



Project Web Scraping

Manajemen Data Statistika

Eka Dicky Darmawan Yanuari
G1501231088

Web Scraping dilakukan pada website Carmudi Indonesia
(<https://www.carmudi.co.id>)



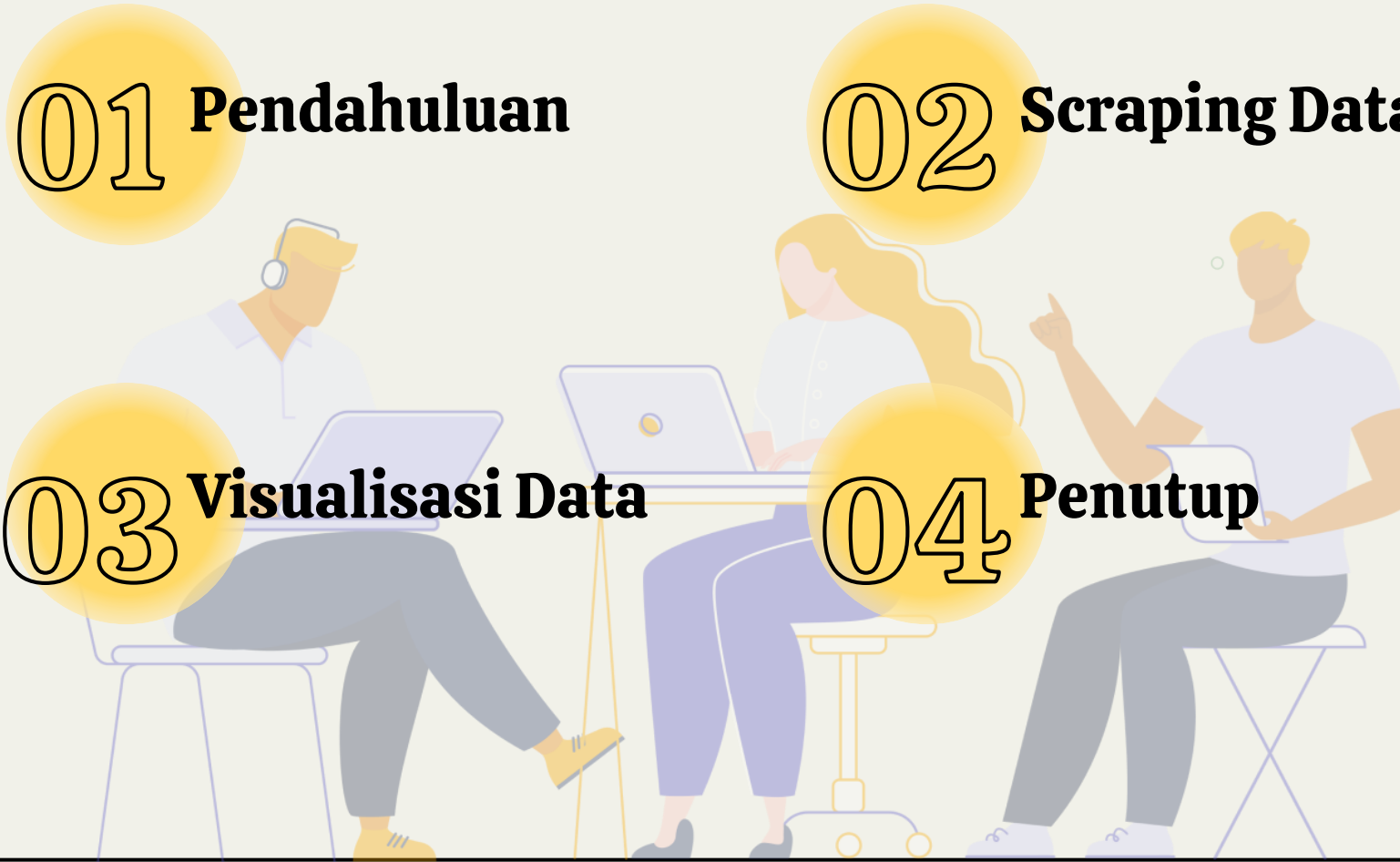
OUTLINE

01 Pendahuluan

02 Scrapping Data

03 Visualisasi Data

04 Penutup



01

?

PENDAHULUAN



mongoDB®



Apa itu Web Scrapping?

Web scraping adalah metode pengambilan data dari sebuah website. Metode ini sangat berguna dalam bisnis online, baik itu untuk riset pasar, riset kompetitor, atau mencari leads.



mongoDB®





Carmudi Indonesia

Proyek kali ini akan melakukan scraping pada situs web "Carmudi Indonesia" yang dapat diakses melalui <https://www.carmudi.co.id/>. Situs web ini adalah platform terkemuka yang menyediakan daftar kendaraan untuk dijual di Indonesia.

Toyota Raize

Salah satu produk kendaraan yang dipasarkan dalam situs ini adalah Toyota Raize. Toyota Raize adalah sebuah SUV Compact yang diproduksi oleh Toyota. Kendaraan ini dikenal dengan desainnya yang sporty dan modern, serta fitur-fitur canggih yang ditawarkan. Scraping data pada Toyota Raize memiliki kepentingan strategis dalam berbagai aspek industri otomotif. Dengan mengumpulkan informasi tentang SUV Compact ini, analis dan dealer dapat memperoleh pemahaman yang lebih dalam tentang tren pasar, preferensi konsumen, dan posisi kompetitif Raize dibandingkan dengan kendaraan lain di segmennya.



Data apa saja yang akan dilakukan scraping?

