Eksplorasi dan Visualisasi Data

Bakti Siregar, M.Sc

2023-11-15

Contents

Kata Pengantar

Selamat datang dalam dunia yang penuh dengan wawasan dan pengetahuan yang mendalam tentang Eksplorasi dan Visualisasi Data! Ebook ini hadir untuk membimbing Anda melalui perjalanan menarik yang mengungkap keajaiban di balik setiap dataset.

Dalam era digital ini, data menjadi elemen kunci dalam pengambilan keputusan, dan kemampuan untuk menggali inti dari setiap informasi adalah suatu keahlian yang sangat bernilai. Dengan tekad untuk memahami dan menguasai konsep eksplorasi dan visualisasi data, Anda akan membuka pintu menuju wawasan yang dapat mengubah cara Anda memandang dan memahami dunia sekitar.

Ebook ini dirancang untuk semua tingkatan pembaca, mulai dari pemula hingga mereka yang sudah memiliki pengalaman dalam analisis data. Setiap babnya didesain dengan teliti untuk memberikan pemahaman yang mendalam tentang konsep-konsep dasar, teknik eksplorasi data yang efektif, serta seni visualisasi data yang memukau.

Terima kasih kepada para penulis dan kontributor yang telah berusaha keras untuk menyajikan informasi yang terkini dan relevan. Semoga ebook ini tidak hanya menjadi panduan praktis, tetapi juga sumber inspirasi bagi setiap pembaca yang ingin mengembangkan keterampilan dalam mengolah dan menyajikan data dengan cara yang menarik dan bermakna.

Tak lupa, apresiasi tinggi kami sampaikan kepada Anda, pembaca setia. Semoga ebook ini memberikan nilai tambah yang signifikan bagi perjalanan Anda dalam dunia eksplorasi dan visualisasi data.

Ringkasan Materi

Data tanpa interpretasi adalah seperti bahasa tanpa arti. Sehingga, buku ini sedimikian hingga agar seni visualisasi data dapat menjadi narasi yang menginspirasi. Berikut ini adalah beberapa poin penting yang menjadi rujukan pembahasan.

- Bagaimana melakukan eksplorasi data secara menyeluruh untuk mendapatkan wawasan mendalam.
- Teknik-tekni visualisasi data yang efektif untuk menyampaikan informasi kompleks dengan jelas.
- Penerapan praktis melalui studi kasus dan proyek-proyek simulasi.

Penulis

- **Bakti Siregar, M.Sc** adalah Ketua Program Studi di Jurusan Statistika Universitas Matana. Lulusan Magister Matematika Terapan dari National Sun Yat Sen University, Taiwan. Beliau juga merupakan dosen dan konsultan Data Scientist di perusahaan-perusahaan ternama seperti JNE, Samora Group, Pertamina, dan lainnya. Beliau memiliki antusiasme khusus dalam mengajar Big Data Analytics, Machine Learning, Optimisasi, dan Analisis Time Series di bidang keuangan dan investasi. Keahliannya juga terlihat dalam penggunaan bahasa pemrograman Statistik seperti R Studio dan Python. Beliau mengaplikasikan sistem basis data MySQL/NoSQL dalam pembelajaran manajemen data, serta mahir dalam menggunakan tools Big Data seperti Spark dan Hadoop. Beberapa project beliau dapat dilihat di link berikut: Rpubs, Github, Website, dan Kaggle.

Asisten Lab

- **Juenzy Hodawya, S.Stat** adalah seorang alumni Statistika yang bersemangat dalam dunia pemrograman dan analisis data. Saat ini Juenzy bekerja di salah satu perusahaan yang melibatkan ilmu olah data yaitu Cost Control Specialist. Selama menajadi mahasiswa Statistika Universitas Matana, Juen berperan dalam membantu mahasiswa dalam memahami konsep-konsep dasar dan kompleks dalam pemrograman R dan Python. Dalam perjalanan waktu, Juen mulai mengambil tanggung jawab lebih besar dalam laboratorium. Ia membantu mengembangkan materi pembelajaran tambahan, menulis modul seperti tutorial online tentang analisis data menggunakan R dan Python. Ia juga aktif dalam berbagai proyek penelitian di bawah bimbingan dosen, yang melibatkan pengolahan data besar untuk analisis statistik, visualisasi, dan menulis jurnal.

Ucapan Terima Kasih

Saya ingin mengucapkan terima kasih yang tulus kepada semua yang telah mendukung dan berkontribusi dalam perjalanan pembuatan modul “Basis Data dan

Penelusuran Data". Modul ini tidak akan mungkin menjadi kenyataan tanpa kerja keras, semangat, dan dukungan yang luar biasa dari berbagai pihak. Terima kasih juga kepada rekan-rekan dan kolega yang telah memberikan masukan, saran, dan diskusi berharga sepanjang perjalanan penulisan modul ini. Kontribusi kalian telah membantu memperkaya isi modul dan menghadirkan sudut pandang yang beragam. Tentu saja,modul ini tidak akan lengkap tanpa rasa terima kasih kepada para peneliti dan praktisi di bidang basis data dan penelusuran data yang telah menciptakan landasan pengetahuan yang menjadi dasar dari modul ini. Pengalaman dan pengetahuan yang kalian bagikan sangat berharga. Saya juga ingin mengucapkan terima kasih kepada keluarga dan teman-teman saya atas dukungan, pengertian, dan dorongan yang tak hentihentinya. Tanpa dukungan kalian, perjalanan menulis modul ini pastinya tidak akan semudah ini.

Akhir kata, semoga modul ini dapat memberikan manfaat dan wawasan baru kepada para pembaca yang ingin mendalami dunia basis data dan penelusuran data. Ucapan terima kasih terakhir saya tujukan untuk semua yang telah berkontribusi, baik secara langsung maupun tidak langsung, dalam menghadirkan modul ini kepada para pembaca.

Masukan & Saran

Semua masukan dan tanggapan Anda sangat berarti bagi kami untuk memperbaiki template ini kedepannya. Bagi para pembaca/pengguna yang ingin menyampaikan masukan dan tanggapan, dipersilahkan melalui kontak dibawak ini!

Email: dsciencelabs@outlook.com

Chapter 1

Visualisasi Data Univariat

Visualisasi data univariat biasanya digunakan untuk melakukan distribusi data dari satu variabel. Dalam hal ini, variabel yang dimaksud dipartisi menjadi dua bagian:

- **Kategoris,** seperti; jenis kelamin, ras, negara, kota, dll.
- **Numerik,** seperti; usia, berat badan, inflasi, suku bunga, dll.

