Automating

Descriptive Statistics

Cheng.Cong@cra-arc.gc.ca

Presented at Ottawa Area SAS User Society (OASUS)

November 23, 2017

Disclaimer

The vis expressed in this presentation are the personal views of the presenting staff and do not necessarily represent the views of Canada Revenue Agency or the Government of Canada.

The presentation is provided for general information purposes intended only as an academic resource and does not constitute professional advice.

Information has been summarized and paraphrased for presentation purposes and the examples are theoretical and have been provided for illustration purposes only.

Overview

1. Multidimensional Statistics and Proc Tabulate
2. Flowchart of Descriptive Statistics (DS) Work
3. Descriptive Statistics Automation Program (DSAP) and Example
4. Summary and Next Steps

1. Multidimensional Statistics and *Proc Tabulate*

1.1 Multidimensional Statistics

- A measure's statistics of interest is usually requested from a multidimensional grouping perspective
- Examples:
  - Median Height and Weight grouped by gender and age range
  - Average Income grouped by province and industrial sector
  - Median and Average Income grouped by gender, major income source and income range
  - Total Tax Payable grouped by Province, Gender and Total Income Range

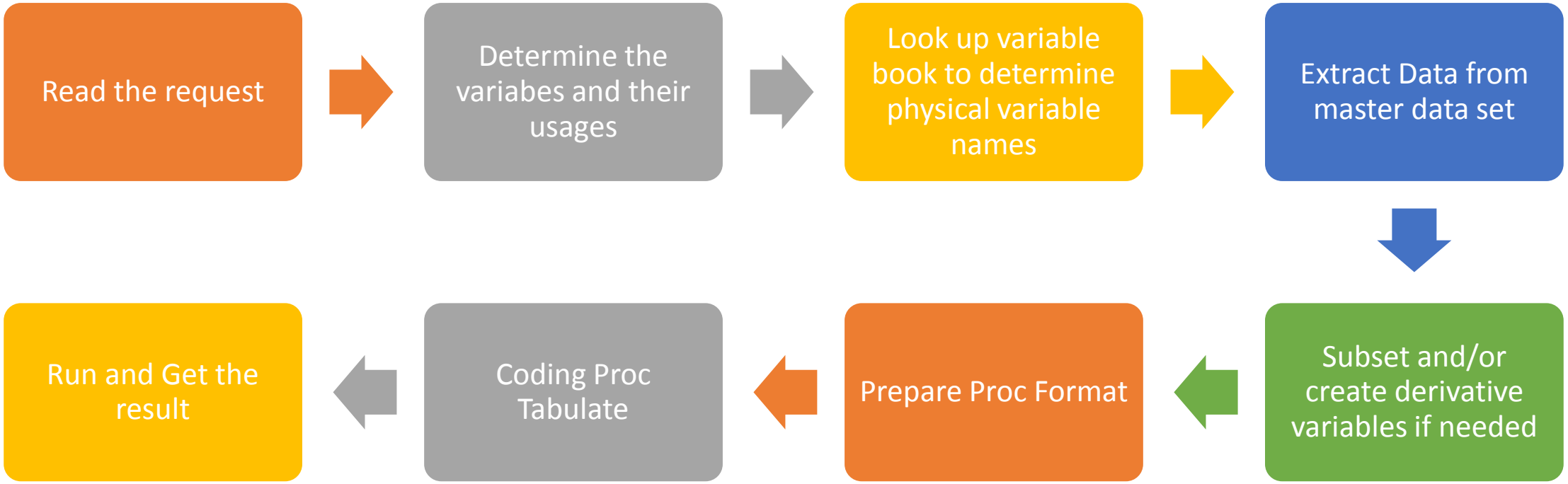
1.2 *Proc Tabulate* – the *omnipotent* solution for multidimensional statistics

- SAS Code for Median Height and Weight grouped by gender and age range

```
Proc tabulate data=lib1.data1 (where=(Year=2016));
Class gender age_range;
Variable height weight;
Table gender all, (age_range all)*Median*(Height Weight);
Format gender sexfmt. age_range agefmt.;
Run;
```

2. Flowchart of a Typical Descriptive Statistics Work

Descriptive Statistics Work Flowchart



Master Data Set - a multiyear demographic and economic micro test data

: Pilot.Mu10yrs]

