MarketAnalysis

Indhresh Achi

2025-04-14

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(readr)
library(tidyr)
library(ggplot2)
```

```
# Importing Dataset
```

```
market <- read.csv("/Users/indhreshachi/Desktop/MarketingAnalysis.csv")
str(market)
```

```
## 'data.frame':    1000 obs. of  9 variables:
##  $ id                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age                 : int  38 21 60 40 65 31 19 43 53 55 ...
##  $ gender              : chr  "Female" "Female" "Female" "Other" ...
##  $ income              : int  99342 78852 126573 47099 140621 57305 54319
108115 34424 45839 ...
##  $ spending_score      : int  90 60 30 74 21 24 68 94 29 55 ...
##  $ membership_years    : int  3 2 2 9 3 3 5 9 6 7 ...
##  $ purchase_frequency  : int  24 42 28 5 25 30 43 27 7 2 ...
##  $ preferred_category  : chr  "Groceries" "Sports" "Clothing" "Home &
Garden" ...
##  $ last_purchase_amount: num  113.5 41.9 424.4 991.9 347.1 ...
```

```
# Data Wrangling
```

```
names(market)
```

```
## [1] "id"                    "age"                   "gender"
## [4] "income"                "spending_score"        "membership_years"
## [7] "purchase_frequency"    "preferred_category"    "last_purchase_amount"
```

```
market <- subset(market, select=-c(id))
head(market)
```

```
##   age gender income spending_score membership_years purchase_frequency
## 1  38 Female  99342             90                3                  24
## 2  21 Female  78852             60                2                  42
## 3  60 Female 126573             30                2                  28
## 4  40  Other  47099             74                9                   5
## 5  65 Female 140621             21                3                  25
## 6  31  Other  57305             24                3                  30
##   preferred_category last_purchase_amount
## 1          Groceries               113.53
## 2             Sports                41.93
## 3           Clothing               424.36
## 4      Home & Garden               991.93
## 5        Electronics               347.08
## 6      Home & Garden                86.85
```

```
summary(market)
```

```
##       age           gender              income       spending_score
##  Min.   :18.00   Length:1000        Min.   : 30004   Min.   :  1.00
##  1st Qu.:30.00   Class :character   1st Qu.: 57912   1st Qu.: 26.00
##  Median :45.00   Mode  :character   Median : 87846   Median : 50.00
##  Mean   :43.78                      Mean   : 88501   Mean   : 50.69
##  3rd Qu.:57.00                      3rd Qu.:116110   3rd Qu.: 76.00
##  Max.   :69.00                      Max.   :149973   Max.   :100.00
##  membership_years purchase_frequency preferred_category
last_purchase_amount
##  Min.   : 1.000   Min.   : 1.0       Length:1000        Min.   : 10.4
##  1st Qu.: 3.000   1st Qu.:15.0       Class :character   1st Qu.:218.8
##  Median : 5.000   Median :27.0       Mode  :character   Median :491.6
##  Mean   : 5.469   Mean   :26.6                          Mean   :492.3
##  3rd Qu.: 8.000   3rd Qu.:39.0                          3rd Qu.:747.2
##  Max.   :10.000   Max.   :50.0                          Max.   :999.7
```

```
summary(market$spending_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   26.00   50.00   50.69   76.00  100.00
```

```
summary(market$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30004   57912   87846   88501  116110  149973
```

```
summary(market$last_purchase_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.4   218.8   491.6   492.3   747.2   999.7
```

```
mean(market$income)
```

```
## [1] 88500.8
```

```
mean(market$spending_score)
```

```
## [1] 50.685
```

```
mean(market$membership_years)
```

```
## [1] 5.469
```

```
mean(market$last_purchase_amount)
```

```
## [1] 492.3487
```

```
sd(market$income)
```

```
## [1] 34230.77
```

```
sd(market$spending_score)
```

```
## [1] 28.95518
```

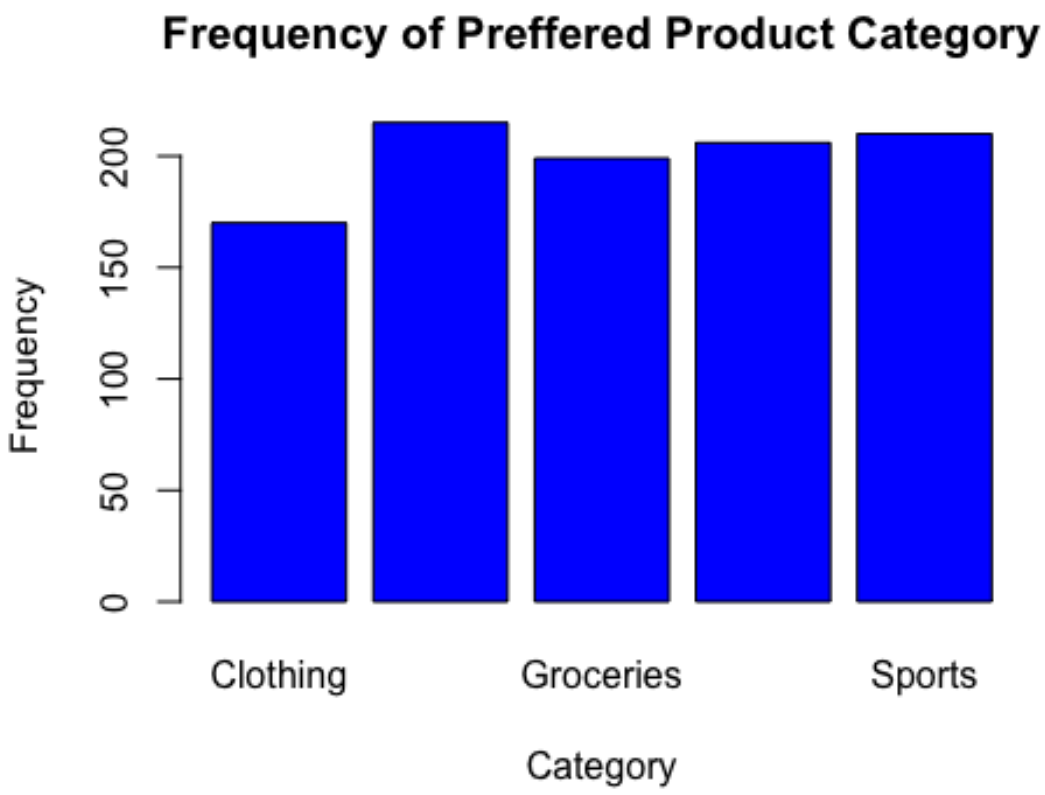
```
sd(market$purchase_frequency)
```

