Washington Hiking Trails Hierarchical Clustering

Dataset worked with:

Data scraped from: https://www.wta.org/go-outside/hikes?b_start:int=1 (https://www.wta.org/go-outside/hikes?b_start:int=1)

Data scraping algorithm and original dataset taken/adapted from: https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-11-24/readme.md (https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-11-24/readme.md)

Hosts of TidyX: https://github.com/thebioengineer/TidyX (https://github.com/thebioengineer/TidyX)

Question:

If you enjoyed a popular hiking trail in Washington and you were looking for a similar hike to the previous hiking trail, then what sibling hiking trails can you recommend for a future Washington hike?

Introduction:

The state of Washington has numerous hiking trails throughout the state, and due to its variety of hiking trails, it can be difficult to determine which hiking trail best fits your favorite set of hiking trails. The way we can combat this concern is by hierarchical clustering of the attributes of hiking trails to determine which hiking trails are most related to one another.

The dataset that is used to answer the question at hand is given by the hosts of TidyX, Ellis Hughes and Patrick Ward, and their data scraping R methods for HTML parsing the Washington Trail Association's (WTA) website. The HTML scraper scrapes the website for key descriptors for each trail that ranges from `name` (name of trail), `location` (location of trail), `length` (length of trail), `gain` (elevation gain through trail), `highpoint` (highest elevation on trail), `rating` (rating of trail), `votes` (number of ratings for the trail), `features` (trail's allowed things to do while on trail), `description` (short descriptor of the trail), `trip_type` (roundtrip, one-way, or of trails), `trip_type_id` (ID version of `trip_type`), `length_total` (total length of trail), `location_general` (general location of trail in Washington), `number_of_features` (numeric number of total features allowed on trail), and `highpoint_type` (classification of highest elevation on trail). Only the name of the trail and all quantitative variables associated with each trail is required for further hierarchical clustering analysis and PCA analysis.

Approach:

To address the question of finding sibling (clustered) trails based on the attributes from the Washington Trail Association (WTA) website, the quantitative variables associated with each trail (further described in the introduction). First, due to the fact that there are many trails with a minimal number of reviews, the dataset is filtered to include only trails with more than 70 reviews (votes). Second, the subsetted columns are scaled to all the quantitative variables associated with each trail. Once the quantitative variables are scaled, the scaled values generate a matrix of `Euclidean` distances between each of the trails' quantitative variables. Then a `complete` clustering method is utilized to find the best matching trails with k = 4 clusters. The dendrogram of the clusters is then generated and the colored dendrograms are shown after clustering is completed. Finally, say if you enjoy a certain trail with a gain in elevation with a certain rating on the WTA website, then you can compare the clustering method to `highpoint_type` (highest elevation of the trail). The top 2 PCA vectors are generated from the

quantitative scaled variables to visualize the importance of the top variables that vary the trails from each other. One last thing to help visualize the differences in attributes of each trail is by plotting the scatter points of each PCA 1 and PCA 2 values for each trail.

```
# Your R code here
# hike_data <- readr::read_rds('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-11-24/hike_data.rds')
hike_data <- readr::read_rds('/Users/azima/Desktop/Data Science Classes/DSC 385 Data Exploration and Visualization/Week 12/hike_updated_data.rds')
hike_data <- hike_data %>%
  mutate(
    trip_type = case_when(
      grepl("roundtrip",length) ~ "roundtrip",
      grepl("one-way",length) ~ "one-way",
      grepl("of trails",length) ~ "trails"),
    trip_type_id = case_when(
      grepl("roundtrip",trip_type) ~ 1, # changed from "roundtrip" to 1
      grepl("one-way",trip_type) ~ 2, # changed from "roundtrip" to 2
      grepl("of trails",trip_type) ~ 3), # changed from "roundtrip" to 3
    length_total = as.numeric(gsub("(\\d+[.]\\d+).*","\\1", length)) * ((trip_type == "one-way") + 1),
    gain = as.numeric(gain),
    highpoint = as.numeric(highpoint),
    rating = as.numeric(rating),
    location_general = gsub("(.*)\\s[-][-].*","\\1",location),
    votes = parse_number(votes), # parse number of ratings for trail
    number_of_features = lengths(features),
    highpoint_type = case_when(
      highpoint >= 7500 ~ "highest elevation greater than 7500 ft",
      highpoint >= 5000 & highpoint < 7500 ~ "highest elevation between 5000 & 7499 ft",
      highpoint >= 2500 & highpoint < 5000 ~ "highest elevation between 2500 & 4999 ft",
      highpoint < 2500 ~ "highest elevation less than 2499 ft"),
  )
```

