UC San Diego

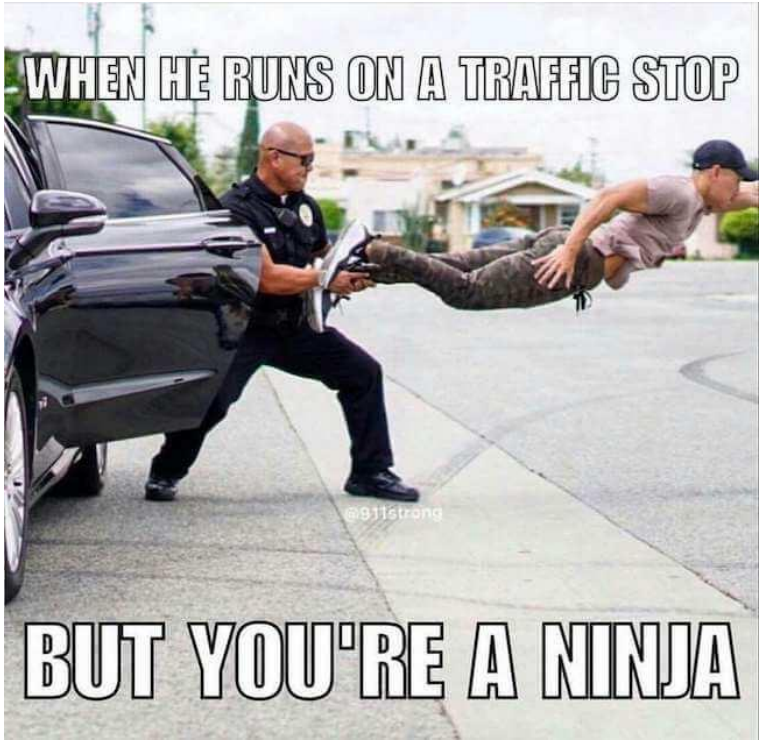
Police Vehicle Stops Data Analysis

ECE 143 - Group 5

Teammates: Param Chordiya, Ninad Ekbote, Divya Sri Dodla, Yi-yang Chen, Yanchen Jing

Agenda

- Overview and Motivation
- Dataset, Data processing and EDA
- Temporal Analysis
- Geospatial Analysis
- Demographic Analysis
- Summary and Conclusion



Overview And Motivation

- San Diego is densely populated urban area with huge vehicle traffic.
- Police department is continuously monitoring for better public safety.
- Can we find few insights which helps them for better resource allocation?
- Many report says that Police department has negative bias towards few demographic groups.
- Does the dataset follow this trend?

UC San Diego

Dataset and Data Processing

Dataset

- The Dataset gives records of car stops/pullovers by the police from Jan 2014 to June 2018 in San Diego.
- [Police Vehicle Stops - City of San Diego Open Data Portal.]
- The data also describes attributes related to the record.
- The Data is in CSV format and can be represented in Tabular form.

Columns are features/Attributes related to the pullover.

A row is a pullover record

	stop_id	stop_cause	service_area	subject_race	subject_sex	subject_age	date_time	date_stop	time_stop	sd_resident	arrested	searched	search_details_id	search_details_type	search_details_description
0	1044975	Moving Violation	110	W	0.0	24	2014-01-01 01:25:00	2014-01-01	1:25	Y	N	N	1208956.0	ActionTaken	Citation
1	1044976	Moving Violation	320	W	0.0	42	2014-01-01 05:47:00	2014-01-01	5:47	Y	N	N	1208957.0	ActionTaken	Verbal Warning
2	1044977	Moving Violation	320	L	0.0	29	2014-01-01 07:46:00	2014-01-01	7:46	Y	N	N	1208958.0	ActionTaken	Verbal Warning
3	1044978	Moving Violation	610	W	0.0	23	2014-01-01 08:10:00	2014-01-01	8:10	Y	N	N	1208959.0	ActionTaken	Citation
4	1044980	Equipment Violation	930	H	0.0	35	2014-01-01 08:35:00	2014-01-01	8:35	N	N	N	1208961.0	ActionTaken	Citation

Data processing

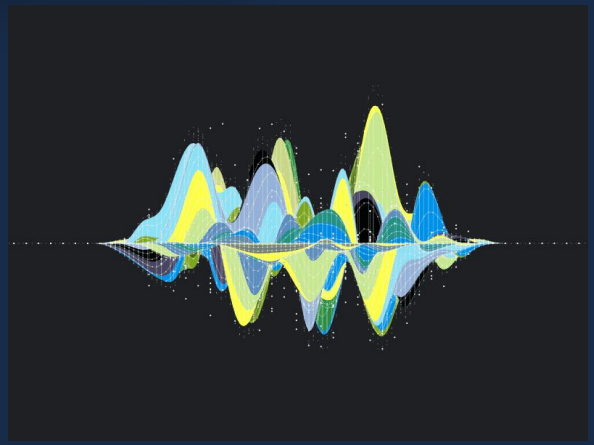
- Combining multiple CSV Files on both rows and Columns.
- Removing Columns that fails to provide any insights.
  - For eg: the ‘stop ID’ fails to provide meaningful insights.
- Removing NaN values.
- Encoded data replaced with original non-decoded entries.
  - For eg: For all races encoded as letters in the CSV data.

    We replaced that with the original race tag.

    ‘W’ in CSV is replaced with ‘White’.

UC San Diego

Exploratory Data Analysis(EDA)



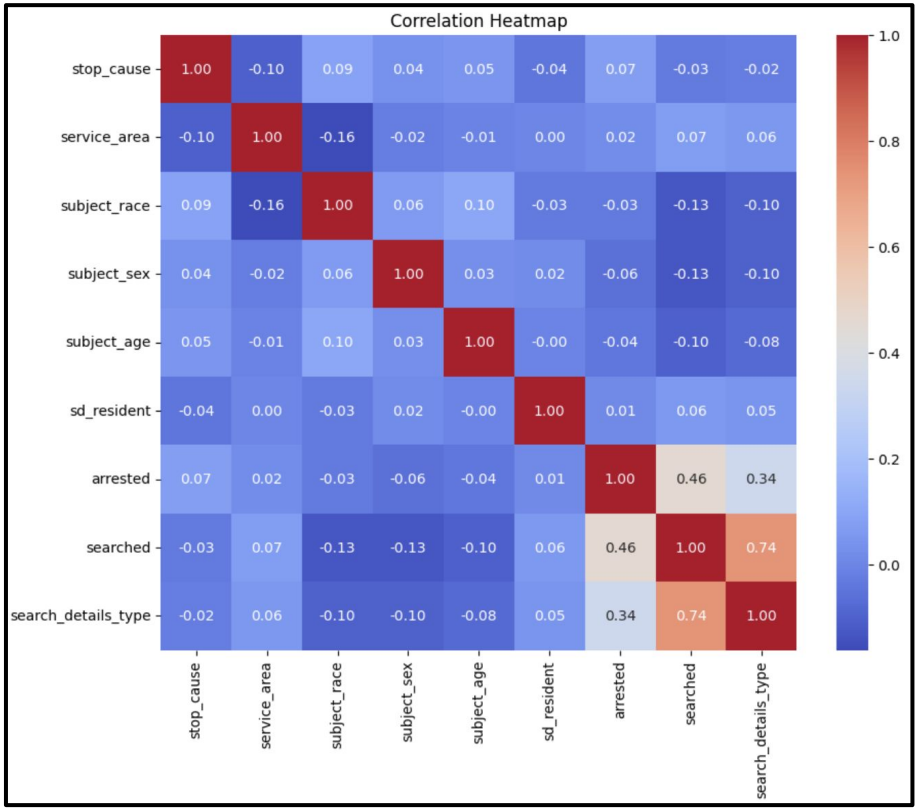
Exploratory Data analysis(EDA)

- The main purpose of EDA is to help look at data before making any assumptions.
- EDA also helps to identify important features/Attributes.
- Important features help us to form meaningful inferences.
- List of features:

stop_id	action	arrested
stop_cause	subject_age	searched
service_area	date_stop	search_details_id
subject_race	time_stop	search_details_type
subject_sex	sd_resident	search_details_description

