Assignment 3

Yujie Wang

2022-10-12

1. Exploratory data analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

- Use Hawks dataset:

```
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 (Q1)

- Let's start by computing some *location estimators* for Hawks' Tail.

1. create a vector called HawksTail, the elements of which are from the Tail column of Hawks data frame.

```
HawksTail<-Hawks%>%select(Tail)%>%unlist()%>%as.vector()
```

2. use the mean and median functions to compute the sample mean and sample median from the vector HawksTail.

```
HawksTail_mean<-mean(HawksTail,na.rm=TRUE)
HawksTail_median<-median(HawksTail,na.rm=TRUE)
print("HawksTail's mean is:")
```

```
## [1] "HawksTail's mean is:"
```

```
print(HawksTail_mean)
```

```
## [1] 198.8315
```

```
print("HawksTail's median is:")
```

```
## [1] "HawksTail's median is:"
```

```
print(HawksTail_median)
```

```
## [1] 214
```

1.2 (Q1)

1. a combination of the summarise(), mean() and median() to compute the sample mean, sample median and trimmed sample mean (with q = 0.5)
2. the Hawk's wing length and Hawk's weight (i.e., the Wing and Weight columns).
3. You may need to remove the NA values.

```
Hawks%>%
  summarise(Wing_mean=mean(Wing,na.rm=TRUE),
            Wing_t_mean=mean(Wing,na.rm=TRUE,trim=0.5),
            Wing_med=median(Wing,na.rm=TRUE),
            Weight_mean=mean(Weight,na.rm=TRUE),
            Weight_t_mean=mean(Weight,na.rm=TRUE,trim=0.5),
            Weight_med=median(Weight,na.rm=TRUE))
```

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

- What can you say by comparing the results of the median and the trimmed mean that you obtain?

Answer: The trimmed mean is equivalent to the median after removing half percent of the data means that the median and trimmed mean is more robust than mean.

1.2 (Q2)

- Combine them with the group_by() function to obtain a breakdown by species.

```
Hawks%>%group_by(Species)%>%
  summarise(Wing_mean=mean(Wing,na.rm=TRUE),
            Wing_t_mean=mean(Wing,na.rm=TRUE,trim=0.5),
            Wing_med=median(Wing,na.rm=TRUE),
            Weight_mean=mean(Weight,na.rm=TRUE),
            Weight_t_mean=mean(Weight,na.rm=TRUE,trim=0.5),
            Weight_med=median(Weight,na.rm=TRUE))
```

```
## # 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 (Q1)

Question:

1. Suppose that a variable of interest X has values $X_1, ...X_n$. Suppose that $X_1, ..., X_n$ has a sample mean A.
2. Let a, b $\in$ R be real numbers and define a new variable $\tilde{X}$ with $\tilde{X}_1, ..., \tilde{X}_n$ defined by $\tilde{X}_i = a \cdot X_i + b \, .for \, i = 1, 2, ..., n.$
3. What is the sample mean of $\tilde{X}_1, ..., \tilde{X}_n$ as a function of a,b and A?

Answer: $mean \, of \, \tilde{X} = a * A + b$

Verification:

```
a<-2
b<-3
#computed by my conclusion
HawksTailmean1<-HawksTail_mean*a+b
print(HawksTailmean1)
```

```
## [1] 400.663
```

```
#computed by HawksTail separately
compute_mean<-function(x){
  return(x*a+b)
}
HawksTail1<-map_dbl(HawksTail,compute_mean)
HawksTailmean2<-mean(HawksTail1,na.rm=TRUE)
print(HawksTailmean2)
```

```
## [1] 400.663
```

The two results are equivalent.

1.3 (Q2)

Question:

Suppose further that $X_1, ...X_n$ has sample variance p and standard deviation q. What is the sample variance of $\tilde{X}_1, ..., \tilde{X}_n$? What is the sample standard deviation of $\tilde{X}_1, ..., \tilde{X}_n$?

Answer:

$Variance\ of\ \tilde{X} = a^2 * p$

$Deviation\ of\ \tilde{X} = a * q$

Verification:

```
a<-2
b<-3
HawksTail_variance<-var(HawksTail,na.rm=TRUE)
HawksTail_deviation<-sd(HawksTail,na.rm=TRUE)
#computer by my conclusion
HawksTailvariance1<-HawksTail_variance*a^2
HawksTaildeviation1<-HawksTail_deviation*a
print("computed by my answer:")
```

```
## [1] "computed by my answer:"
```

```
print(paste("Variance1:",HawksTailvariance1))
```

```
## [1] "Variance1: 5424.1465012701"
```

```
print(paste("Deviation1:",HawksTaildeviation1))
```

```
## [1] "Deviation1: 73.6488051584688"
```

```
#computed by HawksTail separately
HawksTailvariance2<-var(HawksTail1,na.rm=TRUE)
HawksTaildeviation2<-sd(HawksTail1,na.rm=TRUE)
print("computed by new HawksTail:")
```

```
## [1] "computed by new HawksTail:"
```

```
print(paste("Variance2:",HawksTailvariance2))
```

```
## [1] "Variance2: 5424.1465012701"
```

```
print(paste("Deviation2:",HawksTaildeviation2))
```

```
## [1] "Deviation2: 73.6488051584688"
```

1.4

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

```r
hal<-Hawks$Hallux # Extract the vector of hallux lengths
hal<-hal[!is.na(hal)] # remove ay 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:

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

Compute the mean of the original sample and the corrupted sample:

```r
mean(hal)
```

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

```r
mean(corrupted_hal)
```

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

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.

