Assignment 3

EMATM0061: Statistical Computing and Empirical Methods, TB1, 2023

Dr. Rihuan Ke

Introduction

Create an R Markdown for the assignment

It is a good practice to use R Markdown to organize your code and results. You can start with the template called `Assignment03_TEMPLATE.Rmd` which can be downloaded via Blackboard.

You can ***optionally*** submit this assignment by 13:00 on 19 October 2023. Note that this assignment will not count towards your final grade. However, it is recommended that you try to answer the questions to gain a better understanding of the concepts. If you want to your solutions, please generate a PDF file. You can either choose the "PDF" option when creating the R Markdown file (note that this option may require LaTex being installed on your computer), or use R Markdown to output an HTML and convert the HTML file into a PDF file with a browser. We only accept PDF files in the submission of this assignment.

Load packages

We need to load two packages, namely `Stat2Data` and `tidyverse`, before answering the questions. If they haven't been installed on your computer, please use `install.packages()` to install them first.

1. Load the tidyverse package:

```
library(tidyverse)
```

2. Load the Stat2Data package and then the dataset Hawks:

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

1. Exploratory data analysis

This section covers some of the concepts from Lecture 7 on Exploratory Data Analysis.

We will use the Hawks dataset that you have loaded.

```
head(Hawks)
```

```
##   Month Day Year CaptureTime ReleaseTime  BandNumber Species Age Sex Wing
## 1     9  19 1992       13:30               877-76317      RT   I      385
## 2     9  22 1992       10:30               877-76318      RT   I      376
## 3     9  23 1992       12:45               877-76319      RT   I      381
## 4     9  23 1992       10:50               745-49508      CH   I   F  265
## 5     9  27 1992       11:15              1253-98801      SS   I   F  205
## 6     9  28 1992       11:25             1207-55910       RT   I      412
##   Weight Culmen Hallux Tail StandardTail Tarsus WingPitFat KeelFat Crop
## 1    920   25.7   30.1  219           NA     NA         NA      NA   NA
## 2    930     NA     NA  221           NA     NA         NA      NA   NA
```

```
## 3    990   26.7   31.3  235              NA      NA           NA       NA   NA
## 4    470   18.7   23.5  220              NA      NA           NA       NA   NA
## 5    170   12.5   14.3  157              NA      NA           NA       NA   NA
## 6   1090   28.5   32.2  230              NA      NA           NA       NA   NA
```

1.1 Location estimators

(Q1) Let's start by computing some location estimators for Hawks' `Tail`.

First, create a vector called `HawksTail`, the elements of which are from the `Tail` column of Hawks data frame. The first part of the vector should look like:

```
## [1] 219 221 235 220 157 230
```

Second, use the `mean` and `median` functions to compute the sample mean and sample median from the vector `HawksTail`. (note that inputs of the `mean` function are vectors. Type ?mean for further details).

1.2 Combining location estimators with the summarise function

(Q1) Use a combination of the `summarise()`, `mean()` and `median()` to compute the sample mean, sample median and trimmed sample mean (with $q = 0.5$) of the Hawk's wing length and Hawk's weight (i.e., the `Wing` and `Weight` columns). You may need to remove the NA values. What can you say by comparing the results of the median and the trimmed mean that you obtain?

Your result should look something like this:

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

(Q2) Combine them with the `group_by()` function to obtain a breakdown 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.         240      240        420.          378.       378.
## 2 RT           383.         384      384       1094.         1070       1070
## 3 SS           185.         191      191        148.          155        155
```

1.3 Location and dispersion estimators under linear transformations

(Q1) Suppose that a variable of interest X has values $X_1, \cdots, X_n$. Suppose that $X_1, \cdots, X_n$ has a sample mean A. Let $a, b \in \mathbb{R}$ be real numbers and define a new variable $\tilde{X}$ with $\tilde{X}_1, \cdots, \tilde{X}_n$ defined by $\tilde{X}_i = aX_i + b$ for $i = 1, 2, \cdots, n$. What is the sample mean of $\tilde{X}_1, \cdots, \tilde{X}_n$ as a function of a, b and A? (please write down your answer as an expression of a, A, and b. You don't need to use R).

Now using the vector `HawksTail` that you created in Section 1.1 as data and letting $a = 2$ and $b = 3$, verify your conclusion using R codes: Compute the mean of `HawksTail*a+b` and then compare it with the one obtained from the mean of `HawksTail` and your conclusion.

(Q2) Suppose further that $X_1, \cdots, X_n$ has sample variance p and standard deviation q. What is the sample variance of $\tilde{X}_1, \cdots, \tilde{X}_n$? What is the sample standard deviation of $\tilde{X}_1, \cdots, \tilde{X}_n$? (Please write down your results.)

Now using the vector `HawksTail` that you created in Section 1.1 as data and letting $a = 2$ and $b = 3$, verify your result using R codes again.

1.4 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))
}
```

(Q1) *Sample median*:

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.

