SrapingWeb

Tim2

2025-04-28

```
library(mongolite)
library(rvest)
library(dplyr)
library(httr)
library(ggplot2)
library(writexl)
library(scales)
library(tidyr)
```

Scraping Data Pendapatan dan Laba Perusahan besar di duania

```
# URL Wikipedia
url <- "https://en.wikipedia.org/wiki/List_of_largest_companies_by_revenue"

# Baca halaman
page <- read_html(url)

# Ambil semua tabel
tables <- page %>% html_nodes("table")

# Ambil tabel pertama
companies_table <- tables[[1]] %>% html_table(fill = TRUE)

# Perbaiki nama kolom
colnames(companies_table) <- make.names(colnames(companies_table))

# Hapus baris pertama (karena itu header ganda)
companies_table <- companies_table %>% slice(-1)

# Hapus kolom State.owned dan Ref.
companies_table <- companies_table %>%
  select(-State.owned, -Ref.)

# Bersihkan dan konversi tipe data
companies_table <- companies_table %>%
  filter(!is.na(Rank)) %>%
  mutate(
    Revenue = as.numeric(gsub("[\\$,]", "", Revenue)),
    Profit = as.numeric(gsub("[\\$,]", "", Profit)),
    Employees = as.integer(gsub(",", "", Employees))
  )

# Ganti nama kolom Headquarters.note.1. menjadi Headquarters
colnames(companies_table)[colnames(companies_table) == "Headquarters.note.1."] <- "Headquarters"

# Lihat hasil akhir
glimpse(companies_table)
```

```
## Rows: 50
## Columns: 7
## $ Rank         <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", …
## $ Name         <chr> "Walmart", "Amazon", "State Grid Corporation of China", "…
## $ Industry     <chr> "Retail", "Retailinformation technology", "Electricity", …
## $ Revenue      <dbl> 680985, 637959, 545948, 480446, 429700, 476000, 400278, 3…
## $ Profit       <dbl> 19436, 59248, 9204, 106246, 9393, 25250, 14405, 93736, 88…
## $ Employees    <int> 2100000, 1556000, 1361423, 73311, 513434, 1026301, 400000…
## $ Headquarters <chr> "United States", "United States", "China", "Saudi Arabia"…
```

```
# Simpan data ke file Excel
#write_xlsx(companies_table, path = "G:/My Drive/Manajemen data statistik/CompaniesRank.xlsx")
```

Membuat Koneksi dan memasukkan data ke mongoDB

```
atlas_conn <- mongo(
  collection = "companies",
  db = "companies_by_revenue",
  url = "mongodb+srv://amrinajih:Najih9999@dbscraping.e7tk77u.mongodb.net/companies_by_revenue?r
etryWrites=true&w=majority&appName=DBscraping"
)
atlas_conn$remove("{}")
atlas_conn$insert(companies_table)
```

```
## List of 5
##  $ nInserted  : num 50
##  $ nMatched   : num 0
##  $ nRemoved   : num 0
##  $ nUpserted  : num 0
##  $ writeErrors: list()
```

```
# Count documents in collection
doc_count <- atlas_conn$count()
message(paste("Collection contains", doc_count, "documents"))
```

