Tourism Research

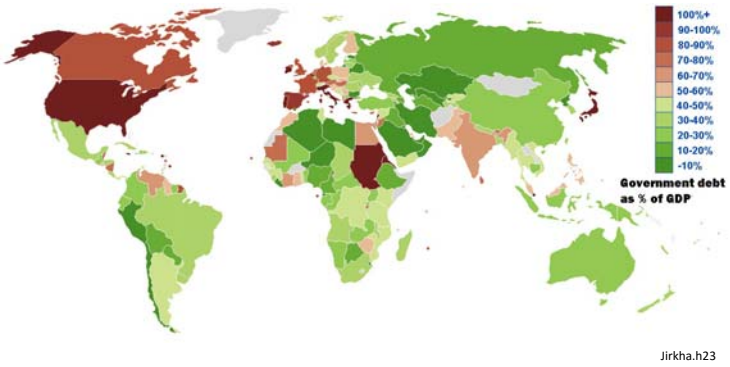
Wk 9 Getting started with SPSS

Lecture 09-1

Today's session…

- Introduces you to the SPSS environment
- Takes you through the preliminary steps from importing data, labelling, to coding, and assigning missing value codes
- Shows you how to recode data into different formats
- Deals with the issue on handling multiple response data

- The importance of this research stage: Growth in time of debt



Jirkha.h23

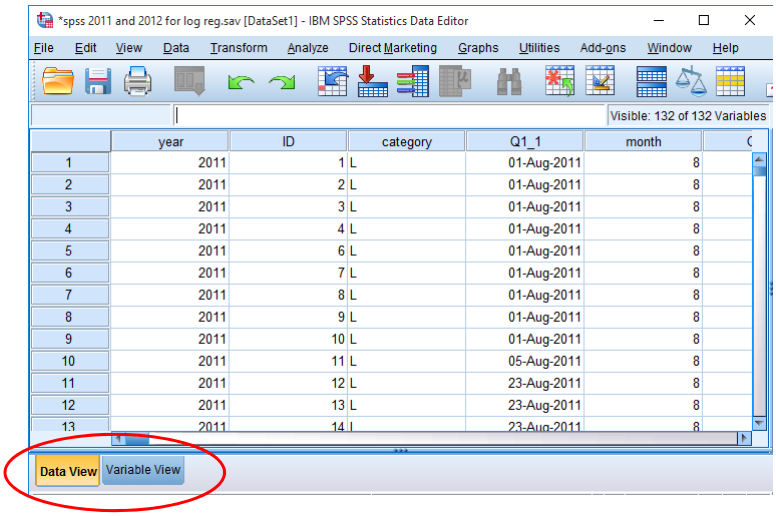
In a much-cited paper, used by conservative governments to defend public budget cuts, Reinhart and Rogoff collected data on public debt relative to gross domestic product and found that once countries crossed the debt threshold of 90% of GDP, economic growth stopped and became negative (-0.1%)

http://www.newyorker.com/rational-irrationality/the-crumbling-case-for-austerity-economics

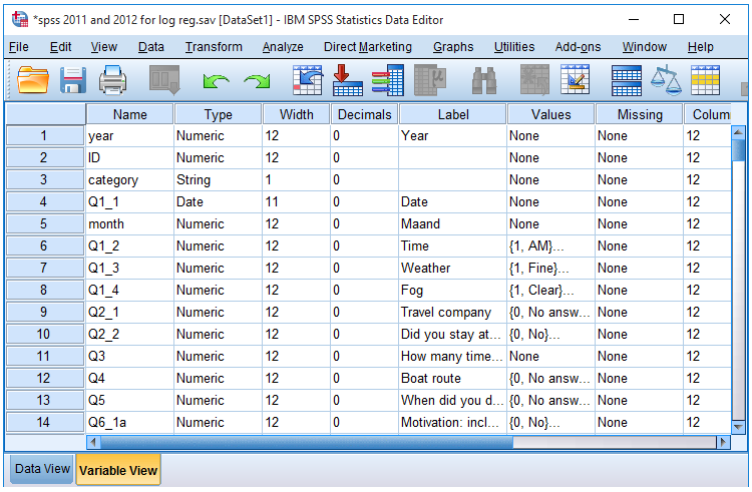
1. Inputting data

- Before attempting any type of data analysis, we need to transfer the data from the questionnaire to a spreadsheet (Excel and/or SPSS). Often it is a good idea to keep the raw data as an Excel sheet and transfer this to SPSS
- IBM SPSS (Statistical Package for the Social Sciences) is the most widely used statistical software package for social sciences, akin to SAS and R
- SPSS is mainly used for quantitative survey data analysis and has tabular and graphical functions
- There are two main ways for getting data into SPSS:
  1. Typing it directly into the spreadsheets
  2. Importing the Excel file

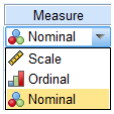
1. Typing data into SPSS:
   - First thing to notice when opening SPSS, there are two tabs: Data and Variable View



- The 'Data View' tab is the one that opens automatically
- In this tab you will code the responses of all your questionnaires
- The columns contain all the variables (i.e. questions or sub-questions of the questionnaire). These are the specific bits of info recorded from every case
- Each row stands for another case or observation (i.e. person that filled in the questionnaire)
- Each cell records a value. This is the particular answer to the question, coded as a numeric value (e.g. 1 = female, 0 = male)



