PH 538: Biostatistical Methods I

SAS: Lab 1 (Descriptive Statistics)

Dr. Fares Qeadan: fqeadan@salud.unm.edu

Objectives:

In this lab students will learn how to use SAS to describe numerical (quantitative) and categorical (qualitative) variables both numerically and graphically.

Background on the data set:

In this Lab, we will be using the *Diabetes and obesity, cardiovascular risk factors* data set. This data set includes 403 African Americans who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia.

The list of variables in the data set:

	Variable	Description
1	id	Subject ID
2	chol	Total Cholesterol
3	stab_glu	Stabilized Glucose
4	hdl	High Density Lipoprotein
5	ratio	Cholesterol/HDL Ratio
6	glyhb	Glycosylated Hemoglobin (A1C)
7	location	County - a factor with levels Buckingham and Louisa
8	age	age in years
9	gender	a factor with levels male and female
10	height	height in inches
11	weight	weight in pounds
12	frame	a factor with levels small, medium and large
13	bp_1s	First Systolic Blood Pressure
14	bp_1d	First Diastolic Blood Pressure
15	waist	waist in inches
16	hip	hip in inches
17	diab	Diabetes status

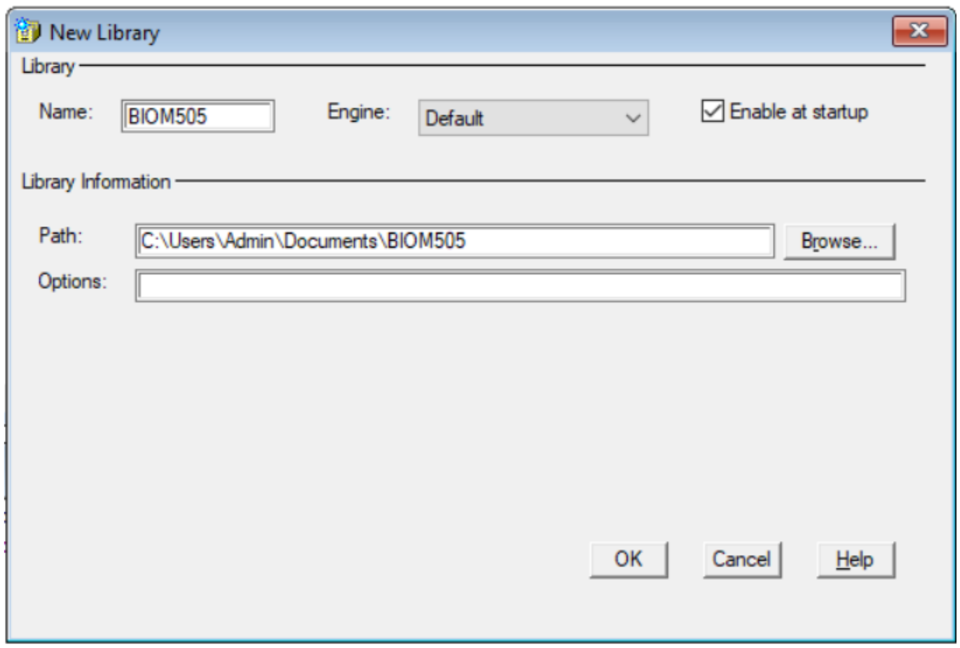
Things to do before starting the analysis of the data

1. **Create a SAS Permanent Library:**

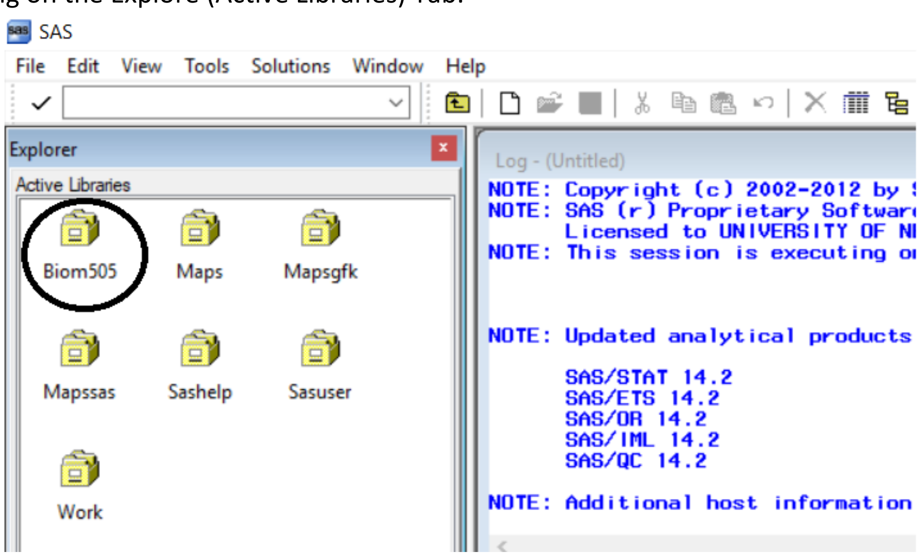
SAS has permanent libraries and one temporary library (the Work library). To create a SAS library, please use the following steps:

a) Firstly, click the button at the top that looks like  and then proceed to step (b)

