ANALYSIS OF 2003-2018 SAN FRANCISCO CRIME DATA

Wanran Li

I. SAN FRANCISCO HIGH IN CRIME

Overall Rank (1 = Safest)	City	Total Score	'Home & Community Safety' Rank	'Natural-Disaster Risk' Rank	'Financial Safety' Rank
145	San Francisco, CA	66.80	149	123	19

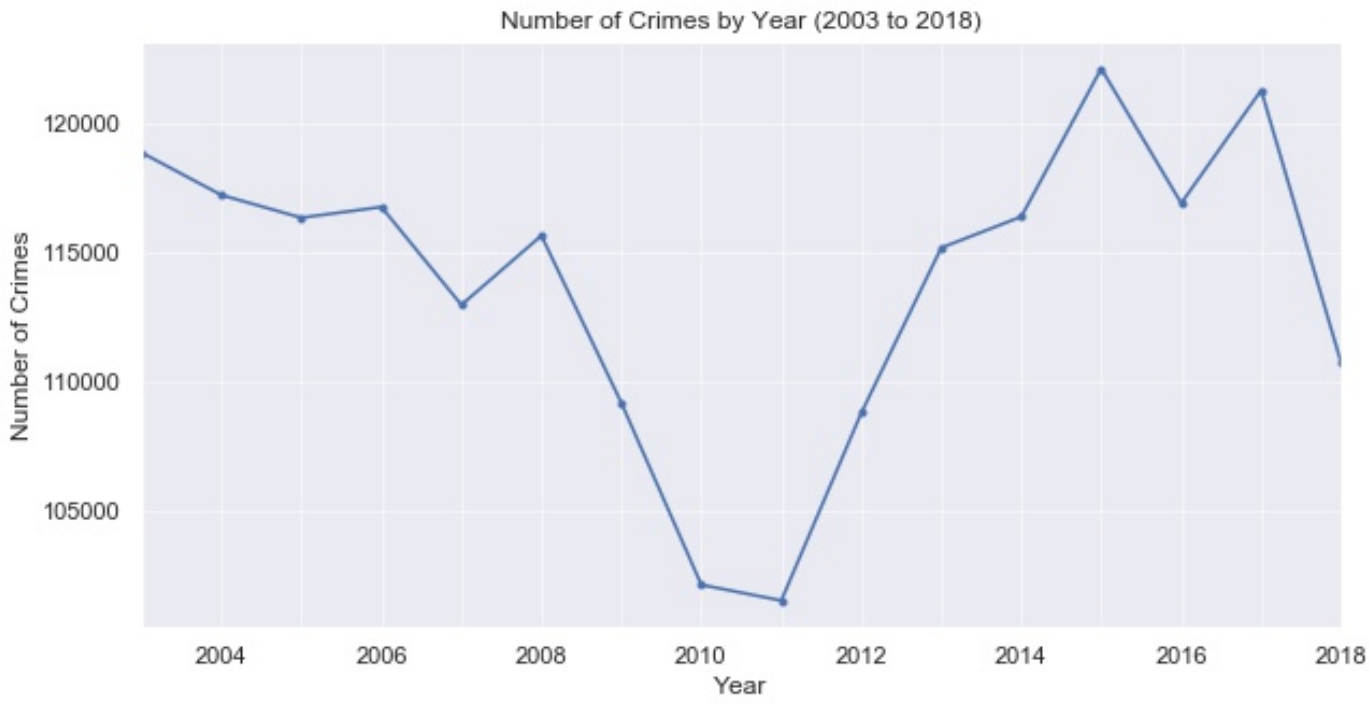
https://wallethub.com/edu/safest-cities-in-america/41926/#main-findings

II. DATA ACQUISITION & CLEANING

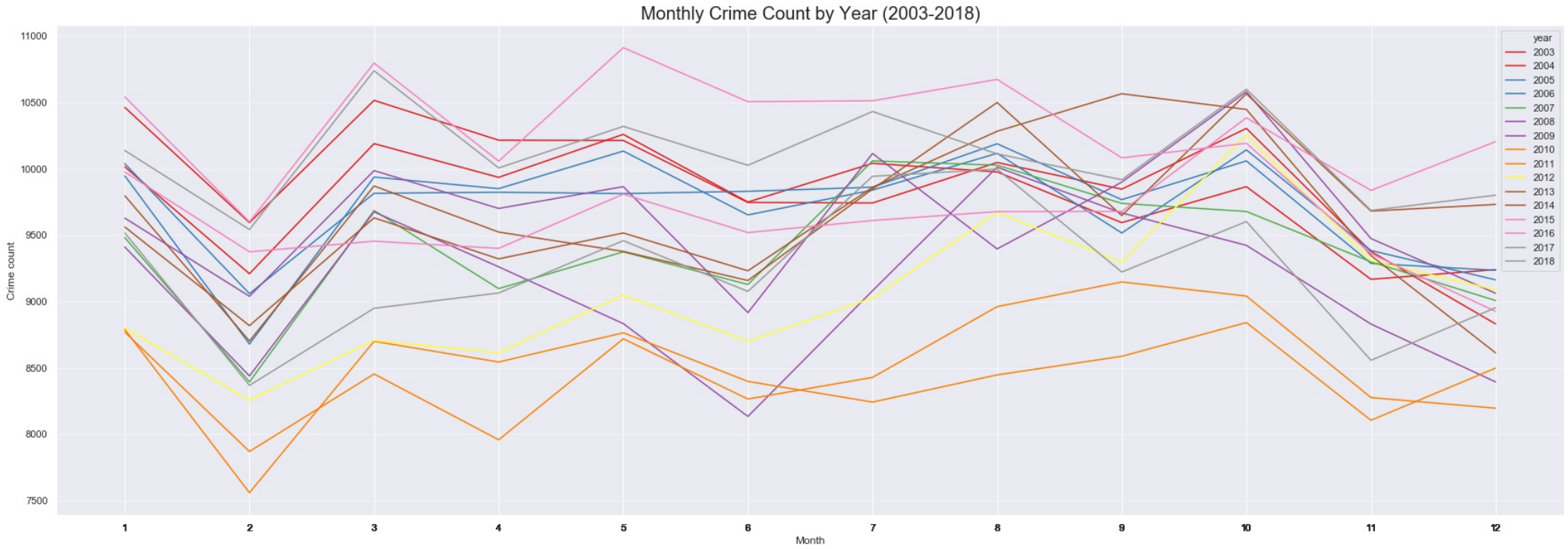
- Combine 2 major datasets: Police Department Incident Reports from 2003 to May 2018 and 2018 to Present
- Data shape (1822068, 17) after removing duplicates and columns with too many null value
- Other data sets:
  - SF Temperature (2014-2018)
  - Unemployment Rate (2003-2018)
  - Homeless Population (2005-2017)
  - SF median household income (2006-2018)

