Welcome!

- Download the slide for today's lab from
  Canvas → CP6025 → Modules → **Week 2 – Research Design** → **Lab2.zip**

- Unzip Lab2.zip → Week2_Lab_DataCleaning.pdf

- When you are done with today's lab activity, save your script and keep it for future references

Credit: Bonwoo Koo

Quick Recap from Lab 1

Data type and data structure in R

- Data type is the type of a single entry (e.g., 1 is an integer)
- Data structure is how data is organized (c(1,2) is a vector)
- Helps you read the help document and figure out what data type to put in.
- Implies what operations you can do (e.g., you can only add up numbers, such as 1+1, but not strings, such as “1” + “1” or “1” + 1).
- Different data type and structures are useful for different types of data storage / operation

Credit: Xiaofan Liang

Quick Recap from Lab 1

Data Type

- Number (as.numeric() or as.integer()), e.g., 1.1, 2, 3, 4
- Character (as.character()), e.g., “one”, “two”, “1”, “2”, “3”
- Factor, e.g., “one”, “two” but “one” < “two”, ordinal data
- Logical, e.g., TRUE, FALSE

Data Structure

- Vector, e.g., c(1,2,3,4) or c(“one”, “two”, 1, 2)
- Matrix, e.g., matrix(1:4, ncol=2, nrow=2)
- List, e.g., list(a=1,b=2) or list(1,2)
- Data Frame (or table) (will review in this lab)

```
     [,1] [,2]
[1,]    1    3
[2,]    2    4
```

```
$a      [[1]]
[1] 1   [1] 1
$b      [[2]]
[1] 2   [1] 2
```


Credit: Xiaofan Liang

Adv. Planning Methods

Lab 2: Data Cleaning & Descriptive Statistics

09.02.2021

Welcome!

- **1 Lab = 1 document (R markdown or html) + 1 slide**
- Document contains the complete material prepared for the lab, mostly by Bonwoo Koo.
- Slide explains parts of the document that I think are important or can be confusing.
- I will go over the slide first. Then, you will have time to read and replicate the document. Once you are done with the document, you can leave!
- Google Doc Queue for Q & A during hands-on
- **3:00-3:30pm is chill time – I will sit outside, come chat**

Credit: Bonwoo Koo

Goals for today

- Download census data
- Working directory
- Reading and examining data in R
- Calculating the descriptive statistics
- Data cleaning with base R
- Data cleaning with tidyverse
- Commenting on script

Download Census data

- We want to download a table called "Ratio of Income to Poverty Level in the Past 12 Months (TableID: C17002)"
- from ACS (American Community Survey, updated every year),
- for all census tracts (a geographic units)
- in the Georgia state.
- This data is already downloaded for you as ACSDT5Y2017.C17002-Data.csv in Lab2 folder. The following process shows where the data come from.

Download Census data

United States Census Bureau

Explore Census Data

The Census Bureau is the leading source of quality data about the nation's people and economy.

nd Tables, Maps, and more ... SEARCH

Advanced Search Help Feedback

Go to https://data.census.gov/cedsci/ and select "Advanced Search"

Tables

Check out our new table display which allows you to dynamically add geographies, topics, or any applicable filters. You can reorder, pin, and hide columns all with simple drag and drop functionality. Tab through different tables to make sure you found the right one, customize it, and then download multiple vintages of it quickly. If you don't see a functionality you need, find a bug, or have a comment, drop us a line at cedsci.feedback@census.gov.

VIEW TABLES

Send Feedback
cedsci.feedback@census.gov

Credit: Bonwoo Koo

Download Census data

United States Census Bureau

Search

// Search / Advanced Search

Advanced Search

Table ID (e.g., DP05)

Narrow search with filters

FIND A FILTER

e.g. 336111 - Automobile Manufacturing

BROWSE FILTERS	TOPICS	INCOME AND POVERTY
Topics	Business and Economy	☑ Income and Poverty
Geography	Education	Income and Earnings
Years	Employment	Poverty
Surveys	Families and Living Arrangements	☐ Wealth and Assets
Codes	Government	
	Health	
	Housing	
	Income and Poverty	
	Populations and People	
	Race and Ethnicity	

Selecting Topics

Sequentially select from the menu
in the order of:

1. Topics
2. Income and Poverty
3. Income and Poverty (check box).

The selection will add this icon next to **Selected Filters:** to indicate that the filter is successfully placed.

Selected Filters: Income and Poverty Topic ✕

CLEAR SEARCH SHOW

Credit: Bonwoo Koo

Download Census data

United States Census Bureau

Search

// Search / Advanced Search

Advanced Search

Table ID (e.g., DP05)

Narrow search with filters

FIND A FILTER

e.g. 336111 - Automobile Manufacturing

Selecting Geography

Then, select the following items in the order of:

1. Geography
2. Tract
3. Georgia
4. All Census Tracts within Georgia

BROWSE FILTERS

- Topics
- Geography
- Years
- Surveys
- Codes

GEOGRAPHY

Show Summary Levels

- Nation
- Region
- Division
- State
- County
- Tract
- Block Group
- Block
- Zip Code Tabulation Area (Five-Digit)
- Elementary School District
- Secondary School District
- Unified School District

WITHIN (STATE)

- Alaska
- American Samoa
- Arizona
- Arkansas
- California
- Colorado
- Commonwealth of the Northern Mariana Islands
- Connecticut
- Delaware
- District of Columbia
- Florida
- Georgia
- Guam

WITHIN (COUNTY)

