Crime Rate Analysis in New York

Data Source

The data set that I will be working with is crime rates in the state of New York (New York State Index Crimes) from New York State Open Data.

I chose this data set from Kaggle because I am interested in what types of crimes are mostly committed in the different counties of New York.

Origin/History: The New York State Division of Criminal Justice Services Mission Statement is to enhance public safety and improve criminal justice. This open data set was started by the development of the Office of Justice Research and Performance (OJRP). The OJRP compartmentalizes other statistical information systems including the State Uniform Crime Reporting (UCR) system. The UCR system has developed this data set in accordance to standard formats from the FBI to implement a similar reporting across states.
These index crimes are categorized based on their seriousness and frequency of occurrence, this data was housed so trends can be conducted on the volume and rate of crime in New York. Index crime data is collected from more than 500 New York State police and sheriff's departments. Each month, participating agencies summarize their reported incident data into several categories described in variables & data types. One important factor that should be mentioned is that although DCJS encourages agencies to contribute to the UCR program, this participation of uploading crime data is not mandatory. If all data were to be uploaded onto this digital platform, the use of statistical analysis and visualization would be very powerful and insightful for preventative measures of crimes. There is only so much the New York State Division of Criminal Justice Services can do without hindering their own work routine.

Variables & Data Types: The crime reporting data set includes several variables of interest. Variables are: County, Agency, Year, Months Reported, Index Total, Violent Total, Murder, Rape, Robbery, Aggravated Assault, Property Total, Burglary, Larceny, Motor Vehicle Theft, and Region.

The different variables at hand are discussed in data type form:

- County is Location where the crime was reported (String)
- Agency is Police Department that reported the crime (String)
- Year is Year the crime incident was reported (Number)
- Months Reported is Number of months an individual agency reported for that year (Number)
- Index Total is Includes sum of Murder, Rape, Robbery, Aggravated Assault, Burglary, Larceny and Motor Vehicle Theft (Number)
- Violent Total is Subtotal includes Murder, Rape, Robbery and Aggravated Assault (Number)
- Murder is One count per victim. The willful killing of one human by another. (Number)
- Rape is One count per victim. (Number)
- Robbery is One count per victim. The taking or attempting to take anything of value from the care, custody, or control of a person. (Number)

- Aggravated Assault is One count per victim. The unlawful attack by one person upon another for the purpose of inflicting severe or aggravated injury. (Number)
- Property Total is Subtotal includes Burglary, Larceny and Motor Vehicle Theft. (Number)
- Burglary is One count per victim. The unlawful entry of a structure to commit a felony or theft. (Number)
- Larceny is One count per victim. The unlawful taking, carrying, leading, or riding away of property. (Number)
- Motor Vehicle Theft is One count per victim. The theft or attempted theft of a motor vehicle, including automobiles, trucks, buses, motorcycles. (Number)
- Region is where the crime was reported. (String)

7 Different Data Types (Shneiderman's Taxonomy)

- According to Shneiderman's Taxonomy, there are 7 different data types. 1-dimensional, 2-dimensional, 3-dimensional, Temporal, Multi-dimensional, Tree, Network.
- Year is considered temporal from Shneiderman's Taxonomy because it is a historical presentation that may create a different data type if all other variables were to follow the order in which the year indicates. For example, starting from 1999, crimes may have centered in a specific part of New York, and as the years go by we can see where the crime hotspots have been migrating to. I have Years included in the dataset and this is a very important factor because there is a distinction in "start and finish time". In hindsight, I believe that this entire data set is following the trend of the variable Year, since all these variables follow suit to the Year variable, I am deciding that the entirety of the data set follows the data type Temporal.

Description of crimes: The categories of murder, rape, robbery and aggravated assault are considered violent crime. Burglary, larceny and motor vehicle theft are considered property crime. Within this set of variables two more refined and subcategory variables are going to be implemented in the Tableau analyzation. With this in mind, I can summarize which subcategory has the most frequency, or is the most occurring.

