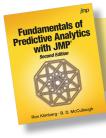


Fundamentals of Predictive Analytics with JMP[®] Second Edition



Ron Klimberg · B. D. McCullough



Fundamentals of Predictive Analytics with JMP[®], Second Edition. Full book available for purchase <u>here</u>.

Contents

About This Book	xiii
About These Authors	xvii
Acknowledgments	xiv
Chapter 1: Introduction	1
Historical Perspective	1
Two Questions Organizations Need to Ask	2
Return on Investment	2
Cultural Change	2
Business Intelligence and Business Analytics	2
Introductory Statistics Courses	4
The Problem of Dirty Data	5
Added Complexities in Multivariate Analysis	5
Practical Statistical Study	6
Obtaining and Cleaning the Data	6
Understanding the Statistical Study as a Story	7
The Plan-Perform-Analyze-Reflect Cycle	7
Using Powerful Software	8
Framework and Chapter Sequence	9
Chapter 2: Statistics Review	
Introduction	11
Fundamental Concepts 1 and 2	12
FC1: Always Take a Random and Representative Sample	12
FC2: Remember That Statistics Is Not an Exact Science	13
Fundamental Concept 3: Understand a Z-Score	14
Fundamental Concept 4	15
FC4: Understand the Central Limit Theorem	15
Learn from an Example	17
Fundamental Concept 5	
Understand One-Sample Hypothesis Testing	

Consider p-Values	21
Fundamental Concept 6:	22
Understand That Few Approaches/Techniques Are Correct—Many Are Wrong	22
Three Possible Outcomes When You Choose a Technique	32
Chapter 3: Dirty Data	33
Introduction	33
Data Set	34
Error Detection	35
Outlier Detection	39
Approach 1	39
Approach 2	41
Missing Values	43
Statistical Assumptions of Patterns of Missing	44
Conventional Correction Methods	45
The JMP Approach	48
Example Using JMP	49
General First Steps on Receipt of a Data Set	54
Exercises	54
Chapter 4: Data Discovery with Multivariate Data	57
Introduction	57
Use Tables to Explore Multivariate Data	59
PivotTables	59
Tabulate in JMP	62
Use Graphs to Explore Multivariate Data	64
Graph Builder	64
Scatterplot	68
Explore a Larger Data Set	71
Trellis Chart	71
Bubble Plot	73
Explore a Real-World Data Set	78
Use Graph Builder to Examine Results of Analyses	79
Generate a Trellis Chart and Examine Results	80
Use Dynamic Linking to Explore Comparisons in a Small Data Subset	83
Return to Graph Builder to Sort and Visualize a Larger Data Set	84
Chapter 5: Regression and ANOVA	87
Introduction	87
Regression	88

Perform a Simple Regression and Examine Results	
Understand and Perform Multiple Regression	
Understand and Perform Regression with Categorical Data	101
Analysis of Variance	107
Perform a One-Way ANOVA	108
Evaluate the Model	109
Perform a Two-Way ANOVA	122
Exercises	129
Chapter 6: Logistic Regression	131
Introduction	
Dependence Technique	
The Linear Probability Model	
The Logistic Function	
A Straightforward Example Using JMP	135
Create a Dummy Variable	135
Use a Contingency Table to Determine the Odds Ratio	136
Calculate the Odds Ratio	138
A Realistic Logistic Regression Statistical Study	147
Understand the Model-Building Approach	147
Run Bivariate Analyses	150
Run the Initial Regression and Examine the Results	152
Convert a Continuous Variable to Discrete Variables	154
Produce Interaction Variables	155
Validate and Use the Model	157
Exercises	158
Chapter 7: Principal Components Analysis	159
Introduction	159
Basic Steps in JMP	160
Produce the Correlations and Scatterplot Matrix	160
Create the Principal Components	
Run a Regression of y on Prin1 and Excluding Prin2	163
Understand Eigenvalue Analysis	163
Conduct the Eigenvalue Analysis and the Bartlett Test	
Verify Lack of Correlation	
Dimension Reduction	165
Produce the Correlations and Scatterplot Matrix	165
Conduct the Principal Component Analysis	165

v

Determine the Number of Principal Components to Select	166
Compare Methods for Determining the Number of Components	167
Discovery of Structure in the Data	167
A Straightforward Example	168
An Example with Less Well Defined Data	169
Exercises	171
Chapter 8: Least Absolute Shrinkage and Selection Operator and Elastic Net	173
Introduction	173
The Importance of the Bias-Variance Tradeoff	174
Ridge Regression	
Least Absolute Shrinkage and Selection Operator	
Perform the Technique	180
Examine the Results	180
Refine the Results	182
Elastic Net	182
Perform the Technique	182
Examine the Results	183
Compare with LASSO	184
Exercises	184
Chapter 9: Cluster Analysis	185
Introduction	185
Example Applications	186
An Example from the Credit Card Industry	186
The Need to Understand Statistics and the Business Problem	
Hierarchical Clustering	187
Understand the Dendrogram	187
Understand the Methods for Calculating Distance between Clusters	187
Perform a Hierarchal Clustering with Complete Linkage	189
Examine the Results	191
Consider a Scree Plot to Discern the Best Number of Clusters	193
Apply the Principles to a Small but Rich Data Set	193
Consider Adding Clusters in a Regression Analysis	196
K-Means Clustering	197
Understand the Benefits and Drawbacks of the Method	197
Choose k and Determine the Clusters	198
Perform k-Means Clustering	200

Change the Number of Clusters	202
Create a Profile of the Clusters with Parallel Coordinate Plots	204
Perform Iterative Clustering	206
Score New Observations	207
K-Means Clustering versus Hierarchical Clustering	208
Exercises	209
Chapter 10: Decision Trees	211
Introduction	211
Benefits and Drawbacks	212
Definitions and an Example	213
Theoretical Questions	213
Classification Trees	214
Begin Tree and Observe Results	215
Use JMP to Choose the Split That Maximizes the LogWorth Statistic	217
Split the Root Node According to Rank of Variables	218
Split Second Node According to the College Variable	219
Examine Results and Predict the Variable for a Third Split	221
Examine Results and Predict the Variable for a Fourth Split	221
Examine Results and Continue Splitting to Gain Actionable Insights	222
Prune to Simplify Overgrown Trees	222
Examine Receiver Operator Characteristic and Lift Curves	223
Regression Trees	224
Understand How Regression Trees Work	225
Restart a Regression Driven by Practical Questions	228
Use Column Contributions and Leaf Reports for Large Data Sets	229
Exercises	230
Chapter 11: k-Nearest Neighbors	231
Introduction	231
Example—Age and Income as Correlates of Purchase	232
The Way That JMP Resolves Ties	234
The Need to Standardize Units of Measurement	234
k-Nearest Neighbors Analysis	235
Perform the Analysis	235
Make Predictions for New Data	236
k-Nearest Neighbor for Multiclass Problems	237
Understand the Variables	237
Perform the Analysis and Examine Results	238

The k-Nearest Neighbor Regression Models	239
Perform a Linear Regression as a Basis for Comparison	239
Apply the k-Nearest Neighbors Technique	239
Compare the Two Methods	240
Make Predictions for New Data	242
Limitations and Drawbacks of the Technique	242
Exercises	243
Chapter 12: Neural Networks	245
Introduction	
Drawbacks and Benefits	
A Simplified Representation	
A More Realistic Representation	
Understand Validation Methods	
Holdback Validation	
k-fold Cross-Validation	251
Understand the Hidden Layer Structure	252
A Few Guidelines for Determining Number of Nodes	252
Practical Strategies for Determining Number of Nodes	252
The Method of Boosting	253
Understand Options for Improving the Fit of a Model	254
Complete the Data Preparation	255
Use JMP on an Example Data Set	
Perform a Linear Regression as a Baseline	
Perform the Neural Network Ten Times to Assess Default Performance	
Boost the Default Model	259
Compare Transformation of Variables and Methods of Validation	
Exercises	
Chapter 13: Bootstrap Forests and Boosted Trees	267
Introduction	
Bootstrap Forests	
Understand Bagged Trees	
Perform a Bootstrap Forest	
Perform a Bootstrap Forest for Regression Trees	274
Boosted Trees	275
Understand Boosting	275
Perform Boosting	

Perform a Boosted Tree for Regression Trees	
Use Validation and Training Samples	
Exercises	
Chapter 14: Model Comparison	
Introduction	285
Perform a Model Comparison with Continuous Dependent Variable	
Understand Absolute Measures	
Understand Relative Measures	
Understand Correlation between Variable and Prediction	
Explore the Uses of the Different Measures	
Perform a Model Comparison with Binary Dependent Variable	290
Understand the Confusion Matrix and Its Limitations	290
Understand True Positive Rate and False Positive Rate	
Interpret Receiving Operator Characteristic Curves	
Compare Two Example Models Predicting Churn	
Perform a Model Comparison Using the Lift Chart	296
Train, Validate, and Test	299
Perform Stepwise Regression	299
Examine the Results of Stepwise Regression	302
Compute the MSE, MAE, and Correlation	302
Examine the Results for MSE, MAE, and Correlation	303
Understand Overfitting from a Coin-Flip Example	303
Use the Model Comparison Platform	305
Exercises	
Chapter 15: Text Mining	319
Introduction	
Historical Perspective	320
Unstructured Data	320
Developing the Document Term Matrix	
Understand the Tokenizing Stage	323
Understand the Phrasing Stage	329
Understand the Terming Stage	330
Observe the Order of Operations	
Developing the Document Term Matrix with a Larger Data Set	
Generate a Word Cloud and Examine the Text	333
Examine and Group Terms	334
Add Frequent Phrases to List of Terms	335
Parse the List of Terms	336

Using Multivariate Techniques	
Perform Latent Semantic Analysis	
Perform Topic Analysis	
Perform Cluster Analysis	
Using Predictive Techniques	
Perform Primary Analysis	
Perform Logistic Regressions	
Exercises	
Chapter 16: Market Basket Analysis	355
Introduction	
Association Analyses	
Examples	
Understand Support, Confidence, and Lift	
Association Rules	
Support	
Confidence	
Lift	
Use JMP to Calculate Confidence and Lift	
Use the A Priori Algorithm for More Complex Data Sets	
Form Rules and Calculate Confidence and Lift	
Analyze a Real Data Set	
Perform Association Analysis with Default Settings	
Reduce the Number of Rules and Sort Them	
Examine Results	
Target Results to Take Business Actions	
Exercises	
Chapter 17: Statistical Storytelling	
The Path from Multivariate Data to the Modeling Process	
Early Applications of Data Mining	
Numerous JMP Customer Stories of Modern Applications	
Definitions of Data Mining	
Data Mining	
Predictive Analytics	
A Framework for Predictive Analytics Techniques	
The Goal, Tasks, and Phases of Predictive Analytics	
The Difference between Statistics and Data Mining	
SEMMA	

References	
Index	

xi

From Fundamentals of Predictive Analytics with JMP[®], Second Edition, by Ron Klimberg and B. D. McCullough. Copyright © 2016, SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED.

Fundamentals of Predictive Analytics with JMP®, Second Edition. Full book available for purchase here.

