CHAPTER 18

Multivariate Data Analysis

CHAPTER SUMMARY

1. MULTIVARIATE ANALYSIS PROCEDURE

I. **Multivariate Analysis**

A. **Multivariate Analysis Defined**

1. General term for statistical procedures that simultaneously analyze multiple measurements on each individual or object under study

B. **Five Techniques for Multivariate Analysis**

1. Multiple regression analysis
2. Multiple discriminant analysis
3. Cluster analysis
4. Factor analysis
5. Conjoint analysis

2. MULTIVARIATE SOFTWARE

1. Running the various types of analyses presented in this text requires appropriate software. Personal computers of today have the power to handle most problems that a marketing researcher might encounter.

2. There is a wide variety of outstanding Window software available for multivariate analysis. SPSS for Windows is one of the best and the most widely used by professional marketing researchers.

3. MULTIPLE REGRESSION ANALYSIS

I. **Multiple Regression Analysis**

A. **Multiple Regression Analysis Defined**

1. Procedure for predicting the level or magnitude of a (metric) dependent variable based on the levels of multiple independent variables

2. **General Equation for Multiple Regression**:

$$\boldsymbol{Y} = a + b_1\boldsymbol{X_1} + b_2\boldsymbol{X_2} + b_3\boldsymbol{X_3} + \ldots b_n\boldsymbol{X_n}$$

where

$\boldsymbol{Y}$ = dependent or criterion variable

a = estimated constant

$\boldsymbol{X_1 - X_n}$ = predictor (independent) variables that influence the dependent variable

$b_1 - b_n$ = coefficients associated with the predictor variables so that a change of one unit in X_n is associated with a change of b_n units in Y; values for the coefficients are estimated from the regression analysis

B. **Applications of Multiple Regression Analysis**

1. Estimating the effects that various marketing mix variables have on sales or market share.

2. Estimating the relationship between various demographic or psychographic factors and frequency of visiting fast food restaurants or other service businesses.

3. Determine the relative influence of individual satisfaction elements on overall satisfaction.

4. Quantifying the relationship between various classification variables, such as age and income, and overall attitude toward a product or service.

5. Determining which variables are predictive of sales of a particular product or service.

C. **Purposes of Multiple Regression Analysis**

1. Predicting the level of the dependent variable, based on given levels of the independent variables

2. Understanding the relationship between the independent variables and the dependent variable

D. **Multiple Regression Analysis Measures**

1. **Coefficient of Determination**, $\boldsymbol{R^2}$–measure of the percentage of the variation in the dependent variable explained by variations in the independent variables

a. This statistic can assume values from 0 to 1

b. ***b* Values or Regression Coefficients**–estimates of the effect of the individual independent variables on the dependent variable.

1) We determine the likelihood that each individual *b* value is the result of chance (H_0: $b_n = 0$)

c. Adjusted R^2-- As models get larger, it is wise to look at a variation of the R^2 statistic called **adjusted R^2**, as the measure of fit for a regression model. The standard R^2 value tends to increase with every predictor variable that is added to the model, regardless whether that variable truly adds to the explanatory power of the model. The adjusted R^2 corrects the coefficient of determination based on the relationship between the number of predictor variables and the overall sample size, producing a more rational estimate of model fit when several independent variables are included.

E. **Dummy Variables**

1. **Dummy Variables**–nominally scaled independent variables

2. **Dichotomous Nominally Scaled Independent Variables--**can be transformed into dummy variables by coding one value (for example, female) as 0 and the other (for example, male) as 1.

3. **Nominally Scaled Independent Variables**–can assume more than two values, a slightly different approach is required.

a. Example: A question regarding racial group with three categories: African American, Hispanic, or Caucasian. Binary or dummy variable coding of responses requires the use of two dummy variables. For example, X_1 (1 for African American, 0 otherwise) and X_2 (1 for Hispanic, 0 otherwise). Then, $X_1 = 0$ and $X_2 = 0$ corresponds to the "left out" or reference category (in this case, Caucasian).

