Paper 205

Data Exploration Made Easy

George C. J. Fernandez, Department of Applied Economics and Statistics /204,
University of Nevada - Reno, Reno NV 89557

ABSTRACT

A comprehensive graphical analysis approach to perform data exploration utilizing the latest capabilities available in SAS® systems are presented here. This practical approach integrates graphical analysis tools available in SAS systems and provides step-wise instructions to perform data exploration quickly without writing SAS program statements or using menu interface by running the SAS macros in background. The main feature of this approach is that analysts can perform complete data exploration quickly by following the steps and using the SAS macro files provided. Since data analysts search for user-friendly but efficient data exploration methods to produce quick results, this approach provides such a tool to achieve their objectives. Using this MACRO APPROACH, SAS users can effectively and quickly perform data exploration and spend more time in interpretation of graphs rather than debugging their program errors etc. Furthermore, by using this approach, data analysts can simplify the steps involved in data exploration and improve quality of statistical analysis.

INTRODUCTION

With the rapid development of improved and user-friendly statistical software and powerful desktop and notebook PCs', computer applications become an important component of data analysis. More and more institutions are totally dependent on statistical software to perform routine data analysis.

Initially, data analyses were usually performed by writing program codes. This approach is considered a more powerful and flexible method of data analysis. Training courses on data analysis were also based on focusing writing program codes and debugging program errors. Even though the program-code based approach is very powerful, this method of training is considered not user friendly and therefore not suitable for average users and a major portion of the learning is spent on debugging program errors. Therefore, training based on writing program codes is considered not suitable for average users.

With the introduction of PC/Windows environment in early 90s, computer users adopted the user-friendly point-and-click and menu-interface approach as the standard form of computer interface. Popular graphical software companies responded to the users needs and developed dynamic, interactive and user-friendly data analysis modules. The SAS Institute also introduced data analysis modules such as SAS/INSIGHT for exploratory data analysis, SAS/LAB for guided analysis, and recently SAS/ANALYST for statistical data analysis. These dynamic and interactive data analysis modules became more popular among the data analysts and were accepted as the industry standards.

However, some limitations were noted in the application of the user friendly dynamic modules. First of all, these modules only have limited functions i.e. Only the most common data analysis options are available in these modules. Also when using these modules, opening multiple windows may clutter the desktop and may confuse the average users. Further more, very limited options are usually available to log and save the steps involved in data analysis. Thus, automating the steps involved in repeated routine analysis become time consuming and inefficient. Finally, the danger of misuse of data analysis increased recently with the availability of the user friendly point-and-click modules. Even in the presence of these limitations, the use of data analysis modules are widely accepted by the users since no other alternative approach is readily available for the users.

As an alternative to the point-and-click menu interface modules for obtaining quick and complete results from data exploration, an approach based on SAS macros is presented here. This macro approach integrates the graphical analysis tools available in SAS systems and provides complete data analysis tasks **quickly** without writing SAS program statements or

using the point and click menu interface by running the SAS macros in the background. The main feature of this approach is users can perform graphical statistical analysis quickly by following the steps and using the SAS macro files provided. Since data analysts continuously search for user-friendly but efficient data analysis methods to produce **quick results**, this approach provides such a tool to achieve their objectives. Using this *MACRO APPROACH*, the analysts can effectively and quickly explore the data, interpretation of graphs and output rather than debugging their program errors etc. Furthermore, by using this approach, the analysts can simplify the steps involved in data analysis and improve the quality of data analysis and report generation.

In this Macro-approach, step-by-step instructions are presented to create SAS data sets, to carry out relevant data analysis, and to create publication quality graphics. All SAS program statements required to perform these analyses are presented as SAS macro(s) and the users are only required to input the appropriate variable names in the Macro-Call window. Each analysis is treated as independent, and complete step-by-step instructions are provided to perform the analysis. Using this approach simplifies the steps involved in data analysis and improves the quality of data analysis.

