2014 DISCOVERY SUMMIT
EXPLORING DATA
INSPIRING INNOVATION

Text Mining Feedback Comments from JMP® Users Group Meeting Participants

Melvin Alexander, Operations Research Analyst, Social Security Administration
Joshua Klick, Economist, Bureau of Labor Statistics

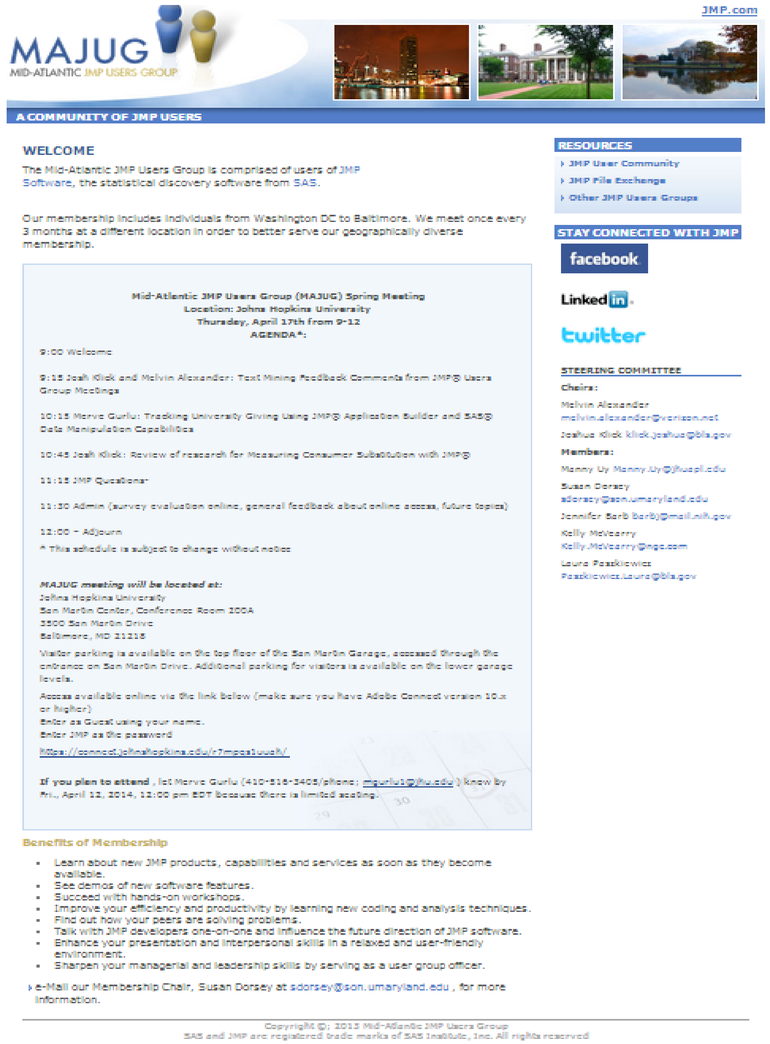
Abstract

- JMP Version 11 introduced the Free Text Command
- Allowed frequency counts of words in comment columns made by MAJUG meeting respondents.
- These word counts, coupled with text mining packages in R (via JSL), helped MAJUG leadership understand meeting aspects that attendees liked, disliked, or needed improvement.
- This presentation reviews ways feedback was useful in planning future meetings that met user-group member's needs.

Agenda

- Purpose: Plan better meetings from feedback comments using text analytic tools of JMP and R integration
- Review MAJUG meeting data and text-analytic methods
- Show returned output from R text mining packages (tm, SnowBallC, Slam, wordcloud, prcomp)
- Show results from JMP free-text analyses from the Analyze > Consumer Research > Categorical Platform
- Summary and Conclusions
- Q & A

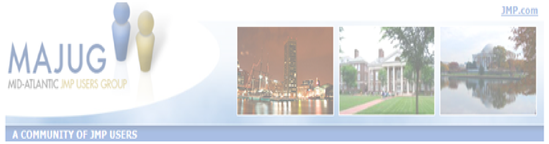
MAJUG meetings are held quarterly and posted on the www.majug.com website.



2014 DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

After each MAJUG meeting, attendees are asked to complete an Evaluation form.

The three questions on the bottom are usually open-ended, unstructured text that respondents write on.



Meeting Feedback/Evaluation Form ***Date:***

Please rate how much you agree with the following statements by circling your response using the scale provided, where 1 = strongly disagree (SD) and 5 = strongly agree (SA).

