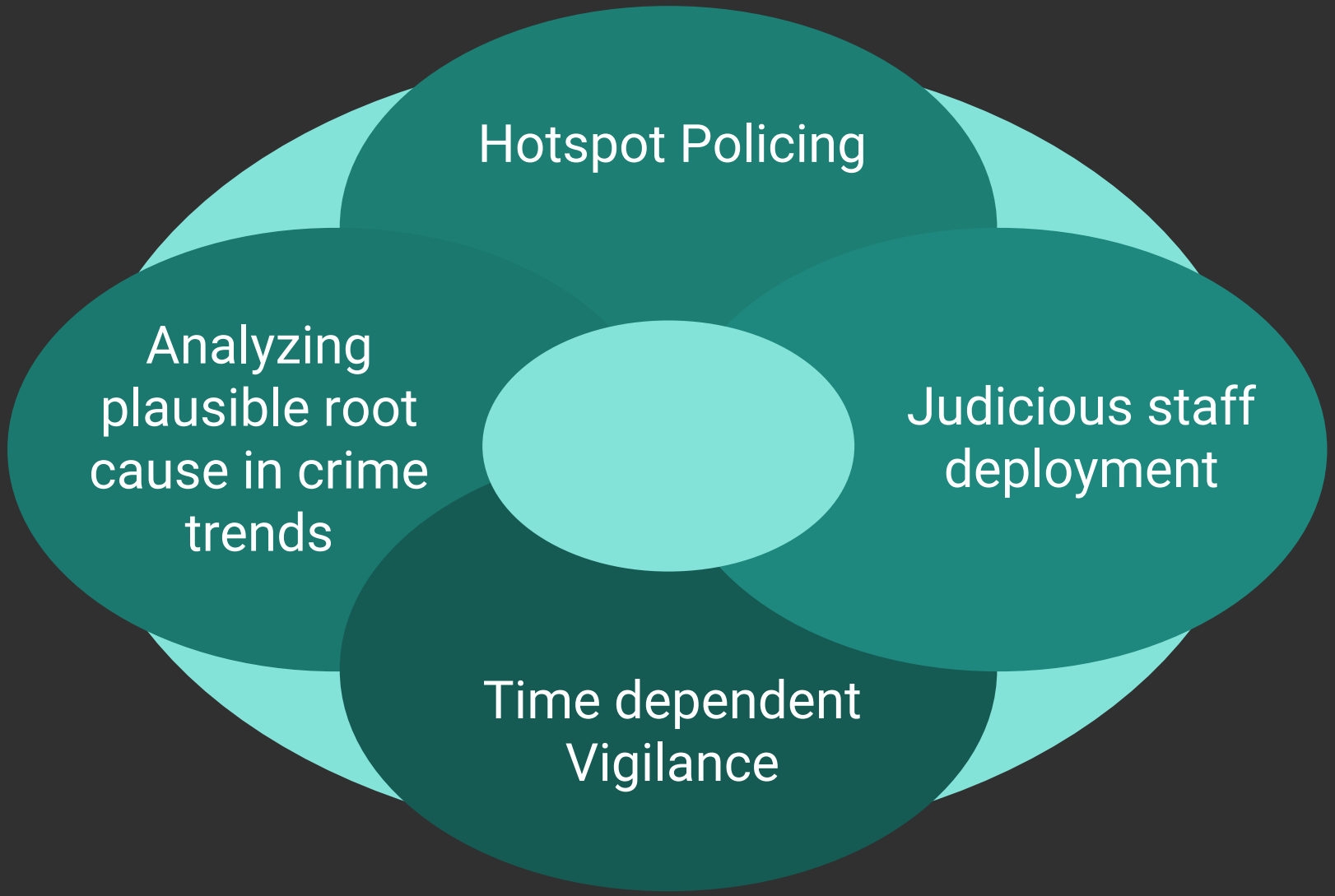
CSE 544 - Project Presentation

NYPD Crime Data Analysis

Anurag Arora (111425080)
Bhavesh Goyal (111482386)
Renu Rani (111482474)
Shayan Ray (111424665)

- *Under guidance of Professor Anshul Gandhi*

Motivation


Hotspot Policing
Analyzing plausible root cause in crime trends
Judicious staff deployment
Time dependent Vigilance

Dataset #1

NYPD Complaint Data

Scope Used:	**2016, 2017**
Net Size:	**131MB**
Total Rows:	**468,000 (approx.)**
Available Columns:	**23**