All SAS macro files used in this approach have *.MAC as file extensions. The macro-call files and the sample data files both have *.SAS as file extensions. The macro files, data files, and macro-call files are stored in "MACRO", "DATA" and "MAC-CALL" folders respectively. The users can easily modify the data files or the macro-call files when creating new SAS data sets, or when submitting the macros. By default in SAS for Windows, when the graph = "DISPLAY"option is used, all the graphs are displayed on the screen and the users can examine the graphs sequentially in the graphics window. By changing the device name to a SAS graphic devices such as "WORD" or "WP" all graphs and SAS OUTPUT can be automatically saved to the default (PLOT and LABWORK) folders respectively or to a user specified folder.

The macro based data analysis is carried out in four steps:

Step 1: Creating SAS data:

i) Identify the sample data used. Open the data file from the disk into the SAS program editor. Modify the SAS statements in the sample data file, enter your data appropriately, and submit your data file to create the SAS data. Or
ii) Enter your data in Microsoft Excel spread sheet and save as a MS excel 4 worksheet format.
(Data from Lotus worksheet and Dbase files also can be used as source)
Start SAS session, open the macro-call file "EXCELSAS.sas" in to the program editor, submit, to open the ExcelSAS window (Figure 1). In the EXCELSAS window, input the excel file name and the file location, and submit. This macro file convert the excel data to a temporary SAS data set . The excel file name will be used as the SAS data name.

Step 2: Running SAS Macro:

Open the appropriate macro-call file from the disk submit to open the macro-call window. In the macro-call window, input the appropriate variable names and submit to execute the SAS macro. Examine the log window for and syntax error and if you find any syntax errors, correct the errors in the macro-call window. Otherwise, examine the SAS graphs produced by the macro by scrolling the graphics window. Next, examine the SAS output by opening the SAS output window. If you are not familiar with the contents of the SAS output or if you need additional information, please refer the relevant SAS manuals or the SAS publications.

Step 3. Saving Graphs and the Output:

To save the SAS graphs and output, go to the macro-call window and change the graph field name to "WP" for WordPerfect or "WORD" for Microsoft Word. Also, if you want to save the graphs to any other location other than the default folder, "A:\plot", specify the correct drive and the folder name in the DIR field. Resubmit your macro-call window and the SAS graphs will be saved

directly to the specified folder. Also, the contents of the OUTPUT window will be saved to the A:\LABWORK folder when you change the graphic device from DISPLAY to any other devices.

Step 4. Project Report:

Start your favorite word processor, open the SAS out put file, edit/modify the contents, add your comments, insert the appropriate graphs in your document, and finish your report. To incorporate the graphics files into your document, follow the directions of your word processing software's.

Advantages of using this Macro-approach:

◘ Minimum Training : Because the same four steps are used in all data analysis, the amount of training required to learn SAS systems is relatively minimum. Also, all the analysis options, new features, and updates can be incorporated in side the macro by the developer. Thus, the end users can get the full benefits of SAS without spending additional time in learning the new features.

◘ Quick and Complete Results: All the required steps needed to perform any given data analysis are incorporated in a SAS macro. Thus, the users can obtain the complete results quickly in one run.

◘ Automating routine analysis: All repeated routine analysis can be performed quickly since the macro has the capability of handling multiple response variables in one run.

A list of SAS macros used in data exploration is given the appendix.

WORKED Example

Data set: Base ball salary data
Source: SAS data files
File type: MS excel 4

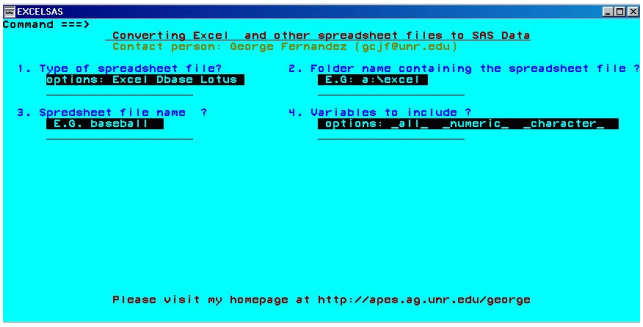
Data exploration performed:

1) Examine the baseball players salary by division using box-plot, frequency histograms, variation plots, and sorted data table.

