**21CSAI01I**

**Introduction to Data Sciences**



**Report by:**

**Group 19**

**Rawan Mohamed 193529**

**Mohamed Hesham Negm 206069**

**Topic:**

* Our data is about crime and hate crime rates in one state in the united states, New-York state. We have two datasets, one about the overall crime rates in the past 30 years in all counties in the state. The other one is about hate crime rates in specific in all counties in the state, in the past 10 years only. This is a small dataset we used in answering certain questions about hate crimes.

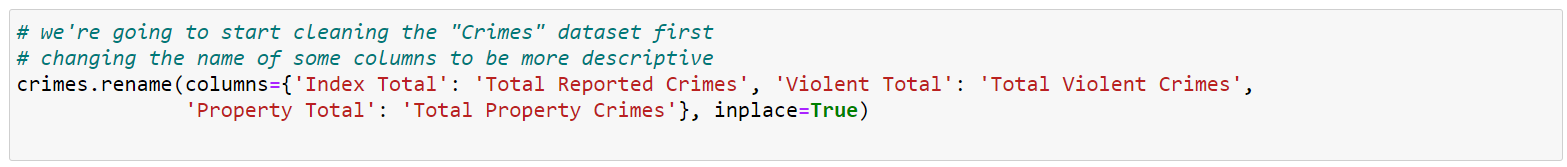
**Questions:**

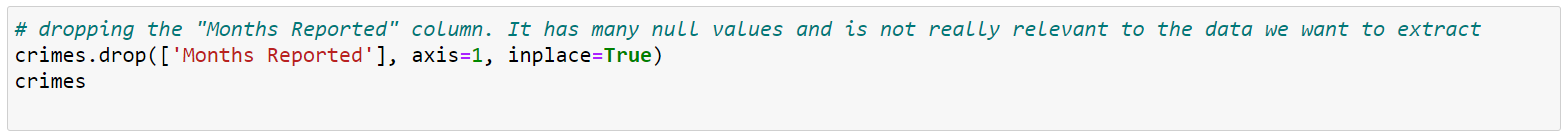
1. Did the covid-19 pandemic affect crime rates in the state?
2. What are the total crimes reported per decade in the 1990s, 2000s, and 2010s? And which decade had the least and most crimes?
3. Which group has the most hate crimes? And how many were against people and how many were against their property?
4. Is there a correlation between rape and gender hate crimes?
5. Do the counties with the most hate crimes have a higher overall crime rate?
6. How does the geographical location affect hate crimes against religious groups?
7. Did the Iran-Iraq war affect crime rates in new york?
8. How did the “me too” movement affect rape rates in the state?

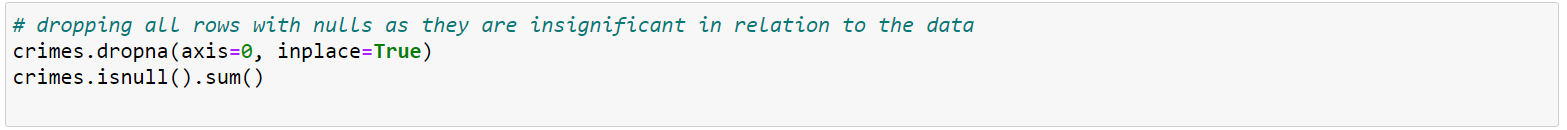
**Project One**

**Pre-processing:**

* After reading our two datasets, we started cleaning and tidying up the crimes dataset. The hate crimes one was almost ready for analysis so it didn’t need much work.
* In the crimes dataset, we changed the names of some columns to make them more descriptive, dropped a column that is irrelevant to us, and dropped the rows with nulls in them, after dropping the nulls, our data went from **21228** rows to **20844** rows.

****

****

****

* We converted the “Region” column to numerical values because it only has two values “New York City” and “Non-New York City’. It specifies if the county in the row is inside the city or not.



