A Data Science Approach to Gender Based Violence in India

Abraham Bauer, Matthew Duffy

Table of Contents

1. Introduction

1.1 Background

Gender-based violence (GBV) has run rampant in communities all over the world. According to the Indian government's National Family Health Survey from 2019-2021, "[t]wenty-nine percent of women age 18-49 have experienced physical violence since age 15, and 6 percent have ever experienced sexual violence in their lifetime" (IIPS and ICF, 2022, p. 639). Regarding GBV, India "is both a critical case and a representative one" (Sukhtankar, Kruks-Wisner & Mangla, 2022, p. 191). India has high rates of crimes against women (CAW), but an excess of GBV plagues every part of the world. One common theme with GBV is a lack of reporting due to low trust and confidence in the police. Also, many of the cases which are reported do not actually get filed due to bias and insensitivities within the police department. Out of the victims of GBV "only 14 percent have sought help for the violence, and 77 percent have never sought any help nor told anyone about the violence they experienced" (IIPS and ICF, 2022, p. 648). In response, there have been initiatives to reform police practices and procedures to combat GBV and its associated struggles. Some of those initiatives include offering a general and female operated women's help desks, hiring more female officers to help with GBV cases, and special trainings to help officers deal with CAW. One article on GBV, which specifically looks at these reforms in India, have found with "mixed and often disappointing results" (Sukhtankar, Kruks-Wisner & Mangla, 2022, p. 191).

Sandip Sukhtankar, Gabrielle Kruks-Wisner, and Akshay Mangla analyzed Indian GBV and its policing and wrote an article with their findings entitled, "Policing in patriarchy: An experimental evaluation of reforms to improve police responsiveness to women in India" (2022). They collected data via police records and various surveys to understand how reforms affect the filing of CAW and found some mixed results. They noticed that the presence of a women's help desk made officers "more likely to register cases of GBV, particularly where female officers run the desks" (p. 191). But they did not find that the presence of female officers mattered much at all, "gender attitudes among the police are difficult to move," and WHD do not alleviate all barriers to women reporting their cases. These findings leave society with more questions and concerns about GBV, so it is time to build on their insights to learn more about the problems GBV causes and what sorts of things can potentially remedy them. More clarification should demonstrate if these reforms help with GBV reporting or are wastes of time and resources.

There are a handful of logistical points to outline regarding the data and the analysis. First off, the variables that describe the reforms, or treatments, exist in every dataset: *treatment, group, regular_whd,* and *women_whd*. The variables *treatment, regular_whd,* and *women_whd* and binary variables that indicate if one of these was present; *treatment* tells whether it was a control or treatment group and *regular_whd* and *women_whd* shows whether something had a regular, or women run helpdesk. The *group* variable is a string that states the classification of treatment and control groups. These variables will often be used when discussing the reforms that could be in place. Moreover, the main types of reports for GBV in India are First Information Reports (FIRs) and Domestic Incident Reports (DIRs) (192, Sukhtankar). FIRs are more common and constitute an investigation while DIRs are a bit rarer and constitute civil proceedings and referrals to social services. Lastly, all analysis will be implemented in Python via the use of Jupyter Notebooks and all data comes from Sukhtankar's article.

1.2 Brief Literature Review

Studying GBV is crucial for society because it is an ethical issue at its core. Andrew Morrison, Mary Ellsberg, and Sarah Bott, authors of "Addressing Gender-Based Violence: A Critical Review of Interventions" (2007), give the United Nations definition of GBV as "any act of gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women, including threats of such acts, coercion or arbitrary deprivations of liberty, whether occurring in public or private life" (p. 25). They also outline that "the consequences of gender-based violence for women's health and well-being, [range] from fatal outcomes such as homicide, suicide, and AIDS-related deaths to nonfatal outcomes such as physical injuries, chronic pain syndrome, gastrointestinal disorders, unintended pregnancies, and sexually transmitted infection" (p. 32). GBV covers a wide array of horrible crimes; women are being objectified and treated as a mere means to sexual pleasure or physical superiority. That violates human rights, and more needs to be done to limit GBV. Police need to improve at preventing CAW before they occur and care for women after these horrific crimes occur. Women need to feel comfortable reporting these crimes so their perpetrators cannot hurt anyone else. Morrison et al. suggest that society must "increase access to justice for women" (p. 25). One actionable way to do that is truly enforcing the law instead of just putting the laws into place. If the police file more reports and make bad actors deal with the consequences of their actions more often, the rates of GBV will naturally diminish. It may also make women feel more confident that their case will be delt with, which would encourage them to report it. Morrison et al. give a helpful background to GBV and shows how interconnected it is with law enforcement; it also evokes a call to make justice more accessible and helpful to women. Better policing can make the world a better place for women everywhere.

