IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

April 5,7 2016

Week 13 Presentation

Week 13 Topic: Agenda

- Final Project Overview
- Proxy statement analysis using Yahoo

- Intro to Text Analytics

Additional FYI items:

- Python
- Wine example in R

Week 13 Topic:
Final Project Overview – 2 Parts

PART 1 (Week 13 Assignment): Document/Table Parsing (Automation)

- This is automation of Week 11 DEF 14A (Proxy) Filling – Executive Compensation Trend analysis. In essence, you are to automate Week 11 assignment using code. Using XML's readHTMLTable, RCurl, or other available functions in R.
- Make sure to collect all available DEF 14A filings available for the companies. Make sure to chart the salary, option awards, and/or stock awards over the available years for the company.
- Highlight any stock trading activity changes before and after filing.

PART 2 (Week 14 Assignment): Corpus & Term Document Matrix creation

- Create a corpus of filings per company. The Corpus is to house the Q&A sections for each filing. Striped of punctuation, etc.
- Generate a Term Document Matrix of the Q&A content.
- Highlight any patterns or trends found in analyzing the results of the Q&A Term Document Matrix statistics.

Week 13 Topic: Yahoo DEF 14A filings

- Yahoo has 18 filings from 1997 to 2015.

EDGAR Search Results

https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=0001011006&type=def+14a&dateb=&owner=exclude&count=40

Home | Latest Filings | Previous Page

U.S. Securities and Exchange Commission

EDGAR Search Results

EDGAR Search Results BETA View

SEC Home » Search the Next-Generation EDGAR System » Company Search » *Current Page*

YAHOO INC CIK#: 0001011006 (see all company filings)
SIC: 7373 - SERVICES-COMPUTER INTEGRATED SYSTEMS DESIGN
State location: CA | State of Inc.: **DE** | Fiscal Year End: 1231
(Assistant Director Office: 3)
Note: Ownership filings are available at this link. Reports containing extracted ownership data are temporarily unavailable.

Business Address
YAHOO! INC.
701 FIRST AVENUE
SUNNYVALE CA 94089
4083493300

Mailing Address
701 FIRST AVENUE
SUNNYVALE CA 94089

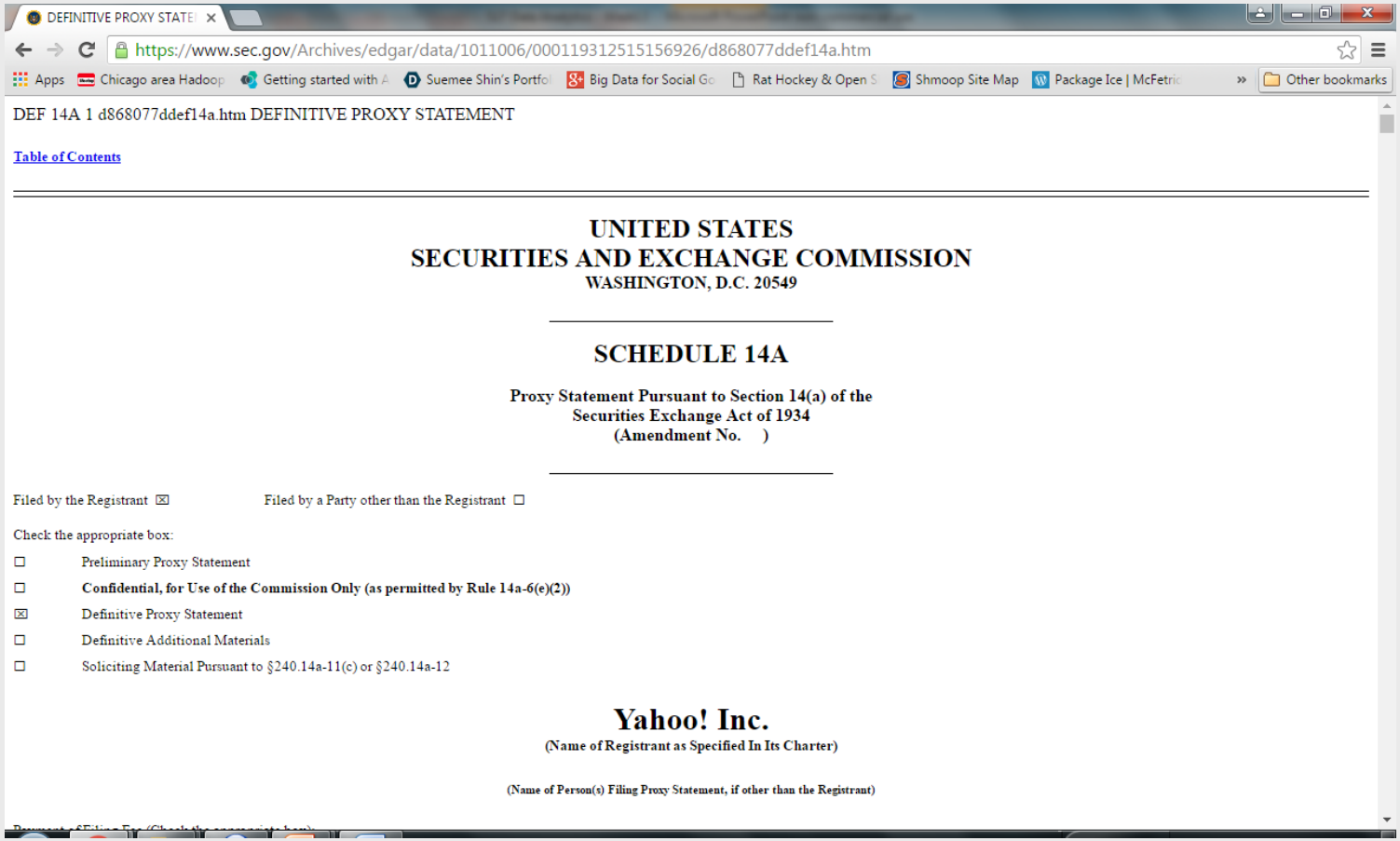
Filter Results: Filing Type: def 14a | Prior to: (YYYYMMDD) | Ownership? include exclude only | Limit Results Per Page 40 Entries | Search | Show All

Items 1 - 18 RSS Feed

Filings	Format	Description	Filing Date	File/Film Number
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-15-156926 (34 Act) Size: 2 MB	2015-04-29	000-28018 15813744
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-14-172132 (34 Act) Size: 2 MB	2014-04-30	000-28018 14798942
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-13-187918 (34 Act) Size: 1 MB	2013-04-30	000-28018 13799046
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-12-258870 (34 Act) Size: 1 MB	2012-06-04	000-28018 12886969
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-11-118858 (34 Act) Size: 1 MB	2011-04-29	000-28018 11795903
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-10-099552 (34 Act) Size: 1 MB	2010-04-29	000-28018 10782547
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-09-092231 (34 Act) Size: 1 MB	2009-04-29	000-28018 09780142
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0000891618-07-000262 (34 Act) Size: 1 MB	2007-04-30	000-28018 07802049
DEF 14A	Documents	Other definitive proxy statements	2006-04-14	000-28018

Week 13 Topic:
Yahoo DEF 14A - html version

https://www.sec.gov/Archives/edgar/data/1011006/000119312515156926/d868077ddef14a.htm:



DEF 14A 1 d868077ddef14a.htm DEFINITIVE PROXY STATEMENT

Table of Contents

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
WASHINGTON, D.C. 20549

SCHEDULE 14A

Proxy Statement Pursuant to Section 14(a) of the Securities Exchange Act of 1934
(Amendment No.)

Filed by the Registrant ☒ Filed by a Party other than the Registrant ☐

Check the appropriate box:

☐ Preliminary Proxy Statement
☐ **Confidential, for Use of the Commission Only (as permitted by Rule 14a-6(e)(2))**
☒ Definitive Proxy Statement
☐ Definitive Additional Materials
☐ Soliciting Material Pursuant to §240.14a-11(c) or §240.14a-12

Yahoo! Inc.
(Name of Registrant as Specified In Its Charter)

(Name of Person(s) Filing Proxy Statement, if other than the Registrant)

Week 13 Topic:
Yahoo DEF 14A - txt version

https://www.sec.gov/Archives/edgar/data/1011006/000119312515156926/0001193125-15-156926.txt:

https://www.sec.gov/Archives/edgar/data/1011006/000119312515156926/0001193125-15-156926.txt

```
<SEC-DOCUMENT>0001193125-15-156926.txt : 20150429
<SEC-HEADER>0001193125-15-156926.hdr.sgml : 20150429
<ACCEPTANCE-DATETIME>20150429170414
ACCESSION NUMBER:               0001193125-15-156926
CONFORMED SUBMISSION TYPE:      DEF 14A
PUBLIC DOCUMENT COUNT:          25
CONFORMED PERIOD OF REPORT:     20150624
FILED AS OF DATE:               20150429
DATE AS OF CHANGE:              20150429
EFFECTIVENESS DATE:             20150429

FILER:

        COMPANY DATA:
                COMPANY CONFORMED NAME:                 YAHOO INC
                CENTRAL INDEX KEY:                      0001011006
                STANDARD INDUSTRIAL CLASSIFICATION:     SERVICES-COMPUTER INTEGRATED SYSTEMS DESIGN [7373]
                IRS NUMBER:                             770398689
                STATE OF INCORPORATION:                 DE
                FISCAL YEAR END:                        1231

        FILING VALUES:
                FORM TYPE:              DEF 14A
                SEC ACT:                1934 Act
                SEC FILE NUMBER:        000-28018
                FILM NUMBER:            15813744

        BUSINESS ADDRESS:
                STREET 1:               YAHOO! INC.
                STREET 2:               701 FIRST AVENUE
                CITY:                   SUNNYVALE
                STATE:                  CA
                ZIP:                    94089
                BUSINESS PHONE:         4083493300

        MAIL ADDRESS:
                STREET 1:               701 FIRST AVENUE
                CITY:                   SUNNYVALE
                STATE:                  CA
                ZIP:                    94089
</SEC-HEADER>
<DOCUMENT>
<TYPE>DEF 14A
<SEQUENCE>1
```

Week 13 Topic: Yahoo's Executive Compensation Table

Target Table in HTML View:

Summary Compensation Table—2012–2014

The following table presents 2012–2014 summary compensation information for our Named Executive Officers. As required by SEC rules, stock awards (RSUs) and option awards are shown as compensation for the year in which they were granted (even if they have multi-year vesting schedules), and are valued based on their grant date fair values for accounting purposes. Accordingly, the table includes stock and option awards granted in the years shown even if they were scheduled to vest in later years, and even if they were subsequently forfeited (such as upon the executive's termination). Therefore, the stock and option columns do *not* report whether the officer realized a financial benefit from the awards (such as by vesting in stock or exercising options).

