DSC 510 Programming Assignment 1
Due Sunday, October 18, 2020 by 11:59 PM
Alex Teboul

The frmgham.csv contains 11,627 patients and 38 variables. The dataset is a version of the Framingham Study and it is a teaching dataset, which is used to explain the outcomes and conditions of patients with various cardiovascular diseases.

Please answer the following questions using SAS/SAS Enterprise Guide, and R for Machine learning parts of problems 3 and 4. If there are issues with missing values, use listwise deletion. Remember for each question, make sure to not only present the syntax and output, but also to explain what the output means in answering the research question.

SAS Initial Work

```
/* #DSC510 Programming Assignment 1 - Alex Teboul */

/* Set Working directory */

LIBNAME LAB '\\apporto.com\dfs\depaul\Users\ateboul_depaul\Documents\DSC510_A1';


/* View what is available in the Library */

PROC CONTENTS DATA=lab._ALL_ NODS;
RUN;

/* Read in Dataset */

PROC IMPORT DATAFILE="\\apporto.com\dfs\depaul\Users\ateboul_depaul\Documents\DSC510_A1\frmgham2.csv"
    OUT=LAB.A1
    DBMS=csv
    REPLACE;
    GETNAMES=YES;
RUN;

/* Check that File was Read in Correctly */

PROC PRINT DATA=LAB.A1; RUN;
```

The SAS System

Obs	RANDID	SEX	TOTCHOL	AGE	SYSBP	DIABP	CURSMOKE	CIGPDAY	BMI	DIABETES	BPMEDS	HEARTRTE	GLUCOSE	educ	PREVCHD	PREVAP	PREVMI	PREVSTRK	PREV
1	2448	1	195	39	106	70	0	0	26.97	0	0	80	77	4	0	0	0	0	
2	2448	1	209	52	121	66	0	0	.	0	0	69	92	4	0	0	0	0	
3	6238	2	250	46	121	81	0	0	28.73	0	0	95	76	2	0	0	0	0	

```
/* Check Structure of the File and Variables */
/* ORDER can order the variables in a variety of ways, varnum orders by variable number */

PROC CONTENTS DATA=LAB.A1 ORDER=varnum; RUN;
```

Variables in Creation Order					
#	Variable	Type	Len	Format	Informat
1	RANDID	Num	8	BEST12.	BEST32.
2	SEX	Num	8	BEST12.	BEST32.
3	TOTCHOL	Num	8	BEST12.	BEST32.
4	AGE	Num	8	BEST12.	BEST32.
5	SYSBP	Num	8	BEST12.	BEST32.
6	DIABP	Num	8	BEST12.	BEST32.
7	CURSMOKE	Num	8	BEST12.	BEST32.
8	CIGPDAY	Num	8	BEST12.	BEST32.
9	BMI	Num	8	BEST12.	BEST32.
10	DIABETES	Num	8	BEST12.	BEST32.
11	BPMEDS	Num	8	BEST12.	BEST32.
12	HEARTRTE	Num	8	BEST12.	BEST32.
13	GLUCOSE	Num	8	BEST12.	BEST32.
14	educ	Num	8	BEST12.	BEST32.
15	PREVCHD	Num	8	BEST12.	BEST32.
16	PREVAP	Num	8	BEST12.	BEST32.
17	PREVMI	Num	8	BEST12.	BEST32.
18	PREVSTRK	Num	8	BEST12.	BEST32.
19	PREVHYP	Num	8	BEST12.	BEST32.
20	TIME	Num	8	BEST12.	BEST32.
21	PERIOD	Num	8	BEST12.	BEST32.
22	HDLC	Num	8	BEST12.	BEST32.
23	LDLC	Num	8	BEST12.	BEST32.
24	DEATH	Num	8	BEST12.	BEST32.
25	ANGINA	Num	8	BEST12.	BEST32.
26	HOSPMI	Num	8	BEST12.	BEST32.
27	MI_FCHD	Num	8	BEST12.	BEST32.
28	ANYCHD	Num	8	BEST12.	BEST32.
29	STROKE	Num	8	BEST12.	BEST32.
30	CVD	Num	8	BEST12.	BEST32.
31	HYPERTEN	Num	8	BEST12.	BEST32.
32	TIMEAP	Num	8	BEST12.	BEST32.
33	TIMEMI	Num	8	BEST12.	BEST32.
34	TIMEMIFC	Num	8	BEST12.	BEST32.
35	TIMECHD	Num	8	BEST12.	BEST32.
36	TIMESTRK	Num	8	BEST12.	BEST32.
37	TIMECVD	Num	8	BEST12.	BEST32.
38	TIMEDTH	Num	8	BEST12.	BEST32.
39	TIMEHYP	Num	8	BEST12.	BEST32.

```
/* Check levels for some of the variables from the upcoming questions */
PROC FREQ DATA=LAB.A1;
    TABLES SEX TOTCHOL HDLC LDLC GLUCOSE DIABETES;
RUN;
```

The FREQ Procedure

SEX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5022	43.19	5022	43.19
2	6605	56.81	11627	100.00

*Note that there are more female participants. 1=Men, 2=Women

DIABETES	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11097	95.44	11097	95.44
1	530	4.56	11627	100.00

*Note that the dataset only has about 5% with diabetes - class imbalance exists.

1. Is there a difference in cholesterol levels between male and female patients?

- **Assumption:** By cholesterol levels the question is referring to TOTCHOL (Serum Total Cholesterol) and not more specifically HDLC and LDLC.
- First I checked to see the split of male and female patients in the dataset. There are 57% female to 43% male patients in the dataset.

