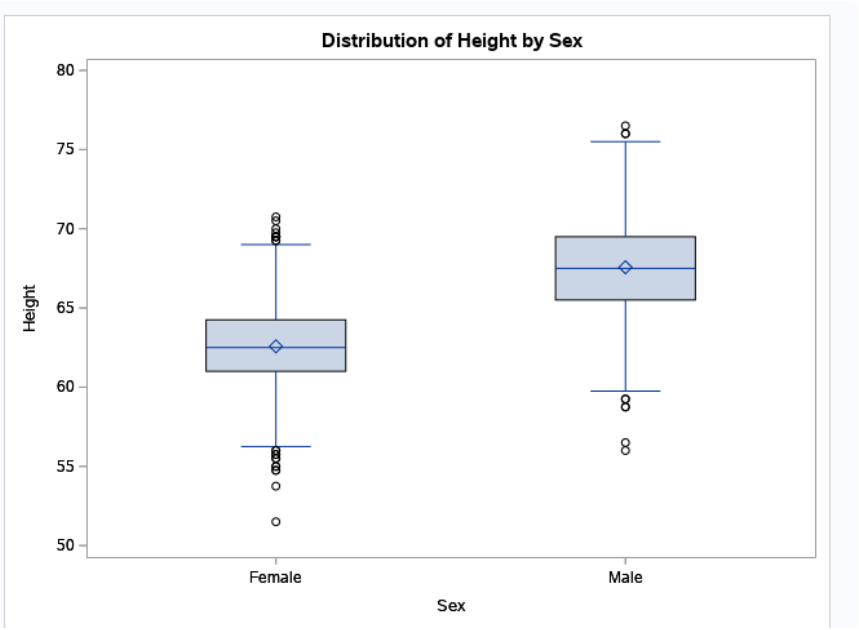
Name	Pranav Nair
UID	2019130042
Branch	TE COMPS
Course	Data Analytics (OE)

EXPERIMENT - 3

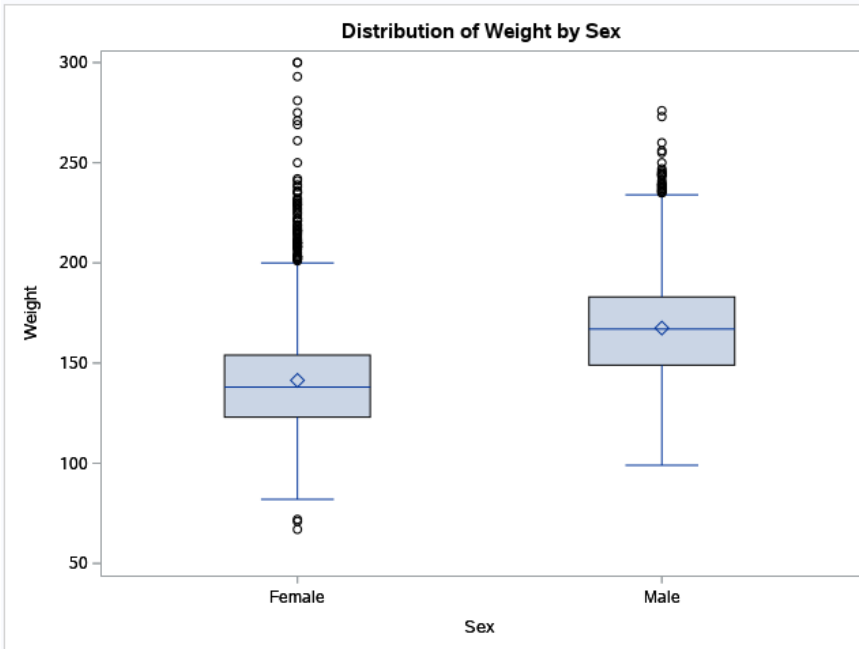
Objective : To explore the SAS software

Dataset : The dataset used for this experiment is the HEART dataset available in the SAS software itself. It contains 5209 rows and 17 columns. It contains features like Survival status of the patient, death cause, gender, age at start, height, weight, diastolic and systolic blood pressure, cholesterol levels, weight status, smoking status, etc.

