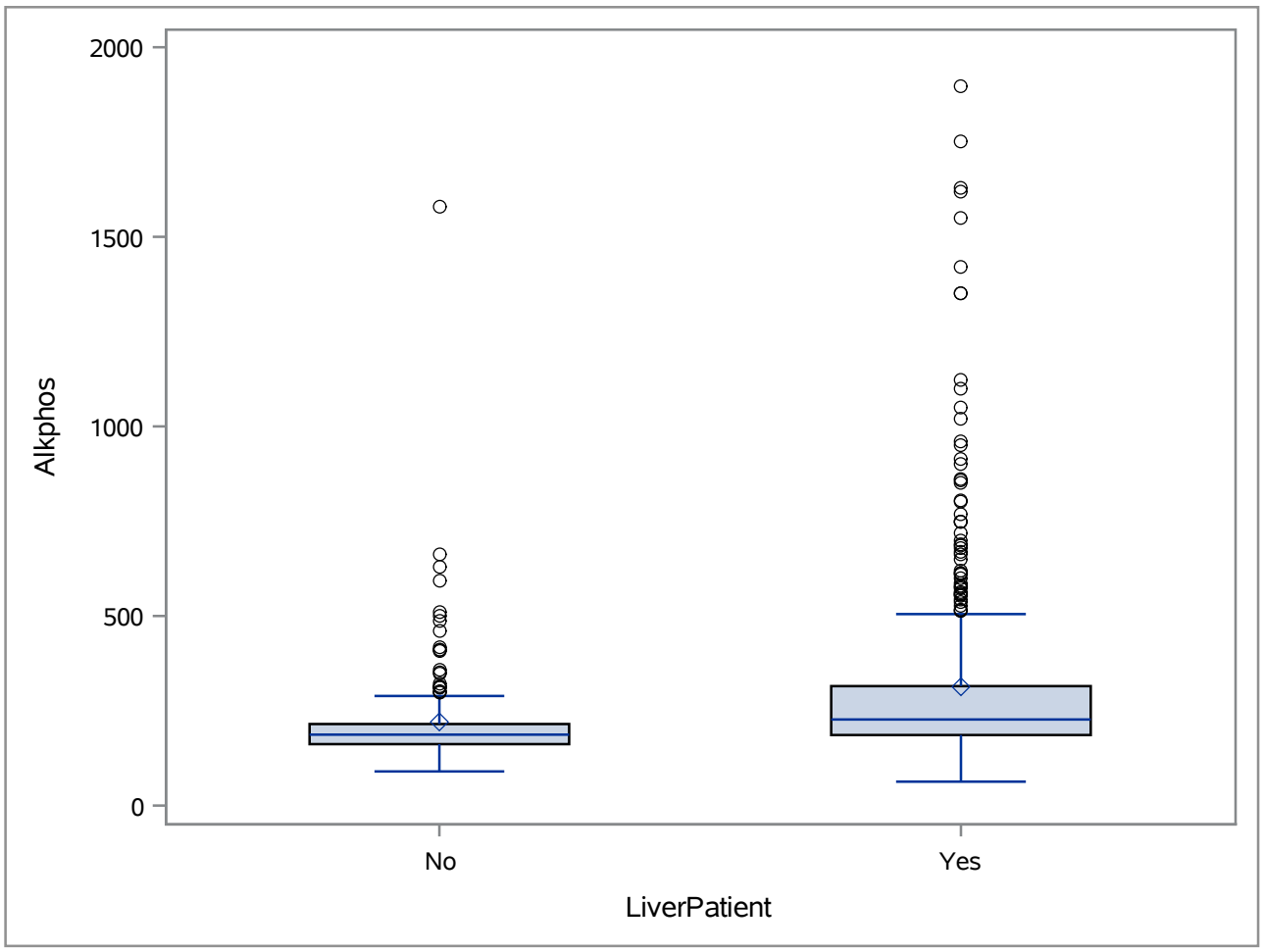
The Indian Liver Patient Dataset is from UCI's Machine Learning Database. It contains information on people who were and were not liver patients, including age, gender, and clinical measures. Some of those clinical measures include total Bilirubin, proteins, Aspartate, and Alkaline Phosphotase. The main reason for analyzing this data is to see if any of these measures are indicators of liver disease.

