# Cluster Analysis on Breast Cancer Dataset

In this project, I run different clustering methods, namely Ward's method and the single linkage method, on the wisconsin breast cancer dataset with Manhattan and Euclidean distance measures. Then I also run KMeans clustering on the data and cross-check the results with our hierarchical clustering methods. Ultimately, the goal is to see if there are certain groups of observations that are similar.

```r
cancer_raw <- read.csv("Data/breast-cancer-wisconsin.csv",as.is=TRUE)

# Clean the data
# Note: the column "bare_nucleoli" has missing values
# We choose to omit rows that have missing (16 rows)
cancer <- as.data.frame(apply(cancer_raw,2,as.numeric))
cancer <- na.omit(cancer)
```

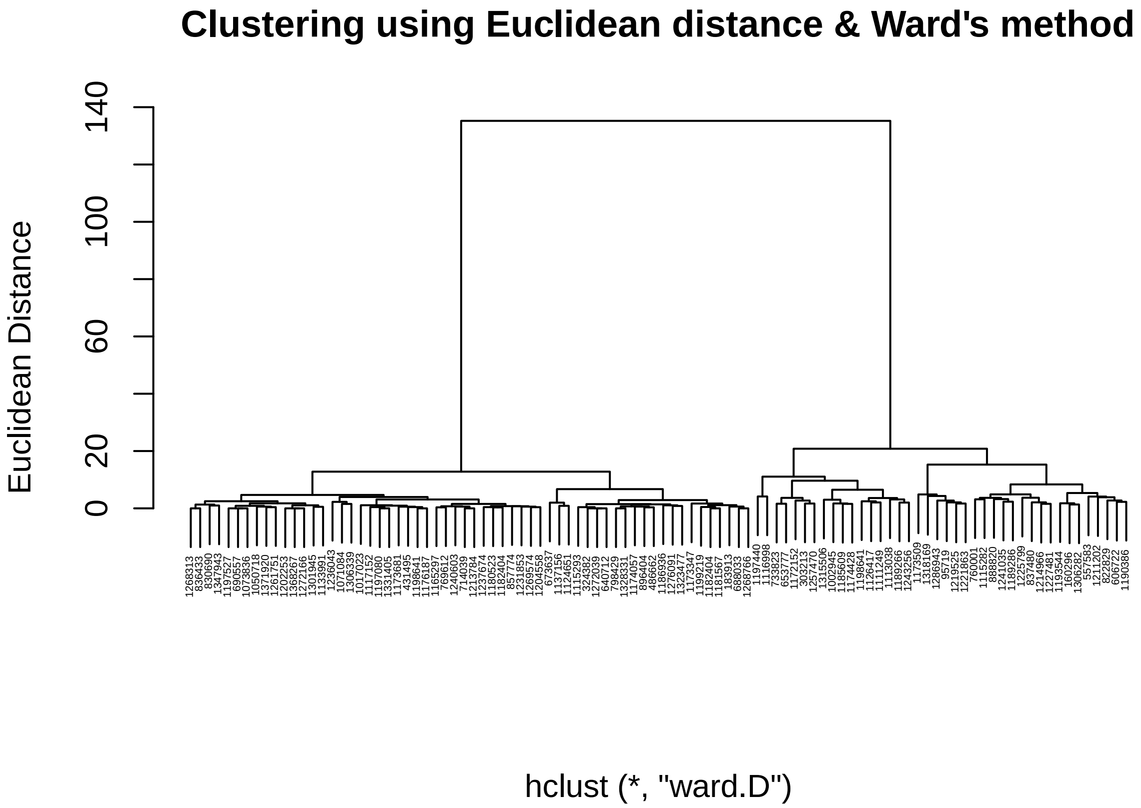
Here, I use the Euclidean distance and Manhattan distance metrics. This is because all our predictor variables are continuous, and these metrics are suitable for continuous variables. Furthermore, since the Minkowski distance is just a generalisation of the Euclidian and Manhattan distances, using these metrics instead will increase the interpretability of our cluster analysis.

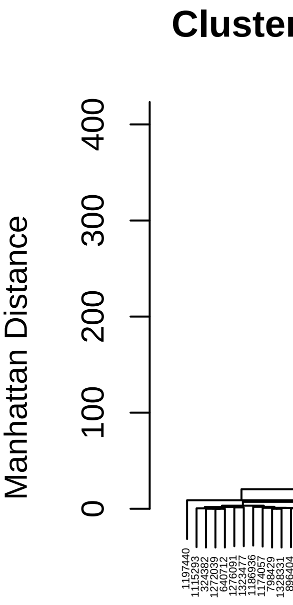
I will also be standardising our data. Although the variables in our original data are already on the same scale, i.e., 1 to 10, standardising our data will increase the interpretability of our results since we are standardising the mean across the predictors.

