

Web Exercise 02: R and R-Studio

DUE Date: September 22, 3:30pm (on Blackboard).

Grade: 20 points

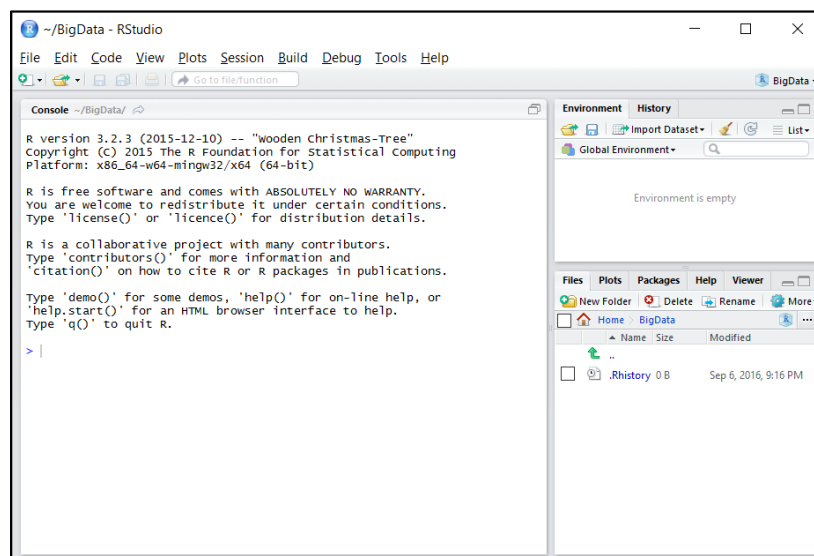
1. Use your own computer or login to the Computer Lab's account.
2. **Install the R Software into your local machine. (If you are using the Lab machine, you can skip this step).**

R is an open source, free software for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis (modified definition from Wikipedia). You can download the R software from here:

<https://cran.r-project.org/>

➔ Select the OS version for your computer, such as "Download R for Windows". ➔ select "base" ➔ Download R 3.3.1 for Windows (or newer version).

3. Install the RStudio (after the installation of R, if you are using the Lab machine, you can skip this step): **RStudio is a free and open-source integrated development environment (IDE) for R**, which provides several key functions (source code editor, code auto-completion, retrieving previous commands, debugger, interpreter, etc.) and graphic user interfaces (GUIs) for programming. To download RStudio, go to this link: <https://www.rstudio.com/>
➔ Download RStudio ➔ Select "RStudio Desktop (Free license)" ➔ select the installer for your OS (Windows or Mac OS). Then install the RStudio.
4. Launch the RStudio first. You will see three windows: Console, Environment, and Files/Plots.



5. Your first task is to use R as a calculator.

Type the following in the Console window. In the R language (and some other programming languages), the “#” sign means ‘**remark**’ for adding comments, notes, and explanation inside the R program. Any texts after the # sign will be ignored during the execution of the R program.

this is your first R exercise.

type basic math calculation as the following (or copy the whole paragraph and paste into the Console).

```
4 + 3          # "+" plus sign
50/4           # "/" divided sign
6*6            # "*" multiply sign
9 + 3*3        # basic calculation procedure
9^3            # "^" is denotes power. 9^3 = 9*9*9
```

After enter the math questions, Press Enter. You will see the Console showing the following results:

```
> 4 + 3        # "+" plus sign
[1] 7
> 50/4          # "/" divided sign
[1] 12.5
> 6*6           # "*" multiply sign
[1] 36
> 9 + 3*3       # basic calculation procedure
[1] 18
> 9^3           # "^" is denotes power. 9^3 = 9*9*9
[1] 729
>
```

There are several arithmetic operators and logical operators in R. You can find more operators in here. <http://www.statmethods.net/management/operators.html>

Arithmetic Operators

+	addition	-	subtraction
*	multiplication	/	division
^ or **	exponentiation	x %% y	modulus (x mod y) 7%%3 is 1
x %/% y	integer division 9%/%		

Logical Operators

<	less than	<=	less than or equal to
>	greater than	>=	greater than or equal to
==	exactly equal to	!=	not equal to
!x	Not x	x y	x OR y
x & y	x AND y	isTRUE(x)	test if X is TRUE

Now you can type three different math operation and see the results from R Console. (Your own exercise.).

The next task is how to enter data in R. There are several ways to enter data. The first one is to enter the data by hand. Let's assume that you have five students and each of them have their ages (X) and their mid-term exam scores (Y).

