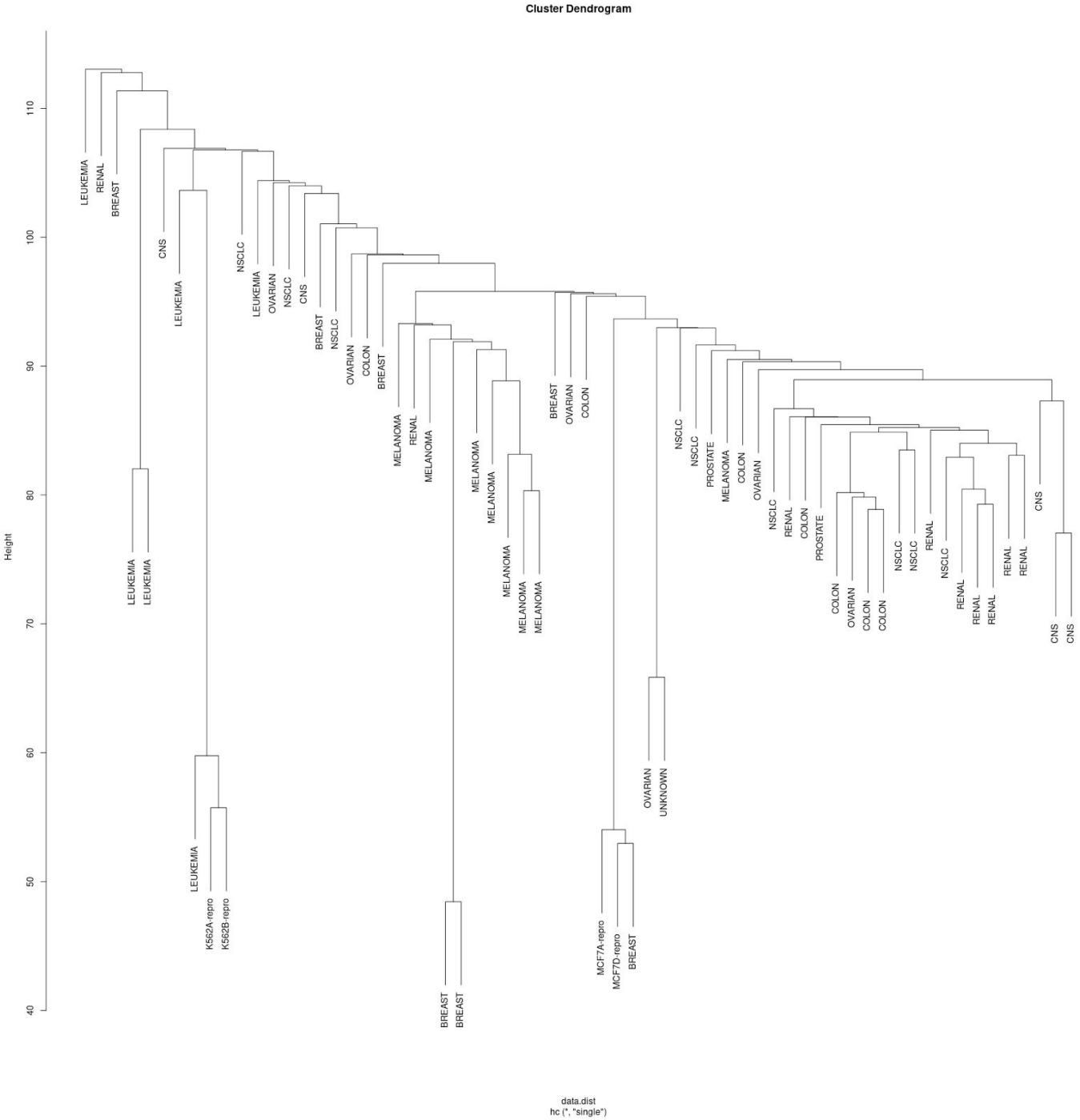