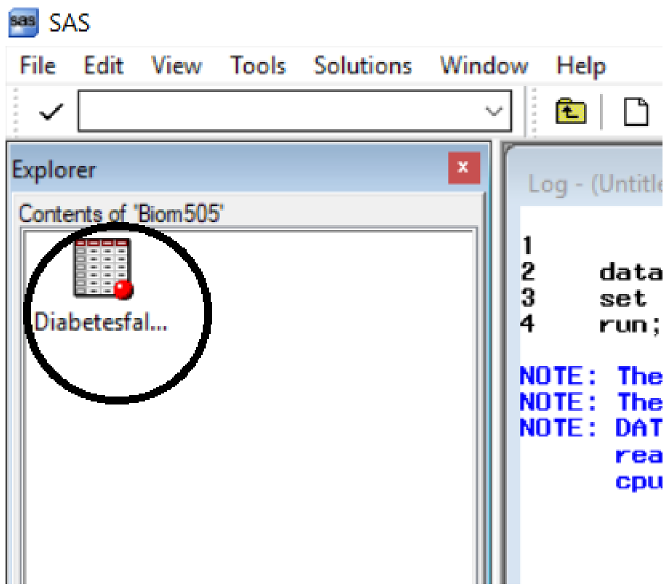
b) Give a name for your new permanent library (e.g. BIOM505), check the box “Enable at startup”, and use “Browse” to specify the path for the folder in which all SAS datasets will be stored (note that the name of the specified folder and that of the library don’t necessarily have to be the same).



c) Click “OK” on the above New Library Window. To verify that BIOM505 was created, you should see the following on the Explore (Active Libraries) Tab:



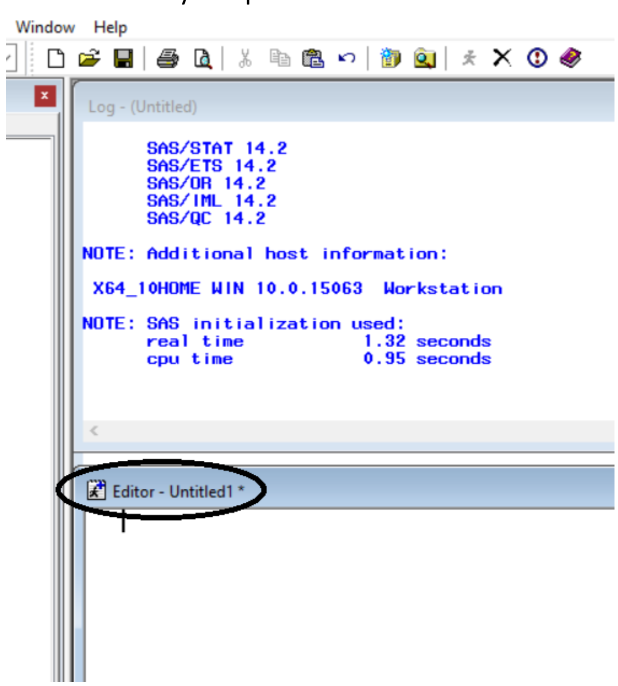
d) Assuming that you have downloaded the *Diabetes and obesity, cardiovascular risk factors* data set into the specified folder in 1 (b) from the class website at: http://www.mathalpha.com/lab1/diabetesfall17.sas7bdat, then if you double click on the newly created library BIOM505 you should be able to see the **diabetesfall17.sas7bdat** dataset:



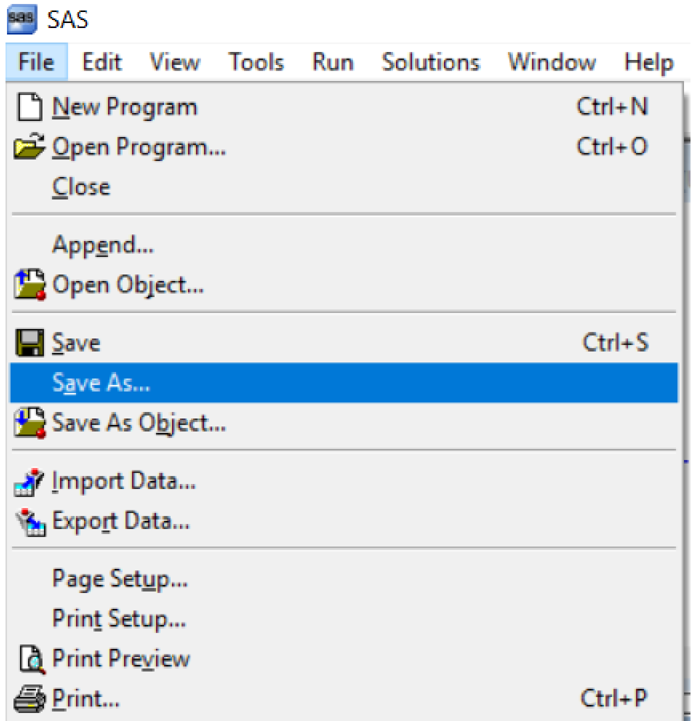
2. Create a SAS program:

To reuse your work, you need to save your SAS syntax into a file. SAS uses program files for this purpose where SAS programs are simply text files whose names end with .sas. There are several ways to create a SAS program as follows:

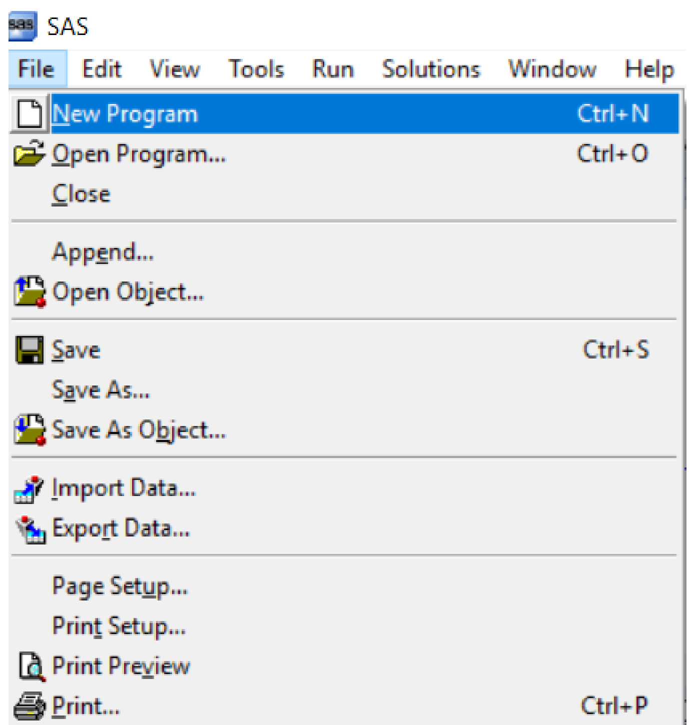
a) Use the SAS **`Editor`** available when you open SAS:



To SAVE your SAS program (call it Lab1) in a desired folder, Click on or from the drop-down menu, click on **File** and then select **save as** to save your file under any name and location you like, say **lab1** and save it at the BIOM505 folder in your computer.



b) A second way to create a new SAS program is to firstly, click the File drop-down menu and click on New Program and then repeat 2(a).