1.1 Variabel Kategori

Distribusi pada suatu variabel kategori tunggal biasanya diplot dengan diagram batang, diagram lingkaran, atau diagram pohon (tetapi ini sangat jarang sekali).

Diagram Batang

Berikut ini adalah contoh yang menunjukkan frekuensi dari dataset `Marriage`, Saya mendapatkannya dari package `mosaicData`. Kita gunakan diagram batang untuk menampilkan distribusi peserta pernikahan berdasarkan `Zodiak`.

```
library(ggplot2)                                    # untuk visualisasi
#setwd("C:/Users/Bakti/Desktop/")                   # jangan lupa mengatur working directory
Marriage<- read.csv("https://raw.githubusercontent.com/Bakti-Siregar/dataset/master/Bookdown-Data
ggplot(Marriage, aes(x = zodiacs)) +                # memplot distribusi dari `Zodiacs`
  geom_bar(fill = "cornflowerblue",
           color= "azure4") +                       # Anda dapat mengganti warna
  theme_minimal() +                                 # menggunakan tema minimal
  labs(x = "Zodiacs",                               # Anda dapat mengganti label dan judul plot
```

```
       y = "Frequency",
       title = "Marriage Participants by Zodiacs")
```

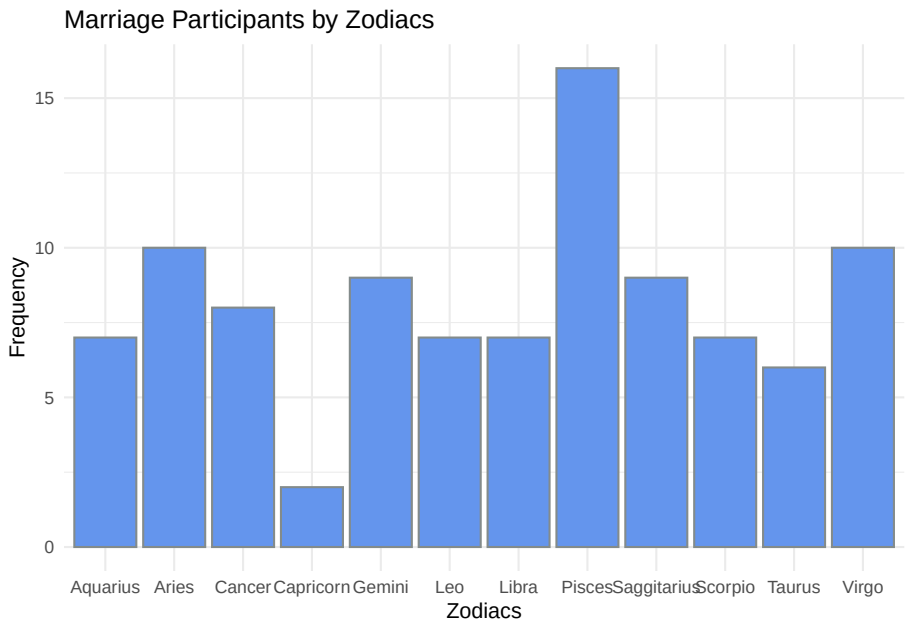
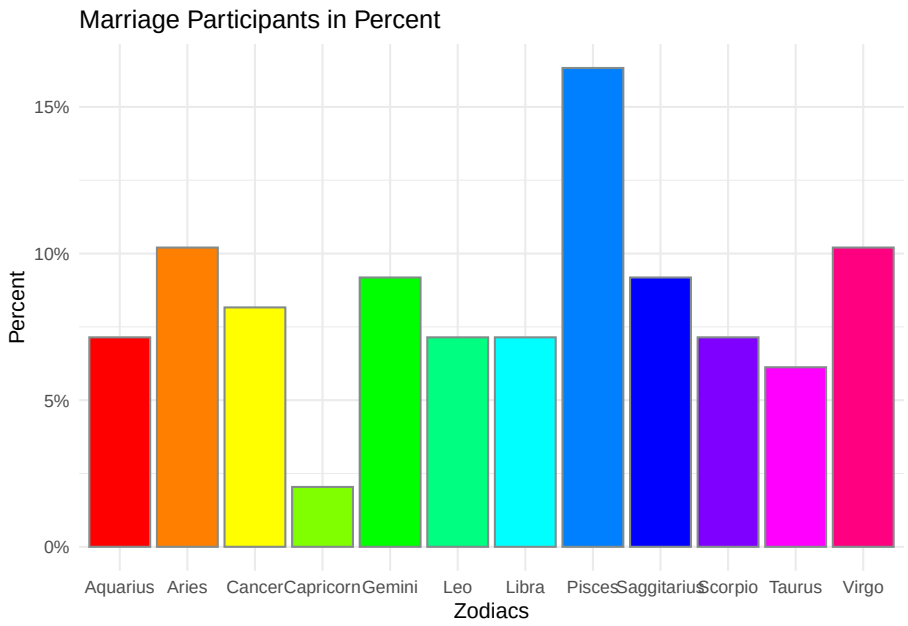


Diagram batang biasanya digunakan untuk menggambarkan persentase dari jumlah. Dalam hal ini diperlihatkan diagram batang pada data (zodiak), pada contoh ini digunakan kode `aes(x=sign)`. Selain itu biasa juga dengan menggunakan `aes(x = sign, y = ..count..)`, di mana ..count.. adalah variabel khusus yang menggambarkan frekuensi dari setiap kategori. Anda dapat menggunakan ini untuk menghitung persentase, dengan menentukan variabel y secara eksplisit seperti diperlihatkan sebagai berikut;

```
library(ggplot2)                                    # untuk visualisasi
ggplot(Marriage,
   aes(x = zodiacs,
       y = ..count.. / sum(..count..))) +
  geom_bar(fill = rainbow(12), color= "azure4") +
  theme_minimal() +                                 # menggunakan tema minimal
  labs(x = "Zodiacs",
       y = "Percent",
  title  = "Marriage Participants in Percent") +
  scale_y_continuous(labels = scales::percent)      # menambahkan simbol % untuk labe
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in
## ggplot2 3.4.0.
```

```
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see
## where this warning was generated.
```



Catatan: Dalam R, warna dapat ditentukan dengan nama (misalnya `col= "red"`) atau dengan triplet RGB heksadesimal (seperti `col = "#FFCC00"`). Anda juga dapat menggunakan sistem warna lain seperti salah satunya diambil dari package `RColorBrewer`.Lebih lanjut

Mengurutkan batang berdasarkan frekuensi sering kali membantu. Pada kode di bawah ini, frekuensi dihitung secara eksplisit. Kemudian fungsi `reorder` digunakan untuk mengurutkan kategori berdasarkan frekuensinya. Opsi `stat="identity"` memberitahu fungsi plot untuk tidak menghitung jumlah, karena mereka diberikan secara langsung.

