Penguins

Maria Dermit

28/07/2020

```
## Getting the data
tuesdata <- tidytuesdayR::tt_load('2020-07-28')
```

```
##
##  Downloading file 1 of 2: `penguins.csv`
##  Downloading file 2 of 2: `penguins_raw.csv`
```

```
tuesdata <- tidytuesdayR::tt_load(2020, week = 31)
```

```
##
##  Downloading file 1 of 2: `penguins.csv`
##  Downloading file 2 of 2: `penguins_raw.csv`
```

```
penguins <- tuesdata$penguins
penguins %>% View()
```

Exploring the data

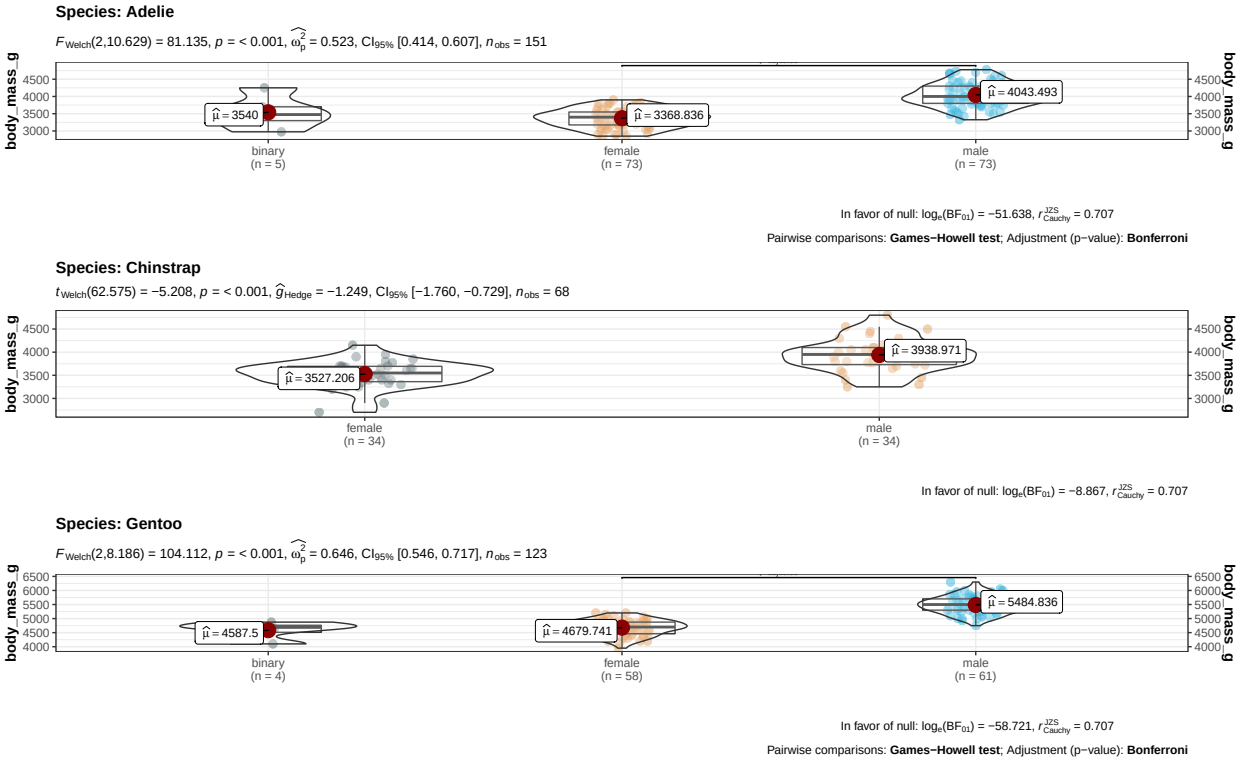
penguins %>% group_by(species,sex) %>% summarise(mean = mean(body_mass_g), n = n())

```
penguins %>%
split(.$species) %>%
 map(~ lm(body_mass_g ~ sex, data = .x))
```

```
## $Adelie
##
## Call:
## lm(formula = body_mass_g ~ sex, data = .x)
##
## Coefficients:
## (Intercept)      sexmale
##      3368.8        674.7
##
##
## $Chinstrap
##
## Call:
## lm(formula = body_mass_g ~ sex, data = .x)
##
## Coefficients:
## (Intercept)      sexmale
##      3527.2        411.8
##
##
```

```
## $Gentoo
##
## Call:
## lm(formula = body_mass_g ~ sex, data = .x)
##
## Coefficients:
## (Intercept)      sexmale
##      4679.7        805.1
```

