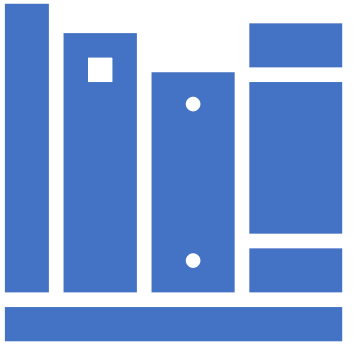
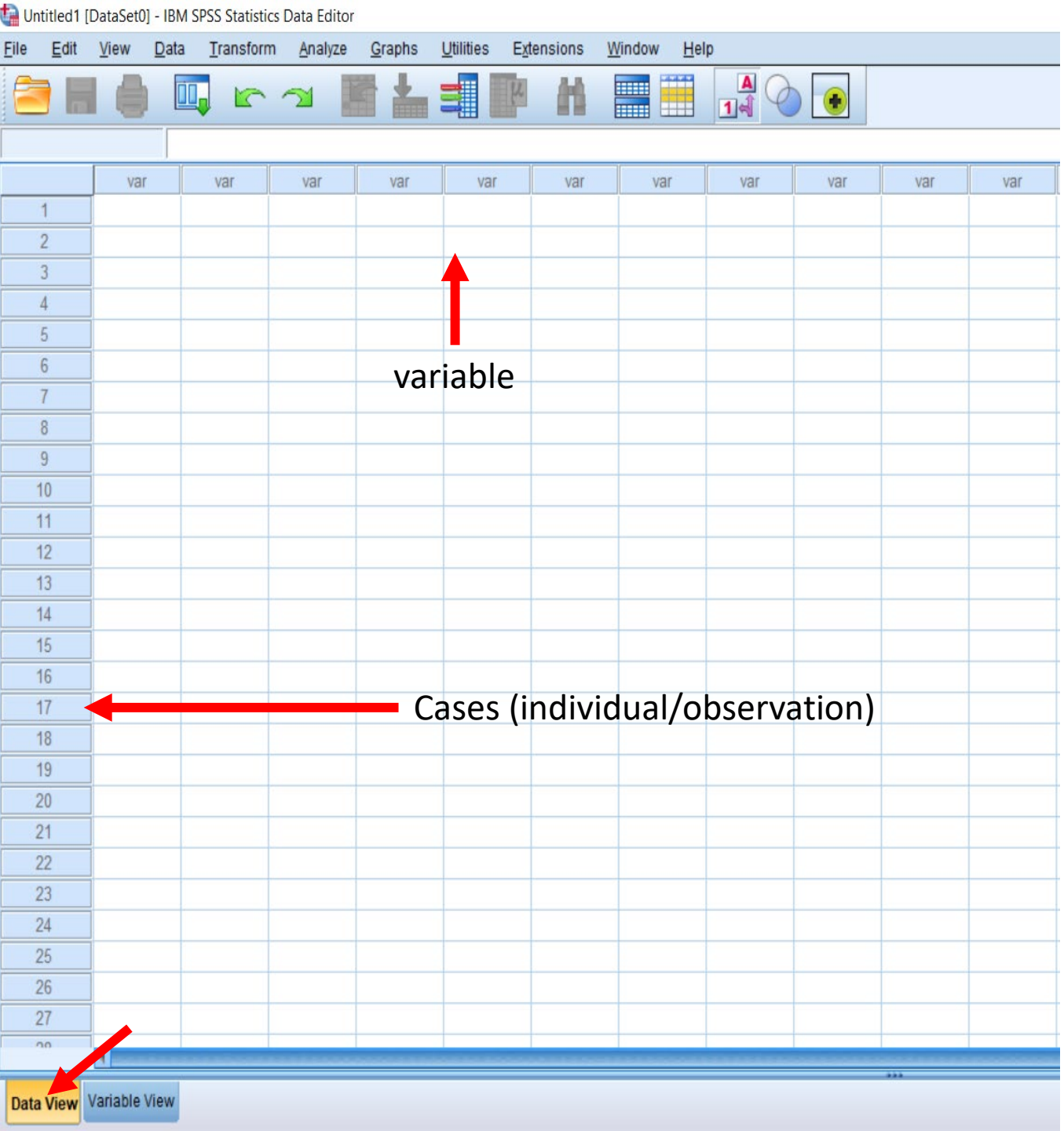


SNEAK PEEK TO SPSS

SPSS: Why?

SPSS: Statistical Package for Social Sciences

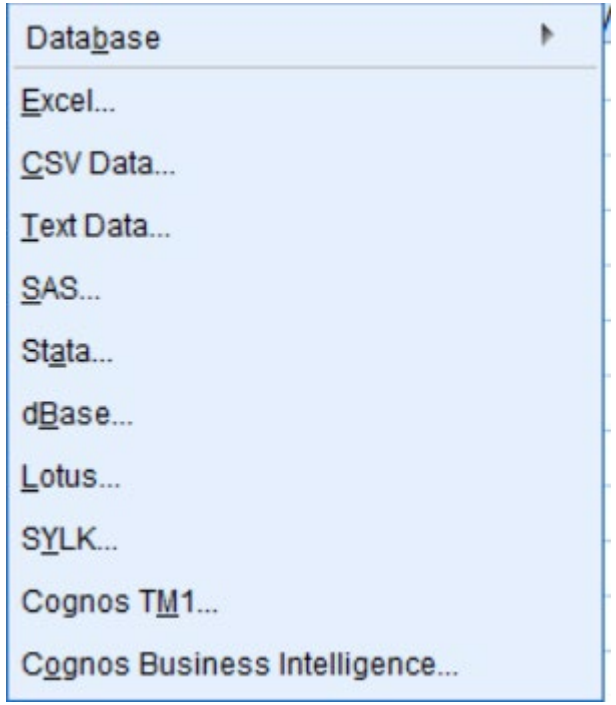
It is a comprehensive and flexible statistical analysis and data management tool. It is one of the most popular statistical package which can perform highly complex data manipulation and analysis with ease. It is designed for both interactive and non interactive users.

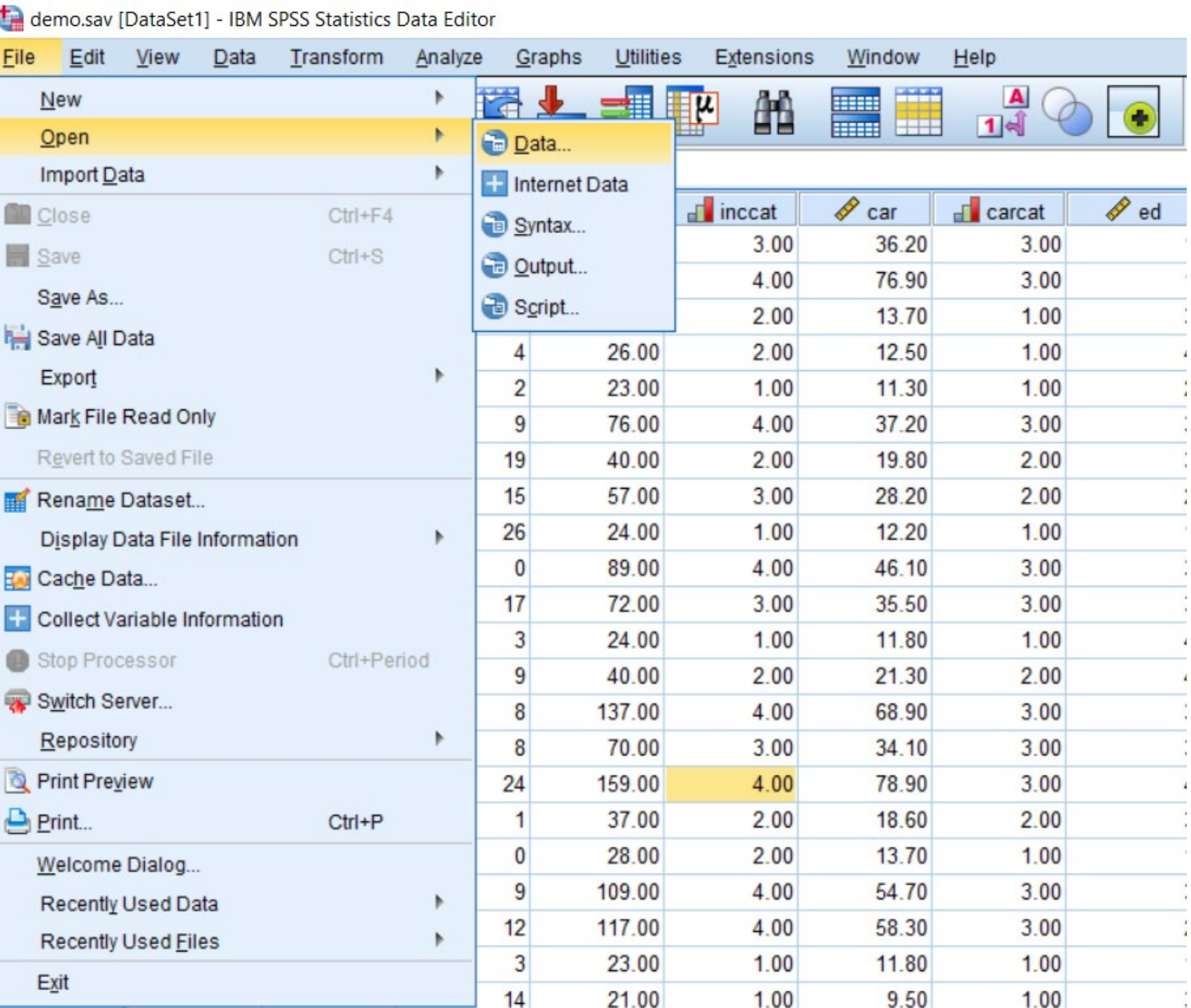


The SPSS Consists of 2 Sheets – One Is the Data View and the Variable View

The Data View is a Spreadsheet which contains rows and columns. The data can be entered in the Data View Sheet either manually or the data can be imported from the data file (Excel, Plain text files or relational (SQL) Databases).

File > Import Data >





File > Open > Data...

A dialog box for opening files is displayed.

By default, IBM® SPSS® Statistics data files (*.sav* extension) are displayed.

Open file: *demo.sav*

demo.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

8 : age 35

	age	marital	address	income	inccat	car	carcat	ed	employ	retire	empcat	jobsat	gender
1	55	1	12	72.00	3.00	36.20	3.00	1	23	0	3	5	f
2	56	0	29	153.00	4.00	76.90	3.00	1	35	0	3	4	m
3	28	1	9	28.00	2.00	13.70	1.00	3	4	0	1	3	f
4	24	1	4	26.00	2.00	12.50	1.00	4	0	0	1	1	m
5	25	0	2	23.00	1.00	11.30	1.00	2	5	0	2	2	m
6	45	1	9	76.00	4.00	37.20	3.00	3	13	0	2	2	m
7	42	0	19	40.00	2.00	19.80	2.00	3	10	0	2	2	m
8	35	0	15	57.00	3.00	28.20	2.00	2	1	0	1	1	f
9	46	0	26	24.00	1.00	12.20	1.00	1	11	0	2	5	f
10	34	1	0	89.00	4.00	46.10	3.00	3	12	0	2	4	m
11	55	1	17	72.00	3.00	35.50	3.00	3	2	0	1	3	f
12	28	0	3	24.00	1.00	11.80	1.00	4	4	0	1	5	m
13	31	1	9	40.00	2.00	21.30	2.00	4	0	0	1	2	f
14	42	0	8	137.00	4.00	68.90	3.00	3	3	0	1	1	f
15	35	0	8	70.00	3.00	34.10	3.00	3	9	0	2	4	m
16	52	1	24	159.00	4.00	78.90	3.00	4	16	0	3	5	m
17	21	1	1	37.00	2.00	18.60	2.00	3	0	0	1	1	m
18	32	0	0	28.00	2.00	13.70	1.00	1	2	0	1	4	f
19	42	0	9	109.00	4.00	54.70	3.00	3	20	0	3	3	f
20	40	1	12	117.00	4.00	58.30	3.00	2	19	0	3	5	f
21	30	0	3	23.00	1.00	11.80	1.00	1	3	0	1	3	m
22	48	0	14	21.00	1.00	9.50	1.00	3	2	0	1	3	m
23	39	1	17	17.00	1.00	8.50	1.00	4	2	0	1	3	m
24	42	1	5	34.00	2.00	16.60	2.00	2	13	0	2	3	f
25	45	1	12	115.00	4.00	57.40	3.00	1	27	0	3	4	f
26	51	1	10	47.00	2.00	23.00	2.00	1	9	0	2	3	m
27	39	1	9	33.00	2.00	16.30	2.00	3	1	0	1	1	m

Alternatively, you can use the Value Labels button for different view of your data set

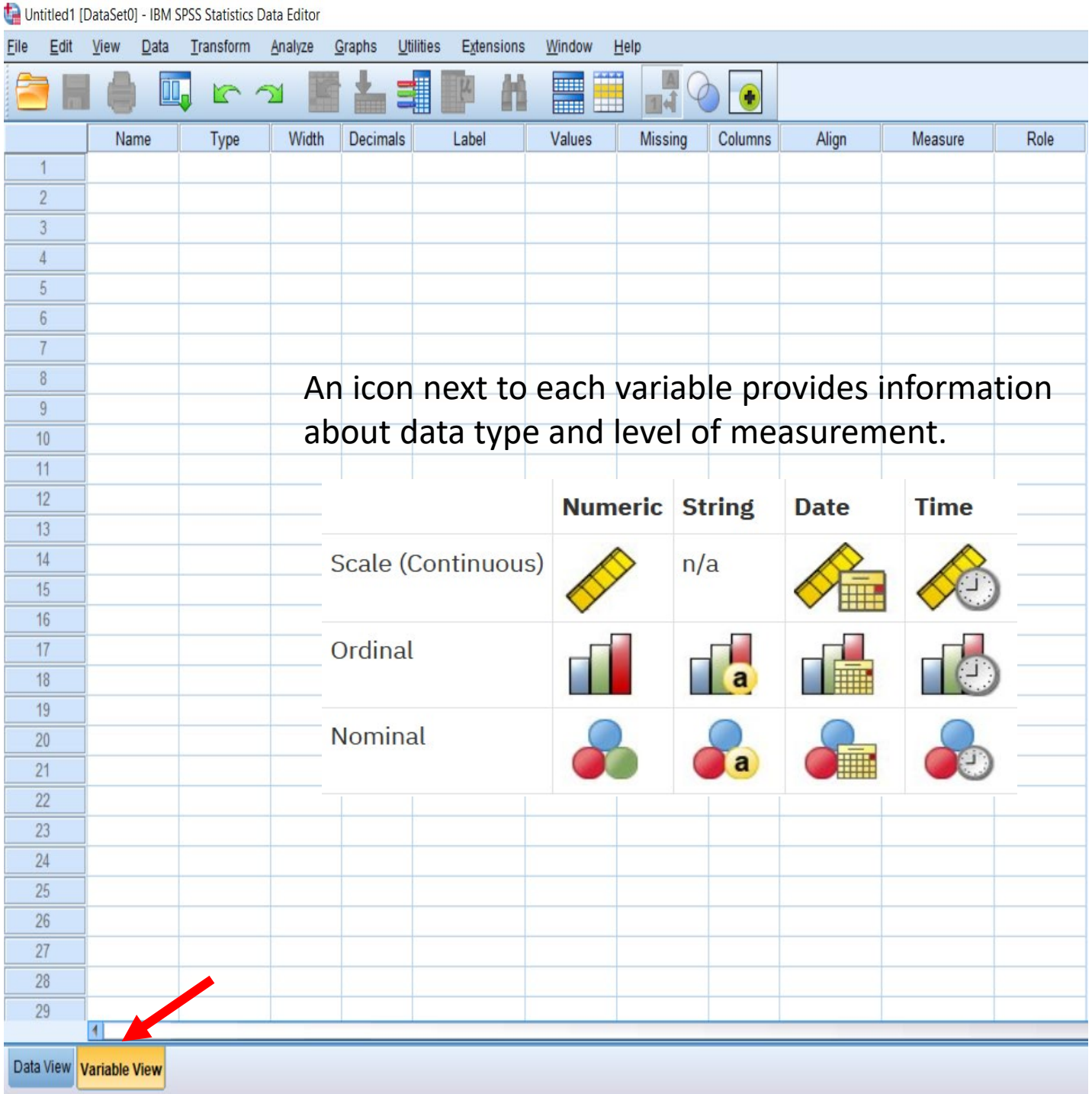
View > Value Labels (tick)

