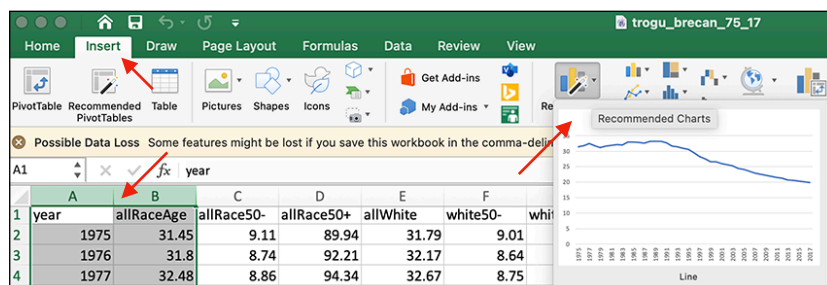


1.2 Generate graphs (breast cancer) step-by-step

Assignment page link: <https://ilearn.sfsu.edu/ay2223/mod/assign/view.php?id=36851>

Below are the steps to create the raw graphs in Excel, Tableau, and R.
See also the iLearn assignment page for more detail.

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

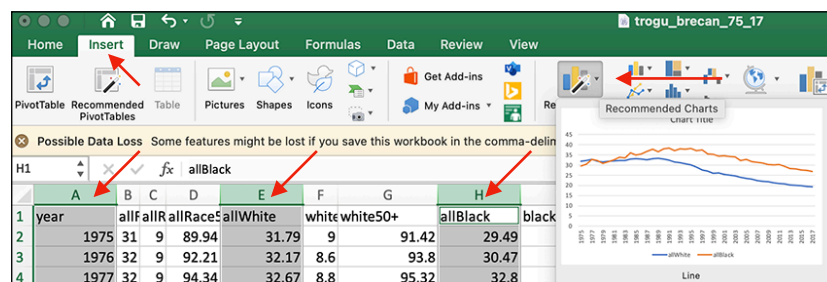


Graph 1 in Excel

(1) Single line graph (US – 1975–2017)

For graph 1, plot a line for female breast cancer death rates (rate/100K) for all US, all races, all ages, for the years 1975 to 2017. In **Excel**, open the dataset file, select the cells from the columns *Year* and *All Races Females All Ages* (in my file the header name is *allRaceAge*). With the cells selected, click —> Insert (button next to Home on left) —> Recommended charts —> Line.

That's it. The wording in the header names might be slightly different in your file. Move the chart so it does not sit on top of the data cells. Select all the empty cells now under the chart (not shown). Print —> Print: Selection --> Save as Adobe PDF --> Press Quality. Save only the PDF page that includes the chart. It might look cropped in Illustrator but Release Clipping Mask should fix it. Or modify page set up or print area to get everything before printing.

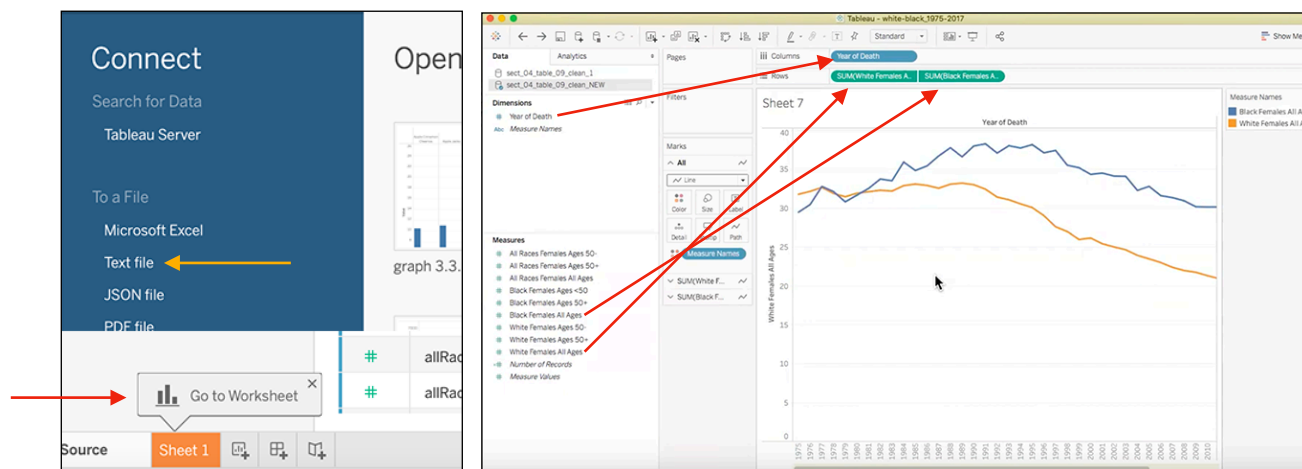


Graph 2 in Excel

(2) Double line graph (US Black & white – 1975–2017)

For the second graph, also in **Excel**, we plot two different lines together, one for Black and the other for White. Repeat the steps used for graph 1 above but select *Year*, *White Female All Ages* (*allWhite*) and *Black Female All Ages* (*allBlack*).

You could play with the various elements: axes, labels, etc, by right-clicking and editing them (Select Data, Format Axis, etc) but for now the basic graph is sufficient. Repeat the steps used for graph 1 for saving the chart (save as Adobe PDF).



After opening **Tableau** (at top-left): --> Connect --> To a file: Text file (csv) --> **lastName_brecan_75_17.csv**

After importing the dataset, create a new "Sheet 1" from bottom left menu (or press New Worksheet).

All the Tables (header names) will be at left, divided into Dimensions (or qualities, at top) and Measures (or quantities, below) however these labels become visible only when you drag items up and down. In my older Tableau videos these are always on. Dimensions are more like categories (qualities) and measures are more like actual quantities (quantitative).