- Within Other Geographies
- All Census Tracts within Georgia
- Appling County, Georgia
- Atkinson County, Georgia
- Bacon County, Georgia
- Baker County, Georgia
- Baldwin County, Georgia
- Banks County, Georgia
- Barrow County, Georgia
- Bartow County, Georgia
- Ben Hill County, Georgia
- Berrien County, Georgia
- Bibb County, Georgia
- Bleckley County, Georgia

Selected Filters: Income and Poverty (Topic) ✕ | All Census Tracts within Georgia (Geography) ✕

Another filter is added

CLEAR | SEARCH | SHOW

Credit: Bonwoo Koo

Download Census data

United States Census Bureau

Search

// Search / Advanced Search

Advanced Search

Table ID (e.g., DP05)

Narrow search with filters

FIND A FILTER

e.g. 336111 - Automobile Manufacturing

BROWSE FILTERS

- Topics
- Geography
- Years
- Surveys
- Codes

YEARS

- [ ] 2022
- [ ] 2019
- [ ] 2018
- [x] 2017
- [ ] 2016
- [ ] 2015
- [ ] 2014
- [ ] 2013
- [ ] 2012
- [ ] 2011
- [ ] 2010
- [ ] 2000

Selecting Years

Then, specify what year you want by selecting

1. Years
2. 2017

Yet another filter is added

Selected Filters: Income and Poverty (Topic) ✕ | All Census Tracts within Georgia (Geography) ✕ | 2017 (Year) ✕

CLEAR | SEARCH | SHOW

Credit: Bonwoo Koo

Download Census data

United States Census Bureau

Search

// Search / Advanced Search

Advanced Search

ratio

Narrow search with filters

FIND A FILTER

e.g. 336111 - Automobile Manufacturing

BROWSE FILTERS

- Topics
- Geography
- Years
- Surveys
- Codes

YEARS

- [ ] 2022
- [ ] 2019
- [ ] 2018
- [x] 2017
- [ ] 2016
- [ ] 2015
- [ ] 2014
- [ ] 2013
- [ ] 2012
- [ ] 2011
- [ ] 2010
- [ ] 2000

Adding a keyword

Finally, narrow down the search by adding a keyword ***ratio*.**

You can do that by typing the word ratio into this search bar.

With the keyword **ratio** typed in the search bar,

Click **SEARCH** button.

Send Feedback
cedsci.feedback@census.gov

Income and Poverty
Topic

All Census Tracts within Georgia
Geography

2017
Year

CLEAR

SEARCH

SHOW

Credit: Bonwoo Koo

Download Census data

United States Census Bureau

Search

ALL TABLES MAPS PAGES

About 22,340 results | **Filter**

Tables

RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017 **Table:** C17002

AGE BY RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017 **Table:** B17024

RATIO OF INCOME TO POVERTY LEVEL OF FAMILIES IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017 **Table:** B17026

MORTGAGE STATUS BY RATIO OF VALUE TO HOUSEHOLD INCOME IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017 **Table:** B25100

RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS BY DISABILITY STATUS
Survey/Program: American Community Survey
Years: 2017 **Table:** C18131

HEALTH INSURANCE COVERAGE STATUS BY RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS BY AGE
Survey/Program: American Community Survey
Years: 2017 **Table:** C27016

PRIVATE HEALTH INSURANCE BY RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS BY AGE
Survey/Program: American Community Survey
Years: 2017 **Table:** C27017

PUBLIC HEALTH INSURANCE BY RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS BY AGE

Send Feedback
cedsci.feedback@census.gov

Selecting a table

This is the data table we are looking for. Click it.

Credit: Bonwoo Koo

Download Census data

United States Census Bureau

Search

ALL | TABLES | MAPS | PAGES

14 Results Filter | Download

RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017
Table: C17002

AGE BY RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017
Table: B17024

RATIO OF INCOME TO POVERTY LEVEL OF FAMILIES IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017
Table: B17026

MORTGAGE STATUS BY RATIO OF VALUE TO HOUSEHOLD INCOME IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017
Table: B25100

RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS BY DISABILITY STATUS
Survey/Program: American Community Survey
Years: 2017
Table: C18131

Send Feedback
cedsci.feedback@census.gov

COVERAGE STATUS

RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
TableID: C17002
Product: 2017: ACS 5-Year Estimates Detailed Tables
Universe: Population for whom poverty status is determined

CUSTOMIZE TABLE

Sorry, that table is too large to display.

DOWNLOAD TABLE

Verifying & downloading

Double-check that the following information are correctly listed in the top panel and click **DOWNLOAD TABLE**:

- Name of the table
- Table ID
- Survey/Program
- Product

- Pay extra attention and make sure it is ACS 5-Year Estimates

Credit: Bonwoo Koo

Download Census data

United States Census Bureau

Search

ALL | TABLES | MAPS | PAGES

14 Results — Filter | Download

RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017
Table: C17002

AGE BY RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017
Table: B17024

RATIO OF INCOME TO POVERTY LEVEL OF FAMILIES IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017
Table: B17026

MORTGAGE STATUS BY RATIO OF VALUE TO HOUSEHOLD INCOME IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
Years: 2017
Table: B25100

RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS BY DISABILITY STATUS
Survey/Program: American Community Survey
Years: 2017
Table: C18131

Send Feedback
cedsci.feedback@census.gov

COVERAGE STATUS

RATIO OF INCOME TO POVERTY LEVEL IN THE PAST 12 MONTHS
Survey/Program: American Community Survey
TableID: C17002

TABLE

Sorry, ti

DOWNLO

Verifying & downloading

Make sure that 2017 is selected and that File Type is CSV.