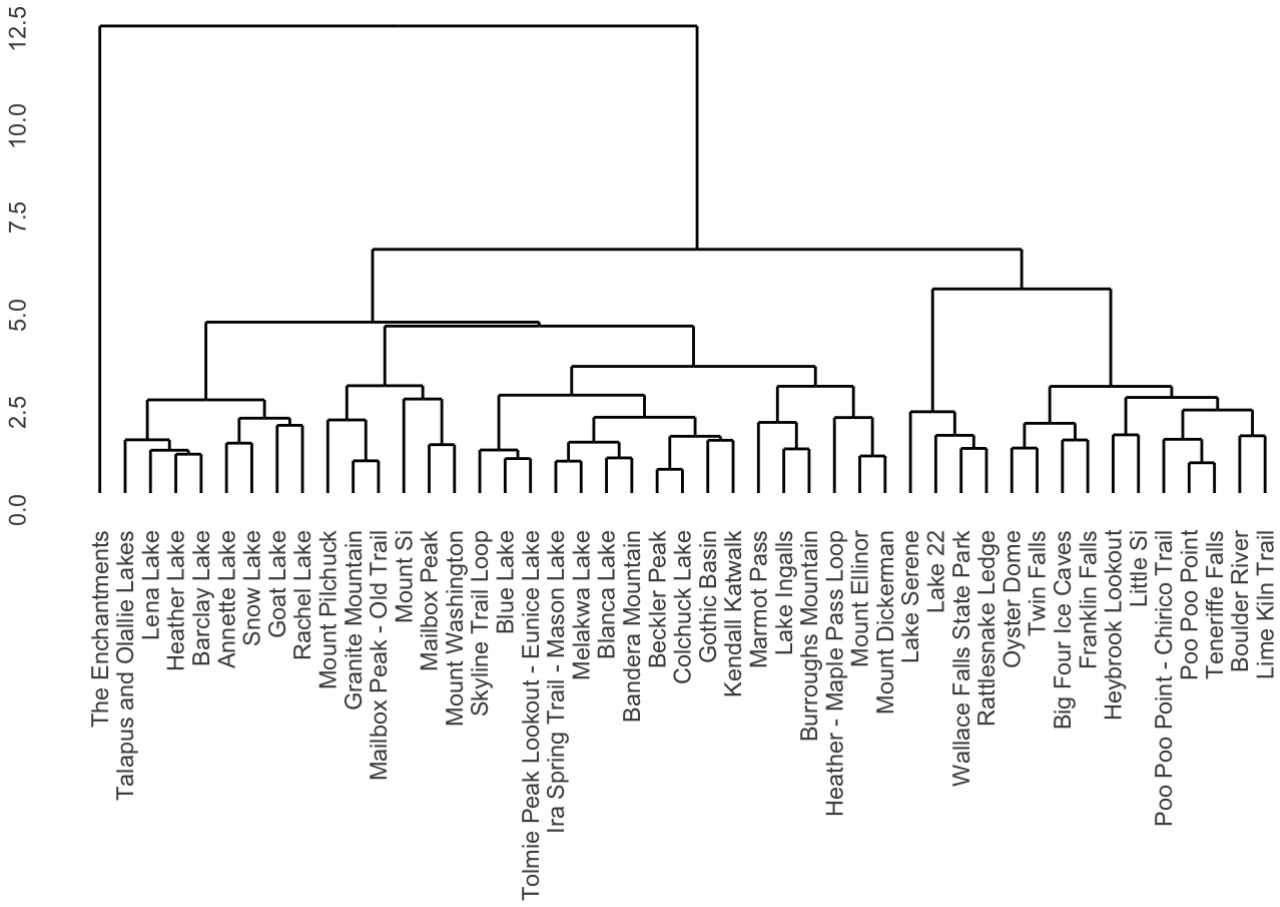
```
hike_data <- hike_data %>%
  filter(votes >= 70) %>%
  mutate(highpoint_type = fct_relevel(highpoint_type, "highest elevation greater than 7500 ft", "highest elevation between 5000 & 7499 ft", "highest elevation between 2500 & 4999 ft", "highest elevation less than 2499 ft"))
hike_data # Table before removing qualitative variables
```

```
## # A tibble: 47 × 15
##    name   location   length   gain highpoint rating votes features description
##    <chr>  <chr>      <chr>   <dbl>     <dbl>  <dbl> <dbl> <list>   <chr>
##  1 Heybr… Central C… 2.6 mi…   850      1700   3.72    78 <chr [3… Heybrook Loo…
##  2 Poo P… Issaquah … 7.2 mi…  1748      2021   3.83   122 <chr [4… Hike railroa…
##  3 Poo P… Issaquah … 3.8 mi…  1760      1850   4.08    80 <chr [5… Hike a short…
##  4 Walla… Central C… 5.6 mi…  1300      1500   4.12   282 <chr [7… An accessibl…
##  5 Oyste… Puget Sou… 5.0 mi…  1050      2025   4.09   158 <chr [5… Oyster Dome …
##  6 Blue … North Cas… 4.4 mi…  1050      6254   4.38    71 <chr [7… At 6254 feet…
##  7 Big F… North Cas… 2.2 mi…   220      1938   3.95   106 <chr [6… This is an e…
##  8 Mount… Snoqualmi… 8.0 mi…  3150      3900   4.02   249 <chr [5… There are ma…
##  9 Ira S… Snoqualmi… 6.5 mi…  2420      4320   4.18   135 <chr [7… Sun drenched…
## 10 Lake … Central C… 8.2 mi…  2000      2521   4.46   357 <chr [6… Lake Serene …
## # … with 37 more rows, and 6 more variables: trip_type <chr>,
## #   trip_type_id <dbl>, length_total <dbl>, location_general <chr>,
## #   number_of_features <int>, highpoint_type <fct>
```