(Q2) *Sample trimmed mean*:

Amend 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 $q = 0.1$.

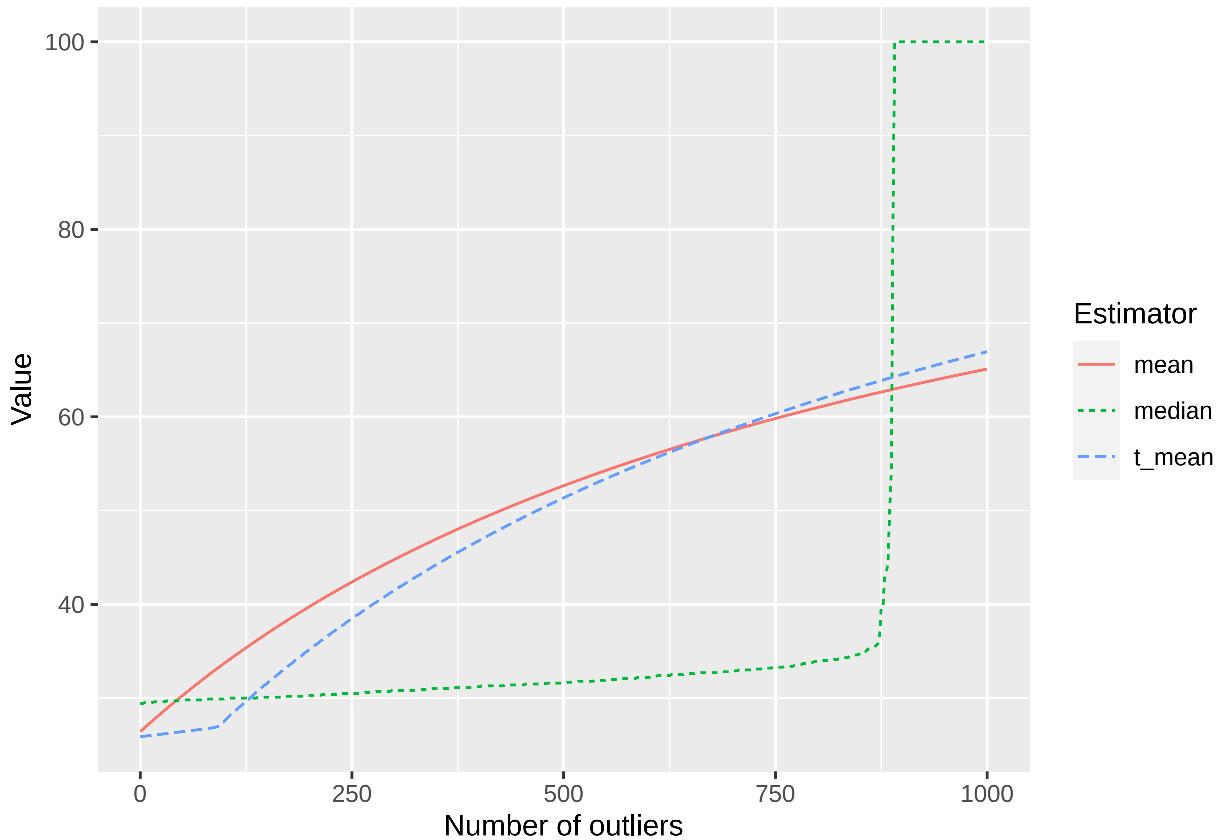
(Q3) *Visualisation*

Now you should have the 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. Recall that the function `pivot_longer()` below is used to reshape the data. Your result should look like:

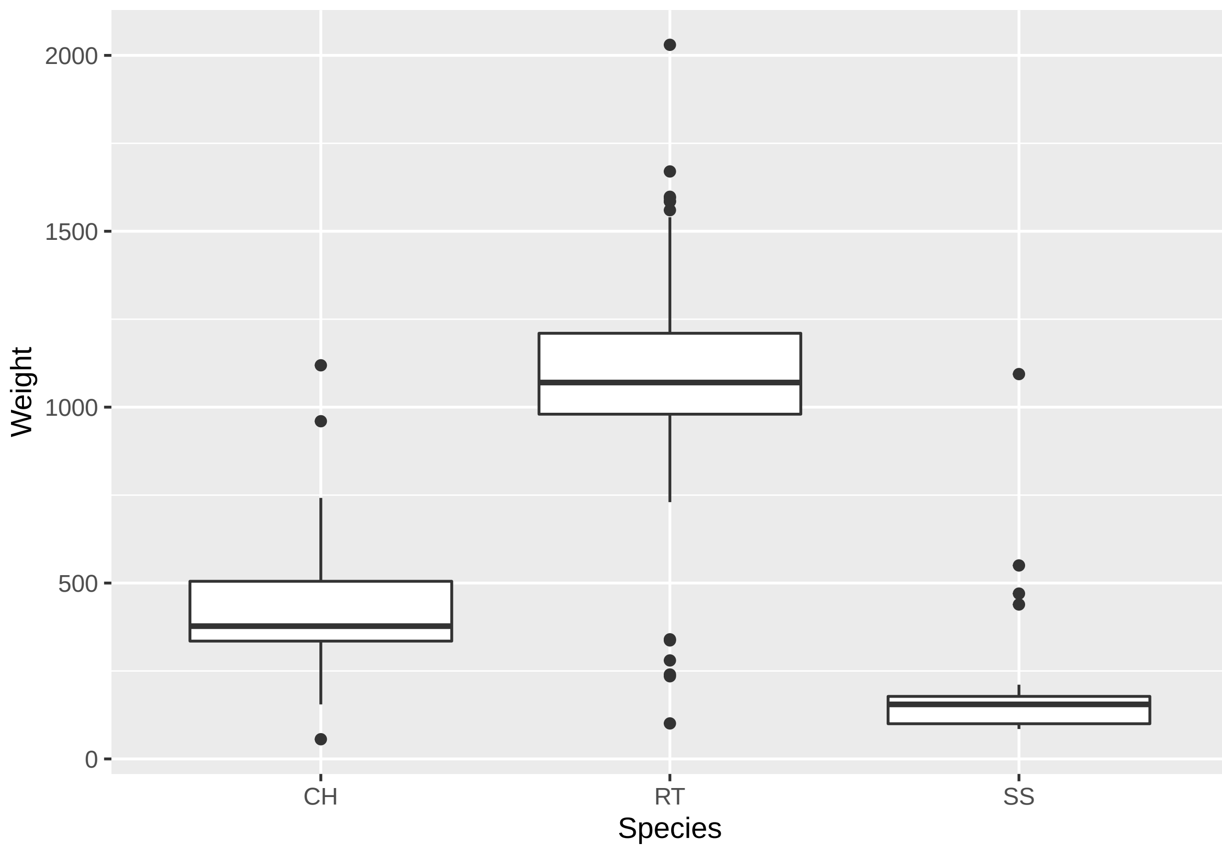
```r
df_means_medians %>%
  pivot_longer(!num_outliers, names_to = "Estimator", values_to = "Value") %>%
  ggplot(aes(x=num_outliers,color=Estimator, linetype=Estimator,y=Value)) +
  geom_line()+xlab("Number of outliers")
```



Which quantity is the most robust when the number of outliers is small? (Note that, in this experiment, the term outliers simply means the artificial data used to corrupt the vector. It is not related to the outliers computed in the next question).

1.5 Box plots and outliers

(Q1) 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 are displayed as individual dots.

(Q2) *quantile and boxplots*

Compute the 0.25-quantile, 0.5-quantile, 0.75-quantile of the `Weight` grouped by Species. Your results should look like this

```
## # A tibble: 3 x 4
##   Species quantile025 quantile050 quantile075
##   <fct>         <dbl>       <dbl>       <dbl>
## 1 CH              335        378.         505
## 2 RT              980       1070         1210
## 3 SS              100        155          178.
```

Now compare these values with the boxplot above. Can you explain which parts of the boxplot these numbers correspond to?

(Q3) *Outliers*

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, an outlier X_i is any numerical value such that the following holds if either of the following holds:

$$X_i < \text{q25} - 1.5 \times \text{IQR}, \quad \text{or}$$
$$X_i > \text{q75} + 1.5 \times \text{IQR}.$$

Create a function called `"num_outliers"` which computes the number of outliers within a sample (with missing values excluded). The function should take a vector as input as output a number.

Test your "num_outliers" function using the code below:

```
num_outliers( c(0, 40,60,185))
```

