STA 9705: Final Exam

Tanay Mukherjee

1 - a)

In the “iris” dataset we have p=4(variables), k=3(groups), n=50(observations).

To test whether there is any significant difference among three groups in terms of those four variables, we could conduct a MONVOA test.

Appropriate test: MANOVA Test

Hypothesis:

H_0: $\mu_1 = \mu_2 = \mu_3$ vs.

H_a: At least one inequality

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for group E = Error SSCP Matrix					
		Characteristic Vector V'EV=1			
Characteristic Root	Percent	X1	X2	X3	X4
32.1919292	99.12	-0.06840592	-0.12656121	0.18155288	0.23180286
0.2853910	0.88	0.00198791	0.17852670	-0.07686357	0.23417227
0.0000000	0.00	0.10268742	-0.19415509	-0.22502879	0.37627520
0.0000000	0.00	-0.24194505	0.10603485	0.10535376	0.00000000

MANOVA Tests for the Hypothesis of No Overall group Effect H = Type III SSCP Matrix for group E = Error SSCP Matrix S=2 M=0.5 N=71		
Statistic	Value	P-Value
Wilks' Lambda	0.02343863	<.0001
Pillai's Trace	1.19189883	<.0001
Hotelling-Lawley Trace	32.47732024	<.0001
Roy's Greatest Root	32.19192920	<.0001

Wilk’s Λ: $\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|} = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} = .0234$

P = 4 (The 4 variables) $V_H = 3-1 = 2$ (df for hypothesis) $V_E = 3(50-1) = 147$ (df for error)

We reject H_0 since $\Lambda = .0234 \le \Lambda_{.05(4,2,147)} = 0.894$ for Wilk’s Λ. We can conclude that there is a difference amongst the specie with respect to the four flower measurements.

Also, with α=0.05, the exact p-values, from the above table are all <.0001. Therefore, all four tests listed in the above table are significant and the null hypothesis of (H_0: $\boldsymbol{\mu}1=\boldsymbol{\mu}2=\boldsymbol{\mu}3$) are rejected.

1 - b)

Discriminant Function for Iris

The CANDISC Procedure

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.984821	0.984508	0.002468	0.969872	32.1919	31.9065	0.9912	0.9912	0.02343863	199.15	8	288	<.0001
2	0.471197	0.461445	0.063734	0.222027	0.2854		0.0088	1.0000	0.77797337	13.79	3	145	<.0001

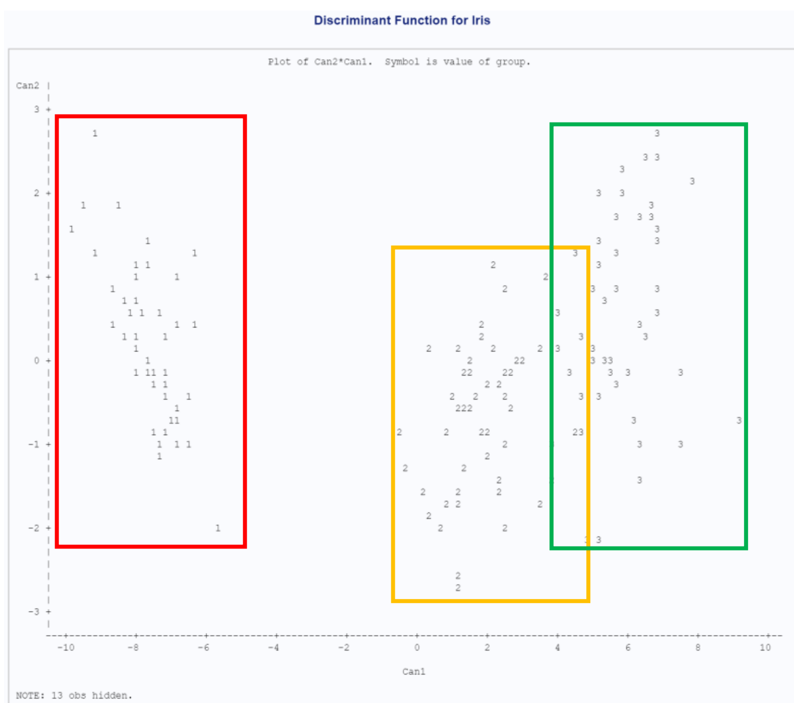
Raw Canonical Coefficients

Variable	Label	Can1	Can2
X1	Sepal Length	-0.829377642	0.024102149
X2	Sepal Width	-1.534473068	2.164521235
X3	Petal Length	2.201211656	-0.931921210
X4	Petal Width	2.810460309	2.839187853

The discriminant functions are:

$$z_1 = -0.829x_1 - 1.534x_2 + 2.201x_3 + 2.810x_4$$

$$z_2 = 0.024x_1 + 2.165x_2 - 0.932x_3 + 2.839x_4$$



The first discriminant function can1 clearly separates the first group of Iris from the second and third group of Iris. Group 2 and 3 of Iris have some overlap but overall are separate. The second discriminant function is ineffective in separating all groups.

1 - c)

Discriminant Function for Iris

The CANDISC Procedure

					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	0.984821	0.984508	0.002468	0.969872	32.1919	31.9065	0.9912	0.9912	0.02343863	199.15	8	288	<.0001
2	0.471197	0.461445	0.063734	0.222027	0.2854		0.0088	1.0000	0.77797337	13.79	3	145	<.0001

Hypothesis:

H_0: $a_1 = 0$ vs.

H_a: $a_1 \neq 0$

Test Statistic:

$$\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$$

$$\Lambda_1 = \left(\frac{1}{1+32.1919}\right)\left(\frac{1}{1+0.2854}\right) = 0.0234$$

p = 4 k = 3 n = 50

$0.0234 < \Lambda_{.05(4,2,147)} = 0.894$

We reject the H_0 and therefore the first discriminant function is significant and move on to test Λ_2.

Hypothesis:

H_0: $\mathbf{a_2} = 0$ vs H_a: $\mathbf{a_2} \neq 0$

Test Statistic:

$$\Lambda_2 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$$

$$\Lambda_2 = \left(\frac{1}{1+0.2854}\right) = 0.778$$

p = 4 k = 3 n = 50

$0.778 < \Lambda_{.05(3,1,146)} = 0.945$

We reject the H_0 and can thus conclude that both discriminant functions are significant.

1 - d)

Hypothesis:

H_0: $\frac{1}{2}$ (μ_2+ μ_3) = μ_1

H_a: $\frac{1}{2}$ (μ_2+ μ_3) $\neq$ μ_1

MANOVA Tests for the Hypothesis of No Overall 1 v/s 2&3 Effect H = Contrast SSCP Matrix for 1 v/s 2&3 E = Error SSCP Matrix S=1 M=1 N=71		
Statistic	Value	P-Value
Wilks' Lambda	0.03273111	<.0001
Pillai's Trace	0.96726889	<.0001
Hotelling-Lawley Trace	29.55196881	<.0001
Roy's Greatest Root	29.55196881	<.0001

Since all MANOVA tests are equivalent to each other for contrasts testing, we will look at just the Wilks' test. The exact p-values in the SAS output indicates rejection of the null hypothesis of the contrast test as there is significant difference between group 1 and group 2 & 3. That is, Iris setosa is different from Iris versicolor and Iris virginica.