We can enter the data into R by using the following method ("<-" is the value assignment operator in R. "c" indicates a column of data (combine values into a Vector or List).

```
Age <- c(23, 20, 21, 22, 30)
Score <- c(83, 99, 80, 79, 90)
```

After entering the two statements, press Enter. Now you can see the two NEW VALUE showing on the Environment Window: Age and Score with their numbers.

Now you can easily summarize the two variables among the five students by typing the following:

```
summary (Age)
summary (Score)
```

You will see the results in the Console (with basic statistic summary of your two variables).

```
> summary (Age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 20.0   21.0   22.0   23.2   23.0   30.0

> summary (Score)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 79.0   80.0   83.0   86.2   90.0   99.0
```

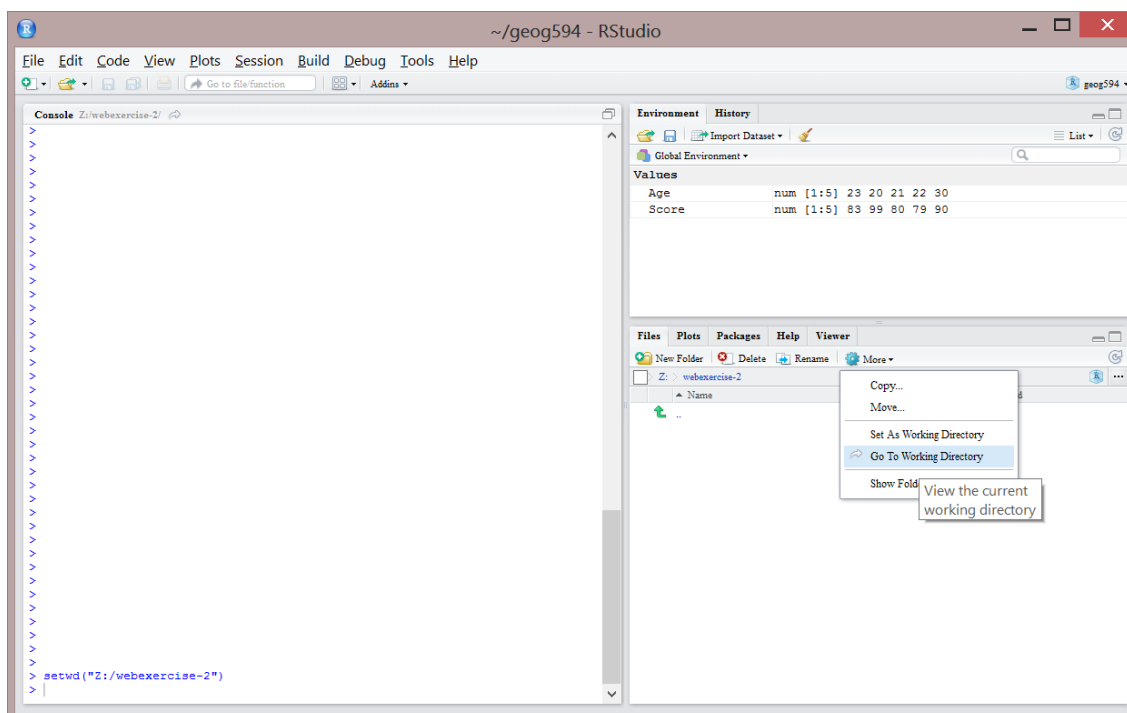
The second way to enter the data into R is to read data directly from a table using “read.table” command. The next task will teach you how to read a table into R.

Before we use the read.table function, we need to **set up the Working Directory** in R. The default working directory in Windows is the “Document” folder inside the User’s Personal Document Folder (such as C:\Users\mtsou\Documents\ in Windows). To easily handle the data in R, please create a dedicate folder in your local drive (D: drive or Z: drive if you are using SAL lab). → create a new folder in Z: (or D: or C:) drive called “**webexercise-2**”.

