Statistical Analysis System

Project: Analysing Heart Study Data

In SAS, sashelp.heart is a built-in dataset that comes with SAS software and contains information related to heart disease patients. This dataset is often used for educational and demonstration purposes. Let's explain the key aspects of the sashelp.heart dataset:

Source: The sashelp.heart dataset is not sourced from real-world data but is instead a synthetic dataset created for teaching and learning purposes within the SAS environment.

Contents: The dataset typically contains information about individuals with heart disease, including variables such as age, gender, cholesterol levels, blood pressure, smoking status, and more. It's designed to serve as a sample dataset for practicing data analysis and statistical techniques.

Usage: SAS users often use sashelp.heart to demonstrate various SAS procedures and techniques, such as data manipulation, statistical analysis, and data visualization. It's a convenient dataset for learning and testing because it's readily available within SAS environments.

Availability: sashelp.heart is usually available in SAS installations by default. Users can access it without the need to import or load an external dataset.

Here are some example variables you might find in the sashelp.heart dataset (please note that the actual variables may vary depending on the SAS version and dataset configuration):

Age: The age of the patients.

Sex: Gender of the patients (e.g., Male or Female).

Chol: Cholesterol levels of the patients.

BP: Blood pressure measurements.

Smoker: Indicates whether the patient is a smoker (e.g., Yes or No).

Chest Pain: Description of chest pain symptoms.

MaxHR: Maximum heart rate.

RestECG: Resting electrocardiogram results.

Disease: Presence or absence of heart disease (e.g., 0 for No Disease, 1 for Disease).

Users can analyze and visualize this dataset using various SAS procedures and techniques to gain insights into heart disease risk factors, correlations, and other related topics. It's

particularly useful for learning SAS programming and data analysis due to its predefined structure and availability within SAS environments.

This code block is using the proc sgplot procedure to create a histogram.

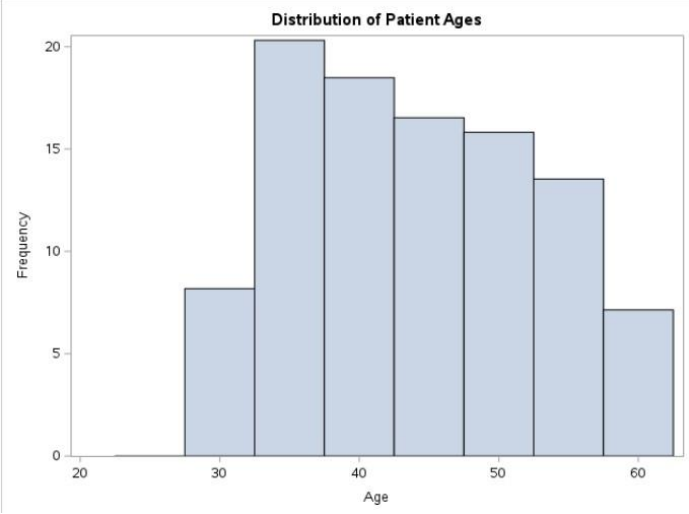
data=Heart specifies the dataset to be used for plotting (Note: the dataset name should be consistent; it should be sashelp.heart instead of Heart).

histogram AgeAtStart / binwidth=5; creates a histogram of the variable AgeAtStart with a bin width of 5 units. This means that the age values in the dataset will be grouped into bins of width 5, and the frequency of each bin will be plotted.

xaxis label="Age"; and yaxis label="Frequency"; are specifying the labels for the x and y-axes of the histogram. title "Distribution of Patient Ages"; adds a title to the histogram.