In "Gender, Crime and Punishment: Evidence from Women Police Stations in India" (2021) by Sofia Amaral, Sonia Bhalotra, and Nishith Prakash, Amaral and her co-authors study how police stations fully staffed by women (WPS) affect the filing of GBV reports. In the article, they quickly identified that GBV is "one of the most widespread violations of human rights, it is at the same time one of the least reported forms of crime" (Amaral, Bhalotra & Prakash, 2021, p. 2). That is a call for research and change; people are being hurt and they do not feel their suffering is worth reporting because they do not trust the police, the people that are meant to protect them. So much pain could be prevented if communities trusted local police (and the police act appropriately of course). Through their study, they found that "the opening of a WPS in a city leads to an immediate and persistent increase in city-level reported violence against women" (p. 4). Specifically, "the implementation of WPS led to a 29 percent increase in the reporting of GBV cases to the police" (p. 23). That increase is promising for the hope of *increased justice* via police reforms, but also hints at how many women do not report their GBV cases. The authors also claim that male officers reduce the effect that the WPS has on the community around it (p. 7). That seems to indicate that reforms involving male officers may be less effective in increasing reporting of GBV. Amaral's article illustrates how big of a problem GBV and its lack of reporting is while providing a strong precedent that some reforms can cause a significant increase in reports.

2. Problem Descriptions

2.1 Question 1

***Q1:** What is the relationship between the number of female staff at a given station with the perceptions pertaining to crimes against women held by all staff at that station?*

The motivation for this question stems from the introduction from Sukhtankar, Kruks-Wisner, and Mangla (2022), where the authors mention that another approach toward improving law enforcement and reducing gender-based violence, aside from creating designated desks or women-only stations, is to simply increase the number of women on staff at regular stations. This builds on theories that "hold that the presence of members of marginalized groups within a public agency improves performance with respect to those groups," although evidence for this type of intervention remains mixed (p. 191). Although the study data provided was not structured to allow for a causal analysis, an exploratory analysis between the number of women working at each station and the attitudes of the officers at that station could help indicate if there is any kind of relationship in either direction, positive or negative, among the population studied, thus offering support for or evidence against the theory of presence improving performance.

2.2 Question 2

***Q2:** Does the creation of a WHD or a women-run WHD cause an increase in dial100 calls to the station?*

The motivation for this question is the theory that the number of calls made to the police station may be one indicator of the degree to which a community trusts the police to help in cases of emergency. Sukhtankar, Kruks-Wisner, and Mangla (2022) use survey instruments to measure changes in satisfaction with experiences with the police and general community feelings of safety, to identify if WHDs improve the perception of the police by constituents. Does the presence of WHDs have any effect on the number of calls to the station, a different form of police interaction?

2.3 Question 3

***Q3:** How do community perceptions and opinions of police change when treatments are implemented?*

A major claim in the study is that perceptions of police, or "low trust in the police" can often cause people not to report cases of GBV (Sukhtankar, Kruks-Wisner, & Mangla, 2022, p. 191). People feel their cases will not be delt with nor respected, so they decide to not waste their time making police reports. Therefore, it is vital to understand what can change the perception of police. If treatments have caused police to file more reports, perhaps people have heard about it and now believe more in the police. If people feel better about police interactions after treatment, the community may be more apt to report FIRs. However, it is possible that the public perceptions are not altered by these recent improvements in policing for GBV. Perhaps more needs to be done to show the public how the police have improved. If treatments are truly increasing the number of reports filed, it will be worth scaling these sorts of treatments so women all over the world see improvements in policing CAW.