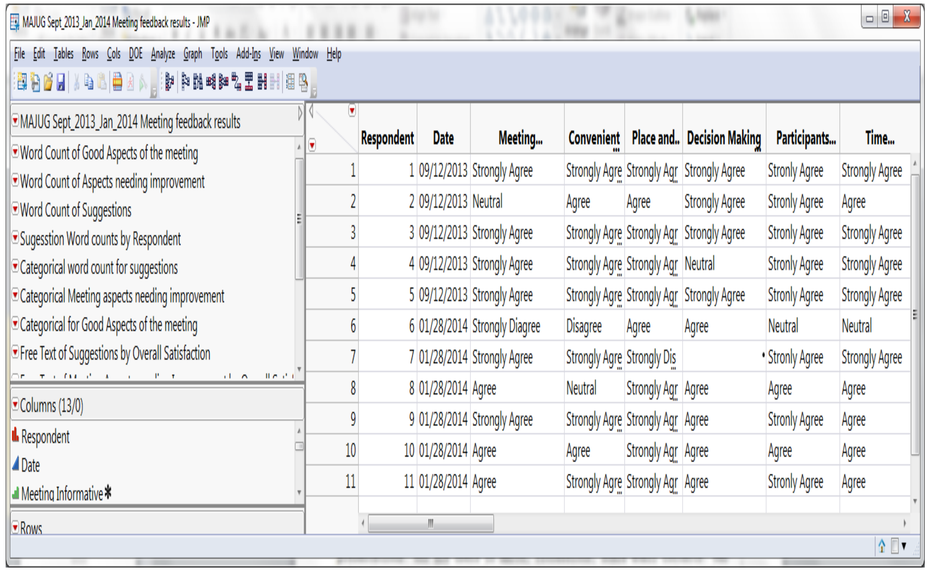
	SD	D	N	A	SA
The meeting topic and program were informative.	1	2	3	4	5
Meeting time was convenient.	1	2	3	4	5
Meeting place and venue were convenient for me.	1	2	3	4	5
Decision-making was shared at this meeting.	1	2	3	4	5
All meeting participants were actively involved.	1	2	3	4	5
Meeting time was used effectively.	1	2	3	4	5
I was satisfied with this meeting overall.	1	2	3	4	5
I enjoyed this meeting.	1	2	3	4	5

What aspects of this meeting were particularly good?

What aspects of this meeting need improvement?

Do you have any suggestions or additional comments about this meeting (use back if necessary)?

A Sample JMP Data Table of Respondent's Feedback from MAJUG Meeting Evaluations is Shown below



	Respondent	Date	Meeting...	Convenient	Place and..	Decision Making	Participants...	Time...
1	1	09/12/2013	Strongly Agree	Strongly Agre	Strongly Agr	Strongly Agree	Stronly Agree	Strongly Agree
2	2	09/12/2013	Neutral	Agree	Agree	Strongly Agree	Stronly Agree	Agree
3	3	09/12/2013	Strongly Agree	Strongly Agre	Strongly Agr	Strongly Agree	Stronly Agree	Strongly Agree
4	4	09/12/2013	Strongly Agree	Strongly Agre	Strongly Agr	Neutral	Stronly Agree	Strongly Agree
5	5	09/12/2013	Strongly Agree	Strongly Agre	Strongly Agr	Strongly Agree	Stronly Agree	Strongly Agree
6	6	01/28/2014	Strongly Diagree	Disagree	Agree	Agree	Neutral	Neutral
7	7	01/28/2014	Strongly Agree	Strongly Agre	Strongly Dis		Stronly Agree	Strongly Agree
8	8	01/28/2014	Agree	Neutral	Strongly Agr	Agree	Agree	Agree
9	9	01/28/2014	Strongly Agree	Strongly Agre	Strongly Agr	Agree	Stronly Agree	Agree
10	10	01/28/2014	Agree	Agree	Strongly Agr	Agree	Agree	Agree
11	11	01/28/2014	Agree	Strongly Agre	Strongly Agr	Agree	Stronly Agree	Agree

MAJUG Meeting Evaluation Comments about Suggested Improvements (text vector of JMP's Suggestions column that was sent to R)

```
Data Table( "MAJUG Sept_2013_Jan_2014 Meeting feedback results" )
//:*/
SuggEX = Column("Suggestions") << Get Values ;
/*:

{
"Improvements on the MAJUG site. Adding previous presentations to the
website, best sources to learn JMP (online or books), and maybe tips and
tricks of using JMP.",
"Rotating location and WebEx access is important",
"next meeting email list of who is planning to come. list of topics of
interest and discuss",
"Time savings. laundry list of topics. Data analytics. Best practices, how to
best summarize. review issues, problems. email beforehand - I'm coming JMP
presentations. Who are users in MAJUG, Professions, share email contacts. Web
value (increase usefulness) What papers/presentations have occurred at MAJUG",
"Query members planning to attend what they want to get out of the meeting so
their concerns, questions, issues can be addressed and discussed",
"", "", "Please start at 10",
"Have coffee break with coffee, more communication between meetings,
suggesting topics",
"MAJUG should have a fee (perhaps $5) to buy refreshments so participants can
get coffee without leaving meeting",
""}
```

