Objectives of data summarization and data reduction psychology essay

Chapter 3

The basic difference between the objectives of data summarization and data reduction depends upon the ultimate research question. In data summarization the ultimate research question may be to better understand the interrelationship among the variables. This may be accomplished by condensing a large number of respondents into a smaller number of distinctly different groups with Q-type factor analysis. More often data summarization is applied to variables in R-type factor analysis to identify the dimensions that are latent within a dataset. Data summarization makes the identification and understanding of these underlying dimensions or factors the ultimate research question.

Data reduction relies on the identification of the dimensions as well, but makes use of the discovery of the items that comprise the dimensions to reduce the data to fewer variables that represent the latent dimensions. This is accomplished by either the use of surrogate variables, summated scales, or factor scores. Once the data has been reduced to the fewer number of variables further analysis may become easier to perform and interpret.

(2) HOW CAN FACTOR ANALYSIS HELP THE RESEARCHER IMPROVE THE RESULTS OF OTHER MULTIVARIATE TECHNIQUES?

Answer

Factor analysis provides direct insight into the interrelationships among variables or respondents through its data summarizing perspective. This gives the researcher a clear picture of which variables are highly correlated

and will act in concert in other analysis. The summarization may also lead to a better understanding of the latent dimensions underlying a research question that is ultimately being answered with another technique. From a data reduction perspective, the factor analysis results allow the formation of surrogate or summated variables to represent the original variables in a way that avoids problems associated with highly correlated variables. In addition, the proper usage of scales can enrich the research process by allowing the measurement and analysis of concepts that require more than single item measures.

(3) WHAT GUIDELINES CAN YOU USE TO DETERMINE THE NUMBER OF FACTORS TO EXTRACT? EXPLAIN EACH BRIEFLY.

Answer

The appropriate guidelines utilized depend to some extent upon the research question and what is known about the number of factors that should be present in the data. If the researcher knows the number of factors that should be present, then the number to extract may be specified in the beginning of the analysis by the a priori criterion. If the research question is largely to explain a minimum amount of variance then the percentage of variance criterion may be most important.

When the objective of the research is to determine the number of latent factors underlying a set of variables a combination of criterion, possibly including the a priori and percentage of variance criterion, may be used in selecting the final number of factors. The latent root criterion is the most

commonly used technique. This technique is to extract the number of factors having eigenvalues greater than 1. The rationale being that a factor should explain at least as much variance as a single variable. A related technique is the scree test criterion. To develop this test the latent roots (eigenvalues) are plotted against the number of factors in their order of extraction. The resulting plot shows an elbow in the sloped line where the unique variance begins to dominate common variance. The scree test criterion usually indicates more factors than the latent root rule. One of these four criterion for the initial number of factors to be extracted should be specified. Then an initial solution and several trial solutions are calculated. These solutions are rotated and the factor structure is examined for meaning. The factor structure that best represents the data and explains an acceptable amount of variance is retained as the final solution.

(4) HOW DO YOU USE THE FACTOR-LOADING MATRIX TO INTERPRET

THE MEANING OF FACTORS?

Answer

The first step in interpreting the factor-loading matrix is to identify the largest significant loading of each variable on a factor. This is done by moving horizontally across the factor matrix and underlining the highest significant loading for each variable. Once completed for each variable the researcher continues to look for other significant loadings. If there is simple structure, only single significant loadings for each variable, then the factors are labeled. Variables with high factor loadings are considered more important than variables with lower factor loadings in the interpretation

phase. In general, factor names will be assigned in such a way as to express the variables which load most significantly on the factor.

(5) HOW AND WHEN SHOULD YOU USE FACTOR SCORES IN CONJUNCTION WITH OTHER MULTIVARIATE STATISTICAL TECHNIQUES?

Answer

When the analyst is interested in creating an entirely new set of a smaller number of composite variables to replace either in part or completely the original set of variables, then the analyst would compute factor scores for use as such composite variables. Factor scores are composite measures for each factor representing each subject. The original raw data measurements and the factor analysis results are utilized to compute factor scores for each individual. Factor scores may replicate as easily as a summated scale, therefore this must be considered in their use.

(6) WHAT ARE THE DIFFERENCES BETWEEN FACTOR SCORES AND SUMMATED SCALES? WHEN ARE EACH MOST APPROPRIATE?

