Assignment 2 for Statistical Computing and Empirical Mathods

Dr. Henry WJ Reeve

Teaching block 1 2021

Introduction

This document describes your first assignment for Statistical Computing and Empirical Methods (Unit EMATM0061) on the MSc in Data Science. Before starting the assignment it is recommend that you first watch video lectures 3, 4 and 5.

You are encouraged to discuss these questions with your colleagues.

Begin by creating an Rmarkdown document with html output. You are not expected to hand in this piece of work, but it is a good idea to get used to using Rmarkdown.

You will need to install the Tidyverse set of packages if you have not already done so:

```
install.packages("tidyverse")
```

Next load the Tidyverse library which is a superset of the ggplot2 library which we shall use for this assignment.

```
library(tidyverse)
```

For the purpose of this assignment we shall use the Hawks data set from the Stat2Data package. This is a brilliant data set containing information about 908 Hawks collected by students and faculty at Cornell College in Mount Vernon, Iowa. First install the "Stat2Data" library:

```
install.packages("Stat2Data")
```

We can then load the data set as follows.

```
library(Stat2Data)
data("Hawks")
```

This will load a data frame called "Hawks" into your environment. We will use a slightly smaller dataframe which we construct as follows:

```
hawksSmall<-drop_na(select(Hawks,Age,Day,Month,Year,CaptureTime,Species,Wing,Weight,Tail))
```

1 Visualisation

We begin by studying some of the concepts from Lecture 3.

1.1 Types of variables

Check how many rows and columns **hawksSmall** using the **dim()** function. Use the **head()** function to display the top 5 rows of the **hawksSmall** data frame. Your result should look something like this:

```
##   Age Day Month Year CaptureTime Species Wing Weight Tail
## 1   I  19     9 1992       13:30      RT  385    920  219
## 2   I  22     9 1992       10:30      RT  376    930  221
## 3   I  23     9 1992       12:45      RT  381    990  235
## 4   I  23     9 1992       10:50      CH  265    470  220
## 5   I  27     9 1992       11:15      SS  205    170  157
```

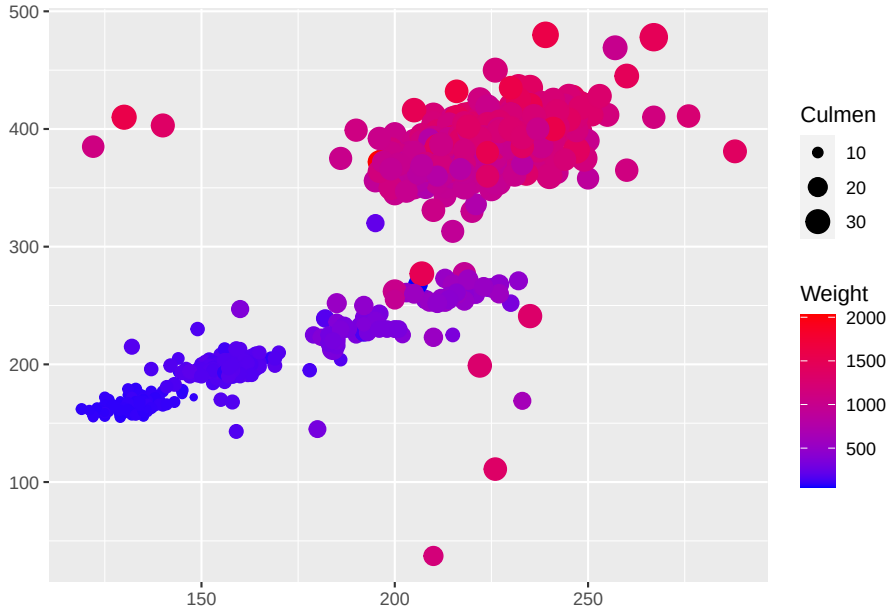
For each of the following variables say whether they continuous, discrete or categorical. Discuss this with your colleagues.

1. Month
2. Species
3. Age
4. Wing
5. Weight

The information available at https://rdrr.io/cran/Stat2Data/man/Hawks.html may help.

