1. Preparing the Dataset

In [17]:
```
library(dplyr)
library(Rtsne)
library(cluster)
library(data.table)
library(ggplot2)
```

In [18]:
```
# Set the common exported dataset directory
dataset_dir <- file.path(getwd(), 'dataset', 'exported')
```

In [19]:
```
# Read the repeat dataset. It returns a data.table type
profile_csv <- file.path(dataset_dir, 'profile')
profile_df <- fread(profile_csv)
```

Convert these columns to type factor.

In [20]:
```
profile_df$gender <- factor(profile_df$gender)
profile_df$city_code <- factor(profile_df$city_code)
```

2. Clustering

All codes below are based on this site: https://www.r-bloggers.com/clustering-mixed-data-types-in-r/

2.1. Gower Distance

Use gower distance because it can handle mixed data types.

Why I didn't use log transform:

1. The dataset is from a pivot table. One customer may have bought only from the bath subcategory, but another one bought from kitchen and bags_men. There are a lot of zeros, which is normal for pivoting to get the sales per customer per subcategory. If I drop those zeroes, I would be left with no data at all.
2. The means of each subcategory are near to each other.
3. When I tested with the log transform, the medoids were only sales for `total_bath` . It did not consider that some clusters have sales on

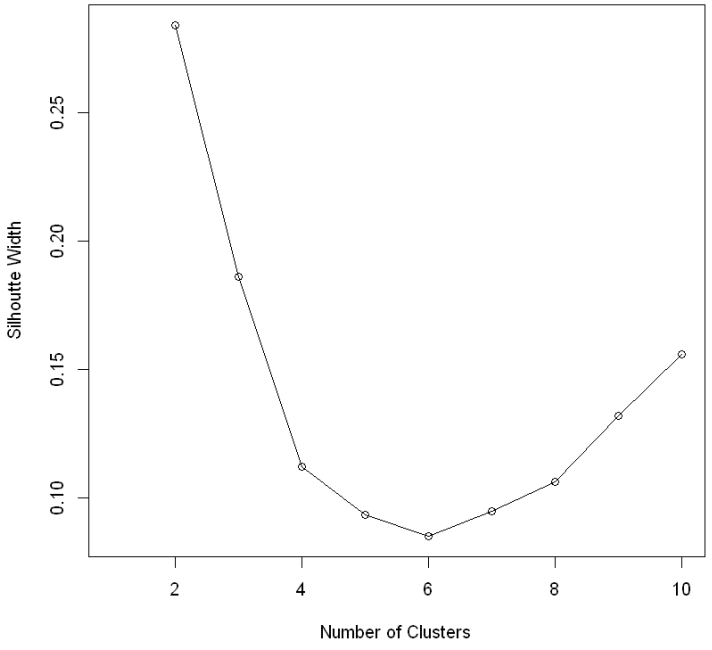
In [21]:
```
gower_dist <- daisy(profile_df[, -c('customer_id')],
                    metric = 'gower')
```

2.2. Silhoutte Width

Use the silhoutte width to choose the final number of clusters to use.

In [22]:
```
# For k ranging from 2 to 10, use PAM
sil_width <- c(NA)
for (i in 2:10) {
    pam_fit <- pam(gower_dist, diss=TRUE, k=i)
    sil_width[i] <- pam_fit$silinfo$avg.width
}
```

In [23]:
```
# Plot
plot(1:10, sil_width,
     xlab = "Number of Clusters",
     ylab = "Silhoutte Width")
lines(1:10, sil_width)
```



Interpretation of the Silhoutte Width

Two clusters would output the best result followed by 3 clusters. However, having 10 clusters also produced a good measurement.

With this, I decided to **go with 10 clusters**. In the e-shop marketing view, grouping the customers into many small clusters is viable. Since this is an e-shop, recommendations in the site are not as costly as wide-reach marketing campaigns like television commercials, billboards, etc. If we are going with onsite recommendations, having 10 clusters would mean 10 targeted marketing campaigns with sure results.

2.3. Partitioning around Medoids

Do PAM with 10 clusters. Print out the summary and medoids of each cluster.

