Week2 Rmarkdown

Tianheng Z

06/10/2021

Include and Install the library

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

```
## Installing package into 'C:/Users/kn21121.UOB/OneDrive - University of Bristol/Documents/R/win-libra
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

```
## Installing package into 'C:/Users/kn21121.UOB/OneDrive - University of Bristol/Documents/R/win-libra
## (as 'lib' is unspecified)
```

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

```
## Installing package into 'C:/Users/kn21121.UOB/OneDrive - University of Bristol/Documents/R/win-libra
## (as 'lib' is unspecified)
```

```
library(ggpubr)
```

Load the data

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

1.Visualization

1.1 Type of variables

Check the dim and head of data

```
dim(hawksSmall)
```

```
## [1] 897   9
```

```
head(hawksSmall,5)
```

```
##   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
```

```
names(hawksSmall)
```

```
## [1] "Age"         "Day"         "Month"       "Year"        "CaptureTime"
## [6] "Species"     "Wing"        "Weight"      "Tail"
```

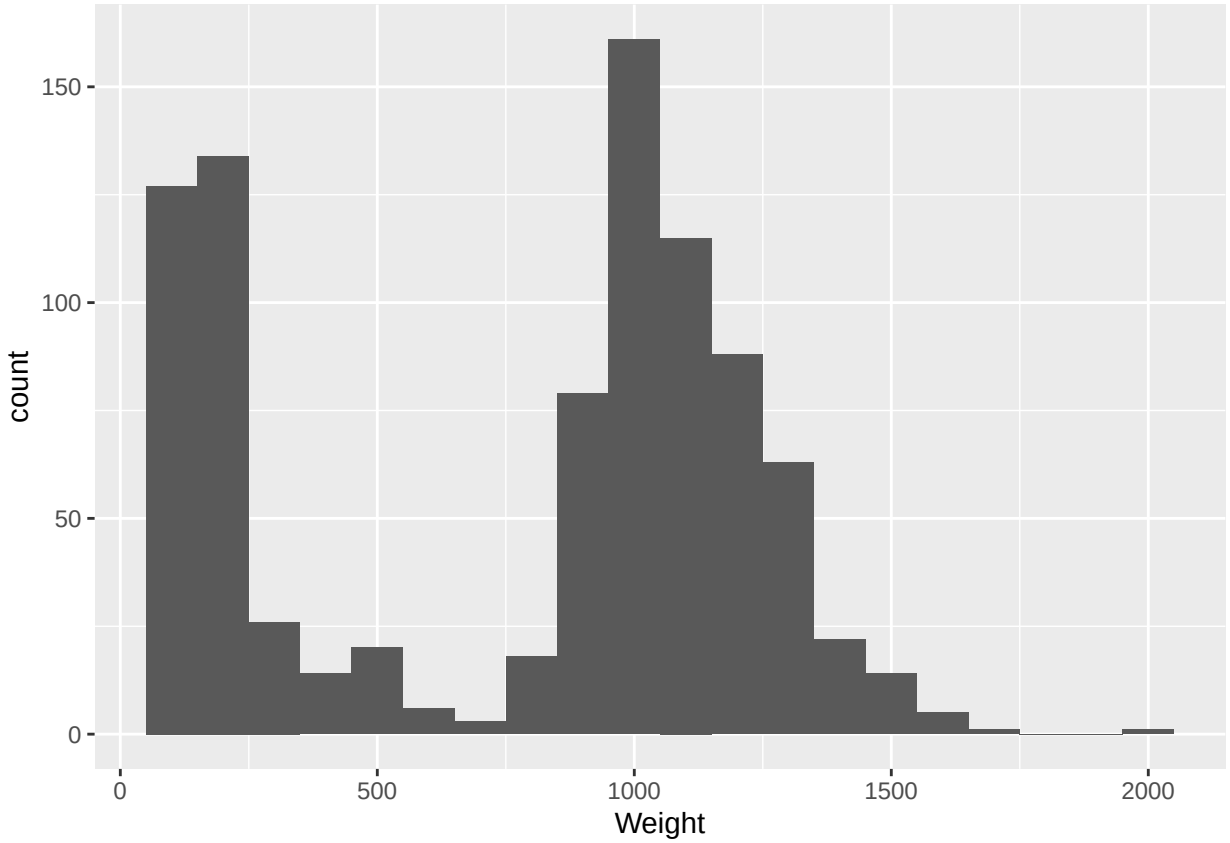
Month-> Categorical, Species->Categorical, Age->discrete, Wing-> Discrete, Weight->Discrete

1.2 What's wrong with the plot

There are too many atttibutes in the plot, making it not easy to understatnd.

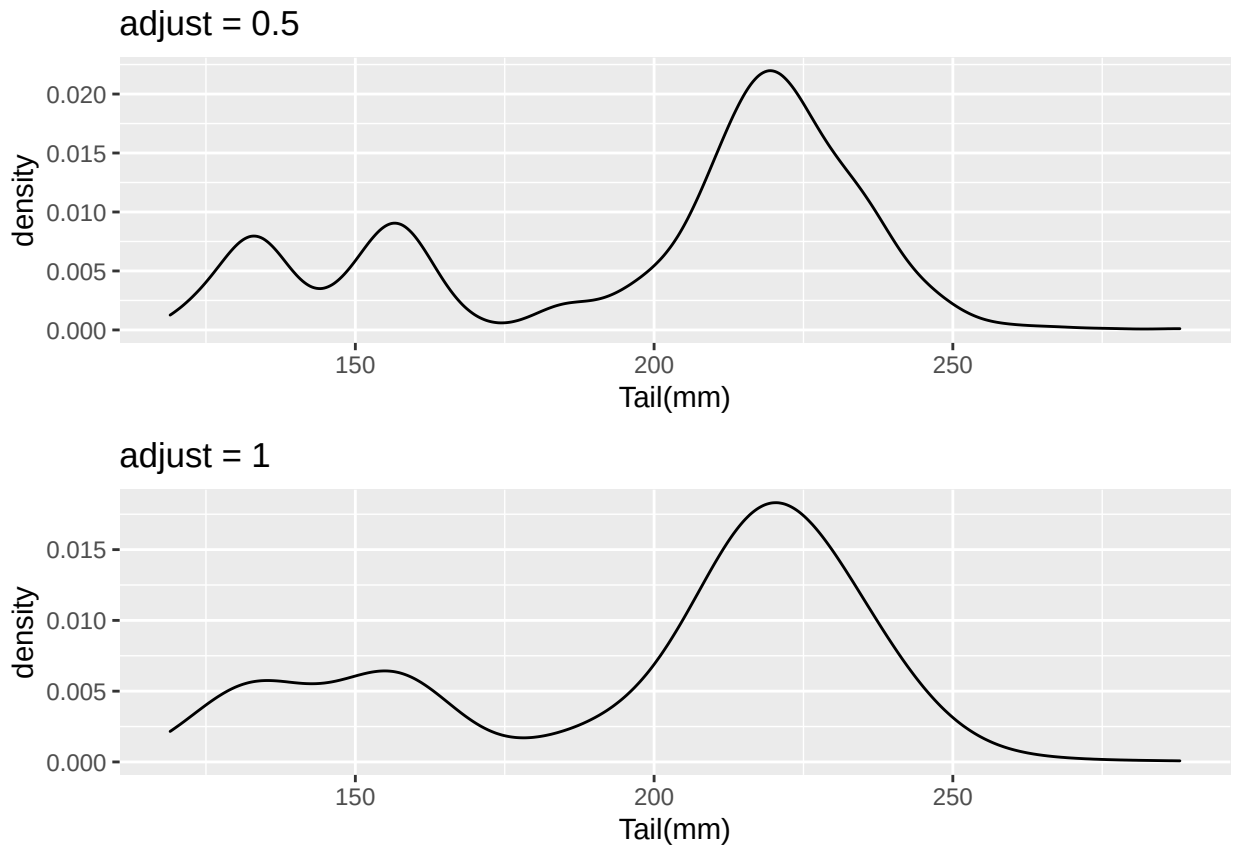
1.3 Generate a histogram

