An Analysis of Crime in NYC

Ali Abbas Causer
Sprint 3 - November 2023

Agenda

Problem Statement

Data

Linear Regression

Classification

For Demo Day

Problem Statement

To understand the underlying features that predict criminal activity in NYC

- What demographic features (excluding race/ethnicity) are most heavily correlated with the type of crime?
    - Why is this useful: Can help provide local public organizations areas to improve (e.g. education, housing, jobs)
- How can the location, demographics and day of week help predict type of crime
    - Why is this useful: “Customer” segmentation to develop a targeted approach for action

Data

5.6M
Arrests

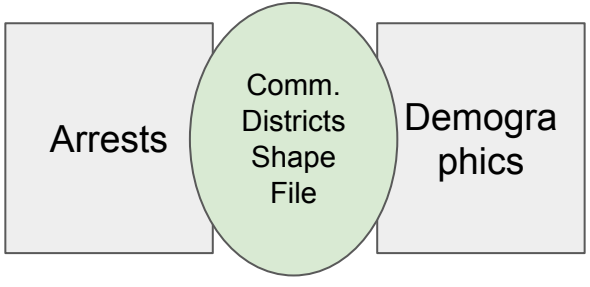
2006 - 2023
~1K / day

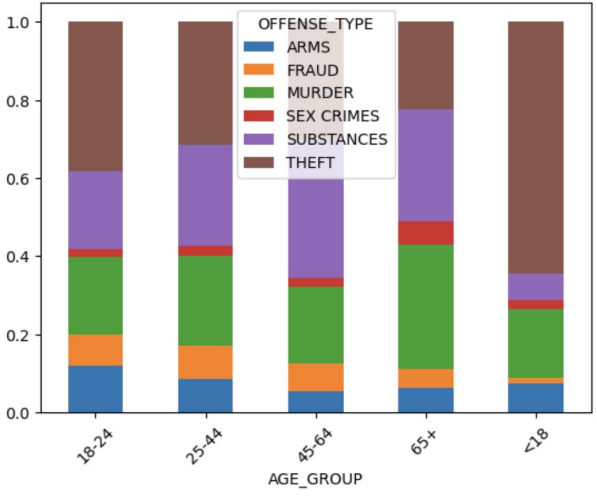
1.6M
Felonies

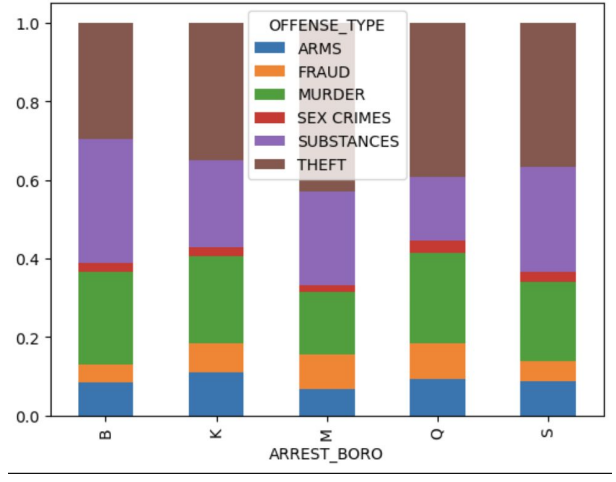
1.5M
Felonies
Non-Null

1.3M
Felonies
In material
categories

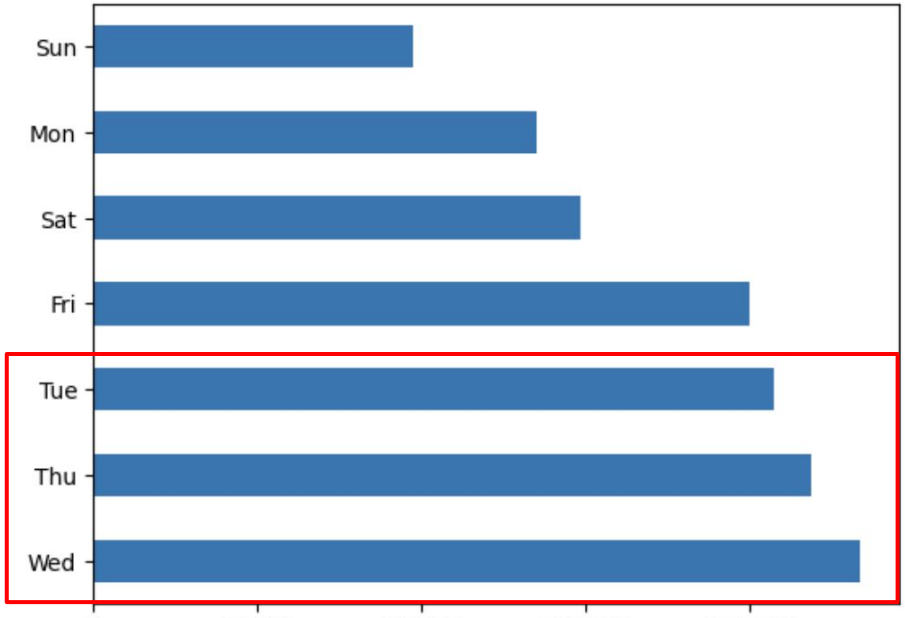
1.2M
Felonies
in 6 categories

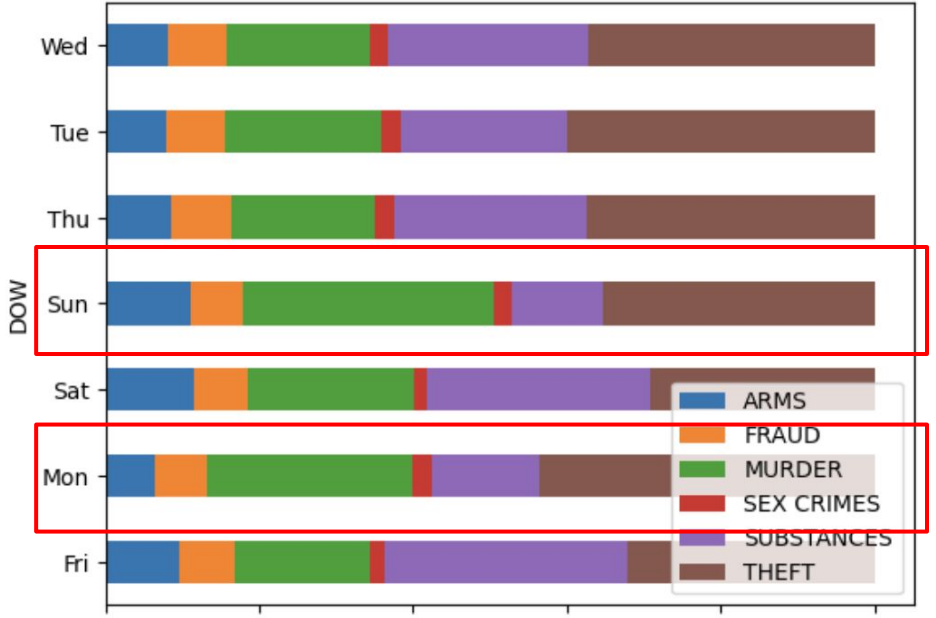






Crime by Day of Week





OLS Regression - crime volume vs. crime rate

OLS Regression Results

Dep. Variable:	OFFENSE_TYPE	**R-squared:**	0.525
Model:	OLS	**Adj. R-squared:**	0.438
Method:	Least Squares	**F-statistic:**	6.022
Date:	Mon, 27 Nov 2023	**Prob (F-statistic):**	1.20e-05
Time:	08:01:49	**Log-Likelihood:**	-608.96
No. Observations:	59	**AIC:**	1238.
Df Residuals:	49	**BIC:**	1259.
Df Model:	9		
Covariance Type:	nonrobust		

