Decoding Safety: Investigating the Role of Predictive Policing in Crime Prevention Across New York City

Sandya Wijaya, Marianne Choi, Richard Soong, Ines Yang

Data 102 Spring 2024: Final Project (Group 62)

1 Introduction

In recent history, predictive policing practices have become more pervasive. Predictive policing is often characterized as a process in which police authorities analyze data trends about crimes in efforts to identify and prepare for future crimes (National Institute of Justice, 2019). Several urban cities throughout the U.S. have adopted such practices, hoping to minimize the prevalence of crime. Predictive policing practices have had varying degrees of success. There have been instances of police forces hyper-patrolling areas for no apparent reason, spreading paranoia amongst the general population and antagonizing various communities or individuals. Due to these effects and controversies, through our project, we hope to explore the effect of predictive policing on crime severity and characteristics about those crimes.

2 Data Overview

Our Kaggle dataset is based in NYC (Morjaria, 2022). After some further research, we discovered that NYC implemented predictive policing in 2013 (Lau, 2016). Our dataset spans 2006-2017, providing substantial data from both before and after the implementation of predictive policing.

Our dataset is a census with each row representing a specific reported crime incident in NYC. Since we lack direct contact with the NYPD, we cannot conclude if the people represented in our dataset might have been aware of the collection and usage of their data. For that same reasoning, we also cannot conclude if any groups of people were systematically left out of our data. We believe that there isn't necessarily selection bias since the dataset was queried almost word-for-word from the NYPD. This logic also rules out the use of convenience sampling. We figured that measurement error could have occurred if the officers who reported the crimes had made clerical mistakes, but we lack such information.

Our dataset presents around 650,000 uncleaned rows, each representing a specific crime incident. All the metadata are present in the kaggle documentation. Our dataset was not modified for differential privacy. Some features that we wish we had access to include economic indicators like suspect's income rate or # of properties owned by the victim. Having these additional variables would have allowed us to explore several questions and inquiries. For example, we could have analyzed how different people's economic situations affected their likelihood of committing a crime before and after the implementation of predictive policing.

2.1 Data cleaning

We first cleaned our data by dropping columns with >95% missing values as they will have little to no effect on our analysis. This included {'LOC_OF_OCCUR_DESC', 'SUSP_SEX', 'SUSP_AGE_GROUP', 'TRANSIT_DISTRICT', 'CMPLNT_TO_TM', 'HOUSING_PSA', 'VIC_AGE_GROUP', 'SUSP_RACE', 'PARKS_NM', 'STATION_NAME', 'HADEVELOPT', 'CMPLNT_TO_DT'}.

Most of these are related to the location of a crime, and a potential reason why it has a high proportion of missing values is because they are simply not applicable – for example, 'HADEVELOPT' contains the "name of NYCHA housing development of occurrence *if applicable*," but it is possible crimes simply did not occur in or near a NYCHA housing development and thus will have a missing value.

Then, we imputed any numeric columns' missing values with that column's median, which is less affected by extreme outlier values than the mean. Though imputation will slightly distort the distribution of the data, it is overall a robust method. Finally, we dropped rows with missing values in non-numeric columns as there is no objective way to impute missing categorical values such as race.

Prior to cleaning, our dataset had 6521496 rows and 36 columns. After, it had 6436152 rows and 24 columns, so this slight decrease will have minimal effect over our analysis.

3 Research Questions

Our causal inference question is "Does predictive policing have a significant impact on crime severity?" If we know there is a causal relationship, this could guide resource allocation strategies – for example, acting as justification for increasing or decreasing investment in predictive policing. To examine the varying degrees of crime severity (violations, misdemeanors, and felonies), we chose to use causal inference as it enables us to isolate the impact of predictive policing from confounders, thus providing more reliable evidence. However, one concern is that causal inference methods rely heavily on the quality and representativeness of the data, so inaccuracies or biases in the dataset could lead to erroneous conclusions.

Our multiple hypothesis testing question is "Is there a significant association between predictive policing and reported crimes features in NYC?" Understanding the associations between predictive policing and reported crime features is crucial for evaluating the potential biases of predictive policing practices. We wanted to look at various aspects – such as race, sex, and age of suspects and victims, types of crime, number of complaints – so we used multiple hypothesis testing to get a more holistic answer. One concern is the increased risk of false positive findings when conducting numerous statistical tests simultaneously.

4 Exploratory Data Analysis

To investigate if the number of offenses differ for different areas, we created a bar graph (Figure 1) of the number of offenses across different boroughs by grouping the dataset by

‘PATROL_BORO’. BRONX has the highest number of complaints with 155395, STATEN ISLAND with lowest at 34945 – nearly 1/5 of the highest.

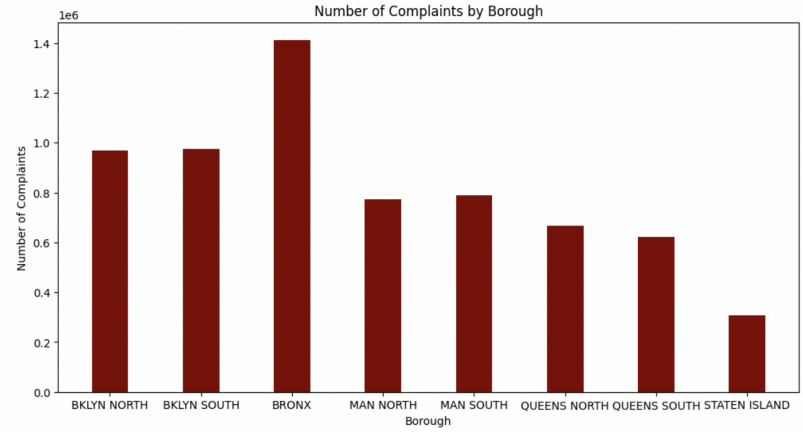


Figure 1: Number of offenses by borough.

We also created a heatmap of the number of offenses across all of NYC using the ‘Latitude’ and ‘Longitude’ column (Figure 2). We see the highest number seen in red around Penn Station in Manhattan in the 34th street area. The rest of NYC has roughly the same level of complaints, seen in green. This further confirms that there is an association between number of crimes and location.

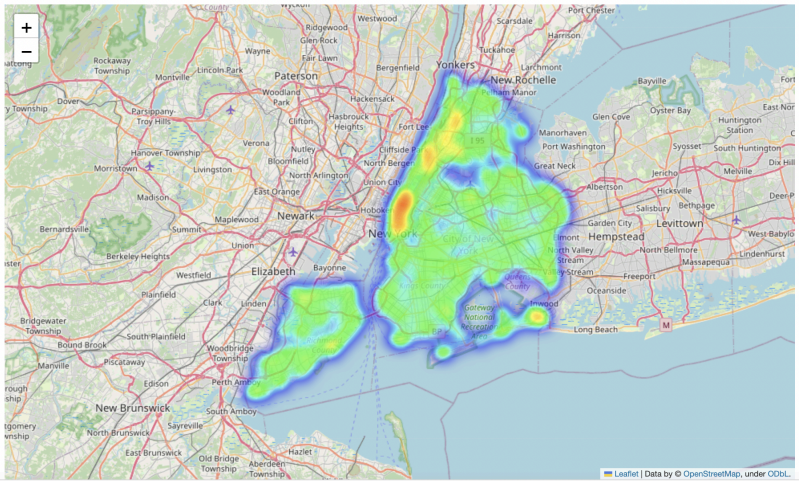


Figure 2: Heatmap of number of offenses across NYC.