III. EDA
3.1 TIMING OF SF CRIME

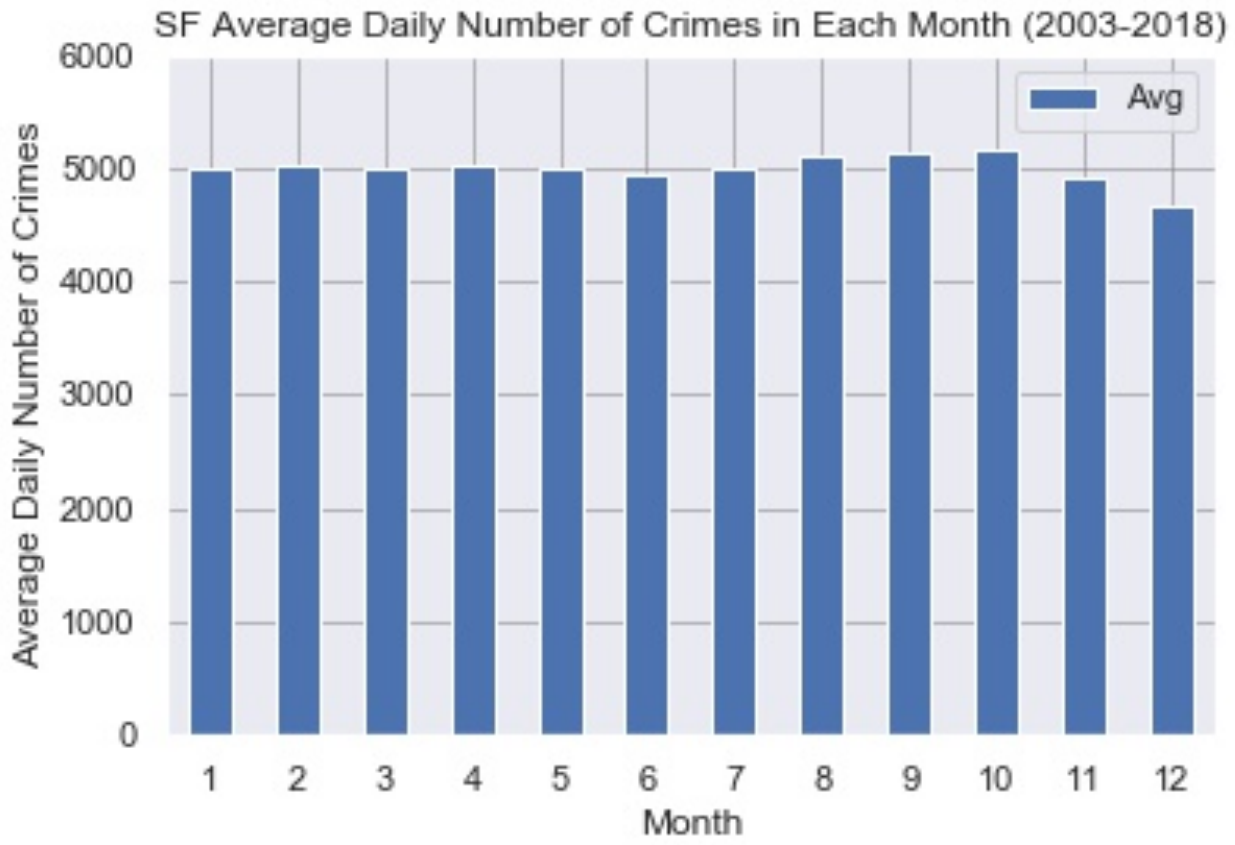
- Annual crime count: V shape



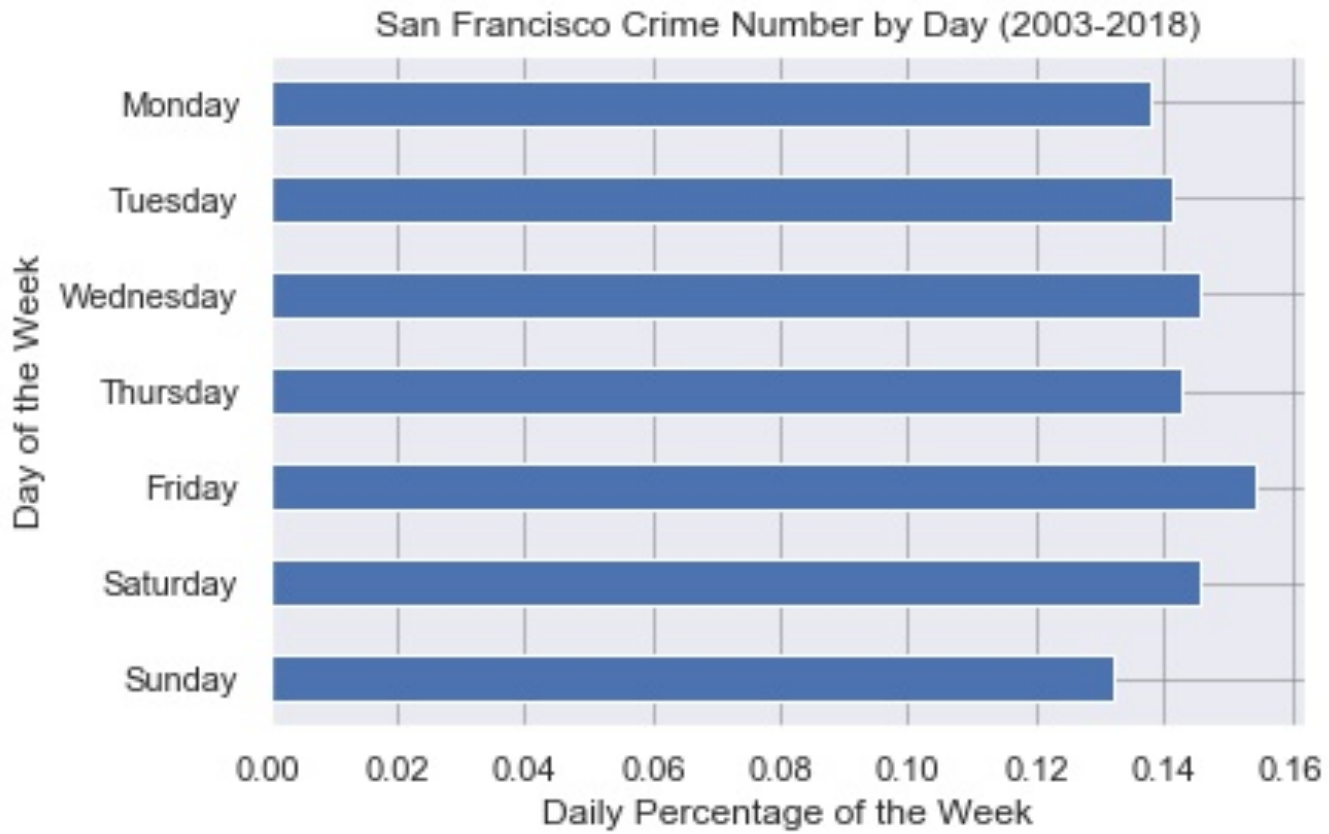
- Monthly trend by year: top 3 months are Oct, Aug & Mar



