Index

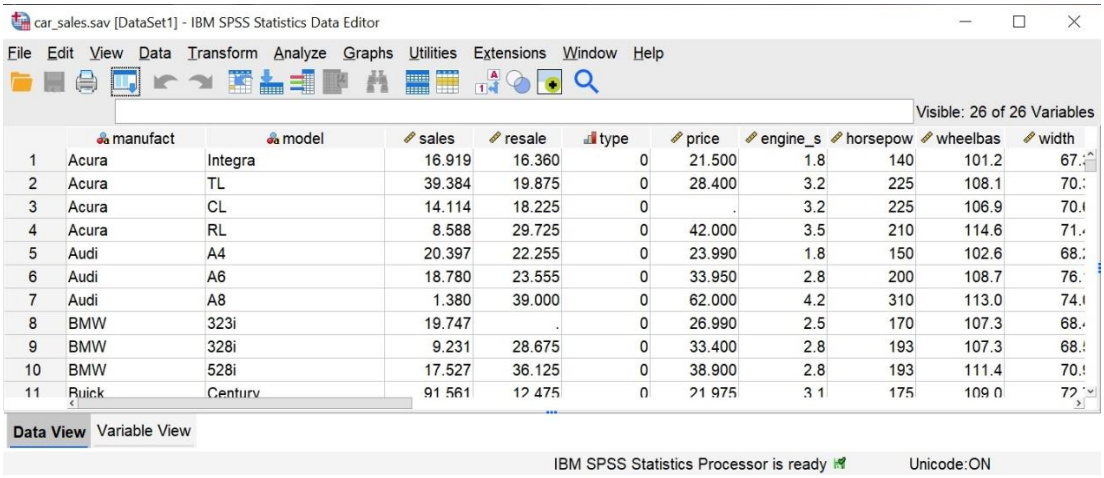
Introduction to IBM SPSS

Statistical Package for Social Sciences **(SPSS)** is a proprietary software of IBM.

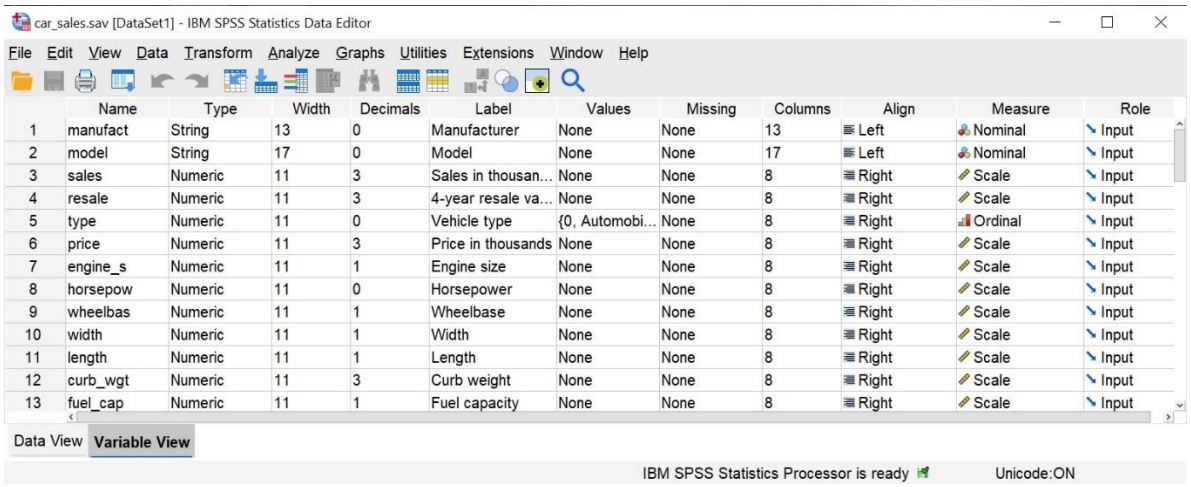
Data Editor: It displays the contents of the data file. This is the window that opens automatically when you start an SPSS session. In this window, you can create new data files or modify existing ones.

1. The data editor provides two views of your data:

- **Data View** - It displays the data view. Each variable is a column & each row is a case.



- **Variable View** – It displays variable definition information, including defined variables and value labels, data type (for e.g. string, date, and numeric), measurement level (nominal, ordinal, scale) and user-defined missing values. Some of the points are discussed below.



The following variable attributes (including others) can be entered and modified in the variable view:

I. **Variable Name**: It is the name given to a variable and each name should be unique: duplication is not allowed. *Variable names cannot contain spaces*. Variable name can be defined with any mixture of upper case and lower case characters and are preserved for display purposes.

II. **Type:** It specifies the type of data for each variable. It can be one of the following;

   a. **Numeric**: A variable whose values are numbers. Values are displayed in standard numeric format.

   b. **Comma**: A numeric value whose values are displayed with commas delimiting every three places and displayed with the period as a decimal delimiter(eg. 1,234 or 1,234,567)

   c. **Dot**: A numeric variable whose values are displayed with periods delimiting every three places and with a comma as a decimal limiter (eg. 1.234 or 1.234.567).

   d. **Scientific Notation**: A numeric variable whose values are displayed with an embedded E and a signed power of 10 exponents. For eg 123, 1.23E2, 1.23E+2 and 1.23+2.

   e. **Date**: A numerical variable whose values are displayed in one of the calendar-date or clock-time formats.

   f. **Dollar**: a numerical variable displayed with a leading dollar sign ($), commas delimiting every three places, and a period as a decimal delimiter.

   g. **Custom Currency**: Numeric variable whose values are displayed in one of the custom currency formats that you have defined on the currency tab of the options dialog box.

   h. **String**: A variable whose values are not numeric and therefore are not used in calculations.

   i. **Restricted Numeric**: A variable whose values are restricted to a nonnegative integer.

III. **Width**: It is used to specify the number of digits or characters that can be entered.

IV. **Decimals**: It is used to specify the number of decimal places in data representation.

V. **Label**: The name of variable can be described in detail here.

VI. **Values**: Nominal or categorical variable can be assigned numerical values. eg. If variable is gn (Gender) and entered as male or female then here assigned values 1 or 2 respectively

VII. **Missing values**: These are user-defined missing values.

VIII. **Columns**: It is used to specify the width of columns (can be changed by dragging also).

IX. **Measurement Level**: It can be specified as follows:

a) **Ordinal**- A categorical variable where sense of ordering (eg. low, medium high) holds. Such type of variables has several ordered categories. Examples of ordinal variables include attitude scores representing the degree of satisfaction or confidence and preference rating scores.

b) **Scale**- A variable can be treated as a scale (continuous) when its value represents ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables include age in years and income in thousands of dollars.

c) **Nominal**- A categorical variable used without any ordering. Such type of variable has several unordered categories. Examples of nominal variables include gender, region, zip code, and religious affiliation.

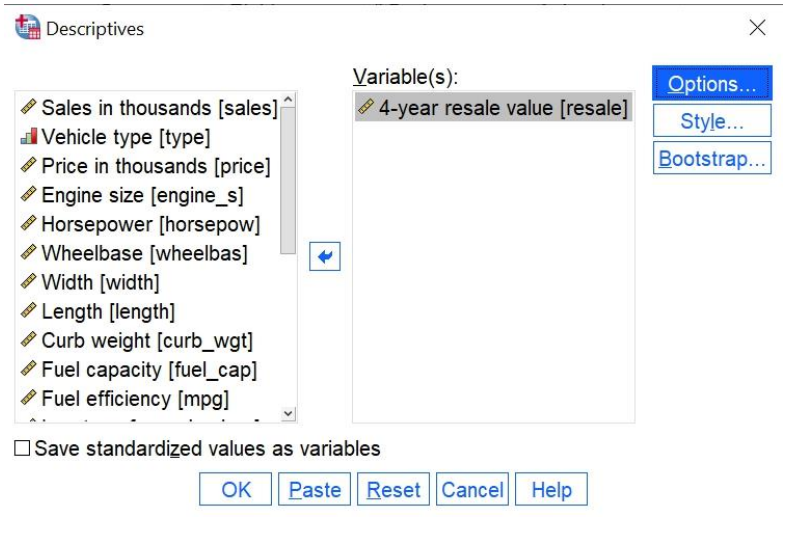
Variable types - IBM Documentation

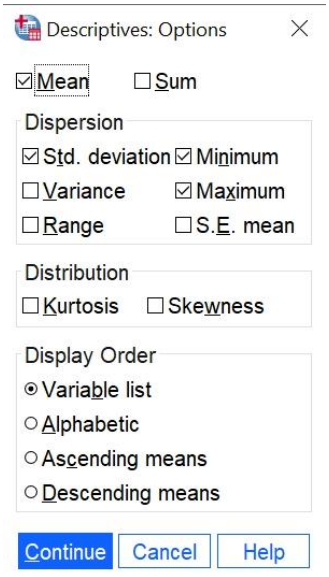
Table 1. Measurement level icons

	Numeric	String	Date	Time
Scale (Continuous)		n/a		
Ordinal				
Nominal				

2. Data: Enter data in data view as per the property of variable specified in variable view various formats, or data (if already stored in other format) can be imported to SPSS editor and the properties of variables can be changed according to need in variable view.

3. Operations: Any operation on data value will be done & observed in data view.





And statistical operation's information will be created in another window

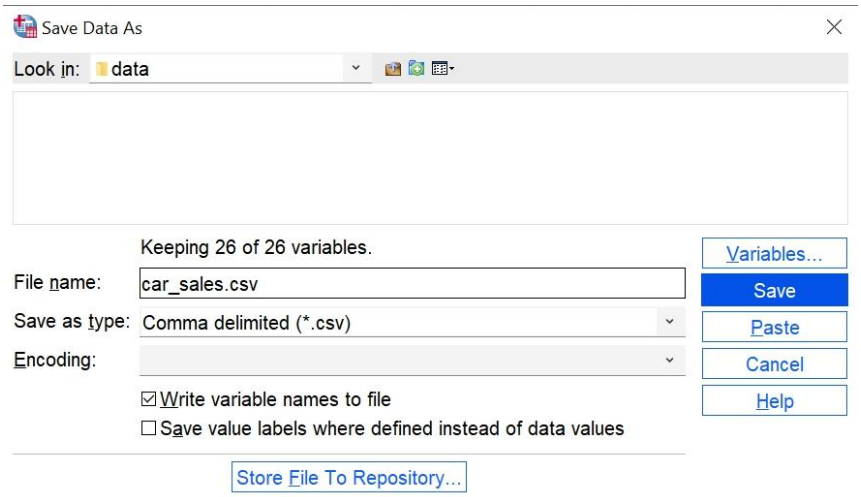
Descriptives

[DataSet1] C:\Users\Administrator\OneDrive\Desktop\Lab_Semester_3_Data\Probability And Statistics\data\car_sales.sav

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
4-year resale value	121	5.160	67.550	18.07298	11.453384
Valid N (listwise)	121				

4. Save: The last step is to 'Save' the files which can be done in various format using 'Export' option under 'File'



5. Saved: File will be saved at desired location in desired format.

Experiment - 1

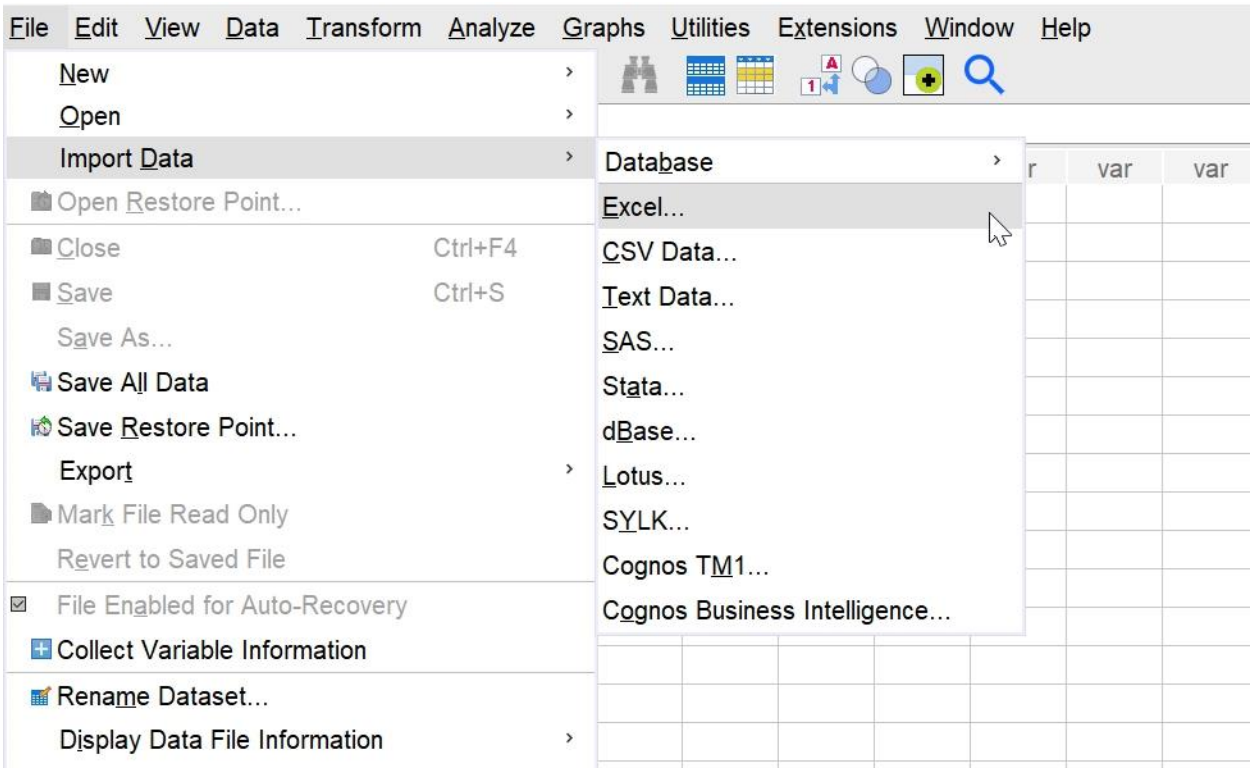
Transportation of Dataset to SPSS Editor

STEPS

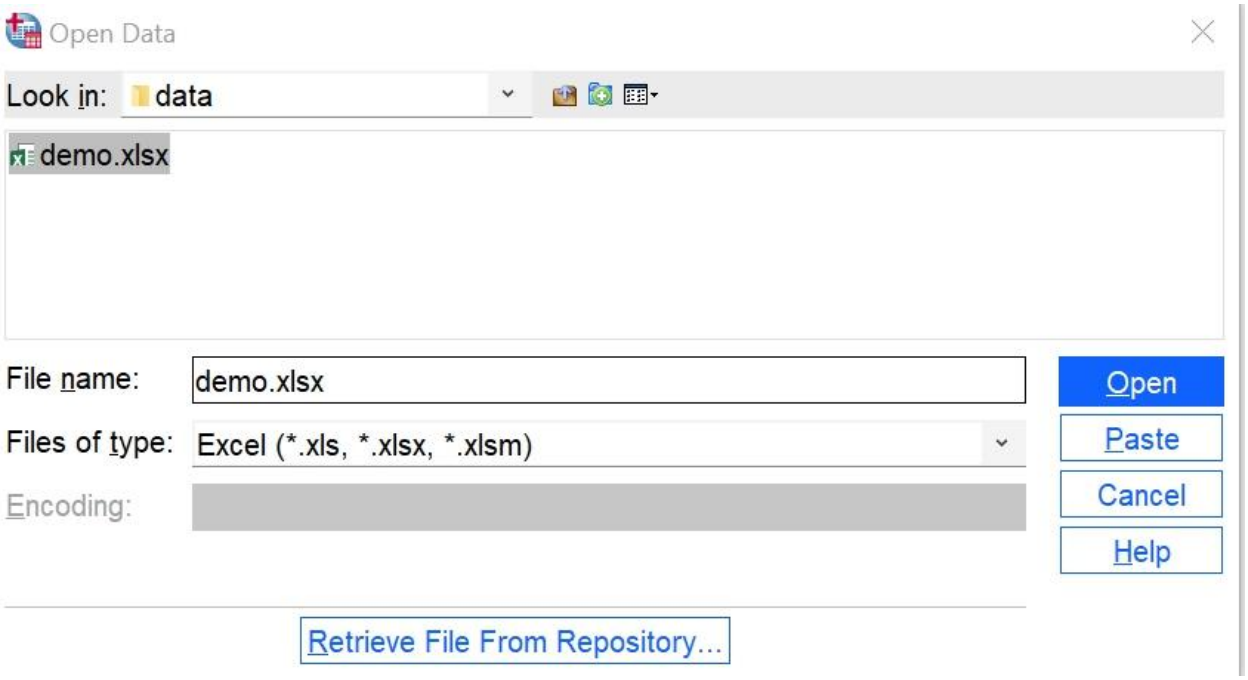
CASE – 1

- Data from an excel file

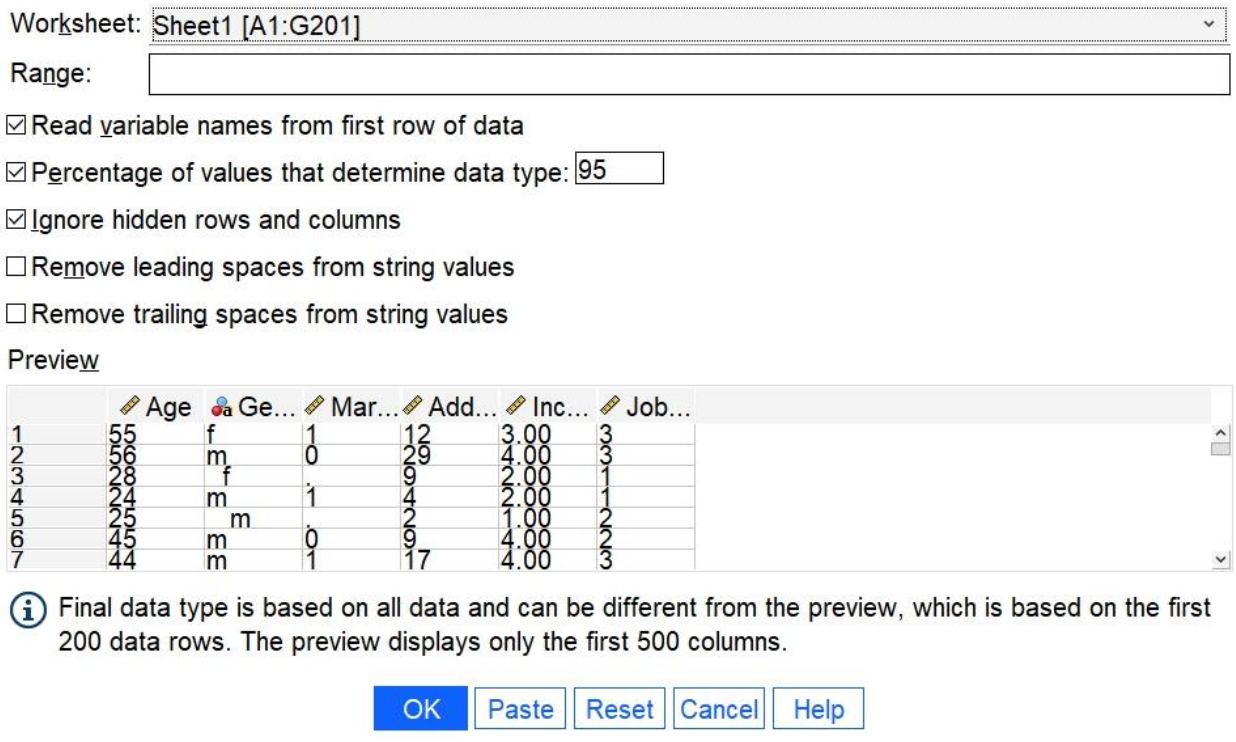
Commands -> File>Open>Import Data>Excel



Commands -> click <file-name> -> Open



Commands -> Press OK



	Age	Ge...	Mar...	Add...	Inc...	Job...
1	55	f	1	12	3.00	3
2	56	m	0	29	4.00	3
3	28	f	.	9	2.00	1
4	24	m	1	4	2.00	1
5	25	m	.	2	1.00	2
6	45	m	0	9	4.00	2
7	44	m	1	17	4.00	3

Data View

*Untitled4 [DataSet3] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	Age	Gender	MaritalStatus	Address	IncomeCategory	JobCategory	var	var
1	55	f	1	12	3.00	3		
2	56	m	0	29	4.00	3		
3	28	f	.	9	2.00	1		
4	24	m	1	4	2.00	1		
5	25	m	.	2	1.00	2		
6	45	m	0	9	4.00	2		
7	44	m	1	17	4.00	3		
8	46	m	.	20	4.00	3		
9	41	m	.	10	2.00	2		
10	29	f	.	4	1.00	2		
11	34	m	0	0	4.00	2		
12	55	f	0	17	3.00	1		
13	28	m	0	9	3.00	1		
14	21	f	1	2	1.00	1		
15	55	f	0	8	4.00	2		
16	35	m	0	8	3.00	2		
17	45	f	0	4	2.00	2		
18	21	m	0	1	2.00	1		
19	32	f	0	0	2.00	1		
20	42	f	0	9	4.00	3		
21	40	f	1	12	4.00	3		
22	36	f	0	6	2.00	1		
23	42	m	1	13	3.00	2		
24	65	m	1	17	2.00	3		
25	52	m	1	5	4.00	3		
26	51	m	1	17	4.00	2		
27	44	m	1	1	2.00	2		
28	26	f	0	6	2.00	1		
29	41	f	1	19	3.00	3		
30	49	m	0	11	2.00	2		
31	64	f	0	36	4.00	3		
32	39	m	0	8	3.00	2		
33	53	m	0	8	4.00	3		
34	38	f	1	17	2.00	2		
35	46	f	1	6	2.00	3		
36	58	m	0	2	3.00	1		
37	25	f	0	0	3.00	1		
38	57	f	0	28	4.00	3		
39	47	m	0	21	2.00	3		
40	21	f	0	0	1.00	1		
41	45	f	1	21	3.00	3		
42	56	f	0	7	4.00	3		
43	24	m	0	2	1.00	1		