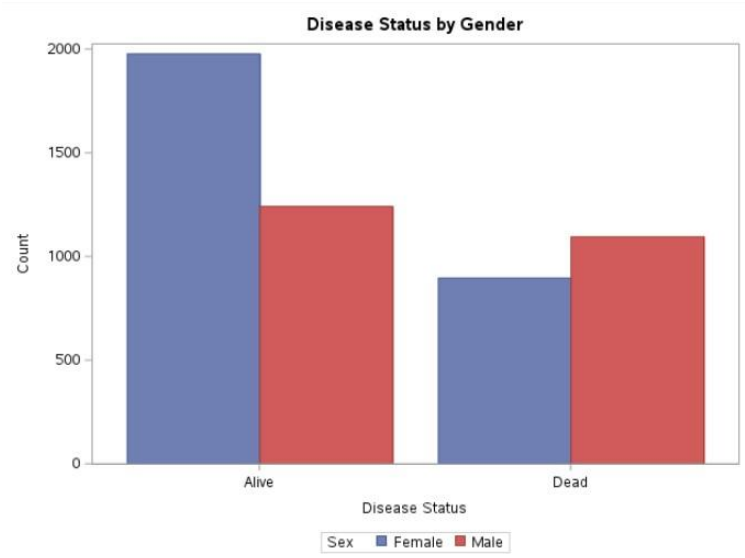
```
/*Loading Data*/
data Heart;
/*Data Set
Selection:*/ set
sashelp.heart; run;
/*Data Printing:*/ proc
print data=sashelp.heart;
run;
/* Create a histogram of patient
ages */ proc sgplot data=Heart;
histogram AgeAtStart / binwidth=5;
xaxis label="Age"; yaxis
label="Frequency"; title
"Distribution of Patient Ages";
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)

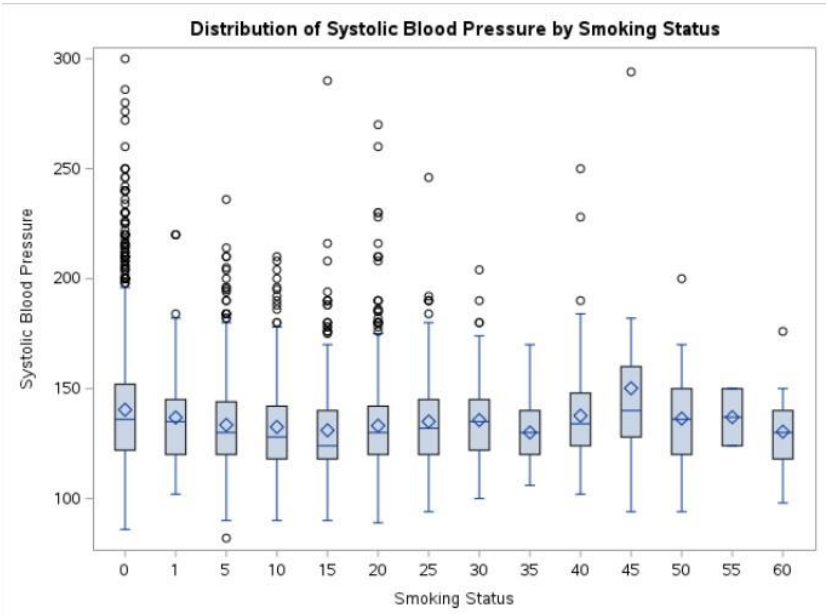


/* Create a bar chart for disease status by gender */

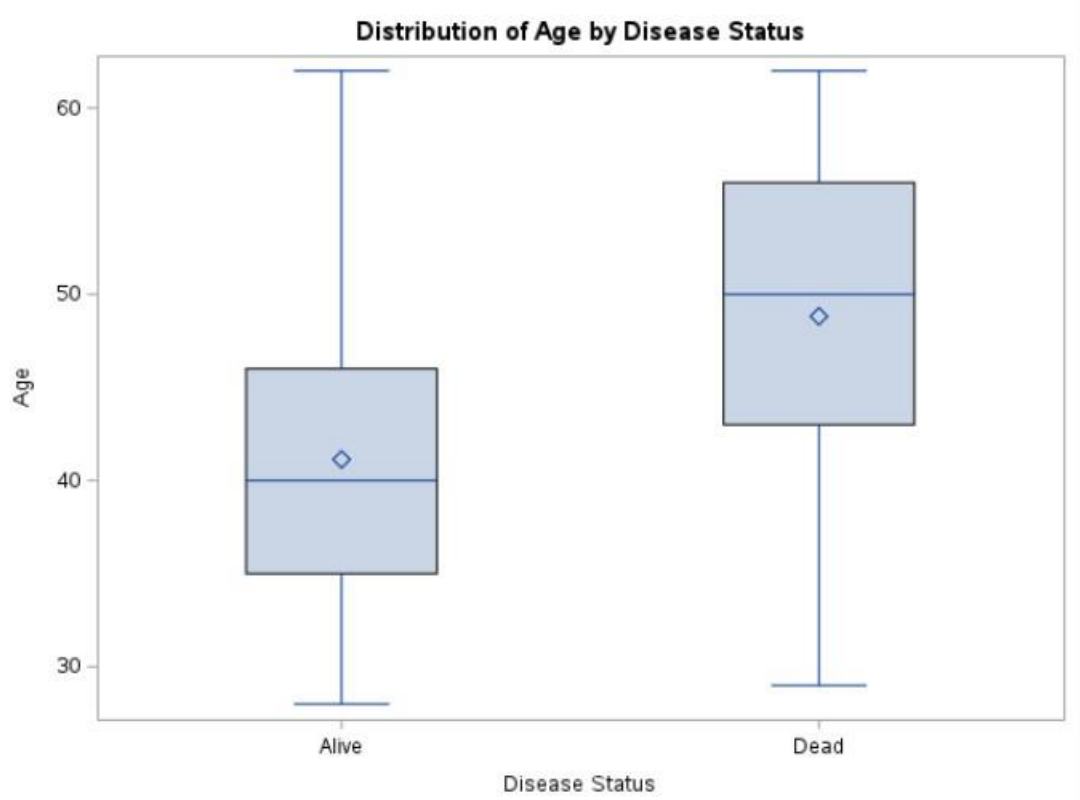
```
proc sgplot data=Heart;  vbar Status / group=Sex groupdisplay=cluster;  xaxis label="Disease Status";  yaxis label="Count";  title "Disease Status by Gender";
run;
```



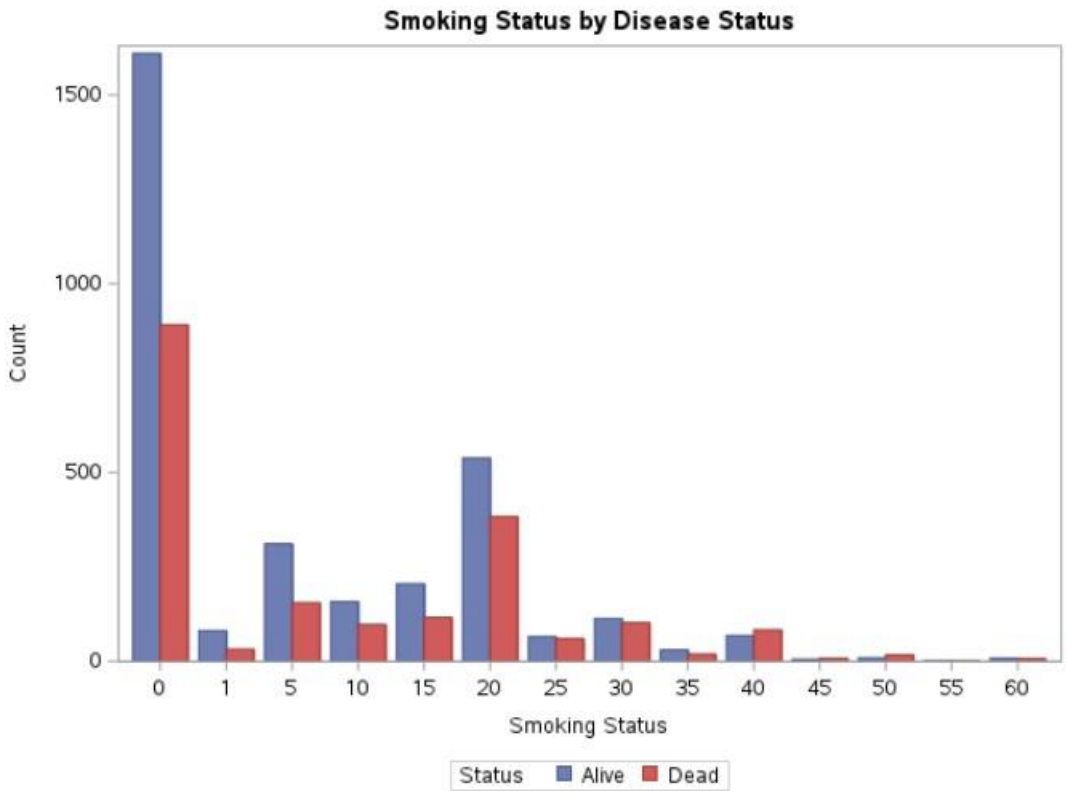
/* Create a box plot of systolic blood pressure (Systolic) by smoking status */ proc sgplot data=Heart; vbox Systolic / category=Smoking; xaxis label="Smoking Status";
yaxis label="Systolic Blood Pressure"; title "Distribution of Systolic Blood Pressure by Smoking Status"; run;



/* Create a box plot of age by disease status */ proc sgplot data=Heart; vbox AgeAtStart / category=Status; xaxis label="Disease Status"; yaxis label="Age"; title "Distribution of Age by Disease Status"; run;



