R Course DQ Lab Academy

Modul 4

Statistic using R for Data Science

Pendahuluan

Statistika adalah ilmu yang mempelajari cara pengumpulan data, menganalisis data untuk mendapatkan kesimpulan informasi sampai dapat dijadikan dasar pembuatan kebijakan. Definisi diatas mirip dengan tugas dari seorang Data Science yaitu mulai dari ekplorasi data, modelling untuk mendapatakan pola yang tersembunyi dari data kemudian menemukan *Insight* untuk dasar kebijakan (*data-driven*).

Kenapa harus belajar statistik ?

Ilmu Statistik fungsinya untuk mengolah data, yang bisa angka maupun bukan angka. Statistik merupakan pondasi awal sebelum belajar Data Science. Alasannya, banyak tools data science merupakan pengembangan dari teknik statistik, mulai dari sederhana sampai yang rumit.

Agar dapat memahami konsep-konsep tersebut, pada bab ini juga disertakan satu dataset file dengan nama data_intro.csv yang akan dijadikan file praktek di R.

Klik tombol Next untuk melanjutkan.

Apa itu statistik?

Apa itu statistik?

- [x] Ilmu yang mempelajari cara pengumpulan data
- [ ] Ilmu yang belajar mengenai sistem database dan security
- [ ] Ilmu yang mempelajari cascading style sheet (css) data
- [x] Ilmu yang mempelajari cara menganalisa data
- [x] Ilmu yang mempelajari bagaimana mendapatkan kesimpulan informasi

Mengapa Statistik itu Penting?

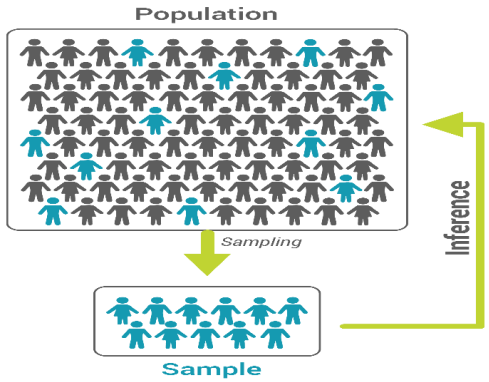
Mengapa Statistik itu Penting?

- [x] Sebagai dasar pengambilan keputusan
- [ ] Sebagai bahan untuk merancang sistem pemograman
- [ ] Sebagai alat perhitungan
- [ ] Sebagai fondasi awal belajar data engineer
- [x] Sebagai materi penunjang Data Science

Statistik sebagai Ilmu Peluang

Sebenarnya statistik merupakan ilmu peluang, yaitu untuk mendapatkan generalisasi populasi dari sampel yang kita miliki. Dalam statistik banyak kaitannya dengan sampel dan populasi, berikut pengertiannya

1. **Sampel** adalah representasi dari sebagian elemen dari populasi
2. **Populasi** adalah total dari semua elemen



Gambaran diatas dapat menjelaskan fungsi dari statistik, yaitu kita dapat mengetahui karakteristik dari populasi melalui sampel yang kita miliki.

Kemudian untuk mengukur karakteristik dari sampel dan populasi, dengan melihat nilai statistik dan parameter. Untuk lebih jelasnya dapat dilihat pengertian berikut

Perbedaan antara statistik dan parameter adalah sebagai berikut:

- **Statistik** adalah nilai estimasi karateristik populasi.
- **Parameter** adalah nilai karakteristik populasi atau bisa disebut karakteristik sebenarnya.

Statistik lebih banyak di cari nilainya daripada parameter, alasannya adalah lebih menghemat biaya,waktu dan tenaga. Selain itu, pengambilan sampel sebenarnya sudah dapat mewakili populasi.

Dan secara praktis, kita tidak mungkin melakukan pengambilan populasi karena dapat bersifat merusak. Contoh: pengambilan sample dari produksi seluruh bola lampu untuk menguji kandungan di dalamnya. Atau pengambilan seluruh populasi udang dari suatu tambak.

Nilai estimasi didapatkan dari data kuantitatif dan kualitatif, untuk mengetahui perbedaanya akan dijelaskan pada subbab selanjutnya.

Klik tombol Next untuk melanjutkan.

Data Kuantitatif dan Kualitatif

Kualitatif dan Kuantitatif

1. **Kuantitatif** adalah data yang dinyatakan dalam bentuk angka
2. **Kualitatif** adalah data yang dinyatakan dalam bentuk bukan angka

Selanjutnya bagaimana untuk mendapatkan nilai karakteristik dari data kuantitatif dan kualitatif, jawabanya yaitu kita harus menggolongkan kedalam skala pengukuran data.

Kenapa harus dilakukan?

Agar data mudah untuk diolah sehingga mendapatkan nilai statistik

Klik tombol Next untuk melanjutkan.

Skala Pengukuran Data

Tiap data perlu suatu standar untuk melakukan pengukuran, ini disebut skala.

Dan berikut adalah jenis-jenis skala pengukuran data:

- **Nominal**: adalah skala yang diberikan hanya sebagai label saja, tidak mengandung pengertian tingkatan.
  Contoh: Jika pria =1 dan wanita = 2, artinya disini 1 dan 2 adalah nominal yang mewakili pria dan wanita. Disini nilai 2 tidak lebih besar dari nilai 1.

- **Ordinal**: adalah skala yang mengandung pengertian tingkatan.
  Contoh: Data kepuasan, 1 = tidak puas, 2 = puas, dan 3 = sangat puas, artinya 1<2<3.

- **Interval**: adalah skala yang mempunyai sifat ordinal dan mengandung jarak(interval).
  Misalnya: harga pakaian merk A 100 ribu, harga pakaian merk C 200 ribu, artinya harga pakian merk A dan C memiliki interval 100 ribu

- **Rasio**: adalah skala yang mempunyai sifat nominal, ordinal, dan interval, serta mempunyai nilai rasio antar objek yang diukur.

Contoh: harga pakaian merk A 100 ribu, harga pakaian merk C 200 ribu. Rasio harga pakaian A dengan pakaian C adalah ½. Sehingga dapat dikatakan bahwa harga pakaian A harganya 2 kali pakaian C.

Dari penjelasan diatas dapat kita simpulkan bahwa data kualitatif dapat kita golongkan menjadi skala nominal dan ordinal. Sedangkan untuk data kuantitatif maka digolongkan menjadi Interval dan Rasio.Skala pengukuran nominal dan ordinal pada R di definisikan sebagai factor atau sering disebut data kategorik sedangkan interval dan rasio di definisikan sebagai numerik.

Klik tombol Next untuk melanjutkan.

Mengapa harus mengambil sampel daripada populasi ?

Mengapa harus mengambil sampel daripada populasi ?

- [ ] Sampel mencakup semua data
- [x] Sampel sudah dapat mewakili populasi
- [x] Pengambilan populasi dapat bersifat merusak
- [x] Lebih menghemat biaya,waktu dan tenaga
- [ ] Sampel dapat memperkecil error

Pernyataan dibawah ini yang benar adalah...

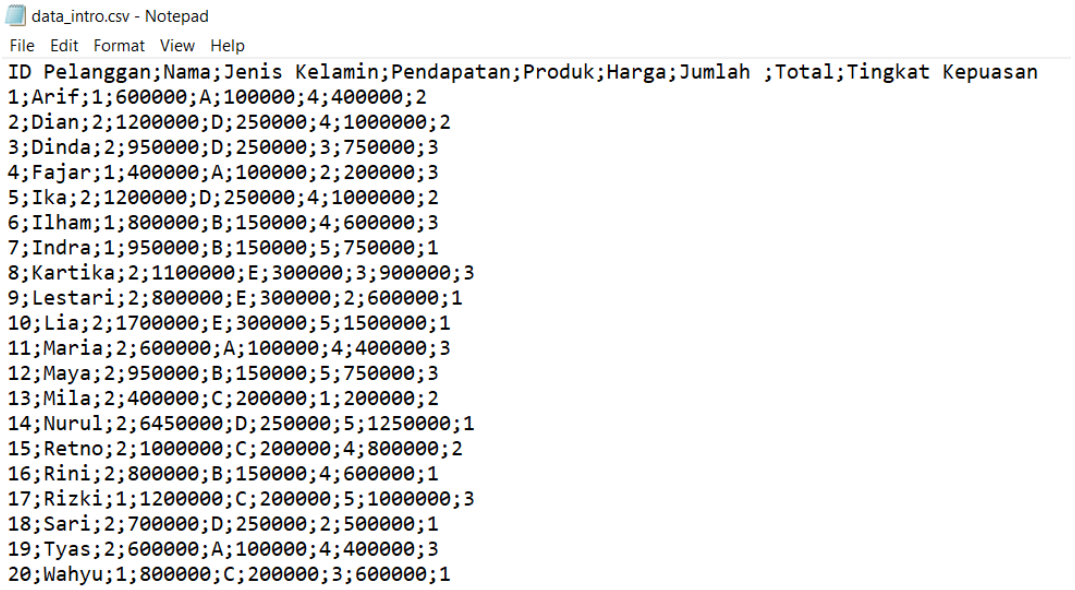
Pernyataan dibawah ini yang benar adalah ?

- [x] Data kuantitatif adalah data yang dapat di lakukan operasi matematika
- [ ] Data Ordinal adalah data yang hanya membedakan antar kategori
- [x] Data Kategorik adalah data Kualitatif
- [x] Semua data harus dalam bentuk angka untuk dilakukan analisis statistik
- [ ] Data Kategorik tidak dapat di analisis statistic

Dataset Tingkat Kepuasan Pelanggan

Dataset yang akan di pakai dalam course ini adalah data tentang kepuasan konsumen terhadap suatu produk pakaian. Dataset ini ada dalam file bentuk format file CSV dengan nama data_intro.csv. Data ini juga dilengkapi karakteristik umum dari konsumen.