Data yang akan dilakukan scraping pada website ini adalah:

- ☐ Jenis Mobil
- ☐ Harga Mobil (Bekas)
- ☐ Odometer
- ☐ Tahun Produksi
- ☐ Jenis Transmisi
- ☐ Lokasi



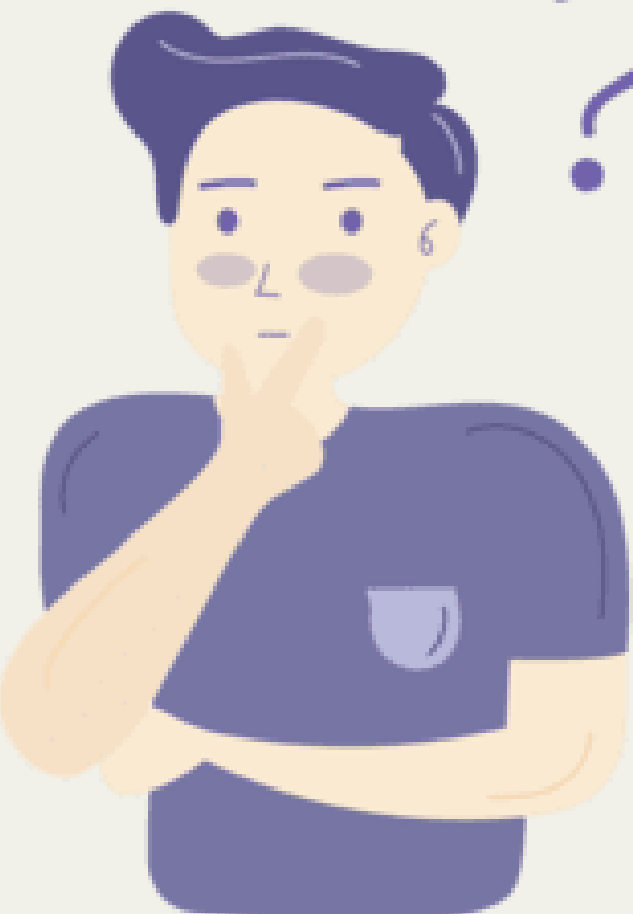
02



SCRAPING DATA



mongoDB®



Muat Library Yang Akan Digunakan

Library(rvest)

Rvest digunakan untuk scraping data menggunakan software Rstudio.

Library(tidyverse)

Tidyverse digunakan untuk membersihkan, merapikan serta visualisasi data.

Library(mongolite)

Mongolite digunakan untuk menghubungkan serta memasukkan data hasil scraping ke server MongoDB.

* Install terlebih dahulu software RStudio dan MongoDB jika belum memiliki aplikasinya *

Masukkan link atau alamat website

```
alamatweb <- 'https://www.carmudi.co.id/mobil-bekas-dijual/toyota/raize/indonesia?page_size=25'  
lamanweb <- read_html(alamatweb)
```

Mengambil variabel data yang diinginkan

```
Jenis_Mobil <- lamanweb %>% html_nodes(".ellipsize") %>%  
  html_text() %>%  
  gsub("\n", "", .)  
Harga_Mobil <- lamanweb %>% html_nodes(".listing__price") %>%  
  html_text() %>%  
  gsub("\n", "", .)  
Odometer <- lamanweb %>% html_nodes(".listing__specs") %>%  
  html_text() %>%  
  gsub("\n", "", .)
```

Membentuk data frame

```
tabeldf <- data.frame(Jenis_Mobil, Harga_Mobil, Odometer, stringsAsFactors = FALSE)  
head(tabeldf)
```

Jenis_Mobil <chr>	Harga_Mobil <chr>
1 2021 Toyota Raize 1.0 GR Sport TSS Wagon - matic	Rp 235.000.000
2 2021 Toyota Raize 1.0 G Wagon	Rp 206.000.000
3 2022 Toyota Raize 1.0 GR Sport Wagon - low km 24ribu	Rp 236.000.000
4 2021 Toyota Raize 1.0 GR Sport Wagon - KM 10RB	Rp 225.000.000
5 2021 Toyota Raize 1.0 G Wagon - bunga 0persen free 1 th asuransi dan garansi 1th	Rp 206.000.000
6 2021 Toyota Raize 1.0 G Wagon	Rp 206.000.000