Data Snapshot

CMPLNT_	CMPLNT_	CMPLNT_	CMPLNT_	CMPLNT_	RPT_DT	KY_CD	OFNS_DE	PD_CD	PD_DESC	CRM_ATP	LAW_CAT	JURIS_DE
149878695	12/30/17	14:10:00	12/30/17	14:15:00	12/30/17	235	DANGEROUS	567	MARIJUANA	COMPLETED	MISDEMEAN	N.Y. HOUSIN
612851223	12/30/17	14:10:00	12/30/17	14:15:00	12/30/17	235	DANGEROUS	567	MARIJUANA	COMPLETED	MISDEMEAN	N.Y. HOUSIN
219841792	12/30/17	9:30:00	12/30/17	10:29:00	12/30/17	359	OFFENSES A	750	RESISTING A	COMPLETED	MISDEMEAN	N.Y. HOUSIN
407001126	12/30/17	2:20:00	12/30/17	3:12:00	12/30/17	106	FELONY ASS	109	ASSAULT 2,1	COMPLETED	FELONY	N.Y. HOUSIN
866078986	12/30/17	1:05:00	12/30/17	1:16:00	12/30/17	125	NYS LAWS-U	847	NY STATE LA	COMPLETED	FELONY	N.Y. POLICE I
701999408	12/30/17	0:20:00	12/30/17	4:08:00	12/30/17	361	OFF. AGNST	639	AGGRAVATE	COMPLETED	MISDEMEAN	N.Y. HOUSIN
649337772	12/30/17	0:10:00	12/30/17	0:20:00	12/30/17	235	DANGEROUS	567	MARIJUANA	COMPLETED	MISDEMEAN	N.Y. POLICE I
836680472	12/29/17	23:40:00	12/29/17	23:45:00	12/30/17	344	ASSAULT 3 &	101	ASSAULT 3	COMPLETED	MISDEMEAN	N.Y. HOUSIN

BORO_NI	ADDR_PC	LOC_OF_(	PREM_TY	PARKS_N	HADEVEL	X_COORD	Y_COORD	Latitude	Longitude	Lat_Lon
MANHATTAN	25	OPPOSITE OI	RESIDENCE - PUBLIC HOU		UPACA (SITE	1,000,868	231,062	40.8008804	-73.939979	(40.800880389, -73.93997859)
BROOKLYN	73	INSIDE	RESIDENCE - PUBLIC HOU		TILDEN	1,009,656	181,287	40.6642394	-73.908425	(40.664239422, -73.908425011)
BRONX	44	INSIDE	RESIDENCE - PUBLIC HOU		CLAREMONT	1,009,143	242,913	40.8333886	-73.910046	(40.833388611, -73.910046026)
BRONX	43	INSIDE	RESIDENCE - PUBLIC HOU		MONROE	1,021,543	239,271	40.8233487	-73.865257	(40.823348717, -73.865256652)
BROOKLYN	81	OPPOSITE OI	STREET		BREVOORT	1,004,856	187,468	40.6812173	-73.925708	(40.681217349, -73.925707802)
BRONX	42	INSIDE	RESIDENCE - PUBLIC HOU		BUTLER	1,009,944	244,103	40.8366525	-73.907147	(40.836652522, -73.907146909)
QUEENS	101	INSIDE	RESIDENCE - APT. HOUSE		BAYSIDE-OCI	1,043,516	157,081	40.5976374	-73.786589	(40.597637353, -73.786588876)
BROOKLYN	90	INSIDE	RESIDENCE - PUBLIC HOU		INDEPENDEN	994,231	196,250	40.7053403	-73.964003	(40.705340265, -73.964002557)

Attribute Descriptions

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)

Dataset #2

Liquor Authority, Active licenses
(*Issued across entire NY State*)

