Penguins Report

By: Anjola Okesola, Cleo Anne Tabacolde, Emily Dalson, Huong Giang Ho, and Lina Montesinos

The Dataset

- A sample of 119 penguins of the Gentoo, Adelie, and Chinstrap species.
- The penguins originate from the Torgersen, Biscoe, and Dream islands.
- Clutch completion, culmen length, culmen depth, flipper length, body mass, and sex represent the data across the penguin species.

Description of Dataset

Variables & Purpose

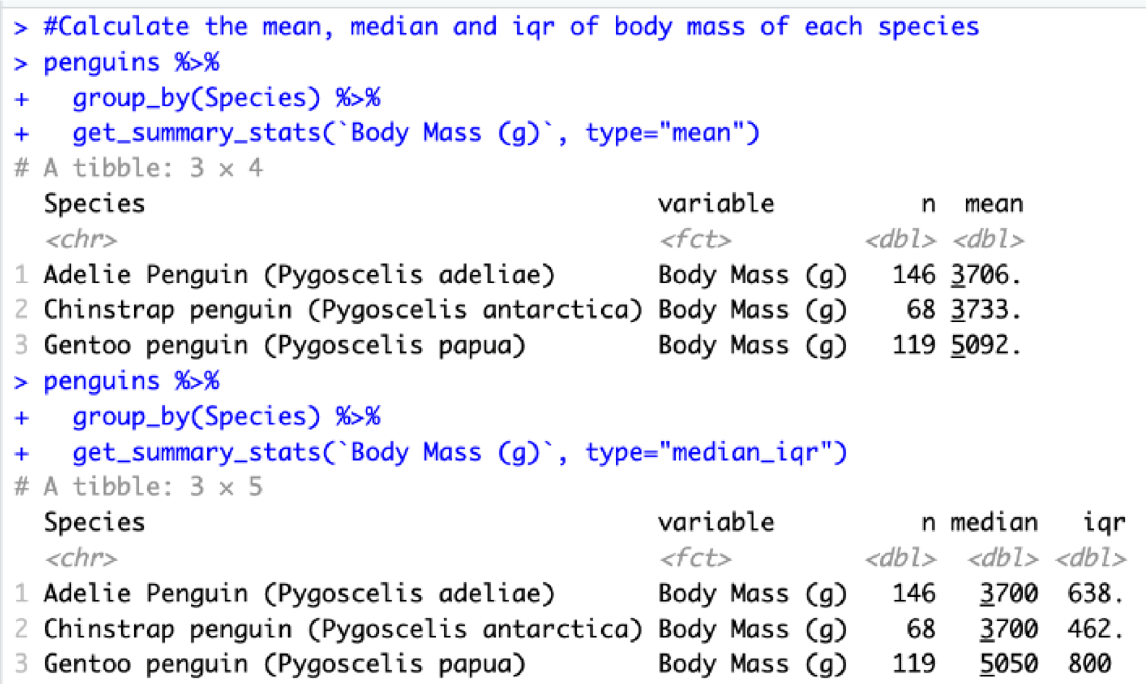
- **Species**: Used to group data and compare average body mass across species.
- **Body Mass:** Measures weight in grams; used to compare between species and sexes.
- **Sex**: Identifies male or female; used to analyze differences in physical traits.
- **Flipper Length**: Length in mm; used to compare size between sexes.
- **Culmen Length** : Beak length in mm; used to examine sex-based differences.
- **Culmen Depth**: Beak depth in mm; used alongside length to study beak variation.

Key Relationships

- Species and Traits: Physical traits vary by species
- Sex and Traits: Males generally have larger body mass, flipper length, and culmen size
- Species and Island: Certain species are tied to specific islands
- Trait Correlations: Physical features are often interrelated

