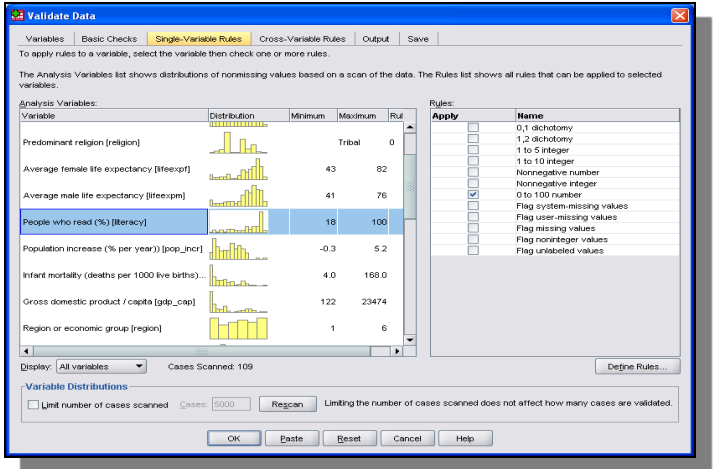
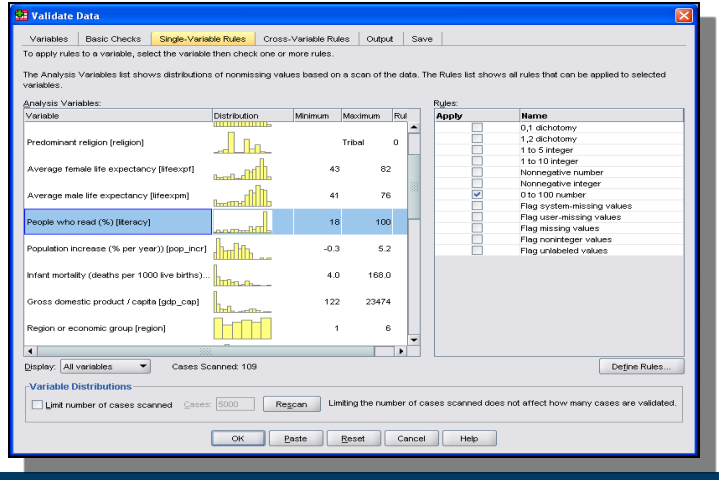
SMARTVISION Europe



Data Cleaning with IBM SPSS Statistics

Jarlath Quinn – Analytics Consultant

SMARTVISION Europe



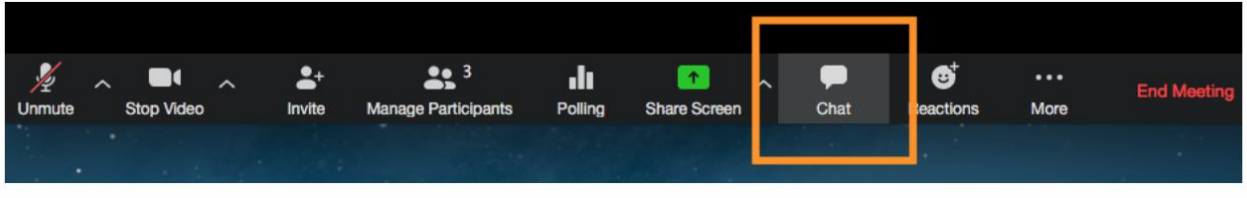
Data Cleaning with IBM SPSS Statistics

Jarlath Quinn – Analytics Consultant

Just waiting for all attendees to join…

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

SMARTVISION
Europe

Data formatting problems

SMARTVISION
Europe

Problems with the data itself

SMARTVISION Europe

Data cleaning with syntax

Additional Resources

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

Smart Vision Europe: Services and Expertise

Consultancy and Help

Guidance and support to help you get started. Embed our expertise alongside your team to help meet your objectives.

Training and Support

Educational support whether onsite or remote/virtual learning. Full day or half-day bite-size courses.