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

(Q4) *Outliers by group*

Now combine your function `num_outliers()` with the functions `group_by()` and `summarise()` to compute the number of outliers for the three samples of hawk weights broken down by species. 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
```

You may want to go back to the above box plot to check the number of dots displayed for each group.

1.6 Covariance and correlation under linear transformations

(Q1) Compute the covariance and correlation between the `Weight` and `Wing` of the `Hawks` data. You can use the cov and cor functions.

(Q2) Suppose that we have a pair of variables: X with values $X_1, \cdots, X_n$ and Y with values $Y_1, \cdots, Y_n$. Suppose that $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ have the sample covariance S and correlation R. Let $a, b \in \mathbb{R}$ be real numbers and define a new variable $\tilde{X}$ with $\tilde{X}_1, \cdots, \tilde{X}_n$ defined by $\tilde{X}_i = aX_i + b$ for $i = 1, 2, \cdots, n$. In addition, let $c, d \in \mathbb{R}$ be real numbers and define a new variable $\tilde{Y}$ with $\tilde{Y}_1, \cdots, \tilde{Y}_n$ defined by $\tilde{Y}_i = cY_i + d$.

What is the covariance between $\tilde{X}_1, \cdots, \tilde{X}_n$ and $\tilde{Y}$ with $\tilde{Y}_1, \cdots, \tilde{Y}_n$ (as a function of S, a, b, c, d)? Assuming that $a \neq 0$ and $c \neq 0$, what is the correlation between $\tilde{X}_1, \cdots, \tilde{X}_n$ and $\tilde{Y}$ with $\tilde{Y}_1, \cdots, \tilde{Y}_n$? Please write down the mathematical expressions.

