MT3 Computer Assignment - Reproducible_Figures_R

2023-12-04

##Q1) Create a figure using the Palmer penguin dataset that is correct but badly communicates the data. Do not make a box plot. ##A) Figure inserted below:

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

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

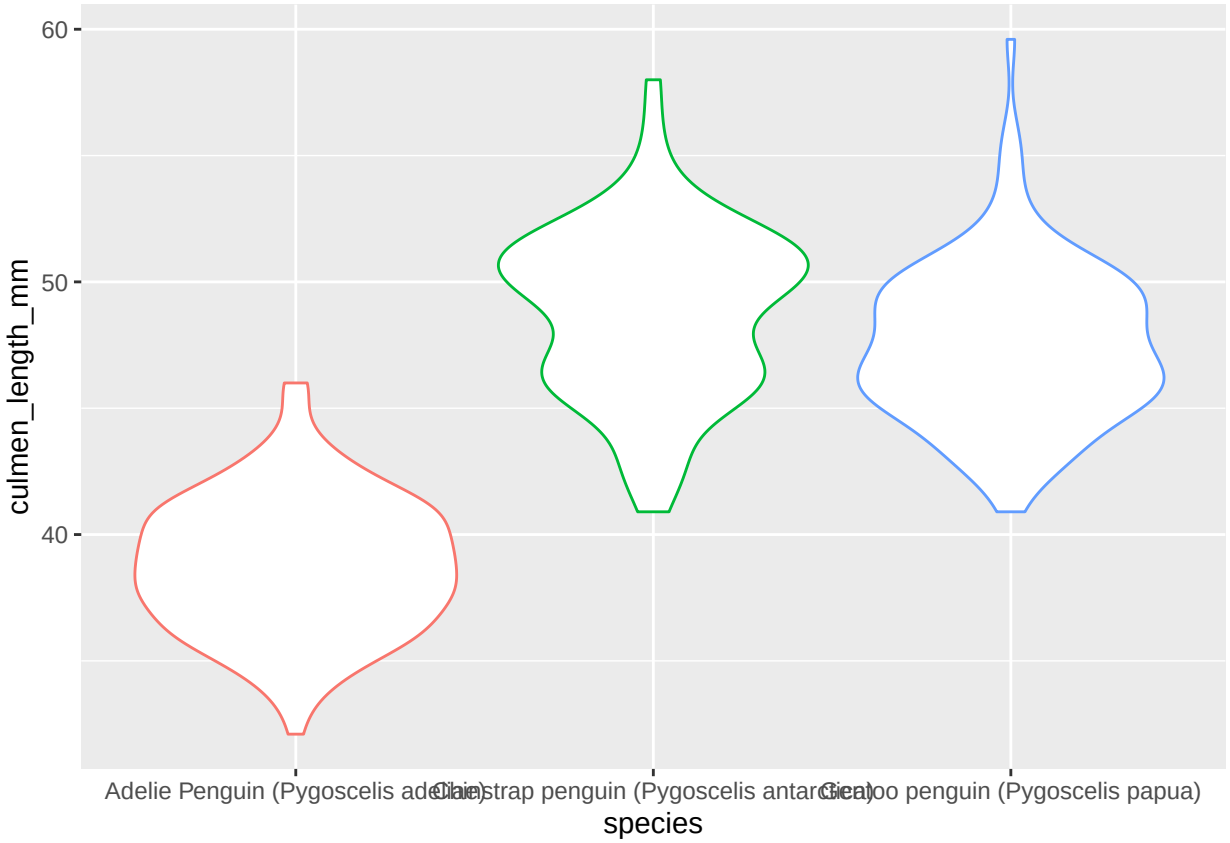
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

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



##B) Write about how your design choices mislead the reader about the underlying data. ##While technically this figure is not incorrect, this violin plot has some issues representing the data. The violin plot does not show individual data points and their distributions so it is difficult to observe the true nature of the data. It is therefore easy to misinterpret the data. Due to the natural complexity of the violin plot it can be inaccesible to those who are not familiar with this plot.

##Q2)Write a data analysis pipeline in your rmd. You should be aiming to write a clear explanation of the steps as well as clear code.

