The Effect of State Mental Health Agency Funding on Suicide in the United States, 2007-2013

Cara Canady
October 22, 2019
SI 618, Project I

Motivation

Suicide is an increasingly concerning public health issue in the United States (Healy, 2019). Further, federal and state funding of public health issues has long been a battle. Although we understand that more resources often lead to improved outcomes, this widely-held belief deserves to be explored in the mental health context (Carey & Harris, 206; Lake & Turner, 2017). Ultimately, data such as this can motivate the need for increased prioritization in state and federal budgets.

Research Questions

- Does state-sponsored mental health funding affect suicide rates?
- Does funding affect suicide rates immediately, or is there some amount of lag?
- Do cuts to mental health funding lead to an increased suicide rate?

Note that research questions evolved from my original project proposal based on feedback from my Lab 6 partner.

Data Sources

To answer these research questions, data regarding funding and suicide rates needed to be obtained.

State Mental Health Agency (SMHA) funding data was retrieved from the Kaiser Family Foundation, and can be located here: https://www.kff.org/other/state-indicator/smha-expenditures-per-capita. According to their website, the Kaiser Family Foundation (KFF) "is a non-profit organization focusing on national health issues, as well as the U.S. role in global health policy." The KFF website also states that the organization strives to be non-partisan and provide health-related data free of charge (KFF, 2019).

This data was exported in .csv for all states by fiscal year, 2007-2013. Each file contained all 50 states including Puerto Rico and Washington DC, their corresponding per capita funding of mental health care ($), and the footnotes associated with each state. Data also included an aggregation for the entire United States. I manually added the fiscal year as a fourth column to the data, although in future iterations of this assignment and with more time and resources present, I would do that programmatically (see *Challenges, Limitations, and Lessons Learned* for further discussion). Therefore, each of the seven files contained 53 rows and 4 columns. Lastly, I made the assumption that all fields present in the .csv were strings; conversion to more appropriate data types can be found in the *Data Manipulation Methods* section. Column headers were not included on the data to allow for easier concatenation.

In the initial project proposal, I had suggested obtaining data regarding suicide rates from the Substance Abuse & Mental Health Data Archive (SAMHDA), National Study on Drug Use and Health (NSDUH). Upon further exploration of this data, I found it overly difficult to manipulate and the SAMHDA dataset insufficiently organized. Therefore, I instead used WISQARS™ (Web-based Injury Statistics Query and Reporting System) data from https://www.cdc.gov/injury/wisqars/index.html. This data originates from the National Center for Injury Prevention and Control, Centers for Disease Control and Prevention. Limitations for this data set can be found in the *Challenges, Limitations, and Lessons Learned* section of this report.

WISQARS™ data included breakdowns for sex, race, state, ethnicity, age group, cause of death, year, deaths (raw count), and population, and crude rate. I elected to export an entire .csv with cause of death as suicide injury for each state broken down by calendar years 2007 to 2013; no data subgroups were present aside from state and year. Analyzing the relationship between subgroup suicide rates and mental health funding represents an area of opportunity in future explorations, although the funding data will need to be equally granular to complete a thorough analysis. In total, 409 rows were exported in a single .csv file. Column headers were included in the data. Similar to the SMHA data from KFF, I treated each field as if it contained all strings.

Data Manipulation Methods

First, outside of the source code, I uploaded the individual KFF/mental health funding files to the Cavium node. Then, I concatenated the multiple files into one file named `state_funding_all.csv`.

Next, within the PySpark shell, I read the two csv files using the `spark.read.format` command. I put headers on the funding file, as upon reading the column names consisted of `_c0`, `_c1`, `_c2`, and `_c3`. I also renamed the column "Crude Rate" in the WISQARS™ data to "CrudeRate" to make manipulation in future steps easier. After structuring the files, I used the command `df.createOrReplaceTempView` to create temporary tables that would be easier for SparkSQL to use.