/* Create a bar chart to visualize smoking status distribution */ proc sgplot data=Heart; vbar Smoking / group=Status groupdisplay=cluster; xaxis label="Smoking Status"; yaxis label="Count"; title "Smoking Status by Disease Status"; run;



/* Create a heatmap to visualize correlations between numerical variables */ proc corr data=Heart noprob nosimple; var AgeAtStart Height Weight Diastolic Systolic Cholesterol; run;

proc sgheatmap data=Heart;

matrix Corr;

run;

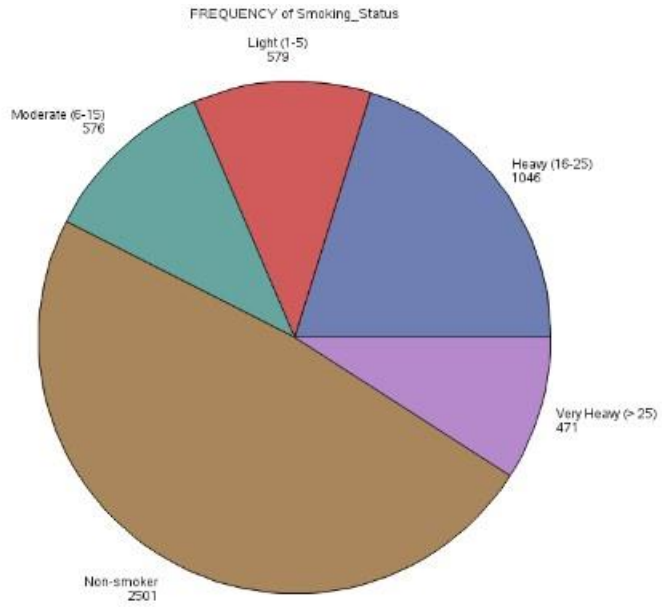
The CORR Procedure

6 Variables:	AgeAtStart Height Weight Diastolic Systolic Cholesterol

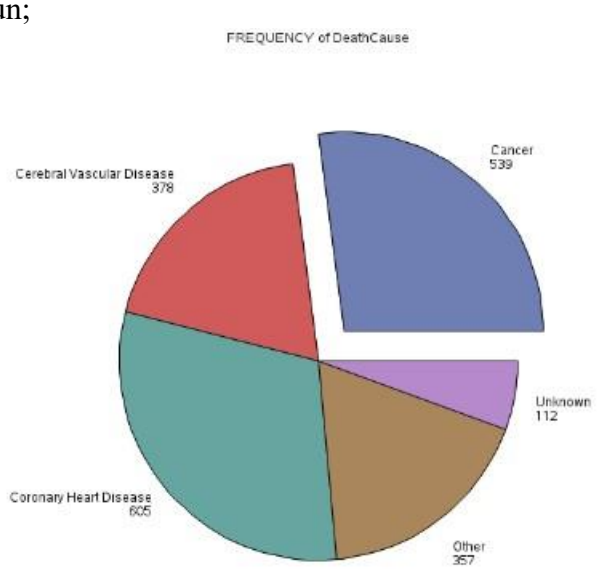
Pearson Correlation Coefficients
Number of Observations

	AgeAtStart	Height	Weight	Diastolic	Systolic	Cholesterol
AgeAtStart Age at Start	1.00000 5209	-0.13173 5203	0.09352 5203	0.27540 5209	0.37938 5209	0.27341 5057
Height	-0.13173 5203	1.00000 5203	0.51739 5199	-0.01425 5203	-0.07113 5203	-0.07959 5051
Weight	0.09352 5203	0.51739 5199	1.00000 5203	0.32757 5203	0.26358 5203	0.07243 5051
Diastolic	0.27540 5209	-0.01425 5203	0.32757 5203	1.00000 5209	0.79606 5209	0.18336 5057
Systolic	0.37938 5209	-0.07113 5203	0.26358 5203	0.79606 5209	1.00000 5209	0.19935 5057