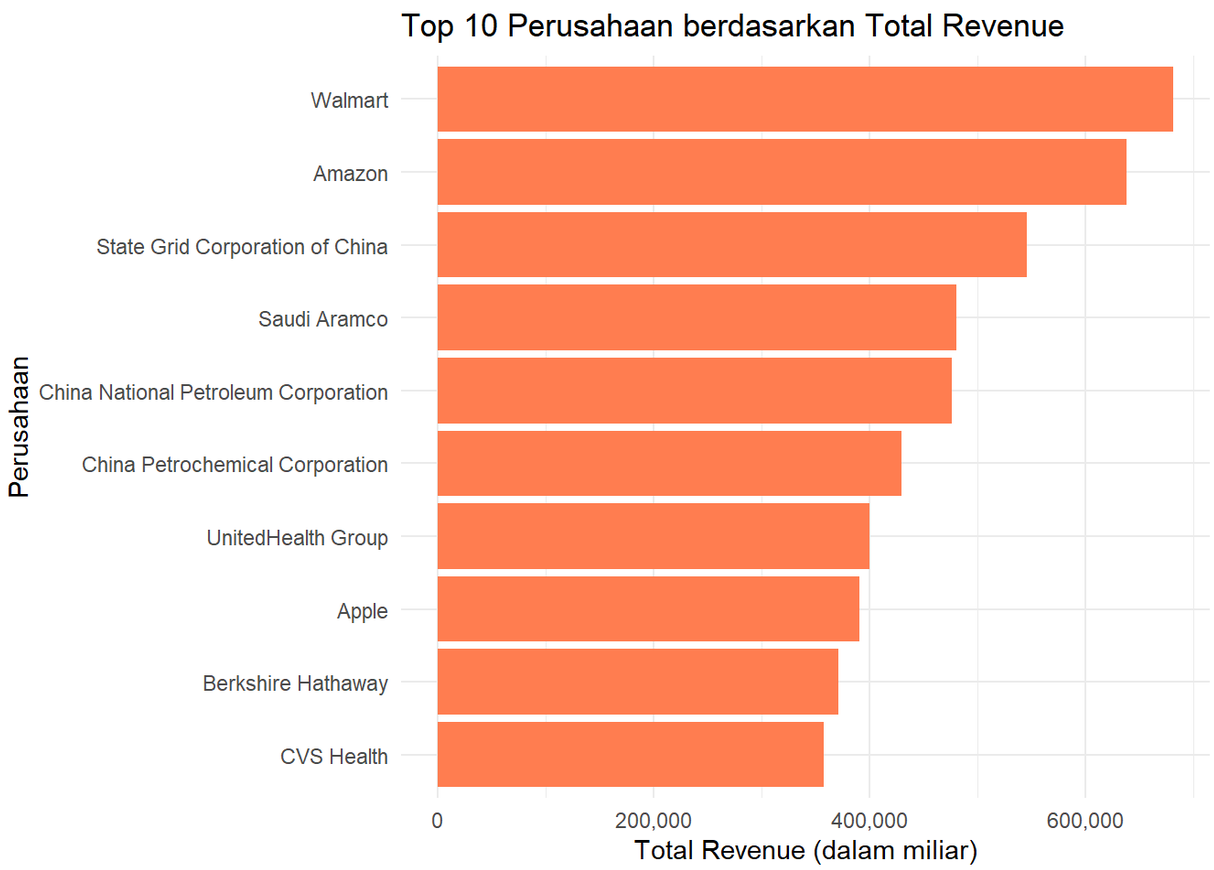
```
## Collection contains 50 documents
```

Top 10 Perusahaan berdasarkan Total Pendapatan

```
comp_by_revenue <- atlas_conn$aggregate('[
  {
    "$group": {
      "_id": "$Name",
      "total_revenue": { "$sum": "$Revenue" }
    }
  },
  {
    "$sort": { "total_revenue": -1 }
  }
]')
print(comp_by_revenue)
```

```
##                                     _id total_revenue
## 1                                Walmart        680985
## 2                                 Amazon        637959
## 3          State Grid Corporation of China        545948
## 4                             Saudi Aramco        480446
## 5     China National Petroleum Corporation        476000
## 6          China Petrochemical Corporation        429700
## 7                       UnitedHealth Group        400278
## 8                                  Apple        391035
## 9                       Berkshire Hathaway        371433
## 10                            CVS Health        357776
## 11                              Alphabet        350018
## 12                        Volkswagen Group        348408
## 13                              ExxonMobil        344582
## 14                                   Vitol        331000
## 15                                   Shell        323183
## 16    China State Construction Engineering        320431
## 17                                  Toyota        312018
## 18                                McKesson        308951
## 19                                 Cencora        262173
## 20                               Microsoft        245122
## 21                               Trafigura        244280
## 22                                 Costco        242290
## 23                          JPMorgan Chase        239425
## 24 Industrial and Commercial Bank of China        222484
## 25                           TotalEnergies        218945
## 26                                Glencore        217829
## 27                                      BP        213032
## 28                         Cardinal Health        205012
## 29                              Stellantis        204908
## 30                                 Chevron        200949
## 31                China Construction Bank        199826
## 32                    Samsung Electronics        198257
## 33                                 Foxconn        197876
## 34                                   Cigna        195265
## 35             Agricultural Bank of China        192398
## 36                          Schwarz Gruppe        180576
## 37   China Railway Engineering Corporation        178563
## 38                                 Cargill        177000
## 39                     Ford Motor Company        176191
## 40                          Bank of China        172328
## 41                        Bank of America        171912
## 42                         General Motors        171842
## 43                        Elevance Health        171340
## 44                              BMW Group        168103
## 45                    Mercedes-Benz Group        165638
## 46                              Tata Group        165500
## 47  China Railway Construction Corporation        160847
## 48                                   Baowu        157216
## 49                               Citigroup        156820
## 50                                    Enel        147100
```

```r
# Filter 10 perusahaan dengan revenue tertinggi
top_companies <- comp_by_revenue %>%
  arrange(desc(total_revenue)) %>%
  slice_head(n = 10)

# Plot
ggplot(top_companies, aes(x = reorder(`_id`, total_revenue), y = total_revenue)) +
  geom_col(fill = "coral") +
  coord_flip() +
  labs(
    title = "Top 10 Perusahaan berdasarkan Total Revenue",
    x = "Perusahaan",
    y = "Total Revenue (dalam miliar)"
  ) +
  scale_y_continuous(labels = scales::comma)+
  theme_minimal()
```