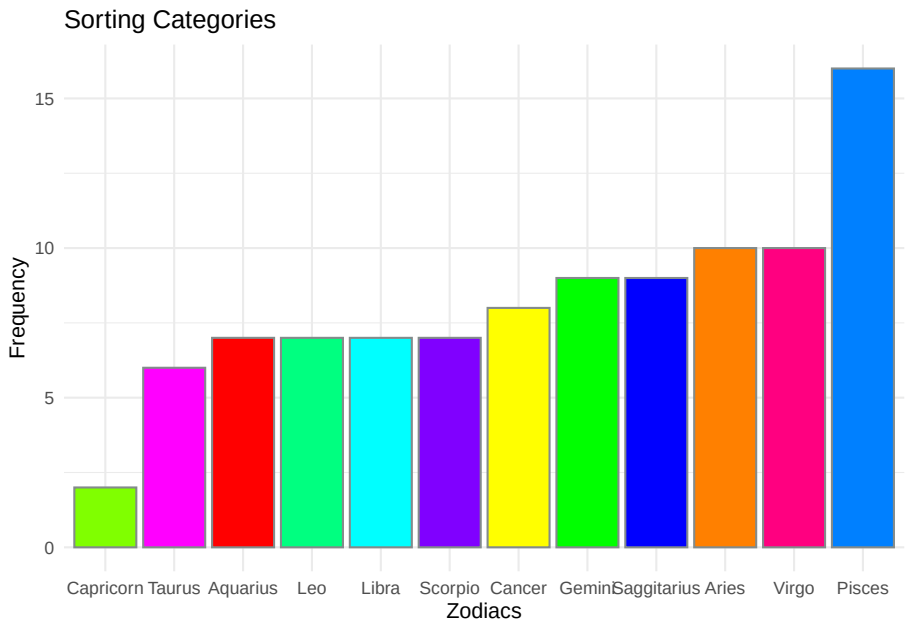
```
library(dplyr)                                    # untuk manipulasi data
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

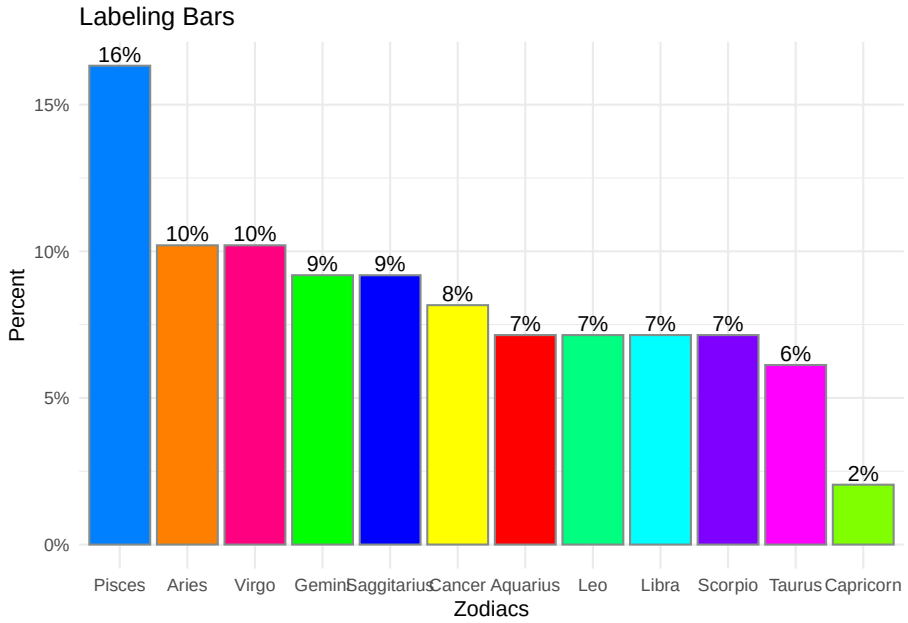
```
library(ggplot2)                                    # untuk visualisasi
plotdata <- Marriage %>%                            # memuat dataset
 count(zodiacs)                                     # jumlah peserta di setiap 'zodia
# menyusun plot batang secara meningkat
ggplot(plotdata,
       aes(x = reorder(zodiacs, n),
           y = n)) +
  geom_bar(stat = "identity",
           fill = rainbow(12),
           color= "azure4") +
   theme_minimal() +                                 # menggunakan tema minimal
  labs(x = "Zodiacs",
       y = "Frequency",
  title  = "Sorting Categories")
```



Jika Anda mungkin ingin memberi label untuk setiap batang dengan nilai numeriknya, ikuti kode berikut:

```
library(dplyr)                                      # untuk manipulasi data
library(ggplot2)                                    # untuk visualisasi
library(scales)                                     # menentukan jeda atau label seca
```

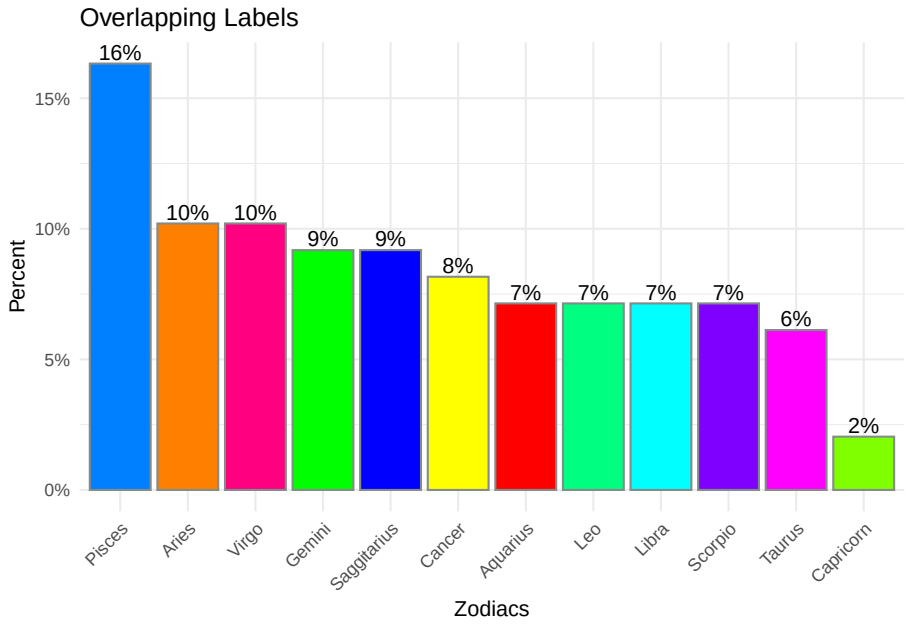
```
plotdata <- Marriage %>%
  count(zodiacs) %>%
  mutate(pct = n / sum(n),
         pctlabel = paste0(round(pct*100), "%"))
# plot batang sebagai persentase, dalam urutan menurun dengan label batang
ggplot(plotdata,
       aes(x = reorder(zodiacs, -pct),
           y = pct)) +
  geom_bar(stat = "identity",
           fill = rainbow(12),
          color = "azure4") +
  geom_text(aes(label = pctlabel),
            vjust = -0.25) +
  theme_minimal() +                                    # menggunakan tema minimal
  scale_y_continuous(labels = percent) +
  labs(x = "Zodiacs",
       y = "Percent",
       title  = "Labeling Bars")
```



Terkadang label kategori mungkin tumpang tindih, ini sangat mengganggu buka? Jadi, Anda dapat memutar label sumbu.

