Tujuan Pembelajaran

Setelah mempelajari modul ini, mahasiswa diharapkan:

1. Mampu memvisualisasi data dengan menggunakan ggplot debgan berbagai variasi geom.
2. Mampu memilih jenis plot yang sesuai dengan data (data kategorik maupun data kontinu).
3. Mampu mengidentifikasi outlier pada data dengan menggunaka ggplot

A. Pendahuluan

Sebelum melakukan analisis data lebih lanjut, eksplorasi data atau biasa disebut *Exploratory Data Analysis* (EDA) perlu dilakukan. Melalui eksplorasi data, atribut atau variabel pada dataset, jenis peubah, nilai atribut, distribusi data, banyaknya data dapat diketahui. EDA bukanlah suatu proses yang formal dengan aturan yang ketat, tetapi kita bebas menyelidiki ide ide yang muncul di benak kita.

EDA adalah hal yang penting dari suatu analisis apapun, karena analisis apapun yang digunakan, kualitas suatu data perlu diselidiki terlebih dahulu. Data cleaning adalah salah satu aplikasi dari EDA. Pada modul ini, akan diperkenalkan teknik eksplorasi data seperti visualisasi atau transformasi dengan menggunakan dataset yang tersedia pada program R, dengan aacuan utama mengacu pada buku yang berjudul " R for data Sience by Wickham H & Grolemund G. Sebelum melanjutkan pembicaraan mengenai visualisasi data dengan menggunakan ggplot, penting untuk mengetahui definisi dari data dan jenis data sebagai berikut:

1. Data dan Jenis Data

Istilah "data" mengacu pada kumpulan informasi hasil pengamatan mengenai atribut dari suatu objek yang bisa berupa entitas, peristiwa, ataupun proses (proses manufaktur industry).

Data dapat dikelompokkan menjadi beberapa jenis skala pengukuran yaitu data nominal, ordinal, interval dan rasio.

Data nominal, merupakan skala pengukuran yang berupa kategori. Misalnya jenis kelamin perempuan diberi simbol 1, dan jenis kelamin laki laki diberi simbol 0. Angka 1 dan 0 hanya sebagai kategori (label) saja.

Data Ordinal merupakan skala pengukuran yang dikelompokkan menjadi beberapa kategori yang memiliki tingkatan (urutan). Misalnya seseorang ditanya mengenai tingkat kesukaan terhadap produk tertentu, jawabannya bisa bertingkat tingkat misalnya tidak suka (1), agak suka (2), suka (3), atau sangat suka (4).

Data interval merupakan data berupa angka-angka dimana jarak antara angka-angka tersebut mempunyai arti, namun data ini tidak mempunyai titik nol mutlak. Misalnya si A mendapat nilai 0 pada ujian mata kuliah tertentu. Hal ini tidaklah berarti bahwa si A tidak tahu sama sekali mata kuliah yang diujikan.

Data rasio, merupakan skala pengukuran yang mempunyai nilai nol mutlak. Misalnya penghasilan seseorang, diberi angka 0 jika sama sekali tidak mempunyai penghasilan. Jika si A memiliki penghasilan 1.000.000 dan si B memiliki penghasilan 500.000, maka dapat dinyatakan bahwa penghasilan si A adalah dua kali lipat dari penghasilan si B.

2. Menyiapkan Data

Ada beberapa cara yang dapat dilakukan terkait penyiapan data. Data dapat dituliskan langsung dalam *datasheet* di R, atau bisa disimpan dalam berbagai format misalnya Excel, atau .csv. Di dalam program R juga tersedia berbagai *database* yang dapat digunakan untuk dianalisis.

3. Bekerja dengan file csv

Salah satu format file yang nyaman untuk bertukar data antara perangkat lunak yang berbeda adalah file *comma-separated value* (CSV) yang diatur dalam baris, dengan satu catatan per baris dan bidang di setiap catatan dipisahkan dengan koma. Kemampuan untuk membaca dan menulis file CSV penting karena *dataset* sering dikumpulkan, diproses, digabungkan, dan dianalisis oleh sekelompok orang dan/atau organisasi, sering menggunakan utilitas perangkat lunak yang berbeda.

Pada program R, cara paling sederhana untuk membaca file CSV adalah dengan menggunakan fungsi read.csv, sedangkan untuk menulis file CSV menggunakan fungsi

write.csv. Parameter yang diperlukan untuk kedua fungsi ini adalah nama file yang akan dibaca atau ditulis, dan nama file ini harus berupa string karakter yang sesuai, secara default, ke file di direktori kerja kita. Sebaliknya, dimungkinkan untuk membaca dari atau menulis ke file di direktori lain dengan memasukkan penunjukan path lengkap dalam file nama.

4. Visualisasi

Sebelum memulai penjelasan tentang visualisasi dengan software R, pernyataan dari salah satu ilmuwan yaitu "John Tukey perlu diutarakan sebagai berikut:

"The simple graph has brough more imporation to the data analyst's mind than any other device."-John Tukey

Grafik sederhana membawa lebih banyak informasi ke benak penganalisis data daripada perangkat mana pun." — John Tukey.

R memilki beberapa sistem untuk pembuatan graf. Salah satu yang paling serbaguna adalah menggunakan ggplot2. Melalui ggplot2, kita dapat membangun grafik lebih cepat dan menerapkannya di banyak tempat.

ggplot2 merupakan adalah satu bagian inti dari "tidyverse". Untuk mengakses kumpulan data, load "tidyverse" dengan menjalankan code:

library (tidyverse)

jika kita run "library(tidyverse), hasilnya

```
> library(tidyverse)
-- Attaching packages --------------------------------------- tid
yverse 1.3.0 --
v ggplot2 3.3.3     v purrr   0.3.3
v tibble  3.0.3     v dplyr   1.0.2
v tidyr   1.0.0     v stringr 1.4.0
v readr   1.3.1     v forcats 0.4.0
-- Conflicts ------------------------------------------ tidyverse
_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Jika kita run code"library(tidyverse)" dan mendapat pesan error " there is no package called "tidyverse", berarti kita harus menginstall package " tidyverse" kemudian me run ulang dengan perintah

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