Name and Principal Position	Year	Salary ($)(1)	Bonus ($)(1)	Stock Awards ($)(2)(3)(4)	Option Awards ($)(2)(3)	Non-Equity Incentive Plan Compensation ($)(5)	All Other Compensation ($)(6)	Total ($)(2)
Marissa A. Mayer	2014	1,000,000	0	11,752,355(7)	28,194,288(7)	1,108,800	28,065	42,083,508
Chief Executive Officer	2013	1,000,000	2,250	8,312,316	13,847,283	1,700,000	73,863	24,935,712
	2012	454,862	0	35,000,002	0	1,120,000	40,540	36,615,404
Ken Goldman	2014	600,000	0	2,813,080	9,327,427	300,000	4,549	13,045,056
Chief Financial Officer	2013	600,000	0	2,597,612	2,290,527	500,000	4,615	5,992,754
	2012	116,667	100,000	7,262,357	0	0	29	7,479,053
David Filo	2014	1	0	0	0	0	0	1
Co-Founder and Chief Yahoo	2013	1	0	0	0	0	0	1
	2012	1	0	0	0	0	0	1
Ronald S. Bell	2014	600,000	0	3,282,107	0	300,000	4,549	4,186,656
General Counsel	2013	600,000	0	3,896,386	0	450,000	4,615	4,951,001
	2012	442,763	206,800	558,300	0	443,200	4,424	1,655,487
Henrique de Castro(8)	2014	27,083	0	0	0	0	1,177,157	1,204,240
Former Chief Operating Officer	2013	600,000	0	0	10,307,359	0	37,001	10,944,360
	2012	84,092	1,100,000	37,999,991	0	0	29	39,184,112

(1) Salary and bonus columns include amounts earned in, or awarded for performance during, the specified year (even if paid out early in the following year).

(2) As required by SEC rules, the stock and option award columns present the aggregate grant date fair value of equity awards granted during the years shown as computed for accounting purposes in accordance with FASB ASC 718. As a result, the stock and option columns (as well as the total column) include awards that have not yet vested, awards that were granted but later forfeited (such as upon the executive's termination), and performance-based awards that failed to vest; therefore, these columns are *not* intended as presentations of pay actually realized by the executive. For information on the assumptions used in the grant date fair value computations, refer to Note 14—"Employee Benefits" in the Notes to Consolidated Financial Statements in our 2014 Form 10-K.

71

Week 13 Topic: Yahoo's Executive Compensation Table

Target Table in Source Code View:

```
6002
6003 <div style="width:97%; margin-top:1.5%; margin-left:1.5%; margin-right:-1.25%">
6004  <P STYLE="margin-top:0pt; margin-bottom:0pt; font-size:9pt; font-family:arial" ALIGN="right">EXECUTIVE COMPENSATION </P>
6005  <p STYLE="margin-top:0pt;margin-bottom:0pt ; font-size:8pt"> </P></div>
6006  <p STYLE="margin-top:0pt;margin-bottom:0pt ; font-size:8pt"> 
6007 </P> <P STYLE="margin-top:0pt; margin-bottom:0pt; font-size:16pt; font-family:arial"><FONT COLOR="#7300ff"><B><A NAME="toc868077_29"></A>COMPENSATION TABLES </B></FONT>
     </P> <P STYLE="font-size:6pt;margin-top:0pt;margin-bottom:0pt"> </P>
6008 <P STYLE="line-height:1.0pt;margin-top:0pt;margin-bottom:2pt;border-bottom:1.00pt solid #000000"> </P> <P STYLE="margin-top:12pt; margin-bottom:0pt; text-indent:6%;
     font-size:10pt; font-family:arial" ALIGN="justify">The tables on the following
6009 pages present compensation information regarding our Chief Executive Officer, Marissa A. Mayer; our Chief Financial Officer, Ken Goldman; our co-founder and Chief Yahoo,
     David Filo; and our General Counsel, Ronald S. Bell. As required by SEC rules,
6010 the tables also include our former Chief Operating Officer, Henrique de Castro, whose service ended during 2014. These five individuals are our &#147;Named Executive
     Officers.&#148; We did not have any other executive officers in 2014. </P>
6011 <P STYLE="margin-top:12pt; margin-bottom:0pt; text-indent:6%; font-size:10pt; font-family:arial" ALIGN="justify">As required by SEC rules, in these tables performance-
     based awards are treated as having been granted in the year in which their
6012 performance goals were established (and if an award has multiple performance periods, the portion relating to each period is treated as a separate grant). </P> <P
     STYLE="margin-top:18pt; margin-bottom:0pt; font-size:16pt; font-family:arial"><FONT
6013  COLOR="#7300ff"><B>Summary Compensation Table&#151;2012&#150;2014 </B></FONT></P> <P STYLE="font-size:6pt;margin-top:0pt;margin-bottom:0pt"> </P>
6014 <P STYLE="line-height:1.0pt;margin-top:0pt;margin-bottom:2pt;border-bottom:1.00pt solid #000000"> </P> <P STYLE="margin-top:12pt; margin-bottom:0pt; text-indent:6%;
     font-size:10pt; font-family:arial" ALIGN="justify"><I></I>The following table
6015 presents 2012&#150;2014 summary compensation information for our Named Executive Officers. As required by SEC rules, stock awards (RSUs) and option awards are shown as
     compensation for the year in which they were granted (even if they have
6016 multi-year vesting schedules), and are valued based on their grant date fair values for accounting purposes. Accordingly, the table includes stock and option awards
     granted in the years shown even if they were scheduled to vest in later years, and
6017 even if they were subsequently forfeited (such as upon the executive&#146;s termination). Therefore, the stock and option columns do <I>not</I> report whether the officer
     realized a financial benefit from the awards (such as by vesting in stock or
6018 exercising options). <I> </I></P> <P STYLE="font-size:12pt;margin-top:0pt;margin-bottom:0pt"> </P>
6019 <TABLE CELLSPACING="0" CELLPADDING="0" WIDTH="100%" BORDER="0" STYLE="BORDER-COLLAPSE:COLLAPSE; font-family:arial; font-size:8pt" ALIGN="center">
6020
6021
6022 <TR>
6023 <TD WIDTH="36%"></TD>
6024 <TD VALIGN="bottom" WIDTH="1%"></TD>
```

ITM - 527

Week 13 Topic: Extract tables from HTML

Using readHTMLTable:

> install.packages("XML")

>library(XML)

> proxy.yahoo.2015.HTML.tables <- readHTMLTable("C:/Users/Desktop/yahoo_proxy2015.html")

> proxy.yahoo.2015.HTML.page[6003:6020] ← *to view Summary Compensation Table section in HTML*

Week 13 Topic:
Select the Executive Compensation Table

> compensation.table <- proxy.yahoo.2015.HTML.tables[[306]] ← *to view Summary Compensation Table in the array. Table # = 306 for 2015 Yahoo DEF 14A filing*

Filter

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
1																
2	Name and Principal Position	Â	Year	Â	Â	Salary($)(1)	Â	Â	Bonus($)(1)	Â	Â	StockAwards($)(2)(3)(4)	Â	Â	OptionAwards($)(2)(3)	Â
3	Marissa A. Mayer	Â	Â	2014	Â Â	Â	Â	1,000,000	Â Â	Â	Â	0	Â Â	Â	Â	11,752,355
4	Chief Executive Officer	Â	Â	2013	Â Â	Â	Â	1,000,000	Â Â	Â	Â	2,250	Â Â	Â	Â	8,312,316
5		Â	Â	2012	Â Â	Â	Â	454,862	Â Â	Â	Â	0	Â Â	Â	Â	35,000,002
6	Ken Goldman	Â	Â	2014	Â Â	Â	Â	600,000	Â Â	Â	Â	0	Â Â	Â	Â	2,813,080
7	Chief Financial Officer	Â	Â	2013	Â Â	Â	Â	600,000	Â Â	Â	Â	0	Â Â	Â	Â	2,597,612
8		Â	Â	2012	Â Â	Â	Â	116,667	Â Â	Â	Â	100,000	Â Â	Â	Â	7,262,357
9	David Filo	Â	Â	2014	Â Â	Â	Â	1	Â Â	Â	Â	0	Â Â	Â	Â	0
10	Co-Founder and Chief Yahoo	Â	Â	2013	Â Â	Â	Â	1	Â Â	Â	Â	0	Â Â	Â	Â	0
11		Â	Â	2012	Â Â	Â	Â	1	Â Â	Â	Â	0	Â Â	Â	Â	0
12	Ronald S. Bell	Â	Â	2014	Â Â	Â	Â	600,000	Â Â	Â	Â	0	Â Â	Â	Â	3,282,107