After importing the dataset and starting a new sheet, make sure "Years" is a Dimension at top, or else drag it there. Then make sure it's a number or a date (it should **not** be a currency, so the year 1975 should **not** look like this: 1,975). If the year has the comma, right-click "Year" dimension (at top) and "Change Data Type" to either "Number" or "Date".

Drag the "Year" dimension to the Columns field at top; Drag the "All white" measure to the Rows field at the top (the line for white rate should appear); Now drag the "All Black" measure also to the Rows field (the line for Black rate should also appear). Right-click the Axis on right and "Synchronize Axis" (the two lines will be on the same scale). Print --> Save as PDF.

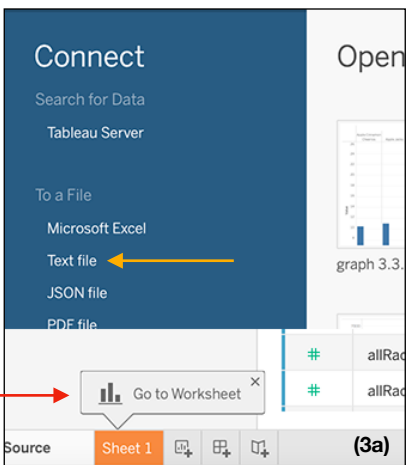
Graph 2 (double line) in Tableau

The video, from 45:40 to 49:23, shows a slight different sequence but the end result is the same. Despite the video's title (Breast Cancer Scatterplot) that section shows a way to switch from bars to lines. The dataset file names and the header names shown in the video may differ from your dataset.

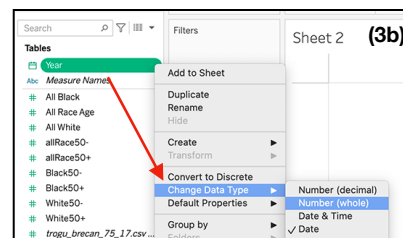


3. Connected scatterplot – 1975–2017

Still in **Tableau** and using the same dataset, for the third graph, we will plot a line connecting dots (the years) that appear to be going backwards. Each year shows the rates for black (Y vertical axis) and white (X horizontal axis) for that particular year. In the first 15 years the rate for black rises steadily while the rate for whites stays about the same. In the next 15 years both rates drop steadily, although in general black rates remain higher than white rates.



Graph 3 in Tableau

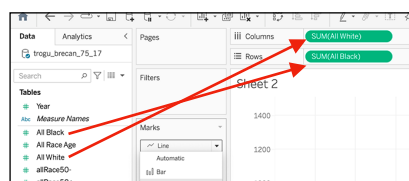


Follow the video from around **53:30** until about **59:00**. After connecting to the dataset (3a) and starting a new worksheet, make sure that "Year" is in the upper section (Dimensions) in the Tables pane at left. If it's in the lower section (Measures) together with all the other variables, drag it to the Dimensions section above. Then (3b above) right-click it --> Change Data Type --> Number (Whole).

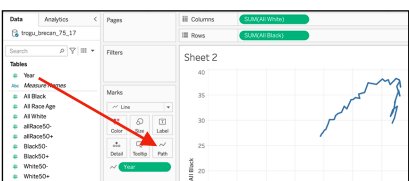
Connected scatterplot reference

Driving Safety, In Fits and Starts

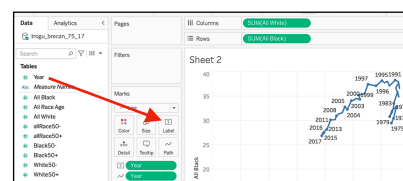
<https://archive.nytimes.com/www.nytimes.com/interactive/2012/09/17/science/driving-safety-in-fits-and-starts.html>
The New York Times, September 17, 2012



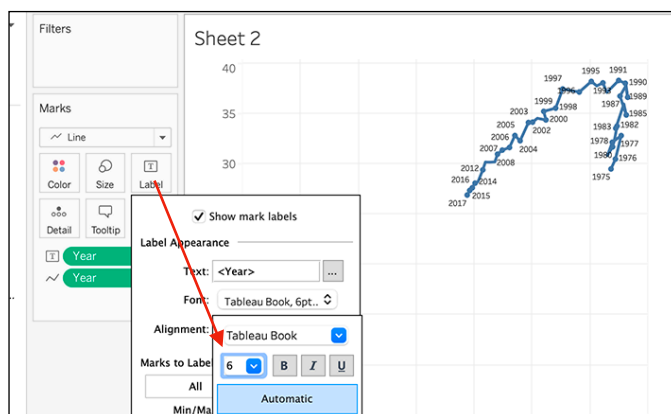
(3c) Next: drag white rate to columns
Change Marks from Automatic to Line
(Shape button will change to Path)



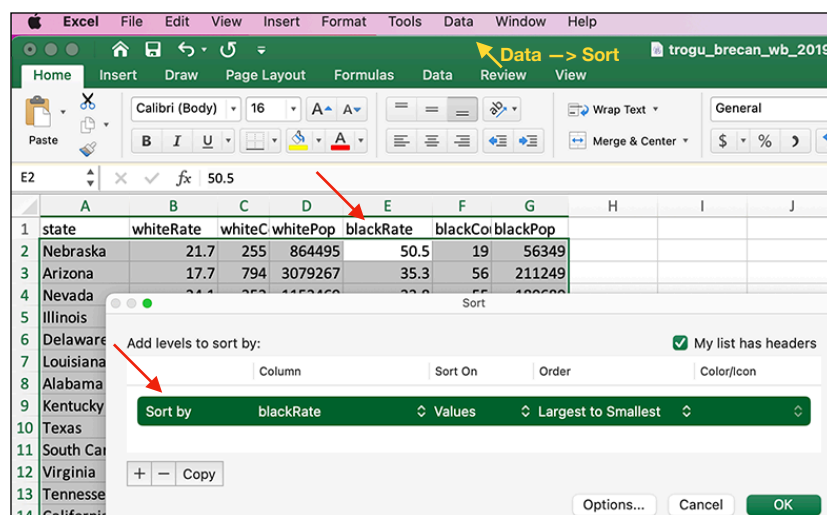
(3d) Drag Years to Path (line will appear);



(3e) Drag Year to Label to add the year labels to dots on the line.



