final

Hannah Cronin

2022-12-18

```
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.3.6      ✔ dplyr   1.0.10
## ✔ tibble  3.1.8      ✔ stringr 1.4.1
## ✔ tidyr   1.2.1      ✔ forcats 0.5.2
## ✔ purrr   0.3.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WB
a
```

```
library(ISLR)
library(cluster)
diabetes <- read_csv("/Users/hannahcronin/Desktop/diabetes_data.csv")
```

```
## Rows: 70692 Columns: 18
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (18): Age, Sex, HighChol, CholCheck, BMI, Smoker, HeartDiseaseorAttack, ...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

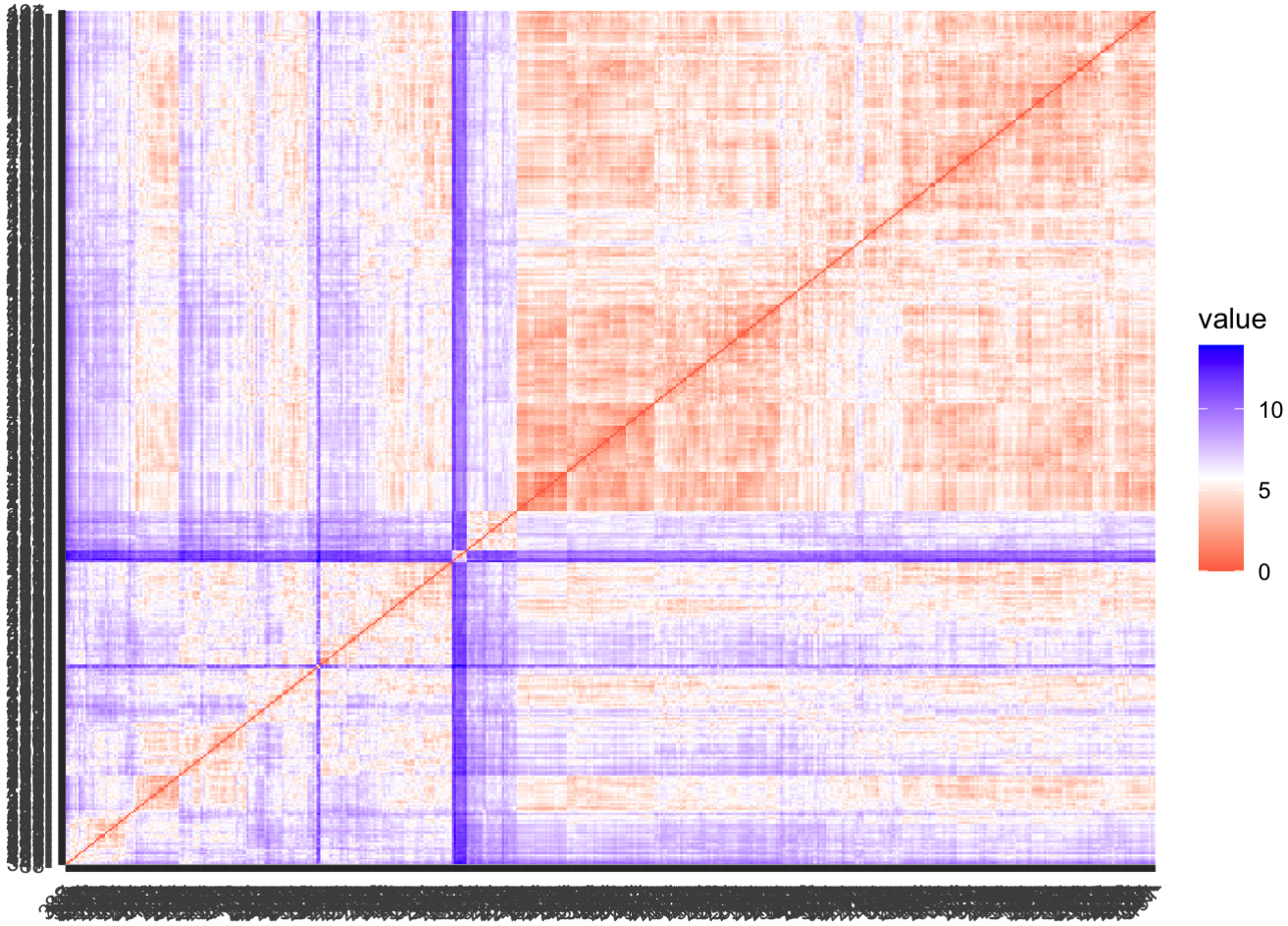
```
set.seed(123)
summary(diabetes) #summary of the 18 variables
```

