# Project 3

**Jason Flores UT EID:** *jf36995*

This is the dataset used in this project:

```
#Data
breed_traits <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da
```

More information about the data set can be found at:

https://github.com/rfordatascience/tidytuesday/blob/master/data/2022/2022-02-01/readme.md

### Part 1

**Question:**

Between bulldogs, french bulldogs, german shepard, golden retrievers and labrador retrievers, what are the similarities and differences of their breed traits?

**Introduction:**

We are working with the `breed_traits` data set which contains trait information on each dog breed and scores for each trait taken from the Vox article. In this data set `breed_traits`, each row is the breed of the dog where the columns are the placement on scale of 1-5 for the breed's tendency to be: Affectionate with family, Good with young children,Good with other dogs, their shedding level, what type of coat grooming frequency required, the drooling level, what king of coat they have and length, how open they are to strangers, playfulness level, how protective they are, their adaptability level, trainability level, energy level, barking level, and how much mental stimulation is needed.

To answer question 1, we use the data set `breed_traits` (already in wide format) and the variables we will be working with is the first five dog breed (first five rows) with all placements score they received. So we exclude column `Coat type` and `Coat Length`. We going to extract a rotation matrix to view contribution of each variable and have an variance explained plot to view the percentages of each variance and lastly a scatter plot is used to determine the overall traits of the dog breed.

**Approach:**

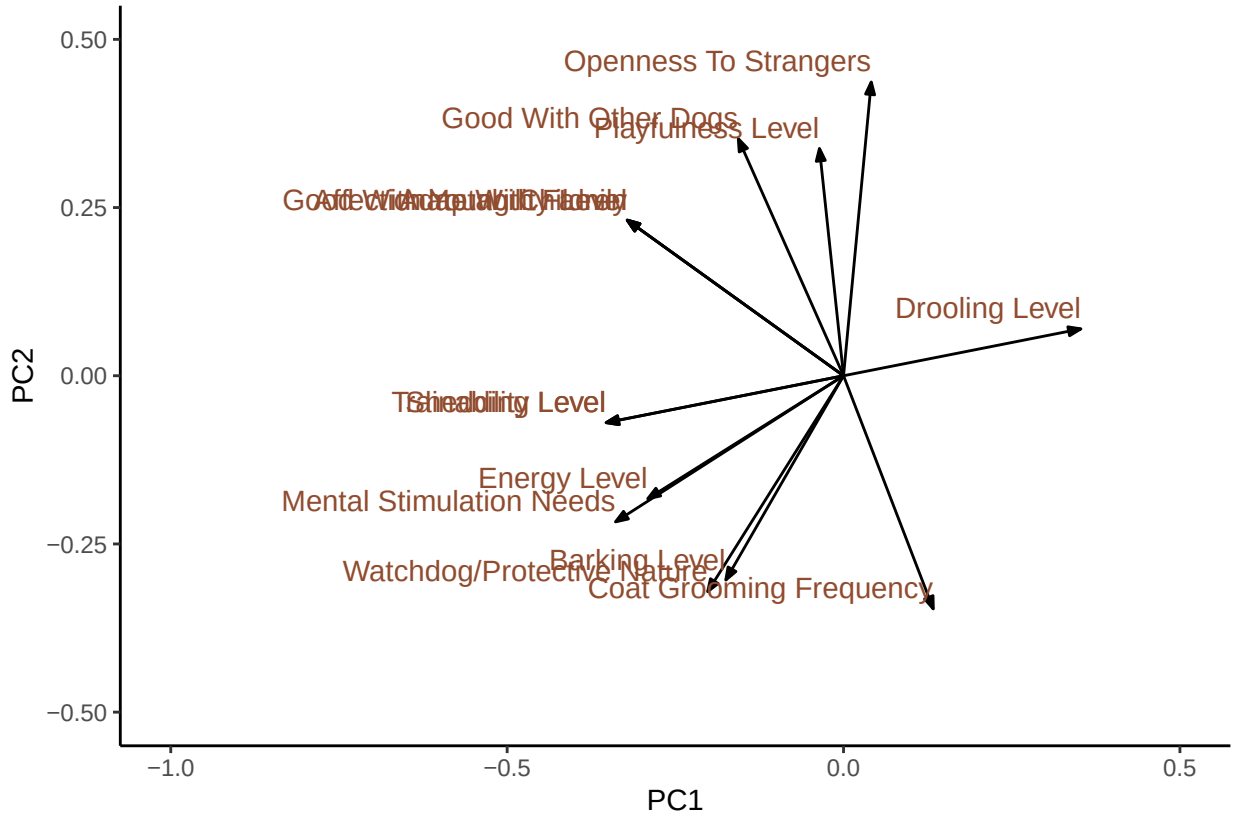
Before we start, we `slice` the data set `breed_traits` to have only the first five dog breeds. From there we want to look the data set in PC coordinates. We start by running a PCA storing the result in `pca_fit_dog` by using `select` to select all numeric variables and `scale` to have the data in unit variance where we finally do PCA by using `prcomp`. After converting the data set to PC coordinates we can now plot the PC coordinates but before we do that want to extract the rotation matrix. This was accomplished by using `tidy(matrix = "rotation")`. An arrow style was defined to view arrows in the plot where the rotation matrix plot was obtained using `geom_segment` and `geom_text`. After obtaining the rotation matrix plot, the equal variance plot was obtained by using `tidy(matrix = "eigenvalues")` where a bar plot (`geom_col`) was used to view the distribution of each variance. Now to view the traits of each dog we plotted the PC coordinates using `augment`, which takes as arguments the fitted model and the original data. We use `geom_point` to obtain the scatter plot and colored to view by dog breed.