Chapter 15: Text Mining

Fundamentals of Predictive Analytics

with JMP

Introduction	319
Historical Perspective	
Unstructured Data	
Developing the Document Term Matrix	321
Understand the Tokenizing Stage	
Understand the Phrasing Stage	
Understand the Terming Stage	
Observe the Order of Operations	
Developing the Document Term Matrix with a Larger Data Set	
Generate a Word Cloud and Examine the Text	
Examine and Group Terms	
Add Frequent Phrases to List of Terms	
Parse the List of Terms	
Using Multivariate Techniques	338
Perform Latent Semantic Analysis	
Perform Topic Analysis	
Perform Cluster Analysis	
Using Predictive Techniques	347
Perform Primary Analysis	
Perform Logistic Regressions	
Exercises	

Introduction

The growth of the amount of data available in digital form has been increasing exponentially. Bernard Marr, in his September 30, 2015, Forbes Magazine Tech post listed several "mindboggling" facts (Marr):

- "The data volumes are exploding, [and] more data has been created in the past two years • than in the entire previous history of the human race."
- "Data is growing faster than ever before and by the year 2020, about 1.7 megabytes of new information will be created every second for every human being on the planet."
- "By then, our accumulated digital universe of data will grow from 4.4 zettabytes today to around 44 zettabytes, or 44 trillion gigabytes."

Historical Perspective

To put this 44 trillion gigabytes forecast into perspective, in 2011 the entire print collection of the Library of Congress was estimated to be 10 terabytes (Ashefelder). The projected 44 trillion gigabytes is approximately 4.4 billion Libraries of Congress.

Historically, because of high costs and storage, memory, and processing limitations, most of the data stored in databases were structured data. Structured data were organized in rows and columns so that they could be loadable into a spreadsheet and could be easily entered, stored, queried and analyzed. Other data that could not fit into this organized structure were stored on paper and put in a file cabinet.

Today, with the cost and the limitation barriers pretty much removed, this other "file cabinet" data, in addition to more departments and more different types of databases within an organization, are being stored digitally. Further, so as to provide a bigger picture, organizations are also accessing and storing external data.

A significant portion of this digitally stored data is unstructured data. Unstructured data are not organized in a predefined matter. Some examples of types of unstructured data include responses to open-ended survey questions, comments and notes, social media, email, Word, PDF and other text files, HTML web pages, and messages. In 2015, IDC Research estimated that unstructured data account for 90% of all digital data (Vijayan).

Unstructured Data

With so much data being stored as unstructured data or as text (*unstructured data* will be considered to be synonymous with *text*), why not leverage the text, as you do with structured data, to improve decisions and predictions? This is where text mining comes in. Text mining and data mining are quite similar processes in that their goal is to extract useful information so as to improve decisions and predictions. The main difference is that text mining extracts information from text while data mining extracts information from structured data.

Both text mining and data mining initially rely on preprocessing routines. Since the data have already been stored in a structured format, the preprocessing routines in a data mining project focus on cleaning, normalizing the data, finding outliers, imputing missing values, and so on. Text mining projects also first require the data to be cleaned. However, differently, text mining projects use natural language processes (a field of study related to human-computer interaction) to transform the unstructured data into a more structured format. Subsequently, both text mining and data mining processes use visualization and descriptive tools to better understand the data and apply predictive techniques to develop models to improve decisions and provide predictions.

Text mining is categorized, as shown in our multivariate analysis framework in Figure 15.1, as one of the interdependence techniques, although text mining also includes the elements of discovery and possibly dependence techniques.

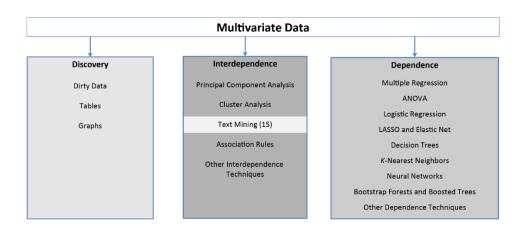


Figure 15.1: A Framework for Multivariate Analysis

Text mining and the process of text mining have several definitions, extensions, and approaches. The text mining process consists of three major steps:

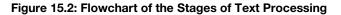
- Developing the document term matrix. The document term matrix (DTM) is a set of zero and 1 variables (also called *indicator variables*) that represent the words in the text. Natural language processing techniques are used to initially develop the DTM. Subsequently, you explore the set of variables and curate the DTM, by grouping words or removing infrequent words, until you are satisfied.
- 2. Using multivariate techniques. Text visualization and the text multivariate techniques of clustering, principal components, and factor analysis (PCA/FA) (similar to the continuous multivariate techniques discussed in Chapters 4, 7, and 9) are used to understand the composition of the DTM.
- 3. Using predictive techniques. If a dependent variable exists, you can use the text multivariate analysis results (along with other structured data) as independent variables in a predictive technique.

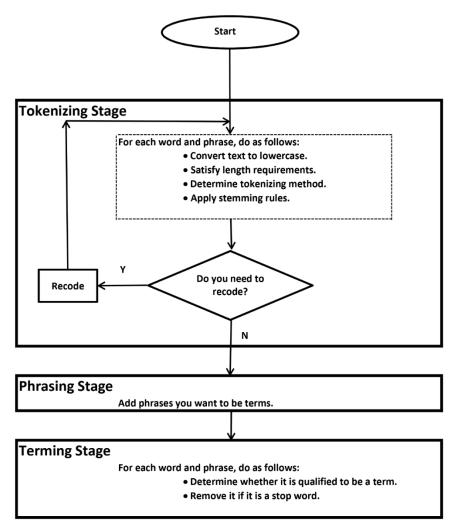
More and more of today's digital data are unstructured data. In this chapter, how the text mining process can leverage information from this unstructured data or text to enhance your understanding of the text, as well as to improve your decisions and predictions, will be discussed.

Developing the Document Term Matrix

The Text Explorer platform in JMP uses a bag of words approach.¹ The order of words is ignored except for phrases, and the analysis is based on the count of words and phrases. The words are processed in three stages to develop the DTM as shown in Figure 15.2:

- 1. tokenizing
- 2. phrasing
- 3. terming





To understand this process of transforming the text into a structured format, open a small data set called toytext.jmp, which contains just one column of text in 14 rows as shown in Figure 15.3.

File Edit Tables	Rows Cols	DOE Analyze (Graph Tools	View Window	w Help	
i 🖽 🔁 🚰 🗔 X	- 🗈 👛 📕 👯) 🗄 🎫 📄 💆	≽ ⊻ 📒			
▼toytext D	<					
Source		text				
		woman				
		Men				
	3	man				
		women				
Columns (1/0)		woman dress				
text	6	women DRESSES				
	7	men cars				
	8	man car				
	9	women shoes				
	10	men beer				
	11	women dresses				
	12	men cars				
	13	mn				
Rows	14	wn				
All rows 14						
Selected 0						
Excluded 0						
Hidden 0						
Labelled 0						

Figure 15.3: Data Table of toytext.jmp File

Each row of words in the variable **text** column is called a *document*. Hence, the toytext.jmp file has 14 documents. The entire set of these 14 documents is called a *corpus*.

Understand the Tokenizing Stage

To access the Text Explorer platform in JMP, complete the following steps:

- 1. Select Analyze ► Text Explorer. The Text Explorer dialog box appears as shown in Figure 15.4.
- 2. Under 1 Columns, click text and click the box Text Columns, and the variable text will be listed.

Figure	15.4:	Text	Explorer	Dialog	Box
--------	-------	------	----------	--------	-----

Text Explorer - JMP Pro					
Select Columns			Cast Selected Co Text Columns	olumns into Roles —	Action OK Cancel
Language	English	•	By	optional	
Maximum Words per Phrase		4			Remove
Maximum Number of Phrases		5000			Recall
Minimum Characters per Word		1			
Maximum Characters per Word		50			Help
Stemming	No Stemming	•			
Tokenizing	Regex	•			
Customize Regex					
Treat Numbers as Words					
					1 □ ▼

Select Options in the Text Explorer Dialog Box

The list of options offered in Figure 15.4 include the following:

- Maximum Words per Phrase, Maximum Number of Phrases and Maximum Characters per Word. As you progress down the flowchart and curate the DTM, you will decide what these limits of words and phrases will be.
- Stemming. Stemming combines related terms with common suffixes, essentially combining words with identical beginnings (called *stems*), but different endings. The stemming process in JMP uses the Snowball string processing language, which is described at http://snowballstem.org. The drop-down arrow provides three options:
 - No Stemming. No terms are combined.
 - Stem for Combining. Terms are stemmed when two or more terms stem to the same term. For example, in your toytext.jmp data set, dress, Dress, and dresses would be stemmed to dress. JMP uses a dot (.) to denote a word's being stemmed.
 - Stem All Terms. All terms are stemmed.
- **Tokenizing**. This is the method used to parse the body of text into terms or tokens. The drop-down arrow provides two options:
 - Basic Words. Text is parsed into terms by using a set of delimiters (such as white space, money, time, URLs, or phone numbers) that typically surround words. To view the default set of delimiters, click the red triangle, select Display Options ► Show Delimiters, and click OK after Text Explorer has been run.
 - Regex (which is short for *regular expression*). Text is decomposed using a set of built-in regular expressions. Regex is an advanced feature (beyond the scope of this book) and a very powerful tool for using regular expressions to identify patterns in the text. The Regex option is a superset of the Basic Words option. That is, when Regex is selected, in addition to the default regular expressions provided by the

Regex option, the default delimiters included by the **Basic Words** option are also included. Furthermore, if **Regex** is selected and you want to add, delete, or edit this set of regular expressions, click the **Customize Regex** check box. Once you click **OK**, the **Regular Expression Editor** dialog box will appear. For more information about building your own **Regex**, see <u>www.regular-expressions.info</u>.

For this example, use the following options:

- 1. In terms of **Minimum Characters per Word**, to avoid words such as "a," "an," and so on, you would normally use at least 2 (and usually 2); but in this small example leave it at the default value of **1**.
- 2. Stem for Combining is recommended; but, initially, with this small example, use No Stemming.
- 3. Select Regex.
- 4. Click OK.

These Text Explorer dialog box selections are the components in the dashed box within the Tokenizing Stage box in Figure 15.2.

The Text Explorer output box will appear as shown in Figure 15.5.

Figure 15.5: Text Explorer Output Box

	ber Numbe						
				Number of Non		tion N	
ofler			•	empty Case:		•	
	12 14	4 22	1.57143	14	4	1.0	000
4 Terr	n and Phra	se l ists					
	i ana i nic						
Term	Count			Phrase	Count	N	
men	4			men cars	2		
wome				women dresses	2	2	
cars	2						
dress							
man	2						
woma							
beer	1						
car	1						
dress	1						
mn	1						
shoes	1						
2110-02							

At the top of Text Explorer output (Figure 15.5), some summary statistics are provided.

Each document is broken into initial units of text called *tokens*. Usually, a token is a word, but it can be any sequence of non-whitespace characters. As shown in Figure 15.5, there are 22 total tokens.

326 Fundamentals of Predictive Analytics with JMP, Second Edition

The basic unit of analysis for text mining is a *term*. Initially, the Text Explorer examines each of the tokens to determine a possible set of useful terms for analysis. As shown in Figure 15.5, the number of initial terms is 12; they are listed below on the left side and sorted by frequency. This number of terms will change as you transform the text.