(3d) Right-click Label to edit year's font size. Further edits are shown in video to make the chart more closely cropped: edit both Axes to start at 15 and end at 45 (Fixed range); Print --> Save as PDF (Don't export to PowerPoint).

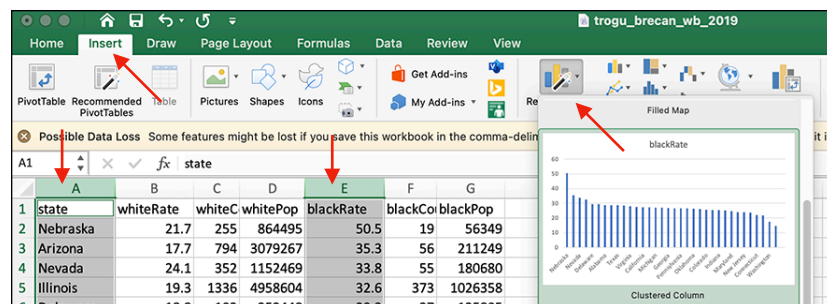


Graph 4 in Excel (black rate)

Charts 4 and 5 show death rates (rate/100K) for individual states for 2018, sorted from high to low.

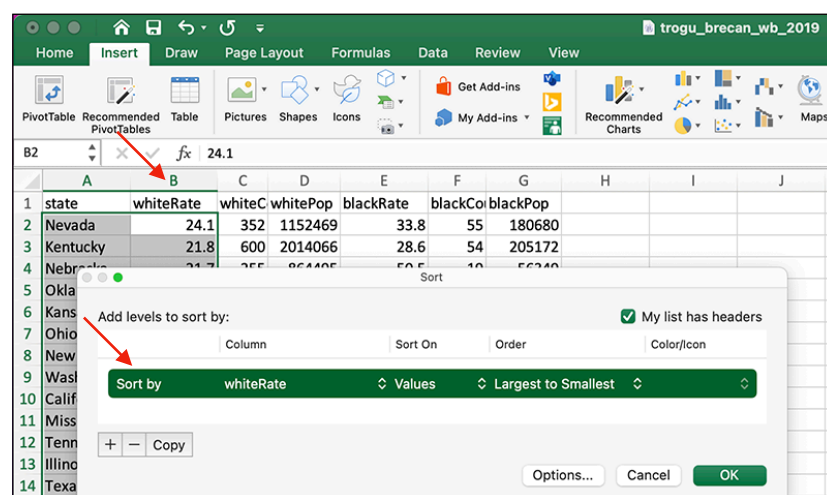
4. Single bar chart (Black – US states 2019 – sorted High to Low)

(4a) In Excel, open dataset file: lastName_brecan_wb_2019.csv then select **BlackRate** cells and sort the column: Data → Sort → Expand selection → Sort by BlackRate → Order: Largest to smallest.



(4b) Then select BlackRate and the State column: → Insert (next to Home) → Recommended charts → Clustered column.

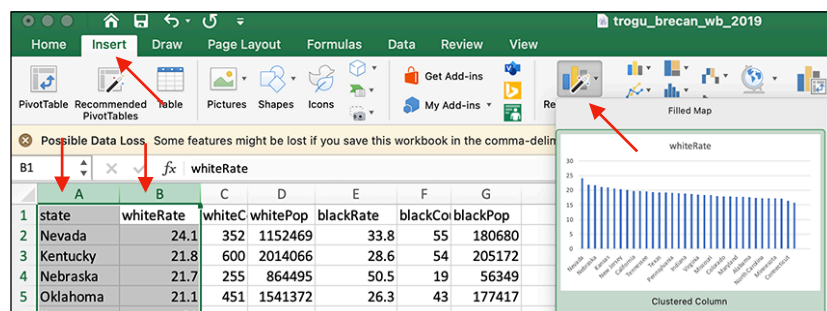
Note that it says clustered even if it shows a single column. Excel calls horizontal bars "bars"; and vertical bars "columns". Enlarge the chart so that every bar has its own State label. Play around with Page Setup (landscape) and select cells under the chart; Print → Print: Selection → Save as Adobe PDF.



Graph 5 in Excel (white rate)

5. Single bar chart (White – US states 2019 – sorted High to Low)

(5a) Using the same dataset file: lastName_brecan_wb_2019.csv select **WhiteRate** cells and sort the column: Data → Sort → Expand selection → Sort by WhiteRate → Order: Largest to smallest.




(5b) Then select WhiteRate and the State column: → Insert (next to Home) → Recommended charts → Clustered column.

Enlarge the chart so that every bar has its own State label. Play around with Page Setup (landscape) and select cells under the chart; Print → Print: Selection → Save as Adobe PDF.

Graph 4 in R (black rate)

Install R. Install and start RStudio.

 The video link at left shows how to create the bar chart (note: the dataset in video is older). The code below is current. You can also open this [annotated R file](#). Just open it in RStudio, not R. You can use the file to avoid retyping everything, just make sure your header names are the same as in the code, or adapt code as needed.

