Korelasi Pearson

Dian Ramadhani

08/01/2020

Korelasi Pearson

Korelasi Pearson memiliki nilai antara nilai -1 sampai dengan 1. Semakin mendekati 1 maka korelasi semakin kuat sedangkan semakin mendekati nol maka korelasi antara dua variabel semakin rendah.

Tanda koefisien korelasi menunjukkan arah hubungan. Tanda negatif (-) menunjukkan hubungan yang berkebalikan. Tanda (+) menunjukkan hubungan yang searah. Berkebalikan artinya semakin meningkat nilai suatu variabel maka variabel lainnya semakin menurun. Searah artinya semakin meningkat nilai suatu variabel maka variabel lainnya ikut meningkat.

Install Packages

Paket merupakan gabungan kode, data, dokumentasi, dan tes, yang telah dibuat oleh suatu pihak. Paket yang telah dibuat dapat digunakan oleh orang lain. Sebagian besar paket tidak tersedia secara otomatis di R sehingga perlu untuk dilakukan tindakan penginstalan.

```
# Menginstall package(s)
install.packages("readr") # membaca file
install.packages("psych") # analisis statistik
install.packages("here") # menampilkan direktori
```

Import Library

Setelah paket diinstal, paket tersebut tidak secara otomatis aktif. Dengan demikian, paket yang telah diinstal selanjutnya akan diaktifkan melalui perintah berikut.

```
# Mengaktifkan package(s)
library(readr)
library(psych)
library(here)
```

Menampilkan Direktori

Apabila ingin mengetahui lokasi pekerjaan saat ini, dapat dilakukan dengan perintah here().

```
# Mengetahui direktori proyek
here()
```

```
## [1] "C:/Users/Dhito/Desktop/r_statistik/r_statistik"
```

Misalnya, pada perangkat ini, pekerjaan berada di direktori “C:/Users/Dhito/Desktop/r_statistik/r_statistik”

Import Data

Langkah pertama yang dilakukan setelah mempersiapkan paket yaitu mengimpor data yang akan digunakan.

Data dicari dengan menggunakan here(“folder”, “subfolder”, "file.csv)

“C:/Users/Dhito/Desktop/r_statistik/r_statistik” telah diwakili oleh perintah “here”

Data yang akan diimpor bernama “correlation_salary.csv”.

Terletak di folder “C:/Users/Dhito/Desktop/r_statistik/r_statistik/data/raw”.

Dengan demikian, data tsb dapat kita impor dengan perintah sebagai berikut:

```
# Mengimport data
df.salary <- read_csv(here("data", "raw", "correlation_salary.csv"))
```

```
## Parsed with column specification:
## cols(
##   yearsexperience = col_double(),
##   absencescore = col_double(),
##   failurescore = col_double(),
##   salary = col_double()
## )
```

Data ini berisi empat variabel yaitu lamanya bekerja (yearsexperience), skor ketidakhadiran (absencescore), skor kegagalan (failurescore), dan jumlah gaji (salary).

Data "correlation_salary.csv" diimpor sebagai tabel bernama "df.salary".

Pada pekerjaan kali ini, kita akan menghitung nilai korelasi antar variabel, memvisualisasikan hasil korelasi, hingga menyimpan hasil.

Eksplorasi Data

Data yang telah diimpor selanjutnya dieksplorasi terlebih dahulu untuk mengetahui strukturnya.

```
# Menampilkan nama kolom
names(df.salary)
```

```
## [1] "yearsexperience" "absencescore"    "failurescore"    "salary"
```

```
# Menampilkan beberapa data teratas
head(df.salary)
```

```
## # A tibble: 6 x 4
##   yearsexperience absencescore failurescore salary
##             <dbl>        <dbl>        <dbl>  <dbl>
## 1             1.1          0.3        1      39343
## 2             1.3          0.6        0.979  46205
## 3             1.5          0.5        0.905  37731
## 4             2            0.7        0.895  43525
## 5             2.2          0.1        0.842  39891
## 6             2.9          0.4        0.811  56642
```

