Practice Assignment: Boxplot Data and Displays

Questions 1–3: These questions reference the *FiveThirtyEight* hate_crimes dataset,[1] which describes hate crime rates within different states. This dataset was also used in Practice Assignment 4.C.

The following table displays selected variables from the full dataset for the first 10 states (including DC):

state_abbrev	median_house_inc	hate_crimes_per_100k_splc	avg_hatecrimes_per_100k_fbi
AL	42278	0.125838926	1.806410489
AK	67629	0.143740118	1.656700109
AZ	49254	0.225319954	3.413927994
AR	44922	0.069060773	0.869208872
CA	60487	0.255805361	2.397985899
CO	60940	0.390523301	2.804688765
CT	70161	0.335392269	3.772701469
DE	57522	0.322754169	1.469979563
DC	68277	1.52230172	10.95347971

The following are descriptions of each variable in the table:

state_abbrev: State abbreviation

median_house_inc: Median household income, 2016

hate_crimes_per_100k_splc: Hate crimes per 100,000 population, Southern Poverty Law Center, Nov. 9–18, 2016

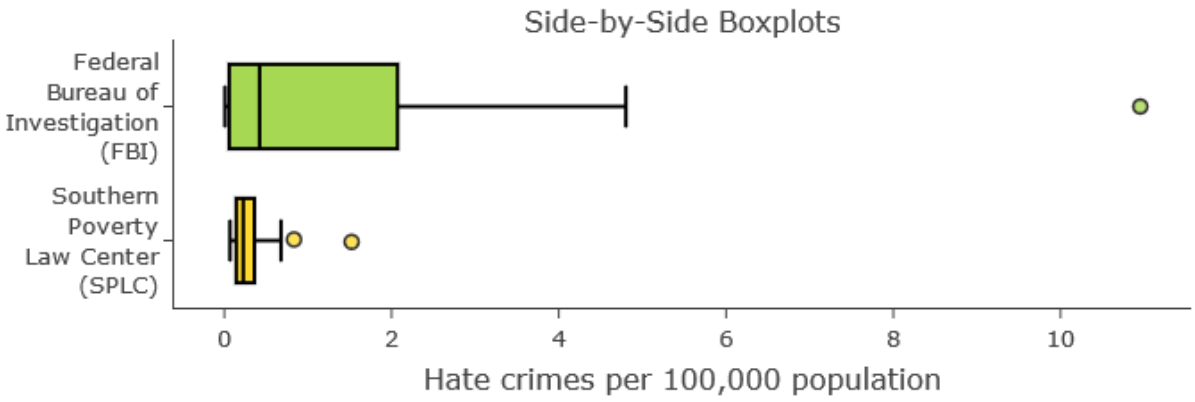
avg_hatecrimes_per_100k_fbi: Average annual hate crimes per 100,000 population, FBI, 2010–2015

1) Go to the *Describing and Exploring Quantitative Variables* tool at https://lumen-learning.shinyapps.io/eda_quantitative/. Select the **Several Groups** tab, locate the drop-down menu under "Dataset," and select "Hate Crimes (Source)."

   Construct parallel boxplots (boxplots displayed side-by-side) to visualize the hate crimes per 100,000 people as reported by the FBI and the hate crimes per 100,000 people as reported by the Southern Poverty Law Center (SPLC). Draw a well-labeled sketch of your visualization. Note: You can ignore states with missing data.

[1] *Hate crimes*. (2018, February 9). FiveThirtyEight. https://github.com/fivethirtyeight/data/tree/master/hate-crimes

Answer:



2) The FBI data describes the average annual hate crime rates from 2010–2015. The SPLC data describes the hate crime rate in 2016 between November 9–18.

   Part A: Compare the two distributions, using your visualization from Question 1.

