

Statistik Deskriptif dengan R

Sebelum memulai dengan konsep dasar dari analisis data, suatu keharusan untuk waspada dari tipe data yang berbeda dan cara untuk mengatur data di file komputer.

1. Beberapa Dasar Ketentuan

- *Populasi* adalah sebuah kumpulan dari subjek (kreaturs, barang, kasus, dan seterusnya)
- *Sampel* adalah kumpulan data dalam studi.
- *Observasi* adalah sebuah studi unit atau subject atau seorang individu.
- *Variabel* adalah kualitas atau kuantitas, terukur atau terekam untuk setiap subjek di dalam sampel (umur, jenis kelamin, tinggi, berat, tingkat merokok, dan lain-lain).
- *Dataset* adalah set nilai semua variabel yang menarik bagi semua individu di dalam pembelajaran.

2. Organisasi Data

Sebuah data biasanya terorganisir (dan disimpan sebagai file komputer) dalam bentuk dari sebuah data matrik;

Sebuah data matrik mewakili jenis kelamin (1-pria; 0-wanita), umur, jumlah anak, berat(kg), dan tinggi (cm) dari 7 individu.

NO	JENIS KELAMIN	UMUR	JUMLAH ANAK	BERAT	TINGGI
1	0	57	1	65	158
2	1	70	3	100	175
3	0	45	0	71	162
4	0	38	2	58	164
5	0	25	1	81	170
6	1	50	4	68	172
7	1	61	0	85	179

Setiap baris dari matriks mewakili satu observasi. Semua baris mempunyai panjang yang sama. Data yang sama sudah direkam untuk semua individu.

Setiap kolom mewakili satu variabel. Sebagai contoh, berat adalah nama dari variabel, mewakili berat badan (dalam kg) dari seorang individu.

3. Tipe-Tipe Data

- Data Numerik
 - Data Diskrit → variabel hanya bisa mengambil nilai-nilai bilangan bulat (0,1,2,dan lain-lain) Contoh: jumlah anak, jumlah teman.
 - Data Kontinu → semua nilai bernomor nyata (seringkali dengan kisaran tertentu) yang memungkinkan. Contoh: badan, berat, umur.
- Kualitatif (non-numerik,kategorikal) Data
 - Data Nominal → Kategori yang tidak ada urutan. Contoh: Kelompok darah, Warna mata.

- Ordinal atau data yang berurutan → kategori yang memiliki urutan atau skala. Contoh: tingkat merokok, sikap (baik, sedang, buruk)

Pengkodean numerik dari data nominal atau yang diurutkan tidak membuat data menjadi numerik!

4. Meringkas atau Memunculkan data

Data kontinu/diskrit

Meringkas lokasi data statistik: mean, median

MEAN

Sampel mean adalah rata-rata aritmatik dari data

Hal ini bisa dihitung dengan menjumlahkan semua nilai data dan membagi dengan jumlah dari total ukuran sampel

Contoh

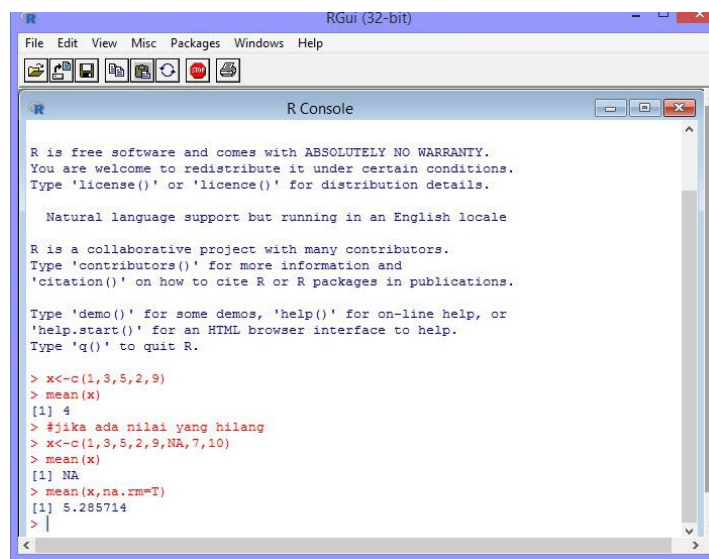
Data: 1 3 5 2 9

Mean: $(1+3+5+2+9)/5=20/5=4$

Secara matematika, untuk variabel X, mean biasanya ditulis sebagai \bar{x} dihitung sebagai

