Statistical Analysis using SAS

Question 1) Conduct a cluster analysis using the CountyData dataset. Explore all options available before you decide on a solution. Comment on the quality of the cluster solution. Interpret the clusters, and also comment on the variables used for the analysis.

Solution:

First, we need to clean the CountyData file so that it doesn't have any null values.

Now, we need to run a cluster analysis by selecting " Cluster Observations ". We do this because, we need to find out how many clusters we need to divide the total data into. We got 2955 clusters, as many as the observations.

Cluster History

Number of Clusters	Clusters Joined		Freq	New Cluster RMS Std Dev	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Tie
2955	Shelby	Osceola	2	0.00396	0.0000	1.00	.	.	707	.	
2954	Clark	Deuel	2	0.00409	0.0000	1.00	.	.	685	.	
2953	Mackinac	Big Stone	2	0.00409	0.0000	1.00	.	.	678	.	
2952	Stark	Brule	2	0.00428	0.0000	1.00	.	.	658	.	
2951	Kalkaska	Missaukee	2	0.00474	0.0000	1.00	.	.	617	.	
2950	Andrew	Iowa	2	0.00529	0.0000	1.00	.	.	565	.	

We look at the “Cubic Clustering Criterion” and see where the value turns from Negative to Positive.

11	CL29	CL21	209	0.0905	0.0132	.599	.598	0.59	441	70.3	
10	CL16	CL61	192	0.1032	0.0167	.583	.586	-1.2	457	69.7	
9	CL14	CL22	853	0.0594	0.0214	.561	.572	-4.0	471	259	
8	CL35	CL13	459	0.0758	0.0242	.537	.555	-6.9	489	197	
7	CL18	CL20	715	0.0649	0.0252	.512	.535	-8.0	515	270	
6	CL9	CL28	1012	0.0667	0.0274	.485	.511	-9.1	555	251	
5	CL7	CL6	1727	0.0712	0.0378	.447	.480	-11	596	284	
4	CL12	CL11	578	0.1035	0.0454	.401	.429	-8.7	660	182	
3	CL5	CL10	1919	0.0842	0.0865	.315	.360	-13	678	504	
2	CL3	CL4	2497	0.0986	0.1369	.178	.223	-13	639	565	
1	CL2	CL8	2956	0.1053	0.1779	.000	.000	0.00	.	639	

For our analysis, we got that at the 11^{th} element. Hence, we choose 11 clusters when performing “K-Means” clustering.

Now, running the “K-Means Clustering” by selecting Maximum Number of Clusters as “11”.

After we run the analysis, the following is observed:

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	1509	0.0690	0.6521		7	0.2770
2	1	.	0		10	1.0889
3	298	0.0683	0.6977		7	0.5023
4	1	.	0		7	1.0935
5	86	0.0887	0.6153		1	0.3667
6	12	0.1075	0.5307		10	0.5576
7	738	0.0654	0.5895		1	0.2770
8	31	0.0776	0.5222		10	0.3741
9	3	0.1120	0.4478		6	0.7534
10	170	0.0832	0.6337		8	0.3741
11	110	0.0936	0.6625		5	0.5090

From the Cluster Summary, we can observe that,

- Cluster 1 has 1509 counties under it. This cluster has the highest number of counties. And the nearest cluster for it is Cluster 7 with distance between their cluster centroids as 0.2770.
- Cluster 2 has only 1 county under it. And the nearest cluster for it is Cluster 10 with distance between their cluster centroids as 1.0889.
- Cluster 3 has 298 counties under it. And the nearest cluster for it is Cluster 7 with distance between their cluster centroids as 0.5023.
- Cluster 4 has only 1 county under it. And the nearest cluster for it is Cluster 7 with distance between their cluster centroids as 1.0935.
- Cluster 5 has 86 counties under it. And the nearest cluster for it is Cluster 1 with distance between their cluster centroids as 0.3667.
- Cluster 6 has 12 counties under it. And the nearest cluster for it is Cluster 10 with distance between their cluster centroids as 0.5576.