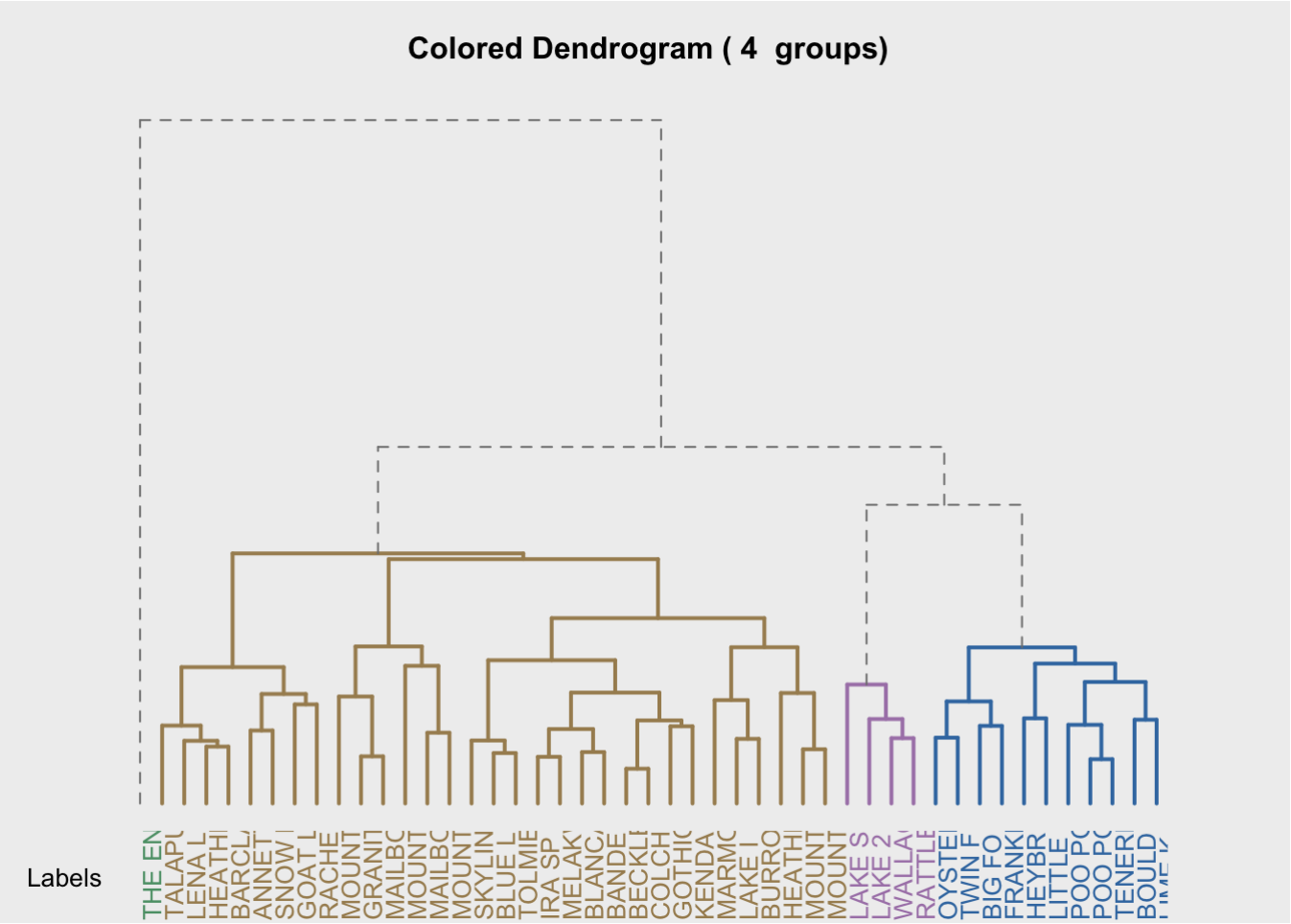
```
clean_hike_data <- hike_data %>%
  select(-c("location", "length", "features", "description", "location_general", "trip_t
ype", "highpoint_type"))

clean_hike_data # Table after removing qualitative variables
```

```
## # A tibble: 47 × 8
##    name   gain highpoint rating votes trip_type_id length_total number_of_featu…
##    <chr> <dbl>     <dbl>  <dbl> <dbl>        <dbl>        <dbl>             <int>
##  1 Heyb…   850      1700   3.72    78            1          2.6                 3
##  2 Poo …  1748      2021   3.83   122            1          7.2                 4
##  3 Poo …  1760      1850   4.08    80            1          3.8                 5
##  4 Wall…  1300      1500   4.12   282            1          5.6                 7
##  5 Oyst…  1050      2025   4.09   158            1          5                   5
##  6 Blue…  1050      6254   4.38    71            1          4.4                 7
##  7 Big …   220      1938   3.95   106            1          2.2                 6
##  8 Moun…  3150      3900   4.02   249            1          8                   5
##  9 Ira …  2420      4320   4.18   135            1          6.5                 7
## 10 Lake…  2000      2521   4.46   357            1          8.2                 6
## # … with 37 more rows
```

```
colors <- c("#5C9E76", "#A78D5F", "#AA83B6",  "#3B79B0")
dist_out <- clean_hike_data %>%
  column_to_rownames(var = "name") %>%
  scale() %>%
  dist(method = "euclidean")

hc_out <- hclust(
  dist_out, method = "complete"
)

# cut dendrogram so there are 4 clusters
cluster <- cutree(hc_out, k = 4)

ggdendrogram(hc_out, rotate = FALSE, color = cluster)
```



