Capstone Project

Agatha Ojonuga Egwemi

Table of Contents

Cuckoo Eggs Dimension

Introduction

This project aims to check for differences between the egg dimensons among six different Cuckoo species namely: Meadow pipit, Hedge sparrow, Tree pipit, Pied wagtail, Robin and Wren. This data is from the DAAG package by Maindonald, Braun, and Braun (2015). It presents measurement on 120 eggs laid in the nest of 6 different species of cuckoos. This data contains the following columns:

length: The length of eggs in millimeters

breadth: The breadth of eggs in millimeters

species: Six species of birth species

Figure 1: Cuckoo Eggs

The following will be performed on the data:

- Data importation
- Exploratory data analysis (EDA).
- Visualization of data using appropriate plots.
- Inferential statistical analysis to see if there's a difference in breadth of eggs laid by the bird species.
- Inferential statistical analysis to see if there's a difference in length of eggs laid by the bird species.
- Results and Discussion

Loading and naming the data

```
library(tidyverse)
library(ggplot2)

Cuckoo_Eggs <- read.csv("https://raw.githubusercontent.com/xrander/bootcamp-test/master/data/
```

I will assign the data imported to a variable name, "**Cuckoo_Eggs**" so it is easier to call out the data. From this point, the imported data will be referred to as "**Cuckoo_Eggs**".

Exploratory Data Analysis

Previewing the data

To get an overview of the data, I will present the first six observations ,the last six observations and ten random observations in a table

For the first six observations:

```
Cuckoo_Eggs |>
  head() |>
  knitr::kable()
```

Table 1: First six observations

length	breadth	species
21.7	16.1	Meadow Pipit
22.6	17.0	Meadow Pipit
20.9	16.2	Meadow Pipit
21.6	16.2	Meadow Pipit
22.2	16.9	Meadow Pipit
22.5	16.9	Meadow Pipit

For the last six observations:

```
Cuckoo_Eggs |>
  tail() |>
  knitr::kable()
```

Table 2: Last six observations

	length	breadth	species
115	20.9	15.9	Wren
116	22.0	16.0	Wren
117	20.0	15.7	Wren
118	20.8	15.9	Wren
119	21.2	16.0	Wren
120	21.0	16.0	Wren

For ten (10) random observations:

```
Cuckoo_Eggs |>
  car::some() |>
  knitr::kable()
```

Table 3: Ten random observations

	length	breadth	species
2	22.6	17.0	Meadow Pipit
10	22.6	17.0	Meadow Pipit
35	21.7	16.2	Meadow Pipit
69	22.8	16.2	Hedge Sparrow
73	23.0	16.7	Hedge Sparrow
81	23.0	17.2	Robin
83	23.9	16.9	Robin
99	24.9	16.8	Pied Wagtail
101	22.1	16.2	Pied Wagtail
107	22.1	16.0	Wren

Data Summary

Getting a simple summary of the data:

i. Number of rows

```
Cuckoo_Eggs |>
  nrow()
```

```
[1] 120
```

ii. Number of columns

```
Cuckoo_Eggs |>
  ncol()
```

```
[1] 3
```

iii. Name of data variables

```
Cuckoo_Eggs |>
  names()
```

```
[1] "length"  "breadth" "species"
```

iv. Structure of data

```
Cuckoo_Eggs |>
  str()
```

```
'data.frame':   120 obs. of  3 variables:
 $ length : num  21.7 22.6 20.9 21.6 22.2 22.5 22.2 24.3 22.3 22.6 ...
 $ breadth: num  16.1 17 16.2 16.2 16.9 16.9 17.3 16.8 16.8 17 ...
 $ species: chr  "Meadow Pipit" "Meadow Pipit" "Meadow Pipit" "Meadow Pipit" ...
```

This gives a summary of the data structure. The Cuckoo egg dimension data has 120 observations with 3 variables namely; length, breadth and species. Length and breadth are numerical variables while Species is a factor with 6 levels.

v. Summary of data

```
Cuckoo_Eggs |>
  summary()
```

```
     length         breadth        species
 Min.   :19.60   Min.   :15.00   Length:120
 1st Qu.:21.90   1st Qu.:16.20   Class :character
 Median :22.35   Median :16.60   Mode  :character
 Mean   :22.45   Mean   :16.55
 3rd Qu.:23.23   3rd Qu.:17.00
 Max.   :25.00   Max.   :17.50
```

All the above rsult can also be gotten fromjust skimming the data:

```
skimr::skim_without_charts(Cuckoo_Eggs)
```

Table 4: Data summary

Name	Cuckoo_Eggs
Number of rows	120
Number of columns	3
_____	
Column type frequency:	
character	1
numeric	2
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
species	0	1	4	13	0	6	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
length	0	1	22.45	1.07	19.6	21.9	22.35	23.22	25.0
breadth	0	1	16.55	0.52	15.0	16.2	16.60	17.00	17.5

This gives a summary of the data, showinh the mean, median, minimum and muximum values of the numeriacal varables. It also shows the amount of obeservations for easch species, showig that the Meadow Pipit has the highest number of observations.