2.4 Question 4

***Q4:** How much does the number of female officers/presence of a female helpdesk affect the number of women that enter the station?*

The authors explain that gender-based reforms seem to increase the quality of service for women by police (p. 197), but is that proportional to the increased amount of people that show up at the station? Seeing if more people are showing up or just more reports are being filed can indicate that women who do not trust the police are not being affected by these changes. Only women who believe the police can help them will go to the police and the reforms merely make more reports get filed. That is an important distinction to understand how these treatments are actually affecting society. It is great if they cause more reports to be filed, but if more people are not coming into the station, then more reforms need to happen. Training the police is important, but it is similarly important to make sure the community feels confident in the police and will utilize them. Seeing if more people (men included) go to the station can indicate an increase in confidence in the police from the entire community.

3. Data Exploration

3.1 Question 1

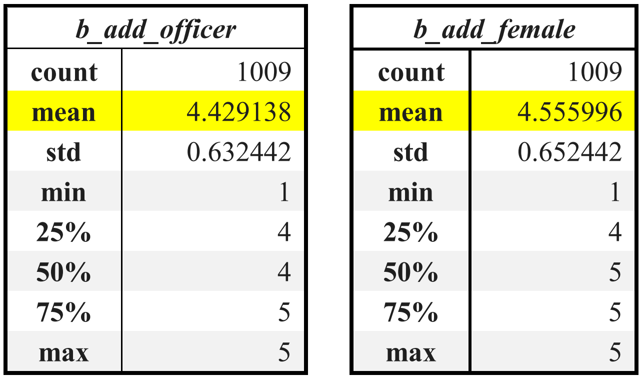
To evaluate our first question, we relied on those data related to the police attitudinal survey, particularly the files *Police_baseline_data* (a.k.a. *P_baseline*), *Police_full_data* (a.k.a. *P_full*), and

Police_station_personnel_data (a.k.a. *P_station*). We also made frequent reference to the baseline and endline Police officer survey instruments. Beginning with *P_full*, we found that out of a total of 1,961 rows, ten endline survey question columns contained between 2 and 29 null values. Removing all rows that contained null values in one or more of these ten columns resulted in a dataset of 1,904 rows, a substantial subset. Later, we used the same procedure to clean a subset of *P_baseline* of null values, leaving a dataset of 1,106 rows. When reviewing *P_station* for null values, it was found that the columns *e_female_weight* and *b_female_weight* contained 16 and 12 NaN values, respectively. In this case, considering **_female_weight* indicated the weight of women interviewees at a given station, a NaN value indicated no women were interviewed at the given station; as such, we tentatively filled each NaN instance with the integer 0.

The values we were interested in studying were the baseline police survey results for each individual, as well as the total number of male and female staff at that individual's station. Baseline responses were chosen to simplify the results by preceding any treatment effects. To have all these values in a single set, we performed a left-merge between *P_full* and *P_station* with *P_full* on the "left," thus matching each single station's information to multiple police survey participants and resulting in a dataset of 1,904 rows (which we will call *P_full_stat*). Examining the contents of this dataset, particularly the indexed baseline question responses, and reviewing the Read Me file provided by the researchers, it became clear that a number of these baseline responses were imputed or otherwise modified in order to connect one-to-one baseline and endline responders. In the process of this imputation, ordinal variables were converted to non-integer float values. Thus, it would be more logical to have merged *P_personnel* with *P_baseline.* Out of curiosity to see if only missing baseline answers were imputed, or if existing baseline responses had been in any way modified, we thus left-merged a subset of rows from *P_baseline* and *P_full_stat* with *P_baseline* on the "left" in order to drop any fully manufactured baseline entries; the resulting dataset, *P_data*, contained 1,009 rows.

It was found that for these remaining "true" baseline cases, the survey results from *P_baseline* and *P_full* corresponded directly, although in *P_full* ordinal textual responses (e.g. those for *b_add_officer*: much more effective – more effective – no difference – less effective – much less effective) had been converted to ordinal numerical scales (e.g. integers 5 – 1). Intriguingly, one question, *b_add_female*, appeared to be reverse-coded when converted to a numerical index. This question asked: "Do you think hiring additional **female** officers will make the police more or less effective in dealing with cases related to women?" The question follows a nearly identical question, *b_add_officer*, which asks the same question without the word female, to ask about adding officers generally. The flipped index on *b_add_female* leads to misleading results. See summary statistics for each question in **Table 3.1**, with the means highlighted to demonstrate the seeming extreme similarity between the two responses in aggregate. Yet, a quick look comparing the respective textual responses, as shown in **Figure 3.1**, reveals the reverse coding hides the fact that *b_add_female* elicited almost a perfectly inverted distribution of answers to *b_add_officer.*