Cholesterol	0.27341 5057	-0.07959 5051	0.07243 5051	0.18336 5057	0.19935 5057	1.00000 5057

/* pie chart for frequencies*/ proc gchart data=heart; pie Smoking_Status;
run;



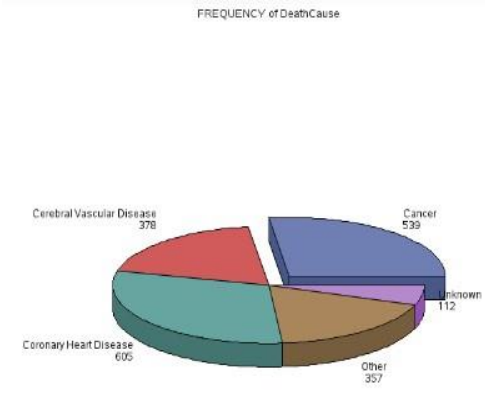
/* pie chart with exploding slice*/ proc gchart data=heart; pie DeathCause / explode='Cancer';
run;

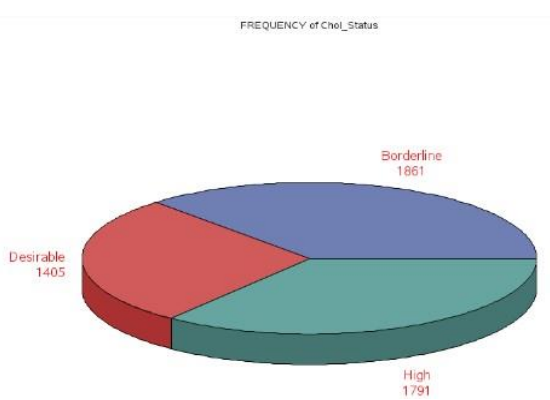


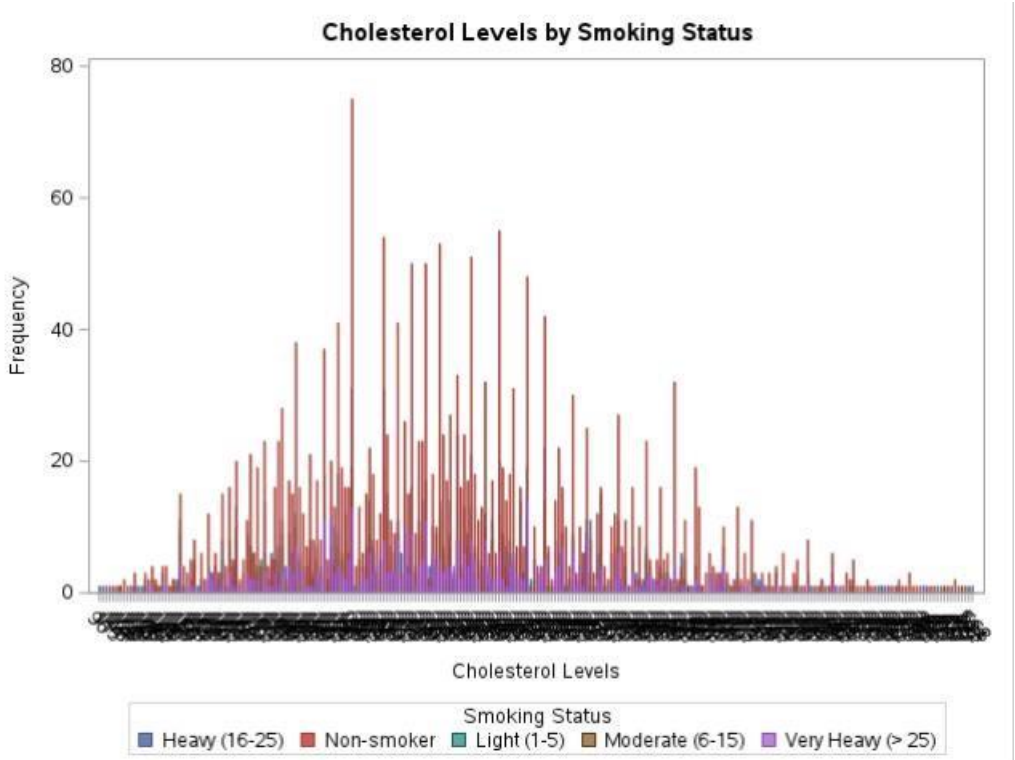
/* 3Dpie chart with exploading slice*/ proc gchart data=heart;
pie3d DeathCause /
explode='Cancer';
run;

proc gchart data=heart; pie3d Chol_Status /
plabel=(h=1.5 color=red);;
run;

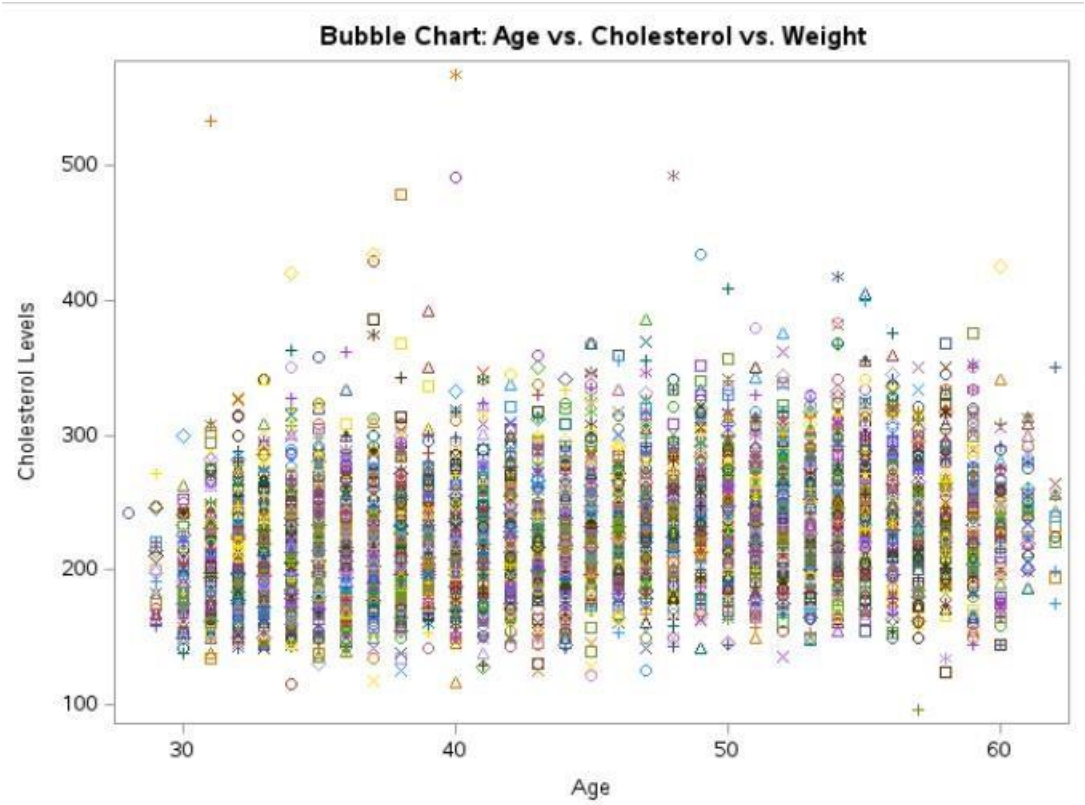
proc sgplot data=heart; vbar Cholesterol /
group=Smoking_Status groupdisplay=cluster; xaxis
label="Cholesterol Levels"; yaxis label="Frequency"; title
"Cholesterol Levels by Smoking Status"; run;







/* Create a bubble chart to visualize Age vs. Cholesterol vs. Weight */ proc sgplot data=heart; scatter x=AgeAtStart y=Cholesterol / group=Weight; xaxis label="Age"; yaxis label="Cholesterol Levels"; title "Bubble Chart: Age vs. Cholesterol vs. Weight"; run;



/* Calculate mean, variance, and standard deviation for selected numerical variables */ proc means data=Sashelp.Heart mean var std;