```
library(ggplot2)                                       # untuk visualisasi
library(scales)                                        # menentukan jeda atau label secara otomatis
```

```
# memplot batang sebagai persentase, dalam urutan menurun dengan label batang
ggplot(plotdata,
       aes(x = reorder(zodiacs, -pct),
           y = pct)) +
  geom_bar(stat = "identity",
           fill = rainbow(12),
           color = "azure4") +
  geom_text(aes(label = pctlabel),
            vjust = -0.25) +
  scale_y_continuous(labels = percent) +
  theme_minimal() +                                    # menggunakan tema minimal
  labs(x = "Zodiacs",
       y = "Percent",
       title  = "Overlapping Labels")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Cara lainnya, Anda dapat menangani situasi ini dengan membalik sumbu `x` dan `y`.

```
library(ggplot2)                                       # untuk visualisasi
library(scales)                                        # menentukan jeda atau label seca
# memplot batang sebagai persentase, dalam urutan menurun dengan label batang
ggplot(plotdata,
       aes(x = reorder(zodiacs, -pct),
```

```
          y = pct)) +
geom_bar(stat = "identity",
         fill = rainbow(12),
         color = "azure4") +
geom_text(aes(label = pctlabel),
          hjust = -0.10) +
scale_y_continuous(labels = percent) +
theme_minimal() +                                    # menggunakan tema minimal
labs(x = "Zodiacs",
     y = "Percent",
     title  = "Overlapping Labels")+
coord_flip()
```

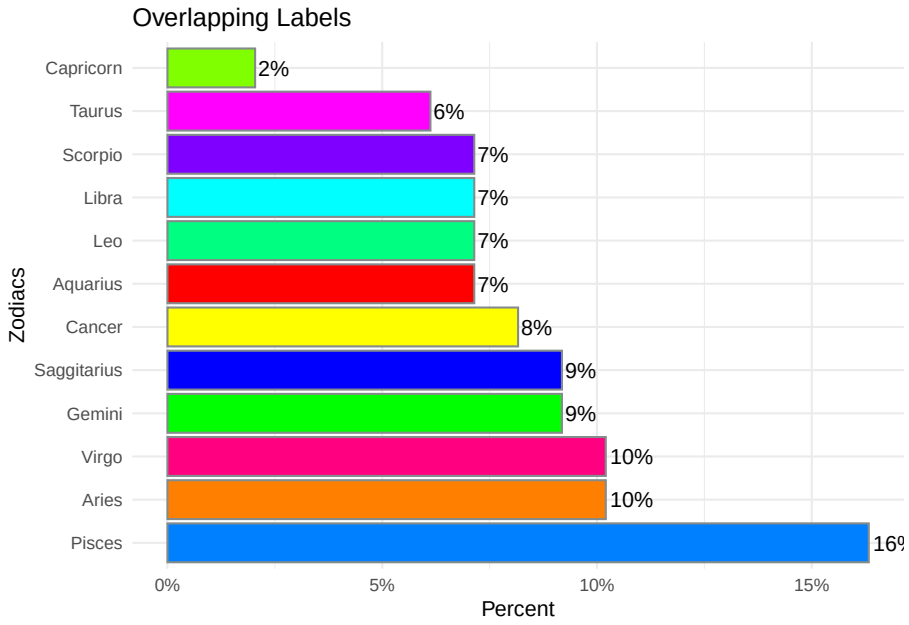


Diagram Pai

Diagram pai bersifat kontroversial di statistik. Jika tujuan Anda adalah membandingkan frekuensi dari kategori, lebih baik Anda menggunakan diagram batang (orang-orang lebih baik dalam menilai panjang batang dari pada volume irisan lingkaran). Jika tujuan Anda untuk membandingkan setiap kategori secara keseluruhan (misalnya berapa porsi partisipan yang merupakan Hispanik (orang Spanyol) dibandingkan dengan semua partisipan), dan jumlah kategorinya kecil, maka diagram pai mungkin cocok untuk Anda. Dibutuhkan sedikit lebih banyak kode untuk membuat diagram pai lebih menarik dalam R.

Ini adalah contoh untuk membuat diagram pai sederhana dengan ggplot2:

```
library(dplyr)                                          # untuk memanipulasi data
library(ggplot2)                                        # untuk visualisasi
library(scales)                                         # menentukan jeda atau label seca
# Persiapan data
plotdata <- Marriage %>%
  count(race) %>%
  arrange(desc(race)) %>%
  mutate(prop = round(n*100/sum(n), 1),
         lab.ypos = cumsum(prop) - 0.5*prop)
# Membuat diagram pai
mycols <- c("#0073C2FF", "#EFC000FF", "#868686FF", "#CD534CFF")
ggplot(plotdata, aes(x = "", y = prop, fill = race)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0)+
  geom_text(aes(y = lab.ypos, label = prop), color = "white")+
  scale_fill_manual(values = mycols) +
  theme_void()+
  labs(title = "Marriage Participants by Race")
```