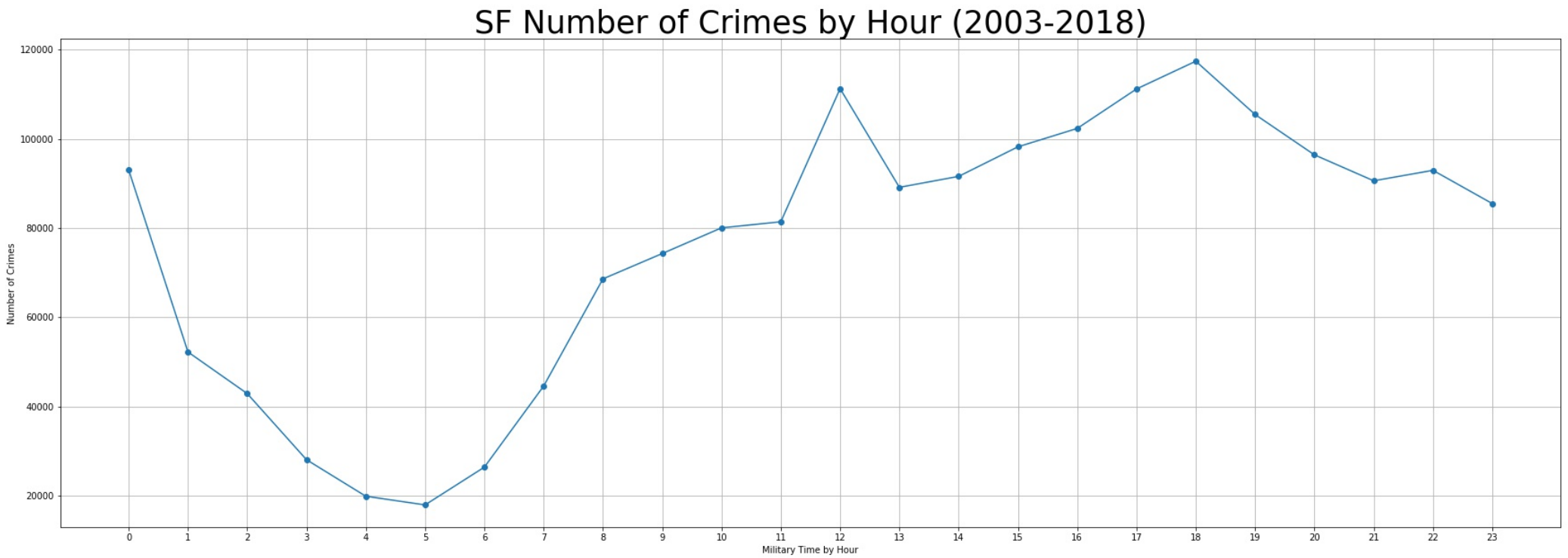
- Average daily # of crimes by month
  - Highest : Oct, Sept, Aug
  - Lowest: Dec Nov



- Crime count by day of the week
  - Highest: Fri, Sat, Wed
  - Lowest: Sun, Mon

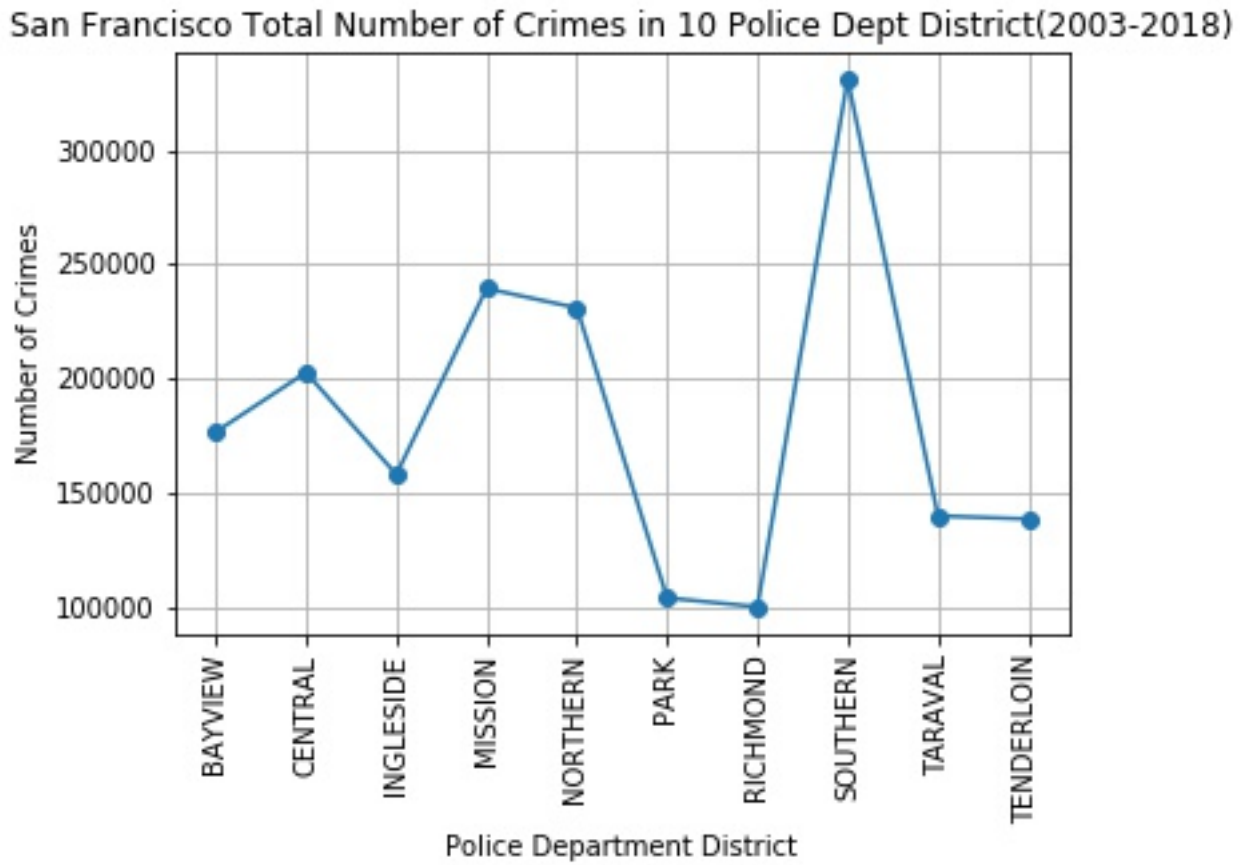


- Crime count by hour
  - Highest: daytime, peaks at 12pm, 18pm
  - Lowest: nighttime, hits bottom at 5am

