Seneca

Academic Year	**2023**
Semester	☒ Fall ☐ Winter ☐ Summer
Course Code - Name	BAN110 - Data Preparation and Handling
Instructor	Muhammad Rehman Zafar
Assessment	Projects

Student ID	**Student Name**	**Role**
113265227	Prashant Rokka	Group Lead
123445231	Sashank Ghimire	Member
143802221	Ankit Yadav	Member
129394235	Anik Bandu Das	Member
142488170	Malathi Chanti	Member

Projects

You are required to choose a project from the list of the projects specified in this document and complete it within groups of **three**.

Since this is a group project, it is required to be done in groups of **3**. Each group should have a Group Lead who would be responsible for submitting the project on Blackboard (Please note that not all the members of the group are required to submit the project separately on Blackboard. **One submission from the Group Lead would be sufficient**).

The detailed requirements for each project are available in this document, so please go through the details and fulfil all the requirements to avoid missing any marks.

Finally, follow the below mentioned instructions carefully.

Instructions:

To obtain maximum marks in this assessment, please ensure the followings:

- Don’t forget to write your name and ID on the first page of this document. The student IDs and names of all the students in the group should be mentioned along with the roles.
- Submit the project by writing your solution in this document under the Solution heading below. Do not use a separate document. Everything related to the project should be included in this document, e.g., code, screenshots etc.
- This project has a weightage of **24%** marks of the course.
- This is a group project so **only 1 submission from the group lead is required.**
- Group Leads are required to submit the project on Blackboard as instructed. Submissions through email will not be accepted.
- The project deadline is **midnight December 5, 2023**. Submissions after the deadline will not be accepted.
- A separate session for presentation and QA for the project will be scheduled.
- Upload presentation slides separately to the Blackboard.

Rubric:
Your assessment will be graded based on the following rubric:

	Excellent (7 - 10)	**Average (4 – 6.9)**	**Poor (<4)**
Project Completion and Code **(14)**	The project was completed without any errors and output is as expected. Fulfills all/most of the requirements for the project.	The project was completed with few errors. Fulfills some of the requirements for the project.	The project is incomplete. Does not fulfill all/most of the requirements.
Presentation and QA **(5)**	The student has a good contribution to the project. Knows the ins and outs of the project. The student has presented his/her part of the project very well. Knows everything / most of his/her part.	The student has an average contribution to the project. Does not know the whole project. The student has averagely presented his/her part of the project. Knows few of the things about his/her part.	The student has no contribution to the project. Does not know anything / most about the project. The student has poorly presented the project. Does not know much about the project.
Report **(5)**	Student has contributed well in preparing the project report and knows all the aspects of the report.	Student has contributed partially in preparing the project report and knows some aspects of the report.	Student has not contributed in preparing the report.

Project Instructions

You are provided with a few datasets however; you are free to pick any dataset you like to work on as a group. You are required to demonstrate at least the following skills in the project:

1. Dataset and task description
2. Data Import
   - This phase requires you to import the data from the provided excel file into SAS using Proc Import.
3. Dataset Characteristics and Cleaning
   - This phase requires you to clean your data before data analysis phase. You should use at least following concepts to complete this phase:
     1. Extract relevant data from the original dataset
     2. Convert a numeric column to character column or vice versa
     3. Create a new column based on existing columns and use it in your analysis
     4. Identify missing values and remove / replace using an appropriate technique
     5. Use built-in SAS function(s) to perform data cleaning, e.g., extracting year from the data column etc.

For example, if:

- Target variable
  1. If categorical, show the frequency distribution of each of the possible values. Interpret. Is the dataset balanced? Any other comment?
  2. If numerical, show the statistics (min, max, mean) and the shape of the distribution of the target variable through a histogram. In some case, numerical target variables need transformation to make data modeling possible.
- Categorical variables
  1. Check and correct errors when necessary.
  2. Check and treat missing values through imputation with the mode.
  3. Create one or more derived variables. Justify why the derived variable is created? Does it answer a specific question? Does it serve for data modeling? Etc..
- Numerical variables
  1. Check (range of values/ less than/larger than) and correct errors by deletion.
  2. Check for missing values and correct through imputation with the mean.
  3. Check the distribution of one or more numerical variables to decide which method to use for outlier detection.
  4. Detect and remove outliers.
  5. Test for normality and plot histogram and QQ plots for a variable with a skewed distribution. Apply a transformation and test for normality again with histogram and QQ plot.

4. Data Analysis
    - This phase requires you to analyze your cleaned dataset to answer at least 3 valid business questions. You are free to pick any business questions you like, however, please keep in mind that picking good business questions to answer would result in better marks.
5. Project Report
    - This phase requires you to create a report in MS Word with the following requirements:
        1. Explain each and every phase of the project (from Phase 1 to 4) along with the screenshots of the output and the related SAS code
        2. Include answers to questions in Phase 4 in your report
        3. Create at least 1 graph / chart in your report which can be simply a Box plot to identify outliers etc.
        4. Make sure not to miss any phase and output of its screenshot

Dataset Options

1. Auto-mpg dataset:
   https://www.kaggle.com/uciml/autompg-dataset
2. Heart disease dataset
   https://www.kaggle.com/ronitf/heart-disease-uci
3. Census income dataset
   https://www.kaggle.com/uciml/adult-census-income
4. Bike sharing dataset
   https://www.kaggle.com/marklvl/bike-sharing-dataset
5. Suicide rates dataset:
   https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
6. Breast Cancer
   https://archive.ics.uci.edu/dataset/14/breast+cancer

You are free to use any other dataset from the following sources. Please make sure the dataset meets the requirements listed in dataset requirements section.

Kaggle: https://www.kaggle.com/datasets
UCI: https://archive.ics.uci.edu/dataset

Solution:

Dataset and Task Description:

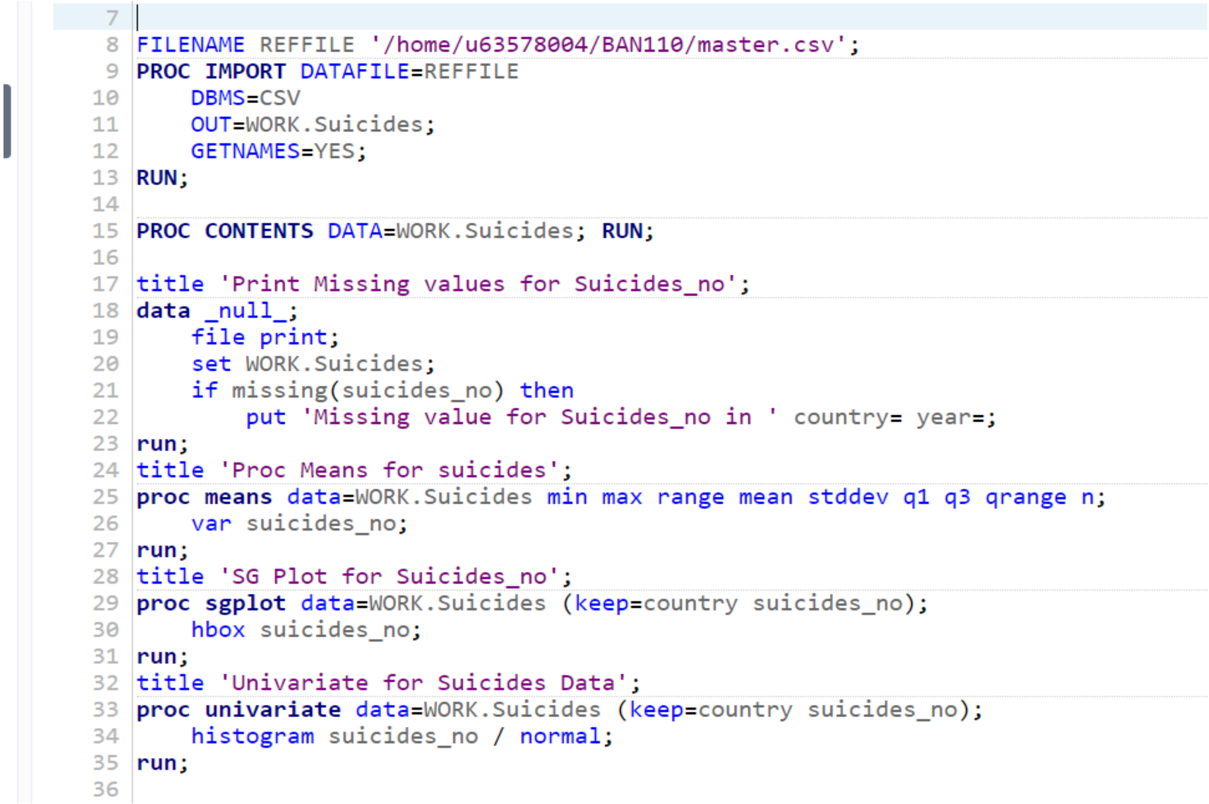
The team has undertaken a comprehensive analysis of suicide rates utilizing a dataset available at Link: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.

The provided SAS code demonstrates adept data preprocessing and analysis, incorporating procedures such as PROC IMPORT, PROC CONTENTS, PROC MEANS, PROC UNIVARIATE, PROC SGPlot, PROC RANK, among others. The code seamlessly handles tasks including data importation, managing missing values, generating insightful summary statistics, detecting outliers, variable transformation, and addressing key business questions. Noteworthy accomplishments include the succinct summarization of male and female suicides, as well as the identification of years exhibiting both the highest and lowest suicide rates.

Data Characteristics, Import Analysis and Cleaning:

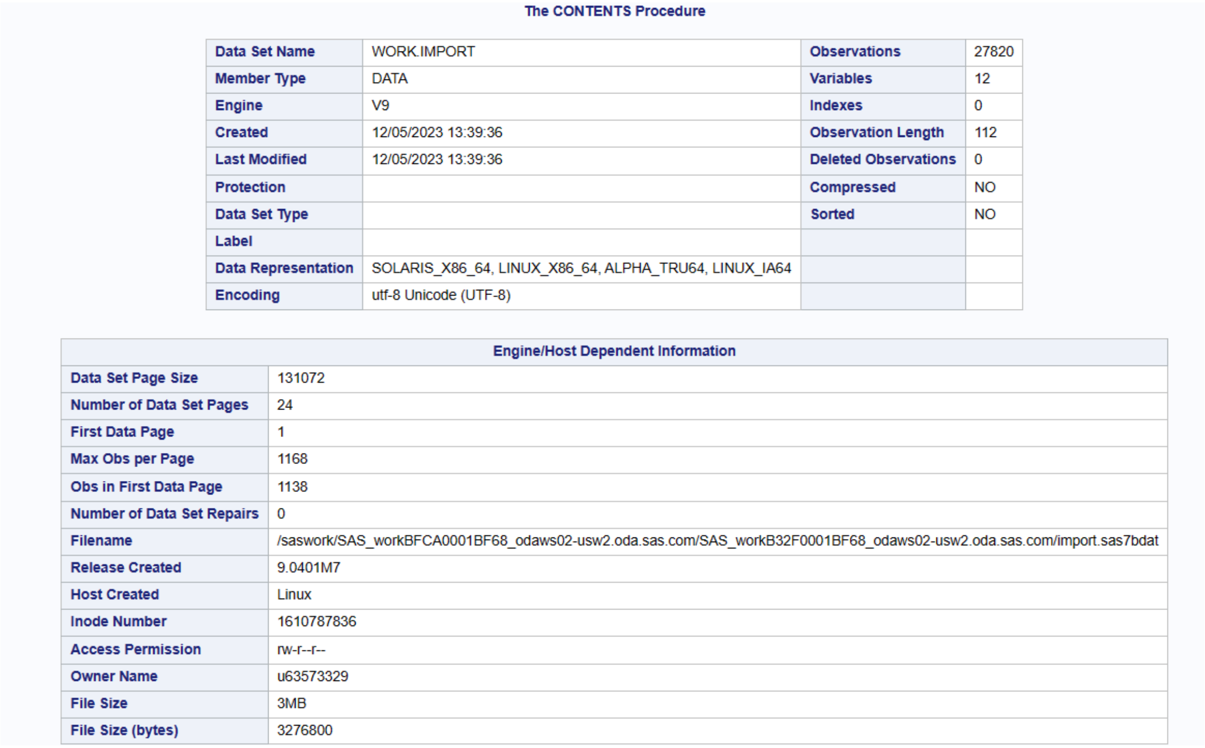
The code begins by exploring the structure of the 'WORK.Suicides' dataset using PROC CONTENTS. It then identifies missing values in the 'suicides_no' variable, computes descriptive statistics, and generates visualizations (box plot and histogram) to understand the distribution of suicide numbers. Outliers are detected and trimmed, contributing to a more robust dataset for analysis.