Tools Data Solutions Window Help

	Year	client_id	birth_dt	sex_cd	race	marital_status	num_of_children	height	weight	eye_color	hair_color	fav_food	fav_sport	is_a_driver	driver_license_expiry
14841	2008	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14842	2009	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14843	2010	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14844	2011	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14845	2012	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14846	2013	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14847	2014	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14848	2015	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14849	2016	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14850	2017	111112595	1969-09-09	Male	Irish	married	5	182	36	Grey	Grey	McDonalds	Tennis	1	2024-05-02
14851	2008	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14852	2009	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14853	2010	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14854	2011	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14855	2012	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14856	2013	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14857	2014	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14858	2015	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14859	2016	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14860	2017	111112596	2000-02-22	Female	Irish	single	2	169	39	Blue	Red	Chinese	Football	0	.
14861	2008	111112597	1982-10-14	Female	Scotish	separated	4	155	67	Blonde	Red	McDonalds	Basketball	0	.

Work Data - Extracted from Master Data Set

	Year	client_id	sex_cd	height	weight	fav_food	prov_of_res	age
8507	2008	111120652	Female	172	23	Chinese	ON	55
8508	2012	111120652	Female	172	23	Chinese	ON	59
8509	2016	111120652	Female	172	23	Chinese	ON	63
8510	2008	111120660	Male	151	47	Chinese	ON	50
8511	2012	111120660	Male	151	47	Chinese	ON	54
8512	2016	111120660	Male	151	47	Chinese	ON	58
8513	2008	111120661	Male	162	44	Chinese	ON	71
8514	2012	111120661	Male	162	44	Chinese	ON	75
8515	2016	111120661	Male	162	44	Chinese	ON	79
8516	2008	111120669	Male	162	47	Pepsi Chips	ON	70
8517	2012	111120669	Male	162	47	Pepsi Chips	ON	74
8518	2016	111120669	Male	162	47	Pepsi Chips	ON	78
8519	2008	111120670	Female	171	43	Pepperoni Pizza	ON	25
8520	2012	111120670	Female	171	43	Pepperoni Pizza	ON	29
8521	2016	111120670	Female	171	43	Pepperoni Pizza	ON	33
8522	2008	111120672	Female	155	72	Chinese	ON	30
8523	2012	111120672	Female	155	72	Chinese	ON	34
8524	2016	111120672	Female	155	72	Chinese	ON	38
8525	2008	111120673	Male	159	55	KFC	ON	75
8526	2012	111120673	Male	159	55	KFC	ON	79
8527	2016	111120673	Male	159	55	KFC	ON	83
8528	2008	111120676	Female	173	52	McDonalds	ON	63
8529	2012	111120676	Female	173	52	McDonalds	ON	67
8530	2016	111120676	Female	173	52	McDonalds	ON	71

SAS Output - Multidimensional Statistics Table

Statistics (2008, 2012, 2016) on height, weight by 2 Features(sex,favorite food)
[prov_of_res eq 1, age ge 18]

		fav_food															
		Pepsi Chips				Pepperoni Pizza				McDonalds				KFC				
		height		weight		height		weight		height		weight		height		weight		he
		N	Median	N	Median	N	Median	N	Median	N	Median	N	Median	N	Median	N	Median	N
Year	sex_cd																	
2008	Female	194	165.00	194	54.00	338	165.00	338	55.00	310	165.00	310	54.00	348	166.00	348	55.00	343
	Male	185	165.00	185	53.00	344	165.00	344	54.00	384	165.00	384	55.00	339	165.00	339	56.00	343
	All	379	165.00	379	53.00	682	165.00	682	54.00	694	165.00	694	55.00	687	165.00	687	55.00	686
2012	sex_cd																	
	Female	201	165.00	201	54.00	357	165.00	357	55.00	328	165.00	328	54.00	370	166.00	370	55.00	361
	Male	198	165.00	198	53.00	351	165.00	351	54.00	407	165.00	407	55.00	360	165.00	360	55.00	357
	All	399	165.00	399	53.00	708	165.00	708	54.00	735	165.00	735	55.00	730	165.00	730	55.00	718
2016	sex_cd																	
	Female	212	165.00	212	53.50	379	165.00	379	55.00	346	165.00	346	54.00	394	165.00	394	55.00	376
	Male	204	165.00	204	53.00	366	165.00	366	54.00	429	165.00	429	55.00	384	165.00	384	55.00	381
	All	416	165.00	416	53.00	745	165.00	745	54.00	775	165.00	775	55.00	778	165.00	778	55.00	757
All	sex_cd																	
	Female	607	165.00	607	54.00	1074	165.00	1074	55.00	984	165.00	984	54.00	1112	165.50	1112	55.00	1080