```r
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))
}
```

1.4 (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.

```r
medians_vect<-c()
for(num_outliers in num_outliers_vect){
  corrupted_hal<-c(hal,rep(outlier_val,times=num_outliers))
  medians_vect<-c(medians_vect,median(corrupted_hal))
}
```

1.4 (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.

```r
t_means_vect<-c()
for(num_outliers in num_outliers_vect){
  corrupted_hal<-c(hal,rep(outlier_val,times=num_outliers))
  t_means_vect<-c(t_means_vect,mean(corrupted_hal,trim=0.1))
}
```

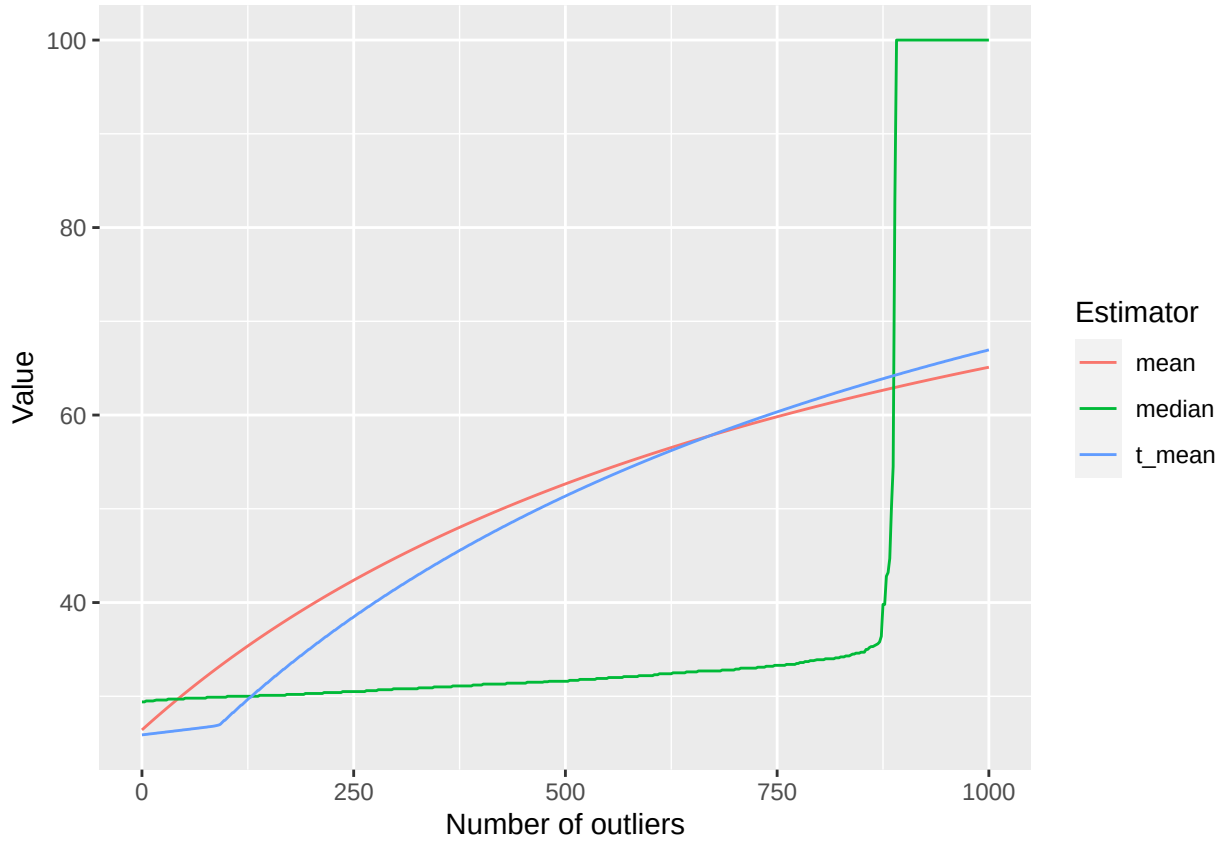
1.4 (Q3) Visualisation

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.

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

Recall that the function pivot_longer() below is used to reshape the data.

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



Question: Which quantity is the most robust when the number of outliers is small?

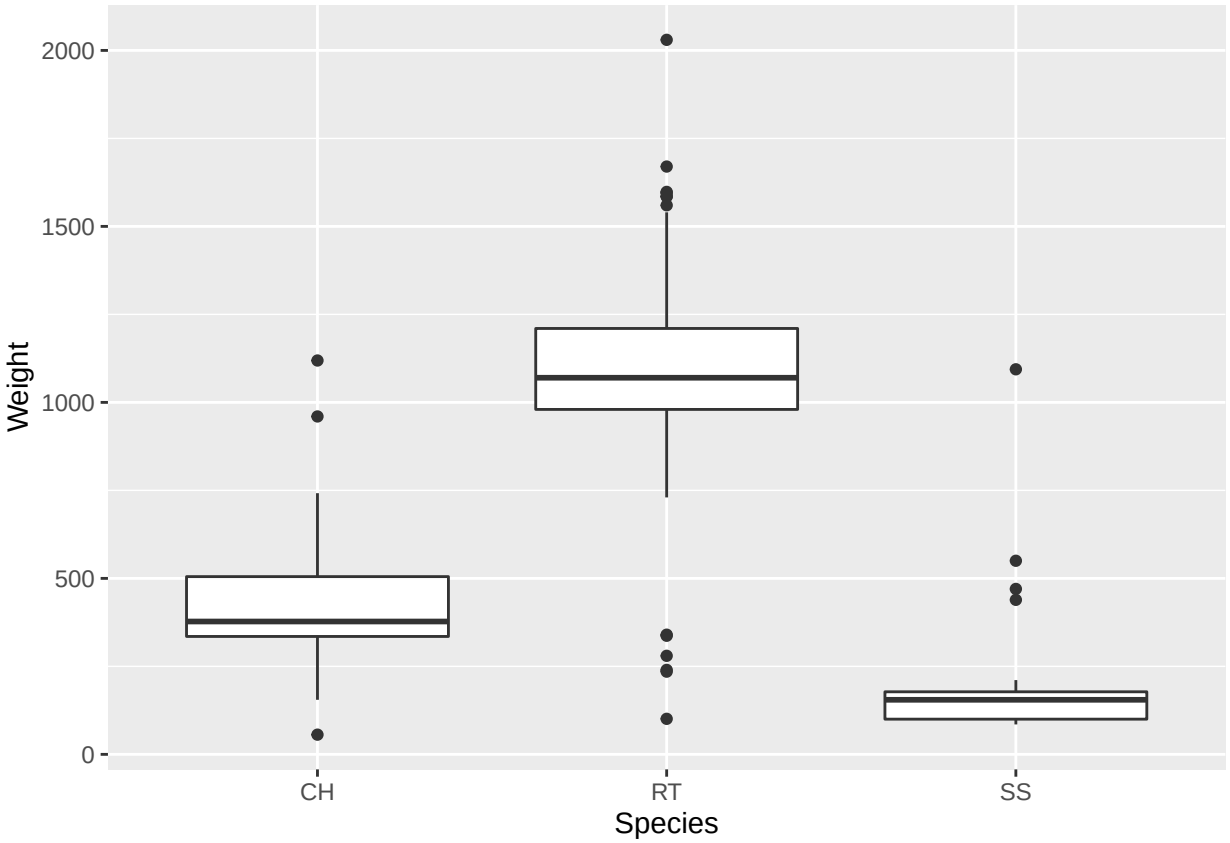
Answer: Median

1.5 (Q1)

Use the functions ggplot() and geom_boxplot() to create a box plot which summarises the distribution of hawk weights broken down by species.