1.2 What's wrong with this plot?

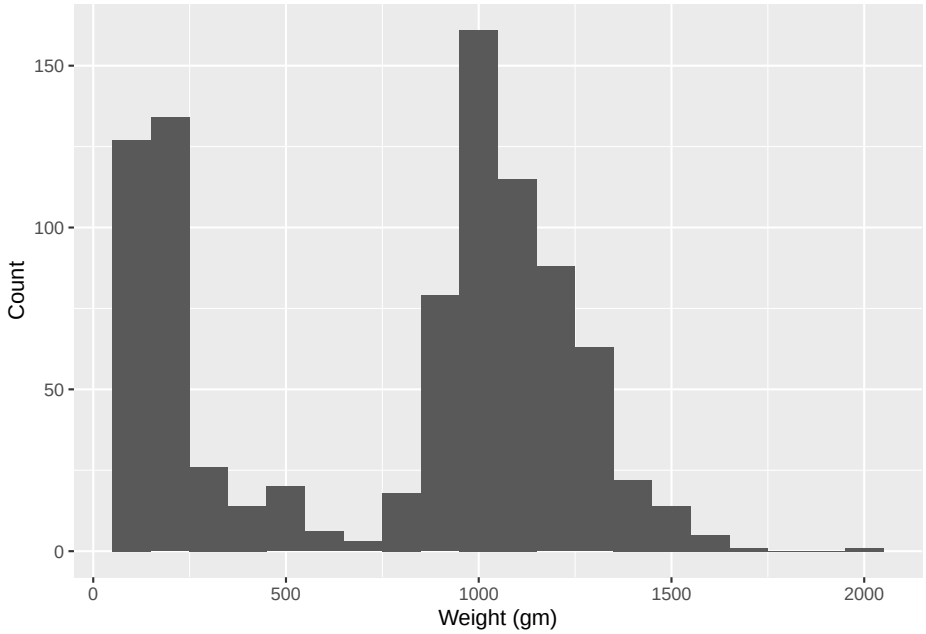
Write down some problems with the plot displayed below.



Hopefully your report doesn't contain plots like this!

1.3 Generate a histogram

Next use a combination of the functions **ggplot()** and **geom_histogram** to create a histogram plot of the weights of the Hawks within the **hawksSmall** data frame with bin widths of 10 grams. Your result should look something like this:

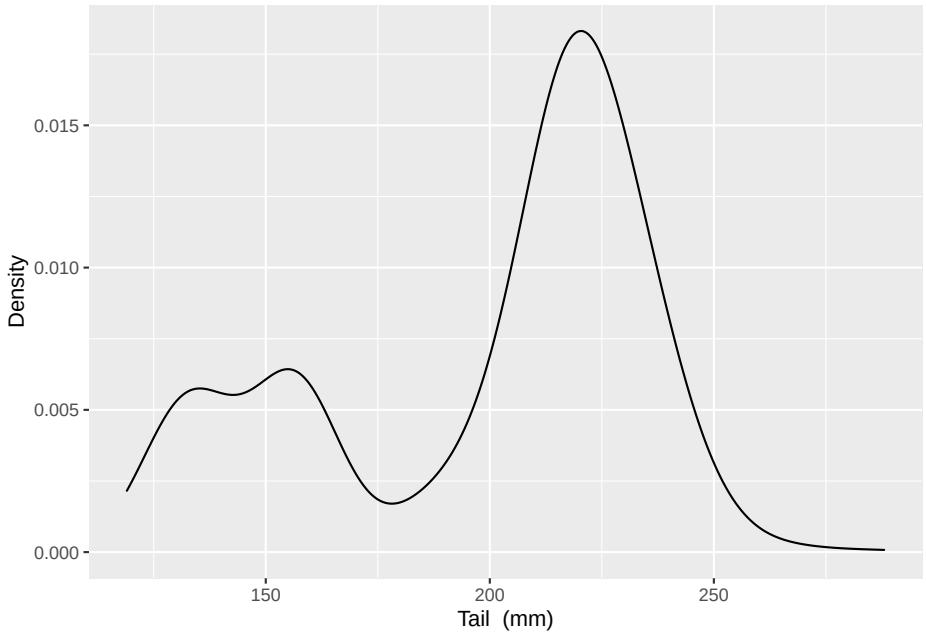


Describe the aesthetic used within this plot.

Which term best describes the shape of the data distribution of Hawk weights: "Unimodal", "Bimodal" or "Trimodal"?

