Roll No. _____

(Do not write anything on question paper except Roll No.)

[This paper consists of Nineteen Pages]

Jagan Institute of Management Studies

End-Term Examination, September, 2016

Trimester IV – PGDM (IB) 2015-17

Business Analytics & Data Mining

ET_IB_BA&DM_2209

Time: 3 Hrs. *M. Marks: 70*

INSTRUCTIONS: Attempt any FIVE questions including Q1 & Q7 which are compulsory.

Q 1 Comment on any **THREE** of the following:

a) i) AHP, ii) Cohort Analysis, iii) Market Basket Analysis.

b) BIG DATA and Importance of BIG DATA, Name sources of BIG DATA.

c) Classification and Regression Trees and Sources of Data, Causes of Bad Quality data, Identifying Data Errors.

d) SEM, NEURAL Networks.

e) Missing Values and how SPSS treats Missing Values.

f) CHAID, RFM Analysis.

g) R programming language and Text Mining. **18**

Q 2 **a)** Why is data Visualization Important. Name some tool/graphs that you can use for data visualization. Support this with appropriate example from the business world. **6**

b) State the Scope of Business Analytics. Why do you think it is gaining strength in the business world? **6**

Q 3 **a)** A company has decided to use Logistic analysis to classify its customers between Loyal and disloyal. The dependent variable Loyalty was coded as 1 for "Loyal" and 0 for "Not Loyal. The independent variables used in the study were "Freq of Purchase" (Freq), "Average Purchase" (AvgPurc), "Years since Purchasing"(Years). From the output obtained

i) State how good is the linear Logistic model. Write the Null and alternate hypothesis. Accept or Reject it.

ii) Develop the Logistic equation.

iii) How good is the Logistic model?

iv) Suppose a person with "Freq of Purchase" (Freq) = 20, "Average Purchase" (AvgPurc) = 25000, "Years since Purchasing"(Years) = 5 years would he be classified as "Loyal" or "Not Loyal". (You may show how to calculate even if you are not able to calculate and suggest on what basis you will categorize the person (assume e-96.245 = 3. 013E-42) **4**

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	18	100.0
	Missing Cases	0	.0
	Total	18	100.0
Unselected Cases		0	.0
Total		18	100.0

a. If weight is in effect, see classification table for the total number of cases.

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Loyalty		Percentage
	Observed		0	1	Correct
Step 0	Loyalty	0	0	9	.0
		1	0	9	100.0
	Overall Percentage				50.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.000	.471	.000	1	1.000	1.000

Variables not in the Equation[a]

			Score	df	Sig.
Step 0	Variables	Freq	7.010	1	.008
		AvgPurc	7.139	1	.008
		Years	7.003	1	.008

a. Residual Chi-Squares are not computed because of redundancies.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	24.953	3	.000
	Block	24.953	3	.000
	Model	24.953	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.750	1.000

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table[a]

Observed			Predicted		
			Loyalty		Percentage Correct
			0	1	
Step 1	Loyalty	0	9	0	100.0
		1	0	9	100.0
	Overall Percentage				100.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Freq	9.478	1203.310	.000	1	.994	13075.018
	AvgPurc	.006	.786	.000	1	.994	1.006
	Years	-5.733	2257.418	.000	1	.998	.003
	Constant	-416.973	51162.247	.000	1	.993	.000

a. Variable(s) entered on step 1: Freq, AvgPurc, Years.

b) A chocolate company wants to draw a perceptual map using an attribute base procedure, of its consumer's perceptions regarding its own brand and two competing brands. Assume that it is Nestle against Cadbury's and Amul. Data was collected from 15 respondents (5 consumers of each brand) on five attributes namely Price, Quality, Availability, Packaging and Taste. The most significant output of the SPSS package is shown here with. i) Draw a perceptual map on the Graph; ii) indicate which brands are closer to which attribute. **4**

	Function	
	1 (X axis)	2 (Y axis)
PRICE	.207	.701
QUALITY	.988	-.454
AVAILABILITY	.999	-.122
PACKAGING	-.398	-.293
TASTE	-.136	.986
1 (Nestle)	2.745	.123

2 (Cadbury)	-1.596	1.073
3 (Amul)	-1.149	-1.196

c) What is the motivation of use of Hadoop for data storage? (or) Explain OLTP and OLAP. **4**

