual User=-4.2738+2.9085 x Temperature

R=0.5002168

Adj.R^2=0.2464

There is a three star with this temperature variable, so it is considered as an important variable. Even the correlation is not strong enough but there is some correlation in data. So temperature is one of the factor helps in determine the number of the casual users will rent the bike in the future in this time period.

Q2)

2.1) 1330/4000

2.2) 690/4000

2.3) 1360/4000

2.4) 240/1050

2.5) 810/1050

2.6)1330/4000+3310/4000-1140/4000 or 0.875

2.7)

	N	V	Q	Marginal Distribution
P				3310/4000
C				690/4000

2.8)

	N	V	Q	Total
P				
C				
Total				
Marginal Distribution	1620/4000	1330/4000	1050/4000	

2.9)

	N	V	Q	Total
P	1360/1620	1140/1330	810/1050	
C	260/1620	190/1330	240/1050	
Total				

Q3)

P(Smokers)=0.20

P(Non-Smoker)=1-P(Smoker)=0.80

P(Smokers and Lung Cancer)=0.20*0.60=0.12

P(Non-Smoker and Lung Cancer)=0.80*0.15=0.120

P(Lung Cancer was a Smoker)=0.50

Q5)

5.1)

A likelihood is generally fixed by our model, and the evidence (i.e. denominator part) is fixed by our data. What we can vary though is a prior, and it is our own choice. So, If the prior and the posterior lie in the same family of distributions then we called is a conjugate prior. The prior is said to be conjugate of the likelihood function. For example, if the prior was normal parameterized by some parameters mu and sigma, then we'd expect the posterior to be also normal but with some other mean and variance.

5.2)
1-We can Get Exact Posterior by using conjugate prior.
2-Easy for Online Learning
For example we have known that Beta Distribution is a conjugate prior of the Bernoulli function. So, the posterior can be easily approximated with this simple formula
Posterior probability=B (N_1+a, N2+b)
Where
B=Beta-Distribution
a,b=constants of beta function
N_1,N_2=parameters of bernoulli function
and you can easily plug it in as you get more and more data.

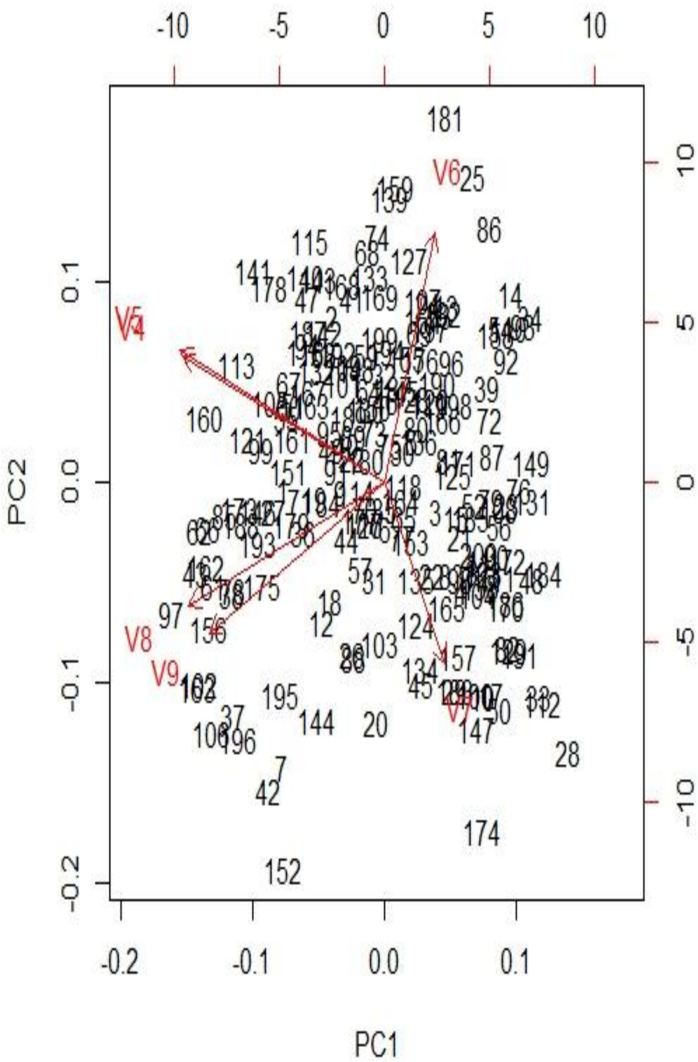
5.3)
Beta Distribution is conjugate of the Bernoulli function
Gamma Distribution is conjugate of the Poission function
Beta Distribution is conjugate of the Binomial function
Dirichlet distribution is conjugate of the Categorical Variable

5.4) Picture is attached

Q6)

6.1)

```
> summary(pZ)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.6702 1.1286 1.0159 0.8568 0.40402 0.08475
Proportion of Variance 0.4649 0.2123 0.1720 0.1224 0.02721 0.00120
Cumulative Proportion  0.4649 0.6772 0.8492 0.9716 0.99880 1.00000
```



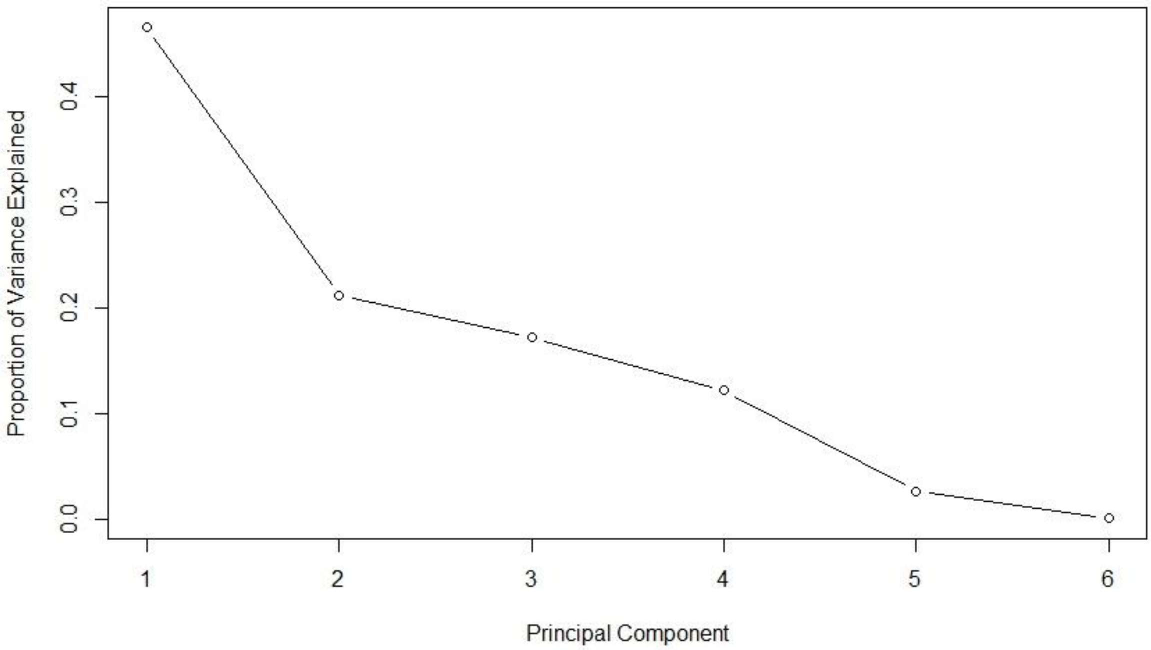
Principal component 1 is basically the reflection of close to 47%of the variations in the data.

