Heart Attack Analysis & Prediction Data_Wrangling

2023-03-01

Import Library

mengimport library yang dibutuhkan

```
library(dplyr)
```

```
##
## 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)
library(tidyr)
library(stringr)
library(readr)
```

Membaca dataset

```
df <- read.csv("C:/Users/Repets/Downloads/archive (7)/heart.csv")
glimpse(df)
```

```
## Rows: 303
## Columns: 14
## $ age      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~
## $ sex      <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~
## $ cp       <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0~
## $ trtbps   <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~
## $ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
## $ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ restecg  <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1~
## $ thalachh <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
## $ exng     <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~
## $ slp      <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~
## $ caa      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0~
## $ thall    <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
## $ output   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

Mengecek informasi dataset

```
str(df)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thall   : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

Mengubah tipe data dari beberapa kolom

```
df$sex <- factor(df$sex, labels = c("Female", "Male"))
df$cp <- factor(df$cp, labels = c("Typical angina", "Atypical angina", "Non-Anginal pain", "Asymptomati
df$fbs <- factor(df$fbs, labels = c("No", "Yes"))
df$restecg <- factor(df$restecg, labels = c("Normal", "ST-T wave abnormality", "Left ventricular hypert
df$exng <- factor(df$exng, labels = c("No", "Yes"))
df$slp <- factor(df$slp, labels = c("Upsloping", "Flat", "Downsloping"))
df$caa <- as.factor(df$caa)
```

karena kesalahan pada saat ingin mentranformasi data pada kolom thall, error tersebut terjadi karena panjang labels yang diberikan pada fungsi factor() tidak sesuai. Fungsi factor() membutuhkan vektor labels yang panjangnya sama dengan jumlah nilai unik pada kolom yang akan diubah menjadi factor. Dalam kasus ini kolom thall memiliki tiga nilai unik yaitu 1, 2, 3.

sehingga jika ingin mengubah kolom thall menjadi factor dengan label tertentu, vektor labels yang diberikan harus memiliki panjang yang sama dengan jumlah nilai unik pada kolom tersebut. jika ingin memberikan label pada masing-masing nilai unik, maka panjang labels harus sama dengan jumlah nilai unik yaitu 3

```
#membersihkan kolom thall
unique(df$thall)
```

```
## [1] 1 2 3 0
```

```
df$thall[df$thall == "?"] <- NA

df$thall <- as.numeric(df$thall)

median_thall <- median(df$thall, na.rm = TRUE)
df$thall[is.na(df$thall)] <- median_thall
```

```r
#mengubah kolom thall menjadi factor dengan label Normal, Fixed defect, reversible defect
df$thall <- factor(df$thall, levels = c(1, 2, 3), labels = c("Normal", "Fixed defect", "reversible defe
```