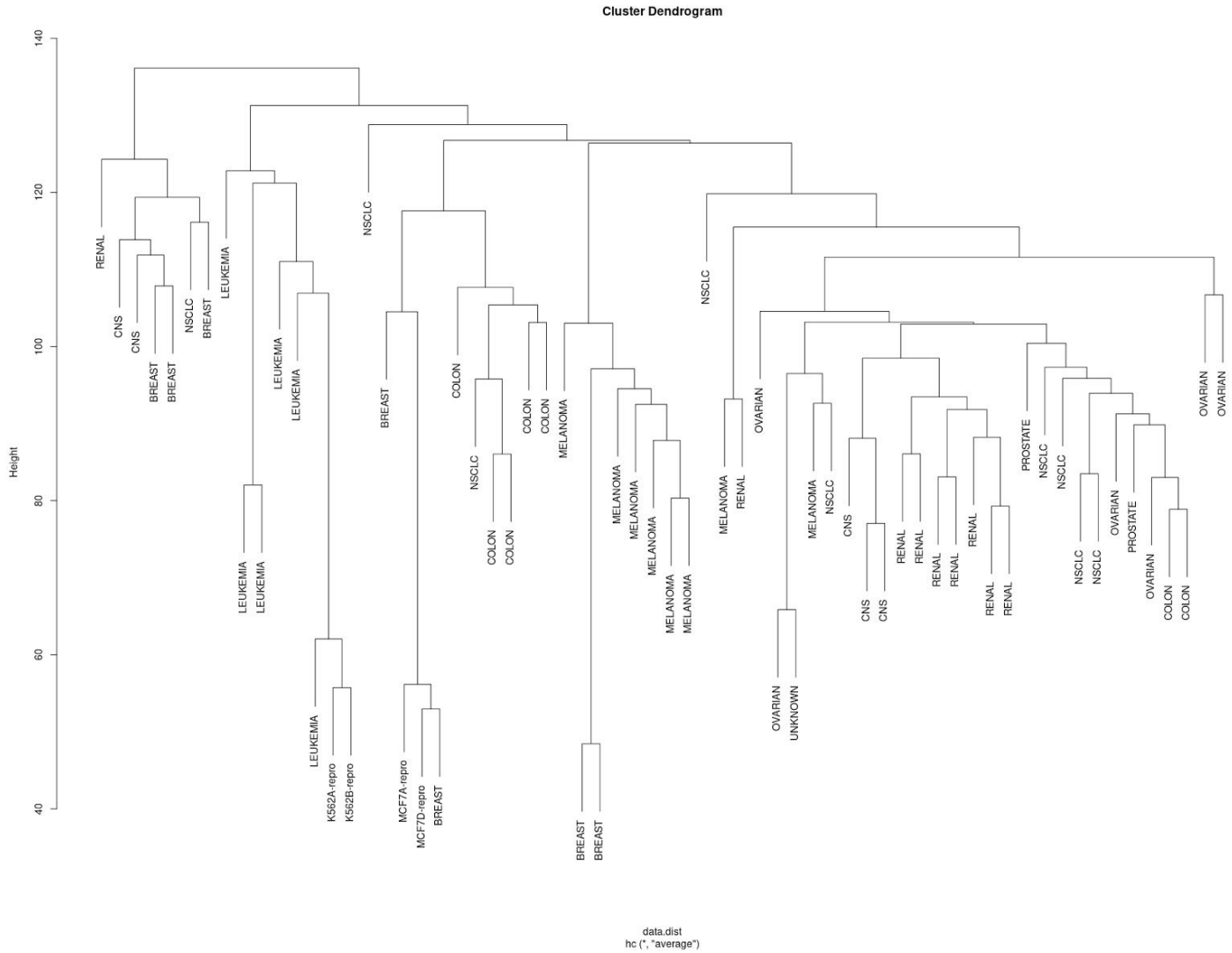
HAC IMPLEMENTATION

