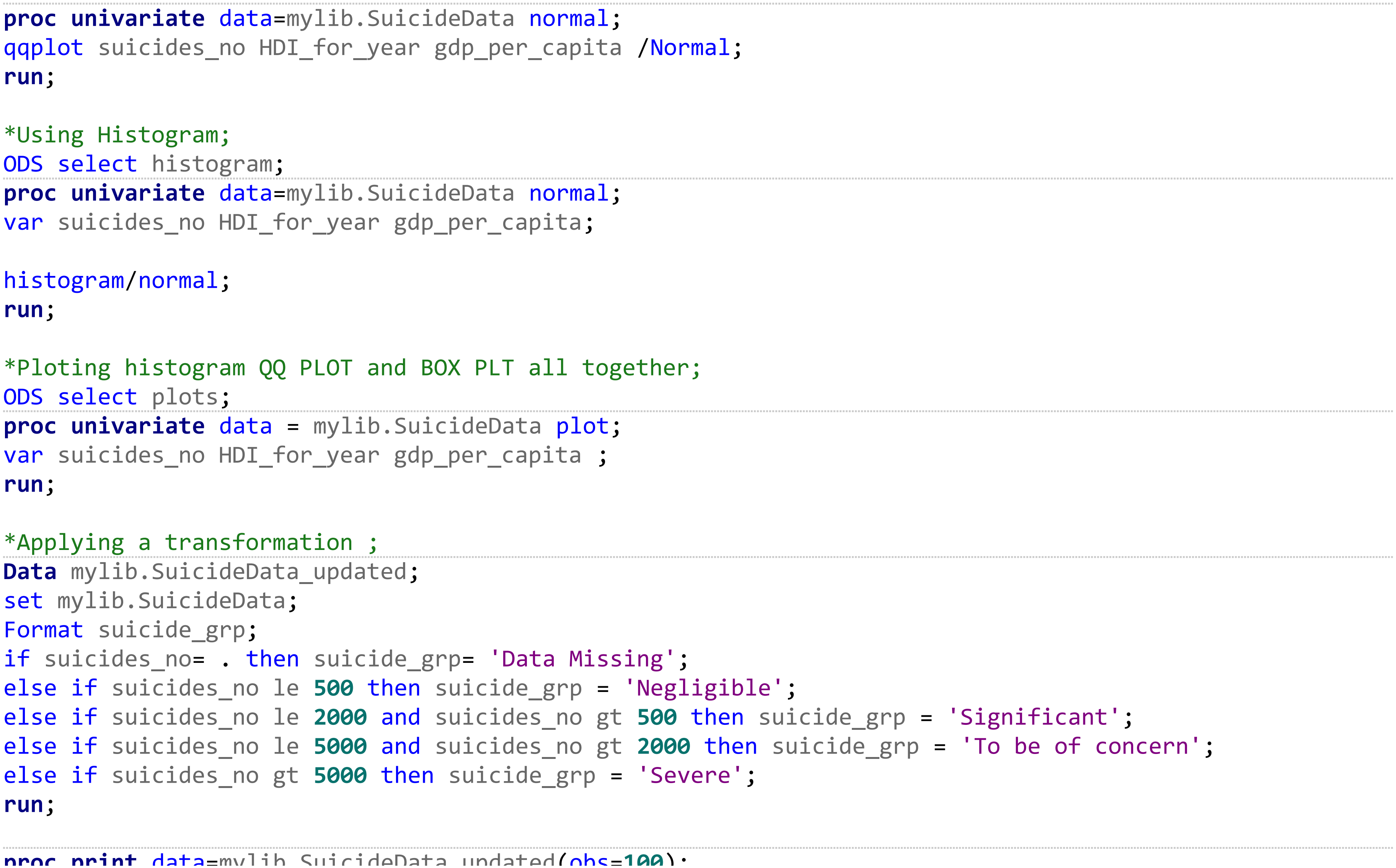
```
*Creating a library;
Libname mylib '/home/u59397413/Assignment110';

*Importing the dataset;
Proc import datafile= '/home/u59397413/Assignment110/SuicideData.xlsx'
out= mylib.SuicideData
dbms=xlsx
replace;
getnames=yes;
run;

PROC PRINT DATA=mylib.SuicideData(obs=10);
RUN;

*Categorical variabales are : Country, SEX, Generation, and Age;
*Check if there are any error in categorical variables using PROC FREQ;
PROC freq data=mylib.SuicideData;
table country sex generation age /nocum nopercent;
run;

*Correcting error in categorical Variable SEX and Generation;
DATA mylib.SuicideData;
set mylib.SuicideData;
IF SEX IN('F','Femail') THEN SEX='female';
ELSE IF SEX IN('Mail','M') THEN SEX='male';
IF generation IN('Gen z','Genn z') THEN generation= 'Generation Z';
run;

PROC FREQ DATA=mylib.SuicideData;
TABLES sex generation /nocum nopercent;
RUN;

*Check for missing values and correcting them;
DATA mylib.SuicideData;
set mylib.SuicideData;
IF SEX= ' ' THEN SEX='N/A';
```

```
run;

PROC FREQ DATA=mylib.SuicideData;
TABLES sex generation age;
RUN;

*Creating a derived variable using proc format;
PROC FORMAT;
Value $age_c '15-24 years'= 'Teen-Young'
'25-34 years'='Young-Middle Thirties'
'35-54 years'='Mid Thirties to Fifties'
'55-74 years'='Fifties to Senior'
'75+ years'= 'Above 75 year';
VALUE $gender 'male'='Male'
'female'='Female';
RUN;

PROC PRINT data=mylib.SuicideData(obs=10);
format sex $gender. age $age_c. ;
RUN;

*Checking the missing values and creating histogram for numerical variables in dataset;
TITLE 'Checking the missing values in numerical variables in dataset';
PROC MEANS DATA= mylib.SuicideData n nmiss mean min max median;
var suicides_no gdp_per_capita HDI_for_year;
RUN;

ods select ExtremeObs Quantiles histogram;
PROC UNIVARIATE DATA=mylib.SuicideData nextrobs=10;
var suicides_no gdp_per_capita HDI_for_year;
histogram/normal;
RUN;
ODS TRACE OFF;

*Using imputation method to treat missing values in numerical dataset;
PROC STDIZE DATA=mylib.SuicideData out=mylib.SuicideData
replace
```

```
method=mean;
var HDI_for_year;
run;

ods select ExtremeObs Quantiles histogram;
PROC UNIVARIATE DATA=mylib.SuicideData nextrobs=10;
var HDI_for_year;
histogram/normal;
RUN;
ODS TRACE OFF;

*Checking errors in data Suicide
This might not be used as I was trying to change data to enter some false values
but was getting repeated error;

data _null_;
file print;
set mylib.SuicideData(keep=Suicides_no gdp_per_capita HDI_for_year);
if notdigit(trimn(Suicides_no)) and not missing (Suicides_no) then
put "invalid value " Suicides_no " for suicide" ;
if notdigit(trimn(gdp_per_capita)) and not missing (gdp_per_capita) then
put "invalid value" gdp_per_capita " for gdp_per_capita" ;
if notdigit(trimn(HDI_for_year)) and not missing (HDI_for_year) then
put "invalid value " HDI_for_year " for HDI_for_year" ;
run;

*Detect the outliers;
*Using 2-Standard variation method;
title "detect the outliers using the 2 standard deviation";
proc means data=mylib.SuicideData noprint ;