```
##       Age              Sex           HighChol        CholCheck
##  Min.   : 1.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 7.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.0000
##  Median : 9.000   Median :0.000   Median :1.0000   Median :1.0000
##  Mean   : 8.584   Mean   :0.457   Mean   :0.5257   Mean   :0.9753
##  3rd Qu.:11.000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :13.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##       BMI            Smoker       HeartDiseaseorAttack  PhysActivity
##  Min.   :12.00   Min.   :0.0000   Min.   :0.0000       Min.   :0.000
##  1st Qu.:25.00   1st Qu.:0.0000   1st Qu.:0.0000       1st Qu.:0.000
##  Median :29.00   Median :0.0000   Median :0.0000       Median :1.000
##  Mean   :29.86   Mean   :0.4753   Mean   :0.1478       Mean   :0.703
##  3rd Qu.:33.00   3rd Qu.:1.0000   3rd Qu.:0.0000       3rd Qu.:1.000
##  Max.   :98.00   Max.   :1.0000   Max.   :1.0000       Max.   :1.000
##      Fruits          Veggies       HvyAlcoholConsump    GenHlth
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:2.000
##  Median :1.0000   Median :1.0000   Median :0.00000   Median :3.000
##  Mean   :0.6118   Mean   :0.7888   Mean   :0.04272   Mean   :2.837
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :5.000
##     MentHlth         PhysHlth        DiffWalk          Stroke
##  Min.   : 0.000   Min.   : 0.00   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 0.000   Median : 0.00   Median :0.0000   Median :0.00000
##  Mean   : 3.752   Mean   : 5.81   Mean   :0.2527   Mean   :0.06217
##  3rd Qu.: 2.000   3rd Qu.: 6.00   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :30.000   Max.   :30.00   Max.   :1.0000   Max.   :1.00000
##      HighBP          Diabetes
##  Min.   :0.0000   Min.   :0.0
##  1st Qu.:0.0000   1st Qu.:0.0
##  Median :1.0000   Median :0.5
##  Mean   :0.5635   Mean   :0.5
##  3rd Qu.:1.0000   3rd Qu.:1.0
##  Max.   :1.0000   Max.   :1.0
```

```
data = sample_n(diabetes, 500) #Limited down the dataset for better performance
summary(data)
```