Q 4 a) The output from FACTOR ANALYSIS using SPSS produced the following result. The analysis was done on 12 variables so that they may be reduced to as few factors as possible. From the output given i) identify how many factors could be extracted and what is the total % of variance (of the original 12 variables) explained by these factors. ii) Which variable out of the 10 are included in factor 1. What information does the Scree Plot gives? **4**

Factor Analysis

Communalities

	Initial	Extraction
Var001	1.000	.670
Var002	1.000	.881
Var003	1.000	.922
Var004	1.000	.943
Var005	1.000	.973
Var006	1.000	.933
Var007	1.000	.948
Var008	1.000	.895
Var009	1.000	.940
Var010	1.000	.745
Var011	1.000	.922
Var012	1.000	.973

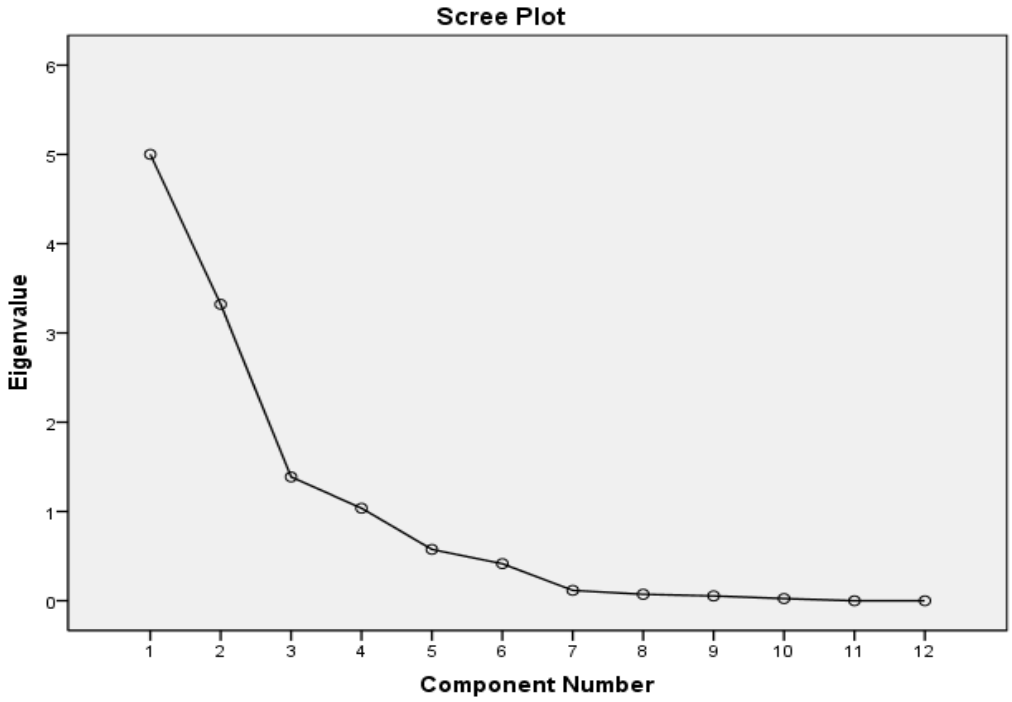
Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.001	41.673	41.673	5.001	41.673	41.673	4.814	40.115	40.115
2	3.321	27.672	69.344	3.321	27.672	69.344	2.159	17.990	58.105
3	1.387	11.558	80.902	1.387	11.558	80.902	1.908	15.902	74.007

4	1.036	8.636	89.538	1.036	8.636	89.538	1.864	15.531	89.538
5	.575	4.790	94.328						
6	.415	3.456	97.785						
7	.116	.964	98.749						
8	.072	.602	99.350						
9	.054	.448	99.798						
10	.024	.202	100.000						
11	2.927E-16	2.439E-15	100.000						
12	-1.497E-15	-1.247E-14	100.000						

Extraction Method: Principal Component Analysis.



