STAT 450 Project

Analysis of Heart Failure

Report Prepared By
Ethan Huang

May 3, 2023

Table of Contents

Introduction

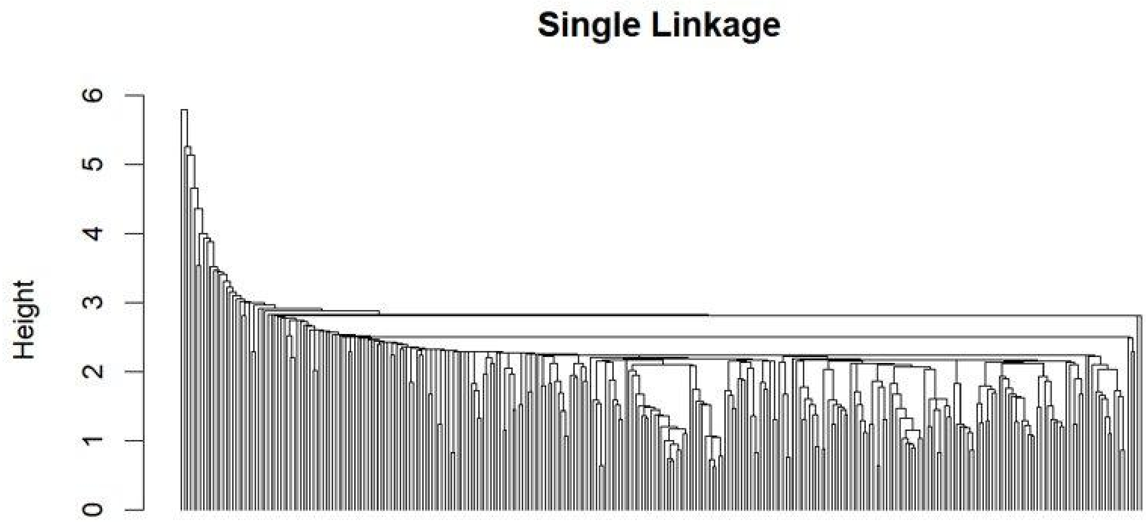
For our project, we used the *Heart Failure Clinical Records* dataset from the UCI Machine Learning Repository. It was originally utilized by Davide Chicco of the Krembil Research Institute to predict survival of patients with heart failure (Chicco). This dataset contains the medical records of 299 patients who had heart failure; the data were collected during their follow-up period. Each patient profile has 13 clinical features. The data contains eight numeric variables and 6 boolean variables. The numeric variables are age, creatinine phosphokinase, the ejection fraction, number of platelets, serum creatinine, serum sodium, time, and age. The six boolean variables reflect the presence of anaemia, high blood pressure, diabetes, gender, smoking, and death.

Throughout our analysis, we seek to answer four main questions. First, what is the optimal way that the data can be grouped? We apply multiple methods of clustering to answer this question. Second, is there a difference in the mean numerical values between smokers and non-smokers? To answer this question, we apply a multivariate analysis of variance to all our numerical values except time ang age, using smoking and non-smoking as our groups. Third, is there a difference in the mean numerical values between those who died and those who did not die? We apply Hotelling's Two-Sample T-Test to all numerical variables except time. Last, how can we use past patient data to classify whether future patients will live or die? We shall utilize linear and quadratic discriminant analysis to answer this question. By answering these questions, we will have a better understanding in determining who survives and who dies.

This report will be broken down as follows: Question 1, Question 2, Question 3, and Question 4. Each section seeks to answer one of the above questions. It will contain an explanation of the method, the results, and interpretation. After all the questions are answered, the conclusion will summarize the overall findings from the study.

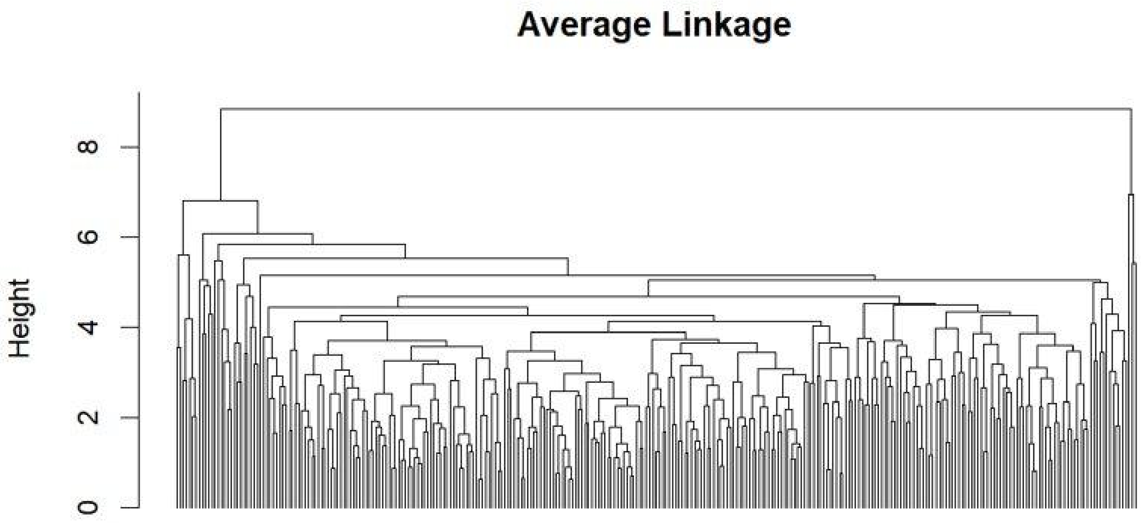
Question 1

Our first task is to find the optimal way that the data can be grouped. To do so, we use a variety of clustering methods to cluster our data into two clusters. The methods are single linkage, average linkage, complete linkage, centroid, and Ward's method. Below are the graphs and tables of the clusters.



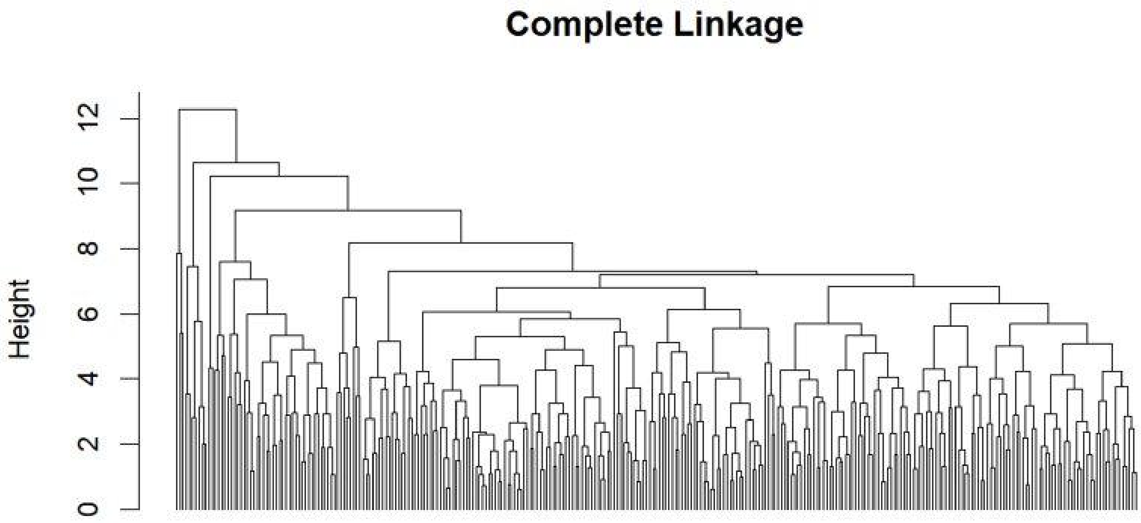
Cluster	age	anaemia	Creatinine phosphokinase	diabetes	Ejection fraction
1	60.8	0.433	570.48993	0.4161	38.
2	60	0	3964	1	62

High blood pressure	platelets	Serum creatinine	Serum sodium	sex	smoking	DEATH
0.352349	263358.03	1.375738	136.59396	0.651007	0.3221477	0.318792
0	263358.03	6.8	146	0	0	1



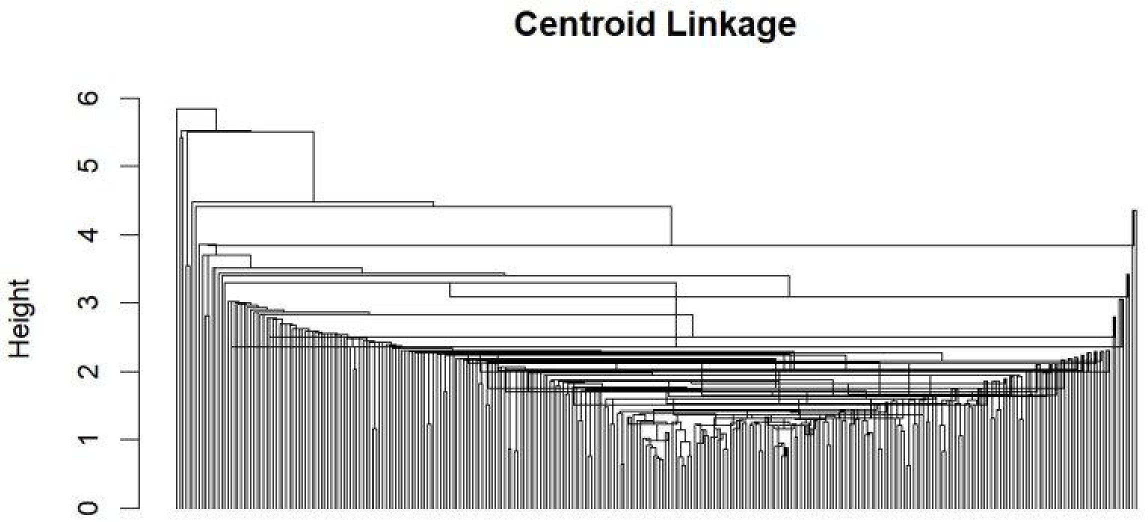
Cluster	Age	Anaemia	Creatinine phosphokinase	diabetes	Ejection fraction
1	60.8	0.429	572.4865	0.4189	37.905
2	64.7	0.667	1504.667	0.3333	55.667

High blood pressure	platelets	Serum creatinine	Serum sodium	sex	smoking	DEATH
0.347973	263316.5	1.322872	136.6047	0.652027	0.320946	0.314189
0.666667	267452.7	8.4	138.6667	0.333333	0.333333	1



