Evan Batzer

HW1

STA 141C

1/14/2019

**1- Computation**

1. Which agencies have the highest median annual spending?

Large agencies, particularly those related to the military and defense, have the highest median annual spending rates. The Defense Security Cooperation Agency, in particular, appears to spend over 14 billion dollars each year on these sorts of expenditures

agency mspend

*<fct>* *<dbl>*

1 Defense Security Cooperation Agency 14495443353.

2 Headquarters, NASA 8107436839.

3 Department of State 4326045181.

4 Defense Human Resources Activity 3668429160.

5 U.S. Special Operations Command 3530342607.

6 Immediate Office of the Secretary of the Navy 3483115983.

1. Qualitatively describe the distribution of median annual spending.

Median annual spending is highly skewed. Many agencies appear to spend relatively few funds in a given fiscal year, while a small number spend very large amounts. Some agencies appear to have positive income according to this dataset – the Air Force Frequency Management Center has a net median spending rate of -$10697 / year, perhaps due to documented payment requests rather than expenditure.

Moreover, quartiles within this distribution reflect its heavy right skew: 1st quartile – $100,000 / year, 2nd quartile – $1,000,000 / year, 3rd quartile -- $12,250,000. A histogram of this distribution is difficult to interpret (Fig 1. Panel 1), as many values are condensed within the first bin.

A screenshot of a cell phone

Description automatically generated

Figure 1. Distribution of median spending per annum (panel 1) and log-transformed median spending per annum (panel 2).

1. Qualitatively describe the distribution of the logarithm of the median annual spending. Plot the histogram.

When log-transformed, our previously highly skewed distribution becomes increasingly normal. While this transformation still produces a distribution with a slight right skew, we have reduced the influence of some of the highest spending agencies and improved the visualization of this distribution in our histogram.

1. Is there a clear separation between agencies that spend a large amount of money, and those which spend less money?

There is perhaps no clear dividing line between agencies that spend a large amount of money and those that do not—this distribution is not bimodal, for instance—but there do appear to be qualitative differences between agencies at different ends of our spending spectrum. Large institutions such as the state department and defense-related departments have very large spending budgets in this datasets, while a smaller and more diverse set of organizations and committees spend much less.

**2 – Reflecting**

1. Qualitatively describe the distribution of the file sizes.

A screenshot of a cell phone

Description automatically generatedLike spending totals, file lengths are highly skewed. This heavily right-tailed distribution likely reflects differences in the total size of the budget and number of transactions within a given agency. Many files are fewer than 1Mb in size unpacked, while 2 files (“0.csv” and “1219.csv”) are well over 1Gb.

1. A close up of a map

   Description automatically generatedHow does the size of the file relate to the number of rows in that file?

The size of the file appears to be linearly related to the total number of rows in that file. In the figure below (“0.csv” not included), a simple linear model shows quite a good fit to this data, which suggests that rows in this dataset, on average, are relatively similar in the total size of content.

1. How long does it take to process all the data?

The R system.time() function indicates that this script requires roughly 10 minutes to complete.

1. user system elapsed
2. 407.64 23.32 600.32
3. Do you think this same approach you took work for 10 times as many files? What if each file was 10 times larger?

This solution will likely still function with an increase in the number of files, though the script would take at least 10 times as long to run. Through processing in batches, we reduce the total use of memory at any one given point in time, so our primary concern would simply be the amount of processing time devoted to reading files and applying our summary statistic function to each.

However, we are likely to face significant problems if the size of files increased 10 times – a size likely beyond the capability of machine memory. As discussed in class, attempting to store a very large file in memory may lead to storage in virtual memory on disk, with associated slowdown in performance. This increase may cause dramatic changes in the total time devoted to data processing, quickly rendering this approach unusable.

1. How do you imagine you could make it faster?

Performance could be improved by distributing the processing job to multiple computers. Because the operations on different files can occur independently, assigning separate computers to processes batches of files independently before combining into a single output may greatly speed up this process. This could be done by distributing the job to multiple machines or running a process in parallel.

Other options might include more efficient coding and use of memory. Optimization of algorithms used to apply the sum function over multiple rows of data can increase performance. For example, based on a presentation I attended given by the developer of data.table in R, differences in the way in which rows and columns are stored in memory could change the speed of this process– from my recollection, using fread() over read.csv() changes the way in which rows and columns are allocated in RAM, the former maximizing speed of access across vertical sets of entries (number of rows), while the latter maximizes access of entries horizontally (columns). When processing very large datasets, this change can dramatically increase reading speeds. While I’ve attempted to use data.table in this assignment, similar sorts of changes are likely possible for many components of my analysis workflow.