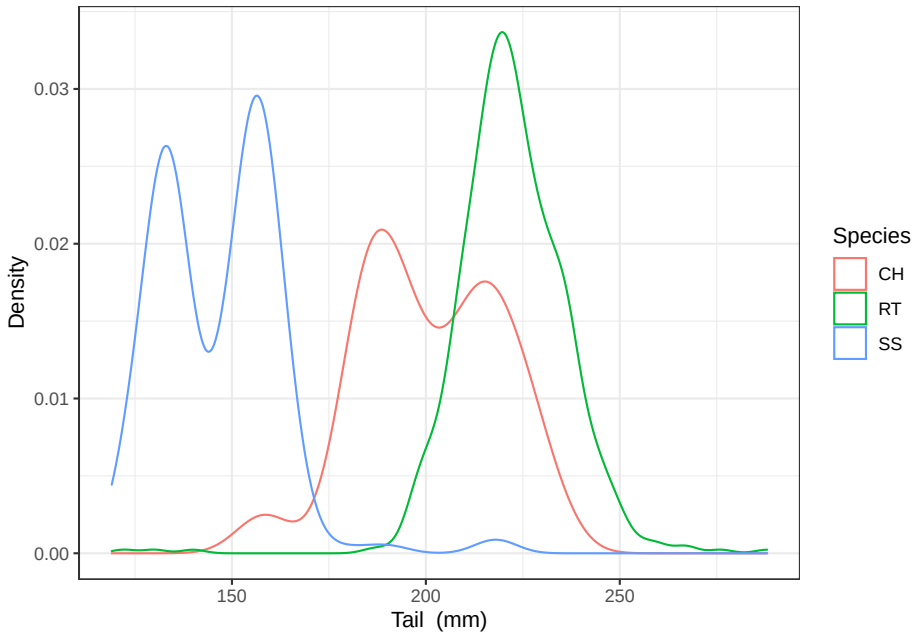
1.4 Generate a density plot

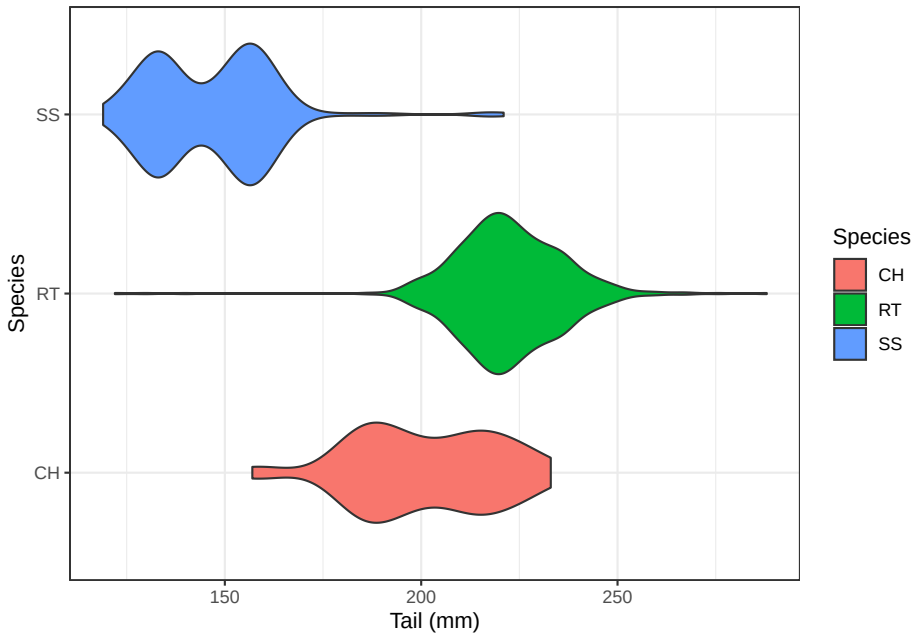
Use a combination of the functions **ggplot()** and **geom_density()** to create a density plot of the tail lengths of the Hawks within the **hawksSmall** data frame. Your result should something like this:



Recreate your plot with the argument `adjust = 0.5` and `adjust = 1`. Describe the role played by the adjust argument within the **geom_density()** function. How many modes does the data distribution of Hawk tail lengths have?

