

1.1 Prepare Datasets (breast cancer) step-by-step

Assignment link: <https://ilearn.sfsu.edu/ay2223/mod/assign/view.php?id=36844>

Below are the steps to download and/or cut and paste the datasets and edit them to make them “clean”. Steps for creating plots in R and saving the script are also given.

 Note: “Play video” symbols next to descriptions are linked to related video content describing some of these steps.

Section: - Breast Table/Figure: Table 4.9: Annual Death Rates

Submit

Download and Print: [Download Printer-friendly PDF](#) [Download data](#)

Table 4.9
Cancer of the Female Breast (Invasive)

Age-adjusted U.S. Death^a Rates by Year, Race and Age

Year of Death	All Races, Females			White Females			Black Females		
	All Ages	Ages <50	Ages 50+	All Ages	Ages <50	Ages 50+	All Ages	Ages <50	Ages 50+
1975-2017	26.94	6.40	80.75	26.77	6.06	81.03	32.72	9.89	92.52
1975	31.45	9.11	89.94	31.79	9.01	91.42	29.49	10.68	78.75
1976	31.80	8.74	92.21	32.17	8.64	93.80	30.47	10.22	83.50

Download and clean dataset file #1

Download link (SEER 1975-2017):
https://seer.cancer.gov/archive/csr/1975_2017/browse_csr.php

(1) Once on the page, select *Breast* and *Table 4.9 Annual Death Rates* from the drop-down menus and click Submit. Scroll down to the resulting Table 4.9. Click *Download data* button.

sect_04_table.09

Possible Data Loss: Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

	A	B	C	D	E	F	G	H	I	J	K
1	Table 4.9										
2	Cancer of the Female Breast (Invasive)										
3											
4	Age-adjusted U.S. Death(a) Rates by Year, Race and Age										
5	Year of Death All Races, Fer All Races, Fer White Femal White Female Black Female Black Females, Ages 50+										
6	1975-2017	26.94	6.4	80.75	26.77	6.06	81.03	32.72	9.89	92.52	
7	1975	31.45	9.11	89.94	31.79	9.01	91.42	29.49	10.68	78.75	
8	1976	31.8	8.74	92.21	32.17	8.64	93.8	30.47	10.22	83.5	
9	1977	32.48	8.86	94.34	32.67	8.75	95.32	32.8	10.52	91.13	
10	1978	31.73	8.69	92.08	31.9	8.46	93.29	32.14	11.37	86.53	

(2) Open the downloaded file `sect_04_table.09.csv` in a text-only editor and/or Excel and start cleaning up. Delete rows 1–4 and 6 (1975–2017) shown in light red. This is the Excel view.

Note: you must install [Excel](#) (available [through school](#)) on your computer. You must also install a text-only editor such as [BBEdit](#) (Mac) or [Notepad++](#) (PC).

sect_04_table.09.csv

```

1 " Table 4.9"
2 " Cancer of the Female Breast (Invasive)"
3
4 " Age-adjusted U.S. Death(a) Rates by Year, Race and Age"
5 "Year of Death","All Races, Females,All Ages","All Races, Females,Ages <50","All Races, Females,Ages 50+"
6 "1975-2017","26.94","6.40","80.75","26.77","6.06","81.03","32.72","9.89","92.52"
7 "1975","31.45","9.11","89.94","31.79","9.01","91.42","29.49","10.68","78.75"
8 "1976","31.80","8.74","92.21","32.17","8.64","93.80","30.47","10.22","83.50"
9 "1977","32.48","8.86","94.34","32.67","8.75","95.32","32.80","10.52","91.13"
10 "1978","31.73","8.69","92.08","31.90","8.46","93.29","32.14","11.37","86.53"

```

Note a: later, when importing the file into R, the program will still not like the “+” and “-” characters in the header names, just as it would not have liked the “<” character, and it will replace them with a period. However, it does not matter since we will only be using the “all ages” columns for black and white female, and not the columns for younger-than-50 and older-than-50 columns.

Note b: also, when importing the file into R, one can check a “quotes” box to remove the quotation marks automatically if they are still present, but it’s good practice to make the final dataset as clean as possible by truly removing them in the original file. In general, anything that is a “string”, like a short phrase that contains punctuation as in these original file headers, does require quotation marks. Hence, keep the header names short and without spaces to simplify things.

(3) Below is the same view of the original file now opened with the text-only editor. Shorten header names and delete all quotation marks, if any. Tip: opening the file in Excel first and doing a simple “Save”, then opening the new saved file in the text-only editor will automatically remove the quotation marks.

Very important: delete any commas (after “All Races” and “Females”) inside header names. Delete also the “<” character (less-than sign) in the header in columns C, F, and I and add a minus sign instead after “50”.

