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**

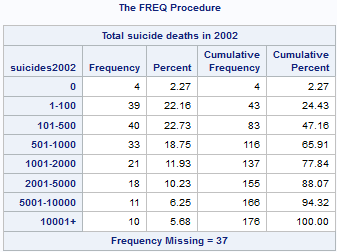
Created GAPMINDER\_RSCH\_EXT data, member of GAP\_DATA library. The GAPMINDER\_RSCH\_EXT contains only rows with data in all research variables, data must be GT than 0. The PROC SQL was used to create final dataset.

1. **Frequencies**

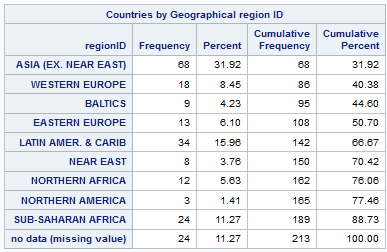
Executed PROC FREQ with corresponding formats, data looks pretty much clean, but not that big, 52 rows out of 213, but we will see if created dataset is useful. ☺

1. **Frequencies outputs:**

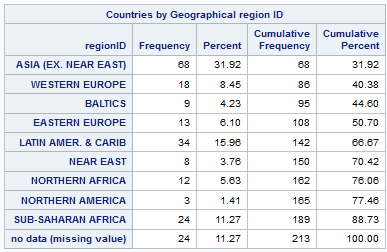
Sample PROC FREQ output for **suicides2002** with missing values and applied formats:



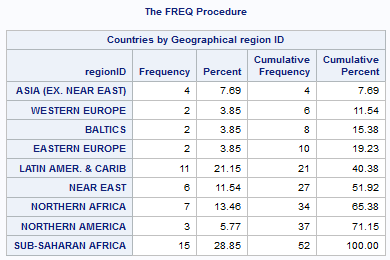
Sample PROC FREQ output for new **regionID** variable, with applied format, missing values substituted to 99 numeric value:

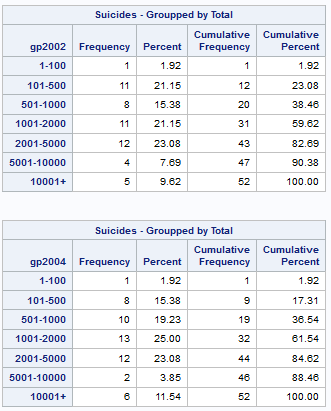


Following PROC FREQ outputs created based on cleaned and verified data, containing most of the data:

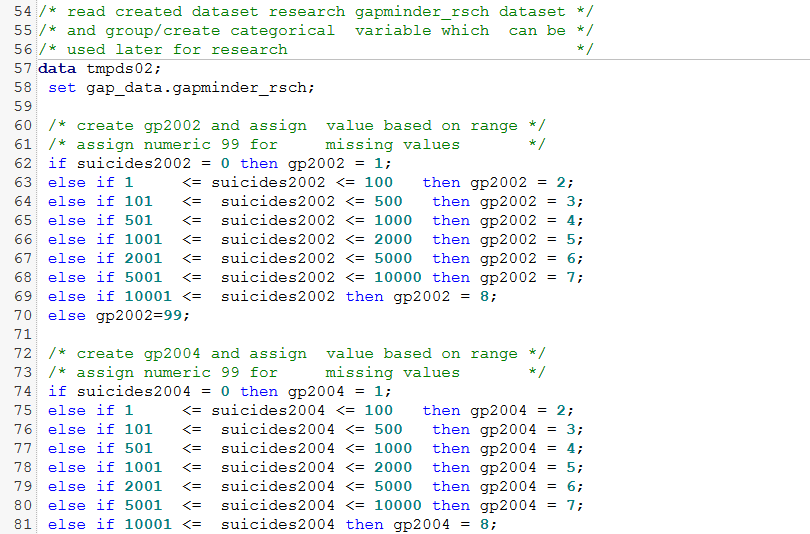


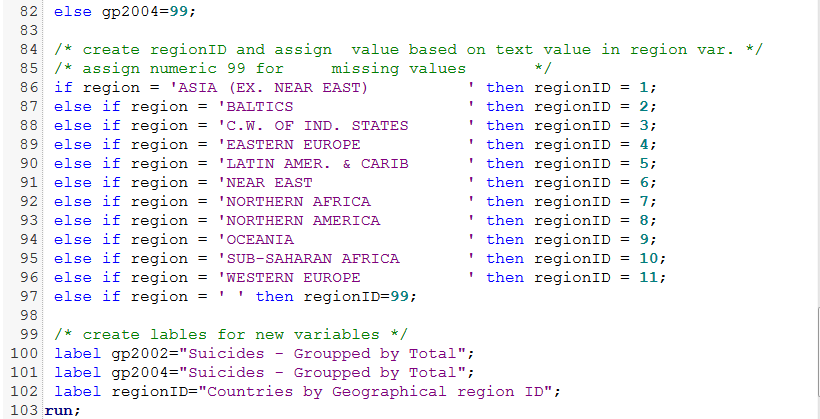
Following represents clean data:

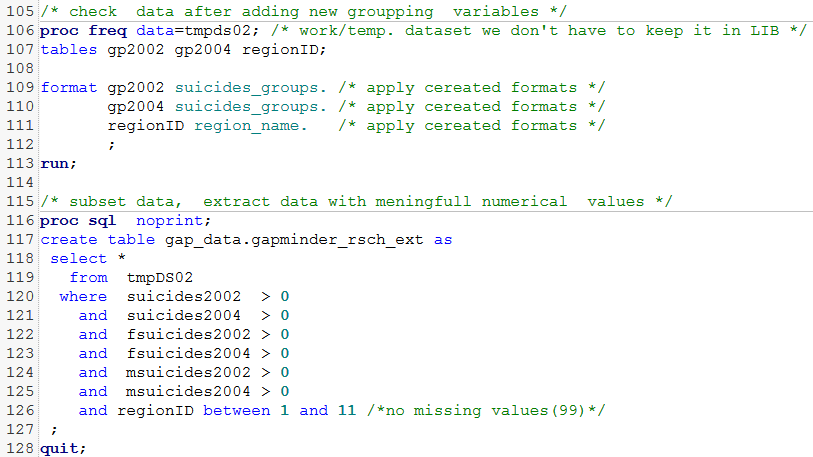


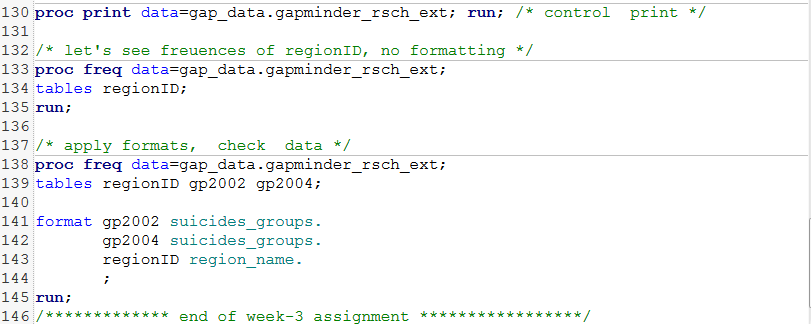


1. **Program code (associated to Week-3 Assignment), described above:**









Document link on [www.github.com](http://www.github.com):

<https://github.com/mapolarbear/Data-Analysis-Interpretation-Wesleyan-SAS/blob/master/Assignment%20-Week-3.docx>

<http://coursera-sas-dm-week3.tumblr.com/post/151117530030/data-management-and-visualization-assignment>