Answer

The key difference between the two is that the factor score is computed based on the factor loadings of all variables loading on a factor, whereas the summated scale is calculated by combining only selected variables. Thus, the factor score is characterized by not only the variables that load highly on a factor, but also those that have lower loadings. The summated scale represents only those variables that load highly on the factor.

Although both summated scales and factor scores are composite measures there are differences that lead to certain advantages and disadvantages for each method. Factor scores have the advantage of representing a composite of all variables loading on a factor. This is also a disadvantage in that it makes interpretation and replication more difficult. Also, factor scores can retain orthogonality whereas summated scales may not remain orthogonal. The key advantage of summated scales is, that by including only those variables that load highly on a factor, the use of summated scales makes interpretation and replication easier. Therefore, the decision rule would be that if data are used only in the original sample or orthogonality must be maintained, factor scores are suitable. If generalizability or transferability is desired then summated scales are preferred.

(7) WHAT IS THE DIFFERENCE BETWEEN Q-TYPE FACTOR ANALYSIS AND CLUSTER ANALYSIS?
Answer

Both Q-Type factor analysis and cluster analysis compare a series of responses to a number of variables and place the respondents into several groups. The difference is that the resulting groups for a Q-type factor analysis would be based on the intercorrelations between the means and standard deviations of the respondents. In a typical cluster analysis approach, groupings would be based on a distance measure between the respondents' scores on the variables being analyzed.

(8) WHEN WOULD THE RESEARCHER USE AN OBLIQUE ROTATION INSTEAD OF AN ORTHOGONAL ROTATION? WHAT ARE THE BASIC DIFFERENCES BETWEEN THEM?

Answer

In an orthogonal factor rotation, the correlation between the factor axes is arbitrarily set at zero and the factors are assumed to be independent. This simplifies the mathematical procedures. In oblique factor rotation, the angles between axes are allowed to seek their own values, which depend on the density of variable clusterings. Thus, oblique rotation is more flexible and more realistic (it allows for correlation of underlying dimensions) than orthogonal rotation although it is more demanding mathematically. In fact, there is yet no consensus on a best technique for oblique rotation.

When the objective is to utilize the factor results in a subsequent statistical analysis, the analyst may wish to select an orthogonal rotation procedure. This is because the factors are orthogonal (independent) and therefore eliminate collinearity. However, if the analyst is simply interested in obtaining theoretically meaningful constructs or dimensions, the oblique factor rotation may be more desirable because it is theoretically and empirically more realistic.

Chapter 4

Multiple Regression Analysis

ANSWERS TO QUESTIONS

(1) HOW WOULD YOU EXPLAIN THE " RELATIVE IMPORTANCE" OF THE PREDICTOR VARIABLES USED IN A REGRESSION EQUATION?

Answer

Two approaches: (a) beta coefficients and (b) the order that variables enter the equation in stepwise regression. Either approach must be used cautiously, being particularly concerned with the problems caused by multi-collinearity.

With regard to beta coefficients, they are the regression coefficients which are derived from standardized data. Their value is basically that we no longer have the problem of different units of measure. Thus, they reflect the impact on the criterion variable of a change of one standard deviation in any predictor variable. They should be used only as a guide to the relative importance of the predictor variables included in your equation, and only over the range of sample data included.

When using stepwise regression, the partial correlation coefficients are used to identify the sequence in which variables will enter the equation and thus their relative contribution.

Chapter 4

(2) WHY IS IT IMPORTANT TO EXAMINE THE ASSUMPTION OF LINEARITY WHEN USING REGRESSION?

Answer

The regression model is constructed with the assumption of a linear relationship among the predictor variables. This gives the model the properties of additivity and homogeneity. Hence coefficients express directly the effect of changes in predictor variables. When the assumption of linearity is violated, a variety of conditions can occur such as multicollinearity, heteroscedasticity, or serial correlation (due to non-independence or error terms). All of these conditions require correction before statistical inferences of any validity can be made from a regression equation.

Basically, the linearity assumption should be examined because if the data are not linear, the regression results are not valid.

(3) HOW CAN NONLINEARITY BE CORRECTED OR ACCOUNTED FOR IN THE REGRESSION EQUATION?