var population;
output out=Mean_Std(drop=_type_ _freq_)
mean=
std= / autoname;
run;

data _null_;
```

```
file print;
set mylib.SuicideData;
if _n_ = 1 then set Mean_Std;
if population lt population_Mean - 2*population_StdDev and not missing(population)
or population gt population_Mean + 2*population_StdDev then
put 'The possible outlliers for population: ' population;
run;

title "detect the outliers using the 2 standard deviation";
proc means data=mylib.SuicideData noprint ;
var suicides_no;
output out=Mean_Std(drop=_type_ _freq_)
mean=
std= / autoname;
run;

data _null_;
file print;
set mylib.SuicideData;
if _n_ = 1 then set Mean_Std;
if suicides_no lt suicides_no - 2*suicides_no_StdDev and not missing(suicides_no)
or suicides_no gt suicide_no_Mean + 2*suicide_no__StdDev then put suicides_no;
run;

*checking outliers using IQR method;
proc means data=mylib.SuicideData noprint;
var suicides_no;
output out=mylib.Tmp
Q1=
Q3=
QRange= / autoname;
run;
data _null_;
file print;
set mylib.SuicideData(keep=suicides_no);
if _n_ = 1 then set mylib.Tmp;
if suicides_no le suicides_no_Q1 - 1.5*suicides_no_QRange and not missing(suicides_no) or
```

```
suicides_no ge suicides_no_Q3 + 1.5*suicides_no_QRange then
put "Possible Outlier for suicide no. " suicides_no;
run;

proc means data=mylib.SuicideData noprint;
var population;
output out=mylib.Tmp
Q1=
Q3=
QRange= / autoname;
run;
data _null_;
file print;
set mylib.SuicideData(keep=suicides_no);
if _n_ = 1 then set mylib.Tmp;
if population le population_Q1 - 1.5*population_QRange and not missing(population) or
population ge population_Q3 + 1.5*population_QRange then
put "Possible Outlier for population. " population;
run;

*Check the distribution of the numerical variables;
*Using BOX PLOt;
PROC SGpLOt data=mylib.SuicideData;
hbox suicides_no ;
run;

PROC SGpLOt data=mylib.SuicideData;
hbox HDI_for_year ;
run;

PROC SGpLOt data=mylib.SuicideData;
hbox gdp_per_capita ;
run;

*Test for normality and plot histogram and QQ plots for a variable with a skewed distribution;
*Using QQ PLOT;
ODS select qqplot;
```



```
proc univariate data=mylib.SuicideData normal;
qqplot suicides_no HDI_for_year gdp_per_capita /Normal;
run;

*Using Histogram;
ODS select histogram;
proc univariate data=mylib.SuicideData normal;
var suicides_no HDI_for_year gdp_per_capita;

histogram/normal;
run;

*Ploting histogram QQ PLOT and BOX PLT all together;
ODS select plots;
proc univariate data = mylib.SuicideData plot;
var suicides_no HDI_for_year gdp_per_capita ;
run;

*Applying a transformation ;
Data mylib.SuicideData_updated;
set mylib.SuicideData;
Format suicide_grp;
if suicides_no= . then suicide_grp= 'Data Missing';
else if suicides_no le 500 then suicide_grp = 'Negligible';
else if suicides_no le 2000 and suicides_no gt 500 then suicide_grp = 'Significant';
else if suicides_no le 5000 and suicides_no gt 2000 then suicide_grp = 'To be of concern';
else if suicides_no gt 5000 then suicide_grp = 'Severe';
run;

proc print data=mylib.SuicideData_updated(obs=100);
```