6 rows | 1-3 of 3 columns

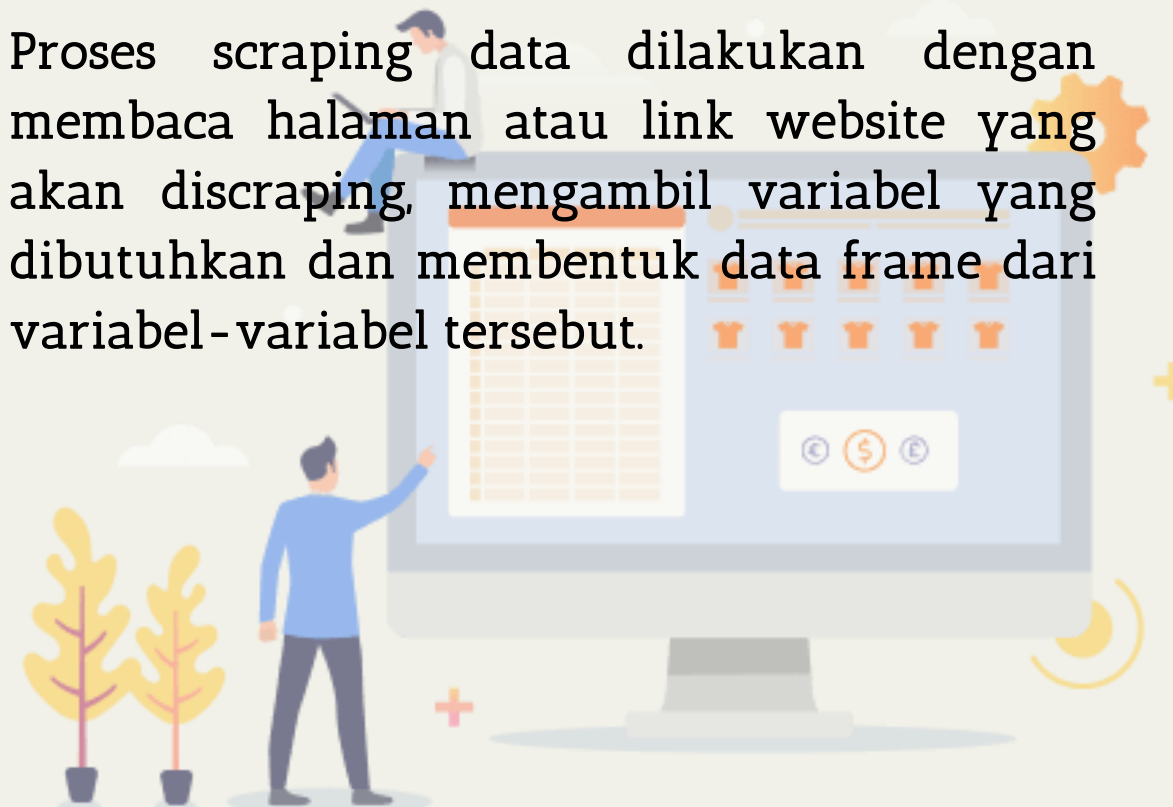
Odometer <chr>
25 - 30K KM Automatic Jawa Timur Dealer Dealer terpercaya memiliki rekam jejak yang terbukti menegakkan praktik jual mobil terbaik disertifikasi ole...
35 - 40K KM Automatic Jawa Barat Sales Agent Dealer terpercaya memiliki rekam jejak yang terbukti menegakkan praktik jual mobil terbaik disertifikas...
15 - 20K KM Automatic Jawa Timur Dealer Dealer terpercaya memiliki rekam jejak yang terbukti menegakkan praktik jual mobil terbaik disertifikasi ole...
10 - 15K KM Automatic Jawa Timur Sales Agent Dealer terpercaya memiliki rekam jejak yang terbukti menegakkan praktik jual mobil terbaik disertifika...
35 - 40K KM Automatic Jawa Barat Sales Agent Dealer terpercaya memiliki rekam jejak yang terbukti menegakkan praktik jual mobil terbaik disertifikas...
35 - 40K KM Automatic DKI Jakarta Sales Agent Dealer terpercaya memiliki rekam jejak yang terbukti menegakkan praktik jual mobil terbaik disertifika...

6 rows | 4-4 of 3 columns



Scraping Data

Proses scraping data dilakukan dengan membaca halaman atau link website yang akan discraping, mengambil variabel yang dibutuhkan dan membentuk data frame dari variabel-variabel tersebut.



Cleaning Data

Data yang telah discraping ternyata masih berantakan. Sehingga perlu dilakukan cleaning data serta merapikan data menjadi satu data frame yang utuh dan rapi untuk kemudian disimpan dalam format csv maupun xlsx.

Membersihkan data

```
tabeldf <- tabeldf %>%
  mutate(
    Tahun = str_sub(Jenis_Mobil, 1, 4),
    Jenis_Mobil = str_sub(Jenis_Mobil, 6),
    Transmisi = str_sub(Odometer, 10, 21),
    Lokasi = str_sub(Odometer, 20, 33),
    Odometer = str_sub(Odometer, 1, 11),
    Harga_Mobil = gsub("[0-9]+%", "", Harga_Mobil),
    Odometer = gsub("Au", "", Odometer),
    Transmisi = gsub("KM", "", Transmisi),
    Transmisi = gsub("M", "", Transmisi),
    Transmisi = gsub("Ja", "", Transmisi),
    Transmisi = gsub("Automatic Ba", "Automatic", Transmisi),
    Transmisi = gsub("Automatic DK", "Automatic", Transmisi),
    Transmisi = gsub("anual w", "Manual", Transmisi),
    Lokasi = gsub("ic", "", Lokasi),
    Lokasi = gsub("c", "", Lokasi),
    Lokasi = gsub("Dea", "", Lokasi),
    Lokasi = gsub("Sal", "", Lokasi),
    Lokasi = gsub("Jawa Timur D", "Jawa Timur", Lokasi),
    Lokasi = gsub("ler", "", Lokasi),
    Lokasi = gsub("De", "", Lokasi),
    Lokasi = gsub("Banten e", "Banten", Lokasi),
    Lokasi = gsub("Banten l", "Banten", Lokasi),
    Lokasi = gsub("awa Timur e", "Jawa Timur", Lokasi),
    Lokasi = gsub("awa Timur l", "Jawa Timur", Lokasi)
  )
head(tabeldf)
```

Jenis_Mobil <chr>	Harga_Mobil <chr>	Odometer <chr>		
1 Toyota Raize 1.0 GR Sport TSS Wagon - matic	Rp 235.000.000	25 - 30K KM		
2 Toyota Raize 1.0 G Wagon	Rp 206.000.000	35 - 40K KM		
3 Toyota Raize 1.0 GR Sport Wagon - low km 24ribu	Rp 236.000.000	15 - 20K KM		
4 Toyota Raize 1.0 GR Sport Wagon - KM 10RB	Rp 225.000.000	10 - 15K KM		
5 Toyota Raize 1.0 G Wagon - bunga Opersen free 1 th asuransi dan garansi 1 th	Rp 206.000.000	35 - 40K KM		
6 Toyota Raize 1.0 G Wagon	Rp 206.000.000	35 - 40K KM		

6 rows | 1-4 of 6 columns

Harga_Mobil <chr>	Odometer <chr>	Tahun <chr>	Transmisi <chr>	Lokasi <chr>
Rp 235.000.000	25 - 30K KM	2021	Automatic	Jawa Timur
Rp 206.000.000	35 - 40K KM	2021	Automatic	Jawa Barat
Rp 236.000.000	15 - 20K KM	2022	Automatic	Jawa Timur
Rp 225.000.000	10 - 15K KM	2021	Automatic	Jawa Timur
Rp 206.000.000	35 - 40K KM	2021	Automatic	Jawa Barat
Rp 206.000.000	35 - 40K KM	2021	Automatic	DKI Jakarta

6 rows | 3-7 of 6 columns

Input Data ke MongoDB

Data scraping yang sudah teratur kemudian dimasukkan ke dalam MongoDB dengan cara menghubungkannya melalui sintaks di RStudio.