**Analysis:**

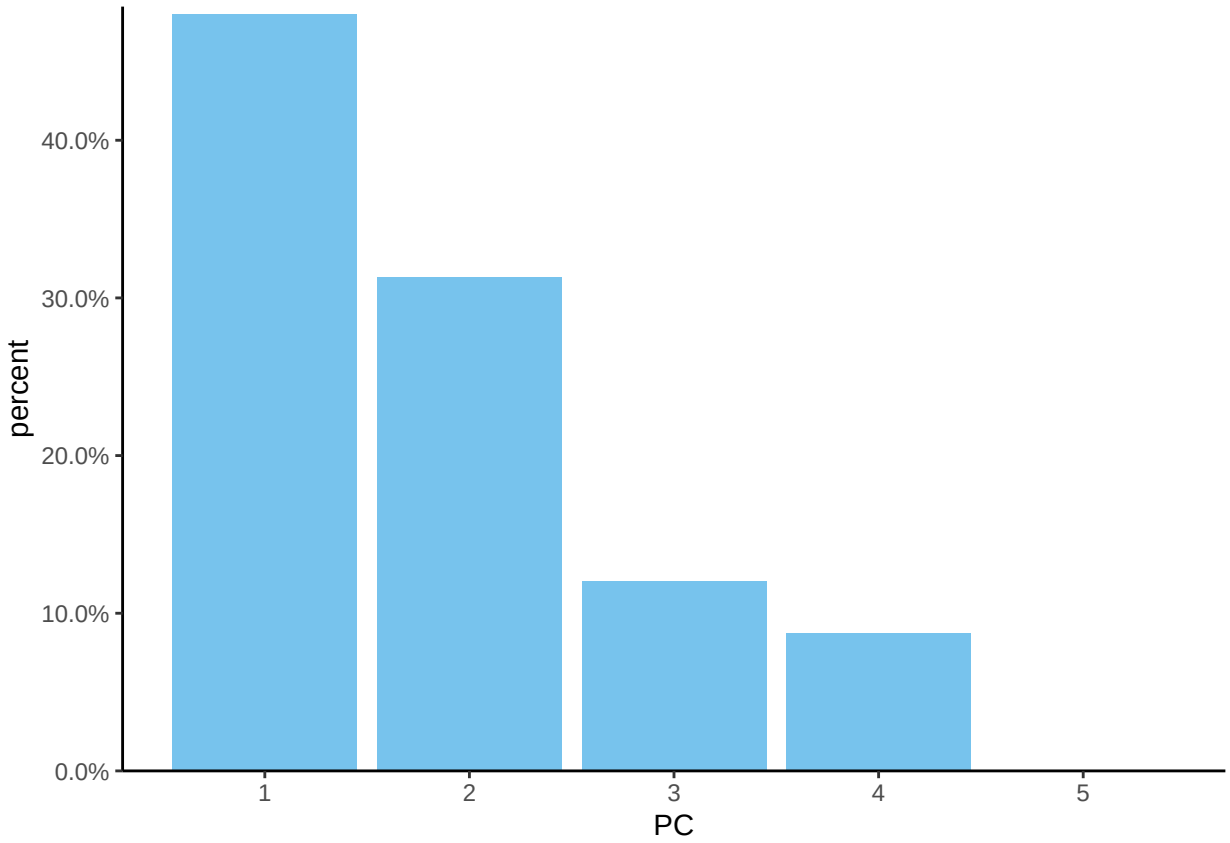
```r
#Simply tidying fist five rows
sum_breed <- breed_traits %>%
  slice(1:5)
sum_breed
```

```
## # A tibble: 5 x 17
##   Breed `Affectionate W~ `Good With Youn~ `Good With Othe~ `Shedding Level`
##   <chr>            <dbl>            <dbl>            <dbl>            <dbl>
## 1 Retr~                5                5                5                4
## 2 Fren~                5                5                4                3
## 3 Germ~                5                5                3                4
## 4 Retr~                5                5                5                4
## 5 Bull~                4                3                3                3
## # ... with 12 more variables: `Coat Grooming Frequency` <dbl>, `Drooling
## #   Level` <dbl>, `Coat Type` <chr>, `Coat Length` <chr>, `Openness To
## #   Strangers` <dbl>, `Playfulness Level` <dbl>, `Watchdog/Protective
## #   Nature` <dbl>, `Adaptability Level` <dbl>, `Trainability Level` <dbl>,
## #   `Energy Level` <dbl>, `Barking Level` <dbl>, `Mental Stimulation
## #   Needs` <dbl>
```

