Web Exercise 02: R and R-Studio

DUE Date: September 15, 5:30pm (on Blackboard).
Grade: 20 points

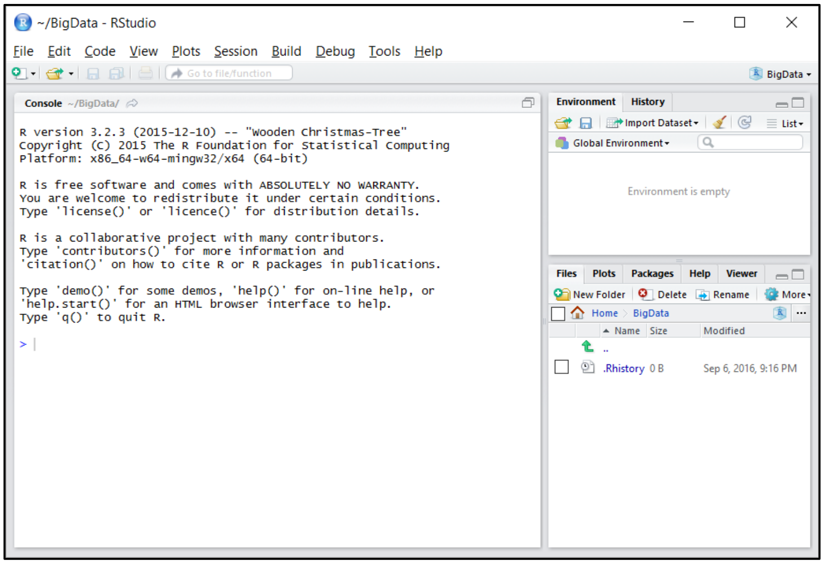
1. Use your own computer or login to the Computer Lab's account.
2. Install the R Software into your local machine. (If you are using the Lab machine, you can skip this step).

   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 3.6.1 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.



```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-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.

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.

> |
```

