STUDYING PHENOMENA AND PROCESSES

Grouping, descriptive statistics and graphical visualization

DATA SETS CLASSIFICATION

- By number of variables there are for each elementary unit (=people, companies, countries, cities, etc.)
- By the kind of measurement (numbers of categories in each case)
- Whether there is a time sequence
- Newly created or was previously created by someone else

- **Univariate** data – just one piece of information for each item.

We can summarize basic properties

- **Bivariate** data – two pieces of information for each item.

+Relationship can be measured

- **Multivariate** – many pieces of information for each item.

+look at the interrelationships among all the items

Company	2008 Profits ($ millions)
McDonald's	$4,313.2
Yum Brands	964.0
Starbucks	315.5
Darden Restaurants	377.2
Brinker International	51.7
Jack in the Box	119.3
Burger King Holdings	190.0
Cracker Barrel Old Country Store	65.6
Wendy's/Arby's Group	–479.7
Bob Evans Farms	64.9

Source: Data are from http://money.cnn.com/magazines/fortune/fortune500/2009/industries/147/index.html, accessed on July 1, 2010.

Company	2008 Profits ($ millions)	Total Return to Investors 2008 (%)
McDonald's	$4,313.2	8.5%
Yum Brands	964.0	–16.0%
Starbucks	315.5	–53.8%
Darden Restaurants	377.2	4.4%
Brinker International	51.7	–44.5%
Jack in the Box	119.3	–14.3%
Burger King Holdings	190.0	–15.4%
Cracker Barrel Old Country Store	65.6	–34.6%
Wendy's/Arby's Group	–479.7	–39.8%
Bob Evans Farms	64.9	–22.3%

Source: Data are from http://money.cnn.com/magazines/fortune/fortune500/2009/industries/147/index.html, accessed on July 1, 2010.

Company	2008 Profits ($ millions)	Total Return to Investors 2008 (%)	Employees	Revenues ($ millions)
McDonald's	$4,313.2	8.5%	400,000	$23,522.4
Yum Brands	964.0	–16.0%	193,200	11,279.0
Starbucks	315.5	–53.8%	176,000	10,383.0
Darden Restaurants	377.2	4.4%	179,000	6,747.2
Brinker International	51.7	–44.5%	100,400	4,235.2
Jack in the Box	119.3	–14.3%	42,700	3,001.4
Burger King Holdings	190.0	–15.4%	41,000	2,455.0
Cracker Barrel Old Country Store	65.6	–34.6%	65,000	2,384.5
Wendy's/Arby's Group	–479.7	–39.8%	70,000	1,822.8
Bob Evans Farms	64.9	–22.3%	49,149	1,737.0

Source: Data are from http://money.cnn.com/magazines/fortune/fortune500/2009/industries/147/index.html, accessed on July 1, 2010.

LEVELS OF MEASUREMENT

- **Nominal**-level variable has values that show difference that subjects have on the characteristic being measured. Simply put, there is no inherent order of categories.
- I.e. religion: Protestant, Catholic, Jewish, other;
- **Ordinal**-level variable has values that show relative differences between subjects on the characteristic being measured. Simply put, there is a meaningful order of categories but we cannot measure the difference between categories.
- I.e. Support for abortion values: oppose, neutral, support.
- ** Nominal + Ordinal =Qualitative data
- **Interval**-level variable has values that communicate exact differences between subjects on the measured characteristic. We can measure both the order and the difference.
- I.e. age: 18, 24, 30
- ***Interval=Quantitative

QUANTITATIVE DATA (NUMBERS)

- **Discrete** quantitative data can assume values only from a list of specific numbers.
- I.e. gender of students, coded 0=male, 1=female;
- Number of kids in household: 1=1kid, 2=2 kids, 3=3 kids, 4=4 or more kids;
- Equipment breakdowns on a factory in the past 24 hours, out of 20 working machines
- **Continuous** quantitative data – all positive numbers, all numbers, all values between 0% and 100%.

Grouping statistical data

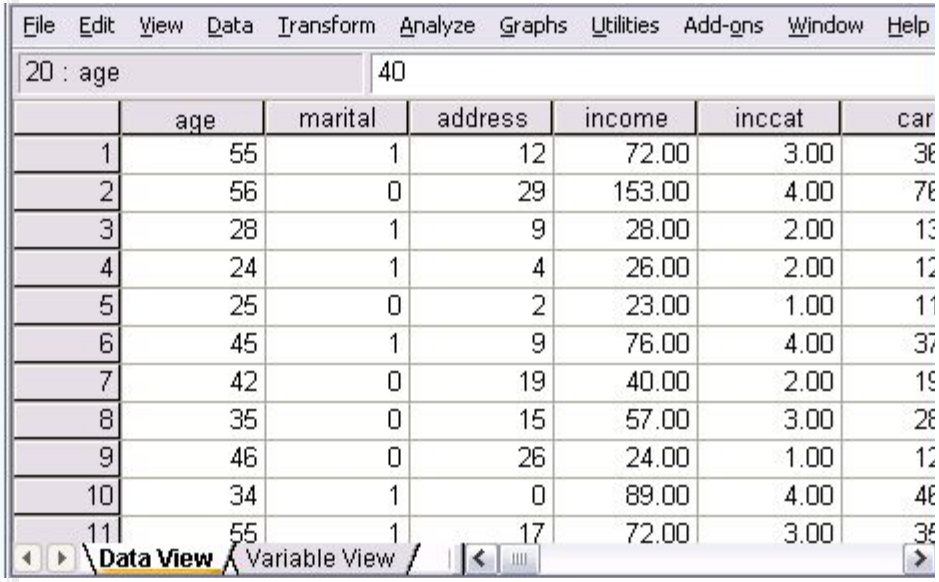
- A data class is group of data which is related by some user defined property.
- For example, if you were collecting the ages of the people you met as you walked down the street, you could group them into classes as those in their teens, twenties, thirties, forties and so on. Each of those groups is called a class.
- Each of those classes is of a certain width and this is referred to as the Class Interval or Class Size. This class interval is very important when it comes to drawing Histograms and Frequency diagrams. All the classes may have the same class size or they may have different classes sizes depending on how you group your data. The class interval is always a whole number.
- Number of intervals: n=1+3,322*lgN (N – number of set values)
- Interval size for equal intervals:
- (Highest Value-Lowest Value)/no of classes
- Size should be a whole number. I.e., if you get 2,7 – your class size is 3

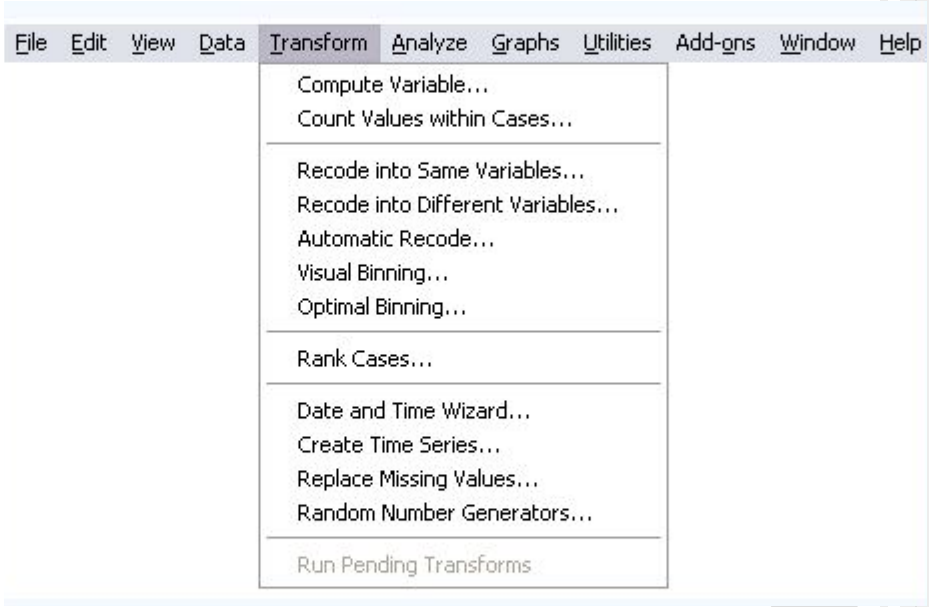
Grouping in Excel

- I.e. you have a raw set of data in excel. You have numbers of different people's ages. Eg. 28 years old -50 ppl, 60 years old – 10 ppl, 14 years old – 10 ppl, etc.. You can sort the data by people's ages and then, using the formula, count the length of an interval. Put the intervals and using the sum function, organize the data into intervals.
- Or function =SUMMPRODUCT

SPSS DATA GROUPING

- We want to group income by less than 25, 25-49, 50-74, 75 and more
- Go Transform-Visual Binning

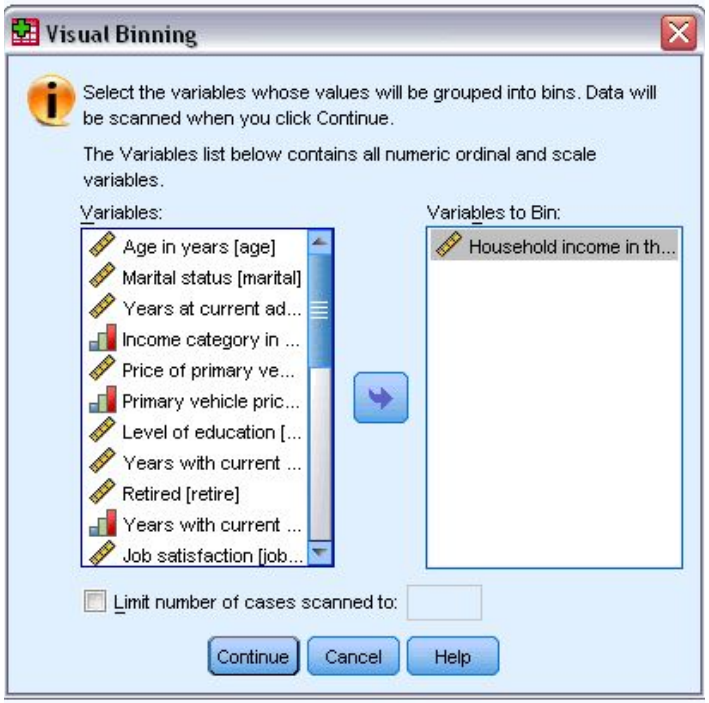


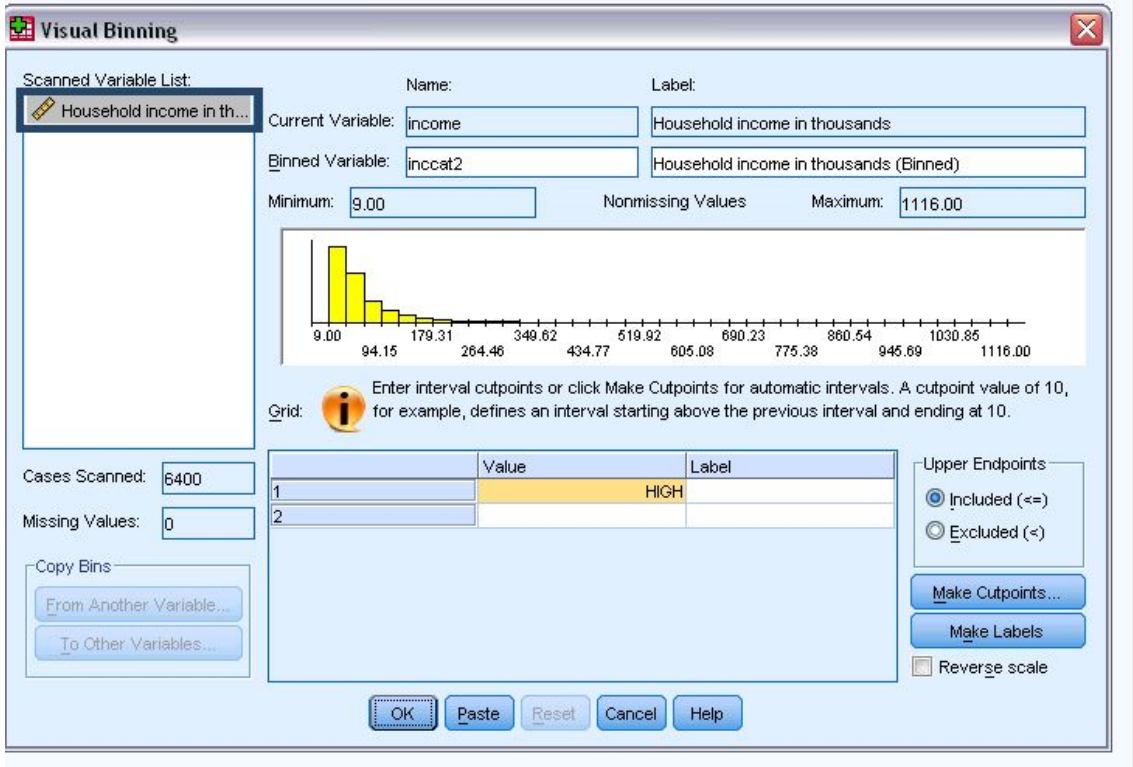


We can select scale or ordinal variable to bin them.