Table 3.1



b_add_officer	
count	1009
mean	4.429138
std	0.632442
min	1
25%	4
50%	4
75%	5
max	5

b_add_female	
count	1009
mean	4.555996
std	0.652442
min	1
25%	4
50%	5
75%	5
max	5

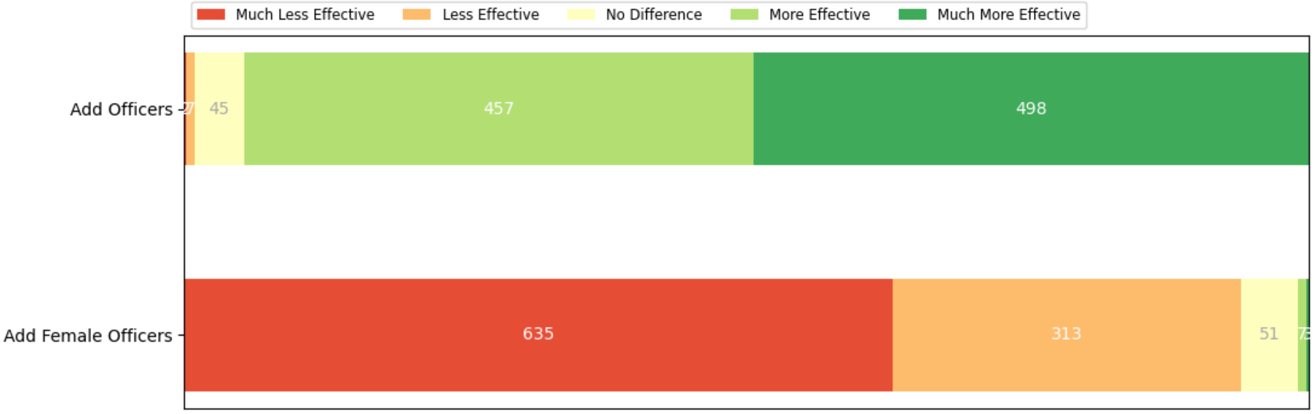


Figure 3.1

There are two potential possibilities: either the results for *b_add_female* are simply intentionally reverse-coded, or there is a problem with the textual data. Arguing for the second case, the two questions about adding officers are followed by the additional question *b_female_better*: “Who do you think is more effective in dealing with cases related to women? [Options: male, female, no difference.]” Overwhelmingly, respondents chose female (815 respondents, or 80.8% of the interviewed population). In addition, reversing the textual coding for *b_add_female* results in the distribution as shown in **Figure 3.2**: in this scenario, the number of respondents who chose “more effective” or “much more effective” remains nearly constant, with a slight increase in confidence in the ability of women officers to make the police more effective in dealing with crimes against women. Further exploration, and possibly conferring with the authors, would be required to determine with confidence whether the numerical or textual indexes are reverse-coded in this situation.