```
ggplot(data=Hawks,aes(x=Species,y=Weight))+
  geom_boxplot()+xlab("Species")+ylab("Weight")
```

```
## Warning: Removed 10 rows containing non-finite values (stat_boxplot).
```



Note the outliers are displayed as individual dots.

1.5 (Q2) Quantile and boxplots

Compute the 0.25-quantile, 0.5-quantile, 0.75-quantile of the Weight grouped by Species.

```
Hawks%>%group_by(Species)%>%
  summarise(quantile025=quantile(Weight,0.25,na.rm=TRUE),
            quantile050=round(quantile(Weight,0.5,na.rm=TRUE),digits=0),
            quantile075=round(quantile(Weight,0.75,na.rm=TRUE),digits=0))
```

```
## # 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
```

Question: Can you explain which parts of the boxplot these numbers correspond to?

Answer: For CH,the top line of the rectangle corresponds to *quantile075*,the line embeded in the rectangle corresponds to *quantile050*, and the bottom line of the rectangle corresponds to *quantile025.RT* and *SS* are same as *CH*

1.5 (Q3) Outliers

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

```
num_outliers<-function(x){
  q25<-quantile(x,0.25,na.rm=TRUE)
  q75<-quantile(x,0.75,na.rm=TRUE)
  IQRrange<-IQR(x,na.rm=TRUE)
  count=0
  # Remove any nans
  outliers<-x[(x>(q75+1.5*IQRrange)|x<(q25-1.5*IQRrange))&(!is.na(x))]
  return(length(outliers))
}
num_outliers(c(0,40,60,185,NA))
```

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

1.5 (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.

```
Hawks%>%group_by(Species)%>%
  summarise(num_outliers_weight=num_outliers(Weight))
```

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

1.6 (Q1)

Compute the covariance and correlation between the Weight and Wing of the Hawks data.