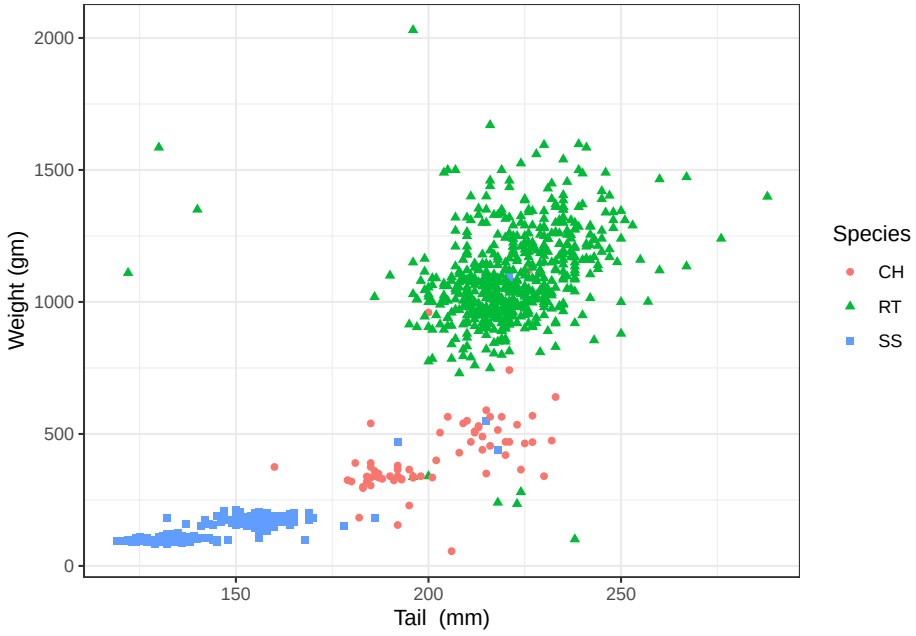
Create the following plots for yourself:





1.5 Scatter plots

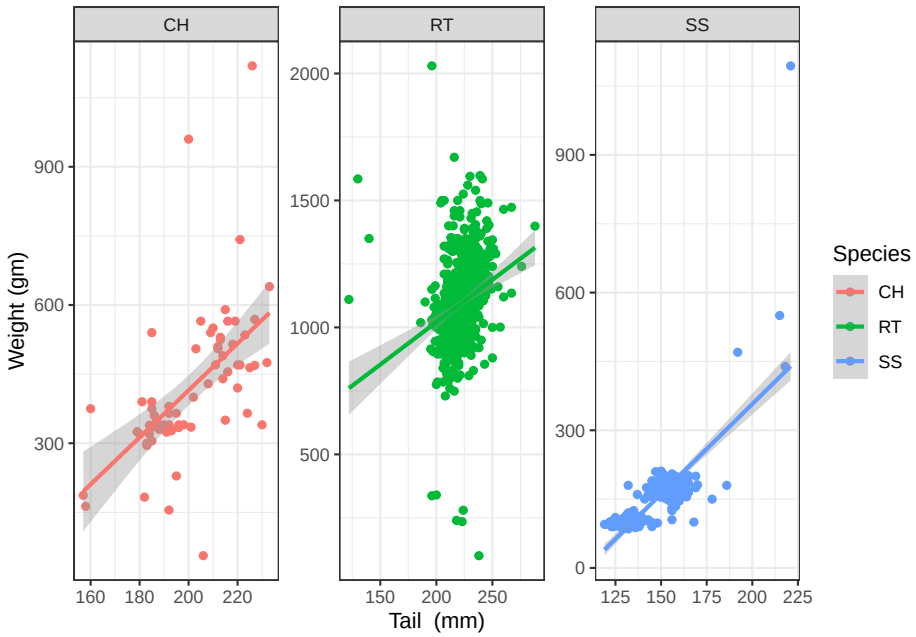
How many aesthetics are present within the following plot? What are the glyphs within this plot?



Generate a plot similar to the above plot using the **ggplot()** and **geom_point()** functions.

1.6 Trend lines and facet wraps

Generate the following plot using the **ggplot()**, **geom_point()**, **geom_smooth()** and **facet_wrap()** functions.



What are the visual cues being used within this plot? Based on the above plot, what can we say about the relationship between the weight of the hawks and their tail lengths?

As an additional challenge, if you have time, you could try adding an annotation to your plot which highlights the weight of the heaviest hawk in the data set.

If you want to learn more about ggplot2 take a look at the ggplot2 gallery https://exts.ggplot2.tidyverse.org/gallery.

2 Data wrangling

We next turn to the data wrangling concepts from Lecture 4.

2.1 Select and filter functions

Use a combination of the **select()** and **filter()** functions to generate a data frame called "hSF" which is a sub-table of the original Hawks data frame with the following characteristics:

1. Your data frame should include the columns:

a) "Wing",

b) "Weight"
c) "Tail".