Selected Suggestions output from the JMP Log of the returned R Term-Document-Matrix

```
Non-/sparse entries: 113/932
Sparsity           : 89%
Maximal term length: 14
Weighting          : term frequency (tf)

                  Docs
Terms              1  2  3  4  5  6  7  8  9 10 11
  $5)              0  0  0  0  0  0  0  0  0  1  0
  (increase        0  0  0  1  0  0  0  0  0  0  0
  (online          1  0  0  0  0  0  0  0  0  0  0
  (perhaps         0  0  0  0  0  0  0  0  0  1  0
  access           0  1  0  0  0  0  0  0  0  0  0
  adding           1  0  0  0  0  0  0  0  0  0  0
  addressed        0  0  0  0  1  0  0  0  0  0  0
  analytics.       0  0  0  1  0  0  0  0  0  0  0
  are              0  0  0  1  0  0  0  0  0  0  0
  attend           0  0  0  0  1  0  0  0  0  0  0
  beforehand       0  0  0  1  0  0  0  0  0  0  0
  best             1  0  0  2  0  0  0  0  0  0  0
  between          0  0  0  0  0  0  0  0  1  0  0
  books),          1  0  0  0  0  0  0  0  0  0  0
  break            0  0  0  0  0  0  0  0  1  0  0
  buy              0  0  0  0  0  0  0  0  0  1  0
  can              0  0  0  0  1  0  0  0  0  1  0
```

Panel below shows how words were correlated (similar) with the word "jmp"

```
 without          0 0 0 0 0 0 0 0 0  1  0
 [1] "best"            "can"             "coffee"          "email"
 [5] "get"             "have"            "jmp"             "list"
 [9] "majug"           "meeting"         "planning"        "presentations"
[13] "topics"          "what"            "who"
 presentations            best             majug       (increase          (online
          1.00            0.94              0.77           0.67              0.67
        adding       analytics.              are     beforehand           books),
          0.67            0.67              0.67           0.67              0.67
        coming        contacts.             data            how               i'm
          0.67            0.67              0.67           0.67              0.67
  improvements          issues,             jmp.        laundry             learn
          0.67            0.67              0.67           0.67              0.67
         majug,           maybe         occurred         papers        practices,
          0.67            0.67              0.67           0.67              0.67
presentations.        previous         problems.   professions,            review
          0.67            0.67              0.67           0.67              0.67
      savings.           share             site.        sources         summarize.
          0.67            0.67              0.67           0.67              0.67
          time            tips           topics.         tricks        usefulness)
          0.67            0.67              0.67           0.67              0.67
         users           using             value            web           website,
          0.67            0.67              0.67           0.67              0.67
         email
          0.56
[1] 95 11
[1]  0 11
```

2014 DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

Selected Suggestions output from the JMP Log of the returned R Document-Term-Matrix

Docs refer to individual respondents. Numbers below each term are the word counts used by each respondent.
Here, 89% of the terms were sparse (i.e., had zeroes in the cells).

```
A document-term matrix (11 documents, 78 terms)

Non-/sparse entries: 93/765
Sparsity           : 89%
Maximal term length: 10
Weighting          : term frequency (tf)

    Terms
Docs access adding address analytics attend beforehand best books break buy can
  1      0      1       0         0      0          0    1     1     0   0   0
  2      1      0       0         0      0          0    0     0     0   0   0
  3      0      0       0         0      0          0    0     0     0   0   0
  4      0      0       0         1      0          1    2     0     0   0   0
  5      0      0       1         0      1          0    0     0     0   0   1
  6      0      0       0         0      0          0    0     0     0   0   0
  7      0      0       0         0      0          0    0     0     0   0   0
  8      0      0       0         0      0          0    0     0     0   0   0
  9      0      0       0         0      0          0    0     0     1   0   0
  10     0      0       0         0      0          0    0     0     0   1   1
  11     0      0       0         0      0          0    0     0     0   0   0
    Terms
Docs coffe coffee come communic concerns contacts data discuss email fee get
  1      0      0    0        0        0        0    0       0     0   0   0
  2      0      0    0        0        0        0    0       0     0   0   0
  3      0      0    1        0        0        0    0       1     1   0   0
  4      0      0    1        0        0        1    1       0     2   0   0
  5      0      0    0        0        1        0    0       1     0   0   1
```

