

Eksplorasi dan Visualisasi Data

Pertemuan 1:
Pengantar Eksplorasi dan Visualisasi Data

Highlights

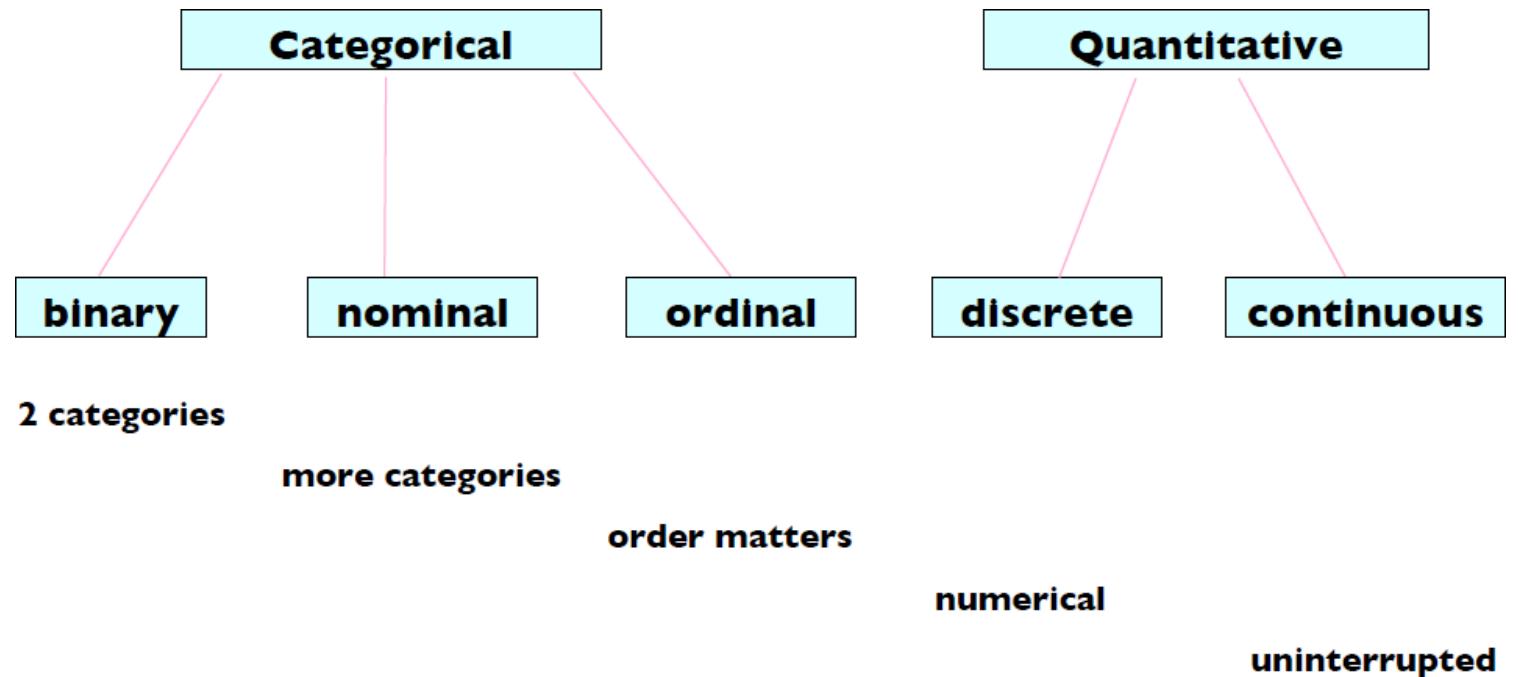
1. Data & Dataset
2. Introduction to Data Exploration
3. Introduction to Data Visualization
4. Good vs Bad Visualization

Background

- 74 zettabytes (1 zettabyte = 10^{21}) of data globally in 2021 (Statista)
- Data is very important, data is new money
- Data scientist is “sexiest job of the 21st century”
- **Data exploration and visualization** = one step closer to your data

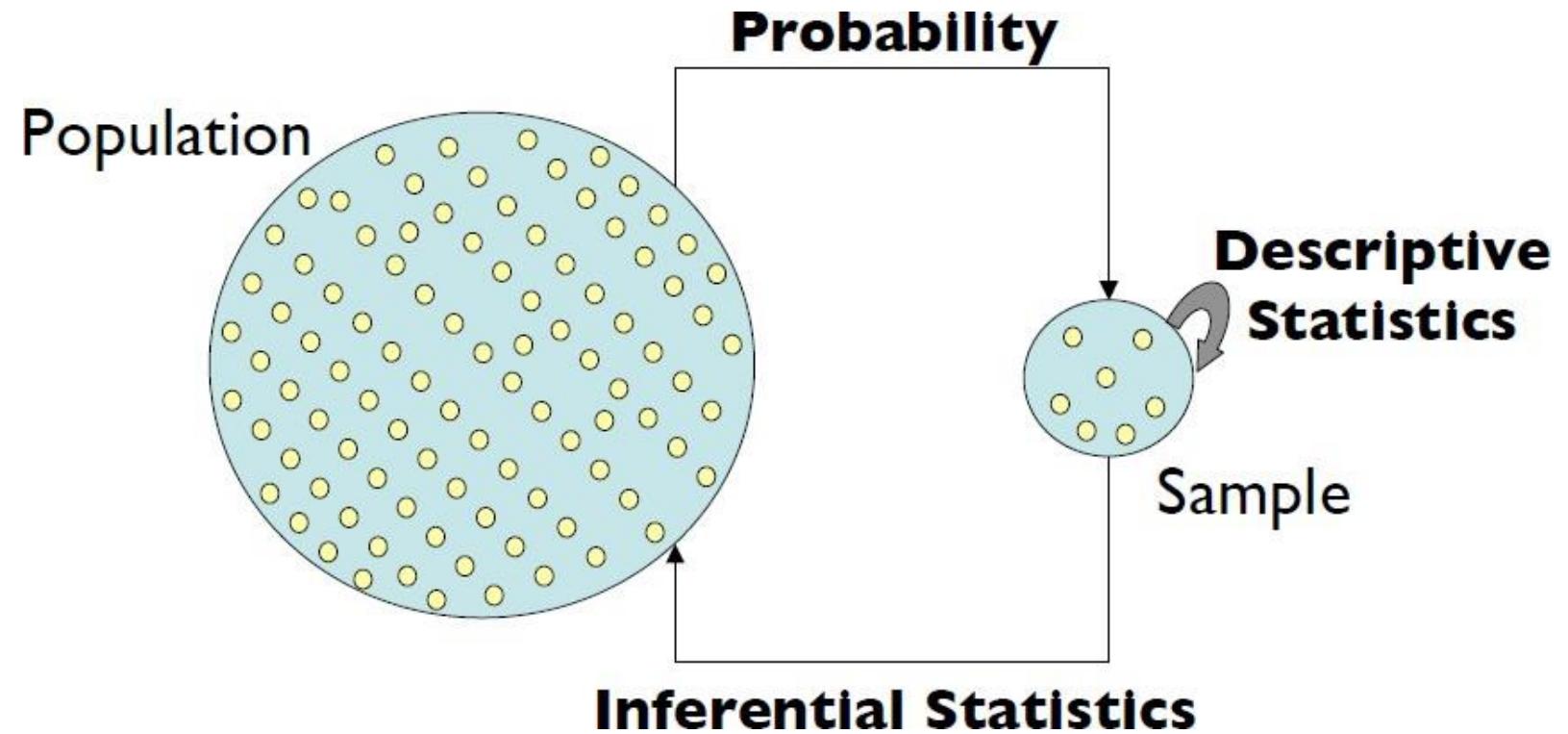
Data

- **A collection of facts** (numbers, words, measurements, observations, etc)
 1. Quantitative/Qualitative
 2. Categorical/Numerical
 3. Univariat/Bivariat/Multivariat



Data Collection

1. Census
2. Sampling



Dataset

- A data set (or dataset) is a **collection of data**.
- Data set refers to a file that contains **one or more records**.
- Usually presented in tabular form (**row and column**)

iris setosa



petal sepal

iris versicolor



petal sepal

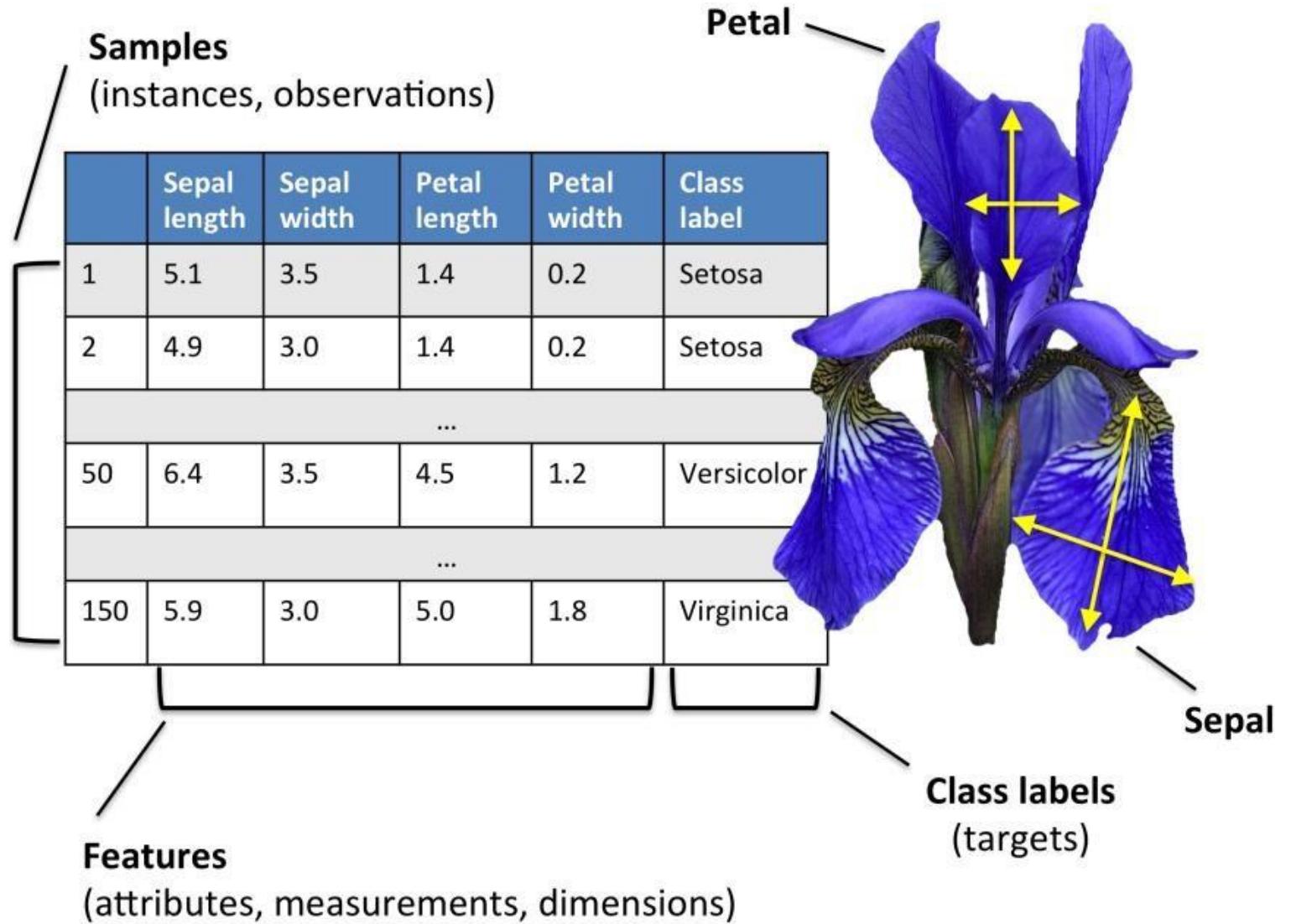
iris virginica



petal sepal

Dataset

- Contains:
 - Data object/
 - Samples/
 - Data points/
- Example:
 - Iris flower



Data Exploration

- A preliminary exploration of the data to **better understand its characteristics**.
- Related to the area of Exploratory Data Analysis (EDA)
- Focus on:
 - Summary statistics
 - Visualization
- Why?
 - Helping to **select the right tool** for preprocessing or analysis
 - Making use of humans' abilities to **recognize patterns**

Data Exploration Tasks

1. Data understanding
2. Preprocessing
 - Join, cleaning, noise, outliers, duplicate, missing value, incomplete data
3. Basic/Summary Statistics
4. Data Visualization
5. Hypotheses
6. Assumption Checking
7. Story Telling (Reporting)

4 EDA Techniques

1. Univariate non-graphical
2. Univariate graphical
3. Multivariate non-graphical
4. Multivariate graphical

Summary Statistic

- Summary statistics are **numbers that summarize properties of the data**
- Summarized properties include **frequency, location, and spread**
- Example:
 - Location - mean
 - Spread - standard deviation
 - Frequency - mode

Data Visualization

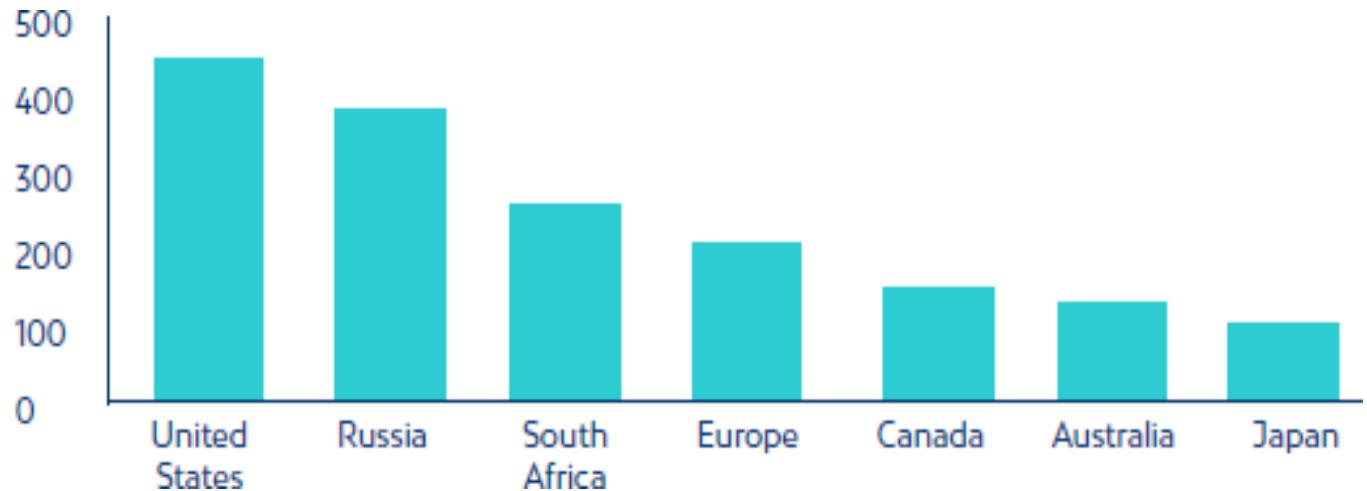
**A (Good) Picture Is
Worth A 1,000 Words**

Data Visualization

- Visualization is **the conversion of data into a visual or tabular format** so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
- Visualization of data is one of **the most powerful techniques** for data exploration.
 - Humans have a well developed ability to analyze large amounts of information that is presented visually
 - Can detect general patterns and trends
 - Can detect outliers and unusual patterns

Data Visualization

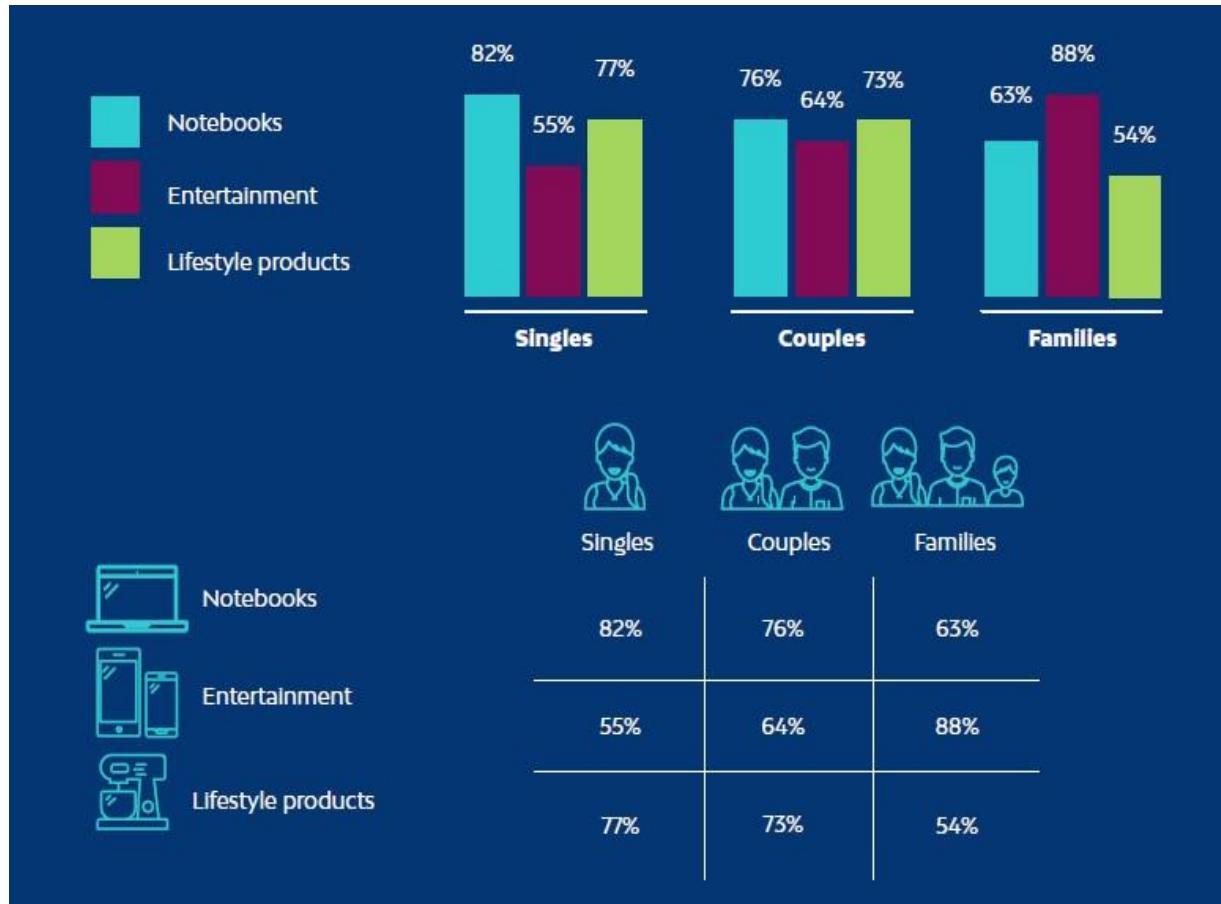
- Main Goal of Data Visualization:
 - Explaining
 - Exploring
 - Analyzing



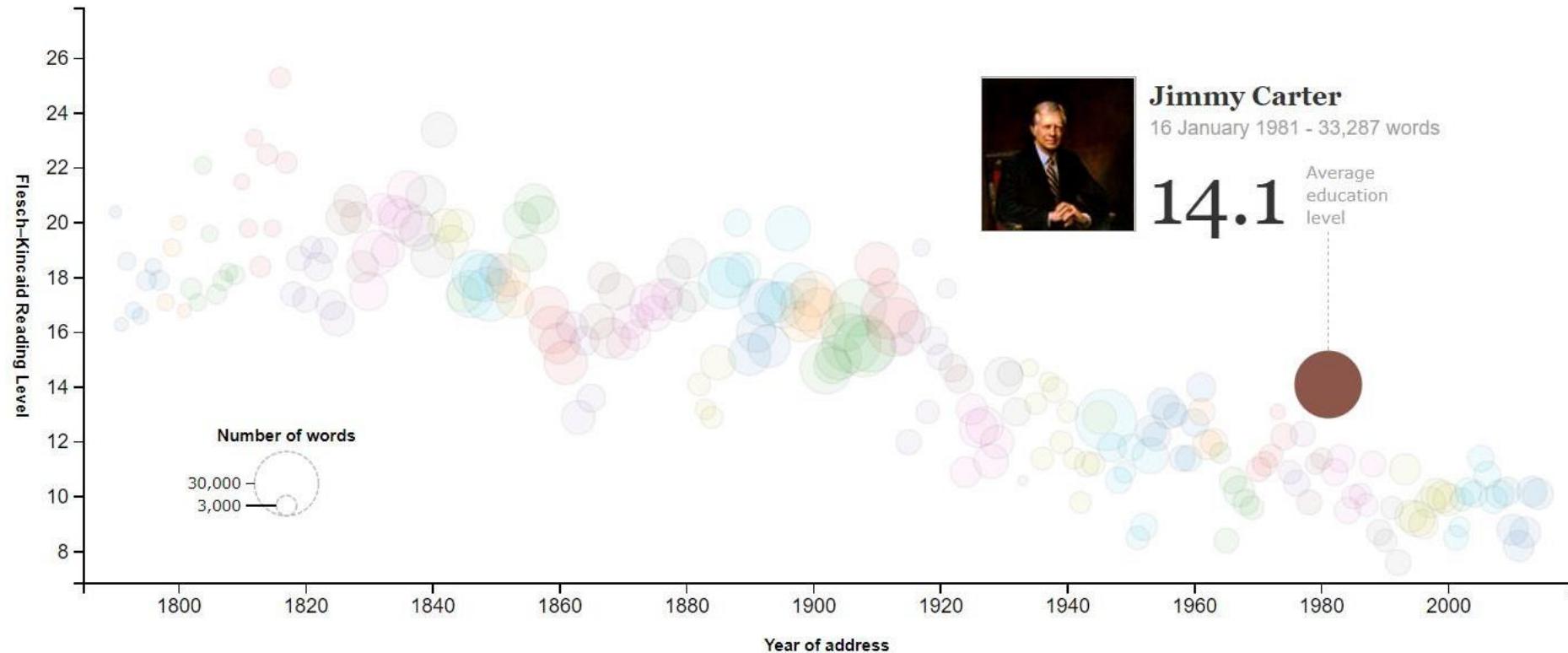
Good Visualization

- Display data accurately and clearly
 - Layout and design
 - Visual variables dan semantics
 - Consistent - colors
 - Simple icon and symbols

Good Visualization



Good Visualization

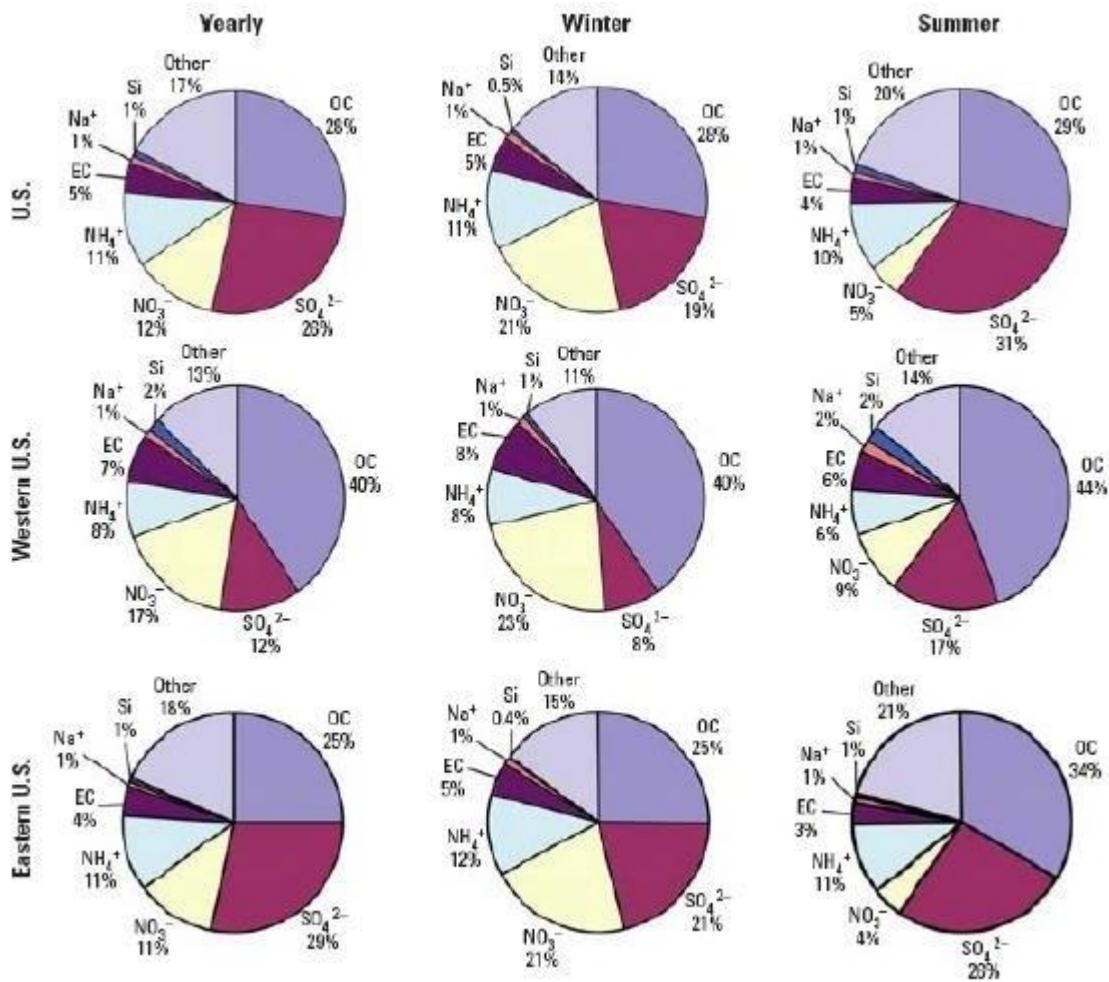
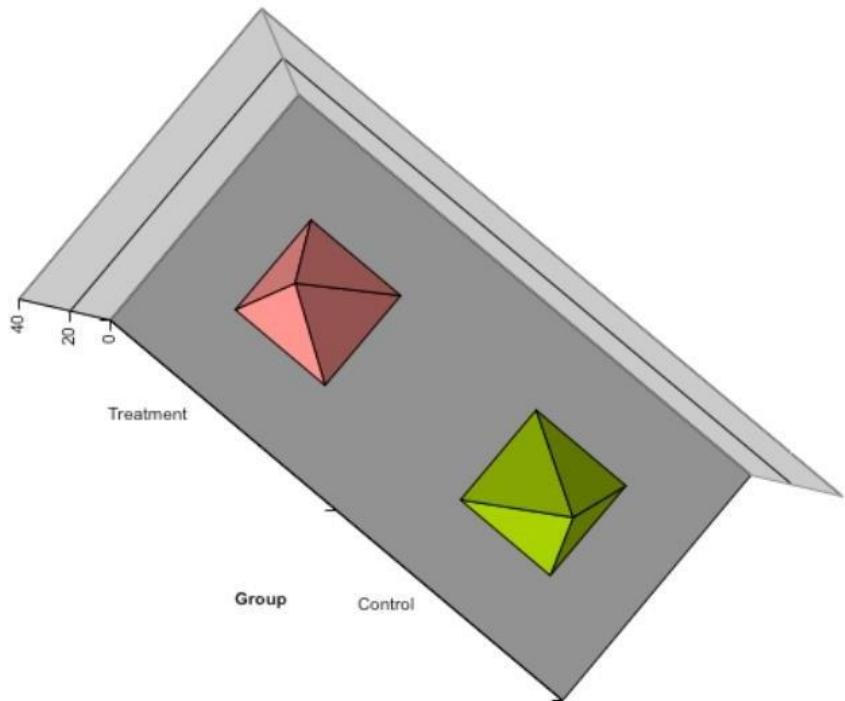


<https://www.theguardian.com/world/interactive/2013/feb/12/state-of-the-union-reading-level>

Bad Visualization

- Display as little/much information as possible
- Obscure what you do show (with chart junk)
- Use pseudo-3d and color gratuitously
- Make a pie chart (preferably in color and 3d)
- Use a poorly chosen scale

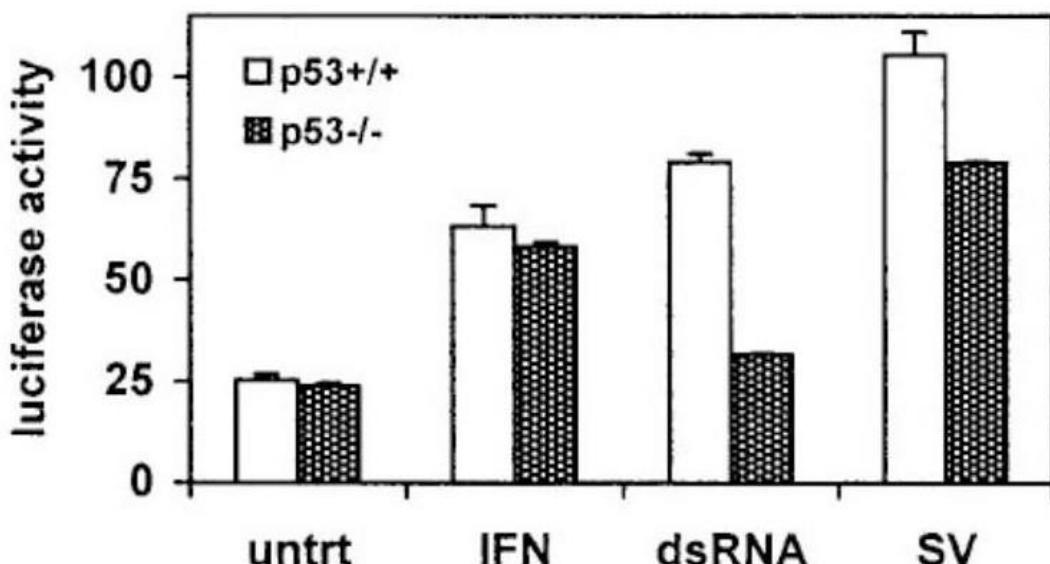
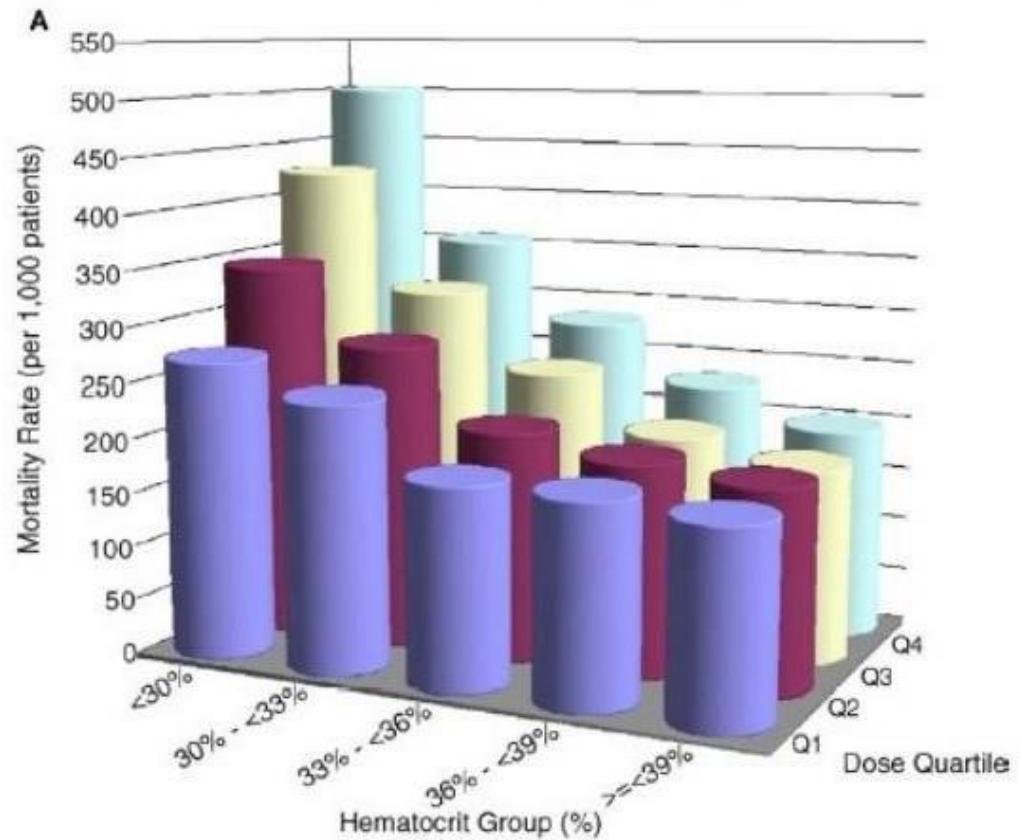
Bad Visualization



From Karl Broman: <http://www.biostat.wisc.edu/~kbroman/>

Bad Visualization

D.J. Cotter et al. / Journal of Clinical Epidemiology 57 (2004) 1086–1095



From Karl Broman: <http://www.biostat.wisc.edu/~kbroman/>

Data Exploration Tools

- Python
- R
- Etc..



Thank You

Eksplorasi dan Visualisasi Data

Pertemuan 2.1:
Pengantar R dan RStudio

Pembahasan

1. Tata cara unduh dan install R dan RStudio
2. Menulis dan membaca data dengan Bahasa R
3. Struktur percabangan dan perulangan
4. Tipe data (vector, matrices, array, data frame, list, factor)
5. Operator aritmatika dan operator logika pada R
6. Pengecakan dan transformasi tipe data

Bahasa R

- Bahasa dan lingkungan untuk komputasi statistik serta grafik
- Lingkungan yang cocok untuk data science
- Program R berupa **interpreter** yang ditulis secara baris per baris seperti program pada Command Line Interface (CLI) atau Shell
- Perangkat lunak khusus untuk manipulasi data, perhitungan matematis, dan tampilan grafis

Fasilitas R

1. Mendukung fasilitas penanganan dan penyimpanan data yang efektif.
2. Menyediakan seperangkat operator untuk manipuasi array, khususnya matriks.
3. Efektif dalam melakukan pekerjaan terkait analisis data, visualisasi data, data science, dan machine learning.
4. Menyediakan banyak teknik statistik (seperti uji statistik, klasifikasi, pengelompokan dan reduksi data).
5. Sangat mudah untuk menggambar grafik menggunakan Bahasa R.
6. Dapat bekerja pada platform yang berbeda (Windows, Mac, Linux).
7. Bahasa pemrograman yang dikembangkan dengan baik, sederhana dan efektif yang mencakup kondisional, perulangan, fungsi rekursif, dan mendukung proses input dan output.

Kelebihan R

1. Gratis dan **Open-Source Software** (OSS).
2. Ketersediaan **package** (perpustakaan fungsi/library) yang lengkap.
3. Fungsi-fungsi yang disediakan memudahkan pekerjaan yang berhubungan dengan **komputasi statistika, eksplorasi, dan visualisasi data**.
4. Memudahkan dalam mempersiapkan data, analisis data, transformasi data, dan mendukung proses import/eksport dalam berbagai format
5. Menghasilkan grafik/visualisasi yang sangat bagus dengan dukungan paket-paket yang lengkap untuk visualisasi data.
6. Memudahkan dalam pembuatan laporan

R Comment

Komentar di R menggunakan kode #.

```
# This is a comment
"Hello World!"
"Hello World!" # This is a comment
# This is a comment
# written in
# more than just one line
"Hello World!"
```

Variabel R

1. R tidak memiliki perintah khusus untuk mendeklarasikan suatu variabel.
2. Variabel dibuat pada saat anda pertama kali menetapkan nilai.
3. Untuk menetapkan nilai ke variabel, anda dapat menggunakan simbol <-.
4. Untuk menampilkan atau mencetak nilai suatu variabel, cukup dengan mengetikkan nama variabelnya atau dengan perintah print().

```
name <- "Agus"
age <- 40
name # output "Agus"
age # output 40

# Assign the same value to multiple variables in one line
var1 <- var2 <- var3 <- "Orange"
```

Variabel R

1. Nama variabel harus dimulai dengan huruf dan dapat berupa kombinasi huruf, angka, titik (.) dan garis bawah (_).
2. Jika dimulai dengan titik (.) tidak boleh diikuti oleh angka.
3. Nama variabel tidak boleh diawali dengan angka atau garis bawah (_)
4. Nama variabel peka terhadap huruf besar/kecil atau case sensitive (usia, Usia, dan AGE adalah tiga variabel berbeda)
5. Kata-kata default tidak dapat digunakan sebagai variabel contoh: TRUE, FALSE, NULL, if, dsb.

