NYC Crime Data Analysis

Anurag Arora[1]
Stony Brook University
ID: 111425080

Bhavesh Goyal[1]
Stony Brook University
ID: 111482386

Renu Rani[1]
Stony Brook University
ID: 111482474

Shayan Ray[1]
Stony Brook University
ID: 111424665

Abstract

In this project, we have analyzed the NYPD Complaint Data from NYC Open Data. Our objective is to empower NYPD with exploratory data analysis and predictions so that they utilize their time and energy in the most effective way. That would result in better hot-spot policing(surveillance of crime prone zones), judicious staff deployment across boroughs, time-dependant vigilance and narrow down plausible root-causes in crime trends.

1 Introduction

New York City aka The Big Apple is the most densely populated city in United States as of 2017. Manhattan (the NYC Borough) as we all know is the heart of the 'Big Apple'. It is one of the world's largest commercial, financial, cultural and entertainment centers. The list of eminent buildings and places of interest is endless, therefore its protection is of supreme importance to the nation. Now that makes the job of the NYPD (New York Police Department) astoundingly challenging. This is enough motivation for us to do exploratory crime/complaint data analysis and empower these policemen with effective information to combat crime.

1.1 Datasets Overview:

DATASET 1 : NYPD-Crime Data :

- Scope: 2016 , 2017.
- Rows: 468,000 (approx)
- Features/Columns: 23
- File Size: 131 MB

Description: Each record has attributes consisting of complaint information(ID, FROM and TO DATE-TIME), offense type, police department responsible for handling it, borough of occurence, latitude and longitude of precise crime location etc.

DATASET 2: Liquor Authority - Active Licenses(issued for NY state)

- Scope: 2017.
- Rows: 51,600 (approx)
- Features/Columns: 21
- File Size: 13 MB

Description: Each record has attributes like license information(ID, TYPE, AGENCY, EFFECTIVE and EXPIRY dates), county where it was issued and premises information (NAME, ADDRESS etc).

1.2 Topics Overview:

1. Predict the number of crimes in NYC boroughs in the next 2 weeks by the hour so that NYPD can deploy staff judiciously across boroughs for the next 2 weeks.

2. Does opening a liquor store in an area impact the Spatial Crime Rate ? Regulating Police patrolling, and identifying a safe(r) place to buy liquor.

3. Are number of hourly crimes happening in 2017, normally distributed ? This would help NYPD be more vigilant at peak hours

4. Is there a relation between crime rate and

poverty? Can we debunk the view that poor people who take food stamps commit more crimes.

1.3 Results Summary:

1. Rolling Auto-Regression with p-value = 6 or 20 seems to perform a better job than rest of the methods applied.

2. Liquor store indeed affects crime rates in their vicinity.

3. Number of crimes per day are normally distributed

4. Schedule of Food Stamp distribution has no relation with crime rate.

2 Dataset Details

For this project, we used two datasets. The major dataset was **NYPD Complaint Map** and the another dataset that we used for our hypothesis "Affect of Liquor Licenses" is **Liquor Authority Quarterly List of Active Licenses Economic Development**.

2.1 Dataset 1: NYPD Complaint Map

This dataset includes all the valid felony, misdemeanor, and violation crimes that happened in the New York City. For our project, we used data of 2016 and 2017 years. This dataset has total 23 columns and around 468K rows which makes it of 131MB (1,). Details of columns descriptions are in Figure 1

2.2 Dataset 2: Liquor License Dataset

The dataset includes the licenses issued for the entire NY State and we have limited our study only to the NYC area. It has 21 columns and around 51.6K Rows i.e. number of licenses issued in NY state. The size of this dataset is around 13MB (2,).

3 Topic 1: Predict the number of crimes in NYC boroughs in the next 2 weeks by the hour

The objective is to identify boroughs and the time of the day in the next 2 weeks with higher risk for proactive patrolling by the NYPD. It is not about overwhelming the boroughs with officers by the hour. The idea is to change the shift of these officers to cope up with the ever-changing landscape of crime in one of the busiest cities in the world. Consequently, enable the NYPD with predictive information to better use their time and energy and improve their efficiency. Note: Experiments were conducted across 5 NYC Boroughs. Only Manhattan data was demonstrated in the presentation. This report includes all 5 borough prediction analytics.

