dslusser_4

David Slusser

10/31/2020

```
######## K-Means for clustering assignment
######## The goal is to predict the university by state and public and private (these ar
e categorical variables)
######## In total there are 1302 colleges with 17 measures
######## We need to load the properties
library(readr) # To load in the dataset
library(tidyverse) # For data manipulation
```

```
## ── Attaching packages ─────────────────────────────────────────────────────
──────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ dplyr   1.0.1
## ✓ tibble  3.0.3     ✓ stringr 1.4.0
## ✓ tidyr   1.1.0     ✓ forcats 0.5.0
## ✓ purrr   0.3.4
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tibble' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## ── Conflicts ────────────────────────────────────────────────────────────
──────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra) # For clustering and plotting
```

```
## Warning: package 'factoextra' was built under R version 4.0.2
```

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

```
library(caret) # For normalizing data
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

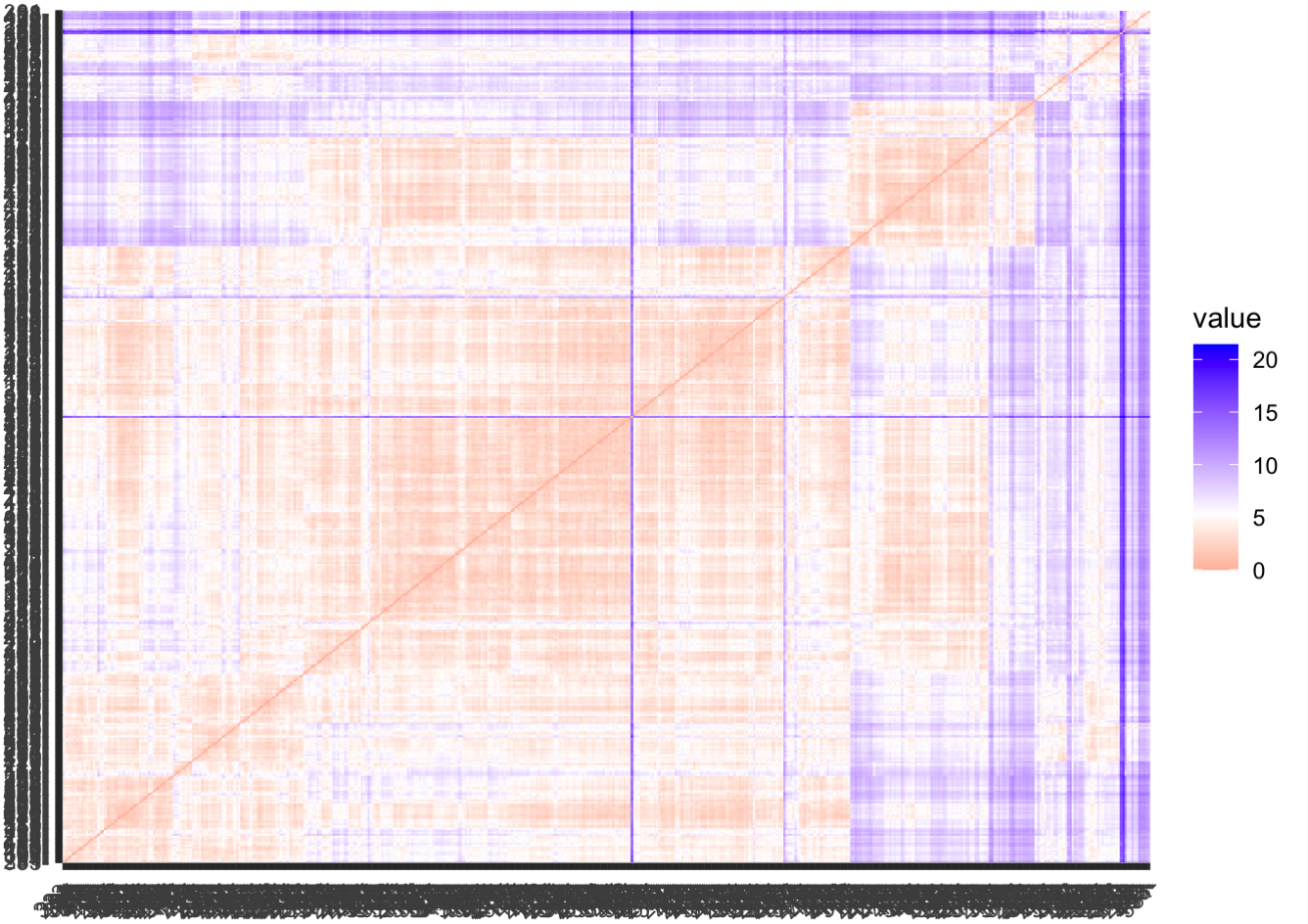
```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(ggplot2) # For plotting
set.seed(1234) # Set the seed to reproduce the same results each time