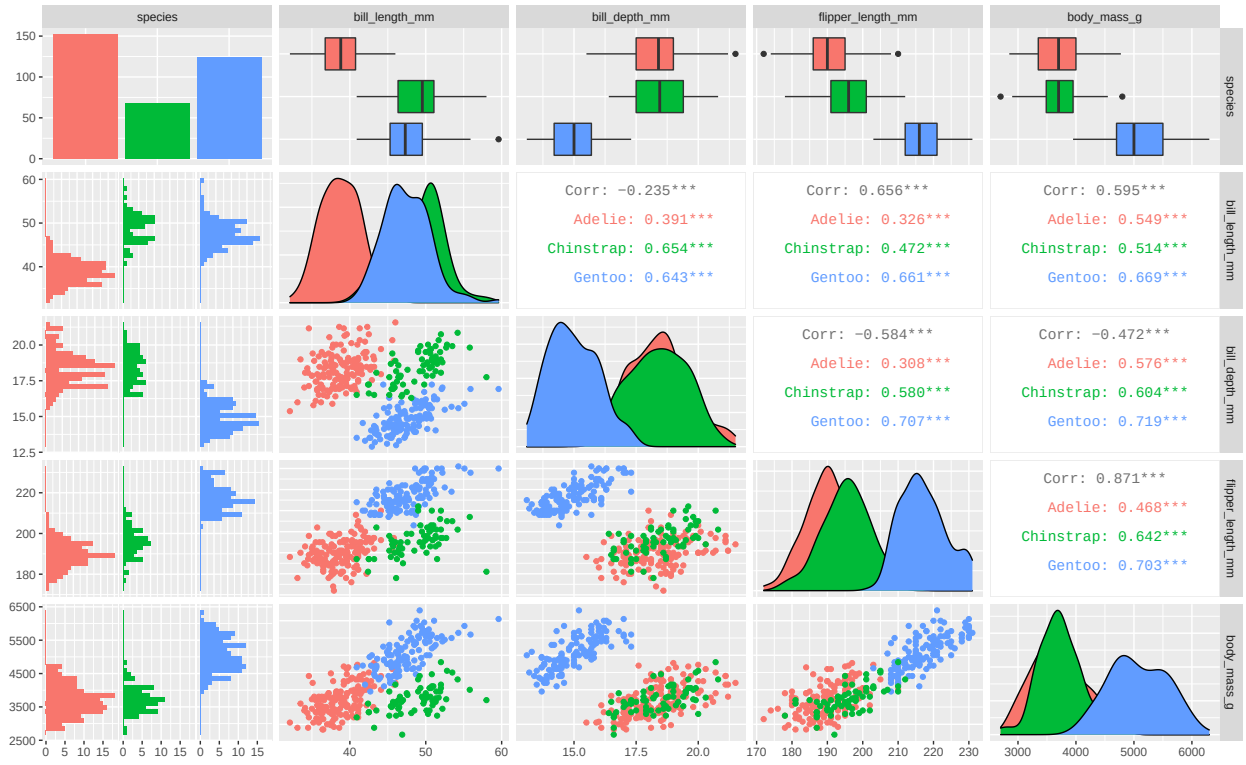
```
#Rename the sex parameter to called "NA" as "binary"
penguins<-penguins %>%  mutate_if(is.character,as.factor)  %>%
mutate(sex = case_when(
sex=="female" ~ "female",
sex=="male" ~ "male",
 TRUE ~ "binary",
))
# plot
ggstatsplot::grouped_ggbetweenstats(
  data = penguins,
  x = sex,
  xlab="",
  y = body_mass_g,
  grouping.var = species, # grouping variable
  pairwise.comparisons = TRUE, # display significant pairwise comparisons
  p.adjust.method = "bonferroni", # method for adjusting p-values for multiple comparisons
  # adding new components to `ggstatsplot` default
  ggplot.component = list(ggplot2::scale_y_continuous(sec.axis = ggplot2::dup_axis())),
  k = 3,
  title.prefix = "Species",
  palette = "default_jama",
  package = "ggsci",
  plotgrid.args = list(nrow = 3),
)
```



