California Highway Patrol Data Analysis

Andy Orfalea
aorfalea@gmail.com

01

Data Management

02

Insights & Visualization

03

Takeaways & Next Steps

Data Management

1. Merge Headers with Daily/Monthly CHP Data

2. Map DMA & Census Information and Merge

3. Data Cleaning - Check Nulls and Schema

4. EDA

5. Feature Engineering

6. Data Visualization

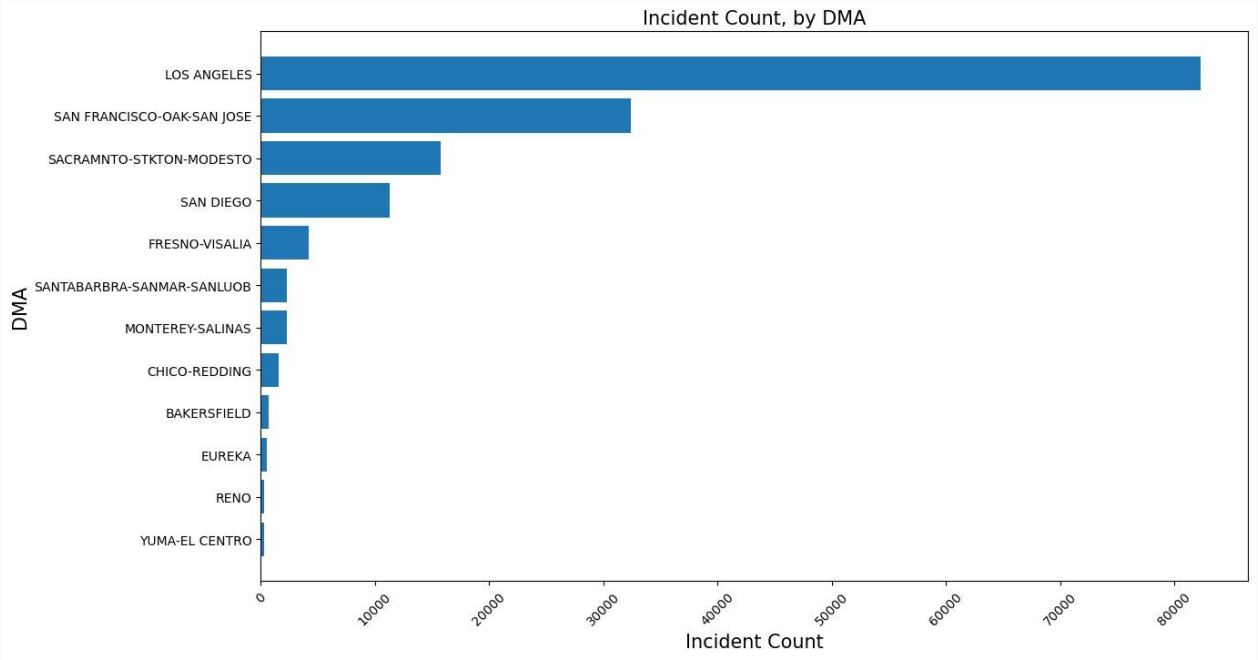
Data Insights and Visualization

Visualizing Where Incidents Occur

Before going to charts, start with a map:
https://public.tableau.com/views/CAHighwayIncidents/Sheet1?:language=en-US&:display_count=n&:origin=viz_share_link