```
#Bimodal data
ggplot(data = hawksSmall,aes(x=Weight)) + geom_histogram(binwidth = 100)
```

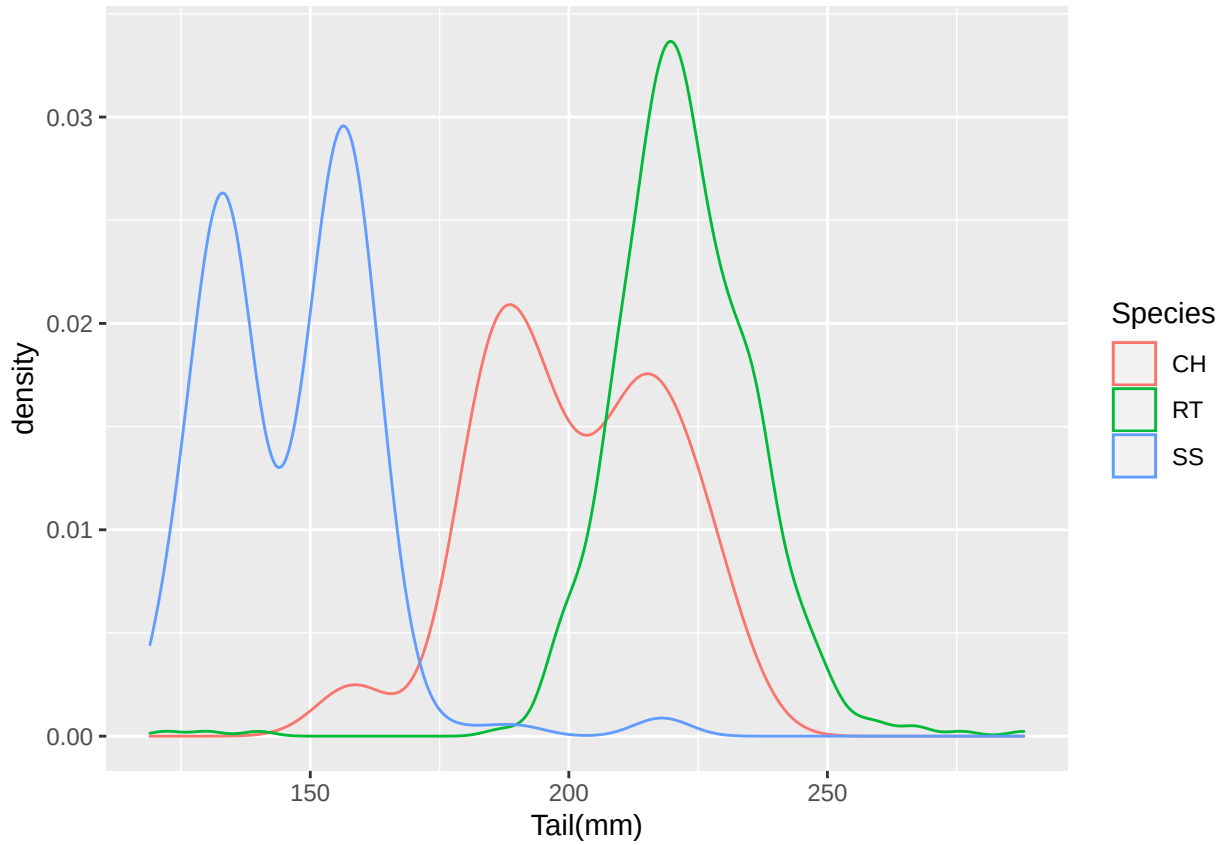


1.4 Generate a density plot

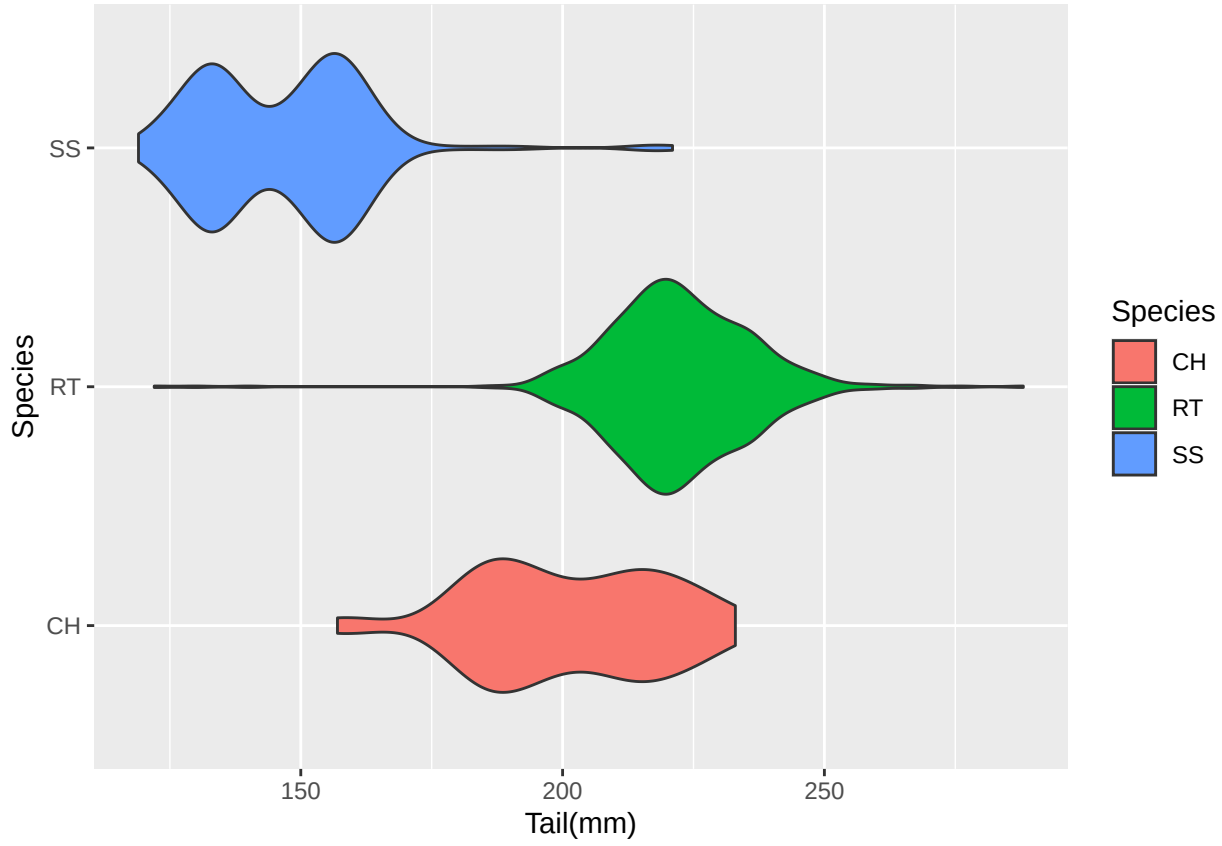
```
density_half <-ggplot(data = hawksSmall,aes(x = Tail)) + geom_density(adjust =
        0.5) + xlab("Tail(mm)") + ggtitle("adjust = 0.5")