Berikut adalah tampilan dari dataset tersebut jika dibuka dengan aplikasi notepad.



Terlihat pemisah antar kolomnya menggunakan tanda titik koma. Terdiri dari sembilan kolom dan 20 baris data

Dan berikut adalah tampilan dari dataset tersebut jika dibuka dengan aplikasi spreadsheet.

	A	B	C	D	E	F	G	H	I	J
1	ID Pelanggan	Nama	Jenis Kelamin	Pendapatan	Produk	Harga	Jumlah	Total	Tingkat Kepuasan	
2	1	Arif	1	600000	A	100000	4	400000	2	
3	2	Dian	2	1200000	D	250000	4	1000000	2	
4	3	Dinda	2	950000	D	250000	3	750000	3	
5	4	Fajar	1	400000	A	100000	2	200000	3	
6	5	Ika	2	1200000	D	250000	4	1000000	2	
7	6	Ilham	1	800000	B	150000	4	600000	3	
8	7	Indra	1	950000	B	150000	5	750000	1	
9	8	Kartika	2	1100000	E	300000	3	900000	3	
10	9	Lestari	2	800000	E	300000	2	600000	1	
11	10	Lia	2	1700000	E	300000	5	1500000	1	
12	11	Maria	2	600000	A	100000	4	400000	3	
13	12	Maya	2	950000	B	150000	5	750000	3	
14	13	Mila	2	400000	C	200000	1	200000	2	
15	14	Nurul	2	6450000	D	250000	5	1250000	1	
16	15	Retno	2	1000000	C	200000	4	800000	2	
17	16	Rini	2	800000	B	150000	4	600000	1	
18	17	Rizki	1	1200000	C	200000	5	1000000	3	
19	18	Sari	2	700000	D	250000	2	500000	1	
20	19	Tyas	2	600000	A	100000	4	400000	3	
21	20	Wahyu	1	800000	C	200000	3	600000	1	

Dataset tersebut terdiri dari sembilan kolom dengan detil berikut:

- **ID Pelanggan**: Kode pelanggan yang sifatnya unik, tidak ada data lain dengan kode yang sama. Kode ini dalam bentuk yang sangat sederhana berupa angka integer (bilangan bulat).
- **Nama**: Nama pelanggan dalam bentuk teks
- **Jenis Kelamin**: Jenis kelamin dari pelanggan, dalam bentuk angka integer. Disini 1 mewakili laki-laki dan 2 mewakili perempuan.
- **Pendapatan**: Nilai pendapatan per bulan dari tiap pelanggan (??).
- **Produk**: Produk yang disurvei.
- **Harga**: Harga produk yang dibeli.
- **Jumlah**: Jumlah produk yang dibeli.
- **Total**: Total harga pembelian.
- **Tingkat Kepuasan**: Indeks tingkat kepuasan pelanggan tersebut terhadap produk yang dibeli.

Dengan data sederhana ini diharapkan dapat mengasah kempuan analisis statistik. Kemampuan analisis statistik akan terlatih dengan *Learning By Doing.* Metode belajar ini sangat efektif untuk pemahaman ilmu statistika.

Membaca Dataset dengan read.csv

Untuk membaca dataset data_intro.csv tersebut kita akan gunakan function read.csv dengan konstruksi berikut:

```
data_intro <-
read.csv("https://academy.dqlab.id/dataset/data_intro.csv",sep="
;")
```

Penjelasan terhadap function di atas adalah sebagai berikut:

Komponen	Deskripsi
data_intro	nama variable yang digunakan untuk menampung pembacaan file dataset data_intro.csv
read.csv	function yang digunakan untuk membaca contoh dataset dengan format file teks (CSV)
https://academy.dqlab.id/dataset/data_intro.csv	lokasi dataset yang terdapat di web DQLab. Jika lokasi file dan aplikasi R terdapat di komputer lokal Anda, maka gantilah dengan lokasi file di lokal. Misalkan c:\data\data_intro.csv
sep=";"	Parameter pemisah (separator) antar kolom data. Kita gunakan tanda titik koma untuk dataset tingkat kepuasan pelanggan.

Tugas Praktek

Lengkapi bagian […1…] pada code editor untuk membaca file seperti yang ditunjukkan pada bagian Lesson.

Jika berjalan dengan lancar maka akan tampil sebagian dataset pada Console sebagai berikut.

	ID.Pelanggan	Nama	Jenis.Kelamin	Pendapatan	Produk	Harga	Jumlah	Total
1	1	Arif	1	600000	A	100000	4	400000
2	2	Dian	2	1200000	D	250000	4	1000000
3	3	Dinda	2	950000	D	250000	3	750000
4	4	Fajar	1	400000	A	100000	2	200000
5	5	Ika	2	1200000	D	250000	4	1000000
6	6	Ilham	1	800000	B	150000	4	600000
7	7	Indra	1	950000	B	150000	5	750000
8	8	Kartika	2	1100000	E	300000	3	900000
9	9	Lestari	2	800000	E	300000	2	600000
10	10	Lia	2	1700000	E	300000	5	1500000
11	11	Maria	2	600000	A	100000	4	400000
12	12	Maya	2	950000	B	150000	5	750000
13	13	Mila	2	400000	C	200000	1	200000
14	14	Nurul	2	6450000	D	250000	5	1250000
15	15	Retno	2	1000000	C	200000	4	800000
16	16	Rini	2	800000	B	150000	4	600000
17	17	Rizki	1	1200000	C	200000	5	1000000
18	18	Sari	2	700000	D	250000	2	500000
19	19	Tyas	2	600000	A	100000	4	400000
20	20	Wahyu	1	800000	C	200000	3	600000

	Tingkat.Kepuasan
1	2
2	2
3	3
4	3
5	2
6	3
7	1
8	3
9	1
10	1
11	3

```
12                3
13                2
14                1
15                2
16                1
17                3
18                1
19                3
20                1
```

Code Editor

#Membaca dataset dengan read.csv dan dimasukkan ke variable data_intro

data_intro <- read.csv("https://academy.dqlab.id/dataset/data_intro.csv", sep=";") #[...1...]

data_intro

Console

```
> #Membaca dataset dengan read.csv dan dimasukkan ke variable data_intro
> data_intro <- read.csv("https://academy.dqlab.id/dataset/data_intro.csv", sep=";")
#[...1...]

> data_intro
   ID.Pelanggan    Nama Jenis.Kelamin Pendapatan Produk  Harga Jumlah   Total
1             1    Arif             1     600000      A 100000      4  400000
2             2    Dian             2    1200000      D 250000      4 1000000
3             3   Dinda             2     950000      D 250000      3  750000
4             4   Fajar             1     400000      A 100000      2  200000
5             5     Ika             2    1200000      D 250000      4 1000000
6             6   Ilham             1     800000      B 150000      4  600000
7             7   Indra             1     950000      B 150000      5  750000
8             8 Kartika             2    1100000      E 300000      3  900000
9             9 Lestari             2     800000      E 300000      2  600000
10           10     Lia             2    1700000      E 300000      5 1500000
11           11   Maria             2     600000      A 100000      4  400000
12           12    Maya             2     950000      B 150000      5  750000
13           13    Mila             2     400000      C 200000      1  200000
14           14   Nurul             2    6450000      D 250000      5 1250000
15           15   Retno             2    1000000      C 200000      4  800000
16           16    Rini             2     800000      B 150000      4  600000
17           17   Rizki             1    1200000      C 200000      5 1000000
18           18    Sari             2     700000      D 250000      2  500000
19           19    Tyas             2     600000      A 100000      4  400000
```

```
20           20   Wahyu             1     800000      C 200000      3  600000
   Tingkat.Kepuasan
1                  2
2                  2
3                  3
4                  3
5                  2
6                  3
7                  1
8                  3
9                  1
10                 1
11                 3
12                 3
13                 2
14                 1
15                 2
16                 1
17                 3
18                 1
19                 3
20                 1
```

Melihat Tipe Data dengan Str

Adalah praktek yang sangat baik untuk mengenal atau melakukan *profile* tiap dataset yang sudah dibaca ke dalam R – dan secara sederhana di R dapat kita lakukan dengan function **str**. Function str akan menyajikan informasi tiap kolom dataset dalam format yang *compact* – satu baris informasi saja per row. Pendekatan singkat dan jelas ini membuat str menjadi function favorit dan efektif untuk mengenal data di tahap awal.

Syntaxnya juga cukup sederhana, cukup masukkan dataset ke dalam function ini seperti pada contoh berikut.

str(data_intro)

Tugas Praktek

Gantilah bagian […1…] pada code editor dengan perintah str yang menggunakan input variable data_intro.