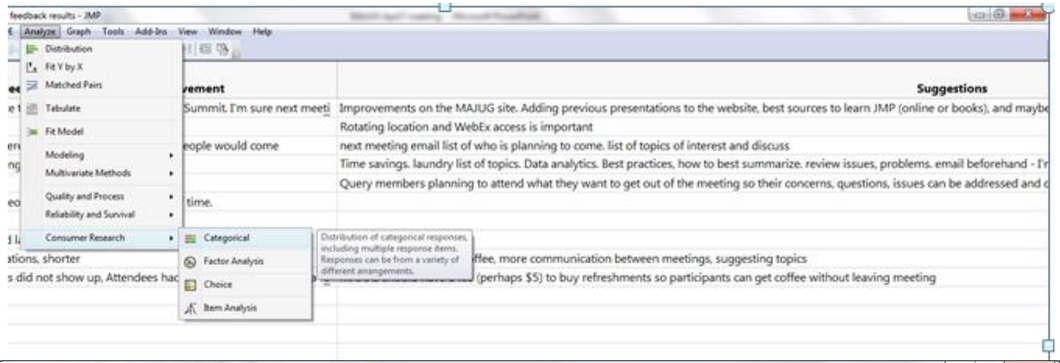
Selected Suggestions output from the JMP Log of the returned Principal Components from R (52% of the amount of information the text was contained in principal components PC1 and PC2)

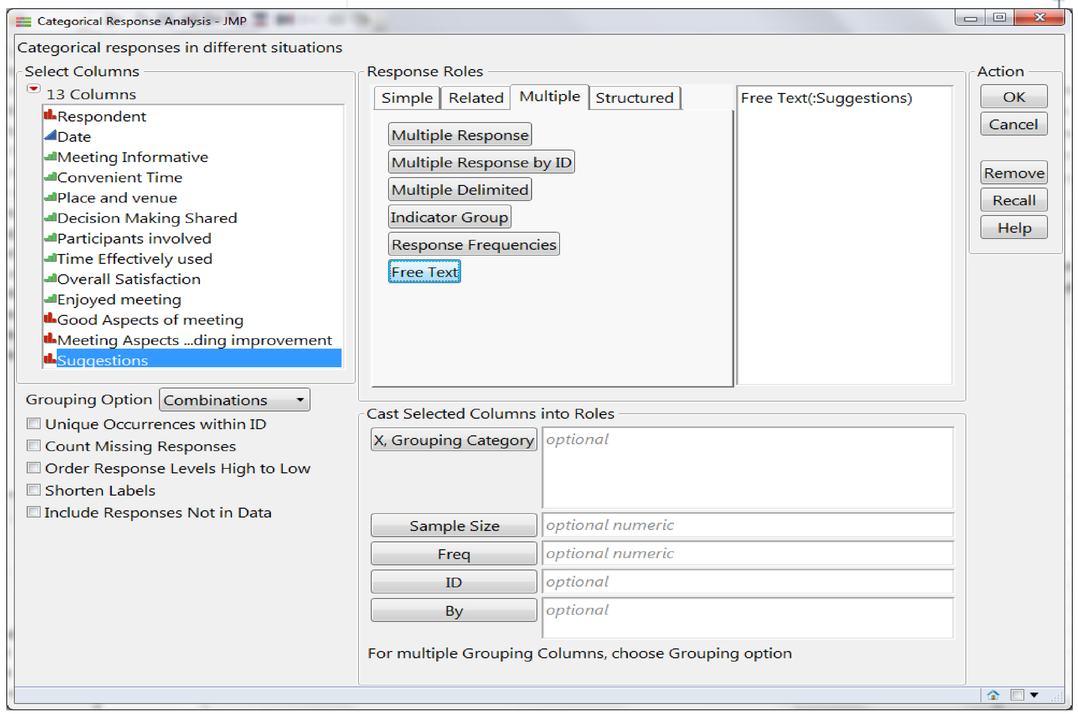
```
Standard deviations:
 [1] 5.071554e+00 3.885909e+00 3.229279e+00 3.001896e+00 2.618718e+00
 [6] 2.355371e+00 1.973126e+00 1.200308e+00 1.049532e-15 3.540632e-16
[11] 6.916402e-32

Rotation:
                     PC1          PC2          PC3         PC4           PC5
access     -0.020553545 -0.013184378  0.076350679  0.05655256 -0.2231014189
adding     -0.014737615  0.248229095 -0.066283135 -0.01748287  0.0323961321
address    -0.033997790 -0.079393501 -0.263347038  0.08049331  0.0599546853
analytics   0.196173283 -0.010958619 -0.015238216 -0.00777396  0.0115480501
attend     -0.033997790 -0.079393501 -0.263347038  0.08049331  0.0599546853
beforehand  0.196173283 -0.010958619 -0.015238216 -0.00777396  0.0115480501
best        0.176061081  0.105518447 -0.045114381 -0.01540069  0.0258733867
books      -0.014737615  0.248229095 -0.066283135 -0.01748287  0.0323961321
break      -0.024458975 -0.024544210  0.138717410  0.04001953  0.3093564957
buy        -0.026381560 -0.049421586 -0.020581754 -0.32050519 -0.0008638968
can        -0.045004111 -0.096013097 -0.211628027 -0.17889429  0.0440436733
coffe      -0.037894298 -0.055130849  0.088053119 -0.20906166  0.2299368072
coffee     -0.024458975 -0.024544210  0.138717410  0.04001953  0.3093564957
come        0.138840967 -0.034828303 -0.014037002  0.04894002  0.0147575239
communic   -0.024458975 -0.024544210  0.138717410  0.04001953  0.3093564957
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     5.0716 3.8859 3.2293 3.0019 2.61872 2.35537 1.97313
Proportion of Variance 0.3297 0.1936 0.1337 0.1155 0.08792 0.07113 0.04991
Cumulative Proportion  0.3297 0.5233 0.6570 0.7726 0.86049 0.93162 0.98153
                           PC8      PC9      PC10      PC11
Standard deviation     1.20031 1.05e-15 3.541e-16 6.916e-32
Proportion of Variance 0.01847 0.00e+00 0.000e+00 0.000e+00
Cumulative Proportion  1.00000 1.00e+00 1.000e+00 1.000e+00
0
```