3. Descriptive Statistics Automation Program (DSAP) and Example Illustration

Descriptive Statistics Automation Program (DSAP)

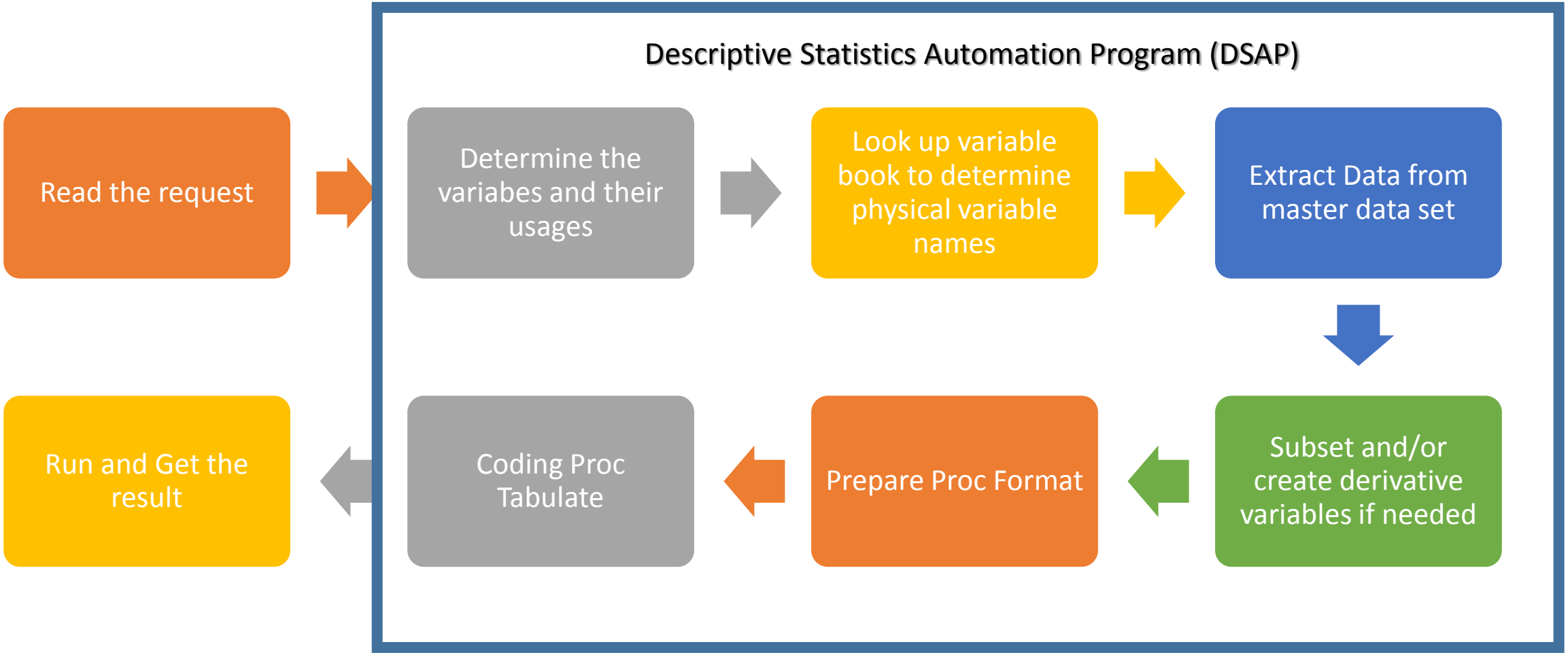
- **Step 1:**

→ Run DSAP macros in SAS (/store source)

- **Step 2:**

→ *Input plain language description of the statistics wanted* to get the result

DSAP Flowchart



Example – *Statistics on Height and Weight for Ontarian Adults*

%let HeyKitty=

Please produce statistics for residents whose province is Ontario and age no younger than 18 on count and median of height and weight by sex and favorite food for the year 2008, 2012 and 2016;

%***Hey_Kitty(***%bquote(&HeyKitty.));

How does DSAP work?