Column	Description
CMPLNT_NUM	Randomly generated persistent ID for each complaint
CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
CMPLNT_TO_DT	Ending date of occurrence for the reported event, if exact time of occurrence is unknown
CMPLNT_TO_TM	Ending time of occurrence for the reported event, if exact time of occurrence is unknown
RPT_DT	Date event was reported to police
KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
JURIS_DESC	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
BORO_NM	The name of the borough in which the incident occurred
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

Figure 1: Columns description of NYPD Complaint MAP dataset

3.1 Methodology

1. **Data Cleaning:** NYPD Complaints from 2017 were considered for all experiments. Inconsistent years(most likely data entry errors)(such as 1846, 1017) and outliers were removed(using Tukey's Rule on CMPLNT_FR_DT column) before performing analysis. Further data segregation carried out was as follows:

   - *Borough specific dataset:* Since these predictions were at a borough level the original 2017 dataset was segmented into 5 smaller datasets each specific to one of the NYC boroughs namely, Manhattan, Bronx, Brooklyn, Queens

CMPLNT_NUM	CMPLNT_FR_DT	CMPLNT_FR_TM	CMPLNT_TO_DT	CMPLNT_TO_TM	RPT_DT	KY_CD
579468609	12/31/2017	23:55:00	10/07/2017	11:30:00	12/31/2017	118
822699000	12/31/2017	23:30:00	09/24/2017	20:15:00	12/31/2017	344
187081621	12/31/2017	23:10:00	12/31/2017	23:15:00	12/31/2017	578
967006956	12/31/2017	23:05:00	12/31/2017	23:15:00	12/31/2017	118

Figure 2: Snapshot of NYPD Complaint MAP dataset

License Serial Number	License Type Name	License Class Code	License Type Code	Agency Zone Office Name	Agency Zone Office Number	County Name (Licensee)	Premises Name
1000080	GROCERY STORE BEER	122	A	New York City	1	BRONX	ERIDANIA CORP
1000090	GROCERY STORE BEER	122	A	New York City	1	BRONX	1098 MINI MARKET INC
1000099	GROCERY STORE BEER	122	A	New York City	1	BRONX	J & A FRUIT & GROCERY CORP
1000167	GROCERY STORE BEER	122	A	New York City	1	BRONX	ALTAGRACIA GROCERY CORP

Figure 3: Snapshot of Liquor License Dataset

and Staten Island.

2. **Data pre-processing:** For each borough dataset, the data was grouped by the ́hourúsing CMPLNT_FR_DT and CMPLNT_FR_TM alongside a count of 1 per entry. (For eg: 2017-01-01 00:01:00 - 1)

3. **Train-Test split:** First 50 weeks of 2017 dataset was considered as the training dataset for each borough. It contains 8424 samples. The last 2 weeks was used as the test dataset. It contains 336 samples to predict in the future. Basically 365 days * 24 hours = 8760 samples based on the hourly grouping of time-series data as part of the pre-processing step. (8424 + 336)

4. **Model accuracy metric:** RMSE (Root Mean Square Error) is the industry standard for accuracy estimator in many prediction problems. It has been used here for all the experiments under this topic.

3.2 Techniques Used : 3+1 Hypothesis/Experiments for all 5 NYC boroughs

1. **Seasonal Last Observed:** The purpose of the seasonal last observed method is to isolate the recurring trends. In other words, decompose the series into the trend effect, seasonal effects, and remaining variability.

2. **EWMA (Exponentially Weighted Moving Average):**It improves on seasonal and moving average methods by assigning weights to the periodic trends.It facilitates using larger data sample size and assigns greater weights to more recent observations.

3. **AR (Auto-Regression) - Rolling Forecast(Linear Regression used to compute Betas) :** It specifies that the output variable depends linearly on its own previous values and on a bias term. Contrary to the moving-average model, the autoregressive model is not always stationary.