The histogram (Figure 3) on the right shows the distribution of the suspect race, while the histogram on the left shows the distribution of the victim race. The most prominent findings from this visualization was that asian / pacific islander victim distribution is much greater than its suspect distribution. However, much of the victim and suspect race is unknown, so it is hard to tell if this histogram shows meaningful data just by visualization alone.

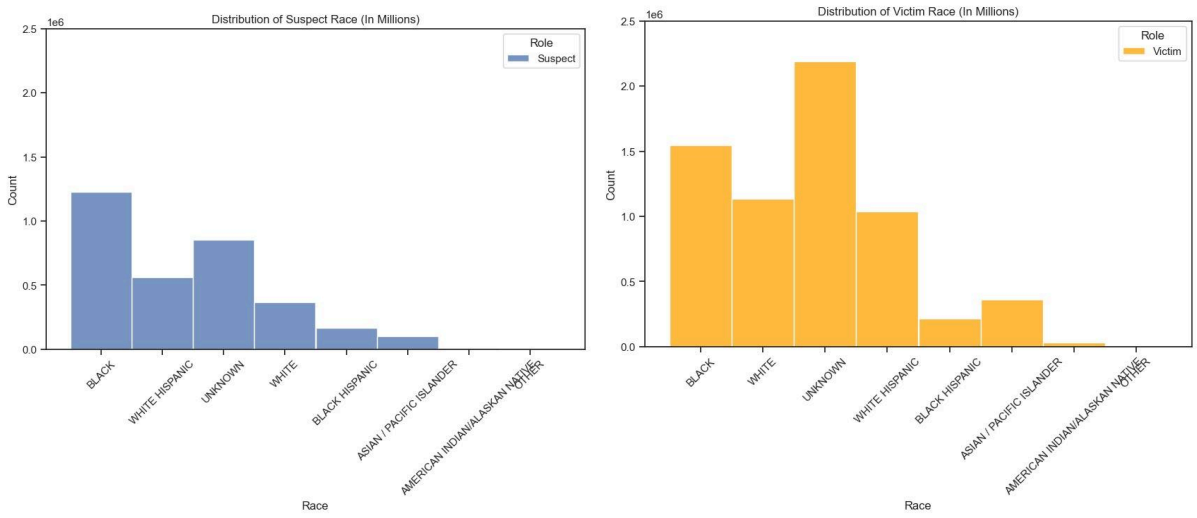


Figure 3: Histogram of the suspect and victim race reported to the NYPD.

5 Research Question #1: Causal Inference

5.1 Methods

5.1.1 Problem setup

The binary treatment Z in this experiment is whether predictive policing is implemented or not. NYC started implementing predictive policing in 2013. We created a binary treatment column (‘treatment’) to identify which rows belonged to which group, taking inspiration from the date column and assigning 0 values to years before 2013 (before the implementation of predictive policing) and 1 values to years on or after 2013.

The outcome Y is the column 'LAW_CAT_CD,’ which details the level of the offense. Every crime is categorized as felony, misdemeanor, or violation Although ‘LAW_CAT_CD’ was originally a categorical variable, we converted it into numerical format to satisfy the constraints of our frequentist models. We then thought about using all of our numerical variables as explanatory variables, though we remembered to account for confounding variables.

We find potential confounders X by assessing which of the columns in our dataset have statistically different distributions in the treatment and non-treatment control group. This process of finding potential confounders is detailed in the following section.

We created a causal DAG by hand-drawing it (Figure 4). As seen in the causal DAG, the confounders X have arrows pointing to both outcome Y and binary treatment Z. Within X, there are arrows between x-coordinate and Longitude, as well as y-coordinate and Latitude as they can be converted to and from each other (Geographic Coordinate Systems, 2010). There are no colliders that we have to ignore.

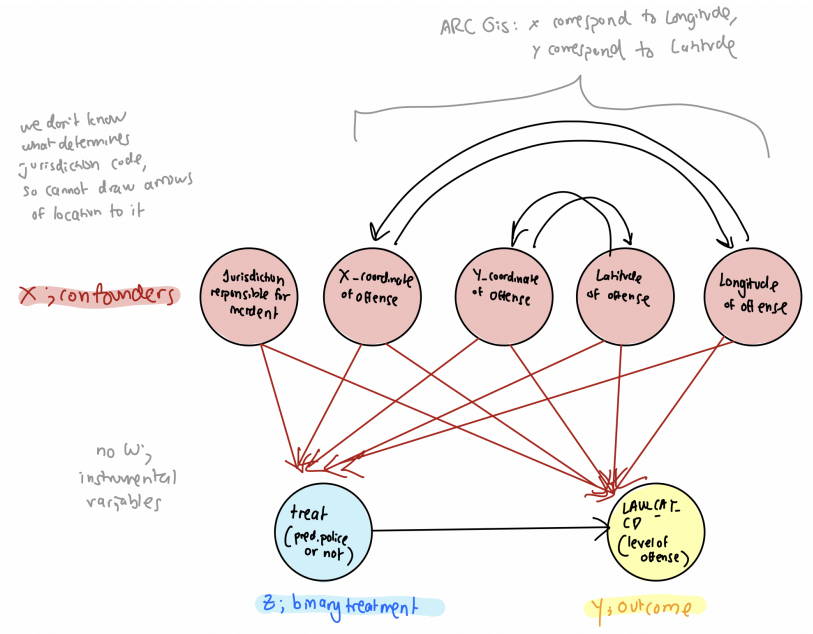
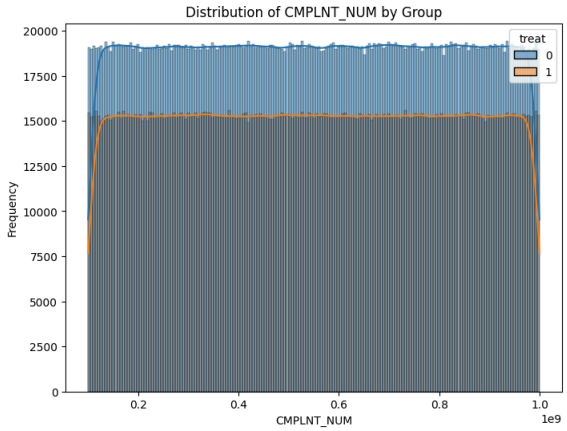


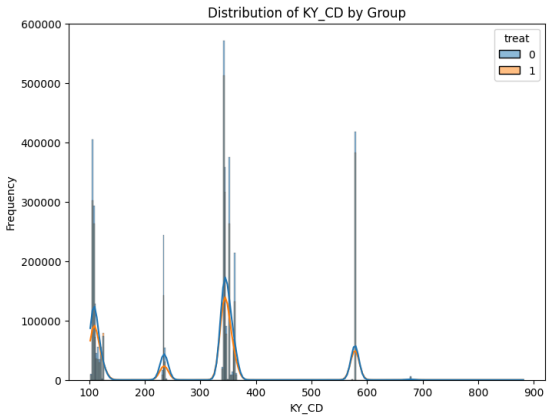
Figure 4: Casual DAG.