Rotated Component Matrix[a]

	Component			
	1	2	3	4

Var001	.106	.182	.758	.226
Var002	-.134	-.920	.078	-.103
Var003	-.156	.754	.531	.219
Var004	.957	-.104	-.073	-.106
Var005	.981	.102	-.013	-.008
Var006	.934	-.228	.018	-.086
Var007	.972	.031	.015	-.056
Var008	-.268	.210	.150	.870
Var009	.064	.117	.033	.960
Var010	-.058	.030	.857	-.079
Var011	-.156	.754	.531	.219
Var012	.981	.102	-.013	-.008

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Component Transformation Matrix

Component	1	2	3	4
1	.946	-.186	-.165	-.209
2	.318	.656	.528	.435
3	.056	-.073	-.599	.795
4	.038	-.728	.578	.367

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

b) The Multidimensional technique was used to graph/position 10 brands of competing mobiles. (Each mobile is represented as VAR001 to VAR0010). The output from SPSS is as shown. i) explain if 2 dimension or 3 dimension are better in positioning the brands more clearly and why ii) suppose the dimension 1 is "price" and dimension 2 is "features". Explain which of the 10 brands are similar on the two dimensions (use the two dimensional output to answer the question), iii) Use the one dimensional output and assuming that it is price then which mobile brand is least priced and which is the most priced. **4**

SPSS output Alscal (Irrelevant details from SPSS command output have been deleted)

Iteration history for the 3 dimensional solution (in squared distances)

For matrix

Stress = .15441 RSQ = .72460

Iteration history for the 2 dimensional solution (in squared distances)