var AgeAtStart Cholesterol Systolic Diastolic;

run;

The MEANS Procedure

Variable	Label	Mean	Variance	Std Dev
AgeAtStart	Age at Start	44.0687272	73.5298379	8.5749541
Cholesterol		227.4174412	2019.20	44.9355238
Systolic		136.9095796	563.5684355	23.7395964
Diastolic		85.3586101	168.3010976	12.9730913

/* Frequency analysis for categorical variables */ proc

freq data=Sashelp.Heart; tables Sex Smoking_Status

DeathCause / nocum nopercent;

run;

The FREQ Procedure

Sex	Frequency
Female	2873
Male	2336

Smoking Status	
Smoking_Status	Frequency
Heavy (16-25)	1046
Light (1-5)	579
Moderate (6-15)	576
Non-smoker	2501
Very Heavy (> 25)	471
Frequency Missing = 36	

Cause of Death	
DeathCause	Frequency
Cancer	539
Cerebral Vascular Disease	378
Coronary Heart Disease	605
Other	357
Unknown	112
Frequency Missing = 3218	

/* Descriptive statistics for categorical variables */ proc freq data=Sashelp.Heart; tables Sex Smoking_Status DeathCause / norow nocol nopercent; run;

The FREQ Procedure

Sex	Frequency	Cumulative Frequency
Female	2873	2873
Male	2336	5209

Smoking Status		
Smoking_Status	Frequency	Cumulative Frequency
Heavy (16-25)	1046	1046
Light (1-5)	579	1625
Moderate (6-15)	576	2201
Non-smoker	2501	4702
Very Heavy (> 25)	471	5173