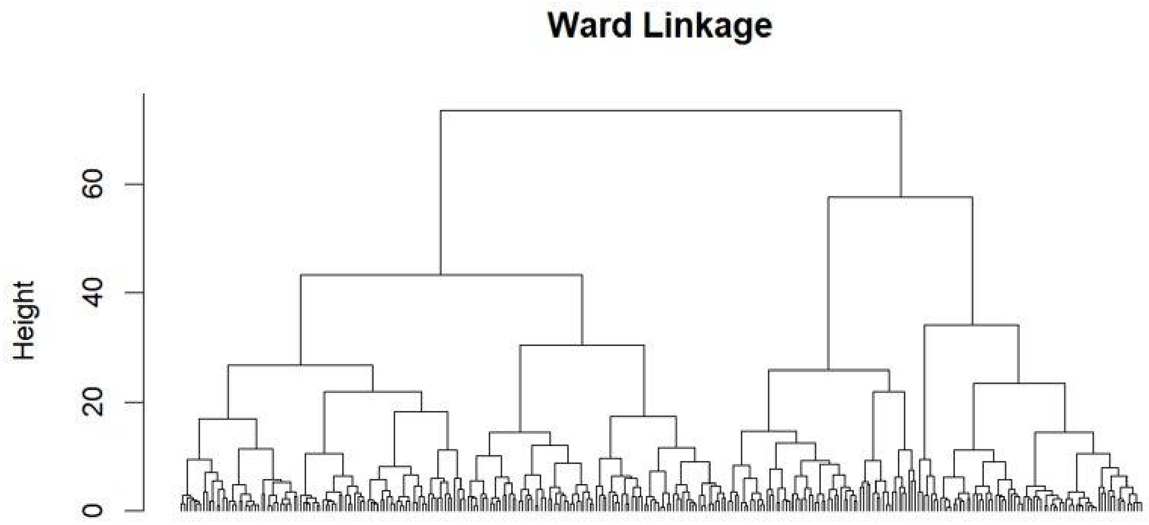
Cluster	Age	Anaemia	Creatinine phosphokinase	diabetes	Ejection fraction
1	60.79505	0.429054	572.4865	0.418919	37.90541
2	64.66667	0.666667	1504.667	0.333333	55.66667

High blood pressure	platelets	Serum creatinine	Serum sodium	sex	smoking	DEATH
0.347973	263316.5	1.322872	136.6047	0.652027	0.320946	0.314189
0.666667	267452.7	8.4	138.6667	0.333333	0.333333	1



Cluster	Age	Anaemia	Creatinine phosphokinase	diabetes	Ejection fraction
1	60.83669	0.432886	570.4899	0.416107	38.00336
2	60	0	3964	1	62

High blood pressure	platelets	Serum creatinine	Serum sodium	sex	smoking	DEATH
0.352349	263358	1.375738	136.594	0.651007	0.322148	0.318792
0	263358	6.8	146	0	0	1

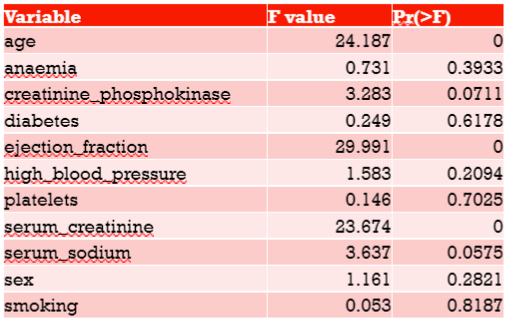


Cluster	Age	Anaemia	Creatinine phosphokinase	diabetes	Ejection fraction
1	61.81395	0.387597	671.6822	0.333333	34.86822
2	60.0902	0.464706	513.6647	0.482353	40.52353

High blood pressure	platelets	Serum creatinine	Serum sodium	sex	smoking	DEATH
0.263566	245997.3	1.707132	135.2248	0.868217	0.674419	0.527132
0.417647	276531.7	1.156176	137.6882	0.482353	0.052941	0.164706

Looking at our two clusters from each method, we conclude that Ward's Method is the best. The proportion of death for clusters 1 and 2 are respectively 0.527132 and 0.164706, meaning that the clusters are broken down more evenly between death=yes and death=no. This contrasts the four other clustering models where the mean proportion of death greatly differs between the clusters (i.e. 0.31 and 1), making it unbalanced. It means that one cluster only contains patients who died, while the other cluster mainly contains patients who survived. In short, we seek to utilize the method that creates the most balance. Moreover, the dendrogram for Ward's method appears to be more organized than those for the other four methods. The clusters do not overlap as much.

We also analyze the p-values of the variables used in this method. Our alpha is 0.05/11 variables = 0.0045.



Variable	F value	Pr(>F)
age	24.187	0
anaemia	0.731	0.3933
creatinine_phosphokinase	3.283	0.0711
diabetes	0.249	0.6178
ejection_fraction	29.991	0
high_blood_pressure	1.583	0.2094
platelets	0.146	0.7025
serum_creatinine	23.674	0
serum_sodium	3.637	0.0575
sex	1.161	0.2821
smoking	0.053	0.8187

We see that the p-values for age, ejection fraction, and serum creatinine are all less than 0.0045. Hence, we reject the null hypothesis of equality of cluster means. We conclude that the means differ among the clusters.

Question 2

In response to our second question, we utilize a one-way MANOVA to compare the difference in the mean numerical values between smokers and nonsmokers, treating smoking as a treatment. Our hypotheses are as follows:

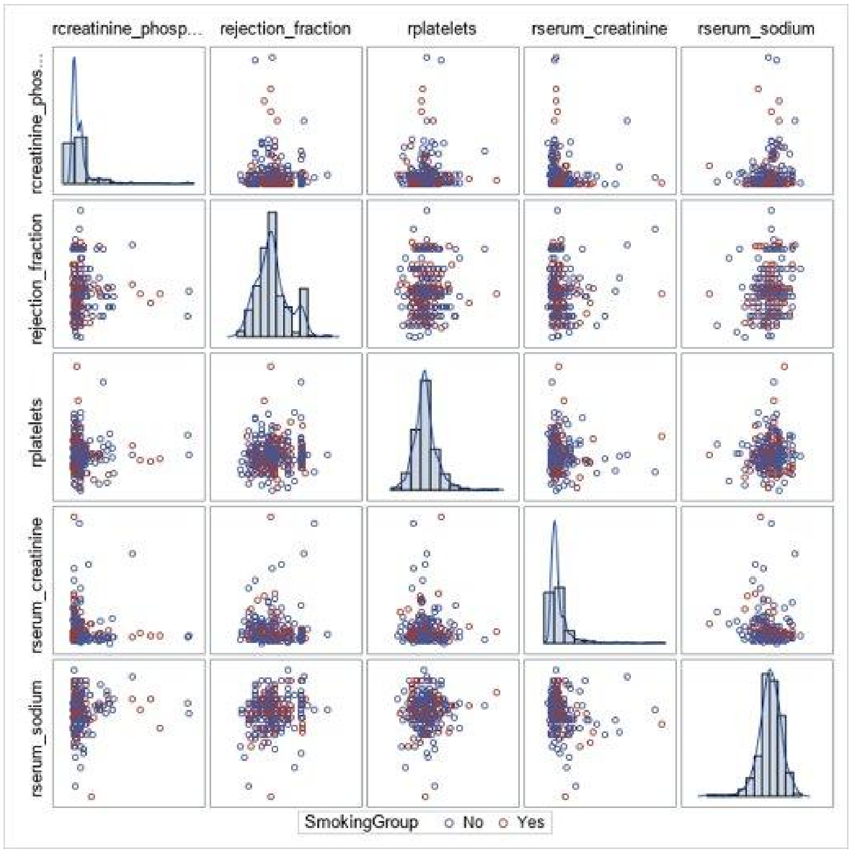
$$H_0 : \mu_{nonsmoking} = \mu_{smokers}$$

Null Hypothesis: The mean vector for nonsmokers is equal to that for smokers.

$$H_A : \mu_{nonsmoking} \neq \mu_{smokers}$$

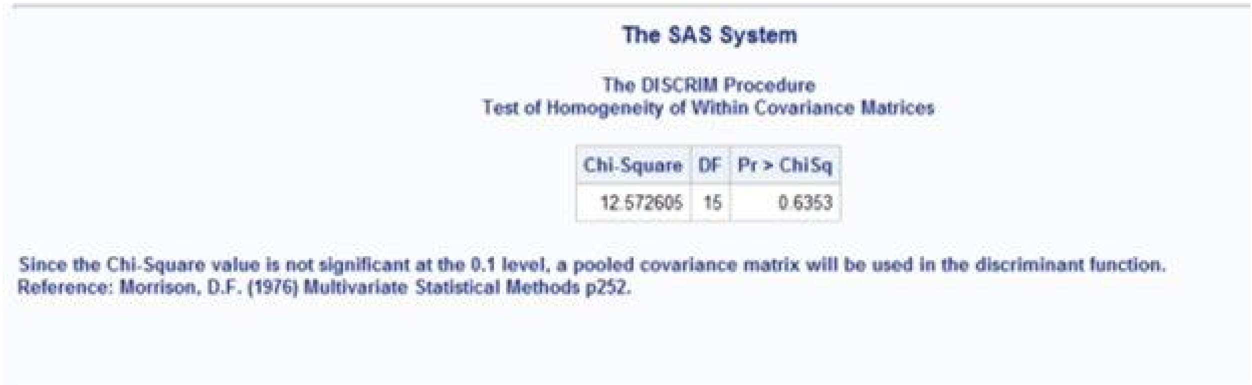
Alternative Hypothesis: The mean vector for nonsmokers is not equal to that for smokers.

However, in order to conduct MANOVA, we must verify normality and equal variance. To verify normality, we use the proc sgscatter function in SAS to plot our residuals.



Based on this scatter plot, we notice that the histograms for creatinine phosphokinase and serum creatinine are slightly skewed to the right. However, the histograms for all the other variables are roughly symmetric. Furthermore, we see that the scatterplots display a subtle elliptical shape, verifying our normality assumption.

Next, we seek to check our assumption of equal variance. To do so, we conduct Bartlett's test. We use the proc discrim procedure in SAS with the pool=test option.



