20 November 2016

QUALITY & QUANTITY OF MEMBERSHIP

MSIS 5223: Deliverable 2

Jaymie Jordan – CWID 10584997

Amber Johnson – CWID 20054679

Niranjan Dakshinamurthy – CWID 11772439

Robert Fritts – CWID 11663382

TABLE OF CONTENTS

EXECUTIVE SUMMARY

According to the American Psychological Association, adolescents' cognitive development is the groundwork of moral reasoning, honesty, and pro-social behaviors such as helping and volunteerism.[1] DeMolay International is a contributable source to aid in such development. A fraternal youth leadership organization, DeMolay International practices core values including personal development, leadership, and community service as well as teaches leadership, public speaking and civic awareness.[2] Though the organization is operating sufficiently, it seeks competitive advantage for enhancing overall membership.

For organizational development, the addressed opportunity is to provide applicable suggestions for increasing member retention, member recruitment, and quality of membership at DeMolay International, Stillwater Chapter, further expanding the Chapter's community impact through membership growth. Key findings may be applicable to other chapters as well, though that is beyond the scope of this analysis.

We anticipate the conducted analysis will identify key predictors of member participation, tenure, and satisfaction based on models, which are developed to influence these drivers and thus impact membership retention and community impact. Key factors are likely to be advisor recruitment and participation of individuals between the ages of 12 - 14 years old. Additionally, most new members join during the summer when the majority of the fun Chapter activities are held, so the expectation is to identify a positive relationship

[1] https://www.apa.org/pi/families/resources/develop.pdf

[2] https://demolay.org

between participation and summer months' activities.

STATEMENT OF SCOPE

The scope of this project is to benefit DeMolay International, Stillwater Chapter, by providing context around their current trends of membership retention, membership recruitment, and membership quality. In combination with supportive research, the intention is for the analysis to contribute to the Chapter's development by accomplishing these three objectives:

- To assess the correlation between the time of the year and Chapter events for optimal participation.
- To correlate the geographic and demographic data of involved individuals and length of membership.
- To describe the relationship between member level, age, and perceptions of the Chapter.

The sample data contains quantitative collections from the Stillwater Chapter throughout a time period of 56 months. Based on the information provided by the sample data, the analysis is generalized for men between the ages of 12 and 21 in Stillwater. In summary, the dataset contains information pertaining to active membership, attendance percentages, and standard Chapter activities. Based on the dataset, the target variables are address data, member level, Chapter expectations and organization quality, and both target variables are predicted by the same variables. The dataset contains several predictor variables, which are categorized in the following seven categories:

Category	Areas Bundled into Category
Membership	Advisor Council Certification
	Advisor Recruitment
	Fundraising
	Founder's Award
	Blue Honor Key
	New Membership
Awards	LCC 1 & 2
	LCC 3 & 4 & 5
	Obligatory Days
	PMC-MSA
	RD
Fun	Athletic Activities
	DeMolay Month Major Activities
	Award of Excellence
	LTC
	Miscellaneous Activity
	Social Events
Ritual	Advisor Participation
	Initiation, DeM, 4th
	Open/Close 1 Ritual
	Individual Qualification
	FT, COL, Mag 7 Qualification
	Performance
	Ritual Tournament Attendance
	Competing
	CDJ/HDJ
Travel	MC, SC, JC (Workshop)

	DeMolays and Sweethearts
	Advisors & Parents
	District Meetings
	Mileage
	Visitation
	MC, SC, JC Conclave
	DeMolays and Sweethearts (Conclave)
	DeMolays and Sweethearts (Conclave)
Education	Annual Financial Report
	Website
	Newspaper
	Television
	Radio Ad
	Billboard
	Chapter Newsletter
	Chapter Checklist
Community Service	Charitable Points
	DeMolay Service Activities
	Masonic/Civic Service

Figure 1 – Categorized Predictor variables

PROJECT SCHEDULE

The project concludes on December 5, 2016 with about 150 hours of time dedicated between four team members for the duration of the project—approximately 98 days. Though the group's initial time estimation was incorrect by nearly 60 hours less, the GANNT chart of deadlines and deliverables remain consistent. The increase is project hours was caused by the misinterpretation of time needed to completed the necessary data preparation and modeling techniques. Bi-weekly meetings are scheduled with weekly status update

emails, and the workload of the project is equally allocated across the team. Holidays are treated equal—depending on the project progress while determine if work is required on those days. Detailed project tasks and milestones with the assignee is documented in Figure 2 - GANTT chart below:

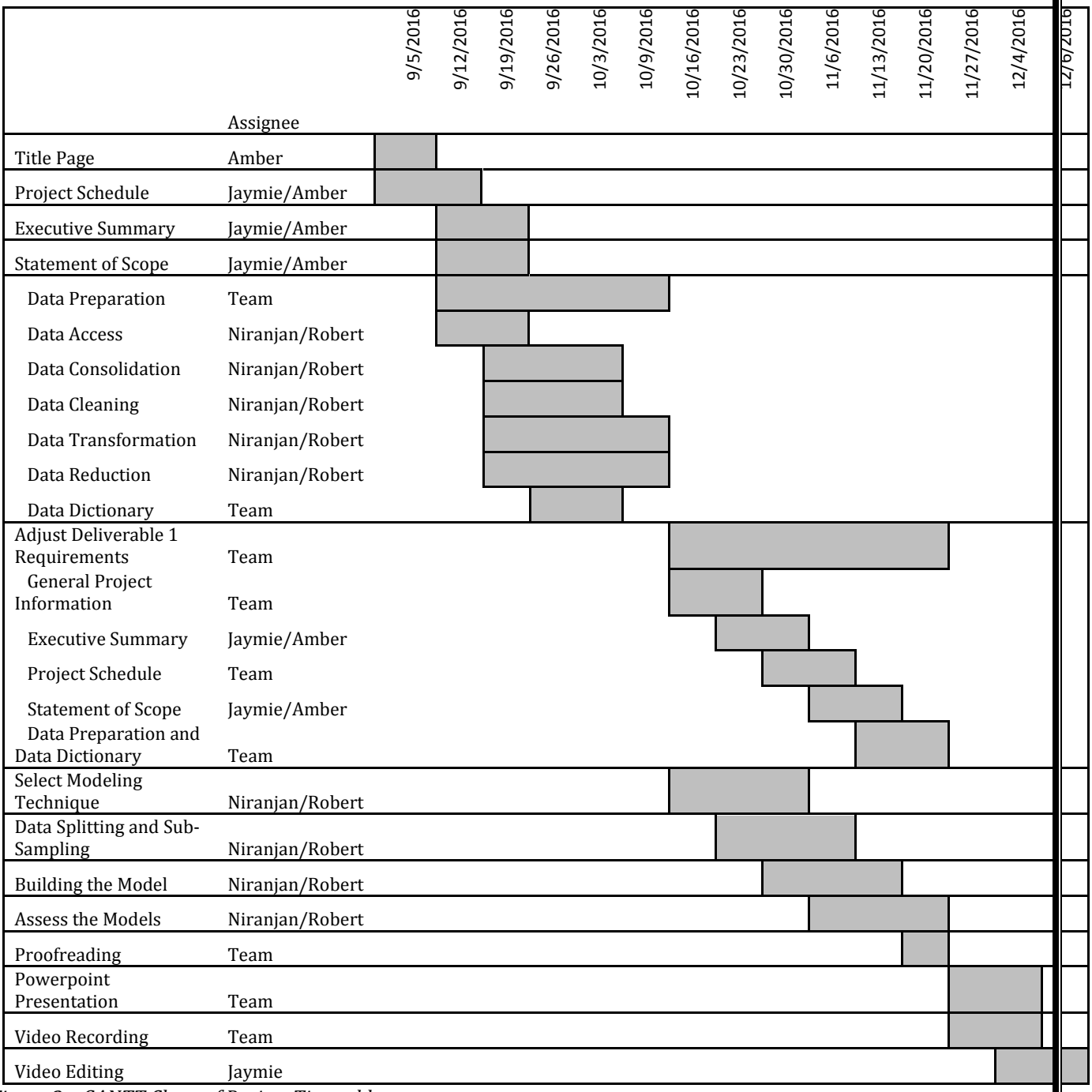


Figure 2 – GANTT Chart of Project Timetable

DATA PREPARATION

The data utilized from the DeMolay Stillwater Chapter is not publicly available. The data was collected in a manner that required minimal cleaning, and cleaning mechanisms occurred during the access portion that otherwise would have to be handled in the cleaning phase.

DATA ACCESS

Data for this project is gathered from internal and external sources of DeMolay Stillwater Chapter. The internal data is provided by the DeMolay Stillwater Chapter and is gathered and maintained by team member Jaymie Jordan in accordance to his role within the organization to maintain information about members and participation.

Objective 1: correlation between time of year and chapter events, relied on the data file "DeMolay.CSV", and it was sourced from the Stillwater Chapter. The data file is 8.11 KB and contains 56 records with 50 attributes. All of the information contained in the "DeMolay.CSV" file is used for the project.

For objective 2: geographic and demographic relationships, the data file is .CSV file titled "Full Demographics 10_1_2016 v2.csv" from the Stillwater Chapter is applied. The external source for objective 2 includes the census government website and Bing maps:

https://geocoding.geo.census.gov/geocoder/geographies/addressbatch?form

https://geocoding.geo.census.gov/geocoder/geographies/onelineaddress

http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?ref=addr&refresh=t#

https://www.bing.com/mapspreview

The Demographics File represents member residence location data (address) and target variables Active Level, Tenure_Yrs and Mem_Attend_Pct in addition to other fields such as date of birth and initiation date that lead to the *_Yrs column values. The demographics information was selected based on its potential to offer insight on the influence of members' locations on the specified target variables – Active Level (active/inactive), Member Attendance Percentage and Tenure (years, decimal).

Additionally, this data contains several census-derived columns, meaning the addresses were referenced to determine latitude, longitude census tract, and census block IDs. The Demographics File consists of 81 records, and the file is in the format of a 16KB CSV file. For this project, the entire data set will be utilized.

With regards to objective 3: relationship between member level, age and perceptions of the Chapter, data gathered through a survey completed by Chapter members is the primary source. The data file is named "Three Week Survey Results.CSV", and the file represents survey results taken from three different polls of members across three weeks. The document contains 57 records in a 5KB CSV file, and the questions gauge members' perceptions of organizational aspects and Chapter activities in addition to their overall perception of the Chapter. Questions are based on a five point Likert scale and a ranking question. The entire data set is used in our project.

The survey data set was solicited, compiled and provided by team member Jaymie Jordan. The original data was obtained through a paper survey completed by Chapter

members, then converted into useable data and combined into the final "Three Week Survey Results.CSV" file.

Data Consolidation

The demographics data was consolidated during the cleaning process due to the amount of manual work required. The survey data file required no consolidation beyond what had been done upstream from the team's analysis, as the file was solicited by and is maintained by the Chapter. Similar to the survey data file, the DeMolay file also required no consolidation beyond what was originally collected and maintained by the Chapter.

Data Cleaning

Regarding the DeMolay.CSV file, data cleaning is unnecessary. The condition of the data is in a state that is appropriate for current analyses. On the other hand, specific to objective 2, extensive manual cleaning was necessary, so it was completed simultaneously with the data consolidation. Majority of the census data (60 percent) was returned through the batch process on the census government website, and manually merged with address data. More than half of the data (40 percent) required manual look up, and was created and cleaned with all remaining data. Overall, the manually compiled data source is a mix of the single line lookup and individual map exploration/inspection. Every record with missing census location information was manually inspected and corrected. A few remain null as a result of the input data and no further cleaning is possible to remove these nulls short of getting new source data, which is not necessary at this time for analysis to continue.

Regarding objective 3 and the data file "Three Week Survey Results.CSV", the survey data file contains no data that needs to be scrubbed based on inspection. However, one field from one survey is missing from the dataset, and this is attributed to missing data—an equivalent "N/A"—as that survey question relates to travel though not all members have experienced travel activities. The survey did not contain a "N/A" answer option, but future surveys will contain this answer choice. Additionally, the missing record will be removed as necessary for reduction analysis, so that every column can be assessed for necessity. The missing field will not be removed from the set for application of the results, understanding it may present itself as an outlier and further adjustment might be required.

DATA TRANSFORMATION

Firstly, Objective 1: correlation between the time of year and Chapter events and data file "DeMolay.CSV", the data contains 50 columns out of which require segregated columns into categorical data and continuous data. Purpose of the segregation: due to the columns resembling points awarded to each individual when an event was completed. These events are related to advisor activities, fun activities, education activities, travel activities, awards, rituals and workshops. The events are ordinal data types with points assigned, for example 25, 50, 75, 100, etc., or in interval counts of 10, for instance 10, 20, 30, etc. Hence, it is important to transform these columns into categorical data.

In addition to the categorical columns, the other fields are related to Date information and integer numeric. These fields were converted to factor data types for dates and to integer for numeric columns.

Pertaining to objective 2 and the data file "Full Demographics 10_1_2016 v2.CSV", transformations to the data frame for location data are expected to be minimal. No transformations have been done up to this point, but are expected for modeling purposes. The expected transformations are converting "Active Level" column to a binary 1/0 value from the current 1/4 value system that is used to denote active/inactive member status.

The "Full Demographics 10_1_2016 v2.CSV" file contains two columns that were constructed during the data access portion of acquiring the data that could be considered transformations of the "Init Date" and "DOB" (date of birth) fields. The derived columns are Tenure_Yrs and Age_Yrs, respectively, and were created in excel while manually cleaning the census data that consolidated the location data. Column values were created through the use of a formula. The value is a decimal representing age and tenure in years (decimal) as the difference in years between the relevant dates and the current date (of analysis). The derivation provides a continuous variable for analysis. Currently, no end date is determined for inactive members; therefore, tenure is open ended. If end dates become available we expect to include the information and complete another transformation such that inactive members' tenure is set and is formatted by the years (decimal) between Init and End dates.

With regards to objective 3 and the data file "Three Week Survey Results.CSV", relatively few transformations are required. The first three columns are date, age, level, and updated the data type to factor (categorical) during the data import, since some of the fields could be denoted as integers. One transformation has been required, and it included the creation of dummy variables corresponding to survey question 25, which pertains to ranking seven components. For each element a corresponding dummy variable was constructed

testing whether it was ranked the value of 1 in order to create a tally of how many components were ranked the highest. The transformation was performed for two effects: data reduction and construct an extra column for testing target variables' dependence. Additional transformations will be detailed as we progress throughout the data interpretation. In addition to the aforementioned transformations, it was determined through the modeling phase for objective three that additional transformations were indeed necessary due to the use of categorical variables and the use of regression as a model. These transformations are similar to the dummy variables mentioned above and were used to create indicator fields for both age (range 12:16,18:20) and membership level (range 1:4) as these are discrete categorical fields with no other acceptable values. The transformations took on typical dummy variable names such as 'age_12', 'age_13', 'lev_1', 'lev_2' and so on. These were specifically required due to the specific objective of understanding the relationship if any between age and membership level on perceptions of the chapter and organization.

