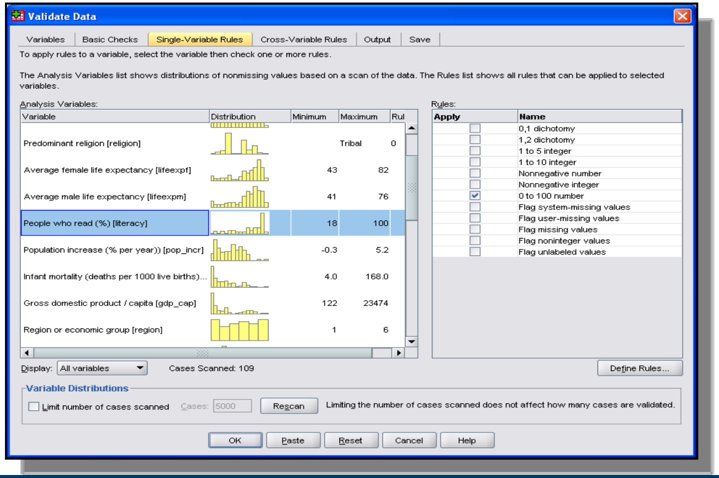


To apply rules to a variable, select the variable then check one or more rules.

The Analysis Variables list shows distributions of nonmissing values based on a scan of the data. The Rules list shows all rules that can be applied to selected variables.

Analysis Variables:

Variable	Distribution	Minimum	Maximum	Rul
Predominant religion [religion]			Tribal	0
Average female life expectancy [lifeexpf]		43	82	
Average male life expectancy [lifeexpm]		41	76	
People who read (%) [literacy]		18	100	
Population increase (% per year)) [pop_incr]		-0.3	5.2	
Infant mortality (deaths per 1000 live births)...		4.0	168.0	
Gross domestic product / capita [gdp_cap]		122	23474	
Region or economic group [region]		1	6	

Rules:

Apply	Name
□	0,1 dichotomy
□	1,2 dichotomy
□	1 to 5 integer
□	1 to 10 integer
□	Nonnegative number
□	Nonnegative integer
☑	0 to 100 number
□	Flag system-missing values
□	Flag user-missing values
□	Flag missing values
□	Flag noninteger values
□	Flag unlabeled values

Display: All variables

Cases Scanned: 109

Variable Distributions

□ Limit number of cases scanned Cases: 5000

Limiting the number of cases scanned does not affect how many cases are validated.

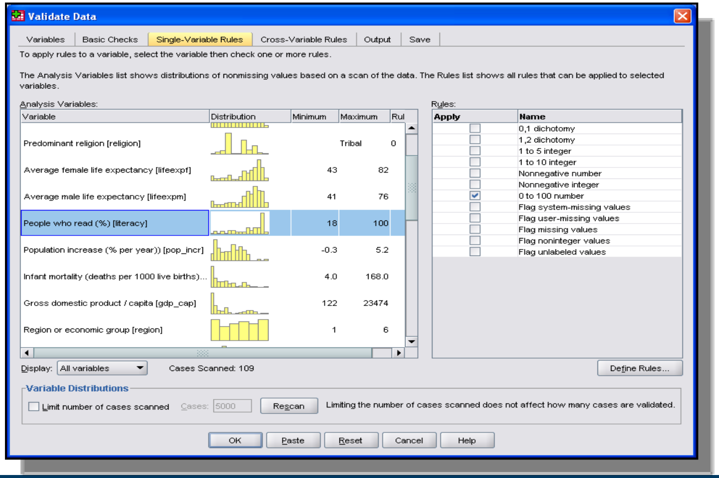
Data Cleaning with IBM SPSS Statistics

Jarlath Quinn – Analytics Consultant

www.sv-europe.com

A SELECT INTERNATIONAL COMPANY

SMARTVISION
Europe



To apply rules to a variable, select the variable then check one or more rules.

The Analysis Variables list shows distributions of nonmissing values based on a scan of the data. The Rules list shows all rules that can be applied to selected variables.

