Multivariate Final Project

Fall 2012 • MKT 7116 • Stephanie Spinella

1. Can we simplify the manager's job? Factor Analysis

Stage 1: Objectives

Upper management has implemented a new emphasis on interactions between customers and employees. They would like to find out if there are broad areas of store operations that can be grouped together for managers to utilize. In the survey taken by 1,405 customers, there are 23 aspects of store operations and policies that are measured. Also included in the survey are an additional 11 questions that deal primarily about employee interactions. The purpose of this analysis is to use an exploratory factor analysis to help identify structures found within a set of variables and examine its reliability. The main objective is to identify broader dimensions of operations. After the analysis is complete, the outcome, which is referred to as the data summarization, is available for use by cluster analysis.

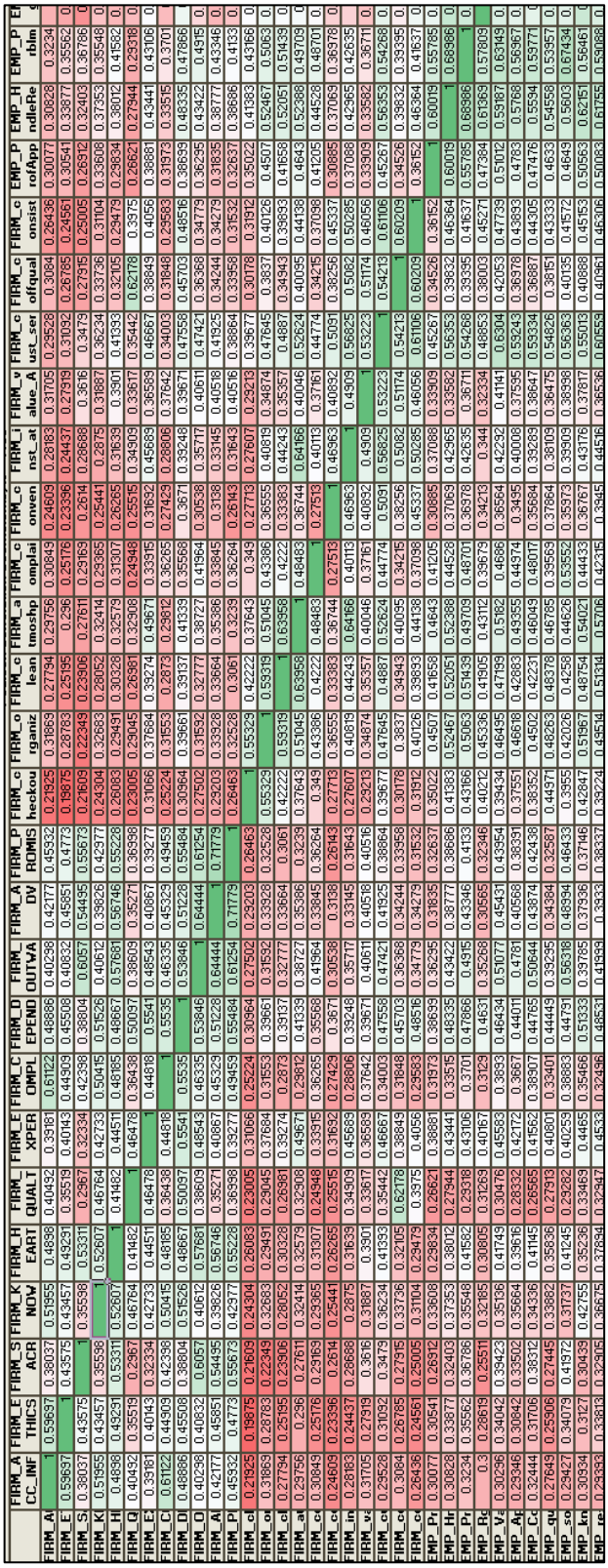
Stage 2: Designing

For this factor analysis, the variables in connection with Firm and Employee survey data are used. The sample size from the survey is 1,405 with 34 variables giving it a ratio of 41:1. That exceeds the minimum requirement of 5:1 for use in factor analysis. Because the analysis will explore the underlying structure of the variables, R-type analysis would be used to understand the structure. It can also be noted that all variables in the data set are metric and range from 1 – 5. Variables used in the analysis deal with Firm and Employee data.

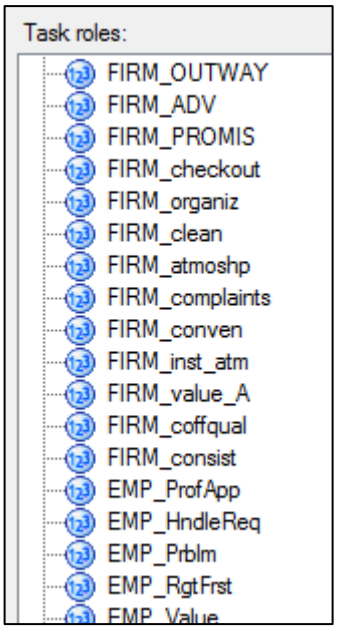
Stage 3: Assumptions

A correlation matrix is performed to gain insight to variable bivariate relationships. It can be seen in the graph below that most variables are correlated, which is logical for the information is already segmented into Firm and Employee groups:

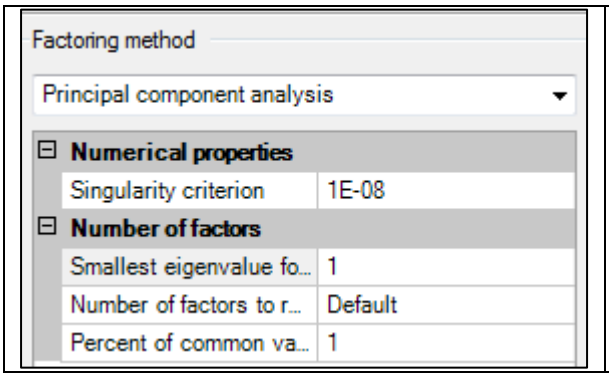
The large amount of correlating variables provides an adequate basis for moving on to the next step of the analysis.