SAS Code:



```
FILENAME REFFILE '/home/u63578004/BAN110/master.csv';
PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=WORK.Suicides;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.Suicides; RUN;

title 'Print Missing values for Suicides_no';
data _null_;
    file print;
    set WORK.Suicides;
    if missing(suicides_no) then
        put 'Missing value for Suicides_no in ' country= year=;
run;
title 'Proc Means for suicides';
proc means data=WORK.Suicides min max range mean stddev q1 q3 qrange n;
    var suicides_no;
run;
title 'SG Plot for Suicides_no';
proc sgplot data=WORK.Suicides (keep=country suicides_no);
    hbox suicides_no;
run;
title 'Univariate for Suicides Data';
proc univariate data=WORK.Suicides (keep=country suicides_no);
    histogram suicides_no / normal;
run;
```

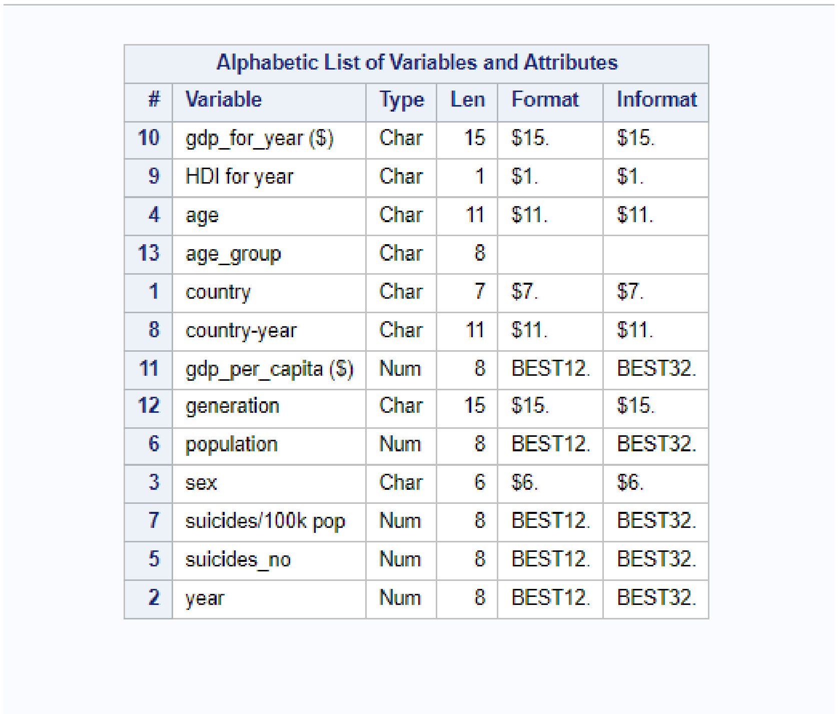
Output:



The CONTENTS Procedure

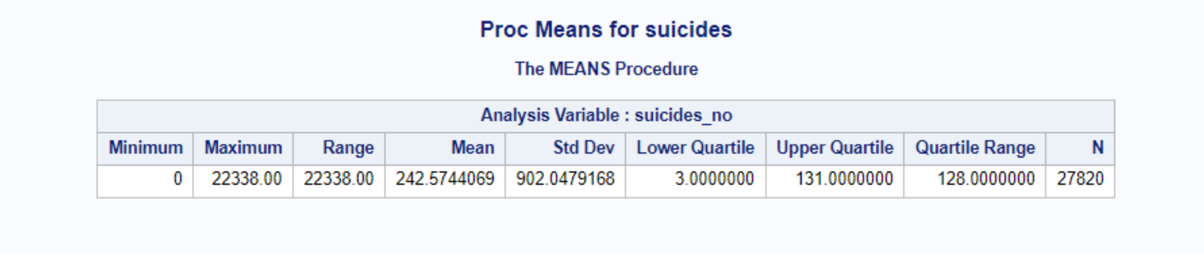
Data Set Name	WORK.IMPORT	Observations	27820
Member Type	DATA	Variables	12
Engine	V9	Indexes	0
Created	12/05/2023 13:39:36	Observation Length	112
Last Modified	12/05/2023 13:39:36	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	24
First Data Page	1
Max Obs per Page	1168
Obs in First Data Page	1138
Number of Data Set Repairs	0
Filename	/saswork/SAS_workBFCA0001BF68_odaws02-usw2.oda.sas.com/SAS_workB32F0001BF68_odaws02-usw2.oda.sas.com/import.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1610787836
Access Permission	rw-r--r--
Owner Name	u63573329
File Size	3MB
File Size (bytes)	3276800



Alphabetic List of Variables and Attributes

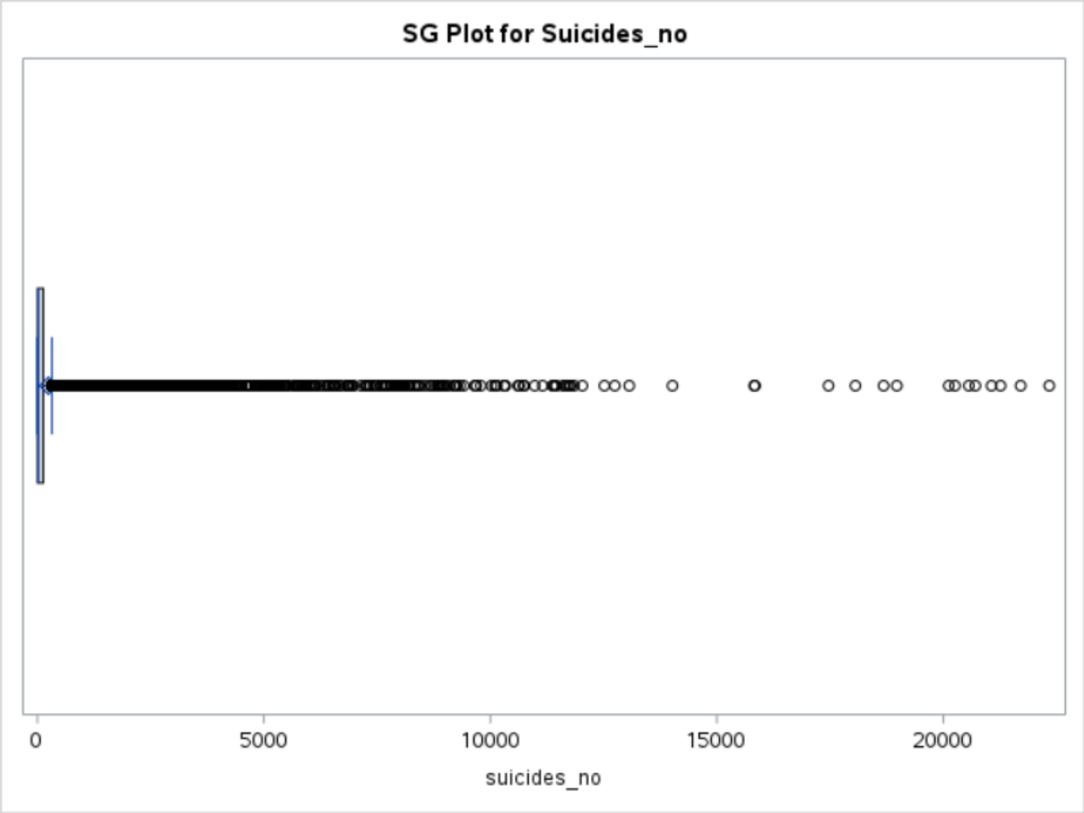
#	Variable	Type	Len	Format	Informat
10	gdp_for_year ($)	Char	15	$15.	$15.
9	HDI for year	Char	1	$1.	$1.
4	age	Char	11	$11.	$11.
13	age_group	Char	8		
1	country	Char	7	$7.	$7.
8	country-year	Char	11	$11.	$11.
11	gdp_per_capita ($)	Num	8	BEST12.	BEST32.
12	generation	Char	15	$15.	$15.
6	population	Num	8	BEST12.	BEST32.
3	sex	Char	6	$6.	$6.
7	suicides/100k pop	Num	8	BEST12.	BEST32.
5	suicides_no	Num	8	BEST12.	BEST32.
2	year	Num	8	BEST12.	BEST32.



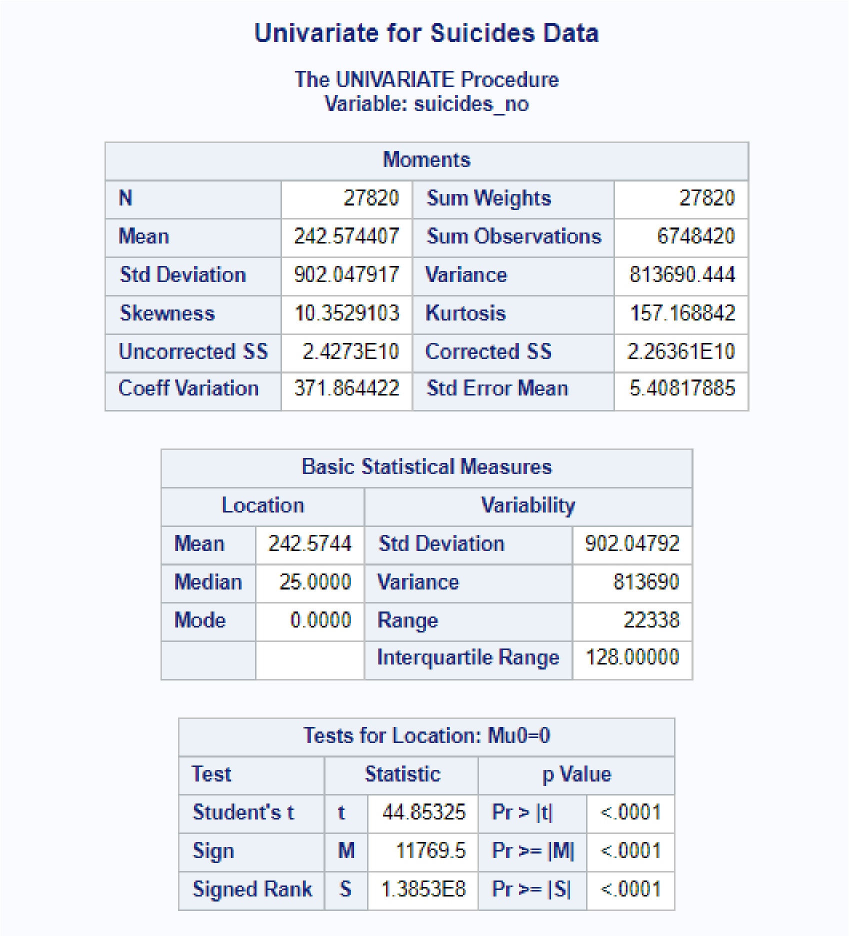
Proc Means for suicides

The MEANS Procedure

Analysis Variable : suicides_no								
Minimum	Maximum	Range	Mean	Std Dev	Lower Quartile	Upper Quartile	Quartile Range	N
0	22338.00	22338.00	242.5744069	902.0479168	3.0000000	131.0000000	128.0000000	27820



SG Plot for Suicides_no



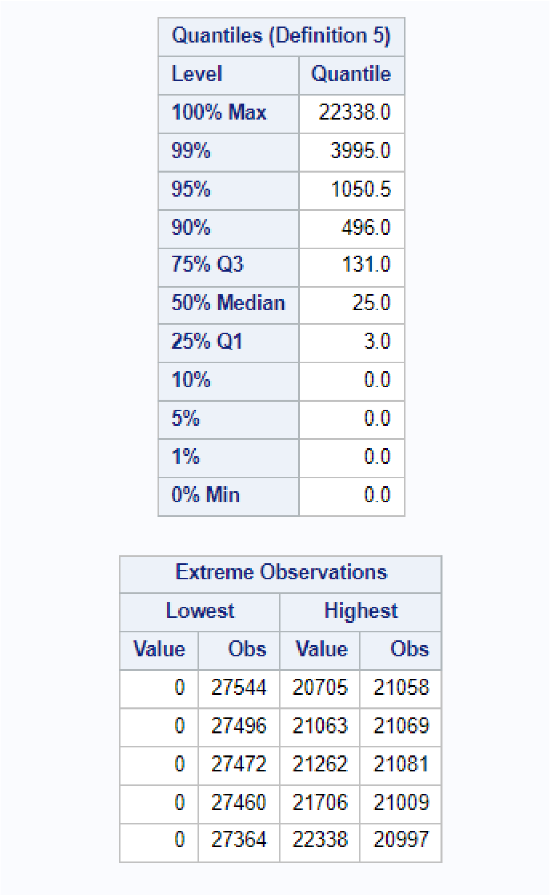
Univariate for Suicides Data

The UNIVARIATE Procedure
Variable: suicides_no

Moments			
N	27820	Sum Weights	27820
Mean	242.574407	Sum Observations	6748420
Std Deviation	902.047917	Variance	813690.444
Skewness	10.3529103	Kurtosis	157.168842
Uncorrected SS	2.4273E10	Corrected SS	2.26361E10
Coeff Variation	371.864422	Std Error Mean	5.40817885