Your first SAS program will involve using data-step programing. Specifically, we need to use **data** and **set** to make a copy of the **diabetesfall17.sas7bdat** dataset from the **BIOM505 library** into the **Work library**:

lab1

```
data diabetes;
 set biom505.diabetesfall17;
 run;
```

Data Analysis

Before initiating any descriptive statistics, let's identify the variable names in the dataset and their type by using **proc contents**. To save the contents of the dataset, we could use the SAS Output Delivery System (ODS):

```
ods rtf;
proc contents data=diabetes;
run;
ods rtf close;
```

The output that you should be getting is:

The SAS System

The CONTENTS Procedure

Data Set Name	WORK.DIABETES	Observations	403
Member Type	DATA	Variables	17
Engine	V9	Indexes	0
Created	09/19/2017 00:55:42	Observation Length	136
Last Modified	09/19/2017 00:55:42	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	481
Obs in First Data Page	403
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\Admin\AppData\Local\Temp\SAS Temporary Files_TD2644_LENOVO-NOTEBOOK_\diabetes.sas7bdat
Release Created	9.0401M4
Host Created	X64_10HOME
Owner Name	BUILTIN\Administrators
File Size	128KB
File Size (bytes)	131072

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
8	age	Num	8			age
14	bp_1d	Num	8			bp_1d
13	bp_1s	Num	8			bp_1s
2	chol	Num	8			chol
17	diab	Num	8			diab
12	frame	Char	6	$6.	$6.	frame
9	gender	Char	6	$6.	$6.	gender
6	glyhb	Num	8			glyhb
4	hdl	Num	8			hdl
10	height	Num	8			height
16	hip	Num	8			hip
1	id	Num	8			id
7	location	Char	10	$10.	$10.	location
5	ratio	Num	8			ratio
3	stab_glu	Num	8			stab_glu
15	waist	Num	8			waist
11	weight	Num	8			weight

1. Numerical Descriptive Statistics for Numerical (Quantitative) Variables:

We will describe the Glycosylated Hemoglobin (A1C) variable *[and other variables]* numerically by providing the following sample statistics:

n, Mean, Median, Mode, Standard deviation (or Variance), Q1, Q3, IQR, Min, Max, Range, Mode

To accomplish this task, we use either PROC MEANS, PROC SUMMARY, or PROC UNIVARIATE

PROC MEANS:

```
proc means data=diabetes n Mean Median Mode Std var Q1 Q3 Min Max Range Mode skew kurt p25 p75 qrange;
var glyhb;
run;
```

The SAS System

The MEANS Procedure

Analysis Variable : glyhb glyhb															
N	Mean	Median	Mode	Std Dev	Variance	Lower Quartile	Upper Quartile	Minimum	Maximum	Range	Skewness	Kurtosis	25th Pctl	75th Pctl	Quartile Range
390	5.5897692	4.8400000	4.4000000	2.2425948	5.0292316	4.3800000	5.6000000	2.6800000	16.1100000	13.4300000	2.2461247	5.1064863	4.3800000	5.6000000	1.2200000

PROC UNIVARIATE:

```
proc univariate data=diabetes;
var glyhb;
run;
```

The SAS System

The UNIVARIATE Procedure
Variable: glyhb (glyhb)

Moments			
N	390	Sum Weights	390
Mean	5.58976923	Sum Observations	2180.01
Std Deviation	2.24259483	Variance	5.02923157
Skewness	2.24612468	Kurtosis	5.10648627
Uncorrected SS	14142.1239	Corrected SS	1956.37108
Coeff Variation	40.1196317	Std Error Mean	0.1135582

Basic Statistical Measures			
Location		Variability	
Mean	5.589769	Std Deviation	2.24259
Median	4.840000	Variance	5.02923
Mode	4.400000	Range	13.43000
		Interquartile Range	1.22000

Note: The mode displayed is the smallest of 2 modes with a count of 6.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	49.22383	Pr > |t|	<.0001
Sign	M	195	Pr >= |M|	<.0001
Signed Rank	S	38122.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	16.11
99%	14.31
95%	10.93
90%	8.99
75% Q3	5.60
50% Median	4.84
25% Q1	4.38
10%	4.00
5%	3.75
1%	2.85
0% Min	2.68

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2.68	37	13.70	59
2.73	337	14.31	63
2.85	321	14.94	363
2.85	305	15.52	33
3.03	308	16.11	399

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	13	3.23	100.00

PROC SUMMARY:

```
proc summary data=diabetes print n Mean Median Mode Std var Q1 Q3 Min Max Range Mode skew kurt p25 p75 qrange;
var glyhb;
run;
```

The SAS System

The SUMMARY Procedure

Analysis Variable : glyhb glyhb															
N	Mean	Median	Mode	Std Dev	Variance	Lower Quartile	Upper Quartile	Minimum	Maximum	Range	Skewness	Kurtosis	25th Pctl	75th Pctl	Quartile Range
390	5.5897692	4.8400000	4.4000000	2.2425948	5.0292316	4.3800000	5.6000000	2.6800000	16.1100000	13.4300000	2.2461247	5.1064863	4.3800000	5.6000000	1.2200000

Remark: One could describe more than one variable at a time as follows:

```
proc means data=diabetes;
var glyhb hip stab_glu chol hdl;
run;
```

