San Francisco Crime Data Analysis

Objective:

1. Preliminary analysis using historical record to understand criminal patterns and trends in San Francisco
2. Provide actionable recommendations to public health & safety agencies (law enforcement, suicide hotlines, etc.) and helpful guidance to local residents and tourists

Data source: Police Department Incident Reports

- Link: https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry
- Time Period: January 2003 - May 2018

Critical Steps in the Analysis

- Section 0: Data collection and preprocessing
- Section 1: Crime category and quantity analysis
- Section 2: Geographical analysis
- Section 3: Weekly cycle analysis (case study: San Francisco downtown)
- Section 4: Monthly cycle analysis (period: 2015 - 2018)
- Section 5: Daily cycle analysis (case study: pre-holiday December 15 of 2015/2016/2017)
- Section 6: Daily cycle analysis (case study: top 3 dangerous district)
- Section 7: Resolution impact analysis

Get DataFrame and SQL

	PdId	IncidntNum	Incident Code	Category	Descript
1	3114751606302	031147516	06302	LARCENY/THEFT	PETTY THEFT FROM A BUILDING
2	5069701104134	050697011	04134	ASSAULT	BATTERY
3	6074729204104	060747292	04104	ASSAULT	ASSAULT
4	7103536315201	071035363	15201	ASSAULT	STALKING
5	11082415274000	110824152	74000	MISSING PERSON	MISSING ADULT
6	4037801104134	040378011	04134	ASSAULT	BATTERY
7	4147669007025	041476690	07025	VEHICLE THEFT	STOLEN TRUCK
8	16010127305073	160101273	05073	BURGLARY	BURGLARY, UNLAWFUL ENTRY
9	17004924306243	170049243	06243	LARCENY/THEFT	PETTY THEFT FROM LOCKED AUTO
10	16065828006244	160658280	06244	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO
11	18032260316100	180322603	16100	DRUG/NARCOTIC	POSSESSION OF HEROIN
12	17612518006244	176125180	06244	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO
13	17629086806374	176290868	06374	LARCENY/THEFT	GRAND THEFT OF PROPERTY
14	17014483765050	170144837	65050	DRIVING UNDER THE INFLUENCE	DRIVING WHILE UNDER THE INFLUENCE OF ALCOHO
15	17614594028150	176145940	28150	VANDALISM	MALICIOUS MISCHIEF, VANDALISM
16	16006703930000	160067039	30000	OTHER OFFENSES	PERMIT VIOLATION, POLICE (GENERAL)
17	17620614406244	176206144	06244	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO

Spark DataFrame version

	category	count
1	LARCENY/THEFT	477975
2	OTHER OFFENSES	301874
3	NON-CRIMINAL	236928
4	ASSAULT	167042
5	VEHICLE THEFT	126228
6	DRUG/NARCOTIC	117821
7	VANDALISM	114718
8	WARRANTS	99821
9	BURGLARY	91067
10	SUSPICIOUS OCC	79087
11	ROBBERY	54467
12	MISSING PERSON	44268

13	FRAUD	41348
14	FORGERY/COUNTERFEITING	22995
15	SECONDARY CODES	22378
16	WEAPON LAWS	21004
17	TRESPASS	19194

Takeaways:

Quantity should not be the only way to measure crimes.

