Faraz Younus

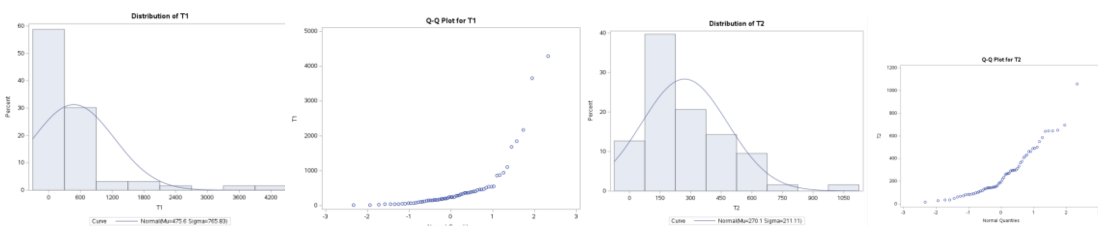
Prof. Yu

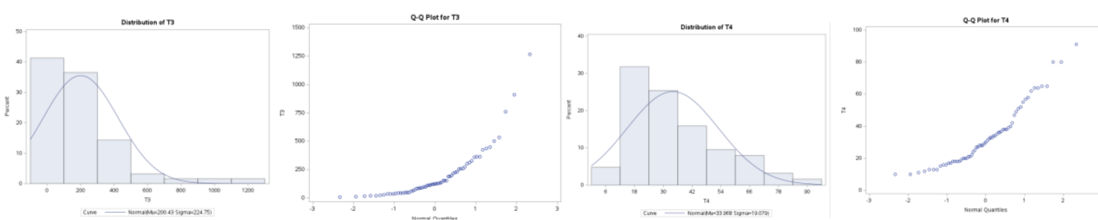
Sta 9750

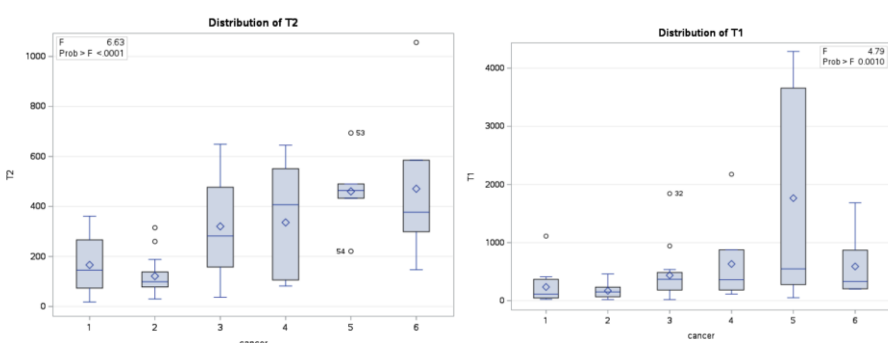
Question 1)

(1A) Let's first consider the factor of cancer only. Is there any difference between the six types of cancers in terms of their survival times? Conduct appropriate tests and list all assumptions for those tests. Which test is the most powerful one? Explain.

- The assumption of normality means that the data must be distributed in a bell-shaped curve. If the data is not normally distributed, the ANOVA test may be less powerful and may not be able to detect significant differences between the groups.
- The assumption of equal variances or homoscedasticity means that the variances of the data must be equal. If the variances are not equal, the ANOVA test may be less powerful and may not be able to detect significant differences between the groups.
- The assumption of independence means that the data must be independent of each other. If the data is not independent, the ANOVA test may be less powerful and may not be able to detect significant differences between the groups.







Just by looking at normal, QQ, and distributions, I can tell that many of the assumptions for ANOVA are violated making ANOVA not that great for our data.

The ANOVA Procedure

Dependent Variable: T1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	10763118.18	2152623.64	4.79	0.0010
Error	57	25599940.90	449121.77		
Corrected Total	62	36363059.08			

Anova on T1 The p-value is less than 0.05, so we can reject the null hypothesis that there is no difference between the six types of cancers in terms of their survival times. This means that there is a significant difference between the survival times of patients with different types of cancer.

Dependent Variable: T2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1016409.660	203281.932	6.63	<.0001
Error	57	1746827.768	30646.101		
Corrected Total	62	2763237.429			

Anova on T2 The p-value is less than 0.05, so we can reject the null hypothesis that there is no difference between the six types of cancers in terms of their survival times. This means that there is a significant difference between the survival times of patients with different types of cancer.

The ANOVA Procedure

Dependent Variable: T3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	452330.782	90466.156	1.92	0.1044
Error	57	2679340.646	47005.976		
Corrected Total	62	3131671.429			

Anova on T3 The p-value is not less than 0.05, so we can **fail to reject the null hypothesis** that there is no difference between the six types of cancers in terms of their survival times.

Dependent Variable: T4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1903.92520	380.78504	1.05	0.3974
Error	57	20664.01131	362.52651		
Corrected Total	62	22567.93651			

Anova on T4 The p-value is not less than 0.05, so we can **fail to reject the null hypothesis** that there is no difference between the six types of cancers in terms of their survival times.