In the R console commend mode: type the following:

```
setwd("Z:/webexercise-2")           # set up working directory
```

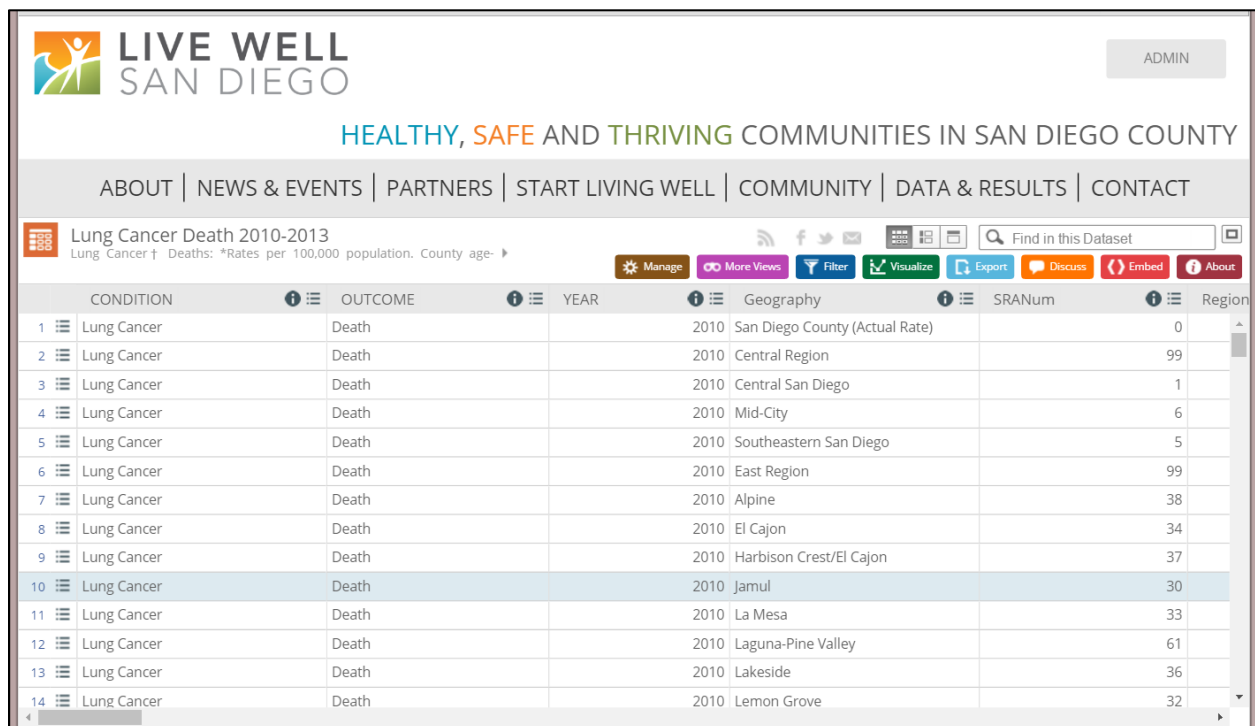
Now your R working director will be in the **Z:/webexercise-2** folder. You can save all your data and the outputs into this folder during this exercise. You can select the [File] window → click on “Files” window → More → Go To Working Directory. Currently, this directory is empty.



Reading Tables and Conduct Cancer Data Analysis

Recently, the County of San Diego has created a very nice **[Live Well] Data portal**, storing many useful public health data. We will try to download the data from the Data portal and then use R for some statistical analysis.

1. Open a web browser to access <https://data.livewellsd.org/>
2. Select the “Non-Communicable (Chronic) Disease” in the Icons.
3. In the new page, type “lung cancer” in the top search textbox.
4. Click on the “Lung Cancer Death 2010-2013”. You will see a data window like the following:



The screenshot shows the Live Well San Diego Data portal interface. At the top, there is a logo and navigation links. Below the logo, the text "HEALTHY, SAFE AND THRIVING COMMUNITIES IN SAN DIEGO COUNTY" is displayed. A navigation bar includes links for ABOUT, NEWS & EVENTS, PARTNERS, START LIVING WELL, COMMUNITY, DATA & RESULTS, and CONTACT. The main content area displays the "Lung Cancer Death 2010-2013" dataset. A search bar and various action buttons (Manage, More Views, Filter, Visualize, Export, Discuss, Embed, About) are visible. The data is presented in a table with columns for CONDITION, OUTCOME, YEAR, Geography, SRANum, and Region. The table lists 14 rows of data, with the 10th row highlighted.