```
## [1] 14.24365
```

```
# Data Visualization
```

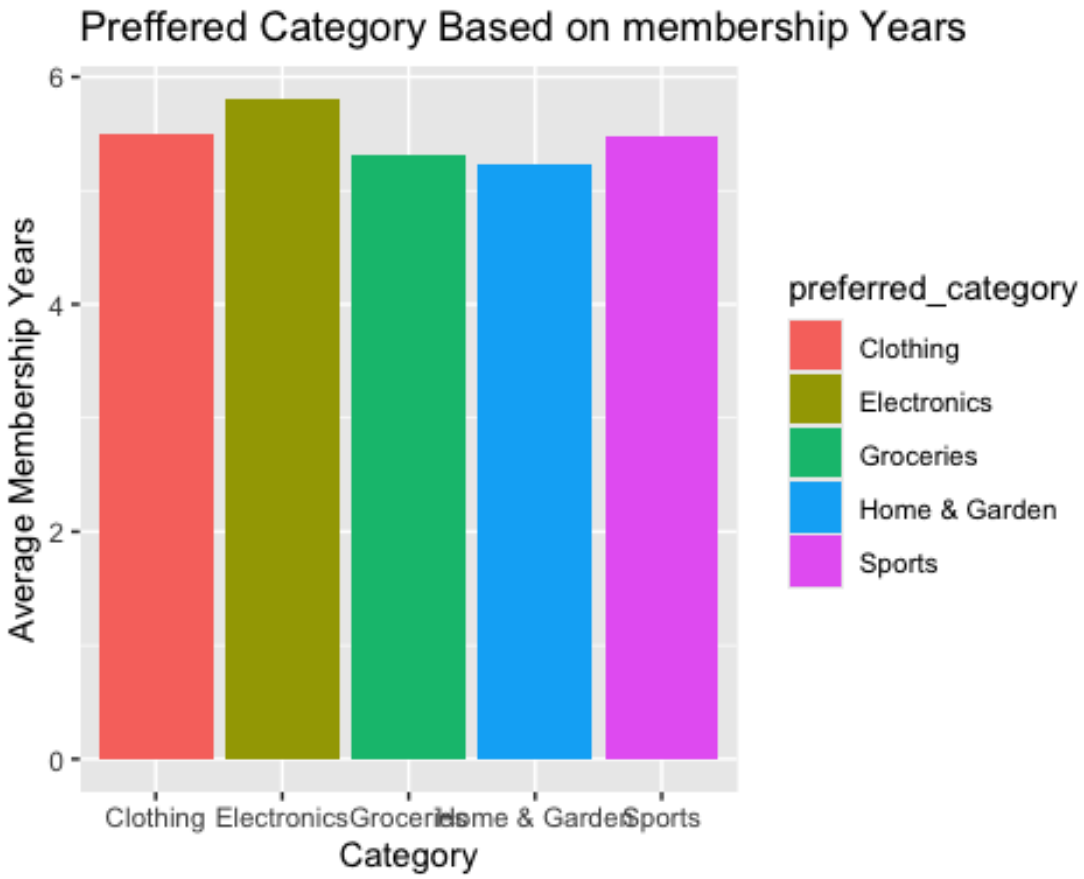
```
boxplot(market$income, main="Income Statistics", col="Green")
```


```r
product_freq = table(market$preferred_category)

barplot(height = product_freq,
        main="Frequency of Preffered Product Category",
        xlab="Category", ylab="Frequency", col="Blue")
```




```r
hist(market$age, main="Distribution of Age", xlab="Age",ylab="Frequency",
col="Red", labels=TRUE)
```


```
member <-  market
member %>%
  group_by(preferred_category) %>%
  mutate(mean_membershipYears = mean(membership_years)) %>%
  select(preferred_category, mean_membershipYears, gender) %>%
  ggplot(member, mapping= aes(fill=preferred_category,y=mean_membershipYears,
x=preferred_category)) +
    geom_bar(position = 'dodge', stat='identity') +
    ggtitle("Preffered Category Based on membership Years")+
  xlab("Category")+
  ylab("Average Membership Years")
```




```
# Machine Learning

## K-Means Clustering

### Elbow Method

library(purrr)

market$gender <- unclass(factor(market$gender))
market$preferred_category <- unclass(factor(market$preferred_category))

head(market)

##   age gender income spending_score membership_years purchase_frequency
## 1  38      1  99342             90                3                  24
## 2  21      1  78852             60                2                  42
## 3  60      1 126573             30                2                  28
## 4  40      3  47099             74                9                   5
## 5  65      1 140621             21                3                  25
## 6  31      3  57305             24                3                  30
##   preferred_category last_purchase_amount
## 1                  3               113.53
## 2                  5                41.93
```

```
## 3                  1               424.36
## 4                  4               991.93
## 5                  2               347.08
## 6                  4                86.85
```

```
cluster_df <- market
names(cluster_df)
```

```
## [1] "age"                  "gender"               "income"
## [4] "spending_score"       "membership_years"     "purchase_frequency"
## [7] "preferred_category"   "last_purchase_amount"
```

```
head(cluster_df)
```

```
##   age gender income spending_score membership_years purchase_frequency
## 1  38      1  99342             90                3                  24
## 2  21      1  78852             60                2                  42
## 3  60      1 126573             30                2                  28
## 4  40      3  47099             74                9                   5
## 5  65      1 140621             21                3                  25
## 6  31      3  57305             24                3                  30
##   preferred_category last_purchase_amount
## 1                  3               113.53
## 2                  5                41.93
## 3                  1               424.36
## 4                  4               991.93
## 5                  2               347.08
## 6                  4                86.85
```

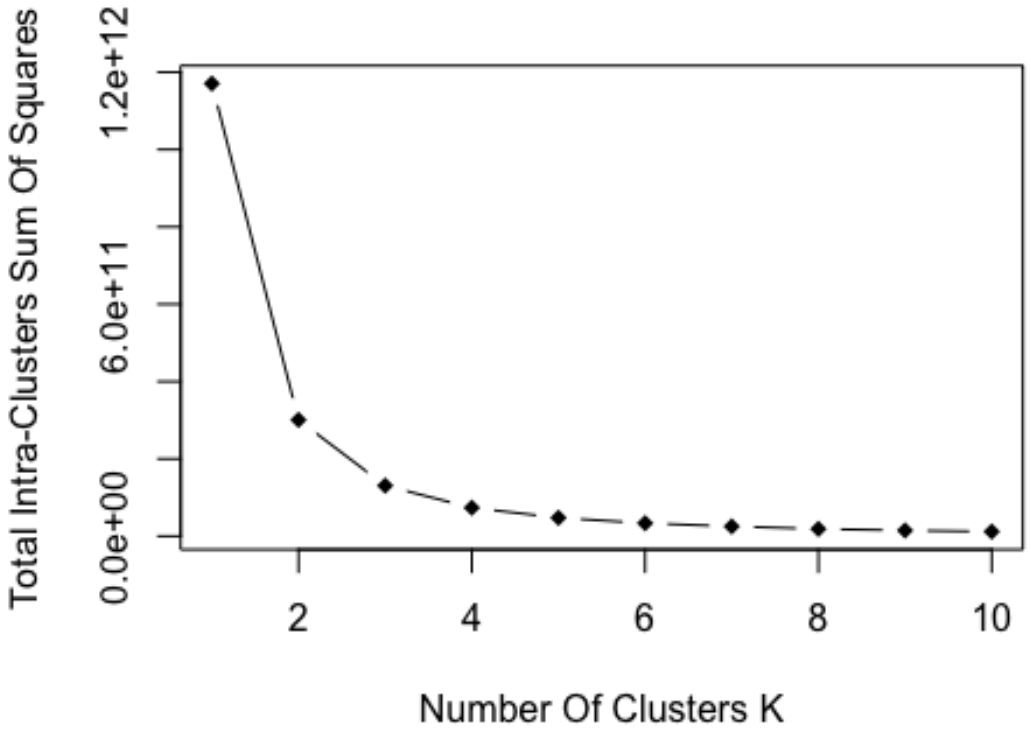
```
set.seed(124)
```