##Firstly you need to install the following packages:

```
install.packages(c("ggplot2", "palmerpenguins", "janitor", "dplyr"))
```

##You then need to load the packages now that they are installed:

```
library(ggplot2)
library(palmerpenguins)
library(janitor)
library(dplyr)
```

##Next you need to load in the data set

```
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
```

```
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

##Cleaning data by firstly looking at the column names

```
names(penguins)
```

```
## [1] "species"           "island"            "bill_length_mm"
## [4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
## [7] "sex"               "year"
```

##Looking at the summary of data to see what is happening

```
summary(penguins)
```

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

##Realise that there are some N/A results and decide to remove them

```
penguins <- na.omit(penguins)
```

##Shortening of island names:

```
Clean_data <- penguins %>%
  mutate(island = substr(island,1,3))
```

##Also shortening species names which can be seen in this pipeline below ##Pipeline for this cleaning:

```r
Clean_data <- penguins %>%
  na.omit() %>%
  mutate(island = substr(island,1,3)) %>%
  mutate(species = substr(species,1,3))
```

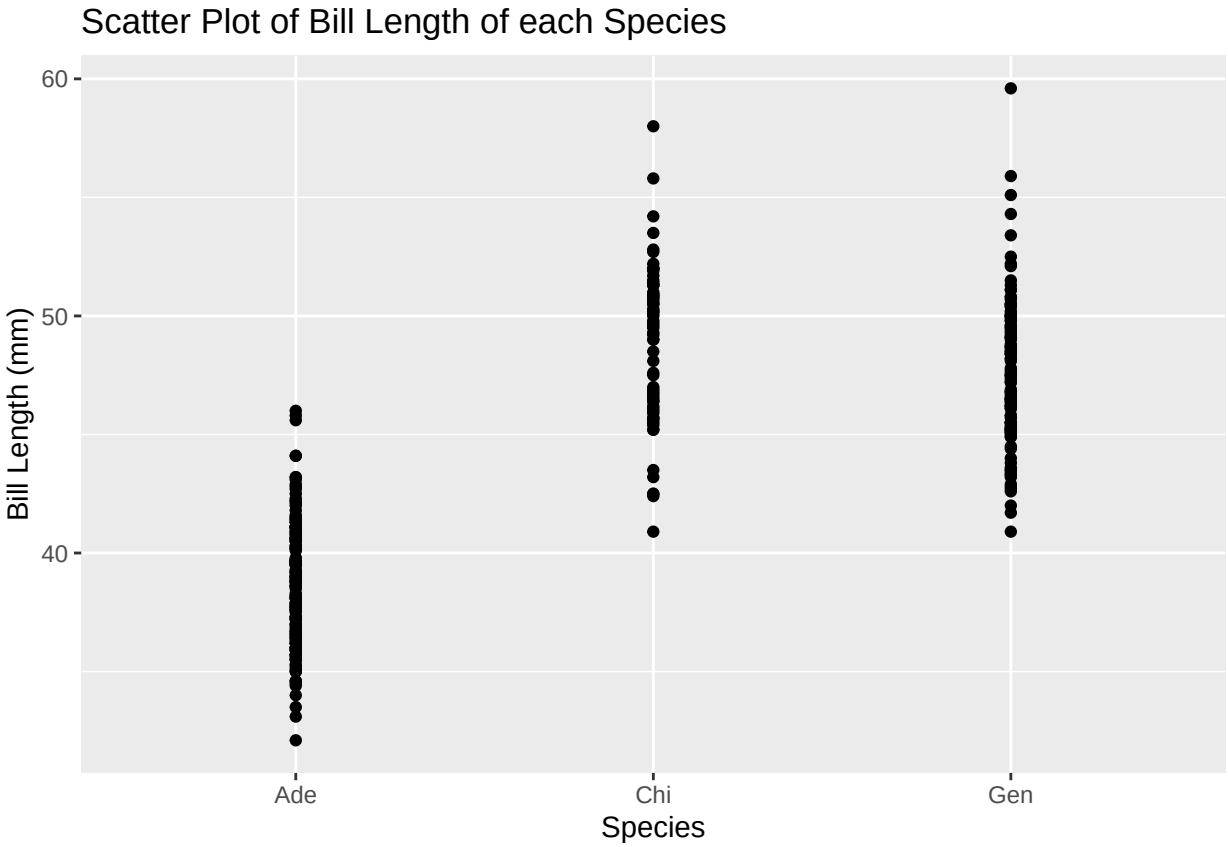
###Observing the cleaning made to the data

```r
head(Clean_data)
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <chr>   <chr>           <dbl>         <dbl>             <int>       <int>
## 1 Ade     Tor              39.1          18.7               181        3750
## 2 Ade     Tor              39.5          17.4               186        3800
## 3 Ade     Tor              40.3          18                 195        3250
## 4 Ade     Tor              36.7          19.3               193        3450
## 5 Ade     Tor              39.3          20.6               190        3650
## 6 Ade     Tor              38.9          17.8               181        3625
## # i 2 more variables: sex <fct>, year <int>
```

##Explanatory figure

