

# 4-Inspeksi-Data

Ashari Ramadhan

11/20/2020

## Inspeksi Data

Inspeksi adalah pemeriksaan secara detail dan cermat terhadap suatu objek apakah sesuai atau tidak dengan aturan dan standar yang telah ditetapkan. Jadi inspeksi data adalah pemeriksaan data secara detail.

### Import Data

```
prop_pr <- read.csv('data/Proporsi Perempuan DPR.csv', stringsAsFactors = T)
```

### Melihat awalan data

Gunakan fungsi head(nama\_data, jumlah). Jika jumlah di kosongkan, secara default akan bernilai 6.

```
# 6 data awal  
head(prop_pr)
```

```
##   id jumlah_kursi_anggota jumlah_kursi_perempuan presentase provinsi_id  
## 1  1                45                1      2.22%             92  
## 2  2                57                6     10.53%             91  
## 3  3                45                4      8.89%             82  
## 4  4                45               12     26.67%             81  
## 5  5                45                8     17.78%             76  
## 6  6                45               12     26.67%             75  
##   nama_provinsi      Pulau  
## 1   Papua Barat    Papua  
## 2         Papua    Papua  
## 3   Maluku Utara Maluku & NTT  
## 4         Maluku Maluku & NTT  
## 5 Sulawesi Barat  Sulawesi  
## 6    Gorontalo    Sulawesi
```

Output di atas menampilkan 6 data pertama

```
# 10 data awal  
head(prop_pr, 10)
```

```
##      id jumlah_kursi_anggota jumlah_kursi_perempuan presentase provinsi_id
## 1      1              45              1      2.22%          92
## 2      2              57              6     10.53%          91
## 3      3              45              4      8.89%          82
## 4      4              45             12     26.67%          81
## 5      5              45              8     17.78%          76
## 6      6              45             12     26.67%          75
## 7      7              45              8     17.78%          74
## 8      8              85             16     18.82%          73
## 9      9              45              7     15.56%          72
## 10    10             45             14     31.11%          71
##      nama_provinsi      Pulau
## 1      Papua Barat      Papua
## 2      Papua            Papua
## 3      Maluku Utara      Maluku & NTT
## 4      Maluku           Maluku & NTT
## 5      Sulawesi Barat    Sulawesi
## 6      Gorontalo        Sulawesi
## 7      Sulawesi Tenggara Sulawesi
## 8      Sulawesi Selatan  Sulawesi
## 9      Sulawesi Tengah   Sulawesi
## 10     Sulawesi Utara     Sulawesi
```

Output di atas menampilkan 10 data pertama

### Melihat data akhir

Kebalikan dari head(), tail(nama\_df, jumlah) digunakan untuk melihat data dari akhir.

```
# 6 data terakhir
tail(prop_pr)
```

```
##      id jumlah_kursi_anggota jumlah_kursi_perempuan presentase provinsi_id
## 29 29              75              13     17.33%          16
## 30 30              55              7     12.73%          15
## 31 31              65             18     27.69%          14
## 32 32              65              6      9.23%          13
## 33 33             100             13     13.00%          12
## 34 34              81             12     14.81%          11
##      nama_provinsi      Pulau
## 29 Sumatera Selatan Sumatera
## 30      Jambi Sumatera
## 31      Riau Sumatera
## 32 Sumatera Barat Sumatera
## 33 Sumatera Utara Sumatera
## 34      Aceh Sumatera
```

### Menampilkan seluruh data

Untuk melihat seluruh data, gunakan fungsi View(nama\_df). Tapi cara ini kurang direkomendasikan apalagi jika data berukuran besar.

```
View(prop_pr)
```

## Melihat dimensi data

Melihat jumlah baris dan kolom dapat menggunakan fungsi berikut

```
dim(prop_pr)
```

```
## [1] 34 7
```

