library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.0 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.1 ✔ tibble 3.1.8  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.0.0 ──  
## ✔ broom 1.0.3 ✔ rsample 1.1.1  
## ✔ dials 1.1.0 ✔ tune 1.0.1  
## ✔ infer 1.0.4 ✔ workflows 1.1.3  
## ✔ modeldata 1.1.0 ✔ workflowsets 1.0.0  
## ✔ parsnip 1.0.4 ✔ yardstick 1.1.0  
## ✔ recipes 1.0.5   
## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ scales::discard() masks purrr::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ recipes::fixed() masks stringr::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ yardstick::spec() masks readr::spec()  
## ✖ recipes::step() masks stats::step()  
## • Learn how to get started at https://www.tidymodels.org/start/

library(cluster)  
library(factoextra)

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

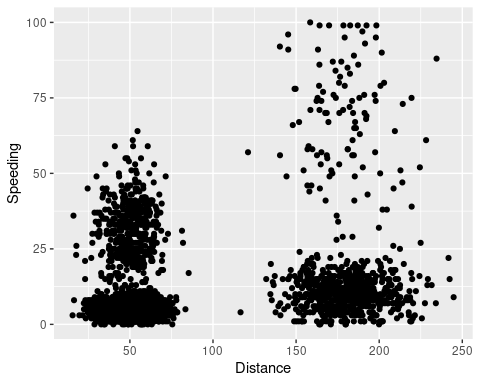
library(dendextend)

##   
## ---------------------  
## Welcome to dendextend version 1.16.0  
## Type citation('dendextend') for how to cite the package.  
##   
## Type browseVignettes(package = 'dendextend') for the package vignette.  
## The github page is: https://github.com/talgalili/dendextend/  
##   
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues  
## You may ask questions at stackoverflow, use the r and dendextend tags:   
## https://stackoverflow.com/questions/tagged/dendextend  
##   
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))  
## ---------------------  
##   
##   
## Attaching package: 'dendextend'  
##   
## The following object is masked from 'package:dials':  
##   
## prune  
##   
## The following object is masked from 'package:stats':  
##   
## cutree

library(readr)  
trucks\_1 <- read\_csv("trucks-1.csv")

## Rows: 4000 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (3): Driver\_ID, Distance, Speeding  
##   
## 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.

trucks <- trucks\_1  
  
ggplot(trucks, aes(Distance, Speeding))+  
 geom\_point()



trucks = trucks %>%   
 select(Distance, Speeding)

Kmeans\_recipe= recipe(~Distance+Speeding, trucks)  
  
trucks\_dummy= Kmeans\_recipe %>%  
 step\_scale(all\_numeric()) %>%  
 step\_center(all\_numeric())  
  
trucks\_dummy = prep(trucks\_dummy, trucks)  
  
trucks\_cleaned = bake(trucks\_dummy, trucks)

summary(trucks\_cleaned)

## Distance Speeding   
## Min. :-1.1319 Min. :-0.7821   
## 1st Qu.:-0.5759 1st Qu.:-0.4903   
## Median :-0.4248 Median :-0.3444   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1947 3rd Qu.:-0.1255   
## Max. : 3.1560 Max. : 6.5127

summary(trucks)

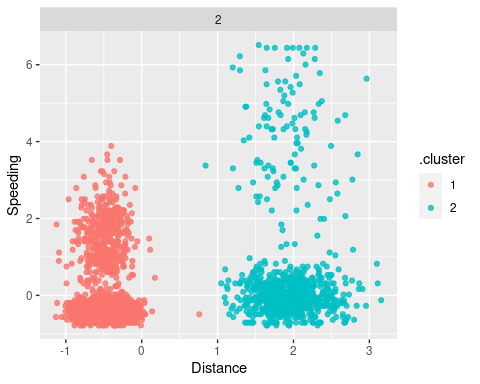
## Distance Speeding   
## Min. : 15.52 Min. : 0.00   
## 1st Qu.: 45.25 1st Qu.: 4.00   
## Median : 53.33 Median : 6.00   
## Mean : 76.04 Mean : 10.72   
## 3rd Qu.: 65.63 3rd Qu.: 9.00   
## Max. :244.79 Max. :100.00

