Hierarchical clustering to match fitness buddies

In the previous notebook, we tried time series matching with the Dynamic Time Wrapping algorithm to match fitness buddies that like to workout at similar times. The second component to this is matching people based on activity levels. Below I use the data on activity to perform hierarchical clustering to cluster people with similar activity levels together.

```
library(tidyverse)
library(lubridate)
library(hms)
activity <- read_csv("dailyActivity_merged.csv")
names(activity)
```

```
##  [1] "Id"                       "ActivityDate"
##  [3] "TotalSteps"               "TotalDistance"
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance"
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"
## [15] "Calories"
```

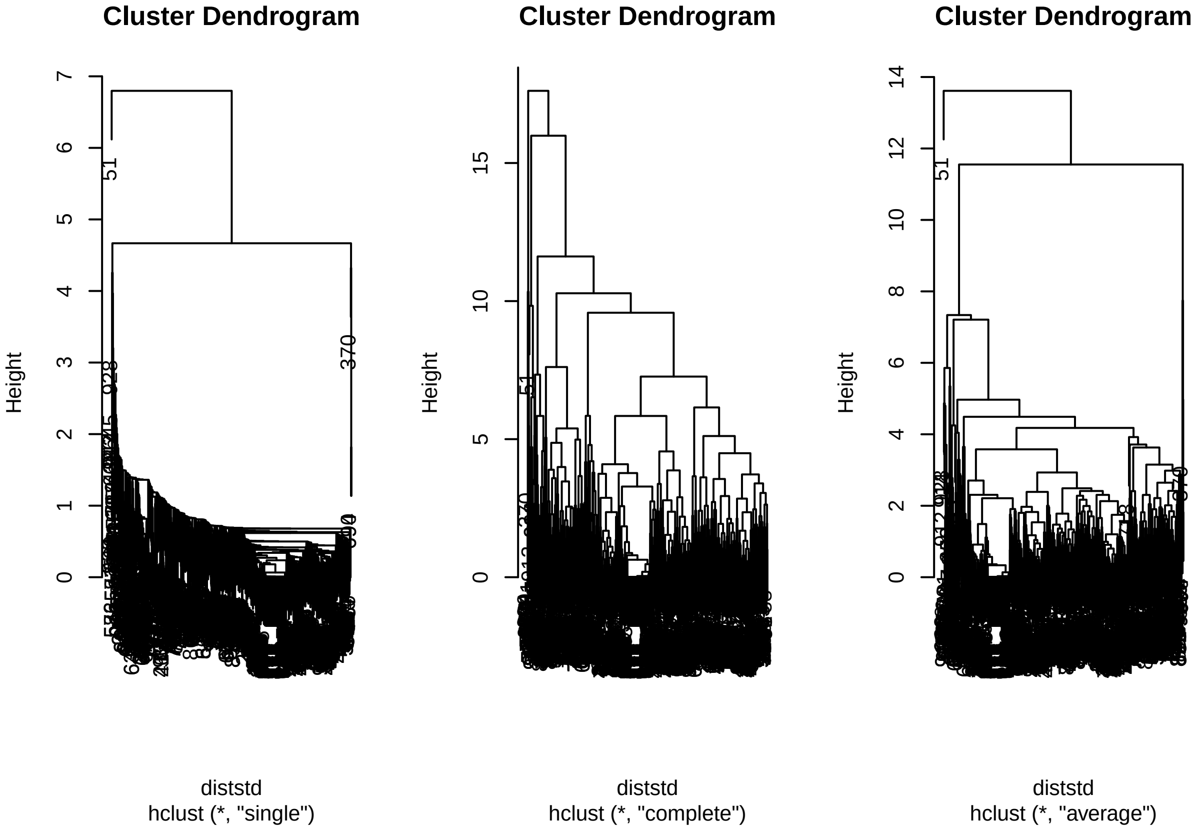
```
activity_new <- activity %>% select(!c("ActivityDate", "TrackerDistance", "LoggedActivitiesDistance",
                                       "Calories"))
clust_data <- activity_new %>% select(!"Id") %>% filter(complete.cases(.))

# create standardized distance
diststd <- dist(scale(clust_data))

# hierarchical clustering (standardized data)
hcstdSL <- hclust(diststd, method="single")
hcstdCL <- hclust(diststd, method="complete")
hcstdAL <- hclust(diststd, method="average")
```

Plotting results from three different methods (single, complete, and average linkage).

```
# plot results (standardized data)
par(mfrow=c(1,3))
plot(hcstdSL)
plot(hcstdCL)
plot(hcstdAL)
```



As we can see, all three methods yield different results. So analysis of which method works the best will need to be done in the future.