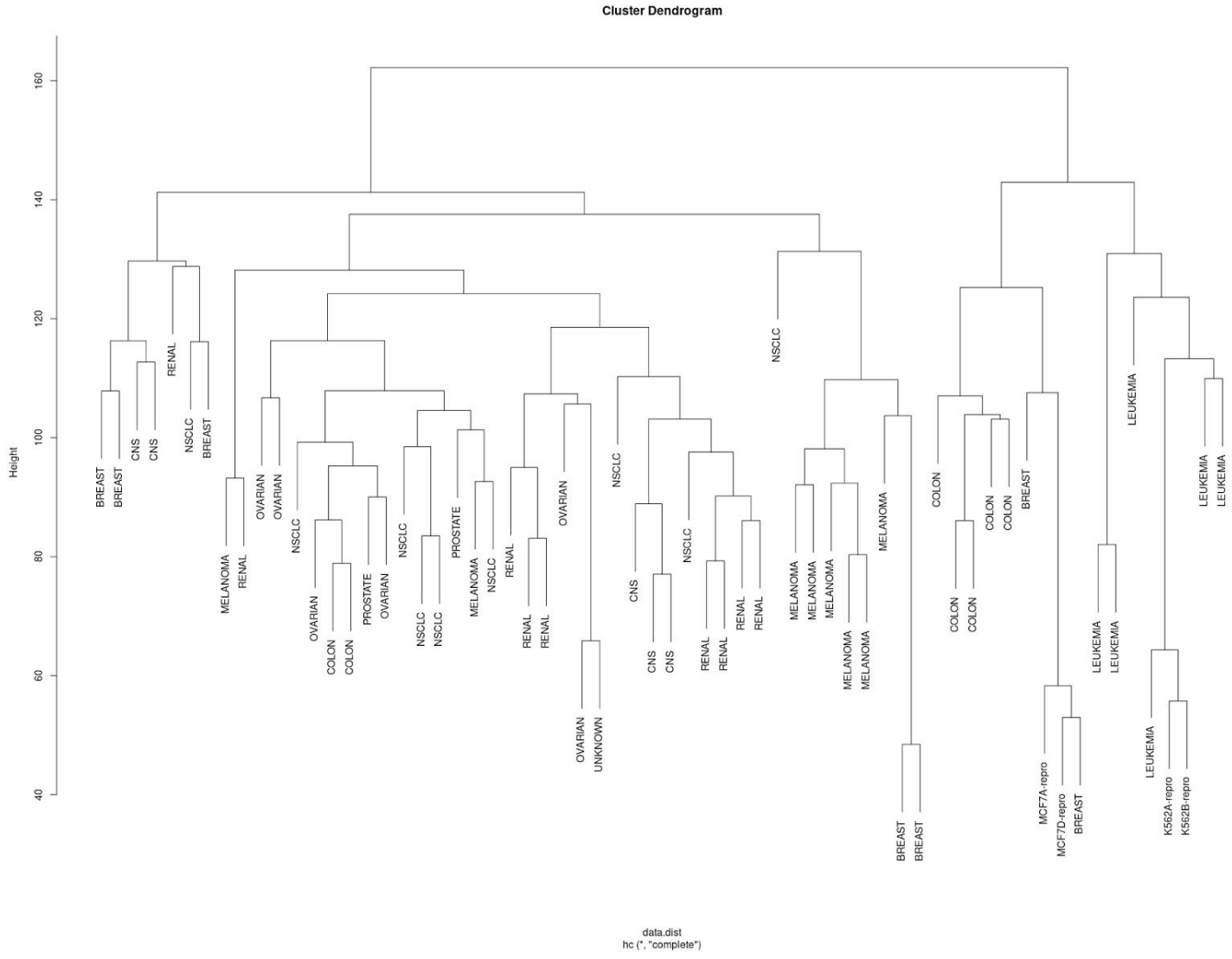
- We implemented HAC functions for ‘average’, ‘single’ , ‘complete’ and ‘centroid’ linkage.
- Applied the function to the dataset and made a dendrogram for each linkage.
- Before applying HAC we have used feature scaling on the dataset as the algorithm involves use of euclidean distance thus scaling ensures more uniform effect of each feature in result.
- The obtained dendrograms are:
- Single Linkage