demo.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

8 : age 35

	age	marital	address	income	inccat	car	carcat	ed	employ	retire	empcat	jobsat	gender
1	55	Married	12	72.00	$50 - $74	36.20	Luxury	Did not co...	23	No	More than 15	Highly sati...	Female
2	56	Unmarried	29	153.00	$75+	76.90	Luxury	Did not co...	35	No	More than 15	Somewhat ...	Male
3	28	Married	9	28.00	$25 - $49	13.70	Economy	Some colle...	4	No	Less than 5	Neutral	Female
4	24	Married	4	26.00	$25 - $49	12.50	Economy	College de...	0	No	Less than 5	Highly diss...	Male
5	25	Unmarried	2	23.00	Under $25	11.30	Economy	High schoo...	5	No	5 to 15	Somewhat ...	Male
6	45	Married	9	76.00	$75+	37.20	Luxury	Some colle...	13	No	5 to 15	Somewhat ...	Male
7	42	Unmarried	19	40.00	$25 - $49	19.80	Standard	Some colle...	10	No	5 to 15	Somewhat ...	Male
8	35	Unmarried	15	57.00	$50 - $74	28.20	Standard	High schoo...	1	No	Less than 5	Highly diss...	Female
9	46	Unmarried	26	24.00	Under $25	12.20	Economy	Did not co...	11	No	5 to 15	Highly sati...	Female
10	34	Married	0	89.00	$75+	46.10	Luxury	Some colle...	12	No	5 to 15	Somewhat ...	Male
11	55	Married	17	72.00	$50 - $74	35.50	Luxury	Some colle...	2	No	Less than 5	Neutral	Female
12	28	Unmarried	3	24.00	Under $25	11.80	Economy	College de...	4	No	Less than 5	Highly sati...	Male
13	31	Married	9	40.00	$25 - $49	21.30	Standard	College de...	0	No	Less than 5	Somewhat ...	Female
14	42	Unmarried	8	137.00	$75+	68.90	Luxury	Some colle...	3	No	Less than 5	Highly diss...	Female
15	35	Unmarried	8	70.00	$50 - $74	34.10	Luxury	Some colle...	9	No	5 to 15	Somewhat ...	Male
16	52	Married	24	159.00	$75+	78.90	Luxury	College de...	16	No	More than 15	Highly sati...	Male
17	21	Married	1	37.00	$25 - $49	18.60	Standard	Some colle...	0	No	Less than 5	Highly diss...	Male
18	32	Unmarried	0	28.00	$25 - $49	13.70	Economy	Did not co...	2	No	Less than 5	Somewhat ...	Female
19	42	Unmarried	9	109.00	$75+	54.70	Luxury	Some colle...	20	No	More than 15	Neutral	Female
20	40	Married	12	117.00	$75+	58.30	Luxury	High schoo...	19	No	More than 15	Highly sati...	Female
21	30	Unmarried	3	23.00	Under $25	11.80	Economy	Did not co...	3	No	Less than 5	Neutral	Male
22	48	Unmarried	14	21.00	Under $25	9.50	Economy	Some colle...	2	No	Less than 5	Neutral	Male
23	39	Married	17	17.00	Under $25	8.50	Economy	College de...	2	No	Less than 5	Neutral	Male



•**Name**: Enter the Unique Identifiable

•**Type:** Can change the Type of Variable (Numeric, Alphabets or Alpha Numeric by selecting the respective type in this column)

•**Width** : Defines the character width this variable should allow (helpful while entering mobile number which allow only 10 character)

•**Decimal:** Defines the Decimal point you required to display,

•**Label:** Can give any name as an Label for that Variable you wanted to assign

•**Value** : This is to define/ Label a Value Wherever you see in the Data, Eg: You can Label “0” in the data as ABSENT for Exam. This helps in reading the data better in data view

•**Missing :** Can mention the Data which you don’t want the SPSS to consider while analyzing, Like “0” Value is considered as Absent, so for analysis it will neglect “0” if its mentioned in Missing, which will be helpful in Mean, Mode Etc,

•**Align** : Can mention the alignment of the data in the data sheet, Left, Right of Middle,

•**Measure:** Define the measure of the Variable that you have entered, Whether Scale, Ordinal or Nominal type of Variable

*demo.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	age	Numeric	4	0	Age in years	None	None	8	Right	Scale	Input
2	marital	Numeric	4	0	Marital status	{0, Unmarrie...	None	8	Right	Nominal	Input
3	address	Numeric	4	0	Years at curren...	None	None	8	Right	Scale	Input
4	income	Numeric	8	2	Household inco...	None	None	8	Right	Scale	Input
5	inccat	Numeric	8	2	Income categor...	{1.00, Under...	None	8	Right	Ordinal	Input
6	car	Numeric	8	2	Price of primary...	None	None	8	Right	Scale	Input
7	carcat	Numeric	8	2	Primary vehicle...	{1.00, Econ...	None	8	Right	Ordinal	Input
8	ed	Numeric	4	0	Level of education	{1, Did not c...	None	8	Right	Ordinal	Input
9	employ	Numeric	4	0	Years with curr...	None	None	8	Right	Scale	Input
10	retire	Numeric	4	0	Retired	{0, No}...	None	8	Right	Nominal	Input
11	empcat	Numeric	4	0	Years with curr...	{1, Less tha...	None	8	Right	Ordinal	Input
12	jobsat	Numeric	4	0	Job satisfaction	{1, Highly di...	None	8	Right	Ordinal	Input
13	gender	String	1	0	Gender	{f, Female}...	None	8	Left	Nominal	Input
14	reside	Numeric	4	0	Number of peop...	None	None	8	Right	Scale	Input
15	wireless	Numeric	4	0	Wireless service	{0, No}...	None	8	Right	Nominal	Input
16	multline	Numeric	4	0	Multiple lines	{0, No}...	None	8	Right	Nominal	Input
17	voice	Numeric	4	0	Voice mail	{0, No}...	None	8	Right	Nominal	Input
18	pager	Numeric	4	0	Paging service	{0, No}...	None	8	Right	Nominal	Input
19	internet	Numeric	4	0	Internet	{0, No}...	8, 9	8	Right	Nominal	Input
20	callid	Numeric	4	0	Caller ID	{0, No}...	None	8	Right	Nominal	Input
21	callwait	Numeric	4	0	Call waiting	{0, No}...	None	8	Right	Nominal	Input
22	owntv	Numeric	4	0	Owns TV	{0, No}...	None	8	Right	Nominal	Input
23	ownvcr	Numeric	4	0	Owns VCR	{0, No}...	None	8	Right	Nominal	Input
24	owncd	Numeric	4	0	Owns stereo/C...	{0, No}...	None	8	Right	Nominal	Input
25	ownpda	Numeric	4	0	Owns PDA	{0, No}...	None	8	Right	Nominal	Input
26	ownpc	Numeric	4	0	Owns computer	{0, No}...	None	8	Right	Nominal	Input
27	ownfax	Numeric	4	0	Owns fax mach...	{0, No}...	None	8	Right	Nominal	Input
28	news	Numeric	4	0	Newspaper sub...	{0, Yes}...	None	8	Right	Nominal	Input
29	response	Numeric	4	0	Response	{0, Yes}...	None	8	Right	Nominal	Input

Data View | Variable View

DESCRIPTIVE STATISTICS: EXPLORING AND SUMARIZING THE DATA

ORGANIZE RAW DATA INTO A MEANINGFUL FORM SO THAT WE CAN UNDERSTAND WHAT THE DATA ARE TELLING US.

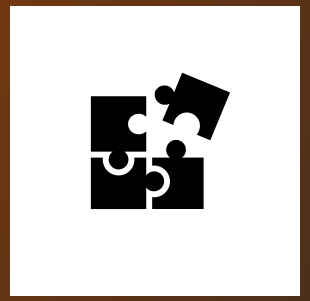
WHAT IS DESCRIPTIVE STATISTICS ?

- To highlight the features and characteristics of a data set by using summary.
- Data is explored and summarized for its:
  i. central tendency: represents the center point or typical value of a dataset.
  ii. Position: location of an individual value (observation) in dataset.
  iii. Variation/Dispersion: variability within the dataset.
  iv. Shape of distribution: distribution (or pattern) within the dataset.

2 ways in exploring and summarizing data in descriptive statistics:

1. Visual: using graphical such as tables, chart, plot and graph.
2. Numerical: using formula.

WHAT WILL YOU LEARN?

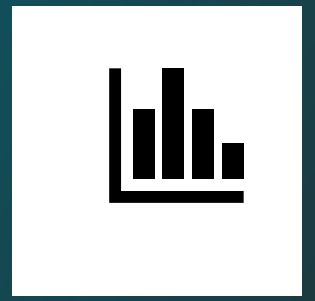


Visual Way for Descriptive Statistics

- Qualitative Data
- Quantitative Data

Numerical Way: Central Tendency & Data Position

- Qualitative Data
- Quantitative Data



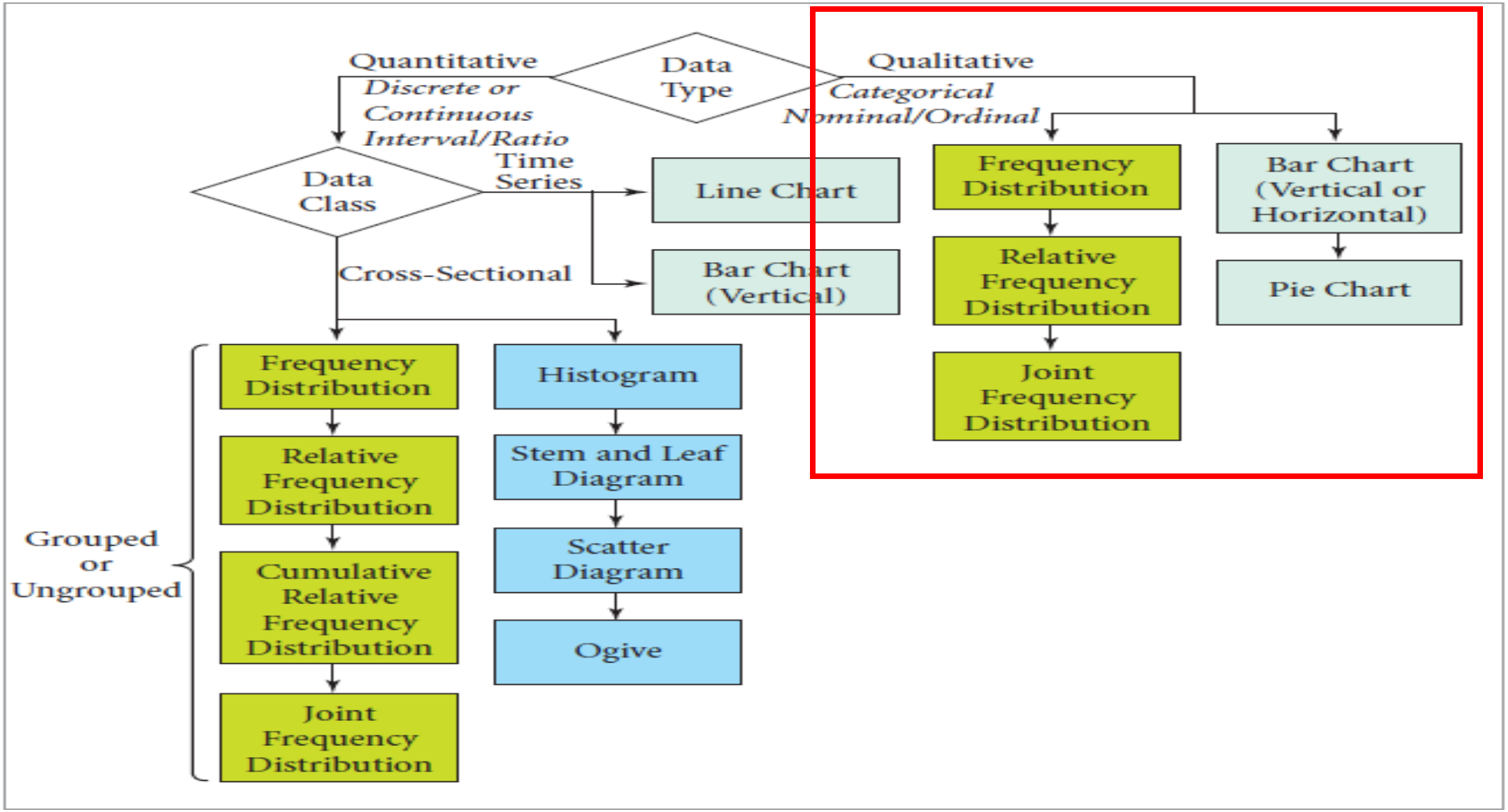
Numerical Way: Data Variation & Shape of Distribution

Quantitative Data

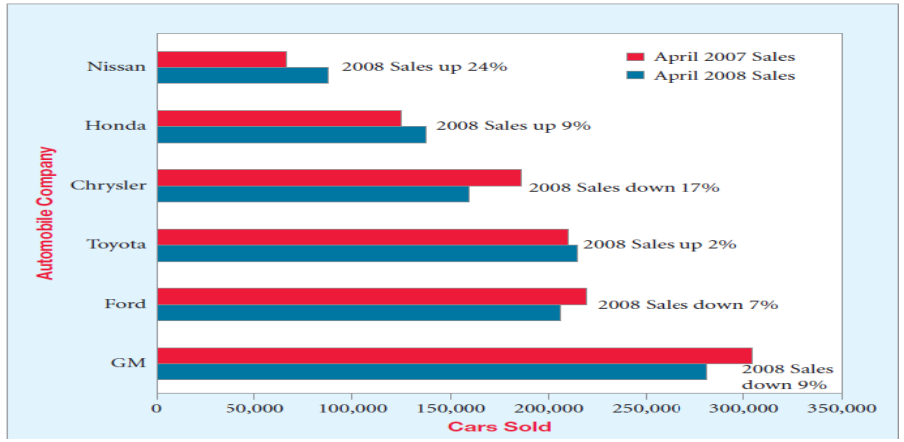
First Topic | Second Topic | Third Topic