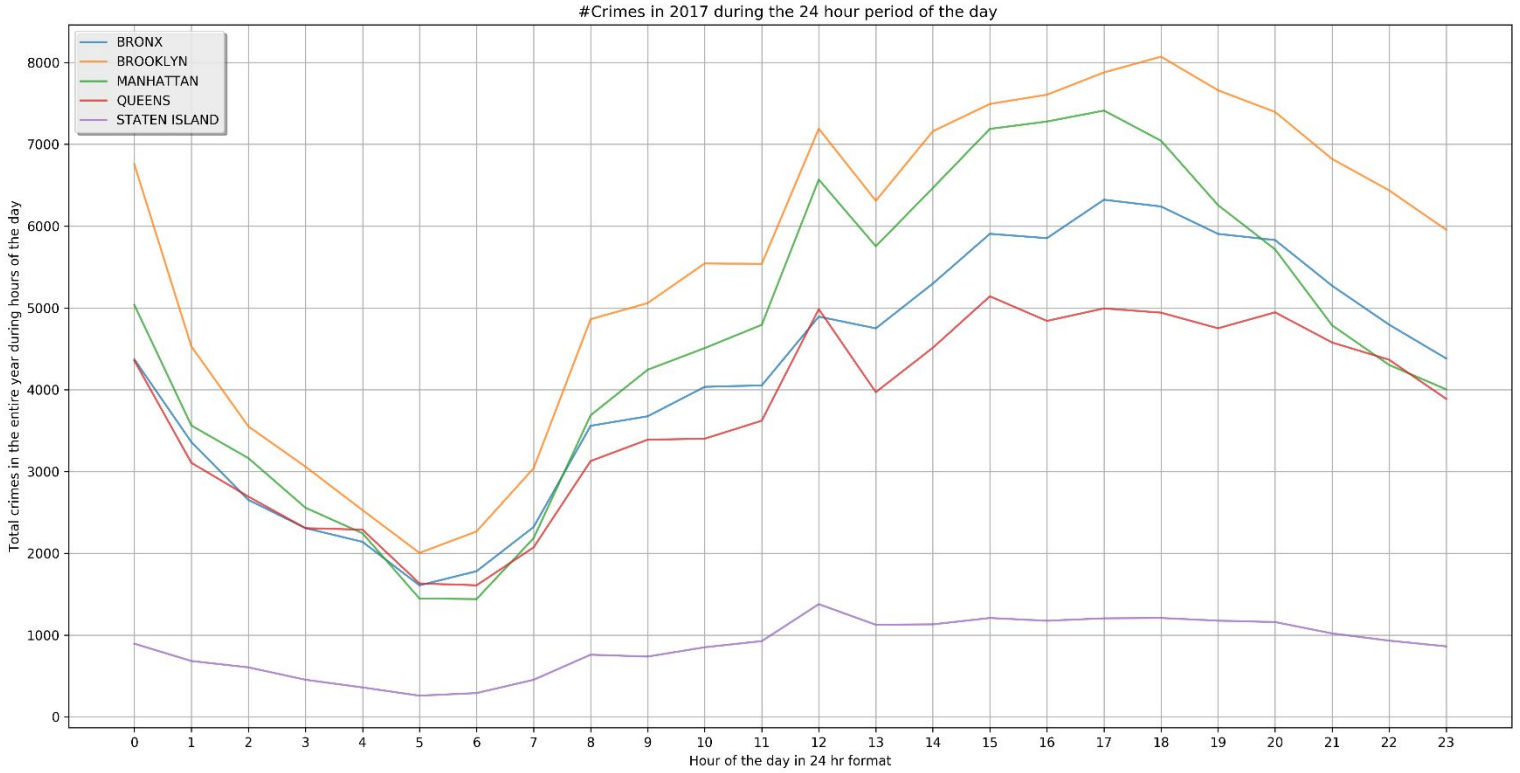
Scope Used:	**2017**
Net Size:	**13MB**
Total Rows:	**51.6k Licenses**
Available Columns:	**21**

Data Snapshot

	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	Doing Business As (DBA)
1	1000080	GROCERY STORE BEER	122	A	New York City	1	BRONX	ERIDANIA CORP	1888 WALTON GROCERY
2	1000090	GROCERY STORE BEER	122	A	New York City	1	BRONX	1098 MINI MARKET INC	JACOBO GROCERY & DELI
3	1000099	GROCERY STORE BEER	122	A	New York City	1	BRONX	J & A FRUIT & GROCERY CORP	J & A FRUIT & GROCERY CTR
4	1000167	GROCERY STORE BEER	122	A	New York City	1	BRONX	ALTAGRACIA GROCERY CORP	
5	1000207	GROCERY STORE BEER	122	A	New York City	1	BRONX	48 BURNSIDE FOOD CORP	FOOD DYNASTY
6	1000218	GROCERY STORE BEER	122	A	New York City	1	BRONX	CRISMELI DELI GROCERY INC	CRISMELI DELI GROCERY
7	1000305	GROCERY STORE BEER	122	A	New York City	1	BRONX	RAMON A RODRIGUEZ	RODRIGUEZ GROCERY
8	1000351	GROCERY STORE BEER	122	A	New York City	1	BRONX	JOSE A DUME	FOREST DELI GROCERY
9	1000407	GROCERY STORE BEER	122	A	New York City	1	BRONX	ANGELO RUIZ	ANGELOS GROCERY STORE
10	1000471	GROCERY BEER, WINE PROD	122	AX	New York City	1	BRONX	WILLIS DELI INC	
11	1000482	GROCERY STORE BEER	122	A	New York City	1	BRONX	ALEJANDRO SANCHEZ	GARDEN MARKET GROC & DELI
12	1000542	GROCERY STORE BEER	122	A	New York City	1	BRONX	NEW ERA FOODS ONE INC	C TOWN SUPERMARKET