Sintaks ini kemudian dimasukkan ke dalam github serta mengatur workflow pada github agar data dapat diinput otomatis setiap hari. Hingga hari ini proses input data masih berjalan secara otomatis.

```
# Memilih satu baris data secara acak
pilih <- sample(1:25,1,replace=F)
mobil_terpilih <- tabeldf[pilih,]

# Koneksi ke MongoDB untuk memasukkan data
atlas_conn <- mongo(
  collection = Sys.getenv("ATLAS_COLLECTION"),
  db          = Sys.getenv("ATLAS_DB"),
  url         = Sys.getenv("ATLAS_URL")
)

atlas_conn$insert(mobil_terpilih)
rm(atlas_conn)
```

Code Blame 28 lines (26 loc) · 767 Bytes Code 55% faster with GitHub Copilot

```
1 name: carmudi_scraping
2
3 on:
4   schedule:
5     - cron: '0 1 * * *'
6     - cron: '0 7 * * *'
7   workflow_dispatch:
8
9 jobs:
10  carmudi-scrape:
11    runs-on: macOS-latest
12    env:
13      ATLAS_URL: ${ secrets.ATLAS_URL }
14      ATLAS_COLLECTION: ${ secrets.ATLAS_COLLECTION }
15      ATLAS_DB: ${ secrets.ATLAS_DB }
16    steps:
17      - name: Start time
18        run: echo "$(date) ** $(TZ=Asia/Jakarta date)"
19      - uses: actions/checkout@v3
20      - uses: r-lib/actions/setup-r@v2
21      - name: Install packages
22        run: |
23          install.packages("rvest", dependencies = TRUE)
24          install.packages("tidyverse", dependencies = TRUE)
25          install.packages("mongolite")
26      shell: Rscript {0}
27      - name: Scrape Carmudi Toyota Raize
28        run: Rscript ScrapingCarmudi.R
```



