

3-Manipulasi Bentuk Data

Ashari Ramadhan

11/19/2020

Dplyr

Dplyr adalah package di R yang dapat digunakan untuk memanipulasi data. Package ini dikembangkan oleh Hadley Wickham dan Roman Francois yang memberikan beberapa fungsi yang mudah digunakan. Package ini sangat berguna ketika digunakan untuk melakukan analisis dan eksplorasi data.

Perintah Dalam Query

Adapun perintah dalam dplyr (dibandingkan dengan perintah pada SQL) adalah sebagai berikut:

Fungsi dalam dplyr	Fungsi dalam SQL	Keterangan
<code>select()</code>	SELECT	Menyeleksi kolom variabel
<code>filter()</code>	WHERE	Menyaring (filter) baris
<code>group_by()</code>	GROUP_BY	Mengelompokkan data
<code>summarise()</code>	tidak ada	Merangkum data
<code>arrange()</code>	ORDER_BY	Mengurutkan data
<code>mutate()</code>	COLUMN ALIAS	Membuat kolom baru
<code>join()</code>	JOIN	Menggabungkan data frame

Figure 1: perintah dplyr

install oackages dplyr atau tidyverse terlebih dahulu jika paket belum tersedia.

```
install.packages('dplyr') #menginstall packages dplyr
```

```
install.packages('tidyverse') #menginstall packages tidyverse
```

```
#load library  
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
```

Buat data seperti di bawah dengan kode R dan beri nama data_latihan

```
data_latihan <- data.frame(
  merk = factor(c('Realme', 'Vivo', 'Vivo', 'Realme', 'Xiaomi', 'Vivo',
                  'Realme', 'Xiaomi', 'Xiaomi')),
  nama_barang = factor(c('Realme C15', 'Vivo y20', 'Vivo y30', 'Realme C11', 'Xiaomi Redmi9',
                         'Vivo y50', 'Realme 7i', 'Xiaomi Mi 9T', 'Xiaomi 6x')),
  harga = c(1890000, 1619900, 2249000, 1540000, 1699000, 3499000, 2798900, 4851000, 2199000),
  penjualan_2016 = c(3, 1, 14, 11, 14, 14, 14, 12, 12),
  penjualan_2017 = c(5, 4, 9, 9, 4, 9, 9, 8, 6),
  penjualan_2018 = c(13, 10, 7, 1, 7, 10, 8, 13, 6),
  penjualan_2019 = c(15, 6, 3, 3, 8, 9, 15, 14, 10)
)
```

Melihat dimensi data

```
dim(data_latihan)
```

```
## [1] 9 7
```

Melihat 6 data pertama

```
head(data_latihan)
```

```
##      merk      nama_barang      harga penjualan_2016 penjualan_2017 penjualan_2018
## 1 Realme      Realme C15 1890000           3           5           13
## 2 Vivo        Vivo y20 1619900           1           4           10
## 3 Vivo        Vivo y30 2249000          14           9            7
## 4 Realme      Realme C11 1540000          11           9            1
## 5 Xiaomi      Xiaomi Redmi9 1699000         14           4            7
## 6 Vivo        Vivo y50 3499000          14           9           10
##      penjualan_2019
## 1              15
## 2               6
## 3               3
## 4               3
## 5               8
## 6               9
```

Melihat struktur data

```
str(data_latihan)
```

```
## 'data.frame':   9 obs. of  7 variables:
##  $ merk          : Factor w/ 3 levels "Realme","Vivo",...: 1 2 2 1 3 2 1 3 3
##  $ nama_barang    : Factor w/ 9 levels "Realme C11","Realme 7i",...: 3 4 5 1 9 6 2 8 7
```

```
## $ harga : num 1890000 1619900 2249000 1540000 1699000 ...
## $ penjualan_2016: num 3 1 14 11 14 14 14 12 12
## $ penjualan_2017: num 5 4 9 9 4 9 9 8 6
## $ penjualan_2018: num 13 10 7 1 7 10 8 13 6
## $ penjualan_2019: num 15 6 3 3 8 9 15 14 10
```

Dari output di atas dapat diperoleh beberapa informasi, seperti type object data_latihan yaitu data.frame, berisi 9 observasi dan 7 variabel, yaitu merk, nama_barang, harga dan seterusnya.