```
iss <- function(k) { kmeans(cluster_df, k,
                            iter.max=100,
                            nstart=100,
                            algorithm="Lloyd")$tot.withinss }
```

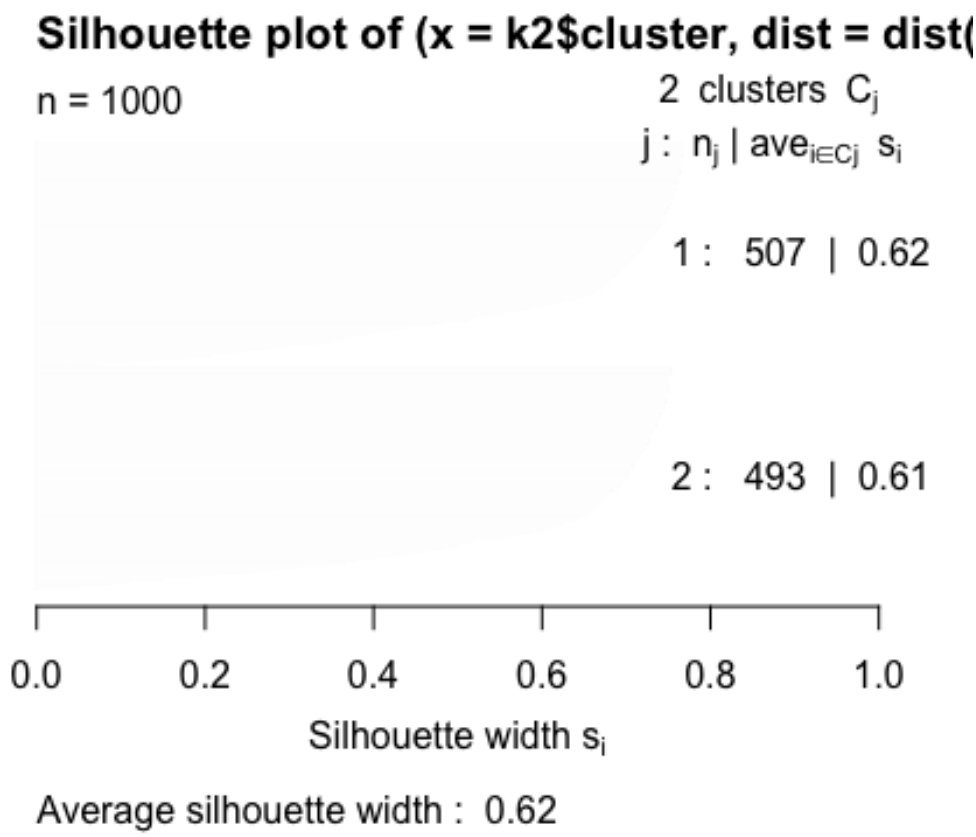
```
k.values <-  1:10
```

```
iss_values <- map_dbl(k.values, iss)
```

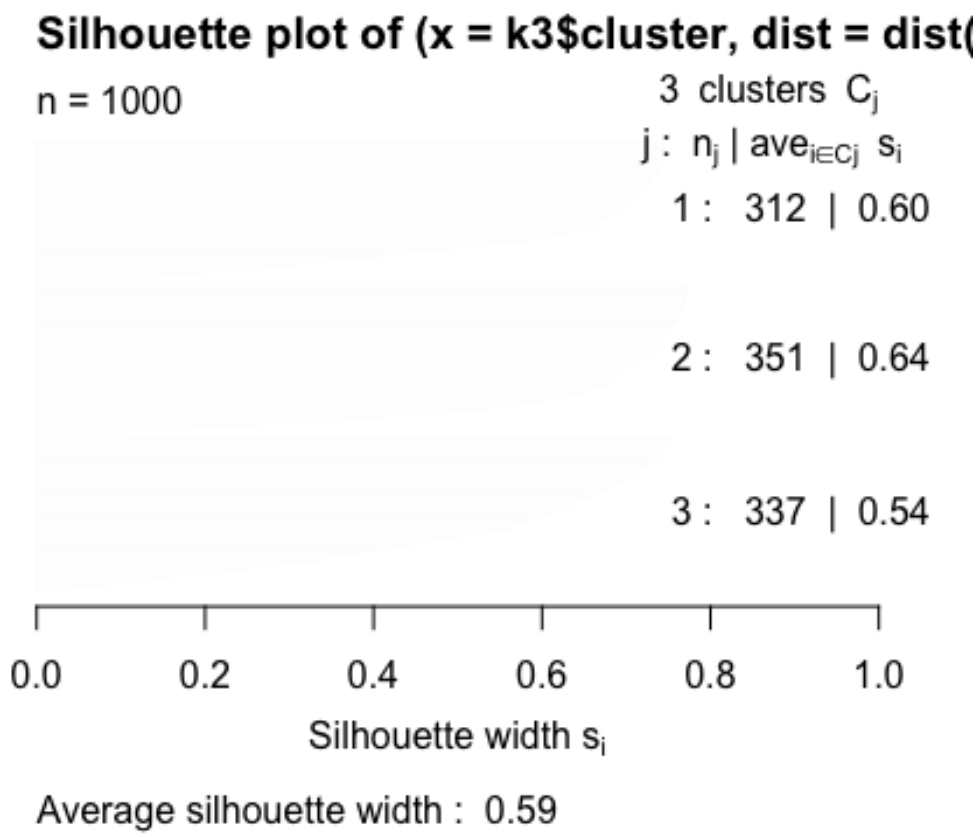
```
plot(k.values, iss_values, type="b", pch=18, frame=TRUE, xlab="Number Of
Clusters K",
     ylab="Total Intra-Clusters Sum Of Squares")
```



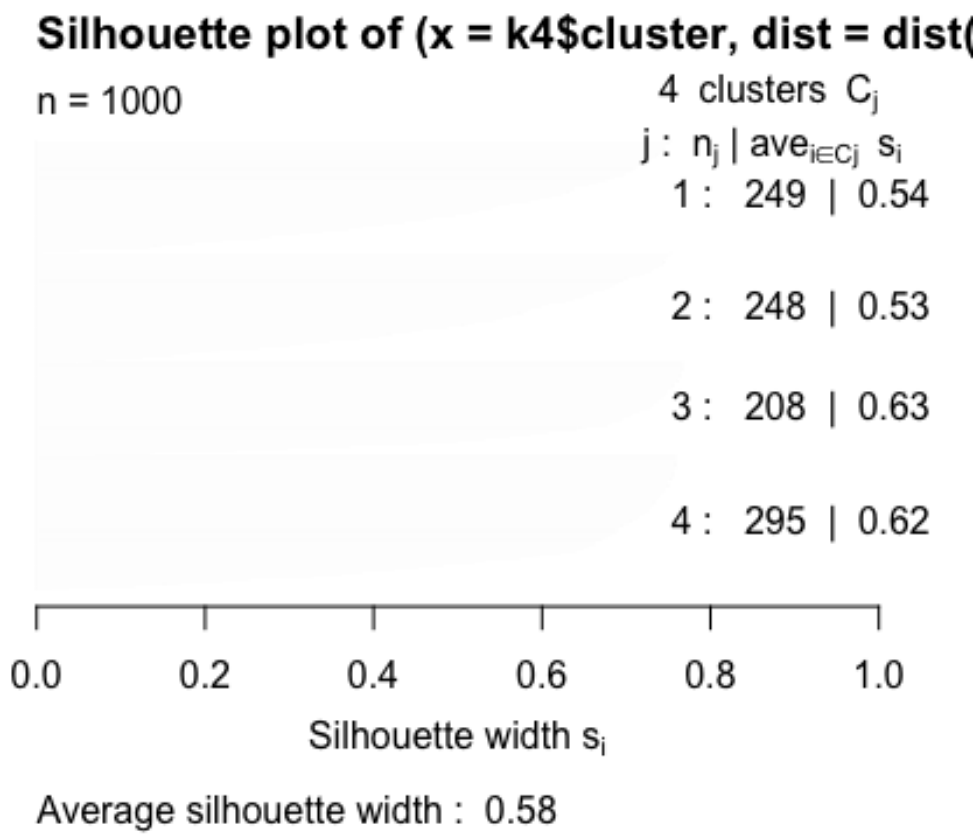