- Cluster 7 has 738 counties under it. And the nearest cluster for it is Cluster 1 with distance between their cluster centroids as 0.2770.

- Cluster 8 has 31 counties under it. And the nearest cluster for it is Cluster 10 with distance between their cluster centroids as 0.3741.

- Cluster 9 has only 3 counties under it. And the nearest cluster for it is Cluster 6 with distance between their cluster centroids as 0.7534.

- Cluster 10 has 170 counties under it. And the nearest cluster for it is Cluster 8 with distance between their cluster centroids as 0.3741.

- Cluster 11 has 110 counties under it. And the nearest cluster for it is Cluster 5 with distance between their cluster centroids as 0.5090.

Cluster Means												
Cluster	Popul	PopDen	PopChg_	Hispan_	Under5_	Age65_75_	Black_	Asian_	BirthRate	HSGrad	ColGrad	IncomeHH
1	0.009433557	0.002624294	0.449700104	0.044015850	0.525855975	0.276327416	0.059673558	0.015736019	0.379938587	0.675183121	0.215178413	0.381793427
2	0.159905801	1.000000000	0.436440678	0.278234086	0.408333333	0.228070175	0.201156069	0.204347826	0.378378378	0.683881064	0.774647887	0.424758419
3	0.005494701	0.003551906	0.417870549	0.019024847	0.560626398	0.263275639	0.516394460	0.008484097	0.431807647	0.408434950	0.130008237	0.243004756
4	0.000006745	0.000167577	0.076271186	0.041067762	0.000000000	1.000000000	0.000000000	0.369565217	0.789789790	0.311424100	0.018108652	0.000000000
5	0.004483751	0.001758603	0.461519511	0.109856263	0.743701550	0.181048552	0.070385805	0.018832154	0.596515120	0.545783746	0.144190726	0.313978345
6	0.094647041	0.058755231	0.431144068	0.161618754	0.506250000	0.190058480	0.056840077	0.568840580	0.434184184	0.730046948	0.396713615	0.521965593
7	0.002246565	0.000712286	0.417602545	0.036678575	0.473667570	0.356511197	0.045185394	0.005596795	0.340629247	0.476739994	0.111329222	0.261695314
8	0.024331358	0.007988197	0.732914161	0.075842883	0.685483871	0.127192982	0.060786873	0.055750351	0.478058704	0.814781160	0.427273317	0.710253625
9	0.638073065	0.049847709	0.459039548	0.332306639	0.644444444	0.175438596	0.209633911	0.151449275	0.531531532	0.644235785	0.398390342	0.429575403
10	0.030792342	0.011108840	0.479586241	0.064947457	0.523137255	0.183126935	0.072186331	0.068836317	0.394228548	0.845564125	0.556385935	0.564684437
11	0.012227442	0.006953090	0.423805855	0.571411238	0.635833333	0.254904306	0.036426695	0.018774704	0.489953590	0.439934557	0.162301079	0.255744975

- Cluster 1 has its highest mean of “0.6751” in the High school graduates category and the lowest mean of “0.002624” in the population density.
- Cluster 2 has its highest mean of “ 1.00” in the Population Density category and the lowest mean of “0.1599” in the population.
- Cluster 3 has its highest mean of “0.5606” in the Under-5 category and the lowest mean of “0.0035” in the population density.
- Cluster 4 has its highest mean of “0.0762” in the Population Chg category and the lowest mean of “0” in the black race and Under-5, which means the cluster entirely didn’t have black race and under 5 aged at all.

- Cluster 5 has its highest mean of “0.7437” in the Under-5 age category and the lowest mean of “0.0017” in the Population Density category.

- Cluster 6 has its highest mean of “0.56” in the Asian race category and the lowest mean of “0.056” in the black race.

- Cluster 7 has its highest mean of “0.4767” in the Highschool grads category and the lowest mean of “0.0007” in the Population Density category.

- Cluster 8 has its highest mean of “0.8147” in the Highschool grads category and the lowest mean of “0.0079” in population density category.