Analysis Variables:

Variable	Distribution	Minimum	Maximum	Rul
Predominant religion [religion]			Tribal	0
Average female life expectancy [lifeexpf]		43	82	
Average male life expectancy [lifeexpm]		41	76	
People who read (%) [literacy]		18	100	
Population increase (% per year)) [pop_incr]		-0.3	5.2	
Infant mortality (deaths per 1000 live births)...		4.0	168.0	
Gross domestic product / capita [gdp_cap]		122	23474	
Region or economic group [region]		1	6	

Rules:

Apply	Name
□	0,1 dichotomy
□	1,2 dichotomy
□	1 to 5 integer
□	1 to 10 integer
□	Nonnegative number
□	Nonnegative integer
☑	0 to 100 number
□	Flag system-missing values
□	Flag user-missing values
□	Flag missing values
□	Flag noninteger values
□	Flag unlabeled values

Display: All variables

Cases Scanned: 109

Variable Distributions

□ Limit number of cases scanned Cases: 5000

Limiting the number of cases scanned does not affect how many cases are validated.

Data Cleaning with IBM SPSS Statistics

Jarlath Quinn – Analytics Consultant

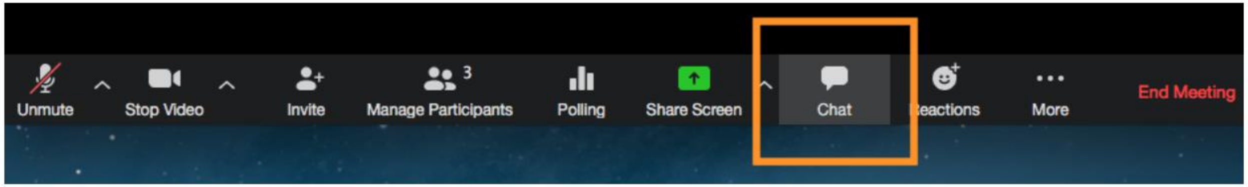
Just waiting for all attendees to join…

www.sv-europe.com

A SELECT INTERNATIONAL COMPANY

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.



SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

SMARTVISION Europe

IBM Gold Business Partner™

predictive SOLUTIONS

DataRobot

- 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

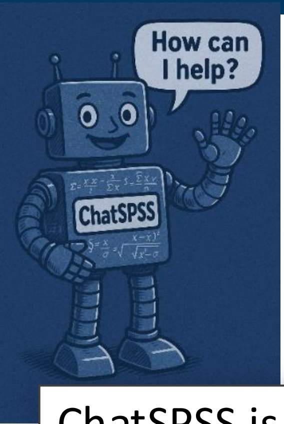
SMARTVISION
Europe

Data cleaning with syntax

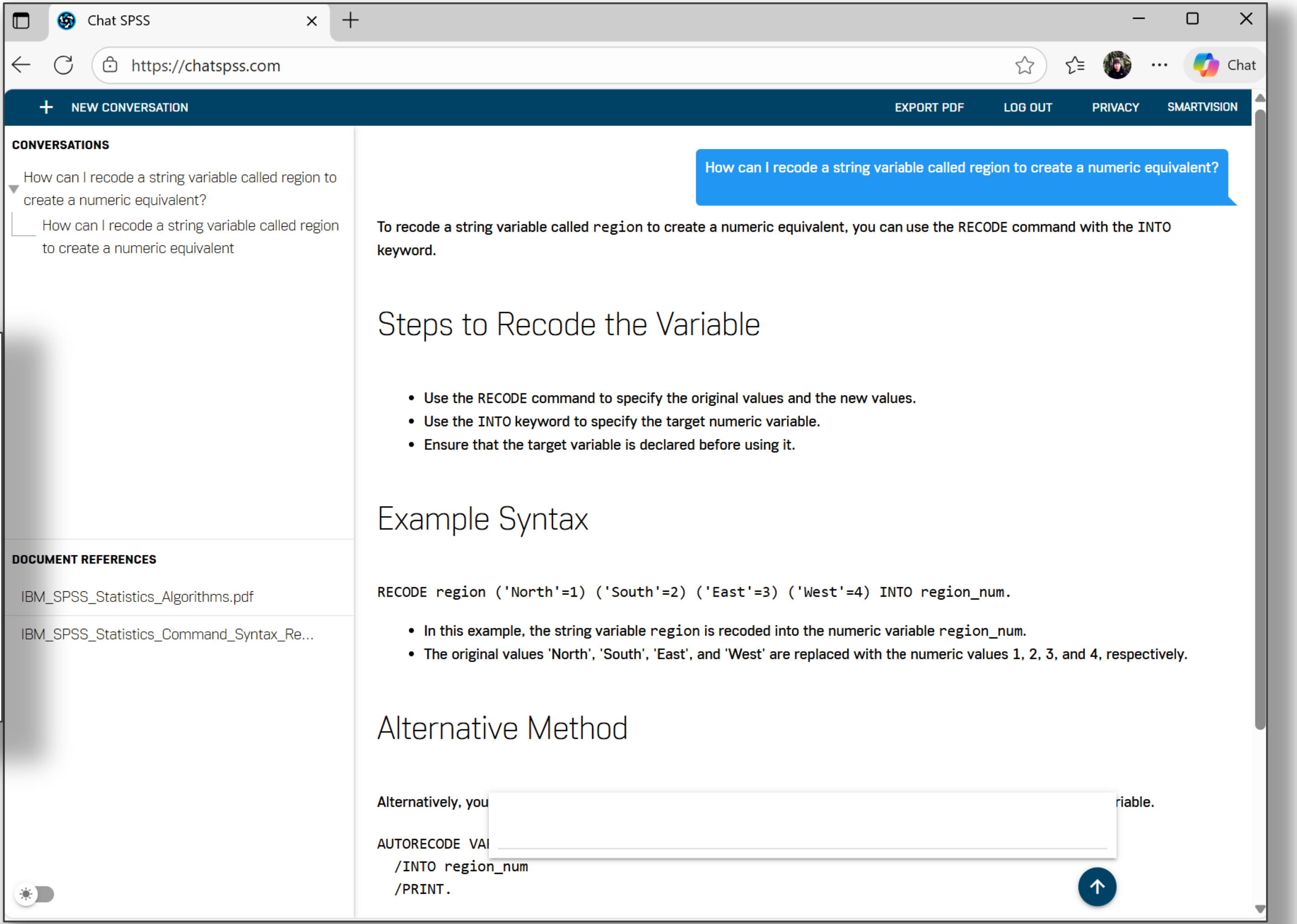
Additional Resources

- 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

Introducing **ChatSPSS:** https://chatspss.com/



ChatSPSS is a free generative AI chatbot developed by Smart Vision Europe Ltd based on official IBM documentation of SPSS procedures



CONVERSATIONS

How can I recode a string variable called region to create a numeric equivalent?

How can I recode a string variable called region to create a numeric equivalent

DOCUMENT REFERENCES

IBM_SPSS_Statistics_Algorithms.pdf

IBM_SPSS_Statistics_Command_Syntax_Re...

How can I recode a string variable called region to create a numeric equivalent?

To recode a string variable called `region` to create a numeric equivalent, you can use the `RECODE` command with the `INTO` keyword.

Steps to Recode the Variable

- Use the `RECODE` command to specify the original values and the new values.
- Use the `INTO` keyword to specify the target numeric variable.
- Ensure that the target variable is declared before using it.