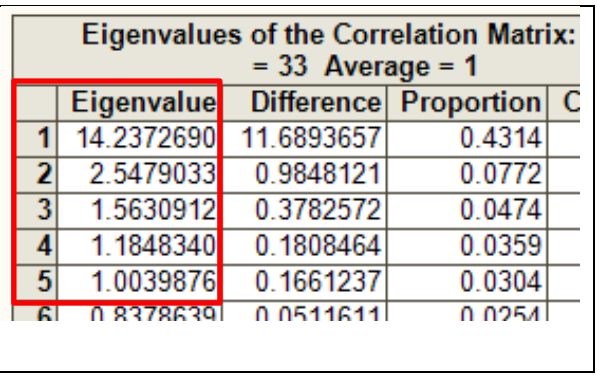


Stage 4: Analysis



Select all Firm and Employee variables as the Task Roles for the factor analysis. After running the factor analysis the first time it was found that Overall MSA = .964 which satisfies the required assumption that MSA should be greater than or equal to 0.5. All variables satisfy the MSA greater than or equal to .50 on an individual level as well which indicates appropriateness of factor analysis.



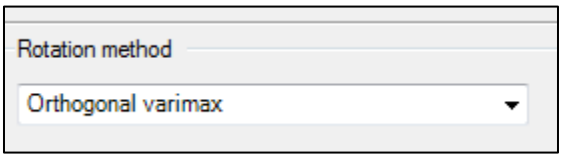


Eigenvalues of the Correlation Matrix: = 33 Average = 1

	Eigenvalue	Difference	Proportion	C
1	14.2372690	11.6893657	0.4314	
2	2.5479033	0.9848121	0.0772	
3	1.5630912	0.3782572	0.0474	
4	1.1848340	0.1808464	0.0359	
5	1.0039876	0.1661237	0.0304	
6	0.8378639	0.0511611	0.0254	

Factoring Method is set as Principal component analysis. Latent root criterion specifies that eigenvalues greater than or equal to 1.0 only be retained. With this criterion in play, five factors will be retained.

The rotation method is set as Orthoginal Varimax. By changing it from its original unrotated state, the variance explained by each factor is more evenly distributed among the factors.



Step 5: Interpreting the Results

The orthogonal varimax will produce rotated factor results which will contain factor loadings for each factor and its corresponding factor loadings, which must be .50 or greater. When first examining the communalities, variable FIRM_cust_serv cross-loaded and will be eliminated from the analysis. Here are the results of the Rotated Factor Pattern after it has been eliminated.

Rotated Factor Pattern					
	Factor1	Factor2	Factor3	Factor4	Factor5
EMP_Approach	0.74858	0.11157	0.26407	0.18495	0.13175
EMP_Value	0.73485	0.11016	0.32244	0.22246	0.15523
EMP_Concern	0.72096	0.09253	0.35753	0.16769	0.12971
EMP_solvprbl	0.69379	0.06771	0.42882	0.17367	0.14318
EMP_respect	0.69272	0.18726	0.137	0.24055	0.26068
EMP_knowdoing	0.67671	0.26583	0.05973	0.22184	0.28409
EMP_HndleReq	0.67185	0.22467	0.12131	0.15781	0.33987
EMP_RgtFrst	0.66278	0.26637	-0.0092	0.19651	0.19966
EMP_Prblm	0.65979	0.17719	0.26092	0.13587	0.31825
EMP_quickeff	0.65865	0.16964	0.10038	0.20268	0.2639
EMP_ProfApp	0.54697	0.23015	0.08661	0.12625	0.33182
FIRM_ACC_INFO	0.08016	0.73843	0.26369	0.06631	0.20154
FIRM_KNOW	0.23236	0.70054	0.15322	0.15825	0.08823
FIRM_ETHICS	0.14344	0.62929	0.34696	0.01362	0.14964
FIRM_COMPL	0.18481	0.61498	0.3541	0.10845	0.14271
FIRM_DEPEND	0.35195	0.57008	0.24858	0.33246	0.10638
FIRM_QUALT	0.13575	0.56879	0.05707	0.55565	-0.03477
FIRM_EXPER	0.29899	0.46099	0.17761	0.3349	0.22546
FIRM_SACR	0.16349	0.24358	0.72205	0.1175	0.06029
FIRM_ADV	0.20667	0.29726	0.71084	0.1694	0.15201
FIRM_OUTWAY	0.32779	0.26614	0.69001	0.19961	0.07294
FIRM_PROMIS	0.19839	0.37627	0.68119	0.13951	0.12038
FIRM_HEART	0.20476	0.49225	0.54281	0.13413	0.09248
FIRM_coffqual	0.25411	0.25788	0.08089	0.76086	0.06981
FIRM_consist	0.3741	0.14773	0.07915	0.67376	0.13672
FIRM_inst_atm	0.18263	0.05941	0.20708	0.62685	0.43528
FIRM_value_A	0.14771	0.10953	0.37645	0.57468	0.24092
FIRM_conven	0.229	0.05313	0.1847	0.51953	0.25897
FIRM_clean	0.32663	0.13494	0.10871	0.20064	0.70249
FIRM_organiz	0.3684	0.22224	0.0526	0.16336	0.66914
FIRM_atmoshp	0.33718	0.12729	0.15053	0.35812	0.62027
FIRM_checkout	0.34683	0.14514	0.07399	0.07486	0.55202
FIRM_complaints	0.37882	0.08441	0.30549	0.17995	0.4203

Factor 1
Perception of Employees

Factor 2
Perception of Products

Factor 3
Perception of Customer Service

Factor 4
Perception of Firm

Factor 5
Perception of Merchandising and Overall Store Appearance

2. What improves customer satisfaction? Multiple Regression

Stage 1: Objectives

The firm is interested in learning more about one of their existing key performance metrics: customer satisfaction. In this assessment, they would like to find out if any of the variables associated with customer perceptions is significant in predicting satisfaction. Upper management would like to create models for each of the categories individual and any composite models developed. Perceptions on Firm, Employee, Customer, CRM program, Customer Buying Behavior and Demographics will be utilized. Because all of the perception variables are metric, a multiple linear regression can be used to predict the dependent variable, customer satisfaction, with the independent variables, the perceptions. The results of this analysis should give upper management a better idea of which variables they need to emphasize to increase satisfaction among their customers.