2) Examine the data for outliers and deviation from normality by distribution and normal probability plots and test statistics for normality.

i) Creating SAS Data from a spreadsheet using the SAS macro ExcelSAS:

Open the macro-call file "Excelsas.sas" from the mac-call folder into the program editor, submit to open the Excelsas window. Input the appropriate



macro variables in the Excelsas window (Figure 1) and submit.

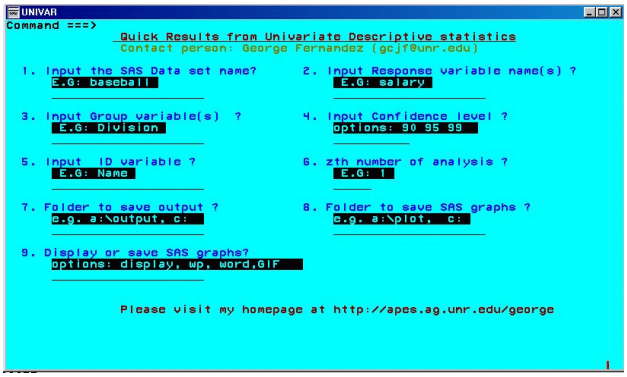


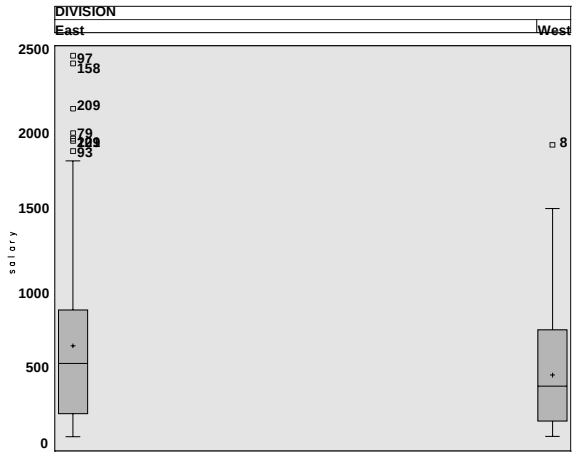
Figure 1. ExcelSAS macro-call window

Figure 2. Univar macro-call window

ii) **Performing exploratory data analysis on baseball players's salary by Division**.

Include the univar.sas mac-call file into the program editor and submit to open the UNIVAR window (Figure 2). Input the appropriate macro parameters in the UNIVAR window and submit to run the exploratory data analysis. The exploratory graphs generated by this macro are given in figures 3-8. This UNIVAR will also produce univariate statistics, trimmed means, confidence intervals, and normality test statistics (Not included in this paper).

Figure 3: Boxplot of Baseball player's salary by division



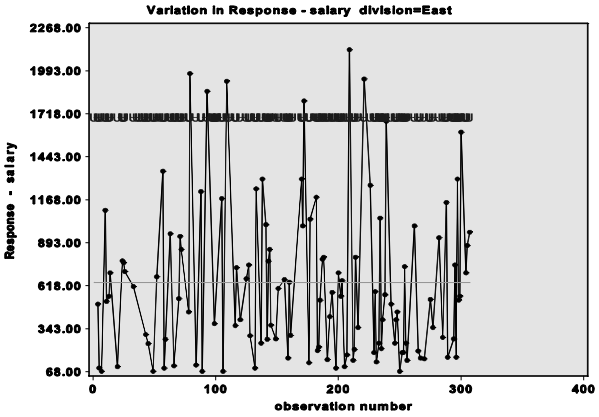


Figure 4.a. Variation in Baseball player's Salary (Division=East)