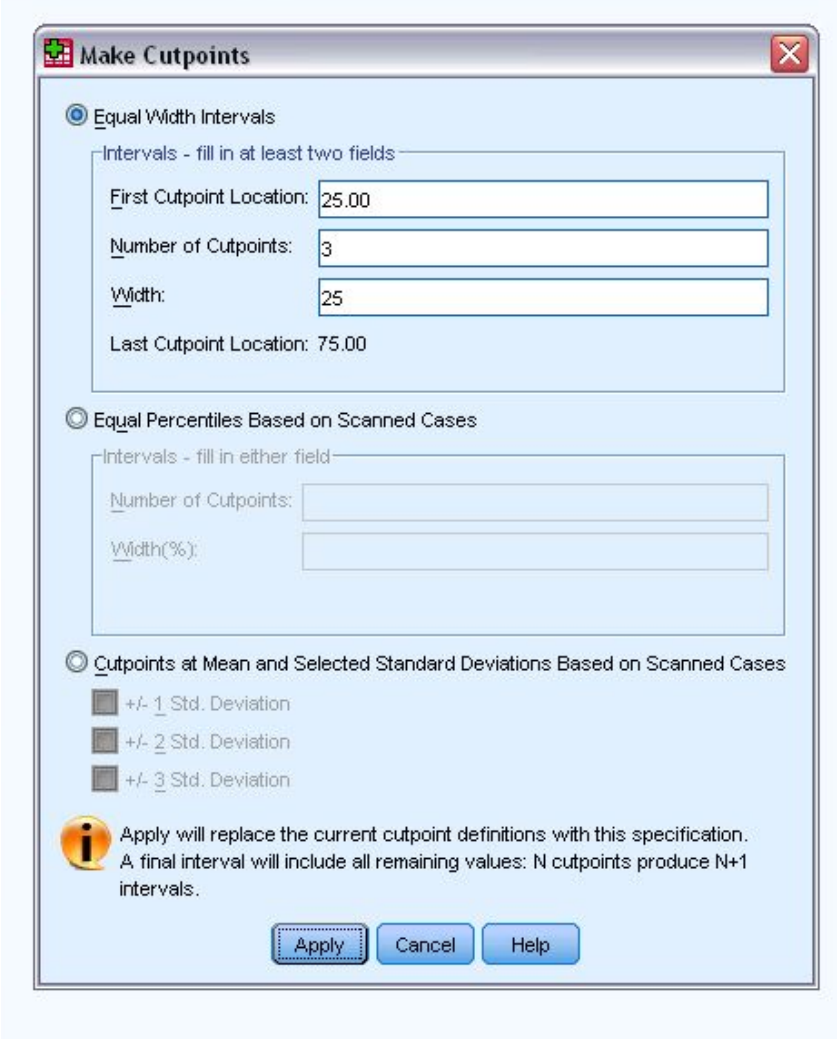
Binning=take two or more contiguous values and group them into a category

Press "make cutpoints"

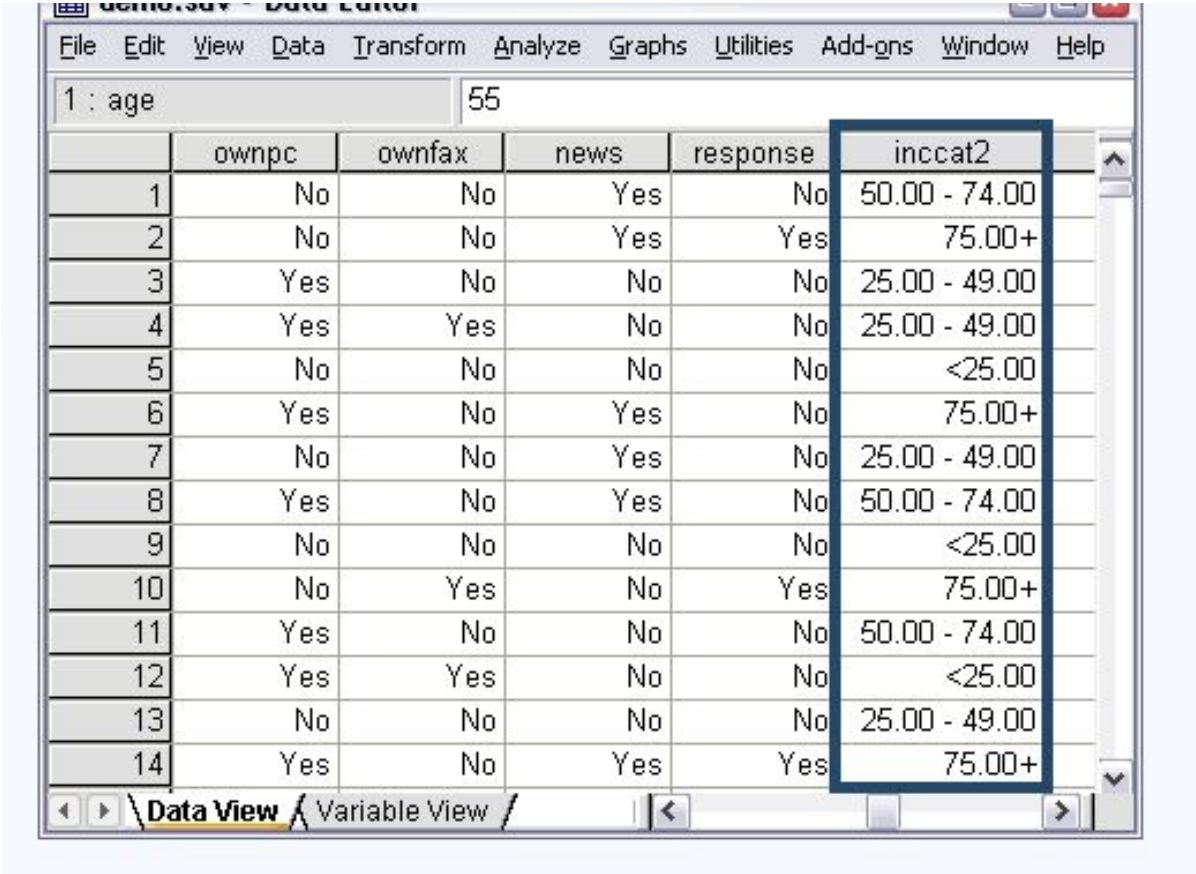




We can select "Equal with Intervals" and put in First Cutpoint Location, Number of Cutpoints and Width .



We can make labels if we want and choose to exclude or include the last number in interval. In the end, we get:



File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : age 55

	ownpc	ownfax	news	response	inccat2
1	No	No	Yes	No	50.00 - 74.00
2	No	No	Yes	Yes	75.00+
3	Yes	No	No	No	25.00 - 49.00
4	Yes	Yes	No	No	25.00 - 49.00
5	No	No	No	No	<25.00
6	Yes	No	Yes	No	75.00+
7	No	No	Yes	No	25.00 - 49.00
8	Yes	No	Yes	No	50.00 - 74.00
9	No	No	No	No	<25.00
10	No	Yes	No	Yes	75.00+
11	Yes	No	No	No	50.00 - 74.00
12	Yes	Yes	No	No	<25.00
13	No	No	No	No	25.00 - 49.00
14	Yes	No	Yes	Yes	75.00+

Data View Variable View

DESCRIPTIVE STATISTICS

- Measures of central tendency - identify the most typical value or best representative of a set of empirical data
- Measures of dispersion- amount of variation around the most representative value

Descriptive statistics: Nominal level

- Central Tendency – **mode** (the most common value of the variable)
- Dispersion – variation ratio

$$v = \frac{\sum(Fnon-modal)}{N}$$

$$v = 1 - \frac{\sum(Fmodal)}{N}$$

Descriptive statistics: Ordinal level

- Central Tendency – mode + **median**=the one in the middle=half the cases with values below the median and half above . Put the data in order and find the middle value.
- Median is the value, which rank is [(N+1)/2] with odd variables, or if variables are even – N/2 and (N+1)/2.
- Dispersion – range [(highest score-lowest score)+1]

Descriptive statistics: Interval level

- Central tendency: mean [=average] $\bar{x} = \frac{\sum x}{n}$
- +Weighed Average $\omega_1 X_1 + \omega_2 X_2 + \ldots + \omega_n X_n = \sum_{i=1}^{n} \omega_i X_i$
- Dispersion – Standard Deviation Sx= $\sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$
- Variation ratio for interval level=

=Standard Deviation/Average

	Quantitative	Ordinal	Nominal
Average, Standard Deviation	+		
Median, Range	+		
Mode, Variation Ratio	+	+	+

- Average in Excel: Insert-Function-Average-Enter
- Weighed average for named columns, where a is weights column: SUMPRODUCT (a;b)/Summ(a).

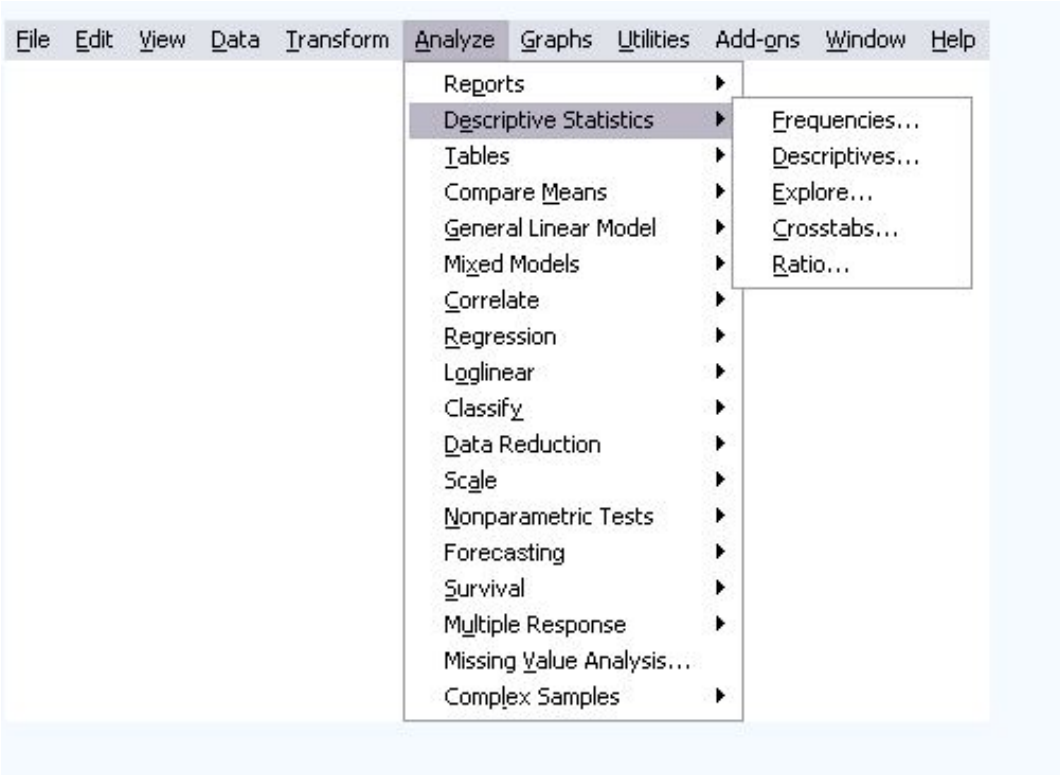
- Median in Excel: =MEDIAN(A1:AN)
- Mode in Excel: =MODE(A1:AN)

- Standard Deviation: =STDEV(A1:AN)
- Or 1) Calculate Avg. 2) Calculate each value's difference from Avg. 3) Square each one, sum the squared ones and divide by N-1 4) square root it all.