In [24]:
```
# Set the seed at a random number (in this case, 7) to always reproduce the result
set.seed(7)
```

In [25]:
```
pam_fit <- pam(gower_dist, diss = TRUE, k = 10)
```

In [26]:
```
pam_results <- profile_df %>%
    dplyr::select(-c(customer_id)) %>%
    mutate(cluster = pam_fit$clustering) %>%
    group_by(cluster) %>%
    do(the_summary = summary(.))
```

In [27]:
```
pam_results$the_summary
```

```
[[1]]
 total_bags_men   total_bags_women   total_bath     total_clothing_kids
 Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0   Median :   0.0   Median :   0.0   Median :   0.0
 Mean   : 470.6   Mean   : 335.7   Mean   : 406.5   Mean   : 348.4
 3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :8121.8   Max.   :6856.5   Max.   :7602.4   Max.   :8141.6

 total_clothing_men total_clothing_women total_furnishing total_kitchen
 Min.   :   0.0     Min.   :   0.0       Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0     Median :   0.0       Median :   0.0   Median :   0.0
 Mean   : 392.3     Mean   : 428.7       Mean   : 378.4   Mean   : 441.5
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :7933.9     Max.   :7083.1       Max.   :7188.0   Max.   :8099.6

  total_tools     gender    city_code     birth_year      cluster
 Min.   :   0.0   F:289   5      :127   Min.   :1970   Min.   :1
 1st Qu.:   0.0   M:  0   2      : 37   1st Qu.:1972   1st Qu.:1
 Median :   0.0           8      : 36   Median :1975   Median :1
 Mean   : 435.5           3      : 31   Mean   :1977   Mean   :1
 3rd Qu.:   0.0           4      : 30   3rd Qu.:1979   3rd Qu.:1
 Max.   :7939.4           6      : 28   Max.   :1992   Max.   :1
                         (Other):  0

[[2]]
 total_bags_men   total_bags_women   total_bath     total_clothing_kids
 Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0   Median :   0.0   Median :   0.0   Median :   0.0
 Mean   : 271.7   Mean   : 309.6   Mean   : 383.4   Mean   : 312.7
 3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :8121.8   Max.   :6577.0   Max.   :6280.8   Max.   :6563.7

 total_clothing_men total_clothing_women total_furnishing total_kitchen
 Min.   :   0.0     Min.   :   0.0       Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0     Median :   0.0       Median :   0.0   Median :   0.0
 Mean   : 289.3     Mean   : 418.9       Mean   : 413.5   Mean   : 348.6
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :8210.1     Max.   :7762.6       Max.   :6547.1   Max.   :6889.7

  total_tools     gender    city_code     birth_year      cluster
 Min.   :   0.0   F: 12   4      :142   Min.   :1970   Min.   :2
 1st Qu.:   0.0   M:165   1      : 11   1st Qu.:1975   1st Qu.:2
 Median :   0.0           7      :  9   Median :1977   Median :2
 Mean   : 524.5           6      :  6   Mean   :1979   Mean   :2
 3rd Qu.:   0.0           9      :  6   3rd Qu.:1983   3rd Qu.:2
 Max.   :7143.8           10     :  3   Max.   :1992   Max.   :2
                         (Other):  0

[[3]]
 total_bags_men   total_bags_women   total_bath     total_clothing_kids
 Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0   Median :   0.0   Median :   0.0   Median :   0.0
 Mean   : 295.7   Mean   : 415.1   Mean   : 419.5   Mean   : 442.6
 3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :5852.1   Max.   :8265.4   Max.   :8215.7   Max.   :7596.9

 total_clothing_men total_clothing_women total_furnishing total_kitchen
 Min.   :   0.0     Min.   :   0.0       Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0.0   1st Qu.:   0.0
```

```
 Median :   0.0     Median :   0.0       Median :   0.0   Median :   0.0
 Mean   : 402.5     Mean   : 265.4       Mean   : 309.8   Mean   : 263.7
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :7630.0     Max.   :8265.4       Max.   :5513.9   Max.   :7160.4

  total_tools     gender    city_code     birth_year      cluster
 Min.   :   0.0   F:  0   5      :142   Min.   :1970   Min.   :3
 1st Qu.:   0.0   M:275   6      : 30   1st Qu.:1972   1st Qu.:3
 Median :   0.0           10     : 28   Median :1974   Median :3
 Mean   : 378.9           1      : 27   Mean   :1976   Mean   :3
 3rd Qu.:   0.0           7      : 24   3rd Qu.:1980   3rd Qu.:3
 Max.   :7994.7           9      : 24   Max.   :1992   Max.   :3
                         (Other):  0

[[4]]
 total_bags_men   total_bags_women    total_bath      total_clothing_kids
 Min.   :   0.0   Min.   :    0.0   Min.   :    0.0   Min.   :   0.0
 1st Qu.:   0.0   1st Qu.:    0.0   1st Qu.:    0.0   1st Qu.:   0.0
 Median :   0.0   Median :    0.0   Median :    0.0   Median :   0.0