Stage 2: Research Design

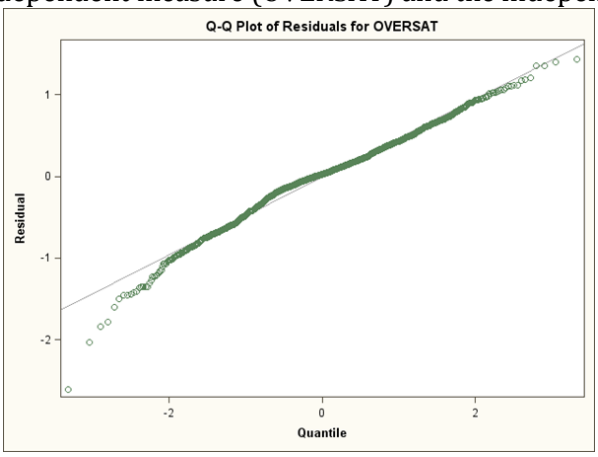
Variables describing customer mindsets about the Firm, Employee, Customer, CRM program, Customer Buying Behavior and Demographics from a survey will be used in this analysis. There are no missing data values present in the set. The preferred ratio for variables to observations is around 15:1 or 20:1 for multiple regression. For this analysis, 1,405 observations and 33 variables gives a 42:1 ratio which satisfies the sample size requirement. All variables are already in a metric format so there is no need for recoding into dummy variables. A baseline was predicted just using the dependent measure of satisfaction and the equation found is $Y = 4.372$. Prediction accuracy is now measured as the degree of improvement compared to the baseline.

Stage 3: Assumptions

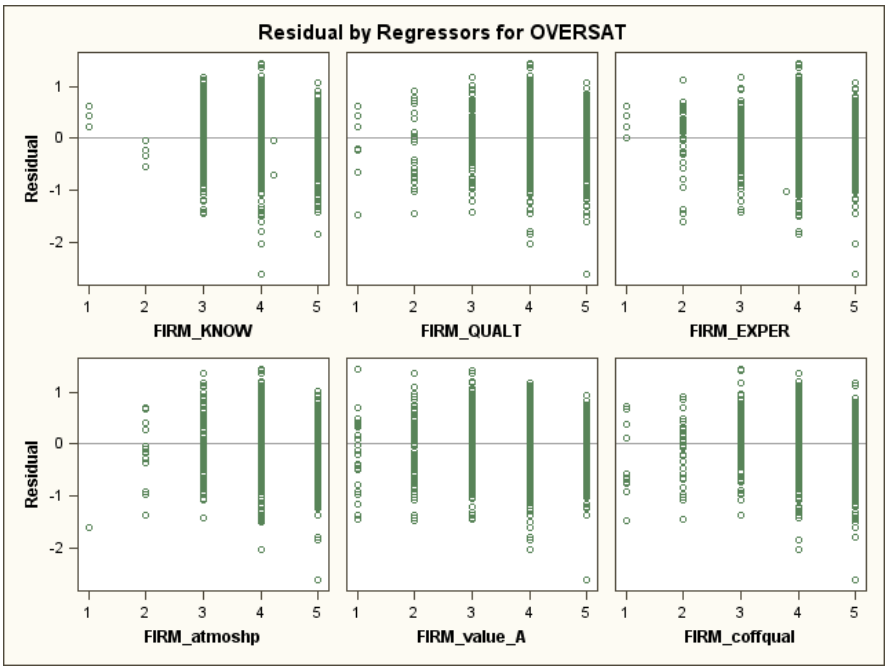
There are four assumptions that must be met when performing multiple regression:

1) Linearity of the phenomenon measured
2) Constant variance of the error terms
3) Independence of the error terms
4) Normality of the error term distribution

The Q-Q plot of Residuals for overall satisfaction is almost linear, indicating there is linearity of the relationship between the dependent measure (OVERSAT) and the independent variables



The Residual Plot shows the residual points for each independent variable and are randomly dispersed around the horizontal axis, which indicates a good fit for a linear model.



Stage 4: Estimating the Regression Model and Assessing Overall Model Fit

A correlation matrix is performed to find the highest correlating bivariate relationships to have a gauge on the existing relationships within the data set.

	OVERSAT
OVERSAT	1.00000
FIRM_coffqual	0.65364
CUST_Recomm	0.57501
FIRM_consist	0.54929
FIRM_cust_serv	0.54418
CUST_SAYPOS	0.52633
CUST_Encourage	0.52171
FIRM_value_A	0.51082
FIRM_inst_atm	0.50942
FIRM_QUALT	0.50784
CUST_Defend	0.50076
FIRM_atmoshp	0.50017
EMP_Value	0.48768
LastTenTimes	0.48491
FIRM_DEPEND	0.47952
FIRM_EXPER	0.47823
EMP_respect	0.47765

EMP_solvprbl	0.45176
EMP_knowdoing	0.45025
EMP_Prblm	0.44626
PurchasePercent	0.44479
FIRM_OUTWAY	0.44356
EMP_HndleReq	0.4434
EMP_Concern	0.43909
EMP_Approach	0.43466
EMP_quickeff	0.42211
FIRM_ADV	0.40481
EMP_RgtFrst	0.40413
FIRM_clean	0.40164
FIRM_organiz	0.39783
FIRM_complaints	0.39235
FIRM_conven	0.39164
FIRM_HEART	0.38574
FIRM_PROMIS	0.38156
CUST_WellInf	0.37538