If there are K categories, K – 1 dummy variables are needed to uniquely identify every category (including K categories would over identify the

model since the last category is represented by "0's" on the previous K – 1 variables)

	X_1	X_2
If person is African American	1	0
If person is Hispanic	0	1
If person is Caucasian	0	0

F. **Potential Use and Interpretation Problems**

1. **Collinearity** – A key assumption when interpreting multiple regression results is that the independent variables are not correlated (collinear) with each other.

a. If they are correlated, then estimated *b* values (regression coefficients) will be less precise (larger standard errors).

b. Check for collinearity by examining the matrix showing the correlations between each variable. If a correlation of .30 or greater exists, check for distortions of *b* values.

c. **Strategies for dealing with collinearity**

1) If two variables are heavily correlated, one variable can be dropped

2) The correlated variables can be combined in some fashion to form a new composite independent variable (create an index or use factor analysis).

2. **Causation**–Regression analysis can show that variables are correlated, but it cannot prove causation.

3. **Scaling of Coefficients**

a. Magnitudes of regression coefficients can be compared directly only if they are scaled in the same units or if the data have been standardized

b. Standardization–achieved by taking each number in a series, subtracting the mean of the series from the number, and dividing the result by the

standard deviation of the series. This process converts any set of numbers to a new set with a mean of 0 and a standard deviation of 1.

4. **Sample Size**

a. The value of R^2 is influenced by the number of predictor variables relative to sample size.

b. Rules of thumb suggest that the number of observations should be equal to at least 10 to 15 times the number of predictor variables.

4. MULTIPLE DISCRIMINANT ANALYSIS

I. Multiple Discriminant Analysis

A. **Multiple Discriminant Analysis Defined**–a procedure for predicting group membership for a (nominal or categorical) dependent variable on the basis of two or more independent variables

1. In multiple discriminant analysis, the dependent variable must be metric; in multiple discriminant analysis, the dependent variable is nominal or categorical in nature.

2. **Goals of Multiple Discriminant Analysis**

a. To determine if there are statistically significant differences between the average discriminant score profiles of the two or more groups.

b. To establish a model for classifying individuals or objects into groups on the basis of their values on the independent variables.

c. To determine how much of the difference in the average score profiles of the two or more groups is accounted for by each independent variable.

3. **General Discriminant Analysis Equation**

$$Z = b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

where

Z = discriminant score

$b_1 - b_2$ = discriminant weights

$X_1 - X_n$ = independent variables

a. Independent variables with large discriminatory power (large differences between groups) will have large weights, and those with little discriminatory power will have small weights.

b. The goal of discriminant analysis is the prediction of a categorical variable.

c. The problem is finding a linear combination of independent variables that shows large differences in group means.

4. **Discriminant Score**–a score that is the basis for predicting to which group a particular object or individual belongs; also called ***Z-score***

5. **Discriminant Coefficient**–estimate of the discriminatory power of a particular independent variable; also called ***discriminatory weight***

B. **Applications of Discriminant Analysis**

1. **Questions Answered by Discriminant Analysis**

a. How are consumers who purchase various brands different from those who do not purchase those brands?

b. How do we target likely buyers for a new product from our database of existing customers in order to conduct the most effective prelaunch marketing campaign?

c. How do consumers who frequent one fast food restaurant differ in demographic and lifestyle characteristics from consumers who frequent another fast food restaurant?

d. How do consumers who have chosen either indemnity insurance, HMO coverage, or PPO coverage differ from one another in regard to health care use, perceptions, and attitudes?

C. **Example of Multiple Discriminant Analysis**

United Wireless's marketing director wants to predict whether or not the five importance ratings used in the regression analysis predict whether or not an individual currently has a wireless telephone. Dummy variables are used for that variable. The results show that the ability to place and receive calls when away from home is the most important variable,

while range of coverage is the least important variable in discriminating between those that currently have and do not have wireless telephone service. The model correctly predicted 73% of all respondents as wireless users or nonusers.

5. CLUSTER ANALYSIS

I. **Cluster Analysis**

A. **Cluster Analysis Defined**–the term cluster analysis is used to refer to a group of techniques used to identify objects or people that are similar in regard to certain variables or measurements.

1. **Purpose**–to classify objects or people into some number of mutually exclusive and exhaustive groups, so that those within a group are as similar as possible to one another.

B. **Procedures for Clustering**

1. **Different Procedures**–but all are similar in their general approach, which involves measuring the similarity between people or objects in regard to their values on the variables used for clustering. Some examples include k-means, two-stage, nearest neighbor, decision trees, ensemble analysis, random forest, BIRCH, and self-organizing neural networks.

