Data Cleaning & Explorations

Abdul-lateef Abiola Hussein

2022-09-26

# Palmer Penguins Cleaning and Exploration

This projects was done to show case my skill in using R programming language for Data cleaning process, exploration and documentation.

## Setting up suitable working environment

Procedures necessary include Installation of palmerpenguins packages which contains the raw data needed for this analysis. Loading the packages needed for data cleaning, data manipulation, visualization such as; ‘tidyverse’, ‘tidyr’,‘dplyr’,‘ggplot’,‘janitor’, and ‘skimr’.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(dplyr)  
library(tidyr)  
library(janitor)

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

library(stringr)  
library(skimr)  
library(palmerpenguins)

## Prepare and processing Phase.

These phases involves inspecting the ‘penguins\_raw’ data, extraction of the needed variables , data cleaning and data validations.

View(penguins\_raw)  
  
colnames(penguins\_raw)

## [1] "studyName" "Sample Number" "Species"   
## [4] "Region" "Island" "Stage"   
## [7] "Individual ID" "Clutch Completion" "Date Egg"   
## [10] "Culmen Length (mm)" "Culmen Depth (mm)" "Flipper Length (mm)"  
## [13] "Body Mass (g)" "Sex" "Delta 15 N (o/oo)"   
## [16] "Delta 13 C (o/oo)" "Comments"

skim\_without\_charts(penguins\_raw)

Data summary

|  |  |
| --- | --- |
| Name | penguins\_raw |
| Number of rows | 344 |
| Number of columns | 17 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 9 |
| Date | 1 |
| numeric | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| studyName | 0 | 1.00 | 7 | 7 | 0 | 3 | 0 |
| Species | 0 | 1.00 | 33 | 41 | 0 | 3 | 0 |
| Region | 0 | 1.00 | 6 | 6 | 0 | 1 | 0 |
| Island | 0 | 1.00 | 5 | 9 | 0 | 3 | 0 |
| Stage | 0 | 1.00 | 18 | 18 | 0 | 1 | 0 |
| Individual ID | 0 | 1.00 | 4 | 6 | 0 | 190 | 0 |
| Clutch Completion | 0 | 1.00 | 2 | 3 | 0 | 2 | 0 |
| Sex | 11 | 0.97 | 4 | 6 | 0 | 2 | 0 |
| Comments | 290 | 0.16 | 18 | 68 | 0 | 10 | 0 |

**Variable type: Date**

| skim\_variable | n\_missing | complete\_rate | min | max | median | n\_unique |
| --- | --- | --- | --- | --- | --- | --- |
| Date Egg | 0 | 1 | 2007-11-09 | 2009-12-01 | 2008-11-09 | 50 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample Number | 0 | 1.00 | 63.15 | 40.43 | 1.00 | 29.00 | 58.00 | 95.25 | 152.00 |
| Culmen Length (mm) | 2 | 0.99 | 43.92 | 5.46 | 32.10 | 39.23 | 44.45 | 48.50 | 59.60 |
| Culmen Depth (mm) | 2 | 0.99 | 17.15 | 1.97 | 13.10 | 15.60 | 17.30 | 18.70 | 21.50 |
| Flipper Length (mm) | 2 | 0.99 | 200.92 | 14.06 | 172.00 | 190.00 | 197.00 | 213.00 | 231.00 |
| Body Mass (g) | 2 | 0.99 | 4201.75 | 801.95 | 2700.00 | 3550.00 | 4050.00 | 4750.00 | 6300.00 |
| Delta 15 N (o/oo) | 14 | 0.96 | 8.73 | 0.55 | 7.63 | 8.30 | 8.65 | 9.17 | 10.03 |
| Delta 13 C (o/oo) | 13 | 0.96 | -25.69 | 0.79 | -27.02 | -26.32 | -25.83 | -25.06 | -23.79 |

peng <- penguins\_raw %>%  
 clean\_names()  
  
colnames(peng)

## [1] "study\_name" "sample\_number" "species"   
## [4] "region" "island" "stage"   
## [7] "individual\_id" "clutch\_completion" "date\_egg"   
## [10] "culmen\_length\_mm" "culmen\_depth\_mm" "flipper\_length\_mm"  
## [13] "body\_mass\_g" "sex" "delta\_15\_n\_o\_oo"   
## [16] "delta\_13\_c\_o\_oo" "comments"

peng\_data<-peng %>%   
 select(c(species,region,island,culmen\_length\_mm,  
 culmen\_depth\_mm,flipper\_length\_mm,body\_mass\_g,sex))  
peng\_data