The R code shows how to sort the data (by Black rate), and also add state name labels at a 45-degree angle.

```
barplot(brecan19$blackRate,
names.arg=brecan19$state)
```

```
sortbyBlack <-
brecan19[order(brecan19$blackRate,
decreasing = TRUE), ]
```

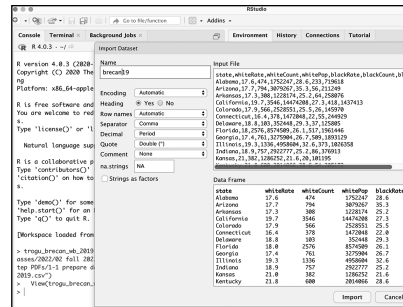
```
barplot(sortbyBlack$blackRate,
names.arg=sortbyBlack$state)
```

```
midpts <-
barplot(sortbyBlack$blackRate,
main="Black chart title -- vertical
labels")
```

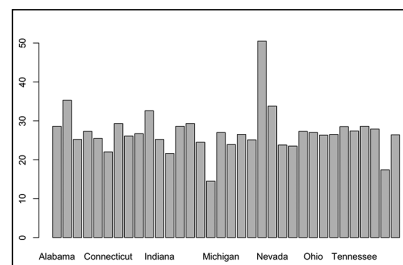
```
text(x=midpts+.8, y=-.7,
sortbyBlack$state, cex=0.5, srt=90,
xpd=TRUE, pos=2)
```

```
barplot(sortbyBlack$blackRate,
main="Black chart title -- 45-degree
labels")
```

```
text(x=midpts+.8, y=-.7,
sortbyBlack$state, cex=0.5, srt=45,
xpd=TRUE, pos=2)
```



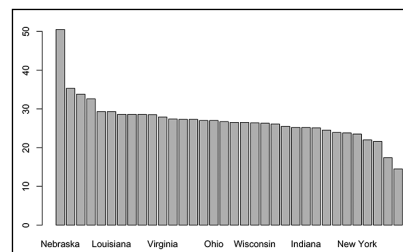
(4c – cont. from 4a & 4b) Import dataset: `lastName_brecan_wb_2019.csv` into RStudio and shorten name to `brecan19`. Make sure to select: heading, Yes. See assignment 1 for additional R tips.



(4d) Open annotated R file or start a new script. Type and run (can be on a single line):

```
barplot(brecan19$blackRate,
names.arg=brecan19$state)
```

Note that the chart is sorted alphabetically by state from left to right because the states column was already sorted in the original file. Whichever column is sorted in the original will be the default rendering, but since we are interested in the black rate (the data shown by the height of the bars) in the next step we will sort the dataset by black rate from high to low.

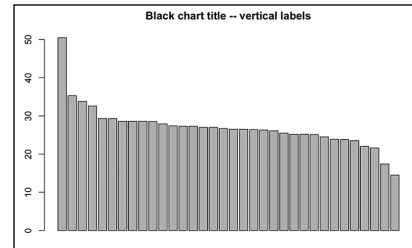


(4e) To create the new sorting by black rate, type:

```
sortbyBlack <-
brecan19[order(brecan19$blackRate,
decreasing = TRUE), ]
```

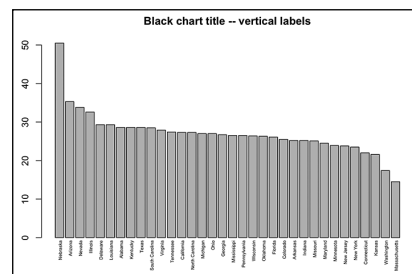
There will be no change in the plot yet. Only a new dataset will be created. Now type and run the original code but with the new name (`sortbyBlack`) to render the new graph:

```
barplot(sortbyBlack$blackRate,
names.arg=sortbyBlack$state)
```



(4f) Since not all names fit horizontally, the next code will define anchor points for the state names to fit vertically. A title is also added, that's the only visible change for now. Type and run:

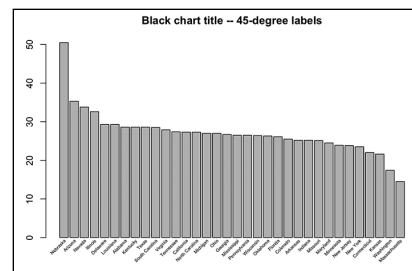
```
midpts <-
barplot(sortbyBlack$blackRate,
main="Black chart title -- vertical
labels")
```



(4g) Type and run:

```
text(x=midpts+.8, y=-.7,
sortbyBlack$state, cex=0.5, srt=90,
xpd=TRUE, pos=2)
```

The x and y anchor values can be adjusted as desired. It's better now because all state names are displayed, but it's not ideal yet. In the next two similar steps, the labels will be at 45 degrees.



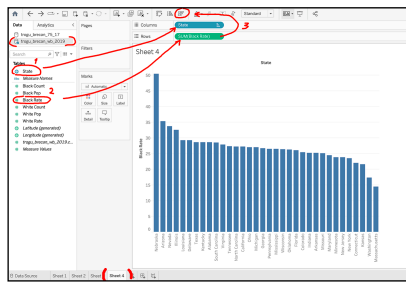
(4h) Type and run the two lines below one after the other:

```
barplot(sortbyBlack$blackRate,
main="Black chart title -- 45-degree
labels")
```

```
text(x=midpts+.8, y=-.7,
sortbyBlack$state, cex=0.5, srt=45,
xpd=TRUE, pos=2)
```

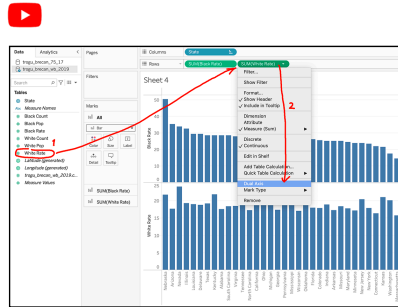
Export the plot to Adobe PDF.

Repeat the process for white rate to practice steps.

