Week 02 Version control

Open and reproducible science: dependable computations and statistics

Homework - solution

4 - R Markdown

4.3

4.3.1

Print the summary of the `penguins` dataset from the `palmerpenguins` R package using `knitr::kable`. How many observations are there? How many different species of penguins?

```
library(palmerpenguins)
library(ggplot2)
```

```
knitr::kable(summary(penguins))
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie :152	Biscoe :168	Min. :32.10	Min. :13.10	Min. :172.0	Min. :2700	female:165	Min. :2007
Chinstrap: 68	Dream :124	1st Qu.:39.23	1st Qu.:15.60	1st Qu.:190.0	1st Qu.:3550	male :168	1st Qu.:2007
Gentoo :124	Torgersen: 52	Median :44.45	Median :17.30	Median :197.0	Median :4050	NA's : 11	Median :2008
NA	NA	Mean :43.92	Mean :17.15	Mean :200.9	Mean :4202	NA	Mean :2008
NA	NA	3rd Qu.:48.50	3rd Qu.:18.70	3rd Qu.:213.0	3rd Qu.:4750	NA	3rd Qu.:2009
NA	NA	Max. :59.60	Max. :21.50	Max. :231.0	Max. :6300	NA	Max. :2009
NA	NA	NA's :2	NA's :2	NA's :2	NA's :2	NA	NA

4.3.2

Order the dataset by `bill_length_mm` (ascending) and print the top 7 lines and the bottom 4 lines using `knitr::kable`.

```
penguins_ordered <- dplyr::arrange(penguins, bill_length_mm)
knitr::kable(rbind(head(penguins_ordered,n=7),tail(penguins_ordered,n=4)))
```

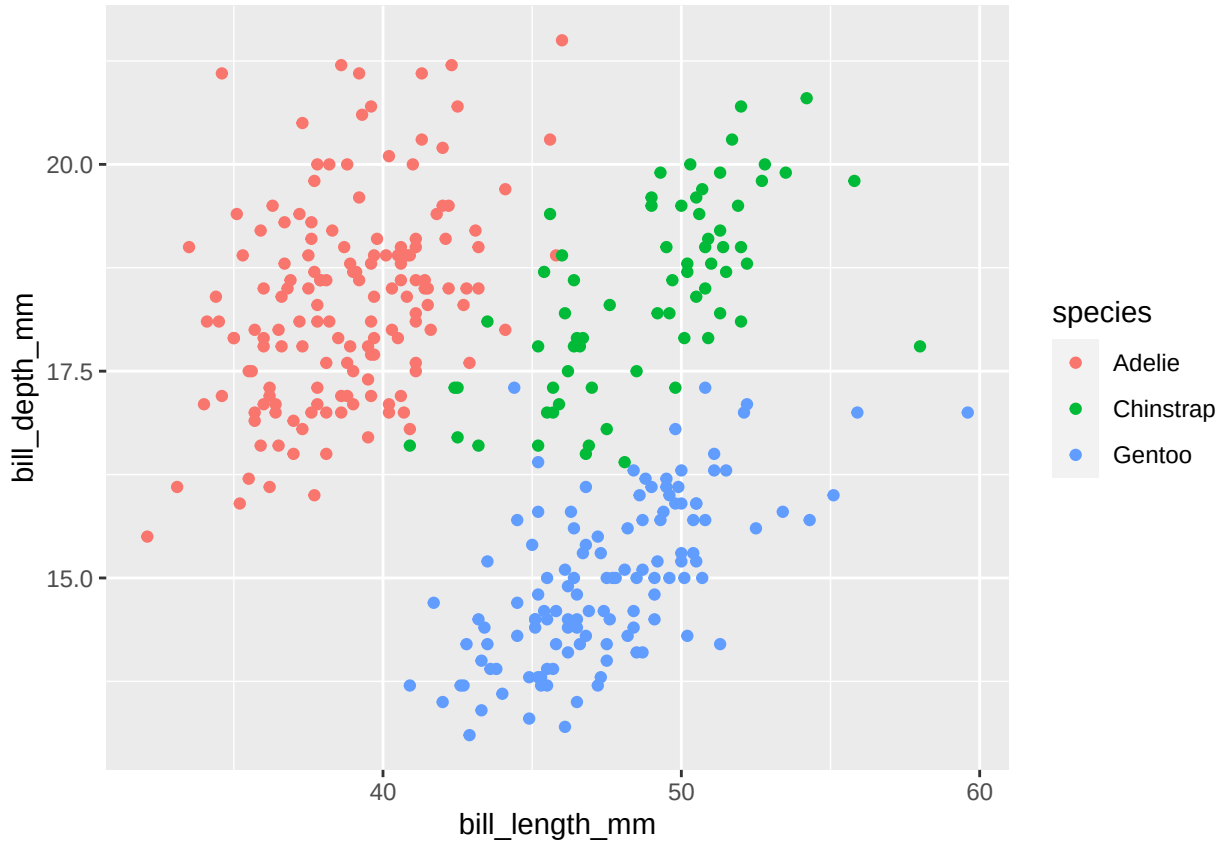
species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Dream	32.1	15.5	188	3050	female	2009
Adelie	Dream	33.1	16.1	178	2900	female	2008
Adelie	Torgersen	33.5	19.0	190	3600	female	2008
Adelie	Dream	34.0	17.1	185	3400	female	2008
Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
Adelie	Torgersen	34.4	18.4	184	3325	female	2007
Adelie	Biscoe	34.5	18.1	187	2900	female	2008
Chinstrap	Dream	58.0	17.8	181	3700	female	2007
Gentoo	Biscoe	59.6	17.0	230	6050	male	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Gentoo	Biscoe	NA	NA	NA	NA	NA	2009

4.3.3

Do a scatter plot (using the package `ggplot2`) of `bill_length_mm` (x-axis) vs. `bill_depth_mm` (y-axis).

