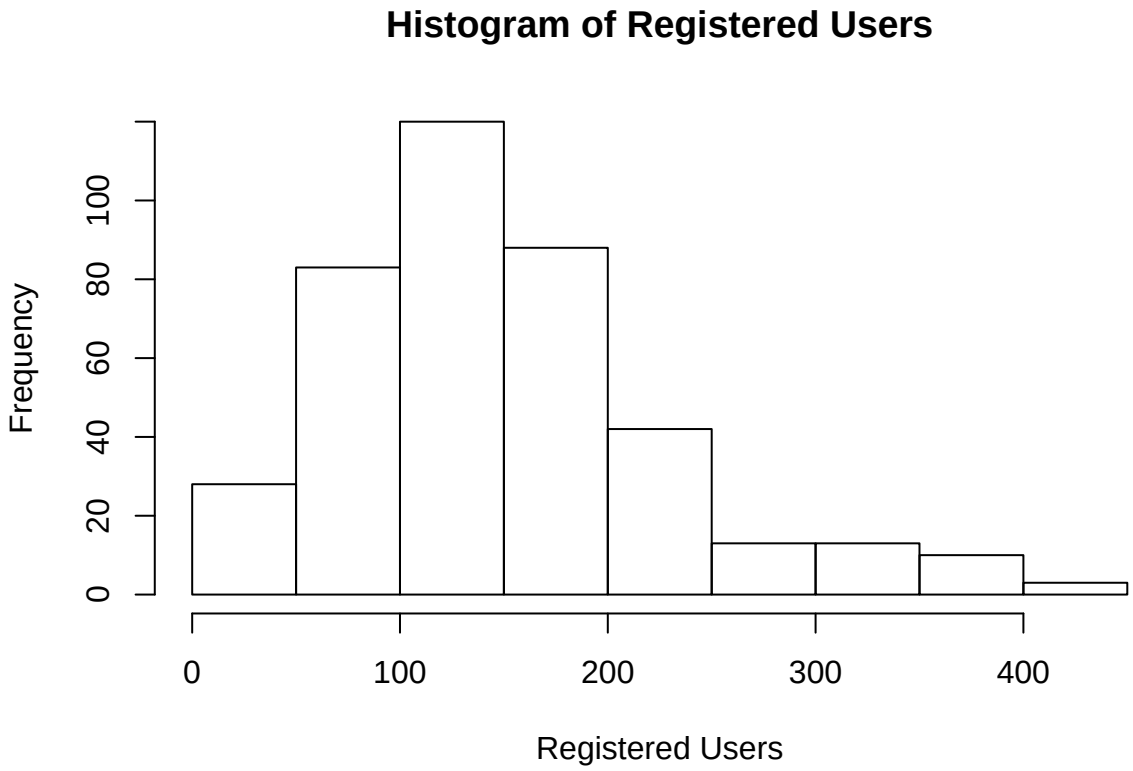
SIT 743: Assignment 1

Student Name: Dhananjay Pandya | Student Number: 218202943

Question 1)

1.1)

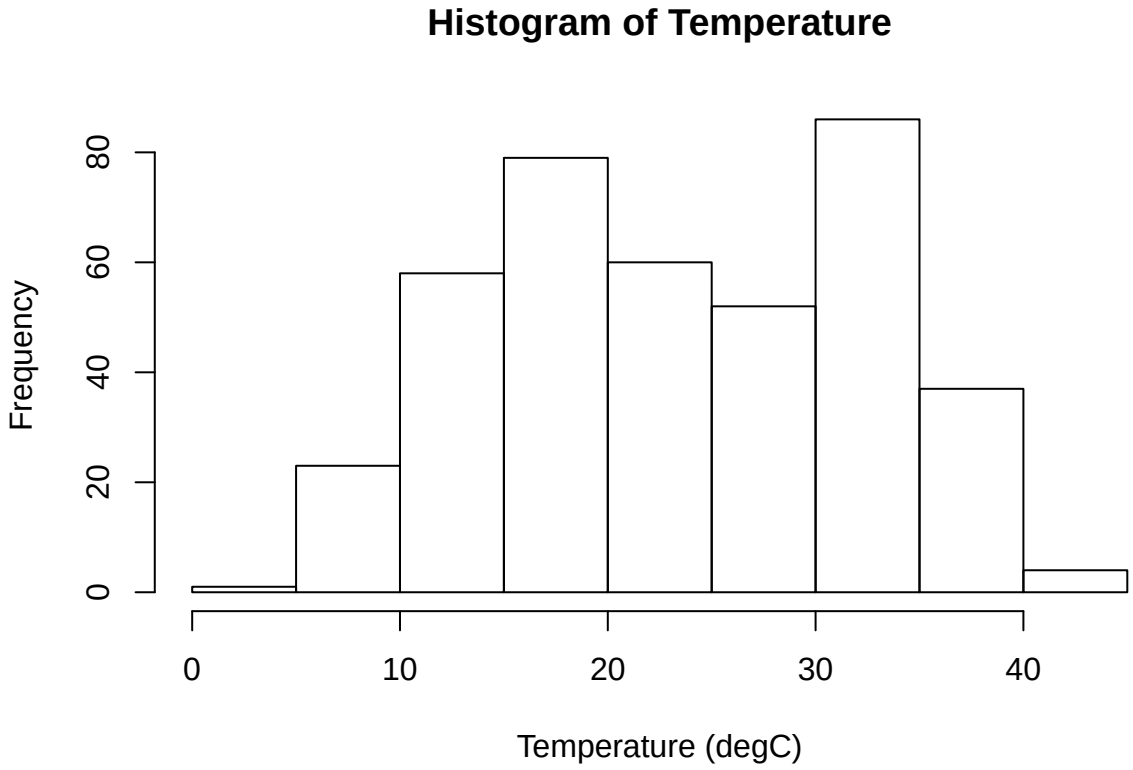


Comments:

1. Shape: This histogram is skewed to the right and only has a single peak. This means that the mean is not a good indicator of the center of the data, since it equally weights ALL data. However, ignoring the outliers in the bins between 250 and 450 registered users, the data is approximately normally distributed around 125 users.

2. Span: The number of registered users is fairly well spread out. The most number of registered users occurs in the bin of 100 to 150 users. The span of this data is affected by the outliers that occur in the bins between 250 and 450 users.

3. Outliers: As mentioned previously, there are some unusually high registered user frequencies occuring between 250 and 450 registered users.