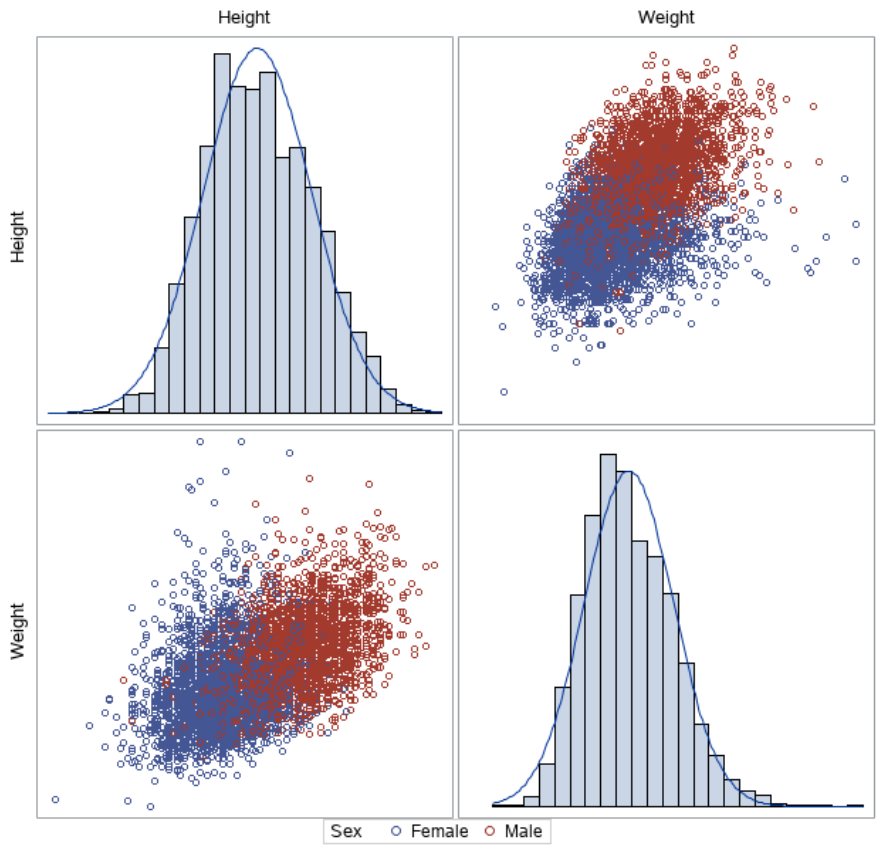
Below I have obtained certain plots using the SAS software which give us insights of the distribution of data.



The plot above shows the outliers for the height of the test subjects categorized based on their gender.



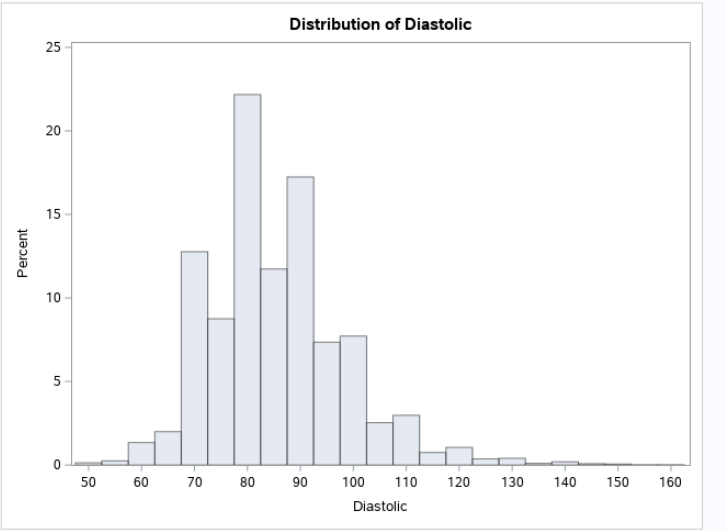
The plot above shows the outliers for the weight of the test subjects categorized based on their gender.

