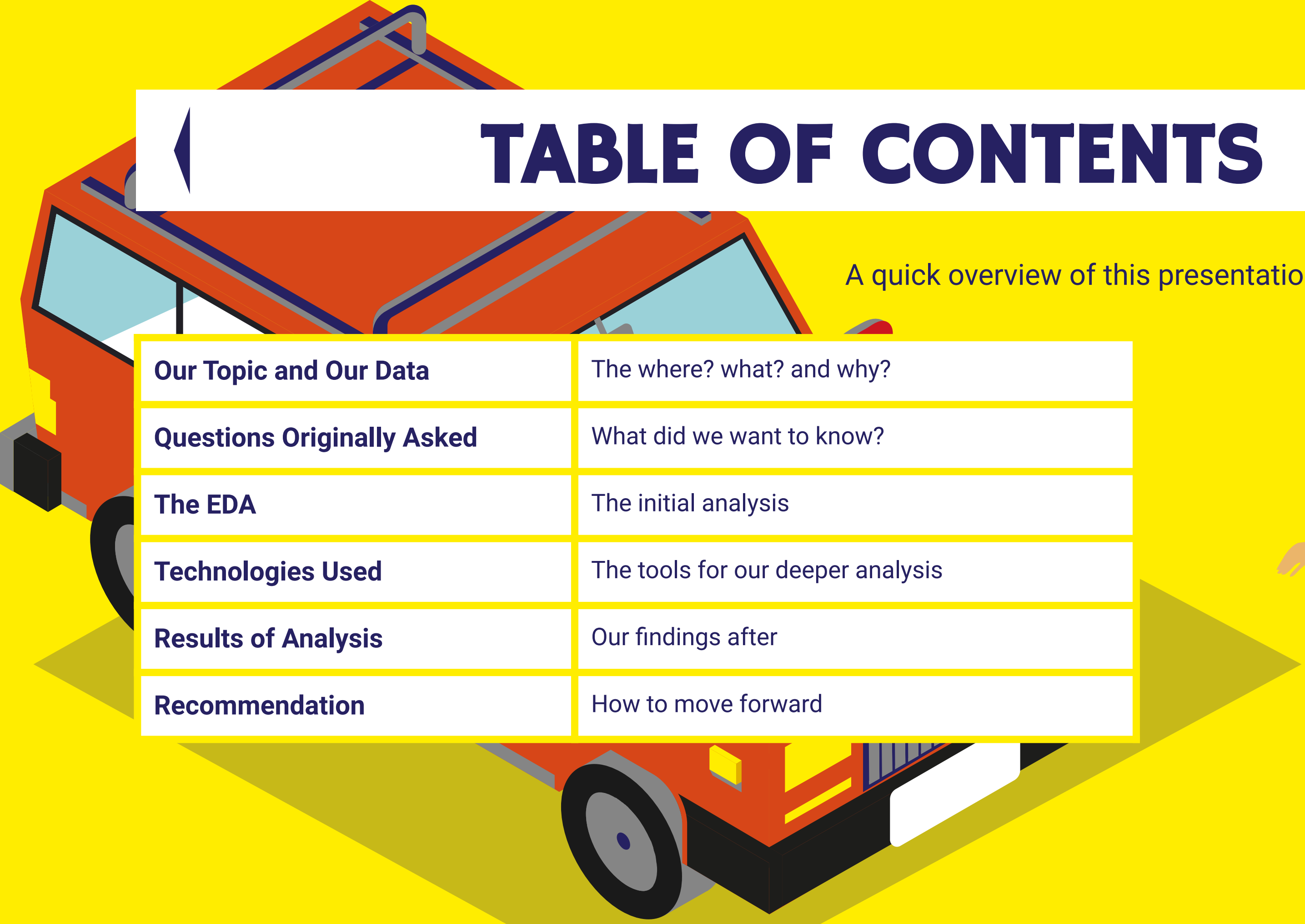
Analysis of

New York City Parking Violations!