Hypothesis:

H_0: $\boldsymbol{\mu_2}$= $\boldsymbol{\mu_3}$

H_a: $\boldsymbol{\mu_2}$$\neq$ $\boldsymbol{\mu_3}$

MANOVA Tests for the Hypothesis of No Overall 2 v/s 3 Effect H = Contrast SSCP Matrix for 2 v/s 3 E = Error SSCP Matrix S=1 M=1 N=71		
Statistic	Value	P-Value
Wilks' Lambda	0.25475426	<.0001
Pillai's Trace	0.74524574	<.0001
Hotelling-Lawley Trace	2.92535143	<.0001
Roy's Greatest Root	2.92535143	<.0001

Since all MANOVA tests are equivalent to each other for contrasts testing, we will look at just the Wilks' test. The exact p-values in the SAS output indicates rejection of the null hypothesis of the contrast test as there is significant difference between group 2 and group 3. That is, Iris versicolor is different from Iris virginica.

Thus, we can say that at an overall level group 1, group 2 and group 3 are all different from each other. That is, Iris setosa, Iris versicolor and Iris virginica are all different from each other.

1 - e)

Pooled Within-Class Standardized Canonical Coefficients			
Variable	**Label**	**Can1**	**Can2**
X1	Sepal Length	-.4269548486	0.0124075316
X2	Sepal Width	-.5212416758	0.7352613085
X3	Petal Length	0.9472572487	-.4010378190
X4	Petal Width	0.5751607719	0.5810398645

The above standardized discriminant function coefficient table shows that X3 contributes the most to separation between the three species. The variable X3 represents the feature 'Petal Length'.

1 - f)

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

Number of Observations and Percent Classified into group				
From group	**1**	**2**	**3**	**Total**
1	50 100.00	0 0.00	0 0.00	50 100.00
2	0 0.00	48 96.00	2 4.00	50 100.00
3	0 0.00	1 2.00	49 98.00	50 100.00
Total	50 33.33	49 32.67	51 34.00	150 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for group				
	1	**2**	**3**	**Total**
Rate	0.0000	0.0400	0.0200	0.0200
Priors	0.3333	0.3333	0.3333	

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

Number of Observations and Percent Classified into group				
From group	**1**	**2**	**3**	**Total**
1	50 100.00	0 0.00	0 0.00	50 100.00
2	0 0.00	47 94.00	3 6.00	50 100.00
3	0 0.00	1 2.00	49 98.00	50 100.00
Total	50 33.33	48 32.00	52 34.67	150 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for group				
	1	**2**	**3**	**Total**
Rate	0.0000	0.0600	0.0200	0.0267
Priors	0.3333	0.3333	0.3333	

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.IRIS
Cross-validation Summary using 5 Nearest Neighbors

Number of Observations and Percent Classified into group				
From group	**1**	**2**	**3**	**Total**
1	50 100.00	0 0.00	0 0.00	50 100.00
2	0 0.00	47 94.00	3 6.00	50 100.00
3	0 0.00	1 2.00	49 98.00	50 100.00
Total	50 33.33	48 32.00	52 34.67	150 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for group				
	1	**2**	**3**	**Total**
Rate	0.0000	0.0600	0.0200	0.0267
Priors	0.3333	0.3333	0.3333	

Misclassification Rate		
Classification Method	Resubstitution	Cross-validation
Linear	0.0200	0.0200
Quadratic	0.0200	0.0267
KNN	0.0267	0.0267

Error rate for Linear Method = (2+1)/150 = 0.0200

Error rate for Quadratic Method = (3+1)/150 = 0.0267

Error rate for KNN (for n = 5) Method = (3+1)/150 = 0.0267

Since resubstition method tends to overfit the data and underestimate the actual error rate, we will just look at the cross-validation method. From the table above, it shows that the linear classification model gives the lowest misclassification rate (0.0200). Therefore, the best approach is linear classification for the iris data.

1 - g)

In the question we have $\boldsymbol{ynew} = \boldsymbol{x}_0' = (\mathbf{5.1}, \mathbf{3.5}, \mathbf{1.75}, \mathbf{0.3})$

The DISCRIM Procedure

Generalized Squared Distance to group			
From group	1	2	3
1	0	89.86419	179.38471
2	89.86419	0	17.20107
3	179.38471	17.20107	0

Linear Discriminant Function for group			
Variable	1	2	3
Constant	-85.20986	-71.75400	-103.26971
X1	23.54417	15.69821	12.44585
X2	23.58787	7.07251	3.68528
X3	-16.43064	5.21145	12.76654
X4	-17.39841	6.43423	21.07911

$$
\begin{aligned}
L1(\boldsymbol{ynew}) &= -85.21 + 23.54X1 + 23.588X2 - 16.43X3 - 17.398X4 \\
&= -85.21+23.54(5.1) + 23.588(3.5) - 16.43(1.75) - 17.398(0.3) \\
&= 83.43
\end{aligned}
$$

$$
\begin{aligned}
L2(\boldsymbol{ynew}) &= -71.75 + 15.698X1 + 7.07X2 + 5.21X3 + 6.43X4 \\
&= -71.75+15.698(5.1) +7.07(3.5) + 5.21(1.75) + 6.43(0.3) \\
&= 44.1013
\end{aligned}
$$

$$
\begin{aligned}
L3(\boldsymbol{ynew}) &= -103.27 + 12.45X1 + 3.685X2+1\ 2.767X3 + 21.079X4 \\
&= -103.27+ 12.45(5.1) + 3.685(3.5) +12.767(1.75) + 21.079(0.3) \\
&= 1.78845
\end{aligned}
$$

The largest $Li(\boldsymbol{ynew})$ is 83.43 with group 1. Therefore, the new observation should be assigned to group 1 (Iris Setosa).

2 - a)

Following table has the covariance matrix:

Simple Statistics					
	JPM	Citi	WF	RDS	EM
Mean	0.0010627806	0.0006554204	0.0016260816	0.0040491252	0.0040386417
StD	0.0208151256	0.0209455772	0.0149657006	0.0268792933	0.0276708181

Covariance Matrix					
	JPM	Citi	WF	RDS	EM
JPM	0.0004332695	0.0002756679	0.0001590265	0.0000641193	0.0000889662
Citi	0.0002756679	0.0004387172	0.0001799737	0.0001814512	0.0001232623
WF	0.0001590265	0.0001799737	0.0002239722	0.0000734135	0.0000605461
RDS	0.0000641193	0.0001814512	0.0000734135	0.0007224964	0.0005082772
EM	0.0000889662	0.0001232623	0.0000605461	0.0005082772	0.0007656742

Total Variance	0.0025841294

We should use the sample covariance matrix opposed to correlation matrix when applying principal component analysis. The diagonal elements of S are not dominated by a single variable and have similar variances. In this case, we want to use S since it is closer to the intent of PCA, where we are trying to form a linear combination with maximal variance.

Following table has the correlation matrix:

Simple Statistics					
	JPM	Citi	WF	RDS	EM
Mean	0.0010627806	0.0006554204	0.0016260816	0.0040491252	0.0040386417
StD	0.0208151256	0.0209455772	0.0149657006	0.0268792933	0.0276708181

Correlation Matrix					
	JPM	Citi	WF	RDS	EM
JPM	1.0000	0.6323	0.5105	0.1146	0.1545
Citi	0.6323	1.0000	0.5741	0.3223	0.2127
WF	0.5105	0.5741	1.0000	0.1825	0.1462
RDS	0.1146	0.3223	0.1825	1.0000	0.6834
EM	0.1545	0.2127	0.1462	0.6834	1.0000

The scale of each variables does not differ much as observed from above two tables, so we can just use covariance matrix S in this case for PCA over correlation matrix R.

