MODUL 1
UJI DATA (1)
ANALISIS MISSING VALUE & OUTLIER

Tujuan dari praktikum modul 1 ini, agar mahasiswa mampu :

1. Mengenali karakteristik *missing value*.
2. Memberikan perlakuan atau solusi pemecahan terhadap data yang *missing*.

Materi

1. Missing Value

Missing value adalah *informasi yang tidak tersedia untuk sebuah objek (kasus). Missing value* terjadi karena informasi untuk sesuatu tentang objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak ada.
Missing value pada dasarnya tidak bermasalah bagi keseluruhan data, apalagi jika jumlahnya hanya sedikit, misal hanya 1 % dari seluruh data. Namun jika persentase data yang hilang tersebut cukup besar, maka perlu dilakukan pengujian apakah data yang mengandung banyak *missing* tersebut masih layak diproses lebih lanjut ataukah tidak.

Untuk lebih jelasnya, aplikasi dalam SPSS terhadap *missing value* akan dijelaskan dengan contoh kasus berikut ini.
Setelah dilakukan survey di 20 region terhadap 5 variabel (jumlah penduduk, jumlah pendapatan daerah, luas lahan pertanian, jumlah pendapatan sektor perdagangan, dan jumlah pendapatan sektor industri) diperoleh data sebagai berikut :

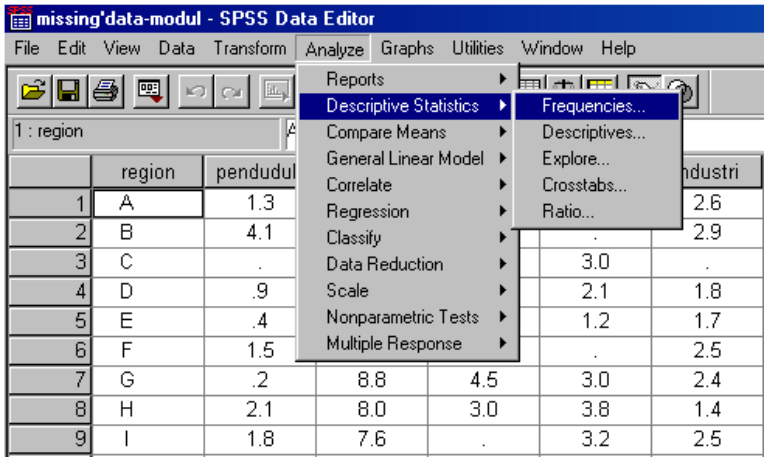
	region	penduduk	pendptan	ptanian	pdgangan	industri
1	A	1.3	9.9	6.7	3.0	2.6
2	B	4.1	5.7	.	.	2.9
3	C	.	9.9	.	3.0	.
4	D	.9	8.6	.	2.1	1.8
5	E	.4	8.3	.	1.2	1.7
6	F	1.5	6.7	4.8	.	2.5
7	G	.2	8.8	4.5	3.0	2.4
8	H	2.1	8.0	3.0	3.8	1.4
9	I	1.8	7.6	.	3.2	2.5
10	J	4.5	8.0	.	3.3	2.2
11	K	2.5	9.2	.	3.3	3.9
12	L	4.5	6.4	5.3	3.0	2.5
13	M	.	.	.	.	2.7
14	N	2.8	6.1	6.4	.	3.8
15	O	3.7	.	.	3.0	.
16	P	1.6	6.4	5.0	.	2.1
17	Q	.5	9.2	.	3.3	2.8
18	R	2.8	5.2	5.0	.	2.7
19	S	2.2	6.7	.	2.6	2.9
20	T	1.8	9.0	5.0	2.2	3.0

Permasalahan yang muncul dari data di atas yaitu :

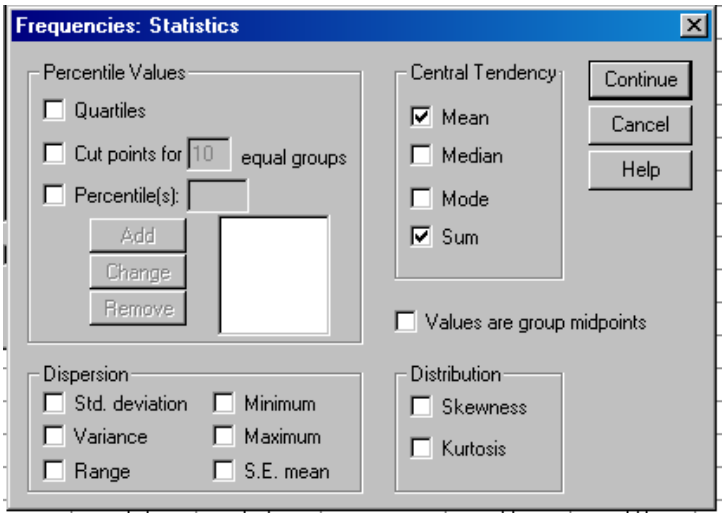
a. Bagaimana deskripsi *missing value* yang terdapat dalam data tersebut ?
b. Bagaimana memperlakukan objek / kasus yang memiliki *missing value* ?