EDA (Finding Important Features)

- Find inter-dependency between features.
- If 2 features are highly correlated use 1.
- ‘Searched’ and ‘Search_ details_type’ highly correlated so we can take only one



EDA(Important Features)

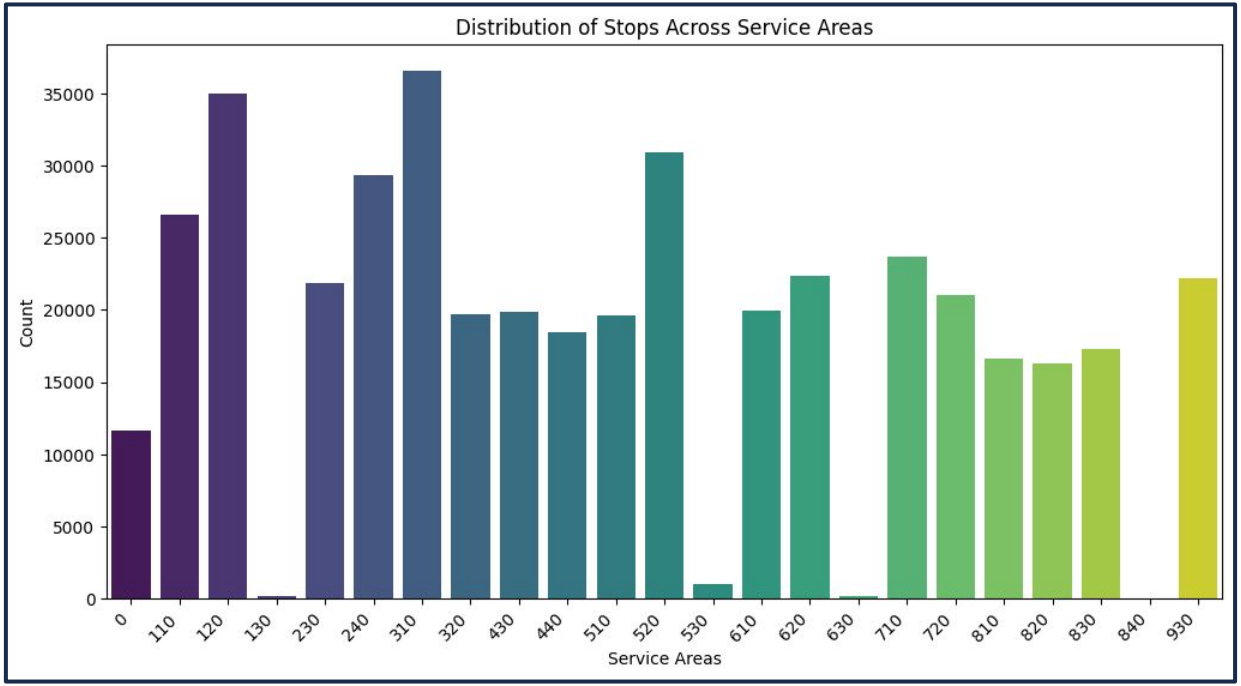
- We can’t include all features in the correlation Heatmap.
- Features related to ‘Time’ are important for Temporal Analysis.
- Binary Features are less likely to have meaningful insights.
- So some of the Important features:

Stop_cause	Subject_race	Date_stop
Service_area	Subject_age	Time_stop

- Later we decompose Date and time in sub columns of months, year, hours for plotting purposes.

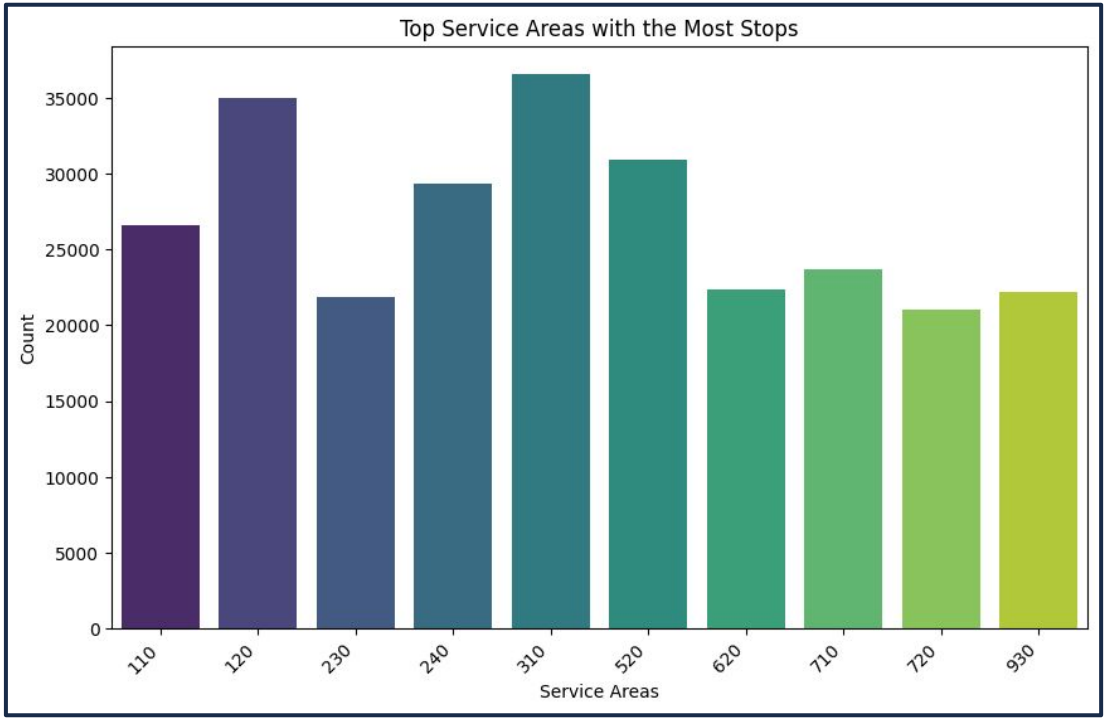
EDA(Service Area)

- Stops per Service Area Codes tell us which area has more pull over rate.
- This information can be used for Geospatial analysis of the Data.



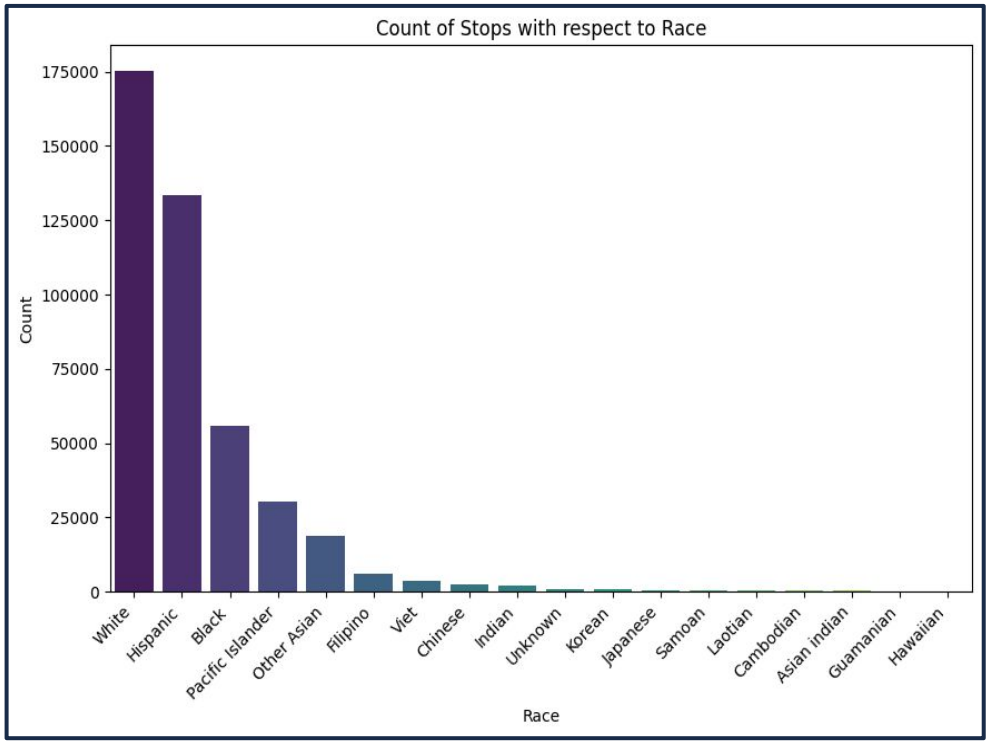
EDA(Service Area)

- Top ten service areas which has the most number of stops.
- We have referred to official San diego Service codes.



