LA & NYC, A DECADE OF CRIME COMPARISON

2010-2020

SAMMY WHITE, ALEX KLEIN, MARLEY AKONNOR

IST 652

Crime Data:

Sources:
1) LA Crime Data
2) NYC Crime Stats

Crime, the number one argument for migration to suburbia. Univariate arguments are common in the court of public opinion but often crumble under the weight of scrutiny. The causality argument for crime in urban areas boasts similar nuance to the argument of shark attacks happening most often in shallow water, that's where the most people are. A uniquely infamous opportunity to add depth to this conversation has occurred over the last 2 years with the onset of the Corona Virus Pandemic. The volatility of crime has often ebbed and flowed with global events, the economy, and the change of political leadership making any specific variable(s) difficult to isolate. In this instance many extraordinary circumstances compounded, the implementation of a curfew, long term isolation, fear of the unknown, mass layoffs, a stagnated economy, and civil unrest born of racial injustice.

The selected data sets from Kaggle, LA Crime Data and NYC Crime Data, provide longitudinal examinations of crime trends over a decade. A greater perspective can be gleaned regarding the types of crimes committed, the amount of crime committed, and what time periods crime was most prevalent between two major cities on opposite coasts. The intention, a comparative analysis demonstrating either a parallel or, inverse relationship with crime patters. These two cities are amongst the largest and most densely populated urban areas in America which make them more prone to visible shifts in crime and the ideal candidates for examination.

Data Cleaning:

The first data set, LA Crime Data, had 2,118,203 rows and 28 columns. After the initial import, the data was soon reduced to the first 10 columns with the row count remaining intact. The rows were then renamed to a more explanatory format: "Record Number", "Date Reported", "Date", "Time Occurred", "Area Number", "Area Name", "District Number", "Part 1-2", "Crm Cd", "Desc". Of those rows, the data frame was further condensed, and the following columns were dropped: "Date Reported", "Area Number", "District Number", "Part 1-2", "Crm Cd." The final data frame for analysis consisted of 2,118,203 rows and 5 columns excluding the index.

```
In [23]: #cleaning up LA
df_LA = df_LA.iloc[:,:10]
LAcolnames = ["Record Number", "Date Reported", "Date", "Time Occured", "Area Number", "Area Name", "District Number
df_LA.columns = LAcolnames
df_LA = df_LA.drop(columns = ["Date Reported", "Area Number", "District Number", "Part 1-2", "Crm Cd"])
df_LA["City"] = "LA"
df_LA.head()
```

Out[23]:

	Record Number	Date	Time Occured	Area Name	Desc	City
0	1307355	02/20/2010 12:00:00 AM	1350	Newton	VIOLATION OF COURT ORDER	LA
1	11401303	09/12/2010 12:00:00 AM	45	Pacific	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA...	LA
2	70309629	08/09/2010 12:00:00 AM	1515	Newton	OTHER MISCELLANEOUS CRIME	LA
3	90631215	01/05/2010 12:00:00 AM	150	Hollywood	VIOLATION OF COURT ORDER	LA
4	100100501	01/02/2010 12:00:00 AM	2100	Central	RAPE, ATTEMPTED	LA

The second data set, NYC Crime Stats, underwent a similar revision. Initially, the data set began with 3,881,989 rows and 19 columns. After the import, the data was soon reduced to columns 2-5 with the row count remaining intact. The row “pd_desc” was also dropped. Two of the remaining columns were then renamed to clearer nomenclature: “Date” and “Desc”. “City” was left the unaltered.

```
In [9]: #cleaning up NYC
df_NYC = df_NYC.iloc[:,2:5]
df_NYC = df_NYC.drop(columns = "pd_desc")
NYCcolumns = ["Date", "Desc"]
df_NYC.columns = NYCcolumns
df_NYC["City"] = "NYC"
df_NYC.head()
```

Out[9]:

	Date	Desc	City
0	2019-01-26	SEX CRIMES	NYC
1	2019-02-06	CONTROLLED SUBSTANCES OFFENSES	NYC
2	2016-01-06	RAPE	NYC
3	2018-11-15	RAPE	NYC
4	2006-09-13	CRIMINAL TRESPASS	NYC

Since a decade is the time horizon being examined, all data prior to the year 2009 was removed. The final data frame for analysis consisted of 2,637,776 rows and 3 columns excluding the index.

```
In [35]: #remove everything from before 2010 in NYC df
df_NYC = df_NYC.loc[df_NYC["Year"]>2009]

In [36]: df_NYC.shape
Out[36]: (2637776, 4)
```

Once the data frames were cleaned and augmented to a satisfactory degree, an additional column, “Year”, was added to the LA Crime Data and NYC Crime Stats respectively.