Initial Planning

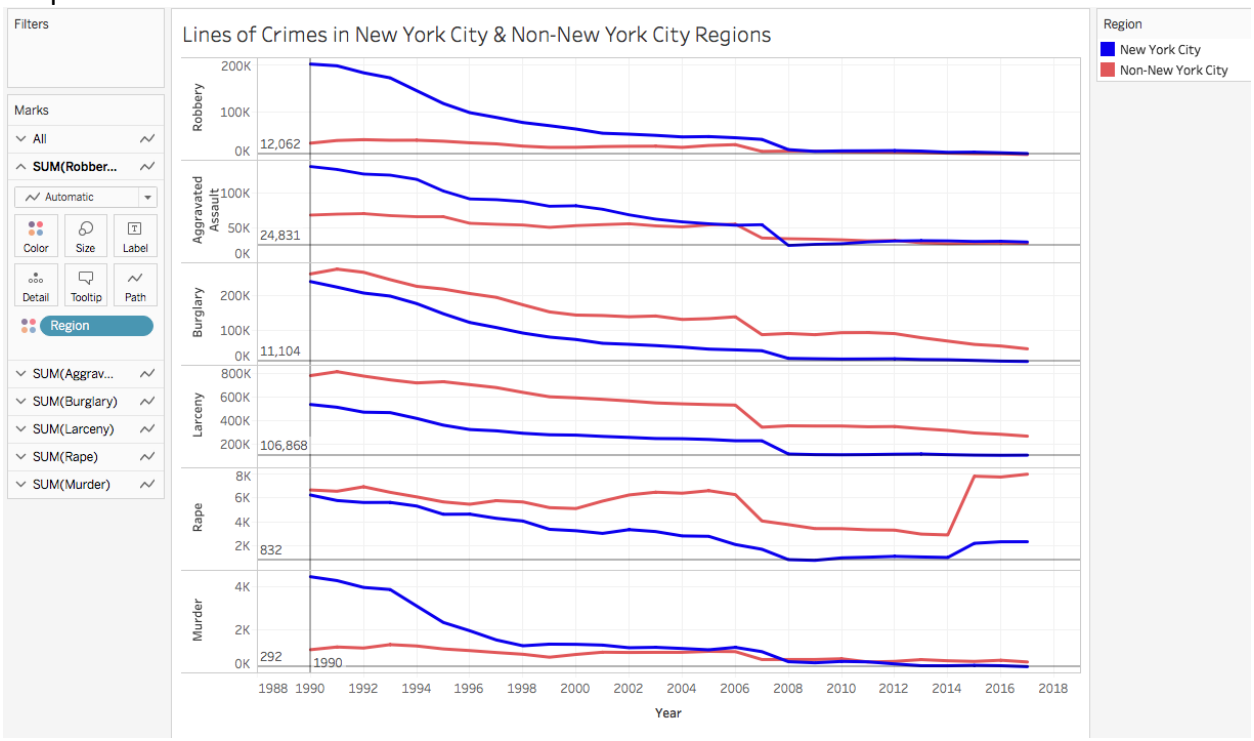
Having dabbled with Tableau and trying out their visualizations, the two visualizations that I am going to implement with my dataset are **packed bubbles** and **lines (continuous)**. Packed bubbles would work very well in showcasing which regions commit the most crimes often, maybe we can also see which years had the most amount of crime rates. This visualization is easy to highlight which crimes are the most committed without having to scan over a lot of numbers or varying degrees of fluctuation. Lines is a powerful and descriptive visualization that can show other lines side by side for easy comparison.

Questions that I will consider when conducting the visualizations are:

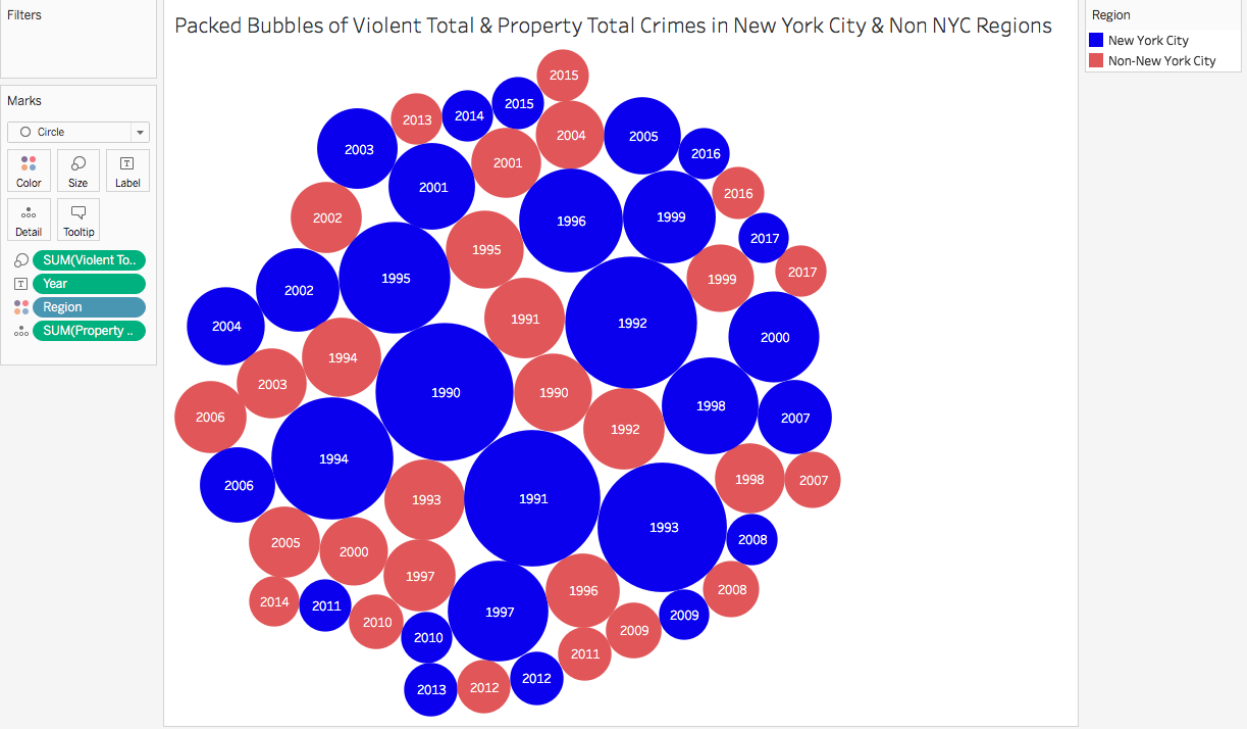
- Crimes can be showcased in order by years so we can see which crimes have been on the rise as the years go by, and we can also discuss which crimes have diminished over time. Which crimes are less frequent now in comparison to previous years? Which subcategory contains more offenders, violent crimes or property crimes? Each crime consists of all offenders of that crime for the appropriate year. Do the 90's have more incidents of violent crimes than the 2000's? Which Police Departments made the most arrests? Since participation in this reporting index is not mandatory there are going to be some missing values which all analysts/researches hate to deal with. How will I deal with the missing values? Most likely these will be completely ignored or given a null value.

Generate Visualization

Graph 1.1



Graph 1.2



Sense Making

It was very interesting working with my New York Crime Rates dataset, and I have to say there were many avenues I could have taken the data into. The two screenshots above are great representations of the huge dataset I uploaded into Tableau.

- Graph 1.1 (Lines) shows the different crimes from top to bottom. Robbery, Aggravated Assault, Burglary, Larceny, Rape, and Murder. Blue lines are New York City regions and Red lines are Non-New York City regions. Right off the bat in the left column next to the name of the crimes we can see an index as to how many crimes are in each category. The crime that stands out the most in sheer numbers is (Larceny), a lot of things go missing in NY. The roof of numbers in that crime tops out at 800 thousand reports of Larceny. The second and third most committed crimes are Robbery and Burglary. These two have numbers upwards of 200 thousand maximum. The crime that has the least incidents is Murder, followed by second to last, Rape. In general view we can see that all types of crimes have decreased as the years go by in New York City regions and non-New York City regions. We can confirm by this graph that the 90's hold the most account of crimes in regard to our current year of 2019. In 1990 we can see that in the New York City regions there was far more Robbery, Aggravated Assault, and Murder cases than in non-New York City regions. Towards 2008 there seems to be an overlap in Robbery, Aggravated Assault, and Murder incidents in both regions of New York. Coming to an end of the description in the graph, one last important observation that I would like to bring to your attention is an obvious spike of Rape crimes from 2014 to the end of 2016. A tad below 4 thousand Rape crimes into a whopping almost 8 thousand

reports in non-New York City regions. In New York City regions there was an increase in Rape but not enough that would be considered a spike.

- Graph 1.2 (Packed Bubbles) lays down a general view of a tally of charges. I would like to reiterate that in this graph I am displaying two subcategories of different oriented crimes. The categories of murder, rape, robbery and aggravated assault are considered violent crime. Burglary, larceny and motor vehicle theft are considered property crime. In reality these two differences cannot be distinguished in the graph since it basically groups them together. But we can attest to the recent graph that in the 1990s it was pretty dangerous to be living in New York City regions. The outer circles are entirely in the 2000 era. Enclosed by the 90s, both New York regions have progressed in society and have diminished crime rates significantly. With a small spike in Rape crimes, this is a great progress for the state of New York.