Comments:

1. Shape: This histogram has twin peaks and is distributed relatively evenly when compared to the registered users histogram. This makes sense because weather in winter is cool (centered around a low temperature) and weather in summer is hot (centered around a high temperature). Days of extreme temperature are uncommon although not unheard of and thus there are some low and high outliers in bins 0-5 degC and 40-45 degC.

2. Span: This histogram spans from ~2.5 degC to ~42.5 degC (based on bins). There are some days of extreme temperatures that affect this and most days, expected temperatures are likely between 12.5 and 32.5 degC depending on the season.

3. Outliers: As mentioned previously, temperatures are seasonal and thereby cyclical. In winter, there are odd days where the temperature is extremely low leading to low outliers in bins from 0 to 10 degC. Similarly, in summer there are odd days where temperatures are very high leading to high outliers in bins from 35 to 45 degC.

1.2)

Casual Users

Measure	Value
Minimum	1.00
1st Quartile	20.00
Median	43.00
3rd Quartile	73.25
Maximum	268.00

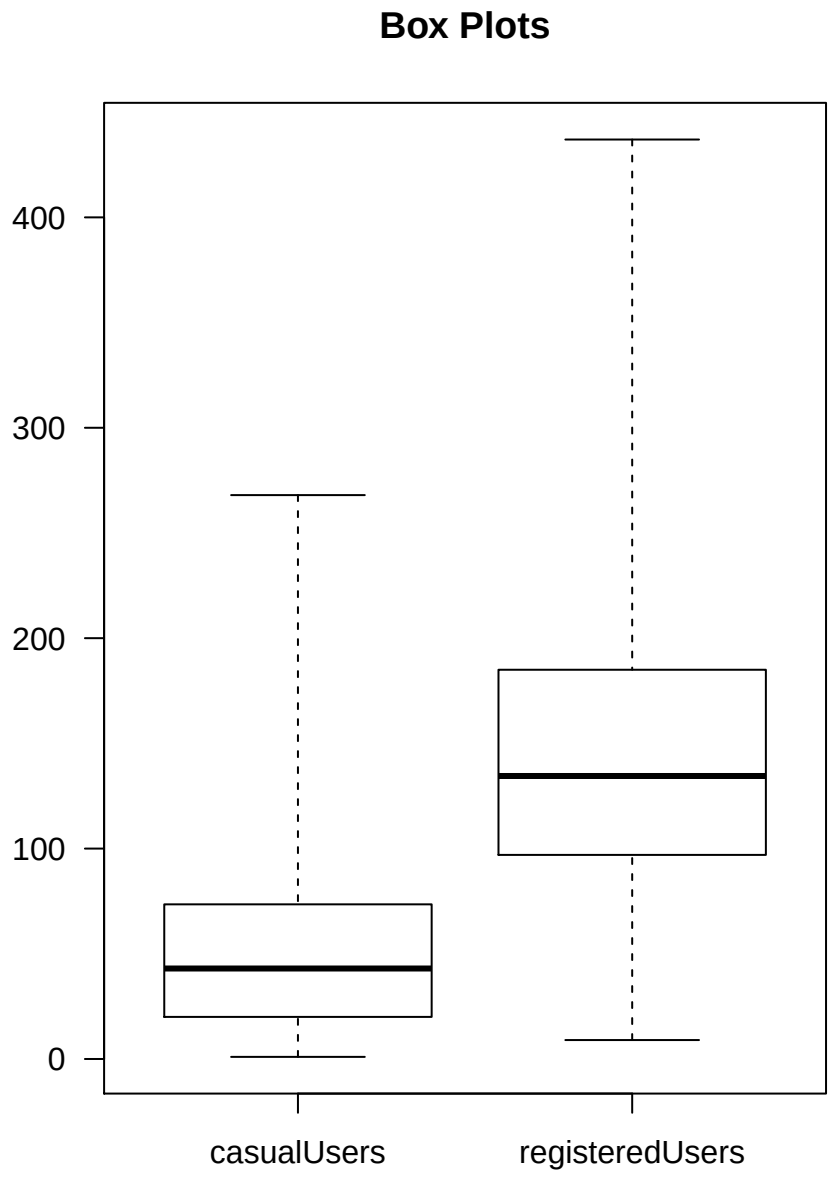
Mean: 59.70

Registered Users

Measure	Value
Minimum	9.0
1st Quartile	97.0
Median	134.5
3rd Quartile	185.0
Maximum	437.0

Mean: 148.9

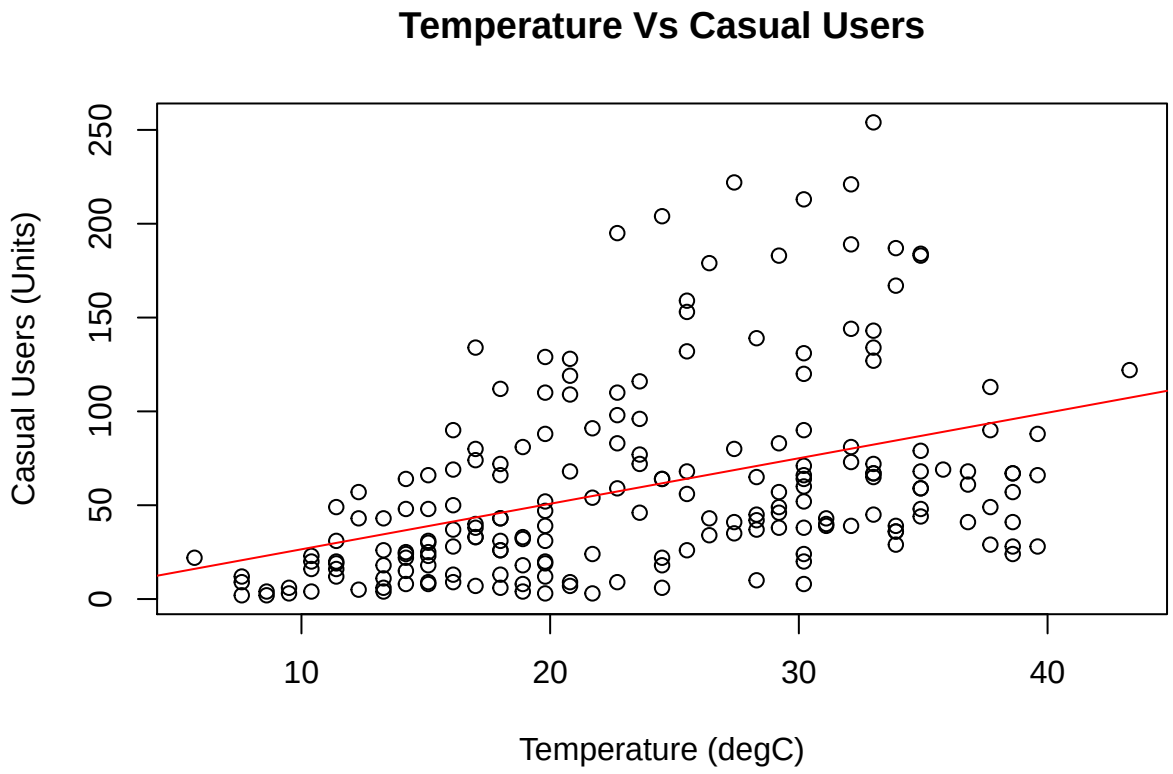
1.3)