* We continued cleaning by changing the datatypes of some columns. We faced trouble changing the datatype because some values had a “,” which caused an error. So we handled it by replacing them with a space and proceeded as follows.

**Text

Description automatically generated**

* Instead of dropping more data, we made another dataframe containing the data we’ll use the most in answering our questions, keeping the rest of the data in the original dataframe in case it’s needed later.

****

* In the hate crimes dataset, we simply dropped columns that had no value whatsoever, they we zeros in all rows, and we converted the “Crime type” column to numerical values because it only had two values and it would be easier for us to analyze that way.

A picture containing chart

Description automatically generated

* We used some aggregation functions like sum, count, max, mean, etc, to gain some general insights about the whole data.

**Answering questions:**

**Question 1:**

* We went on about the first question, which is regarding the covid-19 pandemic, by first finding out how many crimes were reported each year in the last 30 years using the groupby and sum function. Then we visualized using a bar chart since it was the most suitable for us to spot the difference immediately.

**Chart, bar chart

Description automatically generated**

* We then repeated the same with the hate crimes dataset, and also visualized the results.

Chart, bar chart

Description automatically generated

* From the two graphs, we inferred that covid-19 didn’t affect crimes; however, it affected hate crime rates, since we can see in the second graph that in 2020, hate crimes were significantly less than in previous years.

**Question 2:**

* In the second question, we wanted to know which of the last three decades had the most and least crime rates.
* We used “pd.cut” to split the years into three decades, then used the groupby and sum functions to find the total crimes in each decade.
* Then we visualized it using a scatter plot.

**Chart

Description automatically generated**

* We can see that crime rates in the 1990s were much higher than in the other two decades. We can also conclude from the graph that crime rates keep steadily decreasing through the decades.

**Question 3:**

* In this question we wanted to find out who the group that most hate crimes were against, how many were against people, and how many were against their property.
* We first made a new dataframe with all the groups stated and got the group that had the most hate crimes and visualized it.

**Chart, histogram

Description automatically generated**

* We could see from the graph that anti-jewish crimes were overwhelmingly higher than any other crimes.
* Having converted the crime type to numerical values came in handy here, we were able to find the number of crimes in each type as follows:

Text

Description automatically generated

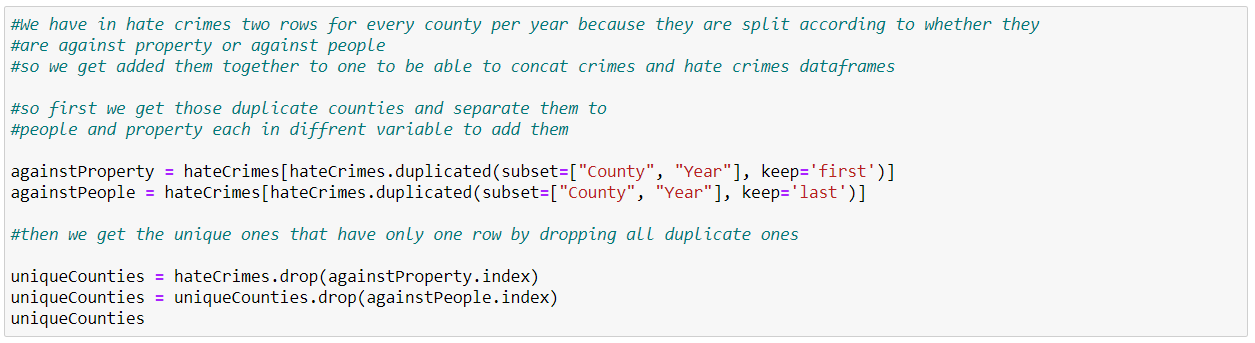
* Then we visualized it and concluded that crimes against Jewish properties are much more than crimes against Jewish people.