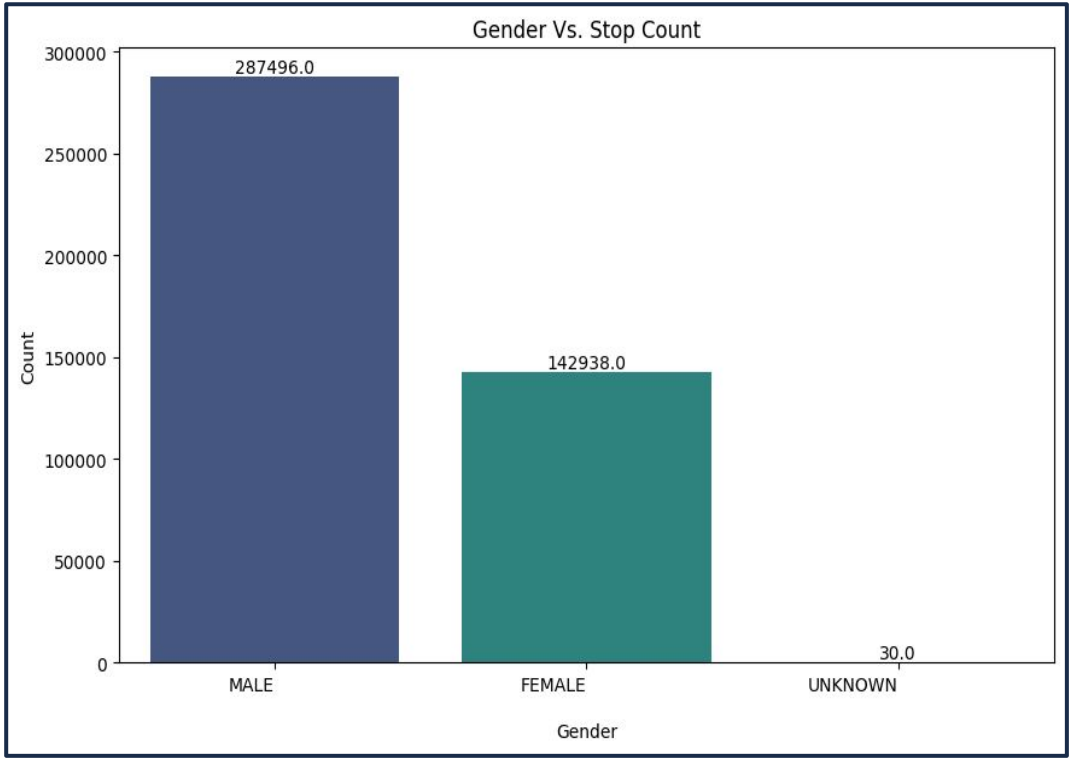
EDA(Subject Race)

- Higher chance of white individuals being pulled over, without considering population demographics.
- San Diego Demographics: Whites 45.9%, Hispanics 33.5%, Blacks 4.7%.
- Pullover rate for blacks almost 3 times in proportion, despite their population being approximately 1/10th of whites.
- City Demographic Profiles San Diego County

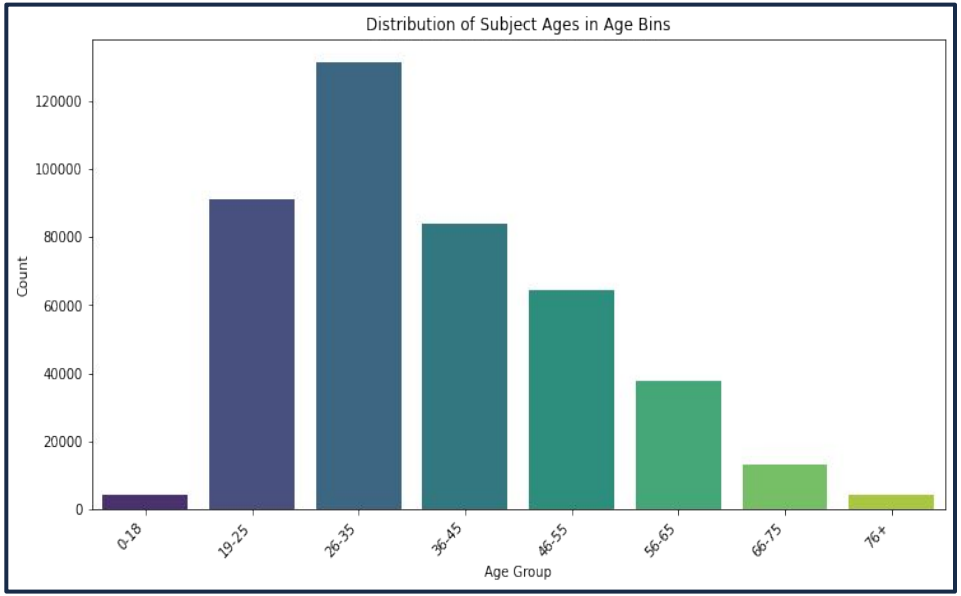


Stops Based on Genders

- Observe that we can find more number of Male Drivers pulled over as compared to Female Drivers.
- But the population of Males and Females is roughly equal in San Diego.
- So we might interpret that Male drivers are twice as likely to drive recklessly or being pulled over.
- City Demographic Profiles San Diego County

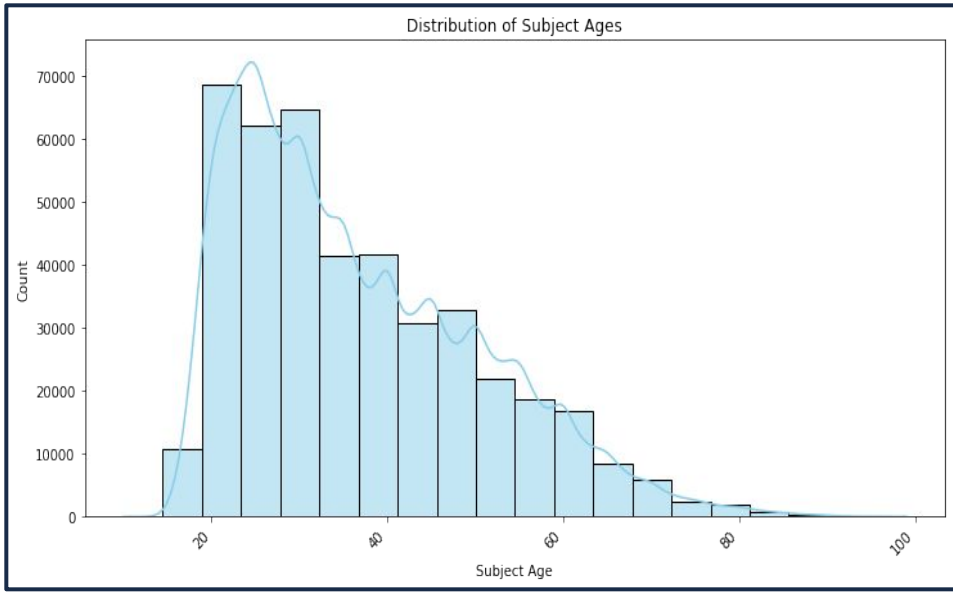


EDA(Subject Age)



- Most stops observed in the age group of 26-35.

- This age group aligns with higher numbers in San Diego demographics.



EDA

- We now have figured out our important features and done basic EDA.