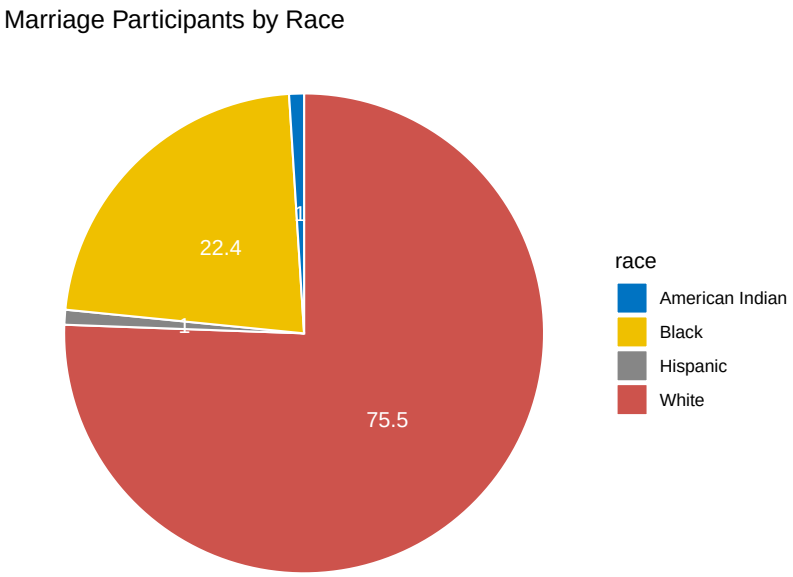
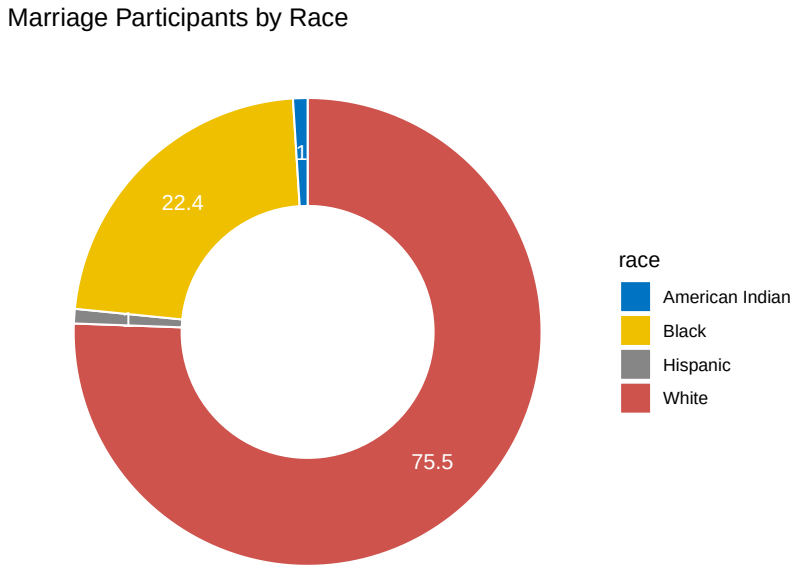


Diagram donat hanyalah diagram pai sederhana dengan lubang di dalamnya. Satu-satunya perbedaan antara kode diagram pai adalah kita menetapkan: `x = 2` dan `xlim = c(0.5, 2.5)` untuk membuat lubang di dalam diagram pai. Selain itu, argumen `width` dalam fungsi `geom_bar()` tidak lagi diperlukan.

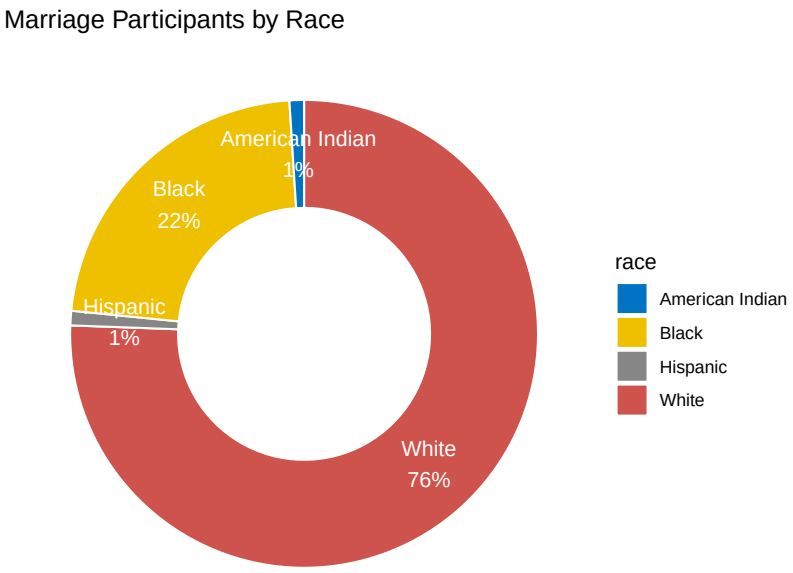
```
library(ggplot2)                                    # untuk visualisasi
library(scales)                                     # menentukan jeda atau label secara otomatis
# membuat diagram donat
ggplot(plotdata, aes(x = 2, y = prop, fill = race)) +
  geom_bar(stat = "identity", color = "white") +
  coord_polar(theta = "y", start = 0)+
  geom_text(aes(y = lab.ypos, label = prop), color = "white")+
  scale_fill_manual(values = mycols) +
  theme_void()+
  xlim(0.5, 2.5)+
  labs(title = "Marriage Participants by Race")
```



Sekarang mari berkreasi dan menambahkan label, sambil menghapus legend.

```
library(ggplot2)                                    # untuk visualisasi
library(scales)                                     # menentukan jeda atau label secara otomatis
# tambahkan label persen
plotdata$percent <- paste0(plotdata$race, "\n",
                           round(plotdata$prop), "%")
# membuat diagram donat dalam persen
ggplot(plotdata, aes(x = 2, y = prop, fill = race)) +
  geom_bar(stat = "identity", color = "white") +
  coord_polar(theta = "y", start = 0)+
  geom_text(aes(y = lab.ypos, label = percent), color = "white")+
```

```
  scale_fill_manual(values = mycols) +
  theme_void()+
  xlim(0.5, 2.5)+
  labs(title = "Marriage Participants by Race")
```