```
##       Age              Sex           HighChol       CholCheck
##  Min.   : 1.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 7.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.000
##  Median : 9.000   Median :0.000   Median :1.000   Median :1.000
##  Mean   : 8.584   Mean   :0.464   Mean   :0.552   Mean   :0.986
##  3rd Qu.:11.000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :13.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
##       BMI            Smoker      HeartDiseaseorAttack  PhysActivity
##  Min.   :17.00   Min.   :0.000   Min.   :0.000        Min.   :0.000
##  1st Qu.:25.00   1st Qu.:0.000   1st Qu.:0.000        1st Qu.:0.000
##  Median :29.00   Median :0.000   Median :0.000        Median :1.000
##  Mean   :30.31   Mean   :0.474   Mean   :0.154        Mean   :0.672
##  3rd Qu.:33.00   3rd Qu.:1.000   3rd Qu.:0.000        3rd Qu.:1.000
##  Max.   :95.00   Max.   :1.000   Max.   :1.000        Max.   :1.000
##      Fruits         Veggies     HvyAlcoholConsump    GenHlth
##  Min.   :0.000   Min.   :0.00   Min.   :0.000     Min.   :1.000
##  1st Qu.:0.000   1st Qu.:1.00   1st Qu.:0.000     1st Qu.:2.000
##  Median :1.000   Median :1.00   Median :0.000     Median :3.000
##  Mean   :0.634   Mean   :0.81   Mean   :0.048     Mean   :2.876
##  3rd Qu.:1.000   3rd Qu.:1.00   3rd Qu.:0.000     3rd Qu.:4.000
##  Max.   :1.000   Max.   :1.00   Max.   :1.000     Max.   :5.000
##     MentHlth          PhysHlth        DiffWalk         Stroke
##  Min.   : 0.000   Min.   : 0.00   Min.   :0.000   Min.   :0.00
##  1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.:0.000   1st Qu.:0.00
##  Median : 0.000   Median : 0.00   Median :0.000   Median :0.00
##  Mean   : 3.896   Mean   : 5.74   Mean   :0.242   Mean   :0.07
##  3rd Qu.: 3.000   3rd Qu.: 7.00   3rd Qu.:0.000   3rd Qu.:0.00
##  Max.   :30.000   Max.   :30.00   Max.   :1.000   Max.   :1.00
##      HighBP          Diabetes
##  Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000
##  Median :1.000   Median :0.000
##  Mean   :0.548   Mean   :0.494
##  3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :1.000   Max.   :1.000
```

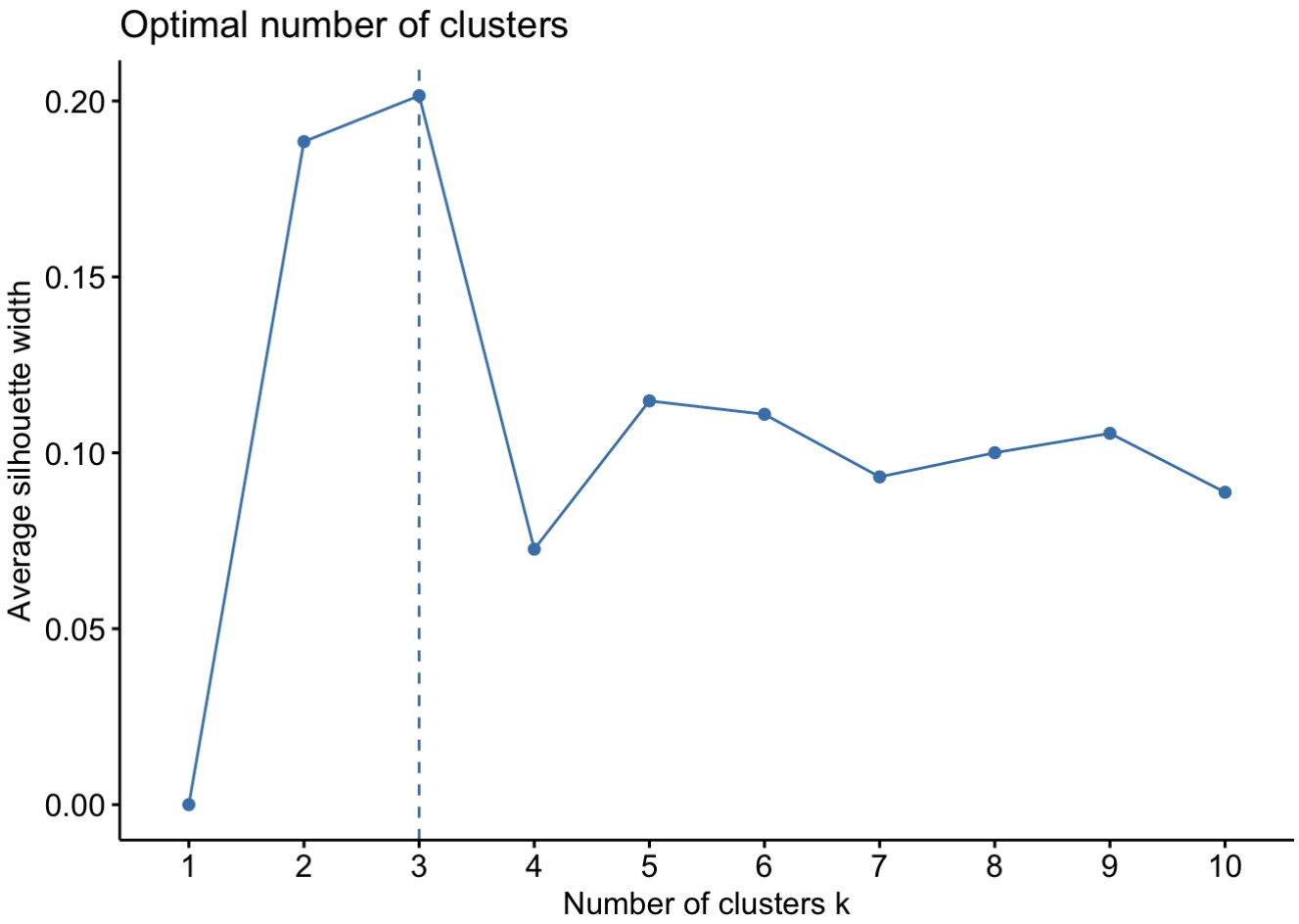
```
df = scale(data)
distance = get_dist(df)
fviz_dist(distance)
```