The FREQ Procedure

SEX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5022	43.19	5022	43.19
2	6605	56.81	11627	100.00

-
- Next, I wanted to check normality for the total cholesterol level variables in the dataset to determine the appropriate method of determining difference between the male and female patients.

```
/* Check for Normality of TOTCHOL and Plot TOTCHOL, as well as Percentiles */

PROC UNIVARIATE DATA=LAB.A1 NORMAL PLOT;
    VAR TOTCHOL;
RUN;
```

-

The UNIVARIATE Procedure
Variable: TOTCHOL

Moments			
N	11218	Sum Weights	11218
Mean	241.162418	Sum Observations	2705360
Std Deviation	45.3680304	Variance	2058.25819
Skewness	0.82044596	Kurtosis	3.38088449
Uncorrected SS	675518640	Corrected SS	23087482.1
Coeff Variation	18.8122307	Std Error Mean	0.42834353

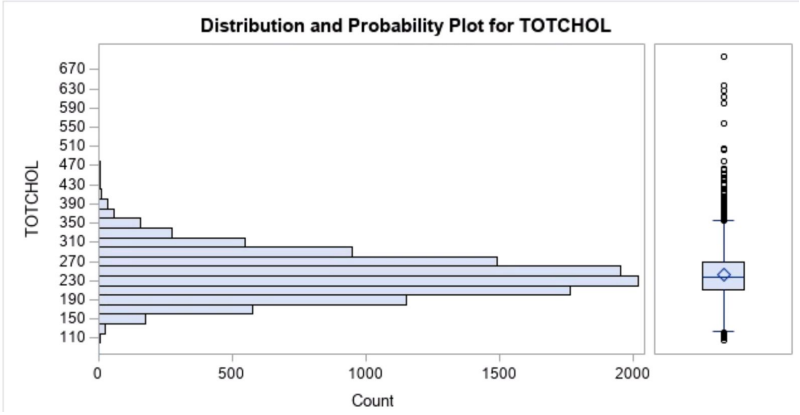
- *Note the 409 missing values in TOTCHOL

Basic Statistical Measures			
Location		Variability	
Mean	241.1624	Std Deviation	45.36803
Median	238.0000	Variance	2058
Mode	240.0000	Range	589.00000
		Interquartile Range	58.00000

- *Large range of values

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.0408	Pr > D	<0.0100
Cramer-von Mises	W-Sq	4.638004	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	29.96839	Pr > A-Sq	<0.0050

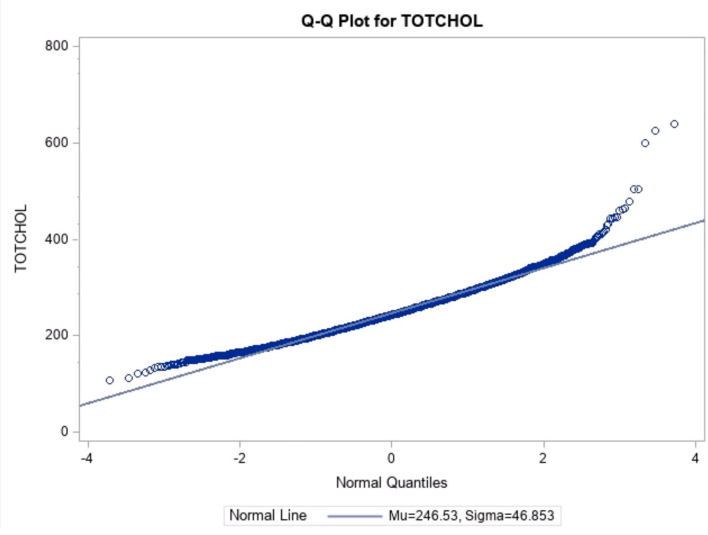
- *The <0.05 confirms that this is not normal. Note Shapiro-Wilks is not shown and it would not work appropriately given the >1,000 patients in the dataset.

- 

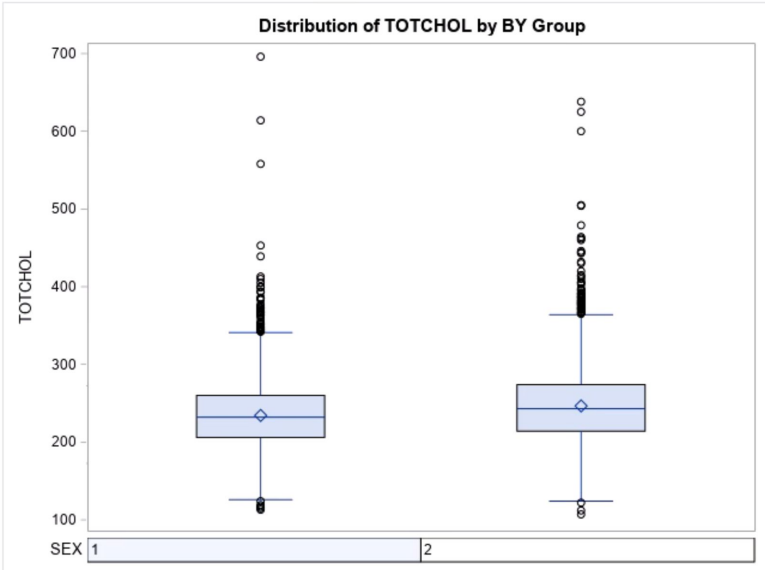
- TOTCHOL is not normal - uneven given by the distribution plot and qq-plot as well.
- ```
  /*Q1. Is there a difference in total cholesterol levels between male and female patients*/
  PROC SORT DATA=Lab.A1;
      BY SEX;RUN;

  PROC UNIVARIATE DATA=LAB.A1 NORMAL PLOT CIPCTLDF;
      BY SEX;
      VAR TOTCHOL;
      HISTOGRAM TOTCHOL / NORMAL;
      QQPLOT / NORMAL (MU=est SIGMA=est);
  RUN;

  /* Check Normality Visually Looking at Boxplots */

  PROC SGPLOT DATA=LAB.A1;
      TITLE "Boxplots of TOTCHOL by Sex";
      VBOX TOTCHOL / Category=SEX;
  RUN;
  ```
- 

*Showing TOTCHOL not normal
  ```

- 

*Appears to be a difference
- I looked at normality in the male and female groups using the method described in class and both are non-normal, so I used the nonparametric t-test.
-

```
/* Mann-Whitney U (Wilcoxon) test - Nonparametric T-Test */

PROC NPAR1WAY DATA=LAB.A1 WILCOXON;
    CLASS SEX;
    VAR TOTCHOL;
RUN;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable TOTCHOL Classified by Variable SEX

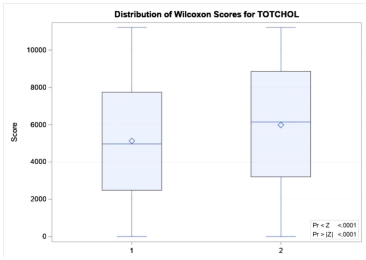
SEX	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	4915	25210064.5	27570692.5	170179.524	5129.20946
2	6303	37717306.5	35356678.5	170179.524	5984.02451

Average scores were used for ties.