	CONDITION	OUTCOME	YEAR	Geography	SRANum	Region
1	Lung Cancer	Death		2010 San Diego County (Actual Rate)		0
2	Lung Cancer	Death		2010 Central Region		99
3	Lung Cancer	Death		2010 Central San Diego		1
4	Lung Cancer	Death		2010 Mid-City		6
5	Lung Cancer	Death		2010 Southeastern San Diego		5
6	Lung Cancer	Death		2010 East Region		99
7	Lung Cancer	Death		2010 Alpine		38
8	Lung Cancer	Death		2010 El Cajon		34
9	Lung Cancer	Death		2010 Harbison Crest/El Cajon		37
10	Lung Cancer	Death		2010 Jamul		30
11	Lung Cancer	Death		2010 La Mesa		33
12	Lung Cancer	Death		2010 Laguna-Pine Valley		61
13	Lung Cancer	Death		2010 Lakeside		36
14	Lung Cancer	Death		2010 Lemon Grove		32

This data include all numbers of lung cancer death incidents in each Sub Regional Areas (SRA) in San Diego County. We will download the data for our further analysis.

Click on [Export] icon in the Data window → Download → CSV for Excel. Save the CSV file into the R working directory folder (for example: **Z:\webexercise-2**). If you are not using the SAL lab computers, please save to your own local R working directory folder.

You can open the file and take a look at the datasets.

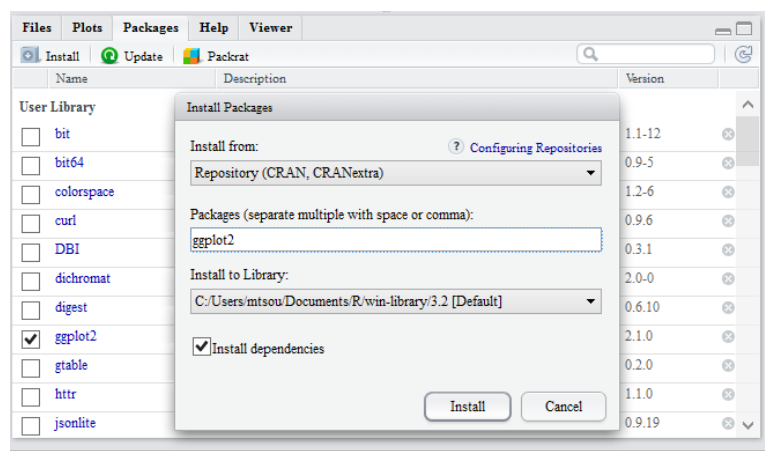
	A	B	C	D	E	F	G	H	I	J	K	L
1	CONDITION	OUTCOME	YEAR	Geography	SRANum	RegionNum	RegionName	Urbanicity	Urbanicity	SESSort	Socioecon	SDNUM
2	Lung Cancer	Death	2010	San Diego County (Actual Rate)	0	0	COUNTY	99		99		99
3	Lung Cancer	Death	2010	Central Region	99	99	REGION CE	99		99		99
4	Lung Cancer	Death	2010	Central San Diego	1	3	CENTRAL	5	Very Urban	2	Low Income	4
5	Lung Cancer	Death	2010	Mid-City	6	3	CENTRAL	5	Very Urban	1	Lowest Income	4
6	Lung Cancer	Death	2010	Southeastern San Diego	5	3	CENTRAL	5	Very Urban	2	Low Income	4
7	Lung Cancer	Death	2010	East Region	99	99	REGION EA	99		99		99
8	Lung Cancer	Death	2010	Alpine	38	5	EAST	1	Rural Area	5	High Income	2
9	Lung Cancer	Death	2010	El Cajon	34	5	EAST	4	Urban Area	2	Low Income	2
10	Lung Cancer	Death	2010	Harbison Crest/El Cajon	37	5	EAST	99		99		2
11	Lung Cancer	Death	2010	Jamul	30	5	EAST	1	Rural Area	6	Highest Income	2
12	Lung Cancer	Death	2010	La Mesa	33	5	EAST	4	Urban Area	2	Low Income	2
13	Lung Cancer	Death	2010	Laguna-Pine Valley	61	5	EAST	1	Rural Area	3	Moderately	2
14	Lung Cancer	Death	2010	Lakeside	36	5	EAST	2	Exurban Area	3	Moderately	2
15	Lung Cancer	Death	2010	Lemon Grove	32	5	EAST	5	Very Urban	2	Low Income	2
16	Lung Cancer	Death	2010	Mountain Empire	62	5	EAST	1	Rural Area	2	Low Income	2
17	Lung Cancer	Death	2010	Santee	35	5	EAST	3	Suburban Area	4	Moderately	2
18	Lung Cancer	Death	2010	Spring Valley	31	5	EAST	4	Urban Area	4	Moderately	2
19	Lung Cancer	Death	2010	North Central Region	99	99	REGION N	99		99		99
20	Lung Cancer	Death	2010	Coastal	11	2	N CENTRAL	3	Suburban Area	5	High Income	4
21	Lung Cancer	Death	2010	Del Mar-Mira Mesa	13	2	N CENTRAL	3	Suburban Area	6	Highest Income	3
22	Lung Cancer	Death	2010	Elliott-Navajo	17	2	N CENTRAL	2	Exurban Area	5	High Income	3
23	Lung Cancer	Death	2010	Kearny Mesa	10	2	N CENTRAL	4	Urban Area	3	Moderately	4
24	Lung Cancer	Death	2010	Miramar	16	2	N CENTRAL	99		99		3
25	Lung Cancer	Death	2010	Peninsula	2	2	N CENTRAL	4	Urban Area	3	Moderately	4
26	Lung Cancer	Death	2010	University	12	2	N CENTRAL	4	Urban Area	4	Moderately	3