34 adalah jumlah baris dan 7 adalah jumlah variabel/kolom

## Melihat struktur data

```
str(prop_pr)
```

```
## 'data.frame': 34 obs. of 7 variables:
## $ id : int 1 2 3 4 5 6 7 8 9 10 ...
## $ jumlah_kursi_anggota : int 45 57 45 45 45 45 45 85 45 45 ...
## $ jumlah_kursi_perempuan: int 1 6 4 12 8 12 8 16 7 14 ...
## $ presentase : Factor w/ 26 levels "0","10.53%","10.77%",...: 16 2 24 20 13 20 13 15 10 2
## $ provinsi_id : int 92 91 82 81 76 75 74 73 72 71 ...
## $ nama_provinsi : Factor w/ 34 levels "Aceh","Bali",...: 25 24 21 20 27 7 30 28 29 31 ...
## $ Pulau : Factor w/ 7 levels "Bali","Jawa",...: 5 5 4 4 6 6 6 6 6 6 ...
```

Berdasarkan output di atas kita mendapatkan gambaran data tentang jumlah observasi, variabel, dan tipe data. Dapat dilihat variabel `jumlah_kursi_anggota` bertipe `int` (integer/angka). Apakah ada yang mengganjal dari struktur data diatas?

## Summary data

Melihat statistik deskriptif data

```
summary(prop_pr)
```

```
##      id      jumlah_kursi_anggota jumlah_kursi_perempuan  presentase
## Min.   : 1.00   Min.   : 0.00      Min.   : 0.000      17.78% : 3
## 1st Qu.: 9.25   1st Qu.: 45.00      1st Qu.: 6.000      10.91% : 2
## Median :17.50   Median : 55.00      Median : 8.000      12.73% : 2
## Mean   :17.50   Mean   : 62.18      Mean   : 9.853      13.00% : 2
## 3rd Qu.:25.75   3rd Qu.: 79.50      3rd Qu.:13.000      18.82% : 2
## Max.   :34.00   Max.   :106.00      Max.   :23.000      26.67% : 2
##                                     (Other):21
##      provinsi_id      nama_provinsi      Pulau
## Min.   :11.00   Aceh      : 1   Bali      : 1
## 1st Qu.:19.50   Bali      : 1   Jawa      : 6
## Median :51.50   Banten    : 1   Kalimantan : 5
## Mean   :47.50   Bengkulu  : 1   Maluku & NTT: 4
## 3rd Qu.:71.75   Daerah Isimewa Yogyakarta : 1   Papua      : 2
## Max.   :92.00   Daerah Khusus Ibukota Jakarta: 1   Sulawesi    : 6
##                                     (Other)    :28   Sumatera    :10
```

Nilai yang diperoleh terdiri dari nilai minimum, kuartil, mean dan nilai maksimum untuk tipe variabel `int`. Adapun untuk tipe data `factor` akan dihitung jumlah `factor` pada data. Contoh pada variabel `Pulau`, “Jawa : 6” artinya terdapat 6 data yang merupakan provinsi di pulau jawa.

## Summary data yang lebih lengkap

Jika anda merasa fungsi `summary()` masih kurang menampilkan statistik deskriptif data, untuk melihat statistik deskriptif yang lebih lengkap kita dapat menggunakan fungsi `skim` pada library `skim`.

```
# install library jika belum tersedia
install.packages('skimr') #cukup sekali install
```

```
#load library
library(skimr)
skim(prop_pr)
```





### Data summary

Name	prop_pr
Number of rows	34
Number of columns	7
Column type frequency:	
factor	3
numeric	4
Group variables	
None	

### Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
presentase	0	1	FALSE	26	17.: 3, 10.: 2, 12.: 2, 13.: 2
nama_provinsi	0	1	FALSE	34	Ace: 1, Bal: 1, Ban: 1, Ben: 1
Pulau	0	1	FALSE	7	Sum: 10, Jaw: 6, Sul: 6, Kal: 5

