**FINAL GROUP PROJECT REPORT**

*ON*

**SUICIDE RATES DATASET CLEANING AND ANALYSIS**

***PREPARED BY: GROUP 4 (ZAA)***

SEPEHR SALEHI ……………………………..…. 130022221

RONGZHAO YI ………………………………….. 176938215

OGHALEOGHENE BLESSING AGBAWHE …. 119888220

CONGHAN ZHENG …………………..………….121647226

OLUWAFEMI SAMSON FAJOYE ……...………119891224

*COURSE:*

**BAN 130 - PROGRAMMING FOR ANALYTICS**

*SUBMITTED TO:*

**DR. OMAR ALTRAD**

**13TH APRIL, 2023.**

**TABLE OF CONTENTS**

1.0 Title Page 1

1.1 Table of Contents 2

2.0 Milestone One 3

2.1 Personal Objectives 3

2.2 Intended Output 3

2.3 Description of the needs of the intended audience 3

2.4 foreseeable challenges 3

3.0 Milestone Two 4

3.1 Description of data to be used 4

3.2 Foreseeable Challenges 4

4.0 Milestone Three 4

4.1 Complete Project’s SAS Code 4

5.0 Milestone Four 6

5.1 SAS Code, Output and Interpretation 6

6.0 Reference 18

**2.0 Milestone 1**

**2.1 Personal objective**

Our dataset is about suicide rates in various countries around the world from 1985 to 2016. The purpose of studying this dataset is to find groups or countries with high suicide rates in the data, to find the causes of this situation, and thus to achieve a reduction in suicide rates.

For this dataset, we have some questions to explore:

1. Is the suicide rate higher for men than for women?
2. Are suicide rates higher among the elderly than among the young?
3. Do suicide rates vary greatly between countries?

**2.2 Intended outcomes**

According to information searched from the Internet, the suicide rate is indeed higher among men and the elderly, and the suicide rate does vary greatly from country to country. However, we still need to explore the data to confirm this claim.

**2.3** **Description of the needs of the intended audience**

The target audience for this project is the government health sector in each country. They need to shape a healthy society to allow for increased citizen happiness. A reduction in suicide rates is a necessary goal to achieve this.

**2.4 Foreseeable challenges**

Probably the biggest challenge is the dataset itself. This dataset contains a very large number of items, and we do not yet have experience working with such datasets in SAS. This may take a while to learn.

**3.0 Milestone Two**

**3.1** **Description of the dataset to be used**

The dataset we used was the suicide rate statistics for each country between 1985 and 2016. This dataset was created by the authors to look for signals of rising suicide rates in different populations around the world.

**3.2 Foreseeable challenges**

There may be some difficulties in importing the data, after all, it is a huge data set. In addition, there may be difficulties in using SAS codes to explore the questions we are asking, as some SAS codes we may not have learned in class.

**4.0 Milestone Three**

**4.1 Complete Project’s SAS Code**

\*importing the dataset;

FILENAME REFFILE '/home/u63025740/BAN110/Suicide Rate Data.csv';

**PROC** **IMPORT** DATAFILE=REFFILE

DBMS=CSV

OUT=Suicide\_dataset;

GETNAMES=YES;

**RUN**;

**PROC** **CONTENTS** DATA=Suicide\_dataset; **RUN**;

**proc** **print** DATA=Suicide\_dataset (obs=**1000**); **RUN**;

\*checking missing values;

**proc** **means** data=Suicide\_dataset NMISS mean;

**run**;

**proc** **format**;

value $missfmt ' '='Missing' other='Not Missing';

**run**;

**proc** **freq** data=Suicide\_dataset;

format \_CHAR\_ $missfmt.;

tables \_CHAR\_ / missing missprint nocum nopercent;

**run**;

\*there is no missing values;

\*EDA;

\*top countries in total number of suicide;

**PROC** **SQL**;

create table total\_num\_suicide as

SELECT country, sum(suicides\_no) as total\_num\_suicide

FROM Suicide\_dataset

group by country;

**QUIT**;

**proc** **sql**;

create table sorted\_total\_num\_suicide as

select country, total\_num\_suicide from total\_num\_suicide

order by total\_num\_suicide desc;

**quit**;

**data** top10\_countries;

set sorted\_total\_num\_suicide (obs=**10**);

**run**;

**proc** **sgplot** data=top10\_countries;

hbar country / response=total\_num\_suicide categoryorder=respdesc;

yaxis display=(nolabel);

xaxis label='total number of suicide';

**run**;

\*mean number of suicides in different years;

**proc** **sql**;

create table suicides\_no as

select year, mean(suicides\_no) as mean from Suicide\_dataset

where **1990**<year<**2016**

group by year;

**quit**;

**proc** **gplot** data=suicides\_no;

plot mean\*year; symbol i=spline;

**run**;

**quit**;

\*calculating the correlation between variables;

**proc** **corr** data=Suicide\_dataset plots=matrix(histogram) PLOTS(MAXPOINTS=**1000000**);

**run**;

\*1. Personal objective;

\*Is the suicide rate higher for men than for women?;

**proc** **means** data=Suicide\_dataset sum nonobs;

class sex;

var suicides\_no;

**run**;

**proc** **SGPLOT** data = Suicide\_dataset;

vbar sex / response=suicides\_no;

title 'sum of suicides separated by gender';

**run**;

**quit**;

\*2. Are suicide rates higher among the elderly than among the young?;

**proc** **means** data=Suicide\_dataset sum nonobs;

class age;

var suicides\_no;

**run**;

**proc** **SGPLOT** data = Suicide\_dataset;

vbar age / response=suicides\_no categoryorder=respdesc;

title 'sum of suicides separated by age';

**run**;

**quit**;

\*3. Do suicide rates vary greatly between countries?;

\*Running ANOVA test

null hypothesis the mean of suicide in different countries is equal

alternative hypothesis the mean on suicide in different countries is different;

**proc** **anova** data = Suicide\_dataset PLOTS;

class country;

model suicides\_no = country;

**run**;

\* since the p-value is less than 0.05 we reject the null hypothesis,

Thus, the mean on suicide in different countries is different;

\*Foreseeable challenges;

**PROC** **SGPLOT** data=suicides\_no;

SCATTER x=year y=mean / markerattrs=(symbol=circlefilled);

REG x=year y=mean /lineattrs=(color=red thickness=**2** pattern=dot);

**RUN**;

**5.0 Milestone Four**

**5.1 SAS Code, Output and Report**

* **Importing of Data**

FILENAME REFFILE '/home/u63025740/BAN110/Suicide Rate Data.csv';

**PROC** **IMPORT** DATAFILE=REFFILE

DBMS=CSV

OUT=Suicide\_dataset;

GETNAMES=YES;