The SAS System

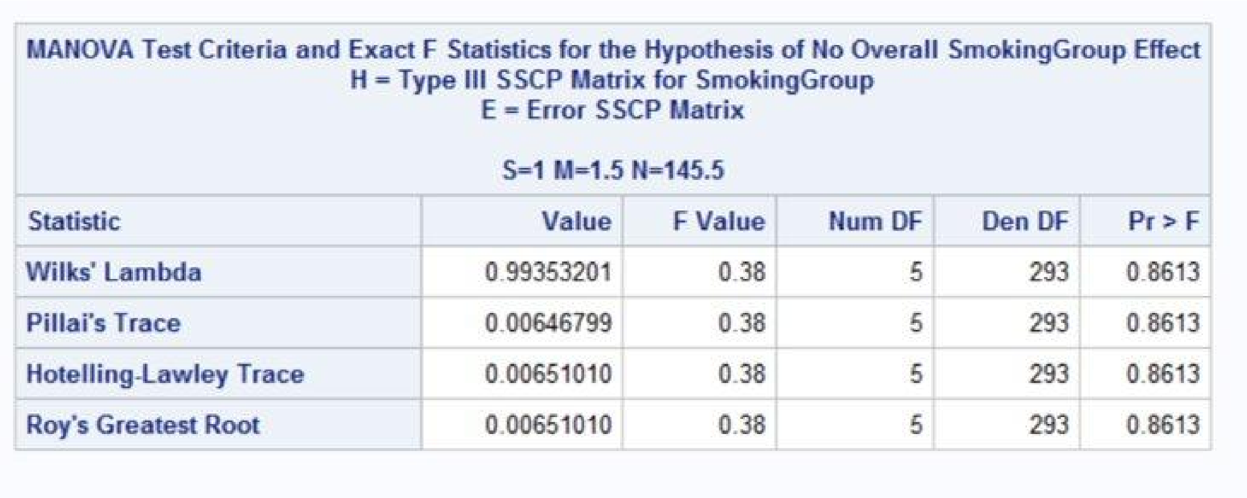
The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
12.572605	15	0.6353

Since the Chi-Square value is not significant at the 0.1 level, a pooled covariance matrix will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

We see that the p-value of the test is 0.6353, which is greater than alpha=0.05. Therefore, we fail to reject the null hypothesis of equal variance. The variance for smokers and non-smokers are the same.

Now, we conduct our MANOVA test as shown below:



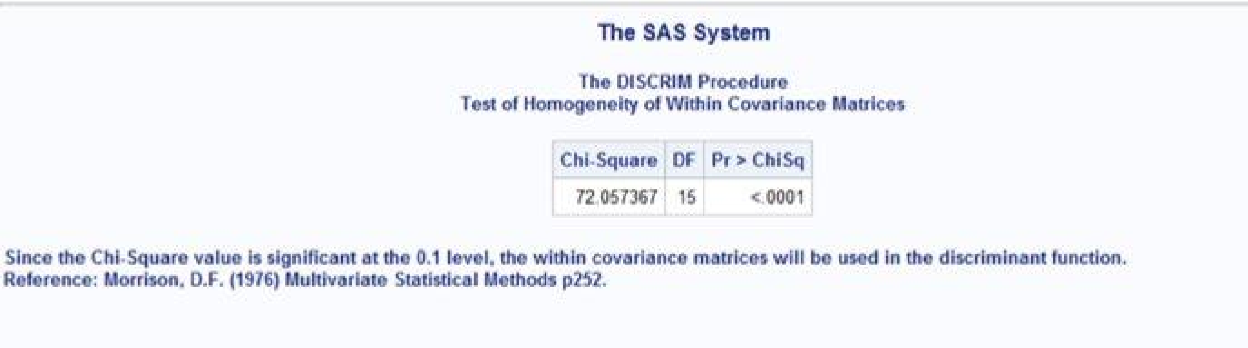
MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall SmokingGroup Effect
H = Type III SSCP Matrix for SmokingGroup
E = Error SSCP Matrix

S=1 M=1.5 N=145.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.99353201	0.38	5	293	0.8613
Pillai's Trace	0.00646799	0.38	5	293	0.8613
Hotelling-Lawley Trace	0.00651010	0.38	5	293	0.8613
Roy's Greatest Root	0.00651010	0.38	5	293	0.8613

We see that the p-values for all these tests are 0.8613, which are greater than alpha=0.05. Hence, we fail to reject the null hypothesis and conclude that the true mean numerical values are the same for smokers and non-smokers.

Furthermore, we were going to do a MANOVA to compare the mean numerical values between males and females, but the equal variance assumption was violated as shown below.



The SAS System

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
72.057367	15	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

We see that the p-value is way less than 0.01, so we reject the null of equal variance and fail to proceed with our MANOVA.

Question 3

Third, we seek to understand the difference in the mean numerical values between those who died and those who did not die. We use the 2-Sample Hotelling's T-Test; we treat those who survived and those who died as two separate groups. We add age to this question because we believe that those who died are generally older than those who did not die. We test the following hypotheses:

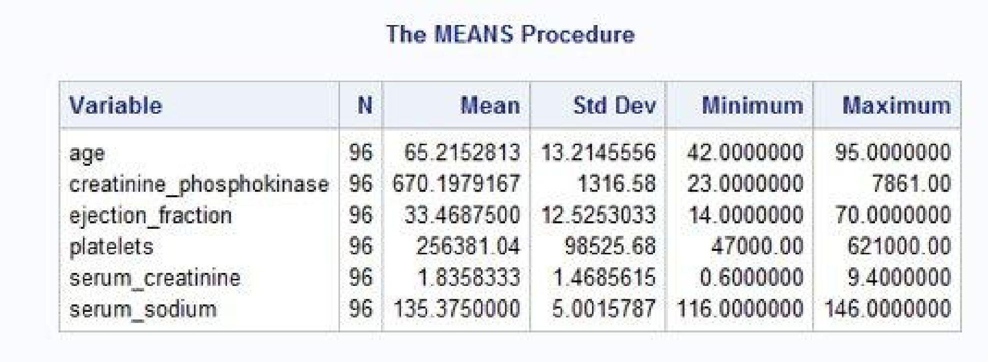
$$H_o : \vec{\mu_1} = \vec{\mu_2}\ vs\ H_a : \vec{\mu_1} \neq \vec{\mu_2}$$

$$H_o : \vec{\mu_1} - \vec{\mu_2} = 0\ vs\ H_a : \vec{\mu_1} - \vec{\mu_2} \neq 0$$

$$\vec{\mu_1}\ is\ death\ occurred$$

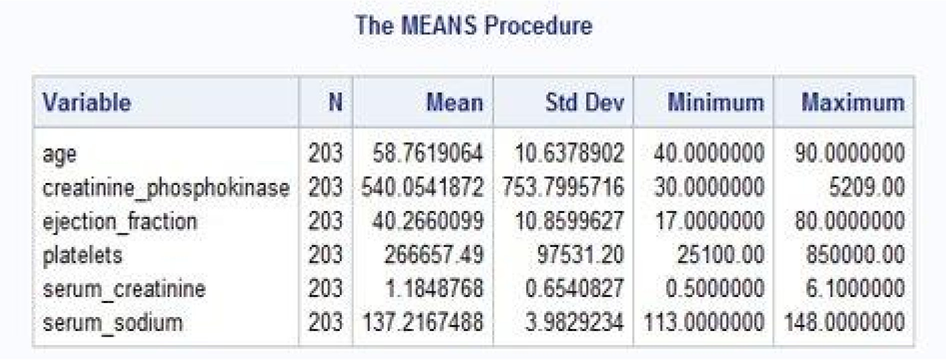
$$\vec{\mu_2}\ is\ death\ did\ not\ occur$$

First, we calculate our sample mean vectors. Below are the mean vectors for those who died and those who did not die, respectively.



The MEANS Procedure

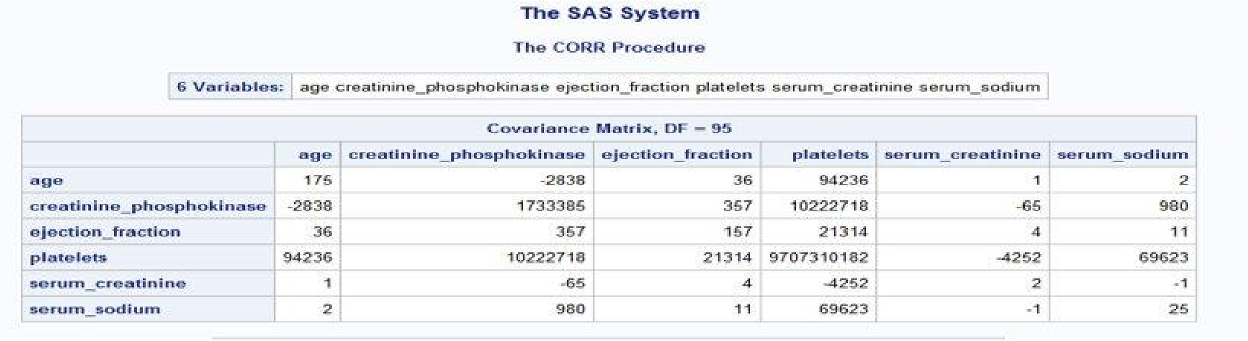
Variable	N	Mean	Std Dev	Minimum	Maximum
age	96	65.2152813	13.2145556	42.0000000	95.0000000
creatinine_phosphokinase	96	670.1979167	1316.58	23.0000000	7861.00
ejection_fraction	96	33.4687500	12.5253033	14.0000000	70.0000000
platelets	96	256381.04	98525.68	47000.00	621000.00
serum_creatinine	96	1.8358333	1.4685615	0.6000000	9.4000000
serum_sodium	96	135.3750000	5.0015787	116.0000000	146.0000000



The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
age	203	58.7619064	10.6378902	40.0000000	90.0000000
creatinine_phosphokinase	203	540.0541872	753.7995716	30.0000000	5209.00
ejection_fraction	203	40.2660099	10.8599627	17.0000000	80.0000000
platelets	203	266657.49	97531.20	25100.00	850000.00
serum_creatinine	203	1.1848768	0.6540827	0.5000000	6.1000000
serum_sodium	203	137.2167488	3.9829234	113.0000000	148.0000000

We also compute the covariance matrices for both groups, respectively.



