Palmer Penguins: A Hierarchical Cluster Analysis

Project 1

Juliana Perez Romero

Background and Problem Definition

I am once again using the Palmer Penguins dataset in order to try and solve a simple question:

- Is it possible to use hierarchical clustering in order to correctly identify a penguins species based only on their measurements?

I also aim to exhibit how the type of linkage you use drastically affects the outcome of your clusters. In this data exploration I will use all three linkage methods:

- Complete Linkage
- Single Linkage
- Average Linkage

Why Hierarchical Clustering ?

The use of hierarchical clustering in more consumer based analysis, such as customer segmentation via purchasing habits, or even clinical research by showing the relationship of symptoms to disease diagnosis shows how powerful of a tool it is. I wanted to test the use of this type of clustering to see if it could recognize a pattern within the different penguin measurements in order to develop a way to categorize penguins into their respective species. To put it simply: Is there a discernible pattern within the measurements that can predict species?

Understanding the Palmer Penguins Data Set

The Palmer Penguins data set was collected by Dr.Kristen Gorman from the years 2007 - 2009. The data itself consist of 344 different penguins belonging to 3 species (Adelie, Chinstrap, and Gentoo) across 3 different islands in the Palmer Archipelago. There are 8 distinct variables within this set. Four of which pertain to the measurements of the penguins including: bill_length_mm, bill_depth_mm, flipper_length_mm, and body_mass_g.

```
| Rows: 344
| Columns: 8
| $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
| $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
| $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
| $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
| $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
| $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
| $ sex               <fct> male, female, female, NA, female, male, female, male~
| $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

Data Cleaning

In order to conduct a hierarchical cluster analysis the vectors must be equal in length. So my first step in preparing the data will be to find and remove observations that are incomplete.