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Dimana x_1, x_2, \dots, x_n menunjukkan pengamatan dari suatu variabel dan n adalah jumlah observasi di sampel.



gambar 1 Menghitung mean dengan aplikasi R

MEDIAN

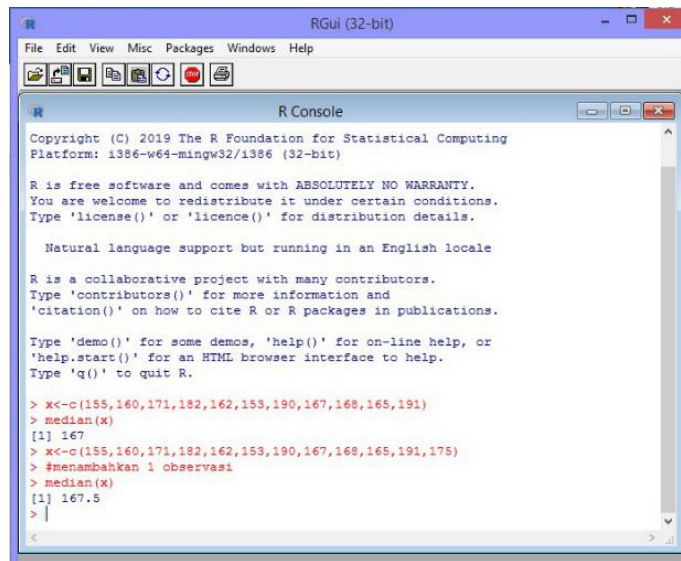
Median adalah titik tengah dari data yang diperintahkan –baik di tengah observasi (apabila data dari observasi ganjil) ataupun rata rata dari 2 nilai tengah dari observasi(jika data observasi genap).

Contoh: 155,160,171,182,162,153,190,167,168,165,191

Data yang diurutkan

154 155 160 162 165 **167** 168 170 171 182 191

Mediannya adalah 167



gambar 2 Mencari median dengan aplikasi R

Keuntungan tapi juga bisa menjadi kerugian dari yaitu tidak terlalu berpengaruh terhadap nilai ekstrim di data. Hal itu tidak berlaku betapa besar atau kecil nya nilai yang lebih besar atau kecil dari median.

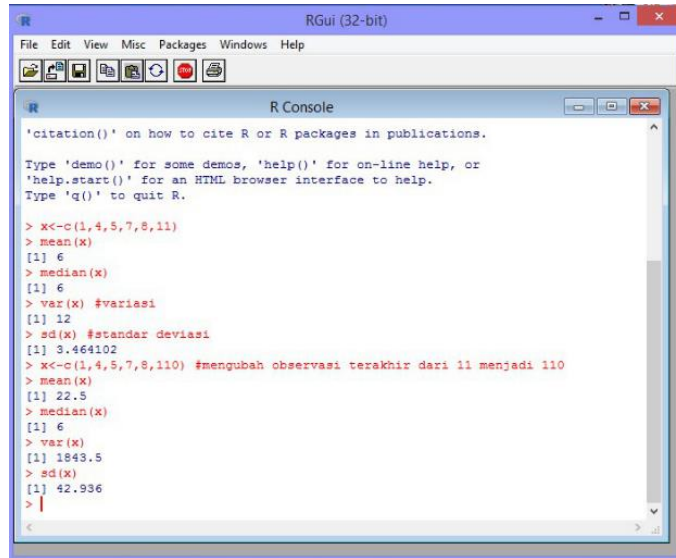
STANDAR DEVIASI

Standar Deviasi adalah jumlah yang menggambarkan varibelitas dari sampel. Kadang bsa dirtikan SD sebagai perkiraan jarak rata rata dari mean. lebih tepatnya,SD didefinisikan sebagai akar kuadrat dari variasi(s^2)(jumlah perbedaan kuadrat dari mean dibagi dengan ukuran sampel minus 1)(yang disebut sebagai sampel variasi, s^2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

seperti mean, standar deviasi sensitif terhadap ekstrim di data.