2014 DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

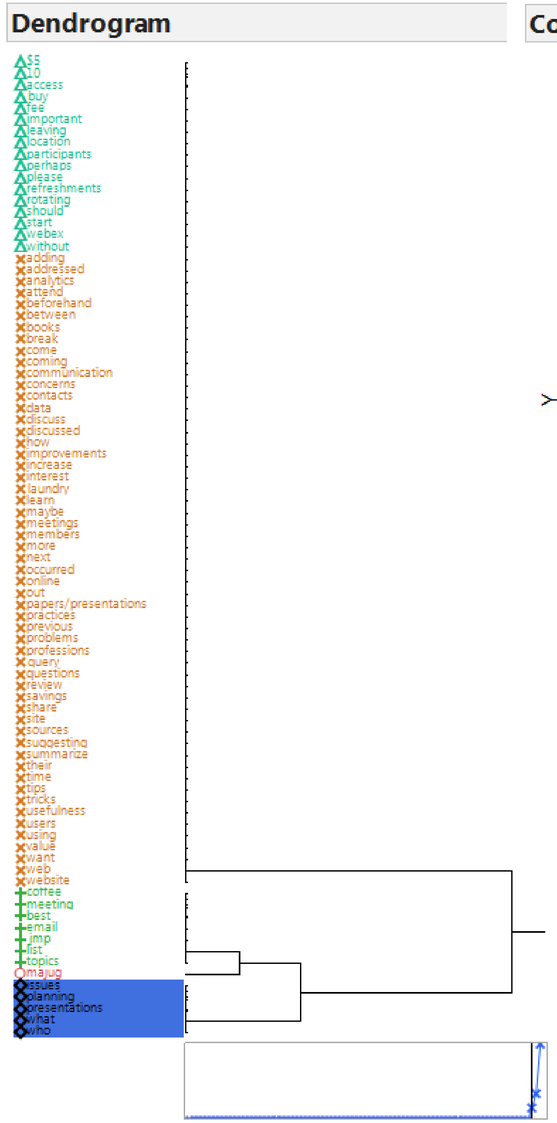
Free-Text Analyses done in JMP by choosing Analyze > Consumer Research > Categorical.