I joined the two temporary tables, named "funding" and "rates," using the `inner join` keyword in SparkSQL and joined on both state and year. Because the state mental health funding was reported by fiscal year and suicide rates were reported by calendar year, for simplicity I assumed parity between fiscal year and calendar year. More discussion about this limitation and its implications can be found in the *Challenges, Limitations, and Lessons Learned* section.

I initially intended to join the data using SparkSQL, create an RDD, and then manipulate, clean, and analyze the data using lambda functions with .map, .filter, .reduceByKey, etc. However, I realized I needed to change data types and reformat the data using SparkSQL in order to create analysis-ready columns; this step needed to happen before the conversion to an RDD. Therefore, I used the SparkSQL command `substring` to strip the `$` from the state mental health data. I also changed numerical data to integers and floats with the `cast` command.

After creating analysis-ready columns with SparkSQL (see *Analysis and Visualization* for more information), I converted each of my queries to RDDs. I then removed "N/A" and "NR" data present in the SMHA data using the following command:

```
.filter(lambda x: x[1] != 'N/A').filter(lambda x: x[1] != 'NR').map(lambda x:
(x[0],x[1],x[2],x[3],x[4],x[5],x[6]))
```

The overall workflow of code used for this project can be found in the figure below.

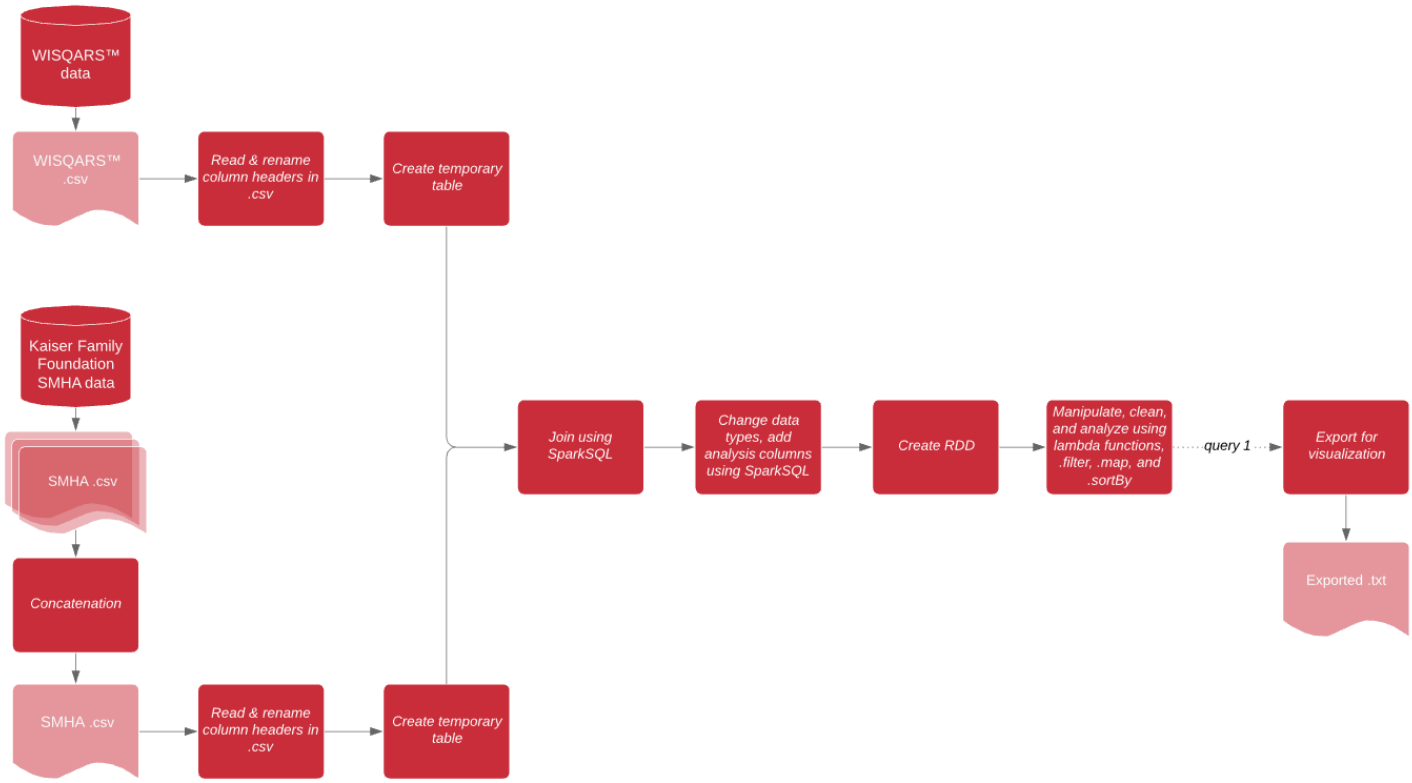


Figure 1: Source code workflow.

Analysis and Visualization

Both the source code analysis and the visualization help answer the three research questions initially posed.