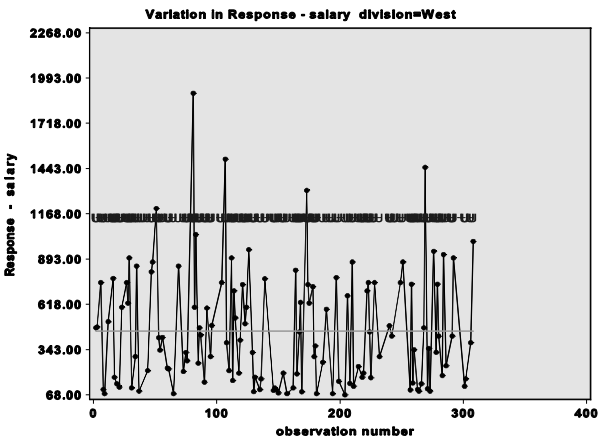


Figure 4.b. Variation in Baseball player's Salary (Division=West)

Exploratory graphs produced by the UNIVAR macro:

1) Boxplot of response variable (Fig. 3): This plot is useful in detecting outliers, examining the spread, and comparing the mean, median and quartiles.

2) Data variation plot (Fig 4). Variation and trend in the response variable is captured in this plot. The sample mean and the (mean ± 2*SD) lines are shown in this plot.

3) Histograms of response variables (Fig. 5) with percent and cumulative percent statistics

4) Data graphics sorted by the response variable (Fig.6)

6) Histogram and normal distribution plot (Fig.7)

7) Normal probability plot (Fig. 8)

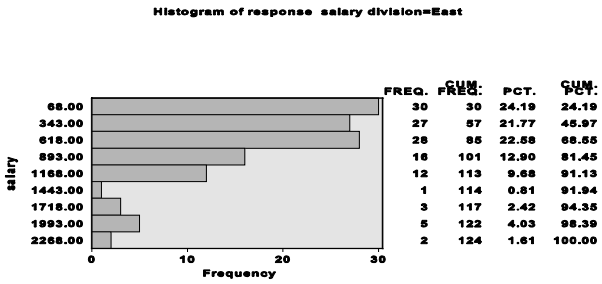


Figure 5a.Frequancy Histogram of player's Salary (Division=East)

Figure 5b.Frequancy Histogram of player's Salary

vision=West)

Figure 6a: Response sorted by salary (Division=East)

Sorted response salary division=East

OBS	NAME	SALARY	min 0.00	max 2460.00
97	Eddie Murray	2460.00		*
158	Jim Rice	2412.50		*
209	Mike Schmidt	2127.33	*	
79	Don Mattingly	1975.00	*	
221	Ozzie Smith	1940.00	*	
109	Gary Carter	1925.57	*	
93	Dave Winfield	1861.46	*	
172	Keith Hernandez	1800.00	*	
239	Rickey Henderson	1670.00	*	
300	Wade Boggs	1600.00	*	
57	Cal Ripken	1350.00	*	
138	Jack Clark	1300.00	*	
170	Kirk Gibson	1300.00	*	
297	Von Hayes	1300.00	*	
226	Paul Molitor	1260.00	*	
133	Jesse Barfield	1237.50	*	
88	Darryl Strawberry	1220.00	*	
182	Leon Durham	1183.33	*	
105	George Bell	1175.00	*	
288	Tony Pena	1150.00	*	
10	Andre Thornton	1100.00	*	
234	Ron Cey	1050.00	*	
177	Keith Moreland	1043.33	*	
141	Jody Davis	1008.33	*	
171	Ken Griffey	1000.00	*	
262	Robin Yount	1000.00	*	
307	Willie Upshaw	960.00	*	
63	Don Baylor	950.00	*	
71	Dwight Evans	933.33	*	
282	Tommy Herr	925.00	*	
305	Willie Randolph	875.00	*	
72	Damaso Garcia	850.00	*	
144	Jim Gantner			