2. Your data frame should contain a row for every hawk such that:

a) They belong to the species of Red-Tailed hawks
b) They have weight at least 1kg.

Make use of the pipe operator to simplify your code.

How many variables does the dataframe hSF have? What would you say to communicate this information to a Machine Learning practitioner?

How many examples does the dataframe hSF have? How many observations? How many cases?

2.2 The arrange function

Use the **arrange()** function to sort the hSF data frame created in the previous section so that the rows appear in order of increasing wing span.

Use the head command to print out the top five rows of your sorted data frame. Your results should look something like this:

```
##    Wing Weight Tail
## 1  37.2   1180  210
## 2 111.0   1340  226
## 3 199.0   1290  222
## 4 241.0   1320  235
## 5 262.0   1020  200
```

2.3 Join and rename functions

The species of Hawks within the data frame have been indicated via a two letter code. The correspondence between these codes and the full names is given by the following data frame.

```
##   species_code species_name_full
## 1           CH          Cooper's
## 2           RT        Red-tailed
## 3           SS     Sharp-shinned
```

Recreate the above data frame containing the correspondence between codes and the full species names and give your data frame an appropriate name.

Use a combination of the functions **left_join()**, the **rename()** and the **select()* functions to create a new data frame called "hawksFullName" which is the same as the "Hawks" data frame except that the Species column contains the full names rather than the two letter codes.

```
## Joining, by = "Species"
```

Use a combination of the **head()** and **select()** functions to print out the top seven rows of the columns "Species", "Wing" and "Weight" of the data frame called "hawksFullName". Do this without modifying the data frame you just created. Your result should something like this:

```
##         Species Wing Weight
## 1    Red-tailed  385    920
## 2    Red-tailed  376    930
## 3    Red-tailed  381    990
## 4      Cooper's  265    470
## 5 Sharp-shinned  205    170
## 6    Red-tailed  412   1090
## 7    Red-tailed  370    960
```

Does it matter what type of join function you use here?

In what situations would it make a difference?

2.4 The mutate function

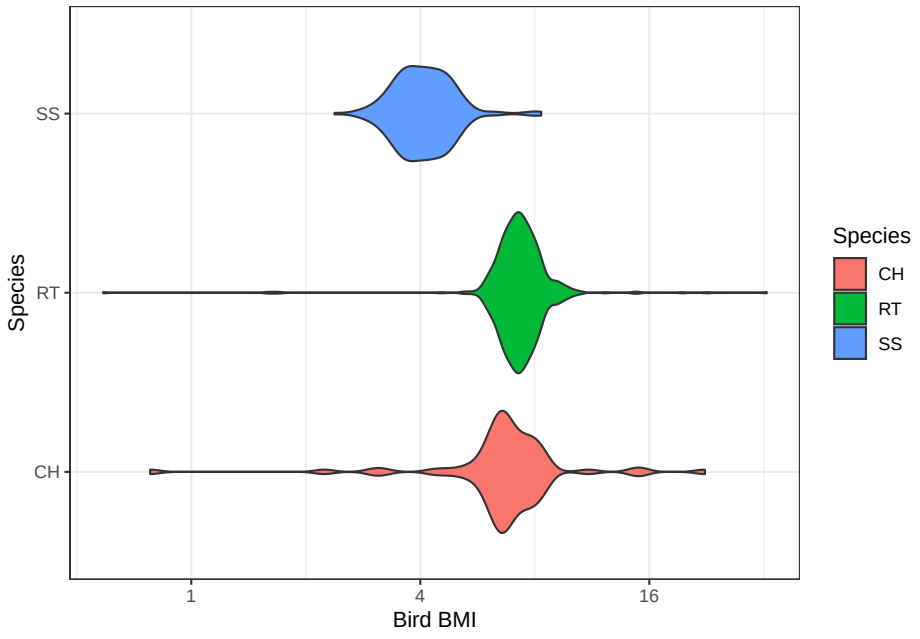
Suppose that the fictitious "Healthy Hawks Society"[1] has proposed a new measure called the "bird BMI" which attempts to measure mass of a hawk standardized by their wing span. The bird BMI is equal to the weight of the hawk (in grams) divided by their wing span (in millimeters) squared. That is,

$$\text{Bird-BMI} := 1000 \times \text{Weight}/\text{Wing-span}^2.$$

Use the **mutate()**, **select()** and **arange()** functions to create a new data frame called "hawksWithBMI" which has the same number of rows as the original Hawks data frame but only two columns - one with their Species and one with their "bird BMI". The rows should appear in descending order of "bird BMI". The top 8 rows of your data frame should look something like this:

```
##   Species  bird_BMI
## 1      RT 852.69973
## 2      RT 108.75741
## 3      RT  32.57493
## 4      RT  22.72688
## 5      CH  22.40818
## 6      RT  19.54932
## 7      CH  15.21998
## 8      RT  14.85927
```

