San Francisco

Crime

- What is the question?
- Data Sources
- Data Cleanup
- What, When, and Where of Crime?
- Modeling
- Conclusion

What is the question we are trying to answer?

San Francisco was famous for hosting some of the world's most notorious criminals on the inescapable island of Alcatraz. Today, the island is a tourist attraction but crime is still a social phenomena.

Is it possible to create a model to predict crime type, time, and/or location?

San Francisco

Founded: 1849

Population: 874,784 (Bay Area: 7m)

Area: 7 miles in both width and length (49 miles squared)

Fun Facts: North America's Oldest China Town

The fortune cookie was born here

City has 50+ hills

What data?

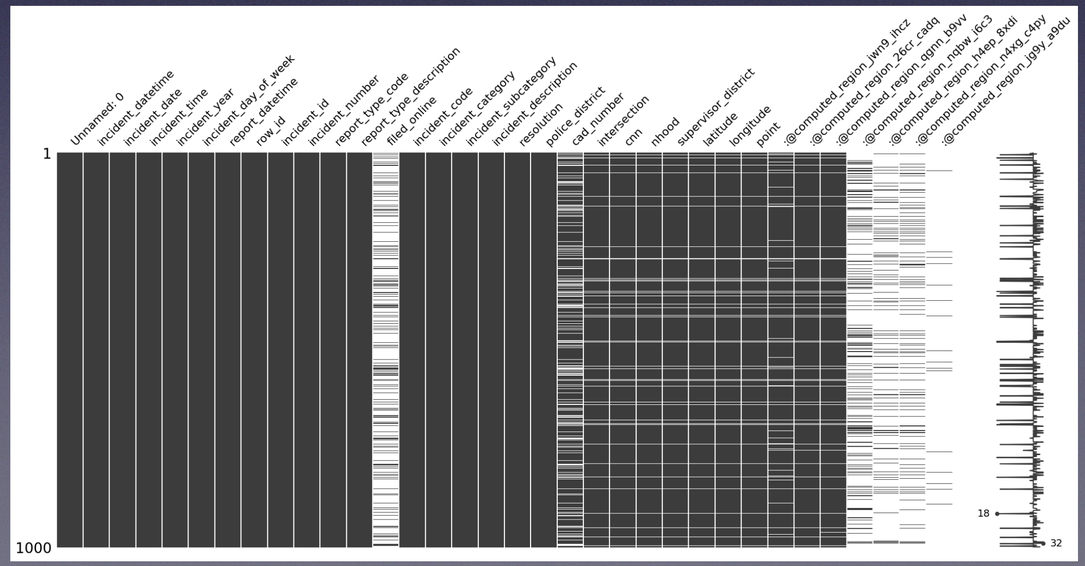
- **Crime**: SFPD Incident Report: 2018 to Present
- **Demographic**: America Community Survey 2020 5 Year
- **Geological**: SF Neighborhood to Census Tract

SFPD Crime Data

- The dataset includes incident reports that have been filed as of January 1, 2018 with an hourly frequency and published daily.
- These reports are filed by officers or self-reported by members of the public using SFPD's online reporting system.
- Coordinates associated with incident locations provided within the dataset are anonymized and reflect the nearest intersection of each occurrence.
- In 2020, 7,414 confidential flagged reports were retained by the department, or about 5.8% of the unredacted dataset. Of the 7,414 reports, 53% were also flagged as domestic violence reports.
- The removal of juvenile related data in 2020 resulted in the retention of 3,581 records, or about 2.82% of the unredacted 2020 dataset.

Data Clean Up

```
$ ...1                          : num [1:664280] 0 1 2 3 4 5 6 7 8 9 ...
$ incident_datetime             : POSIXct[1:664280], format: "2021-07-25 00:00:00" "2022-06-28 23:58:00" "2022-03-11 10:30:00" "2021-05-15 17:47:00" ...
$ incident_date                 : POSIXct[1:664280], format: "2021-07-25" "2022-06-28" "2022-03-11" "2021-05-15" ...
$ incident_time                 : 'hms' num [1:664280] 00:00:00 23:58:00 10:30:00 17:47:00 ...
$ incident_year                 : num [1:664280] 2021 2022 2022 2021 2022 ...
$ incident_day_of_week          : chr [1:664280] "Sunday" "Tuesday" "Friday" "Saturday" ...
$ report_datetime               : POSIXct[1:664280], format: "2021-07-25 13:41:00" "2022-06-28 23:58:00" "2022-03-11 20:03:00" "2021-05-15 17:47:00" ...
$ row_id                        : num [1:664280] 1.06e+11 1.17e+11 1.13e+11 1.03e+11 1.17e+11 ...
$ incident_id                   : num [1:664280] 1057189 1165543 1130480 1030518 1165351 ...
$ incident_number               : num [1:664280] 2.16e+08 2.20e+08 2.26e+08 2.10e+08 2.20e+08 ...
$ report_type_code              : chr [1:664280] "II" "VS" "II" "VS" ...
$ report_type_description       : chr [1:664280] "Coplogic Initial" "Vehicle Supplement" "Coplogic Initial" "Vehicle Supplement" ...
$ filed_online                  : logi [1:664280] TRUE NA TRUE NA NA TRUE ...
$ incident_code                 : chr [1:664280] "06372" "71012" "71000" "07043" ...
$ incident_category             : chr [1:664280] "Larceny Theft" "Other Offenses" "Lost Property" "Recovered Vehicle" ...
$ incident_subcategory          : chr [1:664280] "Larceny Theft - Other" "Other Offenses" "Lost Property" "Recovered Vehicle" ...
$ incident_description          : chr [1:664280] "Theft, Other Property, $50-$200" "License Plate, Recovered" "Lost Property" "Vehicle, Recovered, Motorcycle" ...
$ resolution                    : chr [1:664280] "Open or Active" "Open or Active" "Open or Active" "Open or Active" ...
$ police_district               : chr [1:664280] "Southern" "Out of SF" "Central" "Out of SF" ...
$ cad_number                    : num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ intersection                  : chr [1:664280] NA NA NA NA ...
$ cnn                           : num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ nhood                         : chr [1:664280] NA NA NA NA ...
$ supervisor_district           : num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ latitude                      : num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ longitude                     : num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ point                         : chr [1:664280] NA NA NA NA ...
$ :@computed_region_jwn9_ihcz: num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ :@computed_region_26cr_cadq: num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ :@computed_region_qgnn_b9vv: num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ :@computed_region_nqbw_i6c3: num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ :@computed_region_h4ep_8xdi: num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ :@computed_region_n4xg_c4py: num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
$ :@computed_region_jg9y_a9du: num [1:664280] NA NA NA NA NA NA NA NA NA NA ...
```

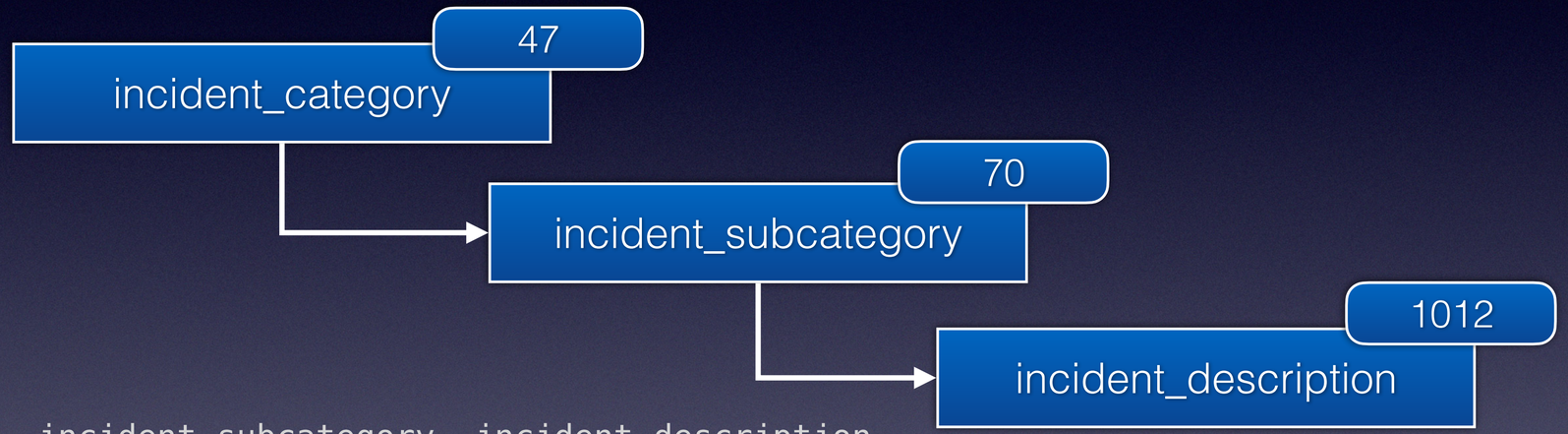


> 600k Observations **Missing Information** **Irrelevant Information**

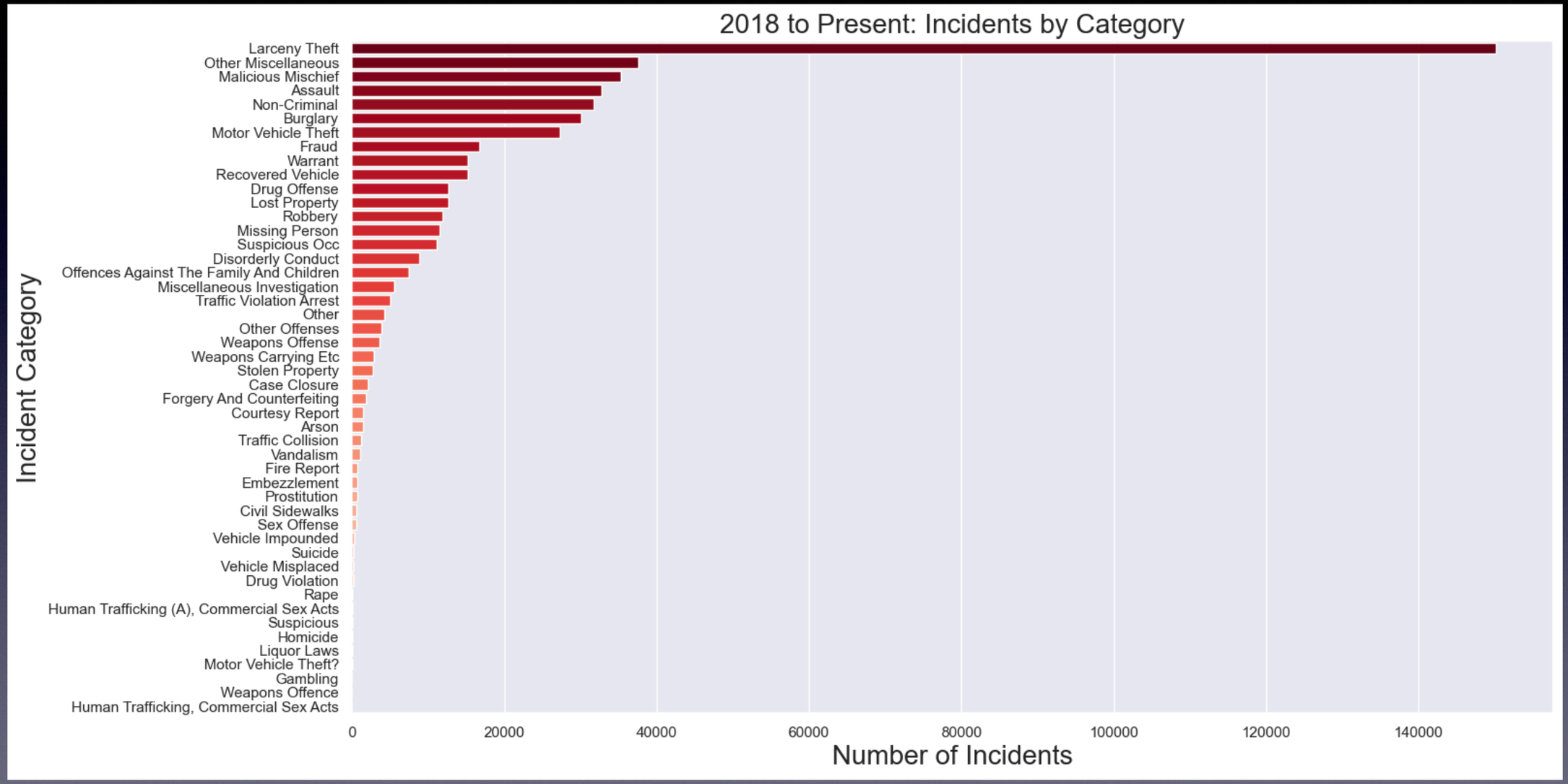
~ 124k Duplicates **26 Columns**

What is crime?

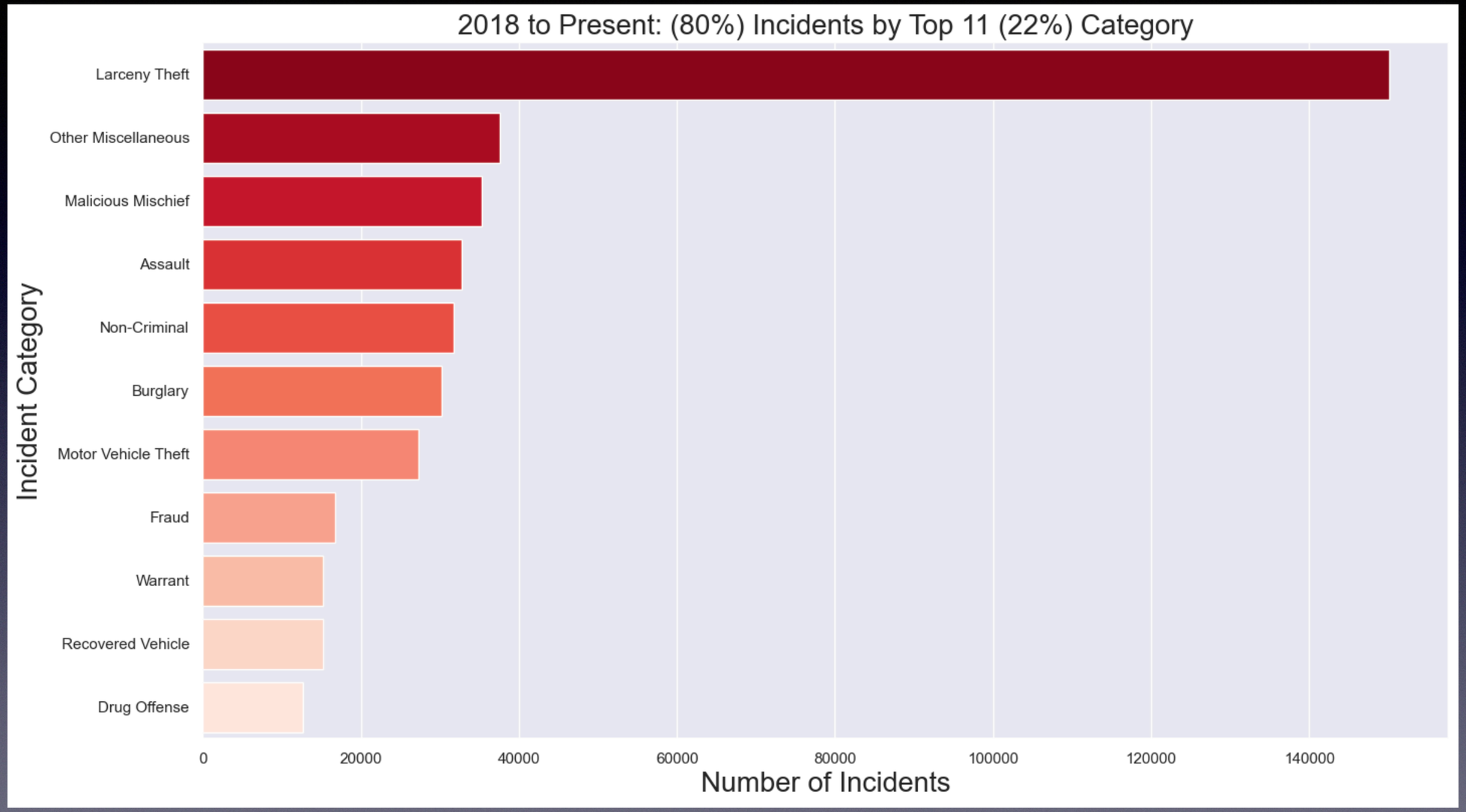
The reports are categorized into the following groups based on the type of incident:

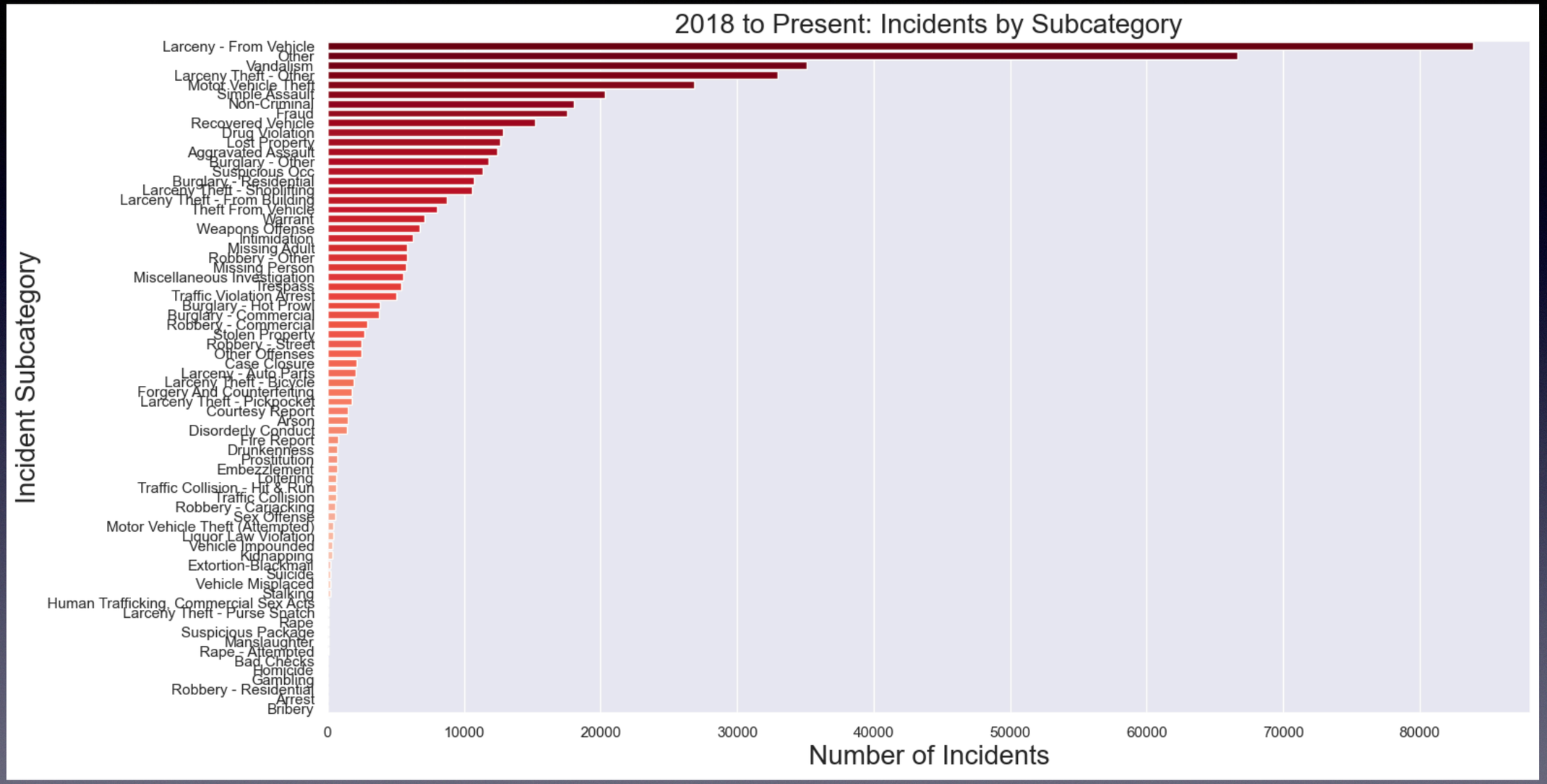


incident_category	incident_subcategory	incident_description	
Arson	Arson	Arson	805
		Arson of Vehicle	351
		Arson with Great Bodily Injury	2
		Arson, Attempted	70
		Arson, Commercial Building	95
			...
Weapons Offense	Weapons Offense	Firearm, Discharging Within City Limits	909
		Weapon, Carrying Concealed	483
		Weapon, Carrying Concealed with Felony or Narcotics Conviction	78
		Weapon, Deadly, Imitation or Laser Scope, Exhibiting	2086
		Weapon, Deadly, Possession W/ Intent To Assault	44

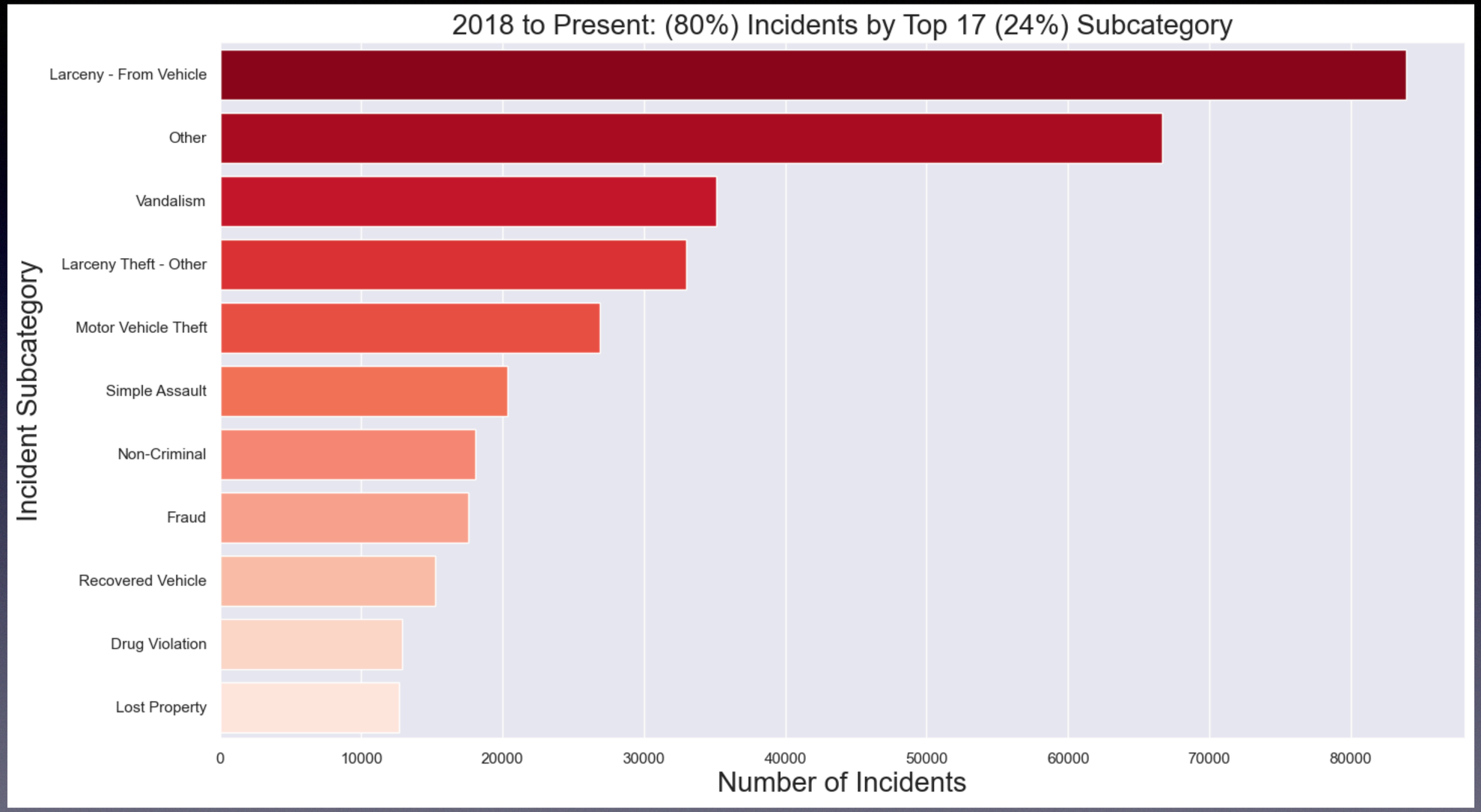


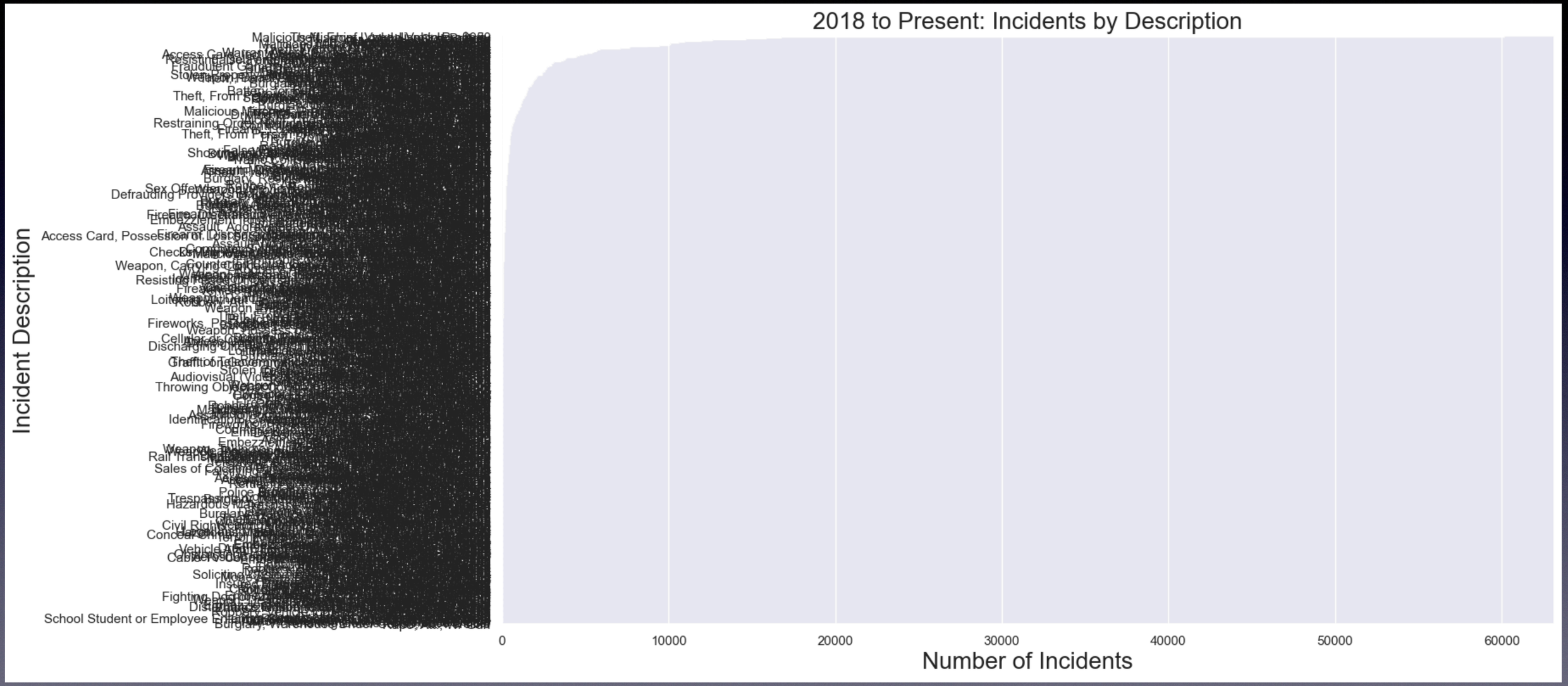
47 Categories


2018 to Present: (80%) Incidents by Top 11 (22%) Category
Incident Category
Larceny Theft
Other Miscellaneous
Malicious Mischief
Assault
Non-Criminal
Burglary
Motor Vehicle Theft
Fraud
Warrant
Recovered Vehicle
Drug Offense
0
20000
40000
60000
80000
100000
120000
140000
Number of Incidents