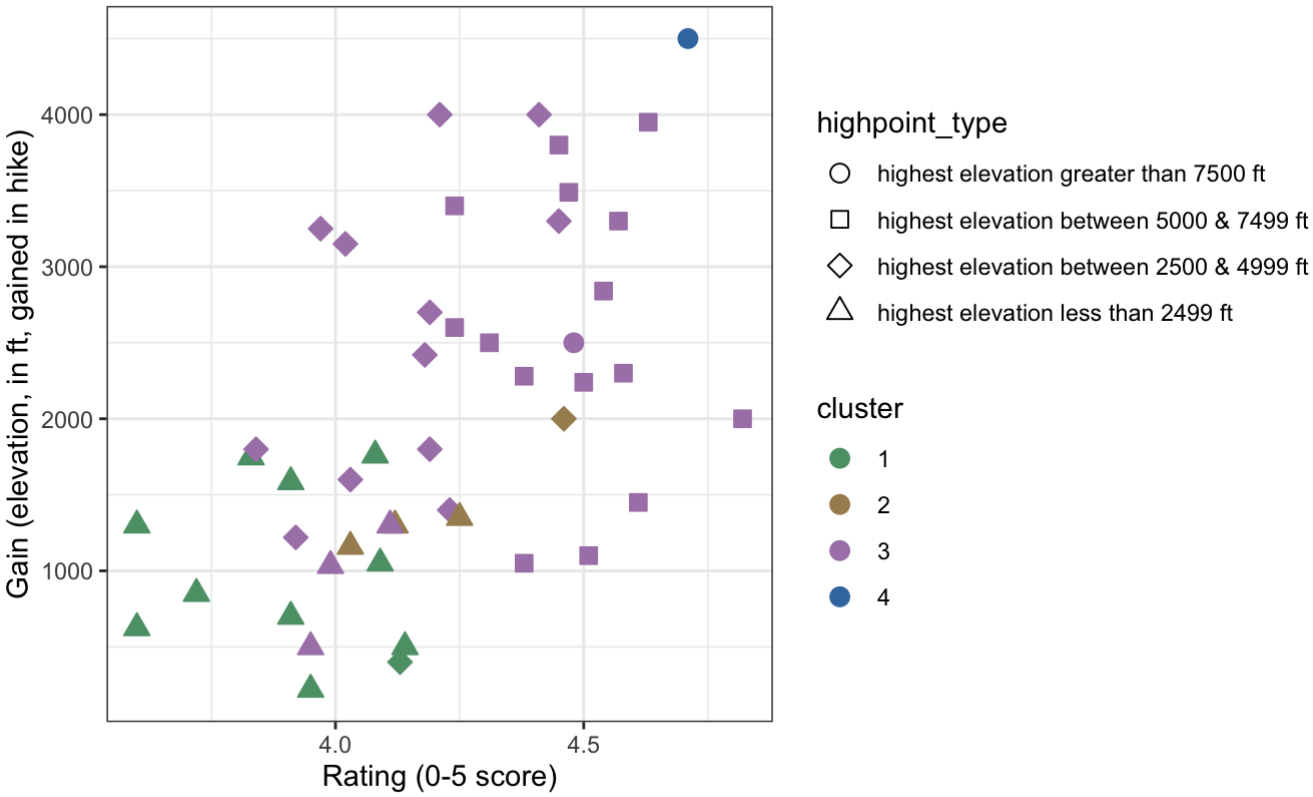

```
# load code of A2R function
source("http://addictedtor.free.fr/packages/A2R/lastVersion/R/code.R")
# colored dendrogram
op = par(bg = "#EFEFEF")
A2Rplot(hc_out, k = 4, boxes = FALSE, col.up = "gray50", col.down = colors, type = "rect
angle", show.labels = TRUE)
```