Descriptive Statistics

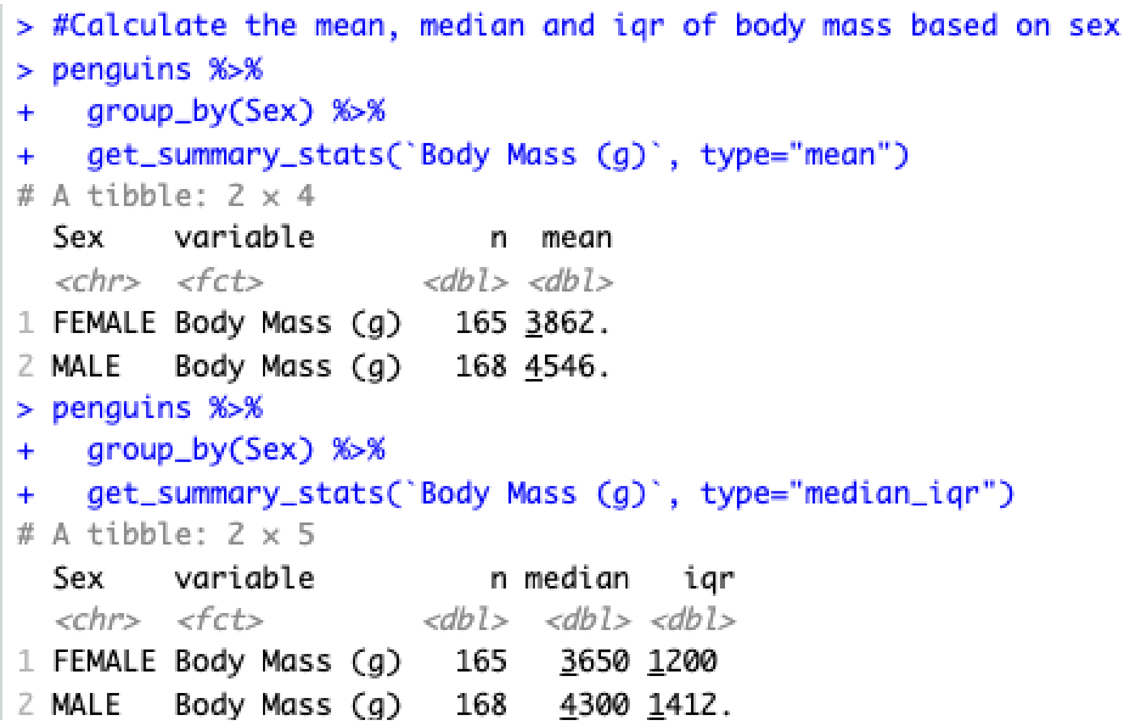
- Mean, median, iqr of penguin species



```
> #Calculate the mean, median and iqr of body mass of each species
> penguins %>%
+   group_by(Species) %>%
+   get_summary_stats(`Body Mass (g)`, type="mean")
# A tibble: 3 x 4
  Species                                   variable          n  mean
  <chr>                                     <fct>         <dbl> <dbl>
1 Adelie Penguin (Pygoscelis adeliae)       Body Mass (g)   146 3706.
2 Chinstrap penguin (Pygoscelis antarctica) Body Mass (g)    68 3733.
3 Gentoo penguin (Pygoscelis papua)         Body Mass (g)   119 5092.
> penguins %>%
+   group_by(Species) %>%
+   get_summary_stats(`Body Mass (g)`, type="median_iqr")
# A tibble: 3 x 5
  Species                                   variable          n median   iqr
  <chr>                                     <fct>         <dbl>  <dbl> <dbl>
1 Adelie Penguin (Pygoscelis adeliae)       Body Mass (g)   146   3700  638.
2 Chinstrap penguin (Pygoscelis antarctica) Body Mass (g)    68   3700  462.
3 Gentoo penguin (Pygoscelis papua)         Body Mass (g)   119   5050  800
```