Wilcoxon Two-Sample Test

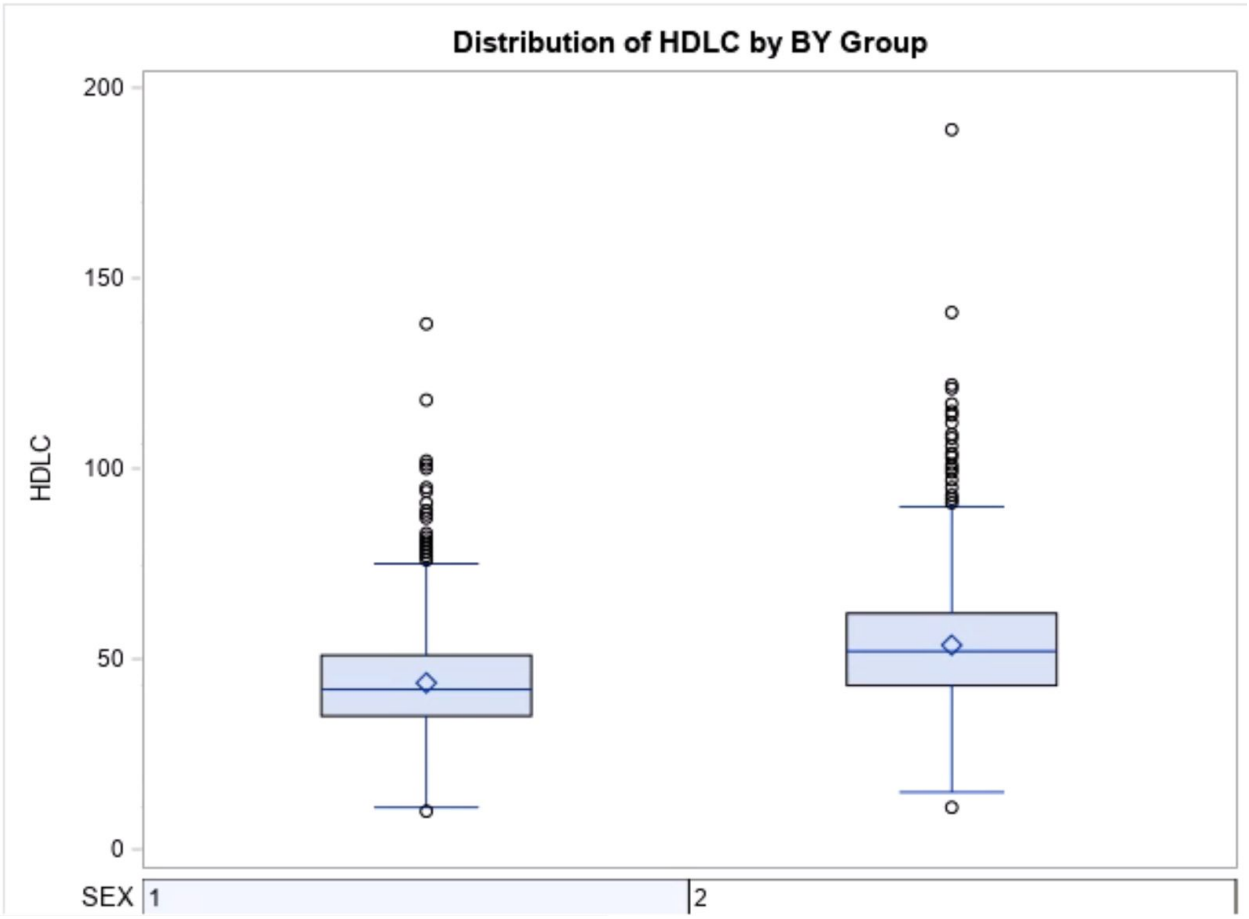
Statistic	Z	Pr < Z	Pr > |Z|	t Approximation Pr < Z	t Approximation Pr > |Z|
25210065	-13.8714	<.0001	<.0001	<.0001	<.0001

Z includes a continuity correction of 0.5.



- **Since p<0.05 we can reject the null hypothesis and argue that male and female patients in this dataset do have a difference in cholesterol levels (Serum Total Cholesterol).**

- In case this question also wanted the same analysis for HDLC and LDLC:

- 

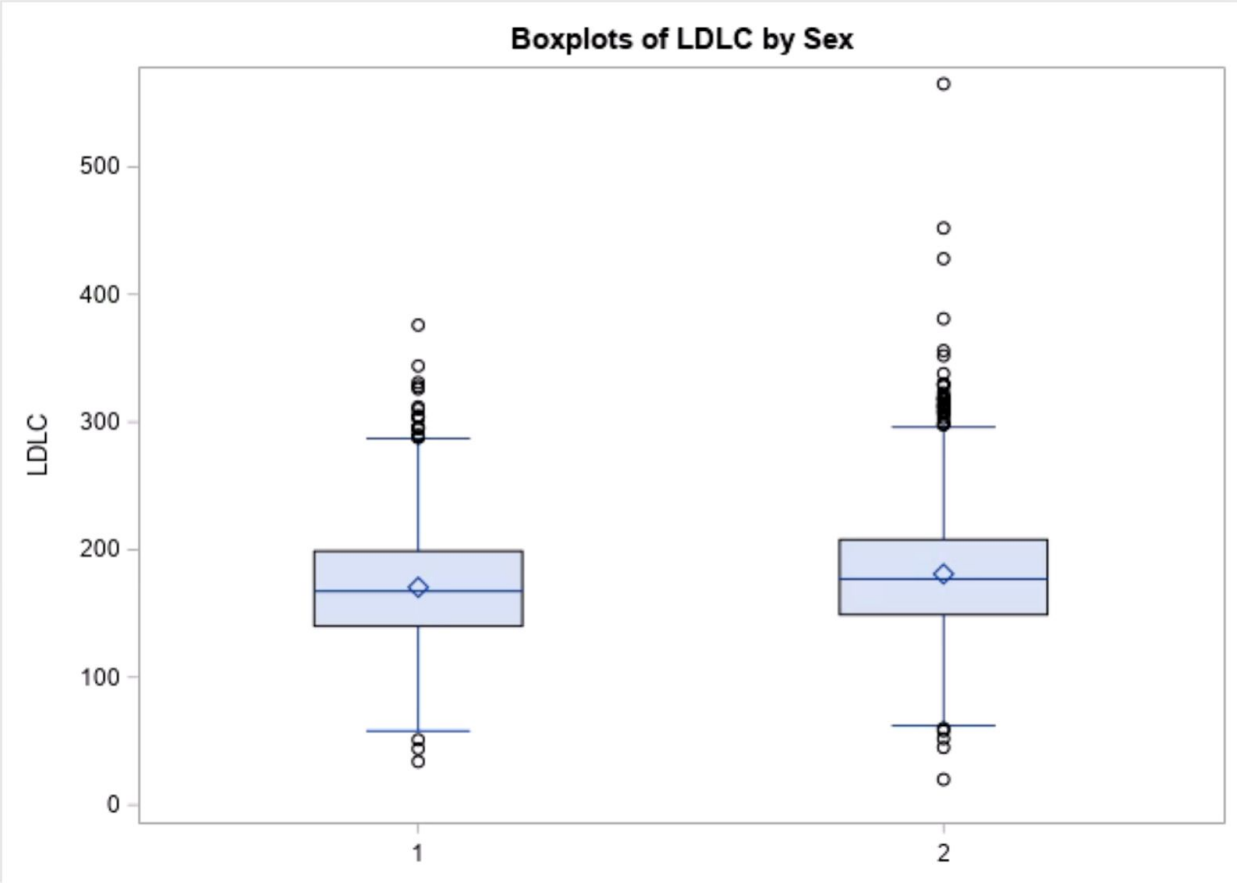
Boxplots of HDLC by Sex

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable HDLC Classified by Variable SEX					
SEX	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	1304	1535224.50	1974256.0	23804.5144	1177.31940
2	1723	3047653.50	2608622.0	23804.5144	1768.80644
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
1535225	-18.4432	<.0001	<.0001	<.0001	<.0001
Z includes a continuity correction of 0.5.					

-
    - There appears to be a difference in HDLC between male and female patients in the dataset as indicated by the nonparametric t-test $p<0.05$.

- 

Boxplots of LDLC by Sex

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable LDLC Classified by Variable SEX					
SEX	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	1304	1834152.50	1973604.0	23798.9287	1406.55867
2	1722	2745698.50	2606247.0	23798.9287	1594.48229
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
1834153	-5.8595	<.0001	<.0001	<.0001	<.0001
Z includes a continuity correction of 0.5.					

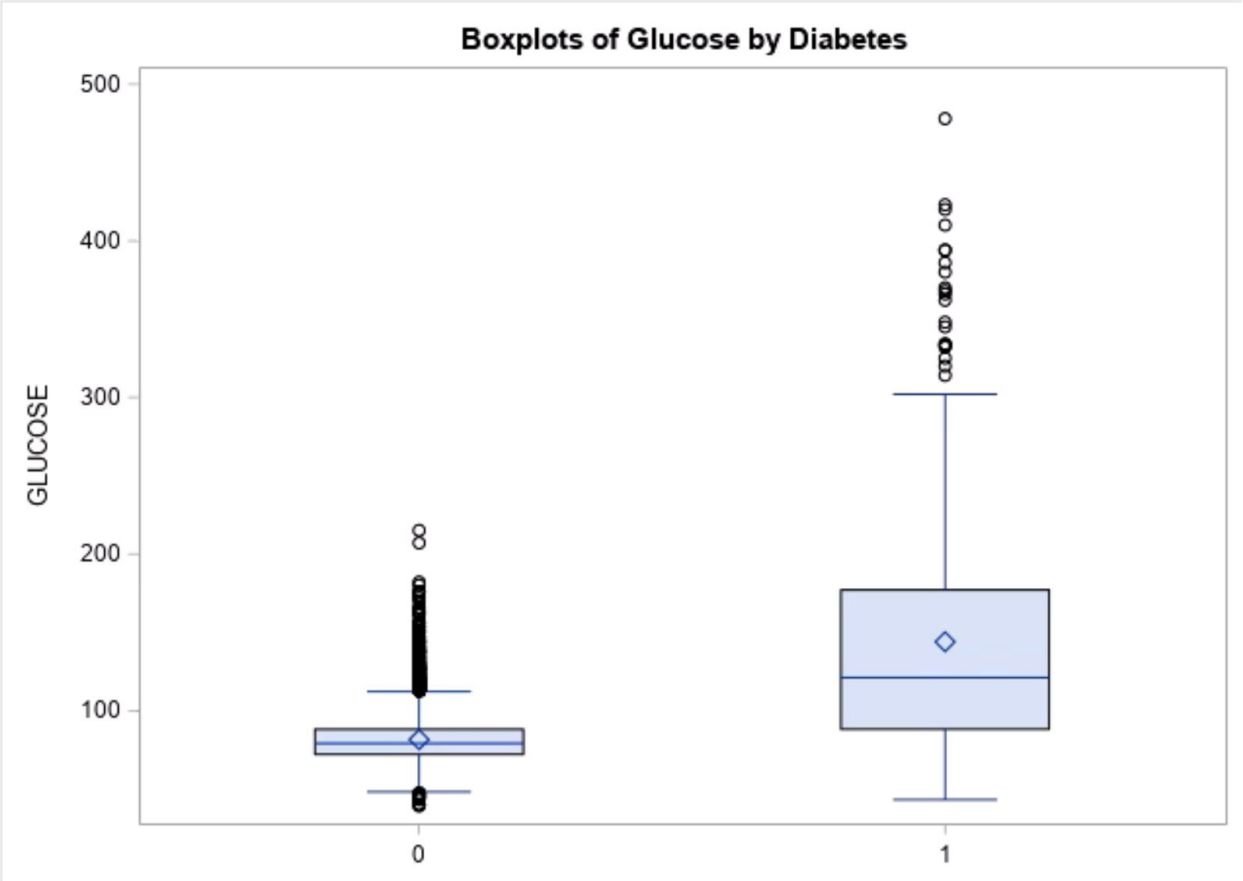
-
    - There appears to be a difference in LDLC between male and female patients in the dataset as well - as indicated by the nonparametric t-test p<0.05.