	Actual Address of Premises (Address1)	City	State	Zip	License Certificate Number	License Original Issue Date	License Effective Date	License Expiration Date	Latitude	Longitude	Location
1	1888 WALTON AVE	BRONX	NY	10452	876158		03/01/2016	02/28/2019	40.84942969	-73.9092608276	(40.849429695°, -73.9092608276°)
2	1098 ANDERSON AVENUE	BRONX	NY	10452	888708	08/31/2004	03/01/2017	02/29/2020	40.834710220	-73.9252384401	(40.8347102202°, -73.9252384401°)
3	11 17 E 171ST STREET	BRONX	NY	10452	893397	09/20/2007	11/01/2017	10/31/2020	40.8412774818	-73.9165823305	(40.8412774818°, -73.9165823305°)
4	10 12 E 176TH STREET	BRONX	NY	10453	892433		07/01/2017	06/30/2020			
5	40 48 WEST BURNSIDE AVE	BRONX	NY	10453	872236	10/15/1997	10/01/2015	09/30/2018	40.853900410	-73.9094397981	(40.8539004109°, -73.9094397981°)
6	297 BROOK AVE	BRONX	NY	10454	892968	12/09/2008	08/01/2017	07/31/2020	40.80896069	-73.9184750776	(40.8089606987°, -73.9184750776°)
7	753 FOX STREET	BRONX	NY	10455	879792	08/10/1995	08/01/2016	07/31/2019	40.815551520	-73.8984235507	(40.8155515206°, -73.8984235507°)
8	751 FOREST AVENUE	BRONX	NY	10456	890290		05/01/2017	04/30/2020	40.818262364	-73.9062670245	(40.8182623646°, -73.9062670245°)
9	279 E 173RD STREET	BRONX	NY	10457	893329		08/23/2017	06/30/2020	40.843720215	-73.9071368378	(40.8437202159°, -73.9071368378°)
10	538 E 178TH STREET	BRONX	NY	10457	892282	07/08/2003	07/01/2017	06/30/2020	40.847674536	-73.895236226	(40.8476745365°, -73.895236226°)
11	393 E 200TH STREET	BRONX	NY	10458	885776	11/13/1995	11/01/2016	10/31/2019	40.867498136	-73.8837294909	(40.8674981367°, -73.8837294909°)
12	668 CRESCENT AVE	BRONX	NY	10458	891389		07/01/2017	06/30/2020	40.854064369	-73.8861074	(40.8540643692°, -73.8861074°)

2017 Hourly Crime Trend - NYC



Summary of Topics covered

- MLE, MME (To find mean and variance)
- Confidence Intervals
- Tukey's Rule with IQR (To remove outliers)
- Permutation Test
- Kolmogorov–Smirnov Test
  - Sample Distribution with Normal
  - Two Sample Distributions
- Seasonal, EWMA, AR (with Linear regression), ARIMA,
- RMSE

Topic #1

Predict the number of crimes in NYC boroughs in the next 2 weeks by the hour

Motivation : NYPD can deploy staff judiciously across boroughs for the next 2 weeks.

Note: Experiments were conducted across 5 NYC Boroughs. Only Manhattan data is demonstrated here. All others will be included in the report.

Steps taken: (All boroughs)

1. **Data Cleaning:**
   - Outlier detection and removal (with CMPLNT_FR_DT)
   - Data filtered by BORO_NM

2. **Data Preprocessing:**
   - Data grouped by the hour using CMPLNT_FR_DT and CMPLNT_FR_TM (e.g., 2017-01-01 00:01:00 - <count>)

