Module 6 Assignment 1

Zackery Bradley

9/21/2021

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.2 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(tidymodels)

## -- Attaching packages -------------------------------------- tidymodels 0.1.3 --

## v broom 0.7.8 v rsample 0.1.0   
## v dials 0.0.9 v tune 0.1.6   
## v infer 1.0.0 v workflows 0.2.3   
## v modeldata 0.1.1 v workflowsets 0.1.0   
## v parsnip 0.1.5 v yardstick 0.0.8   
## v recipes 0.1.16

## -- Conflicts ----------------------------------------- tidymodels\_conflicts() --  
## x scales::discard() masks purrr::discard()  
## x dplyr::filter() masks stats::filter()  
## x recipes::fixed() masks stringr::fixed()  
## x dplyr::lag() masks stats::lag()  
## x yardstick::spec() masks readr::spec()  
## x recipes::step() masks stats::step()  
## \* Use tidymodels\_prefer() to resolve common conflicts.

library(naniar)  
library(ggplot2)  
trucks = trucks <- read\_csv("C:/Users/bradl/OneDrive/Desktop/BAN502/Module 6/trucks.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## Driver\_ID = col\_double(),  
## Distance = col\_double(),  
## Speeding = col\_double()  
## )

library(tidyverse) library(tidymodels) library(naniar) trucks = trucks <- read\_csv(“C:/Users/bradl/OneDrive/Desktop/BAN502/Module 6/trucks.csv”)

**Task 1**

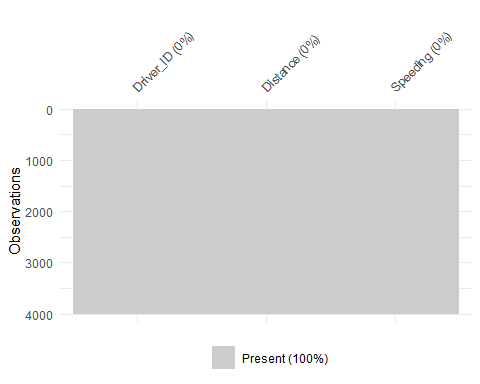
str(trucks)

## spec\_tbl\_df [4,000 x 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Driver\_ID: num [1:4000] 3.42e+09 3.42e+09 3.42e+09 3.42e+09 3.42e+09 ...  
## $ 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 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Driver\_ID = col\_double(),  
## .. Distance = col\_double(),  
## .. Speeding = col\_double()  
## .. )

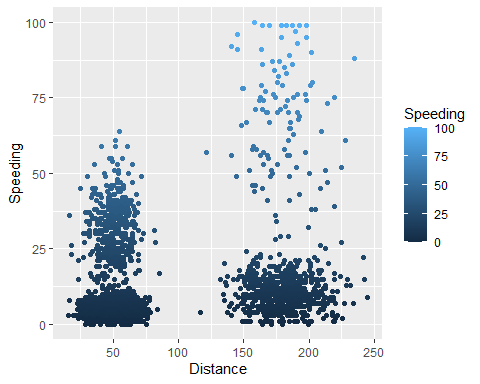
summary(trucks)

## Driver\_ID Distance Speeding   
## Min. :3.423e+09 Min. : 15.52 Min. : 0.00   
## 1st Qu.:3.423e+09 1st Qu.: 45.25 1st Qu.: 4.00   
## Median :3.423e+09 Median : 53.33 Median : 6.00   
## Mean :3.423e+09 Mean : 76.04 Mean : 10.72   
## 3rd Qu.:3.423e+09 3rd Qu.: 65.63 3rd Qu.: 9.00   
## Max. :3.423e+09 Max. :244.79 Max. :100.00

vis\_miss(trucks)



ggplot(trucks, aes(x = Distance, y = Speeding, color = Speeding))+  
geom\_point()

 There seems to be somewhat of a natural cluster in Drivers, as once the traveling distance exceeds around 150 miles, drivers are more inclined to speed.

**Task 2**

trucks\_cleaned = scale(trucks)  
summary(trucks\_cleaned)

## Driver\_ID Distance Speeding   
## Min. :-1.7314 Min. :-1.1319 Min. :-0.7821   
## 1st Qu.:-0.8657 1st Qu.:-0.5759 1st Qu.:-0.4903   
## Median : 0.0000 Median :-0.4248 Median :-0.3444   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.8657 3rd Qu.:-0.1947 3rd Qu.:-0.1255   
## Max. : 1.7314 Max. : 3.1560 Max. : 6.5127

trucks\_cleaned = trucks %>% select(-Driver\_ID)