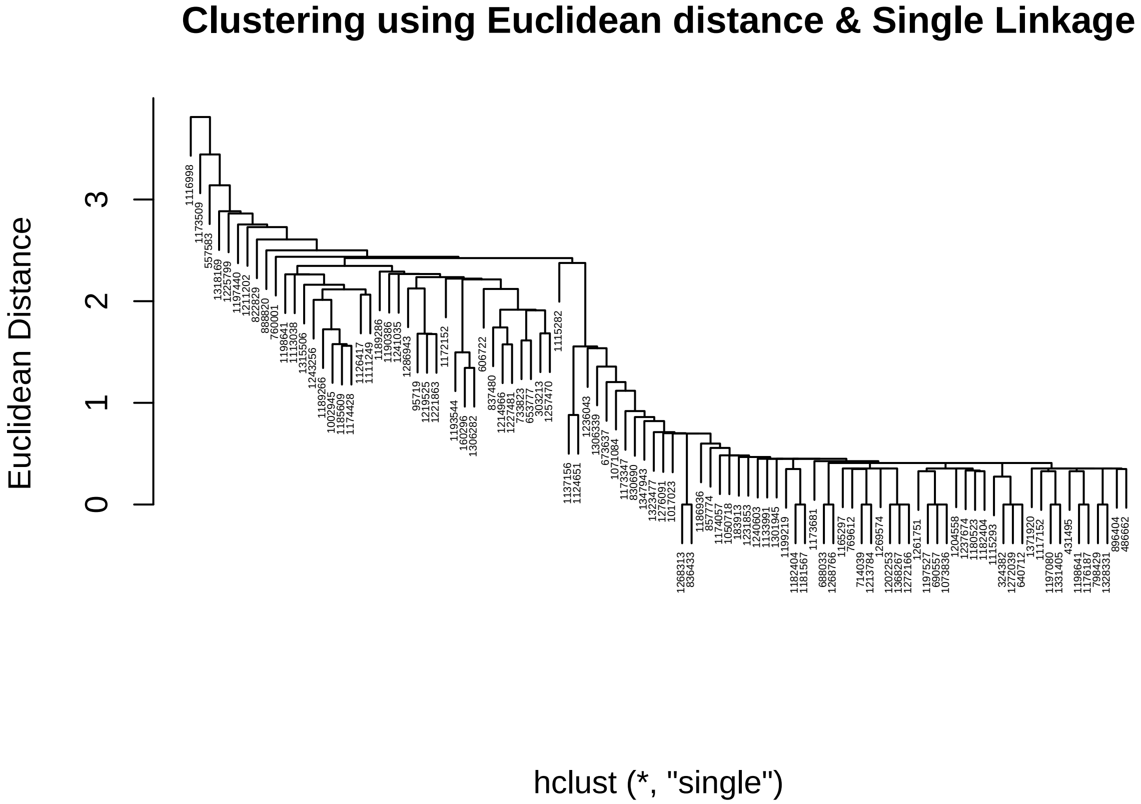
The dataset has over 600 observations, which makes it difficult to produce a readable dendrogram. Hence, we will sample 100 observations from our dataset to use for our analysis in this part.

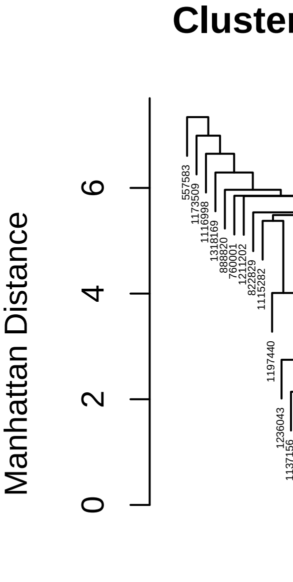
```r
# Standardize Data
scancer <- scale(cancer[,-c(1,11)])

# Sample 100 observations out of the total number of observations
# (This will make the dendrogram more readable)
set.seed(420)
index <- sample(1:nrow(scancer),100)
scancer <- scancer[index,]

# Calculate Distance Matrix
dist1 <- dist(scancer,method="euclidean") # Euclidean
dist2 <- dist(scancer,method="manhattan") # Manhattan

# Perform clustering using Ward's method
clust1_1 <- hclust(dist1,method="ward.D") # Euclidean distance, Ward's method
clust1_2 <- hclust(dist2,method="ward.D") # Manhattan distance, Ward's method

# Perform clustering using Single Linkage method
clust2_1 <- hclust(dist1,method="single") # Euclidean distance, Single Linkage method
clust2_2 <- hclust(dist2,method="single") # Manhattan distance, Single Linkage method
```