VISUAL WAY FOR DESCRIPTIVE STATISTICS: QUALITATIVE VARIABLE

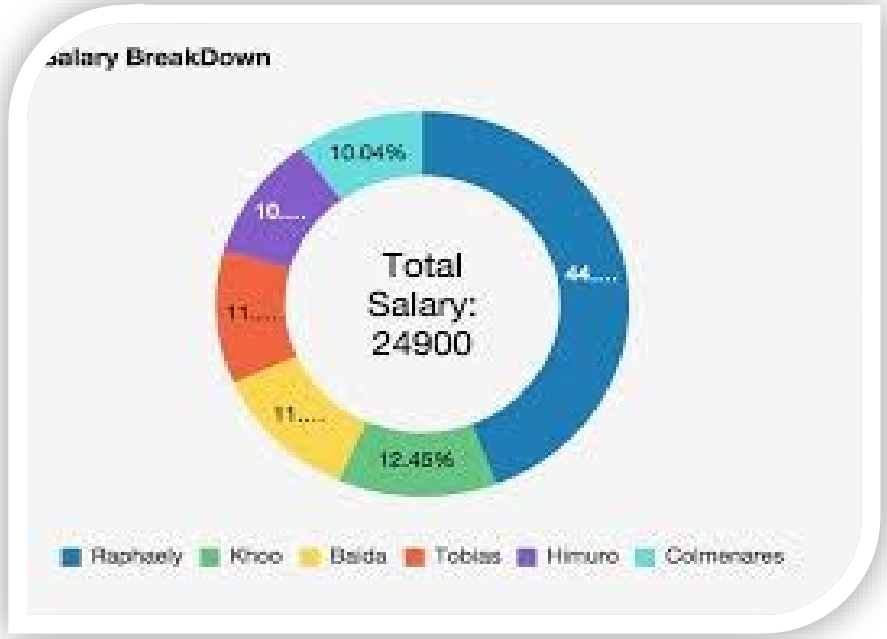
DESCRIPTIVE STATISTICS IN VISUAL WAY



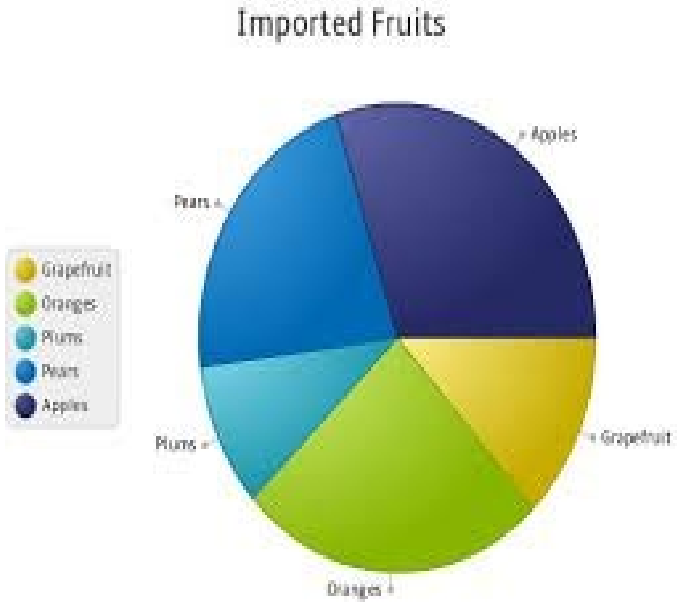
BAR CHART



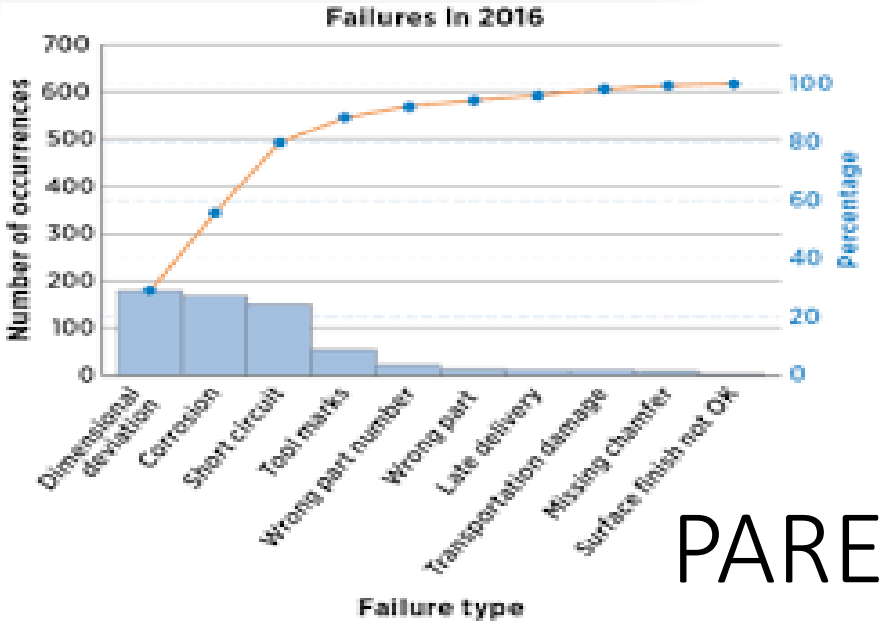
DOUGHNUT CHART



PIE CHART



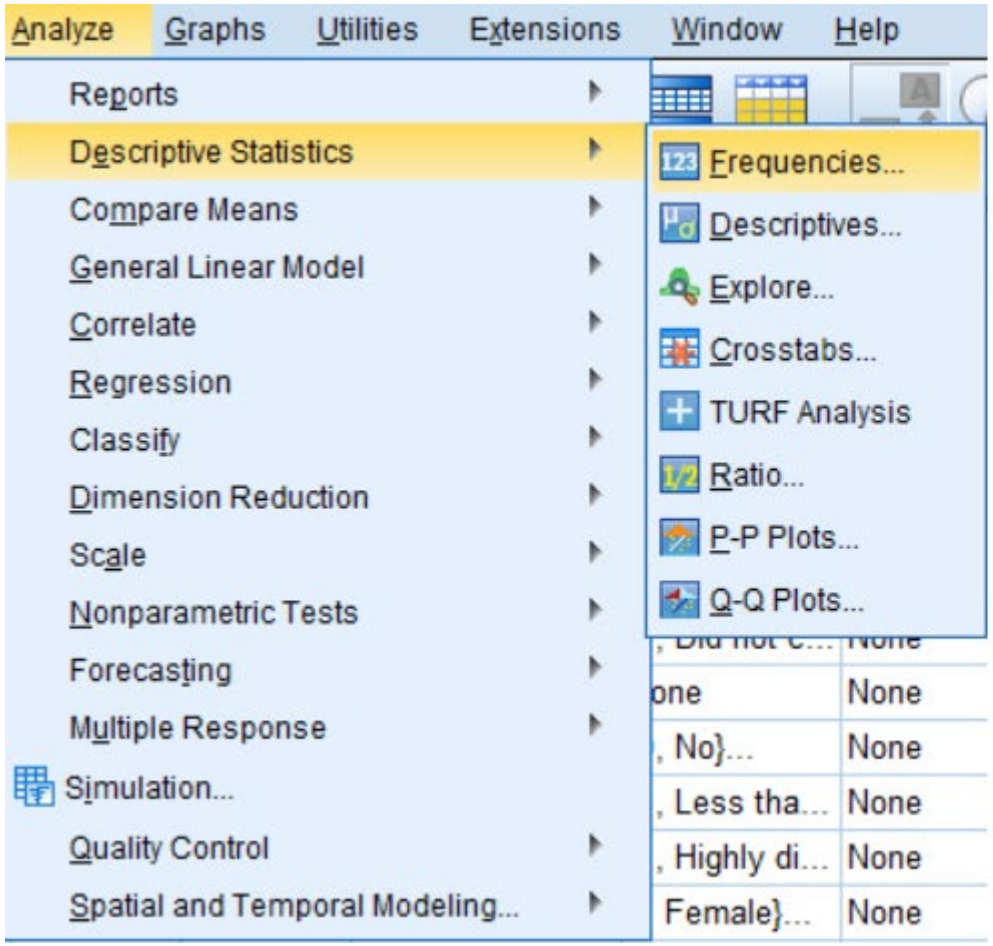
PARETO CHART

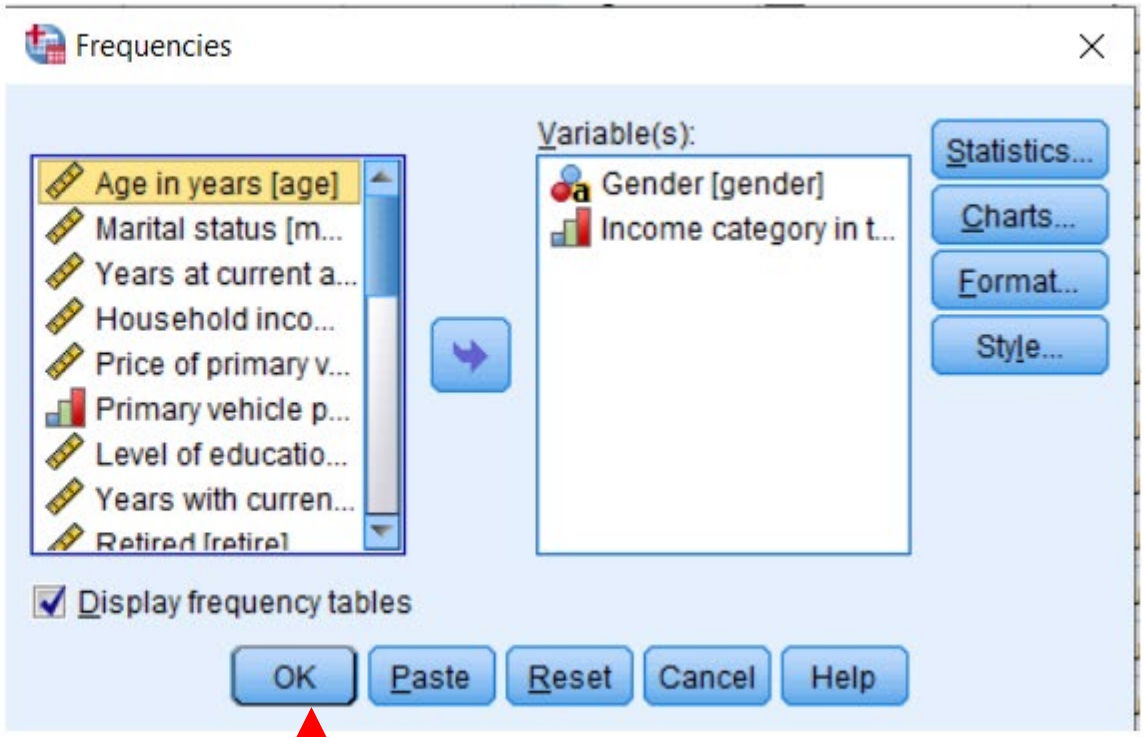


Visual Way for Descriptive Statistics: Qualitative Variable
- Frequency Table for Gender and Income Category

Dataset: demo.sav

Analyze > Descriptive Statistics > Frequencies...





Click OK to run the procedure. Results are displayed in the Viewer window.

Statistics

		Gender	Income category in thousands
N	Valid	6400	6400
	Missing	0	0

Frequency Table

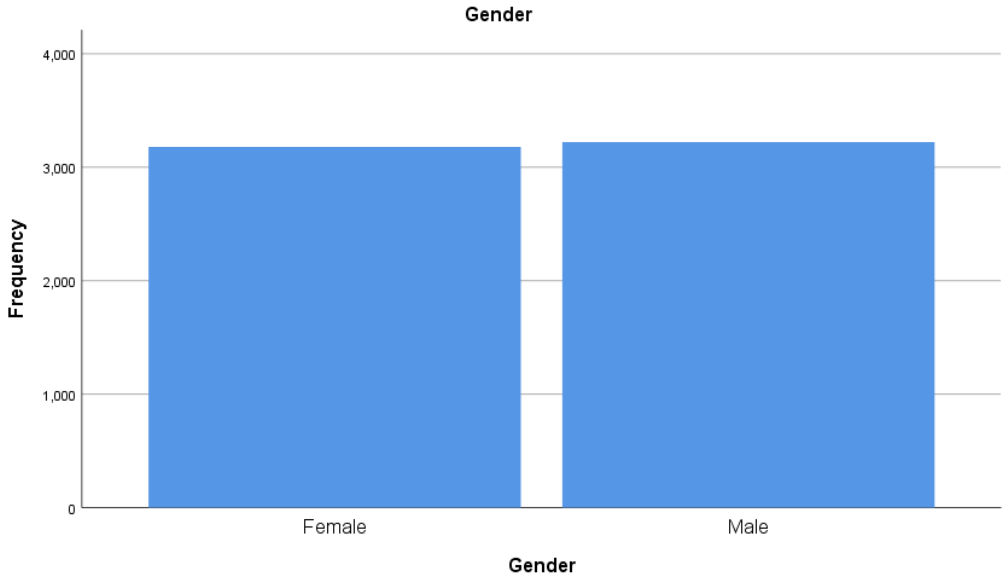
Gender

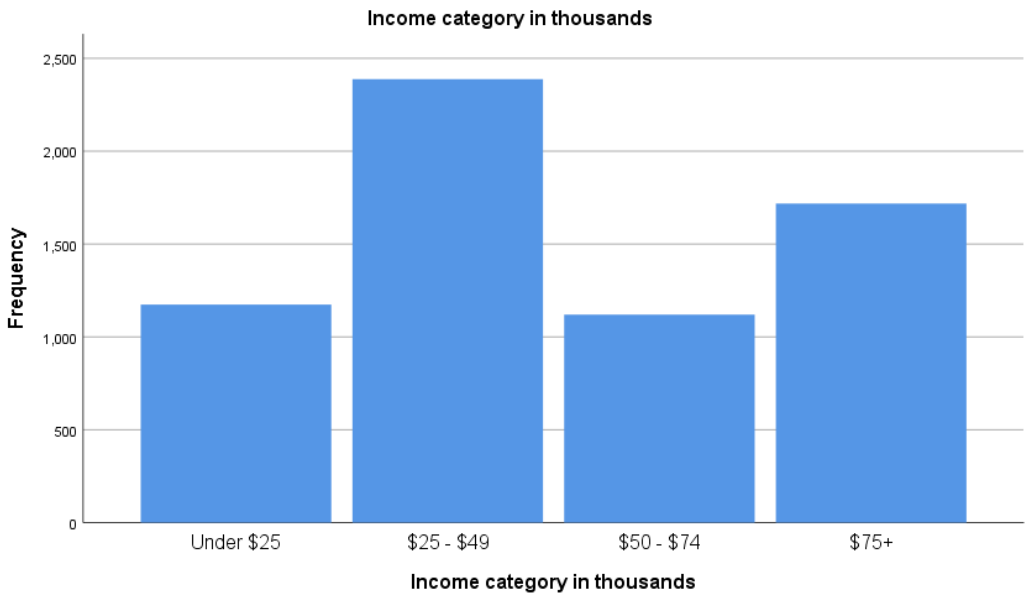
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	3179	49.7	49.7	49.7
	Male	3221	50.3	50.3	100.0
	Total	6400	100.0	100.0	

Income category in thousands

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Under $25	1174	18.3	18.3	18.3
	$25 - $49	2388	37.3	37.3	55.7
	$50 - $74	1120	17.5	17.5	73.2
	$75+	1718	26.8	26.8	100.0
	Total	6400	100.0	100.0	

