# Stanford Poling Dataset

# Nashville

Presenter: Rahul Bajaj

# Stanford Policing Dataset

THE STANFORD
**OPEN POLICING**
PROJECT

# Dataset:

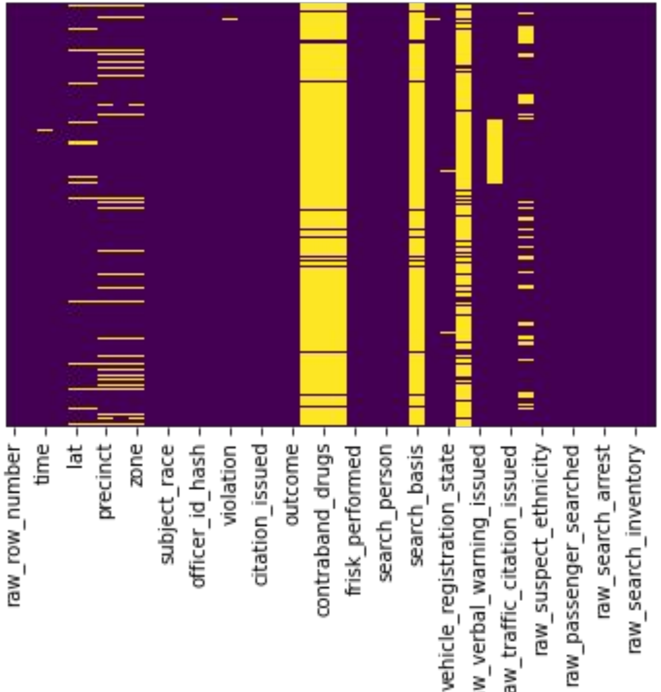
- Initially we have 3092351 rows of data and 42 columns.

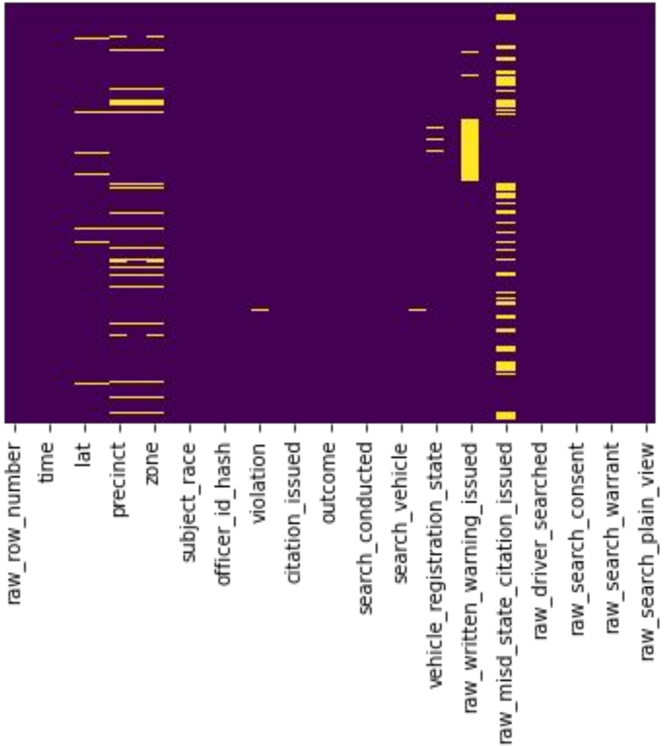
```
In [40]: df.columns

Out[40]: Index(['raw_row_number', 'date', 'time', 'location', 'lat', 'lng', 'precinct',
               'reporting_area', 'zone', 'subject_age', 'subject_race', 'subject_sex',
               'officer_id_hash', 'type', 'violation', 'arrest_made',
               'citation_issued', 'warning_issued', 'outcome', 'contraband_found',
               'contraband_drugs', 'contraband_weapons', 'frisk_performed',
               'search_conducted', 'search_person', 'search_vehicle', 'search_basis',
               'reason_for_stop', 'vehicle_registration_state', 'notes',
               'raw_verbal_warning_issued', 'raw_written_warning_issued',
               'raw_traffic_citation_issued', 'raw_misd_state_citation_issued',
               'raw_suspect_ethnicity', 'raw_driver_searched',
               'raw_passenger_searched', 'raw_search_consent', 'raw_search_arrest',
               'raw_search_warrant', 'raw_search_inventory', 'raw_search_plain_view'],
              dtype='object')
```

# Data Filtering

1. Filter 7 columns that have > 80% NaN values.



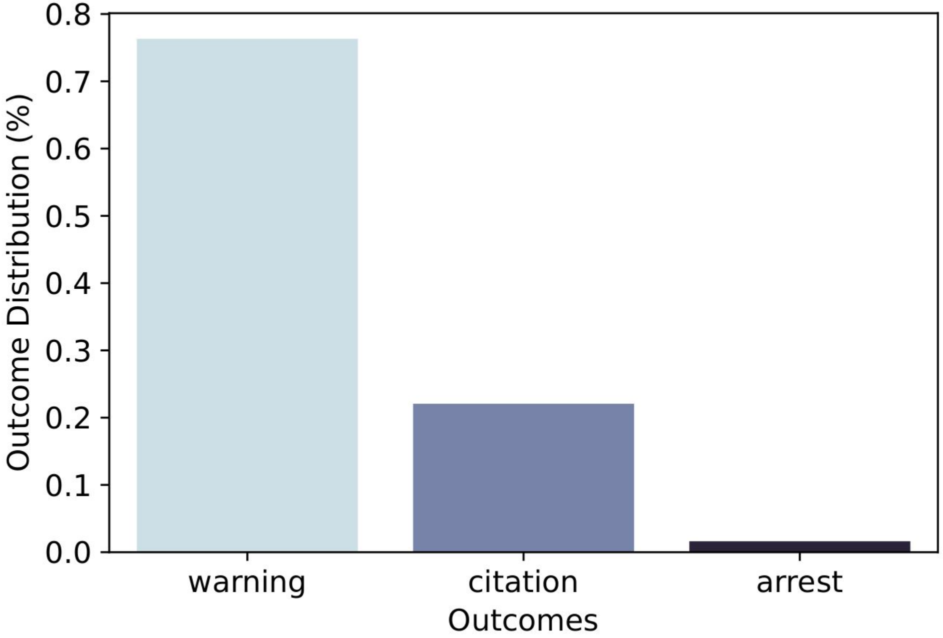
# Data Filtering:

1. We have two fields: ‘date’ and ‘time’ which indicate the date of event and the time at which event occurred. Both these fields are in the string format, to perform calculations on these fields we create  a new column called ‘date_time’ with to_datetime format.
2. We create an index for the ‘date_time’ field to access related data faster.
3. We convert the ‘arrested’  field from string to ‘bool’
4. Finally, we analyse 3079529 rows and 37 columns of data.

# Effect of Race on Crime

1) What are the outcomes of traffic policing?
2) Does the driver race matter?

# What is distribution of outcomes while traffic policing?



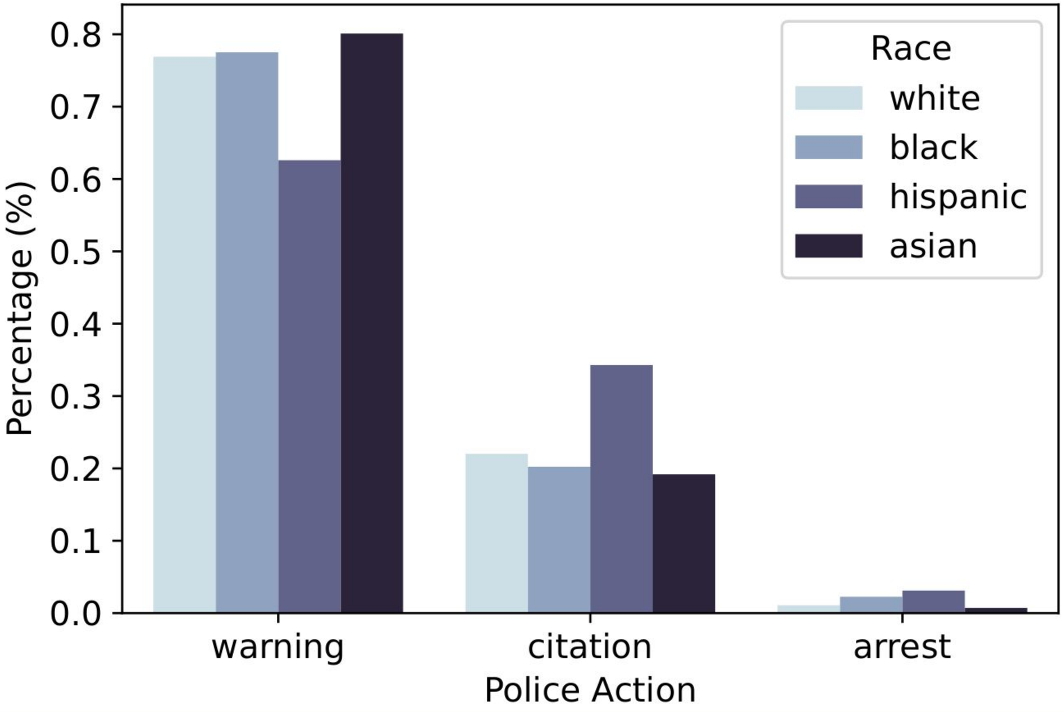
# Outcomes for drivers of a particular race:

Most:

Warnings: Asians

Citation: Hispanic

Arrests: Hispanic

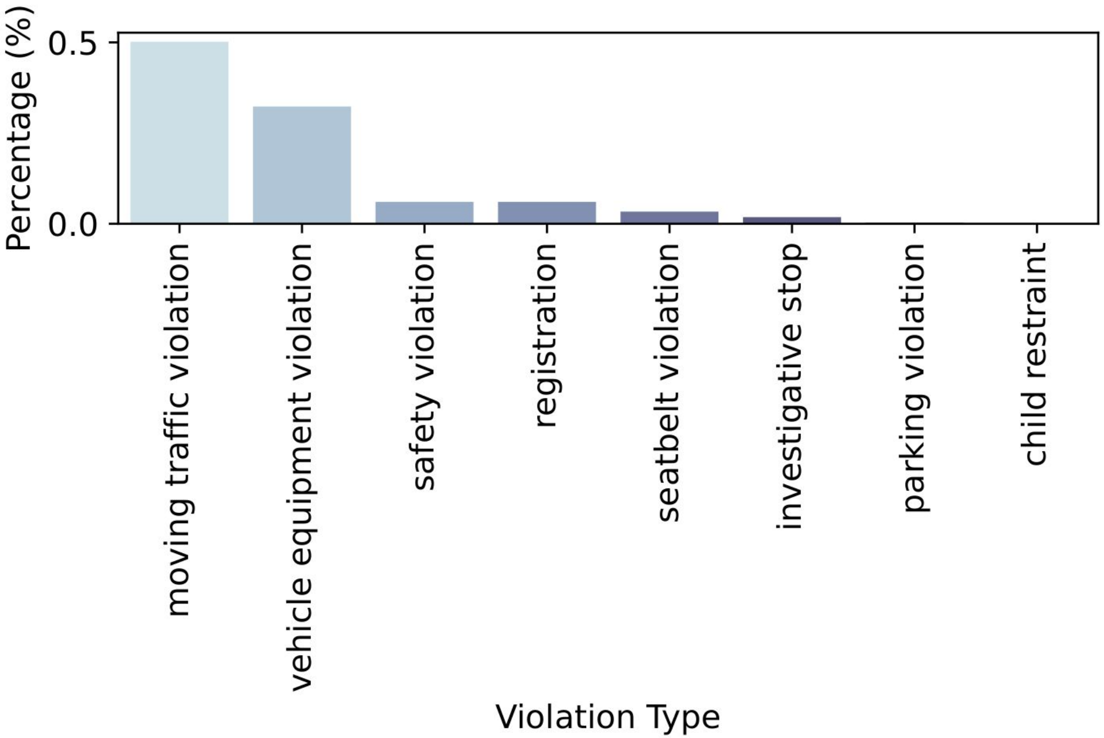


# Does the gender of a driver have an impact on police behavior during a traffic stop?

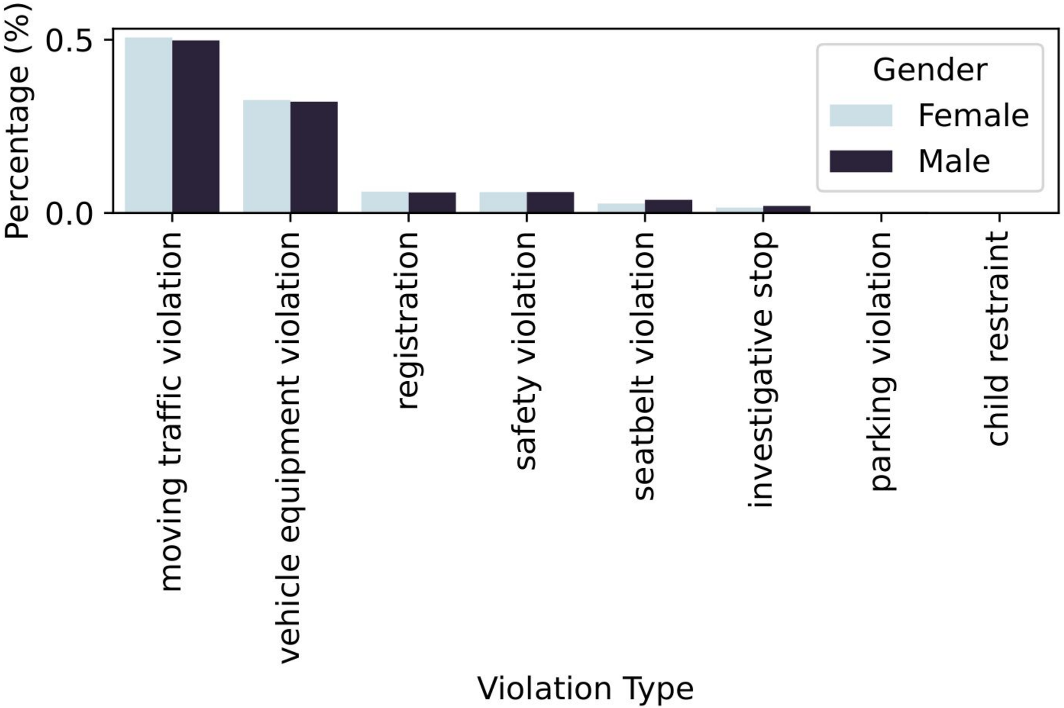
1) Which are the different types of traffic rule violations?
2) Do the genders commit different violations?
3) Does gender affect who gets a speeding ticket?
4) Does gender affect whose vehicle is searched?
5) Does gender affect who is frisked during a search?

# Types of violations

Interesting! Half of all violations are for moving violations, followed by equipment violations and safety violation.