Untuk dapat menjawab permasalahan yang ada, lakukan tahapan pekerjaan dengan menggunakan aplikasi alat SPSS seperti berikut ini.

1. Dari data yang telah dimasukkan, selanjutnya klik menu "**analyze**" dan pilih sub menu "**descriptive statistics**" dan kemudian "**frequencies**" seperti tampilan berikut ini :



2. Masukkan variabel-variabel yang akan dicari numeriknya (penduduk, pendapatan, pertanian, perdagangan, dan industri) pada kotak "**variable(s)**". kemudian klik kotak "*staistics*" dan pilih "**sum**" dan "**mean**" pada *central tendency*. Berikutnya klik "**continue**".



3. Akhiri dengan meng-klik **OK** untuk menampilkan output.

Untuk menjawab "**permasalahan a"**, tampilan output SPSS berikut ini dapat membantu memberikan deskripsi.

Statistics

		jumlah penduduk region (dlm juta jiwa)	pendptan daerah (dlm trilyun rupiah)	luas lahan pertanian (dlm ratusan hektar)	jumlah penerimaan sektor perdagangan (dlm milyar Rp)	jumlah penerimaan sektor industri (dalam milyar Rp)
N	Valid	18	18	9	14	18
	Missing	2	2	11	6	2
Mean		2.178	7.761	5.078	2.857	2.578
Sum		39.2	139.7	45.7	40.0	46.4

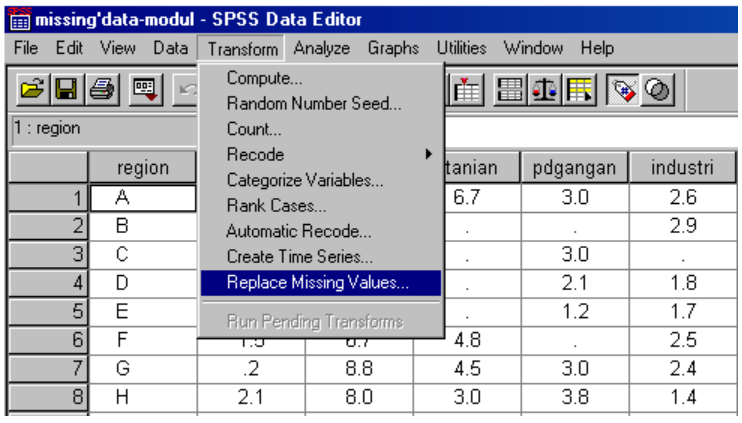
Dari tabel di atas dapat diketahui berapa missing value yang terdapat pada masing-masing variabel data.

- Variabel jumlah penduduk terdapat 2 *missing value* dari 20 objek pengamatan, sehingga 18 objek yang dianggap valid.
- Variabel pendapatan daerah terdapat 2 *missing value* dari 20 objek pengamatan, sehingga 18 objek yang dianggap valid.
- Variabel luas lahan pertanian terdapat 11 *missing value* dari 20 objek pengamatan, sehingga hanya 9 objek yang dianggap valid.
- Variabel penerimaan sektor perdagangan terdapat 6 *missing value* dari 20 objek pengamatan, sehingga 14 objek yang dianggap valid.
- Variabel penerimaan sektor industri terdapat 2 *missing value* dari 20 objek pengamatan, sehingga 18 objek yang dianggap valid.

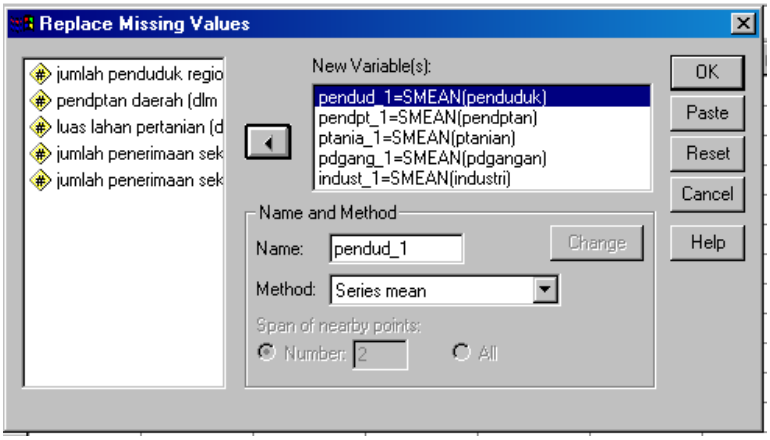
Dengan demikian dapat diketahui berapa persentase validitas data masing-masing variabel.