Scroll down to the bottom of the file and delete any extra, non-data rows. Note that in Excel you might not see any empty rows but they might show up in the text-only version as a series of commas one after the other. So make sure to open the file with the text-only editor and save the final version as a CSV file from inside that program.

```

breast cancer 1975-2017 short header no quotes.csv
~/Desktop/breast cancer prepare datasets 2022-01-30/dataset 1 1975-2017/02 clean dataset/breast cancer 1975-2017 short header no quotes.csv
1 year,allRaceAge,allRace50-,allRace50+,allWhite,white50-,white50+,allBlack,black50-,black50+
2 1975,31.45,9.11,89.94,31.79,9.01,91.42,29.49,10.68,78.75
3 1976,31.80,8.74,92.21,32.17,8.64,93.80,30.47,10.22,83.50
4 1977,32.48,8.86,94.34,32.67,8.75,95.32,32.80,10.52,91.13
5 1978,31.73,8.69,92.08,31.90,8.46,93.29,32.14,11.37,86.53

```

(4) This is the view in text-only editor of the clean file with simplified short header names (note the “camelCase” capitalization and no spaces in the header, or column, names. All quotation marks have also been removed. The name of the file shown is my own working file name.

Age-adjusted U.S. Death Rates by Year, Race and Age

Year of Death	All Races, Females			White Females			Black Females		
	All Ages	Ages <50	Ages 50+	All Ages	Ages <50	Ages 50+	All Ages	Ages <50	Ages 50+
1975-2017	26.94	6.40	80.75	26.77	6.06	81.03	32.72	9.89	92.52
1975	31.45	9.11	89.94	31.79	9.01	91.42	29.49	10.68	78.75
1976	31.80	8.74	92.21	32.17	8.64	93.80	30.47	10.22	83.50

(5) Open the same file in Excel and place it on top of the webpage to visually compare it to the original html table, to make sure the edited file is consistent with the view displayed online. Spot-check a few data points for a few random years just to make sure that nothing shifted in the process.

(6) Open file again in both Excel...

(7) ... and finally in the text-only editor to check for possible “dirty characters” such as stray commas, brackets, and anything else that might look weird. Turn on invisible characters (line breaks, word spaces, etc) which can also reveal unwanted formatting “dirt”.

```

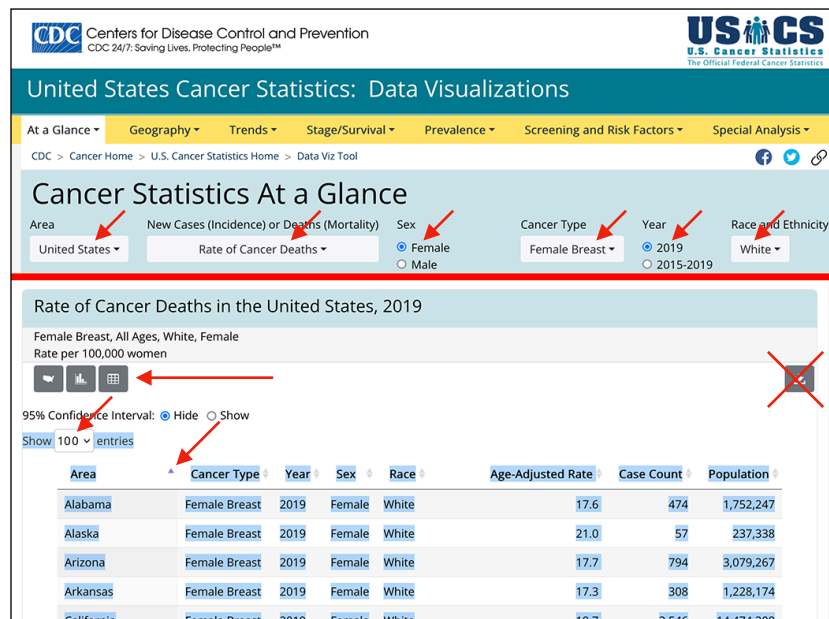
trogu_brecan_75_17.csv
~/Desktop/breast cancer prepare datasets 2022-01-30/dataset 1 1975-2017/02 clean dataset/trogu_brecan_75_17.csv
1 year,allRaceAge,allRace50-,allRace50+,allWhite,white50-,white50+,allBlack,black50-,black50+
2 1975,31.45,9.11,89.94,31.79,9.01,91.42,29.49,10.68,78.75
3 1976,31.80,8.74,92.21,32.17,8.64,93.80,30.47,10.22,83.50
4 1977,32.48,8.86,94.34,32.67,8.75,95.32,32.80,10.52,91.13
5 1978,31.73,8.69,92.08,31.90,8.46,93.29,32.14,11.37,86.53
6 1979,31.21,8.58,90.47,31.48,8.34,92.07,30.82,11.36,81.78
7 1980,31.68,8.57,92.22,31.93,8.39,93.56,31.68,10.82,86.31
8 1981,31.92,8.46,93.33,32.12,8.20,94.77,32.55,11.48,87.74
9 1982,32.19,8.42,94.43,32.31,8.19,95.46,33.75,11.19,92.83
10 1983,32.07,8.16,94.70,32.20,7.89,95.85,33.53,11.17,92.10
11 1984,32.90,8.58,96.57,32.90,8.22,97.55,35.94,12.55,97.21
12 1985,32.90,8.58,96.57,32.90,8.22,97.55,35.94,12.55,97.21

```

I sized screenshots 6 and 7 to visually compare the columns vertically, to make sure they matched. Resize the windows on your screen to do the same.