(Complete Data not shown)

Figure 6b: Response sorted by salary (Division=West)

Sorted response salary division=West

OBS	NAME	SALARY	min 0.00	max 1900.00
81	Dale Murphy	1900.00		*
107	George Brett	1500.00	*	
269	Steve Garvey	1450.00	*	
173	Kent Hrbek	1310.00	*	
51	Carney Lansford	1200.00	*	
83	Dave Parker	1041.67	*	
308	Willie Wilson	1000.00	*	
126	Harold Baines	950.00	*	
276	Tom Brunansky	940.00	*	
284	Terry Kennedy	920.00	*	
29	Brian Downing	900.00	*	
112	Gary Gaetti	900.00	*	
292	Terry Puhl	900.00	*	
48	Carlton Fisk	875.00	*	
210	Mike Scioscia	875.00	*	
251	Rafael Ramirez	875.00	*	
35	Bill Madlock	850.00	*	
69	Doug DeCinces	850.00	*	
164				

(Complete data not shown)

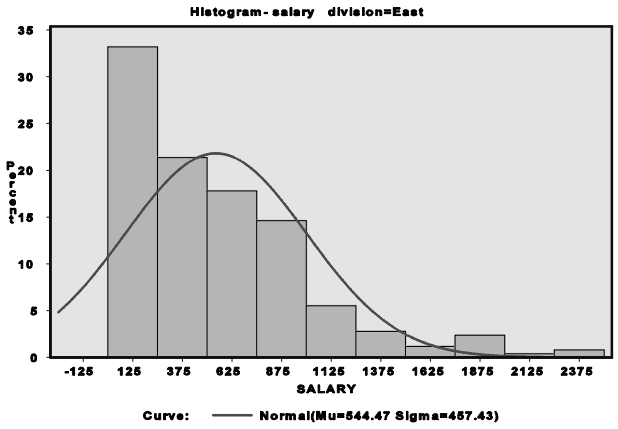


Figure 7 a: Distribution of salary (Division=East)

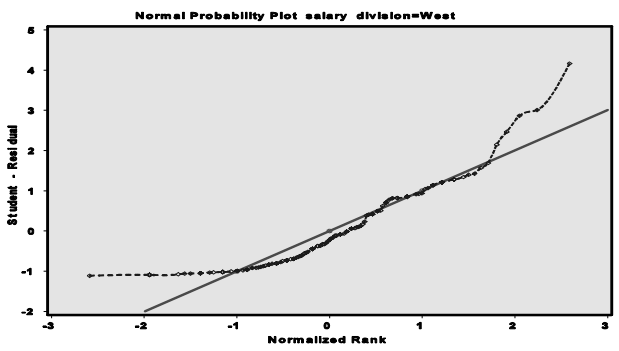


Figure 8 a: Normal probability plot (Division=West)

Figure 7 b: Distribution of salary (Division=West)

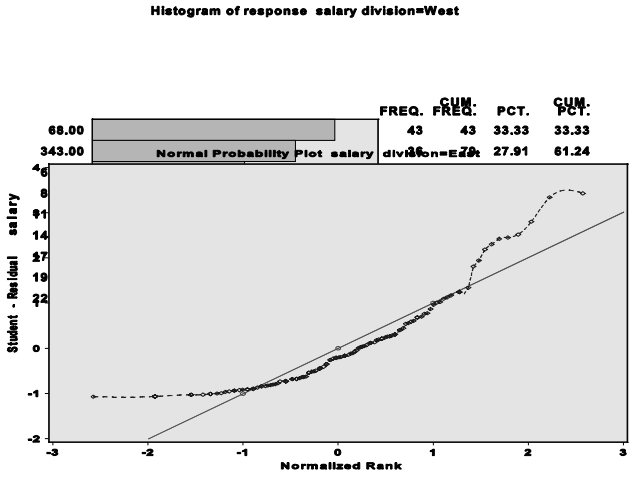


Figure 8 b: Normal probability plot (Division=East)

Summary:

As an alternative to the point-and-click menu interface modules for obtaining quick and complete results from exploratory analysis, a SAS macros based approach is presented here. The Excelsas macro is used to convert an excel file to a SAS data set. The UNIVAR macro is used to produce exploratory graphs. Other SAS macros for performing sampling methods, hypothesis testing, chi-square tests, ANOVA models, and Regression analysis are presented elsewhere (Fernandez 1998).