Frequency Missing = 36		

Cause of Death		
DeathCause	Frequency	Cumulative Frequency
Cancer	539	539
Cerebral Vascular Disease	378	917
Coronary Heart Disease	605	1522
Other	357	1879
Unknown	112	1991
Frequency Missing = 3218		

/* Create a contingency table and perform a chi-square test */ proc freq data=Sashelp.Heart; tables Sex * Smoking_Status / chisq; run;

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Sex by Smoking_Status

Sex	Smoking_Status(Smoking Status)					
	Heavy (16-25)	Light (1-5)	Moderate (6-15)	Non-smoker	Very Heavy (> 25)	Total
Female	339 6.55 11.87 32.41	422 8.16 14.78 72.88	340 6.57 11.90 59.03	1682 32.51 58.89 67.25	73 1.41 2.56 15.50	2856 55.21
Male	707 13.67 30.51 67.59	157 3.03 6.78 27.12	236 4.56 10.19 40.97	819 15.83 35.35 32.75	398 7.69 17.18 84.50	2317 44.79
Total	1046 20.22	579 11.19	576 11.13	2501 48.35	471 9.10	5173 100.00

Frequency Missing = 36

Statistics for Table of Sex by Smoking_Status

Statistic	DF	Value	Prob
Chi-Square	4	743.4890	<.0001
Likelihood Ratio Chi-Square	4	771.5109	<.0001
Mantel-Haenszel Chi-Square	1	40.7641	<.0001
Phi Coefficient		0.3791	
Contingency Coefficient		0.3545	
Cramer's V		0.3791	

Sample Size = 5173
Frequency Missing = 36

/* Cross-tabulation with percentages */