Finally, Click DOWNLOAD.

Download / Print / Share

HIDE

DOWNLOAD | EMBED | SHARE | API | PRINT | MORE DATA

Select Table Vintages

	All	2017
C17002 5-Year	☐	☑

File Type
- (•) CSV
- ( ) PDF

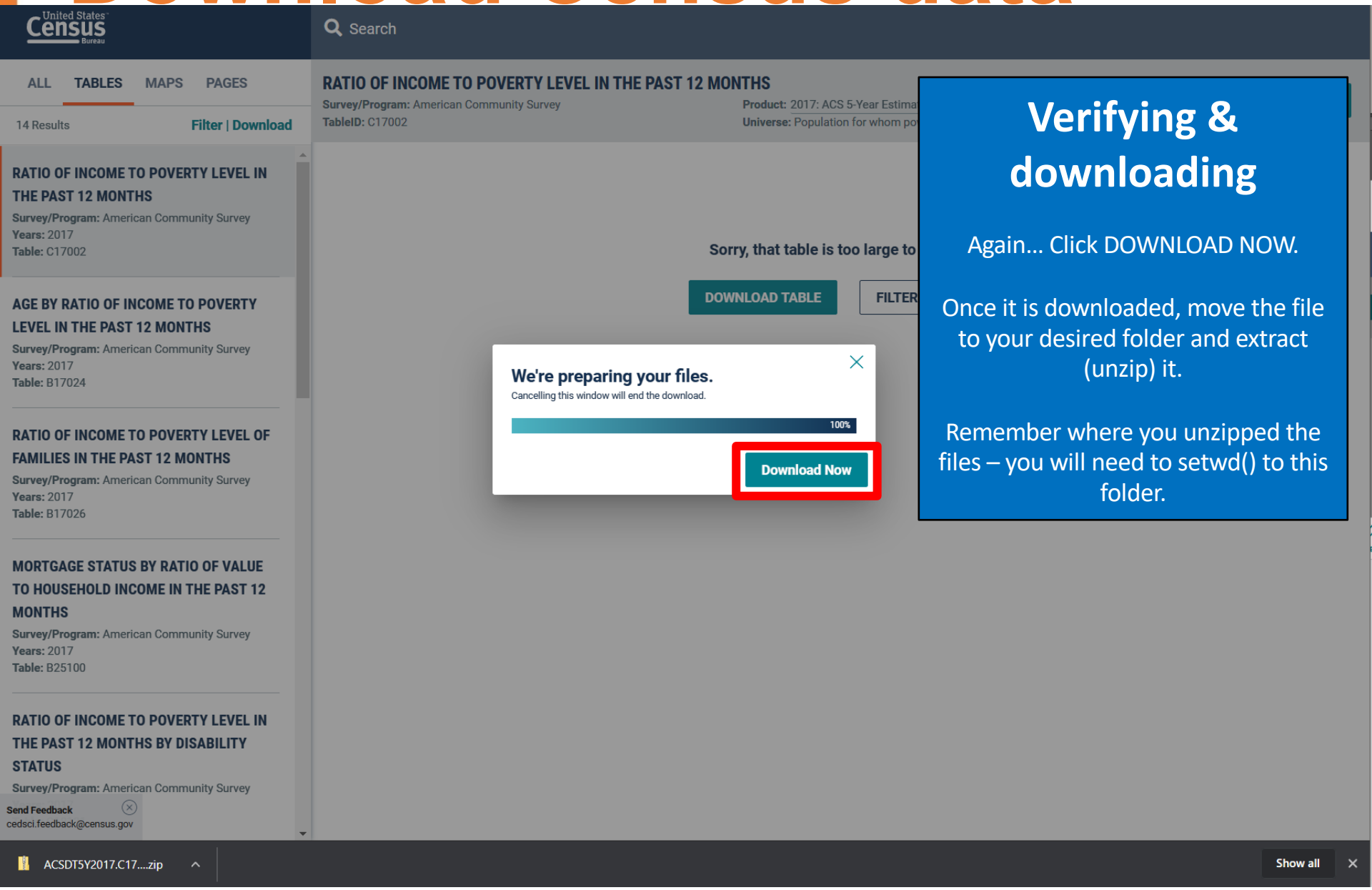
What You're Getting
- 1 .csv files (metadata)
- 1 .csv files (data)
- 1 .txt files (table title)

Uncompressed Estimated Size: 355.2 kB

DOWNLOAD

Credit: Bonwoo Koo

Download Census data



Verifying & downloading

Again... Click DOWNLOAD NOW.

Once it is downloaded, move the file to your desired folder and extract (unzip) it.

Remember where you unzipped the files – you will need to setwd() to this folder.