Missing Data

To check if there is any missing data:

```
is.na(Cuckoo_Eggs) |>
  colSums() |>
  knitr::kable()
```

	x
length	0
breadth	0
species	0

There is(are) no missing value(s) in the imported data.

Single Variable Analysis and Visualization

I will explore and analyse the various variables that are contained in the data set: length, breadth and species.

Exploring Length

```
mean(Cuckoo_Eggs$length)
```

```
[1] 22.45
```

```
median(Cuckoo_Eggs$length)
```

```
[1] 22.35
```

same can also be done with the other numerical variable; breadth. although this is alraedfy contained above from the data summary.

Different lengths and their frequency, arranged in descending order of length

Printing only 20 observations:

```
Cuckoo_Eggs |>
  count(length) |>
  arrange(desc(length)) |>
  knitr::kable() |>
  print(20)
```

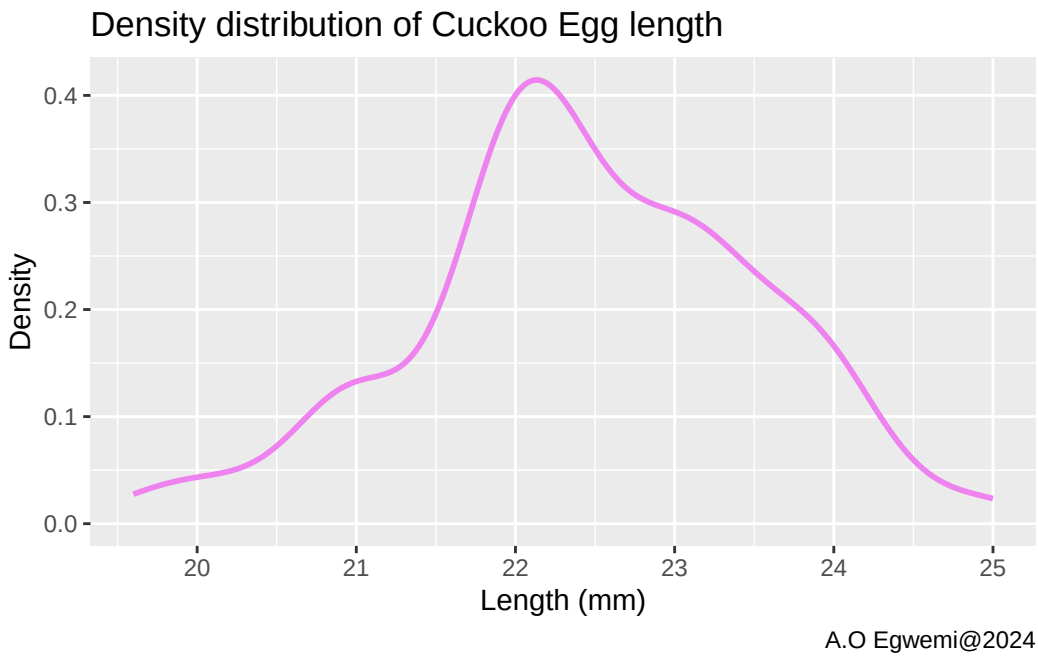
length	n
25.0	1
24.9	1
24.4	1
24.3	1
24.0	7
23.9	2
23.8	4
23.7	1
23.6	1
23.5	2
23.4	3
23.3	6
23.2	1
23.1	3
23.0	9
22.9	1
22.8	4
22.7	1
22.6	4
22.5	2
22.4	5
22.3	6
22.2	4
22.1	6
22.0	13
21.9	4
21.8	4
21.7	3
21.6	1
21.5	1
21.2	1
21.1	2
21.0	4
20.9	4
20.8	1
20.6	1
20.3	1
20.1	1
20.0	1

```
|   19.8|  1|
|   19.6|  1|
```

This shows twenty (20) different lengths and their frequency,. With this, we see that **22.0** has the highest frequency. The lengths are then arranged in a descending order.