memeriksa missing value

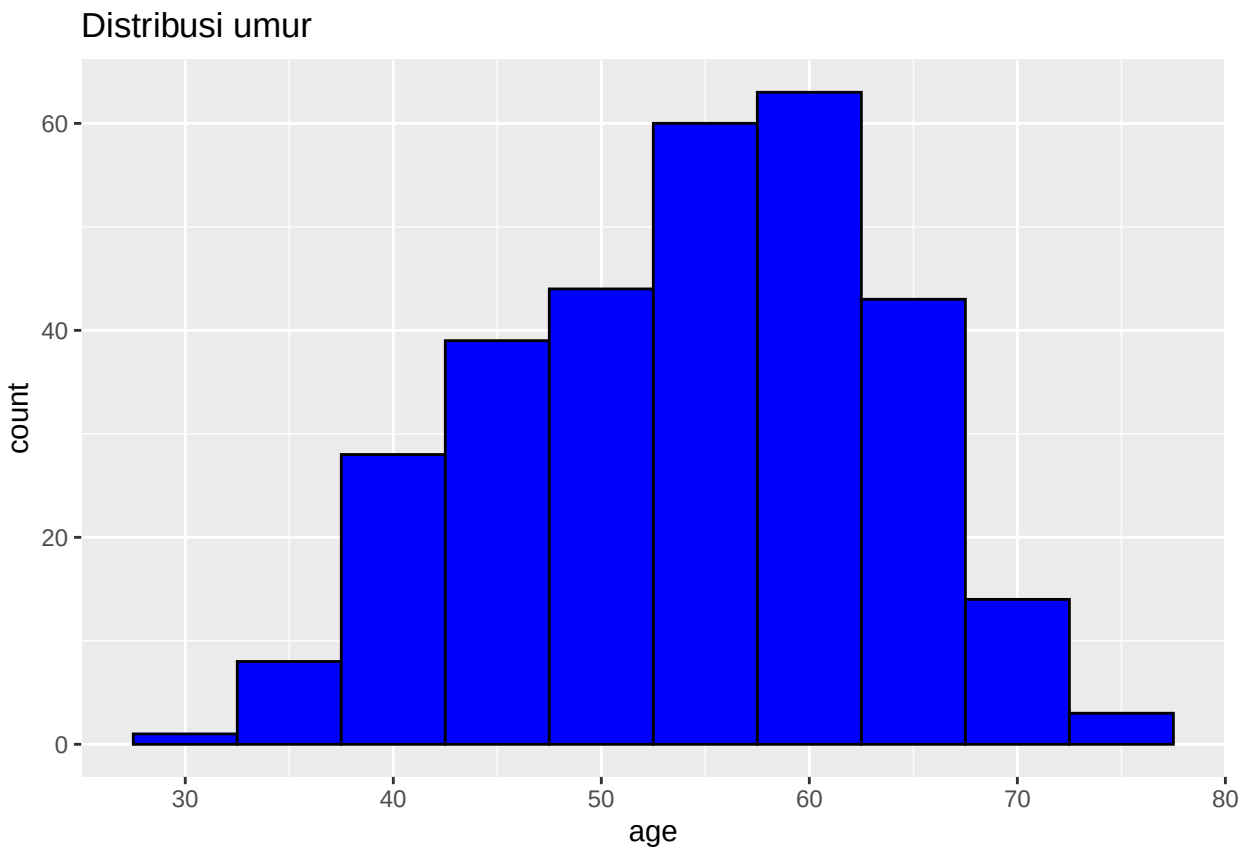
```r
sapply(df, function(x) sum(is.na(x)))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        2        0
```