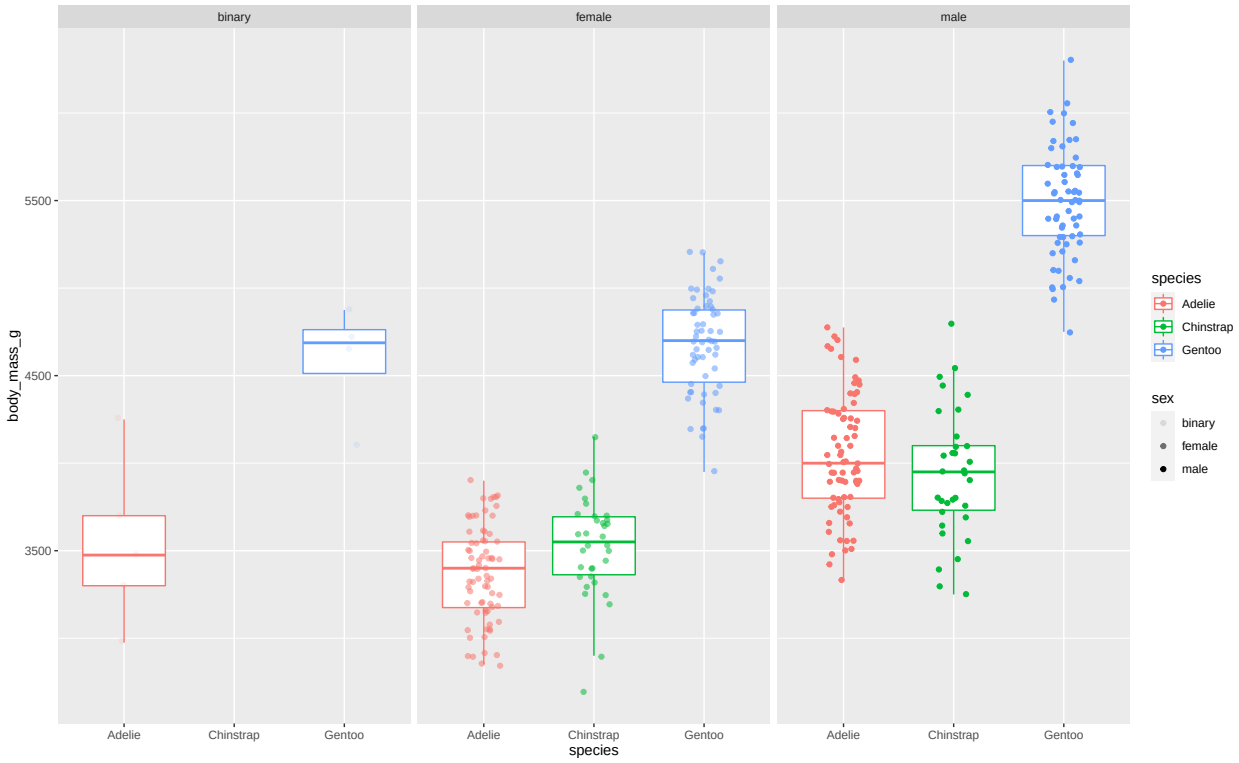

```
#ggsave("/Documents/TidyTuesdays/tidyversewebninar/20200728/plots/grouped_ggbetweenstats.jpg")
```

```
penguins_clean = penguins %>%
  mutate(species = factor(species))

penguins_clean %>%
  select(species,bill_length_mm,bill_depth_mm, flipper_length_mm    ,body_mass_g) %>% ggpairs(mapping =
```




```
penguins_clean %>%
 select(sex,species,bill_length_mm,bill_depth_mm, flipper_length_mm ,body_mass_g) %>%
  ggplot(aes(species, body_mass_g, color = species)) +
  geom_boxplot(outlier.colour = NA) +
  geom_jitter(aes(alpha = sex), width = 0.15) +
  facet_wrap(~sex)
```