Now in the R console, type the following commands to read CSV file:

```
mydata = read.csv("Lung_Cancer_Death_2010-2013.csv")
```

Now you can see the lung cancer data has been imported into the R. with 196 observations and 40 variables. Now go to the [Environment] window click on “mydata”, a data view window will open. (You can also type “View(mydata)” to take a quick look at the Cancer Data.

The next step is to conduct a deeper visualization analysis for this dataset. You will need to install a new library in R, called “**ggplot2**”. Go to the “File/Plots” window, → Select “**Packages**” → click on “**Install**” → type “**ggplot2**” in the Packages textbox, then click on “**Install**” button. You may see a few errors messages, you can ignore them and it will not have impact to this lab exercise.



After installing “ggplot2”, copy the following R-Script into your R Console, and press Enter to execute this R-Script. Try to read each command and understand their meanings.

```
#### Created by Jay Yang, HDMA@SDSU    Sep,2016    ----- ####

#### load required libraries
library(ggplot2)

### Read csv data into R dataframe
cancer_data <- read.csv("Lung_Cancer_Death_2010-2013.csv")

### list all the field names (Variables in the dataset)
names(cancer_data)

### Subset data by specific column
## subset only the year 2010
data_2010 <- cancer_data[cancer_data$YEAR == 2010,]
## subset by other fields (ex. Geography and RegionName)
# data_LaMesa <- cancer_data[cancer_data$Geography == 'La Mesa',]
# data_EAST <- cancer_data[cancer_data$RegionName == 'EAST',]

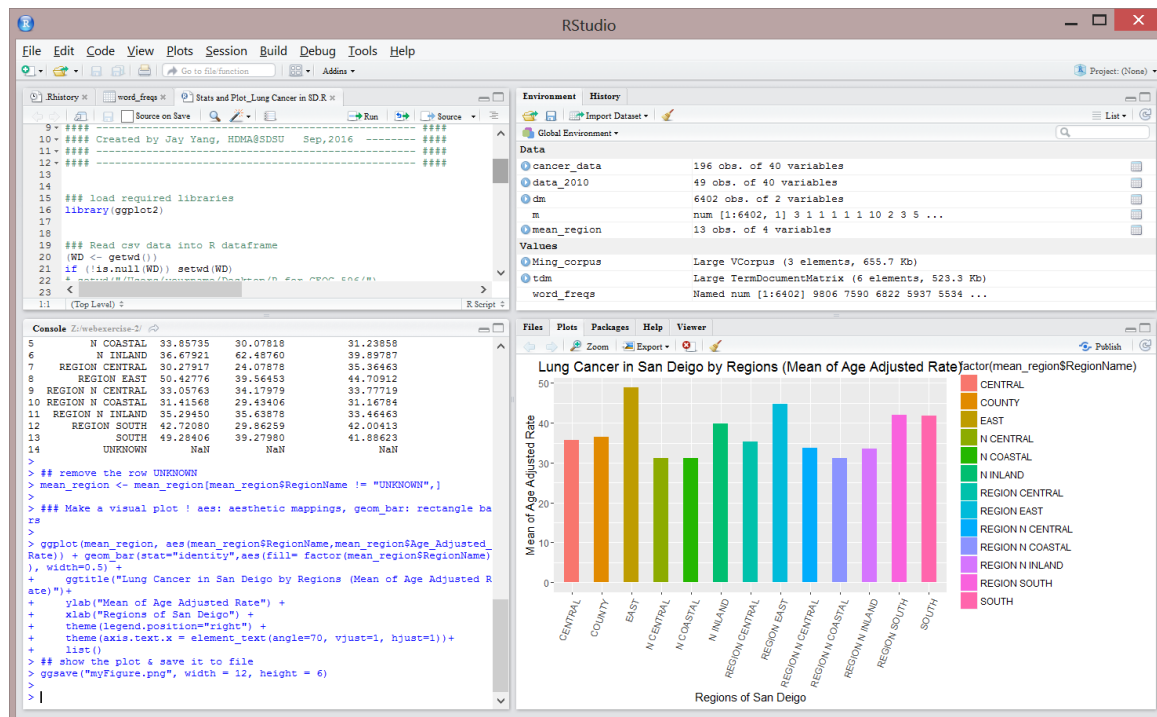
### Generate some statistics
## aggregate region by names, then calculate the mean for three rates
mean_region <- aggregate(data_2010[,
c("Male_Rate","Female_Rate","Age_Adjusted_Rate")], by=list(RegionName =
data_2010$RegionName), FUN=mean, na.rm=TRUE)
print(mean_region)

## remove the row UNKNOWN
mean_region <- mean_region[mean_region$RegionName != "UNKNOWN",]

### Make a visual plot ! aes: aesthetic mappings, geom_bar: rectangle bars

ggplot(mean_region,
aes(mean_region$RegionName,mean_region$Age_Adjusted_Rate)) +
geom_bar(stat="identity",aes(fill= factor(mean_region$RegionName)),
width=0.5) +
  ggtitle("Lung Cancer in San Deigo by Regions (Mean of Age Adjusted Rate)") +
  ylab("Mean of Age Adjusted Rate") +
  xlab("Regions of San Deigo") +
  theme(legend.position="right") +
  theme(axis.text.x = element_text(angle=70, vjust=1, hjust=1)) +
  list()

## show the plot & save it to file
ggsave("myFigure.png", width = 12, height = 6)
```