Answer

Nonlinearity may be corrected or accounted for in the regression equation by three general methods. One way is through a direct data transformation of the original variable as discussed in Chapter 2. Two additional ways are to explicitly model the nonlinear relationship in the regression equation through the use of polynomials and/or interaction terms. Polynomials are power transformations that may be used to represent quadratic, cubic, or higher order polynomials in the regression equation. The advantage of polynomials

over direct data transformations in that polynomials allow testing of the type of nonlinear relationship. Another method of representing nonlinear relationships is through the use of an interaction or moderator term for two independent variables. Inclusion of this type of term in the regression equation allows for the slope of the relationship of one independent variable to change across values of a second dependent variable.

(4) COULD YOU FIND A REGRESSION EQUATION THAT WOULD BE ACCEPTABLE AS STATISTICALLY SIGNIFICANT AND YET OFFER NO ACCEPTABLE INTERPRETATIONAL VALUE TO MANAGEMENT?

Answer

Yes. For example, with a sufficiently large sample size you could obtain a significant relationship, but a very small coefficient of determination-too small to be of value.

In addition, there are some basic assumptions associated with the use of the regression model, which if violated, could make any obtained results at best spurious. One of the assumptions is that the conditions and relationships existing when sample data were obtained remain unchanged. If changes have occurred they should be accommodated before any new inferences are made. Another is that there is a " relevant range" for any regression model. This range is determined by the predictor variable values used to construct the model. In using the model, predictor values should fall within this relevant range. Finally, there are statistical considerations. For example, the effects of multicollinearity among predictor variables is one such consideration.

(5) WHAT IS THE DIFFERENCE IN INTERPRETATION BETWEEN THE REGRESSION COEFFICIENTS ASSOCIATED WITH INTERVAL SCALED PREDICTOR VARIABLES AS OPPOSED TO DUMMY (0, 1) PREDICTOR VARIABLES?

Answer

The use of dummy variables in regression analysis is structured so that there are (n-1) dummy variables included in the equation (where n = the number of categories being considered). In the dichotomous case, then, since n = 2, there is one variable in the equation. This variable has a value of one or zero depending on the category being expressed (e. g., male = 0, female = 1). In the equation, the dichotomous variable will be included when its value is one and omitted when its value is zero. When dichotomous predictor variables are used, the intercept (constant) coefficient (bo) estimates the average effect of the omitted dichotomous variables. The other coefficients, b1 through bk, represent the average differences between the omitted dichotomous variables and the included dichotomous variables. These coefficients (b1-bk) then, represent the average importance of the two categories in predicting the dependent variable.

Coefficients bo through bk serve a different function when metric predictors are used. With metric predictors, the intercept (bo) serves to locate the point where the regression equation crosses the Y axis, and the other coefficients (b1-bk) indicate the effect on the predictor variable(s) on the criterion variable (if any).

(6) WHAT ARE THE DIFFERENCES BETWEEN INTERACTIVE AND CORRELATED PREDICTOR VARIABLES? DO ANY OF THESE DIFFERENCES AFFECT YOUR INTERPRETATION OF THE REGRESSION EQUATION?

Answer

The term interactive predictor variable is used to describe a situation where two predictor variables' functions intersect within the relevant range of the problem. The effect of this interaction is that over part of the relevant range one predictor variable may be considerably more important than the other; but over another part of the relevant range the second predictor variable may become the more important. When interactive effects are encountered, the coefficients actually represent averages of effects across values of the predictors rather than a constant level of effect. Thus, discrete ranges of influence can be misinterpreted as continuous effects.

When predictor variables are highly correlated, there can be no real gain in adding both of the variables to the predictor equation. In this case, the predictor with the highest simple correlation to the criterion variable would be used in the predictive equation. Since the direction and magnitude of change is highly related for the two predictors, the addition of the second predictor will produce little, if any, gain in predictive power.

When correlated predictors exist, the coefficients of the predictors are a function of their correlation. In this case, little value can be associated with the coefficients since we are speaking of two simultaneous changes.

(7) ARE INFLUENTIAL CASES ALWAYS TO BE OMITTED? GIVE EXAMPLES OF WHEN THEY SHOULD AND SHOULD NOT BE OMITTED?

Answer

The principal reason for identifying influential observations is to address one question: Are the influential observations valid representations of the population of interest? Influential observations, whether they be " good" or " bad," can occur because of one of four reasons. Omission or correction is easily decided upon in one case, the case of an observation with some form of error (e. g., data entry).