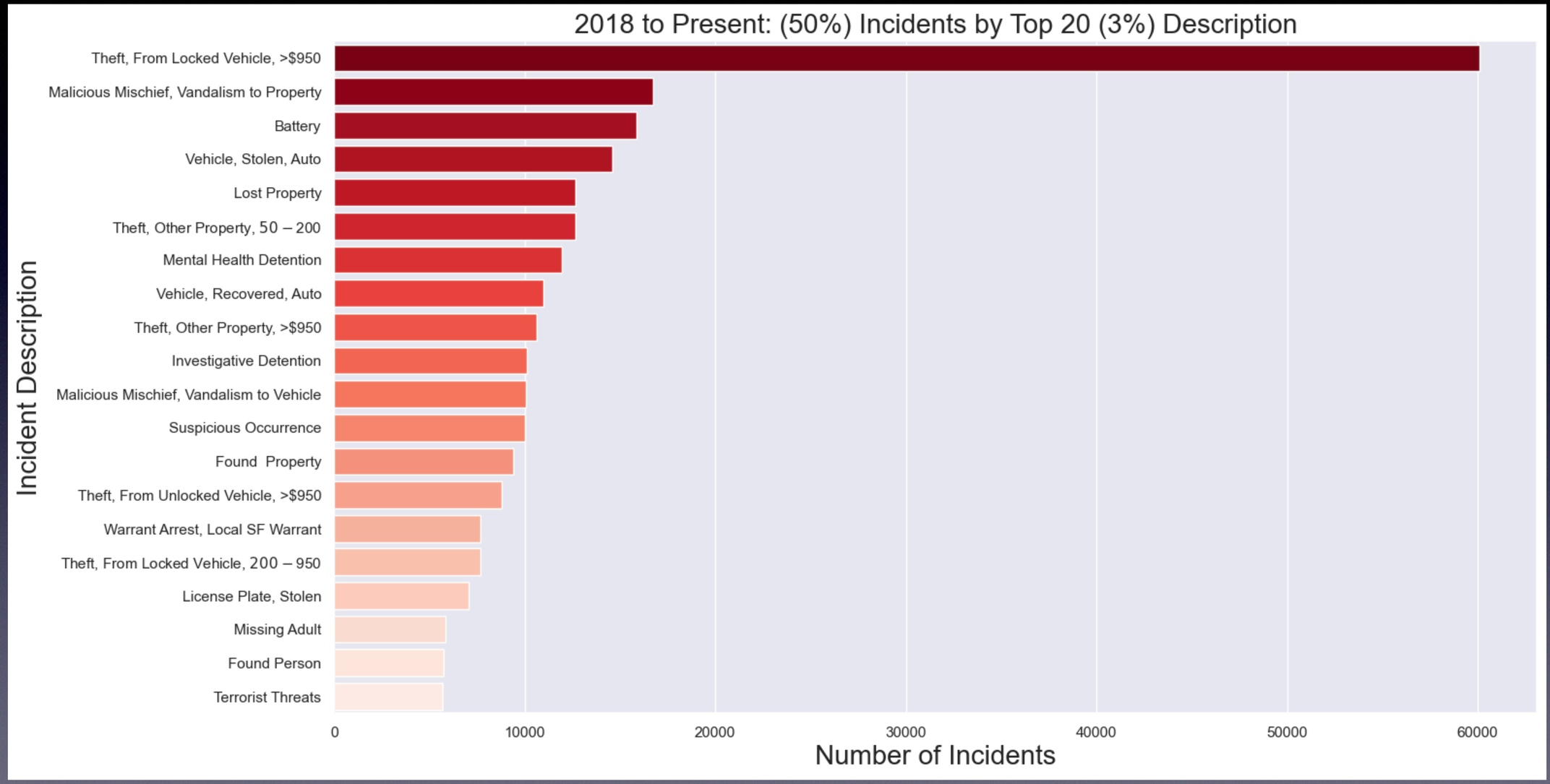


70 Subcategories


2018 to Present: (80%) Incidents by Top 17 (24%) Subcategory
Incident Subcategory
Larceny - From Vehicle
Other
Vandalism
Larceny Theft - Other
Motor Vehicle Theft
Simple Assault
Non-Criminal
Fraud
Recovered Vehicle
Drug Violation
Lost Property
0
10000
20000
30000
40000
50000
60000
70000
80000
Number of Incidents

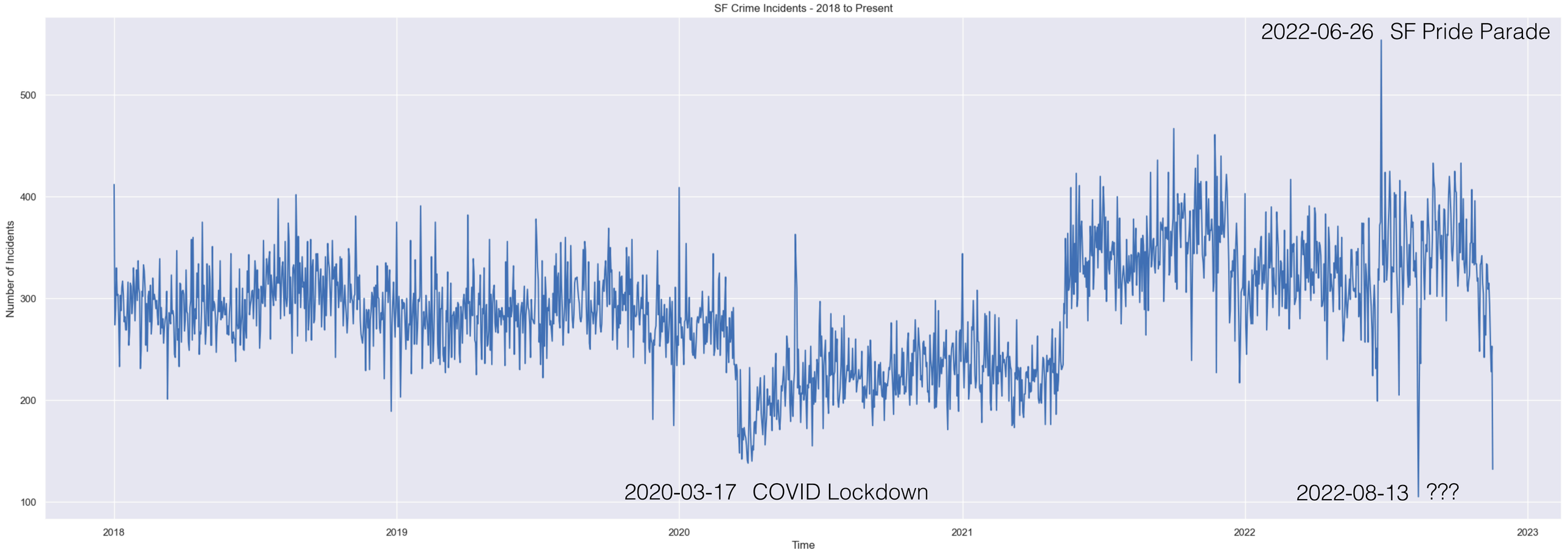


1012 Descriptions (incident codes)



Key take-away, don’t leave valuables in your car when in SF

When is crime?



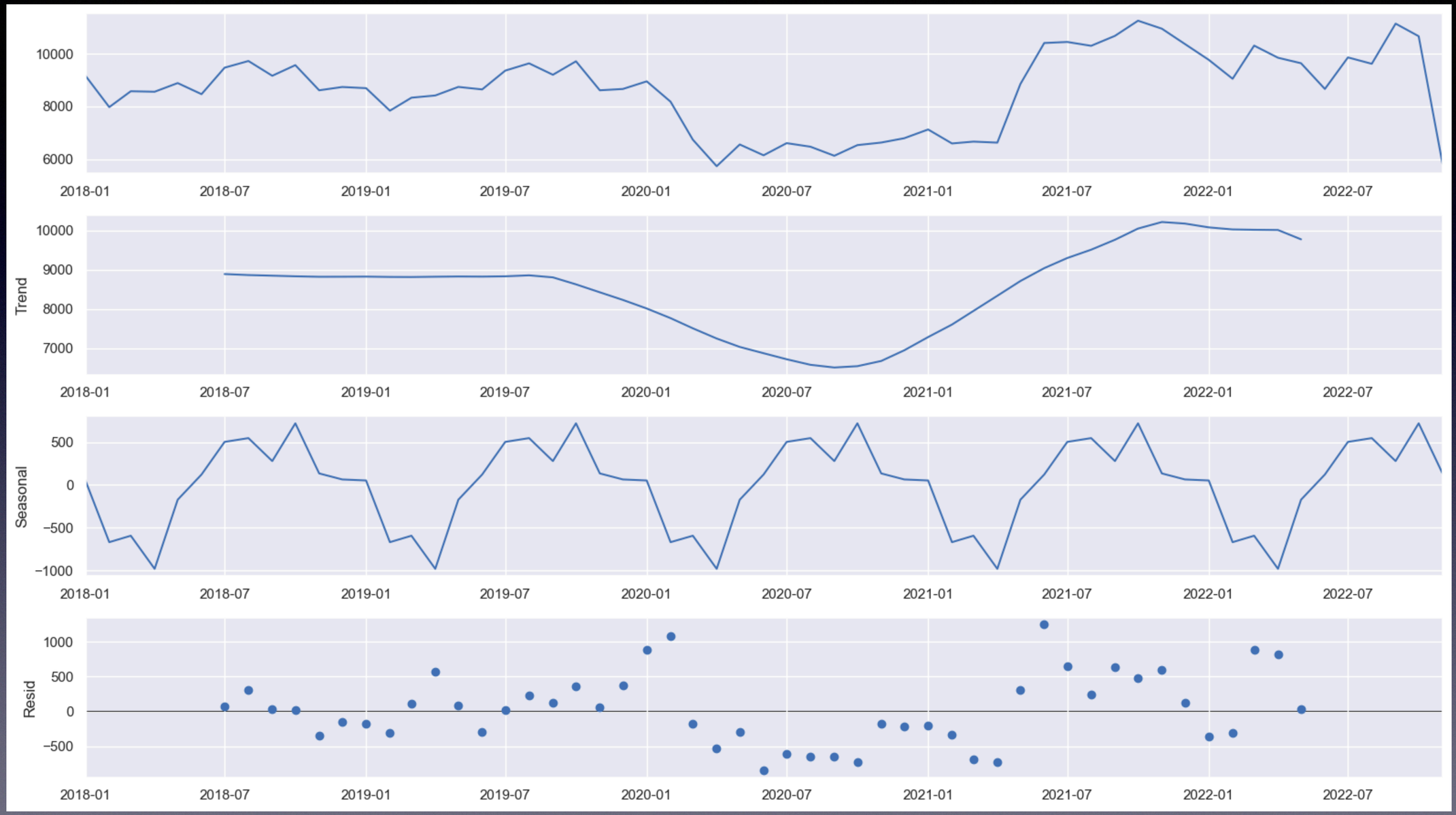
2018 to Present (202211)

Format: YYYY-MM-DD HH:MM:SS

Seasonality?

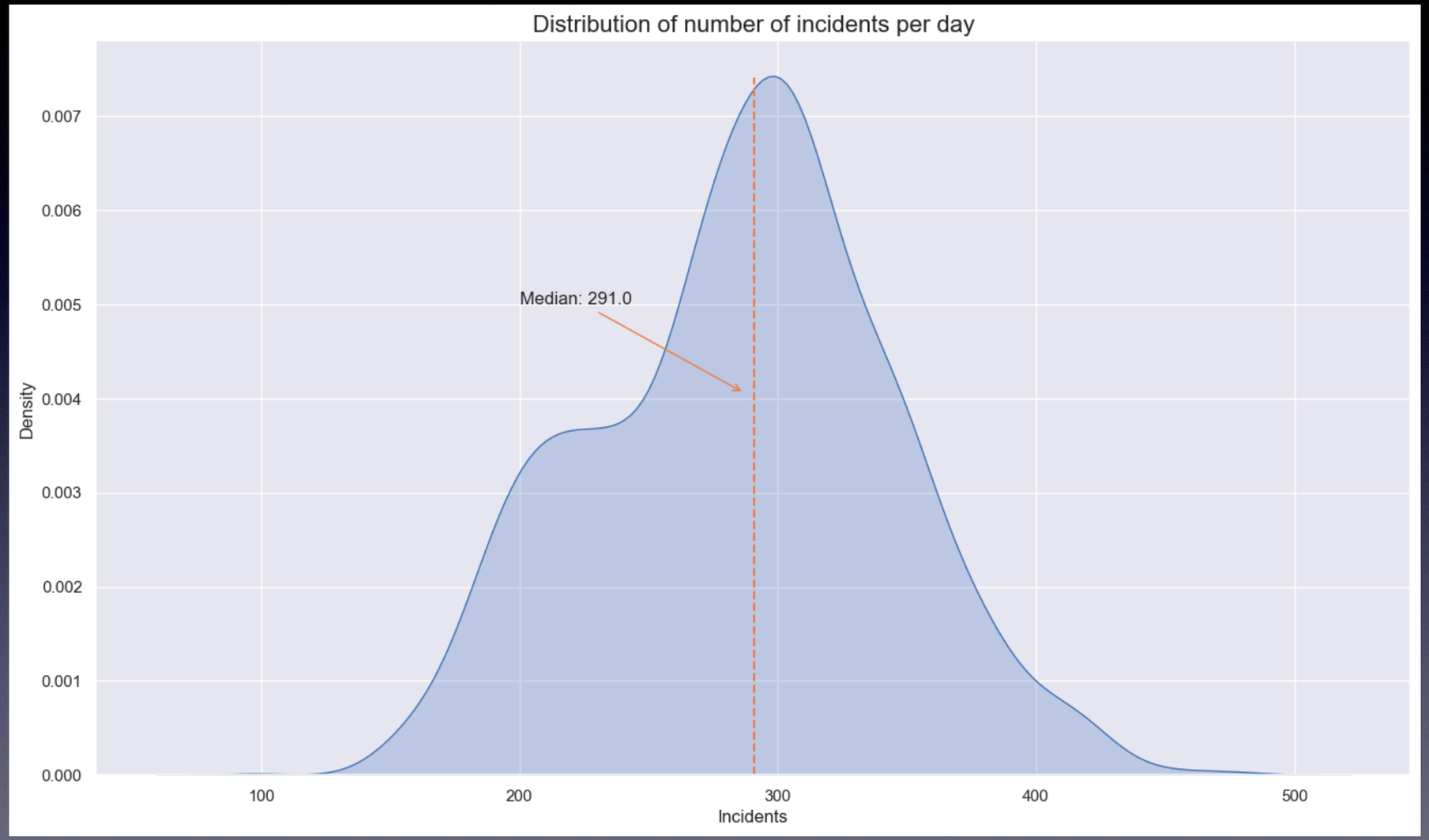
Using daily totals for a Time Series Additive Decomposition is rather noisy and doesn't provide many insights