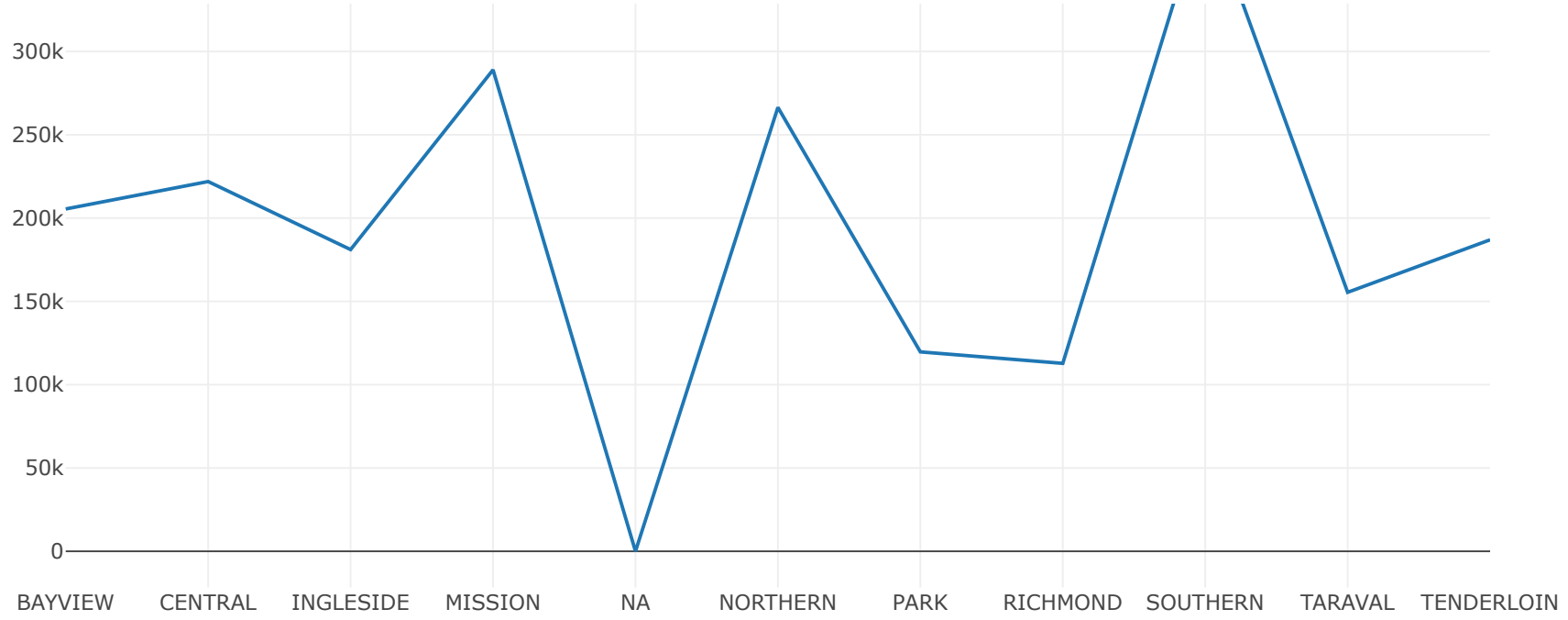
- Certain severe crimes such as murder, rape and missing person didn't even make it to the top 10 but should be categorized as severe and benchmarked against other cities
- Crimes such as theft and assaults, if with certain scale (in monetary value or lives) or when reaching a certain number, should be addressed systematically with government programs
- Certain crimes are hard to define or a combination of several categories and thus grouped as a miscelaneous option called "other offenses". More in-depth analysis should be done

In further investigation, propose prioritizing top items with different dimensions of consideration, such as

- high quantity and consistency (theft)
- high severity (murder)
- public complaints (human feces on sidewalks in early 2018 grouped in vandalism)

Futhermore, geographical analysis should be done on where crimes happen within San Francisco


400k
350k



Spark SQL Version

	PdDistrict	Count
1	SOUTHERN	390692
2	MISSION	288985
3	NORTHERN	266435
4	CENTRAL	221923
5	BAYVIEW	205480
6	TENDERLOIN	186954
7	INGLESIDE	181092
8	TARAVAL	155461
9	PARK	119698
10	RICHMOND	112804
11	NA	1

Takeaways:

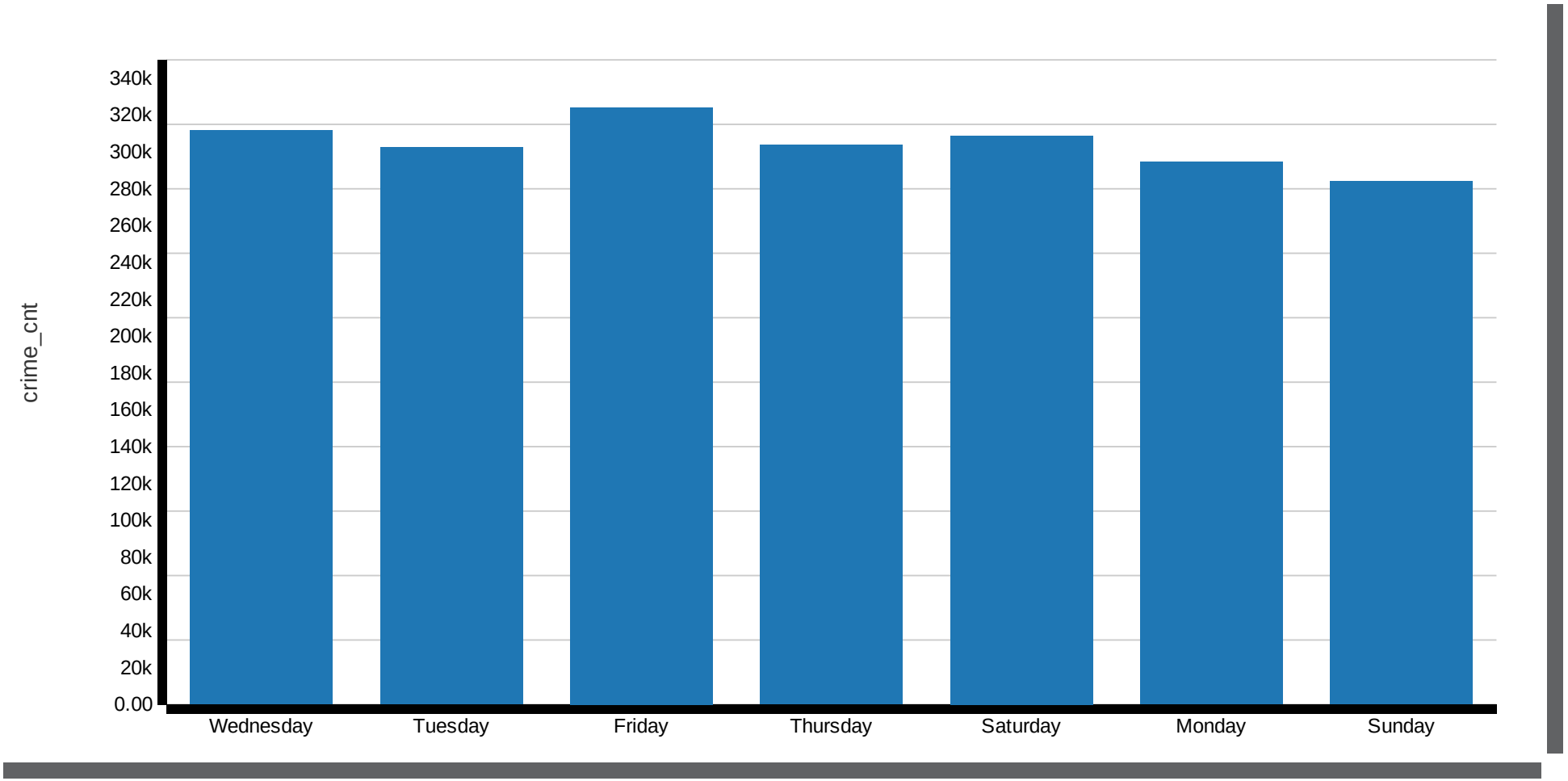
Quantity should not be the only way to measure whether a region is safe or not

- analysis should be done with combination of section 1 (type of crimes)