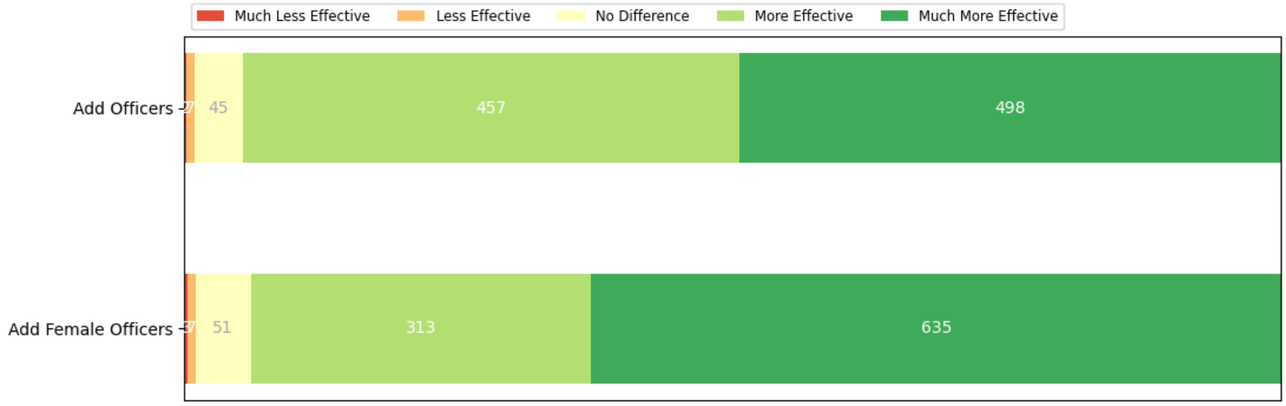


Figure 3.2

3.2 Question 2

In exploring this question, we focused on the admin data related to police station attributes and numbers of call and reports by month, particularly the files *Admin_wide_data* (a.k.a. *A_wide*) and *Police_station_personnel_data* (a.k.a. *P_station*). The key variable of interest in evaluating this question was the change in the number of dial100 calls made to the station between the baseline period and the endline period. To calculate this variable, we took the difference between *e_dial100_count* and *b_dial100_count* for each month at each station. After grouping the data by station, we took the mean of this difference over the 12 months provided, creating a set of 180 averaged changes, one data point per station. The heterogeneity data from *P_station* was then left-merged with our calculated data set on the "left."

To perform a quick examination of the data, we created a histogram of the mean difference in call numbers as calculated above, shown in **Figure 3.3**. The blue vertical line indicates the overall mean of changes in dial100 calls to stations that received treatment, while the red line indicates the same for stations that did not receive treatment. It is worthy of note that, while the mean of the control group is very close to zero, the mean for the treatment group, while small, is more negative. This does not support the initial hypothesis that treatment would lead to increased trust in a station, thus leading to an increase in calls to the station. It would, in fact, imply the opposite.

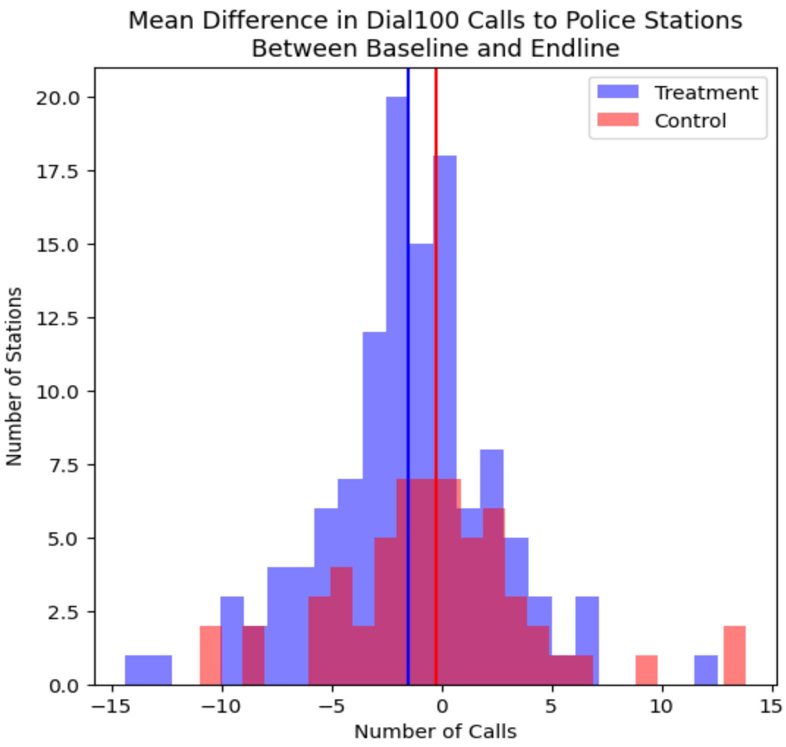


Figure 3.3

To test for the statistical significance of this difference, we used a permutation test to run 999 scenarios comparing the mean of the treatment group to that of the control group, then compared this distribution to the true difference, as shown in **Figure 3.4**. While the visual difference is distinctive, the p-value of the true difference is **0.059**, just above the threshold for statistical significance. As such, we cannot rule out the hypothesis that the difference in means is due to random chance rather than treatment arm.