DATA REDUCTION

Regarding objective 1 and the data file "DeMolay.CSV", Principal Component Analysis (PCA) was first completed for data reduction. PCA was performed separately for the categorical data and the continuous data. Following, Multiple Correspondence analysis (MCA) was completed since it is more suited for analyzing data that contains several categorical variables than PCA. The analysis was done using R code with packages FactoMineR for computing MCA and Factoextra for visualization.

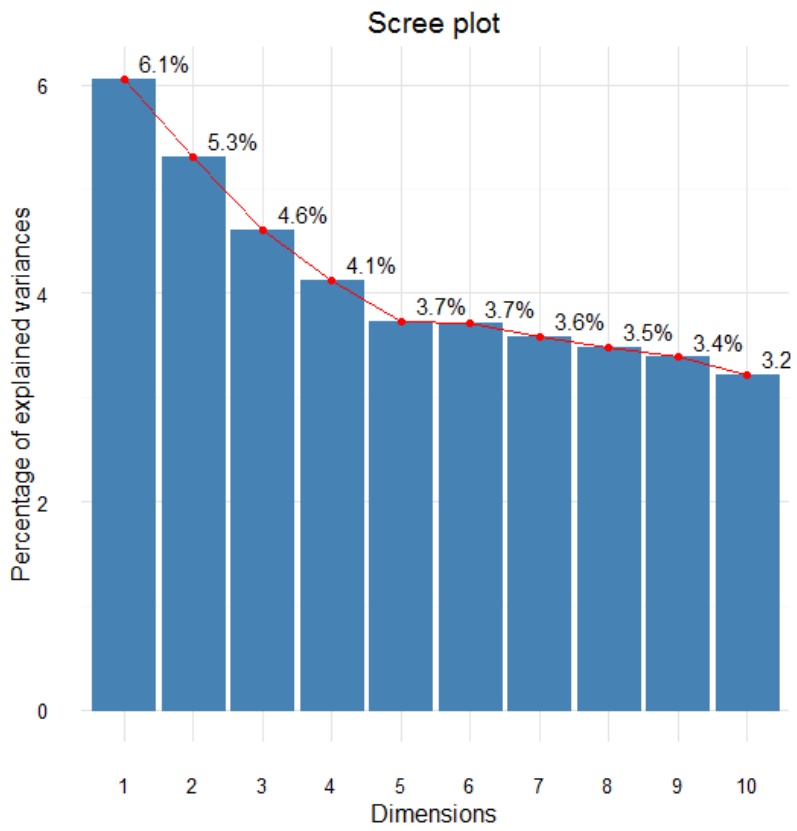


Figure 3 – Scree Plot

Demonstrated by Figure 3, 6.1 percent of the total variance is explained, 5.3 percent by the second dimension, 4.6 percent by the third dimension, which accounts for only 16 percent of the total variance—which is quite low. Furthermore, there is also no drastic decrease in the percentage of variation as seen from the scree plot. This indicates no cut off in considering dimensions and not considering other dimension. Based on this information, the conclusion is that more dimensions will have to be considered for our analysis.

```
$`Dim 1`$quali
                                R2      p.value
trav_District              0.8679236 5.033481e-24
mem_BHK                    0.8655210 8.116032e-24
awards_Obligatory          0.8732531 3.102146e-22
rit_Qual_Ind               0.8800333 2.978239e-18
rit_Advisor_Participation  0.9210959 4.019117e-18
comserv_Masonic            0.8761166 1.228499e-15
fun_March_Activities       0.6155547 7.437036e-11
awards_LCC12               0.5784581 7.922666e-10
rit_Meeting_1Rit           0.5759310 9.235685e-10
rit_Performance            0.6277622 9.704611e-10
edu_Television             0.4665856 5.850805e-08
edu_Radio_Ad               0.3659180 9.307811e-05
mem_New_Members            0.4703764 1.077160e-04
fun_Athletics              0.3591760 1.198642e-04
rit_Meeting_NoRit          0.2839661 1.431331e-04
rit_Qual_Public            0.4605168 3.752885e-04
edu_Billboard              0.2843776 9.146942e-03
```

Figure 4 – MCA

Referencing Figure 4, the table shows that rit_Advisor_Participation has a high R-square value that is linked to dimension1 and can be used to interpret dimension1 rather than the other variables. Further, it is to be seen that the P-value increases down the table. For Figure 5 and Figure 6, rit_Advisor_Participation, has the highest R2 factor and can be used to interpret the respective dimensions.

```
$`Dim 2`
$`Dim 2`$quali
                                R2      p.value
rit_Advisor_Participation 0.8899280 3.040610e-15
rit_Qual_Public           0.7546639 2.503636e-11
awards_Obligatory         0.6544523 3.017488e-11
fun_March_Activities      0.5270610 1.510213e-08
comserv_Masonic           0.7097669 4.055767e-08
rit_Competing             0.4169994 7.699639e-08
rit_Tournament            0.4080423 1.173828e-07
rit_Performance           0.4608011 6.652078e-06
edu_Newspaper             0.4374280 5.414964e-05
awards_LCC345             0.6050608 8.325098e-05
rit_Qual_Ind              0.4742383 2.277467e-04
mem_New_Members           0.4284733 5.003608e-04
comserv_DeMolay_Service   0.2328610 8.895355e-04
edu_Chapter_Newsletter    0.1932102 3.381823e-03
edu_Radio_Ad              0.2372770 7.056312e-03
edu_Website               0.2565061 9.352342e-03
fun_Social                0.3028190 2.133936e-02
fun_Athletics             0.1979207 2.183131e-02
edu_Billboard             0.2292447 3.906618e-02
trav_Visitation           0.2257232 4.256936e-02
trav_Conclave_Councilors  0.1110892 4.413081e-02
```

Figure 5 – MCA

```
$`Dim 3`$quali
                                R2      p.value
trav_Conclave_Councilors  0.4724388 4.367471e-08
trav_Conclave_DeMolays    0.5191920 4.469606e-07
trav_Conclave_Advisors    0.5191920 4.469606e-07
fun_LTC                   0.5202675 1.503251e-06
awards_Obligatory         0.4159656 1.285002e-05
rit_Advisor_Participation 0.6155364 5.272395e-05
fun_Athletics             0.3639395 1.002820e-04
fun_Social                0.4312586 4.539767e-04
edu_Billboard             0.3600633 8.955592e-04
rit_Competing             0.1490107 3.299480e-03
edu_Annual_Financial      0.1929383 3.412156e-03
rit_Tournament            0.1461688 3.639268e-03
edu_Radio_Ad              0.2586954 3.677417e-03
rit_Teams_Assist          0.1409732 4.351675e-03
edu_Newspaper             0.2920447 7.356864e-03
trav_Visitation           0.2911262 7.552867e-03
rit_Performance           0.2603451 8.376122e-03
rit_Qual_Public           0.3521152 1.097984e-02
fun_March_Activities      0.1688528 2.122209e-02
comserv_DeMolay_Service   0.1207886 3.299686e-02
rit_Meeting_NoRit         0.1205294 3.325565e-02
rit_Qual_Ind              0.2977443 4.229878e-02
```

Figure 6 - MCA

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4
Standard deviation     1.2938098 1.0568295 0.9046922 0.6250597
Proportion of Variance 0.4184859 0.2792222 0.2046170 0.0976749
Cumulative Proportion  0.4184859 0.6977081 0.9023251 1.0000000
```

Figure 7 – Continuous Data

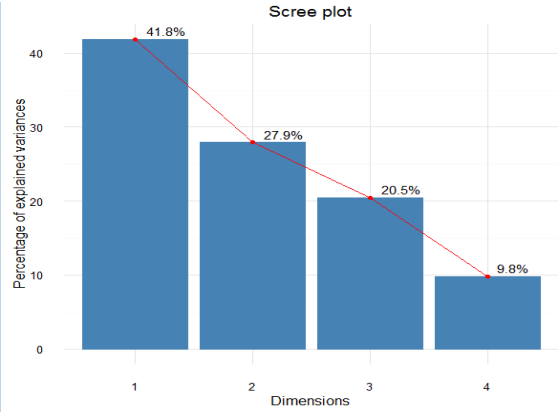


Figure 8 – Continuous Data Scree Plot

According to Figure 7 and Figure 8, it is seen that component 1 and component 2 have Eigen values greater than 1 and, hence, these components will be retained and components 3 and 4 will be rejected. Further component 1 and component2 account for 69.77 percent of the variance.

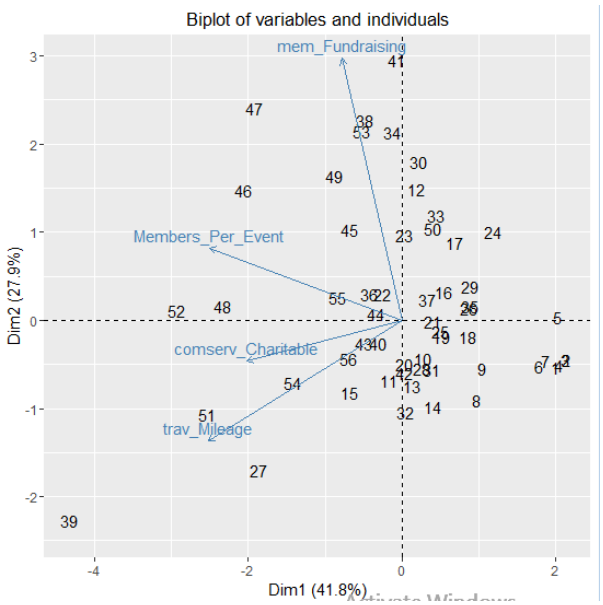


Figure 9 – Bi-plot

Figure 9 provides insight about the variables and observations. The cosine between the vectors is the correlation between the variables. For example, variable trav_Mileage and mem_Fundraising are orthogonal to each other. Comserv_Charitable are almost correlated to each other. While considering dimension1, observations 39, 51, 52 and 48 are in contrast with observations 5, 6 and 7. While considering dimension2, observations 39 and 27 are in contrast with observations 47 and 41.

```
Loadings:
                   Comp.1 Comp.2 Comp.3 Comp.4
Members_Per_Event  -0.602  0.241  0.489  0.584
trav_Mileage       -0.605 -0.402  0.228 -0.648
comserv_Charitable -0.486 -0.137 -0.827  0.248
mem_Fundraising    -0.188  0.872 -0.160 -0.422

                 Comp.1 Comp.2 Comp.3 Comp.4
SS loadings        1.00   1.00   1.00   1.00
Proportion Var     0.25   0.25   0.25   0.25
Cumulative Var     0.25   0.50   0.75   1.00
```

Figure 9 – Loadings Table

Figure 9 shows the level of loading of each variable on different components. For example, trav_Mileage is highly negatively correlated to component 1 with a value of -0.605. Mem_Fundraising is highly positively correlated to component 2 with a value of 0.872.

```
Uniquenesses:
 Members_Per_Event       trav_Mileage comserv_Charitable    mem_Fundraising
             0.802              0.005              0.886              0.985