The majority of the data analysts are not computer programmers and they normally don't show much interest in learning a program-based statistics software. Using this **MACRO APPROACH,** the analysts can effectively and quickly perform complete data exploration and spend more time in interpretation of graphs and output rather than debugging their program errors etc. Furthermore, by using this approach, the analysts can simplify the steps involved in data analysis and improve the quality of statistical analysis and presentation methods.

References:

Fernandez, G.C.J 1998. Quick results from statistical data analysis using SAS macros. 1. Introductory statistical methods. Pages 266. APST270. Introductory statistical methods course Lab guide, Department of Applied Economics and Statistics / 204. University of Nevada Reno Reno NV 89557

****

The following SAS modules are necessary to run these SAS macros:

EXCELSAS macro: SAS/BASE and SAS/ ACCESS

UNIVAR macro: SAS/BASE SAS/STAT, SAS/GRAPH, and SAS/QC

Author's Contact address:

Dr. George C.J. Fernandez
Associate Professor in Applied Statistics
Department of Applied Economics/Statistics/204
University of Nevada- Reno Reno NV 89557.
(775) 784-4206 E-mail: GCJF@unr.edu

Home page: **http://apes.ag.unr.edu/george**

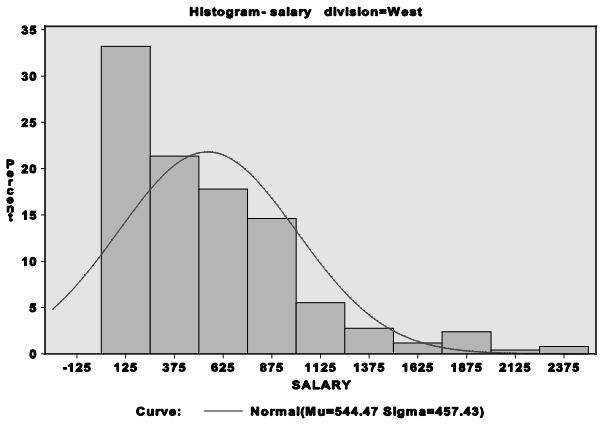
Appendix:

Other SAS macros for performing exploratory analysis

1.. Descriptive charts of one-way data

I. Descriptive frequency charts and graphs
Macro-call file: a:\mac-call\onewayql.sas
II. Descriptive SUM charts and graphs
Macro-call file: a:\mac-call\onewsum.sas
III. Descriptive MEAN charts and graphs
Macro-call file: a:\mac-call\onewmean.sas

2. Descriptive charts of two-way data



I. Descriptive frequency / percentage charts and graphs
Macro-call file: a:\mac-call\Twwyql.sas
II. Descriptive SUM charts and graphs
Macro-call file: a:\mac-call\Twwysum.sas
iii. Descriptive "MEAN" charts and graphs
Macro-call file: a:\mac-call\Twwymean.sas

3. Descriptive charts of Three-way data

I. Descriptive frequency / percentage charts and graphs
Macro-call file: a:\mac-call\Thwyql.sas
II. Descriptive "SUM" charts and graphs
Macro-call file: a:\mac-call\Thwysum.sas
III. Descriptive "MEAN" charts and graphs
Macro-call file: a:\mac-call\Thwymean.sas

4. Descriptive charts of Time series data

1. Trend plots
Macro-call file: a:\mac-call\Trend.sas

****