**RUN**;

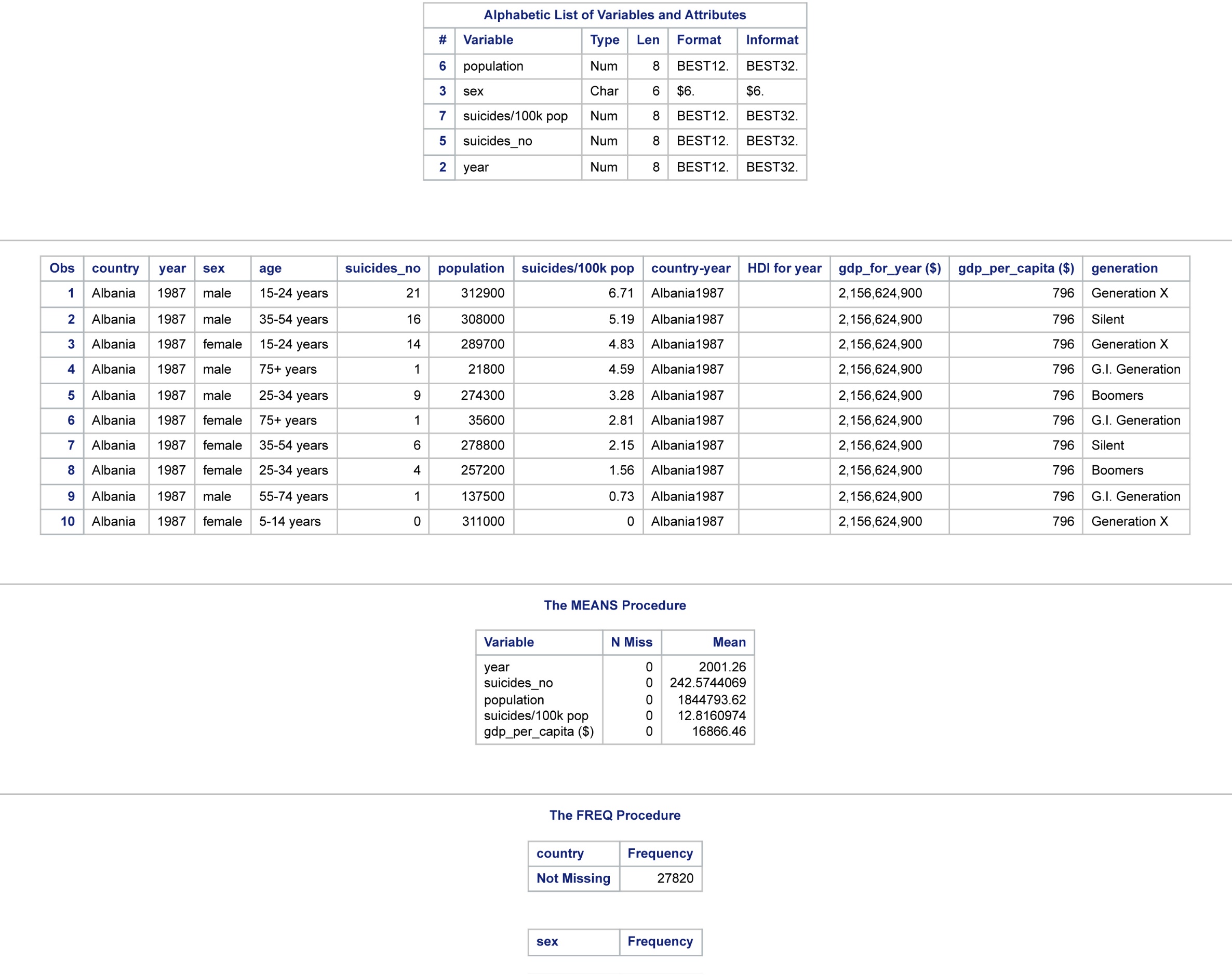
**PROC** **CONTENTS** DATA=Suicide\_dataset;

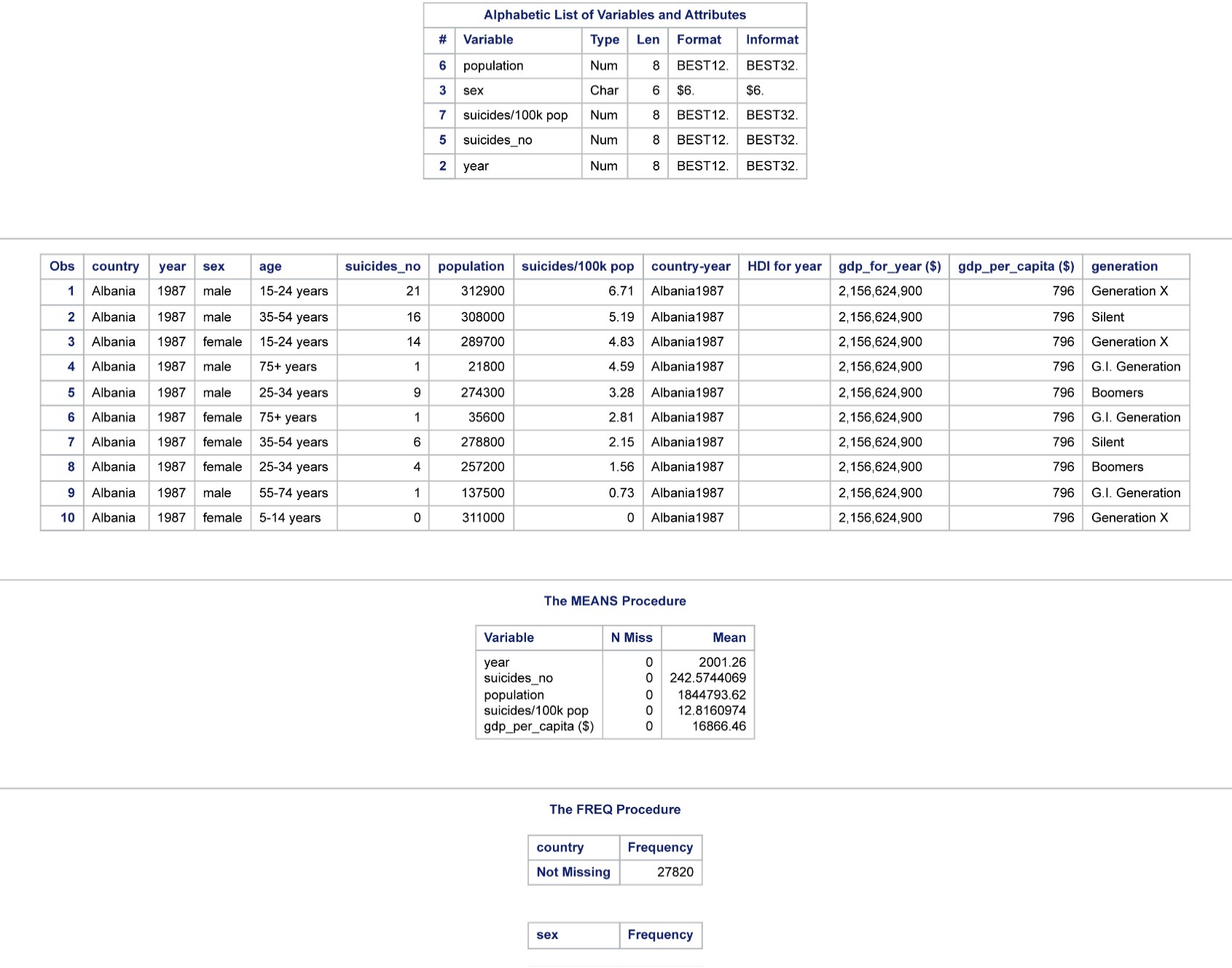
**RUN**;

**proc** **print** DATA=Suicide\_dataset (obs=**1000**);

**RUN**







In this stage, the dataset used for this project was imported into SAS studio using the Proc Import step, as shown above. The dataset is named “Suicide Rates Dataset.” The Proc Contents statement used in the above code instructs SAS to display metadata information about the dataset, such as the number of observations and variables, variable names, data types, formats, and lengths. On the other hand, the Proc Print statement was used to display the actual data values for each observation and variable. Specifically, 10 out of the 27,820 observations in the dataset were displayed.