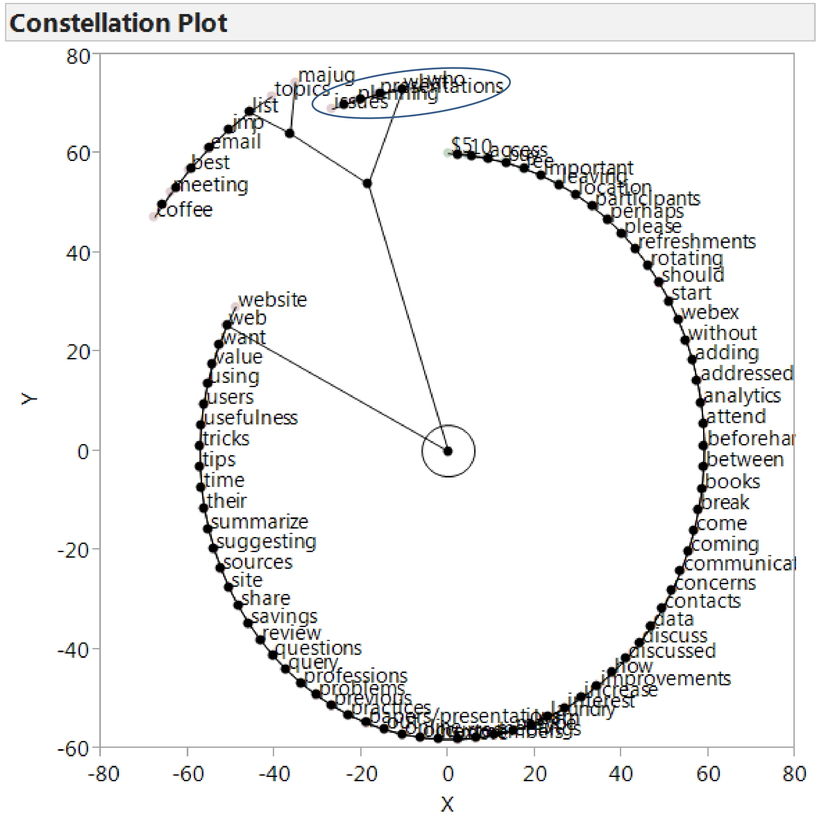
Select Free Text (from Multiple tab) and click OK.





Cluster Dendogram and Constellation Plot of MAJUG Suggestions From JMP (Constellation plot shows how terms were interrelated within clusters. Terms close together tended to load high on the principal components also).

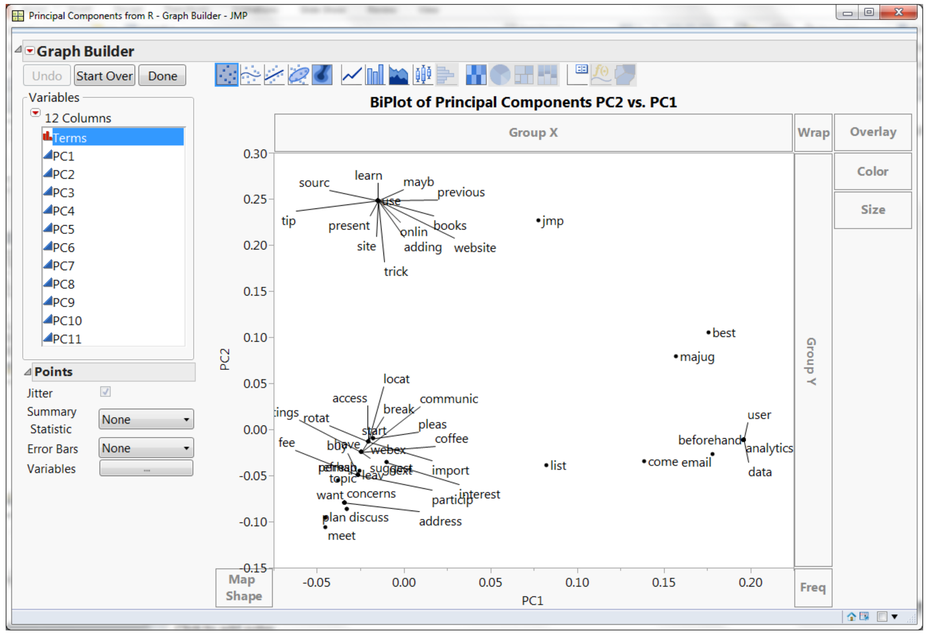




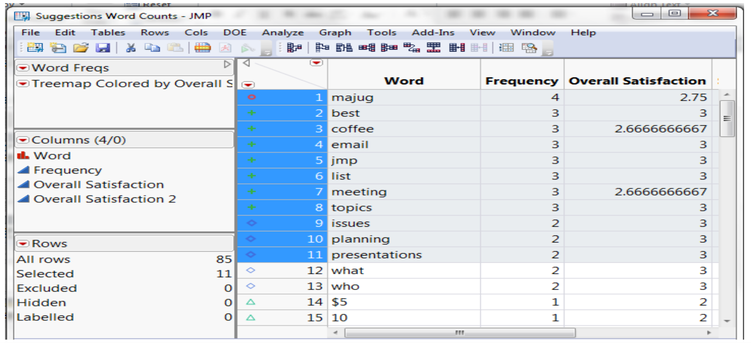
2014 DISCOVERY SUMMIT

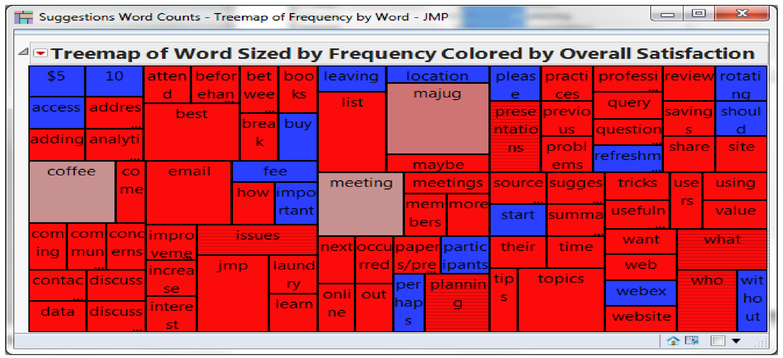
EXPLORING DATA INSPIRING INNOVATION

BiPlot of the Suggestions Principal Components PC2 by PC1 from the Principal Components Data Table