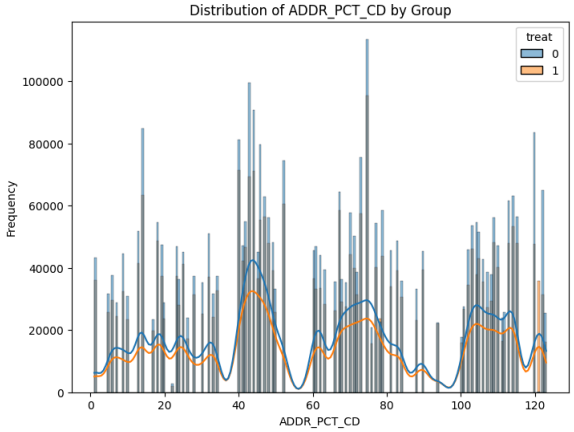
5.1.2 Identifying and evaluating possible confounders

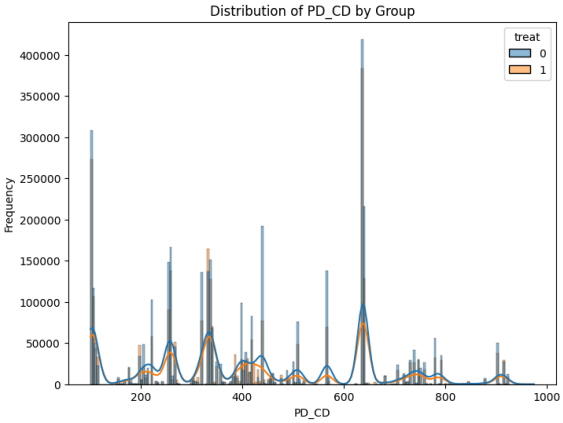
A variable is a confounder if it is correlated with both the treatment (predictive policing) and also the outcome (level of offense). Confounders will have statistically significant differences in distribution between the treatment and control group. To confirm whether this is the case, the variable has to be a quantitative variable so it can be visualized and run tests on. Out of the 24 columns we had, only 9 were numeric – CMPLNT_NUM, KY_CD, ADDR_PCT_CD, PD_CD, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude.

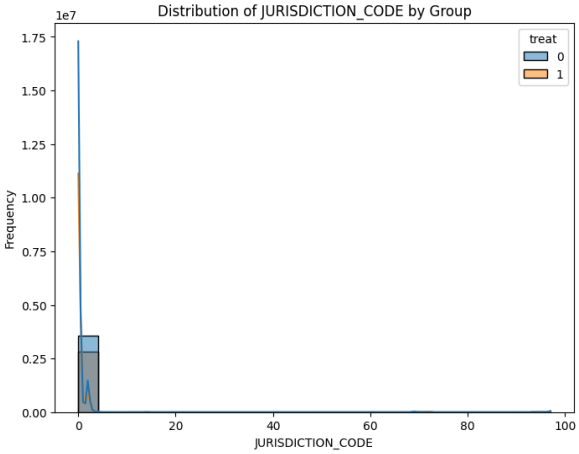
Noticeable differences in the histograms (Figure 5) between the two groups suggest that the variable may be associated with both the treatment and the outcome – and the variable may be a confounding variable. Some aspects to consider when comparing distributions are range and spread, height and density, shape and distribution, and alignment.

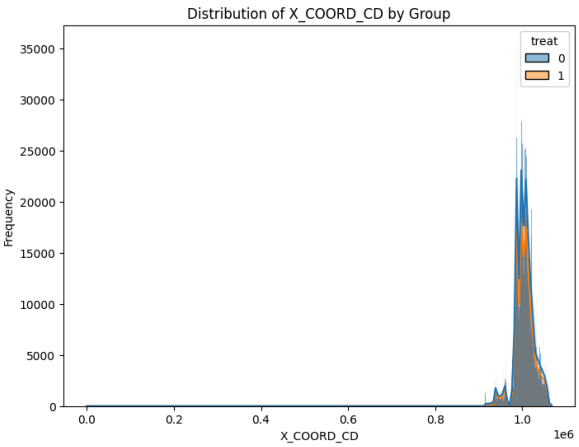

Distribution of CMPLNT_NUM by Group
treat
0
1
Frequency
CMPLNT_NUM
1e9


Distribution of KY_CD by Group
treat
0
1
Frequency
KY_CD


Distribution of ADDR_PCT_CD by Group
treat
0
1
Frequency
ADDR_PCT_CD


Distribution of PD_CD by Group
treat
0
1
Frequency
PD_CD


Distribution of JURISDICTION_CODE by Group
1e7
treat
0
1
Frequency
JURISDICTION_CODE


Distribution of X_COORD_CD by Group
treat
0
1
Frequency
X_COORD_CD
1e6

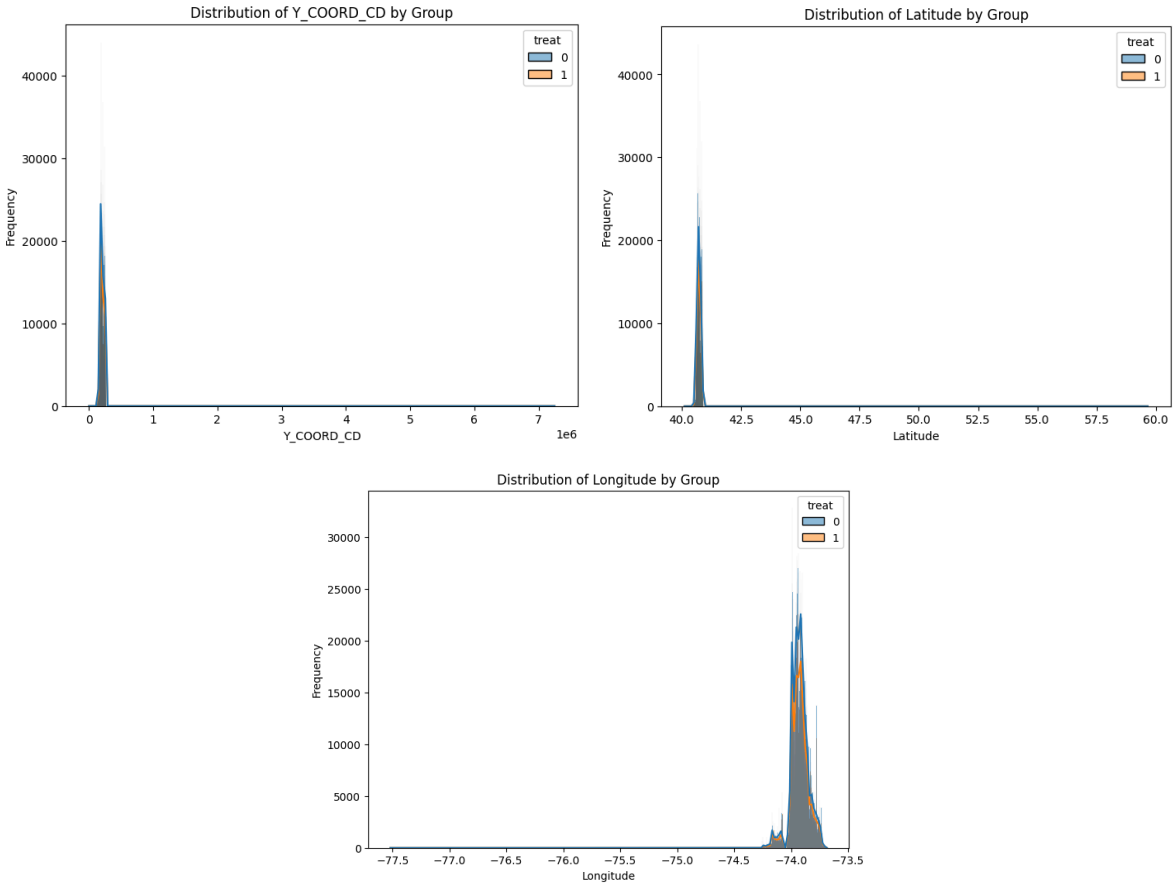


Figure 5: Distribution of each numeric variable for the treatment and non-treatment group.

