Name: Maitri Vora
UID: 2019120068
Batch-B
Class: EXTC
Experiment-2

AIM: Using the SAS software to analyze statistical data.

PROBLEM STATEMENT: Study and understand the workings of SAS studio by referring to the online materials and documentation, etc., and then implement a small problem.

Theory

1) What is SAS?

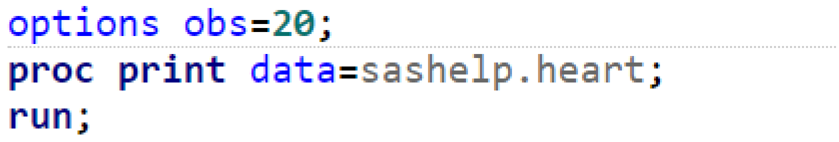
- The full form of SAS is Statistical Analysis Software. It was created in the year 1960 and was used for, business intelligence, Predictive Analysis, Descriptive and Prescriptive Analysis, data management etc. Since then, many new statistical procedures and components were introduced in the software.
- In contrast to other BI solutions on the market, SAS uses considerable programming to transform and analyse data rather than just a simple drag-and-drop method

2) Key features of SAS

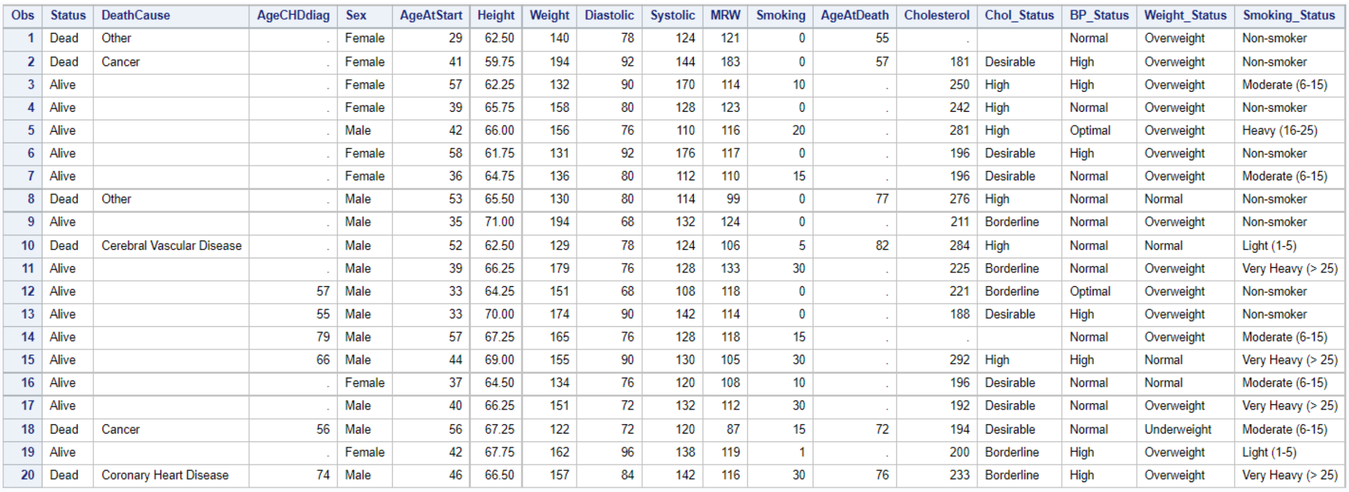
- Easily access raw data files & data from an external database. Read and write almost any data format!
- Manage data using tools for data entry, editing retrieval, formatting & conversion
- Analyze data using descriptive, statistics, multivariate techniques, forecasting, modelling, linear programming
- Advanced analytics helps you to make changes and improvements in business practices.
- Report formation with perfect graphs
- Operations research and project management
- Data updating and modification
- Powerful data handling language
- Excellent data cleansing functions
- Interact with multiple host systems

Code & Outpu

1. I have used the already present database in the SAS studio (My Libraries -> SASHELP -> HEART) I have printed 20 observations from the dataset to know what is in the dataset

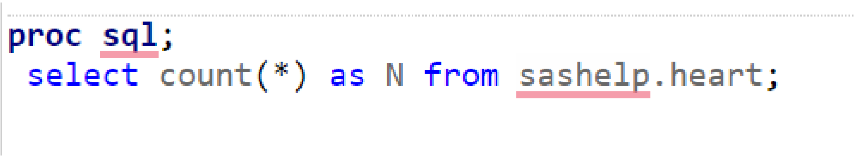


```
options obs=20;
proc print data=sashelp.heart;
run;
```