- Cluster 9 has its highest mean of “0.6444” in the Under-5 age category and the lowest mean of “0.0498” in population density category.

- Cluster 10 has its highest mean of “0.8455” in the Highschool grad category and the lowest mean of “0.0111” in population density category.

- Cluster 11 has its highest mean of “0.6358” in the Under-5 age category and the lowest mean of “0.0069” in population density category.

Approximate Expected Over-All R-Squared =	0.47086

The variance being explained is 47.086%.

The quality of the Cluster analysis is not so great. Because, as clusters always are, vague and is hard to identify the inherent features upon which these clusters are divided into.

Coming to the type of variables involved in this Cluster Analysis, we have used mostly numerical data. That involves the Population in a county, population density in a county and the races(Hispanic, Black and Asian) the population is divided into.

We have also got the variables that represent the Highschool Grads from that county, the average Household Income, the birth rate and the age groups(Age<5, Age between 65 &70) that are present in the county.

Question 2) Use the estress dataset posted to examine the following questions. The dataset lists information from a survey of entrepreneurs regarding economic stress and psychological outcomes. Do entrepreneurial self-efficacy (ese), gender and tenure, moderate the relationship between economic stress and depressed affect?

Solution:

Case1: Interaction effect of Entrepreneurial Self-Efficacy(ese) on relation between Economic Stress and Depressed Affect.

Ho: There is No Interaction Effect
H1: There is Interaction Effect

Here, the variables are:
Y- (Depressed) Affect
X- Economic Stress
M (Moderator)- Entrepreneurial Self-Efficacy(ese)

We run Linear Regression in SAS, choosing (Depressed) Affect(Y) as Dependent variable and Economic Stress(X) & Entrepreneurial Self-Efficacy(ese) as Independent variables. The output is as follows:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	26.81272	8.93757	20.98	<.0001
Error	258	109.89033	0.42593		
Corrected Total	261	136.70305			

Here, The P-value is less than α, hence this is a good model.

Root MSE	0.65263
Dependent Mean	1.59809
R-Square	0.1961
Adj R-Sq	0.1868
AIC	44.35825
AICC	44.59262
SBC	-205.36837

Here, this model explains 18.68% of variation in Depressed Affect.

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.726455	0.975543	-1.77	0.0780
ese	1	0.437343	0.164448	2.66	0.0083
estress	1	0.854554	0.190371	4.49	<.0001
ese*estress	1	-0.119712	0.032299	-3.71	0.0003

Here, the product term(ese*estress) has a P-value of 0.0003, which is less than the α value.

- As P-Value < α $(0.0003<0.0005)$ for the product term, we reject the null that there is No interaction effect. Hence, we can conclude that, Entrepreneurial Self-Efficacy(ese) moderates the relationship between Economic Stress and Depressed Affect.
- The regression line equation can be written as,

Depressed Affect = -1.72264 + 0.437343(ese) + 0.854554(estress) – 0.119712(ese*estress)

  - By observing this equation, we can infer that, the relationship between Depressed Affect and Economic Stress is being moderated by Entrepreneurial Self Efficacy. Here, the **interaction is negatively affecting** the relation between Depressed Effect and Economic stress.
  - For every unit increase in the product term(ese*estress), the Depressed Affect is decreased by 0.119712 times.

Therefore, we can conclude that, the higher levels of Entrepreneurial Self Efficacy reduces the Depressed Affect due to Economic Stress.

Case2: Interaction effect of Gender(sex) on relation between Economic Stress and Depressed Affect.

Ho: There is No Interaction Effect
H1: There is Interaction Effect

Here, the variables are:

Y- (Depressed) Affect
X- Economic Stress
M (Moderator)- Gender(sex)(categorical)

We run Linear Regression in SAS, choosing (Depressed) Affect(Y) as Dependent variable and Economic Stress(X) & Sex(categorical) as Independent variables. The output is as follows:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	15.88628	5.29543	11.31	<.0001
Error	258	120.81677	0.46828		
Corrected Total	261	136.70305			

Here, The P-value is less than α, hence this is a good model.