Week 13 Topic:
Sub-select relevant table data, rename

> compensation.table.data <- compensation.table[3:5, c("V1", "V4","V8","V12","V16")]

	V1	V4	V8	V12	V16
3	Marissa A. Mayer	2014	1,000,000	0	11,752,355
4	Chief Executive Officer	2013	1,000,000	2,250	8,312,316
5		2012	454,862	0	35,000,002

> compensation.table.data <-rename(compensation.table.data, c("V1"="Name and Principal Position", "V4"="Year", "V8"="Salary", "V12"="Stock_Awards", "V16"="Option_Awards"))

	Name and Principal Position	Year	Salary	Stock_Awards	Option_Awards
3	Marissa A. Mayer	2014	1,000,000	0	11,752,355
4	Chief Executive Officer	2013	1,000,000	2,250	8,312,316
5		2012	454,862	0	35,000,002

Week 13 Topic: Other methods

Using getURL:

> install.packages("RCurl")

>library(RCurl)

> proxy.yahoo.2015.HTML.page <- getURL("https://www.sec.gov/Archives/edgar/data/1011006/000119312515156926/d868077ddef14a.htm")

…

Using readLines:

> proxy.yahoo.2015.HTML.page <- readLines("https://www.sec.gov/Archives/edgar/data/1011006/000119312515156926/d868077ddef14a.htm")

>length(proxy.yahoo.2015.HTML.page)

[1] 10261

Week 13 Topic: Installing plyr

Install:

> install.packages("plyr")

Open library:

>library(plyr)

Getting help:

>library(help=plyr)

```
m_ply                   Call function with arguments in array or data
                        frame, discarding results.
maply                   Call function with arguments in array or data
                        frame, returning an array.
mapvalues               Replace specified values with new values, in a
                        vector or factor.
match_df                Extract matching rows of a data frame.
mdply                   Call function with arguments in array or data
                        frame, returning a data frame.
mlply                   Call function with arguments in array or data
                        frame, returning a list.
mutate                  Mutate a data frame by adding new or replacing
                        existing columns.
name_rows               Toggle row names between explicit and implicit.
ozone                   Monthly ozone measurements over Central
                        America.
plyr                    plyr: the split-apply-combine paradigm for R.
plyr-deprecated         Deprecated Functions in Package plyr
progress_text           Text progress bar.
progress_time           Text progress bar with time.
progress_tk             Graphical progress bar, powered by Tk.
progress_win            Graphical progress bar, powered by Windows.
r_ply                   Replicate expression and discard results.
raply                   Replicate expression and return results in a
                        array.
rbind.fill              Combine data.frames by row, filling in missing
                        columns.
rbind.fill.matrix       Bind matrices by row, and fill missing columns
                        with NA.
rdply                   Replicate expression and return results in a
                        data frame.
rename                  Modify names by name, not position.
revalue                 Replace specified values with new values, in a
                        factor or character vector.
```

Week 13 Topic: Installing RCurl

Install:

>install.package("RCurl")

Open library:

>library(RCurl)

Getting help:

>library(help=RCurl)

```
Index:

AUTH_ANY                Constants for identifying Authentication
                        Schemes
CFILE                   Create a C-level handle for a file
CURLHandle-class        Class "CURLHandle" for synchronous HTTP
                        requests
CurlFeatureBits         Constants for libcurl
HTTP_VERSION_1_0        Symbolic constants for specifying HTTP and SSL
                        versions in libcurl
MultiCURLHandle-class   Class "MultiCURLHandle" for asynchronous,
                        concurrent HTTP requests
base64                  Encode/Decode base64 content
basicHeaderGatherer     Functions for processing the response header of
                        a libcurl request
basicTextGatherer       Cumulate text across callbacks (from an HTTP
                        response)
binaryBuffer            Create internal C-level data structure for
                        collecting binary data
chunkToLineReader       Utility that collects data from the HTTP reply
                        into lines and calls user-provided function.
clone                   Clone/duplicate an object
coerce,numeric,NetrcEnum-method
                        Internal functions
complete                Complete an asynchronous HTTP request
curlError               Raise a warning or error about a CURL problem
curlEscape              Handle characters in URL that need to be
                        escaped
curlGlobalInit          Start and stop the Curl library
curlOptions             Constructor and accessors for CURLOptions
                        objects
curlPerform             Perform the HTTP query
curlSetOpt              Set values for the CURL options
curlVersion             Information describing the Curl library
dynCurlReader           Dynamically determine content-type of body from
                        HTTP header and set body reader
```

Week 13 Topic: Installing RCurl (cont.)

Install:

>install.package("RCurl")

Open library:

>library(RCurl)

Getting help:

>library(help=RCurl)

```
fileUpload                  Specify information about a file to upload in
                            an HTTP request
findHTTPHeaderEncoding
                            Find the encoding of the HTTP response from the
                            HTTP header
ftpUpload                   Upload content via FTP
getBinaryURL                Download binary content
getBitIndicators            Operate on bit fields
getCurlErrorClassNames
                            Retrieve names of all curl error classes
getCurlHandle               Create libcurl handles
getCurlInfo                 Access information about a CURL request
getFormParams               Extract parameters from a form query string
getURIAsynchronous          Download multiple URIs concurrently, with
                            inter-leaved downloads
getURL                      Download a URI
guessMIMEType               Infer the MIME type from a file name
httpPUT                     Simple high-level functions for HTTP PUT and
                            DELETE
merge.list                  Method for merging two lists by name
mimeTypeExtensions          Mapping from extension to MIME type
postForm                    Submit an HTML form
reset                       Generic function for resetting an object
scp                         Retrieve contents of a file from a remote host
                            via SCP (Secure Copy)
url.exists                  Check if URL exists
|,BitwiseValue,BitwiseValue-method
                            Classes and coercion methods for enumerations
                            in libcurl
```

Week 13 Topic: Installing stringr

Install:

>*install.package("stringr")*

Open library:

>*library(stringr)*

Getting help:

>*library(help=stringr)*

```
Index:

case                    Convert case of a string.
invert_match            Switch location of matches to location of
                        non-matches.
modifiers               Control matching behaviour with modifier
                        functions.
str_c                   Join multiple strings into a single string.
str_conv                Specify the encoding of a string.
str_count               Count the number of matches in a string.
str_detect              Detect the presence or absence of a pattern in
                        a string.
str_dup                 Duplicate and concatenate strings within a
                        character vector.
str_extract             Extract matching patterns from a string.
str_length              The length of a string.
str_locate              Locate the position of patterns in a string.
str_match               Extract matched groups from a string.
str_order               Order or sort a character vector.
str_pad                 Pad a string.
str_replace             Replace matched patterns in a string.
str_replace_na          Turn NA into "NA"
str_split               Split up a string into pieces.
str_sub                 Extract and replace substrings from a character
                        vector.
str_subset              Keep strings matching a pattern.
str_trim                Trim whitespace from start and end of string.
str_wrap                Wrap strings into nicely formatted paragraphs.
stringr                 Fast and friendly string manipulation.
word                    Extract words from a sentence.

Further information is available in the following vignettes in directory
'C:/Users/sshin/Documents/R/R-3.2.2/library/stringr/doc':
```

Week 13 Topic: Installing XML

Install:

>install.package("XML")

Open library:

>library(XML)

Getting help:

>library(help=XML)

```
Index:

Doctype                     Constructor for DTD reference
Doctype-class               Class to describe a reference to an XML DTD
ExternalReference-class
                            Classes for working with XML Schema
SAXState-class              A virtual base class defining methods for SAX
                            parsing
XMLAttributes-class         Class '"XMLAttributes"'
XMLCodeFile-class           Simple classes for identifying an XML document
                            containing R code
XMLInternalDocument-class
                            Class to represent reference to C-level data
                            structure for an XML document
XMLNode-class               Classes to describe an XML node object.
[.XMLNode                   Convenience accessors for the children of
                            XMLNode objects.
[<-.XMLNode                 Assign sub-nodes to an XML node
addChildren                 Add child nodes to an XML node
addNode                     Add a node to a tree
append.xmlNode              Add children to an XML node
asXMLNode                   Converts non-XML node objects to XMLTextNode
                            objects
asXMLTreeNode               Convert a regular XML node to one for use in a
                            "flat" tree
catalogLoad                 Manipulate XML catalog contents
catalogResolve              Look up an element via the XML catalog
                            mechanism
coerce,XMLHashTreeNode,XMLHashTree-method
                            Transform between XML representations
compareXMLDocs              Indicate differences between two XML documents
docName                     Accessors for name of XML document
dtdElement                  Gets the definition of an element or entity
                            from a DTD.
```

Week 13 Topic: Installing XML (cont.)

