



# Crime Analysis on Homicides of USA



Varun Vikram GRA, Muhammad Abdul Rehman Siddiqui, Vamshi Krishna Ginna, Justin Ku

College of Business + IT, MSIT Program

## INTRODUCTION

Safety and security of every individual is of paramount importance in any civilized society. To improve public safety, an in-depth analysis of crime statistics must be conducted. Data analytics can help law enforcement agencies and governments observe patterns of crimes with respect to region, gender, race, etc., and channelize their efforts and resources in the right direction. A thorough analysis of crime dataset can help us understand the relative importance of various factors in the overall crime scenario. These insights enable us to devise concrete action plan and take precautionary measures to reduce crime rate. Large amount of data cannot be visualized efficiently using traditional visualization techniques, and hence require more advanced tools and techniques. We have used Tableau and related technologies for this analysis to understand the relationships between the occurrence of crimes and other important factors. Through our analysis, we will tell whether a death penalty could be an effective solution to control homicide rates. We will also analyze the trend for homicide rates over the years. Using USA crime data from 1980 to 2014, our analysis focuses on the analyzing the crimes along various dimensions. The dataset includes 638,454 records with 24 features such as Crime Type, Victim & Perpetrator relationships, Weapon, and Age. Thus far, initial findings indicate that 70% of the crimes have been solved. California is the state where maximum number of crimes occurred. Also, 'Handgun' is the most frequently used weapon. We are also investigating the use of data mining techniques such as Decision Tree to identify the patterns in which a crime is solved or not based on the features.

## DATASET

The dataset includes murders from the FBI's Supplementary Homicide Report from 1980 to 2014 and Freedom of Information Act (FOIA). The dataset contains 24 features such as age, race, sex, ethnicity of victims and perpetrators, in addition to the relationship between the victim and perpetrator and weapon used. Our goal is to process the datasets and extract the most important information from them and then draw the relationships between happening of crimes and these factors.

## PROBLEM DEFINITION

To achieve our goal, we try to solve the following problems in our project:

- Has the homicide count decrease over years?
- Are crimes solved improving over the years?
- What are the top 10 cities with their states where crime occurred maximum?
- What are the top 10 weapons used by the Perpetrator?
- Does crime rate follow any seasonality?

## METHODS

To address the problems described above, there are three main steps:

1.) Data collection 2.) Data Preprocessing 3.) Data visualization

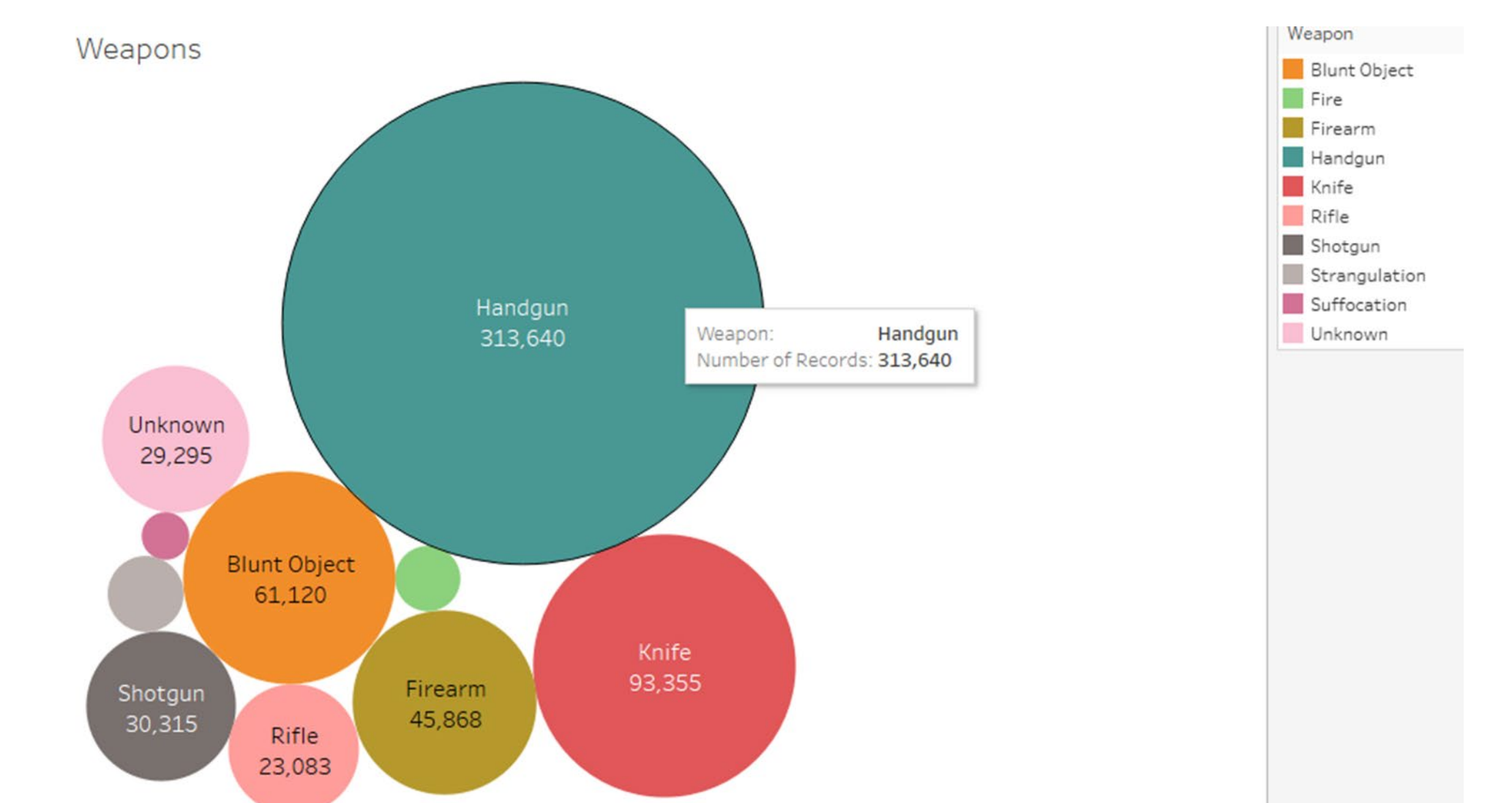
