Data Management and Visualization

**Assignment – Week 4**

**Creating Graphs for Your Data**

**By mapolarbear@gmail.com**

This week assignment is even more interesting, we have to create univariate and bivariate graphs for variables of our study. We also have to interpret created graphs, and write few sentences describing results.

**Research question:** How female and suicide rates are related to various employment / unemployment indicators. And if they are related how (if) female suicide rate relationship to employment / unemployment different versus male suicide rates.

**Hypothesis:** There are a lot of factors affecting suicide behavior, and I think suicide rates and employment rates are related, maybe not fully, but at least partially.

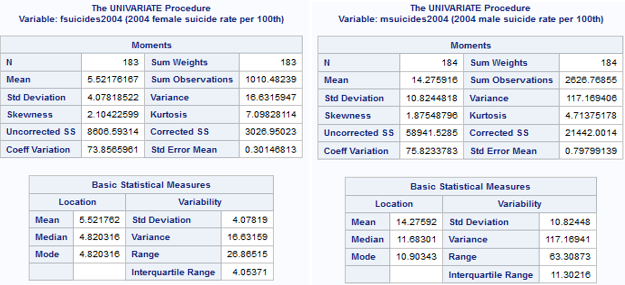
**Variables:**

**fsuicides2004, msuicides2004, employrate\_f2004, employrate\_m2004**

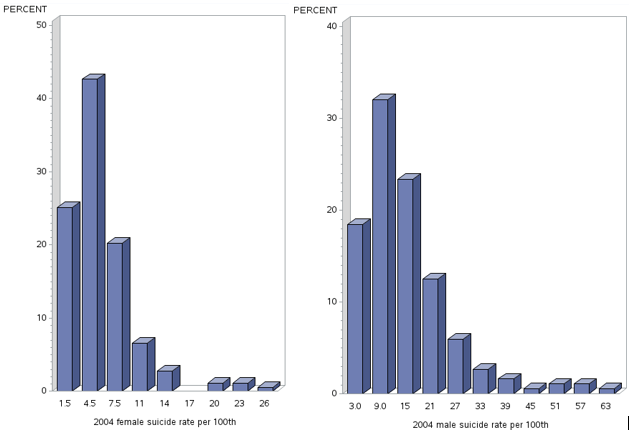
**Dataset:** GAPMINDER, extended with various indicators

**Univariate graphs and interpretations**

Based on data I have, I noticed that **fsuicides2004** and **msuicides2004** variables are not populated for all countries, that there are about 30% of missing values. But I still was able to produce univariate graphs using **GCHART** procedure. Before creating graphs, data of each variable was checked and analyzed with FREQ and **UNIVARIATE** procedures. Here is the sample of Univariate output of **fsuicides2004** and **msuicides2004,** where you can see that **MEAN** values are different, as well as **MEDIAN**, **STANDARD DEVIATION**, **MODE** and **RANGE** of values. The male data, **msuicides2004** has higher values, comparing to **fsuicides2004**, spread of Values (**RANGE**) is also higher in **msuiceds2004, having** more Variance in extreme values also, and will be more “juicy” that fsuicides2004 data. Are men more suicidal than women?