Data Science Recruitment

Help with staff recruitment utilising our technical expertise, extensive global network and decades of experience in the industry.

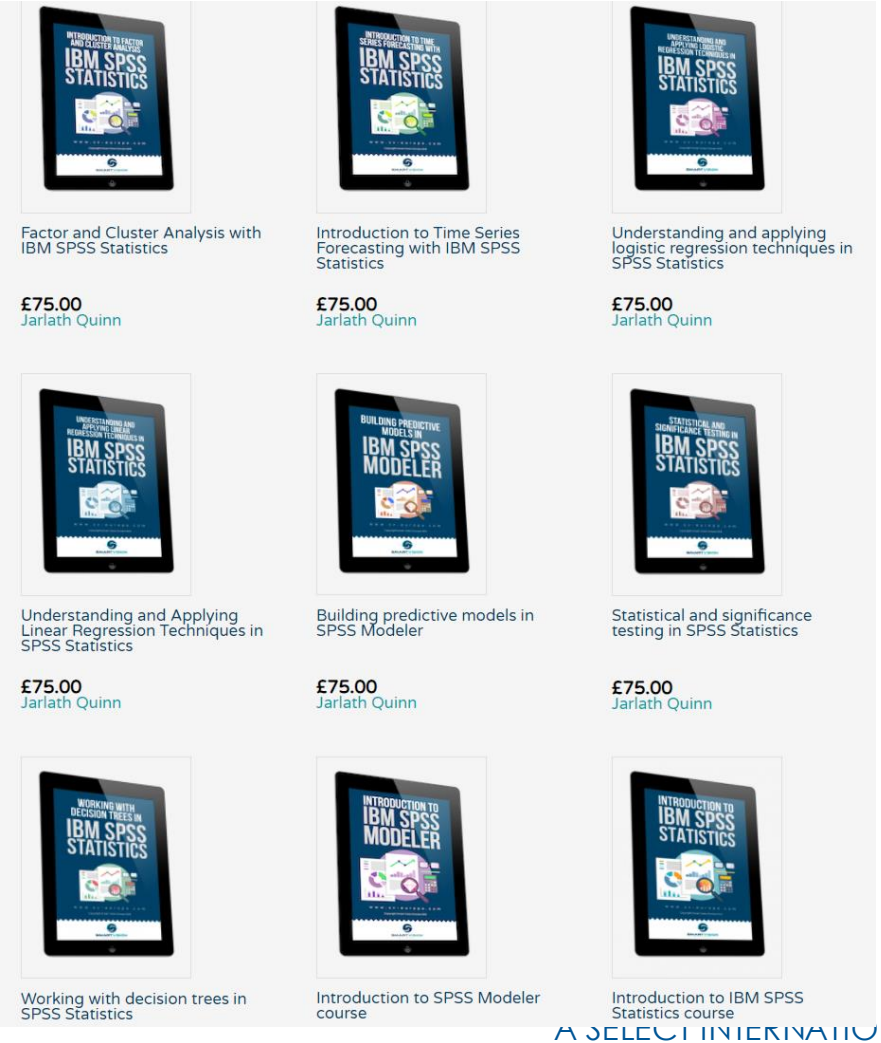
SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Smart Vision Europe: Services and Expertise

- **Software**
  - We can help you manage your existing SPSS licenses
  - You can buy your analytical software from us often with discounts
  - http://www.sv-europe.com/buy-spss-online/
- **Advice and Support**
  - We offer 'no strings attached' technical and business advice relating to analytical activities to anyone
  - Formal technical support services for the IBM SPSS product family
- **Access our online training catalogue**
  - Smart Vision Europe customers gain free access to our library of training materials and recorded self-paced training courses

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



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

SMARTVISION Europe

Contact us:

+44 (0)207 786 3568
info@sv-europe.com
Twitter: @sveurope
Follow us on Linked In

Thank you