Data View Variable View

Variable View

*Untitled4 [DataSet3] - IBM SPSS Statistics Data Editor

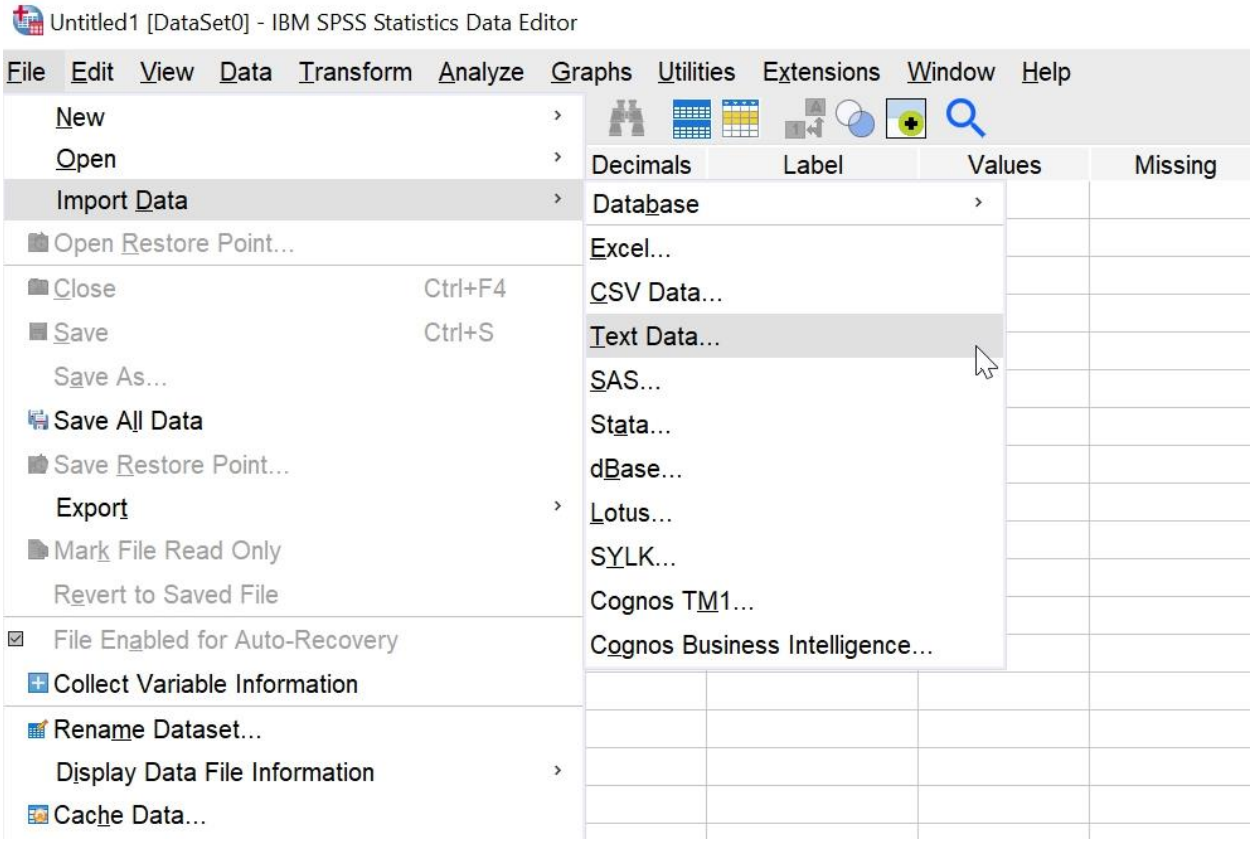
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Age	Numeric	2	0		None	None	12	Right	Scale	Input
2	Gender	String	4	0		None	None	4	Left	Nominal	Input
3	MaritalStatus	Numeric	9	0	Marital Status	None	None	12	Right	Nominal	Input
4	Address	Numeric	2	0		None	None	12	Right	Scale	Input
5	IncomeCate...	Numeric	4	2	Income Category	None	None	12	Right	Nominal	Input
6	JobCategory	Numeric	1	0	Job Category	None	None	12	Right	Nominal	Input
7											

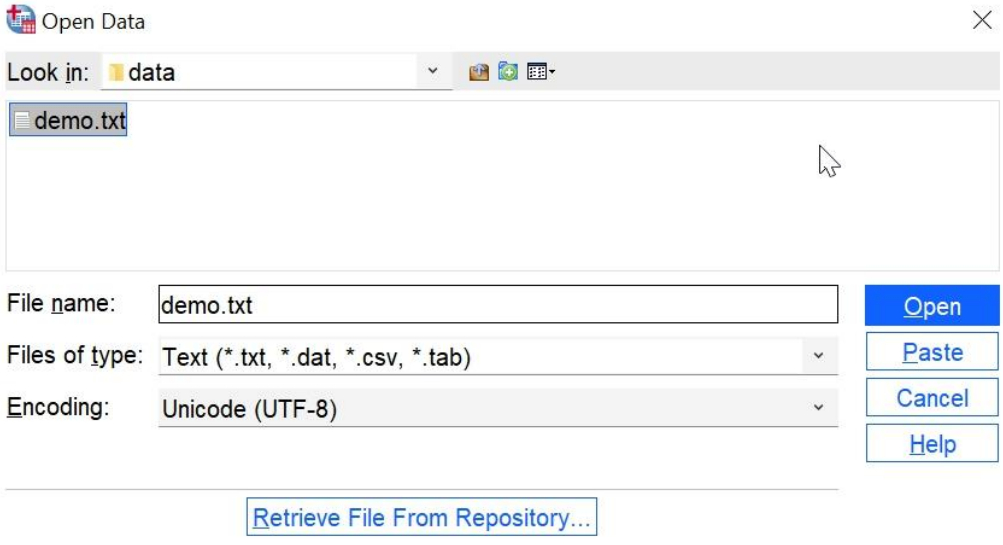
CASE – 2

- Data from an text file

Commands -> File>Open>Import Data>Text Data

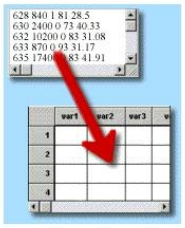


Commands -> click <file-name> -> Open



Commands -> Press Next (Till all the steps end)

Text Import Wizard - Step 1 of 6



Welcome to the text import wizard!

This wizard will help you read data from your text file and specify information about the variables.

Does your text file match a predefined format?

○ Yes [ ] Browse...

◉ No

Text file: C:\Users\Administrator\OneDrive\Desktop\Lab_Semester_3_Data\Probability And Statistics\data\demo...

Name	Age	Gender	Marital Status	Address	Income	Income Category	Job Category
2	55	f	1	12	72.00	3.00	3
3	56	m	0	29	153.00	4.00	3
4	28	f	no answer	9	28.00	2.00	1
5	24	m	1	4	26.00	2.00	1
6	25	m	no answer	2	23.00	1.00	2
7	45	m	0	9	76.00	4.00	2
8	44	m	1	17	144.00	4.00	3
9	46	m	no answer	20	75.00	4.00	3
10	41	m	no answer	10	26.00	2.00	2
11	29	f	no answer	4	19.00	1.00	2
12	34	m	0	0	89.00	4.00	2
13	55	f	0	17	72.00	3.00	1
14	28	m	0	9	55.00	3.00	1
15	21	f	1	2	20.00	1.00	1
16	55	f	0	8	283.00	4.00	2
17	35	m	0	8	70.00	3.00	2

< Back | Next > | Finish | Cancel | Help

Text Import Wizard - Step 2 of 6

How are your variables arranged?

◉ Delimited - Variables are delimited by a specific character (i.e., comma, tab).

○ Fixed width - Variables are aligned in fixed width columns.

Are variable names included at the top of your file?

◉ Yes

Line number that contains variable names: 1

○ No

What is the decimal symbol?

◉ Period

○ Comma

Text file: C:\Users\Administrator\OneDrive\Desktop\Lab_Semester_3_Data\Probability And Statistics\data\demo...

Name	Age	Gender	Marital Status	Address	Income	Income Category	Job Category
2	55	f	1	12	72.00	3.00	3
3	56	m	0	29	153.00	4.00	3
4	28	f	no answer	9	28.00	2.00	1
5	24	m	1	4	26.00	2.00	1
6	25	m	no answer	2	23.00	1.00	2
7	45	m	0	9	76.00	4.00	2
8	44	m	1	17	144.00	4.00	3
9	46	m	no answer	20	75.00	4.00	3
10	41	m	no answer	10	26.00	2.00	2

< Back | Next > | Finish | Cancel | Help

Text Import Wizard - Delimited Step 3 of 6

The first case of data begins on which line number? 2

How are your cases represented?

- ◉ Each line represents a case
- ○ A specific number of variables represents a case: 7

How many cases do you want to import?

- ◉ All of the cases
- ○ The first 1000 cases.
- ○ A random percentage of the cases (approximate): 10 %

Data preview

Name	Age	Gender	Marital Status	Address	Income	Income Category	Job Category
2	55	f	1	12	72.00	3.00	3
3	56	m	0	29	153.00	4.00	3
4	28	f	no answer		9	28.00	2.00 1
5	24	m	1	4	26.00	2.00	1
6	25	m	no answer		2	23.00	1.00 2
7	45	m	0	9	76.00	4.00	2
8	44	m	1	17	144.00	4.00	3
9	46	m	no answer		20	75.00	4.00 3
10	41	m	no answer		10	26.00	2.00 2
11	29	f	no answer		4	19.00	1.00 2
12	34	m	0	0	89.00	4.00	2
13	55	f	0	17	72.00	3.00	1
14	28	m	0	9	55.00	3.00	1

< Back | Next > | Finish | Cancel | Help

Text Import Wizard - Delimited Step 4 of 6

Which delimiters appear between variables?

- ☑ Tab
- ☑ Space
- ☐ Comma
- ☐ Semicolon
- ☐ Other:

What is the text qualifier?

- ◉ None
- ○ Single quote
- ○ Double quote
- ○ Other:

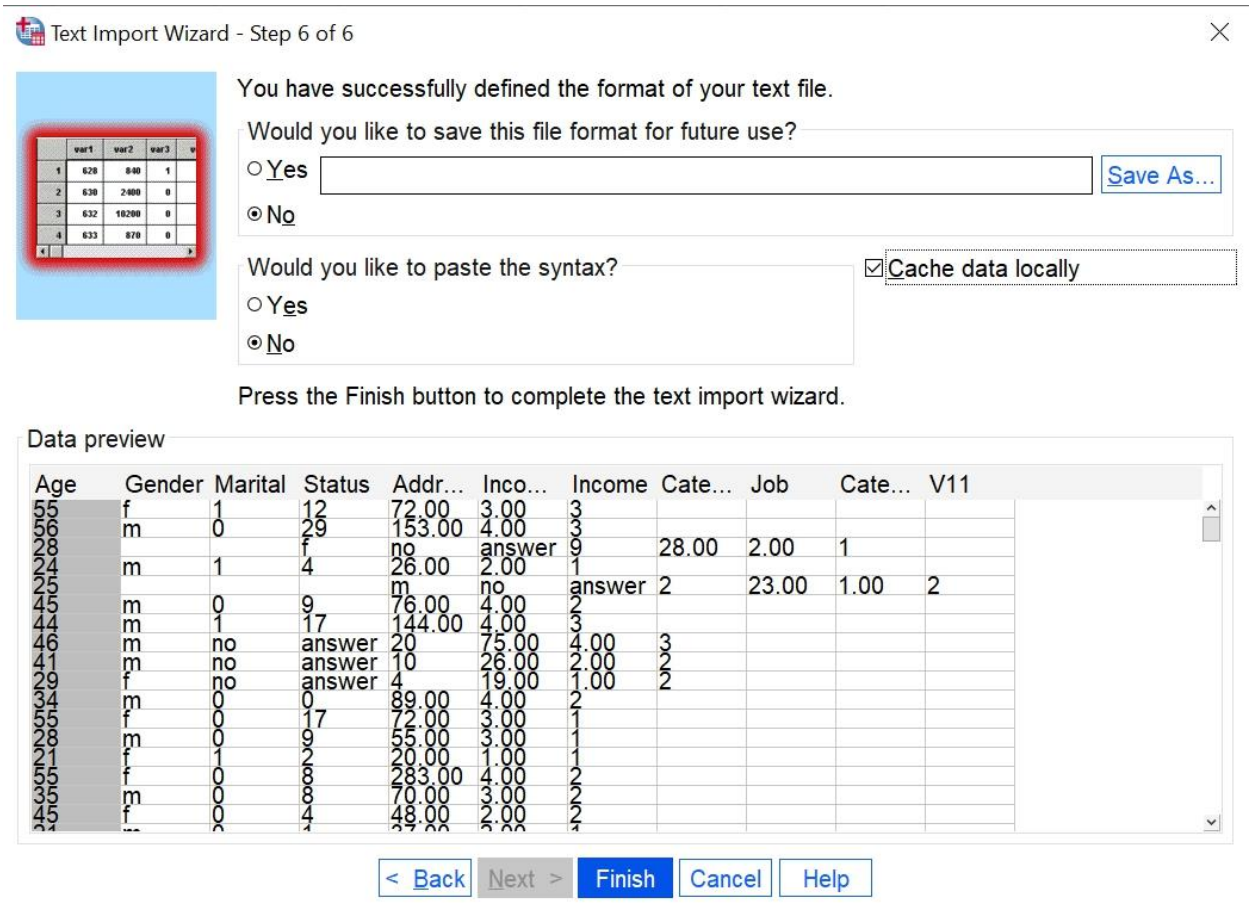
Leading and Trailing Spaces

- ☐ Remove leading spaces from string values
- ☐ Remove trailing spaces from string values

Data preview

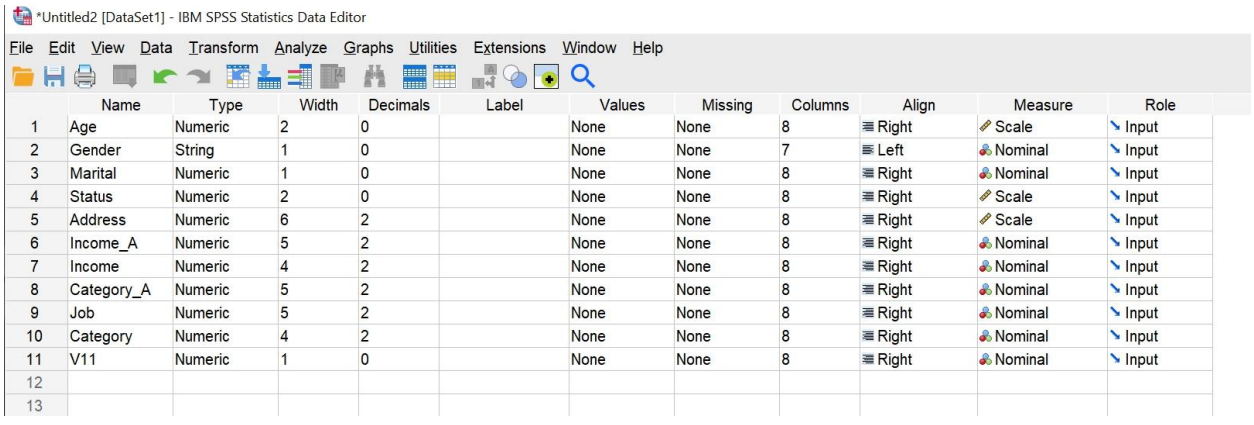
Age	Gender	Marital	Status	Addr...	Income	Income	Cate...	Job	Cate...	V11
55	f	1	12	72.00	3.00	3				
56	m	0	29	153.00	4.00	3				
28			f	no	answer	9	28.00	2.00	1	
24	m	1	4	26.00	2.00	1				
25				m	no	answer	2	23.00	1.00	2
45	m	0	9	76.00	4.00	2				
44	m	1	17	144.00	4.00	3				
46	m	no	answer	20	75.00	4.00	3			
41	m	no	answer	10	26.00	2.00	2			
29	f	no	answer	4	19.00	1.00	2			
34	m	0	0	89.00	4.00	2				
55	f	0	17	72.00	3.00	1				
28	m	0	9	55.00	3.00	1				
21	f	1	2	20.00	1.00	1				
55	f	0	8	283.00	4.00	2				
35	m	0	8	70.00	3.00	2				
45	f	0	4	48.00	2.00	2				
21	m	0	1	37.00	2.00	1				
32	f	0	0	28.00	2.00	1				
42	f	0	9	109.00	4.00	3				

< Back | Next > | Finish | Cancel | Help



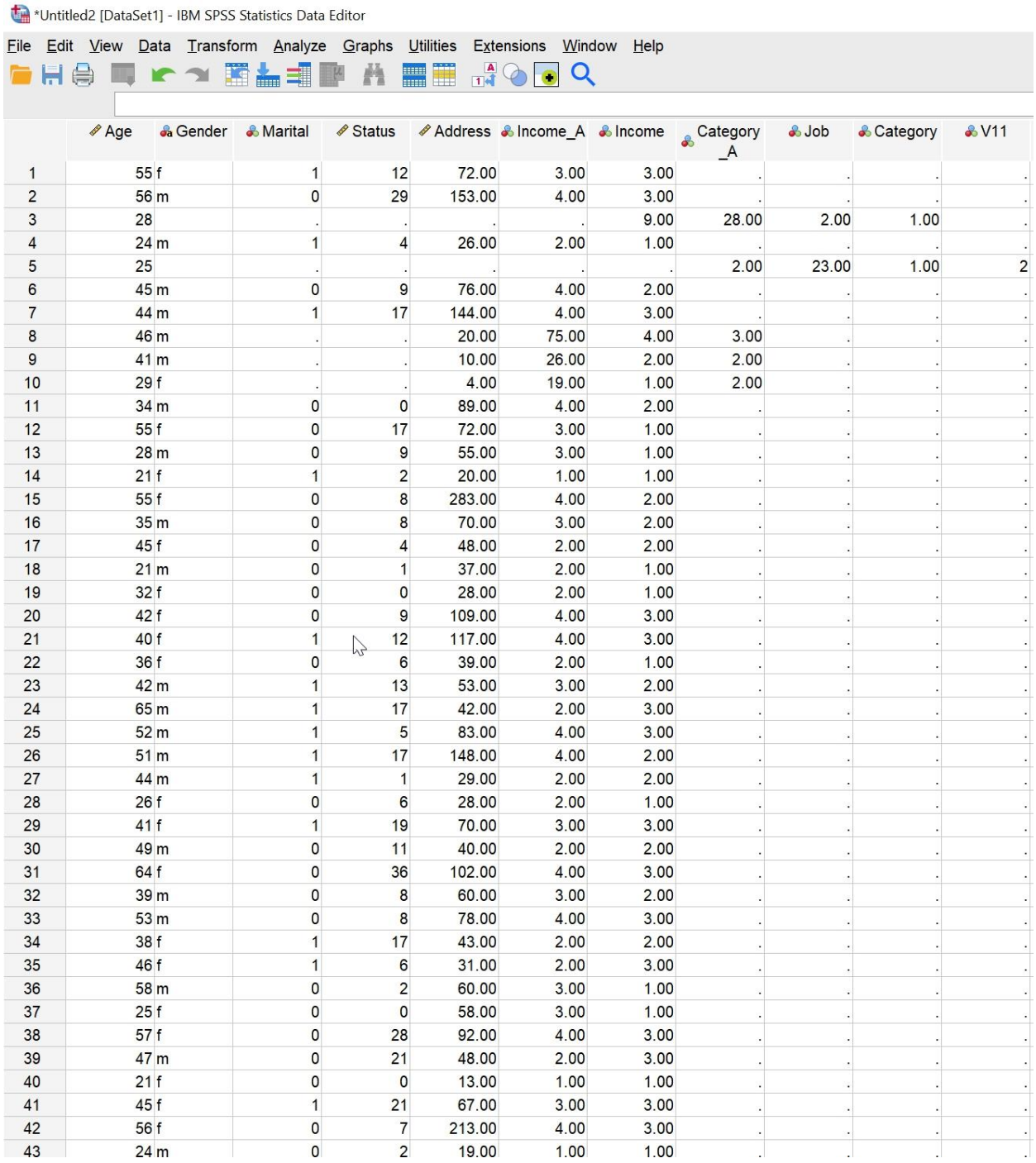
Commands -> Press Finish

Variable View



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Age	Numeric	2	0		None	None	8	Right	Scale	Input
2	Gender	String	1	0		None	None	7	Left	Nominal	Input
3	Marital	Numeric	1	0		None	None	8	Right	Nominal	Input
4	Status	Numeric	2	0		None	None	8	Right	Scale	Input
5	Address	Numeric	6	2		None	None	8	Right	Scale	Input
6	Income_A	Numeric	5	2		None	None	8	Right	Nominal	Input
7	Income	Numeric	4	2		None	None	8	Right	Nominal	Input
8	Category_A	Numeric	5	2		None	None	8	Right	Nominal	Input
9	Job	Numeric	5	2		None	None	8	Right	Nominal	Input
10	Category	Numeric	4	2		None	None	8	Right	Nominal	Input
11	V11	Numeric	1	0		None	None	8	Right	Nominal	Input
12											
13											

Data View



	Age	Gender	Marital	Status	Address	Income_A	Income	Category_A	Job	Category	V11
1	55	f	1	12	72.00	3.00	3.00	.	.	.	.
2	56	m	0	29	153.00	4.00	3.00	.	.	.	.
3	28		.	.	.	.	9.00	28.00	2.00	1.00	.
4	24	m	1	4	26.00	2.00	1.00	.	.	.	.
5	25		.	.	.	.	.	2.00	23.00	1.00	2
6	45	m	0	9	76.00	4.00	2.00	.	.	.	.
7	44	m	1	17	144.00	4.00	3.00	.	.	.	.
8	46	m	.	.	20.00	75.00	4.00	3.00	.	.	.
9	41	m	.	.	10.00	26.00	2.00	2.00	.	.	.
10	29	f	.	.	4.00	19.00	1.00	2.00	.	.	.
11	34	m	0	0	89.00	4.00	2.00	.	.	.	.
12	55	f	0	17	72.00	3.00	1.00	.	.	.	.
13	28	m	0	9	55.00	3.00	1.00	.	.	.	.
14	21	f	1	2	20.00	1.00	1.00	.	.	.	.
15	55	f	0	8	283.00	4.00	2.00	.	.	.	.
16	35	m	0	8	70.00	3.00	2.00	.	.	.	.
17	45	f	0	4	48.00	2.00	2.00	.	.	.	.
18	21	m	0	1	37.00	2.00	1.00	.	.	.	.
19	32	f	0	0	28.00	2.00	1.00	.	.	.	.
20	42	f	0	9	109.00	4.00	3.00	.	.	.	.
21	40	f	1	12	117.00	4.00	3.00	.	.	.	.
22	36	f	0	6	39.00	2.00	1.00	.	.	.	.
23	42	m	1	13	53.00	3.00	2.00	.	.	.	.
24	65	m	1	17	42.00	2.00	3.00	.	.	.	.
25	52	m	1	5	83.00	4.00	3.00	.	.	.	.
26	51	m	1	17	148.00	4.00	2.00	.	.	.	.
27	44	m	1	1	29.00	2.00	2.00	.	.	.	.
28	26	f	0	6	28.00	2.00	1.00	.	.	.	.
29	41	f	1	19	70.00	3.00	3.00	.	.	.	.
30	49	m	0	11	40.00	2.00	2.00	.	.	.	.
31	64	f	0	36	102.00	4.00	3.00	.	.	.	.
32	39	m	0	8	60.00	3.00	2.00	.	.	.	.
33	53	m	0	8	78.00	4.00	3.00	.	.	.	.
34	38	f	1	17	43.00	2.00	2.00	.	.	.	.
35	46	f	1	6	31.00	2.00	3.00	.	.	.	.
36	58	m	0	2	60.00	3.00	1.00	.	.	.	.
37	25	f	0	0	58.00	3.00	1.00	.	.	.	.
38	57	f	0	28	92.00	4.00	3.00	.	.	.	.
39	47	m	0	21	48.00	2.00	3.00	.	.	.	.
40	21	f	0	0	13.00	1.00	1.00	.	.	.	.
41	45	f	1	21	67.00	3.00	3.00	.	.	.	.
42	56	f	0	7	213.00	4.00	3.00	.	.	.	.
43	24	m	0	2	19.00	1.00	1.00	.	.	.	.