 Mean   : 239.2   Mean   :  388.5   Mean   :  433.1   Mean   : 448.3
 3rd Qu.:   0.0   3rd Qu.:    0.0   3rd Qu.:    0.0   3rd Qu.:   0.0
 Max.   :7939.4   Max.   :11418.0   Max.   :10290.9   Max.   :9632.3

 total_clothing_men total_clothing_women total_furnishing total_kitchen
 Min.   :   0.0     Min.   :   0.0       Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0     Median :   0.0       Median :   0.0   Median :   0.0
 Mean   : 371.4     Mean   : 307.4       Mean   : 447.5   Mean   : 489.1
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :8243.3     Max.   :7591.4       Max.   :8044.4   Max.   :7828.9

  total_tools     gender    city_code     birth_year      cluster
 Min.   :   0.0   F:274   7      :149   Min.   :1970   Min.   :4
 1st Qu.:   0.0   M: 23   6      : 41   1st Qu.:1983   1st Qu.:4
 Median :   0.0           4      : 40   Median :1988   Median :4
 Mean   : 440.4           3      : 25   Mean   :1986   Mean   :4
 3rd Qu.:   0.0           8      : 23   3rd Qu.:1990   3rd Qu.:4
 Max.   :8132.8           2      : 19   Max.   :1992   Max.   :4
                         (Other):  0

[[5]]
 total_bags_men   total_bags_women   total_bath     total_clothing_kids
 Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0   Median :   0.0   Median :   0.0   Median :   0.0
 Mean   : 322.2   Mean   : 361.3   Mean   : 227.5   Mean   : 333.7
 3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :6501.8   Max.   :7746.1   Max.   :5317.3   Max.   :6378.1

 total_clothing_men total_clothing_women total_furnishing total_kitchen
 Min.   :   0.0     Min.   :   0.0       Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0     Median :   0.0       Median :   0.0   Median :   0.0
 Mean   : 356.7     Mean   : 331.3       Mean   : 376.1   Mean   : 370.1
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :7542.7     Max.   :7469.8       Max.   :5552.6   Max.   :7116.2

  total_tools     gender    city_code     birth_year      cluster
 Min.   :   0.0   F:187   1      :139   Min.   :1970   Min.   :5
 1st Qu.:   0.0   M: 15   6      : 22   1st Qu.:1979   1st Qu.:5
 Median :   0.0           2      : 16   Median :1980   Median :5
 Mean   : 486.5           3      : 13   Mean   :1981   Mean   :5
 3rd Qu.:   0.0           4      :  8   3rd Qu.:1983   3rd Qu.:5
 Max.   :8011.2           8      :  4   Max.   :1992   Max.   :5
```

```
                        (Other):  0

[[6]]
 total_bags_men   total_bags_women   total_bath     total_clothing_kids
 Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0   Median :   0.0   Median :   0.0   Median :   0.0
 Mean   : 457.1   Mean   : 424.2   Mean   : 325.9   Mean   : 405.5
 3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :7983.6   Max.   :8202.4   Max.   :6563.7   Max.   :9285.3

 total_clothing_men total_clothing_women total_furnishing total_kitchen
 Min.   :   0.0     Min.   :   0.0       Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0     Median :   0.0       Median :   0.0   Median :   0.0
 Mean   : 545.1     Mean   : 270.2       Mean   : 326.9   Mean   : 419.5
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :6986.9     Max.   :6458.7       Max.   :8099.6   Max.   :7480.9

  total_tools     gender    city_code     birth_year      cluster
 Min.   :   0.0   F: 18   2      :138   Min.   :1970   Min.   :6
 1st Qu.:   0.0   M:200   7      : 23   1st Qu.:1980   1st Qu.:6
 Median :   0.0           9      : 18   Median :1983   Median :6
 Mean   : 292.9           6      : 15   Mean   :1982   Mean   :6
 3rd Qu.:   0.0           10     : 14   3rd Qu.:1985   3rd Qu.:6
 Max.   :8204.6           1      : 10   Max.   :1992   Max.   :6
                         (Other):  0

[[7]]
 total_bags_men   total_bags_women   total_bath     total_clothing_kids
 Min.   :   0.0   Min.   :   0     Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0   1st Qu.:   0     1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0   Median :   0     Median :   0.0   Median :   0.0
 Mean   : 423.9   Mean   : 396     Mean   : 353.6   Mean   : 282.5
 3rd Qu.:   0.0   3rd Qu.:   0     3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :7221.2   Max.   :7807     Max.   :6646.6   Max.   :6663.1

 total_clothing_men total_clothing_women total_furnishing total_kitchen
 Min.   :   0.0     Min.   :   0.0       Min.   :   0     Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0     1st Qu.:   0.0
 Median :   0.0     Median :   0.0       Median :   0     Median :   0.0
 Mean   : 353.3     Mean   : 412.7       Mean   : 351     Mean   : 223.9
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0     3rd Qu.:   0.0