Bar Chart di atas menunjukkan 10 perusahaan terbesar di dunia berdasarkan total pendapatan (revenue) dalam satuan miliar USD. Setiap batang merepresentasikan total pendapatan tahunan dari masing-masing perusahaan, dengan perusahaan yang memiliki pendapatan tertinggi berada di urutan paling atas.

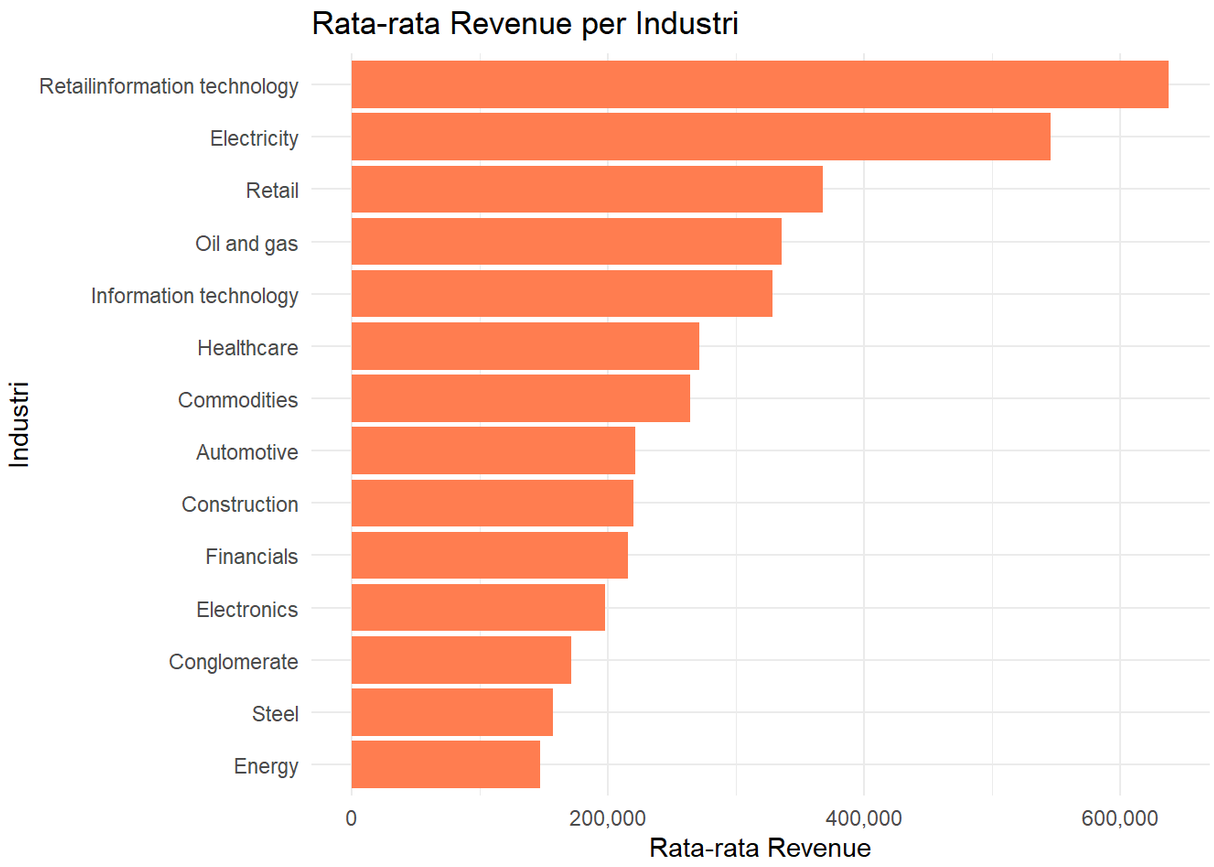
- Walmart menempati posisi teratas sebagai perusahaan dengan pendapatan tertinggi, melebihi 600 miliar USD, menjadikannya pemimpin global dalam sektor ritel.
- Amazon, juga dari sektor ritel dan teknologi, berada di posisi kedua, mencerminkan kekuatan e-commerce dan cloud computing.
- Perusahaan energi seperti State Grid Corporation of China, Saudi Aramco, dan dua entitas energi nasional Tiongkok (China National Petroleum Corporation dan China Petrochemical Corporation) mendominasi posisi selanjutnya. Ini menunjukkan bahwa sektor energi masih memegang kontribusi besar dalam perolehan pendapatan skala global.
- Di bidang kesehatan dan asuransi, UnitedHealth Group dan CVS Health juga masuk 10 besar, mencerminkan skala industri kesehatan yang besar terutama di Amerika Serikat.
- Apple sebagai satu-satunya perusahaan teknologi konsumen dalam daftar menunjukkan bahwa perusahaan produk elektronik dan layanan digital juga mampu bersaing dari sisi pendapatan.
- Berkshire Hathaway, konglomerat yang dipimpin oleh Warren Buffett, juga berada di posisi tinggi berkat diversifikasi portofolio bisnisnya.