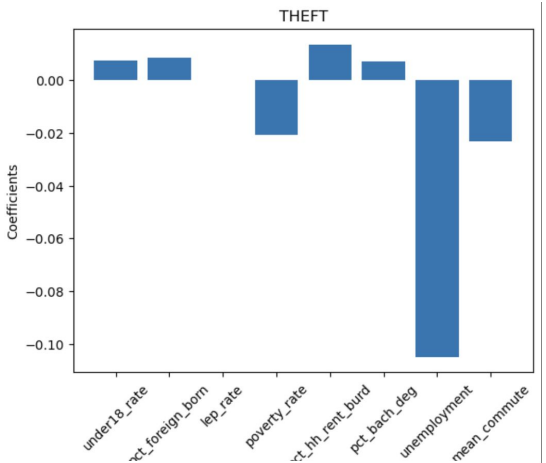
	coef	std err	t	P>|t|	[0.025	0.975]
const	7.678e+04	2.56e+04	2.994	0.004	2.52e+04	1.28e+05
pop_change_00_10	3542.6949	1.11e+04	0.318	0.752	-1.89e+04	2.59e+04
under18_rate	-322.5726	421.557	-0.765	0.448	-1169.723	524.578
pct_foreign_born	276.5874	199.099	1.389	0.171	-123.518	676.693
lep_rate	-448.1322	164.948	-2.717	0.009	-779.608	-116.657
poverty_rate	550.7711	374.988	1.469	0.148	-202.796	1304.338
pct_hh_rent_burd	-255.8517	352.642	-0.726	0.472	-964.511	452.808
pct_bach_deg	-369.4994	161.757	-2.284	0.027	-694.563	-44.436
unemployment	2411.8414	1296.848	1.860	0.069	-194.272	5017.955
mean_commute	-1109.5730	371.572	-2.986	0.004	-1856.274	-362.872

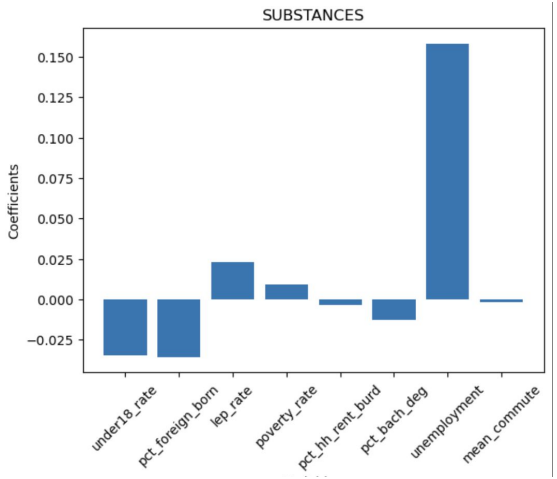
OLS Regression Results

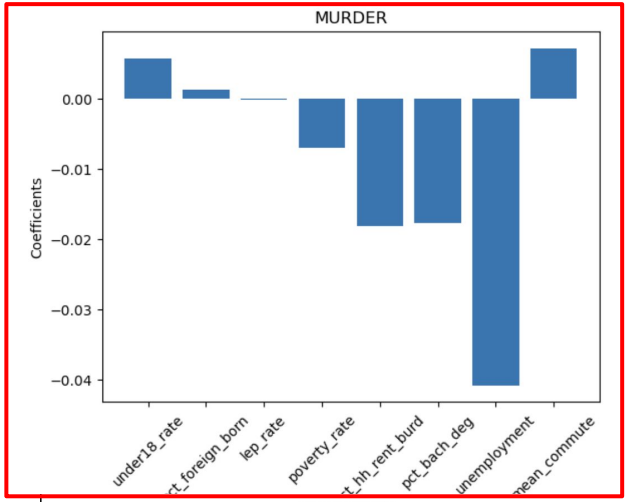
Dep. Variable:	OFFENSE_RATE	**R-squared:**	0.637
Model:	OLS	**Adj. R-squared:**	0.579
Method:	Least Squares	**F-statistic:**	10.98
Date:	Mon, 27 Nov 2023	**Prob (F-statistic):**	8.86e-09
Time:	09:59:05	**Log-Likelihood:**	87.447
No. Observations:	59	**AIC:**	-156.9
Df Residuals:	50	**BIC:**	-138.2
Df Model:	8		
Covariance Type:	nonrobust		