On the right side of Figure 15.5, there is a list of phrases common to the corpus—the phrases "men cars" and "women dresses" occurred twice. A phrase is defined as a sequence of tokens that appear more than once. Each phrase will be considered as to whether it should be a term.

Terms are the units of analysis for text mining. Presently, since you have yet to do anything to the data set and if you want to analyze it, complete the following steps:

- 1. Click the **Text Explorer for text** red triangle, and select **Save Document Term Matrix**; accept all the default values.
- 2. Click OK.

The data table now has 12 new indicator variables, one for each term as shown in Figure 15.6. As you can see in Figure 15.6, one of the first steps that the Text Explorer module does is to convert all the terms to lowercase. In particular, note that **dresses** and **DRESSES** are considered as the same term. By default, the Text Explorer also treats the plural of terms, such as **dress** and **dresses** or **men** and **man**, as different terms or units.

	text	men Binary	women Binary	cars Binary	dresses Binary	man Binary	woman Binary	beer Binary	car Binary	dress Binary	mn Binary	shoes Binary	wn Binary
1	woman	0	0	0	0	0	1	0	0	0	0	0	0
2	Men	1	0	0	0	0	0	0	0	0	0	0	0
3	man	0	0	0	0	1	0	0	0	0	0	0	0
4	women	0	1	0	0	0	0	0	0	0	0	0	0
5	woman dress	0	0	0	0	0	1	0	0	1	0	0	0
6	women DRESSES	0	1	0	1	0	0	0	0	0	0	0	0
7	men cars	1	0	1	0	0	0	0	0	0	0	0	0
8	man car	0	0	0	0	1	0	0	1	0	0	0	0
9	women shoes	0	1	0	0	0	0	0	0	0	0	1	0
10	men beer	1	0	0	0	0	0	1	0	0	0	0	0
11	women dresses	0	1	0	1	0	0	0	0	0	0	0	0
12	men cars	1	0	1	0	0	0	0	0	0	0	0	0
13	mn	0	0	0	0	0	0	0	0	0	1	0	0
14	wn	0	0	0	0	0	0	0	0	0	0	0	1

The current number of 14 terms, or indicator variables, is probably more than what you want to work with. You most likely want to combine some of these terms as well as clean up the data before you proceed. (For now, to delete all 12 indicator variables, select all the indicator variable columns, right-click, and click **Delete Columns**.)

Recode to Correct Misspellings and Group Terms

Examining Figure 15.5 and 15.6 further, you can see that **women** and **woman** as well as **men** and **man** were not combined. You can also see that there are two misspellings: **mn** and **wn**.

To correct the misspelling of mn, click mn in the Text Explorer output, Figure 15.5:

- 1. Right-click the term **mn** in the **Text Explorer output box** and click **Recode**.
- 2. As shown in Figure 15.7, in the New Values box, enter men.
- 3. Click Done.

This misspelling is now included with the term **men**. The count for **men** should have increased from 1 to 5. Also, check the data table. Although **mn** is now included with **men** within the Text Explorer platform, it is still coded as **mn** in the data table.

Figure 15.7: Recode Dialog Box

📝 Recode			×
⊿ ▼ Term		*	Done
Count Old Values (1)	New Values (1)		Cancel
1 mn	men		Undo Redo
			Filter
			Show only Grouped Show only Ungrouped
			Group
		*	Help

To group together the terms men and man, complete the following steps:

- 1. Click the term **men** in the Text Explorer output box, hold down the **Ctrl** key, and click **man**.
- 2. Right-click and click **Recode**. The Recode dialog box appears.
- 3. As shown in Figure 15.8, highlight man and men, and right-click.
- 4. Click Group To men and click Done. The count for men has increased from 2 to 7.

Figure 15.8: The Recode Dialog Box

📝 Recode				×
⊿ ▼ Terr	n		*	Done
Count	Old Values (2)	New Values (2)		Cancel
	man	man		
5	men	men		Undo Redo
				Filter
				Show only Grouped
				Show only Ungrouped
				Group
			*	Help

Similarly, recode wn to women, and group women and woman. The Text Explorer output box will look like Figure 15.9.

Figure 15.9: Text Explorer Output Box

toytext - Te			JMP Pro			• X
Text E	xplorer f	or text				
				Number of Non- empty Cases		
8	14	22	1.57143	14	ļ.	1.0000
⊿ Term a	nd Phras	e Lists				
Term	Count			Phrase	Count	Ν
men women cars dresses beer car dress shoes	7 7 2 1 1 1 1			men cars women dresses		2

The process of combining related terms is called *stemming*. Stemming combines related terms with common suffixes—combining words with identical beginnings (called *stems*), but different endings. To accomplish this, complete the following steps:

- 1. Click the **Text Explorer for text** red triangle.
- 2. Select Term Options ► Stemming ► Stem for Combining.

The Text Explorer output should look similar to Figure 15.10.

Figure 15.10: Text Explorer Output Box after Stemming

Number	Number	Total	Tokens	Number of Non-	Port	tion N
ofTerms	of Cases	Tokens	per Case	empty Cases	empty	per C
6	14	22	1.57143	14	Ļ	1.0
T	I Diana a	- 1 !				
i erm a	nd Phras	e Lists				
Term	Count			Phrase	Count	Ν
men	7			men cars		2
women	7			women dresses	2	2
cari	3					
dress-	3					
beer	1					
shoes	1					

As shown in Figure 15.10, the terms **car** and **cars** have been combined into the one new term **car** and similarly for **dress** and **dresses**. You can check this by clicking one of the stemmed terms **car** or **dress**, right-clicking, and then clicking **Show Text**.

The recoding of terms thus far completed applies only within the Text Explorer platform; that is, the data is not changed in the data table. To verify, click to open the data table and observe that **mn, man, wn**, and **woman** are still listed.

Recoding does affect stemming and should occur before stemming. Hence, it is important that you should try to recode all misspelling and synonyms before executing the Text Explorer platform. Use the **Recode** procedure under the **Cols** option. The terms **woman dress** and **man car** will also need to be recoded.

Understand the Phrasing Stage

If you want any of the phrases to be analyzed as individual concepts and separated from their individual terms, then you can add these phrases to your term list. For example, to add the phrases **men cars** and **women dresses** to your term list, complete the following steps:

- 1. Click **men cars**, hold down the **Shift** key, and click **women dresses** under the list of Phrases.
- 2. Right-click, and then click Add Phrase.

The two phrases are now added to the list of terms. They were both stemmed with the plural phrase that appeared only once as shown in Figure 15.11. They are also dimmed in the phrase list, indicating that they are being treated as terms.

Figure 15.11: Text Explorer Output Box after Phrasing	9	
tovtext - Text Explorer of text - IMP Pro		x

					Number of Non- empty Cases			
(5	14	22	1.57143	14		1.00	00
Term		Cou	nt		Phrase	(Count	
Term		Cou	nt		Phrase		Count	N
men women			4		women dre	sses	2 2	
men car women beer shoes			3 3 1					

330 Fundamentals of Predictive Analytics with JMP, Second Edition

Examining Figure 15.11 further, you can see that the instances of the term **men** that were in the term **men car** have been removed from the count of **men** and similarly for the term **women**. To clearly see what you have done so far, complete the following steps:

- 1. Click the **Text Explorer for text** red triangle.
- 2. Select Save Document Term Matrix.
- 3. Click OK.

Added to the data table are 6 indicator variables, down from your initial 12, one for each term in Figure 15.11, as shown in Figure 15.12.

	text	men Binary	women Binary	men car• Binary	women dress- Binary	beer Binary	shoes Binary
1	woman	0	1	0	0	0	0
2	Men	1	0	0	0	0	0
3	man	1	0	0	0	0	0
4	women	0	1	0	0	0	0
5	woman dress	0	0	0	1	0	0
6	women DRESSES	0	0	0	1	0	0
7	men cars	0	0	1	0	0	0
8	man car	0	0	1	0	0	0
9	women shoes	0	1	0	0	0	1
10	men beer	1	0	0	0	1	0
11	women dresses	0	0	0	1	0	0
12	men cars	0	0	1	0	0	0
13	mn	1	0	0	0	0	0
14	wn	0	1	0	0	0	0

Figure 15.12: Document Text Matrix

Understand the Terming Stage

Stop words are words that can be characterized in one or more of the following ways:

- too common, such as "a,", "an," or "the";
- infrequent (their counts are low); or
- ignorable (not relevant to the analysis).

Create Stop Words

As shown in Figure 15.11, you have 2 terms with counts of 1. To make them into stop words, complete the following steps:

- 1. Click **beer** under the list of Terms.
- 2. Hold down the **Shift** key and click **shoes**.
- 3. Right-click and then click Add Stop Word.

The list of terms now excludes the terms **beer** and **shoes** as shown in Figure 15.13. To see the list of stop words, click the **Text Explorer for text** red triangle, and select **Display Options** \triangleright **Show Stop Words**.

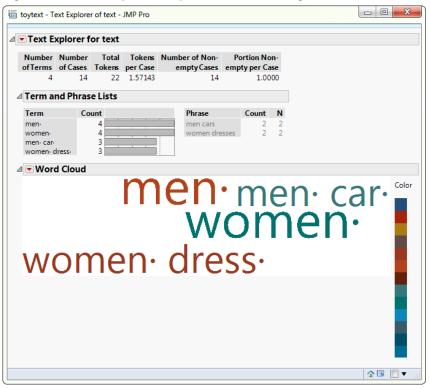


Figure 15.13: Text Explorer Output Box after Terming

Generate a Word Cloud

A visualization of the list of terms is called a *Word Cloud*. To produce the Word Cloud as shown in Figure 15.13, complete the following steps:

- 1. Click the **Text Explorer for text** red triangle.
- 2. Select Display Options ► Show Word Cloud.
- 3. Click the red triangle next to Word Cloud.
- 4. Select Layout ► Centered.
- 5. Again, click the red triangle next to Word Cloud.
- 6. Click Coloring ► Arbitrary Colors.

The Word Cloud is added to the Text Explorer output as in Figure 15.13. The size of each term in the Word Cloud is relative to its frequency.

Observe the Order of Operations

The list of terms, on the left side of Figure 15.13, shows the list of indicator variables that are used in creating the DTM. As you have worked your way through the flowchart (see Figure 15.2), your objective has been to examine and explore the list of terms and phrases to produce a final list of terms that you are satisfied with. There are no definitive approaches to take, particular words to focus on (depending on the objective of the study and domain expertise), nor a definitive measure to say that you have a good list of terms. This is also an iterative process. However, you should be

332 Fundamentals of Predictive Analytics with JMP, Second Edition

aware that the order of operation in the creation of a list of terms can affect the resulting list of terms.

The following general steps are suggested:

- 1. Before executing the Text Explorer, recode all misspellings and synonyms in the data table.
- 2. In the Text Explorer dialog box, select these options:
 - a. Minimum number of characters (use 2)
 - b. Stemming (select Stem for Combining)
 - c. Tokenizing (select **Regex** unless you have some custom regexs that you want to add)
- 3. In the Phrasing Stage, do the following:
 - a. Examine the phrases.
 - b. Specify which phrases you want to be included as terms; in particular, select the most frequent sequence of phrases.
- 4. In the Terming stage, do the following:
 - a. Remove stop words.
 - b. Remove least frequent terms.
 - c. Remove too frequent terms (if any).