One of the advantages of hierarchical clustering is the ability to choose clusters based on the needs of the problem. Below I make a plot with 10 different clusters colored. The number of clusters can be easily changed. This would be one of the other goals as the project progresses.

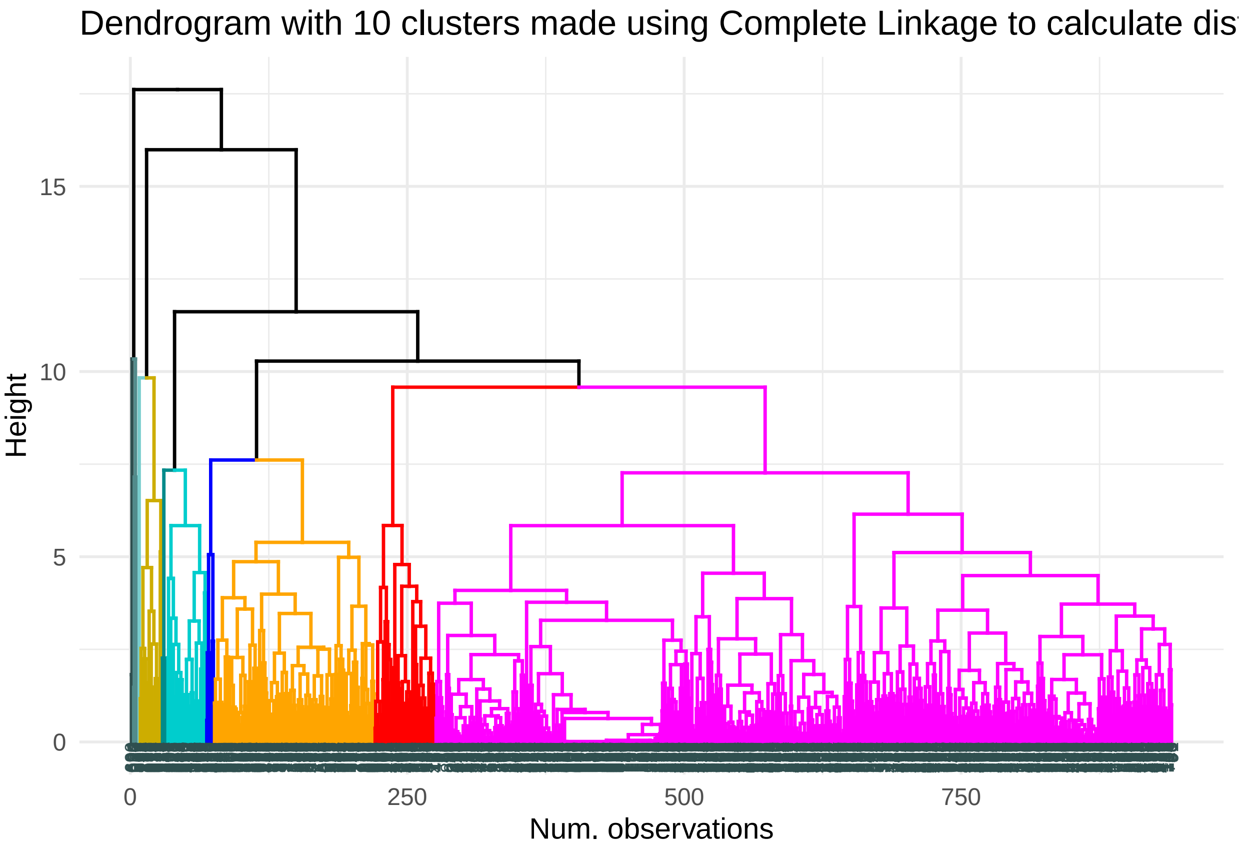
```
library("dendextend")
dendro <- as.dendrogram(hcstdCL)
dendro.col <- dendro %>%
  set("branches_k_color", k = 10, value =   c("darkslategray", "darkslategray4", "darkslategray3", "gol
  set("branches_lwd", 0.6) %>%
  set("labels_colors",
      value = c("darkslategray")) %>%
  set("labels_cex", 0.5)

ggd1 <- as.ggdend(dendro.col)

ggplot(ggd1, theme = theme_minimal()) +
  labs(x = "Num. observations", y = "Height", title = "Dendrogram with 10 clusters made using Complete 
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```


Dendrogram with 10 clusters made using Complete Linkage to calculate dis
Height
15
10
5
0
0
250
500
750
Num. observations