

| o apply rules to a variable, select the variab             |                     |               |               |            |                    |   |
|--|---------------------|---------------|---------------|------------|--------------------|---|
| ne Analysis Variables list shows distribution<br>ariables. | ns of nonmissing va | lues based on | a scan of th  | e data. Th | ne Rules list show | is all rules that can be applied to selected    |
| nalysis Variables:   |                     |               |               |            | Rules:             |   |
| 'ariable   | Distribution        | Minimum       | Maximum       | Rul        | Apply              | Name  |
|  |                     |               |               | -          |                    | 0,1 dichotomy                                   |
| Predominant religion [religion]                            |                     |               | Tribel        | 0          |                    | 1,2 dichotomy                                   |
|  |                     |               |               |            |                    | 1 to 5 integer                                  |
| Average female life expectancy [lifeexpf]                  | - du                | 43 82         |               |            |                    | 1 to 10 integer                                 |
| verage remain are expectancy (meexpi)                      | George Carlling     |               |               |            |                    | Nonnegative number                              |
| Average male life expectancy [lifeexpm]                    | alla                |               |               |            |                    | Nonnegative integer                             |
|  | Beend in            | 41            | 76            |            |                    | 0 to 100 number                                 |
|  |                     |               |               |            |                    | Flag system-missing values                      |
| eople who read (%) [Iteracy]                               | annead an annead    | 18            | 100           |            |                    | Flag user-missing values                        |
|  |                     |               |               |            |                    | Flag missing values                             |
| opulation increase (% per year)) [pop_incr                 |                     | -0.3          | 5.2           |            |                    | Flag noninteger values<br>Flag unlabeled values |
| oparation increase ( in per year)) [pop_incr               |                     |               | 0.2           |            |                    | riag unlabeled values                           |
| nfant mortality (deaths per 1000 live births)              | man.                | 4.0           | 168.0         |            |                    |   |
| Gross domestic product / capita [gdp_cap]                  | bl                  | 122           | 23474         |            |                    |   |
| Region or economic group [region]                          |                     | 1             | 6             | -          |                    |   |
| C s  |                     |               |               | •          |                    |   |
| isplay: All variables 👻 Cases S                            | canned: 109         |               |               |            |                    | Define Rules                                    |
|  |                     |               |               |            |                    |   |
| ∀ariable Distributions                                     |                     |               |               |            |                    |   |
| Limit number of cases scanned Case                         | s: 5000 R           | egcan Lim     | iting the num | ber of ca  | ses scanned doe    | s not affect how many cases are validated       |

# Data Cleaning with IBM SPSS Statistics

Jarlath Quinn – Analytics Consultant

www.sv-europe.com



| ariables.  |              |                                      | a scan of the                                       | data. Ti | ne Rules list show | rs all rules that can be applied to selected   |
|--|--------------|--------------------------------------|---|----------|--------------------|--|
| nalysis Variables:<br>/ariable   | Distribution | Minimum I                            | daximum F   | ul       | Rules:<br>Apply    | Name   |
| Predominanti religion (religion)<br>Average female life expectancy (life<br>Average need life expectancy (life<br>Regels who read (%) (liferasy)<br>Population increase (% per year)) [o<br>Innant mortality (deaths per 1000 live<br>Gross domestic product / capite (gdp |              | 43<br>41<br>18<br>-0.3<br>4.0<br>122 | ribal 0<br>82<br>76<br>100<br>5.2<br>168.0<br>23474 |          |                    | 0,1 dichotomy<br>1,2 dichotomy<br>1 to 5 Integer<br>1 to 10 Integer<br>1 to 10 Integer<br>Mornegaltes integer<br>0 to 100 number<br>Pilg system-missing values<br>Pilg usen-missing values<br>Pilg usen-missing values<br>Pilg usen-missing values<br>Pilg unitabeled values |
| Region or economic group [region]  |              | 1                                    | 6   | -        |                    |  |
| All variables Variable Distributions Umit number of cases scanned  | Cases: 5000  | egcan Limit                          |   | • of ca  | ses scanned doe    | Define Rules   |

# Data Cleaning with IBM SPSS Statistics

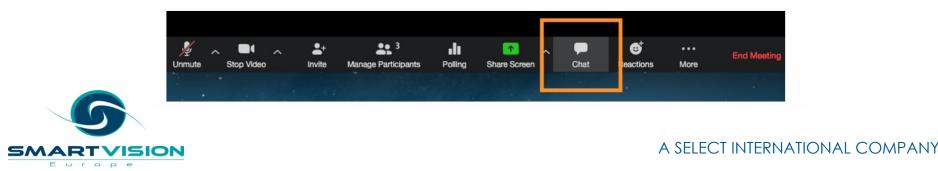
Jarlath Quinn – Analytics Consultant

#### Just waiting for all attendees to join...

www.sv-europe.com

## FAQ's

- Is this session being recorded? Yes
- Can I get a copy of the slides? Yes, we'll email links to download materials after the session has ended.
- Can we arrange a re-run for colleagues? Yes, just ask us.
- How can I ask questions? All lines are muted so please use the chat panel if we run out of time we will follow up with you.