2 - b)

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
JPM	0.222823	0.625226	0.326112	-.662759	0.117660
Citi	0.307290	0.570390	-.249590	0.414094	-.588608
WF	0.154810	0.344505	-.037639	0.497050	0.780304
RDS	0.638968	-.247948	-.642497	-.308869	0.148455
EM	0.650904	-.321848	0.645861	0.216376	-.093718

Diag(**S**)' = (0.2228, 0.570, -0.0376, -0.3089, -0.0937)

The percent of variance explained using **S** is 52.93%, 27.13%, 9.82%, 5.52%, and 4.60%.

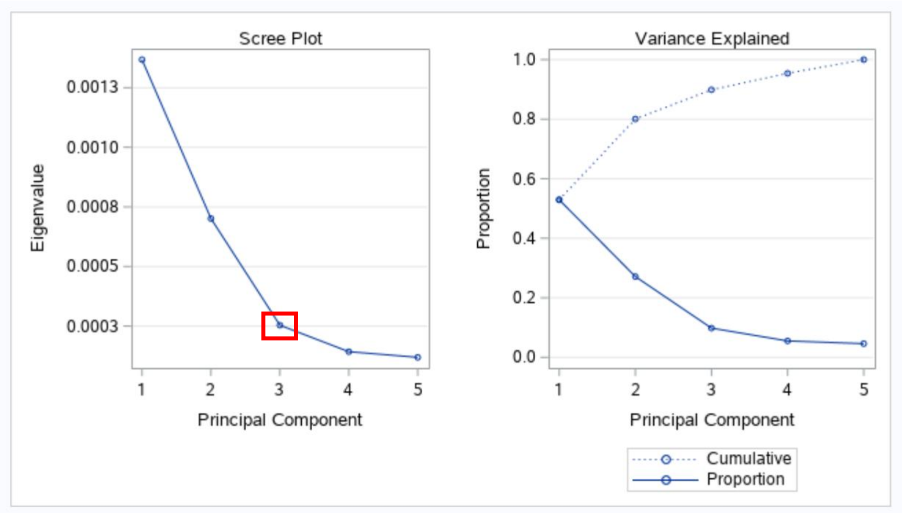
2 - c)

Principal Component Analysis of Stock Returns

Obs	Prin1	Prin2	Prin3	Prin4	Prin5
1	-0.03112	0.012704	0.038332	0.001487	-0.004704
2	0.01657	0.006085	-0.000262	-0.002573	-0.014390
3	-0.01500	-0.009989	-0.008160	0.011957	0.010153
4	-0.02227	0.027530	0.020995	0.000262	0.013370
5	0.04140	-0.014325	0.010564	-0.010819	-0.009560

2 - d)

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



From the scree plot and the table above in for eigenvalues, we should only keep 2 PCs. Because the first 2 PCs account for 80.06% variances and the scree plot has natural break (also, called as elbow point in the scree plot) between the 2nd PC and 3rd PC.

If we account for too much, we run the risk of including components that are sample specific or variable specific.

2 - e)

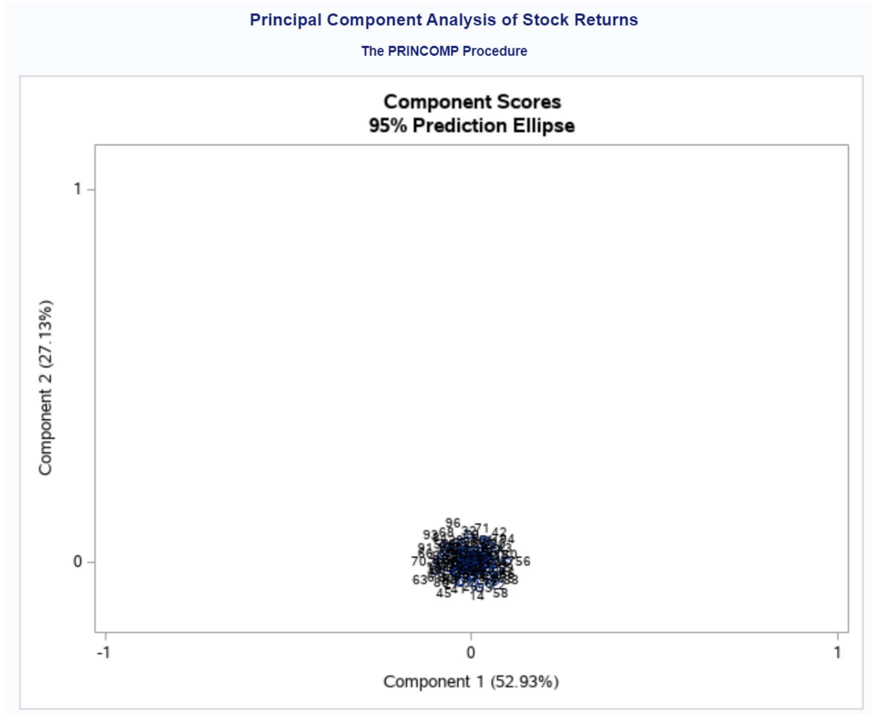
Eigenvectors					
	Prin1	Prin2	Prin3	Prin4	Prin5
JPM	0.222823	0.625226	0.326112	-.662759	0.117660
Citi	0.307290	0.570390	-.249590	0.414094	-.588608
WF	0.154810	0.344505	-.037639	0.497050	0.780304
RDS	0.638968	-.247948	-.642497	-.308869	0.148455
EM	0.650904	-.321848	0.645861	0.216376	-.093718

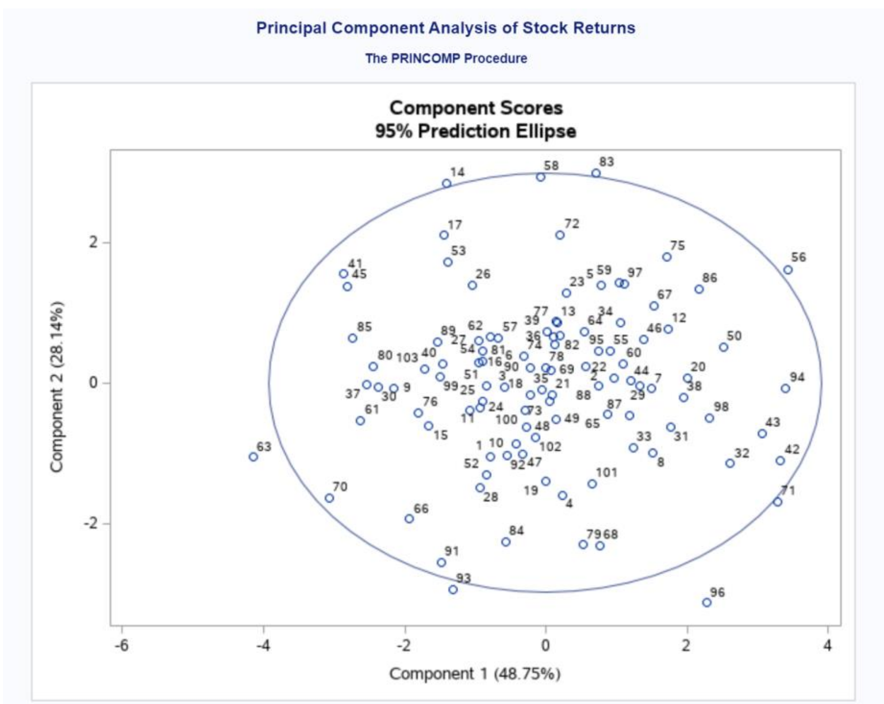
Here, JPM represents JP Morgan, Citi is Citibank, WF is Wells Fargo, RDS is Royal Dutch Shell and EM is Exxon Mobil.