```
ggplot(penguins) +
  geom_point(aes(bill_length_mm,bill_depth_mm,color=species))
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



4.3.4

Fit a linear model (use `lm`) with `bill_length_mm` as a predictor and `bill_depth_mm` as the response. Add the fitted linear relationship to the plot from the previous exercise and display the new plot. Is there an association between `bill_length_mm` and `bill_depth_mm`? What is the sign of the slope?

```
lmfit1 <- summary(lm(bill_depth_mm ~ bill_length_mm, penguins))
lmfit1
```

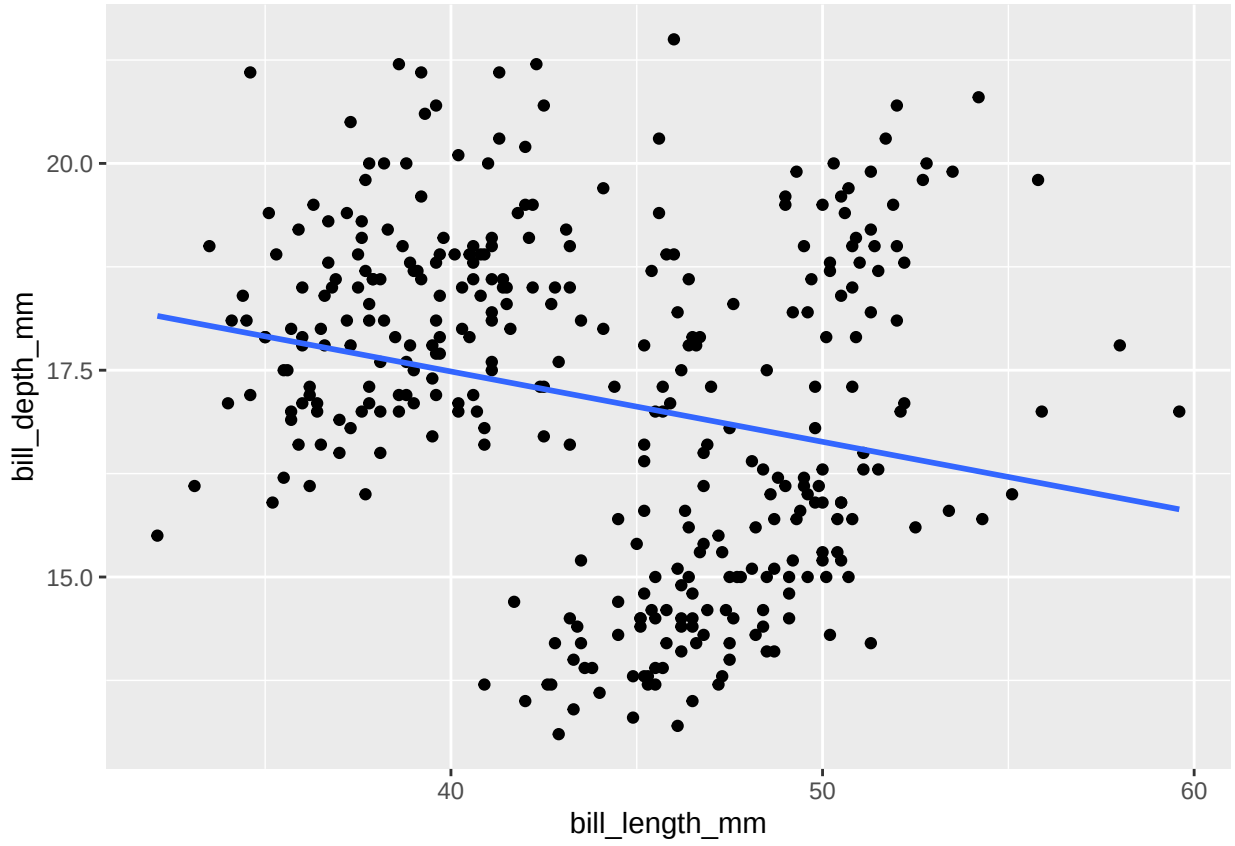
```
##
## Call:
## lm(formula = bill_depth_mm ~ bill_length_mm, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1381 -1.4263  0.0164  1.3841  4.5255
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    20.88547    0.84388  24.749  < 2e-16 ***
## bill_length_mm -0.08502    0.01907  -4.459 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.922 on 340 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.05525,    Adjusted R-squared:  0.05247
## F-statistic: 19.88 on 1 and 340 DF,  p-value: 1.12e-05
```

```
ggplot(penguins) +
  geom_point(aes(bill_length_mm,bill_depth_mm)) +
  geom_smooth(aes(bill_length_mm,bill_depth_mm),method = "lm", se = FALSE)
```