Sedangkan untuk mengetahui bagaimana cara memperlakukan data yang memiliki *missing value*; berikut ini tahapan yang harus dilakukan dalam aplikasi SPSS.

1. Dengan kembali pada tampilan data yang telah di*entry*, selanjutnya dari menu **transform**, pilih sub menu "**Replace missing value**".



2. Masukkan variabel-variabel yang memiliki *missing value* ke bagian **New Variable**. Perhatikan isi di bagian "**name and method**".



3. Akhiri dengan mengklik **OK** untuk menampilkan output dari aplikasi SPSS seperti berikut ini.

Result Variable	Missing Values Replaced	First Non-Miss	Last Non-Miss	Valid Cases	Creating Function
PENDUD_1	2	1	20	20	SMEAN(PENDUDUK)
PENDPT_1	2	1	20	20	SMEAN(PENDPTAN)
PTANIA_1	11	1	20	20	SMEAN(PTANIAN)
PDGANG_1	6	1	20	20	SMEAN(PDGANGAN)
INDUST_1	2	1	20	20	SMEAN(INDUSTRI)

4. Berikutnya setelah memperoleh output tersebut, kembali ke tampilan data sebelumnya. Ternyata terjadi perubahan yang nyata, yaitu cell-cell yang tadinya memiliki *missing value* kini telah terisi oleh suatu nilai (**nilai *mean* dari tiap-tiap variabel**).

	region	penduduk	pendptan	ptanian	pdgangan	industri	pendud_1	pendpt_1	ptania_1	pdgang_1	indust_1
1	A	1.3	9.9	6.7	3.0	2.6	1.30	9.90	6.70	3.00	2.60
2	B	4.1	5.7	.	.	2.9	4.10	5.70	5.08	2.86	2.90
3	C	.	9.9	.	3.0	.	2.18	9.90	5.08	3.00	2.58
4	D	.9	8.6	.	2.1	1.8	.90	8.60	5.08	2.10	1.80
5	E	.4	8.3	.	1.2	1.7	.40	8.30	5.08	1.20	1.70
6	F	1.5	6.7	4.8	.	2.5	1.50	6.70	4.80	2.86	2.50
7	G	.2	8.8	4.5	3.0	2.4	.20	8.80	4.50	3.00	2.40
8	H	2.1	8.0	3.0	3.8	1.4	2.10	8.00	3.00	3.80	1.40
9	I	1.8	7.6	.	3.2	2.5	1.80	7.60	5.08	3.20	2.50
10	J	4.5	8.0	.	3.3	2.2	4.50	8.00	5.08	3.30	2.20
11	K	2.5	9.2	.	3.3	3.9	2.50	9.20	5.08	3.30	3.90
12	L	4.5	6.4	5.3	3.0	2.5	4.50	6.40	5.30	3.00	2.50
13	M	.	.	.	.	2.7	2.18	7.76	5.08	2.86	2.70
14	N	2.8	6.1	6.4	.	3.8	2.80	6.10	6.40	2.86	3.80
15	O	3.7	.	.	3.0	.	3.70	7.76	5.08	3.00	2.58
16	P	1.6	6.4	5.0	.	2.1	1.60	6.40	5.00	2.86	2.10
17	Q	.5	9.2	.	3.3	2.8	.50	9.20	5.08	3.30	2.80
18	R	2.8	5.2	5.0	.	2.7	2.80	5.20	5.00	2.86	2.70
19	S	2.2	6.7	.	2.6	2.9	2.20	6.70	5.08	2.60	2.90
20	T	1.8	9.0	5.0	2.2	3.0	1.80	9.00	5.00	2.20	3.00

Dengan demikian "**permasalahan b**" dapat dijawab. Langkah di atas merupakan salah satu cara memperlakukan adanya *missing value*, yaitu *dengan memasukkan nilai*

mean dari masing-masing variabel tersebut pada cell yang mengandung missing value. Cara lain dalam penanganan *missing value* yaitu :

- Menghilangkan/membuang kasus atau objek yang mengandung *missing value*.
- Menghapus variabel (kolom) yang mengandung *missing value.*

Dengan demikian terdapat 3 (tiga) cara memperlakukan data yang mengandung *missing value*.

