CST 383 – Final Project

Homicide Cases Analysis

Trenton Fengel, Melissa Graham, Ngoc Tran-Dao

Introduction

For our final project, we decided to analyze a dataset of homicide cases. We found the dataset on the website of the nonprofit organization Murder Accountability Project. This organization provides homicide investigators with the case data of thousands of homicides dating back to 1976. The Murder Accountability Project claims to be the most complete accounting of homicides available anywhere. A large portion of the homicide cases in the dataset remain unsolved.

Our goal for this project is to analyze the data and utilize the machine learning concepts we've learned throughout the course to predict the number of homicides in the current year 2022. We believe that the best way to make this prediction is to determine the number of homicides in the previous years and then use linear regression to predict the number of homicides in 2022. We hypothesized that the number of homicides would go down a bit after 2020; we are not entirely sure whether a solid correlation exists between the rise of Covid-19.

Selection of Data

The data that we will be working with is in the form of a comma-separated values (or .csv) file. There are a few datasets from the Murder Accountability Project website, but we downloaded the .csv file that includes the data of each individual homicide case.

There are also 30 columns that specify the various details of a homicide report. The first several fields include the ID code, as well as the state the homicide took place and the agency that filed the report. The next couple of fields include the solved status of the case as well as the year and month that the homicide occurred. Then, there are several columns that specify the age, sex, race, and ethnicity of both the victims and offenders. The last several columns specify the homicide weapon, the circumstances that led to the homicide, and the relationship between the victim and the offender.

We initially broke down the standard deviations, averages, and other aspects of the data:

	Year	Incident	VicAge	OffAge	VicCount	OffCount	FileDate
count	827219.000000	827219.000000	827219.000000	827219.000000	827219.000000	827219.000000	824709.000000
mean	1996.991040	28.947475	47.540888	352.717006	0.128523	0.185658	52603.878351
std	13.008957	110.180594	118.824307	456.035658	0.564474	0.596206	32519.244970
min	1976.000000	0.000000	0.000000	0.000000	0.000000	0.000000	10181.000000
25%	1986.000000	1.000000	22.000000	24.000000	0.000000	0.000000	30180.000000
50%	1996.000000	2.000000	30.000000	38.000000	0.000000	0.000000	40605.000000
75%	2008.000000	10.000000	42.000000	999.000000	0.000000	0.000000	82306.000000
max	2020.000000	999.000000	999.000000	999.000000	21.000000	40.000000	123197.000000

Then, we broke down data according to type:

```
 #   Column        Non-Null Count   Dtype  
---  ------        --------------   -----  
 0   ID            827219 non-null  object 
 1   CNTYFIPS      827219 non-null  object 
 2   Ori           827219 non-null  object 
 3   State         827219 non-null  object 
 4   Agency        827219 non-null  object 
 5   Agentype      827219 non-null  object 
 6   Source        827219 non-null  object 
 7   Solved        827219 non-null  object 
 8   Year          827219 non-null  int64  
 9   StateName     22468 non-null   object 
 10  Month         827219 non-null  object 
 11  Incident      827219 non-null  int64  
 12  ActionType    827219 non-null  object 
 13  Homicide      827219 non-null  object 
 14  Situation     827219 non-null  object 
 15  VicAge        827219 non-null  int64  
 16  VicSex        827219 non-null  object 
 17  VicRace       827219 non-null  object 
 18  VicEthnic     827219 non-null  object 
 19  OffAge        827219 non-null  int64  
 20  OffSex        827219 non-null  object 
 21  OffRace       827219 non-null  object 
 22  OffEthnic     827219 non-null  object 
 23  Weapon        827219 non-null  object 
 24  Relationship  827219 non-null  object 
 25  Circumstance  827219 non-null  object 
 26  Subcircum     32451 non-null   object 
 27  VicCount      827219 non-null  int64  
 28  OffCount      827219 non-null  int64  
 29  FileDate      824709 non-null  float64
 30  MSA           827219 non-null  object 
dtypes: float64(1), int64(6), object(24)
```

We also noticed that our data frame was too big to work with. It takes quite some time to load the file for the IDE to read. Therefore, we performed a drop function of the unnecessary columns to allow the machine to learn quicker. dropped a handful of columns to help with dealing with the volume and processing of relevant data.