Analysis:

```
## Joining, by = "name"
```

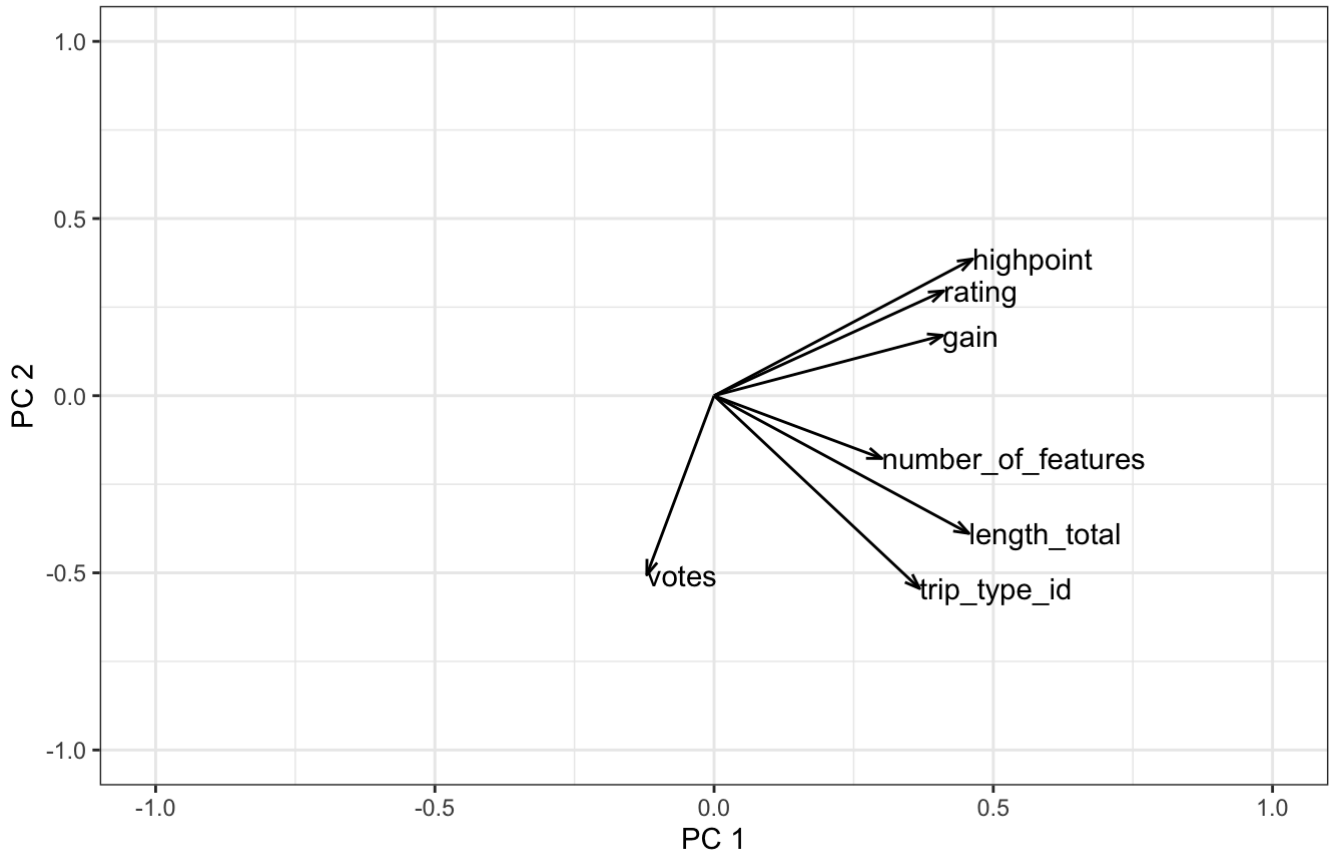
Figure 1. Elevation gain vs rating.
Comparing clustering and highest elevation of trail classification.