2. Outlier

Data *outlier* (pencilan) adalah *data yang secara nyata berbeda dengan data-data yang lain*. Sebagai contoh, data dari 40 mahasiswa Jurusan Teknik Planologi yang mengikuti matakuliah MAP diperoleh rata-rata nilainya 60, sedangkan ada seorang mahasiswa yang mempunyai nilai MAP 100. Jelas dalam hal ini berarti seorang mahasiswa yang memiliki nilai MAP 100 tersebut merupakan **data *outlier***.

Beberapa hal yang mempengaruhi munculnya data *outlier* antara lain :

1. Kesalahan dalam pemasukan data.
2. Kesalahan dalam pengambilan sampel.
3. Memang ada data-data ekstrim yang tidak dapat dihindarkan keberadaannya.

Untuk lebih jelasnya, kasus berikut ini akan menjelaskan pengujian apakah suatu data mengandung data *outlier*, serta bagaimana cara penanganan data *outlier* tersebut.

Setelah dilakukan survey di 20 *region* terhadap 5 variabel (jumlah penduduk, jumlah pendapatan daerah, luas lahan pertanian, jumlah pendapatan sektor perdagangan, dan jumlah pendapatan sektor industri) diperoleh data sebagai berikut :

	region	penduduk	pendptan	ptanian	pdgangan	industri
1	A	1.3	9.9	6.7	3.0	2.6
2	B	4.1	5.7	3.8	9.6	2.9
3	C	7.9	9.9	5.2	3.0	8.9
4	D	.9	8.6	4.9	2.1	1.8
5	E	.4	8.3	5.3	1.2	1.7
6	F	1.5	6.7	4.8	3.4	2.5
7	G	.2	8.8	4.5	3.0	2.4
8	H	2.1	8.0	3.0	3.8	1.4
9	I	1.8	7.6	6.8	3.2	2.5
10	J	4.5	8.0	12.6	3.3	2.2
11	K	2.5	9.2	3.6	3.3	3.9
12	L	4.5	6.4	5.3	3.0	2.5
13	M	3.6	14.5	4.9	4.2	2.7
14	N	2.8	6.1	6.4	4.1	3.8
15	O	3.7	11.9	5.3	3.0	3.5
16	P	1.6	6.4	5.0	3.9	2.1
17	Q	.5	9.2	5.2	3.3	2.8
18	R	2.8	5.2	5.0	3.6	2.7
19	S	2.2	6.7	4.8	2.6	2.9
20	T	1.8	9.0	5.0	2.2	3.0

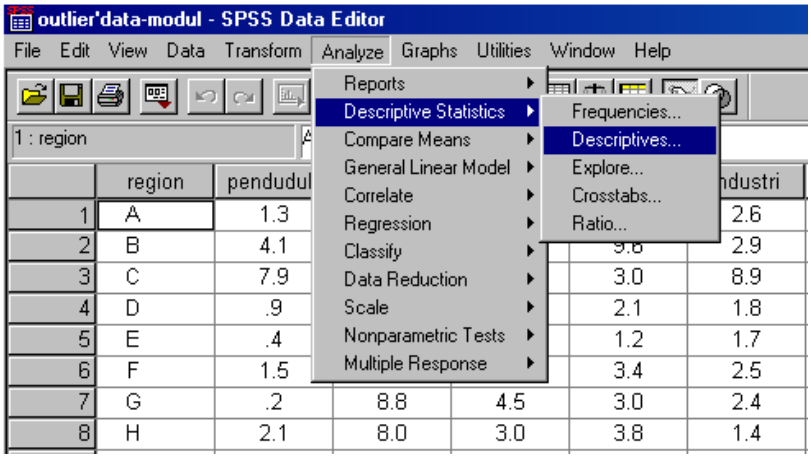
Permasalahan yang muncul dari data di atas yaitu :

a. Apakah data yang telah dikumpulkan tersebut mengandung data *outlier* ? (harus dilakukan deteksi)