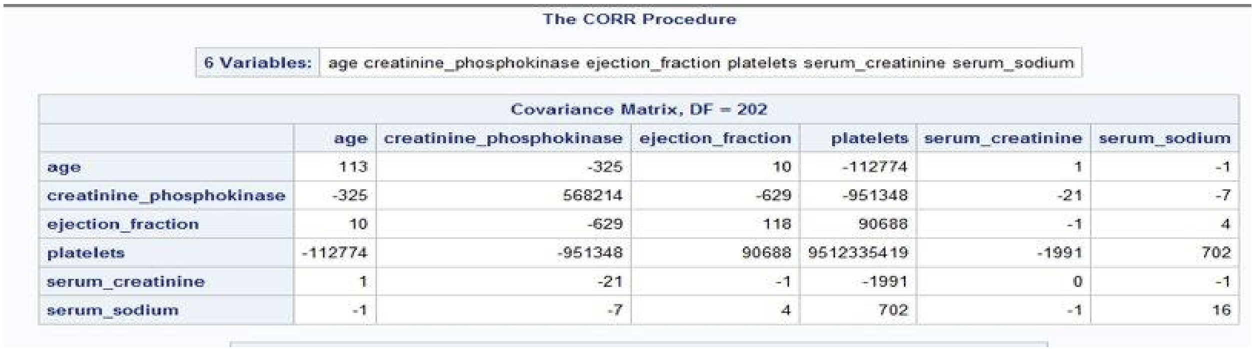
The SAS System

The CORR Procedure

6 Variables:	age creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium

Covariance Matrix, DF = 95

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
age	175	-2838	36	94236	1	2
creatinine_phosphokinase	-2838	1733385	357	10222718	-65	980
ejection_fraction	36	357	157	21314	4	11
platelets	94236	10222718	21314	9707310182	-4252	69623
serum_creatinine	1	-65	4	-4252	2	-1
serum_sodium	2	980	11	69623	-1	25



The CORR Procedure

6 Variables:	age creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium

Covariance Matrix, DF = 202

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
age	113	-325	10	-112774	1	-1
creatinine_phosphokinase	-325	568214	-629	-951348	-21	-7
ejection_fraction	10	-629	118	90688	-1	4
platelets	-112774	-951348	90688	9512335419	-1991	702
serum_creatinine	1	-21	-1	-1991	0	-1
serum_sodium	-1	-7	4	702	-1	16

Next, we use these mean vectors and covariance matrices to compute the T^2 statistic using the following formula, setting Mu 1 - Mu 2 as equal to zero:

$$T^2 = [\vec{\bar{X}}_1 - \vec{\bar{X}}_1 - (\vec{\mu}_1 - \vec{\mu}_2)]^T [\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2]^{-1} [\vec{\bar{X}}_1 - \vec{\bar{X}}_1 - (\vec{\mu}_1 - \vec{\mu}_2)]$$

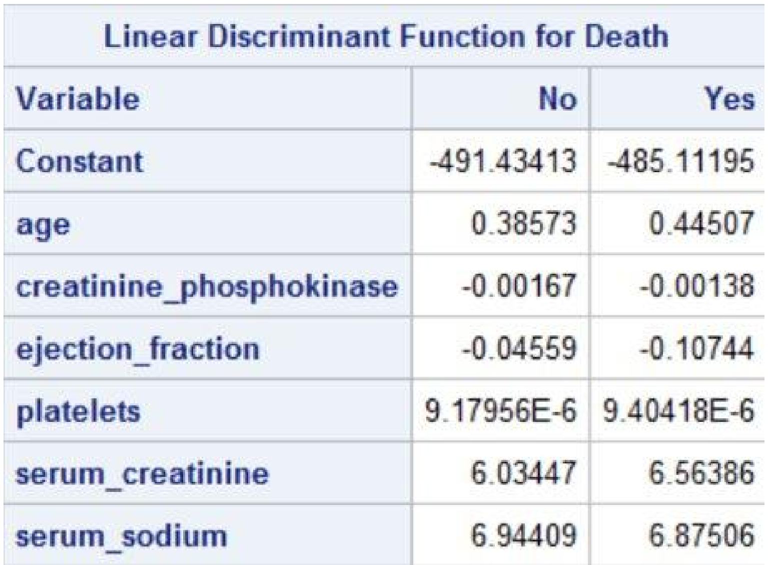
We calculate our T^2, which is equal to 79.94. We then place this into the following formula to compute our observed F Statistic. We also compute our F Critical Value.

$$F_{obs} = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 \sim F_{p, n_1 + n_2 - p - 1}$$

We have 96 observations in sample 1 and 203 observations in sample 2. There are 6 variables. Our observed F Statistic is 12.099, and our critical F statistic is 2.10 when alpha is 0.05. Since 12.09 is greater than 2.10, we reject the null hypothesis and conclude that there is a difference in the mean numeric values between those who died and those who did not die.

Question 4

Last, we use linear and quadratic discriminant analysis to properly classify patients into death=yes and death=no. To do so, we split the data into a training and test set, setting the respective proportions at 80% and 20%. We utilize all the numeric values except time as predictors. We then use proc discrim to fit our LDA model. Below are the equations of the linear discriminant scores.



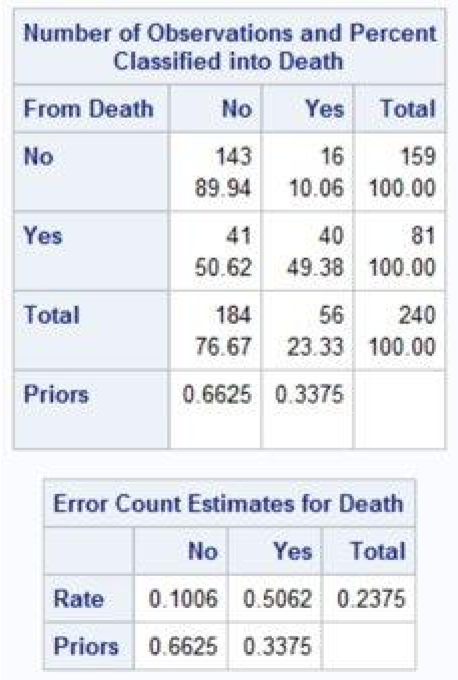
Linear Discriminant Function for Death

Variable	No	Yes
Constant	-491.43413	-485.11195
age	0.38573	0.44507
creatinine_phosphokinase	-0.00167	-0.00138
ejection_fraction	-0.04559	-0.10744
platelets	9.17956E-6	9.40418E-6
serum_creatinine	6.03447	6.56386
serum_sodium	6.94409	6.87506

$$\widehat{d^2_{No}} = -491.43 + 0.386_{age} - 0.0017_{creatinine_{phosphokinase}} - 0.046_{ejection_{fraction}} + 0.0000092_{platelets} + 6.034_{serum_{creatinine}} + 6.944_{serum\_sodium}$$

$$\widehat{d^2_{Yes}} = -485.11 - 0.0014_{creatinine\ phosphokinase} - 0.1074_{ejection\ fraction} + 0.0000094_{platelets} + 6.56_{serum\ creatinine} + 6.875_{serum\ sodium}$$

We have a classification summary and the error counts.

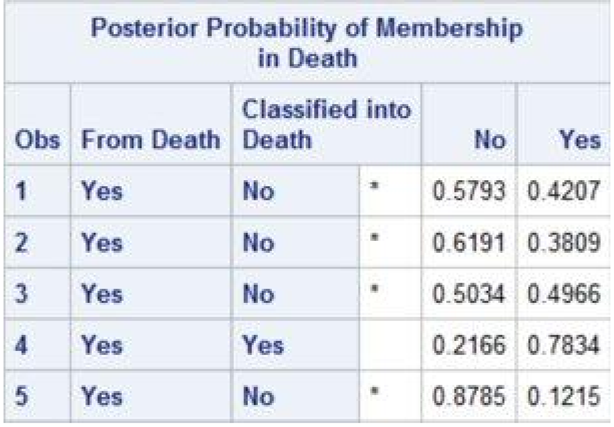


Number of Observations and Percent Classified into Death			
From Death	No	Yes	Total
No	143 89.94	16 10.06	159 100.00
Yes	41 50.62	40 49.38	81 100.00
Total	184 76.67	56 23.33	240 100.00
Priors	0.6625	0.3375	

Error Count Estimates for Death			
	No	Yes	Total
Rate	0.1006	0.5062	0.2375
Priors	0.6625	0.3375	

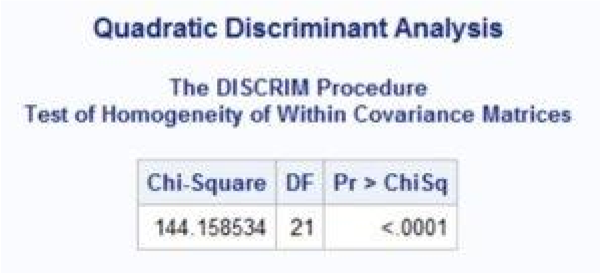
We see that 89.94% of the “No” observations were correctly classified as Death=No, and 49.38 % of the “Yes” observations were correctly classified as Death=Yes. There is a 23.75% error rate.

Then, we apply this model to the test data. Printing out only the first five observations, we see that for all the observations except the fourth, the patients that were assigned to the Death=No group were originally classified as Death=Yes. The fourth observation is the only one that is correctly classified to Death=Yes.



Posterior Probability of Membership in Death					
Obs	From Death	Classified into Death		No	Yes
1	Yes	No	*	0.5793	0.4207
2	Yes	No	*	0.6191	0.3809
3	Yes	No	*	0.5034	0.4966
4	Yes	Yes		0.2166	0.7834
5	Yes	No	*	0.8785	0.1215

Furthermore, we also conduct quadratic discriminant analysis. To do so, the variances must differ by group or else SAS would cease to run.

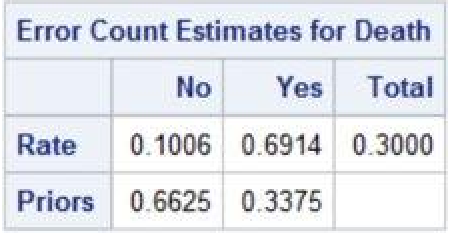


Quadratic Discriminant Analysis

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

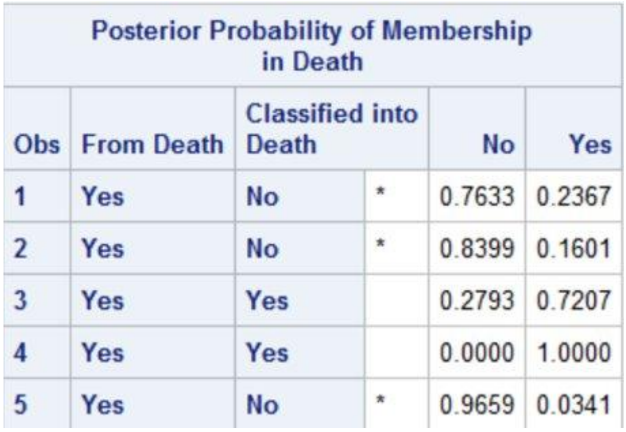
Chi-Square	DF	Pr > ChiSq
144.158534	21	<.0001

We see that the p-value is less than 0.0001, rejecting the null of equal variance. Hence, we continue with our QDA test.