```
library (tidyverse)
```

Menginstall sebuah package hanya dibutuhkan sekali saja, tetapi dibutuhkan untuk me run library() setiap kali memulai sesi baru. Jika kita ingin secara eksplisit tahu darimana suatu fungsi atau dataset berasal, kita bisa menggunakan bentuk khusus:

package::function()

contoh:

ggplot2::ggplot()

Hal ini memberitahu kita bahwa kita menggunakan fungsi ggplot() dari paket ggplot2.

B. STUDI KASUS

Contoh kasus 1:

Gunakan data frame "mpg" yang tersedia pada ggplot2. Data frame pada kolom menunjukkan kumpulan variabel (peubah) sedangkan pada baris berisi data observasi. "mpg" terdiri dari kumpulan observasi yang dikumpulkan oleh "the US Environmental Protection Agency" dengan 30 model mobil.

Gunakan grafik untuk menjawab pertanyaan: Apakah mobil dengan mesin besar menggunakan lebih banyak bahan bakar daripada mobil dengan mesin kecil?

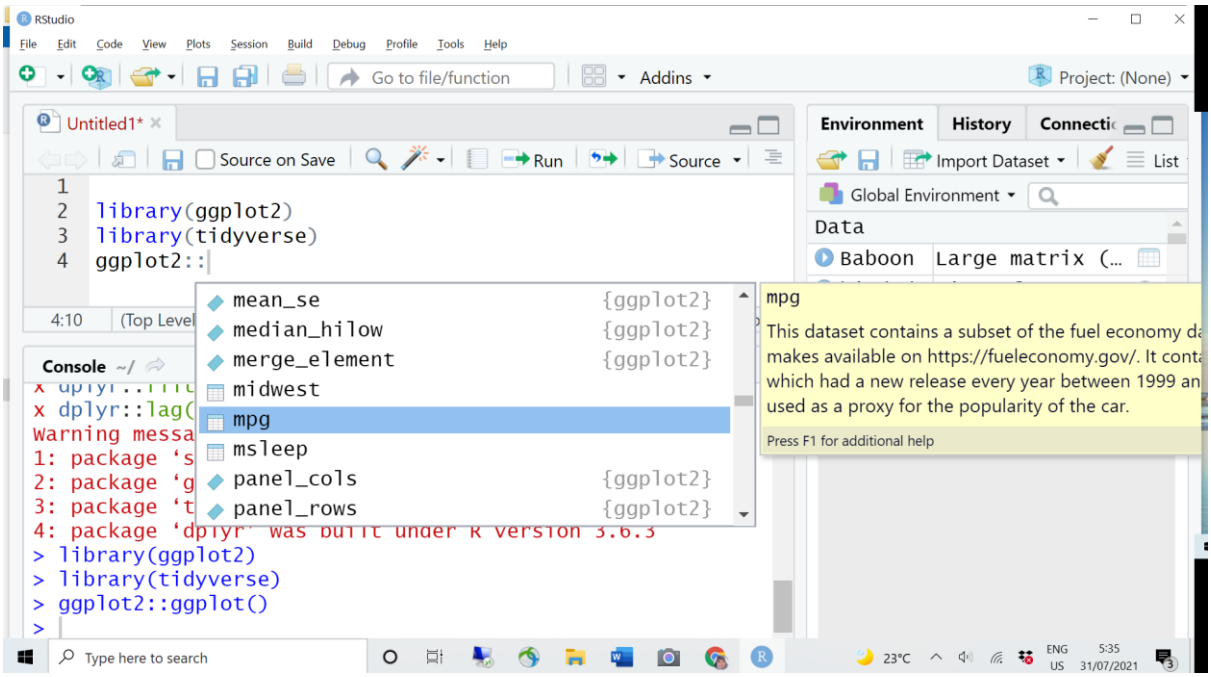
Bagaimana hubungan antara ukuran mesin (displ) dan efisiensi bahan bakar (hwy)?(Apakah hubungannya adalah positif, Negatif, Linier atau Nonlinier?

Penyelesaian:

1. Langkah pertama adalah membuk Rstudio, ketik:

ggplot2::mpg

lalu klik "run"



Hasilnya sebagai brikut:

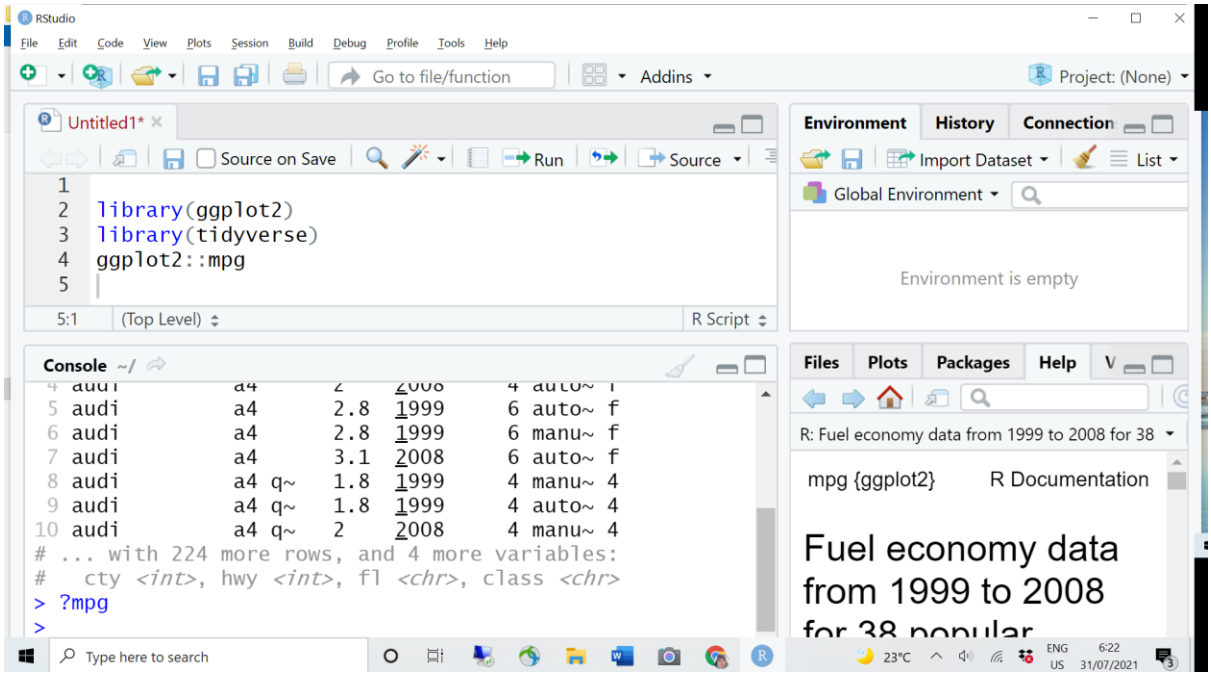
```
> ggplot2::mpg
# A tibble: 234 x 11
   manufacturer model displ  year   cyl trans drv     cty
   <chr>        <chr> <dbl> <int> <int> <chr> <chr> <int>
 1 audi         a4      1.8  1999     4 auto~ f        18
 2 audi         a4      1.8  1999     4 manu~ f        21
 3 audi         a4      2    2008     4 manu~ f        20
 4 audi         a4      2    2008     4 auto~ f        21
 5 audi         a4      2.8  1999     6 auto~ f        16
 6 audi         a4      2.8  1999     6 manu~ f        18
 7 audi         a4      3.1  2008     6 auto~ f        18
 8 audi         a4 q~   1.8  1999     4 manu~ 4        18
 9 audi         a4 q~   1.8  1999     4 auto~ 4        16
10 audi         a4 q~   2    2008     4 manu~ 4        20
# ... with 224 more rows, and 3 more variables:
#   hwy <int>, fl <chr>, class <chr>
```

Beberapa variabel yang tersedia pada data frame “mpg” yaitu manufacturer; model; displ merupakan ukuran mesin mobil (dalam liter); hwy yaitu efisiensi bahan bakar mobil di jalan raya dalam mil per galon (mpg); year; cyl yaitu banyaknya silinder; trans yaitu jenis transmisi dan variabel lainnya.

Sebuah mobil dengan efisiensi bahan bakar rendah mengkonsumsi lebih banyak bahan bakar daripada mobil dengan efisiensi bahan bakar yang tinggi ketika menempuh jarak yang sama.

Untuk mengetahui lebih dalam mengenai data “mpg”, dapat dilakukan melalui dengan menuliskan perintah pada bagian console

?mpg



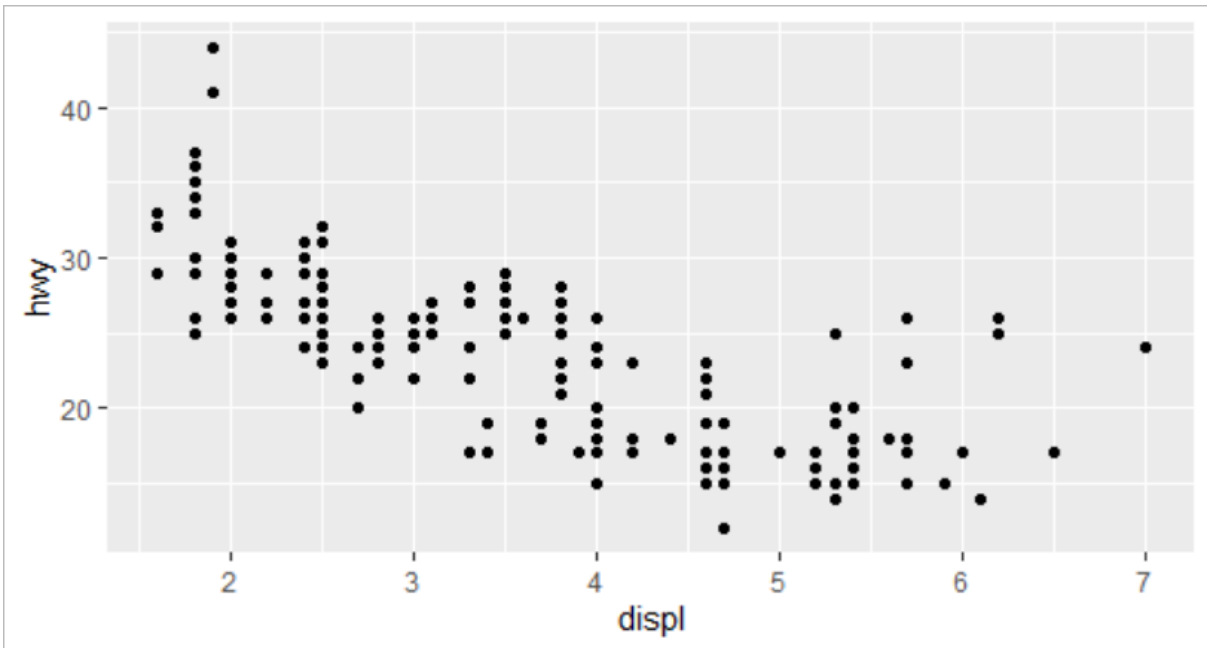
2. Membuat ggplot

Template untuk membuat grafik menggunakan ggplot adalah