2. Is there a relationship between Glucose levels and Diabetes?

```
PROC SGPLOT DATA=LAB.A1;
     TITLE "Boxplots of Glucose by Diabetes";
     VBOX GLUCOSE / Category=DIABETES;
RUN;

/* What is the approriate correlation to use? */

PROC CORR DATA=LAB.A1 PEARSON SPEARMAN KENDALL;
     VAR GLUCOSE;
     WITH DIABETES;
RUN;
```

- 

- **The relationship appears to be that diabetics have a much wider range of Casual Serum Glucose levels as well as a much higher maximum level than those without diabetes.**

Wilcoxon Scores (Rank Sums) for Variable GLUCOSE Classified by Variable DIABETES					
DIABETES	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	9744	48266081.5	49635936.0	60518.4866	4953.41559
1	443	3626496.5	2256642.0	60518.4866	8186.22235
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
3626497	22.6353	<.0001	<.0001	<.0001	<.0001
Z includes a continuity correction of 0.5.					

- *Wilcoxon displayed here, but Spearman Rank below is what should be used given this non-normal data and continuous glucose vs categorical diabetes label.
- The two groups did not appear to be normally distributed and we have categorical data, but the **Spearman Correlation** table does show significance. That said 0.22 is not a strong correlation.

The CORR Procedure

2 Variables:	GLUCOSE DIABETES

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
GLUCOSE	10187	84.12487	24.99378	80.00000	39.00000	478.00000
DIABETES	11627	0.04558	0.20859	0	0	1.00000

Spearman Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	GLUCOSE	DIABETES
GLUCOSE	1.00000 10187	0.22428 <.0001 10187
DIABETES	0.22428 <.0001 10187	1.00000 11627

-
- Would probably be better to split the glucose levels variable in half to have a new variable that is 1 for high glucose and 0 for low glucose so the relationship could be more easily defined.

3. What variables explain any cardiovascular disease? (Use Logistic Regression) In an initial model do not use any training or testing splits. Provide the Odds Ratios and 95% Confidence Intervals, and make sure to explain what they mean in terms of any cardiovascular disease.

The FREQ Procedure

PREVCHD	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	10785	92.76	10785	92.76
1	842	7.24	11627	100.00

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	75.9	Somers' D	0.518
Percent Discordant	24.1	Gamma	0.518
Percent Tied	0.0	Tau-a	0.101
Pairs	518364	c	0.759

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
SEX	1.0000	0.436	0.322	0.589
AGE	1.0000	1.069	1.050	1.088
TOTCHOL	1.0000	1.009	1.000	1.017
HDLC	1.0000	0.978	0.965	0.990
LDLC	1.0000	0.997	0.989	1.005
DIABETES	1.0000	2.160	1.435	3.250
PREVHYP	1.0000	1.954	1.408	2.712
PREVSTRK	1.0000	2.220	1.166	4.226

- Some of the variables that explain cardiovascular disease include sex, age, serum total cholesterol, HDL cholesterol levels, LDL cholesterol levels, diabetes diagnosis, hypertension, and history of stroke. Men tend to develop CVD at a younger age and are typically at higher risk of coronary heart disease. Women tend to be more at risk of stroke, but at an older age. LDL can lead to artery-clogging plaque buildup which is a

symptom of cardiovascular disease. HDL on the other hand can help to clear cholesterol from the blood. So elevated LDL and low HDL could increase risk of cardiovascular disease. Diabetes is another related chronic disease that also sees persistent inflammation of the vasculature. All the selected features make sense in this context.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	AGE	1	1	84.2182	<.0001
2	SEX	1	2	36.9561	<.0001
3	PREVHYP	1	3	22.8178	<.0001
4	LDLC	1	4	15.8042	<.0001
5	DIABETES	1	5	17.0015	<.0001
6	HDLC	1	6	8.6859	0.0032
7	PREVSTRK	1	7	6.3499	0.0117
8	TOTCHOL	1	8	4.2910	0.0383

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.0647	0.6781	80.0011	<.0001
SEX	1	-0.8309	0.1538	29.1885	<.0001
AGE	1	0.0663	0.00900	54.1834	<.0001
TOTCHOL	1	0.00863	0.00420	4.2216	0.0399
HDLC	1	-0.0226	0.00641	12.4719	0.0004
LDLC	1	-0.00273	0.00402	0.4618	0.4968
DIABETES	1	0.7699	0.2086	13.6251	0.0002
PREVHYP	1	0.6699	0.1673	16.0354	<.0001
PREVSTRK	1	0.7975	0.3285	5.8940	0.0152

Run again using training and testing splits using R for machine learning. Make sure to present Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) for comparison with Problem 5 values.

```{r p3b}

frmgham2_clean$PREVCHD <- as.factor(frmgham2_clean$PREVCHD)
log_reg <- glm(
  PREVCHD ~ SEX +AGE +SYSBP +DIABP +BPMEDS +CURSMOKE +TOTCHOL +HDLC +LDLC +BMI +GLUCOSE +DIABETES +HEARTRTE +PREVHYP +PREVSTRK,
  family = "binomial",
  data = frmgham2_clean
  )

# Create training (70%) and test (30%) sets

set.seed(123)  # use a set seed point for reproducibility
split <- initial_split(frmgham2_clean, prop = .7, strata = "PREVCHD")
train <- training(split)
test  <- testing(split)