We then used the Mann-Whitney U test to assess whether two independent samples come from the same distribution. If the p-value is less than my chosen significance level of 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference between the groups. This acts as a quantitative way to help confirm whether the observed differences between the treatment and control groups are statistically meaningful – and in effect determine if it is a possible confounder. The results of these tests can be found in Appendix. With this, we identified the following possible confounders from Kaggle (Figure 6):

# KY_CD	# PD_CD	# JURISDICTION_CODE
Three digit offense classification code	Three digit internal classification code (more granular than Key Code)	Jurisdiction responsible for incident. Either internal, like Police(0), Transit(1), and Housing(2); or external(3),

# X_COORD_CD	# Y_COORD_CD	Latitude	Longitude
X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet	Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees	Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees

Figure 6: Metadata of possible confounders.

We decided to remove KY_CD and PD_CD as confounders because they give more granular information than our target variable (e.g. PD_CD labels a crime as harassment, which is then classified as a violation in target column LAT_CAT_CD). Including these as a feature in our causal analysis may introduce multicollinearity and undermine our results, and/or also cause overfitting and bias in our model.

5.1.3 Associations and unconfoundedness

We conducted a chi-square test which yielded a chi-square value of 1571.5166146758065 with a p-value of 0.0. The magnitude of the chi-square statistic indicates a robust association between the treatment and outcome variables. The minuscule p-value of 0.0 suggests that this association is highly unlikely to be attributable to chance alone. These findings affirm that there is a statistically significant association between the implementation of predictive policing and the level of criminal offenses (misdemeanor, violation, felony) in NYC.

Now that we have set a strong foundation and reasoning for our study, we can utilize various unconfoundedness techniques to estimate the treatment effect. The unconfoundedness assumption means we observe all the relevant confounding variables i.e. there are no unobserved confounders. We first tried outcome regression, then inverse propensity weighting. Since we successfully obtained results from IPW, there is no need to execute the matching process which would have taken more time and given us less significant results.

5.1.4 Outcome Regression - Ordinary Least Squares

We implemented an OLS regression to see if there was a linear relationship between crime severity and predictive policing. For the outcome variable, we converted LAW_CAT_CD into numerical format. We arbitrarily classified violations with a value of 1, misdemeanors with a value of 2, and felonies with a value of 3. We converted LAW_CAT_CD into this numerical format to fit the constraints of OLS regressions as they have trouble dealing with categorical variables.

OLS Regression Results

Dep. Variable:	lawcatcd_numerical	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	2061.
Date:	Sun, 05 May 2024	Prob (F-statistic):	0.00
Time:	00:38:48	Log-Likelihood:	-6.1925e+06
No. Observations:	6436152	AIC:	1.239e+07
Df Residuals:	6436146	BIC:	1.239e+07
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
treatment	-0.0148	0.001	-29.389	0.000	-0.016	-0.014
JURISDICTION_CODE	0.0026	3.79e-05	69.808	0.000	0.003	0.003
X_COORD_CD	2.019e-05	1.15e-06	17.595	0.000	1.79e-05	2.24e-05
Y_COORD_CD	2.816e-05	1.69e-06	16.705	0.000	2.49e-05	3.15e-05
Latitude	-10.3539	0.614	-16.850	0.000	-11.558	-9.150
Longitude	-5.3815	0.318	-16.909	0.000	-6.005	-4.758

Figure 7: Ordinary least squares regression results summary.

From our OLS summary table (Figure 7), the only sizeable effects on LAW_CAT_CD are with latitude and longitude, while the other four variables (treatment, JURISDICTION_CODE, X_COORD_CD, and Y_COORD_CD) all seem to have miniscule effects. Additionally, the log-likelihood is very low and the AIC is very high, implying that the model might not be the best fit.

In addition to our OLS summary table, we measured the average treatment effect of predictive policing on crime severity by visualizing 100 of our bootstrapped estimates of ATE. Our 95% confidence interval ranges from -0.0 to -0 (Figure 8).

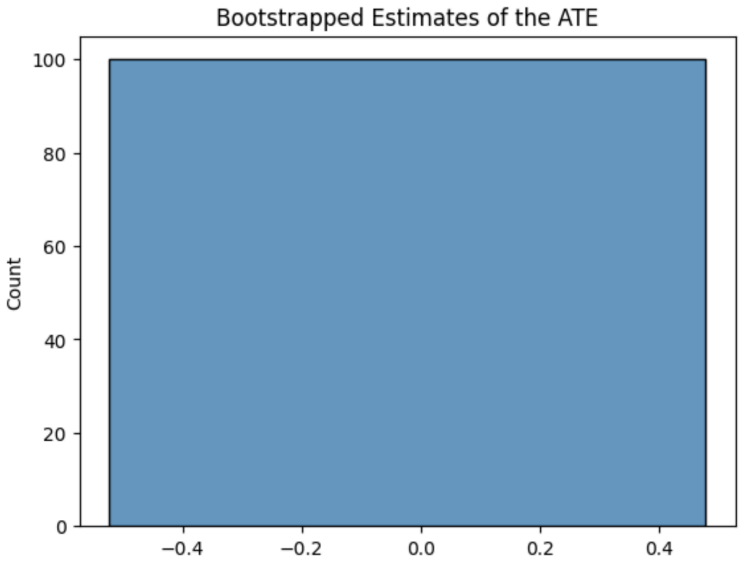


Figure 8: Distribution of bootstrapped estimates of the average treatment effect.

5.1.5 Outcome Regression - Logistic (Revisited)

Since assigning LAW_CAT_CD 1, 2, and 3 values in the aforementioned OLS regression provided suboptimal results, we decided to instead create 3 separate logistic regression models as a better way to model crime severity. We created 3 new columns stemming from LAW_CAT_CD, one illustrating if the crime was a violation (0, 1 values), another illustrating if the crime was a misdemeanor, and another illustrating the crime was a felony. Each of these new columns served as an outcome variable for each of the 3 logistic regressions.

5.1.6 Inverse Propensity Weighting

To apply inverse propensity weighting, we began by transforming the categorical variable 'LAW_CAT_CD' into a numerical format by assigning 1, 2, 3 to misdemeanors, and felonies, respectively. We then calculated the propensity scores by using logistic regression to predict treatment assignment based on covariates. The covariates included 'JURISDICTION_CODE', 'X_COORD_CD', 'Y_COORD_CD', 'Latitude', 'Longitude', while 'treat' represented the presence or absence of predictive policing implementation.

Histograms of propensity scores were then plotted separately for treated and untreated groups to visually inspect the distribution of propensity scores with and without predictive policing. The IPW estimate was calculated as the difference between the average propensity score-weighted 'LAW_CAT_CD' values for treated and untreated groups. This procedure yielded an IPW estimate capturing the causal effect of predictive policing implementation on criminal offense categories, adjusted for confounding covariates.