Descriptive Statistics

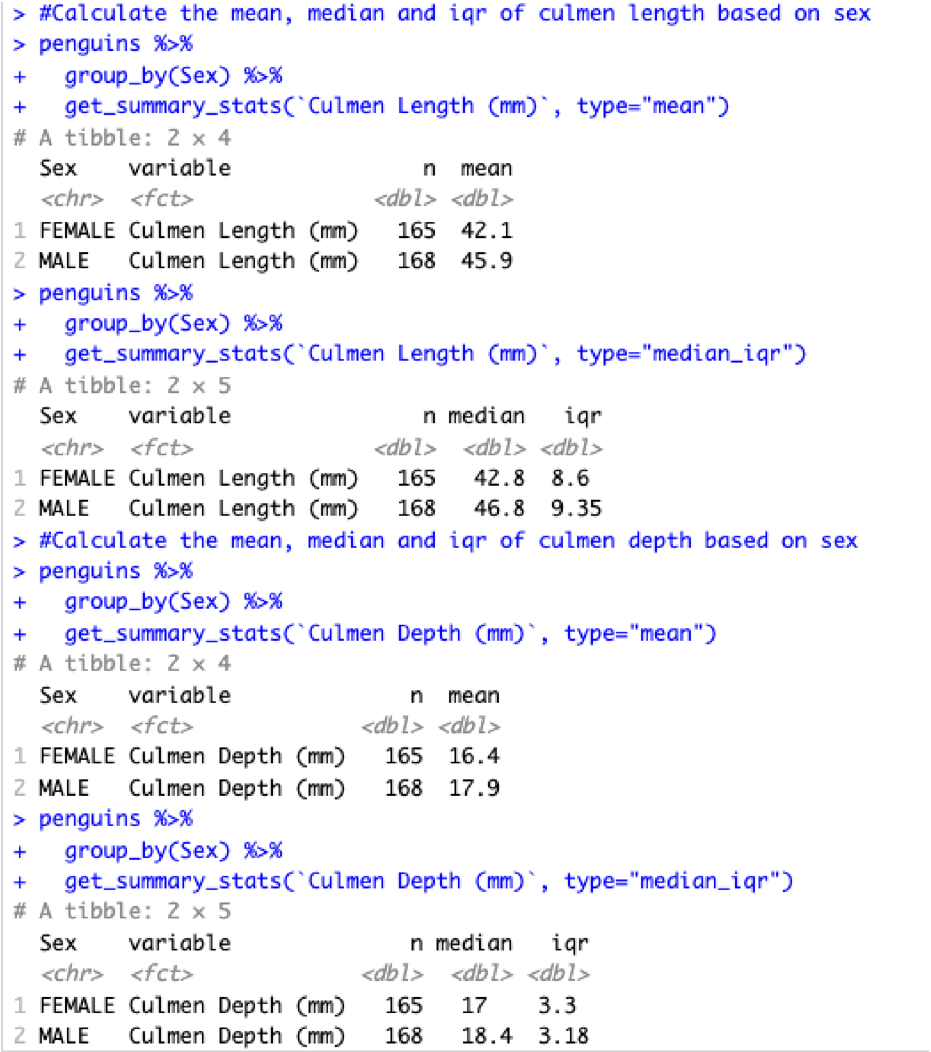
- Mean, median, iqr of body mass on sex



```
> #Calculate the mean, median and iqr of body mass based on sex
> penguins %>%
+   group_by(Sex) %>%
+   get_summary_stats(`Body Mass (g)`, type="mean")
# A tibble: 2 x 4
  Sex    variable          n  mean
  <chr>  <fct>         <dbl> <dbl>
1 FEMALE Body Mass (g)   165 3862.
2 MALE   Body Mass (g)   168 4546.
> penguins %>%
+   group_by(Sex) %>%
+   get_summary_stats(`Body Mass (g)`, type="median_iqr")
# A tibble: 2 x 5
  Sex    variable          n median   iqr
  <chr>  <fct>         <dbl>  <dbl> <dbl>
1 FEMALE Body Mass (g)   165   3650 1200
2 MALE   Body Mass (g)   168   4300 1412.
```

Descriptive Statistics

- Mean, median, iqr of culmen length and depth based on sex



```
> #Calculate the mean, median and iqr of culmen length based on sex
> penguins %>%
+   group_by(Sex) %>%
+   get_summary_stats(`Culmen Length (mm)`, type="mean")
# A tibble: 2 x 4
  Sex    variable               n  mean
  <chr>  <fct>              <dbl> <dbl>
1 FEMALE Culmen Length (mm)   165  42.1
2 MALE   Culmen Length (mm)   168  45.9
> penguins %>%
+   group_by(Sex) %>%
+   get_summary_stats(`Culmen Length (mm)`, type="median_iqr")
# A tibble: 2 x 5
  Sex    variable               n median   iqr
  <chr>  <fct>              <dbl>  <dbl> <dbl>
1 FEMALE Culmen Length (mm)   165   42.8  8.6 
2 MALE   Culmen Length (mm)   168   46.8  9.35
> #Calculate the mean, median and iqr of culmen depth based on sex
> penguins %>%
+   group_by(Sex) %>%
+   get_summary_stats(`Culmen Depth (mm)`, type="mean")
# A tibble: 2 x 4
  Sex    variable              n  mean
  <chr>  <fct>             <dbl> <dbl>
1 FEMALE Culmen Depth (mm)   165  16.4
2 MALE   Culmen Depth (mm)   168  17.9
> penguins %>%
+   group_by(Sex) %>%
+   get_summary_stats(`Culmen Depth (mm)`, type="median_iqr")
# A tibble: 2 x 5
  Sex    variable              n median   iqr
  <chr>  <fct>             <dbl>  <dbl> <dbl>
1 FEMALE Culmen Depth (mm)   165   17    3.3 
2 MALE   Culmen Depth (mm)   168   18.4  3.18
```

