CS 590V: Data Visualization and Exploration

Final Project Report: May 11th, 2017

Kriti Shrivastava (Student ID: 31041848)

Edlab Webpage: http://www-edlab.cs.umass.edu/~kshrivastava/

Dataset: Crime in Context, 1975-2015

https://www.kaggle.com/marshallproject/crime-rates

This data was collected and analyzed under the Marshall Project. They collected the last 40 years' crime reports from 61 police agencies. These reports contain information on the four major crimes that the FBI classifies as violent — homicide, rape, robbery and assault — in 68 police jurisdictions with populations of 250,000 or greater. They calculated the rate of crime in each category and for all violent crime, per 100,000 residents in the jurisdiction, based on the FBI's estimated population for that year.

Metadata:

The table contains around 2.8k records and 15 columns.

File size: 257.75 KB

Column Description:

Column Name	Description	Data Type
report_year	Year the data was reported (1975-2015)	Numeric
agency_code	Code of the agency which reported the data	String
agency_jurisdiction	Jurisdiction area of the agency (City, State code)	String
population	Population of the area	Numeric
violent_crimes	Total number of violent crimes in the area	Numeric
homicides	Number of homicides	Numeric
rapes	Number of rapes	Numeric
assaults	Number of assaults	Numeric
robberies	Number of robberies	Numeric
months_reported	Month of the year for which the data was reported	Numeric
crimes_percapita	Number of crimes per 100,000 residents	Numeric
homicides_percapita	Number of homicides per 100,000 residents	Numeric
rapes_percapita	Number of rapes per 100,000 residents	Numeric

assaults_percapita	Number of assaults per 100,000 residents	Numeric
robberies_percapita	Number of robberies per 100,000 residents	Numeric

Related Work

1. The Marshall project website: https://www.themarshallproject.org/2016/08/18/crime-in-context#.H2vCNHqHi

   In this study, the focus was to answer questions like: Where is violent crime up? Where is it down? They presented the crime trend in selected cities between 2000 to 2015. They also presented 4 different patterns of crime amongst the cities. All the visualizations were done using line graphs only.

2. Kaggle Discussion: https://www.kaggle.com/haimfeld87/crimes-in-us-1975-2015

   In this project, the focus was on the crime trend in New York specifically. Questions like: Why crimes in NYC reduced at the 90's? Is it related to the mayor of NYC? were answered. The visualizations in this project were mainly done using bar charts.

The questions that I have tried to answer and the visualizations used in my project were different from these previous works.

Interest/ Questions answered

Through my analysis and visualizations of this data, I answer the questions like-

- Is crime in America rising or falling?
- Which cities have a higher crime rate? Which cities are safer?
- Is there a relation between the time of the year and the crime rate?
- Is the crime rate dependent on the population of the area? If yes, how?
- What is the lowest and the highest number of crimes reported in a city?
- Is there a relation between various kinds of crimes?
- Which amongst the 4 is the most common type of violent crime?
- What is the overall trend in different types of crimes?, etc.

Few questions like: Which type of crime are you most susceptible to in an area? are not answered directly by one graph but can be interpreted by interacting with multiple visualizations (select state from the geo-spatial visualization and then the answer can be interpreted using the composite line graph showing the trend in different types of crimes).

The specific questions that a visualization answers, are mentioned along with the description of the visualization on the webpage.

Analytics and Computations

1. Data pre-processing (python):
    a. Missing values
        i. Blank rows: There were rows in the data with the location mentioned as "United States" and no other value for any other column. I decided to delete these rows from the table.
        ii. Invalid month value: For few rows, the month column has an invalid value (0). For such rows, I used interpolation to fill the missing value.
    b. Separate the city and state code values from *agency_jurisdiction* column: For creating visualizations at city and state level (ex. Geo-spatial choropleth), I extracted these city and state values from *agency_jurisdiction* column.
    c. Find the latitude and longitude values for the location (city, state code): To plot circles for marking cities on the US map, I had to calculate the latitude and longitude values for every city. I used Google's map API for this task. As the number of records were huge, the requests made per second exceeded the allowed limit. This issue was resolved by adding a sleep timer between consecutive requests. However, inserting a sleep between calls made the script take a long time to complete the task.
    d. Join month and year: Created a new column *date,* by concatenating the valued from the columns *months_reported* and *reported_year*. This column was needed to make graphs with zoomable time axis.
2. Summarization (javascript):
    - For every year and month,
        1. Total number of violent crimes
        2. Total number of rapes, assaults, robberies and homicides
    - For every state,
        1. Total number of violent crimes
        2. Total crime per capita
        3. Total number of rapes, assaults, robberies and homicides