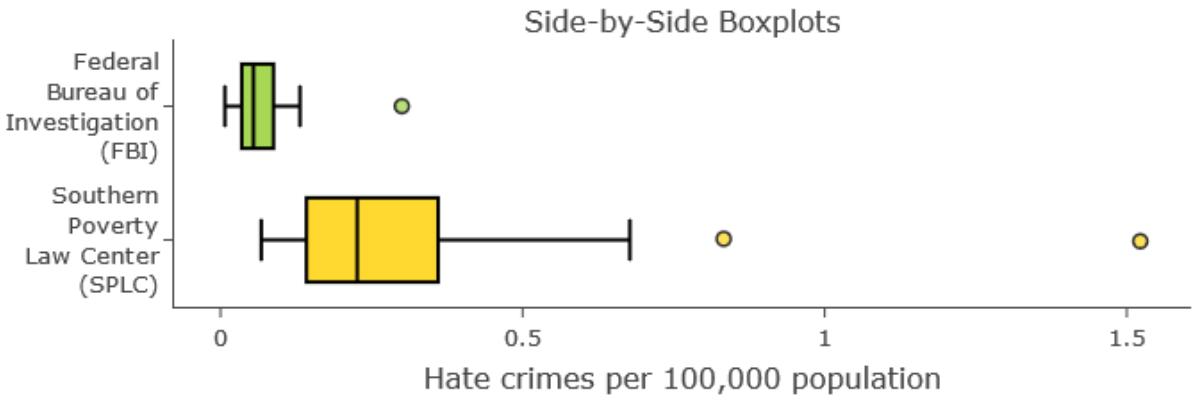
   Answers will vary.

   Sample answer: Overall, the hate crime rates are much lower in the SPLC dataset. For example, all states had hate crime rates lower than 2 per 100,000 people in the SPLC dataset; by contrast, about 25% of states had hate crime rates *higher* than 2 per 100,000 people in the FBI dataset. In addition, hate crime rates tend to vary more between states in the FBI dataset, as it has both a higher range and IQR.

   Part B: Why do you think these two distributions are so different? Explain.

   Answer: The FBI dataset describes *annual* hate crime rates, whereas the SPLC dataset describes rates only over a 10-day period. So, we'd expect to see generally higher hate crime rates in the FBI dataset, due to its longer time window.

3) To put these datasets on the same time scale, we divide the FBI hate crime rates by 36.5—this yields the average 10-day hate crime rates for the years 2010–2015. We visualize this transformed FBI data alongside the SPLC data as follows:



Part A: Comment on what effect our transformation of the FBI dataset had on the parallel boxplots of these distributions.

Answers will vary.

Sample answer: Now, the hate crime rates are much higher in the SPLC dataset. For example, 75% of states in the SPLC dataset (Q1 and above) had hate crime rates that were higher than all but one state in the FBI dataset. So, putting these distributions on the same time scale shows that the state-level hate crime rates were higher in the 10-day period described by the SPLC than in the average hate crime rates per 10-day period for the years 2010–2015 (according to the FBI).

Part B: The 10-day period covered by the SPLC data (Nov. 9–18, 2016) were the 10 days following Donald Trump's presidential election in 2016. Some claim that the election caused a spike in hate crimes. Does the above visualization provide enough evidence to <u>prove</u> this claim? Explain why or why not. If not, describe further information you would need to evaluate this claim.

Answers will vary.

Sample answer: The visualization above does not provide enough evidence to prove this claim. For example, if we had data on more specific 10-day periods from 2010–2015, we may find similar spikes in hate crimes during other election cycles or at random times (due to spurious variation). This evidence would challenge the idea that the spike in hate crimes was specific to and/or caused by Donald Trump's election.

4) In what situation would a dotplot tend to be more effective than a boxplot?

a) When we're working with a very small dataset

b) When the distribution is symmetric

c) When we're especially interested in the median

d) When we're working with a very large dataset

Answer: a

5) Determine whether this statement is true or false: Boxplots tell you the number of data points between the 3rd quartile and the maximum.

Answer: False