Root MSE	0.68431
Dependent Mean	1.59809
R-Square	0.1162
Adj R-Sq	0.1059
AIC	69.19381
AICC	69.42818
SBC	-180.53281

Here, this model explains 10.59% of variation in Depressed Affect.

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.844323	0.187356	4.51	<.0001
estress	1	0.163572	0.037638	4.35	<.0001
sex 0	1	-0.116308	0.295470	-0.39	0.6942
sex 1	0	0	.	.	.
estress*sex 0	1	0.025374	0.062398	0.41	0.6846
estress*sex 1	0	0	.	.	.

Here, the product term(ese*sex 0) has a P-value of 0.6846, which is greater than the α value.

- As P-Value > α (0.6846>0.0005) for the product term, we fail to reject the null that there is No interaction effect. Hence, we can conclude that, Gender **has no interaction effect** on the relationship between Economic Stress and Depressed Affect.

- The regression line equation can be written as,

Depressed Affect = 0.844323 + 0.163572(estress) - 0.116308(sex) + 0.025374(estress*sex)

  - From the equation, we can tell that, For every unit increase in the product term(estress*sex), the Depressed Affect is increased by 0.025374 times which is very small to be considered.

Case 3: Interaction effect of Tenure on relation between Economic Stress and Depressed Affect.

Ho: There is No Interaction Effect
H1: There is Interaction Effect

Here, the variables are:
Y- (Depressed) Affect
X- Economic Stress
M (Moderator)- Tenure

We run Linear Regression in SAS, choosing (Depressed) Affect(Y) as Dependent variable and Economic Stress(X) & Tenure as Independent variables. The output is as follows:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	16.87920	5.62640	12.11	<.0001
Error	258	119.82384	0.46443		
Corrected Total	261	136.70305			

Here, The P-value is less than α, hence this is a good model.

Root MSE	0.68149
Dependent Mean	1.59809
R-Square	0.1235
Adj R-Sq	0.1133
AIC	67.03168
AICC	67.26606
SBC	-182.69494

Here, our model explains 11.33% of variation in Depressed Affect.

Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.842995	0.190541	4.42	<.0001
estress	1	0.175943	0.040216	4.37	<.0001
tenure	1	-0.009745	0.020246	-0.48	0.6307
estress*tenure	1	-0.000000900	0.004167	-0.00	0.9998

Here, the product term(estress*tenure) has a P-value of 0.9998, which is greater than the α value.

- As P-Value > α $(0.9998 > 0.0005)$ for the product term, we fail to reject the null that there is No interaction effect. Hence, we can conclude that, Tenure does not moderate the relationship between Economic Stress and Depressed Affect.

- The regression line equation can be written as,

Depressed Affect = 0.842995 + 0.175943 (estress) – 0.009745(tenure) – 0.000000900 (ese*tenure)

  - From the equation we can infer that, For every unit increase in the product term(estress*tenure), the Depressed Affect is decreased by 0.000000900 times, which is very negligible.

Question 3) Explore the personalitySubset dataset. Find and discuss the principal components that emerge.

Solution:

Exploring the PersonalitySubset dataset, we could figure out that the data is from the Likert Scale which is good but not a greatly constructed Likert scale.

We select the Factor Analysis from the Multivariate Analysis and choose all the variables into the "Analysis Variables".

By running the test, we got the following output:

Eigenvalues of the Correlation Matrix: Total = 10 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	2.53009580	1.17545527	0.2530	0.2530
2	1.35464053	0.03521235	0.1355	0.3885
3	1.31942818	0.26808932	0.1319	0.5204
4	1.05133885	0.13550348	0.1051	0.6256
5	0.91583537	0.21613416	0.0916	0.7171
6	0.69970121	0.05431140	0.0700	0.7871
7	0.64538981	0.09747609	0.0645	0.8516
8	0.54791372	0.04003939	0.0548	0.9064
9	0.50787433	0.08009214	0.0508	0.9572
10	0.42778220		0.0428	1.0000

Here, we should make sure that, we select only factors that have Eigen Value more than 1.
By following that rule, we get to select 4 Factors.