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 
```

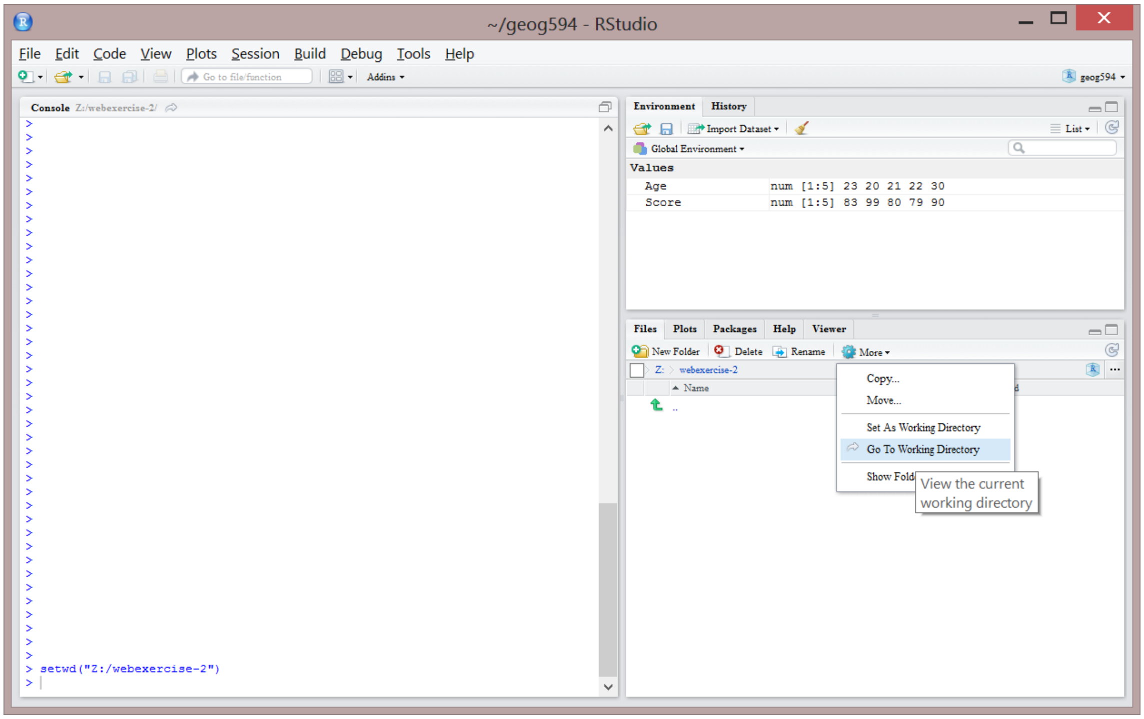
The second way to enter the data into R is to read data directly from a table using "**read.table**" command. The next task will teach you how to read a table into R.

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. → create a new folder in D: or C: called "**webexercise-2**".

In the R console commend mode: type the following:

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

Now your R working director will be in the **D:/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.



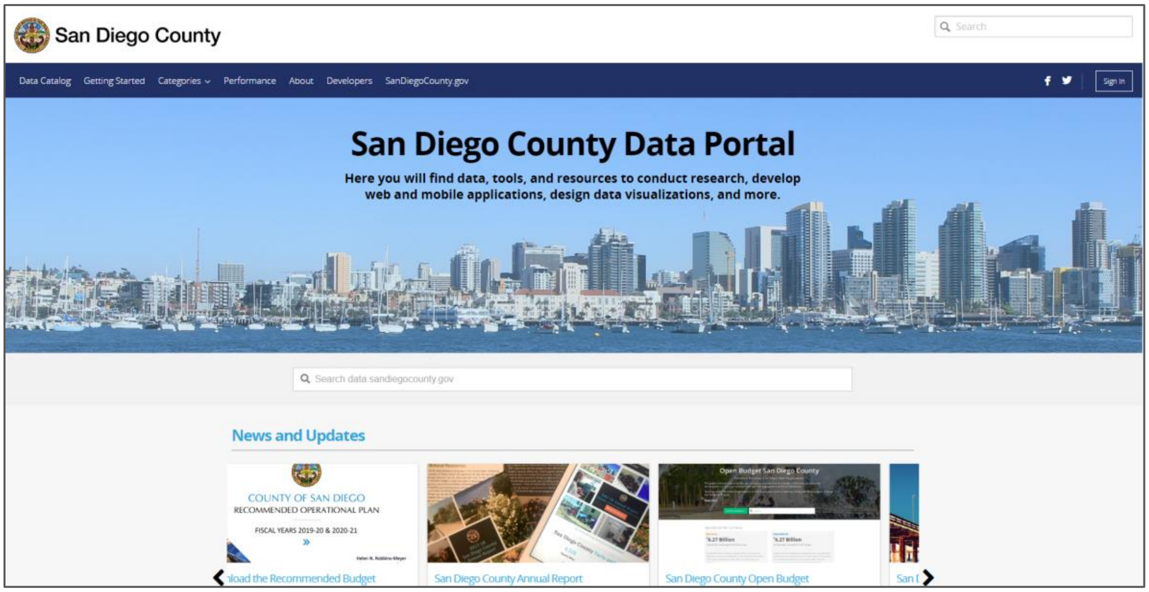
```
Values
  Age          num [1:5] 23 20 21 22 30
  Score        num [1:5] 83 99 80 79 90
```