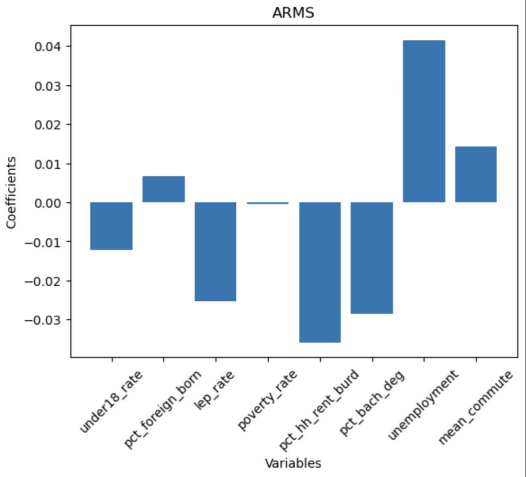
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5269	0.186	2.840	0.007	0.154	0.899
under18_rate	-0.0023	0.003	-0.776	0.441	-0.008	0.004
pct_foreign_born	0.0014	0.001	0.940	0.352	-0.002	0.004
lep_rate	-0.0035	0.001	-2.898	0.006	-0.006	-0.001
poverty_rate	0.0078	0.003	2.808	0.007	0.002	0.013
pct_hh_rent_burd	-0.0015	0.003	-0.566	0.574	-0.007	0.004
pct_bach_deg	-0.0015	0.001	-1.263	0.212	-0.004	0.001
unemployment	0.0183	0.010	1.913	0.061	-0.001	0.038
mean_commute	-0.0102	0.003	-3.954	0.000	-0.015	-0.005