The SAS System

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
glyhb	glyhb	390	5.5897692	2.2425948	2.6800000	16.1100000
hip	hip	401	43.0399002	5.6567132	30.0000000	64.0000000
stab_glu	stab_glu	403	106.6724566	53.0766545	48.0000000	385.0000000
chol	chol	402	207.8457711	44.4455574	78.0000000	443.0000000
hdl	hdl	402	50.4452736	17.2626257	12.0000000	120.0000000

Remark: One could also describe numerical variables within the levels of categorical variables as follows:

```
proc means data=diabetes;
class frame;
var glyhb;
run;
```

The SAS System

The MEANS Procedure

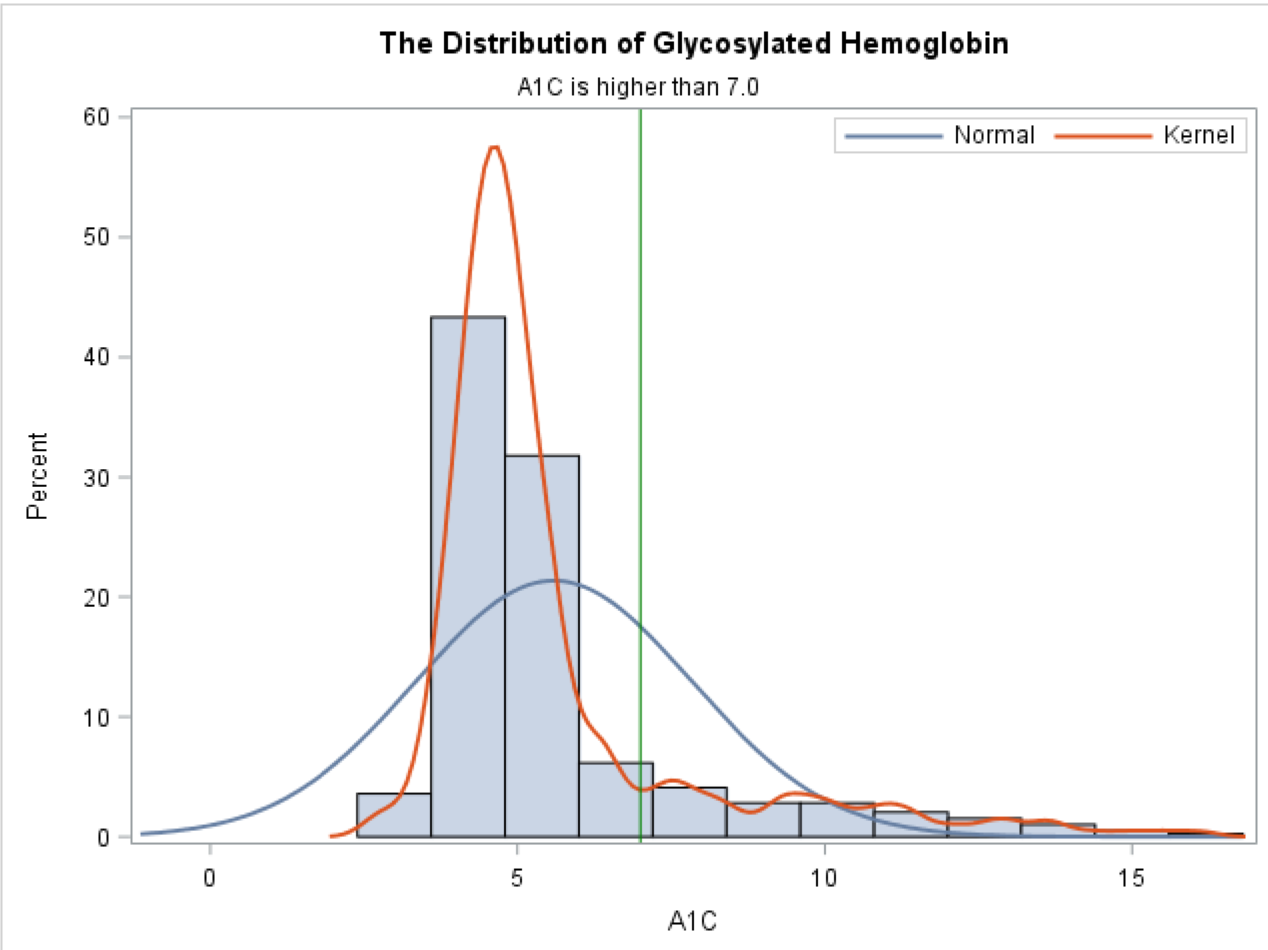
Analysis Variable : glyhb glyhb						
frame	N Obs	N	Mean	Std Dev	Minimum	Maximum
large	103	99	6.1056566	2.2455353	3.5800000	13.7000000
medium	184	178	5.6402809	2.4381130	2.6800000	16.1100000
small	104	102	5.0408824	1.8023824	2.8500000	13.6300000