pada output diatas tidak terdapat missing value pada dataset ini

eksplorasi data

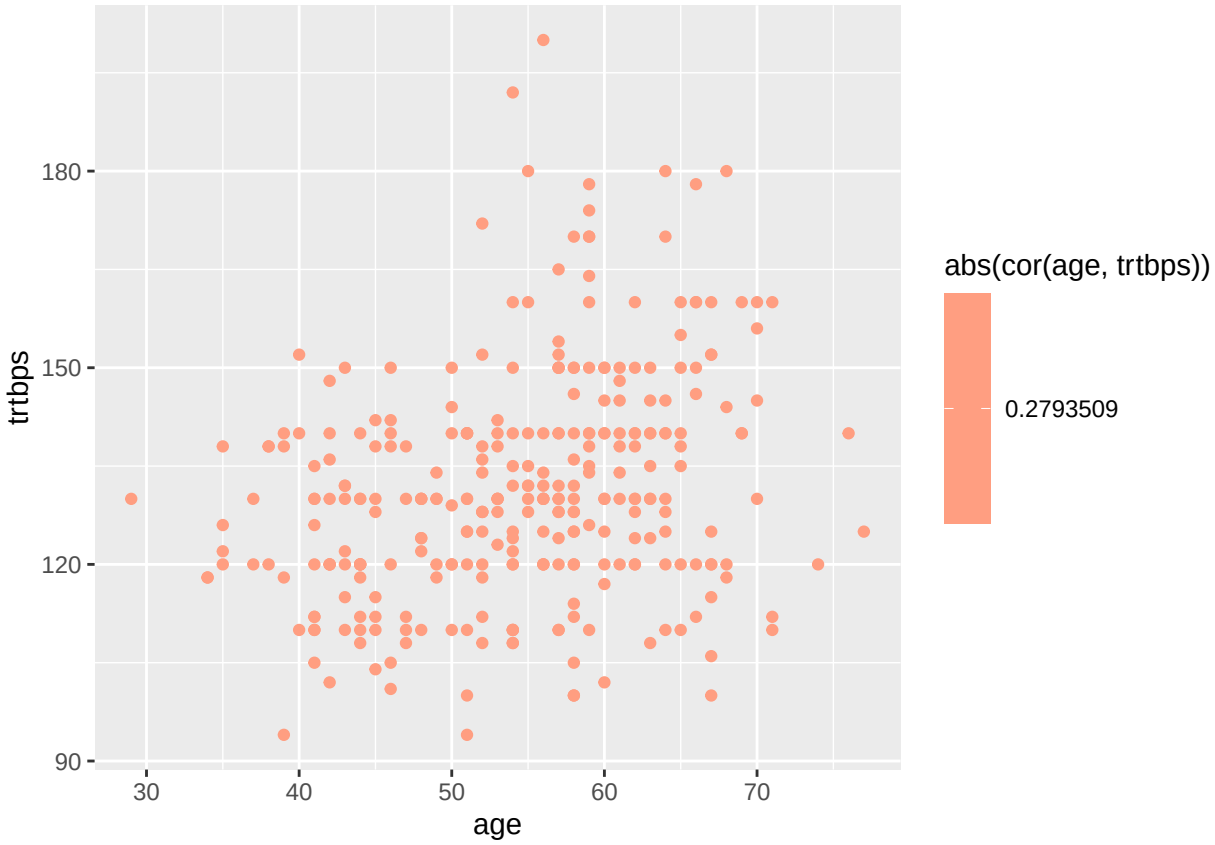
```r
#distribusi umur
ggplot(df, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  ggtitle("Distribusi umur")
```



dari grafik distribusi umur diatas, dapat dilihat bahwa mayoritas pasien berusia antara 40-70 tahun dengan puncak terbanyak pada rentang usia 55-60 tahun

```
#hubungan antara umur dan tekanan darah istirahat
library(ggplot2)

ggplot(data = df, aes(x = age, y = trtbps, color = abs(cor(age, trtbps)))) +
  geom_point() +
  scale_color_gradient(low = "white", high = "red")
```