Install:

>install.package("XML")

Open library:

>library(XML)

Getting help:

>library(help=XML)

getXMLErrors	Get XML/HTML document parse errors
isXMLString	Facilities for working with XML strings
length.XMLNode	Determine the number of children in an XMLNode object.
libxmlVersion	Query the version and available features of the libxml library.
makeClassTemplate	Create S4 class definition based on XML node(s)
names.XMLNode	Get the names of an XML nodes children.
newXMLDoc	Create internal XML node or document object
newXMLNamespace	Add a namespace definition to an XML node
parseDTD	Read a Document Type Definition (DTD)
parseURI	Parse a URI string into its elements
parseXMLAndAdd	Parse XML content and add it to a node
print.XMLAttributeDef	Methods for displaying XML objects
processXInclude	Perform the XInclude substitutions
readHTMLList	Read data in an HTML list or all lists in a document
readHTMLTable	Read data from one or more HTML tables
readKeyValueDB	Read an XML property-list style document
readSolrDoc	Read the data from a Solr document
removeXMLNamespaces	Remove namespace definitions from a XML node or document
saveXML	Output internal XML Tree
setXMLNamespace	Set the name space on a node
startElement.SAX	Generic Methods for SAX callbacks
supportsExpat	Determines which native XML parsers are being used.
toHTML	Create an HTML representation of the given R object, using internal C-level nodes
toString.XMLNode	Creates string representation of XML node
xmlApply	Applies a function to each of the children of an XMLNode
xmlAttributeType	The type of an XML attribute for element from the DTD

Week 13 Topic: Installing tm

Install:

>install.package("tm")

Open library:

>library(tm)

Getting help:

>library(help=tm)

```
Index:

Corpus                  Corpora
DataframeSource         Data Frame Source
DirSource               Directory Source
Docs                    Access Document IDs and Terms
MC_tokenizer            Tokenizers
PCorpus                 Permanent Corpora
PlainTextDocument       Plain Text Documents
Reader                  Readers
Source                  Sources
TermDocumentMatrix      Term-Document Matrix
TextDocument            Text Documents
URISource               Uniform Resource Identifier Source
VCorpus                 Volatile Corpora
VectorSource            Vector Source
WeightFunction          Weighting Function
XMLSource               XML Source
XMLTextDocument         XML Text Documents
ZipSource               ZIP File Source
Zipf_plot               Explore Corpus Term Frequency Characteristics
acq                     50 Exemplary News Articles from the
                        Reuters-21578 Data Set of Topic acq
c.VCorpus               Combine Corpora, Documents, Term-Document
                        Matrices, and Term Frequency Vectors
content_transformer     Content Transformers
crude                   20 Exemplary News Articles from the
                        Reuters-21578 Data Set of Topic crude
findAssocs              Find Associations in a Term-Document Matrix
findFreqTerms           Find Frequent Terms
getTokenizers           Tokenizers
getTransformations      Transformations
inspect                 Inspect Objects
meta                    Metadata Management
plot.TermDocumentMatrix
```

Week 13 Topic: Installing tm (cont.)

Install:

>install.package("tm")

Open library:

>library(tm)

Getting help:

>library(help=tm)

```
plot.TermDocumentMatrix
                        Visualize a Term-Document Matrix
readDOC                 Read In a MS Word Document
readPDF                 Read In a PDF Document
readPlain               Read In a Text Document
readRCV1                Read In a Reuters Corpus Volume 1 Document
readReut21578XML        Read In a Reuters-21578 XML Document
readTabular             Read In a Text Document
readTagged              Read In a POS-Tagged Word Text Document
readXML                 Read In an XML Document
read_dtm_Blei_et_al     Read Document-Term Matrices
removeNumbers           Remove Numbers from a Text Document
removePunctuation       Remove Punctuation Marks from a Text Document
removeSparseTerms       Remove Sparse Terms from a Term-Document Matrix
removeWords             Remove Words from a Text Document
stemCompletion          Complete Stems
stemDocument            Stem Words
stopwords               Stopwords
stripWhitespace         Strip Whitespace from a Text Document
termFreq                Term Frequency Vector
tm_filter               Filter and Index Functions on Corpora
tm_map                  Transformations on Corpora
tm_reduce               Combine Transformations
tm_term_score           Compute Score for Matching Terms
weightBin               Weight Binary
weightSMART             SMART Weightings
weightTf                Weight by Term Frequency
weightTfIdf             Weight by Term Frequency - Inverse Document
                        Frequency
writeCorpus             Write a Corpus to Disk

Further information is available in the following vignettes in
directory 'C:/Users/sshin/Documents/R/R-3.2.2/library/tm/doc':

extensions: Extensions (source, pdf)
tm: Introduction to the tm Package (source, pdf)
```

Week 13 Topic:
Week 12 Assignment

1) Post your, one or two page, coding and analysis steps from Week 11 assignment in the discussion board. Use a word document format. Do not post the company and analysis, just the analysis steps and any coding performed.

2) Post midterm project presentation or word document and annotated code in discussion board.

Week 13 Topic:
Final Project Alternative

WHEN

Friday, April 29
2:00 – 4:00pm

WHERE

The Idea Shop at Illinois Tech

3440 S. Dearborn Chicago, IL

Have you been working on an innovative project during your time at Illinois Tech? Have you participated in an IPRO that solves a local or global issue? Have you created a solution to an everyday problem or developed a really cool invention ?

COME ONE, COME ALL AND REGISTER FOR THE ITM STUDENT INNOVATION & PROTOTYPE COMPETITION!

- **Must be an Illinois Tech Student, all majors are accepted!**
- **Teams may consist of 1-5 members**
- **The solution can be a concept, product, service, etc.**
- **Provide a 5-10 minute presentation; be able to answer inquiries based on judging criteria**
- **Showcase a poster board and demonstrate a prototype**
- **Deadline to register is Sunday, April 17 at midnight**
- **PRIZES! Cool Techie prizes will be given to the winning team!**

Week 13 Topic:
Final Project Google Sign Up Sheet

Once the sign up sheet is up, sign up for the 11 teams/groups, 5 students each. Expect the following:

- Some adjustments may be made dependent on the coding versus analytics capabilities of the team members.
- Teams will need to accommodate online students. *Online students, let me know if there is a preference on time zone, all online student team, mixed online/live student teams, etc.*

Note on Scoring:

- The final project will be graded on a 100 point scale. Use of Discussion Topics is highly recommended. 5 out of the 100 points will be dedicated to the assessment of the teams use of Discussion Topic for information sharing.

Week 13 Topic: Definition of Text Analytics

- The terms text analytics, text data mining, and text mining will be used synonymously in this course.
- Text analytics uses algorithms for turning free-form text into data that can then be analyzed by applying statistical and machine learning methods, as well as natural language processing techniques.
- Text analytics encompasses many sub areas - pattern discovery or exploratory analysis and predictive modeling, as it pertains to text analytics. We will discuss several topics in these areas.

Often the most challenging part of the data mining process is obtaining and preprocessing the data…

Week 13 Topic: Text Mining

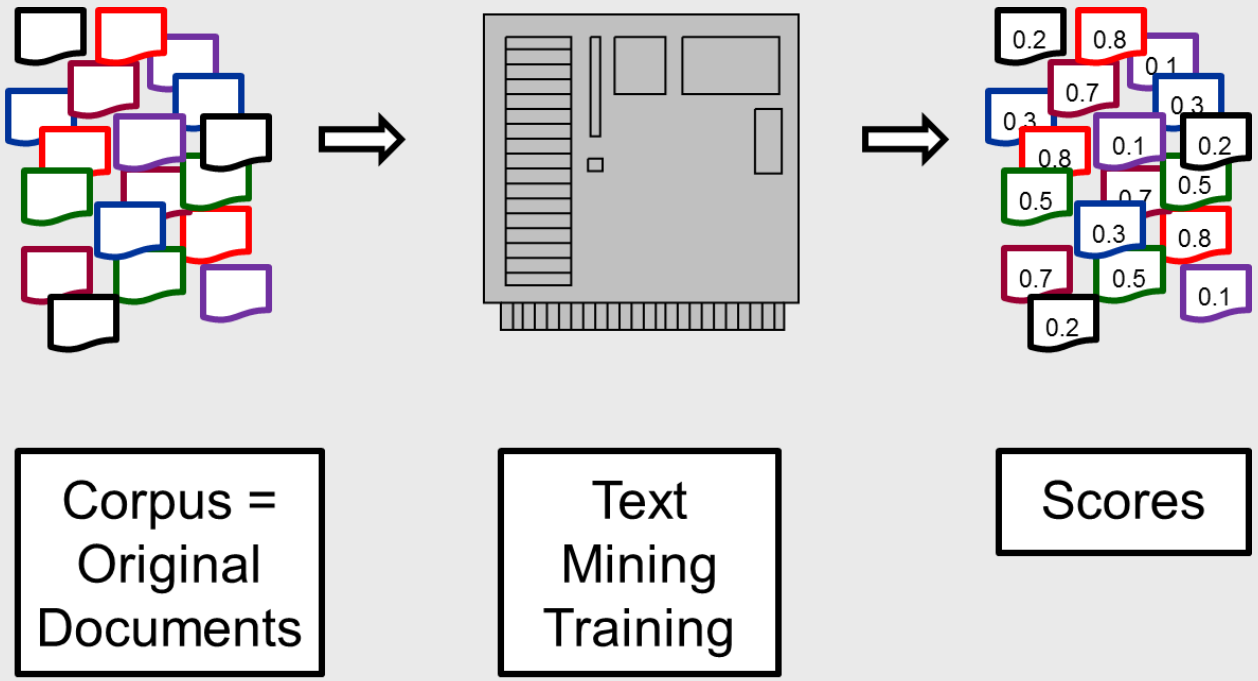
- Text mining as presented here has the following characteristics:
  - operates with respect to a ***corpus*** of documents
  - creates a ***dictionary*** or ***vocabulary*** to identify relevant terms
  - accommodates a variety of ***metrics*** to quantify the contents of a document within the corpus
  - derives a ***structured vector*** of measurements for each document relative to the corpus
  - uses ***analytical methods*** that are applied to the structured vector of measurements based on the goals of the analysis (for example, groups documents into segments)

Week 13 Topic: Text Mining (cont.)