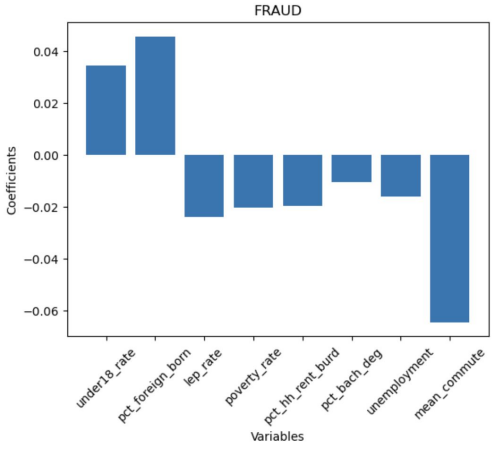
Regression Output

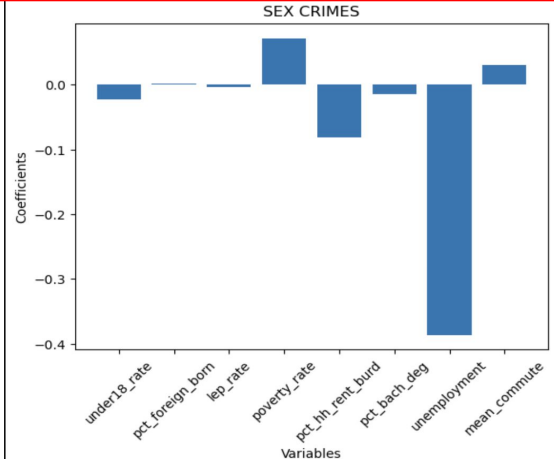




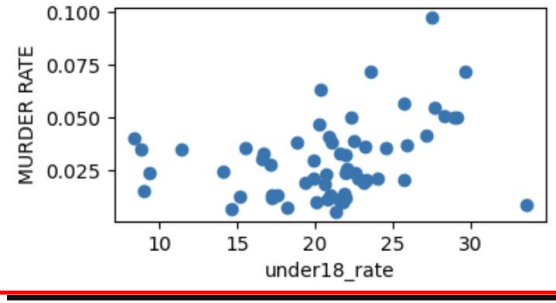


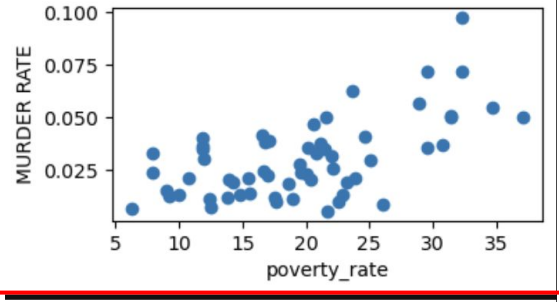


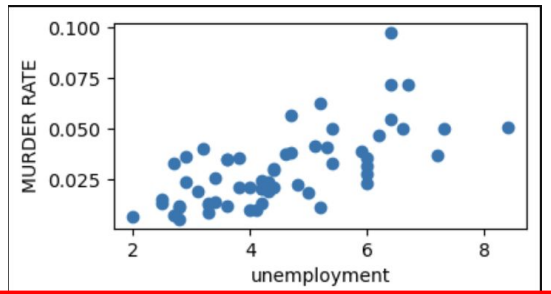


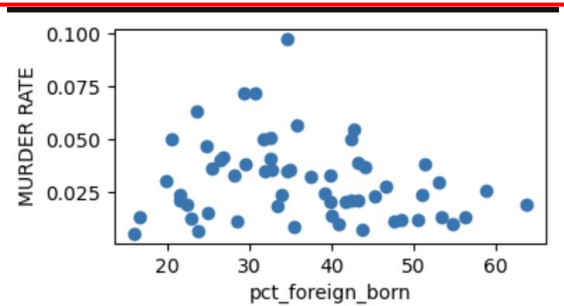


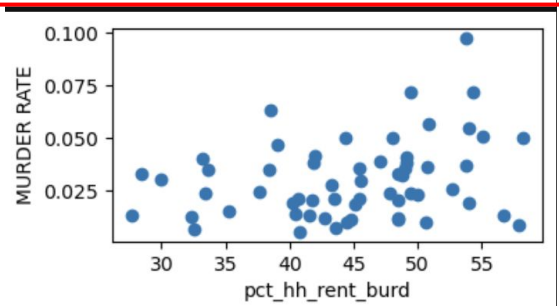
Relationship between independent variables and murder rate