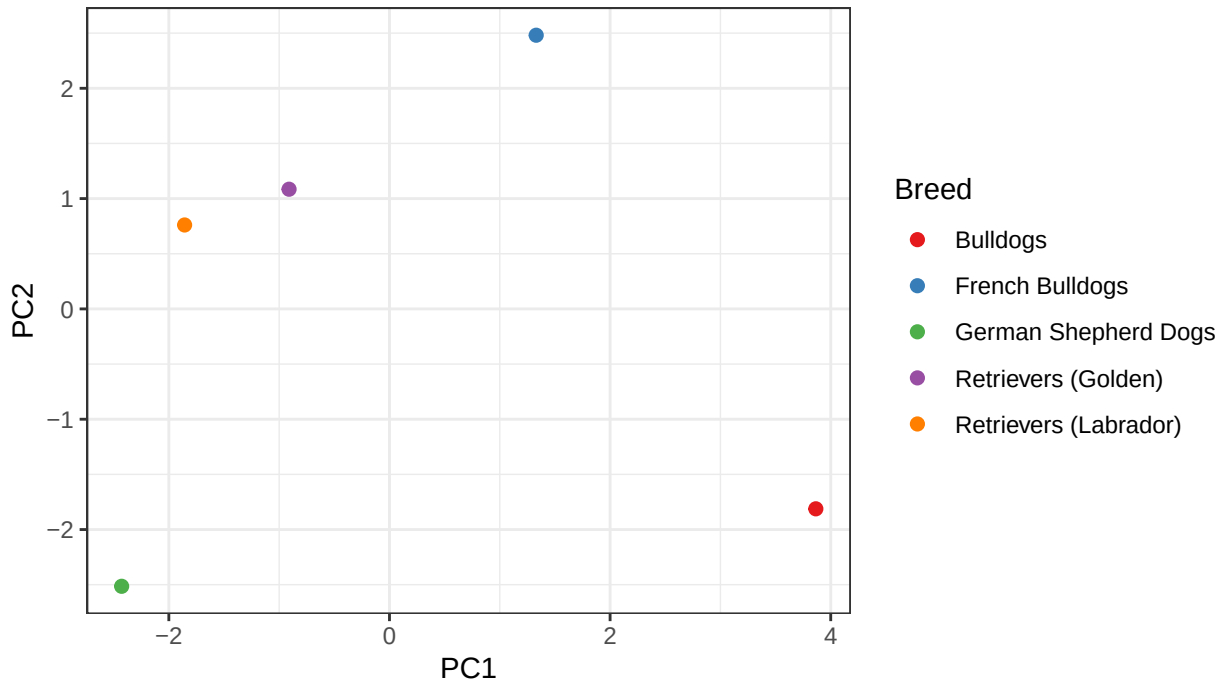
```r
#performing PCA
pca_fit_dog <- sum_breed %>%
  select(where(is.numeric)) %>%
  scale()%>%
  prcomp()

#arrow
  arrow_style <- arrow(
  angle = 20, length = grid::unit(5, "pt"),
  ends = "first", type = "closed"
)
#to wide format
pca_fit_dog %>%
  #extract a rotation matrix
  tidy(matrix = "rotation") %>%
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) %>%

#rotation-plot
ggplot(aes(PC1, PC2)) +
geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style) +
geom_text(aes(label = column),
            hjust = 1, vjust = -0.5,
            color = "#904C2F") +
  xlim(-1.0, 0.5) + ylim(-0.5, 0.5) +
  coord_fixed() +
  theme_classic()
```




```
#eigenvalue plot
pca_fit_dog %>%
  tidy(matrix = "eigenvalues") %>%
  ggplot(aes(PC, percent)) +
  geom_col(fill = "#56B4E9", alpha = 0.8) +
  scale_x_continuous(breaks = 1:9) +
  scale_y_continuous(
    labels = scales::percent_format(),
    expand = expansion(mult = c(0, 0.01))
  ) +
  theme_classic()
```




```
pca_fit_dog %>%
  augment(sum_breed) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = Breed),
             size = 2) +
  coord_fixed()+
  labs(x= "PC1", y= "PC2")+
  theme_bw()+
  scale_color_brewer(palette = "Set1")
```