Developing the Document Term Matrix with a Larger Data Set

Now you will examine a larger and more realistic data set. The data set traffic_violations_dec2014.jmp contains all the electronic traffic violations that occurred in Montgomery County, Maryland, during December 2014 (dataMontgomery: "All electronic traffic violations"). The file contains 16,446 records and 35 columns. The 35 variables are as follows; their dictionary can be found at (dataMontgomery: "Variable dictionary"):

- Date of Stop
- Time of Stop
- Agency
- SubAgency
- Description
- Location
- Latitude
- Longitude
- Accident
- Belts
- Personal Injury
- Property Damage
- Fatal
- Commercial License
- Hazmat
- Commercial Vehicle

- Alcohol
- Work Zone
- State
- Vehicle Type
- Year
- Make
- Model
- Color
- Violation Type
- Charge
- Article
- Contributed to Accident
- Race
- Gender
- Driver City
- Driver State
- DL State
- Arrest Type
- Geolocation

Generate a Word Cloud and Examine the Text

Examine the text in the variable field **Description**:

- 1. Select Analyze ► Text Explorer.
- 2. In the Text Explorer dialog box under 1 Columns, click Description.
- 3. Click the box Text Columns; change the Minimum Characters per Word to 2.
- 4. Click the drop-down arrow for Stemming, and choose Stem for Combining.
- 5. Click OK.
- 6. In the Text Explorer output, click the **Text Explorer for Description** red triangle.
- 7. Click Display Options ► Show Word Cloud.
- 8. Click the red triangle next to Word Cloud,
- 9. Click Layout ► Centered, and again click the red triangle next to Word Cloud.
- 10. Click Coloring ► Arbitrary Colors. The Text Explorer output box with the Word Cloud will appear as shown in Figure 15.14.

Figure 15.14: Text Explorer Output Box

Examine and Group Terms

Under the list of phrases, find the phrase **motor vehicle** and complete the following steps:

- 1. Click the phrase **motor vehicle**.
- 2. Right-click and then click **Select Contains**. This option selects bigger phrases that contain this phrase. Scroll down the list of phrases on the right side of the Text Explorer

output and you can see highlighted the other phrases that contain the phrase **motor** vehicle.

Similarly, under the list of phrases, scroll down until you find the phrase **driving vehicle on highway** and complete the following steps:

- 1. Click the phrase driving vehicle on highway.
- 2. Right-click and this time click **Select Contained**. Scroll up and down the list of phrases; highlighted are all the phrases that contain one or more of the terms in **driving vehicle on highway**.

As you can see in Figure 15.14, you initially have 741 terms with the terms **failure** and **drive**appearing in a little more than one-third of the documents. (Sometimes, if a term occurs too often, it might not be that usefull; you can make that term a stop word. This will not be considered to be the case here.) Continue with the following steps:

- 1. Right-click the term **failure**, and click **Show Text**. A new window appears, with all the text that contains **failure**, which appears to be only the term **failure**.
- 2. Right-click the term **failure**, but this time click **Containing Phrases**. This option selects small phrases that this phrase contains. All the phrases on the right side of Text Explorer output that contain the term **failure** (which you found to be just the term **failure**) will be highlighted.

Now examine the term **drive** as follows:

- 1. Click the **Text Explorer for Description** red triangle.
- 2. Click **Display Options** ► Show Stem Report. Added to the Text Explorer output are two lists—on the left, a list of stemmed terms and their terms used in the stem and, on the right, the list of terms and the term that they are stemmed with.
- 3. Scroll down the left, until you come to **drive** (stemmed terms are in alphabetic order). You can see the terms associated with the stemmed term **drive** are **drive** and **driving**.

(Note that if there is one or more stemmings that you do not like, it is probably best to exit the Text Explorer. Recode those undesired stemmings, and restart the Text Explorer.)

Add Frequent Phrases to List of Terms

Next add the most frequent phrases to your list of terms. Arbitrarily, you decide to include all the phrases occurring more than 500 times.

- 1. Click the first phrase under the list of phrases, which is **driver failure**.
- 2. Hold down the **Shift** key, scroll down, and click the phrase **influence of alcohol**. Notice that all the phrases above are now highlighted.
- 3. Right-click and click Add Phrase.

Several changes occur, and the Text Explorer output will look as shown in Figure 15.15.

Figure 15.15: Text Explorer Output Box

Text E	cplorer f	or Desc	ription					
	Number of Cases 16446	Total Tokens 155114		Number of Non- empty Cases 16446	Portion Non- empty per Case 1.0000			
Term a	nd Phras	e Lists						
Term			Count		Phrase		Count	N
proper- p devic- ins display- light- requir- drive- vel- vehicle drive- vel- suspende registratic property signal- highway lamp- person- d police off card upoi failure to red	icle d n plate- rive- motor icer n demand display- reg	control- hway vehicle	2334 2236 1803 1628 1628 1501 1376 1376 1376 1377 1370 1031 984 938 938 933 926 913 843 843 843 813 813 766 753		failure to traffic cor traffic cor control du failure to obey prop placed tra properly p traffic cor control di obey prop placed tra properly p device ins obey prop placed tra properly q driving m	n highway hicle hicle ure to obey obey ntrol ntrol device evice obey properly perly placed traffic other ontrol device instructions evice instructions evice instructions evice instructions evice instructions evice instructions evice instructions oblaced traffic tructions perly placed placed traffic tructions perly placed other ontrol oblaced other ontrol oblaced other ontrol oblaced other ontrol oblaced other o	2260 2154 1932 1578 1501 1487 1419 1419 1376 1376 1376 1376 1376 1376 1376 1376	2 3 2 2 4 3 2 3 2 4 4 4 4 4 3 3 3 3 2 2 2 2
Word	d Cloud			exceeding	maximum speed		Color	
person d	rive moto chang "fa polic Nghwaywithat r	susper nph in a p	e-traf	railure	stop-sign every worked red stop-line ay. signal. lane r signal lane r signal lane r ine to ober ine to ober ine to ober ine to ober vehicle on high	e police officer police officer conserved with on hwy 26 texts mph property 20 dense theory childle 27 the hyperty conserved with any childle of the hyperty childle of the hyperty childle of the hyperty childle of the hype	ed	
				i i i i i i i i i i i i i i i i i i i	, 9			

Parse the List of Terms

Lastly, parse your list of terms:

- 1. Right-click anywhere under the list of terms and, in the list of options, click **Alphabetic Order**.
- 2. Click the first term 1; hold down the Shift key, and scroll down to and click 99.
- 3. Right-click and click Add Stop Word. All these number terms are now deleted from your list of terms.
- 4. Again, right-click under the list of terms, and deselect **Alphabetic Order**. As before, the terms are now sorted by frequency.

Now delete all the terms that occur fewer than 100 times:

- 1. Scroll down the list of terms until you come to the term secur, which occurs 97 times.
- 2. Click secur:; hold down the Shift key, scroll down to the end, and click vehicle on highway without.
- 3. Right-click and click Add Stop Word.

The Text Explorer output will now look similar to Figure 15.16.

Figure 15.16: Text Explorer Output Box

Totat Explored for Description Mumber Number Total Total Tokers Per Case Portion Non-empty Case 183 1646 15511 9.4312 16354 0.9941 Total Case Total State Per Case Total Case Total Case Total Case Total Non-empty Per Case Total Case	Number of Number of Number of Nome empty Case empty pr Case Portion Nome empty asse empty pr Case 158 16446 155114 9.43172 16354 0.9944 Term and Phrase Lists Term of them and Phrase Lists Phrase Court N drive 2234	traffic-violations_dec2014 - Text	Explorer of Description - JM	P Pro	
of Jeans of Cases Tokens per Case 18 1646 155114 9.43172 16354 0.9944 Term and Phrase Lists Term Count drive 12254 0.9944 Term Count Count drive 12254 0.9944 Phrase Count N vehicle on highway vehicle on highway traffic control device 11578 2 traffic control device 11576 4 proper placed traffic control 1376 4 properly placed traffic 1376 3 traffic to display failure to display on driving motor vehicle 1383 4 traffic to tafp police officer demand by police free properly and demand excelling maximum speed driver vehicle yeb free figures of the properly and driver figures of t	of Cases Tokens per Case empty Cases empty Case 0.9944 Term and Phrase Lists Term and Phrase Lists Term and Phrase Lists Phrase Court N drive- drive- veh 1632 Court Phrase Court N drive- veh 1634 0.9944 Phrase Court N driver- vehicle 1634 0.9944 Phrase Court N driver- vehicle 1634 0.994 Phrase Court 1765 176 <td< th=""><th>Text Explorer for Descri</th><th>ption</th><th></th><th></th></td<>	Text Explorer for Descri	ption		
Term Count Phrase Count N failure 2334	Term Count Phrase Count N failure 2334 100 200 2	of Terms of Cases Tokens p	per Case empty Cases	empty per Case	
failure drive. 2276 drive. 2276 drive. 2276 drive. 2276 drive. 2276 drive. 2276 drive. 2276 drive. 2276 drive. 2276 drive. 2276 drive. 2276 drive. vehicle stop. 1628 drive. stop. 1628 drive. stop. 1628 drive	failure drive. 2260 2 veh drive. 2260 2 2260 2 vehicle on highway driver failure to abey proper placed traffic control device instructions display. light. requir. 1070 drive vehicle 1037 drive vehicle 1038 drive vehicle	Term and Phrase Lists			
drive diteres 1803 veh 1628 ticense 1803 driver 1624 driver 1624 driver 1624 traffic control device 1419 proper placed traffic control 1487 traffic control device 1419 proper placed traffic control 1376 diplay. 1323 drive vehicle 1037 drive vehicle 1038 drive vehicle 1037 drive vehicle 1038 drive vehicle 1038 driving motor vehicle 1088 signal- 926 driver failure to display driving motor vehicle 133 4 driving motor vehicle 133 4 driving motor vehicle 133 4 driver failure to display registration 136 eres driver failure to stop 955 greson driving motor 1855 demand by police officer 815 demand by police officer 816	drive dice matrix display. registration plate proper- placed traffic control device traffic control device traffic control device control device traffic control traffic control t	Term	Count	Phrase	
driver failure to stop card upon demand exceeding maximum speed proper placed traffic control. mph zone person. drive motor vehicle charg speed in highway require failure suspended driver registration plate highway or public uses ended driver failure to obey <u>devic instructions</u> suspended registration driver vehicle on highway	driver failure to stop card upon demand exceeding maximum speed person- driver motor vehicle chang speed highway or vehicle chang speed in highway require failure with display. In the speed highway or public with a posted speed driver failure to obey suppended registration drive- vehicle on highway	drive- license veh driver- stop- driver-failure to obey proper-placed traffic control- devic- instructions display- light- requir- drive-vehicle drive-vehicle drive-vehicle drive-vehicle drive-vehicle on highway suspended registration plate- property signal- highway lamp- person-drive-motor vehicle police officer card upon demand failure to display- registration red	2276 1803 1628 1624 1518 1501 1376 1376 1376 1323 1237 1170 1067 1031 984 978 938 933 926 933 933 926 843 816 813 813 813 813 813 813 813 813 813 813	vehicle on highway driving vehicle traffic control device control device failure to obey properly properly placed traffic control properly placed traffic driving motor driving motor vehicle motor vehicle on highway failure to display driving vehicle on highway failure to stop person driving motor person driving motor person driving motor person driving motor vehicle police officer demand by police officer demand by police officer demand by police officer demand by police officer driving the to stop vehicle on highway without	2154 3 [1932 2 1578 2 1487 2 1419 3 1419 2 1376 4 1376 4 1376 3 1216 2 1088 3 1064 4 1061 3 983 4 956 3 843 4 956 3 843 4 813 3 843 4 816 2 813 4 813 3 744 2 708 2 692 4 691 4
card upon demand exceeding maximum speed proper- placed traffic control. mph zone person- drive- motor vehicle chang speed in highway require- failure with a posted driver- registration plate drive- vehicle vehicle vehicle on bigh mph vehicle control. mph zone suspended registration drive- vehicle on highway	card upon demand exceeding maximum speed proper- placed traffic control. mph zone person- drive- motor vehicle chang speed highway require failure suspended driver registration plate highway or public with a mph zone mph zone jan int lamp- display. progese who may properly with a progese vehicle with a mph zone jan mph vehicle vehicle vehicle on bighway intervent driver failure to obey suspended registration drive- vehicle on highway	■ ■ Word Cloud	Maghawa		
		person drive motor vehicle , highway highway or public us use drive vi mph in a pos	card upon demand ex d traffic contro dang speed requir. failure requir. failure ded driver. registra ehicle veh sted driver. failu devic. instruct uspended registration drive.	imphizone stop-license stop-license minutesered speed stop-license minutesered speed minutesered speed	