Comments:

- The Boxplots for casual and registered users are skewed to the right, although the casual users' skew is more prominent. This indicates that the median is a better measure of central tendency compared to the mean.
- The range of users for both groups is very different. Casual users have a range of 267, which is much smaller than the range of registered users (428). Since the minimum values for both groups are similar (Casual Users: 1; Registered Users: 9), the range indicates that the number of registered users varies considerably more than the number of casual users.
- The interquartile range for casual users is located between 20 and 73.25 users. The interquartile range for registered users is located between 97 and 185 users. This indicates that on any given day, there are likely to be more registered users than casual users. The IQR also confirms that the number of registered users varies considerably more than the number of casual users.
- The median for casual users (43) is much lower than the median for registered users (134.5). This means that on any given day, the number of registered users is likely to be greater than the number of casual users.

1.4)



Comments:

- Overall Trend:
  - The number of casual users increases as the temperature increases.
  - There are three distinct regions of interest: 0 to 20 degC, 20 to 30 degC, and 30 to 45 degC. These are described in more detail below.
- 0 to 20 degC:
  - The number of casual users increases within this group as the temperature gets closer to 20 degC.

  - This group represents the region where the number of riders is at its lowest. This is likely explained by the low temperatures, which are not conducive to bike riding.
- 20 to 30 degC:
  - The number of casual users increases within this group as the temperature gets closer to 30 degC.
  - The number of casual users is at its largest in this group. This is likely because temperatures are optimal for bike riding.
- 30 to 45 degC:
  - The number of casual users decreases within this group as temperatures get closer to 45 degC.
  - The number of casual users is lower compared to '20 to 30 degC' group because conditions are too hot for bike riding.
  - The number of casual users is generally higher compared to '0 to 20 degC' group because warmer conditions are more conducive to bike riding than cooler conditions.

1.5)

For plotted regression line, please see scatter plot in question 1.4.

Linear regression equation: $y = 2.426x + 2.242$

Correlation Coefficient: 0.396

Coefficient of Determination: 0.157

Explanation:

- The gradient of the regression line is 2.426, which indicates that a 1 degree increase in temperature leads to 2.426 more casual bike users.
- The intercept of the regression line is 2.242, which indicates that when temperatures are at 0 degC there would be 2.242 casual bike users on any given day. As a general rule however, it is dangerous to make predictions based on results of a linear regression outside the range of x values. In this case the temperatures range from 2 to 43.3 degC, hence the intercept value must be treated with appropriate caution.
- The correlation coefficient is positive which indicates that there is positive correlation between temperature and casual users. This means that when temperature increases, so does the number of casual users.
- The coefficient of determination is 0.157 which indicates that the regression line does not explain a lot of variation in the dataset. This is common when modelling human behaviour because human behaviour is much harder to embed into a mathematical equation compared to physical phenomena. The reason that the line does not explain all the variation is because of the density of data between 0 and 100 users over the range of temperature. Anything above this range of users is not modelled well by linear regression because the regression coefficients are calculated using the sum of least squares. This method tries to minimise the error between the data points and the regression line. Overall, the data density between 0 and 100 users "pulls" the line towards it and the data above 100 users is ignored. When calculating the coefficient of determination the total variation in the dataset is compared to the variation explained by the regression line. In this case, the line does not explain the high number of casual users in the 20 to 30 degC range and thus the value of the coefficent is low.

Question 2)

2.1)

P(Victoria): 33.25%

2.2)

P(Light Commercial Vehicle): 17.25%

2.3)

P(P AND N): 34%

2.4)

P(C|Q): 22.86%

2.5)

P(Q|P): 24.47%

2.6)

P(V OR P): 87.5%

2.7)

Vehicle Type	Marginal
P	0.8275
C	0.1725

2.8)

State	Marginal
N	0.405
V	0.3325
Q	0.2625

2.9)

Headers	N	V	Q
P	0.8395	0.8571	0.7714
C	0.1605	0.1429	0.2286

Question 3)

3.1)

P(Smoker|Lung Condition) = $\frac{P(LungConditon|Smoker) \times P(LungCondition)}{P(Smokers)}$

P(Smoker|Lung Condition) = $\frac{0.6 \times 0.15}{0.2}$

P(Smoker|Lung Condition) = 0.45 (OR 45%)

Question 4)

a)

Since the N days are iid, the likelihood can be written as

$$p(X|\theta) = p(x_1|\theta) \times p(x_2|\theta) \times .. \times p(x_N|\theta)$$

Now, the expression for number of cars arriving on day 1 is:

$$p(x_1|\theta) = \frac{\theta^{x_1} \times e^{-\theta}}{x_1!}$$

Similarly for day 2:

$$p(x_2|\theta) = \frac{\theta^{x_2} \times e^{-\theta}}{x_2!}$$

And for day N:

$$p(x_N|\theta) = \frac{\theta^{x_N} \times e^{-\theta}}{x_N!}$$

Substituting this into our earlier equation for $p(X|\theta)$

$$p(X|\theta) = \frac{\theta^{x_1} \times e^{-\theta}}{x_1!} \times \frac{\theta^{x_2} \times e^{-\theta}}{x_2!} \times .. \times \frac{\theta^{x_N} \times e^{-\theta}}{x_N!}$$

Combining the products on the RHS, we get

$$p(X|\theta) = \frac{\theta^{x_1+x_2+..+x_N} \times e^{-\theta-\theta-..-\theta}}{x_1! \times x_2! \times .. \times x_N!}$$

