NYC Crime (Big Bang Data Group)

Duc Tran
dt2259@nyu.edu

Aigo Madakimova
am9634@nyu.edu

Bornita Das
bd1599@nyu.edu

ABSTRACT

Crime in the U.S took a drastic change during 2020 at the event of COVID-19 pandemic, the movement of Black Live Matter and the presidential election. With interest to investigate the statistics and trends of crime activities during this unprecedented time, our project focus on data analysis of crime happens in New York City, one of the biggest city in the country. The primary purpose is to look for details within data set to extract meaningful information that can be use to prevent crime in case these events happen again in the future.

ACM Reference Format:
Duc Tran, Aigo Madakimova, and Bornita Das. 2021. NYC Crime (Big Bang Data Group). In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The goal of the project is to extract meaningful information from data set related to crime in New York City. To accomplish this, we perform various steps from data exploration to data cleaning and data analysis. For data exploration, we searched through various data libraries such as: auctus.vida-nyu, Kaggle and Open Data Network. Our group performed data study and cleaning using OpenClean by NYU-VIDA. Lastly we analyze the data with PySpark SQL queries and visualize using Altair, Folium, Tableau and PyWaffle libraries. Our main goal is to answer question such as: Finding out top offense and location, crime rate between different period and location as well as overall statistic of crime between gender, ethnicity and age group.

2 DATA OVERVIEW

The NYPD Arrests Data (Historic) and NYPD Arrest Data (Year to Date) Dataset are a breakdown of every arrest in NYC by NYPD from 2006 to 2020. Here we focus from 2016-2020 for our analyze. The data is manually extracted every quarter and reviewed by Office of Management Analysis and Planning. Each record has 19 columns which are listed below. Historic (2006-2019) has 5 millions rows and Year To Date (2020) has 140413 rows.
NYC Population Census 2019 data and COVID Cases data from NYC Health Github was also used to answer established hypothesis.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

```
Schema
------
  'ARREST_KEY'
  'ARREST_DATE'
  'PD_CD'
  'PD_DESC'
  'KY_CD'
  'OFNS_DESC'
  'LAW_CODE'
  'LAW_CAT_CD'
  'ARREST_BORO'
  'ARREST_PRECINCT'
  'JURISDICTION_CODE'
  'AGE_GROUP'
  'PERP_SEX'
  'PERP_RACE'
  'X_COORD_CD'
  'Y_COORD_CD'
  'Latitude'
  'Longitude'
  'New Georeferenced Column'
```

Figure 1: Data Columns

Some columns that our project would like to focus on are: ARREST_DATE, PD_DESC, OFNS_DESC, ARREST_BORO, ARREST_PRECINCT, AGE_GROUP, PERP_SEX, PERP_RACE, Latitude and Longitude. These columns fit with the information we want to extract from the data set.

3 RESEARCH METHODS

3.1 Data exploration

To find the right data set fit with our goal, we searched through multiple data libraries such as: Kaggle, auctus.vida-nyu, Google data and opendatanetwork. The current data set is found in opendatanetwork.com and is also found in all these other data library platforms. Although there are many NYC crime related data, our team decided to work on NYPD Arrest data set because we are interested on severe crime that involves police.

3.2 Data extraction

Historic dataset is a huge data set ranging from 2006 to 2019, therefore NYU Peel HPC is used to run PySpark SQL queries to extract data that we interested in from 2016 to 2019. Automated script called run.sh executes all PySpark extraction scripts and makes the process easy. After the PySpark execution, 4 new csv contains information with labels of each year are created. Year To Date dataset does not need this extraction process because it only contains 2020. information.

3.3 Data study

Total we have 4 sub-dataset from Historic and 1 dataset from Year To Date. Many of the cleaning data are repetitive with some mild differences, therefore to keep the report simple, we will show the overall data study and cleaning process using Year To Date (2020).

Our first approach with understanding the data is through OpenClean which helps cluster data with similarity in columns and provide some overview structure of the data. The library

profiling feature helps us a lot with detect issues such as empty, distinct, uniqueness values.

	total	empty	distinct	uniqueness	entropy
ARREST_KEY	10000	0	10000	1.000000	13.287712
ARREST_DATE	10000	0	366	0.036600	8.396583
PD_CD	10000	1	166	0.016602	5.380796
PD_DESC	10000	2	158	0.015803	5.349247
KY_CD	10000	2	59	0.005901	4.501480
OFNS_DESC	10000	2	51	0.005101	4.290129
LAW_CODE	10000	0	421	0.042100	6.245291
LAW_CAT_CD	10000	108	4	0.000404	1.043776
ARREST_BORO	10000	0	5	0.000500	2.166298
ARREST_PRECINCT	10000	0	77	0.007700	6.113753
JURISDICTION_CODE	10000	0	21	0.002100	0.636239
AGE_GROUP	10000	0	5	0.000500	1.679094
PERP_SEX	10000	0	2	0.000200	0.643794
PERP_RACE	10000	0	7	0.000700	1.926892
X_COORD_CD	10000	0	5338	0.533800	11.349127
Y_COORD_CD	10000	0	5412	0.541200	11.367285
Latitude	10000	0	5574	0.557400	11.412375
Longitude	10000	0	5574	0.557400	11.412375
New Georeferenced Column	10000	0	5574	0.557400	11.412375

Figure 2: Data Profiling

According to the table through sample of 10000 data, the ARREST_KEY is in correct structure with 100% uniqueness, as well as the distribution of X and Y_COORD_CD, Latitude and Longitude columns. There some blank values in LAW_CAT_CD, which need further study to see if it makes a strong impact on the analysis.

