R Notebook

Packages and Data

```
install.packages('factoextra'); install.packages('cluster')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages('ape'); install.packages('ggdendro')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(ape); library(ggdendro); library(factoextra); library(cluster)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
set.seed(0)
mnist_test = read.csv('mnist_test.csv')
mini_data = mnist_test[sample(200), 2:785]
```

Scale Data and Calculate the distances

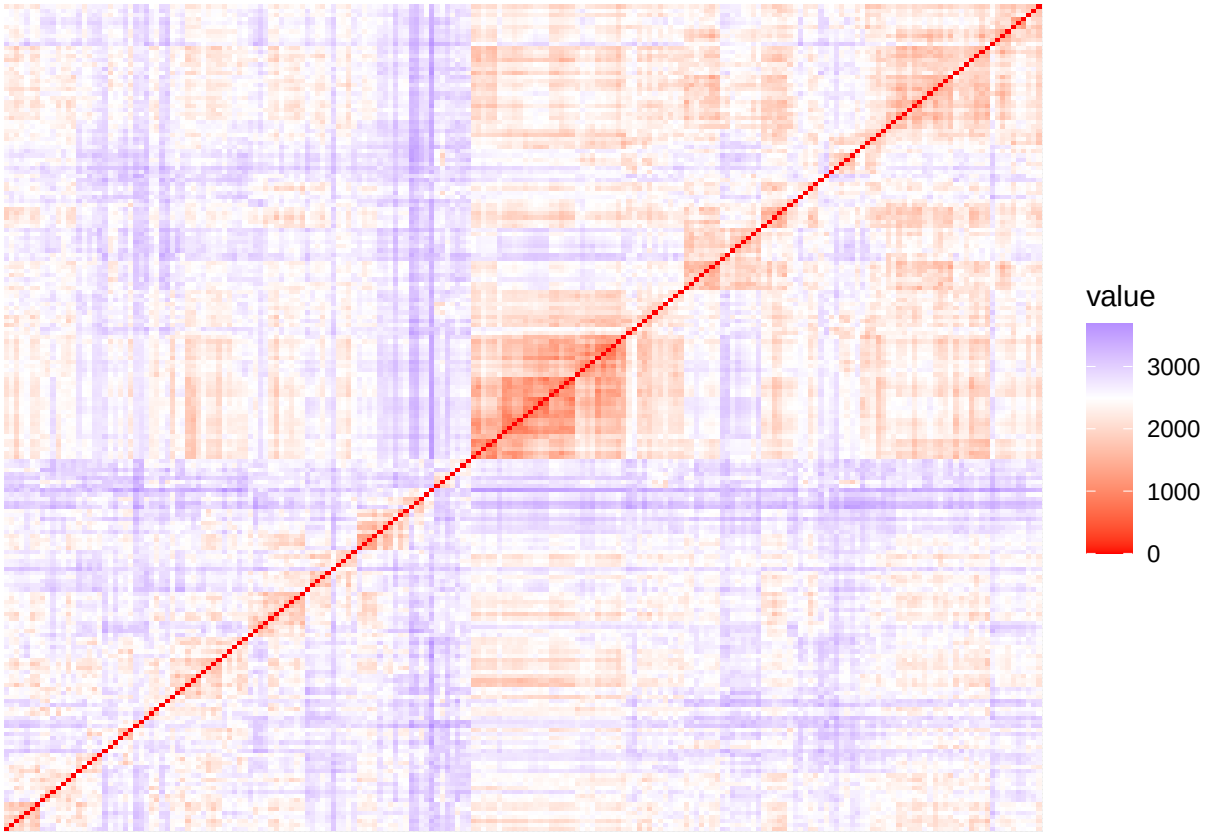
```
scaled_data = mini_data/255 +1e-3
distances = dist(scaled_data)
```

Clusterability

```
clusterability = get_clust_tendency(mini_data, n = 90)
print(clusterability$hopkins)
```