Use the **filter()** function to remove those cases where the bird BMI exceeds 100 from your data frame. Then generate a violin plot of your data which shows the distribution of "bird BMIs" broken down by species. Your result should look something like this:



2.5 Summarize and group-by functions

Using the dataframe "hawksFullName", from problem 3 above, in combination with the **summarize()** and the **groupby** functions, create a summary table, broken down by Hawk species, which contains the following

[1] Both the "Healthy Hawks Society" and the concept of "bird BMI" were made up purely for this assignment.

summary quantities:

1. The number of rows;
2. The mean average wing span in centimeters;
3. The median wing span in centimeters;
4. The trimmed mean average wing span in centimeters (`trim=0.1`);
5. The mean average of the ratio between wing span and tail length.

Your final result should look something like this:

```
## # A tibble: 3 x 6
##   Species       num_rows mn_wing md_wing t_mn_wing tail_wing_ratio
##   <chr>            <int>   <dbl>   <dbl>     <dbl>           <dbl>
## 1 Cooper's            70    244.     240      243.            1.22
## 2 Red-tailed         577    383.     384      385.            1.73
## 3 Sharp-shinned      261    185.     191      184.            1.26
```

Next create a summary table of the following form: Your summary table will show the number of missing values, broken down by species, for the columns Wing, Weight, Culmen, Hallux, Tail, StandardTail, Tarsus and Crop. You can complete this task by combining the **select()**, **group_by()**, **summarize()**, **across()**, **everything()**, **sum()** and **is.na()** functions. You should end with a summary table of the following form:

```
## # A tibble: 3 x 9
##   Species        Wing Weight Culmen Hallux  Tail StandardTail Tarsus  Crop
##   <chr>         <int>  <int>  <int>  <int> <int>        <int>  <int> <int>
## 1 Cooper's          1      0      0      0     0           19     62    21
## 2 Red-tailed        0      5      4      3     0          250    538   254
## 3 Sharp-shinned     0      5      3      3     0           68    233    68
```

You can learn more about dplyr from the Tidyverse website https://dplyr.tidyverse.org/index.html.

3 Exploratory data analysis

We shall now illustrate some concepts from Lecture 5 on Exploratory Data Analysis in the context of our Hawks data set.

3.1 Combining location estimators with the summarise function

Use a combination of the **summarise()**, **mean()** and **median()** to compute the sample mean, sample median and trimmed sample mean (with $q = 0.1$) of the Hawk's wing length and Hawk's weight. Your result should look something like this:

```
##   Wing_mean Wing_t_mean Wing_med Weight_mean Weight_t_mean Weight_med
## 1  315.6375    322.2297      370    772.0802      779.3681        970
```

Combine with the **group_by()** function to obtain a break down by species. Your result should look something like this:

```
## # A tibble: 3 x 7
##   Species Wing_mean Wing_t_mean Wing_med Weight_mean Weight_t_mean Weight_med
##   <fct>       <dbl>       <dbl>    <dbl>       <dbl>         <dbl>      <dbl>
## 1 CH           244.        243.      240        420.          410.       378.
## 2 RT           383.        385.      384       1094.         1089.      1070
## 3 SS           185.        184.      191        148.          140.       155
```

3.2 Location and dispersion estimatiors under linear transformations

Suppose that a variable of interest X_i has values $X_1, \ldots, X_n$. Suppose that $X_1, \ldots, X_n$ has sample mean $\hat{\mu}$. Let $a, b \in \mathbb{R}$ be real numbers and define a new variable $\tilde{X}_i$ with values $\tilde{X}_1, \ldots, \tilde{X}_n$ defined by $\tilde{X}_i = a \cdot X_i + b$ for $i = 1, \cdots, n$. Show that $\tilde{X}_1, \ldots, \tilde{X}_n$ has sample mean $a \cdot \hat{\mu} + b$.