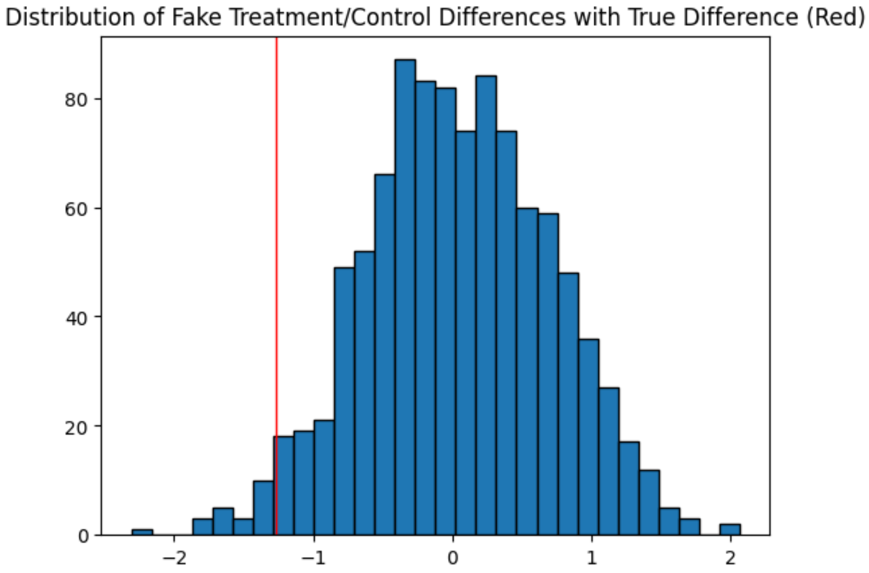


Figure 3.4

3.3 Question 3

The data used to answer this question was the general citizen survey (*Citizen_full data*, aka Citizen) and the police station user survey (*User survey_endline data*, aka User). Citizen was a very messy dataset. It had thousands of null values, but it had a total of 6,519 rows. The large number of rows made it feasible to simply remove the nulls in the columns that had the most missing values (visit indication (*e_visit*) and knowledge of women's help desks (*e_urja_knowledge*) at endline). It seemed important to know how people felt at both baseline and endline, so this decision seemed reasonable. That purge of nulls cut the dataset in half and there were still a few null values. Fortunately they were all numerical, so it was simple to fill in the mean of the column into these which allows the data to maintain trends and keep as much data as possible data. Once the mean was imputed into the null values, each column had 3,294 rows of data to use for analysis. Moreover, User was much cleaner since with less than 50 rows with missing values per column. The most important variables, the user satisfaction metrics, contained all the null values. Since all data was important, it was simple to drop the rows with empty values; the cleaned data ended up with 3,178 rows.

Citizen was important because it measured how people felt about the police at baseline and endline. It measured things like feeling of safety and confidence in police handling CAW on numerical scales. The main variables looked at in this dataset are *b_safety, e_saftey, b_pol_handling* and *e_pol_handling*. These variables are focused on because they give insight into the community perceptions of the police. Furthermore, the metrics were measured for before and after endline, so changes in opinions could be distilled from the data. Citizen is aggregated on the household level, so the baseline and endline differences between each household are present. The only caveat to that is since it is by household, sometimes different people filled out the survey (a husband and wife each filled out one of the surveys one for example). Moreover, the units for these variables are not exactly units per say. They are values on respective scales. Police handling is scaled from -1 to 1 but safety is scaled from one to four. Implement quality (a metric for the successfulness of a treatment *implement_quality*) is scaled from 1 to 9.

User also had metrics describing how people felt about police on a scaled basis, but rather than numbers used ordinal values. The main variables, *visitsats, comfort, respect, resolution,* and *fclitysats*, explain how users feel about the visit to the police station on variations of this scale: very negative, somewhat negative, somewhat positive, and very positive. While they lack hard numbers, they are very intuitive variables which made them great for analysis. This data is at the individual level, which made it a bit more straightforward than Citizen that could have had one row with various people's perceptions.

As seen in **Table 3.2**, Citizen presents substantial differences in means between baseline and endline police handling. For women specifically, there was a 0.19 jump that indicates women generally feel that CAW will be handled by police. Relative to its -1 to 1 scale, that is a 9.5% change from baseline. Another interesting thing to note is that the mean differences between men and women are small. One may expect them to have different perceptions on the police, but the data does not support that idea. Unfortunately though, safety scores went down for both women and men between baseline and endline. So, even though people feel more confident in the police, they felt less safe, which seems strange. The standard deviations on these also inspired me to look at these distributions more closely for police handling, since they are relatively large. **Figure 3.5** shows how there is a lot more positive values than negative on baseline and endline. Endline does see more high positive values and less negative values, but the scores are overwhelmingly popular, which does not align with the claim that people generally do not trust the police (the motivation for this question!).