4. **ARIMA (Autoregressive Integrated Moving Average) - Rolling Forecast :** Non-seasonal ARIMA model has been tried out in this topic. It is denoted by ARIMA (p,d,q) where parameters p, d, and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model.

3.3 Results:

Note(addressing Instructor Feedback): The hyper-parameters used in various predictive techniques below were initially selected based on trial and error. [started off with some high and

low values] Once the parameter values started to produce acceptable RMSE results [the lower RMSE the better], techniques like Gradient Descent was employed manually. The RMSE values for these parameters were gathered and plotted in excel to represent the gradient descent curve over the loss function vs parameter values. The lowest of these values were chosen as the preferred hyper-parameters. Adjacent values were also demonstrated for comparison purposes.

- Graphical representation of original vs predicted values for the next 2 weeks across Manhattan using various time-series predictors and hyper-parameters: (**Figure 4,5**)Note: Additional graphs for all boroughs are available in the github repository.

- Comparison Summary of various time-series predictive techniques across all 5 boroughs: (**Figure 6**)

3.3.1 Conclusion

In this hypothesis, out of all the time-series prediction methods tried out, Auto-regression(Rolling) with p value = 6 and 20 gave the best prediction performance for the test dataset used. Non- Seasonal ARIMA came a close second among the other techniques.

4 Topic 2: Affect of Liquor Licenses

Liquor Stores are a common commodity across crowded neighbourhoods such as NYC. But does every new store make an impact on its surrounding by increasing the number of crimes in the vicinity ? We took a probabilistic approach to analyze effect of every such store opened in 2017 to make a conclusive remark on the effect of its surrounding crime reports. We also present a study to analyze effect of spatial and temporal variation of liquor store and licenses issued with their relationship to the crime reports lodged.

4.1 Methodology

1. **Data Cleaning:** For testing our hypothesis, we filtered the complaints and licenses issued only in 2017. We identified some rows with inconsistent years(such as 1846, 1017) and we removed such entries from the database. Then we further cleaned each of the datasets(Complaint reports and liquor licenses) as follows:

   - *Complaint Dataset:* Location being a crucial aspect in our hypothesis, we filtered out the reports with the location information not available from the year 2017.

   - *Liquor License Dataset:* The dataset was issued for the entire NY State and we have limited our study only to the NYC area. We took steps to filter licenses issued in 2017 in only NYC area and also identify and remove any inconsistent license issue dates. (Such as those found in Crime reports).

2. **Data Pre-Processing:** One of the challenges to relate the crime with liquor stores was to map their locations attributes with each other. For this, we aggregated crime complaints into planar zones centered around the store location and observe the trend in crime before and after the license was issued in that zone. We considered zones of 1 and 2 miles distance around the store for our analysis. We also identified that the two main categories of licenses issued were the On-Premises(Bars/Clubs etc.) and Off-Premises(Grocery Stores), so we limited our study to these classes of two license categories.

3. **Comparison Metrics:** To compare two distributions from the time before and after the liquor store was opened, we used KolmogorovSmirnov test to provedisprove our hypothesis. Since our study involves presenting an affect of distance and time variation for when and where the license was issued, we used two comparison metrics based on the KS-table. These were, $D_{30,\,0.05}$ = 0.1753 and $D_{60,\,0.05}$ = 0.247. Details of the statistics is given in the following sections.

   Next, we aggregated the crime reports for the distributions that were seen to have been affected by a license issue before and after the issue date and also varied the zones(of comparison distance) to actually support our claims of increase in crime around the area.

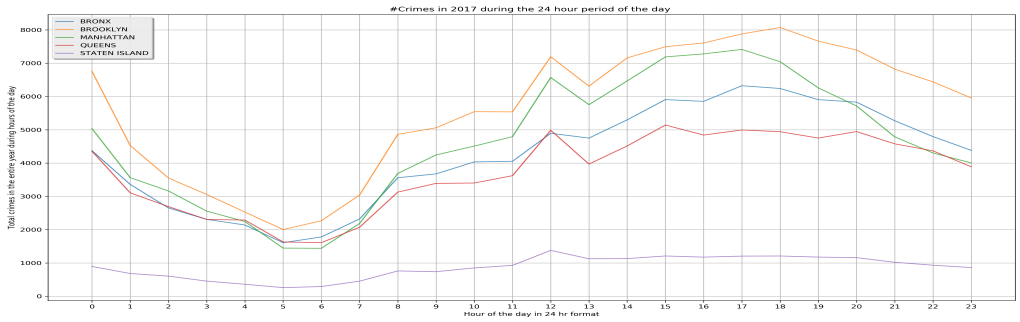

Figure 4: Topic 1: Overall Crime Statistics across all NYC boroughs throughout 2017.