Bar Chart

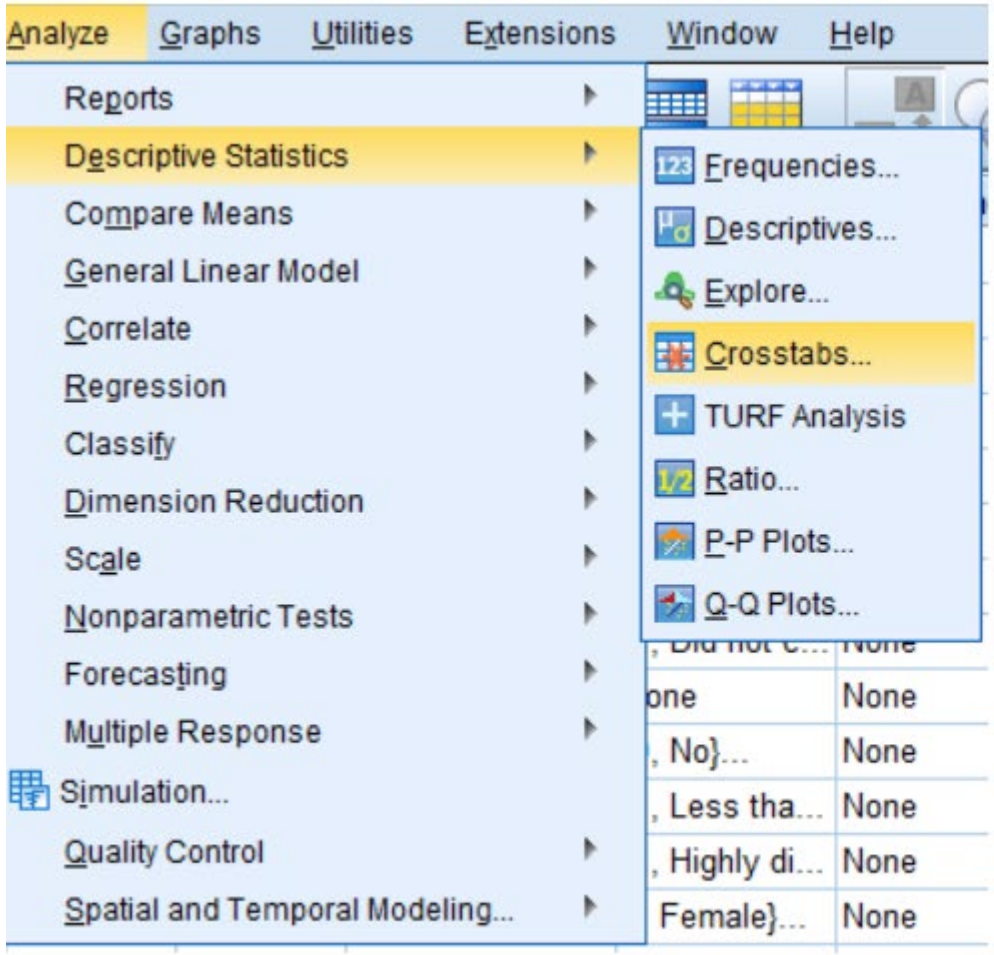


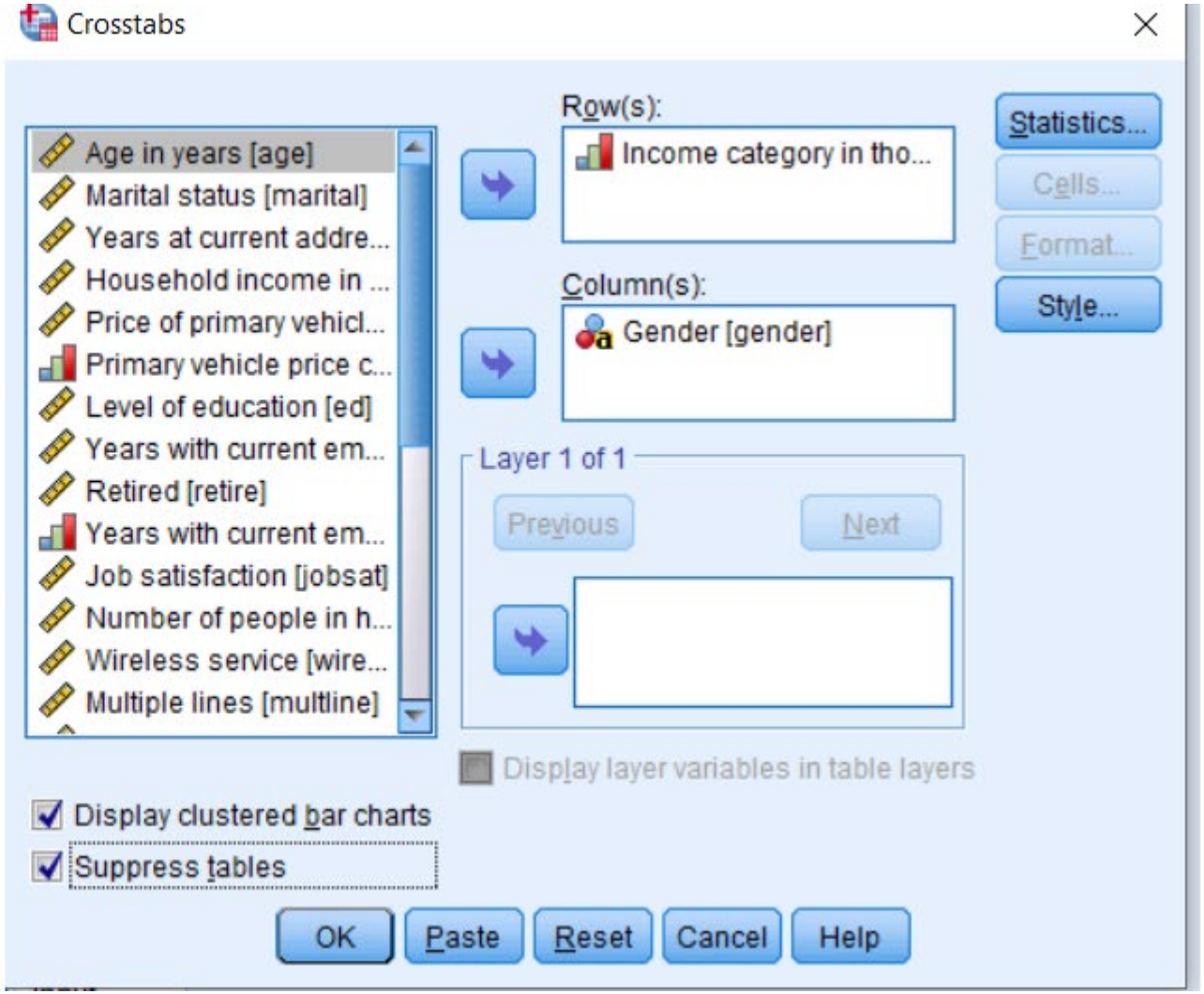


Visual Way for Descriptive Statistics: Qualitative Variable
- Crosstabulation for Gender and Income Category

Dataset: demo.sav

Analyze > Descriptive Statistics > Crosstabs...

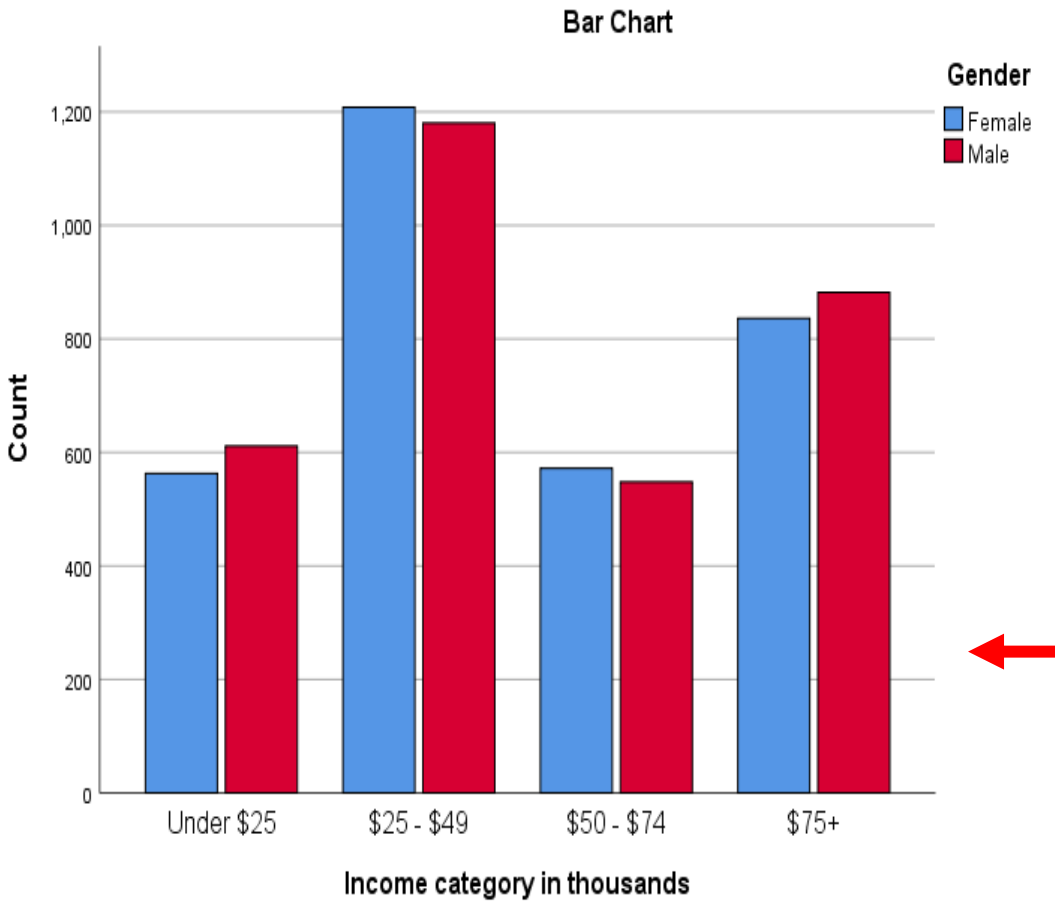




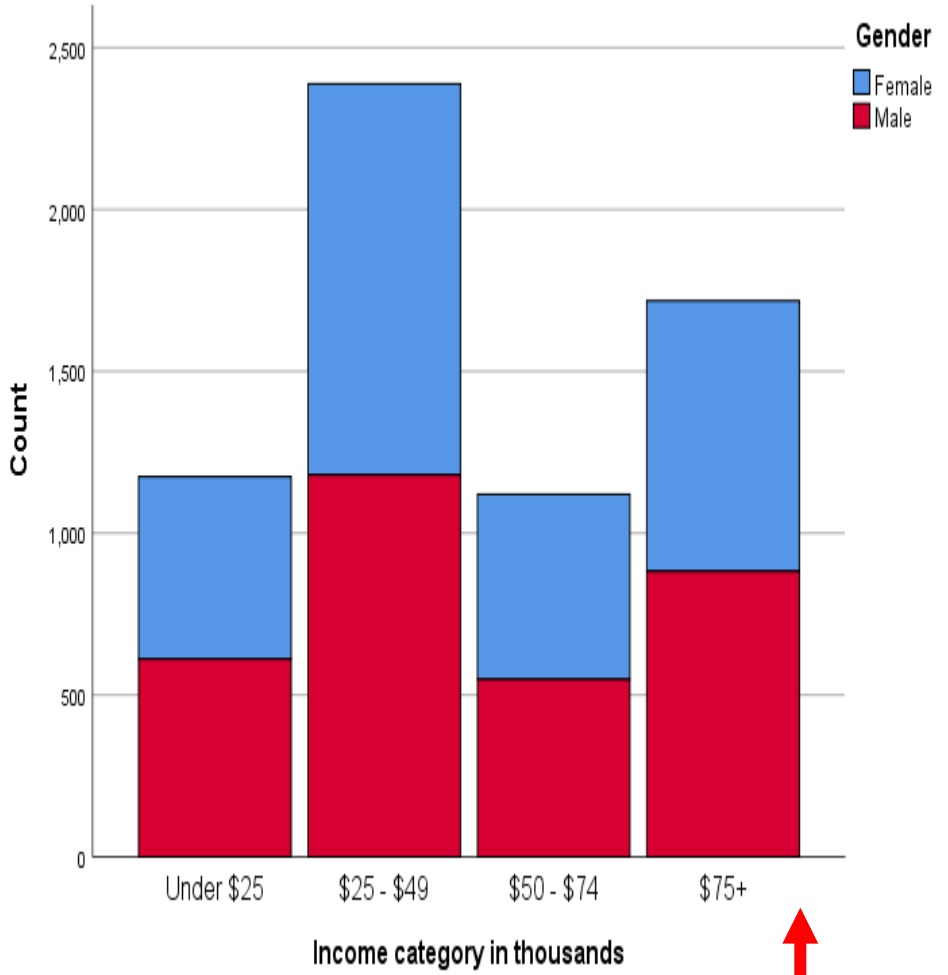
Income category in thousands * Gender Crosstabulation

Count

		Gender		
		Female	Male	Total
Income category in thousands	Under $25	563	611	1174
	$25 - $49	1208	1180	2388
	$50 - $74	572	548	1120
	$75+	836	882	1718
Total		3179	3221	6400



Clustered Bar Chart



Stacked Bar Chart

VISUAL DESCRIPTIVE STATISTICS FOR QUALITATIVE VARIABLE

i. central tendency: the greater the count (frequency)
ii. Position: maximum and minimum value
iii. Variation/Dispersion: variability of count for each category in a variable

Activities:
Data set: Prescription and Medication Pattern Analysis

Purpose of Study: Analyze which medications are commonly prescribed together.

What data type it is? Categorical (drug names, ATC codes).

Problem: No clear understanding of drug prescription trends and potential overuse.