From the coefficients of 1st PC, Royal Dutch Shell and Exxon Mobil contribute the most to the 1st PC. Similarly, for the 2nd PC, we can find that JP Morgan and Citibank contribute the most to the 2nd PC. And the 2nd PC is to compare (JPM, Citi, WF) with (RDS, EM). Since Royal Dutch Shell and Exxon Mobil have relatively larger variances than the other variables, we can see in the 1st PC, these two variables dominate the 1st PC and account for most of the variances.

Another way to interpret this is that the first component Prin1 increases as all the stocks increases, with the greatest contribution from the oil stocks. While the second component Prin2 increases if bank stocks increases.

2 - f)





We check for the about 0 for both PCs and find it to be symmetric and thus it helps us understand the distribution. From the figures above we can say that there are no outliers as most of the variables are clustered together.

3 - a)

Communalities (also known as h2) are the estimates of the variance of the factors, as opposed to the variance of the variable which includes measurement error. Initially the communality estimates are set equal to the R2 between each variable and all others.

The table below shows the initial estimates of communalities:

The FACTOR Procedure
Initial Factor Method: Principal Factors

Prior Communality Estimates: SMC						
X1	X2	X3	X4	X5	X6	X7
0.97154338	0.96837839	0.95390306	0.90511191	0.78074838	0.88487831	0.97679536

Factor loadings: Commonality is the square of the standardized outer loading of an item. Analogous to Pearson's r-squared, the squared factor loading is the percent of variance in that indicator variable explained by the factor. To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor (column) and divide by the number of variables. This is the same as dividing the factor's eigen value by the number of variables.

Interpreting factor loadings: By one rule of thumb in confirmatory factor analysis, loadings should be .7 or higher to confirm that independent variables identified a priori are represented by a particular factor, on the rationale that the .7 level corresponds to about half of the variance in the indicator being explained by the factor. However, the .7 standard is a high one and real-life data may well not meet this criterion, which is why some researchers, particularly for exploratory purposes, will use a lower level such as .6 for the central factor and .25 for other factors. In any event, factor loadings must be interpreted in the light of theory, not by arbitrary cutoff levels.

The table below shows the factor loadings:

Factor Pattern				
		Factor1	Factor2	Factor3
X1	Growth of Sales	0.97487	-0.09011	-0.04976
X2	Profitability of Sales	0.95123	0.03944	-0.32131
X3	New Account Sales	0.93247	0.03339	0.11667
X4	Creativity	0.66532	0.67639	0.32822
X5	Mechanical Reasoning	0.71912	0.16172	-0.00252
X6	Abstract Reasoning	0.65446	-0.61855	0.41953
X7	Mathematical Ability	0.91037	-0.15631	-0.26997

The table below shows the final estimates of communalities:

Final Community Estimates: Total = 6.319028						
X1	X2	X3	X4	X5	X6	X7
0.9609688	1.0096249	0.8842340	1.0078822	0.5432966	0.9869328	0.9260885

The table below shows the proportion of each factor:

Eigenvalues of the Reduced Correlation Matrix: Total = 6.30059331 Average = 0.90008476				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.94159959	4.04012267	0.7843	0.7843
2	0.90147692	0.42552568	0.1431	0.9274
3	0.47595124	0.36700370	0.0755	1.0029
4	0.10894754	0.09443744	0.0173	1.0202
5	0.01451009	0.06674799	0.0023	1.0225
6	-.05223789	0.03741629	-0.0083	1.0142
7	-.08965419		-0.0142	1.0000

Therefore, factor 1 accounts for 78.43% of variance; factor 2 accounts for 14.31% variances; factor 3 account for 7.55% of variance.

The table below shows the specific variances for each factor:

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
4.9415996	0.9014769	0.4759512

The specific variances can be calculated from the final estimates of communalities by equation $\widehat{\psi_i} = 1 - \widehat{h_i}^2$.

The calculations are shown below:

Features	Loadings			Loadings^2			Communality	Specific Variance
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3		
Growth of Sales	0.97487	-0.09011	-0.04976	0.950372	0.00812	0.002476	0.9609688	0.0390312
Profitability of Sales	0.95123	0.03944	-0.32131	0.904839	0.001556	0.10324	1.0096249	-0.0096249
New Account Sales	0.93247	0.03339	0.1166	0.8695	0.001115	0.013596	0.884234	0.115766
Creativity	0.66532	0.67639	0.32822	0.442651	0.457503	0.107728	1.0078822	-0.0078822
Mechanical Reasoning	0.71912	0.16172	-0.00252	0.517134	0.026153	6.35E-06	0.5432966	0.4567034
Abstract Reasoning	0.65446	-0.61855	0.41953	0.428318	0.382604	0.176005	0.9869328	0.0130672
Mathematical Ability	0.91037	-0.15631	-0.26997	0.828774	0.024433	0.072884	0.9260885	0.0739115

3 - b)

There are the unusual estimates on final communalities of value over 1, and this leads to a negative specific variance. When this happens in the iteration, the Heywood method force communality to be 1 to carry out the rest iterations.

Also, to be noted is the eigenvalue for which the cumulative exceeds 1 for variable X4 to X6 before coming back down to 1. The reason why the cumulative exceeds 1 is because of one or more factor that has a non-positive variance.

3 - c)

Rotation serves to make the output more understandable, by seeking so-called "Simple Structure": A pattern of loadings where each item loads strongly on only one of the factors, and much weaker on the other factors. Rotations can be orthogonal or oblique (allowing the factors to correlate).

Varimax rotation is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by extracted factor. Each factor will tend to have either large or small loadings of any particular variable. A varimax solution yields results which make it as easy as possible to identify each variable with a single factor. This is the most common rotation option. However, the orthogonality (i.e., independence) of factors is often an unrealistic assumption. Oblique rotations are inclusive of orthogonal rotation, and for that reason, oblique rotations are a preferred method.

The table below shows estimates of communalities, resulting loading and variances:

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2	3
1	0.78068	0.46289	0.41984
2	-0.07769	0.73850	-0.66976
3	-0.62008	0.49026	0.61250

Rotated Factor Pattern

		Factor1	Factor2	Factor3
X1	Growth of Sales	0.79892	0.36032	0.43917
X2	Profitability of Sales	0.93878	0.31191	0.17615
X3	New Account Sales	0.65303	0.51348	0.44059
X4	Creativity	0.26334	0.96839	0.02734
X5	Mechanical Reasoning	0.55041	0.45107	0.19206
X6	Abstract Reasoning	0.29884	0.05183	0.94601
X7	Mathematical Ability	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

X1	X2	X3	X4	X5	X6	X7
0.9609688	1.0096249	0.8842340	1.0078822	0.5432966	0.9869328	0.9260885

We can see that the final estimates of communalities remain unchanged just as it was before the rotation.