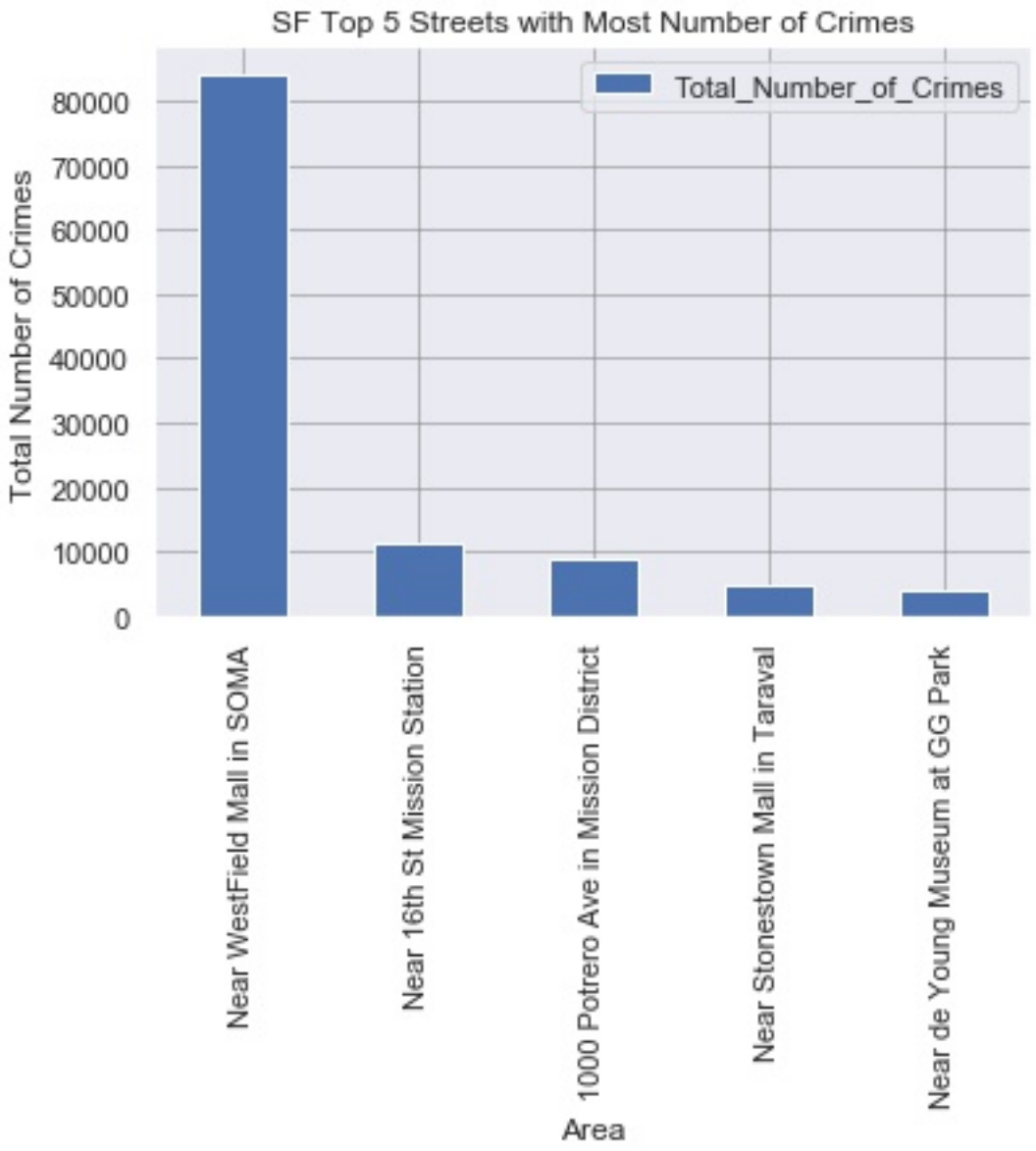


3.2 SF CRIME BY AREA

- Crime count by police districts
  - Tenderloin isn't as dangerous as we thought
  - Highest: Southern, Mission
  - Lowest: Park, Richmond, Tenderloin

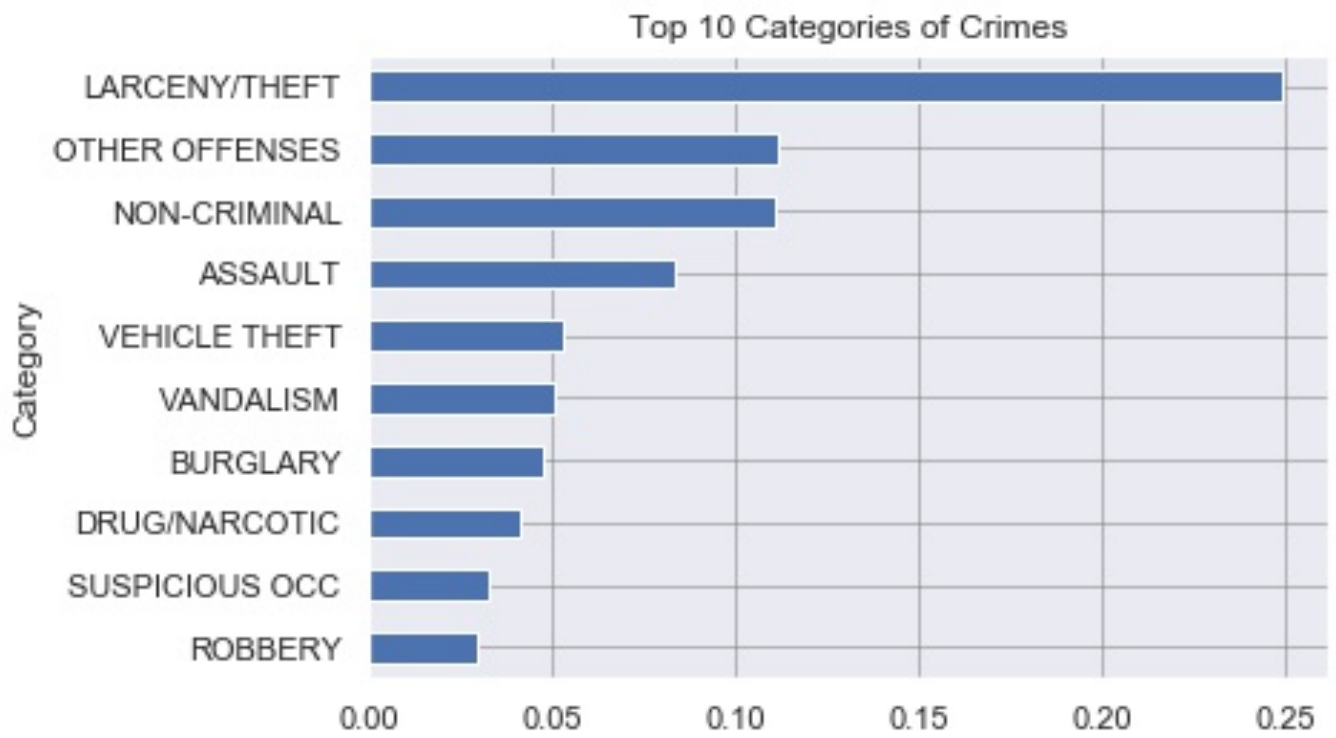


- Streets with most crimes:
  - Near Westfield Mall in SOMA
  - 16th St Mission Station
  - 1000 Potrero Ave in Mission District
  - Near Stonestown Mall in Taraval
  - Near de Young Museum at GG Park