- The concept of a dictionary can be thought of as a *vocabulary*. The document collection has a vocabulary that is the union of all the terms contained in each document. Consequently, text mining uses ***dictionary*** or ***vocabulary*** to refer to the collection of terms that are used in the analysis.
- Terms not in the dictionary are ignored, except possibly for use in determining the relative frequencies of terms in each document. ***Zipf's Law***, discussed in a later chapter, helps identify terms in a dictionary that should be included in an analysis.
- Text mining works with a collection of documents, ***corpus***. The collection can be dynamic, that is, documents can be added to the collection. You can use the collection to train a model, and you can apply the model to new documents coming into the collection.
- New documents are ***scored*** relative to how they compare to the original documents in the collection. If a new document contains a new term, then text mining is ignorant of this new term until that document is used in a new training step.

Week 13 Topic: Text Mining (cont.)

Many commercial text mining products have strong text-analytics capabilities, most lack data mining capabilities beyond text analytics. The ability to score new documents using a decision tree or a neural network presents new opportunities to improve text mining outcomes (for example, making it possible to use variables derived from text analytics in predictive models).



Week 13 Topic:
Data Mining – two broad areas

- Pattern Discovery/Exploratory Analysis (Unsupervised Learning): There is no target variable, and some form of analysis is performed to do the following:
  - identify or define homogeneous groups, clusters, or segments
  - find links or associations between entities, as in market basket analysis

- Prediction (Supervised Learning): A target variable is used, and some form of predictive or classification model is developed.
  - input variables are associated with values of a target variable, and the model produces a predicted target value for a given set of inputs.

Data mining analysts know how a predictive model scores new data. However, some analysts might be unaware that unsupervised learning models (that is, data without a known, available target) can also generate scores, and new data can be scored using the model. For example, a new document is scored by calculating the probability of membership in each cluster, and then it is assigned to the cluster associated with the highest probability.

Week 13 Topic: Text Mining Applications - Unsupervised

- **Information retrieval**
  - finding documents with relevant content of interest
  - used for researching medical, scientific, legal, and news documents such as books and journal articles
- **Document categorization for organizing**
  - clustering documents into naturally occurring groups
  - extracting themes or concepts
- **Anomaly detection**
  - identifying unusual documents that might be associated with cases requiring special handling such as unhappy customers, fraud activity, and so on

Week 13 Topic: Text Mining Applications - Supervised

- Many typical predictive modeling or classification applications can be enhanced by incorporating textual data in addition to traditional input variables.
  - churning propensity models that include customer center notes, website forms, e-mails, and Twitter messages
  - hospital admission prediction models incorporating medical records notes as a new source of information
  - insurance fraud modeling using adjustor notes
  - sentiment categorization from customer comments
  - stylometry or forensic applications that identify the author of a particular writing sample

Week 13 Topic:
Text Mining Signal versus Noise

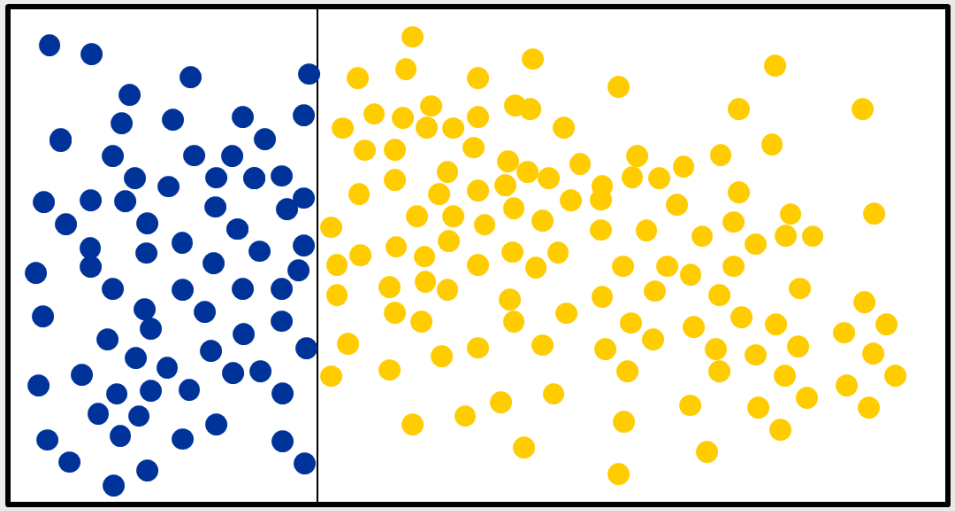
Psychologists know that human beings might react differently to the same stimulus if sufficient time elapses between exposures. On Monday, when you are hungry at lunchtime, you eat a sandwich. Yet, on Tuesday when you are hungry, you opt for a salad. This tendency for different outcomes to occur with similar inputs is attributed to noise, which is unpredictable. You can predict with almost certainty that you will eat lunch next Thursday, but you cannot predict what you will eat with the same certainty.

Analytic experts expect errors in prediction related to noise, so methods are developed to minimize errors in the presence of noise. The incremental value that text mining can provide predictive models should be assessed by comparing the quality of a model without incorporating text mining to that achieved after text mining is added.

- Target = Signal + Noise
- Signal = Systematic Variation = Predictable
- Noise = Random Variation = Unpredictable

Week 13 Topic:
Text Mining Signal versus Noise

The graphic illustrates the pure signal situation. In this case, the training data can be perfectly separated into primary or secondary outcomes using a linear decision boundary. You rarely expect to see this in practice.

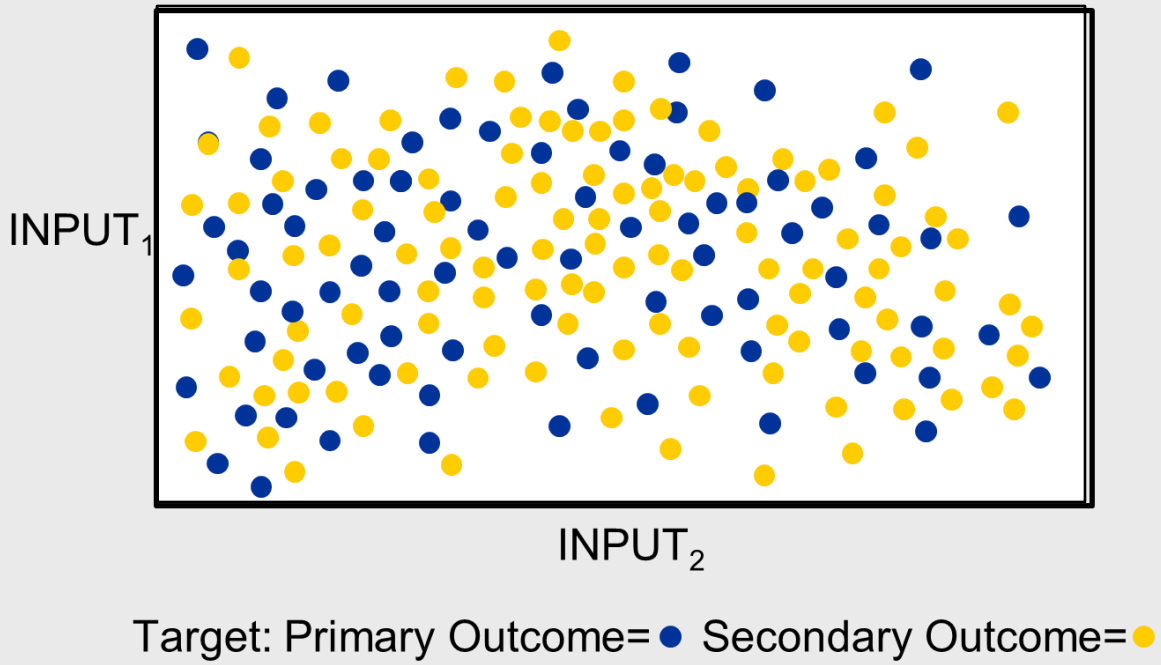


INPUT

Target: Primary Outcome= ● Secondary Outcome= ●

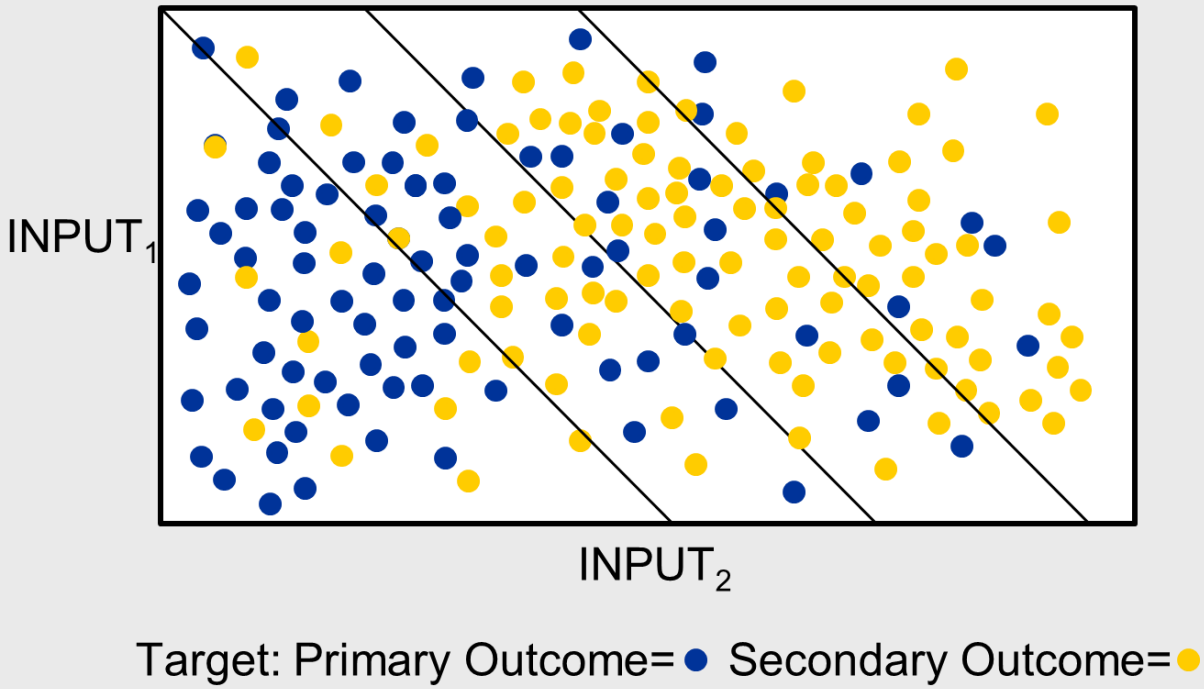
Week 13 Topic:
Text Mining Signal versus Noise

At the other extreme is the pure noise situation. In this case, the training data appears to have no patterns upon which to base a model that can separate the primary outcomes from the secondary outcomes. This situation is more common than you might like. Although pure signal is very rare, pure noise can actually occur in practice.