EMP_ProfApp	0.35765
FIRM_COMPL	0.3412
FIRM_ACC_INFO	0.33894
CommittedCustomer	0.33746
FIRM_KNOW	0.33644
FIRM_SACR	0.32451
FIRM_checkout	0.31511
FIRM_ETHICS	0.31061
CUST_ShareTHGT	0.2894
CUST_Feedbck	0.28719
AgeAsCustomer	0.21828
CRM_Reward	0.21341
CRM_specserv	0.21341
CUST_Sugges	0.19941

CRM_priority	0.17685
TimesPerMonth	0.14772
VisitOthers	0.14741
LCARD_Use	0.07936
AGE	0.06245
GENDER	0.0612
EDUCAT	0.01813
ETHNICITY	-0.06111
OCCUP	-0.07808
SourceInfo	-0.10176
CUST_HappyOCH	-0.12932
CUST_OtherCH	-0.25474
CUST_CompSrv	-0.33981

Now that we have a better idea of the correlations between the independent variables and the dependent variables, a stepwise selection method will be applied in the multiple regression analysis for each group of variables (Firm, Employee, Customer, CRM, Sources of Information about Firm, and Demographics) and also to a composite model consisting of multiple perception variables and of course satisfaction as the dependent measure. Stepwise will be applied because it is a sequential method will automatically select the next added variable as the one with the greatest incremental prediction (partial correlation) in terms of predicting satisfaction and it also removes variables that are non-significant. The following graphs contain information about each stepwise regression that was run for the individual groups of variables.

Perception Group	Variables Used in Equation and Their Reg. Coefficients	Overall Adjusted R^2	Root MSE	F Value	F Significance Level
Firm	Intercept 0.30366 FIRM_QUALT 0.08279 FIRM_EXPER 0.09311 FIRM_OUTWAY 0.0741 FIRM_atmoshp 0.15528 FIRM_value_A 0.08608 FIRM_cust_serv 0.05749 FIRM_coffqual 0.29769 FIRM_consist 0.10333	0.5515	0.4961	216.77	<0.0001
Employee	Intercept 1.22447 EMP_HndleReq 0.0838 EMP_Prblm 0.05895 EMP_RgtFrst 0.06704 EMP_Value 0.16247 EMP_quickeff 0.07358 EMP_solvprbl 0.08288 EMP_knowdoing 0.07809 EMP_respect 0.14189	0.3117	0.6146	80.47	<0.0001

Customer	Intercept 3.20307 CUST_SAYPOS 0.09358 CUST_Defend 0.0936 CUST_CompSrv -0.09475 CUST_Encourage 0.04908 CUST_Sugges -0.05873 CUST_Recomm 0.21599	0.3858	0.5805	147.99	<0.0001
CRM	Intercept 4.0508 CRM_Reward 0.13893	0.0449	0.7240	66.95	<0.0001
Sources of Information about Firm	Intercept 3.31023 AgeAsCustomer 0.07693 ComtdCustomer 0.18313 LastTenTimes 0.07636 PurchasePercnt 0.00275	0.2600	0.6372	124.33	<0.0001
Demographics	Intercept 4.4638 GENDER 0.11113 OCCUP -0.01931 ETHNICITY -0.05419	0.0126	0.7369	6.97	<0.0001

The basis for analyzing model comparison is the adjusted R^2 value from each model. The adjusted R^2 value is used for comparison because the largest R^2 value will be from the model with the largest number of variables whereas the adjusted R^2 value is derived from variances which makes it acceptable to compare models consisting of different numbers of variables. As shown in the chart above, the model with the largest adjusted R^2 value is the one that uses the variables that are concerned with customer mindsets about the firm. Mutually excluding the other variable groups, Firm's R^2 value was 0.554 which means that the variable accounts for 55.4% of the unexplained variance of satisfaction. It also is an indicator of the goodness of fit of the model and how well the regression line approximates the real data pointed. The maximum value of R^2 and Adjusted R^2 is 1.0. The model based on customer perceptions has the second highest Adjusted R^2 of 0.3858 with Employee coming in at a close third with 0.3117.

Now that each group has been given its own regression equation with significant predictors, they will be combined to create a composite model for predicting satisfaction. These are the results of running a stepwise multiple regression analysis using the significant variables found in the previous steps:

Perception Groups Used	ROOT MSE	Overall Adjusted R^2	F Value	F Significance Level
Firm (9) Employee (1) Customer (4) CRM (1) Sources of Info (1) *(Demographics not used)*	0.47333	0.5917	128.17	<0.0001

Variables	Parameter Estimates (b)	Standardized Estimate (Beta)	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	0.742	0.0000	0.117	6.34	<.0001	.	0
FIRM_coffqual	0.244	0.2696	0.024	10.03	<.0001	0.403	2.483
FIRM_atmoshp	0.122	0.1124	0.024	5.06	<.0001	0.590	1.696
FIRM_consist	0.091	0.0909	0.023	3.91	<.0001	0.538	1.859
FIRM_value_A	0.087	0.1143	0.017	5.25	<.0001	0.615	1.627
FIRM_EXPER	0.082	0.0799	0.023	3.52	0.0005	0.564	1.773
CUST_Defend	0.060	0.0936	0.017	3.52	0.0004	0.411	2.431
FIRM_QUALT	0.054	0.0537	0.025	2.18	0.0291	0.480	2.082
EMP_Value	0.047	0.0464	0.023	2.00	0.0456	0.541	1.850
FIRM_OUTWAY	0.045	0.0515	0.022	2.06	0.0391	0.468	2.138
CUST_Recomm	0.038	0.0539	0.021	1.82	0.0696	0.331	3.023
LastTenTimes	0.035	0.1388	0.005	6.66	<.0001	0.669	1.496
CUST_Encourage	0.035	0.0558	0.019	1.85	0.0641	0.320	3.122
CRM_Reward	0.033	0.0510	0.012	2.72	0.0065	0.828	1.207
FIRM_SACR	-0.037	-0.0395	0.021	-1.76	0.078	0.581	1.721
FIRM_KNOW	-0.044	-0.0405	0.023	-1.94	0.0525	0.668	1.496
CUST_Sugges	-0.048	-0.0703	0.014	-3.43	0.0006	0.691	1.448

After figuring out which variables in each perception group were the most statistically significant, they were all ran together in a composite model that was originally comprised of 30 variables. After filtering out the variables that were not significant in predicting satisfaction, the final composite model is comprised of 16 variables that come from the Firm, Employee, Customer, CRM, and Sources of Information about Firm perception groups. The only group not to have a variable in the final model is the Demographics group. This means that the variables such as Gender and Occupation are not significant in predicting satisfaction.