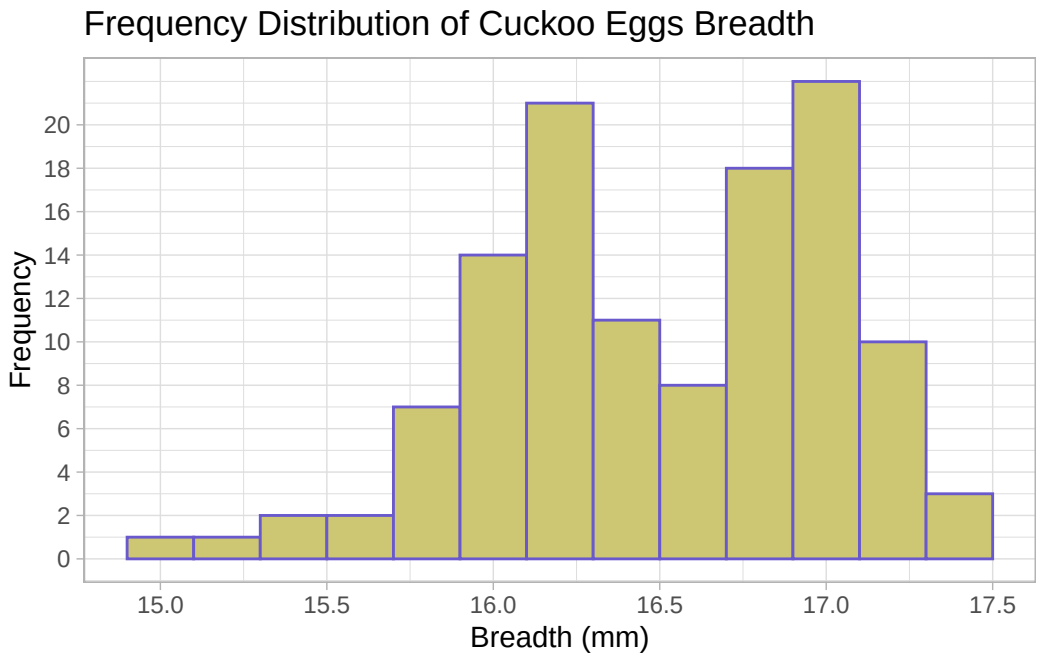
Visualizing Length

```
ggplot(Cuckoo_Eggs, aes(length)) +
  geom_density(
    col = "violet",
    linewidth = 1   )+
  labs(
    x = "Length (mm)",
    y = "Density",
    title = "Density distribution of Cuckoo Egg length",
    caption = "A.O Egwemi@2024"   )
```



Visualizing Breadth with a histogram, showing how many breadth measrements fall within a particular range

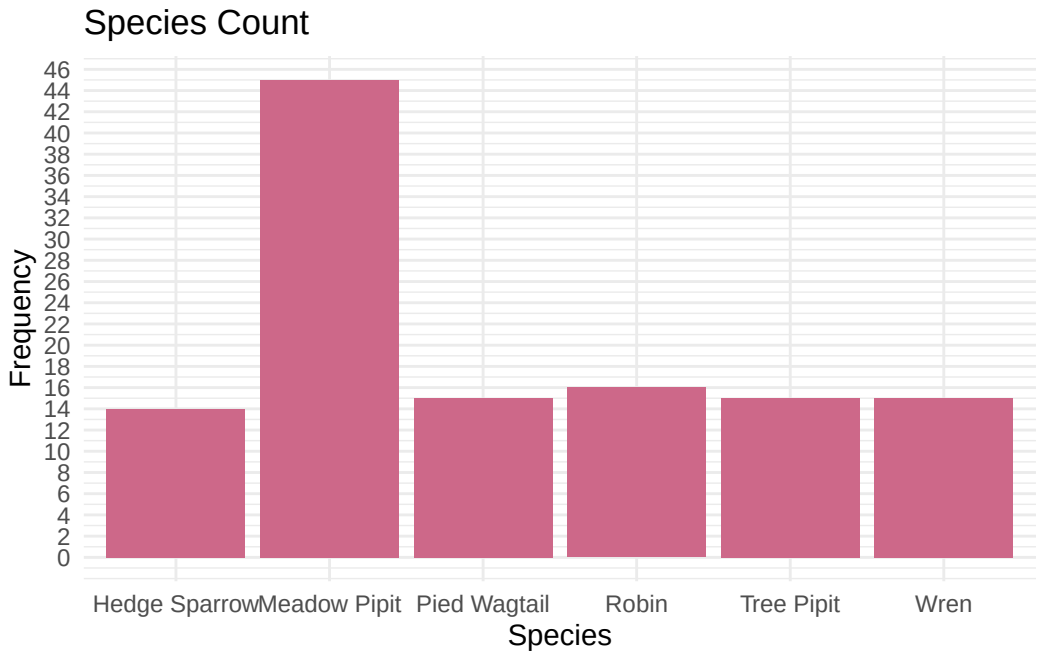
```r
Cuckoo_Eggs |>
  ggplot(aes(breadth))+
  geom_histogram(
    col = "slateblue3",
    fill = "khaki3",
    binwidth = .20   )+
  labs(
    x = "Breadth (mm)",
    y= "Frequency",
    title = "Frequency Distribution of Cuckoo Eggs Breadth"   )+
  theme_light()+
  scale_x_continuous(
    breaks = seq(0,20,.5)   )+
  scale_y_continuous(
    breaks = seq(0,20,2)   )
```



