DATA, DATA & DATA

R untuk Visualisasi dan Analisa Data

Retno Novvitasari Hery Daryono

2019-11-04


tidyverse
ggplot2
blogdown
tidymodels
broom
devtools
withr
Shiny
usethis
tibble
dbplot
tidyr
tidypredict
tidyposterior
bookdown
covr
dplyr

Halo Dunia, ayo kita mulai

Acknowledgement : The material in this slide mostly inspired by Saghir Bashir post on Github

https://github.com/saghirb/Getting-Started-in-R

Tidyverse Package

- adalah kumpulan dari beberapa package yang berfungsi untuk mengimport, memanipulasi, mengeksplorasi, memvisualisasi data dengan cara yang baik sehingga membantu penggunanya untuk lebih produktif

Delapan package dalam dunia Tidy:

1. ggplot2 -> visualisasi data
2. dplyr -> manipulasi data
3. tidyr -> merapikan data
4. readr -> mengimport data
5. purrr -> pemrograman fungsional
6. tibble -> tibble
7. stringr -> strings
8. forcats -> factor

Menginstal Tidyverse

- Instal package **tidyverse** menggunakan salah satu dari 2 opsi

```
install.packages("tidyverse")
```

- atau langsung dari panel Tools di RStudio

Memanggil tidyverse

```
library(tidyverse)
```

ChickWeight Data

- salah satu dataset yang telah terinstal di R
- tentang berat ayam yang diukur mulai tiap dua hari mulai dari menetas sampai hari ke-21
- Terdapat 4 kelompok ayam didasarkan pada asupan proteinnya

Mau tahu lebih banyak tentang ChickWeight ?

(gunakan salah satu fungsi yang telah dipelajari di sesi sebelumnya)

Mengimpor Data (readr)

- menggunakan package readr
- pastikan bahwa data yang akan dimasukkan ke R sudah ada dalam directory

```
CW ← read_csv("ChickWeight.csv")
```

Melihat data

```
CW
```

Mau tahu data nilai data yang lebih lengkap ?

```
glimpse(CW)
```

fungsi glimpse() memungkinkan kita untuk melihat data secara keseluruhan(data.frame)

Tantangan

- berapa jumlah observasi dari data ChickWeight ?
- berapa variabel ?
- ingat bahwa R bersifat case sensitif. Variabel apakah yang pola penulisannya berbeda dengan variabel lainnya ?

Memvisualisasikan data (ggplot2)

- Untuk melihat bagaimana berat ayam berubah berdasarkan jenis makanannya
- Membuat plot: variabel Time sebagai x axis, variabel weight sebagai y axis

```
ggplot (CW, aes(Time, weight)) # plotkosong
```