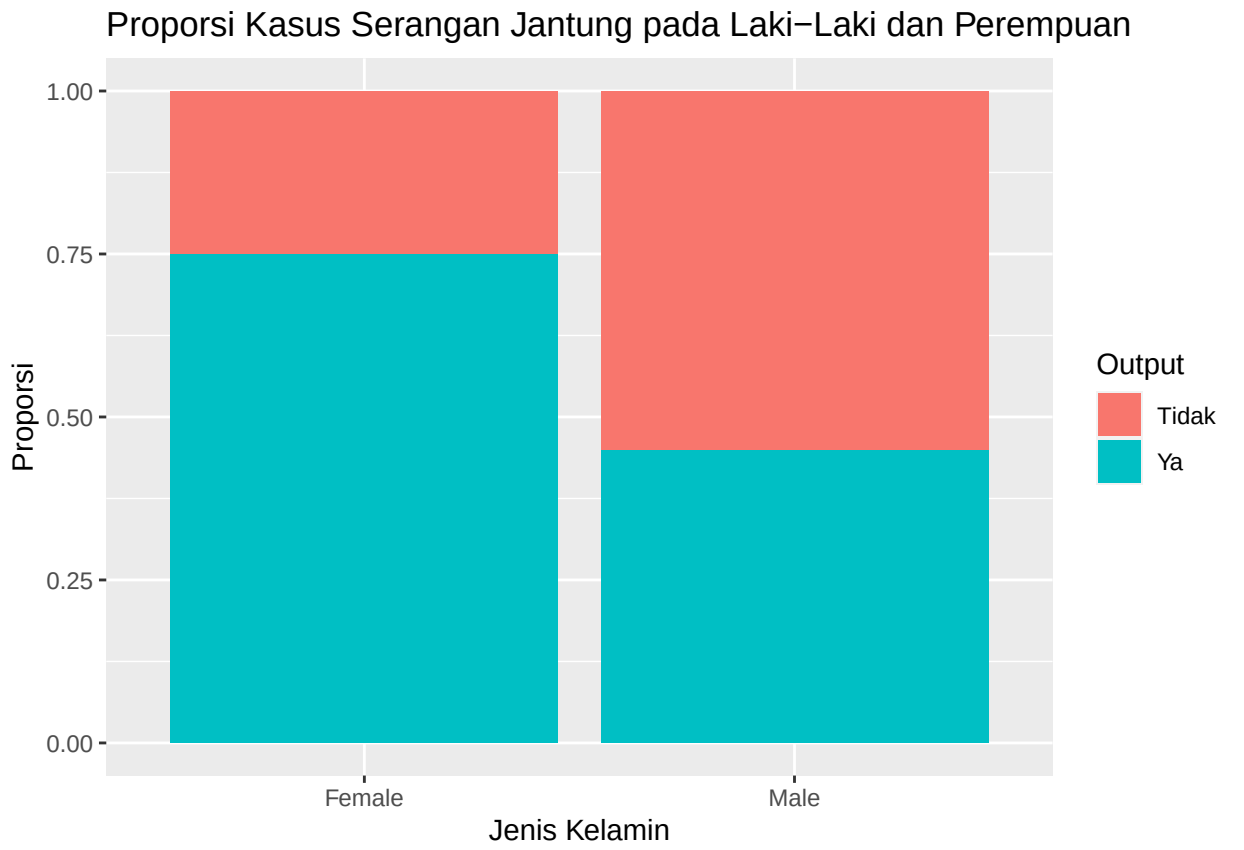
abs(cor(age, trtbps)) digunakkan sebagai nilai warna pada color aes, yang merupakan nilai korelasi(dalam bentuk absolut) antara variabel age, dan trtbps. scale_color_gradient digunakan untuk mengatur skala warna dari rendah ke tinggi. warna yang dipilih adalah dari putih ke merah.

pada scatter plot dengan warna yang menunjukkan korelasi antara variabel age dan trtbps semakin merah warna titik, semakin tinggi nilai korelasinya.

dari grafik tersebut, dapat dilihat bahwa korelasi anatar variabel age dan trtbps cukup rendah, ditunjukkan oleh warna titik yang cenderung tidak berubah. hal ini menunjukan bahwa tidak terlalu ada keterkaitan yang kuat antara usia dan tekanan darah istirahat. sehingga usia tidak dapat dijadikan factor utama dalam memprediksi tekanan darah istirahat. Namun, terdapat beberapa titik dengan warna yang lebih merah yang menunjukan adanya korelasi yang lebih tinggi di antara titik-titik tersebut.

```
#proporsi kasus serangan jantung antara laki-laki dan perempuan
ggplot(data = df, aes(x = factor(sex), fill = factor(output))) +
  geom_bar(position = "fill") +
  labs(x = "Jenis Kelamin", y = "Proporsi", fill = "Output") +
```