Word Frequencies and Treemap of MAJUG Suggestions From JMP

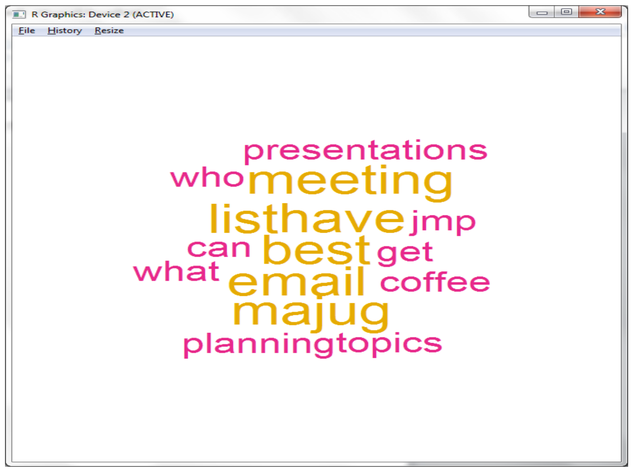




Conclusions

- JMP's Free-Text tools and R help MAJUG leadership identify meeting aspects more analytically.
- Although the sample was small, we were able to explore modern ways to deliver JMP content to MAJUG members.
- MAJUG leadership is testing different content-delivery mechanisms.
- Challenge is to get more knowledgeable expertise, volunteers, time, and resources to provide better value to the MAJUG community.
- The Call for Help continues…

Word Cloud of MAJUG Suggestions from JMP in R



Contact:

Klick.Joshua@bls.gov

Melvin.Alexander@ssa.gov

References

1. Karl, A.T. and Rushing, H. (2013), "Text Mining in JMP with R," http://jmp.com/about/events/summit2013/resources/Paper_Karl_Rushing.pdf
2. Francis, L. and Flynn, M (2010), *Text Mining Handbook*, Casualty Actuarial Society Forum, Spring, 1-60.