End Dataset File #1

Save the file as .csv and rename it:
yourLastName_brecan_75_17.csv



Download two files and combine them into dataset file #2

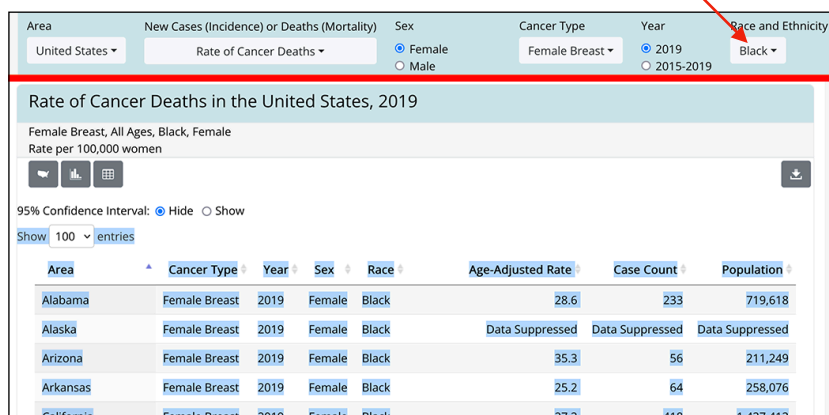
Download link (CDC 2019 white & black):
<https://gis.cdc.gov/Cancer/USCS/#/AtAGlance/>

(8) Once on webpage, select these options: *US, Rate of Cancer Deaths, Female, Female Breast, 2019, White*. The US map will render below by default. Click the “table” button (third from left) to switch to spreadsheet view.

Select *Show 100 entries* and click Area’s sorting button to sort table alphabetically by state name (left column).

Important: the download data button at right will download an incomplete spreadsheet with the state names missing (as of 2022-07-24). Since this download link does not work, select instead all the table text (rows and columns) and paste the selection into a text-only editor. Start selecting a bit above the first row and include a bit after the last row to make sure to capture everything.

Select and cut-paste white data first and save the text file, something like *white2019temp.csv*



(9) Repeat the above steps for *Rate of Cancer Deaths* for black race. Select and cut-paste selection into a new text-only editor file and save, something like *black2019temp.csv*

```

state,cancerType,year,sex,race,blackRate,blackCount,blackPop
1 Alabama,Female Breast,2019,Female,Black,28.6,233,719618
2 Alaska,Female Breast,2019,Female,Black,Data Suppressed,Data Suppressed,Data Suppressed
3 Arizona,Female Breast,2019,Female,Black,35.3,56,211249
4 Arkansas,Female Breast,2019,Female,Black,25.2,64,258076
5 California,Female Breast,2019,Female,Black,27.3,418,1437413

```

(10) Open “black” file in text-only editor. Delete commas (thousands separator) from population data numbers, and then substituting all invisible “tab” characters with commas. Simplify header names and make sure to add “black” to the “Rate”, “Count”, and “Population” header names. Note, sequence in video differs in that I edit white first and then black. Save “black” file.

	A	B	C	D	E	F	G	H	I
1	state	cancerType	year	sex	race	blackRate	blackCount	blackPop	
2	Alabama	Female Brea	2019	Female	Black	28.6	233	719618	
3	Alaska	Female Brea	2019	Female	Black	Data Suppre	Data Suppre	Data Suppressed	
4	Arizona	Female Brea	2019	Female	Black	35.3	56	211249	
5	Arkansas	Female Brea	2019	Female	Black	25.2	64	258076	
6	California	Female Brea	2019	Female	Black	27.3	418	1437413	

(11) Open the “black” file in Excel to check that everything looks good. Later, you will delete columns B,C,D, and E: Cancer Type, Year, Sex, and Race.

1	state,cancerType,year,sex,race,whiteRate,whiteCount,whitePop
2	Alabama,Female Breast,2019,Female,White,17.6,474,1752247
3	Alaska,Female Breast,2019,Female,White,21.0,57,237338
4	Arizona,Female Breast,2019,Female,White,17.7,794,3079267
5	Arkansas,Female Breast,2019,Female,White,17.3,308,1228174
6	California,Female Breast,2019,Female,White,19.7,3546,14474208

(12) Open “white” file in text-only editor and repeat steps 10 and 11. Save “white” file.

