Data Wrangling Summary

Data Import and Clean-Up of Variable Names and Types

During data import, all variable names were converted to lower case and spaces were replaced with underscores. Variable names were shortened where possible and shared variable names were standardized to facilitate the table joins. Other names had to be made more distinct to accommodate planned table merge. In addition, certain variables had to be converted from character to integer or numerical types. All imported files were converted to data frames with names that clearly indicated the nature of the data.

Missing State Level CDC Suicide and Homicide Data

Several states were missing CDC homicide and suicide data for specific years. To assure the same 18 years of data for every state, both the suicide and homicide tables were joined to a complete CDC population data table. A consequent review of the state level data indicated that missing entries were likely not zero values. Further review of state level min, max and mean values suggested replacing the missing values with a state mean. After replacing the missing death count values, the related death rate NA’s were calculated where missing. Data was inspected for consistency and outliers, revealing only that the District of Columbia has an unusually high homicide rate, partially the result of a very small population base.

Letter to Score and Reordered Rankings of Giffords Gun Law Grades

The Giffords gun law data included letter grades, but the analysis would be better served by also having numerical grades. The was addressed by joining the gun law data to a letter grade conversion table and then using a mutate command to populate an equivalent numerical gun law score in the gun law data table. A second issue arose from the gun death ranking variable running from low to high, with highest being the best ranking. This was the opposite of the gun law rankings where low scores were the best. It was decided to maintain the same low-high ranking logic across variables and convert the death rank scores accordingly.

Collapsing Variables for Boston University State Gun Law Data

The state gun law data tracked back to 1991 and included 134 gun law variables. First, the data was filtered to include only the 1999-2016 timeframe covered by the CDC data. Second, and much more complicated, a law category data table was created with specific laws grouped into broader categories reducing the number of gun law variables from 134 to 14. The additional source state law code file facilitated this process, providing the category names and the related variables for the required mutate sequence. The original disaggregated file was retained for potential use at the machine learning phase of the capstone project.

State Region and Subregion Census Data

Census Bureau data for state regions and sub-regions were imported to allow higher level data analysis investigating regional differences in gun ownership, gun laws and gun deaths. A unite command created region and subregion variables combining the code number and the name for simpler sorting of data legends when creating plots.

R-code file is available in GitHub repository here: <https://github.com/datahoundz/Springboard_Data_Science/blob/master/01_data_import.R>