- For every city,
    1. Total number of violent crimes
    2. Total crime per capita
    3. Total number of rapes, assaults, robberies and homicides

3. Clustering (python): Using k-means clustering, I grouped the cities into 4 groups *safe, moderate, unsafe* and *very unsafe* based on
    - Total violent crimes
    - Crimes per capita
    - Total rapes/assaults/homicides/robberies
    - Year and month

    A city can be safe in a year and unsafe in a different year.

Exploratory Analysis Discoveries

Through analysis and visualizations, I discovered the following things-

1. Contrary to what people would assume, the overall violent crime rate across the US is declining.
2. New York saw the most number of violent crimes over the span of 40 years, followed by California.
3. New York, Los Angeles and Chicago are the top three cities in terms of the crimes reported.
4. 24 times in past 40 years, a city became very unsafe because of the extremely high number of violent crimes reported.
5. In general, it can be observed that crime increases with the increase in population of the city. Safer cities have lesser population.
6. The late 1980s and early 1990s saw the most number of violent crimes reports with the peak at 1991.
7. Amongst the different types of crimes, assault is the most common one and homicides are reported the least.
8. Violent crimes reports increased towards the end of the year and the month of December has the most crimes.
9. December 1995 had the highest total violent crime reports. However, May 1990-1995 had the most crime per capita.

These findings are also mentioned along with the description of the corresponding visualization on the webpage.

Implementation

- Python 3.6: Used for data preprocessing and analysis
- MongoDB 3.4: Storing data
- Javascript libraries: DC.js, crossfilter.js and D3.js to make interactive graphs
- Bootstrap 3.3.7: Used to make css layout for html page
- Others: Flask server (render html page with mongoDB data), pymongo (interact with mongoDB using python), sklearn (k-means clustering) and pandas

Visualizations and Interactions

The following visualizations were used:

1. Geo-spatial (US map) heatmap: To show the crime rate in different regions of the US
    - Selection and probing (display the crime statistics of the state)
2. Time series chart: To select a range of years between 1975 to 2015
    - Selection
3. Scatter plot: To show total crime/crime per capita across various cities in the US
    - Zoom and probing (display the details of a city)
4. Pie chart: To show the distribution of cities into *safe, moderate, unsafe* and *very unsafe* based on crime, year and month
    - Selection (group of cities) and probing
5. Splom: To depict the relation between different types of violent crimes
    - Selection (cluster of interest)
6. Composite line Chart: To show the trend in different type of crimes over the years
    - Zoom (view the crime rate trend over a selected year) and probing
7. Scatter plot: To show how the population of an area affects the crime rate
    - Zoom (select population ranges) and probing
8. Area chart: To show the total and average violent crime over time
    - Zoom (view the crime rate trend over a selected year) and probing
9. Heat Map: To show the distribution of total crime/crime per capita over months
    - Selection (month and year of interest) and probing

10. Table: To show details of the top crime records
    - Sorting (sort table based on different columns)
11. Additional map tried: Box plot: To show the minimum, maximum, average crimes reported in the 4 groups of cities: *safe*, *moderate*, *unsafe* and *very unsafe*. I decided to not show this graph on the webpage because I could not change the default colors of the graph according to the color legend followed across the other graphs on the webpage.
    - Selection (select group of interest)

Other features:

1. All visualizations are linked with each other.
2. Parameter selection for heatmap: View visualizations on Total Crimes (default) or Crimes per capita
3. Parameter selection for scatterplot: View cities data based on Total Crimes (default) or Crimes per capita
4. Reset all graphs (clear all filters)