```
#check for NA
sum(is.na(penguins))
```

```
| [1] 19
```

```
#clean data // remove NAs
penguins_cleaned <- na.omit(penguins)
sum(is.na(penguins_cleaned))
```

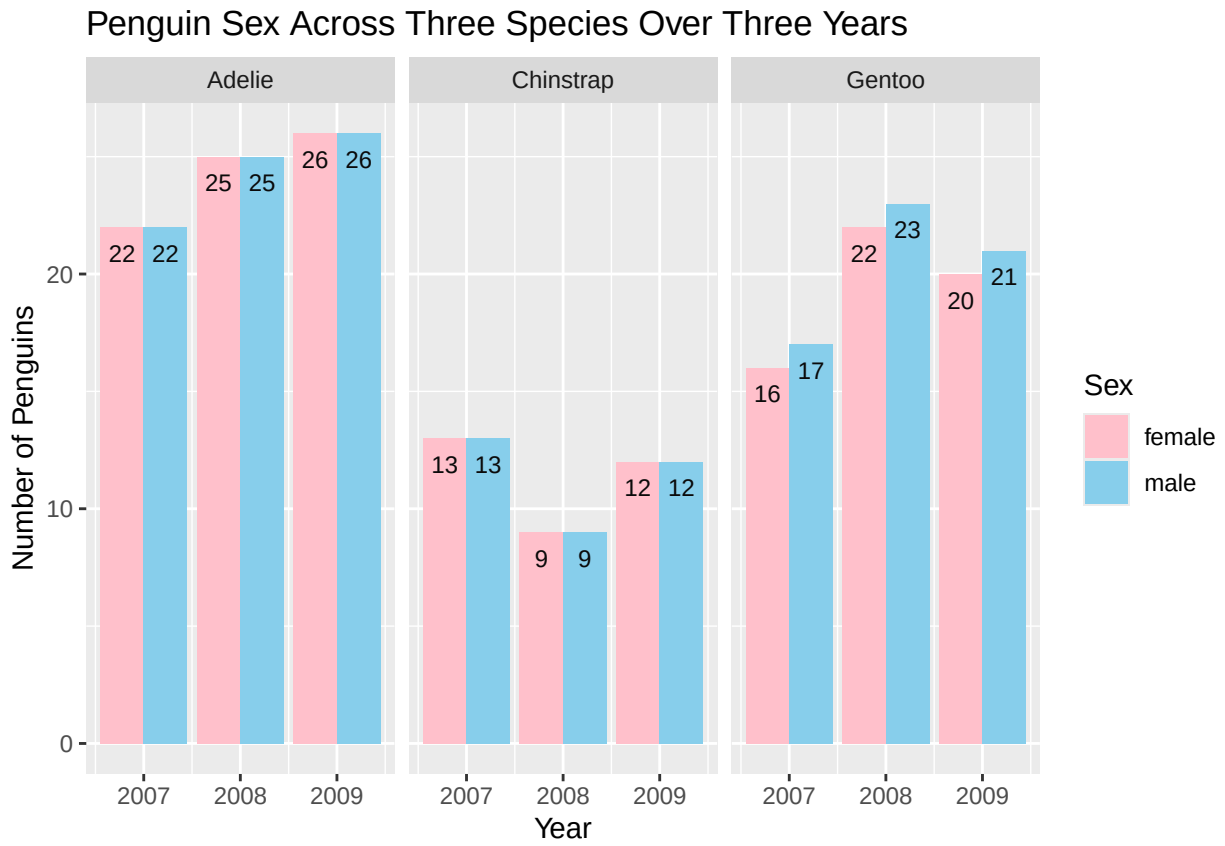
```
| [1] 0
```

After removal of NAs the data frame has lost 11 observations going from 344 total to 333 total. However now the lengths are even and we can proceed with data exploration.

A Look at The Qualitative Data

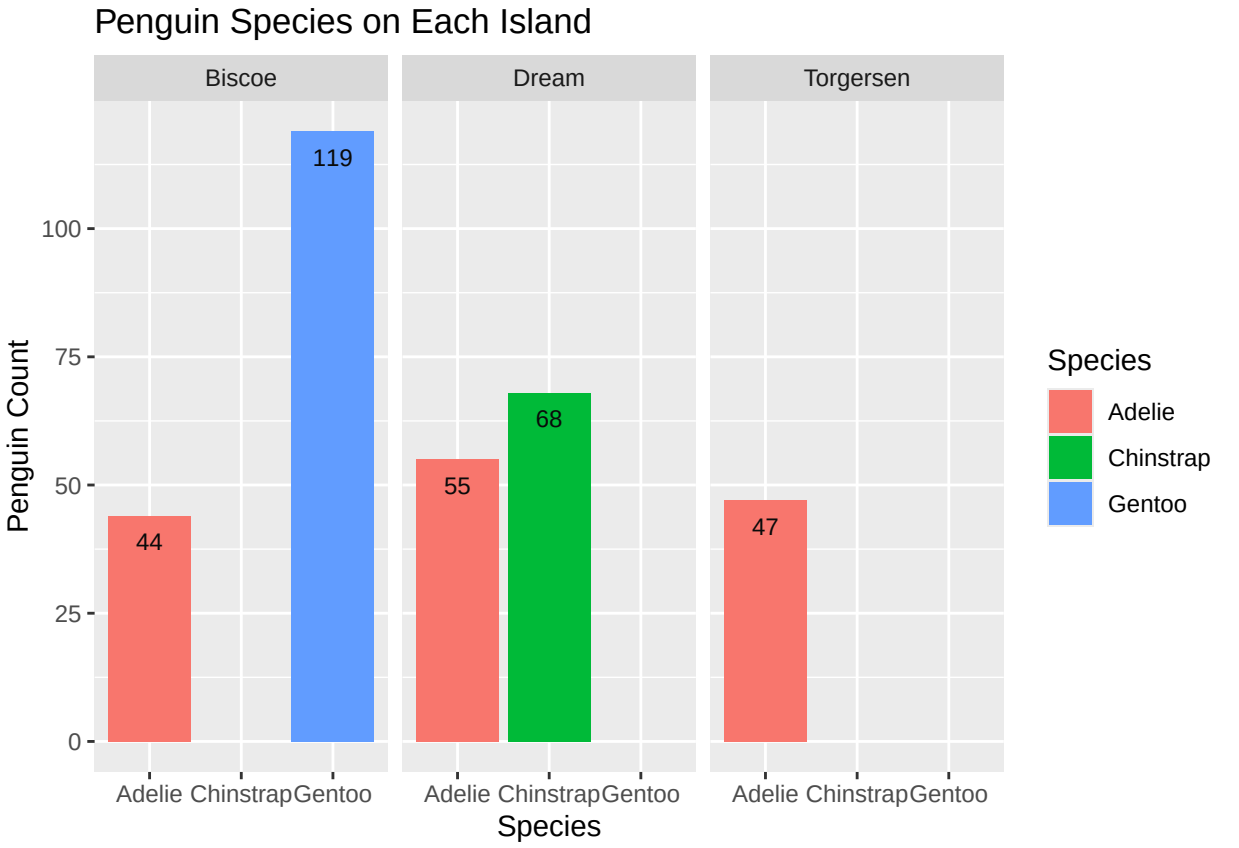
The bulk of my analysis focuses on the use of hierarchical clustering, meaning i will primarily be using the quantitative data provided by this set. As such I wanted to take a look at some of the character class variables.

In this graph we can see the relationship of penguin sex and species over the course of 3 years. The Adelie and Chinstrap species retained 1:1 male to female ratio. The Gentoo however, had a larger male to female ratio, only being by a small margin.



In This bar graph we can see the specific species that reside on each of the 3 islands. The Gentoo species

lives solely on the Biscoe island while the Chinstrap species lives on the Dream island. The Adelie can be found across all three islands.



Exploratory Data Analysis and Corresponding Data Visualizations

I will be using this table as a control for my analysis. This is essentially the goal that I want to achieve and these numbers should be the same in our cluster groups.