Loadings:
                   Factor1
Members_Per_Event   0.445
trav_Mileage        0.997
comserv_Charitable  0.338
mem_Fundraising    -0.123
```

Figure 10 – Factor Analysis

Factor analysis is similar to PCA; factor analysis is used to analyze the uniqueness of the factors where as PCA is used to find the variation among factors. Looking at the Uniqueness of variables in Figure 10, trav_Mileage shows a very low value of 0.005 indicating high variance. Members_Per_Event,comserv-Charitable and mem_Fundraising has a high value of uniqueness or low variance. Factor 1 has inverse loading on mem_Funderaising and high loading on trav_Mileage. Loading helps to use those variables in the factors. The higher the

loading, the better the variable is to be included in the Factor. After analysis on categorical and continuous data performed separately, no variable is removed.

Pertaining to objective 2 and the data file "Full Demographics 10_1_2016 v2.CSV", the final data file after obtaining, creating, consolidating and cleaning requires no additional reduction; since there are two categorical predictors denoting census tract and block constructs, and two continuous predictors that cannot be separated due to their nature of being a set (latitude and longitude). The predictors and targets were chosen by inspection. There are a number of other columns that are in the file and led to the final predictors being derived, but these columns are not considered part of the pertinent data and will not be loaded into an analysis frame. Though, these columns are preserved in the access file for tracing back any issues in the data and for further understanding in case of error as all of the predictor variables are derived from the address data itself. Additionally, state and county codes are preserved for similar reasons and are not to be examined as predictors because they are invariant across the relevant records.

While no column reduction was required, there will be records that need to be removed for analysis. Five records have null data for location information in the predictor columns as a result of bad address (one case) and the address on record is a P.O. Box (four cases). As a result, no location information is available for these members, so they cannot be analyzed for variance on the null values. If null values are possible to be accounted for in modeling techniques, the null cases would be treated as outliers or adding uncertainty. Lastly, the unique ID column contains a duplicate entry. Upon inquiry, the original data source confirmed that the older of the two records is more accurate and the other is an error

and can be removed. Since a member cannot have two different dates of birth, per confirmation from the data source, the younger (more recent) DOB record will be removed from the data frame for analysis. As with the columns, despite the insignificance of these records, they will be preserved within the data file but not loaded into an analysis frame.

Regarding objective 3 and the data file "Three Week Survey Results. CSV", standard data reduction techniques of Principal Components Analysis (PCA) and Factor Analysis (FA) were applied to the survey dataset following cleaning and transformation. Reduction was performed on a subset of questions (Q1-Q24) that were on a 5-point Likert scale. Then reduction was completed separately on the Question 25 components ranked on a 7-point scale. The two reductions were completed separately to resolve issues around standardization or scaling if columns were to be averaged as a result of the analysis. Random sampling to a size of 50 percent of the dataset (28 records of 57 total) created two distinct splits with different variance. Such random sampling is a requirement for PCA and FA to avoid bias of having the same set used to validate itself. Finally, reducing the ranking data (question 25) considered the sum total of each of the native columns as well as each of the associated dummy variables, and compared across variable types and sample sets.

Principal Components Analysis resulted with recommending seven overall factors for the set of columns questions 1 through 24. However, for the ranking question set, comprised of seven components and seven paired dummy variables, PCA suggested required six factors. The number of columns is determined by such factors returned with Eigen values greater than or equal to one. The PCA results are shown below in their respective scree plots:

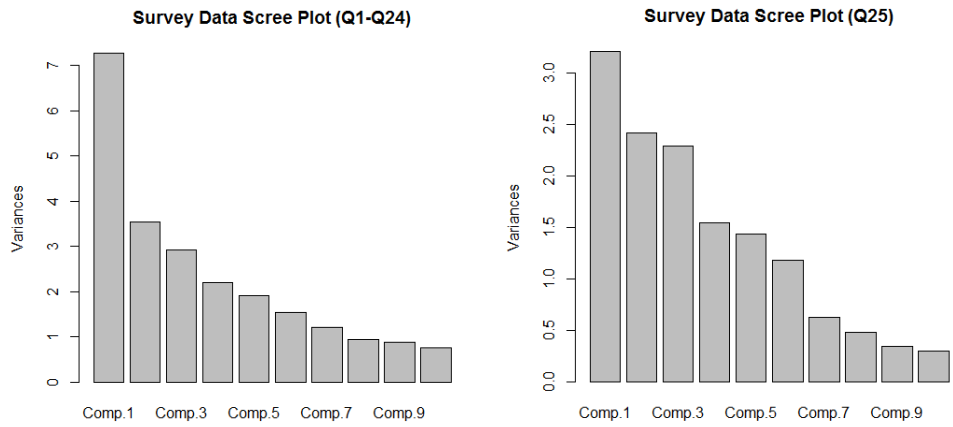


Figure 11 – PCA Results

Factor Analysis was insufficient for the objective due to the inability to produce a result. In each case – reviewing split 2 for questions 1-24 and split 2 for question 25 – the FA technique produced the following error:

> *Error in solve.default(cv) : system is computationally singular: reciprocal condition number = 8.6816e-18*

Typically, the above error occurs when too many columns are correlating with each other (covariant), and those columns can be removed to proceed with the FA. This is typically done through a study of the pairs() output, plotting each variable against each other variable in a grid to spot correlated pairs. Of course, with 20+ independent variables, that is of little use as seen below:

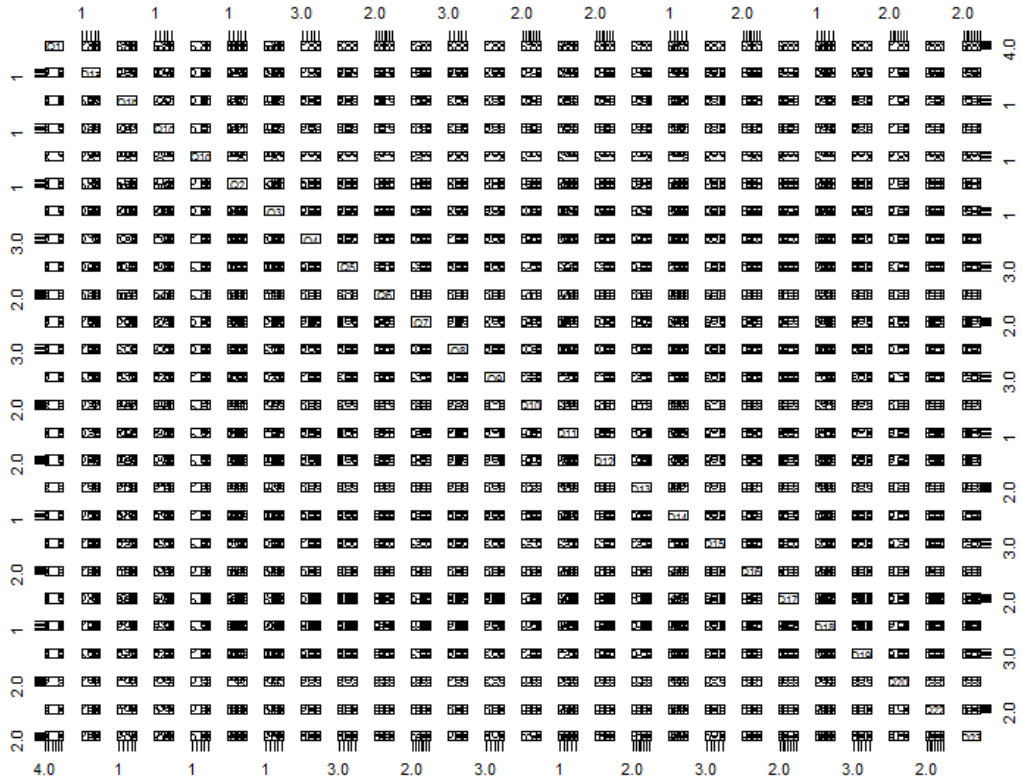


Figure 12 – Plotting for Correlating Pairs

Though the FA unsuccessful, additional reduction techniques will be conducted as well as methods to improve the FA results in order to reduce the overall number of columns before modeling is performed.

Lastly, the sum totals for the ranking question columns (question 25) from both sample sets were analyzed to confirm findings. Expectations were that the sum of all normally ranked (1 to 7, 1 being best) component values, the column with the lowest sum total would likely be the most important. Likewise, expectation was that on the transformed dummy variable columns (if rank=1 then 1, else 0), the sum total with the highest value would be the most important column to test against the target variables. Results were

consistent with expectations across both sets of predictors as well as both random sampled splits, with Q25_fun being top across the board, followed by Q25_membership, also across both sets of variables and samples. Additionally, there was a fair distance between the top two places, so we will select the top (Q25_fun) as the variable to test in the analysis frame.

Objective 3, phase 2 update – Factor Analysis (FA) was revisited at significant length for phase two of the project. It was discovered that the size of the data set (see splitting/sampling section) was the contributing factor for FA failure. New data frames were constructed based on the entire data set and reviewed in whole (Q1-Q25) and also parts of related questions Q1, Q2-Q23 and Q25, excluding Q21 and Q24 target variables. This lead to three different options for reduction making it to the modeling phase with the most appropriate model driving the selection (see model analysis).

Overall, the data file is reduced through the piece parts Factor Analysis, which resulted in 1 necessary variable from Q1 related items, 4 necessary variables from Q2-Q23 related items and 2 variables required for Q25 related items, where of course Q# is the question number from the survey. The final required reduced were Q1B, Q10, Q2, Q23, Q20, Q25_Membership and Q25_Education for the classification tree analysis and the addition of dummy variables for the regression analysis. No records were removed for reduction analysis purposes for either PCA or FA, however one incomplete record was removed for classification and regression analysis due to those analyses failing on null values.

DATA DICTIONARY

FILE NAME: FULL DEMOGRAPHICS 10_1_2016 V2.CSV

Column Name	Source	Description	Data Type	Data Class	Use
ID	DeMolay	Member ID Number, unique per member	Integer	Categorical	
ID_ct	derived	count of mem id #s	Integer	Discrete	
Active Level	DeMolay	Binary value for active (1) or inactive (4) members	Integer	binary	target
Address	DeMolay	member address	string		
City	DeMolay	member city	string		
State	DeMolay	member state	string		
Zip	DeMolay	member zip code	integer		
InitDate	DeMolay	date of member induction	date	continuous	
Tenure	DeMolay	tenure, verbose format	string		
DOB	DeMolay	member date of birth	date	continuous	
Age	DeMolay	member age in whole years	integer	continuous	
Mem_Attend_Pct	DeMolay	member attendance percentage	number	continuous	target
Age_Yrs	derived	member age in years	number	continuous	predictor
Tenure_Yrs	derived	member tenure in years	number	continuous	target
cen_rowid	census data file	row index to find census data for each id in census file	integer	continuous	
cen_MatchResult1	census.gov	match res1 from source	string		
cen_MatchResult2	census.gov	match result2 from source plus manual workaround notes	string		
cen_Latitude	census.gov, bing maps	latitude associated with member address	number	continuous	predictor

cen_Longitude	census.gov, bing maps	longitude associated with member address	number	continuous	predictor
cen_State_cd	census.gov	state FIPS code	integer	categorical	
cen_County_cd	census.gov	county FIPS code	integer	categorical	
cen_Tract_GeoID	census.gov	census tract id	integer	categorical	predictor
cen_Block_GeoID	census.gov	census block id	integer	categorical	predictor

Figure 13 – Data Dictionary: Demographics

FILE NAME: THREE WEEK SURVEY RESULTS.CSV

Column Name	Description	Data Type	Data Class
Date	Date the data was collected	DateTime	Categorical
Age	Age of the survey participant	Integer	Discrete
Level	***MEMBER*** Level* of survey participant. *Differs from active "level" in demo data	Integer	Discrete, ordinal
Q1	1-5 rating of:	Integer	Continuous
Q1a	1-5 rating of:	Integer	Continuous
Q1b	1-5 rating of:	Integer	Continuous
Q1c	1-5 rating of:	Integer	Continuous
Q1d	1-5 rating of:	Integer	Continuous
Q2	1-5 rating of:	Integer	Continuous
Q3	1-5 rating of:	Integer	Continuous
Q4	1-5 rating of:	Integer	Continuous
Q5	1-5 rating of:	Integer	Continuous
Q6	1-5 rating of:	Integer	Continuous
Q7	1-5 rating of:	Integer	Continuous
Q8	1-5 rating of:	Integer	Continuous
Q9	1-5 rating of:	Integer	Continuous

Q10	1-5 rating of:	Integer	Continuous
Q11	1-5 rating of:	Integer	Continuous
Q12	1-5 rating of:	Integer	Continuous
Q13	1-5 rating of:	Integer	Continuous
Q14	1-5 rating of:	Integer	Continuous
Q15	1-5 rating of:	Integer	Continuous
Q16	1-5 rating of:	Integer	Continuous
Q17	1-5 rating of:	Integer	Continuous
Q18	1-5 rating of:	Integer	Continuous
Q19	1-5 rating of:	Integer	Continuous
Q20	1-5 rating of:	Integer	Continuous
Q21	1-5 rating of:	Integer	Continuous
Q22	1-5 rating of:	Integer	Continuous
Q23	1-5 rating of:	Integer	Continuous
Q24	1-5 rating of:	Integer	Continuous
Q25_Membership	Relative Importance/Rank of Membership	Integer	Continuous
Q25_LCC_RD	Relative Importance/Rank of LCC_RD	Integer	Continuous
Q25_Ritual	Relative Importance/Rank of Ritual	Integer	Continuous
Q25_Fun	Relative Importance/Rank of Fun	Integer	Continuous
Q25_Education	Relative Importance/Rank of Education	Integer	Continuous
Q25_Com_Service	Relative Importance/Rank of Com_Service	Integer	Continuous
Q25_Traveling	Relative Importance/Rank of Traveling	Integer	Continuous
Q25_Membership_ranked1	Flag for if item was ranked #1 (1 yes, 0 no)	Integer	Discrete
Q25_LCC_RD_ranked1	Flag for if item was ranked #1 (1 yes, 0 no)	Integer	Discrete
Q25_Ritual_ranked1	Flag for if item was ranked #1 (1 yes, 0 no)	Integer	Discrete
Q25_Fun_ranked1	Flag for if item was ranked #1 (1 yes, 0 no)	Integer	Discrete

Q25_Education_ranked1	Flag for if item was ranked #1 (1 yes, 0 no)	Integer	Discrete
Q25_Com_Service_ranked1	Flag for if item was ranked #1 (1 yes, 0 no)	Integer	Discrete
Q25_Traveling_ranked1	Flag for if item was ranked #1 (1 yes, 0 no)	Integer	Discrete

Figure 14 – Data Dictionary: Survey

FILE NAME: DEMOLAY DATA.CSV

DeMolay Data.csv				
Attribute Name	Description	Data Type	Source	Continuous (Monthly) or Discrete (Yearly)
Members_Per_Event	Number of Members Per Event That Month	Double	Internal	N/A