mongoDB



Studio

03

VISUALISASI DATA



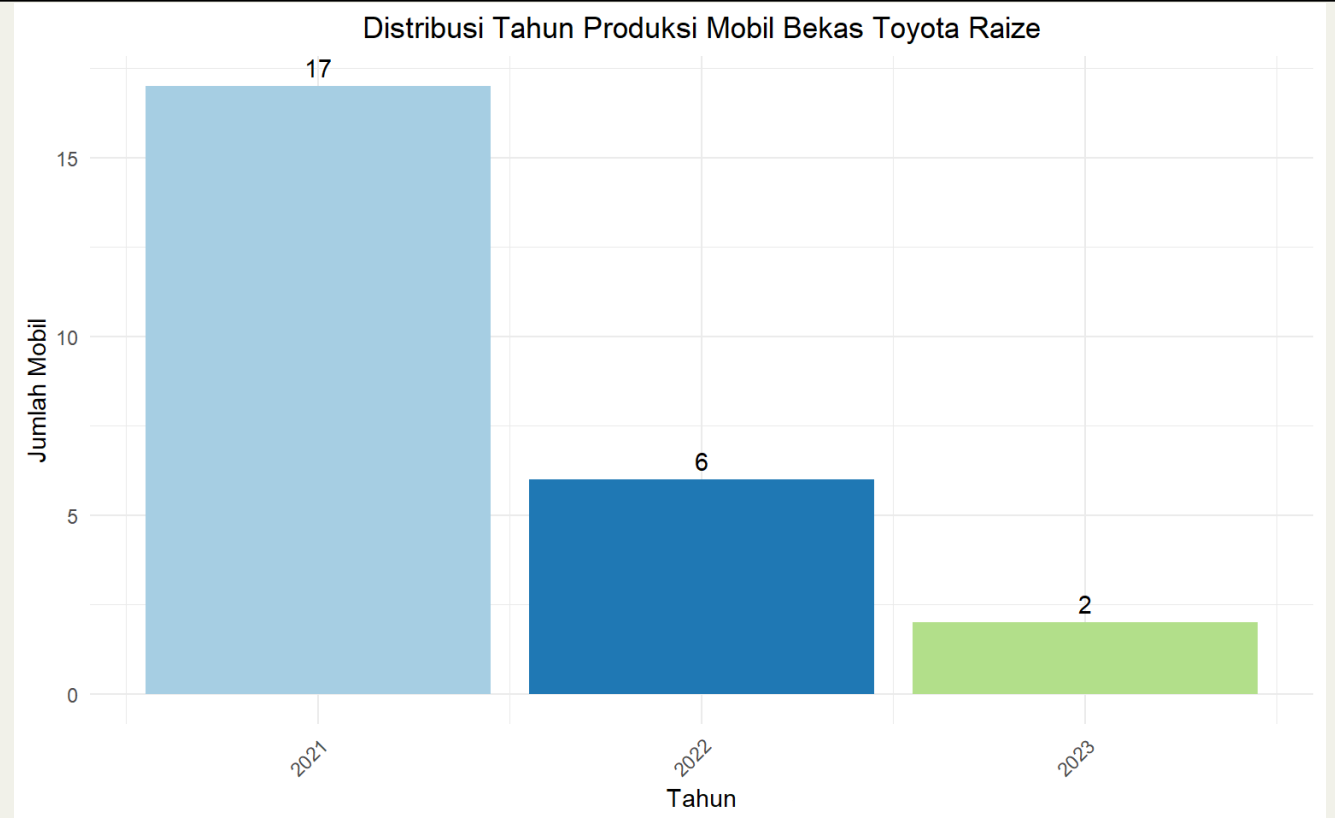
mongoDB®



Studio®

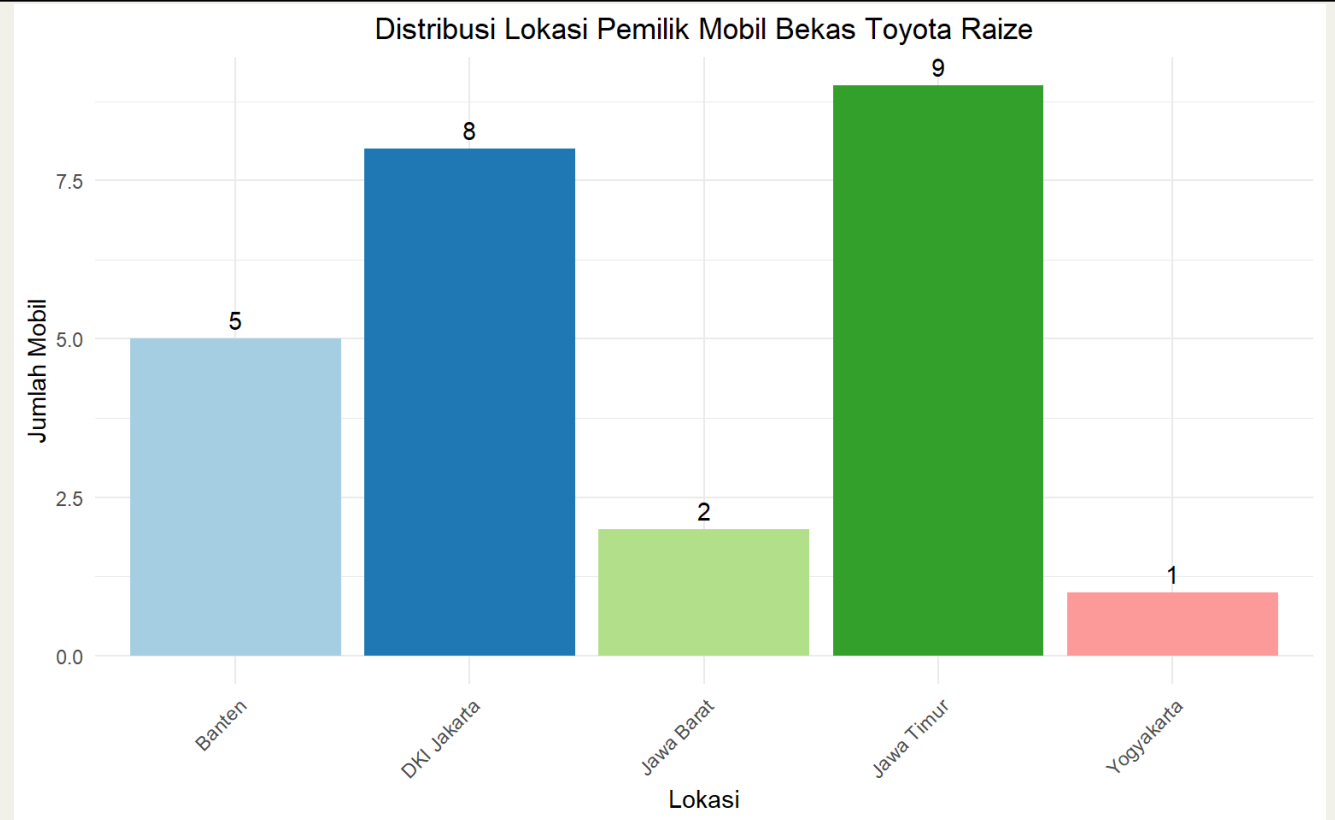
Sebaran Tahun Produksi Mobil

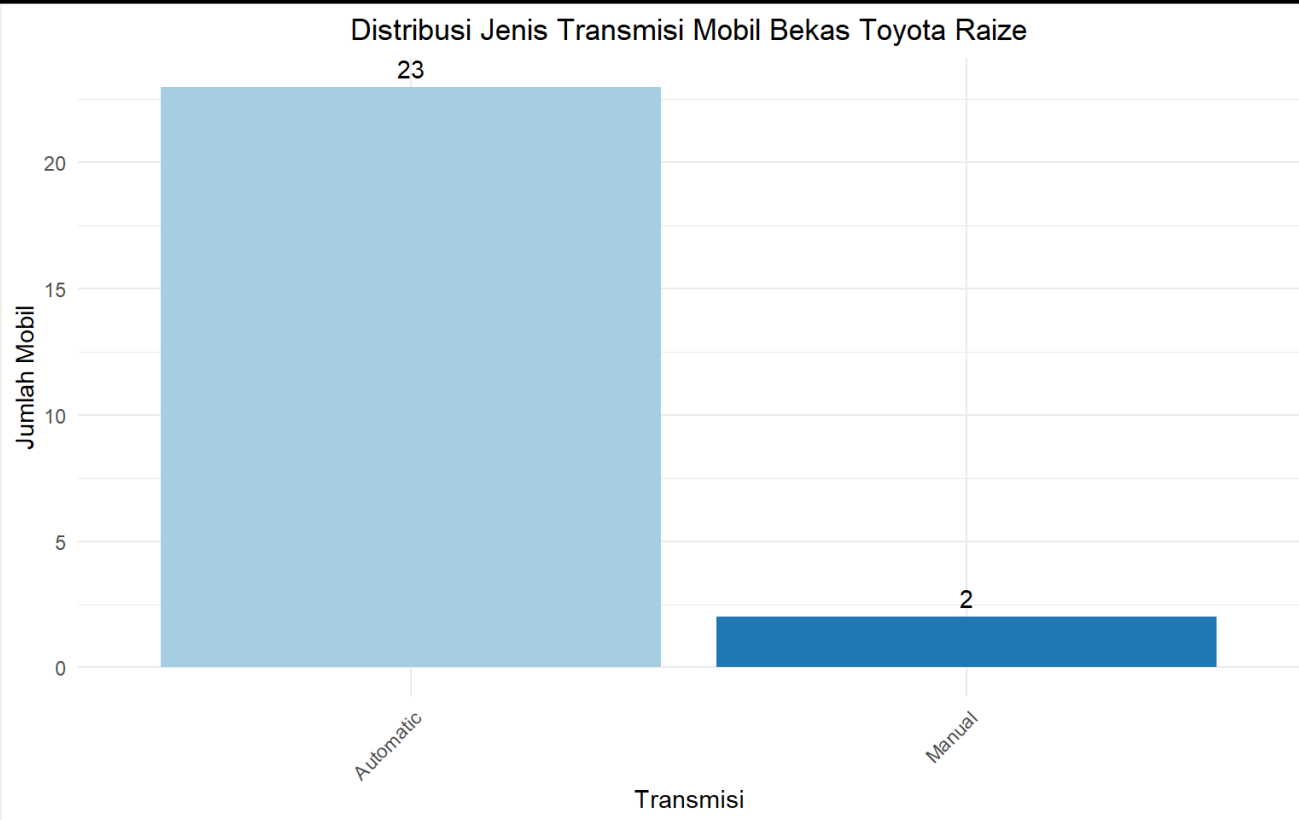
Data menunjukkan bahwa mobil bekas toyota raize dengan tahun produksi 2021 memiliki jumlah tertinggi. Artinya sebanyak 17 dari 25 mobil bekas toyota raize yang dijual pada situs carmudi indonesia merupakan mobil tahun produksi 2021.