gambar 3 Mencari SD dengan aplikasi R

Pendekatan yang paling kuat adalah membagi distribusi data (yang diurutkan) menjadi 4 dan mencari titik mana yang 25%,50%dan75% dari ditribusi hal ini dikenal sebagai kuartil.

Contoh

Sebuah sampel

6 9 9 10 9 10 3 12 7 6 6 4 8 8 3 8 6 4 11

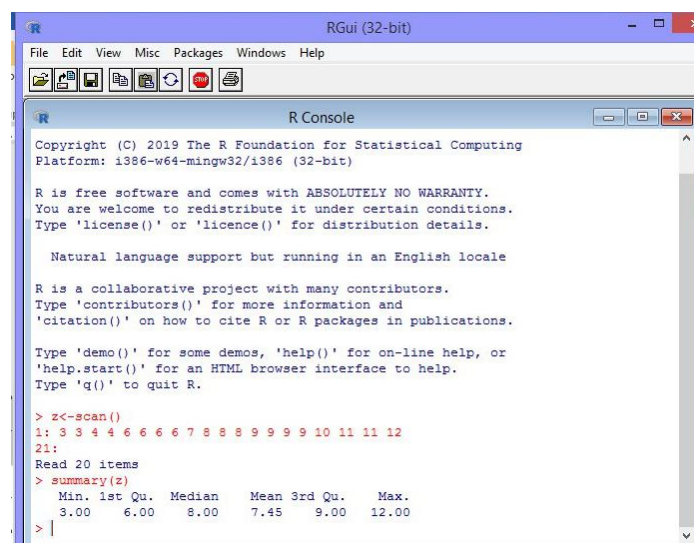
Sampel yang diurutkan

3 3 4 4 6 6 6 6 7 8 8 8 9 9 9 9 10 11 11 12

Sampel yang durutkan dibagi menjadi 4 bagian

3 3 4 4 6 | 6 6 7 8 | 8 8 9 9 9 | 9 10 11 11 12

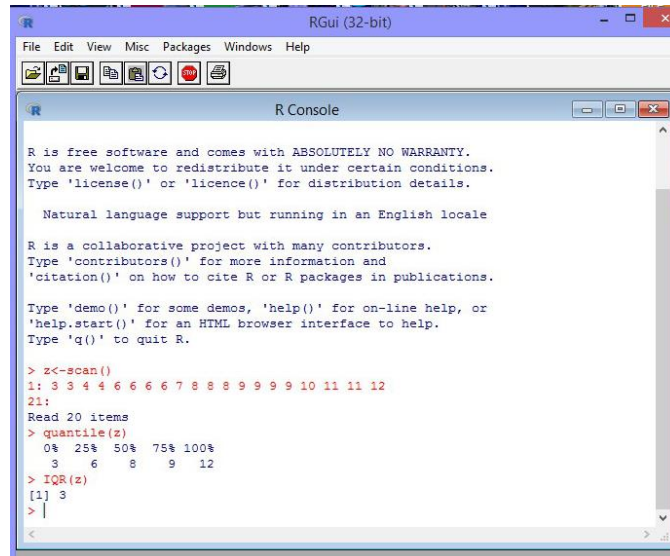
Kuartil : tiitk potong. Kuartil 1 =6, Kuartil 2(median) =8, dan kuartil 3 =9



gambar 4 Summary pada aplikasi R

Di R kamu dapat menggunakan quantile untuk mendapatkan median dan quartile atau kamu bisa menggunakan juga fungsi summary untuk mendapatkan mean.