5.2 Results

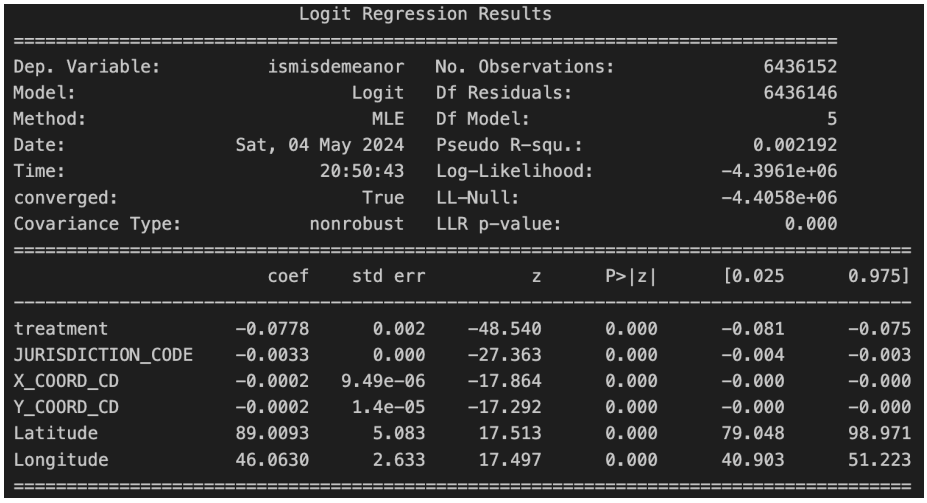
After creating the 3 logistic regression summary tables, we got the following results:

Logit Regression Results

Dep. Variable:	isviolation	No. Observations:	6436152
Model:	Logit	Df Residuals:	6436146
Method:	MLE	Df Model:	5
Date:	Sat, 04 May 2024	Pseudo R-squ.:	0.002021
Time:	20:48:36	Log-Likelihood:	-2.4417e+06
converged:	False	LL-Null:	-2.4466e+06
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
treatment	0.1522	0.002	64.089	0.000	0.148	0.157
JURISDICTION_CODE	-0.0122	0.000	-44.059	0.000	-0.013	-0.012
X_COORD_CD	-0.0002	4.33e-05	-3.826	0.000	-0.000	-8.09e-05
Y_COORD_CD	-0.0002	6.38e-05	-3.865	0.000	-0.000	-0.000
Latitude	89.0148	23.239	3.830	0.000	43.467	134.563
Longitude	46.1312	12.037	3.832	0.000	22.538	69.724

Figure 9: Logistic regression results summary. **Modeling treatment against if a crime is a violation.**

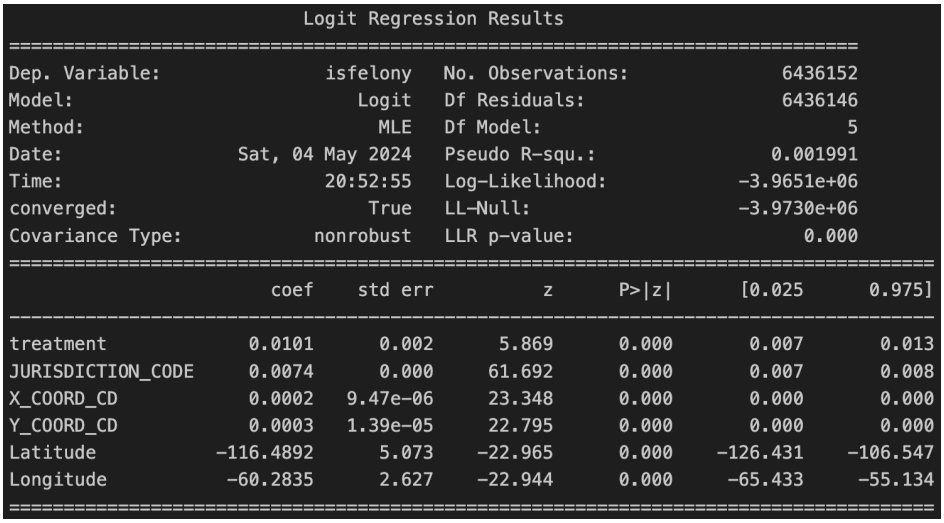


Logit Regression Results

Dep. Variable:	ismisdemeanor	No. Observations:	6436152
Model:	Logit	Df Residuals:	6436146
Method:	MLE	Df Model:	5
Date:	Sat, 04 May 2024	Pseudo R-squ.:	0.002192
Time:	20:50:43	Log-Likelihood:	-4.3961e+06
converged:	True	LL-Null:	-4.4058e+06
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
treatment	-0.0778	0.002	-48.540	0.000	-0.081	-0.075
JURISDICTION_CODE	-0.0033	0.000	-27.363	0.000	-0.004	-0.003
X_COORD_CD	-0.0002	9.49e-06	-17.864	0.000	-0.000	-0.000
Y_COORD_CD	-0.0002	1.4e-05	-17.292	0.000	-0.000	-0.000
Latitude	89.0093	5.083	17.513	0.000	79.048	98.971
Longitude	46.0630	2.633	17.497	0.000	40.903	51.223

Figure 10: Logistic regression results summary. **Modeling treatment against if a crime is a misdemeanor.**



Logit Regression Results

Dep. Variable:	isfelony	No. Observations:	6436152
Model:	Logit	Df Residuals:	6436146
Method:	MLE	Df Model:	5
Date:	Sat, 04 May 2024	Pseudo R-squ.:	0.001991
Time:	20:52:55	Log-Likelihood:	-3.9651e+06
converged:	True	LL-Null:	-3.9730e+06
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
treatment	0.0101	0.002	5.869	0.000	0.007	0.013
JURISDICTION_CODE	0.0074	0.000	61.692	0.000	0.007	0.008
X_COORD_CD	0.0002	9.47e-06	23.348	0.000	0.000	0.000
Y_COORD_CD	0.0003	1.39e-05	22.795	0.000	0.000	0.000
Latitude	-116.4892	5.073	-22.965	0.000	-126.431	-106.547
Longitude	-60.2835	2.627	-22.944	0.000	-65.433	-55.134

Figure 11: Logistic regression results summary. **Modeling treatment against if a crime is a felony.**

IPW yielded an estimate of -0.0099020229564859. Upon visualizing the propensity scores in the treatment and non-treatment group on a histogram, we get the following histograms (Figure 12 and 13).

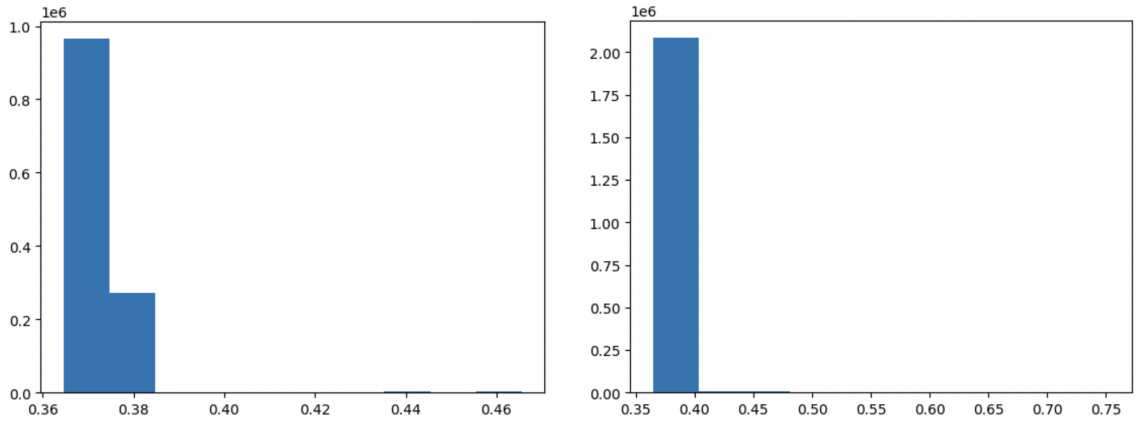


Figure 12 and 13: Distribution of propensity score when receiving the treatment (left) and not receiving the treatment (right).