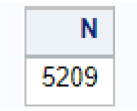


Obs	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol	Chol_Status	BP_Status	Weight_Status	Smoking_Status
1	Dead	Other	.	Female	29	62.50	140	78	124	121	0	55	.		Normal	Overweight	Non-smoker
2	Dead	Cancer	.	Female	41	59.75	194	92	144	183	0	57	181	Desirable	High	Overweight	Non-smoker
3	Alive		.	Female	57	62.25	132	90	170	114	10	.	250	High	High	Overweight	Moderate (6-15)
4	Alive		.	Female	39	65.75	158	80	128	123	0	.	242	High	Normal	Overweight	Non-smoker
5	Alive		.	Male	42	66.00	156	76	110	116	20	.	281	High	Optimal	Overweight	Heavy (16-25)
6	Alive		.	Female	58	61.75	131	92	176	117	0	.	196	Desirable	High	Overweight	Non-smoker
7	Alive		.	Female	36	64.75	136	80	112	110	15	.	196	Desirable	Normal	Overweight	Moderate (6-15)
8	Dead	Other	.	Male	53	65.50	130	80	114	99	0	77	276	High	Normal	Normal	Non-smoker
9	Alive		.	Male	35	71.00	194	68	132	124	0	.	211	Borderline	Normal	Overweight	Non-smoker
10	Dead	Cerebral Vascular Disease	.	Male	52	62.50	129	78	124	106	5	82	284	High	Normal	Normal	Light (1-5)
11	Alive		.	Male	39	66.25	179	76	128	133	30	.	225	Borderline	Normal	Overweight	Very Heavy (> 25)
12	Alive		57	Male	33	64.25	151	68	108	118	0	.	221	Borderline	Optimal	Overweight	Non-smoker
13	Alive		55	Male	33	70.00	174	90	142	114	0	.	188	Desirable	High	Overweight	Non-smoker
14	Alive		79	Male	57	67.25	165	76	128	118	15	.	.		Normal	Overweight	Moderate (6-15)
15	Alive		66	Male	44	69.00	155	90	130	105	30	.	292	High	High	Normal	Very Heavy (> 25)
16	Alive		.	Female	37	64.50	134	76	120	108	10	.	196	Desirable	Normal	Normal	Moderate (6-15)
17	Alive		.	Male	40	66.25	151	72	132	112	30	.	192	Desirable	Normal	Overweight	Very Heavy (> 25)
18	Dead	Cancer	56	Male	56	67.25	122	72	120	87	15	72	194	Desirable	Normal	Underweight	Moderate (6-15)
19	Alive		.	Female	42	67.75	162	96	138	119	1	.	200	Borderline	High	Overweight	Light (1-5)
20	Dead	Coronary Heart Disease	74	Male	46	66.50	157	84	142	116	30	76	233	Borderline	High	Overweight	Very Heavy (> 25)

2. I have used count(*) function in sql to check the number of observations in dataset



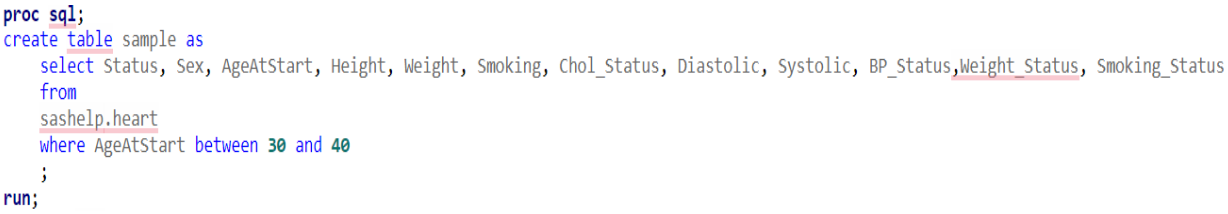
```
proc sql;
 select count(*) as N from sashelp.heart;
```



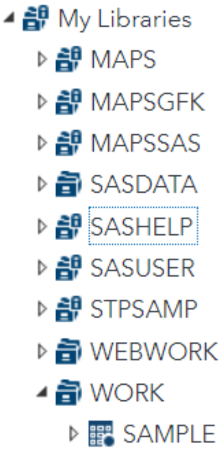
```
   N
5209
```

We have 5209 observations in our data

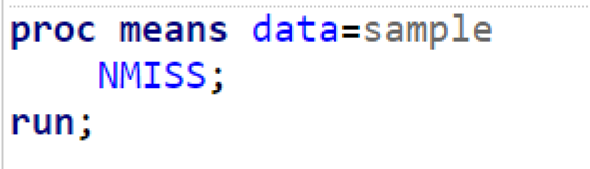
3. I have taken a small sample of the dataset where AgeAt Start is between 30 to 40 and stored it inside the “sample” file inside the work directory in “My Libraries.(My Libraries->work>sample



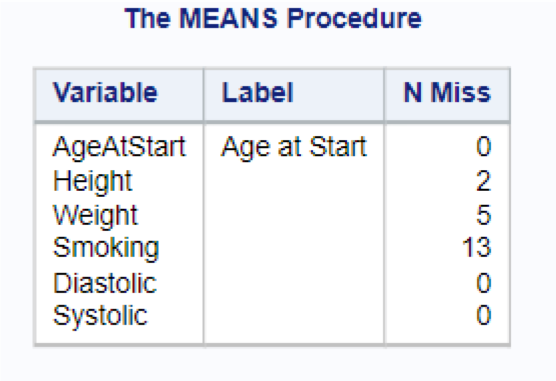
```
proc sql;
create table sample as
    select Status, Sex, AgeAtStart, Height, Weight, Smoking, Chol_Status, Diastolic, Systolic, BP_Status,Weight_Status, Smoking_Status
    from
    sashelp.heart
    where AgeAtStart between 30 and 40
    ;
run;
```



4. Checking number of null values in our “sample” data



```
proc means data=sample
    NMISS;
run;
```

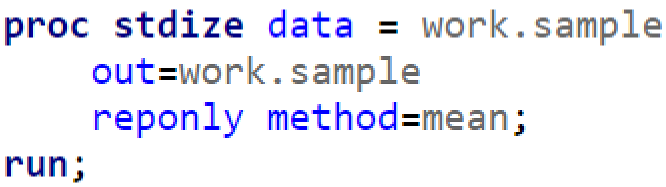


The MEANS Procedure

Variable	Label	N Miss
AgeAtStart	Age at Start	0
Height		2
Weight		5
Smoking		13
Diastolic		0
Systolic		0

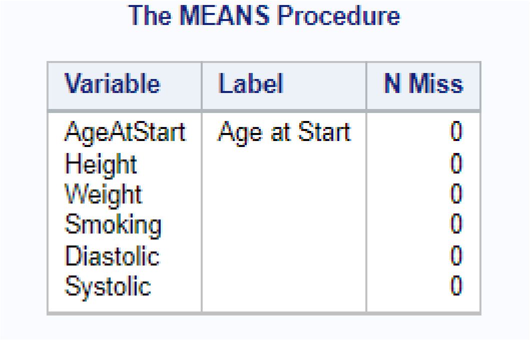
As we can see there are 2 null values in height column, 5 in weight, and 13 in smoking.