(1B)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.SURVTIME
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into cancer							
From cancer	1	2	3	4	5	6	Total
1	0 0.00	10 83.33	2 16.67	0 0.00	0 0.00	0 0.00	12 100.00
2	0 0.00	15 93.75	1 6.25	0 0.00	0 0.00	0 0.00	16 100.00
3	0 0.00	6 37.50	8 50.00	0 0.00	0 0.00	2 12.50	16 100.00
4	0 0.00	3 42.86	2 28.57	1 14.29	0 0.00	1 14.29	7 100.00
5	0 0.00	1 20.00	1 20.00	0 0.00	2 40.00	1 20.00	5 100.00
6	0 0.00	1 14.29	2 28.57	1 14.29	1 14.29	2 28.57	7 100.00
Total	0 0.00	36 57.14	16 25.40	2 3.17	3 4.76	6 9.52	63 100.00
Priors	0.19048	0.25397	0.25397	0.11111	0.07937	0.11111	

Error Count Estimates for cancer							
	1	2	3	4	5	6	Total
Rate	1.0000	0.0625	0.5000	0.8571	0.6000	0.7143	0.5556
Priors	0.1905	0.2540	0.2540	0.1111	0.0794	0.1111	

I have a strong reason to believe that T1and T2 contributed most to the classification because ANOVA said there was difference in Means for those two variables. I believe Linear discriminant function works better if means are different.

(1C)

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.SURVTIME
Resubstitution Summary using Nearest Neighbor

Number of Observations and Percent Classified into cancer							
From cancer	1	2	3	4	5	6	Total
1	12 100.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	12 100.00
2	0 0.00	16 100.00	0 0.00	0 0.00	0 0.00	0 0.00	16 100.00
3	0 0.00	0 0.00	16 100.00	0 0.00	0 0.00	0 0.00	16 100.00
4	0 0.00	0 0.00	0 0.00	7 100.00	0 0.00	0 0.00	7 100.00
5	0 0.00	0 0.00	0 0.00	0 0.00	5 100.00	0 0.00	5 100.00
6	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	7 100.00	7 100.00
Total	12 19.05	16 25.40	16 25.40	7 11.11	5 7.94	7 11.11	63 100.00
Priors	0.19048	0.25397	0.25397	0.11111	0.07937	0.11111	

Error Count Estimates for cancer							
	1	2	3	4	5	6	Total
Rate	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Priors	0.1905	0.2540	0.2540	0.1111	0.0794	0.1111	

I used K means with k=1 and I got 0 percent error rate. It was surprising how accurate it was. Maybe this accuracy rate is deceiving.

(1D) Now let's consider the factor of gender as well as the factor of cancer. Use an appropriate method to answer the following questions: 1. Is there any difference in gender regarding the survival times? 2. Is there any interaction effect between gender and cancer type on the survival times? Show all necessary work.

Dependent Variable: T1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	12740917.73	1158265.25	2.50	0.0135
Error	51	23622141.35	463179.24		
Corrected Total	62	36363059.08			

R-Square	Coeff Var	Root MSE	T1 Mean
0.350381	143.0968	680.5727	475.6032

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gender	1	8892.35	8892.35	0.02	0.8903
cancer	5	11512917.73	2302583.55	4.97	0.0009
gender*cancer	5	1219107.64	243821.53	0.53	0.7552

Source	DF	Type III SS	Mean Square	F Value	Pr > F
gender	1	1206624.52	1206624.52	2.61	0.1127
cancer	5	10729871.07	2145974.21	4.63	0.0015
gender*cancer	5	1219107.64	243821.53	0.53	0.7552

Dependent Variable: T2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1221117.129	111010.648	3.67	0.0007
Error	51	1542120.300	30237.653		
Corrected Total	62	2763237.429			

R-Square	Coeff Var	Root MSE	T2 Mean
0.441915	64.38091	173.8898	270.0952

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gender	1	3016.929	3016.929	0.10	0.7534
cancer	5	1075634.890	215126.978	7.11	<.0001
gender*cancer	5	142465.310	28493.062	0.94	0.4618

Source	DF	Type III SS	Mean Square	F Value	Pr > F
gender	1	53022.5877	53022.5877	1.75	0.1913
cancer	5	748223.2518	149644.6504	4.95	0.0009
gender*cancer	5	142465.3100	28493.0620	0.94	0.4618

Dependent Variable: T3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	480950.391	43722.763	0.84	0.6008
Error	51	2650721.038	51974.922		
Corrected Total	62	3131671.429			

R-Square	Coeff Var	Root MSE	T3 Mean
0.153576	113.7463	227.9801	200.4286

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gender	1	1508.6593	1508.6593	0.03	0.8654
cancer	5	467100.8268	93420.1654	1.80	0.1300
gender*cancer	5	12340.9048	2468.1810	0.05	0.9986

Source	DF	Type III SS	Mean Square	F Value	Pr > F
gender	1	13282.3482	13282.3482	0.26	0.6154
cancer	5	427626.8060	85525.3612	1.65	0.1649
gender*cancer	5	12340.9048	2468.1810	0.05	0.9986

Dependent Variable: T4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	4060.15476	369.10498	1.02	0.4453
Error	51	18507.78175	362.89768		
Corrected Total	62	22567.93651			

R-Square	Coeff Var	Root MSE	T4 Mean
0.179908	56.08140	19.04987	33.96825

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gender	1	6.414072	6.414072	0.02	0.8948
cancer	5	1907.950250	381.590050	1.05	0.3981
gender*cancer	5	2145.790440	429.158088	1.18	0.3307

Source	DF	Type III SS	Mean Square	F Value	Pr > F
gender	1	63.166775	63.166775	0.17	0.6783
cancer	5	1911.095768	382.219154	1.05	0.3971
gender*cancer	5	2145.790440	429.158088	1.18	0.3307

Level of gender	N	T1		T2		T3		T4	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
1	39	484.923077	801.583281	264.666667	225.230915	196.589744	237.174344	33.7179487	19.2325775
2	24	460.458333	720.440927	278.916667	190.253816	206.666667	207.741408	34.3750000	19.2304390

1) From the output All of the P values for gender are over .05 so we fail to reject null hypothesis of means being the same for gender. The Levels of Gender chart above somewhat confirms this.