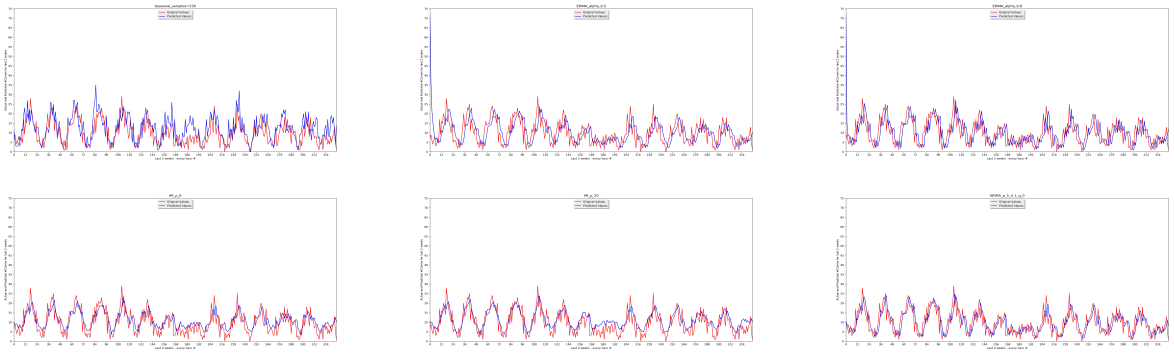

Figure 5: Topic 1: MANHATTAN: Crime frequency predictions using Time-series techniques

NYC Borough	Seasonal (season =336)	EWMA(alpha=0.5)	EWMA(alpha=0.8)	AR(p=20)	AR(p=6)	ARIMA(p=5, q=1, d=0)
Manhattan	5.767	6.454	6.168	4.14	4.317	4.397
Bronx	5.282	5.171	5.174	4.21	4.237	4.301
Brooklyn	6.086	13.155	11.741	4.949	4.887	4.915
Queens	4.773	5.311	5.282	4.003	4.031	4.085
StatenIsland	2.236	2.113	2.192	1.653	1.64	1.694

Figure 6: Topic 1: Tabulated Summary of Results with different Time-Series prediction techniques

GROCERY STORES LICENSES (OFF PREMISES) (3272 Issued)		Spatial Variation	
		1 Mile radius	2 Miles radius
Temporal Variation	30 days	42.47 %	43.46 %
	60 days	51.63 %	54.57 %

Table 1: Space-Time variation, Off-Premises

BARS/NIGHTCLUBS LICENSES (ON PREMISES) (5113 Issued)		Spatial Variation	
		1 Mile radius	2 Miles radius
Temporal Variation	30 days	47.74 %	46.65 %
	60 days	56.05 %	56.6 %

Table 2: Space-Time variation, On-Premises

4.2 Hypothesis Overview

4.2.1 Affect on Crime distribution

In this hypothesis, we check if the crime reports issued in an area are immune to liquor licenses issued. In other words, we test against the null hypothesis that the probabilistic distribution of the number of crimes reported before and after the store was opened, remains the same.

Procedure: To test if the distribution followed by crime complaints change after a license was issued in an area, we compare the number of crimes reported before and after the license date was issued within a specified distance from the store in an interval of fixed time period. We compare the two distributions using the KS-Test to conclusively comment on our hypothesis.