When making conclusions about relative variable importance in the regression equation, the measure that is compared is the Standardized Estimate, which are also called Beta coefficients. These are used because they provide a "scale free" measure of impact. In this model, FIRM_coffqual has the largest Beta value of 0.2696. This variable is the customer's perception of the quality of products and so it can said that when customers believe that the firm provides products are of a high value, the change in the satisfaction will be an increase of 0.2696 due to a change of one standard deviation in FIRM_coffqual.

It is important to test for multicollinearity because while it can be beneficial in revealing suppressor effects, it's generally viewed as harmful because increases in multicollinearity can reduce overall R^2, confound estimation of the regression coefficients, and negatively affect the statistical significance of tests of the coefficients. Diagnosing multicollinearity involves looking at the last two values in the table: Tolerance and Variance Inflation. Tolerance is the amount of variance in an independent variable that is not explained by the other independent variables. Tolerance values of 1.0 indicate

no multicollinearity and as values approach 0 is an indication of greater multicollinearity. The minimum cutoff for tolerance is typically .10. All variables in the model pass this test and the their tolerance values range from 0.32 to 0.828. Variance Inflation Factor (VIF) is the inflation of the variance of the regression coefficients from multicollinearity. VIF values greater than 10 indicate a problem with multicollinearity and VIF values closer to 1 indicate that there is some association between predictor values but generally not enough to cause problems. The great VIF value is this model is 3.122 from variable CUST_Encourage, which was also the variable with the lowest tolerance.

When comparing the final model's results with the individual models, the final composite model improved in prediction accuracy according to the change in Adjusted R^2 and in the Root Mean Square Error (MSE). The final composite model's Adjusted R^2 of 0.5917 was the largest out of all models and the Root MSE was the smallest at 0.47333. These are both an indication of overall goodness of fit of the composite model.

This is the final regression equation that is comprised of the intercept and the regression coefficients is as follows:

Satisfaction = 0.742 + 0.244(FIRM_coffqual) + 0.122 (FIRM_atmoshp) + 0.091 (FIRM_consist) + 0.087 (FIRM_value_A) + 0.082 (FIRM_EXPER) + 0.06 (CUST_Defend) + 0.054 (FIRM_QUALT) + 0.047 (EMP_Value) + 0.045 (FIRM_OUTWAY) + 0.038 (CUST_Recomm) + 0.035 (LastTenTimes) + 0.035 (CUST_Encourage) + 0.033 (CRM_Reward) - 0.037(FIRM_SACR) - 0.044 (FIRM_KNOW) - .048(CUST_Sugges)

3. What makes a customer feel committed? Logistic Regression

Stage 1: Objectives

The firm would like to discover what variables are the best predictor of whether or not a customer feels committed to the firm. When performing the analysis to find these results, two variables that will be emphasized are the length of time that customer has been involved with the firm and whether or not the customer makes use of the company's loyalty program. Predicting commitment can be viewed as a binary variable because customers are either committed or not. This means that a logistic regression will be performed to find out the probability of a customer being a committed customer.

Stage 2: Design and Assumptions

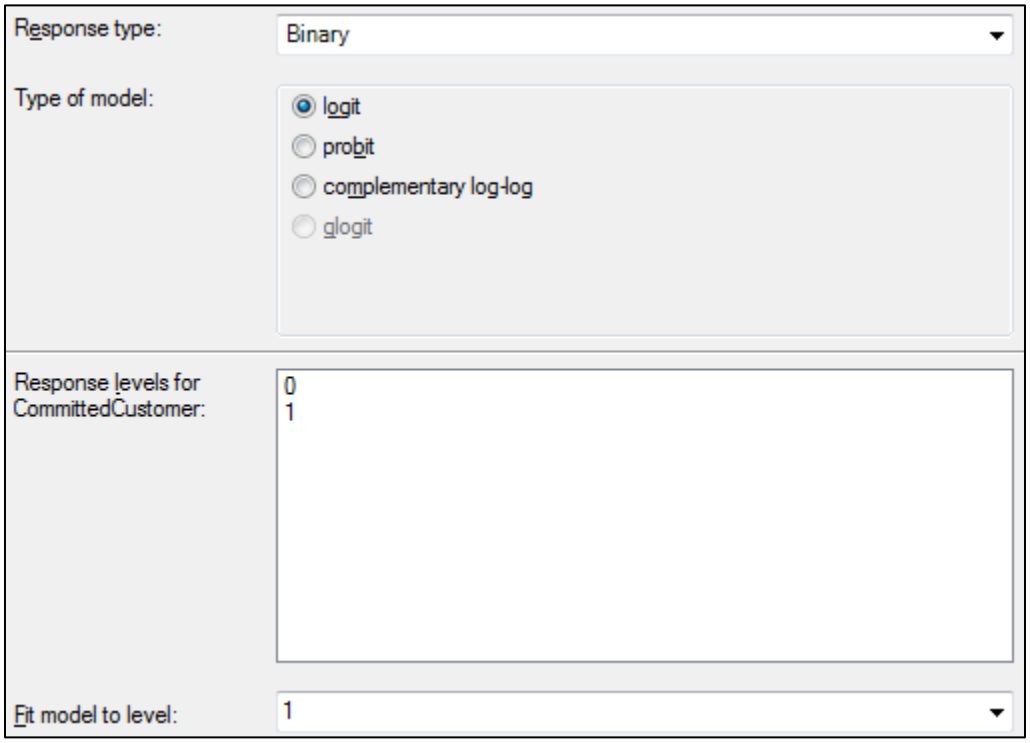
For use in a logistic regression, a sample must contain at least 400 observations and the minimum number of observations per case is 10. The data set satisfies this requirement.