5.3 Discussion

Overall, our results imply that the implementation of predictive policing has minimal causal effect on severity of crime. These two variables do have a high chi-square, showing they are correlated with each other, but correlation does not mean causation. The estimated treatment effect from running outcome regression and inverse propensity weighting were both very small in magnitude, close to 0.

However, there are some limitations with our efforts. Most columns had minimal documentation on how their values were calculated, leading to difficulty for us to draw the arrows in the causal DAG (Figure 4) that helped us identify any colliders. For example, while it may make intuitive sense that our location confounders x-y-coordinate as well as Latitude-Longitude would affect the confounder JURISDICTION_CODE – and led to it being a collider – we cannot confirm this effect and thus still used JURISDICTION_CODE as a confounder in our analysis.

The results from each of our logistic regressions are similar. Every coefficient is statistically significant with t-values greater than 2 or less than -2 (Figure 9, 10, and 11). Similar to the findings from our OLS regressions, the effect of treatment, JURISDICTION_CODE, X_COORD_CD, and Y_COORD_CD are miniscule compared to the effects of latitude and longitude. Predictive policing has the largest positive effect on classifying crimes as felonies, with a treatment coefficient of 0.0101 (Figure 11). Conversely, it has the largest negative effect on classifying crimes as misdemeanors, with a treatment coefficient of -0.0778 (Figure 10). With the effect of treatment being consistently low, we conclude that predictive policing only has a small effect on crime severity. Additionally, our log likelihood values for each of the 3 regressions are in the negative millions indicating that the model is poorly fit, challenging the validity of the model.

Predictive policing's low impact on crime severity along with the overall lack of model fittedness may be attributed to howOLS and logistic regressions only accept numeric variables as viable inputs. This forced us to ignore several categorical variables that may have been causally

related to crime severity, such as offense description (OFNS_DESC) and location description (PREM_TYP_DESC) (Morjaria, 2022).

As we can see from the propensity score histograms yielded from the IPW process, disregarding outliers, the propensity scores are overall slightly higher for non-treatment groups – treatment group scores are all <0.39 but for non-treatment group scores are <=0.40 (Figure 12 and 13). This means that based on the observed characteristics of the non-treatment group, they are more likely to be put in the treatment group.

This difference may be due to the varying application of predictive policing – for example, this policy is more selectively applied to areas with certain characteristics, such as specific demographic profiles. As a result, regions with lower predicted risk may be less likely to be targeted for predictive policing interventions. This makes sense given the background we gained from EDA that the number of offenses vary for different boroughs.

An IPW estimate of ~0.01 means that if we assume that there are no confounding variables then the estimated effect of having predictive policing is that this policy causes there to be ~ -0.01 decrease in intensity, where 1 is the difference in intensity of crime between felony and misdemeanor, and misdemeanor and violation.

6 Research Question #2: Hypothesis Testing

6.1 Methods

6.1.1 Problem setup

The following are our 6 hypotheses we tested using our dataset:

1. Is there a significant difference in crime severity in NYC before and after the implementation of predictive policing in 2013?
2. Is there a significant difference in the proportion of crimes reported per borough in NYC before and after the implementation of predictive policing in 2013?
3. Is there a significant difference in crime suspect sex in NYC before and after the implementation of predictive policing in 2013?
4. Is there a significant difference in crime suspect race in NYC before and after the implementation of predictive policing in 2013?
5. Is there a significant difference in crime suspect age groups in NYC before and after the implementation of predictive policing in 2013?
6. Is there a significant difference in crime level of offense in NYC before and after the implementation of predictive policing in 2013?

In order to comprehensively explore the effect of predictive policing on crime reports, we must test hypotheses on individual features of crime reports, one by one. Multiple hypothesis testing does well because it allows us to compare multiple factors simultaneously, but it also can increase the Type I error rate.

Additionally, we further cleaned the dataset so that data points prior to 2006 were dropped because data before 2006 was very sparse. In order to ensure we had a large enough sample size, we just looked at data from 2006 to 2018.

For hypothesis tests 2 through 6, we obtained extremely low chi-square statistics and extremely high p-values. See Figure 18 for individual results. We failed to reject the null hypothesis and concluded that there is no significant difference in respective features of complaints and implementation of predictive policing.

Lastly, we did not compute the power of an alternative hypothesis test, because our testing was focused on looking at the impact of treatment on a before and after group. Since we are not predicting anything, we do not have a true positive to calculate. While we don't have a defined "true positive", in our paired t-test we calculate "power" as a T statistic, which signals the strength of a significant difference between the two groups.

6.1.2 Hypothesis Test 1: Paired T-Test

Null hypothesis: There is no significant difference in the number of the complaints received per month before and after implementation of predictive policing.
Alternative hypothesis: There is a significant difference.
Test statistic: mean difference in the number of complaints, paired t-test to test for significance. The reason why we chose a paired t-test is because the nature of this question is a comparison between two groups, so we want to see if the two groups (complaints pre 2013 and post 2013) are significantly different from each other.

Conducting some light EDA, we saw a clear difference between the average complaints by month between pre-2013 and post-2013 (Figure 14). Upon performing the chi-square test with threshold 0.05, we obtained a paired t-test T statistic of 23.35 and a p-value of 1.005e-10. Since the T statistic is large and positive, we concluded that there is a significant decrease in the average number of complaints after 2013. From the p-value, we reject the null hypothesis. However, it is important to note that these results do not directly indicate that predictive policing was a cause, but rather points to a correlation in a significant decrease in average complaints per month and the implementation of predictive policing in 2013. This was the only hypothesis test that rejected the null.

	month	count	mean
0	1	217390	36231.666667
1	2	190030	31671.666667
2	3	216101	36016.833333
3	4	199810	33301.666667
4	5	223322	37220.333333
5	6	221721	36953.500000
6	7	229743	38290.500000
7	8	229524	38254.000000
8	9	221244	36874.000000
9	10	225051	37508.500000
10	11	204066	34011.000000
11	12	202496	33749.333333

	month	count	mean
0	1	140427	28085.4
1	2	123762	24752.4
2	3	142092	28418.4
3	4	143911	28782.2
4	5	155584	31116.8
5	6	152963	30592.6
6	7	160298	32059.6
7	8	160587	32117.4
8	9	152515	30503.0
9	10	156412	31282.4
10	11	141554	28310.8
11	12	141052	28210.4

Figure 14: Comparison of the average number of complaints received before and after 2013
The chart on the left shows complaints prior to 2013, and the chart on the right shows complaints after 2013.

