SI618 Project Proposal
Cara Canady
caraca@umich.edu
October 1, 2019

Overarching Research Question:
Does State Mental Health Agency (SMHA) funding affect suicide rates by state?

Summarize and motivate your proposed project:
Suicide is an increasingly concerning public health issue in the United States (Healy, 2019). Further, federal and state spending on public health issues has long been a battle. Although we understand that more resources often leads to improved outcomes, this widely-held belief deserves to be explored in the mental health context (Carey & Harris, 206; Lake & Turner, 2017).

Does state-sponsored mental health spending affect these rates? Do these rates affect spending? Does spending affect rates immediately, or is there some amount of lag? Or, is spending inconsequential on the effect of suicide rates? Does this dataset allow for exploration of the spectrum of suicidality (e.g. ideation, attempt, completion)? Ultimately, data such as this can motivate the need for increased prioritization in state and federal budgets.

I intend to explore this data state by state and year-to-year to understand the nationwide picture of the suicide epidemic. In the future, it may be possible to explore whether or not this data can be predictive in any way. Visualizations will also begin to answer some of these questions, as well.

Describe how you might manipulate and join these two datasets:
SMHA dataset from the Kaiser Family Foundation: https://www.kff.org/other/state-indicator/smha-expenditures-per-capita/?currentTimeframe=0&sortModel=%7B"colId":"Location","sort":"asc"%7D
This data can be exported in .csv, although it will require concatenation in order to develop a longitudinal dataset. Each .csv contains a state and its corresponding spending on mental health care.

Substance Abuse & Mental Health Data Archive (SAMHDA), National Study on Drug Use and Health (NSDUH):
https://datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2013-nsduh-2013-ds0001-nid13699
This data set can also be exported in .csv, but it is available in R as well. It will also require concatenation by year. It contains a number of categories that are referred to in an associated codebook (see pages i19 & i20 for data definitions). Because these categories are in code form, they may require transformation to description to develop logical categorical analysis.

I intend, using sparksql, to inner join these datasets on state and year and draw my conclusions and insights from the product of the join.

As for manipulation, the NSDUH data is large and will likely require some pairing down so that only relevant fields are present (i.e. excluding substance abuse data). In addition to the transforms and manipulations already mentioned, data should also be aggregated to show trends by state, year, and/or spending level.

Describe at least three map-reduce tasks you will perform to gain insights from the datasets (mrjob, spark, sparksql):
After joining these datasets using sparksql, I plan to manipulate, transform, and analyze my data using spark. Note that the NSDUH dataset deserves a good amount of exploratory analysis to fully understand it.

- **Map:** map will certainly be used to get the data into a usable RDD form and will also be used to export the data out into a form (perhaps a pandas dataframe) that can be used by the visualization package. Map can also be used as a preliminary counting step (x, 1) that will then be used by reduceByKey.
- **reduceByKey:** can aggregate by state, year, or even spending level bin to determine total or average spending by year by state.
- **sortBy:** this will reveal *top n* data that is ranked by spending or by rate.

Describe at least one visualization you might create that highlights the insights you hope to gain:
I think a paired bar-line graph with dual axes could show trending data for each metric. As a "stretch goal," I'd like to make this an interactive map where users can select state and/or year for analysis. I would like to use a python library to accomplish this, like ggplot or matplotlib. An example of chart type is included below.

Features of the visualization:

- Line for suicide rate, bar for spending
- Separate y-axis for each metric to indicate amount or frequency and
- Shared x-axis for year

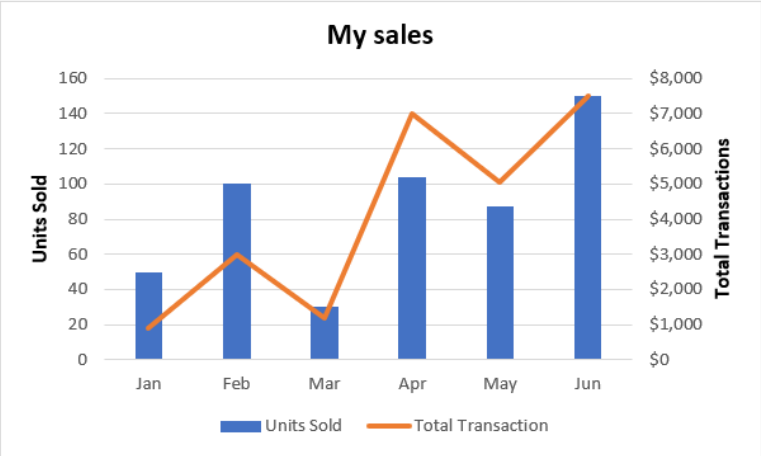

Image retrieved from https://www.microsoft.com/en-us/microsoft-365/blog/2012/06/21/combining-chart-types-adding-a-second-axis/

References:

Healy, Melissa. (2019, June 18). Suicide rates for U.S. teens and young adults are the highest on record. *Los Angeles Times.* Retrieved from https://www.latimes.com/science/la-sci-suicide-rates-rising-teens-young-adults-20190618-story.html.

Lake, J., & Turner, M. S. (2017). Urgent need for improved mental health care and a more collaborative model of care. *The Permanente Journal*, *21*.

Carey, Kevin & Harris, Elizabeth A. (2016, December 12). It Turns Out Spending More Probably Does Improve Education. *The New York Times.* Retrieved from https://www.nytimes.com/2016/12/12/nyregion/it-turns-out-spending-more-probably-does-improve-education.html