By Jun Cho, Mykhal Young, Penwei Wang, Alexander Warshafsky

TABLE OF CONTENTS

A quick overview of this presentation



OUR DATA SOURCE

NYC OpenData

NYC OpenData

Home Data About Learn Alerts Contact Us Blog Sign In

Introducing our new data shaping and exploration experience: Filter, group, aggregate, and more! Try it now Learn more

Parking Violations Issued - Fiscal Year 2023

Parking Violations Issuance datasets contain violations issued during the respective fiscal year. The Issuance datasets are not updated to reflect violation

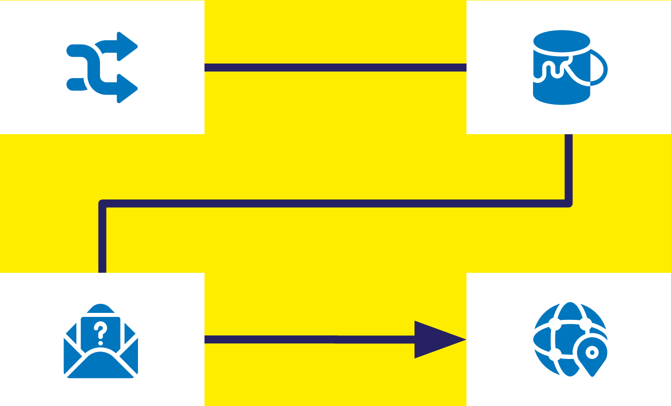
More Views Filter

Summons Number	Plate ID	Registration State	Plate Type	Issue Date	Violation Code	Vehicle Body Type
1484697303	JER1863	NY	PAS	06/10/2022	67	SDN
1484697315	KEV4487	NY	PAS	06/13/2022	51	SUBN
1484697625	H73NYD	NJ	PAS	06/19/2022	63	SDN
1484697674	GJC9296	NY	PAS	06/19/2022	63	SUBN
1484697686	M51PUV	NJ	PAS	06/19/2022	63	SDN
1484697698	H73NYD	NJ	PAS	06/23/2022	63	
1484697728	M51PUV	NJ	PAS	06/23/2022	63	
1484698204	KJJ8637	NY	PAS	06/20/2022	67	SUBN
1484698381	JEC8631	NY	PAS	06/19/2022	98	SUBN
1484698721	K21PNH	NJ	PAS	06/25/2022	10	SUBN
1484698769	LVK1404	PA	PAS	06/05/2022	10	SUBN
1484699683	KSX6366	FL	PAS	06/25/2022	51	SUBN
1484699750	GCX5397	NY	PAS	06/19/2023	63	SUBN
1484703261	KUH5328	NY	PAS	06/09/2022	45	SDN
1484710629	KUH1765	NY	OMS	06/30/2022	14	SDN
1484717909	HJJ9998	NY	PAS	06/10/2022	20	SUBN
1484720581	GHR524	FL	PAS	07/03/2022	68	VAN

QUESTIONS

What relationships are there between the type, color, or maker of a car with violations?

What color cars are more prone to violations (in comparison to the proportion of colors of cars)?



Can vehicle type predict type of violation?

Can machine learning detect these relationships?

OUR DATA

Initial Data

Parking_Violations_Raw	
Summons_Number	float
Plate_ID	varchar(10)
Registration_State	varchar(2)
Plate_Type	varchar(3)
Issue_Date	timestamp
Violation_Code	int
Vehicle_Body_Type	varchar(4)
Vehicle_Make	varchar(5)
Issuing_Agency	varchar(1)
Street_Code1	int
Street_Code2	int
Street_Code3	int
Vehicle_Expiration_Date	int
Violation_Location	int
Violation_Precint	int
Issuer_Precinct	int
Issuer_Code	int

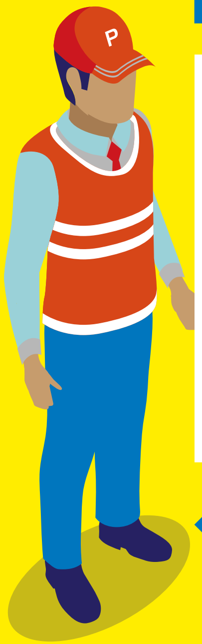
...

E.T.L.

Cleaned Data

Parking_Violations_Clean	
Registration_State	varchar(2)
Plate_Type	varchar(3)
Violation_Code	int
Vehicle_Body_Type	varchar(4)
Vehicle_Make	varchar(5)
Violation_Time	timestamp
Vehicle_Color	varchar(6)
Vehicle_Year	int

DATA CLEANING



- Identify relevant data that can be used in machine learning and discard all the rest.
- Transform data into useable forms.