How to solve?:

1. Descriptive Statistics: TABLE (Frequency and Relative Frequency Distribution)
2. Inferential Statistics: Identify **high-confidence association rules**, e.g.: *Patients taking Metformin often take Insulin.*

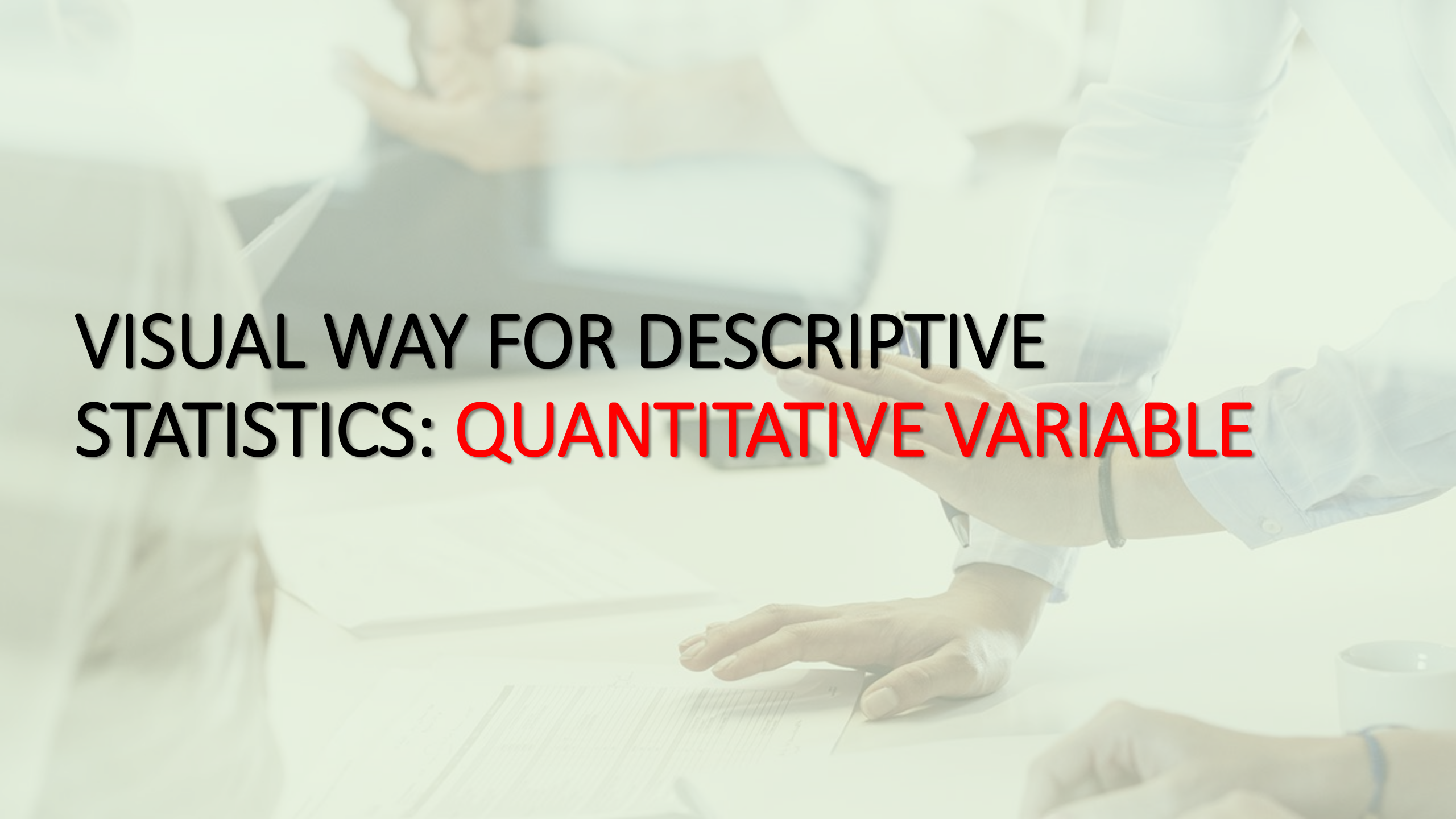
4	2	5	8	8	10	1	4	8	3	4	1	1	3	4
1	4	4	5	4	4	4	9	5	4	4	10	7	11	4
10	2	6	7	10	5	4	6	4	6	2	3	2	4	5
5	4	11	1	4	1	9	2	4	6	6	7	6	2	3
6	5	3	4	5	6	5	3	10	6	5	7	7	4	3
8	2	2	6	5	11	9	9	5	5	6	5	3	1	7
6	6	5	3	8	4	3	3	4	4	4	7	6	4	9
1	6	5	5	4	4	7	5	6	6	9	5	6	10	4
7	5	8	4	4	7	4	6	6	4	4	2	10	4	5
4	11	8	7	9	5	6	4	2	8	4	2	6	6	6
6	4	6	5	7	1	6	9	1	5	9	10	5	5	10
5	4	7	5	7	6	9	5	3	2	1	5	5	5	5
5	9	5	3	2	5	7	2	4	6	4	4	4	4	4
6	5	8	5	5	5	5	5	2	5	5	6	4	6	5
5	7	10	2	2	6	8	3	1	3	5	6	3	3	6
5	4	5	3	3	7	9	4	4	5	10	6	10	5	9
4	3	8	7	1	8	4	3	1	3	6	7	5	5	5
4	7	4	11	6	6	3	7	9	4	4	2	9	7	5
1	6	6	8	3	8	4	4	1	9	3	9	3	4	2
9	5	5	7	10	5	3	4	7	7	6	2	2	4	4
4	7	3	5	4	9	2	3	4	3	2	1	6	4	6
1	8	1	4	3	5	5	10	4	4	4	6	9	2	7
9	4	5	3	6	5	5	3	4	6	5	7	3	6	8
3	6	1	5	7	7	5	4	6	6	6	3	6	9	5
4	5	10	1	5	5	7	8	9	1	6	5	6	6	4
10	6	5	5	5	1	6	5	6	4	7	9	10	2	6
4	4	6	11	9	5	4	4	3	5	4	6	2	6	7
3	5	6	7	4	5	4	6	9	4	3	3	6	9	4
3	7	5	6	11	4	4	8	4	2	8	2	4	2	3
6	5	1	10	5	9	5	4	5	1	4	9	5	4	4

Data set: Prescription and Medication Pattern Analysis → Frequency Table

Table: Market Basket Analysis of item purchased at ABX Pharmacy.

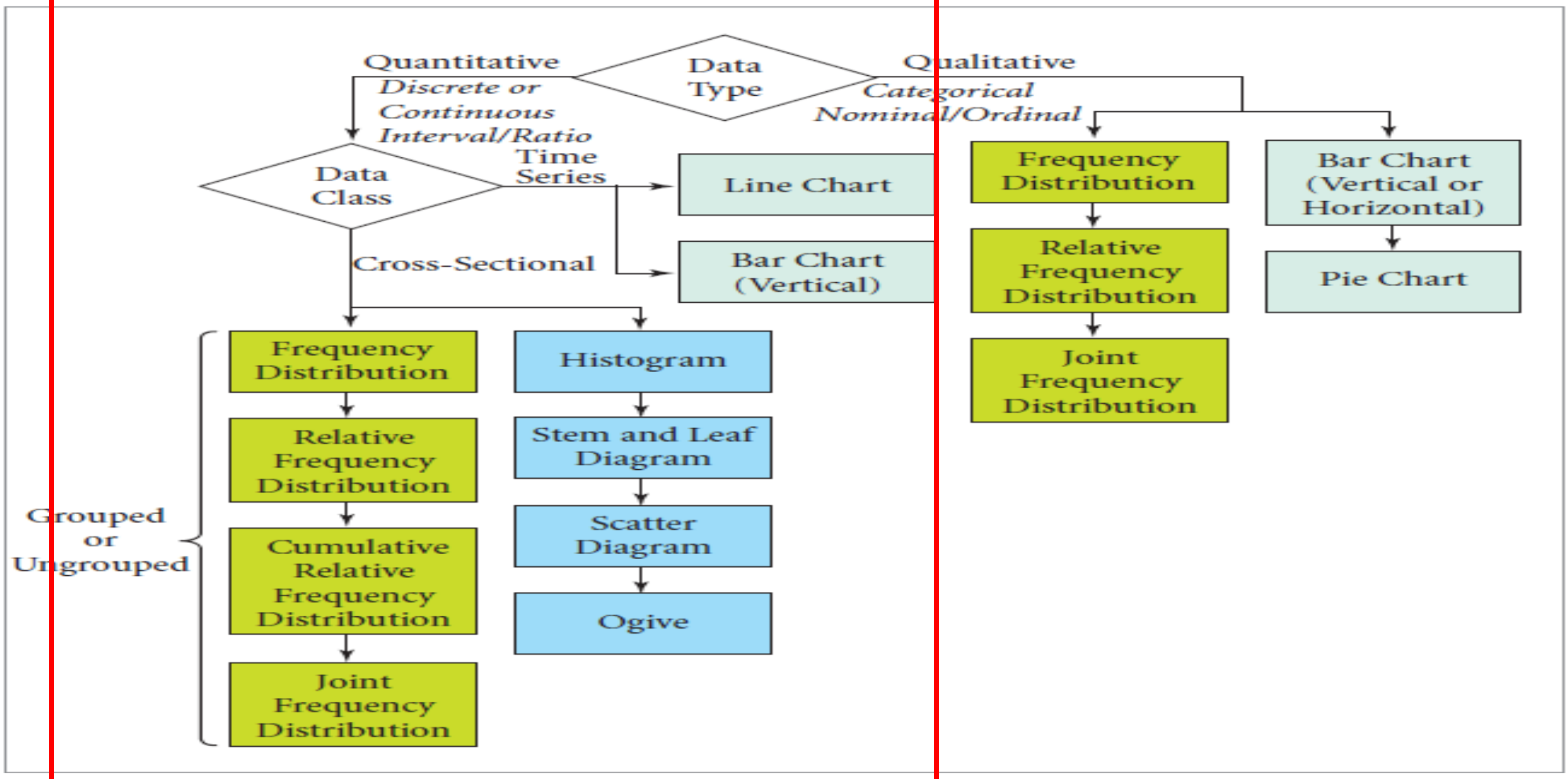
How to construct Relative Frequency Table?

Number of Product Catagories	Frequency
1	25
2	29
3	42
4	92
5	83
6	71
7	35
8	19
9	29
10	18
11	7
	Total = 450



VISUAL WAY FOR DESCRIPTIVE STATISTICS: QUANTITATIVE VARIABLE

DESCRIPTIVE STATISTICS IN VISUAL WAY



PROBLEM: EMERGENCY ROOM PATIENT ARRIVALS

"A hospital administrator wants to determine the typical number of patients who arrive at the emergency department during peak hours. The data in the table below represent the number of patients arriving at the ER during 40 randomly selected 15-minute intervals. For example, in one 15-minute interval, seven patients arrived.".

7	6	6	6	4	6	2	6
5	6	6	11	4	5	7	6
2	7	1	2	4	8	2	6
6	5	5	3	7	5	4	6
2	2	9	7	5	9	8	5

TABLE: Number of Patient Arrivals at the Emergency Room

With the data given, it is possible to use frequency table to describe the data set?

SOLUTION

Number of Customers	Tally	Frequency	Relative Frequency
1	|	1	$\frac{1}{40} = 0.025$
2	𝍸 |	6	0.15
3	|	1	0.025
4	||||	4	0.1
5	𝍸 ||	7	0.175
6	𝍸 𝍸 |	11	0.275
7	𝍸	5	0.125
8	||	2	0.05
9	||	2	0.05
10		0	0.0
11	|	1	0.025

CONSIDER THIS DATA SET. IT IS REASONABLE TO USE FREQUENCY TABLE WITH THIS CONTINUOUS DATA?

PROBLEM: EMERGENCY RESPONSE COMMUNICATION LINKS

"One of the major efforts of the Homeland Security has been to improve the communication between emergency responders, like the police and fire departments. The communications have been hampered by problems involving linking divergent radio and computer systems, as well as communication protocols. While most cities have recognized the problem and made efforts to solve it, Homeland Security recently funded practice exercises in 72 cities of different sizes throughout the United States. The resulting data, already sorted but representing seconds before the systems were linked."

What can you conclude how many seconds which most cities took to link their communications systems?.

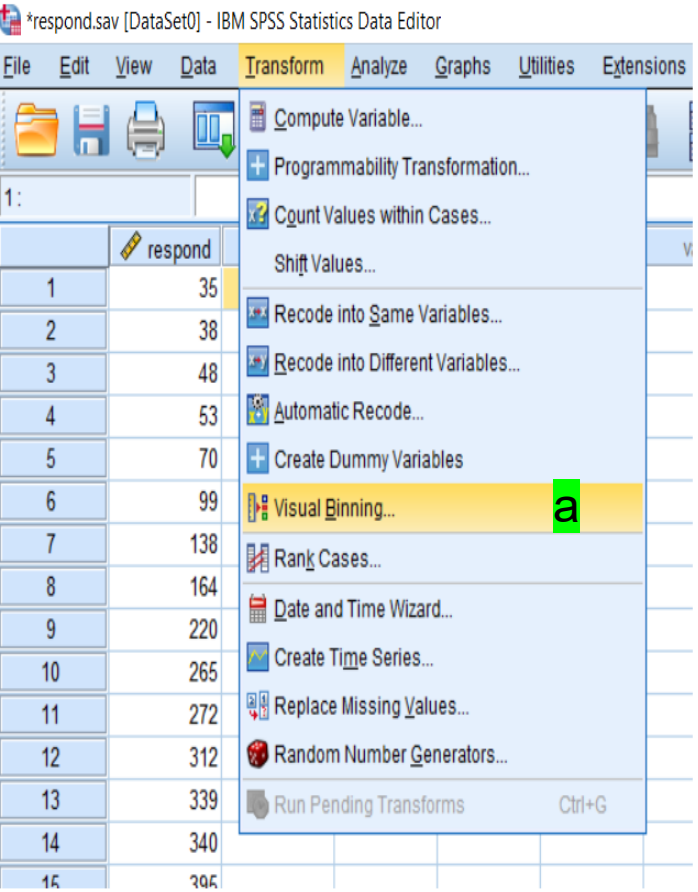
35	339	650	864	1,025	1,261
38	340	655	883	1,028	1,280
48	395	669	883	1,036	1,290
53	457	703	890	1,044	1,312
70	478	730	934	1,087	1,341
99	501	763	951	1,091	1,355
138	521	788	969	1,126	1,357
164	556	789	985	1,176	1,360
220	583	789	993	1,199	1,414
265	595	802	997	1,199	1,436
272	596	822	999	1,237	1,479
312	604	851	1,018	1,242	1,492

TIPS:
USE GROUP/JOIN FREQUENCY TABLE FOR CONTINUOUS DATA

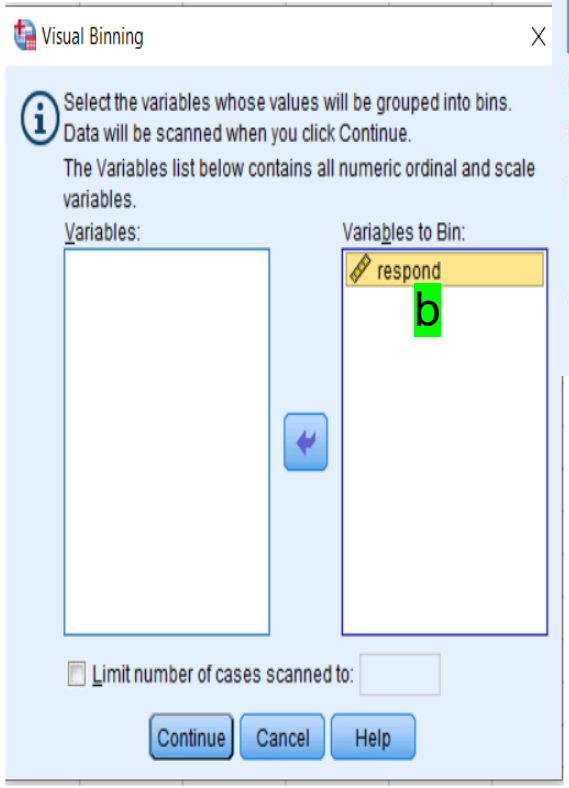
Visual Way for Descriptive Statistics: Quantitative Variable
- Group/Join Frequency Table