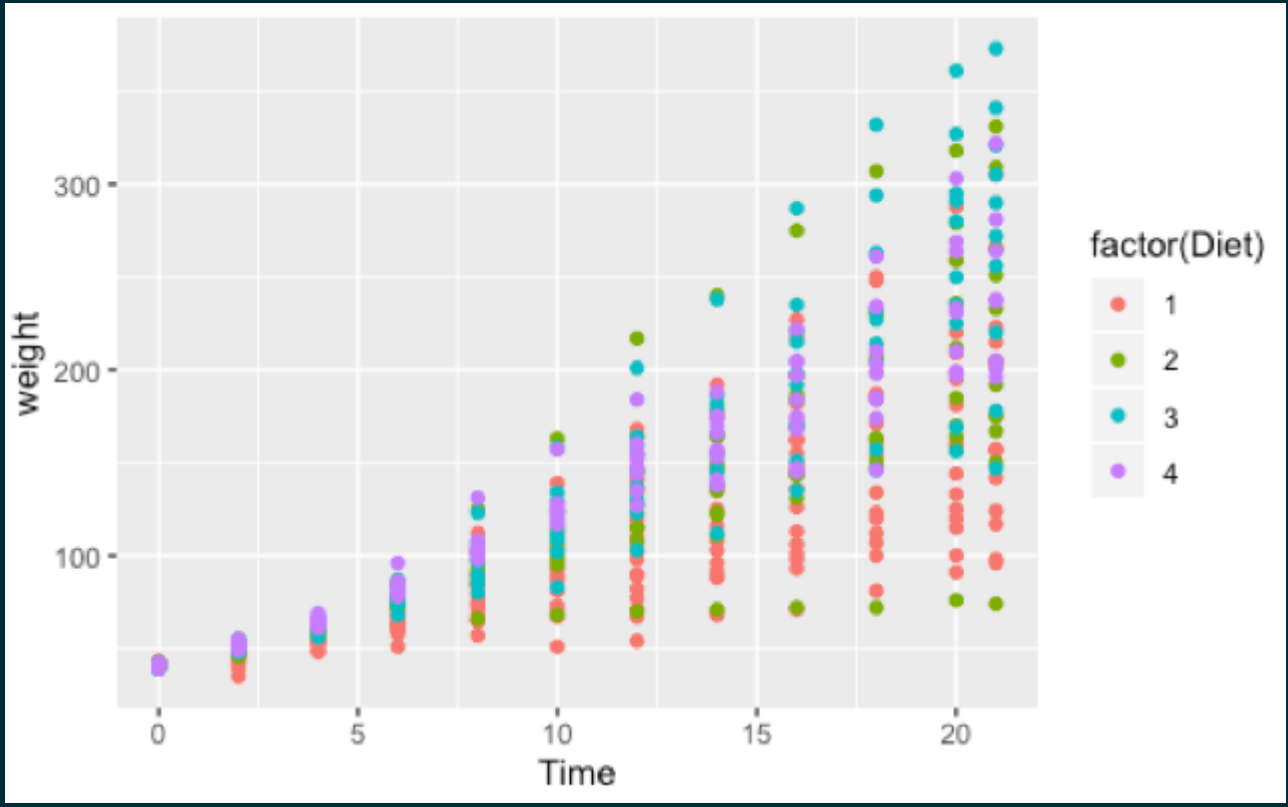
Plot Kedua

- plot yang berisi data
- menambah satu lapis plot di atas plot kosong, yaitu berupa titik (poin)

```
ggplot (CW, aes(Time, weight)) + geom_point()
```

Menambah colour untuk variabel Diet

```
ggplot (CW, aes(Time, weight, colour=factor(Diet))) + geom_point()
```



- INTERPRETASI ?

Variabel faktor

- Dari plot sebelumnya, data perlu dirubah, dari numerik yang kontinyu menjadi kategorikal
- Menggunakan fungsi mutate()

```
CW ← mutate(CW, Diet = factor(Diet))
CW ← mutate(CW, Time = factor(Time))
glimpse(CW)
```

Membuat plot terpisah berdasarkan jenis diet

- fungsi facet_wrap()

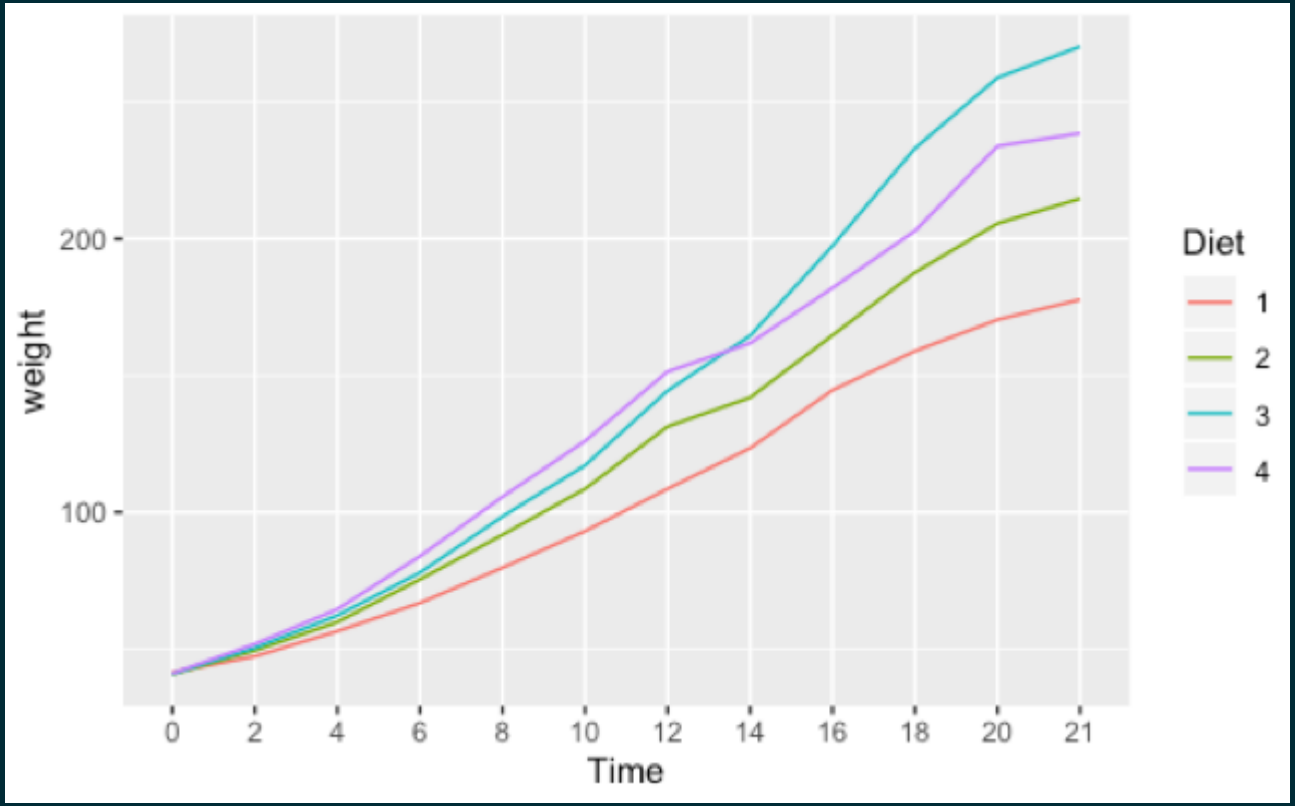
```
ggplot (CW, aes(Time, weight, colour=Diet)) +
          geom_point() +
          facet_wrap(~Diet) +
          theme(legend.position = "bottom") # bisa juga "top", "left
```