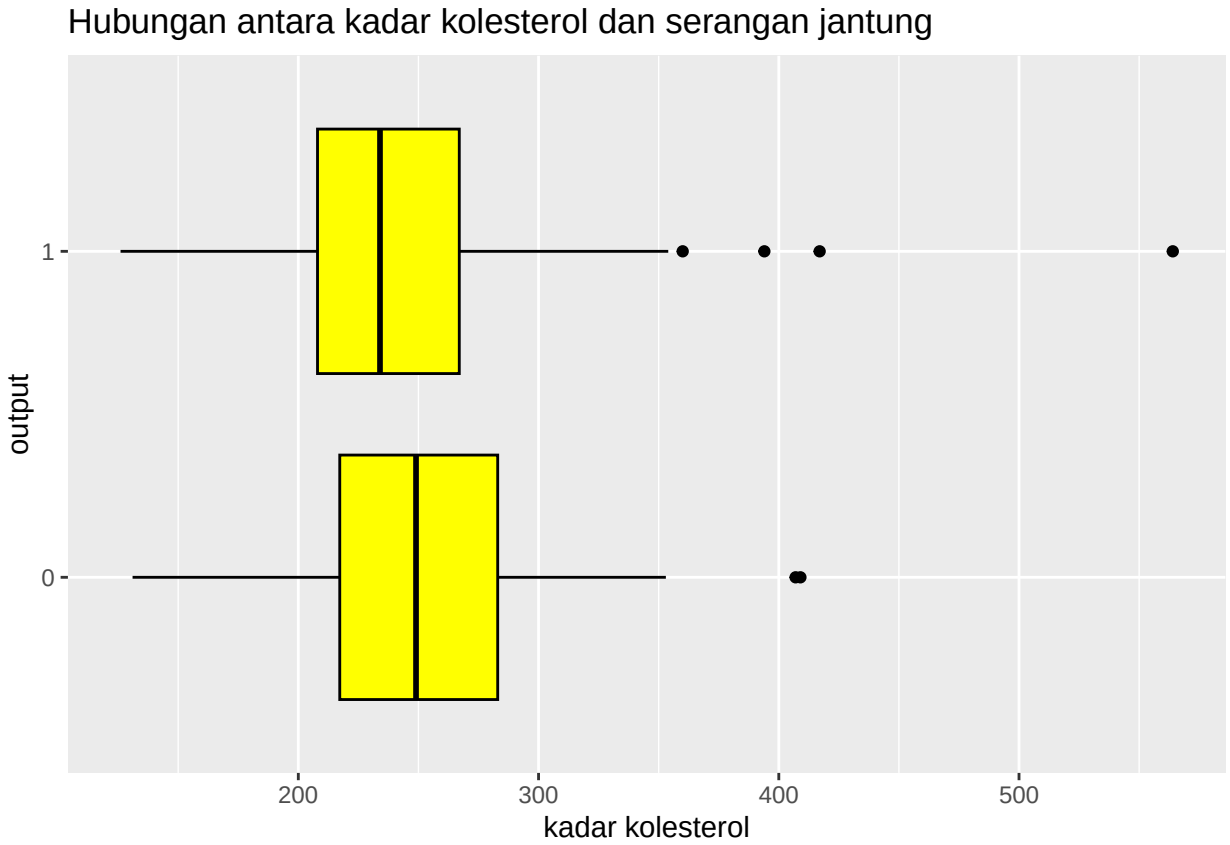
```
  scale_fill_discrete(name = "Output", labels = c("Tidak", "Ya")) +
  ggtitle("Proporsi Kasus Serangan Jantung pada Laki-Laki dan Perempuan")
```



dari grafik proporsi serangan jantung pada laki-laki dan perempuan, terlihat bahwa proporsi kasus serangan jantung pada perempuan yang positif(yang mengalami serangan jantung) cenderung lebih tinggi dibandingkan dengan proporsi pada laki-laki.

