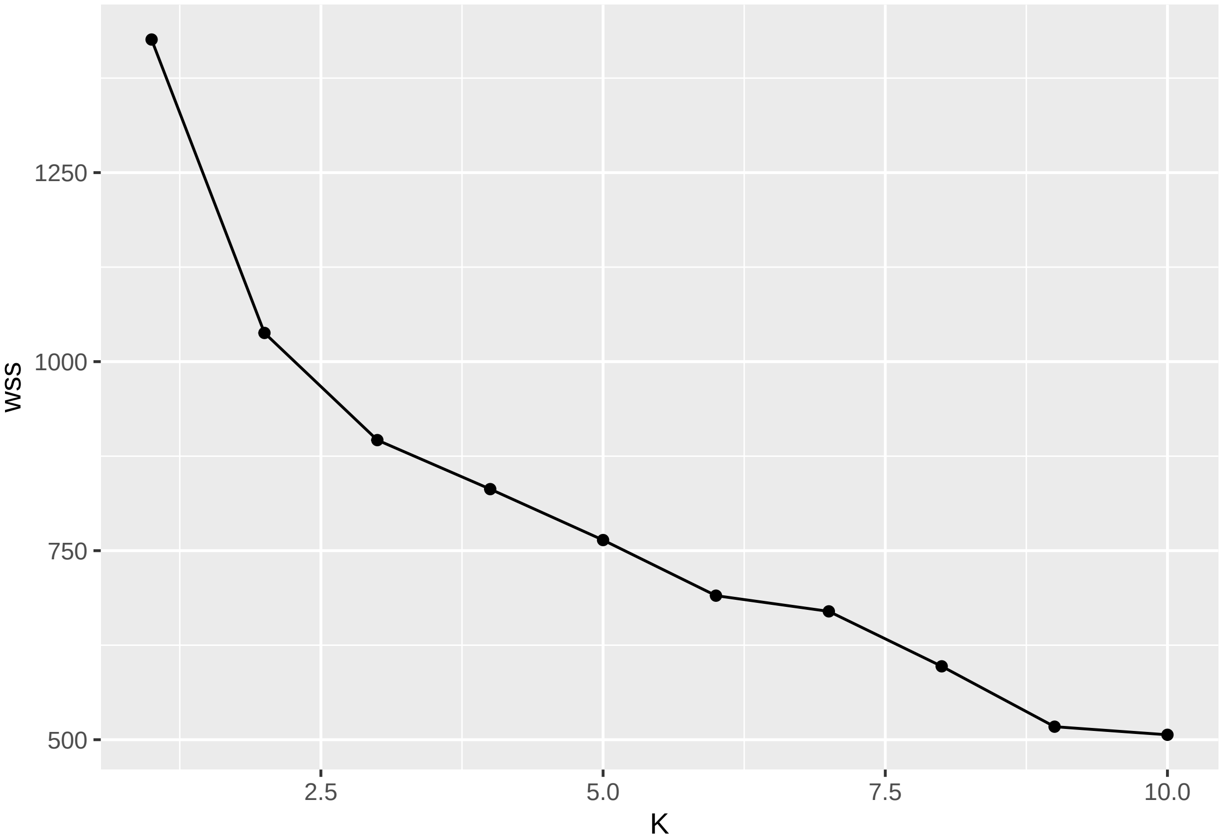
kmeans

2024-12-07

Does geography influence play style?

That is, can we cluster the data such that the clusters represent distinct geographical regions?

```r
library(ggplot2)
coords_df <- read.csv(
  "avg_team_data.csv",
  header=TRUE
)

#Remove variables not to be included in kmeans
kmeans_df <- coords_df[,-c(1:6,ncol(coords_df))]
#Scale data
kmeans_df <- scale(kmeans_df)


#Find k that minimizes within sum of squares (wss)

#Store wss for each k value
wss <- list()
for (i in 1:10) {
  # Fit the model: km.out
  kmeans_wss <- kmeans(kmeans_df, centers = i)$tot.withinss
  # Save the within cluster sum of squares
  wss[[i]] <- kmeans_wss



}


#Scree plot
scree_df <-  data.frame(wss=unlist(wss), k=1:10)

kmeans_scree <- ggplot(scree_df, aes(x = k, y = wss)) +
    geom_point()+
    geom_line() +
    xlab('K')
kmeans_scree
```



```
ggsave("kmeans_scree.png")
```

```
## Saving 6.5 x 4.5 in image
```

Looks like the rate of decrease drops off after k=5. Try clustering with k=5. This seems promising given that it is close to the number of continents. We can creatively group countries together to make 5 groups

```
set.seed(42)
k5 <- kmeans(kmeans_df, centers = 5, nstart = 20)
k5$size
```

```
## [1] 6 8 1 9 8
```

```
#Add clusters to df
coords_df$cluster5 <- as.factor(k5$cluster)
write.csv(coords_df, "avg_team_data_kmeans.csv")
```

There is a group of one! Let's see which country did not fit into any group. Any guesses?