density_one <- ggplot(data = hawksSmall,aes(x = Tail)) + geom_density(adjust =
          1) + xlab("Tail(mm)") +ggtitle("adjust = 1")

ggarrange(density_half,density_one,nrow = 2,ncol = 1)
```




```
ggplot(data = hawksSmall,aes(x = Tail,color = Species)) + geom_density() + xlab("Tail(mm)")
```

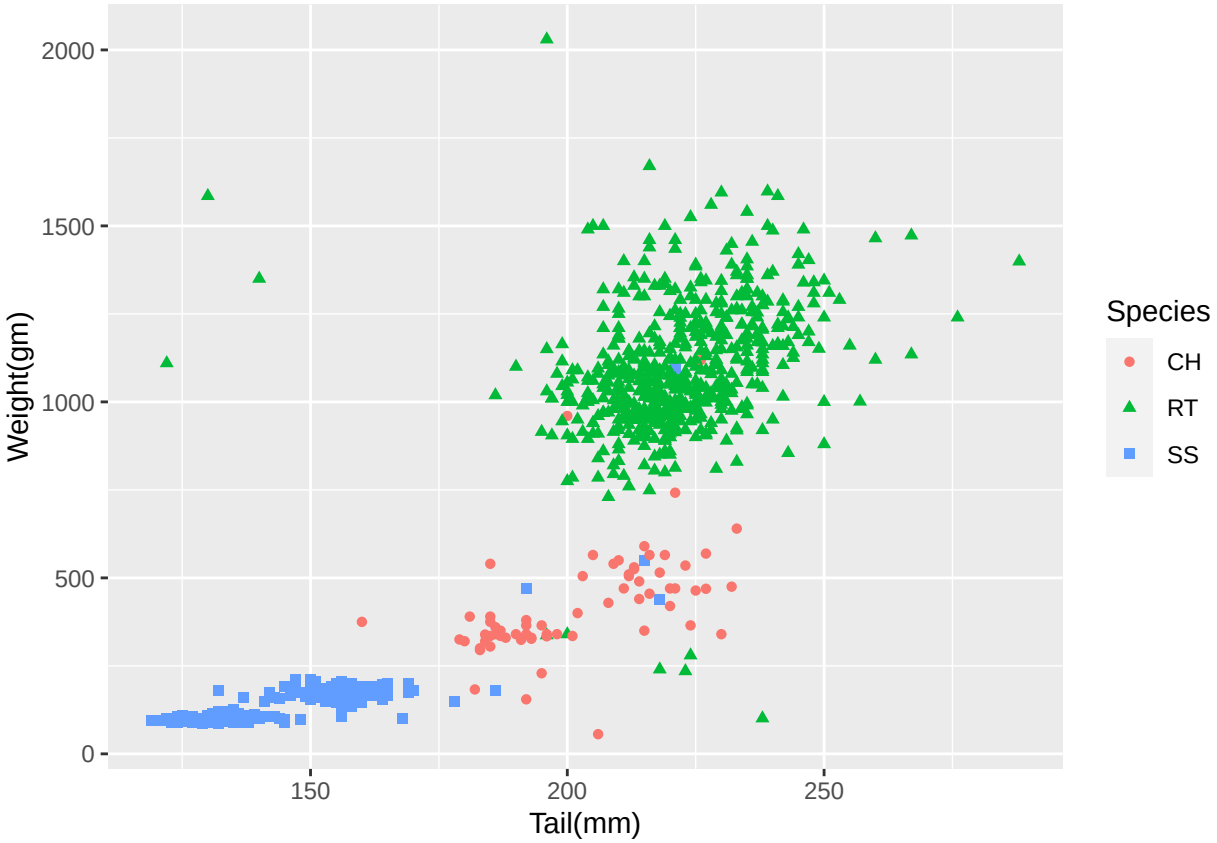