- Banyak poin yang menumpuk satu sama lain. Solusi ?
- Apakah legend masih diperlukan ?
- interpretasi : plot mana yang sedikit variabilitasnya ?

Plot rerata

- memplot rerata weight dari waktu ke waktu

```
ggplot (CW, aes(Time, weight, group=Diet, colour=Diet)) +
          stat_summary(fun.y="mean", geom="line")
```

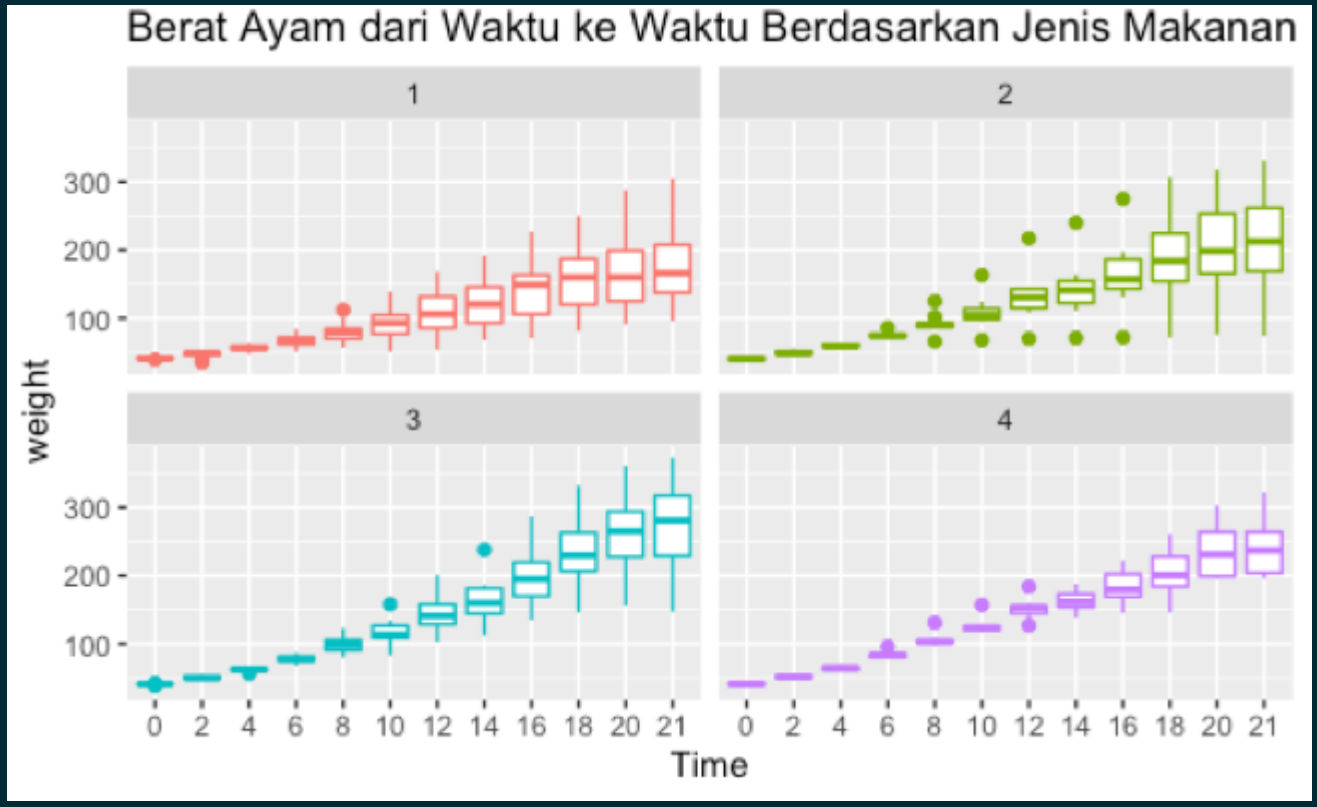


Interpretasi:

- Di akhir eksperimen, diet 3 mempunyai rerata yang paling tinggi, tapi
- bagaimana dengan variasinya (uncertainty) ?