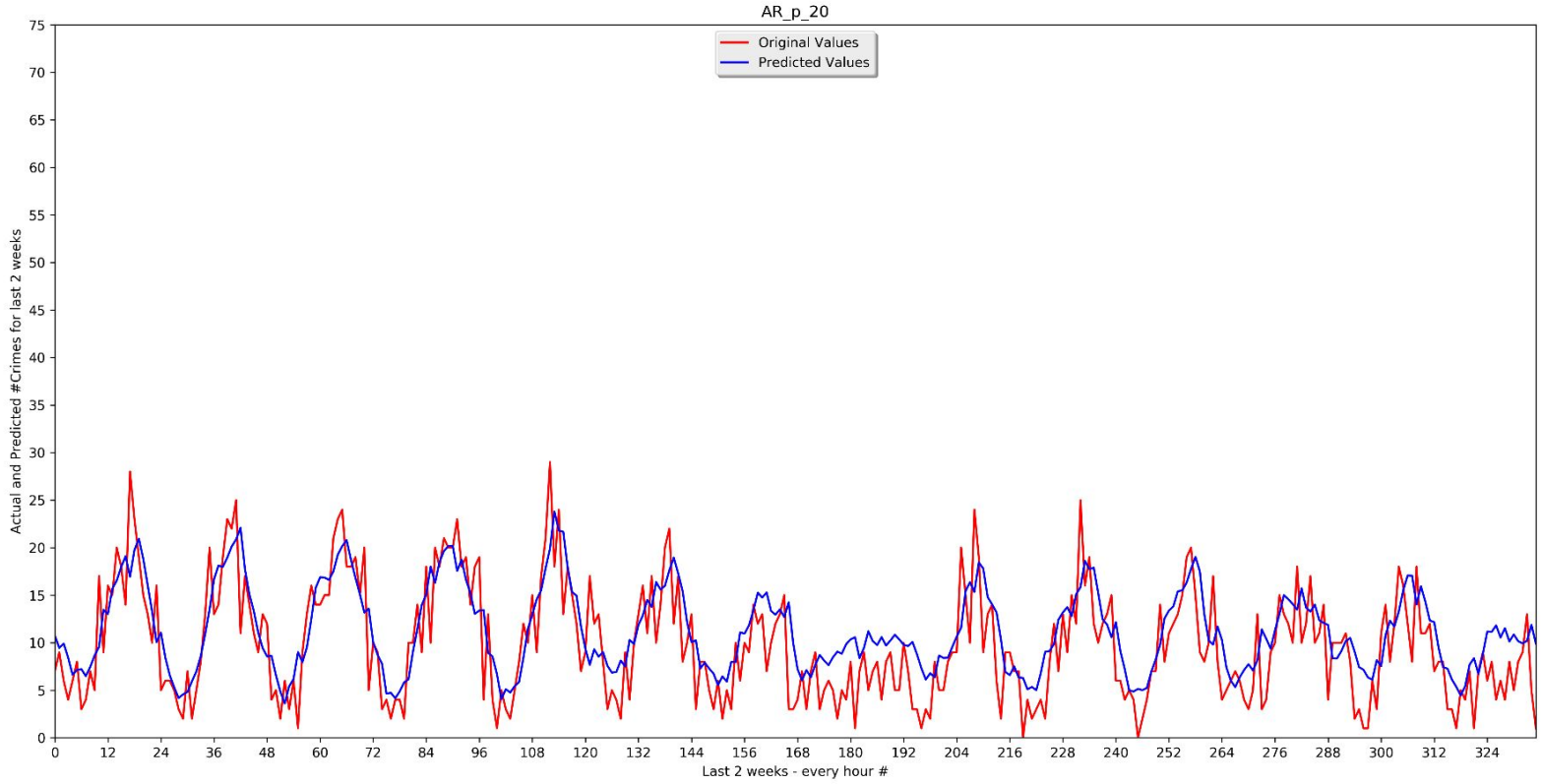
3. **Train-Test Split:**
   - 8424 train samples, 336 test samples(last 2 weeks)

4. **Model accuracy metric:**
   - RMSE(Root Mean Square Error)

Techniques Used

- Seasonal Last Observed
- EWMA (Expected Weighted Moving Average)
- AR (Auto-Regression) - Rolling Forecast(Linear Regression used to compute Betas)
- ARIMA (Autoregressive Integrated Moving Average) - Rolling Forecast

Auto-Regression (p=20)



Prediction Accuracy: RMSE

SEASONAL(season=336)	5.76654
EWMA(alpha=0.5)	6.45367
EWMA(alpha=0.8)	6.16768
AR(p=20)	4.14003
AR(p=6)	4.31661
ARIMA(p=5,q=1,d=0)	4.39698

Topic #2

Does opening a liquor store in an area impact the Spatial Crime Rate?

Null Hypothesis: Liquor stores have no effect on the distribution of the Crime Rate for an area

Motivation: Regulating Police patrolling, and identifying a safe(r) place to buy liquor ;)

Steps Taken

1. **Data Cleaning (2017)**
   - Complaint Data: Limiting analysis to samples with valid location coordinates available.
   - Liquor License Data: Removing Inconsistent License Issue Dates and limiting analysis to NYC licenses.
2. **Data Preprocessing**
   - Aggregating crime reports into (planar)zones of locations for each of liquor-license coordinates
   - Sampling out two major categories of OFF / ON PREMISES license issues for analysis.
3. **Comparison Metric**
   - K-S Testing $\{D_{30, 0.05} = 0.1753, D_{60, 0.05} = 0.2471\}$
4. Compare the aggregated count of crimes to establish conclusions for the ‘hike’ of crime in the area.

Tasks Performed

1. Identified two major Categories of Licenses Issued:
   - Grocery Stores (OFF PREMISES LIQUOR)
   - Bars/Nightclubs (ON PREMISES LIQUOR)
2. Testing effect of a new liquor license on crime in an area w.r.t 1-2 mile radii over a period of 30-60 days
3. Analyzing Spatial variations of distance from liquor stores and the reported crimes
4. Analyzing temporal effect of licenses issued and Crimes reported
5. Performed KS-Test to compare distributions
   - Final Statistics Reported by Aggregating number of crimes analyzed separately for each liquor store

Results

Statistics Overview (Averaged):

- **32.17%** (out of 8385) On + Off Premises **liquor licenses** issued **affected** distribution of Crime reports, as of 2017.
- Out of the affected locations, **49.8%** locations witnessed an **increase** in crime activities that were reported.
- Zones for **On-Premises** saw **more crimes** than Groceries (Off-Premises). Licenses of other types(Distilleries / Farms etc.) didn't pose much effect on distribution(4 - 7 %).
- While both categories saw an increase in crime w.r.t time, **off premises** locations witnessed a **more** significant **increase** in crime **w.r.t distances** from the store.