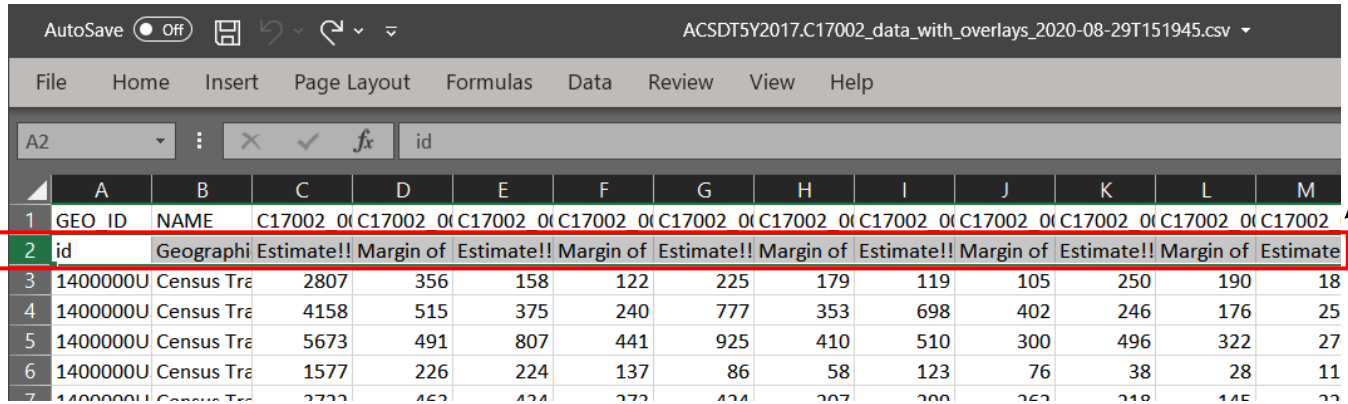
Credit: Bonwoo Koo

Download Census data

Deleting the second header

Open the file named *ACSDT5Y2017.C17002_data_with_overlays_2020-08-29T151945.csv*. The exact name may be slightly different due to the date component in the file name. There will be two different headers (i.e., variable names).

Delete the entire second row so that there will be only one row of header remaining. Then, save and close Excel.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	GEO_ID	NAME	C17002_0(	C17002_0(	C17002_0(	C17002_0(	C17002_0(	C17002_0(	C17002_0(	C17002_0(	C17002_0(	C17002_0(	C17002
2	id	Geographi	Estimate!!	Margin of	Estimate!!	Margin of	Estimate!!	Margin of	Estimate!!	Margin of	Estimate!!	Margin of	Estimate
3	1400000U	Census Tra	2807	356	158	122	225	179	119	105	250	190	18
4	1400000U	Census Tra	4158	515	375	240	777	353	698	402	246	176	25
5	1400000U	Census Tra	5673	491	807	441	925	410	510	300	496	322	27
6	1400000U	Census Tra	1577	226	224	137	86	58	123	76	38	28	11

First header

Second header

→ ***Delete this entire row***

Why are we deleting one of the headers?

The Census Bureau (and sometime other entities too) often provides their data with two rows of header.

The first type of the header is more code-like, not human-readable (e.g., `C17002_001E`), and **the second type** is often more intuitive to human (e.g., `Margin of Error!!Total!!Under .50`).

R users **usually prefer the first type**. It is because (1) the second type of header often contain spaces and special characters which can cause errors and other difficulties, (2) the second type can be very lengthy (and you will have to type them), and (3) when you join the data with ArcGIS, the long variable names are automatically coerced into shorter but somewhat confusing names.

Credit: Bonwoo Koo

Download via Census API in R

- API (Application Programming Interface) is like a chat bot for the database: you tell the API what you want in a specific way, and the API will automatically return you the data.
- No more click and find! But usually requires an API Key.
- Package tidycensus helps you with Census API in R, e.g.,

We want a 2017 table called "Ratio of Income to Poverty Level in the Past 12 Months (TableID: C17002)" from ACS (American Community Survey, updated every year), for all census tracts (a geographic units) in the Georgia state.

```
Table <- get_acs(year = 2017, table = "C17002", geography = "tract", state = "GA")
```

Tidycensus documentation: https://walker-data.com/tidycensus/articles/basic-usage.html

Credit: Xiaofan Liang

Working Directory

- We will read in a data file into R for the lab activities to examine Ratio of Income to Poverty Level in the Past 12 Months for all census tracts in Georgia state.

- This file is located in:

Canvas → CP6025 → Modules → Lab2 → ACSDT5Y2017.C17002-Data.csv

- Keep the Lab2 folder open - we will need it soon.

Credit: Bonwoo Koo

Working Directory

- When you open R-studio, your R session is **in** one of the folders in your computer (usually the Documents folder).

- The following function returns the folder your R session is currently in.

```
getwd()
```

- Is R in the folder where you saved the **ACSDT5Y2017.C17002-Data.csv** data file?

- Type read.csv('ACSDT5Y2017.C17002-Data.csv") in your R script window, what does the error say?

Credit: Bonwoo Koo

Working Directory

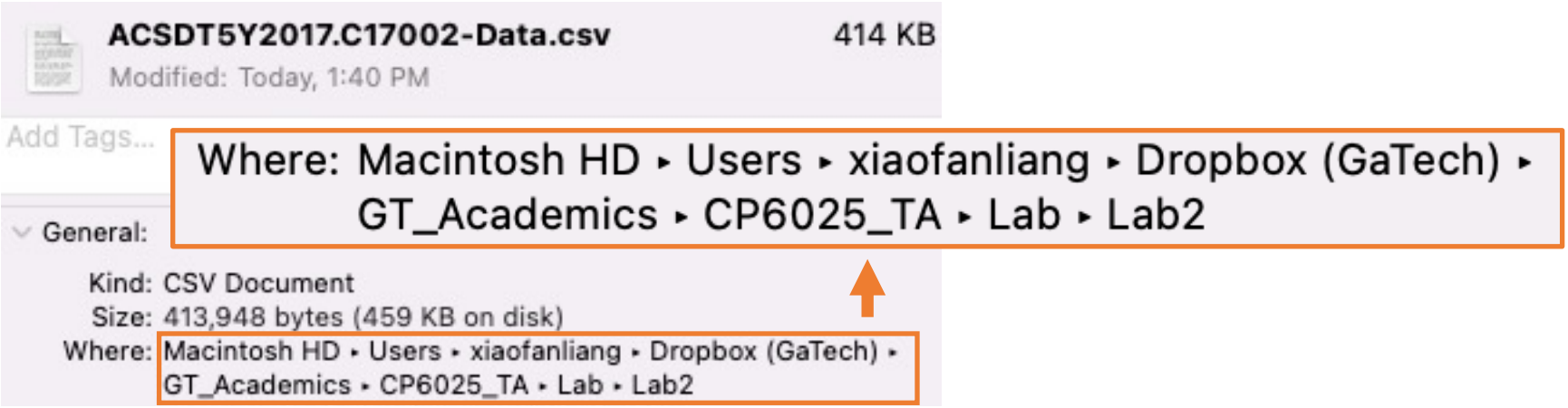
- R cannot find **ACSDT5Y2017.C17002-Data.csv** file because R is in a different folder.