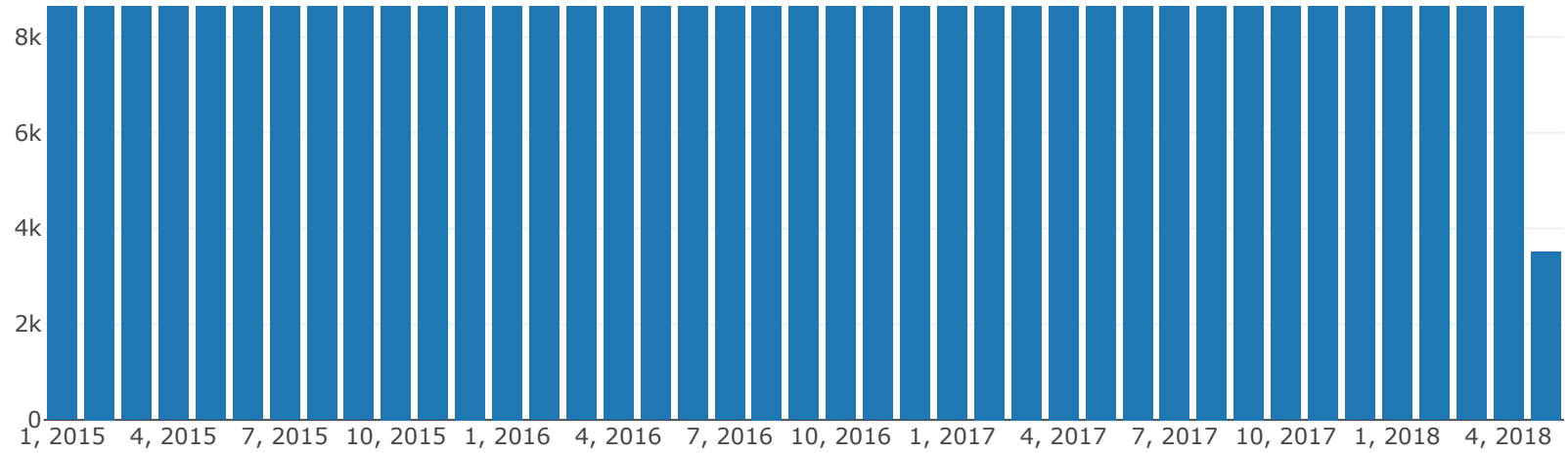
- it is also important to understand the size and location of the district

With 15 years of data, further analysis should be done on the following dimension:

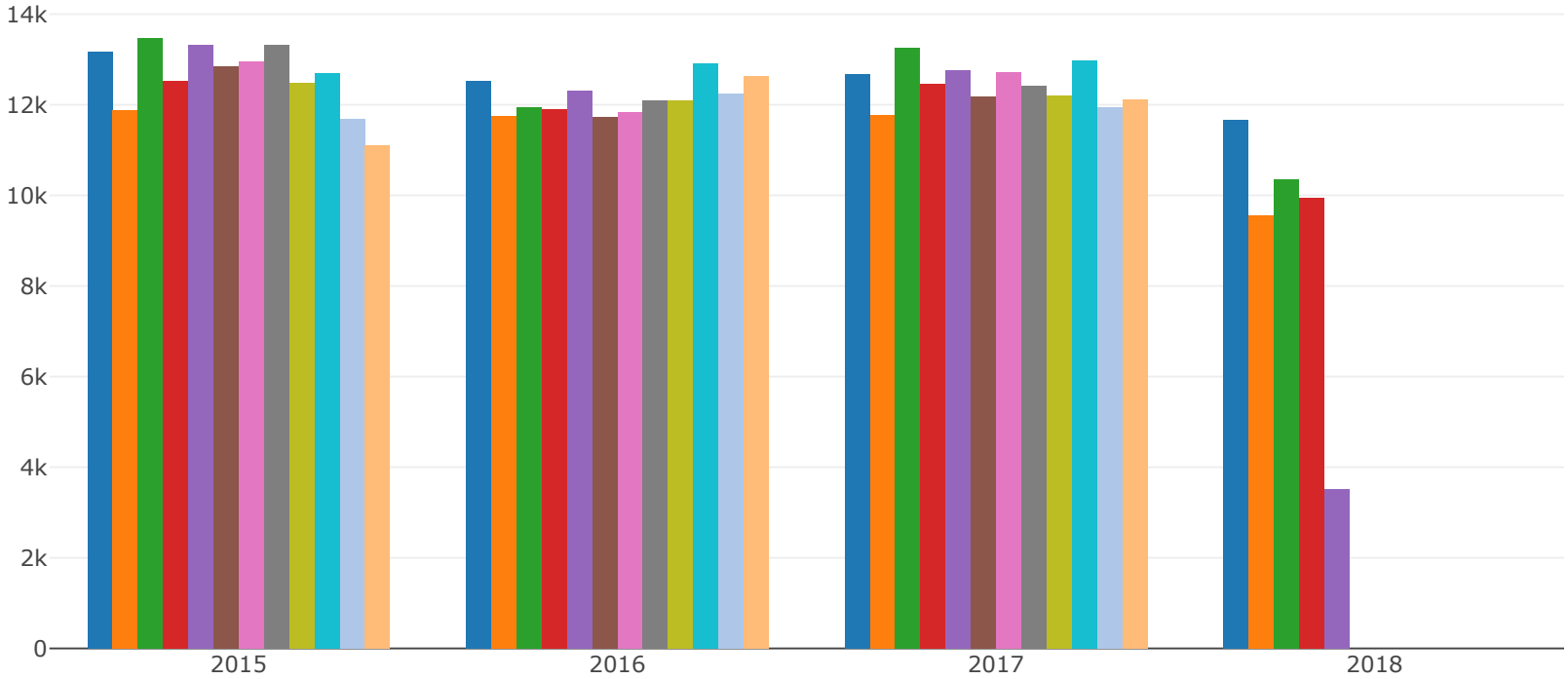
- trend analysis on how crime rates have changed over the years per region
- case studies should be done for certain region of importance to understand crime cycles