Results (Effect on Distribution)

The Tables summarize the percentage of locations affected when the license was issued:

GROCERY STORE LICENSES (OFF PREMISES) (3272 Issued)		Spatial Variation	
		1 Mile radius	2 Miles radius
Temporal Variation	30 days	22.9 %	28.7 %
	60 days	38.29 %	46.4 %

ON PREMISES LIQUOR (5113 Licenses Issued)		Spatial Variation	
		1 Mile radius	2 Miles radius
Temporal Variation	30 days	20.8 %	25.99 %
	60 days	34.5 %	39.8 %

Results (Increase In Crime Reports)

The Tables summarize the percentage increase in crime reports w.r.t spatial and temporal variation:

GROCERY STORE 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 %

ON PREMISES LIQUOR (5113 Licenses Issued)		Spatial Variation	
		1 Mile radius	2 Miles radius
Temporal Variation	30 days	47.74%	46.65%
	60 days	56.05 %	56.6 %

Topic #3

Are number of hourly crimes happening in a year, Normally Distributed?

Null Hypothesis: The Hourly crime rate is normally distributed

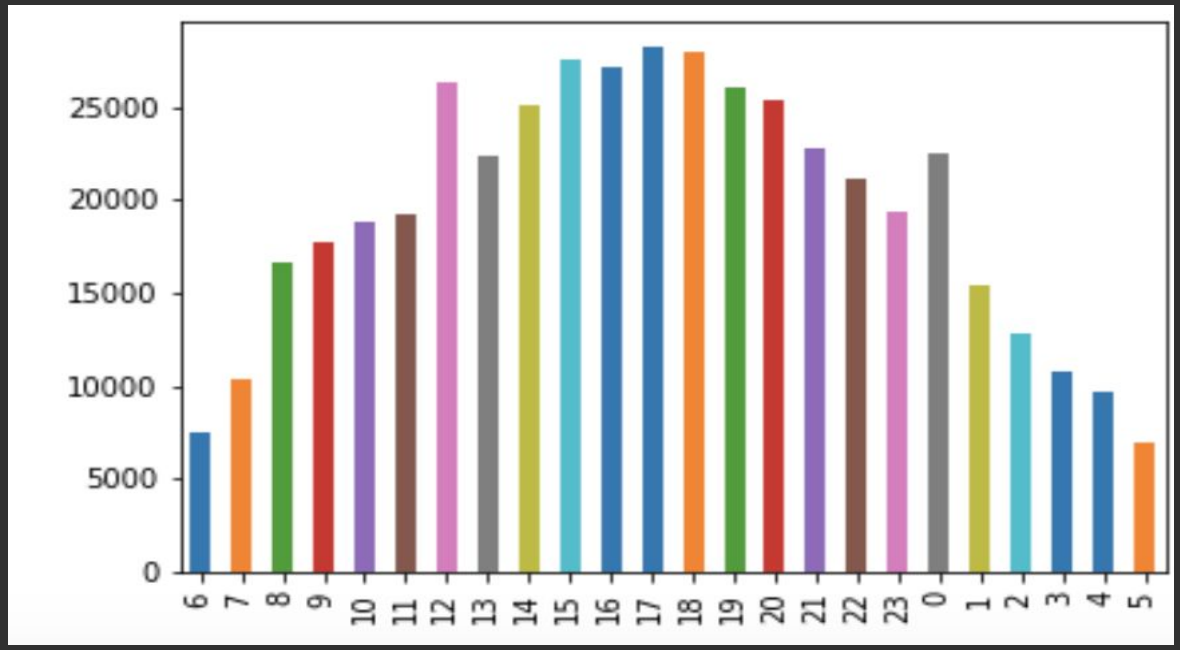
Motivation : NYPD can be more vigilant at peak hours

Steps taken

1. **Data Cleaning (2017)**
   - Removal of incorrect dates from CMPLNT_FR_TM and CMPLNT_FR_DT column *(For instance: Year 1017)*
   - Outlier detection and removal (using Tukey's rule)
2. **Data Preprocessing**
   - Data aggregation by the hour for the year 2017 in 24 hour format.
3. Computing mean and variance using MLE.
4. Providing Confidence Intervals for estimated MLE parameters.
5. Converting the aggregated crime data to standard normal.
6. Performing K-S Test to check whether number of crimes are normally distributed in a day.

Techniques Used

- MLE
- MME
- Confidence Intervals
- Plotting crime data and then guessing mean and variance
- KS-Test
- IQR



MLE Estimators

Used empirical MLE estimators to get mean and variance.