Variabel R

```
# Legal variable names:  
myvar <- "Agus"  
my_var <- "Agus "  
myVar <- "Agus"  
MYVAR <- "Agus"  
myvar2 <- "Agus"  
.myvar <- Agus  
  
# Illegal variable names:  
2myvar <- "Agus"  
my-var <- "Agus"  
my var <- "Agus"  
_my_var <- "Agus"  
my_v@ar <- "Agus"  
TRUE <- "Agus"
```

Tipe Data R

- a. Numerik (10.5, 55, 787)
- b. Integer (1L, 55L, 100L, dimana huruf "L" menyatakan bilangan bulat)
- c. Kompleks (9 + 3i, dimana "i" adalah bagian imajiner)
- d. Karakter (string) diapit oleh "" ("k", "R menarik", "false", "11.5")
- e. Logis (boolean) (BENAR atau SALAH)

R Math: Operasi Matematika

- a. Fungsi **min()** dan **max()** dapat digunakan untuk menentukan nilai terendah atau tertinggi.
- b. Fungsi **sqrt()** untuk mengembalikan akar kuadrat dari suatu angka.
- c. Fungsi **abs()** mengembalikan nilai absolut (positif) dari suatu angka.
- d. Fungsi **ceiling()** membulatkan angka ke atas bilangan bulat terdekat, dan fungsi **floor()** membulatkan angka ke bawah bilangan bulat terdekat.
- e. Dll..

String R

- Karakter atau string digunakan untuk menyimpan suatu nilai teks.
- Sebuah string dikelilingi oleh tanda kutip tunggal atau tanda kutip ganda

```
str <- "Hello"  
str <- "Lorem ipsum dolor sit amet,  
consectetur adipiscing elit."  
  
str # print the value of str  
cat(str)  
nchar(str) #jumlah karakter  
  
str1 <- "Hello"  
str2 <- "World"  
paste(str1, str2) #menggabungkan dua string
```

Operator

A. Operator Matematika

Operator	Nama	Contoh
+	Penjumlahan	$x + y$
-	Pengurangan	$x - y$
*	Perkalian	$x * y$
/	Pembagian	x / y
^	Eksponen (pangkat)	$x ^ y$
%%	Modulus (sisa pembagian)	$x \% \% y$

Operator

B. Operator Perbandingan

Operator	Nama	Contoh
<code>==</code>	Sama dengan	$x == y$
<code>!=</code>	Tidak sama dengan	$x != y$
<code>></code>	Lebih besar dari	$x > y$
<code><</code>	Lebih kecil dari	$x < y$
<code>>=</code>	Lebih besar sama dengan	$x >= y$
<code><=</code>	Lebih kecil sama dengan	$x <= y$

Operator

C. Operator Logika

Operator	Deskripsi
&	Operator AND mengembalikan nilai TRUE jika kedua elemen TRUE
atau	Operator OR mengembalikan nilai TRUE jika salah satu pernyataannya BENAR
!	NOT mengembalikan nilai FALSE jika pernyataan awal bernilai TRUE dan mengembalikan nilai TRUE jika pernyataan awal bernilai FALSE (negasi)

D. Operator Penugasan (<-) untuk menetapkan nilai ke suatu variabel

Percabangan (Kondisi)

- Percabangan adalah suatu perintah (pernyataan) yang memungkinkan suatu perintah (pernyataan) dieksekusi jika suatu kondisi terpenuhi atau tidak terpenuhi.
- Percabangan menggunakan operator kondisional yang akan menghasilkan nilai boolean (benar/true atau salah/false)
- Fungsi: program control, mengatur alur jalannya program sesuai dengan suatu kondisi yang terpenuhi

```
# contoh percabangan sederhana (if) saja
a <- 33
b <- 200

if (b > a) {
  print("b is greater than a")
}
```

Percabangan (Kondisi)

```
# contoh percabangan (if else)
a <- 33
b <- 33
if (b > a) {
  print("b is greater than a")
} else if (a == b) {
  print("a and b are equal")
}
```

```
# percabangan bertingkat
x <- 41
if (x > 10) {
  print("Above ten")
  if (x > 20) {
    print("and also above 20!")
  } else {
    print("but not above 20.")
  }
} else {
  print("below 10.")
}
```

Perulangan (Looping)

- Proses dalam pemrograman yang dilakukan secara berulang-ulang dalam batas yang telah ditentukan
- Fungsi:
 - Mengulang beberapa baris perintah
 - Melakukan suatu proses yang berulang-ulang, seperti mencetak angka dari 1 – 100
- R memiliki dua perintah loop:
 - While loop
 - For loop

While Loop

- Dengan while loop kita dapat mengeksekusi satu set pernyataan selama kondisinya TRUE.

```
i <- 1

while (i < 6) {
    print(i)
    i <- i + 1
}
```

- Pada contoh di atas, loop akan terus menghasilkan angka mulai dari 1 sampai 5. Loop akan berhenti di angka 6 karena $6 < 6$ adalah FALSE.
- Loop while membutuhkan variabel pengindeksan misalnya i, yang kita atur nilainya 1.

For Loop

- Perulangan for digunakan untuk mengulangi suatu nilai dalam urutan (sequence).

```
for (x in 1:10) {  
  print(x)  
}
```

- Output:
 - 1
 - 2
 - 3
 - 4
 - ...
 - 10

Struktur Data R

1. **Vector:** daftar item (lists) yang bertipe sama
2. **Lists:** dapat berisi banyak tipe data yang berbeda
3. **Matrices:** dataset dua dimensi dengan kolom dan baris.
4. **Arrays:** dapat memiliki lebih dari dua dimensi
5. **Data Frame:** data yang ditampilkan dalam format table dan dapat memiliki berbagai jenis data di dalamnya
6. **Factors:** digunakan untuk mengkategorikan suatu data

Vector

```
# Vector of strings
fruits <- c("banana", "apple", "orange")

# Print fruits
fruits

# Vector length
length(fruits)
# Access the first item (banana)
fruits[1]

# Vector of numerical values
numbers <- c(1, 2, 3)

# Print numbers
numbers
numbers <- 1:10
numbers
sort(fruits) # Sort a string
sort(numbers) # Sort numbers
```

List

```
# List of strings
thislist <- list("apple", "banana", "cherry")

# Print the list
thislist

# Access list
thislist[1]

# Update value
thislist[1] <- "blackcurrant"

# list length
length(thislist)

# check item lists
"apple" %in% thislist

# add item
append(thislist, "orange")
```

Matrices

```
# Create a matrix
thismatrix <- matrix(c(1,2,3,4,5,6), nrow = 3, ncol = 2)

# Print the matrix
thismatrix

thismatrix <- matrix(c("apple", "banana", "cherry", "orange"), nrow = 2,
ncol = 2)

thismatrix

# access item
thismatrix[1, 2]

# matrix length
length(thismatrix)
```

Arrays

```
# An array with one dimension with values ranging from 1 to 24
thisarray <- c(1:24)
thisarray

# An array with more than one dimension
multiarray <- array(thisarray, dim = c(4, 3, 2))
multiarray

# Access item array
thisarray <- c(1:24)
multiarray <- array(thisarray, dim = c(4, 3, 2))

multiarray[2, 3, 2]

# looping in array
thisarray <- c(1:24)
multiarray <- array(thisarray, dim = c(4, 3, 2))

for(x in multiarray) {
  print(x)
}
```

Data Frame

```
# Create a data frame
Data_Frame <- data.frame (
  Training = c("Strength", "Stamina", "Other"),
  Pulse = c(100, 150, 120),
  Duration = c(60, 30, 45)
)

# Print the data frame
Data_Frame

# Summarize data
summary(Data_Frame)

# Access item
Data_Frame[1]
Data_Frame[["Training"]]
```

Factors

- Faktor digunakan untuk mengkategorikan suatu data. Contoh faktor adalah:
 1. Demografi: Pria/Wanita
 2. Musik: Rock, Pop, Klasik, Jazz
 3. Pelatihan: Kekuatan, Stamina
- Untuk membuat faktor, gunakan fungsi factor() dan tambahkan vektor sebagai argumennya.

```
# Create a factor
music_genre <- factor(c("Jazz", "Rock", "Classic", "Classic",
"Pop", "Jazz", "Rock", "Jazz"))

# Print the factor
music_genre
```

Factors

```
# Print level
levels(music_genre)

# Factor length
length(music_genre)

# Access factor
music_genre[3]

# Change item value
music_genre[3] <- "Pop"
music_genre[3]
```

Tugas Praktikum

1. Buatlah program sederhana yang berisi beberapa variabel kemudian cetak menggunakan perintah *print()*.
2. Buatlah program sederhana dengan mengimplementasikan struktur percabangan.
3. Buatlah program sederhana dengan mengimplementasikan struktur perulangan.
4. Buatlah struktur data (*R Vectors, Matrices, Arrays, Lists, Data Frame, dan Factors*).



Thank You

Eksplorasi dan Visualisasi Data

Pertemuan 3:
Principles of Analytics Graphics

Outline

1. Show comparisons
2. Show causality, mechanism, explanation, systematic structure
3. Show multivariate data
4. Integrate evidence
5. Describe and document the evidence
6. Content, Content, Content

Introduction

- Pembahasan pada pertemuan ini diambil dari buku Exploratory Data Analysis with R karya Roger D Peng yang terinspirasi dari buku Beautiful Evidence karya Edward Tufte.
- Enam prinsip yang penting untuk membuat grafik data yang informatif dan berguna, diantaranya *Comparisons; Causality, Mechanism, Explanation, Systematic Structure; Data multivariat; Integrate Evidence; Describe and document the evidence;* dan **Konten**. Beberapa dari prinsip-prinsip ini mungkin lebih relevan untuk membuat grafik “final” dibandingkan dengan grafik yang lebih “eksplorasi”.
- **Grafik/ plot eksplorasi** biasanya dibuat dengan cepat dan tujuannya adalah untuk memungkinkan Anda *meringkas data dan menyoroti fitur-fitur umum* serta mengeksplorasi pertanyaan-pertanyaan dasar tentang data dan untuk menilai bukti yang mendukung atau menentang hipotesis tertentu. Pada akhirnya, mereka mungkin berguna untuk menyarankan strategi pemodelan yang dapat digunakan pada “langkah selanjutnya” dari proses analisis data.

Comparisons

- Menunjukkan perbandingan sebenarnya merupakan dasar dari penyelidikan ilmiah yang baik. Bukti suatu hipotesis selalu relatif terhadap hipotesis lain yang bersaing, misalnya “bukti mendukung hipotesis A versus hipotesis B”.
- Ketika dihadapkan dengan klaim atau pernyataan ilmiah, ilmuwan yang baik selalu bertanya **“Dibandingkan dengan Apa?”**.
- Jika penyajian visual (grafik) ingin membantu berpikir, Grafik data umumnya harus mengikuti prinsip yang sama, yaitu menunjukkan perbandingan.

Comparisons (2)

- Sebuah penelitian dari Fakultas Kedokteran Universitas Johns Hopkins dilakukan di rumah tempat tinggal seorang perokok minimal 4 hari dalam seminggu. Di rumah dipasang alat pembersih udara (*air cleaner*). Setiap anak dinilai pada awal dan kemudian 6 bulan kemudian pada kunjungan kedua.
- Tujuannya adalah untuk **meningkatkan hari bebas gejala** anak selama periode 6 bulan sehingga **angka yang lebih tinggi lebih baik** artinya mereka memiliki lebih banyak hari bebas gejala.
- Plot ini menunjukkan perubahan hari bebas gejala pada 47 anak yang terdaftar dalam uji klinis yang **menguji apakah alat pembersih udara dapat memperbaiki gejala terkait asma mereka**. Rata-rata jumlah hari bebas gejala bertambah sekitar 1 hari.

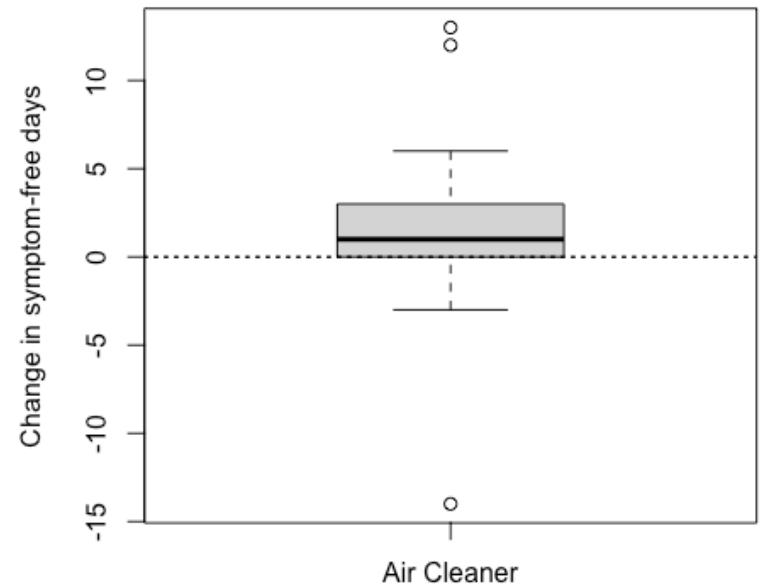


Fig. 1 Change in symptom-free days with air cleaner

Peng, R. D., (2015), *Exploratory Data Analysis with R*, Lean Publishing.

Plot ini tidak menjawab pertanyaan
“dibandingkan dengan apa?”

Comparisons (3)

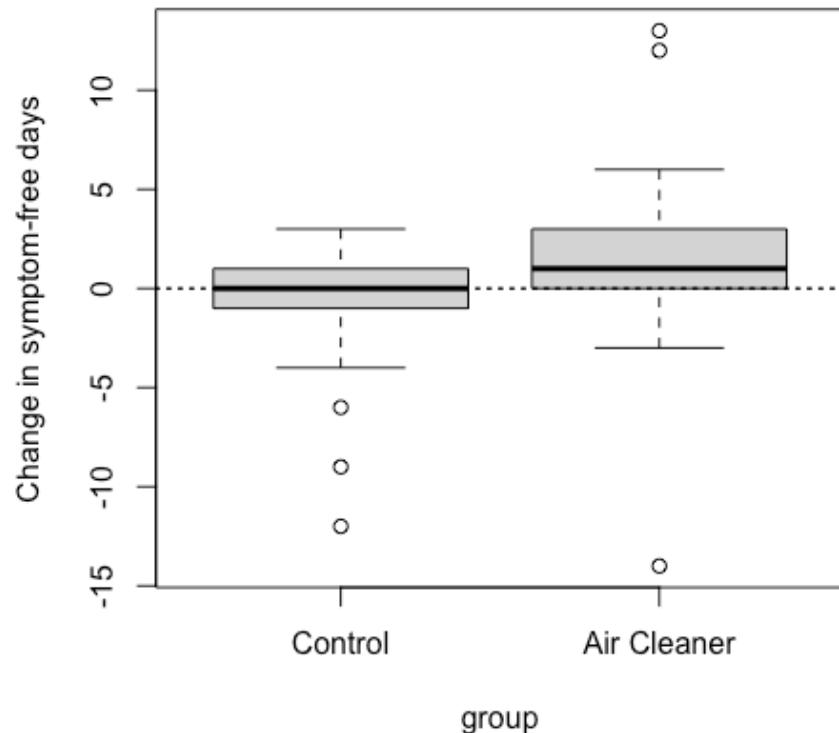


Fig. 2 Change in symptom-free days by treatment group

- Secara khusus, kita tidak tahu apa yang akan terjadi jika anak-anak tidak menerima pembersih udara.
- Peneliti memiliki data tersebut dan dapat menunjukkan kelompok yang menerima pembersih udara dan kelompok kontrol yang tidak.
- Rata-rata anak-anak pada kelompok kontrol hanya mengalami sedikit perubahan dalam hal hari bebas gejala dibandingkan dengan anak-anak yang menerima pembersih udara.

Causality, Mechanism, Explanation, Systematic Structure

- Jika memungkinkan, selalu berguna untuk menunjukkan **kerangka sebab akibat** Anda dalam memikirkan suatu pertanyaan. Seringkali alasan kita mengkaji bukti adalah untuk memahami kausalitas, mekanisme, dinamika, proses, atau struktur sistematis.
- Secara umum, sulit untuk membuktikan bahwa satu hal menyebabkan hal lain bahkan dengan data yang dikumpulkan dengan sangat cermat. Namun, grafik data Anda seringkali berguna untuk **menunjukkan apa yang Anda pikirkan** sehubungan dengan penyebabnya.
- Tampilan seperti ini mungkin dapat memberikan hipotesis atau membantahnya, tetapi yang terpenting, hal tersebut akan memunculkan pertanyaan baru yang dapat ditindaklanjuti dengan data atau analisis baru.



Analisis medis (meliputi pencegahan, diagnosis, intervensi) memerlukan analisis sebab akibat.



Penalaran tentang reformasi dan pengambilan keputusan juga memerlukan logika kausal.

Causality, Mechanism, Explanation, Systematic Structure (2)

Plot Fig. 2 menunjukkan perubahan hari bebas gejala untuk sekelompok anak yang menerima pembersih udara dan sekelompok anak yang tidak menerima, yaitu lebih banyak hari bebas gejala bagi anak yang menerima pembersih udara.

Pertanyaan yang menarik mungkin adalah “[Mengapa anak-anak yang menggunakan alat pembersih udara mengalami peningkatan?](#)” Ini mungkin bukan pertanyaan yang paling penting (Anda mungkin peduli bahwa pembersih udara dapat membantu), tetapi menjawab pertanyaan “mengapa?” mungkin mengarah pada perbaikan atau perkembangan baru.

Causality, Mechanism, Explanation, Systematic Structure (3)

Hipotesis yang dimiliki adalah pembersih udara menghilangkan partikel-partikel udara dari udara. Mengingat semua rumah dalam penelitian ini dihuni oleh perokok, kemungkinan besar terdapat partikel dalam jumlah besar di udara.

Menghirup partikel halus dapat memperburuk gejala asma sehingga masuk akal bahwa mengurangi kehadiran di udara akan memperbaiki gejala asma. Oleh karena itu, kami memperkirakan kelompok yang menerima pembersih udara rata-rata akan mengalami penurunan partikel di udara. Dalam hal ini kami melacak materi partikulat halus, juga disebut PM2.5 yang merupakan singkatan dari materi partikulat yang diameter aerodinamisnya kurang dari atau sama dengan 2,5 mikron.

Causality, Mechanism, Explanation, Systematic Structure (4)

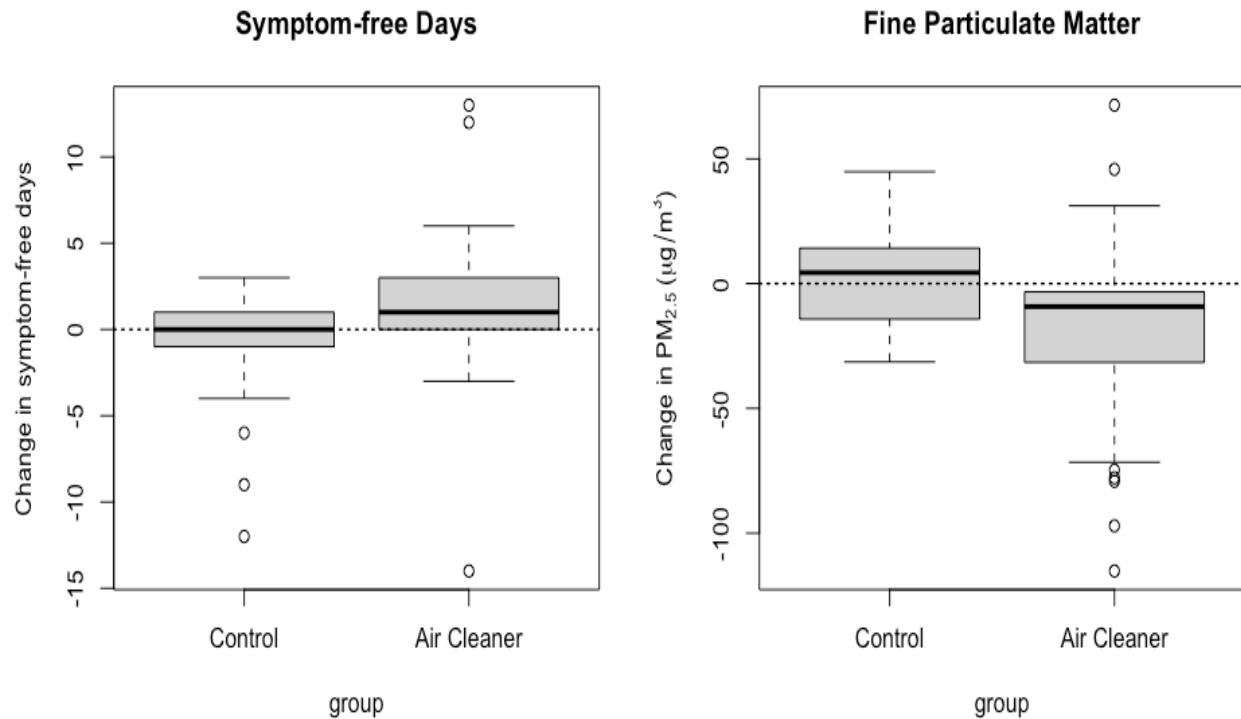
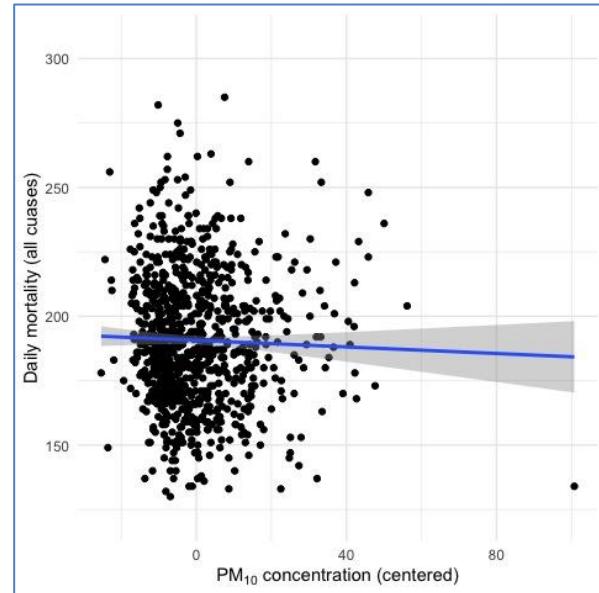
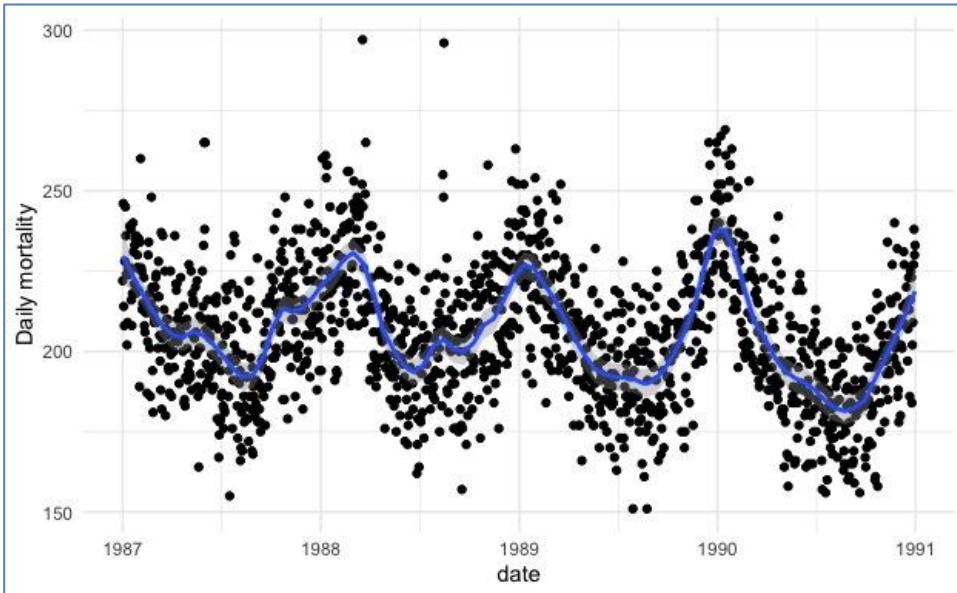


Fig. 3 Change in symptom-free days and change in PM_{2.5} levels in-home

- Dari plot, terlihat **perubahan hari bebas gejala** pada kedua kelompok (kiri) dan perubahan PM_{2.5} pada kedua kelompok (kanan).
- Dari plot sebelah kanan bahwa rata-rata pada kelompok kontrol, kadar PM_{2.5} justru meningkat sedikit, sedangkan pada **kelompok pembersih udara rata-rata kadarnya menurun**. Pola ini konsisten dengan gagasan bahwa pembersih udara meningkatkan kesehatan dengan mengurangi partikel di udara.
- Namun, hal ini **bukanlah bukti konklusif** dari gagasan ini karena **mungkin terdapat faktor perancu lain** yang tidak terukur yang dapat menurunkan tingkat PM_{2.5} dan meningkatkan hari bebas gejala.

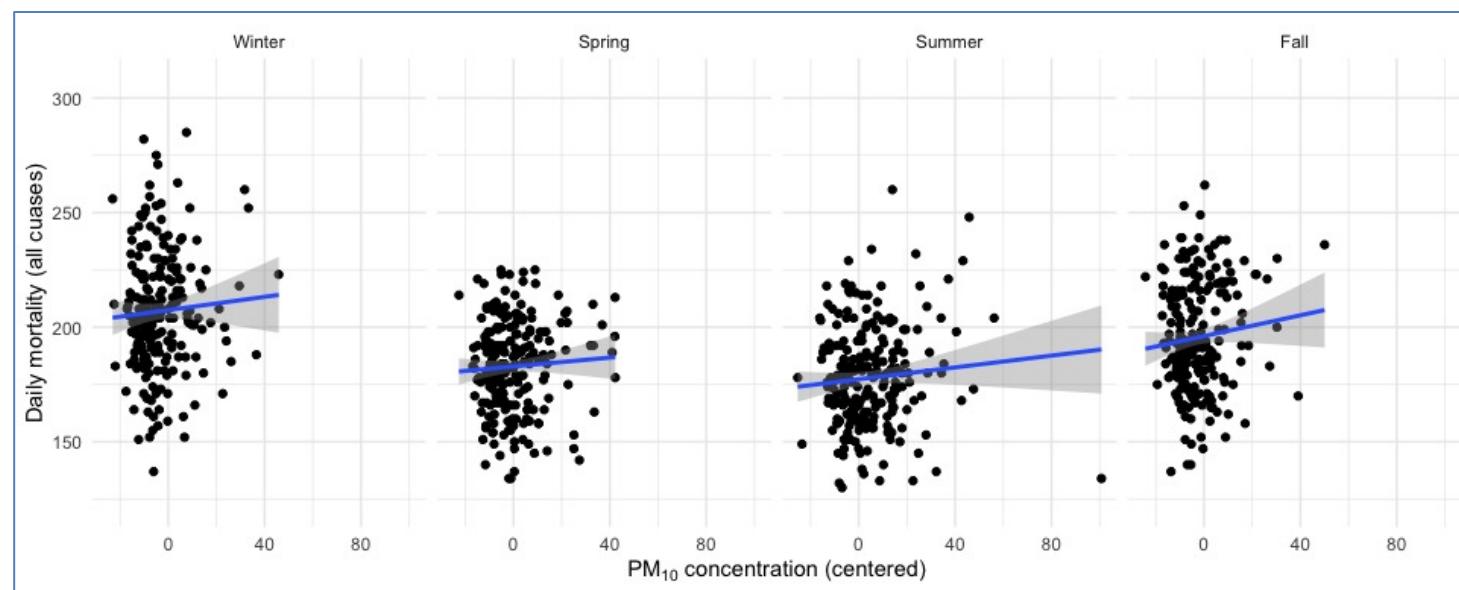
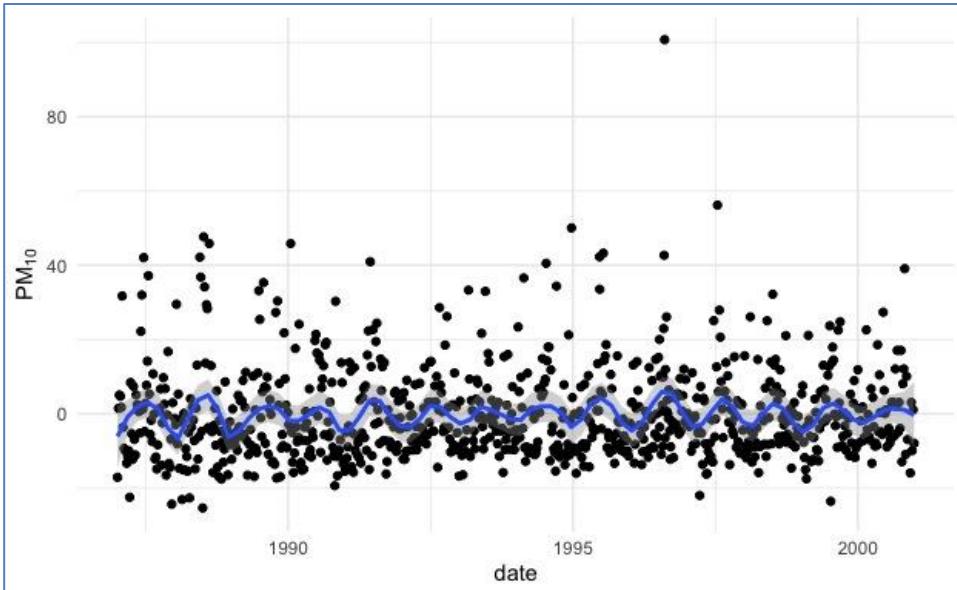
Multivariate Data

- Biasanya ada banyak atribut yang bisa Anda ukur untuk apapun yang mungkin Anda pelajari. Grafik data harus berusaha menampilkan informasi ini sebanyak mungkin secara komprehensif.
- Di samping ini adalah data tentang materi partikulat udara harian (“PM10”) di New York City dan angka kematian dari tahun 1987 hingga 2000. Setiap titik pada plot mewakili rata-rata tingkat PM10 pada hari itu (diukur dalam mikrogram per meter kubik) dan jumlah kematian pada hari itu. Data PM10 berasal dari Badan Perlindungan Lingkungan AS dan data kematian berasal dari Pusat Statistik Kesehatan Nasional AS. Namun, ada faktor lain yang berhubungan dengan angka kematian dan tingkat PM10. Salah satu contohnya adalah musim.



terdapat sedikit hubungan negatif antara kedua variabel

angka kematian cenderung lebih tinggi di musim dingin dibandingkan di musim panas



Peng, R. D., (2015), *Exploratory Data Analysis with R*, Lean Publishing.

Dalam beberapa kasus, Anda mungkin menemukan hubungan yang tidak terduga tergantung pada bagaimana hubungan tersebut diplot atau divisualisasikan.

Integrate Evidence

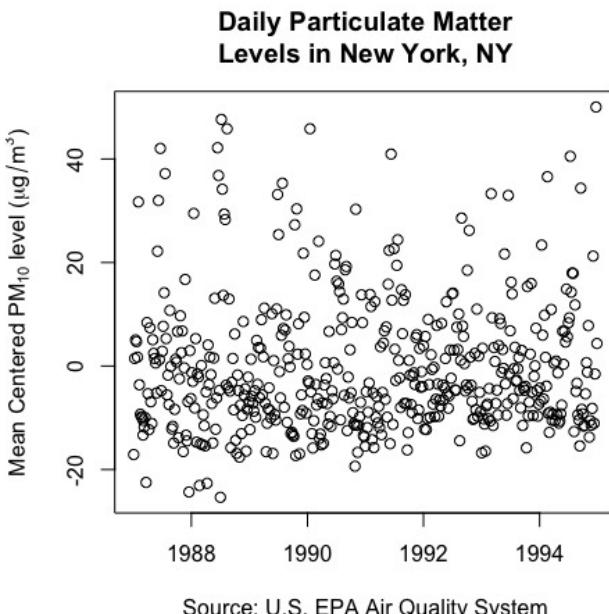
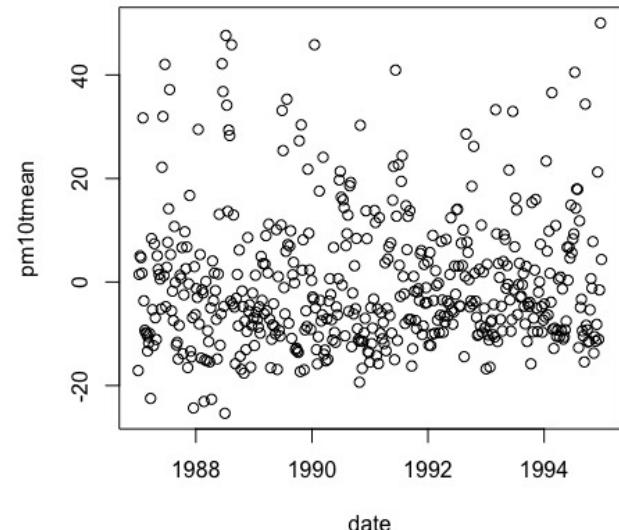
- Hanya karena Anda mungkin membuat grafik data, bukan berarti Anda harus hanya mengandalkan lingkaran dan garis untuk menyampaikan maksud Anda.
- Anda juga dapat memasukkan angka, kata, gambar, dan diagram tercetak untuk menceritakan kisah Anda. Dengan kata lain, grafik data harus menggunakan banyak mode penyajian data secara bersamaan, bukan hanya mode yang familiar bagi Anda atau yang dapat ditangani oleh perangkat lunak.
- Seseorang tidak boleh membiarkan alat yang tersedia menggerakkan analisis; seseorang harus mengintegrasikan bukti sebanyak mungkin ke dalam grafik.

Describe and document the evidence

- Grafik data harus didokumentasikan dengan tepat beserta label, skala, dan sumber.
- Grafik data sebaiknya menceritakan kisah lengkap dengan sendirinya. Anda tidak perlu mengacu pada teks atau deskripsi tambahan saat menafsirkan plot, jika memungkinkan. Idealnya, sebuah plot memiliki semua deskripsi yang diperlukan.
- Anda mungkin berpikir bahwa tingkat dokumentasi ini sebaiknya hanya digunakan untuk plot “final” dan bukan untuk plot eksplorasi, tetapi ada baiknya Anda membiasakan diri untuk mendokumentasikan bukti-bukti Anda sesegera mungkin.
- Bayangkan jika Anda sedang menulis makalah atau laporan, dan grafik data disajikan untuk menyampaikan poin utama. Bayangkan orang yang Anda berikan kertas/laporannya mempunyai waktu yang sangat sedikit dan hanya akan fokus pada grafik. Apakah terdapat cukup informasi pada gambar tersebut agar orang tersebut dapat memahami ceritanya?
- Roger D Peng dalam bukunya menyebutkan cenderung memilih lebih banyak informasi (sangat mendetail) daripada lebih sedikit.

Describe and document the evidence (2)

- Plot di sebelah kiri adalah plot default yang dihasilkan oleh fungsi plot di R.
- Plot di sebelah kanan menggunakan fungsi plot yang sama tetapi menambahkan anotasi seperti judul, label sumbu y, label sumbu x.
- Informasi penting yang disertakan adalah tempat pengumpulan data (New York), satuan pengukuran, skala waktu pengukuran (harian), dan sumber data (EPA).



Content, Content, Content

- Presentasi analitis pada akhirnya bertahan atau gagal tergantung pada **kualitas, relevansi, dan integritas kontennya**.
- Hal ini mencakup pertanyaan yang diajukan dan bukti yang disajikan mendukung hipotesis tertentu.
- Keajaiban visualisasi atau fitur tambahan apapun tidak dapat membuat data yang buruk, atau yang lebih penting, pertanyaan yang bentuknya buruk, bersinar dengan jelas.
- Memulai dengan pertanyaan yang bagus, mengembangkan pendekatan yang tepat, dan hanya menyajikan informasi yang diperlukan untuk menjawab pertanyaan tersebut, sangatlah penting untuk setiap grafik data.