Variasi dari data bisa dirangkum didalam Innerquartile Range (IQR). Jarak antara kuartil pertama dan kuartil ketiga (here: $IQR=9-6=3$)



```
RGui (32-bit)
File Edit View Misc Packages Windows Help

R Console
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

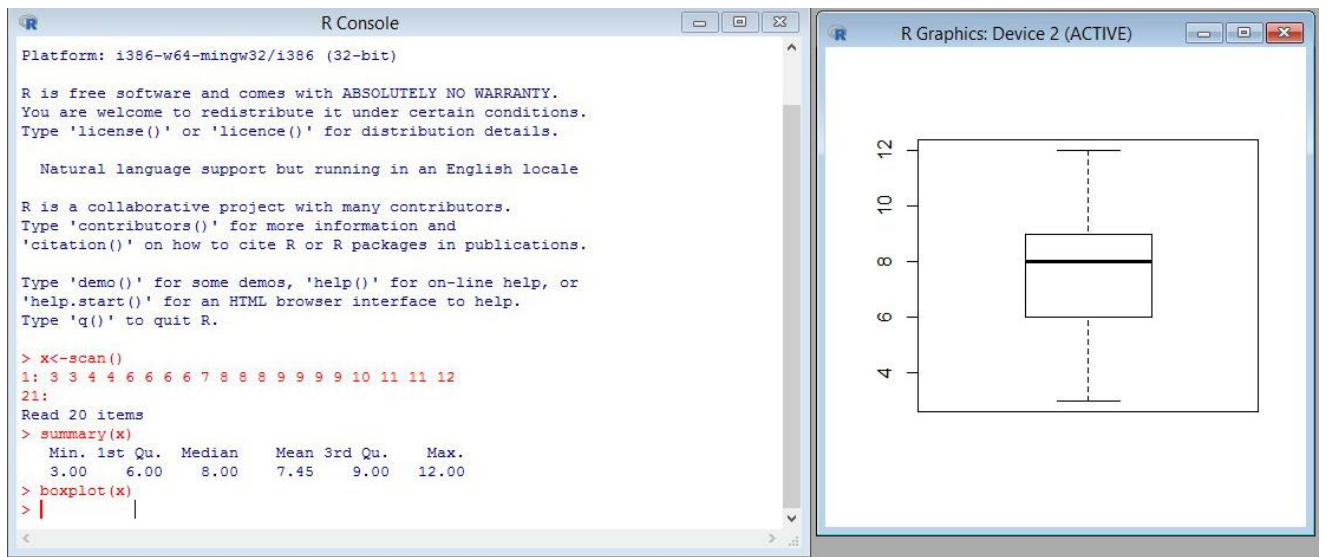
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> z<-scan()
1: 3 3 4 4 6 6 6 6 7 8 8 8 9 9 9 9 10 11 11 12
21:
Read 20 items
> quantile(z)
 0% 25% 50% 75% 100%
 3   6   8   9  12
> IQR(z)
[1] 3
> |
```

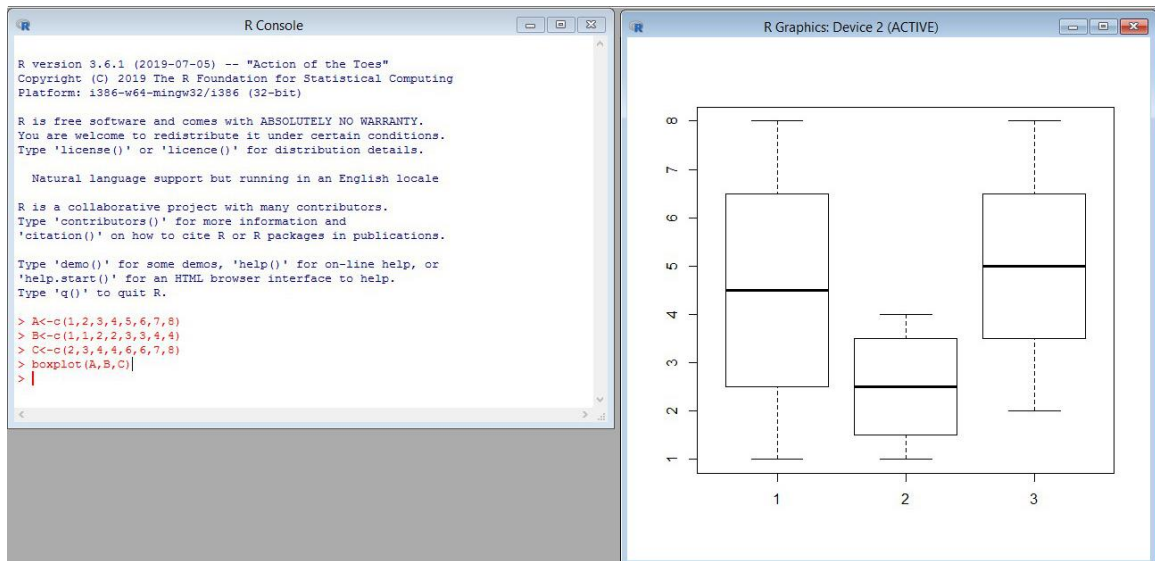
gambar 5 Menggunakan IQR pada aplikasi R