Combination of Principal component 1 and 2 is basically helps in understanding 68%of the variations in the data.

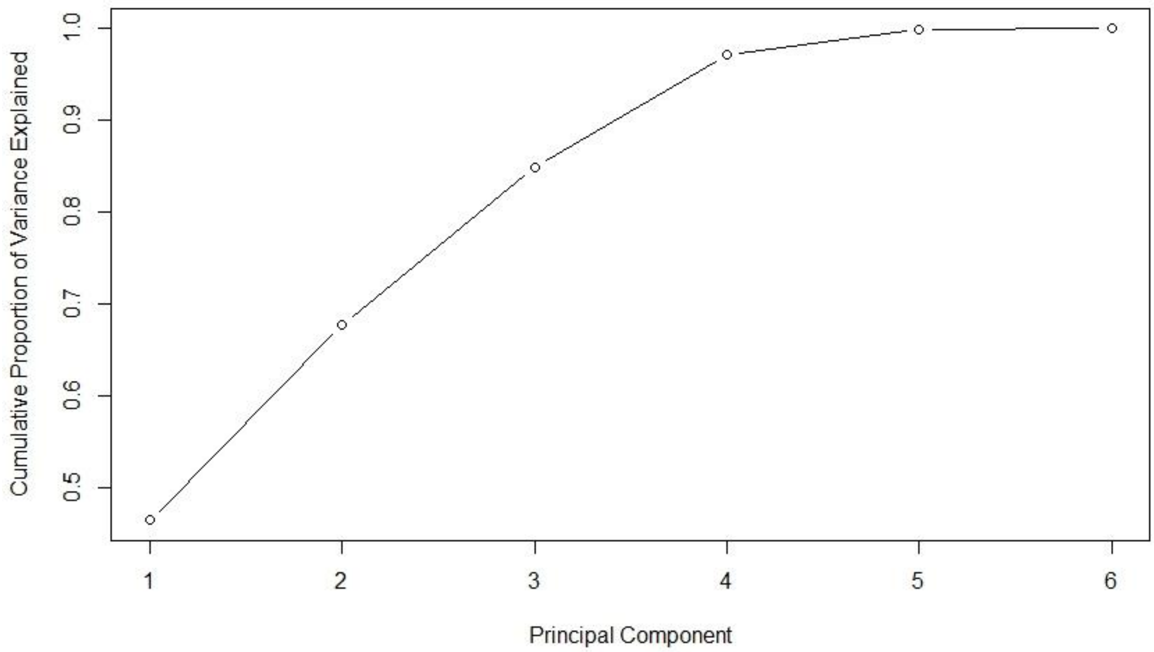
First 4 components collectively describe up to 97% of the data.

Instead of looking for the six dimensions of data, we can infer the good predictions by looking only first 4 dimensions.

6.2)



There are lot of school of thoughts in choosing the best principal component for the data. Generally, more than 85% of the data is captured by the principal component is considered as a good. If we take 3 principal components in this case then we are able to explain very clearly rather than taking 6 dimensions.




```
> #6.2)
> #Conduct Principal Component Analysis
> names(pZ)
[1] "sdev"     "rotation" "center"   "scale"    "x"
> #outputs theprincipal component loading
> pZ$rotation
          PC1        PC2        PC3           PC4         PC5          PC6
V4 -0.5088171  0.3083678  0.3841058 -0.0147539882 -0.06427905 -0.702942201
V5 -0.5130520  0.3226336  0.3536515 -0.0005304772 -0.04991665  0.710720595
V6  0.1275264  0.6123683 -0.3253494  0.7074519480  0.04456861 -0.020377380
V7  0.1543567 -0.4436469  0.6115800  0.6356317617  0.03412140  0.011372531
V8 -0.4929901 -0.3028765 -0.3060210  0.1633981300  0.73802333 -0.014155058
V9 -0.4413100 -0.3704931 -0.3922527  0.2618531924 -0.66749342 -0.001886533
```

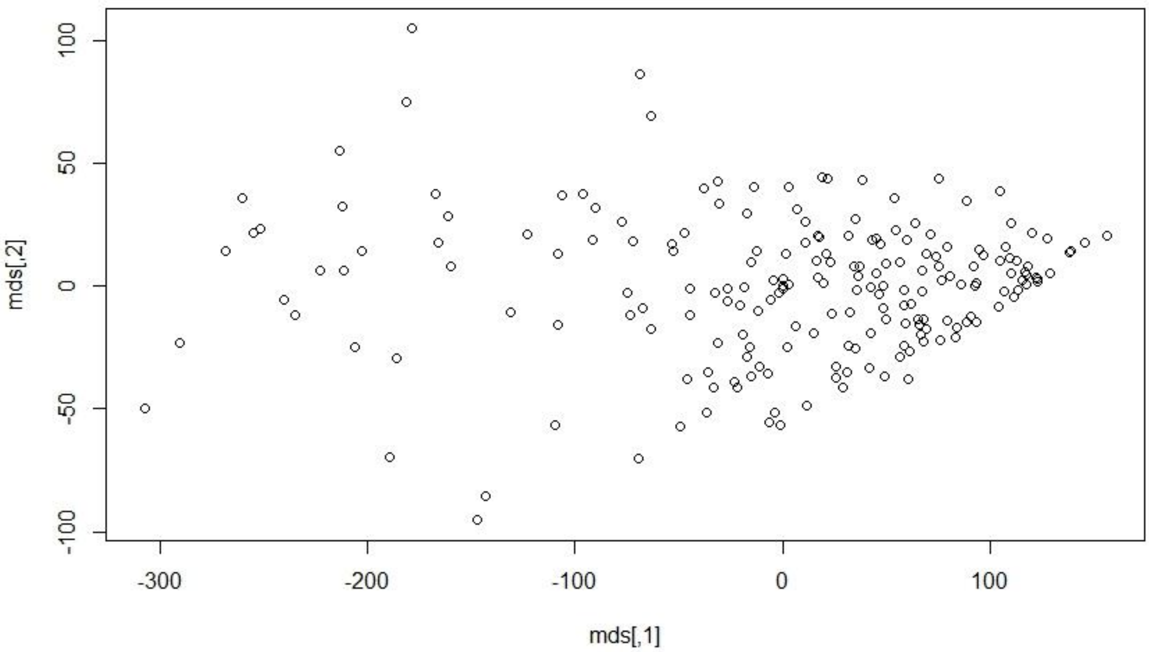
```
> #outputs the mean of variables
> pz$center
      V4       V5       V6       V7       V8       V9
 23.8535  16.9900  57.6150  14.1000  63.9200 151.9150
> #outputs the standard deviation of variables
> pz$scale
       V4        V5        V6        V7        V8        V9
 9.594350 12.138228 17.729310  8.312973 58.309378 84.652454
> #matrix x has the principal component score vectors in a 200x6 dimension
> pz$x
              PC1         PC2          PC3         PC4           PC5           PC6
  [1,] -1.80175705 -0.05711150  0.153869848  0.81523989 -0.3190003587  1.907878e-02
  [2,] -0.91876466  1.32187550 -0.383541810  0.55796074 -0.1435336533 -3.770991e-02
  [3,]  0.93244763 -0.23439394  0.678166884 -0.48201187 -0.2439367569  7.797912e-02
  [4,]  1.39692006 -0.85194120 -0.559061104 -0.25591130  0.5723990012 -2.424928e-02
  [5,]  0.50275590  0.32993310 -0.823450850  0.30819184  0.0876303108  5.085951e-02
  [6,]  0.02582446 -0.36770637 -2.388883149  0.74804968 -1.0550703595  2.679424e-02
  [7,] -1.82109340 -2.26736766 -1.824445333 -0.35928373  0.1533912173  5.459190e-02
  [8,]  1.59809154 -0.70914953  0.186006849 -0.49583474 -0.3133874495 -2.923312e-02
  [9,] -0.77665993 -0.07173260  0.468260741 -0.89464043 -0.1278776741  1.577498e-01
 [10,]  1.76778051 -1.70870501  0.758133422 -0.86591979  0.1965187036 -4.988892e-02
 [11,]  1.68082670 -0.63960680  0.600420362 -0.90572468  0.4278442366 -3.961786e-02
 [12,] -1.12595410 -1.11468033 -2.146054474  0.23255672 -0.3124306785  4.497046e-02
```

```
> prop_varex<-pr_variance/sum(pr_variance)
> prop_varex
[1] 0.464935169 0.212292279 0.172022743 0.122347281 0.027205428 0.001197099
> #proportion of variance explained
```

```
> #compute standard deviation of each principal component
> std_dev<-pz$sdev
> #compute variance
> pr_variance<-std_dev^2
> pr_variance
[1] 2.789611013 1.273753673 1.032136460 0.734083689 0.163232570 0.007182594
```