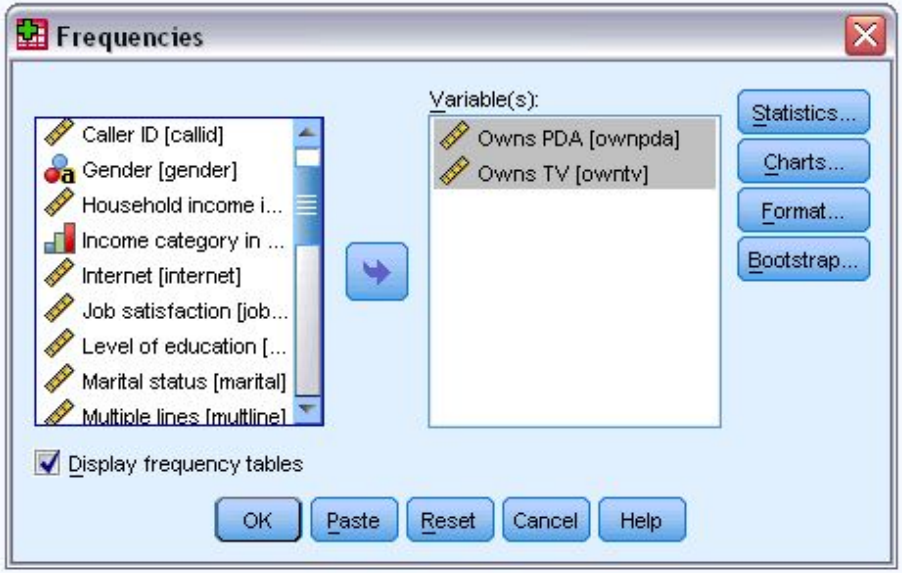
- Or, much, much easier: Service-Data Analysis-Descriptive Statistics – and you get everything you need in one click.
- (If there is no data analysis option – add it in excel properties)

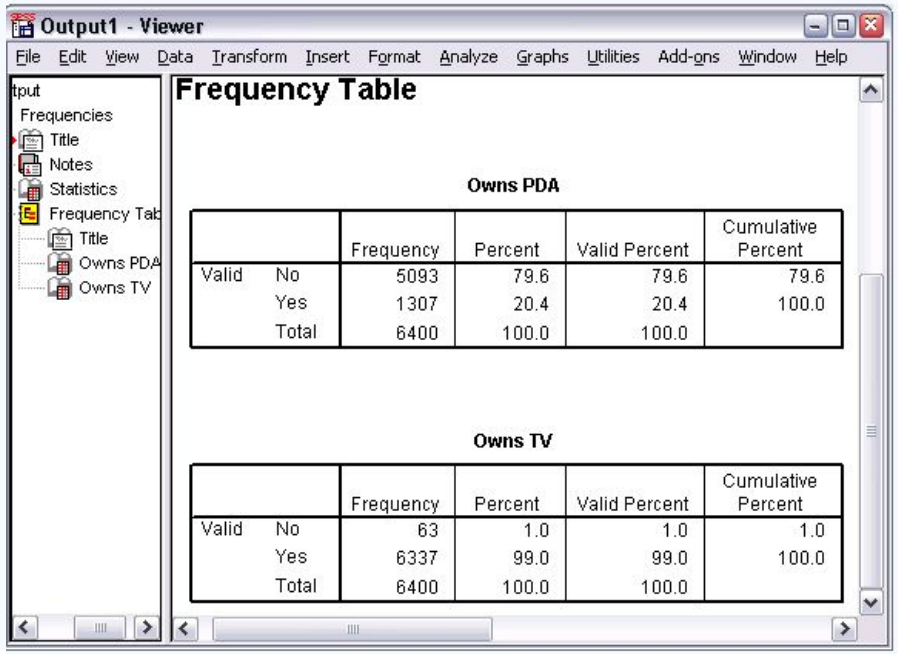
SPSS SUMMARY MEASURES FOR CATEGORICAL DATA

GO ANALYZE – DESCRIPTIVE STATISTICS - FREQUENCIES

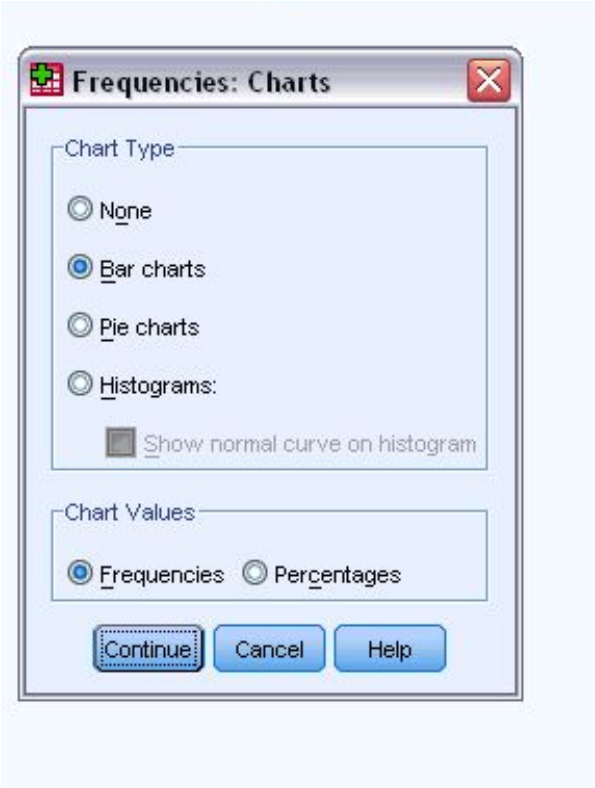


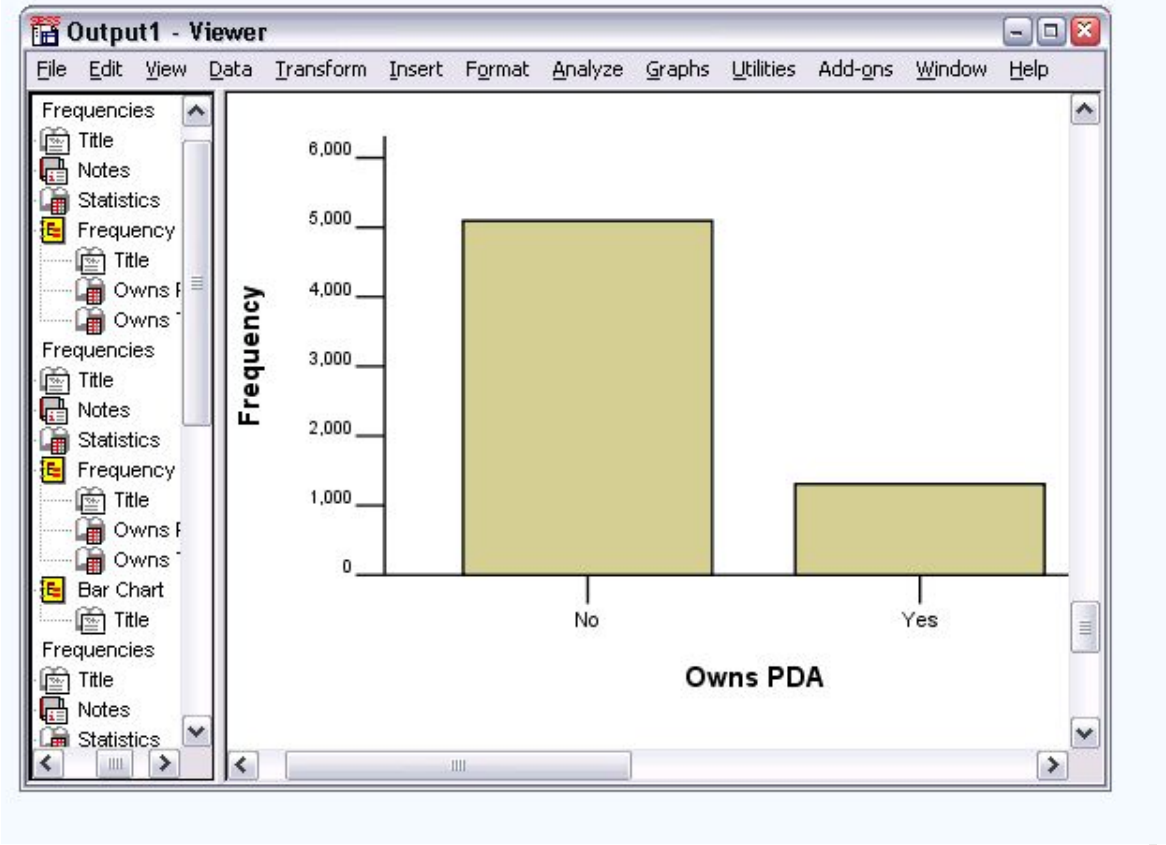
Choose variables, press OK and you get your frequency table





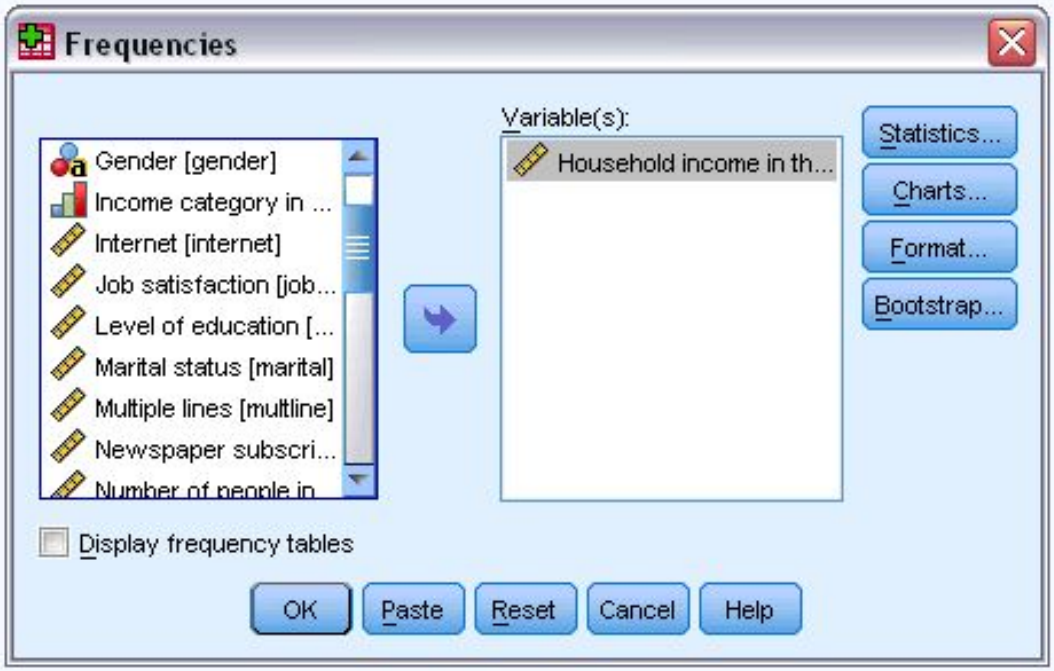
To graphically display press Charts and select the ones you like