Memilih variabel data dengan select

```
data_2016 <- data_latihan %>%
  select(nama_barang, penjualan_2016)
data_2016
```

```
##      nama_barang penjualan_2016
## 1   Realme C15             3
## 2     Vivo y20             1
## 3     Vivo y30            14
## 4   Realme C11            11
## 5 Xiaomi Redmi9            14
## 6     Vivo y50            14
## 7   Realme 7i             14
## 8 Xiaomi Mi 9T            12
## 9   Xiaomi 6x             12
```

Latihan

Coba pilih variabel data nama_barang, penjualan 2018 hingga 2019

Filter harga barang > 2jt

```
filter_2jt <- data_latihan %>%
  filter(harga > 2000000)
filter_2jt
```

```
##      merk nama_barang harga penjualan_2016 penjualan_2017 penjualan_2018
## 1   Vivo   Vivo y30 2249000             14             9             7
## 2   Vivo   Vivo y50 3499000             14             9             10
## 3 Realme   Realme 7i 2798900             14             9             8
## 4 Xiaomi   Xiaomi Mi 9T 4851000            12             8             13
## 5 Xiaomi   Xiaomi 6x 2199000            12             6             6
##      penjualan_2019
## 1             3
## 2             9
## 3            15
## 4            14
## 5            10
```

Latihan

Filter data dengan harga diatas 2jt dibawah 3jt

Jawaban

```
filter_2jt <- data_latihan %>%
  filter(harga > 2000000) %>%
  filter(harga < 3000000)
filter_2jt
```

```
##   merk nama_barang  harga penjualan_2016 penjualan_2017 penjualan_2018
## 1  Vivo   Vivo y30 2249000             14             9             7
## 2 Realme  Realme 7i 2798900             14             9             8
## 3 Xiaomi  Xiaomi 6x 2199000             12             6             6
##   penjualan_2019
## 1                3
## 2                15
## 3                10
```

Filter hanya merk Vivo

```
data_latihan %>%
  filter(merk == 'Vivo')
```

```
##   merk nama_barang  harga penjualan_2016 penjualan_2017 penjualan_2018
## 1 Vivo   Vivo y20 1619900             1             4             10
## 2 Vivo   Vivo y30 2249000             14             9             7
## 3 Vivo   Vivo y50 3499000             14             9             10
##   penjualan_2019
## 1                6
## 2                3
## 3                9
```