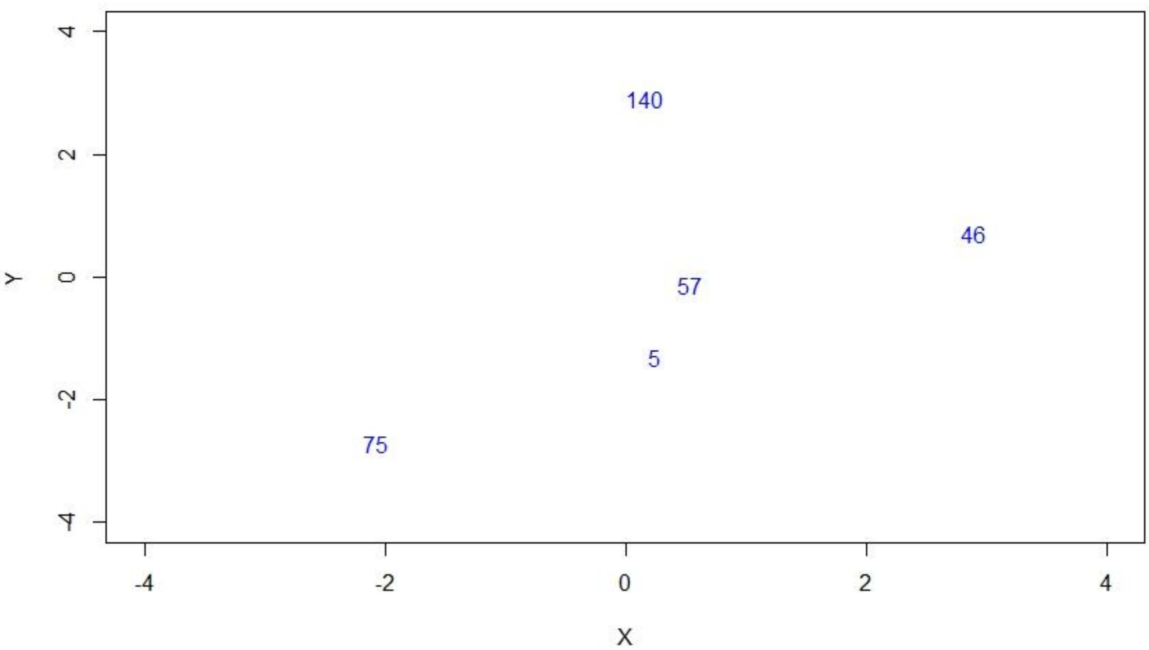
6.3) the scatterplot of the two -dimensional projections does not shows a clear separation between these points. Normally, MDS is used to provide a visual representation of a complex set of relationships that can be scanned at a glance.

CLASSICAL MULTIDIMENSIONAL SCALING

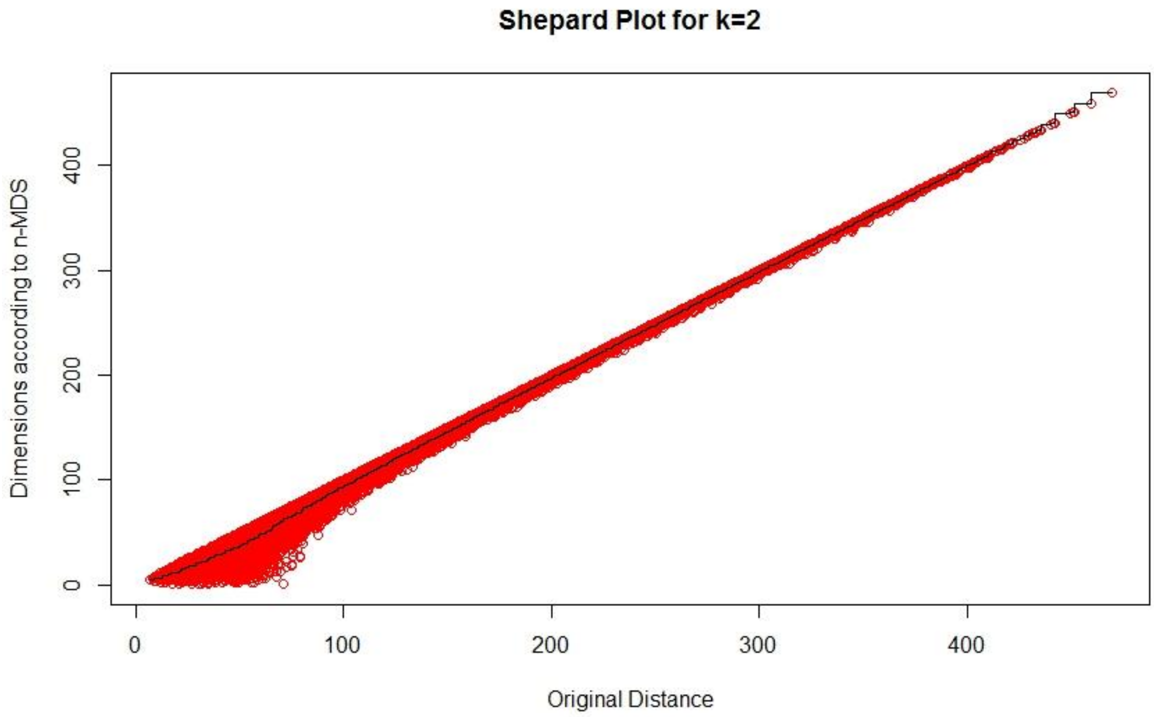


6.4)