However, with the other causes, the answer is not so obvious. A valid but exceptional observation may be excluded if it is the result of an extraordinary situation. The researcher must decide if the situation is one which can occur among the population, thus a representative observation. In the remaining two instances (an ordinary observation exceptional in its combination of characteristics or an exceptional observation with no likely explanation), the researcher has no absolute guidelines. The objective is to assess the likelihood of the observation occurring in the population. Theoretical or conceptual justification is much preferable to a decision based solely on empirical considerations.

Chapter 5

Multiple Discriminant Analysis

ANSWERS TO QUESTIONS

(1) HOW WOULD YOU DIFFERENTIATE BETWEEN MULTIPLE DISCRIMINANT ANALYSIS, REGRESSION ANALYSIS, AND ANALYSIS OF VARIANCE?

Answer

Basically, the difference lies in the number of independent and dependent variables and in the way in which these variables are measured. Note the following definitions:

Multiple discriminant analysis (MDA) – the single dependent (criterion) variable is nonmetric and the independent (predictor) variables are metric.

Regression Analysis – both the single dependent variable and the multiple independent variables are metric.

Analysis of Variance (ANOVA) – the multiple dependent variables are metric and the single independent variable is nonmetric.

(2) WHEN WOULD YOU EMPLOY LOGISTIC REGRESSION RATHER THAN DISCRIMINANT ANALYSIS? WHAT ARE THE ADVANTAGES AND DISADVANTAGES OF THE DECISION?

Answer

Both discriminant analysis and logistic regression are appropriate when the dependent variable is categorical and the independent variables are metric.

In the case of a two-group dependent variable either technique might be applied, but only discriminant analysis is capable of handling more than two groups. When the basic assumptions of both methods are met, each gives comparable predictive and classificatory results and employs similar diagnostic measures. Logistic regression has the advantage of being less affected than discriminant analysis when the basic assumptions of normality and equal variance are not met. It also can accommodate nonmetric dummy-coded variables as independent measures. Logistic regression is limited though to the prediction of only a two-group dependent measure. Thus, when more than two groups are involved, discriminant analysis is required.

(3) WHAT CRITERIA COULD YOU USE IN DECIDING WHETHER TO STOP A DISCRIMINANT ANALYSIS AFTER ESTIMATING THE DISCRIMINANT FUNCTION(S)? AFTER THE INTERPRETATION STAGE?

Answer

a. Criterion for stopping after derivation. The level of significance must be assessed. If the function is not significant at a predetermined level (e. g., . 05), then there is little justification for going further. This is because there is little likelihood that the function will classify more accurately than would be expected by randomly classifying individuals into groups (i. e., by chance).

b. Criterion for stopping after interpretation. Comparison of " hit-ratio" to some criterion. The minimum acceptable percentage of correct classifications usually is predetermined.

(4) WHAT PROCEDURE WOULD YOU FOLLOW IN DIVIDING YOUR SAMPLE INTO ANALYSIS AND HOLDOUT GROUPS? HOW WOULD YOU CHANGE THIS PROCEDURE IF YOUR SAMPLE CONSISTED OF FEWER THAN 100 INDIVIDUALS OR OBJECTS?

Answer

When selecting individuals for analysis and holdout groups, a proportionately stratified sampling procedure is usually followed. The split in the sample typically is arbitrary (e. g., 50-50 analysis/hold-out, 60-40, or 75-25) so long as each " half" is proportionate to the entire sample.

There is no minimum sample size required for a sample split, but a cut-off value of 100 units is often used. Many researchers would use the entire sample for analysis and validation if the sample size were less than 100. The result is an upward bias in statistical significance which should be recognized in analysis and interpretation.

(5) HOW DO YOU DETERMINE THE OPTIMUM CUTTING SCORE?

Answer

a. For equal group sizes, the optimum cutting score is defined by:

ZA + ZB

ZCE = ———-

N

ZCE = critical cutting score value for equal size groups

ZA = centroid for group A

ZB = centroid for Group B

N = total sample size

b. For unequal group sizes, the optimum cutting score is defined by:

NAZA + NBZB

ZCU = ————

NA + NB

ZCU = critical cutting score value for unequal size groups

NA = sample size for group A

NB = sample size for Group B

(6) HOW WOULD YOU DETERMINE WHETHER OR NOT THE CLASSIFICATION ACCURACY OF THE DISCRIMINANT FUNCTION IS SUFFICIENTLY HIGH RELATIVE TO CHANCE CLASSIFICATION?