Research Question

- Will there be a significant difference in culmen depth across the three penguin's species (Gentoo, Adelie, and Chinstrap)

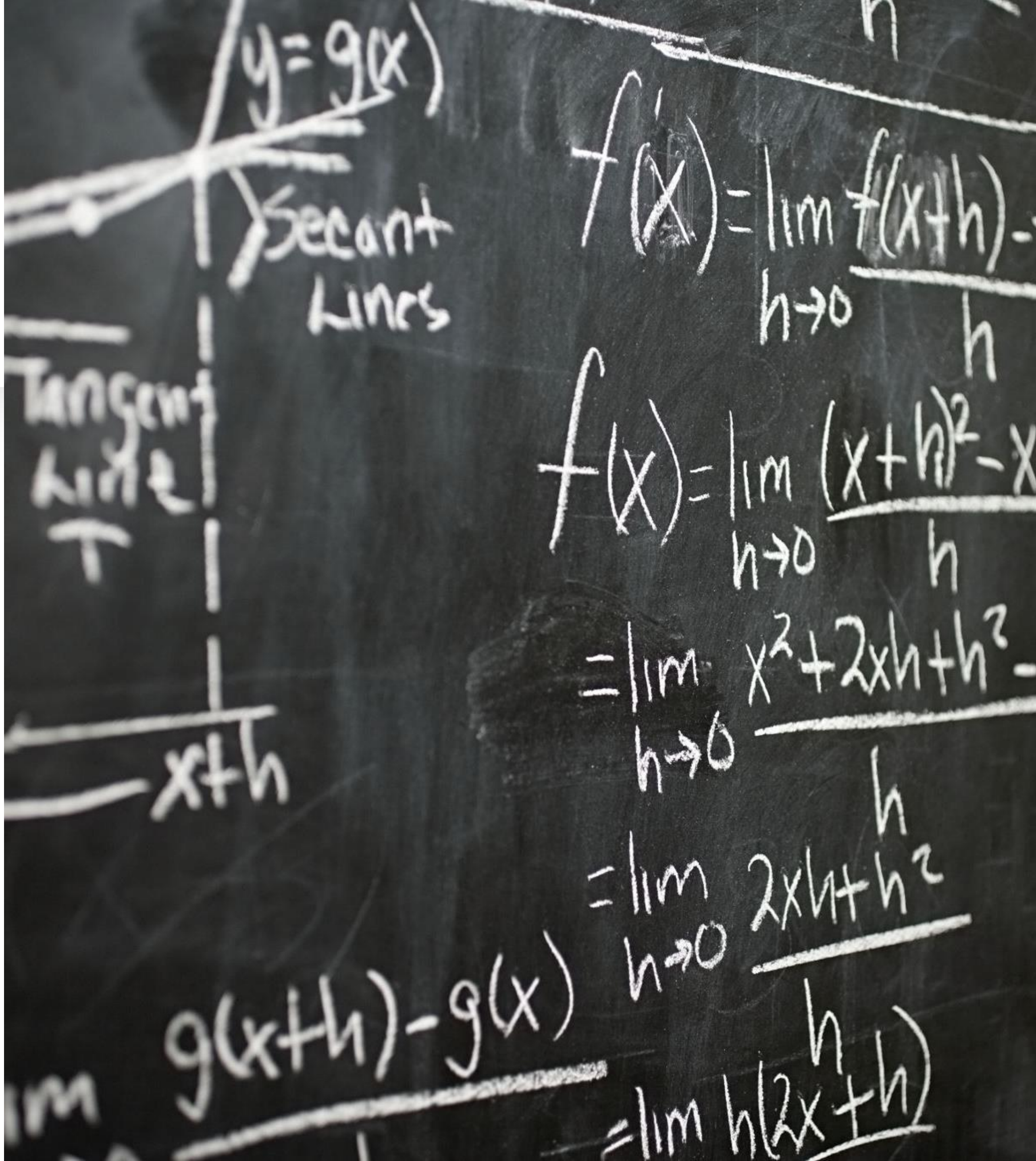
Hypotheses

Null Hypothesis (H_0): There is no difference in culmen depth mean across the three penguin species.