CONCLUSION

- Data from files in excel or text format can be transported to SPSS editor window

PRECAUTIONS

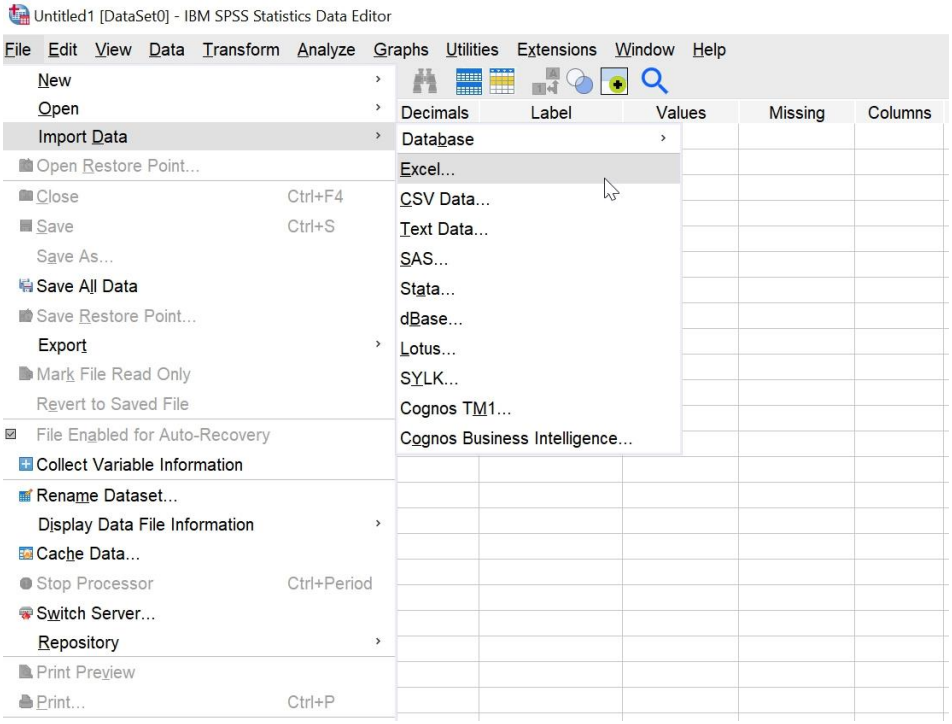
- There should be proper delimiter between data entry in text file
- Extensions of the files should be strictly taken care of

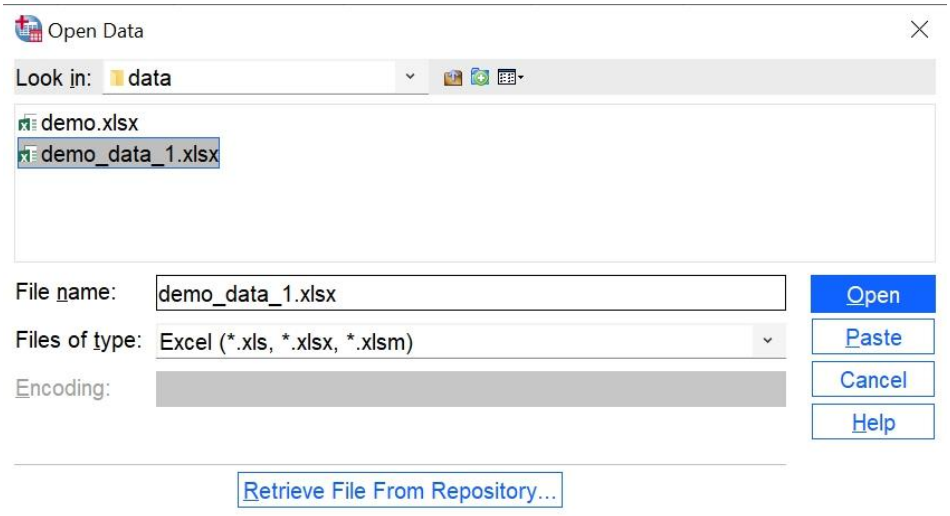
Experiment - 2

Merging of Dataset and providing missing values

STEPS

Command -> File>Import Data>Excel





Data View

*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	name	maths	physics	english	var	var	var
1	Gauss	48	49	43			
2	Alex	.	35	45			
3	Miyazaki	48	39	30			
4	Maya	40	.	42			
5							
6							
7							
8							
9							

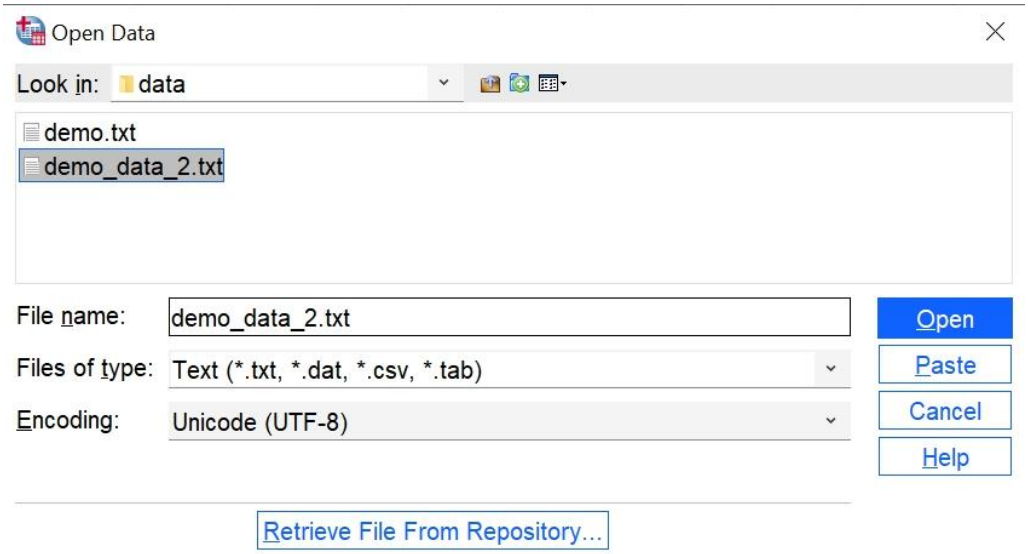
Command -> File>Import Data>Text Data

Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

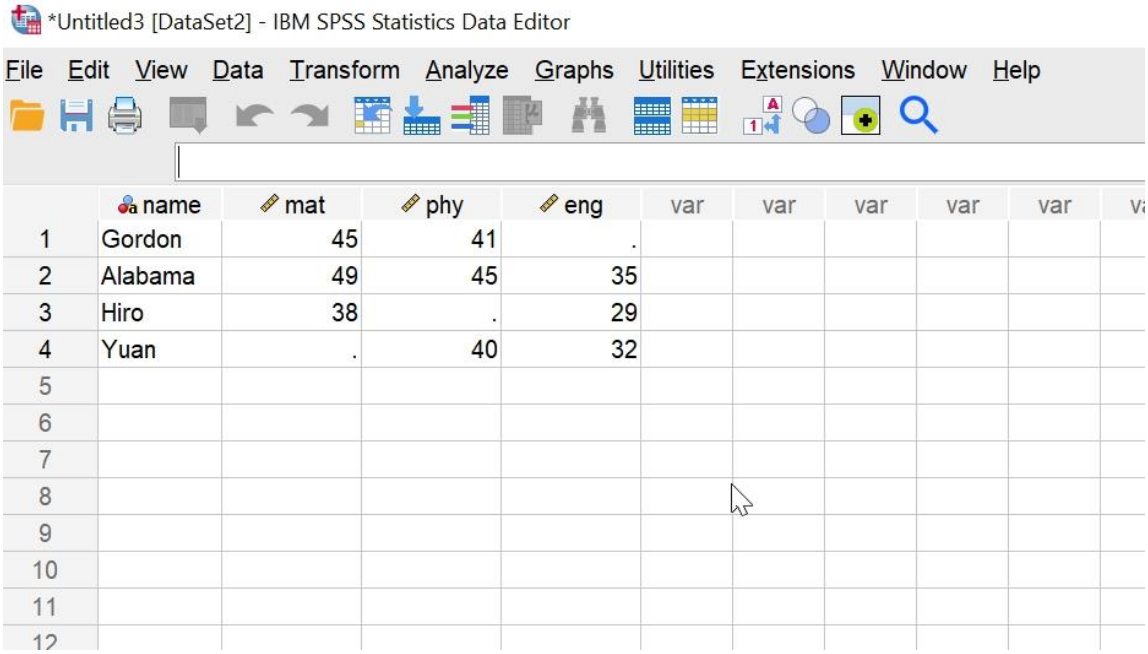
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	Decimals	Label	Values	Missing	Columns

- New
- Open
- Import Data
  - Database
  - Excel...
  - CSV Data...
  - Text Data...
  - SAS...
  - Stata...
  - dBase...
  - Lotus...
  - SYLK...
  - Cognos TM1...
  - Cognos Business Intelligence...
- Open Restore Point...
- Close Ctrl+F4
- Save Ctrl+S
- Save As...
- Save All Data
- Save Restore Point...
- Export
- Mark File Read Only
- Revert to Saved File
- File Enabled for Auto-Recovery
- Collect Variable Information
- Rename Dataset...
- Display Data File Information
- Cache Data...
- Stop Processor Ctrl+Period
- Switch Server...
- Repository
- Print Preview
- Print... Ctrl+P

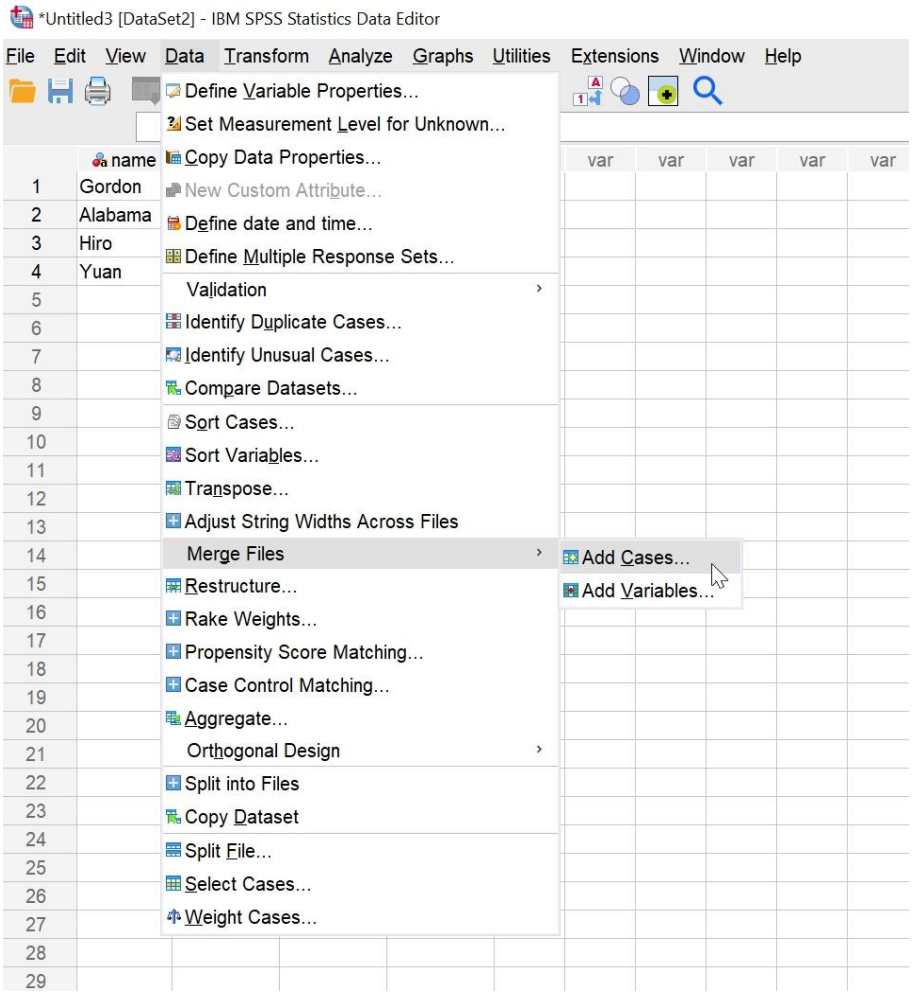


Data View

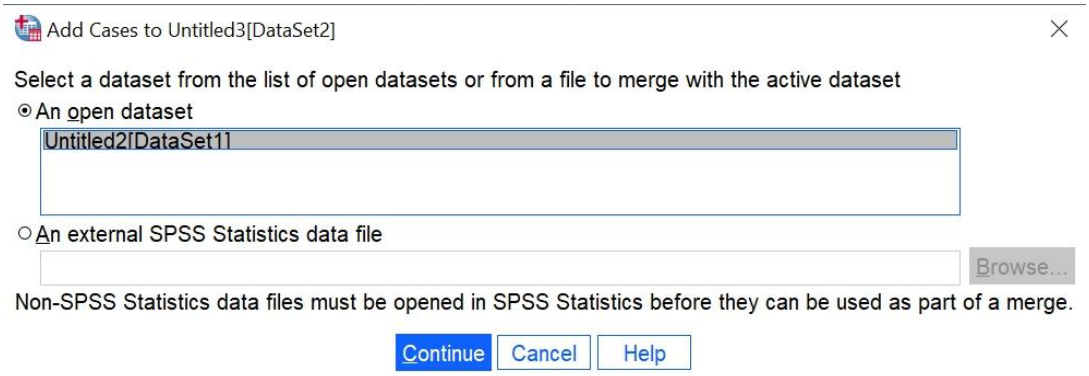


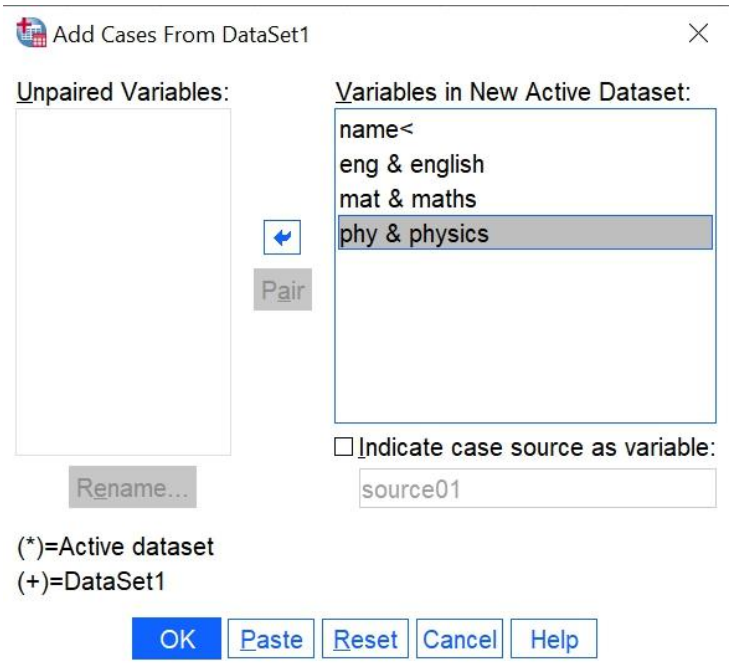
	name	mat	phy	eng	var	var	var	var	var	v
1	Gordon	45	41	.						
2	Alabama	49	45	35						
3	Hiro	38	.	29						
4	Yuan	.	40	32						
5										
6										
7										
8										
9										
10										
11										
12										

Command -> Data>Merge Files>Add Cases



Command -> Click Continue

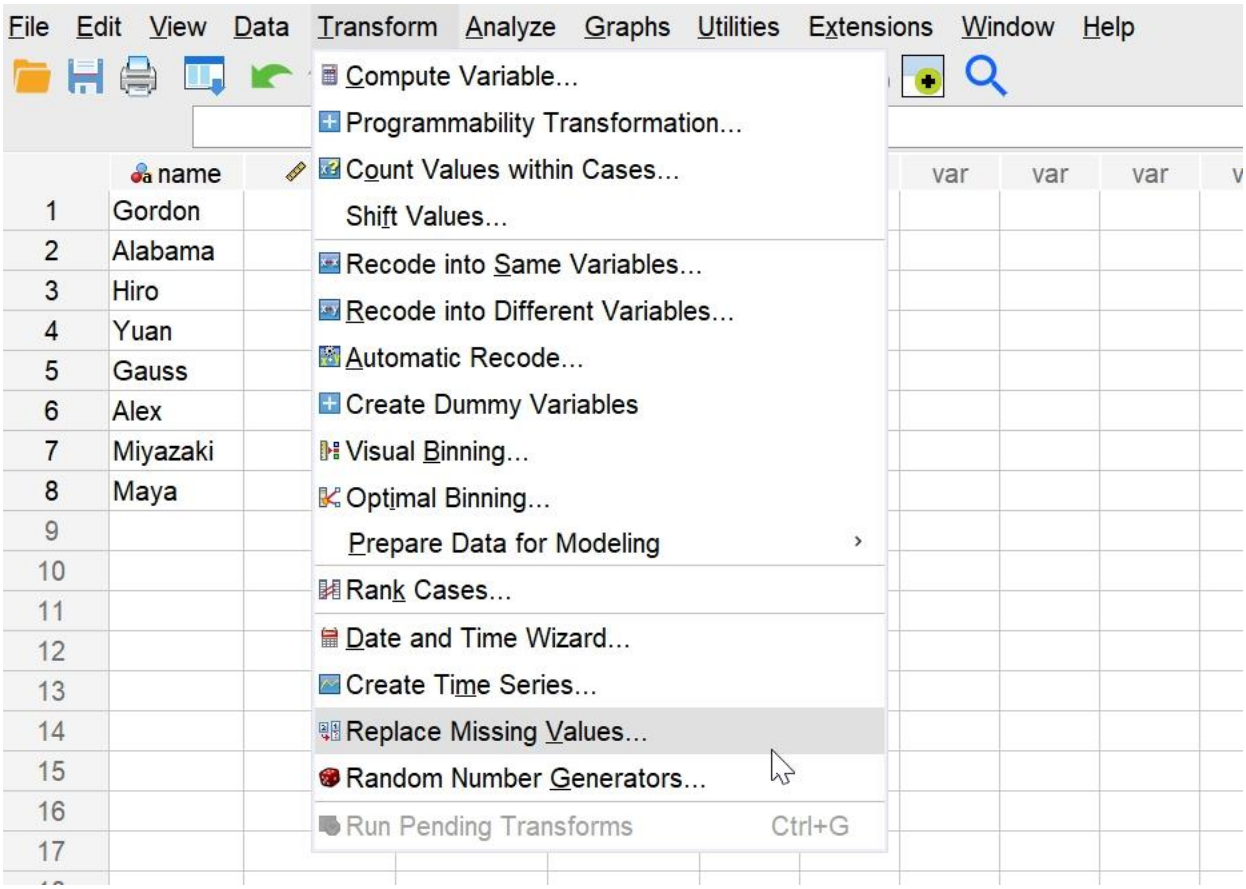


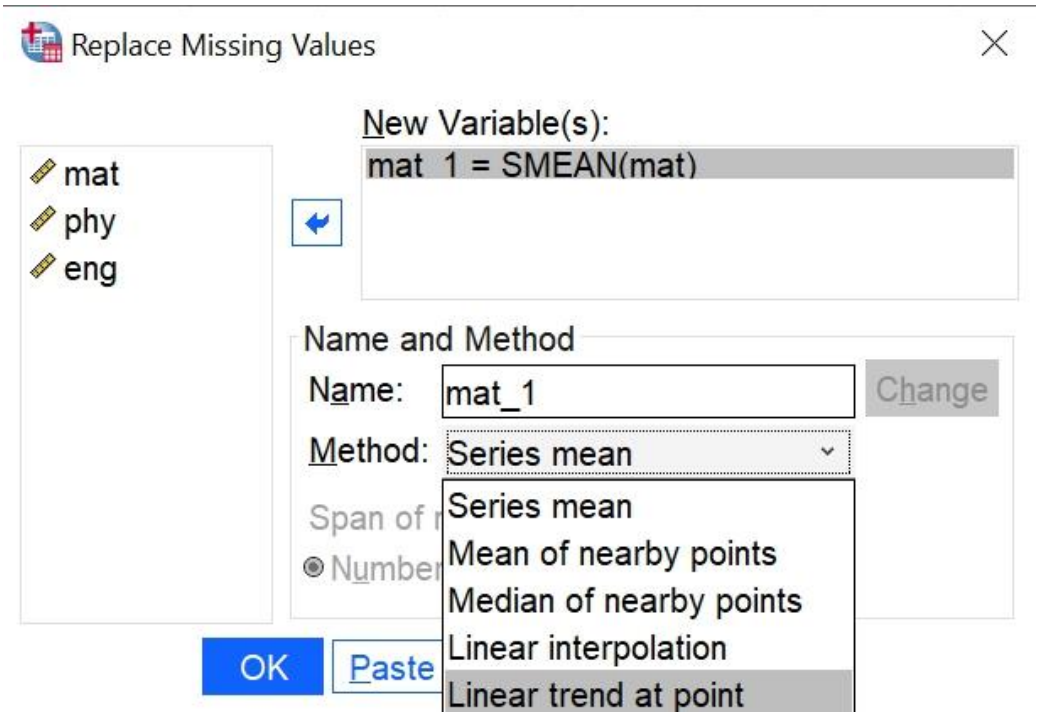


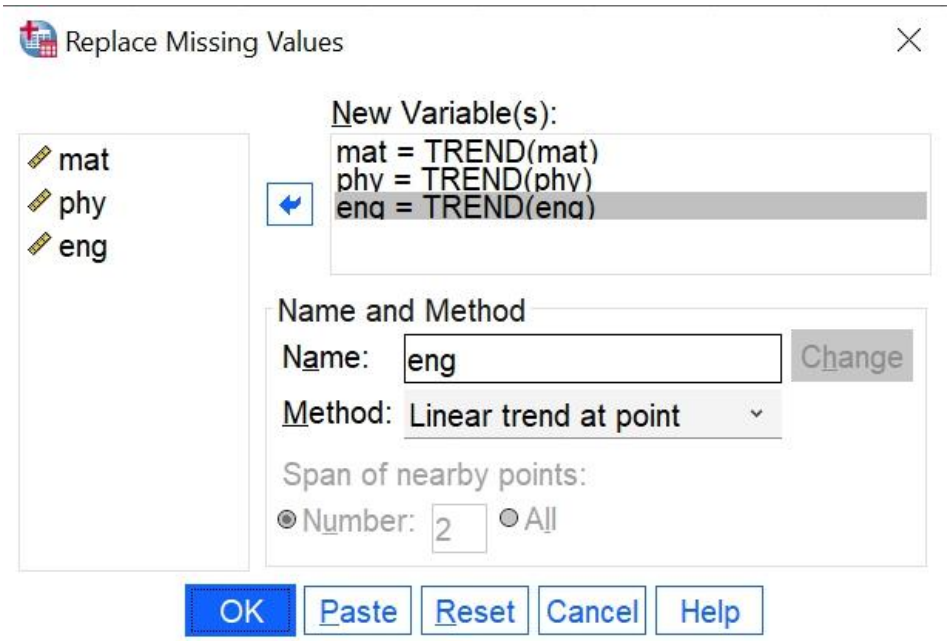
Data View (Some Missing Values)

	name	mat	phy	eng	var
1	Gordon	45	41	.	
2	Alabama	49	45	35	
3	Hiro	38	.	29	
4	Yuan	.	40	32	
5	Gauss	48	49	43	
6	Alex	.	35	45	
7	Miyazaki	48	39	30	
8	Maya	40	.	42	
9					