group_by dan summarise sering digunakan bersamaan

Contoh hitung rata-rata harga handphone berdasarkan merk

```
group_data <- data_latihan %>%
  group_by(merk) %>%
  summarise(rata_rata_harga = mean(harga))
```

```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

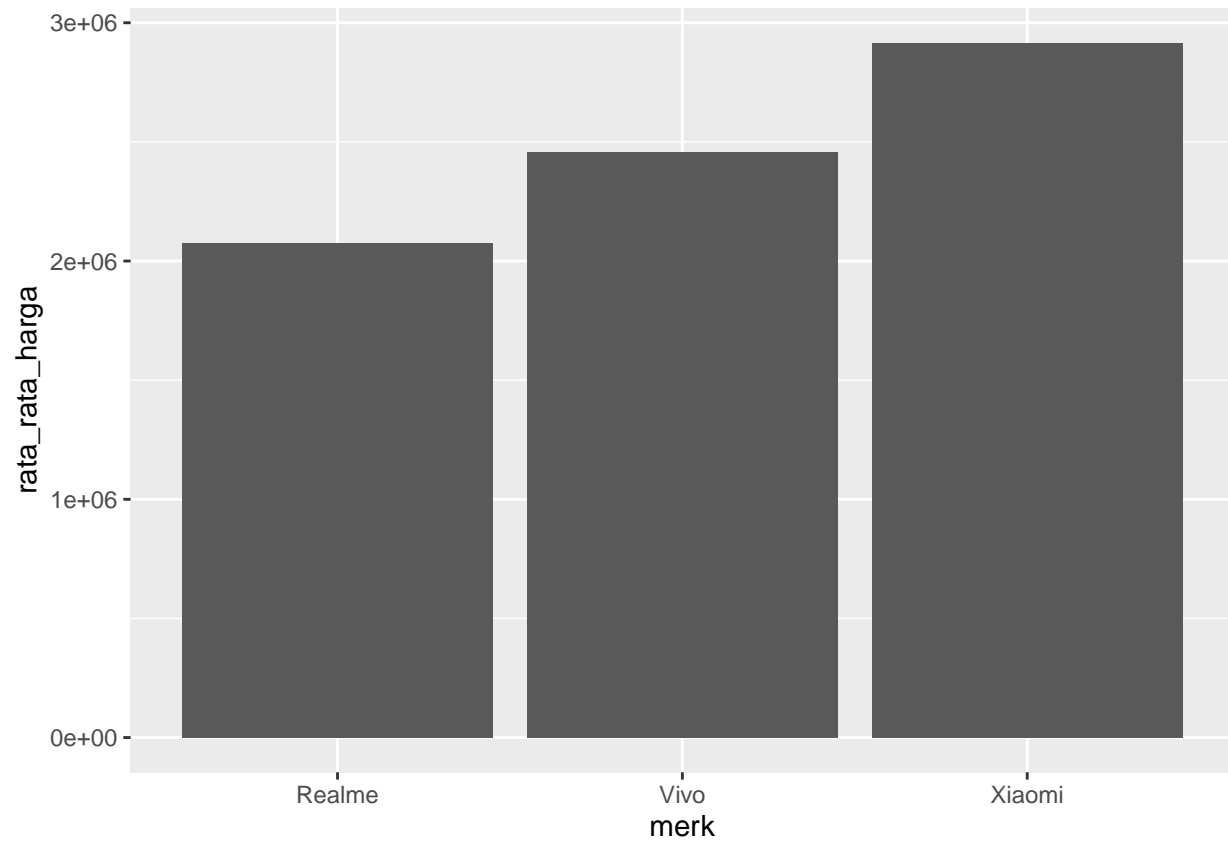
```
group_data
```

```
## # A tibble: 3 x 2
##   merk  rata_rata_harga
```

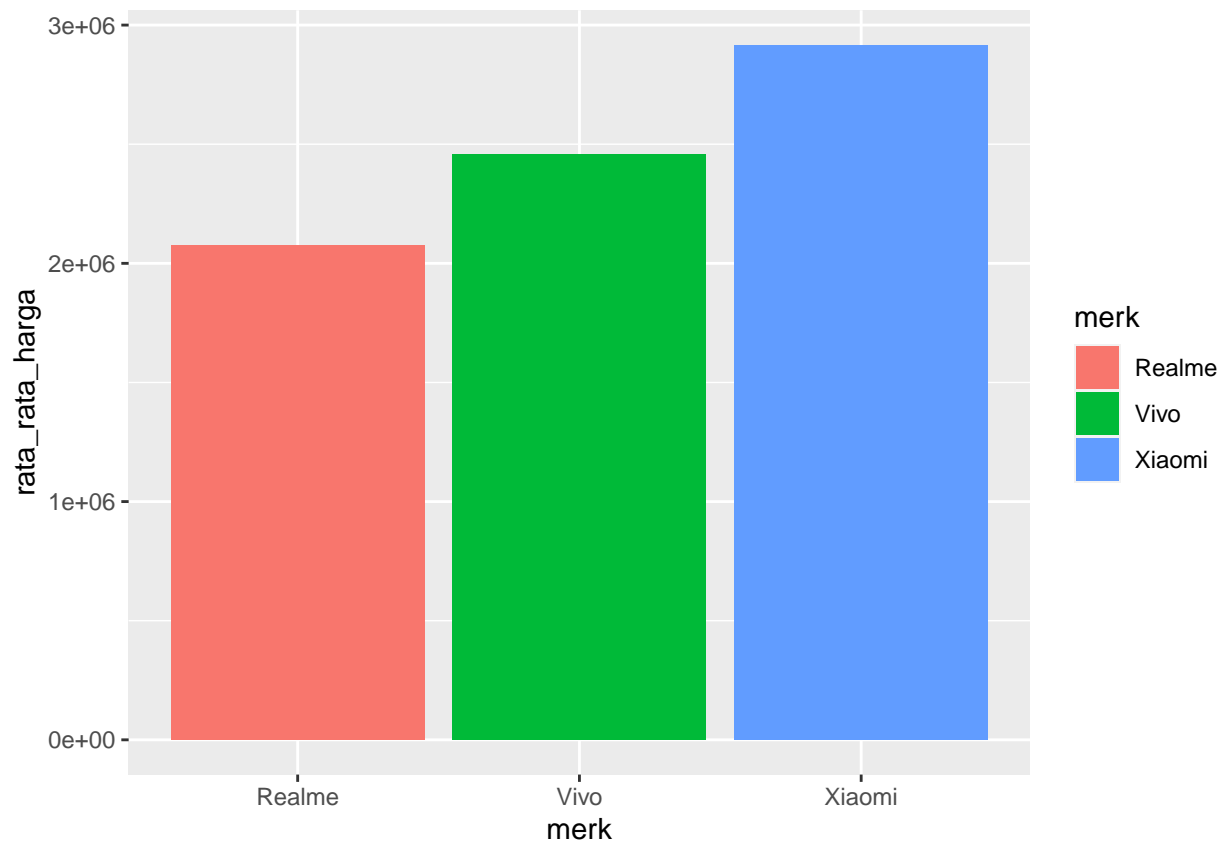
```
##   <fct>          <dbl>
## 1 Realme        2076300
## 2 Vivo          2455967.
## 3 Xiaomi        2916333.
```

Visualisasi harga barang (visualisasi akan dibahas di pertemuan lain)