**Discussion:**

Looking at the variance explained plot, the first principal component explains about 45 % of the total variance in the data set. While the second principle component explains about 30%, third explaining about 15% and the fourth explaining about 10 %. The PCA coordinates plot reveals that French Bulldogs, Golden and Labrador retrievers exhibit high scores on Affectionate with family and good with young children and other dogs, openess to strangers and playfulness level. The bulldog seem to have low scores on those respective traits mentioned above, and the german shepard appear to have both quality of these two. For example, it is not good with other dogs just like the bulldog but it is good with yound children like the other 3 dog breeds. Overall, the first five dog breeds separate mostly along PC 2.

**Part 2**

**Question:**

By using the first 20 dog breeds, how does the affectionate with family score compare to the adaptability level?

**Introduction:**

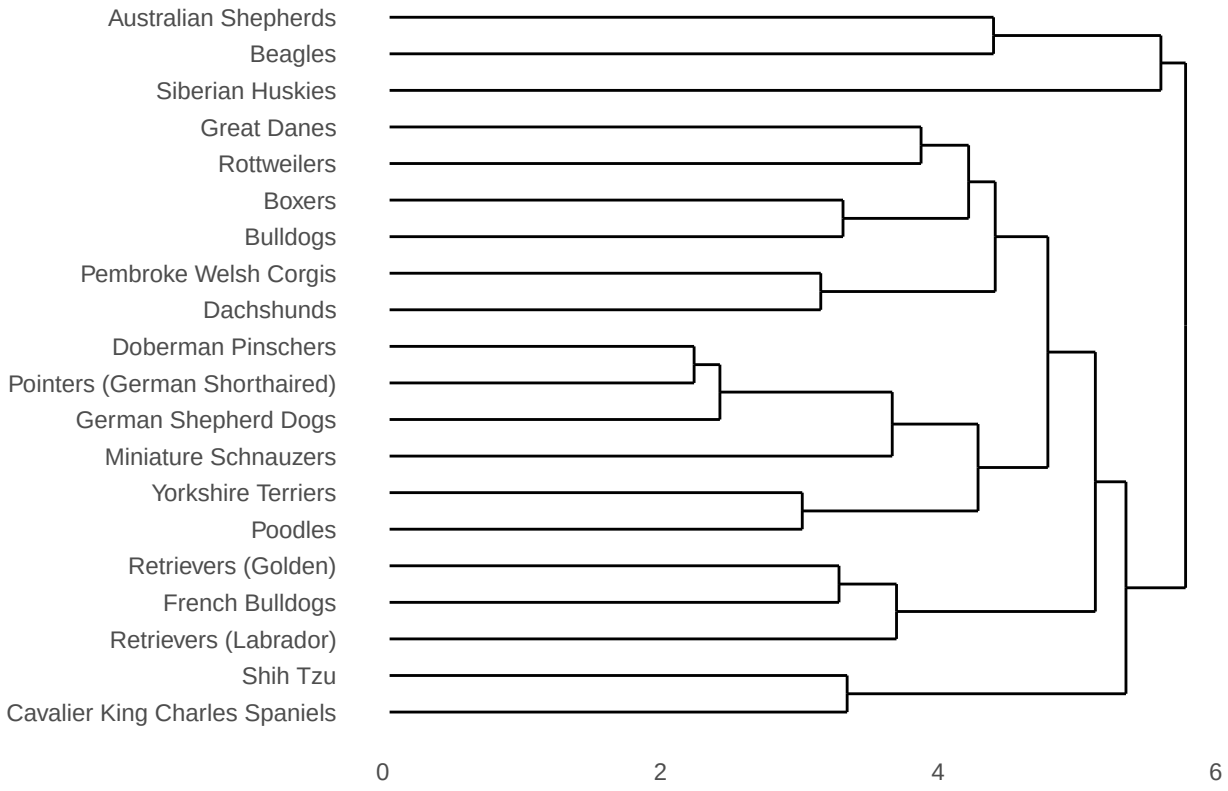
As mentioned in part 1, we will be working with the `breed_traits` data set. To answer question 2 the variables that we will be focusing on are the first 20 dog breeds with the same score variables, but this time we remove any column that is not numeric from the data set, so column `Coat type`, and `Coat length` are removed. A dendrogram plot is obtained to view the hierarchical cluster where its used to plot a scatter plot to view how the dogs ability to show affection towards family relate to its adaptability level.

**Approach:**

First we start by slicing the data set with `slice` to view the first 20 rows of dog breeds. Then we use `select` to remove all non numeric variable being the column `Coat type` and `Coat length`. Now we calculated the distance matrix saving the result as `dist_out` by using `scale` and `dist(method = "euclidean")`. We then create the cluster and plot by using `hc_lust` and `ggdendrogram`. We then cut the dendrogram `k = 5` to have five clusters and used a scatter plot to view how affectionate with family scores relate to the adaptability level using `geom_point` and `position = position_jitter` to prevent overlapping of points.