```
cov(Hawks$Weight,Hawks$Wing,use='complete.obs')
```

```
## [1] 41174.39
```

```
cor(Hawks$Weight,Hawks$Wing,use='complete.obs')
```

```
## [1] 0.9348575
```

1.6 (Q2)

Question: Suppose that we have a pair of variables: X with values $X_1, ..., X_n$ and Y with values $Y_1, ..., Y_n$. Suppose that $X_1, ..., X_n$ relation R. Let $a, b \in R$ be real numbers and define a new variable $\tilde{X}$ with $\tilde{X}_1, ..., \tilde{X}_n$ defined by $\tilde{X}_i = a \cdot X_i + b. for\ i = 1, 2, ..., n$. In addition, Let $c, d \in R$ be real numbers and define a new variable $\tilde{Y}$ with $\tilde{Y}_1, ..., \tilde{Y}_n$ defined by $\tilde{Y}_i = c \cdot Y_i + d$. What is the covariance between $\tilde{X}_1, ..., \tilde{X}_n$ and $\tilde{Y}$ with $\tilde{Y}_1, ..., \tilde{Y}_n$ (as a function of S, a, b, c, d)? Assuming that $a \neq 0$ and $c \neq 0$, what is the correlation between $\tilde{X}_1, ..., \tilde{X}_n$ and $\tilde{Y}$ with $\tilde{Y}_1, ..., \tilde{Y}_n$?

Answer:

The covariance between $\tilde{X}$ *and* $\tilde{Y} = a * c * S$

The correlation between $\tilde{X}$ *and* $\tilde{Y} = R$

Verification:

```
a<-2.4
b<-7.1
c<--1
d<-3
X<-Hawks%>%select(Weight)
Y<-Hawks%>%select(Wing)
#computed by my answer
S<-cov(X,Y,use='complete.obs')
R<-cor(X,Y,use='complete.obs')
newS<-a*c*S
newR<-R
print(paste("The covariance is:",newS))
```

```
## [1] "The covariance is: -98818.535939441"
```

```
print(paste("The correlation is:",newR))
```

```
## [1] "The correlation is: 0.934857460302637"
```

```
#computed by new X and new Y
newX<-a*X+b
newY<-c*Y+d
newS_compute<-cov(newX,newY,use='complete.obs')
newR_compute<-cor(newX,newY,use='complete.obs')
print(paste("The verified covariance is:",newS))
```