We can simplify this equation by using

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

Giving the desired equation

$$p(X|\theta) = \frac{\theta^{N\bar{x}} \times e^{-N\theta}}{x_1! \times x_2! \times .. \times x_N!}$$

b)

$$
\begin{aligned}
L(\theta) &= ln(p(X|\theta) && (1)\\
L(\theta) &= ln(\frac{\theta^{N\bar{x}} \times e^{-N\theta}}{x_1! \times x_2! \times .. \times x_N!}) && (2)\\
L(\theta) &= ln(\theta^{N\bar{x}} \times e^{-N\theta}) - ln(x_1! \times x_2! \times .. \times x_N!) && (3)\\
L(\theta) &= ln(\theta^{N\bar{x}}) + ln(e^{-N\theta}) - ln(x_1! \times x_2! \times .. \times x_N!) && (4)\\
L(\theta) &= N\bar{x}ln(\theta) - N\theta - ln(x_1! \times x_2! \times .. \times x_N!) && (5)\\
& && (6)
\end{aligned}
$$

c)

First, we find the differentiate the log likelihood function $L(\theta)$

$$
\begin{aligned}
\frac{dL(\theta)}{d\theta} &= \frac{d}{d\theta}(N\bar{x}ln(\theta) - N\theta - ln(x_1! \times x_2! \times .. \times x_N!)) && (7)\\
&= \frac{N\bar{x}}{\theta} - N && (8)\\
& && (9)
\end{aligned}
$$

To maximise this, we set the differented function equal to 0

$$
\begin{aligned}
\frac{N\bar{x}}{\theta} - N &= 0 && (10)\\
\frac{N\bar{x}}{\theta} &= N && (11)\\
\theta &= \frac{N\bar{x}}{N} && (12)\\
\theta &= \bar{x} = \hat{\theta} && (13)\\
& && (14)
\end{aligned}
$$

d)

$$
\begin{aligned}
\hat{\theta} &= \bar{x} && (15)\\
&= \frac{1}{N}\sum_{i=1}^{N} x_i && (16)\\
&= \frac{1}{3}(100 + 60 + 70) && (17)\\
&= 76.77 && (18)\\
& && (19)
\end{aligned}
$$

Question 5)

5.1)

Bayes rule states that:

$$p(\mu|D) = \frac{p(D|\mu)p(\mu)}{p(D)}$$

Here, the posterior is: $p(\mu|D)$
The likelihood is: $p(D|\mu)$
The prior is: $p(\mu)$
The normalising constant is: $p(D)$

If the posterior distribution is in the same family as the prior probability distribution, then the prior and posterior are called **conjugate distributions**. The prior is called a **conjugate prior** for the likelihood function.

5.2)

Conjugate priors are useful in Bayesian statistics because they promote simplicity when computing the posterior. If you already know that the prior and posterior are from the same family of probability distributions, there is no need for numerical integration and the process is computationally efficient.

5.3)

Examples:
* Gaussian-Gaussian model
* Dirichlet-Multinomial model
* Beta-Beta model

5.4)

a)

The prior is a Gaussian with $P(\theta) \sim N(m, \tau^2)$
Likelihood is a Gaussian with $P(X|\theta) \sim N(\theta, \sigma^2)$
Therefore, the posterior is a Gaussian with $P(\theta|X) \sim N(\mu_n, \sigma_n^2)$
Where μ_n is the mean of the posterior and σ_n^2 is the variance of the posterior.

$$\mu_n = \sigma_n^2(\frac{n\bar{x}}{\sigma^2} + \frac{m}{\tau^2});\ \frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau^2}\ and,\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where, n is the number of observations

b)

For n=3,

$$\begin{aligned} \frac{1}{\sigma_n^2} &= \frac{n}{\sigma^2} + \frac{1}{\tau^2} && (20)\\ &= \frac{3}{200} + \frac{1}{100} && (21)\\ \sigma_n^2 &= 40 && (22)\\ & && (23) \end{aligned}$$

$$\begin{aligned} \mu_n &= \sigma_n^2(\frac{n\bar{x}}{\sigma^2} + \frac{m}{\tau^2}) && (24)\\ &= 40(\frac{3 \times 1100}{200} + \frac{800}{100}) && (25)\\ &= 980 && (26)\\ & && (27) \end{aligned}$$

Comment:
The posterior variance is less than the variance of the prior and the variance of the likelihood.

c)

$$\begin{aligned} \frac{1}{\sigma_n^2} &= \frac{n}{\sigma^2} + \frac{1}{\tau^2} && (28)\\ &= \frac{15}{200} + \frac{1}{100} && (29)\\ \sigma_n^2 &= 11.76mm && (30)\\ & && (31) \end{aligned}$$

$$\begin{aligned} \mu_n &= \sigma_n^2(\frac{n\bar{x}}{\sigma^2} + \frac{m}{\tau^2}) && (32)\\ &= 11.76(\frac{15 \times 1100}{200} + \frac{800}{100}) && (33)\\ &= 1064.71mm && (34)\\ & && (35) \end{aligned}$$

Comments:

- The variance of the posterior for n=15 is smaller than the variance for n=3.
- The mean of the posterior for n=15 is larger than the mean for n=3.
- As the number of known data points increases, the variance decreases because there is more certainty resulting from known data.
- Since there are 15 days of 1100mm average rainfall instead of 3 days, the posterior function has a higher mean for n=15 than n=3.

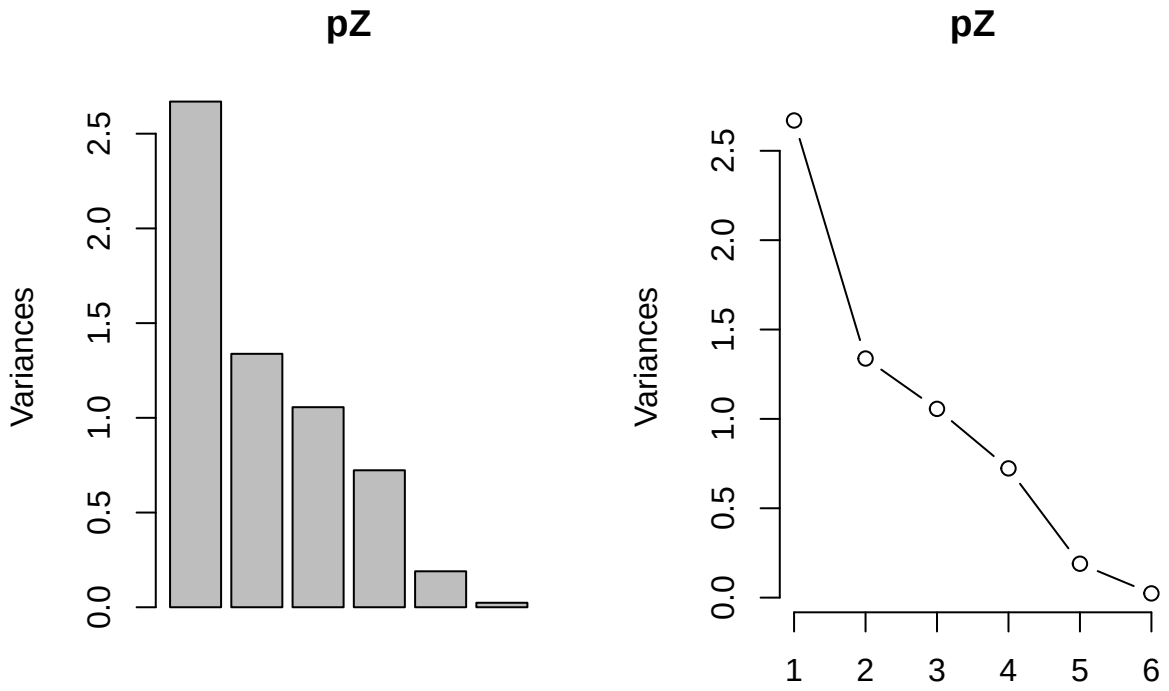
Question 6)