crime_cnt
340k
320k
300k
280k
260k
240k
220k
200k
180k
160k
140k
120k
100k
80k
60k
40k
20k
0.00
Wednesday
Tuesday
Friday
Thursday
Saturday
Monday
Sunday


Plot monthly trend
14k
12k
10k



Show Monthly Count



Extract hours from incident time

	IncidntNum	Category	Descript	DayOfWeek	Date	Time	PdDistrict	Resolut
1	031147516	LARCENY/THEFT	PETTY THEFT FROM A BUILDING	Sunday	09/28/2003	10:00	SOUTHERN	NONE
2	050697011	ASSAULT	BATTERY	Wednesday	06/22/2005	12:20	NORTHERN	NONE
3	060747292	ASSAULT	ASSAULT	Saturday	07/15/2006	00:55	CENTRAL	NONE
4	071035363	ASSAULT	STALKING	Tuesday	09/25/2007	00:01	TARAVAL	NONE
5	110824152	MISSING PERSON	MISSING ADULT	Saturday	09/24/2011	11:00	TARAVAL	LOCATI

Section 6: Daily cycle analysis (case study: top 3 dangerous district)

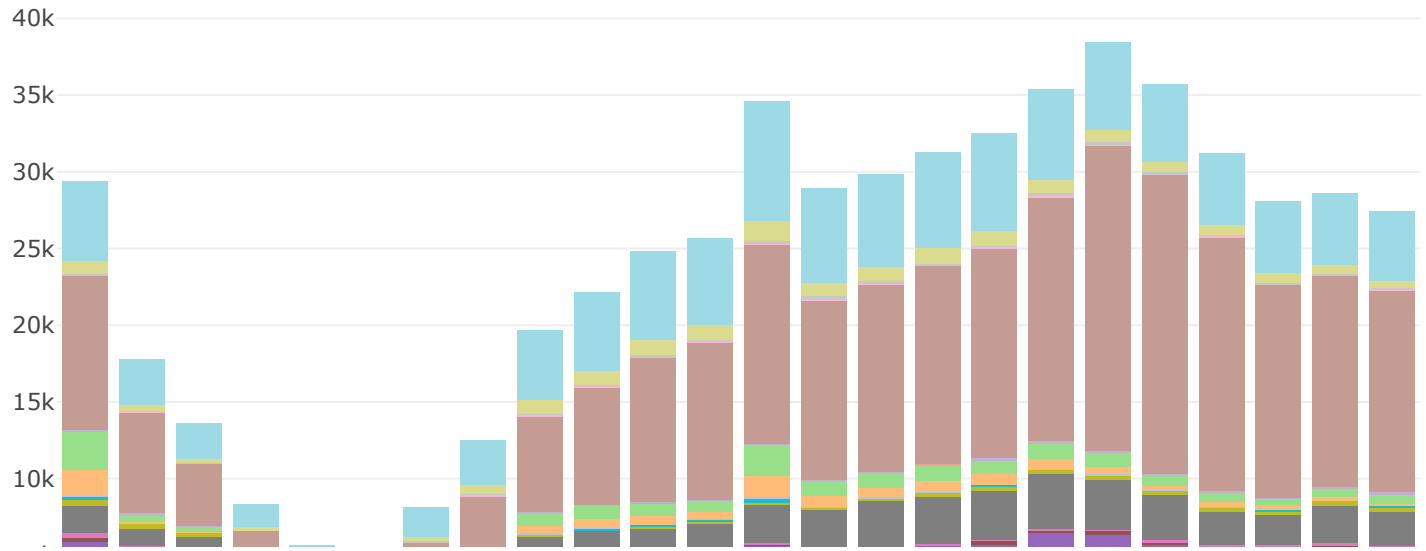
- step 1: find out the regions with highest crime volume
- step 2: drill down to the crime category of each region

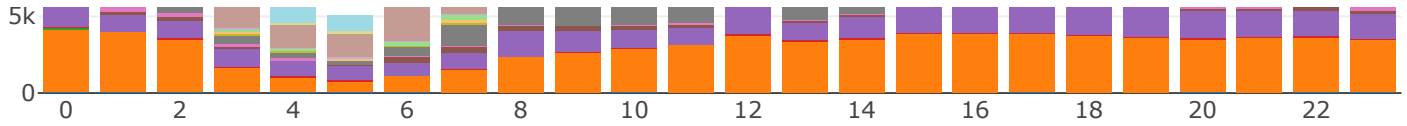
Spark DataFrame Version By Region

	PdDistrict	count
	SOUTHERN	200000

1	SOUTHERN	390692
2	MISSION	288985
3	NORTHERN	266435
4	CENTRAL	221923
5	BAYVIEW	205480
6	TENDERLOIN	186954
7	INGLESIDE	181092
8	TARAVAL	155461
9	PARK	119698
10	RICHMOND	112804
11	NA	1