######## Need to laod in the dataset for university
Universities <- read_csv("~/Desktop/School/Graduate/Machine Learning/HW/Data/Universitie
s.csv")
```

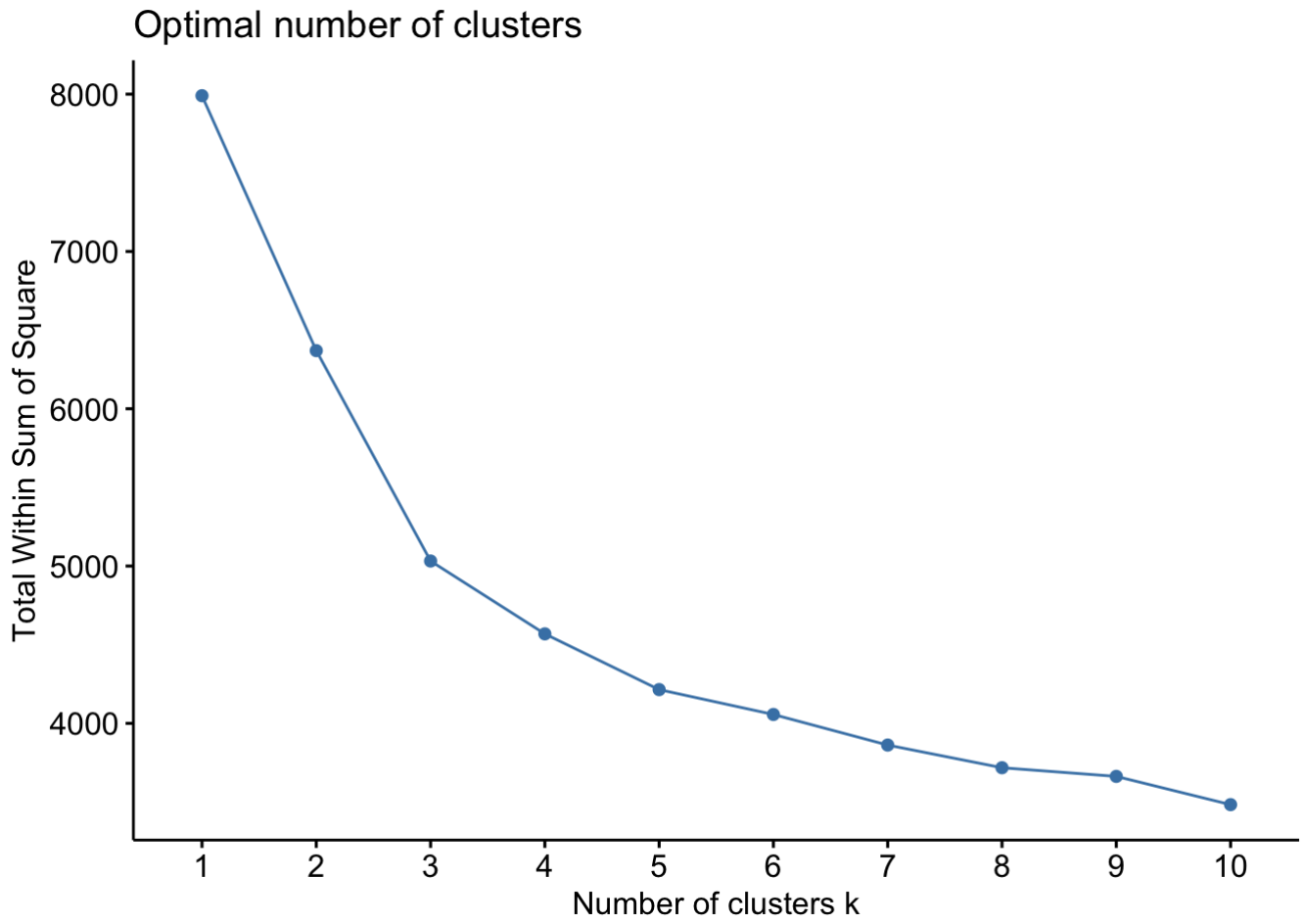
```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   `College Name` = col_character(),
##   State = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