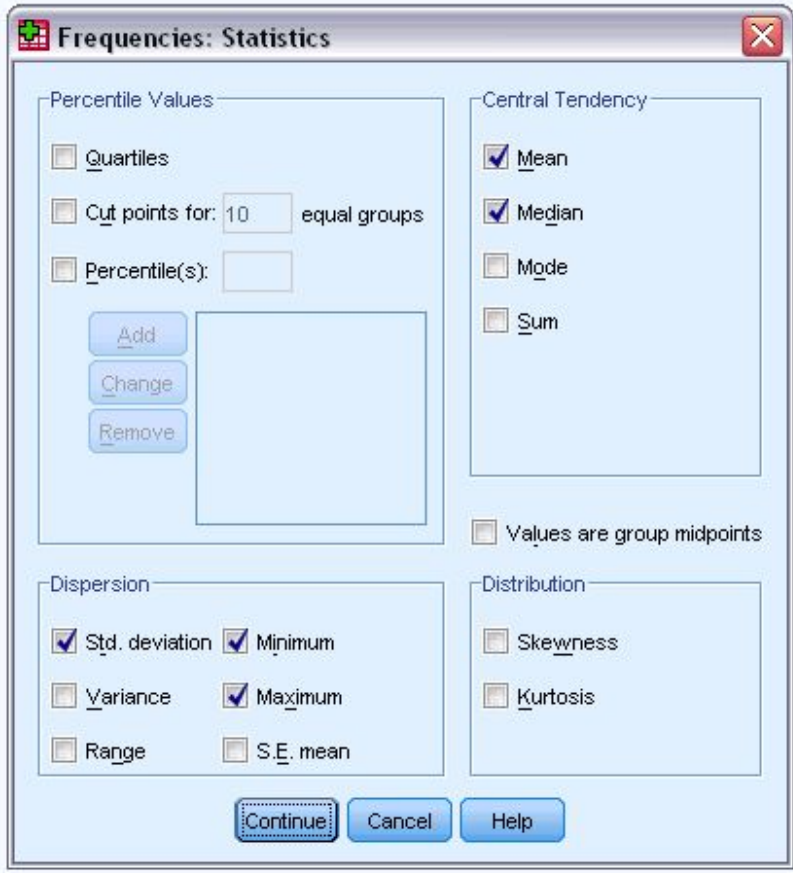




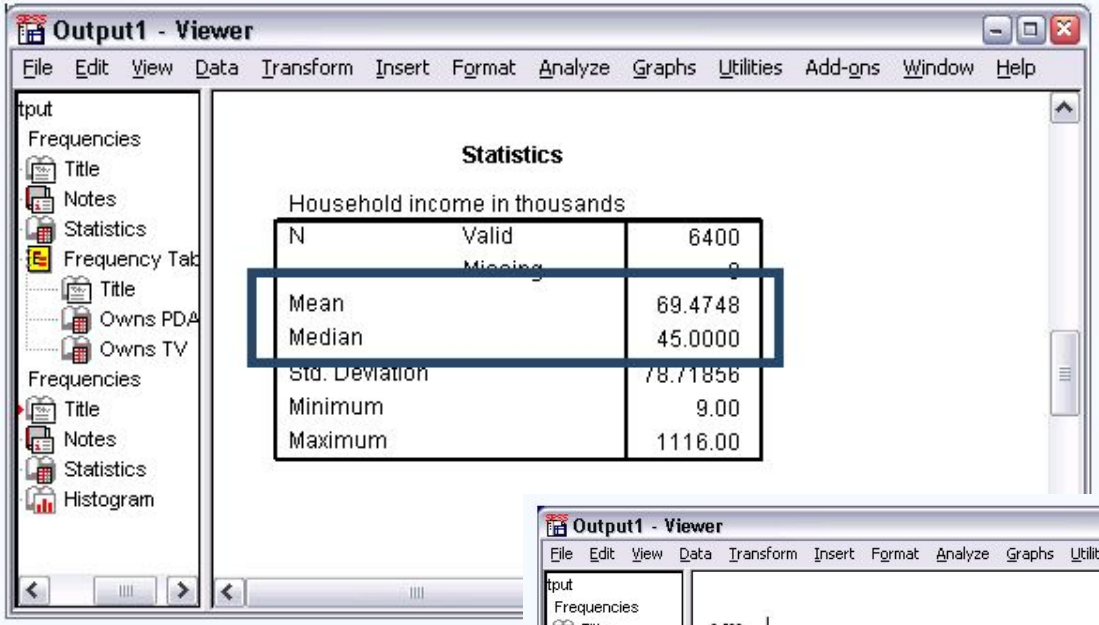
Summary measures for scale variables in SPSS

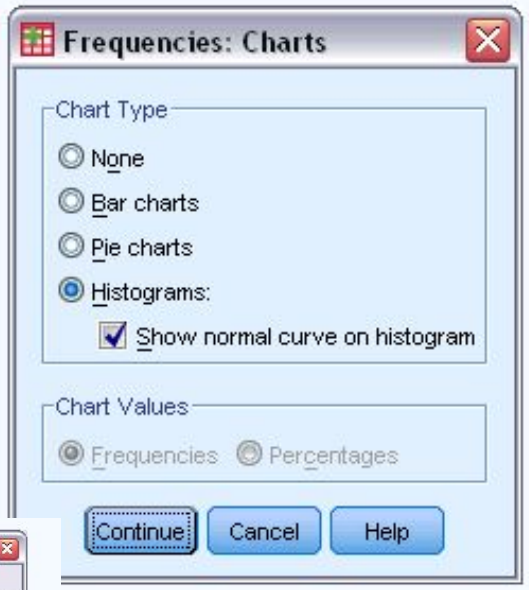
* Go Analyze-Descriptive Statistics-Frequencies
* Choose variables
* Click "statistics", select the ones you need.

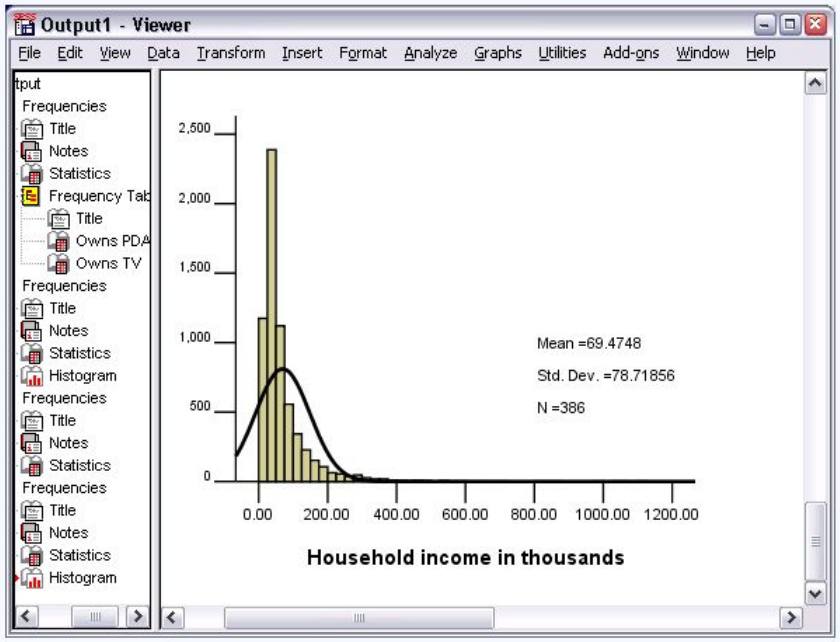




* You get the data in the viewer window
* Go back to frequencies dialog, click charts and choose the ones you want