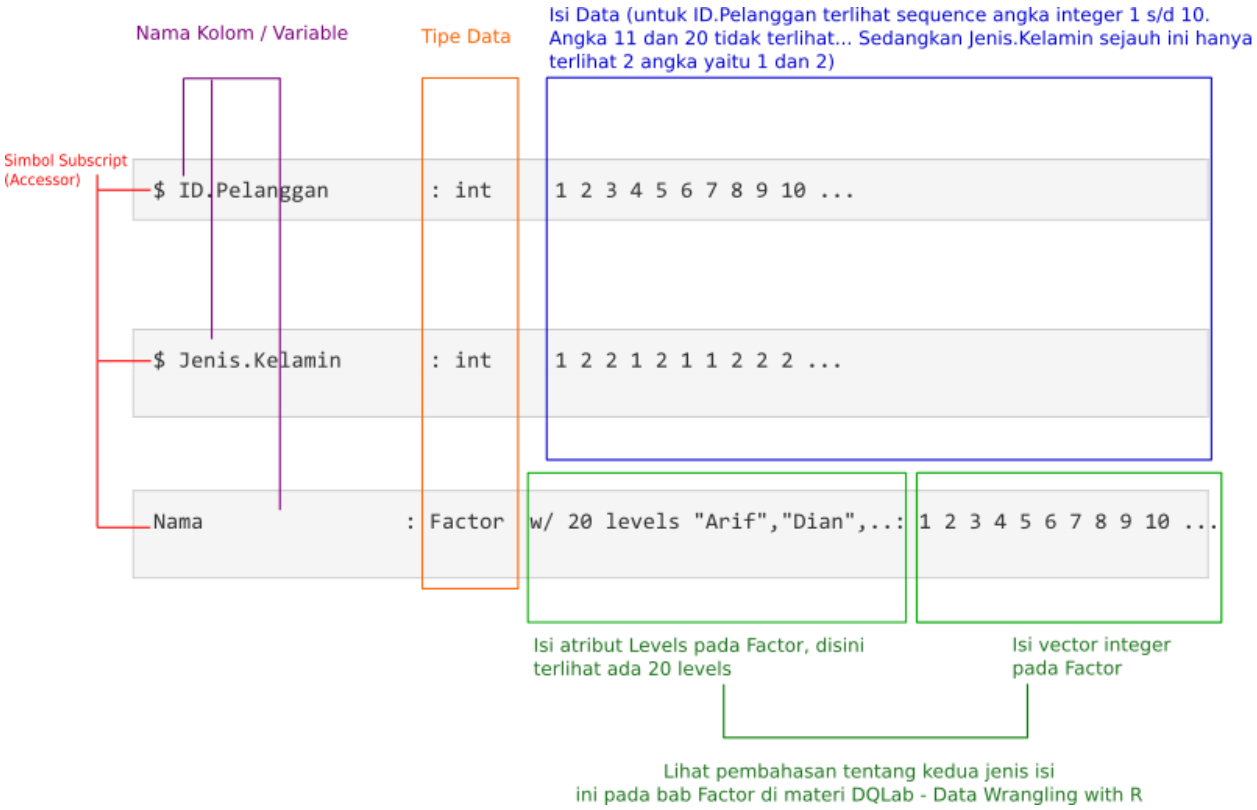
Jika berjalan dengan lancar, maka outputnya sebagian akan terlihat sebagai berikut.

```
> str(data_intro)
'data.frame':    20 obs. of  9 variables:
 $ ID.Pelanggan    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Nama            : Factor w/ 20 levels "Arif","Dian",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Jenis.Kelamin   : int  1 2 2 1 2 1 1 2 2 2 ...
 $ Pendapatan      : int  600000 1200000 950000 400000 1200000 800000 950000 1100000
800000 1700000 ...
 $ Produk          : Factor w/ 5 levels "A","B","C","D",..: 1 4 4 1 4 2 2 5 5 5 ...
 $ Harga           : int  100000 250000 250000 100000 250000 150000 150000 300000 300
000 300000 ...
 $ Jumlah          : int  4 4 3 2 4 4 5 3 2 5 ...
 $ Total           : int  400000 1000000 750000 200000 1000000 600000 750000 900000 6
00000 1500000 ...
 $ Tingkat.Kepuasan: int  2 2 3 3 2 3 1 3 1 1 ...
```

Untuk baris di bawahnya adalah penjelasan dari tiap kolom/variable data yang terdiri dari:

- Nama kolom
- Tipe data kolom
- Isi dari kolom tersebut
- Jika Factor maka ada tambahan indexnya

Berikut penjelasan hasil dalam bentuk ilustrasi dari 3 kolom, yaitu ID.Pelanggan, Nama, dan Jenis.Kelamin.



Code Editor

```
#Membaca dataset dengan read.csv dan dimasukkan ke variable data_intro

data_intro <- read.csv("https://academy.dqlab.id/dataset/data_intro.csv",sep=";")

str(data_intro) #[...1...]
```

Console

```
> #Membaca dataset dengan read.csv dan dimasukkan ke variable data_intro
> data_intro <- read.csv("https://academy.dqlab.id/dataset/data_intro.csv",sep=";")

> str(data_intro) #[...1...]
'data.frame':    20 obs. of  9 variables:
 $ ID.Pelanggan    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Nama            : Factor w/ 20 levels "Arif","Dian",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Jenis.Kelamin   : int  1 2 2 1 2 1 1 2 2 2 ...
 $ Pendapatan      : int  600000 1200000 950000 400000 1200000 800000 950000 1100000
800000 1700000 ...
```

```
 $ Produk          : Factor w/ 5 levels "A","B","C","D",..: 1 4 4 1 4 2 2 5 5 5 ...
 $ Harga           : int  100000 250000 250000 100000 250000 150000 150000 300000 300
000 300000 ...
 $ Jumlah          : int  4 4 3 2 4 4 5 3 2 5 ...
 $ Total           : int  400000 1000000 750000 200000 1000000 600000 750000 900000 6
00000 1500000 ...
 $ Tingkat.Kepuasan: int  2 2 3 3 2 3 1 3 1 1 ...
```

Merubah Tipe Data Kolom ID.Pelanggan menjadi Character

Variabel ID.Pelanggan merupakan kode unik dari setiap variabel dan tidak bisa dicari nilai statistiknya. Sehingga tipe data ID.Pelanggan perlu diubah menjadi character agar tidak ikut di analisis.

Untuk mengubah tipe data ID.Pelanggan menjadi character dapat menggunakan syntax

```
data_intro$ID.Pelanggan <-as.character(data_intro$ID.Pelanggan)
```

Function **as.character** mengubah id tiap pelanggan menjadi string/character - ditandai dengan tanda petik diantara kode unik tersebut.

Tugas Praktek

Gantilah bagian [...1...] pada code editor dengan perintah as.character yang menggunakan input variable data_intro dengan kolom ID.Pelanggan dan Nama. Kemudian pada bagian [...2...] keluarkan output untuk memastikan bahwa output tersebut berupa String.

Jika berjalan lancar maka akan tampil output sebagai berikut

```
> str(data_intro$ID.Pelanggan)
 chr [1:20] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "13" "14" "15" ...
> str(data_intro$Nama)
 chr [1:20] "Arif" "Dian" "Dinda" "Fajar" "Ika" "Ilham" "Indra" "Kartika" "Lestari" "
Lia" ...
```

Code Editor

```
## mengubah data menjadi karakter karena tidak dilakukan analisis statistik pada variabel ID Pelanggan dan nama

data_intro$ID.Pelanggan <- as.character(data_intro$ID.Pelanggan)

data_intro$Nama <- as.character(data_intro$Nama) #[...1...]

## melihat apakah sudah berhasil dalam mengubah variabel tersebut

str(data_intro$ID.Pelanggan) #[...2...]

str(data_intro$Nama)
```

Console

```
> ## mengubah data menjadi karakter karena tidak dilakukan analisis statistik pada va
riabel ID Pelanggan dan nama
> data_intro$ID.Pelanggan <- as.character(data_intro$ID.Pelanggan)

> data_intro$Nama <- as.character(data_intro$Nama) #[...1...]

> ## melihat apakah sudah berhasil dalam mengubah variabel tersebut
> str(data_intro$ID.Pelanggan) #[...2...]
 chr [1:20] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" ...

> str(data_intro$Nama)
 chr [1:20] "Arif" "Dian" "Dinda" "Fajar" "Ika" "Ilham" ...
```

Merubah Sejumlah Kolom menjadi Data Kategorik (Factor)

Pada data_intro beberapa variabelnya bersifat kualitatif yaitu variabel jenis kelamin, produk, dan Tingkat_Kepuasan. Variabel tersebut harus di ubah jenis datanya menjadi faktor untuk mendapatkan karakteristik dari setiap pelanggan (observasi).

Untuk mengubah tipe data menjadi factor dapat menggunakan syntax berikut:

```
data_intro$Jenis.Kelamin <- as.factor(data_intro$Jenis.Kelamin)

data_intro$Produk <- as.factor(data_intro$Produk)

data_intro$Tingkat.Kepuasan <-
as.factor(data_intro$Tingkat.Kepuasan)
```

Tugas Praktek

Gantilah bagian [...1...] pada code editor untuk merubah kolom Jenis.Kelamin, Produk dan Tingkat.Kepuasan menjadi tipe data faktor (Factor). Kemudian gantilah bagian [...2...] pada code editor untuk menampilkan struktur dari kolom Jenis.Kelamin, Produk dan Tingkat.Kepuasan dengan function str.

Jika berjalan dengan lancar maka akan tampil sebagian dataset pada Console sebagai berikut.

```
> str(data_intro$Jenis.Kelamin)
 Factor w/ 2 levels "1","2": 1 2 2 1 2 1 1 2 2 2 ...
> str(data_intro$Produk)
 Factor w/ 5 levels "A","B","C","D",..: 1 4 4 1 4 2 2 5 5 5 ...
> str(data_intro$Tingkat.Kepuasan)
 Factor w/ 3 levels "1","2","3": 2 2 3 3 2 3 1 3 1 1 ...
```

Code Editor

menggunakan functon as.factor

data_intro$Jenis.Kelamin <- as.factor(data_intro$Jenis.Kelamin) #[...1...]

data_intro$Produk <- as.factor(data_intro$Produk) #[...1...]

data_intro$Tingkat.Kepuasan <- as.factor(data_intro$Tingkat.Kepuasan) #[...1...]

Melihat apakah sudah berhasil dalam mengubah variabel tersebut dengan menggunakan function str

str(data_intro$Jenis.Kelamin) #[...2...]

str(data_intro$Produk) #[...2...]

str(data_intro$Tingkat.Kepuasan) #[...2...]