* **Checking for missing values**

\*checking missing values;

**proc** **means** data=Suicide\_dataset NMISS mean;

**run**;

**proc** **format**;

value $missfmt ' '='Missing' other='Not Missing';

**run**;

**proc** **freq** data=Suicide\_dataset;

format \_CHAR\_ $missfmt.;

tables \_CHAR\_ / missing missprint nocum nopercent;

**run**;

\*there is no missing values;





The above SAS code was used to detect any missing values in our dataset. The Proc Means, Proc Format, and Proc Freq procedures were used to compute descriptive statistics, define custom formats, and generate frequency tables, as shown in the above SAS output. No missing values were detected in the dataset.

* **Finding groups or countries with high suicide rates in the data**

\*EDA;

\*top countries in total number of suicide;

**PROC** **SQL**;

create table total\_num\_suicide as

SELECT country, sum(suicides\_no) as total\_num\_suicide

FROM Suicide\_dataset

group by country;

**QUIT**;

**proc** **sql**;

create table sorted\_total\_num\_suicide as

select country, total\_num\_suicide from total\_num\_suicide

order by total\_num\_suicide desc;

**quit**;

**data** top10\_countries;

set sorted\_total\_num\_suicide (obs=**10**);

**run**;

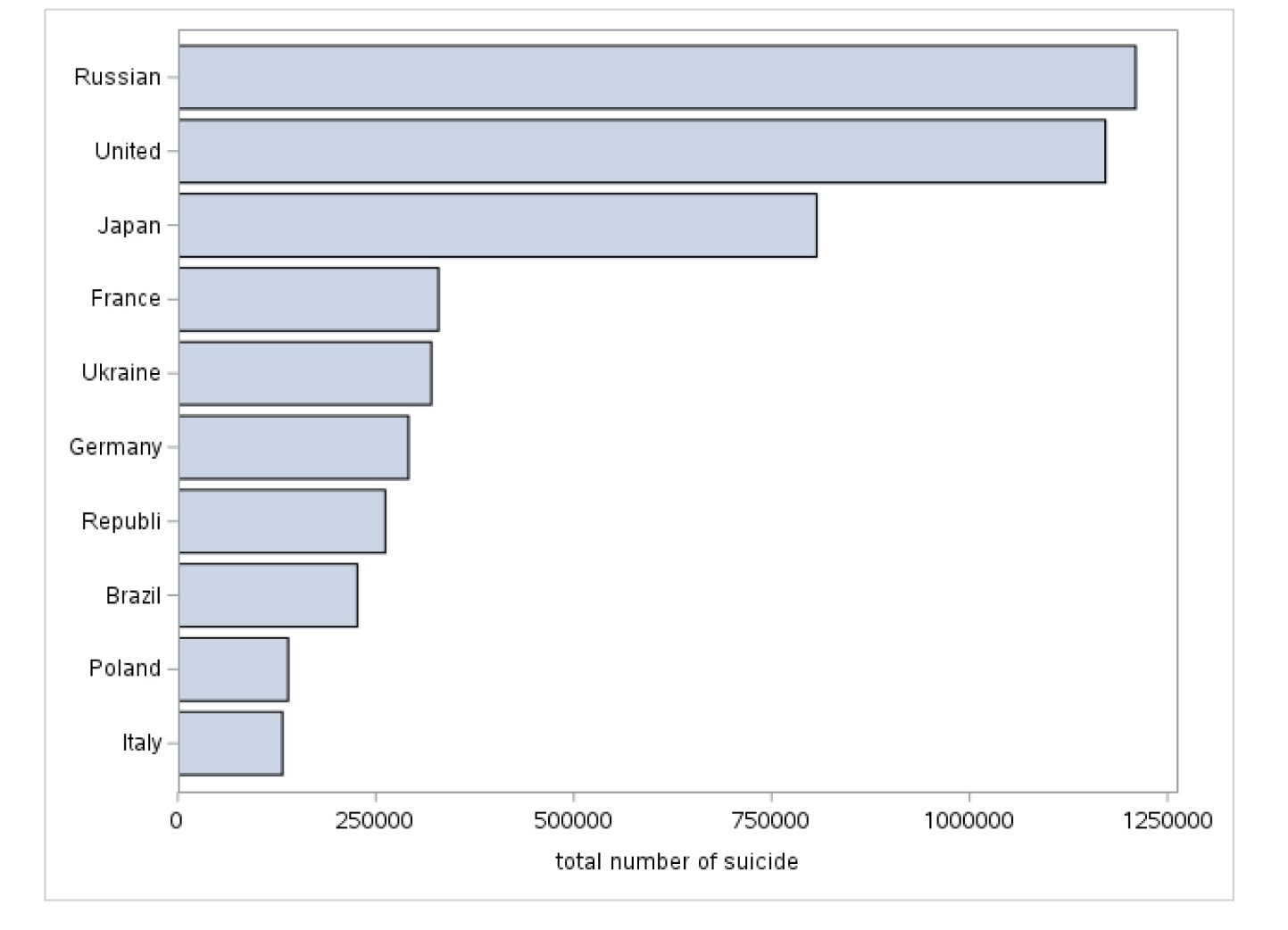
**proc** **sgplot** data=top10\_countries;

hbar country / response=total\_num\_suicide categoryorder=respdesc;

yaxis display=(nolabel);

xaxis label='total number of suicide';

**run**;



Further to our achievement of personal objectives, the Proc SQL procedure was applied in our coding to identify countries with higher suicide rates. Additionally, the Proc SGPLOT procedure was used to generate a horizontal bar chart to visualize the top ten countries with the highest suicide rates. According to the output, Russia tops the list, followed by the United States, Japan, France, Ukraine, Germany, Republic of Congo, Brazil, and Poland, with Italy ranking the lowest.

* **Determining the mean of the top ten countries number of suicides in different years**

\*mean number of suicides in different years;

**proc** **sql**;

create table suicides\_no as

select year, mean(suicides\_no) as mean from Suicide\_dataset

where **1990**<year<**2016**

group by year;

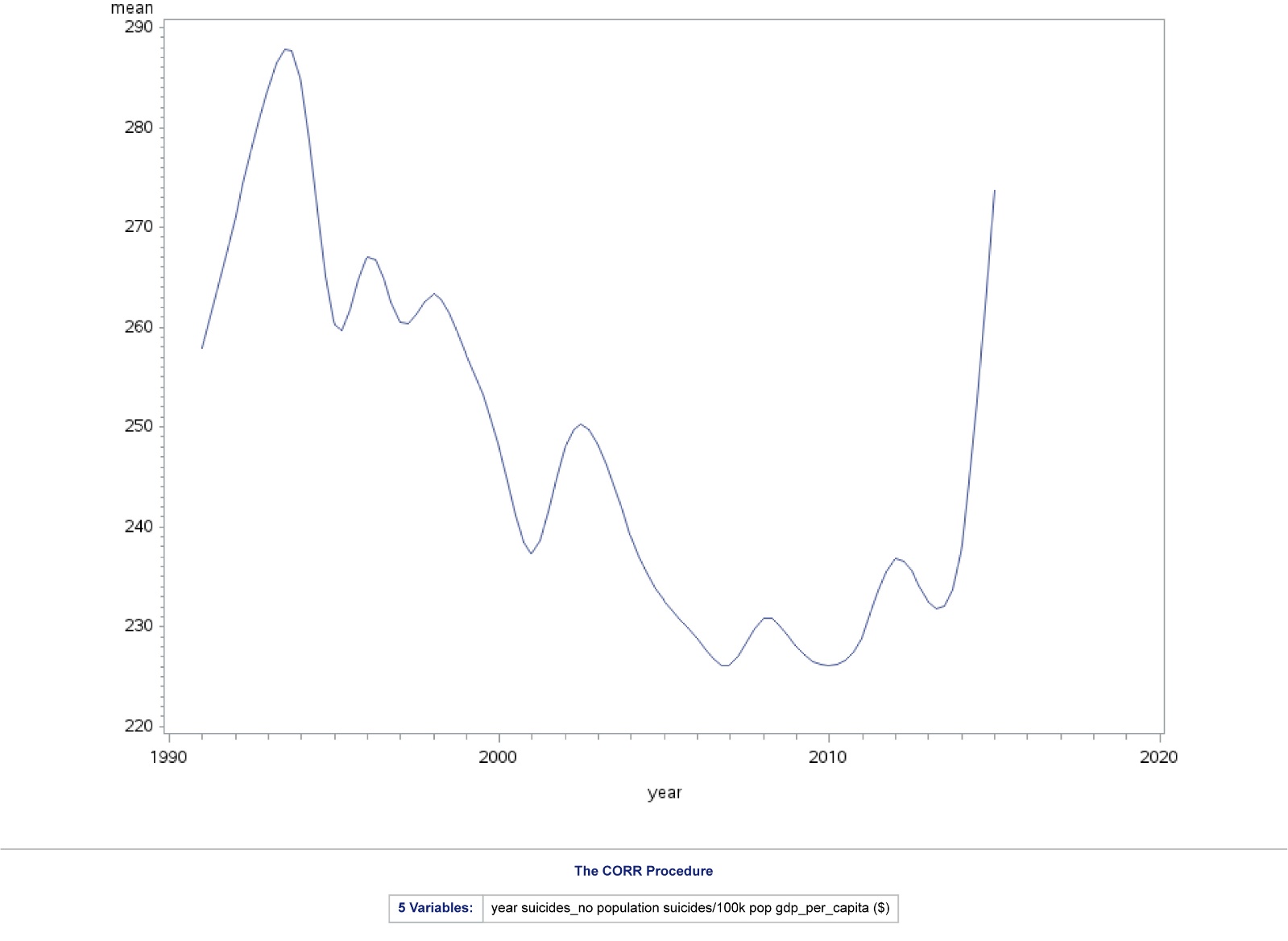
**quit**;