2014 DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

JSL Script To Text Mine Suggestions from the Respondent's Feedback of MAJUG Meeting Evaluations

```
Data Table("MAJUG Sept_2013_Jan_2014 Meeting feedback results") ;
SuggEx = column("Suggestions") << Get Values ;
R Init() ;
R Send(SuggEX);
R Submit("\[
library(wordcloud)
library(tm)
library(SnowballC)
Suggestion <- matrix(SuggEX)
 TEXTFILE = Corpus(VectorSource(c(Suggestion)))
  inspect(TEXTFILE)
 newstopwords <- c("and", "for", "the", "to", "of", "in", "as", "is", "with",
 "an", "then", "by", "they", "than", "he", "she")
 skipWords <- function(x) removeWords(x, stopwords("english"))
funcs <- list(tolower, removePunctuation, removeNumbers, stripWhitespace,
stemDocument, skipWords)
corpus2.proc <- tm_map(TEXTFILE, FUN = tm_reduce, tmFuns = funcs)
#remove new stopwords
TEXTFILE <-tm_map(TEXTFILE, removeWords, newstopwords)
# remove unwanted characters from the text
for(j in 1:length(TEXTFILE)) TEXTFILE[[j]] <- gsub("/", " ",  TEXTFILE[[j]])
for(j in 1:length(TEXTFILE)) TEXTFILE[[j]] <- gsub("-", " ",  TEXTFILE[[j]])
#Document Term Matrix
corpus2a.dtm <- DocumentTermMatrix(corpus2.proc, control =
list(wordLengths = c(3,10)))
dtmDataFrame <- as.data.frame(inspect(corpus2a.dtm))
# find frequent terms which occured more than 2 times
findFreqTerms(corpus2a.dtm, 2)
# Find Associations between words with high association (50%) with jmp
findAssocs(corpus2a.dtm, "jmp", 0.5)
#Term Document Matrix
corpus2b.tdm <- TermDocumentMatrix(TEXTFILE)
inspect(corpus2b.tdm)
# find frequent terms which occured more than 2 times
findFreqTerms(corpus2b.tdm, 2)
# Find Associations between words with high association
(50%) with
# jmp
findAssocs(corpus2b.tdm, "jmp", 0.5)
tdm.common = removeSparseTerms(corpus2b.tdm, 0.1)
# Examine the dimensions of corpus2b.tdm and
#the removed sparse terms in tdm.common
dim(corpus2b.tdm)
dim(tdm.common)
#get principal components
require(stats)
prcomponent <- prcomp(corpus2a.dtm, scale=TRUE,
retx=TRUE, center=TRUE)
prcomponent
summary(prcomp(corpus2a.dtm, scale=TRUE, retx=TRUE,
center=TRUE))
loadings <- prcomponent$rotation  ## note that blank
entries may be small but not zero
loadingnames <- colnames(corpus2a.dtm)
scores <- prcomponent$x
head (scores, n=85)
pcadtm <- as.data.frame(scores)]\");

tdmclust = R Get Graphics(png) ;
New Window ("TDM Cluster", tdmclust ) ;
R Submit("\[
# produce Word cloud of most frequent terms sorted by
decreasing frequency
m <- as.matrix(corpus2b.tdm)
wordFreq <- sort(rowSums(m), decreasing=TRUE)
set.seed(350)
sugwordcloud <-
wordcloud(words=names(wordFreq),freq=wordFreq,
min.freq=2, colors=brewer.pal(6,"Dark2"),
random.order=FALSE) ]\");

//RPCA = R Get (scores ) ;
RPCA = R Get (loadings) ;
As Table (RPCA) << Set Name ("Principal Components
from R") ;
R_Loadingnames = R Get (loadingnames) ;
Data Table("Principal Components from R") << New
Column( "Terms", Character, Nominal, Set
Values(R_Loadingnames))
  << Move Selected Columns({"Terms"}, To First) ;
Data Table ("Principal Components from R");
col = Col1 << Set Name("PC1");
col = Col2 << Set Name("PC2");
col = Col3 << Set Name("PC3");
col = Col4 << Set Name("PC4");
col = Col5 << Set Name("PC5");
col = Col6 << Set Name("PC6");
col = Col7 << Set Name("PC7");
col = Col8 << Set Name("PC8");
col = Col9 << Set Name("PC9");
col = Col10 << Set Name("PC10");
col = Col11 << Set Name("PC11");
Column ("Terms")  << Label(1) ;
/** Produce BiPlot of Principal Components PC2 by
PC1 **/
Data Table ("Principal Components from R") << Graph
Builder( Size( 778, 652 ),
Variables( X( :PC1 ), Y( :PC2 ) ), Elements( Points( X, Y,
Legend( 1 ), Jitter( 1 ) ) ),
SendToReport( Dispatch( {}, "graph title", TextEditBox,
Set Text( "BiPlot of Principal Components PC2 vs. PC1"
)} ),
{Dispatch( {}, "Graph Builder", FrameBox,
{DispatchSeg( Marker Seg( 1 ), label offset({0, -45, -
49}, {1, 33, 64}, {2, 94, 29}, {7, 70, 37}, {8, 28, -48},
{9, -39, -31}, {12, 93, -11}, {14, 74, -61}, {17, 6, 46},
{20, -100, -35},
{23, 80, 42},  {25, 91, 46}, {30, -29, -26}, {33, 19,
-73}, {35, 32, -18},
{37, -142, -44}, {41, 28, 45}, {42, 93, 37}, {45, 57,
-12}, {47, -62, 37},
{48, 75, -5}, {54, -82, -24},  {57, -26, 63}, {58, -99,
-17}, {60, 11, 26},
{63, -121, 31}, {64, -38, 15}, {66, 8, 95}, {69, 5, -
26}, {74, 96, 65} ) )} ) ) );

/* Use Categorical Free Text option on Suggestions
*/
Data Table("MAJUG Sept_2013_Jan_2014 Meeting
feedback results") <<  Categorical(Free Text(
:Meeting Aspects needing improvement,
<<Score Words by Column( :Overall Satisfaction
),<<Save Word Table ), Crosstab Transposed( 1 ),
Legend( 0 ));
Data Table ("Word Freqs")  << Set Row States( [8,
8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8,
8, 8, 8, 8,8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8,
8, 8, 8, 8, 8]) << Set Label Columns( :Word ) ;
Data table ( "Word Freqs" ) << Set
Name("Suggestions Word Counts") <<
Hierarchical Cluster(Columns( :Frequency, :Overall
Satisfaction),Label( :Word ),
Color Clusters( 1 ),Mark Clusters( 1 ),
Number of Clusters( 5 ),Constellation Plot( 1 ) );
 *Create Treemap */
Data Table ("Suggestions Word Counts") <<
Treemap( Categories( :Word ),Coloring( :Overall
Satisfaction ),Sizes( :Frequency ));
R Term() ;
```