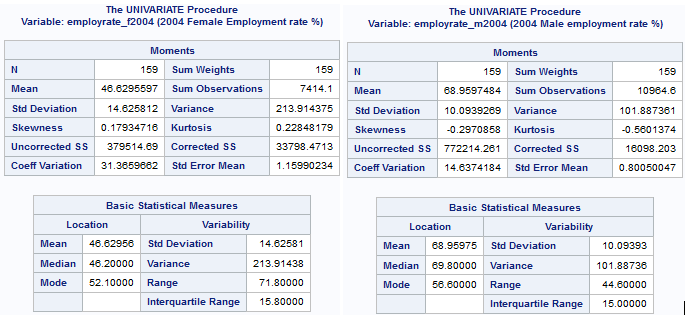


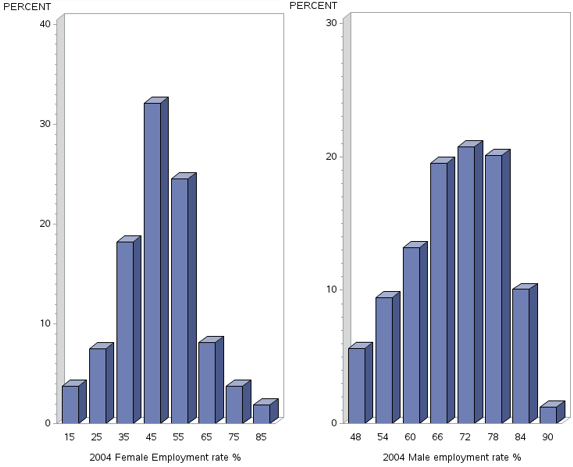
Corresponding bar charts created by GCHART procedure. Here you can see that both charts are **UNIMODAL** and have **right skewness**. As a prove of **UNIVARIATE** numbers, **female data** is peaked at 4.5 rating, at about 40%, having about 25% spike in 1.5 rate and about 20% in nearest 7.5 suicide rates. The highest suicidal rates 20 – 23 are about 1% or less and I think can be treated as extreme values. The male data has more spread and varies from 3.0 with 18% lowest to 63.0 with 1% or less, peaking at 9.0 suicidal rate at 32%.



Here are similar UNIVARIATE statistics for **employrate\_f2004** and **employrate\_m2004**, where we can see difference in major statistical parameters and can conclude that males (**employrate\_m2004**) are more employed that female (**employrate\_f2004**), having higher values in **MEAN,** **MEDIAN**, **STANDARD DEVIATION**, **MODE** and **RANGE.** Both charts are UNIMODAL, with picks at different employment rates, where male data (employrate\_m2004) has strong “highs” in 66 – 78 employment range having about 19% – 21% high for each. Female data (employrate\_f2004) looks different, peaking at 32% for 45 employment rate having closest peaks of 18% at 35 employment rate and 25% at 55 employment rate.

Based on analyzed data I can make a subjective conclusion that men are more employed then women.

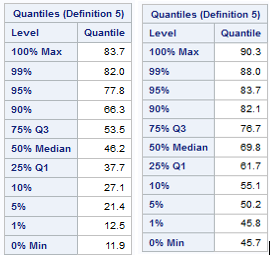




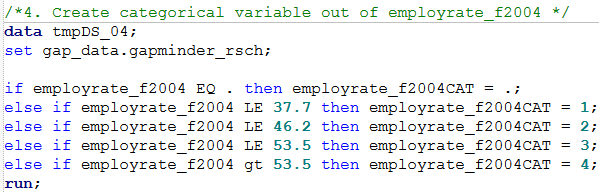
**2. Bivariate graphs and interpretations**

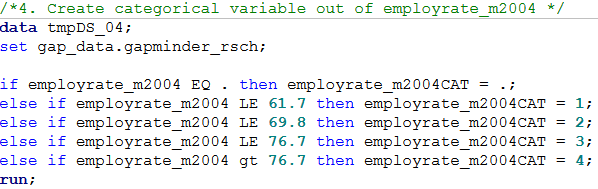
As long as both **variables are quantitate** variables, and to create more interpretable bivariate chart, I’ve created additional categorical variables employrate\_f2004CAT and employrate\_m2004CAT. Both variables where created based on UNIVARIATE Quartiles tables values or 25%, 50%, 75% and above 75% levels and following code added to program:

**Univariate Quantiles tables, employrate\_f2004 (left), employrate (2004) (right):**

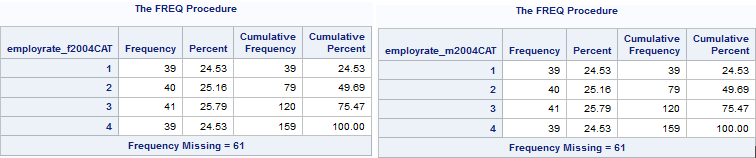


**Code in program:**

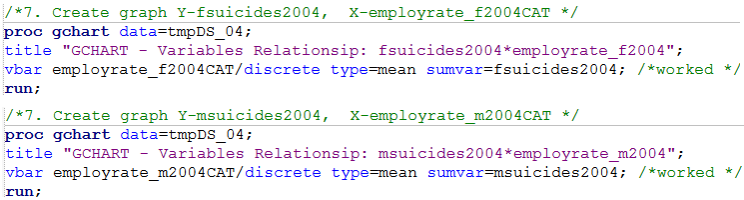


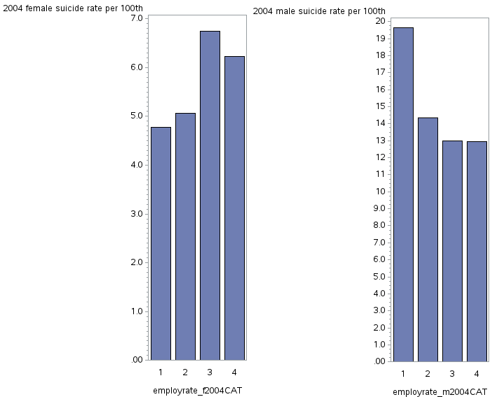


Output of corresponding **FREQ** procedures, data of both variables is evenly spread across corresponding groups respectively:

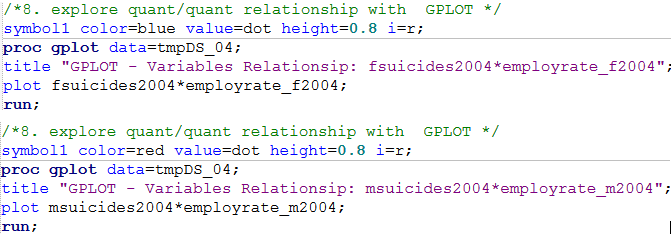


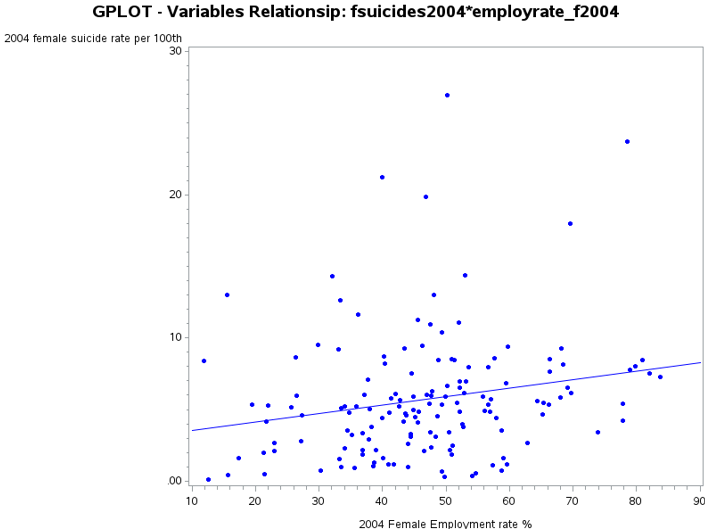
Interesting results created using **GCHART** procedure, where **X-axis** represented by **employrate\_f2004CAT** or **employrate\_m2004CAT** variables and **Y-axis** represented by **fsuicides2004** and **msuicides2004**. As we can see, there is positive and negative correlation between paired variables. In case or **msuicides2004 and employrate\_m2004CAT** correlation is negative (**less male employed, higher suicide rate**), but in female case, when **fsuicides2004 and employrate\_f2004CAT** correlation is positive (**more female employed, higher suicide rate**):

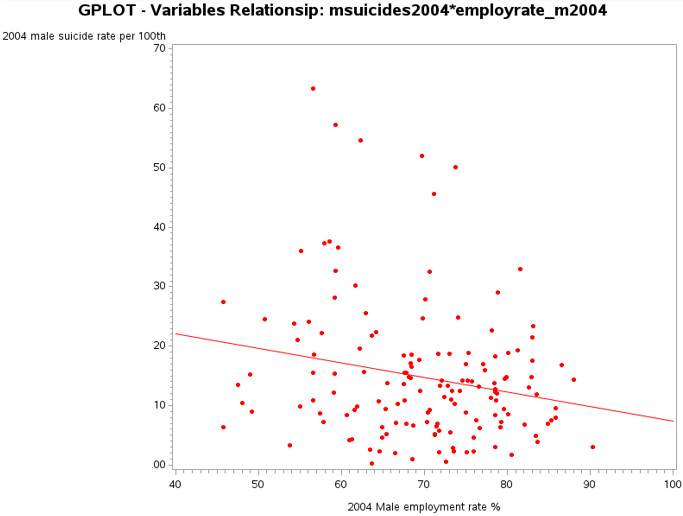




Multivariate graphs using UNIVARIATE procedure, where fsuicides2004 and msuicides2004 are analyzed against employrate\_f2004 and employrate\_m2004 respectively. Visually presented data also shows correlation between corresponding variables, but not that strong, inclinations/declinations are not steep and values are not very much concentrated around corresponding vectors. I think with more suicides data correlation between variables would be much more strong. Here is program code and corresponding graphs:







\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* End of week 4 assignment \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

http://coursera-sas-dm-week4.tumblr.com/post/151798525721/data-management-and-visualization-assignment