Krittika D'Silva and Hudson Poissant
CSE 140, Homework 9

Analyzing Rates of Criminal Activity in America from 1960 to 1999 Relative to the Population Growth

Overview of research questions:

1. *Which states have experienced the largest increase in crime rate relative to their state population change per year?*

The sates with the largest increase in crime rate relative to their state population change per year were:

1. Ohio
2. Michigan
3. South Dakota
4. Wyoming
5. Illinois

These were determined based off of the largest mean squared errors. We noticed that many of these states were very large. This led us to believe that larger states have large increases in the number of crimes that occur in their population. It is interesting however, to note that although Ohio and Michigan are among the larger states in America, the top 5 largest (California, Texas, New York, Florida and Illinois) do not appear on our top 5 list for ratios.

2. *Relative to other states, how quickly (or slowly) has Washington's crime rate increased with respect to its population?*

Washington's "ratio" varied a lot relative to the average. Although we were hoping to pull out meaningful results from our analysis, ultimately, we noticed that we could not pull out a clear conclusion from our data because our results were so varied.

What we are trying to compute and why:

1. We are computing each state's "ratio". A ratio is the change in the number of total criminal incidents from one year to the next divided by its change in population during the same time interval. By examining how this ratio changes yearly from 1960 to 1999 we hope to develop an understanding about how the relative number of crimes has changed from one state compared to another.
2. For this question we are looking at the state of Washington specifically, these calculations are done to highlight where Washington's "ratio" ranks among other states. No additional computations will need to be done to answer this question since we will be using a similar analysis to the one performed in question 1.

Motivation and background:

Often, it is often generalized that as a city becomes increasingly populated and metropolitan, this can correlate with the crime rates in that area. The purpose of our analysis is to extend this idea

to the state level over the course of numerous years. The population of individual states in America ranges from a little over half a million to more than 35 million. Therefore, to avoid giving large states a disadvantage, we will calculate each state's respective "ratio". This number gives us an understanding about the relative rates of change from one state to another. Based on the analysis of our results, if we notice that there were a select few states with low ratios, this gives us an indication that those states have maintained relatively low crime rates despite a growing population. Perhaps these states have effective laws or crime enforcement officers in place. Conversely, states with high ratios indicate that perhaps these states should take steps towards developing increase crime preventative measures.

Dataset:
Our data sets for population were taken from the US Census Bureau website.
1960s to 1970s Population by State
1970s to 1980s Population by State
1980s to 1990s Population by State

The data for the crime incidents per state was taken from:
http://hci.stanford.edu/jheer/workshop/data/

The data for the populations was taken from the US Census Bureau website. An individual file has data for one decade. They include the population totals for the United States as a whole as well as population totals for each state. The format of each file includes two parts with seven columns: the first column has the state the other columns include the yearly totals. The first part includes the first part of the decade and the second part includes the second half of the decade.

The data for crime statistics was taken from a Stanford website. The file is a long list of of crime counts from 1960 until 1999 and includes the state, type of crime, crime, year, and count of the number of crimes committed in that year.

The actual data files that we used can be found below:
https://www.dropbox.com/sh/jp6jdwrxwpd6lek/gzC_7qQf4k

Methodology (algorithm or analysis):
First you will need to compile the information from the files in a usable format. You will need to get the years, states, populations, and total crime for each state.

From the crime data, determine the total number of crimes each year in each state.

Each year, should map to all of the states. Each state (for an individual year) should map to its respective aggregate crimes for that year.
For example:
{ 1960 : { "Alabama" : 1234, "Alaska" : 4567, … }, 1961 : {"Alabama" : 1456, "Alaska" : 4789, …}, 1962 : {"Alabama" : 1567, "Alaska" : 4999, …}, … }

For each year, not including the first year, calculate the change in the total number of crime incidents that occur from one year to the next. The structure of the result should look like this:
For example:
{1961 : {"Alabama" : 222, "Alaska" : 222, …}, 1962 : {"Alabama" : 111, "Alaska" : 210, …}, … }
Note, the first year is not used. Also note that these are not correct values.

Using the data of the population of each state from 1960 to 1999, calculate the population of each state in each year. The format of this data structure should be similar to the ones above.
For example:
{ 1960 : { "Alabama" : 33333, "Alaska" : 22222, … }, 1961 : {"Alabama" : 37999, "Alaska" : 31599, …}, 1962 : {"Alabama" : 45678, "Alaska" : 40000, …}, … }

Similar to the calculations done for the change in total number of crime incidents from one year to the next, calculate the change in population of each state at intervals of 1 year.
For example:
{1961 : {"Alabama" : 466, "Alaska" : 9377, …}, 1962 : {"Alabama" : 7679, "Alaska" : 8401, …}, … }
Note, the first year is not used.

Using the data already calculated up to this point, calculate the ratio for each state for each year. The ratio is defined as the change in crime from one year to the next divided by the change in population from one year to the next. In other words, you must calculate:
(Crime of year2- crime of year1) / (population of year 2 – population of year 1)

If we were to use the example above, the calculated ratios would be:
{1961 : {"Alabama" : 0.47639, "Alaska" : 0.02367, …}, 1962 : {"Alabama" : 0.01445, "Alaska" : 0.02499, …}, … }

Now you will need to calculate the root mean squared (RMS) for each state. The RMS is the squared difference of the state's ratio minus the average ratio. Note that the average ratio will also need to be calculated for every year.

Using the sum of the RMS for each state, determine the top 5 and bottom 5 states.