This R-script will create a “myFigure.png” image inside the working directory. You can check the image. If you like to know more about the function of “aggregate”, you can use the help command to understand their usages and arguments. Your Help Window will display these command information.

help (aggregate)

There are several good graphic demos in R. To Run the demo in R, type the following :

demo (graphics)

Then Press “Enter” to start. You will see several good examples of plots and graphics and their associated R-Scripts by Press “Enter” again.

Creating a Word Cloud for the definition of “Big Data” in this class.

The second task in this exercise is to create a “Word Cloud” (Tag Cloud) image from your class members’ posts about the definition of Big Data in the Blackboard.

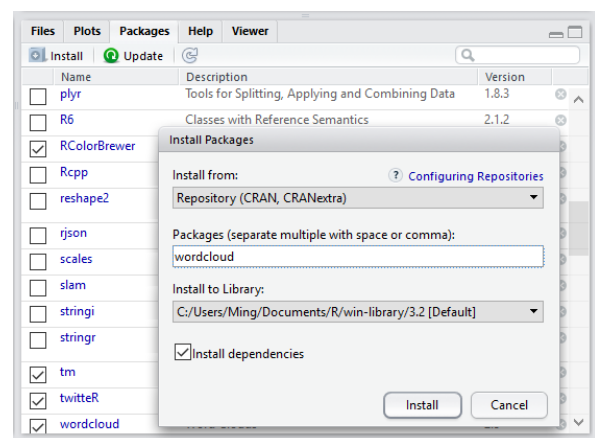
What is Word Cloud (Tag Cloud)? A tag cloud (word cloud, or weighted list in visual design) is a visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance

(usually measured by the number of occurrences) of each tag is shown with font size or color.[2] This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence. When used as website navigation aids, the terms are hyperlinked to items associated with the tag. (cited from Wikipedia: https://en.wikipedia.org/wiki/Tag_cloud).

The first step is to create a plain text ASCII file with everyone's definition about "Big Data".