```
ggplot(data = <DATA>) +
<GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

Untuk menyelesaikan kasus pertama di atas kita bisa menggunakan perintah sebagai berikut:

```
ggplot(data = mpg) +
geom_point(mapping = aes(x = displ, y = hwy))
```



Berdasarkan plot tersebut dapat disimpulkan bahwa terdapat hubungan negatif antara ukuran mesin (displ) dan efisiensi bahan bakar (hwy). Dengan kata lain, mobil dengan mesin besar menggunakan bahan bakar lebih sedikit (lebih efisien).

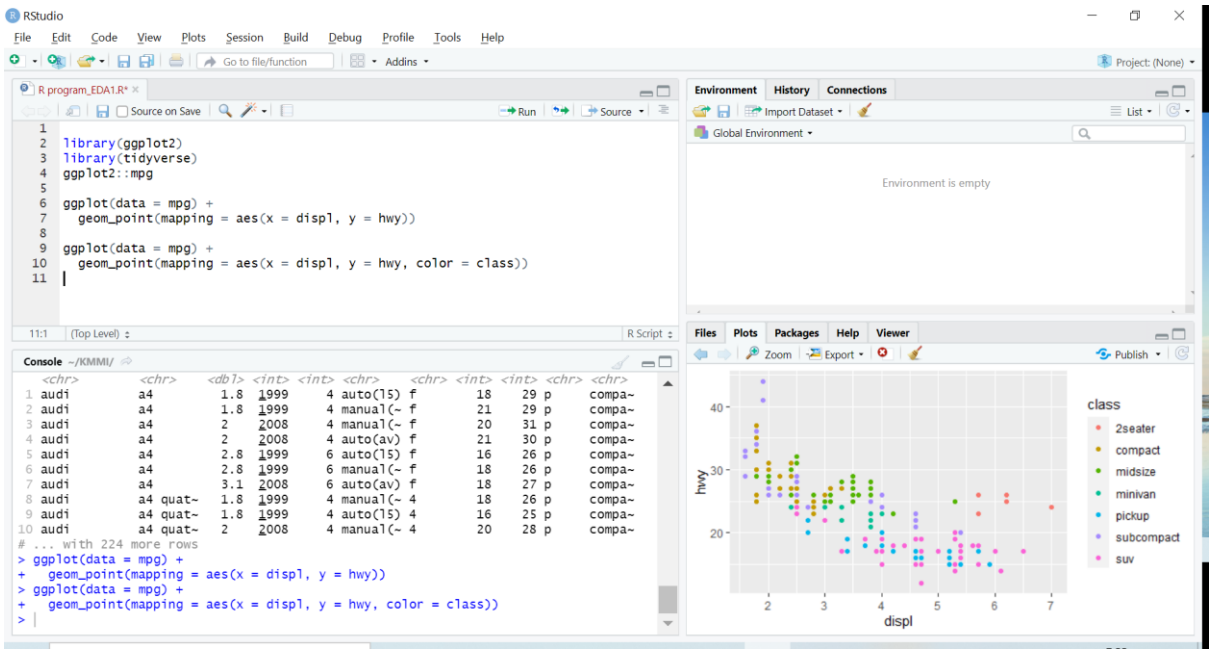
Pemetaan Estetika (Aesthetic Mappings)

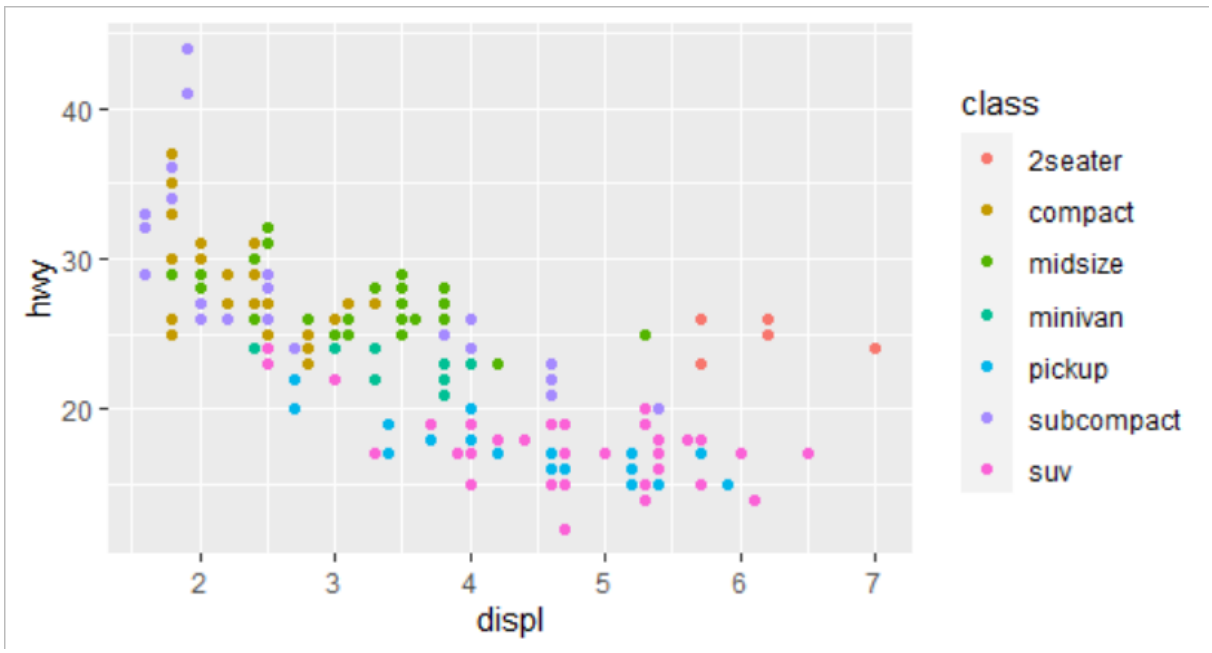
"The greatest value of a picture is when it forces us to notice what we never expected to see." — John Tukey
"Nilai yang paling berharga dari sebuah gambar adalah ketika gambar itu memaksa kita untuk memperhatikan apa yang tidak pernah kita harapkan untuk dilihat." — John Tukey

Estetika mencakup: ukuran, bentuk atau warna plot. Untuk menggambarkan sifat estetika, digunakan kata "tingkat (level)".
Plot awal yang kita buat dapat ditambahkan informasi mengenai variable "class" yaitu variable yang menjelaskan mengenai tipe mobil. Misalnya kita bisa memetakan warna ke dalam variable "class" untuk mengungapkan kelas dari setiap mobil. Perintah tersebut adalah sebagai berikut:

```
ggplot(data = mpg) +
```

```
geom_point(mapping = aes(x = displ, y = hwy, color = class))
```





Dari plot tersebut menunjukkan bahwa terdapat titik titik outlier dari class “2seater” yang merupakan kelas mobil dengan dua tempat duduk.