```
## [1] "The verified covariance is: -98818.535939441"
```

```
print(paste("The verified correlation is:",newR))
```

```
## [1] "The verified correlation is: 0.934857460302637"
```

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

2.1 (Q1)

Questions: Firstly, write down the definition of a random experiment, event and sample space.

Answers:

1. A random experiment is a procedure(real or imagined) which has a well-defined set of possible outcomes and could be repeted arbitrarily many times.
2. A event is a set of possible outcomes of an experiment.
3. A sample space is the set of all possible outcomes of interest for a random experiment.

2.1 (Q2)

Questions: Consider a random experiment of rolling a dice twice. Give an example of what is an event in this random experiment(1). Also, can you write down the sample space as a set(2)? What is the total number of different events in this experiment(3)? Is the empty set considered as an event(4)?

Answers:

(1) The first time is one of {1,2,3,4,5,6} and the second time is also one of {1,2,3,4,5,6}

(2) $\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),$
$(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),$
$(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),$
$(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),$
$(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),$
$(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$

(3) 36

(4) Yes

2.2 (Q1) Set operations

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

Questions:

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 $A \setminus B$ and $A \setminus C$?
4. Are A and B disjoint? Are A and C disjoint?
5. Are B and $A \setminus B$ disjoint?
6. Write down a partition of {1,2,3,4,5,6} consisting of two sets(1). Also, write down another partition of {1,2,3,4,5,6} consisting of three sets(2).

Answers:

1. $A \cup B : \{1,2,3,4,6\}, A \cup C : \{1,2,3,4,5,6\}$
2. $A \cap B : \{2\}$, $A \cap C : \emptyset$
3. $A \setminus B$:{1,3},$A \setminus C$:{1,2,3}
4. A and B are not disjoint,but A and C are disjoint.
5. Yes,B and $A \setminus B$ are disjoint.
6. (1) {1,2,3},{4,5,6},
   (2) {1,2},{3,4},{5,6}

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

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

Questions:

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 $(A \cap B)^c = A^c \cup B^c$.(1) Let's suppose we have a sequence of events $A_1, A_2, ..., A_K \subset \Omega$. Can you write out an expression for $(\cap_{k=1}^K A_k)^c$?(2)
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, $ůůů$, A_K \subset \Omega$. Can you write out an expression for $(\cup_{k=1}^K A_k)^c$?

Answers:

1. A
2. $\emptyset$
3. $A^c := \{w \in \Omega : w \notin A\}, B^c := \{w \in \Omega : w \notin B\}$ $A \subseteq B := \{w \in \Omega : w \in A \implies w \in B\}$ $\implies \{w \in \Omega, w \notin B \implies w \notin A\} \implies B^c \subseteq A^c$
4. (1) Proof:
   Let $M = (A \cap B)^c$ *and* $N = A^c \cup B^c$
   $x \in M \implies x \in (A \cap B)^c$
   $\implies x \notin (A \cap B)$
   $\implies x \notin A$ *or* $x \notin B$
   $\implies x \in A^c$ *or* $x \in B^c$
   $\implies x \in A^c \cup B^c$
   $\implies x \in N$
   Thus, $M \subset N \ldots$ (i)

$y \in N \implies y \in A^c \cup B^c$

$\implies y \in A^c \; or \; y \in B^c$

$\implies y \notin A \; or \; y \notin B$

$\implies y \notin (A \cap B)$

$\implies y \in (A \cap B)^c$

$\implies y \in M$

Thus, $N \subset M \ldots$ (ii)