H_0 : $\mu_1 = \mu_2 = \mu_3$

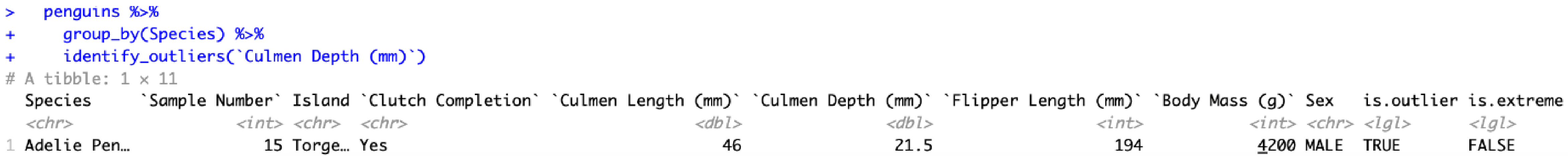
Alternative Hypothesis (H_a): At least one species has a different mean culmen depth.

H_a : $\mu_1 \neq \mu_2 \neq \mu_3$



Analysis Method

Look for outlier: One non extreme outlier is identified



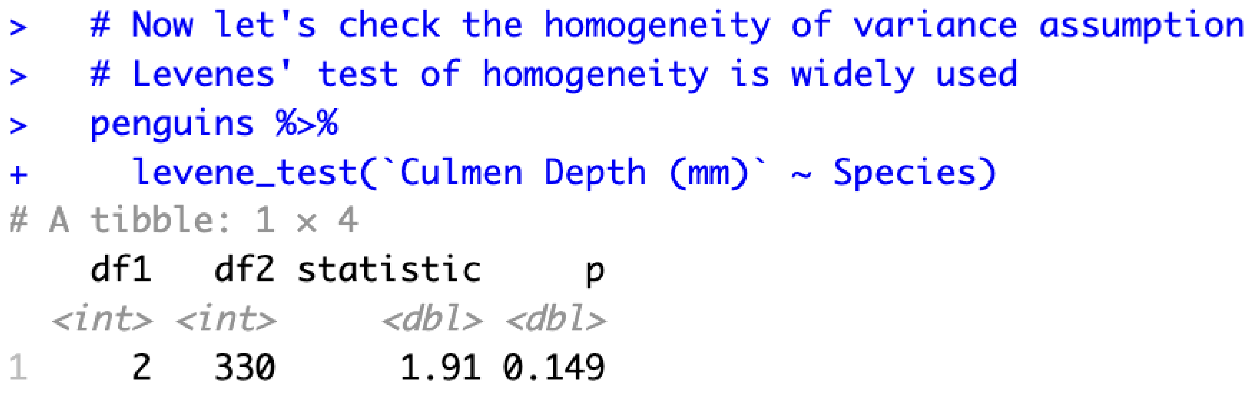
```
>   penguins %>%
+     group_by(Species) %>%
+     identify_outliers(`Culmen Depth (mm)`)
# A tibble: 1 × 11
  Species      `Sample Number` Island `Clutch Completion` `Culmen Length (mm)` `Culmen Depth (mm)` `Flipper Length (mm)` `Body Mass (g)` Sex   is.outlier is.extreme
  <chr>                  <int> <chr>  <chr>                              <dbl>               <dbl>                 <int>           <int> <chr> <lgl>      <lgl>     
1 Adelie Pen…               15 Torge… Yes                                   46                21.5                   194            4200 MALE  TRUE       FALSE     
```

Normality test:

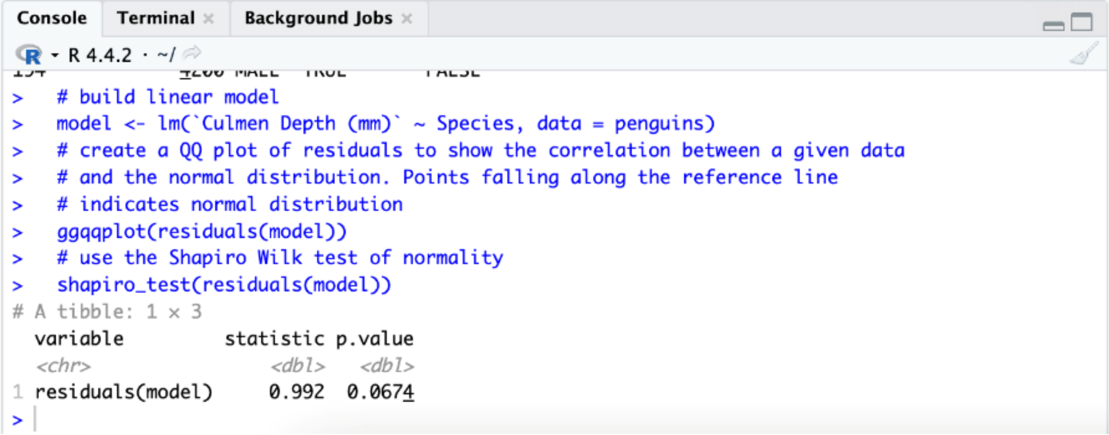
+ Build linear model between species and culmen depth