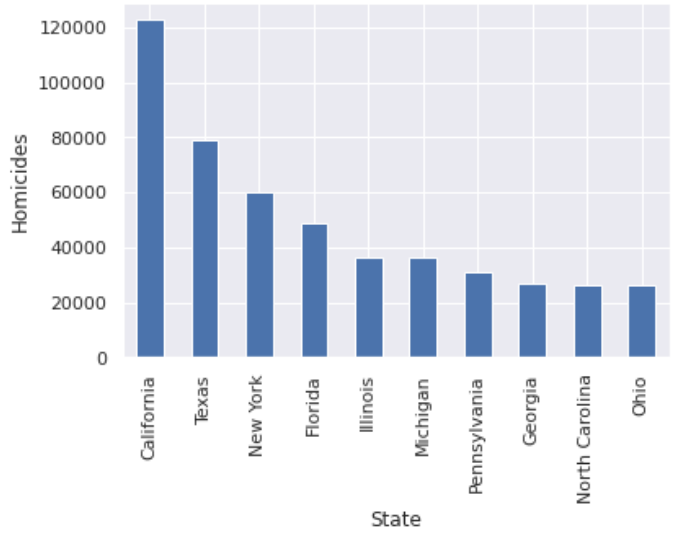
```
#drop the columns that we don't need
df.drop(['ID', 'CNTYFIPS', 'Ori', 'Agency', 'Agentype', 'Source', 'StateName', 'Month', 'Incident', 'ActionType', 'Subcircum', 'FileDate'
df
```

	State	Solved	Year	Homicide	Situation	VicAge	VicSex	VicRace	VicEthnic	OffAge	OffSex	OffRace	OffEthnic	Weapon	R
0	Alabama	No	1976	Murder and non-negligent manslaughter	Single victim/unknown offender(s)	30	Male	Black	Unknown or not reported	38	Unknown	Unknown	Unknown or not reported	Other or type unknown	
1	Alabama	Yes	1977	Murder and non-negligent manslaughter	Single victim/single offender	65	Female	Black	Unknown or not reported	62	Male	Black	Unknown or not reported	Other or type unknown	
2	Alabama	Yes	1977	Murder and non-negligent manslaughter	Single victim/multiple offenders	48	Male	White	Unknown or not reported	52	Male	White	Unknown or not reported	Handgun - pistol, revolver, etc	
3	Alabama	Yes	1977	Murder and non-negligent manslaughter	Single victim/single offender	27	Male	Black	Unknown or not reported	22	Female	Black	Unknown or not reported	Shotgun	
4	Alabama	Yes	1977	Murder and non-negligent manslaughter	Single victim/single offender	17	Female	Black	Unknown or not reported	21	Male	Black	Unknown or not reported	Knife or cutting instrument	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
				Murder and non-	Single				Not of				Not of	Personal weapons,	

Our goals altered a bit throughout initial work on the project; breaking down some of the data with graphs and other visualizations helped us iron out how to do the training required for our predictions.

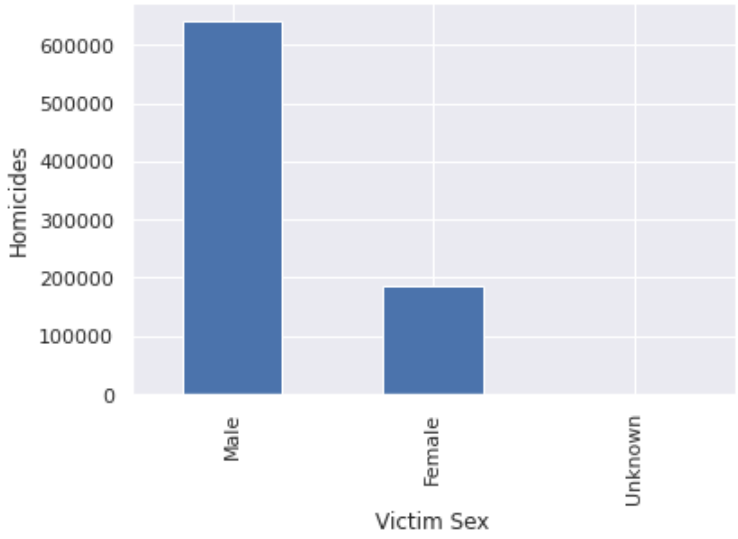
```
#this is a bar graph that represents the ten US states with the most homicides
df.value_counts(subset='State').sort_values(ascending=False)[:10].plot(kind="bar")
plt.xlabel('State')
plt.ylabel('Homicides')
```

```
Text(0, 0.5, 'Homicides')
```