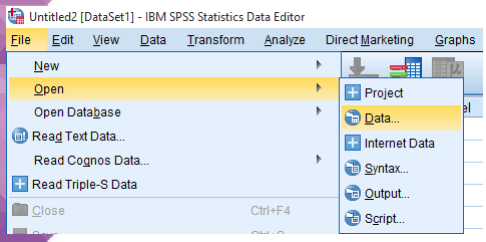
- The 'Variable View' contains all the information on the study variables
- It is a good idea to define the variables before starting to code the answers in Data View
- 'Name' and 'Label' are both used to give a name to the variable. In 'Name' you ought to give a limited coded name to the question, while the 'Label' allows you to include more information (e.g. the actual question)
- 'Type' has to do with the layout of the cell format and will predominantly be numeric
- The 'Measure' column is one of the most important fields and collects data on measurement level of the variable
- 'Values' is a useful field as well, since here you can provide an explanation to the numerical codes

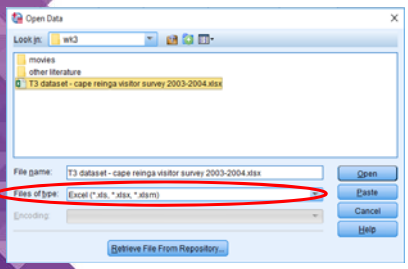


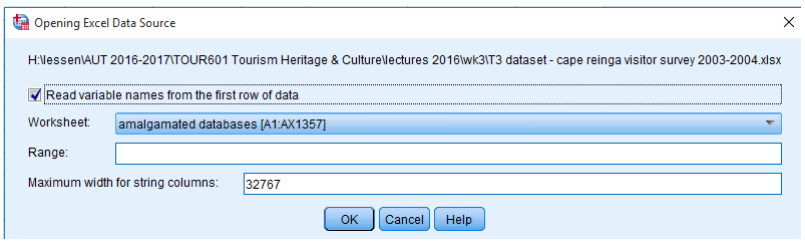
- The measurement level can be 'Scale', 'Ordinal' or 'Nominal'
- Scale refers to what we called ratio variables: i.e. variables with a clear rank to them, equal distance between categories and a true zero point (e.g. age, income, height)
- Ordinal refers to ranked variables (also incorporating interval data here) and is most likely related to Likert scales
- Nominal variables or categorical variables are those that put respondents into different distinguishable categories without a rank order (e.g. gender, employment, nationality)

2. Importing an Excel file:

- It is likely that initial coding (or output from online questionnaires) is collected in an Excel file. These can be easily imported in SPSS
- After importing data in such way, it is still necessary to look to the Variable View tab and define measurement levels and values though





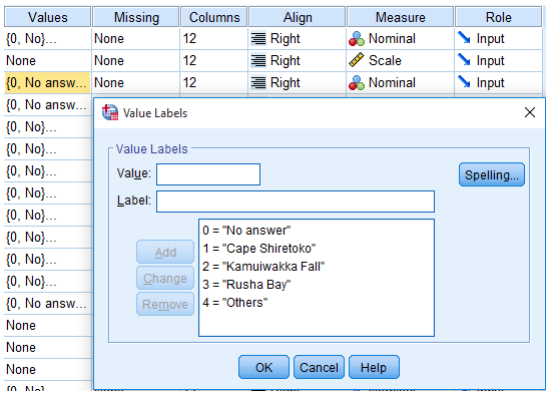


- Go to 'File' → 'Open' → 'Data'
- Make sure to change 'Files of type' to Excel
- You can read the variable names from the first row of the Excel file. So you have to make sure that the Excel file contains either no variable names at the top or only a single row of non-data (i.e. row 1) in order to make the import successful

2. Coding

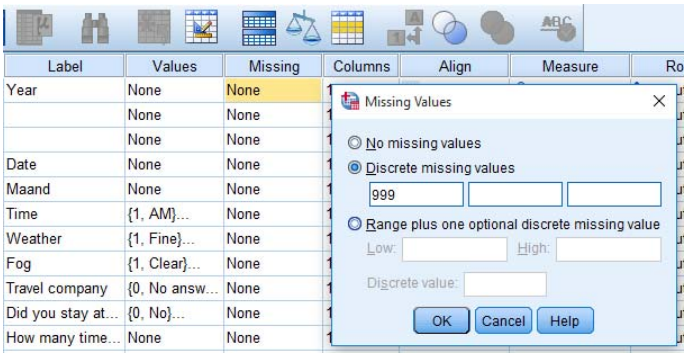
- Normally, you will have already thought about how to code the variables during the development of the questionnaire
- Coding means transforming data collected from its non-computerised form and converting it into number values. Coding ensures that each answer category is uniquely numerically coded
- Coding for quantitative data (ratio and interval/ordinal) should be quite straightforward since the answer is in numbered form already. For ordinal data, the codes have to reflect the rank of answers
- Categorical data, on the other hand, can be coded in multiple ways (e.g. when coding nationality it doesn't matter whether you assign code 1 to New Zealand and code 2 to Germany or the other way around)

- In order to keep a proper overview of the coding you use, and have the SPSS output relate to a meaningful label, instead of a numerical code, it is advisable to design value labels in 'Variable View'