NON-METRIC MULTIDIMENSIONAL SCALING

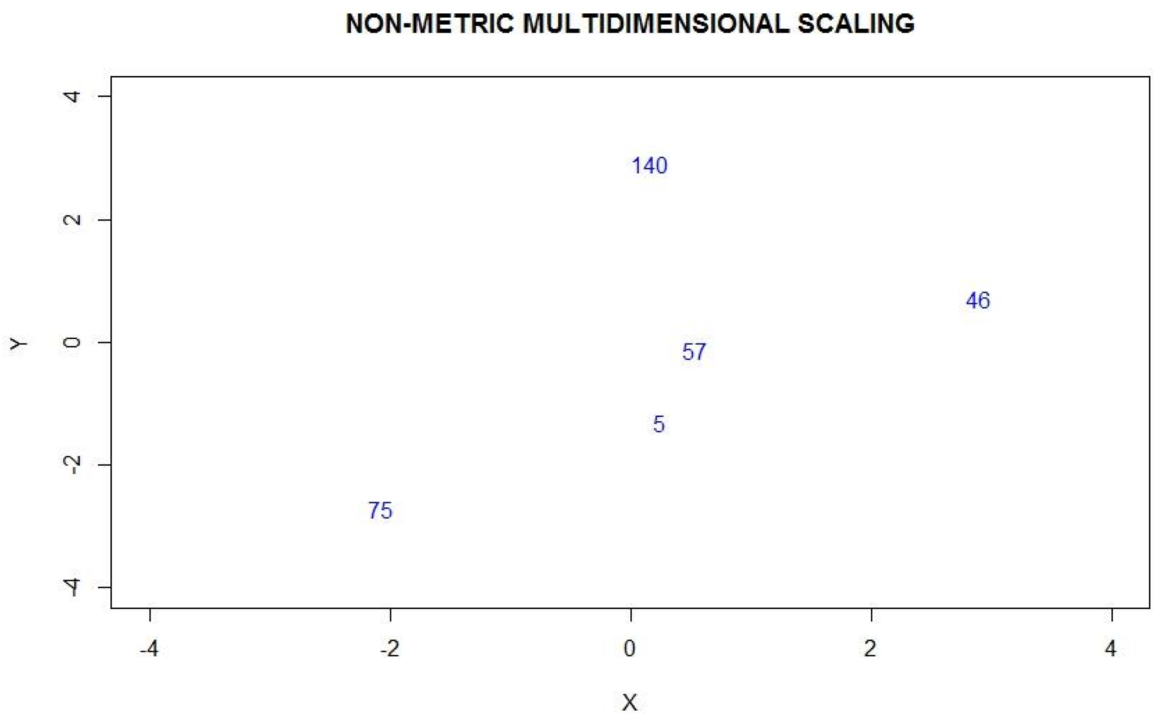


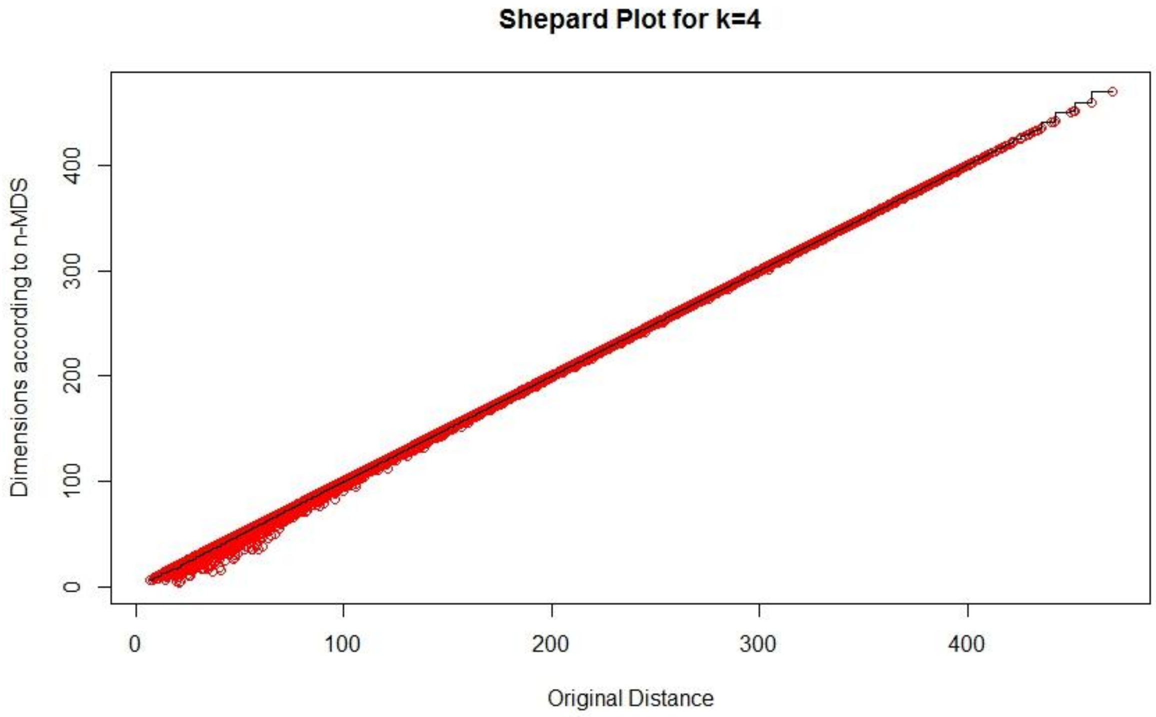
6.5)



I find that there is a good agreement between the observed proximities and the inter-point distances in the Bike share data.

6.6)