Sebaran Lokasi Asal Mobil

Data menunjukkan bahwa mobil bekas toyota raize dari Provinsi Jawa Timur (9) dan DKI Jakarta (8) merupakan yang tertinggi. Artinya sebanyak 17 dari 25 mobil bekas toyota raize yang dijual pada situs carmudi indonesia merupakan mobil yang berasal dari Provinsi Jawa Timur dan DKI Jakarta.

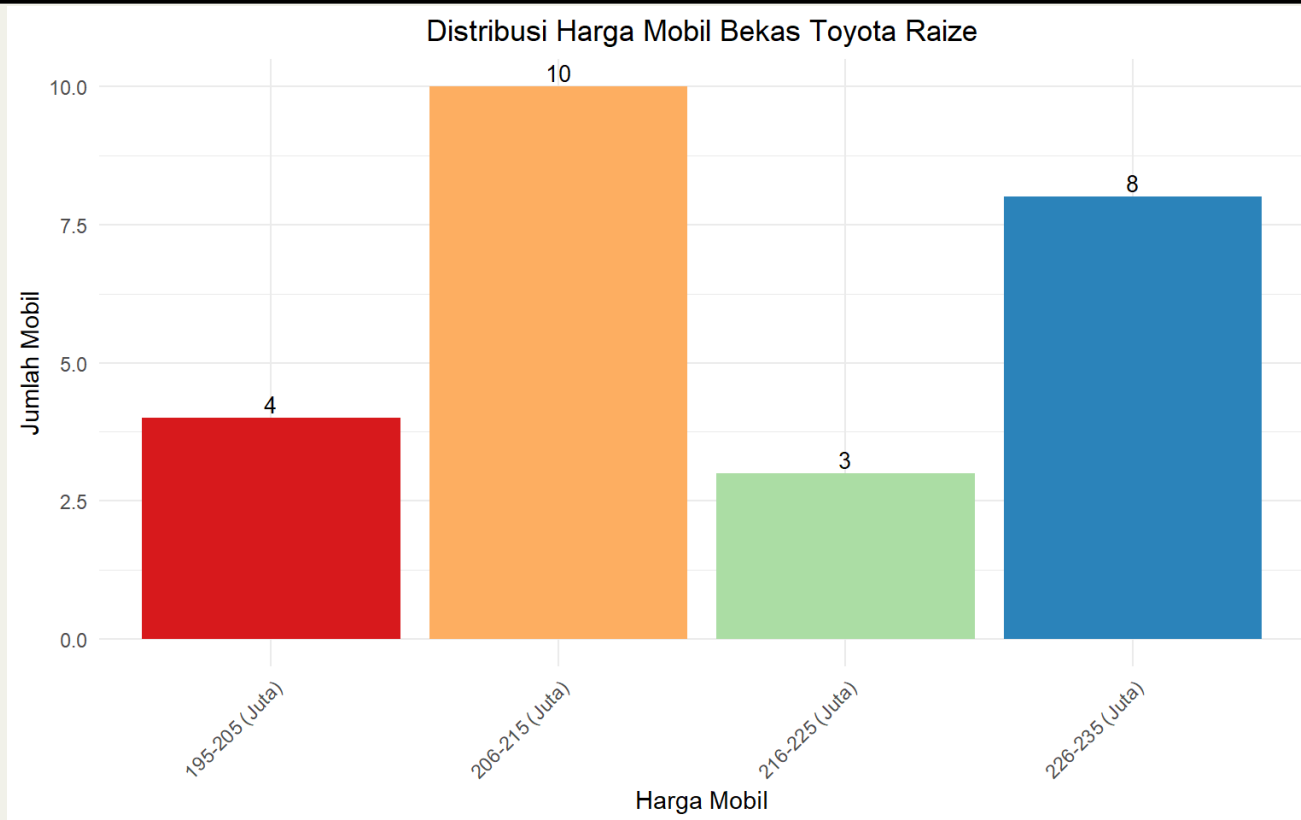




Sebaran Jenis Transmisi Mobil

Data menunjukkan bahwa mobil bekas toyota raize dengan transmisi automatic merupakan yang paling banyak dijual dibanding transmisi manual. Artinya sebanyak 23 dari 25 mobil bekas toyota raize yang dijual pada situs carmudi indonesia merupakan mobil dengan transmisi Automatic.