Logistic regression differs from multiple regression in the fact that it does not require any specific distributional forms of the independent variable. Issues such as homoscedasticity don't matter and it also does not require linear relationships between the independent and dependent variables.

Stage 3: Analysis

A stepwise logistic regression is run to find variables that are significant in predicting whether or not customers will be committed. Customer Committed is selected as the dependent variable and variables dealing with Firm, Employee, Customer, CRM, Source of Information about Firm, and Demographics are entered to be run in the analysis.

Response type is set to binary, type of model is set to logit, and the model will be fit to level 1, in other words it will be set to predict which customers will be committed.



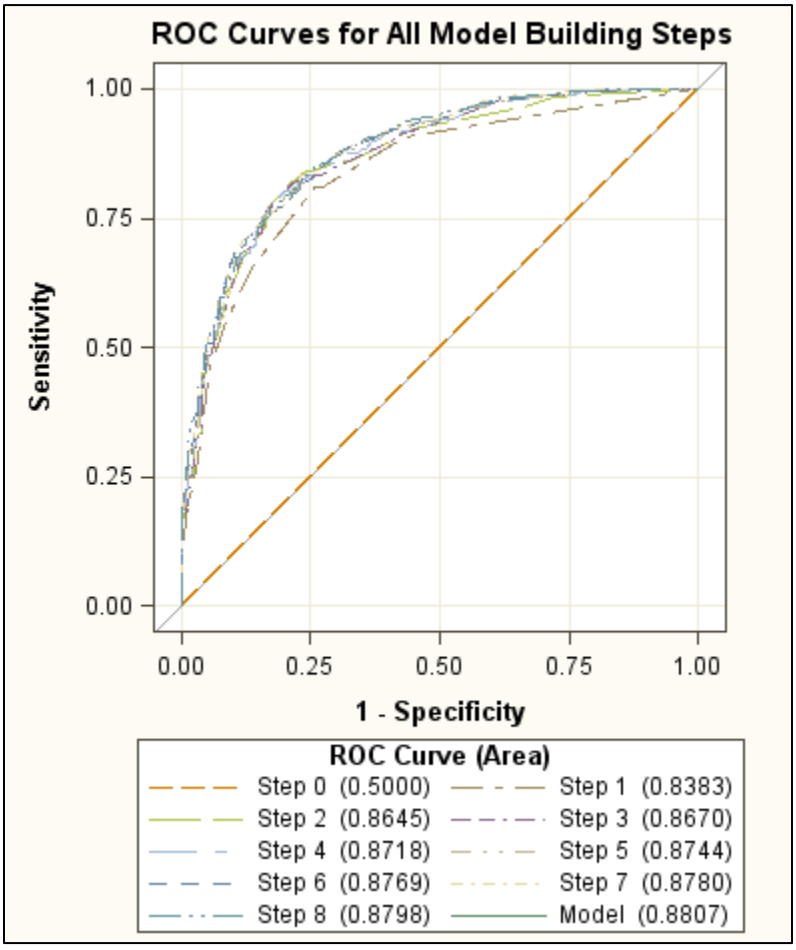
A null model was also created that is solely based on just the dependent variable of customer commitment. The main measure that is analyzed for comparison reason is the likelihood value, which is also referred to as the -2LL value. This likelihood value is a basic measure used to represent the lack of predictive fit and the lower the value, the better the model fit. It can be seen that the -2LL value did decrease from model-to-model (1713.98 → 1090.52). which is an indication of an overall goodness of fit for the model.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1715.979	1105.547
SC	1721.227	1158.025
-2 Log L	1713.979	1085.547

Another resulting table that logistic regression produces is the Deviance/Pearson Goodness of Fit Statistic. Non-significant values in the last column of this table indicates an overall goodness of fit for the model.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	1085.5473	1395	0.7782	1.0000
Pearson	1461.0036	1395	1.0473	0.1069

Both of these measures confirm the use of logistic regression and its overall fit. For a visual representation of overall model fit, an ROC plot is one of the most popular methods. Things to look for in a ROC plot is a high bump a deviance from the diagonal line towards the left.



Logistic regression also has the capability of assign predicted probabilities to customers. Since customer commitment can never truly be predicted, there will be misclassifications and the probability of incorrectly, as well as correctly, assigning the proper commitment prediction to customers can be found in the classification chart that is produced in the results of the logistic regression.

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.300	956	170	250	29	80.1	97.1	40.5	20.7	14.6
0.400	923	225	195	62	81.7	93.7	53.6	17.4	21.6
0.500	887	268	152	98	82.2	90.1	63.8	14.6	26.8
0.600	833	299	121	152	80.6	84.6	71.2	12.7	33.7
0.700	761	338	82	224	78.2	77.3	80.5	9.7	39.9
0.800	653	370	50	332	72.8	66.3	88.1	7.1	47.3
0.900	512	398	22	473	64.8	52.0	94.8	4.1	54.3

Observations with a predicted probability greater than .50 will automatically be assigned 1 and everything else will be assigned 0. In this chart, it can be seen that the sensitivity is rather high for the model which is a good thing because by definition sensitivity is the ability to accurately predict committed customers.