Using Multivariate Techniques

After you have curated the DTM to your satisfaction, you are ready to apply multivariate techniques to understand the underlying structure of the DTM.

These techniques are similar to principal components, factor analysis, and clustering techniques that are applied to continuous data.

Perform Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a family of mathematical and statistical techniques for extracting and representing the terms and phrases from a corpus. The DTM is reduced dimensionally to a manageable size, which makes the analyses go much faster. And the DTM is amenable to using other multivariate techniques, by applying singular value decomposition (SVD).

Understanding SVD Matrices

SVD produces a set of orthogonal columns that are linear combinations of the rows and explains as much of the variation of the data as possible. SVD is an efficient approach to use with large, very sparse matrices, which the DTM typically tends to have. SVD decomposes the DTM into three other matrices:

$$DTM = \mathbf{D} * \mathbf{S} * \mathbf{T}$$

These matrices are defined as follows:

- **D** is an orthogonal document-document matrix of eigenvectors.
- T is an orthogonal term-term matrix of eigenvectors.
- S is a diagonal matrix of singular values.

The singular vectors in \mathbf{D} and \mathbf{T} reveal document-document, document-term, and term-term similarities and other semantic relationships, which otherwise might be hidden.

Many of the singular values in the S matrix are "too small" and can be ignored. So they are assigned values of 0, leaving k nonzero singular values. The representation of the conceptual space of any large corpus requires more than a handful of underlying independent concepts. As a result, the number of orthogonal vectors that is needed is likely to be fairly large. So, k is often several hundred.

Similarities and relationships are now approximated by this reduced model. This process is analogous to using principal components in multivariate analysis. While principal components provide components for the columns, SVD simultaneously provides principal components for both the columns and rows (that is, for the documents and terms).

Plot the Documents or Terms

A common practice is to plot the documents or terms, these singular vectors, and especially the first two vectors, that result from the SVD. Similar documents or terms tend to be plotted closely

together, and a rough interpretation can be assigned to the dimensions that appear in the plot. Complete the following steps:

- 1. Click the **Text Explorer for Description** red triangle, and click **Latent Semantic Analysis, SVD**. The Latent Semantic Analysis Specifications dialog box will appear as shown in Figure 15.17.
- 2. Change the Minimum Term Frequency to 100.
- 3. Click the drop-down arrow for Weighting.
- 4. The **TF_IDF** weighting results are usually more interpretable than the Binary, so click that option.

Regarding weighting options, various methods of the term-frequency counts have been found to be useful, with the **Binary** and **TF_IDF** being the most popular:

- The **Binary** weighting option is the easiest to understand in that it assigns a zero or a 1 to indicate whether the term exists in the document. A disadvantage of the **Binary** option is that it does not consider how often the term occurs in the document.
- The **TF_IDF** weighting option, which is short for *term frequency-inverse document frequency*, does consider the tradeoff between the frequency of the term throughout the corpus and the frequency of the term in the document.

Next you will select one of three SVD approaches:

- Uncentered
- Centered
- Centered and Scaled

The benefits and drawbacks of each are as follows:

- Traditional latent semantic analysis uses an **Uncentered** approach. This approach can be problematic because frequent terms that do not contribute much meaning tend to score high in the singular vectors.
- The **Centered** approach reduces the impact of these frequent terms and reduces the need to use many stop words.
- The **Centered and Scaling** approach is essentially equivalent to doing principal components on the correlation matrix of the DTM. This option explains the variation in the data (not just the variation in the mean). Therefore, it tends to produce more useful results, whereas using just the **Centered** approach is comparable to doing principal components on the covariance matrix of the DTM.

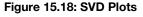
Continue with the example as follows:

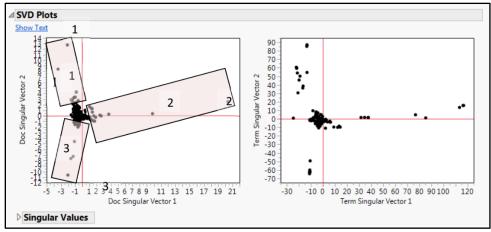
- 1. Click the drop-down arrow for **Centering and Scaling**. Select the **Centered and Scaling** approach.
- 2. Click OK.

Figure	15.17:	Latent	Semantic	Analysis	Specifications	Dialog	Box

Ę	Specifications	×								
	Specifications for Terms and Weights									
	Maximum Number of Terms	168								
	Minimum Term Frequency	100								
	Weighting	TF IDF								
	Number of Singular Vectors	100								
	Centering and Scaling Centered and Scaled									
	ОК	Cancel Help								

The Latent Semantic Analysis output should look similar to Figure 15.18 (except for the highlighted regions).





The plot on the left displays the first two document singular vectors in matrix D. The plot on the right displays the first term singular vectors in T as shown in Figure 15.18. In each graph, the singular vectors have three branches or tendrils. In the document singular vector graph, the three tendrils are highlighted and labeled. To examine the text, complete the following steps:

- 1. Left-click the document singular vector graph, hold down the mouse button, and move to highlight the area labelled 1 in Figure 15.18.
- 2. Just below SVD plots, click Show Text. A window appears with a list of documents.

Examining the document list, you see that major themes are the **drive use handheld** and **failure** of vehicle on highway to display lighted lamps.

Similarly, highlight, the documents in the tendril labeled 2 and click **Show Text**. These documents appear to have several themes of **driving vehicle**, **person drive motor vehicle**, and **negligent driving**. Lastly, highlight the documents in Tendril 3, and you see the terms **driving vehicle on highway without current registration** and **failure of licensee to notify**. (Keep these documents in the third tendril highlighted.)

As you did with the document singular vectors, you can explore the term three tendrils in the term singular vector plot. In general, the document singular vector plot provides more insight than the term singular vector plot.

To examine more SVD plots, complete the following steps:

- 1. Click the **Text Explorer for Description** red triangle.
- 2. Click SVD Scatterplot Matrix.
- 3. Enter 10 for the Number of singular vectors to plot.
- 4. Click OK.

Added to the Text Explorer output are scatterplots of the first 10 singular vectors for the documents and terms. Figure 15.19 shows the top part of this scatterplot.

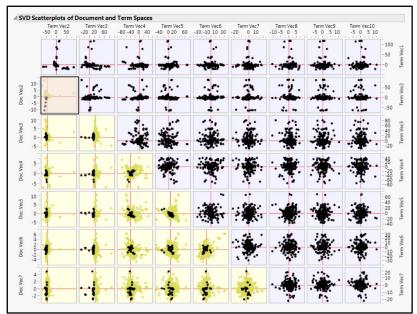


Figure 15.19: Top Portion of SVD Scatterplots of SVD Plots of 10 Singular Vectors

The bottom diagonal plots are graphs of the document singular vectors. The upper diagonal plots are graphs of the term singular vectors. Highlighted in Figure 15.19 is the SVD plot of the first two document singular vectors, which is similar to the plot on the left in Figure 15.18. The highlighted documents in the third tendril are highlighted. And these documents are highlighted in the other document singular vector plots. The term singular vector plot directly above the highlighted document singular vector plot in Figure 15.19 is the plot of the first two-term singular vectors, which is similar to the plot on the left above the highlighted document singular vector plot in Figure 15.19 is the plot of the first two-term singular vectors, which is similar to the plot in Figure 15.18 but rotated 270°.

Perform Topic Analysis

Another way of observing these term themes in the documents is to perform topic analysis. Topic analysis performs a VARIMAX rotation, essentially a factor analysis, on the SVD of the DTM. Complete the following steps:

- 1. Click the red triangle next to Text Explorer for Description.
- 2. From the list options click **Topic Analysis**, **Rotated SVD**.
- 3. Click OK.

The Topic Analysis results are added to the Text Explorer output as shown Figure 15.20.

Topic1 Topic2			Topic3			Topic4			Topic5		
Term	Score	Term	Score	Term	Score	Term		Score	Term		Scor
life imprudent endangering manner negligent person- drive- vehicle	0.34064 0.33945 0.31183	whilemotor motion telephone handheld use hand. address days notify vehicle within administration chang-	0.2965 0.2957 0.2957 0.2957 0.2941 0.2925 -0.1990 -0.1959 -0.1897 0.1879 -0.1837 -0.1864 -0.1652	knowingly devic- cond unfavorable illumin- intersect- red display- line stop-	0.2882 -0.2500 -0.2403 0.2327 0.2135 0.2135 0.1821 0.1740 0.1702 0.1657 0.1353 0.1323 0.1287	person- drive- highway or pu privilege suspended yield property police uniformed turn- highway tab- approach- right		e -0.2920 -0.2867 -0.2408 -0.2354 0.2296 -0.2067 0.1838 0.1811 0.1792 0.1779 -0.1561 0.1520	red signal- speed knowingly uninsured mph flashing steady traffic highway exceeding the posted speed limit		0.3268 0.2245 0.2237 0.2190 0.1911 -0.1852 0.1788 0.1779 0.1711 0.1629 -0.1628 -0.1628 -0.1569
	_	licensee	-0.1617	-	_	way		0.1487	stop		0.139
Topic		Topic		Topic8		Topic9		Topic10 Term Score		6	
Term failure to stop∙	-0.2232	Term drive	Score 0.3205	Term failure	Score 0.2706	Term current		lerm author		Score 0.3214	
hung to stop motor direct- uninsured accident drive- motor turn- emerg- possess- designated half opper- knowingly	0.2119 0.2114 0.2056 0.1949 0.1864 0.1855 -0.1803 -0.1758 -0.1750 0.1677 0.1643 -0.1619 0.1538 0.1434	lic within motor right privilege possess- equip. obstruct- hwy unauthorized reg	-0.2518 0.2417 -0.2373 0.2327 -0.2196 0.2108 0.1936 -0.1933 0.1881 0.1708 -0.1588	fail. speed signal. light. traffic demand visibl. reason- circular motor vehicle safety	0.2652 -0.2402 0.2370 0.2355 0.2254 0.2051 -0.1977 0.1779 0.1650	display- valid- speed stop- sign issued maintain stop- highway suspended permit- obstruct- view reason-	0.2407 -0.2297 0.2021 0.1874 -0.1820 -0.1805 0.1803 -0.1596 0.1403 0.1396 -0.1387 0.1381 -0.1371	drive- motor v vehicle on hig visibl- safety impaired notify alcohol- registration p stop- within child person- licensee adm	ghway		

Figure 15.20: Topic Analysis Output

The scores represent the loading of the term in that topic. The larger a term's score, the stronger its contribution to the topic. Each topic can be examined for major themes or terms that contribute to that topic. In particular, when you examine the first topic in Figure 15.20, it appears that this topic deals with careless and negligent, somewhat like Tendril 2, which you identified in the SVD plots. And topic 2 seems to have similar terms as Tendril 1. If you choose to consider too few topics, then there can be significant overlap between competing information. But if you choose too many topics, then you will have some topics covering the same topics. Use trial and error to decide how many to consider.