	A	B	C	D	E	F	G	H
1	state	cancerType	year	sex	race	whiteRate	whiteCount	whitePop
2	Alabama	Female Brea	2019	Female	White	17.6	474	1752247
3	Alaska	Female Brea	2019	Female	White	21	57	237338
4	Arizona	Female Brea	2019	Female	White	17.7	794	3079267
5	Arkansas	Female Brea	2019	Female	White	17.3	308	1228174
6	California	Female Brea	2019	Female	White	19.7	3546	14474208

(13) Open the “white” file in Excel to check that everything looks good.

	A	B	C	D	E	F	G	H	I
1	state	cancerType	year	sex	race	state	blackRate	blackCount	blackPop
2	Alabama	Female Brea	2019	Female	Black	Alabama	28.6	233	719618
3	Alaska	Female Brea	2019	Female	Black	Alaska	Data Suppre	Data Suppre	Data Suppressed
4	Arizona	Female Brea	2019	Female	Black	Arizona	35.3	56	211249
5	Arkansas	Female Brea	2019	Female	Black	Arkansas	25.2	64	258076
6	California	Female Brea	2019	Female	Black	California	27.3	418	1437413

(14) The next step is to combine both white and black data into a single file, being careful not to scramble the data across the states. Open the “black” file in Excel and copy the state (area) column and paste/insert it before the “blackRate” column.

	A	B	C	D	E	F	G	H	I
1	state	cancerType	year	sex	race	state	blackRate	blackCount	blackPop
2	Alabama	Female Brea	2019	Female	Black	Alabama	28.6	233	719618
3	Alaska	Female Brea	2019	Female	Black	Alaska	Data Suppre	Data Suppre	Data Suppressed
4	Arizona	Female Brea	2019	Female	Black	Arizona	35.3	56	211249
5	Arkansas	Female Brea	2019	Female	Black	Arkansas	25.2	64	258076
6	California	Female Brea	2019	Female	Black	California	27.3	418	1437413

(15) Next, select and copy columns F,G,H, and I, which include both the states and the actual data.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	state	cancerType	year	sex	race	whiteRate	whiteCount	whitePop	state	blackRate	blackCount	blackPop	
2	Alabama	Female Brea	2019	Female	White	17.6	474	1752247	Alabama	28.6	233	719618	
3	Alaska	Female Brea	2019	Female	White	21	57	237338	Alaska	Data Suppre	Data Suppre	Data Suppressed	
4	Arizona	Female Brea	2019	Female	White	17.7	794	3079267	Arizona	35.3	56	211249	
5	Arkansas	Female Brea	2019	Female	White	17.3	308	1228174	Arkansas	25.2	64	258076	
6	California	Female Brea	2019	Female	White	19.7	3546	14474208	California	27.3	418	1437413	

(16) Now open the “white” file and paste those four columns on the right side, after the “white” data. Scroll down to make sure that the two duplicate state columns align and that therefore both white and black data belong in their respective state rows.

At this point “save as” into a new file and name it something like [black_white2019temp.csv](#). Keep this file as a backup file that includes also the “Data Suppressed” rows, which will be deleted in the next cleaner and final version.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	state	cancerType	year	sex	race	whiteRate	whiteCount	whitePop	state	blackRate	blackCount	blackPop	
2	Alabama	Female Brea	2019	Female	White	17.6	474	1752247	Alabama	28.6	233	719618	
3	Alaska	Female Brea	2019	Female	White	21	57	237338	Alaska	Data Suppre	Data Suppre	Data Suppressed	
4	Arizona	Female Brea	2019	Female	White	17.7	794	3079267	Arizona	35.3	56	211249	
5	Arkansas	Female Brea	2019	Female	White	17.3	308	1228174	Arkansas	25.2	64	258076	
6	California	Female Brea	2019	Female	White	19.7	3546	14474208	California	27.3	418	1437413	

(17) Duplicate the file and name it [yourLastName_brecan_wb_2019.csv](#) — this will be the final name of the cleaned file. In this final file, start by deleting the duplicate state column.

	A	B	C
1	state	cancerType	year
2	Alabama	Female Brea	2019
3	Alaska	Female Brea	2019
4	Arizona	Female Brea	2019
5	Arkansas	Female Brea	2019

(18) Next, sort the data: Data —> Sort —> Sort by blackRate —> Largest to Smallest. Make sure to “expand selection” when asked in the prompt.