6.1)

Comments:

- Temperature and 'Feeling' Temperature are closely aligned with PC1 (horizontal axis). Casual users and registered users are also closely aligned with PC1.
- Humidity and windspeed are closely aligned with PC2 (vertical axis).
- Temperature and 'Feeling' Temperature are highly correlated because the angle between them is very small. The same holds true for casual users and registered users.

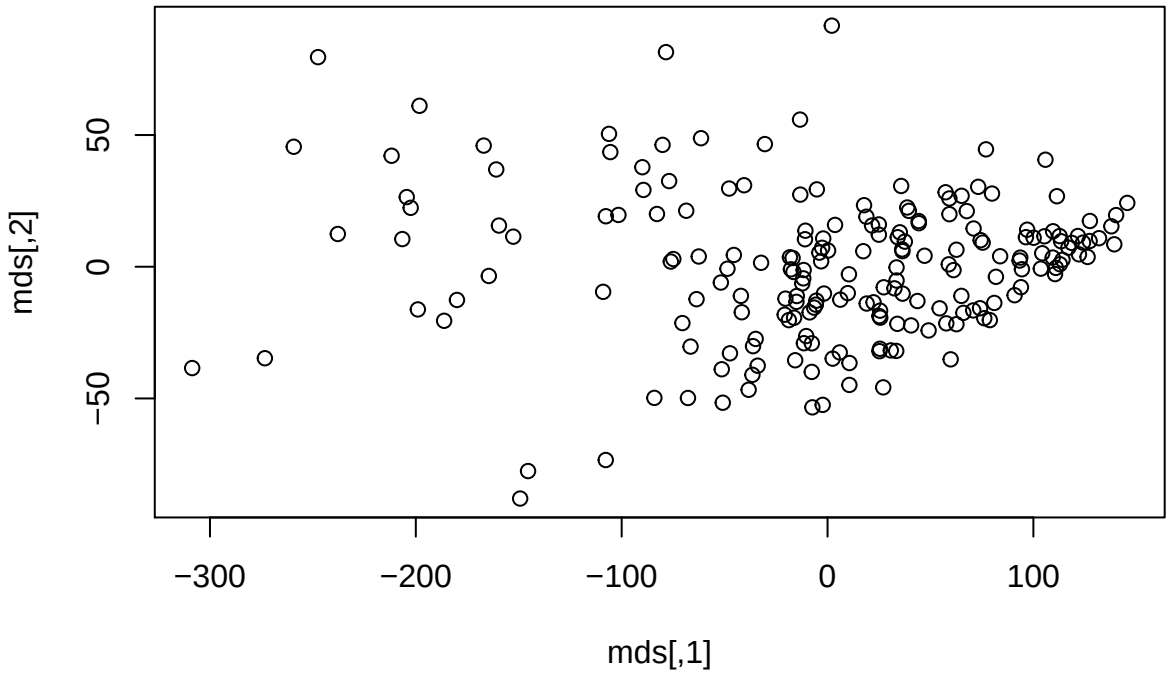
6.2)



Commments:

- These plots show the proportion of variance in the dataset that is explained by each principal component. As expected, the proportion of variance explained by each additonal component decreases. The aim of principal component analysis is to reduce the dimensionality of the data whilst retaining as much information as possible. The amount of information retained is captured by the variation preserved in the dataset.
- In this case, the first three principal components help explain ~85% of the variance in the dataset. I think that preserving 85% of variance while halving the dimensionality of this dataset is a good compromise.

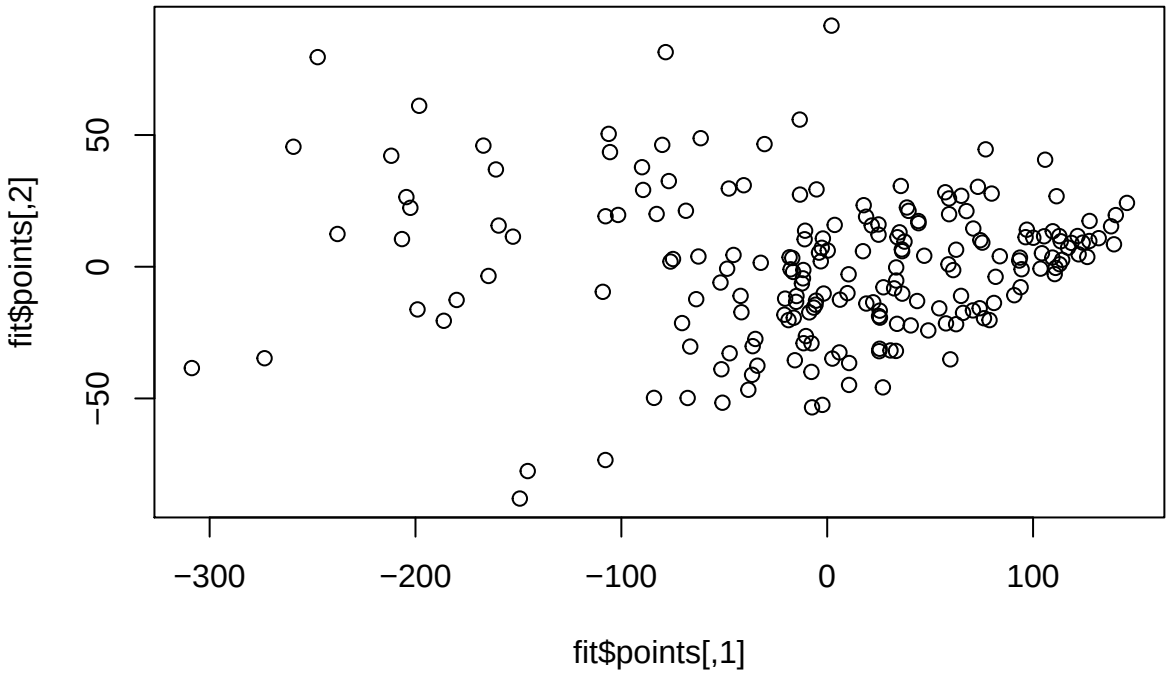
6.3)



Comments:

- In this case, classical MDS has reduced the dimensionality of this data by 4 dimensions.
- The cluster of data points centred around ~(50,0) on the plot above represents points that are very similar. As points get further from the center of the cluster, they become less similar to points near the center of the cluster. The same holds true for any two points on this plot. They closer they are, the more similar they are.