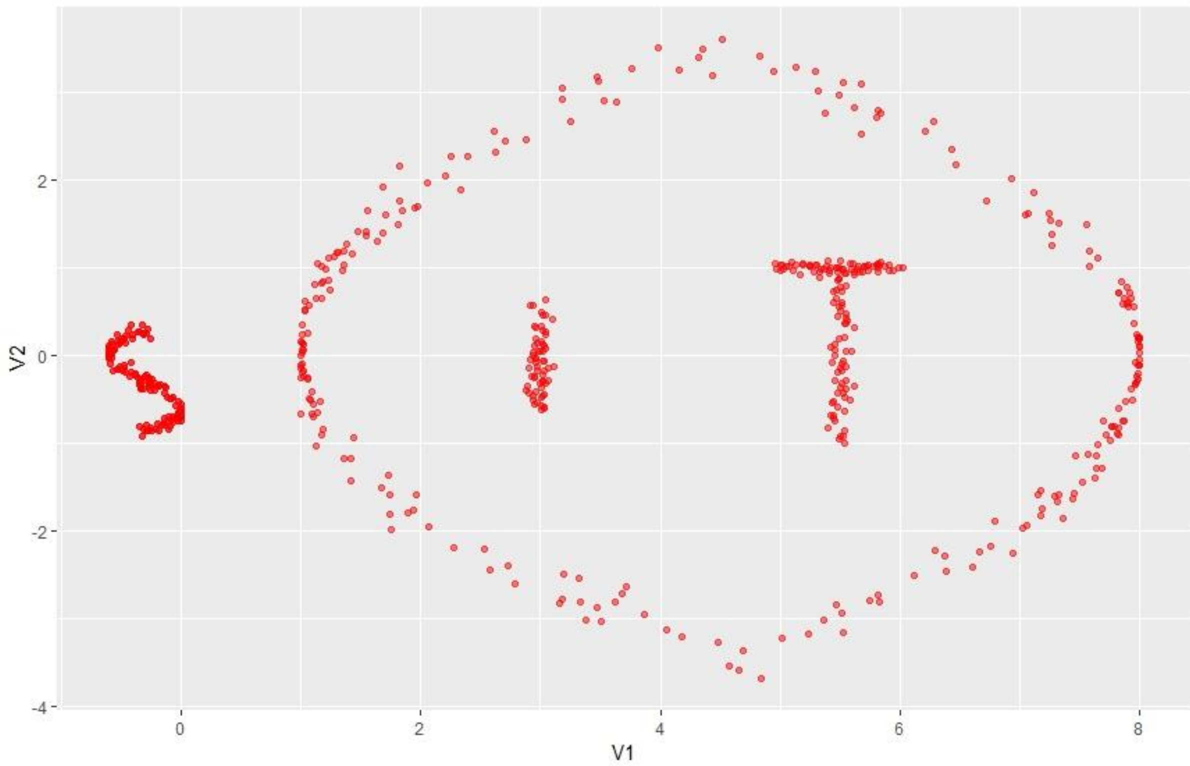


I find that there is a good agreement between the observed proximities and the inter-point distances. It is better Shepard plot for K=2.

Q7)

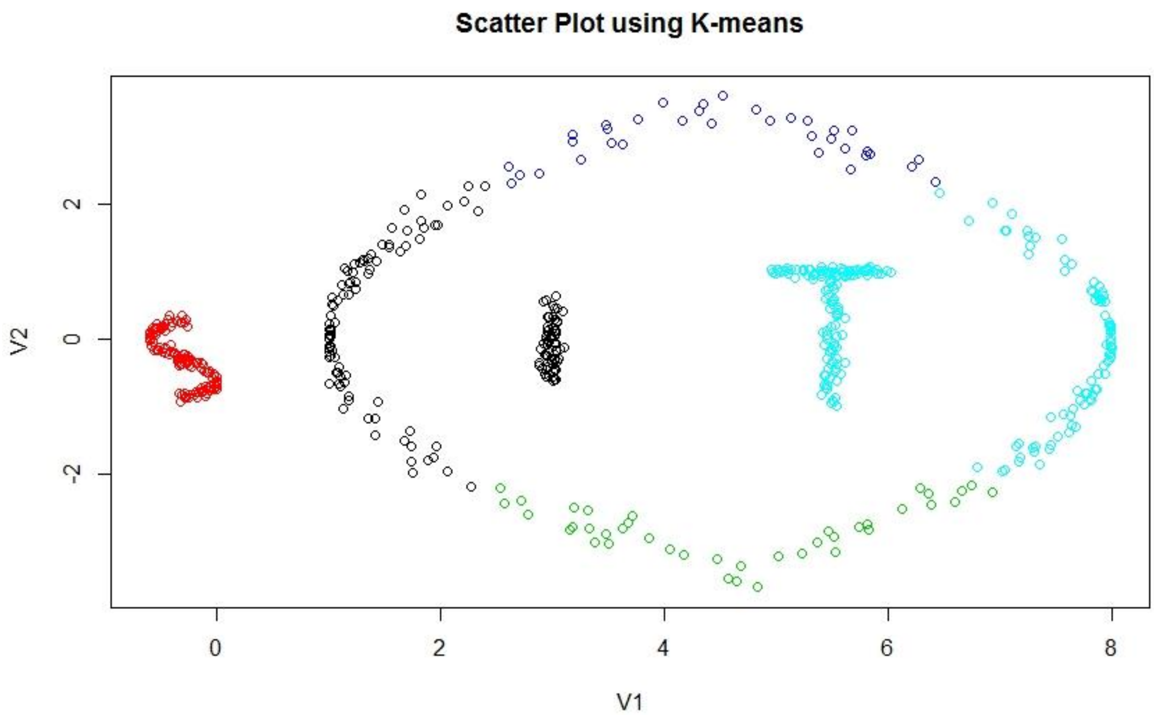
7.1) K-Means Clustering

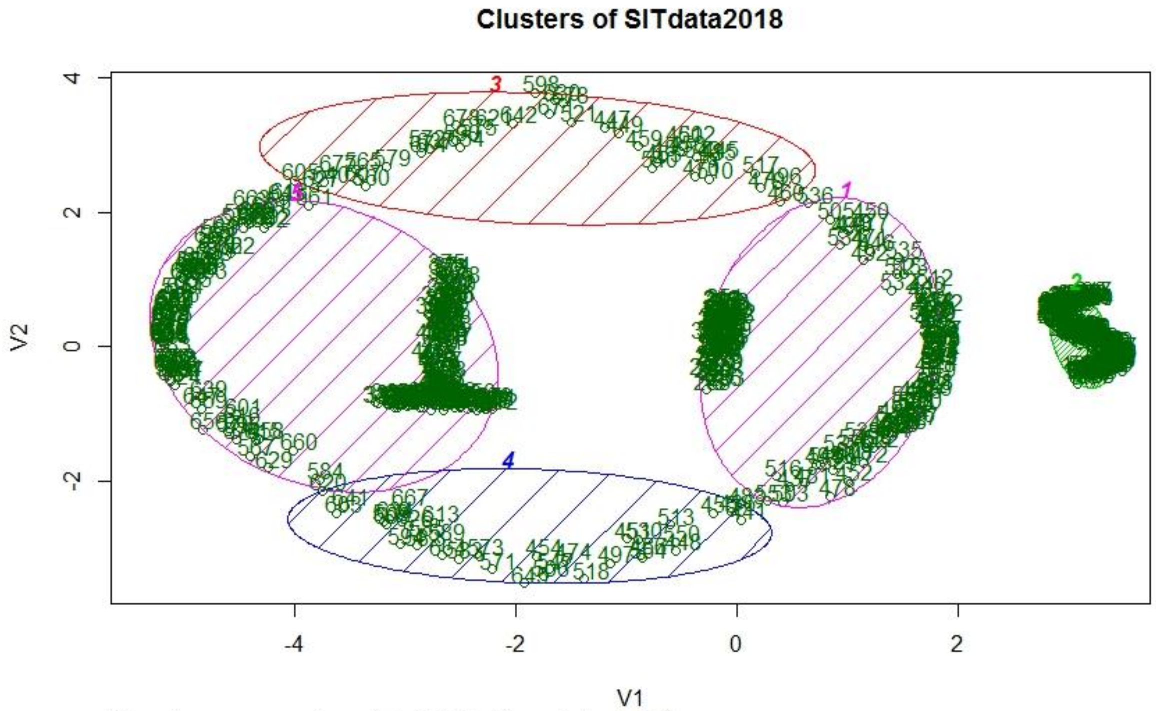
a)



b) Number of Classes=5

c)





