Team Pandas

L.I. Villaranda, Jimmy Fox, Brendan Cartin

September 29, 2020

A New Project

The purpose of this analysis was to find a fun topic we were interested in. So we were going to predict the winner of the Premier League based on the last couple of seasons.

> We were quickly steered away from doing any sort of prediction analysis (who did we think we were?) and soon after, realized this topic was not really going to work at all. Sports are big money, so any site with good information charges for it.

The Information Strikes Back...

Next we moved onto wildfires, wanting to work on something topical that we all cared about. Showing causation of wildfires in California was the main statistic we were interested in.

- We had found (what we thought) was plenty of data
- We (LI) learned how to translate a code from SQLite3 to JSON
- We learned it was an enormous heatmap, while beautiful contained information that was unusable for us.... Again....

Return of the Jedi(information parsers)

At the 11th hour we needed a topic; we had been through a fun topic, we had been through a topic with significance to us, it was time to pick a topic that was doable.

> We found statistics on the FBI website containing crime data for the years 2015-2018 by city. Only cities of 100,000 were included in these datasheets.

It should be noted that this analysis solely looks at crime statistics. This is not an analysis to view correlation between crime and other outliers such as income median, poverty rate, education and other economic and socio-economic circumstances.

Attack of the Questions

Question 1	Question 2	Question 3
What do the crime statistics in VA look like over the course of a four-year period 2015-2018?	What do the crime statistics nationwide look like over the course of a four-year period 2015-2018?	What city has the highest murder rate, according to this data?

Revenge of the Cleanup

PROS

The cleanup process was straightforward, and included: renaming, merging, and formatting columns and then merging the data frames together.
The data sets for our crime analysis did not have as much cleanup and were formatted well.

The data sets we had were small but easier to work with.

Smaller data sets made it easier to navigate and clean the data.

Datasets were straightforward with accurate information.

The hope was to do a more complex analysis of causation of crime in cities.

Revenge of the Cleanup cont...

CONS

The data was good but lacked depth. For example compiling totals is good information but we were not able to compare that data to other datasets.

It would have been more interesting to compare the causation of crime against other data such as median income per city or state, education etc.

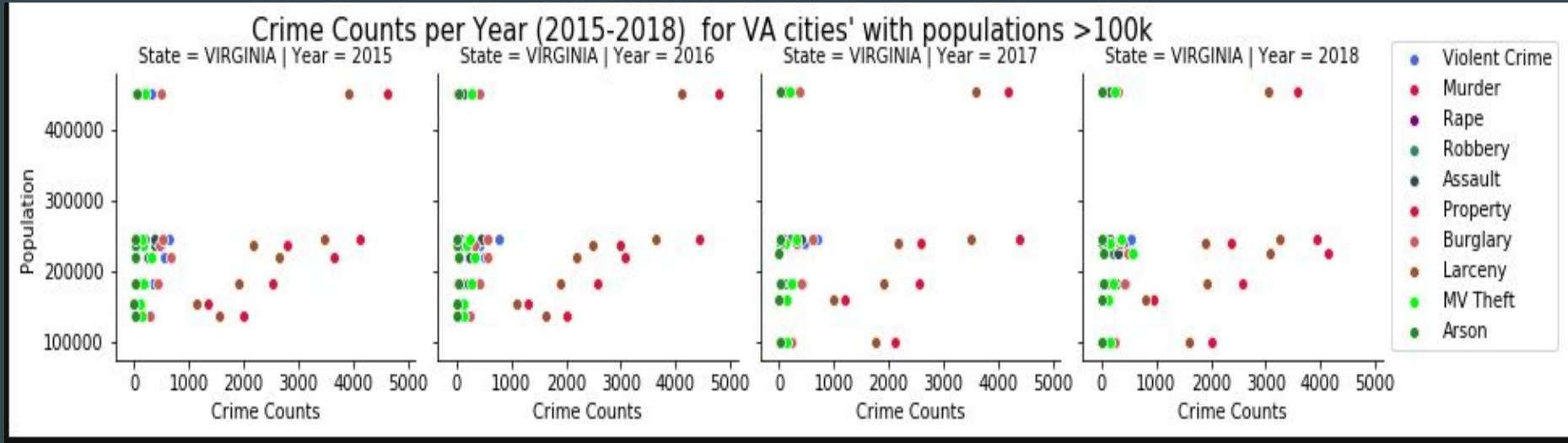
It was difficult to get any sort of interesting information from totaling data by each type of crime.