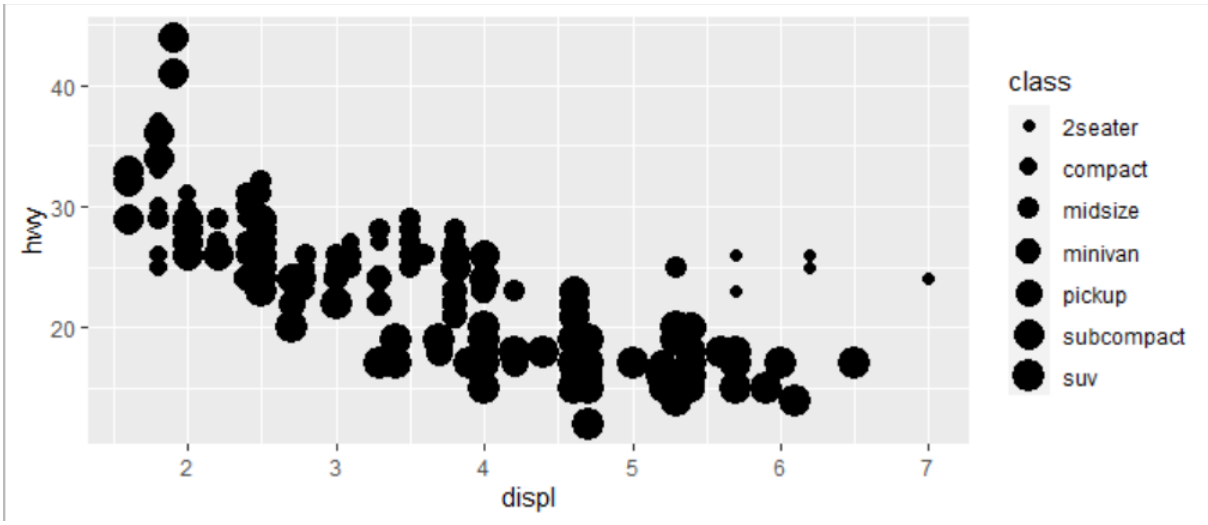
Pada contoh ini, kita memetakan variable “class” ke estetika warna. Akan tetapi, kita juga dapat memetakan “class” ke estetika ukuran.

```
> ggplot(data = mpg) +
```

```
+   geom_point(mapping = aes(x = displ, y = hwy, size = class))
Warning message:
Using size for a discrete variable is not advised.
```

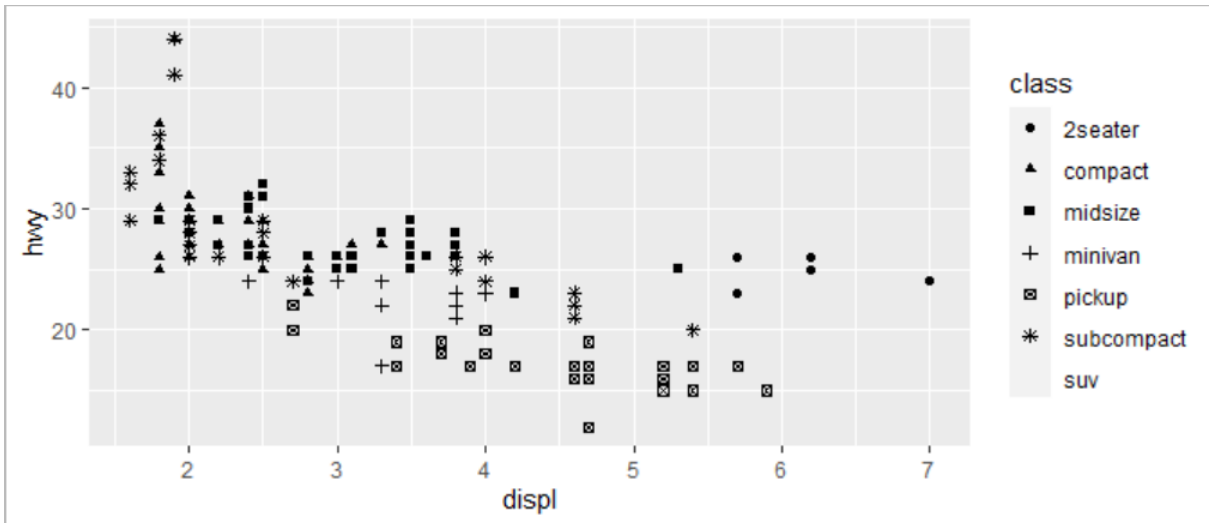
Perhatika bahwa ada "`Warning message: Using size for a discrete variable is not advised.`

Ini menyatakan bahwa tidak direkomendasikan untuk menggunakan ukuran (size) untuk varia bel yang bersifat diskrit.



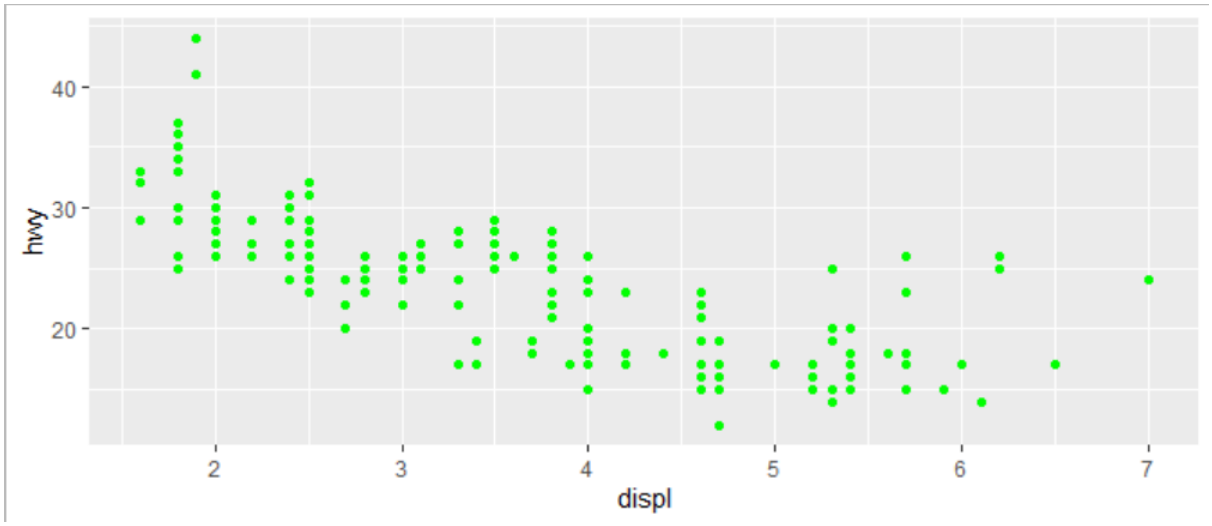
Alternative lain yang dapat digunakan adalah:

```
> ggplot(data = mpg) +
+   geom_point(mapping = aes(x = displ, y = hwy, shape = class))
Warning messages:
1: The shape palette can deal with a maximum of 6 discrete
values because more than 6 becomes difficult to
discriminate; you have 7. Consider specifying shapes
manually if you must have them.
2: Removed 62 rows containing missing values (geom_point).
```



Kita juga dapat mengatur plot kita secara manual dengan membubuhkan warna sesuai dengan selera kita misalnya hijau sebagai berikut:

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), color = "green")
```



Kesalahan Umum

Salah satu masalah umum yang sering terjadi saat membuat grafik ggplot2 adalah meletakkan tanda + di tempat yang salah. Tanda itu harus muncul di akhir baris, bukan di awal. Contoh penulisan kode yang keliru seperti berikut ini.