Command -> Transform>Replace Missing Values







Data View (All Missing Values are replaced)

name	mat	phy	eng
Gordon	45.0	41.0	31.4
Alabama	49.0	45.0	35.0
Hiro	38.0	42.3	29.0
Yuan	44.7	40.0	32.0
Gauss	48.0	49.0	43.0
Alex	44.3	35.0	45.0
Miyazaki	48.0	39.0	30.0
Maya	40.0	38.9	42.0

CONCLUSION

- In this practical we have successfully merged two different types of datasets (excel file and text file) using the “Merge Dataset” feature

- Also, we have filled the missing values in the merged dataset using “Replace Missing Value” feature by using ‘Linear trend at point’ method

Experiment - 3

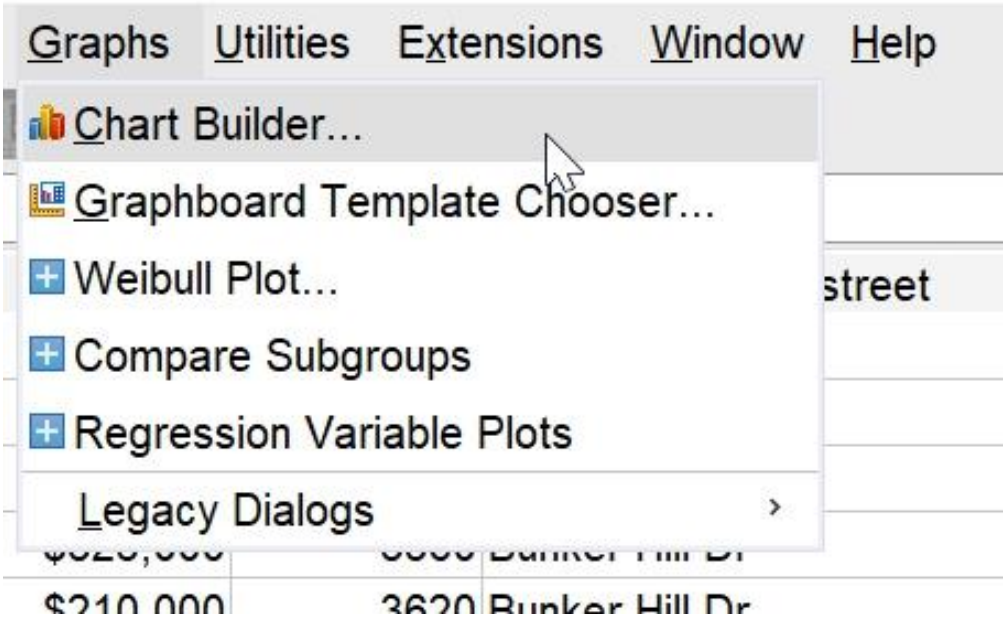
Graphical representation of imported dataset using Bar and Pie charts

STEPS

CASE – 1

- BAR Chart

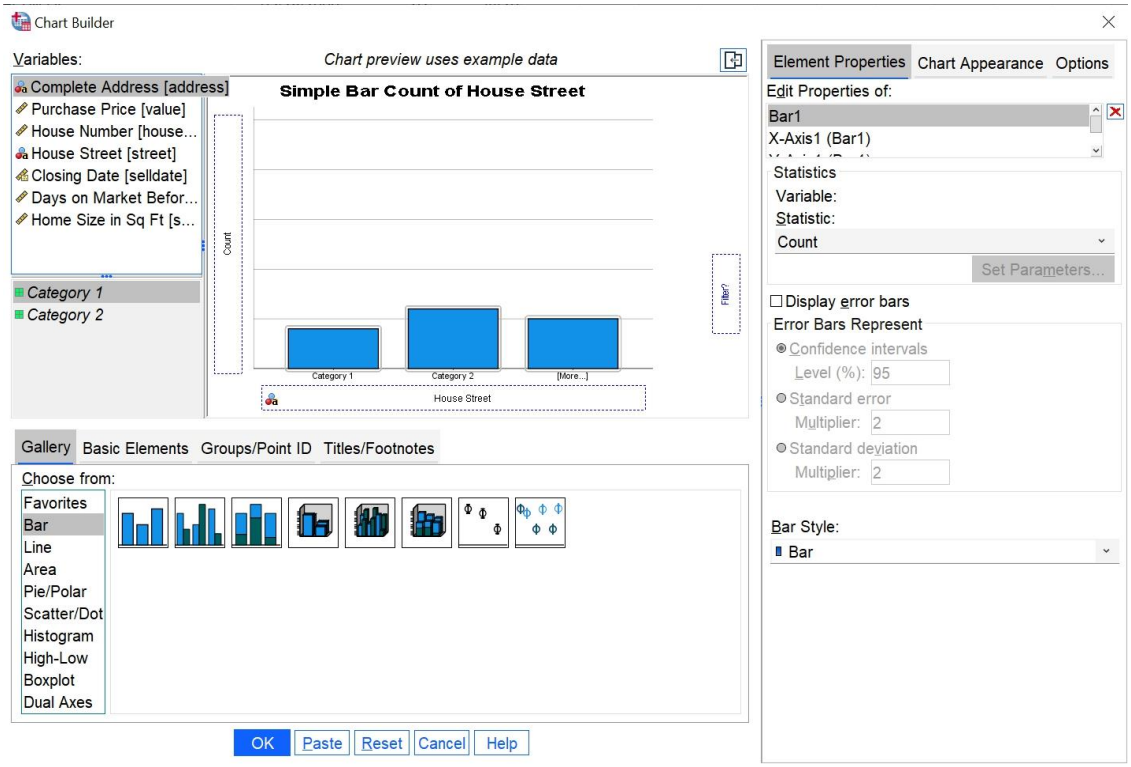
Command -> Graphs>Chart Builder



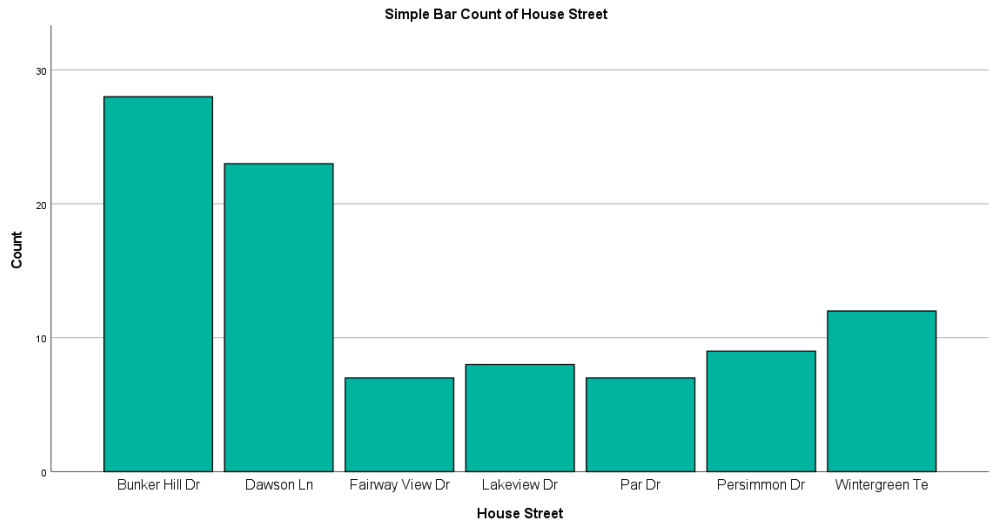
Command -> Drag Bar Chart to Chart Preview Section

Command -> Select Variables>House Street [street] to the X-axis

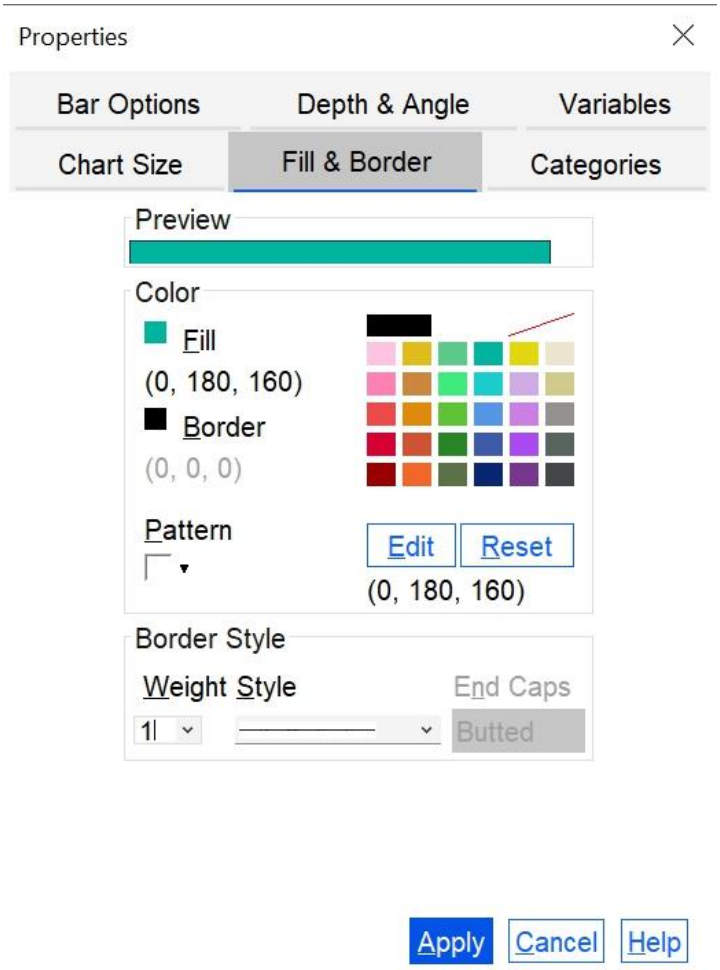
Command -> Select Ok



Bar-Graph



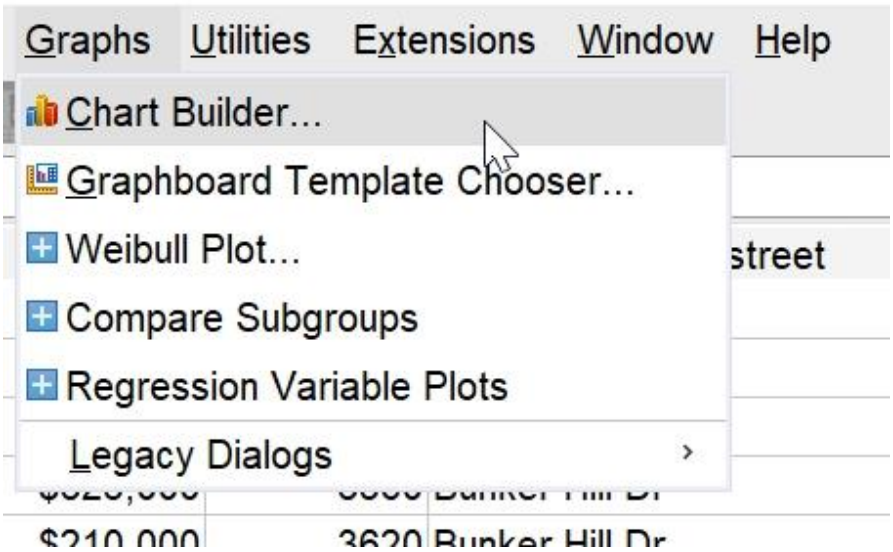
Command –> Double Click Graph to open Properties *settings



CASE – 2

- PIE Chart

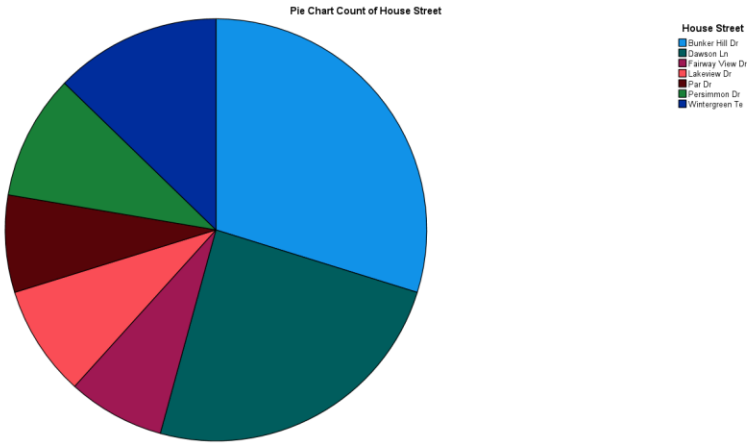
Command –> Graphs>Chart Builder



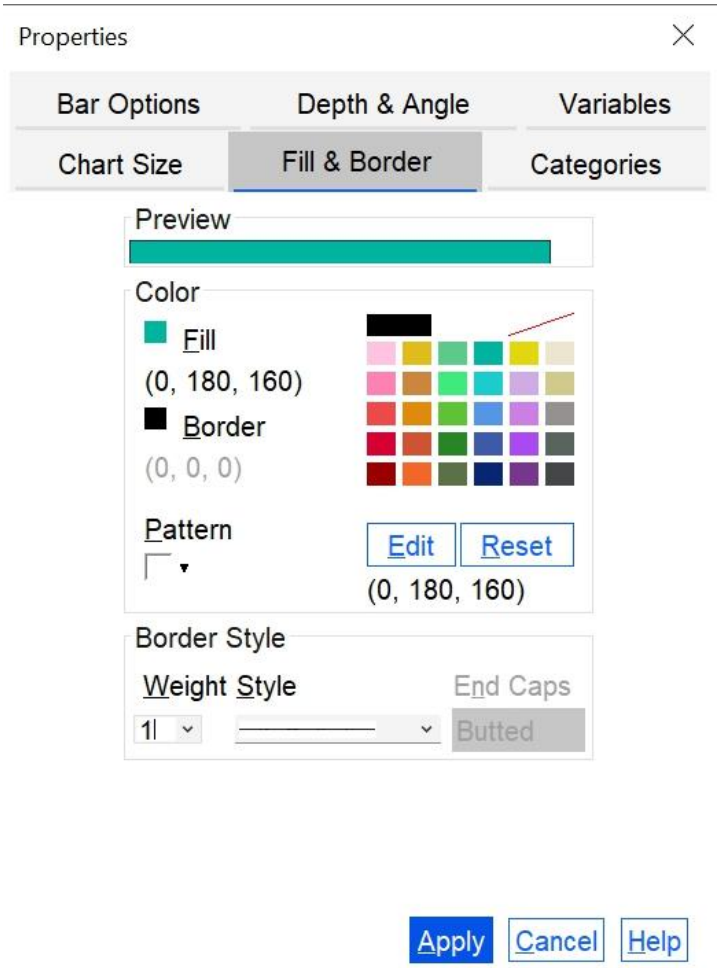
Command -> Drag Pie Chart to Chart Preview Section

Command -> Select Variables>House Street [street] to the X-axis

Command -> Select Ok



Command -> Double Click Graph to open Properties
*settings



CONCLUSION

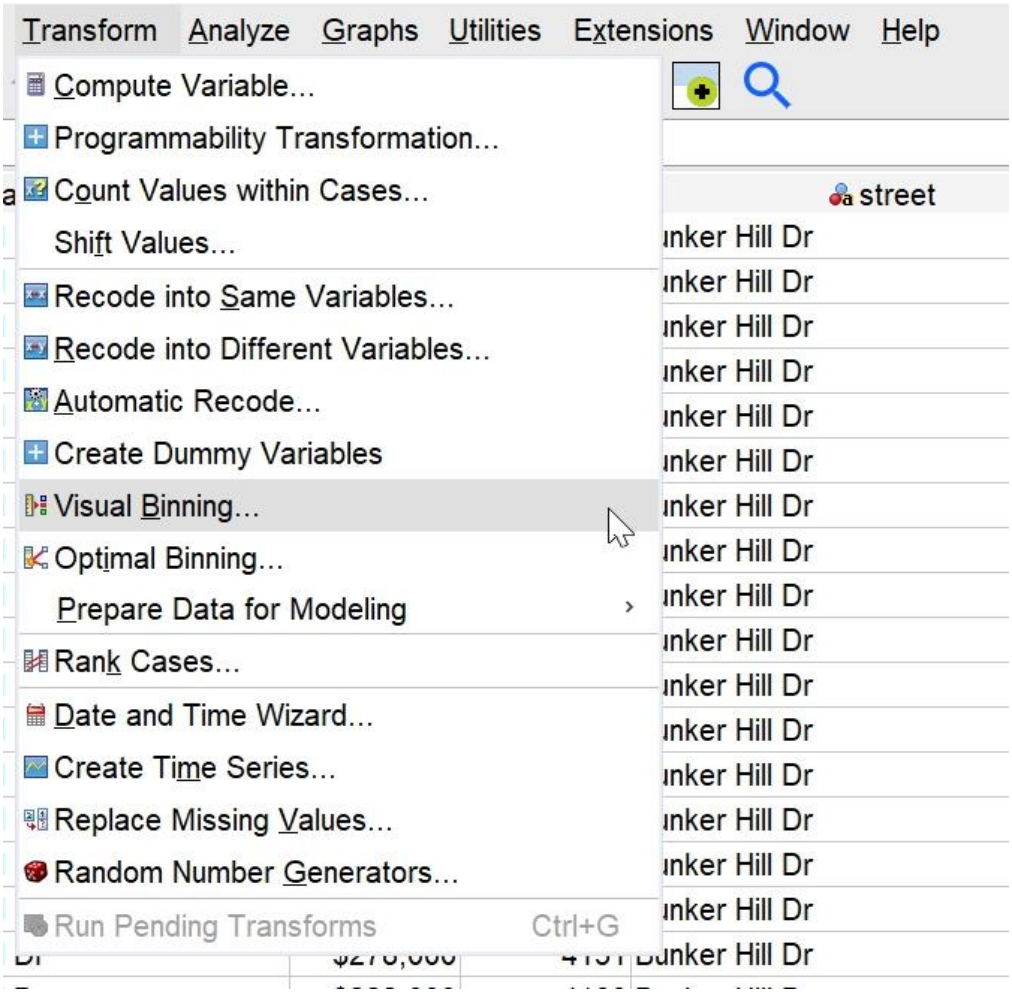
- In this practical we have successfully learnt how to make Bar and Pie Charts using the “Chart Builder” feature
- Also, we have learnt how to edit the ‘Properties’ of the graph i.e. changing border, colors, chart size, etc

Experiment - 4

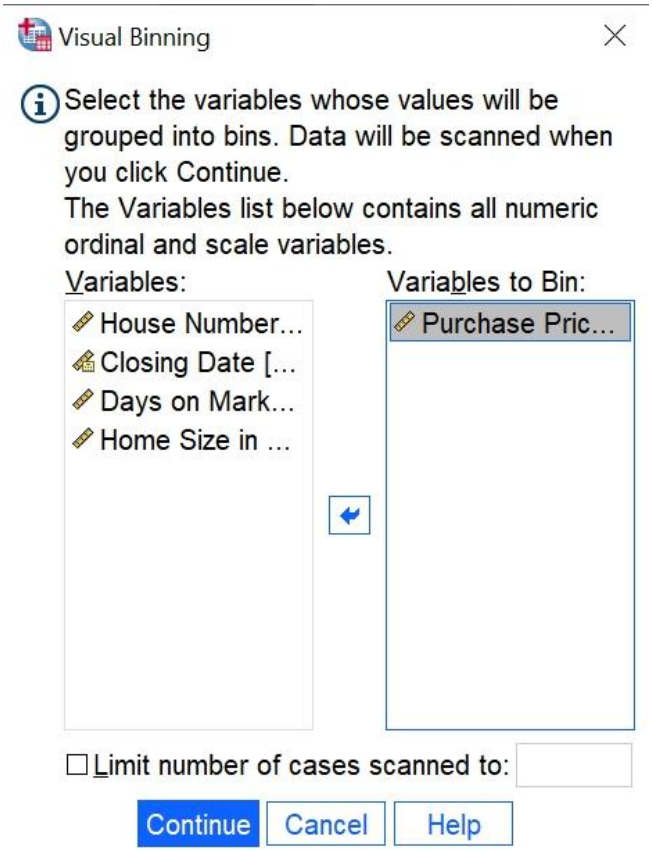
Drawing of Histogram and Distribution Curve