Month	Number Designator of each month starting with January 2012	Integer	Internal	N/A
Year	Year that the data was collected	Integer	Internal	N/A
mem_Advisor_Cert	Awarded if the Advisor Certification paperwork was turned in	Integer	Internal	Discrete
mem_Advisor_Recruitment	Awarded when a new advisor is recruited	Integer	Internal	Continuous
rit_Advisor_Participation	Awarded for each event advisors participate in	Integer	Internal	Continuous
edu_Annual_Financial	Awarded if the Annual Financial Report was turned in	Integer	Internal	Discrete
fun_Athletics	Awarded for each Athletic event	Integer	Internal	Continuous
edu_Website	Awarded for each Website updated	Integer	Internal	Continuous
edu_Newspaper	Awarded for each Newspaper Article	Integer	Internal	Continuous
edu_Radio_Ad	Awarded for each Radio ad	Integer	Internal	Continuous
edu_Billboard	Awarded for each Billboard	Integer	Internal	Continuous
edu_Television	Awarded for each Television spot	Integer	Internal	Continuous
edu_Chapter_Newsletter	Awarded for each Chapter newsletter	Integer	Internal	Continuous
edu_Checklist	Awarded if the yearly Chapter Checklist is filled out	Integer	Internal	Discrete
comserv_Charitable	Awarded for each Charitable donation	Integer	Internal	Continuous
trav_Conclave_Councilors	Awarded for each Councilor that attends the yearly Conclave	Integer	Internal	Discrete

trav_Conclave_DeMolays	Awarded for each DeMolay that attends the yearly Conclave	Integer	Internal	Discrete
trav_Conclave_Advisors	Awarded for each Advisor that attends the yearly Conclave	Integer	Internal	Discrete
fun_March_Activities	Awarded for each March State Activity	Integer	Internal	Discrete
fun_March_Award	Awarded for winning the March Competition	Integer	Internal	Discrete
comserv_DeMolay_Service	Awarded for each DeMolay Service activity	Integer	Internal	Continuous
rit_Teams_Assist	Awarded for each time the chapter assists with a ritual team	Integer	Internal	Continuous
trav_District	Awarded for each attended District Meeting	Integer	Internal	Continuous
mem_Fundraising	Awarded for each dollar fundraised	Integer	Internal	Continuous
mem_Founders	Awarded for a member receiving their Founder's Award	Integer	Internal	Continuous
mem_BHK	Awarded for a member receiving their Blue Honor Key	Integer	Internal	Continuous
fun_LTC	Awarded for each Leadership Training Conference Attended	Integer	Internal	Continuous
awards_LCC12	Awarded for each LCC 1 or LCC 2 completed	Integer	Internal	Continuous
awards_LCC345	Awarded for each LCC 3, LCC 4 or LCC 5 completed	Integer	Internal	Continuous
comserv_Masonic	Awarded for each hour of Masonic Service	Integer	Internal	Continuous
mem_New_Members	Awarded for each new member initiated	Integer	Internal	Continuous
trav_Mileage	Awarded for each mile traveled	Integer	Internal	Continuous
fun_Misc	Awarded for each miscellaneous event	Integer	Internal	Continuous
awards_Obligatory	Awarded for each Obligatory Day observed	Integer	Internal	Continuous
awards_PMCMSA	Awarded for each PMC-MSA award obtained	Integer	Internal	Continuous
rit_Meeting_noRit	Awarded for each meeting no rituals are utilized	Integer	Internal	Continuous
rit_Meeting_1Rit	Awarded for each meeting one ritual is utilized	Integer	Internal	Continuous
rit_Qual_Ind	Awarded for each Individual ritual qualification	Integer	Internal	Continuous
rit_Qual_Public	Awarded for each public ceremony qualified	Integer	Internal	Continuous
rit_Performance	Awarded for each ritual performance	Integer	Internal	Continuous
rit_Tournament	Awarded for each ritual tournament attended	Integer	Internal	Continuous
rit_Competing	Awarded for each ritual tournament competed in	Integer	Internal	Continuous

rit_Judging	Awarded for each ritual judge trained	Integer	Internal	Continuous
awards_RD	Awarded for each RD completed	Integer	Internal	Continuous
fun_social	Awarded for each social event attended	Integer	Internal	Continuous
trav_Visitation	Awarded for each chapter visitation	Integer	Internal	Continuous
trav_Workshop_Councilors	Awarded for each Councilor that attends the yearly Workshop	Integer	Internal	Discrete
trav_Workshop_DeMolays	Awarded for each DeMolay that attends the yearly Workshop	Integer	Internal	Discrete
trav_Workshop_Advisors	Awarded for each Advisor that attends the yearly Workshop	Integer	Internal	Discrete

Column Name	Description	Data Type	Data Class
TRAVEL	Events related to Travel	integer	Continuous
RIT	Events related to Rituals	Integer	Continuous
AWARDS	Events related to Awards	Integer	Continuous
COMSERV	Events related to community service	Integer	Continuous
FUN	Events related to fun	Integer	Continuous
EDUCATION	Events related to education	integer	Continuous
MEMBERSHIP	Events related to membership	integer	Continuous

Figure 15 – Data Dictionary: DeMolay Data

DESCRIPTIVE STATISTICS

To begin with objective 1: correlation between time of year and Chapter events for optimal participation, univariate descriptive analysis was performed for all 9 independent variables to observe the dataset and observe outliers. Based on the analysis, all variables are left skewed except for Education, and all variables contain outliers extending beyond the

third quartile. In order to evaluate the variables relationships, bivariate descriptive analysis is completed.

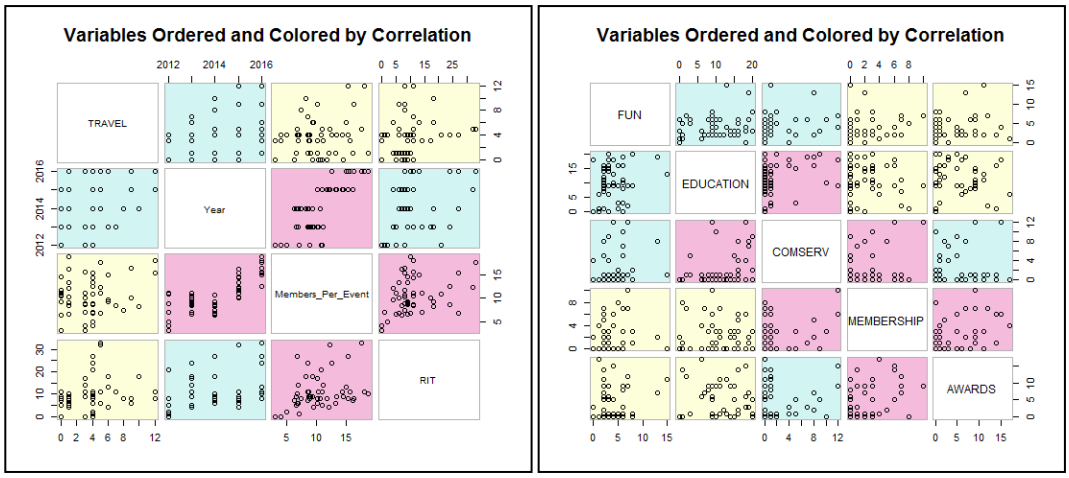


Figure 16 – Objective 1 Bivariate Correlation Matrix

From the descriptive analysis, the data demonstrates that variable transformation would be beneficial. The most applicable variable transformation is log(x+1) because the data contains zeros and is left skewed as well as this transformation provides more normal data with constant variance and reduced skewness. With log(x+3/8) transformation applied, the distribution is normal as showed in Figure 17.

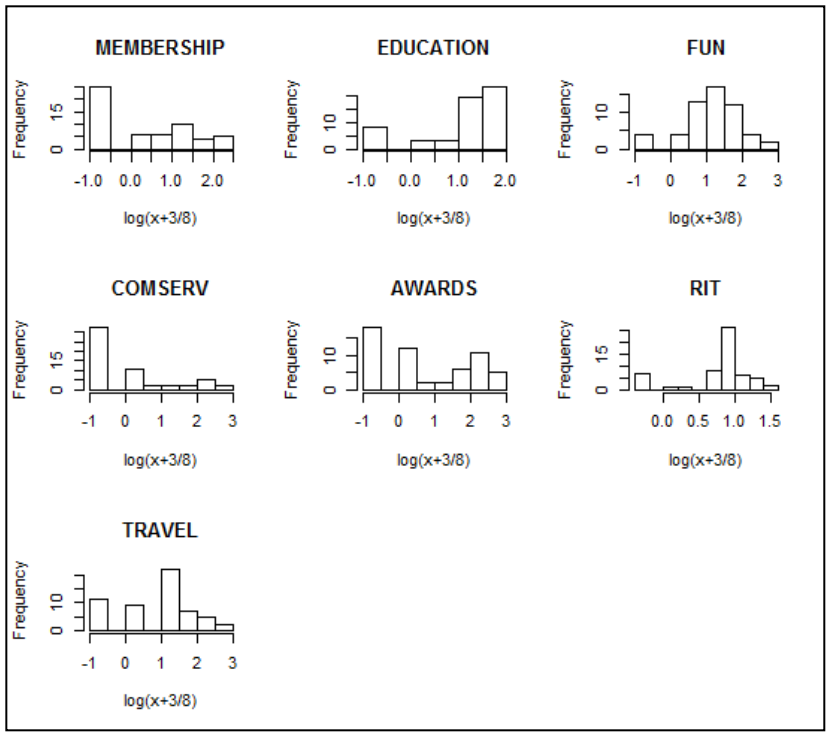
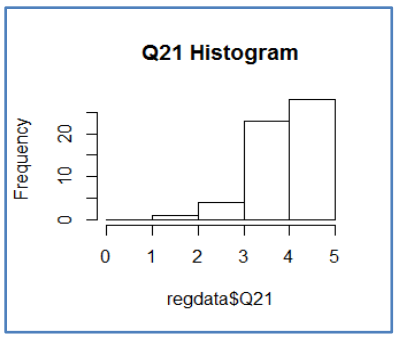
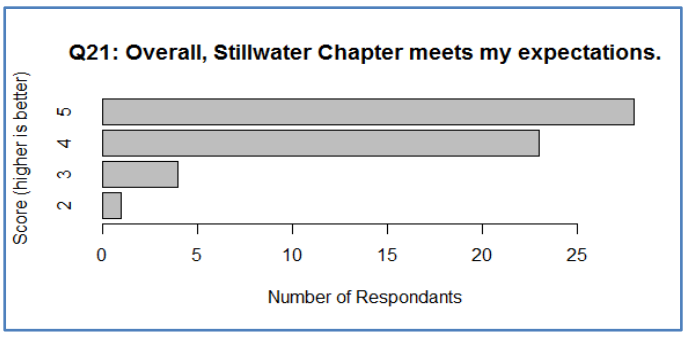


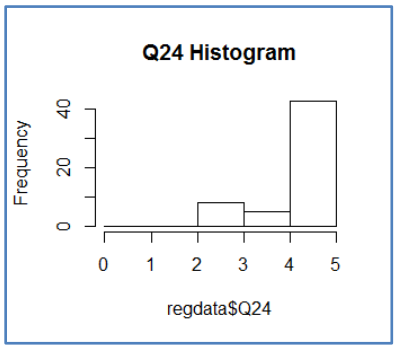
Figure 17 – Objective 1 Variable Transformation with Log(x+1)

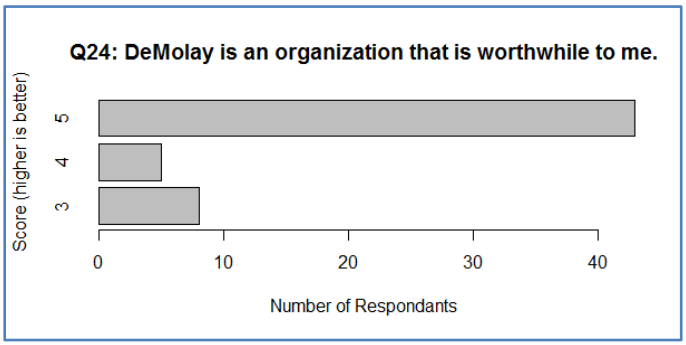
For objective 2, demographic and geographic analysis, the dataset contains 81 samples for 50 active members and 31 inactive members. Within the samples, the average member age for the Stillwater Chapter is about 17 with the average member tenure around 7 years.

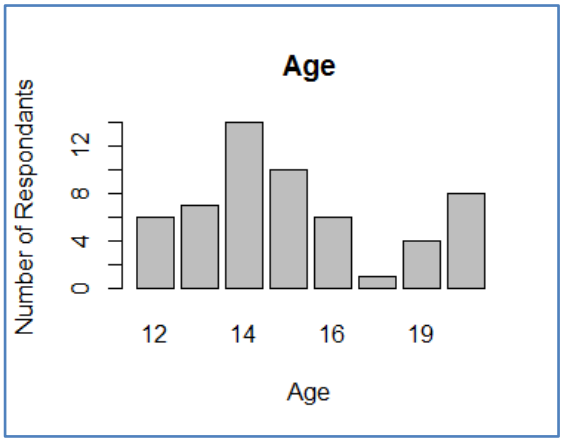
With regard to objective 3, descriptive statistics shed some light on our client question regarding the relationship of age, member level and chapter/organization perceptions. Insight is gained through examining distributions of the 56 records and various targets and predictors. We can see that it may be difficult to discern a relationship due to the skewness of target variables compared to predictor variables via the distributions below:

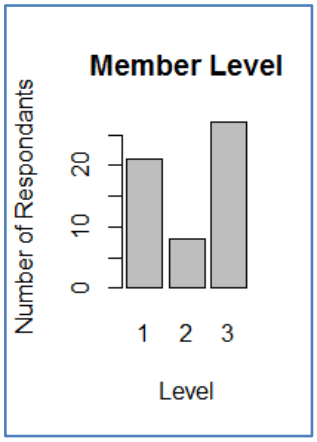












While it is possible that with grouping or clustering among groups that some influence of age or member level exists, but the above charts indicate such heavy concentrations in only a few scores that other indicators may emerge. To the extent that broad descriptive statistics are valuable, all fields from the final frame (regdata) are detailed below.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se

Date*	1	56	2.09	0.82	2	2.11	1.48	1	3	2	-0.16	-1.51	0.11
Age*	2	56	4.11	2.21	4	4.02	1.48	1	8	7	0.52	-0.88	0.3
Level*	3	56	2.11	0.93	2	2.13	1.48	1	3	2	-0.21	-1.83	0.12
Q1	4	56	4.55	0.71	5	4.67	0	2	5	3	-1.82	3.54	0.1
Q1a	5	56	4.45	0.85	5	4.59	0	1	5	4	-1.66	2.97	0.11
Q1b	6	56	4.14	1.26	5	4.35	0	1	5	4	-1.24	0.18	0.17
Q1c	7	56	4.27	1.18	5	4.5	0	1	5	4	-1.55	1.25	0.16
Q1d	8	56	4.39	1.02	5	4.63	0	1	5	4	-2.13	4.21	0.14
Q2	9	56	3.36	1.42	3.5	3.43	2.22	1	5	4	-0.26	-1.33	0.19
Q3	10	56	3.68	1.32	4	3.83	1.48	1	5	4	-0.66	-0.67	0.18
Q4	11	56	4.39	0.8	5	4.52	0	1	5	4	-1.63	3.74	0.11
Q5	12	56	3.98	0.96	4	4.04	1.48	1	5	4	-0.57	-0.24	0.13
Q6	13	56	4	1.06	4	4.11	1.48	2	5	3	-0.45	-1.29	0.14
Q7	14	56	4.38	0.7	4	4.48	1.48	2	5	3	-0.95	0.72	0.09
Q8	15	56	4.27	0.67	4	4.33	1.48	3	5	2	-0.36	-0.88	0.09
Q9	16	56	4.57	0.71	5	4.7	0	3	5	2	-1.3	0.17	0.09
Q10	17	56	4.52	0.79	5	4.67	0	2	5	3	-1.6	1.89	0.11
Q11	18	56	4.07	1.17	4	4.26	1.48	1	5	4	-1.26	0.6	0.16
Q12	19	56	4.2	0.86	4	4.3	1.48	2	5	3	-0.88	0.05	0.12
Q13	20	56	4.14	0.92	4	4.24	1.48	2	5	3	-0.69	-0.65	0.12
Q14	21	56	3.59	1.14	4	3.63	1.48	1	5	4	-0.25	-1.14	0.15
Q15	22	56	4.57	0.74	5	4.72	0	2	5	3	-1.59	1.69	0.1
Q16	23	56	3.59	1.12	4	3.7	1.48	1	5	4	-0.67	-0.15	0.15
Q17	24	56	3.73	1.07	4	3.78	1.48	2	5	3	-0.25	-1.24	0.14
Q18	25	56	3.75	1.08	4	3.85	1.48	1	5	4	-0.6	-0.34	0.14
Q19	26	56	4.38	0.8	5	4.48	0	2	5	3	-0.96	-0.13	0.11
Q20	27	56	4.62	0.82	5	4.83	0	2	5	3	-2.33	4.53	0.11
Q21	28	56	4.39	0.71	4.5	4.5	0.74	2	5	3	-1	0.77	0.09
Q22	29	56	4.12	0.95	4	4.22	1.48	1	5	4	-0.86	0.25	0.13
Q23	30	56	4.16	0.93	4	4.26	1.48	2	5	3	-0.71	-0.65	0.12
Q24	31	56	4.62	0.73	5	4.76	0	3	5	2	-1.54	0.66	0.1
Q25mem	32	56	3.36	1.78	4	3.26	1.48	1	7	6	0.16	-1.03	0.24
Q25lcc	33	56	4.61	1.63	5	4.67	1.48	1	7	6	-0.29	-0.69	0.22
Q25rit	34	56	4.07	2.03	4	4.09	2.22	1	7	6	0.2	-1.26	0.27
Q25fun	35	56	3.25	2.24	2.5	3.09	2.22	1	7	6	0.56	-1.22	0.3
Q25edu	36	56	4.32	1.96	4.5	4.39	2.22	1	7	6	-0.22	-1.37	0.26
Q25com	37	56	3.8	1.89	4	3.8	2.97	1	7	6	-0.01	-1.35	0.25
Q25tra	38	56	4.59	2.05	5	4.72	2.97	1	7	6	-0.36	-1.31	0.27
Q25mem_1	39	56	0.21	0.41	0	0.15	0	0	1	1	1.36	-0.16	0.06
Q25lcc_1	40	56	0.04	0.19	0	0	0	0	1	1	4.87	22.12	0.03
Q25rit_1	41	56	0.11	0.31	0	0.02	0	0	1	1	2.47	4.19	0.04

Q25fun_1	42	56	0.32	0.47	0	0.28	0	0	1	1	0.74	-1.47	0.06
Q25edu_1	43	56	0.09	0.29	0	0	0	0	1	1	2.8	5.97	0.04
Q25com_1	44	56	0.14	0.35	0	0.07	0	0	1	1	1.99	1.98	0.05
Q25tra_1	45	56	0.09	0.29	0	0	0	0	1	1	2.8	5.97	0.04
age_12	46	56	0.11	0.31	0	0.02	0	0	1	1	2.47	4.19	0.04
age_13	47	56	0.12	0.33	0	0.04	0	0	1	1	2.21	2.93	0.04
age_14	48	56	0.25	0.44	0	0.2	0	0	1	1	1.12	-0.75	0.06
age_15	49	56	0.18	0.39	0	0.11	0	0	1	1	1.63	0.68	0.05
age_16	50	56	0.11	0.31	0	0.02	0	0	1	1	2.47	4.19	0.04
age_18	51	56	0.02	0.13	0	0	0	0	1	1	7.09	49.11	0.02
age_19	52	56	0.07	0.26	0	0	0	0	1	1	3.24	8.65	0.03
age_20	53	56	0.14	0.35	0	0.07	0	0	1	1	1.99	1.98	0.05
lev_1	54	56	0.38	0.49	0	0.35	0	0	1	1	0.5	-1.78	0.07
lev_2	55	56	0.14	0.35	0	0.07	0	0	1	1	1.99	1.98	0.05
lev_3	56	56	0.48	0.5	0	0.48	0	0	1	1	0.07	-2.03	0.07

MODEL ASSUMPTIONS

For objective 1: correlation between time of year and Chapter events for participation, VAR time series modeling was utilized. Prior to building the model, it is necessary to check that time series is stationary. To be stationary three things are required:

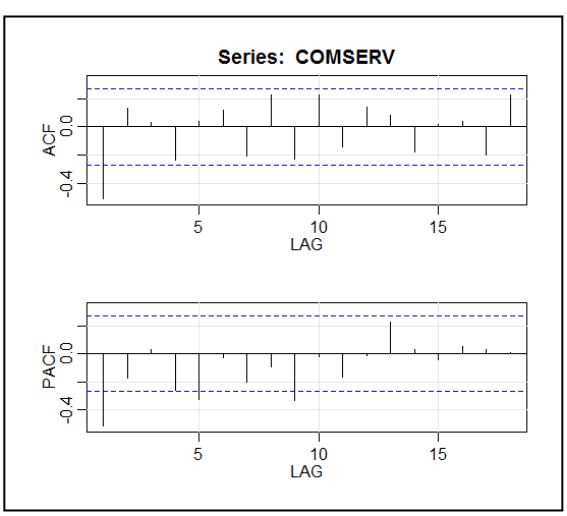
1. Constant mean across all time.

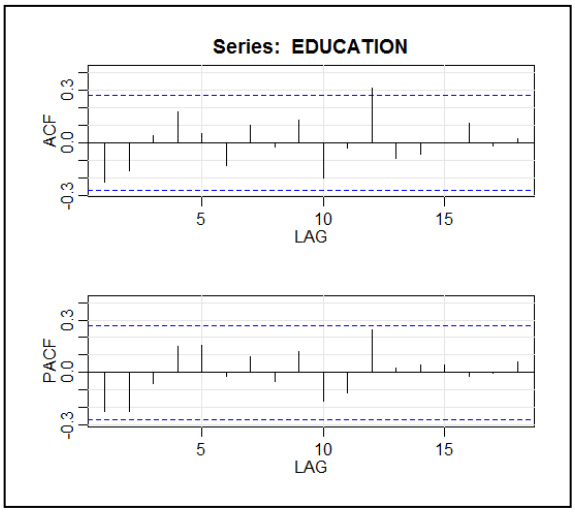
2. Constant variance across all time.

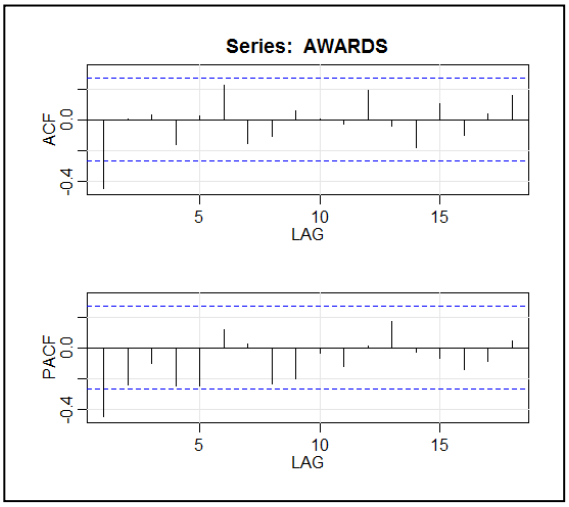
3. Auto covariance between two observations is dependent on distance between the observations.

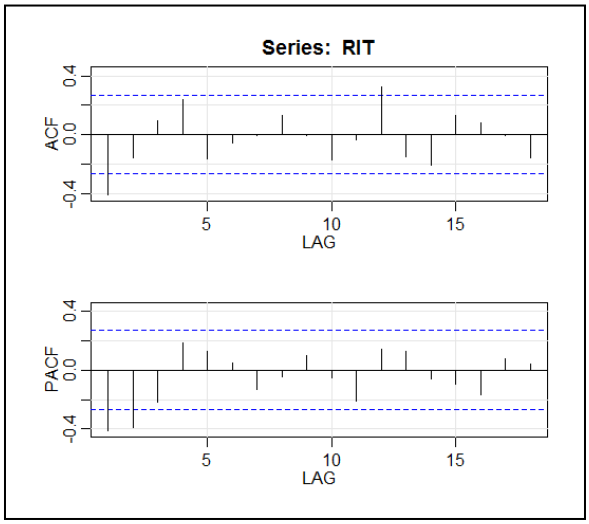
Real data usually does not meet these standards unless white noise is measured, and this can be done using acf2() function in R. ACF (Auto correlation function) and PACF (partial correlation function) will show a quick drop off in correlation after a lag between points. Plots for all the variables can be referenced below in Figure 18. The dotted lines signify the

threshold for each lag. Reviewing the ACF and PACF plot images, the data is stationary since there is no correlation between future and previous values. Therefore, there is no need to transform series to stationary. From the ACF plots there is geometric decay that suggests the use of a pure AR(p) model. Value of p can be considered by looking at PACF where there is a cut off with the threshold dotted lines.









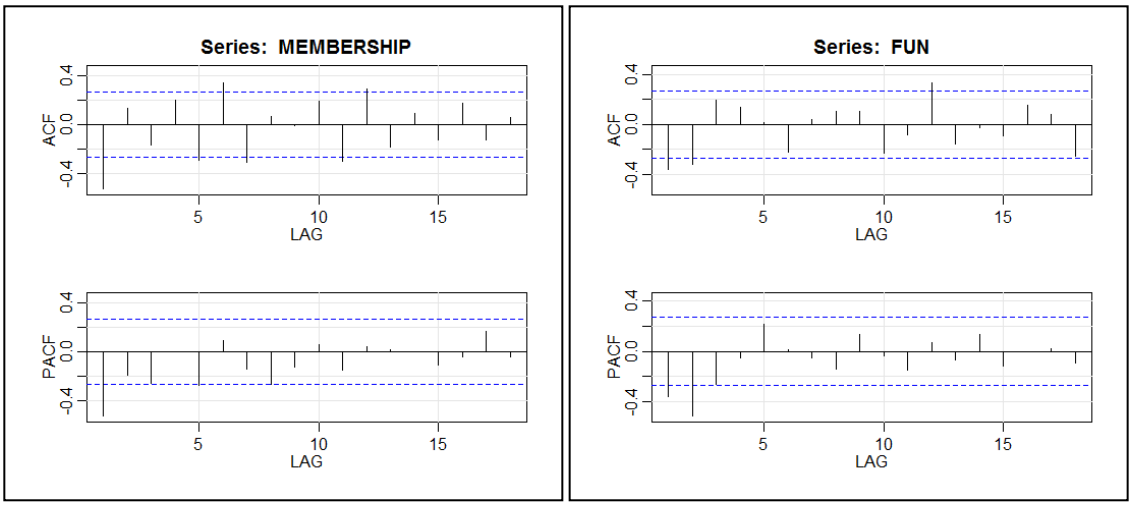


Figure 18 – Objective 1 ACF & PACF Plots

To gain additional confidence for validating stationary, a Dickey Fuller test of Stationary is used with R's package 'forecast' and a function called auto.arima(), which provides accurate ARIMA parameters(p,d,q). Figure 19 shows the results from the Dickey Fuller test using auto.arim(), and from results it is clear that 'd' parameter is zero for all univariate time series, thus indicating no integrative portion.