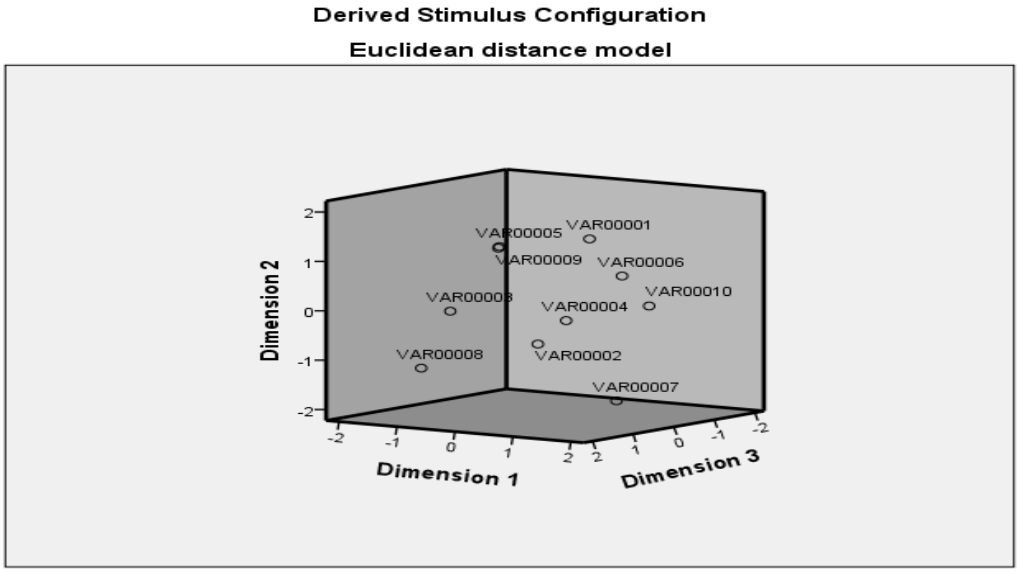
For matrix

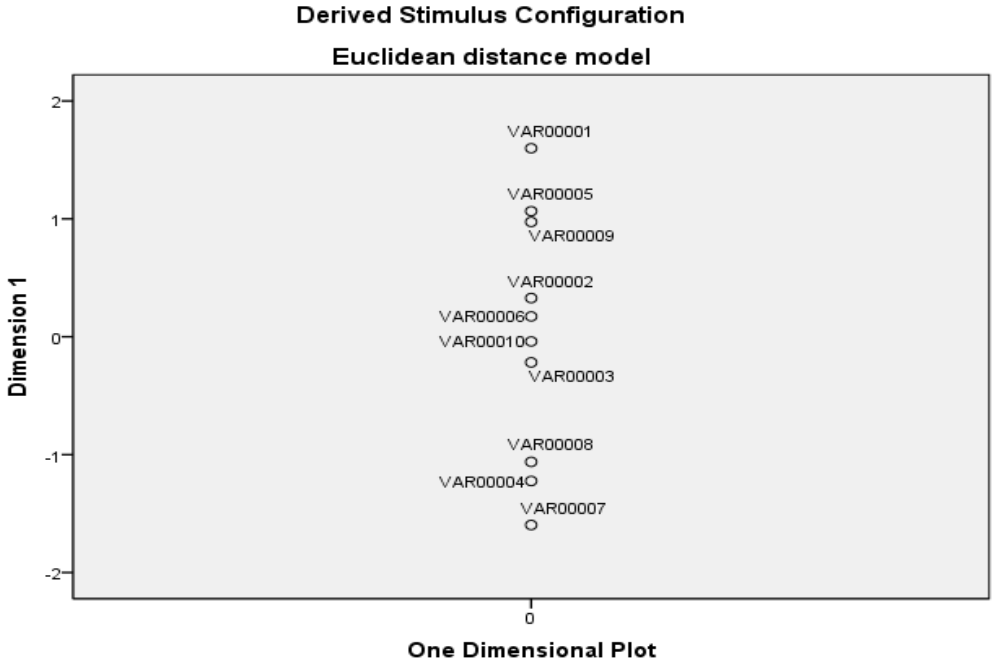
Stress = .23306 RSQ = .61919

Iteration history for the 1 dimensional solution (in squared distances)

For matrix

Stress = .43212 RSQ = .34938





c) An experiment was conducted to see the effect of temperature and Humidity on output. The table of data and the final ANOVA table are given herewith. For the following table, state the null and alternate hypothesis. Accept or reject the null hypothesis stating reason. **4**

	Temp 30	Temp 40	Temp 45
Humidity 10%	500	440	360
	600	450	510

Humidity 8%	300	280	250
	400	350	300
Humidity 6%	200	150	250
	250	275	220

ANOVA: Two-Factor With Replication

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	196769.4	2	98384.72	25.65628	0.000192	4.256495
Columns	12536.11	2	6268.056	1.634553	0.247989	4.256495
Interaction	9838.889	4	2459.722	0.641434	0.646349	3.633089
Within	34512.5	9	3834.722			
Total	253656.9	17				

Q 5 The following table gives hypothetical data on the SENSEX (Y) which is the dependent variable, and share prices of four companies being considered as independent variables;-TATA (X1), RELIANCE(X2), INFOSYS (X3), AIRTEL(X4).

SENSEX (Y)	TATA (X1)	RELIANCE (X2)	INFOSYS (X3)	AIRTEL (X4)
2001	245	338	414	323
2030	177	333	598	340
2226	271	358	656	340
2154	211	372	631	352
2078	196	339	528	380
2080	135	289	409	339
2073	195	334	382	331
1758	118	293	399	311
1624	116	325	343	328
1889	147	311	338	353
1988	154	304	353	518
2049	146	312	289	440
1796	115	283	388	276
1720	161	307	402	207
2056	274	322	151	287
1890	245	335	228	290
2187	201	350	271	355
2032	183	339	440	300
1856	237	327	475	284
2068	175	328	347	337
1813	152	319	449	279
1808	188	325	336	244
1834	188	322	267	253

1973	197	317	235	272
1839	261	315	164	223
1935	232	331	270	272

a) Multiple Regressions was used. And the output is as shown: i) explain the output ii) State the relevant hypothesis (ANOVA table) Accept or reject the null hypothesis stating reason. iii) Develop the Multiple Linear Regression equation. iv) If all the companies have a share price of 200 each then what could be the Sensex rating. **4**

Mutiple Regression for SENSEX (Y)

Regression Statistics	
Multiple R	0.770412704
R Square	0.593535735
Adjusted R Square	0.51611397
Standard Error	105.5652033
Observations	26

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	341731.7833	85432.94583	7.666264599	0.000567546
Residual	21	234024.2551	11144.01215		
Total	25	575756.0385			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	768.5509355	405.4723376	1.895446037	0.071881691	-74.67495141	1611.776822
TATA (X1)	1.435351763	0.652310462	2.200411993	0.039106917	0.078797896	2.79190563
RELIANCE(X2)	1.379955742	1.608971925	0.857663034	0.400761308	-1.966084549	4.725996033
INFOSYS (X3)	0.16572287	0.196176581	0.844763779	0.407764343	-0.242248664	0.573694405
AIRTEL(X4)	1.270892109	0.349602781	3.635245997	0.001548688	0.543853326	1.997930891

b) For the above problem Multiple Regression was used. However this time the analyst decided to use Backward Elimination. The output is as follows :- i) Explain the output ii) State the relevant hypothesis (ANOVA table), and calculate the R^2 value for the last table only. Accept or reject the null hypothesis stating reason. iii) Develop the Multiple Linear Regression equation from the last table iv) if all the **4**

companies (in the last table) have a share price of 300 each then what could be the sensex rating.