Also, the proportion of variances from the SAS output above, we can say that Factor 1 has a share of 50.64%, Factor 2 has a share of 26.35% and Factor 3 has a share of 23.01%

The specific variances can be calculated from the final estimates of communalities by equation $\widehat{\psi_i} = 1 - \widehat{h_i}^2$.

Features	Loadings			Loadings^2			Communality	Specific Variance
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3		
Growth of Sales	0.79892	0.36032	0.43917	0.638273	0.129831	0.19287	0.9609688	0.0390312
Profitability of Sales	0.93878	0.31191	0.17615	0.881308	0.097288	0.031029	1.0096249	-0.0096249
New Account Sales	0.65303	0.51348	0.44059	0.426448	0.263662	0.19412	0.884234	0.115766
Creativity	0.26334	0.96839	0.02734	0.069348	0.937779	0.000747	1.0078822	-0.0078822
Mechanical Reasoning	0.55041	0.45107	0.19206	0.302951	0.203464	0.036887	0.5432966	0.4567034
Abstract Reasoning	0.29884	0.05183	0.94601	0.089305	0.002686	0.894935	0.9869328	0.0130672
Mathematical Ability	0.89026	0.17361	0.32155	0.792563	0.03014	0.103394	0.9260885	0.0739115

After rotation, the loadings and variance explained by each factor have changed, but communalities and specific variances were unchanged. The orthogonal rotation is to rotate the axes so that more points lies close to the new axes in order to make the factors more interpretable. So, the model does not really change. And orthogonal rotation preserves communalities because the distance to the origin are unchanged. And therefore, specific variances are unchanged too.

Also, the proportion of variance, total variance and factor loadings have changed. This is because the variance explained has spread more across all factors.

3 - d)

Rotated Factor Pattern				
		Factor1	Factor2	Factor3
X1	Growth of Sales	0.79892	0.36032	0.43917
X2	Profitability of Sales	$\boxed{0.93878}$	0.31191	0.17615

In this case, the complexity is 1. “Probability of sales” depends most on factor 1, because the loadings are largest (0.93878) among the other two loadings.

3 - e)

With the threshold as 0.6 for factor loadings, the complexity is shown below. The green highlighted columns display the complexity and associated rotated factors.

Feature Index	Features	Loadings			Complexity	Associated Factor
		Factor 1	Factor 2	Factor 3		
X1	Growth of Sales	0.79892	0.36032	0.43917	1	Factor 1
X2	Profitability of Sales	0.93878	0.31191	0.17615	1	Factor 1
X3	New Account Sales	0.65303	0. 51348	0.44059	1	Factor 1
X4	Creativity	0.26334	0.96839	0.02734	1	Factor 2
X5	Mechanical Reasoning	0.55041	0.45107	0.19206	0	----
X6	Abstract Reasoning	0.29884	0.05183	0.94601	1	Factor 3
X7	Mathematical Ability	0.89026	0.17361	0.32155	1	Factor 1

3 - f)

From part 3 - e), we can find that X5 which represents - Mechanical Reasoning, has no associated factor, this means that the model fit is not sufficient for a factor analysis. To improve it, I have changed the number of factors to be 4.

The table below shows estimates of communalities, resulting loading and variances for nfactor =4:

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2	3	4
1	0.70368	0.44014	0.41703	0.37040
2	-0.17952	0.70356	-0.64682	0.23325
3	-0.61537	0.45727	0.63570	-0.09005
4	-0.30649	-0.31965	0.05997	0.89459

Rotated Factor Pattern

		Factor1	Factor2	Factor3	Factor4
X1	Growth of Sales	0.71981	0.33755	0.45295	0.35192
X2	Profitability of Sales	0.82586	0.28257	0.18772	0.43142
X3	New Account Sales	0.62329	0.56983	0.46376	0.16141
X4	Creativity	0.21885	0.89151	0.02264	0.25735
X5	Mechanical Reasoning	0.36033	0.36884	0.19295	0.69268
X6	Abstract Reasoning	0.28377	0.03924	0.88844	0.13731
X7	Mathematical Ability	0.90786	0.18609	0.31514	0.17578

Variance Explained by Each Factor			
Factor1	Factor2	Factor3	Factor4
2.6711226	1.4854945	1.3818451	0.9318195

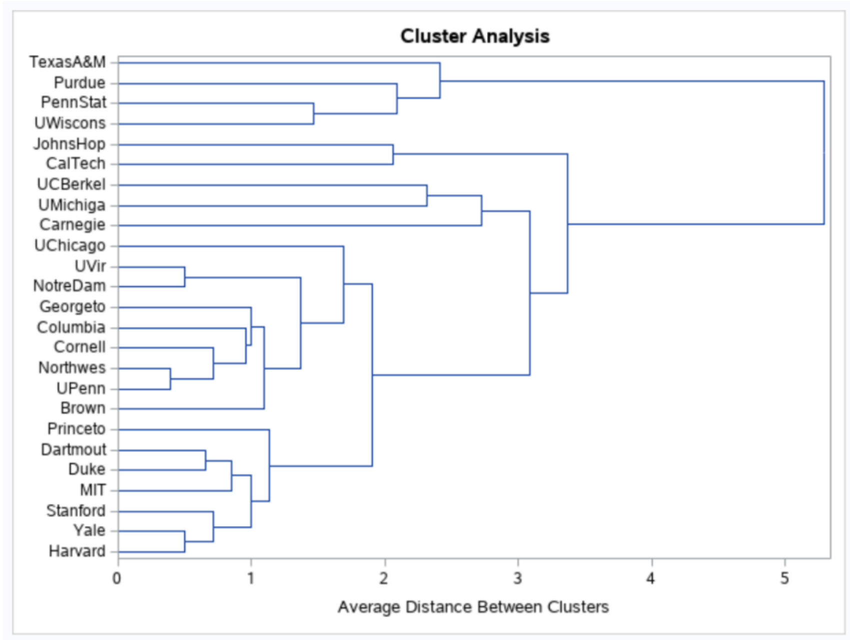
Final Communality Estimates: Total = 6.470282						
X1	X2	X3	X4	X5	X6	X7
0.96107810	0.98324848	0.95432557	0.90942843	0.78291535	0.89023674	0.98904904

With the threshold as 0.6 for factor loadings, the complexity is shown below. The green highlighted columns display the complexity and associated rotated factors. We can see that the X5 which represents - Mechanical Reasoning now has a complexity 1 and is associated to factor 4.