```
### Silhouwette Method

library(cluster)
library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

library(grid)

k2 <- kmeans(cluster_df, 2, iter.max=100,nstart=50,algorithm="Lloyd")
s2 <- plot(silhouette(k2$cluster, dist(cluster_df, "euclidean")))
```



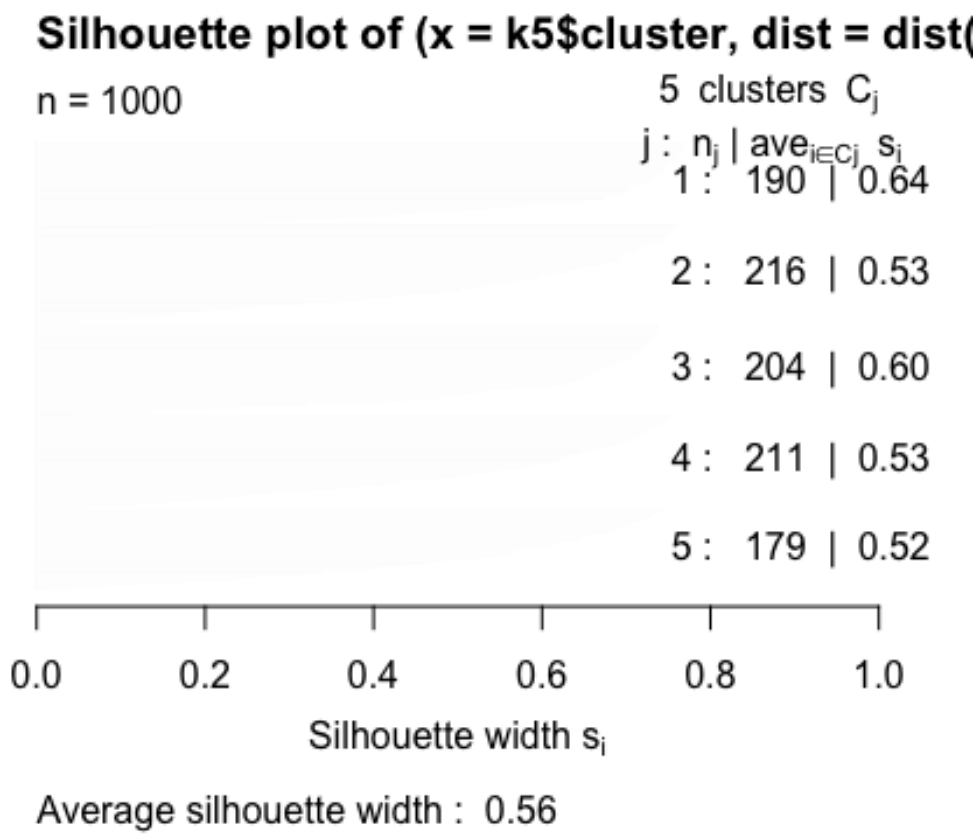

```r
k3 <- kmeans(cluster_df, 3, iter.max=100,nstart=50,algorithm="Lloyd")
s3 <- plot(silhouette(k3$cluster, dist(cluster_df, "euclidean")))
```



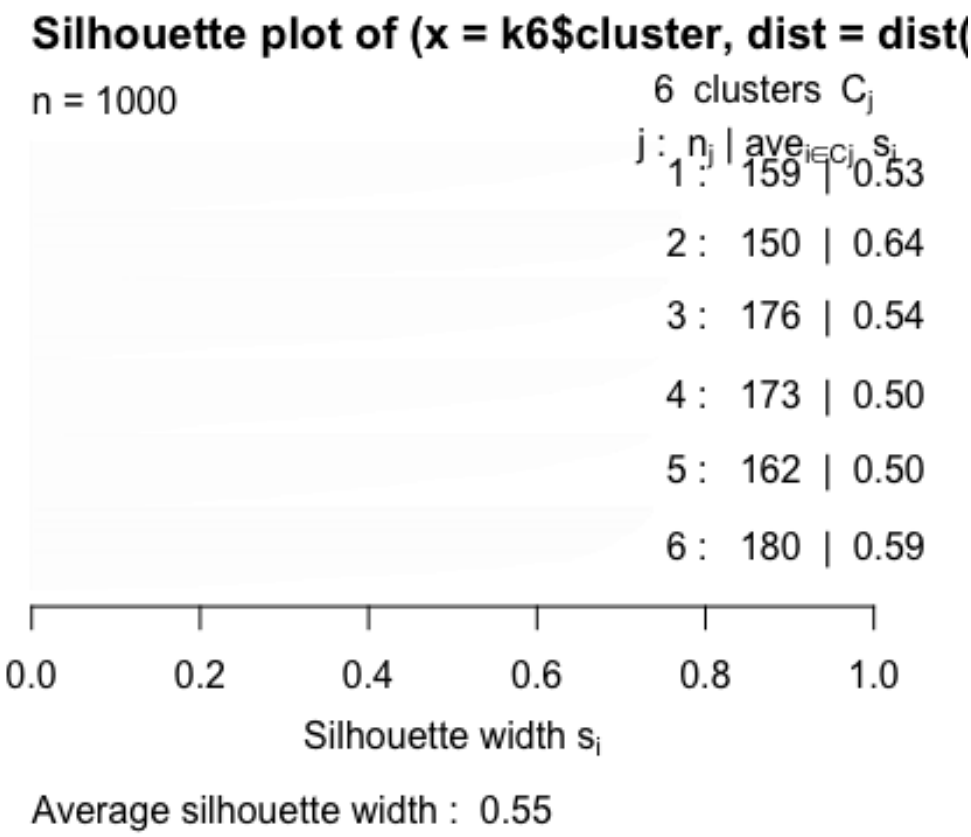

```r
k4 <- kmeans(cluster_df, 4, iter.max=100,nstart=50,algorithm="Lloyd")
s4 <- plot(silhouette(k4$cluster, dist(cluster_df, "euclidean")))
```



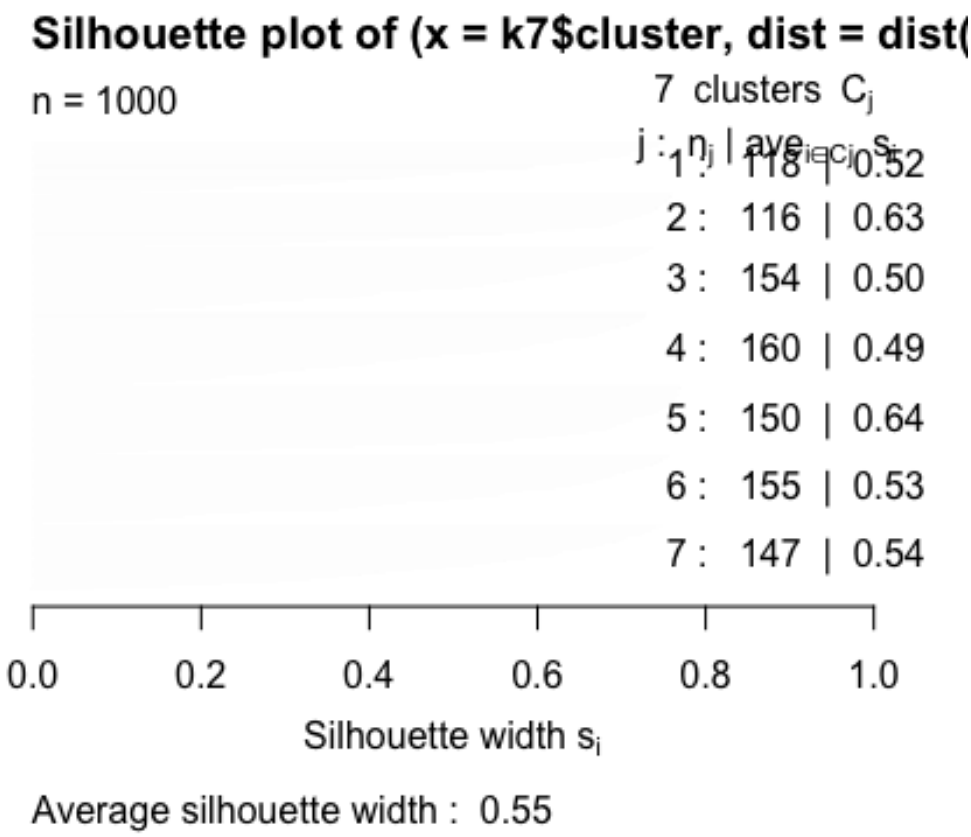

