# Case Study – Palmer Penguins
# Omkar Mankame
# Date – 25 Aug 24

- R library has Palmer Penguins dataset which has three species of penguins with different parameters like flipper length, height, weight, etc.
- The data set has 344 datapoints.

---

**Data sets in package 'palmerpenguins':**

penguins Size measurements for adult foraging penguins near Palmer Station, Antarctica

penguins_raw (penguins) Penguin size, clutch, and blood isotope data for foraging adults near Palmer Station, Antarctica

---

- The aim of this project is to find the relation between flipper length and body mass. A guess would be larger the flipper length more the body mass.
- The same prediction was analyzed using R scattered plot to find the correlation.

**Steps -**

**1.** The Penguins Dataset in R Studio can be installed using install.packages('palmerpenguins') and then using it by library('palmerpenguins')

```
> install.packages('palmerpenguins')
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.3’
(as ‘lib’ is unspecified)
trying URL 'http://rspm/default/__linux__/focal/latest/src/contrib/palmerpenguins_0.1.1.tar.gz
'
Content type 'application/x-gzip' length 3001134 bytes (2.9 MB)
==================================================
downloaded 2.9 MB

* installing *binary* package ‘palmerpenguins’ ...
* DONE (palmerpenguins)

The downloaded source packages are in
        ‘/tmp/RtmpbS5C60/downloaded_packages’
```

**2.** To use the dataset use the code below.

```
> library(palmerpenguins)
> data(package = 'palmerpenguins')
```

**3.** Install additional packages for data analysis – tidyvers which contains ggplot2, dplyr, facets, etc.

```
> install.packages('tidyverse')
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.3’
(as ‘lib’ is unspecified)
trying URL 'http://rspm/default/__linux__/focal/latest/src/contrib/tidyverse_2.0.0.ta
r.gz'
Content type 'application/x-gzip' length 425176 bytes (415 KB)
==================================================
downloaded 415 KB

* installing *binary* package ‘tidyverse’ ...
* DONE (tidyverse)

The downloaded source packages are in
        ‘/tmp/RtmpbS5C60/downloaded_packages’


> library('tidyverse')
── Attaching core tidyverse packages ──────────────────────────────── tidyverse 2.0.
0 ──
✓ dplyr     1.1.2     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.5.1     ✓ tibble    3.2.1
✓ lubridate 1.9.2     ✓ tidyr     1.3.0
```

```
✓ purrr     1.0.1
── Conflicts ─────────────────────────────────────────── tidyverse_conflicts(
) ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package to force all conflicts to become errors
```

**4.** Know your data set – Head gives 6 rows and 8 columns, str shows the internal structure of the dataframe.

```
> head(penguins)
# A tibble: 6 × 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex    year
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int> <fct>  <int>
1 Adelie  Torgers…            39.1          18.7               181        3750 male    2007
2 Adelie  Torgers…            39.5          17.4               186        3800 fema…   2007
3 Adelie  Torgers…            40.3          18                 195        3250 fema…   2007
4 Adelie  Torgers…            NA            NA                  NA          NA NA      2007
5 Adelie  Torgers…            36.7          19.3               193        3450 fema…   2007
6 Adelie  Torgers…            39.3          20.6               190        3650 male    2007
```

```
> str(penguins)
tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1
...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 .
..
```