A plot will be generated that plots each ratio between 1961 and 1999 for each of the top 5 and bottom 5 states.

Similarly, a plot will be created on with the ranking of the state of Washington is shown relative to other states. The other states on the plot will be determined by the results that are produced. For example, if Washington has an extremely high ratio, it will be plotted against states with a range of ratios (high to low) to portray its relative location.

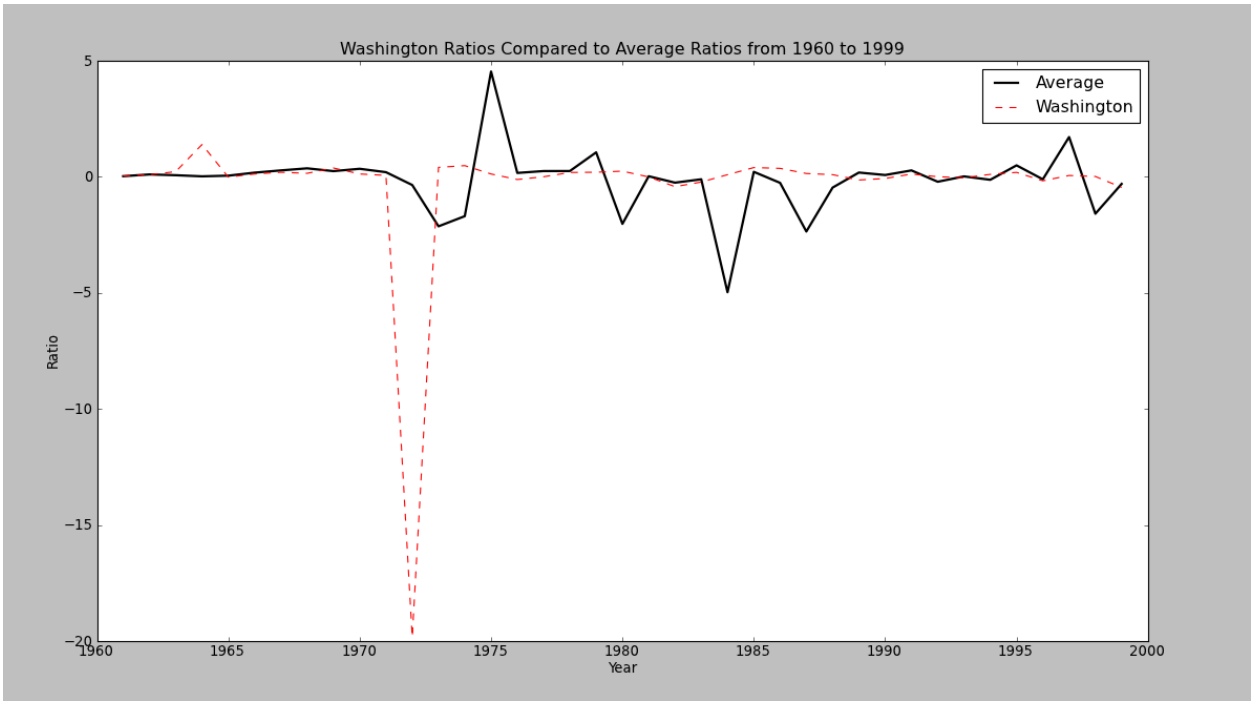
Results:

Question 1: Which states have experienced the largest increase in crime rate relative to their state population change per year?

State	Ratio
Ohio	50216.023057180995
Michigan	32277.496914618423
South Dakota	20161.630937014757
Wyoming	16113.493238791709
Illinois	10636.322399761017

These 5 states had the largest total mean squared error from the time period of 1961 to 1999. Many of these states we very large, crime in a state overall does therefore depend on its total population. We found it difficult to draw assertive statements out of our results, however, because they did vary a lot. For example, Wyoming is the state with the lowest population in America while Ohio has the 7th largest population – they both however, appear to have very high ratios.

Question 2: Relative to other states, how quickly (or slowly) has Washington's crime rate increased with respect to its population?



The average ratio from 1960 to 1999 was relatively constant – it almost always remained between 5 and -5. For Washington, however, we noticed that the ratio varied quite a bit- especially during the 1970s. With time, however, the ratio began to level out; this may lead us to believe that crime preventative measures were successfully implements over the years.

Reproducing your results:

1. From the DropBox, download all files and place them in the same directory as the python script.
2. It is not necessary to run this code from the command line and should be instead run using Idle. No addition work must be done, the figures will appear when the script is run – do note, these plots are not saved after they are closed.

Collaboration:

Hudson and Krittika were partners and worked jointly on this project.

Reflection:

For part I of this assignment, although we found numerous datasets online, we had a really hard time deciding on one that we strongly felt would allow us to conduct an interesting and meaningful analysis. Once we finally found something we thought would be interesting, the rest of the assignment fell into place. There were a few portions of the spec that were unclear, the one that we found the most hindering was the 'Methodology' section since we were unsure of where to draw the line on what to include and what not to include.

After looking over the comments for part I of this assignment, we realized that we would need to make a lot of changes for part II. Our third question in the research questions needed to be changed because we decided analysis on the level of each crime would be too demanding for the time allotted to use. We decided to sum the different crimes and use that for each state for each year. We were able to manage our time and communicate well. We met up every day to work on the assignment for several hours at a time. Coding was difficult because there was one file shared between both of us. Our solution, although probably not the best but it got the job done, was to send sections of code over Facebook. It would have been nice to have a late day for this project due to the huge time commitment. This was a difficult project overall but we definitely learned a lot.