Final Communality Estimates: Total = 10.000000

talkative	finds_fault	does_a_thorough_job	depressed	original	reserved	helpful	careless	relaxed	curious
1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000

The total communality is also equally shared among all the 10 variables.

Now, we run the Factor Analysis again, but this time only choosing 4 Factors as the outputs and by employing Varimax Rotation Method(Oblique rotation type) to ensure that we have a maximum clarity on which variable belongs to which factor.

Rotated Factor Pattern (Standardized Regression Coefficients)					
		Factor1	Factor2	Factor3	Factor4
Values less than 0.3 are not printed.					
does_a_thorough_job	does a thorough job	0.80146	.	.	.
helpful	helpful	0.61771	0.33138	.	-0.39112
original	original	0.37044	0.31743	.	.
careless	careless	-0.66488	.	.	.
talkative	talkative	.	0.80862	.	.
reserved	reserved	.	-0.82826	.	.
relaxed	relaxed	.	.	0.84946	.
depressed	depressed	.	.	-0.77782	.
finds_fault	finds fault	.	.	.	0.76930
curious	curious	.	.	.	0.63229

Here, the variables are separated into clearly observed Bins or Factors.

- Factor 1 has variables like does a thorough job, helpful, original, careless.
- Factor 2 has variables like talkative, reserved.
- Factor 3 has relaxed, depressed.
- Factor 4 has finds fault, curious.

Labelling can be done as follows:

Factor 1: Personality
Factor 2: Nature
Factor 3: Mindset
Factor 4: Evaluating Others

The results are as follows:

Final Communality Estimates: Total = 6.255503									
talkative	finds_fault	does_a_thorough_job	depressed	original	reserved	helpful	careless	relaxed	curious
0.73970998	0.64633582	0.65745676	0.67539108	0.40440515	0.72834588	0.66247460	0.51376620	0.70681158	0.52080631

Here, the communality is not equally shared among the variables. This is because, we have brought down the size of the Factors to 4.

Question 4) The following table contains information on where the privacy disclaimer is located on a web site, for three nationalities. Conduct the appropriate test to infer if nationality of website and location of privacy disclaimer are independent.

Privacy Disclaimer Location and Web Site Nationality

	Nationality of Web Site			
Location	*France*	*UK*	*USA*	Row Total
Home page	56	68	35	159
Order page	19	19	28	66
Client page	6	10	16	32
Other page	12	9	13	34
Col Total	93	106	92	291

Solution: This problem can be framed as a Chi-Square Test of Independence test. Therefore, the null and alternate hypothesis can be set as:

Ho : Variables are Independent
H1 : Variables are Dependent

α is assumed as 0.05.

Here, we have got the Observed values. From them, we have to calculate the Expected values.

These can be calculated by using (Row(total) X Column(total)) / Grand Total.

By calculating that, these are the obtained Expected values:

50.81443299	57.91753	50.26804
21.09278351	24.04124	20.86598
10.22680412	11.65636	10.11684
10.86597938	12.38488	10.74914

From the expected frequency, we have to calculate the Chi-Square statistic, which is obtained by

$$(Fjk - Ejk)^2/Ejk$$

Where, Fjk is Observed Frequency;
Ejk is Estimated Frequency.

Now, by using that formula for each element in the Observed and Expected tables, the output obtained is as follows:

0.529182432	1.75519	4.637401
0.207641765	1.057103	2.439102
1.746965414	0.235367	3.421186
0.118351298	0.925113	0.471328

Now, we have to calculate the Grand total of all the values obtained here, with that, we will be getting the Chi-square Statistic.

The Chi-Square statistic is obtained as: 17.54393146

We have to find the degrees of freedom our problem, that is used for finding the P-value from the Chi-square statistic.