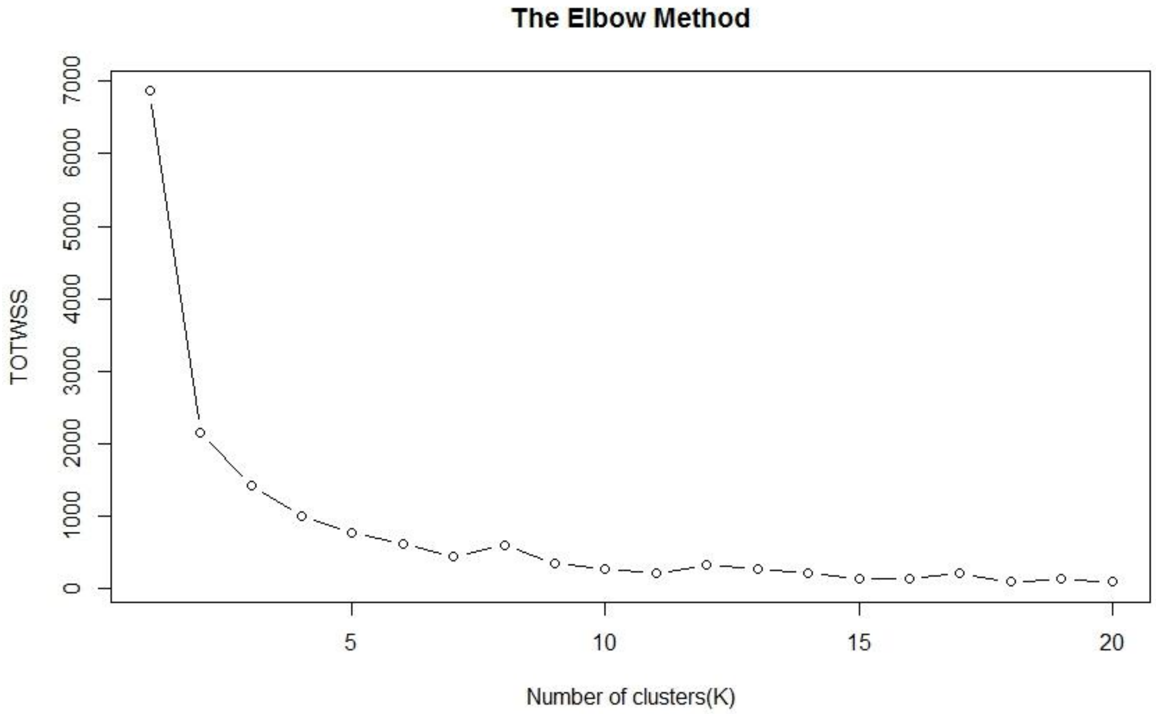
These two components explain 100 % of the point variability.

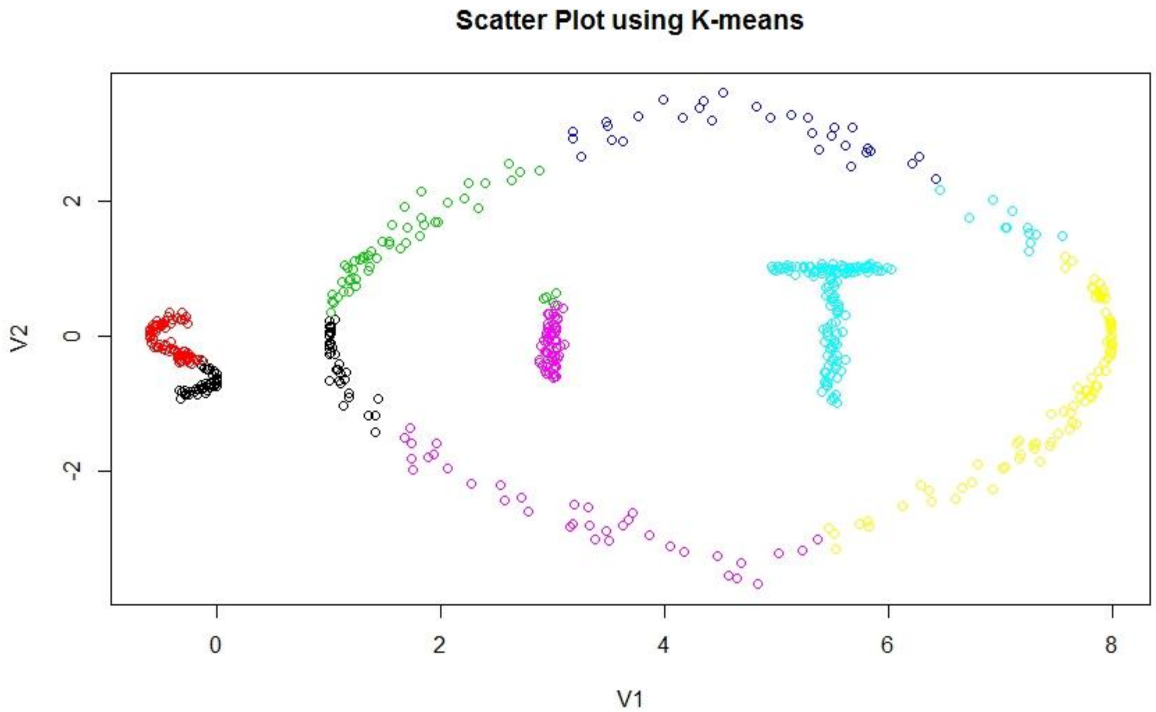
```
K-means clustering with 5 clusters of sizes 150, 248, 40, 35, 210

Cluster means:
        V1         V2
1  2.045373  0.1175819
2 -0.300000 -0.3047486
3  4.628920 -2.8088570
4  4.552600  2.9513163
5  6.363786  0.2351205

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [59] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[117] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[175] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[233] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[291] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[349] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[407] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 1 1 1 1 4 1 1 1 3 1 3 4 3 1 4 1 1 4 4 1 1 3 3 3 3 1 1 1
[465] 3 4 1 1 1 3 1 1 1 4 1 1 1 1 3 1 1 1 4 1 3 1 1 1 3 1 1 1 1 1 1 3 4 1 1 1 1 3 1 1 1 4 1 1 1 3 1 1 4 1 1 1 3 4 1 1 3 1
[523] 1 1 1 1 1 1 1 4 1 1 1 1 1 1 1 1 1 3 1 1 3 1 1 1 4 1 1 4 1 1 1 4 1 1 1 1 1 3 5 5 5 5 3 5 5 5 4 5 4 3 4 5 3 5 5 3 3 5
[581] 5 4 4 5 4 5 5 5 4 3 5 5 5 4 5 5 5 3 5 5 5 5 4 5 3 5 3 5 4 5 5 5 4 5 5 5 5 5 5 4 3 5 5 5 5 4 3 5 5 3 5 5 5 4 5 5 3 5
```