Let $a = 2.4, b = 7.1, c = -1, d = 3$, and let X be the hawk's weight and Y be the hawk's Wing. Verify your conclusion with R codes in a similar way to Section 1.3 (Q1).

2. Random experiments, events and sample spaces, and the set theory

In this exercise, we will learn about Random experiments, events and sample spaces and set theory that were introduced in Lecture 8.

In this section, you are not required to compute your results using R codes. If you want to write math formulas in R-markdown, the document called "Assignment_R MarkdownMathformulasandSymbolsExamples.rmd" provides a list of examples for your reference.

2.1 Random experiments, events and sample spaces

(Q1) Firstly, write down the definition of a random experiment, event and sample space.

(Q2) Consider a random experiment of rolling a dice twice. Give an example of what is an event in this random experiment. Also, can you write down the sample space as a set? What is the total number of different events in this experiment? Is the empty set considered as an event?

2.2 Set theory

Remember that a set is just a collection of objects. All that matters for the identity of a set is the objects it contains. In particular, the elements within the set are unordered, so for example the set {1, 2, 3} is exactly the same as the set {3, 2, 1}. In addition, since sets are just collections of objects, each object can only be either included or excluded and multiplicities do not change the nature of the set. In particular, the set {1, 2, 2, 2, 3, 3} is exactly the same as the set A = {1, 2, 3}. In general there is no concept of "position" within a set, unlike a vector or matrix.

(Q1) Set operations:

Let the sets A, B, C be defined by A := {1, 2, 3}, B := {2, 4, 6}, C := {4, 5, 6}.