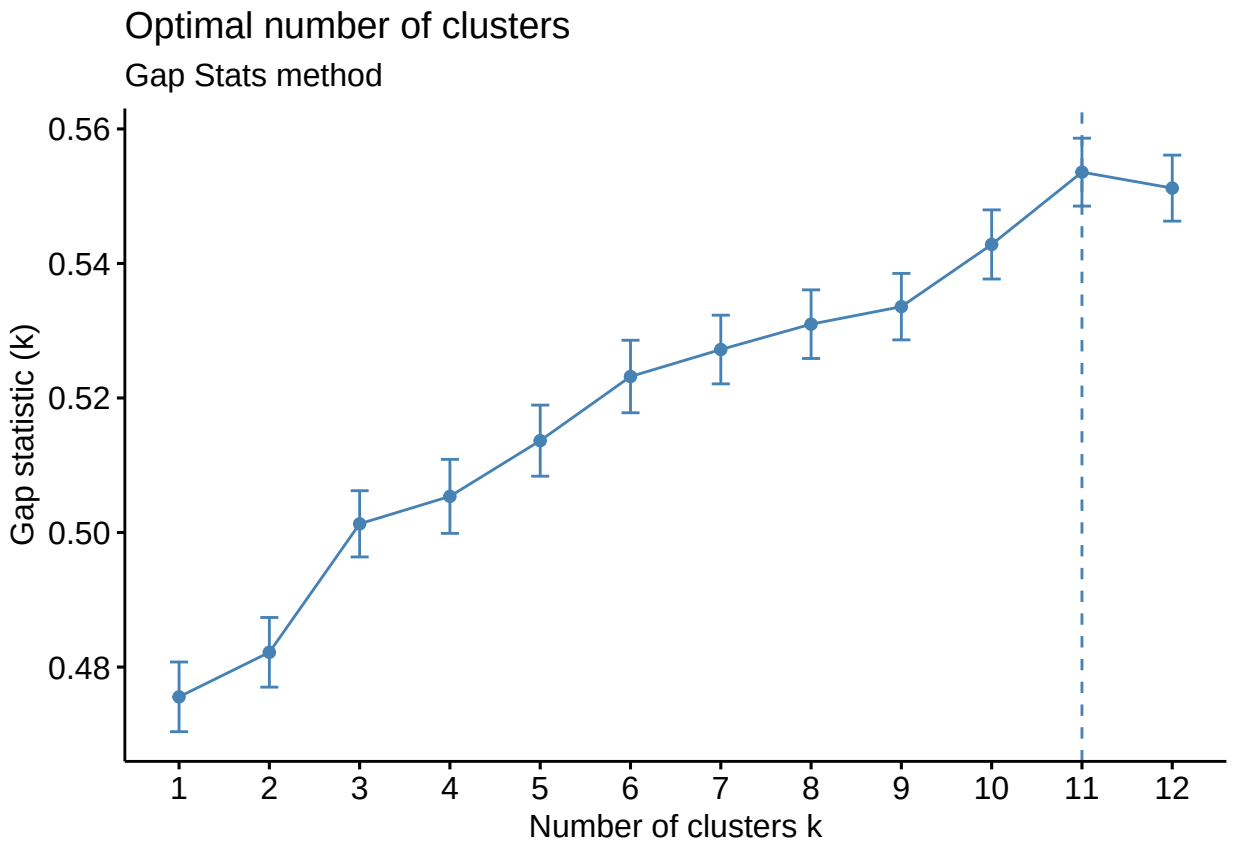
```
## [1] 0.6495675
```

```
clusterability$plot
```