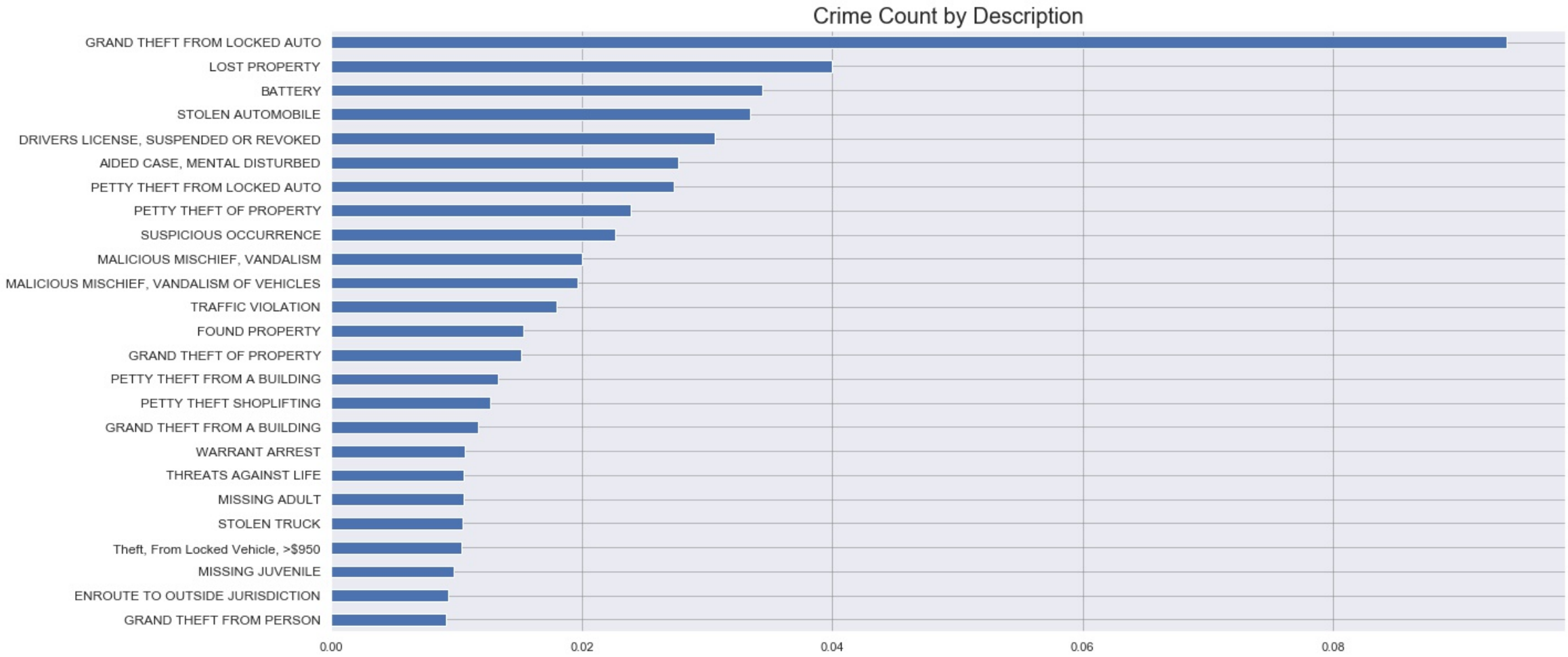


3.3 CATEGORIES AND DECRIPTION OF SF CRIME

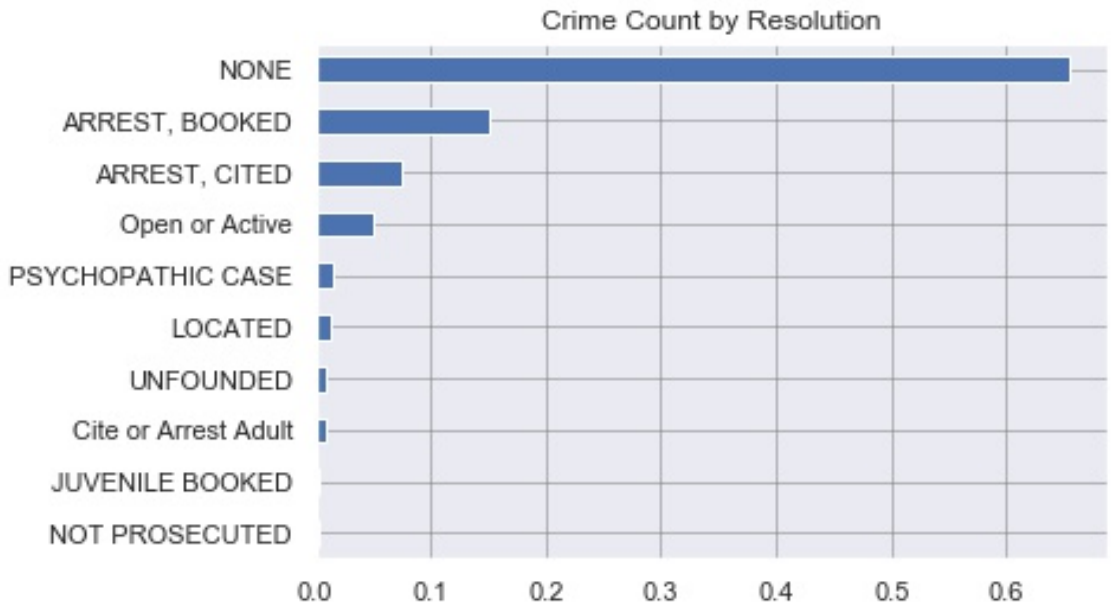
- Most common categories:
  - Larceny/theft
  - Other offenses: mostly traffic related
  - Non-criminal: assault, vandalism, burglary