Question 1: Does state-sponsored mental health funding affect suicide rates?

To answer this question, I went through several iterations of queries to properly join and format the datasets. As mentioned in the previous section, the currency sign needed to be stripped from the spending field, and the data types were changed so that they could be numerically manipulated. Further, I discovered that the reported "crude rate" was not reported as expected. One could easily conclude that a number reported as a "crude rate" is equivalent to a percentage (i.e. a number reported as 12.3 = 12.3% = 12.3×10^{-2}). In fact, though, the "crude rate" was actually a thousandth of this calculation (10^{-5}). Therefore, it was necessary to properly calculate this rate as an "adjusted rate."

The final query (`q1c`) also acted as the basis for Question 2 and Question 3. After confirming that the queries successfully returned the desired results, they were turned into temporary tables for later use. The commands are as follows:

```
q1 = sqlContext.sql('select funding.state, funding.spending, rates.Year, rates.Deaths,
rates.CrudeRate from funding inner join rates on rates.State = funding.state and
rates.Year = funding.fy')

q1b = sqlContext.sql('select funding.state, substring(funding.spending, 2) as
spending_trim, cast(rates.Year as int), rates.Deaths, cast(rates.CrudeRate as float)
from funding inner join rates on rates.State = funding.state and rates.Year =
funding.fy')

q1b.createOrReplaceTempView("funding_rate")

q1c = sqlContext.sql('select state, cast(spending_trim as float) as spending, year,
deaths, (cruderate/100000) as suicide_rate_adj from funding_rate')

q1c.createOrReplaceTempView("funding_rate_clean")
```