6.1.3 Hypothesis Test 2: Chi-Square Test

Null hypothesis: Predictive policing does not have a significant effect on the proportion of victim sex before and after implementation of predictive policing.
Alternative: There is a significant difference.
Test statistic: Difference in proportions, using a chi-square test.

We initially looked at the severity of crime, but we realized that it was too similar to one of our other hypothesis tests. However, we ended up pivoting back to a chi-square test because we wanted to test for a significant difference between the two groups (pre and post 2013).

6.1.4 Hypothesis Test 3: Chi-Square Test

Null hypothesis: There is no significant difference in which borough a report was made before and after predictive policing
Alternative: Certain boroughs experience a significant difference in reports.
Test statistic: Proportion of borough reports before and after predictive policing, and will use a chi-square test to test significance.

We tried three different approaches, and ended up using chi-square testing. We first tried to look at prevention of crime, but since the majority (98%) of crimes were completed, we pivoted to location instead. We initially used logistic regression but realized that predicting based on one feature (borough) alone was not enough to accurately predict pre or post 2013. We ended up pivoting back to a chi-square test because we wanted to test for a significant difference between the two groups (borough reports pre and post 2013).

6.1.5 Hypothesis Test 4

Null hypothesis: There is no significant difference in suspect's race before and after the implementation of predictive policing in 2013.
Alternative: Certain racial groups experience a disproportionate increase in complaints after the implementation of predictive policing, indicating potential targeting or bias.
Test statistic: Proportion of complaints by demographic groups before and after predictive policing.
We initially used Poisson regression to test significance, because we wanted to create a model that predicts a binary categorical variable—whether or not a complaint for a certain racial group comes before the implementation of predictive policing or not. We ended up using a chi-square test, because we wanted to focus more on the relationship between the two categorical variables, suspect's race and predictive policing implementation, as opposed to the occurrences of complaints on different racial groups within fixed timeframes.

6.1.5.1 Hypothesis Test 4: Poisson Regression

For feature engineering, we one-hot-encoded 'SUSP_RACE' and used it to test whether a suspect's race has a statistically significant effect on the likelihood of a complaint coming before 2013; Poisson regression models analyze occurrences of an event over the span of a fixed timeframe. Then we performed standard Poisson procedures (splitting into testing and training, fitting, etc).

```
Optimization terminated successfully.
         Current function value: 0.801886
         Iterations 6
                          Poisson Regression Results
==============================================================================
Dep. Variable:                pre2013   No. Observations:              2626180
Model:                        Poisson   Df Residuals:                  2626172
Method:                           MLE   Df Model:                            7
Date:                Mon, 06 May 2024   Pseudo R-squ.:               0.001026
Time:                        04:59:44   Log-Likelihood:           -2.1059e+06
converged:                       True   LL-Null:                  -2.1081e+06
Covariance Type:            nonrobust   LLR p-value:                     0.000
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            -0.7465        nan        nan        nan         nan         nan
AMERICAN INDIAN/ALASKAN NATIVE   -0.0350        nan        nan        nan         nan         nan
ASIAN / PACIFIC ISLANDER         -0.2397        nan        nan        nan         nan         nan
BLACK                            -0.0915        nan        nan        nan         nan         nan
BLACK HISPANIC                   -0.1614        nan        nan        nan         nan         nan
OTHER                             0.7465        nan        nan        nan         nan         nan
UNKNOWN                           0.0059        nan        nan        nan         nan         nan
WHITE                            -0.0212        nan        nan        nan         nan         nan
WHITE HISPANIC                   -0.1251        nan        nan        nan         nan         nan
=================================================================================================
```

Figure 15: Poisson Regression Model Results.

The coefficients associated with each race category in the model results (Figure 15) denote the estimated change in the pre-2013 predictions, for one-unit change in the suspect's race.

Using the Poisson model, we called predictions on the testing set that yielded the probability of each complaint by a certain racial group having occurred before 2013. With a threshold 0.5 (random chance), we converted these probabilities into binary predictions. We then called a confusion matrix to assess model performance.

```
Confusion Matrix:
[[366226      0]
 [290317      2]]
```

Figure 16: Poisson Regression Model Confusion Matrix.

The model yielded 366,226 true negatives, 0 false positives, 290,317 false negatives, and 2 true positives (Figure 16). Comparing the prediction values with true observations, we obtained an accuracy of 59.6%. We then obtained the p-values associated with each racial category coefficient.

```
P-values for each coefficient:
const                                0.999950
AMERICAN INDIAN/ALASKAN NATIVE       0.999997
ASIAN / PACIFIC ISLANDER             0.999992
BLACK                                1.000000
BLACK HISPANIC                       0.999996
UNKNOWN                              0.999994
WHITE                                0.999996
WHITE HISPANIC                       0.999998
dtype: float64
```

Figure 17: P-Values per Poisson Regression Model Coefficients.

Given that our Poisson regression model predicts correctly 59.6% of the time and that all of the p-values associated with each racial category coefficient is close to 1(Figure 17), we concluded that it is likely that suspect's race has no statistically significant effect on the likelihood of a complaint occuring before 2013.

Notably, such a conclusion does not directly answer our hypothesis question "Is there a significant difference in crime suspect race in NYC before and after the implementation of predictive policing in 2013?". We realized that a chi-square test more suitably answers the question "Is there a significant difference in crime occurrence in NYC based on crime suspect's race?".

6.1.5.2 Hypothesis Test 4: Chi-Square Test

Prior to running the test, we split the data in 2 different datasets that only contain from before 2013 and after 2013, and ensure that 'SUSP_RACE' values for the pre-2013 dataset and

post-2013 dataset contained the same categorical variables. We dropped rows with values NaN or “OTHER”, so this process ensured that we dropped NaN rows.

The number of complaints from the pre-2013 and post-2013 dataset is bound to differ—intuitively, the dataset that covers a longer period of time is going to have significantly more crime counts. We therefore calculated proportions of complaints by suspect race for each dataset and then performed the chi-square test.

6.1.6 Hypothesis Test 5: Chi-Square Test

Null hypothesis: There is no significant difference in the suspect's age group before and after the implementation of predictive policing in 2013.
Alternative: The suspect’s age group changed significantly after the implementation of predictive policing in 2013.
Test statistic: Proportion of suspects by age group before and after predictive policing, and will use a chi-square test to test significance.

Similarly to hypothesis test 4, we split the dataset into pre-2013 values and post-2013 values. There were a 86 counterintuitive values for the ‘SUSP_AGE_GROUP’ column, such as ‘926’, ‘2016’, ‘and ‘709’, which do not fall under the age group category labels such as ‘<18’ or ‘45-64’. We treated the counterintuitive values as missing ‘UNKNOWN’ values. We then calculated proportions of complaints by the suspect's age group for each dataset and performed the chi-square test.

6.1.7 Hypothesis Test 6: Chi-Square Test

Null hypothesis: There is no significant difference in the level of offense of complaints reported in NYC before and after the implementation of predictive policing in 2013
Alternative: The level of offense complaints increased significantly.
Test statistic: The proportion of complaints by level of offense, will use a Chi square test for significance.

After splitting the dataset into pre-2013 values and post-2013 values, we calculated proportions of complaints by crime offense levels for each dataset. There were no NaN values for ‘LAW_CAT_CD’, which is the column indicating level of offense for each complaint such as ‘MISDEMEANOR’, ‘FELONY’, and ‘VIOLATION’. We then performed the chi-square test.