```

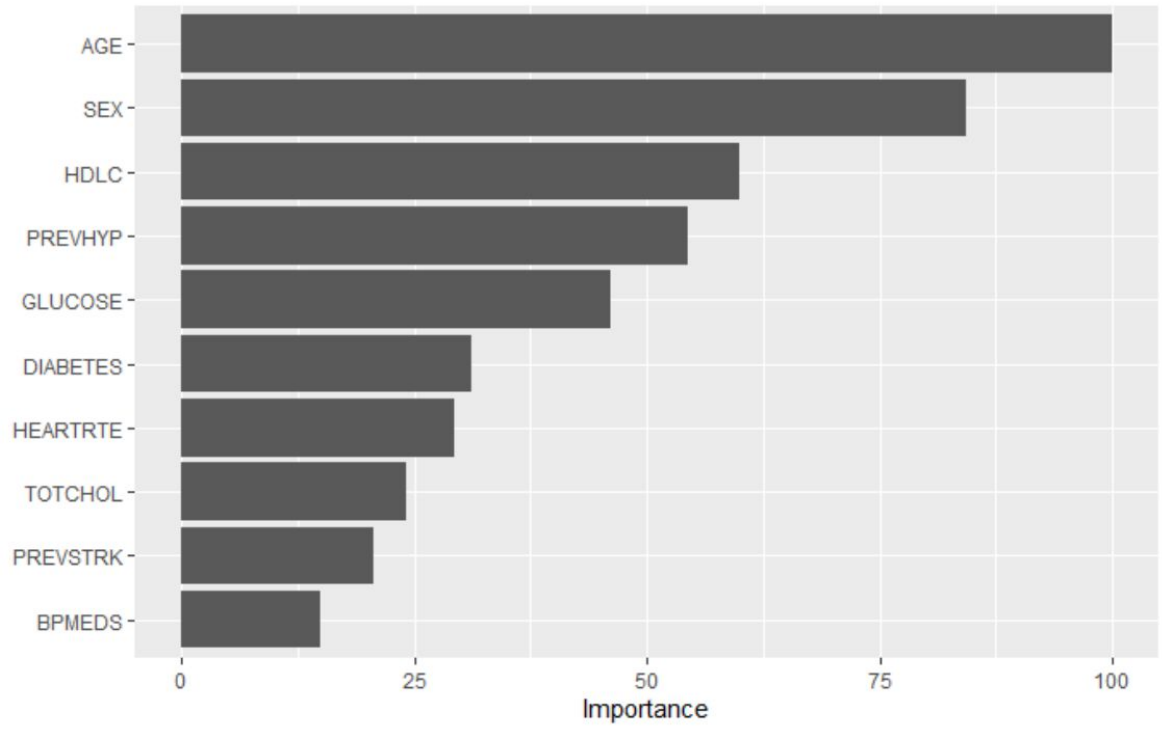
```{r p3b2}

# Create training (70%) and test (30%) sets

summary(log_reg) #Coefficients Not in exponential form

tidy(log_reg) #Coefficients Not in exponential form
```

```{r p3b4}

#For Predicting dependent variable
log_reg = train(
  form = PREVCHD ~ SEX +AGE +SYSBP +DIABP +BPMEDS +CURSMOKE +TOTCHOL +HDLC +LDLC +BMI +GLUCOSE +DIABETES +HEARTRTE +PREVHYP +PREVSTRK,
  data = train,
  method = "glm",
  family = "binomial"
)

```

```{r p3b5}
#Confusion Matrix
confusionMatrix(predict(log_reg, test), as.factor(test$PREVCHD))


#Variables of Importance

vip(log_reg, num_features = 10)
```

```
Call:
glm(formula = PREVCHD ~ SEX + AGE + DIABP + BPMEDS + TOTCHOL +
    HDLC + DIABETES + PREVHYP + PREVSTRK, family = "binomial",
    data = frmgham2_clean)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5007  -0.5107  -0.3578  -0.2333   2.9077

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.784099   0.947237  -5.051 4.40e-07 ***
SEX         -0.828531   0.158733  -5.220 1.79e-07 ***
AGE          0.061767   0.009422   6.556 5.54e-11 ***
DIABP       -0.012415   0.006993  -1.775 0.075817 .
BPMEDS       0.317252   0.191641   1.655 0.097833 .
TOTCHOL      0.005705   0.001559   3.660 0.000252 ***
HDLC        -0.020194   0.005339  -3.782 0.000155 ***
DIABETES     0.800402   0.210775   3.797 0.000146 ***
PREVHYP      0.715835   0.190543   3.757 0.000172 ***
PREVSTRK     0.760732   0.340410   2.235 0.025434 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
               Accuracy : 0.8879
                 95% CI : (0.8615, 0.9108)
    No Information Rate : 0.8924
    P-Value [Acc > NIR] : 0.6736

                  Kappa : 0.0122

 Mcnemar's Test P-Value : 2.517e-14

            Sensitivity : 0.99330
            Specificity : 0.01389
         Pos Pred Value : 0.89307
         Neg Pred Value : 0.20000
             Prevalence : 0.89238
         Detection Rate : 0.88640
   Detection Prevalence : 0.99253
      Balanced Accuracy : 0.50359
```



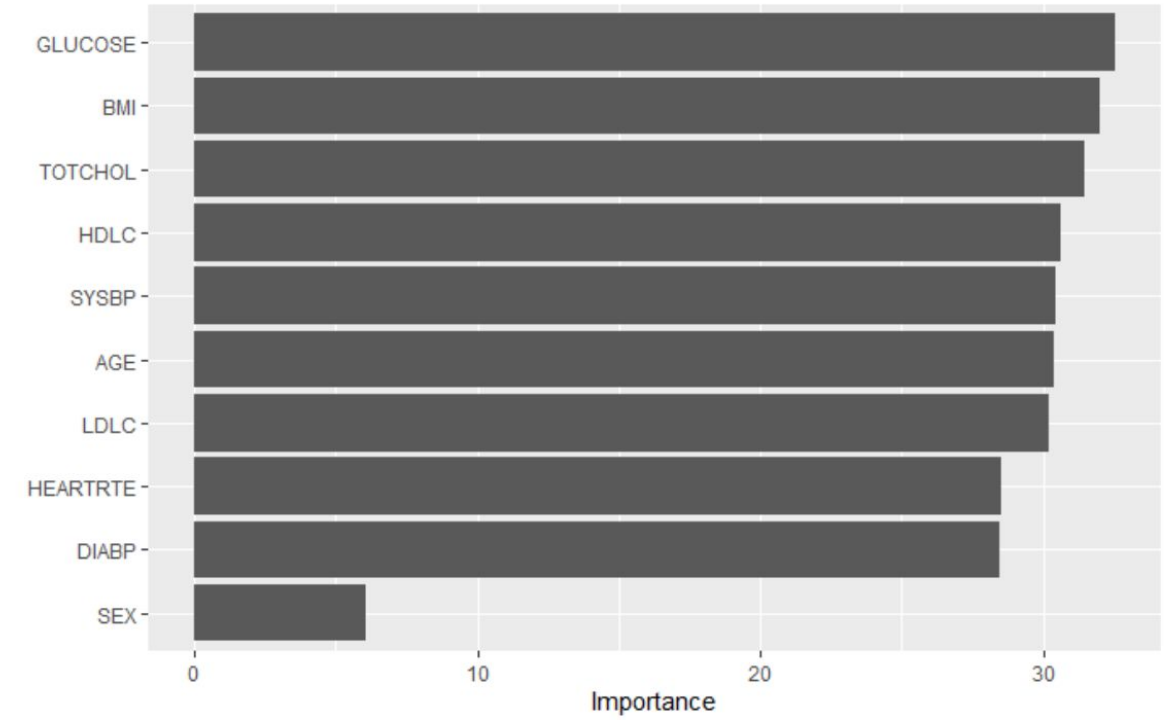
- Similar features selected here as in SAS. Glucose and diabetes are correlated, as are a few of the other features. Could probably pair this down further with similar performance. Blood pressure medication is one that was not in the previous selection but makes sense given the relationship between heart disease and high blood pressure.