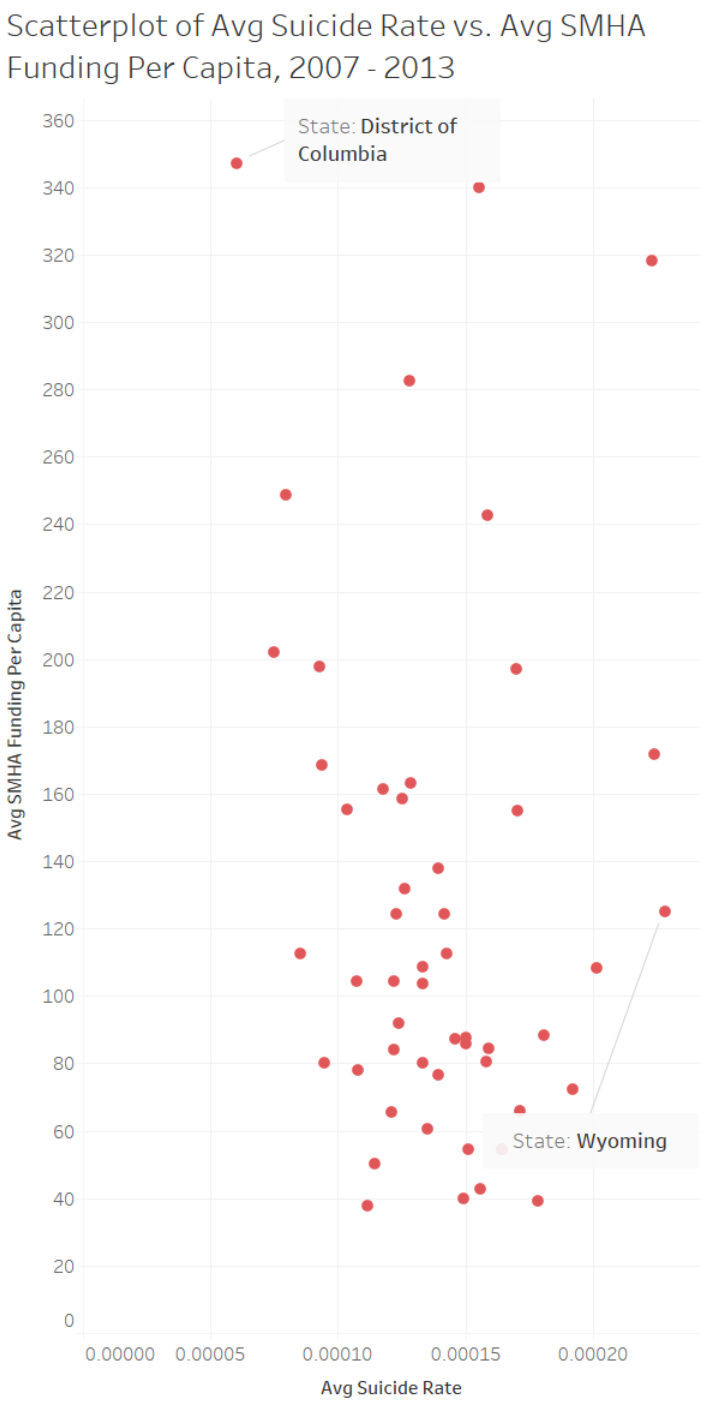
This final query, q1c, was then converted to an RDD, filtered to remove rows with missing or null data, and then analyzed using RDD functions. First, the RDD was mapped to two separate RDDs to include only the state and the spending per capita and the state and the adjusted suicide rate. The values were then mapped so they could be reduced and then manipulated using another map function to find the average suicide rate and spending per capita from 2007 – 2013. The two separate RDDs were then joined on state to reform the complete dataset with both spending per capita and suicide rate. The commands are as follows:

```
mapper_q1c_spend = filterer_q1c.map(lambda x: (x[0],x[1]))
mapper_q1c_rate = filterer_q1c.map(lambda x: (x[0],x[4]))

mapper_values_q1c_spend = mapper_q1c_spend.mapValues(lambda x: (x,1))
mapper_values_q1c_rate = mapper_q1c_rate.mapValues(lambda x: (x,1))
reducer_q1c_spend = mapper_values_q1c_spend.reduceByKey(lambda x,y:
(x[0]+y[0],x[1]+y[1]))
reducer_q1c_rate = mapper_values_q1c_rate.reduceByKey(lambda x,y:
(x[0]+y[0],x[1]+y[1]))
average_q1c_spend = reducer_q1c_spend.map(lambda x: (x[0],x[1][0]/x[1][1]))
average_q1c_rate = reducer_q1c_rate.map(lambda x: (x[0],x[1][0]/x[1][1]))
```

```
q1c_join = average_q1c_rate.join(average_q1c_spend).map(lambda x:
(x[0],x[1][0],x[1][1]))
```

Finally, this data was exported as a .txt so that it could be visualized.