Error Count Estimates for Death			
	No	Yes	Total
Rate	0.1006	0.6914	0.3000
Priors	0.6625	0.3375	

We see that QDA has an error rate of 0.3, which is way higher than that for LDA. We can thus conclude that LDA is significantly stronger than QDA.



Posterior Probability of Membership in Death

Obs	From Death	Classified into Death		No	Yes
1	Yes	No	*	0.7633	0.2367
2	Yes	No	*	0.8399	0.1601
3	Yes	Yes		0.2793	0.7207
4	Yes	Yes		0.0000	1.0000
5	Yes	No	*	0.9659	0.0341

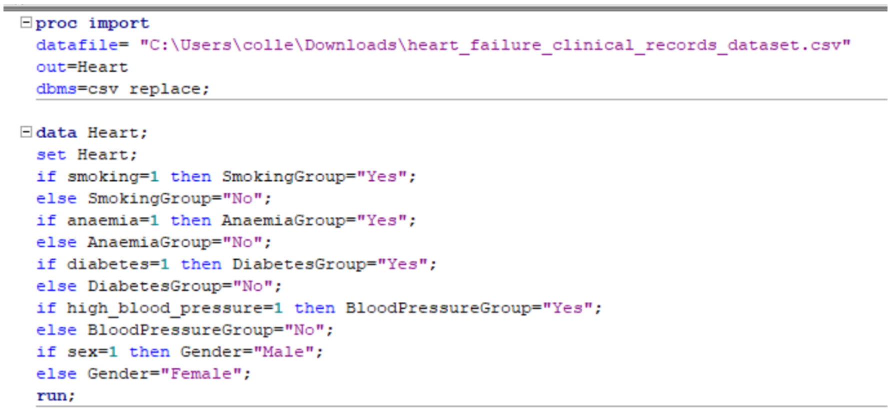
Furthermore, we look at the first five observations of the test set. For observations 1, 2, and 5, the patients who were classified as Death =No were originally classified as Death = Yes. For observations 3 and 4, patients were correctly classified as Death = Yes.

Conclusion

Throughout our study, we were able to successfully answer all four of our questions. First, in finding the optimal way to cluster the data, we found out that Ward's Method is the best clustering method because the clusters contain the most balanced proportions of patients who died. Second, multivariate analysis of variance proves that the mean numerical values between smokers and non-smokers do not significantly differ as the corresponding p-values for all the hypothesis tests are greater than 0.05. Third, Hotelling's 2 Sample T-Test shows that the mean numerical values between death and no death significantly differ. Lastly, we concluded that linear discriminant analysis accurately predicts the occurrence of survival or death, with only a 24% error rate.

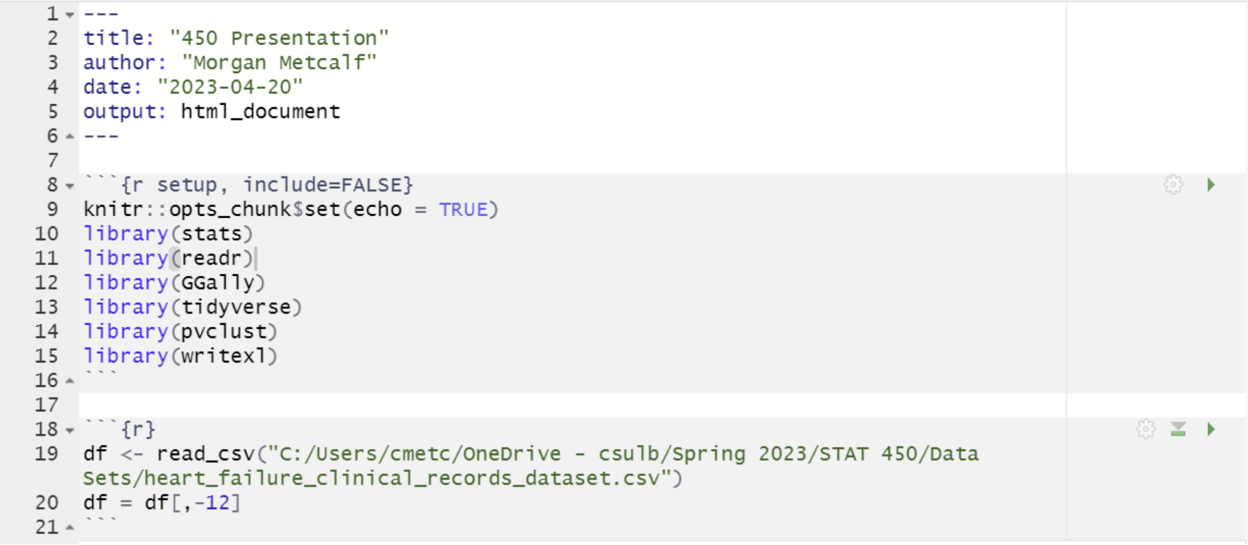
Our study, however, has some research limitations. For this project, we left out the time variable, which refers to when the follow up call was conducted (in terms of the number of days after the initial contact). We did not perform a time-based study, and time could very well be a factor in predicting survival or death. Had we conducted a time-based study, our results could have been different. Nonetheless, the results of this study are useful to help researchers understand who is most likely to die and survive. They can make recommendations to healthcare providers, who are responsible for providing their patients with quality health advice.

Appendix

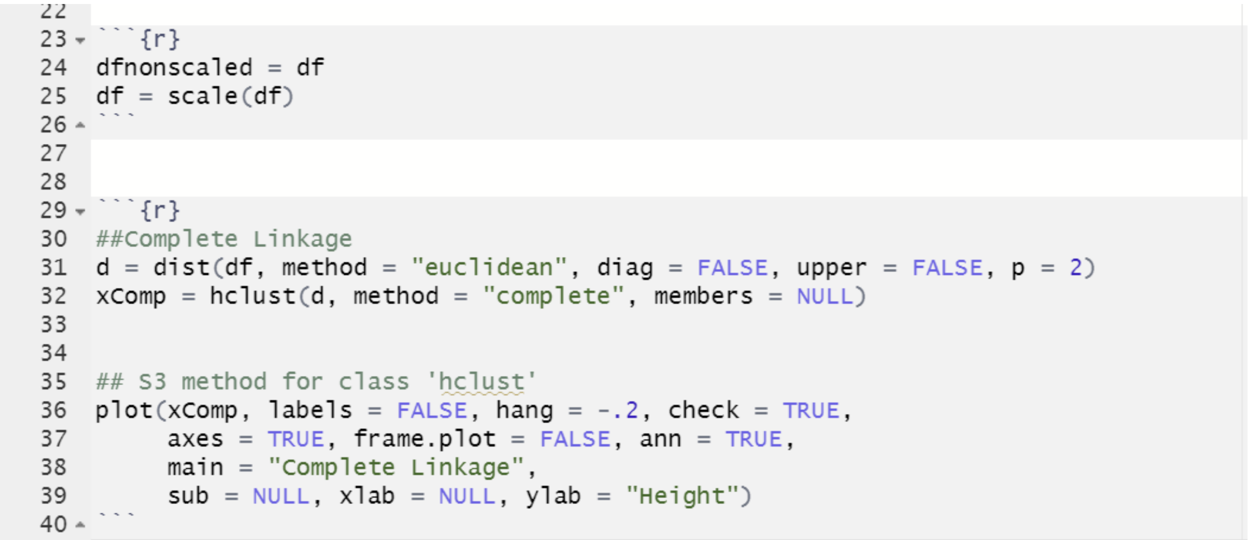


```
proc import
datafile= "C:\Users\colle\Downloads\heart_failure_clinical_records_dataset.csv"
out=Heart
dbms=csv replace;

data Heart;
set Heart;
if smoking=1 then SmokingGroup="Yes";
else SmokingGroup="No";
if anaemia=1 then AnaemiaGroup="Yes";
else AnaemiaGroup="No";
if diabetes=1 then DiabetesGroup="Yes";
else DiabetesGroup="No";
if high_blood_pressure=1 then BloodPressureGroup="Yes";
else BloodPressureGroup="No";
if sex=1 then Gender="Male";
else Gender="Female";
run;
```