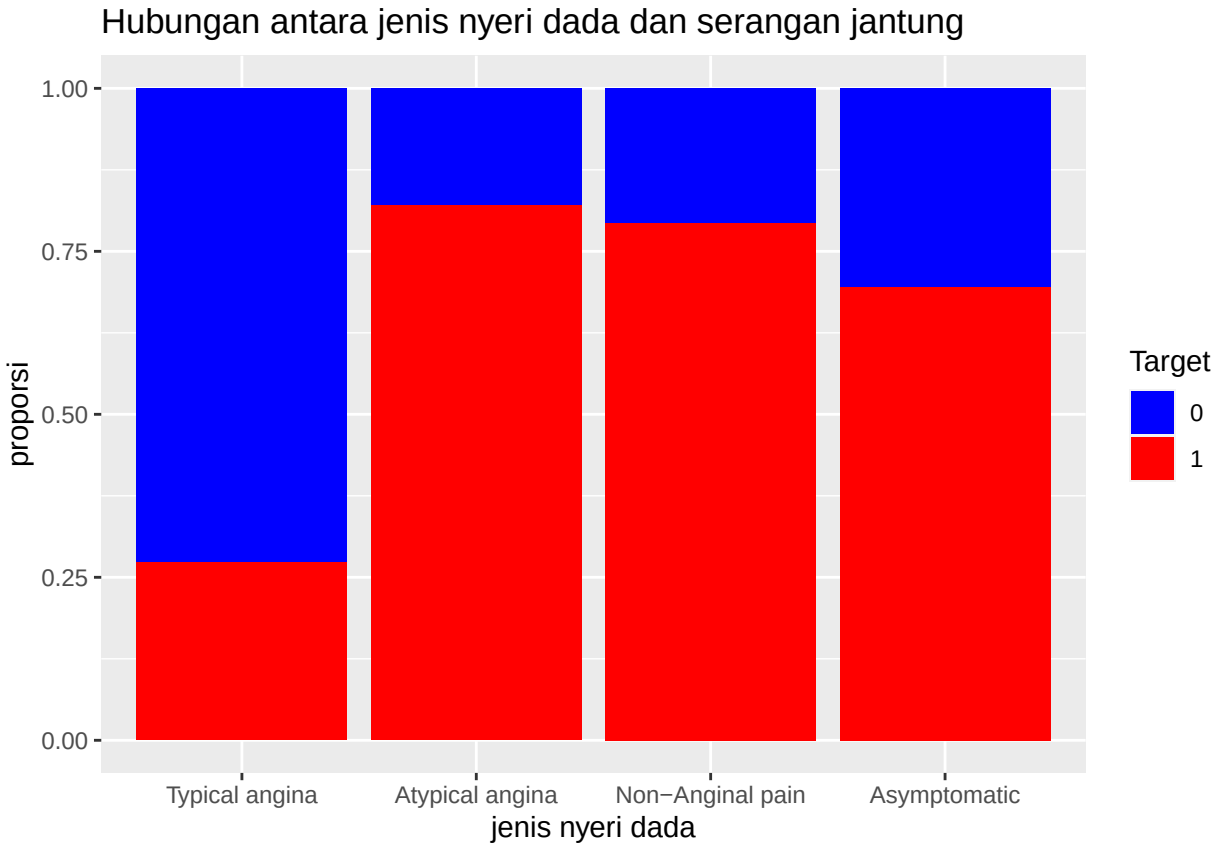
```
df$output <- as.factor(df$output)

ggplot(df, aes(x = chol, y = output)) +
  geom_boxplot(fill = "yellow", color = "black") +
  ggtitle("Hubungan antara kadar kolesterol dan serangan jantung") +
  labs(x = "kadar kolesterol", y = "output")
```



sebelum itu, saya merubah variabel output menjadi factor, sehinggan gglot() dalam memplot data dengan benar. dari grafik hubungan antara kadar kolesterol dan serangan jantung di atas, terlihat bahwa pasien yang mengalami serangan jantung memiliki kadar kolesterol yang lebih tinggi dibandingkan dengan pasien yang tidak mengalami serangan jantung. selain itu, terlihat adanya beberapa outlier pada kedua kelompok yang memiliki kadar kolesterol yang sangat tinggi

```
#hubungan antara jenis nyeri dada dan serangan jantung
ggplot(df, aes(x = cp, fill = factor(output))) +
  geom_bar(position = "fill") +
  ggtitle("Hubungan antara jenis nyeri dada dan serangan jantung") +
  labs(fill = "Target", x = "jenis nyeri dada", y = "proporsi") +
  scale_fill_manual(values = c("blue", "red"))
```



dari grafik hubungan antara jenis nyeri dada dan serangan jantung di atas, terlihat bahwa pasien yang mengalami serangan jantung cenderung memiliki jenis nyeri dada yang lebih sering dan lebih parah dibandingkan dengan yang tidak mengalami serangan jantung. namun, terdapat pula pasien yang tidak mengalami serangan jantung namun mengalami jenis nyeri dada yang parah.