```
> setwd("Z:/webexercise-2")
> 
```

Reading Tables and Conduct Cancer Data Analysis

The County of San Diego has created a very nice **Open Data portal**, providing easy access to public data and information about the county government. 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.sandiegocounty.gov/


San Diego County
Search
Data Catalog
Getting Started
Categories
Performance
About
Developers
SanDiegoCounty.gov
Sign In
San Diego County Data Portal
Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.
Search data.sandiegocounty.gov
News and Updates
COUNTY OF SAN DIEGO
RECOMMENDED OPERATIONAL PLAN
FISCAL YEARS 2019-20 & 2020-21
nload the Recommended Budget
San Diego County Annual Report
San Diego County Open Budget
San I

2. Select the "Data Catalog" in the Icons.
3. Select "non-communicable disease" in the tags and search "lung cancer" from the list or type "lung cancer" in the top search textbox.
4. Click on the "Lung Cancer" → You will see information about the dataset and Table Preview like the following:

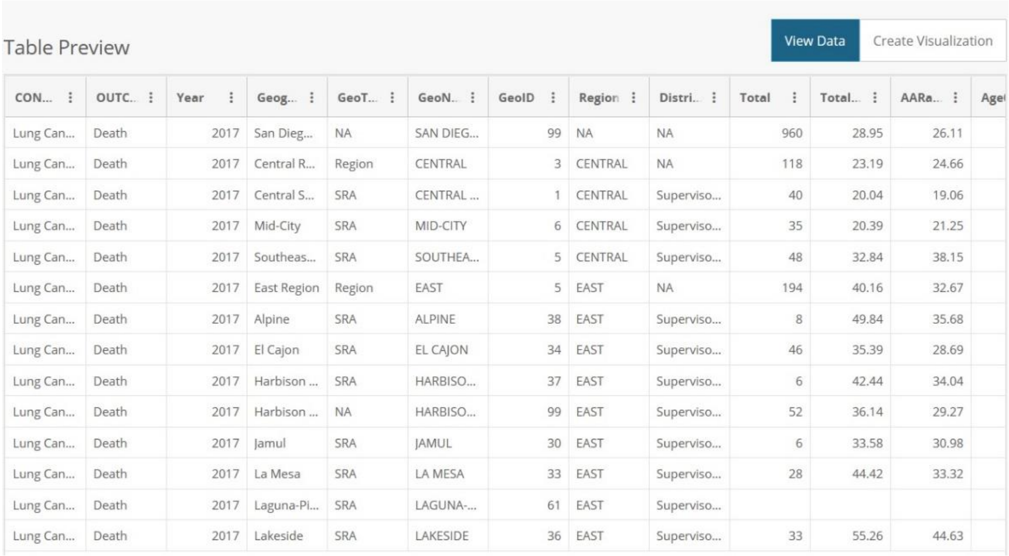


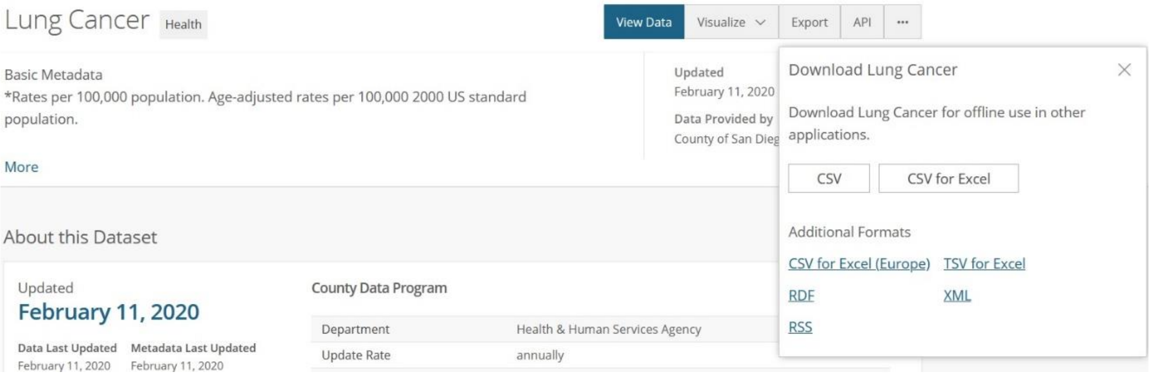
Table Preview

View Data | Create Visualization

CON...	OUTC...	Year	Geog...	GeoT...	GeoN...	GeoID	Region	Distri...	Total	Total...	AARa...	Age
Lung Can...	Death	2017	San Dieg...	NA	SAN DIEG...	99	NA	NA	960	28.95	26.11	
Lung Can...	Death	2017	Central R...	Region	CENTRAL	3	CENTRAL	NA	118	23.19	24.66	
Lung Can...	Death	2017	Central S...	SRA	CENTRAL ...	1	CENTRAL	Superviso...	40	20.04	19.06	
Lung Can...	Death	2017	Mid-City	SRA	MID-CITY	6	CENTRAL	Superviso...	35	20.39	21.25	