Here are the dendrograms for each approach:



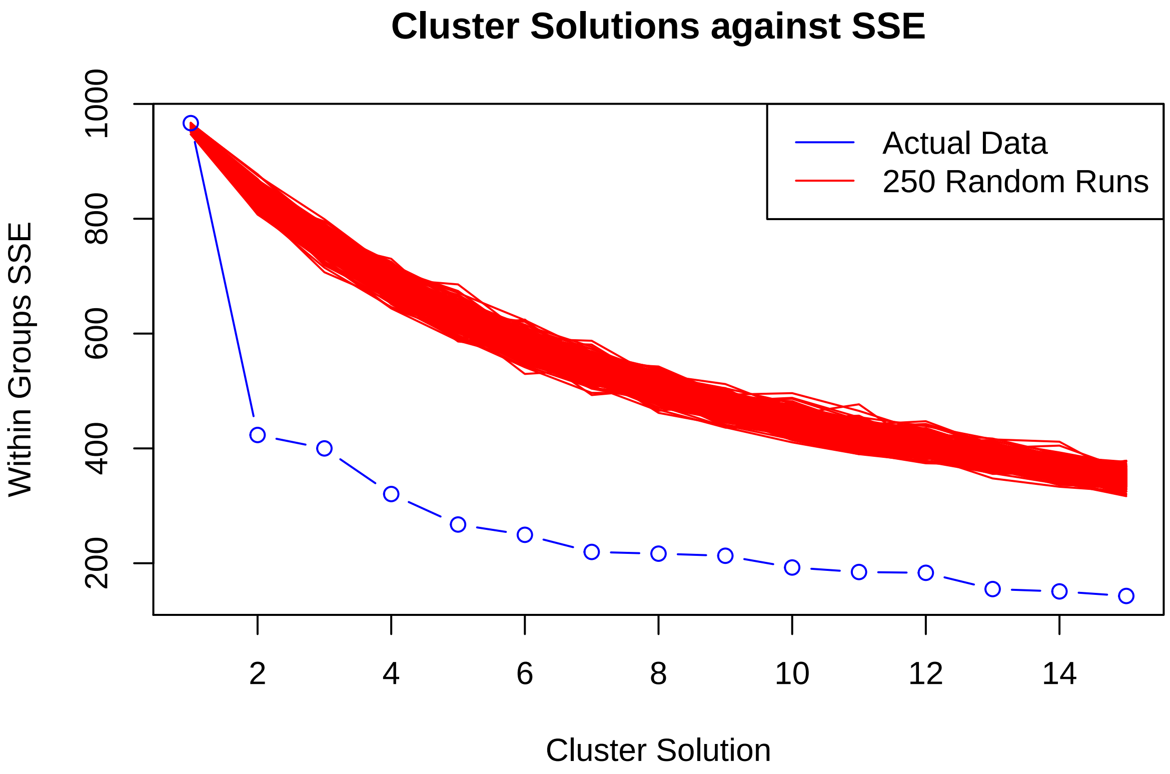






In general, there are 2 groups present when the clustering has been performed. However, when Ward's method was used, there are 2 more distinct groups, i.e., 4 distinct groups. The single linkage method could only yield 2 distinct groups since many groups branched out to other groups, as seen in the following plots:

Here, I use the modified R script by Matt Peeples to perform the k-means clustering algorithm for $1 \leq k \leq 15$.



From the results and scree-plot, we see that there are likely to be 3 groups that exist since 3 is at the point where the ‘elbow’ is. After 3 clusters, the within groups SSE does not seem to improve significantly, i.e., decrease, much as we increase the number of clusters.

Given that the dataset actually classifies the orbservations into the 2 categories of malignant and benign tumours, observing between 2 to 4 clusters in the data seems fairly consistent with the response variable of the original data (which has two categories). Further analysis could be done to examine whether the clustering of the dataset coincides with the classification provided in the dataset.