Thank You

Eksplorasi dan Visualisasi Data

Pertemuan 3.2 :
Managing Data Frame and exploratory graphs

Outline

1. Managing Data Frame (Supplement)
2. Exploratory Graphs
 - a. R built-in function
 - b. ggplot2 package

Managing Data Frame

- Data Frame adalah struktur data utama dalam statistik dan R.
- Struktur dasar data frame adalah terdapat satu observasi per baris dan setiap kolom mewakili variabel, ukuran, fitur, atau karakteristik observasi tersebut.
- R memiliki implementasi internal data frame yang mungkin paling sering Anda gunakan. Namun, ada paket di CRAN yang mengimplementasikan data frame melalui hal-hal seperti database relasional yang memungkinkan Anda mengoperasikan data frame yang sangat besar (tetapi tidak dibahas di sini).
- Mengingat pentingnya mengelola data frame, penting bagi kita untuk memiliki alat yang baik untuk menanganinya, salah satunya paket dplyr.

The dplyr Package

- Salah satu kontribusi penting dari paket dplyr adalah menyediakan “grammar” (khususnya, *verbs*) untuk manipulasi data dan untuk mengoperasikan bingkai data. Anda dapat mengomunikasikan apa yang Anda lakukan ke data frame yang dapat dipahami orang lain (dengan asumsi mereka juga mengetahui tata bahasanya).
- Grammar kunci yang disediakan oleh paket dplyr:
 - `select`: mengembalikan subset kolom data frame, menggunakan notasi fleksibel
 - `filter`: mengekstrak subset baris dari data frame berdasarkan kondisi logis
 - `arrange`: menyusun ulang baris-baris data frame
 - `rename`: mengganti nama variabel dalam data frame
 - `mutate`: menambahkan variabel/kolom baru atau mengubah variabel yang sudah ada
 - `summarise / summarise`: menghasilkan ringkasan statistik dari berbagai variabel dalam kerangka data, mungkin dalam strata
 - `%>%`: operator “pipe” digunakan untuk menghubungkan beberapa tindakan kata kerja menjadi satu *pipeline*

Practice

```

install.packages("dplyr") #from CRAN
install_github("hadley/dplyr") #from GitHub
library(dplyr) # load dplyr into your R session
cars <- mtcars
dim(cars)
names(cars)[1:3]

#subset() function
subset <- select(cars, mpg:disp)
subset <- select(cars, -(mpg:disp))

subset <- select(cars, ends_with("t"))
str(subset)

subset <- select(cars, starts_with("d"))
str(subset)

#filter() function
cars.f <- filter(cars, mpg > 20)
str(cars.f)

summary(cars.f$mpg)
cars.f <- filter(cars, mpg > 20 & wt > 2)
select(cars.f, mpg, wt)

```

```

#arrange() function
cars <- arrange(cars, mpg) #reorder rows according to mpg
cars <- arrange(cars, desc(mpg)) #descending order

#rename() function
cars <- rename(cars, weight = wt, hpwr = hp)
head(cars[, 1:11], 3)

#mutate() function
cars <- mutate(cars, weight = weight - mean(weight, na.rm = TRUE))
head(cars)

#group_by() function
cars.vs <- group_by(cars, vs)

summarize(cars.vs, mpg)

```

Membuat Plot

Scatter Plot

Boxplot

Histogram

Plot Densitas

Dot Plot

Diagram Batang

Visualisasi dengan Fungsi Bawaan R

```
# load data
cars <- mtcars

#Scatterplot
with(cars, plot(mpg, wt))
abline(h = 2.5, lwd = 1, lty = 2)

#Histogram
hist(cars$mpg, col = "green", main = "Miles per gallon")
abline(v = 20, lwd = 2)
abline(v = median(cars$mpg), col = "magenta", lwd = 4)

#Boxplot
boxplot(cars$mpg, col = "blue")
abline(h = 17)

boxplot(mpg ~ vs, data = cars, col = "red") #multiple boxplot

#Barplot
library(dplyr)
table(cars$vs) %>% barplot(col = "wheat")
```

Visualisasi dengan Paket ggplot2

Komponen visualisasi data:

- ✓ data yang akan divisualisasikan
- ✓ estetika visualisasi
- ✓ objek geometris
- ✓ transformasi statistik

Visualisasi data menggunakan paket ggplot2 memungkinkan kita untuk menentukan komponen visualisasi secara terpisah sehingga dapat membuat visualisasi yang kompleks dan berlapis.

Paket ggplot2 memiliki dua fungsi:

1. qplot() atau quickplot() : memiliki antarmuka yang mirip dengan plot()
2. ggplot() : menyediakan tata bahasa lengkap dari antarmuka grafis.

Fungsi qplot()

- Fungsi `qplot()` dapat digunakan sebagai pintasan bagi yang sudah terbiasa menggunakan plot dasar seperti fungsi `plot()`.
- Untuk membuat jenis plot berbeda, cara pemanggilan dapat dilakukan secara konsisten. Hal ini berbeda jika menggunakan plot dasar dari R, misal menggunakan fungsi `hist()` untuk membuat histogram.

Cara pemanggilan fungsi `qplot()`

```
qplot(
  x,
  y,
  ...,
  data,
  facets = NULL,
  margins = FALSE,
  geom = "auto",
  xlim = c(NA, NA),
  ylim = c(NA, NA),
  log = "",
  main = NULL,
  xlab = NULL,
  ylab = NULL,
  asp = NA,
  stat = NULL,
  position = NULL
)
```

facets : menghasilkan beberapa plot kecil yang mewakili subset data;

margins : faset tambahan (TRUE atau FALSE);

geom : menentukan geom yang akan digambar, jika x dan y ditentukan akan terbentuk titik-titik (scatter plot) dan jika hanya x yang ditentukan akan terbentuk histogram;

xlim, ylim : batas sumbu x dan y;

log : variabel mana yang akan diubah lognya;

main : memberikan judul plot;

xlab, ylab : label sumbu x dan label sumbu y;

asp : rasio aspek y/x;

stat, position sudah tidak digunakan lagi.

Fungsi ggplot()

Pemanggilan fungsi ggplot() dilakukan untuk menyediakan dataset dan pemetaan estetika (seperti warna, ukuran, dan lokasi x dan y) dengan aes().

```
a = ggplot(data = NULL, mapping = aes())
```

menambahkan lapisan-lapisan lain dengan menambahkan geom, facet, coordinate, scale, dan theme.

```
a + geom() + facet() + coordinate() + scale() + theme()
```

- Fungsi geom() untuk mewakili titik data serta properti estetika fungsi geom untuk mewakili variabel, misal geom_point(aes(alpha, color, size))
- Fungsi facet() membagi plot menjadi subplot
- Fungsi coordinate() mengatur sistem koordinat dari plot.
- Fungsi scale() memetakan nilai data ke nilai visual sebuah estetika.
- Tema (*theme*) dapat digunakan untuk mengatur komponen plot selain data, seperti judul, latar belakang, garis kisi, font, label, dan legenda.

Instalasi dan *Import ggplot2*

```
install.packages("tidyverse")
library(ggplot2)
```

1

```
install.packages("ggplot2")
library(ggplot2)
```

2

```
install.packages("devtools")
devtools::install_github("tidyverse/ggplot2")
library(ggplot2)
```

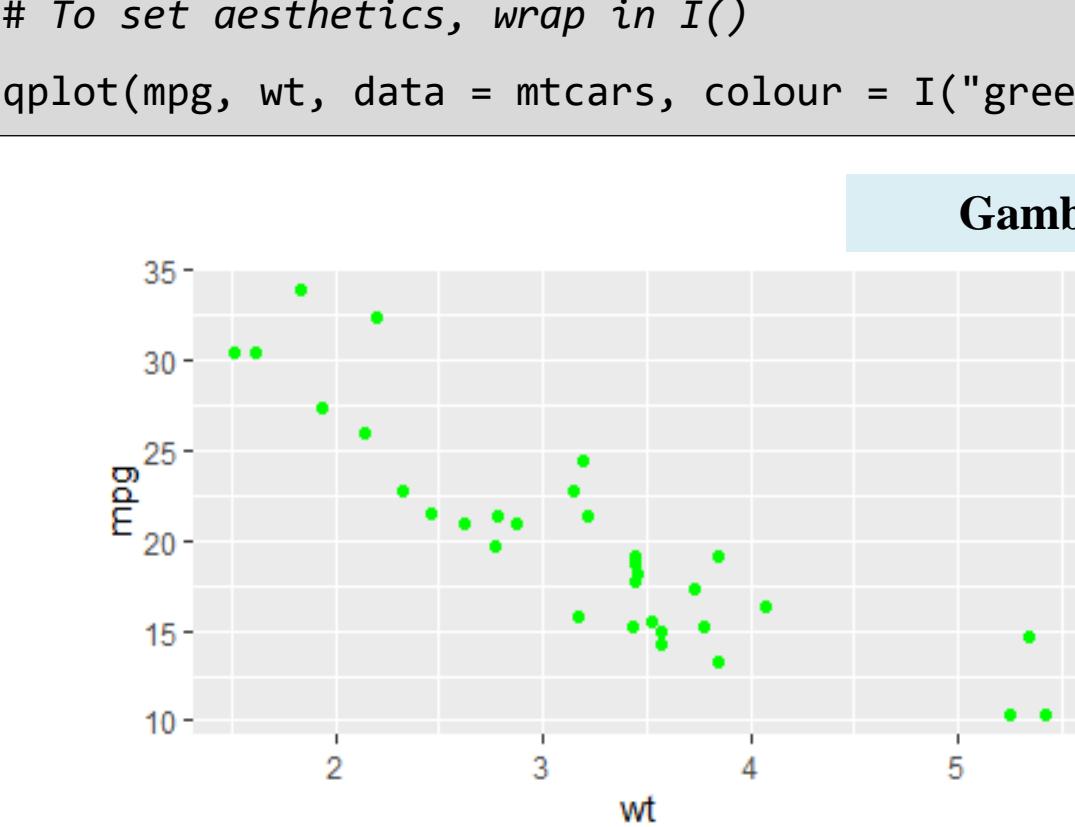
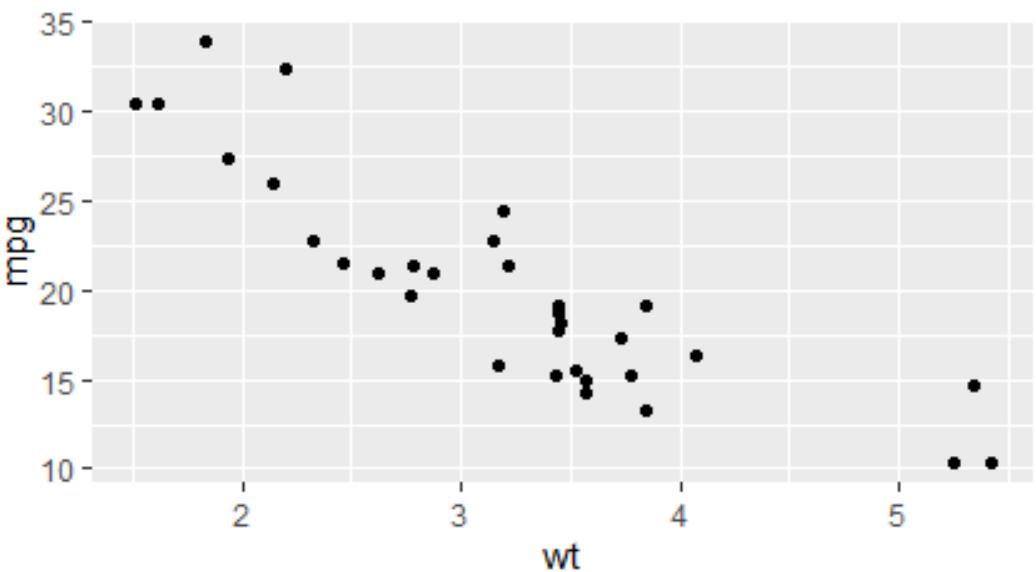
3

a.

Scatter plot antara variabel mpg (konsumsi bahan bakar) dan variabel wt (*weight*) dari data mtcars (*Motor Trend Cars Road Test*)

```
library(ggplot2)
qplot(wt, mpg, data = mtcars)
```

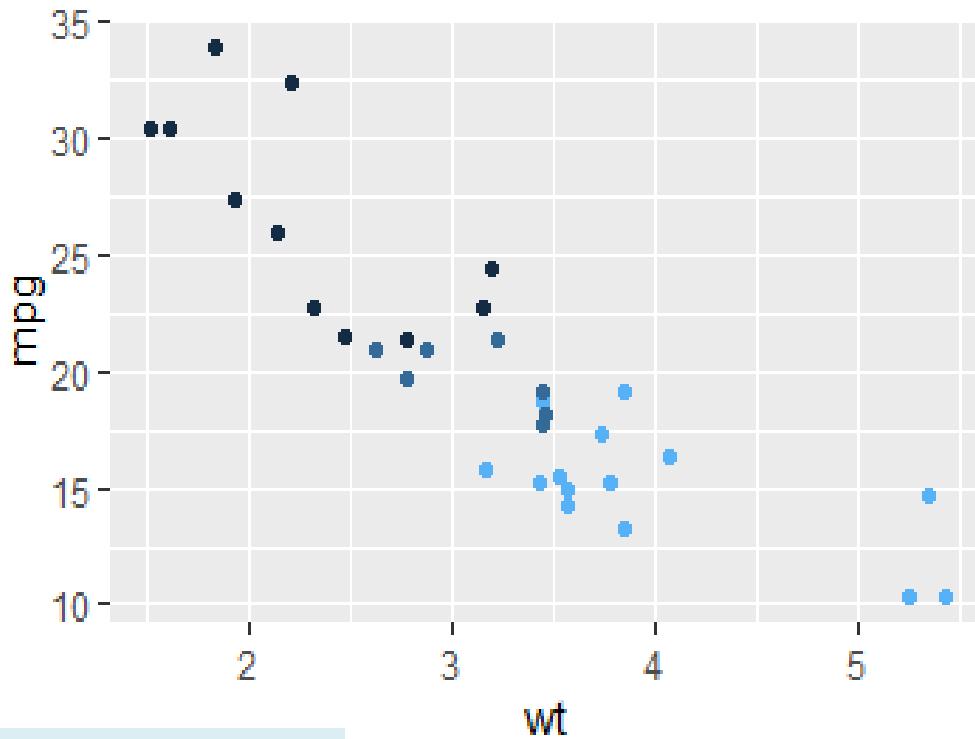
Gambar 1



a.

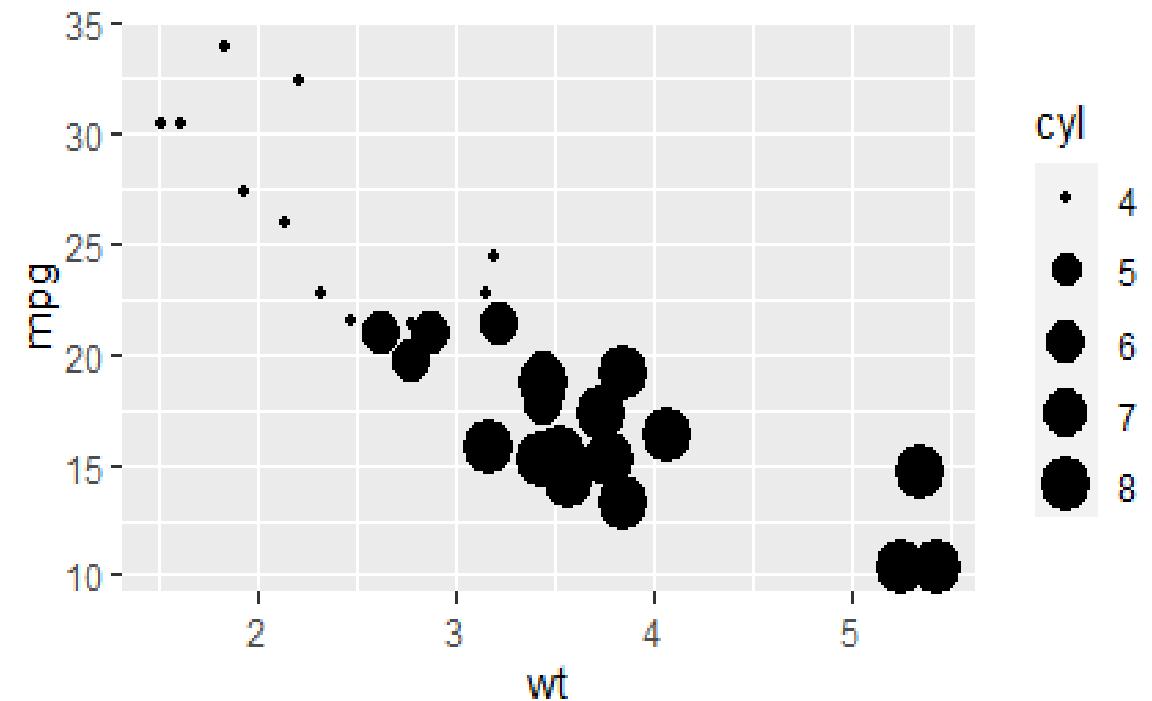
Mobil yang lebih berat cenderung memiliki silinder yang lebih banyak pula.
Konsumsi bahan bakar mobil semakin menurun seiring bertambahnya berat *Motor Trend*

```
library(ggplot2)
qplot(wt, mpg, data = mtcars, colour = cyl)
```



Gambar 3

```
library(ggplot2)
qplot(wt, mpg, data = mtcars, size = cyl)
```

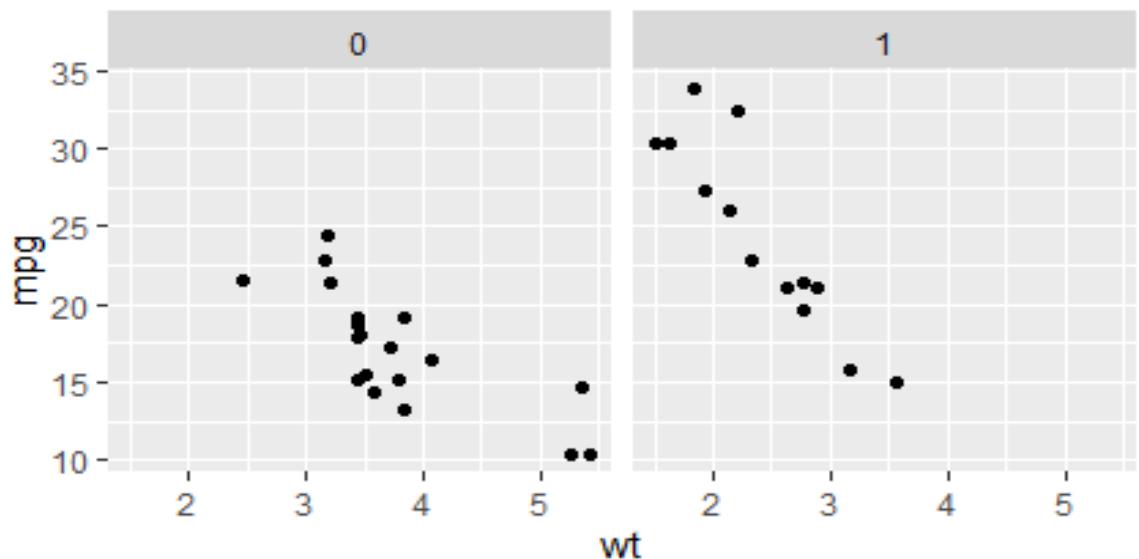


Gambar 4

a. Scatter Plot

```

library(ggplot2)
qplot(wt, mpg, data = mtcars, facets = . ~ am)
  
```



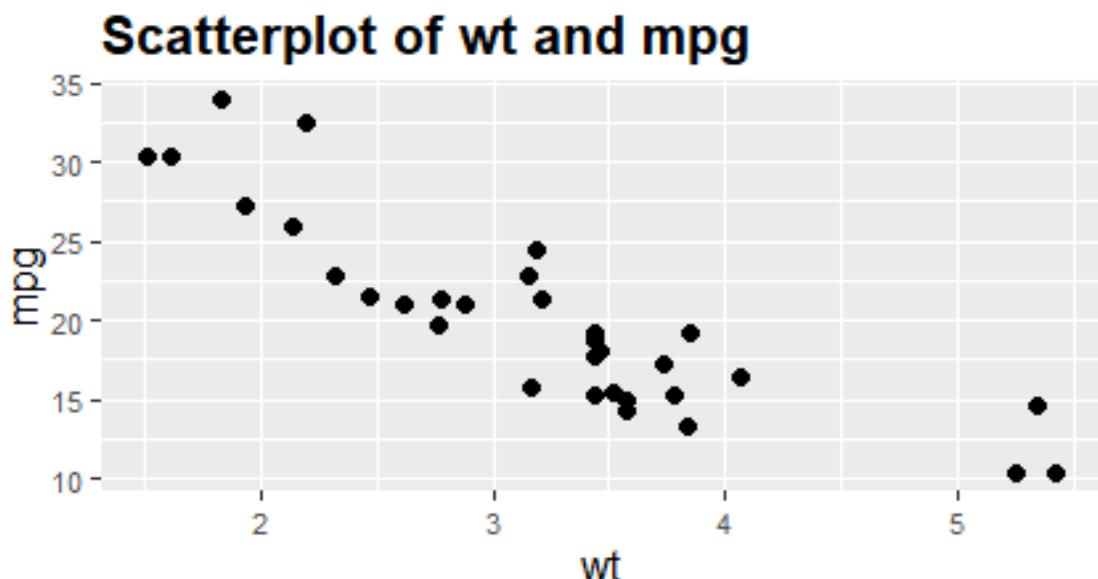
Gambar 5

- ✓ Dari 32 observasi *Motor Trend*, transmisi otomatis (kode 0) memiliki berat sekitar 2,5 sampai 5,5 dalam 1000 lbs dengan konsumsi bahan bakar dari 10 hingga 25 mpg.
- ✓ Untuk transmisi manual (kode 1), berat *Motor Trend* sekitar 1,5 sampai 3,5 dalam 1000 lbs lebih rendah dibandingkan dengan transmisi otomatis.
- ✓ Jadi, konsumsi bahan bakar mobil dengan transmisi manual cenderung lebih banyak karena berat dan konsumsi bahan bakar berbanding terbalik.

a.

Scatter plot dengan fungsi `ggplot()` memuat fungsi `geom_point()` untuk mengatur objek geomteris, fungsi `labs()` untuk mengatur label, dan fungsi `theme()` untuk mengubah tema.

```
ggplot(data=mtcars, aes(x=wt, y=mpg)) +
  geom_point(size=2) + labs(title = "Scatterplot of wt and mpg") +
  theme(axis.text = element_text(size=8),
        axis.title = element_text(size=12),
        plot.title = element_text(size=15, face="bold"))
```



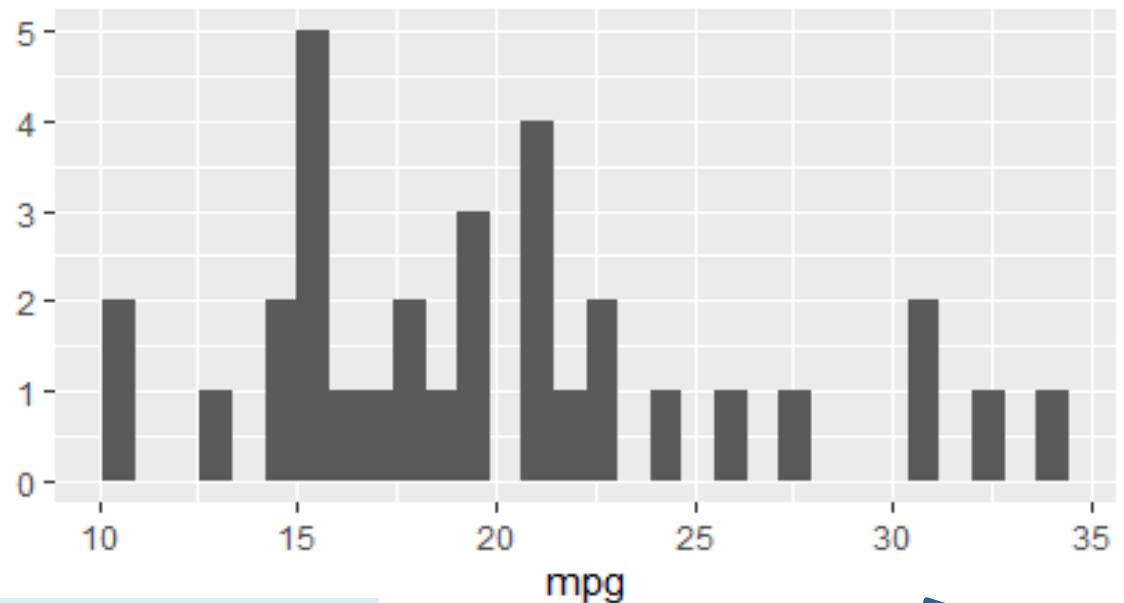
Gambar 6

b. Histogram

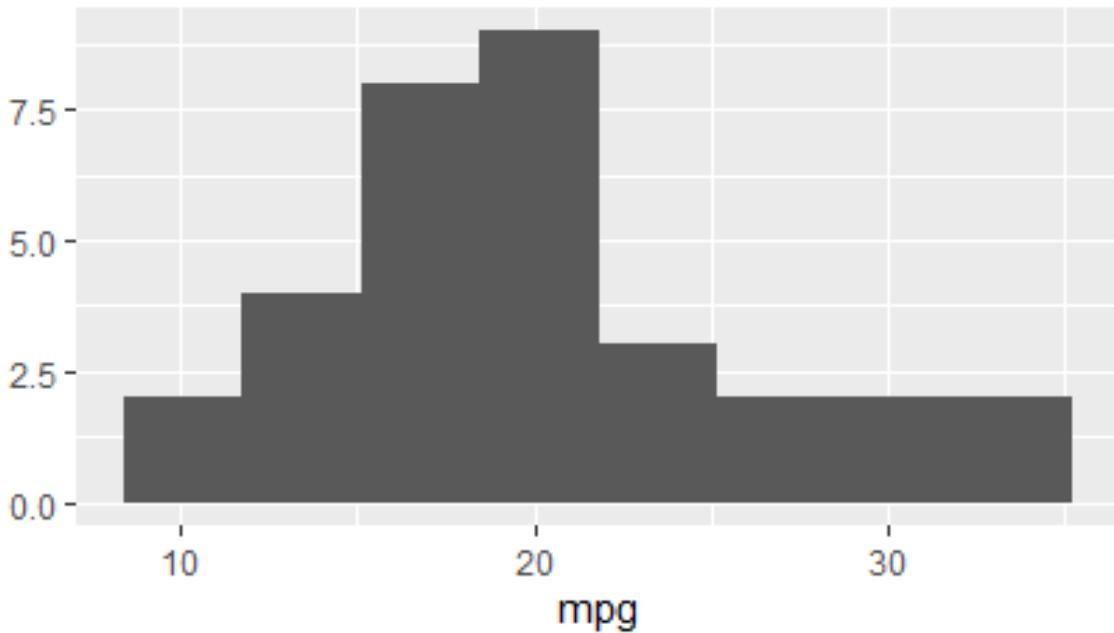
Gambar 8

```
library(ggplot2)
qplot(mpg, data = mtcars)

# atau dengan menentukan geom
qplot(mpg, data = mtcars, geom = "histogram")
```



Gambar 7



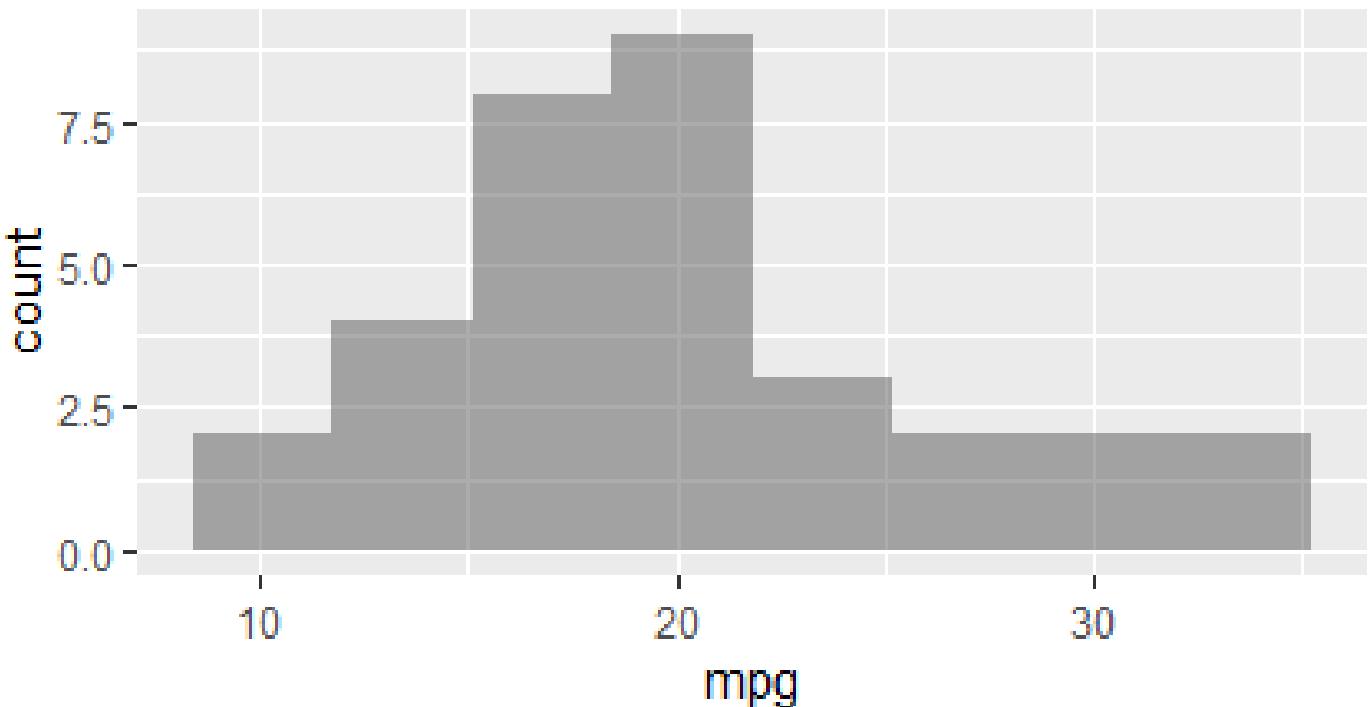
```
library(ggplot2)
qplot(mpg, data = mtcars, geom = "histogram", bins = 8)
```

Batang yang cukup banyak membuat distribusi data kurang terlihat sehingga kode ditambahkan bins untuk mengatur banyaknya batang.

b. Histogram

```
ggplot(data=mtcars, aes(x=mpg)) +  
  geom_histogram(alpha=0.5, bins = 8) +  
  ggtitle ("Histogram of mpg")
```

Histogram of mpg

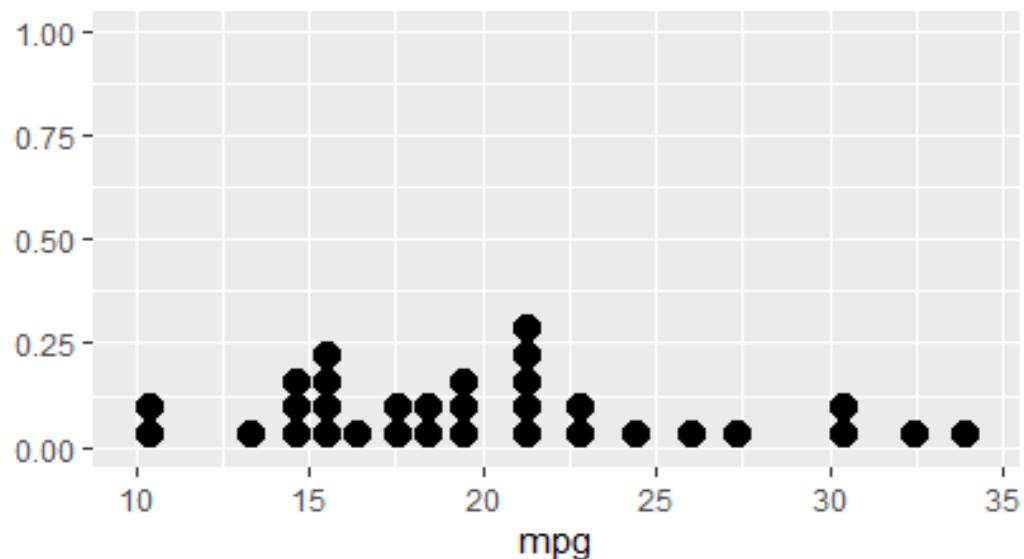


Gambar 9

C. Dot Plot

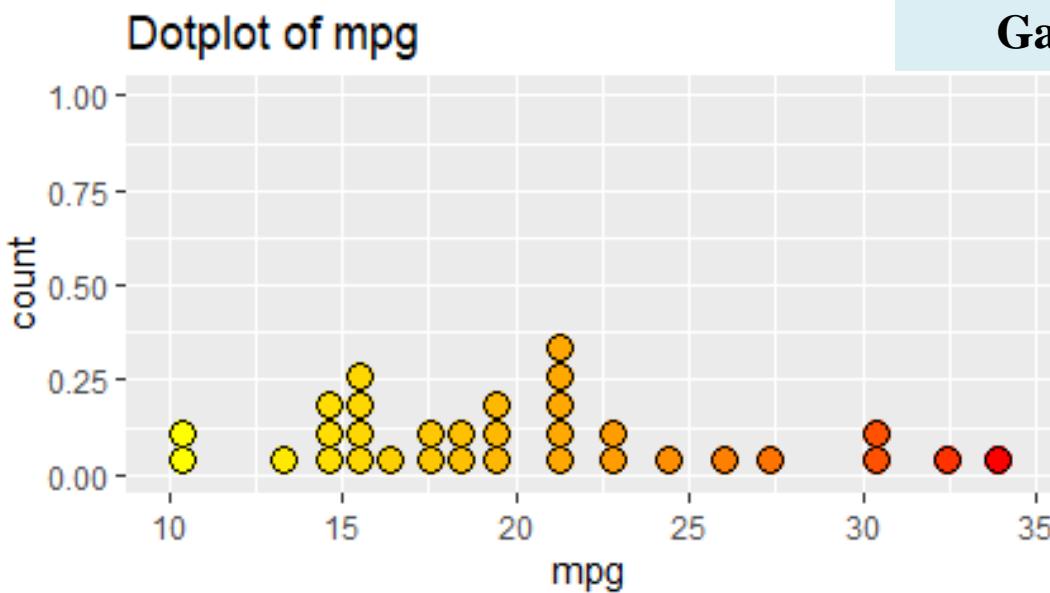
Nilai median sekitar 19 mpg, artinya 50% data berada di bawah 19 mpg dan 50% lainnya berada di atas nilai median.

```
library(ggplot2)
qplot(mpg, data = mtcars, geom = "dotplot")
```



Gambar 10

```
ggplot(data=mtcars, aes(x=mpg)) +
  geom_dotplot(aes(fill=..x..)) +
  ggtitle("Dotplot of mpg")+
  scale_fill_gradient(low="yellow", high="red") +
  guides(fill="none")
```

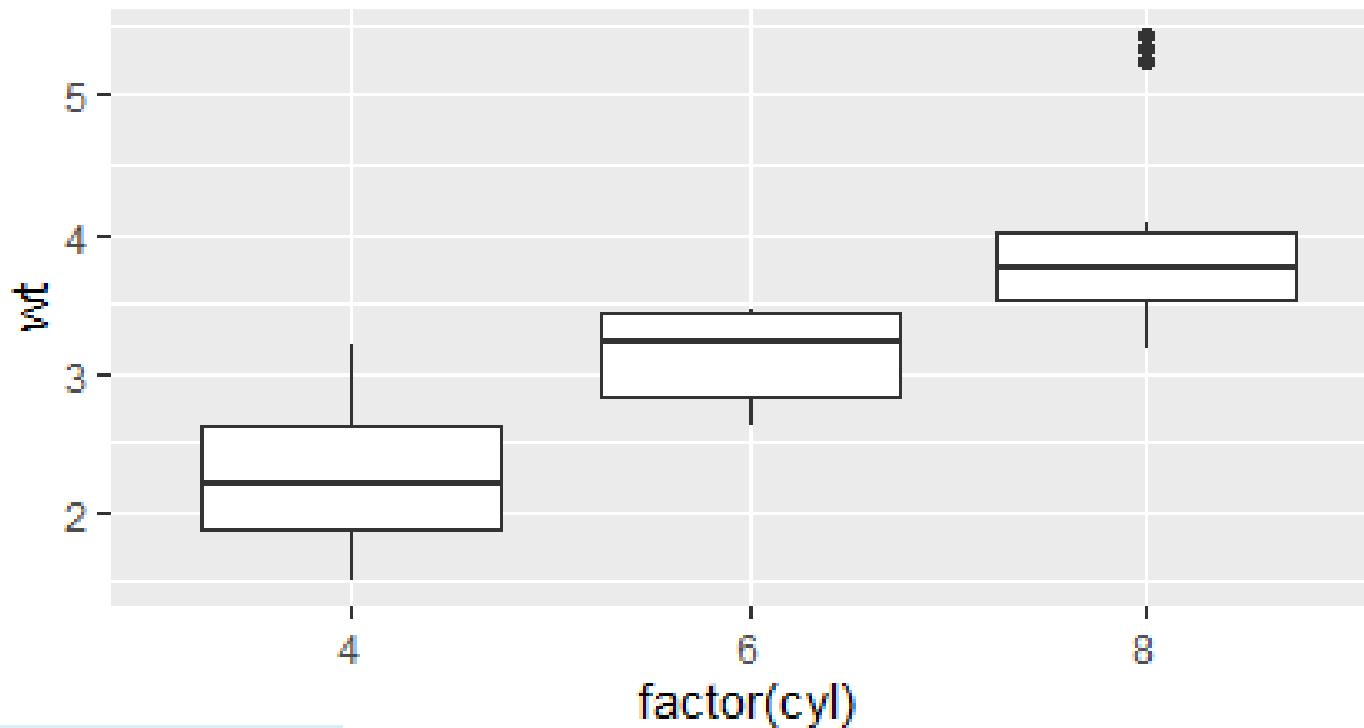


Gambar 11

Bandingkan sebaran data di bawah nilai median dan di atas nilai median, apa yang dapat disimpulkan?

d. Boxplot

```
library(ggplot2)
qplot(factor(cyl), wt, data = mtcars, geom = ("boxplot"))
```



Gambar 12

Hasil boxplot selaras dengan *scatter plot* dalam Gambar 3 dan Gambar 4.