Degrees of Freedom : ((No.of rows)-1 X (No.of Columns)-1)

In our problem, (4-1) x (3-1) = 6

Now, the P-value can be obtained from Excel formula : CHISQ.DIST.RT(Chi-square Stistic, Df)

P-value is obtained as "0.007478972".(<0.05)

As P-value < α, we reject null that these variables are independent. Hence, nationality of website and location of privacy disclaimer are dependent.

Question 5) Open the HEART file available in the SAS library. Run the appropriate model to predict if the person is dead or alive, based on some of the other variables in the dataset.

Solution:

Ho: Dependent Variable is independent of Predictor Value
H1: Dependent Variable is Dependent of Predictor Value

As we are predicting a categorical variable(y), we have to go for Binary Logistic Regression.

Among different variables that are present, I have considered Sex, Cholesterol_Status, BP_Status, Weight_Status and the AgeAtStart as the independent variables.

By selecting the event of interest as "Alive", we run the Binary Logistic Regression in SAS, we have obtained the following results:

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	1386.1354	10	<.0001
Score	1247.7299	10	<.0001
Wald	989.6490	10	<.0001

Since both the Score and Wald P-values are less than α value. Hence, we can say that this is a good model.

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Sex	1	64.0372	<.0001
Chol_Status	2	5.1137	0.0775
BP_Status	2	94.9863	<.0001
Smoking_Status	4	104.4574	<.0001
AgeAtStart	1	714.1179	<.0001

In this table, we look at the individual P-values of each independent variable involved. Sex, BP_Status, Smoking_Status and AgeAtStart are all significant predictors of the Status. But, Chol_Stat is not a significant predictor.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	5.5197	0.2456	505.1402	<.0001
Sex	**Female**	1	0.5853	0.0731	64.0372	<.0001
Sex	**Male**	0	0	.	.	.
Chol_Status	**Borderline**	1	0.1712	0.0775	4.8759	0.0272
Chol_Status	**Desirable**	1	0.1271	0.0901	1.9914	0.1582
Chol_Status	**High**	0	0	.	.	.
BP_Status	**High**	1	-0.7813	0.1117	48.9241	<.0001
BP_Status	**Normal**	1	-0.1531	0.1130	1.8353	0.1755
BP_Status	**Optimal**	0	0	.	.	.
Smoking_Status	**Heavy (16-25)**	1	0.2430	0.1287	3.5637	0.0591
Smoking_Status	**Light (1-5)**	1	0.7299	0.1550	22.1788	<.0001
Smoking_Status	**Moderate (6-15)**	1	0.2821	0.1505	3.5161	0.0608
Smoking_Status	**Non-smoker**	1	0.9994	0.1255	63.3908	<.0001
Smoking_Status	**Very Heavy (> 25)**	0	0	.	.	.
AgeAtStart		1	-0.1254	0.00469	714.1179	<.0001

- From this table, we can clearly infer that, Females are 0.5853 times more likely to be alive when compared to females given
- Dealing with the Blood Pressure, the status of being alive is **reduced by 0.7813** for people with high cholesterol than people with optimal cholesterol levels. And it is **reduced by 0.1531** times for people with Normal Blood pressure levels.
- Coming to Smoking_Status, people belonging to Non-smoker Smoking_Status are 0.9994 time more likely to be alive when compared to Very-Heavy-Smokers.

- It follows the same when compared to Light-Smoking Status people, but this time these are 0.7299 times more likely to be alive than Very-Heavy-Smokers.

- People who smoke moderately are 0.2821 times more likely to be alive than Very-Heavy Smokers.

- And people who smoke heavily are 0.2430 times more likely to be alive than Very-Heavy Smokers.

- The people with more AgeAtStart are less likely to be Alive as it has a negative effect(-0.1254)on the Status of being Alive.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	79.5	Somers' D	0.590
Percent Discordant	20.5	Gamma	0.590
Percent Tied	0.1	Tau-a	0.278
Pairs	6008888	c	0.795

Here, our model explains 79.5 percent variation in the Status of a person being Alive or Not.