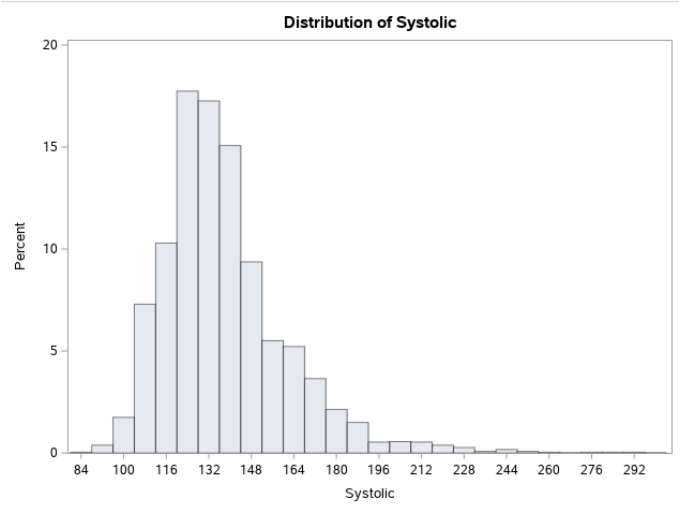


This plot shows the distribution of height and weight and the relationship between the two variables.

Variable	Label	Mean	Std Dev	Minimum	Maximum	N
Height		64.8131847	3.5827074	51.5000000	76.5000000	5203
Weight		153.0866808	28.9154261	67.0000000	300.0000000	5203
Diastolic		85.3586101	12.9730913	50.0000000	160.0000000	5209
Systolic		136.9095796	23.7395964	82.0000000	300.0000000	5209
Smoking		9.3665185	12.0314511	0	60.0000000	5173
AgeAtDeath	Age at Death	70.5364139	10.5594062	36.0000000	93.0000000	1991
Cholesterol		227.4174412	44.9355238	96.0000000	568.0000000	5057

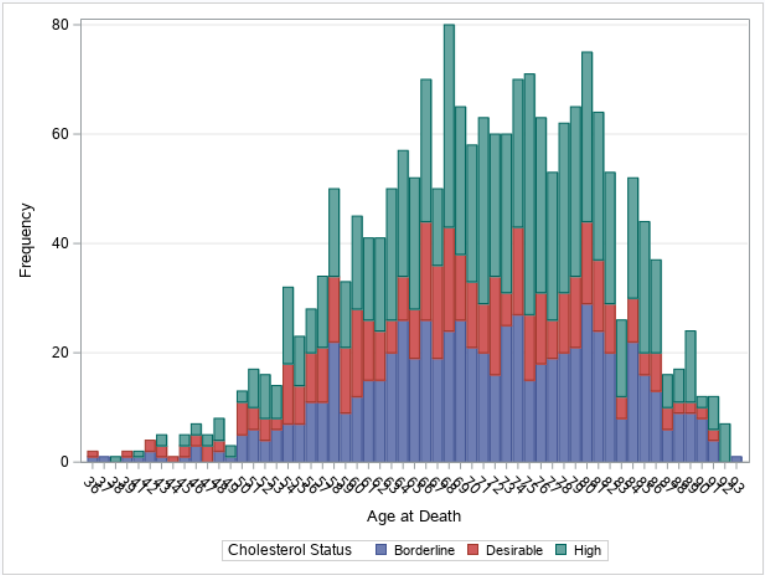
The above table gives us a summary of the numeric data present in the dataset.