5. In SAS, null values are represented as “.” . Since, we have null values in our data, we



```
proc stdize data = work.sample
    out=work.sample
    reponly method=mean;
run;
```

will replace the null values of numeric data with the mean of that column

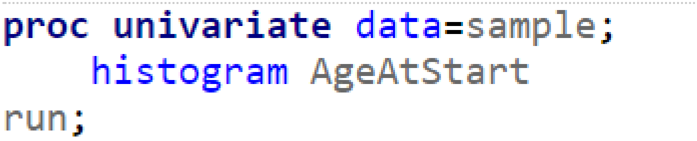


The MEANS Procedure

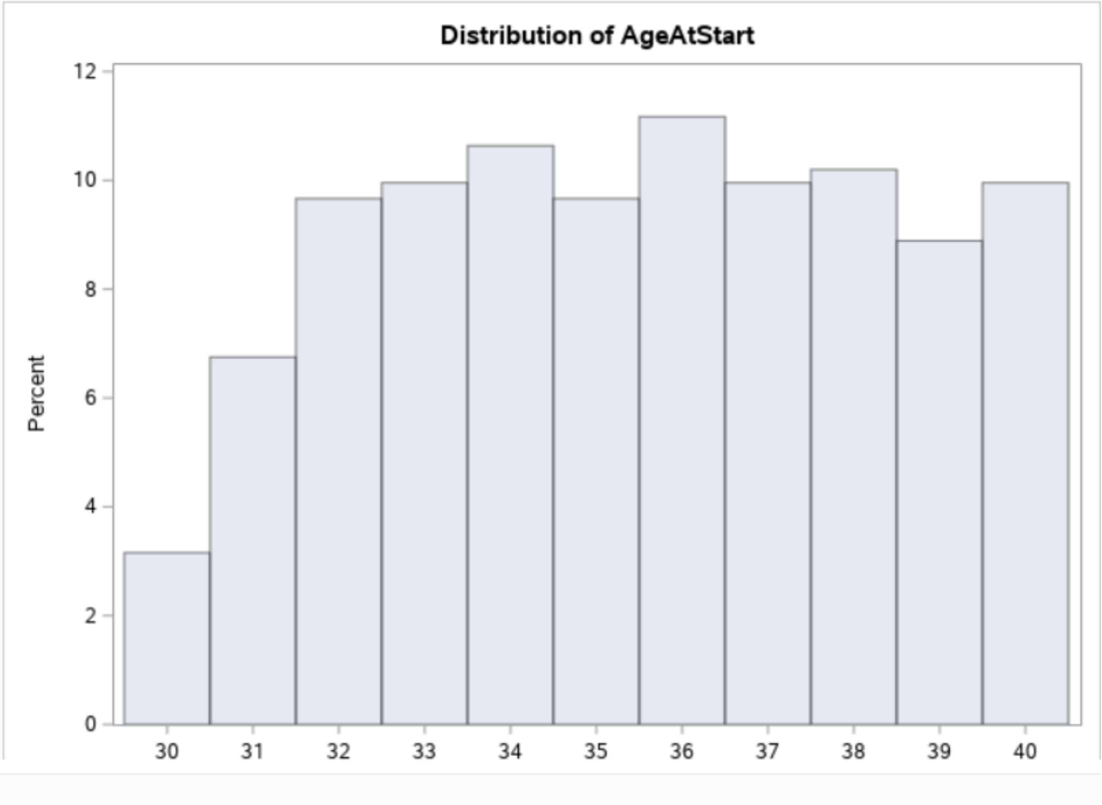
Variable	Label	N Miss
AgeAtStart	Age at Start	0
Height		0
Weight		0
Smoking		0
Diastolic		0
Systolic		0

After substituting null values with the mean of the column we have 0 null values in all columns

6. After filling the null values, next I used the SAS Data representation to view and analyze the data in a better manner. First plot which I am using is the univariate histogram plot which is a graphical display of data using bars of different heights. It groups the various numbers in the data set into many ranges. It also represents the estimation of the probability of distribution of a continuous variable. In SAS the PROC UNIVARIATE is used to create histograms.

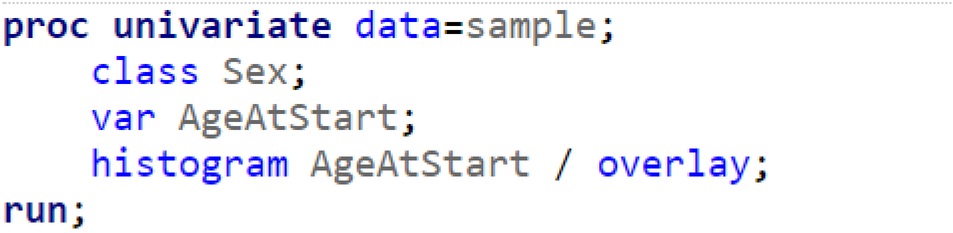


```
proc univariate data=sample;
    histogram AgeAtStart
run;
```