Format: YYYY-MM-DD HH:MM:SS



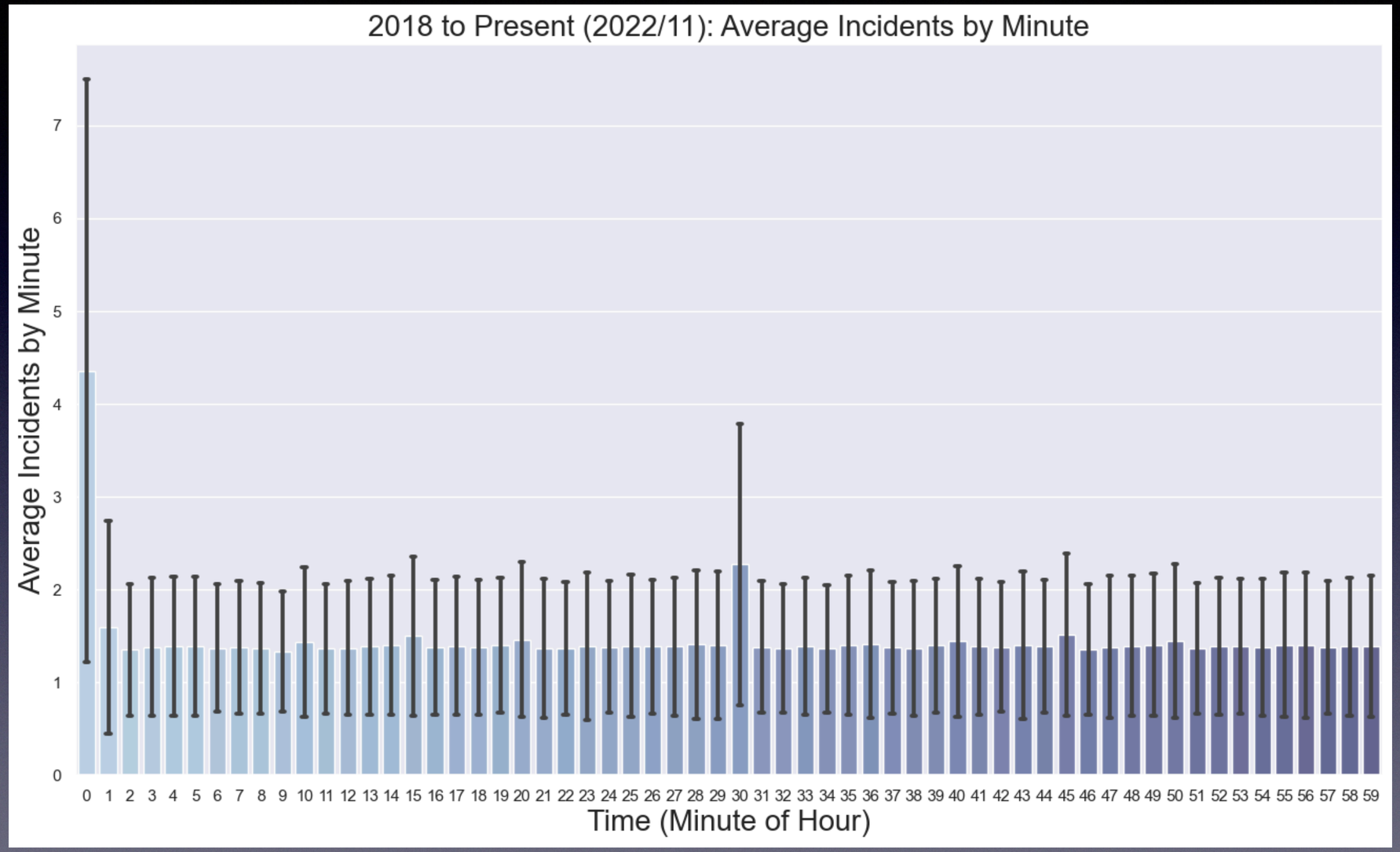
Seasonality

Using monthly totals makes seasonality clearer with distinct trend and yearly peaks around September and troughs in April



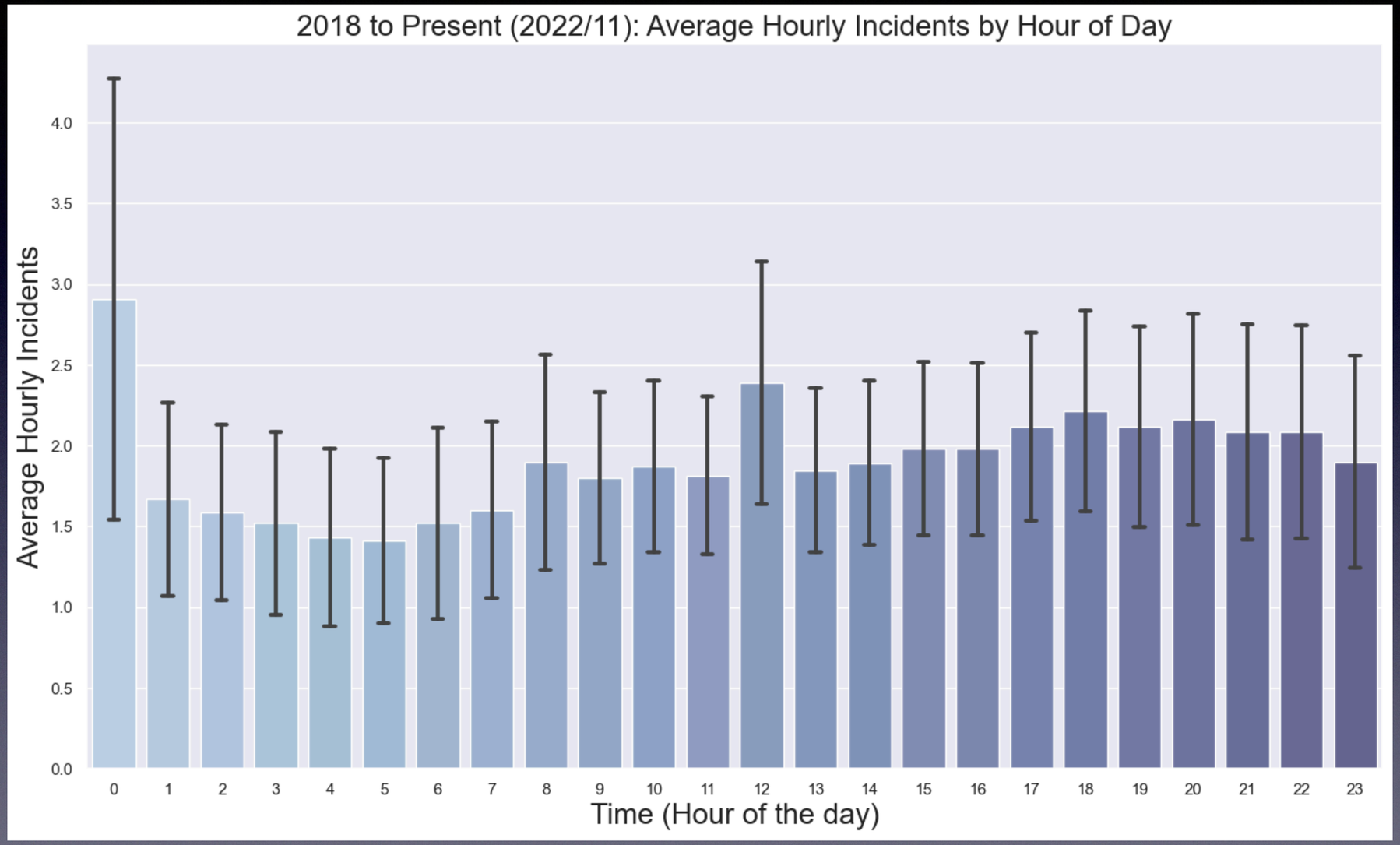
Distribution

There is an average of 291 incident per day



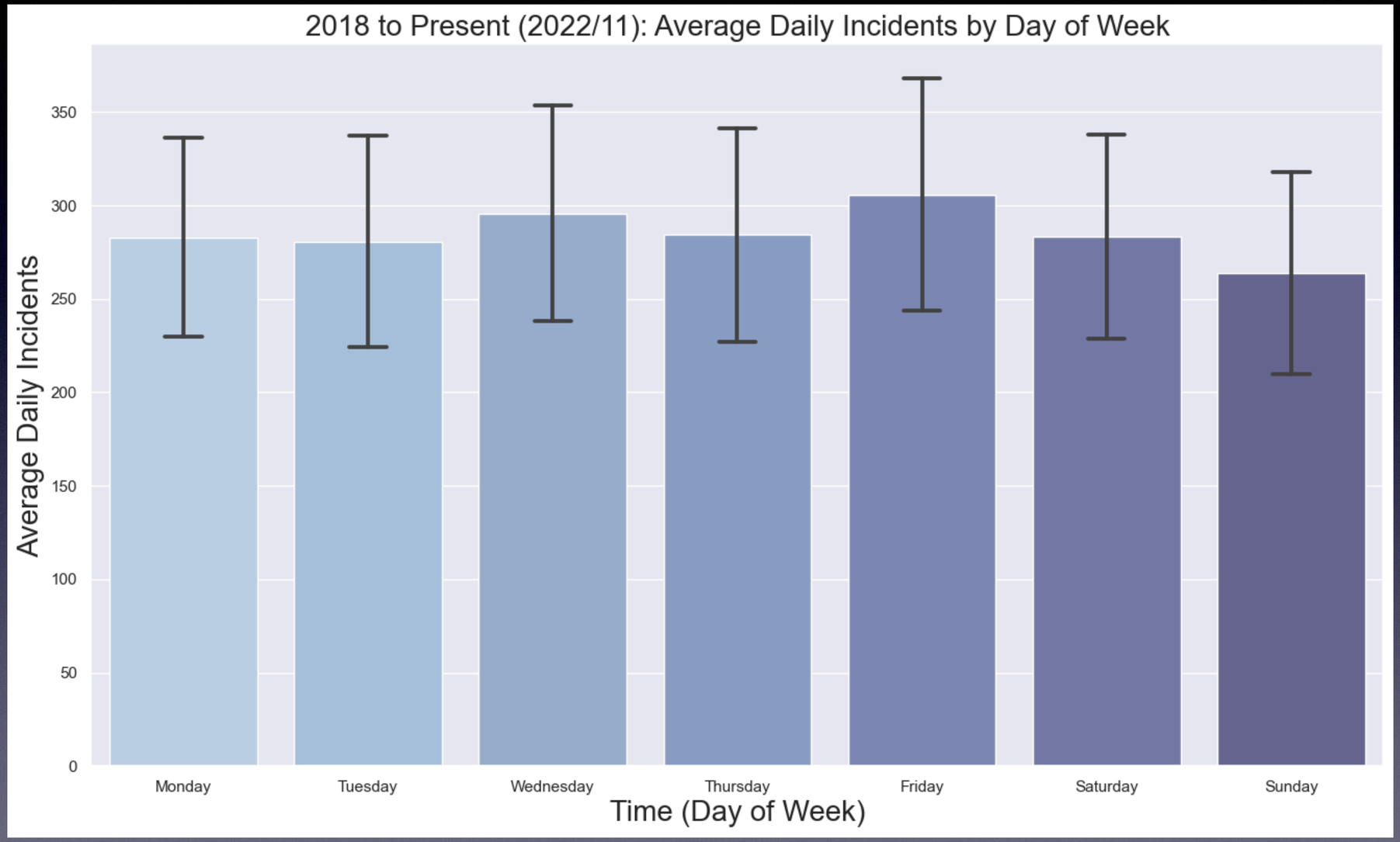
Notice anything?

There is a reporting bias to round incident minute to the top (:00) or bottom (:30) of the hour

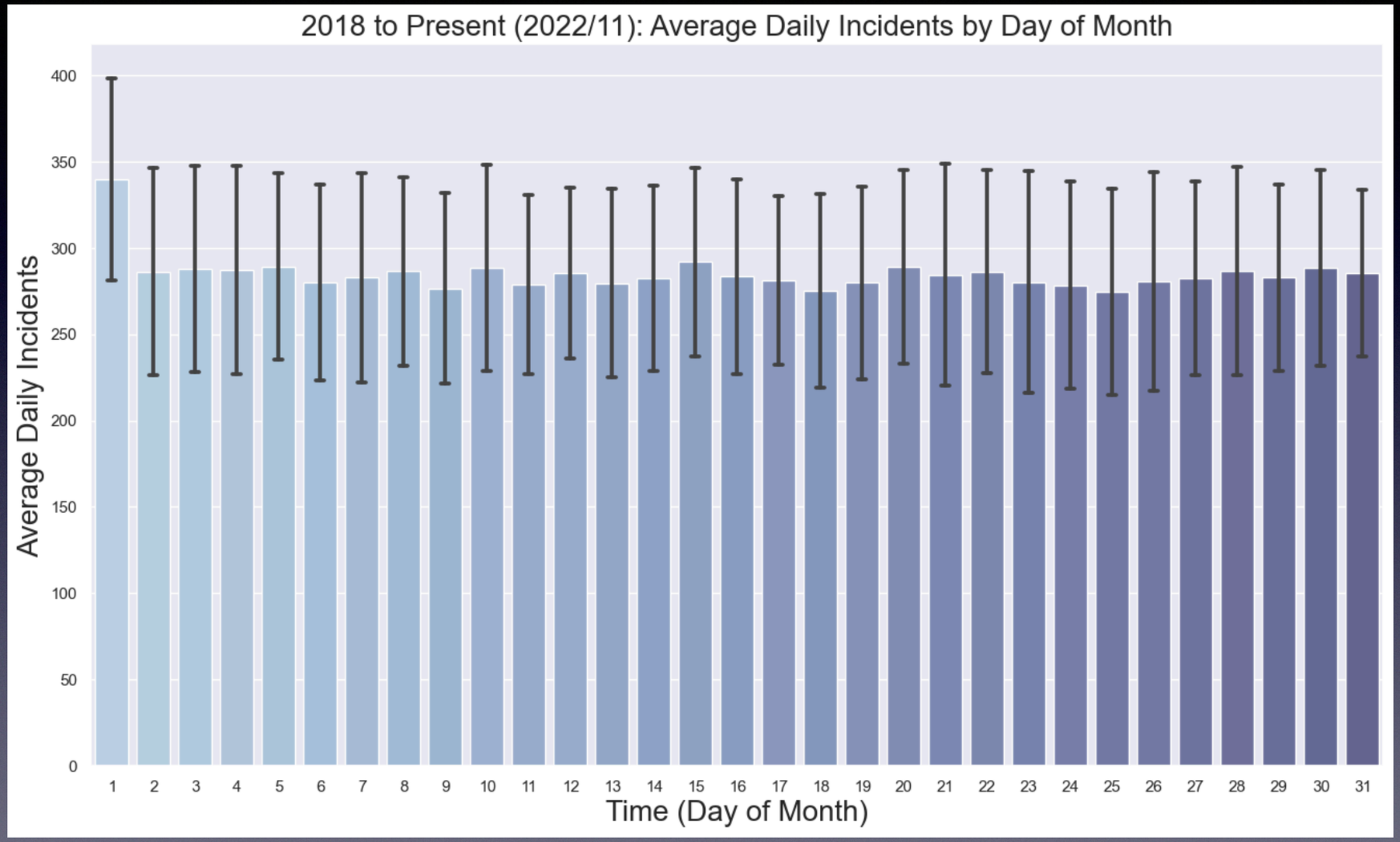


Again?