Perform Cluster Analysis

Cluster analysis is used to further understand the singular vectors. The Text Explorer platform in JMP uses a customized latent class analysis methodology for cluster analysis. This is built just for text. This latent class analysis procedure is good for finding interesting patterns in the text, and because it is sparse matrix based, it is very fast.

Begin the Analysis

To begin, complete the following steps:

- 1. Click the red triangle next to **Text Explorer for Description** and click **Latent Class Analysis**.
- 2. In the Specifications dialog box, change the **Minimum Term Frequency** to **4** and the **Number of Clusters** to **10**.
- 3. Click OK.
- 4. Click the red triangle labelled Latent Class Analysis for 10 Clusters and click Color by Cluster.

The cluster analysis results are shown in Figures 15.21 and 15.22. Because the latent class analysis procedure uses a random seed, your results will be slightly different.

Examine the Results

Scroll back up to the SVD scatterplots. The document SVD plots are now colored by cluster. Documents that tend to cluster together will appear near each other in their SVD plots. (You can increase the size of a plot by grabbing the border with the mouse while holding down the mouse key and dragging out wider.) For example, in Figure 15.23 the upper left portion of the SVD scatterplot is shown.

Look at the group of documents highlighted in **Doc Vec3**. All these documents are in the same cluster and in other plots. You can see that they are close to each other. Match up the color of these documents with cluster colors as shown in Figure 15.21. Click that cluster number; in our case it was Cluster 6. Now all the documents in Cluster 6 are highlighted in the SVD scatterplots.

Scroll further up to **SVD Plots.** In the left document singular vector plot, the Cluster 6 documents are highlighted, which are mostly those documents in Tendril 3 (as shown in Figure 15.18). Click **Show Text**. You see phrases or terms similar to those that you found in Tendril 3. Similarly, you can highlight other clusters of the same colored documents and view the text.

C 424765.56 Show Text																			
Cluster Mixture Prob	abilit	ior																	
	Japint	les																	
Mixture																			
Cluster Probability																			
Cluster1 0.19168																			
Cluster2 0.14078																			
Cluster3 0.13054																			
Cluster4 0.12222																			
Cluster5 0.10395																			
Cluster6 0.07039																			
Cluster7 0.06849																			
Cluster8 0.06253																			
Cluster9 0.05518																			
Cluster10 0.05423																			
Term Probabilities b	v Chur	tor																	
Term Probabilities b		luster Most	Churton																
Term		haracteristic		robable	Cluster1		Cluster2	c	luster3		Cluster4	Chr	ster5	Clus	erfi	Cluster7	0	lust	
failure		luster6	Cluster4		0.0001		0.0001		0.0345		0.4895		0001		021	0.0029	TTT .		
drive-		lusterő	Cluster4 Cluster2		0.0001		0.4146		0.0001		0.2852		0001		492	0.00029			
license		luster0 luster7	Cluster2		0.0000		0.0000		0.0001		0.2832		0001		777	0.7896			
veh		luster/	Cluster4		0.0088		0.0421		0.0330		0.5351		0001		001	0.0506	111	Ξ.	
			Cluster				0.0421		0.2556		0.5351		0001		001	0.0036			
driver		luster8			0.0000														
stop-		luster8	Cluster8		0.1408	E	0.0001		0.0342		0.1241		0000		001	0.0001			
driver-failure to obey		luster3	Cluster3		0.0000		0.0000		0.6979		0.0000		0001		001	0.0001			
proper-placed traffic contr		luster3	Cluster3		0.0000		0.0000		0.6406		0.0000		0000		001	0.0001			
devic- instructions		luster3	Cluster3		0.0000		0.0000		0.6406		0.0000		0000		001	0.0001			
display-		luster10	Cluster		0.0000		0.0000		0.0000		0.0088		0000		876	0.0408			
light-		luster9	Cluster1		0.2432		0.0000		0.0000		0.0223		0000		001	0.0001			
require		luster6	Cluster		0.0523		0.0000		0.0000		0.2182		0000		928	0.0001			
drive-vehicle		luster10	Cluster1		0.0000		0.1377 📃		0.0000		0.0603		1468		224	0.0001			
vehicle		luster4	Cluster4		0.0020		0.0724		0.1427		0.2053		0000		001	0.0001			
drive- vehicle on highway		luster9	Cluster1		0.2141		0.0000		0.0000		0.0000		0000		018	0.0382			
suspended		luster7	Cluster7		0.0000		0.0000		0.0000		0.0071		0000		001	0.8550			
registration plate-		luster10	Cluster1		0.0000		0.0000		0.0000		0.0689		0000		000	0.0001			
property	C	luster7	Cluster7		0.0000		0.0000		0.0000		0.0280		0000		000	0.4658			
signal	C	luster8	Cluster8		0.0845		0.0000		0.0000		0.0377		0000		000	0.0000			
highway	C	luster6	Cluster	i	0.0000		0.0127		0.0000		0.0921	0.	1471	0.3	841	0.0000			
lamp-	c	luster9	Cluster9		0.0740		0.0000		0.0000		0.1176	0.	0000	0.0	000	0.0000			
person- drive- motor vehicle	e C	luster7	Cluster7		0.0000		0.0000		0.0000		0.0000	1 0.	0000	0.0	000	0.7478			
police officer		luster1	Cluster]		0.2578		0.0000		0.0000		0.0015		0000		000	0.0000	ITTI		
card upon demand		luster1	Cluster]		0.2578		0.0000		0.0000		0.0000		0000		000	0.0000	1111		
failure to display- registration		luster1	Cluster1		0.2578		0.0000		0.0000		0.0000		0000		000	0.0000		*	
+																	Þ		
Top Terms per Clust	er		-																
Cluster1			ster2			luster3		Cluste		-	Cluster5		Cluste		-	Cluster7		Clust	
Term	Score			Score			Score		Score				Term	Score			Score		Score 1
	3.7776				driver-failure					mph zone			license		license		11.024		9.3514
drive-vehicle on highway							6.4097				maximum sp		failure		suspended		8.592		6.9448 1
police officer		drive-vehicle			proper-place	o traffic c				mph in a p	osted	4.3513				e-motor ve			6.0867 f
	2.5802			1.6644			3.9843		3.485				display		property		6.3919		5.5054 r
		alcohol		1.4989							the posted s		requir	5.4715	highway or	public use-	4.6583		5.2647
		impaired		1.4817			2.2761		2.3275				highway	4.9751	privilege			traffic	4.0582
		stop- sign			vehicle			equip- 1					police	4.2202			2.194		3.8847
		failure to stop		1.104						drive-vehi	cle		uniformed				1.0249		3.4923
		headlight		0.9937				lamp- 1					demand		revoked			flashing	
stop- sign	1.7605	per		0.9028	telephone		1.4264	reg 1	L.4531	reason-		1.7474	individu	3.8707	drive motor	r	0.6305	intersect-	2.5552 t

Figure 15.21: Top Portion of Latent Class Analysis for 10 Clusters Output Box

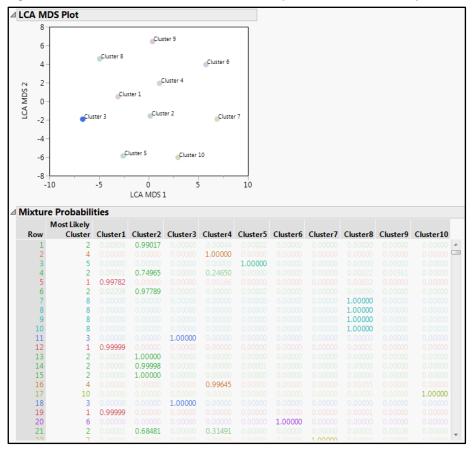


Figure 15.22: Lower Portion of Latent Class Analysis for 10 Clusters Output Box

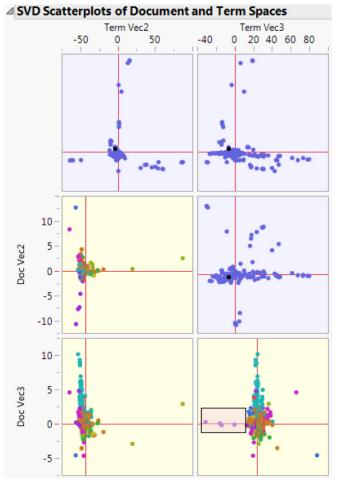


Figure 15.23: Upper Leftmost Portion of SVD Scatterplots

Identify Dominant Terms

The **Term Probabilities by Cluster** report in Figure 15.21 has the term frequency in each cluster. You can examine the terms horizontally or the clusters vertically, focusing on the large frequencies. For example, look at the term **license**, which occurs most often in Clusters 6 and 7. Since you previously have looked at Cluster 6, look at Cluster 7:

- 1. Right-click anywhere inside the report.
- 2. Select Sort Column.
- 3. In the new window, select Cluster 7.
- 4. Click OK.

The most frequent terms in Cluster 7 are toward the top of the report. It seems that the Cluster 7 theme addresses drivers who were driving with suspended licenses.

A scoring algorithm for identifying which terms are most dominant within a cluster is shown in the **Top Terms per Cluster report** (see Figure 15.21). The score is based on the term cluster

frequency relative to its corpus frequency. Larger scores tend to occur more often in the cluster. Look at Cluster 6 and 7. You see the most frequent terms are the terms you have identified earlier.

Occasionally terms score negative numbers, which implies that those terms are less frequent (you do not have any negative scores in our example). Many times when several terms have negative numbers, they occur in the same cluster. This is called a *junk cluster*. Most likely in this junk cluster are blank documents or simply documents that do not fit nicely in other clusters and just are adding noise to the analysis. If a junk cluster occurs, then it may be useful to identify those documents in this cluster and rerun latent class analysis to exclude these junk documents.