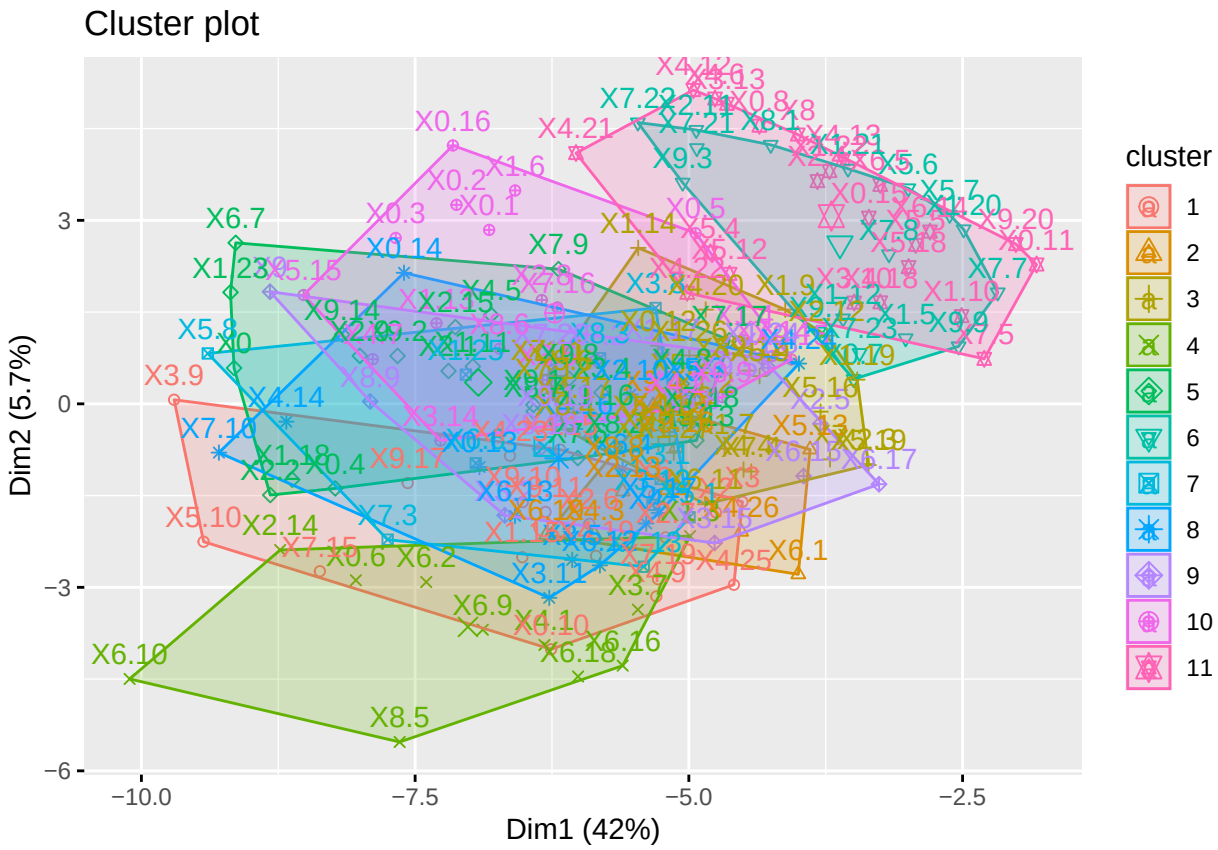


Gap Statistics and Elbow

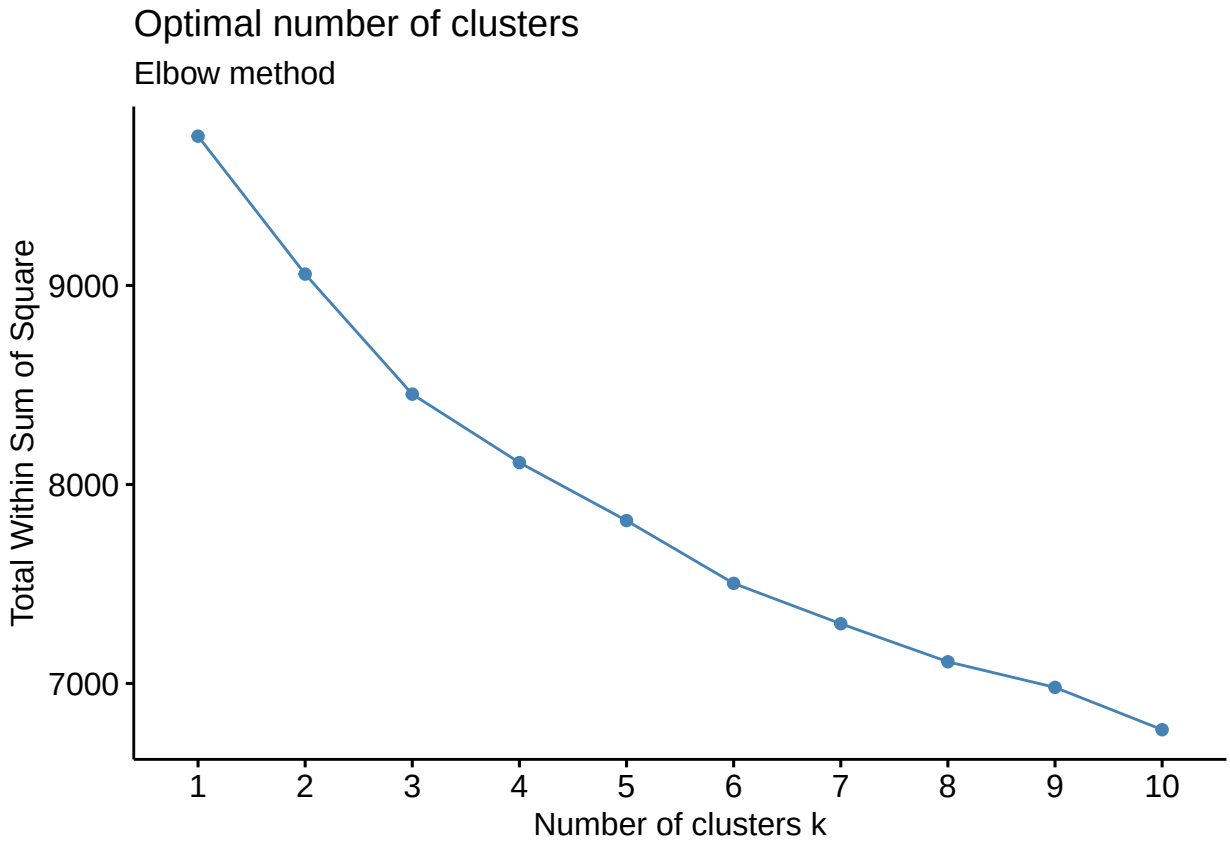
```
#set.seed(42)
fviz_nbclust(scaled_data, kmeans, method = 'gap_stat', k.max = 12, nboot = 50) +
  labs(subtitle = 'Gap Stats method')
```




```
my_kmeans = kmeans(scaled_data, centers = 11, nstart = 100)
row.names(scaled_data) = make.names(mnist_test$label[1:200], unique = T)
fviz_cluster(my_kmeans, scaled_data, stand = F)
```