**5.** Installed ggplot2 package

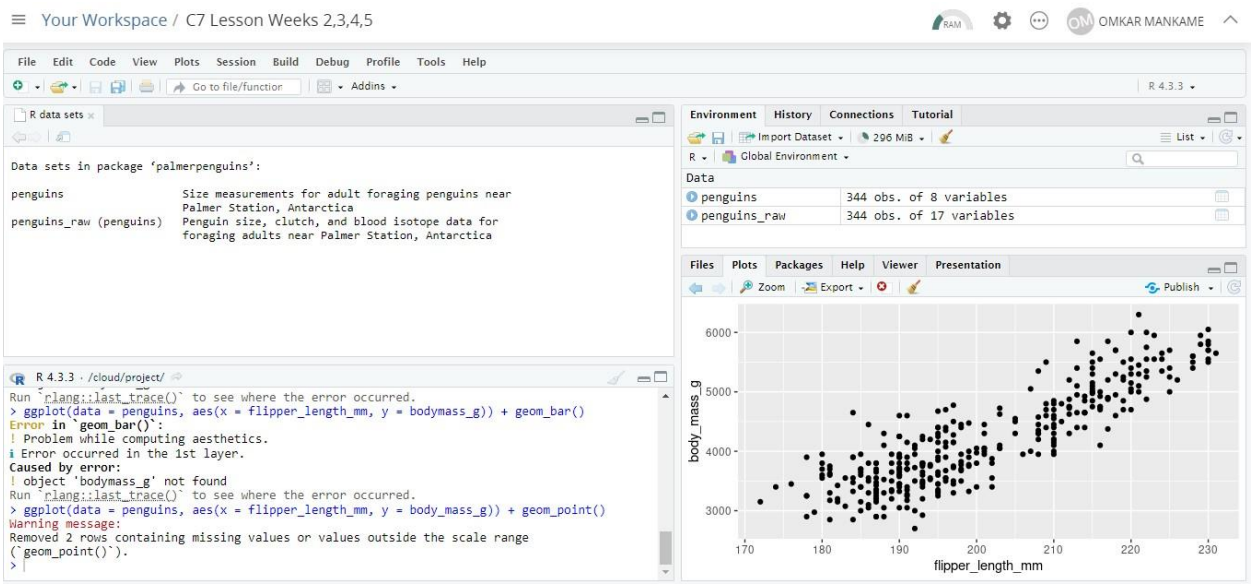
```
> install.packages("ggplot2")
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.3’
(as ‘lib’ is unspecified)
trying URL 'http://rspm/default/__linux__/focal/latest/src/contrib/ggplot2_3.5.1.tar.
gz'
Content type 'application/x-gzip' length 4942224 bytes (4.7 MB)
==================================================
downloaded 4.7 MB

* installing *binary* package ‘ggplot2’ ...
* DONE (ggplot2)
```

```
The downloaded source packages are in '/tmp/RtmpgXHGS2/downloaded_packages'
```

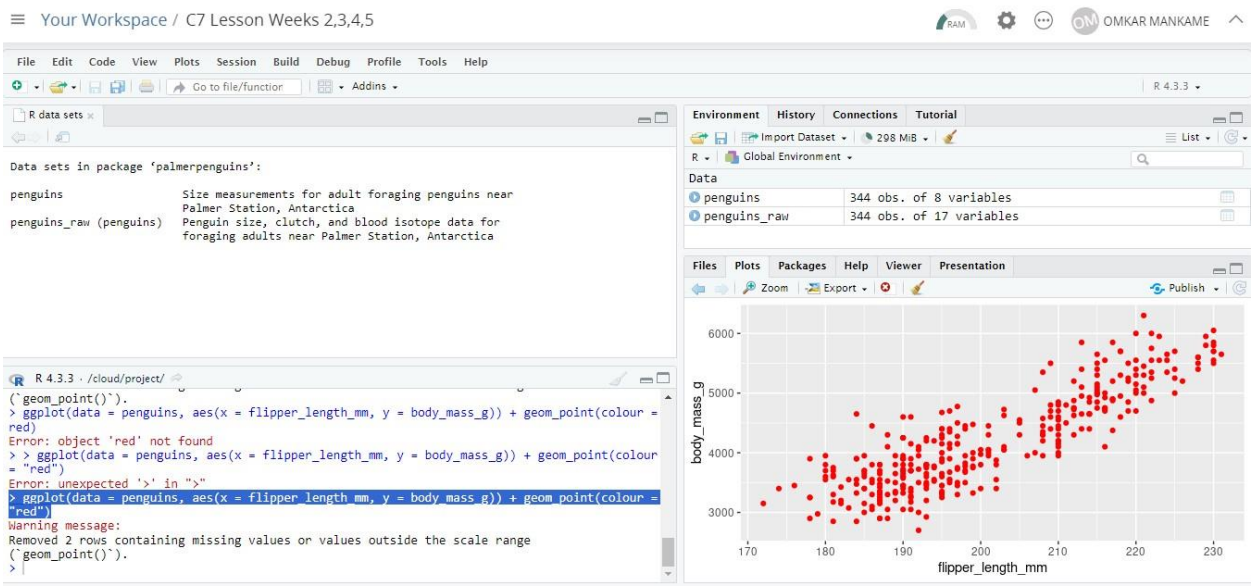
**6.** Created a scattered plot to show the relation between flipper length and body mass.

```
> ggplot(data = penguins, aes(x = flipper_length_mm, y = body_mass_g)) + geom
_point()
Warning message:
Removed 2 rows containing missing values or values outside the scale range
(`geom_point()`).
```