MR-BE

Table of Results for Backward Elimination

Model with all variables entered.

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	341731.7833	85432.94583	7.666264599	0.000567546
Residual	21	234024.2551	11144.01215		
Total	25	575756.0385			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	768.5509355	405.4723376	1.895446037	0.071881691	-74.67495141	1611.776822
TATA (X1)	1.435351763	0.652310462	2.200411993	0.039106917	0.078797896	2.791905633
RELIANCE(X2)	-1.379955742	1.608971925	-0.857663034	0.400761308	-4.725996033	3
INFOSYS (X3)	-0.165722872	0.196176581	-0.844763779	0.407764343	-0.573694405	0.242248664
AIRTEL(X4)	1.270892109	0.349602781	3.635245997	0.001548688	0.543853326	1.997930891

INFOSYS (X3) removed.

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	333779.1283	111259.7094	10.11548418	0.000219087
Residual	22	241976.9102	10998.95046		
Total	25	575756.0385			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	633.9168366	370.4025015	1.711426986	0.101066602	-134.2509355	1402.084609
TATA (X1)	1.217123922	0.595066694	2.045357158	0.052961598	-0.016968868	2.451216711
RELIANCE(X2)	-2.091526868	1.361912583	-1.535727692	0.138863052	-4.915960693	0.732906958
AIRTEL(X4)	1.294574218	0.346201477	3.739366532	0.001136452	0.576596298	2.012552139

RELIANCE(X2) removed.

	df	*SS*	*MS*	*F*	*Significance F*

Regression	2	307838.5486	153919.2743	13.21355806	0.000151102
Residual	23	267917.4899	11648.58652		
Total	25	575756.0385			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1152.218937	157.0676489	7.335813231	1.83557E-07	827.2997495	1477.138124
TATA (X1)	1.805818732	0.468397196	3.855314994	0.000805225	0.836865308	2.774772155
AIRTEL(X4)	1.447009897	0.341321194	4.239437581	0.000310022	0.740933211	2.153086582

No other variables could be removed from the model. Stepwise ends.

c) A retail outlet wants to know the consumer behavioral pattern of the purchase of products in two categories national brand (coded 1) and local brands (coded 2), which would help it to place orders depending on demand and requirements of the customer. The retail outlet uses data from a retail outlet in another location to arrive at a decision about customers visiting at their end. The variables included in the study were Annual Income (in thousands) and Household Size. The data on twenty respondents was analyzed using Discriminant Analysis.

i) Looking at the output State how good is the linear discriminant model. Write the Null and alternate hypothesis. Accept or Reject it.

ii) Develop the linear discriminant equation.

iii) Suppose a person with income of Rs. 18 (thousand) and Household size =4, would he be classified in buying a national brand or local brand. **4**

Discriminant

Group Statistics

Brand		Valid N (listwise) Unweighted	Weighted
1	Annual_Income	10	10.000
	Household_Size	10	10.000
2	Annual_Income	10	10.000
	Household_Size	10	10.000
Total	Annual_Income	20	20.000
	Household_Size	20	20.000

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Annual_Income	.550	14.711	1	18	.001
Household_Size	.625	10.796	1	18	.004

Analysis 1
Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.074[a]	100.0	100.0	.720

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.482	12.405	2	.002

Standardized Canonical Discriminant Function Coefficients

	Function
	1
Annual_Income	.703
Household_Size	-.518

Structure Matrix

	Function
	1
Annual_Income	.872
Household_Size	-.747

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.