- Parsing the input request
- Looking up the knowledge bases
  - Book of Variables – to find physical variable names
  - Book of Code Values – to find the code values
  - Book of Operators – to determine the logical operators
  - Book of SAS Statistics – to find the SAS statistics function names
- Piecing SAS elements together by specialized dataset
- Call Symputx – assign the found SAS elements to corresponding macro variables
- Extracting work data and run proc tabulate

Parsing the request sentence 1

	PosName	PosVal	Sq
1	pos_scope1	26	1
2	pos_scope2	47	3
3	pos_stats1	93	4
4	pos_stats2	97	5
5	pos_analy1	113	6
6	pos_analy2	117	7
7	pos_years1	134	8
8	pos_years2	148	9
9	pos_class1	166	10
10	pos_class2	170	11
11	pos_end	191	12

VIEWTABLE: Work.Posdata

	PosName	posstart	posend
1	analy	117	134
2	class	170	191
3	scope	47	93
4	stats	97	113
5	years	148	166

Parsing the request sentence 2

	Cat	Seq	ElmntSaid
1	analy	1	height
2	analy	2	weight
3	class	1	sex
4	class	2	favorite food
5	scope	1	province is ontario
6	scope	2	age no younger than 18
7	stats	1	count
8	stats	2	median
9	years	1	2008
10	years	2	2012
11	years	3	2016

- Parsing is to decompose, identify and categorize the functional components from the sentence according to the semantic meaning
- DSAP identified the five functional groups of components indicated by the request:
    - Analytical variables
    - Class variables
    - Scope conditions
    - Statistics needed
    - Years needed

Knowledge base 1 - Look up Variable Book

	Cat	Seq	ElmntSaid
1	analy	1	height
2	analy	2	weight

	VARNAME	Description	TYPE	CLASS
17	fil_methd	fil_methd - filing method	1	1
18	first_nm	first_nm - first name	1	0
19	given_name	given_name - given name	2	0
20	hair_color	hair_color - hair color	1	1
21	height	height - height	1	0
22	is_a_driver	is_a_driver - is a driver	1	1
23	last_nm	last_nm - last name	1	0
24	major_inc_source	major_inc_source - major income source	1	1
25	marital_status	marital_status - marital status	1	1

	cat	seq	jecherche	VARNAME	Description	TYPE
1	analy	1	height	height	height - height	1
2	analy	2	weight	weight	weight - weight	1

Knowledge base 2 - Look up Operator Book

.Req_scope]

ols Data Solutions Window Help

	Cat	Seq	Elmnt Said	ce_qui_a_un_op
1	scope	1	province is ontario	province is ontario
2	scope	2	age no younger than 18	age no younger than 18

.Book_of_operators]

ols Data Solutions Window Help

	description	operator	description_length
403	were closer than	lt	17
404	were nearer than	lt	17
405	no smaller than	ge	16
406	no younger than	ge	16
407	no lighter than	ge	16
408	no shorter than	ge	16
409	not poorer than	ge	16
410	not weaker than	ge	16
411	not closer than	ge	16
412	not nearer than	ge	16
413	is no less than	ge	16
414	be no less than	ge	16

	description	operator	description_length	ce_qui_a_un_op	op_position	seq
1	no younger than	ge	16	age no younger than 18	4	2

Knowledge base 3 – Code Value Book

	Cat	Seq	ElmntSaid	ce_qui_a_un_op
1	scope	1	province is ontario	province is ontario
2	scope	2	age no younger than 18	age no younger than 18

VIEWTABLE: Pilot.Book_of_codevalues]