In [10]:
```
#add year column to LA, make the year column an int
year = []
for i in df_LA["Date"]:
    y = i.split()[0]
    y = y[-4:]
    year.append(int(y))

df_LA["Year"] = year
```

In [11]:
```
#add year column to NYC, make the year column an int
year = []
for i in df_NYC["Date"]:
    y = i[:4]
    year.append(int(y))

df_NYC["Year"] = year
```

Finally, the data sets were combined into one merged data framed called “df_merged” by an outer join. The columns "Time Occured", "Area Name", "Record Number" were dropped from the new data frame.

In [13]:
```
#create a combined DF
df_merged = df_LA.merge(df_NYC, how="outer")
df_merged = df_merged.drop(columns = ["Time Occured", "Area Name", "Record Number"])
df_merged.head()
```

Out[13]:

	Date	Desc	City	Year
0	02/20/2010 12:00:00 AM	VIOLATION OF COURT ORDER	LA	2010
1	02/20/2010 12:00:00 AM	VIOLATION OF COURT ORDER	LA	2010
2	02/20/2010 12:00:00 AM	VIOLATION OF COURT ORDER	LA	2010
3	02/20/2010 12:00:00 AM	VIOLATION OF COURT ORDER	LA	2010
4	02/20/2010 12:00:00 AM	VIOLATION OF COURT ORDER	LA	2010

Comparison Questions:

All research questions are comparative in nature between the crime trends of these two major cities. These questions are foundational to the type of analysis and the methodologies used to find answers and patterns. The below 5 questions were proposed and addressed in this report:

Question 1: When was crime most prevalent in both cities?
Question 2: What were the top 10 most common crimes across both cities?
Question 3: When did crime in LA peak?
Question 4: What was the most common crime in LA in 2014?

Question 5: Crime, specifically burglary, was at its highest in 2017. What was the conversation on Twitter like during 2017 in LA regarding burglary?

Description of the Program:

Program 1:

The first program begins by importing the appropriate packages for analysis. In this instance the “csv” module is imported as well as “pandas” as “pd”. The data sets are then instantiated as variables and converted to data frames by the “pd.read_csv” command. Both data set were then cleaned by an initial selection of the most pertinent columns. The remaining columns were renamed in a manner that made them more easily comprehensible. If necessary, additional columns were dropped. At the end of these processes the “head()” command was used to look at a visual representation of the data sets. Unique to the NYC Crime Stats data, all data prior to 2009 was removed to keep the analysis in the selected time window. Both data sets used a for loop to create a year list, populate them with dates, and then append that list to the respective data sets. The data sets were finally combined with an outer join and three columns "Time Occurred", "Area Name", "Record Number" were dropped.

The next phase of the program was the analysis. At this point, the above research questions were introduced and addressed one at a time. Question 1, “When was crime most prevalent in both cities?” was answered by using the “groupby()” function with “Year” as an augment and creating a line plot to visualize the trend. Question 2, “What were the top 10 most common crimes across both cities?” was answered by counting the number of occurrences of different crimes and then finding the crime count for each year by city. This led to an examination of the climbing crime rate in LA showed by a line plot with counts by year. The program then makes note of the 2017 spike in crime in LA and drills down further into the most popular crimes in LA in 2017, the most popular crimes in LA overall, and the most popular crimes in LA before 2017. After this, the comparison with the most pervasive NYC crimes.

Since derivatives of theft/burglary appeared on the LA crime list multiple times, a further investigation was conducted which specifically created a list of burglaries of any type. The list was segmented and then recompiled into a data frame. Once again, a line plot was used to visually represent the count of crimes grouped by year. The program concludes with a pie chart showing the top 10 crimes in LA and their prevalence represented by surface area.