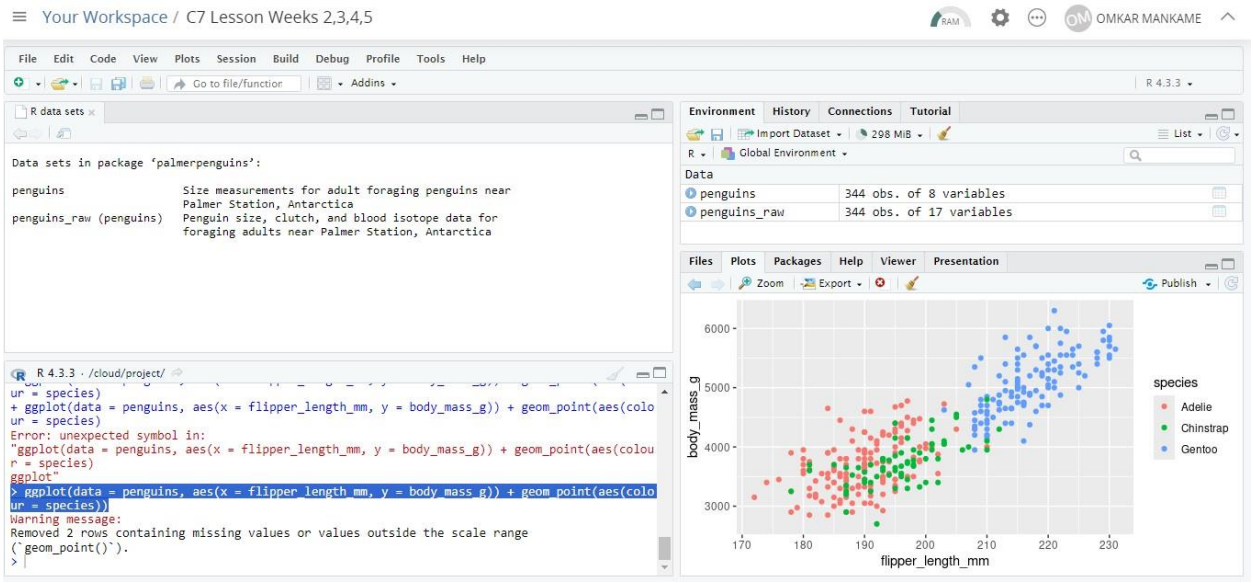

**7.** To change the color of the scattered point to red

```
> ggplot(data = penguins, aes(x = flipper_length_mm, y = body_mass_g)) + geom
_point(colour = "red")
```