The multiple dimensional scaling (MDS) plot of the clusters in Figure 15.22 is produced by calculating the Kullback-Leibler distance between clusters. In natural language processing (NLP), a document is viewed as a probability distribution of terms. The Kullback-Leibler distance is a widely used measure for calculating the distance between two documents or probability distributions (Bigi). An MDS is applied to these Kullback-Leibler distances to create coordinates for the clustering in two dimensions. You can explore the MDS plot to examine clusters that are near one another, as well as clusters that are far away from one another.

The clusters in your MDS plot are fairly well dispersed. Nonetheless, complete the following steps:

- 1. Click Cluster 3. You can see in the **Top Terms per Cluster** report that the main terms are **driver** failure to obey, devic instructions, and proper placed traffic control.
- 2. Scroll back up to the output. You can see where the Cluster 3 documents occur in the SVD plots.
- 3. Click Show text.

Most of the documents appear to have those terms that you detected.

Using Predictive Techniques

If the data set has a dependent variable and you want to do some predictive modeling instead of using the large DTM matrix, you can use the document singular vectors. To determine how many singular vectors to include, complete the following steps:

- 1. Scroll back up the output to find **Singular Values**. (It is just below the **SVD plots** and before the **SVD Scatterplots**.)
- 2. Click the down-arrow next to Singular Values.

In general, as in principal components and factor analysis, you are looking for the elbow in the data. Or another guideline is to include singular values until you reach a cumulative percentage of 80%. Many times with text data, the curve or percentages will only gradually decrease, so to reach 80% you may have to reach several hundred. As you will see, the default is 100. How many to include is a judgment call.

Figure 15.24 shows the first 25 singular values. In this case, the percentages quickly exceed 80%, and there is a sharp decrease. Six singular document vectors are chosen:

- 1. Click the **Text Explorer for Description** red triangle.
- 2. Click Save Document Singular Vectors.

348 Fundamentals of Predictive Analytics with JMP, Second Edition

- 3. In the input box in the new dialog box next to Number of Singular Vectors to Save, enter 6.
- 4. Click **OK**.

Check the data table, and you can see six new singular vector variables.

		<u> </u>	
Figure	15.24:	Singular	Values

⊿ Singula	r Values		
	Singular		
Number	Value	Percent	Cum Perc
1	351.57	25.4519	25.4
2	322.75	21.4496	46.9
3	294.95	17.9134	64.8
4	273.83	15.4395	80.2
5	224.22	10.3518	90.6
6	134.54	3.7275	94.3
7	75.21	1.1649	95.4
8	34.24	0.2414	95.7
9	33.08	0.2253	95.9
10	32.03	0.2113	96.1
11	30.96	0.1974	96.3
12	30.64	0.1933	96.5
13	29.53	0.1796	96.7
14	29.15	0.1750	96.9
15	28.78	0.1705	97.0
16	27.57	0.1565	97.2
17	27.31	0.1535	97.4
18	26.21	0.1415	97.5
19	25.58	0.1347	97.6
20	25.44	0.1333	97.8
21	24.83	0.1269	97.9
22	23.87	0.1173	98.0
23	23.36	0.1124	98.1
24	23.14	0.1102	98.2
25	22.66	0.1057	98.3

Perform Primary Analysis

Before you perform any predictive modeling, do some primary analysis:

- 1. Click **Analyze** ► **Distribution**. The distribution dialog box will appear.
- 2. Click Violation Type; click Y, Columns.
- 3. Click OK.

As shown in the distribution output in Figure 15.25, you can see that most of the violations are equally distributed between **Warning** and **Citation**.

traffic-violatio	ns 🕒		x
⊿ ▼ Distribut ⊿ ▼ Violati		e	
Warning			
ESERO			
Citation			
⊿ Freque	ncies		
Level	Count		
Citation		0.48157	
ESERO Warning		0.45264	
Total		1.00000	
N Missing 3 Le			
		û 🖾 🔲	▼:

Figure 15.25: Distribution Output for Violation Type

Perform Logistic Regressions

Now run a logistic regression, predicting **Violation Type** and using **Property Damage**, **Race**, and **Belts** as independent variables. The results are shown in Figures 15.26 and 15.27. You can see that the three variables are significant in predicting **Violation Type**. However, as the Confusion matrix in Figure 15.27 shows, the misclassification is rather high at 47.6%.

Figure	15.26:	Top	Portion	of	Initial	Logistic	Regression	Output

Nomina	l Logist	ic Fit	for Vi	olat	tion Ty	ре		
⊿ Effect Su	-				-			
Source		LogW	orth					PValu
Property I		_	5.637					0.0000
Race	Jamage		7.208	:		:		0.0000
Belts		4	1.838 💻					0.0000
Remove	Add Edit	F	DR					
Converged in (Gradient, 8	iterat	ions					
Iteration	s							
⊿ Whole N	lodel Te	st						
Model	-LogLike	lihood	ł	DF	ChiSqu	are	Prob>ChiS	9
Difference	-	44.04		14	688.0	895	<.0001	*
Full		16.88						
Reduced	143	60.928	5					
RSquare (U)			0.02	40				
AICc			2806					
BIC			2818					
Observation	-	wgts)	161	41				
Fit Detai	_						1	
△ Lack Of I			1.11		CHIC			
Source Lack Of Fit	28	-Log	Likeliho 19.5		ChiSqua 39.10			
Saturated	28 42				Prob>C			
Fitted	42		14016.8		0.0			
⊿ Paramet	er Estim	ates						
Term			Estim	ate	Std Err	or (ChiSquare	Prob>ChiSq
Intercept		1	1.40528	364	0.11102		160.21	<.0001*
Belts[No]			-0.2153	893	0.0506	95	18.05	<.0001*
Property Da			-1.2401		0.0913		184.36	<.0001*
Race[ASIAN	-				0.07422		25.21	<.0001*
Race[BLACk		(0.08013		0.05440		2.17	0.1408
Race[HISPA					0.05716		71.06	<.0001*
Race[NATIV		AN]			0.22602		0.41	0.5219
Race[OTHE Intercept	Ŋ				0.07652 0.52913		7.34 29.15	0.0067* <.0001*
							0.56	0.4523
Belts[No]				3/5	0,11960	45		
Belts[No] Property Da	mage[No]				0.11969			
Property Da			0.7796	021	0.11969 0.50792 0.14961	54	2.36 0.01	0.1248
	ון ד		0.7796	021 041	0.50792	54 36	2.36	0.1248
Property Da Race[ASIAN	1) (]	(0.7796 -0.0178 0.01292	021 041 935	0.50792 0.14961	54 36 79	2.36 0.01	0.1248 0.9053
Property Da Race[ASIAN Race[BLACk	1] [] [] [] []	AN]	0.7796 -0.0178 0.01292 0.47869 -0.1677	021 041 935 221 705	0.50792 0.14961 0.1189 0.12136 0.50952	54 36 79 24 09	2.36 0.01 0.01 15.56 0.11	0.1248 0.9053 0.9135 <.0001* 0.7420
Property Da Race[ASIAN Race[BLACH Race[HISPA Race[NATIV Race[OTHE	I] (] NIC] E AMERIC R]	AN]	0.7796 -0.0178 0.01292 0.47869 -0.1677 -0.2368	021 041 935 221 705 692	0.50792 0.14961 0.1189 0.12136 0.50952 0.16568	54 36 79 24 09	2.36 0.01 0.01 15.56	0.1248 0.9053 0.9135 <.0001*
Property Da Race[ASIAN Race[BLACk Race[HISPA Race[NATIV Race[OTHEI For log odds	I] NIC] E AMERIC R] of Citatio	AN] n/War	0.7796 -0.0178 0.01292 0.47869 -0.1677 -0.2368 ming, ES	021 041 935 221 705 692	0.50792 0.14961 0.1189 0.12136 0.50952 0.16568	54 36 79 24 09	2.36 0.01 0.01 15.56 0.11	0.1248 0.9053 0.9135 <.0001* 0.7420
Property Da Race[ASIAN Race[BLACK Race[HISPA Race[NATIV Race[OTHEI For log odds	I] NIC] E AMERIC R] s of Citatio ance of I	AN] n/War Estim	0.7796 -0.0178 0.01292 0.47869 -0.1677 -0.2368 ming, ES	021 041 935 221 705 692 SERO	0.50792 0.14961 0.1189 0.12136 0.50952 0.16568	54 36 79 24 09	2.36 0.01 0.01 15.56 0.11	0.1248 0.9053 0.9135 <.0001* 0.7420
Property Da Race[ASIAN Race[BLACk Race[HISPA Race[NATIV Race[OTHEI For log odds	I] NIC] E AMERIC R] s of Citatio ance of I	AN] n/War Estim	0.7796 -0.0178 0.01292 0.47869 -0.1677 -0.2368 ming, ES	021 041 935 221 705 692 SERO	0.50792 0.14961 0.1189 0.12136 0.50952 0.16568 //Warning	54 36 79 24 09	2.36 0.01 0.01 15.56 0.11	0.1248 0.9053 0.9135 <.0001* 0.7420
Property Da Race[ASIAN Race[BLACK Race[HISPA Race[NATIV Race[OTHEI For log odds	I] G NIC] E AMERIC R] of Citatio ance of I celihooc	AN] n/War Estim	0.7796 -0.0178 0.01292 0.47869 -0.1677 -0.2368 ming, ES	021 041 935 221 705 692 SERO	0.50792 0.14961 0.1189 0.12136 0.50952 0.16568 //Warning	54 36 79 24 09 15	2.36 0.01 0.01 15.56 0.11	0.1248 0.9053 0.9135 <.0001* 0.7420
Property Da Race[ASIAN Race[BLACH Race[HISPA Race[NATIV Race[OTHEI For log odds Covaria	I] G NIC] E AMERIC R] of Citatio ance of I celihooc	AN] n/War Estim I Rati	0.7796 -0.0178 0.01292 0.47869 -0.1677 -0.2368 ming, ES nates io Test	021 041 935 221 705 692 SERO	0.50792 0.14961 0.1189 0.12136 0.50952 0.16568 //Warning	54 36 79 24 09 15	2.36 0.01 15.56 0.11 2.04	0.1248 0.9053 0.9135 <.0001* 0.7420
Property Da Race[ASIAN Race[BLACH Race[HISPA Race[NATIV Race[OTHE] For log odds Covaria Effect Lil Source	I] NIC] E AMERIC R] of Citatio ance of I celihooc	AN] n/War Estim I Rati parm	0.7796 -0.0178 0.01292 0.47869 -0.1677 -0.2368 ming, ES nates io Test	021 041 935 221 705 692 SERO	0.50792 0.14961 0.1189 0.12136 0.50952 0.16568 //Warning	54 36 79 24 09 15	2.36 0.01 15.56 0.11 2.04	0.1248 0.9053 0.9135 <.0001* 0.7420

Confusion Matrix									
Training									
Actual									
Violation	Predicted Count								
Туре	Citation	ESERO	Warning						
Citation	4812	0	2961						
ESERO	596	0	466						
Warning	3657	0	3649						

Figure 15.27: Lower Portion of Initial Logistic Regression Output

Rerun the logistic regression, now also including the six singular vectors. Figures 15.28 and 15.29 display the output. The model significantly improved and now has a misclassification rate of 40.7% (not great, but almost 7% better).