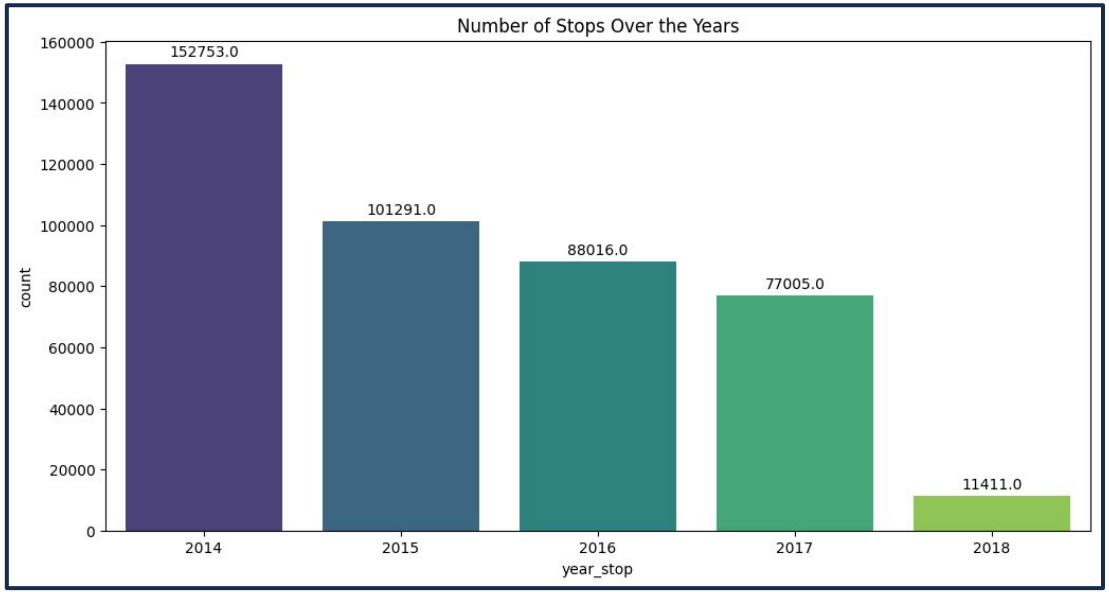
- Further we can move on to work on detailed insights based on
  - Temporal Analysis
  - Geospatial Analysis
  - Demographic Analysis

UC San Diego

Temporal Analysis

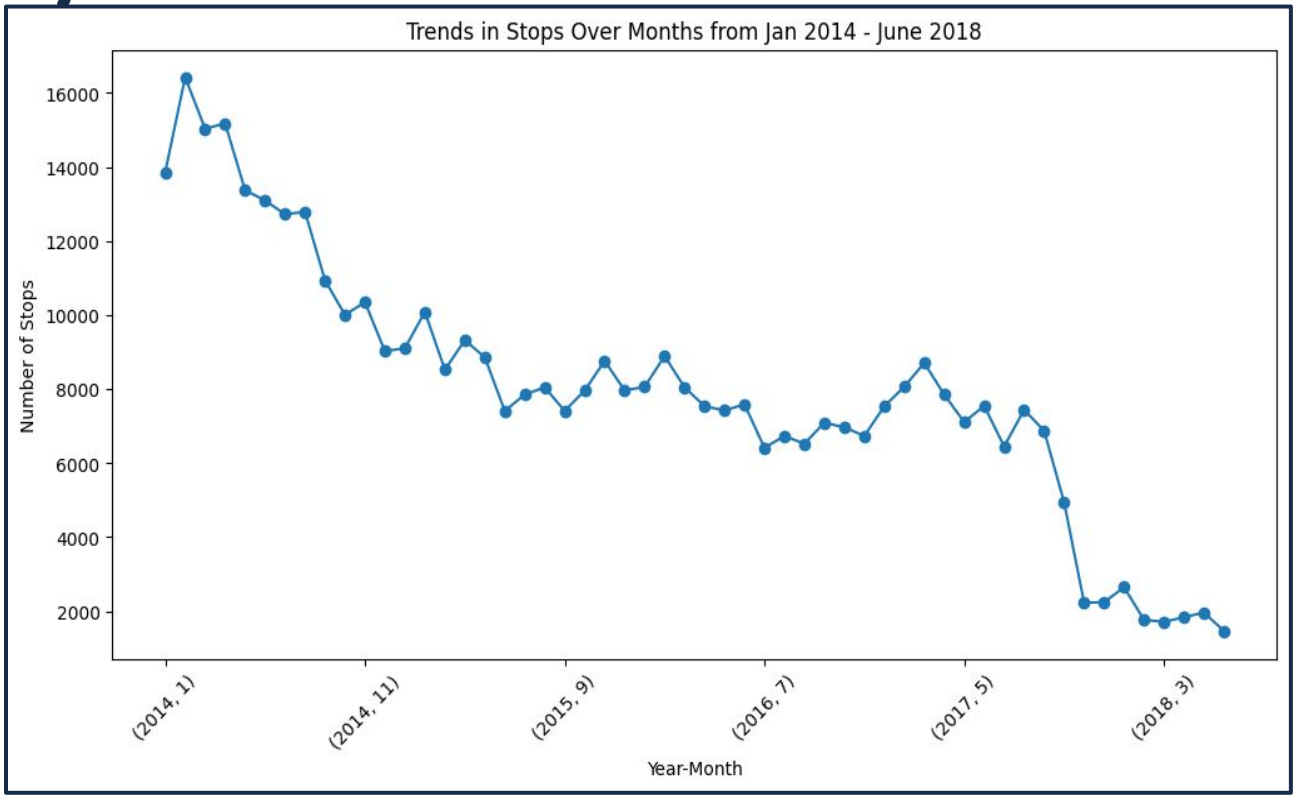
Temporal Analysis

- The number of stops per year is following a negative trend.
- There are two possible reasons for this:
  - More adherence to the law
  - Fewer Cops on roads
- 2018 cannot be considered as the data is collected till June.
- [Census of State and Local Law Enforcement Agencies, 2018 – Statistical Tables]



