Race and Ethnicity in Covid-19 numbers: FL case study

E Cabral Balreira and Grace Stadnyk

Activity 2 - Manipulate Data. Let us determine if there is any relation between COVID-19 cases and race or ethnicity in Florida. We will look at the data from August 1, 2020, though this data is publicly available and so this activity can be updated to reflect more current data.

1. Open the data set in Google sheets - FL Covid-19 and Race Data - Aug 01
2. Save yourself a copy.
3. What is this data? Familiarize yourself with the data. You may want to address the following questions:
    a. Is any data obviously missing?
    b. What does each row tell you? What does each column tell you?
    c. Can you draw any conclusions immediately from looking at the spreadsheet?
    d. What kind of questions can you answer using these data points?
4. Your first goal is to compare the number of cases and the percent of the population that is Hispanic by county.
    a. Select Columns F and J by clicking on the column letter. You will need to hold command (on a Mac) or ctrl (on a PC) to select multiple columns at a time.
    b. Click on Insert > Chart.
    c. Analyze your scatterplot:
        i. Does there seem to be a correlation between these two data sets? If there is a correlation, is it positive or negative? Is it weak or strong?
        ii. Are there any outliers? How would the correlation change if you removed any outliers?

iii. If you think there is a correlation, add a trendline:

1. Double click on your graph.
2. In the Chart Editor panel on the right, click on the Customize tab.
3. Scroll down in this panel and click on the Series section.
4. Check the box for Trendline.
5. Check the box for Show R^2.

iv. What does the R^2 value suggest about the correlation between these two data sets? Does it support or refute your initial interpretation?

d. Based on a-c, can you conclude that counties with larger Hispanic populations can be expected to have a larger number of COVID-19 cases?

e. Does this chart and the data used to create it give you a fair picture on how COVID-19 cases relates to Hispanic populations? Hint: does it omit any potentially important factors?

5. On second thought, your chart from step 4 does not take into account the population in each county. You realize that you should instead be looking at cases per capita. Your second goal is thus to compare the number of cases per capita and the percent of the population that is Hispanic by county.

   a. Label column M CasesPer100k.

   b. Each row in column M should contain the number of Cases in the county divided by the Total Population multiplied by 100,000. For example, in cell M2, type **=100000*F2/H2**.

   c. Copy this formula down to row 68 by either clicking on the bottom right square in cell M2 and dragging directly down to row 68, or highlighting all cells in column M from row 2 to row 68 and then typing command+d (for Mac) or Ctrl+d (for PC). select Columns J and M. Click Insert > Chart.

d. Analyze your scatterplot:

i. Does there seem to be a correlation between these two data sets? If there is a correlation, is it positive or negative? Is it weak or strong?

ii. Are there any outliers? How would the correlation change if you removed any outliers? [Try this, if you'd like, by copying the spreadsheet to a new tab, removing the row(s) corresponding to any outlier(s) and creating a new chart.]

iii. If you think there is a correlation, add a trendline:

1. Double click on your graph.
2. In the Chart Editor panel on the right, click on the Customize tab.
3. Scroll down in this panel and click on the Series section.
4. Check the box for Trendline.
5. Check the box for Show R^2.

iv. What does the R^2 value suggest about the correlation between these two data sets? Does it support or refute your initial interpretation?

e. Based on a-c, can you conclude that counties with larger Hispanic populations can be expected to have a larger number of COVID-19 cases?

f. To what extent does your answer to this question based on Cases per 100k differ from your answer to this question when you were conducting your analysis based on the total number of cases?

6. Repeat step 5 for CasesPer100k and African American, that is columns K and M.

7. Your fourth goal is to see if there is a correlation between the number of cases of COVID-19 and the percentage of the population that is African American OR Hispanic, by county.

a. Label column N "African American or Hispanic"

b. Each row in column N should contain the sum of the corresponding numbers in columns J and K. Thus, in cell N2, type **=J2+K2**. Fill down this formula to row 68, as you did in step 5.

c. Create a new chart by highlighting columns M and N and clicking Insert > Chart. . Note: Make sure the x-axis is African American or Hispanic and Series is CasesPer100k (you can check this by double-clicking your graph and looking in the Chart Editor panel to the right).