- Gold accredited partner to IBM, Predictive Solutions and DataRobot specialising in advanced analytics & big data technologies
- Work with open-source technologies (R, Python, Spark etc.)
- Team each has 15 to 30 years of experience working in the advanced and predictive analytics industry

- Deep experience of applied advanced analytics applications across sectors
  - Retail
  - Gaming
  - Utilities
  - Insurance
  - Telecommunications
  - Media
  - FMCG

## Errors and problems in data

- Data cleaning is an almost universal problem for anyone who works with data
- Errors and irrelevancies in data can occur due to:
  - Data input mistakes such as misplaced keystrokes
  - Inconsistencies in recording information between different data entry operators or due to changes over time
  - Information collected on non-applicable events or subjects
  - Mismatches between database tables
  - Differences in how various systems encode or represent data such as date/time fields



## Challenges in data cleaning

#### • Typical tasks include:

- Identifying records/fields with a high percentage of missing values, a high degree of variability or conversely, too little variability
- Correcting values that are out of range: e.g. people aged 199 or years employed with minus numbers
- Identifying and removing duplicate records
- Ensuring a variables are correctly formatted e.g. removing decimal places from age
- Checking that the values in combinations of variables do not contradict each other or imply errors in the data: e.g. all car drivers should be at least 17 years old
- Creating syntax to correct data issues automatically



#### Two broad classes of data errors

- In SPSS, most issues with data fall into one of two categories:
  - Data formatting issues can be caused by how the data were stored in other systems
  - Problems with the data file itself can be caused by human error or systematic failures

Problems with how the data has been formatted/imported:

- Date/time variables
- Defining missing values
- Variable/value labels
- Variable types strings vs numeric
- Variable names

Problems with the data itself:

- Actual errors
- Irrelevant values/variables/records
- Inconsistencies
- Duplicates
- Illogical relationships





### Data formatting problems

www.sv-europe.com



#### Problems with the data itself

www.sv-europe.com



### Data cleaning with syntax

www.sv-europe.com

### **Additional Resources**

- <u>SPSS FAQs</u> everything from finding out what you have installed to how to merge files or change the language
- <u>Video Guides</u> a wide range of SPSS "how to" topics with mini demos
- <u>SPSS Software</u> information on products, modules and pricing
- <u>Eat your greens</u> blog series on statistical testing and procedures



Online training materials free to Smart Vision customers or available for purchase



£75.00

Factor and Cluster Analysis with IBM SPSS Statistics



Introduction to Time Series Forecasting with IBM SPSS Statistics

£75.00 Jarlath Quinn





Understanding and Applying Linear Regression Techniques in SPSS Statistics

Building predictive models in SPSS Modeler

£75.00 Jarlath Quinn

£75.00 Jarlath Quinn





Introduction to SPSS Modeler course



Understanding and applying logistic regression techniques in SPSS Statistics

£75.00 Jarlath Quinn



Statistical and significance testing in SPSS Statistics

£75.00 Jarlath Quinn



Introduction to IBM SPSS

Statistics course



#### Working with Smart Vision Europe

#### Consulting Services

#### **Project Support**

Purchase 1-2 days of consultancy time to have an expert work alongside you on your own project

#### **Analytics Advice**

Give us 3-5 days to investigate your data & analytical strategy and we'll present our recommendations re: improvements & alternatives

#### **Analytical Deep-Dive**

Let us explore your data landscape to test hypotheses, identify problem areas, find key outcome drivers or develop new applications



## Working with Smart Vision Europe Ltd.

- Sourcing Software
  - You can buy your analytical software from us often with discounts
  - Assist with selection, pilot, implementation & support of analytical tools
  - <u>http://www.sv-europe.com/buy-spss-online/</u>
- Training and Consulting Services
  - Guided consulting & training to develop in house skills
  - Delivery of classroom training courses / side by side training support
  - Identification & recruitment of analytical skills into your organisation
- Advice and Support
  - offer 'no strings attached' technical and business advice relating to analytical activities
    - Technical support services





Contact us:

+44 (0)207 786 3568 info@sv-europe.com

Twitter: @sveurope Follow us on Linked In



# Thank you

www.sv-europe.com