Drill down to Crime Category for the top 3 regions by the hours





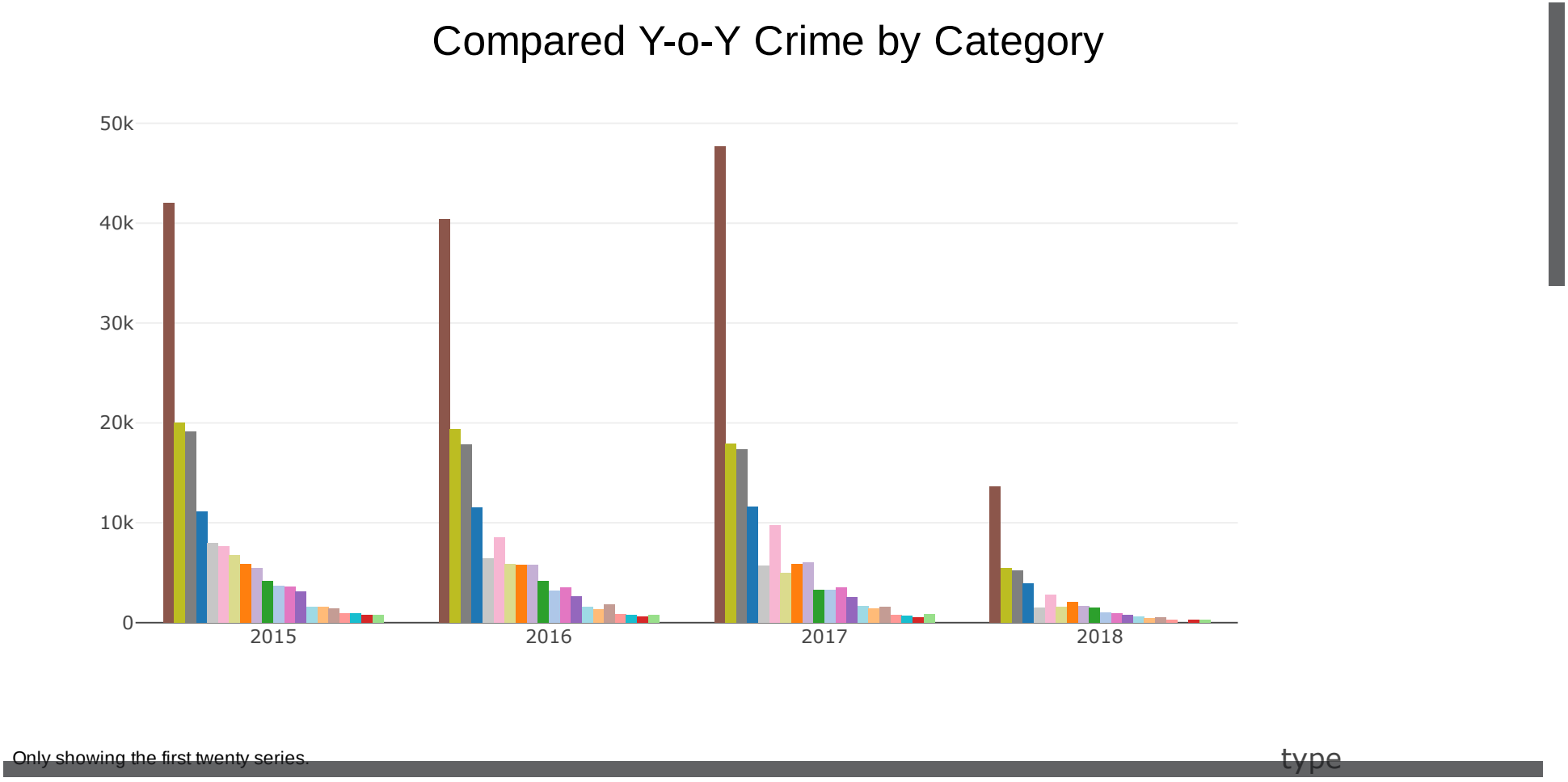
Only showing the first twenty series.

Category

Get Resolution categories

	resolve
1	EXCEPTIONAL CLEARANCE

2	ARREST, BOOKED
3	PROSECUTED FOR LESSER OFFENSE
4	LOCATED
5	UNFOUNDED
6	DISTRICT ATTORNEY REFUSES TO PROSECUTE
7	PSYCHOPATHIC CASE
8	COMPLAINANT REFUSES TO PROSECUTE
9	ARREST, CITED
10	PROSECUTED BY OUTSIDE AGENCY
11	NONE
12	NOT PROSECUTED



Takeaways:

- The top four categories of crimes resolved are STOLEN PROPERTIES, WARRANTS, DRIVING UNDER THE INFLUENCE, DRUG/NARCOTIC.
- The categories which less than 10% of crimes resolved are RECOVERED VEHICLE, VEHICLE THEFT, and LARCENY/THEFT.

- Can increase the police force against theft crimes

Conclusion.

Key Takeaways:

- Quantity should not be the only dimension to measure crimes, or whether a region is safe or not.
- Potential Improvement: Severity, impact and combinations of crimes should be taken into consideration for more insights, with consideration of trend analysis
- Day of week or month is not sensitive to crime rate, but significant events or holiday seasons could increase crime volume of certain categories significantly.
- Recommendation: increase policy patrol during significant events and pre-holiday seasons
- 12:00pm and 18:00pm are peak crime hours and 2:00AM - 6:00AM is non-peak crime hours.
- Recommendation: increase policy patrol during peak hours
- Theft, assault and drug/narcotic are the top categories of crimes with lowest resolution.
- Recommendation: increase policy allocation and crack down of these crime categories with better collaboration of local community