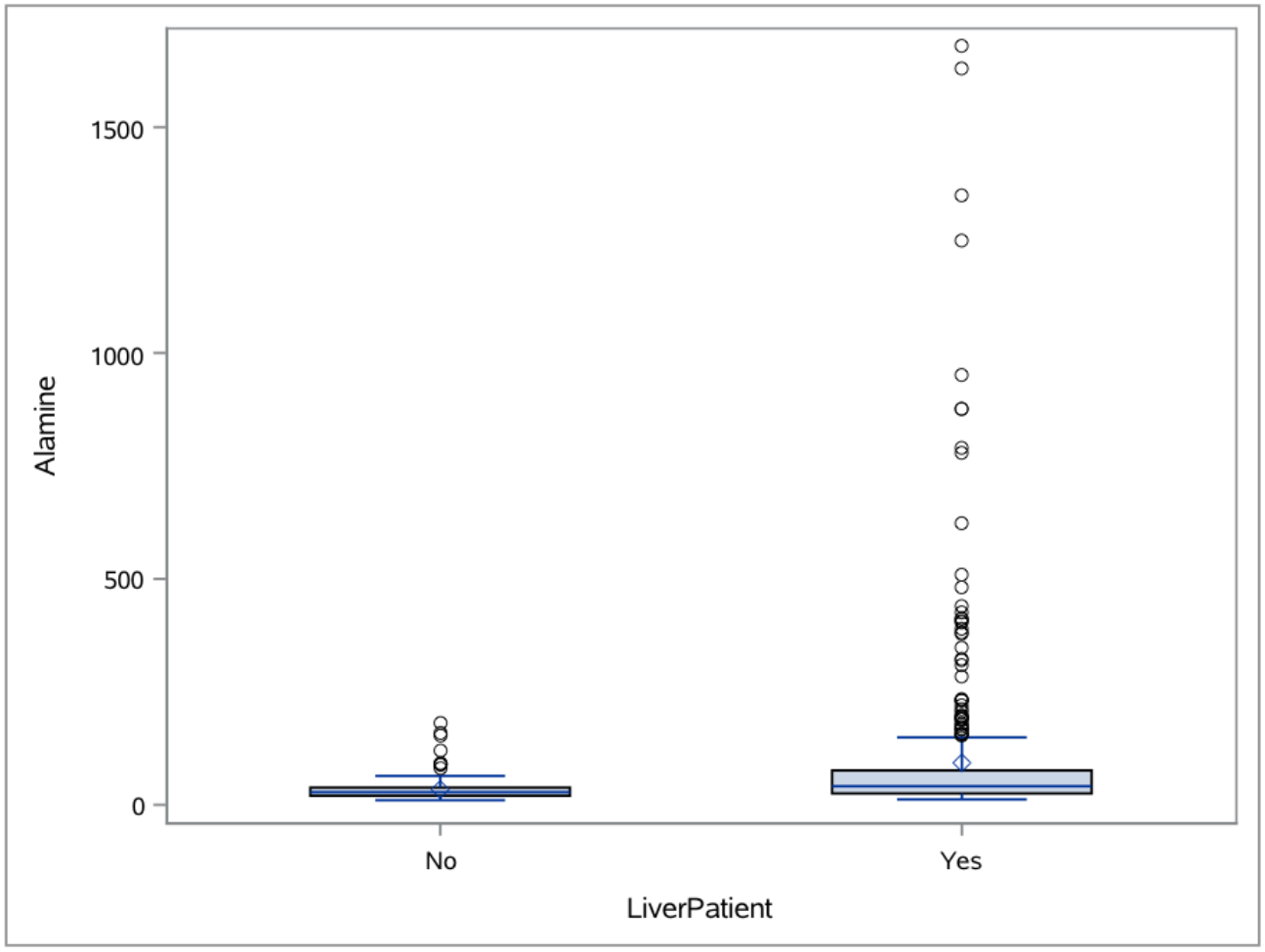
The first thing I did was clean the data. I removed any observations that had missing cell counts as well as any outliers. I then looked at some general statistics and plots to see any difference between those who were liver patients and those who were not.

