Data Management and Visualization

**Assignment – Week 3**

**Making Data Management Decisions**

**By mapolarbear@gmail.com**

This week assignment more interesting, we have to make a decision about how to organize and which way to keep data (research variables), what to do with missing values, aggregate if needed and create additional, derived variables for grouping. I think in the future grouping will help to perform analysis, and create graphs.

1. **Missing values**

Based on data I have, I noticed that **suicides2002** and **suicides2004** variables are not populated for all countries, that there are about 30% of missing values, the situation with **fsuicides2002**, **msuicides2004** and corresponding **msuicides2002** and **msuicides2004** even worth, about 50% of values are missing. I think at current development, missing values cannot participate in my research study and will be potentially eliminated.

1. **Grouping data**

I think it’s interesting to group countries by region, to see stats by region, new variable **region ID** with corresponding format was created. In addition, after analysis of **suicides2002** and **suicides2004** variables, because of very large deviation of values I decided to create dedicated variables **gp2002** and **gp2004** for **suicides2002** and **suicides2004** variables respectively.

Program PROC FORMAT:

/\* Create formats for variables to translate \*/

/\* categorical variables values into something meaningful \*/

proc format;

/\* format for variables - suicides2002 suicides2004 \*/

value suicides\_range 0 = '0'

1-100 = '1-100'

101-500 = '101-500'

501-1000 = '501-1000'

1001-2000 = '1001-2000'

2001-5000 = '2001-5000'

5001-10000 = '5001-10000'

10001-high = '10001+'

;

/\* format for variables - gp2002 gp2004 \*/

value suicides\_groups 1 = '0'

2 = '1-100'

3 = '101-500'

4 = '501-1000'

5 = '1001-2000'

6 = '2001-5000'

7 = '5001-10000'

8 = '10001+'

99 = 'no data (missing value)'

;

/\*format for variable - regionID \*/

value region\_name 1 = 'ASIA (EX. NEAR EAST)'

2 = 'WESTERN EUROPE'

3 = 'BALTICS'

4 = 'EASTERN EUROPE '

5 = 'LATIN AMER. & CARIB'

6 = 'NEAR EAST'

7 = 'NORTHERN AFRICA'

8 = 'NORTHERN AMERICA'

9 = 'OCEANIA'

10 = 'SUB-SAHARAN AFRICA'

11 = 'WESTERN EUROPE'

99 = 'no data (missing value)'

;

run; /\*end of PROC FORMAT \*/

1. **Final data for analysis**

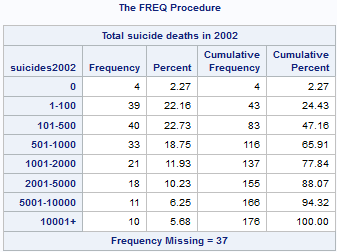
Created GAPMINDER\_RSCH\_EXT data, member of GAP\_DATA library. The GAPMINDER\_RSCH\_EXT contains only rows with data in all research variables, data must be GT than 0. The PROC SQL was used to create final dataset.

1. **Frequencies**

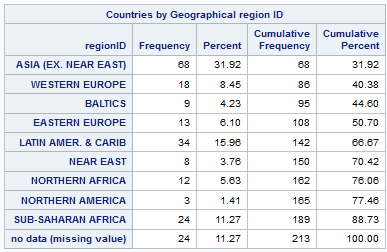
Executed PROC FREQ with corresponding formats, data looks pretty much clean, but not that big, 52 rows out of 213, but we will see if created dataset is useful. ☺

1. **Frequencies outputs:**

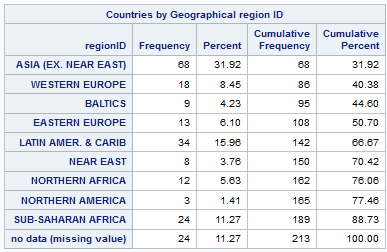
Sample PROC FREQ output for **suicides2002** with missing values and applied formats:



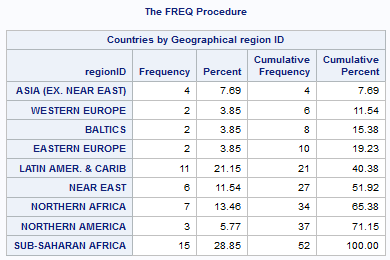
Sample PROC FREQ output for new **regionID** variable, with applied format, missing values substituted to 99 numeric value:

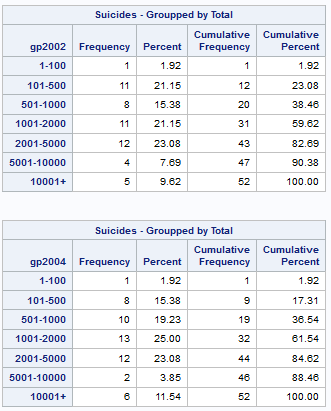


Following PROC FREQ outputs created based on cleaned and verified data, containing most of the data:



Following represents clean data:





1. **Program code (associated to Week-3 Assignment), described above:**

/\*\*\*\*\*\*\*\*\*\*\*\*\* START of week-3 assignment \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Create formats for variables to translate \*/

/\* categorical variables values into something meaningful \*/

proc format;

/\* format for variables - suicides2002 suicides2004 \*/

value suicides\_range 0 = '0'

1-100 = '1-100'