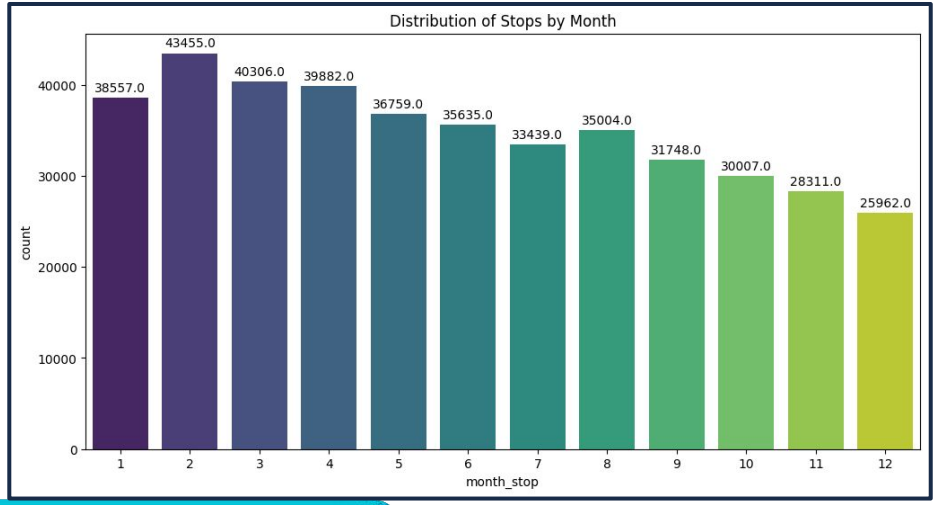
Temporal Analysis

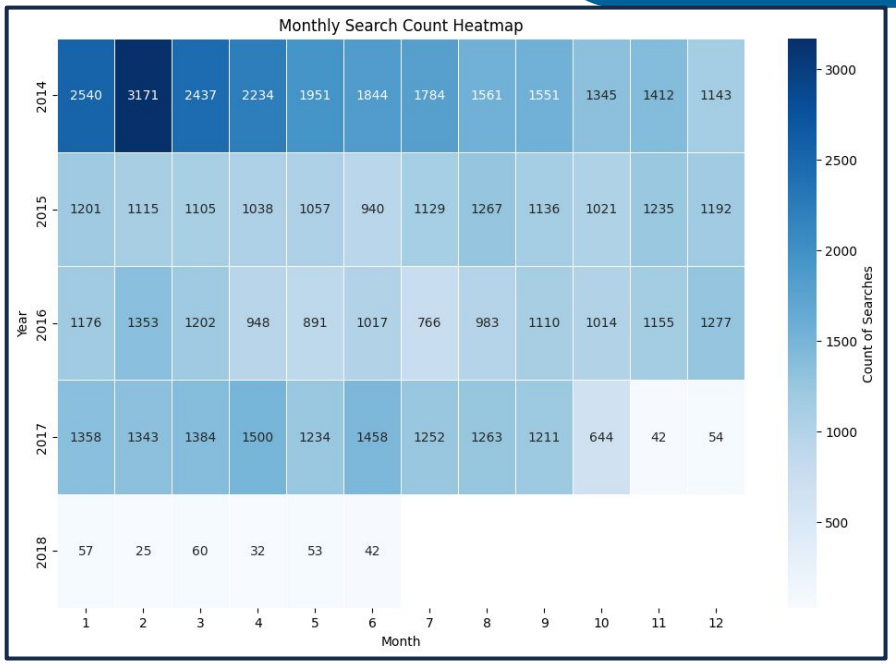
- In Support of previous Bar graph, we can see a decreasing trend over years.



Temporal Analysis

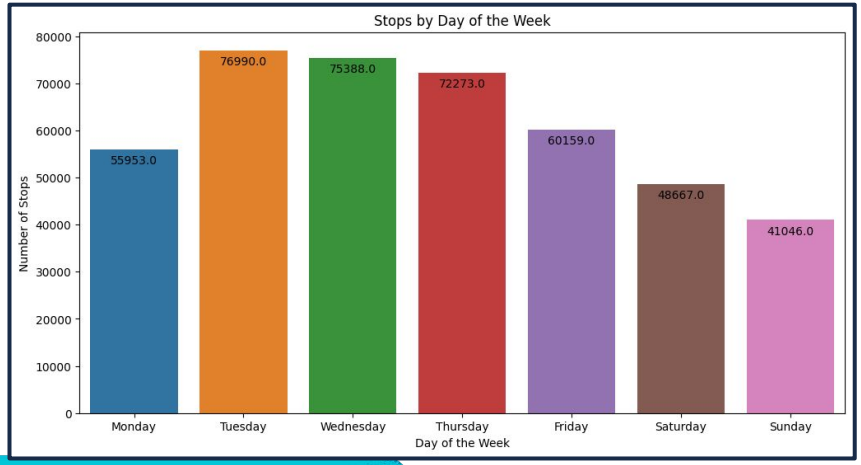
- The number of stops per month is following a negative trend.
- [NBC News Article]

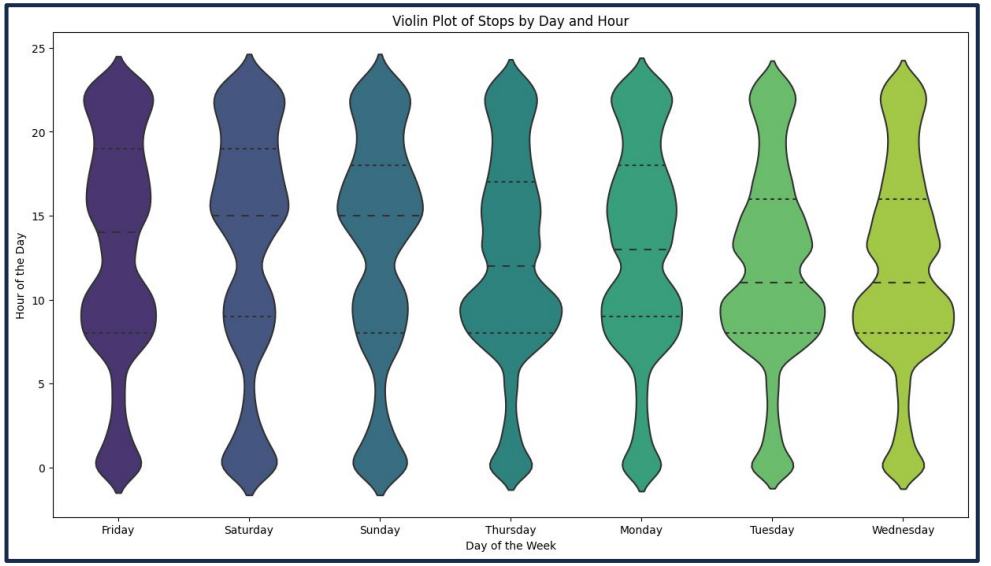




Temporal Analysis

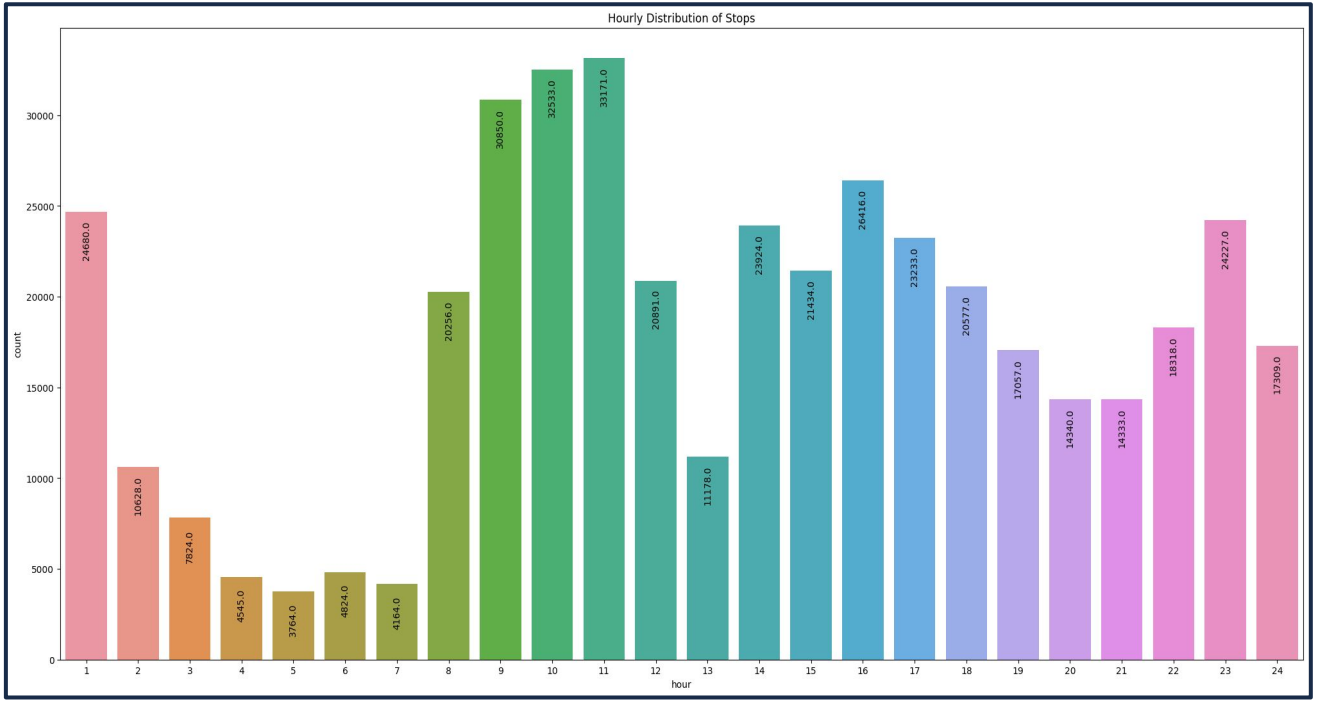
- Here we find that the Stop rate decreases from Tuesday to Sunday, Monday being an exception.
- No special reason but maybe due to public being more attentive to start.