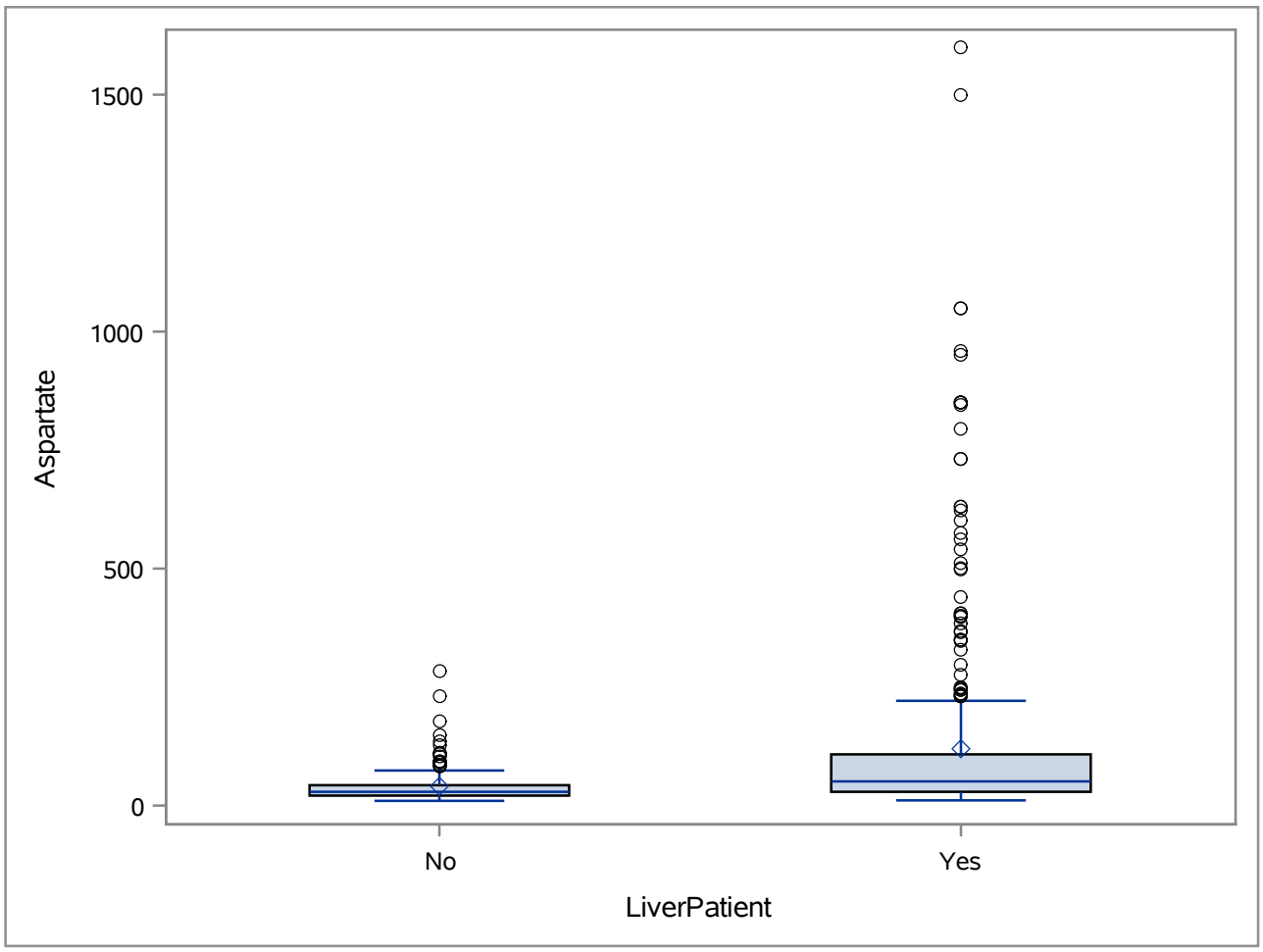
1)By looking at the mean values of clinical measures and ages by liver patients and those who are not, it seems that liver patients have much higher Alkphos, Alamine, and Aspartate levels than non liver patients. Liver patients also have slightly higher TB, DB, and age levels.

The MEANS Procedure

LiverPatient	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
No	165	TB	165	1.1448485	1.0096127	0.5000000	7.3000000
		DB	165	0.3963636	0.5219442	0.1000000	3.6000000
		Alkphos	165	220.6848485	141.5278050	90.0000000	1580.00
		Alamine	165	33.8363636	25.1462287	10.0000000	181.0000000
		Aspartate	165	40.7636364	36.5631580	10.0000000	285.0000000
		TP	165	6.5393939	1.0533163	3.7000000	9.2000000
		ALB	165	3.3393939	0.7785773	1.4000000	5.0000000
		AGRatio	165	1.0295758	0.2872522	0.3700000	1.9000000
		Age	165	41.3636364	17.0591120	4.0000000	85.0000000
Yes	409	TB	409	3.8799511	5.9596798	0.4000000	32.6000000
		DB	409	1.8660147	3.0957346	0.1000000	18.3000000
		Alkphos	409	313.3300733	252.3581160	63.0000000	1896.00
		Alamine	409	92.8508557	184.0133461	12.0000000	1680.00
		Aspartate	409	119.8924205	199.5128458	11.0000000	1600.00
		TP	409	6.4579462	1.1016662	2.7000000	9.6000000
		ALB	409	3.0621027	0.7906354	0.9000000	5.5000000
		AGRatio	409	0.9160636	0.3259254	0.3000000	2.8000000
		Age	409	46.1907090	15.6886937	7.0000000	90.0000000


2000
1500
1000
500
0
Alkphos
No
Yes
LiverPatient


1500
1000
500
0
Alamine
No
Yes
LiverPatient



I then wanted to see if the clinical measures, gender, and ages could predict if a patient had liver disease.

2) By using backwards selection, the best model I obtained includes DB, alamine, TP, ALB, and age. When DB increases by one unit, the odds of having liver disease is 1.720 times higher. When alamine increases by one unit, the odds of having liver disease is 1.016 times higher. When TP increases by one unit, the odds of having liver disease is 1.543 times higher. When ALB increases by one unit, the odds of having liver disease is 0.514 times lower. When age increases by one unit, the odds of having liver disease is 1.018 times lower.