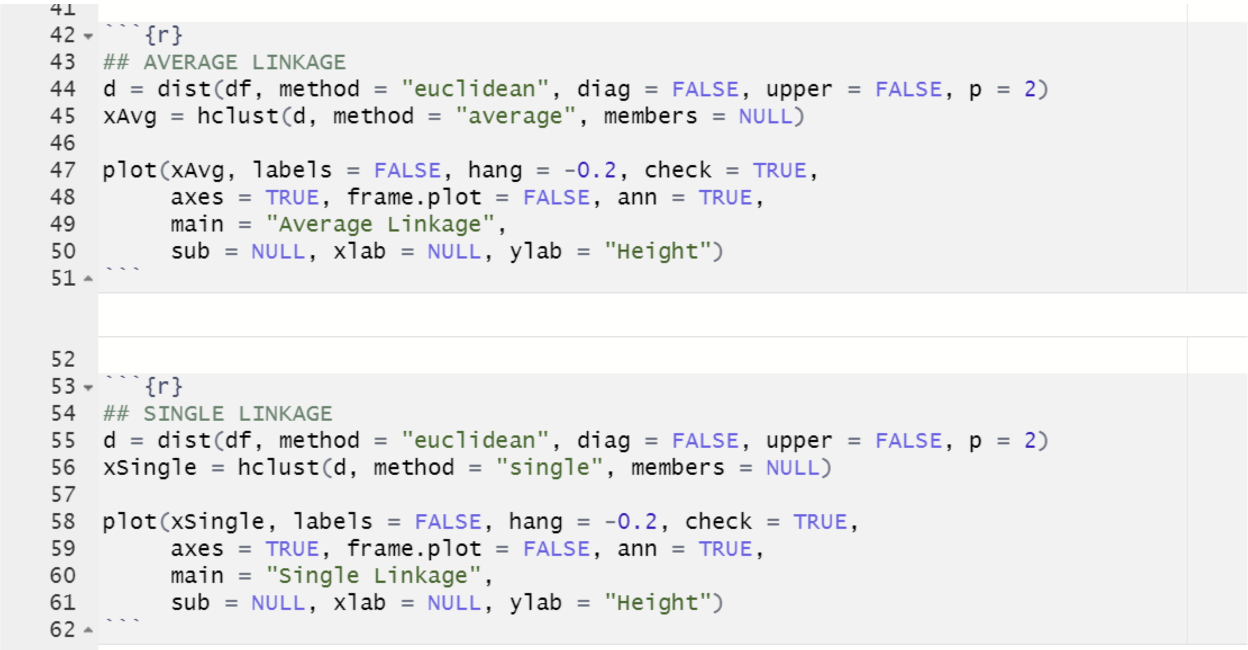
Question 1



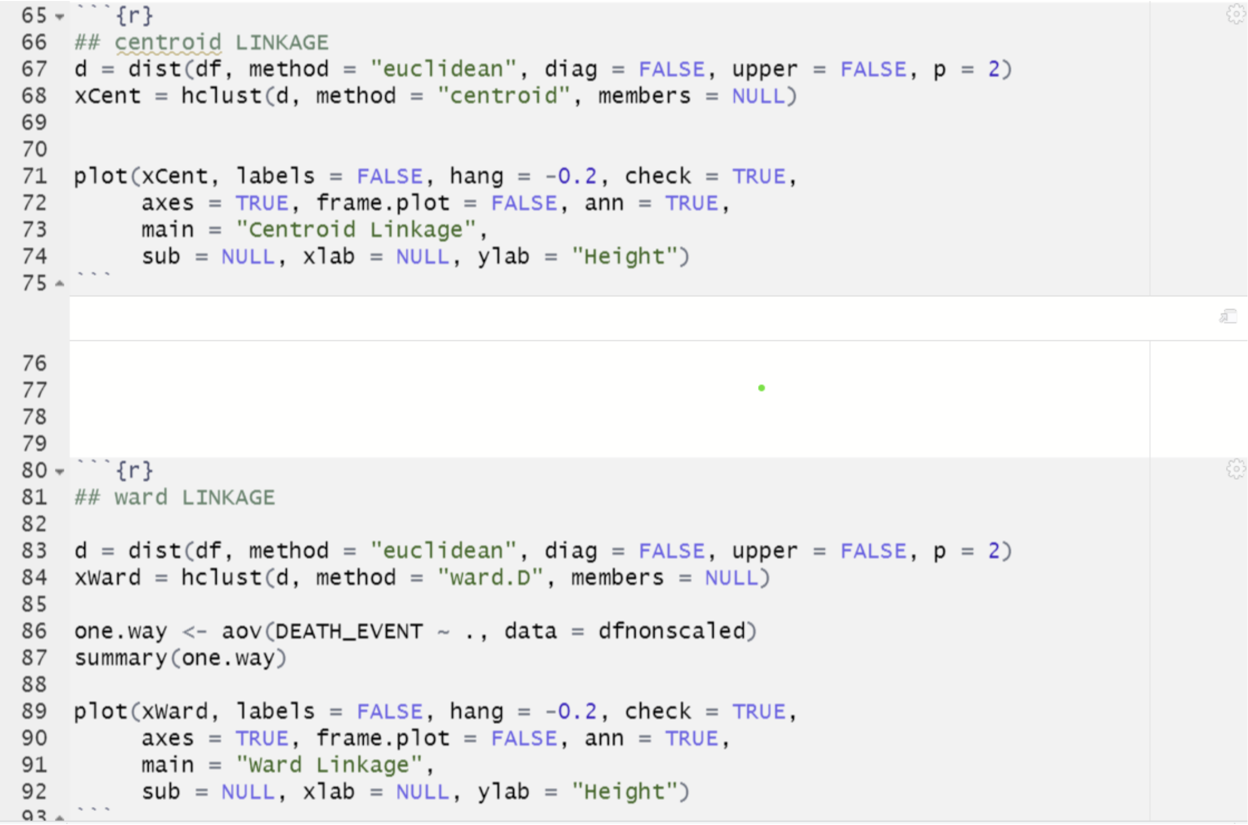
~~~
---
title: "450 Presentation"
author: "Morgan Metcalf"
date: "2023-04-20"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(stats)
library(readr)
library(GGally)
library(tidyverse)
library(pvclust)
library(writexl)
```

```{r}
df <- read_csv("C:/Users/cmetc/OneDrive - csulb/Spring 2023/STAT 450/Data
Sets/heart_failure_clinical_records_dataset.csv")
df = df[,-12]
```
~~~



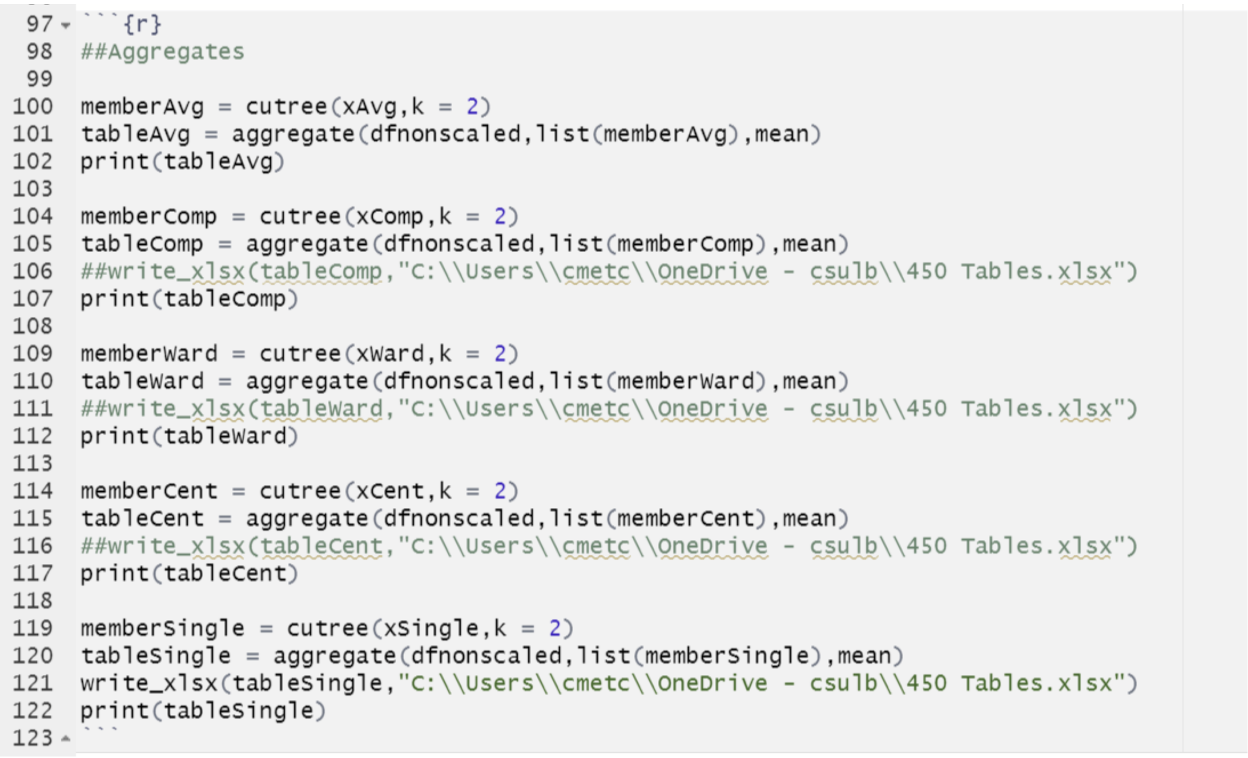
~~~

```{r}
dfnonscaled = df
df = scale(df)
```


```{r}
##Complete Linkage
d = dist(df, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
xComp = hclust(d, method = "complete", members = NULL)


## S3 method for class 'hclust'
plot(xComp, labels = FALSE, hang = -.2, check = TRUE,
     axes = TRUE, frame.plot = FALSE, ann = TRUE,
     main = "Complete Linkage",
     sub = NULL, xlab = NULL, ylab = "Height")
```
~~~



~~~
```{r}
## AVERAGE LINKAGE
d = dist(df, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
xAvg = hclust(d, method = "average", members = NULL)

plot(xAvg, labels = FALSE, hang = -0.2, check = TRUE,
      axes = TRUE, frame.plot = FALSE, ann = TRUE,
      main = "Average Linkage",
      sub = NULL, xlab = NULL, ylab = "Height")
~~~

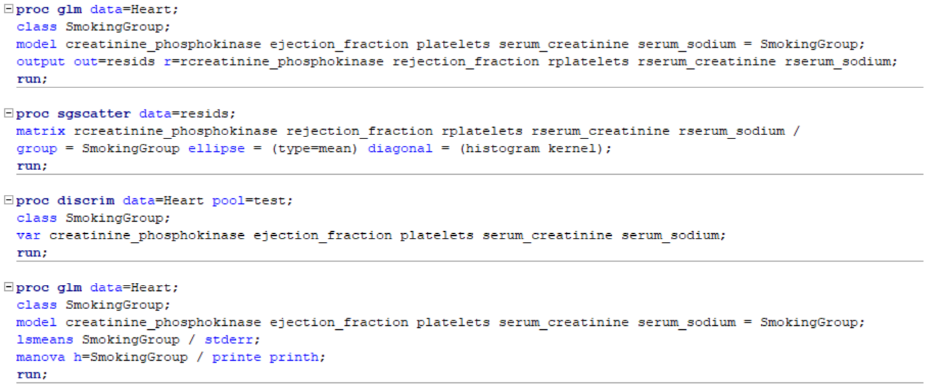
```{r}
## SINGLE LINKAGE
d = dist(df, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
xSingle = hclust(d, method = "single", members = NULL)

plot(xSingle, labels = FALSE, hang = -0.2, check = TRUE,
      axes = TRUE, frame.plot = FALSE, ann = TRUE,
      main = "Single Linkage",
      sub = NULL, xlab = NULL, ylab = "Height")
```
```