```
> auto.arima(TRAVEL)
Series: TRAVEL
ARIMA(0,0,1) with zero mean

Coefficients:
          ma1
      -0.9124
s.e.   0.1345

sigma^2 estimated as 0.6343:  log likelihood=-65.91
AIC=135.82   AICc=136.05   BIC=139.84
>
```

```
> auto.arima(RIT)
Series: RIT
ARIMA(0,0,1) with zero mean

Coefficients:
          ma1
      -0.5247
s.e.   0.0958

sigma^2 estimated as 0.3529:  log likelihood=-49.05
AIC=102.1   AICc=102.33   BIC=106.12
>
```

```
> auto.arima(MEMBERSHIP)
Series: MEMBERSHIP
ARIMA(0,0,1) with zero mean

Coefficients:
          ma1
      -0.9545
s.e.   0.0614

sigma^2 estimated as 0.664:  log likelihood=-67.48
AIC=138.97   AICc=139.2   BIC=142.98
>
```

```
> auto.arima(FUN)
Series: FUN
ARIMA(2,0,2) with zero mean

Coefficients:
          ar1      ar2      ma1     ma2
      -0.0141  -0.6075  -0.6947  0.4732
s.e.   0.1902   0.1476   0.2114  0.1753

sigma^2 estimated as 0.2832:  log likelihood=-41.91
AIC=93.81   AICc=95.04   BIC=103.85
>
```

```
> auto.arima(EDUCATION)
Series: EDUCATION
ARIMA(0,0,1) with zero mean

Coefficients:
          ma1
      -0.2752
s.e.   0.1316

sigma^2 estimated as 0.113:  log likelihood=-17.61
AIC=39.23   AICc=39.46   BIC=43.24
>
```

```
> auto.arima(COMSERV)
Series: COMSERV
ARIMA(0,0,1) with zero mean

Coefficients:
          ma1
      -0.8334
s.e.   0.0952

sigma^2 estimated as 0.7036:  log likelihood=-68.46
AIC=140.92   AICc=141.15   BIC=144.94
>
```

```
> auto.arima(AWARDS)
Series: AWARDS
ARIMA(0,0,1) with zero mean

Coefficients:
          ma1
      -0.7428
s.e.   0.1136

sigma^2 estimated as 1.001:  log likelihood=-77.96
AIC=159.92   AICc=160.15   BIC=163.94
>
```

Figure 19 – Objective 1 Dickey Fuller Test

Additionally, the Augmented Dickey-Fuller Test Results for all univariate analysis shows that data is stationary, and can be referenced in Figure 20. We can validate that our data is appropriate for the time series analysis.

```
        Augmented Dickey-Fuller Test

data:  log_data$MEMBERSHIP
Dickey-Fuller = -3.6407, Lag order = 3, p-value = 0.03778
alternative hypothesis: stationary

> adf.test(log_data$EDUCATION,alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  log_data$EDUCATION
Dickey-Fuller = -0.85977, Lag order = 3, p-value = 0.9511
alternative hypothesis: stationary

> adf.test(log_data$FUN,alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  log_data$FUN
Dickey-Fuller = -3.0567, Lag order = 3, p-value = 0.1485
alternative hypothesis: stationary

> adf.test(log_data$COMSERV,alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  log_data$COMSERV
Dickey-Fuller = -3.7439, Lag order = 3, p-value = 0.02912
alternative hypothesis: stationary

> adf.test(log_data$AWARDS,alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  log_data$AWARDS
Dickey-Fuller = -3.2136, Lag order = 3, p-value = 0.09424
alternative hypothesis: stationary
```

```
> adf.test(log_data$RIT,alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  log_data$RIT
Dickey-Fuller = -2.4215, Lag order = 3, p-value = 0.4046
alternative hypothesis: stationary

> adf.test(log_data$TRAVEL,alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  log_data$TRAVEL
Dickey-Fuller = -3.9937, Lag order = 3, p-value = 0.0163
alternative hypothesis: stationary