Rata-rata pendapatan per Industri

```
avg_rev_by_industry <- atlas_conn$aggregate('[
  {
    "$group": {
      "_id": "$Industry",
      "avg_revenue": { "$avg": "$Revenue" }
    }
  },
  {
    "$sort": { "avg_revenue": -1 }
  }
]')

print(avg_rev_by_industry)
```

```
##                            _id avg_revenue
## 1  Retailinformation technology    637959.0
## 2                   Electricity    545948.0
## 3                        Retail    367950.3
## 4                   Oil and gas    335854.6
## 5        Information technology    328725.0
## 6                    Healthcare    271542.1
## 7                   Commodities    264369.7
## 8                    Automotive    221015.4
## 9                  Construction    219947.0
## 10                   Financials    215828.2
## 11                   Electronics    198066.5
## 12                 Conglomerate    171250.0
## 13                        Steel    157216.0
## 14                       Energy    147100.0
```

```
# Plot rata-rata revenue per industri
ggplot(avg_rev_by_industry, aes(x = reorder(`_id`, avg_revenue), y = avg_revenue)) +
  geom_col(fill = "coral") +
  coord_flip() +
  labs(
    title = "Rata-rata Revenue per Industri",
    x = "Industri",
    y = "Rata-rata Revenue"
  ) +
  scale_y_continuous(labels = scales::comma)+
  theme_minimal()
```