Combing (i) and (ii) we get:$M = N$ means that $(A \cap B)^c = A^c \cup B^c$

(2) $(\cap_{k=1}^{K} A_k)^c = A_1^c \cup A_2^c \cup A_3^c \cup ... \cup A_K^c$

5. Proof:

   $Let \; P = (A \cup B)^c \; and \; Q = A^c \cap B^c$

   $Let \; x \in P, then \; x \in P \implies x \in (A \cup B)^c$

   $\implies x \notin (A \cup B)$

   $\implies x \notin A \; and \; x \notin B$

   $\implies x \in A^c \; and \; x \in B^c$

   $\implies x \in A^c \cap B^c$

   $\implies x \in Q$

   So $P \subset Q \ldots$ (i)

   $y \in Q \implies y \in A^c \cap B^c$

   $\implies y \in A^c \; and \; y \in B^c$

   $\implies y \notin A \; and \; y \notin B$

   $\implies y \notin (A \cup B)$

   $\implies y \in (A \cup B)^c$

   $\implies y \in P$

   So $Q \subset P \ldots$ (ii)

   Combing (i) and (ii) we get:$P = Q$ means that $(A \cup B)^c = A^c \cap B^c$

6. $(\cup_{k=1}^{K} A_k)^c = A_1^c \cap A_2^c \cap A_3^c \cap ... \cap A_K^c$

2.2 (Q3) Cardinality and the set of all subsets:

Question:

Suppose that $\Omega = \{w_1, w_2, ..., w_K\}$ 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., $E := \{A | A \subset \Omega\}$. Give a formula for the cardinality of E in terms of K.

Answer: *The cardinality of* $E : 2^K - 1$

2.2 (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.

Questions:

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 $S_1 := A_1$, $S_2 = A_2 \setminus A_1$, $S_3 = A_3 \setminus (A_1 \cup A_2)$, $S_4 = A_4 \setminus (A_1 \cup A_2 \cup A_3)$. Show that S_1, S_2, S_3, S_4 form a partition of $A_1 \cup A_2 \cup A_3 \cup A_4$.

Answers:

1. $\emptyset$
2. Prove:
   Let $M = A_1 \cup A_2 \cup A_3 \cup A_4$
   Let $x \in M \implies \{x \in A_1 \; or \; x \in A_2 \; or \; x \in A_3 \; or \; x \in A_4\}$
   $S_2 = A_2 \setminus A_1 \implies \{x \in A_2 | x \notin A_1\}$
   $S_3 = A_3 \setminus (A_1 \cup A_2) \implies \{x \in A_3 | x \notin A_1 \; and \; x \notin A_2\}$
   $S_4 = A_4 \setminus (A_1 \cup A_2 \cup A_3) \implies \{x \in A_4 | x \notin A_1 \; and \; x \notin A_2 \; and \; x \notin A_3\}$
   So, S_1 *and* S_2 *and* S_3 *and* S_4 *are disjoint*
   $S_1 \cup S_2 \cup S_3 \cup S_4 = A_1 \cup A_2 \cup A_3 \cup A_4$
   Thus, S_1, S_2, S_3, S_4 are from a partition of $A_1 \cup A_2 \cup A_3 \cup A_4$

2.2 (Q5) Indicator function.

Questions:

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

1. Write down the indicator function ${1_A}^c$ of A^c (use 1_A in your formula).
2. Can you find a set B whose indicator function is ${1_A}^c + 1_A$?
3. Recall that $1_{A \cap B} = 1_A \cdot 1_B$ and $1_{A \cup B} = \max(1_A, 1_B) = 1_A + 1_B - 1_A \cdot 1_B$ 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).

Answers:

1. ${1_A}^c = 1 - 1_A$
2. Ω
3. Prove:
   ${1_A}^c = 1 - 1_A$
   $1_{A \cup B} = \max(1_A, 1_B) = 1_A + 1_B - 1_A \cdot 1_B$
   $1_{(A \cap B)^c} = 1 - 1_{A \cap B} = 1 - 1_A \cdot 1_B$
   $1_{A^c \cup B^c} = 1_{A^c} + 1_{B^c} - 1_{A^c} \cdot 1_{B^c}$

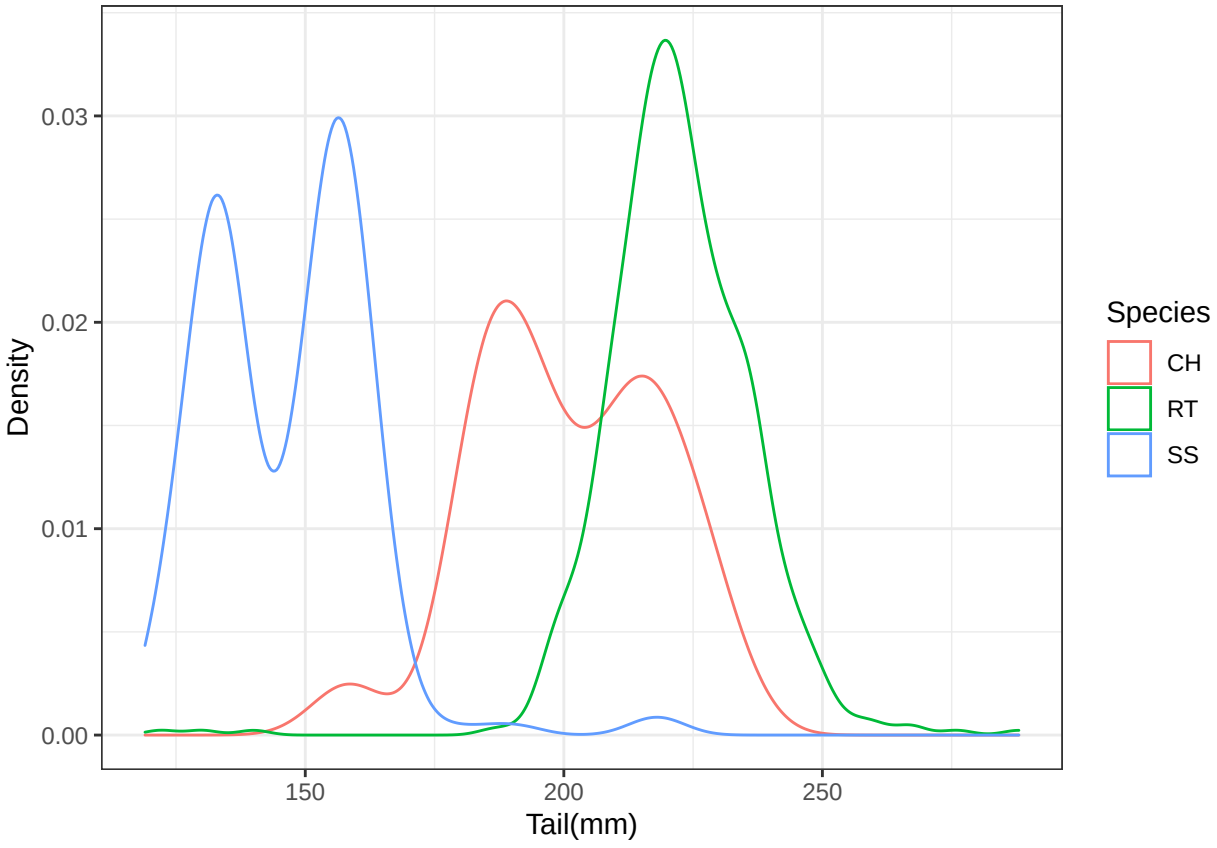
$$= 1 - 1_A + 1 - 1_B - (1 - 1_A) \cdot (1 - 1_B)$$
$$= 2 - 1_A - 1_B - (1 - 1_B - 1_A + 1_A \cdot 1_B)$$
$$= 2 - 1_A - 1_B - 1 + 1_B + 1_A - 1_A \cdot 1_B$$
$$= 1 - 1_A \cdot 1_B$$
$$= 1_{(A \cap B)^c}$$

3. Visualisation

3 (Q1) Density plot:

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

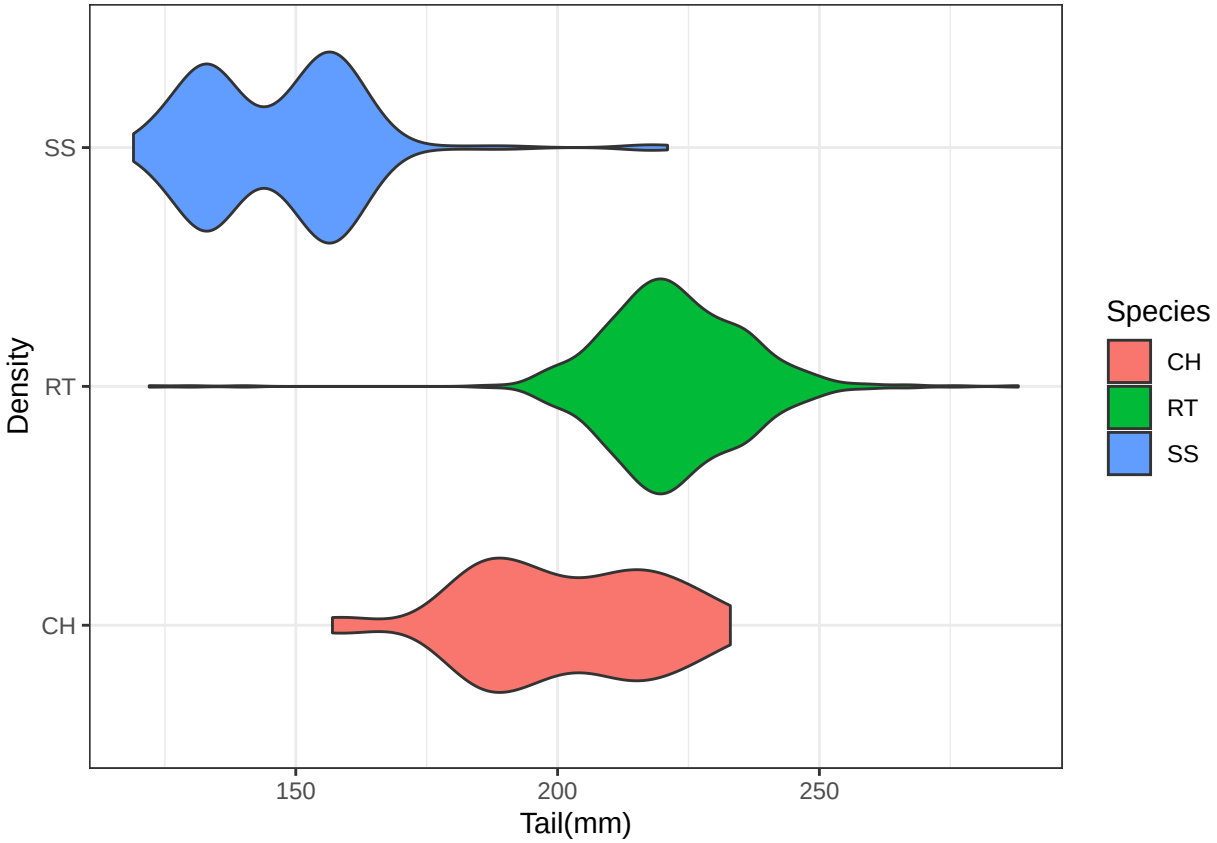
```
ggplot(data=Hawks,aes(x=Tail,color=Species))+
  xlab("Tail(mm)")+ylab("Density")+
  theme_bw()+geom_density()
```



3 (Q2) Violin plot:

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