Temporal Analysis

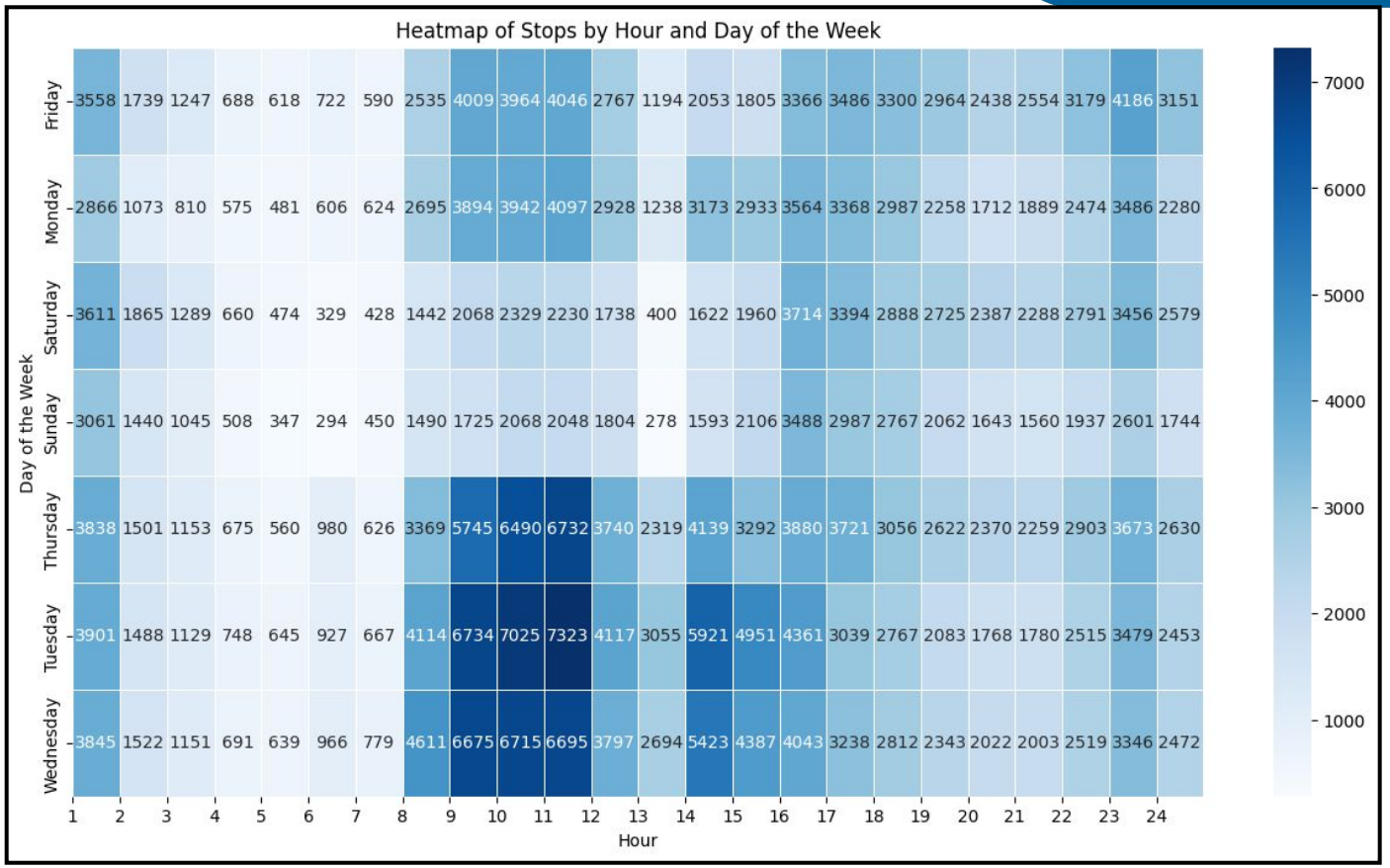
- Bar chart: Hours of the day vs Count.
- Higher number of stops during rush/office hours (8 to 11).
- Mornings are relatively quiet.
- Late nights show a small rise, possibly people returning from work.



Temporal Analysis - Heatmap for Days vs Hours

UC San Diego

- Detailed view by days and hours using heatmap.
- Weekdays, especially office hours, show higher numbers of pullovers.
- Weekends deviate from this pattern with lower pullover rates.



Temporal Analysis - Inferences

In Summary we can say:

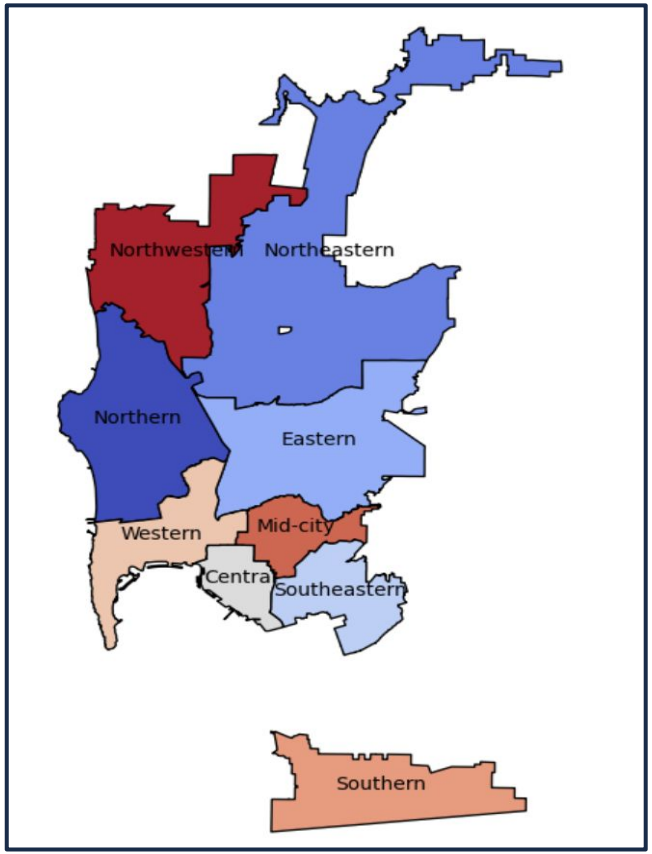
- Decrease in stops and pullovers from 2014 to 2018.
- General trend of more stops on weekdays, particularly during office hours.
- Feb 2014 shows a spike, but excluding it, a downward monthly trend is observed.
- Recommendation: Increase police workforce during office hours, optimize resource allocation from March to December, especially during festive days.

UC San Diego

Geospatial Analysis

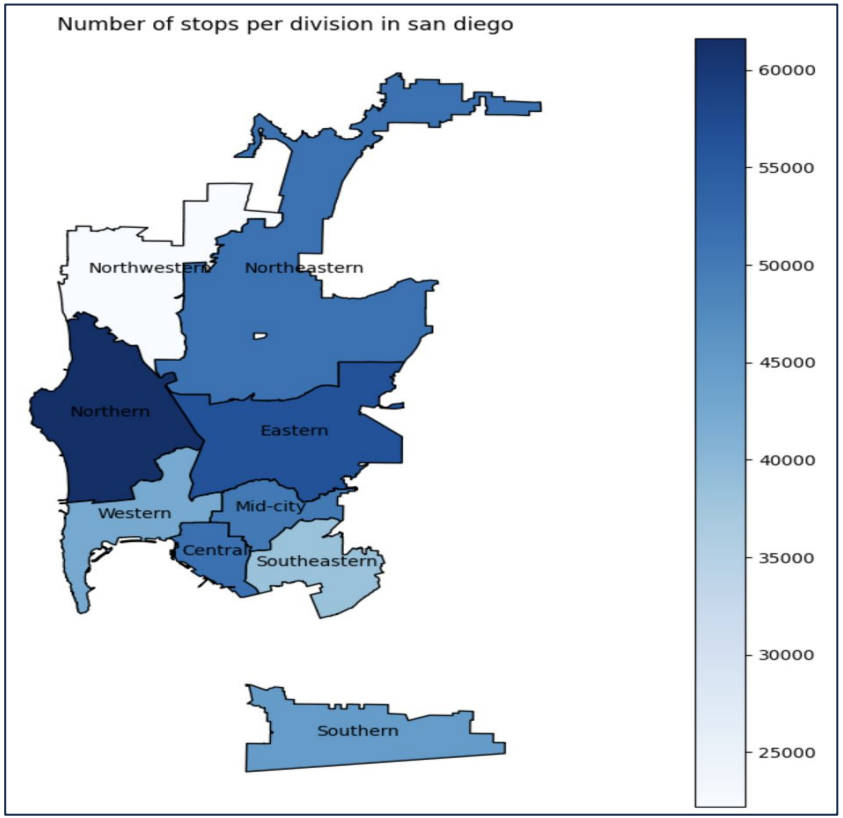
Geospatial Analysis

- Which parts of San Diego has more stops?
- Which part of San Diego has highest probability of arrest?
- Note:- The police departments has divided San Diego into 9 divisions