Feature Index	Features	Loadings				Complexity	Associated Factor
		Factor 1	Factor 2	Factor 3	Factor 4		
X1	Growth of Sales	0.71981	0.33755	0.45295	0.3519	1	Factor 1
X2	Profitability of Sales	0.82586	0.28257	0.18772	0.43142	1	Factor 1
X3	New Account Sales	0.62329	0.56983	0.46376	0.16141	1	Factor 1
X4	Creativity	0.21885	0.89151	0.02264	0.25735	1	Factor 2
X5	Mechanical Reasoning	0.36033	0.36884	0.19295	0.69268	1	Factor 4
X6	Abstract Reasoning	0.28377	0.03924	0.88844	0.13731	1	Factor 3
X7	Mathematical Ability	0.90786	0.18609	0.31514	0.17578	1	Factor 1

4 - a) – (i)

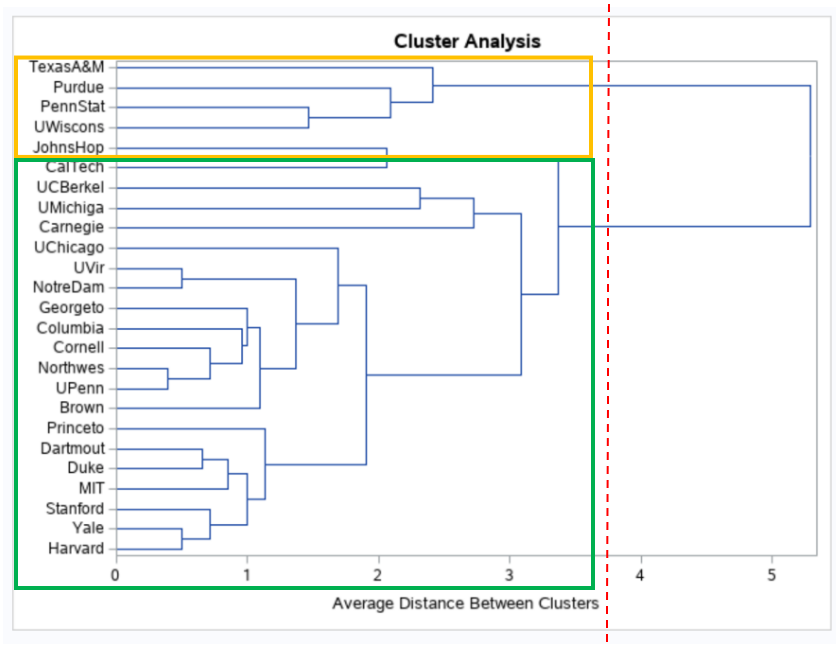
The dendrogram and the resultant cluster history as obtained from SAS:



Cluster History					
Number of Clusters	Clusters Joined		Freq	RMS Distance	Tie
24	UPenn	Northwes	2	0.3949	
23	Harvard	Yale	2	0.4985	
22	NotreDam	UVir	2	0.5036	
21	Duke	Dartmout	2	0.6521	
20	CL24	Cornell	3	0.7147	
19	CL23	Stanford	3	0.7177	
18	MIT	CL21	3	0.8513	
17	CL20	Columbia	4	0.9586	
16	CL17	Georgeto	5	0.9968	
15	CL19	CL18	6	0.999	
14	Brown	CL16	6	1.0972	
13	CL15	Princeto	7	1.1285	
12	CL14	CL22	8	1.3712	
11	UWiscons	PennStat	2	1.4611	
10	CL12	UChicago	9	1.6924	
9	CL13	CL10	16	1.9044	
8	CalTech	JohnsHop	2	2.0559	
7	CL11	Purdue	3	2.0906	
6	UMichiga	UCBerkel	2	2.3157	
5	CL7	TexasA&M	4	2.4139	
4	Carnegie	CL6	3	2.7258	
3	CL9	CL4	19	3.0866	
2	CL3	CL8	21	3.3643	
1	CL2	CL5	25	5.2914	

4 - a) – (ii)

Step 1: Using dendrograms:



By looking at the largest distance we see the difference is mainly between the cluster 1 and cluster 2 as observed from the RMS distance in the cluster history table above. The difference of distance comes out to be 1.93 (5.29 – 3.36). Therefore, we choose 2 clusters and thus g = 2.

Step 2: Using the sophisticated method

To use the method given by Mojena(in 1977), first we have to get $\bar{\alpha}$, which is the average of these distances between each stages, and $s\alpha$ which is the standard deviation of these distances.

Then, we decide "g" when $\boldsymbol{\alpha j} > \boldsymbol{\bar{\alpha}+ks\alpha}$, in other words, we will stop clustering at that stages. $\bar{\alpha}$ is calculated by taking the average of these Norm RMS Distances and $s\alpha$ is the standard distance of these Norm RMS Distances. Using the cluster history table from part 4 – a) (i) above we have:

$\bar{\alpha}$ = 1.636, $s\alpha$ = 1.151, k=1.25 $(given)$, => $\boldsymbol{\bar{\alpha}+ks\alpha}$ **= 3.076**

Check with the above table (highlighted in red), we choose to stop at Cluster 3 (Norm RMS Distance 3.0866 > 3.076), which give us 3 clusters and thus g = 3.

4 - b)

The table below shows the initial seeds:

The FASTCLUS Procedure
Replace=NONE Radius=1.5 Maxclusters=3 Maxiter=10 Converge=0.02

Initial Seeds						
Cluster	x1	x2	x3	x4	x5	x6
1	1.232560743	0.747147830	-1.277417094	-0.422879796	0.841393297	1.134936246
2	1.370988499	1.210255988	-0.719814394	-1.652181530	2.508651168	-0.631501491
3	0.401994205	0.644234905	-0.871887858	0.068840897	-0.324716668	0.803729170

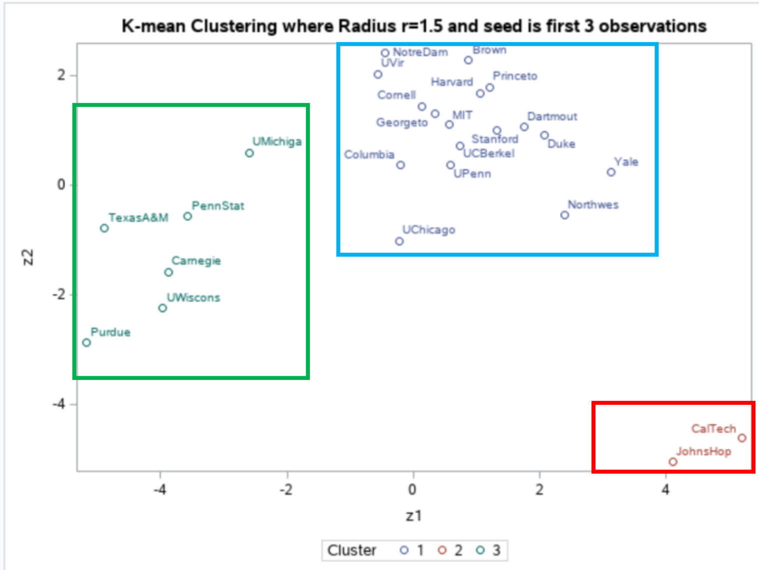
The table below shows the final seeds:

Cluster Means						
Cluster	x1	x2	x3	x4	x5	x6
1	0.393851395	0.468677564	-0.484249617	-0.191481823	0.067408029	0.530970402
2	0.863420059	0.567050212	-0.238248426	-1.529251357	2.339360369	-0.300294415
3	-1.403718973	-1.516936502	1.451456725	1.052282284	-0.970776205	-1.404318001

The table below shows K-means Clustering where Radius r=1.5 and seed is first 3 observations:

Obs	school	CLUSTER	DISTANCE
1	Duke	1	0.50179
2	UPenn	1	0.56667
3	Cornell	1	0.64648
4	Brown	1	0.69071
5	Columbia	1	0.72467
6	Northwes	1	0.75043
7	Dartmout	1	0.81831
8	Stanford	1	0.92317
9	Georgeto	1	1.00782
10	MIT	1	1.02490
11	NotreDam	1	1.29680
12	Princeto	1	1.47603
13	UChicago	1	1.48078
14	Yale	1	1.51533
15	Harvard	1	1.55803
16	UVir	1	1.63375
17	UCBerkel	1	2.35262
18	CalTech	2	1.02797
19	JohnsHop	2	1.02797
20	UWiscons	3	0.75447
21	PennStat	3	1.16158
22	UMichiga	3	1.66774
23	Purdue	3	1.96790
24	TexasA&M	3	2.17147
25	Carnegie	3	2.67807

Plot below shows reasonable clustering. Since each cluster has no points mixed with other clusters



If we compare the scatterplot from the two discriminant functions for the clustering, it is similar to the tree in (a) with the difference being that Cal Tech and Johns Hopkin are far from UCBerkely and U Michigan.