	A	B	C	D	E	F	G	H	I	J	K	L
1	state	cancerType	year	sex	race	whiteRate	whiteCount	whitePop	blackRate	blackCount	blackPop	
2	Alaska	Female Brea	2019	Female	White	21	57	237338	Data Suppre	Data Suppre	Data Suppressed	
3	Hawaii	Female Brea	2019	Female	White	31.3	68	165524	Data Suppre	Data Suppre	Data Suppressed	
4	Idaho	Female Brea	2019	Female	White	18.8	199	843712	Data Suppre	Data Suppre	Data Suppressed	
5	Iowa	Female Brea	2019	Female	White	17.9	381	1457497	Data Suppre	Data Suppre	Data Suppressed	
6	Maine	Female Brea	2019	Female	White	17.8	190	657275	Data Suppre	Data Suppre	Data Suppressed	
7	Montana	Female Brea	2019	Female	White	20	144	480557	Data Suppre	Data Suppre	Data Suppressed	
8	New Hampsh	Female Brea	2019	Female	White	20.1	199	648049	Data Suppre	Data Suppre	Data Suppressed	
9	New Mexico	Female Brea	2019	Female	White	21.7	264	878808	Data Suppre	Data Suppre	Data Suppressed	
10	North Dakota	Female Brea	2019	Female	White	14.4	63	329026	Data Suppre	Data Suppre	Data Suppressed	
11	Oregon	Female Brea	2019	Female	White	20	542	1888940	Data Suppre	Data Suppre	Data Suppressed	
12	Rhode Island	Female Brea	2019	Female	White	18.7	134	462310	Data Suppre	Data Suppre	Data Suppressed	
13	South Dakota	Female Brea	2019	Female	White	17.6	97	377131	Data Suppre	Data Suppre	Data Suppressed	
14	Utah	Female Brea	2019	Female	White	19.5	275	1465880	Data Suppre	Data Suppre	Data Suppressed	
15	Vermont	Female Brea	2019	Female	White	16	78	301932	Data Suppre	Data Suppre	Data Suppressed	
16	West Virginia	Female Brea	2019	Female	White	20.6	269	860249	Data Suppre	Data Suppre	Data Suppressed	
17	Wyoming	Female Brea	2019	Female	White	17.3	59	267327	Data Suppre	Data Suppre	Data Suppressed	
18	Nebraska	Female Brea	2019	Female	White	21.7	255	864495	50.5	19	56349	
19	Arizona	Female Brea	2019	Female	White	17.7	794	3079267	35.3	56	211249	
20	Nevada	Female Brea	2019	Female	White	24.1	352	1152469	33.8	55	180680	
21	Illinois	Female Brea	2019	Female	White	19.3	1336	4958604	32.6	373	1026358	
22	District of C	Female Brea	2019	Female	White	14.3	21	165524	32.5	76	183861	
23	Delaware	Female Brea	2019	Female	White	18.8	103	352448	29.3	37	125805	
24	Louisiana	Female Brea	2019	Female	White	19	411	1495874	29.3	250	821149	

(19) As a result of the sorting, all the rows with “Data Suppressed” will now be on top. Select and delete those rows as well as the District of Columbia row. Note that this action will remove some states completely, even if those states had actual data for “white”.

	A	B	C	D	E	F	G	H	I	J	K
1	state	cancerType	year	sex	race	whiteRate	whiteCount	whitePop	blackRate	blackCount	blackPop
2	Alabama	Female Brea	2019	Female	White	17.6	474	1752247	28.6	233	719618
3	Arizona	Female Brea	2019	Female	White	17.7	794	3079267	35.3	56	211249
4	Arkansas	Female Brea	2019	Female	White	17.3	308	1228174	25.2	64	258076
5	California	Female Brea	2019	Female	White	19.7	3546	14474208	27.3	418	1437413

(20) Next, delete also columns B,C,D, and E. Cancer Type, Year, and Sex are not needed since we know the dataset is for 2019 female breast cancer; and white also needs deleting because this file includes both white and black data.

	A	B	C	D	E	F	G
1	state	whiteRate	whiteCount	whitePop	blackRate	blackCount	blackPop
2	Alabama	17.6	474	1752247	28.6	233	719618
3	Arizona	17.7	794	3079267	35.3	56	211249
4	Arkansas	17.3	308	1228174	25.2	64	258076
5	California	19.7	3546	14474208	27.3	418	1437413
6	Colorado	17.9	566	2528551	25.5	26	145970
7	Connecticut	16.4	378	1472048	22	55	244929

(21) The final file will include 7 columns and 35 rows: 34 data rows (states) plus 1 header row. Note again that the video may show a slightly different number since that data is a bit older.

The seven columns (variables) are:

- A. State (categories, not a true variable)
- B. White rate (deaths per 100K white female pop.)
- C. White count (total death count)
- D. White pop. (total white female population)
- E. Black rate (deaths per 100K black female pop.)
- F. Black count (total death count)
- G. Black pop. (total black female population)

Save the file but open it again in the text-only editor to double check for dirty characters that might have escaped notice.