Looking deeper into the data using K-NN cluster features, we found lots of spelling mistake across multiple columns which creates inconsistency for our analysis. Each K-NN cluster runs on each dataset found different spellings issues. Here is a sample of 2020 cluster runs.

```
Cluster({'KIDNAPPING & RELATED OFFENSES': 44, 'KIDNAPPING AND RELATED OFFENSES': 1})
Cluster({'INTOXICATED & IMPAIRED DRIVING': 2299, 'INTOXICATED/IMPAIRED DRIVING': 346})
Cluster({'OFFENSES AGAINST PUBLIC ADMINI': 4925, 'OFFENSES AGAINST PUBLIC SAFETY': 84})
Cluster({'OTHER STATE LAWS (NON PENAL LA': 536, 'OTHER STATE LAWS (NON PENAL LAW)': 5})
Cluster({'ADMINISTRATIVE CODE': 126, 'ADMINISTRATIVE CODES': 1})
```

Figure 3: Spelling issues 2020

In another study, we found some columns use abbreviation that results in ambiguity. Values in columns ARREST_BORO shows:

```
['B' 'M' 'Q' 'S' 'K']
```

Figure 4: Ambiguous letters

We also ran a test on ARREST_DATE to check if the data is in correct format and we didn't find any issue.

The rest of our study focus on some visualization to better understand the data and what it can offer and also decide what we can focus on. For example: With complex data that we don't understand fully such as longitude and latitude, we created graph HEATMAP of all geolocation data points to see if there are any weird location that is not in NYC. We also dropped columns that we won't use such as X_COORD_CD and Y_COORD_CD.

3.4 Data cleaning

For data cleaning, we created dictionary set of spelling errors as key and correct spelling as values. OpenClean offers a nice update function that clean the data according to our defined dictionary. Majority of our clean related to spelling that located in multiple columns across the data. Here are some changes that we have done for the data (full list is available on our Github repository):

- In order to fix the different representations discovered through clustering of same item in OFNS_DESC column we updated the spelling:
  - Change OTHER STATE LAWS (NON PENAL LA to OTHER STATE LAWS (NON PENAL LAW).
  - Changed KIDNAPPING and KIDNAPPING & RELATED OFFENSES to KIDNAPPING AND RELATED OFFENSES.
  - Changed ADMINISTRATIVE CODES to ADMINISTRATIVE CODE.
  - Changed INTOXICATED IMPAIRED DRIVING, INTOXICATED/IMPAIRED DRIVING to INTOXICATED AND IMPAIRED DRIVING.
- To make the data more interpretive we modified abbreviation:
  - Changed ARREST_BORO to full name of boroughs.
  - Changed PERP_SEX to Male, Female
  - Changed LAW_CAT_CD to full descriptions.
- Drop X_COORD_CD and Y_COORD_CD columns
- Issues in PD_DESC are also found and updated:
  - Standardized all variants of POSSESSION spelling.
  - Removed extra spaces in SALE
  - Standardized all variants of DRUG spelling.

3.5 Data check and testing

In order to make the issues clear besides just cluster, we queries the sub string to have a closer look. For example: KIDNAPPING AND RELATED OFFENSES spellings, we search for KIDNAPPING substring and found other offenses only named KIDNAPPING. Through this we will need to conclude which one we will uses to group our KIDNAPPING offenses.

```
KIDNAPPING & RELATED OFFENSES
KIDNAPPING AND RELATED OFFENSES
KIDNAPPING
```

Figure 5: KIDNAPPING spellings

After each cleaning, we also wrote testing script using python assert function to double check if the data is really fixed and show the final result.

```
check_fix = fix.loc[fix['OFNS_DESC'].isin(kid_val)]
assert all(check_fix['OFNS_DESC'] == 'KIDNAPPING AND RELATED OFFENSES'), "KIDNAPPING spelling is not fixed: " + check_fix['OFNS_DESC'].unique()
print("Successfully fixed: " + check_fix['OFNS_DESC'].unique())
```

```
['Successfully fixed: KIDNAPPING AND RELATED OFFENSES']
```

Figure 6: KIDNAPPING test

4 FIRST STEP CHALLENGE

There are many challenges that we faced while working with the data set. For data exploration, picking the right data set to fit our needs can be difficult. After obtaining the data, we spend tremendous amount of time to study, ask ourselves what we would like to learn from the data so that our data cleaning step fits with our goals and purposes. Through the introduction of OpenClean, we could use this powerful tool to study the data through set of features it provides to us such as profiling, sampling and visualizing portions of the data. The data cleaning step becomes a bit easier, however there are some gray areas we still don't know much about such as the relationship between key code (KY_CD) and offense description (OFNS_DESC), these 2 columns have data that does not connect. For example: with offense ROBBERY key code can be listed as 109 or 105 and not unique only to the description. Although this is an issue, we decide to ignore the KY_CD as we can't find official data on what they are represented. Our goal is to categorize OFNS_DESC as it tell us the most what the crime is about. The biggest challenge we face is to learn about OpenClean and what it can offer. Sometimes it can be frustrating to understand the package library and syntax. Luckily the documentation of functionalities is very well written which helped us execute our queries. Another challenge is to deal with Historic dataset which has 1GB size, initially we thought we could use the entire dataset but later learn that query this large set takes long time and makes our queries complicated. Luckily we have NYU Peel HPC to make this process a lot easier, and we decide to extract data based on year with HPC. One of the requirement is to make the code reproducible, finding the right tools and writing scripts to make the reproducible step comfortable to use is also another challenge we faced.

5 2020 DATA VISUALIZATION AND ANALYSIS

All visualizations are created using Altair, PyWaffle and Folium library. Details of code execution are commented in Google Colab file. We do our analysis by visually observe the data number changes with our charts.

5.1 Top 20 crime

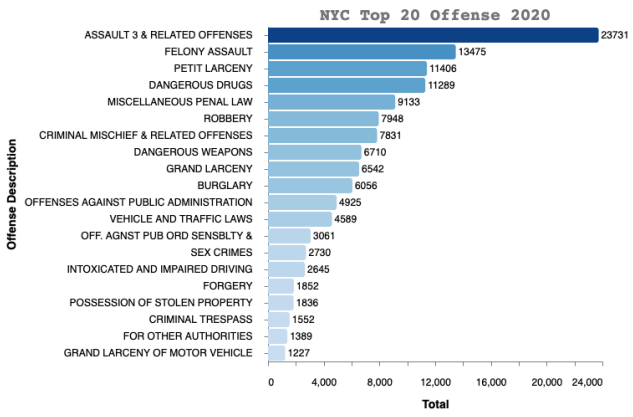

Figure 7: Top 20

Based on NYPD's offense description, there are total of 58 types of offense. Following visualization listed the top 20 crime types among all. The most arrested crimes are assault and related offenses, felony assault, petit larceny, dangerous drugs and penal law. Total crime number is 140413. In the figure above we included total top 20 crimes which add up to 129927 crimes and takes 92% of all committed crime.