Results: We took an averaged observation by varying time period of 30, 60 days and distance radius from the store from 1, 2 mile radius. The average results obtained depict that of all the off-premises and on-premises store locations opened in 2017, 32.17% (out of 8385) locations witnessed a variation in distribution followed by the number of crimes reported in the area. Of the affected locations, 49.8% locations actually saw in increase in crime in the vicinity of the stores. We also observed that the two major categories of licenses issued were most affected and licenses of other types (Distilleries, Farms, other other types of alcohol licenses) showed very low effect on any change in crime distribution(only 4 - 7% of locations)

4.2.2 Temporal correlation from date of issue

Identifying any temporal correlation of time when the license was issued, with the number of crimes reported in the vicinity of the store.

Procedure: We were interested to check if the crime is related to the amount of time that has passed from when a liquor licensed is issued. We hypothesized that since nightclubs and bars are more dependent on time than grocery stores to become a popular attraction in an area, the on-premises stores should witness more crimes when more time has passed from when the license was issued. For this, we performed analyses to compare crime reports in periods of 30 and 60 days from before and after the license was issued and observe (if any) hike in the crime.

Results: From Figure 1 and Figure 2, we can see that both the categories are largely affected by the time lapsed from when the license was issued. There is about 9% increase in crime rate for both the categories and thus our hypothesis that bars would have witnessed more substantial affect is rejected. Thus, both categories witness an increase in crime reports with respect to temporal variation.

4.2.3 Spatial correlation from store

Identifying any spatial correlation of distance from the store for different types of licenses issued and try to correlate the distance with the variation of number of crimes reported in an area.

Procedure: We were also interested to check if the distance from the liquor store has any affect on the crimes in the vicinity. We hypothesized that since Off-premises stores are more prone to neighbourhood crimes away from the store, these should witness more crimes in the areas farther from stores than On-Premises stores which might be areas of 'easy' victims. For this, we analyzed distribution of crimes in planar zones separately of varying radius of 1 and 2 miles from the location where license was issued.

Results: It was found that the spatial variation indeed impact the crime rate from the liquor store. As seen from the results, Off-Premises stores witness a larger percentage increase in the crime as we vary the zonal distance from 1 mile to 2 mile, as compared to On-Premises stores. This is in accordance with our hypothesis.

5 Topic 3: Normal Distribution

In this hypothesis we analyzed whether number of hourly crimes happening in a year are Normally Distributed or not. Our main motivation behind this hypothesis was to find the peak hours to help NYPD police to be more vigilant at that particular time-period. It will help to reduce the crime rate if we are able to find out the timing when most of the

crime happens in a day and police can do patrolling in those areas more carefully.

Null Hypothesis: The Hourly crime rate is normally distributed

5.1 Methodology

1. **Data Cleaning:** First, we removed the incorrect data from the original data. For this hypothesis we used 2017 year's data-set. There were some rows in the data where the complaint data for column "CMPLNT_FR_TM" was entered incorrectly e.g. instead of 2017, year entered was 1017. So we removed such incorrect entries. We also removed the outliers from the data using the Tukeys rule. Here is the procedure to remove outliers:
   1. Find the Q1 (First quartile) and Q3 (Third quartile)
   2. Find the IQR (Interquartile Range). IQR = Q3 - Q1
   3. Lowerbound = Q1 - IQR
   4. Upperbound = Q3 + IQR

   All the data points which are within range (Upperbound - Lowerbound) are valid data and rest all are outliers.

2. **Data Pre-processing:** Since we want to check the hourly crime rate and we have the data for the whole 2017 year. So, we aggregated the data by the hour for the year 2017 in 24 hour format. We created 24 bins for each hour e.g. one bin for all the crimes happening between 1 PM to 2 PM and another bin for crimes between 2 PM to 3 PM and so on.
   We iterate all the crimes and based on time they happened we added them in the corresponding bin. Convert the given time in minutes and the corresponding bin number will be minutes/60

3. **Mean and Variance:** After data cleaning and pre-processing, we got the final data on which we can apply our techniques to check our hypothesis. First, we computed the mean and variance estimators for given data. For that we used three different techniques:

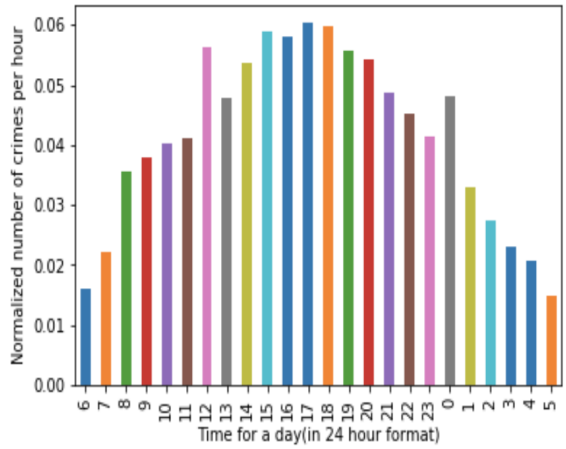

Figure 7: Topic 3: Hourly data distribution

   (a) **MLE**: We used the following formula to get the MLE mean and MLE variance estimators:

$$\hat{\mu}_{mle} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$$

$$\hat{\sigma}_{mle} = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

   We also calculated the Confidence Interval for estimated MLE parameters.

   (b) **MME**: The following formulas are used to find the MME mean and MME variance parameters:

$$\hat{\mu}_{mme} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$$

$$\hat{\sigma}_{mme} = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i)^2 - (\frac{1}{N} \cdot \sum_{i=1}^{N} x_i)^2$$

   (c) **Guess of mean and variance by plotting the graph**: From figure showing hourly data distribution 7, it looks like the peak value is 0.06 so we assumed our mean as 0.5 and variance as 0.0145.

4. **Standard Normal Form:** Converted the given data into standard normal distribution by using following formula:

X - Normal Distribution with mean μ and standard deviation as σ.

$$Z(StandardNormal) = (X - \mu)/\sigma$$

5. **KS Test:** Performed KS Test to check whether number of crime are normally distributed or not. Since our data is already standard normal distributed so we used the "stats.kstest(data, 'norm')" formula to perform KS Test. It gives you the max-distance and the p-value.

5.2 Results

After applying KS test for all the three techniques, we found that that Null hypothesis is accepted i.e. the data is normally distributed.
We used the D value for KS test as 0.26404 for n = 24 and alpha = 0.05

The results for all the three techniques is shown in 12

There is a 95% chance that our CI contains True for estimated MLE. Confidence Interval: (0.0359, 0.0474)

6 Topic 4: Relation between crime rate and poverty

The Supplemental Nutrition Assistance Program (SNAP), commonly known as food stamps, provides assistance to lower income families in USA. Borough of NYC figure in the cities with most food stamp recipients. 21 % of the NYC population i.e around 1.84 million people are currently on food stamps. Generally, people tend to relate crime with poverty. The more poor people live in some community, to more crime rate is expected in that particular locality. In this hypothesis we use NYC crime data and SNAP distribution schedule to analyze whether crime is actually related to poverty or not. Our motivation behind this hypothesis is to debunk the view that poor people commit more crimes.

6.1 Methodology

1. **Data Cleaning:** For testing our hypothesis we are using crime data of first six months of 2016. During these six months, food stamps were distributed in the first two weeks of the month. First, we removed the incorrect dates and time from Complaint from date and Complaint from time columns. There were some rows with incorrect years(1017, etc.).

   Then we aggregate our data in two formats:

   - *6 month aggregated data:* We aggregate number of crimes happening in first 6 months of 2016 in 30 days format.

   - *Monthly aggregated data:* We aggregate number of crimes happening in first 6 months of 2016 in 30 days format for each month separately.

   We then remove outliers from these aggregated data-set using Tukeys rule. This is similar to the Tukey's rule applied in the Normal Distribution hypothesis.

2. **Permutation test:** To compare whether number of crimes are more in the first 15 days of the month or last half of the month, we are using Permutation test. As we are more concerned about the average number of crimes happening in these two time-sets, first 15 days and rest of the month, so we are using permutation test, as permutation test focuses more on the mean of the distribution.

   We are running the test on the data before removing the outliers, after removing the outliers, on 6 month aggregated data and on monthly aggregated data for the all 6 months separately.