Console

```
> ## Mengubah data menjadi factor untuk membedakan data kualitatif dengan menggunakan
functon as.factor
> data_intro$Jenis.Kelamin <- as.factor(data_intro$Jenis.Kelamin) #[...1...]

> data_intro$Produk <- as.factor(data_intro$Produk) #[...1...]

> data_intro$Tingkat.Kepuasan <- as.factor(data_intro$Tingkat.Kepuasan) #[...1...]

> ## Melihat apakah sudah berhasil dalam mengubah variabel tersebut dengan menggunaka
n function str
> str(data_intro$Jenis.Kelamin) #[...2...]
 Factor w/ 2 levels "1","2": 1 2 2 1 2 1 1 2 2 2 ...

> str(data_intro$Produk) #[...2...]
 Factor w/ 5 levels "A","B","C","D",..: 1 4 4 1 4 2 2 5 5 5 ...

> str(data_intro$Tingkat.Kepuasan) #[...2...]
 Factor w/ 3 levels "1","2","3": 2 2 3 3 2 3 1 3 1 1 ...
```

Skala Pengukuran Data

Setelah data diubah jenis tipe datanya, selanjutnya adalah pemeriksaan untuk memastikan apakah tipe data setiap variabel sudah sesuai dengan skala pengukuran masing-masing.

Untuk melihat data dan tipe data dapat menggunakan syntax berikut :

```
data_intro

str(data_intro)
```

Tugas Praktek

Lengkapi bagian [...1...] dan [...2...] pada code editor untuk menampilkan variable **data_intro** dan strukturnya dengan function **str**.

Code Editor

```
## melihat data/ pemanggilan data

data_intro #[...1...]

## melihat tipe data

str(data_intro) #[...2...]
```

Console

```
> ## melihat data/ pemanggilan data
> data_intro #[...1...]
   ID.Pelanggan    Nama Jenis.Kelamin Pendapatan Produk  Harga Jumlah   Total
1             1    Arif             1     600000      A 100000      4  400000
2             2    Dian             2    1200000      D 250000      4 1000000
3             3   Dinda             2     950000      D 250000      3  750000
4             4   Fajar             1     400000      A 100000      2  200000
5             5     Ika             2    1200000      D 250000      4 1000000
6             6   Ilham             1     800000      B 150000      4  600000
7             7   Indra             1     950000      B 150000      5  750000
8             8 Kartika             2    1100000      E 300000      3  900000
9             9 Lestari             2     800000      E 300000      2  600000
10           10     Lia             2    1700000      E 300000      5 1500000
11           11   Maria             2     600000      A 100000      4  400000
12           12    Maya             2     950000      B 150000      5  750000
13           13    Mila             2     400000      C 200000      1  200000
14           14   Nurul             2    6450000      D 250000      5 1250000
15           15   Retno             2    1000000      C 200000      4  800000
16           16    Rini             2     800000      B 150000      4  600000
17           17   Rizki             1    1200000      C 200000      5 1000000
18           18    Sari             2     700000      D 250000      2  500000
```

```
19           19    Tyas             2     600000      A 100000      4  400000
20           20   Wahyu             1     800000      C 200000      3  600000
   Tingkat.Kepuasan
1                 2
2                 2
3                 3
4                 3
5                 2
6                 3
7                 1
8                 3
9                 1
10                1
11                3
12                3
13                2
14                1
15                2
16                1
17                3
18                1
19                3
20                1

> ## melihat tipe data
> str(data_intro) #[...2...]
'data.frame':    20 obs. of  9 variables:
 $ ID.Pelanggan    : chr  "1" "2" "3" "4" ...
 $ Nama            : chr  "Arif" "Dian" "Dinda" "Fajar" ...
 $ Jenis.Kelamin   : Factor w/ 2 levels "1","2": 1 2 2 1 2 1 1 2 2 2 ...
 $ Pendapatan      : int  600000 1200000 950000 400000 1200000 800000 950000 1100000
800000 1700000 ...
 $ Produk          : Factor w/ 5 levels "A","B","C","D",..: 1 4 4 1 4 2 2 5 5 5 ...
 $ Harga           : int  100000 250000 250000 100000 250000 150000 150000 300000 300
000 300000 ...
 $ Jumlah          : int  4 4 3 2 4 4 5 3 2 5 ...
 $ Total           : int  400000 1000000 750000 200000 1000000 600000 750000 900000 6
00000 1500000 ...
 $ Tingkat.Kepuasan: Factor w/ 3 levels "1","2","3": 2 2 3 3 2 3 1 3 1 1 ...
```

Estimasi karakteristik

Ukuran pemusatan (mean,modus,median, presentil)

1. **Modus** adalah nilai yang sering muncul dari suatu distribusi (data nominal-data rasio).
2. **Median** adalah nilai tengah dari suatu distribusi (data interval dan rasio).
3. **Mean** adalah rata-rata aritmatik dari suatu distribusi (data interval dan rasio).

Contoh

Data : 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12

Modus = 9

Median = 9

Mean = 7,81

Estimasi Nilai Statistik Modus

Modus merupakan nilai yang menunjukan nilai yang sering muncul. Modus digunakan untuk data bertipe nominal dan ordinal.

Untuk menampilkan modus dari data dapat menggunakan syntax

```
Mode(data_intro$Produk)
```

Berikut penjelasan function diatas:

- **Mode** akan menampilkan nilai terbanyak pada variabel yang diamati.
- **data_intro$Produk**, merupakan kolom Produk dari variable data_intro.

Untuk menggunakan function Mode tersebut, menggunakan library tambahan bernama **"pracma"**.

Tugas Praktek

Lengkapi bagian [...1...] pada code editor untuk membaca file seperti yang ditunjukkan pada bagian Lesson. Dan lengkapi bagian [...2...] untuk melihat modus pada kolom tingkat kepuasan.

Jika berjalan dengan lancar maka akan tampil sebagian dataset pada Console sebagai berikut.

```
> Mode(data_intro$Produk)
[1] "D"
> Mode(data_intro$Tingkat.Kepuasan)
[1] "3"
```

Code Editor

```
library(pracma)

## carilah modus untuk kolom Produk pada variable data_intro

Mode(data_intro$Produk) #[...1...]

## carilah modus untuk kolom Tingkat.Kepuasan pada variable data_intro

Mode(data_intro$Tingkat.Kepuasan) #[...2...]
```

Console

```
> library(pracma)

> ## carilah modus untuk kolom Produk pada variable data_intro
> Mode(data_intro$Produk) #[...1...]
[1] "D"

> ## carilah modus untuk kolom Tingkat.Kepuasan pada variable data_intro
> Mode(data_intro$Tingkat.Kepuasan) #[...2...]
[1] "3"
```

Estimasi Nilai Statistik Median

Median merupakan nilai tengah dari suatu kumpulan data. median digunakan untuk data bertipe interval dan rasio.

Untuk menampilkan mean dari data dapat menggunakan syntax

```
median(data_intro$Pendapatan)
```

Berikut penjelasan function diatas:

- **median** akan menampilkan nilai tengah pada variabel yang diamati.
- **data_intro$Pendapatan**, merupakan kolom Pendapatan dari variable data_intro.

Tugas Praktek

Lengkapi bagian […1…], [...2...], [...3...], dan [...4...] pada code editor
untuk menghasilkan median dari kolom Pendapatan, Harga, Jumlah dan Total seperti yang diinstruksikan pada comment Lesson.

Code Editor

```
## carilah median untuk kolom Pendapatan dari variable data_intro

median(data_intro$Pendapatan) #[...1...]


## carilah median untuk  kolom Harga dari variable data_intro

median(data_intro$Harga) #[...2...]


## carilah median untuk kolom Jumlah dari variable data_intro

median(data_intro$Jumlah) #[...3...]


## carilah median untuk  kolom Total dari variable data_intro

median(data_intro$Total) #[...4...]
```

Console

```
> ## carilah median untuk kolom Pendapatan dari variable data_intro
> median(data_intro$Pendapatan) #[...1...]
[1] 875000

> ## carilah median untuk  kolom Harga dari variable data_intro
> median(data_intro$Harga) #[...2...]
[1] 2e+05

> ## carilah median untuk kolom Jumlah dari variable data_intro
> median(data_intro$Jumlah) #[...3...]
[1] 4

> ## carilah median untuk  kolom Total dari variable data_intro
> median(data_intro$Total) #[...4...]
[1] 675000
```

Estimasi Nilai Statistik Rata-Rata

Rata-rata merupakan nilai yang menunjukan nilai rata-rata aritmatik. Rata-rata/mean digunakan untuk data bertipe interval dan rasio.

Untuk menampilkan mean dari data dapat menggunakan syntax

```
mean(data_intro$Pendapatan)
```

Berikut penjelasan function diatas:

- **mean** akan menampilkan nilai rata-rata pada variabel yang diamati.
- **data_intro$Pendapatan**, merupakan kolom Pendapatan dari variable data_intro.

Tugas Praktek

Lengkapi bagian [...1...], [...2...], [...3...], dan [...4...] pada code editor untuk membaca file seperti yang ditunjukkan pada bagian Lesson.

Code Editor

```
## carilah mean untuk kolom Pendapatan pada variable data_intro

mean(data_intro$Pendapatan) #[...1...]

## carilah mean untuk kolom Harga pada variable data_intro

mean(data_intro$Harga) #[...2...]

## carilah mean untuk kolom Jumlah pada variable data_intro

mean(data_intro$Jumlah) #[...3...]

## carilah mean untuk kolom Total pada variable data_intro

mean(data_intro$Total) #[...4...]
```

Console

```
> ## carilah mean untuk kolom Pendapatan pada variable data_intro
> mean(data_intro$Pendapatan) #[...1...]
[1] 1160000

> ## carilah mean untuk kolom Harga pada variable data_intro
> mean(data_intro$Harga) #[...2...]
[1] 197500

> ## carilah mean untuk kolom Jumlah pada variable data_intro
> mean(data_intro$Jumlah) #[...3...]
[1] 3.65

> ## carilah mean untuk kolom Total pada variable data_intro
> mean(data_intro$Total) #[...4...]
[1] 710000
```

Penggunaan Mean dan Median

Dari contoh praktik sebelumnya ada perbedaan hasil **Median** dan **Mean** untuk data interval dan rasio.Maka perlu diperhatikan untuk penggunaanya yaitu:
penggunaan **mean** sebaiknya digunakan jika tidak ada **outlier**. Sebaliknya jika ada outlier, maka sebaiknya menggunakan **Median**.