**proc** **gplot** data=suicides\_no;

plot mean\*year; symbol i=spline;

**run**;

**quit**;



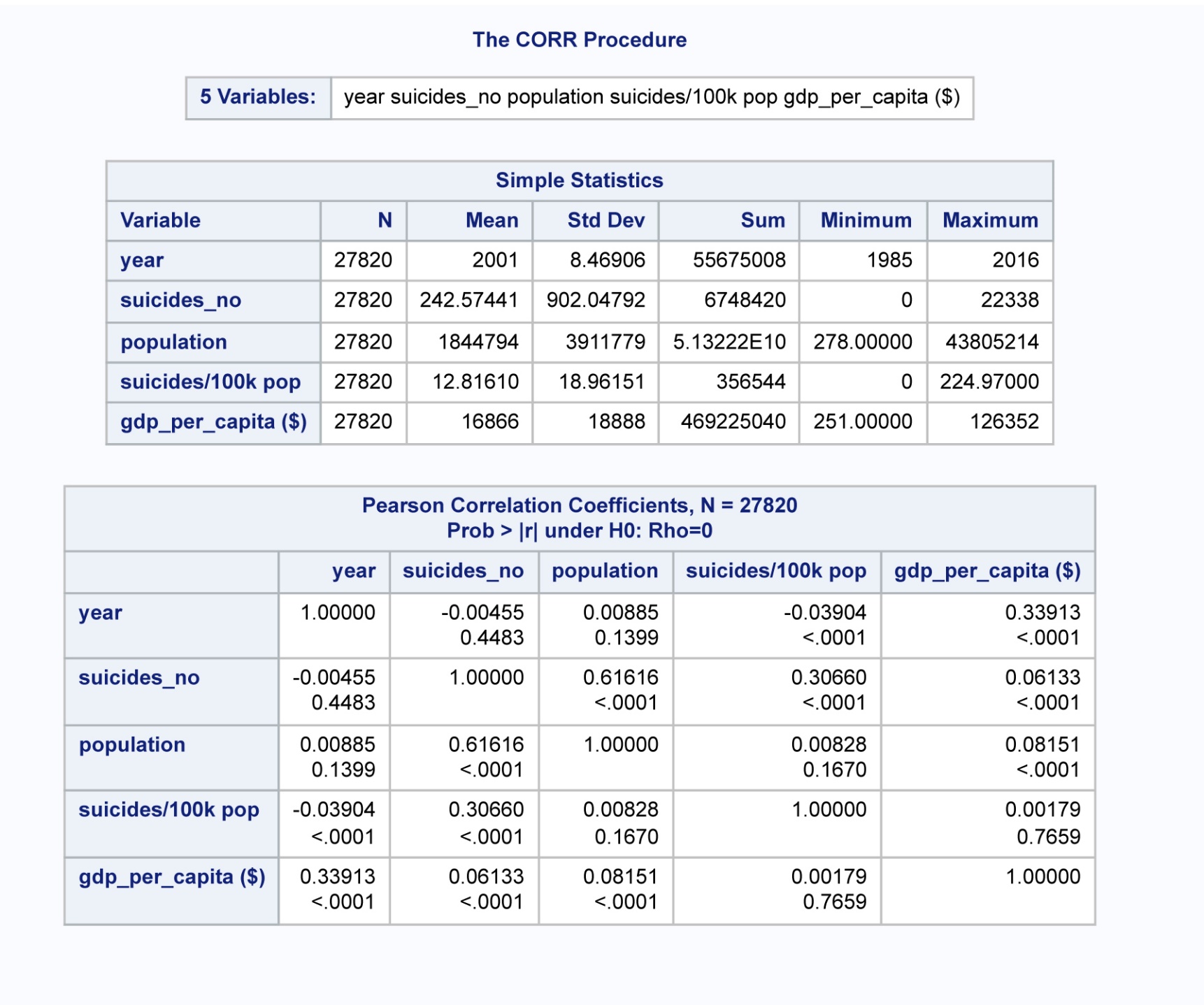
We also checked for differences in the mean number of suicides across the top ten countries between the years 1990 and 2016. We used Proc SQL to instruct SAS to generate the mean of the committed number of suicides within the years 1990 and 2016, while SAS Gplot procedure was used to visualize the trends. The above output revealed that 1993 had the highest mean number of suicides.