Figure 8: Top 20 Percentage

Figure above is another look at the top 20 crime based on percentage using the waffle chart. From the overall look from blue to green, we can tell that assault and related offenses (19.6%), felony assault (9.6%), petit larceny (8.12%), dangerous drugs (8.04%) and penal law (6.5%) which is the top 5 crime takes up to 50% of total committed.

5.2 Top 5 crime VS total based on each month

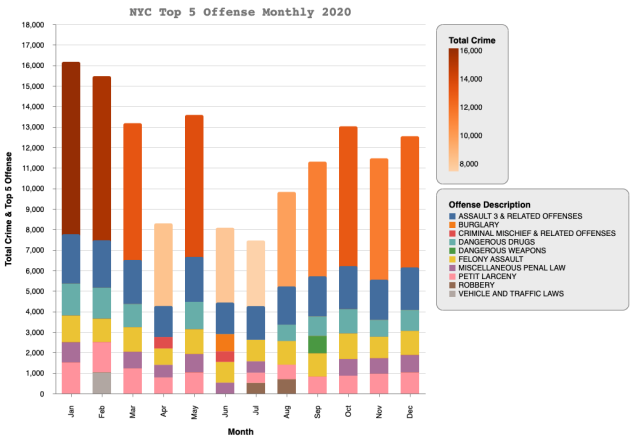

Figure 9: Top 5 Monthly 2020

The total crime label in oranges gradient, we can see that there is a huge drop rate in April, June, and July due to look down period. However March is not affected as the lockdown order by New York governor Andrew Cuomo happen in 20 March 2020.

We see some interesting changes of the top 5 in between each as January, March, October, November and December stays the same while other months has some minor changes. In Febuary, vehicle and traffic laws pops into the top 5 list. In July, August, robbery becomes one of the top 5 offense. In September, dangerous weapons makes it to the list. In accordance with 2020 news timeline, lots of protests that involves violent happens during those months.

5.3 Gender, ethnicity and age group

Figure 10: Top 20

From the chart, color represents age group and majority are from the age of 18-24 and 25-44 committed the top 20 crimes. According to the size, BLACK and WHITE HISPANIC have a larger number of committed crime as well as older age group especially WHITE HISPANIC. WHITE has larger number of under 18 suspects.

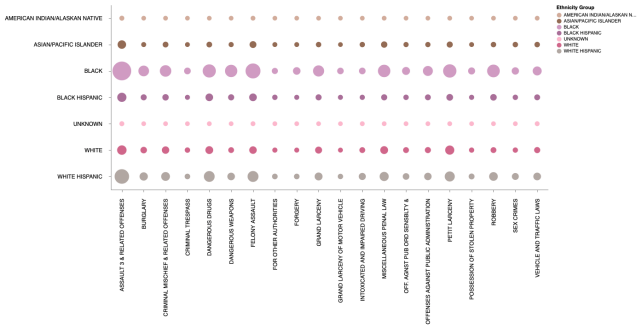

Figure 11: Ethnicity

Figure 11 focus only on total crime and ethnicity which shows that black ethnicity has higher levels of involvement in criminal offending than other ethnicity do. It appears to be considerable evidence that Blacks are disproportionately involved in serious crime. Yet most of this evidence is based on data from dataset 2020 of the general population without normalization.

Let's look at felony assault, felony assault is when a victim suffers a "physical injury". From our figure by hovering to the bubbles, we can see that 6795 of those were black suspects, 3283 white Hispanics, 1257 black Hispanic suspects, and 1179 white suspects. The gap between Black and Hispanics is large. Other races – which includes American Indians, Alaskan natives and Asians is smaller but still considerable.

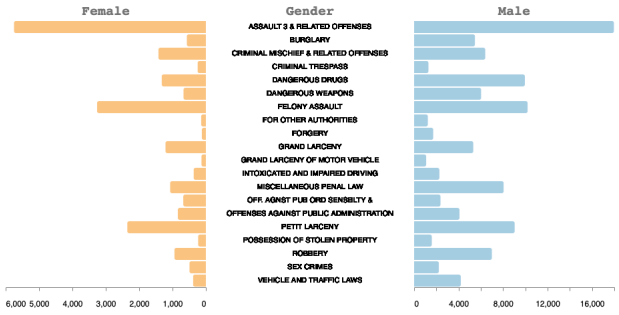

Figure 12: Gender

In figure 12, the number of gender overview of total crimes without consider ethnicity or age group. Male committed more crime than female. Although some offenses, both gender almost committed the same amount.

Our charts are interactive which enable us to scan and read the data in details with selected data interval filter. After scanning, the total of male and female between each ethnicity is quite proportional with the overall total. However, WHITE ethnicity has very interesting number of committed PETIT LARCENY which shown in figure 13.

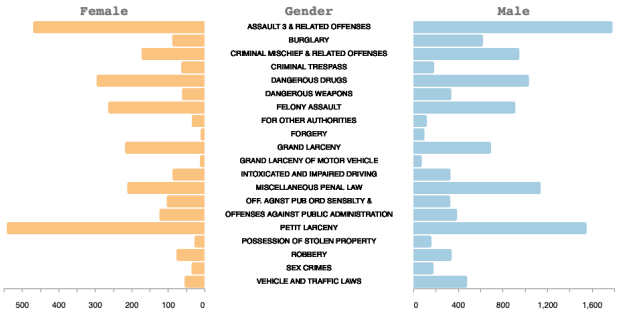

Figure 13: WHITE PETIT LARCENY

5.4 Police precinct

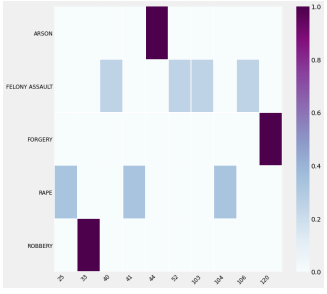

Figure 14: Police precinct

New York City is divided into 77 police precincts, and each precinct is divided into sectors that correspond, as much as possible, with the boundaries of actually established neighborhoods.