1. Create a new folder "**wordcloud**" inside your R working directory (Z:\webexercise-2), then create a new TEXT file called "bigdata.txt" in the new folder (\webexercise-2\wordcloud).
2. Open this "bigdata.txt" in Notepad.
3. Open a web browser, login your Blackboard, go to the 596 Course site, then go to the "Discussions" → "What is Big Data?"
4. Go to each student's post and copy their definitions into the "bigdata.txt". In the Notepad, Press "Enter" after pasting one's definition and start a new line for another student's definition.
5. After copying all students' definition, click on [File] → [Save] to save these texts into the **bigdata.txt** (make sure the bigdata.txt is saved inside the "**\webexercise-2\wordcloud**" folder).

Since creating a Word Cloud will need several additional "packages" (or "libraries") in R. In the RStudio, go to the "**File/Plots**" window, click on "Packages". Select "**twitterR**", "**tm**", "**wordcloud**", and "**RColorBrewer**". If some packages are missing, click on install menu, then type the library name in the Packages, then click on install. After installing all required libraries, you are ready to run the R script. (Ignore some errors messages in the Console).



After installing all required library, you are ready to create your word cloud image. The following is an example of the R script for word cloud. You can modify some contents if you know how to do it.

```
#these are the libraries used in the Word Cloud Tasks
library(twitterR)
library(tm)
library(wordcloud)
library(RColorBrewer)

#Put your text files inside the temp folder under working directory

my_corpus = Corpus(DirSource("wordcloud"))

#You can add or remove STOPWORDS in the list

tdm = TermDocumentMatrix(my_corpus,
  control = list(removePunctuation = TRUE,
    stopwords = c("SDSU", "project", stopwords("english")),
    removeNumbers = TRUE, tolower = TRUE))

# define tdm as matrix
m = as.matrix(tdm)
# get word counts in decreasing order
word_freqs = sort(rowSums(m), decreasing=TRUE)
# create a data frame with words and their frequencies
dm = data.frame(word=names(word_freqs), freq=word_freqs)

# plot wordcloud in R
wordcloud(dm$word, dm$freq, random.order=FALSE, random.color=FALSE, rot.per=
0, colors=brewer.pal(8, "Dark2"))

# save the image in png format - a PNG image Ming_Cloud.png will be created
in the Working Directory

png("WordCloud.png", width=12, height=8, units="in", res=300)
wordcloud(dm$word, dm$freq, random.order=FALSE, random.color=FALSE, rot.per=
0, colors=brewer.pal(8, "Dark2"))

# dev.off will save the output PNG file into the working folder
dev.off()
```

Copy the previous R script, and then PASTE it into the R-Studio Console. Then hit “Enter” to run the script. You may get some warning messages, usually that’s fine and will not affect your word cloud results. When the program stops, you will see the word cloud on the right window. Also, a new PNG file, named “WordCloud.png” will be saved in the R working directory.

Additional Learning Resources:

- R Tutorial: <http://www.cyclismo.org/tutorial/R/>
- DataCamp: <https://www.datacamp.com/courses/free-introduction-to-r>
- R Intro YouTube video: <https://www.youtube.com/watch?v=7cGwYMhPDUY>

After finishing the Web Course, Please use your own words to answer the following questions (next page): **(DO NOT COPY any web resources or Wikipedia texts. We will check your answers with Blackboard tools to verify that your responses are uniquely yours.)** By submitting your answers (paper) to Blackboard, you agree: (1) that you are submitting your paper to be used and stored as part of the SafeAssign™ services in accordance with the [Blackboard Privacy Policy](#); (2) that your institution may use your paper in accordance with your institution's policies; and (3) that your use of SafeAssign will be without recourse against Blackboard Inc. and its affiliates.

SafeAssign accepts files in .doc, .docx, .docm, .ppt, .pptx, .odt, .txt, .rtf, .pdf, and .html file formats only. Files of any other format will not be checked through SafeAssign.

LAB-2 Additional Assignment:

1. Attach your Big Data Definition Word Cloud Image.
2. Go to the Live Well Data Portal (<https://data.livewellsd.org/>) Pick up another data types (other cancer data or other injuries data). Import the data into R and conduct some basic statistical analysis and draw some new visualization graphics (different from the previous example). Attached your R-script for creating the analysis and visual graphics in the report.
3. Select a Webpage or a group of text files, create a word cloud map. In the report, indicate the text sources and the output WordCloud image.
4. In the Word Cloud exercise, we need to manually copy each student's definition in Blackboard into a single text file. Can you think about any better data collection methods or procedure?

Please submit your LAB-2 Answers (in a MS Word or a PDF file format only) to the Blackboard System BEFORE the DUE DATE/TIME.