STEPS

Command -> Transform>Visual Binning

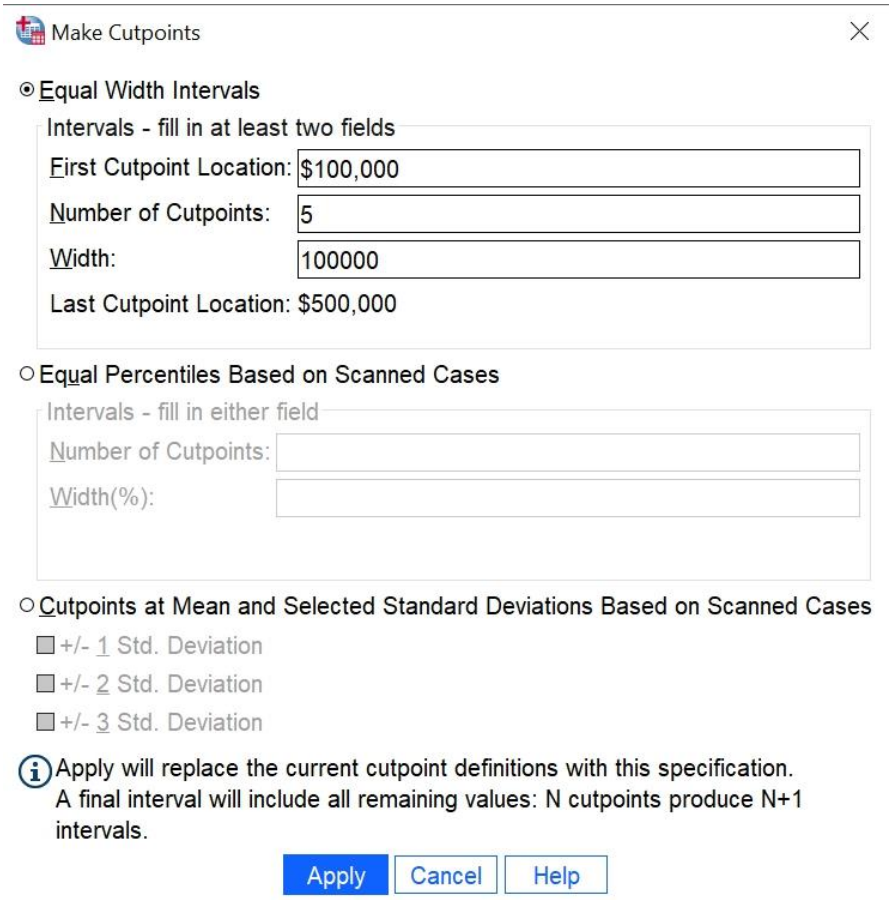


Command -> Click Continue



Command -> Click Make Cutpoints…

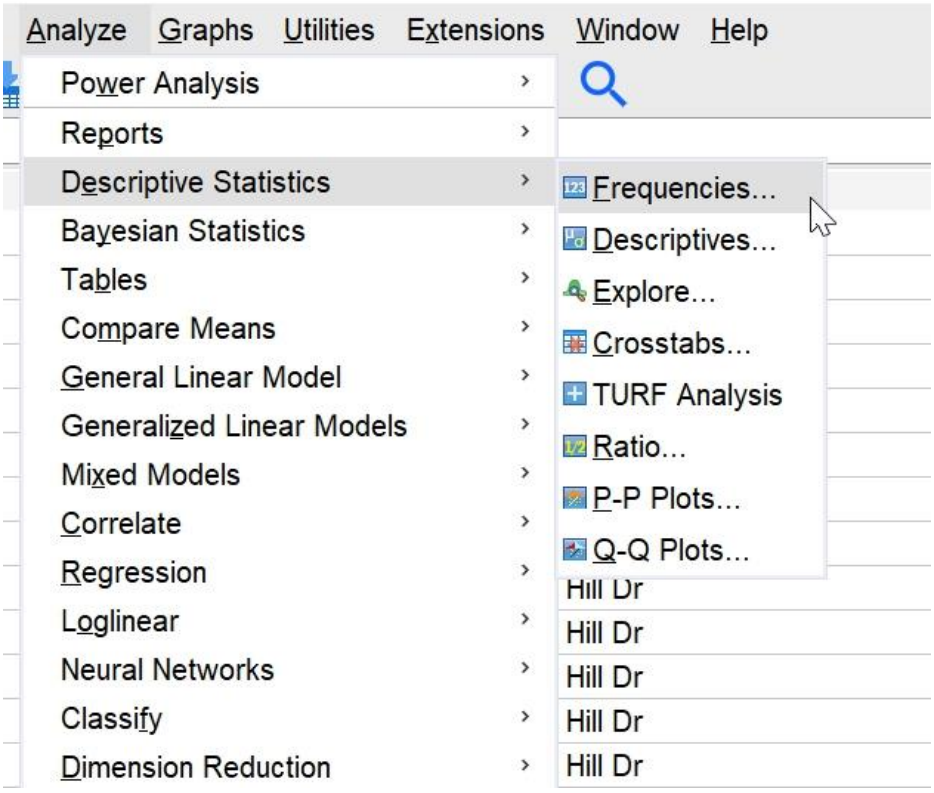
*fill the details as show below



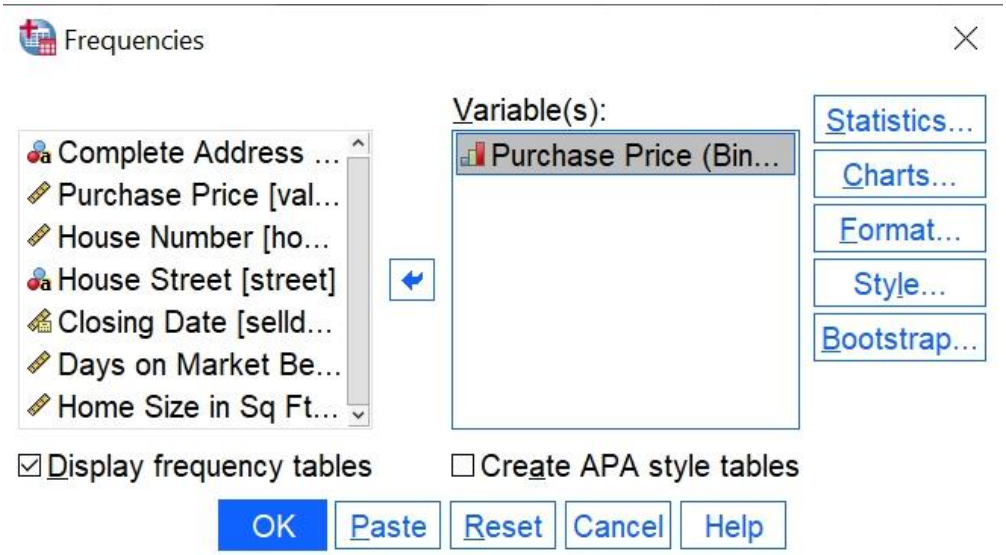
Data View (*new column has been created)

	address	value	housenum	street	selldate	marktime	sqft	price
1	3520 Bunker Hill Dr	$335,000	3520	Bunker Hill Dr	03/13/2000	91	3211	4
2	3521 Bunker Hill Dr	$321,000	3521	Bunker Hill Dr	06/21/2000	54	2810	4
3	3530 Bunker Hill Dr	$300,000	3530	Bunker Hill Dr	01/25/2000	13	3295	3
4	3560 Bunker Hill Dr	$325,000	3560	Bunker Hill Dr	01/01/2000	96	3034	4
5	3620 Bunker Hill Dr	$210,000	3620	Bunker Hill Dr	07/24/2000	18	2265	3
6	3621 Bunker Hill Dr	$416,000	3621	Bunker Hill Dr	03/28/2000	62	4420	5
7	3630 Bunker Hill Dr	$342,000	3630	Bunker Hill Dr	11/15/1999	18	4237	4
8	3631 Bunker Hill Dr	$347,000	3631	Bunker Hill Dr	06/13/2000	133	3933	4
9	3651 Bunker Hill Dr	$284,000	3651	Bunker Hill Dr	03/29/2000	103	2688	3
10	3671 Bunker Hill Dr	$290,000	3671	Bunker Hill Dr	01/01/2000	104	2790	3
11	3700 Bunker Hill Dr	$294,000	3700	Bunker Hill Dr	07/06/2000	46	3937	3
12	3721 Bunker Hill Dr	$235,000	3721	Bunker Hill Dr	09/08/1999	74	2504	3
13	4020 Bunker Hill Dr	$250,000	4020	Bunker Hill Dr	01/01/2000	10	1993	3
14	4030 Bunker Hill Dr	$290,000	4030	Bunker Hill Dr	07/31/2000	13	2894	3
15	4141 Bunker Hill Dr	$247,000	4141	Bunker Hill Dr	01/01/2000	34	3112	3
16	4150 Bunker Hill Dr	$232,000	4150	Bunker Hill Dr	02/03/2000	15	3127	3
17	4151 Bunker Hill Dr	$278,000	4151	Bunker Hill Dr	07/12/2000	66	2585	3
18	4160 Bunker Hill Dr	$222,000	4160	Bunker Hill Dr	01/01/2000	73	1740	3
19	4161 Bunker Hill Dr	$265,000	4161	Bunker Hill Dr	07/26/2000	55	2347	3
20	4170 Bunker Hill Dr	$300,000	4170	Bunker Hill Dr	08/25/2000	85	2939	3
21	4171 Bunker Hill Dr	$274,000	4171	Bunker Hill Dr	06/16/2000	85	2629	3
22	4180 Bunker Hill Dr	$265,000	4180	Bunker Hill Dr	08/07/2000	61	1622	3
23	4201 Bunker Hill Dr	$254,000	4201	Bunker Hill Dr	02/03/2000	1	2655	3

Command -> Analyze>Descriptive Statistics>Frequencies

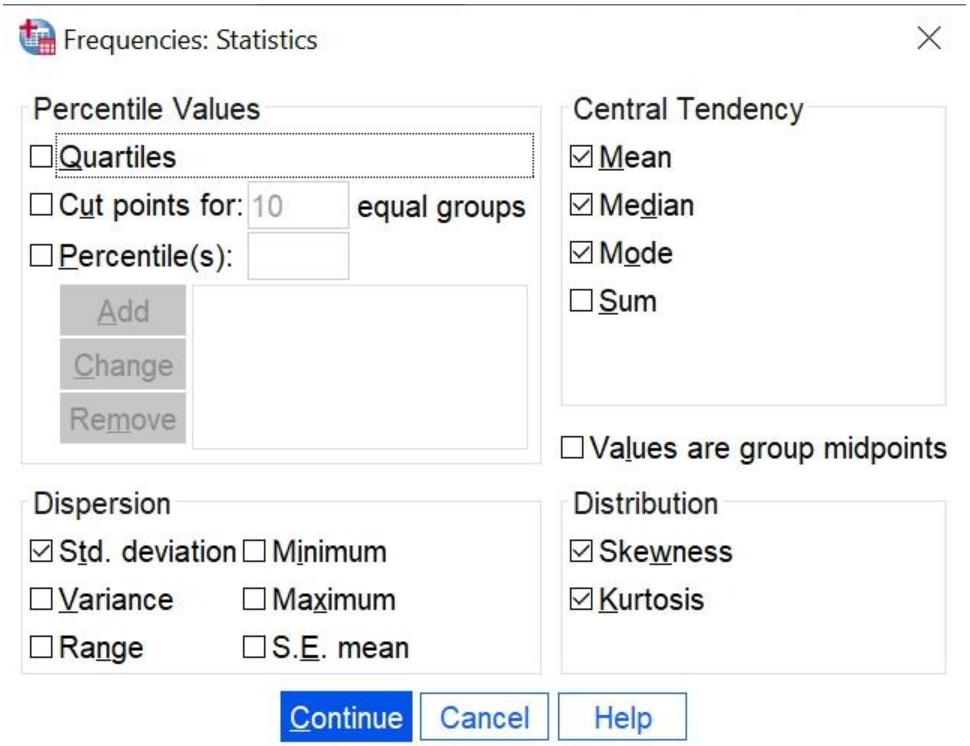


Command -> Click Purchase Price (Binned data)

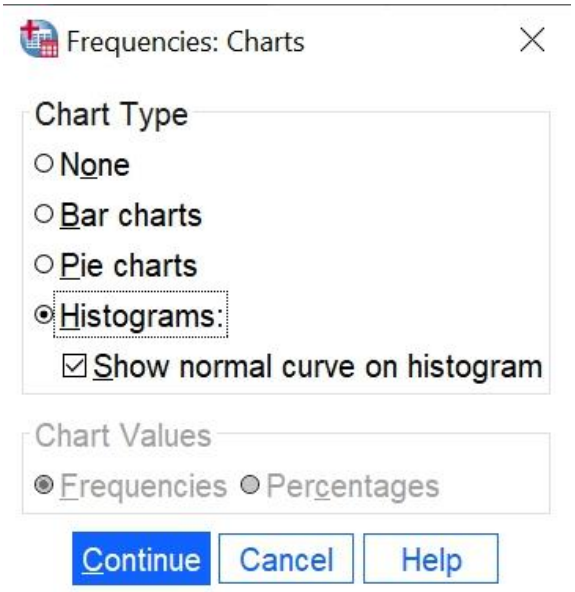


Command -> Click Statistics…

*fill the details as show below



Command -> Click Charts… *check “Histograms” along with “normal curve”



Command -> Click OK

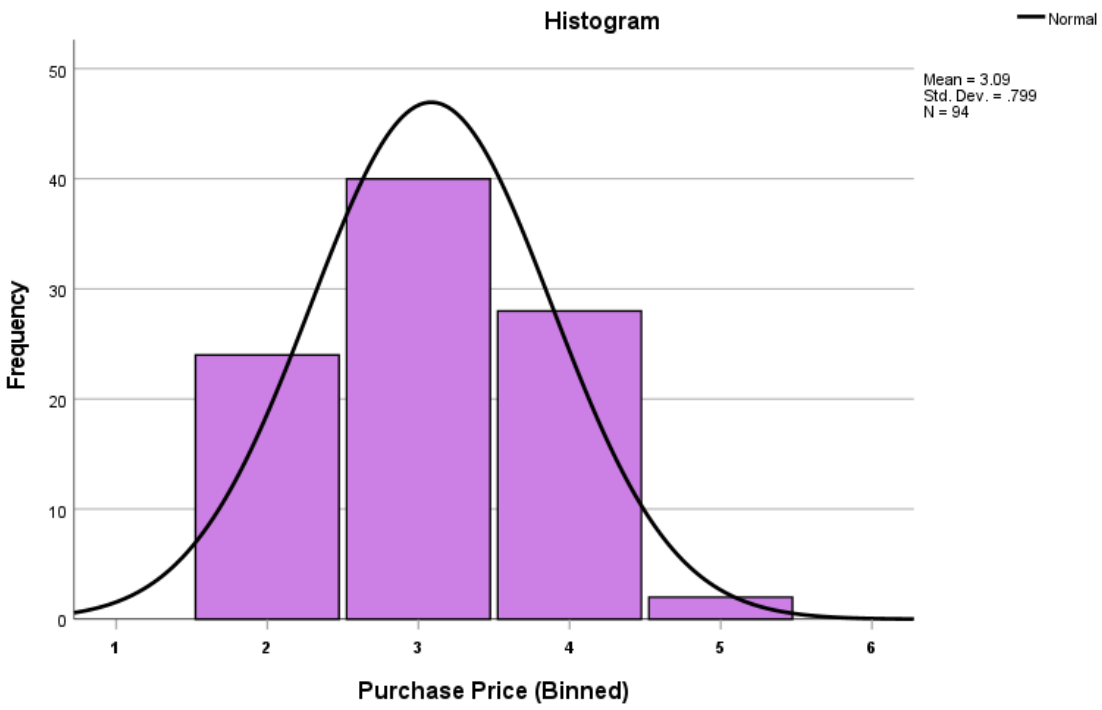
Statistics

Purchase Price (Binned)

N	Valid	94
	Missing	0
Mean		3.09
Median		3.00
Mode		3
Std. Deviation		.799
Skewness		.103
Std. Error of Skewness		.249
Kurtosis		-.850
Std. Error of Kurtosis		.493

Purchase Price (Binned)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	$100,001 - $200,000	24	25.5	25.5	25.5
	$200,001 - $300,000	40	42.6	42.6	68.1
	$300,001 - $400,000	28	29.8	29.8	97.9
	$400,001 - $500,000	2	2.1	2.1	100.0
	Total	94	100.0	100.0	



Experiment - 5

Descriptive statistics like mean, mode, median, variance, skewness, kurtosis, etc.

THEORY

The various statistical constants are to be calculated and interpreted.

$$sample\ mean\ (\bar{x}) = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median

The point in the data set dividing it into two halves is called median.

Mode

It is the value which occurs with the greatest frequency (maybe more than one value).

Variance and standard deviation

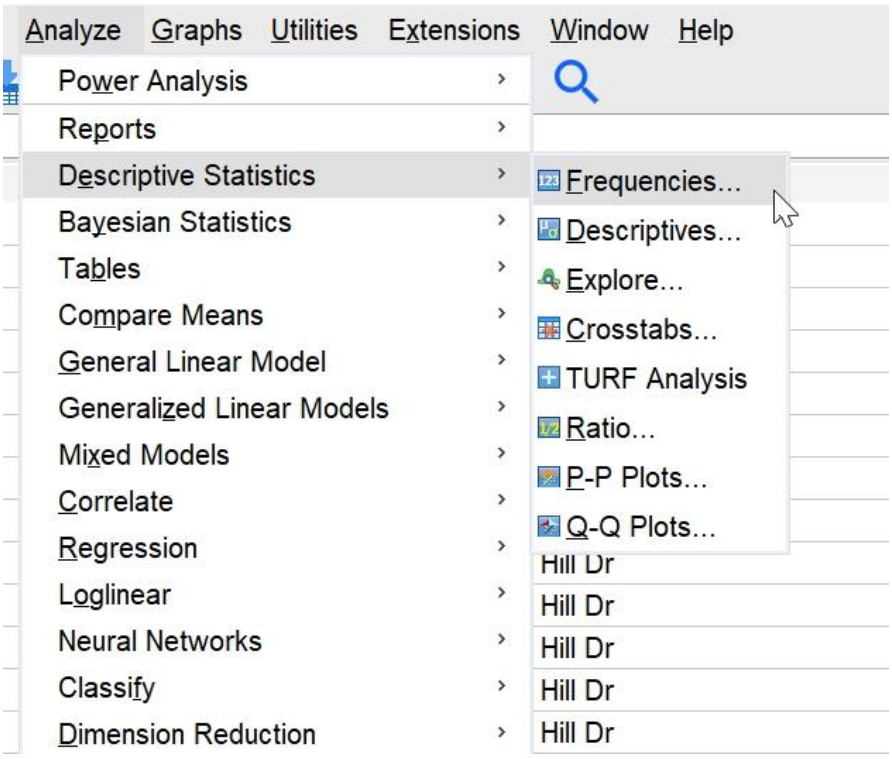
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

STEPS

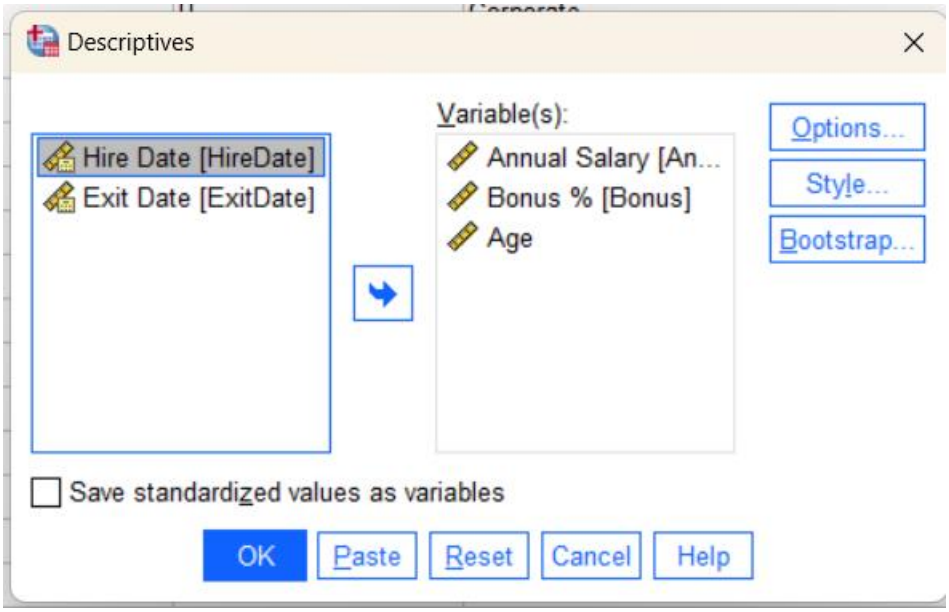
CASE – 1

- Statistical values when all types of cases together

Command -> Analyze>Descriptive Statistics>Descriptives…



Choose the required variable > option and check the desired boxes



Document output window

```
DESCRIPTIVES VARIABLES=AnnualSalary Bonus Age
  /STATISTICS=MEAN STDDEV MIN MAX.
```

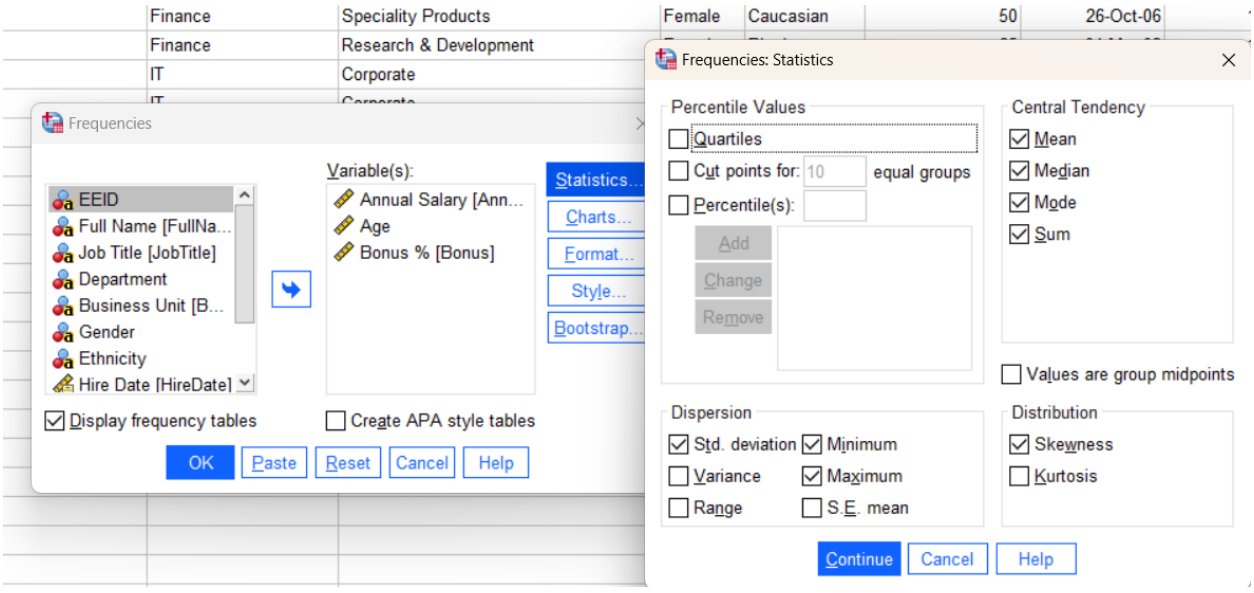
Descriptives

Descriptive Statistics

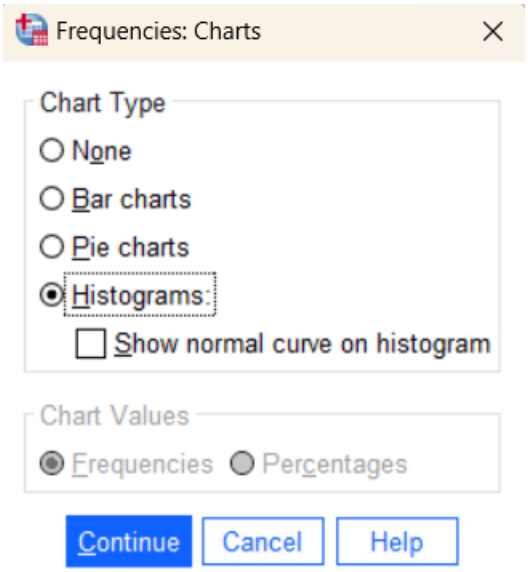
	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	19	41336.0	249270.0	124754.421	50210.0387
Bonus %	19	0.0%	30.0%	9.421%	9.3530%
Age	19	25	65	45.00	14.903
Valid N (listwise)	19				

```
FREQUENCIES VARIABLES=AnnualSalary Age Bonus
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN MEDIAN MODE SUM SKEWNESS SESKEW
  /ORDER=ANALYSIS.
```