Lung Can...	Death	2017	Southeas...	SRA	SOUTHEA...	5	CENTRAL	Superviso...	48	32.84	38.15	
Lung Can...	Death	2017	East Region	Region	EAST	5	EAST	NA	194	40.16	32.67	
Lung Can...	Death	2017	Alpine	SRA	ALPINE	38	EAST	Superviso...	8	49.84	35.68	
Lung Can...	Death	2017	El Cajon	SRA	EL CAJON	34	EAST	Superviso...	46	35.39	28.69	
Lung Can...	Death	2017	Harbison ...	SRA	HARBISO...	37	EAST	Superviso...	6	42.44	34.04	
Lung Can...	Death	2017	Harbison ...	NA	HARBISO...	99	EAST	Superviso...	52	36.14	29.27	
Lung Can...	Death	2017	Jamul	SRA	JAMUL	30	EAST	Superviso...	6	33.58	30.98	
Lung Can...	Death	2017	La Mesa	SRA	LA MESA	33	EAST	Superviso...	28	44.42	33.32	
Lung Can...	Death	2017	Laguna-Pi...	SRA	LAGUNA-...	61	EAST	Superviso...				
Lung Can...	Death	2017	Lakeside	SRA	LAKESIDE	36	EAST	Superviso...	33	55.26	44.63	

This data include all numbers of lung cancer outcome incidents in each Region, Sub Regional Areas (SRA), Supervisor District, and Municipal in San Diego County. Also, when you click ‘Visualize’ → ‘Create Visualization’, you can create charts.

Click on [Export] icon in the Data window → CSV for Excel. Save the CSV file into the R working directory folder (for example: **D:\webexercise-2**). Please save to your own local R working directory folder.



Lung Cancer Health

View Data | Visualize | Export | API | ···

Basic Metadata
*Rates per 100,000 population. Age-adjusted rates per 100,000 2000 US standard population.

More

Updated
February 11, 2020

Data Provided by
County of San Dieg

About this Dataset

Updated
February 11, 2020

Data Last Updated
February 11, 2020

Metadata Last Updated
February 11, 2020

County Data Program

Department	Health & Human Services Agency
