Data exercise: Police searches

Ibraheem Saqib Ellahi

December 29, 2025

Contents

1 Racial Disparities in Search Rates

	subject_race	n	percentage
	<chr>	<int>	<dbl>
1	Black	49861	71.0
2	Hispanic	14113	20.1
3	White	5485	7.81
4	Other	759	1.08

Figure 1: Distribution of Subject Race (1.1)

	model_search_race
Dependent Var.:	search_conducted
subject_raceBlack	0.0255** (0.0078)
subject_raceHispanic	0.0310*** (0.0073)
subject_raceOther	0.0073 (0.0140)
Fixed-Effects:	----------
fe_district_year_month	Yes
fe_beat_weekend_quarter	Yes
S.E.: Clustered	by: FO_EMPLOYEE_ID
Observations	70,213
R2	0.06035
Within R2	0.00040

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 2: Regression Results: Search Rates by Subject Race (1.2)

```
Linear hypothesis test:
subject_raceBlack = 0

Model 1: restricted model
Model 2: search_conducted ~ subject_race | fe_district_year_month + fe_beat_weekend_quarter

  Res.Df Df  Chisq Pr(>Chisq)
1  69382
2  69381  1 10.796   0.001017 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Hypothesis Testing: White and Black (1.3)

```
Linear hypothesis test:
subject_raceBlack - subject_raceHispanic = 0

Model 1: restricted model
Model 2: search_conducted ~ subject_race | fe_district_year_month + fe_beat_weekend_quarter