>
```

Figure 20 – Objective 1 Augmented Dickey-Fuller Test

For objective 2: geographic and demographic analysis, the classification tree was applied due to the several explanatory variables in the file "Demographics Data" contained. For the classification tree to be applicable, the response variable is a continuous measurement with explanatory values either continuous or categorical. For our purpose of objective 2, the categorical regression models serves to confirm if there is a relationship between membership tenure, which is our target variable, and the distance the members live from the Chapter, which are the independent variables and are represented by the multiple census tract geographic ID fields.

Regarding objective 3 – relationship of age, member level and survey responses to perceptions of chapter and organization – two modeling techniques were utilized. These techniques are that of classification trees and multiple regression for a mix of categorical and continuous predictor variables. Due to the mixed use of categorical and continuous data, the types of analyses used do not have any operating assumptions that have to be tested such as for linearity or autocorrelation that we see with linear regression. That is not to say however that there aren't operating assumptions in play. For trees, the standard of best fit at 3 levels will be examined. Similarly for regression modeling, the typical fitness tests still apply – models are only acceptable with significant p-values and adjusted R-squared values with coefficients also requiring significant p-values to be included.

MODELING WITH OBJECTIVES

For objective 1: correlation between time of year and Chapter events, the data utilized was collected during the time period January 2012 to August 2015, which indicates that the dataset is a time series data. Regression model on time series was the model of choice for the

purpose of the objective is to determine the correlation between Chapter events and member participation. Time series modeling involves working on time data to derive insights to make a decision. Model technique selection requires a thorough understanding of number of time series variables involved, stationary and non-stationary series, seasonality in time series, trends, time series domain. As a first step, the data was converted to time series using the TS() function in R. Figure 18 demonstrates the data as it progressed throughout time.

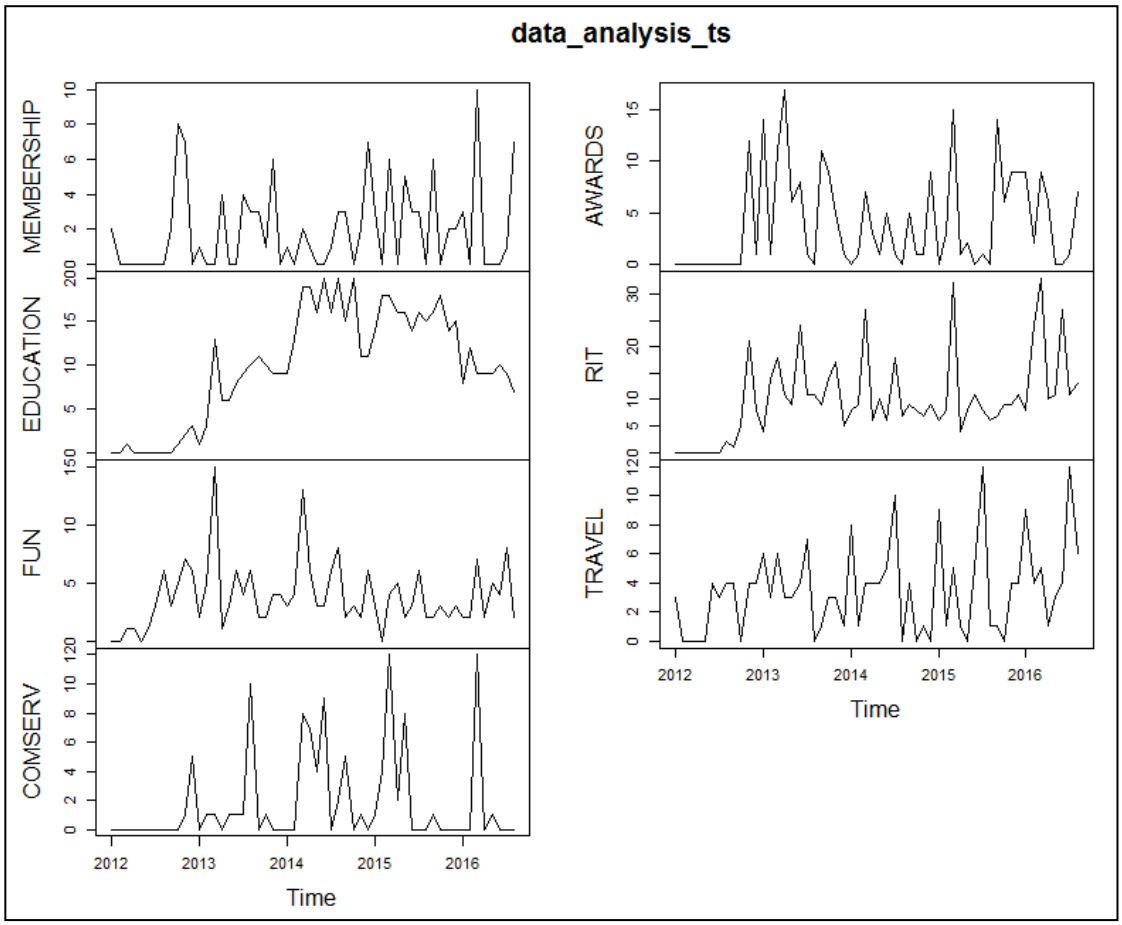


Figure 21 – Objective 1 Time Series Plot

Initially for objective 1, the ARIMA model was selected as the choice to analyze the time series data. The ARIMA model is denoted by three components: p, d, and q. P is the order

of autoregressive model, d is the degree of differencing, and q is the order of moving average model Furthermore, since the data used is seasonal with a period of 12 months, a much more advanced model ARIMA (p, d, q)(P, D, Q)m was also considered because M is the number of periods in each season. Upper case P, D, Q represents the same terms for seasonal component. This approach was a failure as ARIMA model is suited for data with univariate time series, however the data set considered is multivariate time series.

After confirming that multivariate time series is necessary, we applied the analysis through the ARIMAX model with the consideration of testing the VAR model also. Comparing ARIMAX and VAR, VAR is a more applicable model for the time series data because VAR can be estimated using OLS or GLS, which is fast, while ARIMAX requires maximum likelihood which is generally slow.

For objective 2, the goal is to find the correlation between member tenure and geographical/demographical information, so it was decided to utilize regression using categorical data as the model that would most accurately provide this information. Spatial analysis was considered, but was outside the scope of this project. Due to the categorical variables, there were no assumptions that went into this model. The data was compared first by viewing the addresses as geographical locations for active and inactive members through the website BatchGeo, as shown by figure 22.

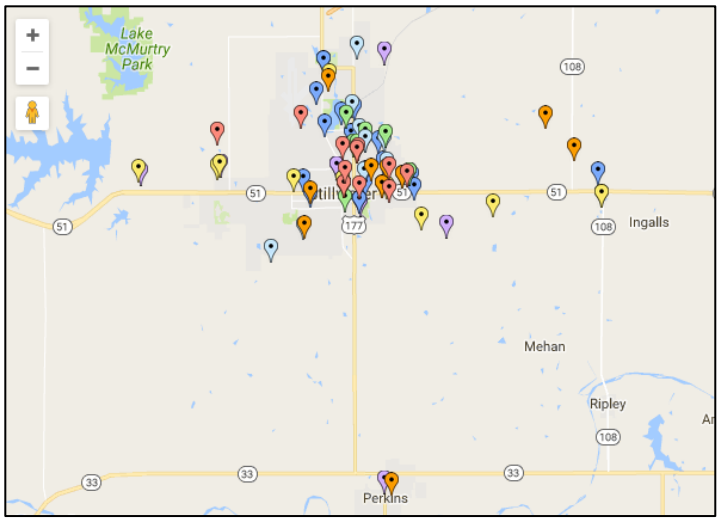


Figure 22 – Objective 2: Graphical Representation of Addresses

From the figure 22, there does seem to be an eye test validation between the correlation between the location of the members' address and tenure. The longest tenured members all live within a small area within Stillwater. As other groups are added according to tenure, the map zooms out to encompass the other addresses. However, when looking at average distance from the Masonic Lodge, the Chapter meeting location, there are the splits based on tenure:

Tenure of Active Members	Average Miles From Lodge	Count of Members
4.02 – 6.06 years	1.28 mi	3
2.54 – 3.51 years	3.55 mi	4
1.87 – 2.12 years	2.22 mi	3
1.53 – 1.85 years	1.99 mi	3
1.07 – 1.43 years	3.99 mi	6
0.12 – 0.85 years	3.15 mi	8
0.08 years	3.60 mi	3

Figure 23 – Objective 2: Table Representation between Tenure and Members' Geographic Location

For the categorical regression, supervised binning was performed to gain initial insight if the tenure was influenced by members' geographic location by imposing categories onto the numerical variables of the census geographical tract IDs. Supervised binning is the selected technique to apply the datasets based on the geo maps.

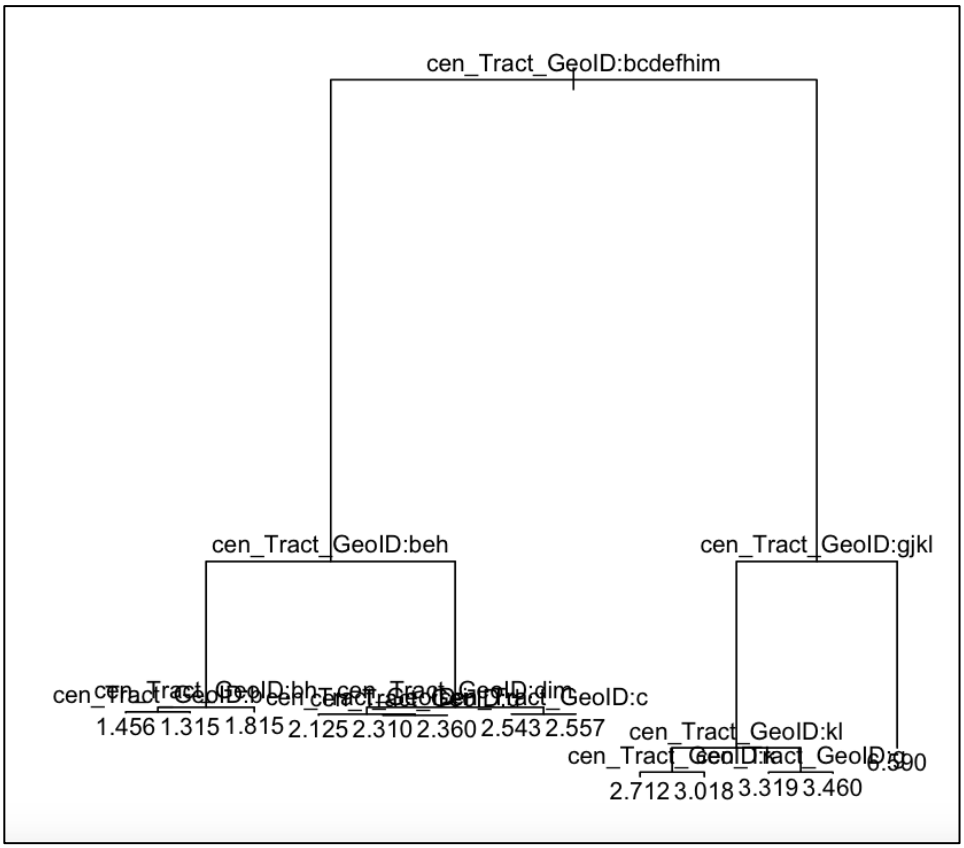


Figure 24 – Objective 2: Classification Tree

After the supervised binning demonstrates the levels, indicator variables are created for the census tract geographic ID field. This creates 13 indicator variables, but the very last dummy variable will be excluded since it populates a 'NA' in the model.

The first regression model applies tenure as the target variable to compare against the censure tract geo IDs. The regression model displays that there is no significant

relationship between the variables and their impact on member tenure. The adjusted R-squared is -0.005156 with a p-value of 0.4895. All the variables have insignificant impact based on their slope and p-value.

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.8980 -1.5086 -0.1180  0.7614  4.9533

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                             2.3600     1.2206   1.934   0.0573 .
demo_data$cen_Tract_GeoID_40027202206   4.2300     2.4412   1.733   0.0877 .
demo_data$cen_Tract_GeoID_40119010101  -0.9044     1.4094  -0.642   0.5232
demo_data$cen_Tract_GeoID_40119010102   0.1827     1.3770   0.133   0.8948
demo_data$cen_Tract_GeoID_40119010200  -0.0500     1.4589  -0.034   0.9728
demo_data$cen_Tract_GeoID_40119010300  -0.5450     1.9299  -0.282   0.7785
demo_data$cen_Tract_GeoID_40119010500   0.1967     1.7262   0.114   0.9096
demo_data$cen_Tract_GeoID_40119010600   0.9586     1.3011   0.737   0.4638
demo_data$cen_Tract_GeoID_40119010700  -1.0450     1.6147  -0.647   0.5197
demo_data$cen_Tract_GeoID_40119010800  -0.2350     1.9299  -0.122   0.9034
demo_data$cen_Tract_GeoID_40119010900   1.1000     1.7262   0.637   0.5261
demo_data$cen_Tract_GeoID_40119011000   0.3525     1.6147   0.218   0.8278
demo_data$cen_Tract_GeoID_40119011101   0.6580     1.3917   0.473   0.6379
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.114 on 68 degrees of freedom
Multiple R-squared:  0.1456,  Adjusted R-squared:  -0.005156
F-statistic: 0.9658 on 12 and 68 DF,  p-value: 0.4895
```

Figure 25 – Objective 2: Regression Model 1 – Tenure & Census Geo Tract ID

To attempt to find a strong relationship, the second regression model uses member participation levels as the target variable to verify if the census geographic IDs have an influence on participation. Similar to the first regression model, there is little significance. The adjusted R-squared value is -0.1295 with a p-value of 0.9957. Additionally, the variables have insignificant slopes and p-values demonstrating that no one variable plays a large role in membership tenure.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.33167 -0.23429 -0.08508  0.24860  0.58206

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           0.191322   0.170504   1.122    0.266
demo_data$cen_Tract_GeoID_40027202206 -0.056588   0.341007  -0.166    0.869
demo_data$cen_Tract_GeoID_40119010101  0.065647   0.196881   0.333    0.740
demo_data$cen_Tract_GeoID_40119010102  0.140345   0.192354   0.730    0.468
demo_data$cen_Tract_GeoID_40119010200  0.053144   0.203791   0.261    0.795
demo_data$cen_Tract_GeoID_40119010300 -0.076406   0.269590  -0.283    0.778
demo_data$cen_Tract_GeoID_40119010500 -0.006137   0.241129  -0.025    0.980
demo_data$cen_Tract_GeoID_40119010600  0.112902   0.181757   0.621    0.537
demo_data$cen_Tract_GeoID_40119010700  0.120843   0.225555   0.536    0.594
demo_data$cen_Tract_GeoID_40119010800  0.132150   0.269590   0.490    0.626
demo_data$cen_Tract_GeoID_40119010900  0.138119   0.241129   0.573    0.569
demo_data$cen_Tract_GeoID_40119011000 -0.024264   0.225555  -0.108    0.915
demo_data$cen_Tract_GeoID_40119011101  0.068540   0.194404   0.353    0.726