Visualizing count of species with a bar chart

```r
ggplot(Cuckoo_Eggs, aes(species)) +
  geom_bar(
    fill= "palevioletred3"   )+
  labs(
    x= "Species",
```

```
    y= "Frequency",
    title = "Species Count"      )+
  scale_y_continuous(
    breaks = seq(0,50,2)   )+
  theme_minimal()
```



This chart shows that the meadow pipit is the species with the highest count/frequency as initially seen in the data summary.

Two Variables Analysis and Visualization

The total and average length/breadth of each species can also be presented in a table and visualized

```
Cuckoo_Eggs |>
  summarise(
    .by = species,
    Total_length = sum(length),
    Average_length = mean(length),
    Total_breadth = sum(breadth),
    Average_breadth = mean(breadth)
```

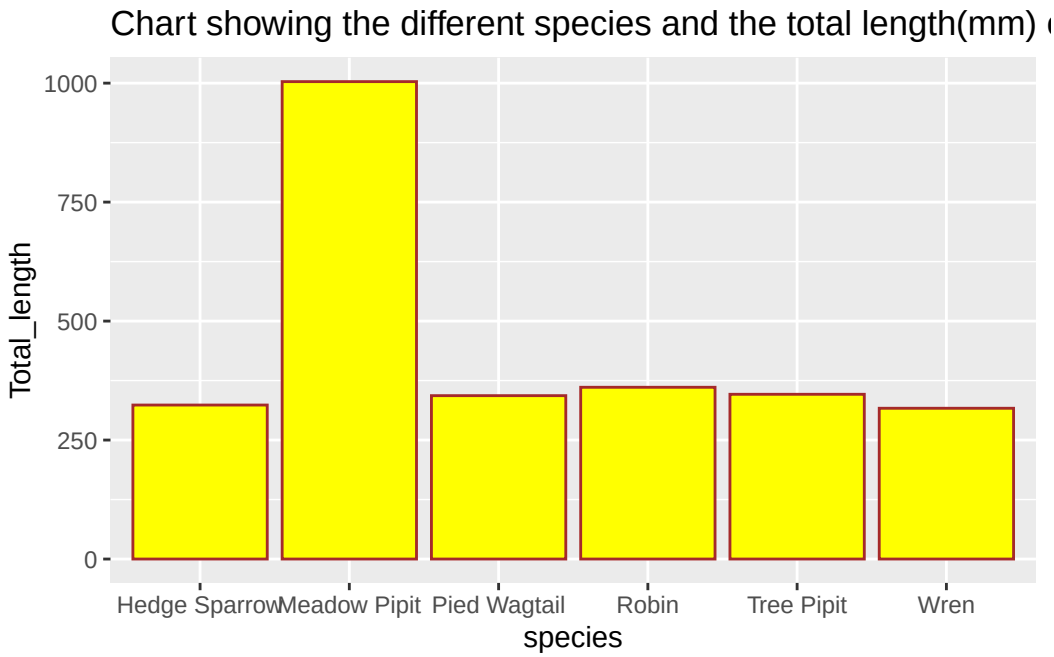
```
    ) |>
  knitr::kable()
```

Table 8: Summary of species Average and total measurements of eggs

species	Total_length	Average_length	Total_breadth	Average_breadth
Meadow Pipit	1003.2	22.29333	753.3	16.74000
Tree Pipit	346.2	23.08000	250.0	16.66667
Hedge Sparrow	323.6	23.11429	234.7	16.76429
Robin	360.9	22.55625	263.2	16.45000
Pied Wagtail	343.3	22.88667	247.5	16.50000
Wren	316.8	21.12000	237.5	15.83333