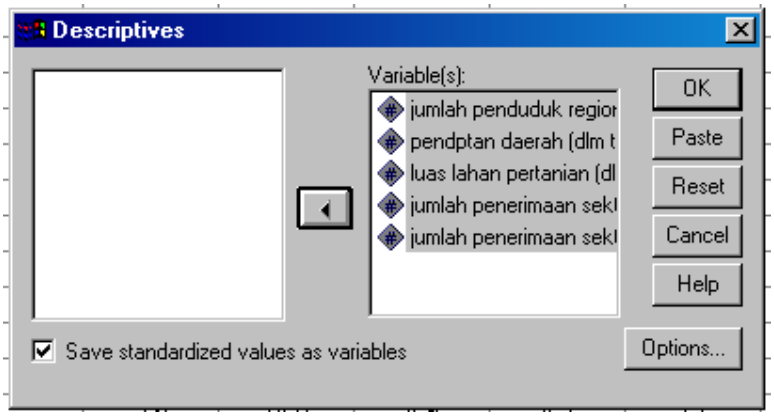
b. Bagaimana memperlakukan data yang mengandung *outlier* tersebut ?

Selanjutnya untuk dapat menjawab permasalahan diatas, harus dilakukan tahapan aplikasi SPSS seperti berikut ini.

1. Dari data yang telah dimasukkan, selanjutnya klik menu "**analyze**" dan pilih sub menu "**descriptive statistics**" dan kemudian "**descriptives**" seperti tampilan berikut ini :



2. Selanjutnya masukkan variabel *penduduk, pendapatan, pertanian, perdagangan, dan industri* ke dalam bagian **Variable(s)**. Aktifkan kotak pilihan **Save standardizedvalues as variables** dengan klik mouse.



3. Akhiri dengan mengklik **OK** untuk menampilkan output aplikasi SPSS seperti berikut ini.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
jumlah penduduk region (dlm juta jiwa)	20	.2	7.9	2.535	1.8248
pendptan daerah (dlm trilyun rupiah)	20	5.2	14.5	8.305	2.2147
luas lahan pertanian (dlm ratusan hektar)	20	3.0	12.6	5.405	1.9286
jumlah penerimaan sektor perdagangan (dlm milyar Rp)	20	1.2	9.6	3.440	1.6136
jumlah penerimaan sektor industri (dalam milyar Rp)	20	1.4	8.9	2.940	1.5405
Valid N (listwise)	20				

Perhatikan kolom **Mean** dan **Std.Deviation** untuk setiap variabel. Sebagai contoh rata-rata jumlah penduduk region adalah 2,535 juta jiwa dengan standar deviasi 1,8248 juta jiwa. Sedangkan rata-rata pendapatan daerah sebesar 8,305 trilyun rupiah dengan standar deviasi 2,2147 trilyun rupiah. Demikian selanjutnya untuk variabel data yang lain.

Selanjutnya untuk melakukan pengujian apakah pada data tersebut terdapat data yang ekstrim (outlier), maka harus dilakukan **standarisasi dengan nilai Z**.

$$Z = \frac{x - \bar{X}}{\sigma}$$

dimana : x = nilai data

$\bar{X}$ = nilai rata-rata

σ = standar deviasi

Sebagai contoh, untuk region A jumlah penduduk 1,3 juta jiwa maka :

$$Z_{penddk'A} = \frac{1,3 - 2,535}{1,8248} = -0,67680$$

sedangkan jika jumlah pendapatan di region A adalah Rp 9,9 trilyun, maka :

$$Z_{pendptan'A} = \frac{9,9 - 8,305}{2,2147} = +0,72020$$

Demikian selanjutnya untuk data yang lain dan variabel lain, sebagaimana hasil standarisasi secara lengkap seperti terlihat pada tampilan tabel data seperti berikut ini.

outlier'data-modul - SPSS Data Editor

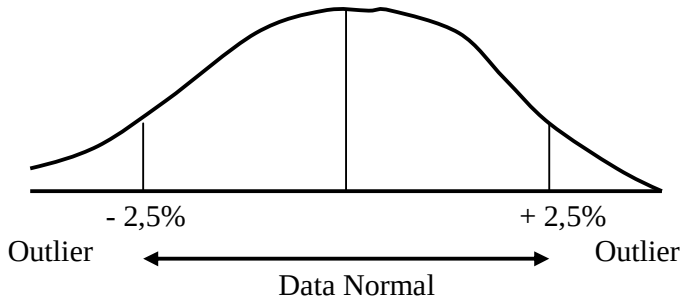
File Edit View Data Transform Analyze Graphs Utilities Window Help