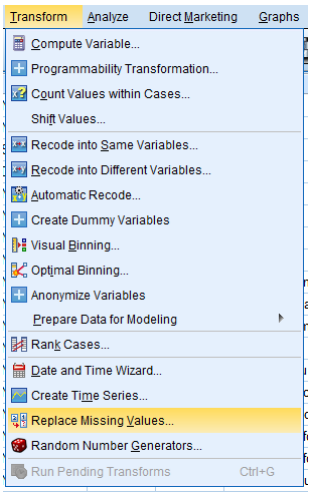
- In the value field, a categorical description can be given to each code. In order to accomplish this, a number gets typed in the 'Value' field and an associated description is given in the 'Label' field
- Generally, the Data View sheet shows the information in numerical values (although this can be toggled), while tabulated results and graphs will make use of the labels

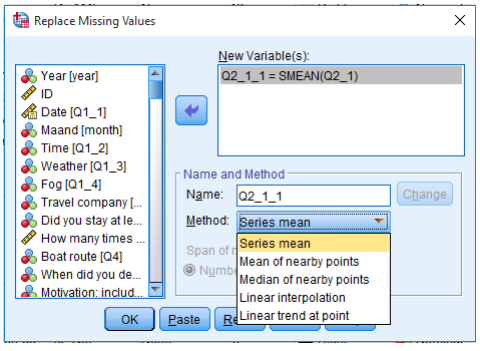
- Missing values are often assigned a specific code (e.g. 999 or 666). It is important to make sure this code does not occur naturally as a genuine answer. In order to make SPSS understand that the code corresponds to a missing (and should thus be ignored), we assign a missing value label in the 'Variable View' tab



- In the Missing field, you can say which discrete code corresponds to a missing value (in this case 999)
- You have the chance to use multiple codes for missing values, or give a range but it is easiest to work with a single code

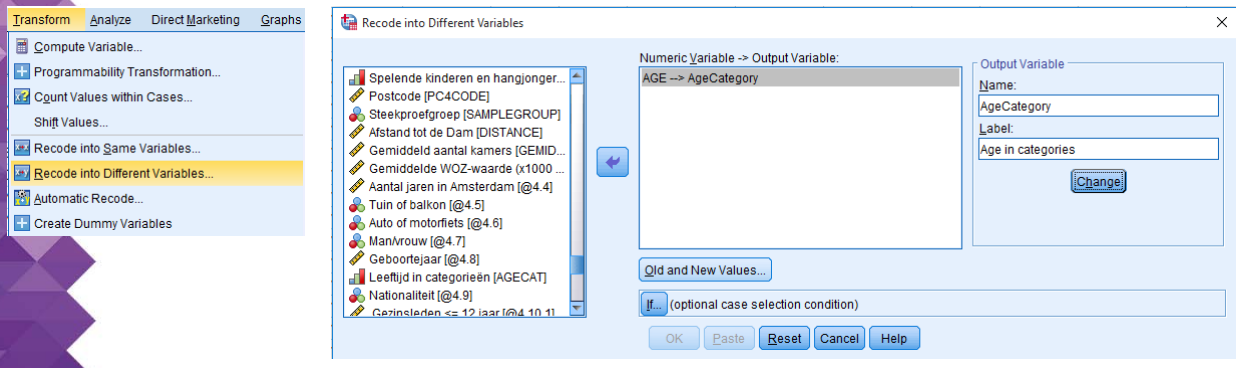
- SPSS also offers the possibility to replace missing values under 'Transform' → 'Replace Missing Values'. However, this should only be considered after analysis of missing value patterns and when the amount of missings is very small

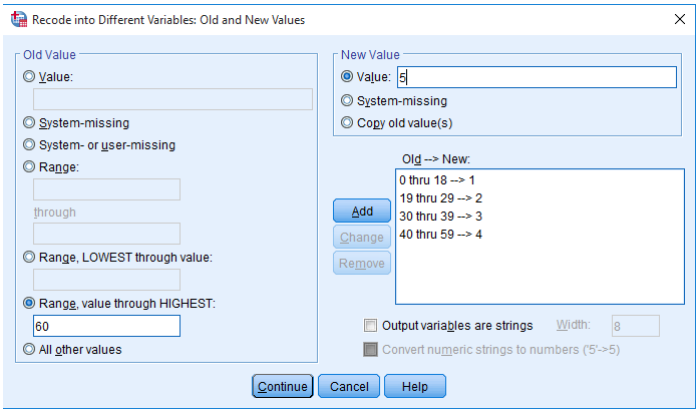




- Replacing a missing field by the series mean is then the choice most often made, although some other methods such as linear interpolation might be considered as well

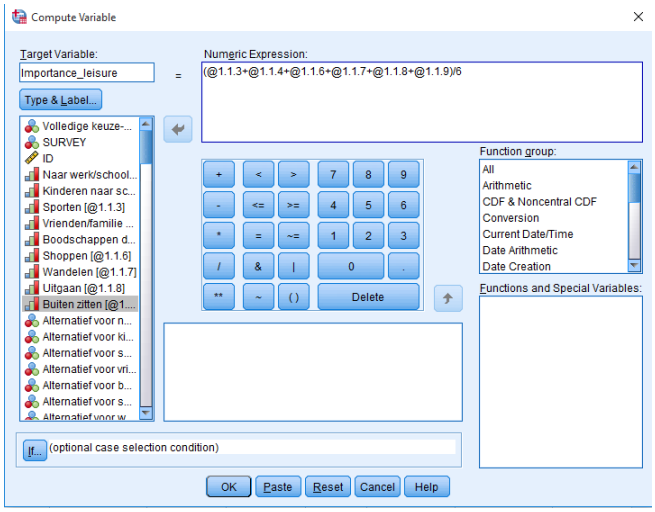
- Another important and related aspect to coding is the option to recode a variable into a new type. This is useful for cases such as age, where you can transform a ratio variable into a limited number of ordinal categories
- Under 'Transform' there are two possibilities to choose from. The safest choice is to 'Recode into Different Variables' since this creates a new variable (i.e. column) while keeping the original data intact. The other option of 'Recoding into Same Variable' will overwrite the original variable and results in a loss of data





