Proposal Proposal

(20points) Summarize and motivate your proposed project.

This project is primarily focused on the gun violence incidence in the United States. Gun-related crimes have always been an issue in this country, and for recent years, most incidence of gun violence would rise a debate regarding whether the right to keep guns should still be kept. Therefore, I'm interested in finding out which counties have the greatest gun violence incidence, and examining what factors that are likely to be associated with it. In this project, unemployment rate and household income are the two factors that will be examined.

(20points) Choose and describe(at least) two different datasets.

1. The first dataset I will use is the gun violence data, which contains information regarding gun violence incidents from 2013-2018. The data is retrieved from Kaggle on https://www.kaggle.com/jameslko/gun-violence-data.
2. The second dataset will be used is the U.S. Unemployment Rate by County: 1990-2016, contributed by the U.S. Department of Labor's Bureau of Labor Statistics. The data is retrieved from Kaggle on https://www.kaggle.com/jayrav13/unemployment-by-county-us.
3. The last dataset is the U.S. Household Income Statistics, which provides the mean, median and standard deviation of household income on a neighborhood scale. The data is retrieved from Kaggle on https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations.

(20points) Describe how you might manipulate and join the two datasets.

Before any further manipulation, I will remove duplicate rows and rows with missing value. Since all three datasets contain county information, and I will join them by this column.

(30points) Describe at least three map-reduce tasks you will perform to gain insights from the datasets (you can use mrjob, spark or sparksql)

1. With MrJob, I will use mapper, combiner, and reducer to count the number of gun incidence within each county, and sort the results from the highest to the lowest. Such task will allow me to find out which county has the greatest gun violence occurrence.
2. With Sparksql, I will map the unemployment rate of each county vs. the gun violence occurrence of the county. I will order the result from

the county with highest gun violence incidence to the lowest to see if gun violence is correlated with unemployment rate.

3. With Sparksql, I will map the average household income of each county vs. the gun violence occurrence of the county. I will order the result from the county with highest gun violence incidence to the lowest to see if household income is correlated with unemployment rate.

(10points) Describe at least one visualization you might create that highlights insights you hope to gain

I will use scatterplot to visualize the relationship between gun violence incidence and unemployment rate. Another scatterplot will be applied to visualize the relationship between gun violence and household income.