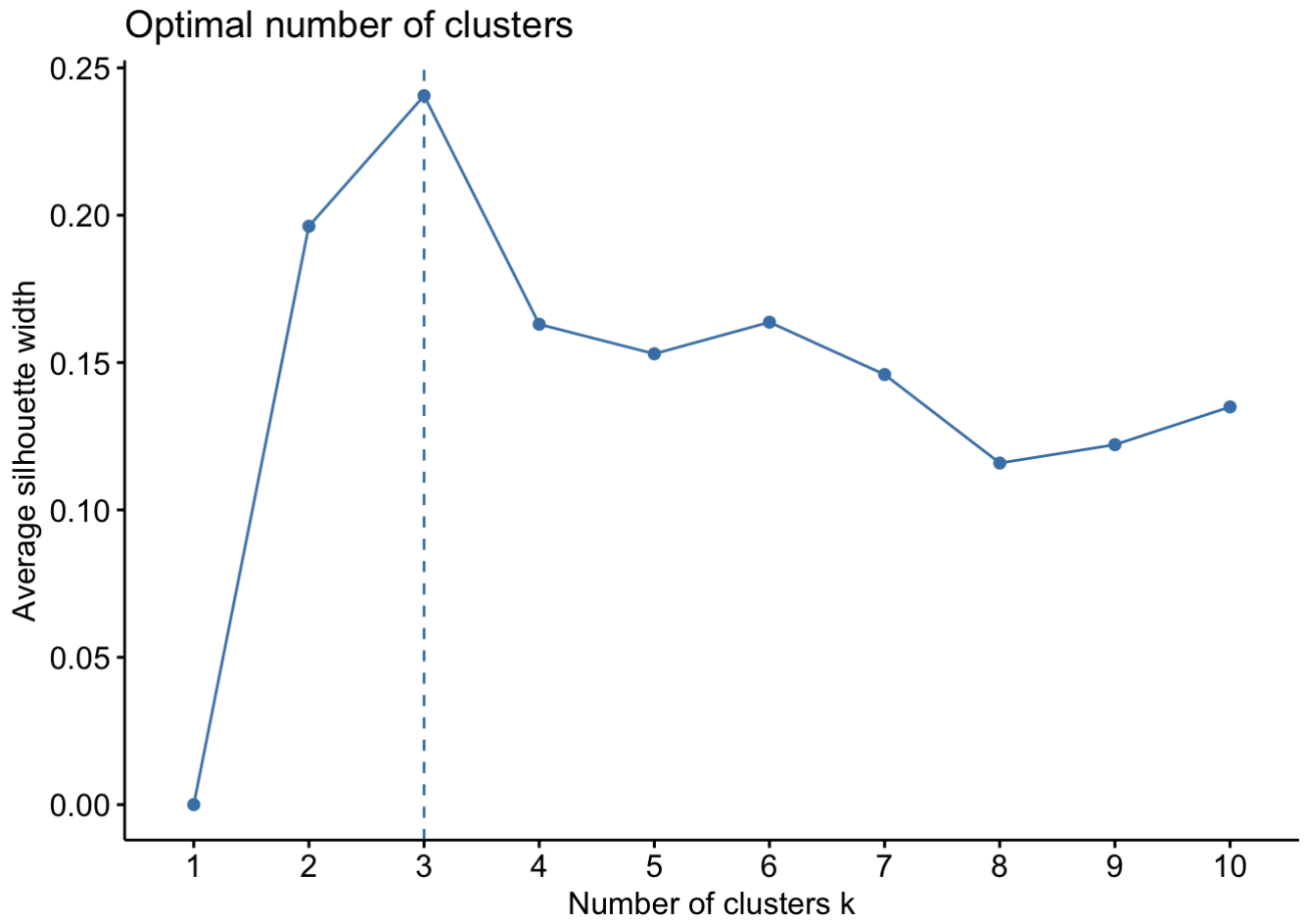
```
######## We need to remove all the missing data from the dataset
Universities1 <- na.omit(Universities) # Rename the data set Universities1

######## Need to normalize all the continuous variavles first
######## The continuous variables are in columns 4-20
######## This means we need to normalize columns 4-20
######## Scaling the data frame (z-score)
Universities_Normal <- scale(Universities1[, 4:20]) # Normalize the continuous variables

######## The preProcess() from the caret package to normalize the variables
######## This will be used for putting each observation into clusters
norm.values <- preProcess(Universities1[, 4:20], method=c("center", "scale"))

######## Get the distance between variables
distance <- get_dist(Universities_Normal) # Calculate the distance
fviz_dist(distance) # Visualize the distance
```




```
######## We want to calculate the distance K that best fits the clustering
######## We will use the factoextra package to see the elbow point and compare it with s
ilhouette
fviz_nbclust(Universities_Normal, kmeans, method = "wss") # Use the normalized data pack
age to find the
```




```
                                                                    # optimal k for elbow p
oint

######## This one is for practice in determing the optimal k-level without institutional
knowledge and to
######## compare with the elbow method
fviz_nbclust(Universities_Normal, kmeans, method = "silhouette") # Use the normalized da
ta package to find the
```




```
                                                                    # optimal k
```