2) There is a little interaction effect between gender and cancer. But, it's not significant. Before Adding the interaction effect We rejected the null hypothesis for T1 and T2 only. After adding the interaction effect also We rejected the null hypothesis for T1 and T2 only. That's why I believe interaction isn't important.

Question 2)

2A) Since the unit of measurement is the same for those stocks, I believe it's best to use covariance matrix. Sometimes it depends on the scenario; I remember covariance matrix had an easier time converging to exact values. PCs are generally computed from covariance matrix rather than correlation. Cov remains closer to the spirit and intent of PCA, especially if further computations on PCs are used. In some cases, the PCs will be more interpretable if correlation matrix is used. For example, if the variances differ widely, the PCs of covariance will be dominated by the variables with large variances. The other variables will contribute little.

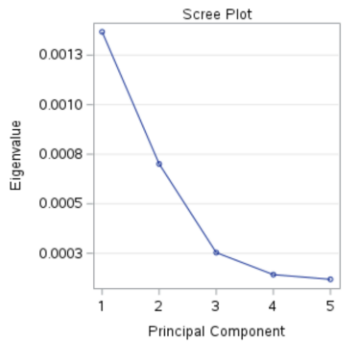
2B)

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	0.00136768	0.00066652	0.5293	0.5293
2	0.00070116	0.00044736	0.2713	0.8006
3	0.00025380	0.00011120	0.0982	0.8988
4	0.00014260	0.00002372	0.0552	0.9540
5	0.00011889		0.0460	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
JP_Morgan	0.222823	0.625226	0.326112	-.662759	0.117660
Citibank	0.307290	0.570390	-.249590	0.414094	-.588608
Wells_Fargo	0.154810	0.344505	-.037639	0.497050	0.780304
Royal_Dutch_Shell	0.638968	-.247948	-.642497	-.308869	0.148455
Exxon_Mobil	0.650904	-.321848	0.645861	0.216376	-.093718



2C)

Obs	Week	JP_Morgan	Citibank	Wells_Fargo	Royal_Dutch_Shell	Exxon_Mobil
1	56	0.018815	0.037969	0.015499	0.05104	0.078416

I consistantly found that Week 56 was the best week after suming up the returns for all stocks by week.

2D)

- There is Kaiser criterion which says keep PCs that are bigger than one.
- The is the elbow scree plot which says keep 2-3 eigenvalues from our data.
- If Cumulative proportion of variance explained is 80 percent then that amount of PCs should be kept
- Retain the PCs with eigenvalues above the average of all eigenvalues.

2E) The first PC we kept maximized the variance captured. The second PC is orthogonal to first PC. This method helps transform a set of possibly correlated stock into a new set of “hybrid” uncorrelated stocks. The principal components are linear combinations of the original stock variables. The banks will be combined into one PC while the oil companies will be combined into another orthogonal PC.

Question 3)

3A)

Dependent Variable: y1

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2520.49121	630.12280	241.32	<.0001
Error	45	117.50399	2.61120		
Corrected Total	49	2637.99520			

Root MSE	1.61592	R-Square	0.9555
Dependent Mean	98.83600	Adj R-Sq	0.9515
Coeff Var	1.63495		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	68.95355	1.35196	51.00	<.0001
x1	1	0.35938	0.07409	4.85	<.0001
x2	1	0.30034	0.09598	3.13	0.0031
x3	1	0.79575	0.13364	5.95	<.0001
x4	1	0.44315	0.03059	14.49	<.0001

Dependent Variable: y2

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4852.96681	1213.24170	321.87	<.0001
Error	45	169.61899	3.76931		
Corrected Total	49	5022.58580			

Root MSE	1.94147	R-Square	0.9662
Dependent Mean	106.62200	Adj R-Sq	0.9632
Coeff Var	1.82089		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	75.41979	1.62433	46.43	<.0001
x1	1	0.11954	0.08902	1.34	0.1860
x2	1	0.86739	0.11531	7.52	<.0001
x3	1	-0.57782	0.16056	-3.60	0.0008
x4	1	0.79513	0.03675	21.63	<.0001

Dependent Variable: y3

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1013.57078	253.39270	153.11	<.0001
Error	45	74.47422	1.65498		
Corrected Total	49	1088.04500			

Root MSE	1.28646	R-Square	0.9316
Dependent Mean	102.81000	Adj R-Sq	0.9255
Coeff Var	1.25130		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	83.70818	1.07632	77.77	<.0001
x1	1	0.56792	0.05899	9.63	<.0001
x2	1	-0.08071	0.07641	-1.06	0.2965
x3	1	0.65839	0.10639	6.19	<.0001
x4	1	0.23258	0.02435	9.55	<.0001

I tried to model multivariate using multiple regression on the 3 variables.