6.2 Hypothesis Overview

1. **Permutation test on aggregated 6 months crime rate data to test the Null hypothesis that crime in first half of the month is equal to crime rate in second half of the month:** In this hypothesis, we check whether crime rate of first 15 days is similar to crime rate in last 15 days for all the crime rate aggregated for first 6 months of 2016. We ran permutation test for 10K and 100K iterations. Number of iterations were varied to check whether number has any effect on the results. In both the cases, number of crimes in second half of the month are not more than number of crimes in first half of the month. To our surprise, crime rate is more in first 15 days, the time of distribution of food stamps, hence, debunking the view that poor people commit more crime. It is obvious

Technique to compute Mean and Variance	Mean	Sigma	Max Distance	p-value	Null Hypothesis
MLE	0.0416	0.0144	0.1281	0.8255	Accepted
MME	0.0416	0.0144	0.1281	0.8255	Accepted
Guess by plotting graph	0.060	0.0145	0.5	4.50e-06	Rejected

Figure 8: Result of Topic 3: Normal Distribution

that, if food stamps are distributed in first two weeks, then there will be less crime, if poor people are the one who commit the crime. But our analysis rejected this general belief.

Results: Null Hypothesis is rejected for 10K iterations and for 100K iterations with a p-value of 0.0316 and 0.03005. One an average, we take a p-value greater than or equal to 0.5 to accept the Null hypothesis. But, the main observation is that, crime rate is more in first 15 days, rather then last 15 days. So, basically, debunking the view that poor people commit more crimes.

2. **Permutation test on monthly aggregated data of first 6 months of 2016:** For this task, we are using monthly aggregated data, instead of running permutation test on aggregated data of first 6 months. We are doing this, because it might be the case that, Null hypothesis might be rejected or accepted in some of the months and only some particular month data is dominating the complete results. So it's better to run permutation test on each month separately also.

   Results: Null Hypothesis is rejected for Jan, Feb, and March, but again crime rate is more in first half of the month. For March, May, and June, Null hypothesis is accepted.

3. **Permutation test on aggregated 6 months crime rate data with outliers removed:** In this hypothesis, we check whether crime rate of first 15 days is similar to crime rate in last 15 days for all the crime rate aggregated for first 6 months of 2016. But, in this Hypothesis we remove all the outliers using Tukey's rule. We did this because we observed in the crime data graph that the number of crime rates on the 1st day of every month is exponentially more than crime rates on any other day. So, we used IQR rule to remove the outliers and then ran permutation test on this aggregated data. We ran permutation test for 10K and 100K iterations. In both the cases, Null hypothesis was accepted i.e number of crimes in first half of the month is equal to number of crimes in last half of the month. Hence, debunking the view that poor people commit more crime.
   Results: Null Hypothesis is accepted for 10K iterations with high p-value of 0.999.

4. **Permutation test on monthly aggregated data for first 6 months of 2016 with outliers removed:** In this hypothesis, we check whether crime rate of first 15 days is similar to crime rate in last 15 days for all of the first 6 months of the 2016 separately. We used IQR rule to remove the outliers and then ran permutation test on this aggregated data. We ran permutation test for 10K and 100K iterations. In both the cases, Null hypothesis was accepted i.e number of crimes in first half of the month are equal to number of crimes in last half of the month. Hence, debunking the view that poor people commit more crime.
   Results: Null Hypothesis is accepted for 10K and 100K iterations with high p-value in between 0.45 and 0.99 for all the six months.

7 Prior Work :

There have been a number of attempts in the past on analyzing and establishing conclusive studies for NYPD Crime data.