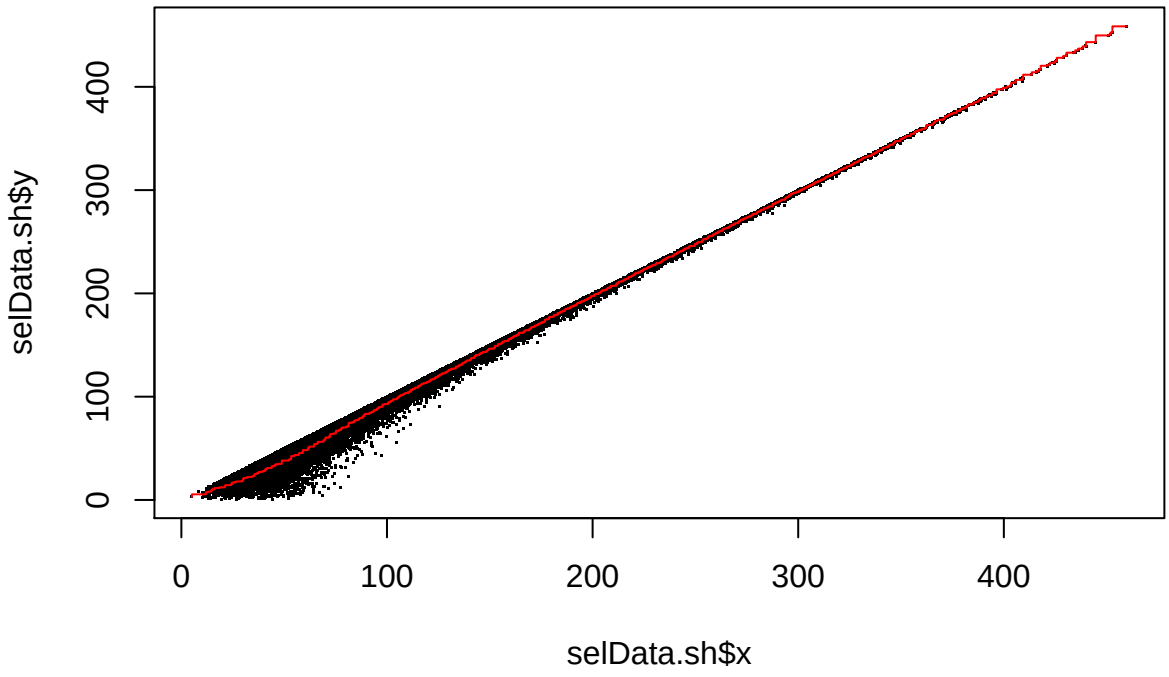
6.4)



Comments:

- The plot shows some clustering with approximately 3 clusters.

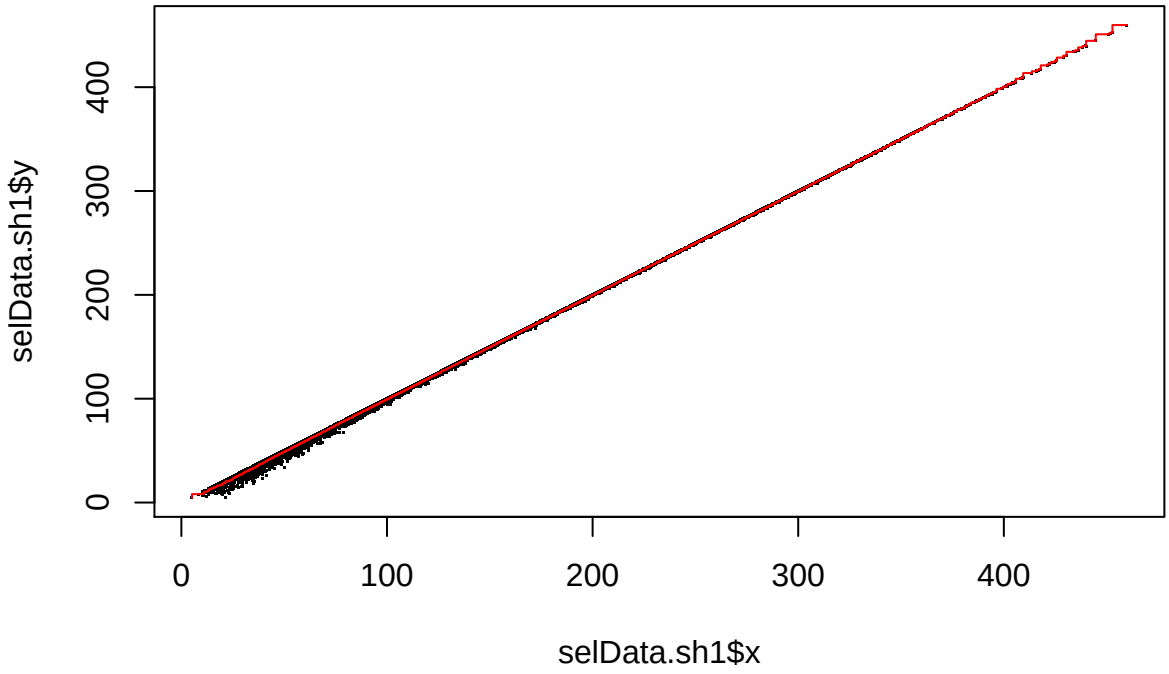
6.5)



Comments:

- In this case the fit of the data is reasonable wherever the points in black lie closer to the red line.
- At the lower end, there are some issues with fit because of the scatter of the original data around the regression. This suggests that the original dissimilarities are not preserved when the dimensionality of this data is reduced.