Residual standard error: 0.2953 on 68 degrees of freedom
Multiple R-squared:  0.03996, Adjusted R-squared:  -0.1295
F-statistic: 0.2359 on 12 and 68 DF,  p-value: 0.9957
```

Figure 26 – Objective 2: Regression Model 2 – Membership Participation & Census Geo Tract ID

With regard to Objective 3 – survey data and perceptions of chapter/organization – as noted previously, two techniques were utilized – classification trees and multiple regression. Additionally, two separate target variables Q21 (chapter perception) and Q24 (organization perception) were identified, so models were built for both using both predictors. Also as noted previously, Factor Analysis led to three different reduction possibilities. This stems from piece part analysis of related questions (referred to as 7var based on resulting number of predictors), analysis of the whole (9var) and minimized analysis of the whole (2var). We've already discussed 7var requiring 7 variables as a result of individual analysis of Q1 related items, Q2-Q23 related items and Q25 related items. The 9var FA result is based on first pass of analyzing the entire set of Q1-Q25, excluding target variables. 10 factors were deemed sufficient per PCA and 10 were attempted via FA. The

result was 9 variables loading well into the factors with factor four failing to load. Due to 7var only describing 50% or so of cumulative variance per FA results, 9var was also kept as a possibility to test. Lastly, 9var was further distilled into 2var by iterating through the process of refining FA by removing variables that don't load into any factors. This was pursued until we were finally down to only a handful of variables left and resulted in a solid Factor Analysis result of 2 variables, but only having a cumulative variance of about 25%. This was kept as a refinement for testing as well.

This led to regression testing of 3 different predictor sets for each question. However it was then discovered that the intercept can be forced off for special cases where there should be no intercept when variables are undefined or zero. Seeing as this is the case with survey data (no answers, no results), it was determined to also test regression models with and without this parameter. Additionally it was determined that two more variants would be tested, those with the dummy variables included and those without. The final result of regression modeling was 3 predictor sets with and without intercept, with and without dummy variables, for a total of 12 models per target variable or 24 total regression models being built and reviewed. Results were difficult to choose between, as half could be discarded due to fit or p-value, but the other half all had highly significant p-values (all <2.2e-16) and very high adjusted R-squared (all 0.97 or better). The half that got discarded were universally those with an intercept and which exhibited high p-values in most cases and in all cases low adjusted R-squared values (at best up to 0.38). Of the half that remained for each target, both the predictor questions, values, coefficients, direction of coefficients and model results were reviewed. The questions all seemed to make sense with respect to their models and other included elements. In the end selections were made based on the best fit

and strong coefficients, which also quite nicely lined up with the need to understand the relationship of age and level to the target variables. This will be expanded on in the results section of the document.

Classification tree analysis was much more straightforward and two trees were considered for each factor set (7var, 2var, 9var) for each target, resulting in 12 total trees to build and review. However in choosing them, that was a bit easier. To remain consistent in comparing the regression and classification models, the predictor set was chosen to align with regression analysis, cutting it down to four trees (1 predictor set, 2 targets, 2 trees each), but in practicality, the trees for each target are a full and pruned version so they're not being weighed against each other.

In the end, the respective models with the best likelihood of addressing the client question were selected and will be discussed in detail in the modeling results section.

Data Splitting and Sampling

When considering data splitting and sampling, it is useful to consider why this should be done. First and foremost, sampling is done to be able to analyze large data sets without examining the entire population, but it is also used to build up data sets to a proper size for analysis. This is often done in conjunction with splitting when the split produces two sample sets that each are too small for use. This use is two-fold, used both for analysis purposes but also to test against each other to validate results (if accurate, should hold for both). This is to rule out confirmation bias within the results.

Due to the size of the entire data set for this project, data splitting is not an effective approach for the modeling techniques. If splitting and sampling were to be applied, the samples would become insignificant with little variation between them due to the amount of oversampling required to build up each split sample set. As a result, the sets would be so similar as to undermine the entire purpose of validation and testing one against would not disprove the effect of confirmation bias.

It should be noted that we did attempt to split and build up sets, but the splits were so small as to affect our ability to move forward with analysis and sampling was required to such an extent as to make validation meaningless as noted above. As a result, splitting and sampling were not successfully pursued for this project. This should be kept in mind when considering results and future applicability.

Model Results

Objective 1, time series was performed. For understanding the model, the data provided must be analyzed. The Residual Standard Error is measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term. Due to the presence of this error term, we are not capable of perfectly predicting our response variable from the predictor one. The Residual Standard Error is the average amount that the response will deviate from the true regression line. Simplistically, degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account residual parameters.

The R-squared statistic (R2) provides a measure of how well the model is fitting the actual data and it takes the form of a proportion of variance. The R2 is a measure of the linear relationship between our predictor variables and our response / target variable. It always lies between 0 and 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). In multiple regression settings, the R2 will always increase as more variables are included in the model. That is why the adjusted R2 is the preferred measure as it adjusts for the number of variables considered. Higher the R2, better the fit the model is.

Figure 27 – Objective 2: Multiple R-Squared & Adjusted R-Squared for all events

EVENTS	Members_Per_Event	TRAVEL	RIT	AWARDS	COMSERV	FUN	EDUCATION	MEMBERSHIP
AdjustedR^2	0.79	0.76	0.808	0.84	0.84	0.78	0.91	0.85
Multiple R^2	0.88	0.86	0.88	0.913	0.911	0.87	0.95	0.91

F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. However, how much larger the F-statistic needs to be depends on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H0: There is no relationship between speed and distance). The reverse is true as if the number of data points that are smaller, the larger F-statistic is required to be able to ascertain that there may be a relationship between predictor and response variables.

Consider the estimation results for equation for, Memebers_Per_Event,

Memebers_Per_Event.lag1, FUN.lag1 and MEMBERSHIP can enter the model at 10 percent significance level.

*Equation: Memebers_Per_Event = 0.53601*Memebers_Per_Event.lag1-0.29006*FUN.lag1+0.59273*MEMBERSHIP.*

Members_Per_Event increases by a factor 0.53601 by considering a previous lag(previous month) of Members_Per_Event. Events related to FUN in the previous lag (previous Month) have a negative impact of -0.29006. MEMBERSHIP has a positive impact on the Memebers_Per_Event by 0.59273.

Members_Per_Event increases by half a time by considering previous members in the event and as membership increases by 60 percent there is more number of members per event. Events related to fun in the previous month have 30 percent reduction in participation of members. Figure 28 is a portion of the VAR results that provides this information:

```
VAR Estimation Results:
=========================
Endogenous variables: Members_Per_Event, TRAVEL, RIT, AWARDS, COMSERV, FUN, EDUCATION,
MEMBERSHIP
Deterministic variables: const
Sample size: 54
Log Likelihood: -668.652
Roots of the characteristic polynomial:
0.9087 0.734 0.5552 0.5552 0.5203 0.5203 0.4501 0.3997 0.3997 0.3742 0.3742 0.2879 0.2879
0.2341 0.2341 0.1647
Call:
VAR(y = y, p = 2, type = c("const", "trend", "both", "none"),
    exogen = analysis_matrix, ic = c("AIC", "HQ", "SC", "FPE"))


Estimation results for equation Members_Per_Event:
==================================================
Members_Per_Event = Members_Per_Event.l1 + TRAVEL.l1 + RIT.l1 + AWARDS.l1 + COMSERV.l1 +
FUN.l1 + EDUCATION.l1 + MEMBERSHIP.l1 + Members_Per_Event.l2 + TRAVEL.l2 + RIT.l2 +
AWARDS.l2 + COMSERV.l2 + FUN.l2 + EDUCATION.l2 + MEMBERSHIP.l2 + const + MEMBERSHIP +
EDUCATION + FUN + COMSERV + AWARDS + RIT + TRAVEL

                    Estimate Std. Error t value Pr(>|t|)
Members_Per_Event.l1  0.53601    0.22194   2.415   0.0220 *
TRAVEL.l1             0.08145    0.11796   0.691   0.4952
RIT.l1                0.08002    0.06107   1.310   0.2000
AWARDS.l1            -0.10207    0.07537  -1.354   0.1857
COMSERV.l1            0.04027    0.11467   0.351   0.7279
FUN.l1               -0.29006    0.14267  -2.033   0.0510 .
EDUCATION.l1          0.16467    0.13754   1.197   0.2406
MEMBERSHIP.l1         0.18374    0.16122   1.140   0.2634
Members_Per_Event.l2  0.34886    0.21332   1.635   0.1124
TRAVEL.l2             0.01598    0.12644   0.126   0.9003
RIT.l2               -0.01597    0.05753  -0.278   0.7832
AWARDS.l2            -0.02299    0.07449  -0.309   0.7597
COMSERV.l2            0.01654    0.11387   0.145   0.8855
FUN.l2                0.15891    0.12835   1.238   0.2253
EDUCATION.l2         -0.12688    0.13042  -0.973   0.3384
MEMBERSHIP.l2        -0.03646    0.14208  -0.257   0.7992
const                 1.00556    1.11307   0.903   0.3735
MEMBERSHIP            0.59273    0.30503   1.943   0.0614 .
EDUCATION            -0.38437    0.74205  -0.518   0.6083
FUN                   0.23143    0.58322   0.397   0.6943
COMSERV              -0.17853    0.33513  -0.533   0.5982
AWARDS                0.24738    0.27842   0.889   0.3813
RIT                  -0.15714    0.67211  -0.234   0.8167
TRAVEL                0.11635    0.31122   0.374   0.7111
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 1.832 on 30 degrees of freedom
Multiple R-Squared: 0.8813, Adjusted R-squared: 0.7903
F-statistic: 9.685 on 23 and 30 DF,  p-value: 1.923e-08


Estimation results for equation TRAVEL:
=======================================
TRAVEL = Members_Per_Event.l1 + TRAVEL.l1 + RIT.l1 + AWARDS.l1 + COMSERV.l1 + FUN.l1 +
EDUCATION.l1 + MEMBERSHIP.l1 + Members_Per_Event.l2 + TRAVEL.l2 + RIT.l2 + AWARDS.l2 +
COMSERV.l2 + FUN.l2 + EDUCATION.l2 + MEMBERSHIP.l2 + const + MEMBERSHIP + EDUCATION + FUN +
COMSERV + AWARDS + RIT + TRAVEL

                     Estimate Std. Error t value Pr(>|t|)
Members_Per_Event.l1 -0.145994   0.178251  -0.819   0.4192
TRAVEL.l1             0.041050   0.094738   0.433   0.6679
RIT.l1                0.037563   0.049052   0.766   0.4498
AWARDS.l1            -0.039722   0.060532  -0.656   0.5167
COMSERV.l1           -0.044210   0.092101  -0.480   0.6347
FUN.l1               -0.147171   0.114590  -1.284   0.2089
```

Figure 28 – Objective 2: Portion of VAR Results

Objective 2: correlation between tenure as the target variable and members' geographic location as the independent variables, the classification regression model demonstrated insignificant values conveying that there is little impact on tenure based on where a member lives in relation to the Chapter. The second model applied membership as

the target variable and the census geographic IDs as the independent variables. With the only change being the target variable in the second classification regression, the model also proved to be insignificant. Therefore, there is little influence on membership participation based on where the member lives in accordance to the Chapter's location.

Modeling and analysis for Objective 3 – effect of age, member level and other survey results on organization and chapter perceptions – results in the consideration of two different model types – regression and classification tree.

The objective 3 regression models are as follows, remembering that Q21 is a target for chapter perception and Q24 is a target for organizational perception:

`Q21 = 0.321387 * Q10 + 0.265127 * Q20 + 0.405963 * Q23`, based on significant coefficients and an overall model fit (adjusted R-squared) of 0.9825 and highly significant p-value less than 2.2*10^-6.

This can be interpreted as saying that chapter perception is positively impacted by members' consideration of questions 10 (agree that self-development is tied to leadership), 20 (agreement with need to meet weekly) and 23 (agreement that there are sufficient events for their age group). It can be further interpreted as Q23 having the most influence, Q10 having roughly a middle influence and Q20 having the least influence among the significant predictors. This may also speak to the skew of each individual component as if one skews low then its multiplier would need to be larger to push the Q21 value higher, which is in line with our previous view of the target variable distributions.

This model was chosen as a result of two factors – it represented the best fit of all compared and while it doesn't address the client question specific to age and level, no other

Q21 regression models did either as all ruled them out based on lack of significant coefficients.

The final interpretation of this model is such that age and level have no significant influence on the chapter perception as compared to other predictors from the survey results. The output of the regression is below:

```
lm(formula = regdata$Q21 ~ 0 + regdata$Q1b + regdata$Q2 + regdata$Q10 +
    regdata$Q20 + regdata$Q23 + regdata$Q25mem + regdata$Q25edu)

Residuals:
     Min       1Q   Median       3Q      Max
-1.79008 -0.17477  0.05826  0.42294  0.98504