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

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).
```



4.3.5

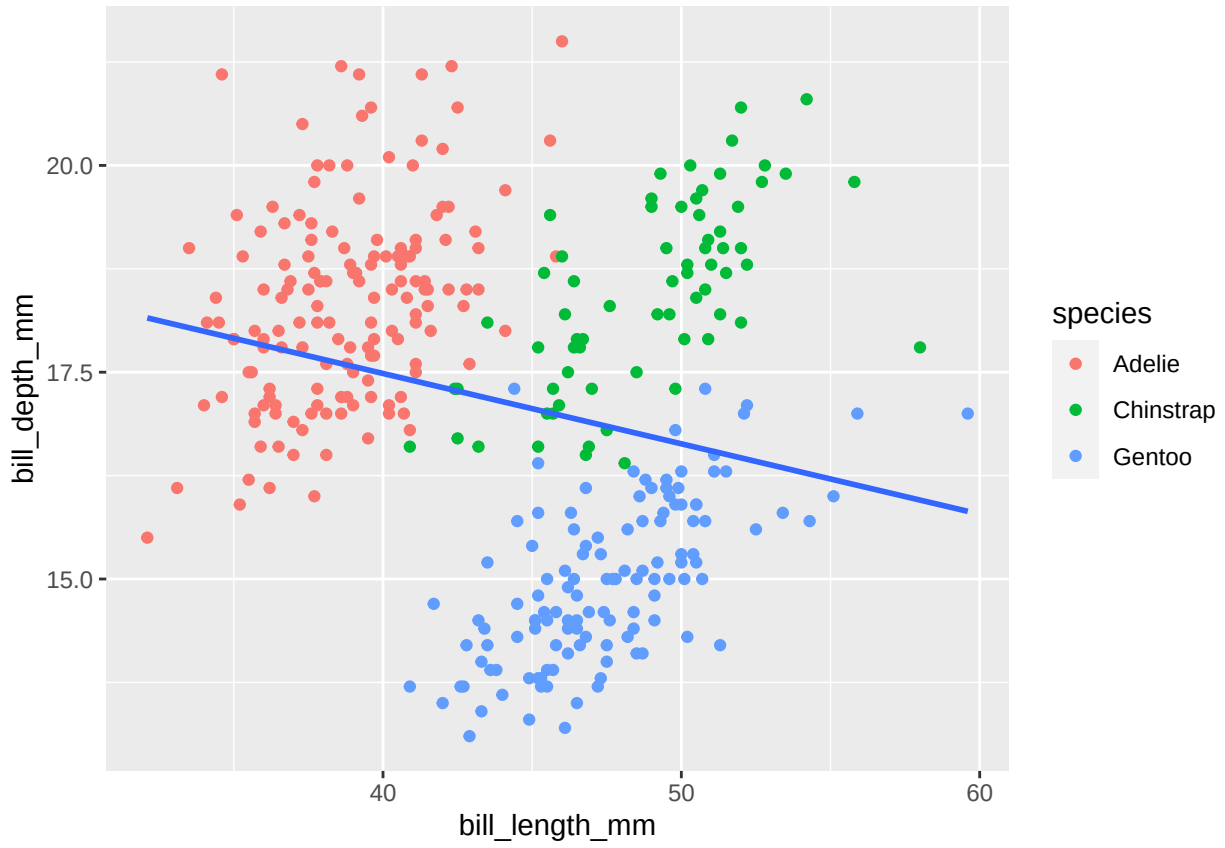
Color the points from the plot from the previous plot according to species. Describe your observations.