Example Syntax

```
RECODE region ('North'=1) ('South'=2) ('East'=3) ('West'=4) INTO region_num.
```

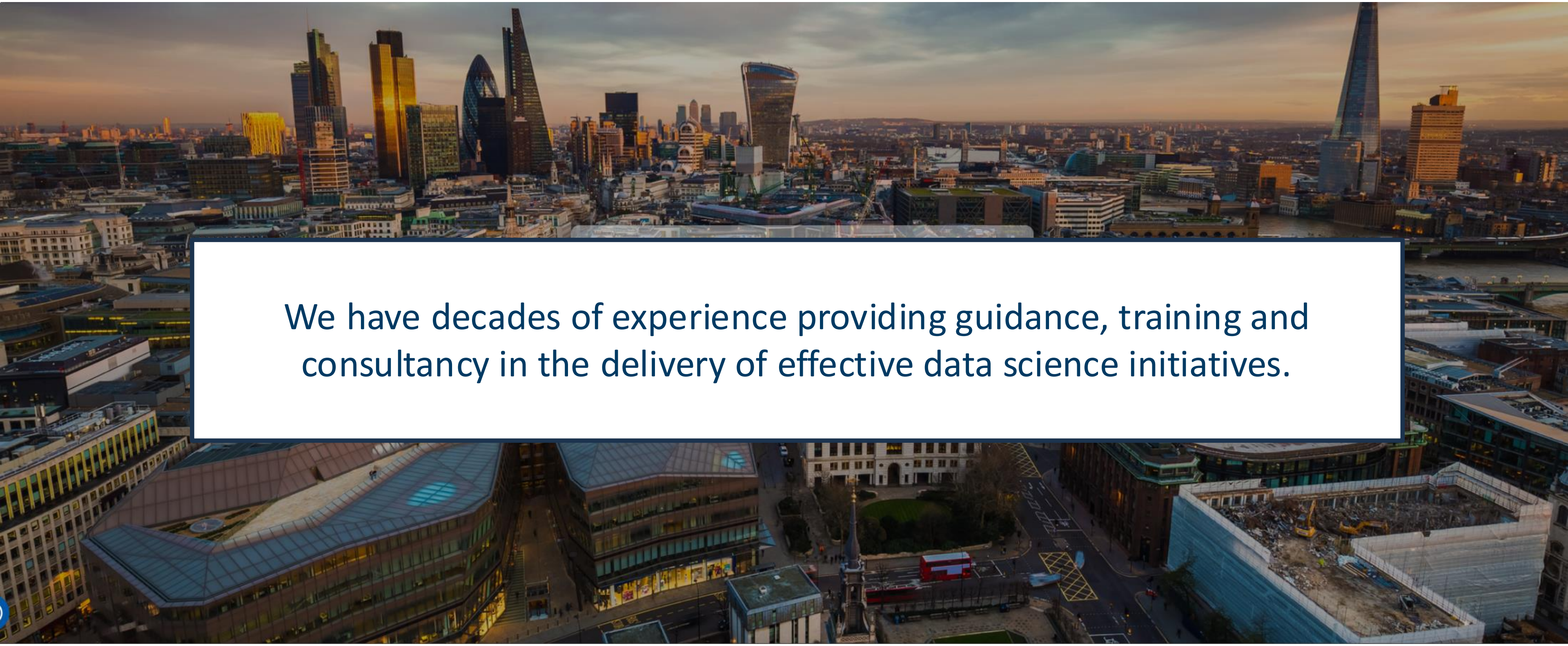
- In this example, the string variable `region` is recoded into the numeric variable `region_num`.
- The original values 'North', 'South', 'East', and 'West' are replaced with the numeric values 1, 2, 3, and 4, respectively.

Alternative Method

Alternatively, you riable.

```
AUTORECODE VA
  /INTO region_num
  /PRINT.
```

Smart Vision Europe: Services and Expertise



We have decades of experience providing guidance, training and consultancy in the delivery of effective data science initiatives.

SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Working with Smart Vision Europe

- **Sourcing Software**
  - You can buy your analytical software from us often with discounts
  - Assist with selection, pilot, implementation & support of analytical tools
  - http://www.sv-europe.com/buy-spss-online/
- **Training**
  - Formal classroom/virtual training
  - Online self-paced training resources
- **Advice and Support**
  - ‘No strings attached’ technical and business advice relating to analytics
  - Tracked technical support services around the IBM SPSS product line

SMARTVISION Europe

Contact us:

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

Thank you