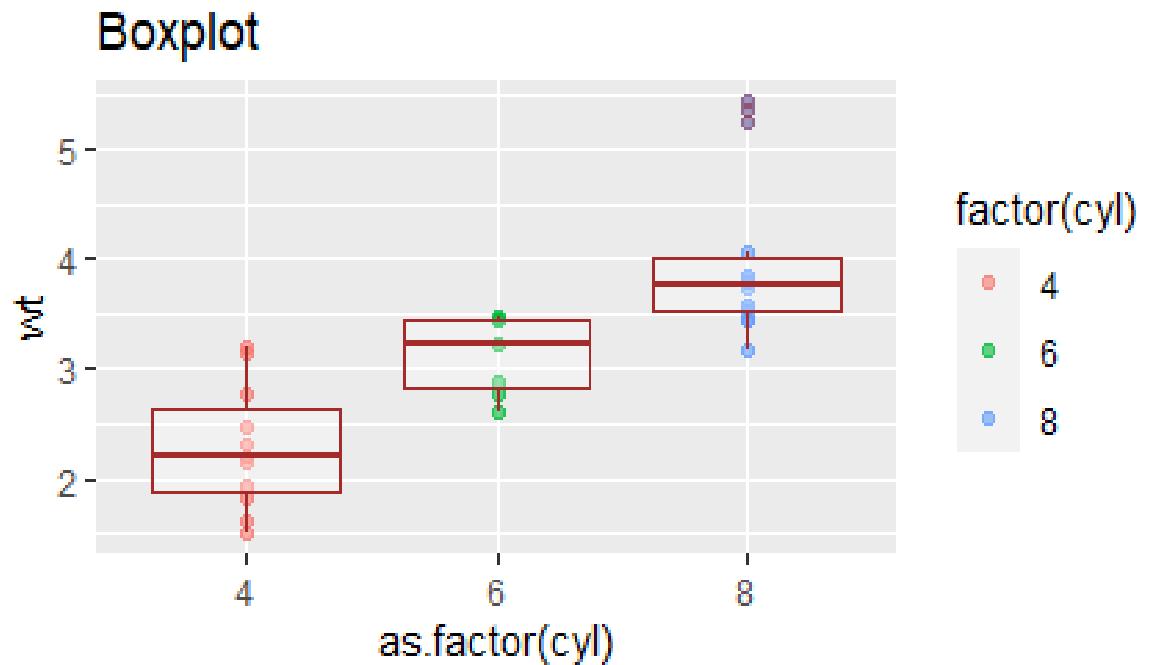
Nilai tengah berat mobil untuk jumlah silinder sebanyak 4 lebih rendah daripada silinder sebanyak 6 dan 8.

Untuk silinder sebanyak 6, *boxplot* menunjukkan adanya **data outlier**.

d. Boxplot

```
library(ggplot2)
ggplot(data=mtcars, aes(x=as.factor(cyl), y=wt)) +
  geom_point(aes(color=factor(cyl)), alpha=0.6) +
  geom_boxplot(alpha=0.3, colour = "brown" ) +
  ggtitle("Boxplot")
# + guides(colour=FALSE) agar legenda tidak muncul
```

- Titik data untuk setiap kategori jumlah silinder berbeda-beda warnanya yang ditunjukkan oleh tanda bulatan di setiap *boxplot*.
- Observasi paling sedikit adalah observasi dengan jumlah silinder sebanyak 6.
- Data *outlier* ditunjukkan dengan titik data yang memiliki *outline* warna coklat (ada di silinder 8).



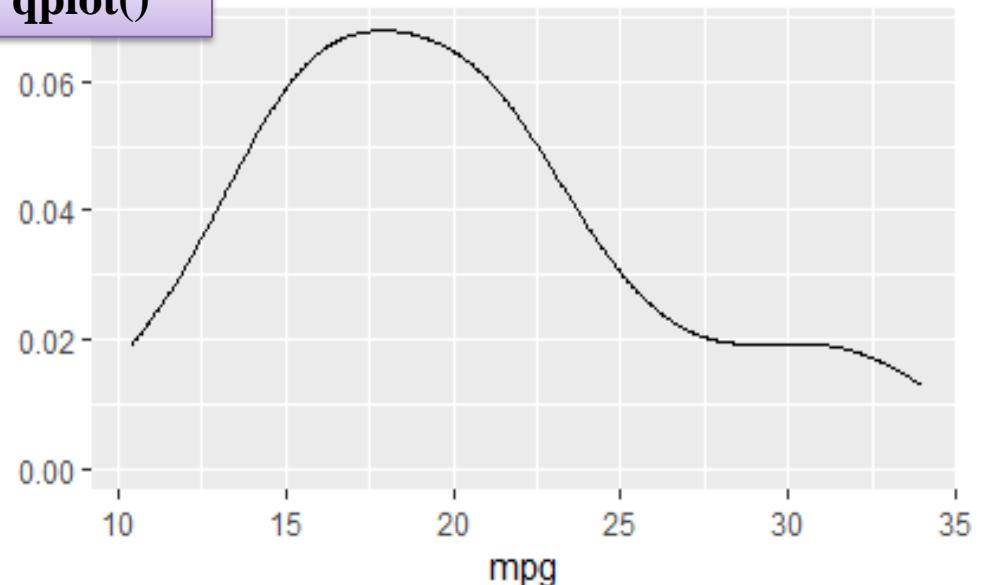
Gambar 13

e. Plot Densitas

```
library(ggplot2)
qplot(mpg, data = mtcars, geom = "density") # Gambar 14(a)

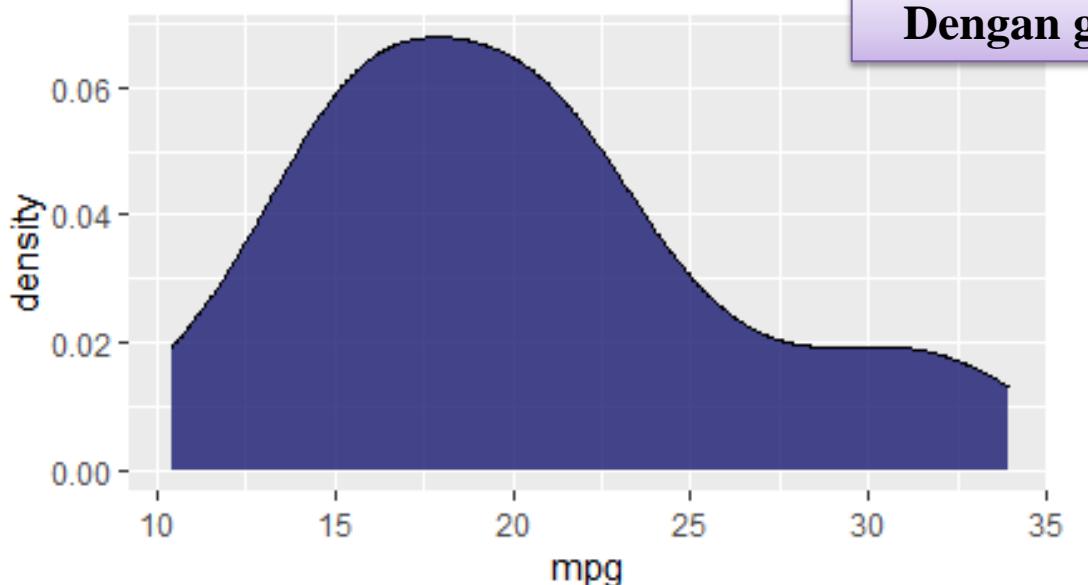
ggplot(data = mtcars, aes(x = mpg)) +
  geom_density(alpha= 0.8, fill = "midnightblue",
               color = "black") # Gambar 14(b)
```

Dengan qplot()



Kurva densitas sesuai dengan histogram pada Gambar 8, yaitu distribusi variabel mpg **miring ke kanan** (memiliki kemiringan positif) yang terlihat dengan adanya ekor pada bagian kanan. Hal ini disebabkan karena adanya dugaan **data outlier**.

Dengan ggplot()



Gambar 14

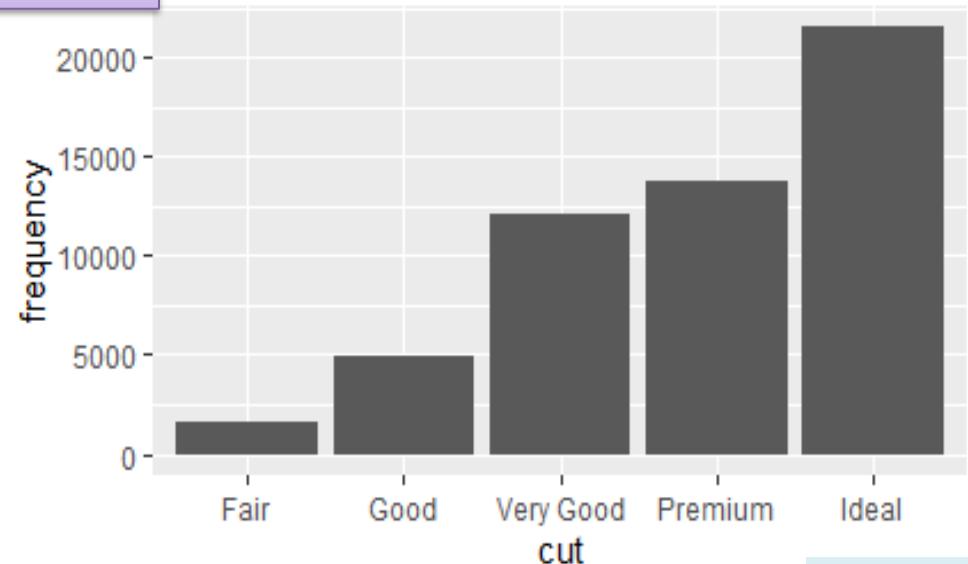
f. Diagram Batang

Diagram batang kualitas pemotongan berlian (cut) yang diambil dari data diamonds, terdiri dari 53.940 observasi, terdapat 5 kategori, yaitu *fair*, *good*, *very good*, *premium*, dan *ideal*.

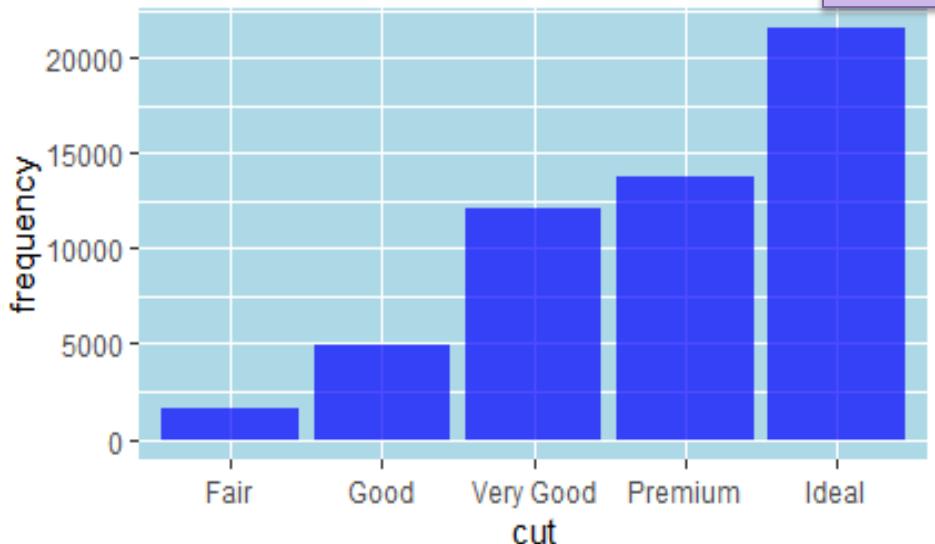
```
# membuat bar chart dengan qplot
qplot(cut, data = diamonds, geom = "bar", ylab = "frequency")
# membuat bar chart dengan ggplot
ggplot(diamonds, aes(cut)) +
  geom_bar(alpha=0.7, fill="blue") + labs(y="frequency") +
  theme(panel.background = element_rect(fill="light blue"))
```

Kualitas pemotongan berlian secara **ideal** memiliki frekuensi terbanyak, yaitu sekitar 22.000 berlian atau **sekitar 40%** dari total observasi.

Dengan `qplot()`



Dengan `ggplot()`

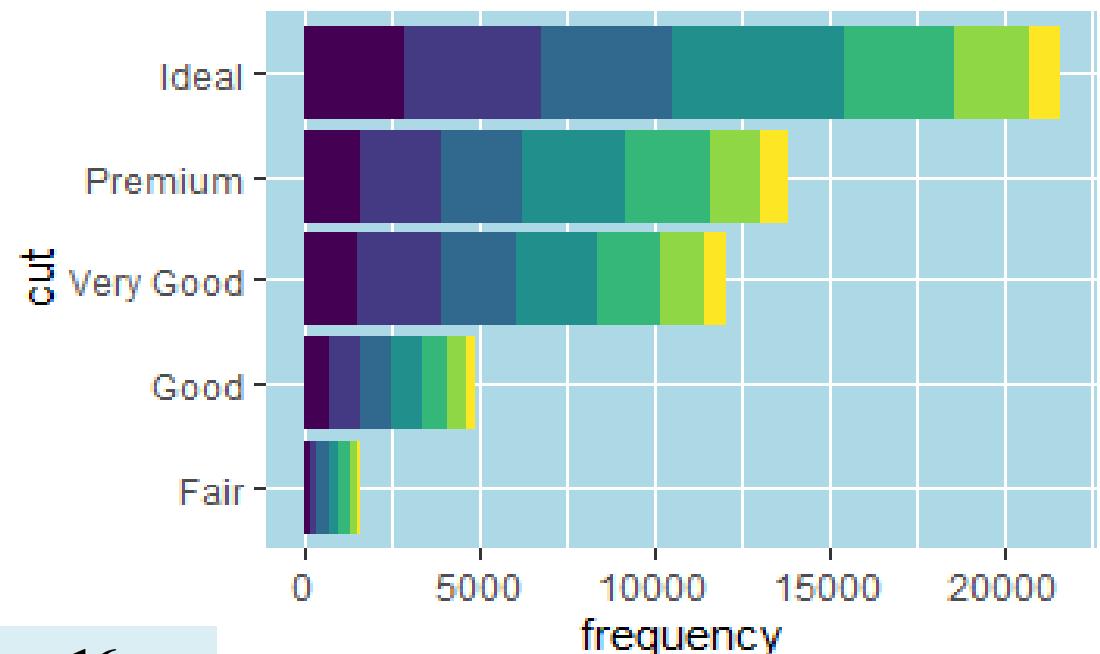


Gambar 15

f. Diagram Batang

Posisi batang diubah menjadi horizontal dengan `parameter_stack` dan menambahkan `variabel warna berlian (color)` ke dalam plot yang sama.

```
ggplot(diamonds, aes(y = cut)) +
  geom_bar(position = position_stack(reverse = TRUE), aes(fill=color)) +
  labs(x="frequency", y="cut") +
  theme(panel.background = element_rect(fill="light blue"))
```



Warna yang terdapat di setiap batang menunjukkan kategori warna berlian, mulai dari yang terbaik (label D) hingga level terburuk (label J).

Gambar 16

Referensi

Dietrich, D., Barry H., & Beibei Y., (2015), *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, John Wiley & Sons, Inc., Indianapolis, Indiana.

Pathak, M. A., (2014), *Beginning Data Science with R*, Springer International Publishing, Switzerland.

Peng, R. D., (2015), *Exploratory Data Analysis with R*, Lean Publishing.

www.rdocumentation.org/packages/ggplot2/versions/3.3.5

www.rdocumentation.org/packages/ggplot2/versions/3.3.5/topics/qplot

www.rdocumentation.org/packages/ggplot2/versions/3.3.5/topics/ggplot

www.tutorialgateway.org/r-ggplot2-density-plot/

<https://github.com/rstudio/cheatsheets/blob/master/data-visualization-2.1.pdf>

TUGAS

1. Melakukan visualisasi data dengan fungsi bawaan R dan paket ggplot2 untuk dataset yang sudah diunduh di UCI / Kaggle.
2. Memilih dan membuat plot yang tepat untuk setiap atribut/ variabel dataset. Gunakan prinsip-prinsip grafik dalam pertemuan sebelumnya.
3. Menginterpretasikan plot yang dibuat.



Thank You



Eksplorasi dan Visualisasi Data

Pertemuan 5:
Visualisasi Data Menggunakan R Studio

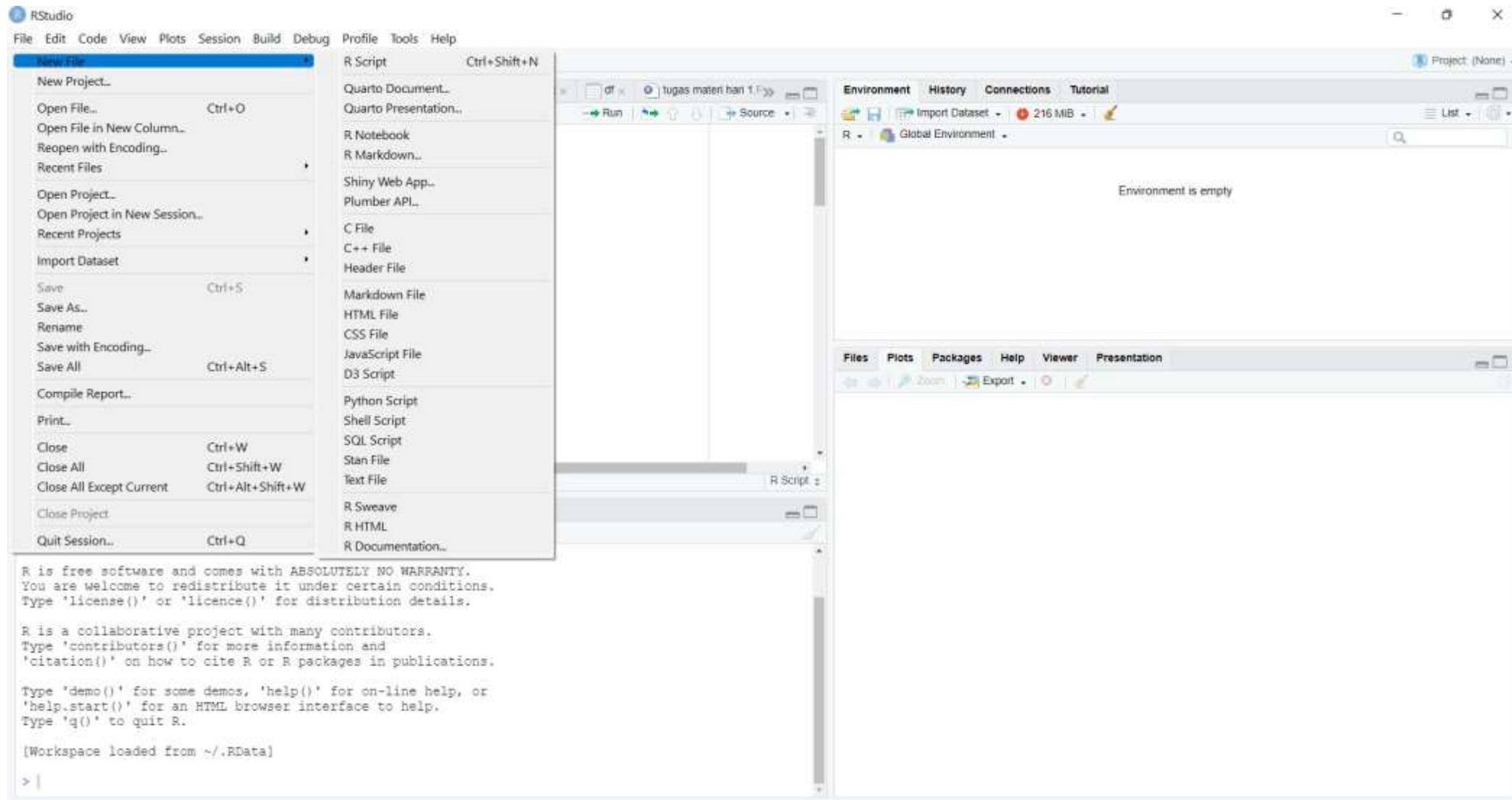
Content

- Menggunakan R Studio
- Membuat Plot dari built-in R
- Visualisasi dasar: scatterplot, bar plot, pie chart
- Membuat layer dan facet
- Visualisasi atribut numerik, atribut kategorik

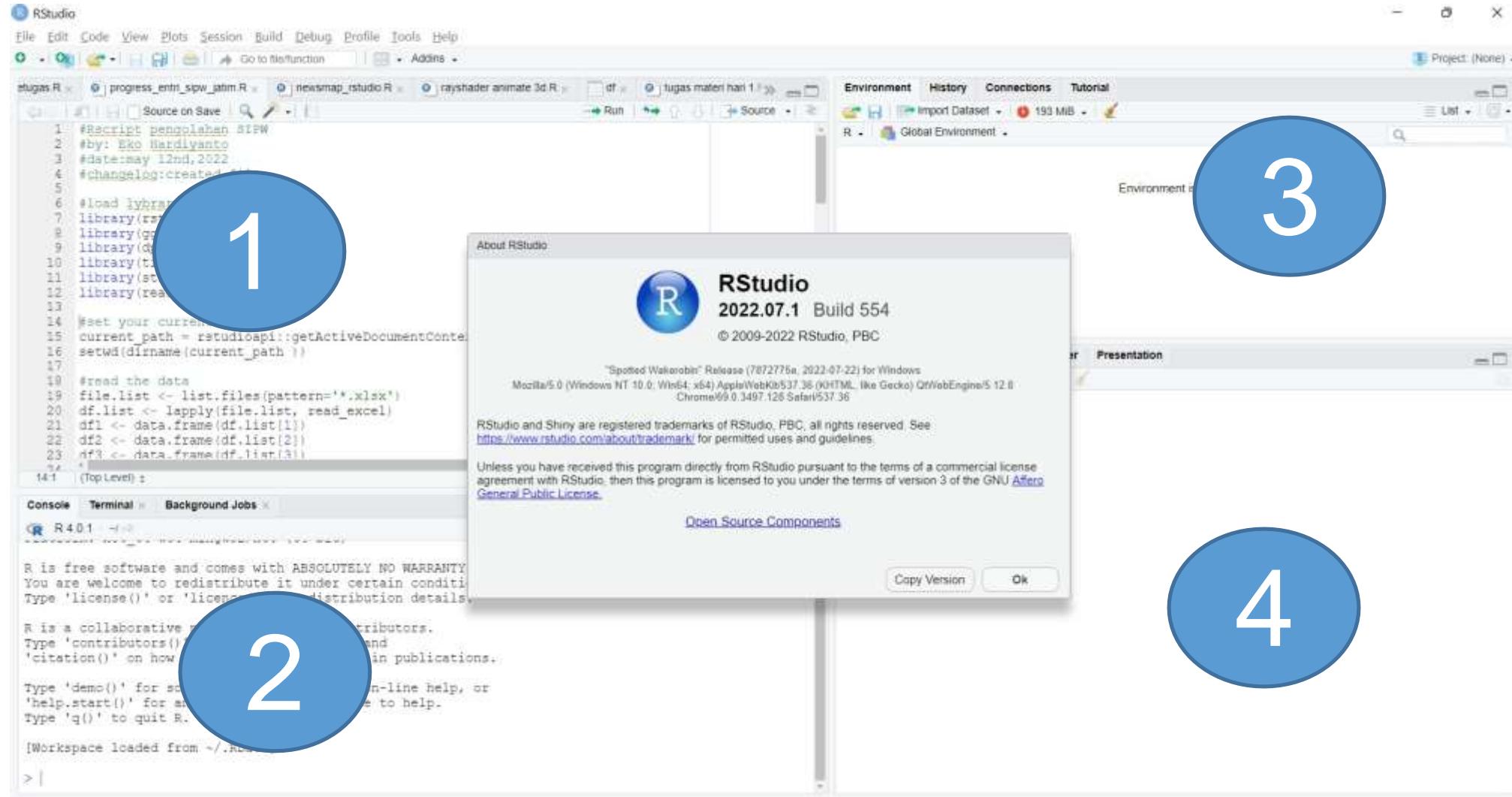
Goals

- Dapat menjelaskan jenis-jenis grafik sesuai jenis atribut/variabel
- Dapat membuat grafik sederhana sesuai dengan jenis atribut/variabel

Menggunakan R Studio



Versi dan layout R Studio



Tentang R

R dapat melakukan proses import data:

- file teks
- spreadsheet excel/csv
- paket statistic
- data spasial
- gambar
- sistem manajemen basis data.

Menu yang digunakan untuk load data 'impor'

Impor data dari csv atau text file

```
"rank", "discipline", "yrs.since.phd", "yrs.service", "sex", "salary"  
"Prof", "B", 19, 18, "Male", 139750  
"Prof", "B", 20, 16, "Male", 173200  
"AsstProf", "B", 4, 3, "Male", 79750  
"Prof", "B", 45, 39, "Male", 115000  
"Prof", "B", 40, 41, "Male", 141500  
"AssocProf", "B", 6, 6, "Male", 97000
```

```
library(readr)  
  
# import data from a comma delimited file  
Salaries <- read_csv("salaries.csv")  
  
# import data from a tab delimited file  
Salaries <- read_tsv("salaries.txt")
```

Impor data dari excel workbooks

```
library(readxl)

# import data from an Excel workbook
Salaries <- read_excel("salaries.xlsx", sheet=1)
```

Impor data dari shape file

```
library(rgdal)  
  
# Spatial data reading
```

```
map_cities_gis <- readOGR(dsn =  
  "./data/cities_gis/city.shp",  
  verbose=FALSE)
```



Jenis Library di R Studio

- Built-in atau Base R (barplot, pie, hist, dll)
- Third Party (dplyr, rgdal, knitr, tidyverse, maptools, ggplot, dll)



Jenis-jenis grafik

- **Grafik Univariat**
 - Kategorik
 - Bar Charts (Diagram Batang)
 - Pie Chart (Diagram Lingkaran)
 - Kuantitatif
 - Histogram
 - Kernel Density Plot
 - Dot Chart (Diagram Titik)
- **Grafik Bivariat**
 - Categorical vs categorical
 - Stacked bar chart
 - Grouped bar chart
 - Segmented bar chart
 - Quantitative vs quantitative
 - Line graph
 - scatterplot
 - Categorical vs quantitative
 - Bar chart
 - Density plot
 - Box plot
 - Ridgeline plot
 - Cleveland dot plot



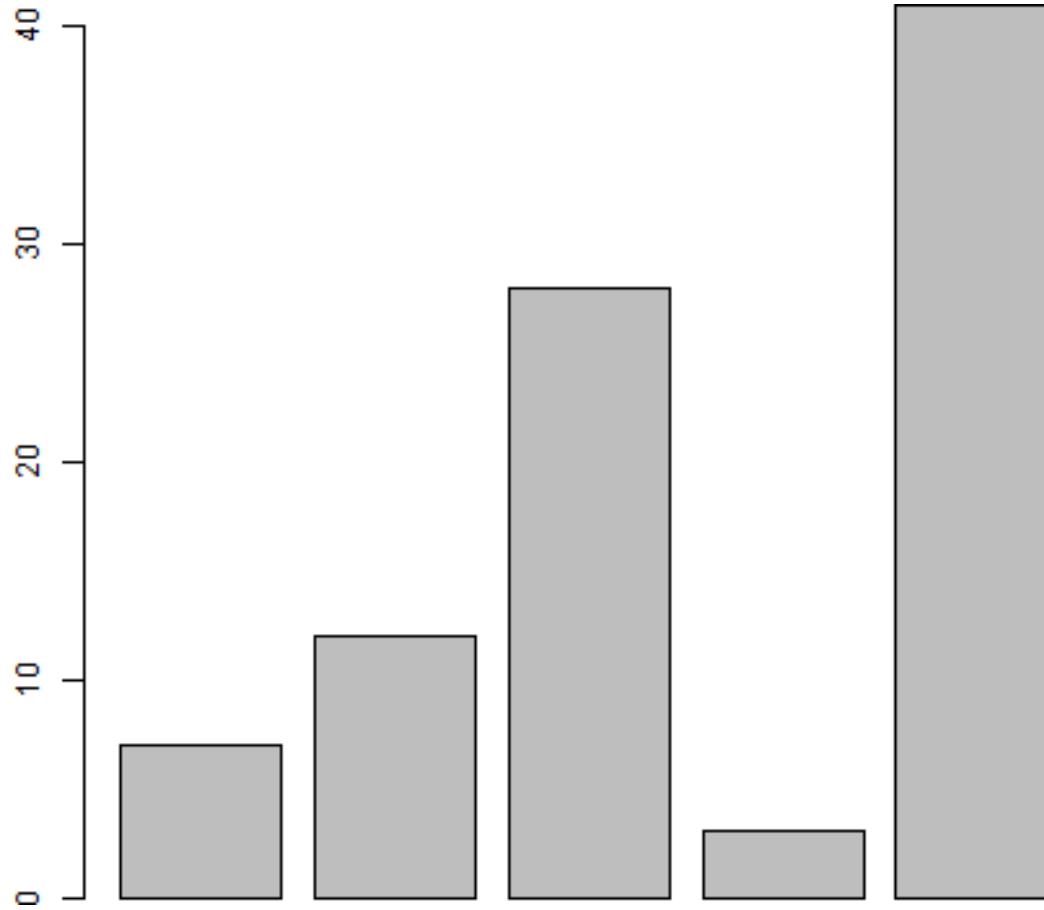
Grafik Univariat – Data Kategori – Barchart

- Untuk membandingkan frekuensi kelompok yang berbeda.
Contoh seperti ini:

```
# Create the data for the chart  
H <- c(7,12,28,3,41)
```

```
# Plot the bar chart  
barplot(H)
```

Output

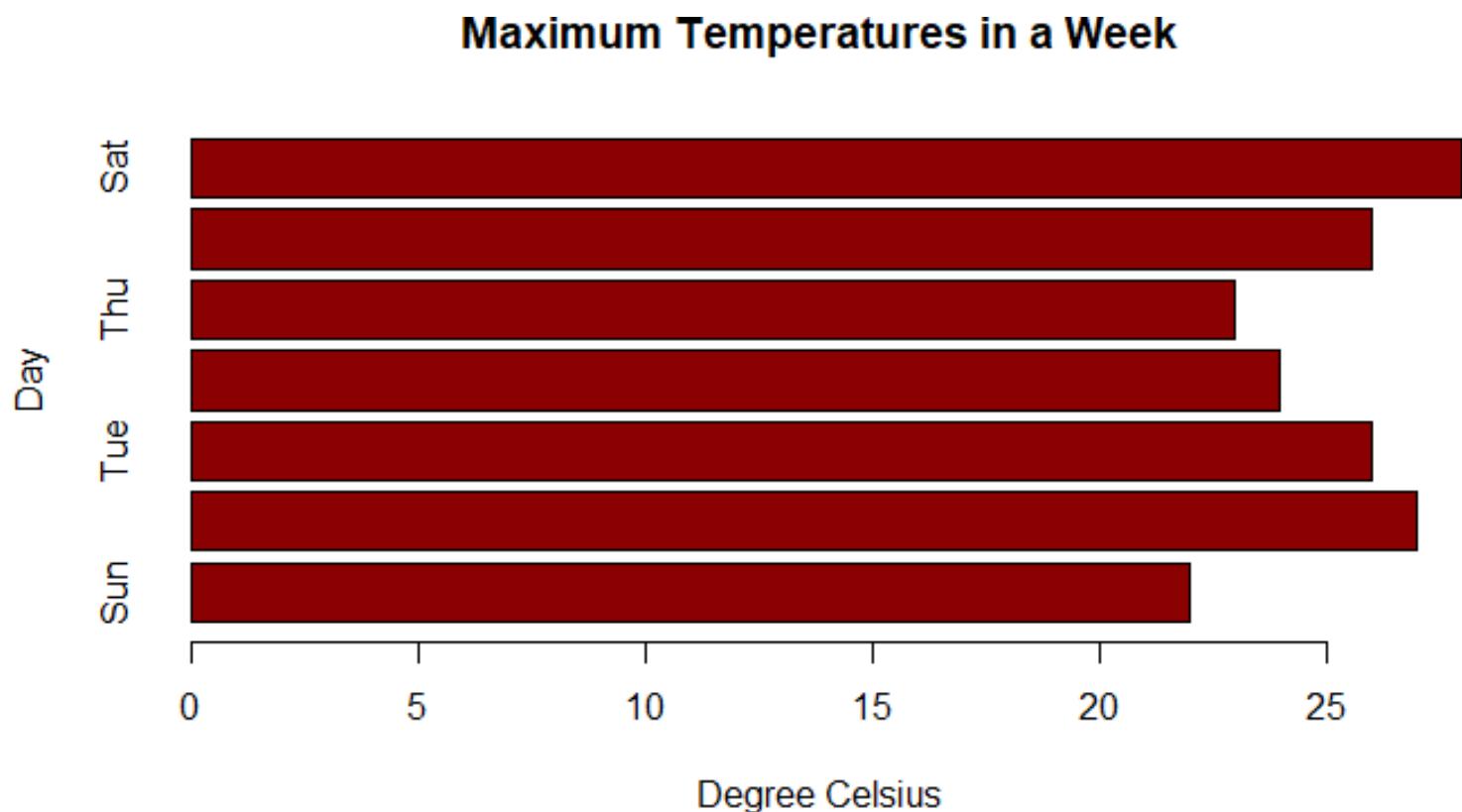


Grafik Univariat – Data Kategori – Barchart Horizontal

- Barchart juga dapat diubah menjadi horizontal. Contoh seperti ini:

```
# barchart with added parameters  
max.temp <- c(22, 27, 26, 24, 23, 26, 28)  
  
Barplot(max.temp,  
        main = "Maximum Temperatures in a Week",  
        xlab = "Degree Celsius",  
        ylab = "Day",  
        names.arg = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"),  
        col = "darkred",  
        horiz = TRUE)
```

Output



Grafik Univariat – Data Kategori – Pie Chart (Diagram Lingkaran) (1)

Diagram lingkaran kontroversial dalam statistik. Jika tujuan anda adalah untuk membandingkan frekuensi kategori, anda lebih baik menggunakan diagram batang (manusia lebih baik dalam menilai panjang batang daripada volume irisan pie).