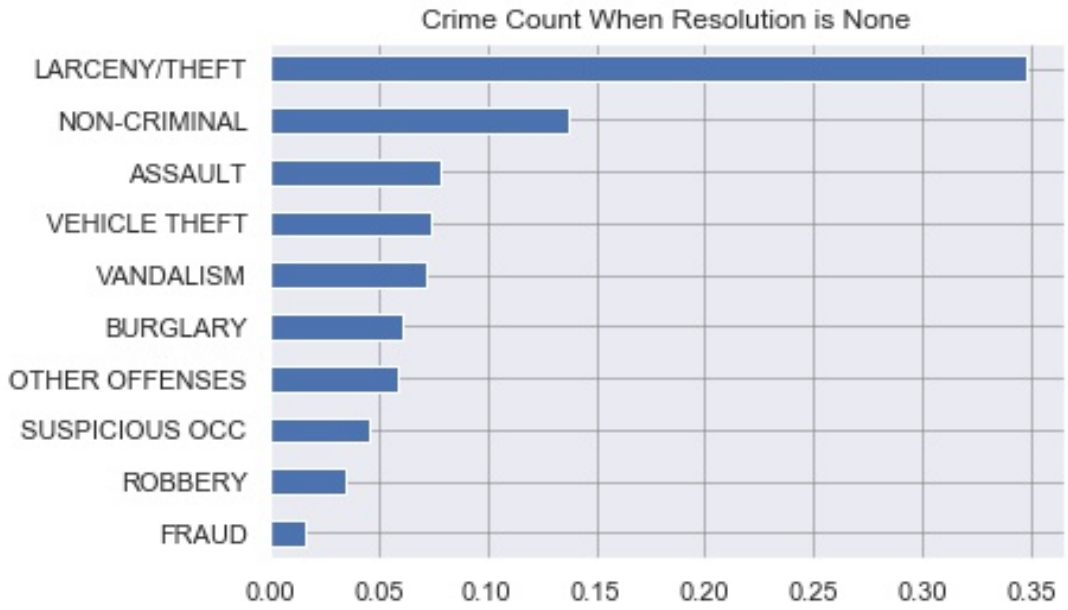


- Most common crimes by description:
  - Most are vehicle related
  - The top 25 types by description below accounts for 50% of crimes



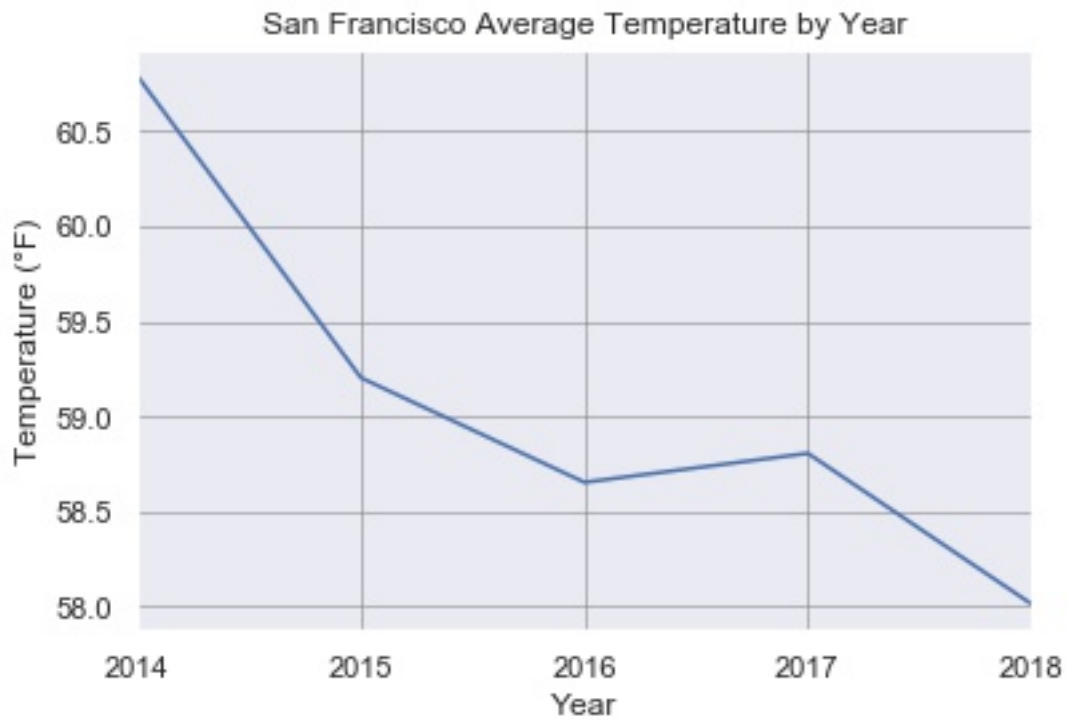
- Crime count by resolution
  - 65%: no actions taken
  - 15%: arrests, booked (police taking suspect's info, fingerprints and mug shots etc.)
  - 7%: arrests, cited (for misdemeanor, e.g. a traffic ticket)