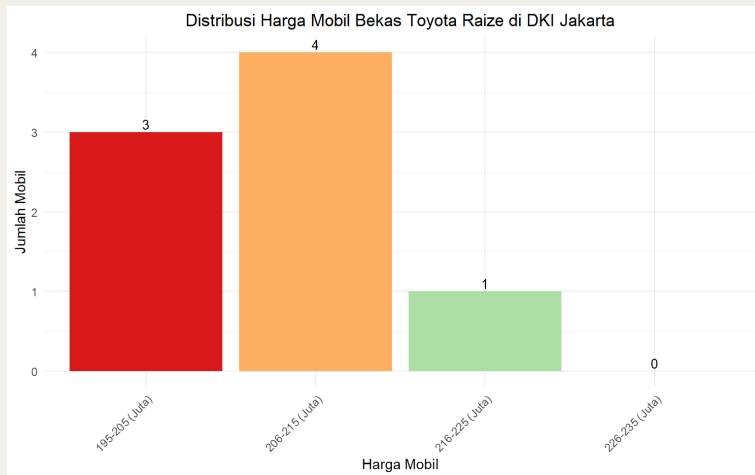
Sebaran Range Harga Mobil

Data menunjukkan bahwa mobil toyota raize sebagian besar memiliki range harga 206-215 juta dan 226-235 juta. Namun yang paling banyak ada pada range harga 206-215 juta rupiah.



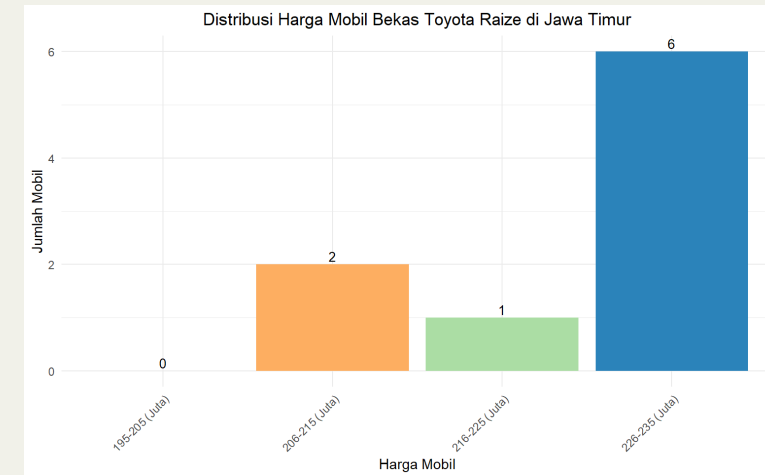
Menarik!

Harga dengan variabel lain memiliki hubungan yang menarik untuk dipahami, oleh karena itu berikut visualisi hubungan harga dan variabel lain.



Harga – Lokasi (DKI Jakarta)

Data menunjukkan bahwa mobil bekas toyota raize dari Provinsi DKI Jakarta sebagian besar memiliki range harga 195-206 juta rupiah, dari 8 mobil tidak ada mobil dengan harga diatas 226 juta rupiah di Provinsi DKI Jakarta.

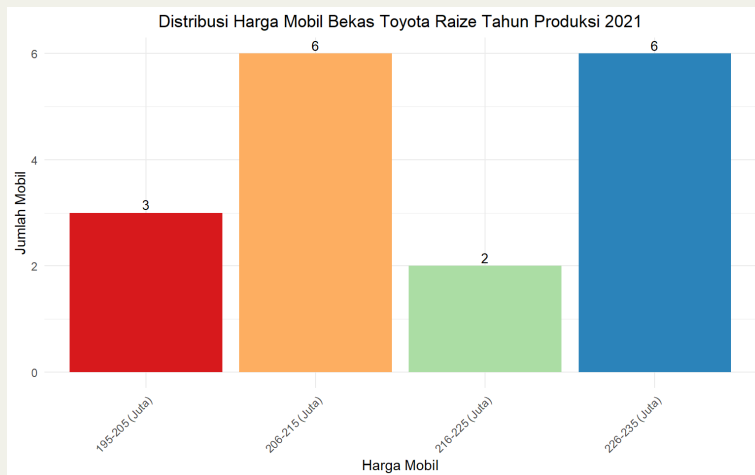


Harga – Lokasi (Jawa Timur)

Berbanding terbalik dengan Provinsi DKI Jakarta, data menunjukkan bahwa mobil bekas toyota raize dari Provinsi Jawa Timur sebagian besar memiliki range harga diatas 226 juta rupiah, sedangkan dari 9 mobil tidak ada mobil dengan harga dibawah 206 juta rupiah.

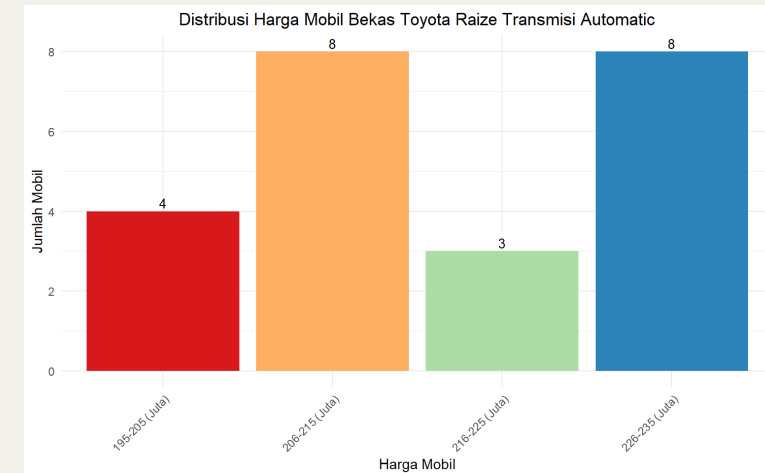
Menarik!