Jika tujuan anda adalah membandingkan setiap kategori secara keseluruhan dan jumlah kategorinya kecil, maka diagram lingkaran mungkin lebih cocok untuk digunakan. Untuk membuat diagram lingkaran yang menarik dengan R dibutuhkan lebih banyak penulisan kode.

Grafik Univariat – Data Kategori – Pie Chart (Diagram Lingkaran) (2)

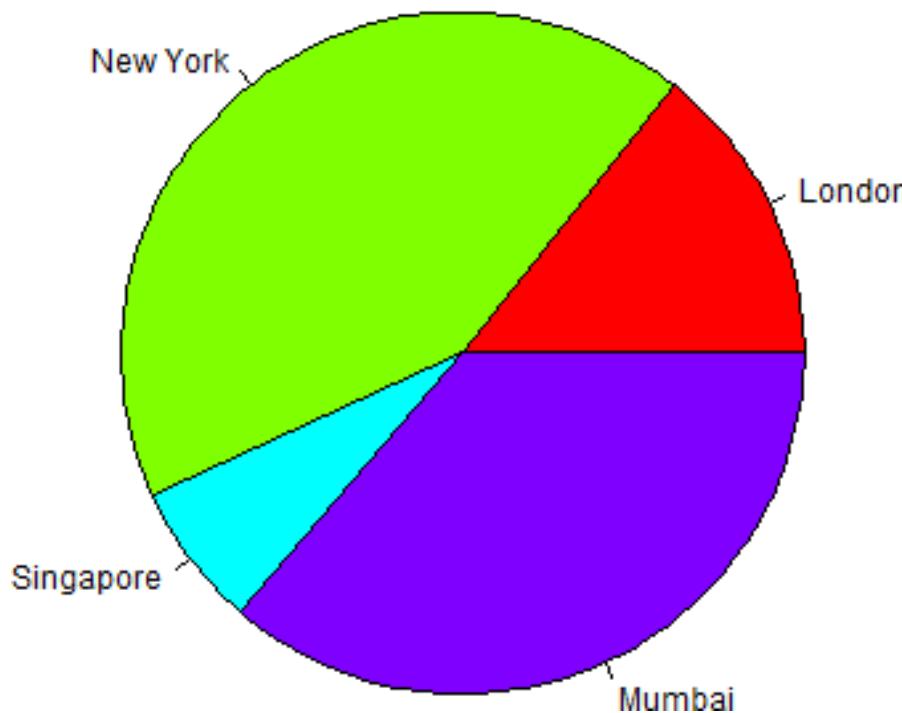
Contoh kode dengan pie chart:

```
# Create data for pie chart graph.  
x <- c(21, 62, 10, 53)  
labels <- c("London", "New York", "Singapore", "Mumbai")
```

```
# Plot the chart with title and rainbow color pallet.  
pie(x, labels, main = "City pie chart", col = rainbow(length(x)))
```

Output

City pie chart



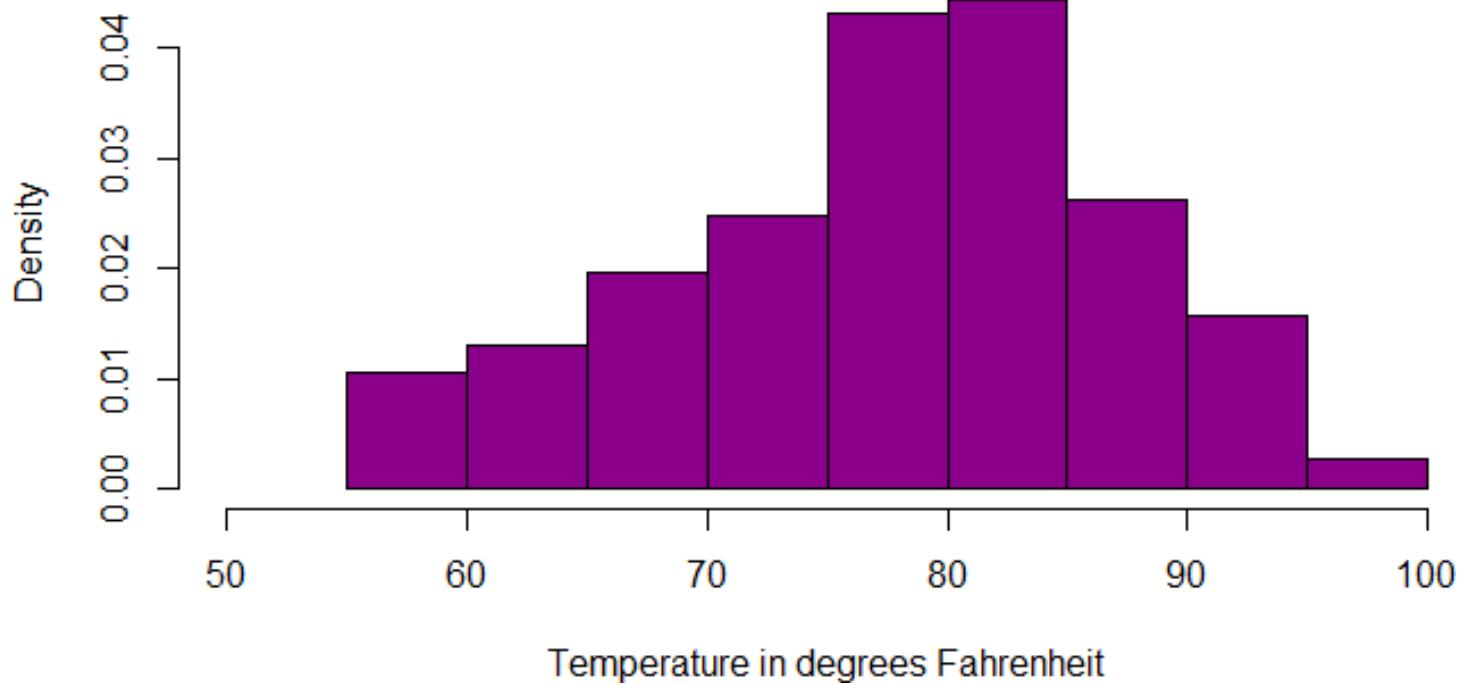
Grafik Univariat – Kuantitatif – Histogram

Distribusi variabel kuantitatif tunggal biasanya diplot menggunakan histogram dan *kernel*

```
Temperature <- airquality$Temp  
# histogram with added parameters  
hist(Temperature,  
      main="Maximum daily temperature at La Guardia Airport",  
      xlab="Temperature in degrees Fahrenheit",  
      xlim=c(50,100),  
      col="darkmagenta",  
      freq=FALSE )
```

Output

Maximum daily temperature at La Guardia Airport



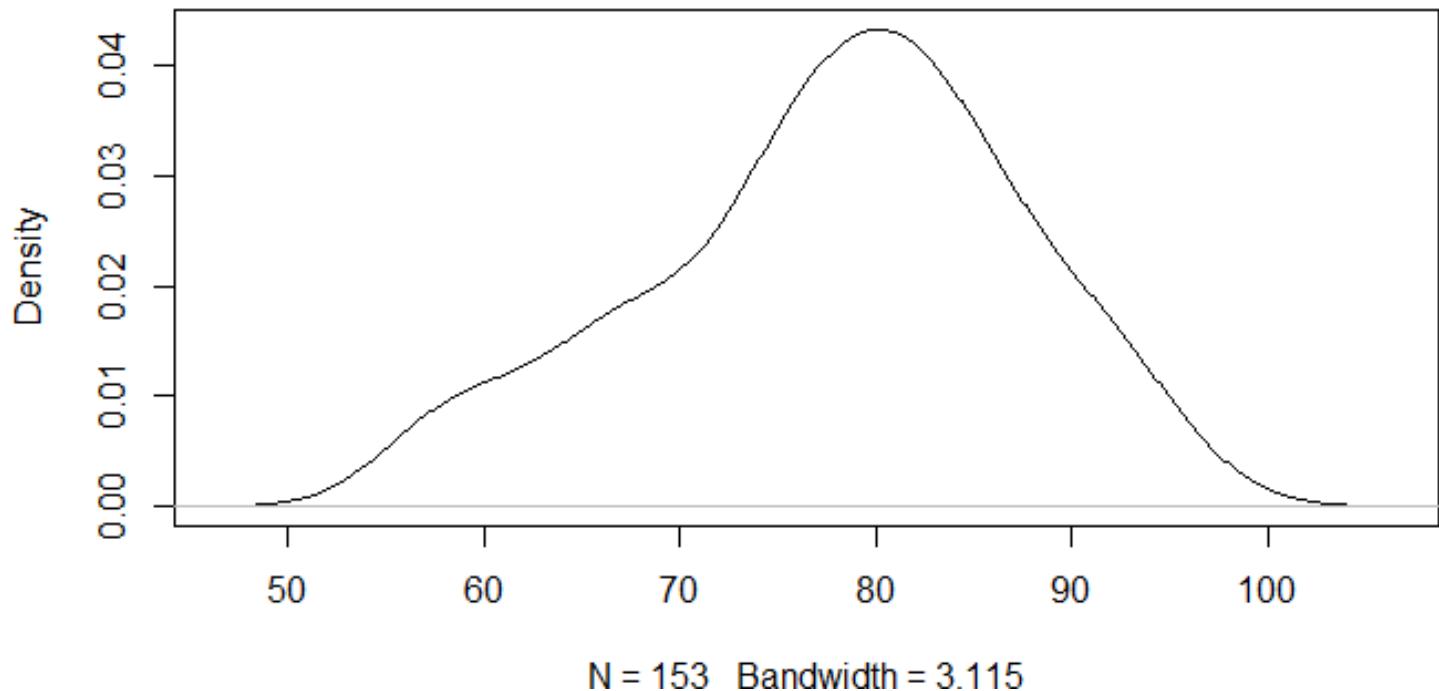
Grafik Univariat – Kuantitatif – Kernel Density Plot

Distribusi variabel kuantitatif tunggal biasanya diplot menggunakan histogram dan *kernel*

```
pressure_density <- density(airquality$Temp)  
plot(pressure_density)
```

Output

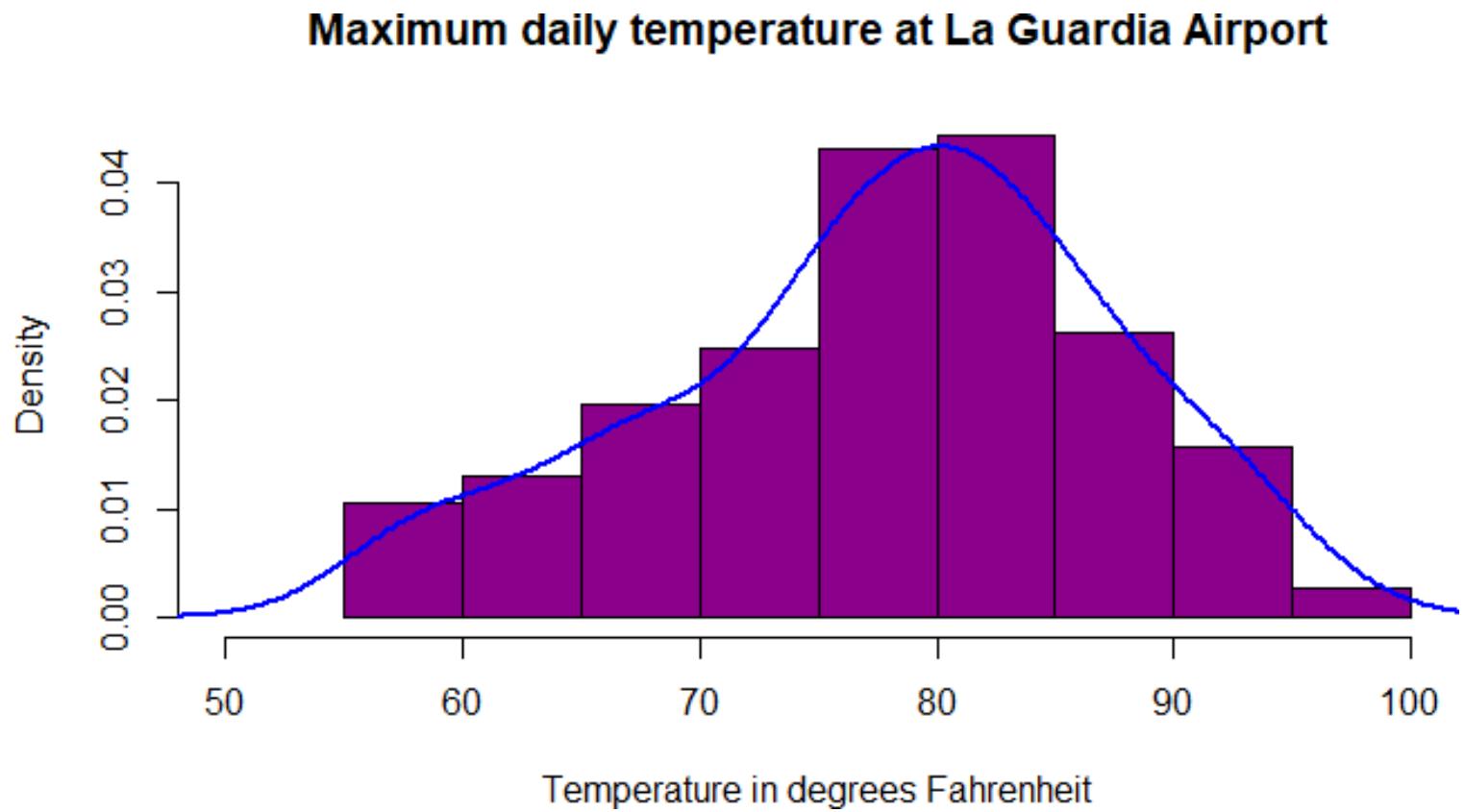
```
density.default(x = airquality$Temp)
```



Grafik Univariat – Kuantitatif – Combined Histogram and Kernel Density Plot

```
Temperature <- airquality$Temp  
hist(Temperature,  
      main="Maximum daily temperature at La Guardia Airport",  
      xlab="Temperature in degrees Fahrenheit",  
      xlim=c(50,100),  
      col="darkmagenta",  
      freq=FALSE  
)  
lines(density(airquality$Temp),col="blue", lwd=2)
```

Output



Grafik Univariat – Kuantitatif – Dot Chart

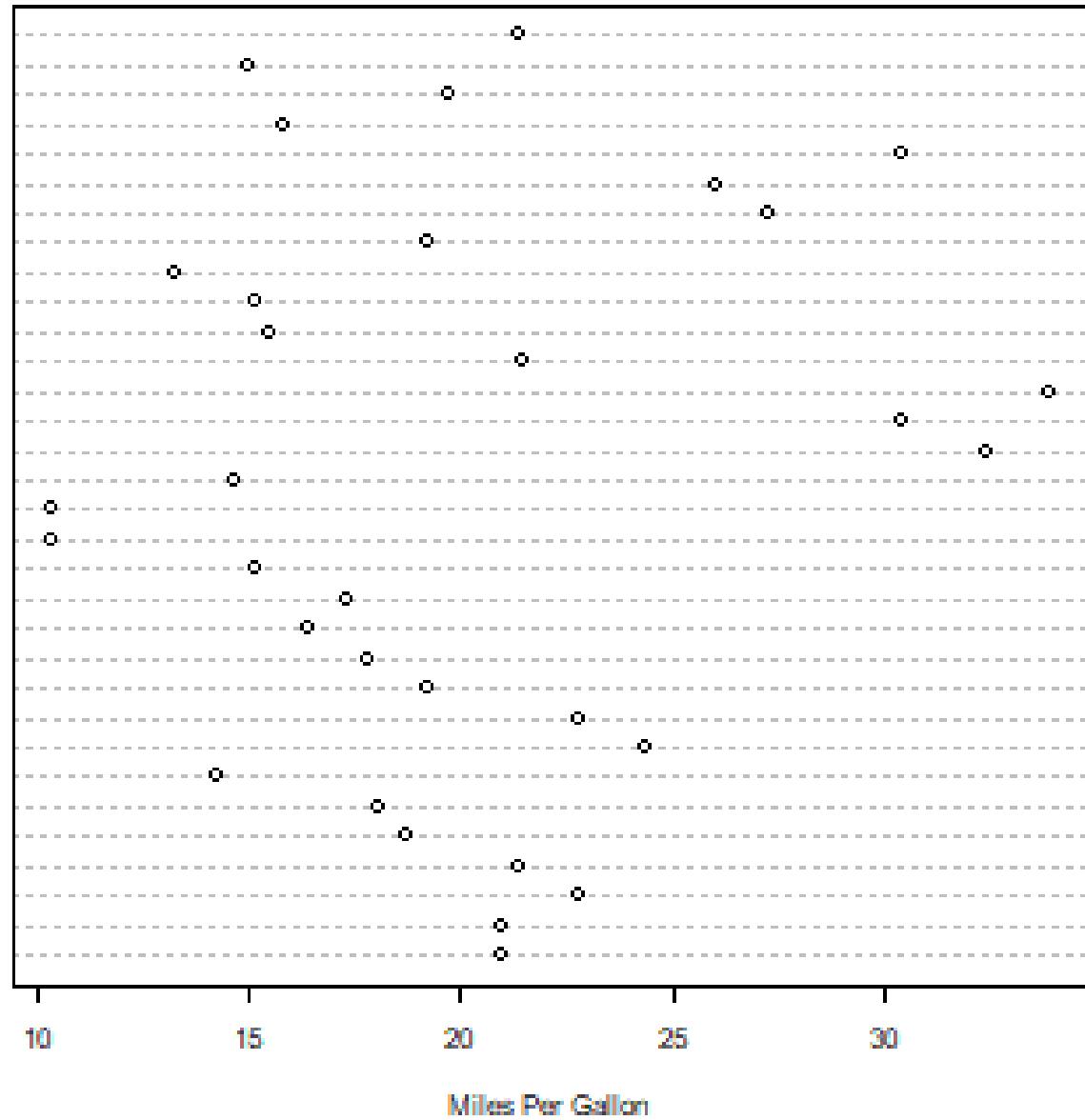
Alternatif lain untuk histogram adalah diagram titik. Variabel kuantitatif dibagi ke dalam beberapa bin, berbeda dengan diagram batang, setiap pengamatan diwakili oleh sebuah titik. Diagram titik bekerja paling baik pada data dengan jumlah kecil

```
dotchart(mtcars$mpg, labels=row.names(mtcars),  
        cex=.5,  
        main="Gas Milleage for Car Model",  
        xlab="Miles Per Gallon"  
        )
```

Output

Volvo 142E
Maserati Bora
Ferrari Dino
Ford Pantera L
Lotus Europa
Porsche 914-2
Fiat X1-9
Pontiac Firebird
Camaro Z28
AMC Javelin
Dodge Challenger
Toyota Corona
Toyota Corolla
Honda Civic
Fiat 128
Chrysler Imperial
Lincoln Continental
Cadillac Fleetwood
Merc 450SLC
Merc 450SL
Merc 450SE
Merc 280C
Merc 280
Merc 230
Merc 240D
Duster 380
Valiant
Hornet Sportabout
Hornet 4 Drive
Datsun 710
Mazda RX4 Wag
Mazda RX4

Gas Milleage for Car Model



Grafik Bivariat

Grafik bivariat (2 Variabel) digunakan untuk menampilkan hubungan antara dua variabel. Jenis grafik akan tergantung pada skala pengukuran variabel (kategorik atau kuantitatif)

Contoh scatterplot, Scatterplot merupakan visualisasi data dalam bentuk titik-titik yang ditampilkan dalam sumbu x dan ya. Sumbu x dan y mewakili nilai dari masing-masing variabel.

Grafik Bivariat – Stacked Bar Chart

Saat memplot hubungan antara dua variabel kategori, *stacked*, *grouped*, atau *segmented* bar charts dapat digunakan.

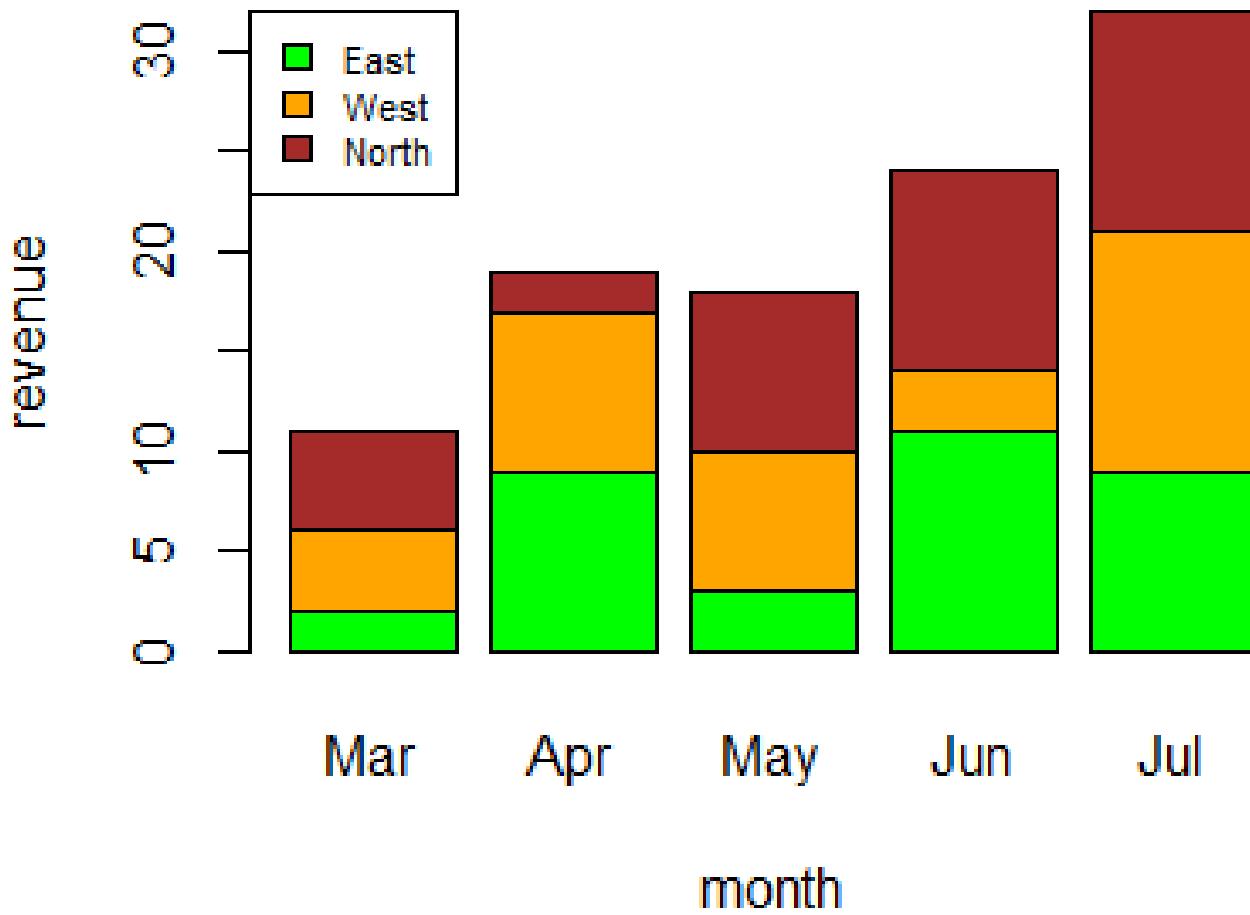
```
# Create the stacked bar chart
colors = c("green","orange","brown")
months <- c("Mar","Apr","May","Jun","Jul")
regions <- c("East","West","North")

# Create the matrix of the values.
Values <- matrix(c(2,9,3,11,9,4,8,7,3,12,5,2,8,10,11),
                  nrow = 3,
                  ncol = 5,
                  byrow = TRUE)
# Create the bar chart
barplot(Values, main = "total revenue",
        names.arg = months,
        xlab = "month",
        ylab = "revenue",
        col = colors)

# Add the legend to the chart
legend("topleft", regions, cex = .7, fill = colors)
```

Output

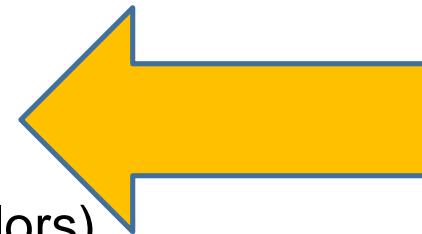
total revenue



Grafik Bivariat – Grouped Bar Chart

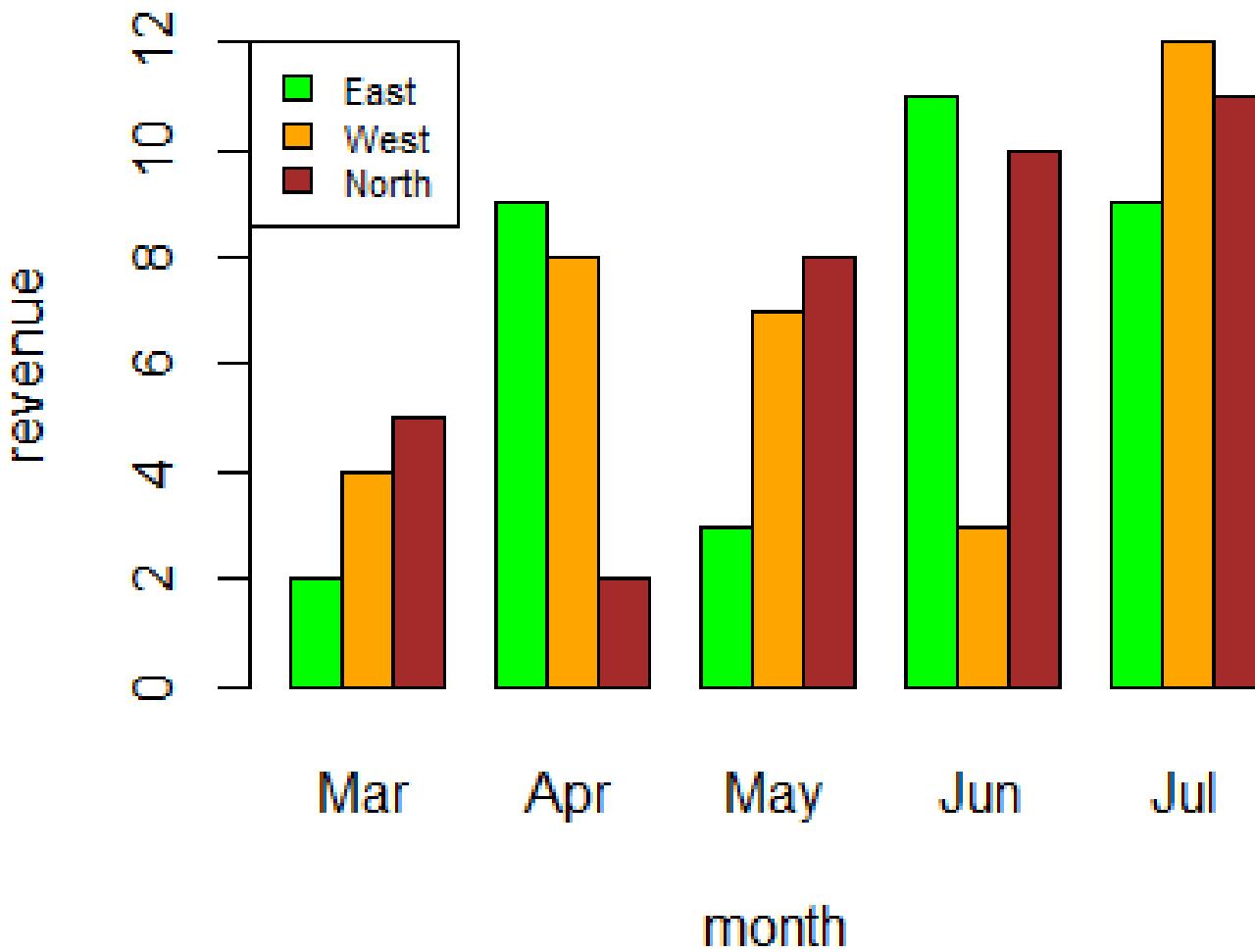
Saat memplot hubungan antara dua variabel kategori, *stacked*, *grouped*, atau *segmented* bar charts dapat digunakan.

```
# Create the grouped bar chart  
colors = c("green", "orange", "brown")  
months <- c("Mar", "Apr", "May", "Jun", "Jul")  
regions <- c("East", "West", "North")  
  
# Create the matrix of the values.  
Values <- matrix(c(2,9,3,11,9,4,8,7,3,12,5,2,8,10,11),  
                  nrow = 3,  
                  ncol = 5,  
                  byrow = TRUE)  
  
# Create the bar chart  
barplot(Values, main = "total revenue",  
        names.arg = months,  
        xlab = "month",  
        ylab = "revenue",  
        col = colors,  
        beside=TRUE)  
  
# Add the legend to the chart  
legend("topleft", regions, cex = .7, fill = colors)
```



Output

total revenue



Grafik Bivariat – Segmented Bar Chart

Saat memplot hubungan antara dua variabel kategori, *stacked*, *grouped*, atau *segmented bar charts* dapat digunakan.

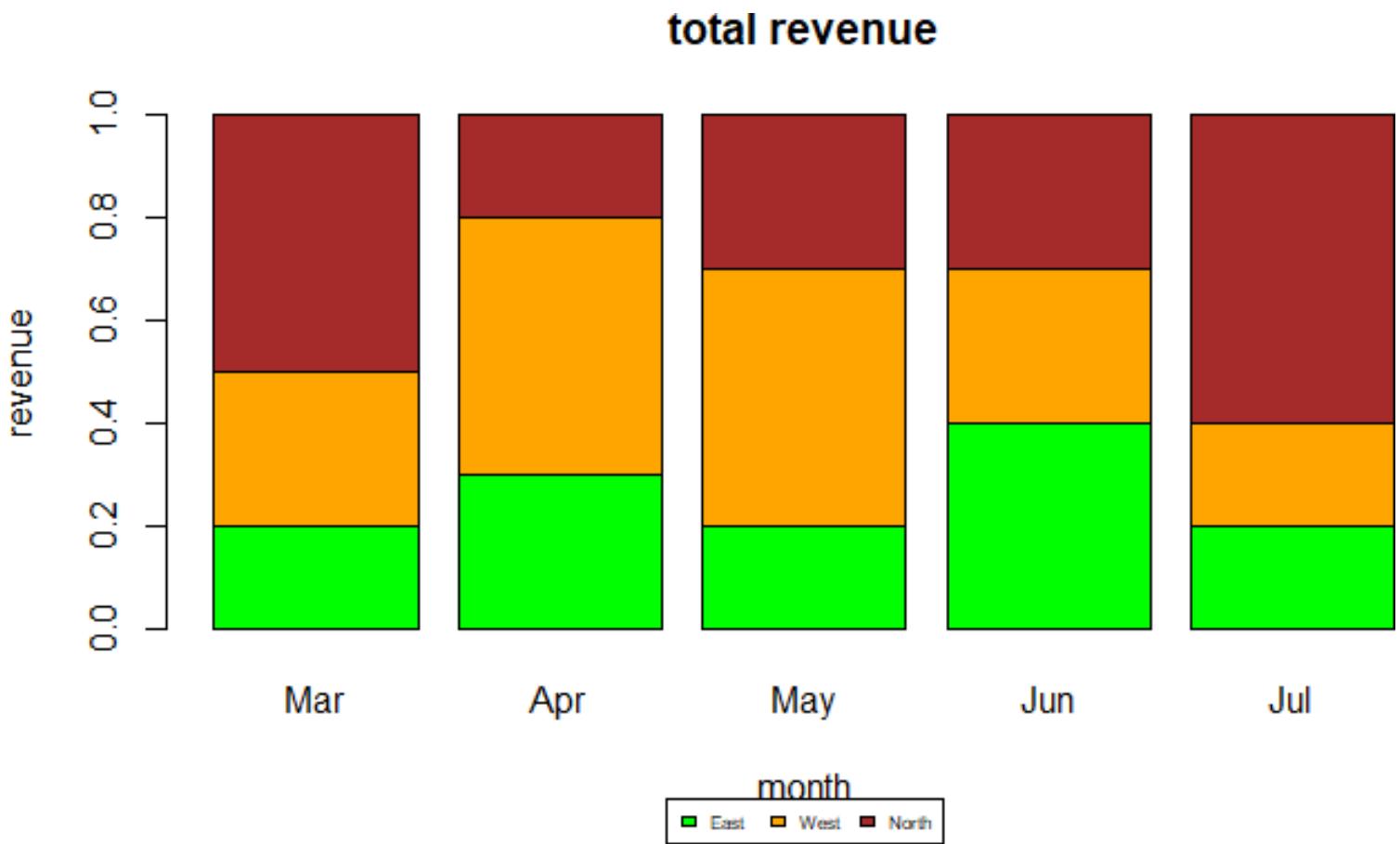
```
# Create the grouped bar chart
colors = c("green", "orange", "brown")
months <- c("Mar", "Apr", "May", "Jun", "Jul")
regions <- c("East", "West", "North")

# Create the matrix of the values.
Values <- matrix(c(0.2,0.3,0.2,0.4,0.2,
                  0.3,0.5,0.5,0.3,0.2,
                  0.5,0.2,0.3,0.3,0.6),
                  nrow = 3,
                  ncol = 5,
                  byrow = TRUE)

# Create the bar chart
barplot(Values, main = "total revenue",
        names.arg = months,
        xlab = "month",
        ylab = "revenue",
        col = colors)

# Add the legend to the chart
legend("topleft", regions, cex = .7, fill = colors)
```