The table below shows the cluster summary:

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	17	0.5116	2.3526	> Radius	2	2.8166
2	2	0.5935	1.0280		1	2.8166
3	6	0.8258	2.6781	> Radius	1	4.1582

We also observe that RMS standard deviation of cluster 3 is 0.826, which is higher than the other two

4 - c)

With g=3 centroids, we have initial seeds listed below:

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=3 Maxiter=50 Converge=0.02

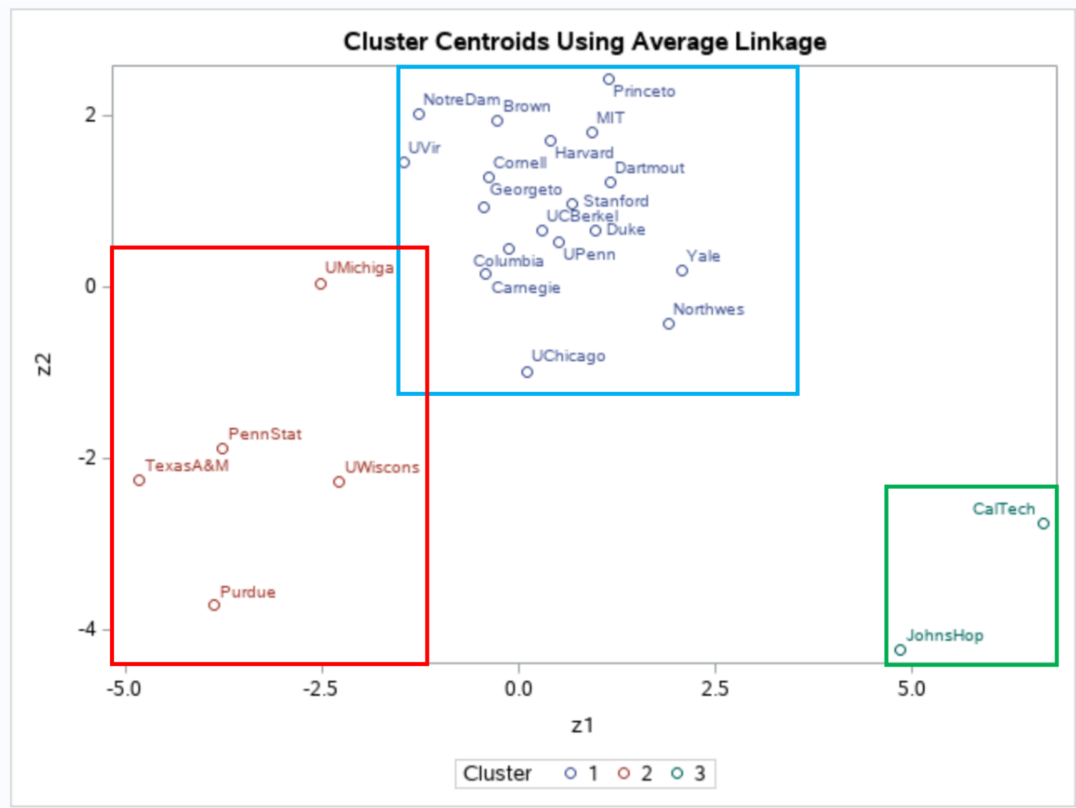
Initial Seeds						
Cluster	x1	x2	x3	x4	x5	x6
1	0.307280476	0.349037307	-0.303613336	-0.177019450	0.008209498	0.379551687
2	-1.891292293	-1.941452315	1.561287560	1.605468064	-1.208675301	-1.652723308
3	0.863420059	0.567050212	-0.238248426	-1.529251357	2.339360369	-0.300294415

With g=3 centroids, we have the final seeds listed below:

Cluster Means						
Cluster	x1	x2	x3	x4	x5	x6
1	0.368669004	0.401246057	-0.401586591	-0.231655082	0.054566204	0.411187451
2	-1.672576438	-1.671305889	1.541011098	1.445658839	-1.132182480	-1.360157058
3	0.863420059	0.567050212	-0.238248426	-1.529251357	2.339360369	-0.300294415

The table below shows k-means clustering for g=3 centroids:

Obs	school	CLUSTER	DISTANCE
1	UPenn	1	0.43075
2	Cornell	1	0.59614
3	Northwes	1	0.63242
4	Duke	1	0.63955
5	Columbia	1	0.66926
6	Brown	1	0.81828
7	Dartmout	1	0.92288
8	Georgeto	1	0.99224
9	Stanford	1	1.03439
10	MIT	1	1.06121
11	NotreDam	1	1.26965
12	UChicago	1	1.37364
13	Princeto	1	1.55672
14	UVir	1	1.58262
15	Yale	1	1.62363
16	Harvard	1	1.67705
17	UCBerkel	1	2.28647
18	Carnegie	1	2.85180
19	UWiscons	2	0.98523
20	PennStat	2	1.06292
21	Purdue	2	1.63680
22	TexasA&M	2	1.80621
23	UMichiga	2	1.95228
24	CalTech	3	1.02797
25	JohnsHop	3	1.02797



The table below shows the cluster summary:

Cluster Summary						
Cluster	**Frequency**	**RMS Std Deviation**	**Maximum Distance from Seed to Observation**	**Radius Exceeded**	**Nearest Cluster**	**Distance Between Cluster Centroids**
1	18	0.5751	2.8518		3	2.7765
2	5	0.7028	1.9523		1	4.4267
3	2	0.5935	1.0280		1	2.7765

We also observe that RMS standard deviation of cluster 2 is 0.703, which is higher than the other two. The clusters in (b) are tighter whereas the cluster in (c) are further apart especially for Cluster 1 and 2. Though, there isn’t necessarily any overlap but we can observe that UMichingan is in close vicinity of the boundaries of Cluster 1 while it actually belongs to Cluster 2. Also, Carnegie moved to cluster 1 represents most of the Ivy League colleges. It is a close call, but it seems cluster based on K-means in part 4. b) is better than the clustering based on centroids as we see in this part 4.c)

<u>APPENDIX:</u>

This section will have the entire SAS code for all the 4 questions.