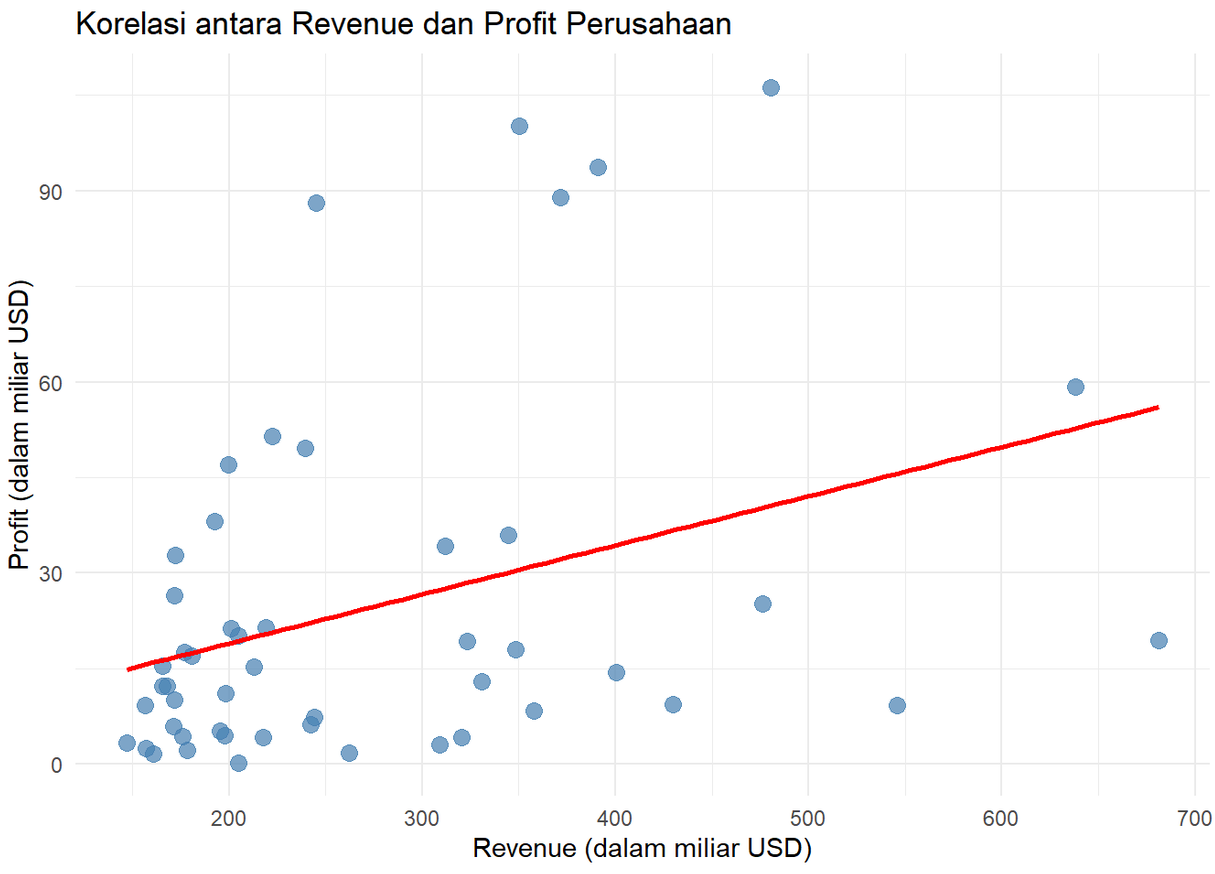


Bar Chart ini menyajikan rata-rata pendapatan (revenue) perusahaan di masing-masing sektor industri, berdasarkan data 50 perusahaan terbesar dunia. Pendapatan yang ditampilkan adalah dalam satuan miliar USD, dan dihitung sebagai rata-rata dari total revenue perusahaan-perusahaan dalam industri yang sama.

- Retail/Information Technology muncul sebagai industri dengan rata-rata pendapatan tertinggi, melebihi 650 miliar USD. Hal ini kemungkinan besar dipengaruhi oleh keberadaan perusahaan raksasa seperti Walmart dan Amazon yang mendominasi kategori tersebut.
- Electricity juga menunjukkan angka rata-rata yang sangat tinggi, mencerminkan skala besar dari perusahaan seperti State Grid Corporation of China.
- Industri lain seperti Retail, Oil and Gas, dan Information Technology juga menempati posisi atas, menegaskan bahwa sektor konsumsi dan energi memainkan peran penting dalam perekonomian global.
- Healthcare, Commodities, dan Automotive memiliki rata-rata pendapatan yang cukup tinggi, tetapi masih berada di bawah sektor energi dan teknologi.
- Sektor seperti Construction, Financials, dan Electronics mencatat rata-rata pendapatan yang sedang, mencerminkan keberagaman skala perusahaan dalam sektor tersebut.
- Industri Steel dan Energy (yang dipisahkan dari “Oil and Gas”) menempati posisi bawah, menunjukkan bahwa perusahaan-perusahaan di sektor ini mungkin memiliki pendapatan yang besar secara individual namun tidak merata di seluruh anggotanya.