 Max.   :7099.6     Max.   :7939.4       Max.   :7774     Max.   :6204.6

  total_tools   gender    city_code     birth_year      cluster
 Min.   :   0   F:191   10     :143   Min.   :1970   Min.   :7
 1st Qu.:   0   M: 19   6      : 19   1st Qu.:1979   1st Qu.:7
 Median :   0           4      : 18   Median :1982   Median :7
 Mean   : 392           3      : 13   Mean   :1982   Mean   :7
 3rd Qu.:   0           8      : 13   3rd Qu.:1985   3rd Qu.:7
 Max.   :7591           2      :  4   Max.   :1992   Max.   :7
                       (Other):  0

[[8]]
 total_bags_men   total_bags_women   total_bath     total_clothing_kids
 Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0   Median :   0.0   Median :   0.0   Median :   0.0
 Mean   : 522.8   Mean   : 374.1   Mean   : 423.9   Mean   : 616.0
 3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.: 468.5
 Max.   :8370.4   Max.   :7743.8   Max.   :5644.3   Max.   :6740.5

 total_clothing_men total_clothing_women total_furnishing total_kitchen
```

```
 Min.   :   0.0     Min.   :   0.0       Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0     Median :   0.0       Median :   0.0   Median :   0.0
 Mean   : 524.4     Mean   : 221.6       Mean   : 379.5   Mean   : 249.8
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :8324.0     Max.   :6263.1       Max.   :7585.8   Max.   :8215.7

  total_tools   gender    city_code     birth_year      cluster
 Min.   :   0   F:143   9      :142   Min.   :1970   Min.   :8
 1st Qu.:   0   M: 17   3      :  5   1st Qu.:1977   1st Qu.:8
 Median :   0           6      :  4   Median :1981   Median :8
 Mean   : 314           8      :  4   Mean   :1981   Mean   :8
 3rd Qu.:   0           4      :  3   3rd Qu.:1984   3rd Qu.:8
 Max.   :6591           2      :  2   Max.   :1992   Max.   :8
                        (Other):  0

[[9]]
 total_bags_men   total_bags_women   total_bath     total_clothing_kids
 Min.   :   0.0   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0   Median :   0.0   Median :   0.0   Median :   0.0
 Mean   : 236.3   Mean   : 379.6   Mean   : 265.2   Mean   : 251.9
 3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :7072.0   Max.   :7226.7   Max.   :6510.7   Max.   :6493.0

 total_clothing_men total_clothing_women total_furnishing total_kitchen
 Min.   :   0.0     Min.   :   0.0       Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0     Median :   0.0       Median :   0.0   Median :   0.0
 Mean   : 549.5     Mean   : 433.5       Mean   : 217.5   Mean   : 408.2
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :8248.8     Max.   :9503.0       Max.   :5184.7   Max.   :6193.5

  total_tools   gender    city_code     birth_year      cluster
 Min.   :   0   F: 19   8      :154   Min.   :1970   Min.   :9
 1st Qu.:   0   M:207   6      : 25   1st Qu.:1977   1st Qu.:9
 Median :   0           7      : 17   Median :1979   Median :9
 Mean   : 407           10     : 13   Mean   :1980   Mean   :9
 3rd Qu.:   0           1      : 11   3rd Qu.:1984   3rd Qu.:9
 Max.   :8227           9      :  6   Max.   :1992   Max.   :9
                        (Other):  0

[[10]]
 total_bags_men   total_bags_women   total_bath     total_clothing_kids
 Min.   :   0.0   Min.   :   0     Min.   :   0.0   Min.   :   0
 1st Qu.:   0.0   1st Qu.:   0     1st Qu.:   0.0   1st Qu.:   0
 Median :   0.0   Median :   0     Median :   0.0   Median :   0
 Mean   : 462.8   Mean   : 259     Mean   : 366.2   Mean   : 301
 3rd Qu.:   0.0   3rd Qu.:   0     3rd Qu.:   0.0   3rd Qu.:   0
 Max.   :7558.2   Max.   :7564     Max.   :6823.4   Max.   :8111

 total_clothing_men total_clothing_women total_furnishing total_kitchen
 Min.   :   0.0     Min.   :   0.0       Min.   :   0.0   Min.   :   0.0
 1st Qu.:   0.0     1st Qu.:   0.0       1st Qu.:   0.0   1st Qu.:   0.0
 Median :   0.0     Median :   0.0       Median :   0.0   Median :   0.0
 Mean   : 331.9     Mean   : 378.7       Mean   : 447.9   Mean   : 380.4
 3rd Qu.:   0.0     3rd Qu.:   0.0       3rd Qu.:   0.0   3rd Qu.:   0.0
 Max.   :7668.7     Max.   :7171.4       Max.   :7420.1   Max.   :7574.8

  total_tools     gender    city_code     birth_year      cluster
 Min.   :   0.0   F: 21   3      :150   Min.   :1970   Min.   :10
 1st Qu.:   0.0   M:301   1      : 48   1st Qu.:1985   1st Qu.:10
 Median :   0.0           10     : 40   Median :1987   Median :10
 Mean   : 510.4           6      : 37   Mean   :1986   Mean   :10
```

```
3rd Qu.:   0.0          9      : 32   3rd Qu.:1990   3rd Qu.:10
Max.   :8623.4          7      : 15   Max.   :1992   Max.   :10
                        (Other):  0
```