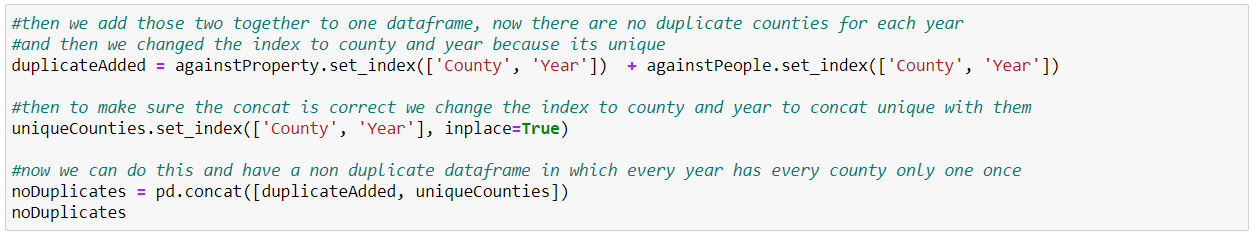
Chart, pie chart

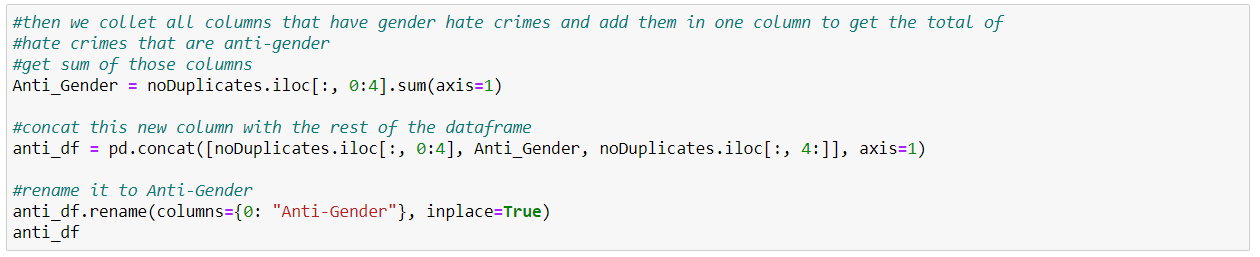
Description automatically generated

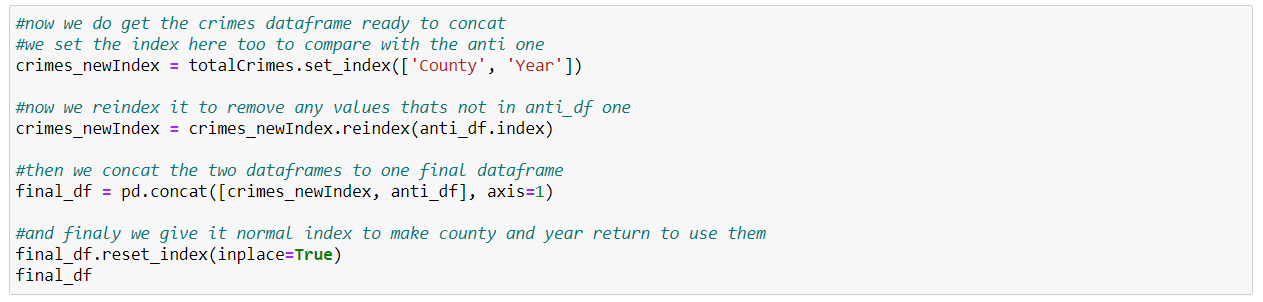
**Questions 4, 5, and 6.**

* In these questions, we needed to merge the crimes and hate crimes dataframes. To concatenate them, we needed to sort out the countries, years, and other repeated columns in both dataframes. We prepared them for merging as follows:









* After this, we had a dataframe of all counties, crimes, and hate crimes. We also summed up all hate crimes against people for their gender and added it to a new column “Anti-Gender”, which will be helpful to us in other questions. It was ready for us to start answering our questions.

**Question 4:**

* We wanted to know if there was a correlation between rape and gender-hate crimes.
* We extracted rape and gender crimes, calculated the correlation, and normalized it.

Graphical user interface, text, application

Description automatically generated

* Then we plotted it. The correlation turned out to be 0.65 which is quite high.