Update Rate	annually

Download Lung Cancer

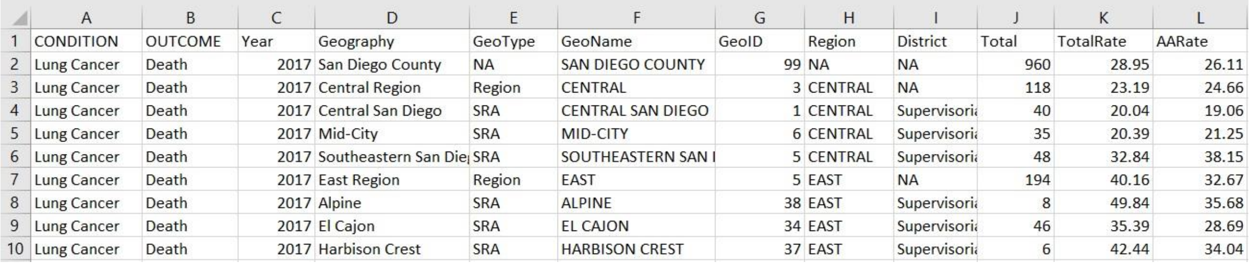
Download Lung Cancer for offline use in other applications.

CSV | CSV for Excel

Additional Formats

CSV for Excel (Europe) | TSV for Excel
RDF | XML
RSS

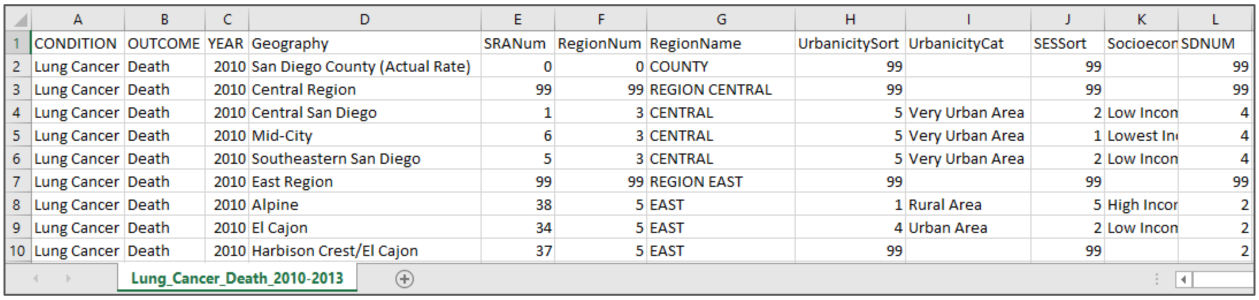
You can open the file and take a look at the datasets.



	A	B	C	D	E	F	G	H	I	J	K	L
1	CONDITION	OUTCOME	Year	Geography	GeoType	GeoName	GeoID	Region	District	Total	TotalRate	AARate
2	Lung Cancer	Death	2017	San Diego County	NA	SAN DIEGO COUNTY	99	NA	NA	960	28.95	26.11
3	Lung Cancer	Death	2017	Central Region	Region	CENTRAL	3	CENTRAL	NA	118	23.19	24.66
4	Lung Cancer	Death	2017	Central San Diego	SRA	CENTRAL SAN DIEGO	1	CENTRAL	Supervisori	40	20.04	19.06
5	Lung Cancer	Death	2017	Mid-City	SRA	MID-CITY	6	CENTRAL	Supervisori	35	20.39	21.25
6	Lung Cancer	Death	2017	Southeastern San Die	SRA	SOUTHEASTERN SAN I	5	CENTRAL	Supervisori	48	32.84	38.15
7	Lung Cancer	Death	2017	East Region	Region	EAST	5	EAST	NA	194	40.16	32.67
8	Lung Cancer	Death	2017	Alpine	SRA	ALPINE	38	EAST	Supervisori	8	49.84	35.68
9	Lung Cancer	Death	2017	El Cajon	SRA	EL CAJON	34	EAST	Supervisori	46	35.39	28.69
10	Lung Cancer	Death	2017	Harbison Crest	SRA	HARBISON CREST	37	EAST	Supervisori	6	42.44	34.04

San Diego County Data Portal incorporated Live Well data portal (back to 2017) and since then, some cancer data format has changed. For example, column name ‘Age_Adjusted_Rate’ is changed to ‘AARate’ and ‘Socioeconomic_Category’ is changed to ‘SES’, and other columns were added. For data analysis using R, we will use ‘Lung_Cancer_Death_2010-2013.csv’ file which was originally downloaded from Live Well data portal. (You can find it from **Google Drive folder inside “Web-Exercises/Data”**)