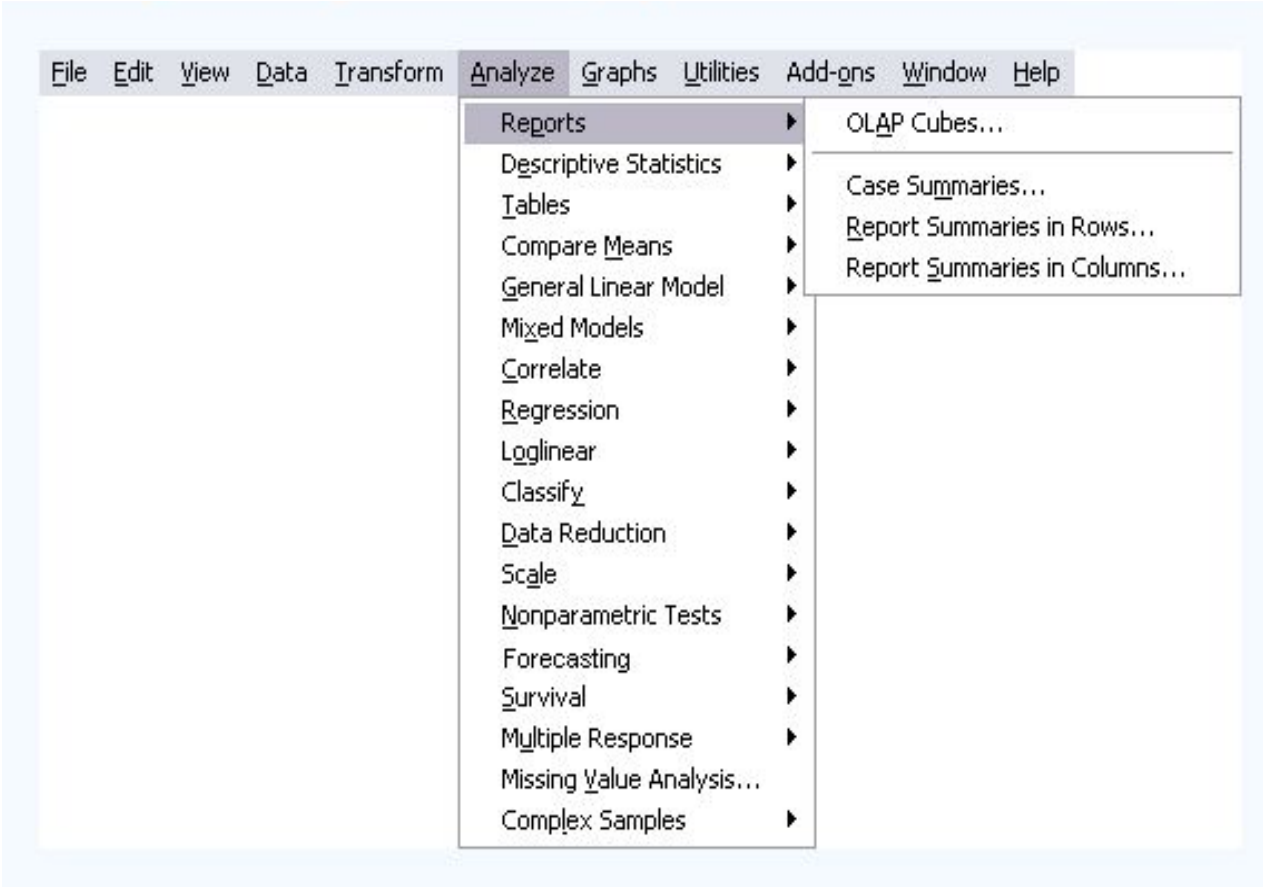




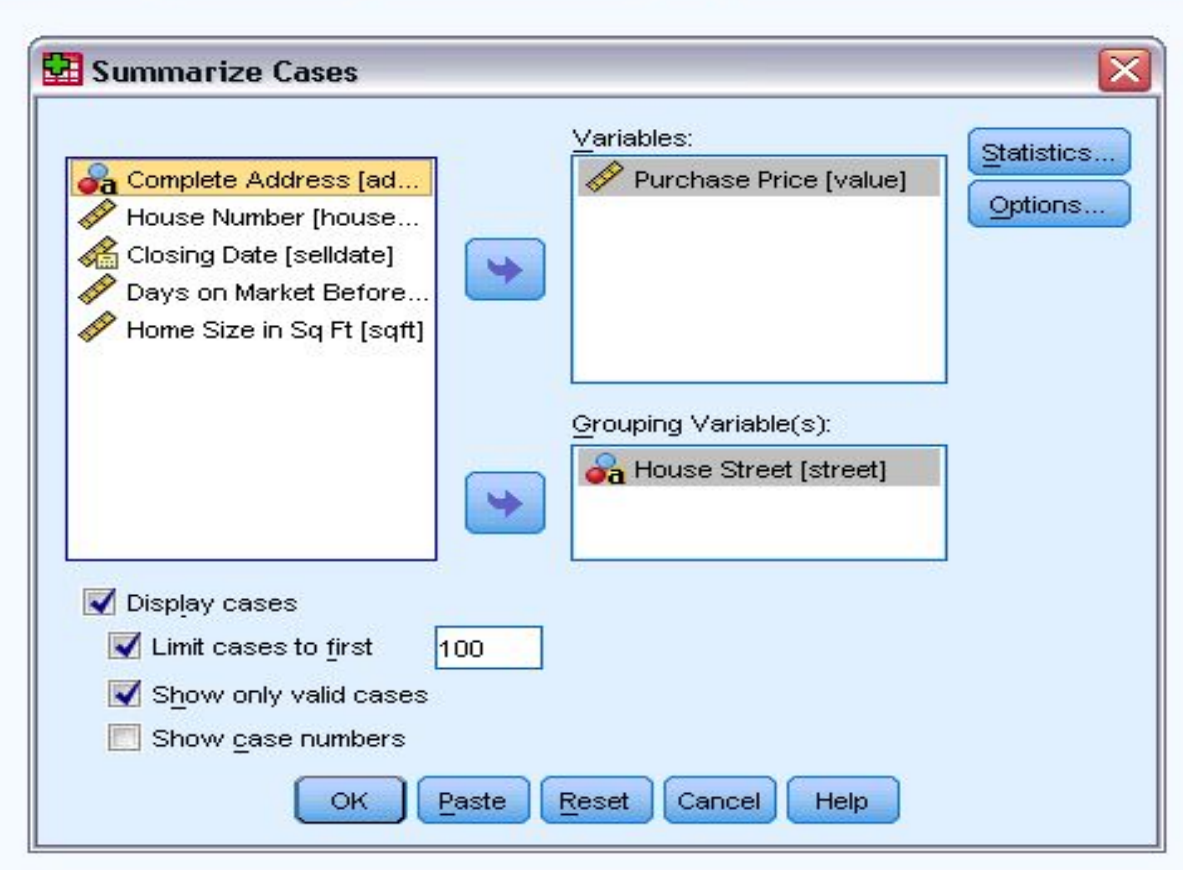


SUMMARY WITH GROUPING VARIABLES IN SPSS

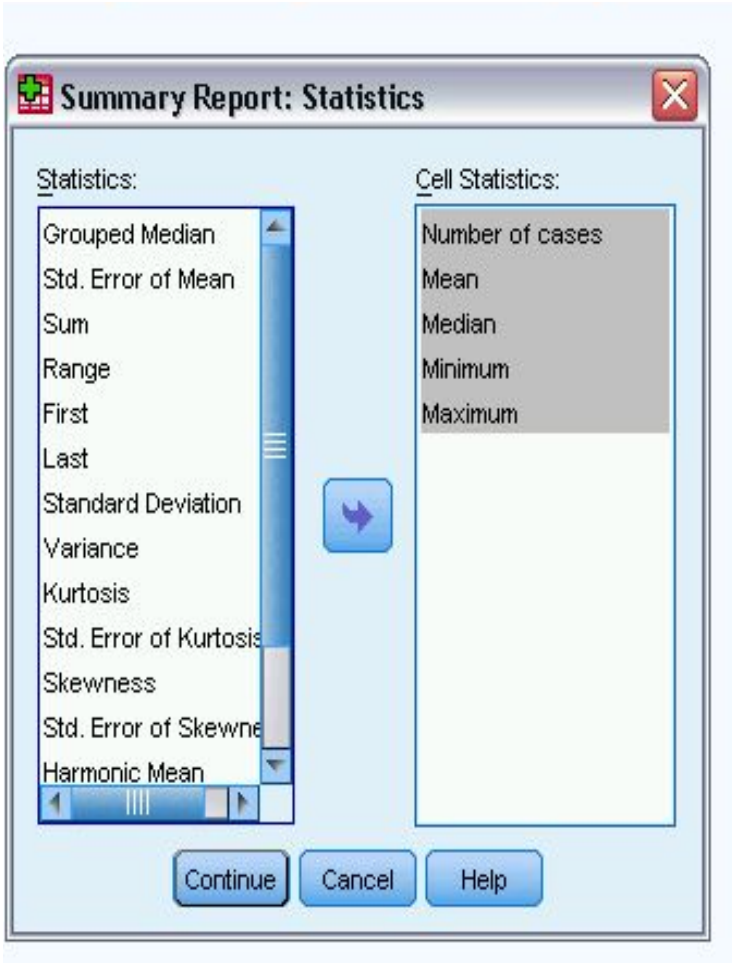
- Analyze – Reports- Case Summaries

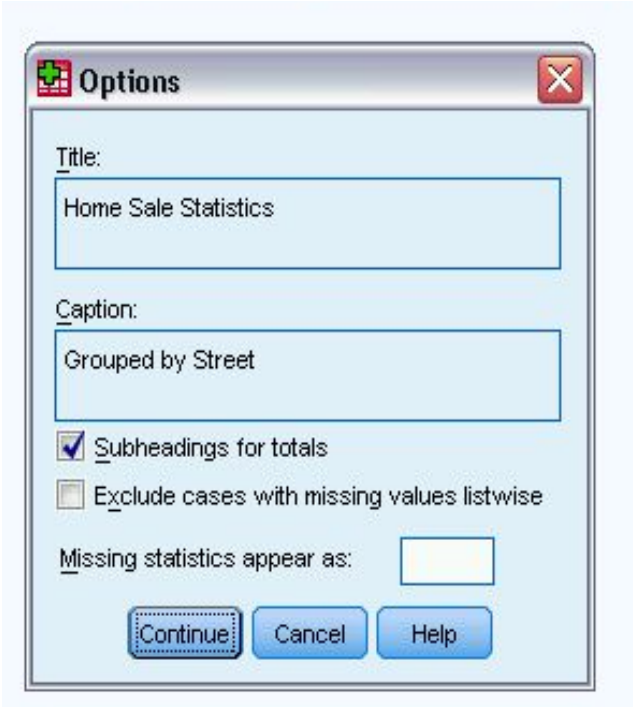


* SELECT THE VARIABLE TO BE SUMMARIZED AND A GROUPING VARIABLE,
* DESELECT “DISPLAY CASES” AND PRESS “STATISTICS”



* SELECT MEAN, MEDIAN, MINIMUM, MAXIMUM (OR ANY OTHER YOU MIGHT NEED)
* CLICK CONTINUE AND CLICK OPTIONS
* IN AN OPTION WINDOW, YOU CAN GIVE A TITLE AND A CAPTION





* YOU GET THE DATA GROUPED BY THE VARIABLE.
* ALL THE DESCRIPTIVES ARE GIVEN FOR EACH VARIABLE, AS WELL AS FOR "TOTAL"

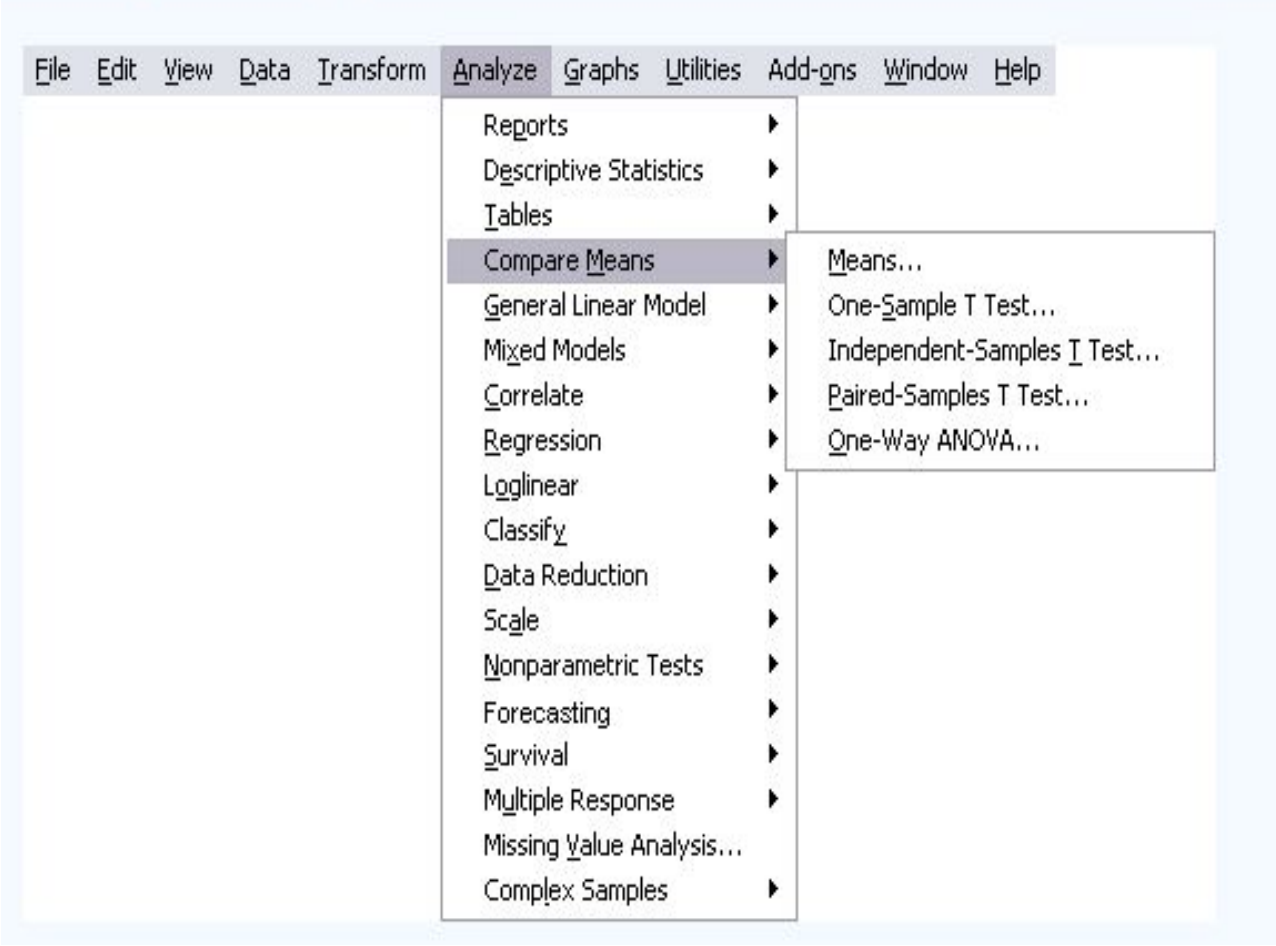
Home Sale Statistics

Purchase Price

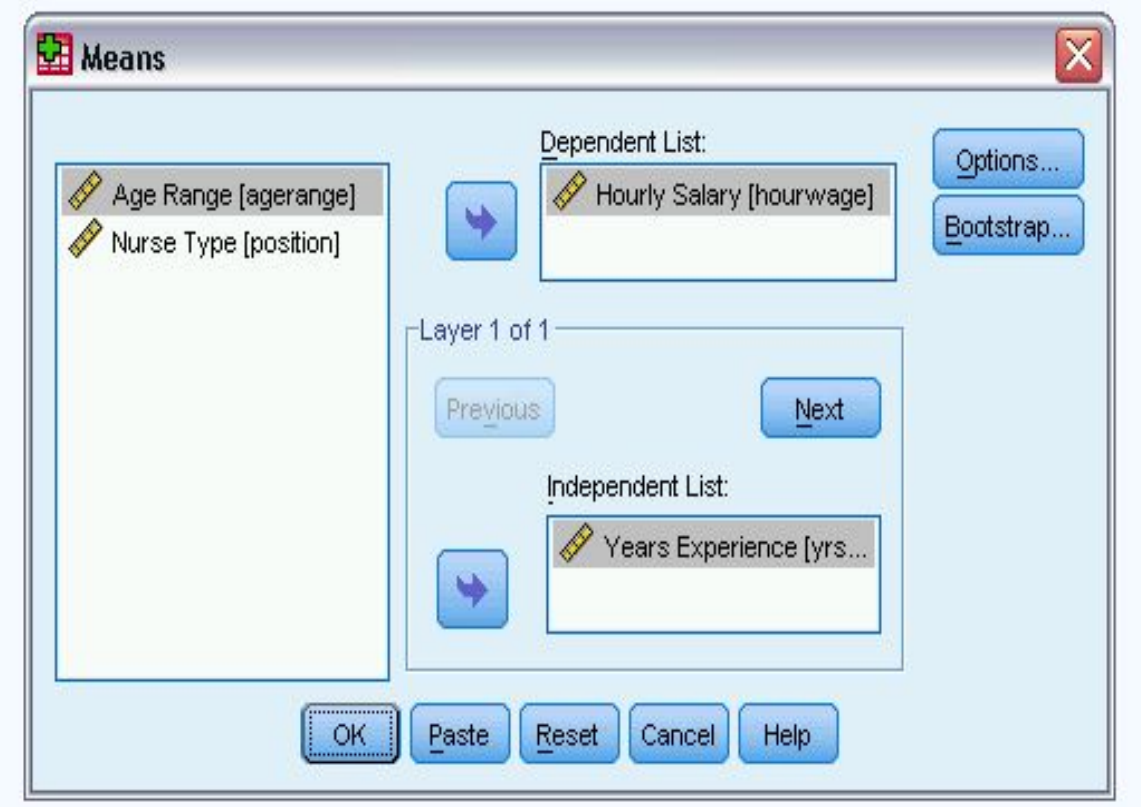
House Street	N	Mean	Median	Minimum	Maximum
Bunker Hill Dr	28	$279,821.43	$281,000.00	$200,000	$416,000
Dawson Ln	23	$131,391.30	$132,000.00	$118,000	$140,000
Fairway View Dr	7	$283,000.00	$271,000.00	$243,000	$343,000
Lakeview Dr	8	$304,000.00	$300,500.00	$289,000	$334,000
Par Dr	7	$303,714.29	$305,000.00	$271,000	$349,000
Persimmon Dr	9	$312,111.11	$300,000.00	$281,000	$351,000
Wintergreen Te	12	$322,833.33	$317,500.00	$273,000	$403,000
Total	94	$256,159.57	$281,000.00	$118,000	$416,000