Parking_Violations_Clean	
Registration_State	varchar(2)
Plate_Type	varchar(3)
Violation_Code	int
Vehicle_Body_Type	varchar(4)
Vehicle_Make	varchar(5)
Violation_Time	timestamp
Vehicle_Color	varchar(6)
Vehicle_Year	int

Parking_Violations_ML	
Registration_State_Group	int
Plate_Type_Group	int
Violation_Code	int
Vehicle_Body_Type_Group	int
Vehicle_Make_Group	int
Violation_Time	datetime
Vehicle_Color_Group	int
Vehicle_Year	int

Parking_Violations_Board	
Registration_State	varchar(2)
Plate_Type_Group_Name	varchar(3)
Violation_Code	int
Vehicle_Body_Type_Group_Name	varchar(4)
Vehicle_Make_Group_Name	varchar(5)
Vehicle_Expiration_Date	int
Violation_Location	int
Violation_Time	datetime
Street_Name	varchar(20)
Vehicle_Color_Group_Name	varchar(6)
Vehicle_Year	int

Postgres Exploration

Overall Violations

	count bigint	violation_code integer
1	3194638	36
2	810066	21
3	434559	38
4	367550	71
5	345813	7
6	315737	14
7	310018	5
8	277004	20
9	237735	40
10	216345	70

Colors

	count bigint	violation_code integer	vehicle_color character varying (3)
1	685661	36	WHT
2	160093	21	WHT
3	110786	38	WHT
4	86793	14	WHT
5	82210	20	WHT
6	82047	71	WHT
7	76118	7	WHT
8	69265	5	WHT
9	61307	69	WHT
10	52523	40	WHT

Manufacturer

	count bigint	violation_code integer	vehicle_make character vary
1	27904	14	FRUEH
2	16332	46	FRUEH
3	11905	20	FRUEH
4	10193	69	FRUEH
5	9624	19	FRUEH
6	7032	47	FRUEH
7	5828	38	FRUEH
8	5691	10	FRUEH
9	4261	48	FRUEH
10	3875	36	FRUEH

Data Analysis

R Studio	Logistic Regression	Chi Squared test	Code	Tableau
Was used to perform logistic regression and chi-squared analysis	Was used to see if there was a relationship between all of our categorical data	Color: significant p-value All Variables: significant p-value	The more variables I added to the Logistic Model the higher the r-squared value	Tableau

Data Analysis

```
car_sub <- subset(cars,Violation_Code=='36', select=c('Violation_Time','Vehicle_Make','Vehicle_Color','Violation_Code'))
car_sub <- subset(cleaned_data,Violation_Code=='36', select=c('Violation_Time','Vehicle_Make','Vehicle_Color','Violation_Code'))
car_subs <- subset(cars,Violation_Code=='46', select=c('Violation_Time','Vehicle_Make','Vehicle_Color','Violation_Code'))
summary(lm(Violation_Code ~ Violation_Time + Vehicle_Make + Vehicle_Color,data=cleaned_data))
save.image("C:/Users/nextg/Downloads/Group_Final_Project-main/Group_Final_Project-main/Notebooks/Final Project Data (Prelim Code).RData")
chi_table <-(cleaned_data$Violation_Code,cleaned_data$Vehicle_Color)
```