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

https://blackboard.stonybrook.edu/bbcswebdav/pid-4369699-dt-content-rid-29771703_1/courses/1184-CSE-512-SEC01-51537/02_MLE_MAP.pdf

Results

Null Hypothesis accepted

$D_{n,\alpha}$ value from KS-table for n = 24, alpha = 0.05 = 0.26404

Technique to compute Mean/Variance	Max - Distance	p-value	Null Hypothesis	Mean	Sigma
MLE	0.128144723996	0.825507472198	Accepted	0.0416	0.0144
MME	0.128144723996	0.825507472198	Accepted	0.0416	0.0144
Guess by plotting graph	0.5	4.50357771409e-06	Rejected	0.060	0.0145

Note: ∃ 95% chance that our CI contains μ_{True} for estimated μ_{MLE}

For our Confidence Interval of (0.0359, 0.0474)

Topic #4

Is there a relation between crime rate and poverty?

Null Hypothesis: Crime is not related to poverty.

Motivation: To debunk the view that poor people commit more crimes.

Trivia: 21% of NYC population is on SNAP

Note: Analyzing the effect of distribution schedule of Food Stamp on crime rate

Steps taken

1. Data Cleaning (2016)
   a. Removal of incorrect dates from CMPLNT_FR_TM and CMPLNT_FR_DT column
   b. Outlier detection and removal (using Tukey's rule)

2. Data Preprocessing:
   a. Data aggregation by the days of the month from Jan 2016 - Jun 2016.

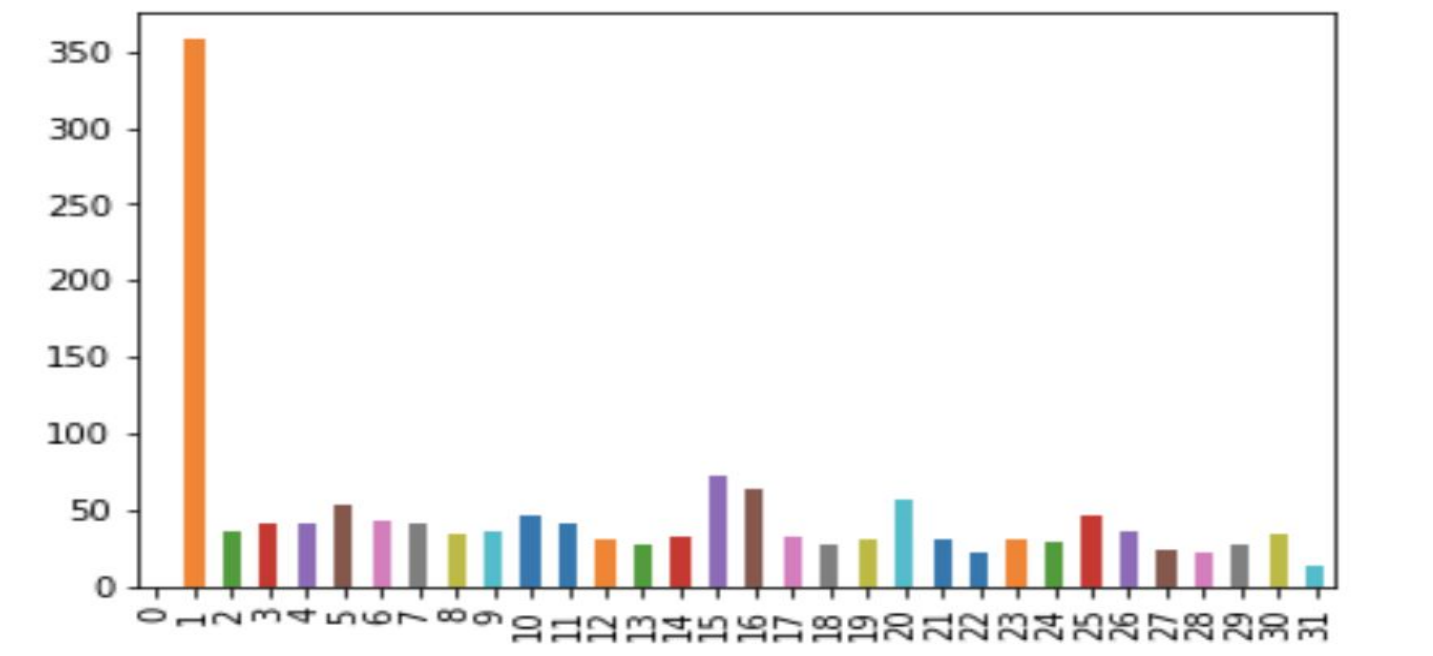
3. Performing Permutation Test to check whether the distribution of crime rates of first half of the month is equal to the distribution of crime rates of the second half of the month.