Program 2:

The second program beings in a similar fashion by importing “snscrape.modules.twitter” as “sntwitter” and “pandas” as “pd”. Next a tweet list is created to address the question of “How many tweets pertaining to the words burglary, burglar, burglarized, stolen, and or stole occurred between June1, 2017 and December 17, 2017?” The length of that list was 1001 tweets. The “tweets_list” was converted into a data frame and the column names 'Date', 'Tweet Id', 'Text', 'User', 'Likes', 'Retweets'

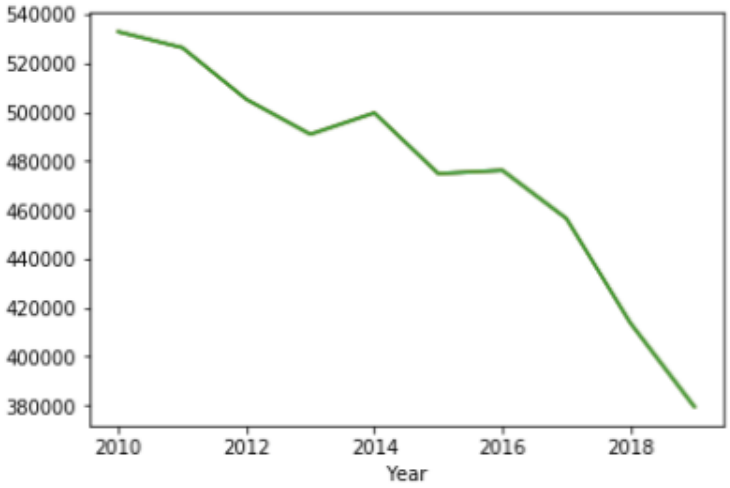
were created. A for loop is created to find all the hashtags and @ mentions used. Those hashtags were then counted and sorted in descending order from most to least used. The output of the top 10 was then printed. At this point "nltk" was imported to aid in better understanding the context. All tokens are created and the first 10 are shown. The tokens are converted to lowercase. After this the "nltk" English corpus of stop words are brought in. This list is used to determine the words that are removed as well as non-keywords. The 30 most common words and their counts are then printed.

Description of Output files and Analysis:

Output file 1: Output I - LA vs NYC Crime
Output file 2: Output II - Twitter Scrape LA Crime 2017

The first output file sought to answer exploration questions 1-4.

Question 1: When was crime most prevalent in both cities?



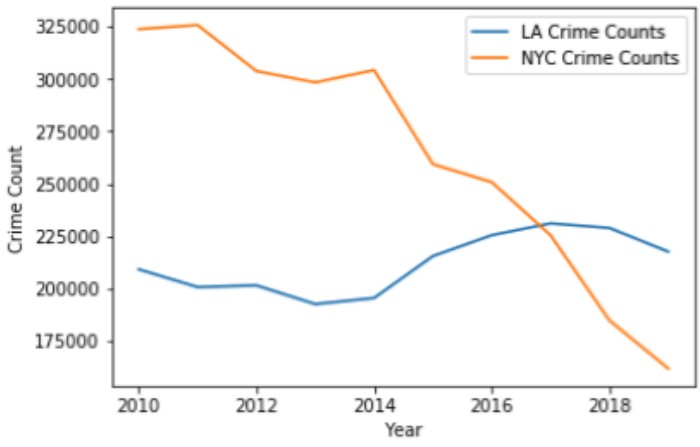
The above plot shows a stark contrast in crime between 2010 and 2020 for both cities combined. In 2010, there were approximately 538,000 crimes ultimately proving to be the peak of crime prevalence in both cities. There was a declining trend with a brief disruption in 2014 where crime peaked, dropped shortly after, plateaued, and then followed a steady decline until 2020. The next natural question for inquiry became question 2:

Question 2: What were the top 10 most common crimes across both cities?

```
df_merged["Desc"].value_counts()[:10]
```

```
DANGEROUS DRUGS                     507493
ASSAULT 3 & RELATED OFFENSES        276060
BATTERY - SIMPLE ASSAULT            190551
BURGLARY                            184308
OTHER OFFENSES RELATED TO THEFT     179166
ROBBERY                             164438
BURGLARY FROM VEHICLE               162182
VEHICLE - STOLEN                    159893
THEFT PLAIN - PETTY ($950 & UNDER)  149874
PETIT LARCENY                       140864
Name: Desc, dtype: int64
```