By using the elbow method, we find that the optimal k level is 3 because that is where it looks like an elbow in the chart. The total within Sum of Square (WSS) is about 5,000. We choose three despite the falling WSS post 3 clusters because there are diminishing marginal returns (DMR) past k=3. At k=3, we balance the tradeoff between overfitting and bias. Because DMR sets in at k=3, this is our optimal value

Similarly, by using the silhouette method, we look for the largest average width, which is at k=3. This confirms what we found in the elbow method and our optimal amount of clusters is k = 3

```
######## Calculate the k-means clustering model
######## We will use k = 3 (3 centers) with 25 restarts
######## Use the normalized data values to calculate the variables
k3 <- kmeans(Universities_Normal, centers = 3, nstart = 25) # k = 3, number of restarts
 = 25

######## Vizualize the output
size <- as.data.frame(k3$size) # Number of universities in each cluster, save as size
centers <- as.data.frame(k3$centers) # Output the center of each cluster, save as center
s
```

1. There are 275 universities in the first cluster, which is universities low acceptance rate, lower tuition both in-state and out-of-state, lower costs for room, board, additional fees, books, higher personal dollars, lower faculty members with a PhD, higher student/faculty ratio, and low graduation rate

2. There are 56 universities in the second cluster, which is universities with high number of applicated, high acceptance rate, really low in-state tuition, but more out-of-state tuition compared to cluster 1, average room costs, lower board costs, higher additional fees and cost of books, above average percent of faculty with a PhD, above average student/faculty ratio, but lower graduation rate
3. There are 150 univerities in the thrid cluster, which is universitie with about average number of applicants and applicants accepted, lower than average new students enrolled, high in-state and out-of-state tuition, high room and board costs, lower additional fees, average book costs, above average percent of faculty with lower student/faculty ratio, and high graduation rate

```
######## Now use the categorical variables (State and private/public) to characterize th
e different clusters
######## We know that there are three clusters, so put each point into 3 clusters
Universities_Clusters <- predict(norm.values, Universities1) # Use the normalized values