Apa itu Outlier ? Outlier adalah data yang jaraknya jauh dari keseluruhan data.

Klik tombol Next untuk melanjutkan.

Ukuran Sebaran Data

Ukuran sebaran yang sering digunakan adalah sebagai berikut:

a. **Range** adalah selisih antara nilai terbesar dan nilai terendah
b. **Varians** adalah simpangan kuadrat data dari nilai rata-ratanya

$$\sigma^2 = \frac{\sum (x-\mu)^2}{(n-1)}$$

c. **Simpangan baku** adalah simpangan data dari nilai rata-ratanya, simpangan baku nama lainnya adalah standard error. Standard error dapat digunakan untuk melihat keakuratan dari hasil estimasi, semakin kecil standard error semakin akurat hasil estimasi.

$$\sigma = \sqrt{\sigma^2}$$

Klik tombol Next untuk melanjutkan.

Estimasi Nilai Sebaran Data Range

Range adalah selisih antara nilai terbesar dan nilai terendah. Untuk menampilkan range dari data dapat menggunakan syntax sebagai berikut.

```
max(data_intro$Jumlah)-min(data_intro$Jumlah)
```

Berikut penjelasan function diatas:

- **max** digunakan untuk mendapatkan nilai maksimal dari data.
- **min** adalah function yang digunakan mendapatkan nilai minimal dari data.

Tugas Praktek

Gantilah bagian […1…] pada code editor dengan perhitungan range dari kolom **Pendapatan** pada variable **data_intro** dengan modifikasi contoh pada Lesson.

Code Editor

```
## carilah range untuk kolom Pendapatan pada variable data_intro

max(data_intro$Pendapatan)-min(data_intro$Pendapatan) #[...1...]
```

Console

```
> ## carilah range untuk kolom Pendapatan pada variable data_intro
> max(data_intro$Pendapatan)-min(data_intro$Pendapatan) #[...1...]
[1] 6050000
```

Estimasi Nilai Sebaran Data Varians

Varians merupakan simpangan kuadrat data dari nilai rata-ratanya. Untuk menampilkan varians dari data dapat menggunakan syntax sebagai berikut

```
var(data_intro$Pendapatan)
```

dimana

- **var** adalah function yang digunakan untuk mendapatkan nilai varians dari data.

Tugas Praktek

Ganti bagian […1…] dengan perintah untuk menghitung nilai varians kolom **Pendapatan** dari variable **data_intro**.

Jika berjalan dengan baik, maka hasilnya akan muncul sebagai berikut.

```
[1] 1.645684e+12
```

Keterangan: e+12 menunjukkan 10 pangkat 12. Jadi nilai di atas lengkapnya adalah 1.645.684.210.526.

Code Editor

Carilah varians untuk kolom Pendapatan dari variable data_intro

var(data_intro$Pendapatan) #[...1...]

Console

```
> ## Carilah varians untuk kolom Pendapatan dari variable data_intro
> var(data_intro$Pendapatan) #[...1...]
[1] 1.645684e+12
```

Estimasi Nilai Sebaran Data Simpangan Baku

Simpangan baku adalah simpangan data dari nilai rata-ratanya, simpangan baku nama lainnya adalah **standard deviasi**. Standard deviasi dapat digunakan untuk melihat keakuratan dari hasil estimasi, semakin kecil standard deviasi semakin akurat hasil estimasi.

Untuk menampilkan simpangan baku dari data dapat menggunakan syntax sebagai berikut

```
sd(data_intro$Jumlah)
```

dimana

- **sd** adalah function yang digunakan untuk mendapatkan nilai simpangan baku dari data.

Tugas Praktek

Lengkapi bagian [...1...] pada code editor untuk mengeluarkan hasil simpangan baku dari kolom Pendapatan dari variable data_intro

Code Editor

```
## Carilah simpangan baku untuk kolom Pendapatan dari variable data_intro

sd(data_intro$Pendapatan) #[...1...]
```

Console

```
> ## Carilah simpangan baku untuk kolom Pendapatan dari variable data_intro
> sd(data_intro$Pendapatan) #[...1...]
[1] 1282842
```

Kesimpulan

Dari pembahasan materi diatas maka kesimpulannya sebagai berikut:

- Statistik merupakan ilmu pengolahan,penyajian dan analisis data.
- Jenis-jenis data yaitu nominal, ordinal, interval, dan rasion.
- Estimasi karakteristik data yang sering digunakan diantaranya mean, median dan modus.
- Jenis sebaran data diantaranya range (jarak), standar deviasi, dan varians.

Dengan menyelesaikan bab pertama ini maka sudah dapat melanjutkan ke bab berikutnya

Analisis Deskriptif pada variable **data_intro**

Analisis Deskriptif adalah proses analisa yang digunakan untuk membangun sebuah hipotesis.

Pada bab ini, analisis deskriptif akan dilakukan pada data sebelumnya dengan tujuan untuk mendapatkan informasi berikut:

- Bagaimana profil pelanggan.
- Bagaimana gambaran produk.
- Membangun hipotesis.

Klik tombol Next untuk mulai melanjutkan ke teori dan praktek untuk melakukan ketiga hal tersebut.

Analisis Deskriptif Menggunakan Nilai Statistik

Untuk melakukan analisis deskriptif setiap variabel pada R, kita dapat menggunakan function berikut.

```
summary(data_intro)
```

Function summary akan menampilkan kesimpulan pada variabel masing-masing. Untuk variabel bertipe character akan menampilkan panjang datanya. Variabel bertipe factor akan menampilkan jumlah data pada masing-masing kelas. Sedangkan untuk variabel bertipe numerik akan memunculkan nilai minimum, Q1,Q2 (median), Q3, mean, dan maximum.

Pengertian dari masing-masing istilah itu adalah sebagai berikut :

- **Minimum** adalah nilai observasi terkecil.
- **Kuartil pertama (Q1)**, yang memotong 25 % dari data terendah.
- **Median (Q2)** atau nilai pertengahan.
- **Kuartil ketiga (Q3)**, yang memotong 25 % dari data tertinggi.
- **Maksimum** adalah nilai observasi terbesar.

Tugas Praktek

Gantilah bagian […1...] pada code editor untuk mendapatkan *summary* dari variable **data_intro**.

Code Editor

```
## carilah summary data dari data_intro

summary(data_intro) #[...1...]
```

Console

```
> ## carilah summary data dari data_intro
> summary(data_intro) #[...1...]
 ID.Pelanggan           Nama           Jenis.Kelamin   Pendapatan      Produk
 Length:20          Length:20          1: 6          Min.   : 400000   A:4
 Class :character   Class :character   2:14          1st Qu.: 675000   B:4
 Mode  :character   Mode  :character                 Median : 875000   C:4
                                                     Mean   :1160000   D:5
                                                     3rd Qu.:1125000   E:3
                                                     Max.   :6450000
     Harga            Jumlah         Total         Tingkat.Kepuasan
 Min.   :100000   Min.   :1.00   Min.   : 200000   1:7
 1st Qu.:150000   1st Qu.:3.00   1st Qu.: 475000   2:5
```

```
 Median :200000   Median :4.00   Median : 675000   3:8
 Mean   :197500   Mean   :3.65   Mean   : 710000
 3rd Qu.:250000   3rd Qu.:4.25   3rd Qu.: 925000
 Max.   :300000   Max.   :5.00   Max.   :1500000
```

Analisis Deskriptif Menggunakan Visualisasi

Setelah melakukan analisis deskriptif sebelumnya, agar lebih jelas bagaimana gambaran/sebaran dari data maka kita perlu membuat grafik dari masing-masing variabel. Grafik disini juga dapat sebagai analisis eskplorasi yang akan membantu dalam membangun hipotesis.

Untuk mendapatkan visualisasi dasar dari setiap variabel pada R bisa menggunakan perintah berikut

```
plot(data_intro$Jenis.Kelamin)
```

```
hist(data_intro$Pendapatan)
```

Berikut penjelasan function diatas:

- **plot** digunakan untuk variabel bertipe **Factor** - function ini menghasilkan grafik Bar Plot.
- **hist** untuk variabel bertipe numerik seperti **int** - function ini menghasilkan grafik Histogram.

Tujuan dari plot dan hist adalah untuk mengetahui sebaran data.

Tugas Praktek

Lengkapi bagian […1…], […2…], […3…], […4...], […5...], […6...], dan […7...] pada code editor untuk untuk melakukan visualisasi data. Petunjuknya ada pada tiap comment dari code editor.