```
ggplot(data = hawksSmall,aes(x = Tail,y = Species,fill = Species)) + geom_violin() + xlab("Tail(mm)")
```



1.5 Scatter plots

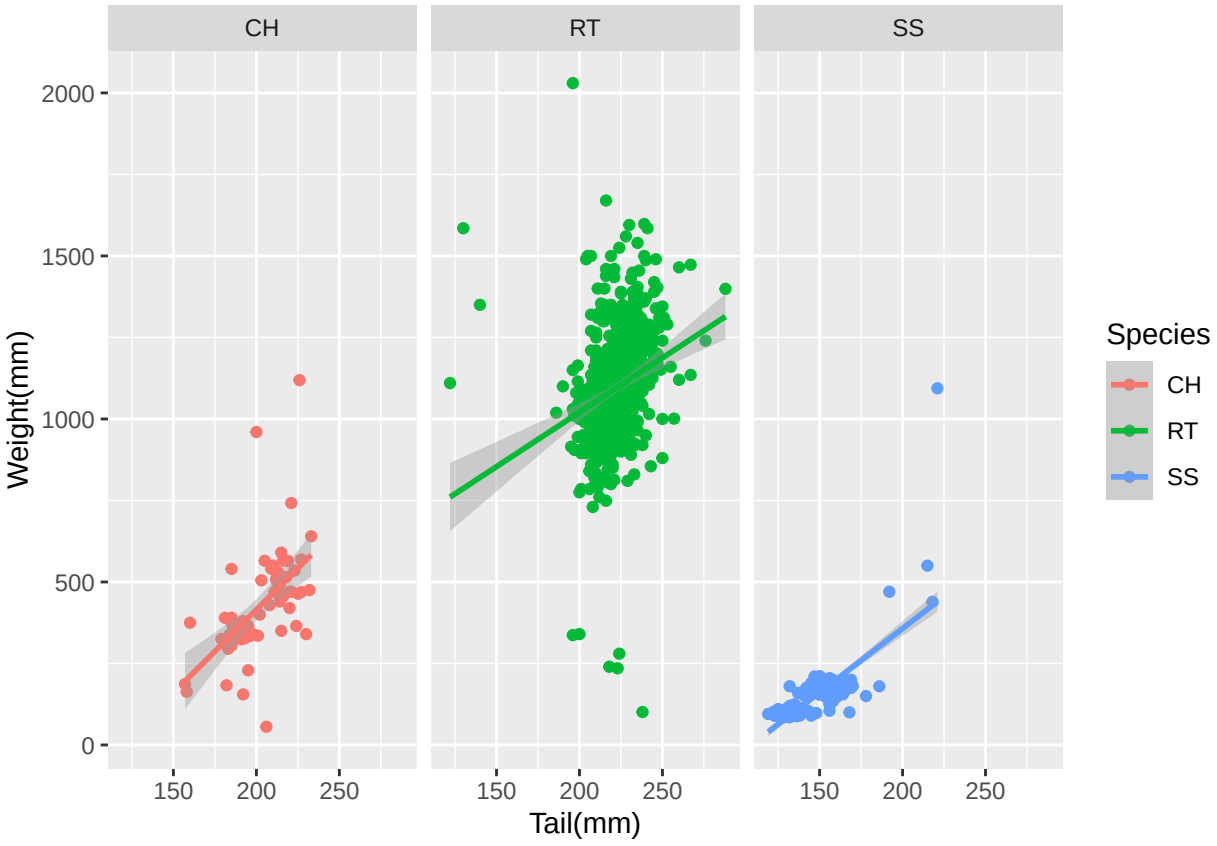
```
ggplot(data = hawksSmall,aes(x = Tail, y = Weight,color = Species,shape = Species))+
  geom_point() + xlab("Tail(mm)") + ylab("Weight(gm)")
```



1.6 Trend Lines and facet wraps

```
ggplot(data = hawksSmall,aes(x = Tail, y =Weight,color = Species)) + geom_point()+ geom_smooth(method =
```

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



2. Data Wrangling

2.1 Select and filter functions

```
hSF <- Hawks %>% filter(Species == "RT", Weight >=1000) %>% select(Wing,Weight,Tail)
dim(hSF)
```

```
## [1] 398   3
```

2.2 The arrange function

```
head(hSF %>% arrange(Wing),5)
```

```
##    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 funtions

```
inter <- data.frame(species_code = c("CH","RT","SS"),
                    Species_name_full = c("Cooper's","Red-tailed","Sharp-shinned"))
```

```
hawksFullname <-left_join(Hawks,rename(inter,Species = species_code),by ="Species" )
hawksFullname %>% select(Species_name_full,Wing,Weight) %>% rename(Species = Species_name_full) %>% hea
```