Incident Count

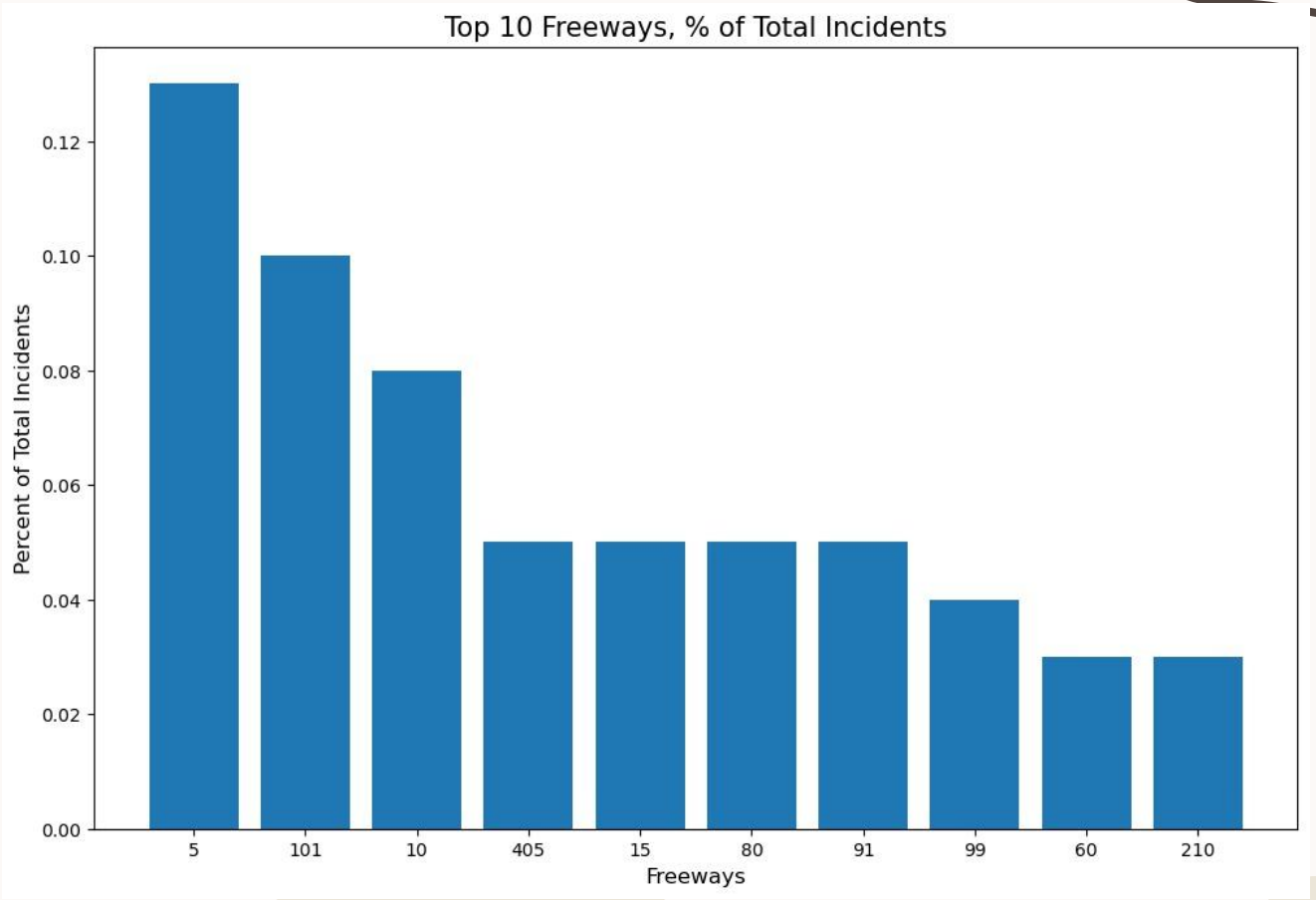
	% of Incidents
LOS ANGELES	0.534525
SAN FRANCISCO-OAK-SAN JOSE	0.210370
SACRAMNTO-STKTON-MODESTO	0.102299
SAN DIEGO	0.073614
FRESNO-VISALIA	0.027276
SANTABARBRA-SANMAR-SANLUOB	0.014875
MONTEREY-SALINAS	0.014849
CHICO-REDDING	0.010427
BAKERSFIELD	0.004460
EUREKA	0.003409
RENO	0.002091
YUMA-EL CENTRO	0.001805