```
ggplot(penguins) +
  geom_point(aes(bill_length_mm,bill_depth_mm,color=species)) +
  geom_smooth(aes(bill_length_mm,bill_depth_mm),method = "lm", se = FALSE)
```

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

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



4.3.6

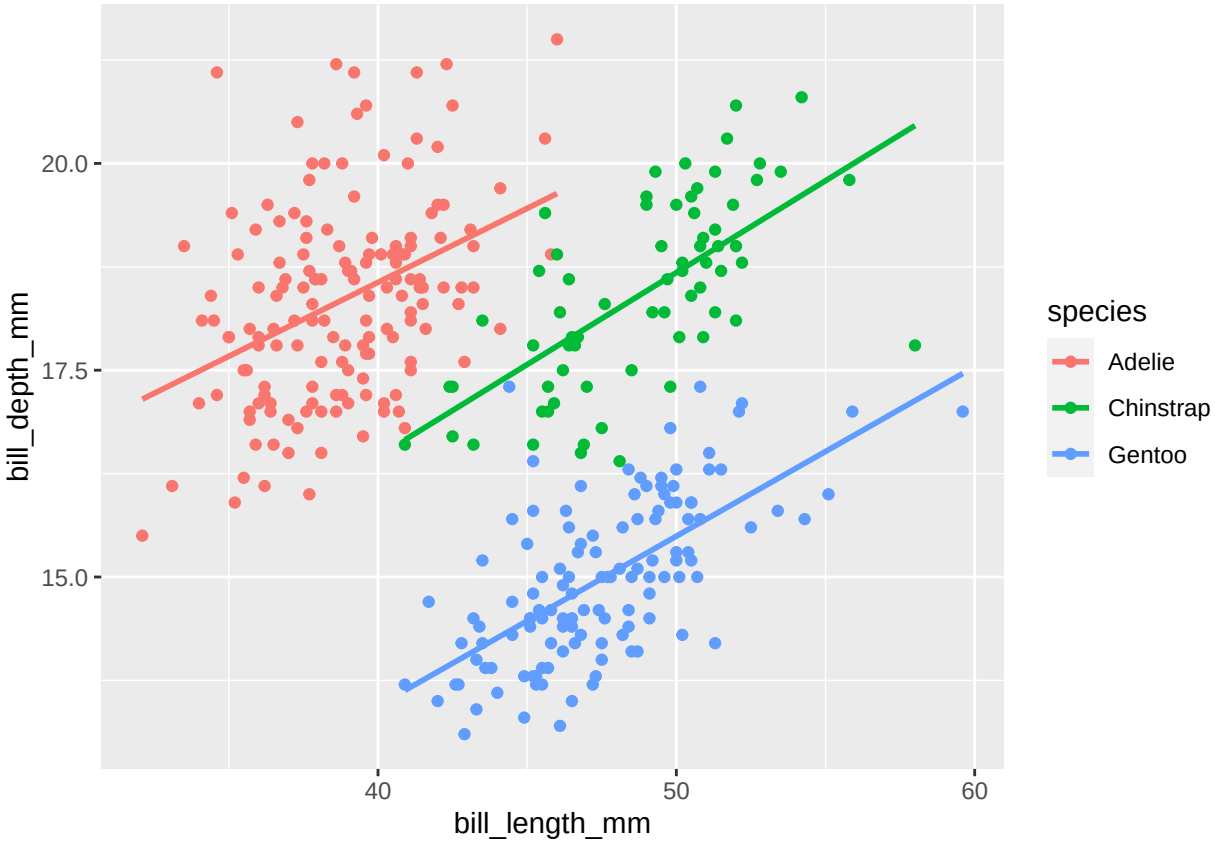
Add `species` as a predictor to your model and replot the previous plot with the new regression lines (there should be three, one for each species). Is there an association between `bill_length_mm` and `bill_depth_mm`? What is the sign of the slope?

```
lmfit2 <- summary(lm(bill_depth_mm ~  bill_length_mm + species, penguins))
ggplot(penguins) +
  geom_point(aes(bill_length_mm,bill_depth_mm,color=species)) +
  geom_smooth(aes(bill_length_mm,bill_depth_mm,color=species),method = "lm", se = FALSE)
```

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

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```


20.0
17.5
15.0
bill_depth_mm
40
50
60
bill_length_mm
species
Adelie
Chinstrap
Gentoo