Cluster Dendrogram
Height
40
50
60
70
80
90
100
110
LEUKEMIA
RENAL
BREAST
CNS
LEUKEMIA
NSCLC
LEUKEMIA
OVARIAN
NSCLC
CNS
BREAST
NSCLC
OVARIAN
COLON
BREAST
MELANOMA
RENAL
MELANOMA
MELANOMA
MELANOMA
MELANOMA
MELANOMA
BREAST
OVARIAN
COLON
NSCLC
NSCLC
PROSTATE
MELANOMA
COLON
OVARIAN
NSCLC
RENAL
COLON
PROSTATE
COLON
OVARIAN
COLON
COLON
NSCLC
NSCLC
RENAL
NSCLC
RENAL
RENAL
RENAL
RENAL
RENAL
CNS
CNS
CNS
LEUKEMIA
LEUKEMIA
LEUKEMIA
K562A-repro
K562B-repro
BREAST
BREAST
MCF7A-repro
MCF7D-repro
BREAST
OVARIAN
UNKNOWN
data.dist
hc (*, "single")

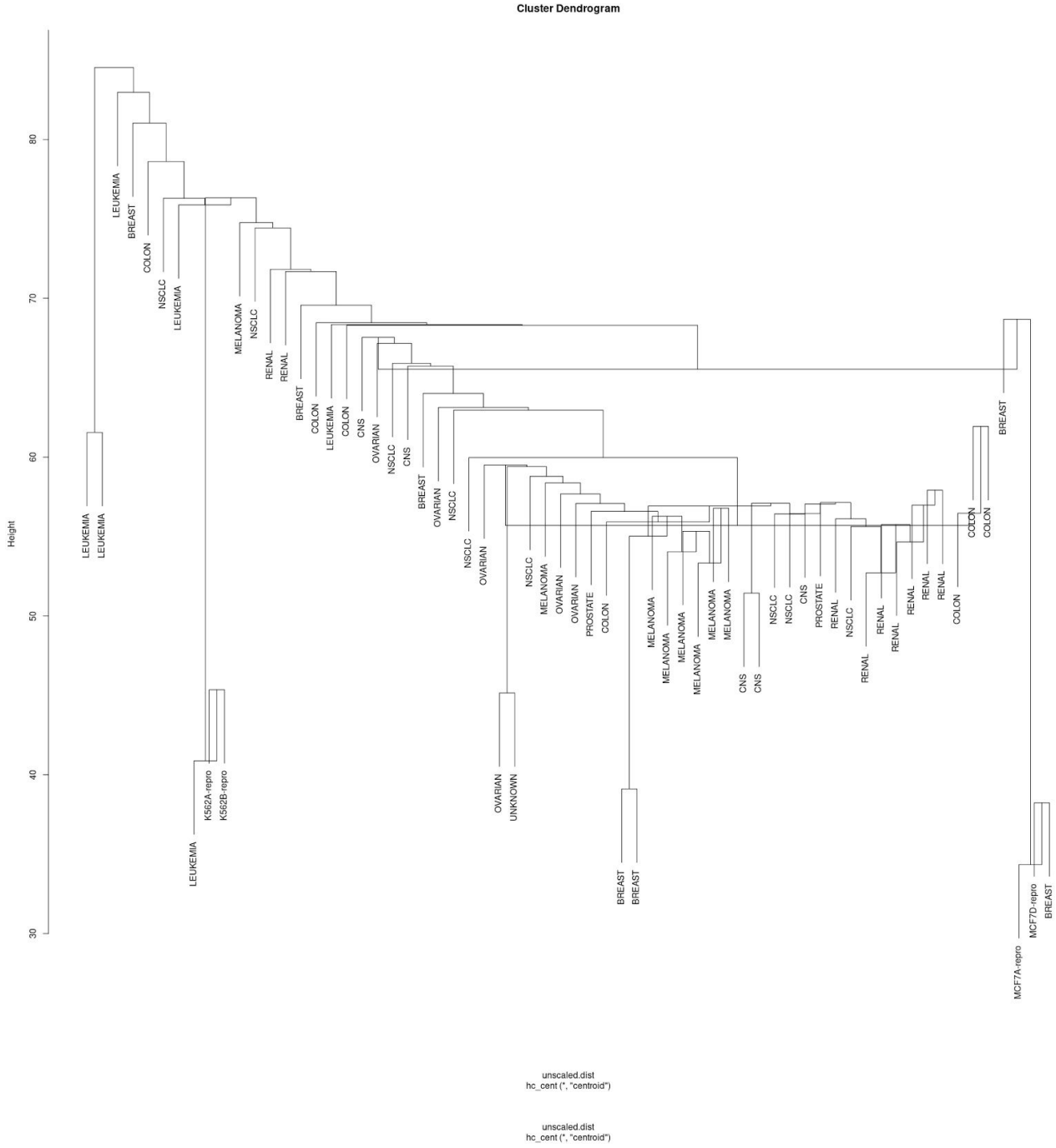
Average Linkage



Complete Linkage



Centroid linkage

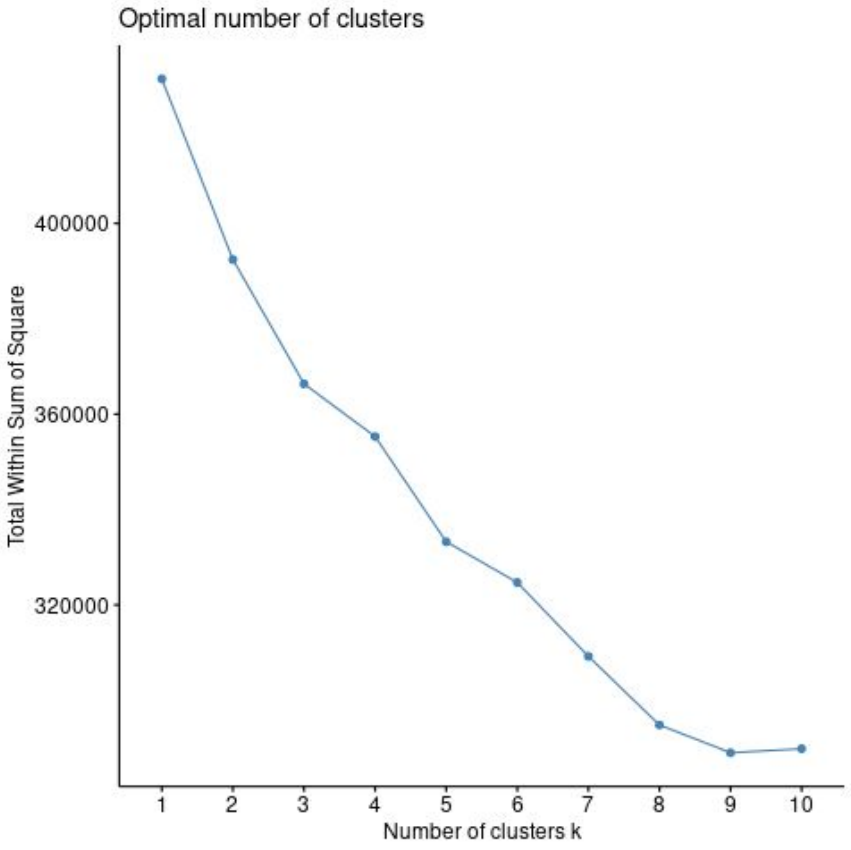


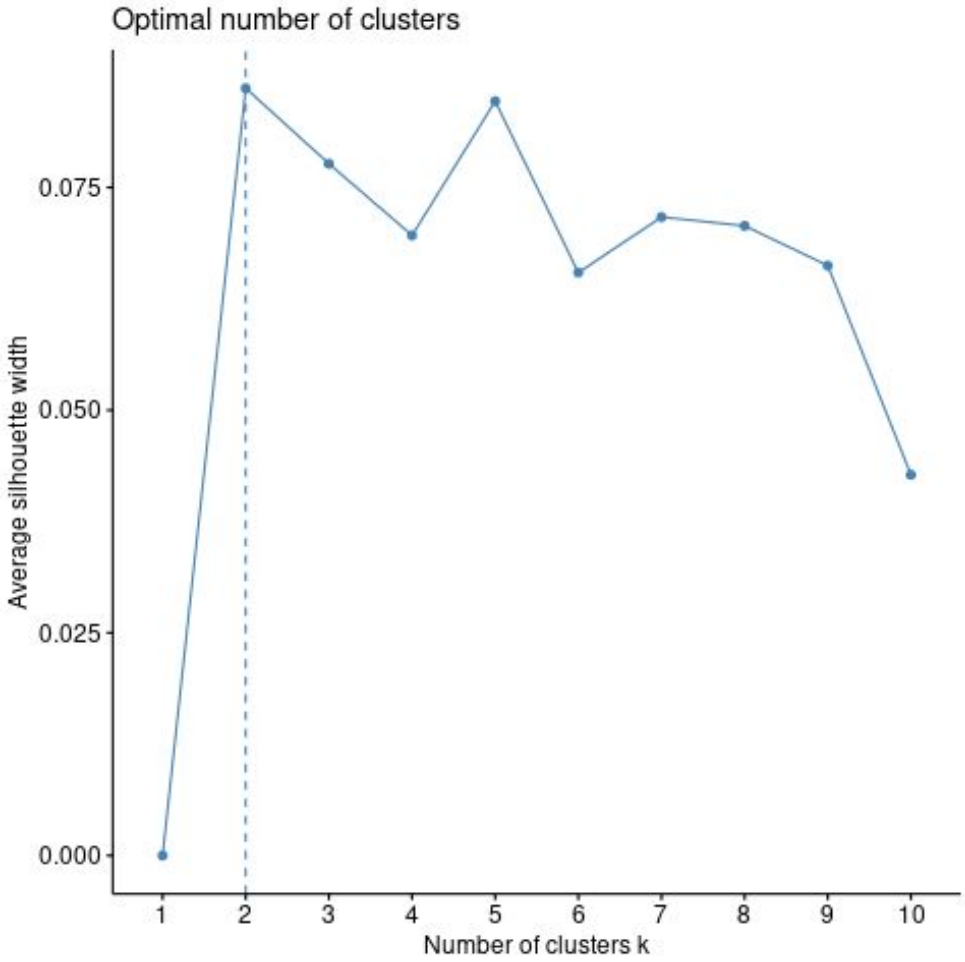
unscaled.dist
hc_cent (*, "centroid")

- Seeing the dendrograms formed by each linkage we find that using complete linkage gives best clusters followed by average linkage.The clusters are well defined and clusters formed are accurate.
- Single and centroid linkage dendrograms are more single branched and the clusters formed are not very clear.