- For the recode, you need to specify the old values on the right, and the new value to the left. Add them to the transformation field until each possible value is covered
- The options 'LOWEST through value' and 'value through HIGHEST' allow you to make sure that the first and/or last category incorporate all lowest and highest values without the need to explicitly state the values. E.g. in the above case, the final category, given label 5, will start from the age of 60 and run all the way to the highest age in the dataset

- A final transformation that might be considered via SPSS is the computation of a new variable. This is useful for simple mathematical computations such as making an average between categories or summing up some variables via 'Transform' → 'Compute Variable'



- The Target Variable will be a newly created variable that you can name
- In the Numeric Expression field you write the mathematical equation you want to perform

3. Handling multiple response data

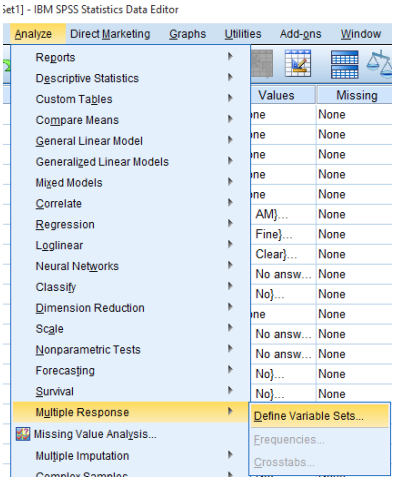
- Multiple response data relates to questions where more than one answer can be given to a survey question, e.g.

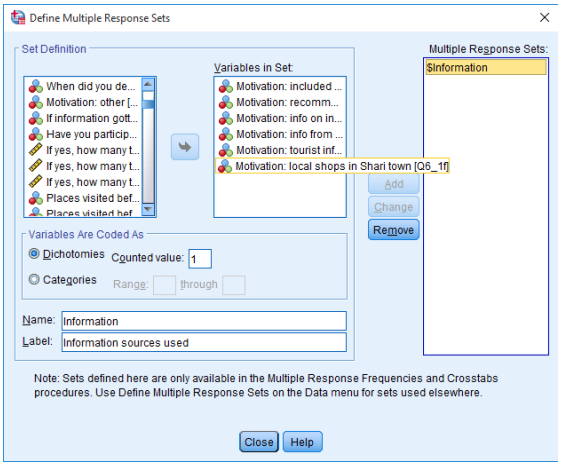
  "How did you get here?" (tick all that apply)

  1) Bus
  2) Bicycle
  3) On foot
  4) Car
  5) Taxi
  6) Train
  7) Other

- Ticking every option that applies makes a difference for data entry and analysis

- Each response needs to be treated as a new variable for the multiple response questions. In first instance, each option is listed in a separate column, giving a value of 0 for not ticked and 1 for ticked
- Then you select 'Analyze' → 'Multiple Response' → 'Define Variable Sets'





- In the Variables in Set field you select all columns (i.e. answer categories that could be chosen)
- You have to tell SPSS how these variables are coded. If you coded them as 0 or 1, you select Dichotomies, Counted value = 1
- Finally you give a name and label to the multiple response set and click 'Add'
- Note that these created multiple response sets can only be used for Frequencies and Crosstabs, they will not be available for other procedures
- In order to generate one of these, go back to 'Analyze' → 'Multiple Response'. Now, after the response set has been created, the options 'Frequencies' and 'Crosstabs' will be available to you

$Information Frequencies

		Responses		Percent of Cases
		N	Percent	
Information sources used[a]	Motivation: included in tour plan	566	31.5%	36.0%
	Motivation: recommended by friends or family	189	10.5%	12.0%
	Motivation: info on internet or from guide book	801	44.6%	51.0%
	Motivation: info from hotels, B&B, etc.	96	5.3%	6.1%
	Motivation: tourist info centre, UNESCO WH centre	140	7.8%	8.9%
	Motivation: local shops in Shari town	4	0.2%	0.3%
Total		1796	100.0%	114.3%

a. Dichotomy group tabulated at value 1.