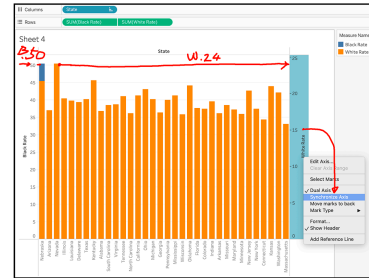


(6a) Connect to dataset (text):
lastName_breachn_wb_2019.csv
New sheet.
Drag States to Columns field
Drag Black Rate to Rows field
(Notice how Tableau renders the header
names in a more friendly way than the actual
dataset names.
Select Black Rate in Rows field and click high-
to-low sorting button (bars will be sorted from
state with highest rate (Nebraska) to lowest
(Massachusetts)).

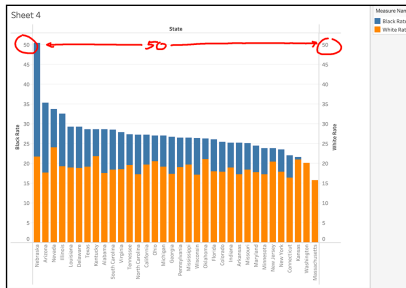
Graph 6: double (clustered, grouped, etc.) bar chart.
(Tableau)



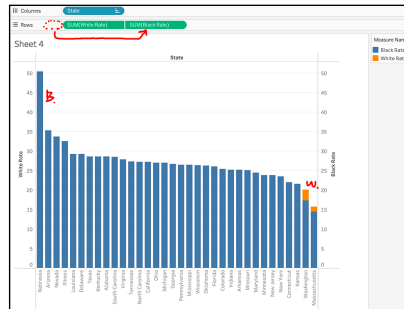
(6b) Drag White Rate to Rows field
(both charts will display)
Right-click White Rate item and select Dual Axis
from drop-down menu.
(Charts will overlap)



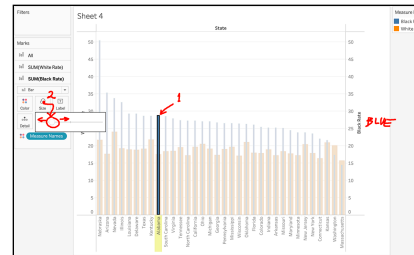
(6c) Black and white vertical scales need to be the
same (Black tops out at 50 and white at about
25). Right-click White axis and select
Synchronize Axis.



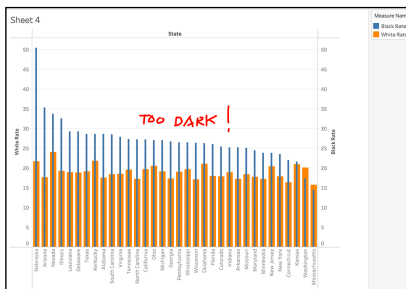
(6d)
Now the scales are same and the Black rate and
White rate bars are proportional to each other.
However the Black (blue) bars could be
confused for being “stacked” on top of the
White (orange) bars, instead of physically
behind and having a full height starting from zero.



(6e) Make the blue bars thinner but first move them
in front of the orange ones.
Switch the position of the Black and White
items at top (first White, then Black).



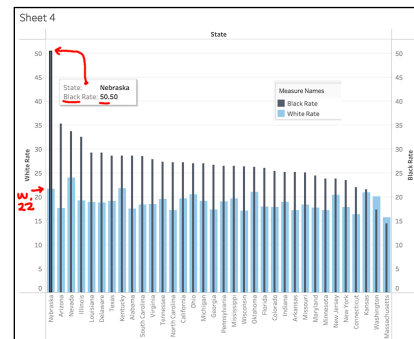
(6f) Select a blue bar, then click the Size button, use
slider to make bar thinner (about one third).
Orange bars will now be visible under the
thinner blue bars.



(6f)
But the default bar colors are still too dark, too
saturated, too everything. In next step, change
the colors so they are less active



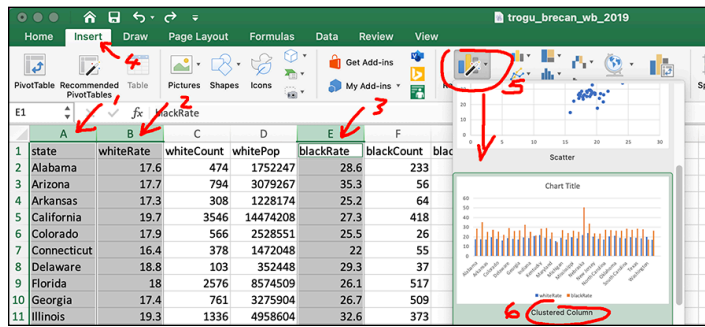
(6g) Select a bar again (1), then click the Color
button and Edit Colors (2), change the color
palette and select Color Blind (3), click color
item in list at the left (current) and then click
color in list at right (4). Use dark gray for the
“black” thin bars in front; use light blue for the
“white” thicker bars in front.



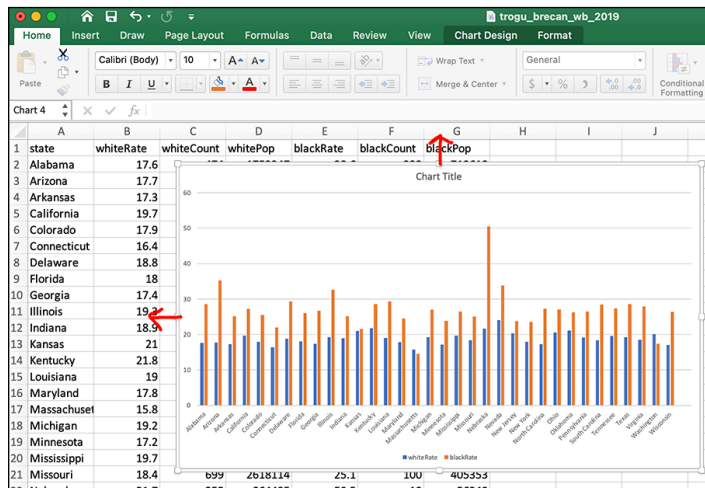
(6g) Final result. It can be improved further in
Illustrator by making the state labels at bottom
at a 45-degree angle and making all typography
black type, not the default Tableau gray. The
latter can be changed in Tableau too. Tableau
charts are interactive and mouse over will reveal
the specific data points.