Tabel "df.salary" memiliki kolom dengan nama yearsexperience, absencescore, failurescore, dan salary.

```
# Menampilkan dimensi tabel
dim(df.salary)
```

```
## [1] 30  4
```

Tabel "df.salary" memiliki dimensi 30 baris dan 4 kolom.

```
# Menampilkan struktur data
str(df.salary)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 30 obs. of  4 variables:
##  $ yearsexperience: num  1.1 1.3 1.5 2 2.2 2.9 3 3.2 3.2 3.7 ...
##  $ absencescore   : num  0.3 0.6 0.5 0.7 0.1 0.4 0.2 0.1 0.4 0.9 ...
##  $ failurescore   : num  1 0.979 0.905 0.895 0.842 ...
##  $ salary         : num  39343 46205 37731 43525 39891 ...
##  - attr(*, "spec")=
```

```
##   .. cols(
##   ..   yearsexperience = col_double(),
##   ..   absencescore = col_double(),
##   ..   failurescore = col_double(),
##   ..   salary = col_double()
##   .. )
```

"df.salary" merupakan data dengan kelas tabel (atau 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame'). Keempat variabel didalamnya termasuk dalam jenis data numerik.

```
# Menampilkan rangkuman data
summary(df.salary)
```

```
##  yearsexperience   absencescore     failurescore         salary
##  Min.   : 1.100   Min.   :0.1000   Min.   :0.01053   Min.   : 37731
##  1st Qu.: 3.200   1st Qu.:0.3000   1st Qu.:0.23158   1st Qu.: 56721
##  Median : 4.700   Median :0.4000   Median :0.38947   Median : 65237
##  Mean   : 5.313   Mean   :0.4933   Mean   :0.45404   Mean   : 76003
##  3rd Qu.: 7.700   3rd Qu.:0.6750   3rd Qu.:0.70526   3rd Qu.:100545
##  Max.   :10.500   Max.   :1.0000   Max.   :1.00000   Max.   :122391
```

Perintah ini menampilkan nilai minimum, Q1, median, mean, Q3, dan maksimum dari masing - masing variabel.

```
# Mengetahui jumlah data kosong
sum(is.na(df.salary))
```

```
## [1] 0
```

Tabel "df.salary" tidak memiliki data kosong.

Jika terdapat data kosong, maka baris dengan data tsb dapat dihilangkan dengan perintah berikut.

```
# Mengetahui baris dengan data kosong
which(is.na(df.salary))
```

```
## integer(0)
```

```
# Menghilangkan baris dengan data kosong
df.salary2 <- na.omit(df.salary)
```

Menghitung Korelasi

Setelah mengeksplorasi data dan memastikan bahwasannya data sudah layak untuk dioleh, kita dapat melakukan perhitungan nilai korelasi melalui perintah berikut.

```
# Menghitung korelasi
correlation <- cor(df.salary, method = "pearson") # lainnya "kendall" "spearman")

# Menampilkan hasil perhitungan korelasi
correlation
```

```
##                 yearsexperience absencescore failurescore     salary
## yearsexperience       1.0000000    0.2407549   -0.9520912  0.9782416
## absencescore          0.2407549    1.0000000   -0.2126867  0.2269139
## failurescore         -0.9520912   -0.2126867    1.0000000 -0.9138655
## salary                0.9782416    0.2269139   -0.9138655  1.0000000
```

```
# Menampilkan hasil perhitungan korelasi dalam bentuk tabel
df.correlation <- data.frame(correlation)
```