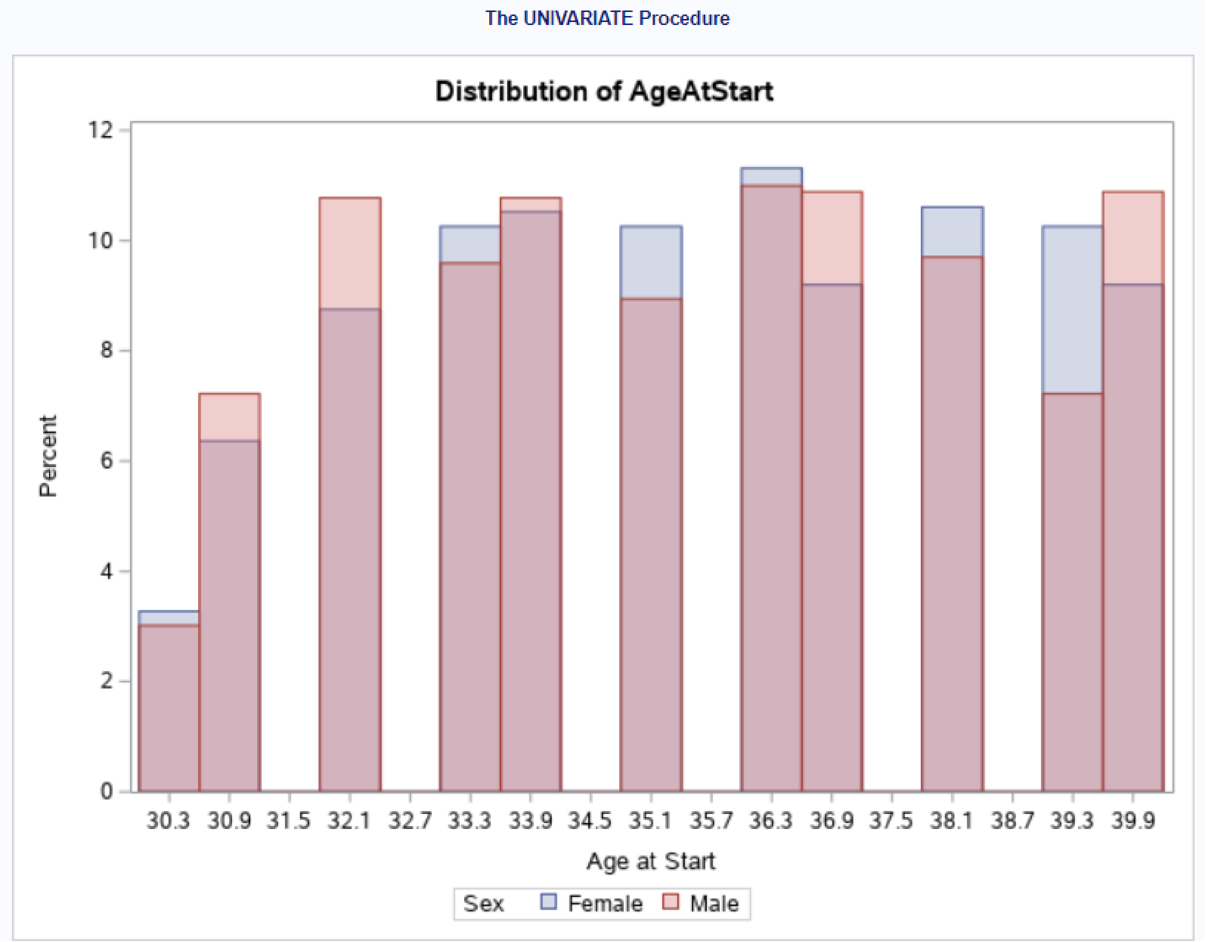


From the above histogram plot, we can evidently say that as the age increases the chances of heart diseases increase with the peak being at 36 after which the curve dips down

7. Next I have compared the chances of heart diseases between males and females. I have used a overlay histogram plot for making the comparison.This plot allows us to quickly see that the required comparison.Notice that the two histograms share an x-axis, which makes it easy to compare the points values between the two genders.