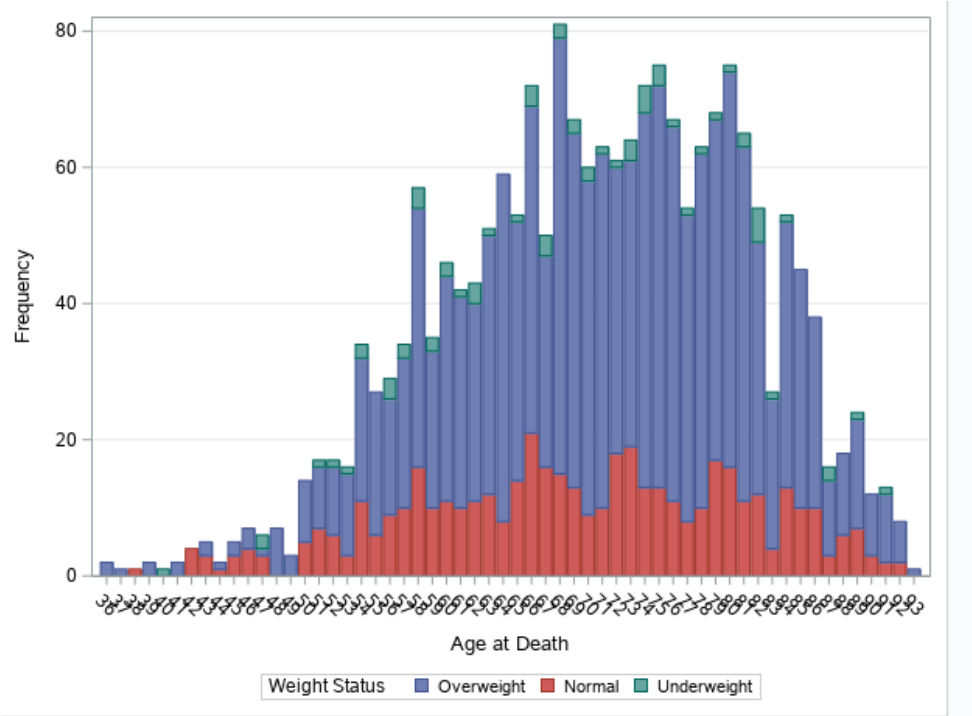




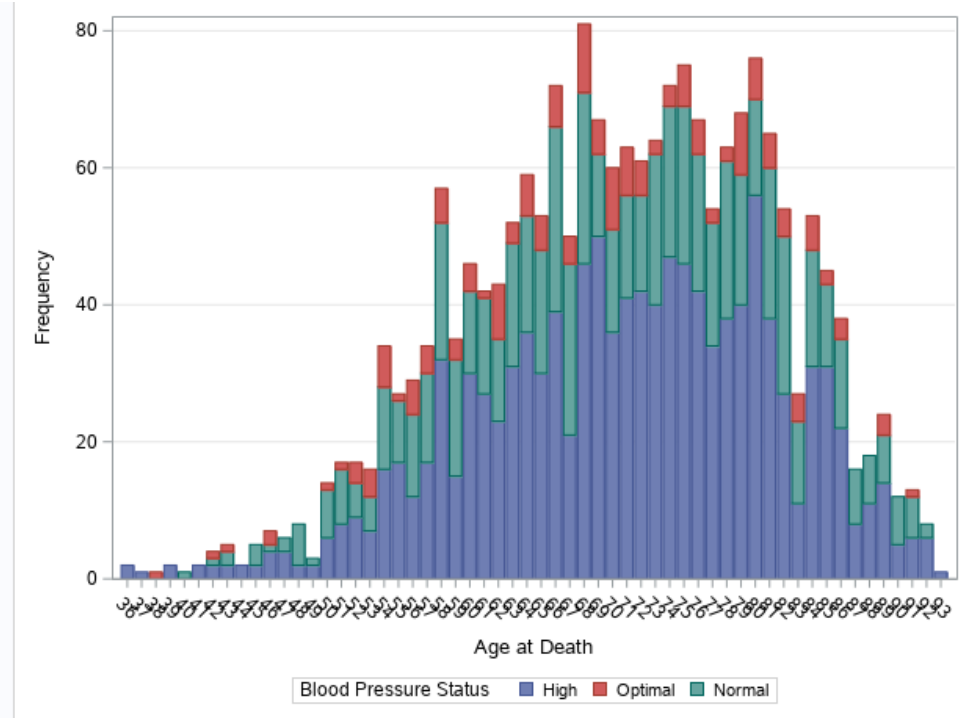
The above two histograms show the frequency distribution of the diastolic and systolic blood pressures of the test subjects.