```
summary(df)
```

```
##       Age                Sex              HighChol         CholCheck
##  Min.   :-2.7384   Min.   :-0.9295   Min.   :-1.109   Min.   :-8.384
##  1st Qu.:-0.5719   1st Qu.:-0.9295   1st Qu.:-1.109   1st Qu.: 0.119
##  Median : 0.1502   Median :-0.9295   Median : 0.900   Median : 0.119
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.000   Mean   : 0.000
##  3rd Qu.: 0.8724   3rd Qu.: 1.0737   3rd Qu.: 0.900   3rd Qu.: 0.119
##  Max.   : 1.5945   Max.   : 1.0737   Max.   : 0.900   Max.   : 0.119
##       BMI               Smoker         HeartDiseaseorAttack  PhysActivity
##  Min.   :-1.6773   Min.   :-0.9483   Min.   :-0.4262      Min.   :-1.4299
##  1st Qu.:-0.6689   1st Qu.:-0.9483   1st Qu.:-0.4262      1st Qu.:-1.4299
##  Median :-0.1646   Median :-0.9483   Median :-0.4262      Median : 0.6979
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000      Mean   : 0.0000
##  3rd Qu.: 0.3396   3rd Qu.: 1.0524   3rd Qu.:-0.4262      3rd Qu.: 0.6979
##  Max.   : 8.1552   Max.   : 1.0524   Max.   : 2.3415      Max.   : 0.6979
##      Fruits           Veggies         HvyAlcoholConsump    GenHlth
##  Min.   :-1.315   Min.   :-2.0627   Min.   :-0.2243   Min.   :-1.6745
##  1st Qu.:-1.315   1st Qu.: 0.4838   1st Qu.:-0.2243   1st Qu.:-0.7819
##  Median : 0.759   Median : 0.4838   Median :-0.2243   Median : 0.1107
##  Mean   : 0.000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.759   3rd Qu.: 0.4838   3rd Qu.:-0.2243   3rd Qu.: 1.0033
##  Max.   : 0.759   Max.   : 0.4838   Max.   : 4.4490   Max.   : 1.8959
##     MentHlth           PhysHlth           DiffWalk           Stroke
##  Min.   :-0.4794   Min.   :-0.5803   Min.   :-0.5645   Min.   :-0.2741
##  1st Qu.:-0.4794   1st Qu.:-0.5803   1st Qu.:-0.5645   1st Qu.:-0.2741
##  Median :-0.4794   Median :-0.5803   Median :-0.5645   Median :-0.2741
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.1103   3rd Qu.: 0.1274   3rd Qu.:-0.5645   3rd Qu.:-0.2741
##  Max.   : 3.2123   Max.   : 2.4528   Max.   : 1.7680   Max.   : 3.6413
##      HighBP            Diabetes
##  Min.   :-1.1000   Min.   :-0.9871
##  1st Qu.:-1.1000   1st Qu.:-0.9871
##  Median : 0.9073   Median :-0.9871
##  Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.9073   3rd Qu.: 1.0111
##  Max.   : 0.9073   Max.   : 1.0111
```

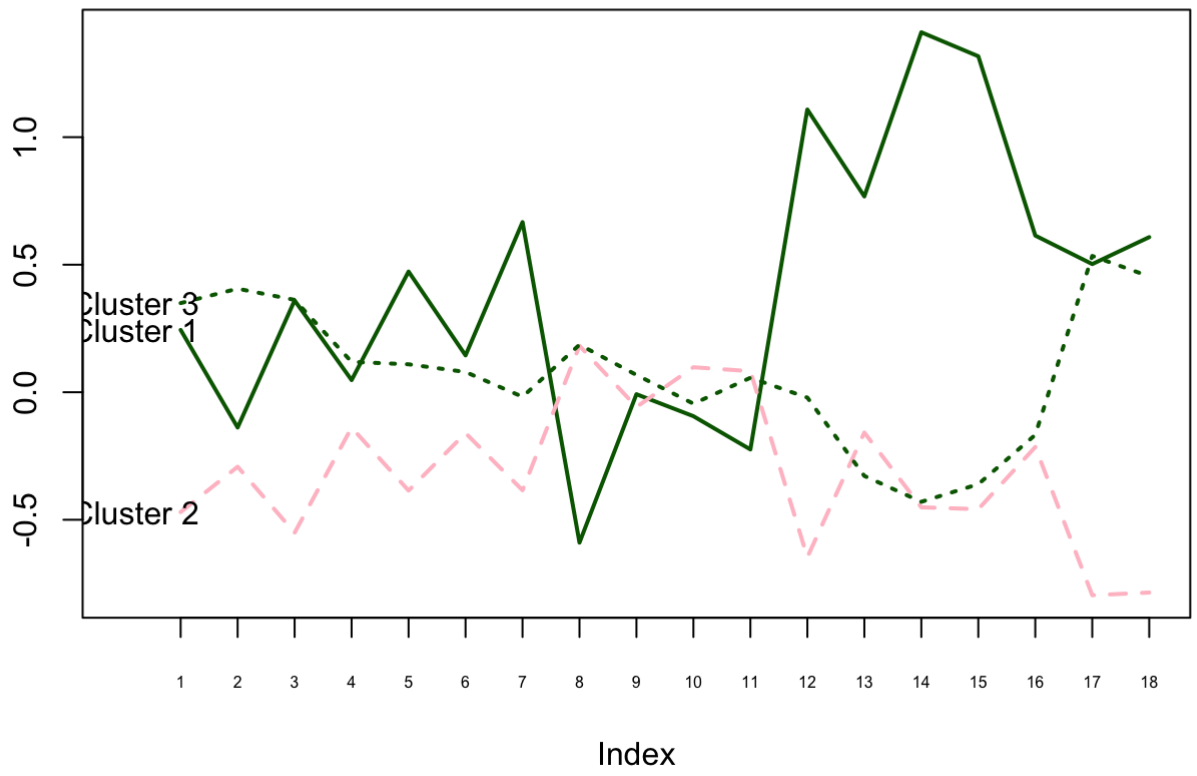
```
fviz_nbclust(df, kmeans, method = "silhouette")
```