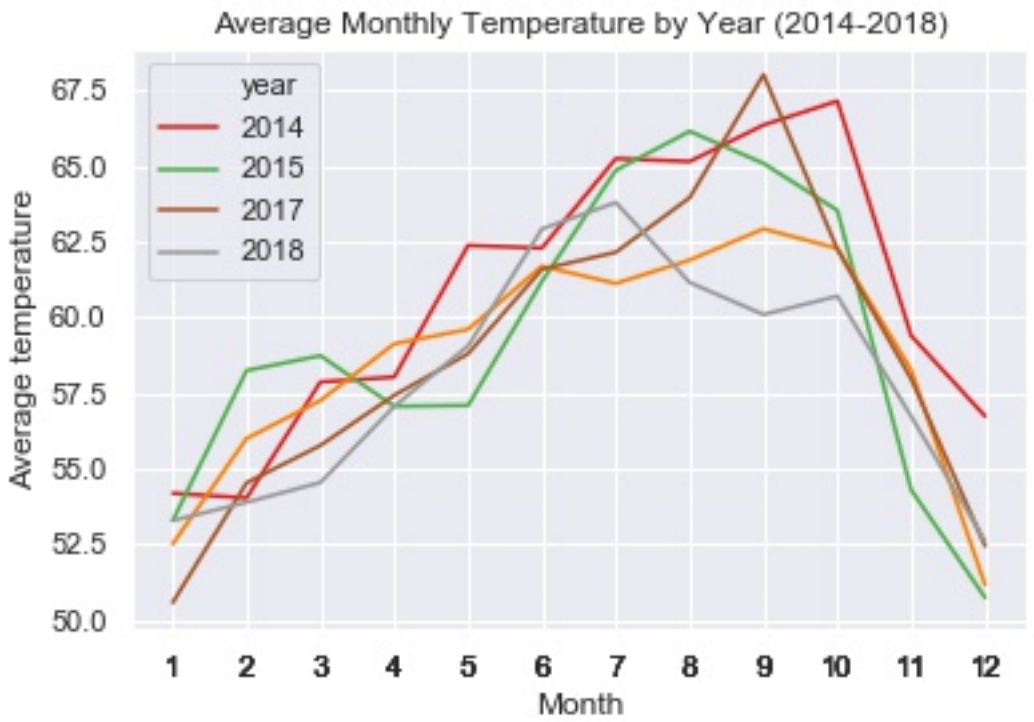




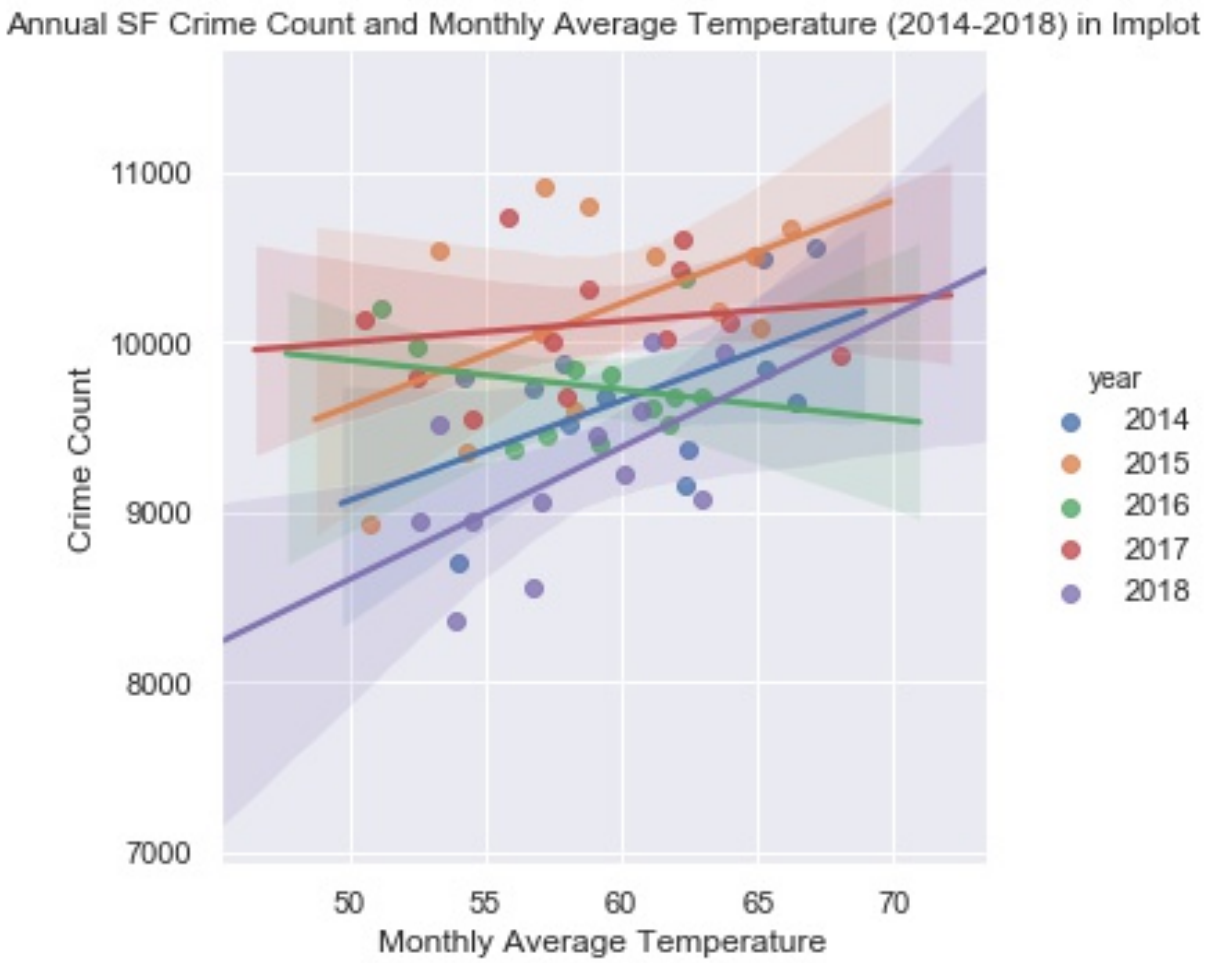
3.4 SF CRIME AND OTHER DATA SETS

- SF temperature (2014-2018)
  - Warmest: fall vs Coldest: winter

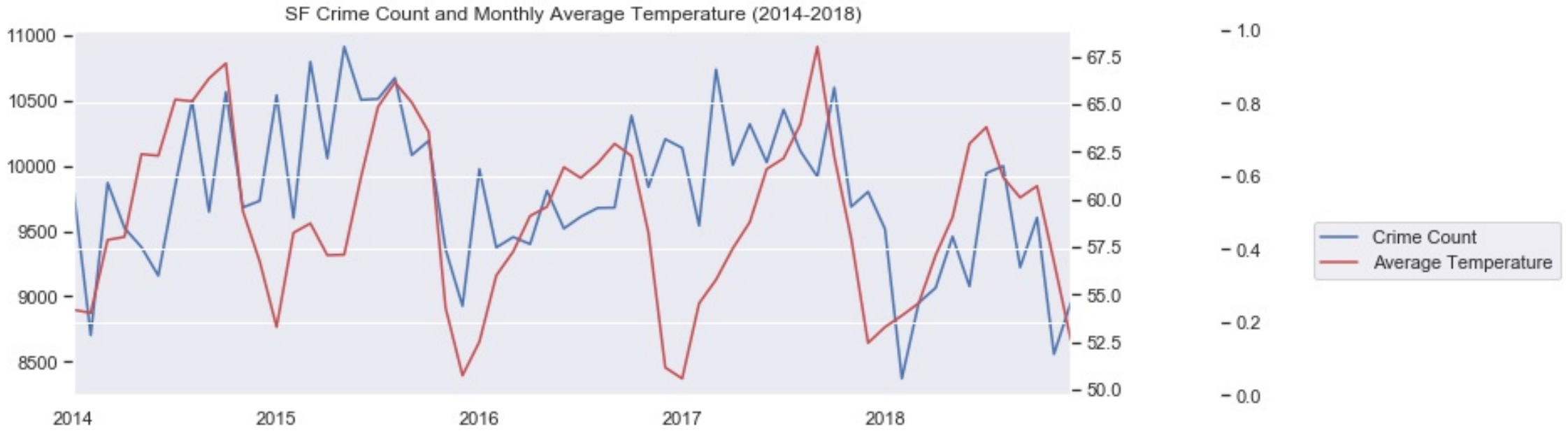




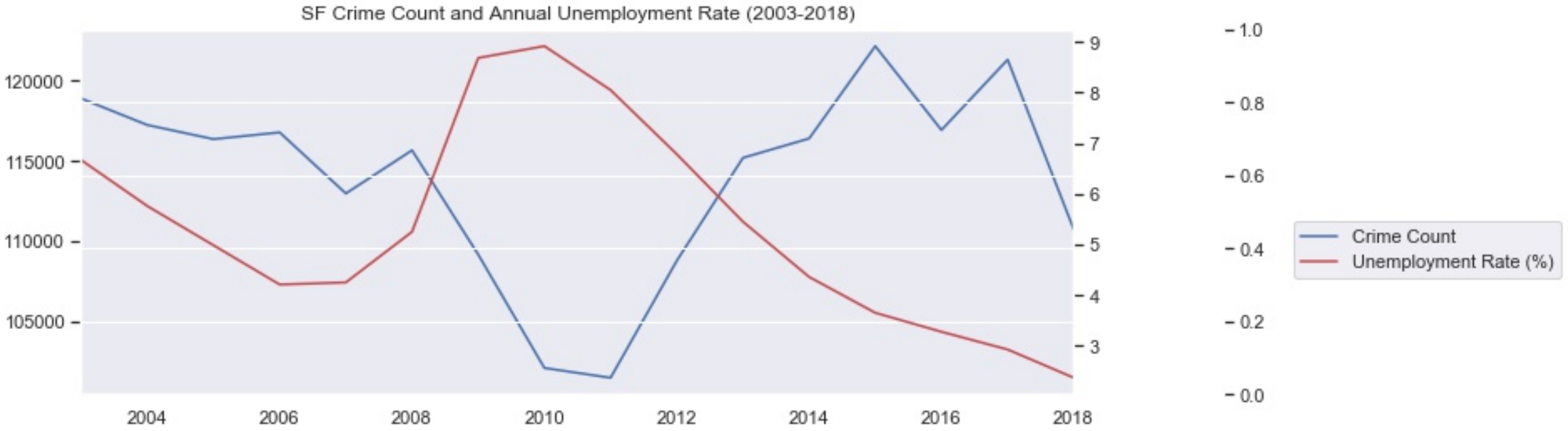
- SF Crime Count & Temperature by Year
  - Positively related in all years except for 2016



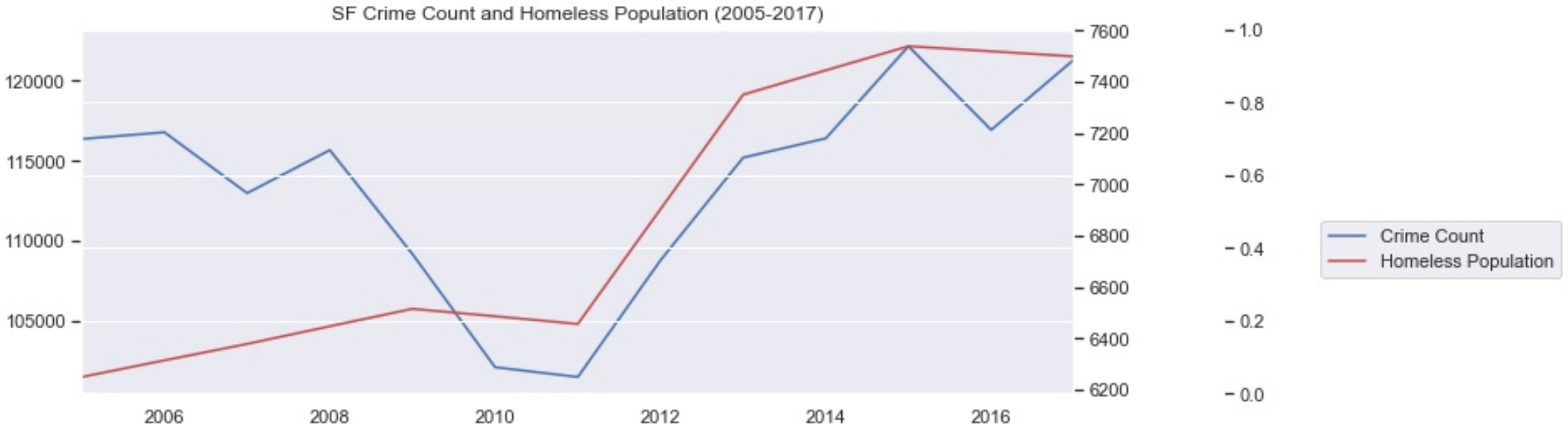
- SF Crime count & temperature by month
  - Lowest crime count in winter: Dec, Nov
  - Highest crime count in fall: Oct, Sept



- SF crime count and unemployment rate (2003-2018)
  - Unemployment rate has some impact, but not strongly correlated



- SF crime count and homeless population
  - Strong positive correlation, especially after 2009



MODELING: XGBOOST

- :To predict crime category based on time, location and resolution (multiclass classification problem)
- XGBoost:
  - Speed execution & better model performance
  - Objective: multi: softprob
  - 3-fold cross validation
  - Accuracy 52.8%

Analysis Neighborhoods	Category	DayOfWeek	PdDistrict	Resolution	SF Find Neighborhoods	X	Y	Day	Month	Year	Hour	Minute
8.0	10	7	1	1	32.0	-122.404795	37.784908	2	12	2018	0	45
36.0	0	6	4	3	19.0	-122.408036	37.786410	1	12	2018	20	30
20.0	14	5	2	3	53.0	-122.416549	37.766871	16	11	2018	1	34
34.0	14	7	1	3	32.0	-122.407015	37.777400	19	8	2018	23	0
0.0	14	1	5	3	0.0	0.000000	0.000000	31	12	2018	1	0

CLASSIFICAITON REPORT

- Precision (accuracy of positive predictions):
  - traffic, assault, sex related
- Recall
  - Theft, sex related, other
- F1-score