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

**Task 3**

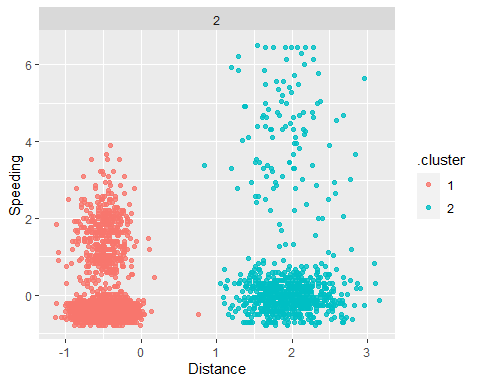
set.seed(64)  
clusts =   
 tibble(k = 2) %>%  
 mutate(  
 kclust = map(k, ~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, tidy),  
 glanced = map(kclust, glance),  
 augmented = map(kclust, augment, trucks\_cleaned)  
 )  
  
clusts

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

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

**plots**

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

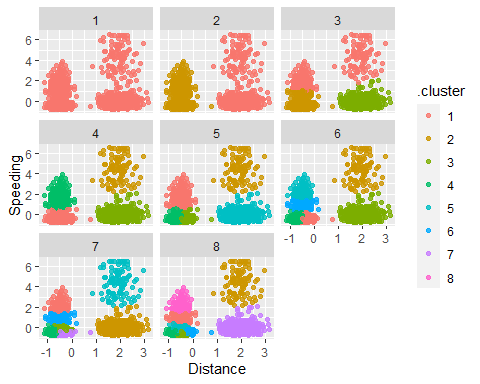
 The cluster shown seems to be pretty natural. There is one possible outlier at around a distance of .75. Longer distance typically generates a higher probability of speeding.

**Task 4**

set.seed(412)  
clusts =   
 tibble(k = 1:8) %>%  
 mutate(  
 kclust = map(k, ~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, tidy),  
 glanced = map(kclust, glance),  
 augmented = map(kclust, augment, trucks\_cleaned)  
 )  
  
clusts

## # A tibble: 8 x 5  
## k kclust tidied glanced augmented   
## <int> <list> <list> <list> <list>   
## 1 1 <kmeans> <tibble [1 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 2 2 <kmeans> <tibble [2 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 3 3 <kmeans> <tibble [3 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 4 4 <kmeans> <tibble [4 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 5 5 <kmeans> <tibble [5 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 6 6 <kmeans> <tibble [6 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 7 7 <kmeans> <tibble [7 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>  
## 8 8 <kmeans> <tibble [8 x 5]> <tibble [1 x 4]> <tibble [4,000 x 3]>

clusters =   
 clusts %>%  
 unnest(cols = c(tidied))  
  
assignments =   
 clusts %>%   
 unnest(cols = c(augmented))  
  
clusterings =   
 clusts %>%  
 unnest(cols = c(glanced))  
  
p2 =   
 ggplot(assignments, aes(x = Distance, y = Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +   
 facet\_wrap(~ k)  
p2

 **Task 5**

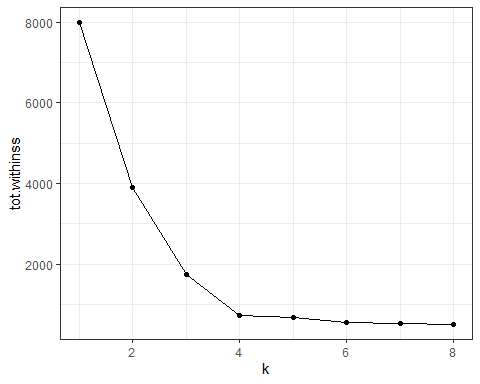
trucks\_scaled = scale(trucks)   
summary(trucks\_scaled)

## Driver\_ID Distance Speeding   
## Min. :-1.7314 Min. :-1.1319 Min. :-0.7821   
## 1st Qu.:-0.8657 1st Qu.:-0.5759 1st Qu.:-0.4903   
## Median : 0.0000 Median :-0.4248 Median :-0.3444   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.8657 3rd Qu.:-0.1947 3rd Qu.:-0.1255   
## Max. : 1.7314 Max. : 3.1560 Max. : 6.5127

\*Create relevant objects\*\*

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

ggplot(clusterings, aes(k, tot.withinss)) +  
 geom\_line() +  
 geom\_point() + theme\_bw()

 Cluster 4 appears to be the best.

**Task 6**

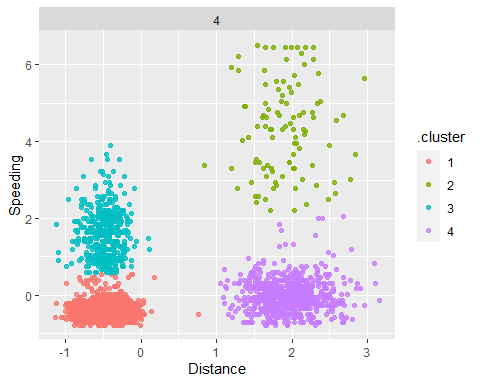
set.seed(64)  
clusts =   
 tibble(k = 4) %>%  
 mutate(  
 kclust = map(k, ~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, tidy),  
 glanced = map(kclust, glance),  
 augmented = map(kclust, augment, trucks\_cleaned)  
 )  
  
clusts

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

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

**plots**

p3 =   
 ggplot(assignments, aes(x = Distance, y = Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +   
 facet\_wrap(~ k)  
p3

 ``` Given the optimal cluster, the majority of drivers in cluster 4 maintain safe speeds on long distances.