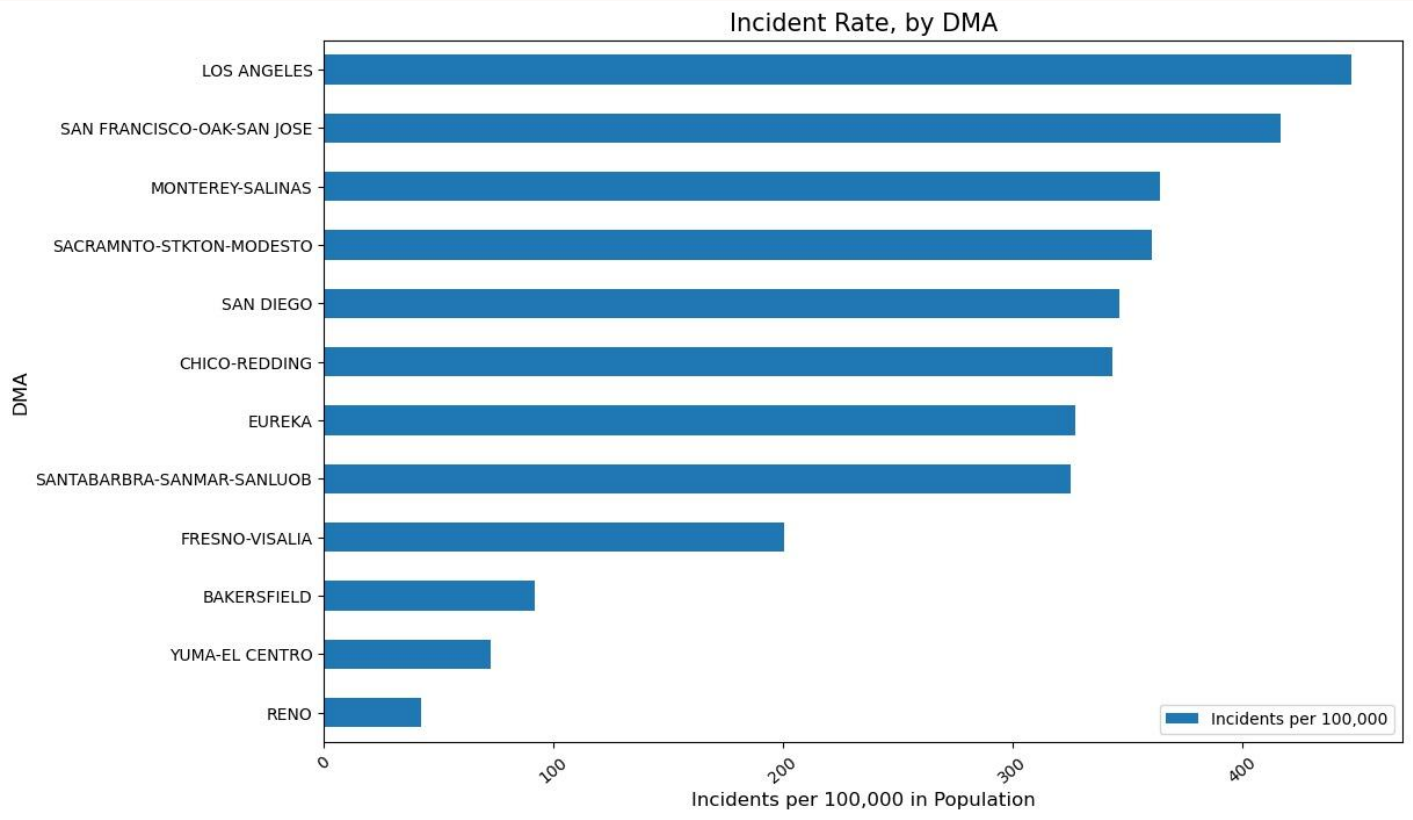
Area Specific

	Number of Incidents
LAFSP	8803
San Diego	7071
Central LA	6374
Santa Ana	5639
Riverside	5196
Orange County FSP	4687
Baldwin Park	4551
South LA	4416
San Jose	4297
East LA	3844