Dataset: respond.sav

Transform > Visual Binning..>Make cutpoints



Select variable

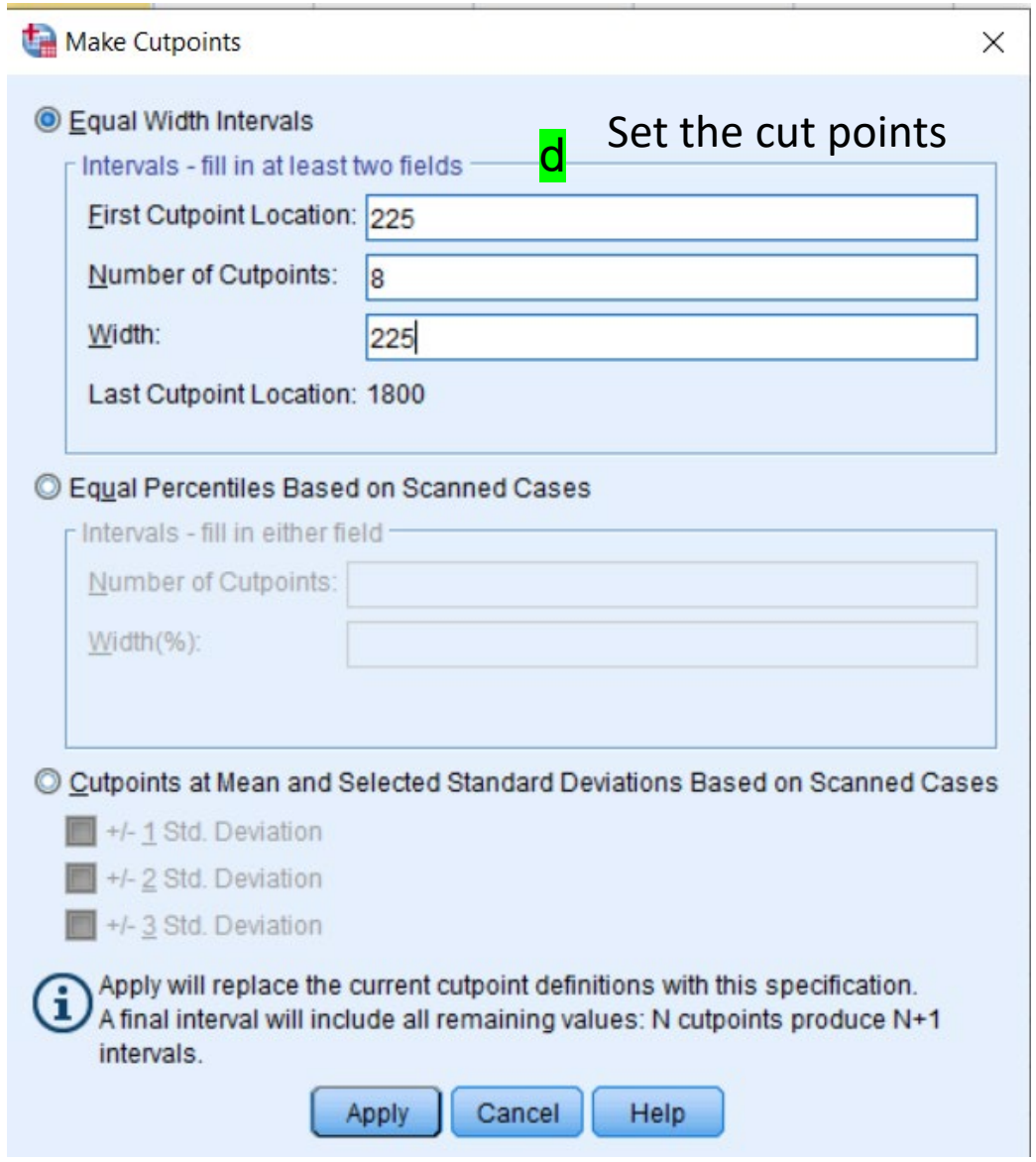


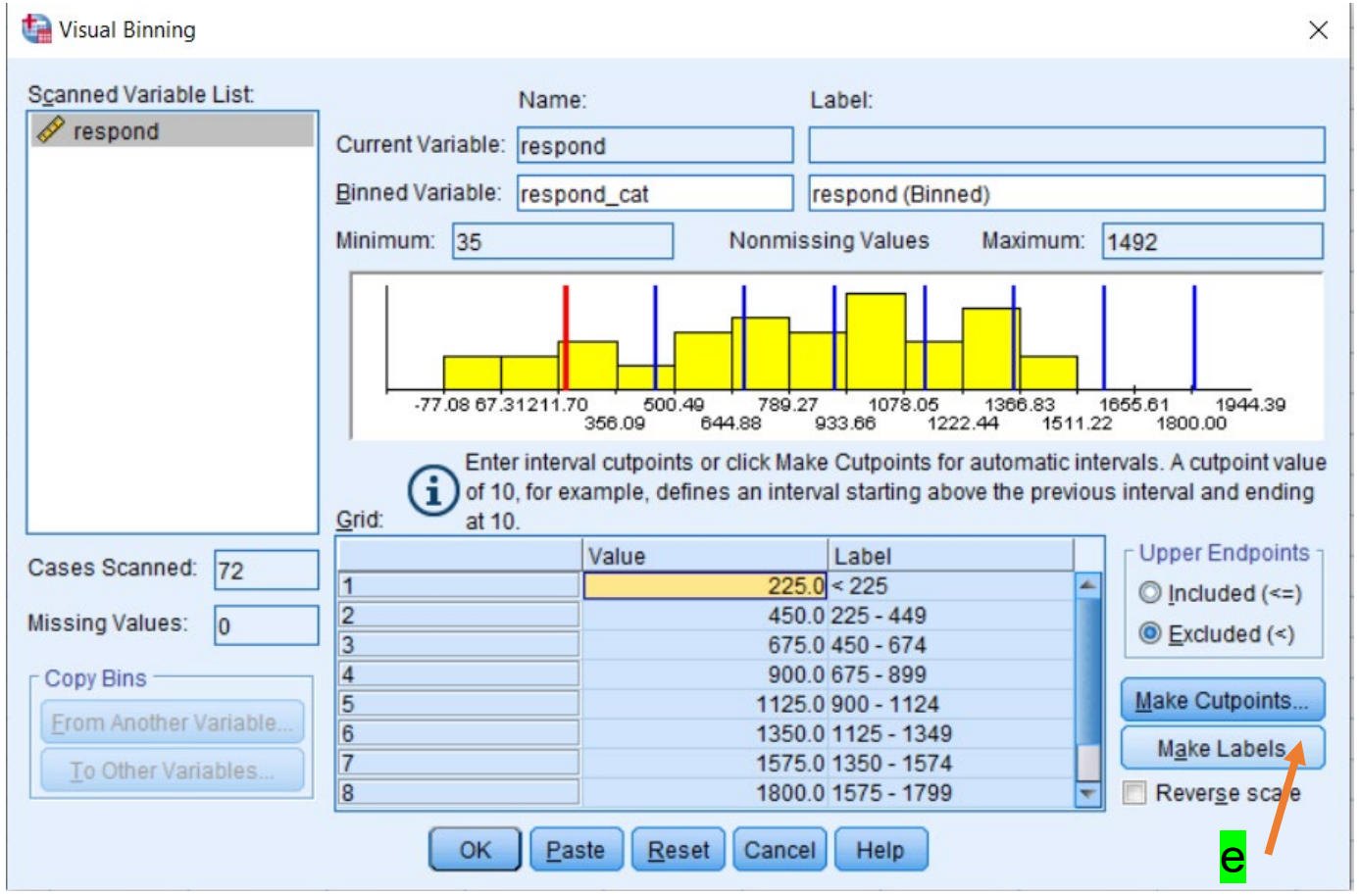
Choose endpoints and make cut points

Visual Way for Descriptive Statistics: Quantitative Variable
- Group/Join Frequency Table

Dataset: respond.sav

Transform > Visual Binning..>Make cutpoints





Make the labels for each category of respond

Visual Way for Descriptive Statistics: Quantitative Variable - Group/Join Frequency Table

Dataset: respond.sav

*respond.sav [DataSet0] - IBM SPSS Statistics Data Editor

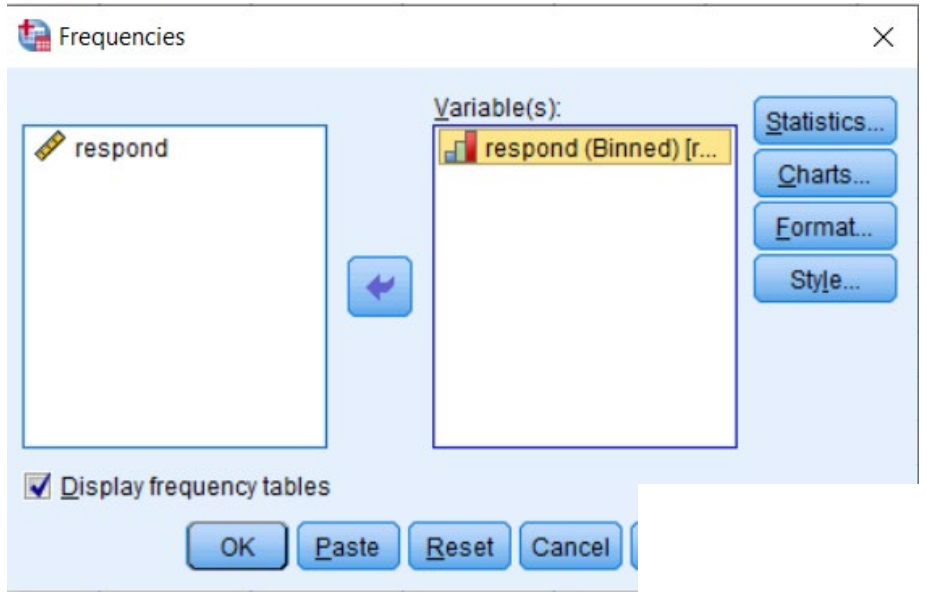
File Edit View Data Transform Analyze

1 : respond_cat | 1

	respond	respond_cat
1	35	< 225
2	38	< 225
3	48	< 225
4	53	< 225
5	70	< 225
6	99	< 225
7	138	< 225
8	164	< 225
9	220	< 225
10	265	225 - 449
11	272	225 - 449
12	312	225 - 449
13	339	225 - 449
14	340	225 - 449
15	395	225 - 449
16	457	450 - 674
17	478	450 - 674
18	501	450 - 674
19	521	450 - 674
20	556	450 - 674
21	583	450 - 674
22	595	450 - 674

To generate frequency table:

Analyze > Descriptive Statistics > Frequencies...



Frequency table for respond

respond (Binned)

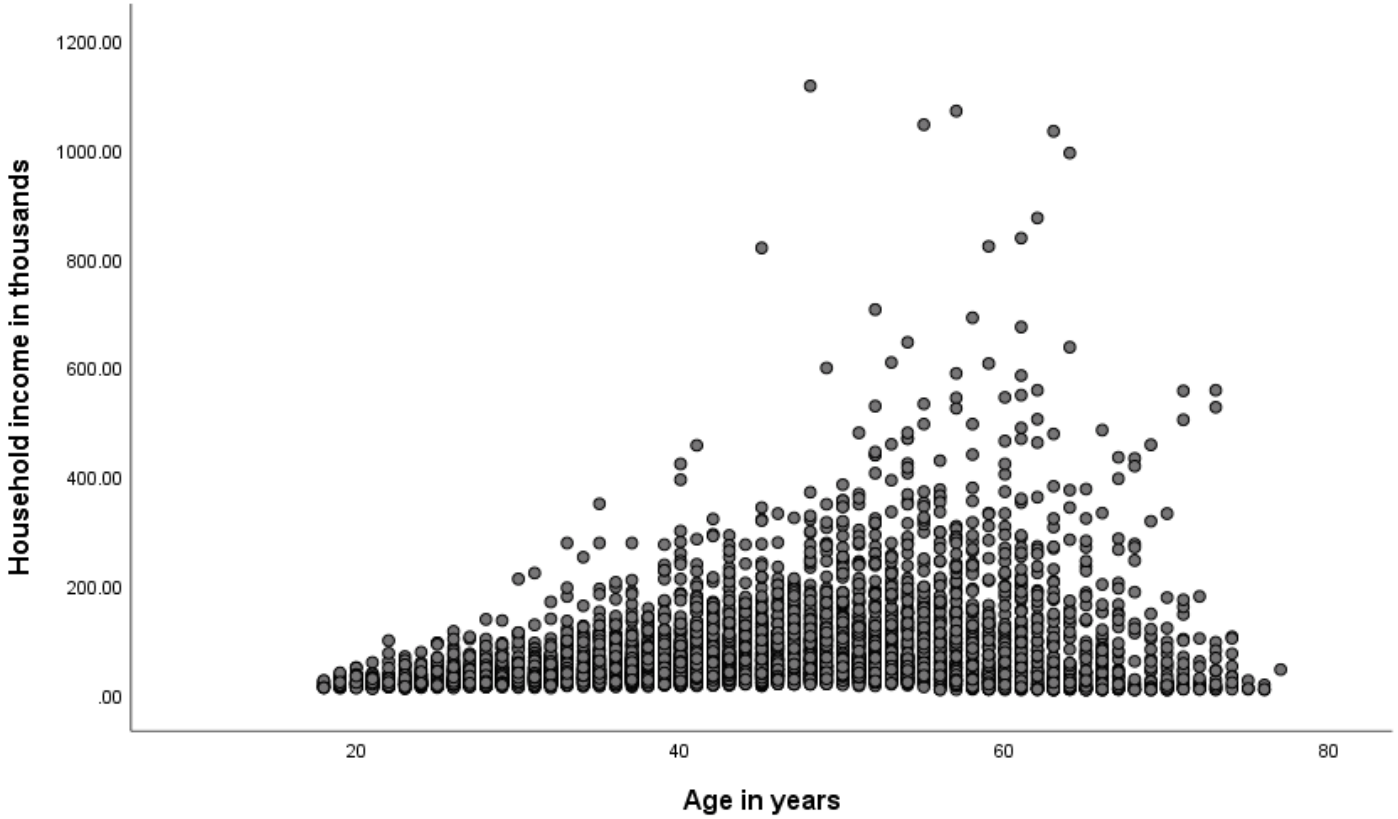
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 225	9	12.5	12.5	12.5
	225 - 449	6	8.3	8.3	20.8
	450 - 674	12	16.7	16.7	37.5
	675 - 899	13	18.1	18.1	55.6
	900 - 1124	14	19.4	19.4	75.0
	1125 - 1349	11	15.3	15.3	90.3
	1350 - 1574	7	9.7	9.7	100.0
	Total	72	100.0	100.0	