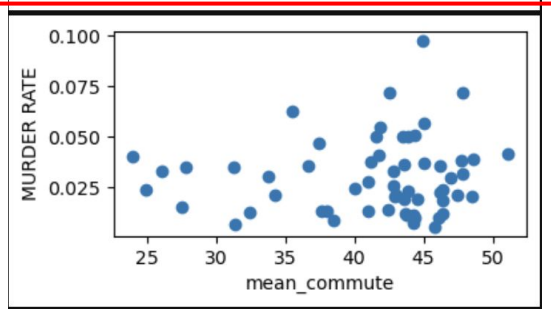


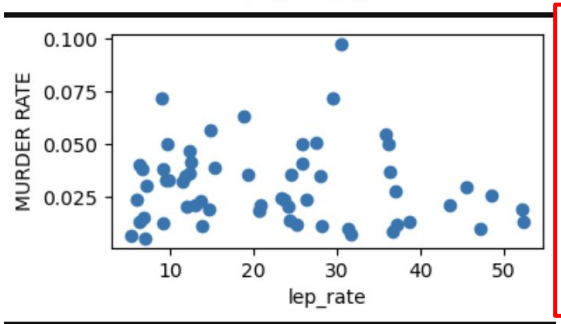


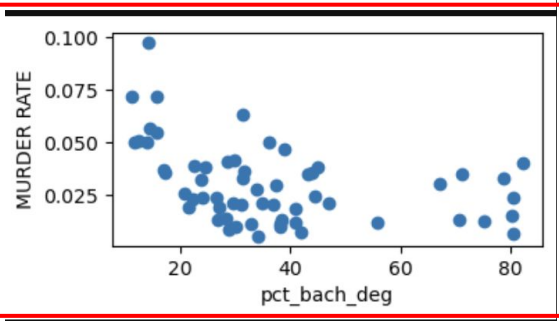


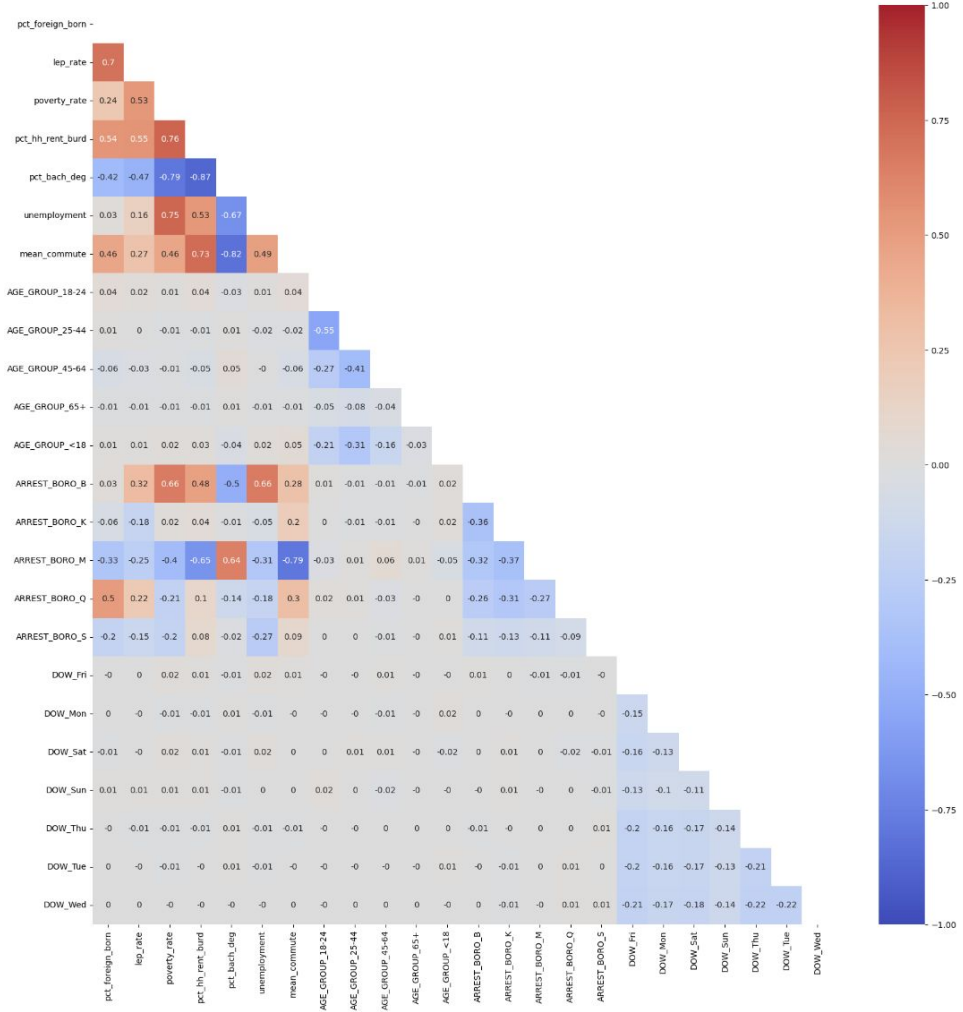










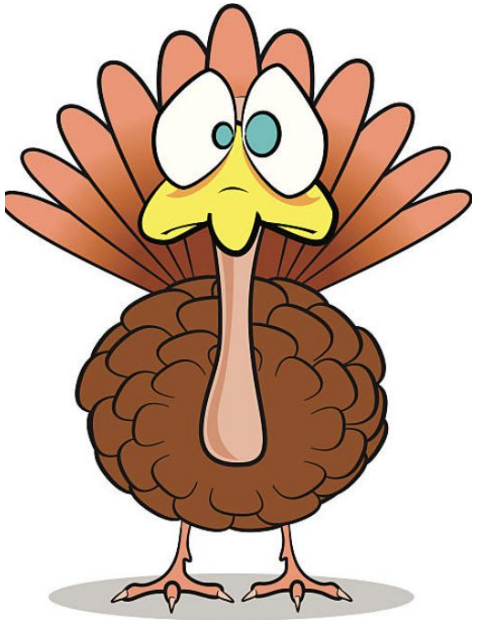


- As expected, rent burden, poverty rate and unemployment have high multicollinearity
- Mean commute has high negative collinearity with percentage bachelor degree

Classification

Independent Variables

- Age Group
- Borough
- Day of Week
- Poverty Rate
- Unemployment Rate
- HH Rent Burden
- Pct Bachelor Degree



	Sprint 3		Sprint 2		Improvement	
Accuracy	Logistic Regression	Decision Tree	Logistic Regression	Decision Tree	Logistic Regression	Decision Tree
Train	42.3	41.1	31.0	30.0	+11.3	+10.1