```
proc freq data=Sashelp.Heart;  tables Sex *
Smoking_Status / nocol nopercent;
run;
```

The FREQ Procedure

Frequency
Row Pct

Table of Sex by Smoking_Status

Sex	Smoking_Status(Smoking Status)					
	Heavy (16-25)	Light (1-5)	Moderate (6-15)	Non-smoker	Very Heavy (> 25)	Total
Female	339 11.87	422 14.78	340 11.90	1682 58.89	73 2.56	2856
Male	707 30.51	157 6.78	236 10.19	819 35.35	398 17.18	2317
Total	1046	579	576	2501	471	5173

Frequency Missing = 36

/* Perform a two-sample t-test for Age by Gender */

```
proc ttest data=Sashelp.Heart;
class Sex;  var
AgeAtStart;
run;
```

The TTEST Procedure

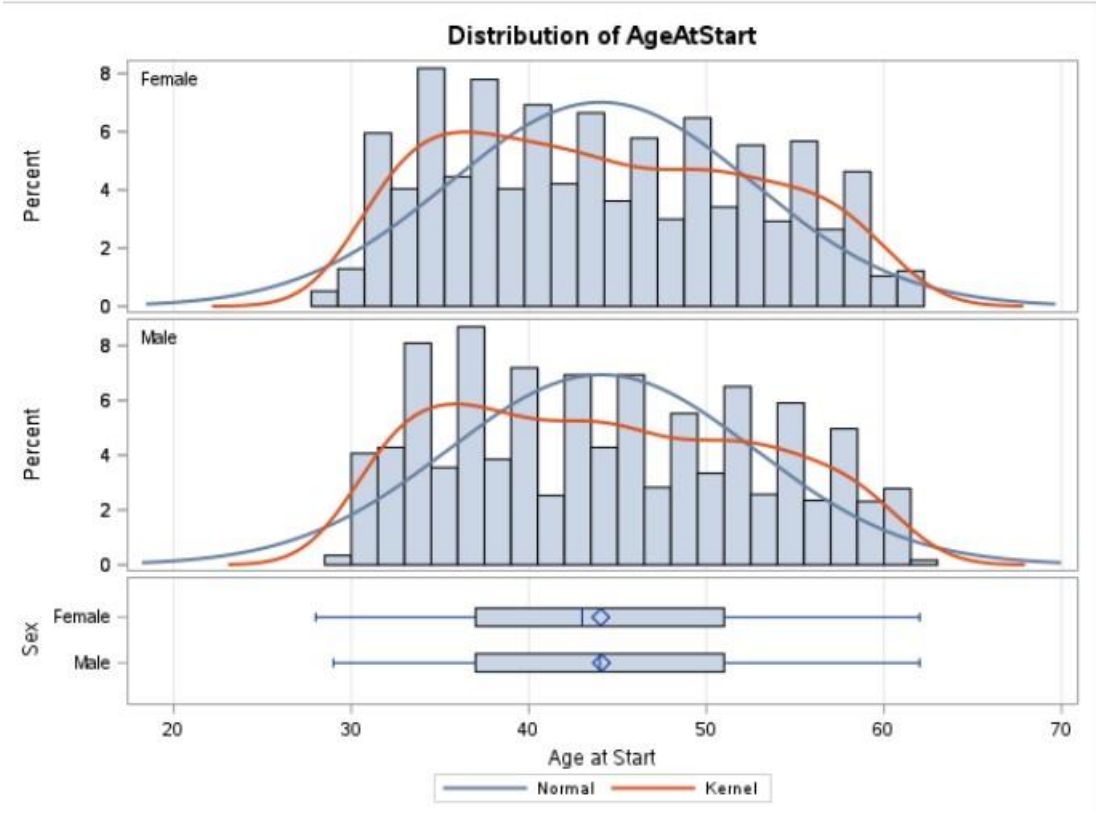
Variable: AgeAtStart (Age at Start)

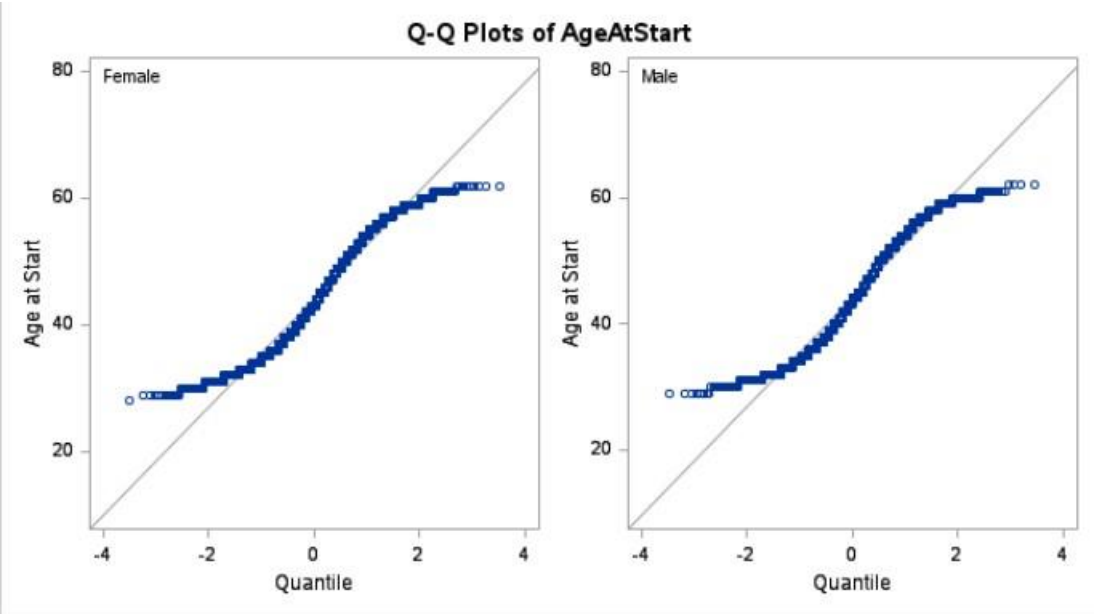
Sex	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female		2873	44.0515	8.5349	0.1592	28.0000	62.0000
Male		2336	44.0899	8.6258	0.1785	29.0000	62.0000
Diff (1-2)	Pooled		-0.0384	8.5758	0.2389		
Diff (1-2)	Satterthwaite		-0.0384		0.2392		

Sex	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		44.0515	43.7393	44.3637	8.5349	8.3198	8.7615
Male		44.0899	43.7399	44.4399	8.6258	8.3853	8.8805
Diff (1-2)	Pooled	-0.0384	-0.5068	0.4300	8.5758	8.4142	8.7437
Diff (1-2)	Satterthwaite	-0.0384	-0.5073	0.4305			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	5207	-0.16	0.8724
Satterthwaite	Unequal	4971.1	-0.16	0.8725

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2335	2872	1.02	0.5898





/* Perform a one-way ANOVA for Cholesterol by Smoking Status */ proc anova data=Sashelp.Heart; class Smoking_Status; model Cholesterol = Smoking_Status; run;

/* code to view the contents of sashelp.heart data*/ proc contents data=sashelp.heart; run;

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
Smoking_Status	5	Heavy (16-25) Light (1-5) Moderate (6-15) Non-smoker Very Heavy (> 25)

Number of Observations Read	5209
Number of Observations Used	5049

The ANOVA Procedure

Dependent Variable: Cholesterol

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	22345.12	5586.28	2.77	0.0257
Error	5044	10168843.77	2016.03		
Corrected Total	5048	10191188.89			

R-Square	Coeff Var	Root MSE	Cholesterol Mean
0.002193	19.74114	44.90020	227.4448

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Smoking_Status	4	22345.11719	5586.27930	2.77	0.0257

The CONTENTS Procedure

Data Set Name	SASHELP.HEART	Observations	5209
Member Type	DATA	Variables	17
Engine	V9	Indexes	0
Created	08/06/2020 06:41:21	Observation Length	168
Last Modified	08/06/2020 06:41:21	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Framingham Heart Study		
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	us-ascii ASCII (ANSI)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	14
First Data Page	1
Max Obs per Page	389
Obs in First Data Page	365
Number of Data Set Repairs	0
Filename	/pbr/sfw/sas/940/SASFoundation/9.4/sashelp/heart.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	134873665
Access Permission	rw-r--r--
Owner Name	odaowner
File Size	960KB
File Size (bytes)	983040

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Label
12	AgeAtDeath	Num	8	Age at Death
5	AgeAtStart	Num	8	Age at Start
3	AgeCHDdiag	Num	8	Age CHD Diagnosed
15	BP_Status	Char	7	Blood Pressure Status
14	Chol_Status	Char	10	Cholesterol Status
13	Cholesterol	Num	8	
2	DeathCause	Char	26	Cause of Death
8	Diastolic	Num	8	
6	Height	Num	8	
10	MRW	Num	8	Metropolitan Relative Weight
4	Sex	Char	6	
11	Smoking	Num	8	
17	Smoking_Status	Char	17	Smoking Status
1	Status	Char	5	
9	Systolic	Num	8	
7	Weight	Num	8	
16	Weight_Status	Char	11	Weight Status

Data Overview:

We loaded the Heart Study data from the sashelp.heart dataset.
then printed the data to get a general understanding of its structure. **Age Distribution:**