Untuk membantu berikut adalah hasil dari perintah str dari variable **data_intro** sehingga Anda bisa memutuskan untuk menggunakan plot atau hist dari kolom terkait.

```
'data.frame':    20 obs. of  9 variables:
 $ ID.Pelanggan    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Nama            : Factor w/ 20 levels "Arif","Dian",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Jenis.Kelamin   : Factor w/ 2 levels "1","2": 1 2 2 1 2 1 1 2 2 2 ...
 $ Pendapatan      : int  600000 1200000 950000 400000 1200000 800000 950000 1100000
800000 1700000 ...
 $ Produk          : Factor w/ 5 levels "A","B","C","D",..: 1 4 4 1 4 2 2 5 5 5 ...
 $ Harga           : int  100000 250000 250000 100000 250000 150000 150000 300000 300
000 300000 ...
 $ Jumlah          : int  4 4 3 2 4 4 5 3 2 5 ...
 $ Total           : int  400000 1000000 750000 200000 1000000 600000 750000 900000 6
00000 1500000 ...
```

```
 $ Tingkat.Kepuasan: Factor w/ 3 levels "1","2","3": 2 2 3 3 2 3 1 3 1 1 ...
```

Code Editor

```
## Carilah sebaran data kolom Jenis.Kelamin dari variable data_intro
plot(data_intro$Jenis.Kelamin) #[...1...]

## Carilah sebaran data dari Pendapatan dari variable data_intro
hist(data_intro$Pendapatan) #[...2...]

## Carilah sebaran data dari Produk dari variable data_intro
plot(data_intro$Produk) #[...3...]

## Carilah sebaran data dari Harga dari variable data_intro
hist(data_intro$Harga) #[...4...]

## Carilah sebaran data dari Jumlah dari variable data_intro
hist(data_intro$Jumlah) #[...5...]

## Carilah sebaran data dari Total dari variable data_intro
hist(data_intro$Total) #[...6...]

## Carilah sebaran data dari Tingkat.Kepuasan dari variable data_intro
plot(data_intro$Tingkat.Kepuasan) #[...7...]
```

Console

(LANGSUNG PRAKTEK)

Kesimpulan Analisis Deskriptif Menggunakan Visualisasi

Dari hasil analisis deskriptif pada praktek sebelumnya kita mendapatkan:

- Profil Pelanggan sebagai berikut:
    1. Sebagian besar pelanggan adalah berjenis kelamin perempuan.
    2. Rata-rata pendapatan pelanggan dalam sebulan adalah 875000 (tidak menggunakan ukuran pemusatan mean, karena pada grafik terdapat outlier. Sehinggan ukuran pemusatan yang dipakai adalah median).
    3. Pelanggan sering membeli produk dalam jumlah 3-4 buah.
    4. Rata-rata total belanja yang sering dihabiskan adalah 710000.
    5. Kebanyakan pelanggan sangat puas kepada produk yang dijual.
- Gambaran produk yang dijual sebagai berikut:
    - Produk yang sering dibeli adalah produk D.
    - Rata-rata harga produk yang terjual sebesar 197500.

Dari hasil statistik deskriptif diatas kita dapat membangun hipotesis, agar analisis data yang kita lakukan kaya informasi yang didapatkan. Pembangunan hipotesis berdasarkan intuisi kita terhadap data yang sudah kita lakukan eksplorasi.

Contoh hipotesis yang dapat kita bangun berdasarkan data diatas adalah sebagai berikut:

1. Apakah ada hubungan pendapatan dengan total belanja?
2. Apakah ada pengaruh suatu produk dengan kepuasan pelanggan?
3. Apakah ada hubungan jenis kelamain dengan total belanja?

Pengenalan Uji Hipotesis

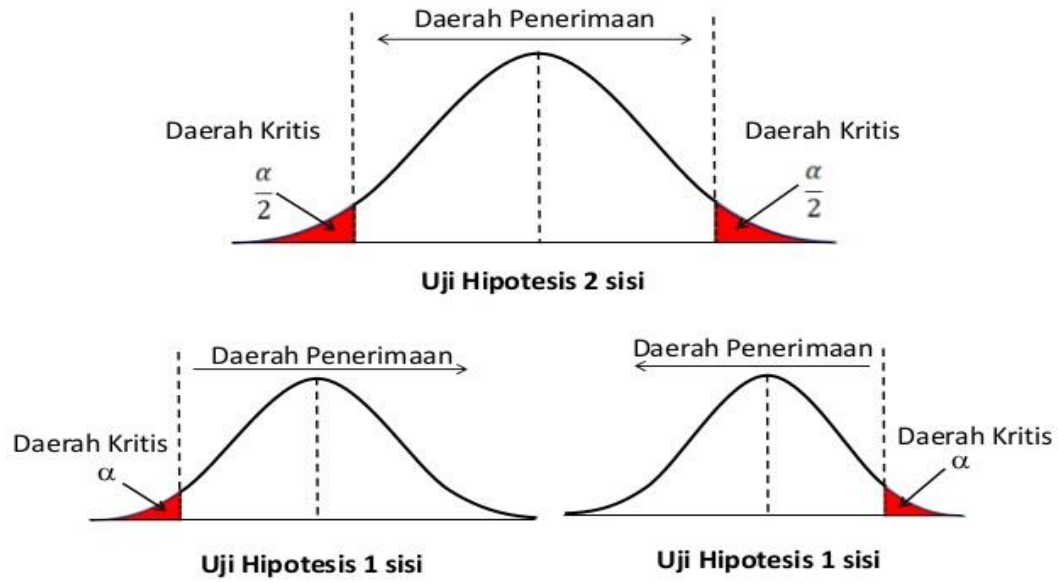
Uji hipotesis adalah metode pengambilan keputusan yang didasarkan dari analisis data. Dalam statistik dapat menguji sebuah hipotesis benar atau salah. Ada 2 jenis hipotesis yaitu hipotesis null (hipotesis nihil) dan hipotesis alternatif. **Hipotesis nihil (Ho)** yaitu hipotesis yang berlawanan dengan teori yang akan dibuktikan. **Hipotesis alternatif (Ha)** adalah hipotesis yang berhubungan dengan teori yang akan dibuktikan.

Dalam melakukan pengujian statistik kita perlu menggunakan metode statistik uji, yang sering digunakan yaitu z-test, t-test, chi-square test, dan f-test. Pada bab kali ini kita tidak akan membahas detail dari setiap statistik uji diatas, tetapi kita akan fokus cara menggunakannya.

Selanjutnya kita harus paham mengenai **p-value** dan **alpha** yang akan digunakan dalam statistik uji.

P-value adalah peluang terkecil dalam menolak Ho. Sedangkan alpha adalah tingkat kesalahan. Nilai alpha biasanya adalah 1%, 5%, dan 10%. Dalam prakteknya alpha 5% sering digunakan, karena lebih moderat.

Hipotesis H0 ditolak jika nilai p-value kurang dari alpha (5%), sedangkan jika p-value lebih dari nilai alpha maka yang H0 diterima. Untuk lebih jelasnya dapat dilihat gambar dibawah ini



Sumber gambar: https://www.slideshare.net/rhandyprasetyo/statistikauji-hipotesis

Perbedaan Statistik Deskriptif dan Statistik Inferensia

Dalam statistik ada 2 jenis analisis data, yaitu **statistik deskriptif** dan **statistik inferensi**. Uji hipotesis, yang dijelaskan pada subbab sebelumnya termasuk kedalam statistik inferensia.

Untuk membedakan antara 2 jenis analisis diatas, maka dapat menyimak penjelasan berikut:

a. **Statistik Deskriptif** adalah statistik yang digunakan untuk analisa data dengan cara menggambarkan data sampel dengan tanpa membuat kesimpulan untuk data populasi. Beberapa hal yang dapat dilakukan adalah penyajian data melalui tabel, grafik, perhitungan modus, median, mean, perhitungan penyebaran data melalui perhitungan rata-rata dan standar deviasi. Statistik Deskriptif digunakan untuk eksplorasi data.
b. **Statistik Inferensia** adalah yaitu statistik yang digunakan untuk menganalisis data sampel dan hasilnya diberlakukan untuk populasi. Beberapa hal yang dapat dilakukan adalah menguji hipotesis dengan statistik uji, seperti chi-square test, student-t test, f-test, z-score test.

Statistik Inferensia dapat digunakan untuk konfirmasi dari hasil statistik deskriptif.

Tujuan Analisis Inferensia

Tujuan Analisis berikutnya dari dataset kita adalah untuk mendapatkan informasi berikut:

- Bagaimana hubungan pendapatan dengan total belanja.
- Bagaimana pengaruh suatu produk dengan kepuasan pelanggan.
- Bagaimana hubungan jenis kelamin dengan total belanja.

Apakah perbedaan antara statistik deskriptif dengan statistik inferensia?

Apakah perbedaan antara statistik deskriptif dengan statistik inferensia?

- [x] Statistik deskriptif sebagai gambaran awal, sedangkan statistik inferensia untuk mengkonfirmasi hasil gambaran awal
- [ ] Statistik deskriptif sebagai konfirmasi atas hasil statistik inferensia
- [ ] Statistik inferensia tidak dapat mengeneralisasi kesimpulan berdasarkan data sampel, sedangkan statistik deskriptif dapat mengeneralisasi kesimpulan
- [ ] Hasil statistik deskriptif berdasarkan peluang, sedangkan statistik inferensia berdasarkan data apa adanya
- [ ] Statistik inferensia untuk melakukan ekplorasi terhadap data, sedangkan statistik deskriptif untuk melakukan gambaran awal

Apakah Fungsi dari membangun sebuah hipotesis dan mengujinya?

Apakah Fungsi dari membangun sebuah hipotesis dan mengujinya?

- [x] Untuk mendapatkan informasi yang lebih akurat
- [x] Untuk menjawab permasalahan berdasarkan data sampel
- [ ] Untuk mendapatkan perhitungan yang tepat
- [ ] Untuk mendapatkan nilai statisik
- [x] Untuk mendapatkan kesimpulan dari data sampel

Analisis Hubungan antar variable

Pada sub-bab ini kita akan membahas cara pengujian hipotesis yang sudah kita susun diatas. Pengujian hipotesis diatas dengan menggunakan analisis inferensia. Ketiga hipotesis diatas dapat digeneralisasi sebagai hipotesis hubungan antar variabel.