4. What variables predict any cardiovascular disease? (Use Machine Learning Algorithm of Your Choice) (Use R for machine learning) Do your answers differ from Problem 4? If so, how?

Random Forest:

```{r p3b5}

library(randomForest)

randomForest <- randomForest(PREVCHD ~ SEX +AGE +SYSBP +DIABP +BPMEDS +CURSMOKE +TOTCHOL +HDLC +LDLC +BMI +GLUCOSE +DIABETES +HEARTRTE
+PREVHYP +PREVSTRK,   data=train)
print(randomForest) # view results

importance(randomForest) # importance of each predictor

vip(randomForest, num_features = 10)

#predict
rfPredict <- predict(randomForest,test)
confusionMatrix(rfPredict, as.factor(test$PREVCHD))

```

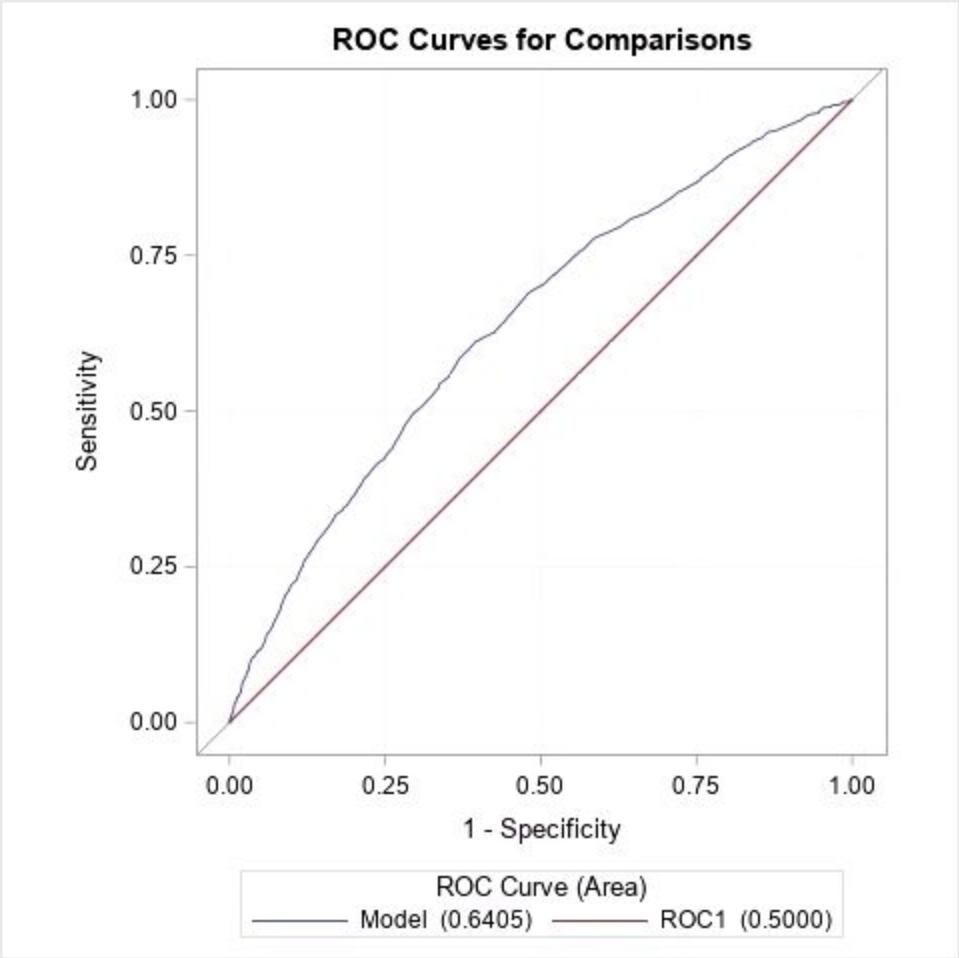
```
               Accuracy : 0.8909
                 95% CI : (0.8648, 0.9135)
    No Information Rate : 0.8924
    P-Value [Acc > NIR] : 0.5804

                  Kappa : -0.003

 Mcnemar's Test P-Value : 2.55e-16

            Sensitivity : 0.9983
            Specificity : 0.0000
         Pos Pred Value : 0.8922
         Neg Pred Value : 0.0000
             Prevalence : 0.8924
         Detection Rate : 0.8909
   Detection Prevalence : 0.9985
      Balanced Accuracy : 0.4992
```



- In this case similar features are selected to predict cardiovascular disease. A difference is that BMI is selected. BMI makes sense to be correlated with cardiovascular disease. A similar feature which I hadn't discussed previously is the heart rate. It is known that high resting heart rate is associated with higher blood pressure and often heart disease itself. Adults with diabetes are more likely to have heart attacks and strokes, so the importance of glucose and diabetes make sense in the model.
- The takeaway of all models is that cardiovascular disease has to do with cholesterol levels (both HDL and LDL), heart rate, blood pressure, diabetes, age, sex, and ultimately an individual's lifestyle habits that contribute to weight (BMI) and these other factors.
- Also of note is the balanced accuracy in both models which is much lower than the accuracy.
- Finally, the models have awful specificity, which I'm not sure how to improve upon when in R to try to balance that out a bit. That said, sensitivity is great.