13 : industri | 2.7

	region	penduduk	pendptan	ptanian	pdgangan	industri	zpendudu	zpendpta	zptanian	zpdganga	zindustr
1	A	1.3	9.9	6.7	3.0	2.6	-.67680	.72020	.67148	-.27269	-.22071
2	B	4.1	5.7	3.8	9.6	2.9	.85765	-1.17625	-.83222	3.81764	-.02597
3	C	7.9	9.9	5.2	3.0	8.9	2.94011	.72020	-.10630	-.27269	3.86895
4	D	.9	8.6	4.9	2.1	1.8	-.89601	.13320	-.26185	-.83046	-.74003
5	E	.4	8.3	5.3	1.2	1.7	-1.17001	-.00226	-.05444	-1.38823	-.80495
6	F	1.5	6.7	4.8	3.4	2.5	-.56720	-.72472	-.31370	-.02479	-.28563
7	G	.2	8.8	4.5	3.0	2.4	-1.27962	.22351	-.46926	-.27269	-.35054
8	H	2.1	8.0	3.0	3.8	1.4	-.23839	-.13772	-1.24703	.22311	-.99969
9	I	1.8	7.6	6.8	3.2	2.5	-.40279	-.31833	.72333	-.14874	-.28563
10	J	4.5	8.0	12.6	3.3	2.2	1.07685	-.13772	3.73071	-.08676	-.48037
11	K	2.5	9.2	3.6	3.3	3.9	-.01918	.40413	-.93592	-.08676	.62319
12	L	4.5	6.4	5.3	3.0	2.5	1.07685	-.86018	-.05444	-.27269	-.28563
13	M	3.6	14.5	4.9	4.2	2.7	.58364	2.79727	-.26185	.47101	-.15580
14	N	2.8	6.1	6.4	4.1	3.8	.14522	-.99564	.51592	.40903	.55827
15	O	3.7	11.9	5.3	3.0	3.5	.63844	1.62328	-.05444	-.27269	.36353
16	P	1.6	6.4	5.0	3.9	2.1	-.51240	-.86018	-.21000	.28508	-.54529
17	Q	.5	9.2	5.2	3.3	2.8	-1.11521	.40413	-.10630	-.08676	-.09088
18	R	2.8	5.2	5.0	3.6	2.7	.14522	-1.40202	-.21000	.09916	-.15580
19	S	2.2	6.7	4.8	2.6	2.9	-.18359	-.72472	-.31370	-.52059	-.02597
20	T	1.8	9.0	5.0	2.2	3.0	-.40279	.31382	-.21000	-.76849	.03895

Perhatikan munculnya notasi "Z" di setiap variabel yang terbentuk.

Selanjutnya untuk mendeteksi ada atau tidaknya data yang ekstrim (outlier), dapat dilakukan pengujian dengan menggunakan kurva distribusi normal (sebagaimana data sudah distandarkan). Dengan menggunakan nilai α = 5% maka kurva dapat digambarkan sebagai berikut :



Berdasarkan kurva di atas dapat dikatakan bahwa suatu data dianggap ***outlier*** apabila nilai Z yang didapat adalah (**z > +2,5**) atau (**z < -2,5**).

Dengan demikian "**permasalahan a**" dapat dijawab dengan mengacu pada tabel hasil aplikasi SPSS atas standarisasi variabel (nilai Z) dan kurva di atas, yaitu :

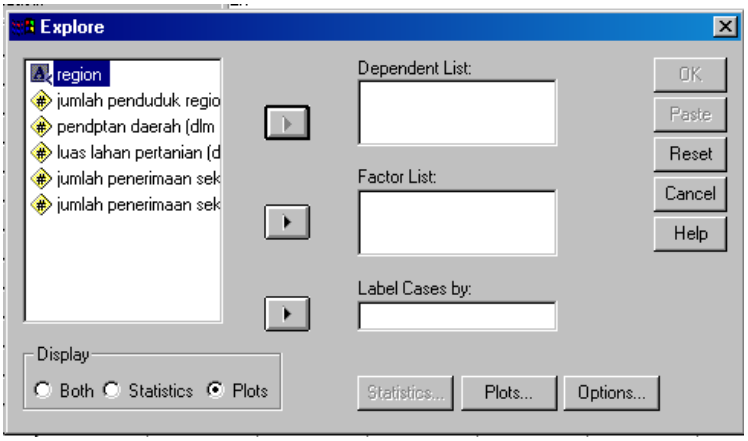
- Pada **variabel jumlah penduduk** nampak bahwa untuk region C memiliki nilai *z = + 2,94011*. Hal ini berarti bahwa **region C adalah data outlier**, yang jika dilihat jumlah penduduk di region C sebesar 7,9 juta jiwa, sedangkan rata-rata jumlah penduduk di ke-20 region tersebut adalah 2,535 juta jiwa. Dengan kata lain region C memiliki *jumlah penduduk jauh melebihi rata-rata jumlah penduduk.*
- Pada **variabel jumlah pendapatan daerah** dapat terlihat bahwa region M memiliki nilai *z = + 2,79727*. Hal ini berarti **region M merupakan data**

outlier (*nilai z > +2,5*). Selain itu nampak pula bahwa jumlah pendapatan daerah di region M adalah sebesar Rp 14,5 trilyun, dimana jumlah tersebut jauh melebihi rata-rata jumlah pendapatan daerah yang sebesar Rp 8,305 trilyun.