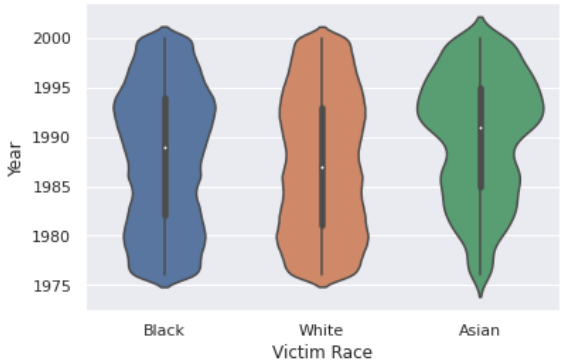
```
#this is a bar graph that represents the number male and female victims
df.value_counts(subset='VicSex').plot.bar()
plt.xlabel('Victim Sex')
plt.ylabel('Homicides')
```

```
Text(0, 0.5, 'Homicides')
```




```
#this is a violin graph that represents the number of victims over time, and is organized by race
graph1 = df[df['VicRace'].isin(['Black', 'White', 'Asian'])]
graph1 = graph1[(graph1.Year >= 0) & (graph1.Year <= 2000)]
sns.violinplot(data=graph1, x='VicRace', y='Year')
plt.xlabel('Victim Race')
plt.ylabel('Year')
```

```
Text(0, 0.5, 'Year')
```



After breaking down the data, we discovered that there are 827,219 homicide records in the dataset. Additionally, we learned that the dataset includes cases from the year 1976 to the year 2020. Another observation is that the minimum and maximum age for both the victims and offenders is 0 and 999 respectively. We noticed that, in the cases where the offender was unidentified, the age was set to 999. Given that the average offender age is 352.7, it seems the values of 999 are throwing off the calculations.

One thing that we noticed while analyzing the data was that the “StateName” and the “Subcircum” fields were mostly null values. To resolve this issue, we filled all the null values with the text “Not reported”.

For “OffAge” and “VicAge” the maximum is 999. We used a replace function to replace them with the median of their own columns.

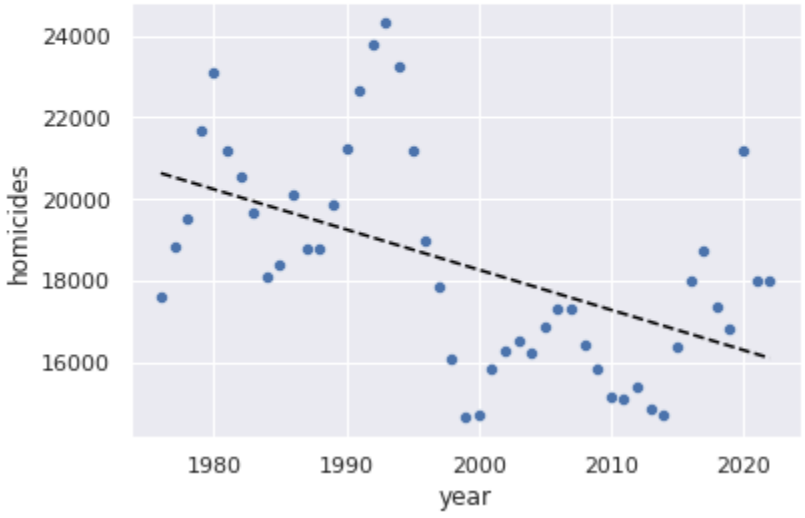
```
#fill the empty index with "Not reported"
df.Subcircum = df.Subcircum.fillna('Not reported')
df.StateName = df.StateName.fillna('Not reported')
#whenever the offender is not known, the offender age is set to 999. Change it to the median age.
df.OffAge = df['OffAge'].replace(to_replace=999,value=df.OffAge.median())
df.VicAge = df['VicAge'].replace(to_replace=999,value=df.VicAge.median())
df
```