Week 13 Topic:
Text Mining Signal versus Noise

The most common situation in practice is a mixture of signal and noise. You can predict more accurately than randomly guessing. How well you predict depends on whether data is dominated by systematic variation or random variation.



Week 13 Topic:
Text Mining – Perfect Separation

Some document collections are well separated for analytic purposes. The hypothetical example shows eight documents, with four that describe national news items exclusively, and the remaining four describing international news items exclusively. Suppose that you could identify a set of terms that are associated with national news and another set of terms associated with international news. These terms could then be used to classify the documents in the corpus.

Document ID	National News # Words	International News # Words	Document Subject
1	3	0	National
2	5	0	National
3	7	0	National
4	8	0	National
5	0	4	International
6	0	5	International
7	0	3	International
8	0	7	International

Perfect Separation: No Mixing of Subjects

Week 13 Topic:
Text Mining – Imperfect Separation

With the same topic and analytic objective, another document collection has documents that might mention a heterogeneous set of news articles. You still get good separation, but noise creeps in due to the fact that a document can include multiple subjects.

Document ID	National News # Words	International News # Words	Document Subject
11	3	1	National
12	8	2	National
13	7	6	Mixed
14	8	1	National
15	1	4	International
16	2	5	International
17	3	3	Mixed
18	1	7	International

Good Separation: Little Mixing of Subjects

Week 13 Topic: Text Mining – Poor Separation

Finally, the example shows that if you have a collection of documents that mention many topics and mixes topics, then trying to classify documents into clean categories is difficult.

Document ID	National News # Words	International News # Words	Document Subject
21	3	4	Mixed
22	8	2	National
23	7	6	Mixed
24	8	1	National
25	4	4	Mixed
26	6	5	Mixed
27	3	3	Mixed
28	1	7	International

Poor Separation: Substantial Mixing of Subjects

Week 13 Topic: Text Analytics References in R and SAS

Using R - Preparation for our next assignment:

- http://handsondatascience.com/TextMiningO.pdf
- https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
- http://cran.us.r-project.org/doc/Rnews/Rnews_2008-2.pdf
- https://rstudio-pubs-static.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html
- http://www.r-bloggers.com/intro-to-text-analysis-with-r/

Using Base SAS - Preparation for our next assignment:

- http://support.sas.com/resources/papers/proceedings12/133-2012.pdf

SAS related FYI as we don't have access to the modules:

- https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf
- https://support.sas.com/resources/papers/Benchmark_R_Mahout_SAS.pdf
- https://support.sas.com/resources/papers/proceedings12/137-2012.pdf

Week 13 Topic: tm package overview

- Initiate package: > *library(tm)*
- Chracteristics:
  - Create a corpus – a collection of text documents
  - Provide various preprocessing operations e.g., stemDoc(), stripWhitespace(), tmTolower()
  - Create a Document-Term matrix
  - Inspect / manipulate the Document-Term matrix (e.g. convert into a data frame needed by classifiers)
  - Train a classifier on pre-classified Document-Term data frame
  - Apply the trained classifier on new text documents to obtain class predictions and evaluate performance

Reference:
http://web.letras.up.pt/bhsmaia/EDV/apresentacoes/Bradzil_Classif_withTM.pdf

Week 13 Topic: Introducing Python

http://en.wikipedia.org/wiki/Python_(programming_language):

- Python was created by *Guido Van Rossem* in 1991 and emphasizes productivity and code readability. Programmers that want to delve into data analysis or apply statistical techniques are some of the main users of Python for statistical purposes.
- The closer you get to working in an engineering environment, the more likely it is you might prefer Python. It's a flexible language that is great to do something novel, and given its focus on readability and simplicity, its learning curve is relatively low.
- Similar to R, Python has packages as well. PyPi (https://pypi.python.org/pypi) is the Python Package index and consists of libraries to which users can contribute. Just like R, Python has a great community but it is a bit more scattered, since it's a general purpose language. Nevertheless, Python for data science is rapidly claiming a more dominant position in the Python universe: the expectations are growing and more innovative data science applications will see their origin here.

Week 13 Topic: Getting started with Python

- You can use Python when your data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database. Being a fully fledged programming language, it's a great tool to implement algorithms for production use.
- While the infancy of Python packages for data analysis was an issue in the past, this has improved significantly over the years. Make sure to install NumPy /SciPy (scientific computing) and pandas (data manipulation) to make Python usable for data analysis. Also have a look at matplotlib to make graphics, and scikit-learn for machine learning.
- Unlike R, Python has no clear "winning" IDE. We recommend you to have a look at Spyder,IPython Notebook and Rodeo to see which one best fits your needs.

Week 13 Topic: Python Pros and Cons

- ***Pro: IPython Notebook:*** The IPython Notebook makes it easier to work with Python and data. You can easily share notebooks with colleagues, without having them to install anything. This drastically reduces the overhead of organizing code, output and notes files. This will allow you to spend more time doing real work.
- ***Pro: A general purpose language:*** Python is a general purpose language that is easy and intuitive. This gives it a relatively flat learning curve, and it increases the speed at which you can write a program. In short, you need less time to code and you have more time to play around with it!
- Furthermore, the Python testing framework is a built-in, low-barrier-to-entry testing framework that encourages good test coverage. This guarantees your code is reusable and dependable.
- ***Pro: A multi purpose language:*** Python brings people with different backgrounds together. As a common, easy to understand language that is known by programmers and that can easily be learnt by statisticians, you can build a single tool that integrates with every part of your workflow.
- ***Pro/Con: Visualizations:*** Visualizations are an important criteria when choosing data analysis software. Although Python has some nice visualization libraries, such as Seaborn, Bokeh and Pygal, there are maybe too many options to choose from. Moreover, compared to R, visualizations are usually more convoluted, and the results are not always so pleasing to the eye.
- ***Con: Python is a challenger:*** Python is a challenger to R. It does not offer an alternative to the hundreds of essential R packages. Although it's catching up, it's still unclear if this will make people give up R?

Week 13 Topic:
Wine example in R – Install skmeans

Follow directions and install the skmeans package. Choose a mirror location close to you.

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> install.packages("skmeans",dependencies = TRUE)
--- Please select a CRAN mirror for use in this session ---
also installing the dependencies 'slam', 'clue'

trying URL 'https://rweb.crmda.ku.edu/cran/bin/windows/contrib/3.2/slam_0.1-32.zip'
Content type 'application/zip' length 111465 bytes (108 KB)
downloaded 108 KB

trying URL 'https://rweb.crmda.ku.edu/cran/bin/windows/contrib/3.2/clue_0.3-50.zip'
Content type 'application/zip' length 782678 bytes (764 KB)
downloaded 764 KB

trying URL 'https://rweb.crmda.ku.edu/cran/bin/windows/contrib/3.2/skmeans_0.2-7.zip'
Content type 'application/zip' length 209565 bytes (204 KB)
downloaded 204 KB