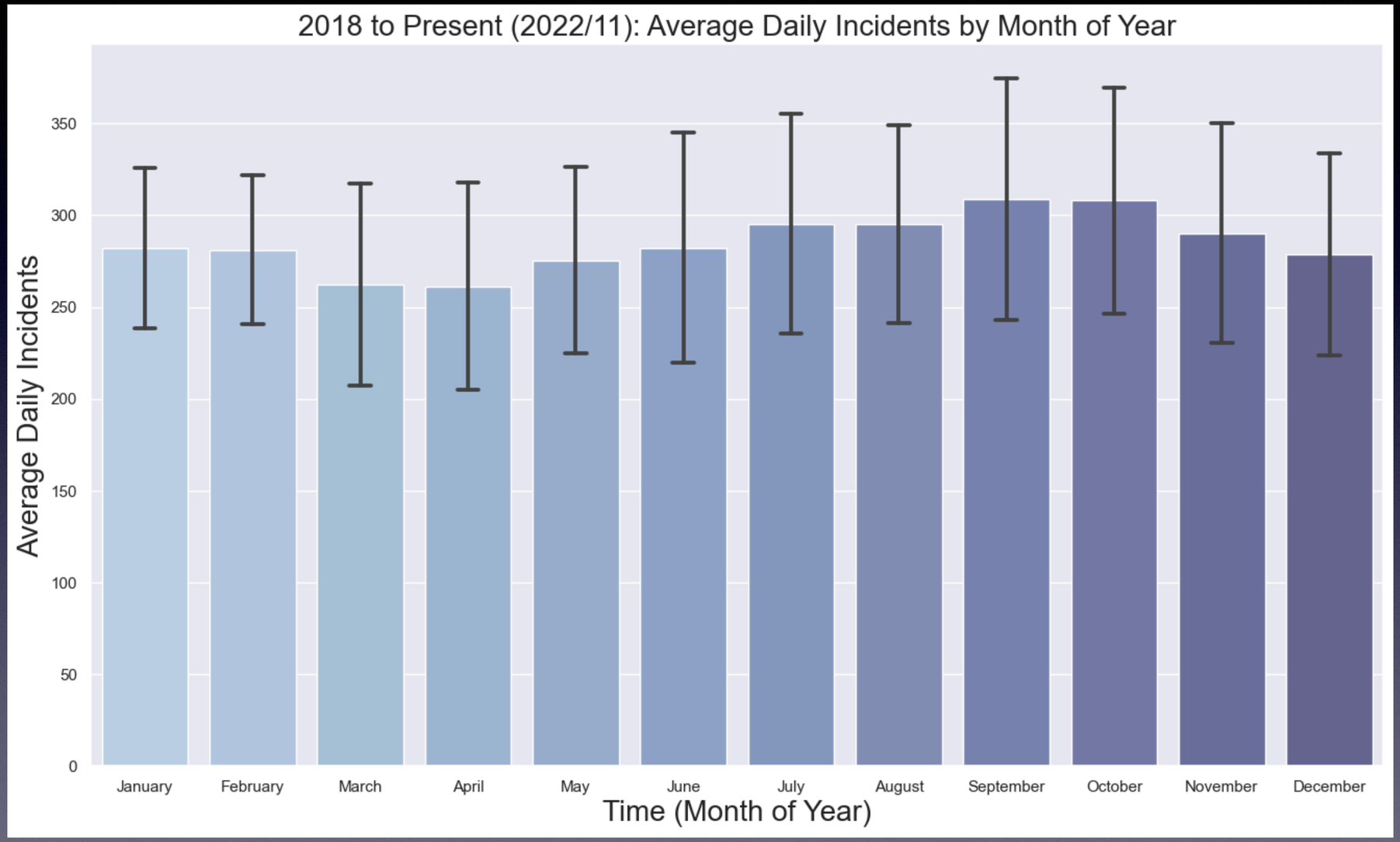
There is a reporting bias to round hours to the bottom (midnight) or middle (noon) of the day with a peak around 5-6pm (commuting hours) and a lull around 5am



Friday seems to be the most active day of the week with Sunday the least active and a small spike on Wednesdays

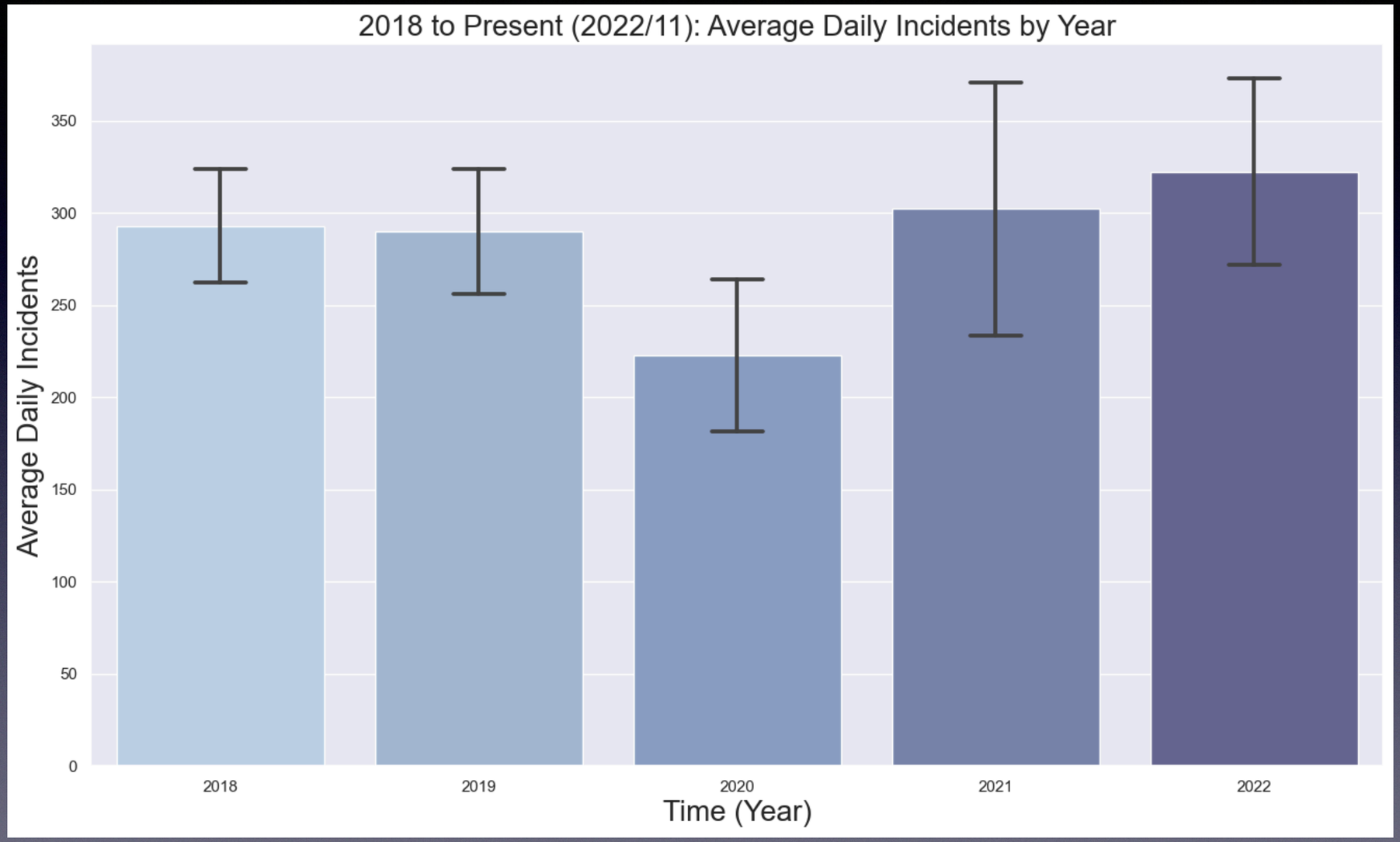


Again, there seems to be a reporting bias to round to the first of the month



Seasonality?

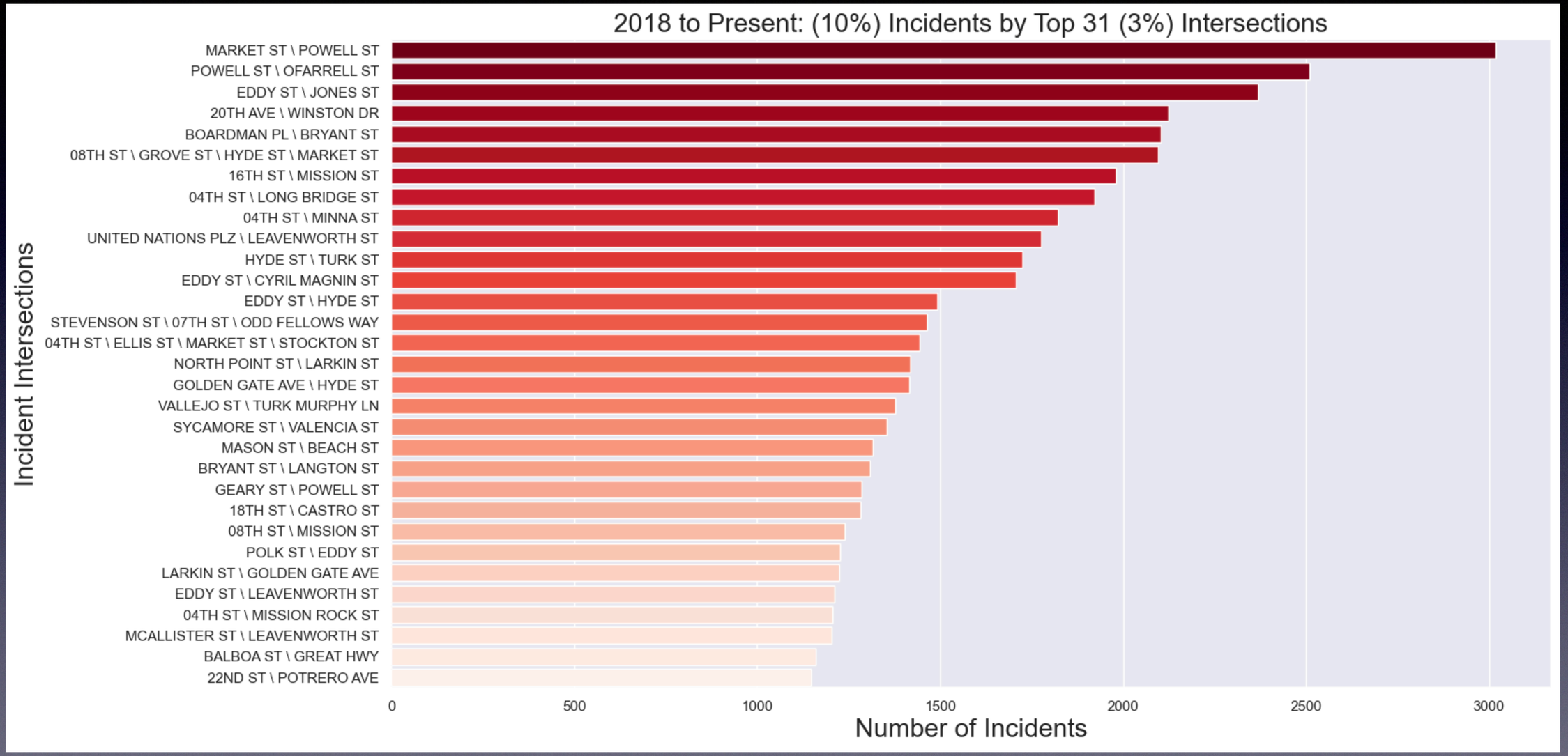
There seems to be a peak around September and October that tappers off until April before increasing again



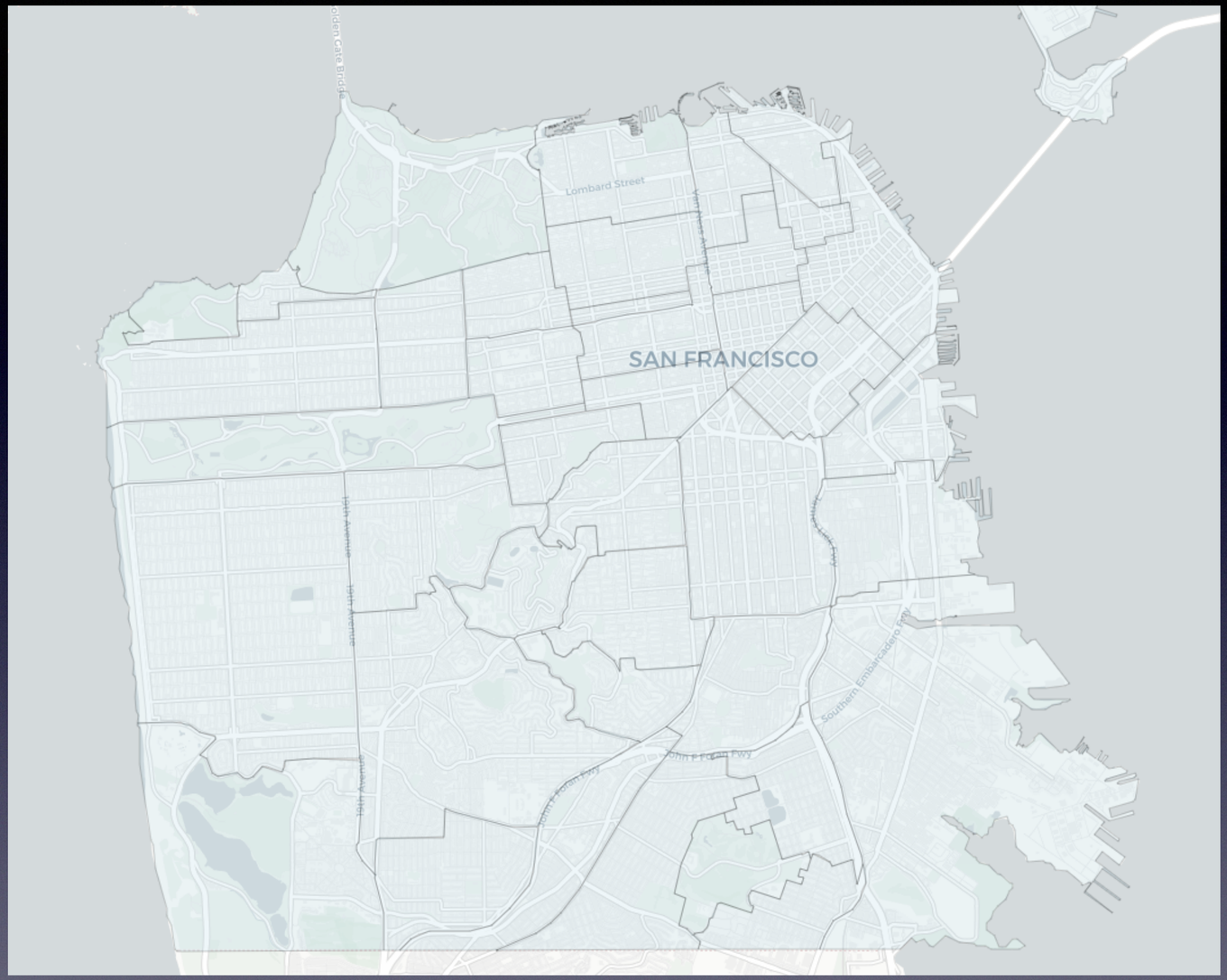
The COVID lockdown during 2020 resulted in the least number of incidents and has now rebounded higher than pre-pandemic

Where is crime?