  Res.Df Df  Chisq Pr(>Chisq)
1  69382
2  69381  1 0.5692     0.4506
```

Figure 4: Hypothesis Testing: Black and Hispanic (1.3)

Table 1: Regression of Search Indicator on Subject Demographics

Dependent Variable:	search_conducted		
Model:	(1)	(2)	(3)
Variables			
Constant	0.1455***		
	(0.0171)		
Race: Black	0.0188	0.0255***	0.0162**
	(0.0171)	(0.0078)	(0.0076)
Race: Hispanic	0.0138	0.0310***	0.0167**
	(0.0129)	(0.0073)	(0.0073)
Race: Other	0.0113	0.0073	0.0042
	(0.0164)	(0.0140)	(0.0139)
Sex: Male			0.0407***
			(0.0051)
Sex: Other/Unknown			-0.1493***
			(0.0272)
Age			-0.0019***
			(0.0002)
Age (Missing Flag)			-0.0901***
			(0.0101)
Design Fixed Effects?	No	Yes	Yes
Mean of Dep. Var.	0.1620	0.1620	0.1620
Fixed-effects			
fe_district_year_month		Yes	Yes
fe_beat_weekend_quarter		Yes	Yes
Fit statistics			
Observations	70,213	70,213	70,213
R^2	0.00020	0.06035	0.06719

Clustered (FO_EMPLOYEE_ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

The regression results in Table 1 indicate a statistically significant racial disparity in the likelihood of being searched. In Model 2, which controls for location and time through design fixed effects, Black subjects are approximately 2.6 percentage points more likely ($p < 0.01$) and Hispanic subjects are 3.1 percentage points more likely ($p < 0.01$) to be searched compared to White subjects. When adding controls for age and gender in Model 3, these coefficients remain economically and statistically significant. The stability of these estimates across specifications suggests that the disparity is not driven solely by differences in the neighborhoods patrolled or the time of day stops occur.

2 Measuring Officer Heterogeneity

Figure 5 shows that most officers are clustered around the mean, though there are a number of outliers. These outliers have a high positive value, which means that their search rates are higher than that of the average officer. Figure 6 shows that similar search rates persist in both halves of stops, supporting the previous results. Table 2 shows a statistically significant relationship of 0.830, implying that officers are consistent across both halves of their searches.

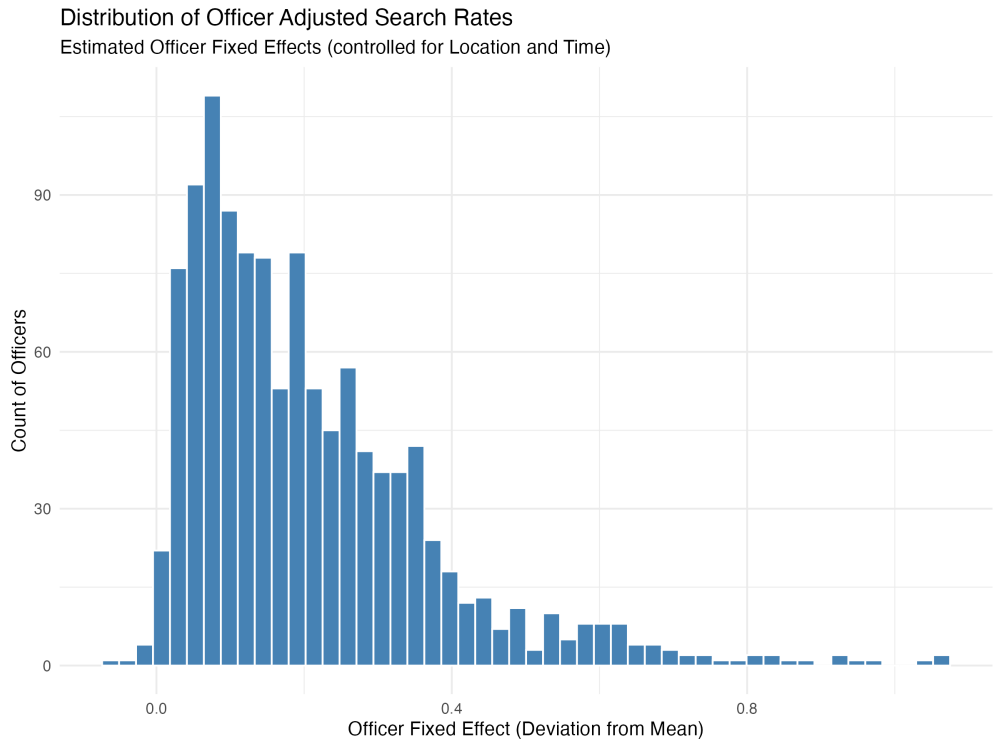

Figure 5: Distribution of Officer Adjusted Search Rates (2.1)

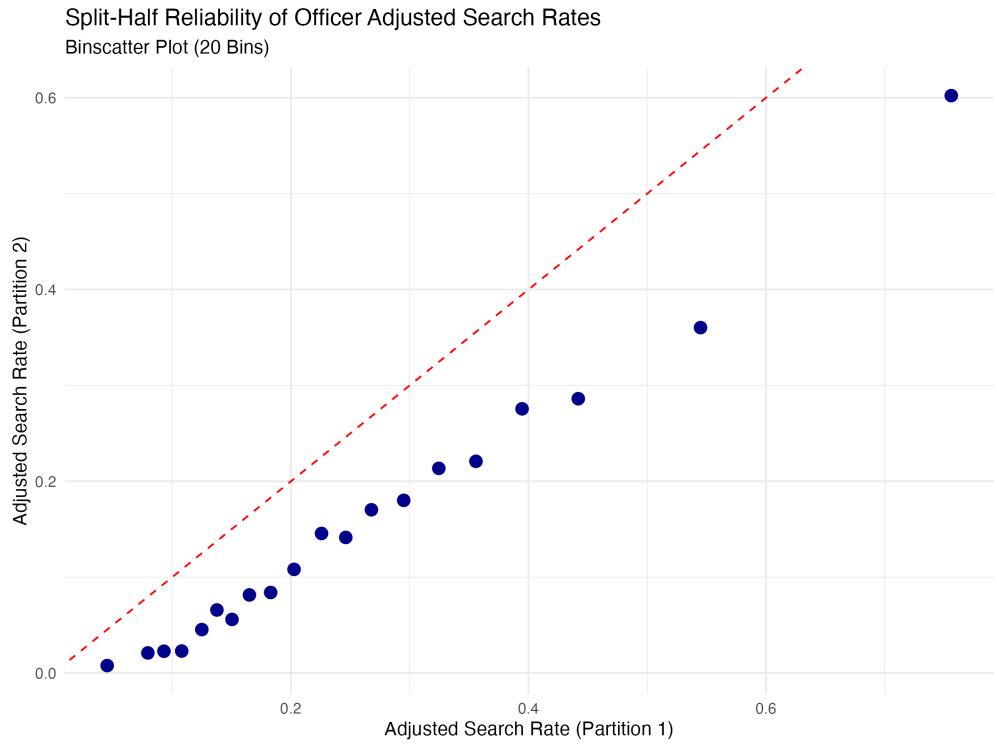

Figure 6: Split-Half Reliability of Officer Adjusted Search Rates (2.2)

Table 2: Reliability of Officer Adjusted Search Rates (Split-Sample) (2.3)

	Dependent variable:
	Adjusted Search Rate (Partition 2)
Adjusted Search Rate (Partition 1)	0.830^{***}
	(0.018)
Constant	-0.058^{***}
	(0.006)
Observations	1,150
R^2	0.654
Note:	*p<0.1; **p<0.05; ***p<0.01

3 Search Rates and Hit Rates

	model_contraband
Dependent Var.:	contraband_found
lo_search	0.2406*** (0.0191)
Fixed-Effects:	----------
fe_district_year_month	Yes
fe_beat_weekend_quarter	Yes
S.E.: Clustered	by: FO_EMPLOYEE_ID
Observations	70,213
R2	0.05236

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 7: Regression Results: Contraband Yield on Search Rate (3.2)

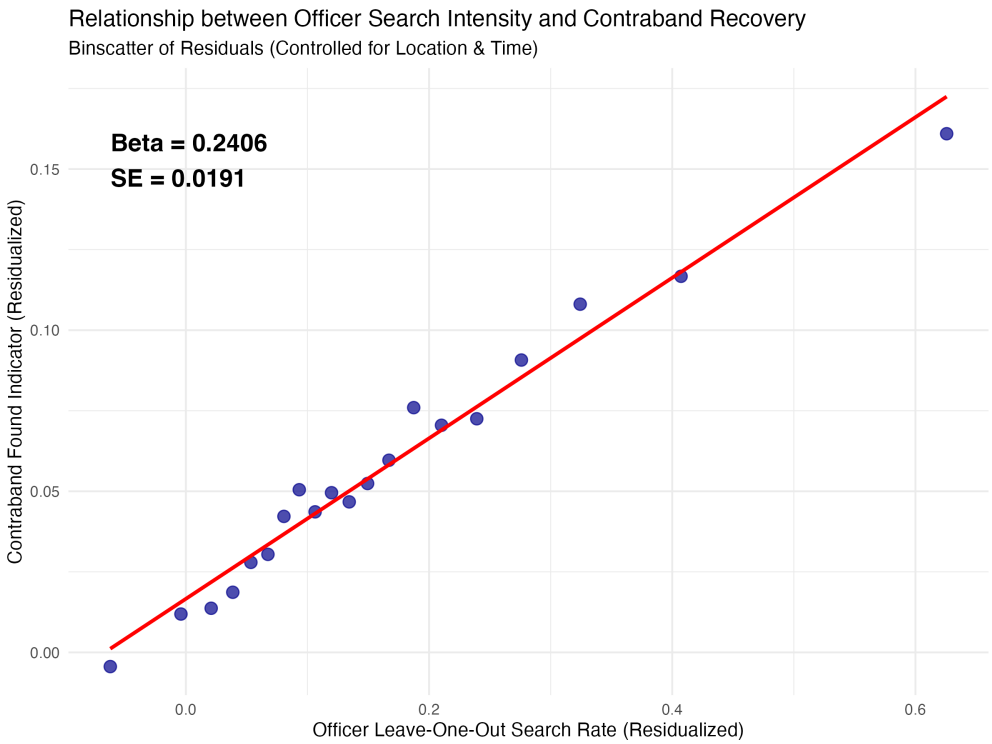

Figure 8: Relationship between Officer Search Intensity and Contraband Recovery (3.2)

Table 3: Correlation between Officer Search Rate and Subject Demographics

Dependent Variable:	lo_search
Model:	(1)
Variables	
Race: Black	-0.0059
	(0.0048)
Race: Hispanic	-0.0003
	(0.0045)
Race: Other	-0.0065
	(0.0068)
Sex: Male	0.0145***
	(0.0030)
Sex: Other/Unknown	-0.0399**
	(0.0202)
Age	-0.0013***
	(0.0001)
Age (Missing Flag)	-0.0431***
	(0.0088)
Fixed-effects	
fe_district_year_month	Yes
fe_beat_weekend_quarter	Yes
Fit statistics	
Observations	70,213
R^2	0.14437

Clustered (FO_EMPLOYEE_ID) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