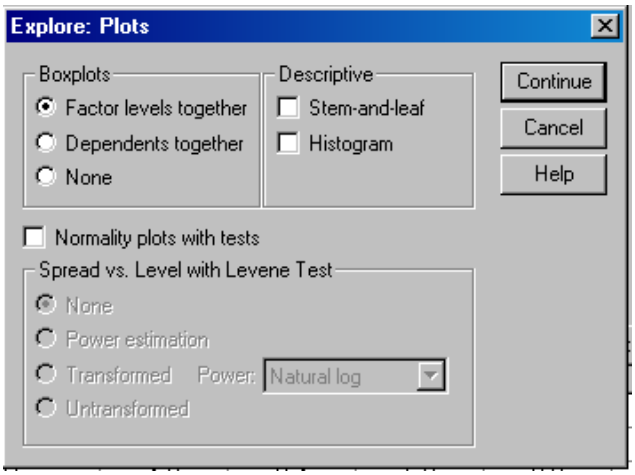
- Demikian juga untuk **variabel luas lahan pertanian**, dimana terdapat data ekstrim (outlier) pada **region J** (*z = + 3,73071*); **variabel jumlah pendapatan sektor perdagangan** outlier terjadi pada **region B** (*z = +3,81764*); dan pada **variabel jumlah pendapatan sektor industri** terkandung data ekstrim (outlier) pada **region C** (*z = + 3,86895*).

Cara lain untuk mengenali adanya data yang ekstrim yaitu dengan melihat dari tampilan box plot data tersebut. Untuk menampilkan output box plot dilakukan beberapa tahapan aplikasi SPSS berikut ini.

1. Dari tampilan tabel data, klik menu "**analyze**" dan pilih sub menu "**descriptive statistics**" kemudian "**explore**". Selanjutnya akan tampak pada layar seperti tampilan berikut ini.

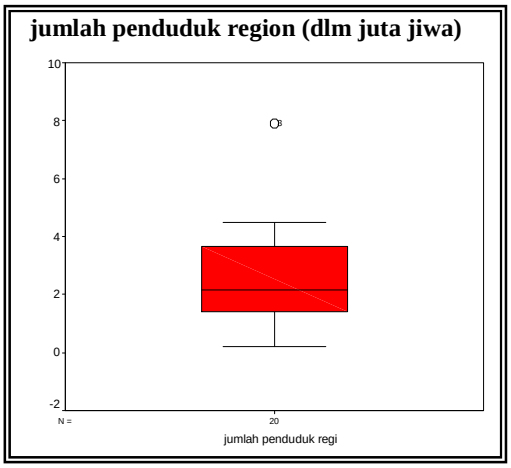


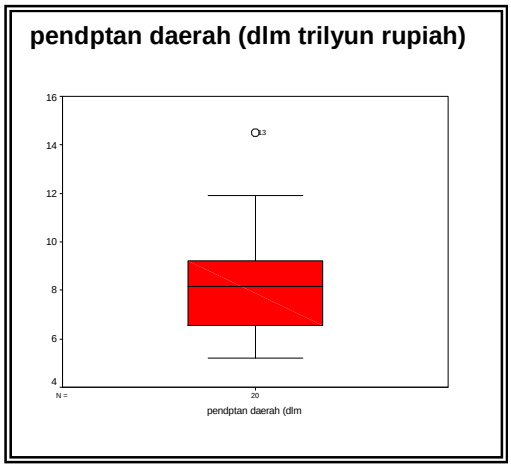
2. Masukkan *variabel jumlah penduduk, pendapatan, luas lahan pertanian, penerimaan sektor perdagangan, dan industri* ke dalam bagian kotak "**Dependent List**". Pada bagian **Display** (*kiri bawah*), klik mouse "**plots**". Kemudian buka kotak "**Plots**" sehingga tampak di layar seperti berikut ini :