We can take a closer look at five crime categories in each Police Department district. Since we have a total of 58 crime categories, plotting all of them can be overwhelming and not informative. Instead, we plot only the five of the citywide crime categories and study their distributions among the 10 PD precincts and 5 crimes such as arson, felony assault, forgery, rape, and robbery. In the figure, we analyze that the 40 Precinct is, where felony assault frequently occurs. 33th, 44th, and 120th Precinct robbery target area among the 10 sample police precincts of New York City.

Forgery crime is most popular in the 120th district in Staten Island. Felony assaults are common in many districts while Robbery happened frequently in Manhattan.

5.5 Heatmap

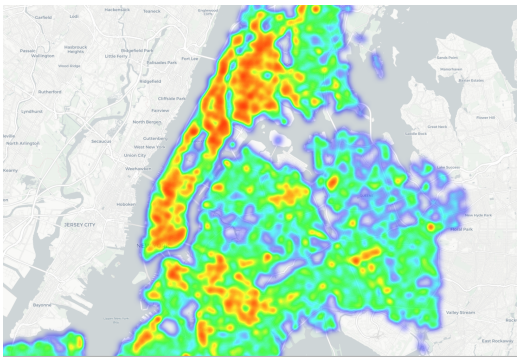

Figure 15: All crimes heat map

Figure 15 is created using Folium library which map crimes location depends on Longitude and Latitude. Yellow areas used to represent places with a little crime, light orange areas represented locations with medium rate of crime and a deeper orange color represented places with a huge rate of crime. We can assume it is dangerous spots in NYC. So we could see the crime distribution density over a map is a very informative using visualization helps people to detect dangerous spots. For example, we can see deep orange color in Manhattan, Bronx and Brooklyn.

5.6 Cluster map based on robbery area

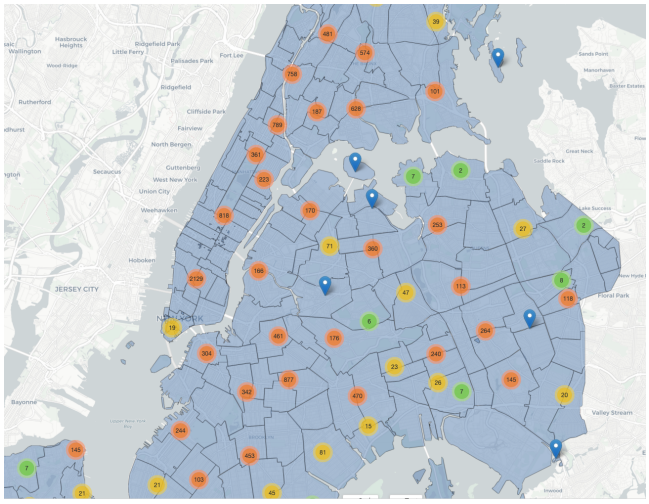

Figure 16: Cluster map based on robbery

As shown in the Figure below, where the round label with numbers is for crime hot-spots and the associated number of incidents. The marker cluster algorithm can help us manage multiple markers at different zooming levels, corresponding to various spatial scales or resolutions. When zooming out, the markers will gather together into clusters to view a broader geographic range on the map yet with a much coarser scale. When viewing the map at a high zooming level, the individual markers can be shown on the map to indicate the exact location of the incident. By using this interactive map, users can look up crimes on the streets. This can provide some preliminary guidance of the distributions of the associated incidents. Economic crises increase unemployment and have a greater impact on vulnerable groups, thus putting additional stressors on people and limiting their opportunities for financial stability, which may in turn trigger a spike in property crime in the later stages of the COVID19 pandemic. Burglary is one of violent crime that many researchers consider to be the best indicator of street-level and neighborhood 'safety' indicator. Moreover, burglary hot spots continue to be the primary 'target' for many of NYPD's crime control strategies. Figure shows many crimes are connected to subway stations, and Burglary occurs mostly in Manhattan. One of evidence we know from mass media when a peaceful protest in Washington Square Park took a violent turn. Protests against police brutality caused property damage and looting took place in the neighborhoods of SoHo, NoHo, East Village, Greenwich Village, and Union Square in Manhattan, as well as in parts of Midtown Manhattan and the Bronx.

5.7 Top 10 robbery area to avoid

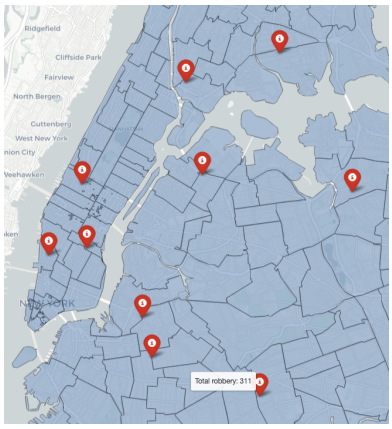

Figure 17: Top 10 robbery area

We map the location with our folium map to show the top 10 location that robbery incidents happen the most for people to be careful when travel. Brooklyn between Sutter Ave and Essex Street has the high number of robbery incidents with total of 311 cases. Follow by Mahattan West 10th street with 220 cases.