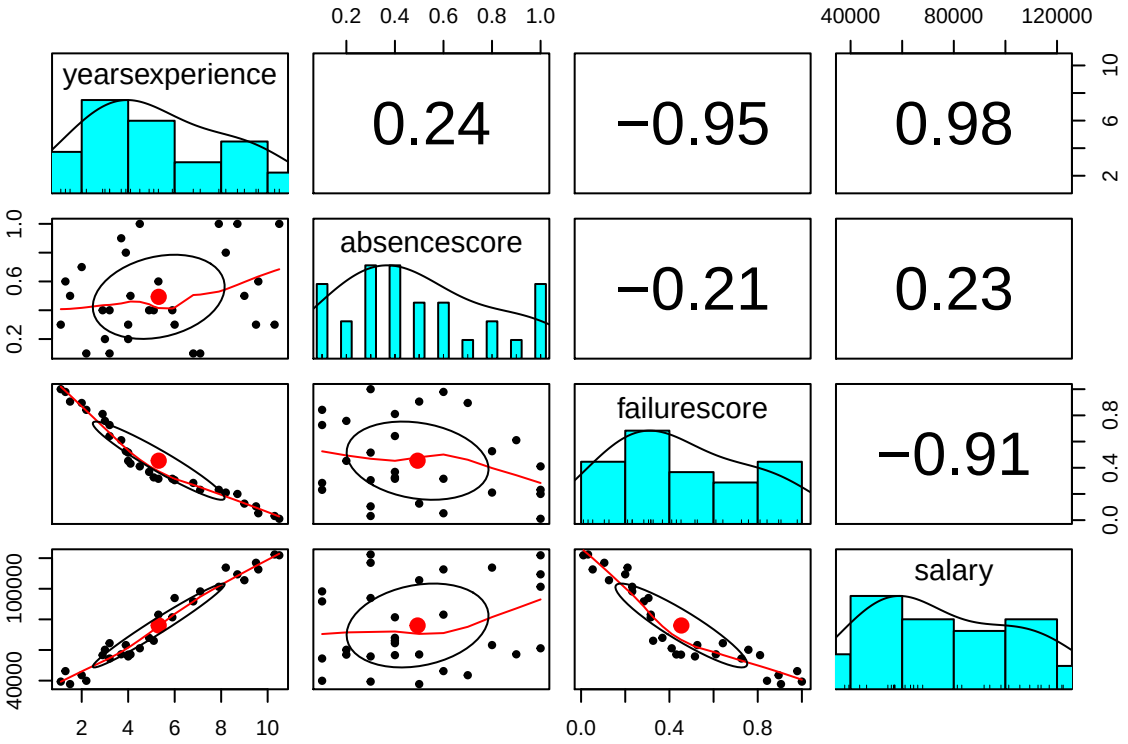
```r
View(df.correlation)
```

Menampilkan hasil korelasi pearson antar empat variabel pada tabel “df.salary”

Visualisasi Korelasi

Selain itu, kita juga dapat memvisualisasikan hasil perhitungan korelasi melalui perintah berikut.

```r
# Visualisasi korelasi
pairs.panels(df.salary, method = "pearson") # lainnya "kendall" "spearman")
```



Tes Korelasi

Tes korelasi antara variabel “absencescore” pada “df.salary” dan “failurescore” pada “df.salary” ditampilkan melalui perintah berikut.

Tes dapat dilakukan secara two.sided, less, maupun greater.

Tes dapat dilakukan dengan metode korelasi pearson, kendall, atau spearman.

Tes juga dapat dilakukan dengan menetapkan tingkat confidence.

```r
# Test korelasi
cor.test(df.salary$absencescore, df.salary$failurescore,
  alternative = c("two.sided", "less", "greater"),
  method = c("pearson", "kendall", "spearman"),
  conf.level = 0.95
)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df.salary$absencescore and df.salary$failurescore
## t = -1.1518, df = 28, p-value = 0.2591
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5321781  0.1598293
## sample estimates:
##        cor
## -0.2126867
```

Simpan Hasil

```
# Menyimpan hasil perhitungan
write_csv(df.correlation, here("data", "processed", "correlation_table.csv"))
```