### Variable type: numeric

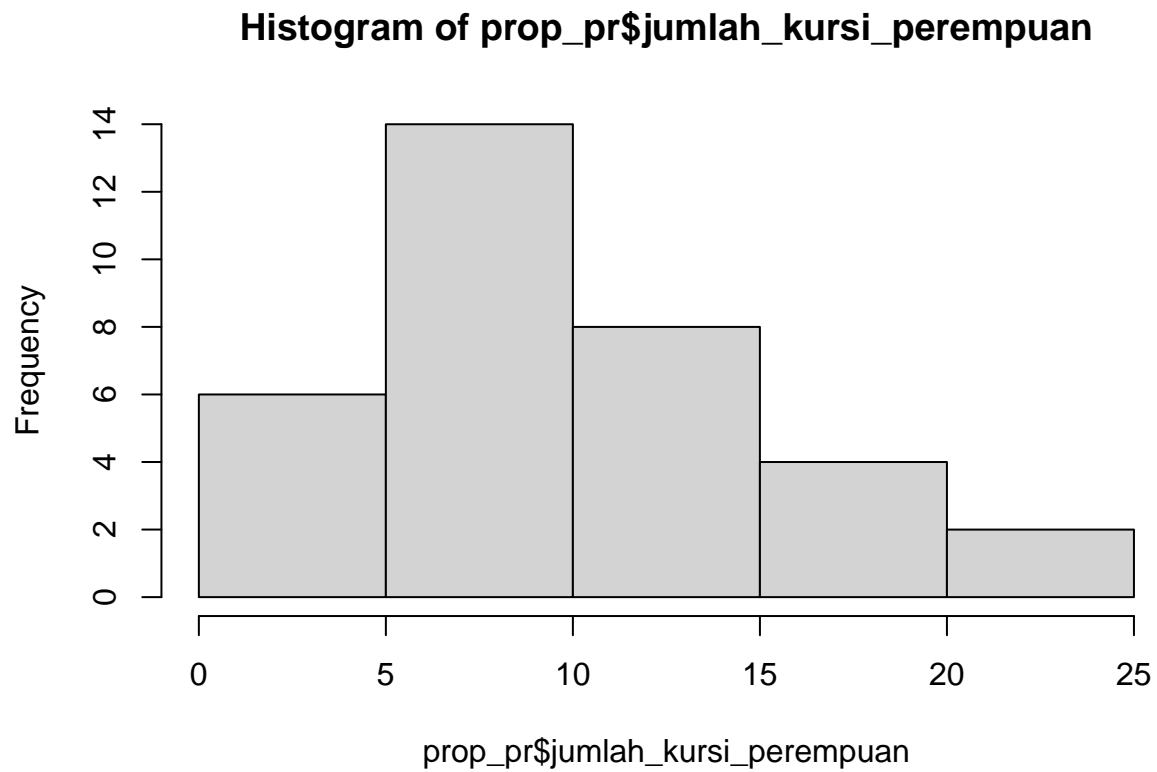
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	17.50	9.96	1	9.25	17.5	25.75	34	
jumlah_kursi_anggota	0	1	62.18	23.00	0	45.00	55.0	79.50	106	
jumlah_kursi_perempuan	0	1	9.85	5.60	0	6.00	8.0	13.00	23	
provinsi_id	0	1	47.50	26.40	11	19.50	51.5	71.75	92	