1. What are the unions $A \cup B$ and $A \cup C$?
2. What are the intersections $A \cap B$ and $A \cap C$?
3. What are the complements $A \backslash B$ and $A \backslash C$?
4. Are A and B disjoint? Are A and C disjoint?
5. Are B and $A \backslash B$ disjoint?
6. Write down an arbitrary partition of {1,2,3,4,5,6} consisting of two sets. Also, write down another partition of {1,2,3,4,5,6} consisting of three sets.

(Q2) Complements, subsets and De Morgan's laws

Let Ω be a sample space. Recall that for an event $A \subseteq \Omega$ the complement $A^c := \Omega \backslash A := \{w \in \Omega : w \notin A\}$. Take a pair of events $A \subseteq \Omega$ and $B \subseteq \Omega$.

1. Can you give an expression for $(A^c)^c$ without using the notion of a complement?
2. What is Ω^c?
3. (Subsets) Show that if $A \subseteq B$, then $B^c \subseteq A^c$.
4. (De Morgan's laws) Show that $(A \cap B)^c = A^c \cup B^c$. Let's suppose we have a sequence of events $A_1, A_2, \cdots, A_K \subseteq \Omega$. Can you write out an expression for $(\cap_{k=1}^K A_k)^c$?
5. (De Morgan's laws) Show that $(A \cup B)^c = A^c \cap B^c$.
6. Let's suppose we have a sequence of events $A_1, A_2, \cdots, A_K \subseteq \Omega$. Can you write out an expression for $(\cup_{k=1}^K A_k)^c$?

(Q3) Cardinality and the set of all subsets:

Suppose that $\Omega = \{w_1, w_2, \cdots, w_K\}$ contains K elements for some natural number K. Here Ω has cardinality K.

Let E be a set of all subsets of Ω, i.e., $E := \{A | A \subset \Omega\}$. Note that here E is a set. Give a formula for the cardinality of E in terms of K.

(Q4) Disjointness and partitions.