```
km = kmeans(df, 3, nstart = 25)
km$cluster
```

```
##   [1] 1 3 3 2 1 1 3 3 1 3 3 1 1 2 3 1 2 1 3 2 3 3 3 2 2 2 3 3 1 3 3 2 1 2 1 2 2
##  [38] 2 3 3 3 3 2 3 2 2 2 2 2 1 2 2 2 1 3 1 3 3 1 3 2 1 3 3 1 1 3 3 2 3 2 2 1 2
##  [75] 3 2 2 3 3 2 1 2 3 1 3 3 2 1 1 3 3 2 3 1 2 3 3 2 2 3 3 2 3 2 3 2 3 1 3 3 2
## [112] 2 3 1 2 1 1 2 3 3 1 1 3 3 2 3 2 2 1 2 2 3 3 3 2 2 3 3 2 3 1 2 1 1 1 1 2 1
## [149] 2 3 2 2 3 1 3 3 3 2 1 3 1 2 1 3 2 1 2 1 3 3 3 3 2 3 3 3 2 3 3 3 3 3 1 2 2
## [186] 3 2 3 2 3 2 1 3 1 3 2 2 2 1 2 3 1 2 2 3 3 2 1 3 2 2 2 3 1 3 2 3 3 3 2 2 2
## [223] 3 3 1 3 2 3 3 2 2 2 2 1 2 1 1 2 1 1 1 2 1 1 3 2 1 1 3 1 2 2 2 2 3 1 2 2 2
## [260] 2 2 1 1 3 2 2 2 3 3 2 2 1 1 2 3 2 1 3 3 3 2 1 3 3 2 1 2 3 2 2 2 1 3 3 1 3
## [297] 2 1 1 3 3 2 2 2 2 3 1 2 3 2 2 1 2 1 2 1 2 3 3 2 2 2 3 3 1 3 3 2 3 3 3 2 1
## [334] 2 1 2 1 2 1 1 1 1 3 1 3 3 1 2 3 1 2 3 2 2 2 2 2 3 2 2 2 3 3 1 3 3 2 3 2 3
## [371] 3 1 2 2 2 3 1 3 1 2 1 1 2 1 2 2 1 3 2 3 2 2 1 2 2 2 2 3 3 3 2 1 2 1 2 2 2
## [408] 1 1 1 3 1 1 3 2 3 2 2 2 3 3 2 3 3 3 3 1 3 2 3 2 3 1 2 3 3 2 1 1 2 1 1 2 3
## [445] 3 2 3 2 3 1 2 2 2 2 3 3 3 3 2 1 3 3 3 2 1 1 3 3 3 2 2 2 3 2 2 3 1 2 2 3 2
## [482] 2 2 2 3 1 3 1 2 3 2 3 2 3 2 3 2 1 2 2
```