101-500 = '101-500'

501-1000 = '501-1000'

1001-2000 = '1001-2000'

2001-5000 = '2001-5000'

5001-10000 = '5001-10000'

10001-high = '10001+'

;

/\* format for variables - gp2002 gp2004 \*/

value suicides\_groups 1 = '0'

2 = '1-100'

3 = '101-500'

4 = '501-1000'

5 = '1001-2000'

6 = '2001-5000'

7 = '5001-10000'

8 = '10001+'

99 = 'no data (missing value)'

;

/\*format for variable - regionID \*/

value region\_name 1 = 'ASIA (EX. NEAR EAST)'

2 = 'WESTERN EUROPE'

3 = 'BALTICS'

4 = 'EASTERN EUROPE '

5 = 'LATIN AMER. & CARIB'

6 = 'NEAR EAST'

7 = 'NORTHERN AFRICA'

8 = 'NORTHERN AMERICA'

9 = 'OCEANIA'

10 = 'SUB-SAHARAN AFRICA'

11 = 'WESTERN EUROPE'

99 = 'no data (missing value)'

;

run;

/\* group and check data before subsetting \*/

proc freq data=gap\_data.gapminder\_rsch;

tables suicides2002 suicides2004;

format suicides2002 suicides\_range. /\* apply cereated formats \*/

suicides2004 suicides\_range.; /\* apply cereated formats \*/

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* read created dataset research gapminder\_rsch dataset \*/

/\* and group/create categorical variable which can be \*/

/\* used later for research \*/

data tmpds02;

set gap\_data.gapminder\_rsch;

/\* create gp2002 and assign value based on range \*/

/\* assign numeric 99 for missing values \*/

if suicides2002 = 0 then gp2002 = 1;

else if 1 <= suicides2002 <= 100 then gp2002 = 2;

else if 101 < suicides2002 <= 500 then gp2002 = 3;

else if 501 < suicides2002 <= 1000 then gp2002 = 4;

else if 1001 < suicides2002 <= 2000 then gp2002 = 5;

else if 2001 < suicides2002 <= 5000 then gp2002 = 6;

else if 5001 < suicides2002 <= 10000 then gp2002 = 7;

else if 10001 < suicides2002 then gp2002 = 8;

else gp2002=99;

/\* create gp2004 and assign value based on range \*/

/\* assign numeric 99 for missing values \*/

if suicides2004 = 0 then gp2004 = 1;

else if 1 <= suicides2004 <= 100 then gp2004 = 2;

else if 101 < suicides2004 <= 500 then gp2004 = 3;

else if 501 < suicides2004 <= 1000 then gp2004 = 4;

else if 1001 < suicides2004 <= 2000 then gp2004 = 5;

else if 2001 < suicides2004 <= 5000 then gp2004 = 6;

else if 5001 < suicides2004 <= 10000 then gp2004 = 7;

else if 10001 < suicides2004 then gp2004 = 8;

else gp2004=99;

/\* create regionID and assign value based on text value in region var. \*/

/\* assign numeric 99 for missing values \*/

if region = 'ASIA (EX. NEAR EAST) ' then regionID = 1;

else if region = 'BALTICS ' then regionID = 2;

else if region = 'C.W. OF IND. STATES ' then regionID = 3;

else if region = 'EASTERN EUROPE ' then regionID = 4;

else if region = 'LATIN AMER. & CARIB ' then regionID = 5;

else if region = 'NEAR EAST ' then regionID = 6;

else if region = 'NORTHERN AFRICA ' then regionID = 7;

else if region = 'NORTHERN AMERICA ' then regionID = 8;

else if region = 'OCEANIA ' then regionID = 9;

else if region = 'SUB-SAHARAN AFRICA ' then regionID = 10;

else if region = 'WESTERN EUROPE ' then regionID = 11;

else if region = ' ' then regionID=99;

/\* create lables for new variables \*/

label gp2002="Suicides - Groupped by Total";

label gp2004="Suicides - Groupped by Total";

label regionID="Countries by Geographical region ID";

run;

/\* check data after adding new groupping variables \*/

proc freq data=tmpds02; /\* work/temp. dataset we don't have to keep it in LIB \*/

tables gp2002 gp2004 regionID;

format gp2002 suicides\_groups. /\* apply cereated formats \*/

gp2004 suicides\_groups. /\* apply cereated formats \*/

regionID region\_name. /\* apply cereated formats \*/

;

run;

/\* subset data, extract data with meningfull numerical values \*/

proc sql noprint;

create table gap\_data.gapminder\_rsch\_ext as

select \*

from tmpDS02

where suicides2002 > 0

and suicides2004 > 0

and fsuicides2002 > 0

and fsuicides2004 > 0

and msuicides2002 > 0

and msuicides2004 > 0

and regionID between 1 and 11 /\*no missing values(99)\*/

;

quit;

proc print data=gap\_data.gapminder\_rsch\_ext; run; /\* control print \*/

/\* let's see freuences of regionID, no formatting \*/

proc freq data=gap\_data.gapminder\_rsch\_ext;

tables regionID;

run;

/\* apply formats, check data \*/

proc freq data=gap\_data.gapminder\_rsch\_ext;

tables regionID gp2002 gp2004;

format gp2002 suicides\_groups.

gp2004 suicides\_groups.

regionID region\_name.

;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\* end of week-3 assignment \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/