Tasks Performed

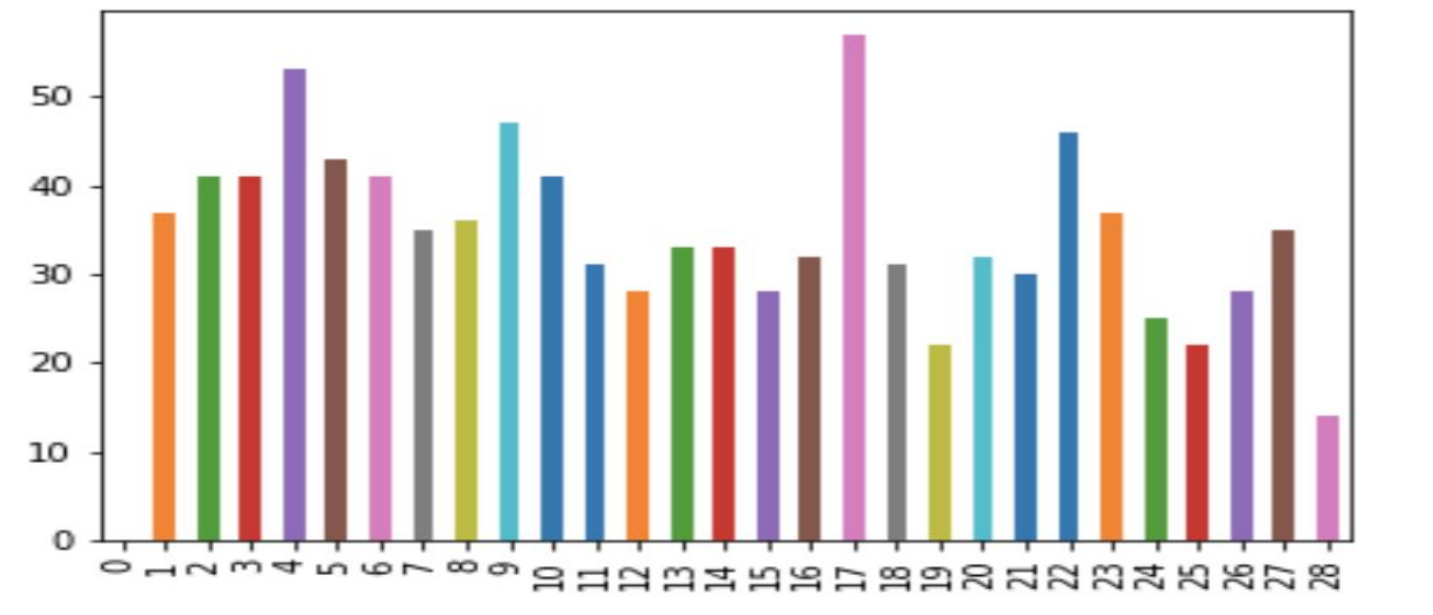
1. Performed Permutation-Test
   a. By aggregating 6 months data
   b. Separately analyzed data for each month

2. With/without removing outliers

3. Ran Permutation test for 10K and 100K iterations



With Outliers



Without Outliers

Results

Null Hypothesis accepted

For 10K iterations

Without removing outliers:

	Aggregated data (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	0.6988
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

With removing outliers:

	Aggregated data (Jan 2016 – Jun 2016)	Jan 2016	Feb 2016	March 2016	April 2016	May 2016	June 2016
p-value	0.999	0.6694	0.5498	0.9469	0.4464	0.8863	0.9983
Result	A	A	A	A	A	A	A

Results

For 100K iterations

Without removing outliers:

	Aggregated data (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

With removing outliers:

	Aggregated data (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

Conclusions

Hypothesis #1

AR does a better job at predicting number of crimes by the hour

Hypothesis #2

Liquor stores indeed affect crime rates in their vicinity.

Hypothesis #3

Number of crimes per day are normally distributed

Hypothesis #4

Schedule of Food stamp distribution has no relation with crime rates

Thank You!

Questions?

Fun Fact: *Statistics recommends, a safer place to grab the next cold one might be your nearest, newly opened 7/11 than the flashy bar everyone is talking about!*

References

- NY State Approved Liquor Licenses: https://data.ny.gov/Economic-Development/Liquor-Authority-Quarterly-List-of-Active-Licenses/hrvs-fxs2
- Crime Hotspots across NYC: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Map-Year-to-Date-/2fra-mtpn
- Supporting arguments for SNAP: https://www.usich.gov/resources/uploads/asset_library/myths-homeless.pdf
- Visualization Tool to identify major categories of Licenses: https://data.ny.gov/d/83cw-i36h/visualization
- Complaint dataset column descriptors courtesy of cityofnewyork.us: https://data.cityofnewyork.us/api/views/5uac-w243/files/5622c761-4377-4060-a628-ee39ad5c5497?download=true&filename=NYPD_Incident_Level_Data_Column_Descriptions.csv