Canonical Discriminant Function Coefficients

	Function
	1
Annual_Income	.323
Household_Size	-.331
(Constant)	-4.250

Unstandardized coefficients

Functions at Group Centroids

Brand	Function

	1
1	.983
2	-.983

Unstandardized canonical discriminant functions evaluated at group means

Classification Statistics

Classification Processing Summary

Processed		20
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		20

Prior Probabilities for Groups

Brand	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1	.500	10	10.000
2	.500	10	10.000
Total	1.000	20	20.000

Classification Results[a]

			Predicted Group Membership		
		Brand	1	2	Total
Original	Count	1	9	1	10
		2	2	8	10
	%	1	90.0	10.0	100.0
		2	20.0	80.0	100.0

Q 6 a) A consumer durable goods company wants to know various features and services the consumers perceive when purchasing through "online shopping". The purpose is to determine how the various features of "online shopping" marketing go together in terms of its convenience features, risk reducing features and son on. 10 statements are made in order to measure the variables of perception. The 10 statements are given below. The researcher asked the respondents to rate the following question on a 5 point scale (1=highly agree, 2= agree, 3= neither agree nor disagree, 4= disagree, 5= highly disagree). The questions are as follows. Results were obtained for 20 respondents.

i) The company should provide toll free number.

ii) The reputation of the company should be good.

iii) They should have discount schemes based on quantity.

iv) The company should provide guarantee for the products.
v) The company should give a trial period.
vi) The online site should be attractive and eye-catchy.
vii) The company should make on-time-delivery.
viii) Number of years in business is an important factor.
ix) Advertisements play an important role in decision-making.
x) It should have the facility to return the good, if not satisfied.

After using "Hierarchical Clustering" the researcher used the "K-means clustering technique". Based on the researcher results obtained from SPSS. i) How many clusters (or target markets) have been identified in the output? ii) How many people out of 20 are in cluster 4? iii) Discuss the features / characteristics of cluster 2.

Quick Cluster

Initial Cluster Centers

	Cluster			
	1	2	3	4
Var001	1	1	5	3
Var002	2	5	5	1
Var003	2	1	3	4
Var004	5	1	2	4
Var005	4	1	1	4
Var006	4	1	3	3
Var007	4	1	2	3
Var008	1	1	5	5
Var009	1	2	5	5
Var010	2	2	4	3

Iteration History[a]

	Change in Cluster Centers			
Iteration	1	2	3	4
1	2.926	2.330	.000	2.653
2	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 6.000.

Cluster Membership

Case Number	Cluster	Distance
1	1	2.182
2	4	2.836
3	4	3.408
4	4	3.747

5	1	2.926
6	4	1.478
7	4	3.670
8	1	2.750
9	1	2.272
10	2	3.207
11	2	2.268
12	2	2.330
13	4	2.653
14	2	2.619
15	2	3.505
16	3	.000
17	2	2.000
18	2	2.000
19	4	3.054
20	1	3.027

Final Cluster Centers

	Cluster			
	1	2	3	4
Var001	2	1	5	3
Var002	3	4	5	2
Var003	2	1	3	3
Var004	5	2	2	3
Var005	5	2	1	3
Var006	5	2	3	2
Var007	4	2	2	3
Var008	2	2	5	4
Var009	3	2	5	4
Var010	3	2	4	3

b)

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		5.888	7.692	5.459
2	5.888		6.130	4.660
3	7.692	6.130		4.524
4	5.459	4.660	4.524	

c)

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Var001	4.926	3	1.111	16	4.435	.019
Var002	5.917	3	.925	16	6.396	.005

Var003	3.726	3	.786	16	4.742	.015
Var004	8.429	3	.907	16	9.291	.001
Var005	9.993	3	.486	16	20.574	.000
Var006	8.238	3	.568	16	14.507	.000
Var007	7.974	3	.664	16	12.004	.000
Var008	10.640	3	.914	16	11.638	.000
Var009	6.012	3	1.282	16	4.689	.016
Var010	1.190	3	1.664	16	.715	.557

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster

Cluster	1	5.000
	2	7.000
	3	1.000
	4	7.000
Valid		20.000
Missing		.000

b) XYZ Paint Company identified the attributes which are important to their customers and also classified each of the attributes into their levels. Based on this thy want to us the technique of conjoint analysis to determine from a potential customers point of view, how important each attribute is to him, they also want to know how much utility the customer derives from a given combination of these levels of attributes. It also helps to understand the feasible offerings from the marketers' point of view. The three important attributes identified for the paint are given in the table. The variables have been dummy coded as following.