```
ggplot(data = mpg)
```

```
+ geom_point(mapping = aes(x = displ, y = hwy))
```

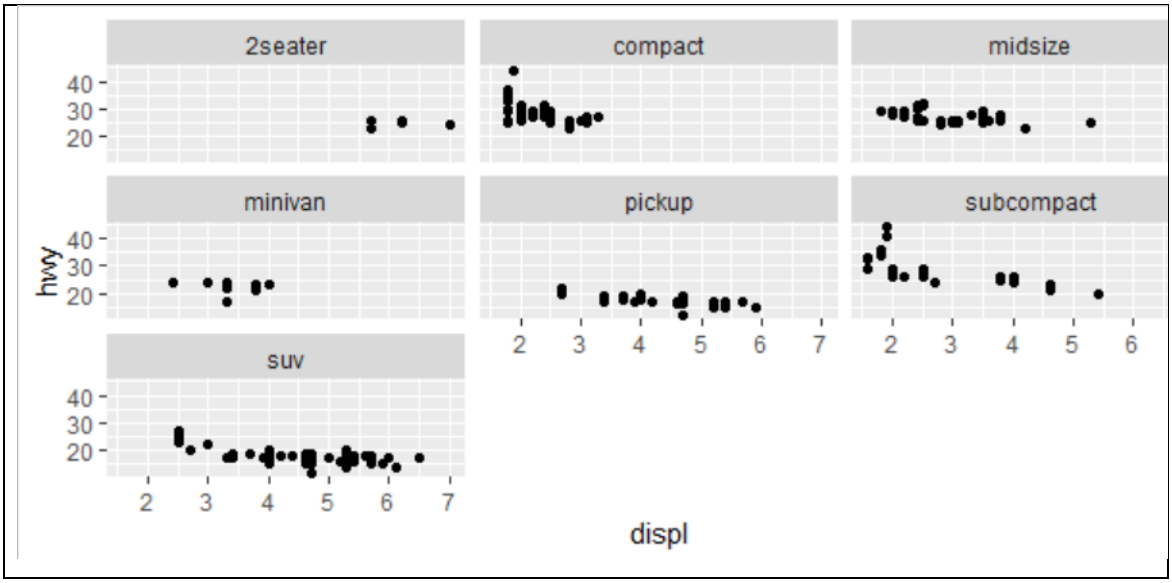
FACET

Salah satu cara untuk menambahkan variabel tambahan adalah dengan estetika. Cara lain, khususnya berguna untuk variabel kategori, adalah dengan membagi plot menjadi faset, subplot yang masing-masing menampilkan satu subset data.

Untuk membagi plot dengan satu variabel, gunakan facet_wrap(). Argumen pertama facet_wrap() harus berupa rumus, yaitu dengan ~ diikuti dengan nama variabel (di sini "rumus" adalah nama struktur data di R, bukan sinonim untuk "persamaan").

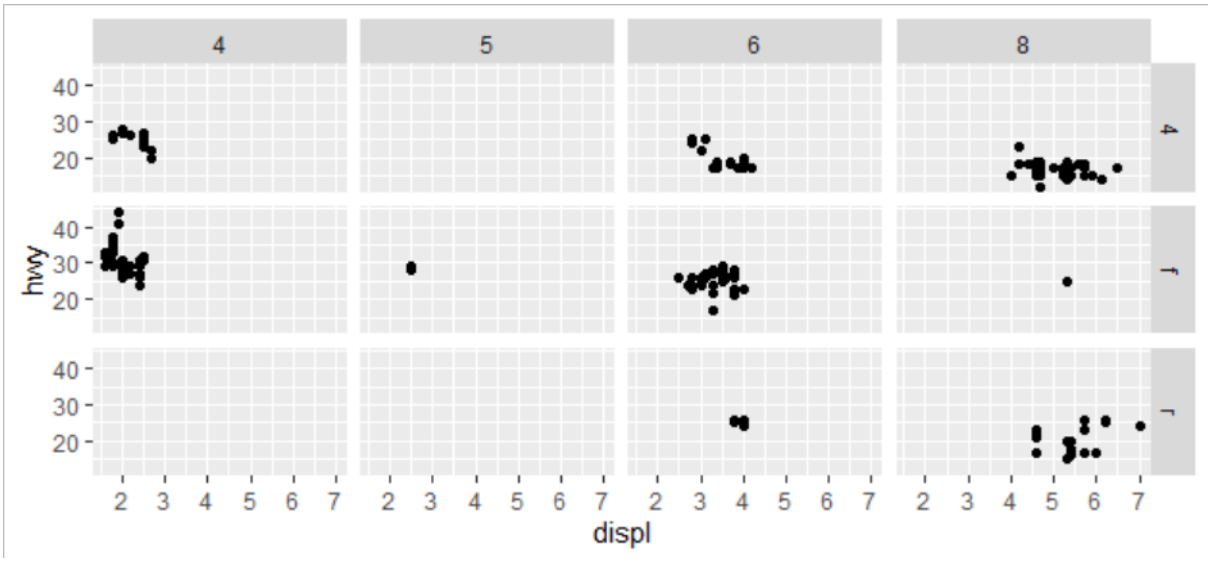
Catatan: variabel yang diberikan ke facet_wrap() harus diskrit.

```
#facet
ggplot(data = mpg) +
geom_point(mapping = aes(x = displ, y = hwy)) +
facet_wrap(~ class, ncol = 3)
```



Untuk faset plot pada kombinasi dua variabel, tambahkan facet_grid()

```
ggplot(data = mpg) +
geom_point(mapping = aes(x = displ, y = hwy)) +
facet_grid(drv ~ cyl)
```



Geometric Objects

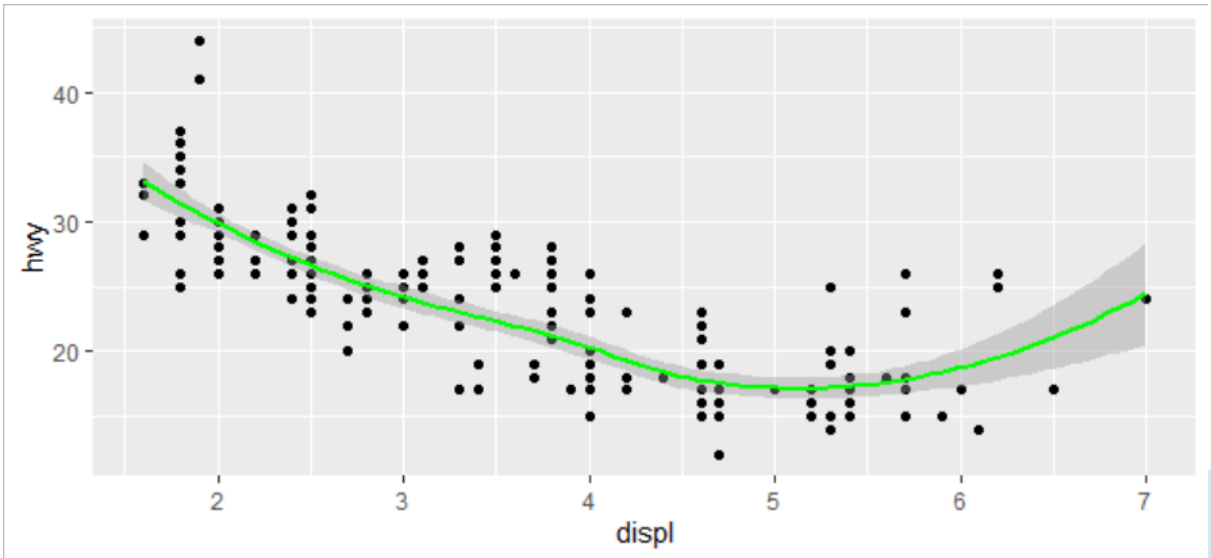
ggplot2 menyediakan lebih dari 40 jenis geom, dan paket ekstensi menyediakan lebih banyak lagi.

Berikut ini contoh penggunaan geom yang lain:

Untuk menampilkan beberapa geom dalam plot yang sama, tambahkan beberapa fungsi geom ke ggplot() seperti berikut ini:

```
> #multiple geoms in the same plot
> ggplot(data = mpg) +
+   geom_point(mapping = aes(x = displ, y = hwy)) +
+   geom_smooth(mapping = aes(x = displ, y = hwy),color = "green")
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Hasilnya sebagai berikut:



Perhatikan bahwa 2 perintah R code berikut ini menghasilkan plot yang yang sama

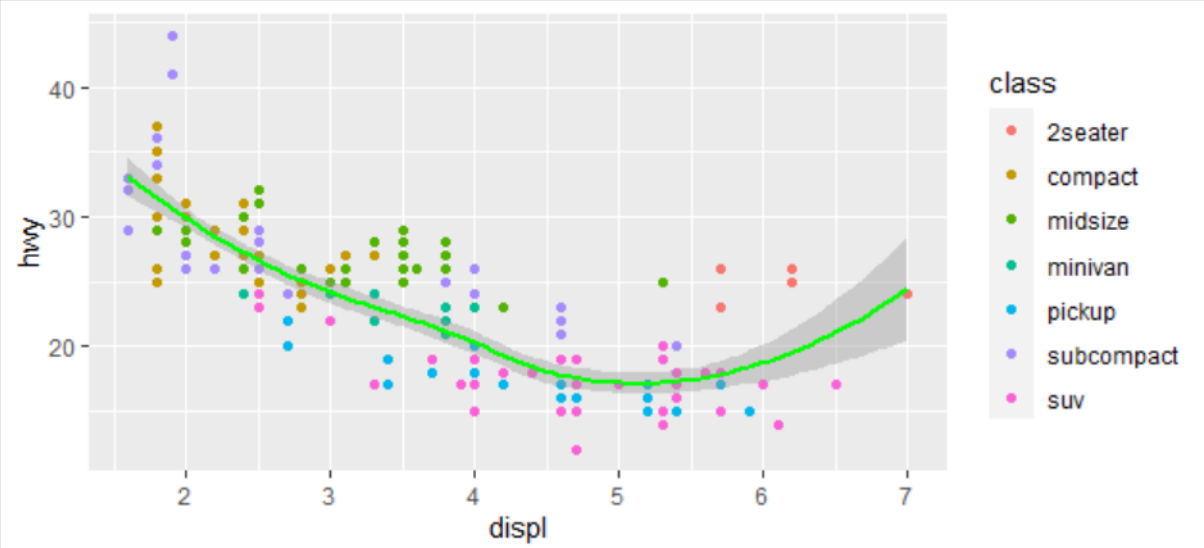
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
geom_point() +
geom_smooth()

ggplot() +
geom_point(data = mpg, mapping = aes(x = displ, y = hwy)) +
geom_smooth(data = mpg, mapping = aes(x = displ, y = hwy))
```

Alternative lainnya adalah menampilkan estetika yang berbeda pada layer yang berbeda pula sebagai berikut:

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
geom_point(mapping = aes(color = class)) +
geom_smooth(color = "green")
```

Hasilnya:



Exploratory Data Analysis

Ada 2 jenis pertanyaan yang bermanfaat untuk membuat penemuan pada data kita yaitu:

Jenis variasi apa yang terjadi pada variable kita dan jenis kovariasi apa yang terjadi di antara variable kita?

Hal penting yang perlu didefinisikan adalah:

-variabel: kuantitas, kulitas atau sifat yang dapat diukur.

-Nilai (value) adalah keadaan suatu variable. Ketika diukur nilai suatu variable dapat berubah dari pengukuran ke pengukuran.

-Pengamatan (observasi) adalah serangkaian pengukuran yang dilakukan di bawah kondisi yang sama.

Visualisasi Sebaran

Visualisasi distribusi variable tergantung pada jenis variable (kategori atau kontinu). Dalam R, variabel kategori biasanya disimpan sebagai faktor atau vektor karakter.

Jika distribusi variabel adalah kategori, maka gunakan diagram batang (bar chart).

Contoh kasus 2

- Gunakan data "diamonds" yang tersedia pada R. Buatkan diagram batang (bar chart) untuk variable kualitas potongan berlian (cut)!

Untuk melihat data gunakan perintah sebagai berikut:

```
> data = diamonds
> data
# A tibble: 53,940 x 10
   carat cut       color clarity depth table price     x     y     z
   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
 1 0.23  Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
 2 0.21  Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
 3 0.23  Good      E     VS1      56.9    65   327  4.05  4.07  2.31
 4 0.290 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
 5 0.31  Good      J     SI2      63.3    58   335  4.34  4.35  2.75
 6 0.24  Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
 7 0.24  Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
 8 0.26  Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
 9 0.22  Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
10 0.23  Very Good H     VS1      59.4    61   338  4     4.05  2.39
# ... with 53,930 more rows
```

Untuk mengetahui penjelasan mengenai data gunakan perintah di Console:

?diamonds

Penjelasan mengenai data tersebut adalah sebagai berikut:

Description

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

Usage

```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price

: price in US dollars (\$326–\$18,823)

carat

: weight of the diamond (0.2–5.01)

cut

: quality of the cut (Fair, Good, Very Good, Premium, Ideal)

color

: diamond colour, from D (best) to J (worst)

clarity

: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

x

: length in mm (0–10.74)

y

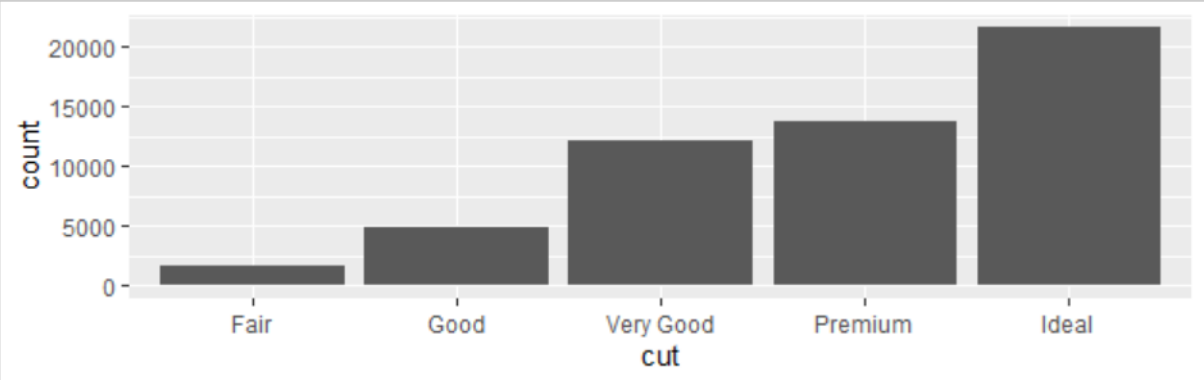
: width in mm (0–58.9)

z

: depth in mm (0–31.8)

depth

```
        total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
table
        width of top of diamond relative to widest point (43–95)
```



Sumbu y menandakan banyaknya pengamatan pada setiap nilai variable x. Untuk menghitung nilai variable x dapat digunakan perintah sebagai berikut:

```
> diamonds %>%
+   count(cut)
# A tibble: 5 x 2
  cut           n
  <ord>     <int>
1 Fair       1610
2 Good       4906
3 Very Good 12082
4 Premium   13791
5 Ideal     21551
```

Berbeda dengan varibel kategori, untuk memeriksa distribusi dari variabel kontinu (variable yang berisi angka-angka) digunakan histogram.

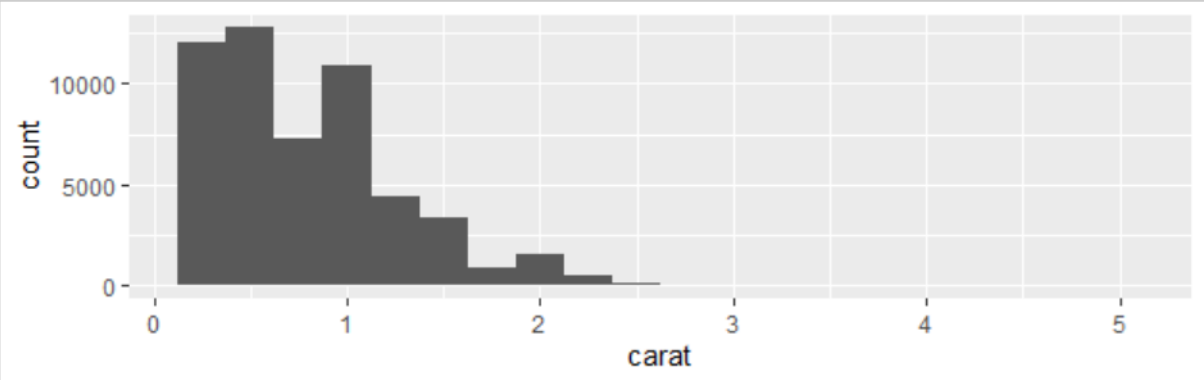
- Gunakan data “diamonds” yang tersedia pada R. Buatkan histogram untuk variable berat berlian (carat)!

Penyelesaian:

Untuk membuat histogram, kita bisa menggunakan geom_histogram sebagai berikut:

```
#Histogram
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = carat), binwidth = 0.25)
```

Hasilnya sebagai berikut:




```
> diamonds %>%
+   count(cut_width(carat, 0.25))
# A tibble: 19 x 2
   `cut_width(carat, 0.25)`     n
   <fct>                    <int>
 1 [0.125,0.375]            12024
 2 (0.375,0.625]            12763
 3 (0.625,0.875]             7286
 4 (0.875,1.12]             10934
 5 (1.12,1.38]               4412
 6 (1.38,1.62]               3398
 7 (1.62,1.88]                912
 8 (1.88,2.12]               1530
 9 (2.12,2.38]                465
10 (2.38,2.62]                152
11 (2.62,2.88]                 24
12 (2.88,3.12]                 27
13 (3.12,3.38]                  2
14 (3.38,3.62]                  3
15 (3.62,3.88]                  2
16 (3.88,4.12]                  3
17 (4.12,4.38]                  1
18 (4.38,4.62]                  1
19 (4.88,5.12]                  1
```

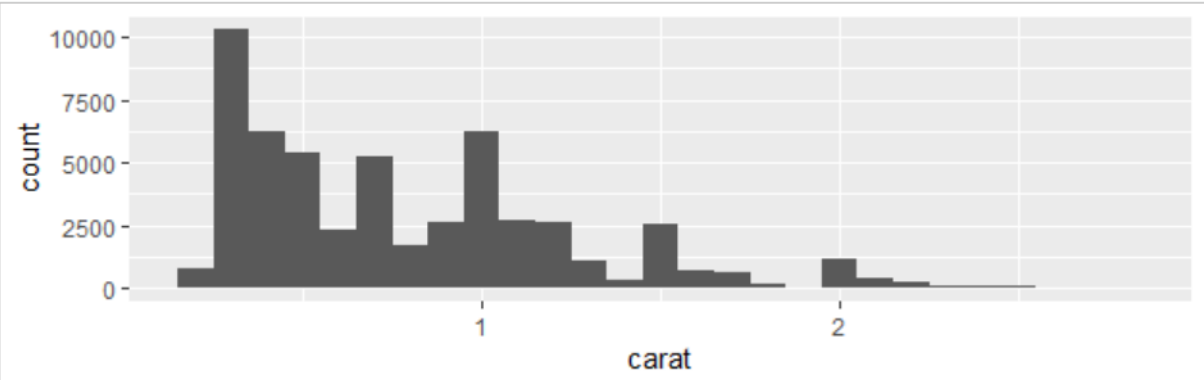
Histogram membagi sumbu-x ke dalam bin dengan spasi yang sama dan kemudian menggunakan ketinggian batang untuk menampilkan jumlah pengamatan yang ada di setiap

bin. Pada grafik di atas, batang tertinggi menunjukkan bahwa hampir 12763 pengamatan memiliki nilai karat antara 0,375 dan 0,625 yang merupakan tepi kiri dan kanan batang.

Berikut ini adalah perintah dan tampilan grafik ketika hanya menampilkan berlian dengan ukuran kurang dari 3 karat dan dengan memilih bandwith yang lebih kecil sebagai berikut:

Nama data baru yaitu smaller.

```
> smaller <- diamonds %>%
+   filter(carat < 3)
> smaller
# A tibble: 53,900 x 10
   carat cut       color clarity depth table price     x     y     z
   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
 1 0.23  Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
 2 0.21  Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
 3 0.23  Good      E     VS1      56.9    65   327  4.05  4.07  2.31
 4 0.290 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
 5 0.31  Good      J     SI2      63.3    58   335  4.34  4.35  2.75
 6 0.24  Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
 7 0.24  Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
 8 0.26  Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
 9 0.22  Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
10 0.23  Very Good H     VS1      59.4    61   338  4     4.05  2.39
# ... with 53,890 more rows
> ggplot(data = smaller, mapping = aes(x = carat)) +
+   geom_histogram(binwidth = 0.1)
```



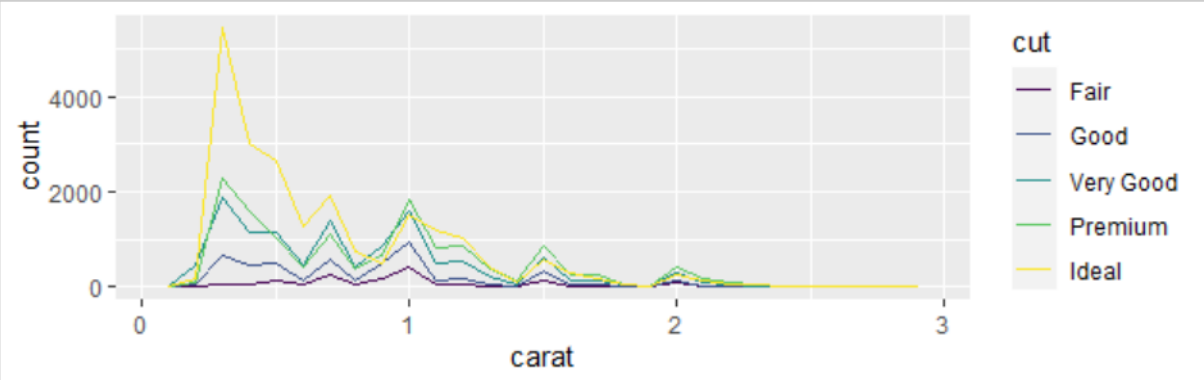
Lebar interval dalam histogram dapat diatur dengan argumen binwidth yang diukur dalam satuan variabel x. Direkomendasikan untuk menjelajahi berbagai binwidth ketika bekerja dengan histogram, karena binwidth yang berbeda dapat mengungkapkan pola yang berbeda.

Kita dapat membuat beberapa histogram dalam plot yang sama dengan menggunakan geom_freqpoly() bukan dengan geom_histogram.

geom_freqpoly() melakukan perhitungan yang sama seperti geom_histogram(), tetapi digunakan garis bukan batang. Jauh lebih mudah untuk memahami garis yang tumpang tindih dibandingkan menggunakan batang.

Diberikan contoh sebagai berikut:

```
> ggplot(data = smaller, mapping = aes(x = carat, colour = cut)) +
+   geom_freqpoly(binwidth = 0.1)
```

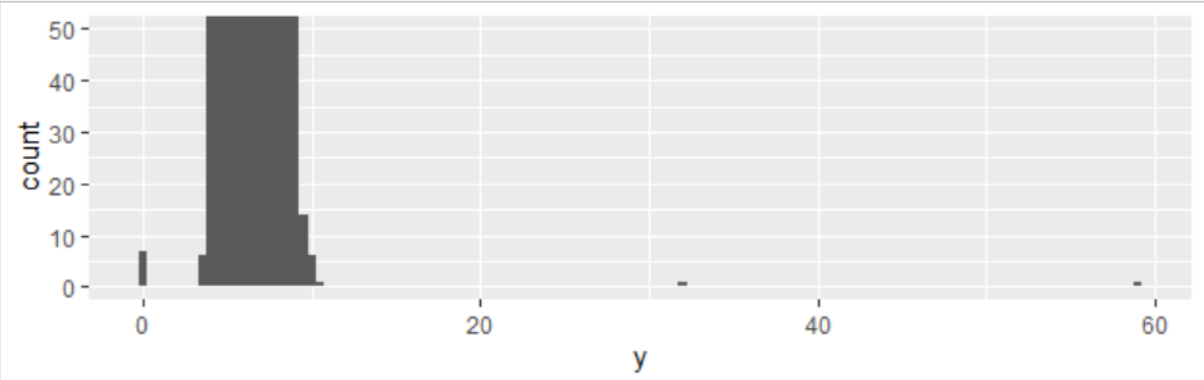


Pencilan (outlier)

Pencilan adalah pengamatan yang tidak biasa; titik data yang tampaknya berbeda dengan pola lainnya. Outlier dapat terjadi karena kesalahan entri data; atau justru outlier menyarankan ilmu baru yang penting. Ketika terdapat banyak data, outlier terkadang sulit dilihat dalam histogram. Misalnya, ambil distribusi variabel y dari dataset “diamonds”. Satu-satunya bukti outlier adalah batas lebar yang luar biasa pada sumbu x.

Untuk memudahkan melihat nilai outlier, kita perlu memperbesar nilai sumbu y yang kecil dengan coord_cartesian():

```
ggplot(diamonds) +
  geom_histogram(mapping = aes(x = y), binwidth = 0.5) +
  coord_cartesian(ylim = c(0, 50))
```



Dari plot ini, terlihat ada 3 data outlier yaitu nilai: 0, ~30 dan ~60. Hal ini dapat lebh jelas dilihat dengan perintah:

```
> unusual <- diamonds %>%
+   filter(y < 3 | y > 20) %>%
+   select(price, x, y, z) %>%
+   arrange(y)
> unusual
# A tibble: 9 x 4
  price     x     y     z
  <int> <dbl> <dbl> <dbl>
1  5139  0      0    0
2  6381  0      0    0
3 12800  0      0    0
4 15686  0      0    0
5 18034  0      0    0
6  2130  0      0    0
7  2130  0      0    0
8  2075  5.15  31.8  5.12
9 12210  8.09  58.9  8.06
```

Variabel y mengukur salah satu dari tiga dimensi berlian ini, dalam mm. Berlian tidak mungkin memiliki lebar 0 mm, jadi nilai ini pasti salah. Kita juga menduga bahwa ukuran 32 mm dan 59mm tidak masuk akal.

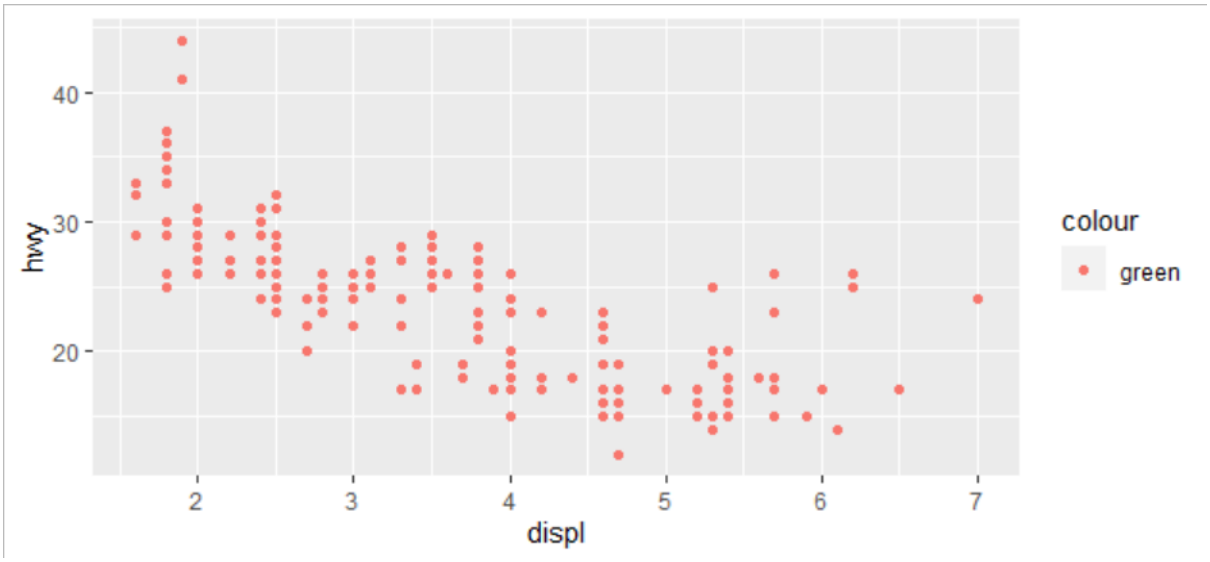
C. Latihan Mandiri

Pertanyaan terkait dengan data kasus 1 yaitu data "mpg"

1. Run ulang ggplot(data=mpg). Apa yang dapat Anda simpulkan?
2. Ada berapa baris (dan kolom yang ada dalam data ”mpg”?
3. Apa yang dijelaskan oleh variabel ”drv”?
4. Apa yang dijelaskan oleh variabel ”cty”?
5. Buat diagram pencar antara varibel ”hwy” dengan ”cyl”!
6. Variabel apa saja yang merupakan variabel kategori?
7. Variabel apa saja yang merupakan variabel kontinu?

   Petunjuk: gunakan perintah *?mpg* pada console untuk menjawab pertanyaan 2 sampai 7!
8. Jelaskan apa yang salah denga R code berikut! Mengapa warna titiknya tidak berwarna hijau?

```
ggplot(data = mpg) +
geom_point(mapping = aes(x = displ, y = hwy, color = "green"))
```



9. Geom apa yang digunakan untuk line chart, area chart dan histogram?
10. Silahkan run R code beriktu ini

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

Hasil apa yang diperoleh dengan memasukkan se =FALSE? Apa yang terjadi ketika perintah se=FALSE dihapus?

Pertanyaan terkait dengan data kasus 2 yaitu data "diamonds"

1. Lakukan eksplorasi distribusi dari masing-masing variabel x, y, dan z dalam data "diamonds". Apa yang dapat Anda pelajari? Pikirkan tentang berlian dan bagaimana dimensi panjang, lebar, dan dalamnya.

2. Lakukan eksplorasi distribusi harga. Apakah Anda menemukan sesuatu yang tidak biasa atau mengejutkan? (Petunjuk: Pikirkan baik-baik tentang lebar bin dan pastikan Anda mencoba berbagai nilai.)

3. Berapa banyak berlian 0,99 karat? Berapa banyak berlian 1 karat? Menurut Anda apa penyebab perbedaan tersebut?

4. Bandingkan dan kontraskan coord_cartesian() vs xlim() atau ylim() saat memperbesar histogram. Apa yang terjadi jika Anda membiarkan binwidth tidak disetel? Apa yang terjadi jika Anda mencoba dan memperbesar sehingga hanya setengah "bar" yang ditampilkan?

D. Rangkuman

Modul ini telah memperkenalkan Exploratory Data Analysis (EDA) dengan memperkenalkan berbagai cara untuk visualisai data dengan menggunakan ggplot melalui software R baik untuk variable kategori maupun variable kontinu. Telah diperkenalkan juga berbagai fungsi geom yang dapat digunakan untuk berbagai data. Cara untuk mendeteksi outlier pada data juga diilistrasikan dengan menggunakan ggplot.

F. Daftar Pustaka

1. Wickham, H. &Grolemund, G. 2017. R for data Science. Andira Publisher: Makassar.
2. Pearson, R.K. 2018. Exploratory Data Analysis Using R. CRC Press: Boca Raton, London