- Jenis plot apa yang bisa dipakai untuk membantu interpretasi ?

Box-whisker plot untuk memvisualisasikan variasi

```
ggplot (CW, aes(Time, weight, colour=Diet)) +
  facet_wrap(~Diet) +
  geom_boxplot() +
  theme(legend.position = "none") +
  ggtitle("Berat Ayam dari Waktu ke Waktu Berdasarkan Jenis Makanan")
```

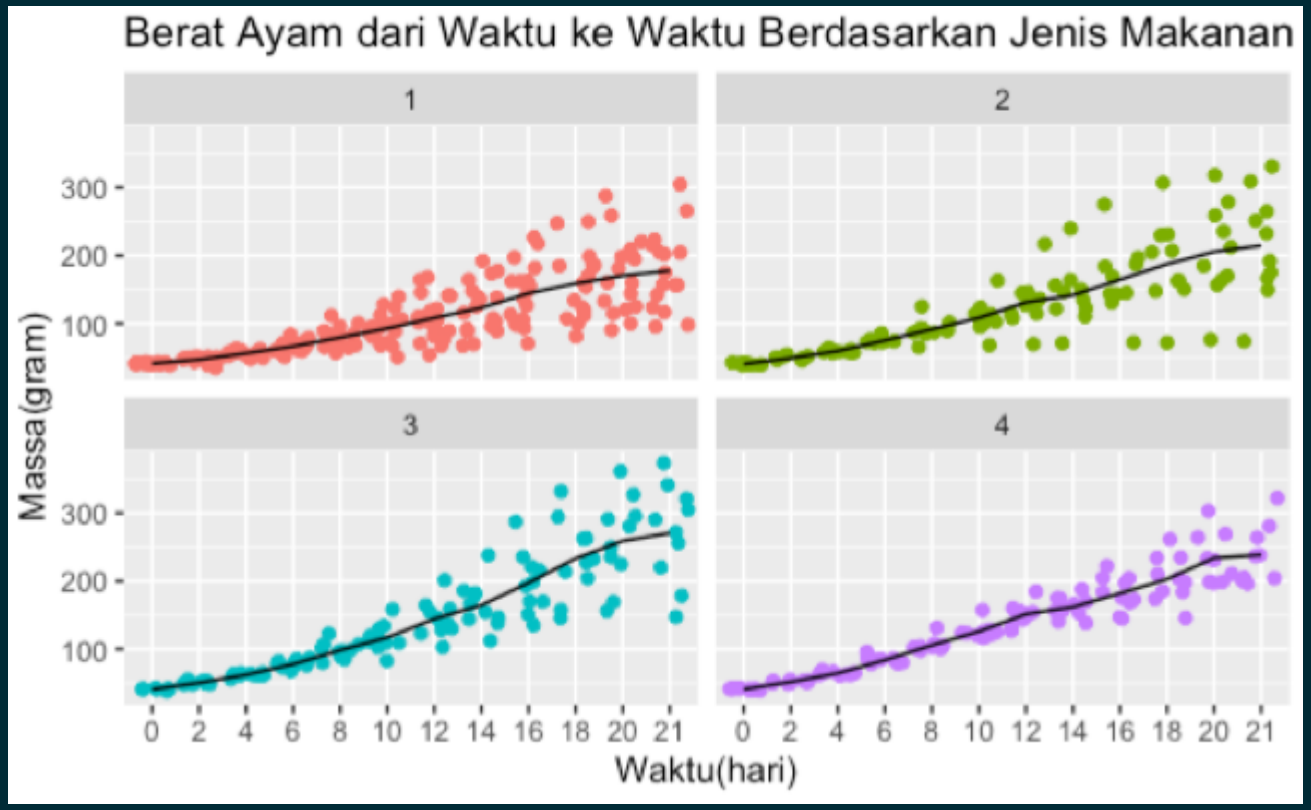


- Interpretasi
  1. Diet mana yang sepertinya mempunyai rerata tertinggi ?
  2. Diet mana yang variasinya paling kecil ?

