Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India
(Autonomous College Affiliated to University of Mumbai)

Academic SEM: VII **Year: 2022-23**

Experiment: EDA Using SAS

Name:	Tripathi Vinayak Ramprakash
UID:	2019110067
Class:	BE ETRX
Batch:	A
Subject:	Data Analytics Lab

Objective: Getting Familiar with SAS Studio and Perform Exploratory Data Analysis using SAS

System Requirements: SAS Studio

DataSet:

The Dataset considered in this experiment is the inbuilt dataset from SASHELP. It is the data about the medical information about people with attributes like weight BP, Cholesterol, Smoking, Status(Alive or Dead), Cause of death.

Code:

```
data first;
     set sashelp.heart;
run;
title 5 point Summary;
proc means data=sashelp.heart mean median mode std var min max;
run;

Title NUmber of missing values;
proc means data=sashelp.heart nmiss;
run;

proc print =sashelp.heart;
where status = "Dead";
run;

title Getting Number of Distinct Values;
proc sql;
select count(distinct Status) as Status,
       count(distinct DeathCause) as DeathCause,
       count(distinct Sex) as Sex,
       count(distinct Chol_status) as Chol_status,
       count(distinct DeathCause) as Smoking_status
  from sashelp.heart;
quit;

title Correlation of the Attributes;
proc corr data=sashelp.heart;
run;
```

```
title Frequency of the Categorical Values
proc freq data=Sashelp.Heart;
   tables _CHARACTER_;    /* _ALL_ is the defaul */
run;

proc print data= HeartNumeric(obs=5);
run;

proc means data=HeartNumeric nmiss;
run;

ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=sashelp.heart;
     vbox  AgeAtStart / category=;
     yaxis grid;
run;
ods graphics / reset;

title "Scatter Plot of Height and weight";
proc sgplot data=sashelp.heart;
    scatter x = Height  y = Weight;
run;


title "Systolic Outlier";
proc sgplot data=sashelp.heart;
     vbox  Systolic / category=status;
     yaxis grid;
run;
ods graphics / reset;

title "Diastolic Outlier";
proc sgplot data=sashelp.heart;
     vbox  Diastolic / category=status;
     yaxis grid;
run;
ods graphics / reset;


title "Cholestrol ranges";
proc sgplot data=sashelp.heart;
     vbox  Cholesterol / category=Chol_Status;
     yaxis grid;
run;

DATA dead;
   SET sashelp.heart;
   IF (Status = "Dead") THEN OUTPUT;
RUN;
proc print =dead;
run;

title Diastolic BP Histogram grouped by BP_Status;
```

```
ods graphics / reset;
proc sort data=SASHELP.HEART out=_HistogramTaskData;
     by BP_Status;
run;
proc sgplot data=_HistogramTaskData;
     by BP_Status;
     histogram Diastolic /;
     yaxis grid;
run;
proc sgplot data=_HistogramTaskData;
     by BP_Status;
     histogram Systolic /;
     yaxis grid;
run;
```

Output: Please Consider the SAS Output File

Interpretation:

1. First we found out what types of data are there in the dataset. It consisted of Numerical data(weight, Height, Age) and Categorical data(Sex, BP_Status)
2. We found out the mean of the numerical data. The mean age at which the people getting diagnosed is 63 years. This is due to the fact that older people are more likely to get affected by CHD. THe Average and median BP of CHD patient shows a abnormally high figure, this clearly indicates that CHD Patients have high BP.
3. By looking closely to the attribute of smoking the people who are found to have CHD smokes on average 9 cigarette per day.
4. There were some of the values missing also. After getting the correlation of the attributes that there are no correlation among the features so we can say that the feature are independent of each other.
5. However there lies some correlation of order 60% in the feature weight and height.
6. Next getting knowing about the outliers we found that the people died were having higher diastolic and systolic pressure than the people alive. This clearly indicates that the people died had reached critical stage of CHD.
7. Similarly when we see outliers for the Cholesterol ranges there were no outliers in the borderline since the range of the borderline is very small. For the desirable cholesterol we found that the outlier exists in the downward side this indicates that people are trying to have as much lower cholesterol levels as possible. For high cholesterol we found that due CHD the cholesterol level shows large variation of the higher side

Conclusion:

- We got introduced to SAS
- We understood how to load dataset in SAS, do computation and plot graphs.