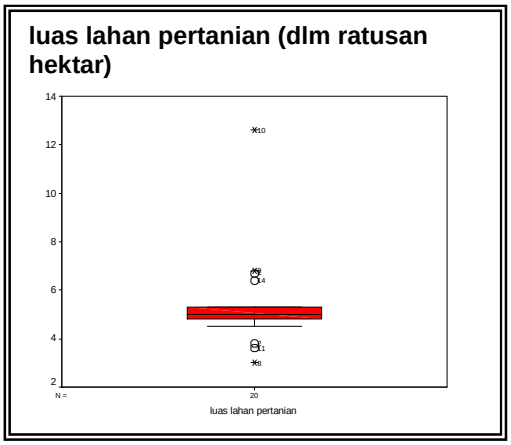


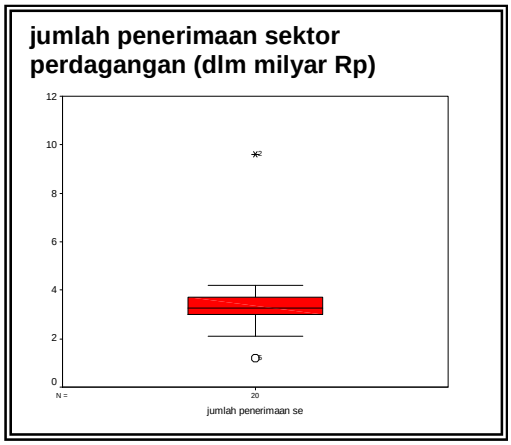
Tampak berbagai bentuk penyajian plot. Oleh karena hanya diinginkan tampilan box plot, maka **non-aktifkan** pilihan "**steam and leaf**" pada bagian *Descriptives*.

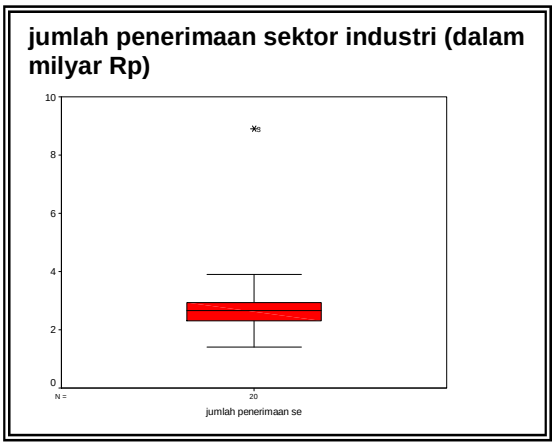
3. Abaikan bagian yang lainnya dan akhiri dengan mengklik **OK** untuk menampilkan output tampilan box plot.











Dari kelima tampilan box plot di atas dapat terlihat bahwa masing-masing variabel adanya data ekstrim (outlier), yaitu pada box plot tersebut tampak keberadaannya jauh di luar sebaran data (box plot). Masing-masing data ekstrim dalam box plot tersebut diberikan tanda objek mana yang menjadi outlier data (hal tersebut akan sangat terlihat pada sreen komputer yang lebih lebar daripada tampilan di atas).

Selanjutnya untuk menjawab "**permasalahan b**" tentang penanganan terhadap adanya data ekstrim (outlier) adalah :

1. Data outlier *dihilangkan,* karena dianggap tidak mencerminkan sebaran data yang sesungguhnya, atau mungkin didapat karena kesalahan pengambilan data, kesalahan *inputing*, dan sebagainya.
2. Data outlier tetap *dipertahankan* karena dianggap memang terdapat data yang seperti itu, atau tidak dapat dikatakan ada kesalahan pada proses sampling maupun inputing data. Namun *pada saat melakukan analisis hendaknya data outlier tersebut dipisahkan* dari data yang lain; karena akan mempengaruhi hasil analisis. Selanjutnya *data outlier tersebut perlu dilakukan analisis tersendiri*.

Tugas Praktikum

Dari data yang tersimpan dalam direktori D:\MAP, pada folder data modul 1, terdapat serangkaian data yang belum diteliti apakah mengandung *missing value* dan *outlier*. Oleh karena itu, lakukanlah tahapan-tahapan dalam menangani *missing value* dan *outlier*, serta :

a. Bagaimana deskripsi *missing value* yang terdapat dalam data tersebut ?
b. Bagaimana memperlakukan objek / kasus yang memiliki *missing value* ?
c. Apakah data yang telah dikumpulkan tersebut mengandung data *outlier* ? (harus dilakukan deteksi)
d. Bagaimana memperlakukan data yang mengandung *outlier* tersebut ?

Untuk itu, gunakan analisis *missing value* dan *outlier.*