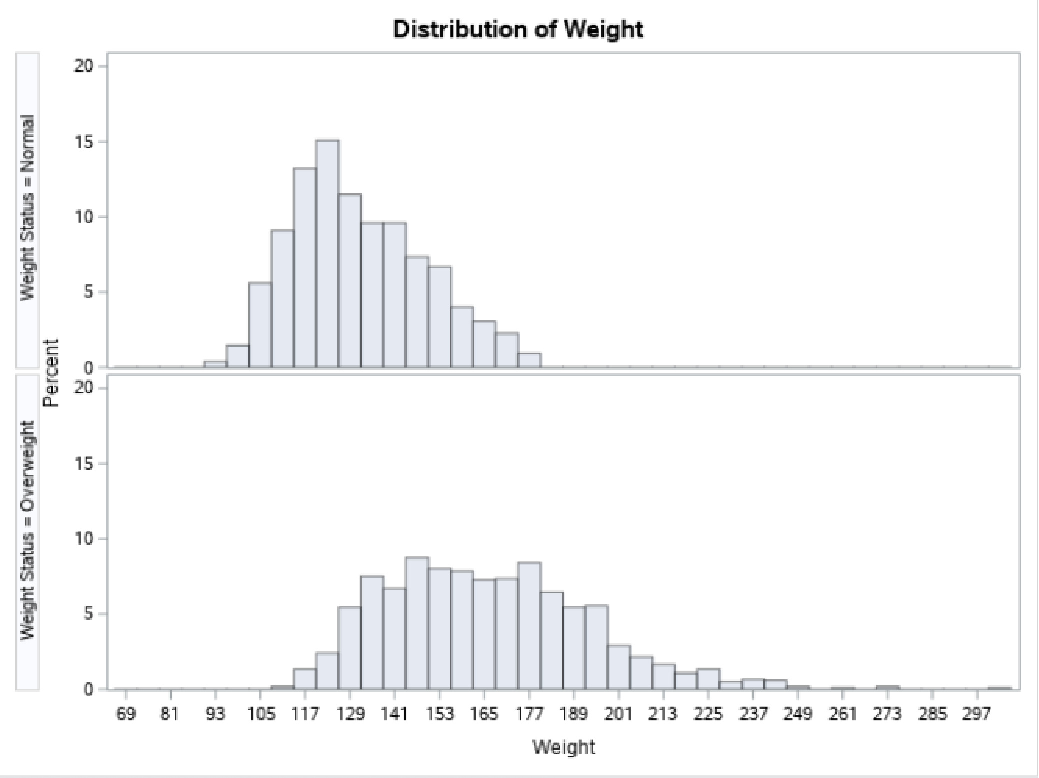


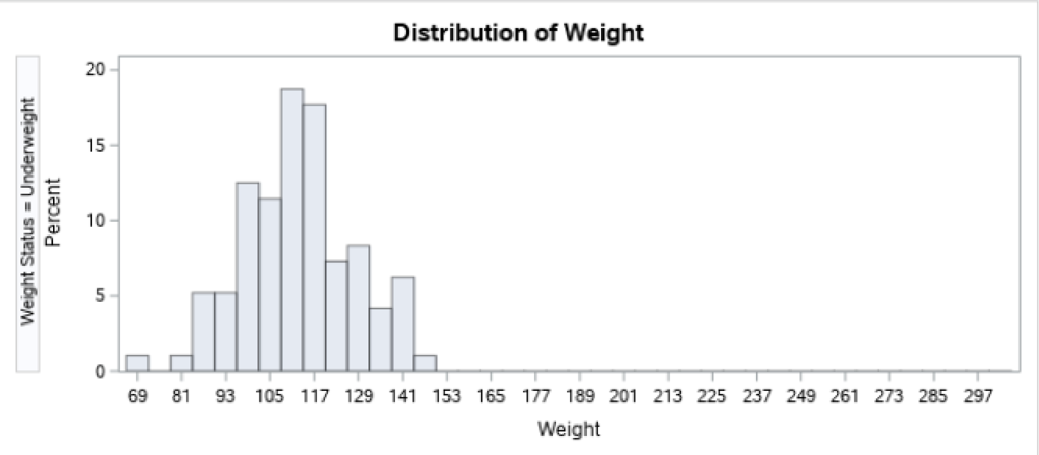
```
proc univariate data=sample;
    class Sex;
    var AgeAtStart;
    histogram AgeAtStart / overlay;
run;
```



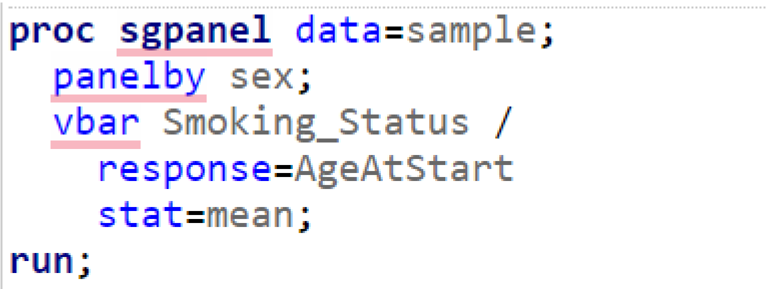
As observed from above overlay histogram plot, for age 33,35,36,38 and 39 more females are affected as compared to males, and for other age values more males are affected than females.

8. From the below graph we can see that in our data, we have overweight, underweight and a normal weighted person. As we can see 117-120 pounds have highest frequency in normal weighted group. The maximum weight in underweight category is between141-153, 177-189 for normal weighted person and higher ranges are for overweighted person.