```
library(ggplot2)
ggplot(group_data, aes(merk, rata_rata_harga)) +
  geom_bar(stat = 'identity')
```



```
ggplot(group_data, aes(merk, rata_rata_harga, fill=merk)) +
  geom_bar(stat = 'identity')
```



Arrange

Mengurutkan data, misal ingin mengurutkan smartphone dari termurah ke termahal

```

arrange_data_min <- data_latihan %>%
  arrange(harga)
arrange_data_min

```

```

##      merk   nama_barang   harga penjualan_2016 penjualan_2017 penjualan_2018
## 1 Realme   Realme C11 1540000             11             9             1
## 2 Vivo     Vivo y20 1619900              1             4            10
## 3 Xiaomi  Xiaomi Redmi9 1699000          14             4             7
## 4 Realme   Realme C15 1890000              3             5            13
## 5 Xiaomi   Xiaomi 6x 2199000            12             6             6
## 6 Vivo     Vivo y30 2249000            14             9             7
## 7 Realme   Realme 7i 2798900            14             9             8
## 8 Vivo     Vivo y50 3499000            14             9            10
## 9 Xiaomi  Xiaomi Mi 9T 4851000          12             8            13
##   penjualan_2019
## 1              3
## 2              6
## 3              8
## 4             15
## 5             10
## 6              3

```

```
## 7          15
## 8           9
## 9          14
```

Mengurutkan data, dari termahal ke murah

```
arrange_data_max <- data_latihan %>%
  arrange(desc(harga))
arrange_data_max
```

```
##      merk      nama_barang      harga penjualan_2016 penjualan_2017 penjualan_2018
## 1 Xiaomi  Xiaomi Mi 9T 4851000          12           8           13
## 2 Vivo     Vivo y50 3499000          14           9           10
## 3 Realme   Realme 7i 2798900          14           9           8
## 4 Vivo     Vivo y30 2249000          14           9           7
## 5 Xiaomi   Xiaomi 6x 2199000          12           6           6
## 6 Realme   Realme C15 1890000          3           5          13
## 7 Xiaomi   Xiaomi Redmi9 1699000        14           4           7
## 8 Vivo     Vivo y20 1619900           1           4          10
## 9 Realme   Realme C11 1540000         11           9           1
##      penjualan_2019
## 1          14
## 2           9
## 3          15
## 4           3
## 5          10
## 6          15
## 7           8
## 8           6
## 9           3
```

Mutate, untuk membuat kolom baru

Contoh hitung buat kolom pendapatan 2016, dengan mengalikan harga dan penjual_2016

```
mutate_data <- data_latihan %>%
  mutate(pendapatan_2016 = harga * penjualan_2016)
mutate_data
```