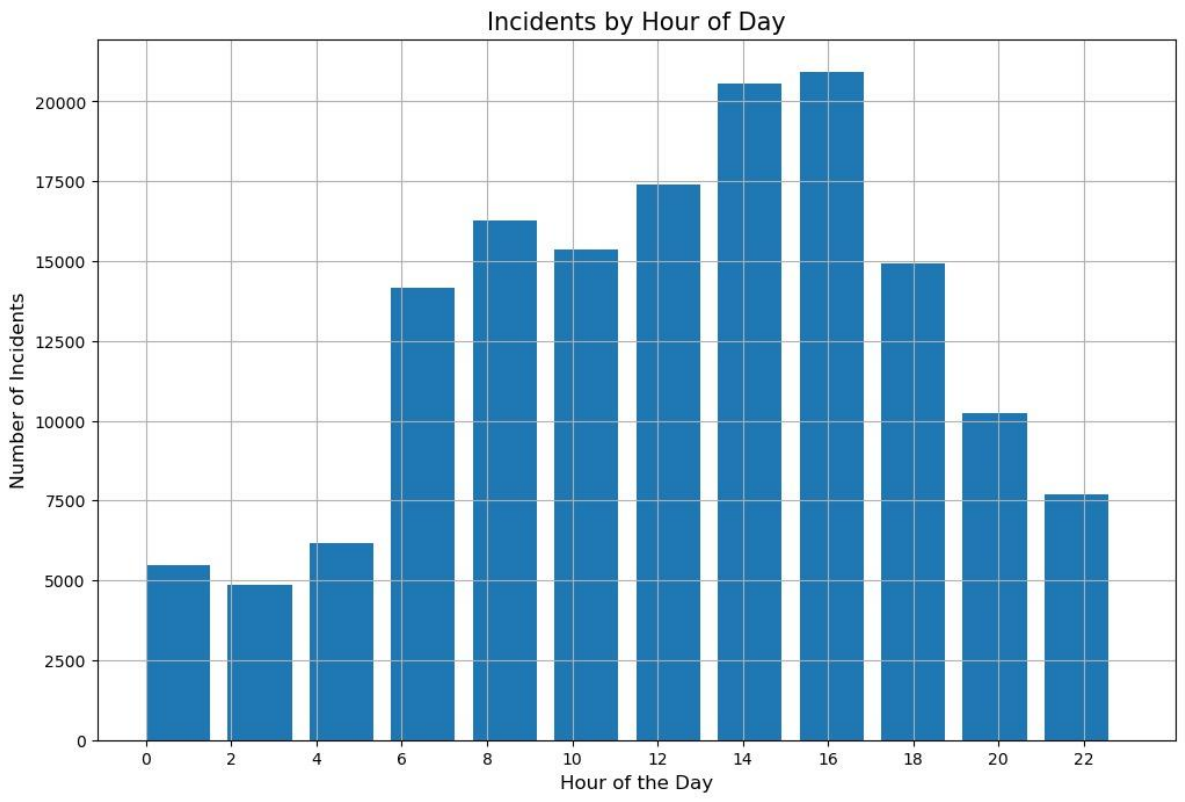

I-5
I-101
I-10


Top 10 Freeways, % of Total Incidents
Percent of Total Incidents
0.00
0.02
0.04
0.06
0.08
0.10
0.12
5
101
10
405
15
80
91
99
60
210
Freeways