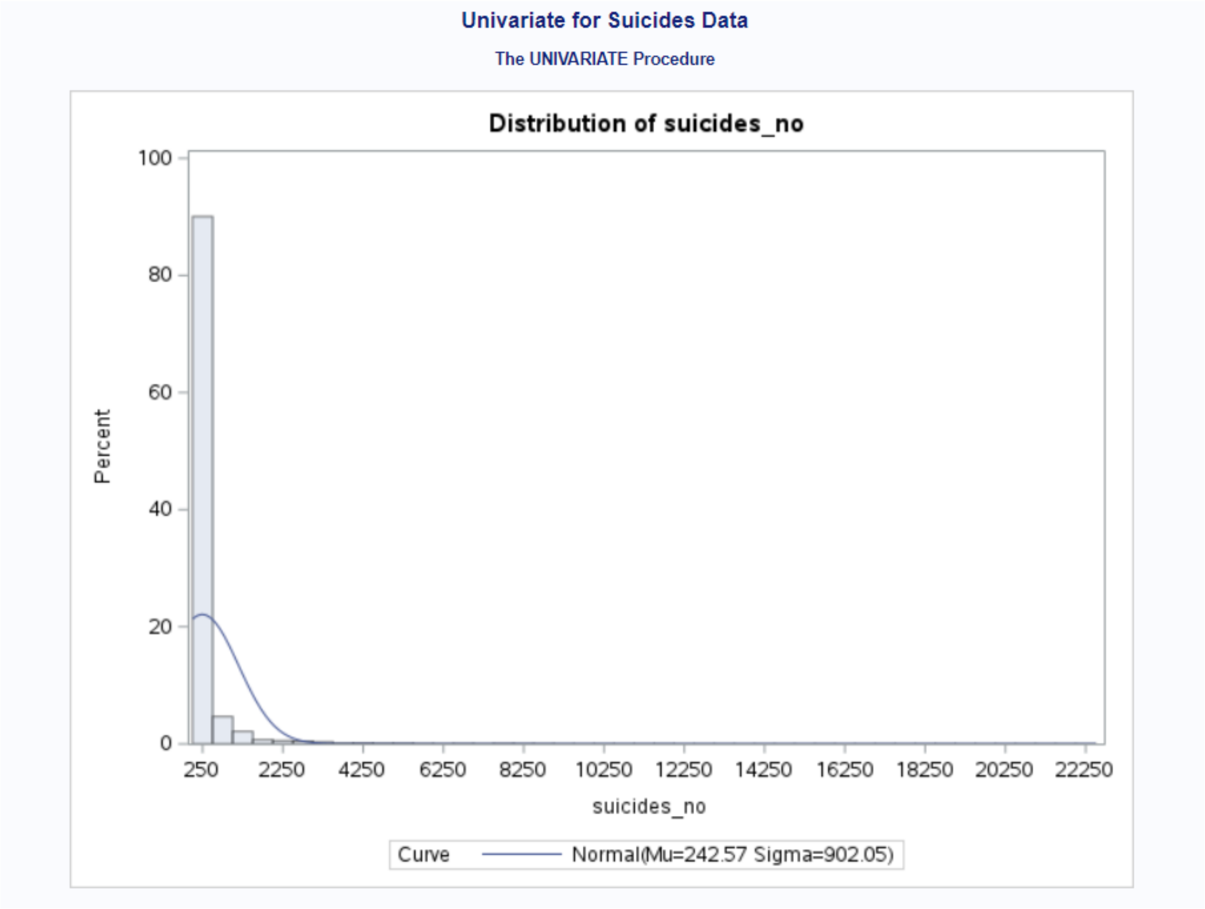
Basic Statistical Measures			
Location		Variability	
Mean	242.5744	Std Deviation	902.04792
Median	25.0000	Variance	813690
Mode	0.0000	Range	22338
		Interquartile Range	128.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	44.85325	Pr > |t|	<.0001
Sign	M	11769.5	Pr >= |M|	<.0001
Signed Rank	S	1.3853E8	Pr >= |S|	<.0001



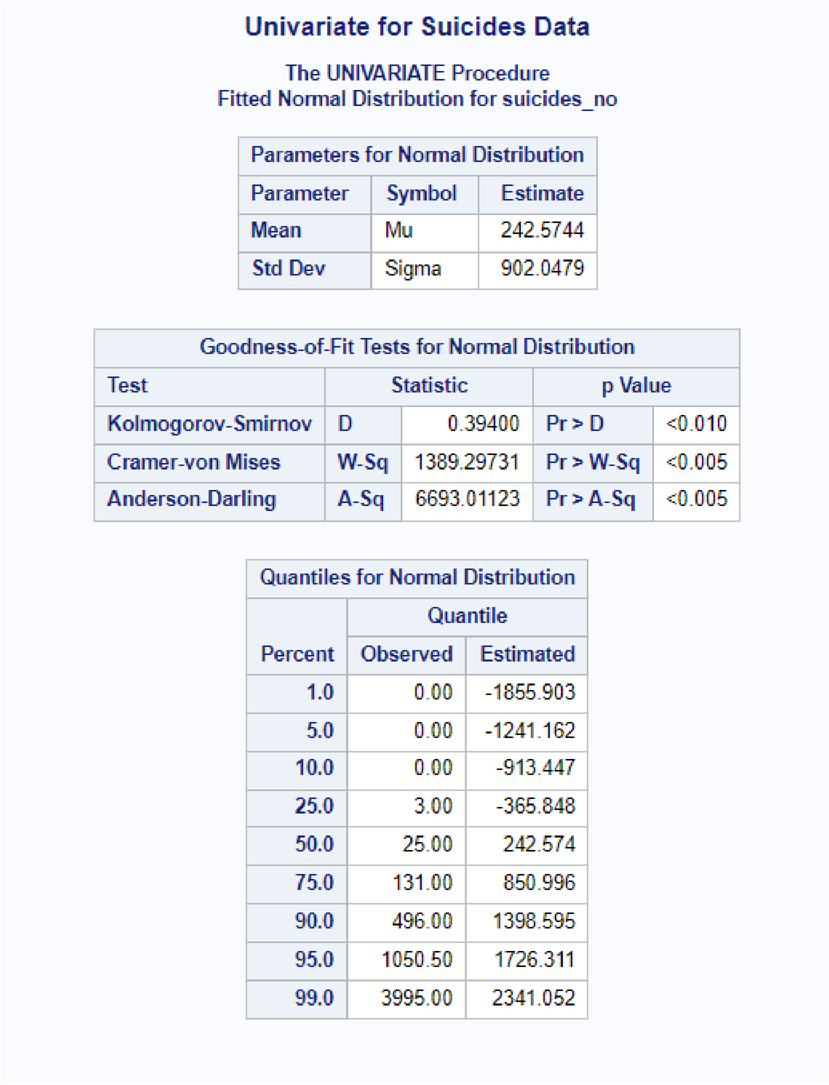
Quantiles (Definition 5)	
Level	Quantile
100% Max	22338.0
99%	3995.0
95%	1050.5
90%	496.0
75% Q3	131.0
50% Median	25.0
25% Q1	3.0
10%	0.0
5%	0.0
1%	0.0
0% Min	0.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	27544	20705	21058
0	27496	21063	21069
0	27472	21262	21081
0	27460	21706	21009
0	27364	22338	20997



Univariate for Suicides Data

The UNIVARIATE Procedure



Univariate for Suicides Data

The UNIVARIATE Procedure
Fitted Normal Distribution for suicides_no

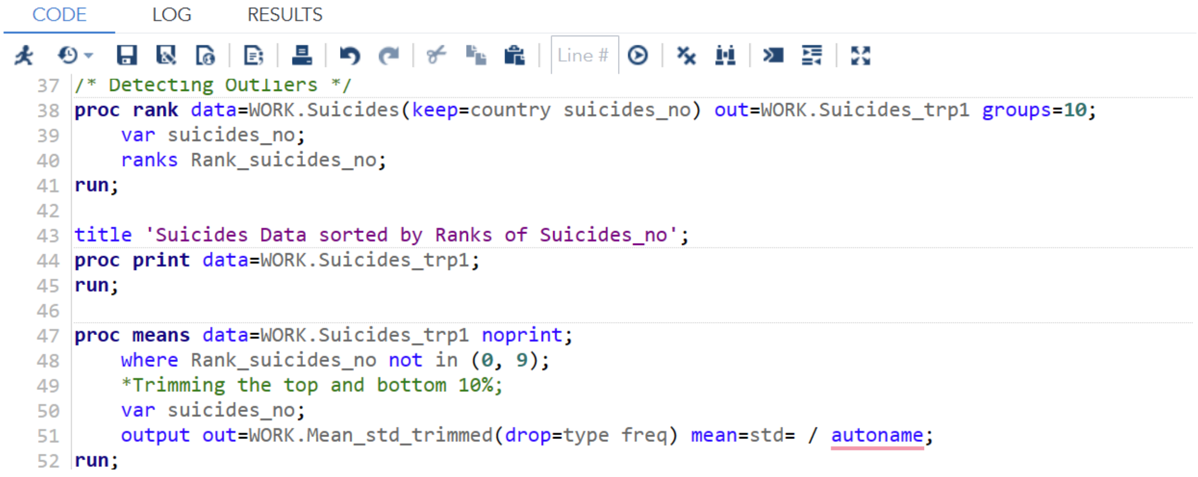
Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	242.5744
Std Dev	Sigma	902.0479

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.39400	Pr > D	<0.010
Cramer-von Mises	W-Sq	1389.29731	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	6693.01123	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	0.00	-1855.903
5.0	0.00	-1241.162
10.0	0.00	-913.447
25.0	3.00	-365.848
50.0	25.00	242.574
75.0	131.00	850.996
90.0	496.00	1398.595
95.0	1050.50	1726.311
99.0	3995.00	2341.052

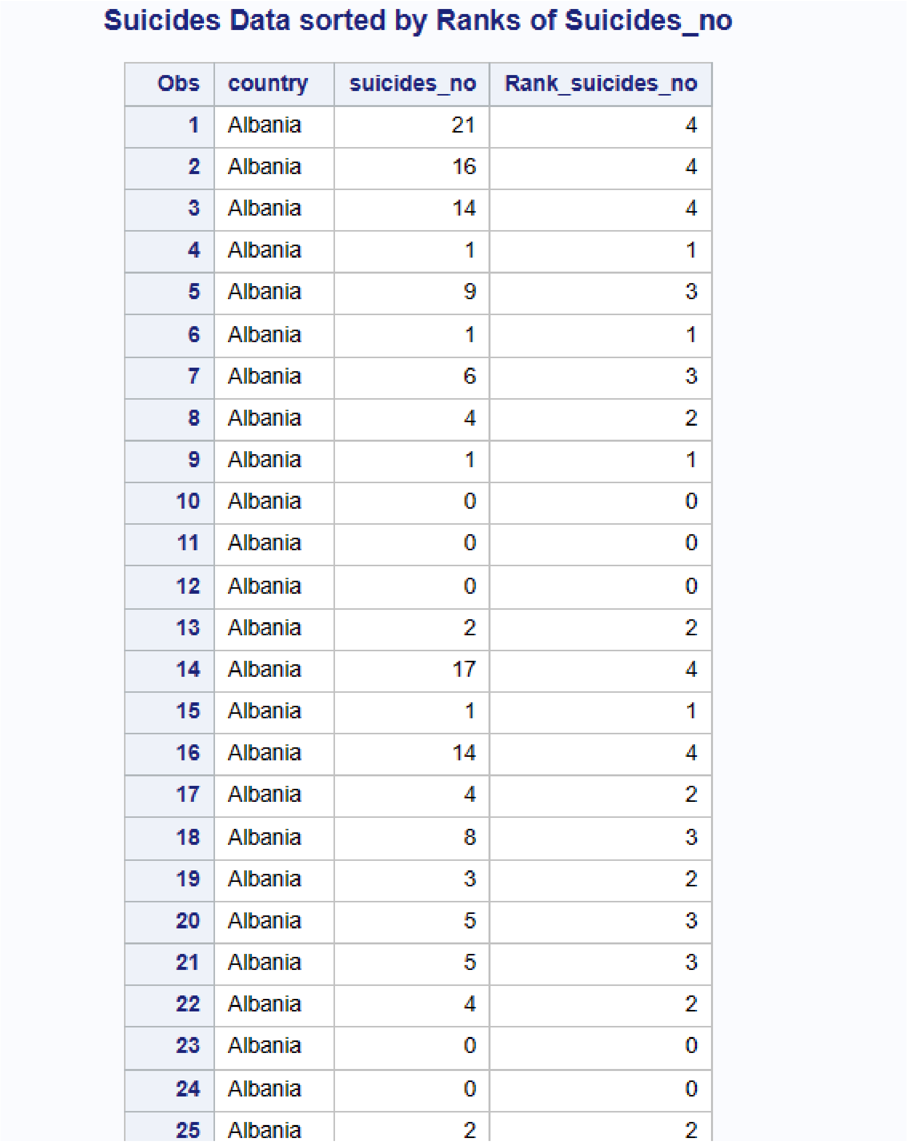
Detecting Outliers:

SAS Code:



```
/* Detecting Outliers */
proc rank data=WORK.Suicides(keep=country suicides_no) out=WORK.Suicides_trp1 groups=10;
    var suicides_no;
    ranks Rank_suicides_no;
run;

title 'Suicides Data sorted by Ranks of Suicides_no';
proc print data=WORK.Suicides_trp1;
run;

proc means data=WORK.Suicides_trp1 noprint;
    where Rank_suicides_no not in (0, 9);
    *Trimming the top and bottom 10%;
    var suicides_no;
    output out=WORK.Mean_std_trimmed(drop=type freq) mean=std= / autoname;
run;
```