```
##      merk      nama_barang      harga penjualan_2016 penjualan_2017 penjualan_2018
## 1 Realme   Realme C15 1890000           3           5          13
## 2 Vivo     Vivo y20 1619900           1           4          10
## 3 Vivo     Vivo y30 2249000          14           9           7
## 4 Realme   Realme C11 1540000         11           9           1
## 5 Xiaomi   Xiaomi Redmi9 1699000        14           4           7
## 6 Vivo     Vivo y50 3499000          14           9          10
## 7 Realme   Realme 7i 2798900          14           9           8
## 8 Xiaomi   Xiaomi Mi 9T 4851000        12           8          13
## 9 Xiaomi   Xiaomi 6x 2199000          12           6           6
##      penjualan_2019 pendapatan_2016
## 1          15          5670000
## 2           6          1619900
```

```
## 3          3      31486000
## 4          3      16940000
## 5          8      23786000
## 6          9      48986000
## 7         15      39184600
## 8         14      58212000
## 9         10      26388000
```

Referensi

<https://muhammadilhammubarak.wordpress.com/2018/05/01/manipulasi-data-dengan-librarydplyr-di-r/>

<https://rpubs.com/arumprimandari/368022>

Data Tidying

<https://garrettgman.github.io/tidying/>

Bentuk manipulasi selanjutnya adalah mengorganisir bentuk data
ada 4 fungsi yang sering digunakan

`Spread()`



Figure 2: image.png

gather()

country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4

Figure 3: image.png

separate()

memisahkan kolom

unity

menggabungkan kolom

Melihat ulang data

```
head(mutate_data)
```

```
##      merk   nama_barang   harga penjualan_2016 penjualan_2017 penjualan_2018
## 1 Realme   Realme C15 1890000           3           5           13
## 2 Vivo     Vivo y20 1619900           1           4           10
## 3 Vivo     Vivo y30 2249000          14           9            7
## 4 Realme   Realme C11 1540000          11           9            1
## 5 Xiaomi   Xiaomi Redmi9 1699000         14           4            7
## 6 Vivo     Vivo y50 3499000          14           9           10
##   penjualan_2019 pendapatan_2016
## 1             15         5670000
## 2              6         1619900
## 3              3         31486000
## 4              3         16940000
## 5              8         23786000
## 6              9         48986000
```

Memilih semua variabel kecuali variabel pendapatan_2016 dan melihat 6 data terakhir

```
tidying_data <- mutate_data %>%
  select(!pendapatan_2016)
tail(tidying_data)
```

```
##      merk      nama_barang      harga penjualan_2016 penjualan_2017 penjualan_2018
## 4 Realme      Realme C11 1540000          11           9           1
## 5 Xiaomi      Xiaomi Redmi9 1699000          14           4           7
## 6 Vivo        Vivo y50 3499000          14           9          10
## 7 Realme      Realme 7i 2798900          14           9           8
## 8 Xiaomi      Xiaomi Mi 9T 4851000          12           8          13
## 9 Xiaomi      Xiaomi 6x 2199000          12           6           6
##      penjualan_2019
## 4              3
## 5              8
## 6              9
## 7             15
## 8             14
## 9             10
```

Mengganti nama kolom

```
colnames(tidying_data) <- c('merk', 'nama_barang', 'harga', '2016', '2017',
                             '2018', '2019')
head(tidying_data)
```

```
##      merk      nama_barang      harga 2016 2017 2018 2019
## 1 Realme      Realme C15 1890000     3    5   13   15
## 2 Vivo        Vivo y20 1619900     1    4   10    6
## 3 Vivo        Vivo y30 2249000    14    9    7    3
## 4 Realme      Realme C11 1540000    11    9    1    3
## 5 Xiaomi      Xiaomi Redmi9 1699000    14    4    7    8
## 6 Vivo        Vivo y50 3499000    14    9   10    9
```

Gather

```
library(tidyr)
gather_data <- tidying_data %>%
  gather('tahun', 'penjualan', 4:7)
head(gather_data)
```

```
##      merk      nama_barang      harga tahun penjualan
## 1 Realme      Realme C15 1890000  2016           3
## 2 Vivo        Vivo y20 1619900  2016           1
## 3 Vivo        Vivo y30 2249000  2016          14
## 4 Realme      Realme C11 1540000  2016          11
## 5 Xiaomi      Xiaomi Redmi9 1699000  2016          14
## 6 Vivo        Vivo y50 3499000  2016          14
```