5. Download ‘Lung_Cancer_Death_2010-2013.csv’ file from Google Drive folder and copy it into your R working directory.



	A	B	C	D	E	F	G	H	I	J	K	L
1	CONDITION	OUTCOME	YEAR	Geography	SRANum	RegionNum	RegionName	UrbanicitySort	UrbanicityCat	SESSort	Socioecor	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 CENTRAL	99		99		99
4	Lung Cancer	Death	2010	Central San Diego	1	3	CENTRAL	5	Very Urban Area	2	Low Incon	4
5	Lung Cancer	Death	2010	Mid-City	6	3	CENTRAL	5	Very Urban Area	1	Lowest In	4
6	Lung Cancer	Death	2010	Southeastern San Diego	5	3	CENTRAL	5	Very Urban Area	2	Low Incon	4
7	Lung Cancer	Death	2010	East Region	99	99	REGION EAST	99		99		99
8	Lung Cancer	Death	2010	Alpine	38	5	EAST	1	Rural Area	5	High Incor	2
9	Lung Cancer	Death	2010	El Cajon	34	5	EAST	4	Urban Area	2	Low Incon	2
10	Lung Cancer	Death	2010	Harbison Crest/El Cajon	37	5	EAST	99		99		2

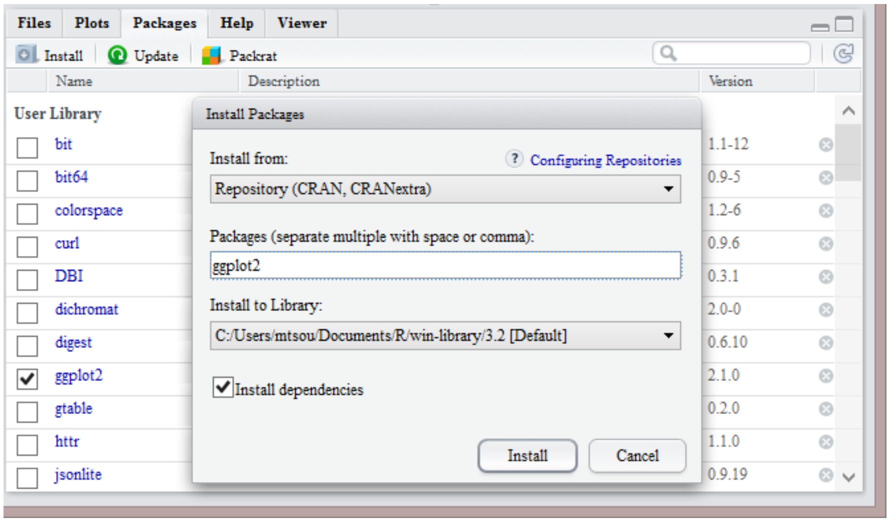
Lung_Cancer_Death_2010-2013

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.



Name	Description	Version
User Library		
□ bit		1.1-12
□ bit64		0.9-5
□ colorspace		1.2-6
□ curl		0.9.6
□ DBI		0.3.1
□ dichromat		2.0-0
□ digest		0.6.10
☑ ggplot2		2.1.0
□ gtable		0.2.0
□ httr		1.1.0
□ jsonlite		0.9.19

Install from: Repository (CRAN, CRANextra)

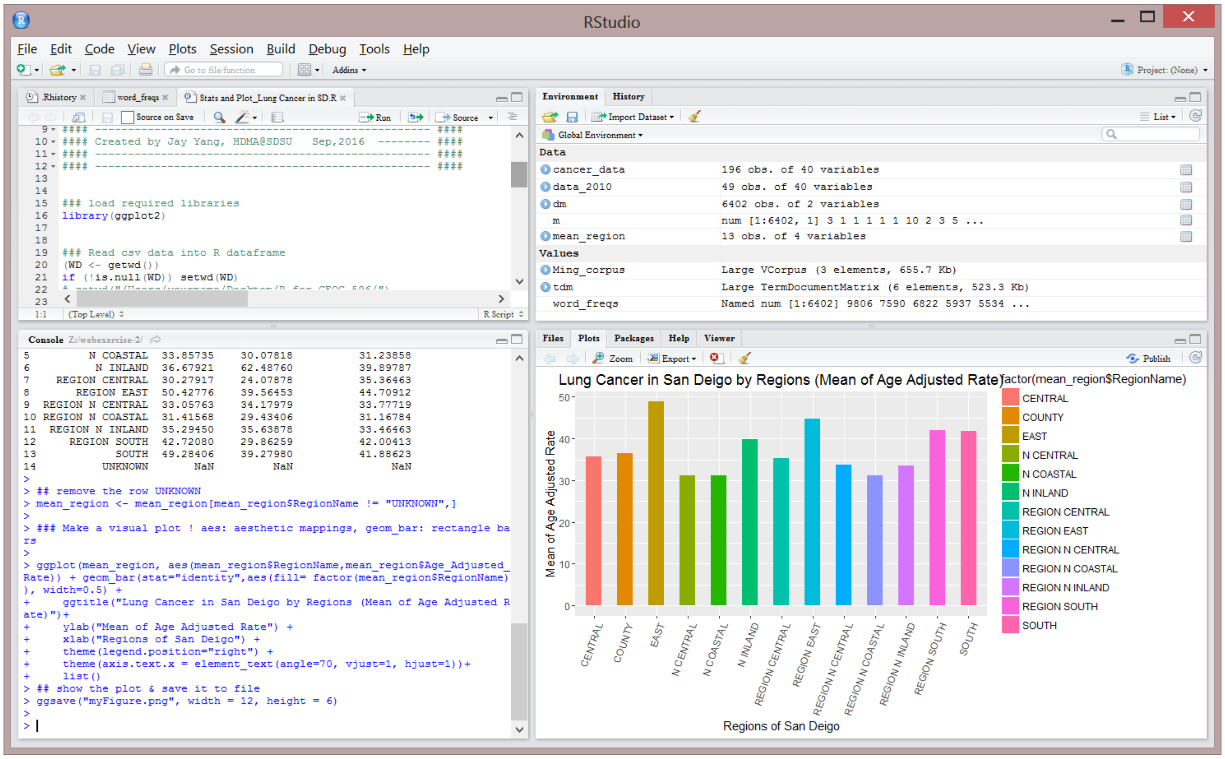
Packages (separate multiple with space or comma): ggplot2

Install to Library: C:/Users/mtsou/Documents/R/win-library/3.2 [Default]

☑ Install dependencies

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)
```



