Presentation Transcript

Hello, today I will be presenting our statistical products that use the FBI’s Crime US data.

The contributors of these products are Ruofan Chen, Liquan Zhong, and myself.

Now, I will give a very brief overview of our products. First, we have an interactive map, which shows crime rates by state; rates are per 100,000 inhabitants. The client will get to hover over any of the 50 states, excluding Puerto Rico and District of Columbia. The client will also get to choose which years to analyze, anywhere from 1995 to 2019. A rank bar plot is also included, and it includes the ranks for the 50 states, as well as the ranks for Puerto Rico and District of Columbia.

Next, we have an animated bar plot, which shows the crime rates by area. These areas are divided into metropolitan areas, non-metropolitan areas, and non-metropolitan counties. The client will get to choose any state to analyze except NJ, RI, Hawaii, Puerto Rico, and District of Columbia as there was not enough data collected for these states. This barplot transforms over a range of 10 years, from 2010-2019. A table showing the top states with the highest crime rates is also included.

Next, we have a state trends graph, which shows the crime rates by crime type, state and year. The client will get to choose any of the 50 states, as well as Puerto Rico and District of Columbia. This graph has a range of 25 years, from 1995-2019. Violent crime and property crime boxplots are also included, and they show the statistics for the specific subcategory of each crime type.

We also have a clustering map, which shows the states that have comparable violent and property crime rates. Clustering allows you to find and analyze the groups that have formed organically. The k-means clustering method was used, which is a type of [unsupervised learning](https://blogs.oracle.com/ai-and-datascience/post/supervised-vs-unsupervised-machine-learning), and is used when there is unlabeled data (i.e., data without defined categories or groups). This map displays the 50 states, and it covers a range of 15 years (from 2005, to 2019) – the client will be able to choose which year he or she would like to analyze. A bubble plot is included, showing the actual rates, as well as the population numbers, and it includes the District of Columbia as well as Puerto Rico.

Lastly, for the second part of this presentation, we have a regression analysis. Variables that may affect crime rates were studied. This analysis covers a range of 24 years; from 1995 to 2018, and it includes the 50 states as well as the District of Columbia. A time series regression for each state was done, as well as a panel data regression analysis for each year and a Granger causality test. A heat map displaying the distance of the coefficients was also made. All of these will be explained more in detail by Ruofan.

\*CHANGE SLIDES\*

Now, for the fun part, we will get to show how our products can be used.

Here is our interactive map, which is quite intuitive and user friendly. The client can choose which year and which crime type to analyze, and will get to see which state has the highest crime rate in the given year, or to see which state has the lowest crime rate in the given year. Whatever suits the clients needs. The darker color represents higher crime rates, while the lighter color represents lower crime rates. The rank bar plot helps the client easily see which states rank the highest for each crime type and each year. For example, in 2019, we see that New Mexico has the highest rate of burglary crimes, while New Hampshire has the lowest rate of burglary crimes. One thing that may be of interest to the client, is that District of Columbia is almost always among the top three states with the highest crime rates. Actually, in the early 1990s, Washington, D.C., was known as the nation's "murder capital". However, after the 1991 peak there was a downward trend through the 2000’s. However, as we can see, it is still among the top areas in the US with the highest murder and assault crime rates.

So in this second analysis, we have the barplot animation, which shows the change in crime rates over the range of ten years, specifically, between 2010 and 2019. The client can choose the state and crime type that he or she would like to analyze. For example, we can choose Alaska as the state and assault as the crime type. Here we can see that assault in the metropolitan area was steadily increasing from 2014-2018, and then decreased in 2019. In most states, the metropolitan areas have the highest crime rates, but that is mostly due to the fact that they have higher populations. This barplot is also quite intuitive, and not very hard to interpret. This plot can help the client understand which specific area is the main reason for the variation of the total number of state crimes over the years. In addition, a table is included, showing the top states with the highest rates for each crime type. Not surprisingly, D.C can be found in almost all of these lists, such as assault and murder. However, as mentioned previously, D.C was not included as a state option for the barplot due to the lack of data, but still included in this table.

For the third analysis, we allow the client to analyze the state crime trends over a span of 25 years. The numbers in these plots are calculated as the total occurrence rate per 100,000 inhabitants. One thing to note is that Puerto Rico did not have the total rate for each crime from 1995 to 1998 in the original FBI dataset, but we obtained the total crime counts by using “The World Bank Data” and calculating the rank. One thing to note before analyzing this data, is that the FBI expanded its definition of rape in 2013, and there were two definitions of rape between 2013 and 2016, and after 2017, only the expanded definition of rape was included. We only included the new expanded definition of rape in our analysis, though it is good to keep this in mind to not falsely believe there are sudden spikes of rape crimes around 2013, as it is only truly due to the expanded definition.

So now to analyze the data, first we have the line trend graph, with each line representing a different crime type. The client will get to choose which state here as well as the crime type he or she would like to analyze, by clicking on the crime type and unclicking on the rest. This will give a much better understanding of the crime trend he or she is interested in. For example, we can choose South Dakota, and analyze the assault crime trend by unclicking everything except assault. Here, we can see that there is a slight increase in the assault rates from 2010 to 2019.

One thing that may be interesting to do next, is to analyze in which area this increase is occurring in, so the client may go back to the barplot, and click South Dakota and assault there. And then analyze which specific area assault rates are increasing in. This is an example of one of the many ways our client can use our products interchangeably.

So here, we have two box and whisker plots, showing the summary statistics for each type of violent crime or property crime. The client can also select which crime type he would like to analyze, by unclicking the crimes he is not interested in. For example, we can unclick everything except murder, and if we hover our mouse over this plot, we can read the summary stats such as the median, maximum, minimum, and interquartile ranges. If I click every crime, we can see that assault usually has the highest rates compared to the other violent crimes, and this is actually the case for most states. It is the same idea with the boxplot for property crimes. Usually, larceny theft is the most common type of property crime among all states, as was observed from this data. The client can also get to see these ‘outliers’ here in the trend analysis directly.

So next, we have a cluster analysis; specifically, a map grouping the states that have similar, or comparable, violent and property crime rates. Clustering is a way to group objects that have similar characteristics that we may not have known before, and can be used for pattern recognition. As I briefly explained before, the k-means cluster analysis method was used, and Mahalanobi’s distance was used to calculate the distance, which I will not go into much detail about, due to time. Here, our client can choose which year he or she would like to analyze, anywhere from 2005 to 2019. The number of colors to group the states was chosen to be three. And each color here represents a cluster – so for example if I choose 2019, I can see that these states in purple have similar violent and property crime rates. But their rates are different than the rates of the states that are colored in green. Now, we can see the clusters in the bubble plot as well, if we hover the mouse over the circle, we can see the actual numbers of the property and violent crime rates for each state; as well as the population numbers. Actually, a larger sized bubble means a larger population.

Now, from this analysis, we can only say that the states that are grouped have similar patterns of crime, but we cannot explain why. If we observe the results for each year, we can find that some states do change from one group to another (for example, if we change the year from 2019 to 2018, TN becomes grouped with Alaska and New Mexico). However, most states in the Northeast, for example, usually stay in the same group – meaning they almost always have similar violent and property crime rates. This can mean that the violent and property crime pattern might be related with location.

In any case, Ruofan will later help us understand more about the causes of these crime patterns in more detail.

Now, what is left, are the tables that contain the organized datasets we used for these analyses. Table 1 was used for the map and trend analysis, while table 2 was used for the barplot animation.

And that is all for my part! Next, Ruofan will present our last product, the regression analysis.