Peta Pohon

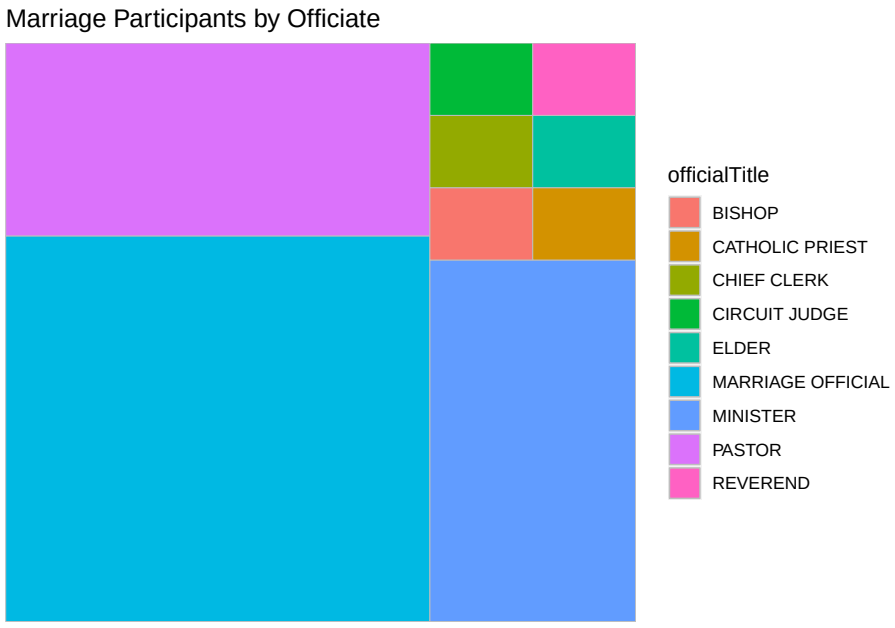
Sebuah alternatif untuk diagram pai adalah peta phon. Tidak seperti diagram pai, peta pohon dapat menangani variabel kategorikal yang memiliki banyak tingkatan.

```
library(ggplot2)                                    # untuk visualisasi
library(treemapify)                                 # untuk visualisasi
```

```
## Warning: package 'treemapify' was built under R
## version 4.3.2
```

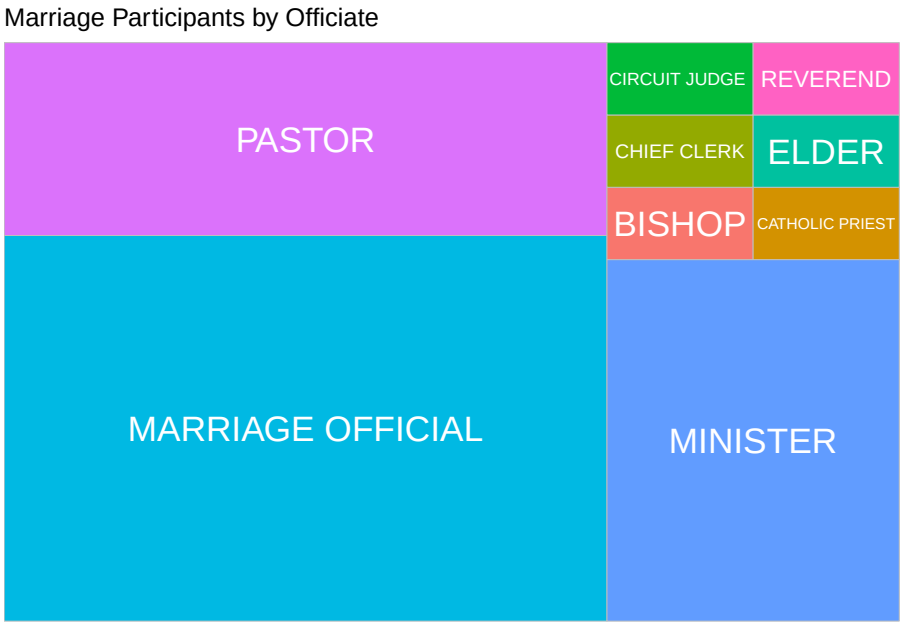
```
library(scales)                                     # menentukan jeda atau label seca
plotdata <- Marriage %>%
  count(officialTitle)
ggplot(plotdata,
       aes(fill = officialTitle,
```

```
          area = n)) +
  geom_treemap() +
  labs(title = "Marriage Participants by Officiate")
```



Berikut ini adalah versi yang lebih berguna dengan label.

```
ggplot(plotdata,
       aes(fill = officialTitle,
           area = n,
           label = officialTitle)) +
  geom_treemap() +
  geom_treemap_text(colour = "white",
                    place = "centre") +
  labs(title = "Marriage Participants by Officiate") +
  theme(legend.position = "none")
```



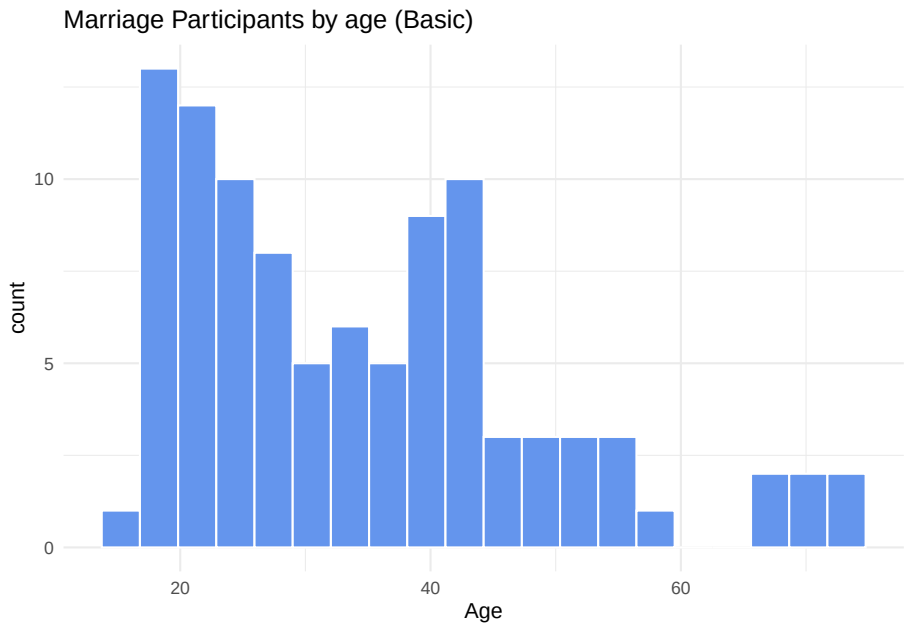
1.2 Variabel Kontinu

Distribusi variabel kuantitatif tunggal biasanya diplot dengan histogram, plot densitas kernel, atau plot titik.