```
```{r}
## centroid LINKAGE
d = dist(df, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
xCent = hclust(d, method = "centroid", members = NULL)


plot(xCent, labels = FALSE, hang = -0.2, check = TRUE,
      axes = TRUE, frame.plot = FALSE, ann = TRUE,
      main = "Centroid Linkage",
      sub = NULL, xlab = NULL, ylab = "Height")
```

```{r}
## ward LINKAGE

d = dist(df, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
xWard = hclust(d, method = "ward.D", members = NULL)

one.way <- aov(DEATH_EVENT ~ ., data = dfnonscaled)
summary(one.way)

plot(xWard, labels = FALSE, hang = -0.2, check = TRUE,
      axes = TRUE, frame.plot = FALSE, ann = TRUE,
      main = "Ward Linkage",
      sub = NULL, xlab = NULL, ylab = "Height")
```
```
```



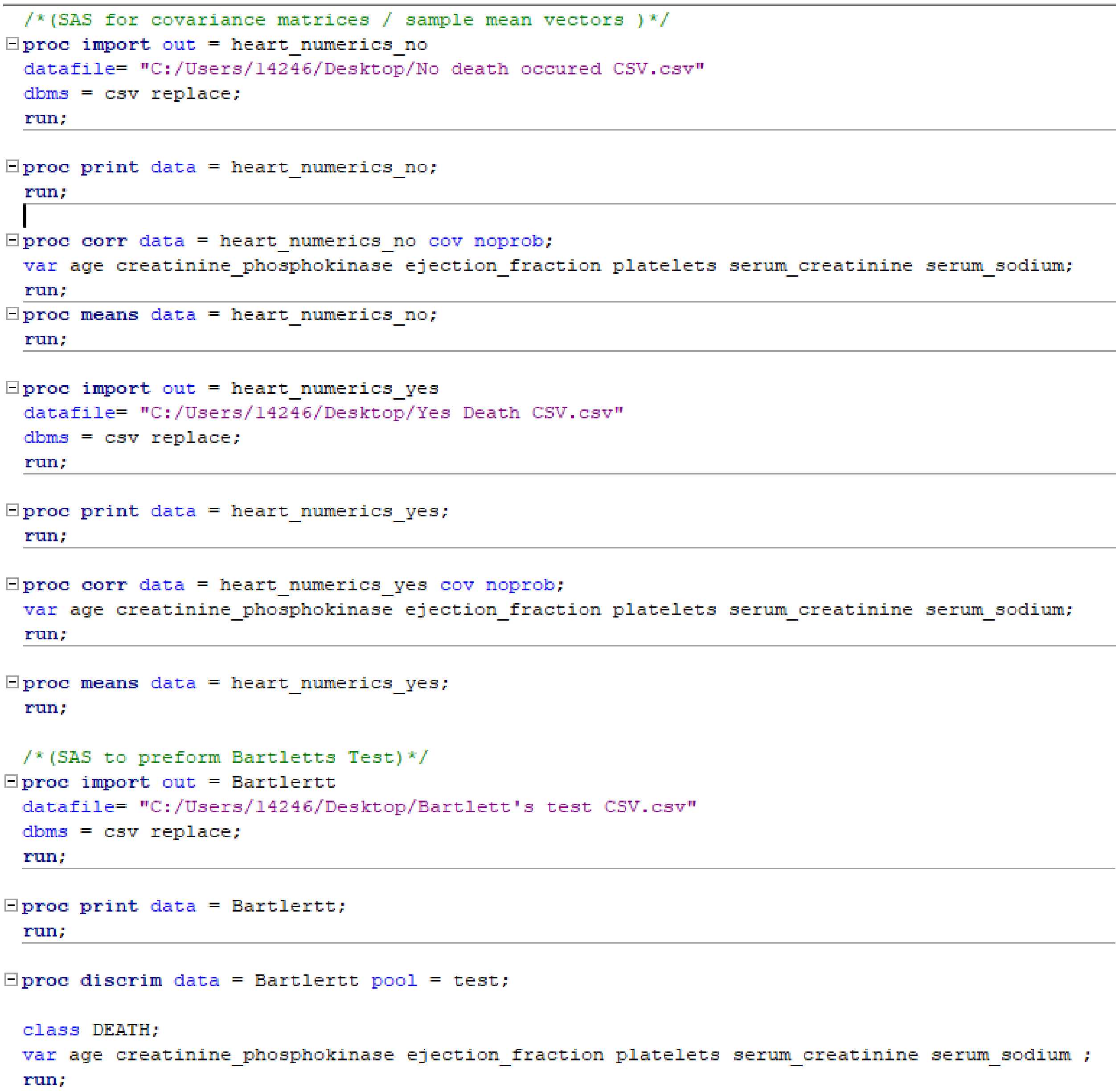
~~~
```{r}
##Aggregates

memberAvg = cutree(xAvg,k = 2)
tableAvg = aggregate(dfnonscaled,list(memberAvg),mean)
print(tableAvg)

memberComp = cutree(xComp,k = 2)
tableComp = aggregate(dfnonscaled,list(memberComp),mean)
##write_xlsx(tableComp,"C:\\Users\\cmetc\\OneDrive - csulb\\450 Tables.xlsx")
print(tableComp)

memberWard = cutree(xWard,k = 2)
tableWard = aggregate(dfnonscaled,list(memberWard),mean)
##write_xlsx(tableWard,"C:\\Users\\cmetc\\OneDrive - csulb\\450 Tables.xlsx")
print(tableWard)

memberCent = cutree(xCent,k = 2)
tableCent = aggregate(dfnonscaled,list(memberCent),mean)
##write_xlsx(tableCent,"C:\\Users\\cmetc\\OneDrive - csulb\\450 Tables.xlsx")
print(tableCent)

memberSingle = cutree(xSingle,k = 2)
tableSingle = aggregate(dfnonscaled,list(memberSingle),mean)
write_xlsx(tableSingle,"C:\\Users\\cmetc\\OneDrive - csulb\\450 Tables.xlsx")
print(tableSingle)
```
~~~

Question 2

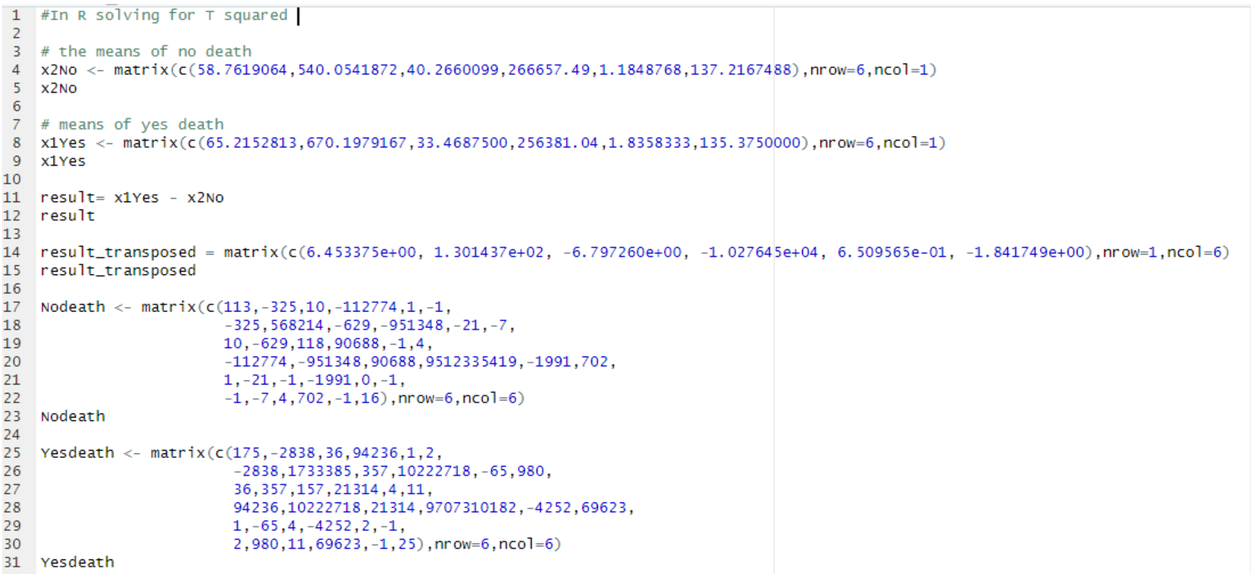


```
proc glm data=Heart;
 class SmokingGroup;
 model creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium = SmokingGroup;
 output out=resids r=rcreatinine_phosphokinase rejection_fraction rplatelets rserum_creatinine rserum_sodium;
 run;

proc sgscatter data=resids;
 matrix rcreatinine_phosphokinase rejection_fraction rplatelets rserum_creatinine rserum_sodium /
 group = SmokingGroup ellipse = (type=mean) diagonal = (histogram kernel);
 run;

proc discrim data=Heart pool=test;
 class SmokingGroup;
 var creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium;
 run;

proc glm data=Heart;
 class SmokingGroup;
 model creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium = SmokingGroup;
 lsmeans SmokingGroup / stderr;
 manova h=SmokingGroup / printe printh;
 run;
```

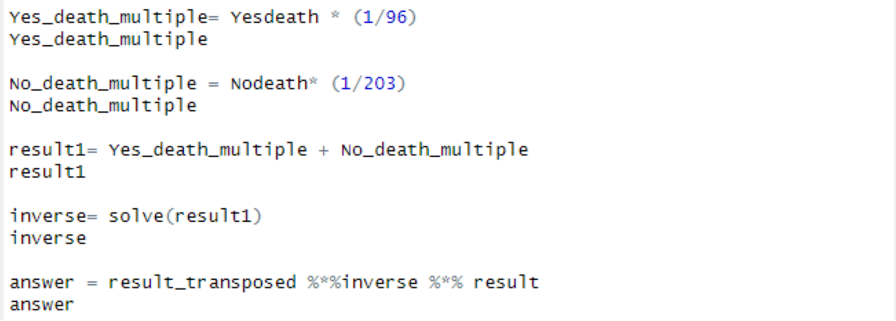
Question 3



```
/*(SAS for covariance matrices / sample mean vectors )*/
proc import out = heart_numerics_no
datafile= "C:/Users/14246/Desktop/No death occured CSV.csv"
dbms = csv replace;
run;

proc print data = heart_numerics_no;
run;

proc corr data = heart_numerics_no cov noprob;
var age creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium;
run;
proc means data = heart_numerics_no;
run;

proc import out = heart_numerics_yes
datafile= "C:/Users/14246/Desktop/Yes Death CSV.csv"
dbms = csv replace;
run;

proc print data = heart_numerics_yes;
run;

proc corr data = heart_numerics_yes cov noprob;
var age creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium;
run;

proc means data = heart_numerics_yes;
run;

/*(SAS to preform Bartletts Test)*/
proc import out = Bartlertt
datafile= "C:/Users/14246/Desktop/Bartlett's test CSV.csv"
dbms = csv replace;
run;

proc print data = Bartlertt;
run;

proc discrim data = Bartlertt pool = test;

class DEATH;
var age creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium ;
run;
```