In [28]:
```
profile_df[pam_fit$medoids, ]
```

A data.table: 10 × 13

customer_id	total_bags_men	total_bags_women	total_bath	total_clothing_kids	total_clothing_men	total_clothi
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
269430	0.000	0.000	0.00	0.00	110.5	
269136	0.000	0.000	0.00	0.00	0.0	
271969	0.000	0.000	0.00	77.35	0.0	
270344	0.000	0.000	320.45	0.00	0.0	
272759	0.000	0.000	0.00	0.00	0.0	
275049	0.000	184.535	0.00	0.00	0.0	
271405	98.345	0.000	0.00	0.00	0.0	
272047	0.000	0.000	0.00	384.54	0.0	
267222	0.000	0.000	0.00	0.00	0.0	
275026	0.000	0.000	0.00	0.00	0.0	

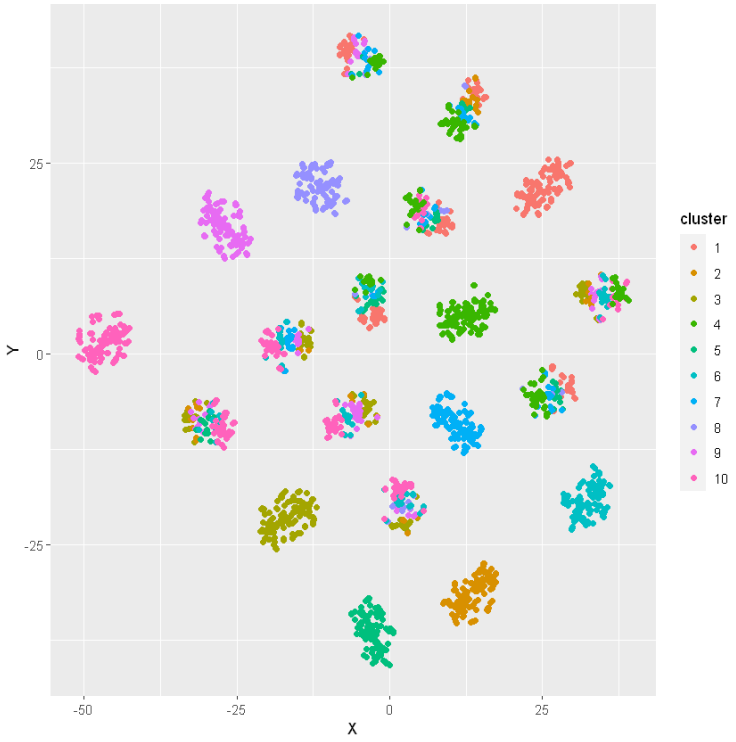
2.4. Plotting the Clusters

Use t-distributed stochastic neighbor embedding to plot the clusters using the gower distance.

In [29]:
```
tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
```

In [30]:
```
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering),
         customer_id = profile_df$customer_id)
```

In [31]:
```
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster))
```



2.5. Exporting the Cluster Results

Export the cluster results in CSV format. The file will then be read in the analysis python notebook.

In [32]:
```
write.csv(tsne_data, file.path(dataset_dir, 'profile_tsne_data'), row.names=TRUE)
```