Output

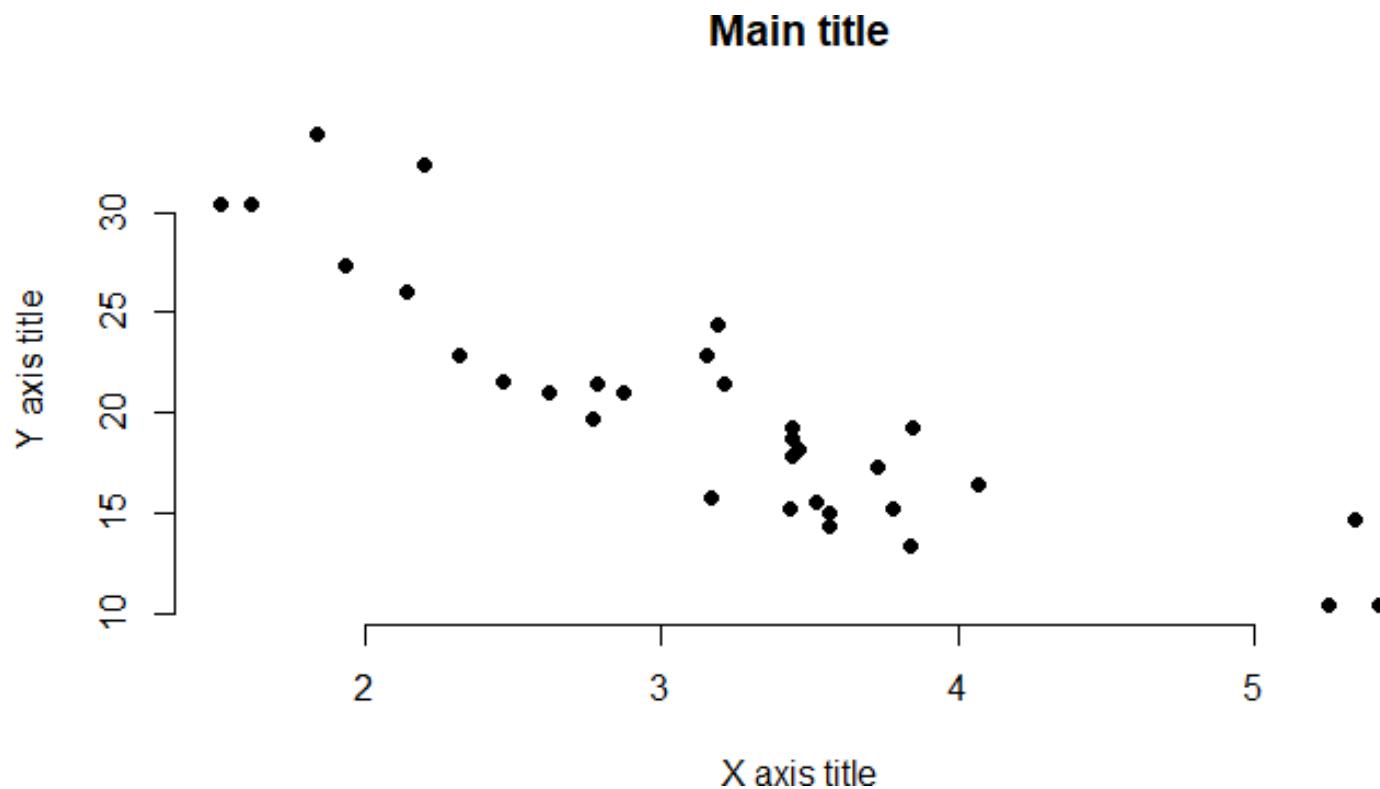


Grafik Bivariat – Scatter Plot

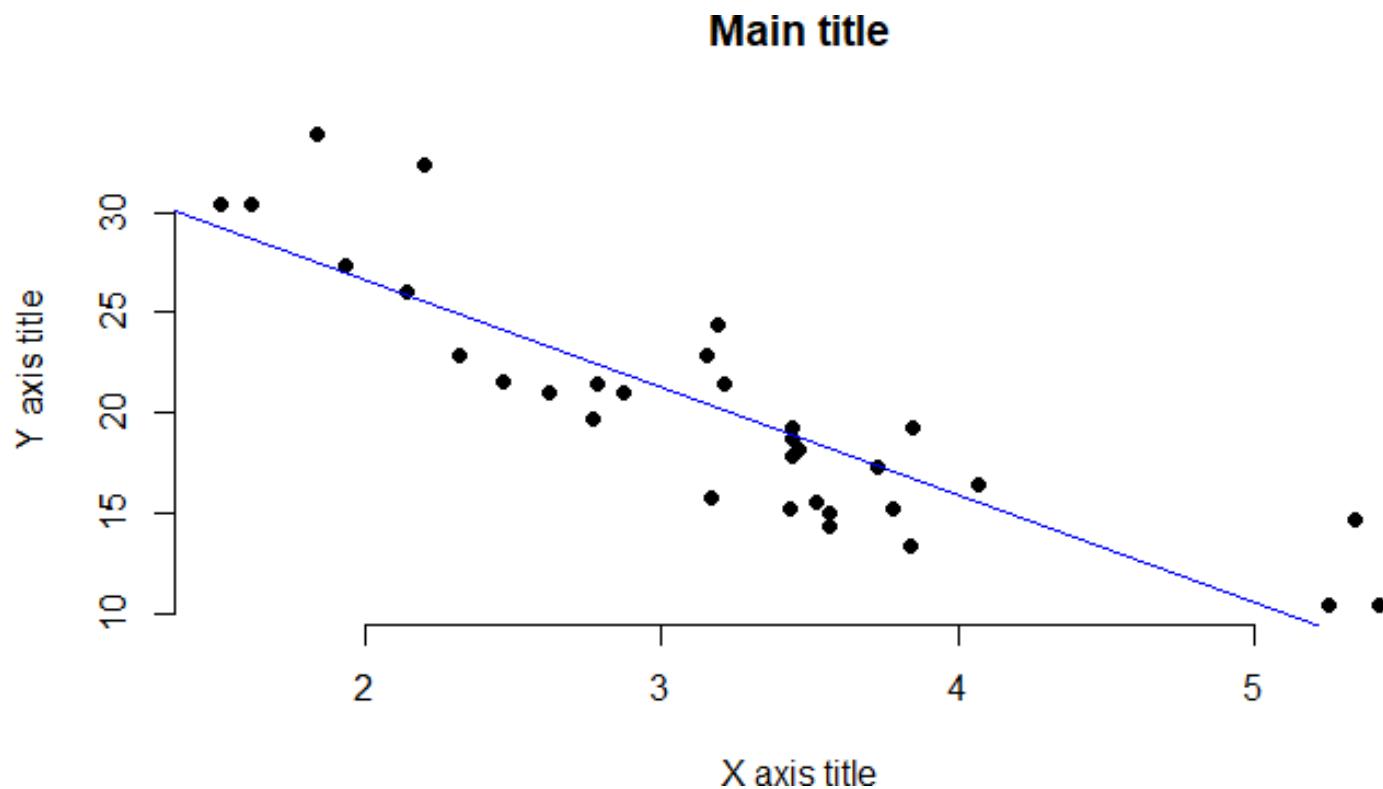
Hubungan antara dua variabel kuantitatif biasanya ditampilkan menggunakan scatterplots dan line graphs (grafik garis)

```
#scatter plot#
x <- mtcars$wt
y <- mtcars$mpg
# Plot with main and axis titles
# Change point shape (pch = 19) and remove frame.
plot(x, y, main = "Main title",
      xlab = "X axis title", ylab = "Y axis title",
      pch = 19, frame = FALSE)
# Add regression line
plot(x, y, main = "Main title",
      xlab = "X axis title", ylab = "Y axis title",
      pch = 19, frame = FALSE)
abline(lm(y ~ x, data = mtcars), col = "blue") #linear regression model
```

Output



Output



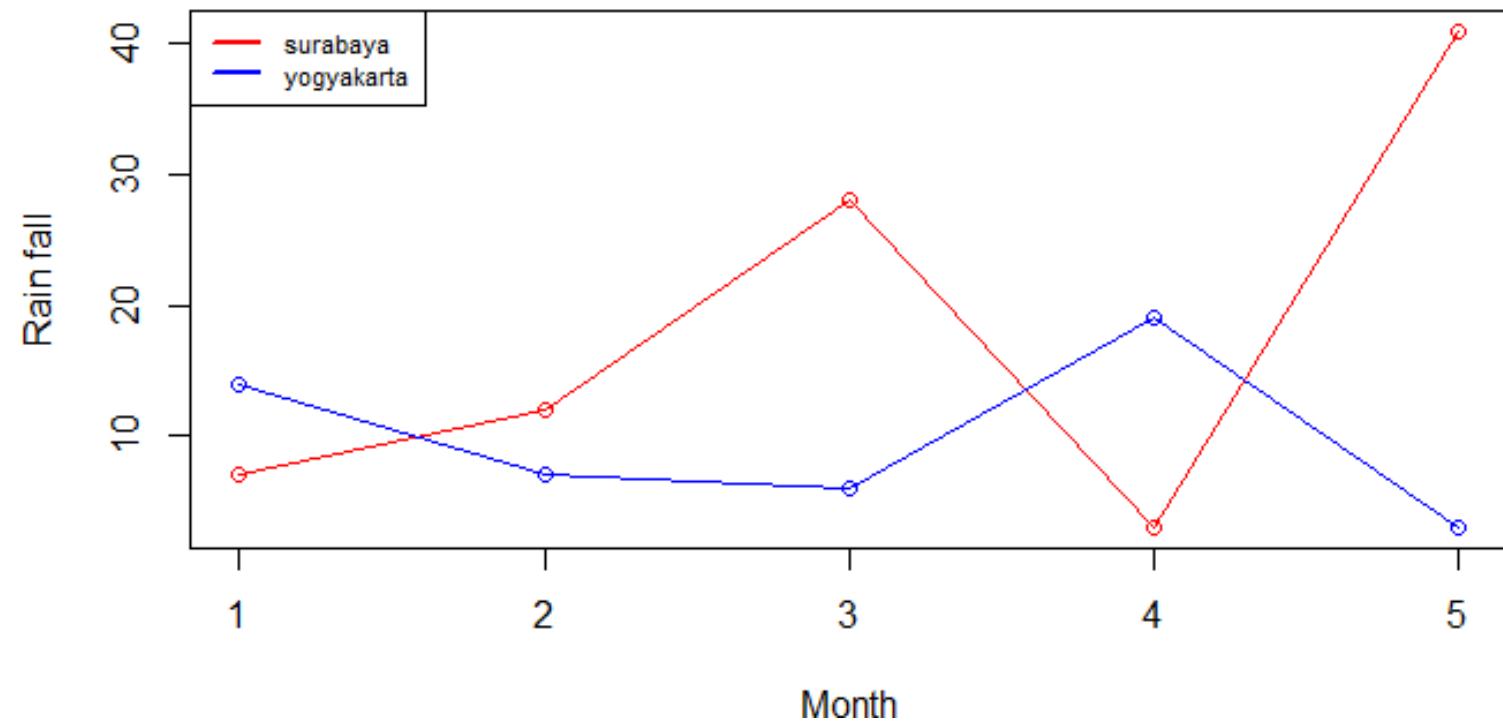
Grafik Bivariat – Line Graph

Hubungan antara dua variabel kuantitatif biasanya ditampilkan menggunakan scatterplots dan line graphs (grafik garis)

```
# Create the data for the chart.  
surabaya <- c(7,12,28,3,41)  
yogyakarta <- c(14,7,6,19,3)  
# Plot the bar chart.  
plot(surabaya,type = "o",col = "red", xlab = "Month", ylab = "Rain fall", main =  
"Rain fall chart")  
lines(yogyakarta, type = "o", col = "blue")  
legend("topleft", legend = c("surabaya", "yogyakarta"),  
cex = .7,  
lty = c(1, 1), # Line types  
col = c("red", "blue"), # Line colors  
lwd = 2 # Line width  
)
```

Output

Rain fall chart



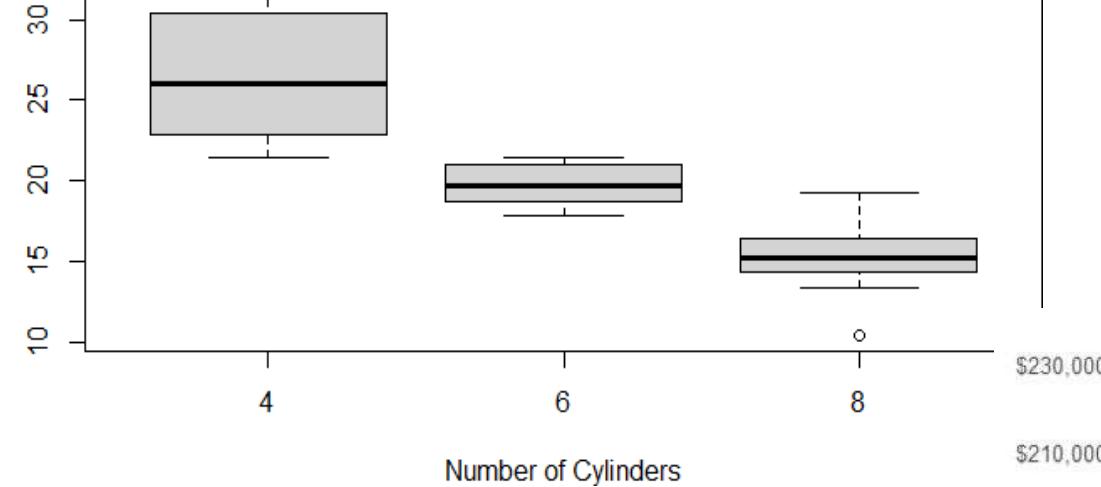
Grafik Bivariat – Box Plot

Box plot dapat menampilkan persentil (25%), median, dan persentil (75%) dari suatu distribusi. Box plot berguna untuk membandingkan beberapa kelompok (yaitu, tingkat variabel kategori) pada variabel numerik

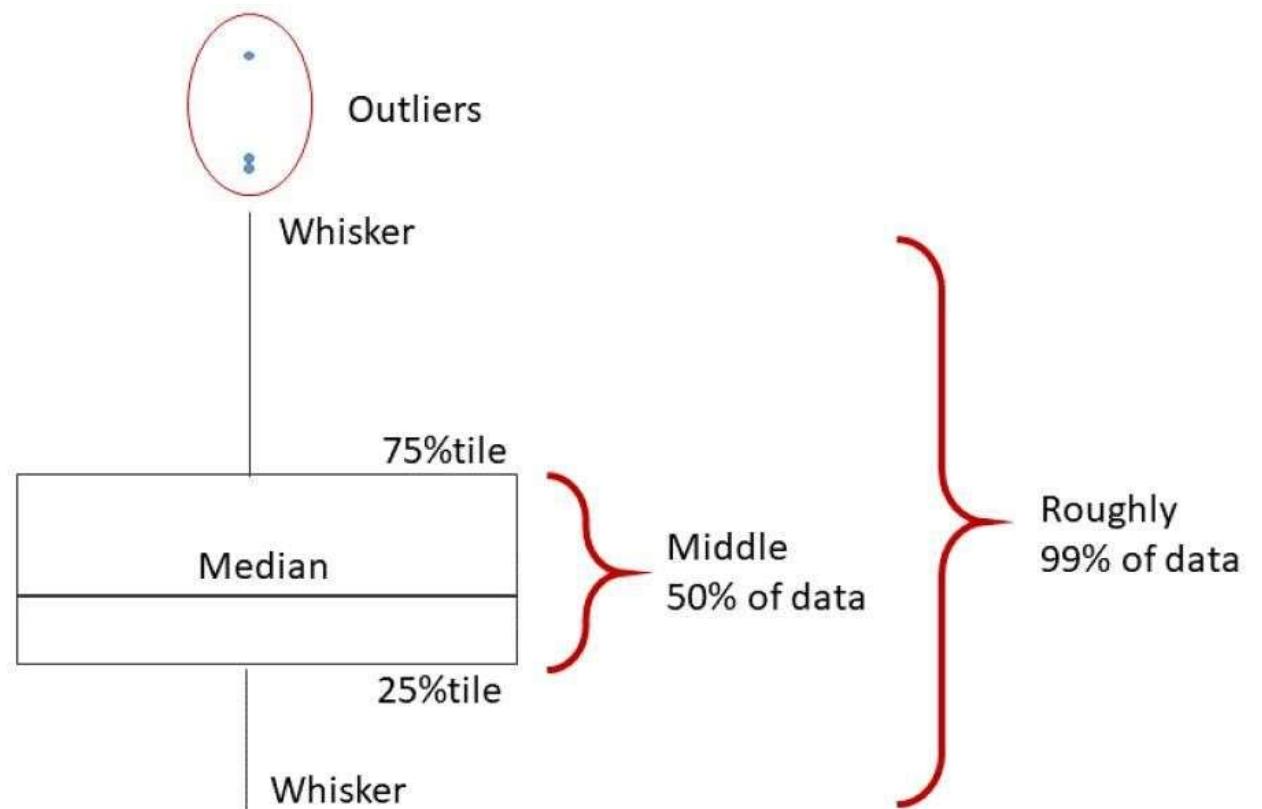
```
# Boxplot of MPG by Car Cylinders  
boxplot(mpg~cyl,data=mtcars, main="Car Milage Data",  
        xlab="Number of Cylinders", ylab="Miles Per Gallon")
```

Output

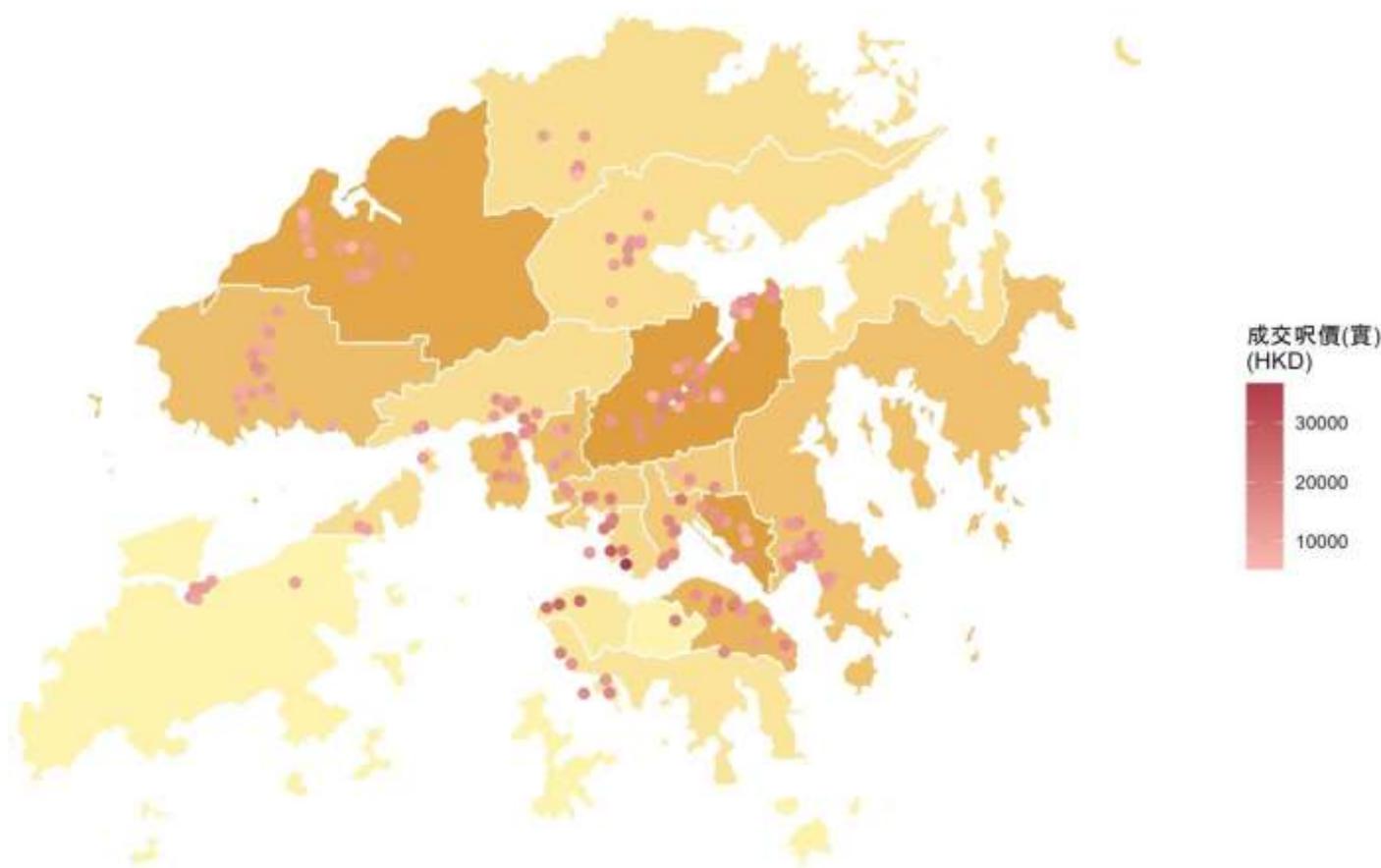
Miles Per Gallon



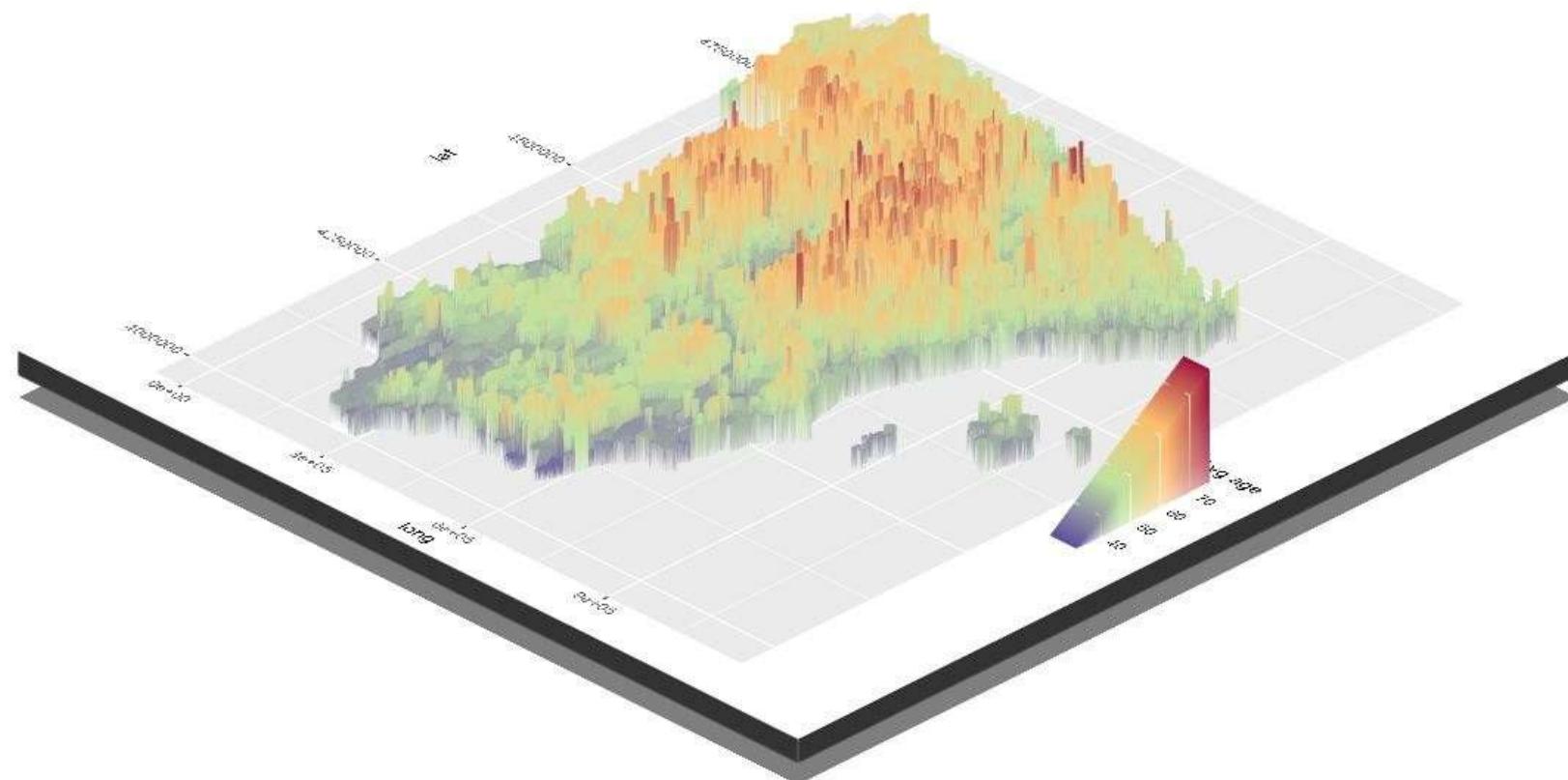
\$230,000
\$210,000
\$190,000
\$170,000
\$150,000
\$130,000
\$110,000
\$90,000
\$70,000
\$50,000



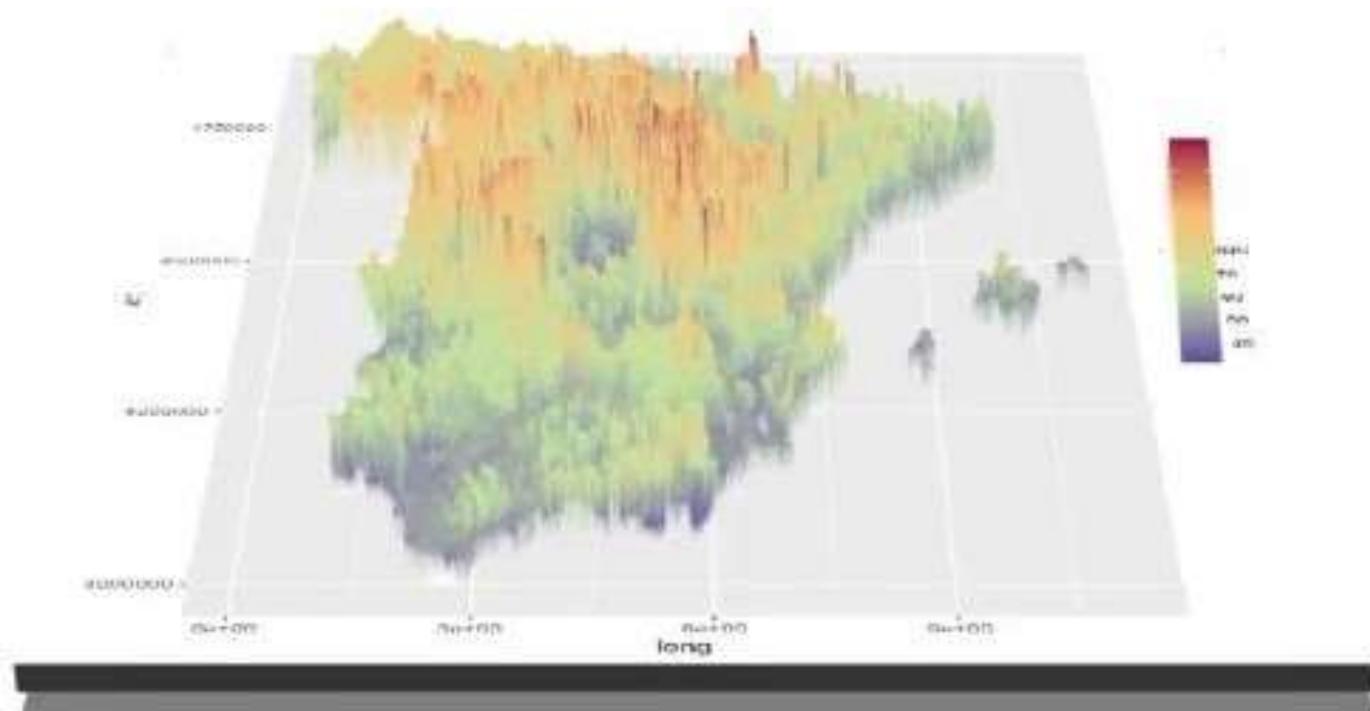
Import Map



Import Map (2)



Import Map Animated





Terima Kasih

Eksplorasi dan Visualisasi Data

Pertemuan 4:
Visualisasi berdasarkan Tipe Data

Pembahasan

1. Atribut
2. Tipe Data
3. Measure of Central Tendency and Dispersion
4. Skewness and Kurtosis
5. Visualisasi Atribut Kategorik dan Numerik / Kuantitatif

Latar Belakang

- Hal pertama yang harus dilakukan pada dataset apa pun yang anda miliki adalah mengenalinya terlebih dahulu.
- Selanjutnya, yang harus anda lakukan adalah mengidentifikasi tipe data yang anda miliki untuk setiap variabel, misalnya, apakah termasuk dalam tipe data numerik atau kategorik.
- Proses ini juga membantu anda untuk menentukan jenis visualisasi apa yang dapat dibuat dari data tersebut.
- Proses ini juga dilakukan tidak hanya untuk membiasakan diri anda dalam memahami semua data yang telah dikumpulkan, tetapi juga untuk mengurangi beban kerja selama proses analisis nanti.

Atribut

- Atribut mewakili **karakteristik atau ciri** dari objek data.
- Kata lain dari atribut adalah **dimension, feature, and variable**.
- Istilah **dimension** umumnya digunakan dalam data warehouse, **feature** untuk pembelajaran mesin (machine learning), dan istilah **variable** lebih digunakan dalam bidang statistika.
- Contoh atribut yang menjelaskan objek/entitas pelanggan:
 - ID pelanggan, nama dan alamat.
- Distribusi data yang melibatkan satu atribut (atau variabel) disebut sebagai **univariat**.
- Distribusi data yang melibatkan dua atribut disebut sebagai **bivariat** (secara umum 2 atribut atau lebih disebut **multivariat**).

Tipe Data

- Tipe data berdasarkan skala pengukuran data:

Jenis	Deskripsi	Contoh
Kategorik atau kualitatif	Nominal Skala pengukuran paling rendah dan angka-angka dalam skala nominal tidak mengukur besaran (tidak bermakna kuantitas) tetapi hanya sebagai penggolongan saja (agar dapat dibedakan).	Jenis kelamin, agama, negara, kode pos, warna.
	Ordinal Memiliki karakteristik skala nominal , tetapi data dikelompokkan ke dalam kategori/ nilai huruf, pangkat kelompok dengan adanya urutan/ peringkat yang bermakna. Misal: angka 1 memiliki nilai kepuasan: lebih tinggi dibandingkan 0. Namun, jarak 1 (sangat tidak puas), 2 antara 0 dan 1 tidak terdefinisi dengan jelas. (tidak Di skala ini, data tidak memiliki sifat selisih/ puas), 3 (cukup puas), 4 interval dan perbandingan rasio yang yang (puas), dan 5 (sangat puas) bermakna.	Tingkat pendidikan, ranking, Tingkat pendidikan, ranking, militer, tingkat

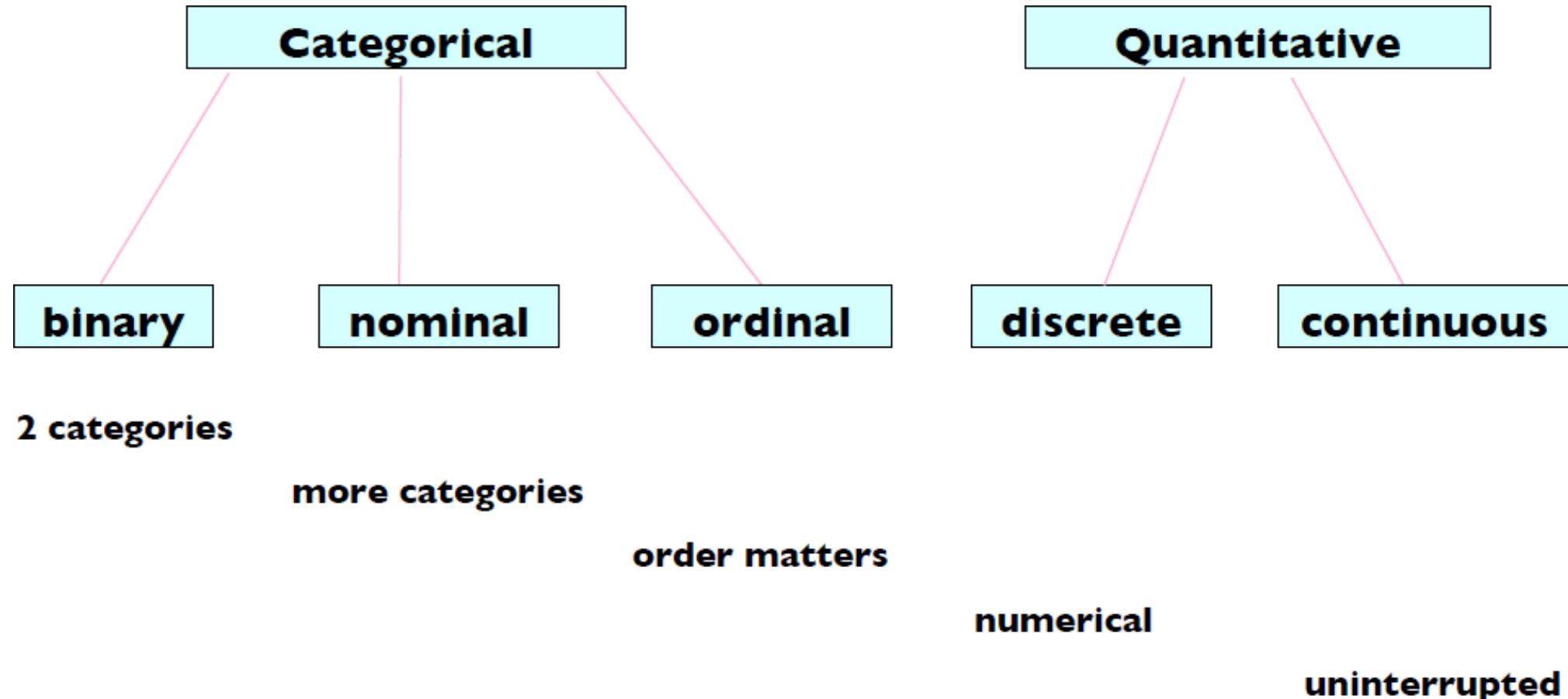
Tipe Data

- Tipe data berdasarkan skala pengukuran data:

Jenis	Deskripsi	Contoh
Numerik atau kuantitatif	Interval	Memiliki semua karakteristik skala ordinal , ditambah dengan adanya jarak antara nilai-nilai yang bermakna. Akan tetapi, data dengan skala interval tidak memiliki nol mutlak dan sifat perbandingan yang bermakna.
	Rasio	Memiliki semua karakteristik skala interval serta adanya nol mutlak . Pada skala ini, kita dapat melakukan perbandingan rasio yang bermakna antara nilai-nilai.

Pemahaman tentang skala pengukuran data sangat penting dalam statistika karena mempengaruhi jenis analisis yang dapat dilakukan

Tipe Data



Data Biner

- Data biner hanya memiliki dua kategori seperti ya/tidak, kepala/ekor, dan sebagainya.
- Saat menggunakan data biner dalam analisis, umumnya dikodekan ke dalam format 0 dan 1.
- Data biner sebenarnya adalah **tipe khusus dari data nominal**, dengan hanya dua kategori saja.

Data Nominal

- Data nominal juga sering disebut sebagai data kategorik, ini adalah data yang berisi beberapa kelompok (kategori) yang dapat diberikan nilai numerik tetapi **tidak memiliki urutan** alami.
- Misalnya, berbagai jenis kendaraan seperti sepeda, mobil, truk, kapal, pesawat, dapat diberi nomor 1 sampai 5 untuk memudahkan analisis.
- Namun, nilai-nilai itu sendiri tidak ada artinya dan tidak memiliki urutan, mobil memiliki nilai 1 lebih besar dari sepeda hanya berdasarkan penugasan dan bukan karena "lebih baik."

Data Ordinal

- Di sisi lain, data ordinal juga dapat diberikan nilai numerik tetapi **memiliki urutan alami**.
- Misalnya, reaksi terhadap bahan kimia bisa bernilai “tidak ada”, “ruam”, atau “melepuh” dan mereka dapat diberi nilai 1 hingga 3 dengan 3 jelas lebih parah dari 1, tetapi tidak harus 3 kali lebih parah.
- Contoh lain dari data ordinal adalah skala Likert pada kuesioner yang bernilai tidak setuju sampai sangat setuju.

Data Diskrit

- Data diskrit adalah jenis data kuantitatif yang mengandalkan aspek jumlah.
- Data diskrit hanya berisi nilai-nilai terbatas, mencakup nilai-nilai yang hanya **dapat dihitung** dalam bilangan bulat atau tidak dapat dipecah menjadi pecahan atau desimal.
- Contoh:
 - jumlah mahasiswa di UNAIR
 - jumlah mobil di tempat parkir
 - jumlah komputer di laboratorium komputer
 - jumlah hewan di kebun binatang
 - dll.

Data Kontinu

- Data kontinu adalah data yang diperoleh dari hasil **pengukuran**, yaitu data yang besarannya dapat menempati semua nilai yang ada di antara dua titik (nilai dari suatu rentang tertentu).
- Contoh:
 - jarak tempuh dari rumah A ke kampus C UNAIR (km)
 - hasil panen petani A (ton)
 - prestasi belajar mahasiswa (IPK)
 - dll.
- Contoh lainnya adalah data terkait usia, tinggi atau berat badan seseorang, suhu, waktu, uang, pendapatan dll.

Tipe Data

- Data juga dapat dikategorikan berdasarkan **sumber datanya** menjadi:
 - Data primer
 - Data sekunder
- Data berdasarkan **waktu pengumpulannya**:
 - Cross-sectional Data
 - Time Series Data
 - Data Longitudinal
 - Data Panel
- Data yang tergantung terhadap **lokasi**:
 - Data Spasial / Geospasial (lokasi geografis)

Data Primer

- Data primer merupakan jenis data yang dikumpulkan oleh peneliti langsung dari sumber utama.
- Data didapat secara langsung misalnya melalui eksperimen, sensor, observasi langsung, atau survey.
- Data primer bersifat real-time dan berupa data mentah.
- Interpretasi data primer biasanya lebih baik dan kuat dibandingkan dengan data sekunder.
- Hal ini disebabkan pengambilan data primer dilakukan secara spesifik untuk menjawab suatu hipotesis tertentu.

Data Sekunder

- Data sekunder adalah jenis data yang tidak diambil atau dikumpulkan secara langsung oleh peneliti.
- Data sekunder didapat misalnya dari database/datawarehouse atau berasal dari penelitian-penelitian sebelumnya.
- Data sekunder biasanya bersifat lampau dan terkadang sudah tidak relevan.
- Proses pengumpulannya lebih cepat, murah, dan mudah, tetapi tidak selalu ada (tersedia).
- Dari segi akurasinya data sekunder juga biasanya kurang akurat dan kurang bisa diandalkan dibandingkan dengan data primer.
- Hal ini disebabkan proses pengumpulan datanya tidak dilakukan secara langsung melainkan kita hanya tinggal menggunakan data yang sudah ada atau sudah diolah.
- Hal ini tentunya dapat menimbulkan bias yang cukup tinggi.