package 'slam' successfully unpacked and MD5 sums checked
package 'clue' successfully unpacked and MD5 sums checked
package 'skmeans' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\sshin\AppData\Local\Temp\RtmpiSQ8Hl\downloaded_packages
> |
```

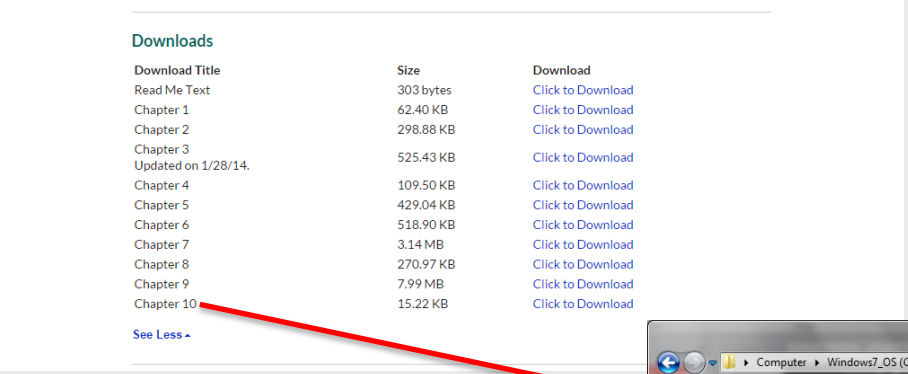
HTTPS CRAN mirror

- 0-Cloud [https]
- Austria [https]
- Chile [https]
- China (Beijing 4) [https]
- Colombia (Cali) [https]
- France (Lyon 2) [https]
- France (Paris 2) [https]
- Germany (Münster) [https]
- Iceland [https]
- Mexico (Mexico City) [https]
- Russia (Moscow) [https]
- Spain (A Coruña) [https]
- Switzerland [https]
- UK (Bristol) [https]
- UK (Cambridge) [https]
- USA (CA 1) [https]
- USA (KS) [https]
- USA (MD) [https]
- USA (MI 1) [https]
- USA (TN) [https]
- USA (TX) [https]
- USA (WA) [https]
- (HTTP mirrors)

OK Cancel

Week 13 Topic:
Wine example in R – WineKMC.csv

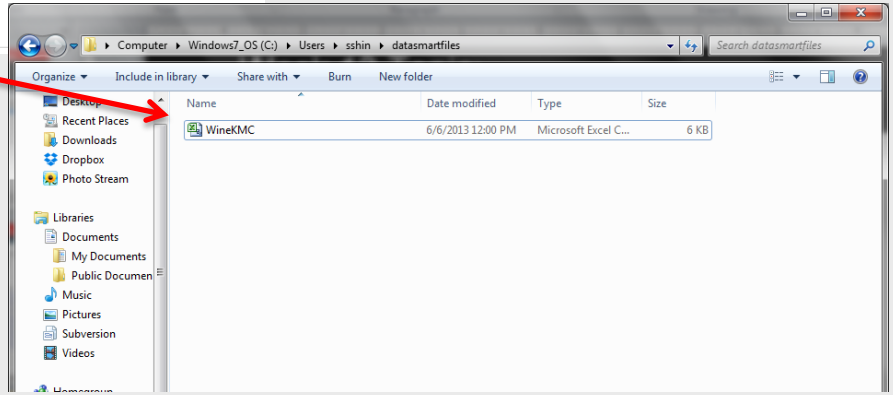
From the http://www.wiley.com/go/datasmart website, download Chapter 10 WineKMC.csv file in the R working directory. Remember, this is a different file for use with R code. The previous download was WineKMC.xls.



Downloads

Download Title	Size	Download
Read Me Text	303 bytes	Click to Download
Chapter 1	62.40 KB	Click to Download
Chapter 2	298.88 KB	Click to Download
Chapter 3 Updated on 1/28/14.	525.43 KB	Click to Download
Chapter 4	109.50 KB	Click to Download
Chapter 5	429.04 KB	Click to Download
Chapter 6	518.90 KB	Click to Download
Chapter 7	3.14 MB	Click to Download
Chapter 8	270.97 KB	Click to Download
Chapter 9	7.99 MB	Click to Download
Chapter 10	15.22 KB	Click to Download

See Less



Week 13 Topic:
Wine example in R – WineKMC.csv

WineKMC - Microsoft Excel non-commercial use

Offer	Mth	Varietal	MinQty	Disc	Origin	PastPeak	Adams	Allen	Anders	Bailey	Baker	Barnes	Bell	Bennett	Brooks	Brown	Butler	Campbell	Ca
1	Jan	Malbec	72	56	France	FALSE											1		
2	Jan	Pinot Noir	72	17	France	FALSE							1					1	
3	Feb	Espumant	144	32	Oregon	TRUE									1				
4	Feb	Champagr	72	48	France	TRUE											1		
5	Feb	Cab. Sauv.	144	44	NZ	TRUE													
6	Mar	Prosecco	144	86	Chile	FALSE													
7	Mar	Prosecco	6	40	Australia	TRUE				1	1					1			
8	Mar	Espumant	6	45	S. Africa	FALSE								1	1				
9	Apr	Chardonn	144	57	Chile	FALSE		1											
10	Apr	Prosecco	72	52	CA	FALSE					1	1							
11	May	Champagr	72	85	France	FALSE									1				
12	May	Prosecco	72	83	Australia	FALSE													
13	May	Merlot	6	43	Chile	FALSE													
14	Jun	Merlot	72	64	Chile	FALSE													
15	Jun	Cab. Sauv.	144	19	Italy	FALSE													
16	Jun	Merlot	72	88	CA	FALSE													
17	Jul	Pinot Noir	12	47	Germany	FALSE							1						
18	Jul	Espumant	6	50	Oregon	FALSE	1												
19	Jul	Champagr	12	66	Germany	FALSE					1								
20	Aug	Cab. Sauv.	72	82	Italy	FALSE													
21	Aug	Champagr	12	50	CA	FALSE						1							
22	Aug	Champagr	72	63	France	FALSE						1			1		1		
23	Sept	Chardonn	144	39	S. Africa	FALSE													
24	Sept	Pinot Noir	6	34	Italy	FALSE			1				1					1	
25	Oct	Cab. Sauv.	72	59	Oregon	TRUE													
26	Oct	Pinot Noir	144	83	Australia	FALSE			1				1					1	

Week 13 Topic:
Wine example in R – Load winedata

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> getwd()
[1] "C:/Users/sshin/datasmartfiles"
> setwd("C:/Users/sshin/datasmartfiles")
> winedata <- read.csv("WineKMC.csv")
> winedata[,1:9]
   Offer  Mth   Varietal MinQty Disc    Origin PastPeak Adams Allen
1      1  Jan     Malbec     72   56    France    FALSE    NA    NA
2      2  Jan Pinot Noir     72   17    France    FALSE    NA    NA
3      3  Feb  Espumante    144   32    Oregon     TRUE    NA    NA
4      4  Feb  Champagne     72   48    France     TRUE    NA    NA
5      5  Feb Cab. Sauv.    144   44        NZ     TRUE    NA    NA
6      6  Mar   Prosecco    144   86     Chile    FALSE    NA    NA
7      7  Mar   Prosecco      6   40 Australia     TRUE    NA    NA
8      8  Mar  Espumante      6   45 S. Africa    FALSE    NA    NA
9      9  Apr Chardonnay    144   57     Chile    FALSE    NA     1
10    10  Apr   Prosecco     72   52        CA    FALSE    NA    NA
11    11  May  Champagne     72   85    France    FALSE    NA    NA
12    12  May   Prosecco     72   83 Australia    FALSE    NA    NA
13    13  May     Merlot      6   43     Chile    FALSE    NA    NA
14    14  Jun     Merlot     72   64     Chile    FALSE    NA    NA
15    15  Jun Cab. Sauv.    144   19     Italy    FALSE    NA    NA
16    16  Jun     Merlot     72   88        CA    FALSE    NA    NA
17    17  Jul Pinot Noir     12   47   Germany    FALSE    NA    NA
18    18  Jul  Espumante      6   50    Oregon    FALSE     1    NA
19    19  Jul  Champagne     12   66   Germany    FALSE    NA    NA
20    20  Aug Cab. Sauv.     72   82     Italy    FALSE    NA    NA
21    21  Aug  Champagne     12   50        CA    FALSE    NA    NA
22    22  Aug  Champagne     72   63    France    FALSE    NA    NA
23    23 Sept Chardonnay    144   39 S. Africa    FALSE    NA    NA
24    24 Sept Pinot Noir      6   34     Italy    FALSE    NA    NA
25    25  Oct Cab. Sauv.     72   59    Oregon     TRUE    NA    NA
26    26  Oct Pinot Noir    144   83 Australia    FALSE    NA    NA
27    27  Oct  Champagne     72   88        NZ    FALSE    NA     1
28    28  Nov Cab. Sauv.     12   56    France     TRUE    NA    NA
29    29  Nov  P. Grigio      6   87    France    FALSE     1    NA
30    30  Dec     Malbec      6   54    France    FALSE     1    NA
31    31  Dec  Champagne     72   89    France    FALSE    NA    NA
32    32  Dec Cab. Sauv.     72   45   Germany     TRUE    NA    NA
>
>
```

Week 13 Topic:
Wine example in R – Replace w ‘0’