Methods & Results

To make a prediction about the number of homicide cases in 2022, we started by creating a smaller dataset with two columns. One column stores the years and the other column stores the corresponding number of homicide cases for that year. After plotting the data it became apparent that the dataset only includes cases from the year 1976 to the year 2020. We worked around this by creating a row for 2021 and another for 2022 and we set the number of cases for both years to the median number of cases. After preparing the data we used linear regression to make predictions, but the curve didn’t seem to fit the data very well.

```
[ ] #plot the linear regression line
    predict = reg.predict(X)
    sns.scatterplot(x='year',y='homicides',data=df1)
    plt.plot(X,predict,color='black',linestyle='dashed')
```

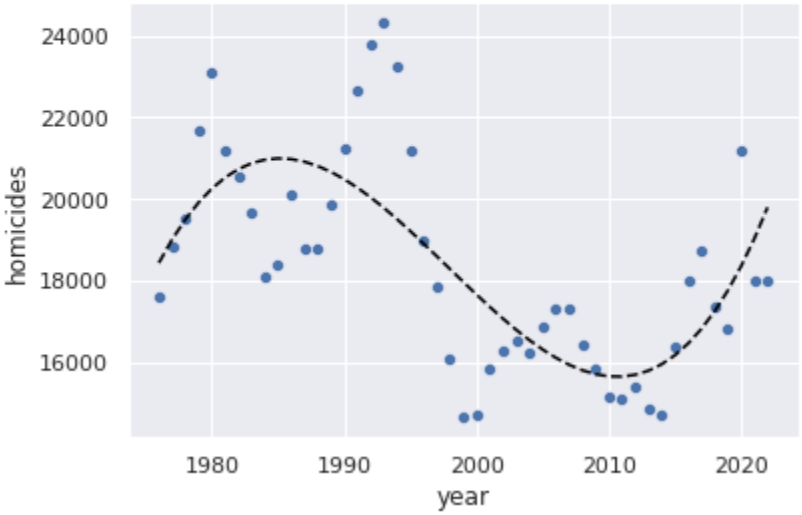
```
[<matplotlib.lines.Line2D at 0x7fb107d23790>]
```



We then decided to use the polynomial features with a degree of 3. After plotting the predictions of the data, we decided the curve fit a lot better. To figure out the predicted number of homicides we looked at the last value of the polynomial curve.

```
#plot the polynomial
predict = reg2.predict(X_poly)
sns.scatterplot(x='year',y='homicides',data=df1)
plt.plot(X,predict,color='black',linestyle='dashed')
```

```
[<matplotlib.lines.Line2D at 0x7fb0ffed2e90>]
```



After implementing the polynomial features of linear regression, we were able to create a prediction curve that seems to align with the data. The number of homicides predicted for 2022 is 19,794. This is down from the 21,180 homicides reported in 2020, but still relatively high compared to the homicide numbers of the years before the COVID-19 pandemic. This confirms our hypothesis that the predicted homicide numbers for 2022 will be lower than 2020, but still high possibly due to the pandemic.

Discussion & Summary

According to our results, the number of homicides in 2022 are about 19,794 incidents. Coincident enough with our research, there is an article that shows crime rates had increased in 2020 about 25%, and it was still climbing in the early of 2021. With our polynomial regression graph, the homicide numbers are increasing for the next couple years.

Some of the interesting facts from the data would be that men tend to be involved in homicides more than women. Those are from victim sex and offender sex. With some of the predictions and looking through the information from the data frame, we found that homicide cases are a consistent, slowly increasing trend for a couple more years. In 44% of all homicide cases, the victim and the offender are strangers to one another. It would have been a neat extra study to look at circumstances related to the 56% homicides where victims actually were acquainted [or more] with the offenders.