Histogram

Menggunakan dataset Marriage, mari kita plot usia dari peserta pernikahan.

```
library(ggplot2)                                    # untuk visualiSaSi
ggplot(Marriage, aes(x = age)) +
  geom_histogram(fill = "cornflowerblue",
                 color = "white",bins = 20) +
  theme_minimal() +                                 # menggunakan tema minimal
  labs(title="Marriage Participants by age (Basic)",
       x = "Age")
```



Sebagian besar peserta tampaknya berusia 20-an tahun dengan kelompok lain berusia 40-an tahun, dan kelompok yang lebih kecil berusia 60-an dan 70-an tahun. Ini akan menjadi distribusi multimodal. Warna histogram dapat diganti menggunakan dua opsi:

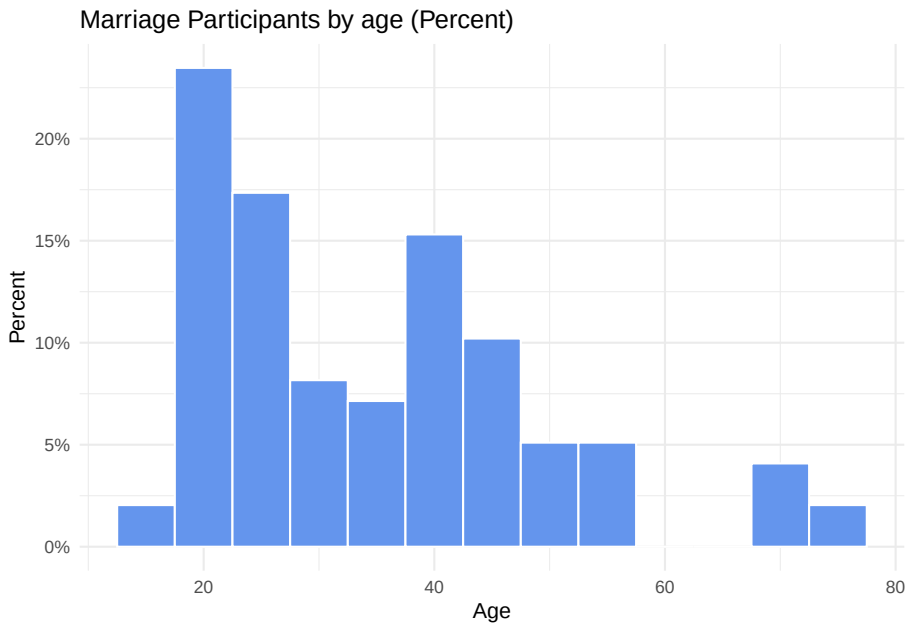
- fill - warna isi untuk batang.
- color - warna batas di sekitar batang.

Cara lainnya adalah Anda dapat menggunakan **binwidth**, lebar nampan yang diwakili oleh batang.

```
library(ggplot2)                                    # untuk visualisaSi
library(scales)                                     # menentukan jeda atau label secara otomatis
ggplot(Marriage,
       aes(x = age,
           y= ..count.. / sum(..count..))) +
  geom_histogram(fill = "cornflowerblue",
                 color = "white",
                 binwidth = 5) +
  theme_minimal() +                                 # menggunakan tema minimal
  labs(title="Marriage Participants by age (Alternative Bins and bandwidths)",
       y = "Percent",
       x = "Age") +
  scale_y_continuous(labels = percent)
```

Seperti diagram batang, sumbu y dapat mewakili jumlah atau persen dari total.

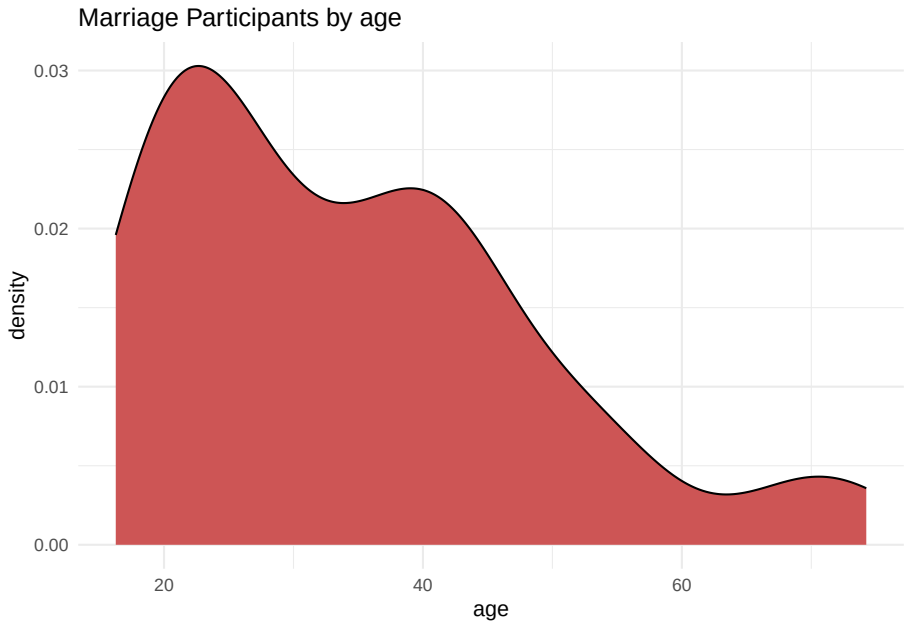
```
library(ggplot2)                                    # untuk visualisasi
library(scales)                                     # menentukan jeda atau label seca
ggplot(Marriage,
       aes(x = age,
           y= ..count.. / sum(..count..))) +
  geom_histogram(fill = "cornflowerblue",
                 color = "white",
                 binwidth = 5) +
  theme_minimal() +                                 # menggunakan tema minimal
  labs(title="Marriage Participants by age (Percent)",
       y = "Percent",
       x = "Age") +
  scale_y_continuous(labels = percent)
```



Plot Densitas Kernel

Alternatif untuk histogram adalah plot densitas Kernel. Secara teknis, perkiraan densitas kernel adalah metode non-parametrik untuk memperkirakan fungsi densitas probabilitas dari variabel acak kontinu. (Apa??) Pada dasarnya, kita mencoba untuk menggambar histogram yang diperhalus, di mana area di bawah kurva sama dengan satu.

```r
library(ggplot2)                                    # untuk visualisaSi
ggplot(Marriage, aes(x = age)) +
  geom_density(fill = "indianred3") +
  theme_minimal() +                                 # menggunakan tema minimal
  labs(title = "Marriage Participants by age")
```