Chart, histogram

Description automatically generated

**Question 5:**

* To find if the countries with higher crime rates also had high hate crime rates, we did as we did in the previous question and extracted the two columns we need in a new df and renamed them to be more descriptive.

Graphical user interface, text, application

Description automatically generated

* We then calculated the correlation, normalized it, and plotted the results.

Chart, scatter chart

Description automatically generated

* The correlation was 0.8 which is very high, so we can conclude that yes, crimes and hate crimes do affect each other.

**Question 6:**

* Since we had the “Region” column which we converted to numerical values in the cleaning phase, we know if each county is in new york city or not. So we want to know how that difference in the region affects hate crimes against religious groups, in other words, is there more hate on religion in NYC or outside it (still within the state).

Text

Description automatically generated

* We extracted the columns that are anti-religion hate crimes and separated the crimes that occurred inside the city and those that occurred outside it.
* Turns out that hate crimes are much more inside new york city than outside it. In other words, NYC is the area with the most hate crimes in New York state.

Chart, histogram

Description automatically generated

**Project Two**

* Now, we’re going to scrape data from the website: <https://www.disastercenter.com/crime/nycrime.htm>, clean it, and integrate it with ours.

**Scrapping and cleaning:**

* First, we scrapped the data from the table we want. It was only the first table that we were interested in, so we used “find” to get it.

Graphical user interface, text, application, email

Description automatically generated

* We faced an issue with some values, some of them had “\n”, or “\xa0” which made it hard to get the value alone, so we solved this by replacing all that we didn’t want like this.

Text

Description automatically generated with medium confidence

* After adding each column in a list using a for loop, we added them to a df like this, and arranged the columns so they would match the data we have.

Text

Description automatically generated with medium confidence

**Integration:**

* We integrated the data by first dropping columns that we didn’t need anymore, extracting only the rows that we were missing in our old data and adding them. After additional processing, we had a new df called “new\_totalCrimes” with both our old and new data.

**Text

Description automatically generated**

**Answering remaining questions:**

**Question 7:**

* We wanted to know if the war had any effect on crime rates, so we made a heat map to see the difference in rates. And concluded that indeed the war had a great effect on crime rates.

**A picture containing rectangle

Description automatically generated**

**Question 8:**

* We wanted to see if the “me too” movement affected rape rates. So after visualizing it, we saw that it caused an increase in rape crimes.**Chart

  Description automatically generated**

**Hypothesis Testing:**

* We used a t test in our test since these are numerical values.

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

* We concluded that the alternative hypothesis is correct.

**Conclusion:**

After all our analysis, we’ve answered many questions that provided us with important patterns and insights.

* We found that the covid-19 pandemic caused a decrease in hate crime rates in new york.
* The 1990s had the highest crime rate in the last 3 decades, while the 2020s had the least rates. This tells us that fortunately, crime rates are steadily decreasing over time.
* We also found that unexpectedly, Jewish people face the most hate crimes in the whole state, especially their properties. This tells us that sadly, New York state is a relatively unsafe place for Jewish people.
* The correlation between gender and rape hate crimes is also quite high in the state. Meaning that the two crimes usually affect each other.
* Another correlation we found is overall crimes and hate crimes. The rate of overall crime in one place somehow affects hate crime rates, and vice-versa.
* Another important pattern we found is religious hate crimes. We found that inside New york city itself, hate crimes against religious people and properties were almost twice the crimes against religious people outside the city, but still inside the state. This tells us that since NYC is a very big and versatile city, it tends to be quite unsafe for most minorities.
* After scrapping and integrating new data, we found that the war years, from 1980 to 1988, had an overwhelming increase in crime rates in the state. Until now, crime rates have never reached this peak again.
* Lastly, we found that the “me too” movement that gained attention in 2017, caused a great increase in rape crimes. This is because the movement encouraged people to come forward about their experiences and report the crime, not because it caused an increase in the rape itself.