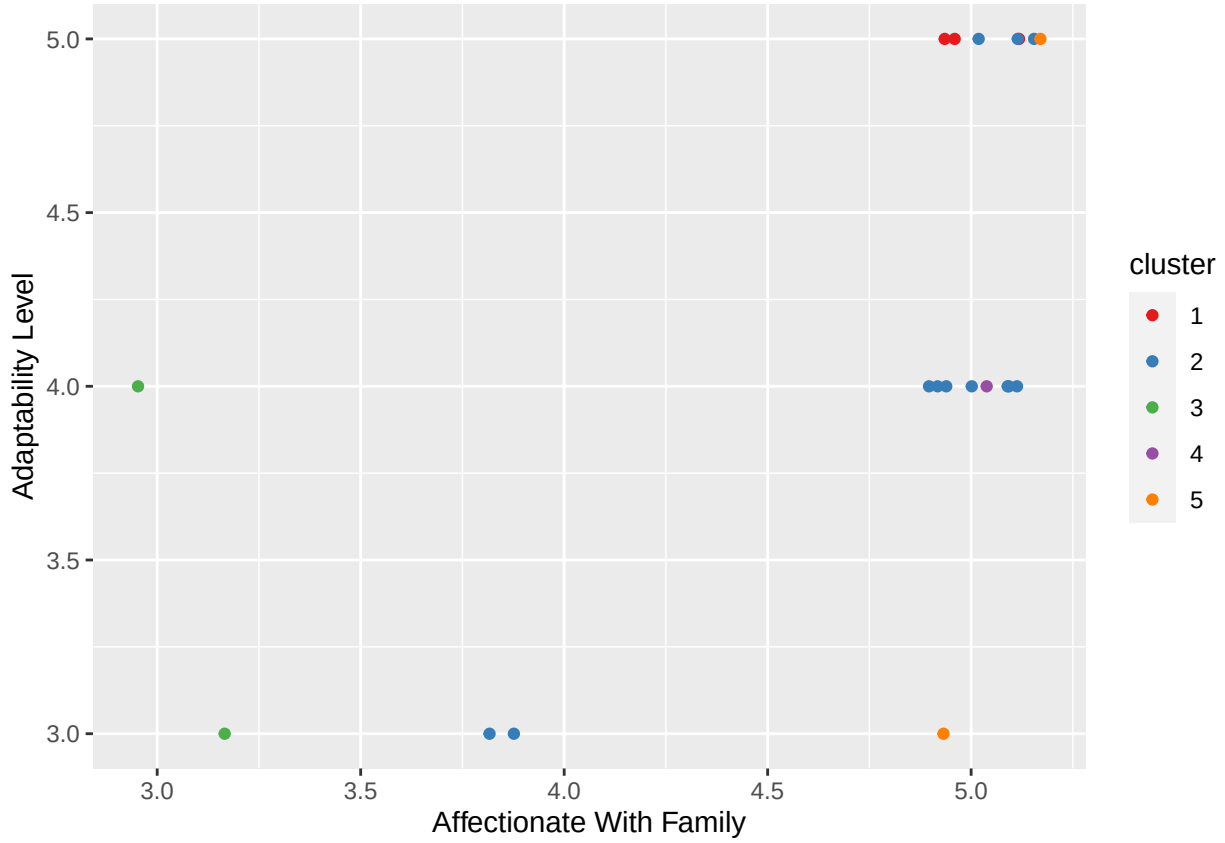
**Analysis:**

```
#tidy
top20breed <- breed_traits %>%
  slice(1:20) %>%
  select(- starts_with("Coat"))

#Dist_out
dist_out <- top20breed %>%
  column_to_rownames(var = "Breed") %>%
  scale()%>%
  dist(method = "euclidean")

#hc_out
hc_out <- hclust(
  dist_out,method = "average"
)

#dendrogram
ggdendrogram(hc_out, rotate = TRUE)
```



```r
#cluster at 5
cluster <- cutree(hc_out, k=5)

#scatterplot
top20breed%>%
  left_join(
    tibble(
      Breed = names(cluster),
      cluster = factor(cluster)
    )
  )%>%
  ggplot(aes(`Affectionate With Family`,`Adaptability Level`))+
  geom_point(aes(color = cluster),
             position = position_jitter(width = 0.2, height = 0))+
  labs(
    x = "Affectionate With Family",
    y = "Adaptability Level"
  )+
 scale_color_brewer(palette = "Set1")
```

```
## Joining, by = "Breed"
```




**Discussion:**

The clusters are not well separated, but it appears that all clusters expect 3 has an affectionate with family and adaptability score of 4 or higher. Additionally, there is a lot of cluster specifically cluster 1 that has a max score for adaptability and affectionate with family. It can be concluded that the dogs that score high on

adaptability level have a high affectionate towards its family. This observation is reasonable since caring and loving the dog leads to trust which can help them to adapt to its surroundings. Furthermore, as seen in part 1, dogs with high affectionate with family tend to be more open to strangers and are good with young children and other dogs.