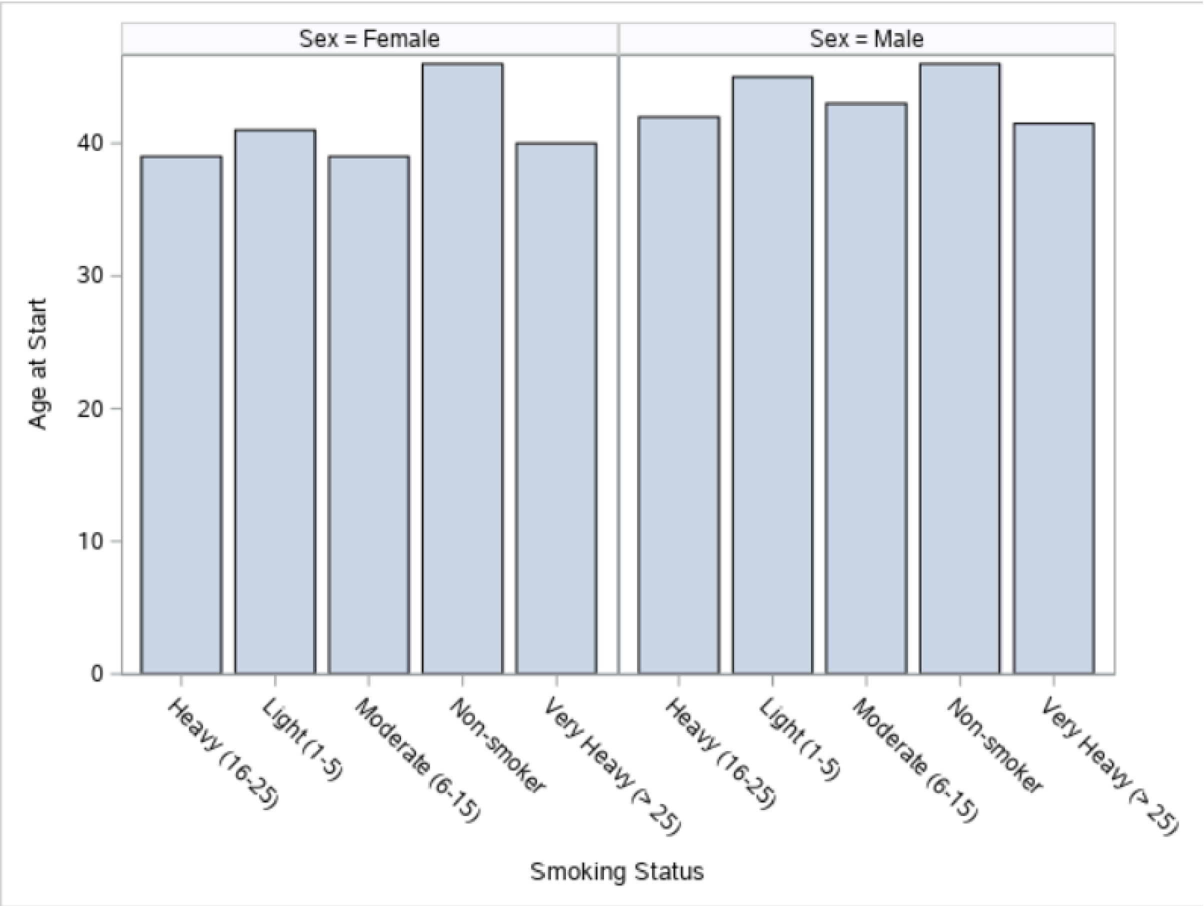




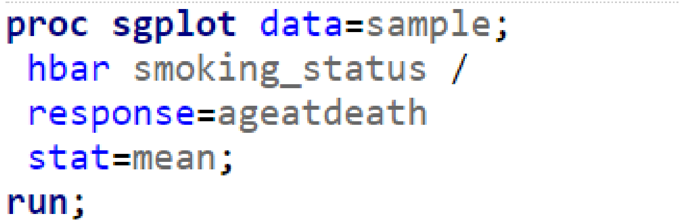
9. As it can be seen the risk of having heart disease at a young age is high for heavy, moderate and very heavy smokers. Non smokers face the risk of heart disease in their late 40's. The risk of having disease is almost same for males and females for all types of smokers.



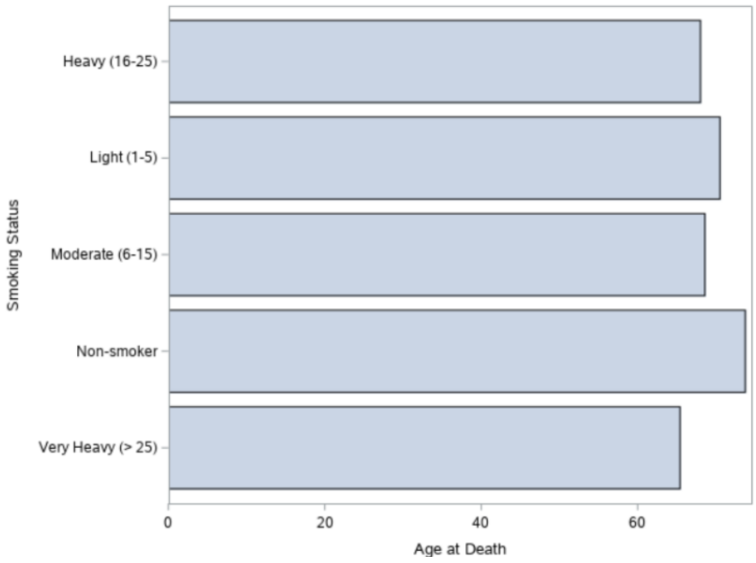
```
proc sgpanel data=sample;
  panelby sex;
  vbar Smoking_Status /
    response=AgeAtStart
    stat=mean;
run;
```



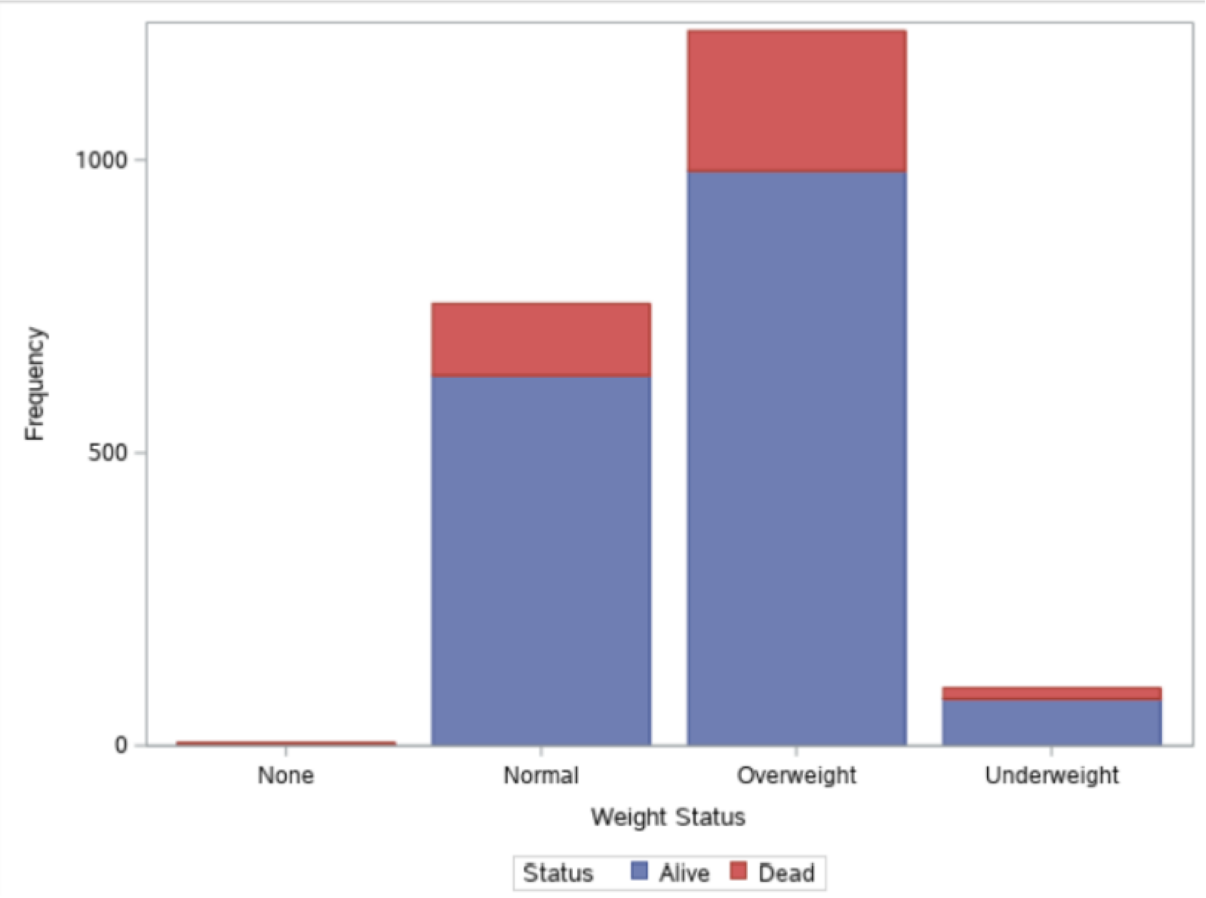
10. From the below plot we can see that, a non smoker lives the most as compared to heavy, moderate and very heavy smokers. Very heavy smokers, die comparatively early than others.



```
proc sgplot data=sample;
 hbar smoking_status /
 response=ageatdeath
 stat=mean;
run;
```