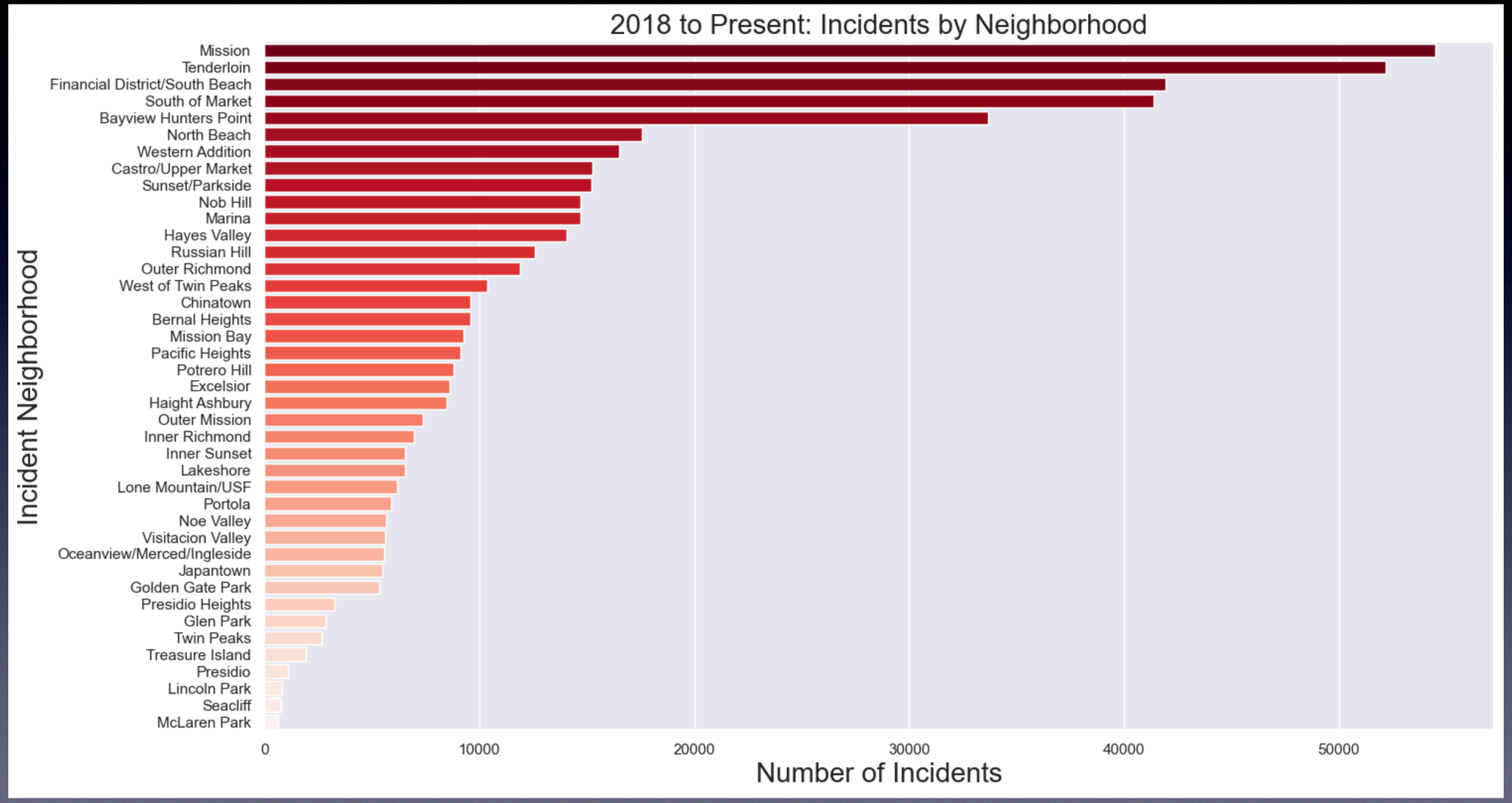
SF Map with intersections



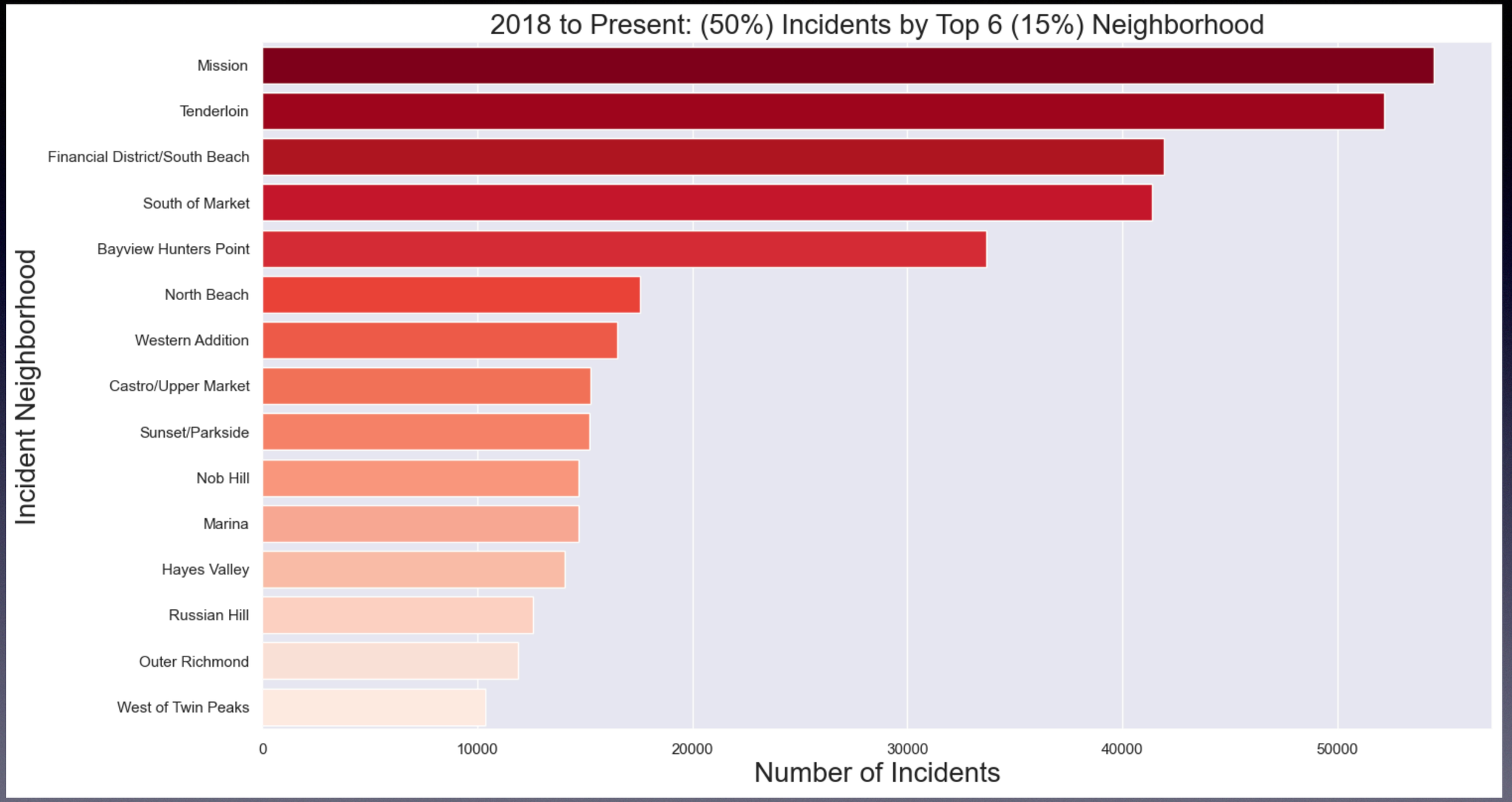
6367 Unique Intersections

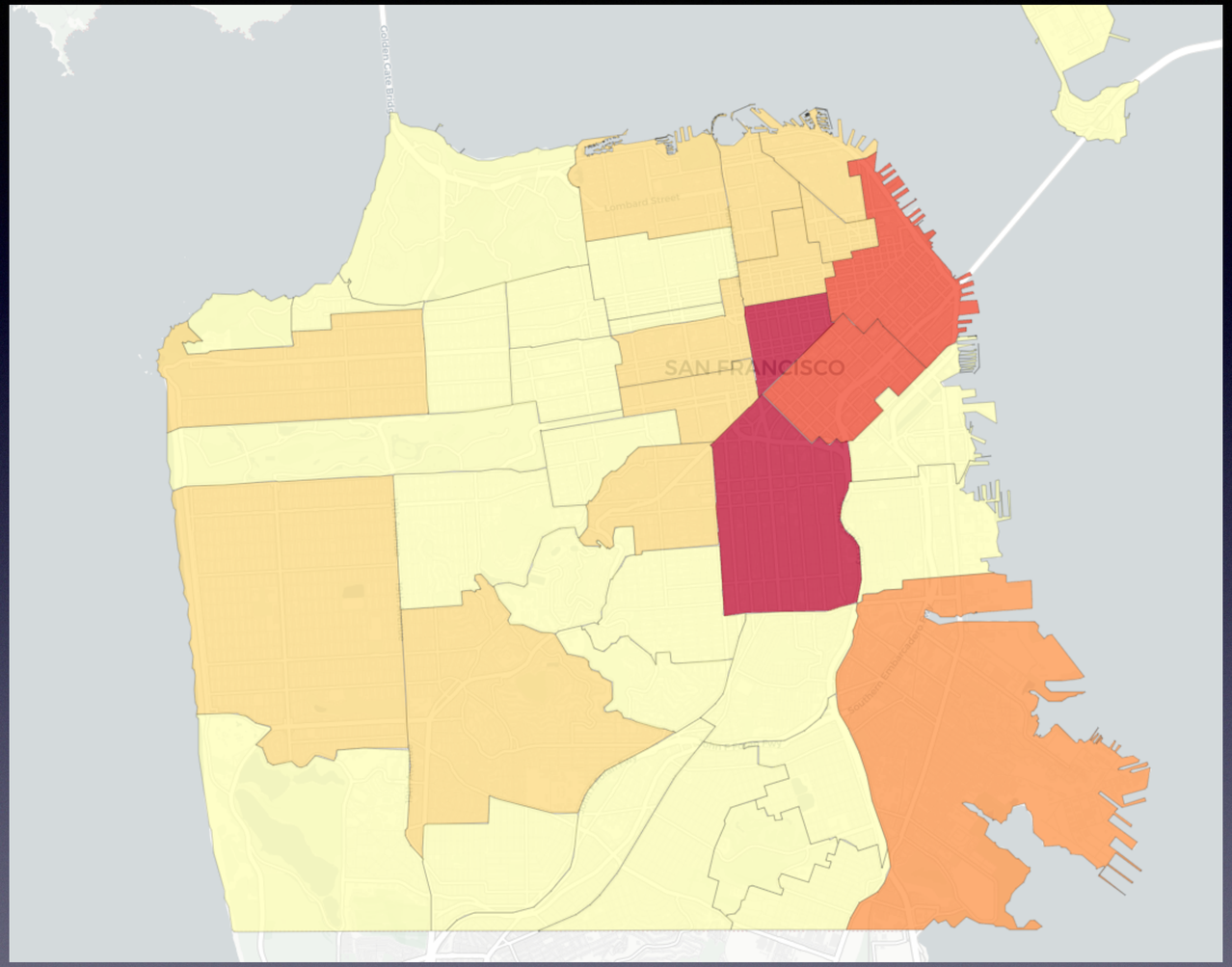


40 Neighborhoods



40 Neighborhoods

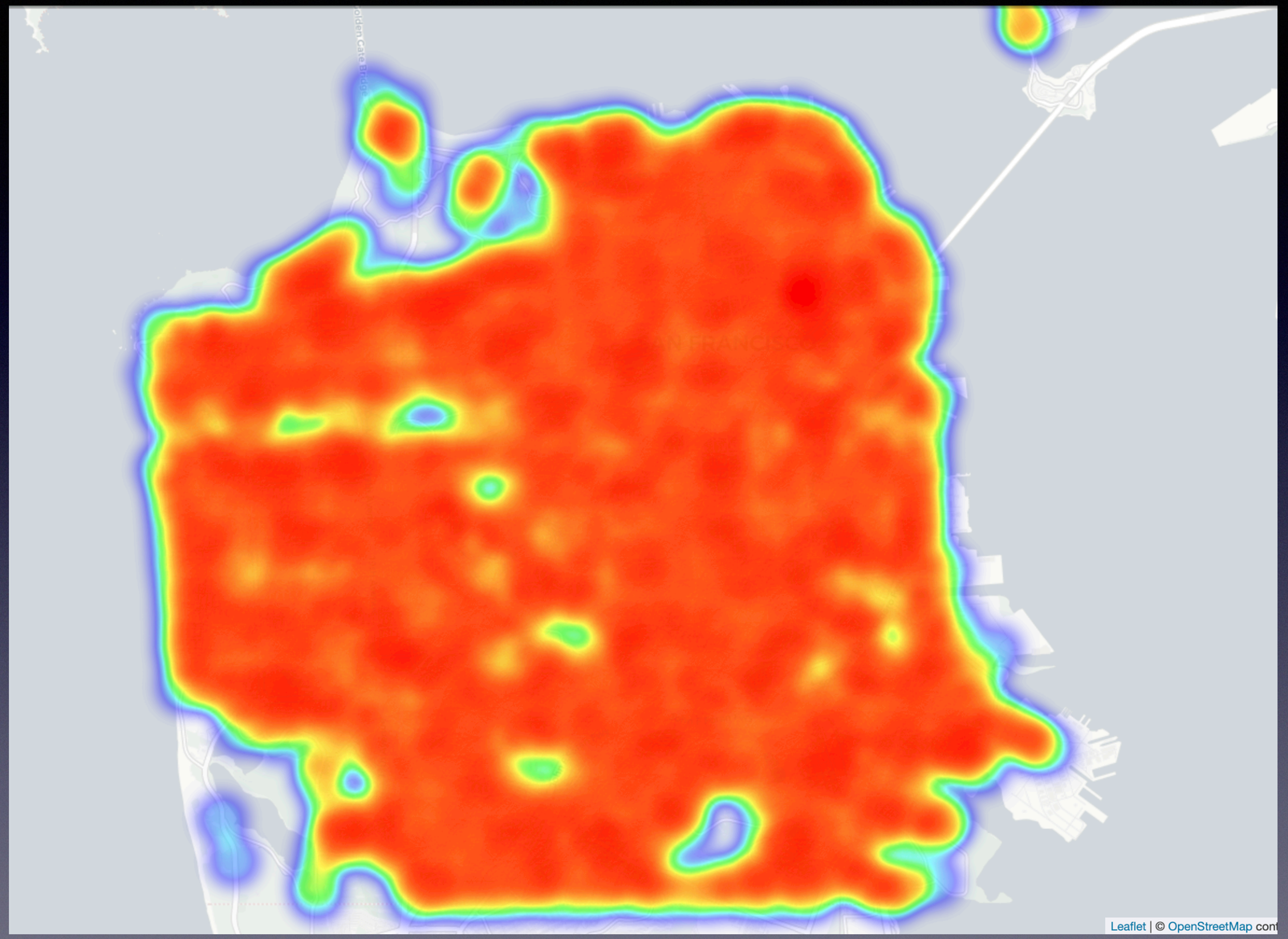

2018 to Present: (50%) Incidents by Top 6 (15%) Neighborhood
Incident Neighborhood
Mission
Tenderloin
Financial District/South Beach
South of Market
Bayview Hunters Point
North Beach
Western Addition
Castro/Upper Market
Sunset/Parkside
Nob Hill
Marina
Hayes Valley
Russian Hill
Outer Richmond
West of Twin Peaks
0
10000
20000
30000
40000
50000
Number of Incidents