Table 3.2

	e_gender	*Female*	*Male*
b_pol_handling	**mean**	0.35	0.41
	std	0.47	0.46
e_pol_handling	**mean**	0.54	0.50
	std	0.44	0.50
b_safety	**mean**	3.32	3.51
	std	0.47	0.44
e_safety	**mean**	3.19	3.23
	std	0.50	0.47

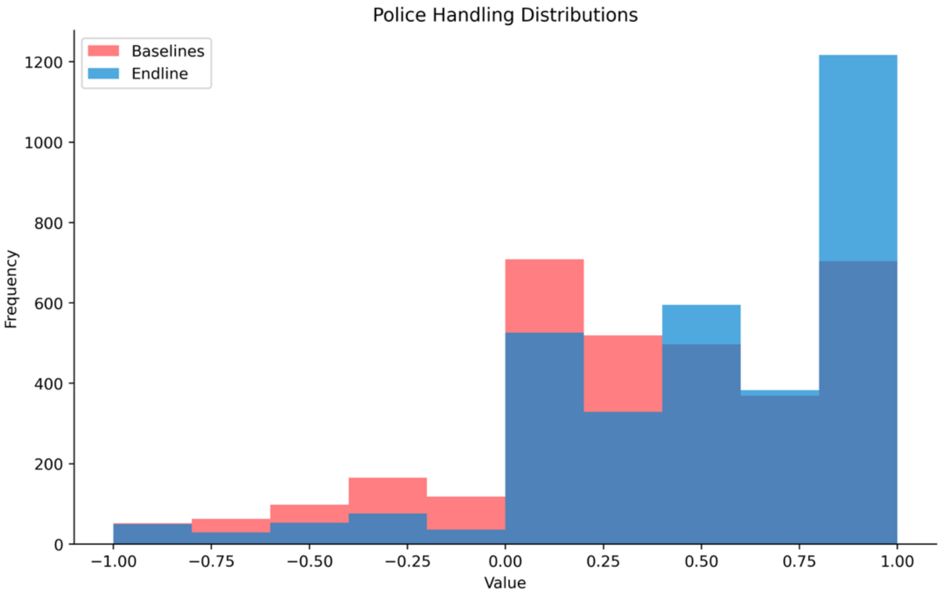


Figure 3.5

In User, that trend of positive scores continues. **Figure 3.6** shows the survey distributions for *visitsats*, or visit satisfaction scores, based on control and treatment groups. Almost 85% of the control group is satisfied and almost 88% of the treatment group is satisfied. Based on the data, people seem to have better satisfaction scores with the treatment, but people are generally satisfied with their visit to the station regardless. This finding also directly opposes what the authors claim in the paper about a lack of community trust in the police: "…social barriers, including stigma, can deter women from reporting crimes, as can low trust in the police" (191, Sukhtankar). *Low trust* and a satisfaction rate over 80% do not align.