2. **Scatter Plots**– in the case of two clustering variables, the dots indicate the positions of consumers with respect to the variables. The distance between any pair of dots is negatively related to how similar the corresponding individuals are (the smaller the distance between two dots, the more similar the individuals).

3. **Computer algorithms**–the basic idea behind most of the algorithms is to start with some arbitrary cluster boundaries and modify the boundaries until a point is reached where the average inter-point distances within clusters are as small as possible relative to average distances between clusters.

6. FACTOR ANALYSIS

I. Factor Analysis

A. **Factor Analysis Defined**

1. Procedure for simplifying data by reducing a large set of variables to a smaller set of factors or composite variables by identifying underlying dimensions of the data

a. **Objective**–to summarize the information contained in a large number of measures into a smaller number of summary measures called *factors*

b. No dependent variable

c. If after the data has been analyzed there are several measures of a concept, they can be used to form an average score on the concept.

B. **Factor Scores**

1. **Factor**–technical definition "a linear combination of variables"–weighted summary score of a set of related variables.

2. **Factor Analysis**–each measure is first weighted according to how much it contributes to the variation of each factor

3. **Factor Score**–calculated on each subject in the data set.

4. **Relative Sizes**–of the scoring coefficients are used in determining relative importance of each variable.

C. **Factor Loadings**

1. Correlation between each factor score and each of the original variables

a. Because the loadings are correlation coefficients, values near +1 or –1 indicate a close positive or negative association

2. **Nature of the Factors Derived**–determined by examining the factor loadings

D. **Naming Factors**

1. **Identify Factors**–the next step is to "name" the factors–name should communicate the concept that the researcher feels the questions are measuring

E. **Number of Factors to Retain**

1. **Final Results** –one factor or up to as many factors as there are variables

2. **Decision**–determined by the percent of the variation explained by each factor

3. **When to Stop**–stop factoring when additional factors no longer make sense

7. CONJOINT ANALYSIS

I. **Conjoint Analysis**

A. **Conjoint Analysis Defined**

1. Procedure used to quantify the value that consumers associate with different levels of product/service attributes

2. **Conjoint Analysis**

a. Is not a completely standardized procedure

b. A typical conjoint analysis application involves presenting various product or service combinations in a carefully controlled exercise, then estimating the relative value of each feature tested. They type of conjoint approach (e.g. ratings-based, discrete choice, graded pairs, dual choice, full profile, partial profile, adaptive choice, etc.) impacts how the exercise is presented and what statistical procedures are most appropriate for analyzing the results.

B. **Example of Conjoint Analysis**

1. Golf Ball Manufacturer – Titleist, a major manufacturer of golf balls conducted a focus group recently and determined from this group, past research studies, and personal experience that the three most important features are:

a. Average driving distance

b. Average ball life

c. Price

2. Approach

a. **Traditional Approach**

b. **Considering Features Conjointly**

1) Respondents are asked to evaluate features conjointly or in combination, so that advantages for one attribute can only be chosen at the expense of another attribute.

2) This allows researchers to examine acceptable tradeoffs

c. **Estimating Utilities**

1) The researcher calculates a set of values, referred to as **utilities**, for each attribute levels.

- Utilities for this simple example can be computed using ordinary least square regression.

d. **Simulating Buyer Choice**

1) Three steps discussed–collecting trade-off data, using the data to estimate buyer preference structures, and predicting choice–are the basis of any conjoint analysis application.

2) One of the most common approaches to conducting conjoint analysis is the use of a discrete choice or choice-based conjoint exercise. Two or more products are shown side-by-side with details provided on each key attribute being tested. Respondents are asked to select a single product from amongst the options shown. The exercise is repeated multiple times in order to present a wide variety of product designs, but no individual sees more than a fraction of the sometimes thousands or even millions of possible product combinations.

C. **Limitations of Conjoint Analysis**

1. Conjoint analysis suffers from a certain degree of artificiality

2. Respondents may be more deliberate in their choice processes in this context than in a real situation

3. The survey may provide more product information than respondents would get in a real market situation

4. If key attributes or popular options within key attributes are excluded from the study, demand estimates could be severely impacted.

5. Testing too many attributes or features will diminish the amount of attention that can be given to each individual's most desired features, reducing measurement precision.