```
plot(c(0), xaxt = 'n', ylab = "", type = "l",
     ylim = c(min(km$centers), max(km$centers)), xlim = c(0, 18))
axis(1, at = c(1:18), labels = names(df), cex.axis=.5)
for (i in c(1:3))
lines(km$centers[i,], lty = i, lwd = 2, col = ifelse(i %in% c(1, 3, 5),
                            "dark green", "pink"))
text(x = 0.2, y = km$centers[, 1], labels = paste("Cluster", c(1:3)))
```




```
km$centers #numerical descriptions of each cluster
```

```
##          Age        Sex   HighChol   CholCheck        BMI     Smoker
## 1  0.2442670 -0.1383048  0.3597775  0.04758719  0.4730734  0.1444879
## 2 -0.4698172 -0.2921027 -0.5508830 -0.13862167 -0.3855515 -0.1601777
## 3  0.3494865  0.4059815  0.3620837  0.11903943  0.1095272  0.0793504
##   HeartDiseaseorAttack PhysActivity      Fruits     Veggies HvyAlcoholConsump
## 1           0.66690051   -0.5895075 -0.00777263 -0.09394280       -0.22431970
## 2          -0.38429158    0.1820935 -0.05794248  0.09800218        0.08251491
## 3          -0.01787665    0.1863218  0.06774620 -0.04494666        0.05659067
##       GenHlth   MentHlth   PhysHlth   DiffWalk     Stroke     HighBP   Diabetes
## 1  1.10828912  0.7676839  1.4111857  1.3172197  0.6142881  0.5024585  0.6080735
## 2 -0.64667223 -0.1581143 -0.4511577 -0.4584437 -0.2147531 -0.7958521 -0.7852500
## 3 -0.02101259 -0.3281298 -0.4295184 -0.3605317 -0.1670993  0.5343506  0.4542008
```