+ Create qq plot: points do fall along the reference line indicate normal distribution

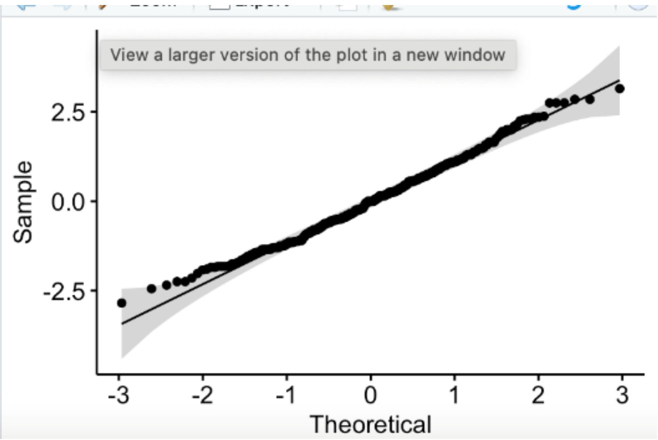
+ Shapiro Wilk test: p = 0.067 > 0.05 => normal distribution



```
>   # Now let's check the homogeneity of variance assumption
>   # Levenes' test of homogeneity is widely used
>   penguins %>%
+     levene_test(`Culmen Depth (mm)` ~ Species)
# A tibble: 1 × 4
    df1   df2 statistic     p
  <int> <int>     <dbl> <dbl>
1     2   330      1.91 0.149
```



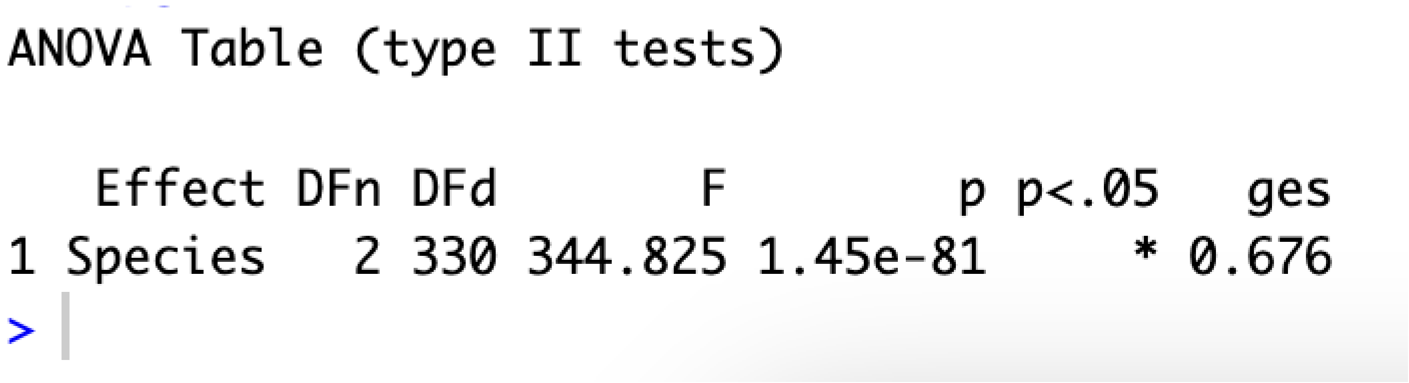
```
194    4200 MALE   TRUE        FALSE
>   # build linear model
>   model <- lm(`Culmen Depth (mm)` ~ Species, data = penguins)
>   # create a QQ plot of residuals to show the correlation between a given data
>   # and the normal distribution. Points falling along the reference line
>   # indicates normal distribution
>   ggqqplot(residuals(model))
>   # use the Shapiro Wilk test of normality
>   shapiro_test(residuals(model))
# A tibble: 1 × 3
  variable        statistic p.value
  <chr>               <dbl>   <dbl>
1 residuals(model)    0.992  0.0674
> 
```