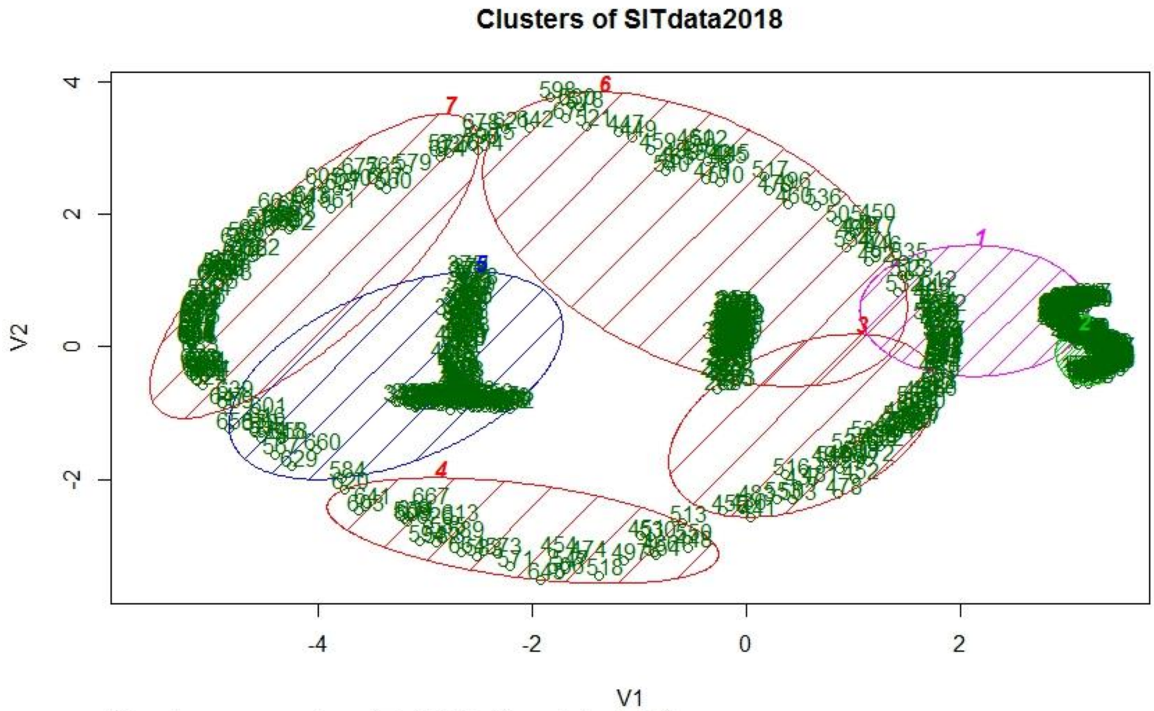
```
Within cluster sum of squares by cluster:
[1] 234.72319  43.11805  76.13197  50.65142 426.61966
 (between_SS / total_SS =  87.9 %)
```

The letter T and left curved portion is considered as one cluster. Same thing is happen for the right round curve and Letter I. The k-means clustering is not able to distinguish as two different labels.

d)


The Elbow Method
TOTWSS
0
1000
2000
3000
4000
5000
6000
7000
5
10
15
20
Number of clusters(K)

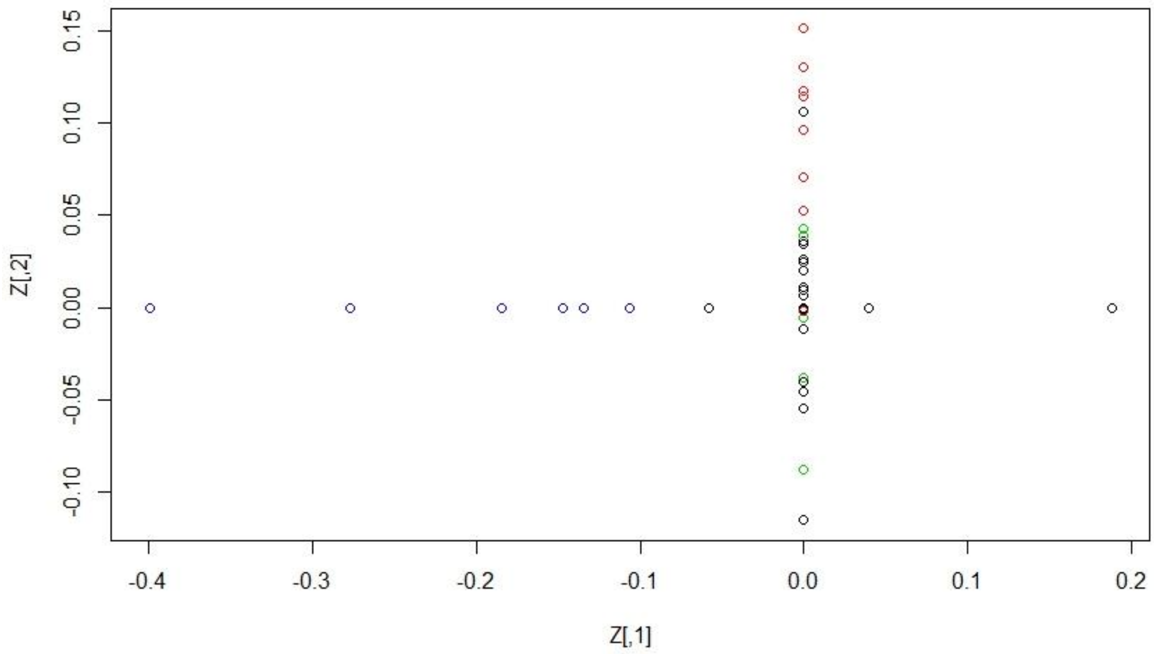

Scatter Plot using K-means
V2
-2
0
2
0
2
4
6
8
V1

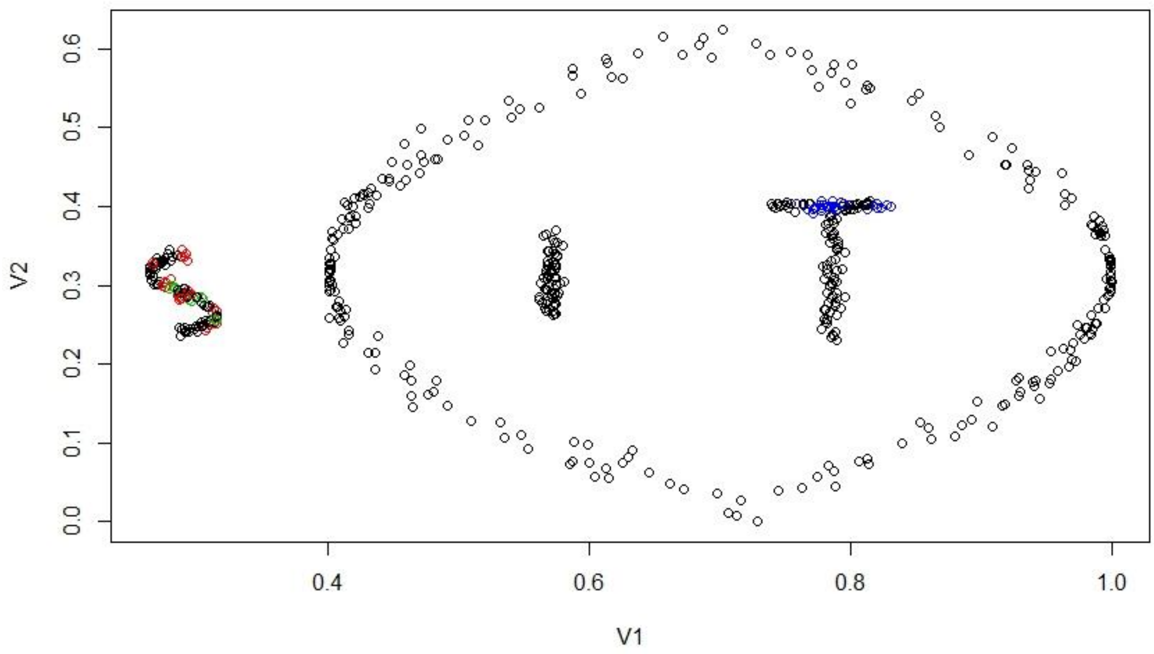


These two components explain 100 % of the point variability.