Command -> Analyze>Descriptive Statistics>Frequencies…



Command -> Choose the required variable > Histogram > Continue



Document output window

Frequencies

Statistics

		Annual Salary	Age	Bonus %
N	Valid	19	19	19
	Missing	0	0	0
Mean		124754.421	45.00	9.421%
Median		113527.000	50.00	9.000%
Mode		41336.0[a]	27[a]	0.0%
Std. Deviation		50210.0387	14.903	9.3530%
Skewness		.592	-.079	.658
Std. Error of Skewness		.524	.524	.524
Minimum		41336.0	25	0.0%
Maximum		249270.0	65	30.0%
Sum		2370334.0	855	179.0%

a. Multiple modes exist. The smallest value is shown

Frequency Table

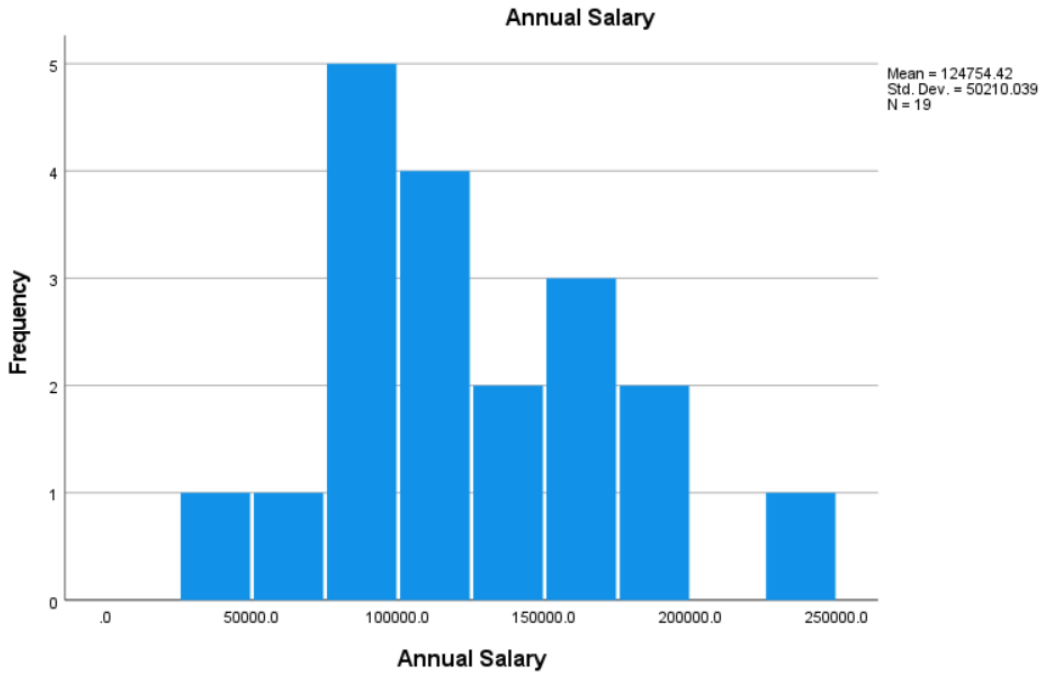
Annual Salary

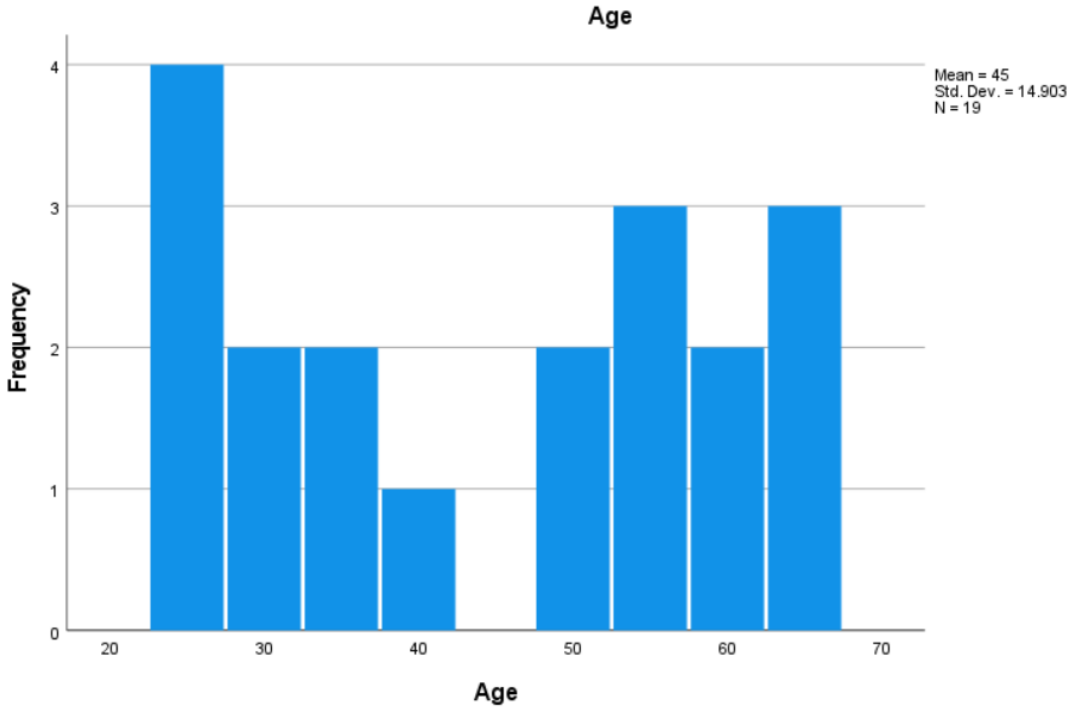
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	41336.0	1	5.3	5.3	5.3
	50994.0	1	5.3	5.3	10.5
	77203.0	1	5.3	5.3	15.8
	84913.0	1	5.3	5.3	21.1
	95409.0	1	5.3	5.3	26.3
	97078.0	1	5.3	5.3	31.6
	99975.0	1	5.3	5.3	36.8
	105086.0	1	5.3	5.3	42.1
	109851.0	1	5.3	5.3	47.4
	113527.0	1	5.3	5.3	52.6
	119746.0	1	5.3	5.3	57.9
	141604.0	1	5.3	5.3	63.2
	146742.0	1	5.3	5.3	68.4
	154828.0	1	5.3	5.3	73.7
	157333.0	1	5.3	5.3	78.9
	163099.0	1	5.3	5.3	84.2
	175837.0	1	5.3	5.3	89.5
	186503.0	1	5.3	5.3	94.7
	249270.0	1	5.3	5.3	100.0
	Total	19	100.0	100.0	

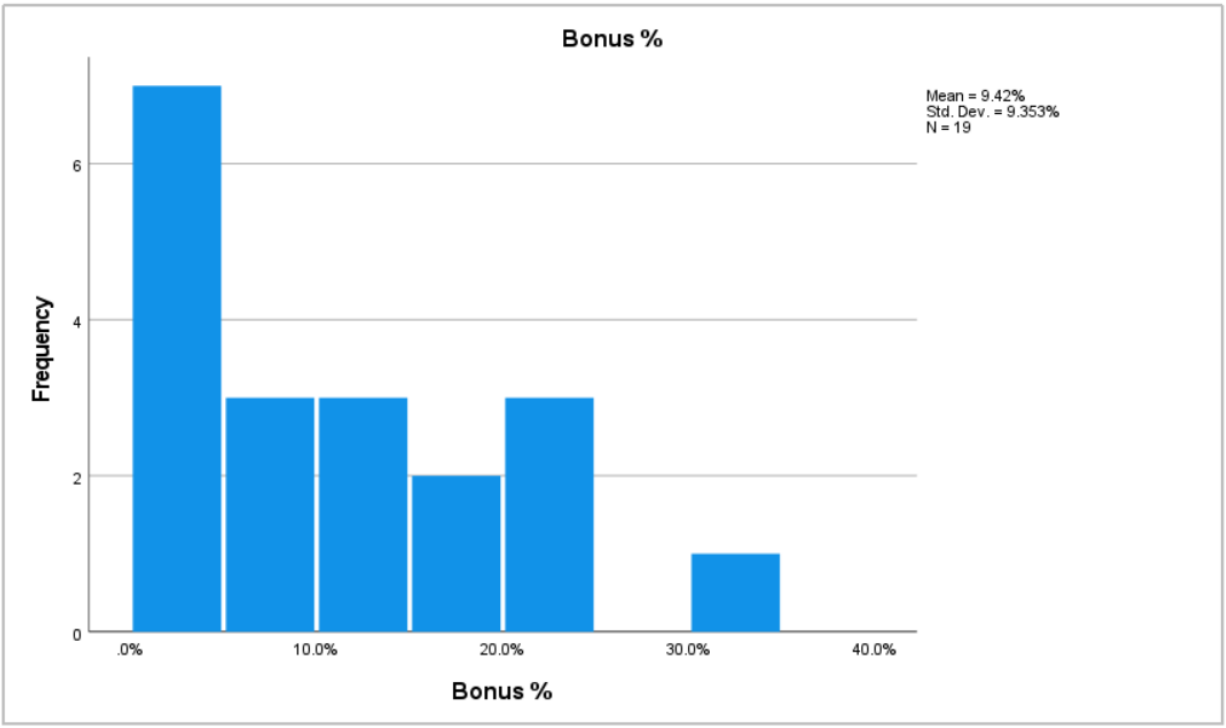
Age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	25	1	5.3	5.3	5.3
	26	1	5.3	5.3	10.5
	27	2	10.5	10.5	21.1
	29	1	5.3	5.3	26.3
	31	1	5.3	5.3	31.6
	34	1	5.3	5.3	36.8
	36	1	5.3	5.3	42.1
	41	1	5.3	5.3	47.4
	50	1	5.3	5.3	52.6
	51	1	5.3	5.3	57.9
	55	2	10.5	10.5	68.4
	57	1	5.3	5.3	73.7
	59	2	10.5	10.5	84.2
	64	2	10.5	10.5	94.7
	65	1	5.3	5.3	100.0
	Total	19	100.0	100.0	

Histogram


Annual Salary
Mean = 124754.42
Std. Dev. = 50210.039
N = 19
Frequency
0
1
2
3
4
5
.0
50000.0
100000.0
150000.0
200000.0
250000.0
Annual Salary


Age
Mean = 45
Std. Dev. = 14.903
N = 19
Frequency
0
1
2
3
4
20
30
40
50
60
70
Age

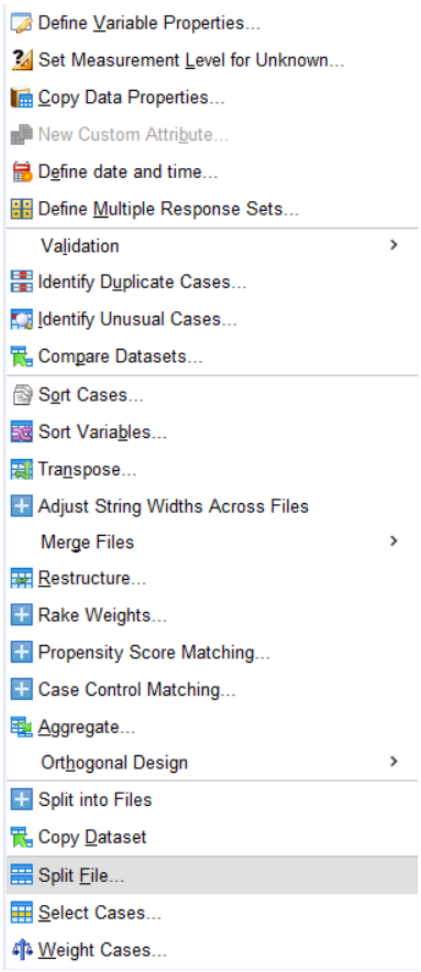


CASE – 2

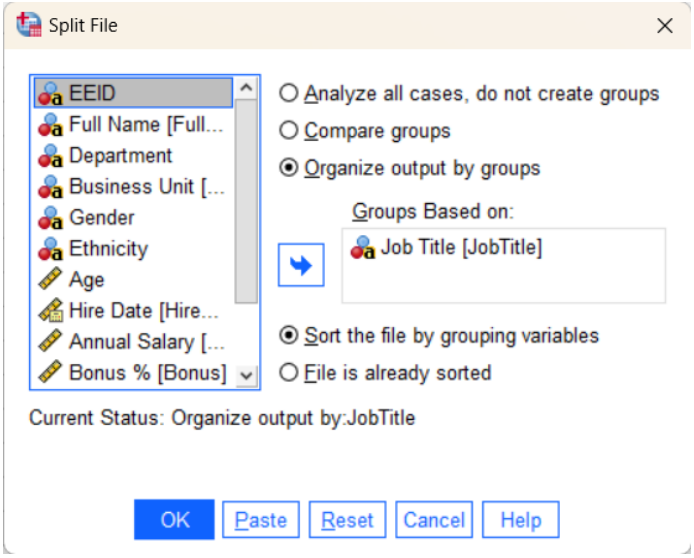
- Statistical values for all type of cases separately.

Suppose we are interested to have the statistical details with respect to all type of “Job Title” separately then follow

Command -> Data>Split file



Select the desired variable (e.g. “Job Title”) for which the splitter information is to be obtained, also select organize output by groups.



