Project: Investigate a Dataset

Dataset Information

FBI Gun Data (1998-2017)
Source: Github

The data comes from the FBI's National Instant Criminal Background Check System. The NICS is used by to determine whether a prospective buyer is eligible to buy firearms or explosives. Gun shops call into this system to ensure that each customer does not have a criminal record or isn't otherwise ineligible to make a purchase.

Data Wrangling

• First, I made sure that there are no duplicated rows in the dataset. Indeed, the dataset was already "clean".
• Then, I took to finding any null values. There we no null values in the non-numeric columns ("month" and "state"), so I filled in every NaN cell with "0" instead. The necessary checks were performed to confirm that there were no null values anymore.
• Thirdly, I converted the column "month" from string to "datetime" for consistency.

After performing the necessary cleaning, I started to make the dataset more convenient and statistically "rich".
• I confirmed that all numeric columns were the sum of the Totals columns, to better understand the logic of the dataset.
• Then, I noticed that all the numeric columns were floats with ".0" at the end, so there were no "partial" checks. It means that I can easily convert them to ints.
• Also, I created separate columns "year" and "quarter" to be able to better understand annual and quarterly trends.
• Finally, in order to better group all the columns into sub-categories, I have grouped all columns related to handguns, long guns, other types, and permit-related columns as well.

Data Analysis Process

To analyse the dataset, I have asked two state-specific questions:

Question 1: How did the split by gun type change for the state with the highest number of background checks in 1998-2017?

Question 2: Is there any seasonality in the trend for long gun and handgun background checks in the state of New York in the first decade of the XXI century (2000-2009)?

Data Investigation

Question 1

The lion's share of necessary dataset preparations was made at the Data Wrangling stage, however there were several other actions to perform before I could analyse the data.
• First, I had to query the state with the highest number of background checks throughout the whole period of the data frame, which is 1998-2017. I found out that it was "Kentucky".
• I created a list of gun-only items **['total_long_gun','multiple','total_handgun','total_other']** so that non-gun items **[permit-related columns and 'admin']** are excluded.
• Then, I created a separate Kentucky dataset by using pandas "query" functionality.
• After that, I had to filter data by years 1999 and 2016 in two separate subsets. These subsets were grouped by gun type using the **groupby()** functionality.

Question 2

Same as for Question 1, the major part of data preparation was done during Data Wrangling process, however there were several other actions to perform before I could analyse the data.
• First, I had to query the information related strictly the state of New York and related to the period of 2000-2009 by using the "query" functionality

• Then, I grouped it by month using the **groupby()** function to see the aggregated monthly trend for long guns and handguns
• To further analyse the seasonality of long gun background checks, the same data (now, only having long gun background checks) was grouped by quarter – to see the quarterly trends.

Summary statistics and Communicating final results

Question 1

In order to correctly communicate my findings, two big pie charts were created using the **plot.pie()** function: one for the year 1999 and the other one for the year 2016. Proper color schemes, size and explode options were chosen to show the end user the changes between two time periods in the most descriptive way:

- x-axis label names were adjusted to be more "presentable";
- title was added;
- y-axis labels were removed.

Question 2

Two major things were done to better describe the seasonality for log guns and handguns in the state of New York:
• 2 plots were built using the **plt.plot()** function that showed the trends on a monthly and quarterly basis. The monthly trend for handguns already did not show any seasonality, so it was excluded from the quarterly analysis. Several improvements were made to achieve a better visibility:

- The y-axis (number of background checks) was adjusted to better capture the fluctuations in background checks per quarter;
- The x- and y-labels were added and formatted;
- The y-axis legend was relocated;
- The title was added

• Two additional aggregated tables were added to confirm the observations in the line charts. The one showed the months with the highest number of background checks, and the other one – with the lowest.
All these visualizations helped answer the question in an exhaustive way.