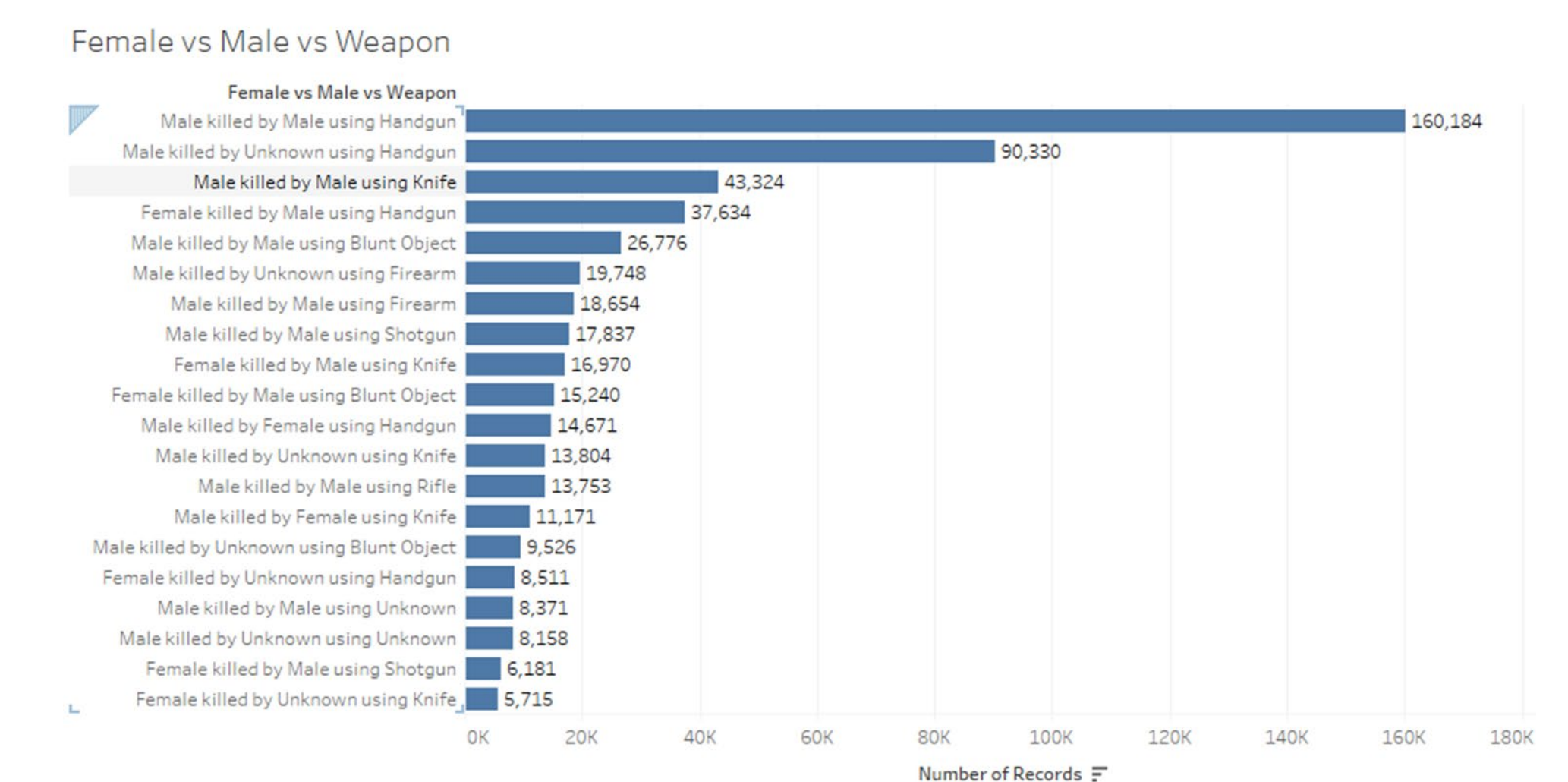
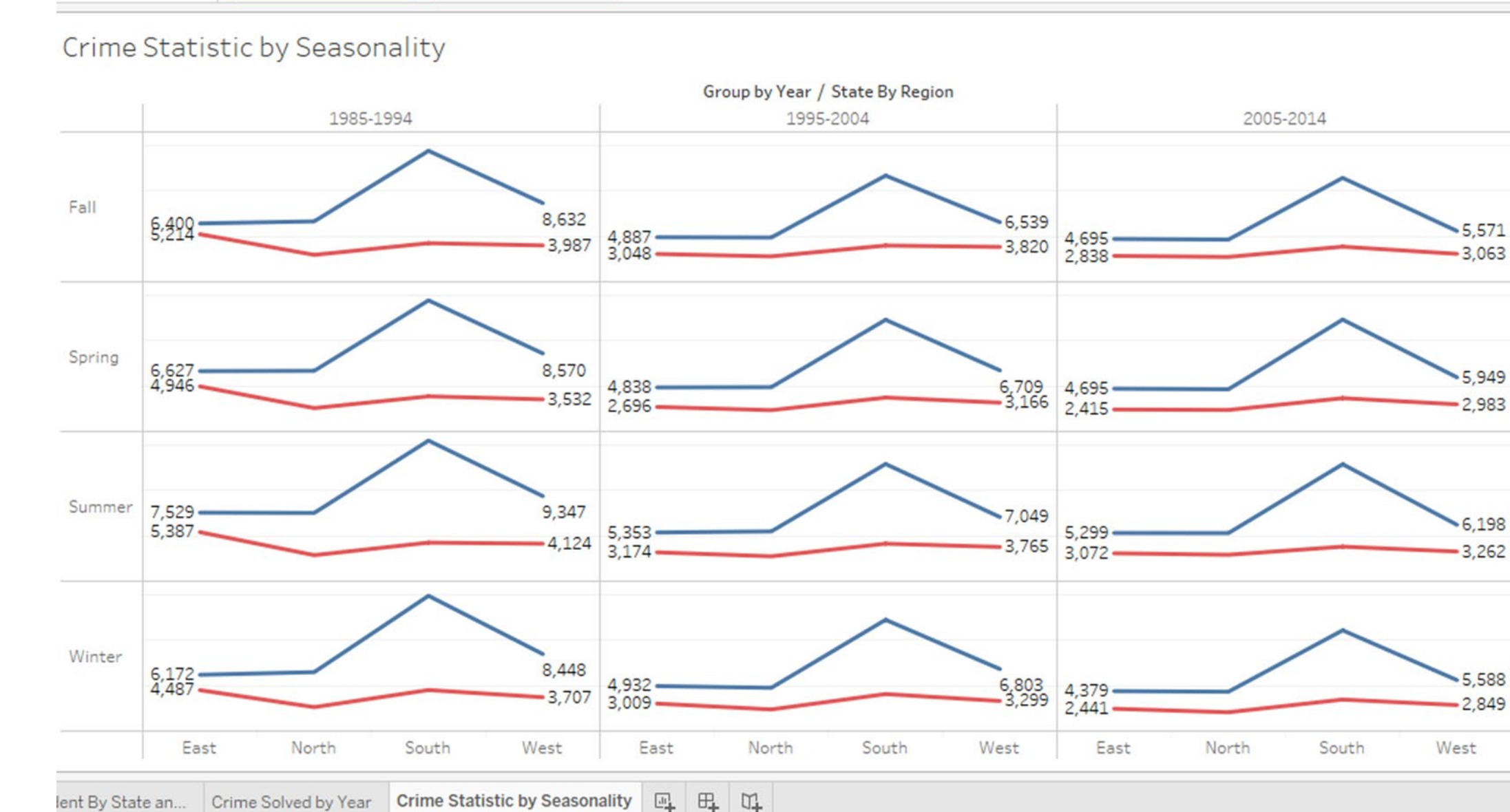
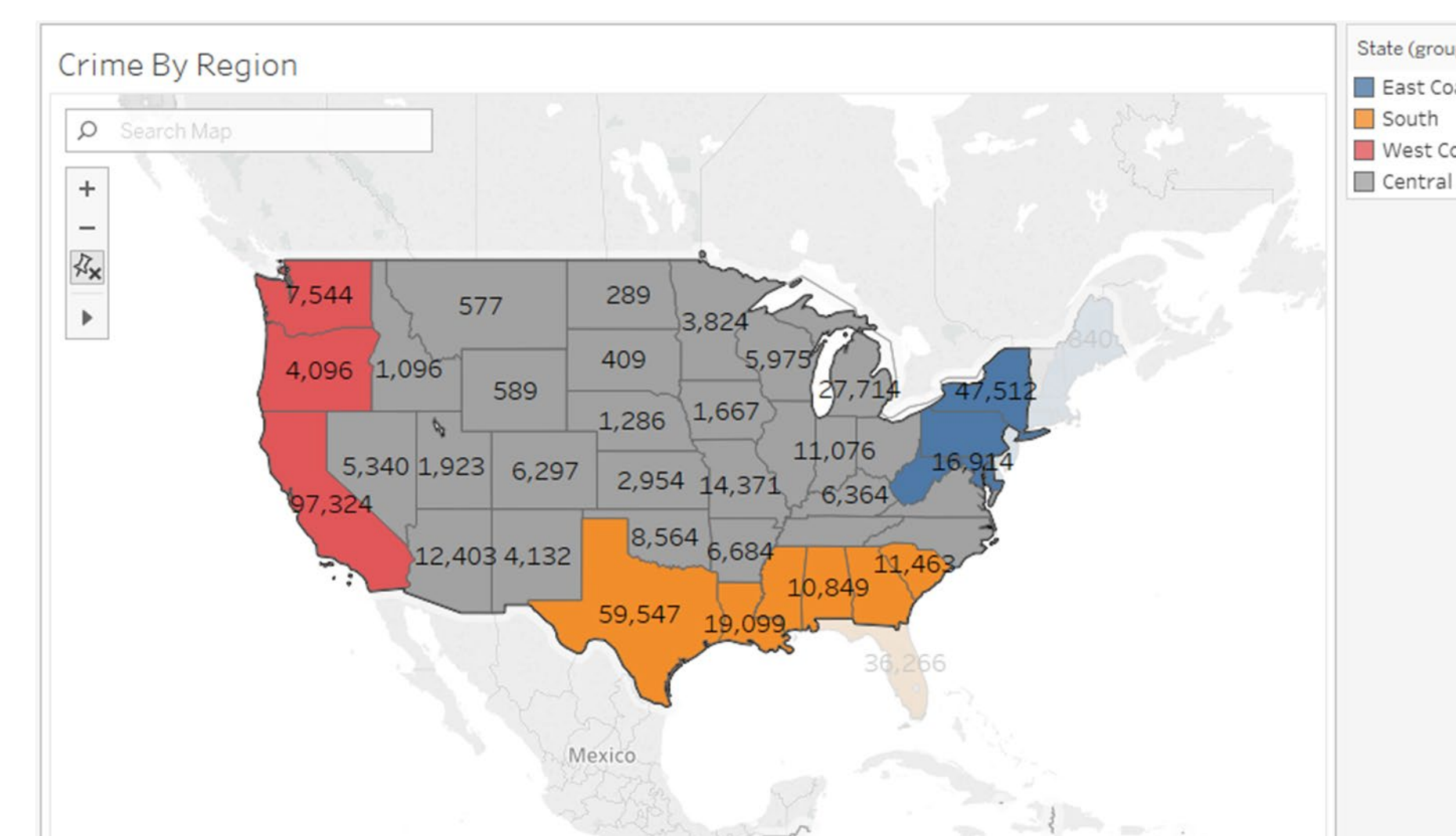