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
regdata$Q1b    -0.115624   0.103714  -1.115   0.2704
regdata$Q2      0.110851   0.076874   1.442   0.1557
regdata$Q10     0.321387   0.120399   2.669   0.0103 *
regdata$Q20     0.265127   0.096365   2.751   0.0083 **
regdata$Q23     0.405963   0.081228   4.998 7.79e-06 ***
regdata$Q25mem  0.042855   0.043509   0.985   0.3295
regdata$Q25edu -0.005662   0.040132  -0.141   0.8884
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5883 on 49 degrees of freedom
Multiple R-squared:  0.9847,    Adjusted R-squared:  0.9825
F-statistic: 450.3 on 7 and 49 DF,  p-value: < 2.2e-16
```

Additionally, Q24 is described as follows:

Q24 = **-0.15383** * Q2 + 0.44976 * Q20 + **-0.20340** * Q25mem + 3.56375 * Q25mem_1 + 5.07630 * Q25lcc_1 + 3.14715 * Q25rit_1 + 3.51611 * Q25fun_1 + 3.76541 * Q25edu_1 + 3.27437 * Q25com_1 + 3.95988 * Q25tra_1 + 0.74794 * age_13 + 0.51924 * lev_1 + 0.60896 * lev_2, based on significant coefficients and an overall model fit (adjusted R-squared) of 0.9877 and highly significant p-value less than 2.2*10^-6.

This can be interpreted as saying that organization perception (Q24) is somewhat negatively impacted by questions 2 (agreement to 1 business meeting per month) and

25mem (membership rank). It should be noted that this is likely a dampening effect to offset an overshoot caused by other factors in the regression, rather than a broad statement that those questions are negatively correlated with org perception. This is inferred because the negative relationship was not observed in any other regression models coupled with the very large influence of dummy variables as noted here. Additionally Q20 wields some influence as well (agreement to meet weekly). Finally with this regression, we see the influence of the dummy variables at play here. Each _1 indicator is a flag for a member having ranked that particular aspect (membership, fun, etc) as #1 among its peers. Lastly, the influence of age (for 13 year olds) and level (for 1s and 2s) is noticed as well with seemingly large bumps for each of those indicators, meaning 13 year old level 1s and 2s rank things typically much higher than their peers.

This regression model was chosen as a result of it offering the best fit of all others attempted. The final interpretation of the Q24 regression is that while it does show some influence from age and level, it is very limited with respect to age in that only 13 year olds seem to have that influence. The regression model output is observed below:

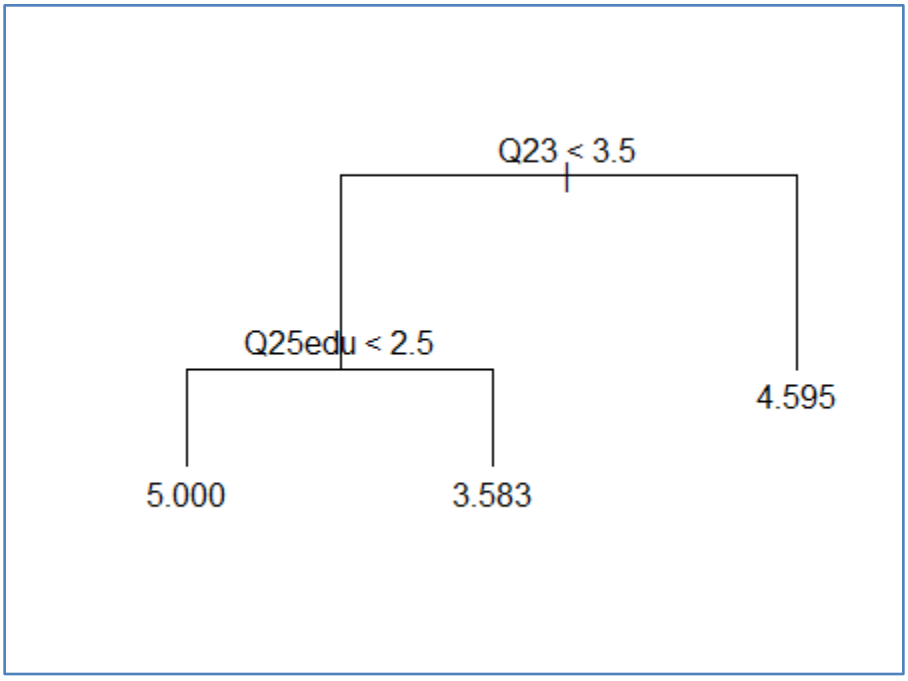
```
lm(formula = regdata$Q24 ~ 0 + regdata$Q1b + regdata$Q2 + regdata$Q10 +
    regdata$Q20 + regdata$Q23 + regdata$Q25mem + regdata$Q25edu +
    regdata$Q25mem_1 + regdata$Q25lcc_1 + regdata$Q25rit_1 +
    regdata$Q25fun_1 + regdata$Q25edu_1 + regdata$Q25com_1 +
    regdata$Q25tra_1 + regdata$age_12 + regdata$age_13 + regdata$age_14 +
    regdata$age_15 + regdata$age_16 + regdata$age_20 + regdata$lev_1 +
    regdata$lev_2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23988 -0.22819 -0.00441  0.32808  0.74278
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
regdata$Q1b      -0.01643    0.14267  -0.115 0.908996
regdata$Q2       -0.15383    0.08445  -1.822 0.077315 .
regdata$Q10       0.17944    0.16333   1.099 0.279657
regdata$Q20       0.44976    0.12417   3.622 0.000942 ***
regdata$Q23      -0.16077    0.12378  -1.299 0.202757
regdata$Q25mem   -0.20340    0.07921  -2.568 0.014806 *
regdata$Q25edu   -0.07043    0.05696  -1.237 0.224719
regdata$Q25mem_1  3.56375    0.91456   3.897 0.000435 ***
regdata$Q25lcc_1  5.07630    1.01393   5.007 1.68e-05 ***
regdata$Q25rit_1  3.14715    0.90480   3.478 0.001402 **
regdata$Q25fun_1  3.51611    0.88783   3.960 0.000363 ***
regdata$Q25edu_1  3.76541    0.98028   3.841 0.000509 ***
regdata$Q25com_1  3.27437    0.97939   3.343 0.002025 **
regdata$Q25tra_1  3.95988    0.94922   4.172 0.000197 ***
regdata$age_12   -0.31123    0.38336  -0.812 0.422520
regdata$age_13    0.74794    0.39475   1.895 0.066658 .
regdata$age_14    0.06606    0.33929   0.195 0.846775
regdata$age_15   -0.05014    0.34579  -0.145 0.885573
regdata$age_16    0.12763    0.36053   0.354 0.725528
regdata$age_20    0.29289    0.36763   0.797 0.431153
regdata$lev_1     0.51924    0.26952   1.927 0.062433 .
regdata$lev_2     0.60896    0.28963   2.103 0.042987 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5191 on 34 degrees of freedom
Multiple R-squared:  0.9925,    Adjusted R-squared:  0.9877
F-statistic: 205.4 on 22 and 34 DF,  p-value: < 2.2e-16
```

As noted previously, the other type of models considered are classification tree models. The pruned version of each tree is provided below.

Q21 (chapter perception):

```
> prune_q21tree2_class
node), split, n, deviance, yval
      * denotes terminal node

1) root 56 27.360 4.393
  2) Q23 < 3.5 14  8.357 3.786
    4) Q25edu < 2.5 2  0.000 5.000 *
    5) Q25edu > 2.5 12  4.917 3.583 *
  3) Q23 > 3.5 42 12.120 4.595 *
```

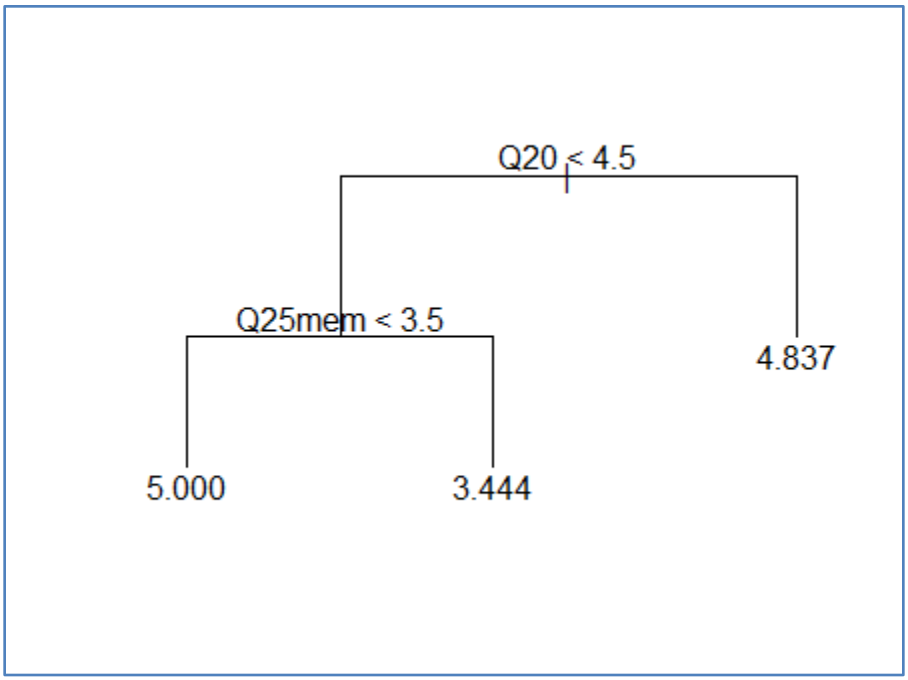


Interpretation of this tree takes into account both text and graph. Q23 is the first split with 42 members selecting 4 or higher and Q21 averaging 4.6 among them. For those that select 3 or less for Q23, the next branch is based on their ranking of education with those highly ranking it (<2.5) only being 2 people and always ranking Q21 a 5. Attention should

be drawn to those that rank education as > 2.5 (less important), as those folks have a much lower average (3.6) and the group contains 12 members.

Q24 (org perception):

```
> prune_q24tree2_class
node), split, n, deviance, yval
      * denotes terminal node

1) root 56 29.120 4.625
  2) Q20 < 4.5 13 10.920 3.923
    4) Q25mem < 3.5 4  0.000 5.000 *
    5) Q25mem > 3.5 9  4.222 3.444 *
  3) Q20 > 4.5 43  9.860 4.837 *
```



This tree should be evaluated in a similar fashion to the previous, noting the large number of high Q24 values split along Q20, but something to focus on being driven by those selecting 4 or less on Q20 and valuing Q25mem (importance of membership) less than the midpoint and ranking it a 4 or higher (worse). This group of members has 9 in the group with an average of 3.4 for Q24. As a result the chapter can learn what to focus on to keep

their progress up but also what to focus on to shift dissatisfied members but also to understand where there are diminishing returns in chasing one or two people in deeper levels (broader tree not displayed).

FINAL MODEL

For objective 1: correlation between time of year and Chapter events for participation, the final model selected was with VAR and it was applied to find correlation between variables. VAR Select() is a function that is used select the appropriate lag in the regression equation. Usually last 20 lags are considered for selecting the appropriate lag, later AIC, BIC, SC or FPE is used as a selection criteria. It is clear that lag=4 is the best lag as demonstrated by Figure 19. Also, the Roots() function provide the eigen value of companion matrix. VAR(p) model is stable if the characteristic polynomial has no roots in or on the complex circle, and this is equivalent to the condition that all eigen value of companion matrix have modulus less than 1. Even though the highest eigen value(0.915) is near 1, the model is stable.

```
> VARselect(log_data, lag.max = 20, type = "const" ,exogen=analysis_matrix)
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     4      4      4      5

$criteria
                  1             2             3    4    5    6    7    8    9   10   11
AIC(n)  -5.203163e+02  -5.117405e+02  -5.221588e+02 -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
HQ(n)   -5.187043e+02  -5.093762e+02  -5.190422e+02 -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
SC(n)   -5.156977e+02  -5.049666e+02  -5.132295e+02 -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
FPE(n)  1.561626e-226  2.287677e-222  1.276870e-225  NaN    0    0    0    0    0    0    0
         12   13   14   15   16   17   18   19   20
AIC(n) -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
HQ(n)  -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
SC(n)  -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
FPE(n)    0    0    0    0    0    0    0    0    0
```

```
> roots(var_result)
 [1] 0.9157663 0.6874756 0.5364735 0.5364735 0.4806319 0.4806319 0.3771113 0.3771113
 [9] 0.3401201 0.3141134 0.3141134 0.2672973 0.2672973 0.1808705 0.1808705 0.1558277
>
```

Figure 20 – Objective 1 VAR Final Model

Based on the modeling for objective 1, we can make the below conclusions:

1. Events that relate to travel is influenced by membership events and there is 47 percent more participation in travel events when there is more membership.

2. Events related to ritual are likely to be increased by 165 percent with participation in community service. Fun events during last month and the month before have a negative impact. There is drop of 60 percent and 70 percent events related ritual by fun events in the last two months.

3. Events related to community services decrease by 18 percent due to Fun events in the last month

4. Fun events drop by 10 percent due to events relate to Awards.

5. Events on education decrease by 12 percent and 13 percent due to events on ritual on previous two months. Community service events are likely to increase by 93 percent due to education events. Also, events on education increase by 23 percent due to educational events on the previous two months. Membership decreases by 56 percent due to educational events.

6. Membership events drop down by 31 percent due to events on travel.

With regards to objective 2: correlation between membership tenure and demographic/geographic data, we can conclude that there is little significance between tenure and geographic location of members' homes. Two models were performed, classification regression as the tenure as the target variable and a second model with membership participation as the second target variable. The final model selected for objective is selected based on the purpose of the data we are trying to compare. Both regression models performed were insignificant, therefore, both models state that tenure of members and members' participation are not influenced by members' living location in relation to the Chapter. Model 1 categorical regression is selected as the final modeling technique because it explains the relationship between members' geographic location and members' tenure. Though the regression proves that there is no influence on tenure based on members' geographic location, the final regression result is:

Tenure_yrs = 2.36 + 4.23cen_Tract_GeoID_40027202206 – 0.9044cen_Tract_GeoID_40119010101 + 0.1827cen_Tract_GeoID_40119010102 – 0.05cen_Tract_GeoID_40119010200 – 0.545en_Tract_GeoID_40119010300 + 0.1967en_Tract_GeoID_40119010500 + 0.9586cen_Tract_GeoID_40119010600 – 1.045cen_Tract_GeoID_40119010700 – 0.235cen_Tract_GeoID_40119010800 + 1.1cen_Tract_GeoID_40119010900 + 0.3525cen_Tract_GeoID_40119011000 + 0.06854cen_Tract_GeoID_40119011101

Overall, the classification regression validates that when determining how to extend membership tenure for the Chapter, geographic data is not a factor is consider.

Finally, regarding objective 3, we can put forth several observations and recommendations. First, a note on model selection. The model types developed were selected based on the question (influence of specific predictors) but also based on the perceived intent of the question (what levers should be pulled to drive satisfaction). It is our opinion that instead of one model being superior to the other that they are in fact complementary to one another. The previous stage of analysis left us with the best possible regression models for both target questions Q21 (chapter perception) and Q24 (organization perception) as well as associated classification trees for the same set of predictors, where the regression models selected were based on the best possible fit and also addressed the specific predictors inquired about where possible.

Observation 1 - Lack of Relationship – The specific question for objective 3 inquires as to the relationship of age and membership level with regard to perceptions of the chapter and organization. Based on our observations and regression modeling, these factors are insignificant for chapter perception (Q21) and have limited effect for organization perception (Q24). Furthermore, neither of those factors appear in the pruned tree for either question. As a result, we conclude that neither age nor level are as important as other factors from the survey responses.

Caveat 1 – A cautionary note - It should also be noted that we may be seeing the impact of survivor bias, where the only people polled are a biased sample set because the people that are no longer a part of the chapter were not surveyed. As a result our sample may suffer from selection bias due to the survivors being predisposed to higher satisfaction

than those that were not able to be surveyed because they were no longer participants. Such analysis is beyond the scope of this project but should be considered along with these results.

Recommendation 1 – The implied question – While the direct question with objective 3 revolves around relationships of predictors to targets, the implicit question is what should be done with that information. Based on our observations and particularly on the tree splits being so skewed and the histograms of Q21 and Q24 target variables, we recommend that the client consider both sides, not just what to do to move people from dissatisfied to satisfied but what to do to protect the very high number of members already ranking perceptions of org and chapter high. Based on the distribution of scores, results are highly skewed with 80%+ selecting a score of 4 or 5. While it is good to focus on detractors and what can be done to satisfy everyone, the imbalance here indicates attention should be paid to defending those already satisfied as the chapter has more downside risk than upside reward. Balancing the two will be important so as to not undo the good work that has led here.

Recommendation 2 – Model Selection – Instead of either/or, consider both/and. We recommend that the chapter utilize both the tree map and regression to understand not just what levers to pull or how sensitive perceptions are to those changes, but also in utilizing the tree or perhaps a deeper version of it, understand economies of scale and how many members stand to be impacted by the change. Said differently, while the regression can help predict the satisfaction of any single member based on this population data, the tree provides the ability to understand the impact to the group as a whole. By utilizing both in conjunction with one another, reasonable choices can be made understanding the cost/benefit of chasing

after a tree branch with only a few members associated versus another tree branch with a dozen members associated that could be moved from dissatisfied to satisfied. Lastly, utilizing these together will allow for both offense (satisfy the dissatisfied) and defense (protect those already satisfied) strategies to be balanced based on potential impact.

Recommendation 3 – Celebrate the win – While it is important to be vigilant on improving customer, consumer or member sentiment, it is also important to keep an eye on the objective and acknowledge success. The very high number of very satisfied members indicates that the chapter should recognize where it is on its success path and not lose sight of the fact they have achieved something significant. This can then be utilized to appreciate the gains they've made and move forward motivated to defend them.

CONCLUSION

All objectives detailed in initial project scoping have been addressed. Our final recommendation is that the chapter embrace and extend their data collection and monitoring in support of extending this and other analyses. It is our hope that the insights gained through this analysis will lead the chapter to further success and maintenance of same. All code has been submitted to the client as part of a working folder to aid in this regard as we conclude the team's analysis of the chapter's data.