Visualisasi Korelasi Pendapatan dan Laba

```
companies_df <- atlas_conn$aggregate('[
  {
    "$match": {
      "Revenue": { "$ne": null },
      "Profit": { "$ne": null }
    }
  },
  {
    "$project": {
      "_id": 0,
      "Name": 1,
      "Revenue": 1,
      "Profit": 1
    }
  }
]')


# Ubah satuan ke miliar USD (opsional)
companies_df <- companies_df %>%
  mutate(
    Revenue = Revenue / 1000,
    Profit = Profit / 1000
  )

# Plot
ggplot(companies_df, aes(x = Revenue, y = Profit)) +
  geom_point(color = "steelblue", size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, color = "red", linewidth = 1) +
  labs(
    title = "Korelasi antara Revenue dan Profit Perusahaan",
    x = "Revenue (dalam miliar USD)",
    y = "Profit (dalam miliar USD)"
  ) +
  scale_x_continuous(labels = comma) +
  scale_y_continuous(labels = comma) +
  theme_minimal()
```



Scatter plot (diagram sebar) ini menampilkan hubungan antara pendapatan (revenue) dan laba bersih (profit) dari 50 perusahaan terbesar di dunia. Setiap titik pada grafik merepresentasikan satu perusahaan, dengan: Sumbu horizontal (X) menunjukkan pendapatan tahunan dalam miliar USD.Sumbu vertikal (Y) menunjukkan laba bersih

dalam miliar USD.Garis merah adalah garis regresi linear yang memperkirakan arah dan kekuatan hubungan antara pendapatan dan laba.

- Garis tren merah menunjukkan bahwa secara umum terdapat hubungan positif antara revenue dan profit, artinya semakin tinggi pendapatan sebuah perusahaan, cenderung semakin tinggi pula labanya.
- Namun, sebaran titik relatif menyebar cukup luas, terutama di rentang revenue menengah ke atas (200–600 miliar USD), menunjukkan bahwa tingginya pendapatan tidak selalu diikuti oleh laba yang besar.
- Beberapa perusahaan memiliki pendapatan sangat tinggi namun profit yang relatif kecil, atau bahkan lebih kecil dibanding perusahaan lain dengan revenue sedang.
- Terdapat pula beberapa outlier di atas grafik (misalnya perusahaan dengan profit > 90 miliar USD), yang kemungkinan adalah perusahaan dengan efisiensi atau margin keuntungan yang sangat tinggi.