6. The presentation of information (e.g. the order in which attributes are listed; whether pictures are used for some attributes, but not others; how price is displayed; etc.) can greatly impact what features a respondent focuses on and ultimately how they make their decisions.

7. It is important to either be as neutral as possible in the presentation of a conjoint exercise or else try to replicate how the product or service is actually evaluated and compared in the marketplace in order to avoid biasing results.

D. Big Data and Hadoop

1. Big Data:

a. "Big data" is the term used to describe large and complex data sets. Companies have been collecting transaction based information since the beginning of the computer age.

b. However, the sheer volume of information has grown exponentially over the last few years and the types of information now being generated does not easily fit into traditional hierarchical database structures.

c. Big data describes the new data capture and management approaches that are designed to handle the higher volume, faster acquisition rates and broader array of data types. Most of the tools for big data are still evolving and individuals with the skills to capitalize on them are in short supply.

2. Hadoop:

a. Hadoop is an open-source platform distributed by Apache for managing large amounts of information across hundreds or thousands of networked computers.

b. Each computer works independently on a small portion of the total dataset so that a task such as clustering several billion records can be

handled in a fraction of the time taken for more conventional database structures.

c. There are numerous backup copies of each data chunk so that any failure can be immediately picked up by another computer with access to the same information.

d. Google and Yahoo have had a hand in developing the platform and underlying technology for Hadoop as they sought ways to store and access the vast array of search information they were collecting.

E. Predictive Analytics

1. Predictive analytics describes a wide array of tools and techniques that are used to extract and analyze information from data sets. Statistics, machine learning, database management and computer programming all play a part in identifying patterns and transforming data into insights.

2. Predictive analytics can apply to big data or traditional databases, observational data like loyalty card usage, Internet sources like social media text and web tracking data or primary survey research results. Fraud detection, trend analyses, targeted direct marketing, predicting heavy users and likely buyers are just some of the applications for predictive analytics.

F. **Predictive Analytics Process**

1. Acquiring a Data Set:

a. Before applying predictive analytics, an organization must assemble a target data set relevant to the problem of interest.

b. Predictive analytics can only uncover patterns and relationships that exist in the available data.

c. Typically, the data set must be large enough to include all the patterns and combinations that are likely to be found in the real world.

d. In the past, assembling such large data sets was very costly and time consuming. Today, most companies capture terabytes of information on their customers as a normal course of business and many social media companies provide access to massive amounts of data in real time for anyone to tap into.

e. In addition, third-party vendors provide a wide variety of data elements that can be purchased for just about any household or company in the United States.

2. Preprocessing:

a. Once assembled, the data set must be cleaned in a process where observations that contain excessive noise, errors and missing data are edited or excluded.

b. Data transformations may be used to smooth out irregular distributions and minimize extreme values.

c. Imputing missing values from comparable records and building predictive models to fill-in missing information is often used.

d. Linking multiple data sets is also part of pre-processing available data.

3. Modeling: a variety of techniques may be employed as part of the modeling process:

a. Clustering: This is the task of discovering groups and structures in the data that are similar in certain selected sets of variables.

b. Classification:

1) Readily available information like demographics and geography might be used to classify individuals on key behaviors such as purchase frequency or product preference.

2) Proprietary information such as online ads viewed or previous products purchased can be very effective at predictive future behaviors whenever such information is available.

3) Customer segments identified through clustering might also be modeled in order to predict which segment new customers and prospects belong. Successful models

c. Estimation:

1) Calculations such as risk scores, fraud detection, retention rates, lifetime value and likelihood to purchase rates may be calculated for individuals or groups.

2) These calculations can be used to predict future outcomes based on limited present-day data. They can also be used to monitor individuals or groups in order to detect changes in behavior that allow the organization to react before customers or revenues are lost.

G. **Validating Results**

1. A final step of knowledge discovery from the target data and modeling is to attempt to verify the patterns produced by the predictive modeling algorithms in a wider data set.

2. In the evaluation process, the patterns or models identified in the wider data set are applied to a test data set that was not used to develop the predictive modeling algorithm. The resulting output is compared to the desired output.