```
Analysis of Variance Table

Model 1: search_conducted ~ 1
Model 2: search_conducted ~ subject_race + age_clean + sex_clean
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  70217 9519.5
2  70211 9439.8  6    79.614 98.692 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9: Joint Significance Test for Race, Age, and Gender (3.4)

Table 3 examines whether officers with a higher propensity to search (proxied by their leave-one-out search rate, `lo_search`) are more likely to stop specific demographic groups. The coefficients for subject race are statistically indistinguishable from zero, suggesting that "high-search" officers do not stop minority subjects at disproportionately higher rates than "low-search" officers, conditional on the fixed effects. However, the coefficient for `Sex: Male` is positive and significant (0.0145), and the coefficient for `Age` is negative and significant (-0.0013). This implies that officers who conduct searches more frequently tend to stop male and younger subjects at a higher rate than their peers who search less often.

3.1 Measure of Subject's Socioeconomic Status

To construct a measure of a subject's socioeconomic status using the existing ISR dataset and external sources, I proposes a strategy focusing on vehicle valuation and location of stop:

1. **Vehicle Valuation (Individual Level):** The ISR dataset contains the fields `VEHICLE_MAKE`, `VEHICLE_MODEL`, and `VEHICLE_YEAR`. A direct proxy for wealth can be constructed by estimating the market value of the subject's vehicle. I can assign an estimated dollar value to the vehicle involved in the stop using an external dataset.

2. **Location of Stop (Neighborhood Level):** While individual income is not recorded, the location of the stop is available via `BLOCK_CD`, `BEAT_CD`. Using the U.S. Census Bureau's TIGER/Line shapefiles, I can join the stop locations to their corresponding Census Tracts and assign *Median Household Income* and *Poverty Rate* data to the stop. While a stop location does not guarantee the subject lives there, it can serve as a viable proxy for neighborhood socioeconomic status.

4 Work Summary

1. **Time taken:** I spent a total of 4 hours and 30 minutes on the exercise. This includes approximately 45 minutes on preparing the dataset and understanding the tasks, 1 hour on Part 1, 45 minutes on Part 2, 1 hour 30 minutes on Part 3, and 30 minutes on review.

2. **Approximate runtime:** 8.2 seconds.

3. **Generative AI:** My use of AI was limited to brainstorming ways to address certain errors and writing code for repetitive tasks like defining clean dictionary labels for regression tables.