we created a histogram of patient ages, showing the distribution of ages at the study's start. The histogram suggests that the study included a wide range of ages, with a higher concentration of participants in their 40s and 50s. **Disease Status by Gender:**

We created a bar chart to visualize the distribution of disease status by gender.
The chart provides insights into how disease status varies between male and female participants.
Systolic Blood Pressure by Smoking Status:

We created a box plot to visualize the distribution of systolic blood pressure based on smoking status.
This analysis shows whether there are differences in systolic blood pressure between smokers and non-smokers.
Age by Disease Status:

We created a box plot to visualize the distribution of age based on disease status.
This analysis helps understand how age relates to disease status in the study population.
Smoking Status by Disease Status:

We created a bar chart to visualize the distribution of smoking status by disease status.

It shows the proportion of smokers and non-smokers among individuals with different disease statuses.

Correlation Analysis:

We performed a correlation analysis to explore relationships between numerical variables. The heatmap visually represents the correlations between variables such as age, height, weight, blood pressure, and cholesterol. **Pie Charts:**

We created pie charts to visualize the distribution of smoking status and causes of death. These charts provide a quick overview of the proportions within these categories. **3D Pie Charts:**

We created 3D pie charts, including an exploding slice for the "Cancer" category in the cause of death data.

These charts add a visual element to the distribution of causes of death, emphasizing specific categories.

Cholesterol Levels by Smoking Status:

We created a bar chart to explore how cholesterol levels vary with smoking status.

This chart can help identify differences in cholesterol levels between smokers and nonsmokers.

Bubble Chart:

We created a bubble chart to visualize relationships between age, cholesterol levels, and weight. It can be used to identify patterns or clusters within these three variables. **Descriptive Statistics:**

We calculated mean, variance, and standard deviation for selected numerical variables.

This provides a summary of the central tendency and variability of key variables.

Frequency Analysis:

We conducted frequency analyses for categorical variables like sex, smoking status, and causes of death.

These analyses give insight into the distribution of categorical variables. **Chi-Square Test:**

We performed a chi-square test to explore the relationship between sex and smoking status.

This test assesses whether there's a significant association between the two variables. **Two-Sample T-Test:**

We conducted a two-sample t-test to compare the means of age between genders.

It helps determine whether there's a statistically significant difference in age between males and females.

One-Way ANOVA:

We performed a one-way ANOVA to examine the effect of smoking status on cholesterol levels. It tests whether there are statistically significant differences in cholesterol levels among different smoking groups.

In conclusion, our analysis of the Heart Study data in SAS provides valuable insights into various aspects of the dataset, including demographics, health measures, and relationships between variables. The specific findings and conclusions will depend on the results generated by each analysis, and further interpretation may be necessary for a comprehensive understanding of the data.