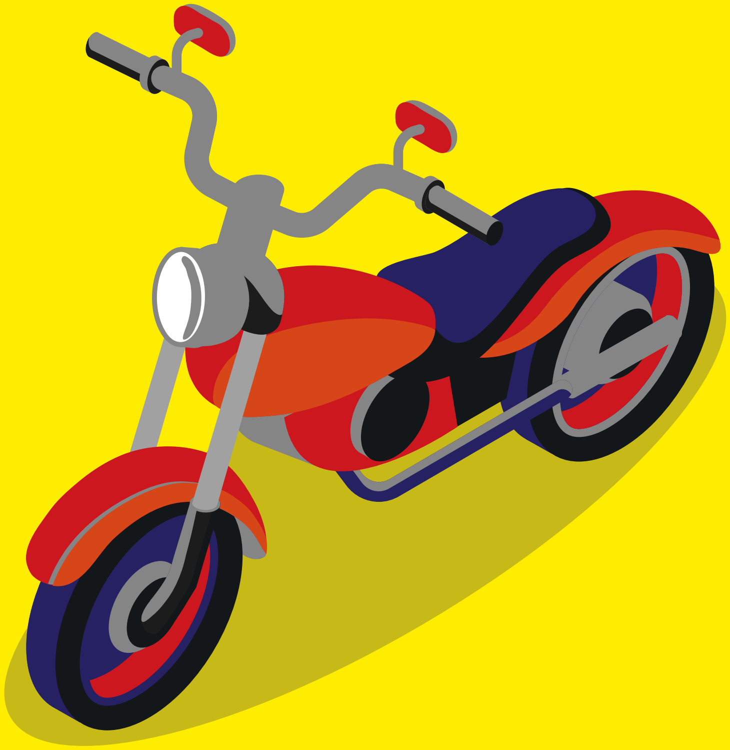
ENCODING FOR ML

	Registration State	Plate Type	Body Type	Make	Color
	Top Six: 'NY': 0 'FL': 1 'VA': 2 'GA': 3 'OH': 4 'NJ': 5	Top Eight: ‘PAS’: 0 ‘COM’: 1 ‘OMT’: 2 ‘SRF’: 3 ‘OMS’: 4 ‘APP’: 5 ‘ORG’: 6 ‘SPO’: 7	Top Eight: 'SUBN': 0 '4DSD': 1 VAN': 2 'PICK':3 'DELV':4 '2DSD':5 'REFG':6 'SDN':7	Top Eight: 'HONDA':0 'FORD':1 'TOYOT':2 'NISSA':3 'CHEVR':4 'ME/BE':5 'BMW':6 'JEEP':7 'FRUEH':8 'HYUND':9 'SUBAR':10 'LEXUS':11	Top Ten: 'RED': 1 'BLK': 2 'BLU': 3 'WHT': 4 'GRN': 5 'GRY': 6 'ORG' : 7 'BRN' : 8 'OTH' : 9

36

Speeding in School Zone

The largest number of violations



RESULTS

Logistic Regression

max_iter = 500
solver = ‘saga’

```
Accuracy: 0.5764
Precision: 0.4249
Recall: 0.0181
F1 Score: 0.0348
```

Random Forest

```
Accuracy: 0.5995
Precision: 0.5828
Recall: 0.1708
F1 Score: 0.2642
```

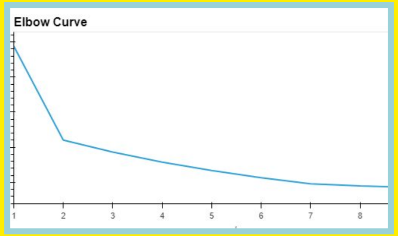
warm_start = true
n-estimators = 1000

Logistic Regression with Scaler

StandardScaler
max_iter = 500
solver = ‘saga’

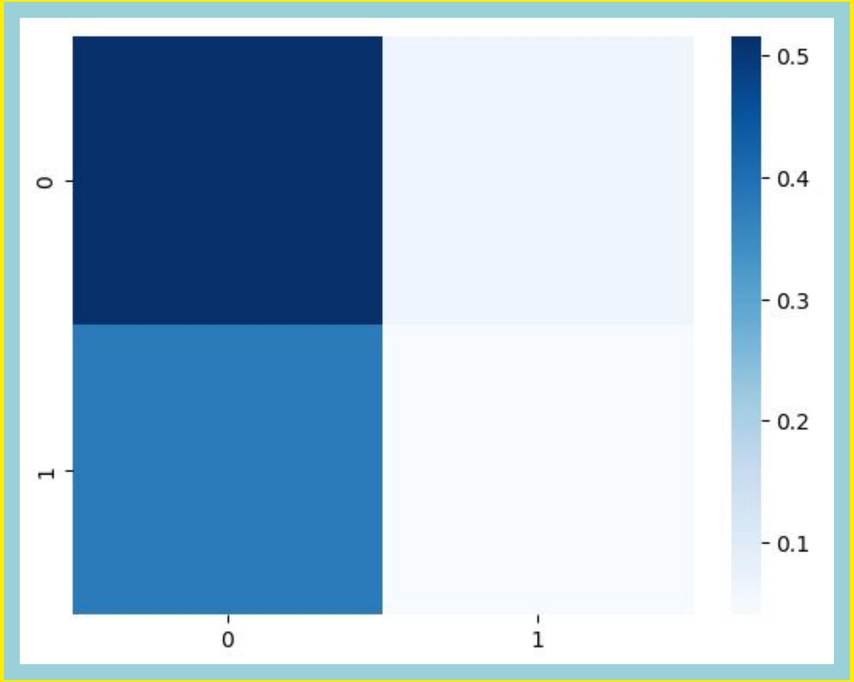
```
Accuracy: 0.5764
Precision: 0.4249
Recall: 0.0181
F1 Score: 0.0348
```

K-Means



n_clusters = 7

Random Forest Confusion Matrix



We see that the proportion of True Negative (0,0) is really high.

This means our model is good at predicting which car types **would not** be caught speeding in a school zone.

If only we had more time...

Bins
Make better categorical bins

Advice
Wished we talked with TAs more earlier

Moving Violations
More interesting, more human

Reduce Noise
Test models without some input vars.

Models
Make bigger Neural Network

Better Data
Could have found something cleaner