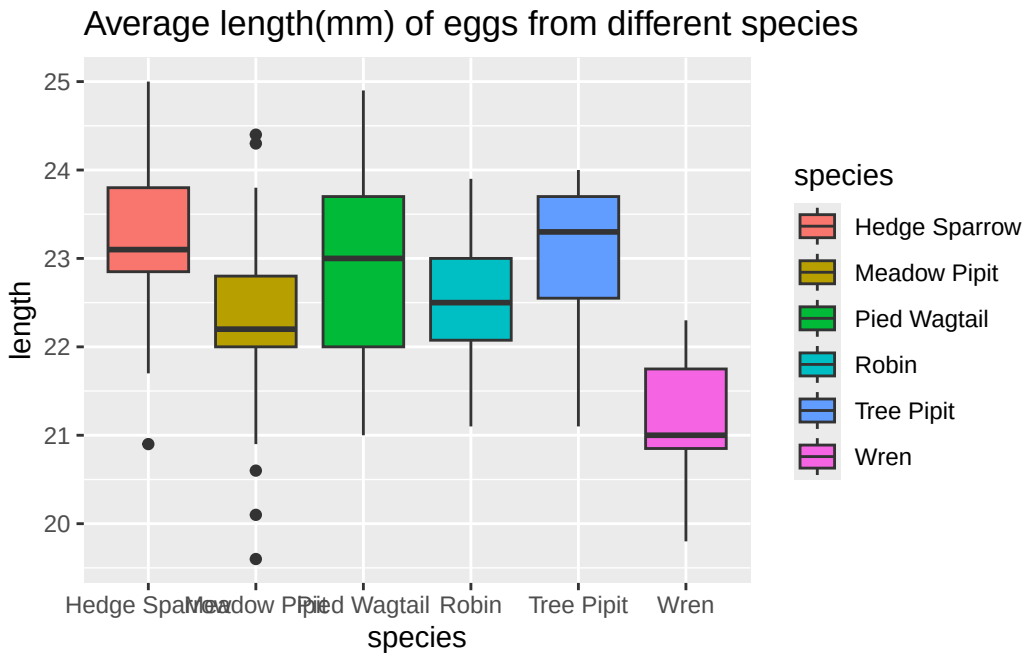
Visualizing the total length and species:

```
Cuckoo_Eggs |>
  summarise(
    .by = species,
    Total_length = sum(length),
    Average_length = mean(length),
    Total_breadth = sum(breadth),
    Average_breadth = mean(breadth)     ) |>
  ggplot(aes(species,Total_length))+
  geom_bar(stat = "identity", col="brown", fill = "yellow")+
  labs(
    title = "Chart showing the different species and the total length(mm) of their eggs"   )
```



This shows that the Meadow pipit has the longest length of eggs. However to know which species egg is best in terms of length, we use the average length of egg. This can be visualized using a boxplot. This shows that the Tree Pipit has the highest median length of eggs. This suggests that on average, it has the longest eggs This can also be visualized.

```
Cuckoo_Eggs |>
  ggplot(aes(species, length, fill = species))+
  geom_boxplot()+
  labs(
    title = "Average length(mm) of eggs from different species"
    )
```



Correlation

Checking to see if the length and breadth of each species is correlated:

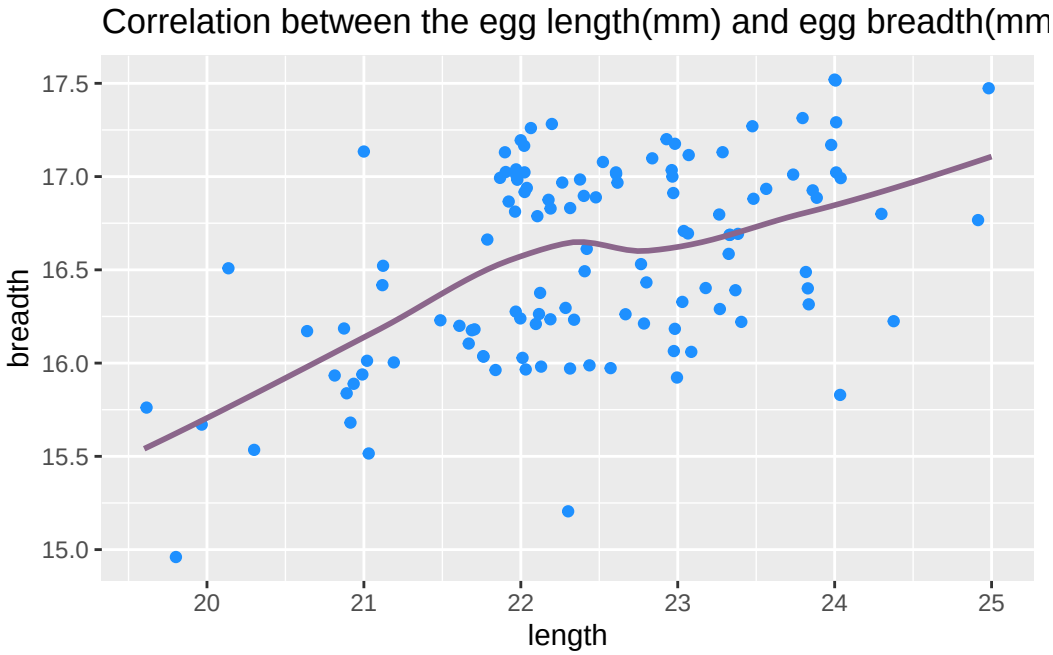
```
cor(Cuckoo_Eggs$length, Cuckoo_Eggs$breadth)
```

```
[1] 0.5017711
```