Suppose further that $X_1, \ldots, X_n$ has sample variance S_X^2. What is the sample variance of $\tilde{X}_1, \ldots, \tilde{X}_n$? What is the sample standard deviation of $\tilde{X}_1, \ldots, \tilde{X}_n$?

3.3 Robustness of location estimators

In this exercise we shall investigate the robustness of several location estimators: The sample mean, sample median and trimmed mean.

We begin by extracting a vector called "hal" consisting of the talon lengths of all the hawks with any missing values removed.

```
hal<-Hawks$Hallux # Extract the vector of hallux lengths
hal<-hal[!is.na(hal)] # Remove any nans
```

To investigate the effect of outliers on estimates of location we generate a new vector called "corrupted_hall" with 10 outliers each of value 100 created as follows:

```
outlier_val<-100
num_outliers<-10
corrupted_hal<-c(hal,rep(outlier_val,times=num_outliers))
```

We can then compute the mean of the original sample and the corrupted sample as follows.

```
mean(hal)
```

```
## [1] 26.41086
```

```
mean(corrupted_hal)
```

```
## [1] 27.21776
```

Now let's investigate what happens as the number of outliers changes from 0 to 1000. The code below generates a vector called "means_vect" which gives the sample means of corrupted samples with different numbers of outliers. More precisely, means_vect is a vector of length 1001 with the i-th entry equal to the mean of a sample with $i - 1$ outliers.

```
num_outliers_vect<-seq(0,1000)
means_vect<-c()

for(num_outliers in num_outliers_vect){
  corrupted_hal<-c(hal,rep(outlier_val,times=num_outliers))
  means_vect<-c(means_vect,mean(corrupted_hal))
}
```

Copy and modify the above code to create an additional vector called "medians_vect" of length 1001 with the i-th entry equal to the median of a sample "corrupted_hal" with $i-1$ outliers.

Ammend the code further to add an additional vector called "t_means_vect" of length 1001 with the i-th entry equal to the trimmed mean of a sample with $i-1$ outliers, where the trimmed mean has a trim fraction $q = 0.1$.

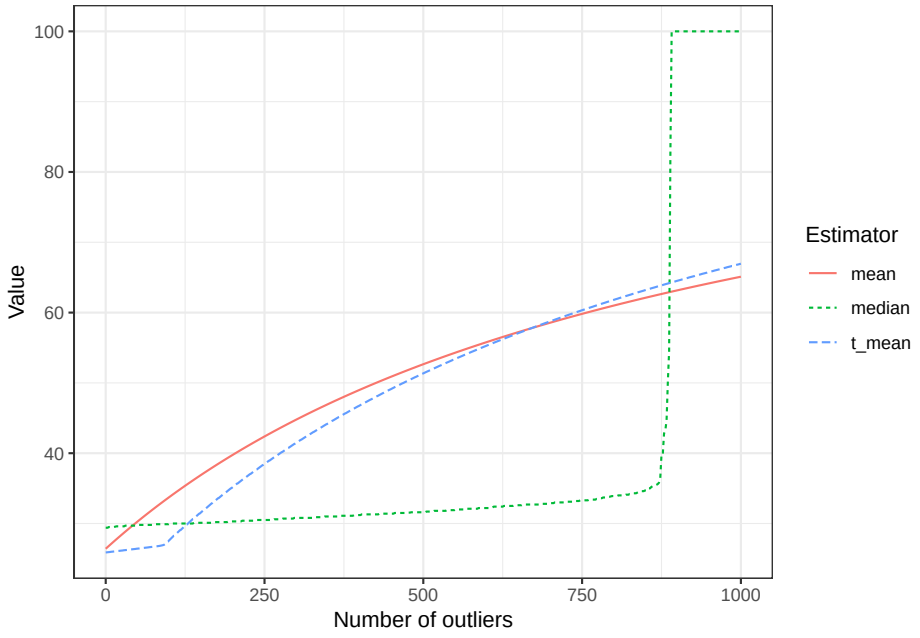
You should now have four vectors: "num_outliers_vect", "means_vect", "medians_vect" and "t_means_vect". Combine these vectors into a data frame with the following code.

```
df_means_medians<-data.frame(num_outliers=num_outliers_vect,
                             mean=means_vect,t_mean=t_means_vect,
                             median=medians_vect)
```

Now use the code below to reshape and plot the data. The function **pivot_longer()** below is used to reshape the data. Don't worry if this operation is unclear at this stage. Its use will be explained soon.