6 TRENDS DATA VISUALIZATION AND ANALYSIS FROM 2016 TO 2020

6.1 Monthly Trends

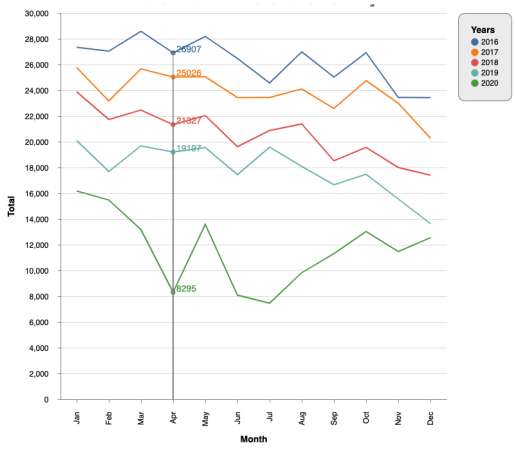

Figure 18: Crime rate 2016-2020 monthly

We begin with an assessment of overall crime trends. The dataset includes all crime incidents from 2016 to 2020. We would like to start our temporal analysis for the general trends of the crimes on the scale of years and months.

NYC shows a dramatic pattern during 2016. It had the highest rate of crime significantly higher in February 2016, dropped to the same level as last year in the transitional month of May, and then remained significantly in June through July. We have fluctuation during August through October, starting from October we can see how the crime rate goes down. The overall crime rate in 2017 - 2019 follows the slightly similar pattern. February is the month where less crime occurs and goes slowly down. So perhaps this is a seasonal phenomenon. The usual seasonal pattern is October months to occur in the winter and it turns into the low-crime November and December in the winter.

Crime decreased significantly during the early months of 2020. Please note an upward violent crime trend in May similar to what is being reported in the media. Crime peaked in October before declining through the rest of the year.

6.2 Weekday Trends

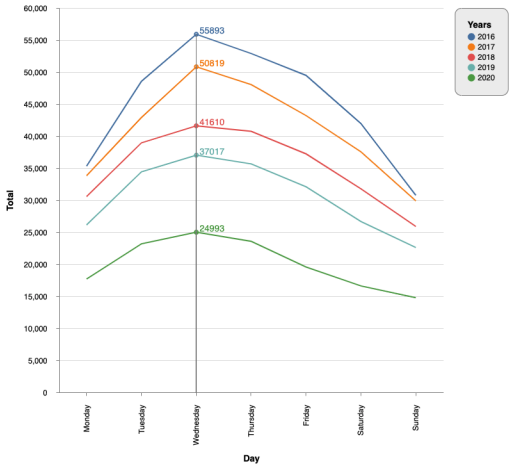

Figure 19: Crime rate 2016-2020 weekday

While the trend on the year and month scale can provide valuable information about the overall changing of the security environment, we are also interested in more detailed analysis that can possibly imply patterns that we can learn from. Day of the week is an important feature that has a potential impact on the crime categories. We can plot the pattern for each crime changing with the day of the week. Figure shows the overall crime rate and five consecutive years.

The overall graph indicates that Wednesday period has the highest crime counts. In addition, we can see the low crime rate patterns on weekends.

6.3 Borough Trends

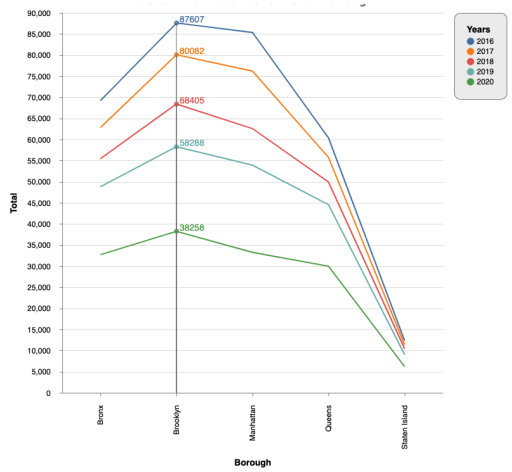

Figure 20: Crime rate 2016-2020 borough

People in NYC always say that some boroughs are safer than others. However, there's a disproportionate amount of people who live in Manhattan and Brooklyn with fewer living in Staten Island and Queens, so it is difficult to assume that any of these boroughs are safer in terms of different populations. The interactive map below shows the number of crime rates per population. Visually, we can tell that Staten Island is safer compared to other parts of the city. Brooklyn has the highest normalized crime rates than Manhattan.

6.4 Normalized Borough Trends

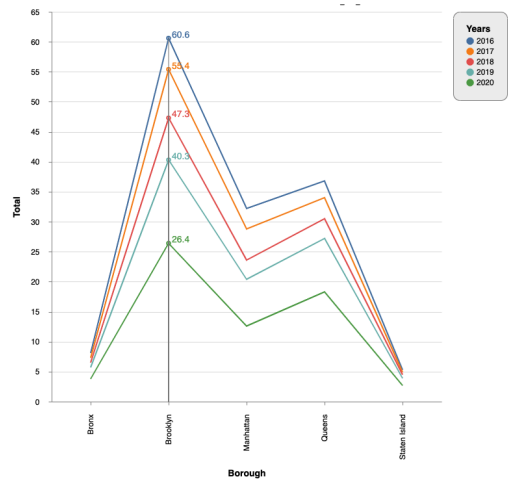

Figure 21: Crime rate 2016-2020 borough normalize per 1000 people

It is bias to consider crime rate without population of borough. Therefore we include an NYC borough dataset to calculate crime rate per 1000 peopple. The result is surprising showing Bronx has a low rate as Staten Island while Brooklyn is still the top dangerous place. This data eliminate the steoreotype the Bronx is a dangerous place.

6.5 Top 5 Crime Trend

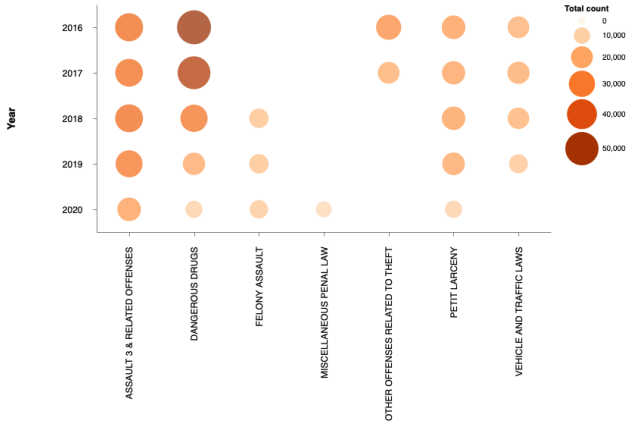

Figure 22: Crime rate 2016-2020 top 5 trend

Here, we can see the crime incident that has happened the most frequently in New York City during five years.There are seven type of crime in New York State: Assault 3 and related offences, dangerous drugs, felony assault, Miscellaneous penal law, Other offences related to theft, Petit Larceny and Vehicle and traffic laws.