```r
ggplot(Clean_data, aes(x = species, bill_length_mm)) +
  geom_point() +
  labs(title = "Scatter Plot of Bill Length of each Species",
       x = "Species",
       y = "Bill Length (mm)")
```



Null hypothesis: there is no significant difference in mean bill length between the three different species of penguin.

##Alternative hypothesis: there is signficant difference in mean bill length between atleast two of the three species.

##Run a statistical test. I decided to do an ANOVA test looking at the differences between mean bill length across the different species.

```
model <- lm(bill_length_mm ~ species, data = Clean_data)
anova_result <- anova(model)
```

##It is time to analyse the results of the anova...

```
print(anova_result)
```

```
## Analysis of Variance Table
##
## Response: bill_length_mm
##            Df Sum Sq Mean Sq F value    Pr(>F)
## species     2 7015.4  3507.7   397.3 < 2.2e-16 ***
## Residuals 330 2913.5     8.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

##This anova has produced an incredibly small p value it almost pracitcally 0. This tells us that there is strong evidence that the results are highly statistically significant. The null hypothesis is infact wrong and there are differences in mean bill length among species.

I decided to conduct a post-hoc test using the Tukey method as the results of my anova suggest there are significant differences and I want to see which species differ from eachother in particular.

```
library(agricolae)

# Conduct Tukey's HSD post-hoc test
tukey_result <- HSD.test(model, "species")

# Print the results
print(tukey_result)
```

```
## $statistics
##    MSerror  Df     Mean       CV
##   8.828839 330 43.99279 6.754143
##
## $parameters
##    test  name.t ntr StudentizedRange alpha
##   Tukey species   3         3.329537  0.05
##
## $means
##     bill_length_mm      std   r        se  Min  Max    Q25   Q50    Q75
## Ade       38.82397 2.662597 146 0.2459095 32.1 46.0 36.725 38.85 40.775
## Chi       48.83382 3.339256  68 0.3603275 40.9 58.0 46.350 49.55 51.075
## Gen       47.56807 3.106116 119 0.2723819 40.9 59.6 45.350 47.40 49.600
##
## $comparison
## NULL
##
## $groups
##     bill_length_mm groups
## Chi       48.83382      a
## Gen       47.56807      b
## Ade       38.82397      c
##
## attr(,"class")
## [1] "group"
```

##The results of this test suggest Chinstrap has a significantly different mean bill length compared to Gentoo and Adelie. Gentoo and Adelie do not show a significant difference in mean bill length.