Print and Save as PDF.

Graph 6: double bar chart. (Excel)



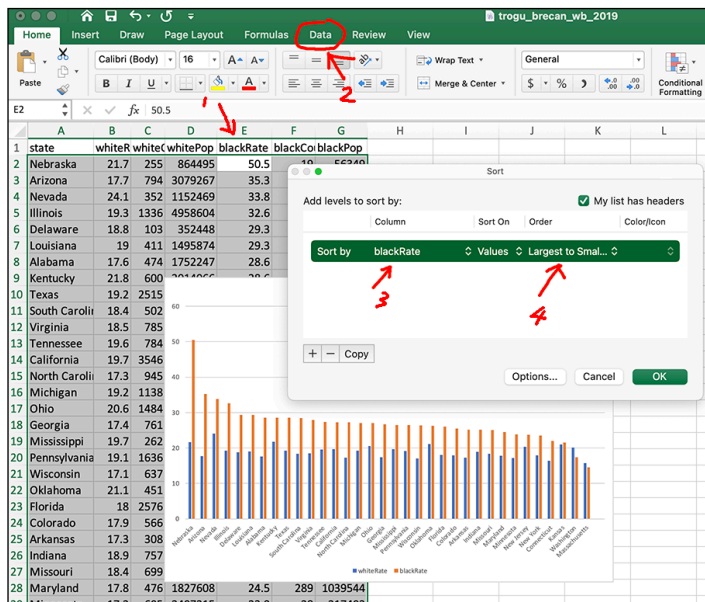
(6h) Open this file in Excel: lastName_breacn_wb_2019.csv
Select State, white rate, and black rate columns;
Click Insert —> Recommended charts —> Clustered column.
Similar to the double line graph, but with states in horizontal axis instead of years. While for years you should not change the order of the dates, here you will change the order of the states from sorted by alphabetical to sorted by black rate. Since it's a double chart, only the "black" set of bars will be sorted high-to-low. The other set (white) will just follow the order determined by the first sorting.



(6i) Expand the resulting chart so that all state names are visible

Note the current alphabetical sorting. This is a very bad default setting for the chart if the spreadsheet happens to be already sorted alphabetically!

Sort the chart in the next step.



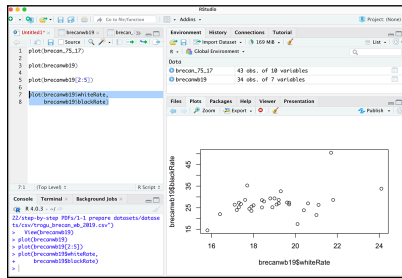
(6j) Select black rate. Click Data —> Sort. In the prompt (not shown) confirm Expand Selection (all cells will become selected).

In sorting pane, select black rate and largest to smallest.

After the chart is sorted, move it away from over the data cells and put it on top of blank cells. Select blank cells under chart and Print —> Selection —> Save as Adobe PDF.

Important: choose Save as Adobe PDF, not just Save as PDF. That way, the type will render properly and not scrambled when the file is opened in Illustrator.

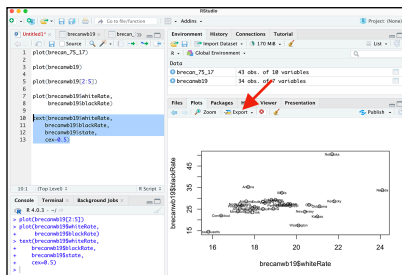
Graph 7. Scatterplot US states 2019 (RStudio)



(7a) In RStudio, you have already draw this plot as Scatterplot #3 in Assignment 1.1. The steps are repeated here. After importing the file: `lastName_brecan_wb_2019.csv` and shortening the name to `brecan2019`:

Type and run:

```
plot(brecanwb19$whiteRate,
     brecanwb19$blackRate)
```

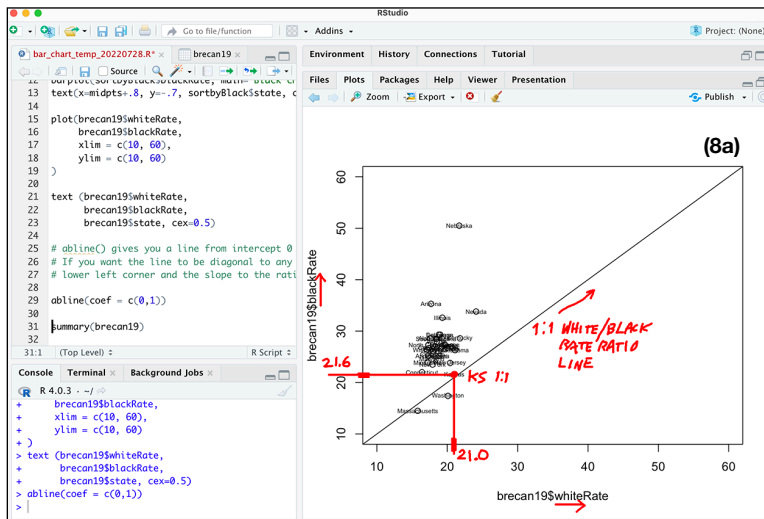


(7b) Then, add the names of the states on top of each data point. Type and run:

```
text(brecanwb19$whiteRate,
     brecanwb19$blackRate,
     brecanwb19$state,
     cex=0.5)
```

Save the plot (Export to PDF). Experiment with the proportions of the plot as you export it. Try for example a square, the nominal format of the next plot.