6.6)



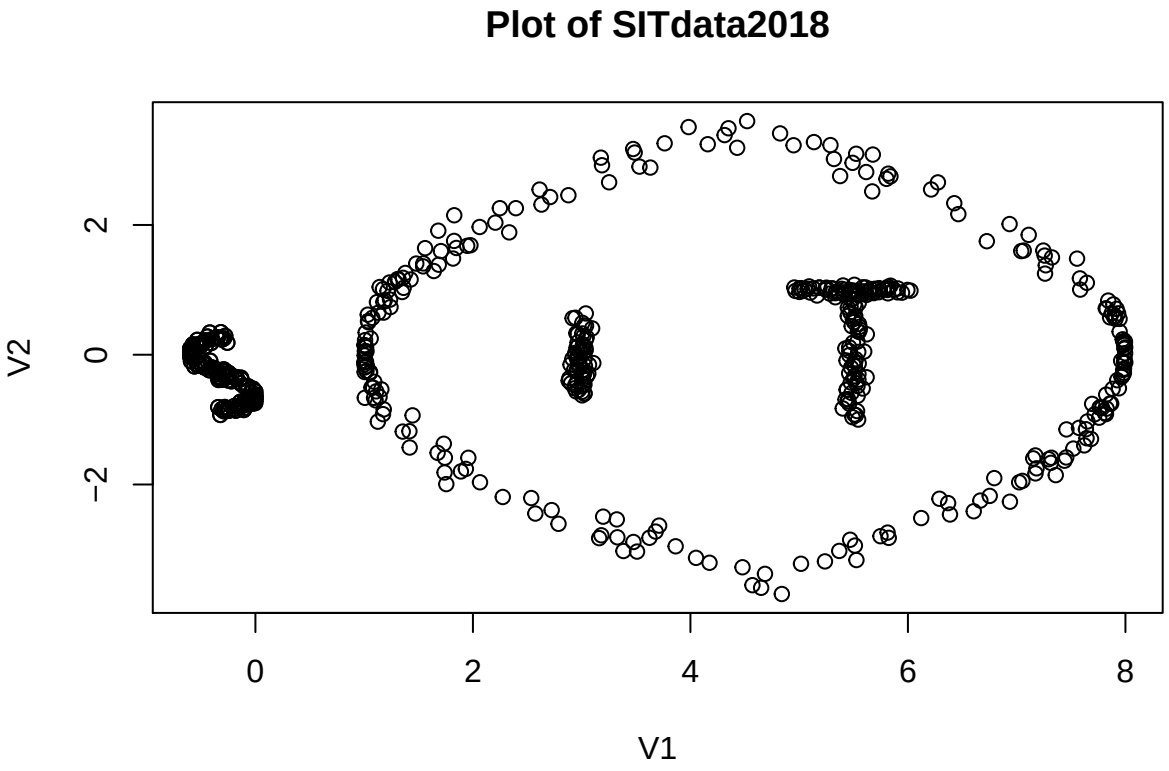
Comments:

- This data has a very good fit as evident by the lack of scatter around the regression line.
- Comparing this to the results for k=2, the fit for k=4 is much better. This is because more dimensions are retained using k=4 and thus more information is retained regarding the dissimilarities between the original dataset.

Question 7)

7.1)

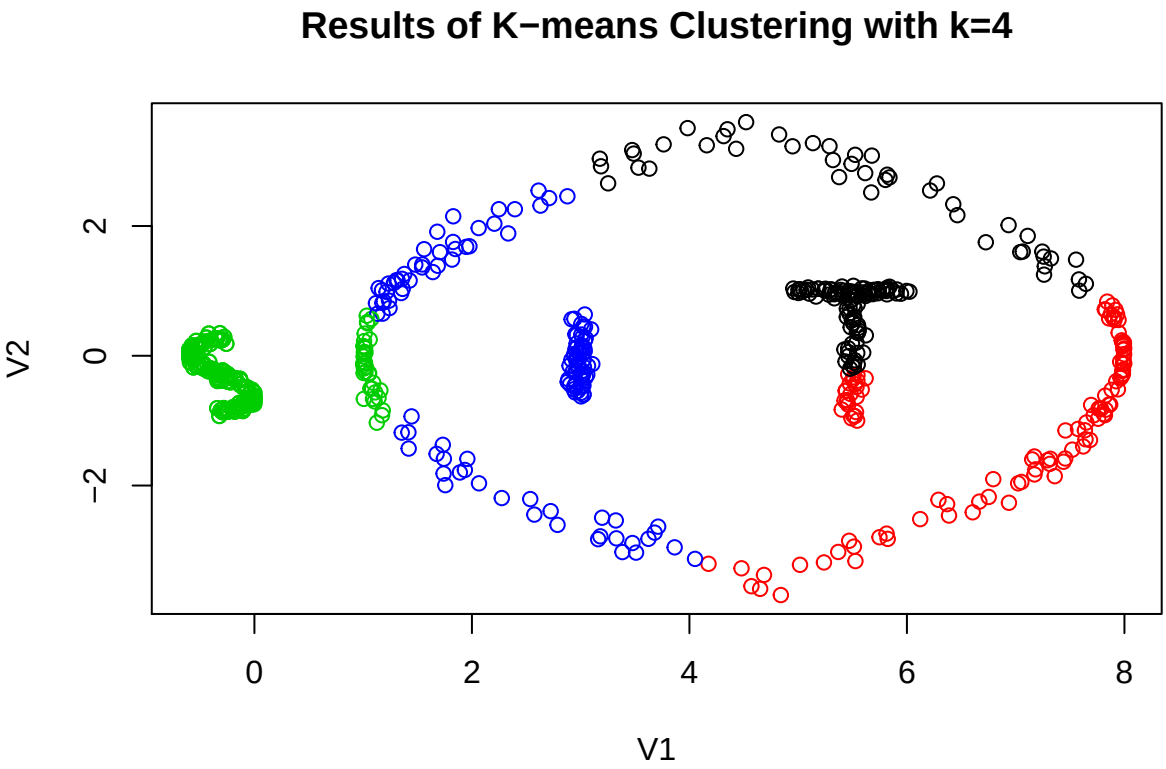
a)



b)