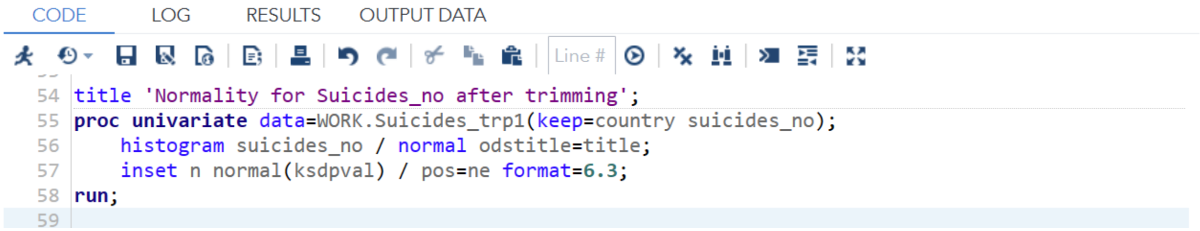
Output:



Suicides Data sorted by Ranks of Suicides_no

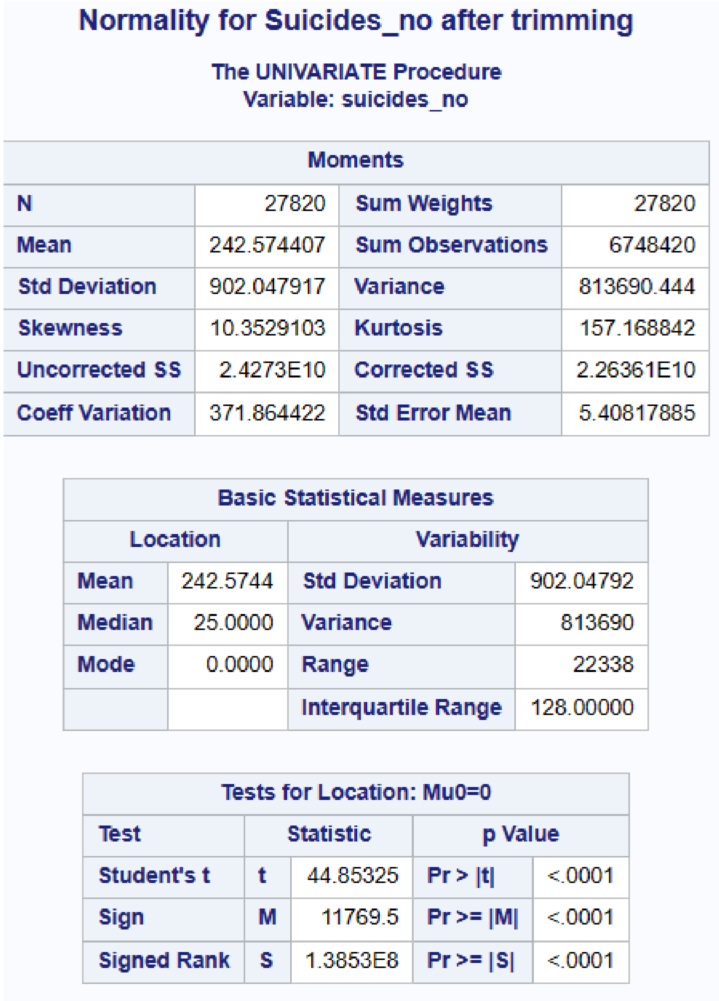
Obs	country	suicides_no	Rank_suicides_no
1	Albania	21	4
2	Albania	16	4
3	Albania	14	4
4	Albania	1	1
5	Albania	9	3
6	Albania	1	1
7	Albania	6	3
8	Albania	4	2
9	Albania	1	1
10	Albania	0	0
11	Albania	0	0
12	Albania	0	0
13	Albania	2	2
14	Albania	17	4
15	Albania	1	1
16	Albania	14	4
17	Albania	4	2
18	Albania	8	3
19	Albania	3	2
20	Albania	5	3
21	Albania	5	3
22	Albania	4	2
23	Albania	0	0
24	Albania	0	0
25	Albania	2	2

SAS Code:



```
title 'Normality for Suicides_no after trimming';
proc univariate data=WORK.Suicides_trp1(keep=country suicides_no);
    histogram suicides_no / normal odstitle=title;
    inset n normal(ksdpval) / pos=ne format=6.3;
run;
```

Output:



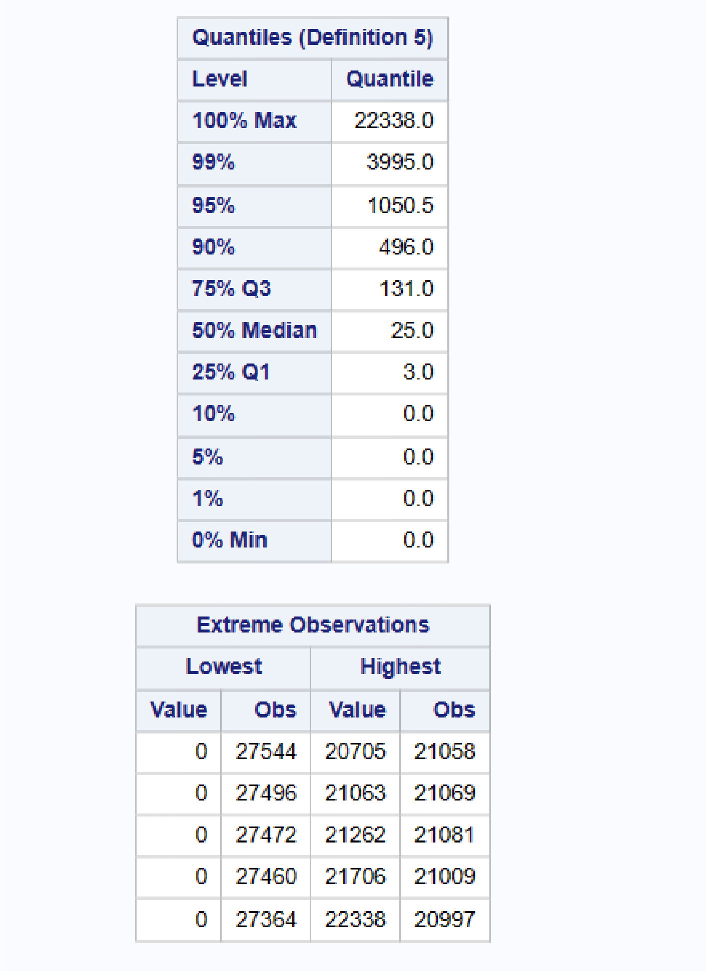
Normality for Suicides_no after trimming

The UNIVARIATE Procedure
Variable: suicides_no

Moments			
N	27820	Sum Weights	27820
Mean	242.574407	Sum Observations	6748420
Std Deviation	902.047917	Variance	813690.444
Skewness	10.3529103	Kurtosis	157.168842
Uncorrected SS	2.4273E10	Corrected SS	2.26361E10
Coeff Variation	371.864422	Std Error Mean	5.40817885

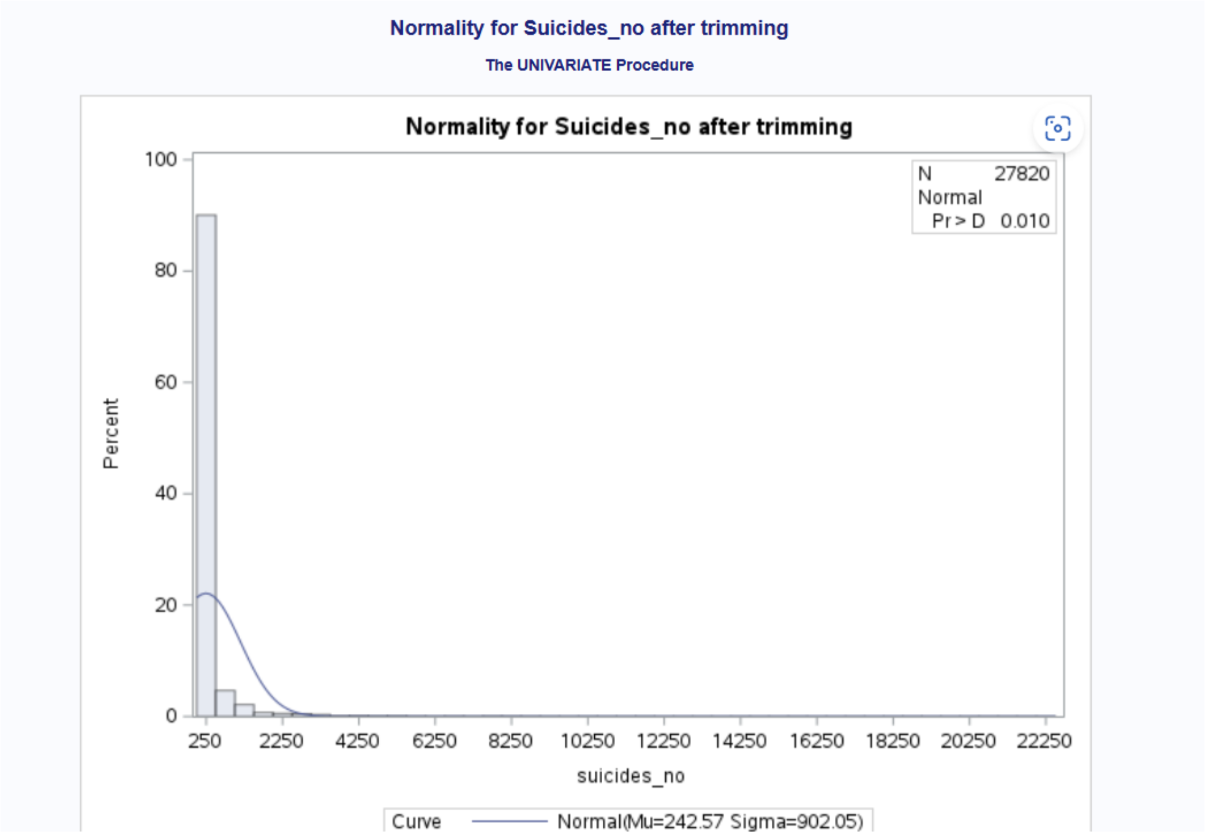
Basic Statistical Measures			
Location		Variability	
Mean	242.5744	Std Deviation	902.04792
Median	25.0000	Variance	813690
Mode	0.0000	Range	22338
		Interquartile Range	128.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	44.85325	Pr > |t|	<.0001
Sign	M	11769.5	Pr >= |M|	<.0001
Signed Rank	S	1.3853E8	Pr >= |S|	<.0001



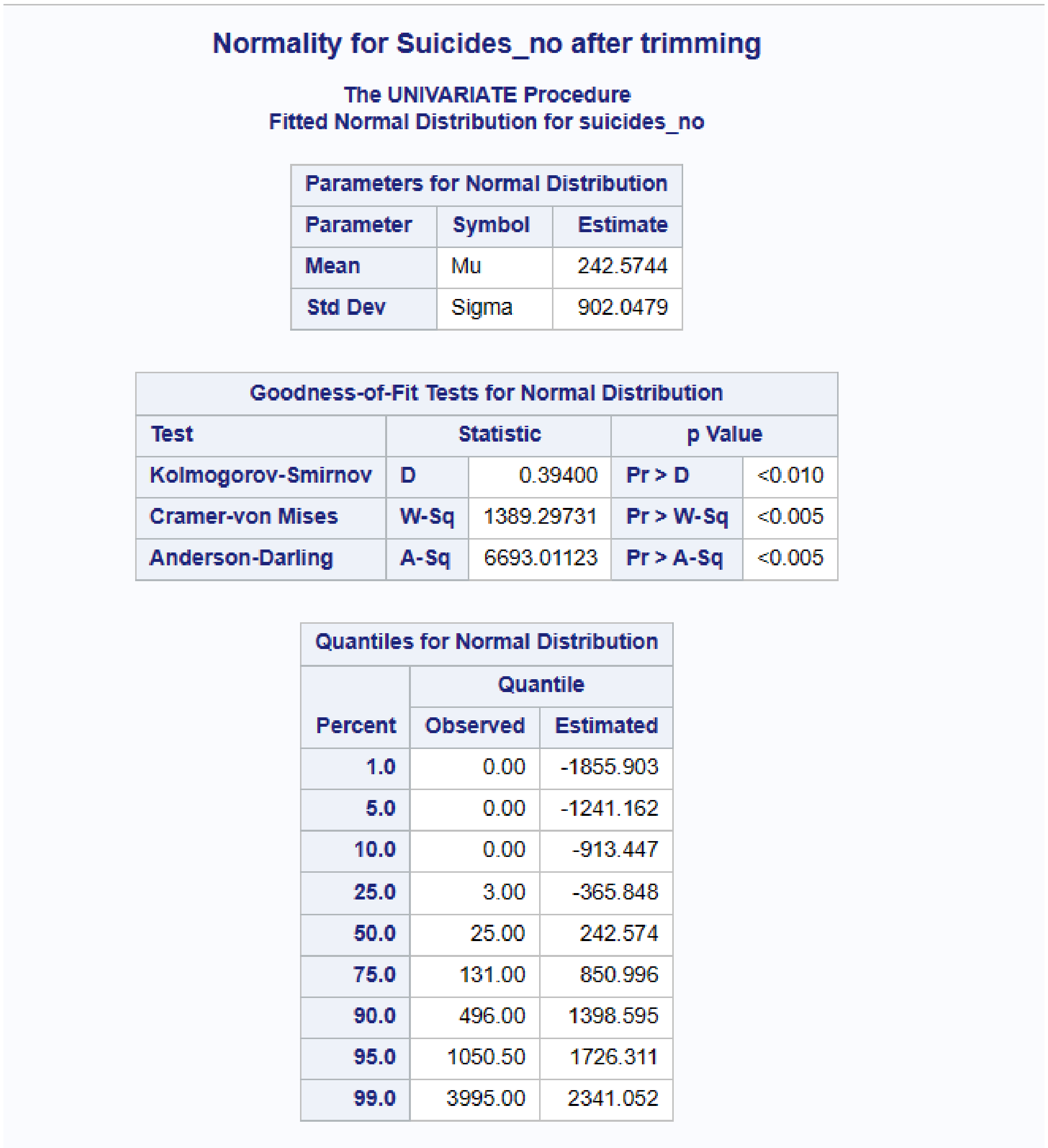
Quantiles (Definition 5)	
Level	Quantile
100% Max	22338.0
99%	3995.0
95%	1050.5
90%	496.0
75% Q3	131.0
50% Median	25.0
25% Q1	3.0
10%	0.0
5%	0.0
1%	0.0
0% Min	0.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	27544	20705	21058
0	27496	21063	21069
0	27472	21262	21081
0	27460	21706	21009
0	27364	22338	20997