6.1.8 Controlling for Error Rates

To correct for multiple hypothesis tests, we used two different error rate control methods: Bonferroni Correction and Benjamini-Hochberg correction.

Bonferroni correction is a FWER (family-wise error rate) correction method that controls the probability of a false discovery amongst each individual hypothesis test to be under a threshold value. The strength of the Bonferroni procedure is that it minimizes Type 1 errors, because it controls FWER by using a threshold value of alpha divided by total number of hypothesis tests for each individual hypothesis test. Concomitantly, it follows that Bonferroni

correction has more likelihood to make Type 2 errors, because its conservative decisions prioritize avoiding false discoveries in borderline cases. Upon applying Bonferroni correction on our 6 hypothesis tests, we made 1 total discovery, out of which 0 were false (Figure 19).

Next we tried the more lenient Benjamini-Hochberg correction method, which controls the proportion of false discoveries amongst rejected hypotheses of significant p-values. Upon applying the BH correction on our 6 hypothesis tests, we made 1 total discovery, out of which 0 were false (Figure 20). For our case, the BH correction method is more suitable. Our dataset does not contain a wide span of time for before and after the implementation of predictive policing. It is unlikely that the effects of a policing method takes place immediately, so it is important to prioritize the detection of any discovery, as opposed to conservatively minimizing false discoveries. If we had more than 6 hypotheses, using the BH correction method would be even more preferable, because its threshold for false discovery proportion becomes more lenient.

6.2 Results

6.2.1 Multiple Hypothesis Testing Results

	statistic	p-value
Hypothesis 1	23.35	1.01E-10
Hypothesis 2	0.00257648729	0.9999652444
Hypothesis 3	0.00015003868	0.9999999972
Hypothesis 4	0.00592514579	0.9999999957
Hypothesis 5	0.139338346	0.999633173
Hypothesis 6	0.01456996641	0.992741488

Figure 18: Multiple Hypothesis Testing Results.

6.2.1 Error Rate Control Results

```
            Decision = 0  Decision = 1
Truth = 0              5             0
Truth = 1              0             1

total discoveries: 1
fraction of discoveries which were actually false: 0.000
```

Figure 19: Bonferroni Decision Results.

```
           Decision = 0  Decision = 1
Truth = 0             5             0
Truth = 1             0             1

total discoveries: 1
fraction of discoveries which were actually false: 0.000
[ True False False False False False]
```

Figure 20: Benjamini-Hochberg Decision Results.

6.3 Discussion

After applying the correction procedures, there was only 1 discovery that remained significant– our first hypothesis test. The error correction did not change our initial discoveries. However, this is expected because our p-values were very extreme – either almost 1 or almost 0.

From the individual tests, we can make the decision to consider the average number of complaints as an important feature that is impacted by predictive policing. From the aggregate results, we saw that the other features did not change much after 2013.

Some of the limitations of our analysis is that we only looked at complaint data to analyze the impact of predictive policing. However, it would have been very beneficial to get data on arrests made on site, which could have given us more insight on the behavior of the police officers. By solely focusing on complaint data, we may have unintentionally engaged in p-hacking, since we were choosing which features to analyze, rather than analyzing them all. Although it is not comprehensive, we mitigate this risk using a large number of hypothesis tests, (6 in total), and prioritized controlling error rates using statistical methods discussed in Lab 2.

Another limitation is that for hypothesis test 5, dropping NaN and counterintuitive 'SUSP_AGE_GROUP' values resulted in more than 70% data points to be 'UNKNOWN'. The significant lack of data likely has skewed the significance of our test results.

It was hard to find comprehensive crime data with many features and a large enough sample size. We were lucky because NYC already had data before and after the implementation of predictive policing, however most states did not have this data. It would be very interesting to compare crime reports by state, especially those with and without predictive policing. If we had more data, we will be able to make more generalizable conclusions about the impact of predictive policing, rather than state-specific.

7 Conclusion

Our multiple hypothesis tests indicate that while there was only one significant result in multiple hypothesis testing, suggesting a notable difference in crime report demographics, other factors such as race, sex, and age did not exhibit significant disparities. Our causal inference analysis found the impact of predictive policing on the level of offenses to be minimal. Although these are primarily applicable to NYC which is where our dataset is based, our findings represent a small glimpse of the impact of predictive policing. Rather than use our findings as evidence to

make decisions in the real world, our findings should be a consideration in developing a cautious approach that ensures both effectiveness and ethical considerations.

We did not merge different data sources, relying solely on the extensive NYPD dataset. The analysis was hindered by the inability to include categorical variables into our causal analysis, limiting the comprehensiveness of confounders and model fitting, so future work should focus on quantifying these. Another future step is to build a model using the significant features we found to predict whether a report came from pre-2013 or post-2013, to identify which features changed the most after predictive policing.

Throughout the project, we learned the iterative nature of data science projects, requiring extensive revision and reflection on project goals.

References

Lau, Tim. Predictive Policing Explained. Brennan Center for Justice, 2016, https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained.

Geographic Coordinate Systems. Esri. "Geographic Coordinate Systems." ArcGIS Help, 2010, https://help.arcgis.com/en/geodatabase/10.0/sdk/arcsde/concepts/geometry/coordref/coordsys/geographic/geographic.htm#:~:text=In%20this%20case%2C%20longitude%20values,on%20the%20spherical%20reference%20system.

Morjaria, Mr., "New York City Police Crime Data (Historic)." Kaggle, 2022, https://www.kaggle.com/datasets/mrmorj/new-york-city-police-crime-data-historic?resource=download.

National Institute of Justice. Predictive Policing: Forecasting Crime Through Technology. U.S. Department of Justice, 2019, https://www.ojp.gov/pdffiles1/nij/230414.pdf.

Appendix

```
Mann-Whitney U test for CMPLNT_NUM:
   U statistic: 5114410191191.0
   p-value: 0.7822998936876222
   Conclusion: There is no statistically significant difference between the groups.

Mann-Whitney U test for ADDR_PCT_CD:
   U statistic: 5111014536975.5
   p-value: 0.2405321316351211
   Conclusion: There is no statistically significant difference between the groups.

Mann-Whitney U test for KY_CD:
   U statistic: 5159983540311.0
   p-value: 1.725453542751991e-87
   Conclusion: There is a statistically significant difference between the groups.

Mann-Whitney U test for PD_CD:
   U statistic: 4986467450188.0
   p-value: 0.0
   Conclusion: There is a statistically significant difference between the groups.

Mann-Whitney U test for JURISDICTION_CODE:
   U statistic: 5143243277543.0
   p-value: 1.0984289459614415e-36
   Conclusion: There is a statistically significant difference between the groups.

Mann-Whitney U test for X_COORD_CD:
   U statistic: 5153885674355.0
   p-value: 8.719115198252066e-66
   Conclusion: There is a statistically significant difference between the groups.

Mann-Whitney U test for Y_COORD_CD:
   U statistic: 5140433761664.0
   p-value: 4.82782784552511e-30
   Conclusion: There is a statistically significant difference between the groups.

Mann-Whitney U test for Latitude:
   U statistic: 5140521416197.5
   p-value: 3.1401918823300707e-30
   Conclusion: There is a statistically significant difference between the groups.

Mann-Whitney U test for Longitude:
   U statistic: 5153958376104.0
   p-value: 5.11141316788845e-66
   Conclusion: There is a statistically significant difference between the groups.
```

Figure A1: Mann-Whitney U Tests.