State	Avg SMHA Funding Per Capita ($)	Avg Suicide Rate
Alabama	76.7	1.39e-04
Alaska	318.1	2.23e-04
Arizona	196.9	1.70e-04
Arkansas	42.6	1.56e-04
California	155.4	1.03e-04
Colorado	88.3	1.81e-04
Connecti..	197.6	9.29e-05
Delaware	104.4	1.22e-04
District o..	347.1	6.04e-05
Florida	39.8	1.49e-04
Georgia	50.1	1.15e-04
Hawaii	163.2	1.29e-04
Idaho	39.1	1.78e-04
Illinois	80.2	9.47e-05
Indiana	80.2	1.33e-04
Iowa	131.9	1.26e-04
Kansas	124.3	1.42e-04
Kentucky	54.6	1.51e-04
Louisiana	65.6	1.21e-04
Maine	339.9	1.55e-04
Maryland	168.6	9.35e-05
Massach..	112.4	8.51e-05
Michigan	124.4	1.23e-04
Minneso..	161.5	1.18e-04
Mississi..	103.6	1.33e-04
Missouri	87.4	1.46e-04
Montana	171.9	2.24e-04
Nebraska	77.9	1.08e-04
Nevada	72.3	1.92e-04
New Ha..	138.0	1.39e-04
New Jers..	202.0	7.49e-05
New Mex..	108.4	2.02e-04
New York	248.7	7.96e-05
North Ca..	158.5	1.25e-04
North Da..	85.7	1.50e-04
Ohio	84.0	1.22e-04
Oklahoma	54.6	1.64e-04
Oregon	155.0	1.70e-04
Pennsylv..	282.6	1.28e-04
Rhode Isl..	104.5	1.07e-04
South Ca..	60.4	1.35e-04
South Da..	84.3	1.59e-04
Tenness..	87.6	1.50e-04
Texas	37.8	1.11e-04
Utah	65.8	1.71e-04
Vermont	242.5	1.59e-04
Virginia	91.8	1.24e-04
Washing..	112.7	1.43e-04
West Vir..	80.5	1.58e-04
Wisconsin	108.7	1.33e-04
Wyoming	125.1	2.28e-04

Figure 2: Scatterplot of average suicide rate versus average SMHA funding per capita

Tableau was used to visualize the data. Figure 2 reveals that higher spending doesn't necessarily equate to a lower suicide rate, as several states represented at the top of the plot (higher spending) have varying average suicide rates. States in the West (Wyoming, Montana, Arizona, New Mexico) lie on the right/bottom-right side of the chart, meaning that they have higher suicide rates and less funding per capita.

Question 2: Does funding affect suicide rates immediately, or is there some amount of lag?
This question required a complicated SQL self-join in order to find the rate of change from one year to another. It also took me a fair amount of time to figure out how to refer to the fields so that "year1" represented one year and "year2" represented the subsequent year. I subtracted the initial value from the final value and divided the outcome by the initial value to calculate the rate of change. The commands are as follows:

```
q2 = sqlContext.sql('select a.state, a.year as year1, (b.year) as year2, a.spending as
spending_year1, b.spending as spending_year2, (b.spending - a.spending)/(a.spending)
as spend_change, a.suicide_rate_adj as suicide_rate_adj_year1, b.suicide_rate_adj as
suicide_rate_adj_year2, ((b.suicide_rate_adj -
a.suicide_rate_adj)/(a.suicide_rate_adj)) as suicide_rate_change from
funding_rate_clean a inner join funding_rate_clean b on (a.year+1 = b.year) and
a.state = b.state order by a.state, a.year')

q2.createOrReplaceTempView("funding_rate_calc")
```

Similar to the format of Question 1, I created an RDD, cleaned the data using the filter command mentioned in the *Data Manipulation Methods* section, and mapped the values for potential future use. To analyze this data, I sorted it using the sortBy RDD method. I sorted the data twice: once to find the states with the highest suicide rates and again to find the states with the greatest change in funding to examine its effect on suicide rate. The source code is below.

```
mapper_q2 = filterer_q2.map(lambda x: (x[0],x[1],x[2],x[3],x[4],x[5],x[6],1))
sorter_q2rate = filterer_q2.sortBy(lambda x: x[6],ascending=False)
sorter_q2change = filterer_q2.sortBy(lambda x: x[5], ascending=False).sortBy(lambda x:
x[0], ascending=False)
sorter_q2change.take(6)
```

Results are below.