5. Using an ROC Curve, determine the optimal cutoff for Systolic Blood Pressure? Using the cutoff point, create a Kaplan Meier curve for the outcome of Stroke.

```
ODS GRAPHIC ON;
PROC LOGISTIC DATA = LAB.A1;
 MODEL PREVCHD (EVENT='1')=SYSBP/OUTROC
 ROC; ROCCONTRAST;
RUN;
ODS GRAPHIC OFF;


DATA ROCDATA;
     SET ROCDATA;

     /* Note:  Remember to change the cu
     /* cutoff = (logit+Intercept)/slope
     /* Choose cutoff with maximum Youde

     logit=log(_prob_/(1-_prob_));
     cutoff=(logit+1.1058)/-0.0515;
     prob= _prob_;
     Sensitivity = _SENSIT_;
     Specificity = 1-_1MSPEC_;
     Youden= _SENSIT_+ (1-_1MSPEC_)-1;
RUN;

PROC SORT DATA=ROCDATA DESCENDING;
     BY Youden;
RUN;

PROC PRINT DATA=ROCDATA; RUN;
```

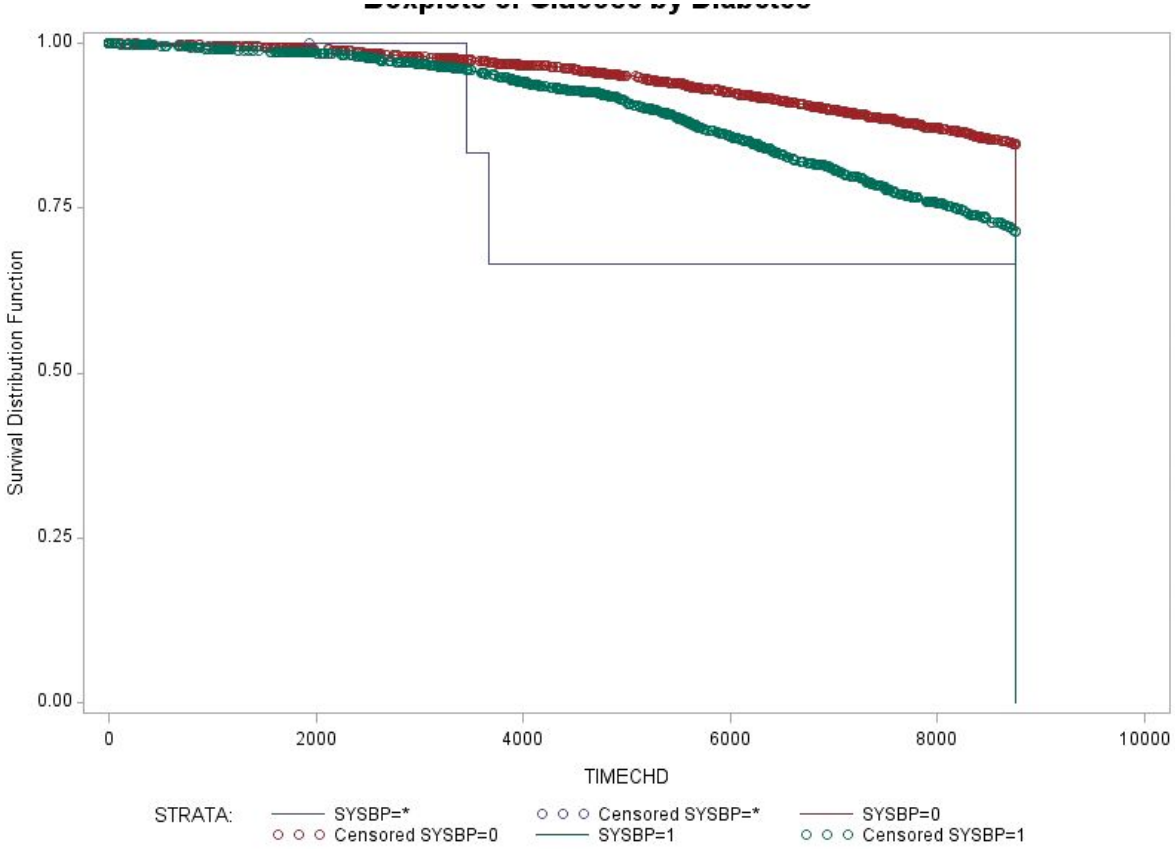


ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Model	0.6405	0.00986	0.6212	0.6598	0.2810	0.2839	0.0377
ROC1	0.5000	0	0.5000	0.5000	0	.	0

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Model	1	202.9434	<.0001

- ROC is pretty small to use, but given that it is medical data and it is above 0.6 we can use it.

- **140 cutoff variable <140 0 above 1 strata variable SYS BP**



Summary of the Number of Censored and Uncensored Values					
Stratum	SYSBP	Total	Failed	Censored	Percent Censored
1	*	8	6	2	25.00
2	0	7155	5604	1551	21.68
3	1	4464	2859	1605	35.95
Total		11627	8469	3158	27.16

*** Wasn't sure how to get rid of the missing values in SAS. Code below.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	237.9905	2	<.0001
Wilcoxon	226.6565	2	<.0001
-2Log(LR)	4.1432	2	0.1260

- It does appear to be significant for the composite end point. So high blood pressure really does appear to put individuals at higher risk of experiencing any type of cardiovascular related disease incident.

```
/* Cutoff <140 = 0 else 1*/
PROC FORMAT;
     VALUE SYSBP_cut

          80 - 139 = "0"
          140 - 300 = "1";
RUN;

DATA LAB.A1;
     set LAB.A1;
     FORMAT SYSBP SYSBP_cut.;
     LABEL SYSBP_cut =  "SYSBP Cutoff";
RUN;

PROC FREQ DATA=LAB.A1;
     TABLES SYSBP;
RUN;

/* Check Variables were created */

PROC CONTENTS DATA=LAB.A1 ORDER=varnum; RUN;

************************************************************;
/* Kaplan-Meier Curve Analysis */

/* time_var = time to event variable (i.e. Time to Any Cardiovascular Event)
   censor_var = censored variable (i.e. Any Cardiovascular Event)
   strata_var = Strata Variable (i.e. Gender, Cutoff-Variable) */
PROC LIFETEST DATA=LAB.A1 PLOTS=survival(atrisk=0 to 365 by 60);
   TIME TIMECHD*ANYCHD(1);
   STRATA SYSBP;
RUN;
```