Y1 = x1 +x2+x3+x4

Y2 = x1 +x2+x3+x4

Y3 = x1 +x2+x3+x4

All of the Anova's had P value less then .05. The R^2 of the regression was good for prediction purpose. Standard errors were't bad. I think most of the tests are statistically useful for performance.

3B)

Variable: y1

Moments

N	50	Sum Weights	50
Mean	98.836	Sum Observations	4941.8
Std Deviation	7.33734535	Variance	53.8366367
Skewness	-0.586368	Kurtosis	-0.5670708
Uncorrected SS	491065.74	Corrected SS	2637.9952
Coeff Variation	7.42375789	Std Error Mean	1.03765733

Variable: y2

Moments

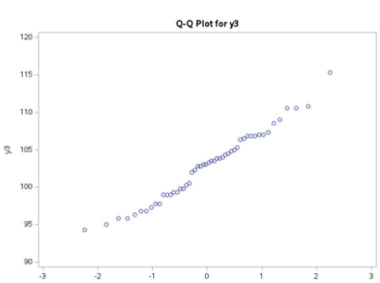
N	50	Sum Weights	50
Mean	106.622	Sum Observations	5331.1
Std Deviation	10.1243148	Variance	102.501751
Skewness	-0.0667732	Kurtosis	-1.1185544
Uncorrected SS	573435.13	Corrected SS	5022.5858
Coeff Variation	9.49552141	Std Error Mean	1.43179434

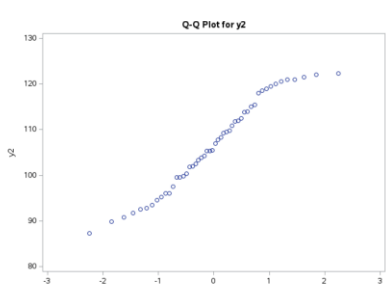
Variable: y3

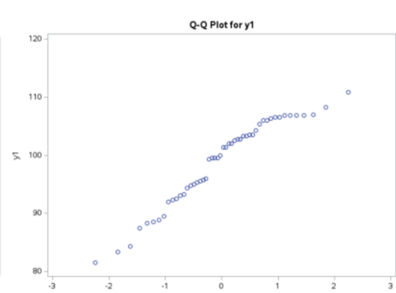
Moments

N	50	Sum Weights	50
Mean	102.81	Sum Observations	5140.5
Std Deviation	4.71221816	Variance	22.205
Skewness	0.22562755	Kurtosis	-0.3069492
Uncorrected SS	529582.85	Corrected SS	1088.045
Coeff Variation	4.58342395	Std Error Mean	0.66640828

The Kurtosis is small and near 0 which is a good thing for normality assumption!







The QQ plots also confirm normality!

3C)

Were do stepwise model selection to see which tests are best for which performance measure. There are ways to find reduced models that don't hurt overall accuracy.

For first performace measure 4th test is probably the best measure.

For second performace measure, 4th test is also the best measure.

For third performace measure, 4th and second test are the best measures

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	68.95355	1.35196	6792.43757	2601.27	<.0001
x1	0.35938	0.07409	61.43177	23.53	<.0001
x2	0.30034	0.09598	25.57001	9.79	0.0031
x3	0.79575	0.13364	92.58581	35.46	<.0001
x4	0.44315	0.03059	548.00956	209.87	<.0001

Bounds on condition number: 1.9804, 28.288

All variables left in the model are significant at the 0.1500 level.

All variables have been entered into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		1	0.8599	0.8599	95.5309	294.63	<.0001
2	x2		2	0.0459	0.9058	51.1952	22.88	<.0001
3	x3		3	0.0264	0.9322	26.5263	17.90	0.0001
4	x1		4	0.0233	0.9555	5.0000	23.53	<.0001

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	75.82927	1.60933	8514.65011	2220.17	<.0001
x2	0.94326	0.10140	331.90448	86.54	<.0001
x3	-0.61672	0.15930	57.48392	14.99	0.0003
x4	0.80409	0.03646	1865.74813	486.49	<.0001

Bounds on condition number: 1.8856, 14.624

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		1	0.8917	0.8917	98.3160	395.19	<.0001
2	x2		2	0.0617	0.9534	18.0538	62.31	<.0001
3	x3		3	0.0114	0.9649	4.8033	14.99	0.0003

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	83.34162	1.02013	11074	6674.45	<.0001
x1	0.53739	0.05149	180.75408	108.94	<.0001
x3	0.63980	0.10506	61.53599	37.09	<.0001
x4	0.22455	0.02316	155.91474	93.97	<.0001

Bounds on condition number: 1.7596, 13.42

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		1	0.7269	0.7269	133.564	127.74	<.0001
2	x1		2	0.1464	0.8733	39.2981	54.32	<.0001
3	x3		3	0.0566	0.9299	4.1158	37.09	<.0001

3D)

$$H_0 : \rho_2 = \rho_3 = \cdots = \rho_s = 0 \quad \text{vs} \quad H_a : \text{At least } \rho_2 \neq 0$$

					Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.994483	0.994021	0.001572	0.988996	89.8745	86.5063	0.9621	0.9621	0.00214847	87.39	12	114.06	<.0001
2	0.878107	0.872097	0.032704	0.771071	3.3682	3.1956	0.0361	0.9982	0.19524127	18.53	6	88	<.0001
3	0.383606	0.366795	0.121835	0.147153	0.1725		0.0018	1.0000	0.85284669	3.88	2	45	0.0278

Correlations Among the Original Variables

Correlations Among the VAR Variables

	x1	x2	x3	x4
x1	1.0000	0.5907	0.1469	0.4126
x2	0.5907	1.0000	0.3860	0.5746
x3	0.1469	0.3860	1.0000	0.5664
x4	0.4126	0.5746	0.5664	1.0000

Correlations Among the WITH Variables

	y1	y2	y3
y1	1.0000	0.9261	0.8840
y2	0.9261	1.0000	0.8425
y3	0.8840	0.8425	1.0000

Correlations Between the VAR Variables and the WITH Variables

	y1	y2	y3
x1	0.5720	0.5415	0.7004
x2	0.7081	0.7459	0.6375
x3	0.6744	0.4654	0.6411
x4	0.9273	0.9443	0.8526

Critical value for First, second and third canonical correlations.

simplify $\dfrac{1.96}{\sqrt{\left(1-0.994483^2\right)}}$

Solution

$18.68484\ldots$

simplify $\dfrac{1.96}{\sqrt{\left(1-0.878107^2\right)}}$

Solution

$4.09643\ldots$

simplify $\dfrac{1.96}{\sqrt{\left(1-0.383606^2\right)}}$

Solution

$2.12236\ldots$

Since the Aproximate F value > the above Critical values we conclude all canonical cor **of them are significant.**

3E)

Canonical Correlation Analysis

Standardized Canonical Coefficients for the VAR Variables

	V1	V2	V3
x1	0.2755	-0.7600	0.9739
x2	0.1040	0.6823	-0.4803
x3	0.1916	-1.0607	-0.5996
x4	0.6621	0.7199	0.1194

Standardized Canonical Coefficients for the WITH Variables

	W1	W2	W3
y1	0.4577	-1.2772	-2.7673
y2	0.2119	2.4517	1.0480
y3	0.3688	-1.1229	1.8067

Standardized Variance of the VAR Variables Explained by

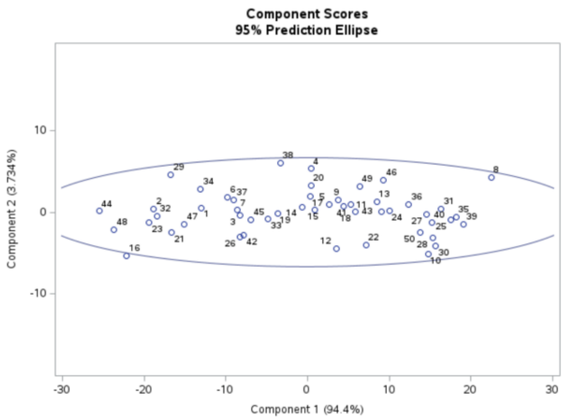
	Their Own Canonical Variables			The Opposite Canonical Variables	
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.5594	0.5594	0.9890	0.5533	0.5533
2	0.0983	0.6578	0.7711	0.0758	0.6291
3	0.1919	0.8497	0.1472	0.0282	0.6573

Standardized Variance of the WITH Variables Explained by

	Their Own Canonical Variables			The Opposite Canonical Variables	
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.9206	0.9206	0.9890	0.9105	0.9105
2	0.0463	0.9670	0.7711	0.0357	0.9462
3	0.0330	1.0000	0.1472	0.0049	0.9511

I still think X4 is the best variable because it's most correlated with V1 and V2 which have the biggest cononical R^2. X4 which is math ability also had the largest partial R^2 in stepwise regression 86%, 89%, and 73% respectivly for y1 y2 y3.

I calculated the 95^2 prediction ellipse and for two components below.

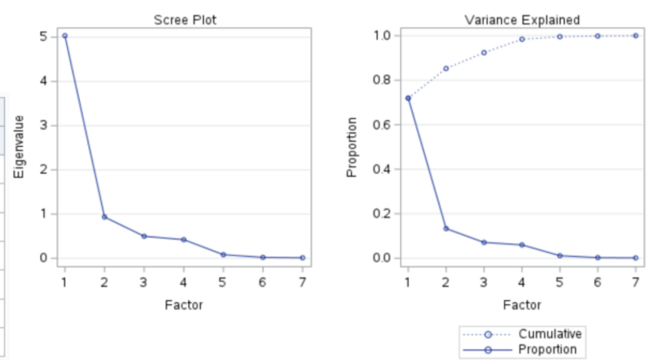


4)

4A)

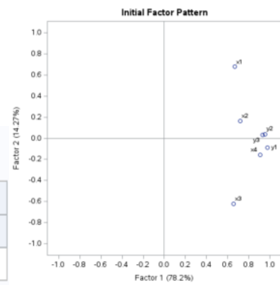
- Choose m factors necessary for the variance accounted for to achieve a predetermined percentage, say 80%, of the total variance
- Choose m = number of eigenvalues greater than the average eigenvalue. For R the average is 1.
- Use the scree plot based on the eigenvalues of R. If the graph drops sharply, followed by a straight line with much smaller slope, choose m equal to the number of eigenvalues before the straight line begins.