Save as .csv file:

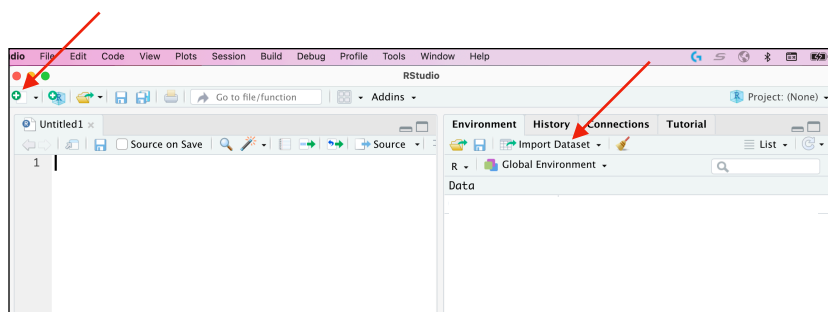
yourLastName_breacan_wb_2019.csv

End Dataset File #2

Plot matrix & scatterplot in R/RStudio

[Download and install R](#)

[Download and Install RStudio](#)



(22) Start RStudio (R will automatically run in the background)

Start a new R script (top left icon with green + sign)

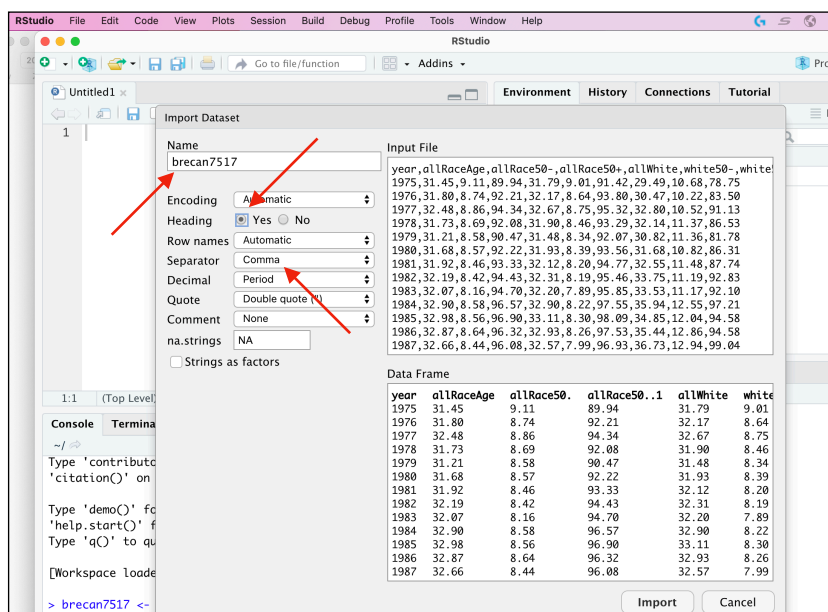
New file will be:

Untitled1

Then, in the Environment, right pane: Import dataset, from Text, base.

Find and import file 1:

[yourLastName_brecan_75_17.csv](#)



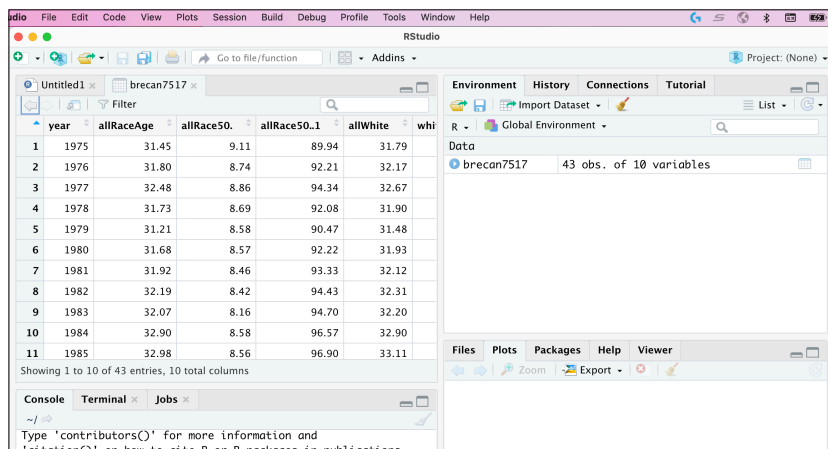
(23) When the file preview opens in the importing window, shorten the file name to [brecan7517](#) (top left in import window)