```
ggplot(data=Hawks,aes(x=Tail,y=Species,fill=Species))+
  xlab("Tail(mm)")+ylab("Density")+theme_bw()+geom_violin()
```



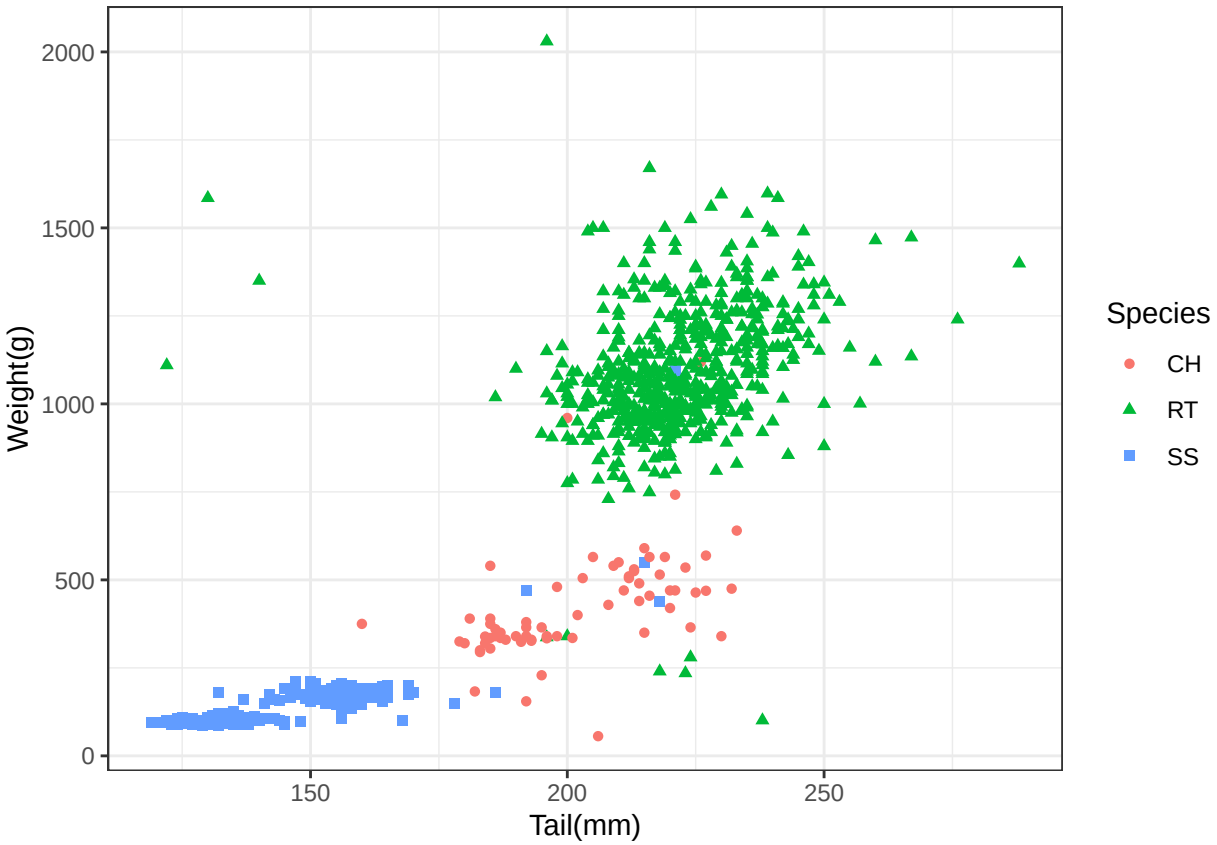
3 (Q3) Scatter plot:

Questions: 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?