Suppose we have a sample space Ω, and events A_1, A_2, A_3, A_4 are subsets of Ω.

1. Can you think of a set which is disjoint from every other set? That is, find a set $A \subseteq \Omega$ such that $A \cap B = \emptyset$ for all $B \subseteq \Omega$.
2. Define events $S_1 := A_1$, $S_2 = A_2 \backslash A_1$, $S_3 = A_3 \backslash (A_1 \cup A2)$, $S_4 = A_4 \backslash (A_1 \cup A2 \cup A_3)$. Show that S_1, S_2, S_3, S_4 form a partition of $A_1 \cup A_2 \cup A_3 \cup A_4$.

(Q5) Indicator function.

Suppose we have a sample space Ω, and the event A is a subset of Ω. Let $\mathbf{1}_A$ be the indicator function of A.

1. Write down the indicator function $\mathbf{1}_{A^c}$of A^c (use $\mathbf{1}_A$ in your formula).
2. Can you find a set B whose indicator function is $\mathbf{1}_{A^c} + \mathbf{1}_A$?
3. Recall that $\mathbf{1}_{A \cap B} = \mathbf{1}_A \cdot \mathbf{1}_B$ and $\mathbf{1}_{A \cup B} = \max(\mathbf{1}_A, \mathbf{1}_B) = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_A \cdot \mathbf{1}_B$ for any $A \subseteq \Omega$ and $B \subseteq \Omega$. Combining this with the conclusion from Question (Q5) 1, use indicator functions to prove $(A \cap B)^c = A^c \cup B^c$ (De Morgan's laws).

(Q6) Uncountable infinities (this is an optional extra).

This is a challenging optional extra. You may want to return to this question once you have completed all other questions.

Show that the set of numbers $\Omega := [0, 1]$ is uncountably infinite.

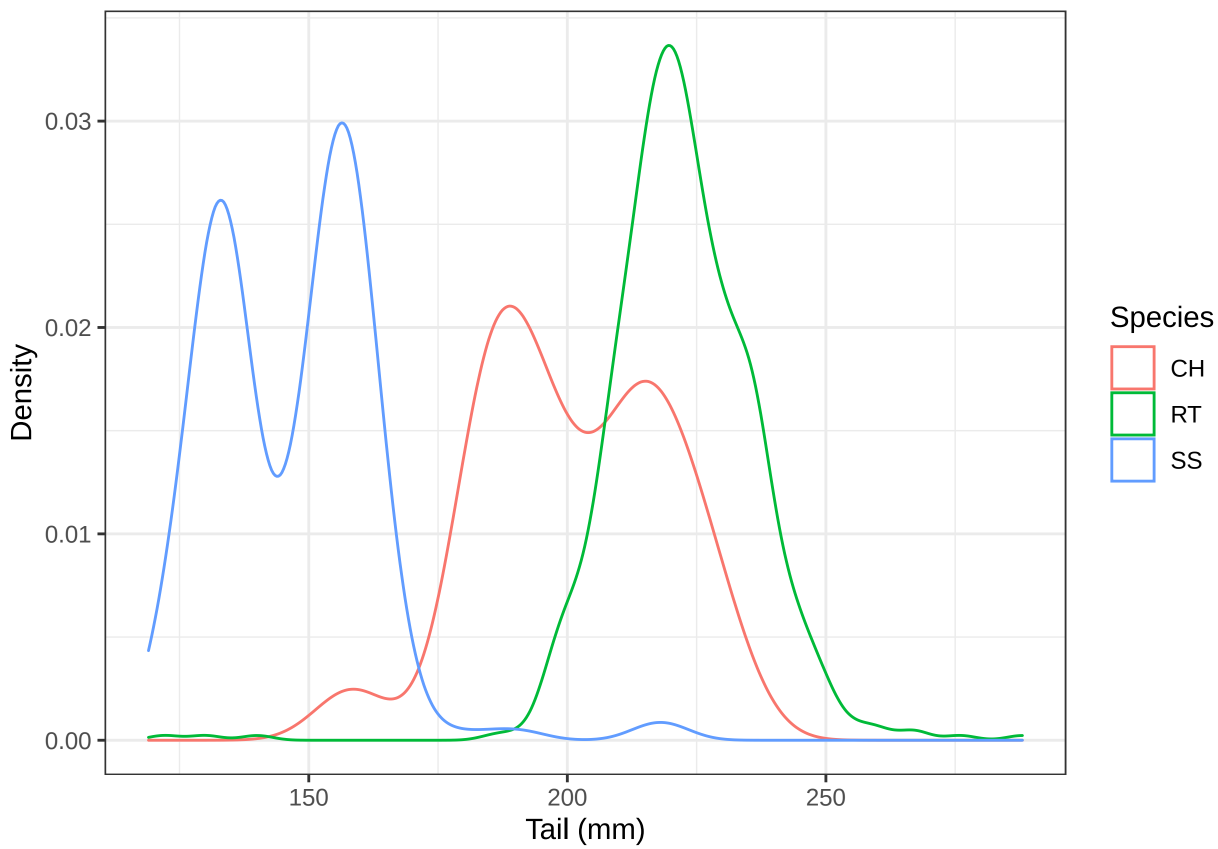
3. Visualisation

The last part of this assignment is a continuation of the visualisation experiment in Assignment 2 and covers parts of the concepts of data visualisation from Lecture 6.