Eigenvalues of the Correlation Matrix: Total = 7 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	5.03459779	4.10108165	0.7192	0.7192
2	0.93351614	0.43559640	0.1334	0.8526
3	0.49791975	0.07667426	0.0711	0.9237
4	0.42124549	0.34020506	0.0602	0.9839
5	0.08104043	0.06069980	0.0116	0.9955
6	0.02034063	0.00900087	0.0029	0.9984
7	0.01133977		0.0016	1.0000



I believe One or two factors will be enough. This is because after 2 factors all of the eigen values are much less than 1. Furthermore Eivenvalues 1 and 2 account for 85 percent of the information. Also the scree plot flattens after 2.

4B)

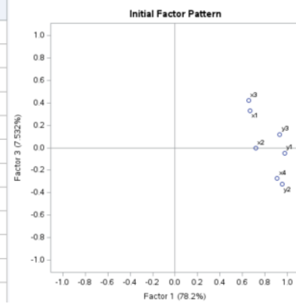


The FACTOR Procedure

Initial Factor Method: Iterated Principal Factor Analysis

Prior Communality Estimates: SMC						
y1	y2	y3	x1	x2	x3	x4
0.97154338	0.96837839	0.95390306	0.90511191	0.78074838	0.88487831	0.97679536

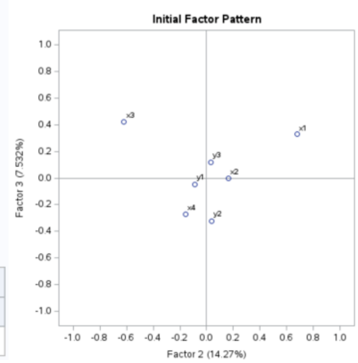
Iteration	Change	Communalities						
1	0.1602	0.96090	0.98036	0.91394	0.89492	0.62059	0.88437	0.96294
2	0.0506	0.96068	0.99329	0.90283	0.90223	0.57003	0.88810	0.95387
3	0.0147	0.96144	1.00000	0.89934	0.91336	0.55533	0.89191	0.94610
4	0.0117	0.96200	1.00000	0.89774	0.92510	0.55071	0.89531	0.93986
5	0.0115	0.96249	1.00000	0.89661	0.93658	0.54890	0.89844	0.93559
6	0.0110	0.96283	1.00000	0.89558	0.94761	0.54785	0.90134	0.93260
7	0.0105	0.96305	1.00000	0.89458	0.95815	0.54705	0.90404	0.93043
8	0.0101	0.96318	1.00000	0.89362	0.96822	0.54634	0.90655	0.92880
9	0.0096	0.96326	1.00000	0.89269	0.97784	0.54569	0.90891	0.92752
10	0.0092	0.96332	1.00000	0.89180	0.98705	0.54508	0.91113	0.92648
11	0.0088	0.96335	1.00000	0.89096	0.99587	0.54450	0.91323	0.92559
12	0.0041	0.96337	1.00000	0.89016	1.00000	0.54397	0.91522	0.92482



Factor Pattern			
	Factor1	Factor2	Factor3
y1	0.97487	-0.09011	-0.04976
y2	0.95123	0.03944	-0.32131
y3	0.93247	0.03339	0.11667
x1	0.66532	0.67639	0.32822
x2	0.71912	0.16172	-0.00252
x3	0.65446	-0.61855	0.41953
x4	0.91037	-0.15631	-0.26997

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
4.9415996	0.9014769	0.4759512

Final Communality Estimates: Total = 6.319028						
y1	y2	y3	x1	x2	x3	x4
0.9609688	1.0096249	0.8842340	1.0078822	0.5432966	0.9869328	0.9260885

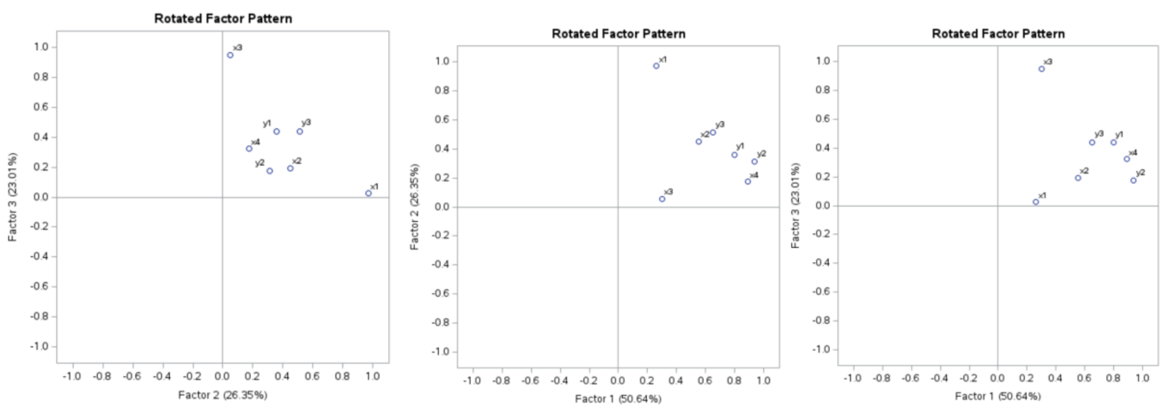


4C)