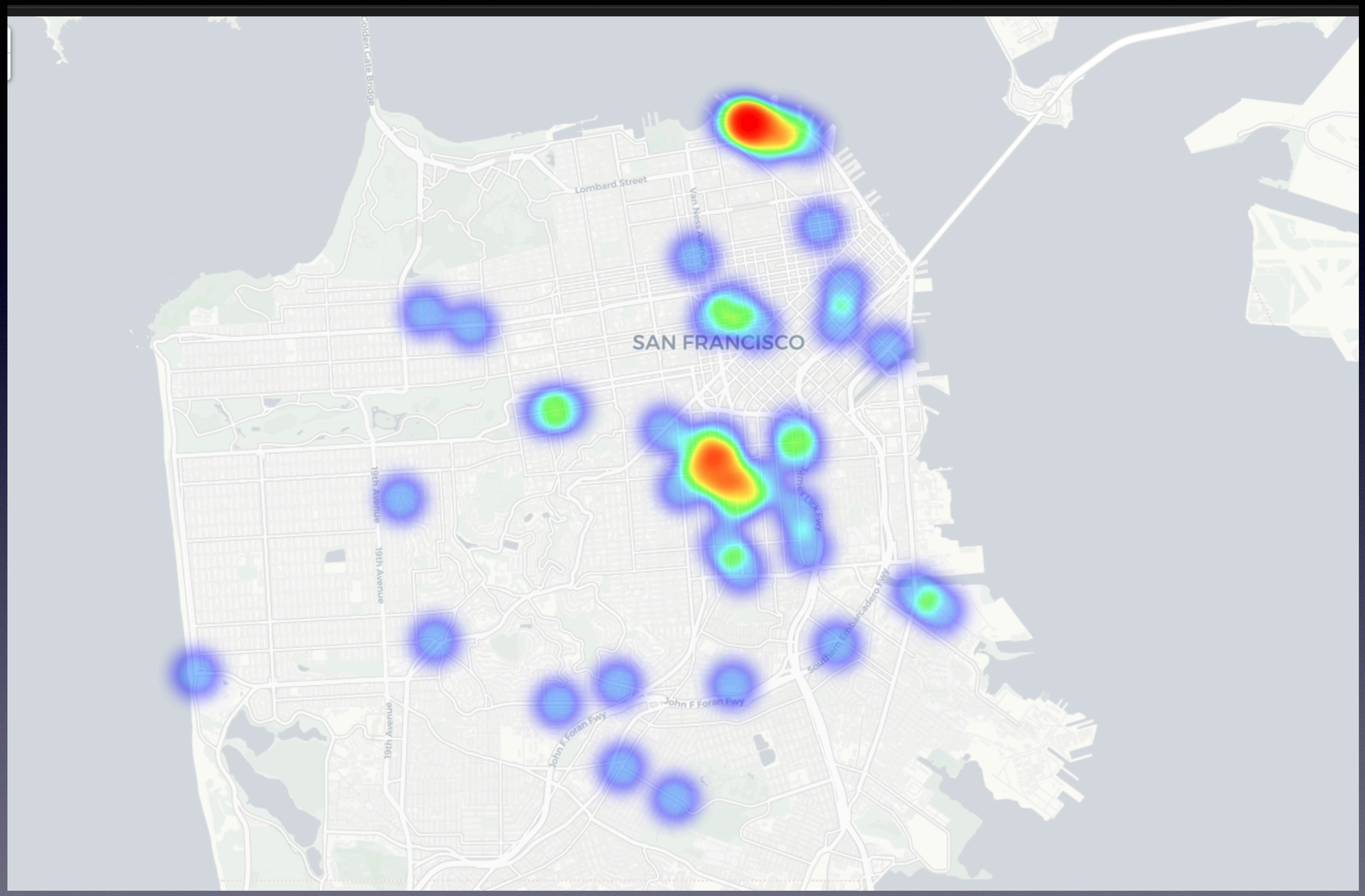
Heatmap

Heatmap of total incidents by neighborhood highlighting Tenderloin and Mission then South of Market and Financial District as highest occurrences of crime

Heatmap by Time?



Crime Category: Larceny Theft



Crime Category: Liquor Laws

Demographic Data

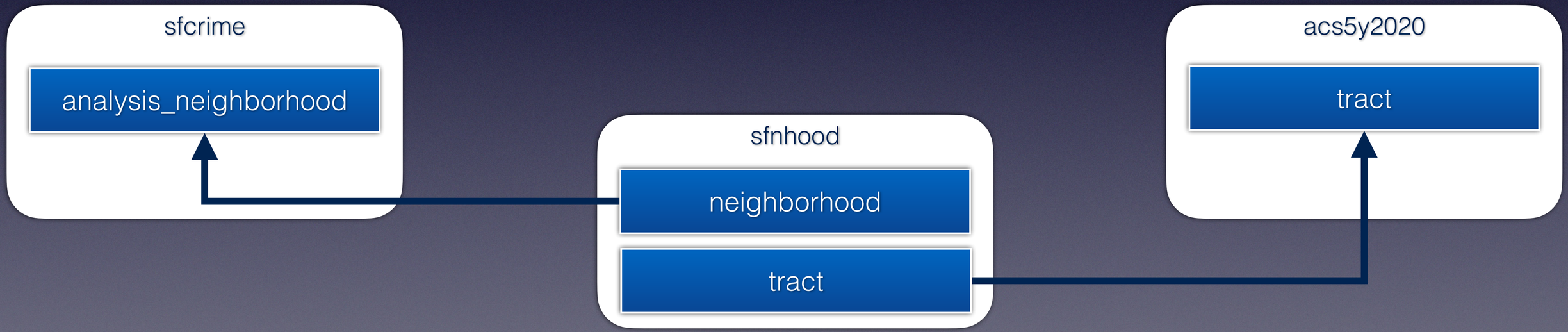
- Social
  - Educational Attainment (CHCI)
- Housing
  - Median Number of Rooms
  - Rental Units
  - Rental Size
  - Median Home Price
- Economic
  - Median Household Income
  - Unemployment
  - Poverty
- Demographic
  - Population
  - Race
  - Age

Any others? Weather, Holidays, Events, Disaster, etc.

40 Neighborhoods with Crime Data

242 Census Tracts with Demographic Data

Crime / Demographic Data



Insights into crime?

Dependent Variable

Independent Variables

Observations

Train / Test

Machine Learning Models

OLS Regression Results

Dep. Variable:	crime_rate	R-squared:	0.963
Model:	OLS	**Adj. R-squared:**	**0.917**
Method:	Least Squares	F-statistic:	21.06
Date:	Wed, 23 Nov 2022	Prob (F-statistic):	9.40e-09
Time:	13:18:31	Log-Likelihood:	-123.76
No. Observations:	41	AIC:	293.5
Df Residuals:	18	BIC:	332.9
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-127.3935	125.039	-1.019	0.322	-390.091	135.304
chci	**0.6648**	**0.258**	**2.574**	**0.019**	**0.122**	**1.207**
a85a	-0.9544	1.453	-0.657	0.520	-4.007	2.099
a7584	0.8287	0.870	0.952	0.354	-1.000	2.657
a6574	-0.1390	1.402	-0.099	0.922	-3.083	2.806
a5564	-0.6101	1.041	-0.586	0.565	-2.797	1.577
a3554	**1.9880**	**1.061**	**1.873**	**0.077**	**-0.242**	**4.218**
a2034	0.6673	0.599	1.115	0.280	-0.591	1.925
mage	0.6508	1.440	0.452	0.657	-2.375	3.677
pop	**-0.0075**	**0.003**	**-2.483**	**0.023**	**-0.014**	**-0.001**
pop_white	-0.1523	0.973	-0.157	0.877	-2.196	1.892
pop_black	-0.1645	1.075	-0.153	0.880	-2.423	2.094
pop_asian	0.1583	0.987	0.160	0.874	-1.916	2.233
pop_latino	-0.1159	1.001	-0.116	0.909	-2.220	1.988
h_med_rooms	0.2592	3.041	0.085	0.933	-6.130	6.649
h_rental_units	-0.0240	0.153	-0.157	0.877	-0.346	0.298
h_rentel_size	**13.7870**	**10.735**	**1.284**	**0.215**	**-8.767**	**36.341**
h_med_rent	-0.0017	0.008	-0.230	0.821	-0.018	0.014
h_med_homeprice	-1.187e-06	7.76e-06	-0.153	0.880	-1.75e-05	1.51e-05
eco_clf	-0.4060	0.718	-0.565	0.579	-1.915	1.103
eco_unemp	-0.2932	1.173	-0.250	0.805	-2.757	2.170
eco_med_hincome	**-0.0003**	**0.000**	**-2.209**	**0.040**	**-0.001**	**-1.42e-05**
eco_poverty	-0.6266	0.536	-1.169	0.258	-1.753	0.500

Omnibus:	4.803	Durbin-Watson:	1.993
Prob(Omnibus):	0.091	Jarque-Bera (JB):	3.558
Skew:	0.673	Prob(JB):	0.169
Kurtosis:	3.521	Cond. No.	1.33e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.33e+08. This might indicate that there are
strong multicollinearity or other numerical problems.

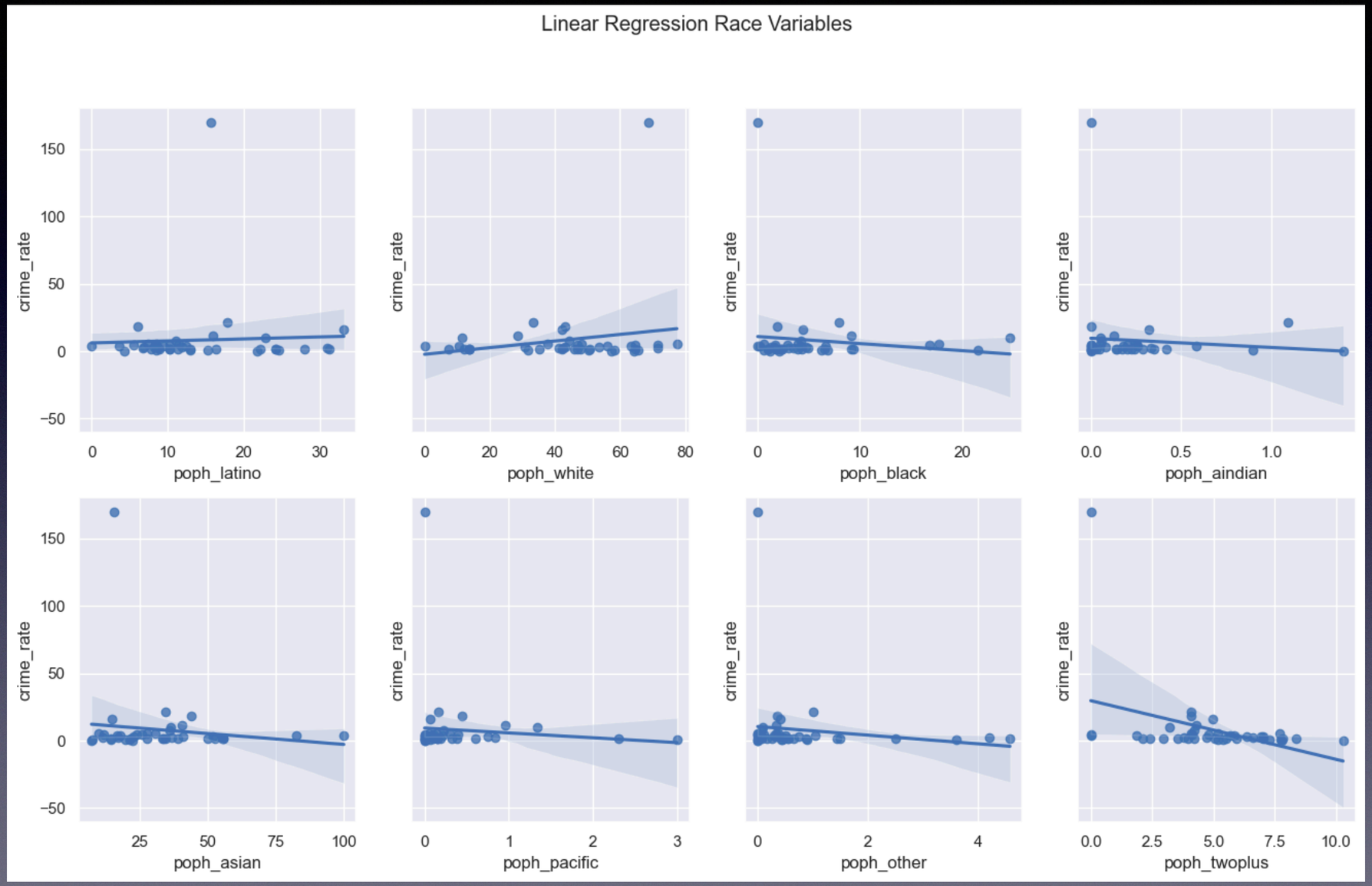
Adjusted R Squared: **91%**

Positive Correlations

- CHCI
- Age (35 - 54)
- Rental Household Size

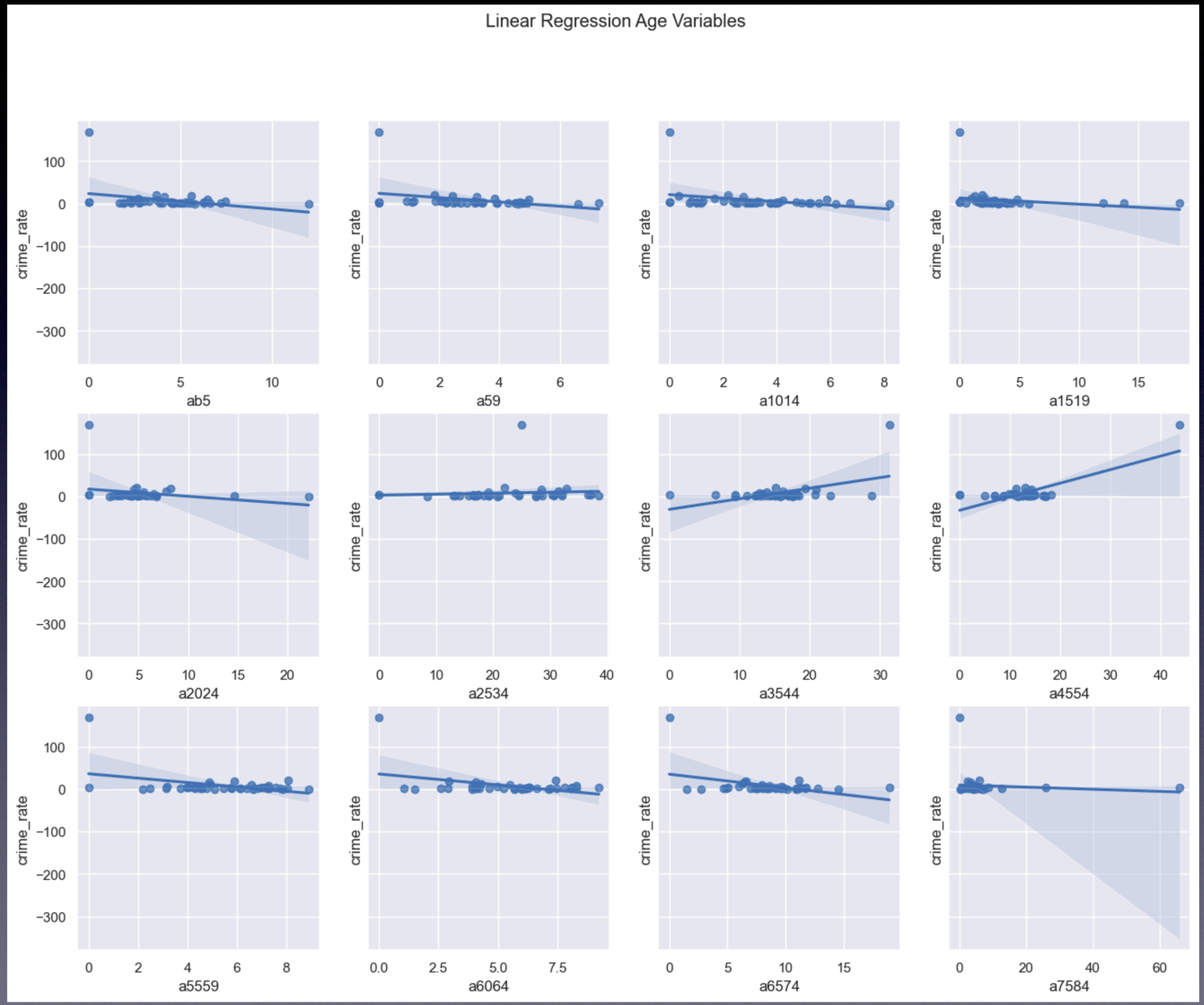
Negative Correlations

- Population
- Median House Price



Outliers?