The below plot shows us the patient's age at death and categorized on the basis of their cholesterol levels.

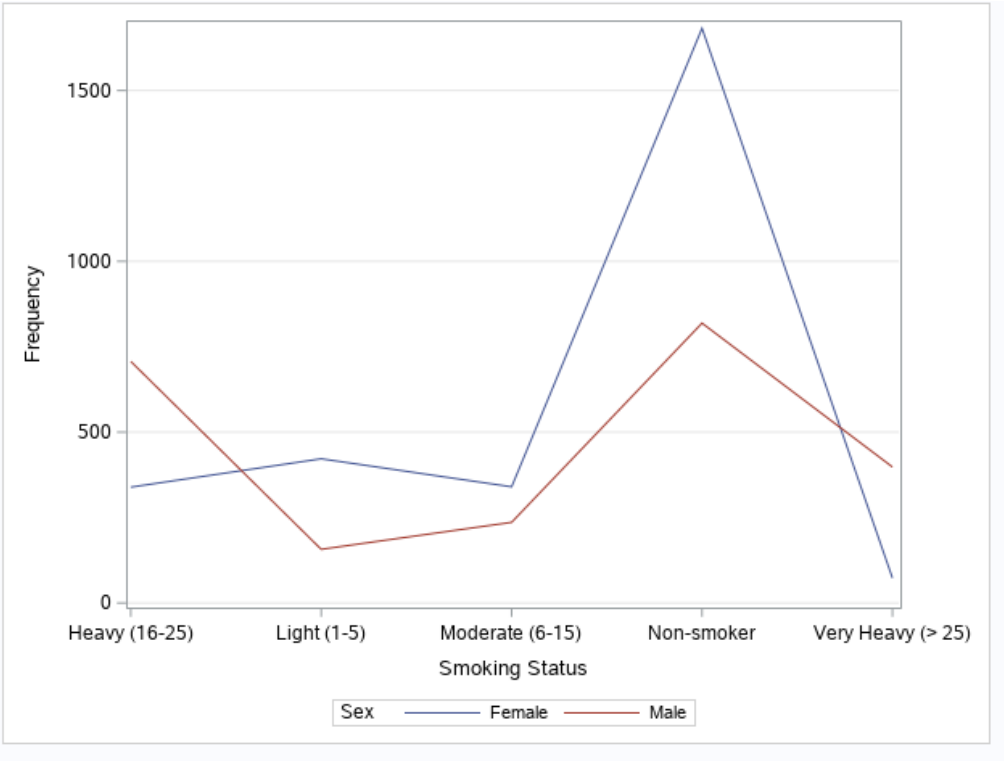




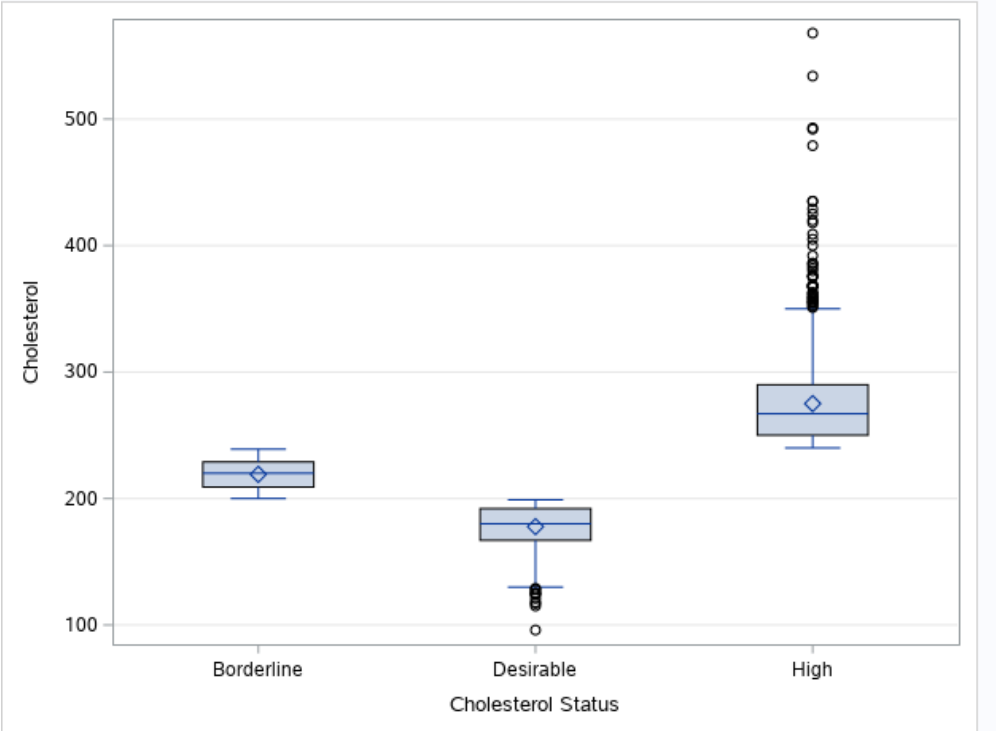
The below plot shows us the patient's age at death and categorized on the basis of their weight status.



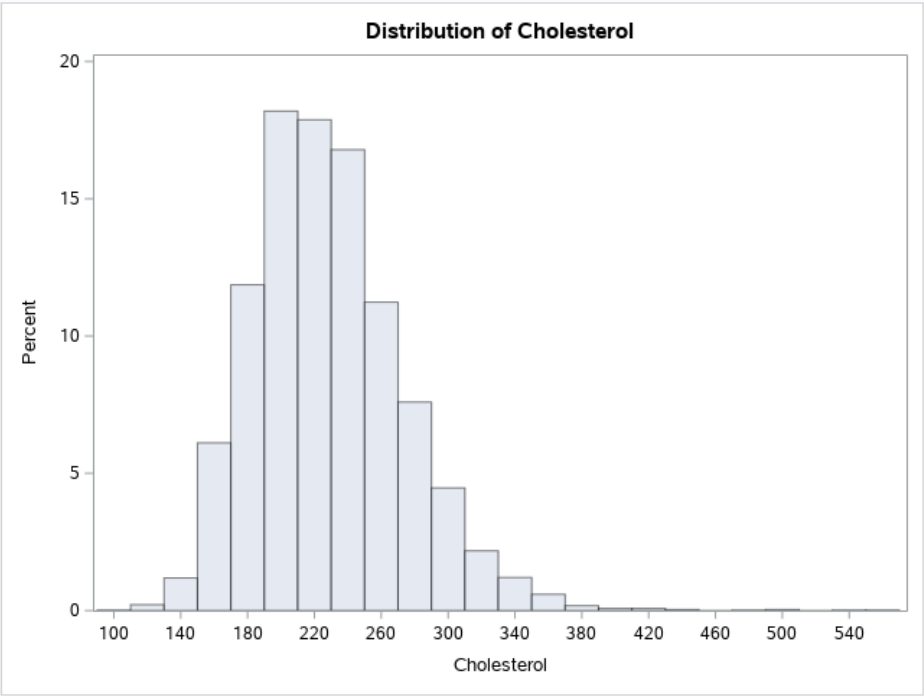
The below plot shows us the patient's age at death and categorized on the basis of their blood pressure levels.



This line plot shows the smoking status of the male and female candidates.



The above box plots give an insight of the patients cholesterol status



The cholesterol levels of the patient's recorded are given in the above histogram.