The output suggest that the length and breadth are positively correlated (the absence of a negative sign), but the value also suggests that the correlation is not strong.

This can be visualized to see a trend line:

```
Cuckoo_Eggs |>
  ggplot(aes(length,breadth))+
  geom_jitter(col= "dodgerblue")+
  geom_smooth(col = "plum4", se = FALSE, method = "loess", formula = "y~x")+
  labs(
    title = "Correlation between the egg length(mm) and egg breadth(mm) across different spec
    )
```



This plot shows that the trend line slopes upwards, indicating a positive correlation (as length of egg increases, the egg breadth may also increase as well). However, there are still a lot of points scattered around the trend line, suggesting a not so strong correlation between egg length and egg breadth.

Inferential Statistical Analysis

Breadth

```
breadth_anova <- aov(breadth~species, data = Cuckoo_Eggs) |>
  anova()

##Viewing the ANOVA table
breadth_anova
```

```
Analysis of Variance Table

Response: breadth
           Df Sum Sq Mean Sq F value    Pr(>F)
species     5 10.373 2.07457  10.885 1.412e-08 ***
Residuals 114 21.727 0.19059
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##Getting a summary of the ANOVA table
summary(breadth_anova)
```

```
       Df             Sum Sq          Mean Sq          F value          Pr(>F)
 Min.   :  5.00   Min.   :10.37   Min.   :0.1906   Min.   :10.89   Min.   :0
 1st Qu.: 32.25   1st Qu.:13.21   1st Qu.:0.6616   1st Qu.:10.89   1st Qu.:0
 Median : 59.50   Median :16.05   Median :1.1326   Median :10.89   Median :0
 Mean   : 59.50   Mean   :16.05   Mean   :1.1326   Mean   :10.89   Mean   :0
 3rd Qu.: 86.75   3rd Qu.:18.89   3rd Qu.:1.6036   3rd Qu.:10.89   3rd Qu.:0
 Max.   :114.00   Max.   :21.73   Max.   :2.0746   Max.   :10.89   Max.   :0
                                                   NA's   :1       NA's   :1
```

Post-hoc test

```
##Installing the package used to run a post-hoc test
install.packages("TukeyC", repos = "https://cran.rstudio.com/")
```