The LOGISTIC Procedure

Model Information	
Data Set	WORK.LIVER2
Response Variable	LiverPatient
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	690.638	589.209
SC	694.990	637.088
-2 Log L	688.638	567.209

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	121.4288	10	<.0001
Score	71.1155	10	<.0001
Wald	53.9247	10	<.0001

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	690.638	587.224
SC	694.990	630.750
-2 Log L	688.638	567.224

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	121.4140	9	<.0001
Score	70.0223	9	<.0001
Wald	53.7659	9	<.0001

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	117.4463	6	<.0001
Score	62.7850	6	<.0001
Wald	51.9754	6	<.0001

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	690.638	585.861
SC	694.990	611.977
-2 Log L	688.638	573.861

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	114.7765	5	<.0001
Score	62.7107	5	<.0001
Wald	51.2081	5	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.6709	0.7758	4.6386	0.0313
DB	1	0.5425	0.1740	9.7253	0.0018
Alamine	1	0.0157	0.00391	16.1446	<.0001
TP	1	0.4339	0.1755	6.1082	0.0135
ALB	1	-0.6652	0.2498	7.0878	0.0078
Age	1	0.0180	0.00636	8.0401	0.0046

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
DB	1.720	1.223	2.419
Alamine	1.016	1.008	1.024
TP	1.543	1.094	2.177
ALB	0.514	0.315	0.839
Age	1.018	1.006	1.031

I then wanted to see how the clinical measures , age, and gender relate to total proteins. To do that, I created a general linear model.

3) In order to obtain the best model, I used backwards selection. I started with all the terms in the model, looked at the type 3 analysis, removed the most insignificant term, and then I refitted the model. I did this until all the terms were significant. The final model includes DB, Alamine, Aspartate, ALB, and AGRatio as predictors for Total Proteins. When DB, Aspartate, and ALB increase, total proteins increase, but when Alamine and AGRatio increase, total proteins decrease.

The GLM Procedure

Dependent Variable: TP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	74	400.9098889	5.4177012	19.20	<.0001
Error	334	94.2667859	0.2822359		
Corrected Total	408	495.1766748			

R-Square	Coeff Var	Root MSE	TP Mean
0.809630	8.226435	0.531259	6.457946

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TB	1	0.2489776	0.2489776	0.88	0.3483
DB	1	0.0646814	0.0646814	0.23	0.6324
Alkphos	1	0.0654924	0.0654924	0.23	0.6303
Alamine	1	2.0041541	2.0041541	7.10	0.0081
Aspartate	1	0.7483514	0.7483514	2.65	0.1044
ALB	1	288.9718152	288.9718152	1023.87	<.0001
AGRatio	1	53.4555659	53.4555659	189.40	<.0001
Gender	1	0.0151422	0.0151422	0.05	0.8170
Age	66	13.8485240	0.2098261	0.74	0.9278

The GLM Procedure

Dependent Variable: TP

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TB	1	0.1693399	0.1693399	0.63	0.4291
DB	1	0.1791510	0.1791510	0.66	0.4161
Alkphos	1	0.0001833	0.0001833	0.00	0.9792
Alamine	1	2.6823425	2.6823425	9.92	0.0018
Aspartate	1	0.9367798	0.9367798	3.47	0.0634
ALB	1	356.0312366	356.0312366	1317.23	<.0001
AGRatio	1	71.1140180	71.1140180	263.10	<.0001
Gender	1	0.0000072	0.0000072	0.00	0.9959

The GLM Procedure

Dependent Variable: TP

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TB	1	0.1699572	0.1699572	0.63	0.4277
DB	1	0.1804845	0.1804845	0.67	0.4137
Alkphos	1	0.0001944	0.0001944	0.00	0.9786
Alamine	1	2.6832072	2.6832072	9.95	0.0017
Aspartate	1	0.9375310	0.9375310	3.48	0.0629
ALB	1	362.5406113	362.5406113	1344.66	<.0001
AGRatio	1	71.4922588	71.4922588	265.16	<.0001

Dependent Variable: TP

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TB	1	0.1698605	0.1698605	0.63	0.4272
DB	1	0.1802961	0.1802961	0.67	0.4134
Alamine	1	2.6907263	2.6907263	10.00	0.0017
Aspartate	1	0.9381447	0.9381447	3.49	0.0625
ALB	1	362.7043233	362.7043233	1348.62	<.0001
AGRatio	1	73.2702282	73.2702282	272.44	<.0001

The GLM Procedure

Dependent Variable: TP

3)Final model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	386.8913029	77.3782606	287.97	<.0001
Error	403	108.2853720	0.2686982		
Corrected Total	408	495.1766748			

R-Square	Coeff Var	Root MSE	TP Mean
0.781320	8.026717	0.518361	6.457946

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DB	1	13.3919909	13.3919909	49.84	<.0001
Alamine	1	2.7828237	2.7828237	10.36	0.0014
Aspartate	1	1.0925022	1.0925022	4.07	0.0444
ALB	1	362.5345141	362.5345141	1349.23	<.0001
AGRatio	1	73.8558776	73.8558776	274.87	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.031557048	0.11140347	27.21	<.0001
DB	0.062649873	0.00887422	7.06	<.0001
Alamine	-0.000829866	0.00025787	-3.22	0.0014
Aspartate	0.000490597	0.00024330	2.02	0.0444
ALB	1.609156431	0.04380826	36.73	<.0001
AGRatio	-1.746258400	0.10532909	-16.58	<.0001

4)Since the researcher is interested in groupings of the measures, I used cluster analysis. However, the measures did not cluster nicely, so I first used principal component analysis to reduce the dimensions. I ended up keeping the first 3 principal components. By looking at the graphs, liver patients and non liver patients overlap quite a bit. Liver patients tend to be higher in principal components 1 and 2, while on average they are about equal to non liver patients in component 3.