## # A tibble: 344 × 8  
## species region island culme…¹ culme…² flipp…³ body\_…⁴ sex   
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 Adelie Penguin (Pygoscel… Anvers Torge… 39.1 18.7 181 3750 MALE   
## 2 Adelie Penguin (Pygoscel… Anvers Torge… 39.5 17.4 186 3800 FEMA…  
## 3 Adelie Penguin (Pygoscel… Anvers Torge… 40.3 18 195 3250 FEMA…  
## 4 Adelie Penguin (Pygoscel… Anvers Torge… NA NA NA NA <NA>   
## 5 Adelie Penguin (Pygoscel… Anvers Torge… 36.7 19.3 193 3450 FEMA…  
## 6 Adelie Penguin (Pygoscel… Anvers Torge… 39.3 20.6 190 3650 MALE   
## 7 Adelie Penguin (Pygoscel… Anvers Torge… 38.9 17.8 181 3625 FEMA…  
## 8 Adelie Penguin (Pygoscel… Anvers Torge… 39.2 19.6 195 4675 MALE   
## 9 Adelie Penguin (Pygoscel… Anvers Torge… 34.1 18.1 193 3475 <NA>   
## 10 Adelie Penguin (Pygoscel… Anvers Torge… 42 20.2 190 4250 <NA>   
## # … with 334 more rows, and abbreviated variable names ¹​culmen\_length\_mm,  
## # ²​culmen\_depth\_mm, ³​flipper\_length\_mm, ⁴​body\_mass\_g

peng\_cleaned<-peng\_data %>%   
 drop\_na(sex, culmen\_length\_mm,culmen\_depth\_mm,flipper\_length\_mm,body\_mass\_g)  
  
sep<- peng\_cleaned %>%   
 separate(species, into = c("a","b"))

## Warning: Expected 2 pieces. Additional pieces discarded in 333 rows [1, 2, 3, 4,  
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].

peng\_cleaned\_data <- sep %>%   
 unite("species", c(a,b), sep = " " )  
  
skim\_without\_charts(peng\_cleaned\_data)

Data summary

|  |  |
| --- | --- |
| Name | peng\_cleaned\_data |
| Number of rows | 333 |
| Number of columns | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 4 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| species | 0 | 1 | 14 | 17 | 0 | 3 | 0 |
| region | 0 | 1 | 6 | 6 | 0 | 1 | 0 |
| island | 0 | 1 | 5 | 9 | 0 | 3 | 0 |
| sex | 0 | 1 | 4 | 6 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| culmen\_length\_mm | 0 | 1 | 43.99 | 5.47 | 32.1 | 39.5 | 44.5 | 48.6 | 59.6 |
| culmen\_depth\_mm | 0 | 1 | 17.16 | 1.97 | 13.1 | 15.6 | 17.3 | 18.7 | 21.5 |
| flipper\_length\_mm | 0 | 1 | 200.97 | 14.02 | 172.0 | 190.0 | 197.0 | 213.0 | 231.0 |
| body\_mass\_g | 0 | 1 | 4207.06 | 805.22 | 2700.0 | 3550.0 | 4050.0 | 4775.0 | 6300.0 |

peng\_cleaned\_data <-peng\_cleaned\_data %>%   
 mutate(body\_mass\_kg = body\_mass\_g/1000, sex= tolower(sex)) %>%   
 select(-body\_mass\_g)

## Analysis Phase

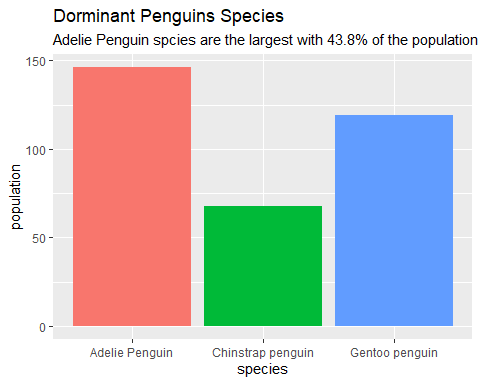
This phase involves manipulating the cleaned penguins datasets in order to extract answers from the heart of the data.