From the bubble chart, Assault 3 and related offences is the most popular level of crime during five consecutive years. The second popular one is dangerous drugs. Surprisingly decreasing over the years and appears very low on 2020. Next popular crime in NYC is petit larceny there was an initial increase, followed by a decrease.

Past 3 years, felony assault crimes in the city have taken a dramatic turn. We could see the distributions of the past 3 years stayed around the same level.Noticeable increases Miscellaneous penal law such as resisting arrest, disorderly conduct, and public intoxication are serious charges. We can see in 2020 it was the most common crime and NYPD recorded 9133 cases. Other offences related to theft and vehicle and traffic laws showing significantly low-frequent crime in NYC crime rate by 2020.

6.6 Top 5 Crime Location Robbery Trends

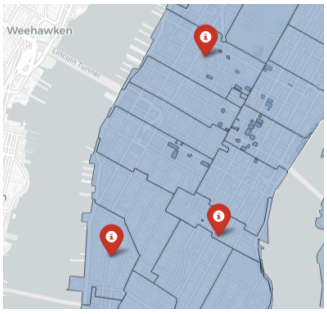

Figure 23: Manhattan area

The Manhattan area has 3 location in top 5 only in 2020.

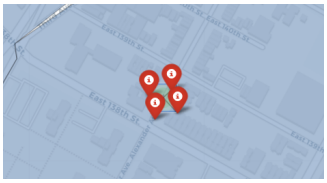

Figure 24: North Manhattan area

Up North New York at East 128th St has top robbery cases happen at the same places over 2017-2020.

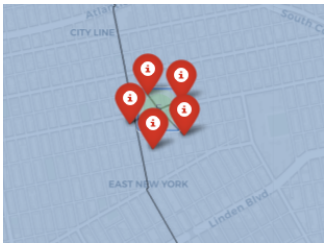

Figure 25: Brooklyn area

Another area in East New York lower Brooklyn has top robbery cases in consecutive 5 years.

7 ANALYSIS AND VISUALIZATION CHALLENGE

7.1 Challenges

- Altair is a new tool and we have lots of difficulty learning how to use it. The feature of stacking, concatenation, interaction are absolutely amazing. However, we faced difficult with buggy features and relied on Github to QA or issues to answer our questions.
- Data with multi dimensions are hard to represents and we have to think of creative way to best illustrate it. For example: gender, ethnicity and age group
- Sometimes it a bit hard to know what we want to queries.
- We also try many tools to fit our needs, Google Colab is one of the biggest contribution of our cleaning and visualization.

8 EXPERIMENTATION WITH TABLEAU

We combined the NYC Crime data for all the years 2016-2020 in Tableau and tried to analyse the whole set here.

8.1 Crime Trends Overall

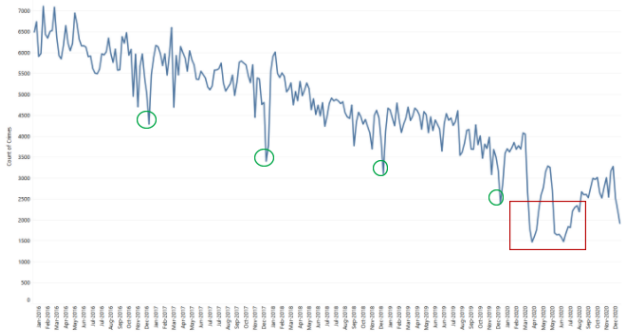

Figure 26: Crime Trends Overall

It is observed that the total no. of arrests has a decreasing trend over the time period 2016 – 2020. Seasonality prevails over the downward spikes in all the years with the exception of 2020. Arrests drop significantly in December every year, probably due to it being a holiday season with more public activity in general.

However, in 2020 the pattern changes. We see a downward trend in Mar 2020. We suspect that this irregularity is due to the onset of the pandemic.

To confirm our hypothesis, we juxtaposed the NYC Crime graph on the NYC COVID cases for 2020. It appears that COVID does appear to have some impact on the Crime rate. When COVID cases were seen rising the crime rates dipped and vice-versa.

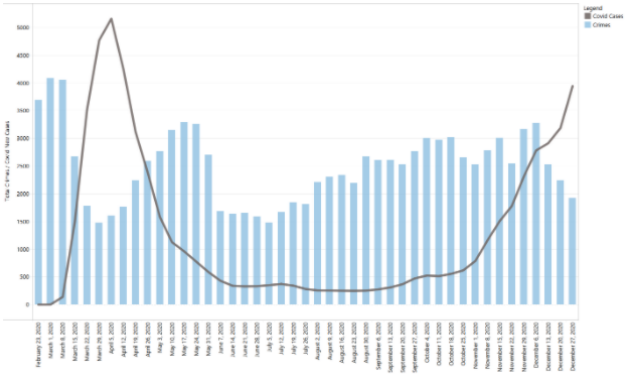

Figure 27: Covid and Crime Trends

Further probe on the chart Crime Rate 2016-2020 by LAW_CAT_CD, shows the decrease in overall arrests is attributed by a steady decline in MISDEMEANOR while FELONY has

remained constant over time. This shows serious charges have not seen much of a change over time.