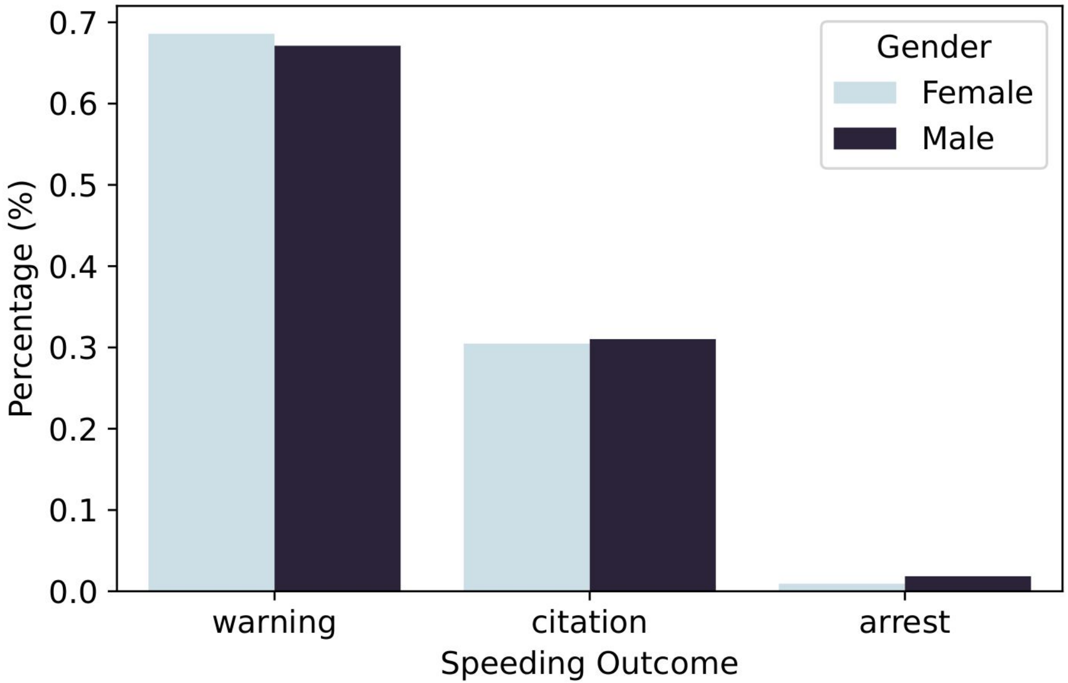


# Whether male and female drivers tend to commit different types of traffic violations?



# Does gender affect who gets a speeding ticket?

Interesting! The data fails to show that gender has an impact on who gets a ticket for speeding.



# Does gender affect whose vehicle is searched?

```
policing_df.groupby('subject_sex').search_conducted.mean()
```

```
subject_sex
female     0.023516
male       0.053492
Name: search_conducted, dtype: float64
```

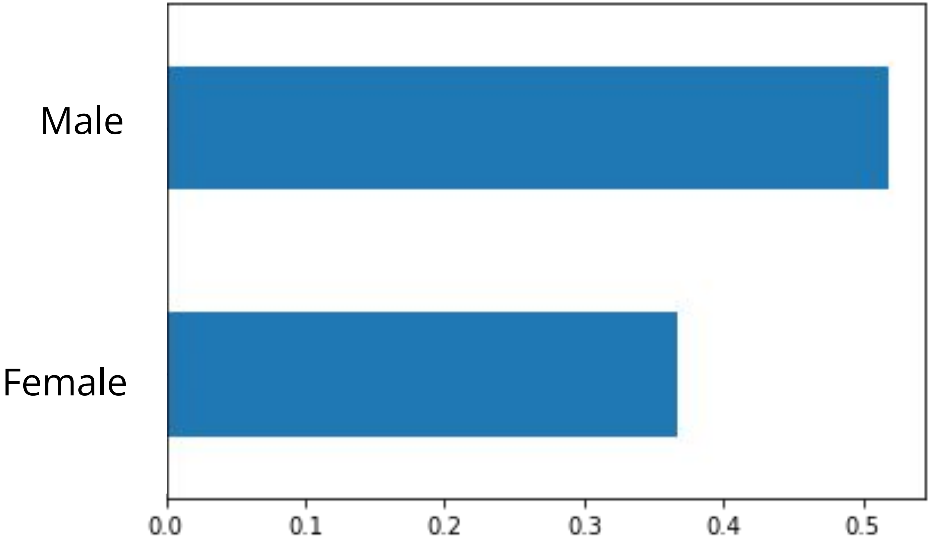
Wow! Male drivers are searched more than twice as often as female drivers.

Even though men and women commit similar kinds of crime, searches are performed more on men for each violation type when compared to women!

| violation | subject_sex | |
|---|---|---|
| child restraint | female | 0.030345 |
| | male | 0.071795 |
| investigative stop | female | 0.094319 |
| | male | 0.184111 |
| moving traffic violation | female | 0.019689 |
| | male | 0.046677 |
| parking violation | female | 0.024733 |
| | male | 0.047719 |
| registration | female | 0.025768 |
| | male | 0.055277 |
| safety violation | female | 0.023334 |
| | male | 0.048620 |
| seatbelt violation | female | 0.031903 |
| | male | 0.060277 |
| vehicle equipment violation | female | 0.024817 |
| | male | 0.055152 |

# Does gender affect who is frisked during a search?

Interesting! The frisk rate is higher for males than for females.