Conclusion

- This lecture focused on the elementary process of importing data and defining labels and codes
- While not leading to any initial analysis in itself, these initial steps are essential in order to ensure that the subsequent data analysis can run smoothly
- The next step after data input and coding will be to draw some descriptive statistics

Tourism Research

Wk 10 Descriptive data analysis and data exploration with graphs in SPSS

Lecture 10-1

Today's session…

- Primarily focuses on descriptive statistics using SPSS
- Introduces the graphical function of SPSS to give a first overview of data characteristics of single variables as well as looking into exploratory statistics (mean, median, mode) and simple frequency tables for single variable analysis
- Describes how we can use both graphs and other exploratory statistics to start identifying patterns and similarities/differences between groups of variables

- The use of descriptives: A standard deviation of 76.2 in age of respondents

Table 3: Descriptive statistics of the sample

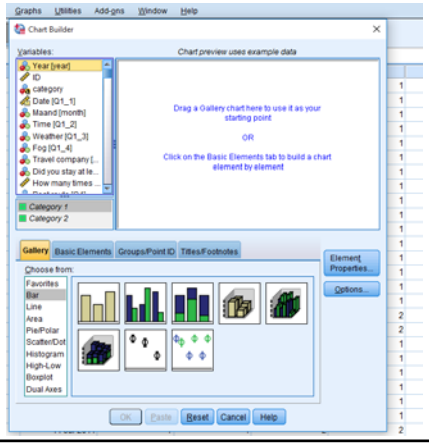
Variable	Frequency / Mean *(std. dev)*	N
Female	49.1%	701
Belgian nationality	97.0%	699
University degree	23.3%	681
Household income after tax:		543
≤€1,000	3.5%	
€1,001-2,000	31.7%	
€2,001-3,000	32.2%	
€3,001-4,000	21.9%	
>€4,000	10.7%	
Employment:		666
Employee, not management	37.5%	
Employee, management	17.3%	
Self-employed	6.8%	
Teacher	5.0%	
Unemployed	3.0%	
Housekeeper	1.1%	
Retired	27.6%	
Student	0.3%	
Other	1.5%	
Age	54.1 *(76.2)*	701
Length of residence	34.2 *(20.3)*	684
Average distance to study area	4,223.0 *(2,389.4)*	699

In one of my own draft research papers, the age of respondents was found to have a mean age of 54.1 and a standard deviation around this mean of 76.2. Any thoughts on this?

1. Individual variable analysis

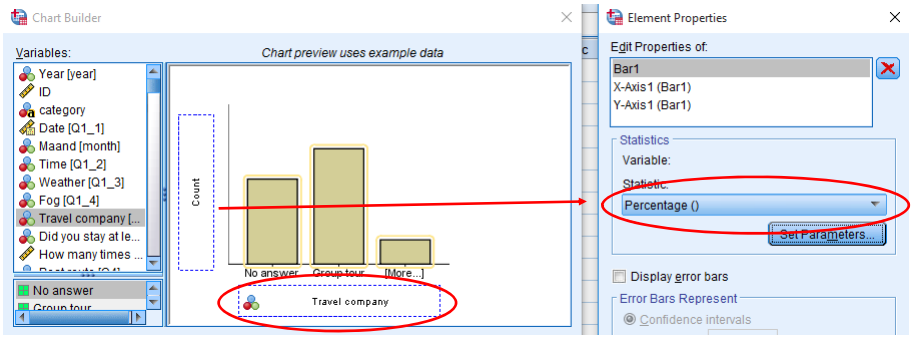
a. <u>Exploring individual variables with graphs:</u>

- In SPSS go to 'Graphs' → 'Chart Builder' to open basic functionalities

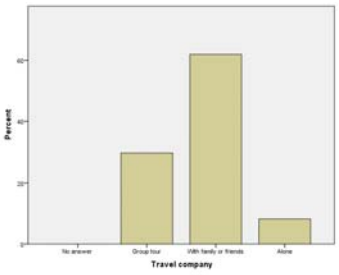


- The programme will remind you of the need to set your measurement levels properly and add labels to your nominal variables
- In the bottom field, a choice can be made to work with a pre-defined type of graph through 'Gallery' or to create a graph from scratch through 'Basic Elements'. For most purposes, the 'Gallery' will be all you need
- The most common types of graphs are included. For nominal and ordinal variables, the most likely choices will be 'Bar' or 'Histogram', while for scale variables a 'Boxplot' can be interesting as well
- Start by dragging the chosen graph style to the canvas. Then you can drag the variable you want to visualize to the canvas as well

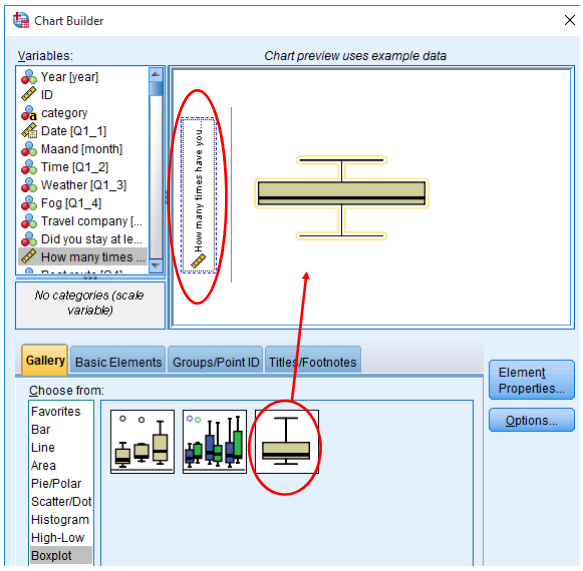
1. Constructing a ‘Histogram’-graph to show frequencies of ‘Travel Company’ (nominal)