Graph 8. Scatterplot US states 2019 (same length axes – RStudio)



(8a) Repeat steps as in #7 but make the two X and Y axes the same length (from rate 10 to rate 60). Note that since the plot area in R is like a rubber sheet, you want to export this plot to a square format when saving the PDF. On screen it's not square, and it doesn't have to be, however the two axes will have the same data range length.

```
plot(brecan19$whiteRate,
     brecan19$blackRate,
     xlim = c(10, 60),
     ylim = c(10, 60))
```

```
text(brecan19$whiteRate,
     brecan19$blackRate,
     brecan19$state, cex=0.5)
```

```
abline(coef = c(0,1))
```

The last line of code (`abline`) adds a diagonal line to the graph that shows the spots where the rates between Black and White would be the same ratio (proportion), or 1:1 and therefore no disparity would exist between the groups. As you see most states above the line show a disparity.

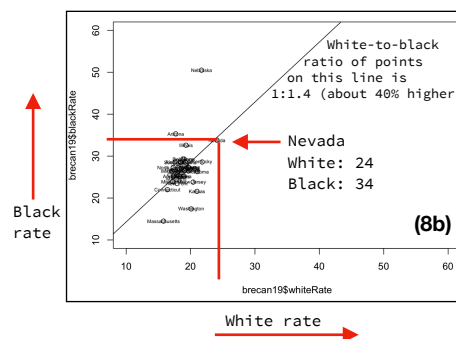
(8b) Changing the `abline` code to 0:1.5 moves the line nearer to where most states are (black rate is about 50% higher).

`abline()` gives you a line from intercept 0 with slope 1 in an existing plot. If you want the line to be diagonal to any plot just set the intercept to the lower left corner and the slope to the ratio of increase between the two axis.

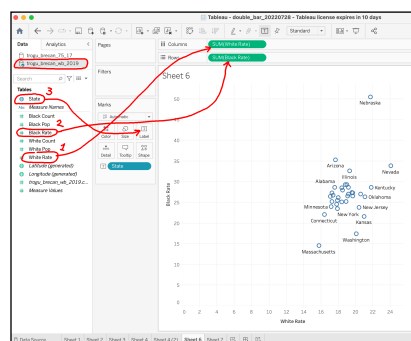
Change the last line of code to:

```
abline(coef = c(0,1.5))
```

to get a different sloped line →



Graph 9. Scatterplot US states 2019 (Tableau)



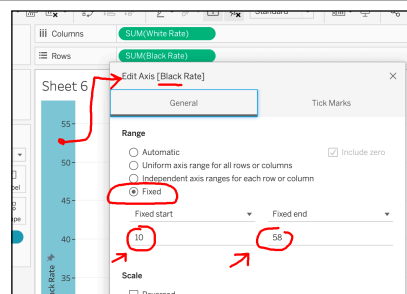
(9a) use again dataset
last_name_brecan_wb_2019.csv

After you connect to the text file, start a new sheet (bottom menu). Plot the two variables Black rate and White rate for 2019:

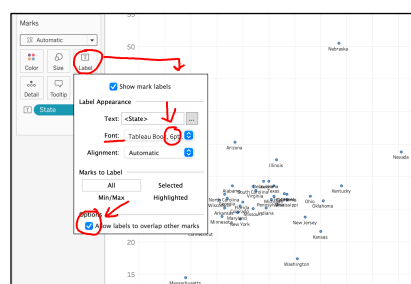
Drag Black to the rows fields

Drag White to the columns field

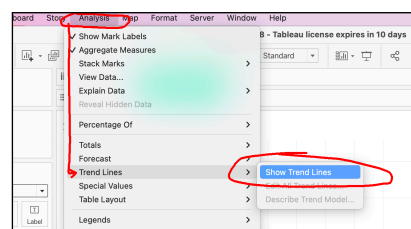
Drag States to Label in the Marks field (or drag to plot area).



(9b) Edit the axes, right-click left Black axis and make the range fixed: 10 to 54 (this is roughly the range for Black rate; Repeat the step for horizontal axis but set the range from 15 to 26 (not shown).



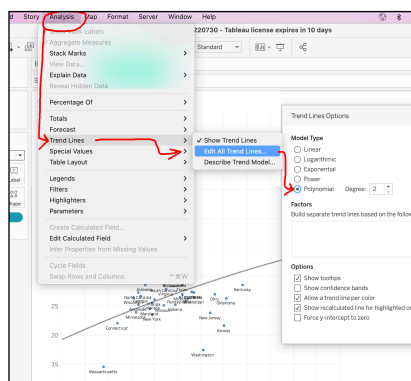
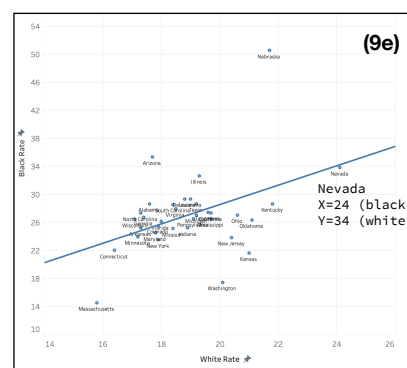
(9c) After setting the axes lengths, click Size (next to Label) to make the dots smaller. Then, make the state labels smaller: click Label, change the font size to 6 pt. and check box "Allow labels to overlap other marks". This displays all names regardless of whether they overlap each other. The labels' positions can be edited later in Illustrator



(9d) Add a "trend" line (similar to abline in R in #8):

Analysis —> Trend Lines —> Show trend lines. This will display a default straight (linear) trend line, shown at right (9e).