- You need to tell R to navigate to the folder you save the CSV file using setwd()

- Setting directory/path is like pointing directions for R: a path (e.g., C:/Users/xiaofanliang/Dropbox (GaTech)/) tells R to first go to *C drive*, go to *Users* folder, then *xiaofanliang* folder, then *Dropbox (GaTech)* folder.

Working Directory

- How do I find the path to a file?
  - Mac: Right click on the file -> get info (No need to include Macintosh HD in the path).
  - Window: Right click on the file -> Properties (or Copy as Path) or double click on folder directory. Change \ to / in the path.



/Users/xiaofanliang/Dropbox (GaTech)/GT_Academics/CP6025_TA/Lab/Lab2

Credit: Xiaofan Liang

Working Directory

- The folder that your R session is currently in is called

Working Directory

getwd()

"show me the current working directory"

setwd()

"Change the working directory to a different folder"

Credit: Bonwoo Koo

Working Environment

- R searches for files and saves outputs in the working directory.
- Let's change your working directory to a folder where you saved the data for today's lab.
- Write the following in your script window and run **(hit control + enter)**:

```
setwd("path-to-your-folder")
```

e.g., setwd("C:/Users/bkoo34/Dropbox (GaTech)/CP6025")

- To verify that you are in the folder where our data is located:

```
dir()
```

- Do you see the CSV you saved (i.e., ACSDT5Y2017.C17002-Data.csv)?

Credit: Bonwoo Koo

Reading data in R

- We are now in the right folder. R can now read the data.
- Before using R, let's first take a look at the data **in Excel**. Double click "ACSDT5Y2017.C17002-Data.csv" to open it in Excel.

A	B	C	D	E	F
GEO_ID	NAME	C17002_001E	C17002_001EA	C17002_001M	C17002_001MA
1400000US13001950100	Census Tract 9501, Appling County, Georgia	2807	null	356	null
1400000US13001950200	Census Tract 9502, Appling County, Georgia	4158	null	515	null
1400000US13001950300	Census Tract 9503, Appling County, Georgia	5673	null	491	null
1400000US13001950400	Census Tract 9504, Appling County, Georgia	1577	null	226	null
1400000US13001950500	Census Tract 9505, Appling County, Georgia	3722	null	463	null
1400000US13003960100	Census Tract 9601, Atkinson County, Georgia	2106	null	255	null
1400000US13003960200	Census Tract 9602, Atkinson County, Georgia	4698	null	281	null
1400000US13003960300	Census Tract 9603, Atkinson County, Georgia	1460	null	212	null

- Each row is one Census tract (similar to a neighborhood)
- Each column is one variable. C17002 is the table. 001 is variable #1 (total population) in the table. E represents estimate, EA is estimate annotation, M is margin of error, and MA is margin of error annotation.
- For example, this cell shows that "the first neighborhood has a total population estimate of 2807"

Credit: Xiaofan Liang

Reading data in R

- In R script window, write the following and run it (control + enter)

```
pov.data <- read.csv("ACSDT5Y2017.C17002-Data.csv")
```

Notice that this is a relative path, relative to the path that you set in the setwd(). Alternatively, without using the setwd(), you can do read.csv("Users/xiaofanliang/Dropbox (GaTech)/GT_Academics/Lab/Lab2/ACSDT5Y2017.C17002-Data.csv")