		Var1	var2
	5 years	1	0
Life	4 years	0	1
	3 years	-1	-1

		Var3	Var4
Price	Rs. 50	1	0
	Rs. 60	0	1
	Rs. 70	-1	-1

		Var5	Var6
Colour	Green	1	0
	Blue	0	1
	Cream	-1	-1

The data input from 27 people were taken and Var7 (Variable 7)

represents rating given to each combination of the design. Var7 was taken as the dependent variable and the balance 6 variables were taken as the independent variable. The conjoint analysis was run as regression model and the result obtained from SPSS are as follows:
i) identify the utility of each levels of **life, price, colour** ii) calculate the utility for each product mix and the net utility. Which out of **life, price, colour** is offering the most utitlty iii) If you offer life of paint 4 years and price Rs.60 and Colour Blue the what is the expected utility. **4**

Regression Output:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.978218
R Square	0.956911
Adjusted R Square	0.943984
Standard Error	1.878569
Observations	27

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1567.42	261.2366	74.0252	1.34E-12
Residual	20	70.58045	3.529022		
Total	26	1638			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	9.131944	0.447372	20.41244	7.31E-15	8.198744	10.06515
Var 1	5.969551	0.577645	10.33429	1.81E-09	4.764605	7.174497
Var 2	4.594551	0.577645	7.953938	1.27E-07	3.389605	5.799497
Var 3	3.540705	0.508385	6.964609	9.25E-07	2.480232	4.601178
Var 4	1.340705	0.508385	2.637183	0.015801	0.280232	2.401178
Var 5	-10.8547	0.735128	-14.7657	3.22E-12	-12.3882	-9.32125
Var 6	1.736645	0.532061	3.263999	0.003884	0.626786	2.846504

c) Output form excel command for AHP performed on a variety of cars is as follows (previous tables are not shown and only the final table is being shown)

	Style	Reliability	Fuel Economy					Importance (Benefits)	
versa	0.19	0.18	0.23		Style	0.27	versa	0.19	
sx4	0.12	0.24	0.30		Reliability	0.61	sx4	0.21	

hondacity	0.33	0.24	0.24		Fuel Economy	0.12	hondacity	0.26	
carolla	0.36	0.35	0.23				carolla	0.34	

	Importance (Benefits)			cost	Normalised cost	Benefits/cost
versa	0.19			900000		
sx4	0.21			850000		
hondacity	0.26			950000		
carolla	0.34			1400000		
				4100000		

i) If only Benefits were considered which car needs to be purchased? **1**

ii) Calculate the column of "Normalised cost" and "Benefits/Cost" and suggest which car needs to be purchased. **3**

Q 7 Read the case and answer the questions given at the end.

Case Study: Database Technology Used to Improve Airport Security

The Transportation Security Administration is evaluating various approaches to harness database technology in its efforts to improve airport security. Unfortunately, any such system may require airlines to invest in additional information systems technology – at a time when they are suffering from a lack of revenue one major airline has already declared a bankruptcy.

One idea is to develop a database system that links every airline reservation system in the country with a number of private and government databases. Data mining and predictive analysis would be used to sort through personal travel histories, the backgrounds of passengers aboard particular flights, and a wealth of other data to assign numerical threat ratings to individuals. Warnings would be sent electronically to workers at airport screening locations to inspect individuals with high threat ratings more closely.

Another approach would allow prescreened "trusted travelers" to pass through airport security checkpoints quickly, avoiding long lines and congestion, This system would devote more time and resources to screening other travelers whose level of risk is higher or unknown. Those who apply for the "trusted traveler" program would have to pass a background check using data from a number of state and federal databases. Once at the airport, "trusted traveler" passengers would be identified, perhaps by scanning their fingerprints or retinas or requiring

some form of identification card. (The federal government is considering developing a security ID card for airline passengers that would rely on biometric identification and be linked to government database). The system would also cross-check the passenger's identification with the FBI's watch list database and a federal passenger profiling system known as Computer Assisted Passenger Screening. Provided everything was clear, they could then proceed to his or her airplane using expedited security check in procedures.

Questions:

a) Which approach is best in terms of improving airport security: assigning a threat rating to individuals or prescreening individuals to identify "trusted travelers"? Why do you think this approach is best? **3**

b) Discuss the various "Data Mining Techniques" that you are aware of. Explain in detail any three. **10**

c) Identify specific data that could be used to assign numeric threat ratings to individuals. Describe how the system would work. **3**

****