Dari penjelasan sebelumnya, kita akan melakukan analisis hubungan antar variable yaitu:

1. Variabel pendapatan dengan total belanja
2. Variabel pengaruh jenis produk dengan kepuasan pelanggan
3. Variabel jenis kelamain dengan total belanja

Klik tombol Next untuk melanjutkan.

Hubungan Antara Variabel Numerik

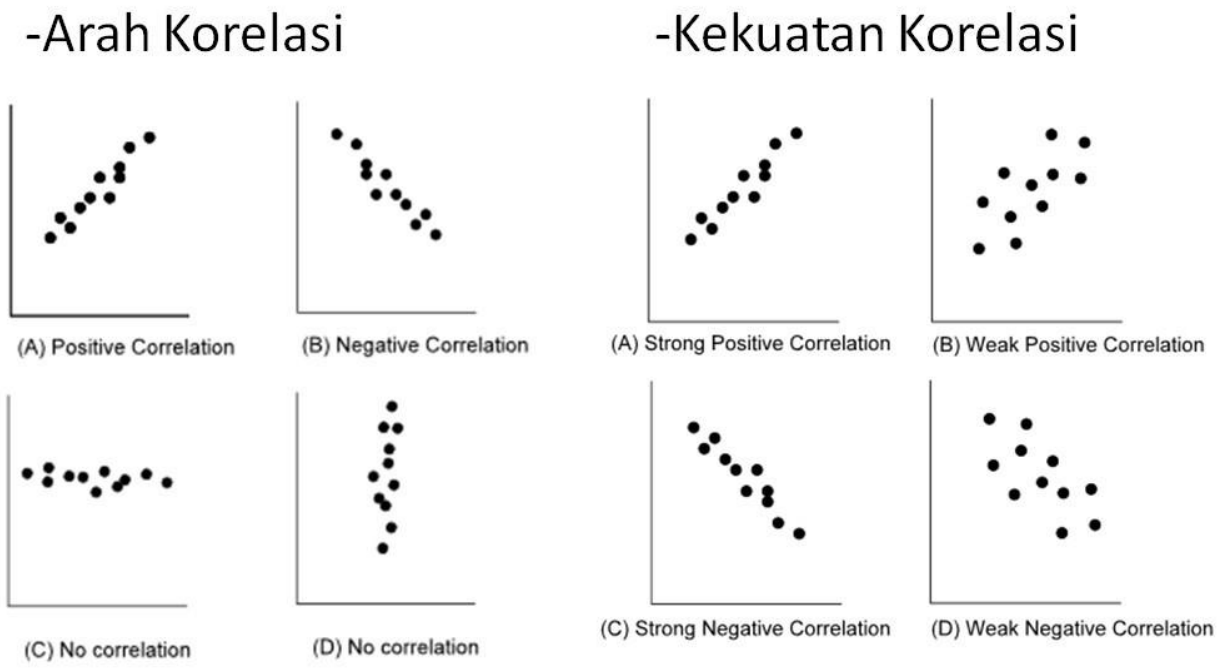
Berdasarkan hasil kasus sebelumnya, kita akan melihat hubungan antara data numerik dan numerik.

Ada dua cara untuk melihat hubungan antar variabel, yaitu dengan grafik **scatter plot** dan **analisis korelasi**. Grafik scatter plot untuk melihat arah hubungan, poisitif dan negatif. Sedangkan analisis korelasi adalah untuk menguji/konfirmasi apakah kedua variabel tersebut memang berhubungan dan sebearapa kuat hubungannya.

Rentang nilai koefisien korelasi antara -1 sampai 1. Korelasi kuat ketika mendakati -1 atau 1, sedangkan dikatakan lemah jika mendekati 0. Untuk mengetahui ada hubungan atau tidaknya menggunakan analisis korelasi, dengan hipotesis sebagai berikut

- Hipotesis nihil (null): tidak ada hubungan antara kedua variabel.
- Hipotesis alternatif: ada hubungan antara kedua variabel.

Berikut gambaran yang lebih jelasnya.



Sumber gambar: https://dsmlmdblog.blogspot.com/2016/03/pengertian-dan-perhitungan-korelasi.html

Scatter Plot

Sebelum melakukan analisis korelasi sebaiknya kita melihat hubungan dari dua variabel numerik menggunakan scatter plot. Scatter plot dapat disebut juga analisis deskriptif.

Untuk melakukan scatter plot pada R menggunakan perintah plot seperti berikut.

```
plot(data_intro$Pendapatan,data_intro$Total)
```

Variabel pertama yaitu data_intro$Pendapatan akan diplot untuk sumbu x,sedangkan variabel kedua yaitu data_intro$Total untuk sumbu y.

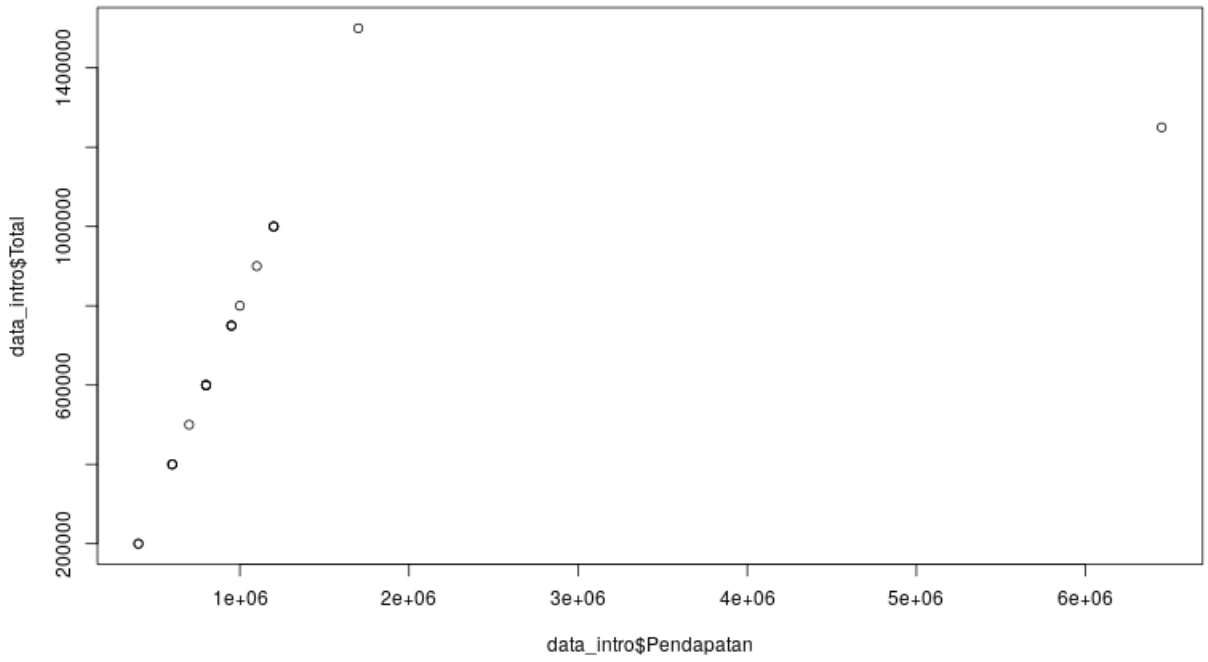
Tugas Praktek

Lengkapi bagian [...1...] pada code editor seperti yang ditunjukkan pada bagian Lesson.

Code Editor

```
plot(data_intro$Pendapatan, data_intro$Total) #[...1...]
```

Console



Hubungan Pendapatan dengan Total Belanja dengan cor.test

Setelah melihat hubungan variabel pendapatan dengan total belanja menggunakan scatter plot diatas maka kita akan mengujinya, apakah benar-benar pendapatan memiliki pengaruh positif terhadap total belanja

Untuk melakukan uji korelasi pada R menggunakan perintah

```
cor.test(data_intro$Pendapatan,data_intro$Total)
```

Berikut penjelasan function diatas:

- Function **cor.test** digunakan untuk melihat hubungan secara statistik.
- Pada korelasi test untuk mengujinya kita memakai t-test. Dengan hipotesis sebagai berikut:
  - **H0** : tidak ada hubungan antara pendapatan dan total belanja.
  - **Ha** : terdapat hubungan antara pendapatan dan total belanja

Tugas Praktek

Lengkapi bagian [...1...] pada code editor untuk seperti yang ada pada bagian Lesson.

Code Editor

```
#Gunakan cor.test untuk mencari hubungan Pendapatan dengan Total Belanja
cor.test(data_intro$Pendapatan, data_intro$Total) #[...1...]
```

Console

```
> #Gunakan cor.test untuk mencari hubungan Pendapatan dengan Total Belanja
> cor.test(data_intro$Pendapatan, data_intro$Total) #[...1...]

        Pearson's product-moment correlation

data:  data_intro$Pendapatan and data_intro$Total
t = 3.1168, df = 18, p-value = 0.005957
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2026033 0.8197871
sample estimates:
      cor
0.5920437
```

Hubungan Antara Variabel Kategorik

Hubungannya diantara keduanya dapat dilihat dengan menggunakan tabulasi silang dan dapat juga dilihat kecenderungannya. Pada hubungan antara variabel kategorik dan kategorik tersebut tidak bisa diketahui seberapa kuat hubungan diantara keduanya dan bagimana pengaruhnya (**positif** atau **negatif**). Untuk mengetahui ada hubungan atau tidaknya menggunakan uji statistik **chi-square test,** dengan hipotesis sebagai berikut:

- **Null hipotesis**: tidak ada hubungan antara kedua variabel
- **Hipotesis Alternatif alternatif**: ada hubungan antara kedua variabel

Hubungan Produk dengan Tingkat Kepuasan dengan chisq.test

Berdasarkan kasus diatas kita akan melihat hubungan antara data kategorik dan kategorik, yaitu variabel jenis produk dan tingkat kepuasan. Sebelum menguji hubungannya, sebaiknya dilakukan tabulasi silang sebagai analisis deskriptif. Selanjutnya analisis inferensia yaitu menguji apakah ada hubungan maka dapat digunakan **chi-square test**.