**8.** To mark different colors for different species

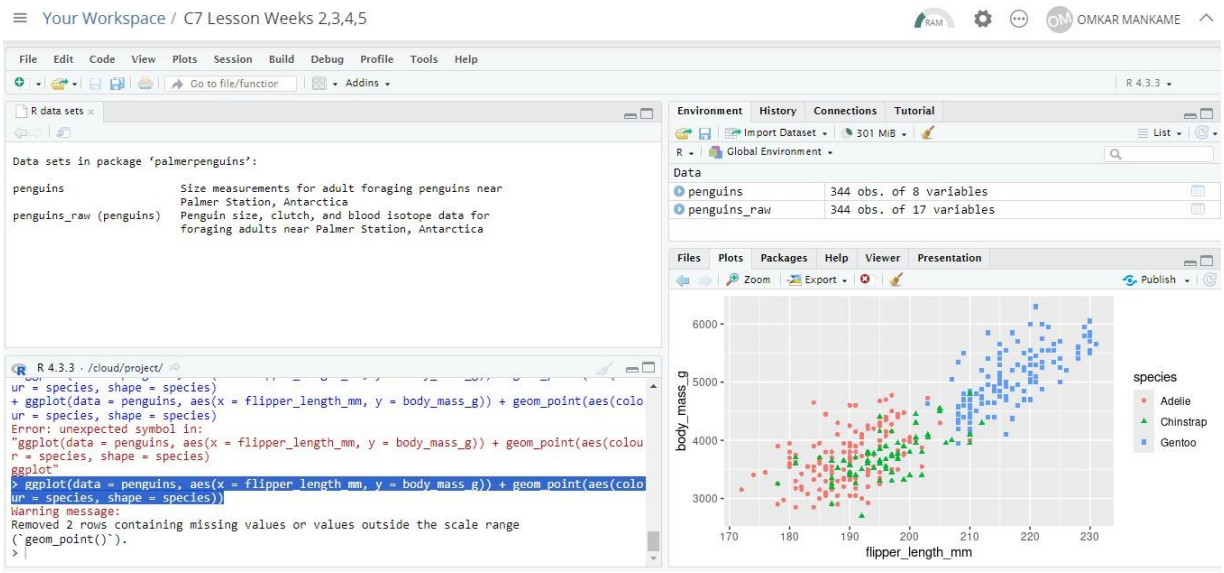
```
> ggplot(data = penguins, aes(x = flipper_length_mm, y = body_mass_g)) + geom
_point(aes(colour = species))
```




<u>The plot shows that Gentoo penguins are the largest. R has created automatic legends for the plot to help us understand the color coding.</u>

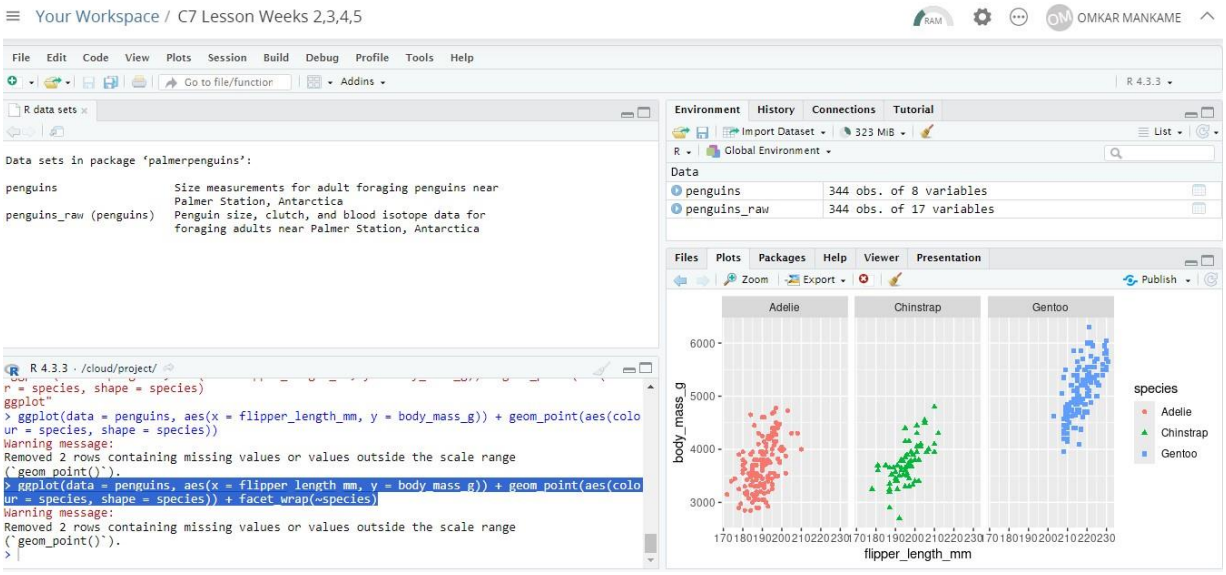
**9.** To create different colors and shapes for different species in the scattered plot shape was added in aesthetics.

```
> ggplot(data = penguins, aes(x = flipper_length_mm, y = body_mass_g)) + geom
_point(aes(colour = species, shape = species))
```