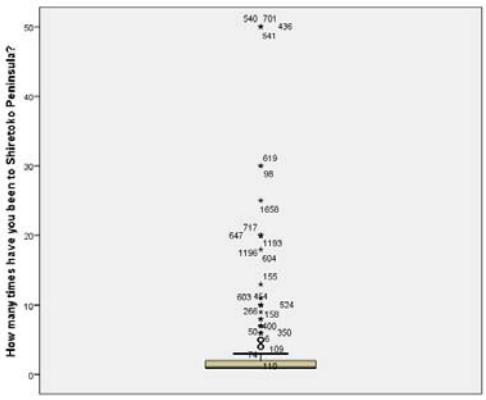
- After dragging the simple histogram-chart to the canvas, you then drag the variable ‘Travel company’ to the horizontal X-axis
- The Y-axis standard is set to Count. If you want to show a different statistic, you can change this in the right field under ‘Statistics’. Don’t forget to click ‘Apply’ when you do!
- The constructed graph will then show in the output viewer of SPSS



2. Constructing a ‘Boxplot’-graph to show frequencies of ‘Times visited Shiretoko’ (ratio)

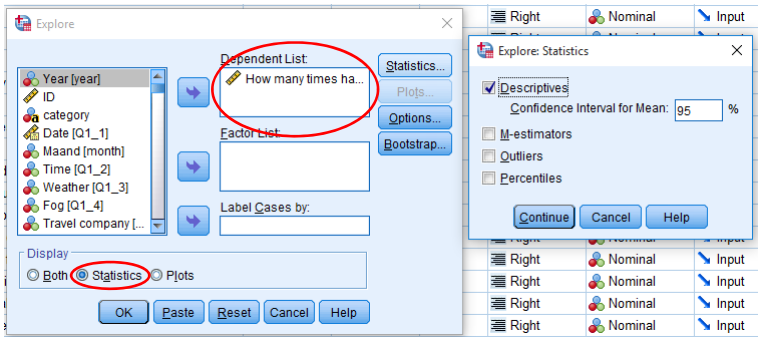


- After dragging the simple boxplot-chart to the canvas, you then drag the variable ‘Times visited’ to the vertical Y-axis
- The boxplot shows the median, the upper and lower quartile (i.e. middle 50% of data), and the top and bottom 25% in the whiskers, as well as meaningful outliers



b. Exploring individual variables with descriptive statistics:

1. Getting measures of central tendency via 'Analyze' → 'Descriptive Statistics' → 'Explore'



- For an analysis of individual variables, drag the variables you want to investigate to the 'Dependent List' field. The other two fields stay empty
- You can select basic plots, statistics or both. Now we are just interested in the statistics
- Clicking on the 'Statistics' button will open a new window that asks you which statistics you want to explore. For our purposes it is enough to select 'Descriptives'

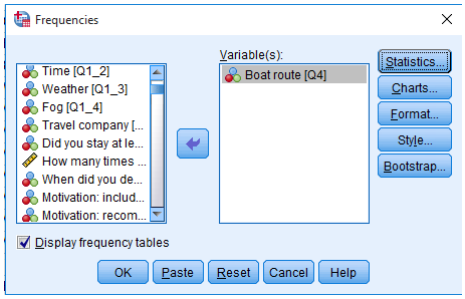
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
How many times have you been to Shiretoko Peninsula?	1678	98.5%	25	1.5%	1703	100.0%

Descriptives

			Statistic	Std. Error
How many times have you been to Shiretoko Peninsula?	Mean		1.94	.077
	95% Confidence Interval for Mean	Lower Bound	1.79	
		Upper Bound	2.09	
	5% Trimmed Mean		1.51	
	Median		1.00	
	Variance		9.967	
	Std. Deviation		3.157	
	Minimum		1	
	Maximum		50	
	Range		49	
	Interquartile Range		1	
	Skewness		10.490	.060
	Kurtosis		139.620	.119

2. Getting frequency distribution table via 'Analyze' → 'Descriptive Statistics' → 'Frequencies'



Statistics

Boat route

N	Valid	1382
	Missing	321

Boat route

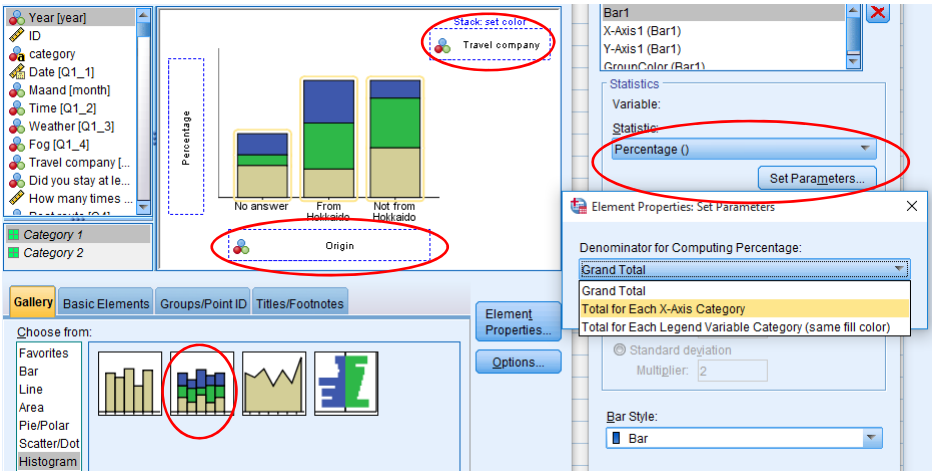
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Cape Shiretoko	854	50.1	61.8	61.8
	Kamuiwakka Fall	370	21.7	26.8	88.6
	Rusha Bay	151	8.9	10.9	99.5
	Others	7	.4	.5	100.0
	Total	1382	81.2	100.0	
Missing	System	321	18.8		
Total		1703	100.0		