```r
k5 <- kmeans(cluster_df, 5, iter.max=100,nstart=50,algorithm="Lloyd")
s5 <- plot(silhouette(k5$cluster, dist(cluster_df, "euclidean")))
```



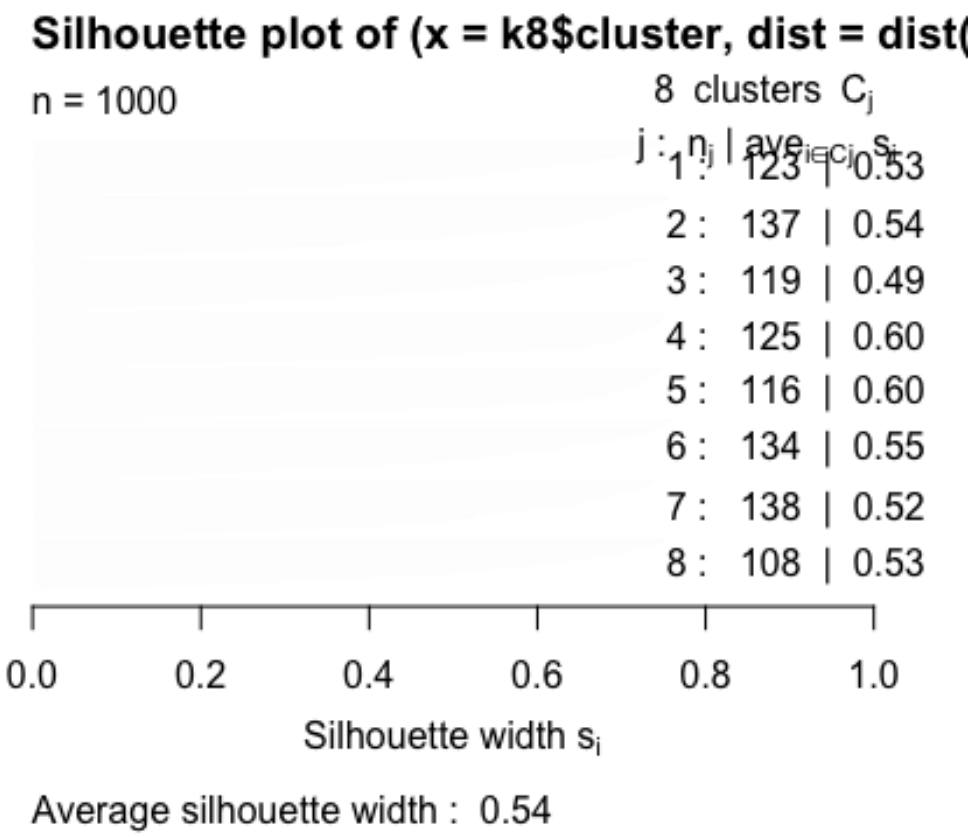

```r
k6 <- kmeans(cluster_df, 6, iter.max=100,nstart=50,algorithm="Lloyd")
s6 <- plot(silhouette(k6$cluster, dist(cluster_df, "euclidean")))
```



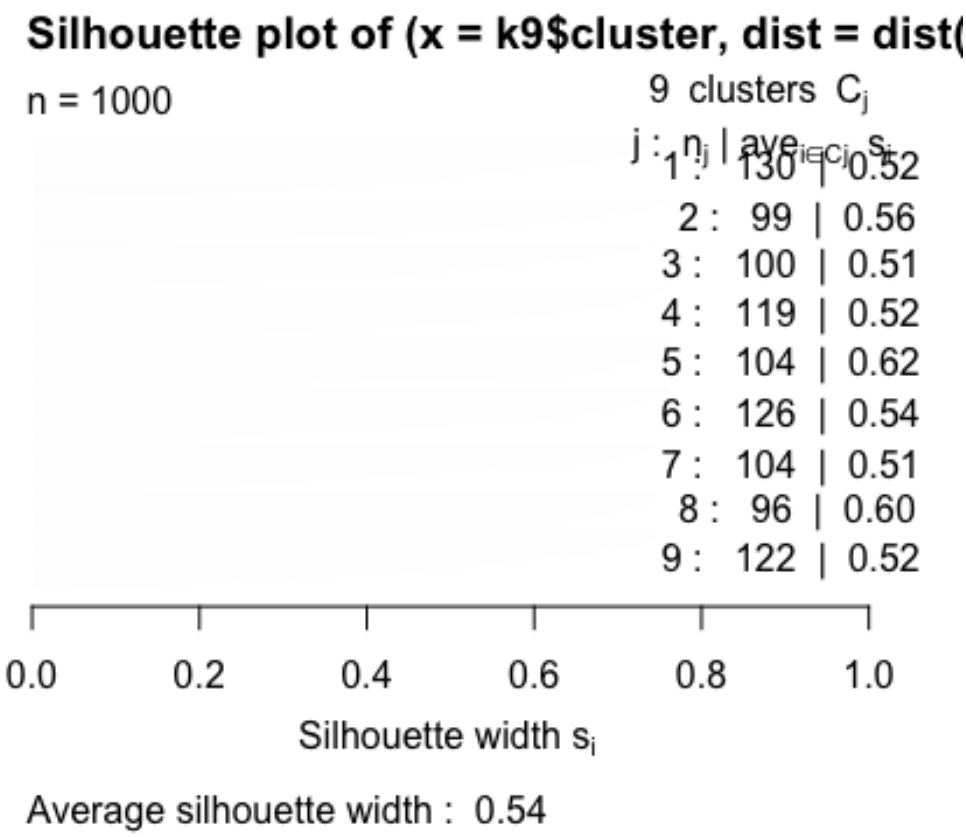