Comparing the two datasets with a greater gap in time may have showed more differences, as the data from the FBI goes back to 1995.
During exploration, one quirk that was discovered was that only one year would show the approximate population count. We resolved this by filling the column with the previous year's data (i.e. The population of Huntsville, AL (in the 2016 data) was listed as "190,106" in 2015, but the 2016 row was blank.)

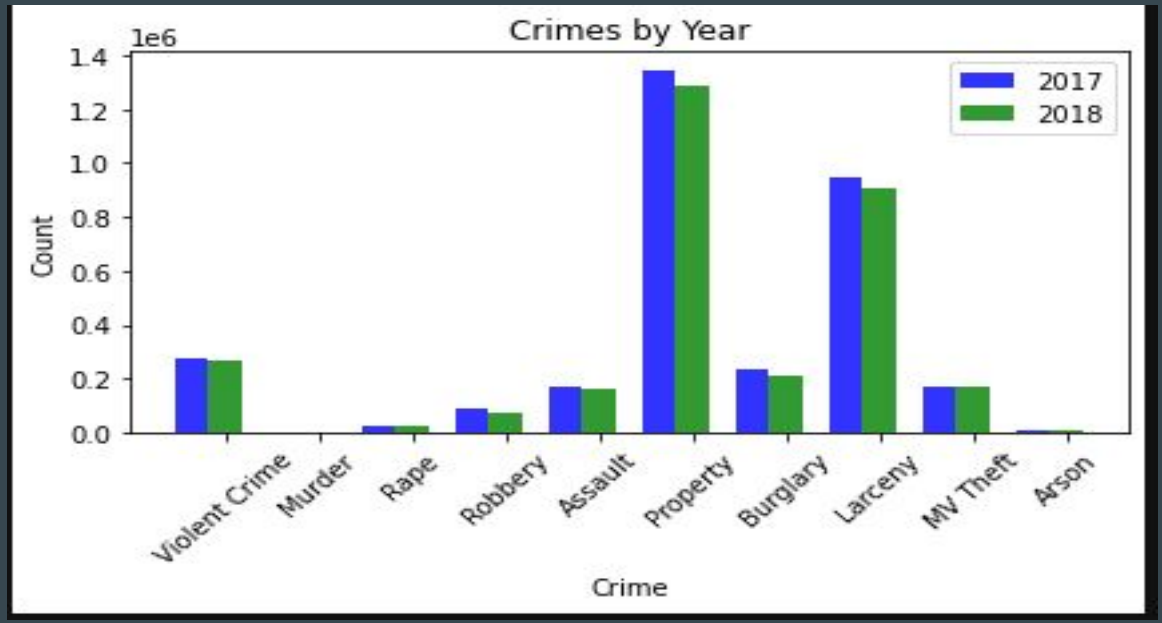
The Analysis Awakens

Now that team Pandas cleaned up the data frame, we were off to analysis. After some formatting, our main goal was to groupby “Year” to get totals for all the crimes, by year.

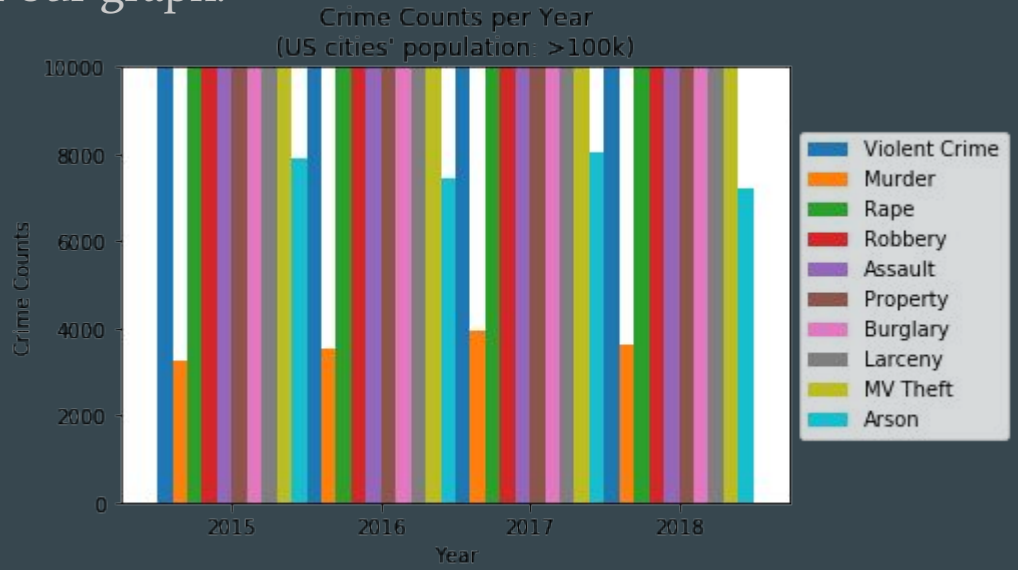
Our first question was regarding crime statistics in our state, Virginia. We made scatterplots for each of the years for all crimes, using Population and Crime Counts as the x and y respectively.