Number of classes/clusters = 4

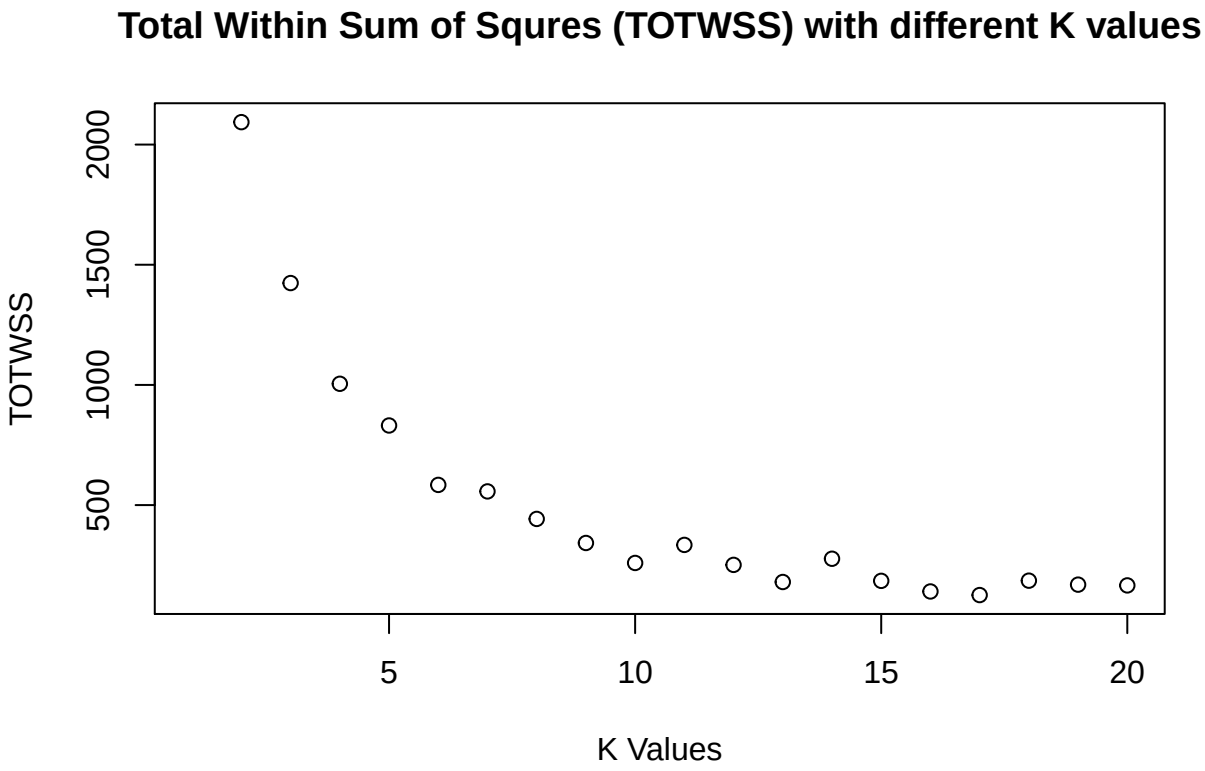
c)



Comments:

- The clustering generated by the K-means algorithm is very poor. The 'S', 'I', 'T' should form one cluster each, however this is clearly not the case. Similarly for the ellipsoid, this should be one cluster, however this is also not the case. This is because the algorithm has difficulty clustering data with non-convex shapes.

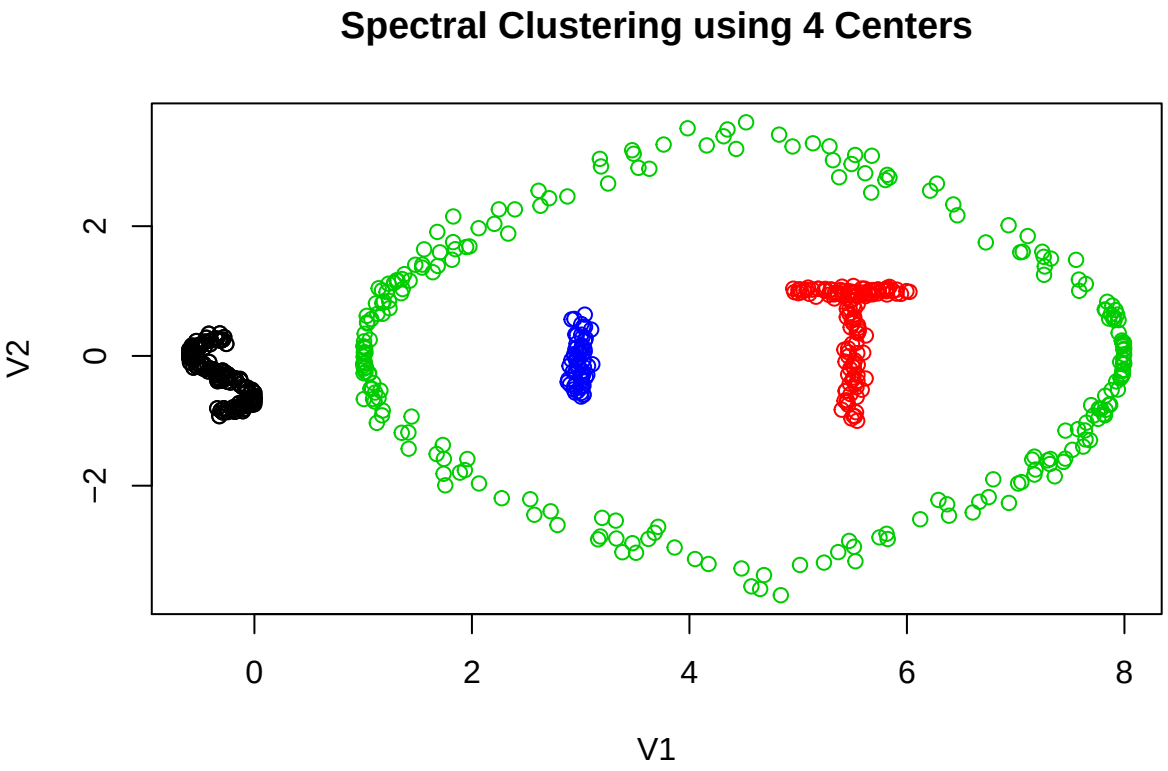
d)



Comments:

- This graph can be used to find the optimal number of clusters using the elbow method. In this method, we look to find an "elbow" on the plot of K Values against Total Within Sum of Squares. The elbow point is the point at which incremental increase in the number of clusters gives minimal reduction in the total within sum of squares. The total within sum of squares represents the closeness of each data point to the center of the cluster.
- Beyond K=6, the total within sum of squares diminishes too slowly. Hence, for K-means this is the suggested number of clusters.

7.2)



Comments:

- Whilst K-means assumes that clustering is always spherical, Spectral Clustering uses the Gaussian kernal to find the similarity matrix and then the four smallest eigenvectors of the Laplacian are used.
- In this case, the Spectral Clutering does a better job of grouping the data compared to the K-means.
- Spectral Clustering is perfect. The 'S', 'I', 'T' and the ellipsoid around 'IT' should be one cluster each. As evident from the plot above, this is clearly the case.