The PRINCOMP Procedure

Observations	574
Variables	9

Simple Statistics

	TB	DB	Alkphos	Alamine	Aspartate	TP	ALB	AGRatio	Age
Mean	3.093728223	1.443554007	286.6986063	75.8867596	97.1463415	6.481358885	3.141811847	0.9486933798	44.80313589
StD	5.207381226	2.710174466	229.8697592	158.1327321	173.2352339	1.087699383	0.796476222	0.3192156619	16.22748161

Correlation Matrix

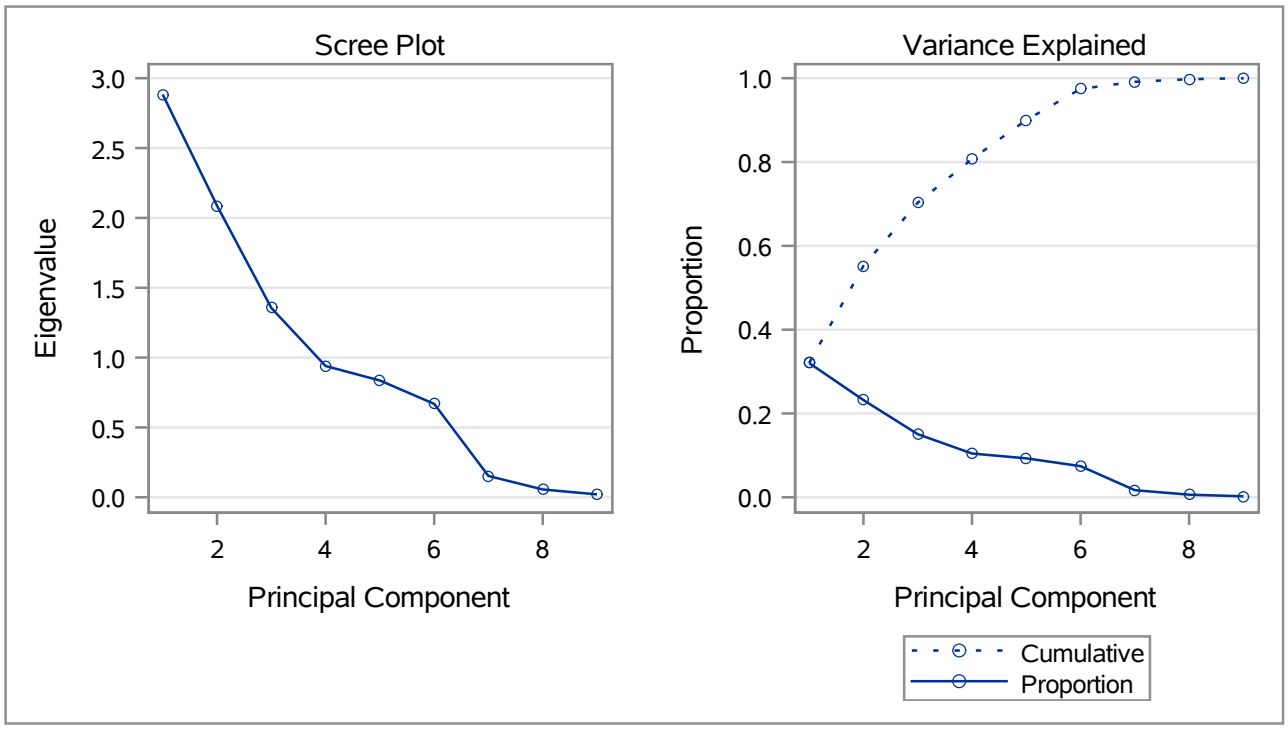
	TB	DB	Alkphos	Alamine	Aspartate	TP	ALB	AGRatio	Age
TB	1.0000	0.9785	0.2475	0.2455	0.3538	-.0148	-.2371	-.2041	0.0122
DB	0.9785	1.0000	0.2466	0.2270	0.3247	-.0097	-.2279	-.1917	0.0187
Alkphos	0.2475	0.2466	1.0000	0.1198	0.1372	-.0294	-.1651	-.2375	0.0863
Alamine	0.2455	0.2270	0.1198	1.0000	0.8400	-.0405	-.0261	-.0086	-.1028
Aspartate	0.3538	0.3247	0.1372	0.8400	1.0000	-.0464	-.1036	-.0817	-.0750
TP	-.0148	-.0097	-.0294	-.0405	-.0464	1.0000	0.7874	0.2408	-.1875
ALB	-.2371	-.2279	-.1651	-.0261	-.1036	0.7874	1.0000	0.6888	-.2641
AGRatio	-.2041	-.1917	-.2375	-.0086	-.0817	0.2408	0.6888	1.0000	-.2156
Age	0.0122	0.0187	0.0863	-.1028	-.0750	-.1875	-.2641	-.2156	1.0000

