Clustering

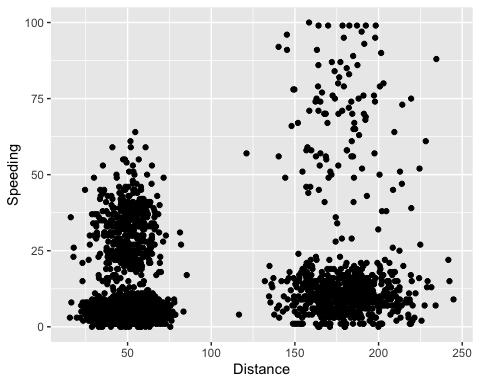
Morgan Nanni

2/22/2022

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

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

 There are some natural clusters, there is a cluster of distances below 75 miles and speeding percentages below 12% (approximately), distances between 150-200 and speeding below 25%, and distances below 75 miles and speeding percentages between 12-50%. The relationship between the 2 variables doesn’t appear to be very strong.

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

set.seed(64)  
clusts =   
 tibble(k = 1:2) %>%  
 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: 2 × 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]>

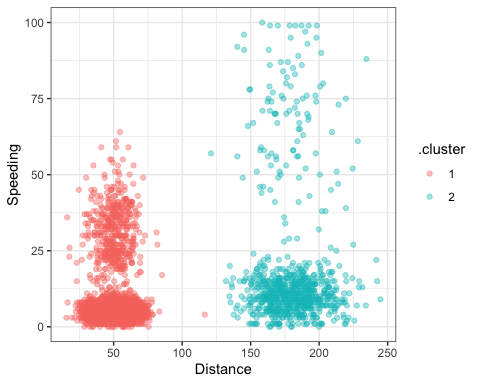
clusters =   
 clusts %>%  
 unnest(cols = c(tidied))  
  
assignments =   
 clusts %>%   
 unnest(cols = c(augmented))  
  
clusterings =   
 clusts %>%  
 unnest(cols = c(glanced))

set.seed(64)  
clusters = kmeans(trucks\_cleaned, 2)

trucks = augment(clusters,trucks)  
head(trucks)

## # A tibble: 6 × 4  
## Driver\_ID Distance Speeding .cluster  
## <dbl> <dbl> <dbl> <fct>   
## 1 3423311935 71.2 28 1   
## 2 3423313212 52.5 25 1   
## 3 3423313724 64.5 27 1   
## 4 3423311373 55.7 22 1   
## 5 3423310999 54.6 25 1   
## 6 3423313857 41.9 10 1

ggplot(trucks, aes(x=Distance, y=Speeding, color = .cluster)) +  
 geom\_point(alpha = 0.4) + theme\_bw()

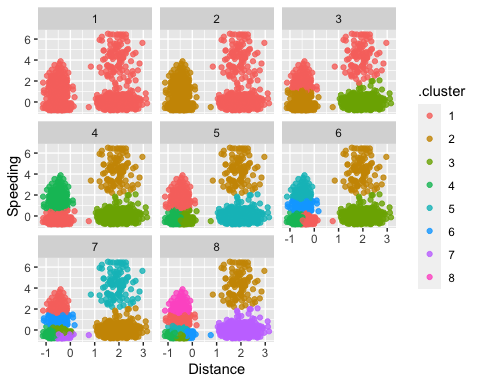
 The 2 clusters are very clearly different from each other. Cluster 1 consists of mostly drivers that drive less than 100 miles on average a day. Cluster 2 is the opposite, consisting of drivers that drive over 125 miles per day. The bulk of both cluster’s points are under 25% of time spent speeding. Cluster 2 has points reaching higher percentages of time spent speeding all the way up to 100% while cluster 1 points don’t reach past about 65%. Cluster 1 points are also more spread out above 25% speeding time going to 100% compared to cluster 2 points that stays more compact as speeding percentage increases.

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

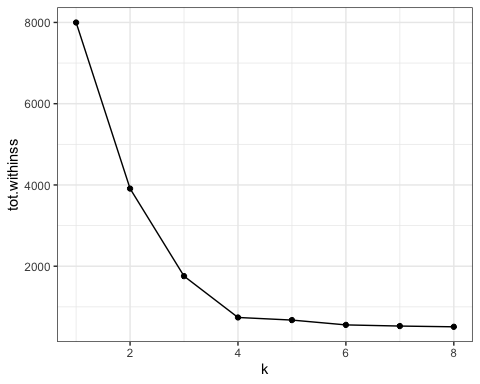
## # A tibble: 8 × 5  
## k kclust tidied glanced augmented   
## <int> <list> <list> <list> <list>   
## 1 1 <kmeans> <tibble [1 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 2 2 <kmeans> <tibble [2 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 3 3 <kmeans> <tibble [3 × 5]> <tibble [1 × 4]> <tibble [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 = Distance, y = Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +   
 facet\_wrap(~ k)  
p1

 The plots above showing k values 1:8 don’t define a clear view of which k value is ideal. The k values of 2,3 and 4 show the most separate and defined clusters while 5,6,7 and 8 show to have clusters on top of each other and connected. The plots do show that higher k values would be less ideal as the increase in clusters would decrease their significance.

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

 Based on the line graph, either 3 or 4 clusters seem to be appropriate for this dataset.

set.seed(64)  
clusts =   
 tibble(k = 1:4) %>%  
 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: 4 × 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]>

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

set.seed(64)  
clusters = kmeans(trucks\_cleaned, 4)

trucks = augment(clusters,trucks)  
head(trucks)

## # A tibble: 6 × 4  
## Driver\_ID Distance Speeding .cluster  
## <dbl> <dbl> <dbl> <fct>   
## 1 3423311935 71.2 28 3   
## 2 3423313212 52.5 25 3   
## 3 3423313724 64.5 27 3   
## 4 3423311373 55.7 22 3   
## 5 3423310999 54.6 25 3   
## 6 3423313857 41.9 10 1

ggplot(trucks, aes(x=Distance, y=Speeding, color = .cluster)) +  
 geom\_point(alpha = 0.4) + theme\_bw()

