

INCREASE THE DEFENSIVENESS OF YOUR CODE: REGULAR EXPRESSIONS

Valeriia Oreshko, Covance

Daryna Khololovych, Intego Group, LLC



Copyright © 2018 Covance. All Rights Reserved

Public



Content

- ▶ Regular expressions: quick introduction
- ▶ Examples of using regular expressions
 - Removing non-printable characters
 - Date validation
 - Deleting duplicates from a string
 - Dividing string into parts
- ▶ Conclusion



Regular Expressions: Quick Introduction



Regular Expressions: Introduction

- ▶ Regular expression (RE) - special string for describing a text pattern
- ▶ The string processing tasks in SAS® can be done using Perl regular expressions
- ▶ SAS has PRX family of functions to manipulate with strings using RE:
 - **PRXPARSE** creates RE id
 - **PRXMATCH** searches for a pattern match
 - **PRXCHANGE** performs a pattern-matching replacement



Examples of Using Regular Expressions



Removing Non-printable Characters

Goal: to get rid of non-printable characters while converting from one type of data or code to another

```
non_print=prxparse('s/[\a\b\e\f\n\r\t\v]+/ /');  
call prxchange(non_print, -1, var);
```

Metacharacter	Meaning
\a	Alert
\b	Backspace
\e	Escape
\f	page break
\n	new line
\r	carriage return
\t	horizontal tabulation
\v	vertical tabulation

Date Validation

Goal: to validate date reported in character format considering different rules (format, possible number of days in each month, leap years etc.)

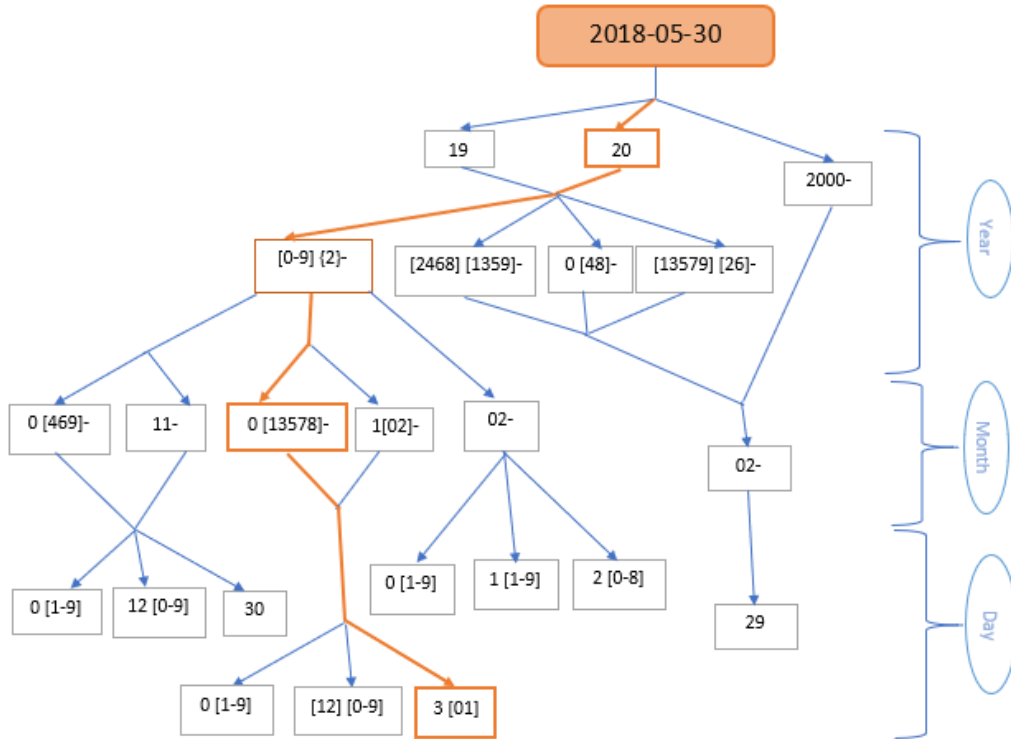
date	status
2018-70-01	NOT VALID
2018-12-32	NOT VALID

Simple example

```
data date_valid(drop = d_form);  
  set source;  
  d_form = prxparse  
    ("/(\\d{4}) - ((0[1-9])|(1[12])) - ((([012][09]$)|(3[01]$)))/o");  
  if prxmatch(d_form, date) eq 0 then status = "NOT VALID";  
run;
```



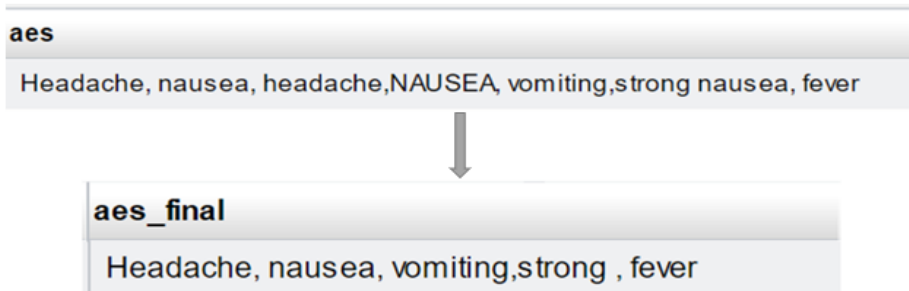
Date Validation



Scheme of regular expression for date validation

Deleting Duplicates from a String

Goal: to delete duplicate words:



RE used to delete duplicates:

```
d_ch="s/ (\\b\\w+\\b) (.*) (\\b\\1+\\b) /\\1\\2/io";  
d_det= " / (\\b\\w+\\b) (.*) (\\b\\1+\\b) /io";  
sp_com="s/ ( (\\, ) (\\s*) (?=\\, ) ) | (\\, \\s*$) //";
```

Deleting Duplicates from a String

```
do i=1 to &n;  
    aes=prxchange(d_ch, -1, compbl(aes));  
    if not prxmatch(d_det, compbl(aes))  
    then leave;  
end;
```

Result after loop:

aes

Headache, nausea, ,, vomiting, strong , fever

```
aes_final = prxchange(sp_com, -1, aes);
```

Dividing String into Parts

Goal: to divide string into several parts using “ATC” as a delimiter

var0	num1	num2	num3	num4	num5
ATC L ATC L04 ATC L04A ATC L04AC ATC L04AC01	L	L04	L04A	L04AC	L04AC01
ATC C10AA05	C10AA05				
J05AE05					

Used RE:

`/(?<=atc\|)\w*/io` - indicates word after “atc|”



Dividing String into Parts

```
data example3_prxnext;  
  set base;  
  atc_ptrn=prxparse("/(?<=atc\\|)\\w*/io");  
  start=1;  
  call prxnext(atc_ptrn,st,length(var0),var0,pos,ln);  
  array num $200 num1-num%eval(&n);  
  do i = 1 to &n while (pos > 0);  
    num[i]= substr(var0, pos, ln);  
    call prxnext(atc_ptrn,st,length(var0),var0,pos,ln);  
  end;  
run;
```

Conclusion and Q&A



Conclusion and Q&A

- ▶ The advantages of regular expressions:
 - Code independency on the new loads of raw data
 - Conciseness of code
 - Ability to identify and handle data errors caused by human error while gathering information
- ▶ Questions?





Covance Inc., headquartered in Princeton, NJ, USA, is the drug development business of Laboratory Corporation of America Holdings (LabCorp). COVANCE is a registered trademark and the marketing name for Covance Inc. and its subsidiaries around the world.