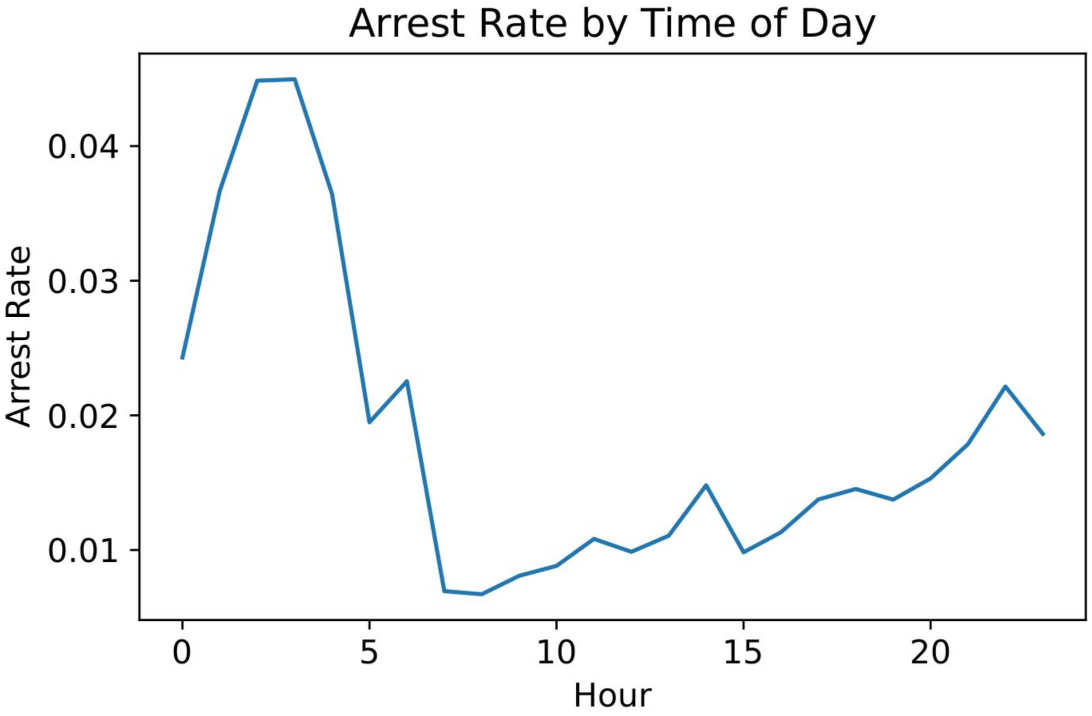


# Effect of Time and Drugs on Crime

1) Does time of the day affect arrest rate?
2) What is the trend for drugs found vs search conducted?

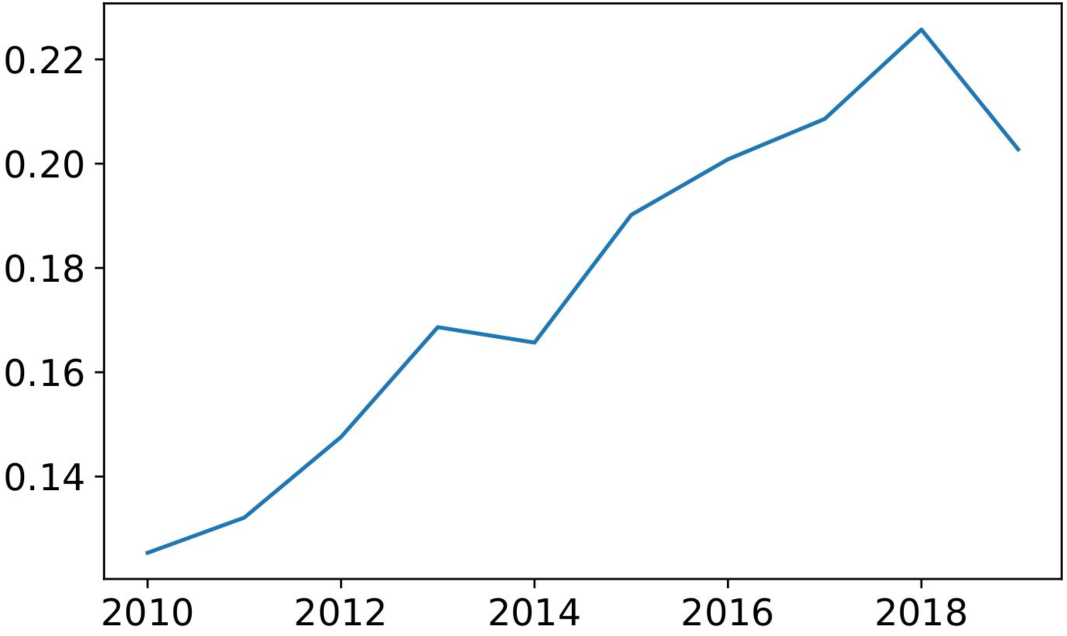
# Does time of the day affect arrest rate?

The arrest rate has a significant spike overnight, and then dips in the early morning hours.