Colored by k = 4 hierarchical clustering and
shape identifies the highest elevation of a hiking trail.

```
pca_hike <- clean_hike_data %>%
  select(where(is.numeric)) %>% # retain only numeric columns
  scale() %>%                   # scale to zero mean and unit variance
  prcomp()

arrow_style <- arrow( # Set up the PCA arrow vector
  angle = 20, length = grid::unit(6, "pt"),
  ends = "first", type = "open"
)

pca_hike %>% # get rotation matrix of PC vectors
  tidy(matrix = "rotation") %>%
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) %>%
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style
  ) +
  geom_text(aes(label = column), hjust = 0, vjust = 0.5) +
  xlim(-1, 1) + ylim(-1, 1) +
  theme(legend.position="none") +
  labs(x = "PC 1", y = "PC 2", title = "Figure 2. Rotation plot of PC components 1 and
 2.", caption = "The vectors identify the variability of the data in each column.") +
  theme_bw()
```

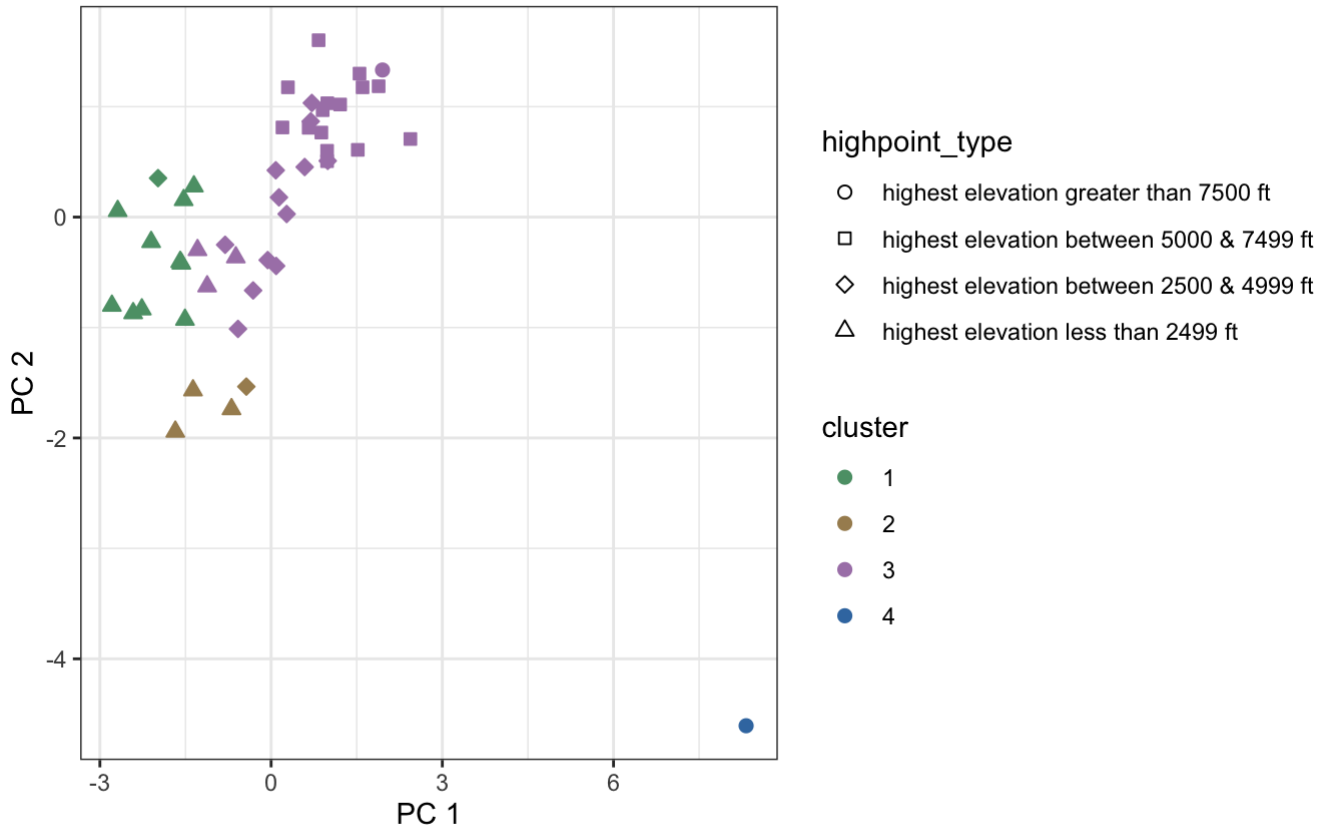
Figure 2. Rotation plot of PC components 1 and 2.