H. **Applying the Results**

1. Once the models and calculations are in place and have been validated, they are applied to existing and future customer records to improve the efficiency and effectiveness of marketing efforts. For example, specific information captured from a new sales inquiry can be used to classify an individual into the correct market segment. Based on their market segment, the most appropriate product offering can be prepared and the marketing messages can be adjusted to most resonate with that individual. Purchasing prospect lists with specific information

appended to each record allows an organization to avoid wasting marketing dollars on unlikely purchasers (based on applied predictive models) and focus resources on the most likely buyers and those with the greatest potential lifetime value.

QUESTIONS FOR REVIEW AND CRITICAL THINKING

1. Distinguish between multiple discriminant analysis and cluster analysis. Give several examples of situations in which each might be used.

Multiple discriminant analysis analyzes the relationships between a set of metric independent variables and a nominal or categorical dependent variable. It can test a hypothesized relationship and it describes how the independent variables discriminate between the groups of the dependent variable. Cluster analysis is a statistical tool used for clustering people or objects based on a particular criteria or variable in the study. For example, we might have 15 different measures of benefits and want to cluster people into benefit groups for market segmentation. The same concept could be used for personal values, attitudes toward rating healthy alternatives, or the types of restaurants frequented. Multiple discriminant analysis could be used for segmenting users from nonusers, light from heavy users, patrons from nonpatrons, and a host of other dependent categorical variables. A number of independent variable sets such as benefits, attributes, knowledge, and preferences could be used to predict group membership.

2. What purpose does multiple regression analysis serve? Give an example of how it might be used in marketing research. How is the strength of multiple regression measures of association determined?

Multiple regression analysis is used to examine the relationship between two or more metric predictor variables and one metric dependent variable. It can also be used to generate predictions for the dependent variable, given a combination of values for the independent variables. Multiple regression analysis has many applications in marketing research. One general application relates to determining the effects of various marketing variables on sales or market share. The

Coefficient of Determination or R^2 provides a measure of the percentage of variation in the dependent variable explained by variation in the independent variable(s).

3. What is a dummy variable? Give an example using a dummy variable.

The term "dummy variable" describes the coding of nominally scaled independent variables so that they can be used in regression analysis. An example of a dummy variable for a measure of location of birth would be

0 = born in the United States, 1 = born outside the United States

4. Describe the potential problem of collinearity and multiple regression. How might a researcher test for collinearity? If collinearity is a problem, what should the researcher do?

Collinearity refers to the condition when a significant correlation exists between two or more independent variables. This condition reduces the statistical power of significance tests for the regression coefficients. One can test for collinearity by examining the correlation matrix. If there is a value higher than .30, the researcher should consider corrective action. This correction might be accomplished be dropping one of the correlated variables, or collapsing the correlated variables into a single variable.

5. A sales manager examined age data, education level, a personality measure that indicated introvertedness / extrovertedness, and levels of sales attained by the company's 120-person sales force. The technique used was multiple regression analysis. After analyzing the data, the sales manager said, "It is apparent to me that the higher level of education and the greater the degree of extrovertedness a salesperson has, the higher will be an individual's level of sales. In other words, a good education and being extroverted cause a person to sell more." Would you agree or disagree with the sales manager's conclusions? Why?

The manager should consider whether age and education are correlated (collinearity), as older salespersons may have greater education and thus the "education effect" may really be an "age/ experience effect." It is also plausible that the extroverted salespersons are also older, as greater

experience generally leads to greater confidence and performance. Another important issue is whether the manager has correctly specified the regression model. It is likely that many other factors that have a significant impact on salesperson performance have not been included in the model. Any of these could be confounded with the included variables. Perhaps most importantly, the manager should keep in mind that causation can never be proven with statistical evidence alone.

6. The factors produced and the result of the factor loadings from factor analysis are mathematical constructs. It is the task of the researcher to make sense out of these factors. The following table lists four factors produced from a study of cable TV viewers. What label would you put on each of these factors? Why?

See the text (p489) for the table.

- Factor 1 Variety of programming or repetitive programming. All of the questions deal with the movie channels playing the same movies over and over again.
- Factor 2 Emotional programming or "Tear-jerking" programming. All of the items are about programming that elicits an emotional response.
- Factor 3 Religious programming. These questions measure the viewers' opinions of religious programs.
- Factor 4 Home entertainment. The items measure the viewers' preferences for viewing movies at home.

7. The following table is a discriminant analysis that examines responses to various attitudinal questions from cable TV users, former cable TV users, and people that have never used cable TV. Looking at the various discriminant weights, what can you say about each of the three groups?