- For our purposes here it will be enough to just select the variables we want to make a frequency table of
- Make sure the variables are either ordinal or nominal. Frequency tables for scale variables don't make much sense and are too hard to interpret
- In the 'Statistics' tab, you can again select for a number of descriptive central tendency statistics if wanted

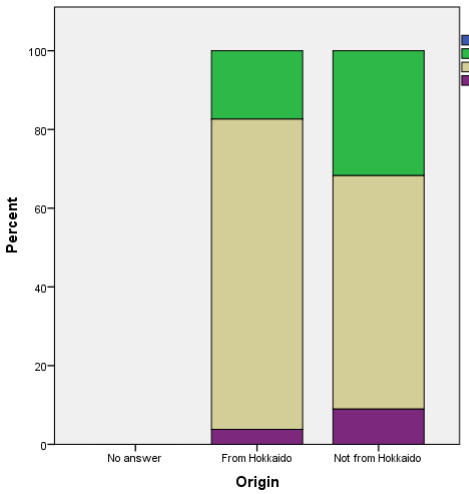
2. Multiple variable analysis

a. <u>Exploring multiple variables with graphs:</u>
   - Basically the same graphs that are useful to identify distribution of individual variables, can also be applied in order to understand potential group differences
   - The key difference will be the inclusion of a categorical/nominal variable to split the data

1. Constructing a stacked histogram to show frequencies of 'Travel Company' (nominal) by 'Origin' (nominal)

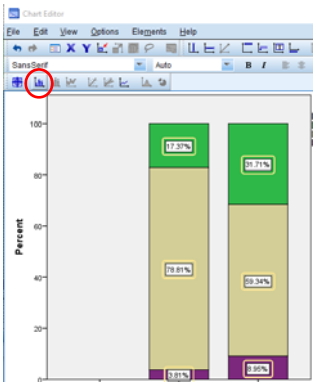


- The constructed graph will then show in the output viewer of SPSS

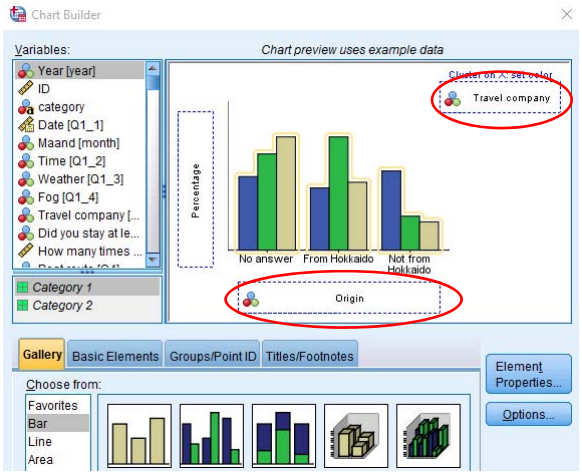


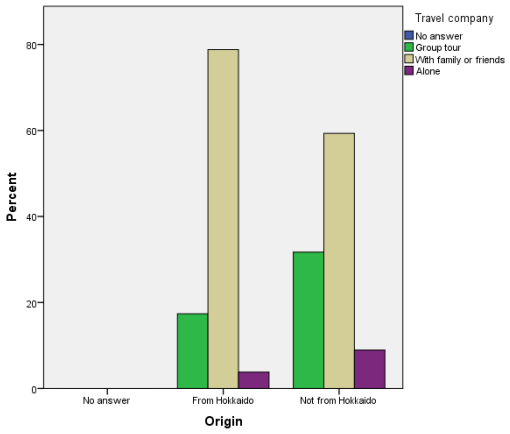
- The classification variable goes in the X-axis and the variable to be compared across classes is set in the 'Stack set color' field
- In order to improve comparability, it is advisable to change the Statistics to 'Percentage' and 'Total for Each X-Axis Category'. This way each X-Axis classification will add up to 100%.

- Double-click on graph in the output window to open the chart editor and add data labels to the columns



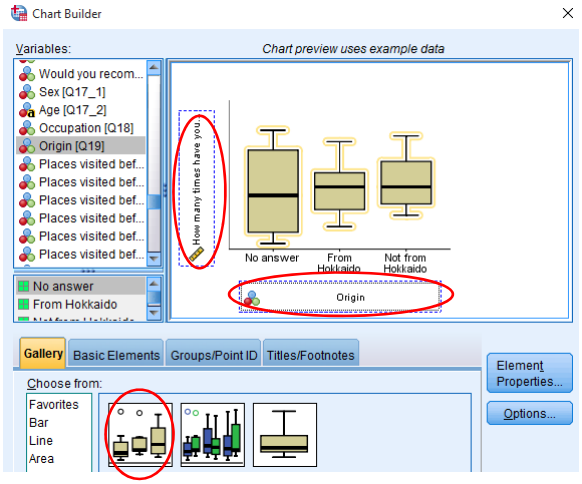
2. Constructing a clustered bar chart as another way to show frequencies of 'Travel Company' (nominal) by 'Origin' (nominal)