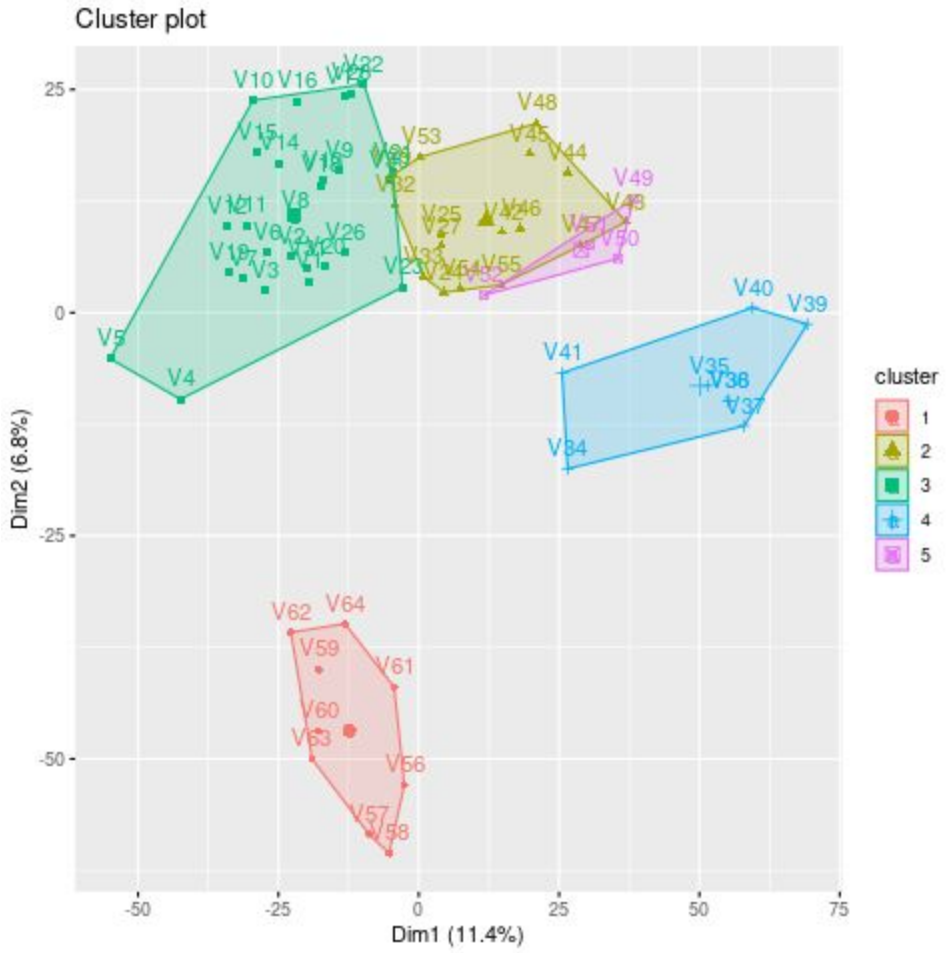
Implementing K-Means

- We have used elbow method and silhouette width to identify the k(the number of clusters).We find that 5 is the best number of clusters.


Optimal number of clusters
Total Within Sum of Square
400000
360000
320000
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Optimal number of clusters
Average silhouette width
0.075
0.050
0.025
0.000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

- After this we have applied kmeans to the dataset,Obtained and visualised the results



COMPARING K-MEANS AND HIERARCHICAL AGGLOMERATIVE CLUSTERING

- Distance used: Hierarchical clustering can virtually handle any distance metric while k-means rely on euclidean distances.

- Stability of results: k-means requires a random step at its initialization that may yield different results if the process is re-run. That wouldn't be the case in hierarchical clustering.
- Number of Clusters: While you can use elbow plots, Silhouette plot etc. to figure the right number of clusters in k-means, hierarchical too can use all of those but with the added benefit of leveraging the dendrogram for the same.