Descriptive Statistics: Quantitative Variable

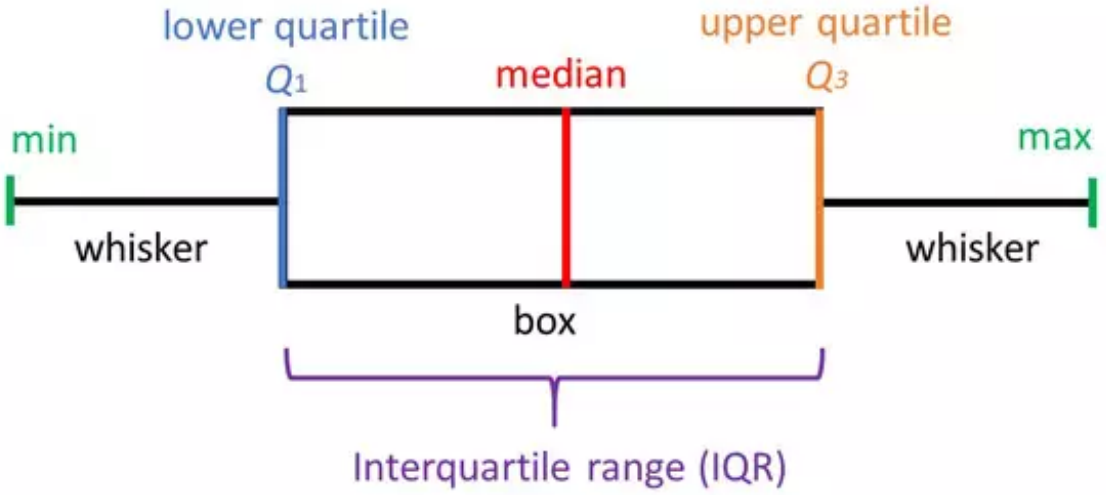
Data Set: *demo.sav*

Graph > Legacy Dialogs > Scatter/Dot > Simple Scatter

SCATTER PLOT



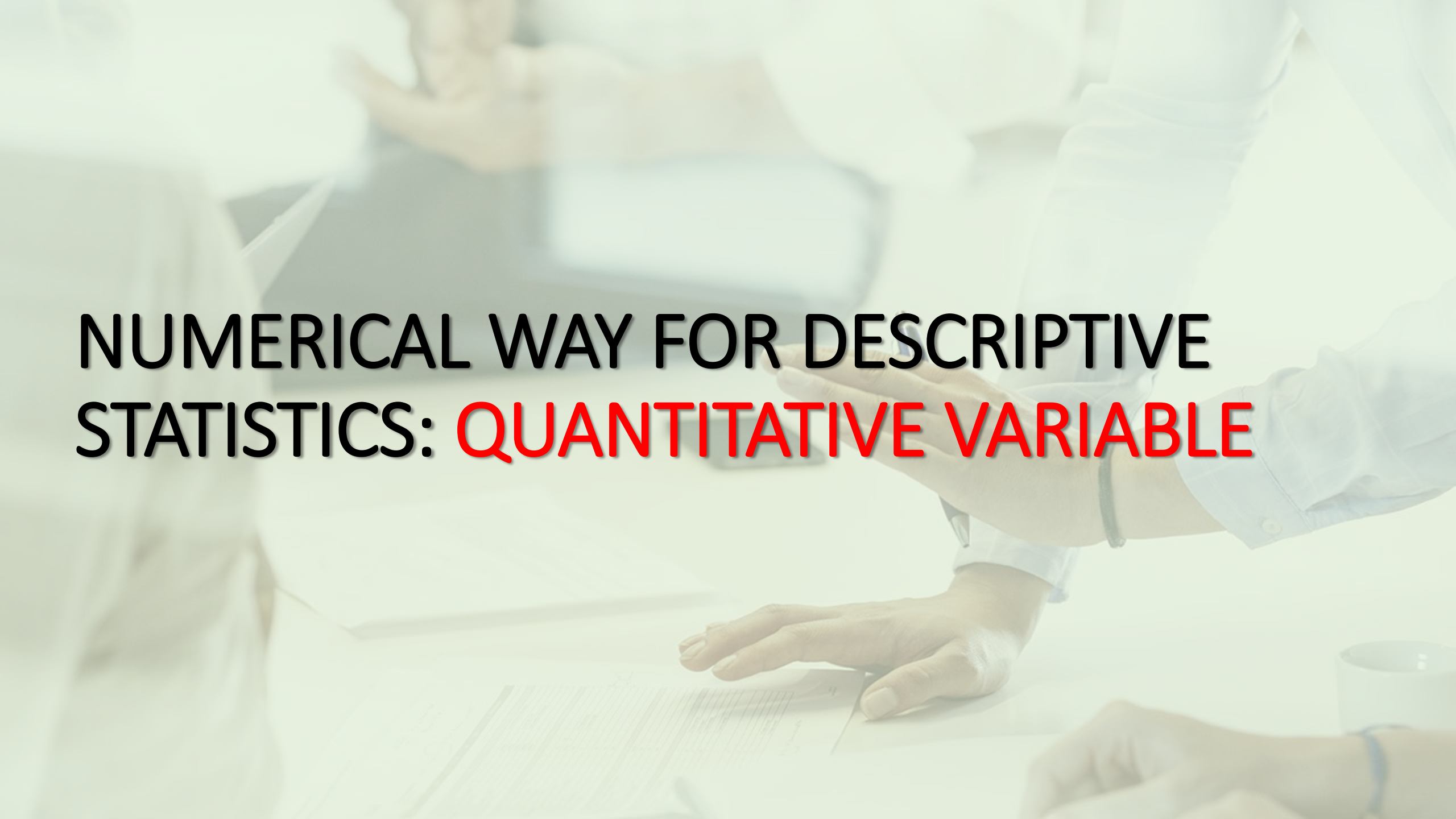
BOX PLOT



- Also known as box and whisker plots
- To visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles).
- Box has the width that ranges from Q1 to Q3 and vertical line through the box is placed at the median.
- Can be used to identify outliers.
- Quartiles : Those value that divide the data set into four equal-sized groups.
- The 1st quartile (Q1) = 25th percentile, 2nd quartile (Q2) = median, 3rd quartile (Q3) = 75th percentile and 4th quartile = 100th percentile
- IQR (Interquartile Range) = range from 1st quartile (Q1) to 3rd quartile (Q3)

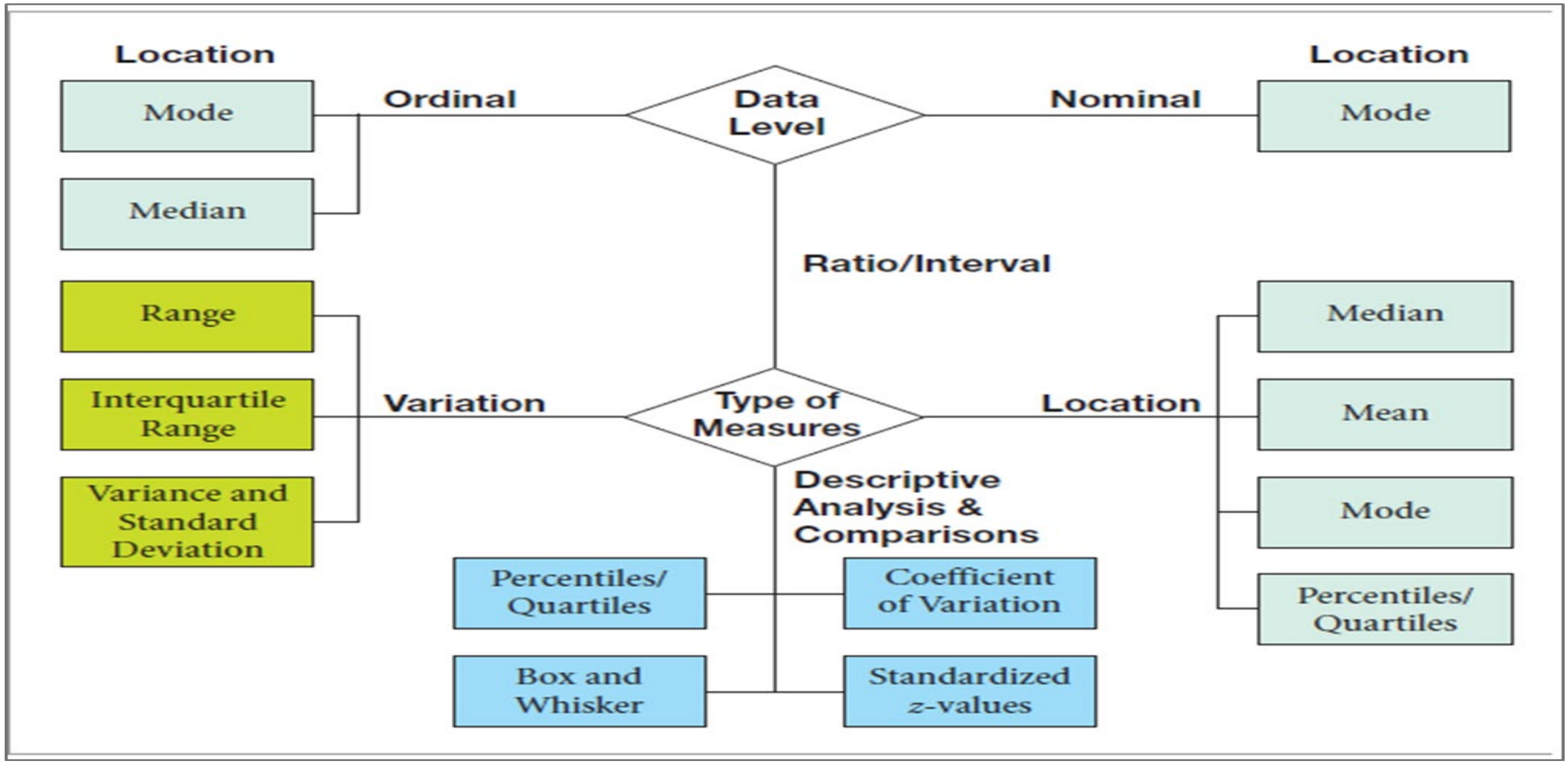
- Visual way descriptive statistics for <u>qualitative variable</u>:
  i. central tendency: the greater the count (frequency)
  ii. Position: maximum and minimum value
  iii. Variation/Dispersion: variability of count for each category in a variable (frequency table)

- Visual way descriptive statistics for quantitative variable:
  i. central tendency: the greater the count (frequency table)
  ii. Position: Box n Whisker Plot (4V)
  iii. Variation/Dispersion: histogram, box n whisker plot and scatter plot
  iv. Shape of distribution: distribution (or pattern) in histogram, box n whisker plot



NUMERICAL WAY FOR DESCRIPTIVE STATISTICS: QUANTITATIVE VARIABLE

DESCRIBING DATA IN NUMERICAL



MEDIAN

Arrange the data point from smallest to largest. Then compute median using equation 3.3:

Median Index

$$i = \frac{1}{2}n \qquad (3.3)$$

where:

i = The index of the point in the data set corresponding to the median value

n = Sample size

If i is not an integer, round its value up to the next highest integer. This next highest integer then is the position of the median in the data array.

If i is an integer, the median is the average of the values in position i and position $i + 1$.

Example 1:

Personnel manager has hired 10 new employees. The age of the employees is as follows: 23 25 25 34 35 45 46 47 52 54

Thus, median index = $\frac{1}{2}n = \frac{1}{2}(10) = 5$, therefore the $M_d = \frac{35+45}{2} = 40$

Example 2:

Customers at a restaurant are asked to rate the service they receives on a scale of 1 to 100. A total of 15 customers were asked to provide the ratings. The data are presented as follows:

60 68 75 77 80 80 80 85 88 90 95 95 95 95 99

Thus, median index = $\frac{1}{2}n = \frac{1}{2}(15) = 7.5$ (round up to 8), therefore the M_d= 85

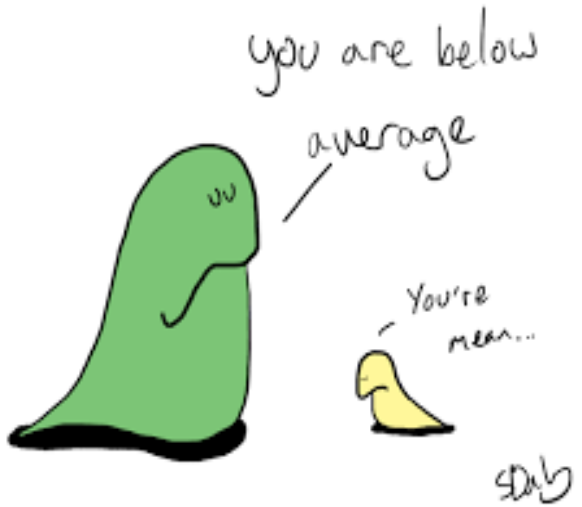
MEAN

Population Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

where:

μ = Population mean (mu)
N = Population size
x_i = ith individual value of variable x



Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where:

$\bar{x}$ = Sample mean (pronounced "x-bar")
n = Sample size

ALERT!!!

The Mean measure can be affected by extreme values (for example: income/salary data)

MEASURES OF DATA POSITION

Measure of data position: describe the relative position of a data value of a numerical variable to the other values of the variable.

Common measures are:

1. Percentiles
   - The pth percentile in a data array is a value that divides the data set into two parts.
   - The lower segment contains at least p% and the upper segment contains at least (100 - p)% of the data.
   - The 50th percentile is the median.
2. Quartiles
   - Those value that divide the data set into four equal-sized groups.
   - The 1st quartile (Q1) = 25th percentile, 2nd quartile (Q2) = median, 3rd quartile (Q3) = 75th percentile and 4th quartile = 100th percentile
   - IQR = range from 1st quartile (Q1) to 3rd quartile (Q3)

DESCRIBING DATA IN NUMERICAL: QUANTITATIVE VARIABLE (INTERVAL & RATIO)

Data Set: *demo.sav*

Central Tendency, Position and Dispersion

Analyze > Descriptive Statistics > Descriptive > ..

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Price of primary vehicle	6400	95.70	4.20	99.90	30.1284	21.92692	480.790
Years at current address	6400	56	0	56	11.56	9.938	98.767
Valid N (listwise)	6400						

DESCRIBING DATA IN NUMERICAL: QUANTITATIVE VARIABLE (INTERVAL & RATIO): LOOKING AT PROPORTION

Data Set: *demo.sav*

Example: Gender Proportion for Number of Years at current address

Analyze > Descriptive Statistics > Explore > ..

Descriptives

b

	Gender			Statistic	Std. Error
Years at current address	Female	Mean		11.40	.176
		95% Confidence Interval for Mean	Lower Bound	11.06	
			Upper Bound	11.75	
		5% Trimmed Mean		10.62	
		Median		9.00	
		Variance		98.085	
		Std. Deviation		9.904	
		Minimum		0	
		Maximum		56	
		Range		56	
		Interquartile Range		14	
		Skewness		1.052	.043
		Kurtosis		.705	.087
	Male	Mean		11.71	.176
		95% Confidence Interval for Mean	Lower Bound	11.37	
			Upper Bound	12.06	
		5% Trimmed Mean		10.94	
		Median		9.00	
		Variance		99.422	
		Std. Deviation		9.971	
		Minimum		0	
		Maximum		53	
		Range		53	
		Interquartile Range		14	
		Skewness		1.030	.043
		Kurtosis		.650	.086

Case Processing Summary

a

		Cases					
		Valid		Missing		Total	
	Gender	N	Percent	N	Percent	N	Percent
Years at current address	Female	3179	100.0%	0	0.0%	3179	100.0%
	Male	3221	100.0%	0	0.0%	3221	100.0%

Percentiles

c

			Percentiles						
		Gender	5	10	25	50	75	90	95
Weighted Average (Definition 1)	Years at current address	Female	.00	1.00	3.00	9.00	17.00	26.00	31.00
		Male	.00	1.00	4.00	9.00	18.00	26.00	31.00
Tukey's Hinges	Years at current address	Female			3.00	9.00	17.00		
		Male			4.00	9.00	18.00		

Extreme Values

d

	Gender			Case Number	Value
Years at current address	Female	Highest	1	1998	56
			2	5336	51
			3	317	50
			4	3731	50
			5	1796	49
		Lowest	1	6380	0
			2	6376	0
			3	6307	0
			4	6299	0
			5	6243	0[a]
	Male	Highest	1	5078	53
			2	6259	52
			3	4437	51
			4	2002	50
			5	2179	50
		Lowest	1	6383	0
			2	6348	0
			3	6297	0
			4	6284	0
			5	6143	0[a]

a. Only a partial list of cases with the value 0 are shown in the table of lower extremes.

NUMERICAL WAY FOR DESCRIPTIVE STATISTICS:
DATA VARIATION FOR QUANTITATIVE VARIABLE

WHY NEED TO EXPLORE AND SUMMARIZE DATA VARIATION?

Variation = A set of data exhibits variation if all the data is not in the same value.