set.seed(64)  
clusts =   
 tibble(k=2) %>%  
 mutate(  
 kclust = map(k,~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, glance),  
 glanced = map(kclust, glance),   
 augmented = map(kclust, augment, trucks\_cleaned)  
 )  
  
clusts

## # A tibble: 1 × 5  
## k kclust tidied glanced augmented   
## <dbl> <list> <list> <list> <list>   
## 1 2 <kmeans> <tibble [1 × 4]> <tibble [1 × 4]> <tibble [4,000 × 3]>

clusters =   
 clusts %>%   
 unnest(cols = c(tidied))  
  
assignments =  
 clusts %>%  
 unnest(cols= c(augmented))  
  
clusterings =   
 clusts %>%  
 unnest(cols = (glanced))

p1 =  
 ggplot(assignments, aes(x = Distance, y = Speeding)) +  
 geom\_point(aes(color= .cluster), alpha = 0.8) +  
 facet\_wrap(~k)  
p1

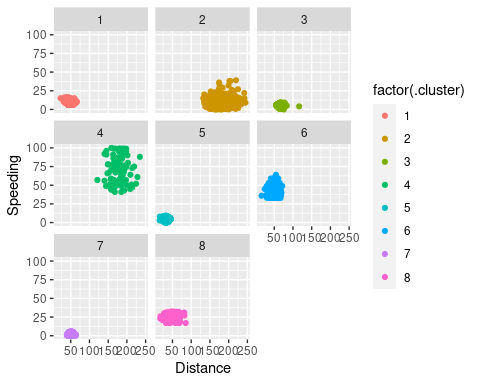


set.seed(412)  
clusters = kmeans(trucks\_cleaned, 8)

trucks = augment(clusters, trucks)  
str(trucks)

## tibble [4,000 × 3] (S3: tbl\_df/tbl/data.frame)  
## $ Distance: num [1:4000] 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding: num [1:4000] 28 25 27 22 25 10 20 8 34 19 ...  
## $ .cluster: Factor w/ 8 levels "1","2","3","4",..: 8 8 8 8 8 1 8 1 6 8 ...

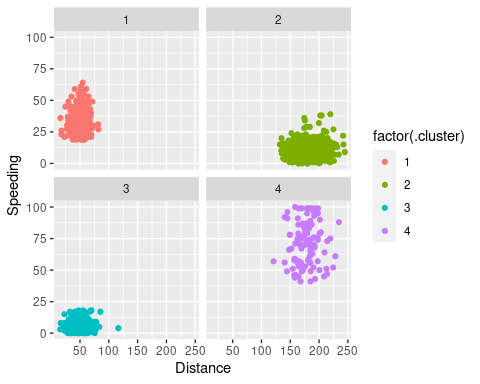
ggplot(trucks, aes(x= Distance, y= Speeding, color= factor(.cluster))) + geom\_point() + facet\_wrap(~factor(.cluster))



set.seed(412)  
clusters = kmeans(trucks\_cleaned, 4)  
  
trucks = augment(clusters, trucks)  
str(trucks)

## tibble [4,000 × 3] (S3: tbl\_df/tbl/data.frame)  
## $ Distance: num [1:4000] 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding: num [1:4000] 28 25 27 22 25 10 20 8 34 19 ...  
## $ .cluster: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 3 1 3 1 1 ...

ggplot(trucks, aes(x= Distance, y= Speeding, color= factor(.cluster))) + geom\_point() + facet\_wrap(~factor(.cluster))



set.seed(412)  
clusters = kmeans(trucks\_cleaned, 1)  
  
trucks = augment(clusters, trucks)  
str(trucks)

## tibble [4,000 × 3] (S3: tbl\_df/tbl/data.frame)  
## $ Distance: num [1:4000] 71.2 52.5 64.5 55.7 54.6 ...  
## $ Speeding: num [1:4000] 28 25 27 22 25 10 20 8 34 19 ...  
## $ .cluster: Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...

ggplot(trucks, aes(x= Distance, y= Speeding, color= factor(.cluster))) + geom\_point() + facet\_wrap(~factor(.cluster))