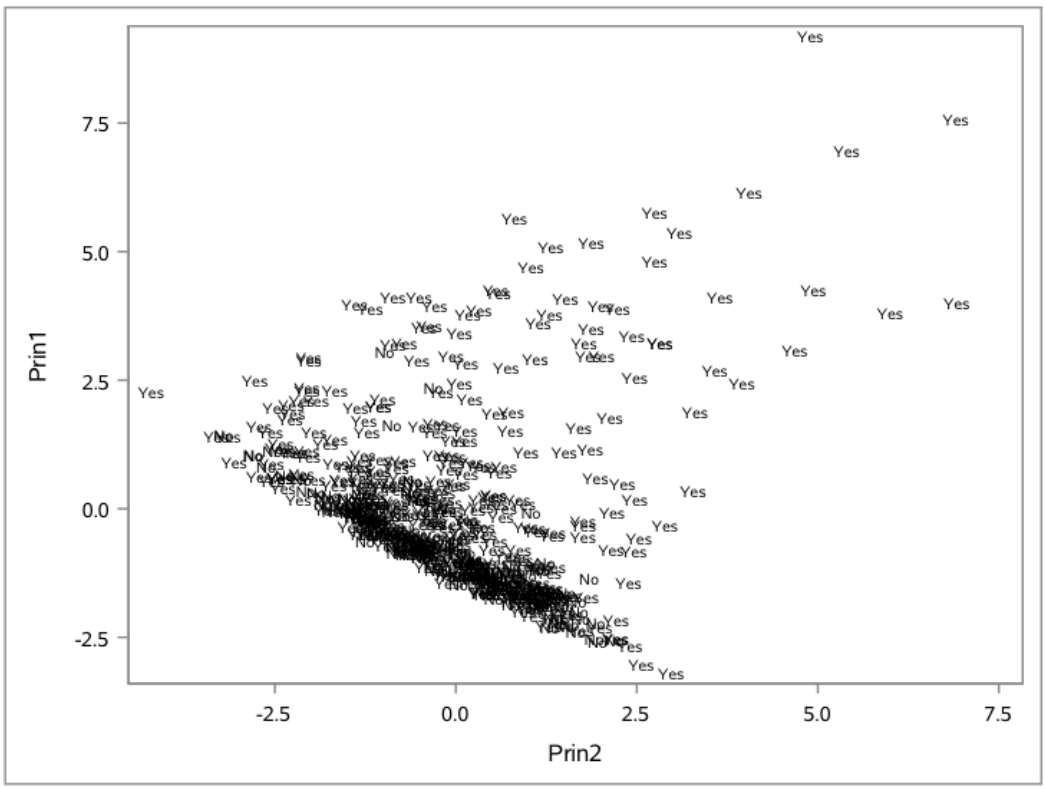
Eigenvalues of the Correlation Matrix

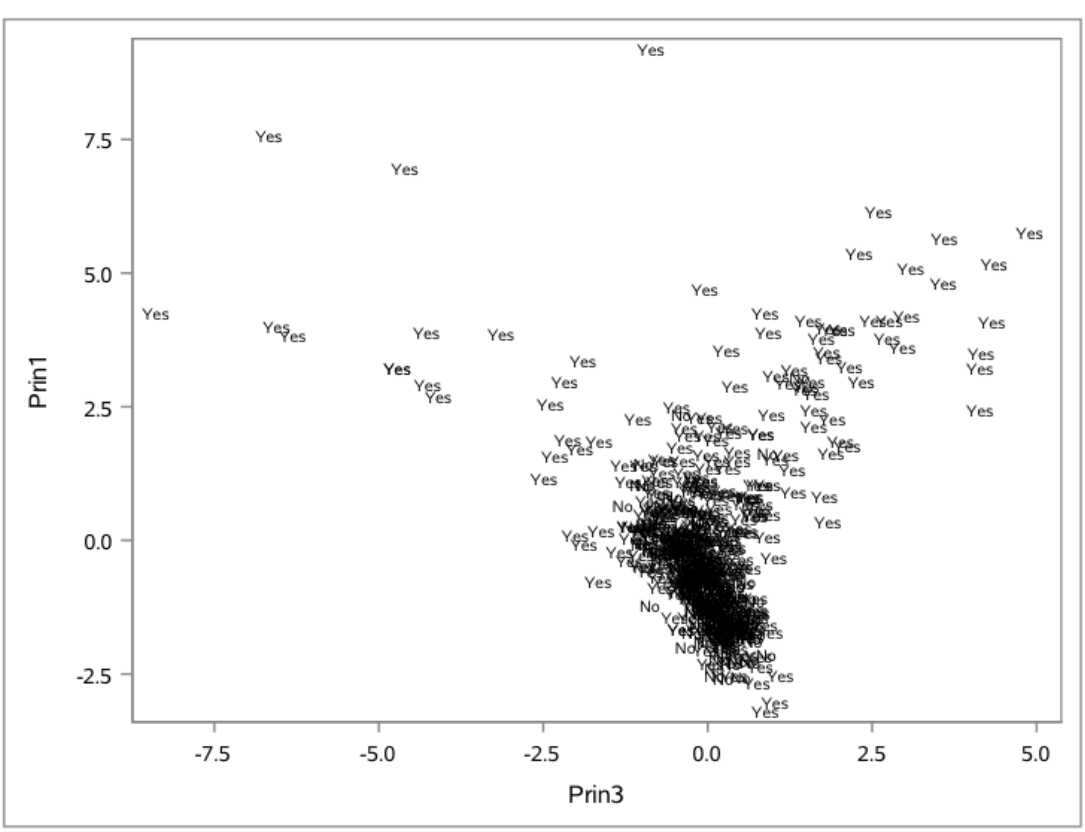
	Eigenvalue	Difference	Proportion	Cumulative
1	2.88320909	0.79557510	0.3204	0.3204
2	2.08763400	0.73243269	0.2320	0.5523
3	1.35520131	0.41539455	0.1506	0.7029
4	0.93980676	0.10379195	0.1044	0.8073
5	0.83601482	0.16718562	0.0929	0.9002
6	0.66882920	0.51626025	0.0743	0.9745
7	0.15256894	0.09669020	0.0170	0.9915
8	0.05587874	0.03502162	0.0062	0.9977
9	0.02085712		0.0023	1.0000

The PRINCOMP Procedure

Eigenvectors									
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**	**Prin9**
TB	0.454464	0.224481	0.411895	-.230522	0.093106	0.045908	-.016909	-.059923	0.711069
DB	0.446128	0.220277	0.431265	-.232524	0.099979	0.060736	-.074309	-.033487	-.701472
Alkphos	0.246103	0.004487	0.180647	0.733213	-.418127	0.440242	0.021892	0.007196	0.001804
Alamine	0.289811	0.375345	-.523846	0.116613	0.086248	-.030628	-.685783	0.089768	0.019470
Aspartate	0.346997	0.361291	-.455647	0.085747	0.100346	-.092343	0.713333	-.070932	-.040288