```
| # A tibble: 3 x 2
| # Groups:   species [3]
|   species       n
|   <fct>     <int>
| 1 Adelie      146
| 2 Chinstrap    68
| 3 Gentoo      119
```

Since I will only be working with measurements I am also going to create a subset of the cleaned data that only contains those columns.

```
penguins_clustering <- penguins_cleaned %>%
  select (bill_length_mm:body_mass_g)
```

Hierarchical Clustering: Complete Linkage

I will be explaining this process in depth for Complete Linkage. The other linkages methods are almost identical so I will be going over those more briefly.

Taking a look at our measurement variables bill length, bill depth, and flipper length are all measured using millimeters. However, body mass is measured using grams. This means that currently our data is spread across two distinct scales.

- In order to continue we have to scale our data to normalize the values so that all the data points are measured fairly and body mass doesn't disproportionately weigh the data.

```
penguins_scaled <- scale(penguins_clustering)
```

- Now we will take the Euclidean distance of each scaled data point. This allows us to quantify how similar or dissimilar two data points are from one another.

```
penguins_distance <- dist(penguins_scaled)
```

- Now we can use the `hclust()` function to create a hierarchical clustering structure
  - `method = "complete"`
  - in complete linkage the distance is determined by the maximum distance between two points

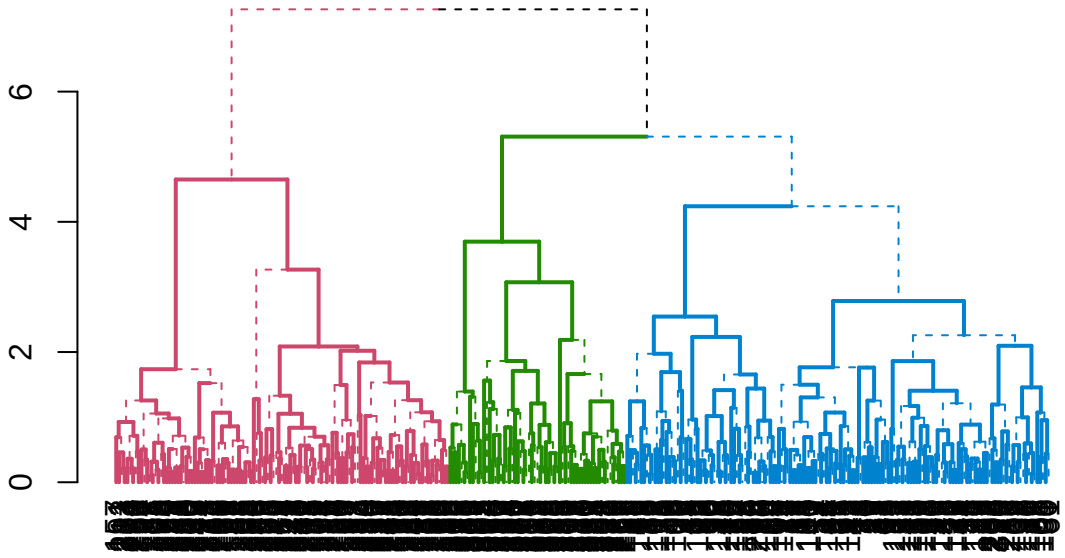
```
penguins_complete <- hclust(penguins_distance, method = "complete")
```

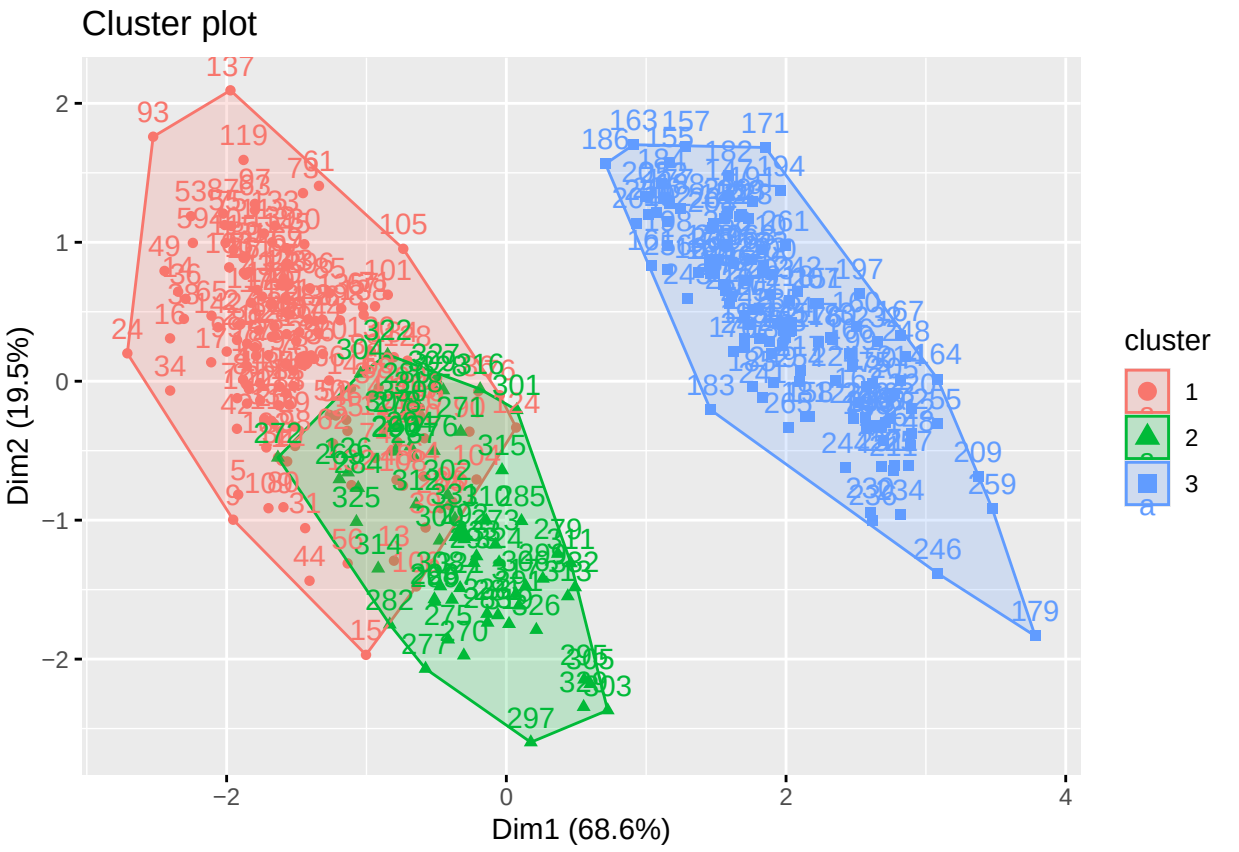
- Partition data into 3 clusters using `cutree()`
  - we specify k = 3 as we want to create three groups representing the 3 species