BOXPLOT

Boxplot adalah tampilan grafikal dari median dan kuartil. Boxplot memberikan tampilan distribusi dari data. Ini biasa digunakan untuk membandingkan data antar kelompok yang berbeda



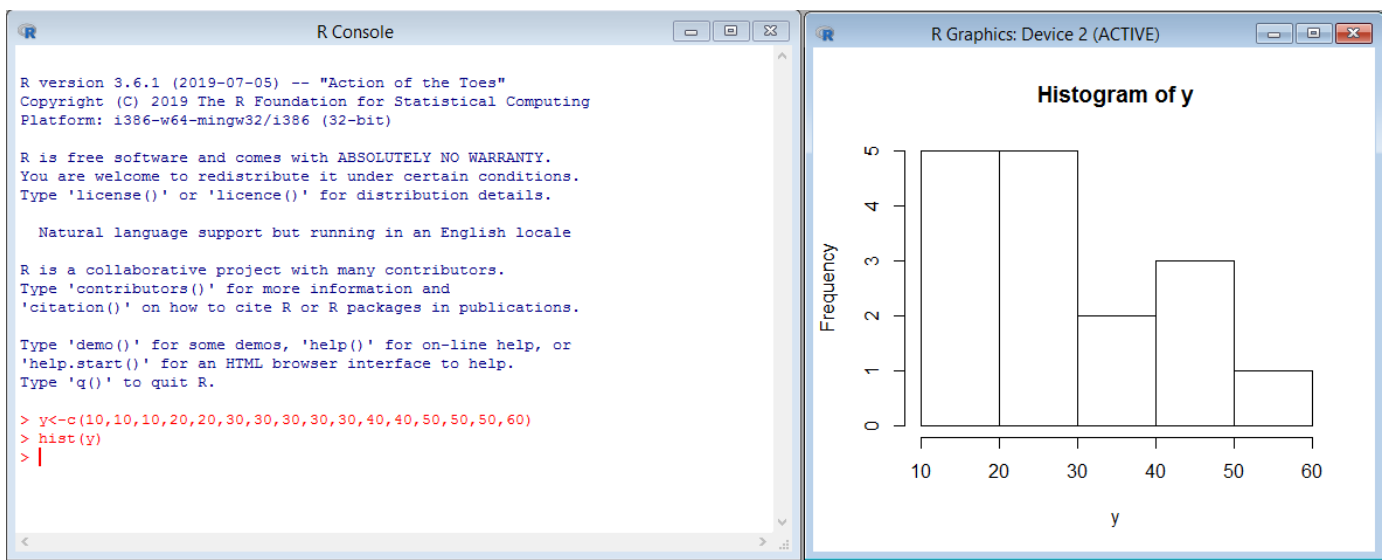
gambar 6 Penggunaan boxplot pada aplikasi R



gambar 7 Boxplot dengan tiga nilai pada aplikasi R

HISTOGRAM

Cara lain untuk melihat distribusi adalah membuat histogram data. Untuk mendapatkan histogram skala dari variabel dibagi menjadi interval berturut turut dengan panjang dan jumlah yang sama dari setiap interval yang dihitung.