Dapat dilihat output `skim` lebih lengkap dibanding dibanding `summary()`

## Plot Data

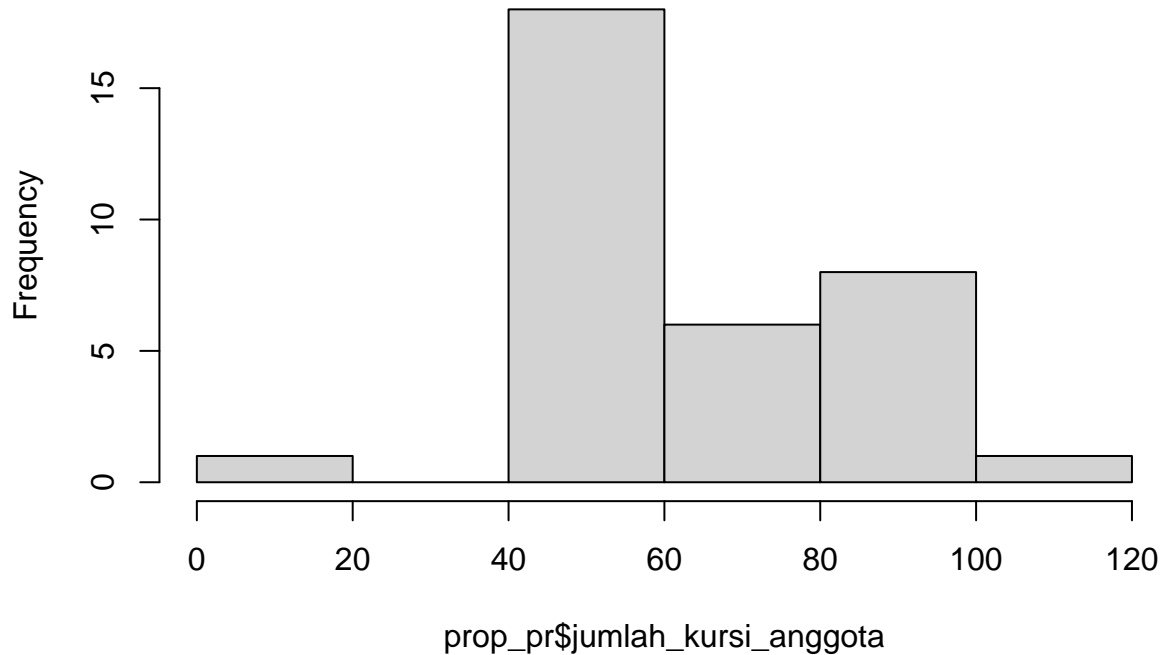
Kita akan membuat histogram jumlah\_kursi\_perempuan di DPR

```
hist(prop_pr$jumlah_kursi_perempuan)
```



```
hist(prop_pr$jumlah_kursi_anggota)
```

## Histogram of prop\_pr\$jumlah\_kursi\_anggota



Pada output di atas diketahui bahwa jumlah kursi perempuan DPR berkisar antara 5-10 kursi, sedangkan kursi total kursi anggota berkisar antara 40-60 kursi.

### Mengapa inspeksi data penting?

Pada kenyataannya tidak semua data siap untuk di olah, kebanyakan data harus di manipulasi bentuknya terlebih dahulu. Dengan menginspeksi data kita mengetahui anomali pada data, seperti kesalahan tipe data, data kosong nama variabel yang tidak sesuai dan sebagainya. Contoh data yang kotor:

```
data_pilpres <- read.csv('data/Hasil Pilpres 2014.csv', stringsAsFactors = T)
str(data_pilpres)
```

```
## 'data.frame':   68 obs. of  7 variables:
## $ id          : int   1 2 3 4 5 6 7 8 9 10 ...
## $ jumlah_suara: Factor w/ 68 levels "1.032.354","1.089.290",...: 2 26 11 7 62 22 41 21 28 36 ...
## $ presentase  : Factor w/ 68 levels "23,08%","26,63%",...: 45 19 68 36 31 39 22 29 8 16 ...
## $ urutan     : int   1 1 1 1 1 1 1 1 1 1 ...
## $ pasangan    : Factor w/ 2 levels "JOKO WIDODO & JUSUF KALLA",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ id_provinsi : int   11 12 13 14 15 16 17 18 19 21 ...
## $ nama        : Factor w/ 34 levels "Aceh","Babel",...: 1 34 32 26 10 33 5 18 2 17 ...
```

Apa yang salah dari data di atas? Variabel jumlah\_suara dan presentase yang merupakan angka dibaca sebagai tipe factor. Hal yang menyebabkan ini adalah tanda titik “.” pada jumlah suara dan tanda koma “,” dan persen “%” pada variabel presentase.