Figure 15.28: Top Portion with Singular Values Logistic Regression Output

Nomina	-		t for V	ola	tion Ty	ре		
Whole N								
Observation	ns (or Su	ım Wgts	;) 161	.41				
Fit Detai	ils							
Lack Of	Fit							
Source	D	E alor	al ikaliho	bod	ChiSqua	ra		
Lack Of Fit			12512/		25024			
Saturated	2742		933.		Prob>C			
Fitted	2/42	-	13445		1.00			
		-			2.01			
Paramet	erEsti	mates						
Term			Estim					Prob>ChiSq
Intercept			1.4586		0.1151		160.40	<.0001*
Belts[No]		1-1	-0.2198		0.0524		17.55	<.0001*
Property Da		10]			0.09442		177.01	<.0001*
Race[ASIAN			-0.3301				18.51 1.02	<.0001* 0.3128
Race[BLAC Race[HISPA					0.05625		69.92	<.0001*
Race[NATI\			0.17876		0.2335		0.59	0.4441
Race[OTHE		acorting			0.07932		9.92	0.0016*
Singular Ve	-		0.14541				183.73	<.0001*
Singular Ve			-0.0201	244	0.01092		3.39	0.0655
Singular Ve			-0.0540)743	0.00838	93	41.55	<.0001*
Singular Ve	ctor 4		-0.0439	174	0.00921	66	22.71	<.0001*
Singular Ve			0.24647	847	0.01136	55	470.31	<.0001*
Singular Ve	ctor 6				0.01785		13.84	0.0002*
Intercept					0.52986		26.74	<.0001*
Belts[No]					0.11999		0.44	0.5063
Property Da		10]	0.70561		0.5083		1.93	0.1651
Race[ASIAN Race[BLAC	-		0.00355		0.15007 0.1192		0.00	0.9812 0.9528
Race[HISPA	-				0.12166		15.98	<.0001*
Race[NATI\	-	ICANI			0.51044		0.14	0.7127
Race[OTHE					0.16607		1.88	0.1702
Singular Ve	-				0.02314		1.37	0.2425
Singular Ve			0.08472	2824	0.02508	63	11.41	0.0007*
Singular Ve			-0.13	405	0.01988	55	45.44	<.0001*
Singular Ve			-0.0131				0.56	0.4539
Singular Ve			0.07425		0.021		11.45	0.0007*
Singular Ve					0.03386		3.04	0.0812
For log odd			-	SERC)/Warning	9		
Covaria	ance o	f Estir	nates					
Effect Li	keliho	od Rat	tio Tes	ts				
					L-R			
Source		Nparm	DF	Ch	iSquare	Pro	b>ChiSq	
Belts		2			3456469		<.0001*	
Property Da	amage	2			2.080134		<.0001*	
Race		10			3.627373		<.0001*	
Singular Ve		2			3.348504		<.0001*	
Singular Ve		2			7949328		<.0001*	
Singular Ve		2			2697129		<.0001*	
Singular Ve Singular Ve		2			1297768		<.0001* <.0001*	
Singular Ve		2			6591688		<.0001*	
Singular ve	0.010	2	2	210	0001000		10001	

Confusion Matrix									
Training									
Actual									
Violation	Predicted Count								
Туре	Citation	ESERO	Warning						
Citation	4704	0	3069						
ESERO	380	0	682						
Warning	2446	0	4860						

Figure 15.29: Lower Portion with Singular Values Logistic Regression Output

Exercises

- 1. In the aircraft_incidents.jmp file is data for airline incidents that were retrieved on November 20th, 2015 from http://www.ntsb.gov/layouts/ntsb.aviation/Index.aspx. For the Final Narrative variable, use the Text Explorer to produce a DTM by phrasing and terming.
- 2. Create a Word Cloud. As in problem 1, similarly produce a DTM by phrasing and terming, and create a Word Cloud except for the variable Narrative Cause. In the file Nicardipine.jmp is data from adverse events from this drug. For the Reported Term for the Adverse Event variable, use the Text Explorer to produce a DTM by phrasing and terming.
- 3. Create a Word Cloud. In the Airplane_Crash_Reports.jmp file is one variable, NTSB Narrative, that summarizes the crash report. For this variable, use the Text Explorer to produce a DTM by phrasing and terming.
- 4. Create a Word Cloud. In the FDA_Enforcement_Actions.jmp file, the variable Citation Description describes the violation. For this variable, use the Text Explorer to produce a DTM by phrasing and terming.
- 5. Create a Word Cloud. The traffic-violation_jun2015.jmp is similar to the file used in the chapter except that the data is for June 2015 only. For the variable Description, use the Text Explorer to produce a DTM by phrasing and terming.
- 6. Create a Word Cloud. How does this compare to data for December 2014? Perform Latent Semantic Analytics, Topic Analysis, and Cluster Analysis on the DTM you produced in Problem 1. Perform Latent Semantic Analytics, Topic Analysis, and Cluster Analysis on the DTM you produced in Problem 2.
- 7. Perform Latent Semantic Analytics, Topic Analysis, and Cluster Analysis on the DTM that you produced in Problem 3. Perform Latent Semantic Analytics, Topic Analysis, and Cluster Analysis on the DTM you produced in Problem 4. Perform Latent Semantic Analytics, Topic Analysis, and Cluster Analysis on the DTM you produced in Problem 5. Perform Latent Semantic Analytics, Topic Analytics, Topic Analysis, and Cluster Analysis, and Cluster Analysis on the DTM you produced in Problem 5. Perform Latent Semantic Analytics, Topic Analysis, and Cluster Analysis, and Cluster Analysis on the DTM you produced in Problem 6. How does this compare to data for December 2014?
- 8. Similar to the predictive model that you did in the chapter, create a predictive model for violation type. How does this compare to data for December 2014?

¹ The authors would like to thank Daniel Valente and Christopher Gotwalt for their guidance and insight in writing this chapter.

About This Book

What Does This Book Cover?

This book focuses on the business statistics intelligence component of business analytics. It covers processes to perform a statistical study that may include data mining or predictive analytics techniques. Some real-world business examples of using these techniques are as follows:

- target marketing
- customer relation management
- market basket analysis
- cross-selling
- market segmentation
- customer retention
- improved underwriting
- quality control
- competitive analysis
- fraud detection and management
- churn analysis

Specific applications can be found at <u>http://www.jmp.com/software/success</u>. The bottom line, as reported by the KDNuggets poll (2008), is this: The median return on investment for data mining projects is in the 125–150% range. (See <u>http://www.kdnuggets.com/polls/2008/roi-data-mining.htm</u>.)

This book is *not* an introductory statistics book, although it does introduce basic data analysis, data visualization, and analysis of multivariate data. For the most part, your introductory statistics course has not completely prepared you to move on to real-world statistical analysis. The primary objective of this book is, therefore, to provide a bridge from your introductory statistics course to practical statistical analysis. This book is also not a highly technical book that dives deeply into the theory or algorithms, but it will provide insight into the "black box" of the methods covered. Analytics techniques covered by this book include the following:

- regression
- ANOVA
- logistic regression
- principal component analysis
- LASSO and Elastic Net
- cluster analysis
- decision trees
- *k*-nearest neighbors
- neural networks

- bootstrap forests and boosted trees
- text mining
- association rules

Is This Book for You?

This book is designed for the student who wants to prepare for his or her professional career and who recognizes the need to understand both the concepts and the mechanics of predominant analytic modeling tools for solving real-world business problems. This book is designed also for the practitioner who wants to obtain a hands-on understanding of business analytics to make better decisions from data and models, and to apply these concepts and tools to business analytics projects.

This book is for you if you want to explore the use of analytics for making better business decisions and have been either intimidated by books that focus on the technical details, or discouraged by books that focus on the high-level importance of using data without including the how-to of the methods and analysis.

Although not required, your completion of a basic course in statistics will prove helpful. Experience with the book's software, JMP Pro 13, is not required.

What's New in This Edition?

This second edition includes six new chapters. The topics of these new chapters are dirty data, LASSO and elastic net, *k*-nearest neighbors, bootstrap forests and boosted trees, text mining, and association rules. All the old chapters from the first edition are updated to JMP 13. In addition, more end-of-chapter exercises are provided.

What Should You Know about the Examples?

This book includes tutorials for you to follow to gain hands-on experience with SAS.

Software Used to Develop the Book's Content

JMP Pro 13 is the software used throughout this book.

Example Code and Data

You can access the example code and data for this book by linking to its author page at <u>http://support.sas.com/authors</u>. Some resources, such as instructor resources and add-ins used in the book, can be found on the JMP User Community file exchange at <u>https://community.jmp.com</u>.

Where Are the Exercise Solutions?

We strongly believe that for you to obtain maximum benefit from this book you need to complete the examples in each chapter. At the end of each chapter are suggested exercises so that you can practice what has been discussed in the chapter. Professors and instructors can obtain the exercise solutions by requesting them through the authors' SAS Press webpages at http://support.sas.com/authors.

We Want to Hear from You

SAS Press books are written by SAS Users for SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit <u>http://support.sas.com/publishing</u> to do the following:

- sign up to review a book
- recommend a topic
- request authoring information
- provide feedback on a book

Do you have questions about a SAS Press book that you are reading? Contact the author through saspress@sas.com or http://support.sas.com/author_feedback.

SAS has many resources to help you find answers and expand your knowledge. If you need additional help, see our list of resources: <u>http://support.sas.com/publishing</u>.

About These Authors



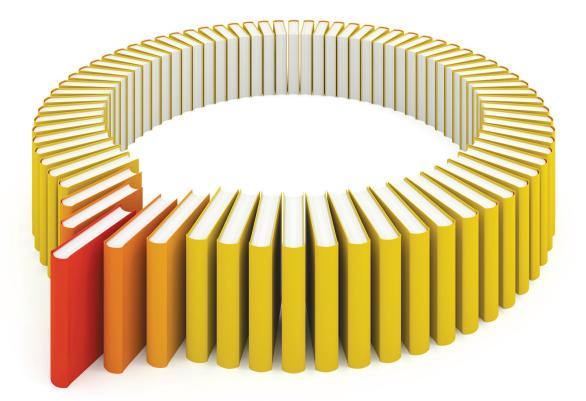
Ron Klimberg, PhD, is a professor at the Haub School of Business at Saint Joseph's University in Philadelphia, PA. Before joining the faculty in 1997, he was a professor at Boston University, an operations research analyst at the U.S. Food and Drug Administration, and an independent consultant. His current primary interests include multiple criteria decision making, data envelopment analysis, data visualization, data mining, and modeling in general. Klimberg was the 2007 recipient of the Tengelmann Award for excellence in scholarship, teaching, and research. He received his PhD from Johns Hopkins University and his MS from George Washington University.



B. D. McCullough, PhD, is a professor at the LeBow College of Business at Drexel University in Philadelphia, PA. Before joining Drexel, he was a senior economist at the Federal Communications Commission and an assistant professor at Fordham University. His research interests include applied econometrics and time series analysis, statistical and econometrics software accuracy, research replicability, and data mining. He received his PhD from The University of Texas at Austin.

Learn more about these authors by visiting their author pages, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more: <u>http://support.sas.com/publishing/authors/klimberg.html</u> <u>http://support.sas.com/publishing/authors/mccullough.html</u>

xviii



Gain Greater Insight into Your SAS[®] Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.





SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. @ indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0413