```
tail(gather_data)
```

```
##      merk      nama_barang      harga tahun penjualan
## 31 Realme      Realme C11 1540000  2019           3
## 32 Xiaomi      Xiaomi Redmi9 1699000  2019           8
## 33 Vivo        Vivo y50 3499000  2019           9
```

```
## 34 Realme      Realme 7i 2798900 2019      15
## 35 Xiaomi     Xiaomi Mi 9T 4851000 2019      14
## 36 Xiaomi     Xiaomi 6x 2199000 2019      10
```

Variabel 2016-2019 digabung menjadi satu di variabel tahun dan nilai-nilainya di simpan di variabel penjualan

Spread

Kembalikan dari gather

```
spread_data <- gather_data %>%
  spread(tahun, penjualan)
tail(spread_data)
```

```
##      merk      nama_barang      harga 2016 2017 2018 2019
## 4   Vivo      Vivo y20 1619900      1    4    10    6
## 5   Vivo      Vivo y30 2249000     14    9    7    3
## 6   Vivo      Vivo y50 3499000     14    9   10    9
## 7  Xiaomi     Xiaomi 6x 2199000     12    6    6   10
## 8  Xiaomi     Xiaomi Mi 9T 4851000     12    8   13   14
## 9  Xiaomi     Xiaomi Redmi9 1699000     14    4    7    8
```

Data kembali ke bentuk semula

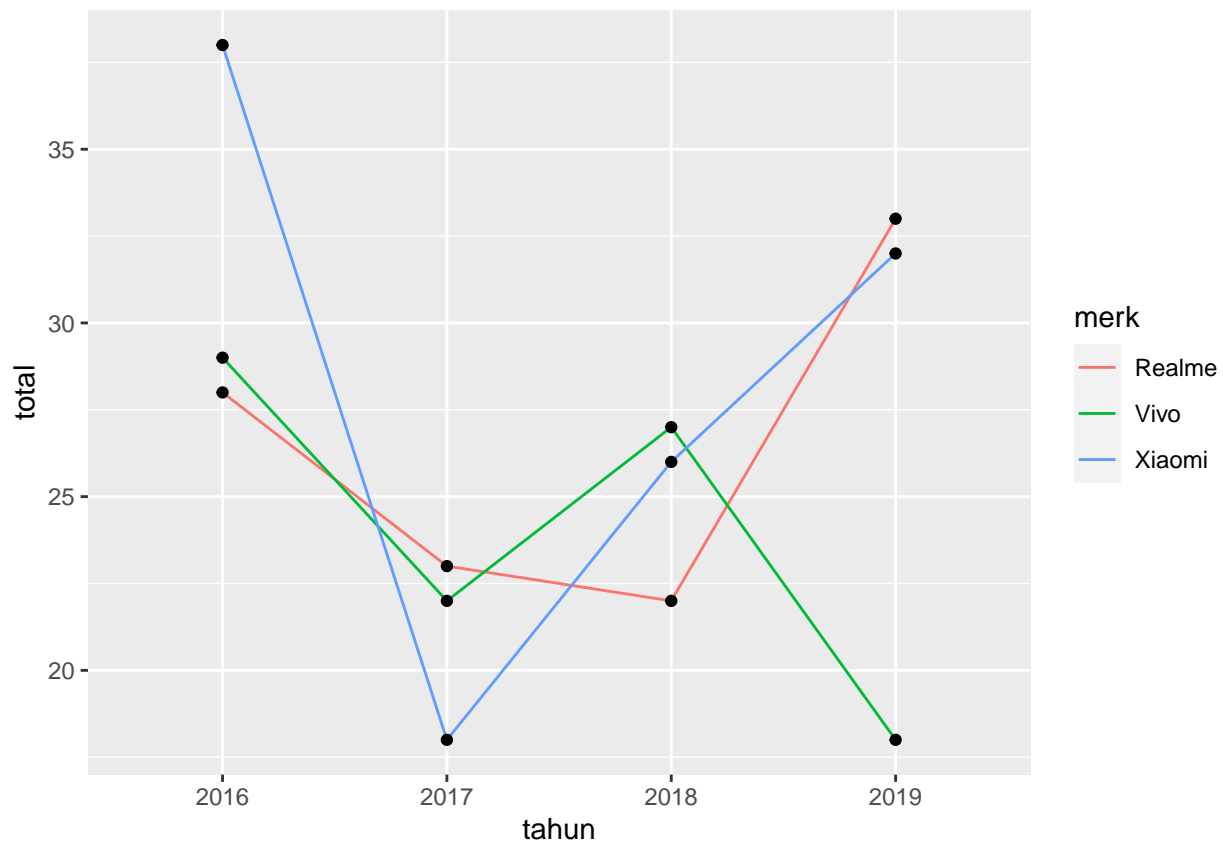
Penerapan

Contoh penggunaan kita ingin membuat plot line berbandingan penjualan dari tahun 2016-2019 sesuai data di atas, agar fungsi ggplot dapat membuat plot tersebut, data spread di atas harus di susun (manipulasi) bentuknya agar sesuai dengan yang di inginkan fungsi ggplot