Edit View Tools Data Solutions Window Help

	vamame	vardescription	format	type	class	codevalue	description
78	phone_type	phone type	phone_type_simufmt	1	1	4	phone_type-4-others
79	prov_of_res	province of residence	prov_simufmt	1	1	1	prov_of_res-1-ontario
80	prov_of_res	province of residence	prov_simufmt	1	1	2	prov_of_res-2-quebec
81	prov_of_res	province of residence	prov_simufmt	1	1	3	prov_of_res-3-british columbia
82	prov_of_res	province of residence	prov_simufmt	1	1	4	prov_of_res-4-alberta
83	prov_of_res	province of residence	prov_simufmt	1	1	5	prov_of_res-5-monetoba
84	prov_of_res	province of residence	prov_simufmt	1	1	6	prov_of_res-6-saskatchewan
85	prov_of_res	province of residence	prov_simufmt	1	1	7	prov_of_res-7-nova scotia
86	prov_of_res	province of residence	prov_simufmt	1	1	8	prov_of_res-8-new brunswick
87	prov_of_res	province of residence	prov_simufmt	1	1	9	prov_of_res-9-newfoundland and labrador
88	prov_of_res	province of residence	prov_simufmt	1	1	10	prov_of_res-10-prince edward island
89	prov_of_res	province of residence	prov_simufmt	1	1	11	prov_of_res-11-north territory
90	prov_of_res	province of residence	prov_simufmt	1	1	12	prov_of_res-12-yukon
91	prov_of_res	province of residence	prov_simufmt	1	1	13	prov_of_res-13-nunavut
92	prov_of_res	province of residence	prov_simufmt	1	1	14	prov_of_res-14-others
93	race	race	race_simufmt	1	1	1	race-1-first nation

SAS - [VIEWTABLE: Work.Voici_scopecodevalue_1]

File Edit View Tools Data Solutions Window Help

	cat	seq	jecherche	vamame	vardescription	format	type	class	codevalue	description
1	ScopeCodeValue	1	ontario	prov_of_res	province of residence	prov_simufmt	1	1	1	prov_of_res-1-ontario

Knowledge base 4 - Look up SAS Stats Book

: Work.Req_stats]

Tools Data Solutions Window Help

	Cat	Seq	ElmntSaid
1	stats	1	count
2	stats	2	median

: Pilot.Book_of_stats]

v Tools Data Solutions Window Help

	Stat	Description
8	mean	mean - average
9	min	min - minimum
10	mode	mode - mode
11	n	n - count
12	nmiss	nmiss - number of missing values
13	range	range - range
14	skewness	skewness - skew
15	stddev	stddev - standard deviation

BLE: Work.Voici_stats]

View Tools Data Solutions Window Help

	cat	seq	jecherche	Stat	Description
1	stats	1	count	n	n - count
2	stats	2	median	median	median - median, p50

Piecing elements together by specialized datasets – e.g. all variable names

Work.Varname_needed]

Tools Data Solutions Window Help

	cat	seq	jecherche	VARNAME	Description
1	analy	1	height	height	height - height
2	analy	2	weight	weight	weight - weight
3	class	1	sex	sex_cd	sex_cd - sex code
4	class	2	favorite food	fav_food	fav_food - favorite food
5	scope	1	province	prov_of_res	is
6	scope	2	age	age	no younger than

Piecing elements together by specialized datasets – Observation Filters

Work.Scope_decompose]

Tools Data Solutions Window Help

	op_position	description_length	ce_que_je_cherche	jecherche
1	9	3	province	province
2	4	16	age	age

: Work.Voici_scope]

v Tools Data Solutions Window Help

	ce_qui_a_un_op	scope_head	operator	scope_tail	op_position	description_length
1	province is ontario	province	eq	ontario	9	3
2	age no younger than 18	age	ge	18	4	16

Key Coding Techniques and Traps

- Key Coding Techniques
  - ❑ Macros – top-down vs bottom-up approach
  - ❑ && and &&& Macro Variables
  - ❑ Call symputx

- Traps
  - ❑ Global and local macro variables
  - ❑ Initialization of parsing datasets
  - ❑Clean working datasets before rerunning

4. Summary and Next Steps

Summary and Next Steps

- DSAP is a preliminary Artificial Intelligence (AI) program specializing in producing custom descriptive statistics by SAS
- Knowledge bases play important role to relate the human language phrases to SAS language elements
- Next steps:
  - More language recognition e.g. French, Spanish and Chinese
  - Taking input mistakes such as misspelling and word missing
  - Self-learning in knowledge-base refreshment
  - Multimedia input channels e.g. microphone voice input, GUI, etc.

Thank You!

Cheng.Cong@cra-arc.gc.ca