Normality for Suicides_no after trimming

The UNIVARIATE Procedure



Normality for Suicides_no after trimming

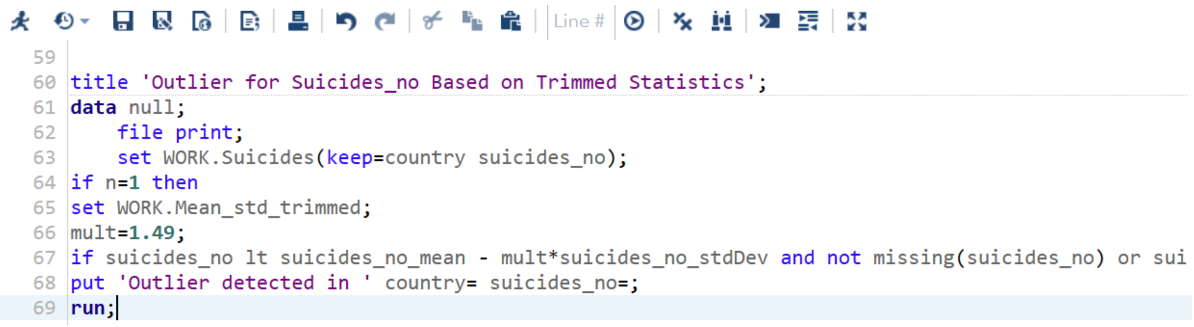
The UNIVARIATE Procedure
Fitted Normal Distribution for suicides_no

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	242.5744
Std Dev	Sigma	902.0479

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.39400	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	1389.29731	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	6693.01123	**Pr > A-Sq**	<0.005

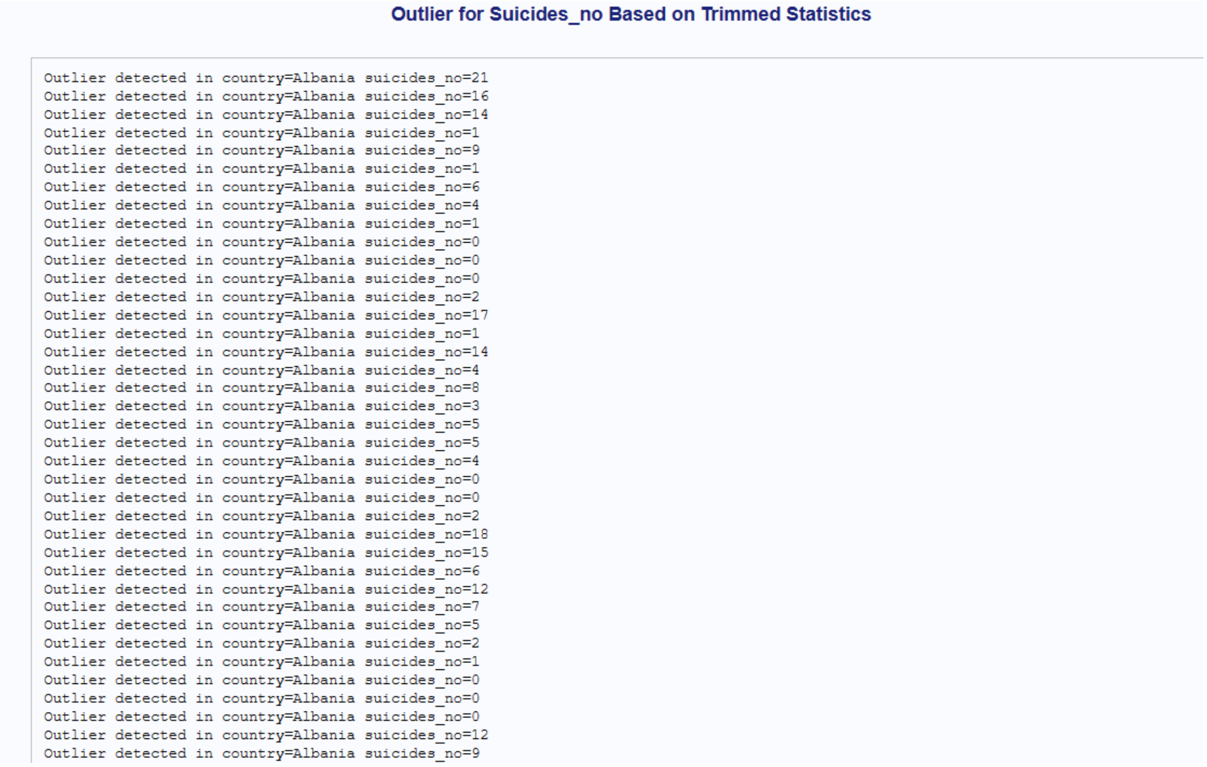
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	0.00	-1855.903
5.0	0.00	-1241.162
10.0	0.00	-913.447
25.0	3.00	-365.848
50.0	25.00	242.574
75.0	131.00	850.996
90.0	496.00	1398.595
95.0	1050.50	1726.311
99.0	3995.00	2341.052

SAS Code:



```

title 'Outlier for Suicides_no Based on Trimmed Statistics';
data null;
    file print;
    set WORK.Suicides(keep=country suicides_no);
if n=1 then
set WORK.Mean_std_trimmed;
mult=1.49;
if suicides_no lt suicides_no_mean - mult*suicides_no_stdDev and not missing(suicides_no) or sui
put 'Outlier detected in ' country= suicides_no=;
run;
```

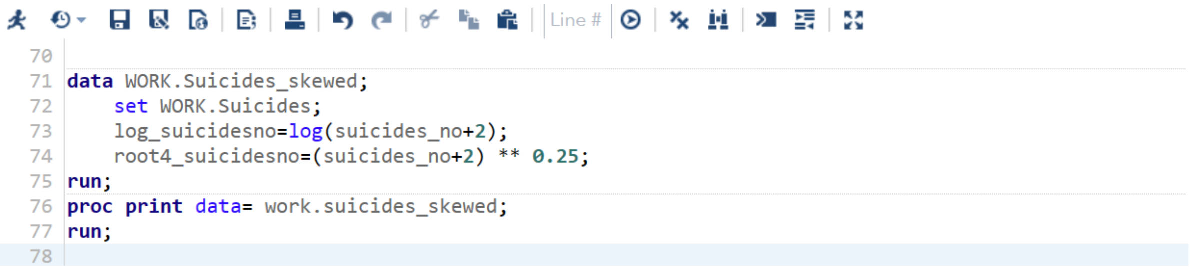
Output:



Outlier for Suicides_no Based on Trimmed Statistics

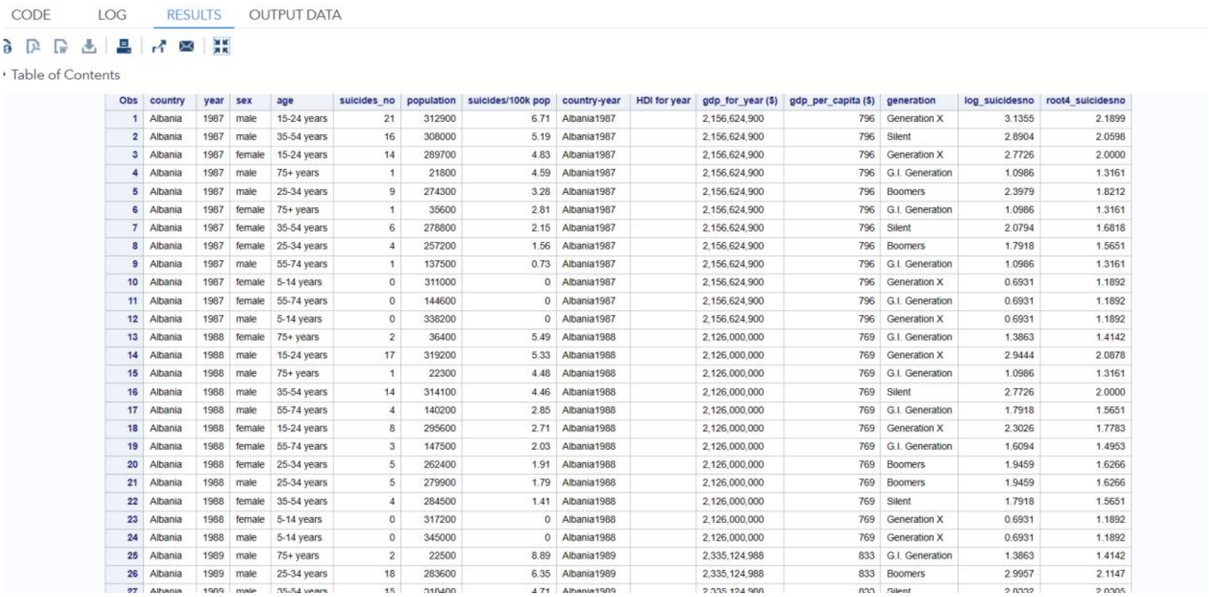
```
Outlier detected in country=Albania suicides_no=21
Outlier detected in country=Albania suicides_no=16
Outlier detected in country=Albania suicides_no=14
Outlier detected in country=Albania suicides_no=1
Outlier detected in country=Albania suicides_no=9
Outlier detected in country=Albania suicides_no=1
Outlier detected in country=Albania suicides_no=6
Outlier detected in country=Albania suicides_no=4
Outlier detected in country=Albania suicides_no=1
Outlier detected in country=Albania suicides_no=0
Outlier detected in country=Albania suicides_no=0
Outlier detected in country=Albania suicides_no=0
Outlier detected in country=Albania suicides_no=2
Outlier detected in country=Albania suicides_no=17
Outlier detected in country=Albania suicides_no=1
Outlier detected in country=Albania suicides_no=14
Outlier detected in country=Albania suicides_no=4
Outlier detected in country=Albania suicides_no=8
Outlier detected in country=Albania suicides_no=3
Outlier detected in country=Albania suicides_no=5
Outlier detected in country=Albania suicides_no=5
Outlier detected in country=Albania suicides_no=4
Outlier detected in country=Albania suicides_no=0
Outlier detected in country=Albania suicides_no=0
Outlier detected in country=Albania suicides_no=2
Outlier detected in country=Albania suicides_no=18
Outlier detected in country=Albania suicides_no=15
Outlier detected in country=Albania suicides_no=6
Outlier detected in country=Albania suicides_no=12
Outlier detected in country=Albania suicides_no=7
Outlier detected in country=Albania suicides_no=5
Outlier detected in country=Albania suicides_no=2
Outlier detected in country=Albania suicides_no=1
Outlier detected in country=Albania suicides_no=0
Outlier detected in country=Albania suicides_no=0
Outlier detected in country=Albania suicides_no=0
Outlier detected in country=Albania suicides_no=12
Outlier detected in country=Albania suicides_no=9
```

SAS Code:



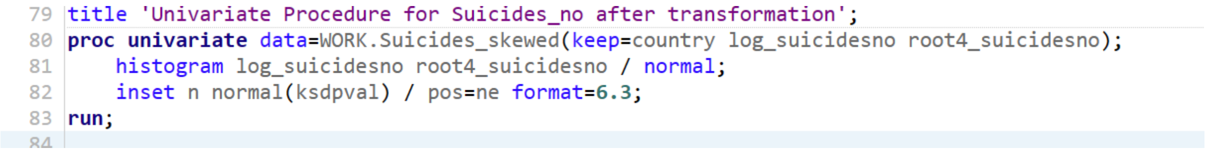
```

data WORK.Suicides_skewed;
    set WORK.Suicides;
    log_suicidesno=log(suicides_no+2);
    root4_suicidesno=(suicides_no+2) ** 0.25;
run;
proc print data= work.suicides_skewed;
run;

```

Output:



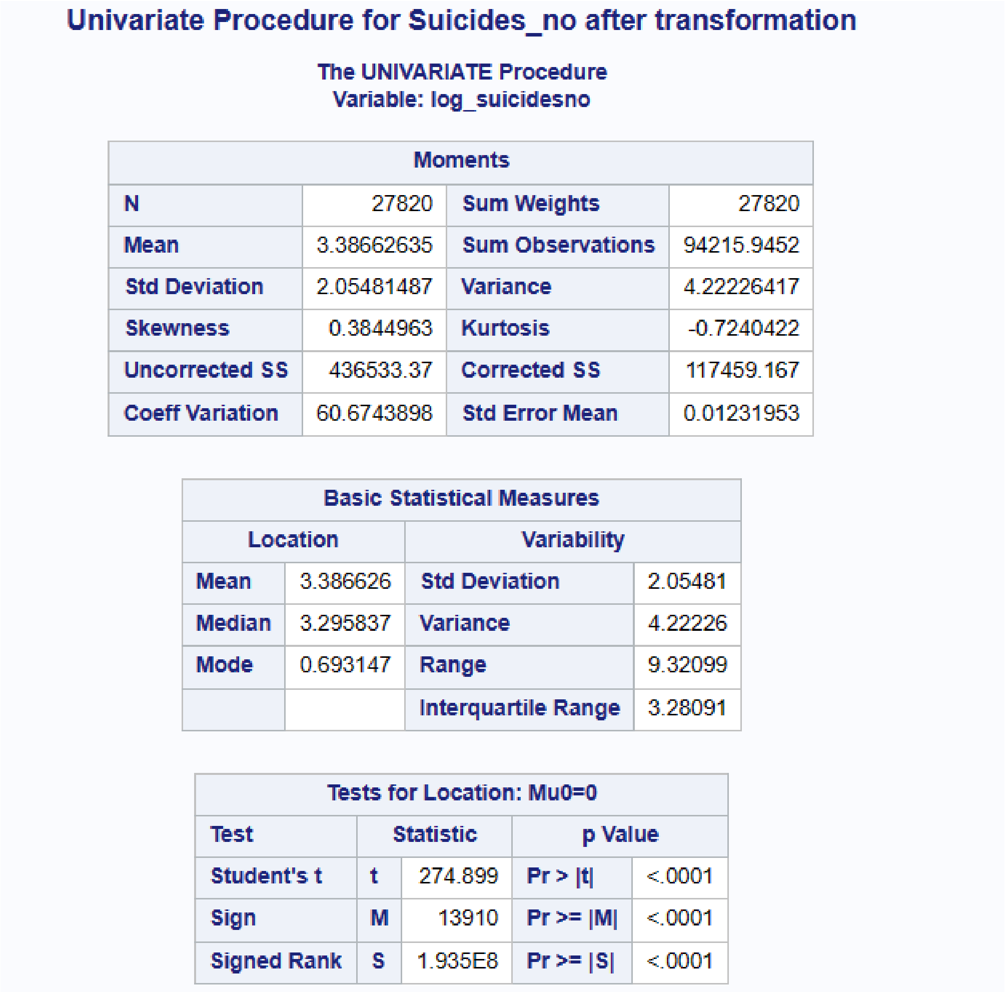
Obs	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation	log_suicidesno	root4_suicidesno
1	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987		2,156,624,900	796	Generation X	3.1355	2.1899
2	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987		2,156,624,900	796	Silent	2.8904	2.0598
3	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987		2,156,624,900	796	Generation X	2.7726	2.0000
4	Albania	1987	male	75+ years	1	21800	4.59	Albania1987		2,156,624,900	796	G.I. Generation	1.0986	1.3161
5	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987		2,156,624,900	796	Boomers	2.3979	1.8212
6	Albania	1987	female	75+ years	1	35600	2.81	Albania1987		2,156,624,900	796	G.I. Generation	1.0986	1.3161
7	Albania	1987	female	35-54 years	6	278800	2.15	Albania1987		2,156,624,900	796	Silent	2.0794	1.6818
8	Albania	1987	female	25-34 years	4	257200	1.56	Albania1987		2,156,624,900	796	Boomers	1.7918	1.5651
9	Albania	1987	male	55-74 years	1	137500	0.73	Albania1987		2,156,624,900	796	G.I. Generation	1.0986	1.3161
10	Albania	1987	female	5-14 years	0	311000	0	Albania1987		2,156,624,900	796	Generation X	0.6931	1.1892
11	Albania	1987	female	55-74 years	0	144600	0	Albania1987		2,156,624,900	796	G.I. Generation	0.6931	1.1892
12	Albania	1987	male	5-14 years	0	338200	0	Albania1987		2,156,624,900	796	Generation X	0.6931	1.1892
13	Albania	1988	female	75+ years	2	36400	5.49	Albania1988		2,126,000,000	769	G.I. Generation	1.3863	1.4142
14	Albania	1988	male	15-24 years	17	319200	5.33	Albania1988		2,126,000,000	769	Generation X	2.9444	2.0878
15	Albania	1988	male	75+ years	1	22300	4.48	Albania1988		2,126,000,000	769	G.I. Generation	1.0986	1.3161
16	Albania	1988	male	35-54 years	14	314100	4.46	Albania1988		2,126,000,000	769	Silent	2.7726	2.0000
17	Albania	1988	male	55-74 years	4	140200	2.85	Albania1988		2,126,000,000	769	G.I. Generation	1.7918	1.5651
18	Albania	1988	female	15-24 years	8	295600	2.71	Albania1988		2,126,000,000	769	Generation X	2.3026	1.7783
19	Albania	1988	female	55-74 years	3	147500	2.03	Albania1988		2,126,000,000	769	G.I. Generation	1.6094	1.4953
20	Albania	1988	female	25-34 years	5	262400	1.91	Albania1988		2,126,000,000	769	Boomers	1.9459	1.6266
21	Albania	1988	male	25-34 years	5	279900	1.79	Albania1988		2,126,000,000	769	Boomers	1.9459	1.6266
22	Albania	1988	female	35-54 years	4	284500	1.41	Albania1988		2,126,000,000	769	Silent	1.7918	1.5651
23	Albania	1988	female	5-14 years	0	317200	0	Albania1988		2,126,000,000	769	Generation X	0.6931	1.1892
24	Albania	1988	male	5-14 years	0	345000	0	Albania1988		2,126,000,000	769	Generation X	0.6931	1.1892
25	Albania	1989	male	75+ years	2	22500	8.89	Albania1989		2,335,124,988	833	G.I. Generation	1.3863	1.4142
26	Albania	1989	male	25-34 years	18	283600	6.35	Albania1989		2,335,124,988	833	Boomers	2.9957	2.1147
27	Albania	1989	male	35-54 years	15	310400	4.71	Albania1989		2,335,124,988	833	Silent	2.8332	2.0305

SAS Code:



```
title 'Univariate Procedure for Suicides_no after transformation';
proc univariate data=WORK.Suicides_skewed(keep=country log_suicidesno root4_suicidesno);
    histogram log_suicidesno root4_suicidesno / normal;
    inset n normal(ksdpval) / pos=ne format=6.3;
run;
```

Output:



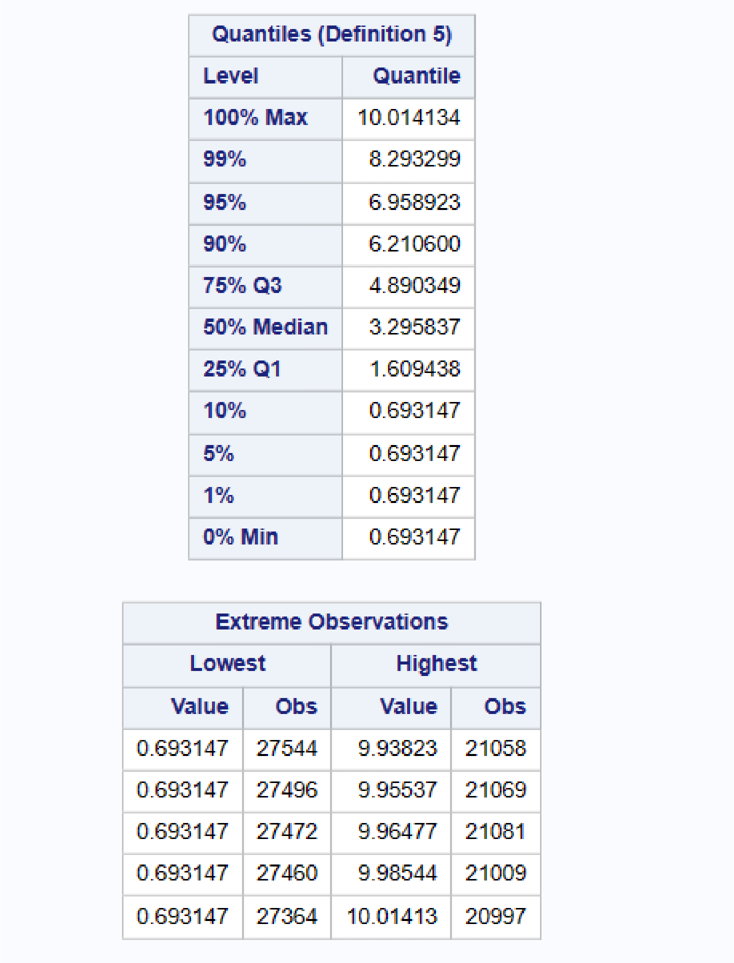
Univariate Procedure for Suicides_no after transformation

The UNIVARIATE Procedure
Variable: log_suicidesno

Moments			
N	27820	Sum Weights	27820
Mean	3.38662635	Sum Observations	94215.9452
Std Deviation	2.05481487	Variance	4.22226417
Skewness	0.3844963	Kurtosis	-0.7240422
Uncorrected SS	436533.37	Corrected SS	117459.167
Coeff Variation	60.6743898	Std Error Mean	0.01231953

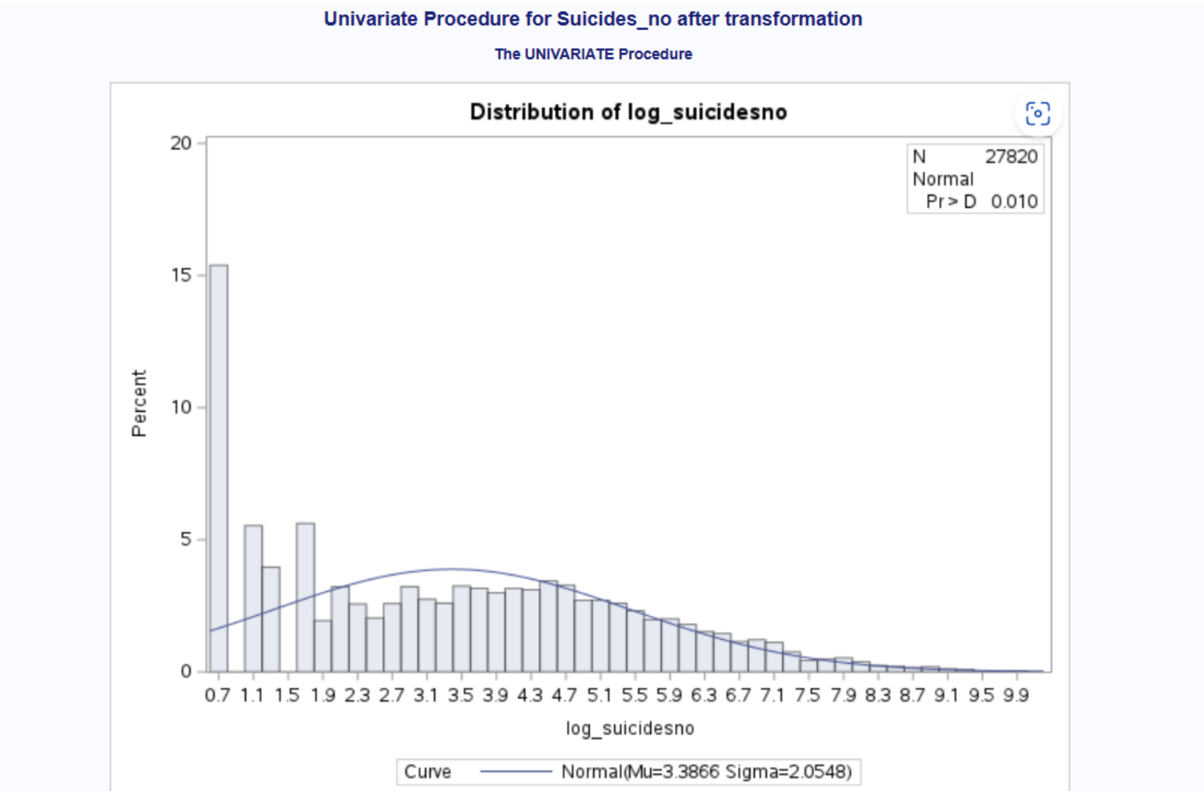
Basic Statistical Measures			
Location		Variability	
Mean	3.386626	Std Deviation	2.05481
Median	3.295837	Variance	4.22226
Mode	0.693147	Range	9.32099
		Interquartile Range	3.28091

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	274.899	Pr > |t|	<.0001
Sign	M	13910	Pr >= |M|	<.0001
Signed Rank	S	1.935E8	Pr >= |S|	<.0001