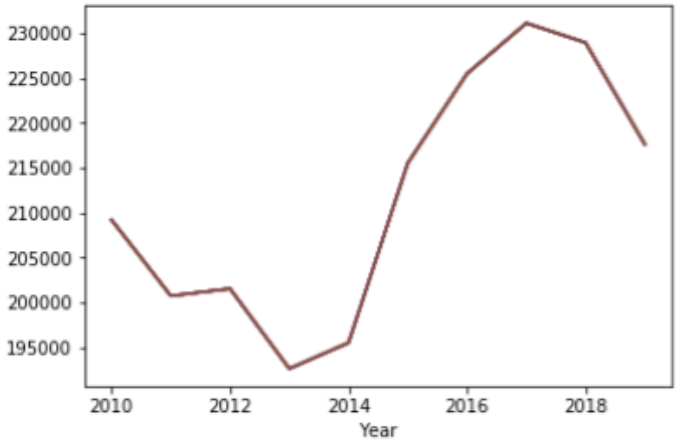
The largest count of crime committed is classified as “Dangerous Drugs” at 507, 493. This makes sense intuitively as the US has undergone various magnitudes of opioid epidemics. Drug related crimes tend to be more personal, isolation wise and can be committed in the privacy of a dwelling making them more likely to occur. The next two categories of crimes occupying spots two and three are “Assault 3 & Related Offenses” at 276,060 instances and “Battery – Simple Assault” at 190,551 instances respectively. The next two are par for the course in larger cities. People often time live right on top of one another and interact more frequently which would lead to more altercations. Interestingly, the next 7 crimes are all related to burglary/theft in some capacity. In the same vein as the most popular type of crime, theft is most often executed in isolation. It could simply be an extension of a nefarious part of human nature or, it could be an extension of poverty which is often prevalent in major cities. When people lack the means, they seek it elsewhere.



When the data is separated it is revealed that the initial line plot was heavily influenced by NYC’s crime trends. LA in that decade had considerably less crime until around 2017 when the overall crime volume in LA exceeded that of NYC handedly. This

unforeseen movement in the line was the impetus for a more in-depth look at when specially in the arch in occurred.

Question 3: When did crime in LA peak?



This plot allows for a clearer look at crime trends. It is apparent that crime jumped drastically from 2013-2017. In 2013 there were less than 195,000 crimes and that number rose alarmingly to 230,000 in 2017. Various sources at the time suggest that this was due to emboldened racial supremacist groups committing hate crimes toward Asians, who comprise a large portion of LA's population. All of this on the heels of Donald Trump's election to president at the time.

```
#most popular crimes in LA 2017
LA2017 = df_LA.loc[df_LA["Year"]==2017]
LA2017["Desc"].value_counts()[:5]
```

```
BATTERY - SIMPLE ASSAULT                19092
VEHICLE - STOLEN                        18791
BURGLARY FROM VEHICLE                   18067
BURGLARY                                15300
THEFT PLAIN - PETTY ($950 & UNDER)      14746
Name: Desc, dtype: int64
```

While this presupposition is a theory, the data suggests that there could be a correlation. When the LA data was broken down into the top 5 most highly occurring crimes in 2017, the number 1 crime was "Battery – Simple Assault." This alone without the context of racial motivation detailed in the crimes is not damning evidence but spurred further investigation.

```
#most popular crimes in LA overall
df_LA["Desc"].value_counts()[:5]
```

```
BATTERY - SIMPLE ASSAULT                190551
BURGLARY FROM VEHICLE                   162182
VEHICLE - STOLEN                        159893
THEFT PLAIN - PETTY ($950 & UNDER)      149874
BURGLARY                                147716
Name: Desc, dtype: int64
```

Most popular crimes in LA Overall

```
#most popular crimes in LA before 2017
LApre2017 = df_LA.loc[df_LA["Year"]<2017]
LApre2017["Desc"].value_counts()[:5]
```