2. Graphical Descriptive Statistics for Numerical (Quantitative) Variables:

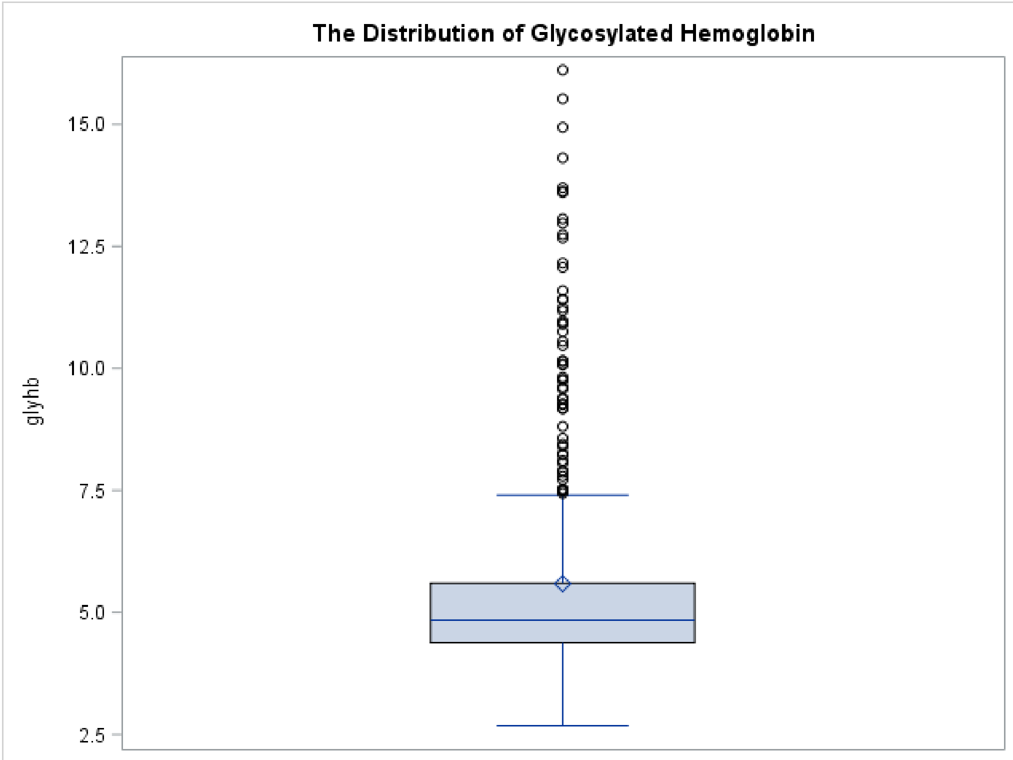
We will describe the Glycosylated Hemoglobin (A1C) variable *[and other variables]* graphically by providing the following presentations:

Histogram, Box-plot, Stem and leaf and Scatter plot.

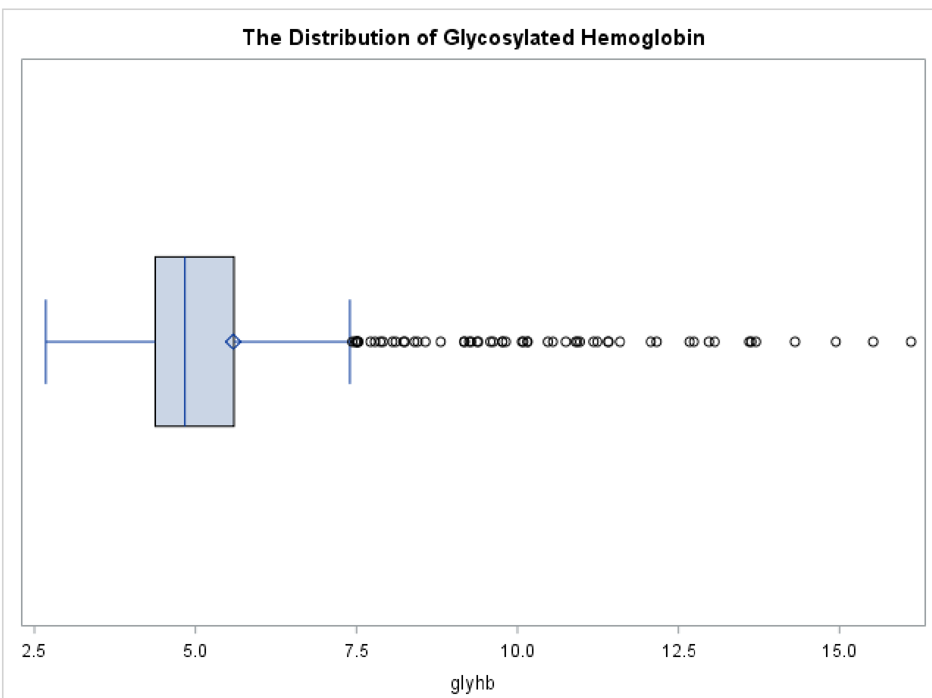
```
proc sgplot data=diabetes;
histogram glyhb;
density glyhb;
density glyhb / type=kernel;
keylegend / location=inside position=topright;
Title "The Distribution of Glycosylated Hemoglobin";
xaxis label="A1C";
refline 7 /axis=x label=" A1C is higher than 7.0" lineattrs=(color=green);
run;
```



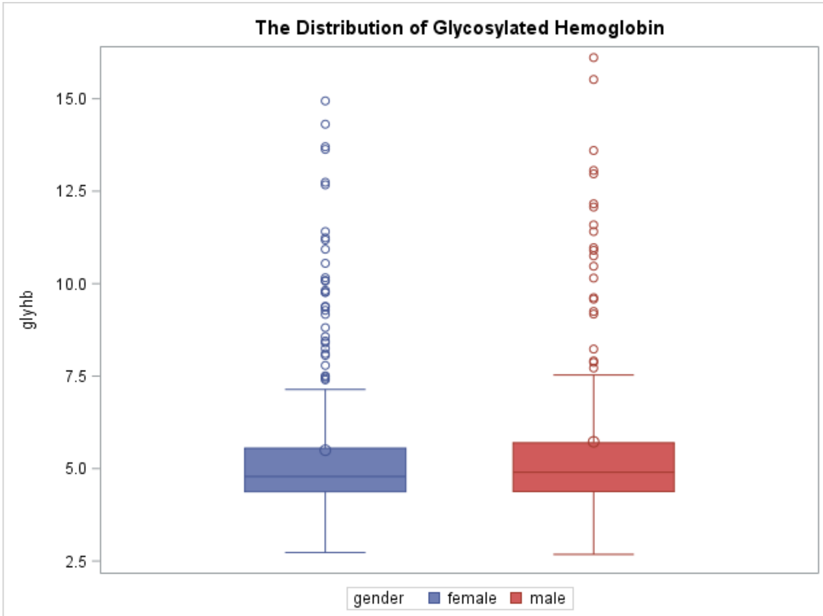

```
proc sgplot data=diabetes;
vbox glyhb;
*hbox  glyhb;
run;
```




```
proc sgplot data=diabetes;
*vbox glyhb;
hbox  glyhb;
run;
```



```
proc sgplot data=diabetes;
vbox glyhb/group=gender;
run;
```