###1.Which specie of penguins is dominant.

peng\_cleaned\_data %>%   
 group\_by(species) %>%   
 count(species) %>%   
 mutate(percent\_species= n\*100/333)

## # A tibble: 3 × 3  
## # Groups: species [3]  
## species n percent\_species  
## <chr> <int> <dbl>  
## 1 Adelie Penguin 146 43.8  
## 2 Chinstrap penguin 68 20.4  
## 3 Gentoo penguin 119 35.7

peng\_cleaned\_data %>%   
 ggplot(aes(x=species, fill= species)) + geom\_bar(show.legend= FALSE)+  
 labs(title = "Dorminant Penguins Species", y= "population",  
 subtitle = "Adelie Penguin spcies are the largest with 43.8% of the population")

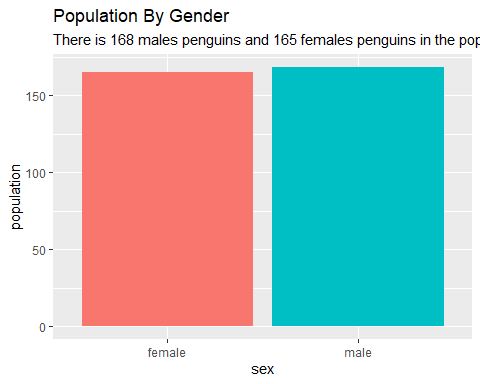


### #2. Which gender dominates the population

peng\_cleaned\_data %>%   
 group\_by(sex) %>%   
 count(sex) %>%   
 mutate(percent\_sex = n\*100/333)

## # A tibble: 2 × 3  
## # Groups: sex [2]  
## sex n percent\_sex  
## <chr> <int> <dbl>  
## 1 female 165 49.5  
## 2 male 168 50.5

peng\_cleaned\_data %>%   
 ggplot() + geom\_bar(mapping = aes(x= sex, fill= sex), show.legend = FALSE) +   
 labs(title ="Population By Gender",y= "population",   
 subtitle = "There is 168 males penguins and 165 females penguins in the population.")



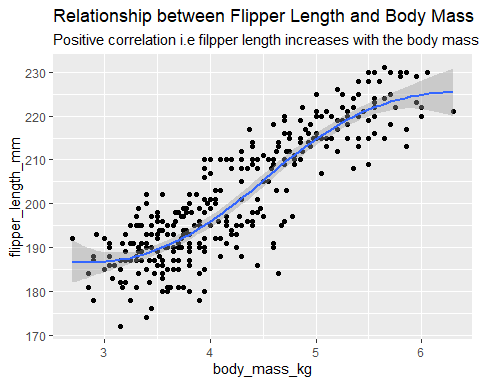
### 3.What’s the correlation between flipper\_length and body\_mass\_kg

peng\_cleaned\_data %>%   
 select(flipper\_length\_mm,body\_mass\_kg)

## # A tibble: 333 × 2  
## flipper\_length\_mm body\_mass\_kg  
## <dbl> <dbl>  
## 1 181 3.75  
## 2 186 3.8   
## 3 195 3.25  
## 4 193 3.45  
## 5 190 3.65  
## 6 181 3.62  
## 7 195 4.68  
## 8 182 3.2   
## 9 191 3.8   
## 10 198 4.4   
## # … with 323 more rows

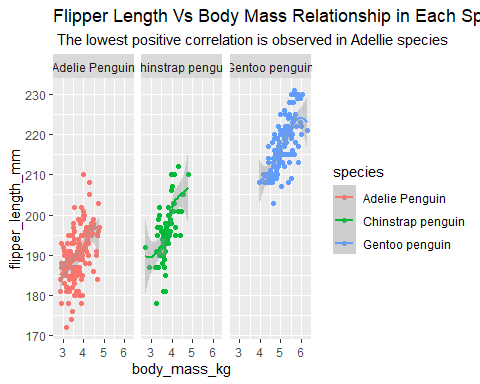
peng\_cleaned\_data %>%   
 ggplot(aes(x= body\_mass\_kg,y =flipper\_length\_mm)) + geom\_point()+  
 geom\_smooth(method = "loess") +  
 labs(title = "Relationship between Flipper Length and Body Mass", subtitle =  
 "Positive correlation i.e filpper length increases with the body mass ")