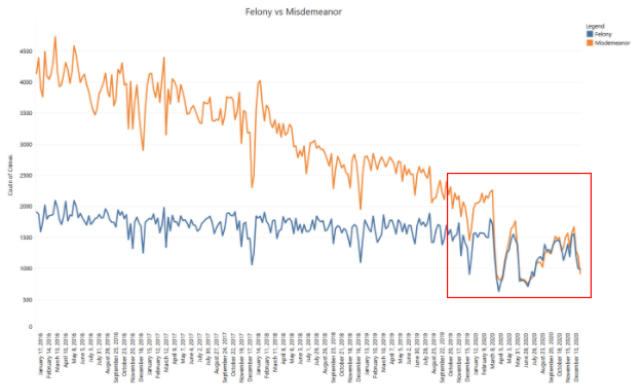

Figure 28: Felony VS Misdemeanor

Another fact worth noting is that, never again in the time frame of 2016-2020 have arrests under law category MISDEMEANOR and FELONY have aligned, with the exception of 2020.

This could lead us to hypothesize that either police/system was so overwhelmed by COVID related emergencies that only very severe cases of MISDEMEANOR led to arrests, maybe due to social distancing and other COVID related protocols. Or this could also mean people reached out for support to each other in the trying times. Also, public activity came to a virtual standstill which could have contributed to the sudden drop in arrests. This makes a good subject for a dedicated data analysis project as there is no one simple answer to this.

8.2 Crime Trends By Ethnicity

On dissecting the data at a more granular level of arrests committed by each race in every borough, we met an interesting graph.

The thing that stands out in this graph is Borough of Staten Island. The trend of arrests per borough gives a consistent trend over 2016-2020 for each race, with Staten Island being an outlier. White has a crime rate comparable to that of Black in this borough.

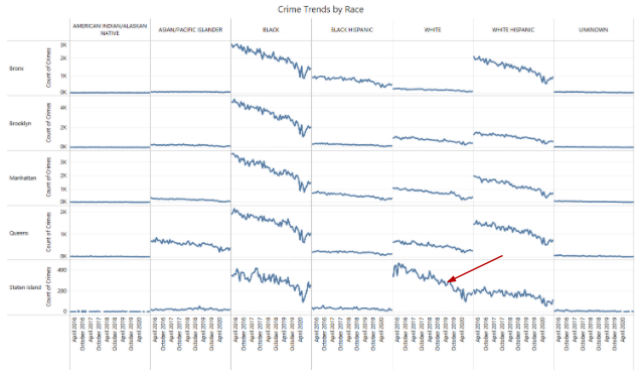

Figure 29: Ethnicity Trends

To investigate this deviation, we decided to normalise the arrests by population of each race in a borough. The population data was fetched from NYC 2019 Census. The NYC 2019 Census data had the population distribution by race organised in a different fashion than the NYC Crime PREP_RACE column. So, the normalisation effort was limited to 3 races, namely - BLACK, WHITE, AMERICAN INDIAN/ALASKAN NATIVE, which was satisfying for our purpose.

Bronx	AMERICAN INDIAN/ALASKAN NATIVE	0.145886
	BLACK	2.498957
	WHITE	0.181405
Brooklyn	AMERICAN INDIAN/ALASKAN NATIVE	0.408004
	BLACK	2.613589
	WHITE	0.395582
Manhattan	AMERICAN INDIAN/ALASKAN NATIVE	0.332583
	BLACK	5.791108
	WHITE	0.386982
Queens	AMERICAN INDIAN/ALASKAN NATIVE	0.477816
	BLACK	2.555573
	WHITE	0.319768
Staten Island	AMERICAN INDIAN/ALASKAN NATIVE	0.510051
	BLACK	4.394112
	WHITE	0.578191

Figure 30: Ethnicity Trends Normalized

The normalised rate does not support the observations of the graph rather presents a stark contrast to the absolute arrest count for Staten Island as seen in the Crime Trends by Race chart.

Thus, having analyzed this data, we could not arrive at a concrete conclusion on which race is prone to commit more crimes. When we do not know if the same individual is committing multiple crimes in the same borough we also do not know what proportion of the actual population of a particular race is involved in crimes. Addition of this layer of granularity could possibly impact the outcome of the normalized percentage of crimes by race in a substantial manner.

9 ASSOCIATION RULES ON OFFENSES RELATED TO ARRESTS BY BOROUGH

An attempt to fit NYC Crime data to MARKET-BASKET Analysis. In this part of our analysis we dive in to find if some types of offenses/crimes are likely to occur together in the same borough on the same calendar day by using Association Rules.

We consider NYC as the "market", each 'ARREST_DATE' to be "basket" and all 'OFNS_DESC'of that particular day as "items". The "baskets" were passed to the Apriori Algorithm to find the "frequent_items" to determine confidence in the association. This experiment was carried out for each 'ARREST_BORO'. We leveraged the apriori and association_rules method from mlxtend.frequent_patterns library to conduct this experiment.

Steps taken to mine Associative Rule from the NYC Crime data:

(1) Consolidate the items into 1 transaction per row - by grouping data on 'ARREST_DATE', 'OFNS_DESC' to retrieve count of 'ARREST_KEY' (rows).
(2) Each item in a transaction is one hot encoded - there are a lot of zeros in the data and any positive values are converted to a 1 and anything less than 0 is set to 0.

(3) Apply Apriori algorithm to each basket to find the frequent itemsets, while tuning the minimum support to see some relevant results. [minimum support = 0.5]
(4) Find Association Rules from the frequent datasets.
(5) Repeats steps 1-4 for each Borough in 'ARREST_BORO'.

antecedents	consequents	antecedent support	consequent support	support	confidence
(ASSAULT 3 & RELATED OFFENSES)	(BURGLARY)	1.000000	0.934426	0.934426	0.934426
(BURGLARY)	(ASSAULT 3 & RELATED OFFENSES)	0.934426	1.000000	0.934426	1.000000
(ASSAULT 3 & RELATED OFFENSES)	(CRIMINAL MISCHIEF & RELATED OFFENSES)	1.000000	0.969945	0.969945	0.969945
(CRIMINAL MISCHIEF & RELATED OFFENSES)	(ASSAULT 3 & RELATED OFFENSES)	0.969945	1.000000	0.969945	1.000000
(ASSAULT 3 & RELATED OFFENSES)	(CRIMINAL TRESPASS)	1.000000	0.655738	0.655738	0.655738