 # Que - 1

```
TITLE 'Question 1';
TITLE 'Iris dataset';



TITLE 'Q1(a)';
DATA iris;
  INFILE "/folders/myfolders/data/iris.txt" DLM=' ';
  INPUT X1 X2 X3 X4 group;
  LABEL X1='Sepal Length' X2='Sepal Width' X3='Petal Length' X4='Petal Width';

PROC GLM;
  CLASS group;
  MODEL X1 X2 X3 X4= group;
  MANOVA H=group/PRINTE PRINTH MSTAT=EXACT;
RUN;



TITLE 'Q1(b)';
PROC FORMAT;
  VALUE group 1 = ' Iris setosa' 2 = ' Iris versicolor' 3 = 'Iris virginica';
RUN;



TITLE 'Discriminant Function for Iris';
PROC CANDISC OUT=CAND MSTAT=EXACT;
```

```
  CLASS group;
RUN;
PROC PRINT DATA=CAND;
RUN;
PROC PLOT DATA=CAND;
  PLOT CAN2*CAN1=group;
RUN;



TITLE 'MANOVA Test for Iris';
TITLE'Q1(c)';
PROC GLM;
  CLASS group;
  MODEL X1 X2 X3 X4 = group;
  CONTRAST '1 v/s 2&3'
    group -1 .5 .5;
  CONTRAST '2 v/s 3'
    group 0 1 -1;
  MANOVA H=group/PRINTE PRINTH MSTAT=EXACT;
RUN;



TITLE'Q1(f)';
DATA iris;
  INFILE "/folders/myfolders/data/iris.txt" DLM=' ';
  INPUT X1 X2 X3 X4 group;

PROC DISCRIM LIST crossvalidate;
   CLASS group;
RUN;
```

```
proc discrim data=iris pool=no list crossvalidate;
class group;
var X1 X2 X3 X4;
RUN;


TITLE 'Discriminant Analysis of Iris Data';
proc discrim data=iris method=npar k=5 crossvalidate;
class group;
var X1 X2 X3 X4;
RUN;
```

 # Que - 2

```
TITLE 'Question 2';
TITLE 'Stock price dataset';

DATA stock;
  INFILE "/folders/myfolders/data/stock_price.txt" DLM='09'x;
  INPUT JPM Citi WF RDS EM;
  RUN;


TITLE 'Principal Component Analysis of Stock Returns';
/* PCA USING S*/
PROC PRINCOMP COV OUT=RESULTS plots(ncomp =2)=score(ellipse);
  VAR JPM Citi WF RDS EM;
RUN;


proc print data=RESULTS;
var PRIN1 PRIN2 PRIN3 PRIN4 PRIN5 ;
RUN;


/* PCA USING R */
PROC PRINCOMP plots(ncomp =2)=score(ellipse);
  VAR JPM Citi WF RDS EM;
RUN;
```

 # Que - 3

```
TITLE 'Question 3';
TITLE 'Salesman dataset';

DATA salesman;
  INFILE "/folders/myfolders/data/salesman.txt" DLM=' ';
  INPUT X1 X2 X3 X4 X5 X6 X7;
  LABEL X1='Growth of Sales' X2='Profitability of Sales' X3='New Account Sales' X4='Creativity'
       X5='Mechanical Reasoning' X6='Abstract Reasoning' X7='Mathematical Ability' ;
RUN;


/* PRINCPICAL COMPONENT METHOD */
TITLE 'Factor Analysis of Salesman';
PROC FACTOR METHOD=PRIN NFACT=3 ROTATE=VARIMAX PLOTS=ALL;
  VAR X1-X7;
RUN;

/* Iterated Principal Factor METHOD */
PROC FACTOR METHOD=PRINIT NFACT=3 PRIORS=SMC HEYWOOD MAXITER=100
ROTATE=VARIMAX CORR PLOTS=ALL;
RUN;


/* PRINCIPAL FACTOR METHOD */
PROC FACTOR METHOD=PRIN PRIORS=SMC NFACT=3 ROTATE=VARIMAX PLOTS=ALL;
  VAR X1-X7;
RUN;
```

```
/* PRINCIPAL FACTOR METHOD with 4 Factors */
TITLE 'Q3(f)';
PROC FACTOR METHOD=PRIN PRIORS=SMC NFACT=4 ROTATE=VARIMAX PLOTS=ALL;
  VAR X1-X7;
RUN;
```

Que - 4

```
/*Question 4 university*/

DATA university;
  INFILE "/folders/myfolders/DATA/university.txt" DLM=' ';
  INPUT school $ x1 x2 x3 x4 x5 x6;
  LABEL X1='Average SAT' X2='Top 10%' X3='% Accepted' X4='Student Faculty Ratio' X5='Estimated Annual Expense' X6='Graduation Rate %';
run;

/*standardization of data */
proc standard data=university out=university mean=0 std=1;
var x1 x2 x3 x4 x5 x6;
run;
proc print data=university;
title '---------------------- Standardized Data ----------------------';
run;

/* Part a average linkage method for hierarchical clustering  */
proc cluster data=university outtree=tree_school method=average nonorm;
title '-------------- Part A: Hierarchical Clustering Wtih Average Linkage ----------------------';
   var x1 x2 x3 x4 x5 x6;
   id school;
run;

/* Before K-mean clustering  */
/*  Use principal components to guess the number of initial clusters to use   */
proc princomp data=university out=ProPC;
title '-------------- PCA to see how many cluster to use for k-mean clustering ----------------------';
```

```
var  x1 x2 x3 x4 x5 x6;
run;
proc sgplot         data=ProPC;
scatter y = prin2 x = prin1 / datalabel=school;
label prin2 = 'Z2' prin1='Z1';
run;quit;



/* Part b K-mean cluster:  First 3 observations for getting seeds */
proc fastclus data=university radius=1.5 maxc=3 replace=none maxiter=10 out=Clus_out ;
title '--------- Part B: K-mean Clustering, Seeds=first 3 observations with Radius r=1.5 ----------';
var x1 x2 x3 x4 x5 x6;
id school;
run;
proc sort data=Clus_out;
by cluster distance;
run;

proc means data=newdata;
by cluster;
output out=Seeds mean=x1 x2 x3 x4 x5 x6;
var x1 x2 x3 x4 x5 x6;
run;

proc candisc data=Clus_out noprint out=ProCan(keep=school cluster Can1 Can2);
class cluster;
var  x1 x2 x3 x4 x5 x6;
run;
```

```
proc sgplot data=ProCan;
scatter y=Can2 x=Can1 / group=cluster datalabel=school;
label Can1="z1" Can2="z2";
run;quit;

proc print data=Clus_out;
var school cluster distance;
run;

/* Part C: Use Average Linkage to get cluster centriods   */
title '-------- Part C: Use Average Linkage to get cluster centriods as seeds ------------';
proc cluster data=university method=average outtree=ProTree noprint;
var x1 x2 x3 x4 x5 x6;
id school;
run;
proc tree data=ProTree nclusters=3 out=newdata noprint;
id school;
copy x1 x2 x3 x4 x5 x6;
run;
proc sort data=newdata;
by cluster;
run;
proc print data=newdata;
var school cluster;
run;
proc means data=newdata;
by cluster;
output out=Seeds mean=x1 x2 x3 x4 x5 x6;
var x1 x2 x3 x4 x5 x6;
run;
```

```
proc fastclus data=university maxc=3 maxiter=50 seed=Seeds out=Clus_out;
var x1 x2 x3 x4 x5 x6;
id school;
run;
proc sort data=Clus_out;
by cluster distance;
run;
proc print data=Clus_out;
var school cluster distance;
run;

proc candisc data=Clus_out noprint out=ProCan(keep=school cluster Can1 Can2);
class cluster;
var x1 x2 x3 x4 x5 x6;
run;
proc sgplot data=ProCan;
scatter y=Can2 x=Can1 / group=cluster datalabel=school;
label Can1="z1" Can2="z2";
run;quit;
```