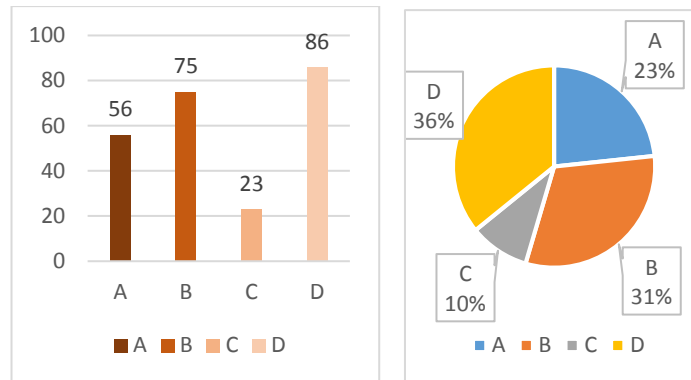
gambar 8 Histogram pada aplikasi R

Untuk nominal data, mean dan standar deviasi tidak terlalu diperukan, begitu juga median dan presentil. Untuk melihat distribusi data. Lihat ke tabel frekuensi

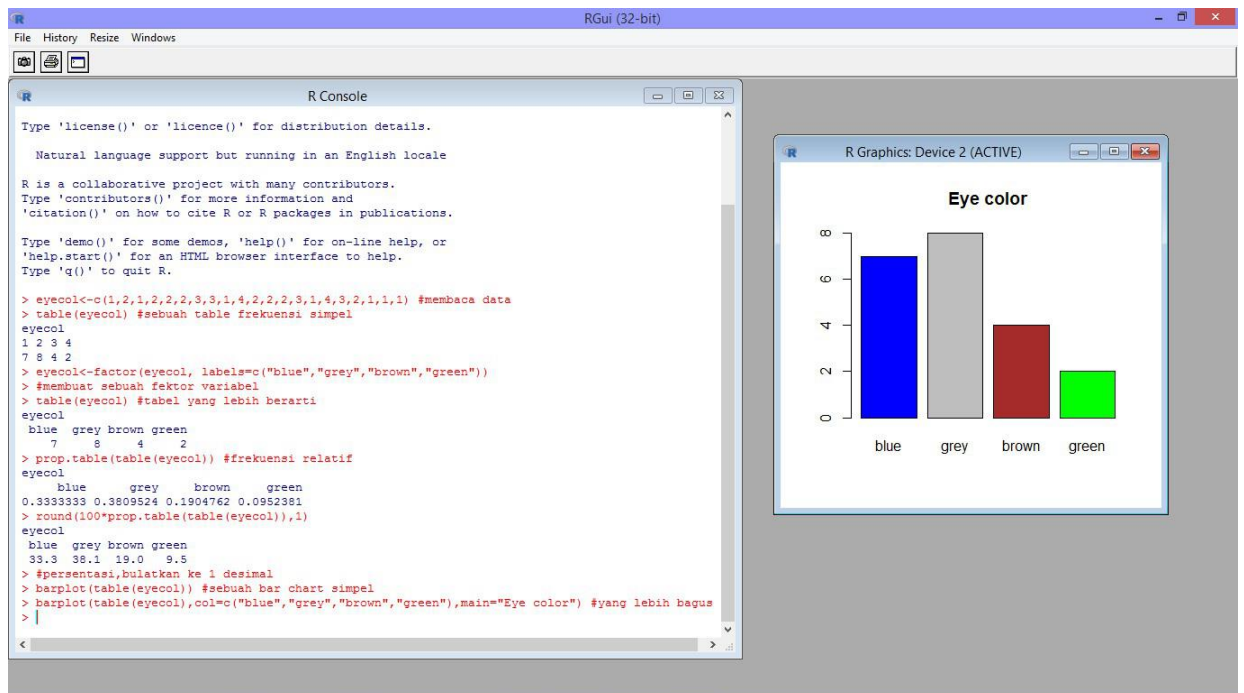
Contoh: golongan darah dari 240 individu