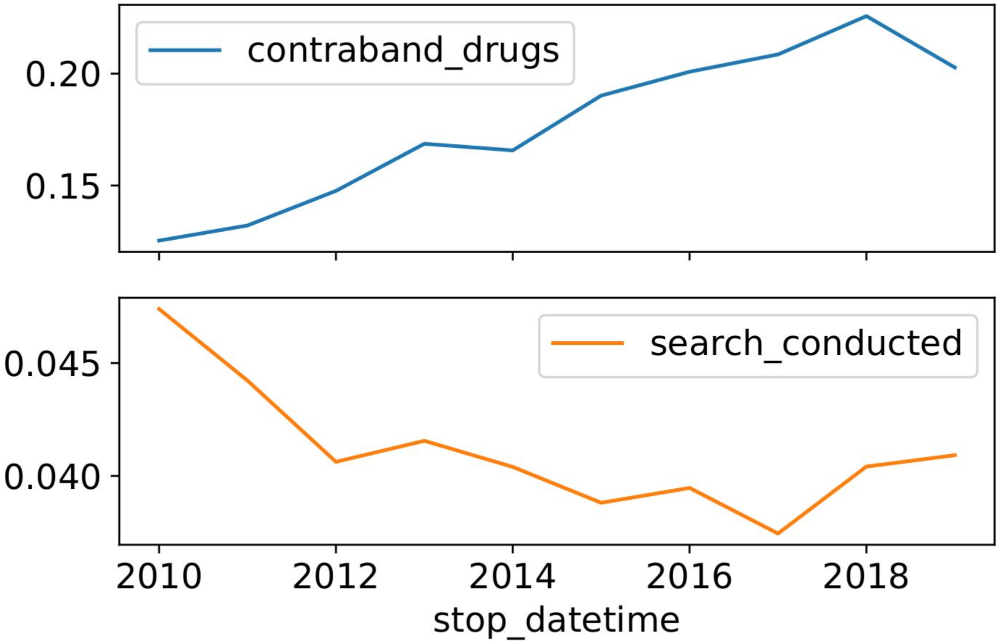
# Annual rate of drug related stops

The rate of drug-related stops nearly doubled over the course of 10 years

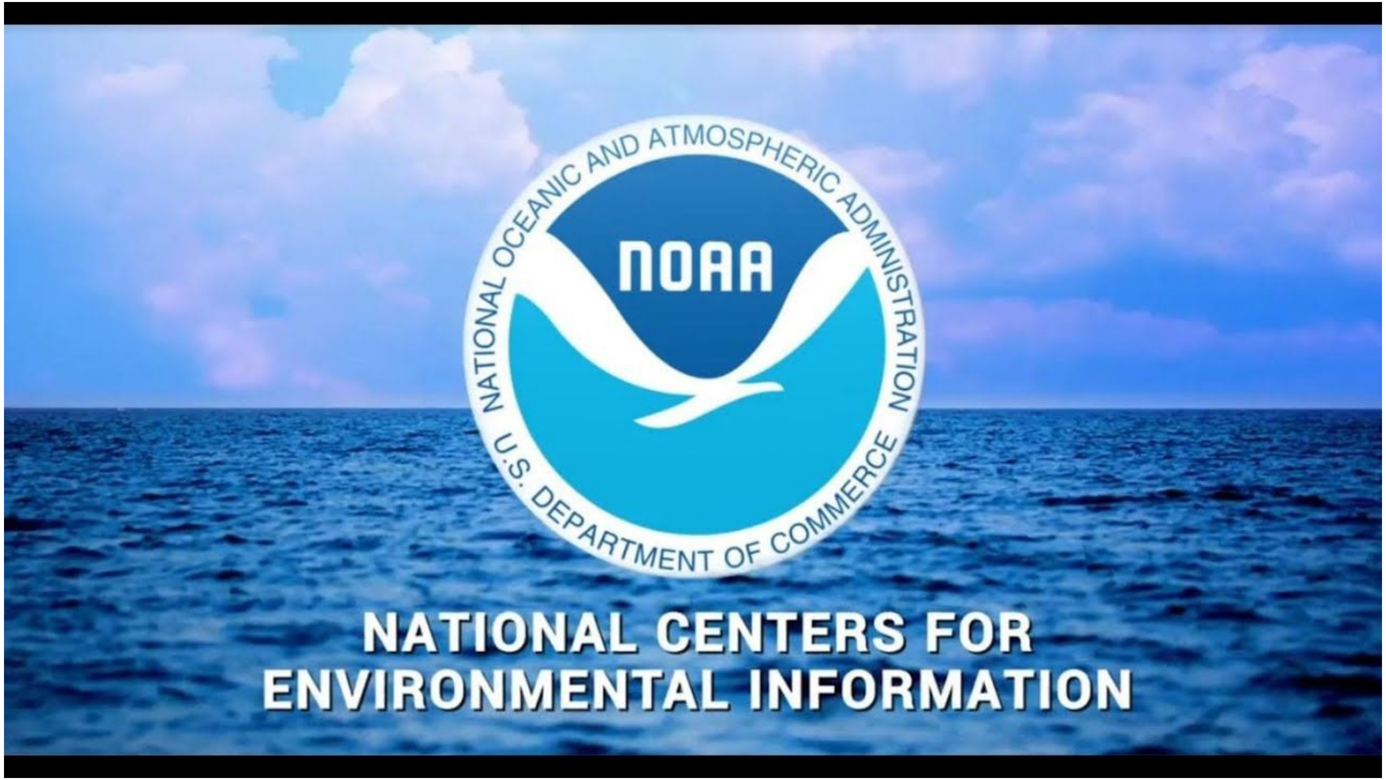


# Search Conducted V/s Drugs Found Annually

Wow! The rate of drug-related stops increased even though the search rate decreased.



# Effect of weather on Policing



# Dataset

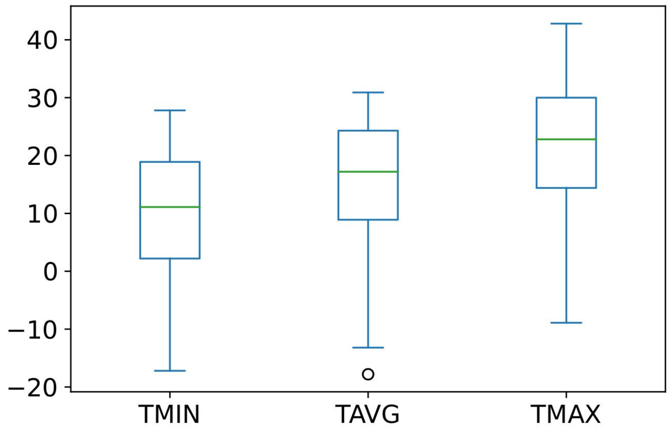
We consider the data from a single weather station (Nashville International Airport) in Nashville. This is not ideal, but it will still give us a general idea of the weather throughout the state.