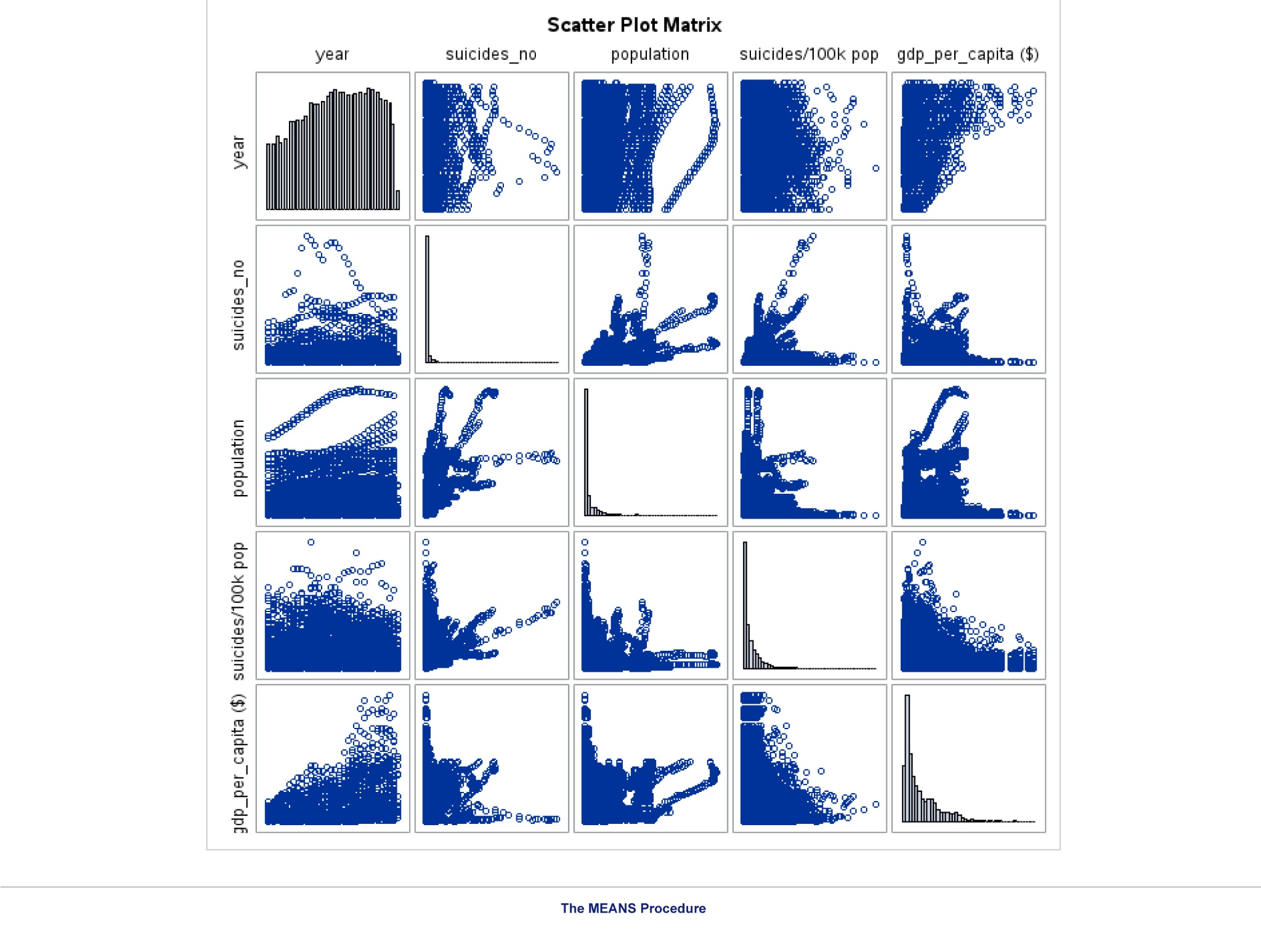
* **Determining if there is any relationship between the variables by calculating the correlation coefficients**

\*calculating the correlation between variables;

**proc** **corr** data=Suicide\_dataset plots=matrix(histogram) PLOTS(MAXPOINTS=**1000000**);

**run**;





Statistical calculation was also carried out by us to determine the level of relationship or association between five variables of the selected dataset. Proc Corr procedure was applied to instruct SAS to generate the Pearson correlation coefficient between year, suicides number, population suicides/100k pop, and GDP per capital of the ten selected countries.The SAS output reveal 61.62% (0.61616) representing moderate positive correlation between suicide number and population. The other variables as indicated in the above table has a weak positive relationship with each other.

Following the above analysis result, a significant relationship was statistically proven among the five variables.

* **Answering our personal objective questions**

**Q1- Is the suicide rate higher for men than the women?**

\*1. Personal objective;

\*Is the suicide rate higher for men than for women?;

**proc** **means** data=Suicide\_dataset sum nonobs;

class sex;

var suicides\_no;

**run**;

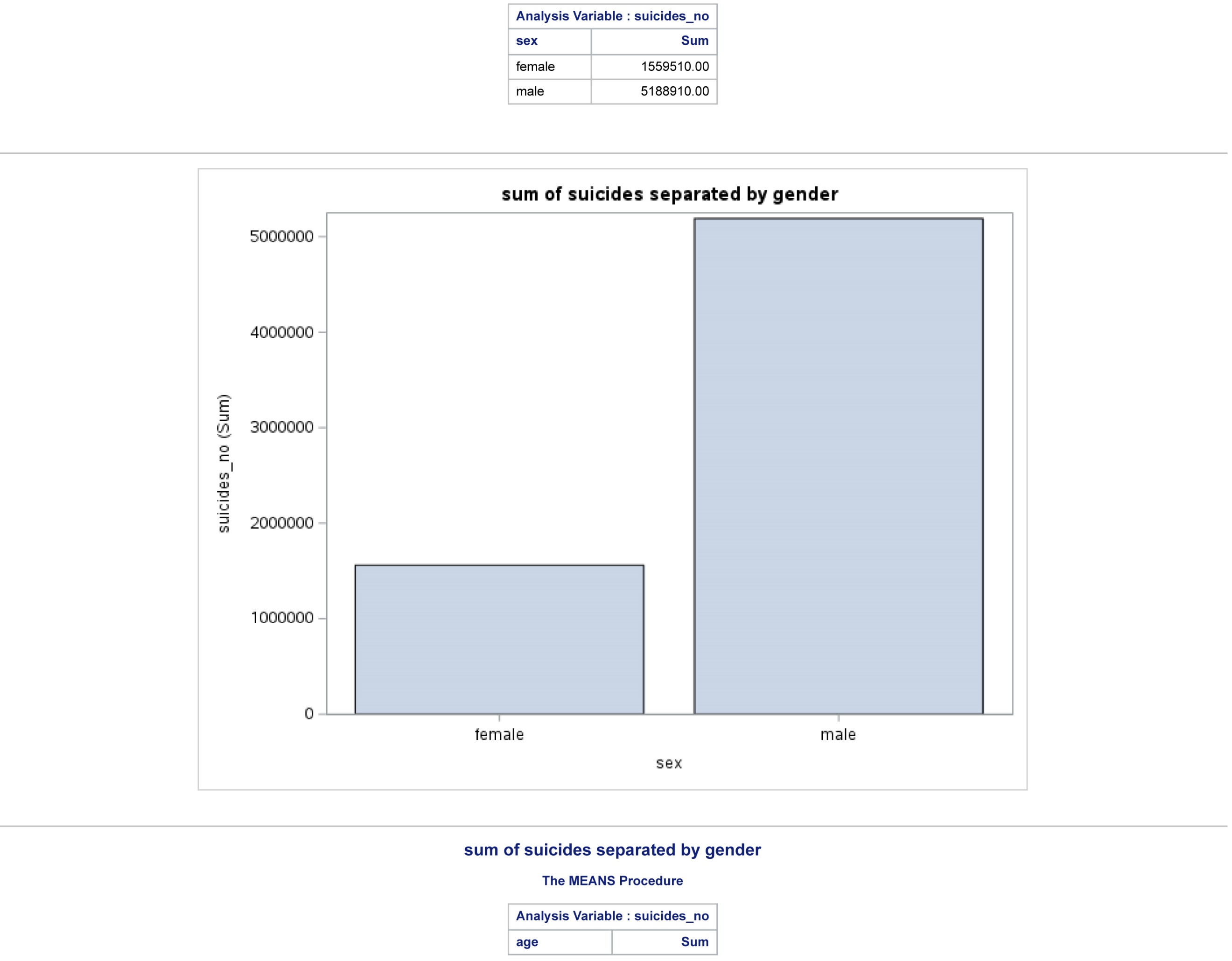
**proc** **SGPLOT** data = Suicide\_dataset;