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
11    11  May  Champagne     72   85    France    FALSE    NA    NA
12    12  May   Prosecco     72   83 Australia    FALSE    NA    NA
13    13  May     Merlot      6   43     Chile    FALSE    NA    NA
14    14  Jun     Merlot     72   64     Chile    FALSE    NA    NA
15    15  Jun Cab. Sauv.    144   19     Italy    FALSE    NA    NA
16    16  Jun     Merlot     72   88        CA    FALSE    NA    NA
17    17  Jul Pinot Noir     12   47   Germany    FALSE    NA    NA
18    18  Jul  Espumante      6   50    Oregon    FALSE     1    NA
19    19  Jul  Champagne     12   66   Germany    FALSE    NA    NA
20    20  Aug Cab. Sauv.     72   82     Italy    FALSE    NA    NA
21    21  Aug  Champagne     12   50        CA    FALSE    NA    NA
22    22  Aug  Champagne     72   63    France    FALSE    NA    NA
23    23 Sept Chardonnay    144   39 S. Africa    FALSE    NA    NA
24    24 Sept Pinot Noir      6   34     Italy    FALSE    NA    NA
25    25  Oct Cab. Sauv.     72   59    Oregon     TRUE    NA    NA
26    26  Oct Pinot Noir    144   83 Australia    FALSE    NA    NA
27    27  Oct  Champagne     72   88        NZ    FALSE    NA     1
28    28  Nov Cab. Sauv.     12   56    France     TRUE    NA    NA
29    29  Nov  P. Grigio      6   87    France    FALSE     1    NA
30    30  Dec     Malbec      6   54    France    FALSE     1    NA
31    31  Dec  Champagne     72   89    France    FALSE    NA    NA
32    32  Dec Cab. Sauv.     72   45   Germany     TRUE    NA    NA
>
> winedata[is.na(winedata)]<-0
> winedata[1:10,8:17]
   Adams Allen Anders Bailey Baker Barnes Bell Bennett Brooks Brown
1      0     0      0      0     0      0    0       0      0     0
2      0     0      0      0     0      0    1       0      0     0
3      0     0      0      0     0      0    0       0      1     0
4      0     0      0      0     0      0    0       0      0     0
5      0     0      0      0     0      0    0       0      0     0
6      0     0      0      0     0      0    0       0      0     0
7      0     0      0      1     1      0    0       0      0     1
8      0     0      0      0     0      0    0       1      1     0
9      0     1      0      0     0      0    0       0      0     0
10     0     0      0      0     1      1    0       0      0     0
> library(skmeans)
> ?skmeans
starting httpd help server ... done
>
```

Week 13 Topic:
Wine example in R – ?skmeans

SAS Information C | SAS Studio | Results: Characteri | Results: Characteri | Results: psam_h17 | Results: psam_h17 | R: Compute Spheri

127.0.0.1:18778/library/skmeans/html/skmeans.html

Apps | Chicago area Hadoo... | Getting started with... | Suemee Shin's Portf... | Big Data for Social ... | Rat Hockey & Open... | Shmoop Site Map | Package Ice | McFet... | Other bookmarks

skmeans {skmeans} R Documentation

Compute Spherical k-Means Partitions

Description

Partition given vectors *x_b* by minimizing the spherical *k*-means criterion *∑_{b,j} w_b u_{bj}^m d(x_b, p_j)* over memberships and prototypes, where the *w_b* are case weights, *u_{bj}* is the membership of *x_b* to class *j*, *p_j* is the *prototype* of class *j* (thus minimizing *∑_b w_b u_{bj}^m d(x_b, p)* over *p*), and *d* is the cosine dissimilarity *d(x, p) = 1 - \cos(x, p)*.

Usage

```
skmeans(x, k, method = NULL, m = 1, weights = 1, control = list())
```

Arguments

`x`
: A numeric data matrix, with rows corresponding to the objects to be partitioned (such that row *b* contains *x_b*). Can be a dense matrix, a simple triplet matrix (package **slam**), or a dgTMatrix (package **Matrix**). Zero rows are not allowed.

`k`
: an integer giving the number of classes to be used in the partition.

`method`
: a character string specifying one of the built-in methods for computing spherical *k*-means partitions, or a function to be taken as a user-defined method, or `NULL` (default value). If a character string, its lower-cased version is matched against the lower-cased names of the available built-in methods using `pmatch`. See **Details** for available built-in methods and defaults.

`m`
: a number not less than 1 controlling the softness of the partition (as the "fuzzification parameter" of the fuzzy *c*-means algorithm). The default value of 1 corresponds to hard partitions; values greater than one give partitions of increasing softness obtained from a generalized soft spherical *k*-means problem.

`weights`
: a numeric vector of non-negative case weights. Recycled to the number of objects given by `x` if necessary.

`control`
: a list of control parameters. See **Details**.

Details

Week 13 Topic:
Wine example in R - Transpose & Cluster

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> ncol(winedata)
[1] 107
> winedata.transposed <- t(winedata[,8:107])
> winedata.tramsposed[1:10,1:10]
Error: object 'winedata.tramsposed' not found
> ncol(winedata)
[1] 107
> winedata.transposed <- t(winedata[,8:107])
> winedata.transposed[1:10,1:10]
        [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
Adams      0    0    0    0    0    0    0    0    0     0
Allen      0    0    0    0    0    0    0    0    1     0
Anders     0    0    0    0    0    0    0    0    0     0
Bailey     0    0    0    0    0    0    1    0    0     0
Baker      0    0    0    0    0    0    1    0    0     1
Barnes     0    0    0    0    0    0    0    0    0     1
Bell       0    1    0    0    0    0    0    0    0     0
Bennett    0    0    0    0    0    0    0    1    0     0
Brooks     0    0    1    0    0    0    0    1    0     0
Brown      0    0    0    0    0    0    1    0    0     0
> winedata.clusters <- skmeans(winedata.transposed,3,method="genetic")
> winedata.clusters
A hard spherical k-means partition of 100 objects into 3 classes.
Class sizes: 17, 36, 47
Call: skmeans(x = winedata.transposed, k = 3, method = "genetic")
> str(winedata.clusters)
List of 7
 $ prototypes: num [1:3, 1:32] 0.09 0 0.167 0.331 0 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3] "1" "2" "3"
  .. ..$ : NULL
 $ membership: NULL
 $ cluster   : Named int [1:100] 2 3 1 2 3 3 1 2 3 2 ...
  ..- attr(*, "names")= chr [1:100] "Adams" "Allen" "Anders" "Bailey" ...
 $ family    :List of 7
  ..$ description: chr "spherical k-means"
  ..$ D          :function (x, prototypes)
  ..$ C          :function (x, weights, control)
  ..$ init       :function (x, k)
  ..$ e          : num 1
```

Week 13 Topic:
Wine example in R – Transpose & Count

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> aggregate(winedata.transposed,by=list(winedata.clusters$cluster),sum)
  Group.1 V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32
1       1  2  7  0  0  0  0  0  0  0   1   0   1   0   0   0   1   7   0   0   0   0   0   1  12   0  12   1   0   0   0   0   0
2       2  0  0  2  0  0  1 16 17  0   2   1   1   6   0   0   0   0  13   1   0   2   1   0   0   0   0   1   1  16  17   0   0
3       3  8  3  4 12  4 11  3  3 10   4  12   3   0   9   6   4   0   1   4   6   2  20   4   0   6   3   7   5   1   5  17   4
> winedata.clustercounts <-t(aggregate(winedata.transposed,by=list(winedata.clusters$cluster),sum)[,2:33])
> winedata.clustercounts
    [,1] [,2] [,3]
V1     2    0    8
V2     7    0    3
V3     0    2    4
V4     0    0   12
V5     0    0    4
V6     0    1   11
V7     0   16    3
V8     0   17    3
V9     0    0   10
V10    1    2    4
V11    0    1   12
V12    1    1    3
V13    0    6    0
V14    0    0    9
V15    0    0    6
V16    1    0    4
V17    7    0    0
V18    0   13    1
V19    0    1    4
V20    0    0    6
V21    0    2    2
V22    0    1   20
V23    1    0    4
V24   12    0    0
V25    0    0    6
V26   12    0    3
V27    1    1    7
V28    0    1    5
V29    0   16    1
V30    0   17    5
V31    0    0   17
V32    0    0    4
```

Week 13 Topic:
Wine example in R – Join to get Info

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
>
>
>
> winedata.desc.plus.counts[order(-winedata.desc.plus.counts[,8]),]
    Offer  Mth   Varietal MinQty Disc    Origin PastPeak  1  2  3
V24    24 Sept Pinot Noir      6   34     Italy    FALSE 12  0  0
V26    26  Oct Pinot Noir    144   83 Australia    FALSE 12  0  3
V2      2  Jan Pinot Noir     72   17    France    FALSE  7  0  3
V17    17  Jul Pinot Noir     12   47   Germany    FALSE  7  0  0
V1      1  Jan     Malbec     72   56    France    FALSE  2  0  8
V10    10  Apr   Prosecco     72   52        CA    FALSE  1  2  4
V12    12  May   Prosecco     72   83 Australia    FALSE  1  1  3
V16    16  Jun     Merlot     72   88        CA    FALSE  1  0  4
V23    23 Sept Chardonnay    144   39 S. Africa    FALSE  1  0  4
V27    27  Oct  Champagne     72   88        NZ    FALSE  1  1  7
V3      3  Feb  Espumante    144   32    Oregon     TRUE  0  2  4
V4      4  Feb  Champagne     72   48    France     TRUE  0  0 12
V5      5  Feb Cab. Sauv.    144   44        NZ     TRUE  0  0  4
V6      6  Mar   Prosecco    144   86     Chile    FALSE  0  1 11
V7      7  Mar   Prosecco      6   40 Australia     TRUE  0 16  3
V8      8  Mar  Espumante      6   45 S. Africa    FALSE  0 17  3
V9      9  Apr Chardonnay    144   57     Chile    FALSE  0  0 10
V11    11  May  Champagne     72   85    France    FALSE  0  1 12
V13    13  May     Merlot      6   43     Chile    FALSE  0  6  0
V14    14  Jun     Merlot     72   64     Chile    FALSE  0  0  9
V15    15  Jun Cab. Sauv.    144   19     Italy    FALSE  0  0  6
V18    18  Jul  Espumante      6   50    Oregon    FALSE  0 13  1
V19    19  Jul  Champagne     12   66   Germany    FALSE  0  1  4
V20    20  Aug Cab. Sauv.     72   82     Italy    FALSE  0  0  6
V21    21  Aug  Champagne     12   50        CA    FALSE  0  2  2
V22    22  Aug  Champagne     72   63    France    FALSE  0  1 20
V25    25  Oct Cab. Sauv.     72   59    Oregon     TRUE  0  0  6
V28    28  Nov Cab. Sauv.     12   56    France     TRUE  0  1  5
V29    29  Nov  P. Grigio      6   87    France    FALSE  0 16  1
V30    30  Dec     Malbec      6   54    France    FALSE  0 17  5
V31    31  Dec  Champagne     72   89    France    FALSE  0  0 17
V32    32  Dec Cab. Sauv.     72   45   Germany     TRUE  0  0  4
>
>
>
```