Harga dengan variabel lain memiliki hubungan yang menarik untuk dipahami, oleh karena itu berikut visualisi hubungan harga dan variabel lain.



Harga – Tahun Produksi (2021)

Data menunjukkan bahwa mobil bekas toyota raize tahun produksi 2021 sebagian besar memiliki range harga 206-215 juta dan 226-235 juta rupiah.



Harga – Transmisi (Automatic)

Data menunjukkan bahwa mobil bekas toyota raize dengan transmisi automatic sebagian besar ada pada range harga 206-215 juta dan 226-235 juta rupiah.

04

PENUTUP



mongoDB®



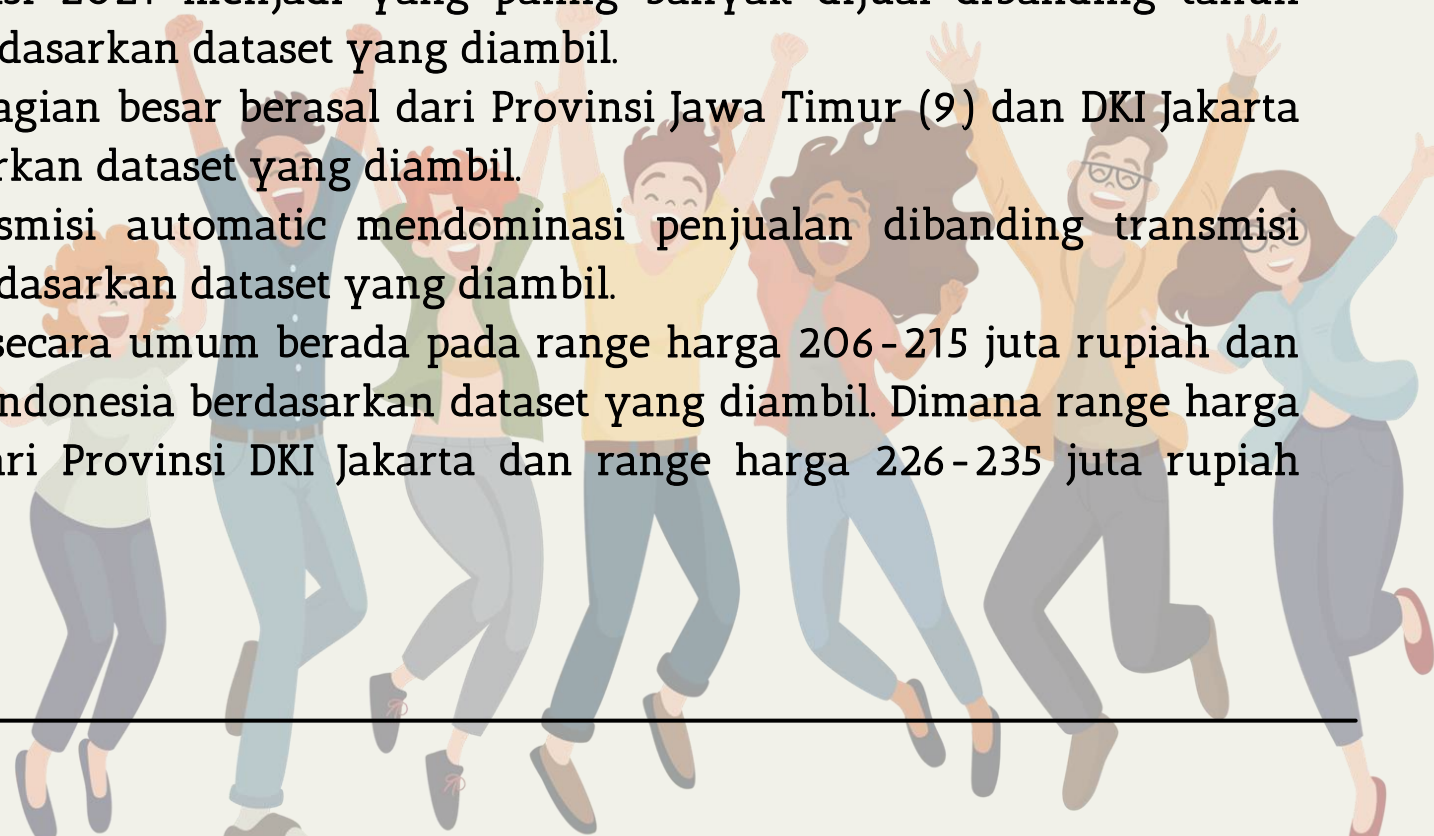
Studio®



Kesimpulan

Berdasarkan analisis dan visualisasi sebelumnya dapat ditarik beberapa kesimpulan diantaranya sebagai berikut:

- ✓ Mobil bekas toyota raize tahun produksi 2021 menjadi yang paling banyak dijual dibanding tahun lainnya pada situs carmudi indonesia berdasarkan dataset yang diambil.
- ✓ Mobil bekas toyota raize yang dijual sebagian besar berasal dari Provinsi Jawa Timur (9) dan DKI Jakarta (8) pada situs carmudi indonesia berdasarkan dataset yang diambil.
- ✓ Mobil bekas toyota raize dengan transmisi automatic mendominasi penjualan dibanding transmisi manual pada situs carmudi indonesia berdasarkan dataset yang diambil.
- ✓ Sebaran harga mobil bekas toyota raize secara umum berada pada range harga 206-215 juta rupiah dan 226-235 juta rupiah pada situs carmudi indonesia berdasarkan dataset yang diambil. Dimana range harga 206-215 juta rupiah banyak berasal dari Provinsi DKI Jakarta dan range harga 226-235 juta rupiah banyak berasal dari Provinsi Jawa Timur.






mongoDB®



Studio®

Thank you!

 ekadik7@gmail.com

 ekadicky

 ekdrmw n