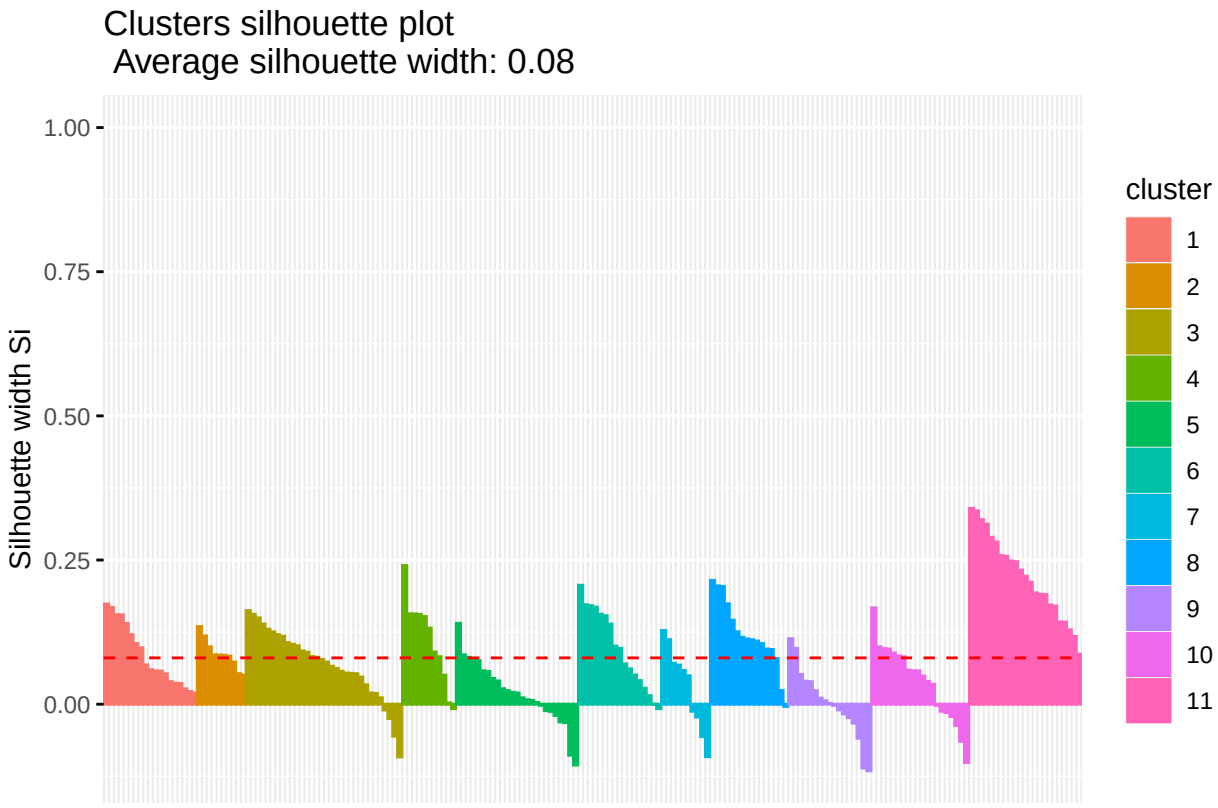

```
fviz_nbclust(scaled_data, kmeans, method = 'wss') + labs(subtitle = 'Elbow method')
```



Performance Validation

```
silhouette = silhouette(my_kmeans$cluster, distances)
fviz_silhouette(silhouette)
```

```
##    cluster size ave.sil.width
## 1        1   19          0.08
## 2        2   10          0.09
## 3        3   32          0.07
## 4        4   11          0.11
## 5        5   25          0.02
## 6        6   17          0.10
## 7        7   10          0.03
## 8        8   16          0.12
## 9        9   17          0.00
## 10      10   20          0.04
## 11      11   23          0.22
```



A fancy dendrogram

```
hc = hclust(distances, method = 'ward.D2')
nclust = cutree(hc, 11)
colours = c('red', 'green', 'yellow', 'blue', 'black', 'pink',
            'orange', 'purple', 'brown', 'grey', 'chocolate')
plot(as.phylo(hc), type = 'fan', tip.color = colours[nclust],
     label.offset = 1, cex = 0.5)
```