```
ggplot(data=Hawks,aes(x=Tail,y=Weight))+
  geom_point(aes(color=Species,shape=Species))+
  xlab("Tail(mm)")+ylab("Weight(g)")+theme_bw()
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```



Answers:

1. Four:(1) Tail-horizontal position (2) Weight-vertival position (3) Species-colour (4) Species-shape
2. Points,Triangle,Square
3. (1) Position (2) Shape (3) Color

3 (Q4) Trend lines and facet wraps:

Questions: 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.

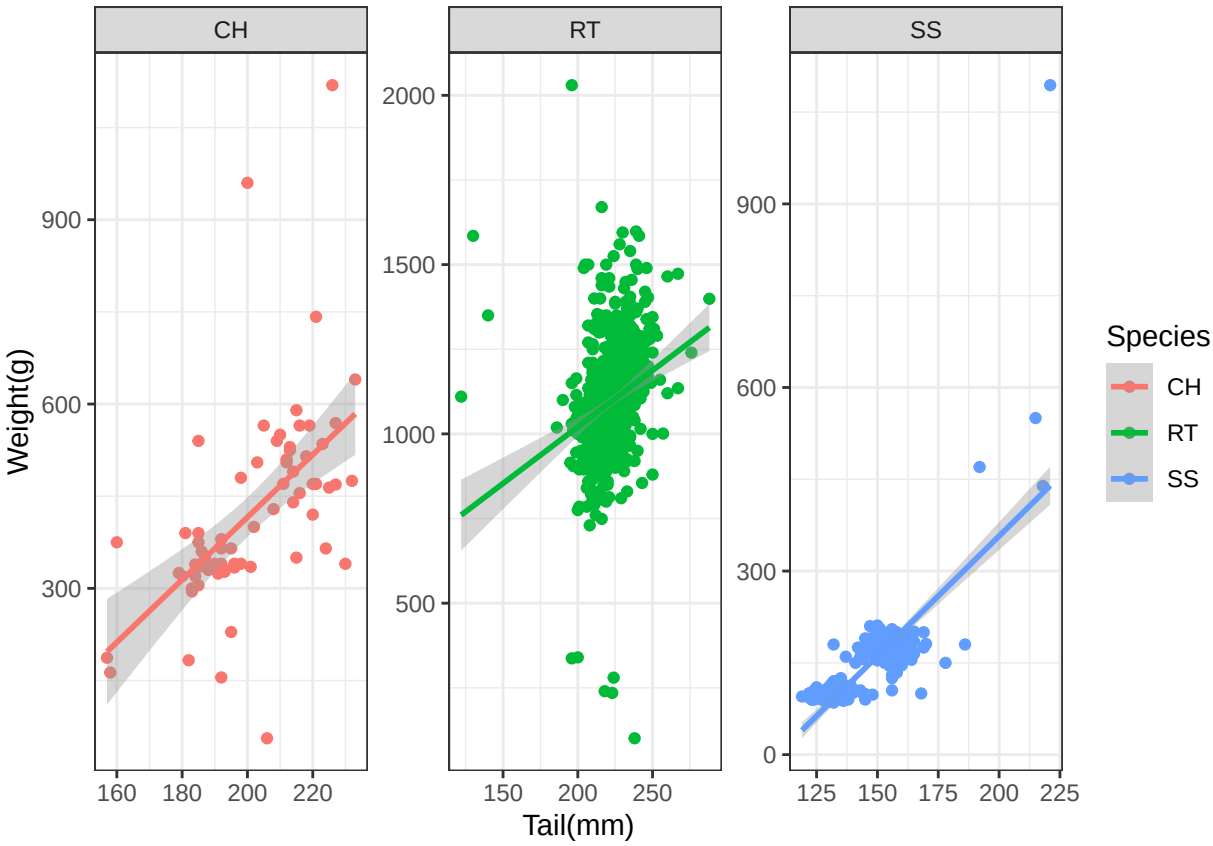
1. What are the visual cues being used within this plot?
2. Based on the plot below, what can we say about the relationship between the weight of the hawks and their tail lengths?

```
ggplot(data=Hawks,aes(x=Tail,y=Weight))+
  geom_point(aes(color=Species))+
  xlab("Tail(mm)")+ylab("Weight(g)")+theme_bw()+
  geom_smooth(aes(color=Species),method="lm")+
  facet_wrap(~Species,scales="free")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 10 rows containing non-finite values (stat_smooth).

## Warning: Removed 10 rows containing missing values (geom_point).
```



Answers:

1. (1) Position (2) Angle (3) Shade (4) Colour (5) Direction

2. There is a positive relationship between Weight and Tail,means that the longer the tail is,the heavier the hawk is.

3 (Q5) 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.

```
heaviest_hawk<-Hawks%>%
  filter(Weight==max(Weight,na.rm=TRUE))%>%
  select(Species,Weight,Tail)
heaviest_hawk
```

```
##   Species Weight Tail
## 1      RT   2030  196
```

```
ggplot(data=Hawks,aes(x=Tail,y=Weight))+
  geom_point(aes(color=Species))+
  xlab("Tail(mm)")+ylab("Weight(g)")+
  theme_bw()+
  geom_curve(x=heaviest_hawk$Tail,xend=heaviest_hawk$Tail,
             y=heaviest_hawk$Weight-300,yend=heaviest_hawk$Weight,
             arrow=arrow(length=unit(0.1,'cm')),curvature=0.5)+
  geom_text(x=heaviest_hawk$Tail,y=heaviest_hawk$Weight-300,
            label="heaviest hawk",size=2)
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```