In the Analysis of Maximum Likelihood Estimates, the variables that were chosen to be in the model are displayed and it can be seen that they are all statistically significant. LoyaltyProgram was not found to be significant in this model.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	8.3285	0.8007	108.1866	<.0001
FIRM_consist	1	-0.3228	0.1365	5.5937	0.0180
CUST_SAYPOS	1	-0.5719	0.0939	37.1081	<.0001
CUST_OtherCH	1	0.2340	0.0840	7.7622	0.0053
CUST_Sugges	1	-0.2282	0.0757	9.0935	0.0026
CRM_Reward	1	-0.1499	0.0699	4.6007	0.0320
LastTenTimes	1	-0.3090	0.0546	31.9814	<.0001
PurchasePercent	1	-0.0121	0.00435	7.7663	0.0053
TimesPerMonth	1	-0.0355	0.0101	12.3760	0.0004
VisitOthers	1	-1.0330	0.2496	17.1347	<.0001

4. Are there customer segments? Cluster Analysis

Stage 1: Objectives

Cluster analysis is a group of multivariate techniques whose primary purpose is to group objects based on the characteristics they possess. Differing from other multivariate analyses, clustering is always descriptive and does not involve statistics or inferences. Cluster analysis will always create clusters, regardless if the clusters actually exist or not and the solution cannot be considered generalizable because the solutions are totally dependent upon the variables used in the analysis. The primary goal of cluster analysis is to partition a set of objects into two or more groups based on the similarity of the objects for a set of specified characteristics. For this analysis, the objective is to find if customer segments can be derived and the segments will be divided into three groups based on customer perceptions: store, employee, and customer (themselves). Upper management has made clear that the customer segments should be as equal as possible and that customer satisfaction should be varied from segment to segment.

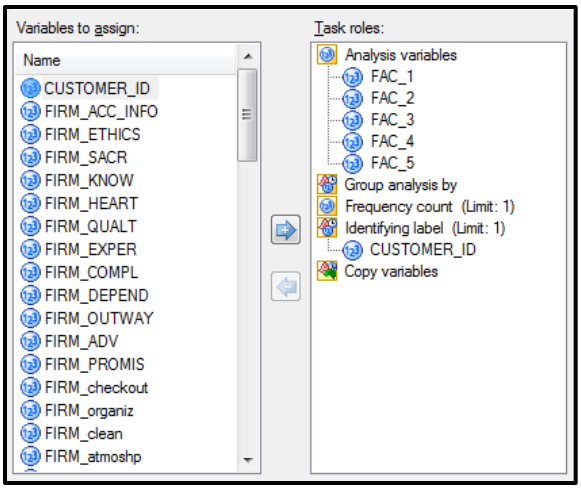
Stage 2: Research and Design

From the factor analysis that was completed to simplify the job of the manager, five different factors were derived from the analysis. These are the five named factors that were derived along with their correlated variables:

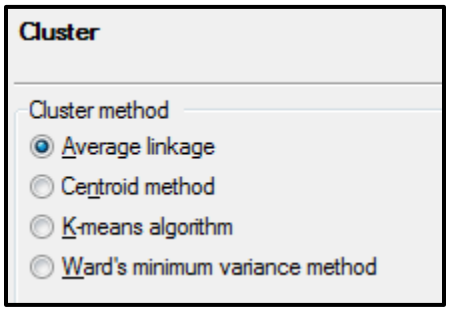
Factor	Variables		
1) Perception of Employees	EMP_APPROACH EMP_VALUE EMP_CONCERN EMP_SOLVPRBL	EMP_RESPECT EMP_KNOWDOING EMP_HNDLEREQ EMP_RGTFRST	EMP_PRBLM EMP_QUICKEFF EMP_PROFAPP
2) Perception of Products	FIRM_ACC_INFO FIRM_KNOW FIRM_ETHICS	FIRM_COMPL FIRM_DEPEND FIRM_QUALT	FIRM_EXPER
3) Perception of Customer Service	FIRM_SACR FIRM_ADV	FIRM_OUTWAY FIRM_PROMIS	FIRM_HEART
4) Perception of Firm	FIRM_COFFQUAL FIRM_CONSIST	FIRM_INST_ATM FIRM_VALUE_A	FIRM_CONVEN
5) Perception of Merchandising and Overall Store Appearance	FIRM_CLEAN FIRM_COMPLAINTS	FIRM_ATMOSHP FIRM_CHECKOUT	FIRM_ORGANIZ

Stage 3: Analysis

After deriving these factors, their summated scales were added into the data set so that clusters can be formed directly from these factors instead of the original variables. The cluster analysis variables will be FAC_1 through FAC_5 and the identifying label will be CUSTOMER_ID.



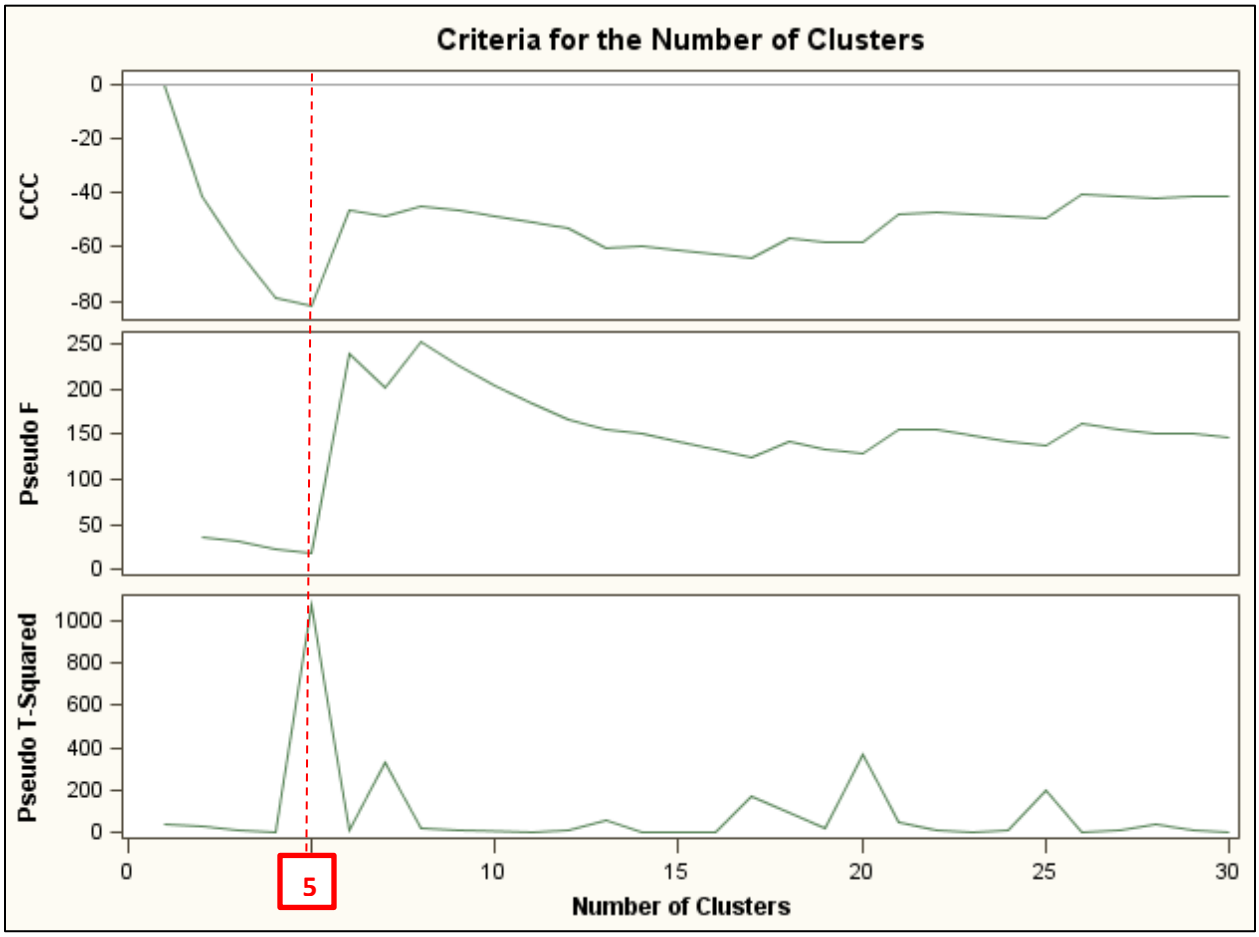
We will use an Average Linkage as the cluster method because it is based on the average similarity of all individuals in a cluster. This method tends to generate clusters with small within-cluster variation and is less affected by outliers.