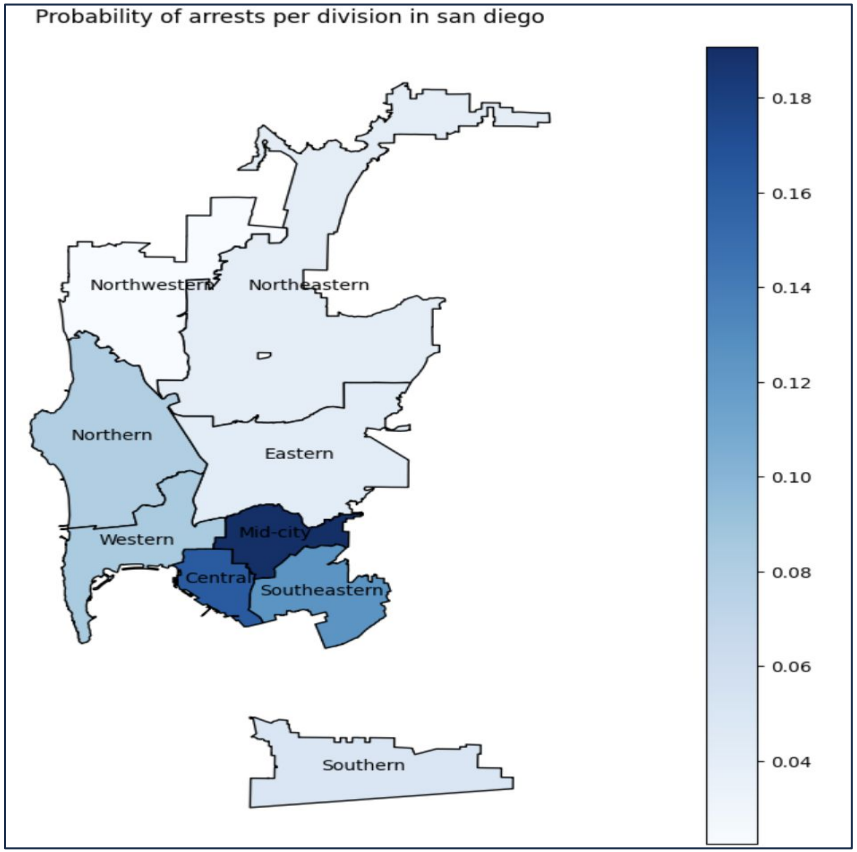
Geospatial Analysis - Most Stops

- Northern division has highest stops.
- Many tourist attractions.
- More population compared to other regions.



Geospatial Analysis - Most Arrests

- Mid-city division has highest probability of Arrests.
- Mid-city has highest probability of arrests, this could be because the area experiences high crime rate, rules violations.



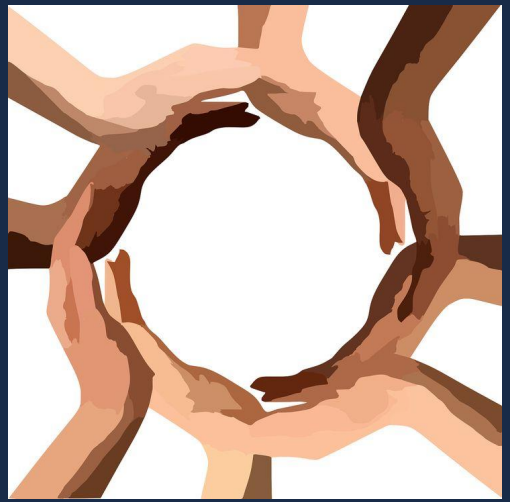
Geospatial Analysis - Inference

- Northern division has more stops, due huge population and many tourist attractions.
- Helps to identify the requirement of resource allocation.
- Mid-city has highest probability of arrests, this could be because the area experiences high crime rate, rules violations.

UC San Diego

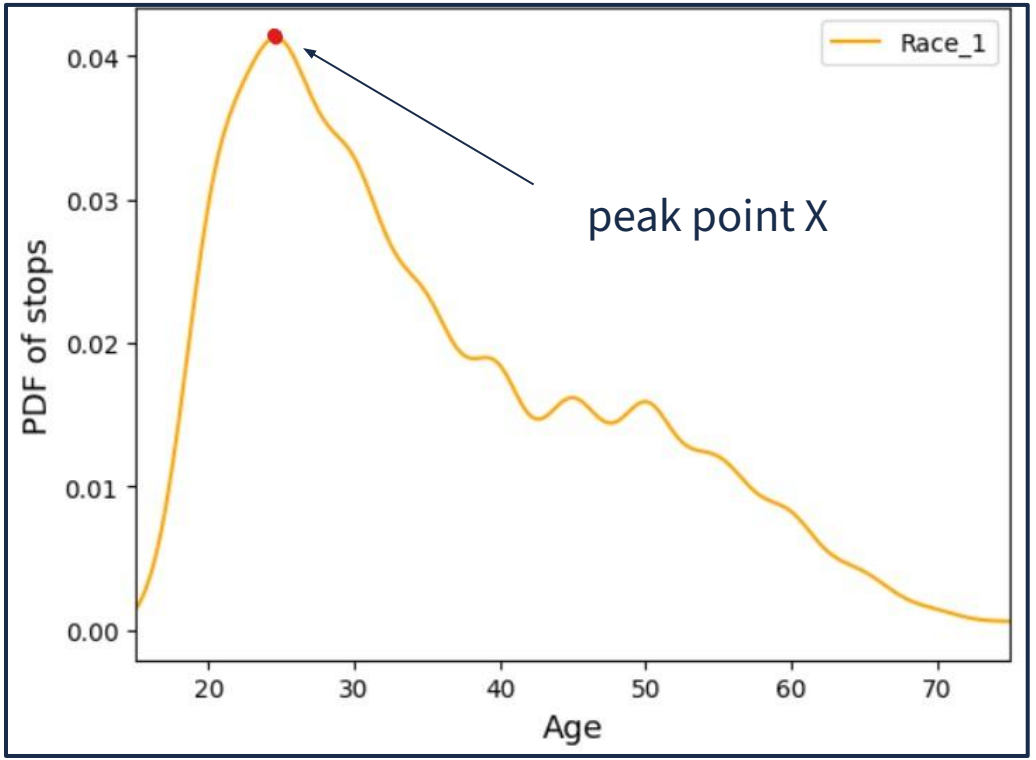
Demographic Analysis

Main Question :- Is there Bias and Discrimination against a Race or age group?