TP	-.240471	0.436977	0.322852	0.324544	0.040598	-.515158	0.042788	0.522253	0.006911
ALB	-.391144	0.471466	0.138359	0.162170	0.119771	0.016445	-.069199	-.748210	-.007863
AGRatio	-.327948	0.345141	-.030142	-.192445	0.233332	0.723524	0.088619	0.386559	0.013131
Age	0.105348	-.311928	0.073522	0.406911	0.847679	0.042188	-.009912	-.012356	0.005477




Prin1
Prin2
7.5
5.0
2.5
0.0
-2.5
-2.5
0.0
2.5
5.0
7.5
Yes


Prin1
Prin3
7.5
5.0
2.5
0.0
-2.5
-7.5
-5.0
-2.5
0.0
2.5
5.0
Yes

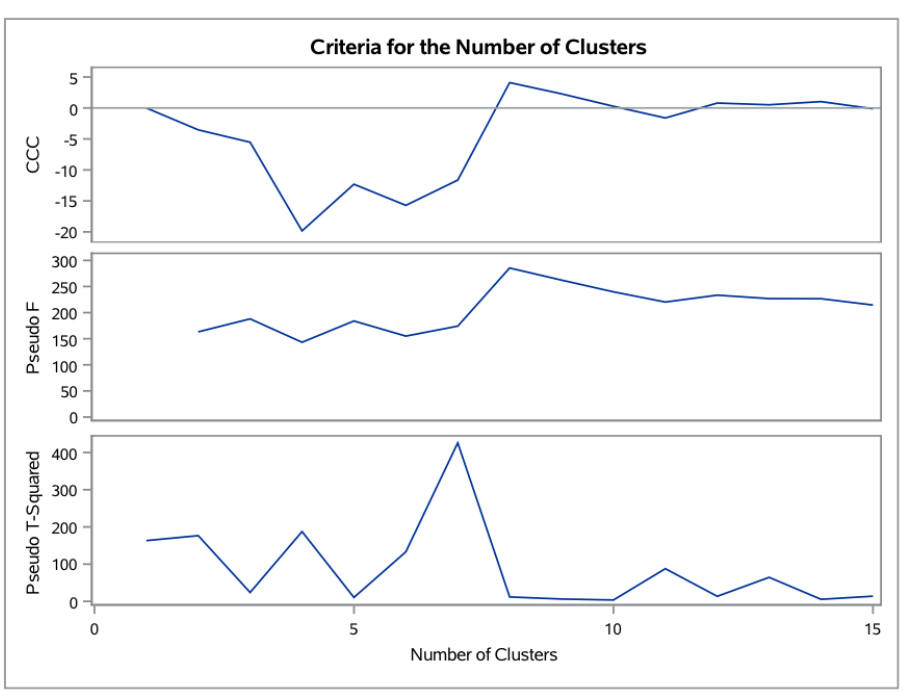
4)After doing cluster analysis on the principal components, the data still did not cluster well. I ended up choosing 8 clusters. Clusters 2,3, and 5-8 contained all or mostly all liver patients, while clusters 1 and 4 contained about 2 out of 3 liver patients in each. Since the clusters with mostly liver patients are high in principal component 1 (which compares TB and DB to ALB and AGRatio) and 2 (which compares TP and ALB to age), liver patients tend to have higher TP and DB.

