Module 6 - Clustering Assignment

Darya Jordan

2022-06-26

## Clustering

Libraries

Read in data

trucks = 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.

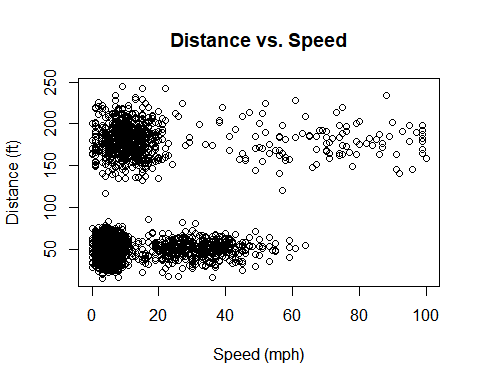
str(trucks)

## spec\_tbl\_df [4,000 × 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()  
## .. )  
## - attr(\*, "problems")=<externalptr>

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

plot(trucks$Speeding, trucks$Distance, xlab='Speed (mph)', ylab='Distance (ft)',   
 main='Distance vs. Speed')



trucks<- select(trucks, -c(Driver\_ID))  
kmeans\_recipe = recipe(~ Distance + Speeding, trucks)   
  
trucks\_dummy = kmeans\_recipe %>%   
 step\_dummy(all\_nominal(), one\_hot = TRUE) %>%  
 step\_scale(all\_numeric()) %>%  
 step\_center(all\_numeric())   
  
trucks\_dummy = prep(trucks\_dummy, trucks) #prepares the recipe  
  
trucks\_cleaned = bake(trucks\_dummy, trucks) #applies the recipe and yields a data frame  
summary(trucks\_cleaned)

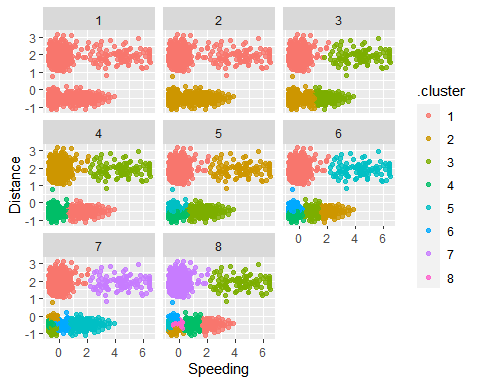
## Distance Speeding   
## Min. :-1.1319 Min. :-0.7821   
## 1st Qu.:-0.5759 1st Qu.:-0.4903   
## Median :-0.4248 Median :-0.3444   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1947 3rd Qu.:-0.1255   
## Max. : 3.1560 Max. : 6.5127

set.seed(64)  
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

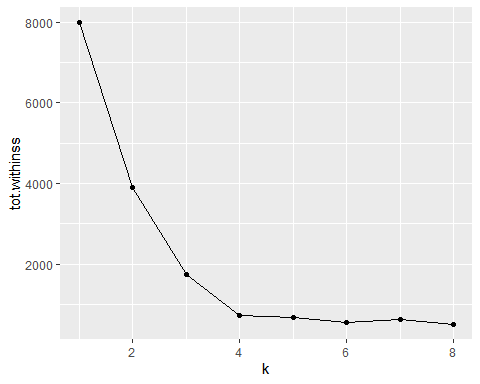
## # A tibble: 8 × 5  
## k kclust tidied glanced augmented   
## <int> <list> <list> <list> <list>   
## 1 1 <kmeans> <tibble [1 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 2 2 <kmeans> <tibble [2 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 3 3 <kmeans> <tibble [3 × 5]> <tibble [1 × 4]> <tibble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

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

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



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



Let’s try 2 clusters

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

Attach cluster to dataset

trucks = augment(clusters, trucks)  
ggplot(trucks, aes(x=Speeding,y=Distance,color=factor(.cluster))) + geom\_point()