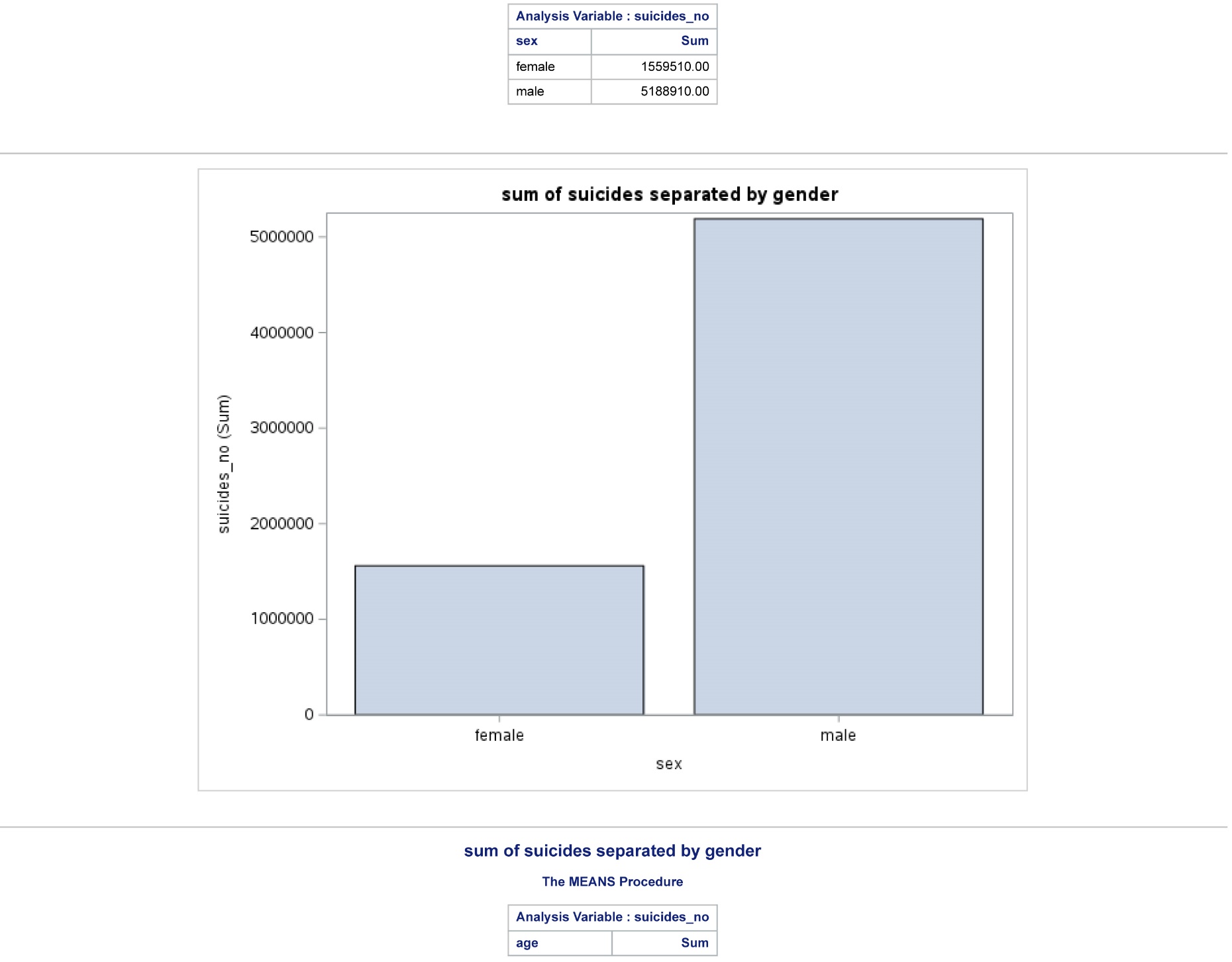
vbar sex / response=suicides\_no;

title 'sum of suicides separated by gender';

**run**;

**quit**;





The SAS output answer our first personal question revealing male has the highest number of suicide rate with a total of 5,188,910.00 representing 76.89% of the sample size. The female counterpart on the other hand, commit lesser suicide with 23.22% (1,559,510.00) of the sample size.

**Q2 Are suicide rates higher among the elderly than among the young?**

\*2. Are suicide rates higher among the elderly than among the young?;

**proc** **means** data=Suicide\_dataset sum nonobs;

class age;

var suicides\_no;

**run**;

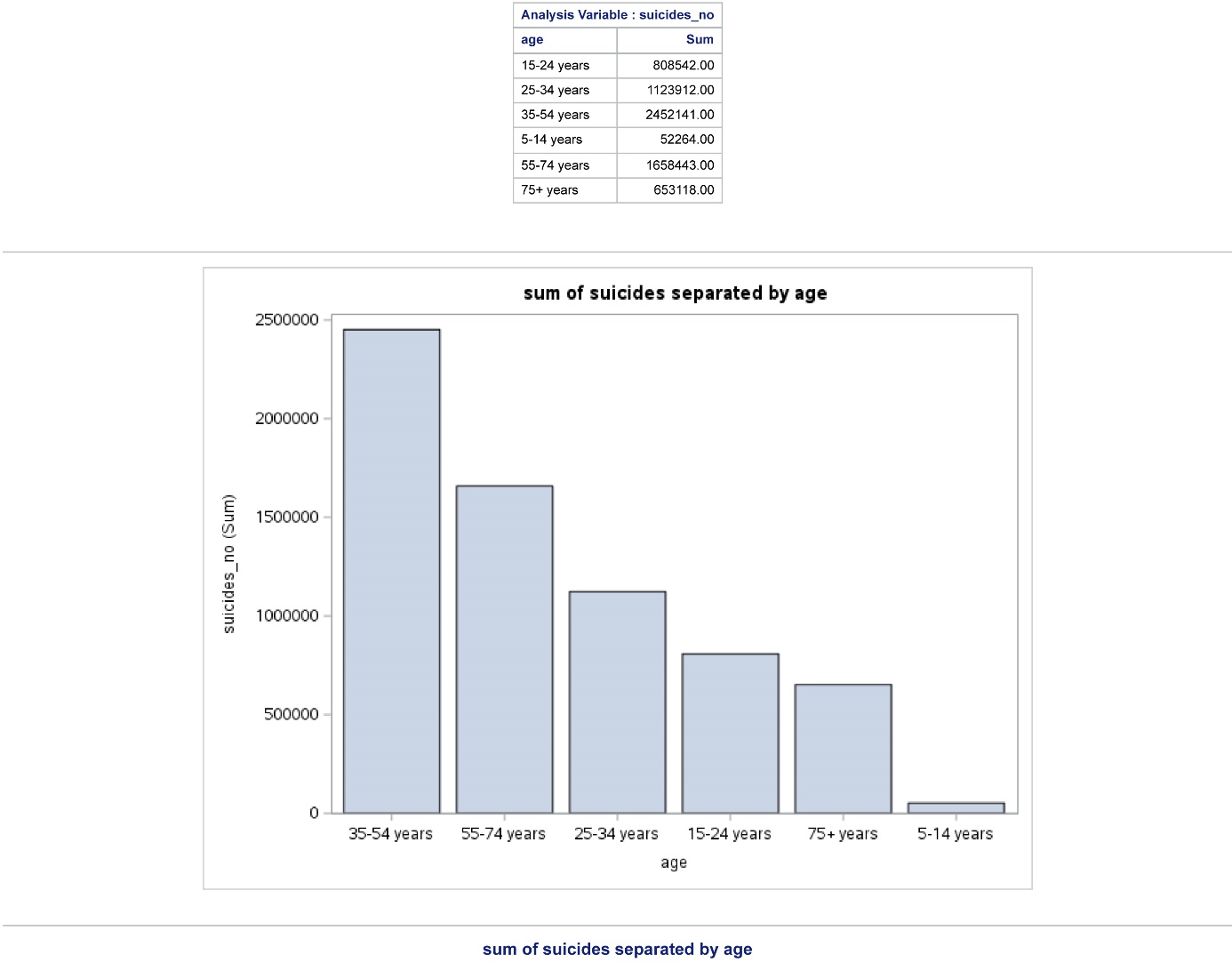
**proc** **SGPLOT** data = Suicide\_dataset;