Grouped by Street

* YOU CAN ALSO LAYER YOUR DATA USING SEVERAL VARIABLES:
* ANALYZE-COMPARE MEANS - MEANS



* ADD THE VARIABLE YOU WANT TO EXAMINE TO “DEPENDENT LIST”
* ADD THE VARIABLES YOU WANT TO GROUP BY TO “INDEPENDENT LIST”
* IF YOU WANT TO HAVE MORE THAN ONE, PRESS NEXT AND ADD THEM



WHAT YOU GET

- Data, grouped by two variables. You get info bout hourly salary, grouped by “years of experience” and “nurse type”.
- What you can see is that nursing salaries seem to take into account both experience and type of work…

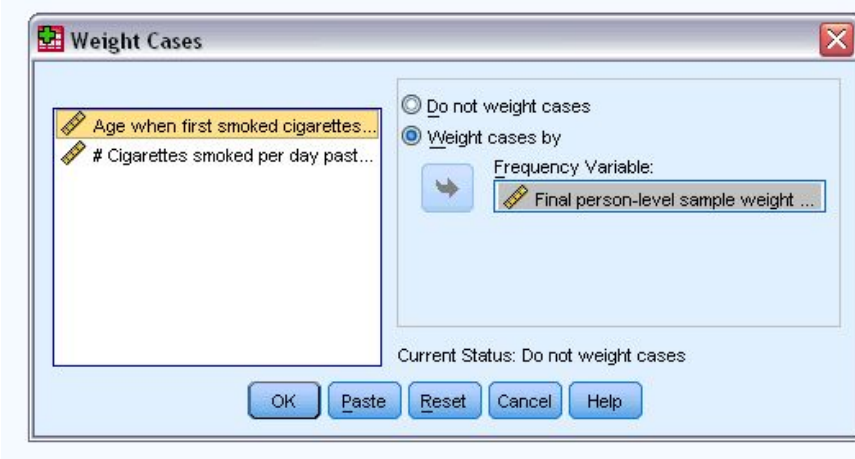
Hourly Salary

Years Experience	Nurse Type	Mean	N	Std. Deviation
5 or less	Hospital	19.0753	147	3.37129
	Office	15.9882	74	3.98762
	Total	18.0416	221	3.86667
6-10	Hospital	19.4846	313	3.35218
	Office	17.7082	147	4.32447
	Total	18.9169	460	3.77816
11-15	Hospital	20.2412	518	3.41065
	Office	18.3784	234	4.57662
	Total	19.6616	752	3.90528
16-20	Hospital	21.1369	471	3.29487
	Office	18.7373	258	4.23293
	Total	20.2876	729	3.82786
21-35	Hospital	21.8601	350	3.48989
	Office	20.1471	189	4.82372
	Total	21.2594	539	4.08669
36 or more	Hospital	22.0641	146	3.14466
	Office	20.6534	64	4.38931
	Total	21.6342	210	3.61826
Total	Hospital	20.6764	1945	3.49582
	Office	18.6859	966	4.58852
	Total	20.0159	2911	4.00309

- You can also select certain cases that follow the rule you choose (using if=, if> and any functions), as well as sort your data.
- (Data-Select Cases; Data-Sort)

One-way ANOVA (Means comparison) as a Bivariate Descriptive Statistic

- Data-Weight Cases
- Analyze-Compare Means-Means
- Choose Statistics you want and check “Anova Table and eta”, as well as “Test for linearity”.
- Anova table shows tests for linear, nonlinear and combined relationship.
- Significance is lower than 0,05 – there is linear relationship.

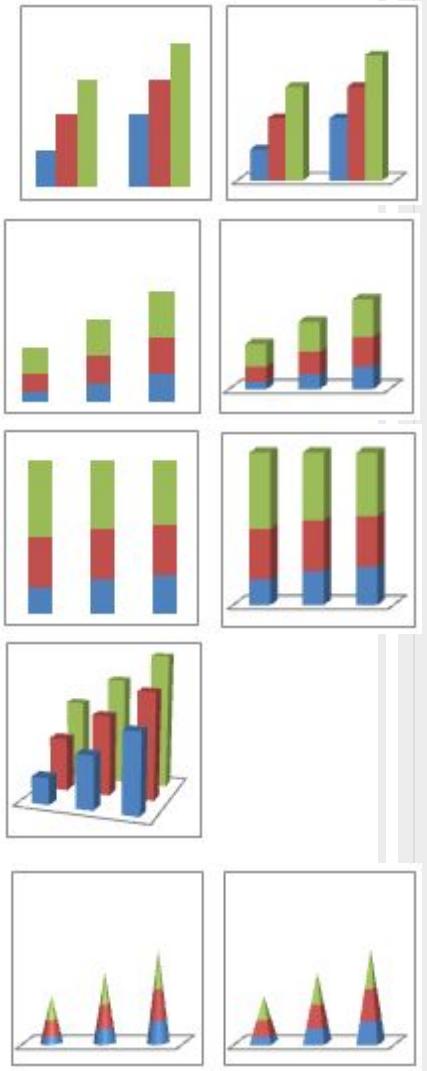


			Sum of Squares	df	Mean Square	F	Sig.
Age when first smoked a cigarette * # Cigarettes smoked per day past 30 days	Between Groups	(Combined)	1974.095	4	493.524	21.158	.000
		Linearity	1321.500	1	1321.500	56.655	.000
		Deviation from Linearity	652.595	3	217.532	9.326	.000
	Within Groups		125841.1	5395	23.326		
	Total		127815.2	5399			

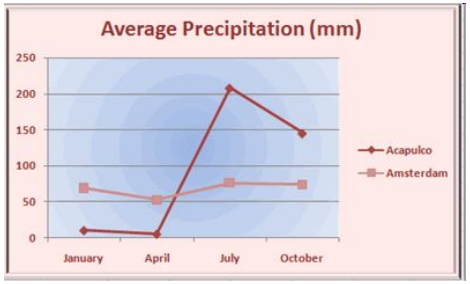
Graphical visualization in Excel and SPSS

Column charts - used to show amounts or the number of times a value occurs.

1. Grouping Histograms - compare values in each category
2. Cumulative Histograms – show the parts of total and their relation to total. They are useful when "totals" are important for us.
3. Standardized Cumulative Histograms. Compare percent that each category contribute to total. Useful for three or more variables and when we want to see each variable's percent of contribution to the total.
4. 3d Histograms compare different values. Can be used to compare data both in categories and in sets.
5. Cone, pyramid and cylinder can be used analogically

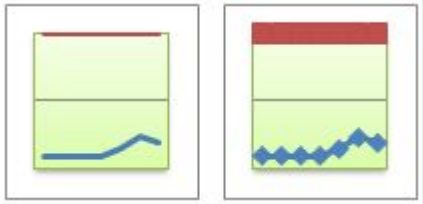


Graphs

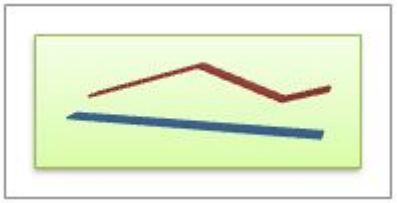


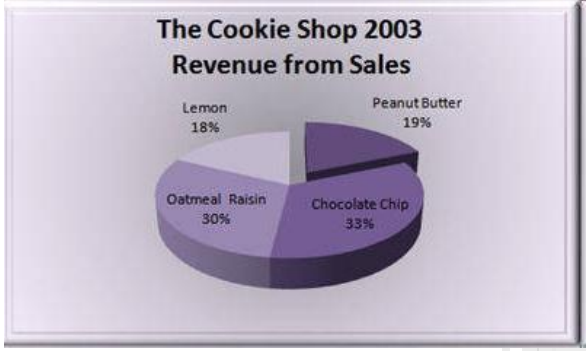
- Can show continuous change of values over time on the same scale. Are perfect for trend visualization

1. Graphs and graphs with markers
2. Cumulative graphs– to show dynamics in contribution of each category to the total.
3. Normalized cumulative graphs – to demonstrate dynamics in percent contribution of each category to the total



4. 3d graph

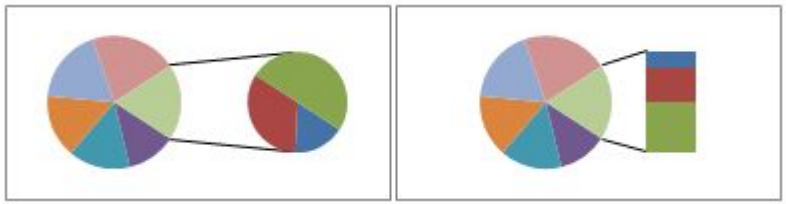


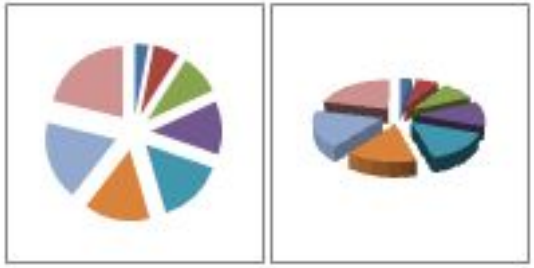


Pie-charts

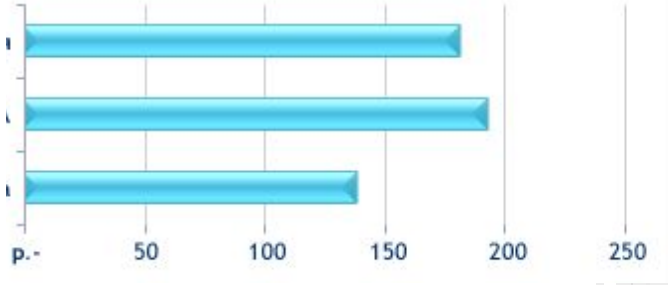
- They are used to chart only one variable at a time. As a result, it can only be used to show percentages.
- The circle of pie charts represents 100%. The circle is subdivided into slices representing data values. The size of each slice shows what part of the 100% it represents.
- All the values should not be lower than 0. There should be no more than 7 categories.

1. Secondary pie chart and secondary histogram show data regarding one of the sectors of a pie chart.
2. Exploded pie chart concentrates on each value



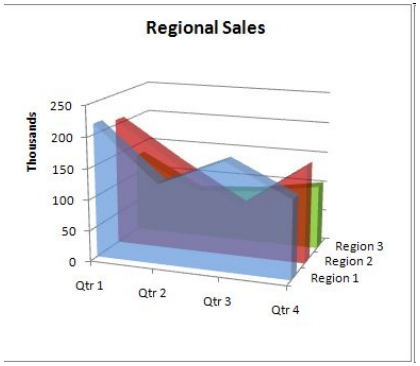


Bar charts

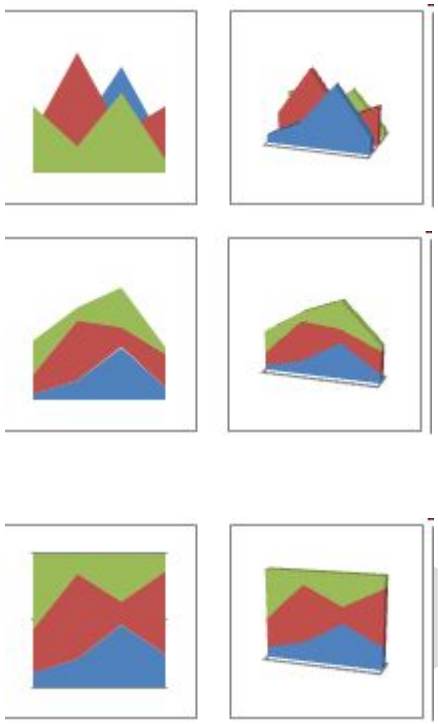


- Are almost the same as histograms, they illustrate comparison of different elements
- These are useful when the axis lables are long, yet we want to see the difference between values.
- The types are the same as in histograms.