Fig. Association Rule for OFNS_DESC in Manhattan

antecedents	consequents	antecedent support	consequent support	support	confidence
(BURGLARY)	(ASSAULT 3 & RELATED OFFENSES)	0.745902	1.000000	0.745902	1.000000
(ASSAULT 3 & RELATED OFFENSES)	(BURGLARY)	1.000000	0.745902	0.745902	0.745902
(ASSAULT 3 & RELATED OFFENSES)	(CRIMINAL MISCHIEF & RELATED OFFENSES)	1.000000	0.989071	0.989071	0.989071
(CRIMINAL MISCHIEF & RELATED OFFENSES)	(ASSAULT 3 & RELATED OFFENSES)	0.989071	1.000000	0.989071	1.000000
(DANGEROUS DRUGS)	(ASSAULT 3 & RELATED OFFENSES)	0.931694	1.000000	0.931694	1.000000

Fig. Association Rule for OFNS_DESC in Bronx

antecedents	consequents	antecedent support	consequent support	support	confidence
(BURGLARY)	(ASSAULT 3 & RELATED OFFENSES)	0.781421	1.000000	0.781421	1.000000
(ASSAULT 3 & RELATED OFFENSES)	(BURGLARY)	1.000000	0.781421	0.781421	0.781421
(ASSAULT 3 & RELATED OFFENSES)	(CRIMINAL MISCHIEF & RELATED OFFENSES)	1.000000	0.978142	0.978142	0.978142
(CRIMINAL MISCHIEF & RELATED OFFENSES)	(ASSAULT 3 & RELATED OFFENSES)	0.978142	1.000000	0.978142	1.000000
(ASSAULT 3 & RELATED OFFENSES)	(CRIMINAL TRESPASS)	1.000000	0.418033	0.418033	0.418033

Fig. Association Rule for OFNS_DESC in Queens

antecedents	consequents	antecedent support	consequent support	support	confidence
(ASSAULT 3 & RELATED OFFENSES)	(BURGLARY)	1.000000	0.907104	0.907104	0.907104
(BURGLARY)	(ASSAULT 3 & RELATED OFFENSES)	0.907104	1.000000	0.907104	1.000000
(ASSAULT 3 & RELATED OFFENSES)	(CRIMINAL MISCHIEF & RELATED OFFENSES)	1.000000	0.994536	0.994536	0.994536
(CRIMINAL MISCHIEF & RELATED OFFENSES)	(ASSAULT 3 & RELATED OFFENSES)	0.994536	1.000000	0.994536	1.000000
(CRIMINAL TRESPASS)	(ASSAULT 3 & RELATED OFFENSES)	0.546448	1.000000	0.546448	1.000000

Fig. Association Rule for OFNS_DESC in Brooklyn

Figure 31

antecedents	consequents	antecedent support	consequent support	support	confidence
(ASSAULT 3 & RELATED OFFENSES)	(CRIMINAL MISCHIEF & RELATED OFFENSES)	0.870879	0.651099	0.579670	0.665615
(CRIMINAL MISCHIEF & RELATED OFFENSES)	(ASSAULT 3 & RELATED OFFENSES)	0.651099	0.870879	0.579670	0.890295
(ASSAULT 3 & RELATED OFFENSES)	(DANGEROUS DRUGS)	0.870879	0.579670	0.521978	0.599369
(DANGEROUS DRUGS)	(ASSAULT 3 & RELATED OFFENSES)	0.579670	0.870879	0.521978	0.900474
(ASSAULT 3 & RELATED OFFENSES)	(FELONY ASSAULT)	0.870879	0.708791	0.620879	0.712934

Fig. Association Rule for OFNS_DESC in Staten Island

Figure 32

The outcome of the experiment came to a similar consensus for most boroughs. Among 'OFSN_DESC' "BURGLARY, ASSAULT3 RELATED OFFENSE" appeared to be very closely associated to each other with very high confidence, indicating they are likely to occur on the same day, if anyone of them occurs on a given day. In simpler terms, when a burglary happens it is likely that someone/victim will be assaulted in some form to carry out the crime.

10 CHALLENGES

While analysing the data we noticed sudden deviations and came up with certain hypotheses. To prove the hypotheses additional research or scavenging of other datasets was needed.

For the hypothesis that abnormal drop in crime rates in Mar 2020, we had to find a COVID dataset which was easy to merge with the NYC Crime Dataset in terms of structure of data. While the hypothesis provided for a sudden decline in arrests in the MISDEMEANOR law category during 2020, we could not find any simple, direct answer on the internet to substantiate it.

The experiment with Data Mining was challenging from the standpoint of morphing our dataset to fit the Market-Basket-Analysis prototype. After testing the transformation of different units as "basket" and "items" both conceptually and through code, we could finally arrive with the arrangement that we have run our final experiments with.

11 GITHUB

https://github.com/duketran1996/NYC-Crime

12 ONLINE RESOURCES

Data set Year To Date: https://www.opendatanetwork.com/dataset/data.cityofnewyork.us/uip8-fykc

Data set Historic: https://www.opendatanetwork.com/dataset/data.cityofnewyork.us/8h9b-rp9u

Data set borough population: https://www.opendatanetwork.com/dataset/data.cityofnewyork.us/xywu-7bv9

Data set covid cases: https://github.com/nychealth/coronavirus-data/blob/master/trends/data-by-day.csv

Data set NYU population:
https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork,bronxcountybronxboroughnewyork,kingscountybrooklynboroughnewyork,newyorkcountymanhattanboroughnewyork,queenscountyqueensboroughnewyork,richmondcountystatenislandboroughnewyork/PST045219#

Market basket:
https://pbpython.com/market-basket-analysis.html

Open clean documents: https://openclean.readthedocs.io/

Altair documents: https://altair-viz.github.io/index.html

PyWaffle documents:
https://pywaffle.readthedocs.io/en/latest/index.html

Folium documents:
https://python-visualization.github.io/folium/