```
km$withinss #numerical descriptions of each cluster
```

```
## [1] 2307.851 2637.065 2214.767
```

```
km$size #numerical descriptions of each cluster
```

```
## [1] 119 198 183
```

Cluster 3 is the most homogenous and is also the second largest.

Cluster 2 younger, contains more females, lower cholesterol, does not get their cholesterol checked as frequently, has a lower BMI, less smokers, less history of Heart disease/attacks, more active, eats less fruits, eats more veggies, drinks more, better general health, less mental/physical illness/injury days, less difficulty walking, less stroke, lower blood pressure, and less diabetes.

Alias for Cluster Tw0 is Tend to be Young/In Better Health

Cluster 3 is older, contains more men, has higher cholesterol, tends to get their cholesterol checked, a mid-range BMI, some smokers, lesser history of heart dieases/attack, more physical days, eats

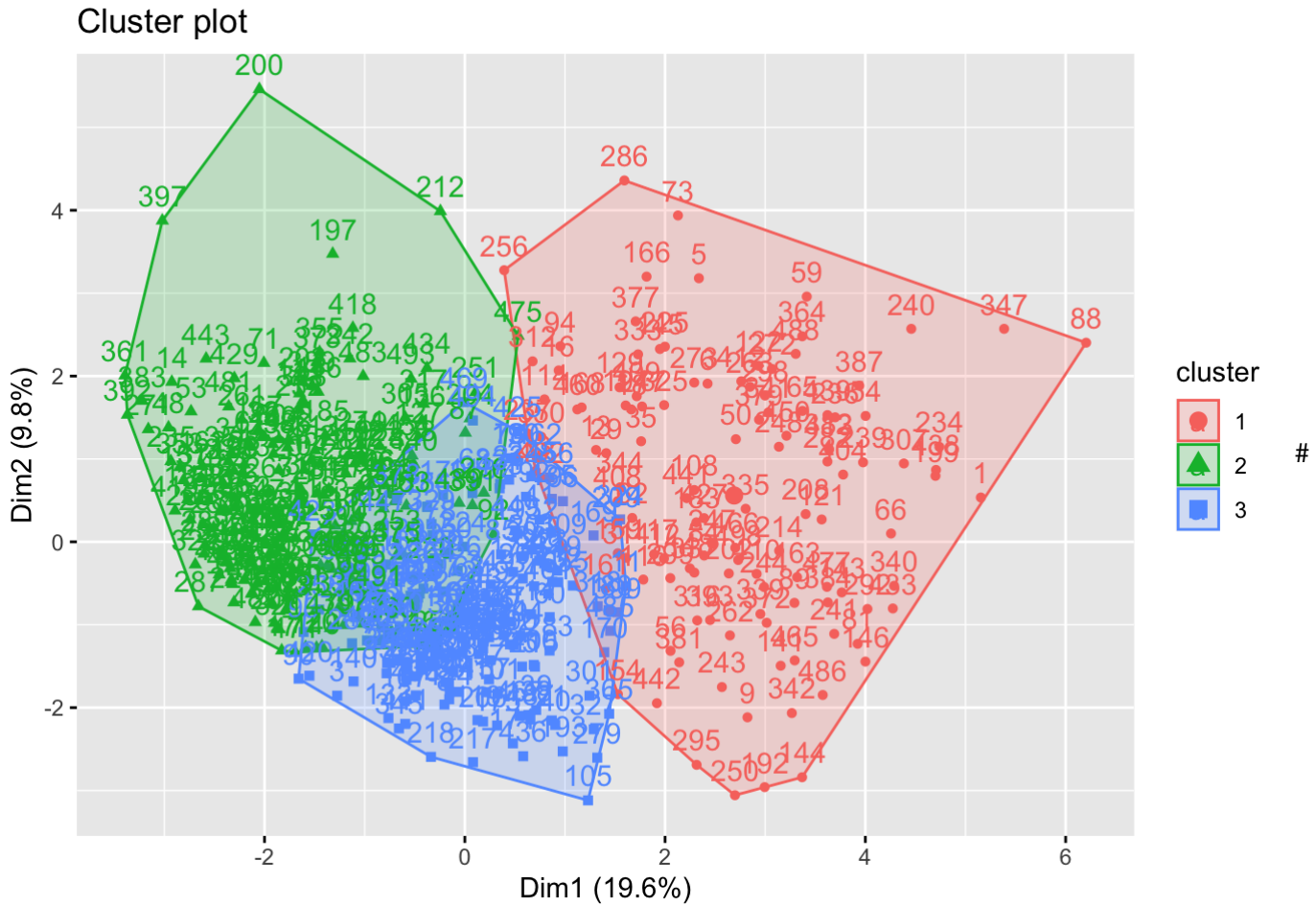
more fruits, eats less veggies, tends to drink in heavier amounts, general health is average, has some poor mental health days, has some injury/illness days, has less difficulty walking, has some strokes, has high blood pressure, and some diabetes.

Alias for Cluster Three is Tend to be Older/In Moderate Health

Cluster 1 is middle aged (between the other clusters), contains more females, has higher cholesterol, tends to get their cholesterol checked, has the highest BMI, has more smokers, has history of heart disease/attack, less physical activity, eats less fruit, eats less veggies, tend not to be heavy drinkers, poorer general health, more mental/physical illness/injury days, has more difficulty walking, has more history of strokes, has some high blood pressure, and has the most diabetes.

Alias for Cluster One is Tend to be Middle-Aged/Poorer Health

```
fviz_cluster(km, data = df)
```



Clusters 2 and 3 have a lot of overlap and share a lot of similarities, Cluster 1 has a little overlap but is more separate.