The CLUSTER Procedure
Complete Linkage Cluster Analysis

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Norm Maximum Distance	Tie
15	CL18	CL42	35	0.0063	.843	.844	-.13	215	14.0	1.4199	
14	CL34	CL23	12	0.0028	.840	.836	1.03	227	5.5	1.5369	
13	CL28	CL22	189	0.0111	.829	.827	0.52	227	64.5	1.5924	
12	CL15	CL17	43	0.0087	.820	.817	0.79	233	13.6	1.7883	
11	CL19	CL16	269	0.0240	.796	.805	-1.6	220	87.8	1.8294	
10	CL26	CL27	5	0.0036	.793	.791	0.27	240	3.7	1.8586	
9	CL20	CL35	10	0.0050	.788	.775	2.28	262	6.3	1.9484	
8	CL21	CL14	16	0.0086	.779	.756	4.09	285	11.9	1.9982	
7	CL11	CL13	458	0.1311	.648	.732	-12	174	426	2.2231	
6	CL12	CL24	84	0.0711	.577	.702	-16	155	133	2.6491	
5	CL8	OB197	17	0.0131	.564	.663	-12	184	10.6	3.0539	
4	CL7	CL6	542	0.1341	.430	.607	-20	143	187	3.1503	
3	CL9	CL10	15	0.0328	.397	.448	-5.5	188	23.8	3.2532	
2	CL4	CL5	559	0.1751	.222	.253	-3.5	163	177	3.8765	
1	CL2	CL3	574	0.2220	.000	.000	0.00	.	163	5.9117	

The CLUSTER Procedure
Complete Linkage Cluster Analysis



The FREQ Procedure

Frequency	Table of CLUSTER by LiverPatient			
		LiverPatient		
	CLUSTER	No	Yes	Total
	1	81	188	269
	2	0	10	10
	3	8	33	41
	4	75	114	189
	5	1	42	43
	6	0	16	16
	7	0	5	5
	8	0	1	1
	Total	165	409	574

5) Since the researcher is interested in classification, I used discriminant analysis. In order to obtain the best model, I used stepwise discriminant analysis. The final model includes DB, Alkphos, age, and Aspartate. The researcher was correct, it is difficult to classify the four groups. A female who is not a liver patient had a classification error rate of 0.1837, while the three other groups had error rates above .7. There is a large gap between females, while a smaller gap between men. Male liver patients and non liver patients may have similar levels of DB, Alkphos, age, and Aspartate.

The STEPDISC Procedure

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	DB		0.0668	13.60	<.0001	0.93318583	<.0001	0.02227139	<.0001
2	2	Alkphos		0.0291	5.69	0.0008	0.90598909	<.0001	0.03180427	<.0001
3	3	Age		0.0248	4.81	0.0026	0.88355074	<.0001	0.03959133	<.0001
4	4	Aspartate		0.0267	5.18	0.0015	0.85997941	<.0001	0.04768164	<.0001

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
1061.083394	30	<.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.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.LIVER2
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into cell					
From cell	FemaleNo	FemaleYes	Male No	Male Yes	Total
FemaleNo	40 81.63	0 0.00	9 18.37	0 0.00	49 100.00
FemaleYes	50 54.95	13 14.29	13 14.29	15 16.48	91 100.00
Male No	80 68.97	10 8.62	25 21.55	1 0.86	116 100.00
Male Yes	113 35.53	34 10.69	83 26.10	88 27.67	318 100.00
Total	283 49.30	57 9.93	130 22.65	104 18.12	574 100.00
Priors	0.25	0.25	0.25	0.25	

Error Count Estimates for cell					
	FemaleNo	FemaleYes	Male No	Male Yes	Total
Rate	0.1837	0.8571	0.7845	0.7233	0.6371
Priors	0.2500	0.2500	0.2500	0.2500	

While it is difficult to tell, the results show that high TB and DB levels are an indicator of liver disease. As shown in the model in (2), when DB increases by one unit, the odds of having liver disease is 1.720 times higher.When TP increases by one unit, the odds of having liver disease is 1.543 times higher. However, as shown in (4) and (5), there is quite a bit of overlap between liver patients and non liver patients.