Gentoo male penguins are the the chubbiest among these three species. It looks like more body weight is associated with longer flipper length and with less bill depth (so more pointy mouth).Lets do a linear regression model for body weigths of these cute penguins.

Modeling

```
line_fit <- lm(body_mass_g~flipper_length_mm, data=penguins_clean)
line_smry <- summary(line_fit)
line_smry
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = penguins_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1058.80  -259.27   -26.88   247.33  1288.69
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5780.831    305.815  -18.90   <2e-16 ***
## flipper_length_mm    49.686      1.518   32.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.3 on 340 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.7583
## F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

The R^2 value for this model is 0.759, which means that this model explains 75.9% of the variance (not sufficient). We gain some information from our hypothesis test using this model.

We can use step function to choose a the model with lowest AIC.

```
step_fit=step(lm(data=penguins_clean, body_mass_g ~ .),trace=0,steps=10000)
step_smry <- summary(step_fit)
step_smry
```

```
##
## Call:
## lm(formula = body_mass_g ~ species + bill_length_mm + bill_depth_mm +
##     flipper_length_mm + sex + year, data = penguins_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -811.16 -181.48   -7.36  183.57  874.76
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       78566.673  41406.696   1.897 0.058635 .
## speciesChinstrap   -295.945     81.477  -3.632 0.000325 ***
## speciesGentoo       924.181    135.715   6.810 4.58e-11 ***
## bill_length_mm       20.875      7.059   2.957 0.003329 **
## bill_depth_mm        64.386     19.720   3.265 0.001208 **
## flipper_length_mm    17.748      3.080   5.762 1.89e-08 ***
## sexfemale            49.597    100.018   0.496 0.620305
## sexmale             421.614    103.024   4.092 5.36e-05 ***
## year                -40.067     20.701  -1.936 0.053774 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 288.4 on 333 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8737, Adjusted R-squared:  0.8707
## F-statistic:   288 on 8 and 333 DF,  p-value: < 2.2e-16
```

```
best_fit <-lm(formula = body_mass_g ~ species + bill_length_mm + bill_depth_mm +
    flipper_length_mm + sex + year, data = penguins_clean)

best_smry <- summary(best_fit)
best_smry
```

```
##
## Call:
## lm(formula = body_mass_g ~ species + bill_length_mm + bill_depth_mm +
##     flipper_length_mm + sex + year, data = penguins_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -811.16 -181.48   -7.36  183.57  874.76
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       78566.673  41406.696   1.897 0.058635 .
## speciesChinstrap   -295.945     81.477  -3.632 0.000325 ***
```