Levenes' test of homogeneity: check the homogeneity of variance assumption: p = 0.149 > 0.05 => non-significant, confirming homogeneity of variance

Analysis Method

Compute ANOVA test:
p = 1.45e-81<0.05 => there is significant difference among the three groups

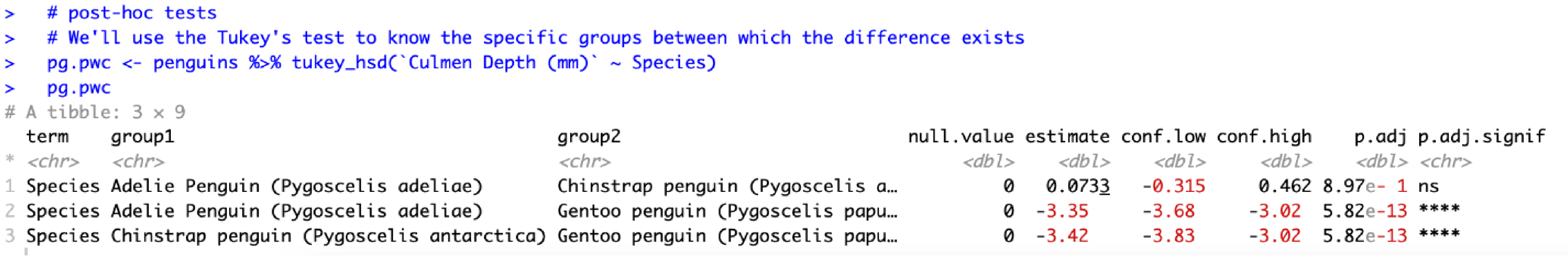


```
ANOVA Table (type II tests)

    Effect DFn DFd       F        p p<.05   ges
1 Species   2 330 344.825 1.45e-81     * 0.676
> 
```

Post-hoc tests:
Tukey's test to know the significant between each pair

+ Adelie vs. Chinstrap: p = 8.97e- 1 = 0.897 (ns) > 0.05=> No significant difference between these two species
+ Adelie vs. Gentoo: p = 5.82e-13 < 0.05 => Extremely significant difference
+ Chinstrap vs. Gentoo: p = 5.82e-13 < 0.05 => Extremely significant difference