Answer

Some chance criterion must be established. This is usually a fairly straight-forward function of the classifications used in the model and of the sample size. The authors then suggest the following criterion: the classification accuracy (hit ratio) should be at least 25 percent greater than by chance.

Another test would be to use a test of proportions to examine for significance between the chance criterion proportion and the obtained hit-ratio proportion.

(7) HOW DOES A TWO-GROUP DISCRIMINANT ANALYSIS DIFFER FROM A THREE-GROUP ANALYSIS?

Answer

In many cases, the dependent variable consists of two groups or classifications, for example, male versus female. In other instances, more than two groups are involved, such as a three-group classification involving low, medium, and high classifications. Discriminant analysis is capable of handling either two groups or multiple groups (three or more). When two classifications are involved, the technique is referred to as two-group discriminant analysis. When three or more classifications are identified, the technique is referred to as multiple discriminant analysis.

(8) WHY SHOULD A RESEARCHER STRETCH THE LOADINGS AND CENTROID DATA IN PLOTTING A DISCRIMINANT ANALYSIS SOLUTION?

Answer

Plots are used to illustrate the results of a multiple discriminant analysis. By using the statistically significant discriminant functions, the group centroids can be plotted in the reduced discriminant function space so as to show the separation of the groups. Plots are usually produced for the first two significant functions. Frequently, plots are less than satisfactory in

illustrating how the groups differ on certain variables of interest to the researcher. In this case stretching the discriminant loadings and centroid data, prior to plotting the discriminant function, aids in detecting and interpreting differences between groups. Stretching the discriminant loadings by considering the variance contributed by a variable to the respective discriminant function gives the researcher an indication of the relative importance of the variable in discriminating among the groups. Group centroids can be stretched by multiplying by the approximate F-value associated with each of the discriminant functions. This stretches the group centroids along the axis in the discriminant plot that provides more of the accounted-for variation.

(9) HOW DO LOGISTIC REGRESSION AND DISCRIMINANT ANALYSES EACH HANDLE THE RELATIONSHIP OF THE DEPENDENT AND INDEPENDENT VARIABLES?

Answer

Discriminant analysis derives a variate, the linear combination of two or more independent variables that will discriminate best between the dependent variable groups. Discrimination is achieved by setting variate weights for each variable to maximize between group variance. A discriminant (z) score is then calculated for each observation. Group means (centroids) are calculated and a test of discrimination is the distance between group centroids.

Logistic regression forms a single variate more similar to multiple regression.

It differs from multiple regression in that it directly predicts the probability of

an event occurring. To define the probability, logistic regression assumes the relationship between the independent and dependent variables resembles an S-shaped curve. At very low levels of the independent variables, the probability approaches zero. As the independent variable increases, the probability increases. Logistic regression uses a maximum likelihood procedure to fit the observed data to the curve.

(10) WHAT ARE THE DIFFERENCES IN ESTIMATION AND INTERPRETATION BETWEEN LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS?

Answer

Estimation of the discriminant variate is based on maximizing between group variance. Logistic regression is estimated using a maximum likelihood technique to fit the data to a logistic curve. Both techniques produce a variate that gives information about which variables explain the dependent variable or group membership. Logistic regression may be comfortable for many to interpret in that it resembles the more commonly seen regression analysis.

(11) EXPLAIN THE CONCEPT OF ODDS AND WHY IT IS USED IN PREDICTING PROBABILITY IN A LOGISTIC REGRESSION PROCEDURE.

Answer

One of the primary problems in using any predictive model to estimate probability is that is it difficult to " constrain" the predicted values to the appropriate range. Probability values should never be lower than zero or

higher than one. Yet we would like for a straight-forward method of estimating the probability values without having to utilize some form of nonlinear estimation. The odds ratio is a way to express any probability value in a metric value which does not have inherent upper and lower limits. The odds value is simply the ratio of the probability of being in one of the groups divided by the probability of being in the other group. Since we only use logistic regression for two-group situations, we can always calculate the odds ratio knowing just one of the probabilities (since the other probability is just 1 minus that probability). The odds value provides a convenient transformation of a probability value into a form more conducive to model estimation.