Example: Consider two hospitals (Hospital A and Hospital B) monitoring the daily blood pressure readings (in mmHg) of patients in a critical care unit.

Hospital A	Hospital B
110 mmHg	125 mmHg
140 mmHg	126 mmHg
150 mmHg	124 mmHg
115 mmHg	125 mmHg
135 mmHg	126 mmHg

Table: Daily average systolic blood pressure readings in two hospitals

Hospital A:

Mean = 130.0 mmHg
Median = 135.0 mmHg

Hospital B:

Mean = 125.2 mmHg
Median = 125.0 mmHg

Summary: The descriptive statistics (mean and median) are equal, and the distribution of blood pressure readings at both hospitals appears symmetrical. Therefore, at first glance, the two hospitals seem to have similar patient health outcomes.

HOWEVER!!!: There is a HUGE variation in daily blood pressure readings in Hospital A, whereas Hospital B maintains a stable range of values. While Hospital B keeps systolic blood pressure readings almost the same every day, Hospital A has considerable fluctuations—some days are high-risk, while others are low.

THUS: Looking at only measures of central tendency (mean/median) can be misleading. We need MEASURES OF VARIATION to understand the true health stability of patients.

Measures of variation

Range

Interquartile Range (IQR)

Variance & Standard Deviation

Coefficient of Variation (CV)

1. RANGE

- The range is a measure of variation that is computed by finding the difference between the maximum and minimum values in a data set.
- Range = Maximum value – Minimum value
- But range sensitive to extreme values.

2. INTERQUARTILE

- A measure of variation that tends to overcome the range's susceptibility to extreme values
- Interquartile = Third Quartile(Q_3)– First Quartile(Q_1)
- But the measure did not use all the available data in its computation

3 & 4. VARIANCE AND STANDARD DEVIATION

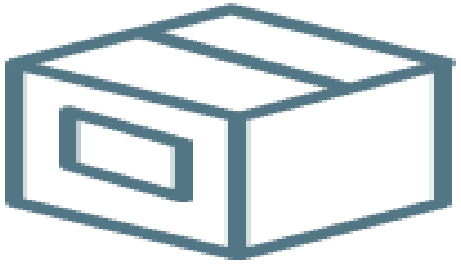
- The measure (variance and standard deviation) incorporate all the values in the data set.

- These two measures are closely related.
  - The standard deviation is the square root of the variance.

- Standard deviation – is the original unit, thus the standard deviation are usually used to measure variation in a population and sample.
  - Why? – because dealing with original units are easier compare with the square of the unit (variance)

NUMERICAL WAY FOR DESCRIPTIVE STATISTICS: COMPARING VARIATION FOR QUANTITATIVE VARIABLE

TIPS TO DESCRIBE VARIATION IN YOUR DATASET



If 2 different sample/population having **SAME MEAN** can use **standard deviation** to measure variation

If 2 different sample/population having **NOT SAME MEAN** can use coefficient of variation (CV) to measure variation

COEEFICIENT OF VARIATION

- If 2 different sample having same mean ➔ can use standard deviation to measure variation

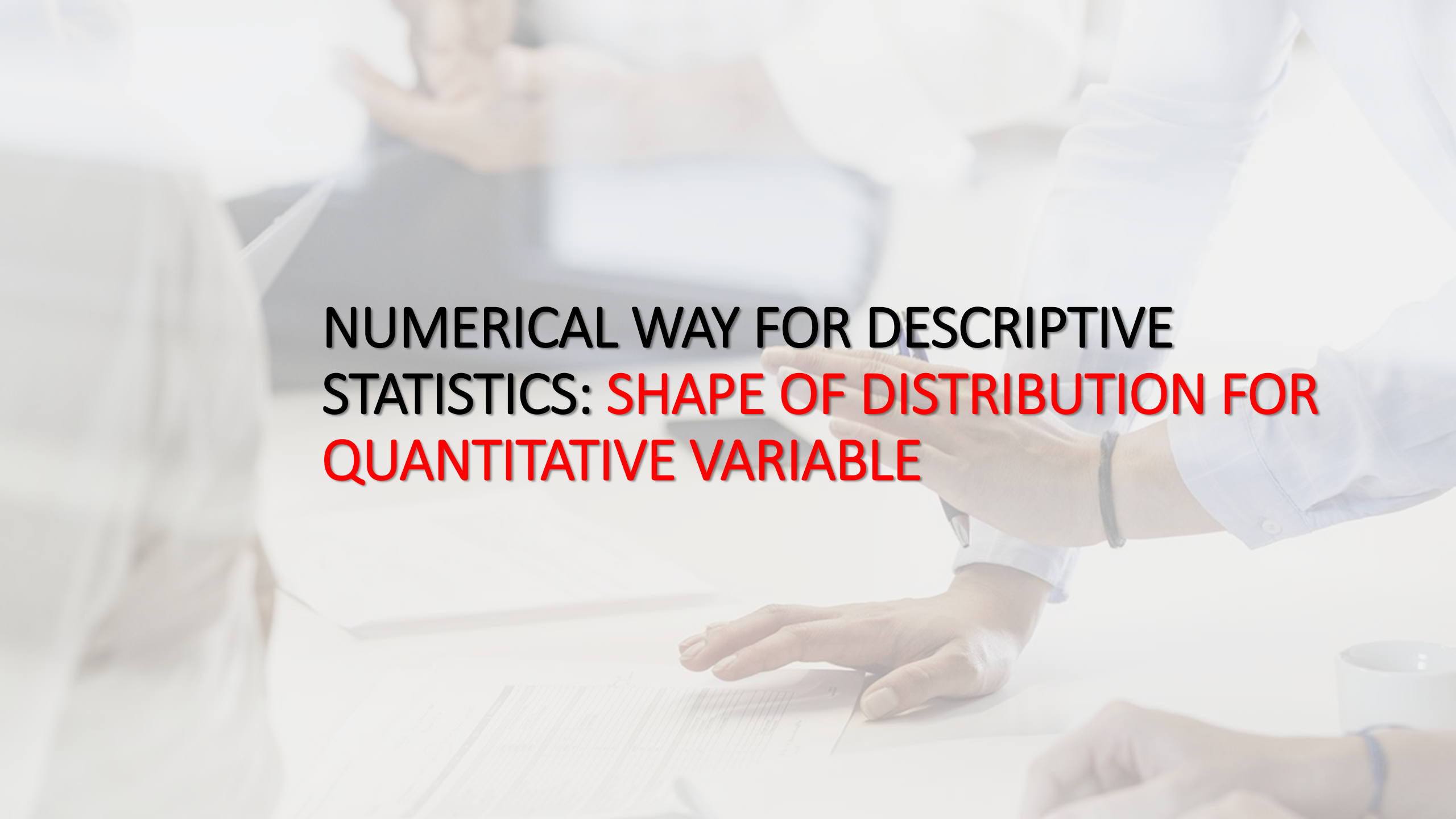
- BUT, if 2 distribution did not have same mean, it does not make sense to measure based on standard deviation.
  - need other measurement to measure the variation ➔ COEEFICIENT OF VARIATION

- Coefficient of variation (CV):

Ratio of standard deviation and mean

$$\text{Coefficient of Variation} = \frac{\sigma}{\mu} \times 100\%$$

  - the distribution with the largest *CV* is said to have the greatest relative spread.



NUMERICAL WAY FOR DESCRIPTIVE STATISTICS: SHAPE OF DISTRIBUTION FOR QUANTITATIVE VARIABLE

DESCRIPTIVE STATISTICS FOR SHAPE OF DISTRIBUTION

- Data in population or sample can be either:
  - Symmetric (Normal – Bell Shaped) Distribution – data sets whose values are evenly spread around the center.
  - Non-symmetric Distribution - data sets whose values are not evenly spread around the center.
- Statistics measure for shape of distribution: SKEWNESS & KURTOSIS
  - Skewness : refers to degree of symmetry
  - Kurtosis: refers to degree of outliers' presence in the distribution

SKEWNESS

Skewness statistics – implies the direction of skewness.

The higher the absolute value, the more the data are skewed.

When the data is highly skewed, median is a useful measure of the center

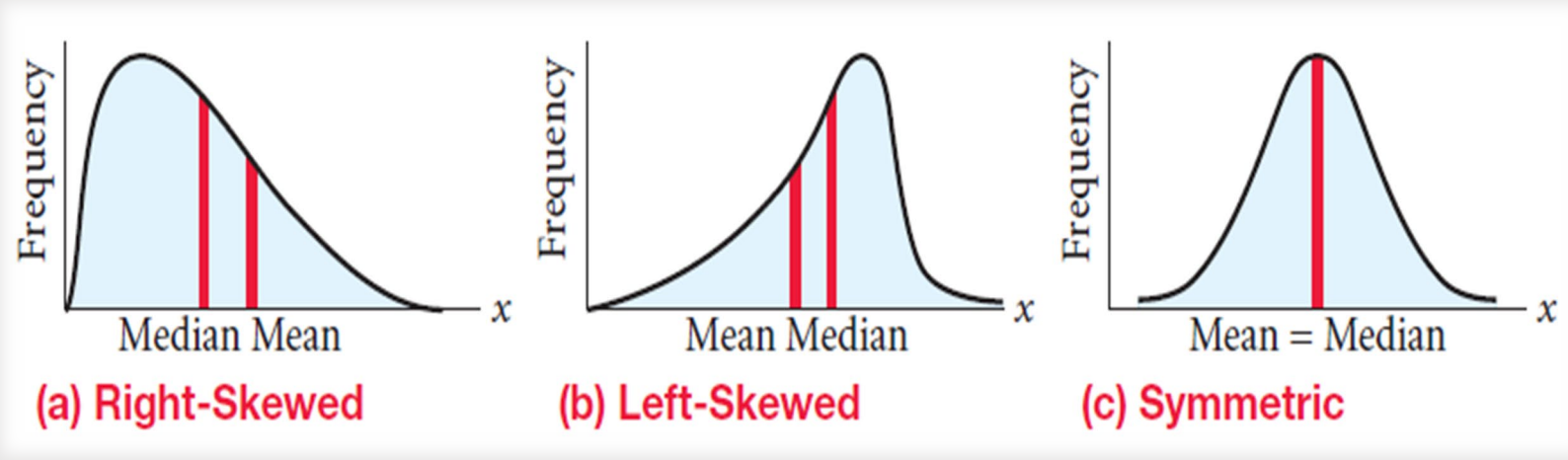
A positive skewness indicates that the size of the right-handed tail is larger than the left-handed tail and vice versa.

The rule of thumb:

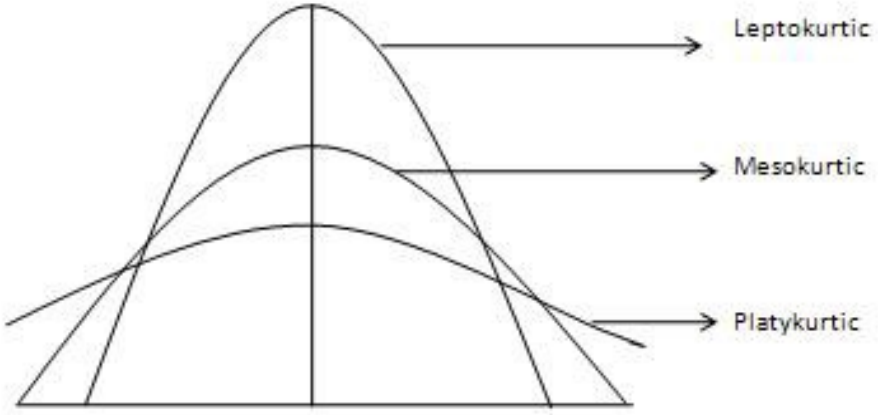
If the skewness is between -0.5 and 0.5, the data are fairly symmetrical

If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed

If the skewness is less than -1 or greater than 1, the data are highly skewed



KURTOSIS



- Kurtosis statistics – implies **how tall and sharp the central peak is**, relative to a standard bell curve.
- **Type of Kurtosis:**
  - **Mesokurtic (close to 0)**: Like that of the normal distribution.
  - **Leptokurtic (greater than zero):** Distribution is longer, tails are fatter. Peak is higher and sharper than Mesokurtic, which means that data are heavy-tailed or profusion of outliers.
  - **Platykurtic: (less than zero):** Distribution is shorter, tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers.
    The reason for this is because the extreme values are less than that of the normal distribution.

DESCRIBING DATA IN NUMERICAL: QUANTITATIVE VARIABLE (SHAPE OF DISTRIBUTION)

Analyze > Descriptive Statistics > Descriptive > ..

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Price of primary vehicle	6400	95.70	4.20	99.90	30.1284	21.92692	480.790	1.216	.031	.524	.061
Years at current address	6400	56	0	56	11.56	9.938	98.767	1.041	.031	.675	.061
Valid N (listwise)	6400										

What can you summarize in terms of the shape of distribution for female and male?

Descriptives

	Gender			Statistic	Std. Error
Years at current address	Female	Mean		11.40	.176
		95% Confidence Interval for Mean	Lower Bound	11.06	
			Upper Bound	11.75	
		5% Trimmed Mean		10.62	
		Median		9.00	
		Variance		98.085	
		Std. Deviation		9.904	
		Minimum		0	
		Maximum		56	
		Range		56	
		Interquartile Range		14	
		Skewness		1.052	.043
		Kurtosis		.705	.087
	Male	Mean		11.71	.176
		95% Confidence Interval for Mean	Lower Bound	11.37	
			Upper Bound	12.06	
		5% Trimmed Mean		10.94	
		Median		9.00	
		Variance		99.422	
		Std. Deviation		9.971	
		Minimum		0	
		Maximum		53	
		Range		53	
		Interquartile Range		14	
		Skewness		1.030	.043
		Kurtosis		.650	.086

DESCRIBING DATA IN NUMERICAL

i. Central tendency: represents the center point or typical value of a dataset.
   - Mean, Mode, Median

ii. Position: location of an individual value (observation) in dataset.
   - Quartile (Percentile): Q1 (25^{th} percentile), Q2 (50^{th} percentile), Q3 (75^{th} percentile), Q4 (100^{th} percentile)

iii. Variation/Dispersion: variability within the dataset.
   - Range, IQR, Variance, Standard Deviation

iv. Shape of distribution: distribution (or pattern) within the dataset.
   - Kurtosis, Skewness

Try this....

PROBLEM: Patient Analysis

Healthcare_Dataset.xlsx containing information about 200 patients, including their age, gender, BMI, blood pressure, cholesterol levels, diabetes status, heart rate, and smoking history. Using the provided healthcare dataset, perform descriptive statistical analysis to summarize key characteristics of the patient population. Answer the following:

1. Central Tendency:

1. What are the mean, median, and mode of **Age** and **BMI**?
2. What does the mean tell you about the distribution of these variables?

2. Dispersion & Variability:

1. Calculate the range, variance, and standard deviation for **Systolic Blood Pressure** and **Diastolic Blood Pressure**.
2. Interpret these values: Are the blood pressure values highly variable?

3. Categorical Data Analysis:

1. What percentage of the patients have **High Cholesterol Levels**?
2. What proportion of patients are classified as **Diabetic**?

4. Exploratory Data Analysis:

1. Construct a histogram for **BMI**. What does the shape of the histogram indicate about the distribution?
2. Create a boxplot for **Heart Rate**. Are there any outliers? What do they suggest?

- **Bonus Question:** Is there a noticeable difference in **BMI** between smokers and non-smokers? Support your answer with descriptive statistics.


LESSON 1 END

THANK YOU