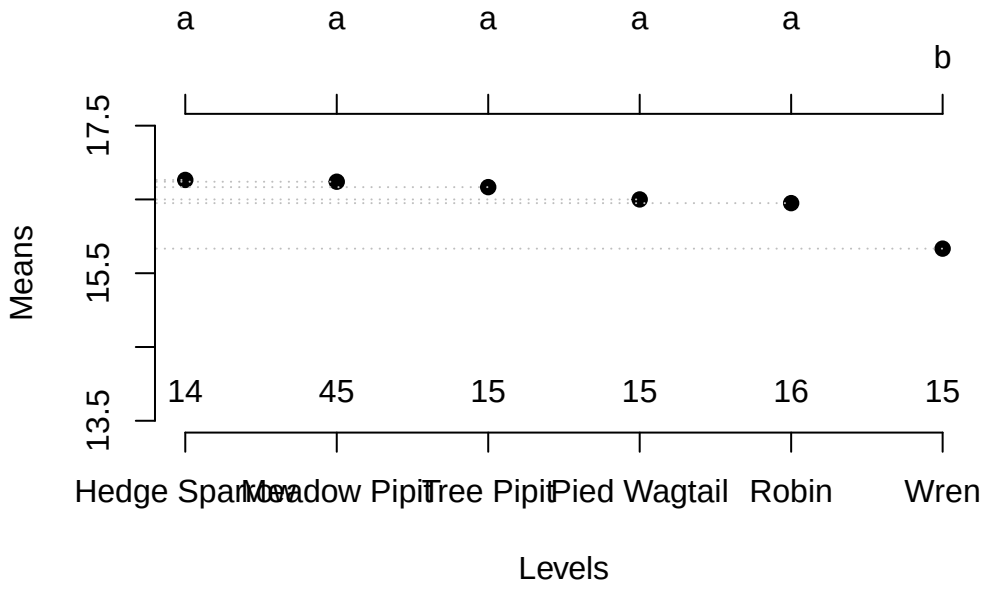
```
package 'TukeyC' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
    C:\Users\HP\AppData\Local\Temp\RtmpQlKB69\downloaded_packages
```

```
library(TukeyC)
breadth_aov <-aov(breadth~species, data = Cuckoo_Eggs)
TukeyC(breadth_aov)
```

```
Results
              Means G1 G2
Hedge Sparrow 16.76  a
Meadow Pipit  16.74  a
Tree Pipit    16.67  a
Pied Wagtail  16.50  a
Robin         16.45  a
Wren          15.83     b

Sig.level
 0.05

Diff_Prob
              Hedge Sparrow Meadow Pipit Tree Pipit Pied Wagtail Robin  Wren
```

```
Hedge Sparrow         0.000        0.024      0.098        0.264 0.314 0.931
Meadow Pipit          1.000        0.000      0.073        0.240 0.290 0.907
Tree Pipit            0.991        0.993      0.000        0.167 0.217 0.833
Pied Wagtail          0.581        0.442      0.901        0.000 0.050 0.667
Robin                 0.368        0.210      0.738        1.000 0.000 0.617
Wren                  0.000        0.000      0.000        0.001 0.002 0.000

MSD
              Hedge Sparrow Meadow Pipit Tree Pipit Pied Wagtail Robin  Wren
Hedge Sparrow         0.000        0.387      0.470        0.470 0.463 0.470
Meadow Pipit          0.387        0.000      0.377        0.377 0.368 0.377
Tree Pipit            0.470        0.377      0.000        0.462 0.455 0.462
Pied Wagtail          0.470        0.377      0.462        0.000 0.455 0.462
Robin                 0.463        0.368      0.455        0.455 0.000 0.455
Wren                  0.470        0.377      0.462        0.462 0.455 0.000
```

```r
##Visualizing the differences
TukeyC(breadth_aov) |>    plot()
```



Length

```r
length_anova <- anova(aov(length ~ species, data= Cuckoo_Eggs))

##Viewing the ANOVA table
length_anova
```

```
Analysis of Variance Table

Response: length
           Df Sum Sq Mean Sq F value    Pr(>F)
species     5  42.81  8.5620  10.449 2.852e-08 ***
Residuals 114  93.41  0.8194
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
##Getting a summary of the ANOVA table
summary(length_anova)
```

```
       Df             Sum Sq         Mean Sq          F value          Pr(>F)
 Min.   :  5.00   Min.   :42.81   Min.   :0.8194   Min.   :10.45   Min.   :0
 1st Qu.: 32.25   1st Qu.:55.46   1st Qu.:2.7550   1st Qu.:10.45   1st Qu.:0
 Median : 59.50   Median :68.11   Median :4.6907   Median :10.45   Median :0
 Mean   : 59.50   Mean   :68.11   Mean   :4.6907   Mean   :10.45   Mean   :0
 3rd Qu.: 86.75   3rd Qu.:80.76   3rd Qu.:6.6264   3rd Qu.:10.45   3rd Qu.:0
 Max.   :114.00   Max.   :93.41   Max.   :8.5620   Max.   :10.45   Max.   :0
                                                   NA's   :1       NA's   :1
```