```
penguins_complete_cut <- cutree(penguins_complete,k = 3)
```

- We will create a dendrogram using `dendextend()` package
- We will create a cluster plot using `factoextra()` package

```
#Dendrogram
penguins_complete_dend <- penguins_complete_dend %>%
          color_branches(k = 3) %>%
          set("branches_lwd", c(2,1,2)) %>%
          set("branches_lty", c(1,2,1))
 plot(penguins_complete_dend)

#Cluster Plot
 fviz_cluster(list(data = penguins_scaled, cluster = penguins_complete_cut))
```


0
2
4
6




```
table(penguins_complete_cut, penguins_cleaned$species)
```

```
|
| penguins_complete_cut Adelie Chinstrap Gentoo
|                     1    145         6      0
|                     2      1        62      0
|                     3      0         0    119
```

Complete Linkage Results:

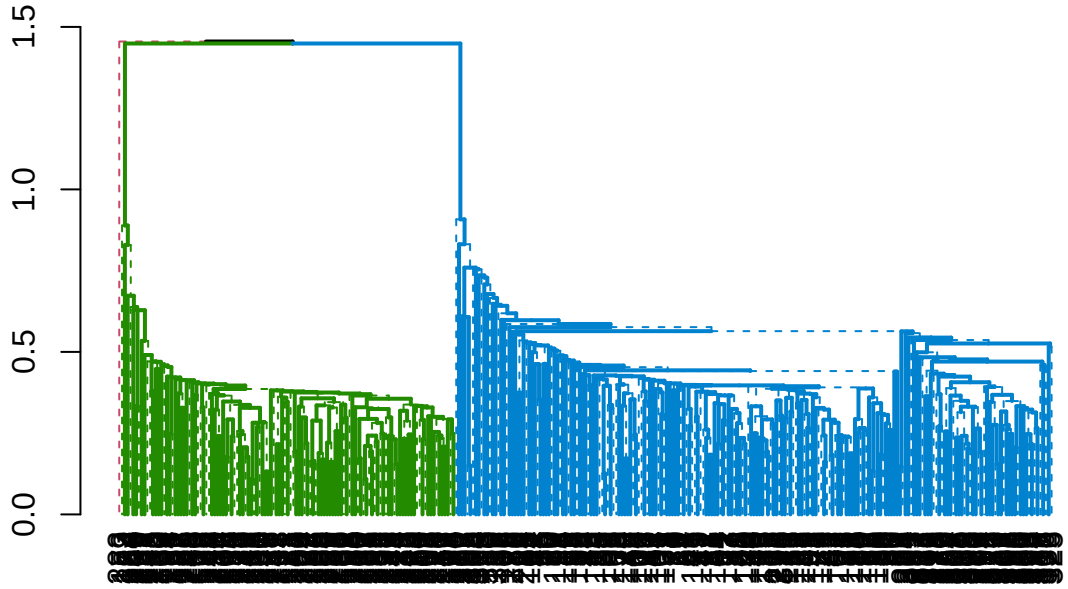
We can see that the clustering was able to find a discernible pattern in measurements for the Gentoo penguin species. Both the table and cluster plot show all 119 observations present in cluster 3. However the same cannot be said for the Adelie and Chinstrap species. The clustering was unable to find a pattern significant enough to isolate all the members of either species. Creating an overlap.

Hierarchical Clustering: Single Linkage

In single linkage the distance between any two points is determined by the minimum distance
- `method = "single"`

```
penguins_scaled <- scale(penguins_clustering)
penguins_distance <- dist(penguins_scaled)
penguins_single <- hclust(penguins_distance, method = "single")

penguins_single_dend <- as.dendrogram(penguins_single)

penguins_single_dend <- penguins_single_dend %>%
```

```
          color_branches(k = 3) %>%
          set("branches_lwd", c(2,1,2)) %>%
          set("branches_lty", c(1,2,1))
plot(penguins_single_dend)
```