Select Heading: Yes

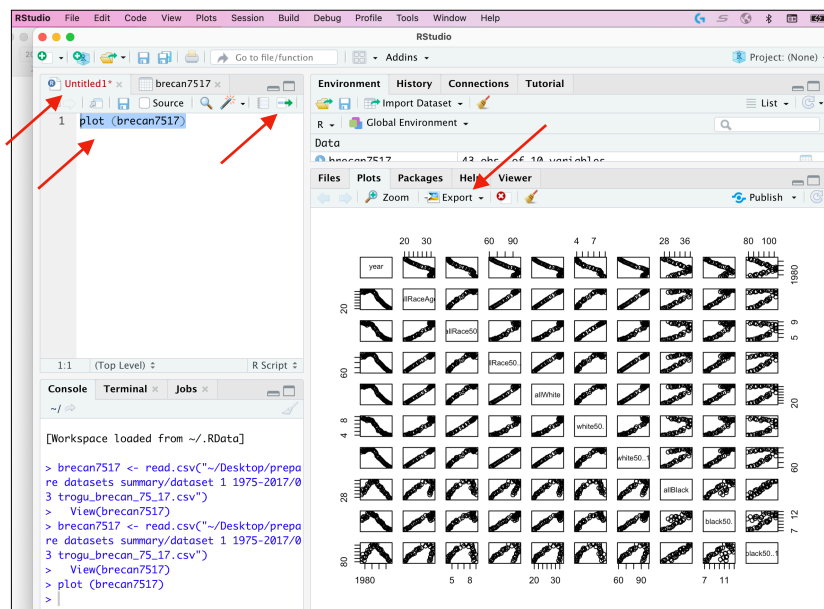
Separator: Comma

The original raw file is shown in the “Input File” pane on top, while the “Data Frame” pane at bottom shows how the program is interpreting the original file.

Make sure to select Yes for Heading, otherwise the program will generate its own headers named V1, V2, etc, and push the original header down to row #2. This will create big trouble as text will now occupy the cells in row #2 where numbers are normally expected to be. Notice also that the program does not like the “-” and “+” characters in the header names, so it inserts dots and other characters instead.



(24) The dataset will now appear in a tab of the same left pane as the script will be.



(25) Click back the **Untitled1** tab to go back to the script window; type:

`plot (brecan7517)`

The shortened filename is inside the parentheses.

Run the line of code; click icon with green arrow after selecting line or placing cursor anywhere on the line of code.

The plot matrix should render in plot window.

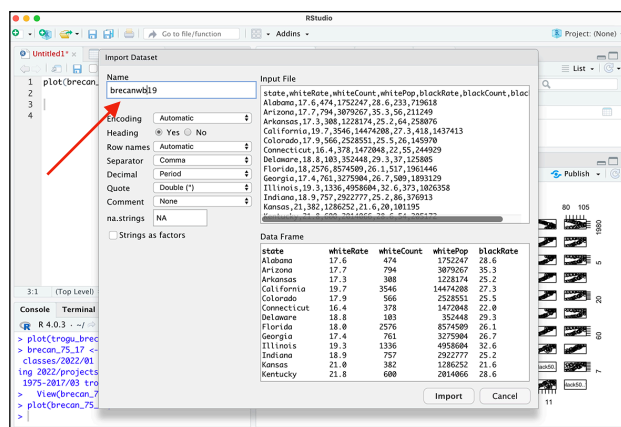
Each small graphic is a plot made by using two separate column variables. Each little circle is a spatial data point given by the combination of two other X and Y data point values (variables). All possible combinations are displayed. We will discuss this more in class.

Expand the window to get more detail and Export the plot to PDF.

Name the file:

yourLastName_7517_plot1.pdf

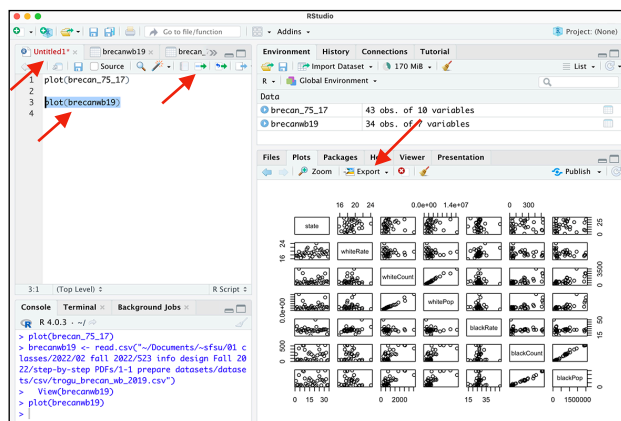
End plot matrix #1



(26) Import the second dataset:
yourLastName_brecan_wb_2019.csv

Repeat the import steps above but shorten this file name to:

brecanwb19



(27) After importing, click tab **Untitled1** (R script) and type the same plot line but with new name in parentheses:

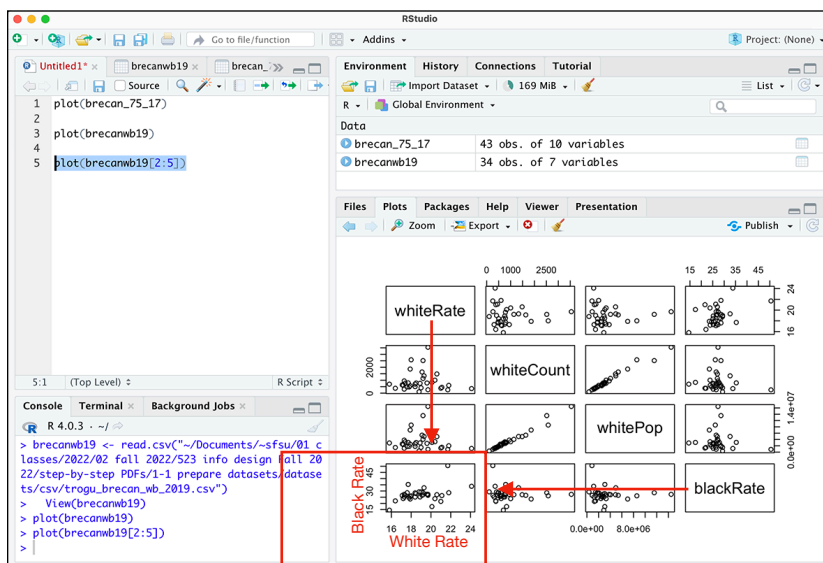
`plot (brecanwb19)`

Run the line and the new matrix plot should render in the plot window. Save the PDF (export) of the plot.

Save file as:

yourLastName_wb_19_plot2.pdf

End plot matrix #2

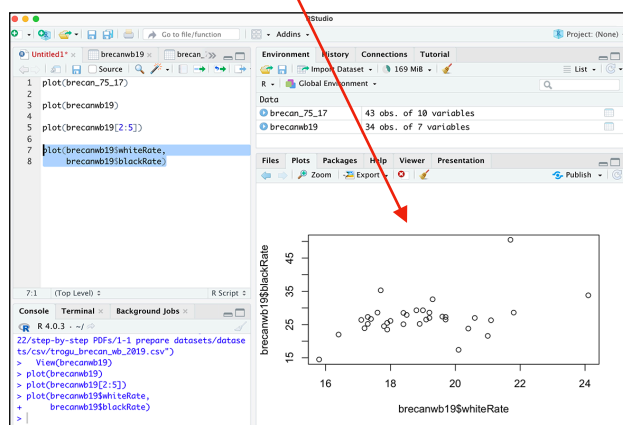


(28) Zoom in on plot matrix 2 by plotting only column 2 through 5. Column 2 is whiteRate and column 5 is blackRate. In **Untitled1** (R script) type and run:

```
plot(breanwb19[2:5])
```

Can you start to see the patterns? For each individual graph, the horizontal x-axis is labeled by the label found either above or below it in the matrix; the vertical y-axis is labeled by the label found on either left or right of it. Each dot represents a state. Thus, the plot comparing the white cancer death rate with the black rate is the plot on the lower-left-hand corner.

Can you identify plots that would not make sense to visualize?

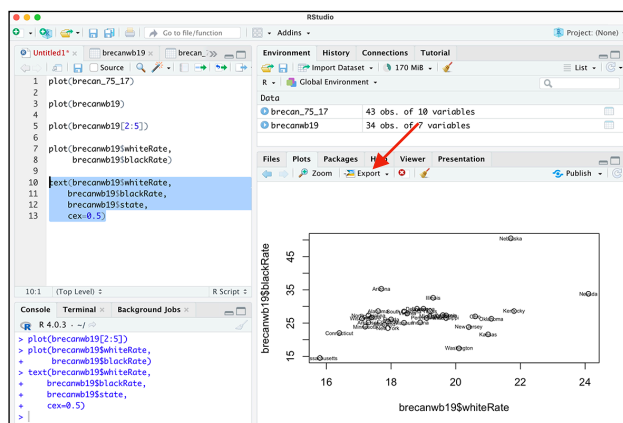


(29) Type and run this code (returns don't matter so this is still just a single line:

```
plot(breanwb19$whiteRate,
     breanwb19$blackRate)
```

Note that the horizontal white scale is longer physically but shorter in real data terms: 8 points from 16 to 24. By contrast, the vertical black scale is shorter physically but much longer in real terms: 35 points from 15 to 50. The plot's proportions can be changed by stretching the window and also on export.

Do you think this design works?



(30) Add the names of the states on top of each data point. Type and run:

```
text(breanwb19$whiteRate,
     breanwb19$blackRate,
     breanwb19$state,
     cex=0.5)
```

Plot looks very raw but it can be made better later in Illustrator. Save the PDF (export) the plot. Save file as:

yourLastName_wb_19_scatterplot3.pdf

End scatterplot #3

Save the R **Untitled1** script as:

yourLastName_brean.R

In iLearn, upload the **two datasets files**, the **three plot PDF files**, and the **R script file**.