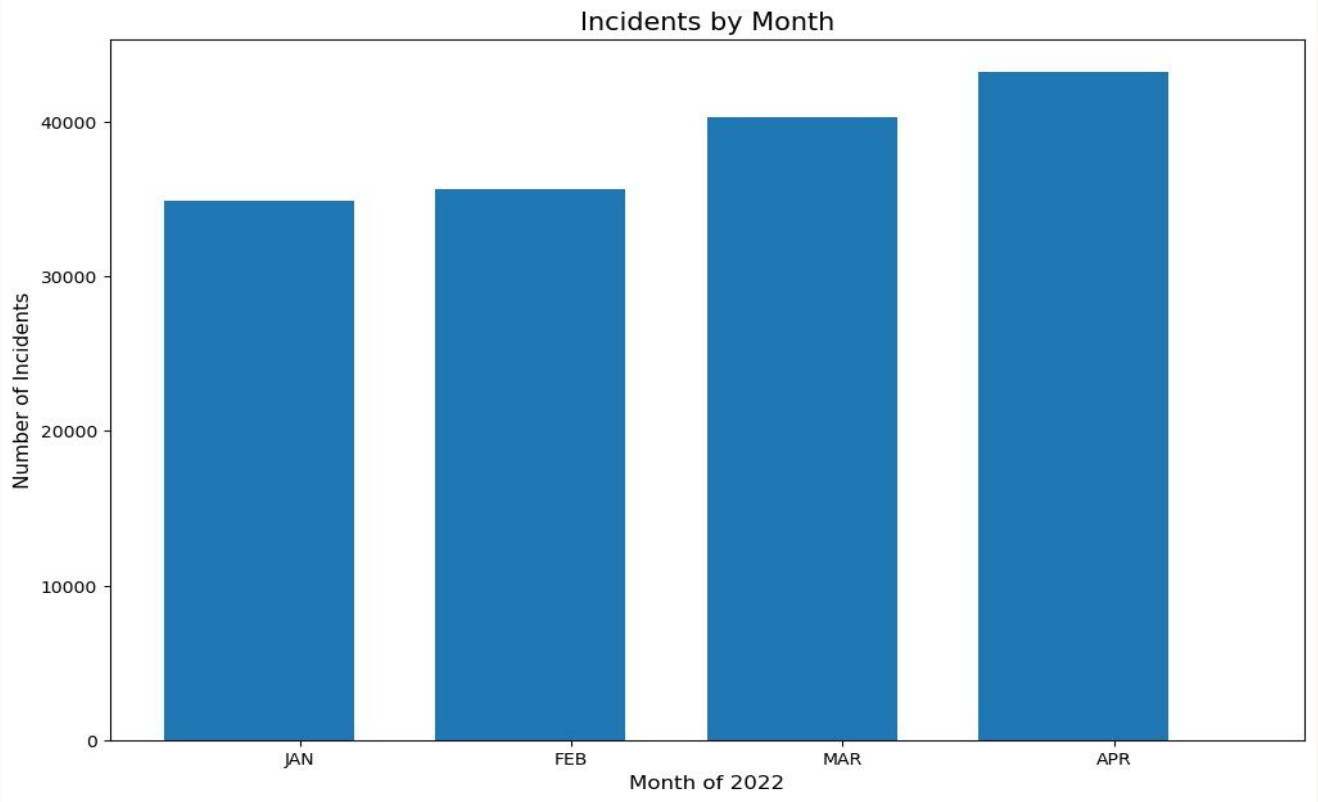
Normalizing For Population

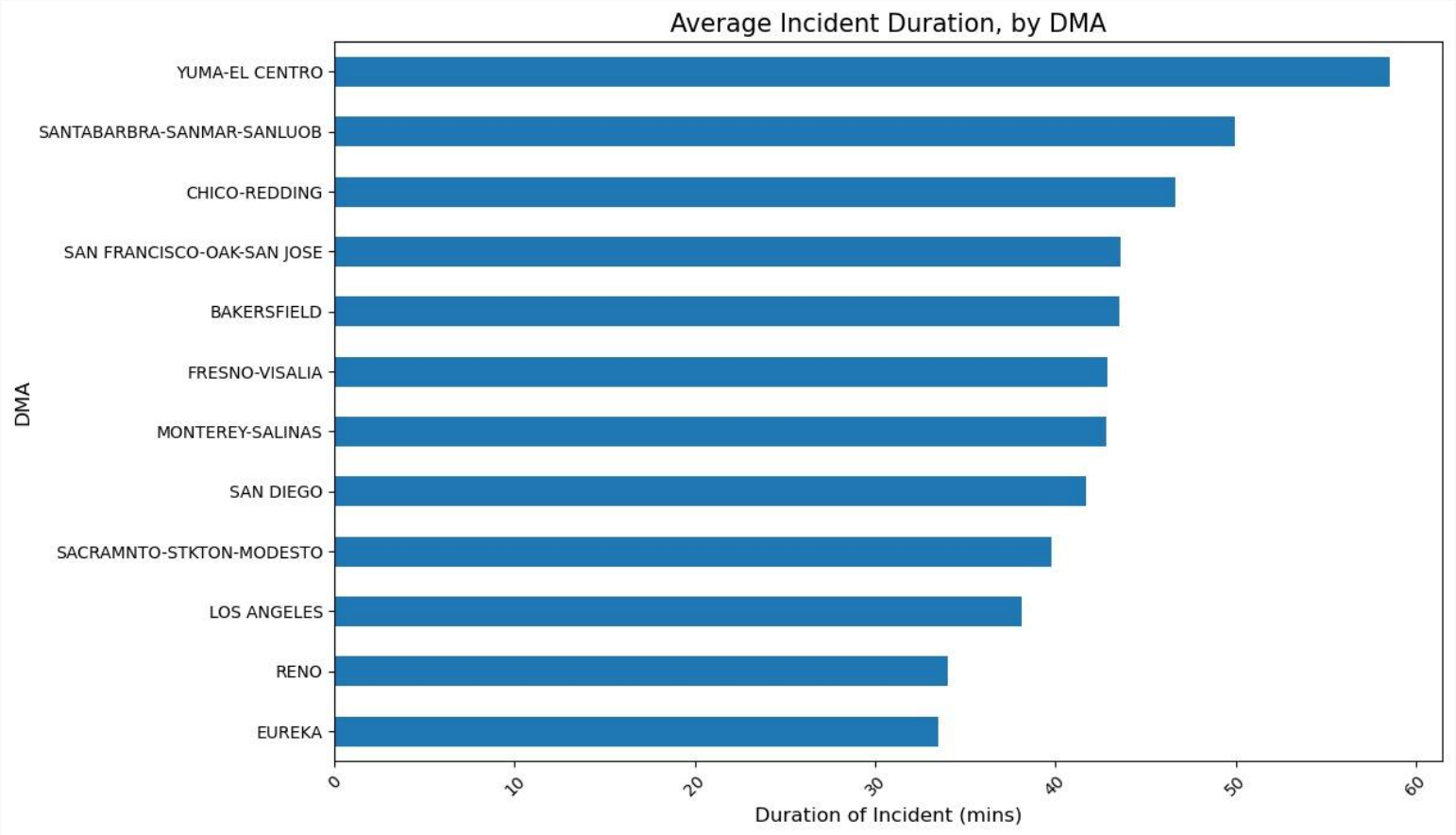


Hour of the Day

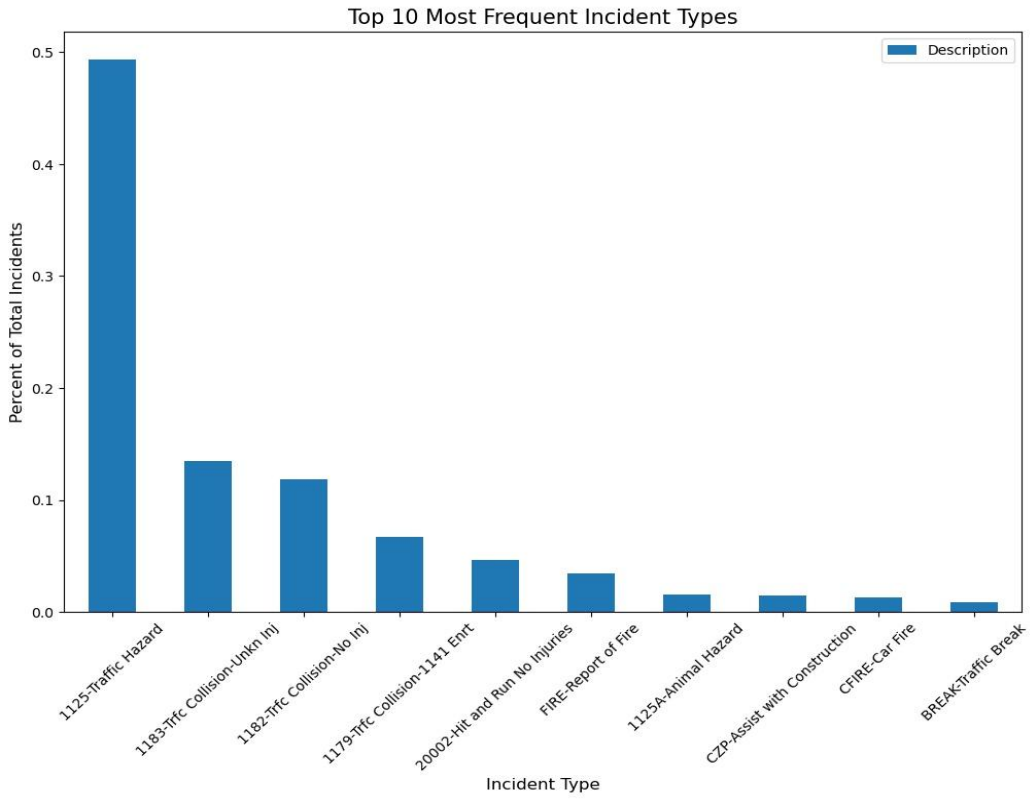


Seasonality




Average Incident Duration, by DMA
YUMA-EL CENTRO
SANTABARBRA-SANMAR-SANLUOB
CHICO-REDDING
SAN FRANCISCO-OAK-SAN JOSE
BAKERSFIELD
FRESNO-VISALIA
MONTEREY-SALINAS
SAN DIEGO
SACRAMNTO-STKTON-MODESTO
LOS ANGELES
RENO
EUREKA
DMA
0
10
20
30
40
50
60
Duration of Incident (mins)

Incident Type



Which DMA has the most collisions?

	index	DMA	Collision Count	Incident Count	% Collisions