```r
k7 <- kmeans(cluster_df, 7, iter.max=100,nstart=50,algorithm="Lloyd")
s7 <- plot(silhouette(k7$cluster, dist(cluster_df, "euclidean")))
```



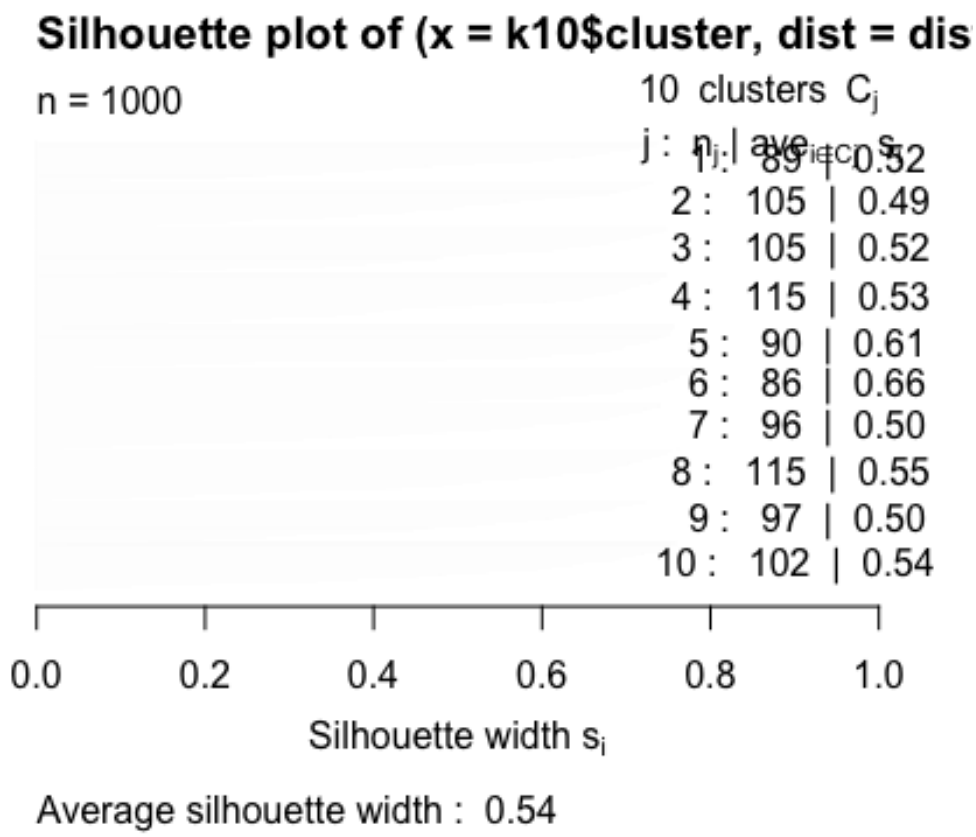

```r
k8 <- kmeans(cluster_df, 8, iter.max=100,nstart=50,algorithm="Lloyd")
s8 <- plot(silhouette(k8$cluster, dist(cluster_df, "euclidean")))
```



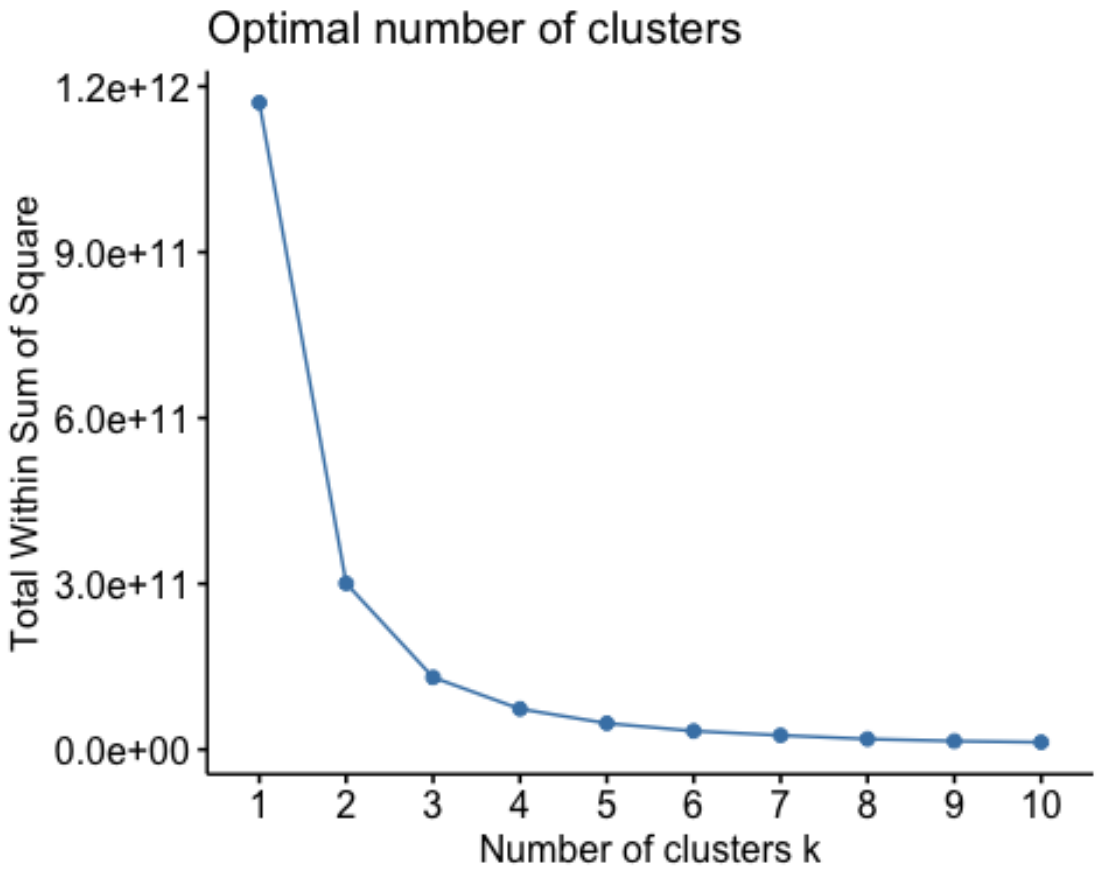

```r
k9 <- kmeans(cluster_df, 9, iter.max=100,nstart=50,algorithm="Lloyd")
s9 <- plot(silhouette(k9$cluster, dist(cluster_df, "euclidean")))
```



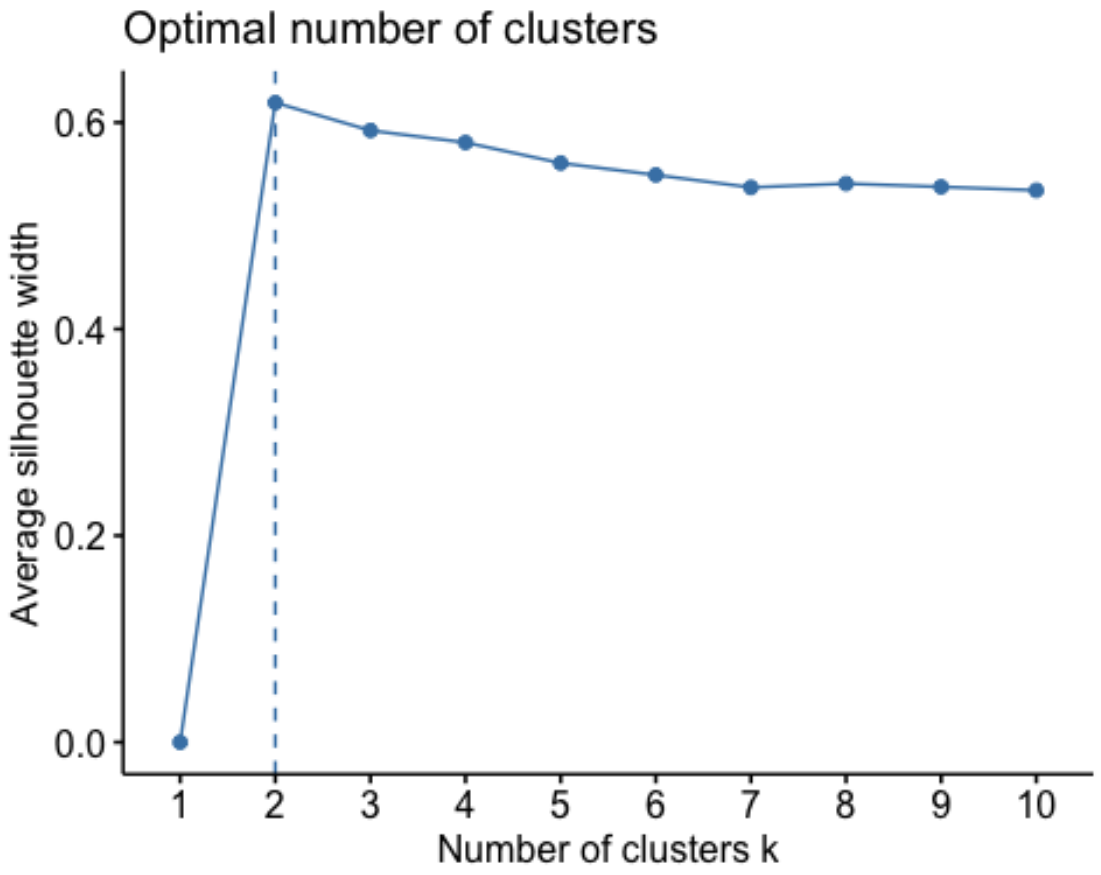