F-TEST

Now we conduct a F-test where the null hypothesis is that the means of cholesterol for male and female is same, the results of the tests are as follows :

Class Level Information		
Class	Levels	Values
Sex	2	Female Male

Number of Observations Read	5209
Number of Observations Used	5057

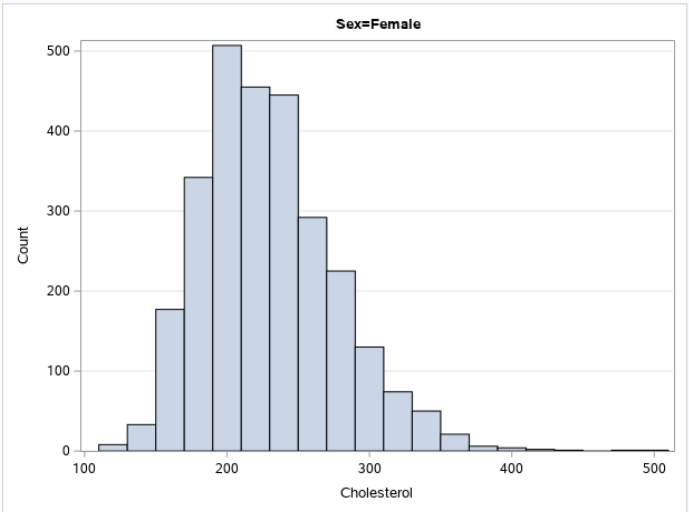
Dependent Variable: Cholesterol

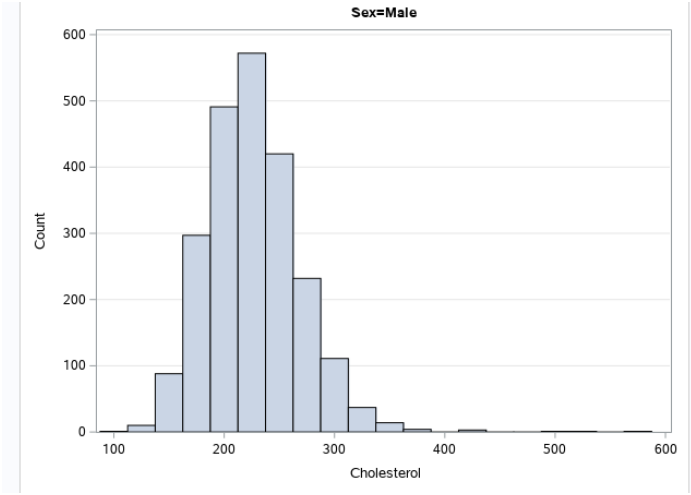
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7768.13	7768.13	3.85	0.0498
Error	5055	10201313.65	2018.06		
Corrected Total	5056	10209081.78			

R-Square	Coeff Var	Root MSE	Cholesterol Mean
0.000761	19.75348	44.92287	227.4174

The level of significance taken is 0.05
Now since we get the p value using the f value , just less than 0.05, we can say that the means of these two groups are not the same, that is we can reject the null hypothesis.

Level of Sex	N	Cholesterol	
		Mean	Std Dev
Female	2774	228.541817	46.9216942
Male	2283	226.051248	42.3672387





Inference :

In this experiment, I have explored the various functionalities of the SAS software. SAS is a command-driven statistical software suite widely used for statistical data analysis and visualization. SAS's full form is Statistical Analysis Software. It allows you to use qualitative techniques and processes which help you to enhance employee productivity and business profits. SAS is also used for advanced analytics like business intelligence, crime investigation, and predictive analysis. In SAS, data is extracted & categorized which helps you to identify and analyze data patterns. It is a software suite which allows you to perform advanced analysis, Business Intelligence, Predictive Analysis, data management to operate effectively in the competitive & changing business conditions. Moreover, SAS is platform independent which means you can run SAS on any operating system either Linux or Windows.