Document output window

```
SORT CASES  BY JobTitle.
SPLIT FILE SEPARATE BY JobTitle.
DESCRIPTIVES VARIABLES=AnnualSalary Bonus Age
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

Job Title = Account Representative

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	1	50994.0	50994.0	50994.000	.
Bonus %	1	0.0%	0.0%	0.000%	.
Age	1	57	57	57.00	.
Valid N (listwise)	1				

a. Job Title = Account Representative

Job Title = Analyst

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	1	41336.0	41336.0	41336.000	.
Bonus %	1	0.0%	0.0%	0.000%	.
Age	1	25	25	25.00	.
Valid N (listwise)	1				

a. Job Title = Analyst

Job Title = Computer Systems Manager

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	1	84913.0	84913.0	84913.000	.
Bonus %	1	7.0%	7.0%	7.000%	.
Age	1	26	26	26.00	.
Valid N (listwise)	1				

a. Job Title = Computer Systems Manager

Job Title = Controls Engineer

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	1	109851.0	109851.0	109851.000	.
Bonus %	1	0.0%	0.0%	0.000%	.
Age	1	27	27	27.00	.
Valid N (listwise)	1				

a. Job Title = Controls Engineer

Double-click to activate

Job Title = Director

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	3	163099.0	186503.0	175146.333	11717.2765
Bonus %	3	20.0%	24.0%	21.333%	2.3094%
Age	3	50	65	59.67	8.386
Valid N (listwise)	3				

a. Job Title = Director

Job Title = Manager

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	3	105086.0	119746.0	112786.333	7358.0120
Bonus %	3	6.0%	10.0%	8.333%	2.0817%
Age	3	27	59	38.33	17.926
Valid N (listwise)	3				

a. Job Title = Manager

Job Title = Sr. Analyst

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	3	77203.0	97078.0	89896.667	11024.6664
Bonus %	3	0.0%	0.0%	0.000%	0.0000%
Age	3	31	55	40.00	13.077
Valid N (listwise)	3				

a. Job Title = Sr. Analyst

Job Title = Sr. Manger

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	4	141604.0	157333.0	150126.750	7260.0641
Bonus %	4	10.0%	15.0%	13.250%	2.3629%
Age	4	36	64	51.50	11.676
Valid N (listwise)	4				

a. Job Title = Sr. Manger

Job Title = Technical Architect

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	1	99975.0	99975.0	99975.000	.
Bonus %	1	0.0%	0.0%	0.000%	.
Age	1	59	59	59.00	.
Valid N (listwise)	1				

a. Job Title = Technical Architect

Job Title = Vice President

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Annual Salary	1	249270.0	249270.0	249270.000	.
Bonus %	1	30.0%	30.0%	30.000%	.
Age	1	41	41	41.00	.
Valid N (listwise)	1				

a. Job Title = Vice President

CONCLUSION

- From this practical we have learnt to how to analyze data using “Descriptive Statistics like mean, mode, median, variance, skewness, kurtosis, etc.
- Also we have tackled two cases (cases together and cases separately) by using the “Data>Split File” features to analyze data together as well separated on some particular basis (here “Job Title”)

Experiment - 6

Correlation between two Random Variables

Theory and Methodology

As a measure of intensity or degree of linear relationship between two variables Karl Pearson coefficient of correlation denoted as r_{xy} or $r(X,Y)$ between two random variables X & Y s defined as

If (x_i, y_i) , $i = 1,2, \dots n$ is the bivariate distribution, then

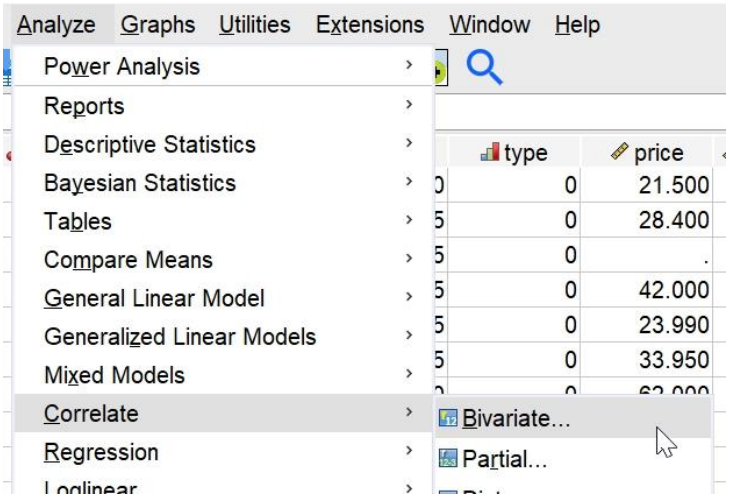
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

Open the desired file in SPSS editor File->Open->file path

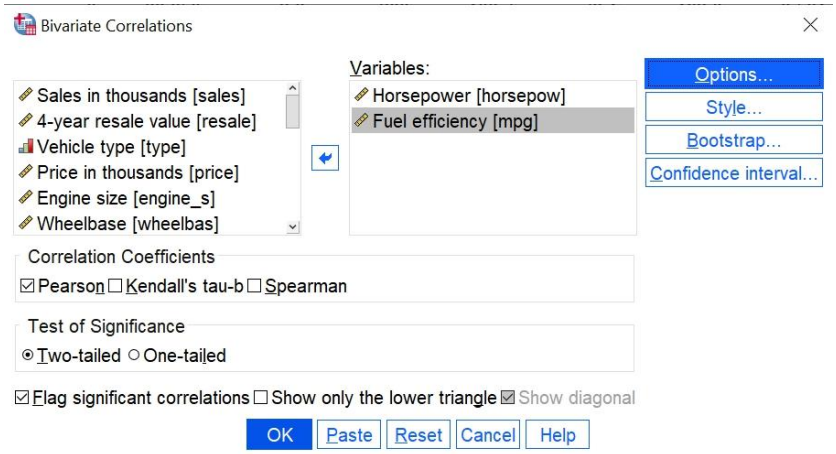
(car_sales.sav from repository is used and the variables Horsepower of the engine & Fuel efficiency is considered)

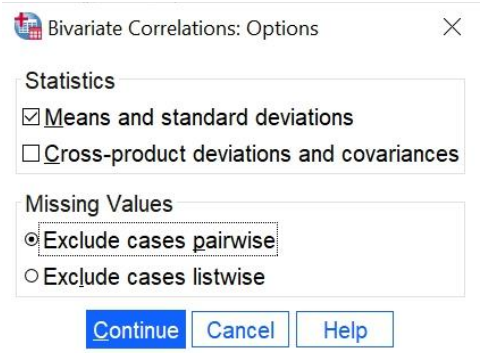
Case1. Missing values dealt as Exclude cases pairwise

Now follow the path Analyze->Correlate->Bivariate



Click option and check the desired boxes.





Document Output Window

Descriptive Statistics

	Mean	Std. Deviation	N
Horsepower	185.95	56.700	156
Fuel efficiency	23.84	4.283	154

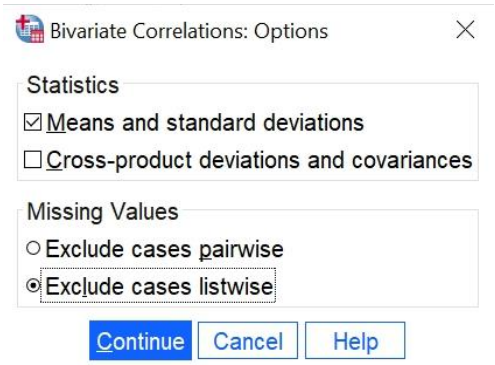
Correlations

		Horsepower	Fuel efficiency
Horsepower	Pearson Correlation	1	-.611**
	Sig. (2-tailed)		<.001
	N	156	154
Fuel efficiency	Pearson Correlation	-.611**	1
	Sig. (2-tailed)	<.001	
	N	154	154

**. Correlation is significant at the 0.01 level (2-tailed).

Case2. Missing values dealt as Exclude cases list wise

Now follow the path Analyze->Correlate->Bivariate



Document Output Window

Descriptive Statistics

	Mean	Std. Deviation	N
Horsepower	185.66	57.006	154
Fuel efficiency	23.84	4.283	154

Correlations[b]

		Horsepower	Fuel efficiency
Horsepower	Pearson Correlation	1	-.611**
	Sig. (2-tailed)		<.001
Fuel efficiency	Pearson Correlation	-.611**	1
	Sig. (2-tailed)	<.001	

**. Correlation is significant at the 0.01 level (2-tailed).

b. Listwise N=154

Case2. Missing values before calculation of correlation*

*(Note : if the missing values are dealt before finding the correlation by series mean, then the total number of cases has become 157)

Document Output Window

Descriptive Statistics

	Mean	Std. Deviation	N
Horsepower	186.74	57.379	157
Fuel efficiency	23.86	4.269	157

Correlations

		Horsepower	Fuel efficiency
Horsepower	Pearson Correlation	1	-.584**
	Sig. (2-tailed)		<.001
	N	157	157
Fuel efficiency	Pearson Correlation	-.584**	1
	Sig. (2-tailed)	<.001	
	N	157	157

**. Correlation is significant at the 0.01 level (2-tailed).

Experiment - 7

Regression Analysis

Theory and Methodology

Let (x_i, y_i) , $i = 1,2, \ldots n$ be a bivariate sample, assuming one of them Independent say(x) and other dependent on first say(y), we predict the value of y by fitting a curve to data. If the curve is line then it is called line of regression and there is said to be a linear regression.

The curve fit for Y on x is called regression curve of y on x and similarly curve fit for X on y is called regression curve of x on y

Let the line of regression of $Y\ on\ x$ be

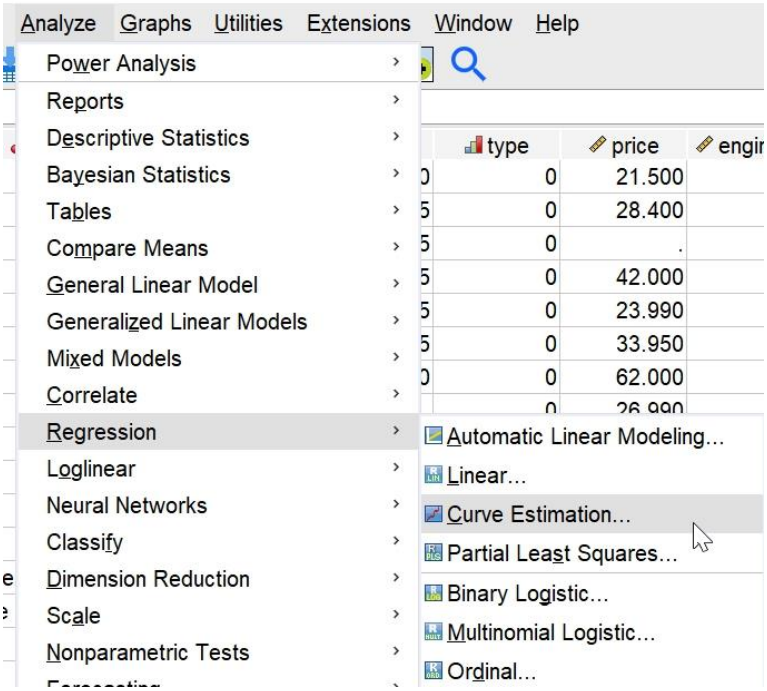
$$Y = a + bx$$

`Open the desired file in SPSS editor` File->Open->file path

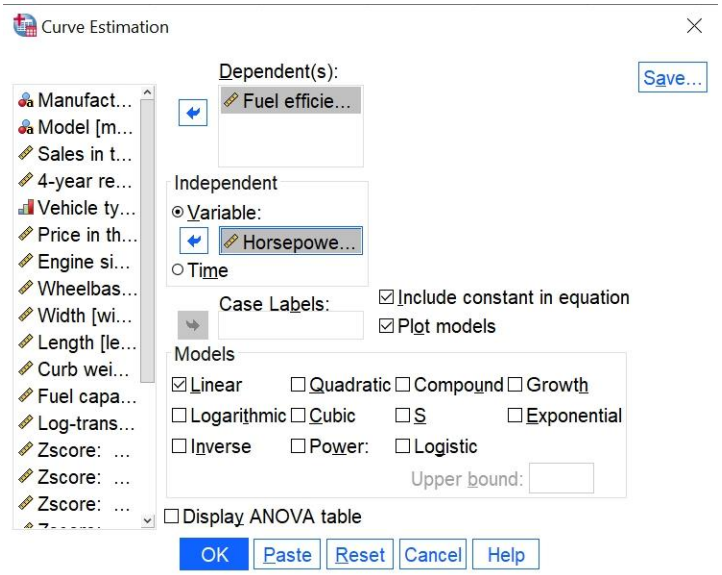
(car_sales.sav from repository is used and the variables Horsepower of the engine & Fuel efficiency is considered)

Case1. Using curve fitting path

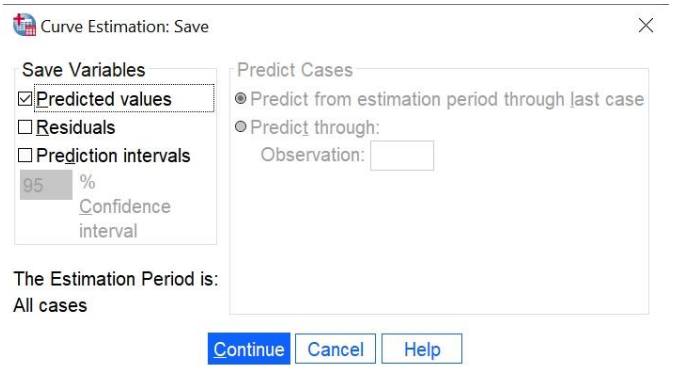
`Now follow the path` Analyze->Regression->Curve Estimation



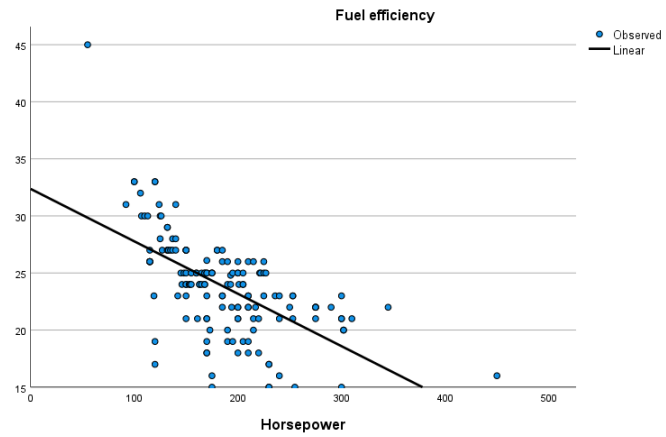
Choose dependent and independent variables and check the other desired boxes (Linear Model). Click save



Check the predicted values



Document Output Window



Model Description

Model Name		MOD_2
Dependent Variable	1	Fuel efficiency
Equation	1	Linear
Independent Variable		Horsepower
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified

Case Processing Summary

	N
Total Cases	157
Excluded Cases[a]	3
Forecasted Cases	0
Newly Created Cases	0

a. Cases with a missing value in any variable are excluded from the analysis.

Variable Processing Summary

		Variables	
		Dependent	Independent
		Fuel efficiency	Horsepower
Number of Positive Values		154	156
Number of Zeros		0	0
Number of Negative Values		0	0
Number of Missing Values	User-Missing	0	0
	System-Missing	3	1

Model Summary and Parameter Estimates

Dependent Variable: Fuel efficiency

	Model Summary					Parameter Estimates	
Equation	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.374	90.743	1	152	<.001	32.371	-.046

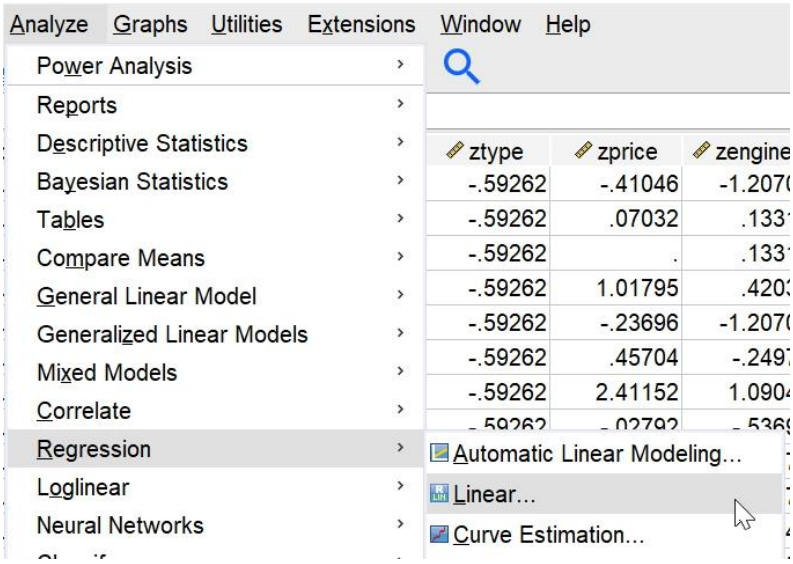
The independent variable is Horsepower.

Also in data sheet a new variable for predicted values has been created.

FIT_1
25.94066
22.03627
22.03627
22.72528
25.48132
23.18462
18.13189
24.56264
23.50616
23.50616
24.33297
21.34727
22.95495
22.95495

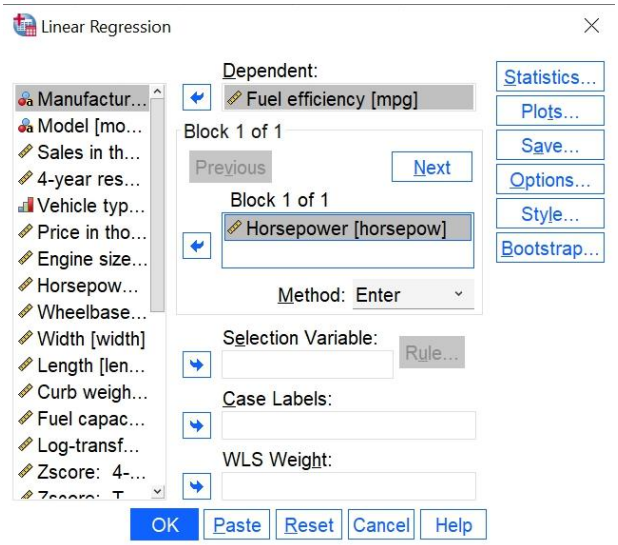
Case2. Using Linear Model Approach

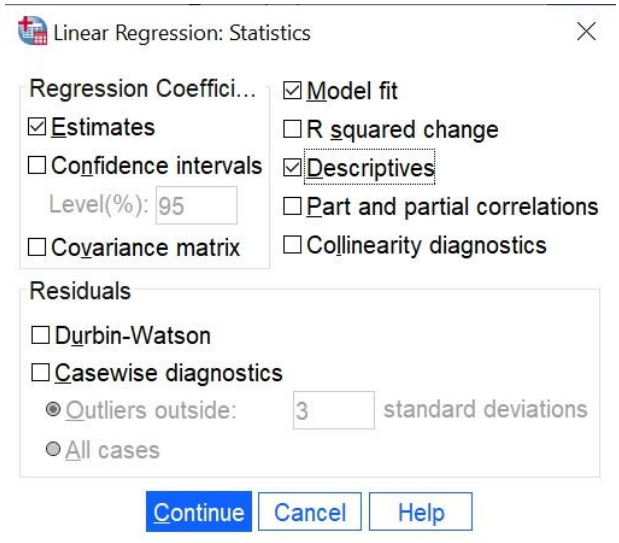
Now follow the path Analyze->Regression->Linear



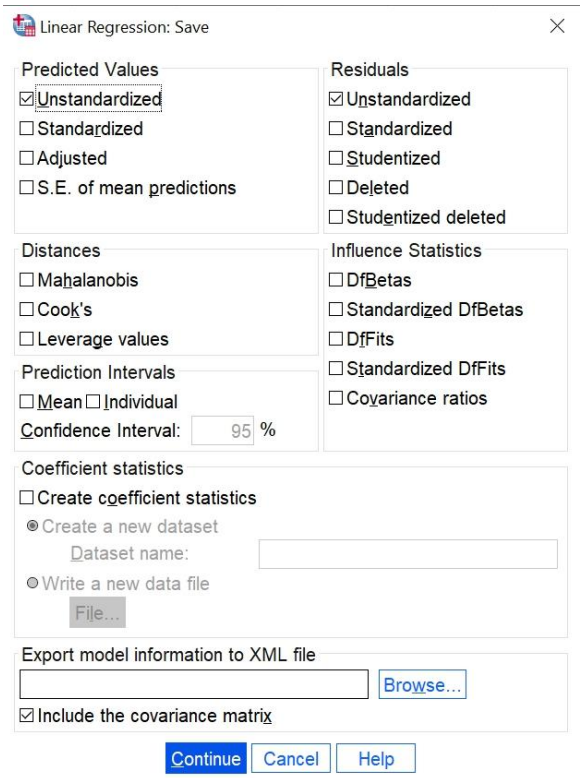
Choose dependent and independent variables and click statistics,

Check the desired boxes and continue

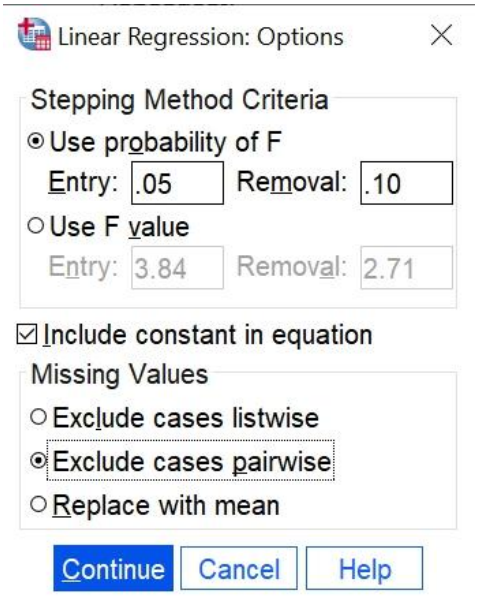




Click Save and check the desired boxes



Click Option



Document Output Window

Descriptive Statistics

	Mean	Std. Deviation	N
Fuel efficiency	23.84	4.283	154
Horsepower	185.95	56.700	156

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1049.053	1	1049.053	90.743	<.001[b]
	Residual	1757.226	152	11.561		
	Total	2806.279	153			

a. Dependent Variable: Fuel efficiency

b. Predictors: (Constant), Horsepower

Correlations

		Fuel efficiency	Horsepower
Pearson Correlation	Fuel efficiency	1.000	-.611
	Horsepower	-.611	1.000
Sig. (1-tailed)	Fuel efficiency	.	<.001
	Horsepower	.000	.
N	Fuel efficiency	154	154
	Horsepower	154	156

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	32.431	.942		34.421	<.001
	Horsepower	-.046	.005	-.611	-9.526	<.001

a. Dependent Variable: Fuel efficiency

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Horsepower[b]	.	Enter

a. Dependent Variable: Fuel efficiency

b. All requested variables entered.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	11.65	29.89	23.84	2.619	156
Residual	-9.889	15.109	-.014	3.389	154
Std. Predicted Value	-4.657	2.309	.000	1.000	156
Std. Residual	-2.908	4.444	-.004	.997	154

a. Dependent Variable: Fuel efficiency

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.611[a]	.374	.370	3.400

a. Predictors: (Constant), Horsepower

b. Dependent Variable: Fuel efficiency

Two new variables for predicted values (PRE 1) and difference between predicted and actual (RES_1) have been created.

PRE_1	RES_1
25.96548	2.03452
22.04006	2.95994
22.04006	3.95994
22.73278	-.73278
25.50367	1.49633
23.19460	-1.19460
18.11464	2.88536
24.58004	1.51996
23.51787	.48213
23.51787	1.28213
24.34913	.65087

Conclusion:

The Data set of Car sales from repository is used, and a regression line fitted between the variables (Horse power-independent variable and Fuel efficiency-dependent variable). the value of constant is.... and coefficient is ..... . Explain further the regression line properties exhibited in the case.

Precautions:

1. Variable Properties are to be defined carefully.
2. Missing values if any should be taken care of.

Experiment - 8

Hypothesis Testing 'T'- test

Theory and Methodology

Null Hypothesis **:** Any hypothesis we wish to test and is denoted by H_0 (hypothesis of no difference)

Alternative Hypothesis **:** Any hypothesis which is complementary to null hypothesis is called and alternative hypothesis, denoted by H_1

't' - test for single mean

If a random sample x_i , $i = 1,2, \dots n$ of size 'n' and is drawn from a normal population with mean μ then we have the statistic

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \quad , \quad or \quad t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

't' – test for difference of means

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Paired 't'- test for difference of means

Here we consider the increments $d_i = x_i - y_i$, $i = 1,2, \dots n$ under the null hypothesis H_0 that increments are due to fluctuations of sampling, that is, drug is not responsible for these instruments;

The test statistic $\boldsymbol{t = \frac{\bar{d}}{S/\sqrt{n}}}$ where

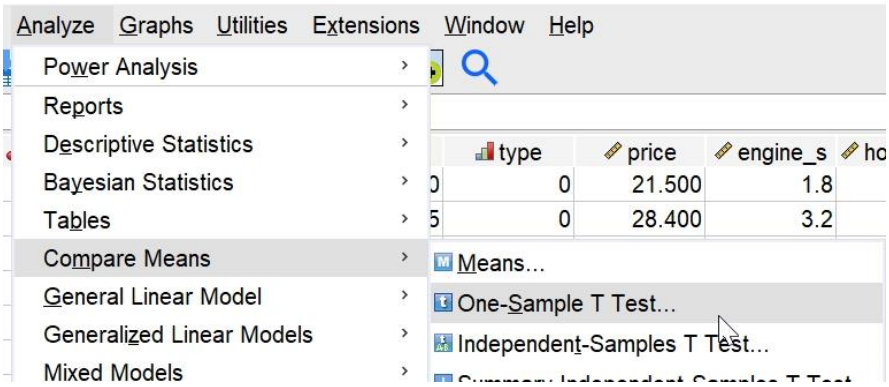
$\bar{d} = \frac{\sum d_i}{n}$ $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right]$ with $n-1$ df

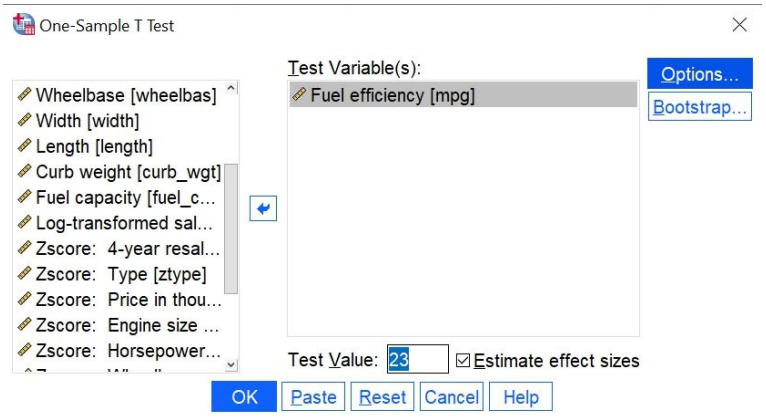
Case1. Test for single mean

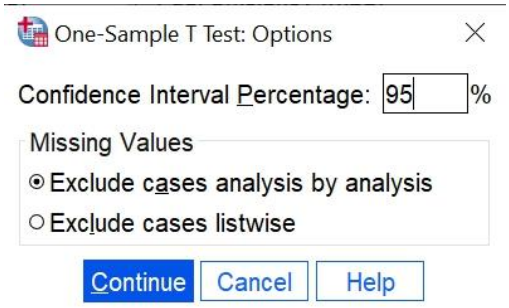