For "users," the most discriminating variables are A19 (easygoing on repairs) and A18 (no repair service). For "formers," the most discriminating variables are A4 and A18 (burned out on repeats and no repair service, respectively). For the "nevers," the most discriminating variables are A7

and A19 (breakdown complainer and easygoing on repairs, respectively). These results suggest that concerns about, or reactions to, service failure (breakdowns/repairs) are the most predictive of whether a consumer is a user, a former user, or never a user.

8. The following table shows regression coefficients for two dependent variables. The first dependent variable is willingness to spend money for cable TV. The independent variables are responses to attitudinal statements. The second dependent variable is stated desire never to allow cable TV in their homes. By examining the regression coefficients, what can you say about persons willing to spend money for cable TV and those that will not allow cable TV in their home?

See the text (p490) for table.

Those who are willing to spend money for cable television enjoy watching movies and comedy, and they are likely to do so late at night. They may be somewhat lonely ("forlorn") and may have a greater need for external sources of "stimulation," such as might be offered by television programming. They are also dissatisfied with the service level and wish the cable stations offered more variety. Those who will not allow cable in their homes do not enjoy watching sporting events, object to sex on television, and do not feel a need for many choices in their television programming.

REAL-LIFE RESEARCH

Case 18.1 – Satisfaction Research for Pizza Quik

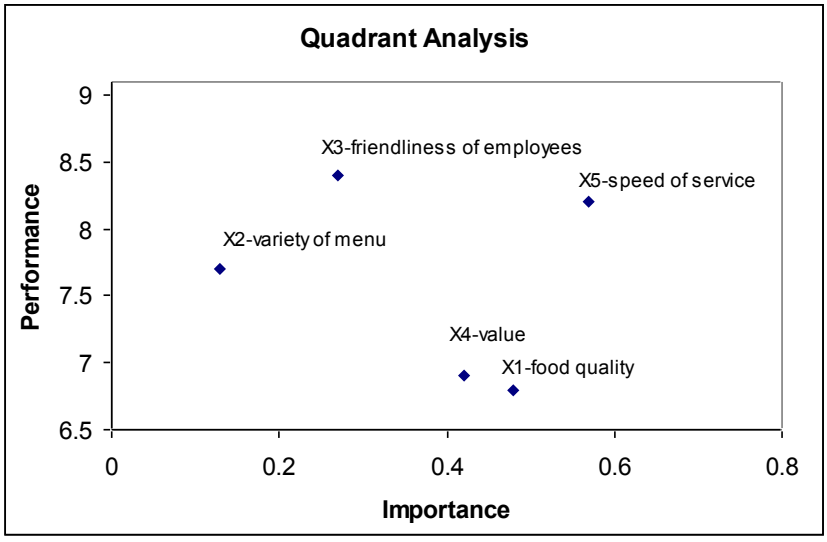
Key Points:

- Pizza Quik wants to know what performance factors define quality for their customers, which of these factors are most important to their customers, and how well Pizza Quik delivers these quality factors.
- Survey data was collected and a regression analysis conducted. The regression equation was:
    - $S = .48X_1 + .13X_2 + .27X_3 + .42X_4 + .57X_5$.
    - Average ratings were S = 7.3, $X_1 = 6.8$, $X_1 = 7.7$, $X_3 = 8.4$, $X_4 = 6.9$, and $X_5 = 8.2$.

○

Questions

1. Plot the importance and performance scores in a matrix. One axis would be importance from low to high, and the other would be performance from low to high.



2. Which quadrant should you pay the most attention to? Why?

In principle, you would attend to all quadrants for a full understanding of your situation. In terms of attractiveness, the upper right quadrant contains the elements that are rated by customers as the most important service factors and where performance is greatest (being good at the right things).

3. Which quadrant or quadrants should you pay the least attention to? Why?

The upper left quadrant contains items that are unimportant to consumers but where performance is high (being good at the wrong things).

4. Based upon your analysis, where would you advise the company to focus its efforts? What is the rationale behind this advice?

Quality of food is the second most important service factor and holds the lowest customer appraisal rating. This would be a good place to start. One problem is that we don't know how much it would cost to improve any of these service factors. What we would want to do is to work

on the factor that would produce the highest return on investment for achieving an improved score.