Blood Group	N	%
A	56	23.3%
B	75	31.2%
AB	23	9.6%
O	86	35.8%

Dan ini merupakan gambaran grafik- dari bar chart dan pie chart:



Tips: pie chart yang tidak sesuai harus dihindari



gambar 9 Contoh tabel berwarna

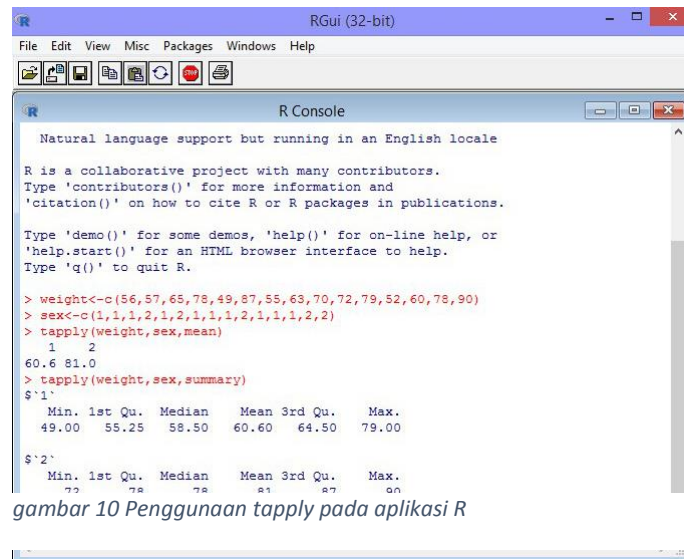
Beberapa catatan untuk statistik grafik

- Sebuah grafik harus memuat informasi yang berguna dan lebih efisien daripada rangkuman numerik yang lain
- Sebuah grafik yang bagus harus memiliki informasi yang baik antara rasio. Hindari detail yang mewah yang tidak mengandung informasi namun hanya membuat grafik lebih komplikasi (besar dan lebih berwarna)

- Berikan perhatian kepada skala dari grafik

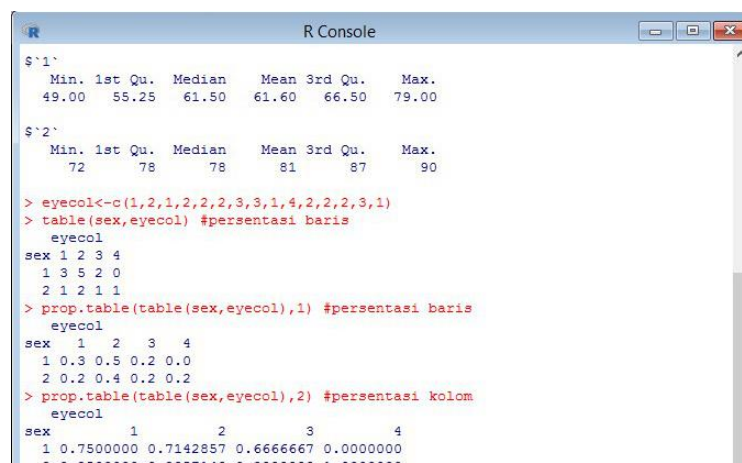
5. R: Deskriptif Statistik dari 2 kelompok, 2-dimensi tabel

Untuk membandingkan mean atau deskriptis statistik yang lain dalam subgrup yang berbeda dari sample. Di r kamu bisa menggunakan fungsi `tapply` untuk itu. Fungsi ini mengambil 3 argumen yaitu numerik, variabel, kelompok kategorikal variabel dan fungsi untuk mengaplikasikan.



gambar 10 Penggunaan tapply pada aplikasi R

Atau ketika anda mempunyai 2 kategorikal variabel, anda mungkin tertarik dengan kontigensi tabel 2-dimensi



gambar 11 kontigensi tabel 2-dimensi pada aplikasi R