```
penguins_single_cut <- cutree(penguins_single,3)


fviz_cluster(list(data = penguins_scaled, cluster = penguins_single_cut))
```


```
table(penguins_single_cut, penguins_cleaned$species)
```

```
|
| penguins_single_cut Adelie Chinstrap Gentoo
|                   1    146        67      0
|                   2      0         0    119
|                   3      0         1      0
```

Single Linkage Results:

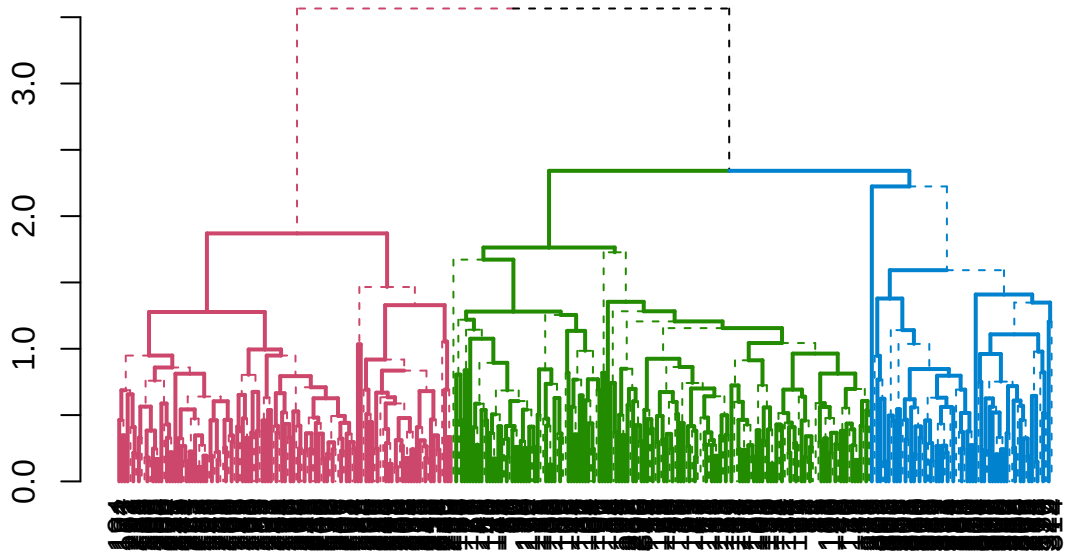
Single linkage performed poorly compared to complete linkage. It was unable to find a pattern to discern Adelie from Chinstrap. Instead it combined both species into one cluster and actually saw them existing as one species rather than two separate ones, with the exception of one outlier grouped into its own cluster. However much like complete linkage, single was also able to completely isolate the Gentoo species.

Hierarchical Clustering: Average Linkage

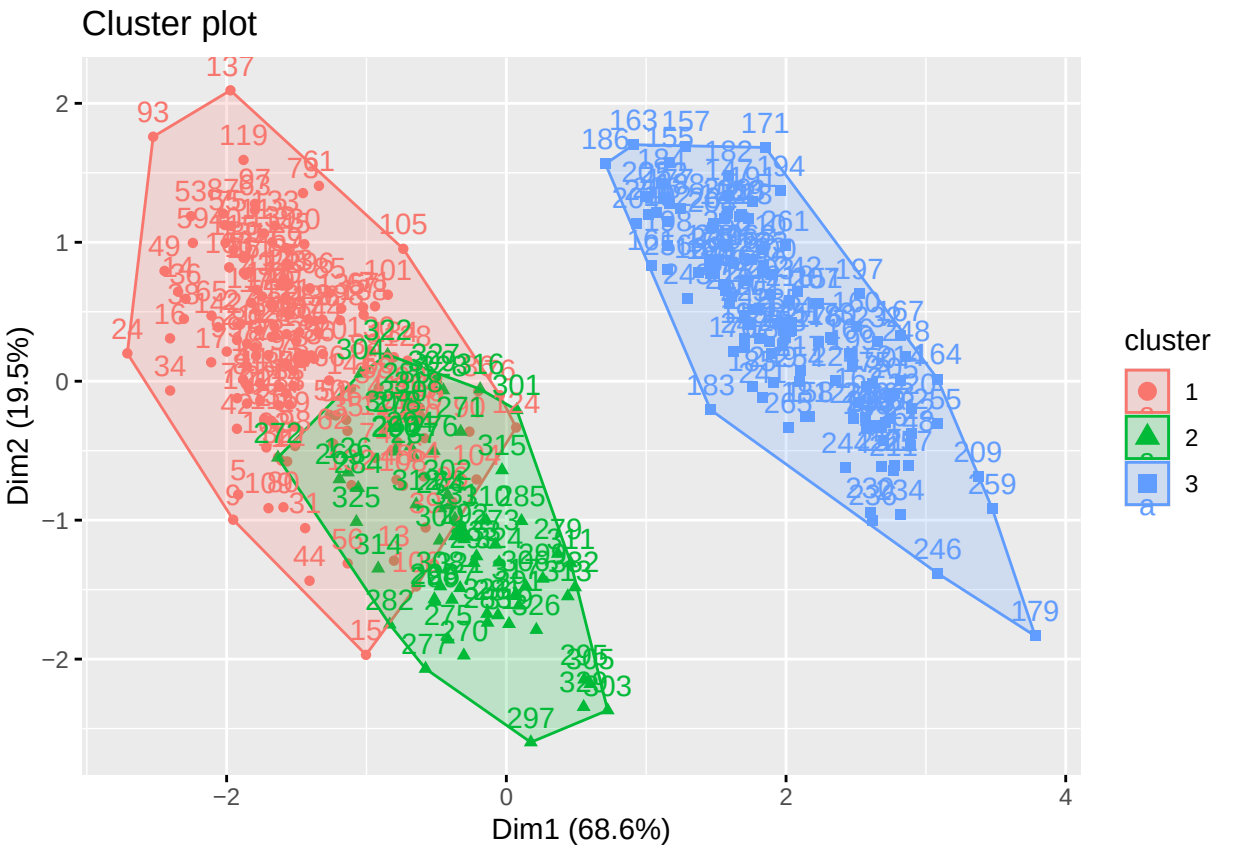
In average linkage the distance between any two points is determined by the average distance
- `method = "average"`