A very high Adjusted R-Squared seemed suspicious.



Outliers?

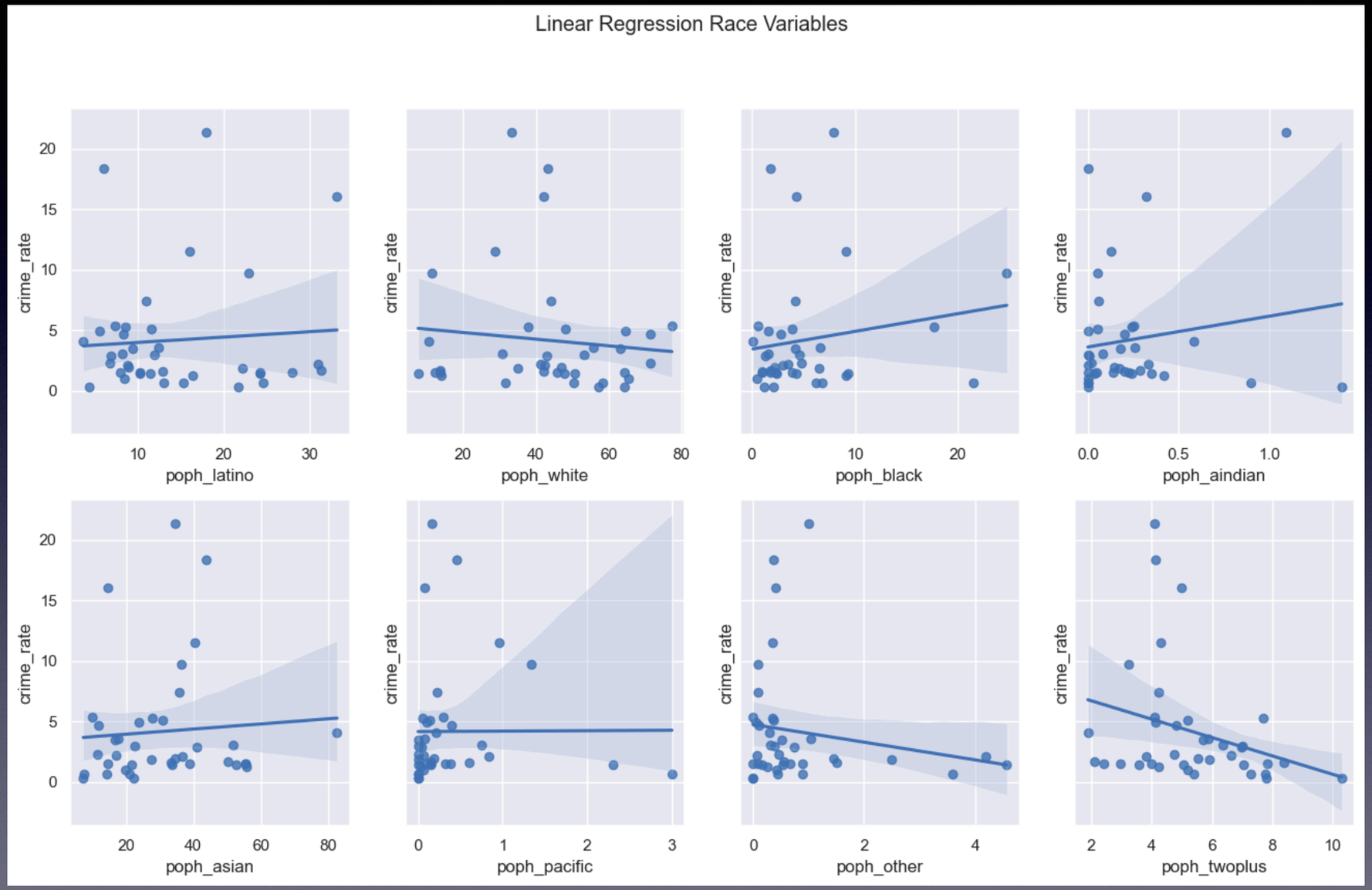
Notice the anomalies?

Parks have very little ACS population though a significant number of crime incidents, creating outliers

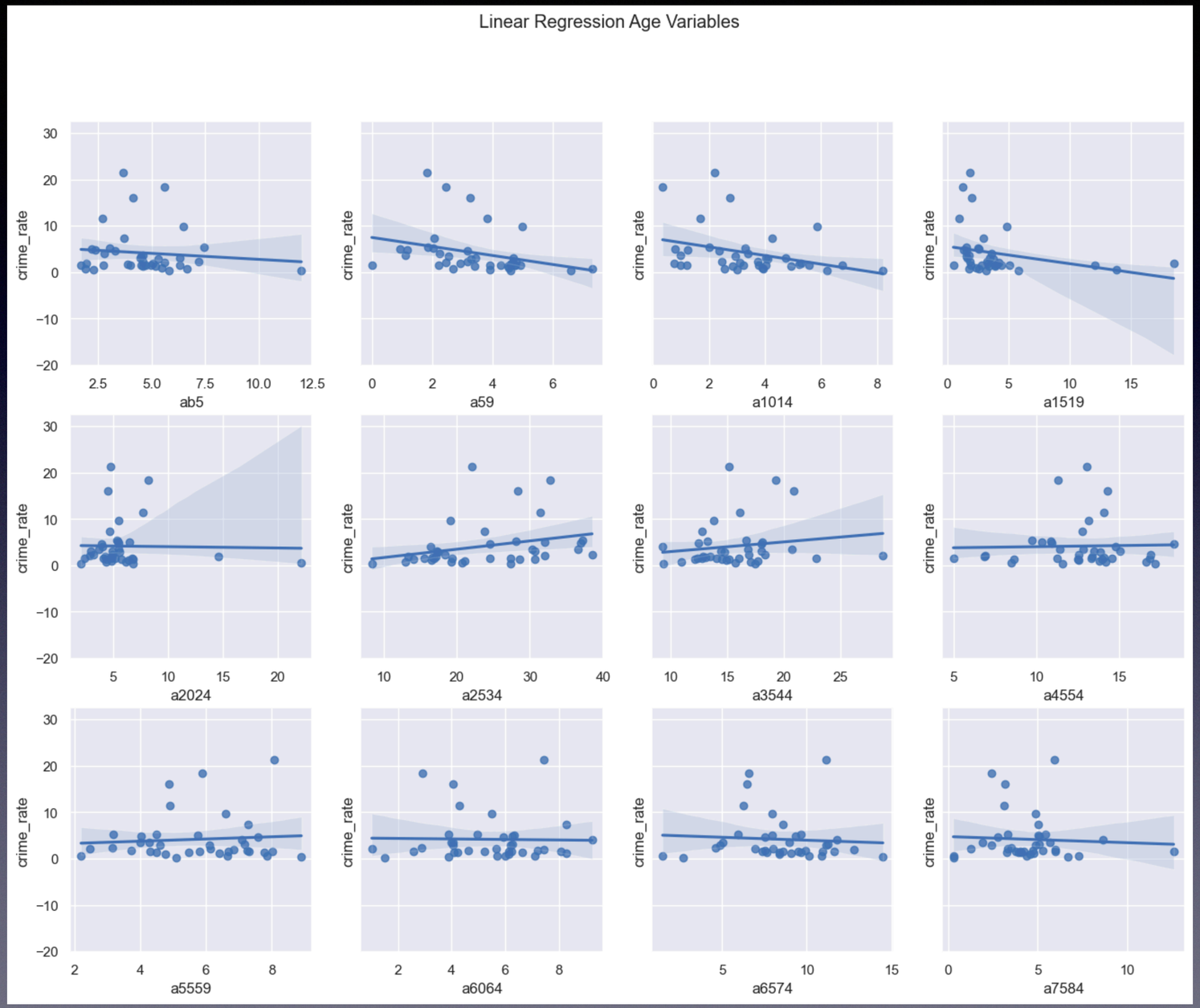
neighborhood	Population	incidents	crime_rate
Golden Gate Park	32.000000	5353	167.281250
Tenderloin	2477.166667	52194	21.070040
Financial District/South Beach	2296.300000	41928	18.258938
Mission	3436.705882	54504	15.859373
South of Market	3590.285714	41372	11.523317
Bayview Hunters Point	3498.181818	33660	9.622141
North Beach	2386.800000	17545	7.350846
Marina	2798.444444	14715	5.258278
Western Addition	3185.571429	16501	5.179918
Nob Hill	2916.333333	14719	5.047091
Russian Hill	2605.285714	12594	4.834019
Castro/Upper Market	3305.428571	15242	4.611202
Lincoln Park	185.000000	808	4.367568
Chinatown	2385.000000	9603	4.026415
McLaren Park	153.000000	607	3.967320
Hayes Valley	3963.600000	14057	3.546523
Pacific Heights	2661.444444	9138	3.433474
Sunset/Parkside	5102.437500	15225	2.983868
Potrero Hill	2969.000000	8802	2.964635
Outer Richmond	4158.636364	11903	2.862236
Haight Ashbury	3836.200000	8494	2.214170
Bernal Heights	4358.166667	9568	2.195419
Mission Bay	4443.333333	9258	2.083571
West of Twin Peaks	5497.857143	10367	1.885644
Lakeshore	3592.000000	6530	1.817929
Excelsior	5122.500000	8631	1.684919
Inner Richmond	4550.600000	6936	1.524195
Japantown	3624.000000	5492	1.515453
Noe Valley	3771.333333	5670	1.503447
Outer Mission	4929.200000	7365	1.494157
Portola	4060.750000	5877	1.447270
Lone Mountain/USF	4337.500000	6170	1.422478
Visitacion Valley	3975.000000	5624	1.414843
Inner Sunset	4758.500000	6539	1.374173
Oceanview/Merced/Ingleside	4555.833333	5554	1.219096
Presidio Heights	3481.666667	3277	0.941216
Twin Peaks	4034.000000	2674	0.662866
Glen Park	4327.000000	2854	0.659579
Treasure Island	3184.000000	1901	0.597048
Seacliff	2416.000000	763	0.315811
Presidio	4073.000000	1070	0.262706

Remove Observations:

Neighborhood: Golden Gate Park
Population: 32
Incidents: 5353
Crime Rate: **167**



Better representation with positive correlation for all races except White, Other Race, and Two Plus Races



Better distribution with negative correlation for younger ages (<20 years) and strong positive correlation for young adults (<45)

OLS Regression Results

Dep. Variable:	crime_rate	R-squared:	0.809
Model:	OLS	Adj. R-squared:	**0.529**
Method:	Least Squares	F-statistic:	2.888
Date:	Wed, 23 Nov 2022	Prob (F-statistic):	0.0194
Time:	13:20:11	Log-Likelihood:	-82.556
No. Observations:	38	AIC:	211.1
Df Residuals:	15	BIC:	248.8
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-109.4247	57.781	-1.894	0.078	-232.582	13.733
chci	0.2230	0.129	1.727	0.105	-0.052	0.498
a85a	-1.0017	0.772	-1.298	0.214	-2.646	0.643
a7584	-0.0195	1.076	-0.018	0.986	-2.314	2.275
a6574	1.1228	0.768	1.462	0.164	-0.514	2.760
a5564	0.9737	0.635	1.533	0.146	-0.380	2.328
a3554	0.5748	0.536	1.073	0.300	-0.567	1.717
a2034	**0.8145**	**0.296**	**2.752**	**0.015**	**0.184**	**1.445**
mage	1.0738	0.672	1.597	0.131	-0.359	2.507
pop	-0.0022	0.002	-1.380	0.188	-0.006	0.001
pop_white	-0.5874	0.523	-1.123	0.279	-1.703	0.528
pop_black	0.2051	0.505	0.407	0.690	-0.870	1.280
pop_asian	-0.4314	0.500	-0.863	0.402	-1.497	0.635
pop_latino	-0.0632	0.487	-0.130	0.898	-1.102	0.975
h_med_rooms	**-3.1406**	**1.551**	**-2.025**	**0.061**	**-6.446**	**0.164**
h_rental_units	0.1953	0.111	1.764	0.098	-0.041	0.431
h_rentel_size	9.7535	4.980	1.959	0.069	-0.861	20.368
h_med_rent	**0.0081**	**0.004**	**2.194**	**0.044**	**0.000**	**0.016**
h_med_homeprice	**1.292e-05**	**5.11e-06**	**2.531**	**0.023**	**2.04e-06**	**2.38e-05**
eco_clf	-0.4151	0.341	-1.217	0.242	-1.142	0.312
eco_unemp	-0.2503	0.555	-0.451	0.659	-1.433	0.933
eco_med_hincome	-1.234e-05	6.99e-05	-0.177	0.862	-0.000	0.000
eco_poverty	-0.0475	0.271	-0.175	0.863	-0.626	0.531

Omnibus:	1.461	Durbin-Watson:	1.923
Prob(Omnibus):	0.482	Jarque-Bera (JB):	1.160
Skew:	0.422	Prob(JB):	0.560
Kurtosis:	2.858	Cond. No.	1.36e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.36e+08. This might indicate that there are
strong multicollinearity or other numerical problems.

Adjusted R Squared: **52%**

Positive Correlations

- Age (20 - 34)
- Median House Price
- Rental Household Size

Negative Correlations

- Median Household Room

Next Steps

- Crime is concentrated in certain locations
- Crime is seasonal
- A time series forecast should be possible
- What other variables influence crime are there?
- With some more data wrangling, there are still a lot of insights to be gained with additional models