from the preProcess()
                                                              # function to normalize the
 Universities data

x <- kmeans(Universities_Clusters[, 4:20], 3) # Estimate the cluster for each observatio
n (We know optimal k = 3)
Universities_Clusters$c3 <- x$cluster # Estimate the cluster for each observation for th
e categorical variables

######## Look at the public/private observations
######## Use the tidyverse to get counts by public/private for each cluster
Universities_Clusters %>%
  group_by(`Public (1)/ Private (2)`, c3) %>% #c3 is the variable we named in for out cl
uster
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

```
## `summarise()` regrouping output by 'Public (1)/ Private (2)' (override with `.groups`
argument)
```

```
## # A tibble: 6 x 4
## # Groups:   Public (1)/ Private (2) [2]
##   `Public (1)/ Private (2)`    c3     n   freq
##                       <dbl> <int> <int>  <dbl>
## 1                         1     1    84 0.656
## 2                         1     2    41 0.320
## 3                         1     3     3 0.0234
## 4                         2     1   191 0.557
## 5                         2     2     5 0.0146
## 6                         2     3   147 0.429
```

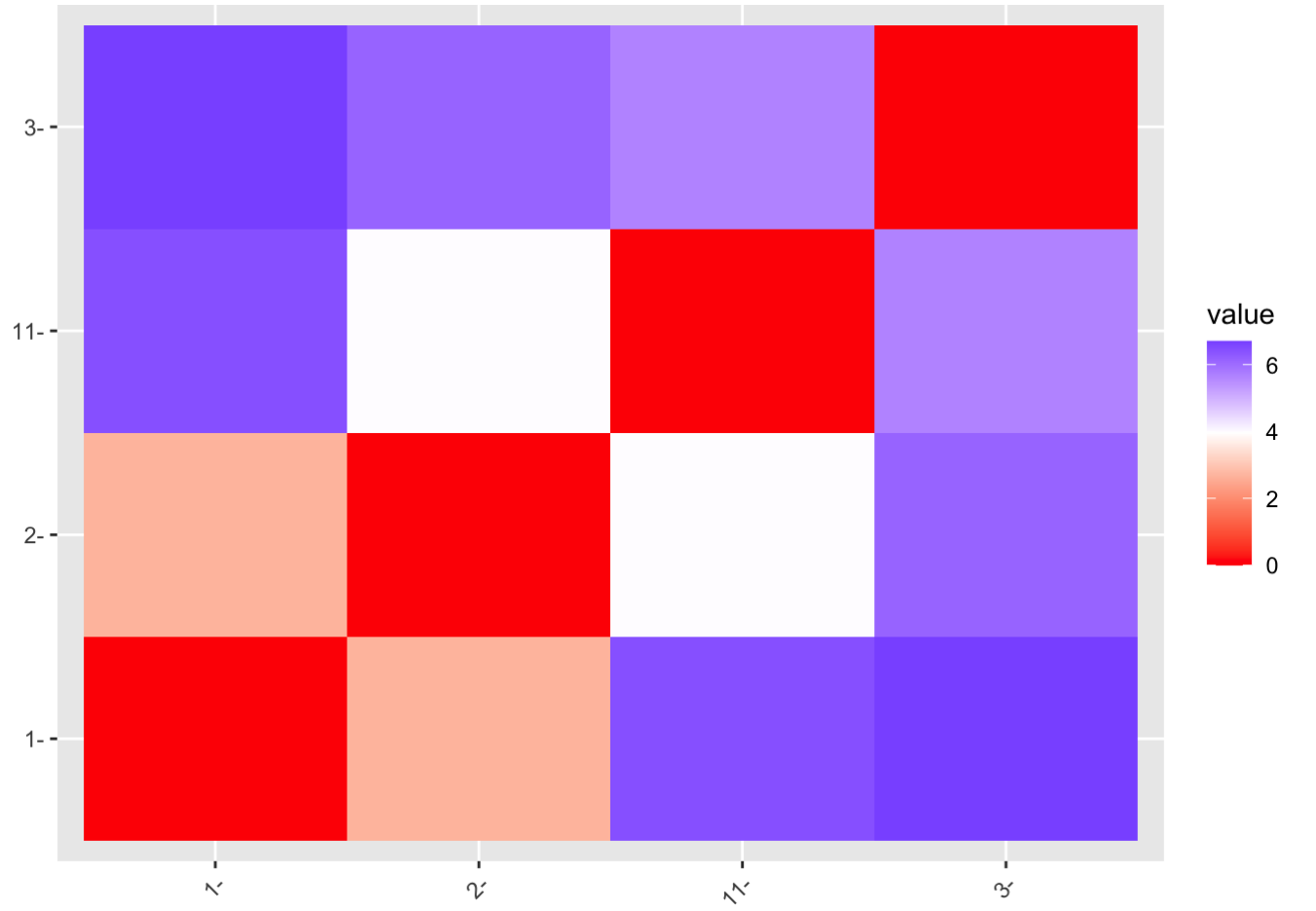
- Looking at public school: 3 (0.0234%) were in cluster 1, the low acceptance rate and low tuition rate. There were 41 (32%) in cluster 2; high number of applicants, low in-state tuition, low graduation rate. There were 84 (65.6%) that were in cluster 3; high acceptance rate, high tuition and room and board, and high graduation

- Looking at private school: 147 (42.9%) were in cluster 1, the low acceptance rate and low tuition rate. There were 5 (14.6%) in cluster 2; high number of applicants, low in-state tuition, low graduation rate. There were 191 (55.7%) that were in cluster 3; high acceptance rate, high tuition and room and board, and high graduation

This signals that cluster 1 is primarily private schools and cluster 2 is public schools, where cluster three is a miz of both.

- Other external information could be for profit and not for profit univerities; where the incentives for education are changed. Splitting that up by private and public could yield a difference. Historical prestige is another that would can explain some of the clusters. The for profit is likely a cluster 3.