## `geom\_smooth()` using formula 'y ~ x'



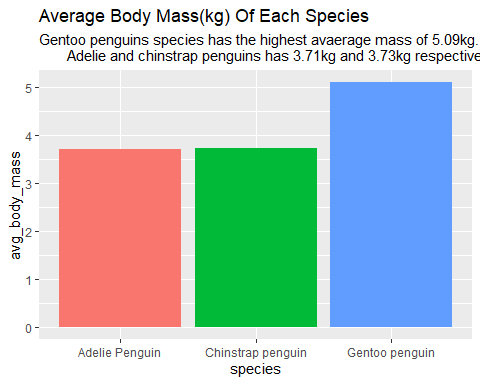
peng\_cleaned\_data %>%   
 ggplot(aes(x= body\_mass\_kg,y =flipper\_length\_mm, color= species)) + geom\_point()+  
 geom\_smooth(method = "loess") + facet\_wrap(~species) +  
 labs(title = "Flipper Length Vs Body Mass Relationship in Each Species",  
 subtitle= " The lowest positive correlation is observed in Adellie species")

## `geom\_smooth()` using formula 'y ~ x'



### 4.Which species has the highest average body\_mass\_kg

peng\_cleaned\_data %>%   
 group\_by(species) %>%   
 summarise(avg\_body\_mass = mean(body\_mass\_kg)) %>%   
 ggplot(aes(x= species, y= avg\_body\_mass, fill= species)) + geom\_col(show.legend = FALSE) +  
 labs(title = "Average Body Mass(kg) Of Each Species",  
 subtitle = "Gentoo penguins species has the highest avaerage mass of 5.09kg.  
 Adelie and chinstrap penguins has 3.71kg and 3.73kg respectively.")

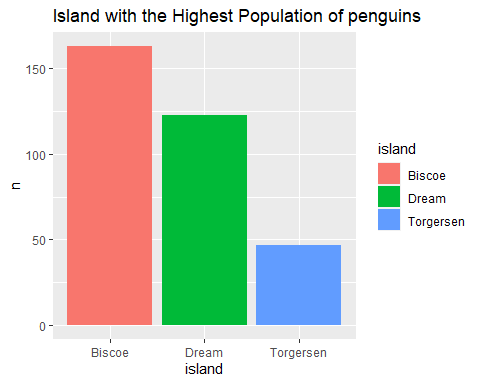


### 5.Which island has the highest population of penguins and why?

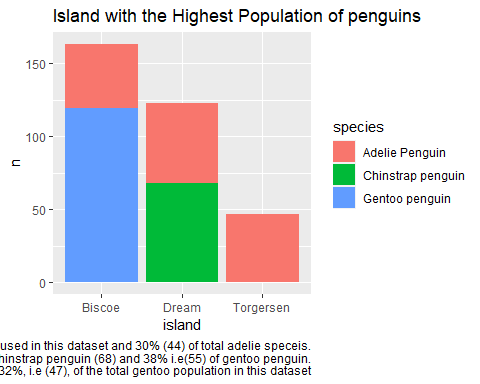
peng\_cleaned\_data %>%   
 group\_by(island) %>%   
 count(island) %>%   
 arrange(-n, .by\_group = TRUE)

## # A tibble: 3 × 2  
## # Groups: island [3]  
## island n  
## <chr> <int>  
## 1 Biscoe 163  
## 2 Dream 123  
## 3 Torgersen 47

peng\_cleaned\_data %>%   
 group\_by(island) %>%   
 count(island) %>%   
 arrange(-n, .by\_group = TRUE) %>%   
 ggplot(aes(x= island, y= n, fill = island)) + geom\_col() +   
 labs(title = "Island with the Highest Population of penguins")



peng\_cleaned\_data %>%   
 group\_by(island,species) %>%   
 count(species) %>%   
 arrange(-n, .by\_group = TRUE) %>%   
 ggplot(aes(x= island, y= n, fill = species)) + geom\_col() +   
 labs(title = "Island with the Highest Population of penguins",  
 caption ="Biscoe island contains all the gentoo penguins (119) used in this dataset and 30% (44) of total adelie speceis.  
 Dream island contains all Chinstrap penguin (68) and 38% i.e(55) of gentoo penguin.  
 Torgersen island contains only gentoo penguins which is 32%, i.e (47), of the total gentoo population in this dataset")