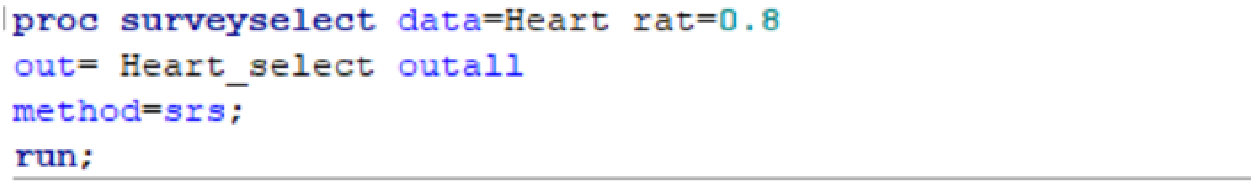
```
#In R solving for T squared

# the means of no death
x2No <- matrix(c(58.7619064,540.0541872,40.2660099,266657.49,1.1848768,137.2167488),nrow=6,ncol=1)
x2No

# means of yes death
x1Yes <- matrix(c(65.2152813,670.1979167,33.4687500,256381.04,1.8358333,135.3750000),nrow=6,ncol=1)
x1Yes

result= x1Yes - x2No
result

result_transposed = matrix(c(6.453375e+00, 1.301437e+02, -6.797260e+00, -1.027645e+04, 6.509565e-01, -1.841749e+00),nrow=1,ncol=6)
result_transposed

Nodeath <- matrix(c(113,-325,10,-112774,1,-1,
                    -325,568214,-629,-951348,-21,-7,
                    10,-629,118,90688,-1,4,
                    -112774,-951348,90688,9512335419,-1991,702,
                    1,-21,-1,-1991,0,-1,
                    -1,-7,4,702,-1,16),nrow=6,ncol=6)
Nodeath

Yesdeath <- matrix(c(175,-2838,36,94236,1,2,
                     -2838,1733385,357,10222718,-65,980,
                     36,357,157,21314,4,11,
                     94236,10222718,21314,9707310182,-4252,69623,
                     1,-65,4,-4252,2,-1,
                     2,980,11,69623,-1,25),nrow=6,ncol=6)
Yesdeath
```



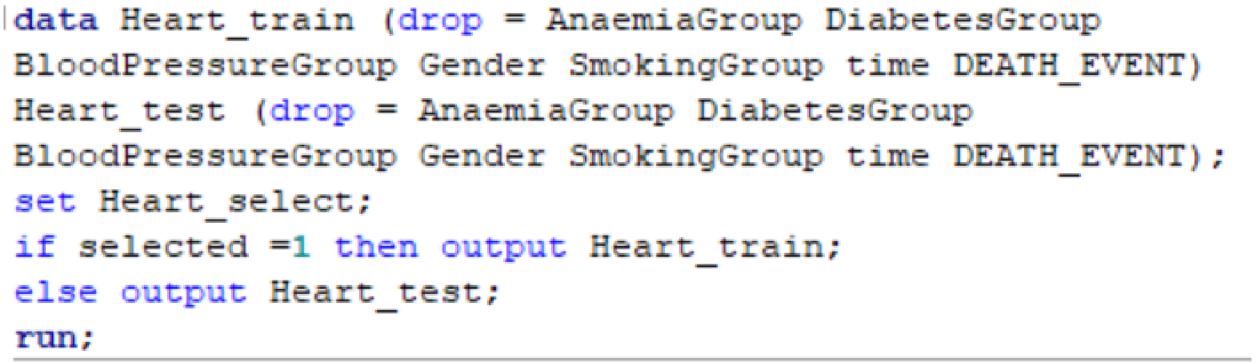
```
Yes_death_multiple= Yesdeath * (1/96)
Yes_death_multiple

No_death_multiple = Nodeath* (1/203)
No_death_multiple

result1= Yes_death_multiple + No_death_multiple
result1

inverse= solve(result1)
inverse

answer = result_transposed %*%inverse %*% result
answer
```

Question 4

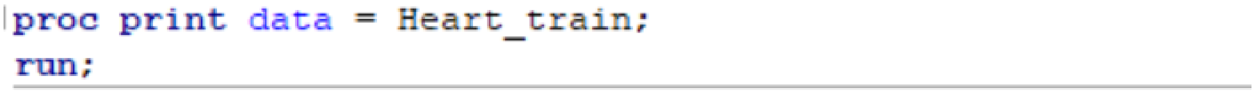
Splitting data into training set and test set



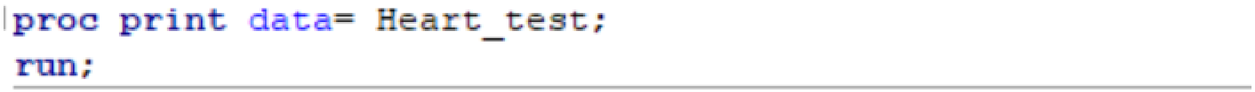
```
proc surveyselect data=Heart rat=0.8
out= Heart_select outall
method=srs;
run;
```



```
data Heart_train (drop = AnaemiaGroup DiabetesGroup
BloodPressureGroup Gender SmokingGroup time DEATH_EVENT)
Heart_test (drop = AnaemiaGroup DiabetesGroup
BloodPressureGroup Gender SmokingGroup time DEATH_EVENT);
set Heart_select;
if selected =1 then output Heart_train;
else output Heart_test;
run;
```

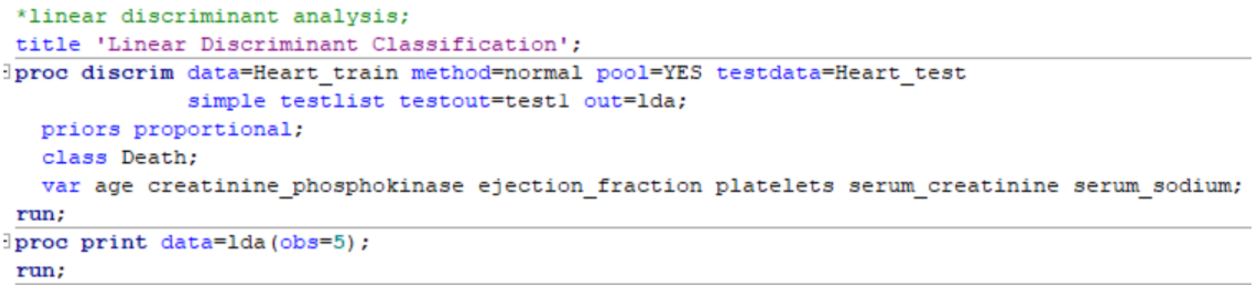


```
proc print data = Heart_train;
run;
```



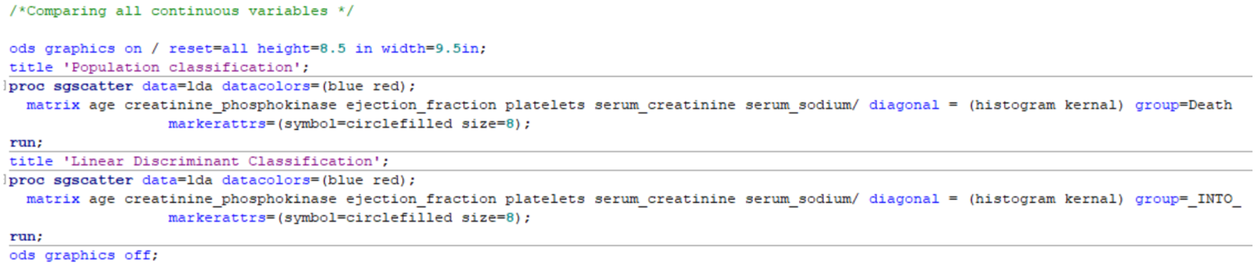
```
proc print data= Heart_test;
run;
```

Running LDA



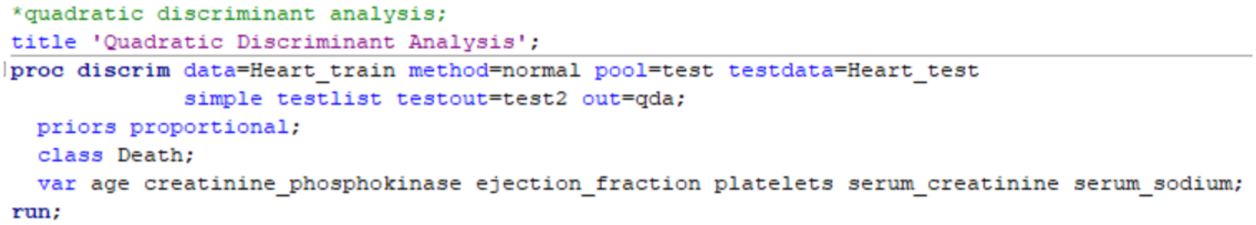
```
*linear discriminant analysis;
title 'Linear Discriminant Classification';
proc discrim data=Heart_train method=normal pool=YES testdata=Heart_test
             simple testlist testout=testl out=lda;
  priors proportional;
  class Death;
  var age creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium;
run;
proc print data=lda(obs=5);
run;
```

Creating graphical display



```
/*Comparing all continuous variables */

ods graphics on / reset=all height=8.5 in width=9.5in;
title 'Population classification';
proc sgscatter data=lda datacolors=(blue red);
  matrix age creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium/ diagonal = (histogram kernal) group=Death
                 markerattrs=(symbol=circlefilled size=8);
run;
title 'Linear Discriminant Classification';
proc sgscatter data=lda datacolors=(blue red);
  matrix age creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium/ diagonal = (histogram kernal) group=_INTO_
                 markerattrs=(symbol=circlefilled size=8);
run;
ods graphics off;
```

Running QDA



```
*quadratic discriminant analysis;
title 'Quadratic Discriminant Analysis';
proc discrim data=Heart_train method=normal pool=test testdata=Heart_test
             simple testlist testout=test2 out=qda;
  priors proportional;
  class Death;
  var age creatinine_phosphokinase ejection_fraction platelets serum_creatinine serum_sodium;
run;
```

References

Dataset Source:
https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

Original Journal Article:
Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak* 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5