```
## speciesGentoo       924.181    135.715   6.810 4.58e-11 ***
## bill_length_mm       20.875      7.059   2.957 0.003329 **
## bill_depth_mm        64.386     19.720   3.265 0.001208 **
## flipper_length_mm    17.748      3.080   5.762 1.89e-08 ***
## sexfemale            49.597    100.018   0.496 0.620305
## sexmale             421.614    103.024   4.092 5.36e-05 ***
## year                -40.067     20.701  -1.936 0.053774 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 288.4 on 333 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8737, Adjusted R-squared:  0.8707
## F-statistic:   288 on 8 and 333 DF,  p-value: < 2.2e-16
```

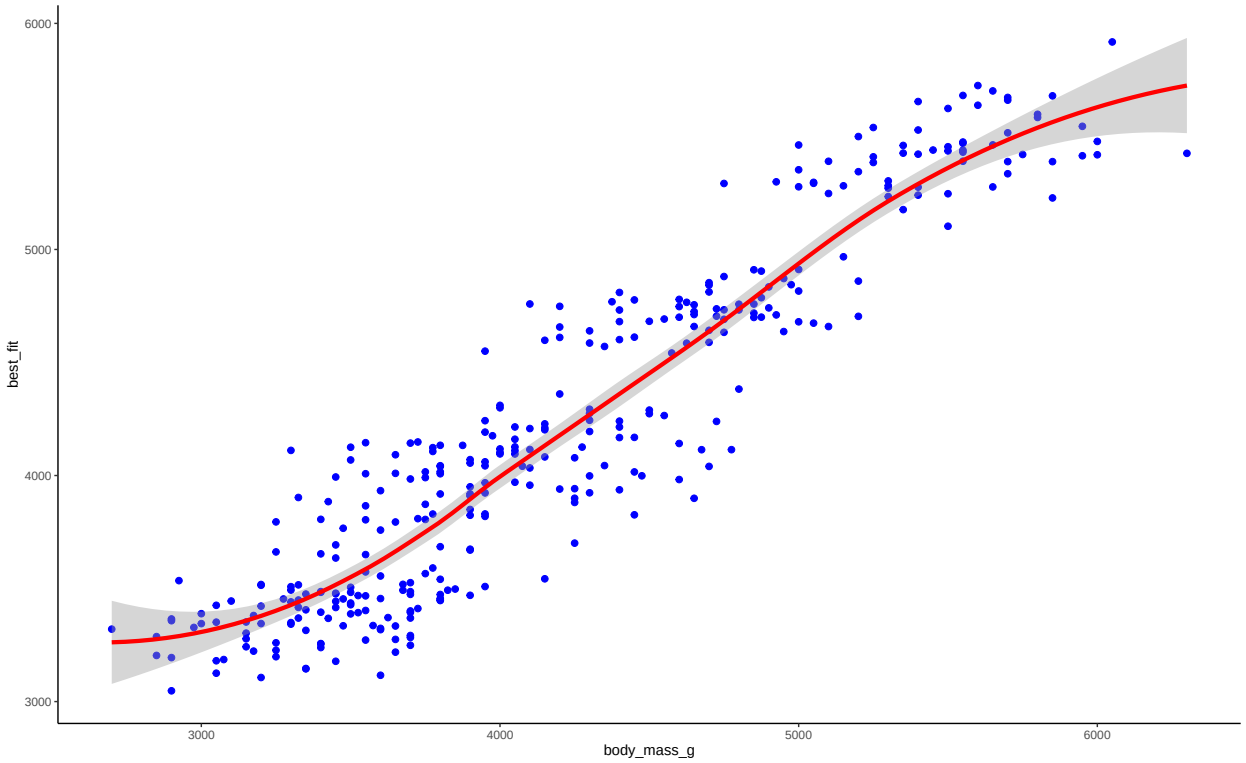
This model captures 87.64% of total variance in body weight. The p-value: < 2.2e-16 This looks like a more robust highly predictive model.

```
library(broom)
glance(best_fit) %>%
select( sigma, logLik, AIC, BIC, df.residual) %>%
  kable()
```

sigma	logLik	AIC	BIC	df.residual
288.3741	-2417.893	4855.786	4894.135	333

```
penguins_clean$best_fit <- stats::predict(best_fit, newdata=penguins_clean)
err <- stats::predict(best_fit, newdata=penguins_clean, se = TRUE)
pred.int <-predict(best_fit, newdata = penguins_clean, interval = "confidence")
penguins_clean_conf <- cbind(penguins_clean, pred.int)

g <- ggplot(penguins_clean_conf)
g <- g + geom_point(aes(x=body_mass_g, y = best_fit), size = 2, colour = "blue")
g <- g + geom_smooth(data=penguins_clean_conf, aes(x=body_mass_g, y=best_fit, ymin=lwr, ymax=upr), size
            colour = "red", se = TRUE, stat = "smooth") +
  theme_classic()
  g
```


6000
5000
4000
3000
best_fit
3000
4000
5000
6000
body_mass_g