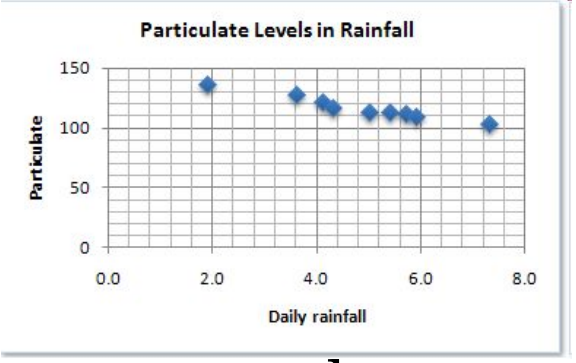
Area chart



- Area charts are much like line charts, but they display different colors in the areas below the lines. This colorful and visual display distinguishes the data more clearly.
- Area charts emphasize the magnitude of change over time and can be used to draw attention to the total value across a trend.
- **2-D area and 3-D area charts** display the trend of values over time or other category As a general rule, you should consider using a line chart instead of a nonstacked area chart, because data from one series can be obscured by data from another series.
- **Stacked area charts** display the trend of the contribution of each value over time or other category data.
- **100% stacked area charts** display the trend of the percentage that each value contributes over time or other category data.

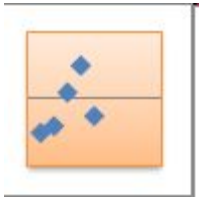


XY (SCATTER) CHARTS

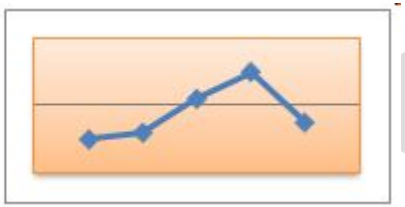


- Scatter charts show the relationships among the numeric values in several data series, or plots two groups of numbers as one series of xy coordinates.

1. Scatter with only markers to compare pairs of values.
2. Scatter with smooth lines and scatter with smooth lines and markers
3. Scatter with straight lines and scatter with straight lines and markers



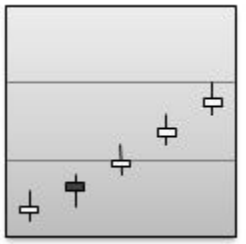




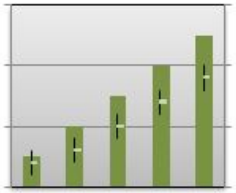
Stock charts

- Most often used to illustrate the fluctuation of stock prices. However, this chart may also be used for scientific data. For example, you could use a stock chart to indicate the fluctuation of daily or annual temperatures. You must organize your data in the correct order to create stock charts.

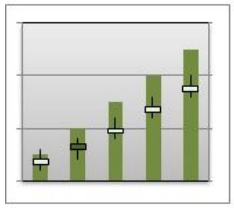
1. **High-low-close** The high-low-close stock chart is often used to illustrate stock prices. It requires three series of values in the following order: high, low, and then close.



2. **Open-high-low-close** This type of stock chart requires four series of values in the correct order (open, high, low, and then close).



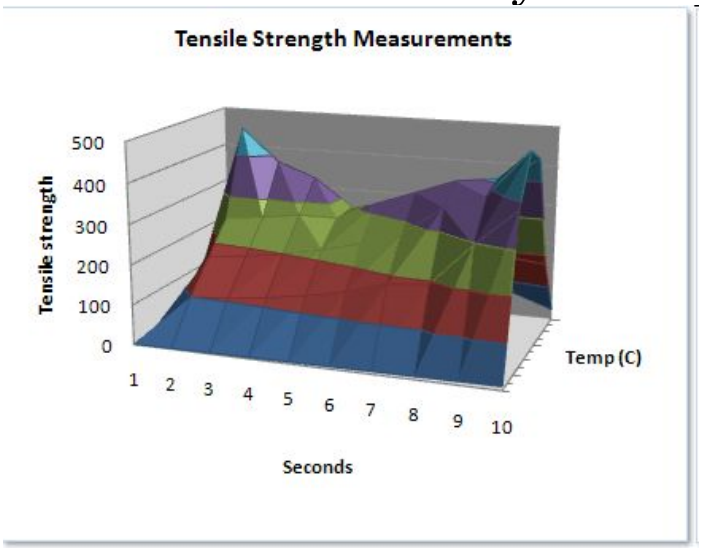
3. **Volume-high-low-close** This type of stock chart requires four series of values in the correct order (volume, high, low, and then close). It measures volume by using two value axes: one for the columns that measure volume, and the other for the stock prices.



4. **Volume-open-high-low-close** This type of stock chart requires five series of values in the correct order (volume, open, high, low, and then close).

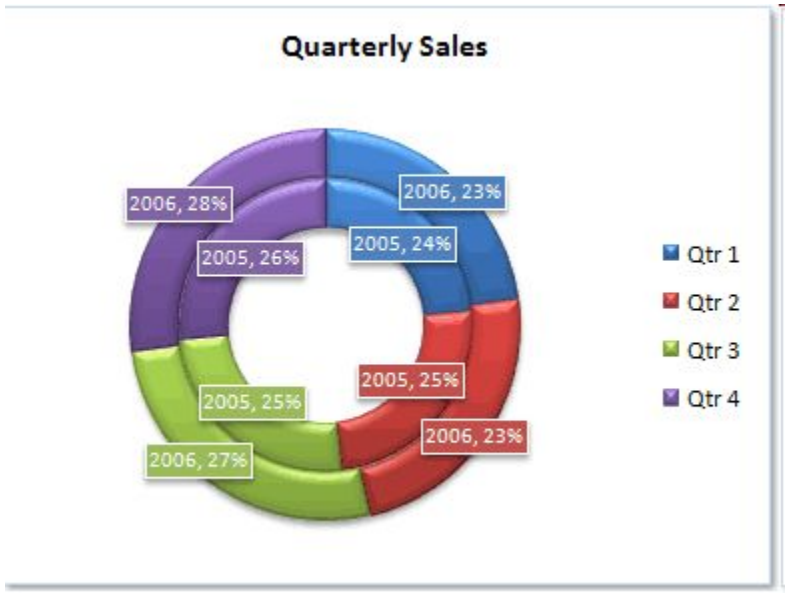
Surface charts

- A surface chart is useful when you want to find optimum combinations between two sets of data. As in a topographic map, colors and patterns indicate areas that are in the same range of values.
- You can use a surface chart when both categories and data series are numeric values.
- Color bands in a surface chart do not represent the data series; they represent the distinction between the values. This chart shows a 3-D view of the data, which can be imagined as a rubber sheet stretched over a 3-D column chart. It is typically used to show relationships between large amounts of data that may otherwise be difficult to see.



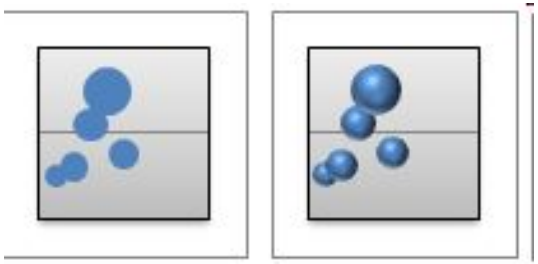
Doughnut charts

- Like a pie chart, a doughnut chart shows the relationship of parts to a whole, but it can contain more than one data series.

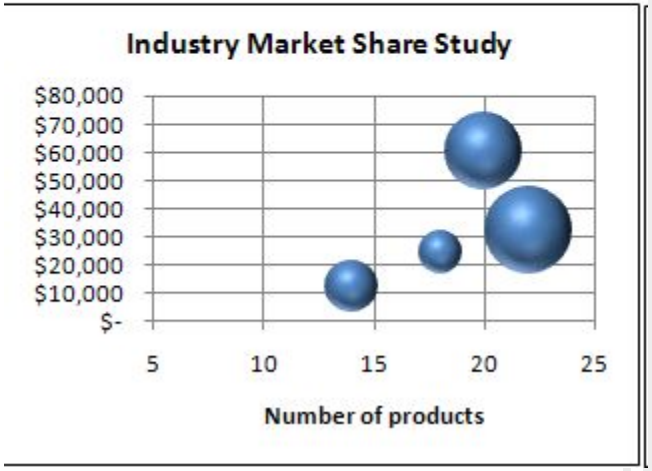


Bubble chart

- **Bubble or bubble with 3-D effect** Both bubble chart types compare sets of three values instead of two. The third value determines the size of the bubble marker. You can choose to display bubbles in 2-D format or with a 3-D effect.

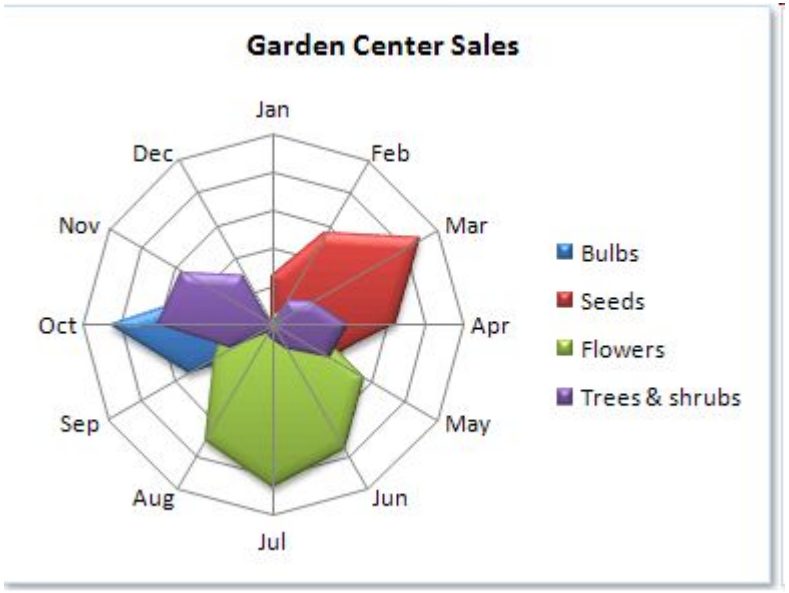


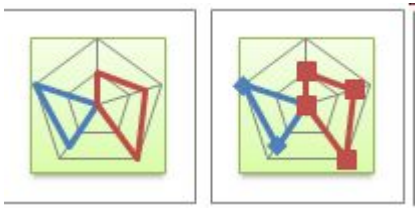
Number of products	Sales	Market Share %
14	$12,200.00	15%
20	$60,000.00	33%
18	$24,400.00	10%
22	$32,000.00	42%



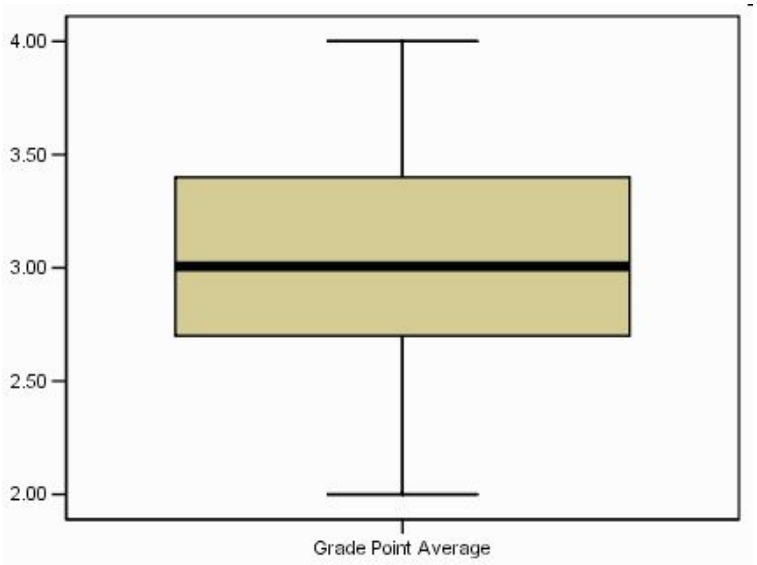
Radar chart

- Radar charts compare the aggregate values of several data series. Radar charts display changes in values relative to a center point.