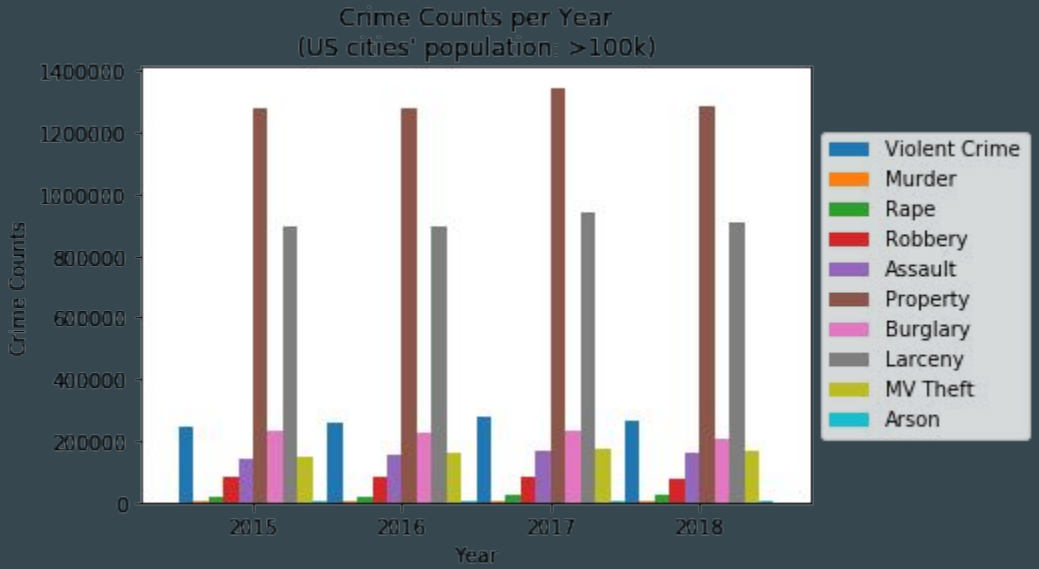
Next we wanted to see a data visualization of crime statistics for the United States as a whole. We started with a bar chart of crimes for 2017-2018.



We then decided to change the x axis to the year and graph the crimes, since the crime counts varied greatly it was hard to visualize using one graph. This is a good visual of the bottom half of our graph.



The below graphs gives a nice overview using a bar chart. However due to major differences in between crime categories some are not showing up as well.



The Last Jedi (Information Parsed)

Where is the murder capital of the United States for the years 2015-2018?

Chicago

The Rise of Team Pandas

Finding a topic with a dataset the team was interested in was more difficult than we thought. This caused a few days of working and cleaning files only to realize we would need to change course.

With extra time additional datasets could be used to correlate crime in cities to other factors such as median income, education stats, employment rates, poverty levels.

It would have been more interesting to focus on a few cities with the high crime. Cities are very diverse economically and socioeconomically. There are major gaps in resources and income as well as other factors that play into crime. Plotting and graphing data in a major city like Chicago may show this.

Coming up with questions to ask on this dataset was a challenge.

The crime statistics for larger cities (over 100,000) people did not change from the years 2015-2018, due to data being so similar from year to year

Chicago having the most murders reported is not surprising because the murder rates were much higher than all the other cities that we had analyzed and plotted/

Being able to predict the winner of a professional sports league would have been much more exciting and perhaps even lucrative.