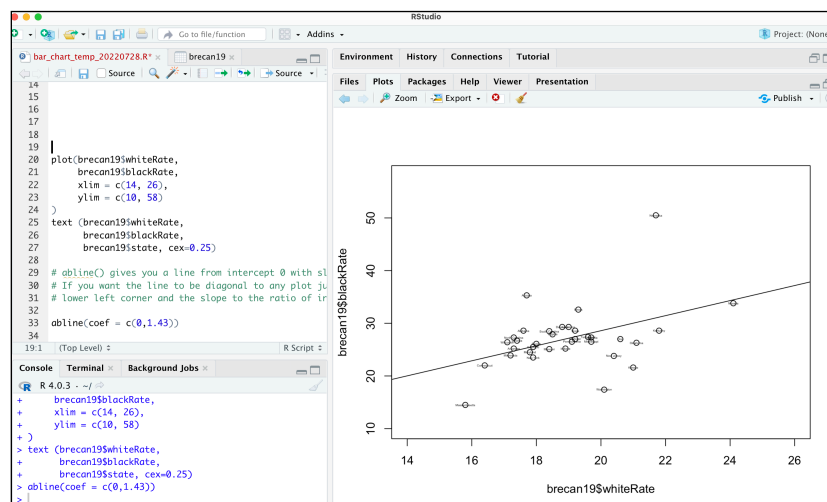
Make a note of the approximate coordinates (X,Y) of the beginning (14,20) and end of the line (26,36). If you divide Y/X (20/14 & 36/26) in both cases you get approximately 1.4 which is the simple ratio (on the line) between the White rate and the Black rate. That is, the Black rate is approximately 1.4 times the White rate (approximately 40% higher). You can see that Nevada, which has the highest rate for White, lies on the line and the values are 24 and 34. Divide 34 by 24 and the result is about 1.4. So if every state dot where on that line, every state's W:B ratio would be the same: 1:1.4.



(9f) Change the trend line from "linear" to "polynomial". The line will bend slightly upward but it's still about the same line. If you, unlike me, studied statistics, you probably understand the difference. It's a kind of average line because if you add up all the Black values from all the states and divide the total by the total for White values, the result is about the same: 1.4 or 1:1.4 (Black is 40% higher).

Save the plot. Print, save as PDF.

Some reflections...



This, just for comparison, is the same plot as # 9 but done again in R (was 8b), using again fixed axes but setting the range of White horizontal axis close to the actual values (smaller than for Black). The “abline” is also set at 1.43, the general average. Below is how the code was modified from #8b, by changing the values for x and y limit, the value for the size of the label type (0.25) but leaving the ratio of the line's slope unchanged (1.43).

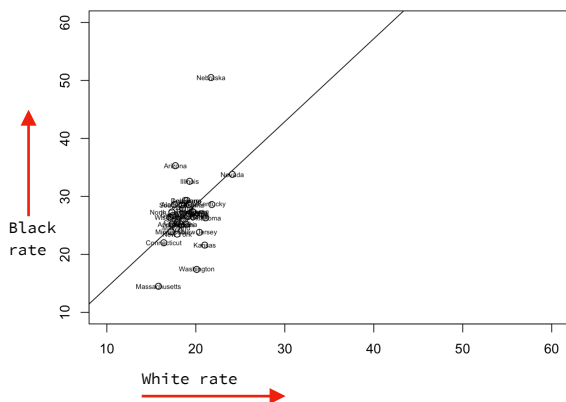
```

plot(brean19$whiteRate,
     brean19$blackRate,
     xlim = c(14, 26),
     ylim = c(10, 58)
)

text(brean19$whiteRate,
     brean19$blackRate,
     brean19$state, cex=0.25)

abline(coef = c(0,1.43))

```



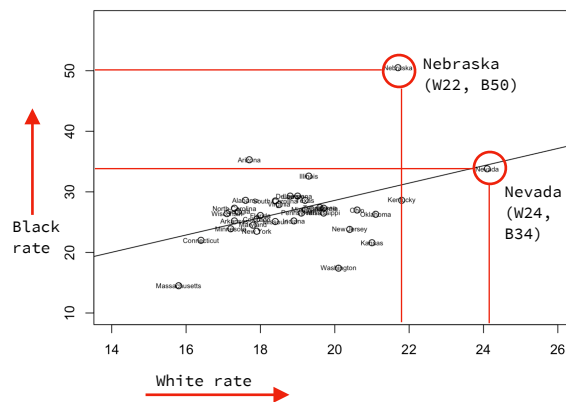
Compare these two versions of the scatterplot in R. Both show the same data and the same trend line (slope = 1.43).

Since the proportions (aspect ratio) of the box can be stretched in any direction, it's easy to see how one can make the same data look more or less extreme.

The data range length of the axes in the chart above left is the same (from 10 to 60) pushing the dots closer to the Black axis than to the White axis. This also puts the dots very close to each other.

In the chart on the right, by making the White axis range closer to the actual data (from 14 to 26), the dots become more legible.

Note however, that both boxes share the same distortion in that the axes themselves are not physically proportional to each other. In the chart at left, although both axes start at 10 and end at 60, the box is not a square as it would be if the divisions (the tic mark spaces) where themselves spaced by the same physical distance.



In the chart on the right, the true White “data range” from 14 to 26 (22 units) is actually physically longer (the length of the box) than the much longer Black “data range” which goes from 10 to 50 (40 units). Because the proportions changed, the angle of the trend line also changed, and the line does not look as dramatic as in the chart on the left.

So one should perhaps invert the axes and plot White on the vertical Y axis and Black on the horizontal X axis?

Try and see what happens.

No matter the shape of the box, some data points stick out (outliers), Nebraska for the highest Black rate (50) and Nevada for the highest White rate (24). Both are clearly noticeable in the chart on the right, however Nevada is not as noticeable in the chart on the left.