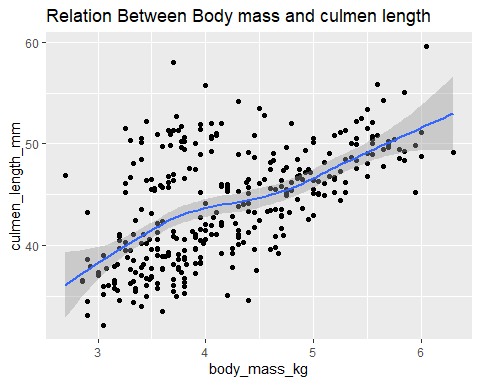
### 6. What is the correlation between the body\_mass\_kg and culmen\_length\_mm in general

peng\_cleaned\_data %>%   
 select(culmen\_length\_mm,body\_mass\_kg)

## # A tibble: 333 × 2  
## culmen\_length\_mm body\_mass\_kg  
## <dbl> <dbl>  
## 1 39.1 3.75  
## 2 39.5 3.8   
## 3 40.3 3.25  
## 4 36.7 3.45  
## 5 39.3 3.65  
## 6 38.9 3.62  
## 7 39.2 4.68  
## 8 41.1 3.2   
## 9 38.6 3.8   
## 10 34.6 4.4   
## # … with 323 more rows

peng\_cleaned\_data %>%   
 ggplot(aes(x=body\_mass\_kg,y=culmen\_length\_mm)) + geom\_point() +  
 geom\_smooth(method = "loess") +  
 labs(title = "Relation Between Body mass and culmen length")

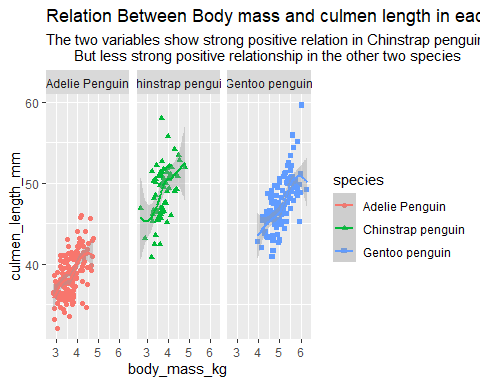
## `geom\_smooth()` using formula 'y ~ x'



### 6b. What is the correlation between the body\_mass\_kg and culmen\_length\_mm (faceting by species)

peng\_cleaned\_data %>%   
 ggplot(aes(x=body\_mass\_kg,y=culmen\_length\_mm, line = species, shape= species, color=species)) +   
 geom\_jitter() + geom\_smooth(method = "loess") +facet\_wrap(~species) +  
 labs(title= "Relation Between Body mass and culmen length in each species",  
 subtitle ="The two variables show strong positive relation in Chinstrap penguins.  
 But less strong positive relationship in the other two species")

## `geom\_smooth()` using formula 'y ~ x'



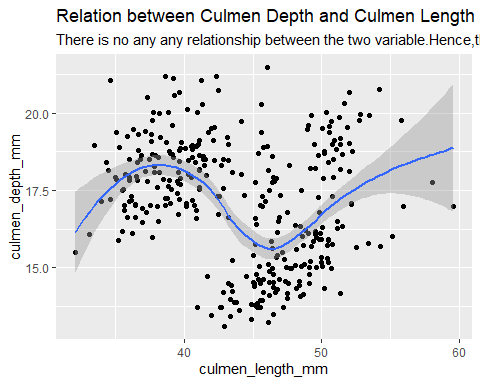
### 7. What is the correlation between the culmen\_depth\_mm and culmen\_length\_mm in general

peng\_cleaned\_data %>%   
 select(culmen\_length\_mm,culmen\_depth\_mm)

## # A tibble: 333 × 2  
## culmen\_length\_mm culmen\_depth\_mm  
## <dbl> <dbl>  
## 1 39.1 18.7  
## 2 39.5 17.4  
## 3 40.3 18   
## 4 36.7 19.3  
## 5 39.3 20.6  
## 6 38.9 17.8  
## 7 39.2 19.6  
## 8 41.1 17.6  
## 9 38.6 21.2  
## 10 34.6 21.1  
## # … with 323 more rows

peng\_cleaned\_data %>%   
 ggplot(aes(x=culmen\_length\_mm,y=culmen\_depth\_mm)) +   
 geom\_jitter() + geom\_smooth(method = "loess") +  
 labs(title = "Relation between Culmen Depth and Culmen Length",  
 subtitle = "There is no any any relationship between the two variable.Hence,the relationship is non-linear")

## `geom\_smooth()` using formula 'y ~ x'



#### Thanks :)