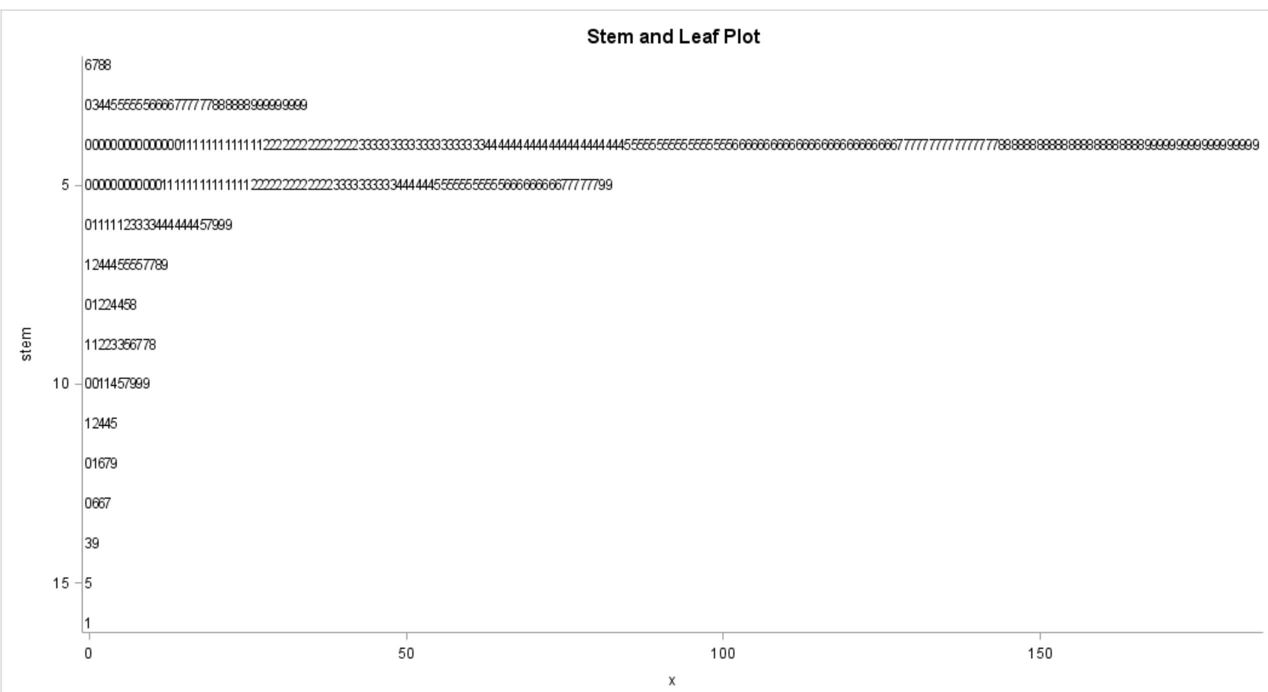

```
ods graphics off;
ods select Plots SSPlots;
proc univariate data=diabetes plot;
var glyhb;
run;
```

Stem and Leaf Plot

The UNIVARIATE Procedure
Variable: glyhb (glyhb)

```
                        Histogram                            #          Boxplot
   16.5+*                                                    1             *
       .*                                                    1             *
       .*                                                    2             *
       .*                                                    4             *
       .**                                                   5             *
       .**                                                   5             *
       .***                                                 10             *
    9.5+***                                                 11             0
       .**                                                   8             0
       .****                                                13             0
       .******                                              23             |
       .*********************                               83          +--+--+
       .***********************************************    185          *-----*
       .*********                                           35             |
    2.5+*                                                    4             |
        ----+----+----+----+----+----+----+----+----+--
        * may represent up to 4 counts
```

```
data diabetes;
set diabetes;
stem=floor(glyhb);
leaf=floor((glyhb-stem)*10);
run;
proc sort data=diabetes out=stemleafSort;
by glyhb;
run;
data stemleafGraph;
  set stemleafSort;
  by stem;
  zero=0;
  retain x 0;
    if first.stem then x=0;
    else x+1;
run;
ods graphics on / width=11in height=6in;
title 'Stem and Leaf Plot';
proc sgplot data=stemleafGraph noautolegend noborder;
  text x=x y=stem text=leaf / textattrs=(size=9) strip;
  yaxis reverse;
run;
```



3. Numerical Descriptive Statistics for Categorical (Qualitative) Variables:

We will describe the Diabetes status variable *[and other variables]* numerically by providing the frequencies and relative frequencies through contingency tables:

```
proc freq data=diabetes;
table diab;
run;
```

The FREQ Procedure

diab				
diab	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	330	84.62	330	84.62
1	60	15.38	390	100.00
Frequency Missing = 13				

Note that we could also find the sample proportion of diabetes by gender as follows:

```
proc freq data=diabetes;
table gender*diab/chisq;
run;
```

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of gender by diab

gender(gender)	diab(diab) 0	1	Total
female	194 49.74 85.09 58.79	34 8.72 14.91 56.67	228 58.46
male	136 34.87 83.95 41.21	26 6.67 16.05 43.33	162 41.54
Total	330 84.62	60 15.38	390 100.00
Frequency Missing = 13			

Statistics for Table of gender by diab

Statistic	DF	Value	Prob
Chi-Square	1	0.0941	0.7591
Likelihood Ratio Chi-Square	1	0.0938	0.7594
Continuity Adj. Chi-Square	1	0.0270	0.8695
Mantel-Haenszel Chi-Square	1	0.0938	0.7594
Phi Coefficient		0.0155	
Contingency Coefficient		0.0155	
Cramer's V		0.0155	

How to read the Table on the left?