9	9	SAN FRANCISCO-OAK-SAN JOSE	11385	32401	0.351378
4	4	LOS ANGELES	27217	82327	0.330596
7	7	SACRAMNTO-STKTON-MODESTO	5051	15756	0.320576
3	3	FRESNO-VISALIA	1302	4201	0.309926
0	0	BAKERSFIELD	194	687	0.282387
8	8	SAN DIEGO	3129	11338	0.275975
5	5	MONTEREY-SALINAS	630	2287	0.275470
6	6	RENO	86	322	0.267081
10	10	SANTABARBRA-SANMAR-SANLUOB	571	2291	0.249236
1	1	CHICO-REDDING	339	1606	0.211083
11	11	YUMA-EL CENTRO	48	278	0.172662
2	2	EUREKA	90	525	0.171429

Takeaways

- LA and the Bay Area have the highest populations but also the highest rates of incident per 100,000 people. Advertising dollars might be most efficient in these areas.
- LA and the Bay Area have the most collisions, would could produce the most leads. Investing advertising dollars in this area and procuring more law firm partnerships around areas could be beneficial

Next Steps and Curiosities

- Continue to investigate seasonality over all 12 months of the year. This data can help Walker schedule staff more accurately to handle call volume.
- Look into Duration. Is there a correlation between longer duration and ultimate leads for Walker?
- Which incident type is ultimately the biggest driver of revenue for Walker?

THANK YOU