Central Tendency Measures

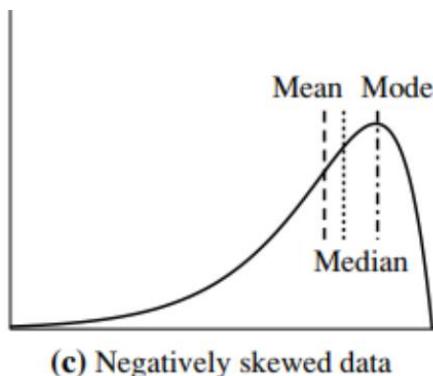
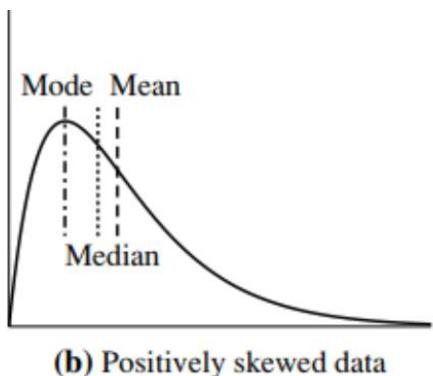
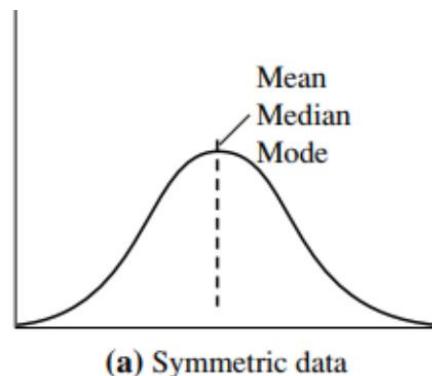
- Central tendency measures are statistics that represent the center or average of a data set.
- Central tendency measures can become a short summary of the data " or central value of a distribution.
- Which when follows a bell-shaped curve, the average or common data lies in the central part.
- The three most common measures of central tendency are:
 - **Mean** is the sum of all values in a data set divided by the number of values;
 - **Median** is the middle value when data is arranged in ascending or descending order;
 - **Mode** is the value that appears most frequently in a data set.
- When should we use mean, median, or mode?

Measures of Dispersion

- To measure how data spreads from the central tendency
- These measures provide insights into the distribution of data points and how they are scattered around the center.
- Common measures of dispersion include:
 - **Range** is the simplest measure of dispersion and is calculated as the difference between the maximum and minimum values in the data set;
 - **Interquartile Range (IQR)** is a measure of variability, based on dividing a data set into quartiles, is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) in a data set;
 - **Variance** measures the average squared deviation of each data point from the mean and It is calculated as the average of the squared differences between each data point and the mean;
 - **Standard deviation** is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.
- How to choose which measure of dispersion to use?

Skewness

- **Skewness** is the degree of distortion from the symmetrical bell curve or the normal distribution.
- Skewness measures the asymmetry of data distribution, namely the degree to which data tends to be skewed to one side of the middle value (mean, median, or mode).
- If skewness is **positive**, then the data tends to be skewed to the right (long tail on the right) and has a mean value that is greater than the median.
- If skewness is **negative**, then the data tends to be skewed to the left (long tail on the left) and has a mean value that is smaller than the median.
- A skewness value around zero indicates that the data has a **symmetrical** (not skewed) distribution.



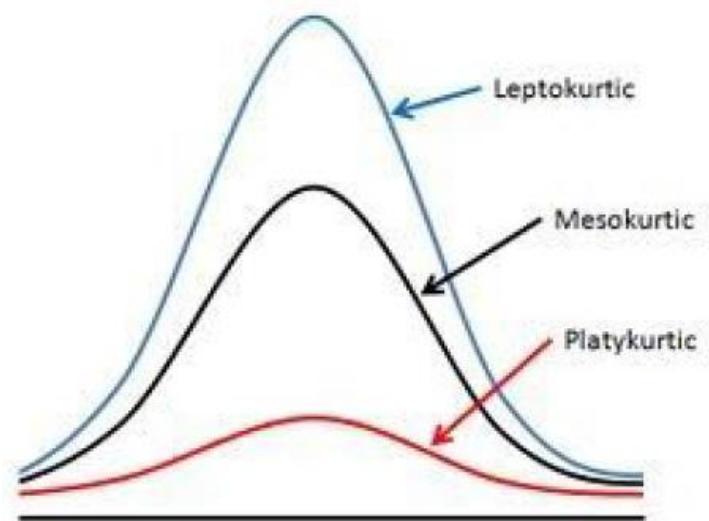
Kurtosis

- It is the sharpness of the **peak** of a frequency-distribution curve. It is actually the measure of outliers present in the distribution.
- **Positive kurtosis (leptokurtic)** indicates that the peak of the distribution is **sharper** (taller) than a normal distribution and the **tails are thicker**, which indicates the data has more outliers than would be expected in a normal distribution.
- **Negative kurtosis (platykurtic)** indicates that the peak of the distribution is **flatter and broader** (lower) than a normal distribution and the **tails are thinner**, indicating the data has fewer extreme values than would be expected in a normal distribution.
- **Zero kurtosis (mesokurtic)** indicates that the data distribution has the **same peak shape** as a normal distribution.

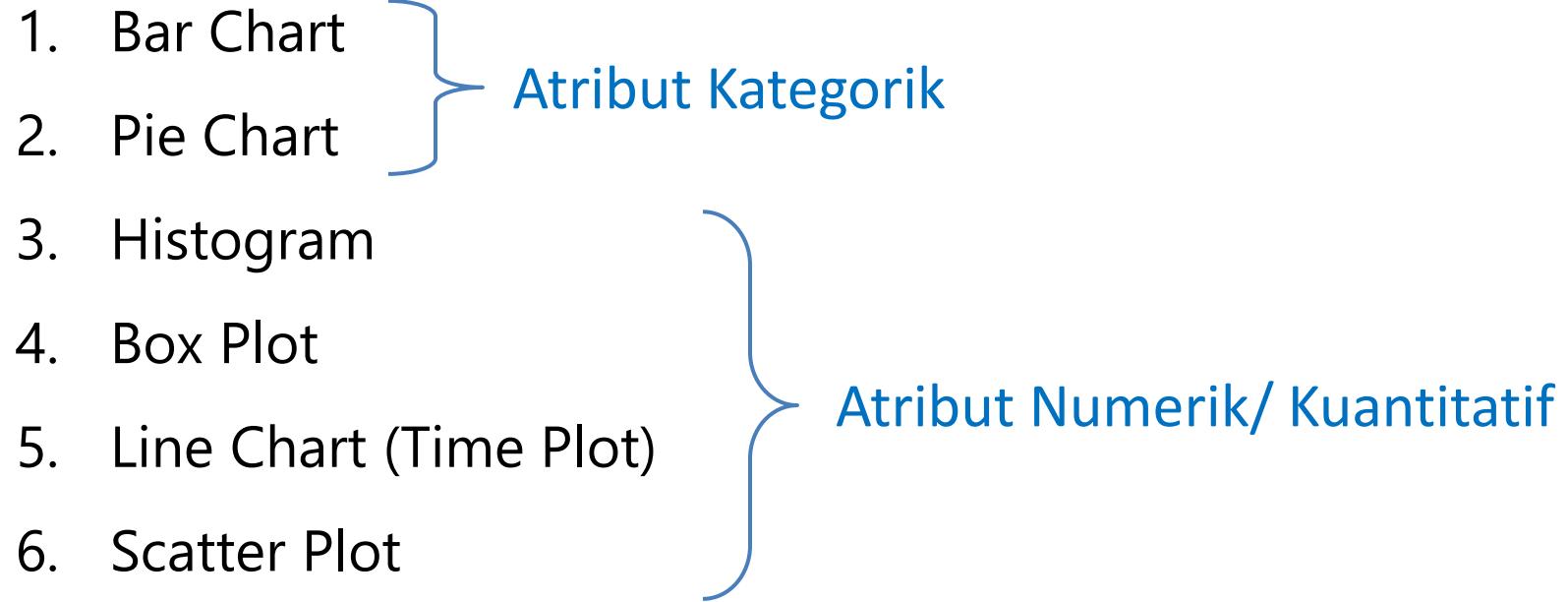
Leptokurtic → Absolut Kurtosis > 3
 Platykurtic → Absolut Kurtosis < 3
 Mesokurtic → Absolut Kurtosis = 3

$\text{Relative Kurtosis} = \text{Absolut Kurtosis} - 3$

Absolut Kurtosis distribusi normal = 3
 sehingga relative kurtosis = 0



Visualisasi dengan Grafik/ Plot

- 
- 1. Bar Chart
 - 2. Pie Chart
 - 3. Histogram
 - 4. Box Plot
 - 5. Line Chart (Time Plot)
 - 6. Scatter Plot

- Kita sering kali perlu menampilkan beberapa visualisasi secara bersamaan.
- Parameter `mflow` dari fungsi `par()` memungkinkan kita menggambar **banyak plot (multiple plots)** dalam satu kotak. Parameter ini mengambil vektor dari dua elemen yang menunjukkan jumlah baris dan kolom plot

Visualisasi dengan Grafik/ Plot

Univariate Graphs	Bivariate Graphs
<ul style="list-style-type: none"> Grafik univariat adalah jenis grafik yang digunakan untuk menggambarkan distribusi atau karakteristik dari satu variabel tunggal. Fokus utama adalah untuk memahami sebaran atau perilaku variabel tunggal tanpa mempertimbangkan hubungan dengan variabel lain. <u>Contoh</u>: histogram, diagram batang, diagram lingkaran, dan box plot yang menggambarkan satu variabel. 	<ul style="list-style-type: none"> Grafik bivariat adalah jenis grafik yang digunakan untuk menggambarkan hubungan antara dua variabel. Fokus utama dari grafik bivariat adalah untuk memvisualisasikan pola, korelasi, atau interaksi antara dua variabel. <u>Contoh</u>: scatter plot, matrix scatter plot, grafik garis dengan dua garis, dan heatmap yang menggambarkan hubungan atau perbandingan antara dua variabel.

Bivariate graphs vs multiple plots?



Thank You

SID205

Pedoman Penggunaan Warna

Untuk Visualisasi

(Week 6)

Mohammad Ghani, S.Si., M.Si., Ph.D.

mohammad.ghani@ftmm.unair.ac.id

Faculty of Advanced Technology and Multidiscipline

Universitas Airlangga

Warna di R

Di R, warna dapat ditentukan dengan nama (misalnya `col = "red"`) atau sebagai triplet RGB heksadesimal (seperti `col = "#FFCC00"`). Anda juga dapat menggunakan sistem warna lain seperti yang diambil dari paket `RColorBrewer` .

Nama Warna Bawaan di R (Built-in color names in R):

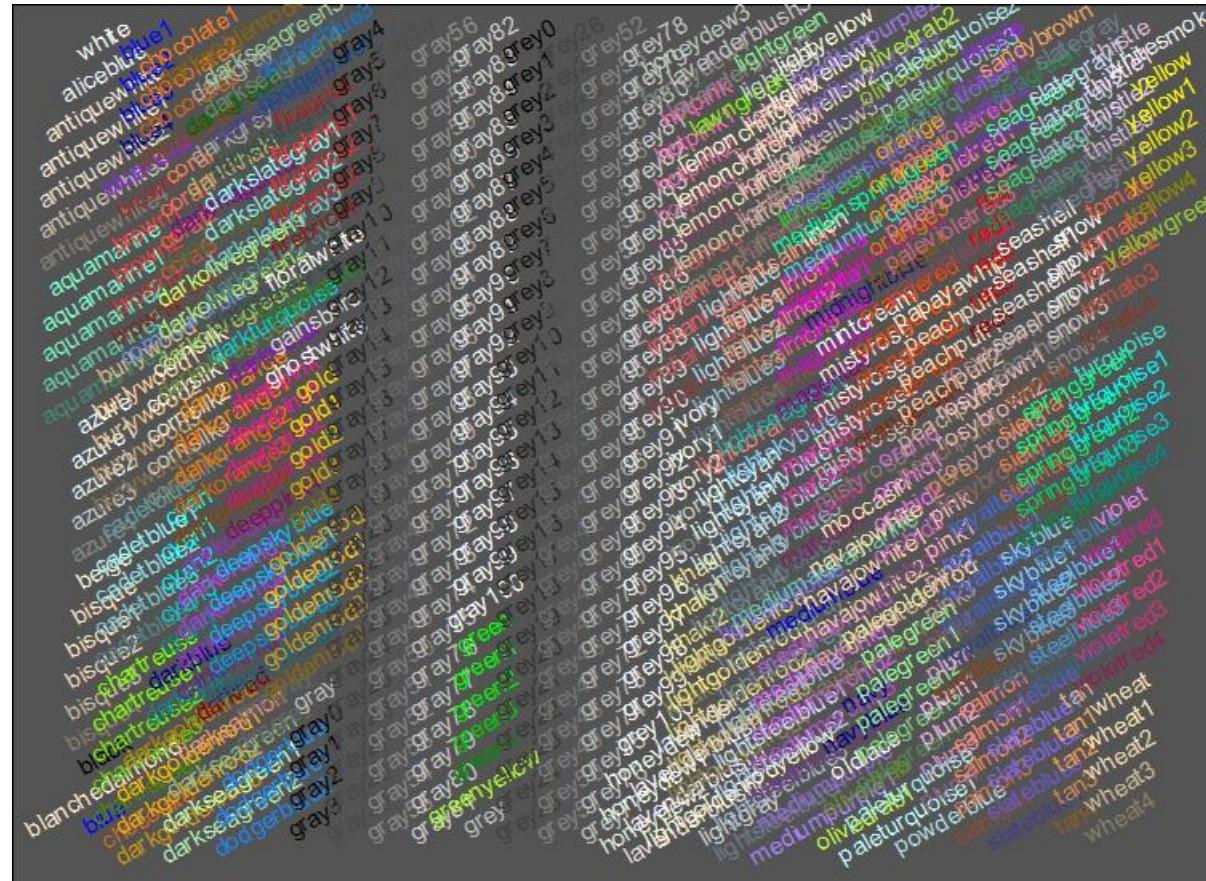
Kita akan gunakan fungsi R kustom berikut untuk menghasilkan plot nama warna yang tersedia di R:

Warna di R

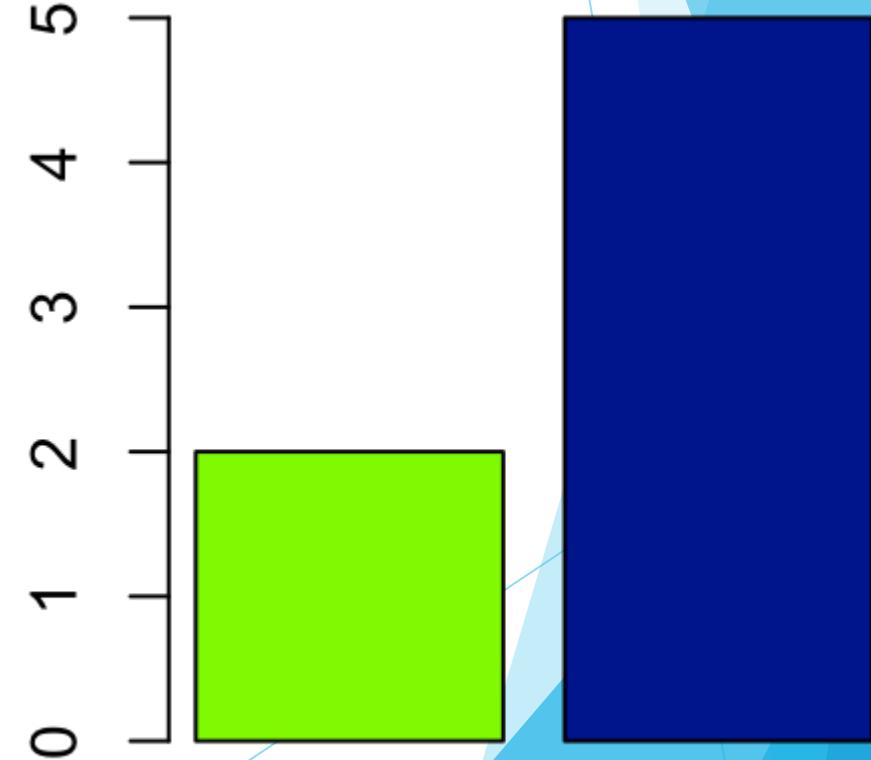


Warna di R

```
howCols(cl= colors(), bg="gray33", rot=30,
cex=0.75)
```

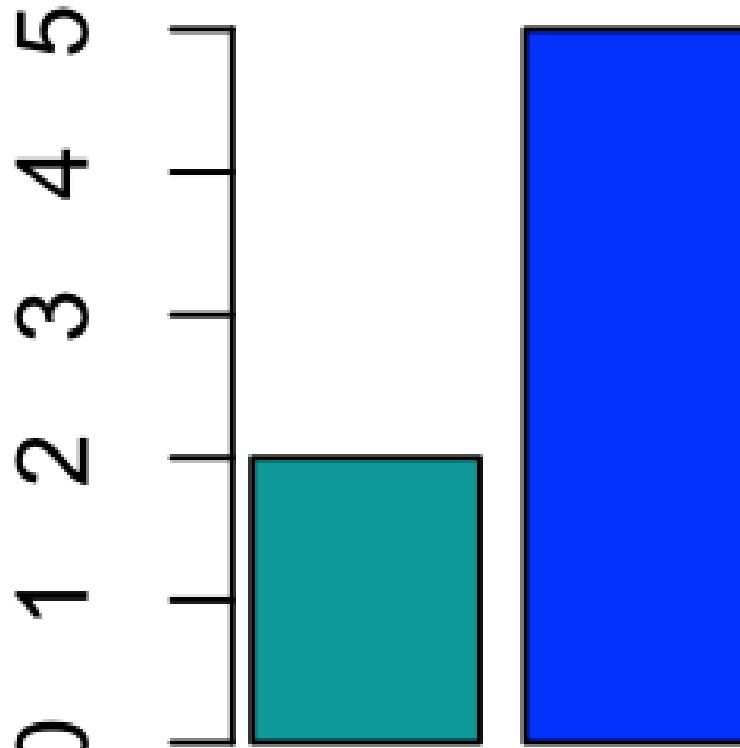


```
# Barplot using color names
barplot(c(2,5),
col=c("chartreuse", "blue4"))
```



Warna di R

```
# Barplot using hexadecimal color code
barplot(c(2,5), col=c("#009999",
 "#0000FF"))
```



(Source: <http://www.visibone.com>)

FFF FFF	CCC CCC	999 999	666 666	333 333	000 000	FFC C00	FF9 900	FF6 600	FF3 300													
99C C00						CC9 900	FFC C33	FFC C66	FF9 966	FF6 633	FF3 300											
CCF F00	CCF F33	333 300	666 600	999 900	CCC C00	FFF F00	CC9 933	CC6 633	330 000	660 000	990 000	CC0 000	FF0 000	FF3 366	FF6 033	FF0 033	FF0 033	FF0 033	FF0 033			
99F F00	CCF F66	99C C33	666 633	999 933	CCC C33	FFF F33	996 600	993 300	663 333	993 333	CC3 333	FF3 333	CC3 366	FF6 699	FF0 066	FF0 066	FF0 066	FF0 066	FF0 066			
66F F00	99F F66	66C C33	669 900	999 966	CCC C66	FFF F66	996 633	663 300	996 666	CC6 666	FF6 666	990 033	CC3 399	FF6 399	CC3 6CC	FF6 099	FF0 099	FF0 099	FF0 099			
33F F00	66F F33	339 900	66C C00	99F F33	CCC C99	FFF F99	CC9 966	CC6 600	CC9 999	FF9 999	FF3 399	CC0 066	FF0 066	FF3 3CC	FF0 0CC	FF0 0CC	FF0 0CC	FF0 0CC	FF0 0CC			
00C C00	33C C00	336 600	669 933	99C C66	CCF F99	FFF FCC	FFC C99	FFC 933	FF9 CCC	FF9 9CC	CC6 699	993 366	660 033	CC0 099	330 099	CC0 099	330 099	CC0 099	330 099			
33C C33	66C C66	00F F00	33F F33	66F F66	99F F99	CCF FCC											CC9 9CC	996 699	993 399	990 099	663 366	660 066
006 600	336 633	009 900	339 933	669 966	99C C99											FFC CFF	FF9 9FF	FF6 6FF	FF3 3FF	FF0 0FF	CC6 6CC	CC3 3CC
003 300	00C C33	006 633	339 966	66C C99	99F FCC	CCC FFF	339 9FF	99C CFF	99C CFF	996 9FF	CC9 6CC	996 399	663 066	330 066	990 0CC	CC0 0CC	330 0CC	CC0 0CC	330 0CC	CC0 0CC		
00F F33	33F F66	009 933	00C C66	33F F99	99F FFF	99F CCC	006 6CC	669 9CC	999 9FF	999 9CC	993 3FF	660 0CC	660 099	660 3FF	660 0FF	660 0FF	660 0FF	660 0FF	660 0FF			
00F F66	66F F99	009 C66	00C 966	66F FFF	66C CCC	66C 999	003 366	336 699	666 6FF	666 6CC	666 699	666 099	666 366	666 399	666 366	666 399	666 366	666 399	666 366			
00F F99	66F FCC	009 C99	00C FFF	66F CCC	99F 999	66C 666	336 699	336 399	336 3FF	336 3CC	336 399	336 366	336 399	336 366	336 399	336 366	336 399	336 366	336 399			
00F FCC	33F FCC	00F FFF	00C CCC	009 999	006 666	003 333	339 9CC	336 6CC	000 0FF	000 0CC	000 099	000 066	000 033	000 3FF	000 0FF	663 663	330 330	330 0CC	330 0CC			
00C C99						009 9CC	33C CFF	66C CFF	669 9FF	336 6FF	003 3CC											

Menggunakan palet RColorBrewer

Ada **3 jenis palet** : sekuensial, divergen, dan kualitatif.

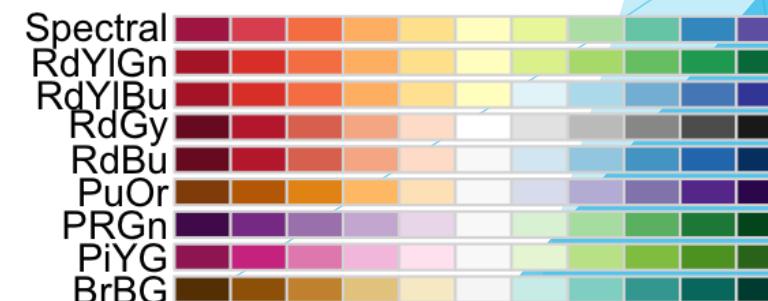
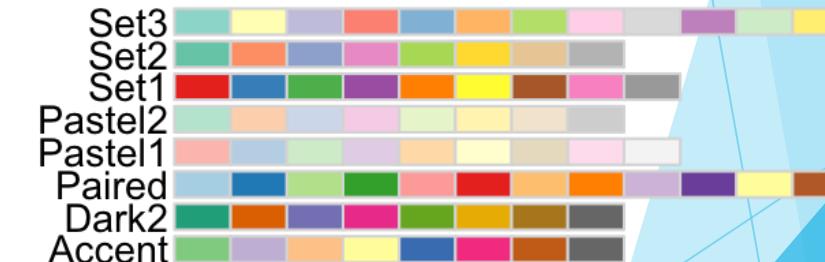
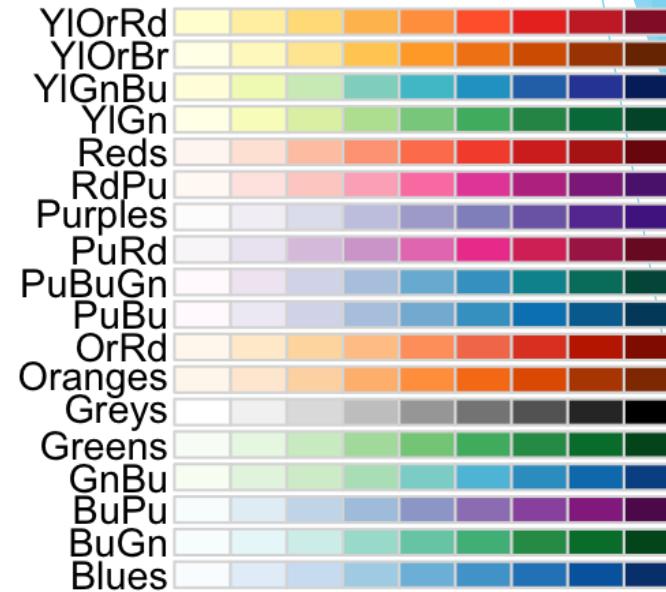
1. Palet sekuensial cocok untuk data terurut yang berkembang dari rendah ke tinggi (gradien). Nama paletnya adalah: Blues, BuGn, BuPu, GnBu, Greens, Greys, Oranges, OrRd, PuBu, PuBuGn, Purd, Purples, RdPu, Reds, YIGn, YIGnBu YIOrBr, YIOrRd.

2. Palet divergen memberikan penekanan yang sama pada nilai kritis rentang menengah dan ekstrem di kedua ujung rentang data. Palet divergen adalah : BrBG, PiYG, PRGn, PuOr, RdBu, RdGy, RdYIBu, RdYIGn, Spectral

3. Palet kualitatif paling cocok untuk mewakili data nominal atau kategoris. Mereka tidak menyiratkan perbedaan besaran antar kelompok. Nama paletnya adalah: Accent, Dark2, Paired, Pastel1, Pastel2, Set1, Set2, Set3

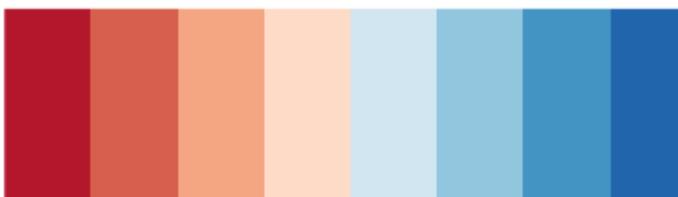
```
library("RColorBrewer")
```

```
display.brewer.all()
```



Menggunakan palet RColorBrewer

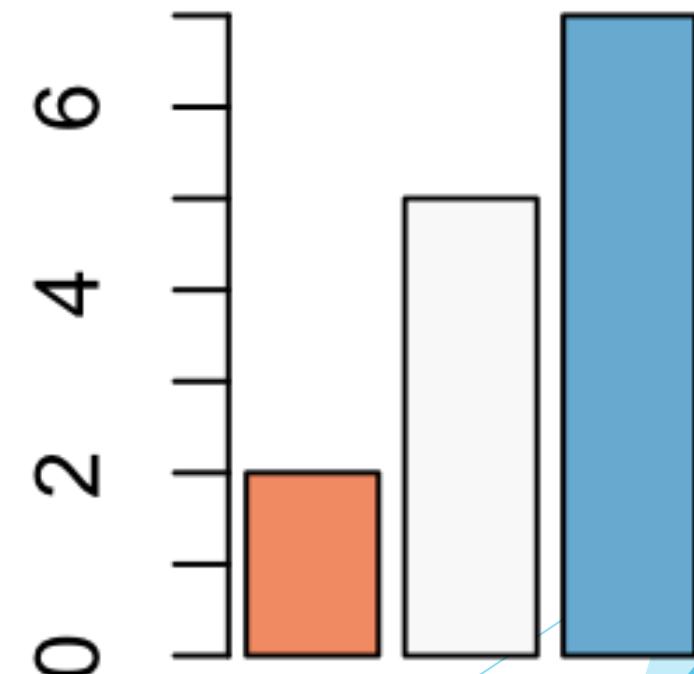
```
# View a single  
RColorBrewer palette by  
specifying its name  
display.brewer.pal(n = 8,  
name = 'RdBu')
```



```
# Hexadecimal color  
specification brewer.pal(n =  
8, name = "RdBu")
```

```
[1] "#B2182B" "#D6604D" "#F4A582"  
"#FDDBC7" "#D1E5F0" "#92C5DE" "#4393C3"  
"#2166AC"
```

```
# Barplot using RColorBrewer  
barplot(c(2,5,7),  
col=brewer.pal(n = 3, name =  
"RdBu"))
```

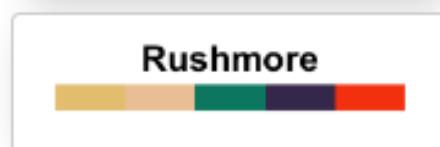
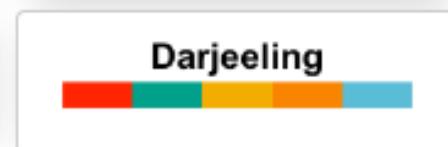
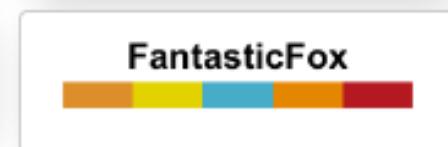
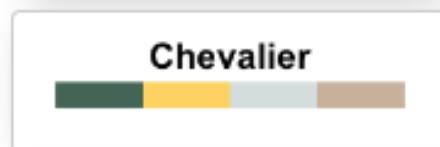
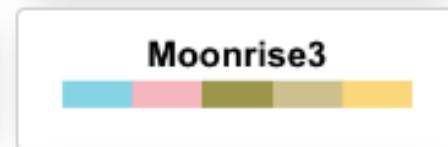
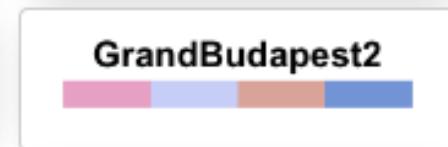
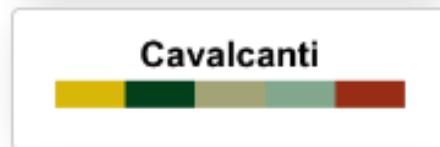


Gunakan palet warna Wes Anderson

```
# View a single Wes  
Anderson  
wes_palette("Royal2")
```

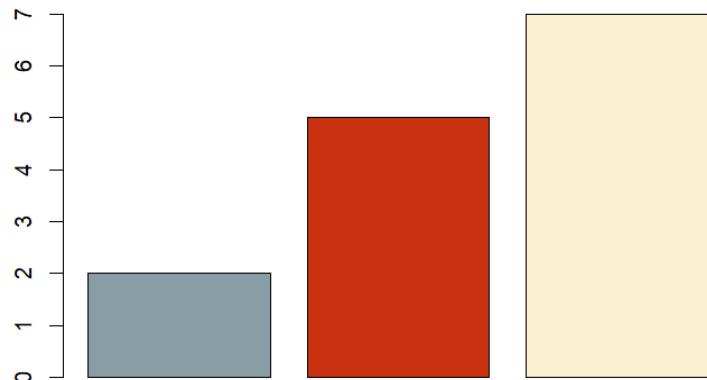
```
# View a single Wes  
Anderson  
wes_palette("Royall")
```

```
# View a single Wes  
Anderson  
wes_palette("Rushmore")
```

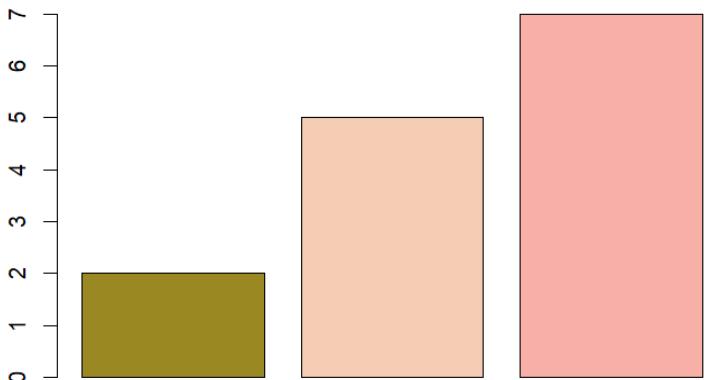


Gunakan palet warna Wes Anderson

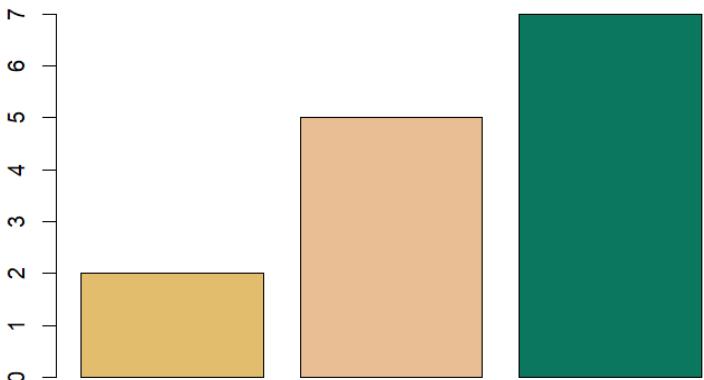
```
# simple barplot  
barplot(c(2,5,7),  
col=wes_palette(n=3,  
name="Royall1"))
```



```
# simple barplot  
barplot(c(2,5,7),  
col=wes_palette(n=3,  
name="Royal2"))
```



```
# simple barplot  
barplot(c(2,5,7),  
col=wes_palette(n=3,  
name="Rushmore"))
```

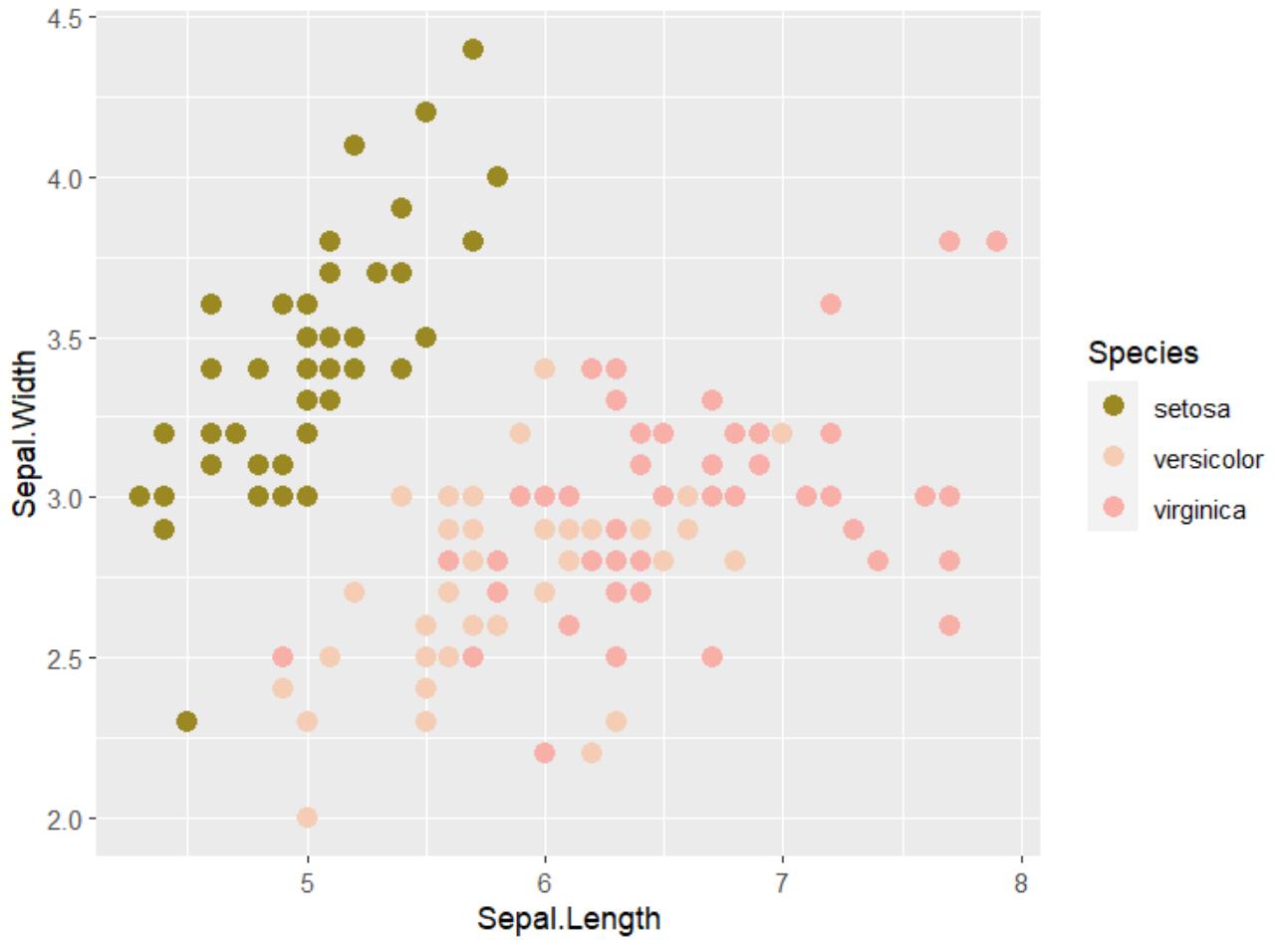


Gunakan palet warna Wes Anderson

```

library(ggplot2) ggplot(iris,
aes(Sepal.Length, Sepal.Width,
color = Species)) +
geom_point(size = 3) +
scale_color_manual(values =
wes_palette(n=3, name="Royal2"))

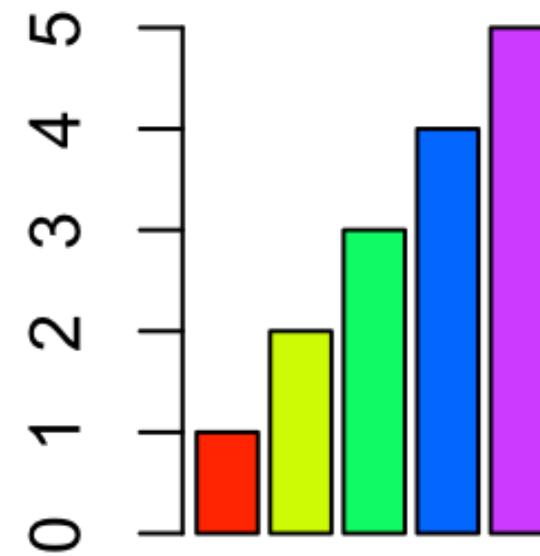
```