Chapter 7

Cluster Analysis

ANSWERS TO QUESTIONS

(1) WHAT ARE THE BASIC STAGES IN THE APPLICATION OF CLUSTER ANALYSIS?

Answer

Partitioning – the process of determining if and how clusters may be developed.

Interpretation – the process of understanding the characteristics of each cluster and developing a name or label that appropriately defines its nature.

Profiling – stage involving a description of the characteristics of each cluster to explain how they may differ on relevant dimensions.

(2) WHAT IS THE PURPOSE OF CLUSTER ANALYSIS AND WHEN SHOULD IT BE USED INSTEAD OF FACTOR ANALYSIS?

Answer

Cluster analysis is a data reduction technique that's primary purpose is to identify similar entities from the characteristics they possess. Cluster analysis identifies and classifies objects or variables so that each object is very similar to others in its cluster with respect to some predetermined selection criteria.

As you may recall, factor analysis is also a data reduction technique and can be used to combine or condense large numbers of people into distinctly different groups within a larger population (Q factor analysis).

Factor analytic approaches to clustering respondents are based on the intercorrelations between the means and standard deviations of the respondents resulting in groups of individuals demonstrating a similar response pattern on the variables included in the analysis. In a typical cluster analysis approach, groupings are devised based on a distance measure between the respondent's scores on the variables being analyzed.

Cluster analysis should then be employed when the researcher is interested in grouping respondents based on their similarity/dissimilarity on the variables being analyzed rather than obtaining clusters of individuals who have similar response patterns.

(3) WHAT SHOULD THE RESEARCHER CONSIDER WHEN SELECTING A SIMILARITY MEASURE TO USE IN CLUSTER ANALYSIS?

Answer

The analyst should remember that in most situations, different distance measures lead to different cluster solutions; and it is advisable to use several measures and compare the results to theoretical or known patterns. Also, when the variables have different units, one should standardize the data before performing the cluster analysis. Finally, when the variables are intercorrelated (either positively or negatively), the Mahalanobis distance measure is likely to be the most appropriate because it adjusts for intercorrelations and weighs all variables equally.

(4) HOW DOES THE RESEARCHER KNOW WHETHER TO USE HIERARCHICAL OR NONHIERARCHICAL CLUSTER TECHNIQUES? UNDER WHICH CONDITIONS WOULD EACH APPROACH BE USED?

Answer

The choice of a hierarchical or nonhierarchical technique often depends on the research problem at hand. In the past, hierarchical clustering techniques were more popular with Ward's method and average linkage being probably the best available. Hierarchical procedures do have the advantage of being fast and taking less computer time, but they can be misleading because undesirable early combinations may persist throughout the analysis and lead to artificial results. To reduce this possibility, the analyst may wish to cluster

analyze the data several times after deleting problem observations or outlines.

However, the K-means procedure appears to be more robust than any of the hierarchical methods with respect to the presence of outliers, error disturbances of the distance measure, and the choice of a distance measure. The choice of the clustering algorithm and solution characteristics appears to be critical to the successful use of CA.

If a practical, objective, and theoretically sound approach can be developed to select the seeds or leaders, then a nonhierarchical method can be used. If the analyst is concerned with the cost of the analysis and has an a priori knowledge as to initial starting values or number of clusters, then a hierarchical method should be employed.

Punj and Stewart (1983) suggest a two-stage procedure to deal with the problem of selecting initial starting values and clusters. The first step entails using one of the hierarchical methods to obtain a first approximation of a solution. Then select candidate number of clusters based on the initial cluster solution, obtain centroids, and eliminate outliers. Finally, use an iterative partitioning algorithm using cluster centroids of preliminary analysis as starting points (excluding outliers) to obtain a final solution.

Punj, Girish and David Stewart, " Cluster Analysis in Marketing Research: Review and Suggestions for Application," Journal of Marketing Research, 20 (May 1983), pp. 134-148.

(5) HOW CAN YOU DECIDE HOW MANY CLUSTERS TO HAVE IN YOUR SOLUTION?

Answer

Although no standard objective selection procedure exists for determining

the number of clusters, the analyst may use the

(5) HOW CAN YOU DECIDE HOW MANY CLUSTERS TO HAVE IN YOUR SOLUTION?

https://assignbuster.com/objectives-of-data-summarization-and-data-reduction-psychology-essay/