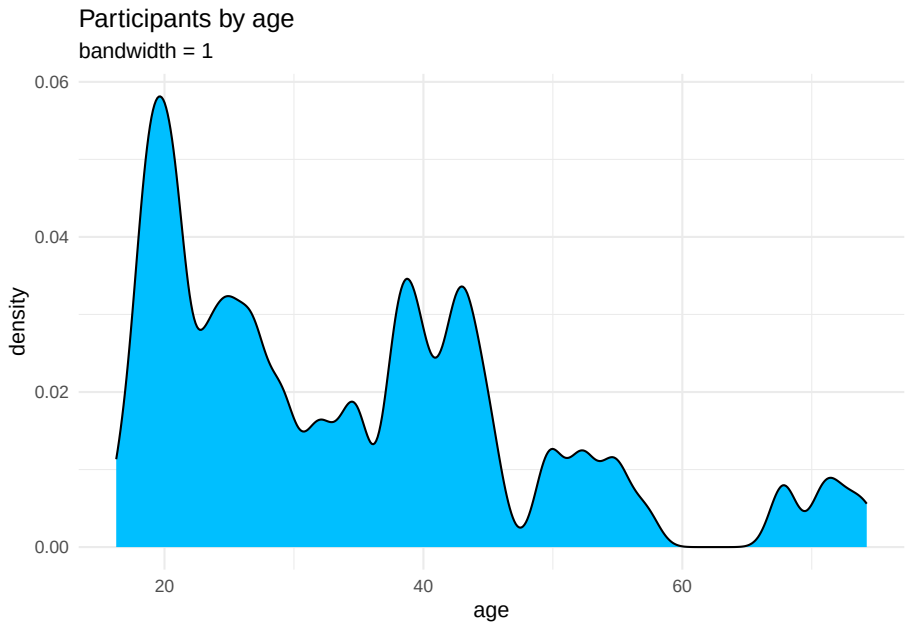
Grafik menunjukkan distribusi dari nilai. Sebagai contoh, perbandingan kasus antara 20 dan 40 tahun akan diwakili oleh area di bawah kurva antara 20 dan 40 pada sumbu x. Seperti diagram sebelumnya, kita juga dapat menggunakan `fill` dan `color` untuk menentukan warna isian dan batasannya.

Parameter Penghalusan (Smoothing)

Tingkat kehalusan dikontrol oleh parameter bandwidth `bw`. Untuk menemukan nilai default untuk variabel tertentu, gunakan fungsi `bw.nrd0`. Nilai yang lebih besar akan menghasilkan penghalusan yang lebih banyak, sedangkan nilai yang lebih kecil akan menghasilkan penghalusan yang lebih sedikit.

```r
library(ggplot2)                                    # untuk visualisaSi
bw.nrd0(Marriage$age)                               # default bandwidth untuk variabel usia
ggplot(Marriage, aes(x = age)) +
  geom_density(fill = "deepskyblue",
               bw = 1) +
```

```
  theme_minimal() +                                    # menggunakan tema manual
  labs(title = "Participants by age",
       subtitle = "bandwidth = 1")
```



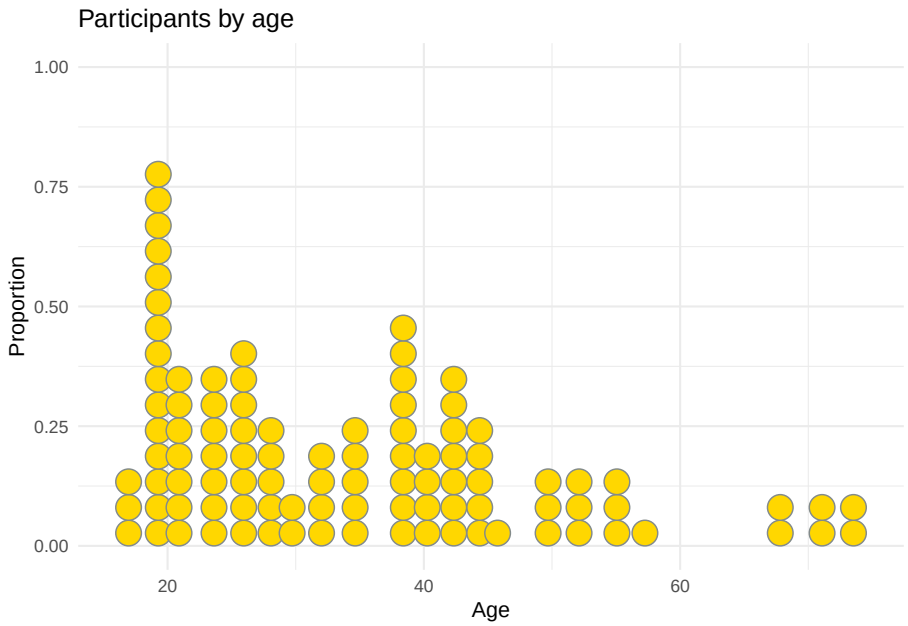
Plot densitas kernel memungkinkan Anda dengan mudah melihat skor mana yang paling sering dan mana yang relatif jarang. Namun sulit untuk menjelaskan arti sumbu y kepada seorang non-ahli statistik. (Tapi itu akan membuatmu terlihat sangat pintar!)

Diagram Titik

Alternatif lain untuk histogram adalah diagram titik. Sekali lagi, variabel kuantitatif dibagi menjadi beberapa kelompok, tetapi bukan berbentuk batang, setiap pengamatan ditentukan oleh sebuah titik. Secara default, lebar dari sebuah titik sama dengan lebar bin, dan titik-titik bertumpuk, dengan setiap titik mewakili satu pengamatan. Ini bekerja dengan baik jika jumlah pengamatan kecil (katakanlah, kurang dari 150). Opsi `fill` dan `color` dapat digunakan untuk menentukan warna isian dan batasan masing-masing titik.

```
library(ggplot2)                                       # untuk visualiSasi
ggplot(Marriage, aes(x = age)) +
  geom_dotplot(fill = "gold",
               color = "azure4",
```

```
              binwidth = 2) +
theme_minimal() +                                    # menggunakan tema minimal
labs(title = "Participants by age",
     y = "Proportion",
     x = "Age")
```



Ada lebih banyak pilihan yang tersedia. Klik di sini untuk detail dan contoh.

1.3 Praktikum

Chapter 2

Visualisasi_Data_Univariat

2.1 Praktikum

Chapter 3

Visualisasi Data Multivariat

3.1 Praktikum

Chapter 4

Referensi