Untuk melakukan tabulasi dan uji statistik chi-square test pada R tahapannya sebagai berikut

```
table(data_intro$Produk,data_intro$Tingkat.Kepuasan)
chisq.test(table(data_intro$Produk,data_intro$Tingkat.Kepuasan))
```

Perintah **table** untuk melihat tabulasi antar variabel kategorik, sedangkan perintah **chisq.test** digunakan untuk melihat hubungan secara statistik.

Dengan hipotesis sebagai berikut :

- **H0** : tidak ada hubungan antara jenis produk dan tingkat kepuasan.
- **Ha** : terdapat hubungan antara jenis produk dan tingkat kepuasan

Tugas Praktek

Gantilah bagian [...1...] dan [...2...] masing-masing untuk mencari tabulasi antar variabel kategorik dan melihat hubungan secara statistik dengan **chi-square test**.

Code Editor

```
## Carilah tabulasi silang antara kolom jenis produk (Produk) dan tingkat kepuasan
(Tingkat.Kepuasan) dari variable data_intro
table(data_intro$Produk, data_intro$Tingkat.Kepuasan) #[...1...]

## Analisis bagaimana hubungan jenis produk dengan tingkat kepuasan mengunakan
uji korelasi
chisq.test(table(data_intro$Produk, data_intro$Tingkat.Kepuasan)) #[...2...]
```

Console

```
> ## Carilah tabulasi silang antara kolom jenis produk (Produk) dan tingkat kepuasan
(Tingkat.Kepuasan) dari variable data_intro
> table(data_intro$Produk, data_intro$Tingkat.Kepuasan) #[...1...]

   1 2 3
 A 0 1 3
 B 2 0 2
 C 1 2 1
 D 2 2 1
```

```
  E 2 0 1

> ## Analisis bagaimana hubungan jenis produk dengan tingkat kepuasan mengunakan uji
korelasi
> chisq.test(table(data_intro$Produk, data_intro$Tingkat.Kepuasan)) #[...2...]

        Pearson's Chi-squared test

data:  table(data_intro$Produk, data_intro$Tingkat.Kepuasan)
X-squared = 7.95, df = 8, p-value = 0.4384
```

Hubungan Antara Variabel Kategorik dan Variabel Numerik

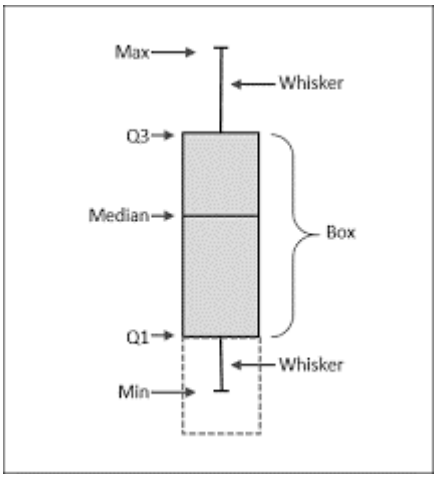
Hubungannya diantara keduanya dapat dilihat dengan membandingkan rata-rata pada setiap kategori. Jika nilai rata-ratanya berbeda maka kedua variabel memiliki hubungan. Pada hubungan antara variabel kategorik dan numerik tidak bisa diketahui seberapa kuat hubungan diantara keduanya dan bagimana pengaruhnya (positif atau negatif).

Untuk mengetahui ada hubungan atau tidaknya menggunakan uji statistik **t-test**, dengan hipotesis sebagai berikut

- **Null hipotesis**: tidak ada hubungan antara kedua variabel
- **Hipotesis Alternatif alternatif**: ada hubungan antara kedua variabel

Hubungan Jenis Kelamin dengan Total Belanja dengan t.test

Berdasarkan kasus diatas kita akan melihat hubungan antara data kategorik dan numerik, yaitu variabel jenis kelamin dan total belanja. Sebelum menguji hubungannya, sebaiknya dilihat perbedaan rata-rata total belanja untuk laki-laki dan perempuan dengan visualisasi yaitu menggunakan boxplot. Boxplot grafik statistik dari data dengan komponen lima ukuran statistik yaitu Min, Q1,Q2, Q3, dan Max. Untuk lebih jelasnya mengenai boxplot dapat dilihat gambar dibawah ini



Selanjutnya analisis inferensia yaitu untuk mengetahui apakah ada perbedaan rata-rata total belanja pada laki-laki dan perempuan maka digunakan statistik uji t-test.

Untuk melakukan visualisasi boxplot dan uji statistik t-test pada R tahapannya sebagai berikut

```
boxplot(Total~Jenis.Kelamin,data = data_intro)
t.test(Total~Jenis.Kelamin,data = data_intro)
```

Function **boxplot** digunakan untuk melihat secara grafik rata-rata total belanja pada laki-laki dan perempuan, sedangkan perintah **t.test** digunakan untuk melihat hubungan secara statistik. Penggunaan kedua fungsi diatas yaitu variabel pertama yang bertipe numerik, sedangkan variabel kedua variabel kategorik. Hipotesis t-test sebagai berikut :

- Null hipotesis : tidak ada perbedaan rata-rata total belanja antara laki-laki dan perempuan
- Hipotesis alternatif : ada perbedaan rata-rata total belanja antara laki-laki dan perempuan

Tugas Praktek

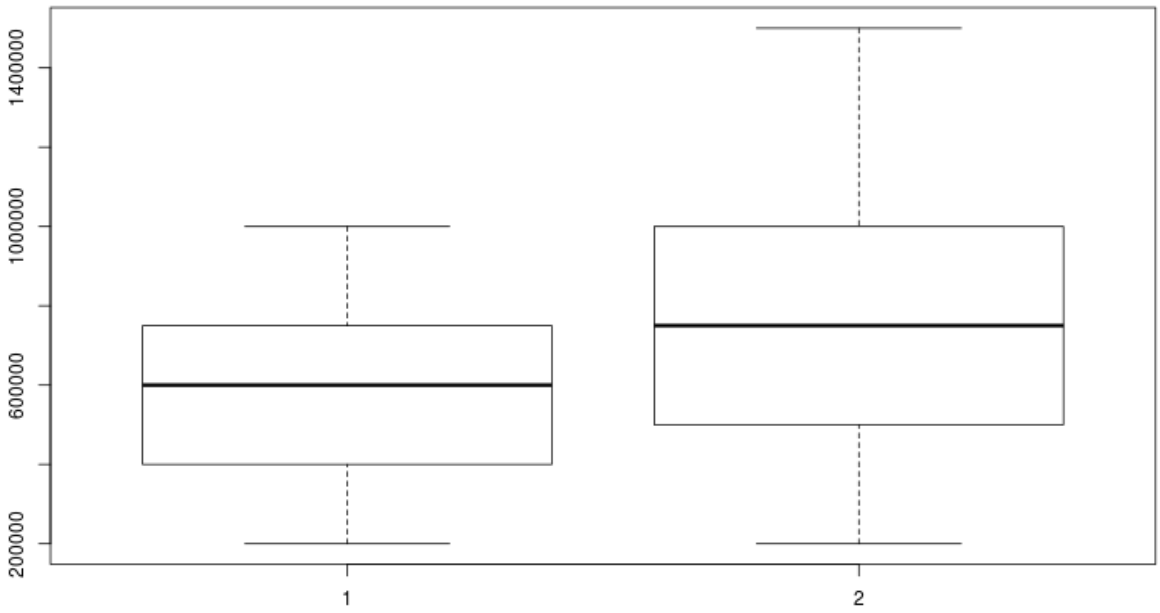
Lengkapi bagian […1...] dan […2…] pada code editor untuk membaca file seperti yang ditunjukkan pada bagian Lesson.

Code Editor

```
## carilah boxplot antara variabel jenis kelamin dengan total belanja
boxplot(Total~Jenis.Kelamin, data=data_intro) #[...1...]

## analisis bagaimana hubungan jenis kelamin dengan total belanja mengunakan uji statistik t-test
t.test(Total~Jenis.Kelamin, data=data_intro) #[...2...]
```

Console



Pernyataan berikut yang benar adalah

Pernyataan berikut yang benar adalah

- [x] Untuk menguji hubungan variabel numerik dengan kategorik menggunakan t-test
- [x] Untuk menguji hubungan variabel kategorik dengan kategorik menggunakan chi-square test
- [ ] Untuk menguji hubungan variabel numerik dengan numerik menggunakan f-test
- [ ] Untuk Menguji hubungan variabel kategorik dengan numerik menggunakan chi-square test
- [x] Untuk menguji hubungan variabel numerik dengan numerik menggunakan cor.test

Cara analisis yang runtut adalah...

Cara analisis yang runtut adalah:

1. *Analisis Inferensia*
2. *Eksplorasi data*
3. *Membuat Hipotesis*
4. *Memberi Kesimpulan*
5. *Memberikan rekomendasi berdasarkan analisis data*

- [ ] 1,2,3,4,5
- [x] 2,3,1,4,5
- [ ] 3,1,2,4,5
- [ ] 4,5,3,1,2
- [ ] 3,2,1,4,5

Kesimpulan

Selamat! Dengan menyelesaikan bab kedua ini maka Anda sudah menyelesaikan course Introduction to Statictics with R yang singkat namun padat ini!

Dari pembahasan materi diatas maka kesimpulannya sebagai berikut:

- Sebelum memulai menganalisis data harus dilihat summary per tiap variabel.
- Analisis Deskriptif digunakan untuk membangun sebuah hipotesis.
- Analisis Inferensia digunakan untuk menguji hipotesis.

Klik tombol Next untuk mengakhiri course ini.