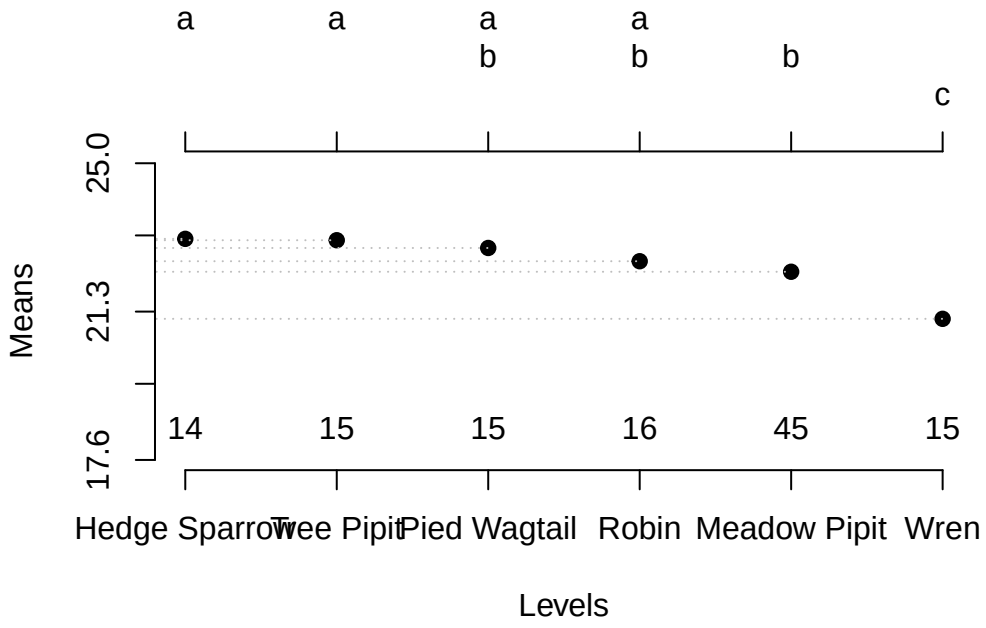
Post-hoc test

```r
length_aov <-aov(length ~ species, data= Cuckoo_Eggs)
TukeyC(length_aov)
```

```
Results
              Means G1 G2 G3
Hedge Sparrow 23.11  a
Tree Pipit    23.08  a
Pied Wagtail  22.89  a  b
Robin         22.56  a  b
Meadow Pipit  22.29     b
Wren          21.12        c
```

```
Sig.level
 0.05

Diff_Prob
              Hedge Sparrow Tree Pipit Pied Wagtail Robin Meadow Pipit  Wren
Hedge Sparrow         0.000      0.034        0.228 0.558        0.821 1.994
Tree Pipit            1.000      0.000        0.193 0.524        0.787 1.960
Pied Wagtail          0.984      0.992        0.000 0.330        0.593 1.767
Robin                 0.545      0.594        0.912 0.000        0.263 1.436
Meadow Pipit          0.042      0.048        0.247 0.918        0.000 1.173
Wren                  0.000      0.000        0.000 0.000        0.000 0.000

MSD
              Hedge Sparrow Tree Pipit Pied Wagtail Robin Meadow Pipit  Wren
Hedge Sparrow         0.000      0.975        0.975 0.960        0.803 0.975
Tree Pipit            0.975      0.000        0.958 0.943        0.782 0.958
Pied Wagtail          0.975      0.958        0.000 0.943        0.782 0.958
Robin                 0.960      0.943        0.943 0.000        0.764 0.943
Meadow Pipit          0.803      0.782        0.782 0.764        0.000 0.782
Wren                  0.975      0.958        0.958 0.943        0.782 0.000
```

```r
##Visualizing the differences
TukeyC(length_aov) |>
  plot()
```



Summary

From the results of the ANOVA on the breadth of eggs, the p-value is less than 0.05, implying that there is significant difference in the breadths of the eggs of differnet species. The Tukey's post-hoc test showed where exactly the difference lies. The eggs of all species except the Wren have similar mean breadths ranging from 16.45 to 16.76 which is not significantly different. However the egg of the Wren has a mean breadth of 15.83 which is significantly different and smaller from the other species. This is also denoted by the grouping "b". This is also visualized with the plot.

For the length of eggs, the ANOVA result also implies a significant difference in the egg length of the different species. To confirm where the differences exist, the post hoc test was performed and it showed that the Wren egg length was significantly different from all other species. The Robin and Pied Wagtail grouped as 'b' had mean egg lengths that were significantly different from the Wren but not different from the other species. The Meadow Pipit's eggs are significantly shorter than those of Hedge Sparrow and Tree Pipit but not significantly different from Robin and Pied Wagtail, positioning it in group "b".

From this analysis, we can conclude that significant differences exist in the dimensions of the eggs laid by the the differnt cuckoo species. The egg of the Wren can be considered the smallest having the lowest mean length and breadth.

References

Maindonald, John H, W John Braun, and Maintainer W John Braun. 2015. "Package 'DAAG'." *Data Analysis and Graphics Data and Functions*.