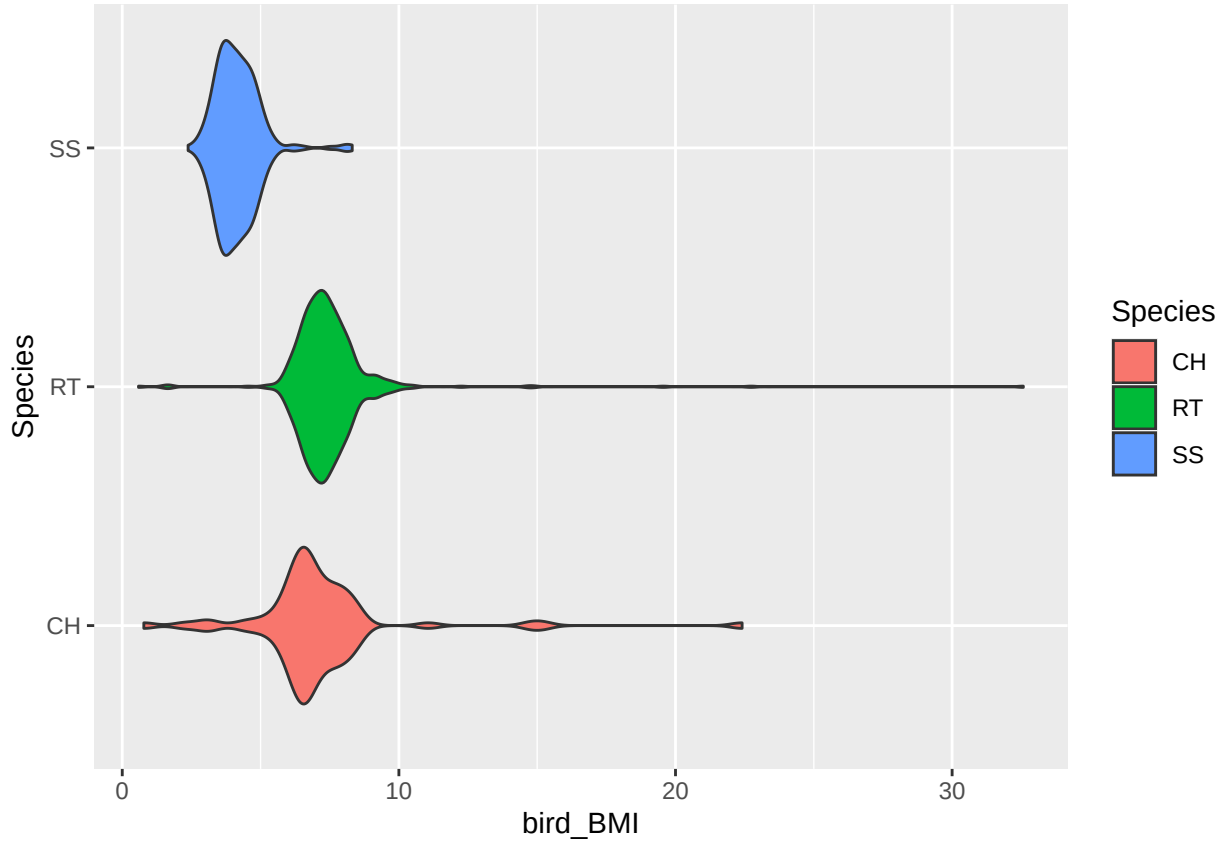
```
##         Species Wing Weight
## 1    Red-tailed  385    920
## 2    Red-tailed  376    930
## 3    Red-tailed  381    990
## 4      Cooper's  265    470
## 5 Sharp-shinned  205    170
## 6    Red-tailed  412   1090
## 7    Red-tailed  370    960
```

2.4 The mutate function

```
hawksWithBMI <- Hawks %>% mutate(bird_BMI = 1000 * Weight/(Wing^2) ) %>%
  select(Species,bird_BMI) %>% arrange(desc(bird_BMI))
head(hawksWithBMI,8)
```

```
##   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
```

```
hawksWithBMI %>% filter(bird_BMI < 100) %>%
  ggplot(aes(y=Species,x = bird_BMI,fill = Species))+
  geom_violin()
```



2.5 Summarize and group-by functions

```
hawksFullname %>%
  group_by(Species) %>%
  summarise(num_rows = n(),mn_wing = mean(Wing,na.rm = 1),md_wing = median(Wing,na.rm = 1),
            t_mn_wing = mean(Wing,trim=0.1,na.rm = 1),
            tail_wing_ratio = mean(Wing/Tail,na.rm = 1))
```

```
## # 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 CH            70    244.     240      243.            1.22
## 2 RT           577    383.     384      385.            1.73
## 3 SS           261    185.     191      184.            1.26
```

```
hawksFullname %>% select(Wing,Weight,Culmen,Hallux,Tail,StandardTail,Tarsus,Crop,Species_name_full) %>%
```

```
## # 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
```

3. Exploratory data analysis

3.1 Combining location estimators with the summarise function

```
Hawks %>% summarise(Wing_mean = mean(Wing,na.rm = 1),
                    Wing_t_mean = mean(Wing,trim = 0.1,na.rm = 1),
                    Wing_med = median(Wing,na.rm = 1),
                    Weight_mean = mean(Weight,na.rm = 1),
                    Weight_t_mean = mean(Weight,na.rm = 1,trim = 0.1),
                    Weight_med = median(Weight,na.rm = 1)
                    )
```

```
##   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
```

```
Hawks %>% group_by(Species) %>% summarise(Wing_mean = mean(Wing,na.rm = 1),
                    Wing_t_mean = mean(Wing,trim = 0.1,na.rm = 1),
                    Wing_med = median(Wing,na.rm = 1),
                    Weight_mean = mean(Weight,na.rm = 1),
                    Weight_t_mean = mean(Weight,na.rm = 1,trim = 0.1),
                    Weight_med = median(Weight,na.rm = 1)
                    )
```

```
## # 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 transformation