State	Year 1	Year 2	SMHA Funding Year 1	Funding Year 2	Funding Change	Adjusted Suicide Rate Year 1	Adjusted Suicide Rate Year 2	Adjusted Suicide Rate Change
Mississippi	2012	2013	106.6100006	55.95000076	-0.475189941	0.0001374	0.0001299	-0.054585154
North Carolina	2012	2013	134.7799988	97.08000183	-0.279715071	0.0001318	0.0001304	-0.010622181
Hawaii	2010	2011	169.9900055	126.6299973	-0.255073868	0.0001522	0.0001313	-0.137319324
Nevada	2008	2009	81.37999725	64	-0.21356596	0.000199	0.0001881	-0.054773878
Tennessee	2008	2009	98.12000275	78.30999756	-0.201895685	0.0001557	0.0001502	-0.035324293
Hawaii	2009	2010	212.1499939	169.9900055	-0.198727267	0.0001299	0.0001522	0.171670557
Idaho	2009	2010	44	36.63999939	-0.167272741	0.0001956	0.000185	-0.054192203
Michigan	2009	2010	142.8399963	119.2300034	-0.16528979	0.0001181	0.0001278	0.082133724
Rhode Island	2009	2010	107.1900024	90.51000214	-0.15561153	0.000112	0.0001225	0.093750019
Louisiana	2012	2013	65.51000214	55.5	-0.152801127	0.0001232	0.000126	0.022727329

Figure 3: States with the greatest funding change

Feedback from Lab 6 suggested that I look closely at a state to examine how year-on-year changes affected suicide rates. I used Wyoming, due to its high suicide rate.

State	Year 1	Year 2	SMHA Funding Year 1	Funding Year 2	Funding Change	Adjusted Suicide Rate Year 1	Adjusted Suicide Rate Year 2	Adjusted Suicide Rate Change
Wyoming	2007	2008	99.55000305	142.4600067	0.431039702	0.0001888	0.0002271	0.202860174
Wyoming	2008	2009	142.4600067	154.6499939	0.085567785	0.0002271	0.0001983	-0.126816349
Wyoming	2009	2010	154.6499939	133.2400055	-0.138441573	0.0001983	0.0002324	0.171961667
Wyoming	2010	2011	133.2400055	115.8499985	-0.130516409	0.0002324	0.0002326	0.000860605
Wyoming	2011	2012	115.8499985	111.4800034	-0.03772115	0.0002326	0.0002966	0.275150454
Wyoming	2012	2013	111.4800034	118.8000031	0.065661997	0.0002966	0.0002215	-0.253202976

Figure 4: A closer look at Wyoming

These results are surprising, as they indicate that changes to funding do not necessarily affect suicide rates immediately. Figure 3 shows that decreases to funding do not have an overall positive effect on adjusted suicide rate (at least in the top 10 records).

There may be a lag in effect. As observed in Wyoming, a 43% funding increase does not produce a reduction in suicide rate until the year after. Still, this reduction is not sustained.

More investigation into this relationship is recommended, especially in Western states suicide is more prevalent.

Question 3: Do cuts to mental health funding lead to an increased suicide rate?