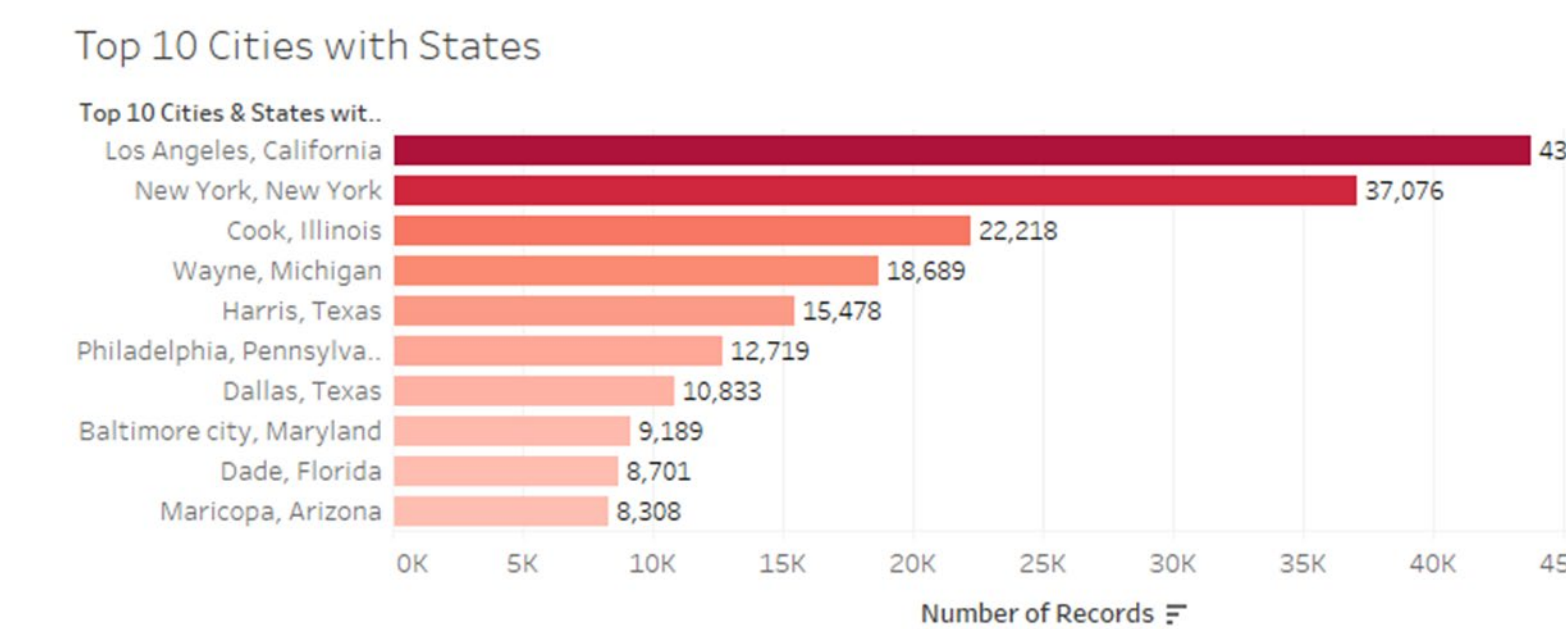
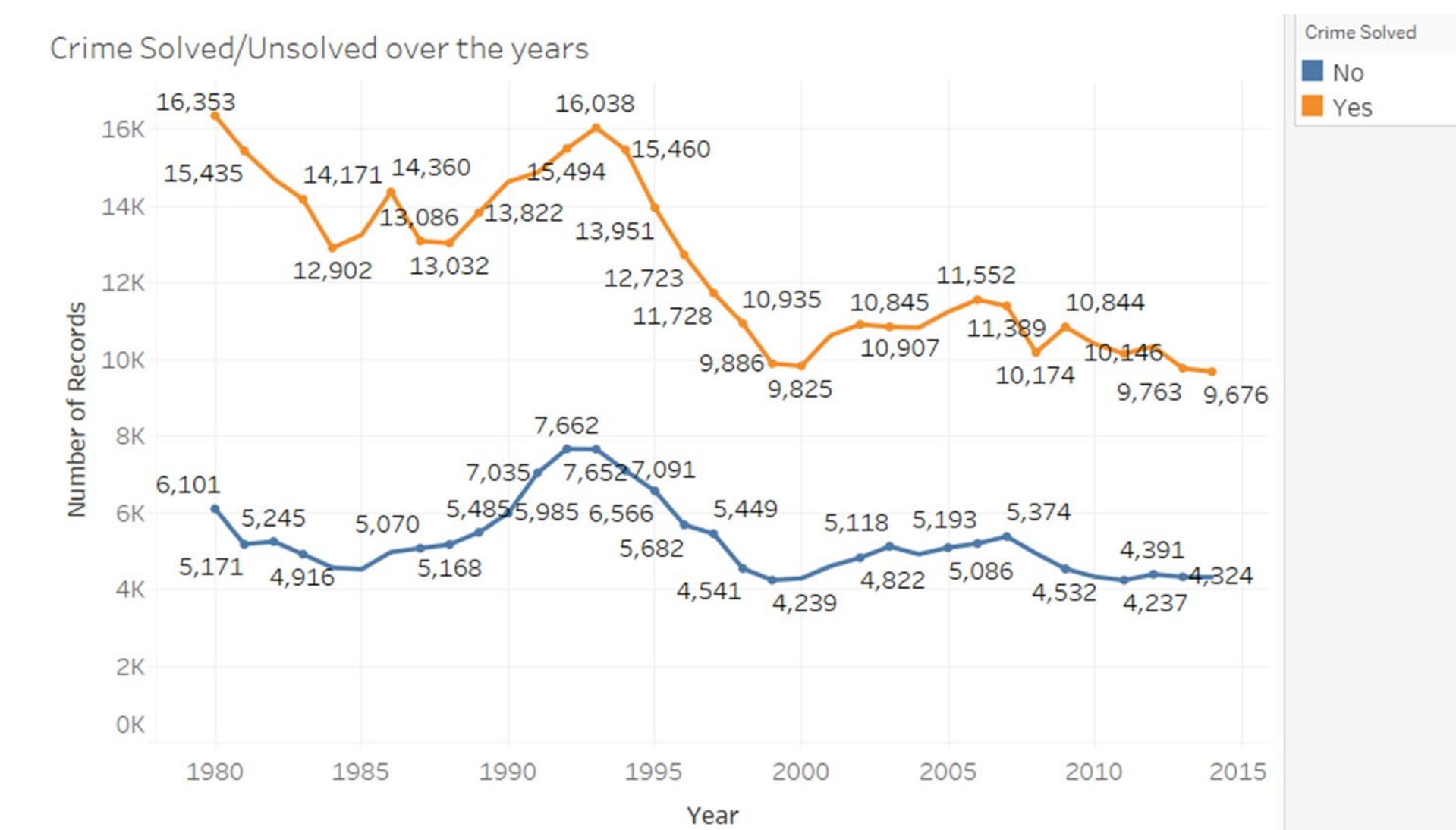
1.) Data collection – In this phase, we have collected data of homicide reports from FBI and FOIA requests. There are no missing and null values, however, the dataset contains unknown values (records of either victim or perpetrator).