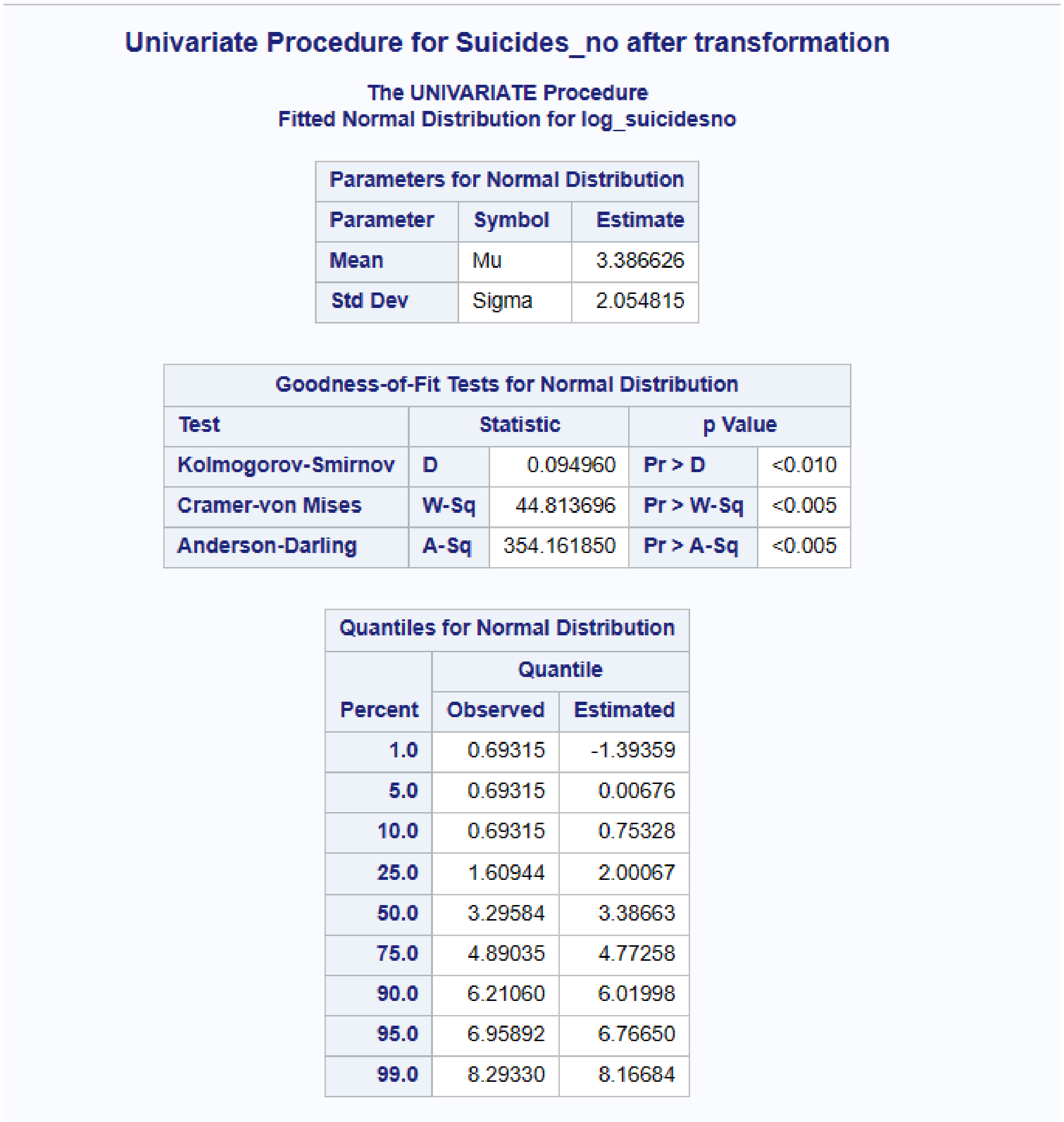
Quantiles (Definition 5)	
Level	Quantile
100% Max	10.014134
99%	8.293299
95%	6.958923
90%	6.210600
75% Q3	4.890349
50% Median	3.295837
25% Q1	1.609438
10%	0.693147
5%	0.693147
1%	0.693147
0% Min	0.693147

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.693147	27544	9.93823	21058
0.693147	27496	9.95537	21069
0.693147	27472	9.96477	21081
0.693147	27460	9.98544	21009
0.693147	27364	10.01413	20997



Univariate Procedure for Suicides_no after transformation

The UNIVARIATE Procedure



Univariate Procedure for Suicides_no after transformation

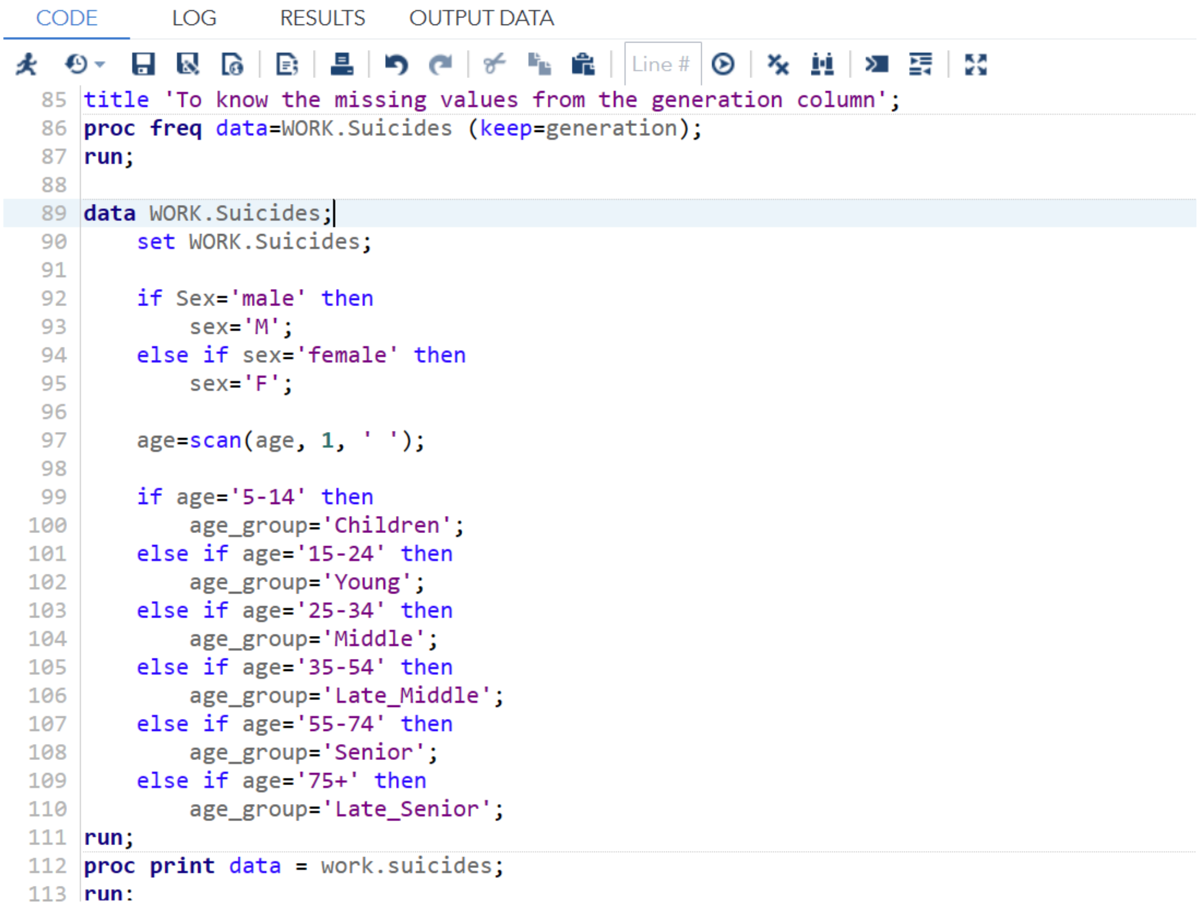
The UNIVARIATE Procedure
Fitted Normal Distribution for log_suicidesno

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	3.386626
Std Dev	Sigma	2.054815

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.094960	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	44.813696	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	354.161850	**Pr > A-Sq**	<0.005

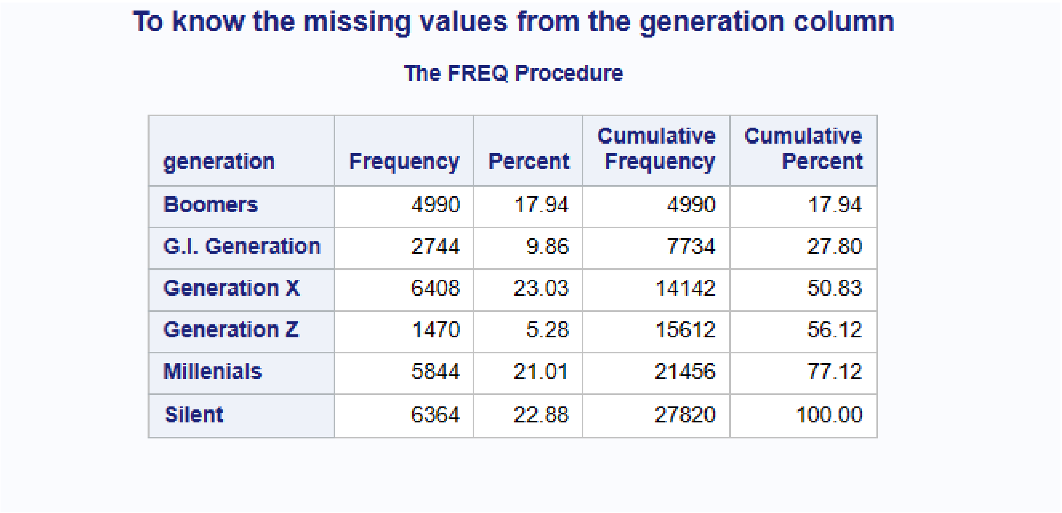
Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	0.69315	-1.39359
5.0	0.69315	0.00676
10.0	0.69315	0.75328
25.0	1.60944	2.00067
50.0	3.29584	3.38663
75.0	4.89035	4.77258
90.0	6.21060	6.01998
95.0	6.95892	6.76650
99.0	8.29330	8.16684

SAS Code:



```
title 'To know the missing values from the generation column';
proc freq data=WORK.Suicides (keep=generation);
run;

data WORK.Suicides;
    set WORK.Suicides;

    if Sex='male' then
        sex='M';
    else if sex='female' then
        sex='F';

    age=scan(age, 1, ' ');

    if age='5-14' then
        age_group='Children';
    else if age='15-24' then
        age_group='Young';
    else if age='25-34' then
        age_group='Middle';
    else if age='35-54' then
        age_group='Late_Middle';
    else if age='55-74' then
        age_group='Senior';
    else if age='75+' then
        age_group='Late_Senior';
run;
proc print data = work.suicides;
run;
```

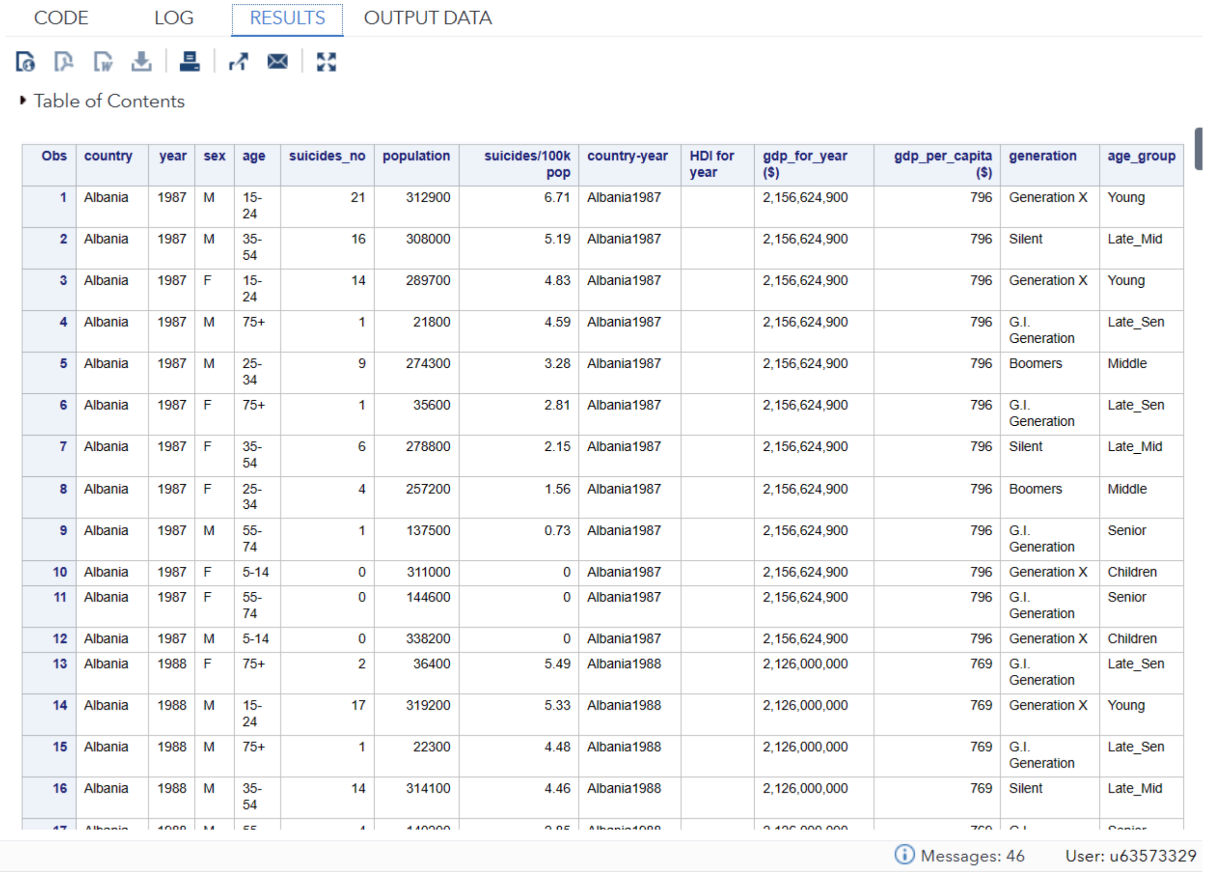
Output:



To know the missing values from the generation column

The FREQ Procedure

generation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Boomers	4990	17.94	4990	17.94
G.I. Generation	2744	9.86	7734	27.80
Generation X	6408	23.03	14142	50.83
Generation Z	1470	5.28	15612	56.12
Millenials	5844	21.01	21456	77.12
Silent	6364	22.88	27820	100.00



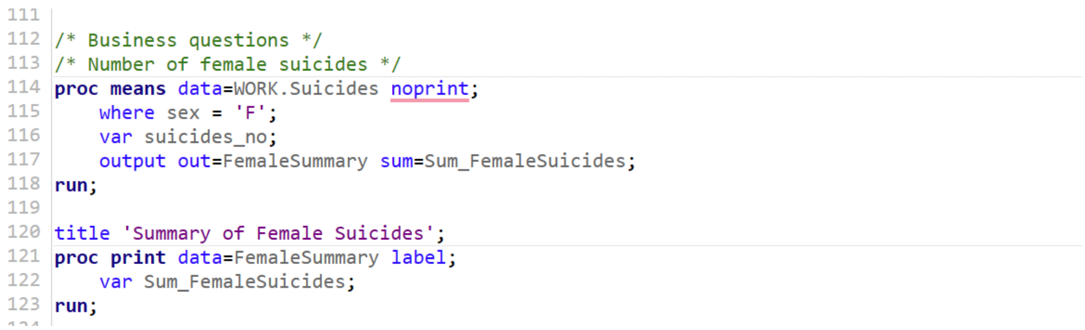
Obs	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation	age_group
1	Albania	1987	M	15-24	21	312900	6.71	Albania1987		2,156,624,900	796	Generation X	Young
2	Albania	1987	M	35-54	16	308000	5.19	Albania1987		2,156,624,900	796	Silent	Late_Mid
3	Albania	1987	F	15-24	14	289700	4.83	Albania1987		2,156,624,900	796	Generation X	Young
4	Albania	1987	M	75+	1	21800	4.59	Albania1987		2,156,624,900	796	G.I. Generation	Late_Sen
5	Albania	1987	M	25-34	9	274300	3.28	Albania1987		2,156,624,900	796	Boomers	Middle
6	Albania	1987	F	75+	1	35600	2.81	Albania1987		2,156,624,900	796	G.I. Generation	Late_Sen
7	Albania	1987	F	35-54	6	278800	2.15	Albania1987		2,156,624,900	796	Silent	Late_Mid
8	Albania	1987	F	25-34	4	257200	1.56	Albania1987		2,156,624,900	796	Boomers	Middle
9	Albania	1987	M	55-74	1	137500	0.73	Albania1987		2,156,624,900	796	G.I. Generation	Senior
10	Albania	1987	F	5-14	0	311000	0	Albania1987		2,156,624,900	796	Generation X	Children
11	Albania	1987	F	55-74	0	144600	0	Albania1987		2,156,624,900	796	G.I. Generation	Senior
12	Albania	1987	M	5-14	0	338200	0	Albania1987		2,156,624,900	796	Generation X	Children
13	Albania	1988	F	75+	2	36400	5.49	Albania1988		2,126,000,000	769	G.I. Generation	Late_Sen
14	Albania	1988	M	15-24	17	319200	5.33	Albania1988		2,126,000,000	769	Generation X	Young
15	Albania	1988	M	75+	1	22300	4.48	Albania1988		2,126,000,000	769	G.I. Generation	Late_Sen
16	Albania	1988	M	35-54	14	314100	4.46	Albania1988		2,126,000,000	769	Silent	Late_Mid
17	Albania	1988	M	55-	4	140200	2.85	Albania1988		2,126,000,000	769	G.I.	Senior

Messages: 46 User: u63573329

Data Analysis and Business Question:

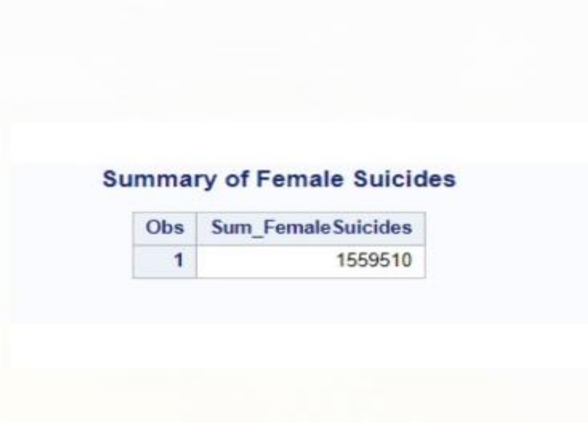
This section involves transforming the 'suicides_no' variable through logarithmic and root transformations, addressing missing values in the 'generation' column, and conducting analyses to answer specific business questions. Business questions include summarizing female and male suicides separately, as well as identifying the year with the highest and lowest suicide rates. The results of these analyses provide valuable insights into the dataset and support decision-making processes.

SAS Code:



```

/* Business questions */
/* Number of female suicides */
proc means data=WORK.Suicides noprint;
    where sex = 'F';
    var suicides_no;
    output out=FemaleSummary sum=Sum_FemaleSuicides;
run;

title 'Summary of Female Suicides';
proc print data=FemaleSummary label;
    var Sum_FemaleSuicides;
run;
```

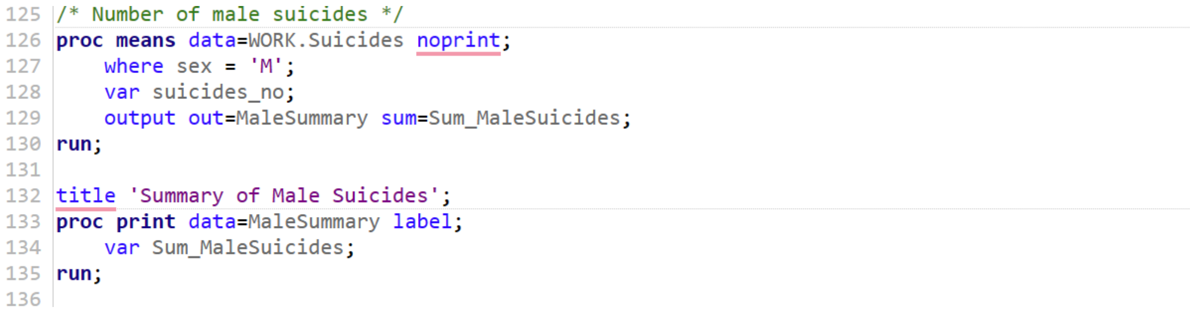
Output:



Summary of Female Suicides

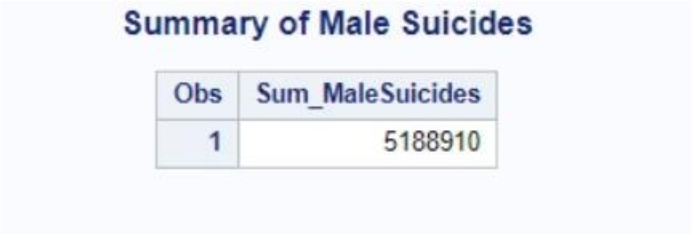
Obs	Sum_FemaleSuicides
1	1559510

SAS Code:



```
/* Number of male suicides */
proc means data=WORK.Suicides noprint;
    where sex = 'M';
    var suicides_no;
    output out=MaleSummary sum=Sum_MaleSuicides;
run;

title 'Summary of Male Suicides';
proc print data=MaleSummary label;
    var Sum_MaleSuicides;
run;

```

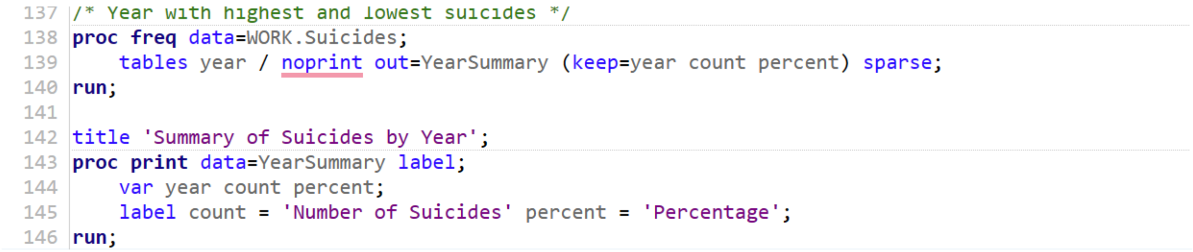
Output:



Summary of Male Suicides

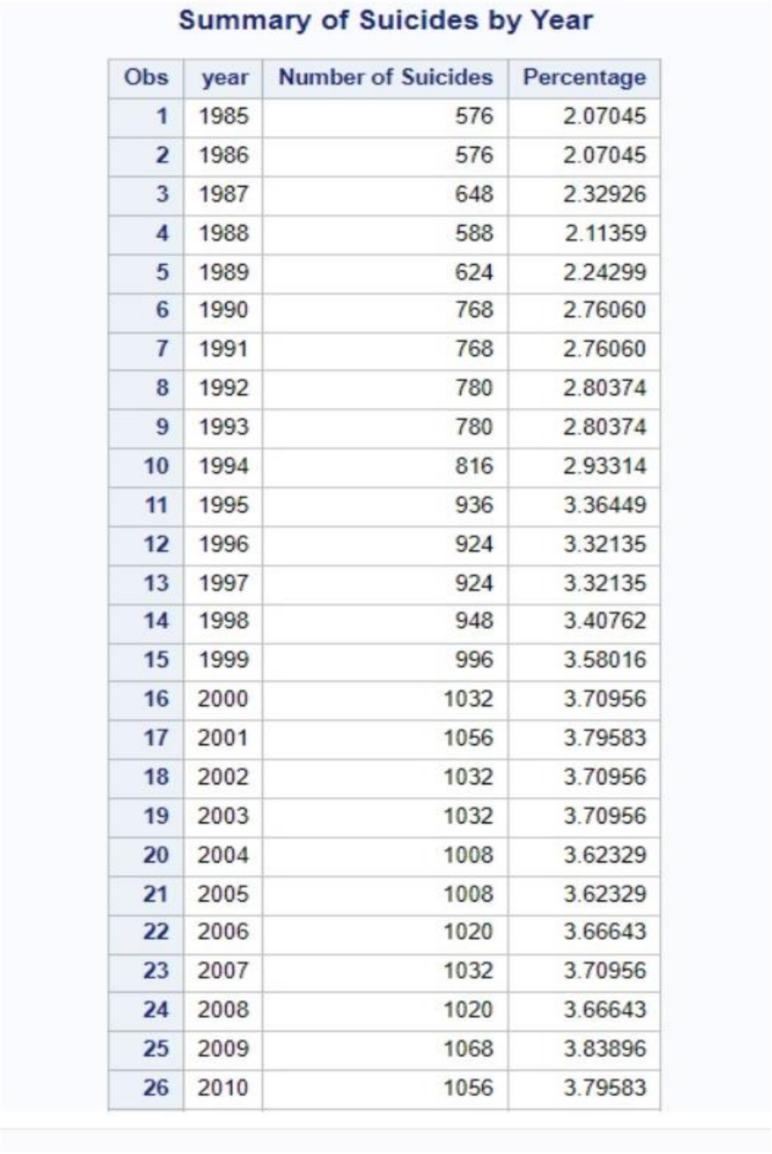
Obs	Sum_MaleSuicides
1	5188910

SAS Code:



```
/* Year with highest and lowest suicides */
proc freq data=WORK.Suicides;
    tables year / noprint out=YearSummary (keep=year count percent) sparse;
run;

title 'Summary of Suicides by Year';
proc print data=YearSummary label;
    var year count percent;
    label count = 'Number of Suicides' percent = 'Percentage';
run;
```

Output:



Summary of Suicides by Year

Obs	year	Number of Suicides	Percentage
1	1985	576	2.07045
2	1986	576	2.07045
3	1987	648	2.32926
4	1988	588	2.11359
5	1989	624	2.24299
6	1990	768	2.76060
7	1991	768	2.76060
8	1992	780	2.80374
9	1993	780	2.80374
10	1994	816	2.93314
11	1995	936	3.36449
12	1996	924	3.32135
13	1997	924	3.32135
14	1998	948	3.40762
15	1999	996	3.58016
16	2000	1032	3.70956
17	2001	1056	3.79583
18	2002	1032	3.70956
19	2003	1032	3.70956
20	2004	1008	3.62329
21	2005	1008	3.62329
22	2006	1020	3.66643
23	2007	1032	3.70956
24	2008	1020	3.66643
25	2009	1068	3.83896
26	2010	1056	3.79583