The vectors identify the variability of the data in each column.

```
pca_hike %>%
  augment(cluster_plot) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = cluster, shape = highpoint_type, fill = cluster), size = 2) +
  scale_color_manual(values = c(colors[1], colors[2], colors[3], colors[4])) + # manuall
y change colors to the colors from color pallette
  scale_fill_manual(values = c(colors[1], colors[2], colors[3], colors[4])) + # manually
change colors to the colors from color pallette
  scale_shape_manual(values = c(21, 22, 23, 24)) +
  labs(x = "PC 1", y = "PC 2", title = "Figure 3. PC 2 versus PC 1 colored by hierarchic
al clustering", caption = "Colored by k = 4 hierarchical clustering and \nshape identifi
es the highest elevation of a hiking trail.") +
  theme_bw()
```

Figure 3. PC 2 versus PC 1 colored by hierarchical clustering



Colored by k = 4 hierarchical clustering and shape identifies the highest elevation of a hiking trail.

Discussion:

In the Discussion section, interpret the results of your analysis. Identify any trends revealed (or not revealed). Speculate about why the data looks the way it does.

Based on the the hierarchical clustering dendrograms in the analysis section, you can note that there is one trail that is uniquely different clustered from the rest of the trails, likely because of its extremely long length (36 miles) of `The Enchantments` trail and uniquely high `highpoint` (highest elevation of the trail) that is far and away the most unique trail from the subsetted trails with more than 70 reviews on the WTA website. Cluster 3 is the largest cluster that have second-highest elevation ranges and gains in similar heights (in ft) to each other trail in the cluster. Cluster 3 also has the higher rated trails compared to clusters 2 and 1. Cluster 2 is the smallest cluster that is mostly related to trails that are in between the lowest elevations and medium-high elevations of cluster 3. Cluster 1 is the second largest cluster; it seems to related to the lowest elevation trails and also the lowest rated trails.

It seems from Figure 1, that if you enjoy trails with high ratings and medium-high elevation trails, then you would enjoy cluster 3 trails. When looking at Figure 2, the top 2 PCA vectors seem to be heavily dependent on the length of the trail, elevation, and number of features. It seems that if you enjoy trails with large number of features on it, then you would likely be in cluster 2 or 4. For Figure 3, it shows when looking at the points plotted by PCA values there are likely 3-4 clusters that exist based on the PCA attributes passed into the PCA decomposition.

Overall, it seems that with the attributes from the WTA website we can determine siblings (clusters) for most trails with 70 or more reviews, except for one trail - `The Enchantments` . If you might be looking for a trail like `The Enchantments` then you may need to look for trails outside of Washington because it seems to be uniquely independent from all other trails in Washington with 70 or more reviews. All other trails have siblings that you can find similar-like trails.