  - Theft, sex related, others

category	precision	recall	f1-score	support
aided_case	0.511406	0.197675	0.285136	12475.000000
alcohol_related	0.000000	0.000000	0.000000	15.000000
arson	0.000000	0.000000	0.000000	174.000000
assault	0.777778	0.005012	0.009960	4190.000000
drug_related	0.555556	0.191489	0.284810	1175.000000
fraud	0.549550	0.048355	0.088889	2523.000000
gambling	0.000000	0.000000	0.000000	3.000000
murder	0.000000	0.000000	0.000000	6.000000
others	0.501497	0.444893	0.471502	18074.000000
robbery	0.428270	0.039494	0.072319	5140.000000
sex_related	0.681648	0.491892	0.571429	370.000000
suicide	0.000000	0.000000	0.000000	28.000000
theft	0.537347	0.926723	0.680257	32029.000000
traffic_related	0.900000	0.007494	0.014864	1201.000000
vandalism	0.000000	0.000000	0.000000	96.000000
accuracy	0.528407	0.528407	0.528407	0.528407
macro avg	0.362870	0.156868	0.165278	77499.000000
weighted avg	0.535325	0.528407	0.452504	77499.000000

IV. LIMITATIONS

- Result can be improved if:
  - the model runs on gpu so that we can fit more years of historical data into the model,
  - or if we add more categorical features such as temperature and homeless data
- We could explore more about types of crimes based on location and time
- Recording categories in a more consistent manner. (e.g. "Stolen Property" and "Larceny/Theft" are the same)
- Our accuracy of studying the correlation between crime count and homeless population can be improved if the homeless count is performed yearly instead of every other year, so we don't need to use estimate for the years in between