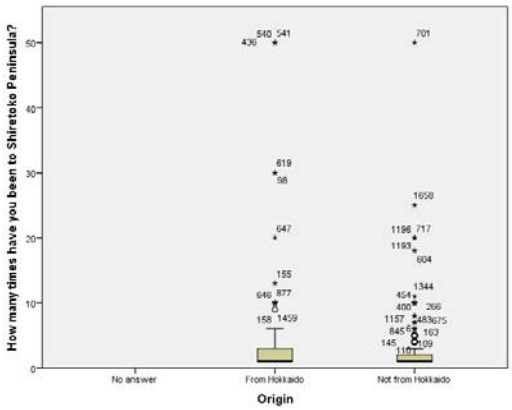




- Analysis very similar to stacked histogram, only output looks a bit different

3. Constructing a simple boxplot show differences in median and spread of 'Times visited Shiretoko' (ratio) by 'Origin' (nominal)

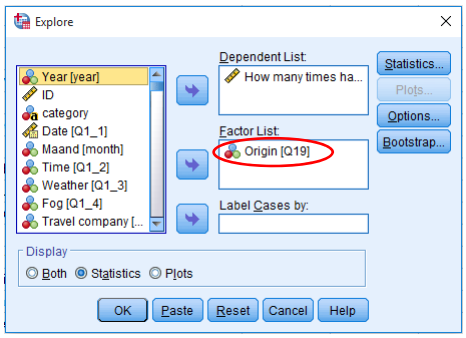




- The classification variable 'Origin' goes into the X-axis field and the ratio variable in the Y-Axis field
- The results show the difference in spread of data for these two categories

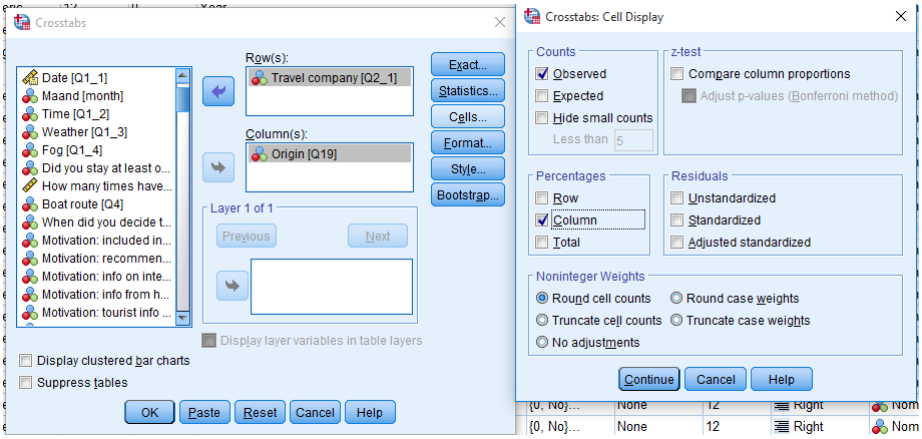
b. Exploring multiple variables with descriptive statistics:

1. Getting comparative measures of central tendency via 'Analyze' → 'Descriptive Statistics' → 'Explore'



- Similar to the individual variable analysis, the variable under investigation is dragged to the 'Dependent List' field
- However, this time we also drag a grouping variable into the 'Factor List' field
- The result will be that descriptive statistics are given not for the whole of data but for each category of the 'Origin' variable. Therefore this gives us a tool to compare central tendency values such as median and mean

2. Getting bivariate crosstabs via 'Analyze' → 'Descriptive Statistics' → 'Crosstabs'



- Generally, the independent variable will go into the 'Column(s)' field and the dependent variable will go into the 'Row(s)' field
- Clicking on the 'Cells' tab gives you a new window in which you can select the option to show percentages on the row or column total. For interpretation it will make most sense to calculate a percentage on 'Column'

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Travel company * Origin	1655	97.2%	48	2.8%	1703	100.0%

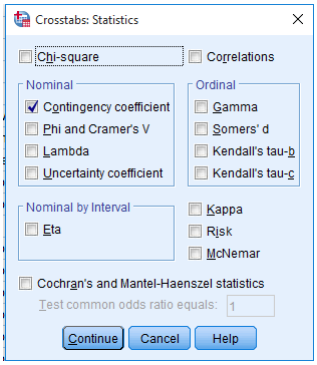
Travel company * Origin Crosstabulation

			Origin		
			From Hokkaido	Not from Hokkaido	Total
Travel company	Group tour	Count	41	450	491
		% within Origin	17.4%	31.7%	29.7%
	With family or friends	Count	186	842	1028
		% within Origin	78.8%	59.3%	62.1%
	Alone	Count	9	127	136
		% within Origin	3.8%	8.9%	8.2%
Total		Count	236	1419	1655
		% within Origin	100.0%	100.0%	100.0%

- Finally, the 'Statistics' tab under Crosstabs allows for a first venture into inferential statistics
- Depending on the measurement level of the table variables, different statistics on association and correlation can be selected, offering a statistical significance test to the difference in frequencies found
- E.g. in this case the Contingency coefficient shows that with a significance level < 0.05, the association between origin of tourists and travel company is significant. We could therefore conclude that people from outside the Hokkaido region travel differently in terms of travel company than locals

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Contingency Coefficient	.140	.000
N of Valid Cases		1655	



Conclusion

- Descriptive statistics are the first step in data analysis. At this point possible mistakes in data input can be identified and initial patterns in the data might become clear
- Descriptive analysis can use both graphs and tables to showcase the data and concentrate on a single variable or introduce a classification variable as well
- Crosstabs are probably the most useful descriptive analysis to start recognising associations on a very basic level