Plot Final

- geom_jitter dan rerata, berdasarkan jenis Diet

```
ggplot (CW, aes(Time, weight, group=Diet, colour=Diet)) +
  facet_wrap(~Diet) +
  geom_jitter () +
  stat_summary(fun.y="mean", geom="line", colour="black") +
  theme(legend.position = "none") +
  ggtitle("Berat Ayam dari Waktu ke Waktu Berdasarkan Jenis Makanan")
  xlab("Waktu(hari)") +
  ylab("Massa(gram)")
```


Berat Ayam dari Waktu ke Waktu Berdasarkan Jenis Makanan
1
2
3
4
Massa (gram)
300
200
100
0 2 4 6 8 10 12 14 16 18 20 21
Waktu(hari)

Wrangling Data dalam Tidyverse

- Konsep Piping (%>%) hint: Ctrl + Shift + M (**Windows**)
- Idealnya, biasanya kita melakukan serangkaian wrangling data untuk mendapatkan data akhir yang kita inginkan. %>% berperan penting dalam hal ini.
- %>% dapat diartikan "kemudian"

```
CW21 ← CW %>%
  filter(Time %in% c(0,21)) %>%
  rename(Weight = weight) %>%
  mutate(Group = factor(str_c("Diet", Diet))) %>%
  select(Chick, Group, Time, Weight) %>%
  arrange(Chick, Time)
```

- Kita membuat data baru namanya CW21 menggunakan dataset CW, kemudian memilih data waktu 0 dan 21 hari, kemudian menamai ulang variabel weight menjadi Weight, kemudian membuat variabel baru yaitu Group, kemudian memilih variabel Chick, Group, Time, dan Weight, kemudian mengatur data dengan Chick, disusul Time

Saatnya Latihan (Latihan 3)

- Buatlah sebuah data frame baru dengan nama CWd14, menggunakan dataset CW, dengan dengan memilih hanya dua jenis diet yaitu diet 1 dan 4, kemudian menamai variabel Chick dengan nama baru yaitu Chicken, kemudian memilih variabel antara lain Chicken, Group, Time, dan weight, kemudian mengatur urutan variabel yaitu Chicken terlebih dahulu, disusul dengan Time
- Tampilkan CWd14

```
CWd14 ← CW %>%
  filter(Diet %in% c(1,4)) %>%
  rename(Chicken = Chick) %>%
  select(Chicken, Diet, Time, weight) %>%
  arrange(Chicken, Time)
```

```
CWd14
```

Berat Ayam : Summary Statistics

- dari plot yang kita buat sebelumnya, kita menyimpulkan bahwa diet 3 mempunyai rerata tertinggi, akan tetapi diet 3 mempunyai variasi paling besar
- Selanjutnya, kita lakukan summary statistik untuk analisa lebih lanjut
- Menngetahui jumlah observasi dan rerata berdasarkan waktu dan jenis diet
- membuat tibble baru dengan nama CW2

```
CW2 ← CW %>%
  group_by(Diet, Time) %>%
  summarise(N = n(), Mean = mean(weight)) %>%
  arrange (Diet, Time)
CW2
```

- Apakah jumlah ayam dan rerata berat saja sudah cukup ?

Menambahkan komponen lainnya

- SD weight, Median weight, nilai paling kecil (min) dan paling besar (max) dari weight pada penghitungan di hari 0 dan 21

```
CW3 ← CW %>%
  filter(Time %in% c(0, 21)) %>% # memfilter Time di hari 0 dan 21
  group_by(Diet, Time) %>% # dikelompokkan berdasarkan Diet dan Waktu
  summarise(N = n(), # jumlah ayam
            Mean = mean(weight), #rerata berat ayam
            SD = sd(weight), # standar deviasi berat ayam
            Median = median(weight), #median berat ayam
            Min = min(weight), #nilai paling kecil dari berat ayam
            Max = max(weight)) %>% #nilai paling besar dari berat aya
  arrange(Diet, Time) #diatur Diet terlebih dahulu, kemudian Time
```

Analisa Data, Diet mana yang mempengaruhi pertambahan berat badan ayam ?