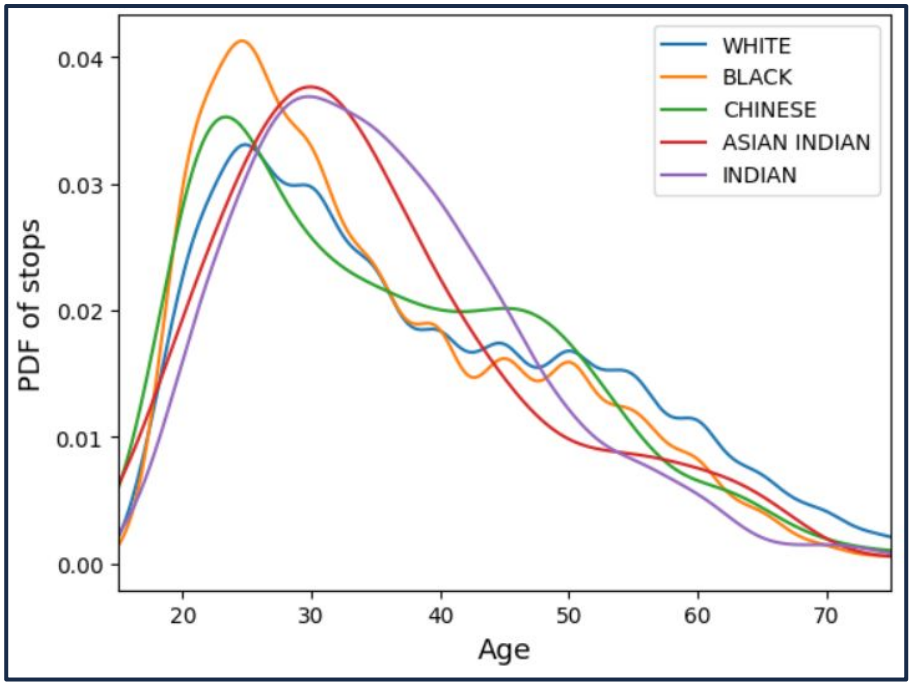
Demographic Analysis

- Point X has coordinates.
  - (27,0.04)
- P(stopping|Race, Age) = Y_coordinate.
- Probability for current example
  - P(stopping|Race_1, 27) = 0.04
- More probability means more bias towards one age group or Race.



Demographic Analysis

- From the combined graph, P(stopping|Black, 27) = 0.04 which is the highest.
- This means that a young black person will get stopped by the cops more often.
- Ideally the PDF for Races should be same. Only then we can say that Bias against anyone group is absent. But that is not the case here.



Demographic Analysis

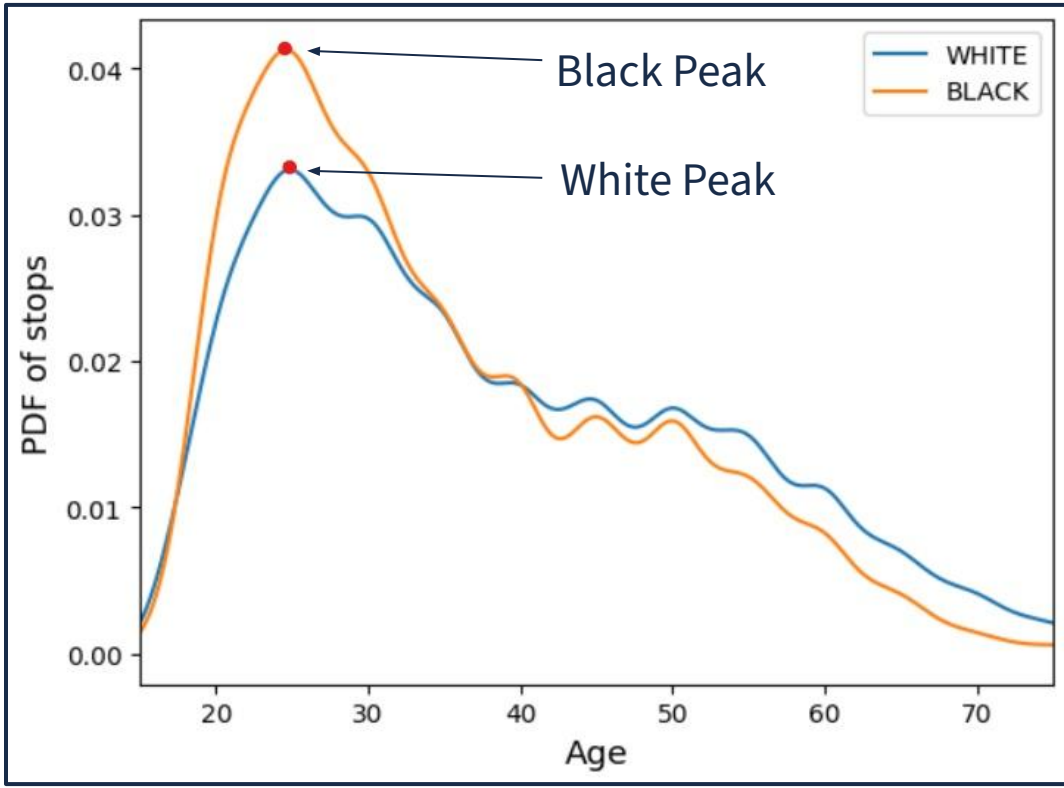
- P(stopping|Black,27) = 0.04
- P(stopping|White,27) ~ 0.032

  i.e

  P(stopping|Black, 27)>

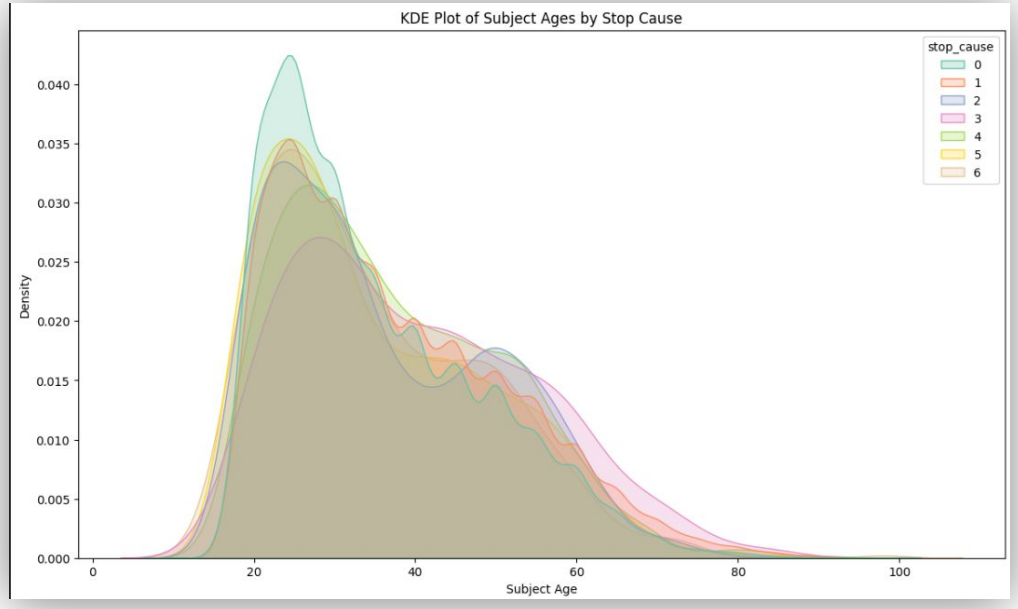
  P(stopping|White, 27)

- This means a young black person is more likely to get stopped than a young white person.



Age Vs Stop Cause

- Same trends as observed in the age vs Race graph.
- Youngsters are more likely to be stopped by the police than compared elderly folks.
- People usually stop driving cars/vehicles after 65. Hence a sudden drop in %.



Demographic Analysis-Inferences

- The probability tells that a young Black person is more likely to get stopped by the cops.
- Does this mean that there is a Race/age bias?
- The answer is no, we still need more data to analyse.

Summary/Conclusion

- **Temporal Analysis**
  By the trend of stops in a day and a year, we suggest that increasing police workforce on commute hours and holidays from March to December help to optimize the resource of police.

- **Geospatial Analysis**
  The division with higher stops require more resource allocations.

- **Demographic Analysis**
  The probability distribution indicates that there is a bias, but more evidence is needed to back the statement.

Outliers - Always follow the rules!!

Reference

- City Demographic Profiles San Diego County
- Racist Comments, Excessive Force and Offensive Behavior Revealed in San Diego Police Department Internal Affairs Cases
- https://voiceofsandiego.org/wp-content/uploads/2014/01/Vehicle-Stop-form.pdf
- RIPA police stop data - race of persons stopped - City of San Diego Open Data Portal

UC San Diego