```
a = c(1,2,3,4,5,6,7,8,9,10)
mean(a)
```

```
## [1] 5.5
```

```
var(a)
```

```
## [1] 9.166667
```

```
mean(5 *a + 10)
```

```
## [1] 37.5
```

```
var(5 *a +10)
```

```
## [1] 229.1667
```

3.3 Robustness of location estimators

```
hal<-Hawks$Hallux     # Extract the vector of hallux lengths
hal<-hal[!is.na(hal)] # Remove any nans
outlier_val<-100
num_outliers<-10
corrupted_hal<-c(hal,rep(outlier_val,times=num_outliers))
mean(hal)
```

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

```
mean(corrupted_hal)
```

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

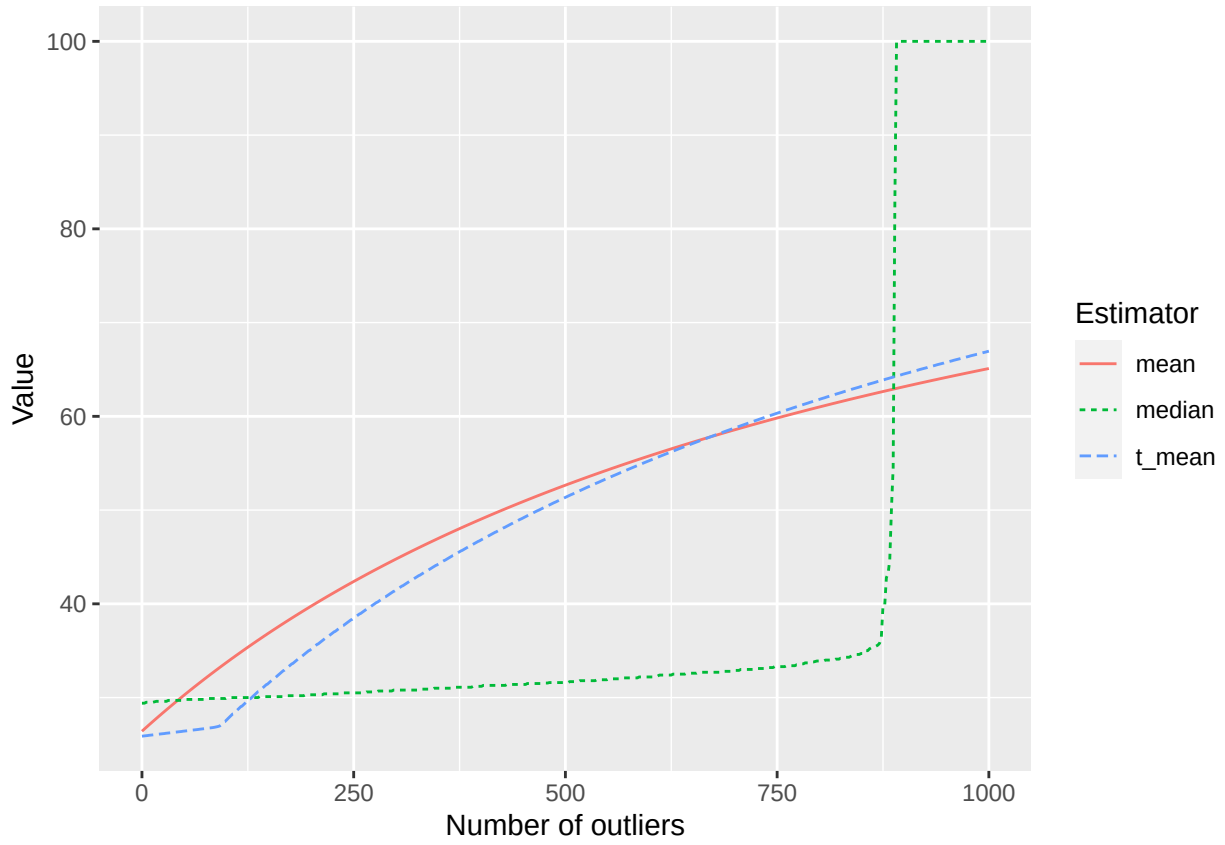
```
num_outliers_vect<-seq(0,1000)
means_vect<-c()
for(num_outliers in num_outliers_vect){
corrupted_hal<-c(hal,rep(outlier_val,times=num_outliers))
means_vect<-c(means_vect,mean(corrupted_hal))
}
```

```
num_outliers_vect<-seq(0,1000)
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))
}
```