and then,

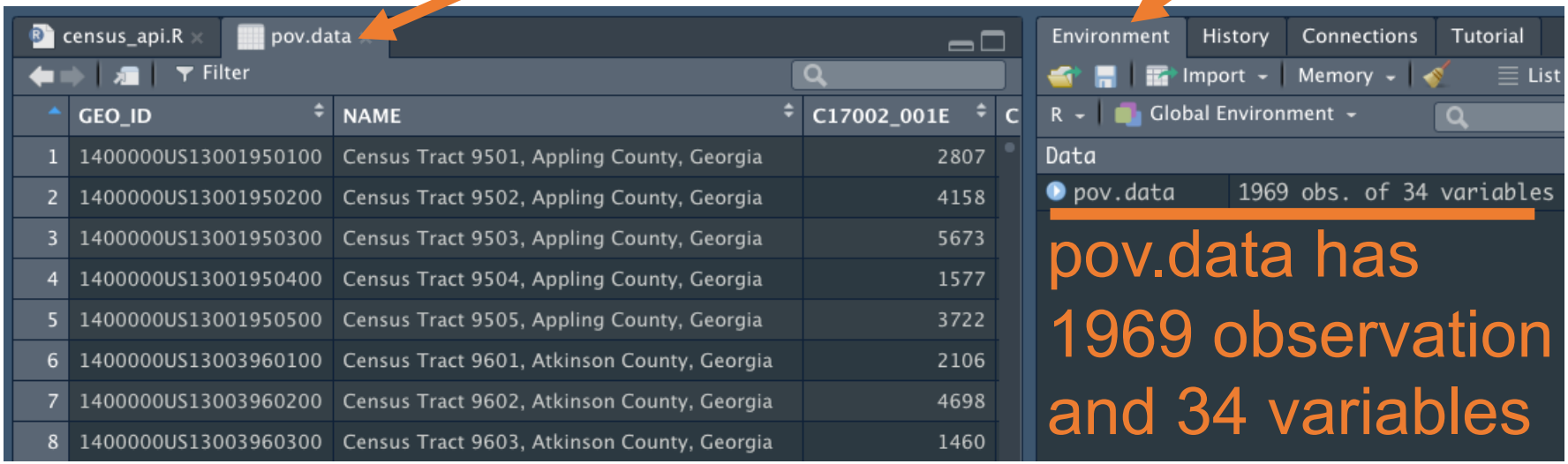
```
head(pov.data)
```

head() shows the first six rows of the data.
Useful when you want to take a quick glance of your data

Credit: Xiaofan Liang

Examining data in R

- After read.csv, pov.data shows up in R environment. Click on pov.data, an Excel-like table pops up



	GEO_ID	NAME	C17002_001E
1	1400000US13001950100	Census Tract 9501, Appling County, Georgia	2807
2	1400000US13001950200	Census Tract 9502, Appling County, Georgia	4158
3	1400000US13001950300	Census Tract 9503, Appling County, Georgia	5673
4	1400000US13001950400	Census Tract 9504, Appling County, Georgia	1577
5	1400000US13001950500	Census Tract 9505, Appling County, Georgia	3722
6	1400000US13003960100	Census Tract 9601, Atkinson County, Georgia	2106
7	1400000US13003960200	Census Tract 9602, Atkinson County, Georgia	4698
8	1400000US13003960300	Census Tract 9603, Atkinson County, Georgia	1460

Credit: Xiaofan Liang

Calculating the descriptive statistics

- Today, we are interested in the mean value of total population, which is the variable named 'C17002_001E' in pov.data.

- We need to extract only the needed variable from the data.frame

Credit: Bonwoo Koo

$ operator

pov.data

A	B	C	D	E	F
GEO_ID	NAME	C17002_001E	C17002_001EA	C17002_001M	C17002_001MA
1400000US13001950100	Census Tract 9501, Appling County, Georgia	2807	null	356	null
1400000US13001950200	Census Tract 9502, Appling County, Georgia	4158	null	515	null
1400000US13001950300	Census Tract 9503, Appling County, Georgia	5673	null	491	null
1400000US13001950400	Census Tract 9504, Appling County, Georgia	1577	null	226	null
1400000US13001950500	Census Tract 9505, Appling County, Georgia	3722	null	463	null
1400000US13003960100	Census Tract 9601, Atkinson County, Georgia	2106	null	255	null
1400000US13003960200	Census Tract 9602, Atkinson County, Georgia	4698	null	281	null
1400000US13003960300	Census Tract 9603, Atkinson County, Georgia	1460	null	212	null

GEO_ID

A
GEO_ID
1400000US13001950100
1400000US13001950200
1400000US13001950300
1400000US13001950400
1400000US13001950500
1400000US13003960100
1400000US13003960200
1400000US13003960300

+

NAME

B
NAME
Census Tract 9501, Appling County, Georgia
Census Tract 9502, Appling County, Georgia
Census Tract 9503, Appling County, Georgia
Census Tract 9504, Appling County, Georgia
Census Tract 9505, Appling County, Georgia
Census Tract 9601, Atkinson County, Georgia
Census Tract 9602, Atkinson County, Georgia
Census Tract 9603, Atkinson County, Georgia

+

C17002_001E (total population estimate)

C
C17002_001E
2807
4158
5673
1577
3722
2106
4698
1460

+ ...

Credit: Xiaofan Liang; Bonwoo Koo

$ operator

pov.data

A	B	C	D	E	F
GEO_ID					
1400000US13001950100					
1400000US13001950200					
1400000US13001950300					
1400000US13001950400					
1400000US13001950500					
1400000US13003960100					
1400000US13003960200					
1400000US13003960300	Census Tract 9603, Atkinson County, Georgia	1460	null	212	null

pov.data$C17002_001E is a R-way of saying: "give me the variable named C17002_001E from the data.frame called pov.data"

GEO_ID

A
GEO_ID
1400000US13001950100
1400000US13001950200
1400000US13001950300
1400000US13001950400
1400000US13001950500
1400000US13003960100
1400000US13003960200
1400000US13003960300

+

NAME

B
NAME
Census Tract 9501, Appling County, Georgia
Census Tract 9502, Appling County, Georgia
Census Tract 9503, Appling County, Georgia
Census Tract 9504, Appling County, Georgia
Census Tract 9505, Appling County, Georgia
Census Tract 9601, Atkinson County, Georgia
Census Tract 9602, Atkinson County, Georgia
Census Tract 9603, Atkinson County, Georgia

+

$C17002_001E (total population estimate)

C
C17002_001E
2807
4158
5673
1577
3722
2106
4698
1460

+ ...