To comprehensively answer the third question, it seemed prudent to look at the question from two different perspectives. First, I looked at how often the phenomenon of decreased funding and increased suicide rate occurred in the same year. I also looked at instances where funding increased and suicide rate decreased. The query used for execution is listed below:

```
q3a = sqlContext.sql('select state, year1, year2, spend_change, suicide_rate_change
from funding_rate_calc where (spend_change < 0 and suicide_rate_change > 0)')
q3b = sqlContext.sql('select state, year1, year2, spend_change, suicide_rate_change
from funding_rate_calc where (spend_change > 0 and suicide_rate_change < 0)')
```

After converting these queries to RDDs, I counted the rows in each query and divided them by the row count of q2. This result gave me the instances of occurrences instead of individual states. Although out of scope for this analysis, one could examine this on a state level to examine the frequency of this phenomenon in individual states. The code used for this analysis is found below:

```
dec_spend_inc_suicide = float(q3a_rdd.count())/(q2_rdd.count())
inc_spend_dec_suicide = float(q3b_rdd.count())/(q2_rdd.count())
```

Calculation	% of Records
Decreasing spending, increasing suicide	24.83
Increasing spending, decreasing suicide	19.93

Figure 5: Frequency of phenomena occurring in the dataset

Nearly 25% of records contain the phenomena of decreased spending and increased suicide; nearly 20% of records contain an increase in spending in the same year that suicide rates decrease. These phenomena should be further investigated.

In summary, I attempted to get the entirety of my code to work as expected and to run. I fortunately did not encounter many insurmountable barriers. However, challenges, limitations, and lessons learned can be found in the following section.

Challenges, Limitations, and Lessons Learned

Challenges

As mentioned earlier, the "crude rate" present in the WISQARS™ dataset wasn't representative of the traditional sense of a "crude rate." I did not realize this until my project was nearly complete; however, this forced me to go through the entirety of my code to ensure that values reported and calculated were reading as expected.

Additionally, when attempting to strip the currency sign from the SMHA data, I found that the RIGHT and LEFT functions (which I typically use for SQL trimming) were not available. I was forced to use another method, which was time consuming and frustrating. Finally, I landed on substring; this was successful in producing the desired results.

Data types also presented a challenge. Because so many numerical fields were present in my data sets, each needed to be converted to a float or integer data type to allow for manipulation. Missing data was also reported differently throughout datasets, so my results needed to be carefully checked to ensure that these missing records were adequately removed.

Limitations

One of the biggest assumptions I made in this project was equating fiscal year with calendar year. As the SMHA data was reported on a fiscal year basis and the WISQARS™ data was reported by calendar year, the months of each year do not line up. However, because this is a transparent assumption made throughout the project, it seemed reasonable.

The absence of large numbers also presents a limitation. In some instances, the number of deaths due to suicide injury is small, which results in an even smaller suicide rate per capita. However, the smallest number of deaths is 29, which does not seem problematically small. Another limitation is my knowledge of statistics; in future iterations of this project, I would like to definitively qualify these small numbers as problematic or not. The absence of true statistical analysis prevents my work from being defensible, so an opportunity presents itself for richer analysis in the future.

Further, the absence of rigorous statistics and other variables in the dataset present the possibility of a spurious correlation. Perhaps factors outside of state mental health agency funding accounted for a rise or fall in suicide rates, such as unemployment or substance use. These other factors represent the potential for confounding variables.

Lessons Learned

With more time and resources, instead of manually adding a column for fiscal year, I would programmatically add this column; since I did this initially back in late September, I did not realize this opportunity until the writing of this report. Further, there are several opportunities for richer analysis. I look forward to continuing to work on and think about this project throughout this course.

This project helped me become much better at error detection and interpretation of Spark error messages. From data type errors to incorrectly referring to an element in an RDD, I did not get discouraged by the verbose error messages; I was able to comb through them to decipher exactly where I made my mistake.

Conclusion

Money is not a panacea, but it does provide resources and infrastructure to tackle large, systemic problems. Given the gravity of suicide, the relationship between state mental health agency funding and suicide rates should be further explored, especially in the Western United States.

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.

Kaiser Family Foundation. "About Us," *Kaiser Family Foundation*. https://www.kff.org/about-us/. Retrieved 2019, October 1.

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 Funding More Probably Does Improve Education. *The New York Times.* Retrieved from https://www.nytimes.com/2016/12/12/nyregion/it-turns-out-funding-more-probably-does-improve-education.html