```
num_outliers_vect<-seq(0,1000)
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))
}
```

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

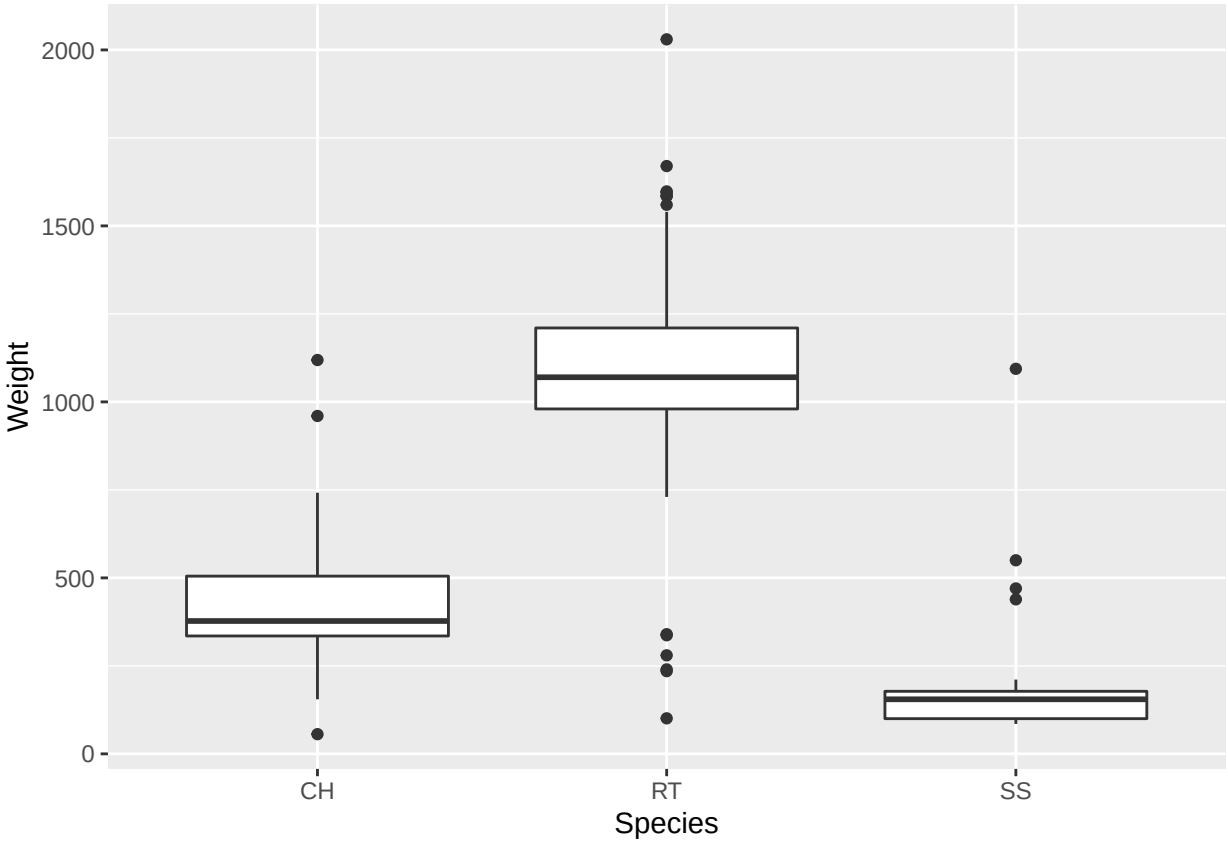
```
df_means_medians%>%
pivot_longer(!num_outliers, names_to = "Estimator", values_to = "Value")%>%
ggplot(aes(x=num_outliers,color=Estimator,
linetype=Estimator,y=Value))+
geom_line()+xlab("Number of outliers")
```



3.4 Box plots and outliers

```
Hawks %>% group_by(Species) %>% ggplot(aes(x = Species,y = Weight)) + geom_boxplot()
```

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




```
num_outliers <-function(sample)
{

  out_num <- 0
  q25 <- quantile(sample,prob = .25)
  q75 <- quantile(sample,probs = .75)
  sec <- q75-q25
  for(num in sample)
  {
    if(num <q25-1.5*sec | num > q75 + 1.5*sec)
    {
      out_num = out_num +1
    }
  }
  out_num
}

Hawks %>%filter(!is.na(Weight)) %>% group_by(Species) %>% summarise(outlier_weights = num_outliers(Weigh
```

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

3.5 Covariance and Correlation under linear transformation

```
set.seed(10000)
X = rnorm(100,5,5)
Y = rnorm(100,10,10)
cov(X,Y)
```

```
## [1] 7.871955
```

```
cov(X,Y)/sd(X)/sd(Y)
```

```
## [1] 0.1545011
```

```
new_X = 5*X + 10
new_Y = 2*Y + 4
cov(new_X,new_Y)
```

```
## [1] 78.71955
```

```
cov(new_X,new_Y)/sd(new_X)/sd(new_Y)
```

```
## [1] 0.1545011
```