Name of data.frame $Name of variable

Credit: Xiaofan Liang

Descriptive Stats in R

- mean(pov.data$C17002_001E)
- median(pov.data$C17002_001E)
- min(pov.data$C17002_001E)
- max(pov.data$C17002_001E)

What is the data type of pov.data$C17002_001E?

Will the function still work if the column has NAs?

Tips: mean(pov.data$C17002_001E, na.rm=T)

Credit: Xiaofan Liang

Data Cleaning with Base R

- In a real project or a study, data **NEVER** come in a neatly cleaned form.
- More than 50~60% of my time as a researcher is put into cleaning data.
- All the statistical knowledge does not mean much if you don't know how to clean your data.

Credit: Bonwoo Koo

Common Operations in Data Cleaning

1. **Subsetting row/column**
   - Index by position
   - Index by condition
   - Index by name
2. **Changing variable names**
3. **Create new variables based on existing ones (see lab)**

Credit: Bonwoo Koo

Common Operations

- Square brackets [ ] after a vector or data.frame mean you want to subset it or access some parts of it.
- If it is a vector (i.e., 1-dimensional), you need one index

vector[index]

Credit: Bonwoo Koo

Subsetting

```
my.vector <- c(1,3,6,8,12,7,5)
```

my.vector

1	3	6	8	12	7	5

my.vector[3]

1	3	6	8	12	7	5

my.vector[2:4]

1	3	6	8	12	7	5

Credit: Bonwoo Koo

Subsetting

my.vector[c(TRUE, FALSE, FALSE, TRUE, TRUE, FALSE, TRUE)]

my.vector	1	3	6	8	12	7	5
index	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
Output:	1	8	12	5			

Only those elements in my.vector that are <u>at the same position as TRUE</u> in the index can go through.

Credit: Bonwoo Koo

Subsetting

A vector of TRUEs and FALSEs are very useful because ..

>, < , <=, =>, ==, and %in% are questions.
R answers with TRUEs and FALSEs

E.g., ***my.vector > 6*** *is the same as asking,*
"are elements in my.vector larger than 6?"

Credit: Bonwoo Koo

Subsetting

A vector of TRUEs and FALSEs are very useful because ..

>, < , <=, =>, ==, and %in% are questions.
R answers with TRUEs and FALSEs

E.g., ***my.vector > 6*** *is the same as asking,*
"are elements in my.vector larger than 6?"

my.vector	=	1	3	6	8	12	7	5
my.vector > 6	=	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE

Credit: Bonwoo Koo

Subsetting

Only those elements in my.vector that are <u>at the same position as TRUE</u> can go through.

my.vector[my.vector > 6]

my.vector	1	3	6	8	12	7	5
my.vector > 6	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE

output	1	3	6	8	12	7	5

Credit: Bonwoo Koo

Subsetting

- If it is a data.frame, you need two indices.

data.frame[index for rows, index for columns]

my.df[,]

	a	b	Avengers
Row 1	a	1	Peter
Row 2	b	2	Natasha
Row 3	c	3	Hulk
Row 4	d	4	Thor

my.df[2:3 , 1]

	a	b	Avengers
Row 1	a	1	Peter
Row 2	b	2	Natasha
Row 3	c	3	Hulk
Row 4	d	4	Thor

my.df[, "Avengers"]

	a	b	Avengers
Row 1	a	1	Peter
Row 2	b	2	Natasha
Row 3	c	3	Hulk
Row 4	d	4	Thor

my.df[c(TRUE, FALSE, TRUE, FALSE) ,]

	a	b	Avengers
Row 1	a	1	Peter
Row 2	b	2	Natasha
Row 3	c	3	Hulk
Row 4	d	4	Thor

Credit: Bonwoo Koo

Subsetting + Filtering

- Subset dataframe so that it only returns rows when column b <= 2.

data.frame[index for rows, index for columns]

mydf[mydf$b <= 2,]

	a	b	Avengers
Row 1	a	1	Peter
Row 2	b	2	Natasha

mydf$b <= 2
Returns a vector of c(TRUE, FALSE …)

mydf[mydf$b <= 2,]
Returns rows where row indices are TRUE

Credit: Xiaofan Liang

Change column names

- colnames() allows you to access column names of a matrix or a data.frame

colnames(my.df)

a	b	Avengers
a	1	Peter
b	2	Natasha
c	3	Hulk
d	4	Thor

colnames(my.df)[2]

a	b	Avengers
a	1	Peter
b	2	Natasha
c	3	Hulk
d	4	Thor

colnames(my.df)[2] <- "Defenders"

a	Defenders	Avengers
a	1	Peter
b	2	Natasha
c	3	Hulk
d	4	Thor

Inserting a new name, "Defenders", to the 2nd position of my.df's variable names

Credit: Bonwoo Koo

Data cleaning with Tidyverse

- tidyverse is a series of packages that agree to use %>% (pipe syntax) to do data wrangling and cleaning.
- install.packages("tidyverse") and Library("tidyverse") will install and load all affiliated packages
- tidyverse packages have other names for base R functions that are more intuitive (e.g., filter).
- What is %>% (pipe syntax)?

filter(mydf, b<=2) *is equivalent of* mydf %>% filter(b<=2)