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


### load required libraries
library(ggplot2)


### Read csv data into R dataframe
(WD <- getwd())
if (!is.null(WD)) setwd(WD)
# setwd("/Users/yourname/Desktop/R for GEOG 596/")
```

```
Data
cancer_data      196 obs. of 40 variables
data_2010        49 obs. of 40 variables
dm               6402 obs. of 2 variables
m                num [1:6402, 1] 3 1 1 1 1 1 10 2 3 5 ...
mean_region      13 obs. of 4 variables
Values
Ming_corpus      Large VCorpus (3 elements, 655.7 Kb)
tdm              Large TermDocumentMatrix (6 elements, 523.3 Kb)
word_freqs       Named num [1:6402] 9806 7590 6822 5937 5534 ...
```

```
5          N COASTAL  33.85735    30.07818          31.23858
6           N INLAND  36.67921    62.48760          39.89787
7     REGION CENTRAL  30.27917    24.07878          35.36463
8        REGION EAST  50.42776    39.56453          44.70912
9   REGION N CENTRAL  33.05763    34.17979          33.77719
10  REGION N COASTAL  31.41568    29.43406          31.16784
11   REGION N INLAND  35.29450    35.63878          33.46463
12      REGION SOUTH  42.72080    29.86259          42.00413
13             SOUTH  49.28406    39.27980          41.88623
14           UNKNOWN       NaN         NaN               NaN
>
> ## remove the row UNKNOWN
> mean_region <- mean_region[mean_region$RegionName != "UNKNOWN",]
>
> ### Make a visual plot ! aes: aesthetic mappings, geom_bar: rectangle ba
rs
>
> ggplot(mean_region, aes(mean_region$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 R
ate)")+
+     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 commend 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.

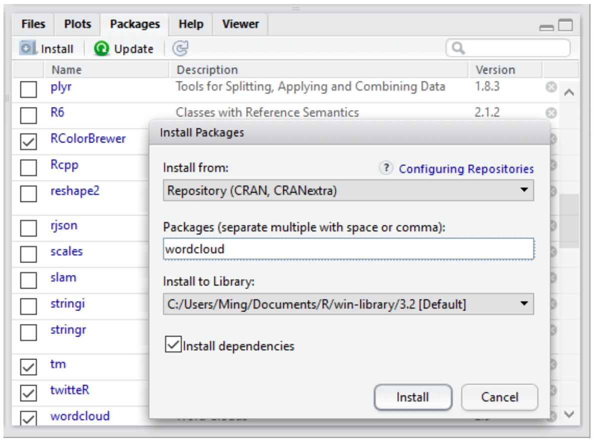
The second task in this exercise is to create a “Word Cloud” (Tag Cloud) image from your class members’ posts about the definition of Big Data in the Blackboard.