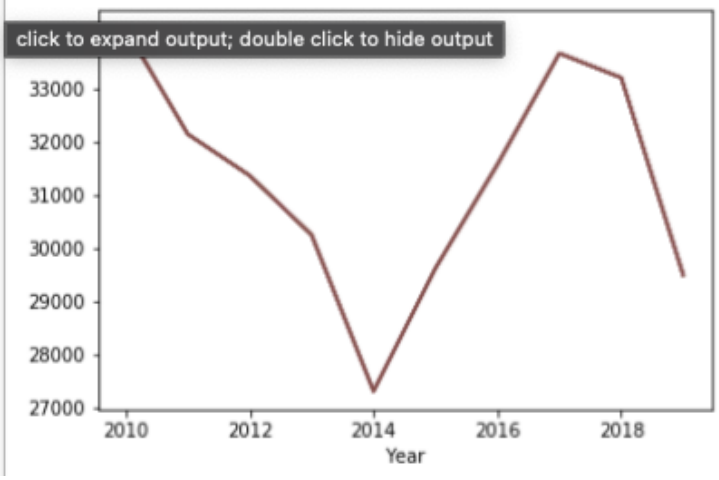
```
BATTERY - SIMPLE ASSAULT                133136
BURGLARY FROM VEHICLE                   109349
VEHICLE - STOLEN                        108710
BURGLARY                                104922
THEFT PLAIN - PETTY ($950 & UNDER)      104311
Name: Desc, dtype: int64
```

Most popular crimes in LA prior to 2017

If the spike is to be attributed to a shift in political power, then it could potentially stand to reason assault would have moved into the number 1 spot from a lower position. When the totality of crimes in LA was studied, it was shown that "Battery – Simple Assault" was still in the first spot by 30,000 instances with a total of 190,551. This number would suggest that assaults have simply historically been the most committed crime in LA, but a fair counter argument could have been that the number was the result of the massive 2017 outlier. The next step was to appraise assaults prior to 2017. The data still revealed a 24,000 instance gap between assault and the next most common crime of "Burglary from Vehicle." In summation, though it is possible to attribute the 2017 spike to political instigation, more and specific data would be needed to make any concrete conclusions.

Question 4: What was the most common crime in LA in 2014?

```
<matplotlib.axes._subplots.AxesSubplot at 0x3016ac9b0>
```



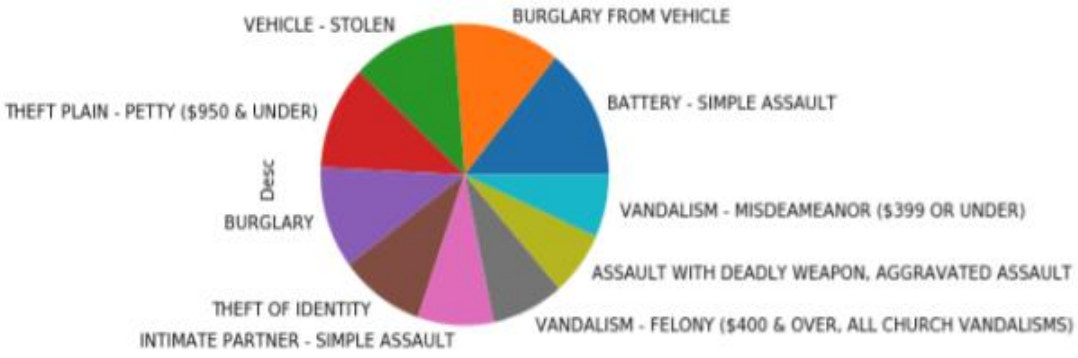
Moving in the opposite direction, 2014 was one of the lowest crime points in the decade for crime in LA. This piqued an inquiry into if this low may have also influenced a shift in the ordinal placement of the most common crimes over the decade.

```
df_LA.loc[df_LA["Year"]==2014]["Desc"].value_counts()[:5]
```

click to expand output; double click to hide output

```
BATTERY - SIMPLE ASSAULT                18414
THEFT PLAIN - PETTY ($950 & UNDER)      15750
BURGLARY                                13963
VEHICLE - STOLEN                        13699
BURGLARY FROM VEHICLE                   13080
Name: Desc, dtype: int64
```