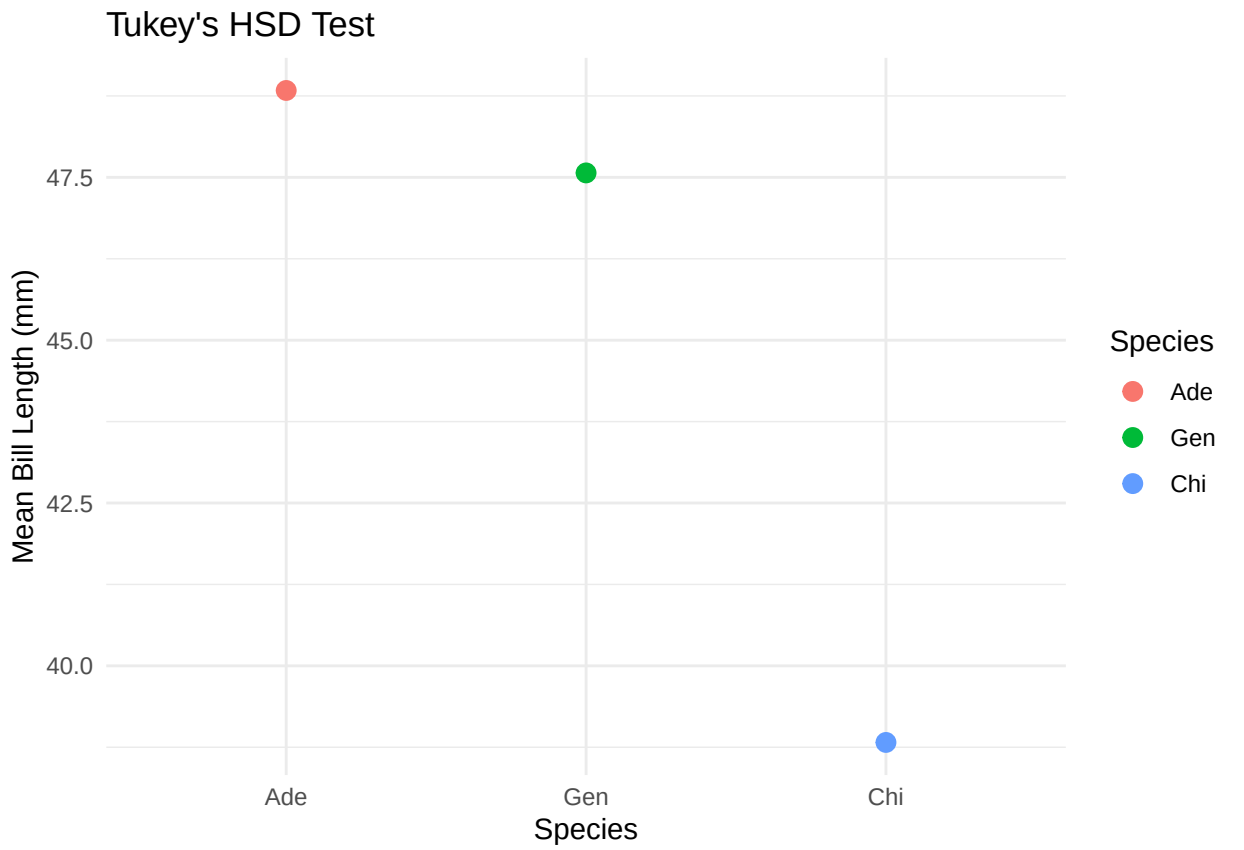
##In order to visualise the results of these statistical tests I chose to create a dot-plot:

```
library(ggplot2)

tukey_df <- as.data.frame(tukey_result$groups)
```

```r
tukey_df$groups <- factor(tukey_df$groups, levels = c("a", "b", "c"), labels = c("Ade", "Gen", "Chi"))

ggplot(tukey_df, aes(x = groups, y = bill_length_mm, color = groups)) +
  geom_point(position = position_dodge(width = 0.8), size = 3) +
  labs(title = "Tukey's HSD Test",
       x = "Species",
       y = "Mean Bill Length (mm)",
       color = "Species") +
  theme_minimal()
```



The results of my statistical tests as shown in the figure above clearly present that there is a significant difference in mean bill length between at least two of the three species. We can therefore reject our null hypothesis that states there is no significant differences in mean bill length between the three species.

##Q3) Open Science:

##A) My github link: https://github.com/BioBabe2002/Reproducible-Figures-R

##B) Partner's github link: https://github.com/Elephant34/ReproducibleScience

##C) Reflect on your experience running their code. What elements of their code helped you to understand their data pipeline? Did it run and did you need to fix anything? What suggestions would you make for improving their code to make it more understandable or reproducible and why? If you needed to alter your partner's figure using their code, do you think that would be easy or difficult and why?

My partner's code was successful in allowing me to understand their data pipeline for a number of reasons. Firstly, their code was continously annotated. This let me know exactly what function a line of code had so that I could follow their process.. Secondly, they handily make note of where figures can be found in their repo which was helpful to compare to when running the code myself. Their code ran very smoothly without need for alteration. I would argue that the removal of N/A values could be included in their "DataCleaning" block but this is a minimal comment. It would also be useful to contain session information indicating the versions of software used. It would also be easy to alter my partner's figure using their code as the steps appear to be very transparent and possible for me to integrate my own desired code.

##D) Reflect on your own code based on your experience with your partner's code and their review of yours.

##Their review was very plausible. I avoided added in code to install given packages but in hindsight I could have made note that these packages should be installed and even produced code to only run install.packages if the package is not already installed which would look hopefully something like this:

```
##firstly checking that the Agricolae package is installed:
if (!requireNamespace("Agricolae", quietly = TRUE)) {
  # then if it is not installed, it will install
  install.packages("Agricolae", dependencies = TRUE)
}
```

##My partner also comments that by splitting the pipeline into seperate functions this would improve readibility and I agree. By doing so it would make my code easier to follow along and reproduce. ##Writing code for other people has helped me to generate my clarity in my work. When coding for myself it is easy to make brief notes that only I would understand or lack thereof hoping I will know what it means when I return to my work. By outlining descriptions of my steps taken it ensures that my work is useful. ## Taking my partner's advice I have uploaded my functions to improve reproducibility.