Exploring the Dataset:

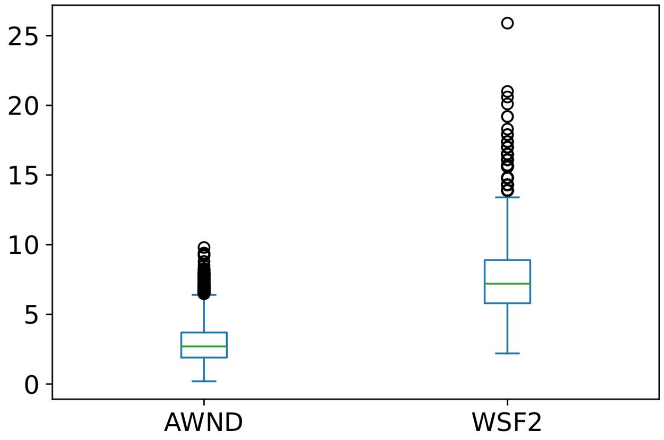
STATION: station ID
TAVG, TMIN, TMAX: Temperature
AWND, WSF2: Wind speed
WT01 .. WT22: Bad weather conditions (0-20)

# Data Exploration

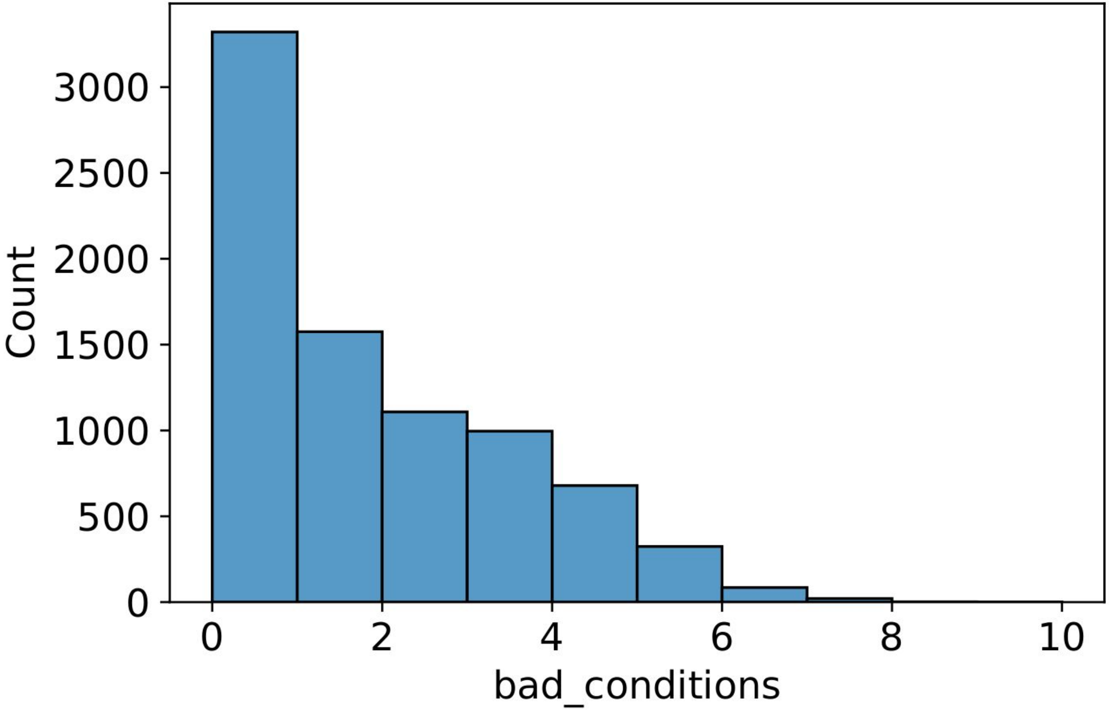
TAVG, TMIN, TMAX: Temperature




AWND, WSF2: Wind speed



# Data processing




0 : GOOD

1 - 4 : BAD

5 - 9 : WORST

# Does weather affect the arrest rate?

The arrest rate increases
as the weather gets worse

```
violation                   rating
child restraint             bad       0.036522
                            good      0.025097
                            worse     0.181818
investigative stop          bad       0.079678
                            good      0.085474
                            worse     0.079046
moving traffic violation    bad       0.014834
                            good      0.014513
                            worse     0.018629
parking violation           bad       0.011006
                            good      0.010638
                            worse     0.024390
registration                bad       0.017268
                            good      0.017018
                            worse     0.015903
safety violation            bad       0.014873
                            good      0.014501
                            worse     0.017791
seatbelt violation          bad       0.017129
                            good      0.016980
                            worse     0.013739
vehicle equipment violation bad       0.014714
                            good      0.014477
                            worse     0.017332
Name: arrest_made, dtype: float64
```

# Predict arrests made

- Logistic Regression: uses a generalized linear equation to describe the directed dependencies among a set of variables.
    - A number of statistical assumptions must be met.
    - Overfitting a concern (rule of ten), as well as outliers.
- Random Forest: Top-down induction based approach to classification and prediction. Averages many decision trees (CARTs) together.
    - No statistical assumptions; can handle multicollinearity.
    - Robust to overfitting and outliers.

We perform Logistic Regression (linear) and Random Forest (non-linear) classifier to predict whether arrests would be made or not.