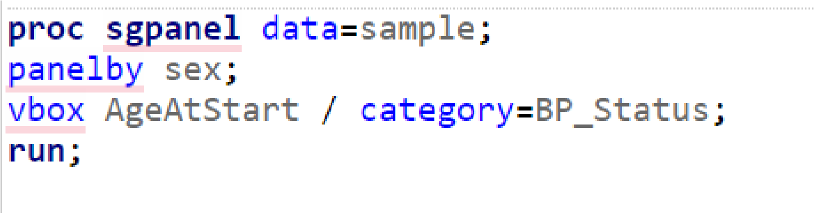


11. As we can see from below graph, smoking and weight is also related in determining when a person will die.

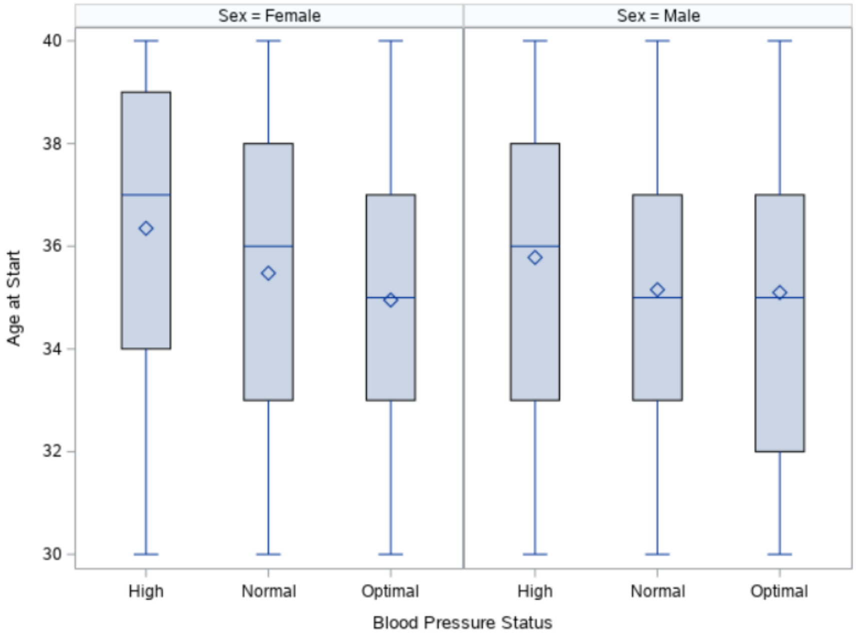


12.Box Plots in Vertical Panel

We can divide the Boxplots of a variable into many vertical panels(columns). Each panel holds the boxplots for all the categorical variables. But the boxplots are further grouped using another third variable which divides the graph into multiple panels.



```
proc sgpanel data=sample;
panelby sex;
vbox AgeAtStart / category=BP_Status;
run;
```