Rotated Factor Pattern			
	Factor1	Factor2	Factor3
y1	0.79892	0.36032	0.43917
y2	0.93878	0.31191	0.17615
y3	0.65303	0.51348	0.44059
x1	0.26334	0.96839	0.02734
x2	0.55041	0.45107	0.19206
x3	0.29884	0.05183	0.94601
x4	0.89026	0.17361	0.32155

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
3.2001820	1.6648642	1.4539816

Final Communality Estimates: Total = 6.319028						
y1	y2	y3	x1	x2	x3	x4
0.9609688	1.0096249	0.8842340	1.0078822	0.5432966	0.9869328	0.9260885



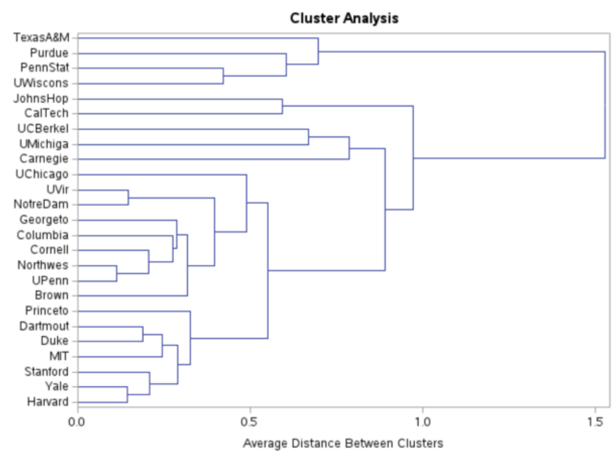
4D)

Rotated Factor Pattern			
	Factor1	Factor2	Factor3
y1	0.79892	0.36032	0.43917
y2	0.93878	0.31191	0.17615
y3	0.65303	0.51348	0.44059
x1	0.26334	0.96839	0.02734
x2	0.55041	0.45107	0.19206
x3	0.29884	0.05183	0.94601
x4	0.89026	0.17361	0.32155

The first factor is able to explain the y1 y2 y3 and x4 really well. The other factors factor 2 and 3 are better at explaining x1 and x3 at .60 threshold.

Question 5)

5A)



Cluster History

Number of Clusters	Clusters Joined		Freq	Norm RMS Distance	Tie
24	UPenn	Northwes	2	0.114	
23	Harvard	Yale	2	0.1439	
22	NotreDam	UVir	2	0.1454	
21	Duke	Dartmout	2	0.1882	
20	CL24	Cornell	3	0.2063	
19	CL23	Stanford	3	0.2072	
18	MIT	CL21	3	0.2458	
17	CL20	Columbia	4	0.2767	
16	CL17	Georgeto	5	0.2878	
15	CL19	CL18	6	0.2884	
14	Brown	CL16	6	0.3167	
13	CL15	Princeto	7	0.3258	
12	CL14	CL22	8	0.3958	
11	UWiscons	PennStat	2	0.4218	
10	CL12	UChicago	9	0.4886	
9	CL13	CL10	16	0.5497	
8	CalTech	JohnsHop	2	0.5935	
7	CL11	Purdue	3	0.6035	
6	UMichiga	UCBerkel	2	0.6685	
5	CL7	TexasA&M	4	0.6968	
4	Carnegie	CL6	3	0.7869	
3	CL9	CL4	19	0.891	
2	CL3	CL8	21	0.9712	
1	CL2	CL5	25	1.5275	

Root-Mean-Square Distance Between Observations	3.464102

5B)

– Cubic clustering criterion (CCC): a measure of the deviation of the clusters from the distribution expected if data points were obtained from a uniform distribution. A larger CCC, a better clustering.

– Pseudo-F statistic: a ratio of the mean sum of squares between groups to the mean sum of squares within group. A larger pseudo-F, a better clustering.

With two, three, 4 clusters respectively:

	2 clusters	3 clusters	4 clusters
Pseudo F Statistic =	30.51	24.12	23.18
Approximate Expected Over-All R-Squared =	0.20454	0.34659	0.46247
Cubic Clustering Criterion =	14.529	13.066	12.949

I choose two clusters because two clusters has the highest Pesudo F statistics & Cubic Custer Criterion,

Obs	col	CLUSTER	DISTANCE
1	Duke	1	0.54121
2	UPenn	1	0.59359
3	Columbia	1	0.68464
4	Dartmout	1	0.74145
5	Northwes	1	0.75373
6	Stanford	1	0.82377
7	MIT	1	0.83861
8	Cornell	1	0.88630
9	Brown	1	0.94298
10	Georgeto	1	1.21093
11	Yale	1	1.36407
12	UChicago	1	1.38184
13	Princeto	1	1.40509
14	Harvard	1	1.46222
15	NotreDam	1	1.53344
16	UVir	1	1.85867
17	JohnsHop	1	2.36741
18	UCBerkel	1	2.48512
19	CalTech	1	3.03489
20	UWiscons	2	0.75447
21	PennStat	2	1.16158
22	UMichiga	2	1.66774
23	Purdue	2	1.96790
24	TexasA&M	2	2.17147
25	Carnegie	2	2.67807

SAS code:

Q1

```
data survtime;
  infile "/home/u63223421/SURVTIME.txt" delimiter=' ';
  input cancer gender t T1 T2 T3 T4;
run;


PROC UNIVARIATE data=survtime;
  VAR T1 T2 T3 T4;
  HISTOGRAM T1 T2 T3 T4 / NORMAL;
  QQPLOT T1 T2 T3 T4 / NORMAL;
RUN;
proc anova data=survtime;
  class cancer;
  model T1 T2 T3 T4 = cancer;
  means cancer / tukey;
run;
PROC DISCRIM LIST pool=Yes ;
   CLASS cancer ;
   VAR T1 T2 T3 T4;
   PRIORS proportional;
RUN;
PROC DISCRIM LIST pool=Yes Method=NPAR k=1 ;
   CLASS cancer ;
   VAR T1 T2 T3 T4;
   PRIORS proportional;
RUN;
PROC GLM DATA=survtime;
  CLASS gender cancer;
  MODEL T1-T4 = gender|cancer;
  LSMEANS gender / DIFF;
  LSMEANS gender*cancer / DIFF;
RUN;
```

```
proc glm data=survtime;
class gender cancer;
model T1 T2 T3 T4 = gender cancer gender*cancer;
run;
proc glm data=survtime;
   class Gender Cancer;
   model T1 T2 T3 T4 = Gender|Cancer;
   means Gender Cancer / hovtest=levene;
run;
```

Q2)

```
DATA stock_data;
   INFILE '/home/u63223421/stock_price.txt' DLM='09'x ;
   INPUT JP_Morgan Citibank Wells_Fargo Royal_Dutch_Shell Exxon_Mobil;
RUN;
PROC PRINCOMP Cov DATA=stock_data OUTSTAT=pc_stats;
   VAR JP_Morgan Citibank Wells_Fargo Royal_Dutch_Shell Exxon_Mobil;
RUN;
/* Week with highest gain = 56*/
data stock_data;
   infile '/home/u63223421/stock_price.txt' dlm='09'x ;
   input JP_Morgan Citibank Wells_Fargo Royal_Dutch_Shell Exxon_Mobil;
   Week = _N_;
   Total_Return = sum(of JP_Morgan Citibank Wells_Fargo Royal_Dutch_Shell Exxon_Mobil);
run;
proc sort data=stock_data;
   by descending Total_Return;
run;

proc print data=stock_data(obs=1);
   var Week JP_Morgan Citibank Wells_Fargo Royal_Dutch_Shell Exxon_Mobil;
run;
/* Question D  */
proc princomp data=stock_data n=5 outstat=pc_stats;
   var JP_Morgan Citibank Wells_Fargo Royal_Dutch_Shell Exxon_Mobil;
run;

proc print data=pc_stats;
   var Variable Eigenvalue Proportion Cumulative;
run;
```

Q3)

```
DATA salesman;
   INFILE '/home/u63223421/salesman.txt' DLM=' ' ;
   INPUT y1 y2 y3 x1 x2 x3 x4;
RUN;
proc reg data=salesman;
model y1 y2 y3 = x1 x2 x3 x4;
run;
PROC UNIVARIATE data=salesman;
  VAR y1 y2 y3;
  HISTOGRAM y1 y2 y3 / NORMAL;
  QQPLOT y1 y2 y3 / NORMAL;
RUN;
/*  this code says that kurtiosis and skewneess is near zero so normality of residual
 is present*/
PROC REG data=salesman;
  MODEL y1 y2 y3 = x1 x2 x3 x4 / selection=stepwise;
RUN;
/*  */
PROC CANCORR data=salesman ALL MStat= exact;
  WITH y1 y2 y3;
  VAR x1 x2 x3 x4;
RUN;
PROC PRINCOMP COV OUT=results plots(ncomp = 2)=score(ellipse);
    VAR y1 y2 y3;
RUN;
```

Q4)

```
DATA salesman;
   INFILE '/home/u63223421/salesman.txt' DLM=' ' ;
   INPUT y1 y2 y3 x1 x2 x3 x4;
RUN;
/*  A*/
PROC FACTOR data=salesman method=principal scree rotate=none;
VAR y1 y2 y3 x1 x2 x3 x4;
RUN;
/*  B*/
PROC FACTOR METHOD=PRIN PLOTS=scree;
RUN;
/*Iterated Principal Factor Method*/
PROC FACTOR METHOD=PRINIT NFACT=3 PRIORS=SMC HEYWOOD MAXITER=100 plots= all;
RUN;
/* D */
PROC FACTOR METHOD=PRINIT NFACT=3 PRIORS=SMC HEYWOOD MAXITER=100  rotate=varimax plots= all;
```

Q5)

```
data college1;
  infile '/home/u63223421/university.txt' delimiter='' ;
  input col $ SAT top acrate sfr annualexpenses gradrate;
run;

proc standard data=college1 out=college mean=0 std=1;
  var  SAT top acrate sfr annualexpenses gradrate;
run;
```

```
proc tree data=ProTree nclusters=4 out=newdata noprint;
  id col;
  copy SAT top acrate sfr annualexpenses gradrate;
run;
proc sort data=newdata;
  by CLUSTER;
run;
proc print data=newdata;
  var col CLUSTER;
run;
```

```
/*  Method #1 for getting seeds:  Random 5 observations   */
proc fastclus data=college maxc=2 replace=random maxiter=10 radius=1.25 out=Clus_out;
  var SAT top acrate sfr annualexpenses gradrate;
  id col;
run;

proc sort data=Clus_out;
  by CLUSTER distance;
run;

proc print data=Clus_out;
  var col CLUSTER distance;
```