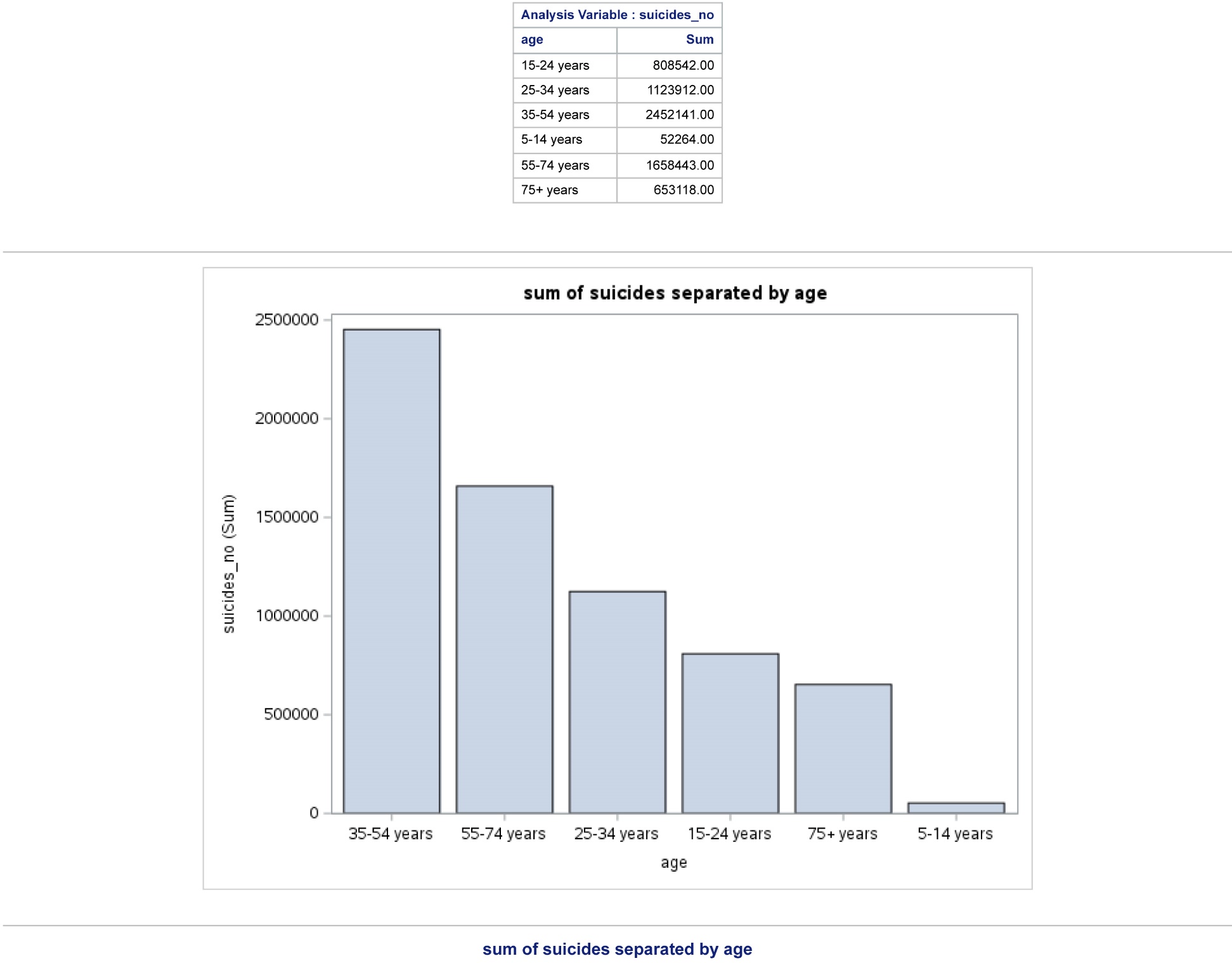
vbar age / response=suicides\_no categoryorder=respdesc;

title 'sum of suicides separated by age';

**run**;

**quit**;





**Q3 Do suicide rates vary greatly between countries?**

3. Do suicide rates vary greatly between countries?;

\*Running ANOVA test

null hypothesis the mean of suicide in different countries is equal

alternative hypothesis the mean on suicide in different countries is different;

**proc** **anova** data = Suicide\_dataset PLOTS;

class country;

model suicides\_no = country;

**run**;

\* since the p-value is less than 0.05 we reject the null hypothesis,

Thus, the mean on suicide in different countries is different;

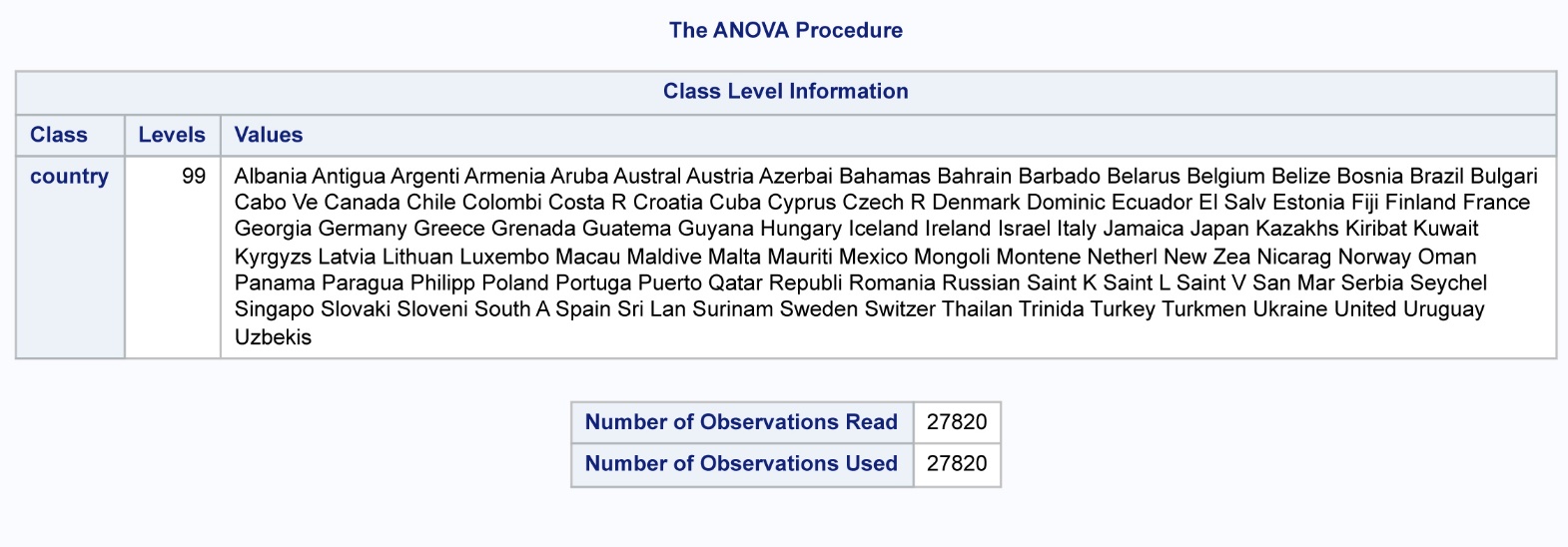
\*Foreseeable challenges;