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" (class members in Fall 2020.

1. Create a new folder "**wordcloud**" inside your R working directory (D:\webexercise-2), (\webexercise-2\wordcloud).
2. Open the "**BDA594-2020-Definition-text.csv**" file inside the Class Shared Google Driver folder **\BDA-GEOG-594-BigDataScience\Web-Exercises\Data (link: https://drive.google.com/open?id=1iUUsVbTywFKMStI7FaeMpPGvH_BGshFd)**
3. Remove the name field in the CSV file and save the definition texts as "**bigdata.txt**" into your local **"\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).



	Name	Description	Version
□	plyr	Tools for Splitting, Applying and Combining Data	1.8.3
□	R6	Classes with Reference Semantics	2.1.2
☑	RColorBrewer		
□	Rcpp		
□	reshape2		
□	rjson		
□	scales		
□	slam		
□	stringi		
□	stringr		
☑	tm		
☑	twitteR		
☑	wordcloud		

Install Packages

Install from: Repository (CRAN, CRANextra)

Configuring Repositories

Packages (separate multiple with space or comma): wordcloud

Install to Library: C:/Users/Ming/Documents/R/win-library/3.2 [Default]

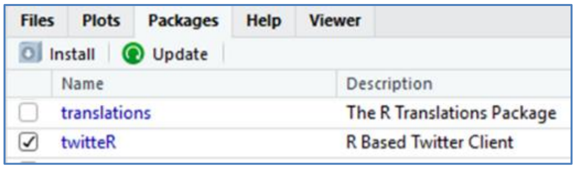
☑ Install dependencies

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(twitteR)
library(tm)
library(wordcloud)
library(RColorBrewer)
library(NLP)

#Put your text files inside the temp folder(wordcloud) under working
directory(D:\webexercise-2)

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 above 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.

If you get a warning message such as "Warning message: package 'twitteR' was built under R version 3.5.3", you can load libraries by checking the package from package list.



	Name	Description
□	translations	The R Translations Package
☑	twitteR	R Based Twitter Client

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. Rename the WordCloud.png to "**BigData-Definition-V1.png**".
One problem in this WordCloud image is that "Big" and "Data" are too prominent in the image. One possible way to correct this is to add the two words into the STOPWORD list. You can modify the R scrip by changing the stopwords command into the following:

```
stopwords = c("Big", "Data", stopwords("english")),
```

Then you can run the R script again. A new WordCloud.png will be created. Rename this image to "**BigData-definition-V2.png**"

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
- **R Shiny is an R package to build interactive web apps from R libraries and JavaScripts. It is a very powerful platform for web-based data analytic tools.** https://shiny.rstudio.com/

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 the two Big Data Definition Word Cloud Images (texts from the students in the current semester) .
2. Go to the San Diego County Data Portal **(**https://data.sandiegocounty.gov/) 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 (similar or different from the previous example). You will need to revise R-script that is used for previous Cancer analysis. **Attach your new R-script and the new data file** for creating the analysis and **the visual graphics** in the report.
3. Select a Webpage or a group of text files, create a word cloud map with some selected "Stopwords". In the report, indicate the text sources, the selected "stopwords", and the output of WordCloud images.
4. Introduce **R-Shiny** and its major functions (at least 200 words) using your own languages (don't copy/paste the definition from the Internet). Find one on-line example of R-Shiny applications and provide the URL (Web address) and the application screen shot.
5. There are many **data science** oriented R packages (libraries) available (for data manipulation, data visualization, or machine learning). Please identify **THREE packages** (libraries) in R related to data science and introduce each of them briefly (at least 100 words for each packages).

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.