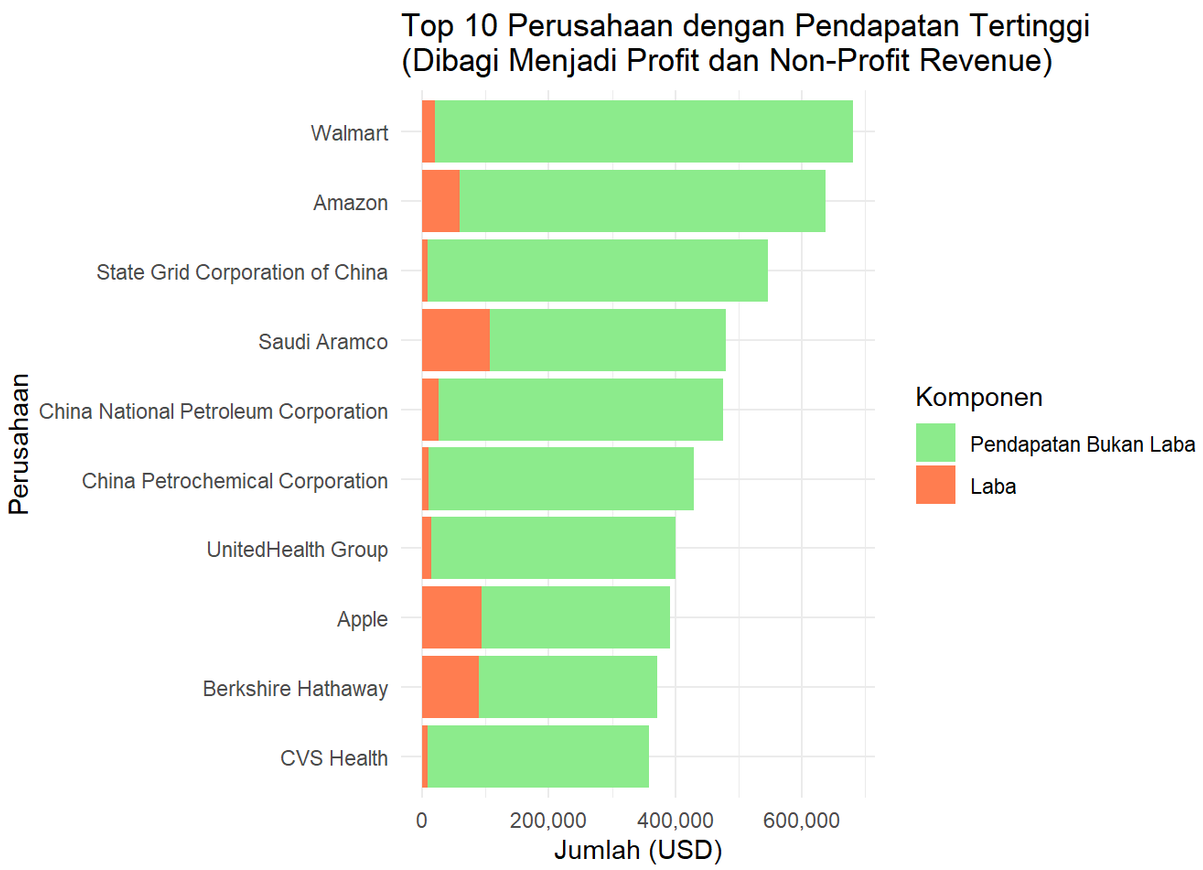
```
# 10 Perusahaan dengan jumlah karyawan terbesar
top_employees_df <- atlas_conn$aggregate('[
  {
    "$match": {
      "Employees": { "$ne": null }
    }
  },
  {
    "$project": {
      "_id": 0,
      "Name": 1,
      "Employees": 1
    }
  },
  {
    "$sort": { "Employees": -1 }
  },
  {
    "$limit": 10
  }
]')

# Plot bar chart
ggplot(top_employees_df, aes(x = reorder(Name, Employees), y = Employees)) +
  geom_col(fill = "coral") +
  coord_flip() +
  labs(
    title = "Top 10 Perusahaan dengan Jumlah Pegawai Terbanyak",
    x = "Perusahaan",
    y = "Jumlah Pegawai"
  ) +
  scale_y_continuous(labels = comma) +
  theme_minimal()
```

Menampilkan 10 perusahaan dengan jumlah karyawan terbanyak di dunia, berdasarkan data yang tersedia. Sumbu horizontal menunjukkan jumlah pegawai (dalam satuan individu), sedangkan sumbu vertikal menampilkan nama perusahaan.