```r
k10 <- kmeans(cluster_df, 10, iter.max=100,nstart=50,algorithm="Lloyd")
s10 <- plot(silhouette(k10$cluster, dist(cluster_df, "euclidean")))
```



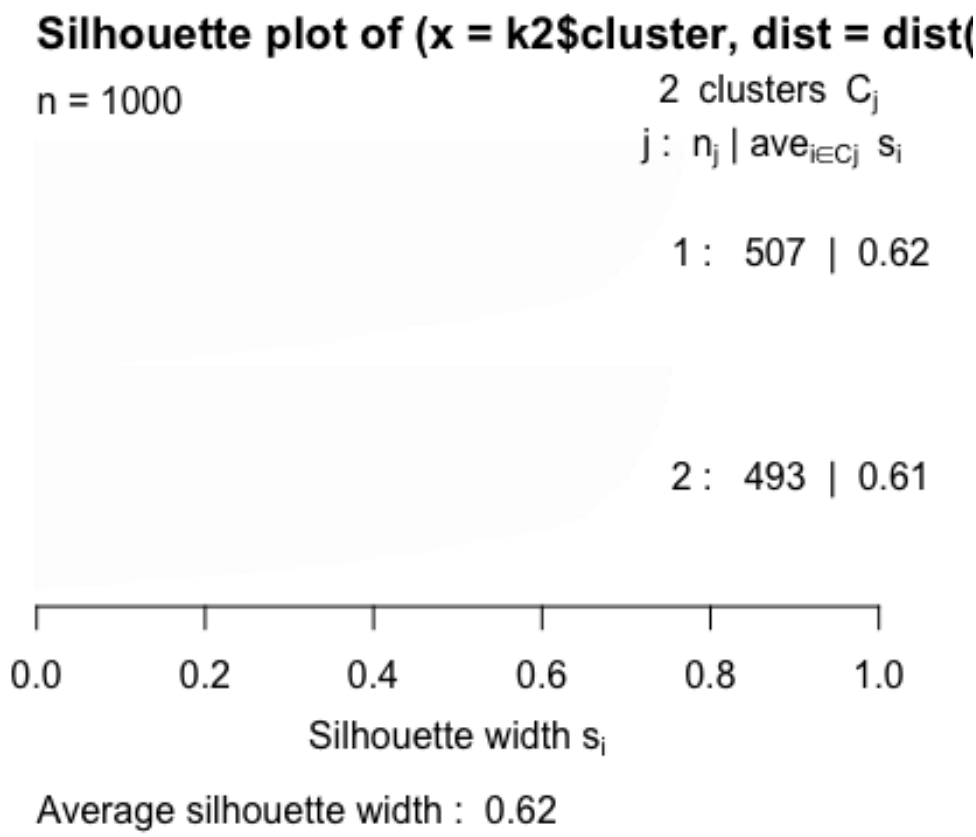

```
library(ggplot2)
library(NbClust)
library(factoextra)

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

library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

fviz_nbclust(cluster_df, kmeans, method = "wss")
```




```
fviz_nbclust(cluster_df, kmeans, method = "silhouette")
```




```
### Gap Statistic Method


k2 <- kmeans(cluster_df, 2, iter.max=100,nstart=50,algorithm="Lloyd")
s2 <- plot(silhouette(k2$cluster, dist(cluster_df, "euclidean")))
```



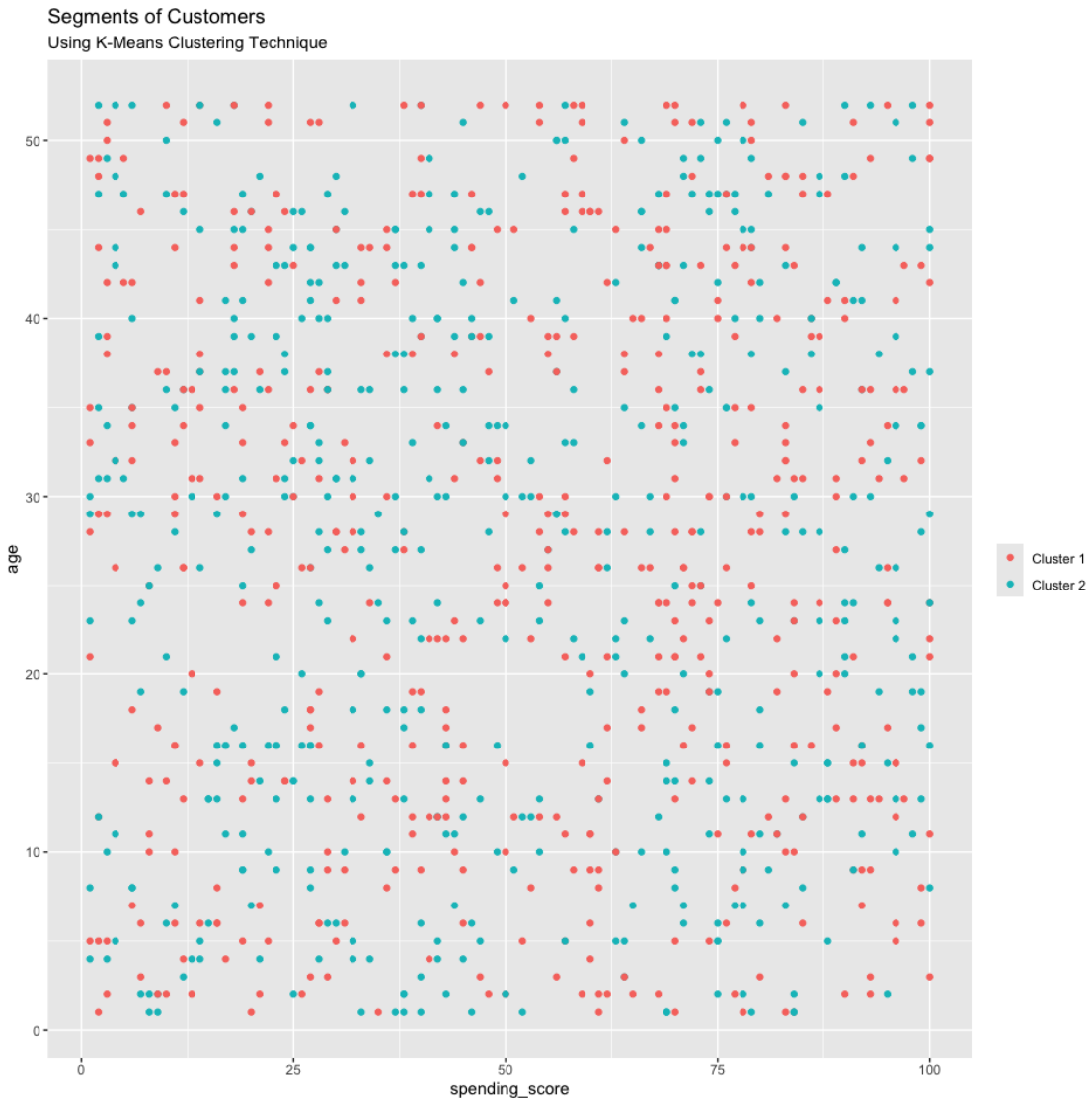