From different colours we can get to know how many different clusters we needed to analyse all the shapes in our data.

7.2) Spectral Clustering





From spectral clustering we are able to find all the patterns in the SIT dataset. There is some wrong prediction over S and T but overall it identifies all the shapes/curves in our data.

Q3) $P(\text{Smokers}) = 0.20$

$P(\text{Non-Smoker}) = 1 - P(\text{Smoker}) = 0.80$

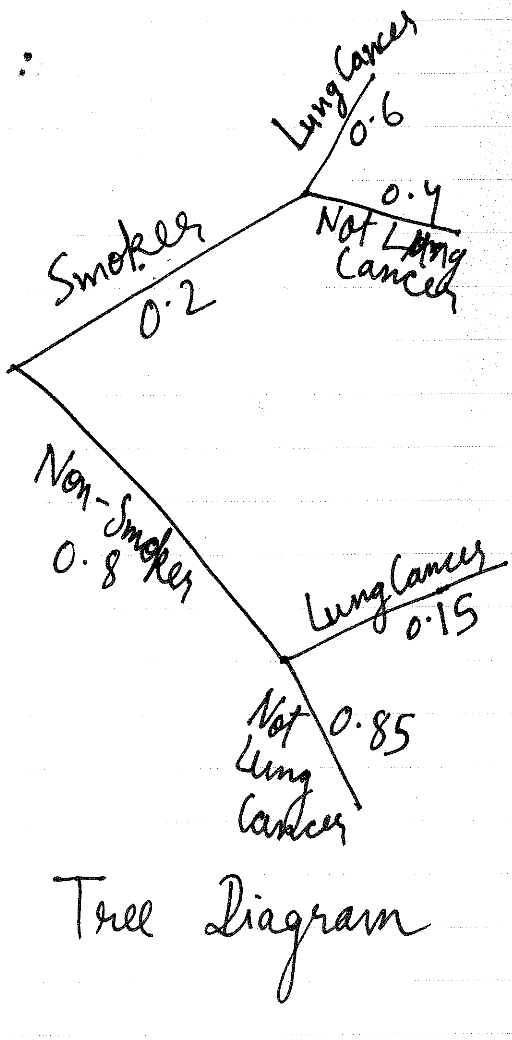
$P(\text{Smokers and Lung Cancer}) = 0.20 * 0.60 = 0.12$

$P(\text{Non-Smoker and Lung Cancer}) = 0.80 * 0.15 = 0.120$

P(Lung Cancer was a Smoker) =

$$\frac{P(\text{Smoker and Lung Cancer})}{P(\text{Smoker and Lung Cancer}) + P(\text{Non-Smoker and Lung Cancer})}$$

$$= \frac{0.120}{0.120 + 0.120} = \frac{0.120}{0.240} = 0.50$$



Tree Diagram

Ques 4 -

$$x_i \sim \text{Poid}(\theta)$$
$$\text{Poid}(\theta) = p(x_i|\theta) = \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

x_i are iid

(a) Expression for likelihood function

$$p(X|\theta) = \frac{\prod_{i=1}^{N} \theta^{x_i} e^{-\theta}}{x_i!}$$

$$= \frac{\theta^{\sum_{i=1}^{N} x_i} e^{-\sum_{i=1}^{N}\theta}}{\sum_{i=1}^{N} x_i!}$$

$$= \frac{\theta^{\sum_{i=1}^{N} x_i} e^{-N\theta}}{x_1! \, x_2! \, x_3! \ldots\ldots x_N!}$$

$$= \frac{\theta^{N\bar{x}} e^{-N\theta}}{x_1! \, x_2! \, x_3! \ldots\ldots x_N!} \quad \text{where } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

(b) Expression for log likelihood

$$L(\theta) = \ln(p(X|\theta))$$

$$= \ln\left(\frac{\theta^{N\bar{x}} e^{-N\theta}}{x_1! \, x_2! \, x_3! \ldots\ldots x_N!}\right)$$

$$= (N\bar{x}\ln\theta - N\theta) - (\ln(x_1) + \ln(x_2) + \ln(x_3) + \ldots\ldots + \ln(x_N))$$

$$= N\bar{x}\ln\theta - N\theta - \sum_{i=1}^{N} \ln(x_i!)$$

$$= -N\theta - \sum_{i=1}^{N} \ln(x_i!) + N\bar{x}\ln(\theta)$$

(c) Maximum Likelihood Estimation

Differentiating $L(\theta)$ w.r.t θ

$$\frac{dL(\theta)}{d\theta} = -N + \frac{N\bar{x}}{\theta}$$

$$\therefore \frac{dL(\theta)}{d(\theta)} = 0$$

$$-N + \frac{N\bar{x}}{\theta} = 0$$

$$\frac{\cancel{N}\bar{x}}{\theta} = \cancel{N}$$

$$\bar{x} = \theta$$

$$\Rightarrow \hat{\theta} = \bar{x} \quad \text{where } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

(d) $\hat{\theta} = \bar{x}$

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$x_1 = 100$, $x_2 = 60$, $x_3 = 70$, $N = 300$

$$\hat{\theta} = \frac{1}{3}(100 + 60 + 70)$$

$$\hat{\theta} = \frac{1}{3}(230) = 76.67$$

MLE for given data is 76.67

5) Prior $\sim N(800, 100^2)$ $\bar{x} = 1100$

Likelihood $\sim N(\theta, 200^2)$

Posterior $\sim N(\mu_N, \sigma^2_N)$

μ_N is mean of posterior

σ^2_N is variance of posterior

N is Number of observations.

a) $$\mu_N = \sigma^2_N\left(\frac{n\bar{x}}{200^2} + \frac{800}{100^2}\right)$$ Posterior $\propto$ likelihood X Prior.

$$\frac{1}{\sigma^2_N} = \frac{n}{200^2} + \frac{1}{100^2}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

b) $n=3$

$$\frac{1}{\sigma^2_N} = \frac{n}{200^2} + \frac{1}{100^2}$$

$$\frac{1}{\sigma^2_N} = \frac{3}{200^2} + \frac{1}{100^2}$$

$$\sigma^2_N \approx 5714.29$$

$$\sigma_N = 75.59$$

$$\mu_N = 5714.29\left(\frac{3300}{200^2} + \frac{800}{100^2}\right)$$

$$= 928.57$$

Mean $= 928.57$

Standard Deviation $= 75.59$

c) $n = 15$

$$\frac{1}{\sigma^2_N} = \frac{15}{200^2} + \frac{1}{100^2}$$

$$\sigma^2_N = 2105.26$$

$$\sigma_N = 45.88$$

$$\mu_N = 2105.26\left(\frac{16500}{200^2} + \frac{1}{100^2}\right)$$

$$= 868.63$$

Mean $= 868.63$

Standard deviation $= 45.88$