%>% assumes the first input in the function is in front of the %>%

Credit: Xiaofan Liang

Data cleaning with Tidyverse

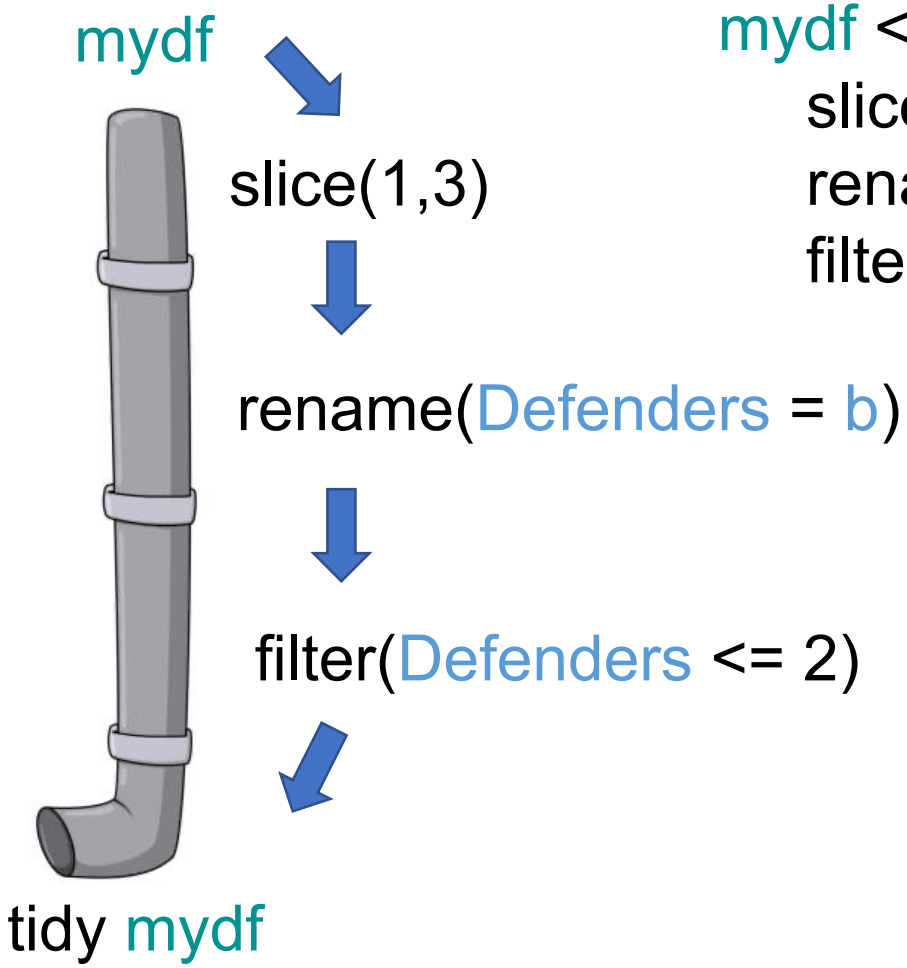
- Why do people use Tidyverse or %>% (pipe syntax)?

	Base R	Tidyverse
Subset 1st and 3rd row rows	mydf[c(1,3),]	mydf %>% slice(1,3)
Rename columns	colnames(mydf)[2] <- "Defenders"	mydf %>% rename(Defenders = b)
Subset and filter by condition	mydf[mydf$b <= 2,]	mydf %>% filter(Defenders <= 2)
All at once	mydf <- mydf[c(1,3),] mydf <- mydf[mydf$b <= 2,] colnames(mydf)[2] <- "Defenders"	mydf <- mydf %>% slice(1,3) %>% rename(Defenders = b) %>% filter(Defenders <= 2)

Tips: you don't have to remember all the syntax at once!! Go to R cheatsheet: https://www.rstudio.com/resources/cheatsheets/ (e.g., Base R, Advanced R, Data Transformation with dplyr and tidyr (both are included in Tidyverse))

Credit: Xiaofan Liang

Data cleaning with Tidyverse




```
mydf <- mydf %>%
    slice(1,3) %>%
    rename(Defenders = b) %>%
    filter(Defenders <= 2)
```


Credit: Xiaofan Liang

Commenting on script

- # is used for documentation.
- # makes anything written after it invisible to R.
- R will not run anything that is written after #.

```
##################################
### Date: Aug 21, 2019          ###
### Title: CP6025 Lab 2 material ###
### Author: Bonwoo Koo          ###
##################################

# This code is prepared as an example of
# commenting on your R script.

# Load required libraries
library(swirl)
library(sf)
library(tidyverse)

# Set the working directory
setwd("C:/Users/Bonwoo Koo/Dropbox/School/CP6025/Labs/Lab2")

# Check what is in the new working directory
dir()

# Read the data file into R
testdata <- read.csv("testdata.csv")

# Examine the data
head(testdata)
str(testdata)

# Print descriptive statistics to check the distribution
# and whether there are any missing values
summary(testdata)
```

Submit assignment R code

- When you save your script, it will create a file with .R extension.

NHTS2017_Dec14_TidyCode.R	1/18/2019 (Friday) 10:3...	R File	73 KB
NHTS2017_functions_and_dictionaries.R	7/16/2018 (Monday) 3:...	R File	10 KB

- In the assignments, you will be asked to submit your code. Submit this .R file.

- In the assignments, you will also be asked to submit plots (export or screenshot) in the future. Don't submit .R file instead.

Credit: Bonwoo Koo

Start working!

Download Modules/Week 2/Lab2.zip → unzip Lab2.zip → open lab2.html

Start reading through the document and try to replicate each step in your R-Studio

You don't need to submit your replication. The lab familiarizes you with basic functions for assignment 1.

Let me know if you have questions