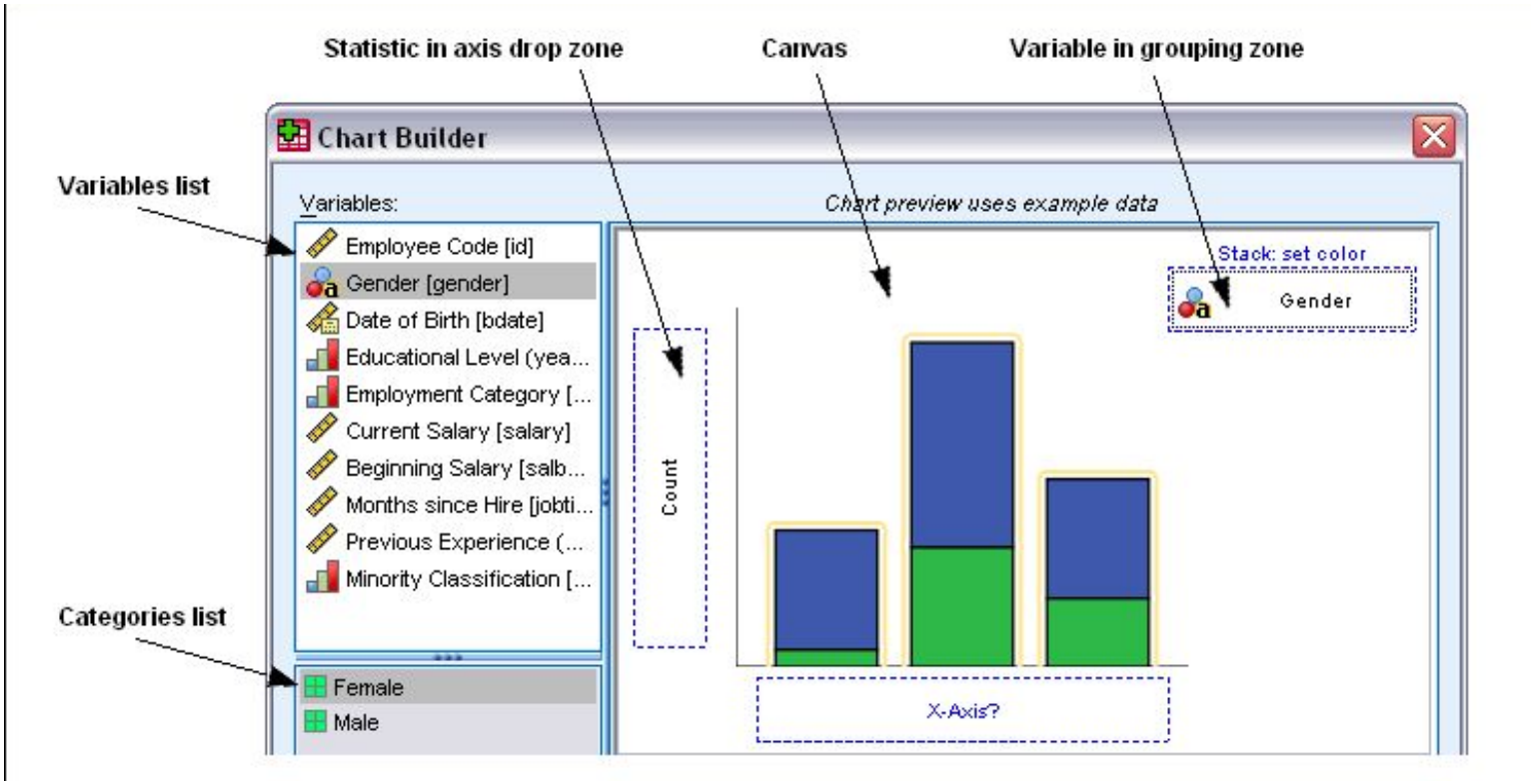




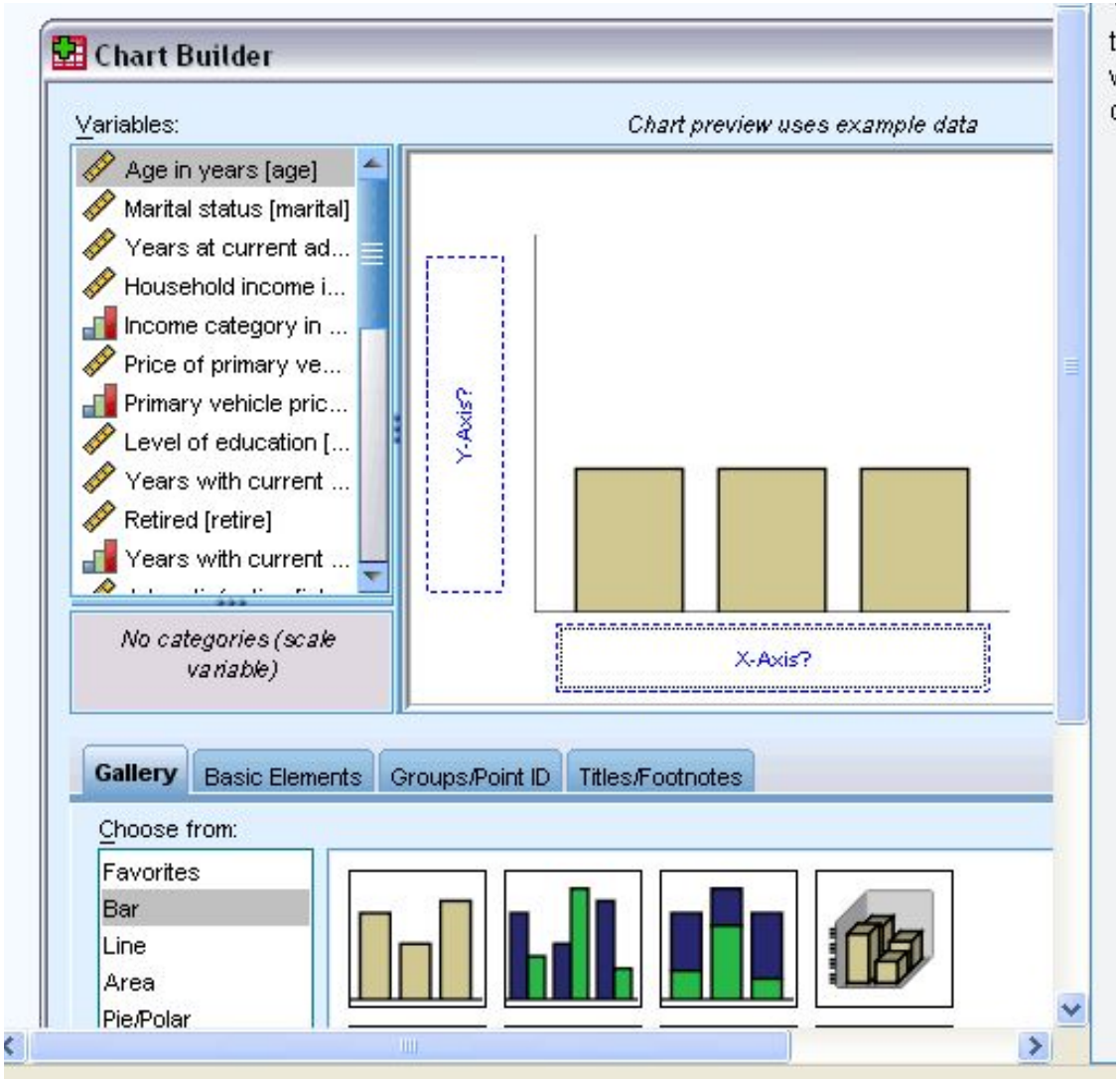
- SPSS has the same graphical visualization types plus a boxplot option.
- A boxplot shows the five statistics: min, 1st quartile, median, 3rd quartile, maximum.
- It is useful for displaying the distribution of a scale variable and pinpoiting outliers (unusual data values).



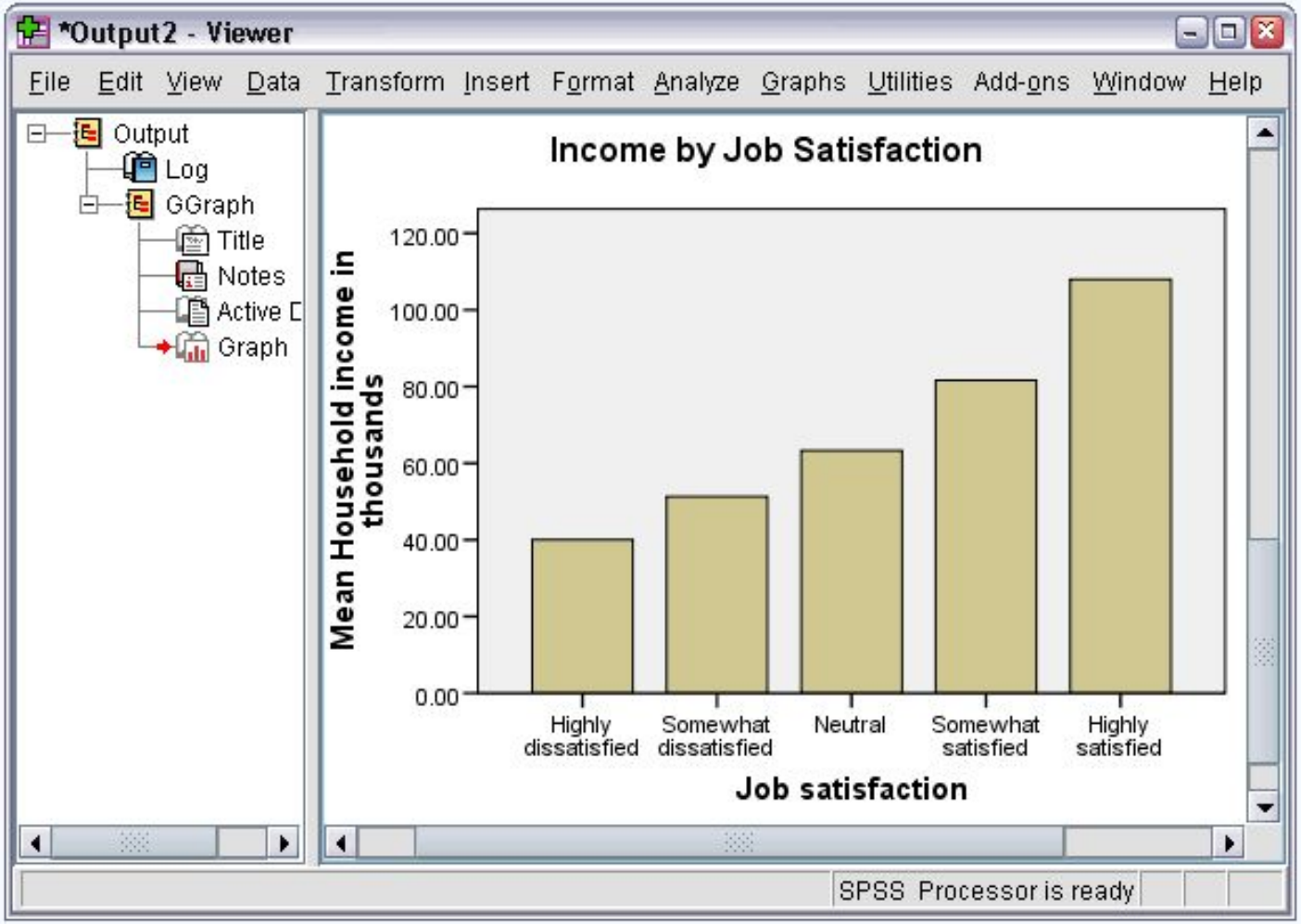
Graphical visualization in SPSS



* Add variables, dragging them from the variables list to the canvas.
* Choose graph type below



I.e. you drag Job Satisfaction to X axis, Household Income to Y axis, choose the type as Bar. That's what you get in the output window:



Any questions?