```
######## Now use tufts university to predict what cluster it is in
######## We only know some of the values, so only use the values to predict
######## Calculate the Euclidean distance of each record from each cluster

######## Select Tufts University
######## First apply the normalized values
Universities[, 4:20] <- predict(norm.values, Universities[, 4:20]) # Replace first two c
olumns with normalized values

Tufts <- Universities %>%
  select(-c(State, `Public (1)/ Private (2)`, `# PT undergrad`)) %>% # Select all variab
les but state,
                                                                    # pu
blic/private, # of PT Undergrad
  filter(`College Name` == "Tufts University") %>% # Select only Tufts
  select(-c(`College Name`))

centers <- centers %>%
  select(-c(`# PT undergrad`)) # Don't select number of part time undergrad students

####### Calculate the distance between clusters
distance <- get_dist(rbind(Tufts, centers)) # get the distance matrix
fviz_dist(distance) # View it
```




```
Universities_Clusters %>% # Calculate the normalized value for PT undergrad
  filter(c3 == 1) %>% # We only want cluster 1
  summarise(Tufts.PT.Undergrad = mean(`# PT undergrad`)) # Mean number of PT undergrad
```

```
## # A tibble: 1 x 1
##   Tufts.PT.Undergrad
##                <dbl>
## 1             -0.122
```

We find that Tufts belongs in cluster 1, so we want the average number of PT undergrad. The normalized value is -0.273, or that Tuft's University has a below average number of part time undergrad students