```
penguins_scaled <- scale(penguins_clustering)
penguins_distance <- dist(penguins_scaled)
penguins_avg <- hclust(penguins_distance, method = "average")

penguins_avg_dend <- as.dendrogram(penguins_avg)

penguins_avg_dend <- penguins_avg_dend %>%
```

```
          color_branches(k = 3) %>%
          set("branches_lwd", c(2,1,2)) %>%
          set("branches_lty", c(1,2,1))
plot(penguins_avg_dend)
```




```
penguins_avg_cut <- cutree(penguins_avg,3)


fviz_cluster(list(data = penguins_scaled, cluster = penguins_avg_cut))
```




```
table(penguins_avg_cut, penguins_cleaned$species)
```

```
|
| penguins_avg_cut Adelie Chinstrap Gentoo
|                1    144         5      0
|                2      2        63      0
|                3      0         0    119
```

Average Linkage Results:

Average linkage performed on par with complete linkage. It was also able to isolate the Gentoo species. It created three distinct clusters but had the same faults as complete where the Adelie and Chinstrap species were too closely grouped that there exists overlap for some observations.

Conclusion and Future Analysis

From the tests conducted we can see that the Palmer Penguin data set is no suitable for hierarchical clustering. While it was able to isolate the Gentoo species in all three trials it was not able to replicate those same results for the other two species. The reason why becomes more obvious when we take a look at the mean measurements of the species:

```
| # A tibble: 3 x 5
|   species   flipper_mean bill_depth_mean bill_length_mean body_mass_mean
|   <fct>            <dbl>           <dbl>            <dbl>          <dbl>
| 1 Adelie             190.            18.4             38.8          3706.
| 2 Chinstrap          196.            18.4             48.8          3733.
| 3 Gentoo             217.            15               47.6          5092.
```

Here we can see that the Adelie and Chinstrap Species on average are similar in size and proportions. The Gentoo, while similar in bill depth and length, has such a large margin for body_mass that the model is able to use that as a metric for isolating the species.

So overall unless you are trying to identify two very distinct penguins species, this model is unable to predict species if they are similar in size. In the future I would like to try with a data set that has more quantitaive data. Perhaps the introduction of these variables would be able to improve the model.