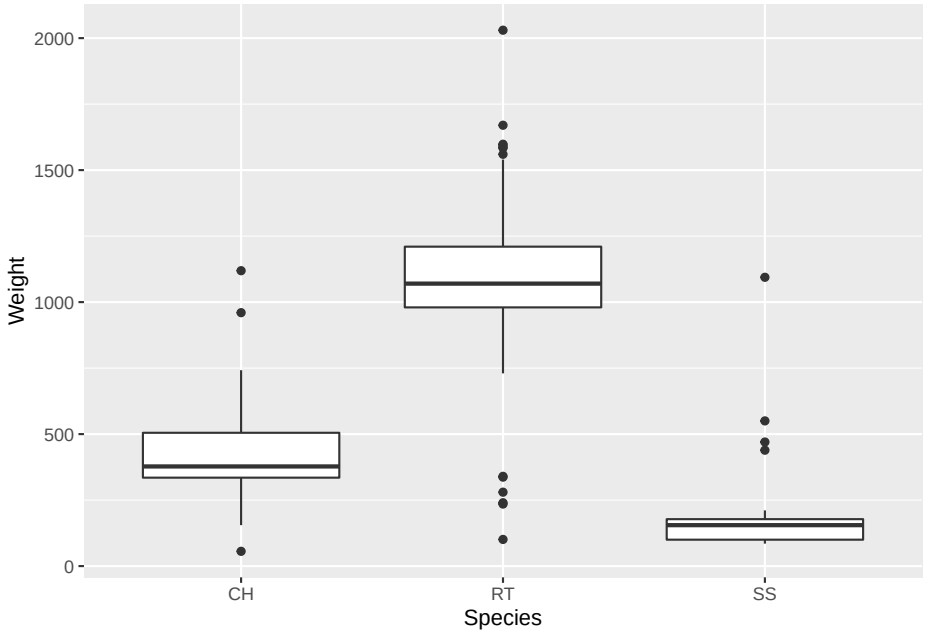
```
df_means_medians%>%
  pivot_longer(!num_outliers, names_to = "Estimator", values_to = "Value")%>%
  ggplot(aes(x=num_outliers,color=Estimator,
                   linetype=Estimator,y=Value))+
  geom_line()+xlab("Number of outliers")
```

The output of your code should look as follows:



3.4 Box plots and outliers

Use the functions **ggplot()** and **geom_boxplot()** to create a box plot which summarises the distribution of hawk weights broken down by species. Your plot should look as follows:



Note the outliers displayed as individual dots.

Suppose we have a sample $X_1, \cdots, X_n$. Let $q25$ denote the 0.25-quantile of the sample and let $q75$ denote the 0.75-quantile of the sample. We can then define the interquartile range, denoted IQR by $IQR := q75 - q25$. In the context of boxplots and outlier X_i is any numerical value such that the following holds if either of the following holds:

$$X_i < q25 - 1.5 \times IQR$$

$$X_i > q75 + 1.5 \times IQR.$$

Create a function called "num_outliers" which computes the number of outliers within a sample (with missing values excluded).

Now combine your function **num_outliers()** with the functions **group_by()** and **summarise()** to compute the number of outlier for the three samples of hawk weights broken down by specied. Your result should look as follows:

```
## # A tibble: 3 x 2
##   Species num_outliers_weight
##   <fct>                 <int>
## 1 CH                        3
## 2 RT                       13
## 3 SS                        4
```

3.5 Covariance and correlation under linear transformations

Suppose that we have a pair of variables: X_i with values $X_1, \ldots, X_n$ and Y_i with values $Y_1, \ldots, Y_n$. Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ have sample covariance $\Sigma_{X,Y}$. Let a, $b \in \mathbb{R}$ be real numbers and define a new variable $\tilde{X}_i$ with values $\tilde{X}_1, \ldots, \tilde{X}_n$ defined by $\tilde{X}_i = a \cdot X_i + b$ for $i = 1, \cdots, n$. In addition, let c, $d \in \mathbb{R}$

be real numbers and define a new variable $\tilde{Y}_i$ with values $\tilde{Y}_1, \ldots, \tilde{Y}_n$ defined by $\tilde{Y}_i = c \cdot Y_i + d$ for $i = 1, \cdots, n$. What is the covariance between the pair of variables $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \ldots, \tilde{Y}_n$?

Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ have correlation $\rho_{X,Y}$. What is the correlation between the pair of variables $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \ldots, \tilde{Y}_n$?