# Classification performance of arrests made

Split data into train and test based on arrests made.
Y train (arrest made = 34752, arrest not made = 2113592)
Y test (arrest made = 14894, arrest not made = 905826)

Due to data imbalance, even though accuracy is very high but f1-score is very low

## Logistic Regression

Train Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 1.00 | 1.00 | 2113592 |
| 1 | 0.97 | 0.47 | 0.63 | 34752 |
| accuracy | | | 0.99 | 2148344 |
| macro avg | 0.98 | 0.73 | 0.81 | 2148344 |
| weighted avg | 0.99 | 0.99 | 0.99 | 2148344 |

---

Test Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 1.00 | 1.00 | 905826 |
| 1 | 0.93 | 0.44 | 0.60 | 14894 |
| accuracy | | | 0.99 | 920720 |
| macro avg | 0.96 | 0.72 | 0.80 | 920720 |
| weighted avg | 0.99 | 0.99 | 0.99 | 920720 |

## Random Forest

Train Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 1.00 | 1.00 | 2113592 |
| 1 | 0.97 | 0.47 | 0.63 | 34752 |
| accuracy | | | 0.99 | 2148344 |
| macro avg | 0.98 | 0.73 | 0.81 | 2148344 |
| weighted avg | 0.99 | 0.99 | 0.99 | 2148344 |

---

Test Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.99 | 1.00 | 1.00 | 905826 |
| 1 | 0.93 | 0.44 | 0.60 | 14894 |
| accuracy | | | 0.99 | 920720 |
| macro avg | 0.96 | 0.72 | 0.80 | 920720 |
| weighted avg | 0.99 | 0.99 | 0.99 | 920720 |

# Classification performance of arrests made

Randomly sampled 45,000 for each group to make a balanced dataset. Split data into train and test based on arrests made.
Y train (arrest made = 31500, arrest not made = 31500)
Y test (arrest made = 13500, arrest not made = 13500)

Even though we see a reduced accuracy from the previous results, here we observe that both the classes are performing equally well.

## Logistic Regression

Train Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.91 | 0.84 | 31500 |
| 1 | 0.89 | 0.75 | 0.81 | 31500 |
| accuracy | | | 0.83 | 63000 |
| macro avg | 0.84 | 0.83 | 0.83 | 63000 |
| weighted avg | 0.84 | 0.83 | 0.83 | 63000 |

----------------------------------------------

Test Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.91 | 0.84 | 13500 |
| 1 | 0.89 | 0.75 | 0.81 | 13500 |
| accuracy | | | 0.83 | 27000 |
| macro avg | 0.84 | 0.83 | 0.83 | 27000 |
| weighted avg | 0.84 | 0.83 | 0.83 | 27000 |

## Random Forest

Train Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.81 | 0.93 | 0.87 | 31500 |
| 1 | 0.92 | 0.78 | 0.85 | 31500 |
| accuracy | | | 0.86 | 63000 |
| macro avg | 0.87 | 0.86 | 0.86 | 63000 |
| weighted avg | 0.87 | 0.86 | 0.86 | 63000 |

----------------------------------------------

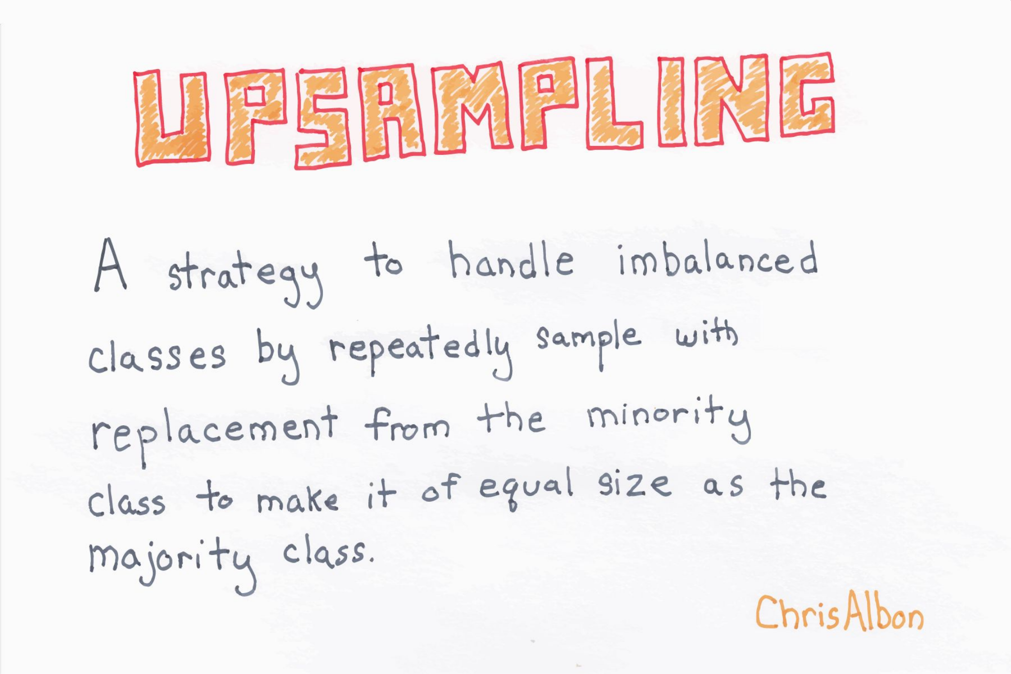
Test Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.79 | 0.91 | 0.84 | 13500 |
| 1 | 0.89 | 0.75 | 0.82 | 13500 |
| accuracy | | | 0.83 | 27000 |
| macro avg | 0.84 | 0.83 | 0.83 | 27000 |
| weighted avg | 0.84 | 0.83 | 0.83 | 27000 |

# Improvements in the model:

We could further improve this model by:

1. Either upsampling the smaller class.
2. Using additional quantitative metrics.
3. Using more algorithms to get an ensemble prediction score.



Thank you!

Code and Slides:
https://github.com/rabajaj0509/stanford-policing-dataset