`Open the desired file in SPSS editor` File > Open > file path

(car_sales.sav from repository is used and the variables Fuel efficiency is considered)

$H0$: $Average\ is$ 23 km/lt or (= 23 $km\ per\ lt$) $H1$: $\mu \neq$ 23 $km\ per\ lt$)







Document Output View

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Fuel efficiency	154	23.84	4.283	.345

One-Sample Test

	Test Value = 23					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Fuel efficiency	2.444	153	.016	.844	.16	1.53

One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval	
				Lower	Upper
Fuel efficiency	Cohen's d	4.283	.197	.037	.356
	Hedges' correction	4.304	.196	.037	.354

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

*Note the p value is 0.016 which is less than the assigned value 0.05 hence null hypothesis is rejected, i.e., claim μ = 23 $km\ per\ lt$ cannot be accepted.

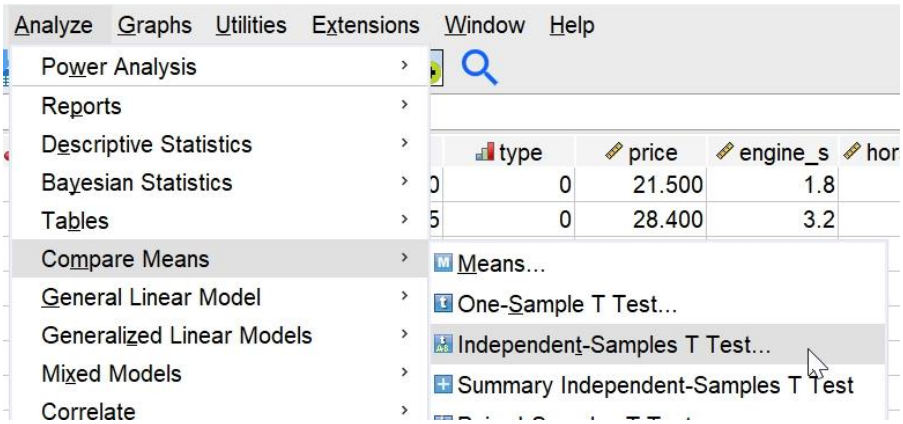
Case2. Test for difference of means

`Open the desired file in SPSS editor` File > Open > file path

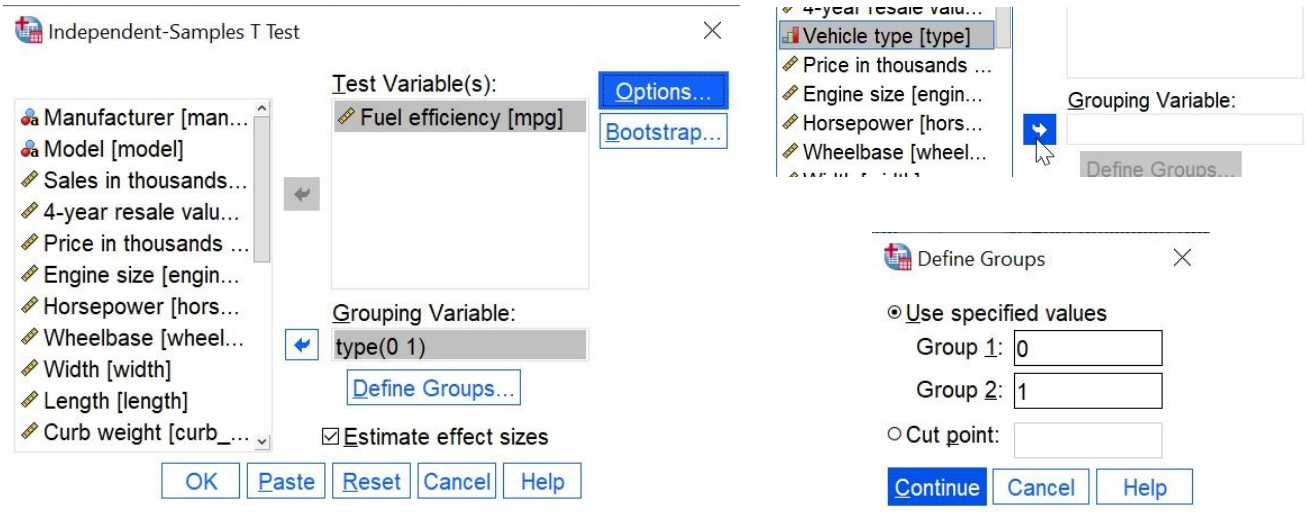
(car_sales.sav from repository is used and the variables Fuel efficiency is considered)

$H0 : Fuel\ efficiency\ of\ Automobiles = Fuel\ efficiency\ of\ Trucks$ ($A = \mu T$)

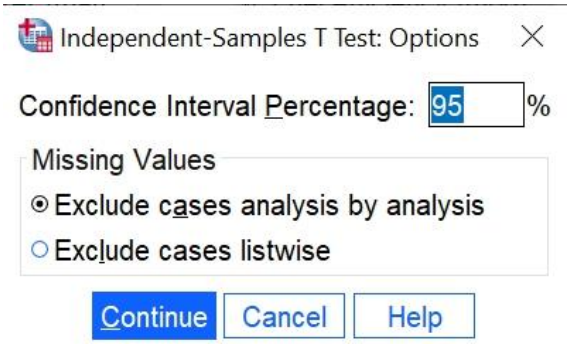
`Now follow the path` Analyze>Compare Means>Independent Sample Test



`Transfer the the desired variable to test variable then transfer the variable w.r.t. which the means are to be tested (in present case Automobiles & truck are considered)`



Enter the desired level of significance



Document Output View

Group Statistics

	Vehicle type	N	Mean	Std. Deviation	Std. Error Mean
Fuel efficiency	Automobile	114	25.30	3.646	.341
	Truck	40	19.70	3.107	.491

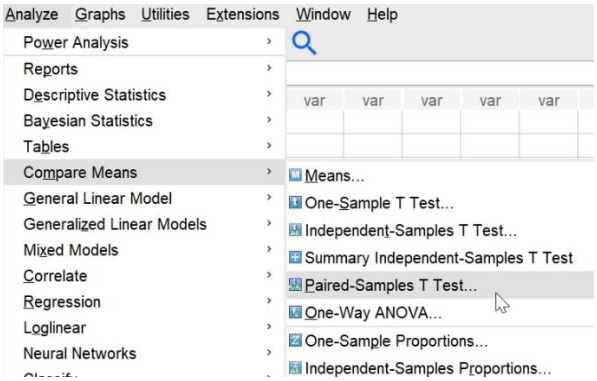
Independent Samples Test

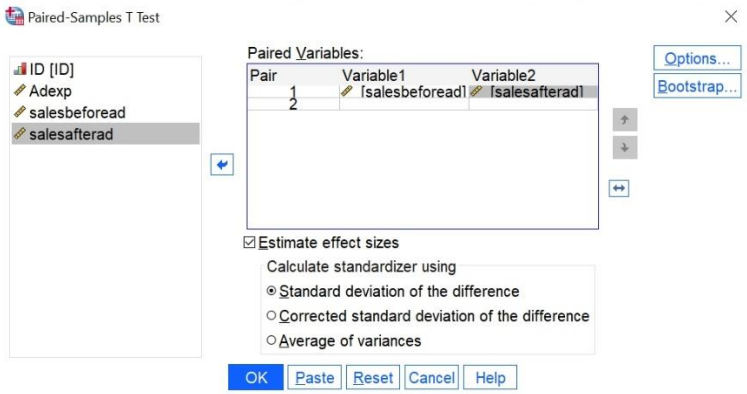
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Fuel efficiency	Equal variances assumed	.004	.948	8.664	152	<.001	5.597	.646	4.321	6.874
	Equal variances not assumed			9.356	79.405	<.001	5.597	.598	4.407	6.788

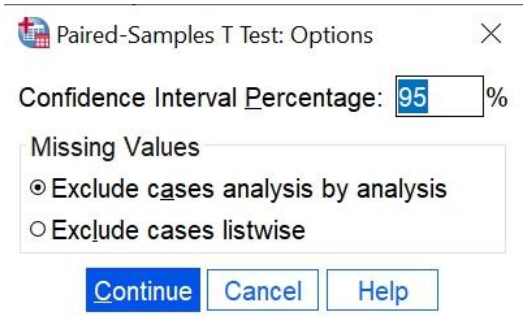
*(Levene's test is to decide the equivalence of the variances) in the present case in first row reading of p-value is 0.948 > 0.05 hence equivalence of variances can be accepted (Now we will read all information from first row). Also for the t-test p-value is almost zero <0.05 hence null hypothesis is rejected.

Case3. Paired 't'-test for difference of means

Open the desired file in SPSS editor File > Open > file path







H_0: $avg\ sales\ before\ ad = avg\ sales\ after\ ad$ ($b = \mu a$) or { $\mu b - \mu a = 0$ }

Document Output View

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	salesbeforead	72.9974	20	2.75234	.61544
	salesafterad	74.1420	20	2.74647	.61413

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	salesbeforead & salesafterad	20	.990	<.001

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	salesbeforead - salesafterad	-1.14459	.39674	.08871	-1.33027	-.95891	-12.902	19	<.001

Precautions:

1. Variable Properties are to be defined carefully.
2. Missing values if any should be taken care of.
3. Null hypothesis is defined carefully and level of significance chosen appropriately.

Experiment – 9

Chi Square Test

Theory and Methodology

Chi Square test in SPSS is test for association (or independence) of two nominal variables.

Case1. Chi-squared test of independence in contingency tables

A cell containing information of two characteristics resulting into a bivariate data. The table displaying data is called contingency table. Horizontal and vertical display of data is referred as $r \times c$ table.

Our interest is to see whether two characteristics are independent, thus

H_0: Two characteristics are independent. H_1: they are dependent.

O_{ij} : observed frequency of ith row and jth column respectively.

e_{ij} : expected frequency of ith row and jth column respectively.

Then

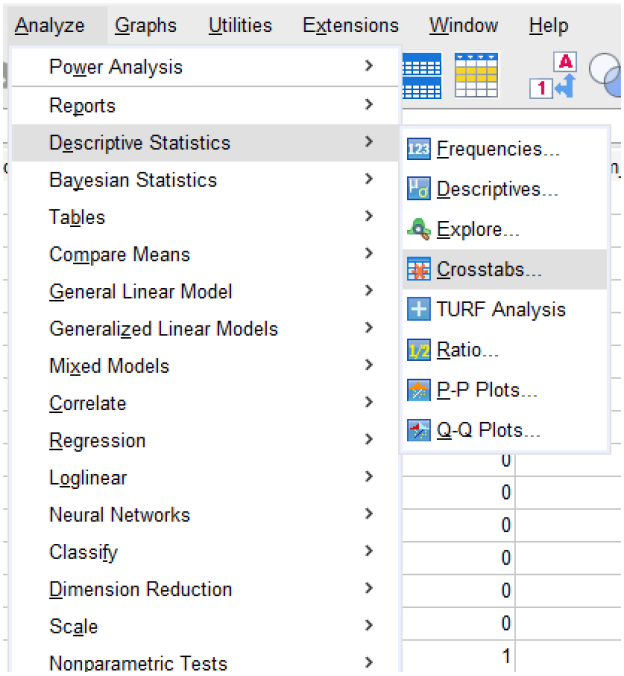
$So, \quad \chi^2 = \sum_i \sum_j \frac{(O_{ij}-e_{ij})^2}{e_{ij}}$ with $(r-1) \times (c-1)$ degree of freedom (df).

Open the desired file in SPSS editor File > Open > file path

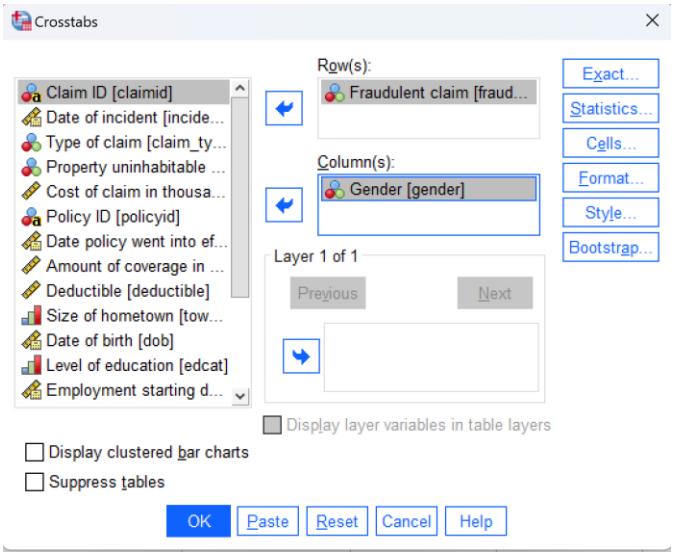
(insurance_claims.sav from repository is used and Fraudulent claim & Gender variables are considered for their association)

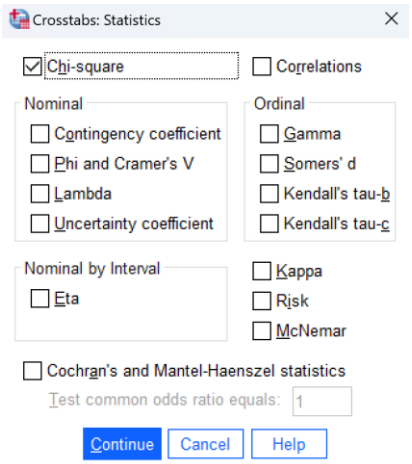
H_0: Fraudulent claims are independent of Gender

H_1: Fraudulent claim is Gender dependent



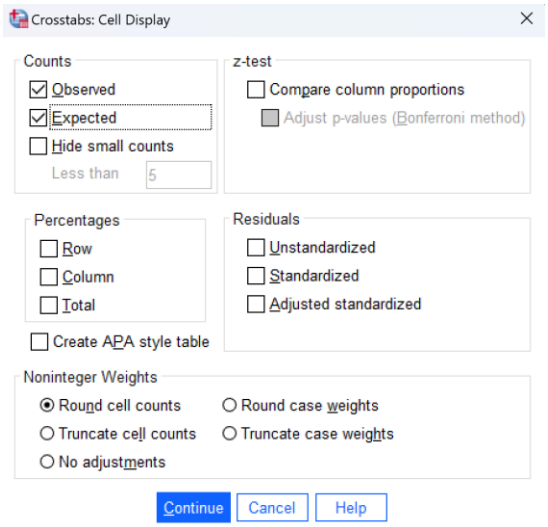
Transfer the desired variables in rows & column box (their order does not matter).





Then click Statistics tab and check Chi-square box.

click continue. Then click Cells button



Document Output View

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Fraudulent claim * Gender	4415	100.0%	0	0.0%	4415	100.0%

Fraudulent claim * Gender Crosstabulation

			Gender		
			Male	Female	Total
Fraudulent claim	No	Count	1964	1988	3952
		Expected Count	1946.9	2005.1	3952.0
	Yes	Count	211	252	463
		Expected Count	228.1	234.9	463.0
Total		Count	2175	2240	4415
		Expected Count	2175.0	2240.0	4415.0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.820[a]	1	.093		
Continuity Correction[b]	2.657	1	.103		
Likelihood Ratio	2.824	1	.093		
Fisher's Exact Test				.095	.051
Linear-by-Linear Association	2.819	1	.093		
N of Valid Cases	4415				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 228.09.

b. Computed only for a 2x2 table

The Pearson Chi-Square p-value 0.093 is greater than 0.05, therefore Null Hypothesis is not rejected. It is concluded that fraudulent claim is not gender biased.

Case2. Chi-square test for goodness of fit

If a population has a specified theoretical distribution. The test is based on how good a fit we have between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution.

A goodness of fit test between observed and expected frequencies is based on the quantity

$$\mathcal{X}^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$$

O_i: Observed Frequency of the cell
e_i: Expected Frequency of the cell

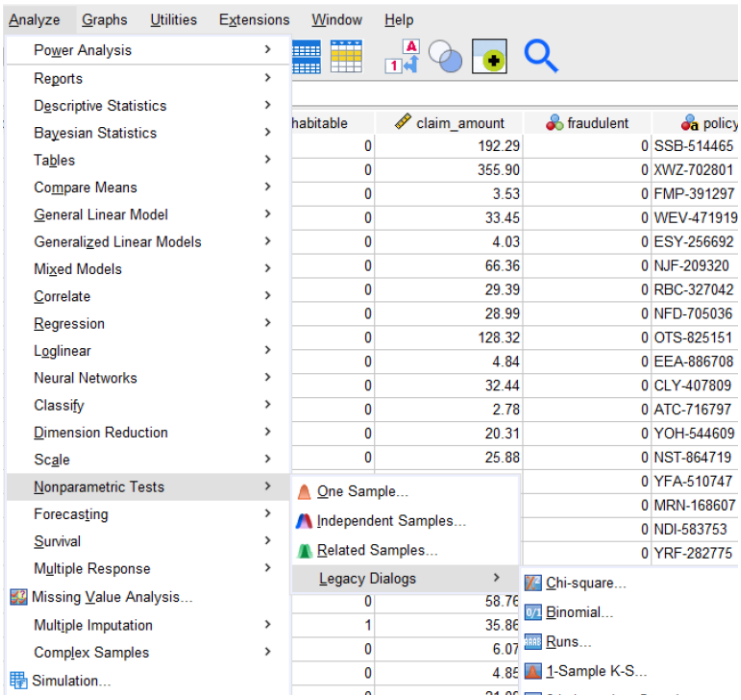
Where $\mathcal{X}^2$ is a value of a random variable whose sampling distribution is approximated very closely by the Chi square distribution with $v = k - 1$ degree of freedom (df).

Critical values

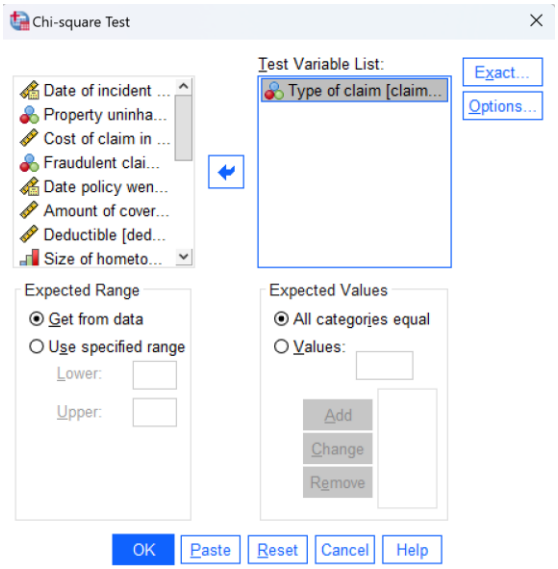
If the observed frequencies differ considerately from the expected frequencies, the $\mathcal{X}^2$ value will be large and fit is poor. A good fit leads to the acceptance of H_0, where as a poor fit leads to rejection.

H_0: Claims are equally distributed
H_1: It is not equally distributed

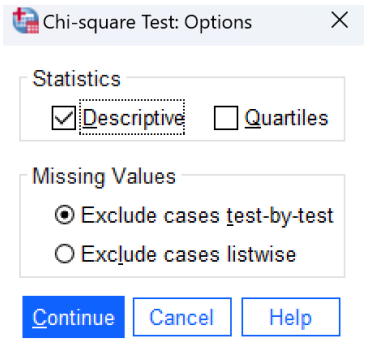
Now follow the path Analyze > Nonparametric Tests > Legacy dialogs > Chi Square



Transfer the desired variable to Test variable list box



Click option and check the desired box (Descriptive under statistics)



Document Output View

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Type of claim	4415	3.06	1.535	1	5

Chi-Square Test

Frequencies

Type of claim

	Observed N	Expected N	Residual
Wind/Hail	1054	883.0	171.0
Water damage	627	883.0	-256.0
Fire/Smoke	1039	883.0	156.0
Contamination	404	883.0	-479.0
Theft/Vandalism	1291	883.0	408.0
Total	4415		

Test Statistics

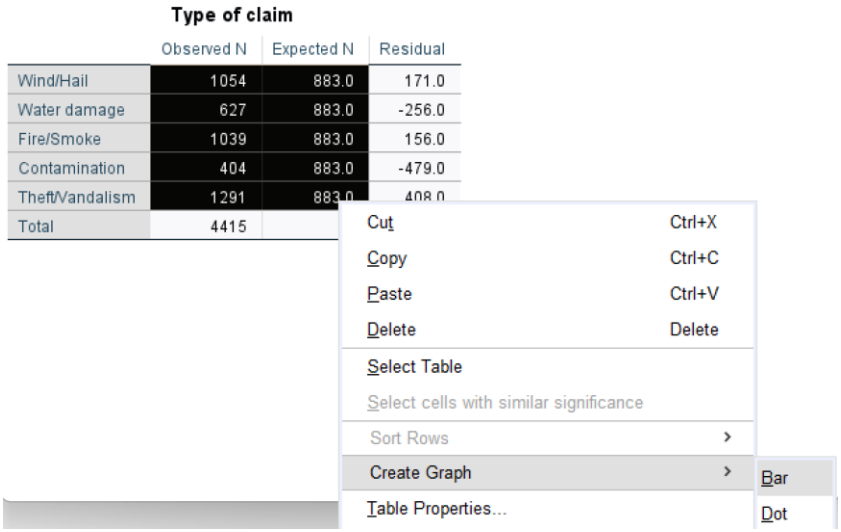
	Type of claim
Chi-Square	583.259[a]
df	4
Asymp. Sig.	<.001

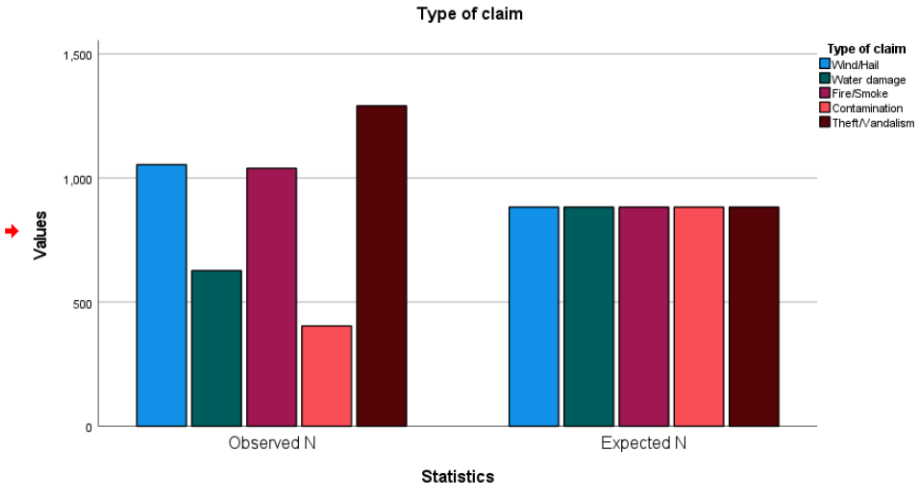
a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 883.0.

The Test Statistics shows the p-value almost zero which is less than 0.05, therefore the null hypothesis is rejected in favor of alternate, i.e., type of claim are not equally distributed. This can be visualized through graph also.

Double click the frequency table to bring into editable mode

Select both columns of observed and expected frequency and right-click >Create Graph>Bar





Conclusion:

Give information of data set used for testing. Also mentioning the p-value write the decision about rejection or non-rejection of null hypothesis.

Precautions:

1. Variable Properties are to be defined carefully.
2. Missing values if any should be taken care of.
3. Null hypothesis is defined carefully and level of significance chosen appropriately.

Experiment - 10

ANOVA One-Way

Theory and Methodology

Let k random samples of size $n_1, n_2 \ldots n_k$ are selected from normal populations with mean $\mu_1, \mu_2 \ldots \mu_k$, and common variance σ^2 . And let the means are depending on one factor then One Way ANOVA is applied

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ and H_1: at least two means are different.

Procedure: the total variation can be attributed to variation within the class and variation between the class.

Total sum of squares = Sum of square within the class + sum of square between the class

$TSS = SSB + SSW$

The degree of freedom (df) for TSS is $(N-1)$, for SSB is $(k-1)$, & for SSW is $(N-k)$.

Now mean square (MS) for each variation is obtained by dividing each sum of square by its respective degree of freedom (df), i.e.,

$$MSS = \frac{TSS}{N-1}, \qquad MSB = \frac{SSB}{k-1}, \qquad MSW = \frac{SSW}{N-k}$$

The value of F - Statistic will be

$$F = \frac{\sum_i \sum_j n_i (x_i - \bar{x})^2}{\sum_i \sum_j (x_{ij} - \bar{x})^2} \times \frac{N-k}{k-1}, \qquad v_1 = k-1, \qquad v_2 = N-k$$

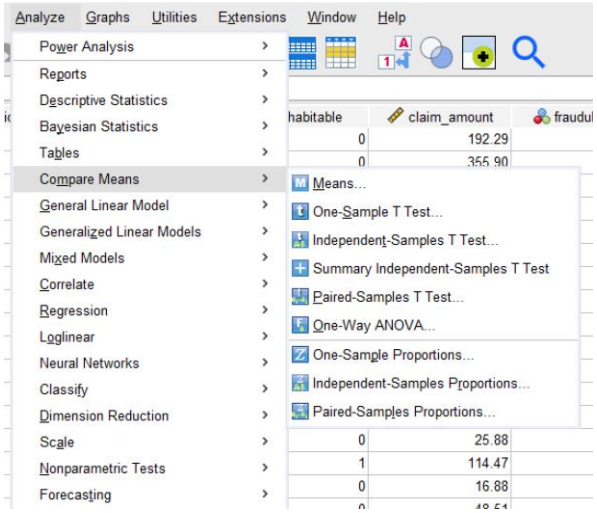
Open the desired file in SPSS editor File > Open > file path

*(insurance_claims.sav from repository is used and claim amount & Type of claim variables are considered) (see the data 5-types of claims are there)

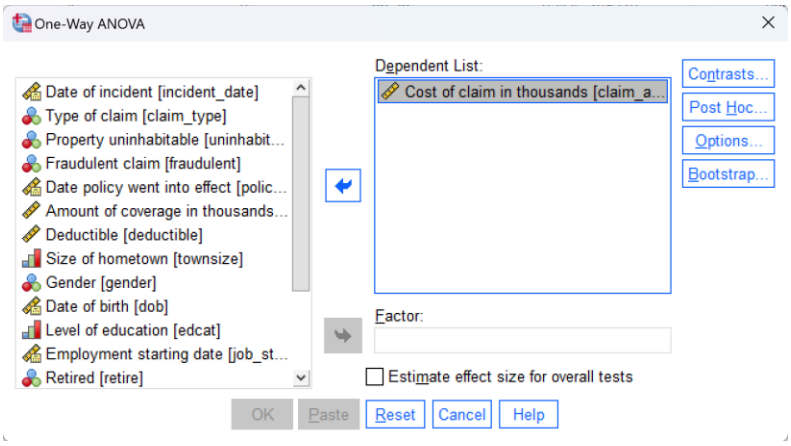
H_0: average claim amount in each category is same. Ie., $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5$

H_1: At least one of these differs

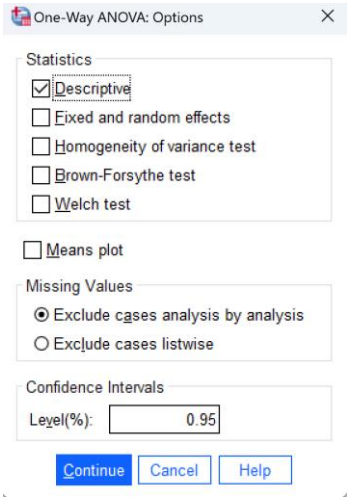
Now follow the path Analyze > Compare Means



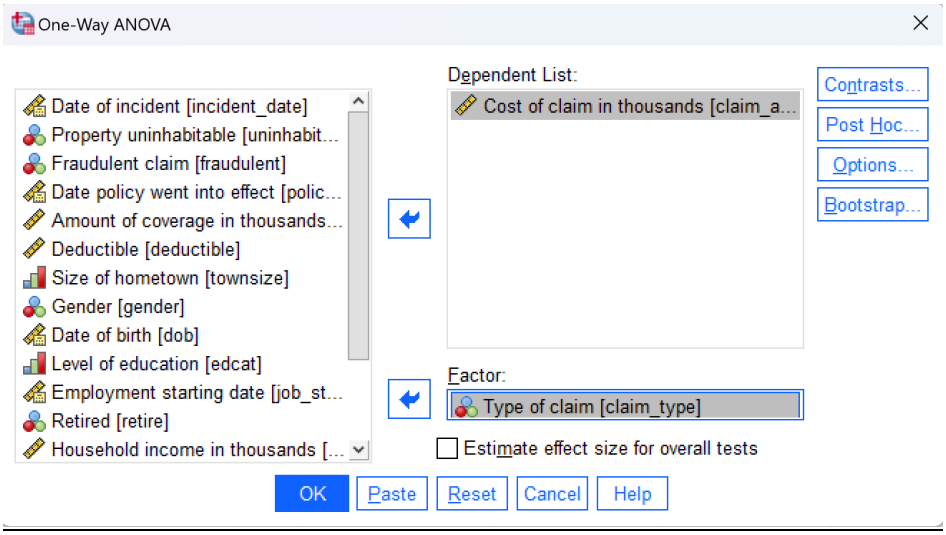
Transfer Cost of claim to dependent list and Type of claim to factor list



Click Option and select the Descriptive to display statistics



Continue & click OK.



Document Output View

Oneway

Descriptives

Cost of claim in thousands

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Wind/Hail	1054	16.7819	20.51640	.63195	15.5419	18.0219	1.46	184.38
Water damage	627	35.2896	47.55906	1.89933	31.5598	39.0195	2.31	504.94
Fire/Smoke	1039	171.5793	204.53371	6.34537	159.1281	184.0305	11.59	1635.00
Contamination	404	202.2067	227.66948	11.32698	179.9393	224.4740	18.69	1662.00
Theft/Vandalism	1291	17.4795	25.27650	.70348	16.0994	18.8596	1.57	576.91
Total	4415	73.0109	144.40137	2.17323	68.7503	77.2715	1.46	1662.00

ANOVA

Cost of claim in thousands

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	25043718.32	4	6260929.579	412.125	<.001
Within Groups	66995929.16	4410	15191.821		
Total	92039647.48	4414			

Since the p-value of F-statistics is almost zero which is less than 0.05, hence the Null Hypothesis is rejected, i.e., at least one of the mean is different.

Conclusion:

Give information of data set used for testing. Also mentioning the p-value write the decision about rejection or non-rejection of null hypothesis.

Precautions:

1. Variable Properties are to be defined carefully.
2. Missing values if any should be taken care of.
3. Null hypothesis is defined carefully and level of significance chosen appropriately.