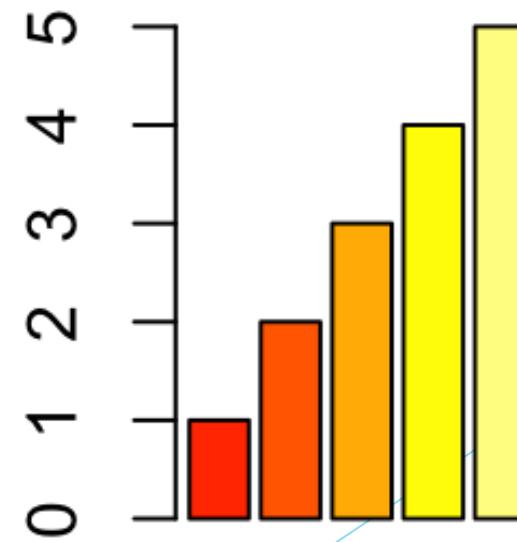
Sebuah vektor dari n warna yang bersebelahan

Anda juga dapat menghasilkan vektor n warna yang bersebelahan menggunakan fungsi **rainbow(n)** , **heat.colors(n)** , **terrain.colors(n)** , **topo.colors(n)** , dan **cm.colors(n)**

```
# Use rainbow colors barplot(1:5,  
col=rainbow(5))
```

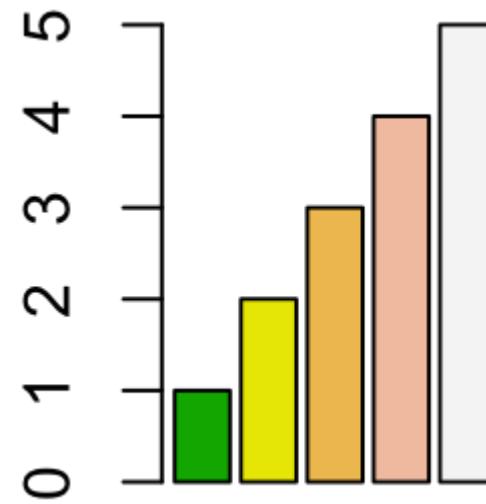


```
# Use heat.colors barplot(1:5,  
col=heat.colors(5))
```

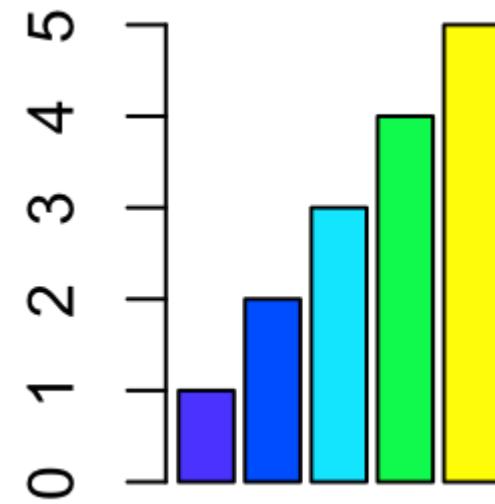


Sebuah vektor dari n warna yang bersebelahan

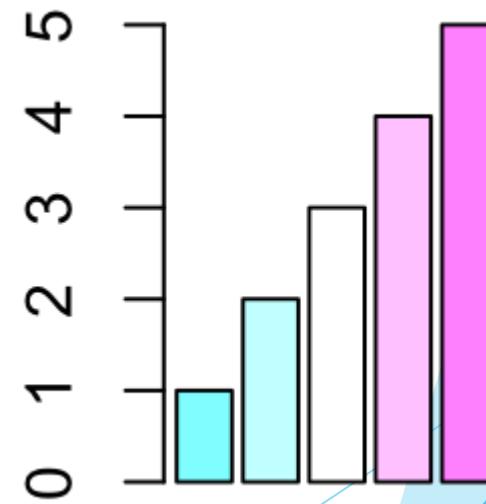
```
# Use terrain.colors  
barplot(1:5,  
col=terrain.colors(5))
```



```
# Use topo.colors  
barplot(1:5,  
col=topo.colors(5))
```

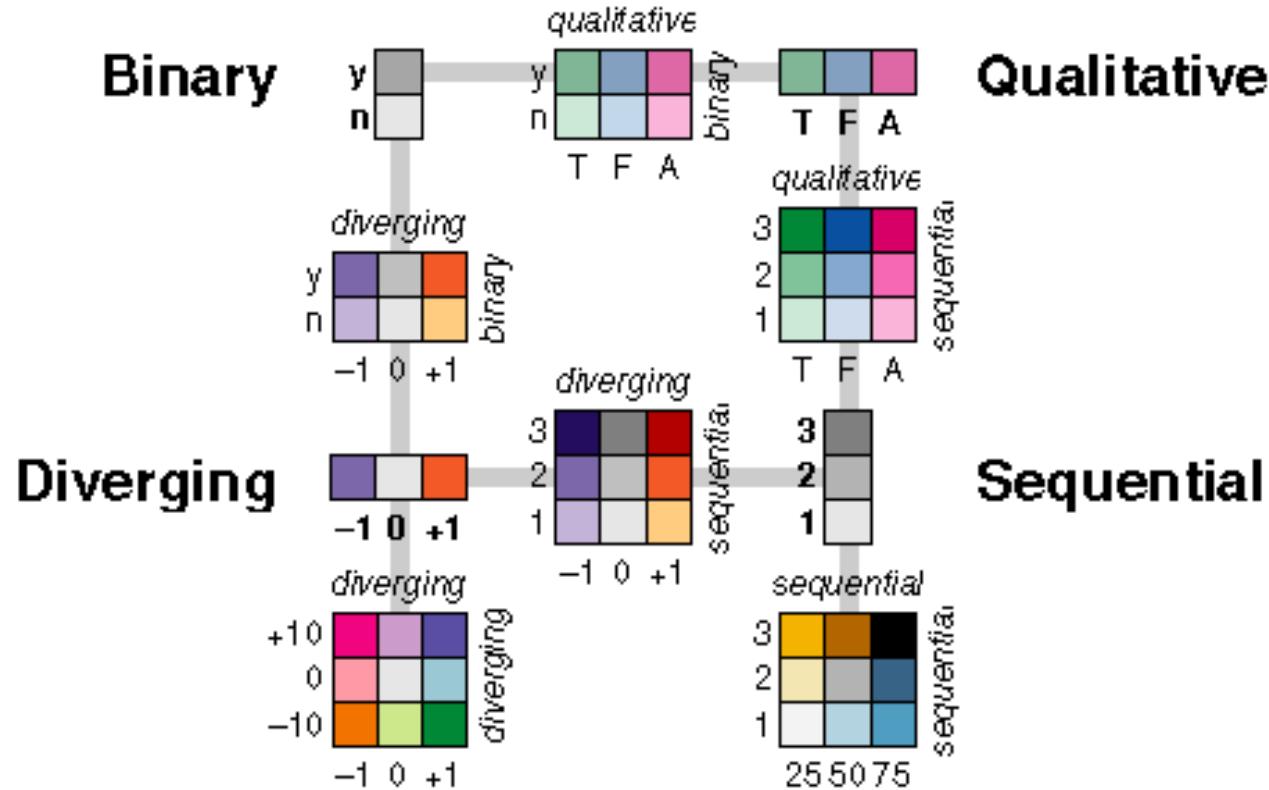


```
# Use cm.colors  
barplot(1:5,  
col=cm.colors(5))
```



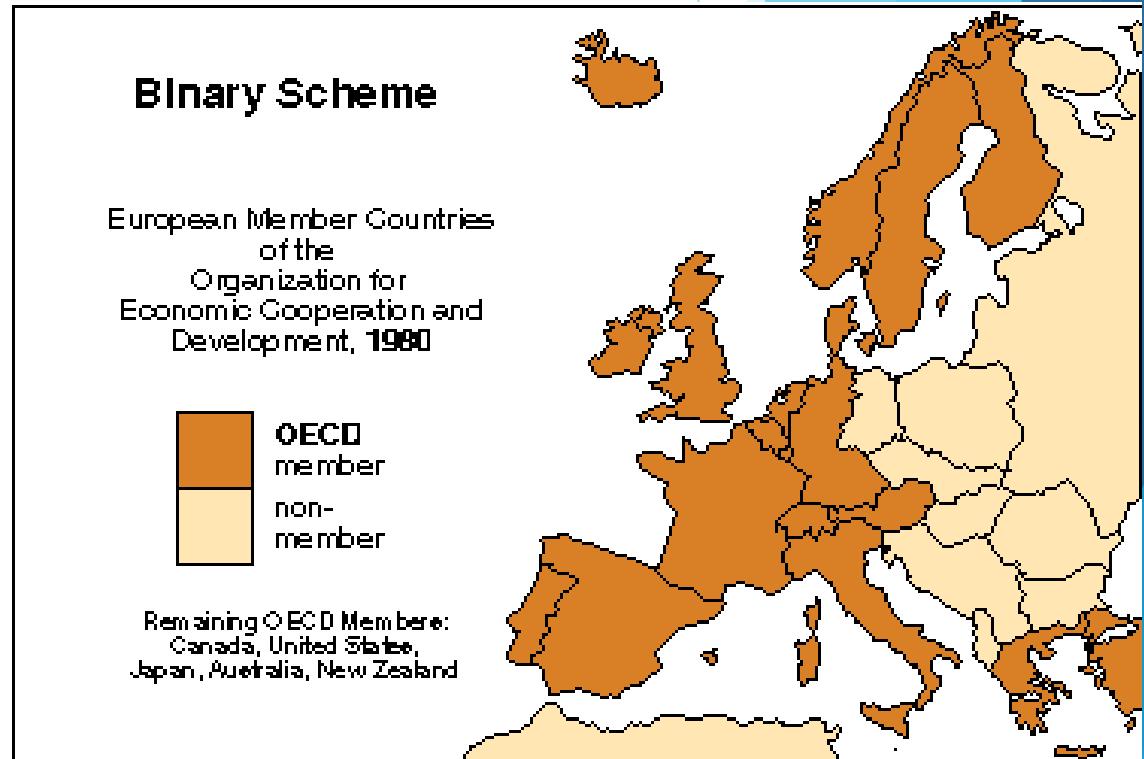
Jenis dan Kombinasi Skema Warna: Gambaran Umum

Pilih skema warna yang menarik di bawah ini untuk melihat contoh penggunaannya.



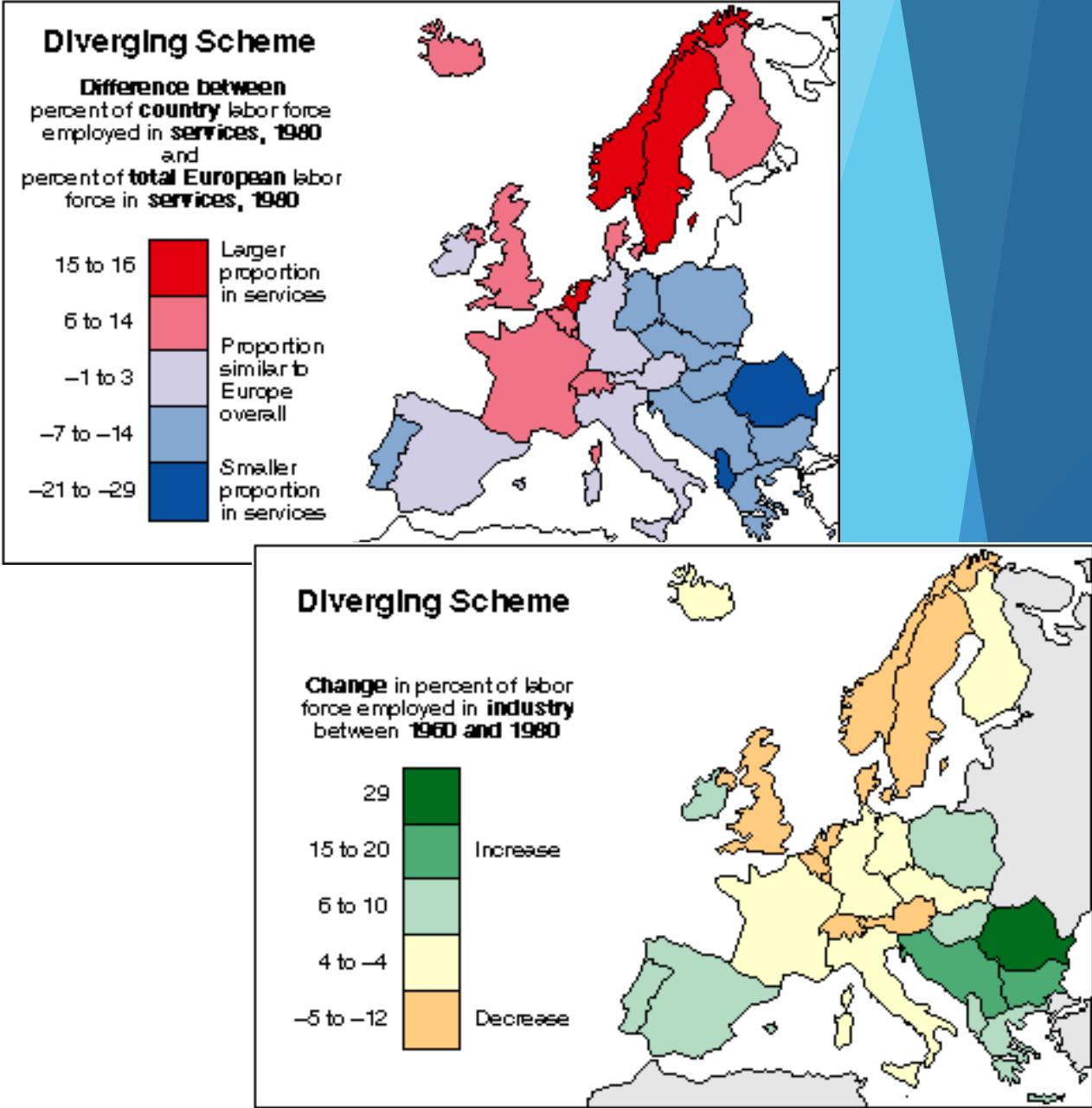
Skema Warna Biner

Skema biner menunjukkan perbedaan nominal yang hanya dibagi menjadi dua kategori. Perbedaan persepsi utama antara dua kategori skema biner mungkin merupakan langkah ringan, tidak seperti penggunaan rona untuk variabel kualitatif multi-nilai. Daerah perkotaan yang tergabung dan tidak tergabung diwakili dengan baik oleh skema warna biner.



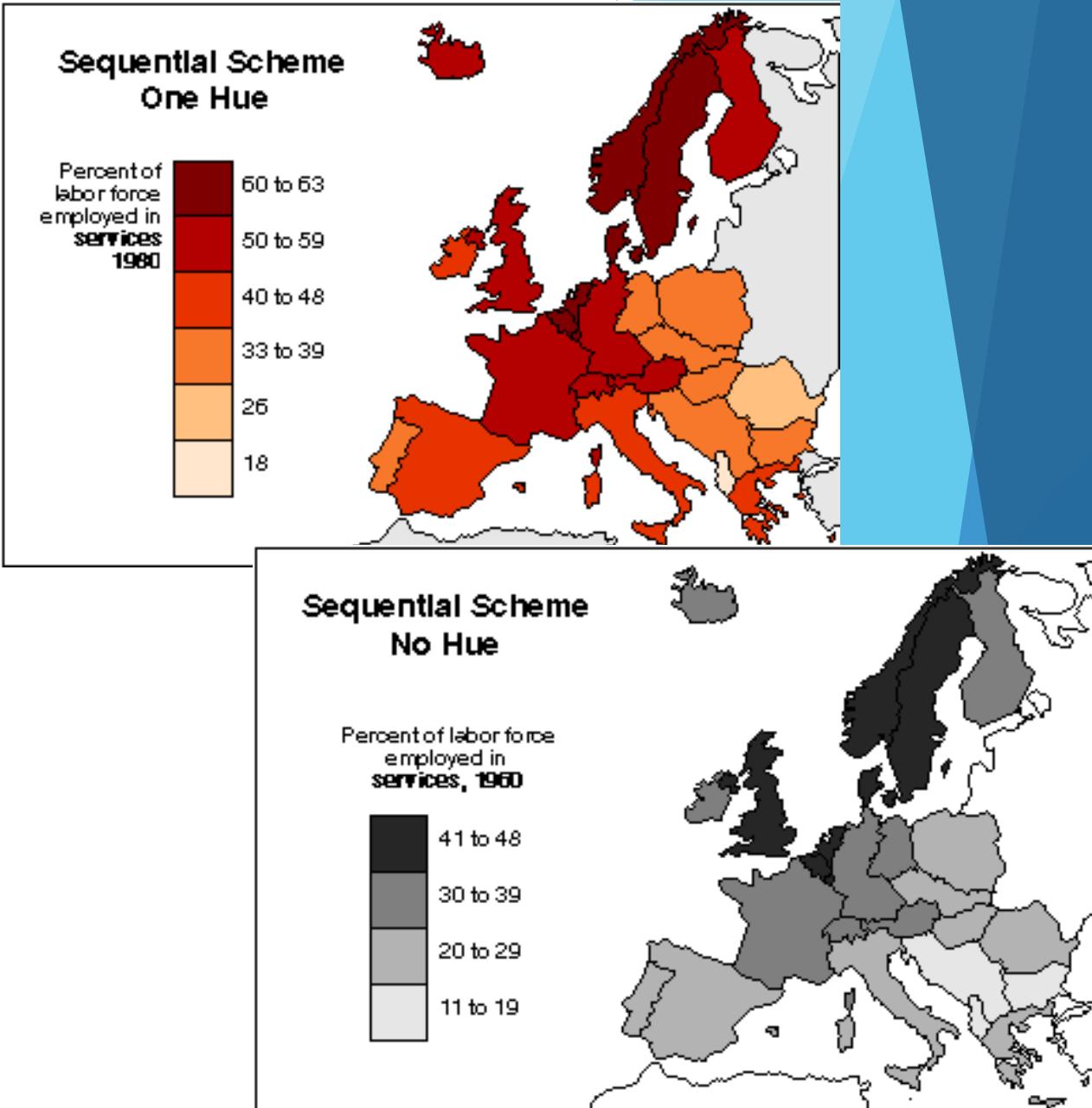
Skema Warna Divergen

Skema divergen memungkinkan penekanan tampilan data kuantitatif menjadi progresi keluar dari titik tengah kritis dari rentang data. Skema divergen tipikal memasangkan skema sekuensial berdasarkan dua rona berbeda sehingga mereka menyimpang dari warna terang bersama, untuk titik tengah kritis, menuju warna gelap rona berbeda di setiap ekstrem. Penyimpangan di atas dan di bawah rata-rata angka kematian akibat penyakit, misalnya, terwakili dengan baik oleh skema warna yang berbeda.



Skema Warna Sekuensial

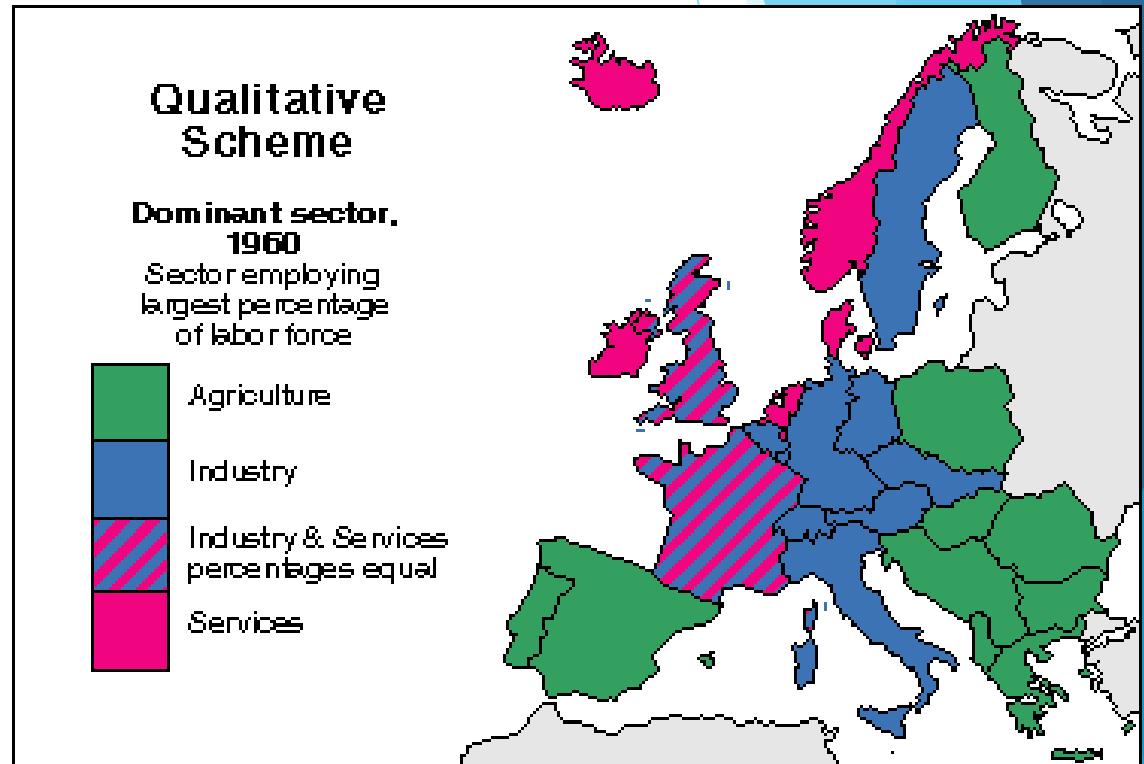
Kelas data sekuensial diatur secara logis dari tinggi ke rendah, dan urutan kategori yang bertahap ini harus diwakili oleh langkah-langkah ringan sekuensial. Nilai data rendah biasanya diwakili oleh warna terang dan nilai tinggi diwakili oleh warna gelap. Transisi antar rona dapat digunakan dalam skema berurutan, tetapi progresi terang-ke-gelap harus mendominasi skema. Kategori kemiringan medan atau kepadatan penduduk, misalnya, terwakili dengan baik oleh skema warna berurutan.



Skema Warna Kualitatif

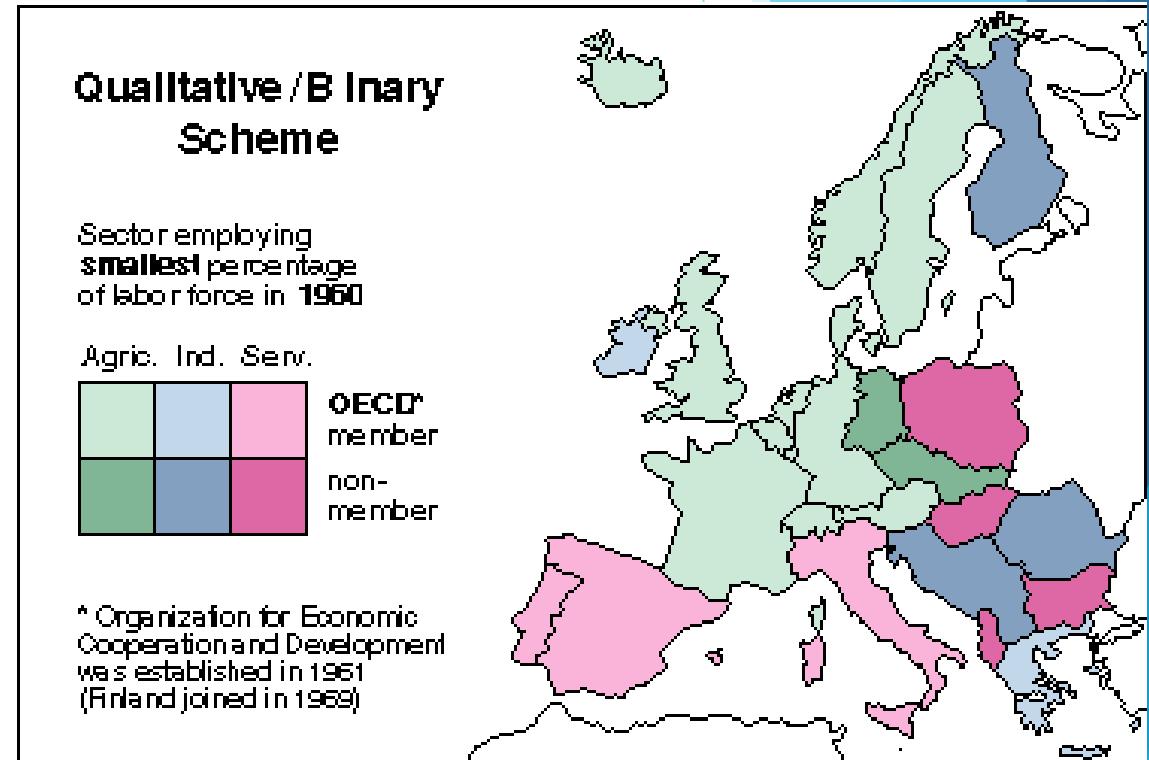
Skema kualitatif menggunakan perbedaan warna untuk mewakili perbedaan nominal, atau perbedaan jenis. Kecerahan warna yang digunakan untuk kategori kualitatif harus serupa tetapi tidak sama.

Tetapkan warna paling terang, paling gelap, dan paling jenuh dalam skema ke kategori yang menjamin penekanan pada peta. Data tentang penggunaan lahan lahan, misalnya, terwakili dengan baik oleh skema warna kualitatif.



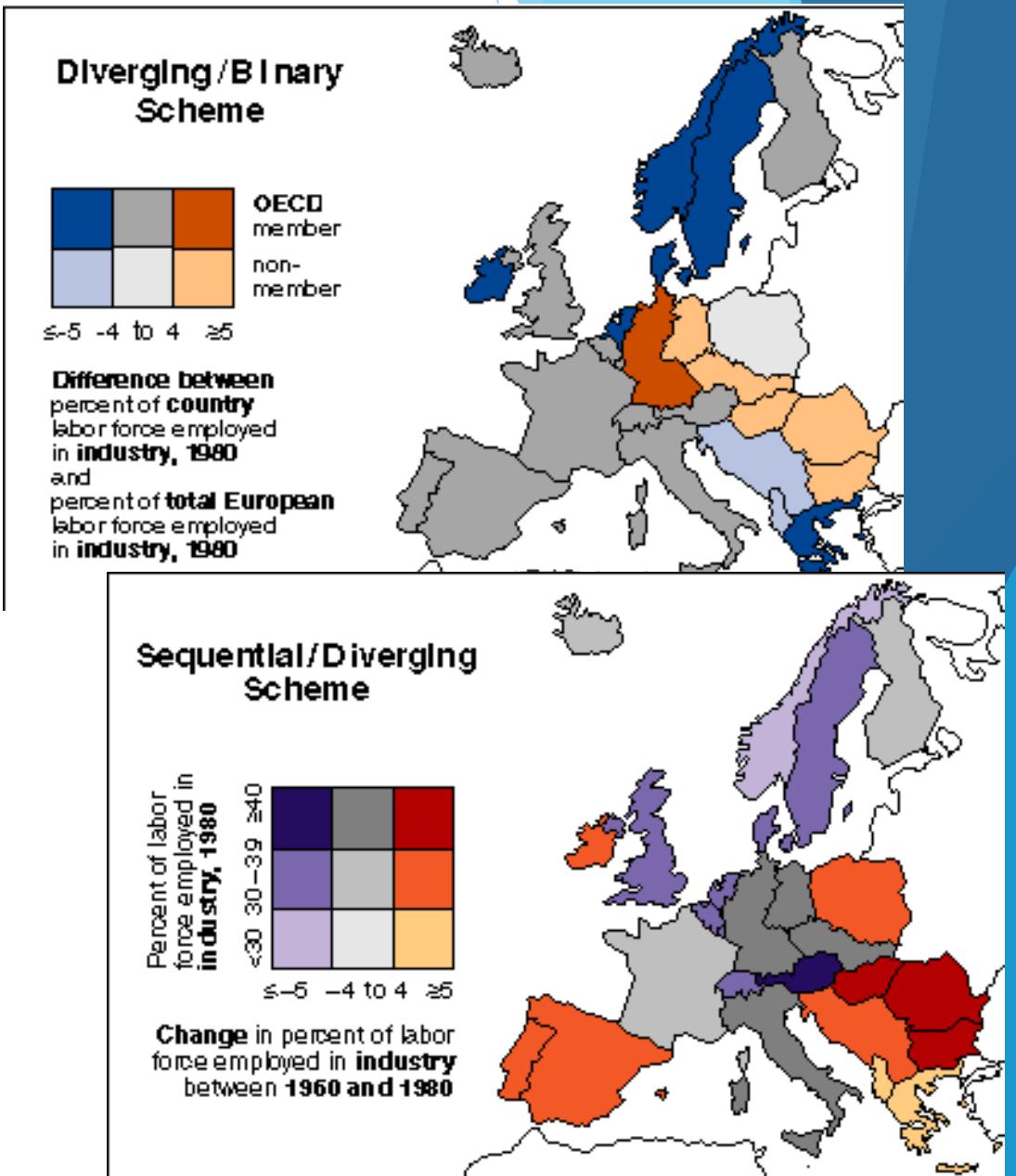
Skema Warna Kualitatif-Biner

Dalam skema kualitatif/biner, versi terang dan gelap dari setiap rona variabel kualitatif sesuai dengan kategori variabel biner. Skema biner/biner adalah bagian dari skema kualitatif/biner dengan satu perbedaan biner diwakili oleh perbedaan rona dan yang lainnya oleh perbedaan ringan. Peta vegetasi multi-warna (kualitatif) dengan warna yang lebih gelap untuk vegetasi di lahan publik dan warna yang lebih terang untuk vegetasi di lahan pribadi (biner) sangat cocok dengan skema warna kualitatif/biner.



Skema Warna Divergen-Biner & Divergen-Sekuensial

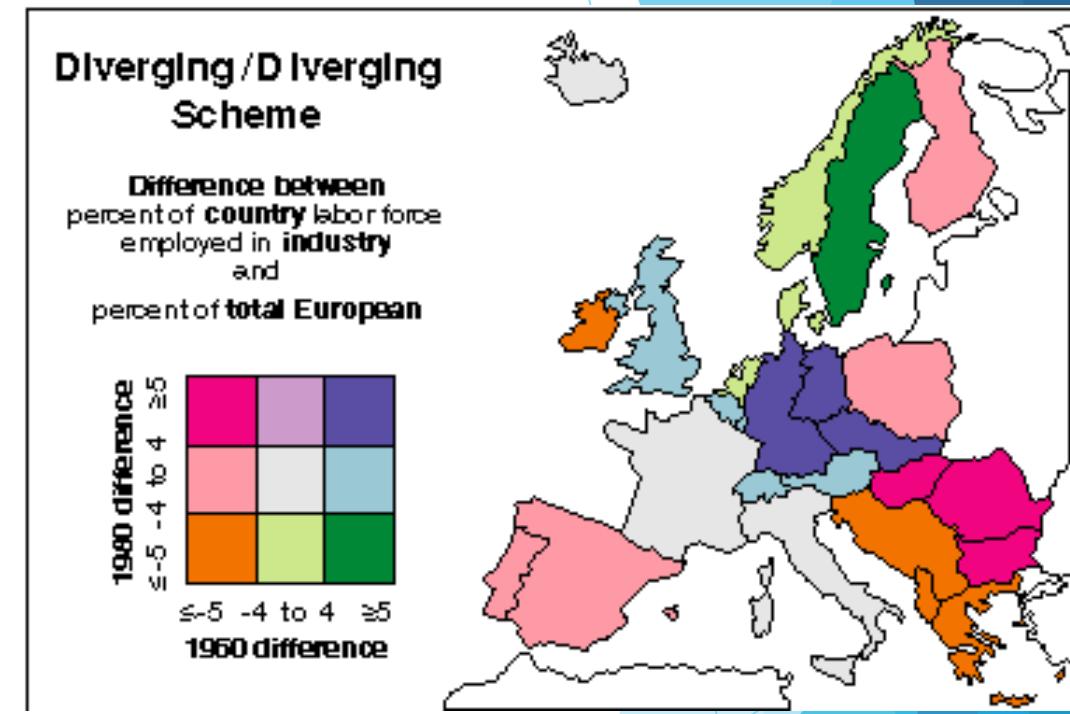
Skema divergen/biner dan divergen/sekuensial memiliki karakteristik persepsi yang sama. Keberhasilan skema bergantung pada rentang kontras besar yang tersedia dalam dimensi ringan. Langkah ringan yang besar digunakan untuk variabel biner atau sekuensial. Langkah ringan yang lebih kecil, yang didukung oleh perubahan warna, mewakili komponen divergen skema dalam setiap langkah kecerahan besar dari variabel perbandingan. Misalnya, data tentang tingkat kanker di atas dan di bawah tingkat rata-rata (diverging) dan tingkat polusi udara (sekuensial) terwakili dengan baik oleh skema warna divergen/urutan.



Skema Warna Divergen-Divergen

Skema divergen/divergen adalah satu-satunya skema dua variabel yang berangkat dari gagasan overlay langsung skema komponen satu variabel. Tempatkan rona agak gelap yang berbeda di masing-masing dari empat sudut legenda. Keempat rona ini mewakili kategori yang ekstrim untuk kedua variabel. Tempatkan warna yang sangat terang atau putih di tengah legenda, buat warna terang yang sesuai untuk kelas yang berisi nilai kritis atau titik tengah dari kedua variabel.

Warna yang tersisa lebih terang dari sudut, karena mengandung titik tengah dari salah satu dari dua variabel, dan mereka adalah warna transisi yang terletak di antara warna yang berdekatan. Lingkaran warna pada dasarnya direntangkan di sekeliling legenda dan kecerahannya disesuaikan sebagai respons terhadap nilai kritis dalam rentang data kedua variabel. Daerah di atas dan di bawah garis kemiskinan pada tahun 1960 dan 1990, misalnya, terwakili dengan baik oleh skema warna yang berbeda.



Skema Warna Sekuensial-Sekuensial

Skema sekuensial/sekuensial adalah campuran logis dari semua kombinasi warna dalam dua skema sekuensial. Dengan demikian, skema didasarkan pada dua warna. Campuran rona dapat membentuk rona ketiga (urutan magenta dan cyan menghasilkan berbagai transisi ungu-biru). Jika dua rona yang disilangkan merupakan komplemen perkiraan, campurannya menghasilkan diagonal abu-abu netral dan warna transisi desaturasi. Perbedaan ringan yang sistematis di seluruh skema adalah penting; tidak bergantung pada rona untuk menyampaikan pesan besarnya. Gunakan transisi rona untuk menunjukkan perbedaan dalam proporsi dua variabel yang dipetakan. Misalnya, data pencapaian pendidikan yang disilangkan dengan kategori tingkat kejahatan terwakili dengan baik oleh skema warna berurutan/berurutan.