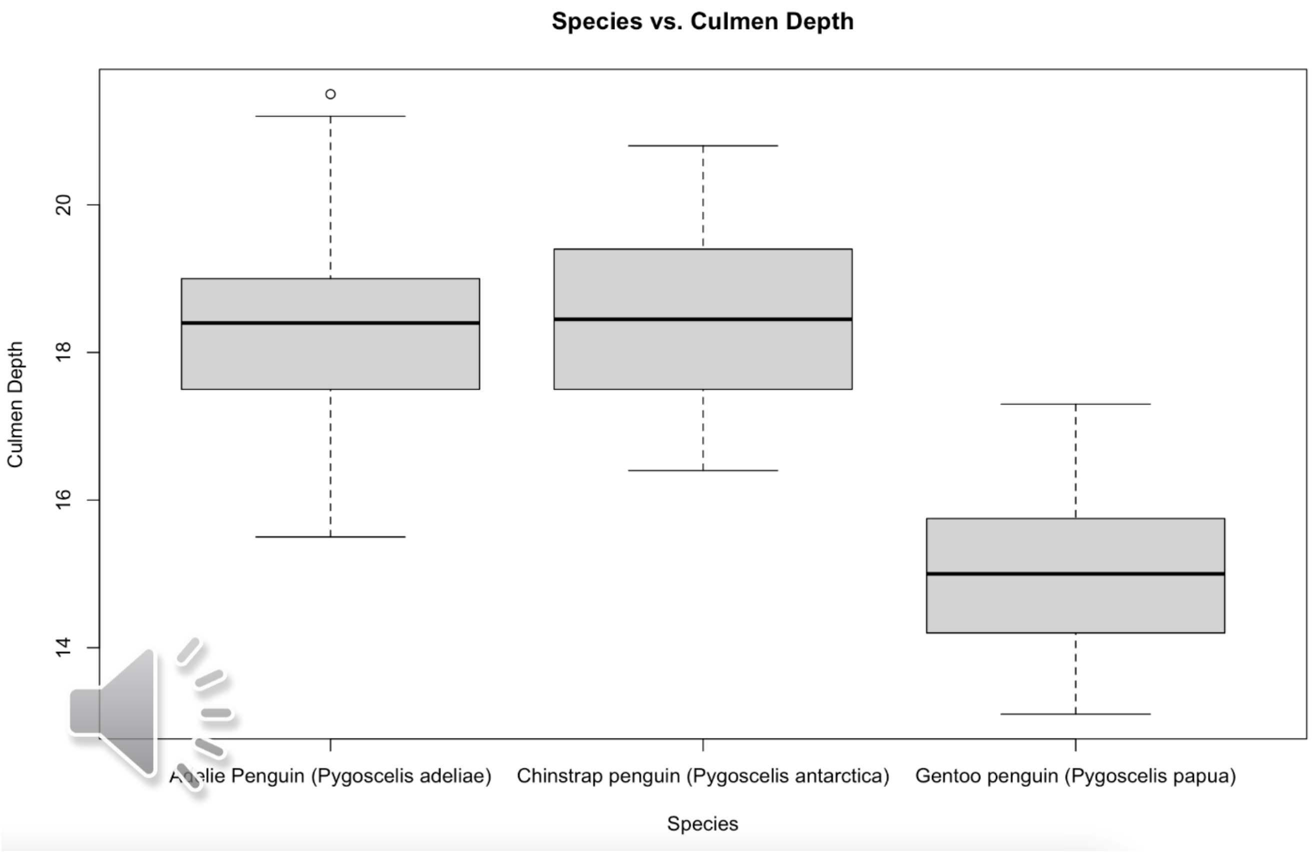


```
>   # post-hoc tests
>   # We'll use the Tukey's test to know the specific groups between which the difference exists
>   pg.pwc <- penguins %>% tukey_hsd(`Culmen Depth (mm)` ~ Species)
>   pg.pwc
# A tibble: 3 × 9
  term    group1                             group2                             null.value estimate conf.low conf.high    p.adj p.adj.signif
* <chr>   <chr>                              <chr>                                   <dbl>    <dbl>    <dbl>     <dbl>    <dbl> <chr>
1 Species Adelie Penguin (Pygoscelis adeliae)       Chinstrap penguin (Pygoscelis a…          0   0.0733   -0.315     0.462 8.97e- 1 ns
2 Species Adelie Penguin (Pygoscelis adeliae)       Gentoo penguin (Pygoscelis papu…          0  -3.35     -3.68     -3.02  5.82e-13 ****
3 Species Chinstrap penguin (Pygoscelis antarctica) Gentoo penguin (Pygoscelis papu…          0  -3.42     -3.83     -3.02  5.82e-13 ****
```

Results

1. Outlier Analysis
A preliminary outlier analysis using the interquartile range (IQR) method identified one non-extreme outlier (Culmen Depth = 21.5 mm) in the Adelie penguin group. The outlier was retained for analysis as it did not exceed the threshold for extreme outliers.

2. Normality Analysis
Normality of distribution was evaluated using both graphical and statistical methods: QQ plot inspection revealed that the distribution is normal with data falling along the reference line. Shapiro Wilk test also confirmed normality in distribution with p = 0.067 > 0.05



3. Homogeneity of variance assumption
Levene's test of homogeneity indicated no significant heterogeneity across groups (p = 0.149),satisfying the assumption for ANOVA.

4. One-Way ANOVA
A one-way ANOVA revealed significant differences in culmen depth among the three penguin species ($p = 1.45 \times 10^{-81}$).

5. Post-Hoc Comparisons
Pairwise comparisons using Tukey's test showed that Gentoo penguins exhibit significantly different culmen depth from both Adelie and Chinstrap species (p = 5.82e-13 < 0.05). No significant difference was detected between Adelie and Chinstrap penguins (p = 0.897 > 0.05)

Cleo

Interpretation of Results

- Statistical analysis shows culmen depth significantly differs among species
- Gentoo penguins have distinct culmen depth compared to Adelie and Chinstrap
- No significant difference between Adelie and Chinstrap
- Culmen depth is a reliable trait for identifying Gentoo penguins but less effective for distinguishing Adelie vs. Chinstrap
- Other physical traits (like body mass, flipper length, etc.) also vary by species and sex

Suggestions:

- Focus further analysis on interactions between sex and physical traits
- Aim to better understand gender-based differences across species

Cleo