Jika kita paksa membuat hist untuk data jumlah\_suara, sudah pasti akan error.

```
hist(data_pilpres$jumlah_suara)
Error in hist.default(data_pilpres$jumlah_suara) : 'x' must be numeric
```

Mengetahui anomali data, berarti kita langkah apa yang kita harus lakukan, yakni menghapus tanda “.”, “,” dan “%”.

```
library(stringr) #library untuk manipulasi string
# menghapus tanda . (titik) pada jumlah suara
data_pilpres$jumlah_suara <- str_replace_all(data_pilpres$jumlah_suara, "[[:punct:]]", "")

#mengganti tanda , (koma) menjadi tanda . (titik)
data_pilpres$presentase <- str_replace_all(data_pilpres$presentase, ",", ".")

#menghapus tanda % (persen)
data_pilpres$presentase <- str_replace_all(data_pilpres$presentase, "%", "")
str(data_pilpres)
```

### Mengubah struktur dan tipe data

```
## 'data.frame': 68 obs. of 7 variables:
## $ id : int 1 2 3 4 5 6 7 8 9 10 ...
## $ jumlah_suara: chr "1089290" "2831514" "1797505" "1349338" ...
## $ presentase : chr "54.39" "44.76" "76.92" "50.12" ...
## $ urutan : int 1 1 1 1 1 1 1 1 1 1 ...
## $ pasangan : Factor w/ 2 levels "JOKO WIDODO & JUSUF KALLA",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ id_provinsi : int 11 12 13 14 15 16 17 18 19 21 ...
## $ nama : Factor w/ 34 levels "Aceh","Babel",...: 1 34 32 26 10 33 5 18 2 17 ...
```

Pada output di atas tanda titik “.” pada variabel jumlah\_suara sudah hilang. Begitupun pada variabel presentase tanda “,” telah diganti menjadi “.” dan tanda “%” sudah di hapus. Selanjutnya ubah tipe data ke integer/numeric

```
data_pilpres$jumlah_suara <- as.numeric(data_pilpres$jumlah_suara)
data_pilpres$presentase <- as.numeric(data_pilpres$presentase)
data_pilpres$pasangan <- as.factor(data_pilpres$pasangan)
data_pilpres$nama <- as.factor(data_pilpres$nama)
str(data_pilpres)
```

```
## 'data.frame': 68 obs. of 7 variables:
## $ id : int 1 2 3 4 5 6 7 8 9 10 ...
## $ jumlah_suara: num 1089290 2831514 1797505 1349338 871316 ...
## $ presentase : num 54.4 44.8 76.9 50.1 49.2 ...
## $ urutan : int 1 1 1 1 1 1 1 1 1 1 ...
## $ pasangan : Factor w/ 2 levels "JOKO WIDODO & JUSUF KALLA",...: 2 2 2 2 2 2 2 2 2 2 ...
## $ id_provinsi : int 11 12 13 14 15 16 17 18 19 21 ...
## $ nama : Factor w/ 34 levels "Aceh","Babel",...: 1 34 32 26 10 33 5 18 2 17 ...
```

```
head(data_pilpres,7)
```

```
##   id jumlah_suara presentaseurut pasangan
## 1 1      1089290      54.39 1 PRABOWO SUBIANTO & M. HATTA RAJASA
## 2 2      2831514      44.76 1 PRABOWO SUBIANTO & M. HATTA RAJASA
## 3 3      1797505      76.92 1 PRABOWO SUBIANTO & M. HATTA RAJASA
## 4 4      1349338      50.12 1 PRABOWO SUBIANTO & M. HATTA RAJASA
## 5 5       871316      49.25 1 PRABOWO SUBIANTO & M. HATTA RAJASA
## 6 6      2132163      51.26 1 PRABOWO SUBIANTO & M. HATTA RAJASA
## 7 7       433173      45.27 1 PRABOWO SUBIANTO & M. HATTA RAJASA
##   id_provinsi  nama
## 1      11    Aceh
## 2      12   Sumut
## 3      13  Sumbar
## 4      14   Riau
## 5      15   Jambi
## 6      16  Sumsel
## 7      17 Bengkulu
```

Akhirnya data telah bersih dan siap untuk di olah

```
hist(data_pilpres$jumlah_suara, breaks = 100)
```