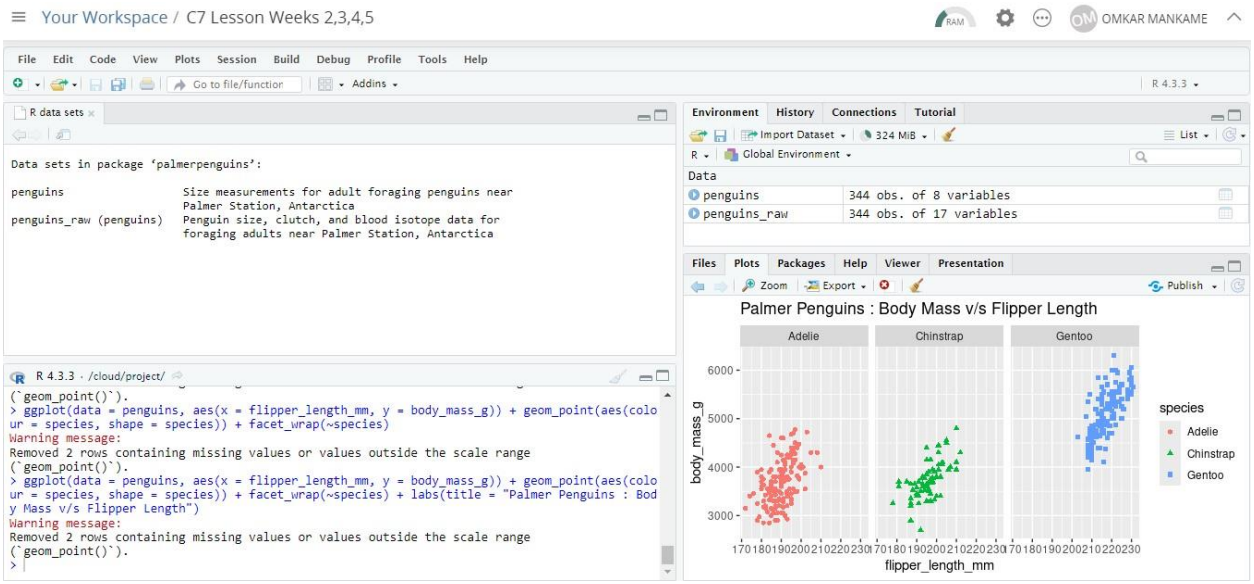

**10.** Now the subsets of the plot were created for each species using facet wrap.

```
> ggplot(data = penguins, aes(x = flipper_length_mm, y = body_mass_g)) + geom
_point(aes(colour = species, shape = species)) + facet_wrap(~species)
```



**11.** Now a title was given to our plots

```
> ggplot(data = penguins, aes(x = flipper_length_mm, y = body_mass_g)) + geom
_point(aes(colour = species, shape = species)) + facet_wrap(~species) + labs(
title = "Palmer Penguins : Body Mass v/s Flipper Length")
```




**12.** The analysis was then saved using R Markdown. It is a tool to document analysis in Rstudio.

First the package was installed

```
> install.packages("rmarkdown")
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.3’
(as ‘lib’ is unspecified)
trying URL 'http://rspm/default/__linux__/focal/latest/src/contrib/rmarkdown_2.28.tar
.gz'
Content type 'application/x-gzip' length 2618894 bytes (2.5 MB)
==================================================
downloaded 2.5 MB

* installing *binary* package ‘rmarkdown’ ...
* DONE (rmarkdown)

The downloaded source packages are in
        ‘/tmp/Rtmp8aC7uw/downloaded_packages’
```

**13.** R Markdown script was created by using the file in RStudio.