Here is a correct statement: 14.91% of females were found to have diabetes

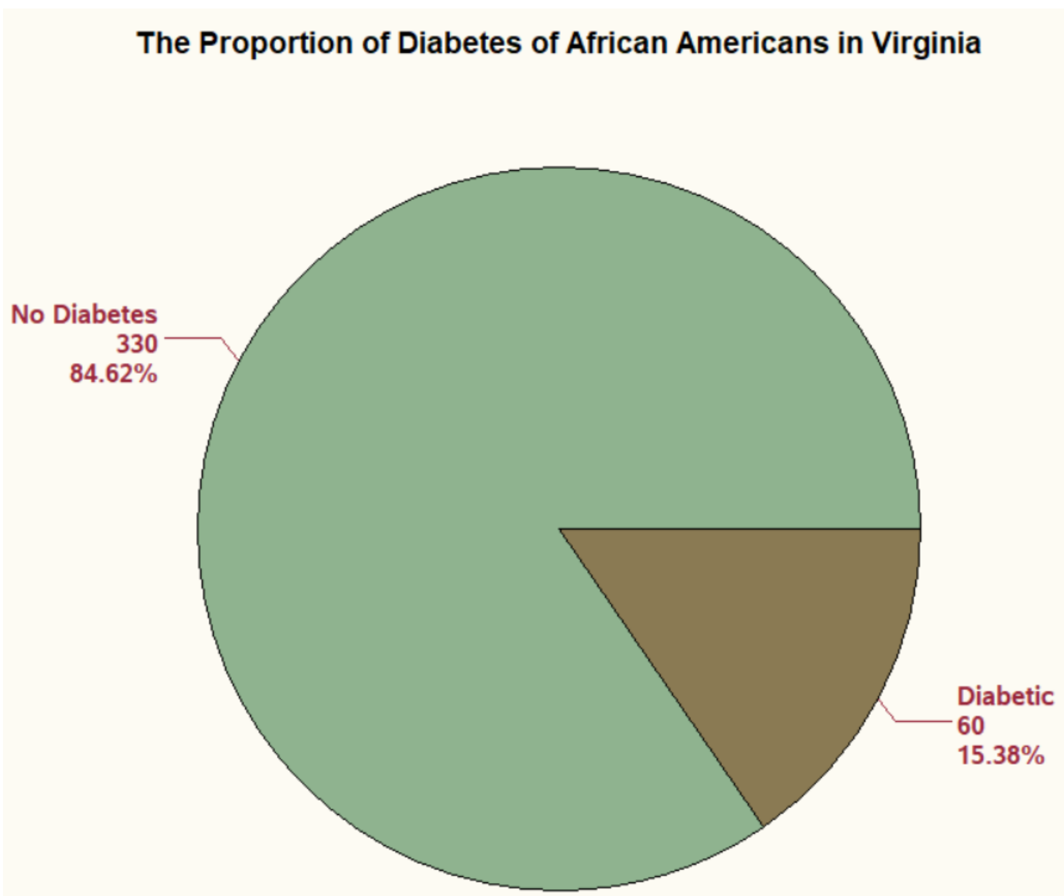
Here is a correct statement: 56.67% of subjects with diabetes were females.

Remark: Note that there are three different percentages one could obtain, the total one, the row one and the column one and each one of them has a different denominator and hence different interpretation.

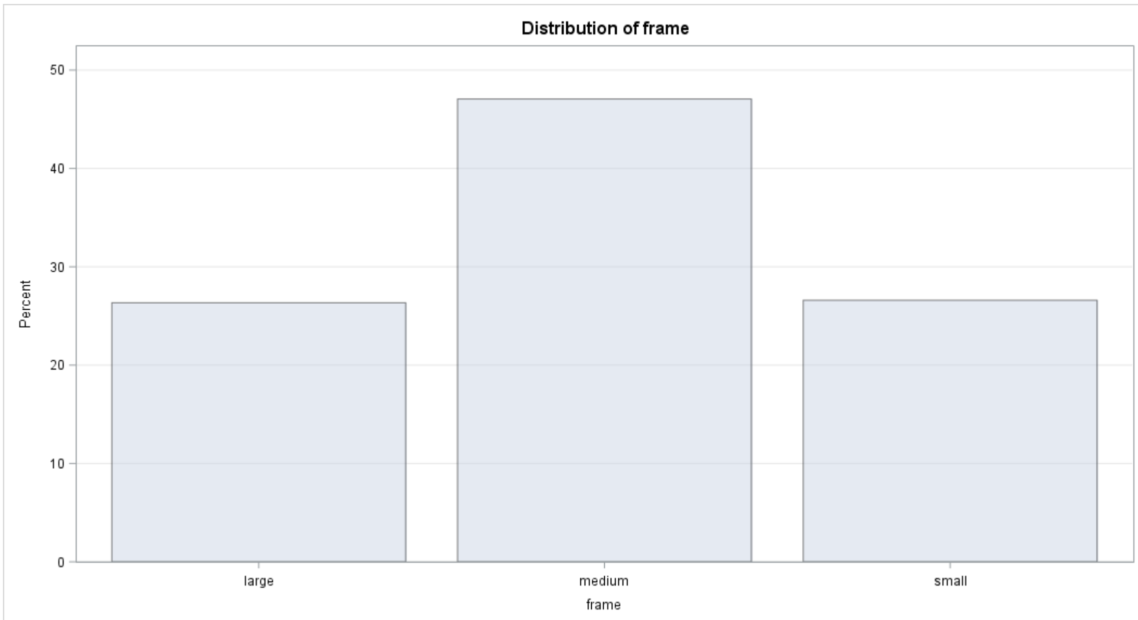
4. Graphical Descriptive Statistics for Categorical (Qualitative) Variables:

We will describe the Diabetes status variable *[and other variables]* graphically by providing the pie and bar charts:

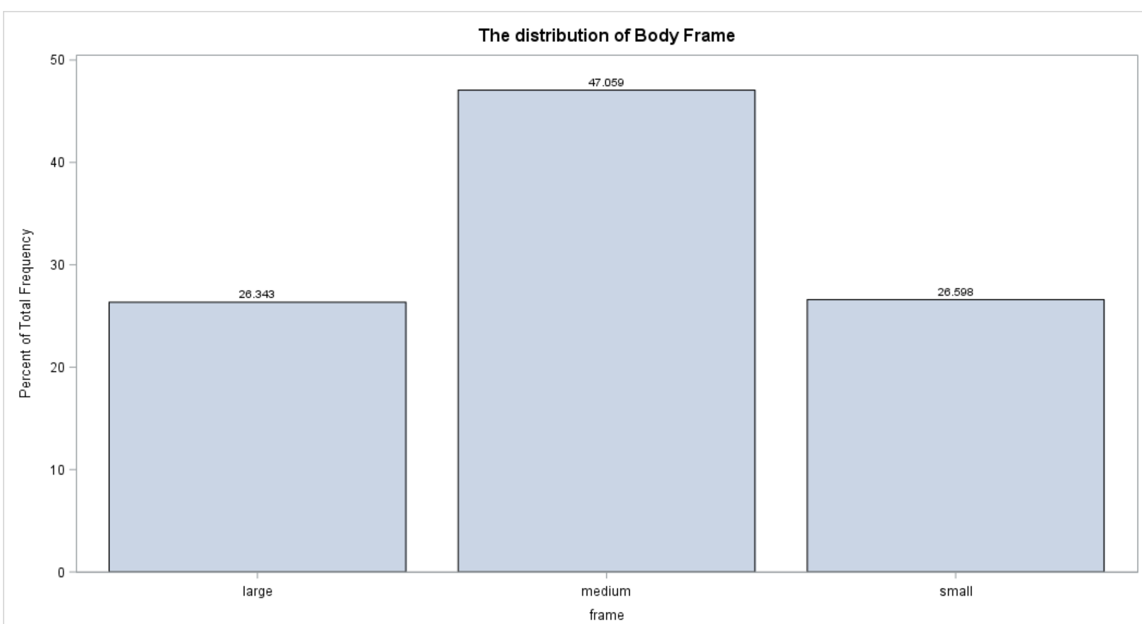
```
proc format;
value diab
0="No Diabetes"
1="Diabetic";
run;

proc gchart data=diabetes;
title "The Proportion of Diabetes of African Americans in Virginia";
   pie diab / percent=arrow
              slice=arrow
              noheading
              plabel=(font='Albany AMT/bold' h=1.3 color=depk);
format diab diab.;
run;
quit;
```



```
title "The distribution of Body Frame";
ods graphics on;
proc freq data=diabetes;
tables frame / plots=FreqPlot(scale=Percent) out=Freq1Out; /* save Percent variable */
run;
```




```
proc sgplot data=freq1out;
vbar frame/response=percent datalabel;
run;
```



Data Management:

1. Please create a BMI variable from the given weight and height variables?

```
data diabetes;
set diabetes;
BMI=(weight/(height*height))* 703;
RUN;
```

2. Please create a BMI categorical variable from the BMI numeric one? Note that, in public health, BMI for adults is often divided into four categories:
    1. Underweight if BMI<18.5
    2. normal weight if BMI is within [18.5, 25)
    3. overweight if BMI is within [25, 30)
    4. obese if BMI ≥ 30

```
data diabetes;
set diabetes;
if bmi<18.5 & age>=18 then BMI_cat=1;
if bmi>=18.5 & bmi<25 & age>=18 then BMI_cat=2;
if bmi>=25 & bmi<30 & age>=18 then  BMI_cat=3;
if bmi>=30 & age>=18 then BMI_cat=4;
if bmi=. then BMI_cat=.;
run;
```

```
data diabetes;
set diabetes;
if bmi lt 18.5 and age ge 18 then BMI_cat=1;
if bmi ge 18.5 and bmi lt 25 and age ge 18 then BMI_cat=2;
if bmi ge 25 and bmi lt 30 and age ge 18 then  BMI_cat=3;
if bmi ge 30 and age ge 18 then BMI_cat=4;
if bmi=. then BMI_cat=.;
run;
```

3. Get the contingency table for BMI categories and cross tab it with diabetes status?

```
proc freq data=diabetes;
table BMI_cat;
format bmi_cat bmi.;
run;
```

The FREQ Procedure

BMI_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Underweight	9	2.27	9	2.27
Normal weight	113	28.46	122	30.73
Overweight	123	30.98	245	61.71
Obese	152	38.29	397	100.00
Frequency Missing = 6				

```
proc freq data=diabetes;
table BMI_cat*diab/chisq;
format bmi_cat bmi. diab diab.;
run;
```

The FREQ Procedure

Table of BMI_cat by diab

Frequency Percent Row Pct Col Pct	diab(diab)		
BMI_cat	No Diabetes	Diabetic	Total
Underweight	9 2.34 100.00 2.76	0 0.00 0.00 0.00	9 2.34
Normal weight	100 26.04 91.74 30.67	9 2.34 8.26 15.52	109 28.39
Overweight	99 25.78 83.19 30.37	20 5.21 16.81 34.48	119 30.99
Obese	118 30.73 80.27 36.20	29 7.55 19.73 50.00	147 38.28
Total	326 84.90	58 15.10	384 100.00

Frequency Missing = 19

Statistics for Table of BMI_cat by diab

Statistic	DF	Value	Prob
Chi-Square	3	8.3066	0.0401
Likelihood Ratio Chi-Square	3	10.1244	0.0175
Mantel-Haenszel Chi-Square	1	7.7070	0.0055
Phi Coefficient		0.1471	
Contingency Coefficient		0.1455	
Cramer's V		0.1471	

Effective Sample Size = 384
Frequency Missing = 19