In Assignment 2, we have learned how to create univariate plots using ggplot2 histogram and density plot functions. In this assignment, we will explore bivariate and multivariate plots.

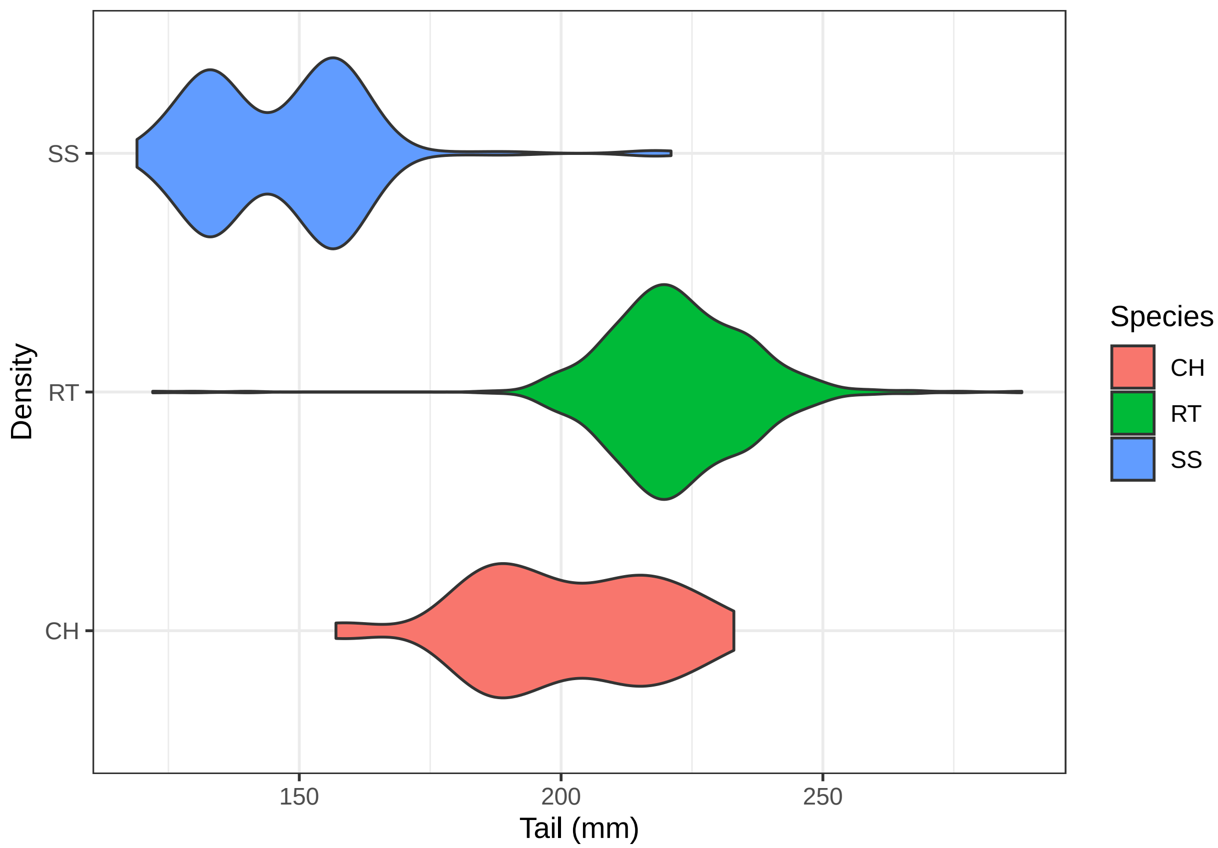
(Q1) Density plot:

Use the ggplot and geom_density() functions to create the following density plot for the three species.



(Q2) Violin plot:

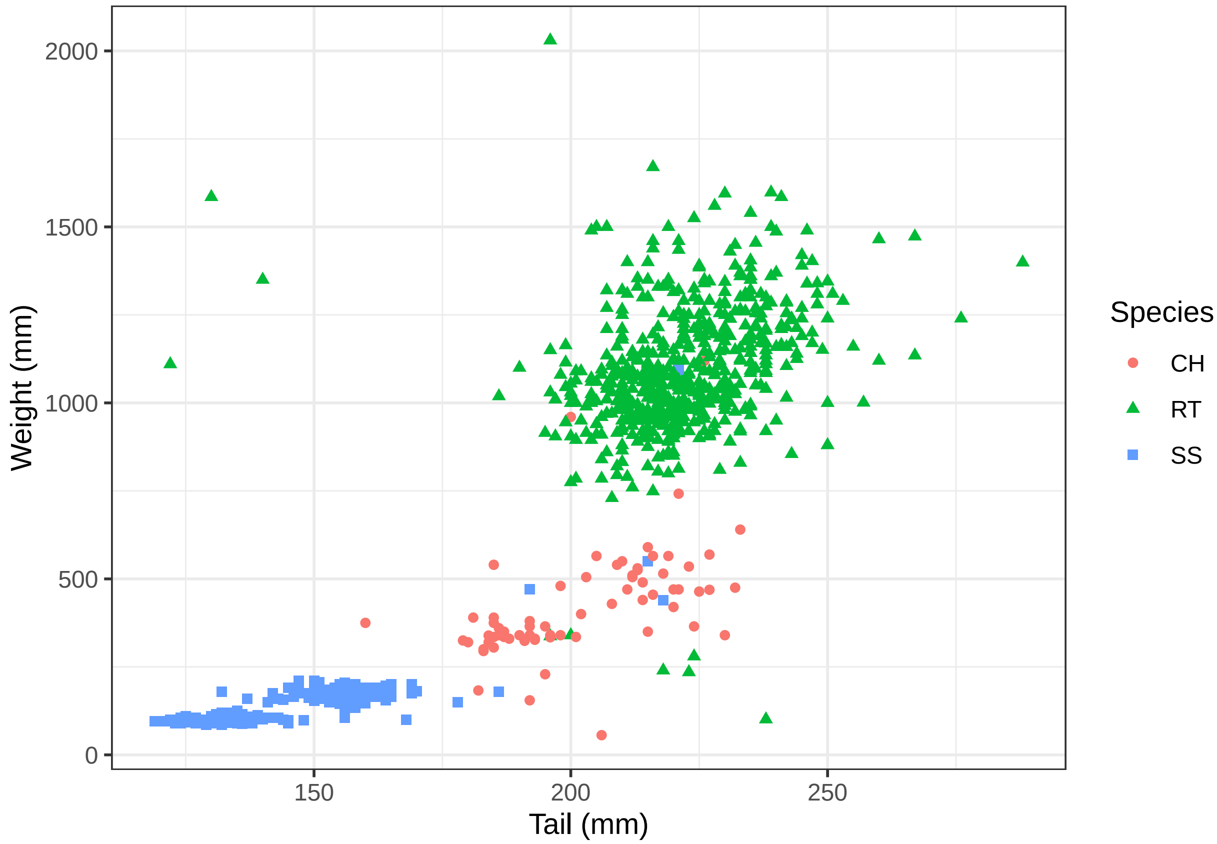
Use the ggplot and geom_violin() functions to create the following violin plot for the three species.



Q(3) Scatter plot

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

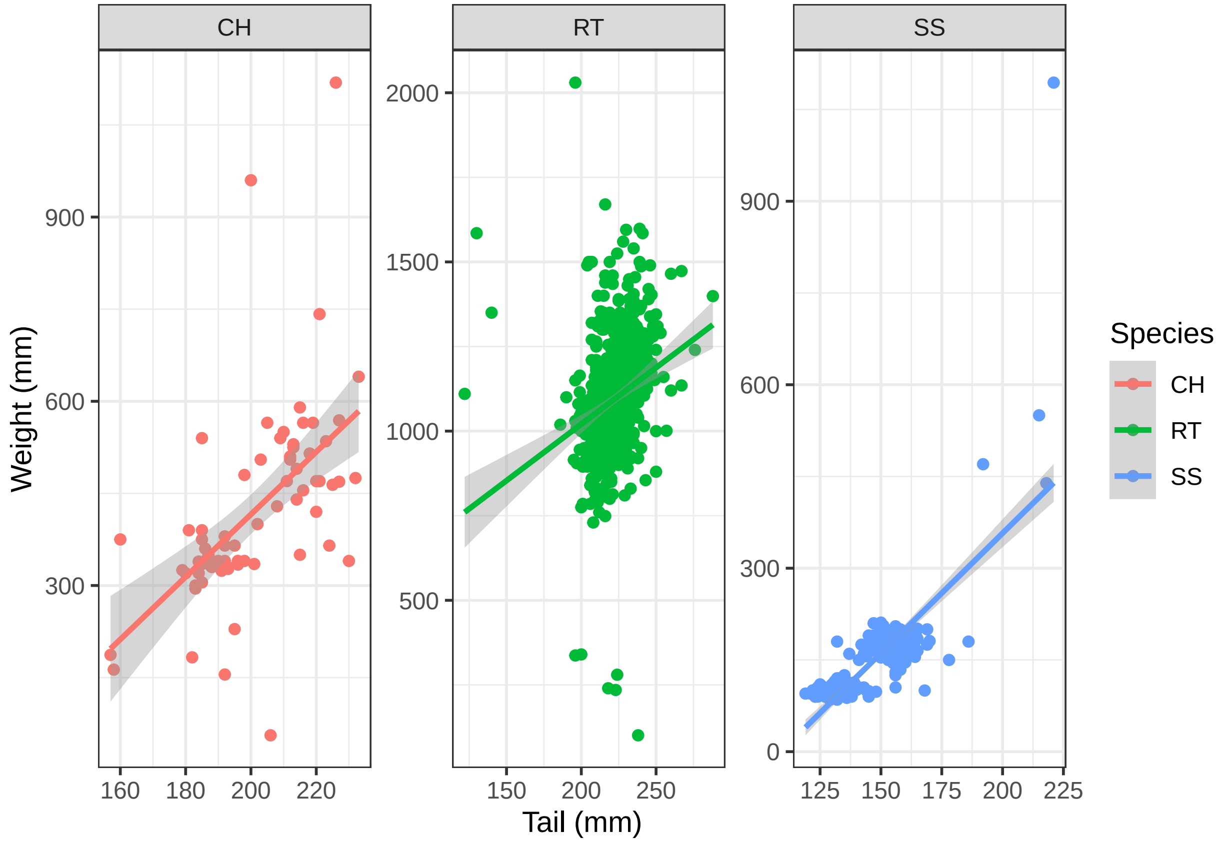
1. How many aesthetics are present within the following plot?
2. What are the glyphs within this plot?
3. What are the visual cues being used within this plot?



Q(4) Trend lines and facet wraps:

Generate the following plot using the ggplot(), geom_point(), geom_smooth() and facet_wrap() functions. Note that in the facet plot, the three panels use different scales.

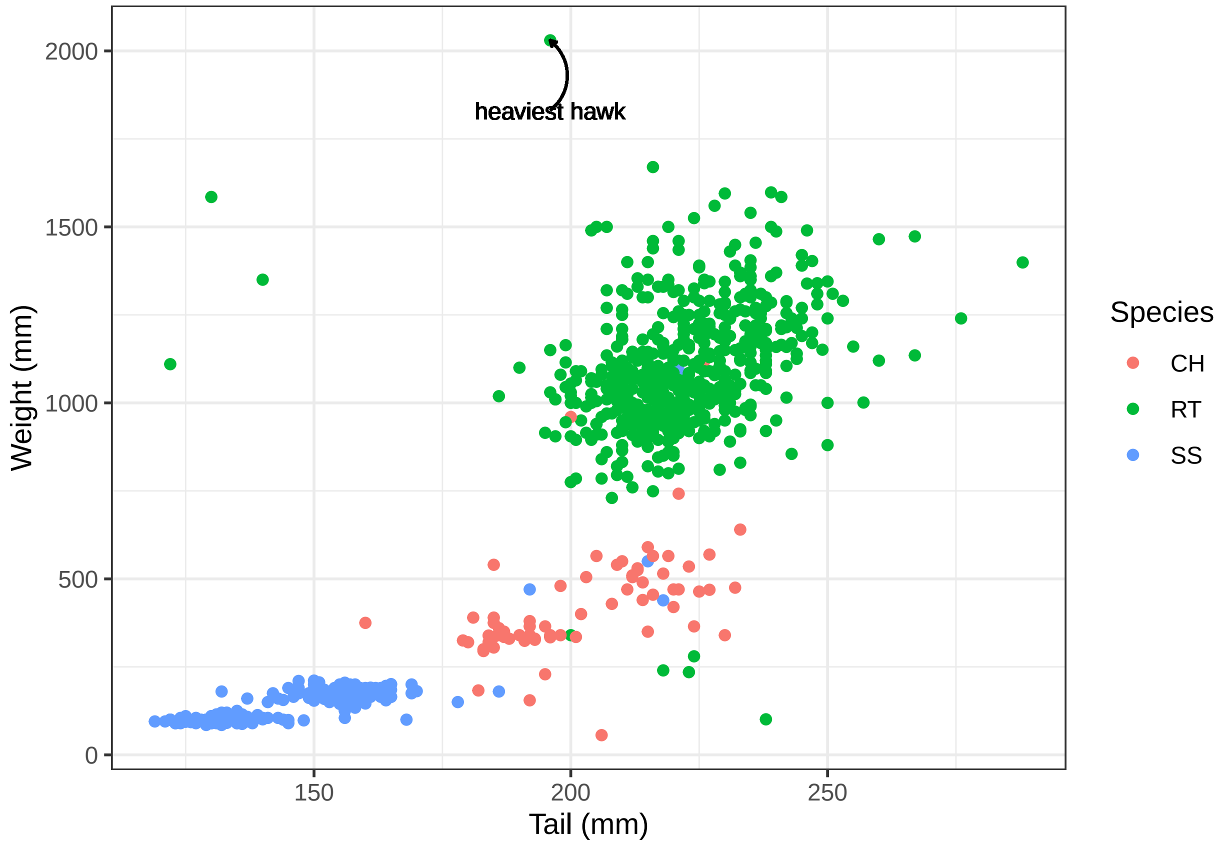
1. What are the visual cues being used within this plot?
2. Based on the plot below, what can we say about the relationship between the weight of the hawks and their tail lengths?



Q(5) Adding annotations

First, compute the `Weight` and the `Tail` of the heaviest hawk in the dataset. You can use filter() and select() function to select proper data.

Second, reuse the code that you create from Q(3), adding an arrow and an annotation to indicate the heaviest hawk. Your result should look similar to this:


heaviest hawk
Weight (mm)
2000
1500
1000
500
0
150
200
250
Tail (mm)
Species
CH
RT
SS