2.) Data Preprocessing – Data in the real world is “dirty”. It can be incomplete, noisy and inconsistent. In our dataset, there are a few values which can easily be identified as an outlier. For example, there are values for victim's age which are greater than 100, and in one case it is 900. So, we simply ignored this value and removed it from our dataset.

3.) Data visualization - It is a process of converting data into a graphical or visual representation. It is common for us to efficiently process visual information which can greatly speed the data analysis process. Throughout our research, we have tried to answer questions based on the visualizations we have created. We have included bar charts, line graphs, maps, and bubble charts to clearly tell the story.

In the process of data visualization, we first created a time-series analysis about how the crime rate is changing from 1980 to 2014. We also checked crime solved with respect to time. After gaining an insight, we checked if there is any seasonality shown by the dataset. To do this, we created different groups of month (Fall, Spring, Summer and Winter). We then took cities and states into consideration to gain an insight of the top 10 cities with maximum number of crime occurred with their states. Furthermore, we visualized the relationship between victim, perpetrator and weapons used. Using filled maps, and by dividing states into 4 different regions, we were able to show how the crime was spread across the states and regions. Finally, we identified the top 10 weapons being used.

## RESULTS



## DISCUSSION

The results above indicate that “Los Angeles” is the city with maximum number of crime. Also, we have grouped the states in 4 different regions and then computed the total number of crimes occurred in those particular regions. Also, with time, there is a decrease in solving the crime rates from 1980 to 2014. The most common gun which was used maximum number of times by the perpetrator to attack the victim was “Handgun”. As far as the seasonality is concerned, there is no clear evidence in the graph which is more inclined towards a particular season. The results can be used to understand and guide in the real life. However, in fact, the relationships among all of the factors are more complex than we did here. We hope to do more in the future by considering more factors to analysis and predict crime.