- One such study involves identification of hotspots for the areas that witness larger number of crimes than the other. The tool has been

	Aggr(Jan 2016 - Jun 2016)	Jan 2016	Feb 2016	March 2016	April 2016	May 2016	June 2016
p-value	0.0316	0.0002	0.0344	0.5031	0.0304	0.1403	0.698
Result	R	R	R	A	R	A	A
More crime in	first 15 days	first 15 days	first 15 days	N/A	first 15 days	N/A	N/A

Figure 9: Result of Topic 4: Permutation test results for 10K iterations (R - Rejected, A - Accepted)

	Aggr(Jan 2016 - Jun 2016)	Jan 2016	Feb 2016	March 2016	April 2016	May 2016	June 2016
p-value	0.03005	0.00036	0.03367	0.50865	0.02859	0.14294	0.69639
Result	R	R	R	A	R	A	A
More crime in	first 15 days	first 15 days	first 15 days	N/A	first 15 days	N/A	N/A

Figure 10: Result of Topic 4: Permutation test results for 100K iterations (R - Rejected, A - Accepted)

	Aggr(Jan 2016 - Jun 2016)	Jan 2016	Feb 2016	March 2016	April 2016	May 2016	June 2016
p-value	0.9999	0.6694	0.5498	0.9469	0.4464	0.8863	0.9983
Result	A	A	A	A	A	A	A

Figure 11: Result of Topic 4: Permutation test results for 10K iterations with outliers removed(R - Rejected, A - Accepted)

	Aggr(Jan 2016 - Jun 2016)	Jan 2016	Feb 2016	March 2016	April 2016	May 2016	June 2016
p-value	0.99874	0.67467	0.54318	0.94626	0.44954	0.88701	0.99828
Result	A	A	A	A	A	A	A

Figure 12: Result of Topic 4: Permutation test results for 100K iterations with outliers removed (R - Rejected, A - Accepted)

developed by the data.cityofnewyork themselves and can be found here

- Another one of NYPD's efforts was to provide a borough wise crime statistics tool for the type of crime witnessed in the area. The NYPD also provides weekly updated statistics that are categorized by police borough and precinct with the number of crimes witness. Found here.

- Center for Civic Innovation, Manhattan Institute also makes use of the dataset to evaluate explanations that witness for the sharp decline in crime in New York City during the 1990s. The authors attribute the decline to various socio-economic factors, such as an economy, declining use of hard drugs and other factors. The study could be found here

- According to an article published in Politico, NYPD has also developed a predictive software that models crime forecasts based on temporal patterns, weather, environmental risk factors (locations of bars, bus stops, etc.) and other factors. While we couldn't find conclusive evidence that the dataset was completely the one we went for, but since the studies are performed by NYPD themselves, it is highly likely that this data has been used in their predictive models. Article Link

8 Future Work :

8.1 Topic 1: Time Series Prediction

The combination of all 3 tasks below in the future would have led to more interesting results.

- Seasonal ARIMA: Non-seasonal ARIMA was tried out. Seasonal ARIMA might have led to more interesting results.

- Hyper-parameter tuning: Each of the techniques had hyper-parameters from 1 to a few. Many combinations are yet to be tried which would have solidified the conclusion even more.

- Big-Data Analytics: In general more data samples (big-data) lead to better predictions. Now that is not always true for Time-Series predictions. This would have been very interesting to combine data from previous years (2016,2015 and so on) or isolate them year-wise just like 2017 and figure out the predictive results in each case.

8.2 Topic 2: Liquor Store impact

Given more time, we planned to provide a more comprehensive analysis of the liquor stores and identify a set of 'risky bars' around a location. Moreover, rather than just comparing number of crimes to classify 'risk' of a location, we could use logistic regression based methods that could given probability of occurrence of crime at a location given other factors such as time of day, date etc. The results could then be used to develop a mobile application that the user could use to identify any bars around him or even safer paths to go from a destination to another(Example: A more safer path/bars in user's path from 42nd Street to 54th in NYC)

8.3 Topic 4: Food Stamps impact

Currently, we are only focusing on distribution schedule of SNAP and not on specific community wise distribution schedule. It might be the case that, communities with more SNAP recipients actually see a spike in crime rate during the non-distribution period. Currently, crime rates are for all the NYC borough, because SNAP distribution data for each specific borough or small community is not publicly available.

9 Source Code Repository :

Project Source code containing .ipynb files and graphical outputs for all topics. https://github.com/shayanraysbu/CSE544-ProbStats-FinalProject

References

[1] NYC Open Data website
[2] NYC Liquor Dataset
[3] Git Repo