- Walmart berada di urutan pertama dengan lebih dari 2 juta karyawan, menjadikannya perusahaan dengan jumlah tenaga kerja terbesar secara global. Hal ini mencerminkan skala ritel fisik Walmart yang sangat luas di berbagai negara.
- Amazon menyusul di posisi kedua, juga dengan lebih dari 1,5 juta pegawai, mencerminkan besarnya jaringan logistik dan teknologi yang dikelolanya.
- State Grid Corporation of China—perusahaan penyedia listrik terbesar di dunia—berada di posisi ketiga, menunjukkan skala infrastruktur dan layanan publik di sektor energi.
- Tata Group dan China National Petroleum Corporation juga menempati posisi atas, mencerminkan keragaman sektor dan penyebaran tenaga kerja dari konglomerat India serta perusahaan energi besar dari Tiongkok.
- Volkswagen Group dan Foxconn menandai peran besar industri otomotif dan manufaktur elektronik dalam menciptakan lapangan kerja berskala besar.
- Perusahaan ritel Schwarz Gruppe (pemilik jaringan supermarket seperti Lidl dan Kaufland), serta perusahaan minyak China Petrochemical Corporation (Sinopec), turut masuk dalam daftar.
- Agricultural Bank of China melengkapi daftar sebagai institusi keuangan dengan tenaga kerja besar, mengindikasikan skala operasional bank milik negara tersebut di seluruh wilayah Tiongkok.

```
stacked_df <- atlas_conn$aggregate('[
  {
    "$match": {
      "Revenue": { "$ne": null },
      "Profit": { "$ne": null }
    }
  },
  {
    "$project": {
      "_id": 0,
      "Name": 1,
      "Revenue": 1,
      "Profit": 1
    }
  },
  {
    "$sort": { "Revenue": -1 }
  },
  {
    "$limit": 10
  }
]')

stacked_df_long <- stacked_df %>%
  mutate(
    NonProfitRevenue = Revenue - Profit,
    Profit = Profit
  ) %>%
  select(Name, Profit, NonProfitRevenue) %>%
  pivot_longer(cols = c("NonProfitRevenue", "Profit"),
               names_to = "Component",
               values_to = "Value")

# Buat kolom NonProfitRevenue
stacked_df <- stacked_df %>%
  mutate(
    NonProfitRevenue = Revenue - Profit
  )

# Hitung total revenue untuk sorting
total_rev <- stacked_df %>%
  mutate(Total = Revenue) %>%
  select(Name, Total)

# Ubah ke format long
stacked_df_long <- stacked_df %>%
  select(Name, Profit, NonProfitRevenue) %>%
  pivot_longer(cols = c("NonProfitRevenue", "Profit"),
               names_to = "Component",
               values_to = "Value") %>%
  left_join(total_rev, by = "Name")

# Visualisasi
ggplot(stacked_df_long, aes(x = Value, y = reorder(Name, Total), fill = Component)) +
  geom_col() +
  labs(
    title = "Top 10 Perusahaan dengan Pendapatan Tertinggi\n(Dibagi Menjadi Profit dan Non-Profi
t Revenue)",
    x = "Jumlah (USD)",
    y = "Perusahaan",
    fill = "Komponen"
  ) +
  scale_x_continuous(labels = comma) +
  scale_fill_manual(
    values = c("NonProfitRevenue" = "lightgreen", "Profit" = "coral"),
    labels = c("NonProfitRevenue" = "Pendapatan Bukan Laba", "Profit" = "Laba")
```

```
) +
theme_minimal()
```



Grafik batang horizontal ini menampilkan 10 perusahaan dengan total pendapatan (revenue) tertinggi di dunia, sekaligus memvisualisasikan bagaimana komposisi pendapatan tersebut terbagi menjadi dua komponen: Laba (Profit) ditampilkan sebagai segmen berwarna jingga (orange). Pendapatan selain laba (Non-Profit Revenue) ditampilkan sebagai segmen berwarna hijau muda, yaitu selisih antara total revenue dan profit. Dengan kata lain, panjang total bar = total revenue, sedangkan distribusi warna menunjukkan berapa besar bagian dari pendapatan yang benar-benar menjadi laba dan berapa sisanya adalah biaya operasional, beban, atau non-profit income.

- Walmart dan Amazon memiliki total revenue yang sangat besar, namun hanya sebagian kecil dari pendapatan mereka yang menjadi laba. Ini mencerminkan margin keuntungan yang tipis, umum terjadi di industri ritel.
- Saudi Aramco, meskipun revenue-nya lebih kecil dibanding Walmart, memiliki segmen laba yang jauh lebih besar, menunjukkan margin profit yang tinggi, tipikal di industri minyak dan gas.
- Apple dan Berkshire Hathaway juga menampilkan porsi laba yang relatif besar dibanding total pendapatan, mengindikasikan efisiensi dan profitabilitas tinggi.
- Sebaliknya, perusahaan seperti CVS Health dan State Grid Corporation of China menunjukkan rasio laba yang sangat kecil terhadap total revenue, menandakan bisnis yang sangat padat biaya operasional atau dengan struktur margin rendah.

```
rm(atlas_conn)
```