Test	42.5	41.1	30.2	29.8	+12.3	+11.3

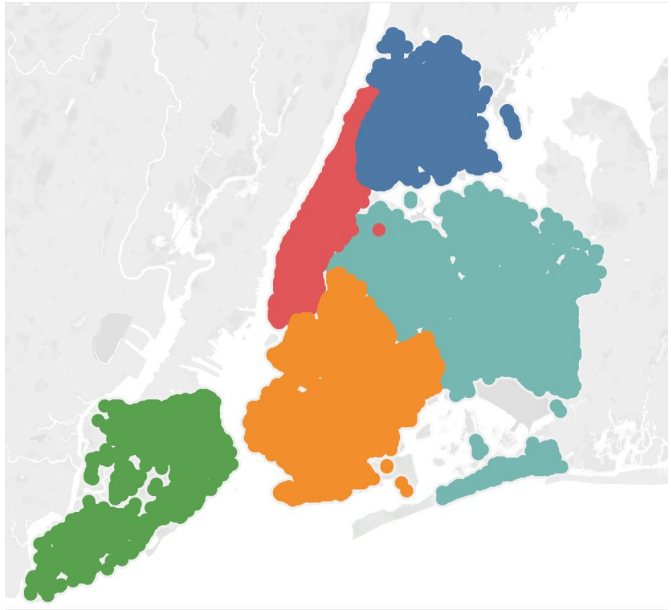
For Demo Day

Tool to predict type of crime

Day of Week | Borough | Poverty Rate | Pct Bachelor | Rent Burden

LIKELY CRIME

Interactive map for crime type



Closing Thoughts

With more time I would:

- Plot out decision tree
- Identify additional attributes that would move the needle on model accuracy and add precision and recall
- Run grid search to identify best model type and hyperparameters

Thoughts:

- Identify intersection in your data where there is most variability
- Be ethical - frame your question clearly and contextualize model outputs, this is where human interpretation is critical before making decisions