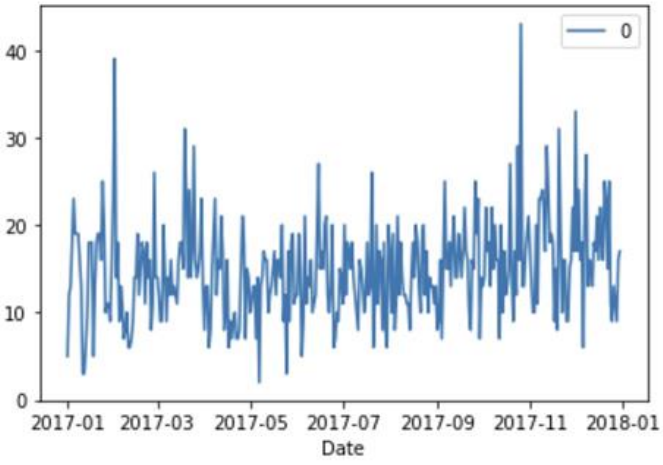
Even in 2014, the order and crimes remained consistent with “Battery – Simple Assault” still at the top. This solidified that it was the most committed crime in LA during the last decade. However, if all iterations of theft/burglary are combined, they comprise a massive portion of the pie.



Question 5: Crime, specifically burglary, was at its highest in 2017. What was the conversation on Twitter like during 2017 in LA regarding burglary?

```
Hashtag: #LA Count 22
Hashtag: #stolen Count 14
Hashtag: #HelpSamantha Count 13
Hashtag: #HelpShareMyStory Count 13
Hashtag: #LosAngeles Count 11
Hashtag: #hollywood Count 9
Hashtag: #Oscars Count 9
Hashtag: # Count 8
Hashtag: #Repost Count 8
Hashtag: #Dodgers Count 8
```

Hashtags are often used thematically at the end of a tweet to give it a category of sorts. While useful with full context, without it, makes having a proper understanding of what's going on quite difficult to ascertain. Above are the top 10 hashtags for tweets containing the words "burglary", "burglar", "burglarized", "stolen" and or "stole" in the year 2017. The insinuation is that perhaps 3 of the hashtags "stolen", "Help Samantha", and "HelpShareMyStory" pertain to burglary in the manner intended. The other 7, seemingly are the opinions of sports or movie fans who believe their team, or celebrity was robbed of a rightful victory.



The above line plot shows spikes in conversations pertaining to theft over the one-year time horizon from 2017 to 2018. January and November had the most noticeable spikes up to 40 with the median density around 18 conversations at any given time.

```
Message: Growing up, my biggest dream was to be photographed for GQ. Hahahahahhaaaa NOT. I stole someone else's dream Lolz! #GQ
#ThorRagnarok https://t.co/GSBGnkQE6T
Likes: 18171

Message: Tell me why my postmates guy cancelled my order when it was on the way... dude straight up stole my food lol... wtf
Likes: 4010

Message: $5000 TO ANYONE THAT FINDS MY LAPTOPS, HARD DRIVES, PHONE OR ANYTHING ELSE STOLEN. EMAIL - WillSingesShit@gmail.com
Likes: 2748
```

The top 3 most liked messages regarding theft are above. The first 2 unsurprisingly have comedic elements but pertain to theft in an abstract and literal way none the less. Not surprising for Twitter. The final tweet from Mr. Singesshit had a tone of sincerity and urgency however, also had the lowest number of likes.

```
stole: 3033 tweets
https: 2082 tweets
stolen: 1831 tweets
amp: 393 tweets
someone: 368 tweets
car: 345 tweets
got: 303 tweets
like: 296 tweets
one: 272 tweets
get: 232 tweets
```

Finally, the top 10 keywords. Unsurprisingly the world “stole” is first with 3033 tweets, “stolen” is in third place with 1831, and the rest save the word “car” began to get somewhat arbitrary.

Conclusion:

Crime has and will likely continue to be a concern surrounding major cities. In New York crime has been on a steady decline since 2010. LA featured a much lower crime rate initially but had a steady rise after 2013 that peaked in 2017 and began to decline thereafter. The predominant crime in NYC was “Dangerous Drugs” and in LA the number 1 crime was “Battery – Simple Assault” followed by an amalgam of theft-based crimes. Though the 2017 peak saw an increase in assaults, it was unclear what the prime causality was as assaults were the most popular crime for the entire decade including prior to 2017. The twitter exploration confirmed that these crimes were noticed and became a point of conversation on social media. The words “stole” and “stolen”

appeared in over 3,000 tweets. Though these crimes are an intrinsic part of society, the data shows the overall volumes are decreasing.