```
#### Principle Cluster Analysis


cluster_df$gender <-  as.numeric(as.factor(cluster_df$gender))
cluster_df$income<-  as.numeric(as.factor(cluster_df$income))
cluster_df$spending_score <-
as.numeric(as.factor(cluster_df$spending_score))
cluster_df$membership_years <-
as.numeric(as.factor(cluster_df$membership_years))
cluster_df$purchase_frequency <-
as.numeric(as.factor(cluster_df$purchase_frequency))
cluster_df$age <-  as.numeric(as.factor(cluster_df$age))
cluster_df$preferred_category <-
as.numeric(as.factor(cluster_df$membership_years))


pcclust <- prcomp(cluster_df)

summary(pcclust)

## Importance of components:
##                           PC1      PC2      PC3      PC4      PC5
PC6
```

```
## Standard deviation     300.6385 282.8503 28.95435 15.07344 14.17479
4.02324
## Proportion of Variance   0.5265   0.4660  0.00488  0.00132  0.00117
0.00009
## Cumulative Proportion    0.5265   0.9925  0.99741  0.99873  0.99990
1.00000
##                            PC7       PC8
## Standard deviation     0.7985 2.828e-16
## Proportion of Variance 0.0000 0.000e+00
## Cumulative Proportion  1.0000 1.000e+00

pcclust$rotation

##                                  PC1           PC2           PC3
PC4
## age                    2.607660e-03  0.0017654858 -1.317267e-02 -9.629322e-
01
## gender                 1.839258e-04  0.0001482290 -9.140367e-05 -1.390063e-
03
## income                -5.301956e-01  0.8478751478 -1.547707e-04  8.926455e-
05
## spending_score        -1.137952e-03 -0.0008940636 -9.998951e-01  1.145524e-
02
## membership_years       6.606343e-05 -0.0004008535 -2.675151e-03  4.238224e-
03
## purchase_frequency     9.519659e-04  0.0007083614 -4.468859e-03  2.694163e-
01
## preferred_category     6.606343e-05 -0.0004008535 -2.675151e-03  4.238224e-
03
## last_purchase_amount   8.478700e-01  0.5301914488 -1.392805e-03  2.729881e-
03
##                                  PC5           PC6           PC7
PC8
## age                   -0.2693985906  0.0001554523 -1.612163e-03
0.000000e+00
## gender                -0.0010050193  0.0057998828  9.999817e-01 -3.858849e-
15
## income                 0.0001134268  0.0005313637 -3.102146e-05 -9.654210e-
19
## spending_score         0.0079273694 -0.0036895633 -4.576340e-05  7.492871e-
17
## membership_years      -0.0145876809  0.7069264676 -4.109133e-03 -7.071068e-
01
## purchase_frequency    -0.9627725370 -0.0215017078 -4.690893e-04 -7.878378e-
17
## preferred_category    -0.0145876809  0.7069264676 -4.109133e-03  7.071068e-
01
## last_purchase_amount   0.0019935819  0.0002395662 -2.302576e-04  6.424736e-
19
```

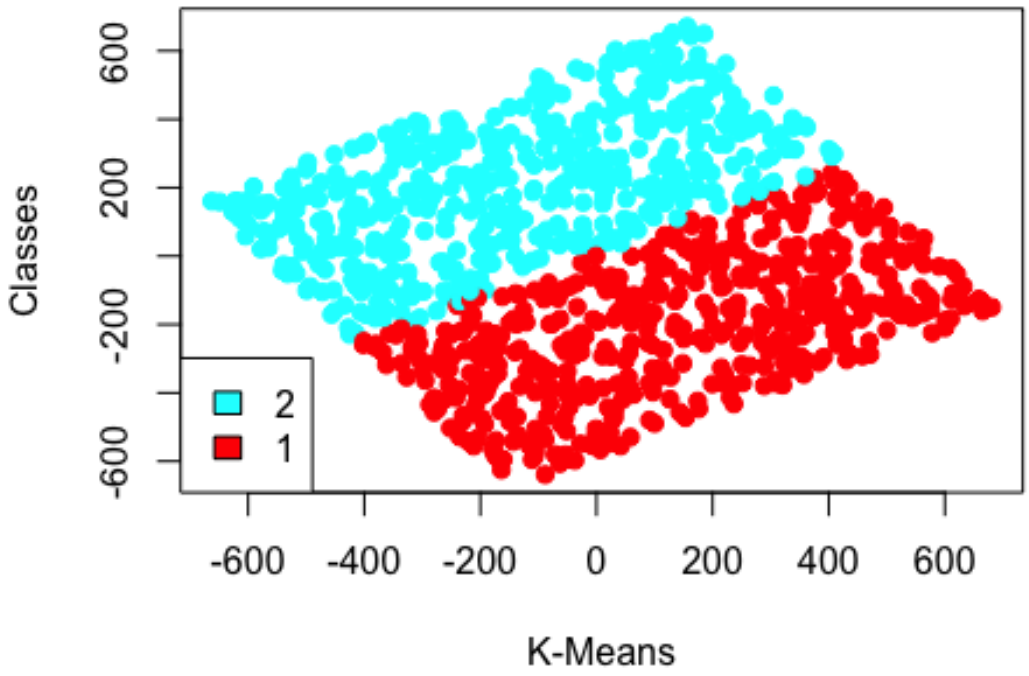
```r
set.seed(42)

glimpse(cluster_df)
```

```
## Rows: 1,000
## Columns: 8
## $ age                  <dbl> 21, 4, 43, 23, 48, 14, 2, 26, 36, 38, 6, 51,
12, …
## $ gender               <dbl> 1, 1, 1, 3, 1, 3, 3, 2, 2, 1, 3, 3, 1, 3, 2,
2, 2…
## $ income               <dbl> 592, 410, 819, 143, 921, 241, 216, 669, 38,
132, …
## $ spending_score       <dbl> 90, 60, 30, 74, 21, 24, 68, 94, 29, 55, 16,
91, 8…
## $ membership_years     <dbl> 3, 2, 2, 9, 3, 3, 5, 9, 6, 7, 7, 1, 3, 4, 5,
6, 5…
## $ purchase_frequency   <dbl> 24, 42, 28, 5, 25, 30, 43, 27, 7, 2, 24, 49,
27, …
## $ preferred_category   <dbl> 3, 2, 2, 9, 3, 3, 5, 9, 6, 7, 7, 1, 3, 4, 5,
6, 5…
## $ last_purchase_amount <dbl> 113.53, 41.93, 424.36, 991.93, 347.08, 86.85,
191…
```

```r
ggplot(cluster_df, aes(x=spending_score, y=age )) +
  geom_point(stat='identity', aes(color=as.factor(k2$cluster))) +
  scale_color_discrete(name=' ',
                       breaks=c("1","2"),
                       labels=c('Cluster 1', 'Cluster 2')) +
  ggtitle("Segments of Customers", subtitle="Using K-Means Clustering
Technique")
```




```
kcols = function(vec){cols=rainbow(length(unique(vec)))
  return (cols[as.numeric(as.factor(vec))])}

digCluster <- k2$cluster;

dignm <- as.character(digCluster)

plot(pcclust$x, col=kcols(digCluster), pch=19, xlab="K-Means",
ylab="Classes")
legend("bottomleft", unique(dignm), fill=unique(kcols(digCluster)))
```


Classes
600
200
-200
-600
2
1
-600
-400
-200
0
200
400
600
K-Means