- Melakukan wrangling data untuk menghitung pertambahan berat ayam, yaitu dengan menghitung selisih berat ayam usia nol dan usia 21

Berat Ayam saat menetas (CWnol)

- Membuat data bernama CWnol, mengambil time 0 pada data CW,
- kemudian mengaturnya berdasarkan Chick,
- lalu merename weight menjadi Weightnol

```
CWnol ← CW %>%
  filter(Time == 0) %>%
  arrange(Chick) %>% # untuk mengetahui missing value atau tidak pada
  rename(Weightnol = weight) %>%
  select(Chick, Diet, Weightnol)
```

Berat ayam di akhir eksperimen (CWAkhir)

- membuat CW akhir dgn memfilter time = 21 pada data CW,
- merename weight menjadi Weightakhir

```
CWakhir ← CW %>%
  filter(Time == 21) %>%
  arrange(Chick) %>%
  rename(Weightakhir = weight) %>%
  select(Chick, Diet, Weightakhir)
```

Menghilangkan Data yang tidak matching

- Ada selisih jumlah observasi pada antara CWnol dan CW akhir
- Menghilangkan data yang hilang dengan membuat data baru bernama CWNolfix

```
CWnolfix ← CWnol %>%
  filter(!Chick %in% c(8, 15, 16, 18, 44))
```

merename Chick menjadi Chickid agar tidak sama dengan data CWnol,

- karena setelah ini kedua data akan digabungkan membuat data baru bernama CWakhirfix

```
CWakhirfix ← CWakhir %>%
  rename(Chickid = Chick) %>%
  select(Chickid, Weightakhir)
```

Menggabungkan CWnol dan CWakhirfix fungsi bind_cols()

```
CWfix ← bind_cols(CWnolfix, CWakhirfix)
```

Membuat variabel baru dengan nama Delta (Weightakhir - Weightnol)

```
CWdelta ← CWfix %>%
  mutate(
    Delta = Weightakhir - Weightnol,
    Diet = Diet %>% as.factor
  ) %>%
  arrange(Diet) %>%
  select(Diet, Weightnol, Weightakhir, Delta)
```

Plot Pertambahan Berat Badan Ayam

```
ggplot(CWdelta, aes(Diet, Delta)) + geom_boxplot() +
  xlab("Diet") +
  ylab("Pertambahan Berat (g)") +
  ggtitle("Pertambahan Berat Ayam Berdasarkan Jenis Diet") +
  theme_bw()
```

Descriptive statistics

```
desc ← CWdelta %>% group_by(Diet) %>%
  summarise(
    count = n(),
    mean = mean(Delta, na.rm = TRUE),
    sd = sd(Delta, na.rm = TRUE),
    Rerata.Delta = round(mean, 2),
    SD.Delta = round(sd, 2)
  ) %>%
  rename(Jumlah = count) %>%
  select (Diet, Jumlah, Rerata.Delta, SD.Delta)
```

Apakah ada beda pertambahan berat ayam berdasarkan jenis Diet ?

Analisis Varian

```
chickModel ← aov(Delta ~ Diet, data = CWdelta) ## membuat model
summary(chickModel)
```

- Hasil ANOVA, F-value nya 0,00655 (**). Hipotesis nol rejected.
- Dilanjutkan dengan Uji Post Hoc

Post Hoc Test (Tukey)

Menggunakan package multcomp

```
library(multcomp)
Posthoc ← glht(ChickModel, linfct = mcp(Diet = "Tukey"))
summary(Posthoc)
```

Making a nice table

Package formattable

```
library(formattable)
formattable(desc,
            align =c("l","c","c","r"),
            list(`Diet` = formatter(
              "span", style = ~ style(color = "black",font.weight = "
              `Rerata.Delta` = color_tile("white", "orange")
            ))
```

Diet	Jumlah	Rerata.Delta	SD.Delta
1	16	136.19	58.87
2	10	174.00	78.99
3	10	229.50	71.11
4	9	197.67	43.92

Big Thanks to :

All the participants,

Saghir Basir, for an inspiring page guide **Getting Started in R**

https://github.com/saghirb/Getting-Started-in-R

Yihui Xie, Creator of the R package **xaringan**

https://github.com/yihui/xaringan

Garrick Aden Buie, Creator of this nice theme [**xaringanthemer**]

https://github.com/gadenbuie/xaringanthemer