    - Visual Pitfalls in regard to Bresciani and Eppler's article.
        - My data set has some flaws that are discussed in the article. One major pitfall disadvantage is "Inconsistency", the Police Departments in New York are not required to submit their crime rates at all. I had mentioned this earlier and this kind of put a slight dent in analyzing several counties. For the most part it didn't entirely affect my graphs since I did not analyze counties or police departments specifically. All police departments should submit their data reports to the (UCR) program so the crime data set can grow all together in completeness instead of having values missing. The second pitfall that I noticed in my data set is "redundancy" I know that in many fields' redundancy is a bad thing to have because it is already implemented once, there is no need for there to be anymore representation of it. In the variable (County) and (Agency) there are several instances of the same county and same Police Department. This is because the variable (Year) changes along with them. Either the variable (Year) has to be implemented in many columns, such as 1990, 1991,1992, 1993 etc. so that there may only be one instance of the county and the police department, but then again there are different Police Departments in the same city. This would need to have some thought put into it to avoid such redundancy, but for the most part it is organized in an intuitive manner.

Reflect

I have learned a lot from working with this project and using Tableau side by side to make sense of the data set. My initial statements were met with some clarity and sense. Although there were some flaws with the data, inconsistency, redundancy. I made full use of Tableau's visualization tools to present to you what my findings are. There were a lot of crimes being committed in the early 1990's, diminishing significantly into the 2000's. Rape crimes have spiked recently in both regions of New York and there needs to be preventative measures to see this crime go down. In hindsight, data from all police departments should be submitted (mandatory) to the (UCR) program so that visualizations can be accurately displayed.

Appendix

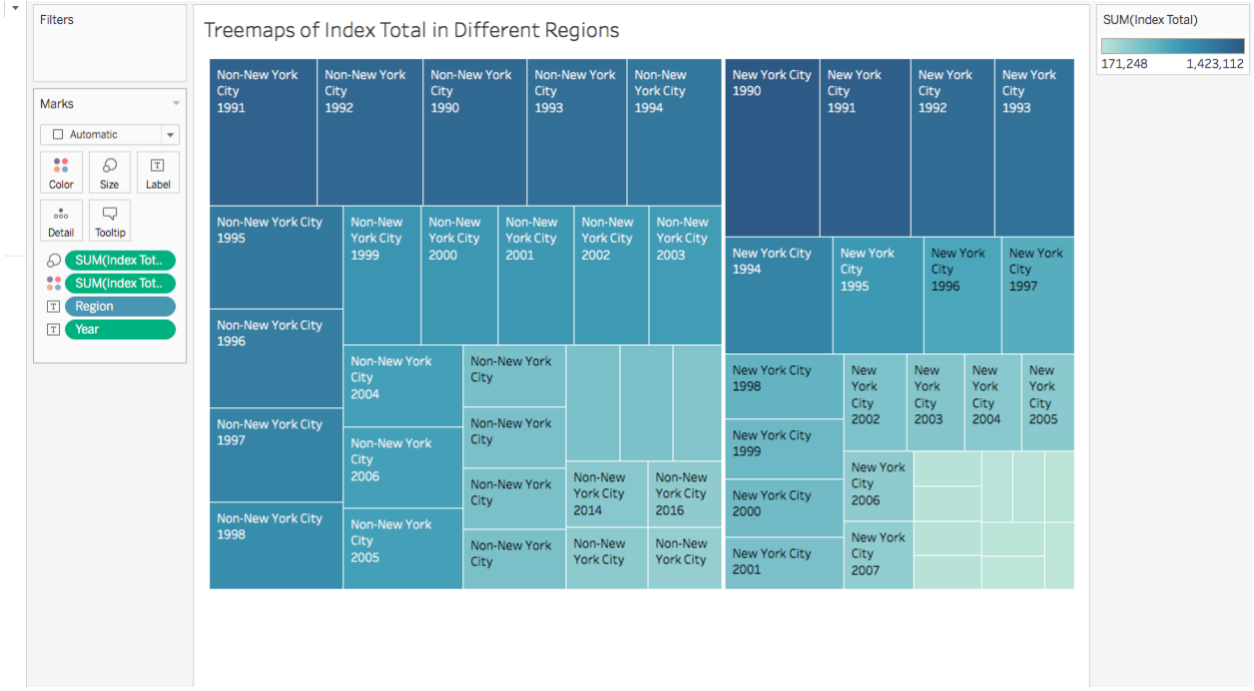
References:

- **URL**: (https://www.kaggle.com/new-york-state/new-york-state-index-crimes)
- The Pitfalls of Visual Representations: A Review and Classification of Common Errors Made While Designing and Interpreting Visualizations – Sabrina Bresciani and Martin J. Eppler
- The Eyes Have It - Ben Shneiderman

Additional Graph:

There is one other graph I would like to show.

Graph 1.3



Graph 1.3 shows which years were the most affected by crimes and the two different regions of New York. 1990 and 1991 still identify as the most concentrated sections of the treemaps which indicate heavy crime flow.