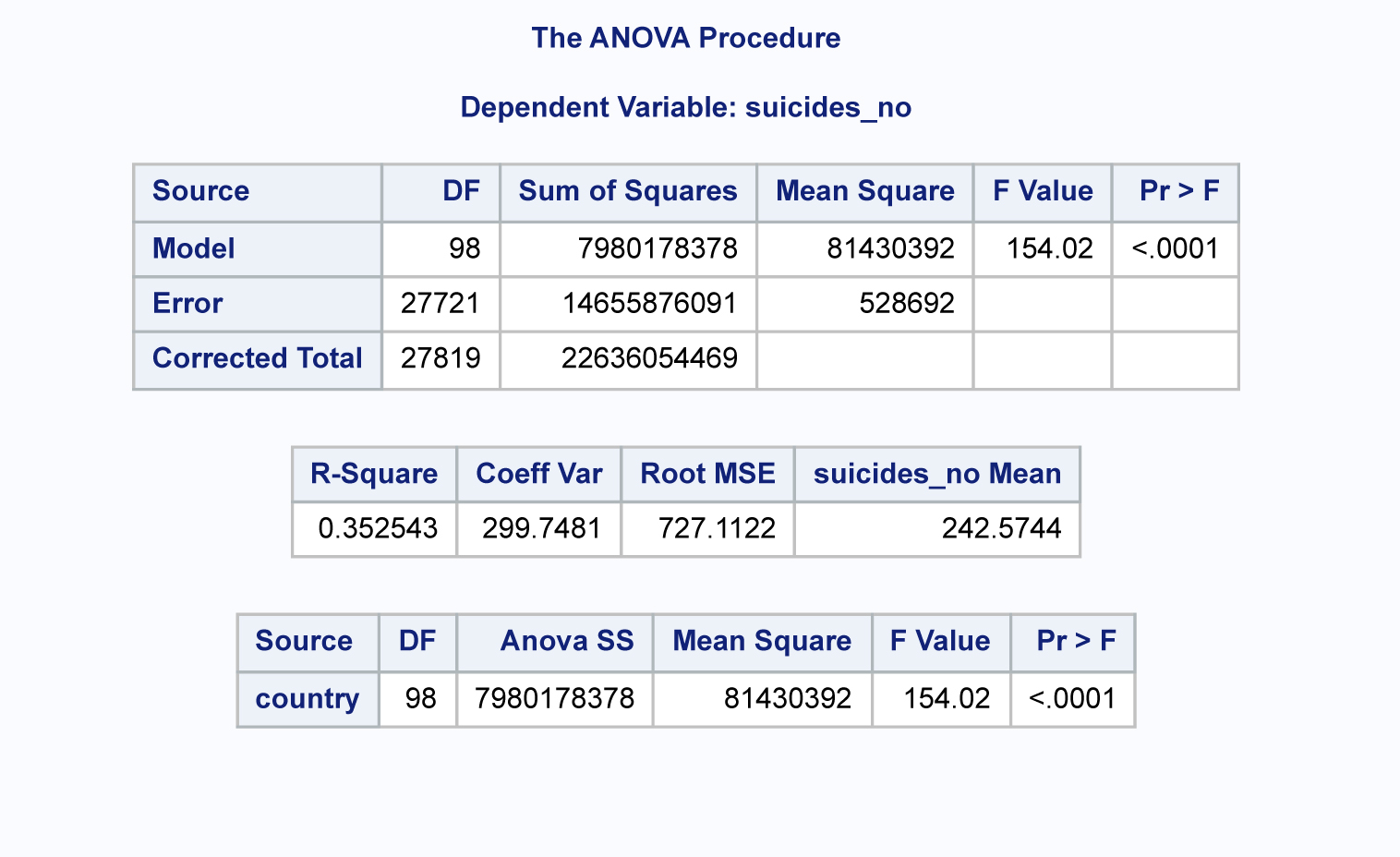
**PROC** **SGPLOT** data=suicides\_no;

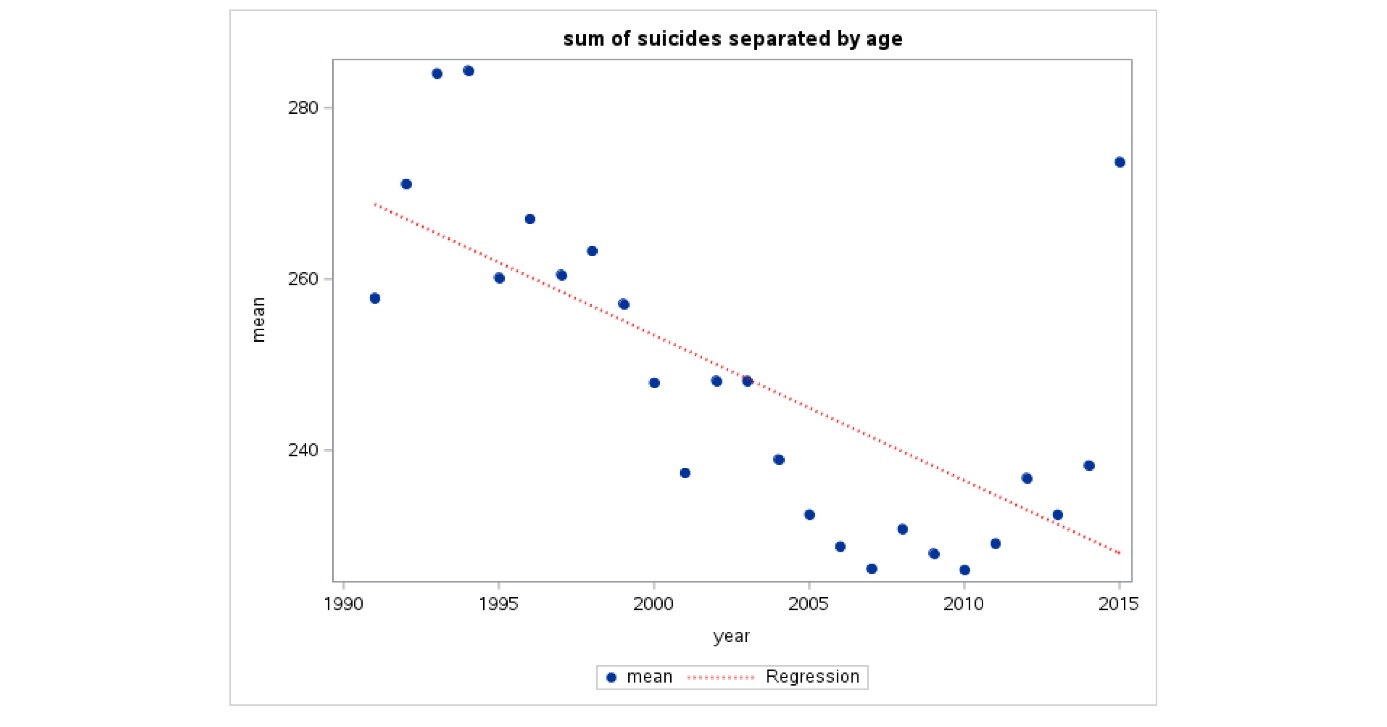
SCATTER x=year y=mean / markerattrs=(symbol=circlefilled);

REG x=year y=mean /lineattrs=(color=red thickness=**2** pattern=dot);

**RUN**;



****



Attempt solving our third explored questions, using ANOVA to test the suicide differences among the different countries.

From the SAS result indicated in the above table, the calculated P-Value is <.0001 which is less than standard value <0.05.

This means that null hypothesis is rejected in favour of the alternative hypothesis to conclude that the mean on suicide in different countries is different.

**6.0 References**