```
line_data <- gather_data %>%
  group_by(merk, tahun) %>%
  summarise(total = sum(penjualan))
```

```
## 'summarise()' regrouping output by 'merk' (override with '.groups' argument)
```

```
line_data$tahun <- as.factor(line_data$tahun)
ggplot(line_data, aes(x = tahun, y= total, group = merk)) + geom_line(aes(color=merk)) +
  geom_point(aes(x = tahun, y= total))
```



Unite

Menggabung beberapa kolom menjadi satu kolom

```
head(mutate_data)
```

```
##      merk   nama_barang   harga penjualan_2016 penjualan_2017 penjualan_2018
## 1 Realme   Realme C15 1890000          3          5          13
## 2 Vivo     Vivo y20 1619900          1          4          10
## 3 Vivo     Vivo y30 2249000         14          9           7
## 4 Realme   Realme C11 1540000         11          9           1
## 5 Xiaomi   Xiaomi Redmi9 1699000        14          4           7
## 6 Vivo     Vivo y50 3499000         14          9          10
##   penjualan_2019 pendapatan_2016
## 1             15          5670000
## 2              6          1619900
## 3              3          31486000
## 4              3          16940000
## 5              8          23786000
## 6              9          48986000
```

```
unite_data <- mutate_data %>%
  unite('merk_nama', nama_barang, merk, sep = '/')
head(unite_data)
```

```
##      merk_nama  harga penjualan_2016 penjualan_2017 penjualan_2018
## 1  Realme C15/Realme 1890000          3          5          13
## 2    Vivo y20/Vivo 1619900          1          4          10
## 3    Vivo y30/Vivo 2249000         14          9          7
## 4  Realme C11/Realme 1540000         11          9          1
## 5 Xiaomi Redmi9/Xiaomi 1699000        14          4          7
## 6    Vivo y50/Vivo 3499000        14          9          10
##  penjualan_2019 pendapatan_2016
## 1          15          5670000
## 2           6          1619900
## 3           3          31486000
## 4           3          16940000
## 5           8          23786000
## 6           9          48986000
```

Dari output di atas terlihat kolom merk dan nama digabung menjadi variabel merk_nama

Separate

```
separate_data <- unite_data %>%
  separate(merk_nama, into=c('nama_barang', 'merk'), sep = '/')
head(separate_data)
```

```
##      nama_barang  merk  harga penjualan_2016 penjualan_2017 penjualan_2018
## 1  Realme C15 Realme 1890000          3          5          13
## 2    Vivo y20  Vivo 1619900          1          4          10
## 3    Vivo y30  Vivo 2249000         14          9          7
## 4  Realme C11 Realme 1540000         11          9          1
## 5 Xiaomi Redmi9 Xiaomi 1699000        14          4          7
## 6    Vivo y50  Vivo 3499000        14          9          10
##  penjualan_2019 pendapatan_2016
## 1          15          5670000
## 2           6          1619900
## 3           3          31486000
## 4           3          16940000
## 5           8          23786000
## 6           9          48986000
```

Kebalikan dari unite, separate membagi kolom menjadi beberapa kolom, sebagai contoh variabel merk_nama di bagi menjadi menjadi variabel nama_barang dan merk menggunakan separator pemisah '/'. Sehingga Realme C15/Realme akan menjadi Realme C15 dan Realme