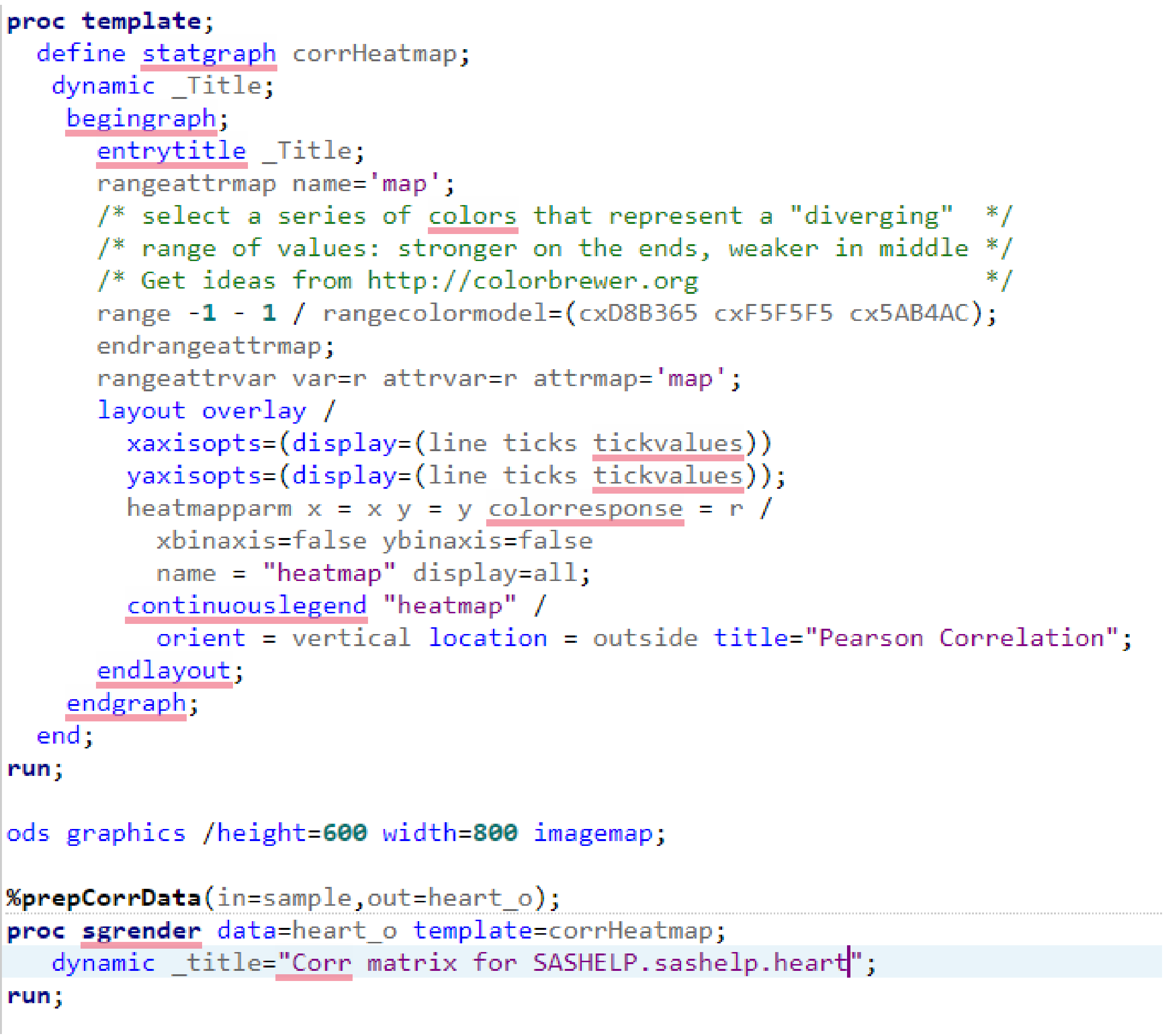


As we can see from above graph, for high,normal and optimal BP Status, women are at a higher average risk at an young age than the men. The variability of data for optimal BP status for female is almost normally distributed as compared to other classes and gender.

13. Heatmap

HeatMaps is about replacing numbers with colors because the human brain understands visuals better than numbers, text, or any written data.Heatmaps can describe the density or intensity of variables, visualize patterns, variance, and even anomalies. Heatmaps show relationships between variables. These variables are plotted on both axes. We look for patterns in the cell by noticing the color change. It only accepts numeric data and plots it on the grid, displaying different data values by varying color intensity.

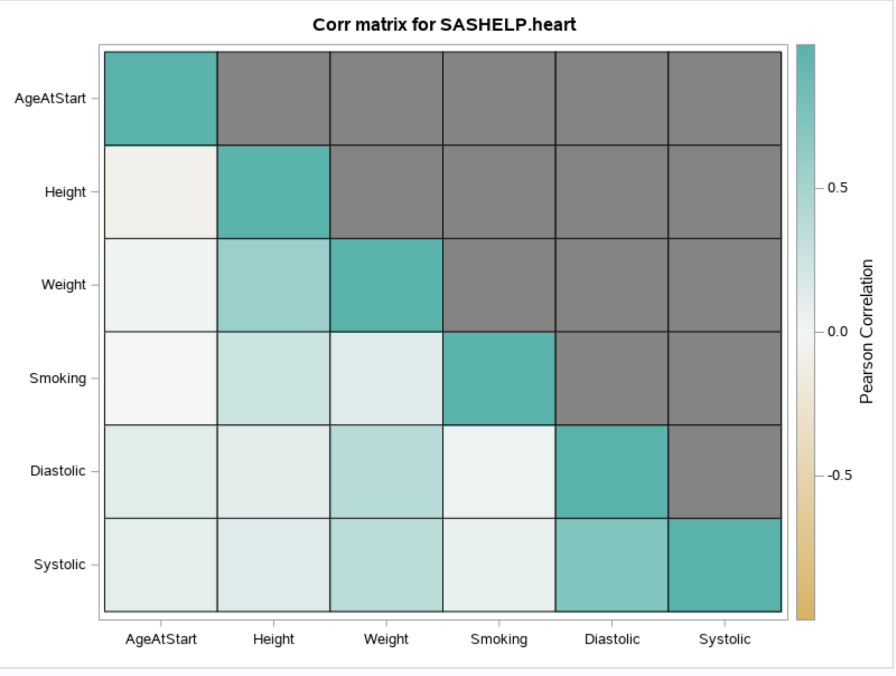
From the output heatmap show below, we can see that there is high correlation between systoli and diastolic. There is no correlation between height and AgeAtStart. There is somewhat positive correlation between weight at AgeAtStart which indicates that if a person is overweighted his/her chances of getting a heart disease is higher. There is a negative correlation between AgeAtStart and Smoking. This makes sense, as a person's smoking score increases, he/she faces diseasess at younger age.



```
proc template;
  define statgraph corrHeatmap;
   dynamic _Title;
    begingraph;
      entrytitle _Title;
      rangeattrmap name='map';
      /* select a series of colors that represent a "diverging"  */
      /* range of values: stronger on the ends, weaker in middle */
      /* Get ideas from http://colorbrewer.org                   */
      range -1 - 1 / rangecolormodel=(cxD8B365 cxF5F5F5 cx5AB4AC);
      endrangeattrmap;
      rangeattrvar var=r attrvar=r attrmap='map';
      layout overlay /
        xaxisopts=(display=(line ticks tickvalues))
        yaxisopts=(display=(line ticks tickvalues));
        heatmapparm x = x y = y colorresponse = r /
          xbinaxis=false ybinaxis=false
          name = "heatmap" display=all;
        continuouslegend "heatmap" /
          orient = vertical location = outside title="Pearson Correlation";
      endlayout;
    endgraph;
  end;
run;

ods graphics /height=600 width=800 imagemap;

%prepCorrData(in=sample,out=heart_o);
proc sgrender data=heart_o template=corrHeatmap;
   dynamic _title="Corr matrix for SASHELP.sashelp.heart";
run;
```



CONCLUSION :

After learning about SAS studio, its features, functionalities and also implementing them, I was able to conclude that :

- SAS provides us various functionalities like Data Management, Statistical Analysis, Report formation with perfect graphics etc that I have used in the above experiment.
- So in my case I loaded the already existing dataset by just making some changes in the original dataset to create my own data and then after cleaning the values, filling missing values I performed various kind of data analysis by plotting the various utility functions for plotting that SAS provides.