After running the cluster analysis, the first output examined is the chart of Cluster History. The most noticeable item on this chart is the *Pseudo T-squared* value for when five clusters are formed. When six clusters were joined together this value is 7.8 but when five are formed, this number dramatically increases to 1074. It can also be noted that when five clusters are joined, the R-Square value drops dramatically as well, from .461 to .047. These are both indications that the ideal number of clusters for this data set is five.

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Norm RMS Distance
.....										
7	CL9	CL11	757	0.0952	.464	.747	-49	202	329	1.0253
6	CL7	CL44	760	0.0032	.461	.729	-46	239	7.8	1.0636
5	CL6	CL8	1395	0.4138	.047	.706	-82	17.3	1074	1.1728
4	CL10	CL89	4	0.0019	.045	.673	-79	22.1	4.7	1.2669
3	CL15	CL46	6	0.0032	.042	.622	-61	30.7	8.3	1.4053
2	CL5	CL4	1399	0.0165	.026	.518	-41	36.7	24.1	1.9619
1	CL2	CL3	1405	0.0255	.000	.000	0.00	.	36.7	2.0082

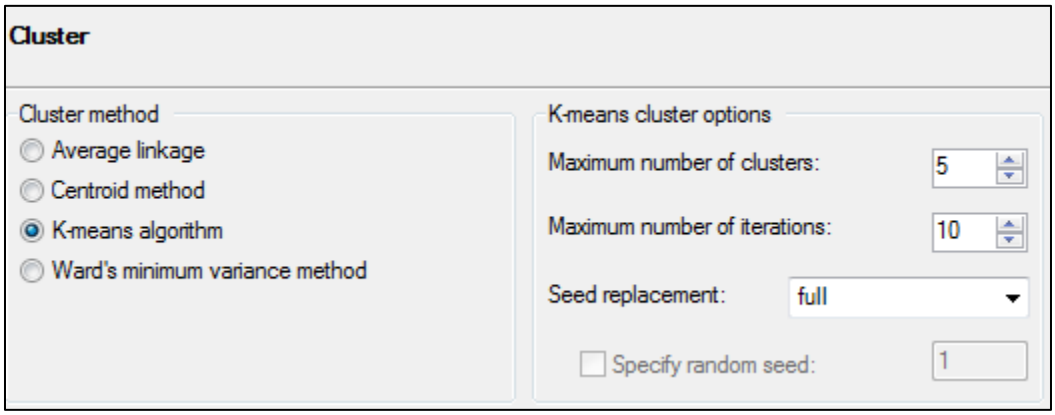
The next items that are examined in the results of the cluster analysis are the CCC, Pseudo F, and Pseudo T-Squared plots. These also give an indication to how many clusters are segmented by the analysis, which in this case is also five.



After running an one-way frequency plot though, an issue is found. There seems to be one dominant cluster and that is not a true reflection of the segments found within the data set.

CLUSNAME	Frequency	Percent	Cumulative Frequency	Cumulative Percent
CL10	2	0.14	2	0.14
CL15	4	0.28	6	0.43
CL46	2	0.14	8	0.57
CL5	1395	99.29	1403	99.86
CL89	2	0.14	1405	100.00

We will now run a cluster analysis done with K-means, which work by portioning the data into user-specified number of clusters and then iteratively reassigning observations to clusters until some numerical criterion is met. The same variables are used but the cluster method will now be set to K-means algorithm. After selecting this method, the K-means cluster options become available and for these we will select 5 for the maximum number of clusters and 10 for the maximum number of iterations. Seed replacement option stays at "full."



The following table is the result of running the cluster analysis with the cluster method set to K-means algorithm.

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	4	0.6536	1.6680		5	3.0215
2	199	0.4123	2.4482		4	1.2170
3	386	0.3232	1.5423		5	1.2160
4	411	0.3266	1.8809		5	0.8989
5	405	0.3441	1.5984		4	0.8989

An one-way ANOVA is run between the clusters and Overall Satisfaction to check for differences among clusters.

Cluster	Mean of OVERSAT
1	5
2	3
3	5
4	4
5	5