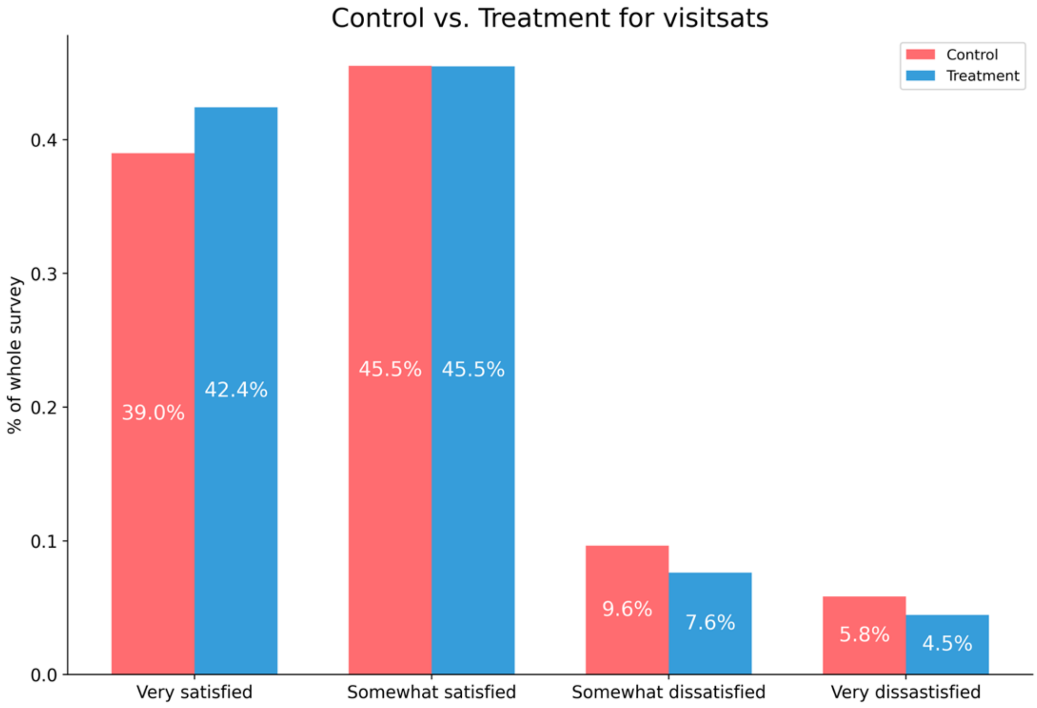


Figure 3.6

3.4 Question 4

The data used to answer this question came from administrative data detailing the number of FIRs and DIRs stations reported (*Admin_long data.dta*, aka Admin), CCTV (closed-circuit television) data (*Cctv_full data.dta*, aka CCTV), police personnel data (*Police_personnel data.dta*, aka Personnel). The most important vairbales from Admin are FIR and DIR counts (*fir_overall_count* and *dir_count*) which gives supporting evidence and context for this question. Admin had no nulls values, which was a nice change of pace. CCTV on the other hand came with many null values in important variables such as the average amount of female foot traffic at endline (*eavg_women*). Since they are numeric, the nulls were imputed with the mean average female foot traffic. Then all superfluous (and there were many of them) columns were discarded which left CCTV with 3,416 rows of non-null data (so no rows were discarded). Then Personnel was cleaned in the same fashion as in Question 1 above. CCTV and Personnel were then joined on the police station ID to create one dataset known as CCTV_Personnel. The combined table allowed the dataset to describe how many people on average walked into a given station on a given day and what the staff makeup in that station looked like, which addresses the question quite nicely.

Admin gives a helpful insight into the reporting, which offers context to the landscape of this question. Admin collected data monthly per station. **Tables 3.3 and 3.4** shows average FIR and DIR counts based on the *group*. DIRs see a drastic increase when there is a women's help desk or female officers present. But the number of FIRs is almost identical across the treatment groups. While they seem to balance each other out, it is important to note there are a lot more FIRs filed than DIRs. So, while it is a large relative jump, the lack of movement in FIR count seems more telling of the situation. It is great to see a few more DIRs filled out per month, but it is still less than one per month on average.

Table 3.3

	group	control	regular mhd	women officers
dir_count	**mean**	0.02	0.79	0.71
	std	0.35	2.75	2.27
	min	0.00	0.00	0.00
	max	8.00	33.00	32.00

Table 3.4

	group	control	regular mhd	women officers
fir_overall_count	**mean**	33.59	33.36	33.03
	std	24.34	23.60	21.28
	min	0.00	2.00	0.00
	max	271.00	419.00	235.00

CCTV_Personnel can demonstrate how different factors, especially those staffing a station, affects the amount of people that come to the station; they key variables are (*eavg_women, dayofweek, timeofday,* and *e_female_staff, e_female_officers*). The various reforms should help all staff members deal with CAW and filing of FIRs and DIRs, but that may not get people in the door. Getting people to the station is a good measure of how many new people are going to the station with their cases. It may not be for GBV, but it may. Most importantly though, if community perceptions are on the rise, more people will go to the station (that is what is being assessed). FIR and DIT counts may indicate a growth in reporting but cannot decipher if the growth is due to new people trying to use police services. Moreover, The data is taken on a daily level for each station and gives a baseline and endline average foot traffic for both women and men. The units for this dataset are people, but the entire dataset aggregates into averages by day of the week and time of the day.

The question requires exploration into how women have reacted to the presence of female officers. **Table 3.4** shows that endline foot traffic for women is generally higher than baseline on every day of the week. The improvements seem small, but those increments over a year indicate some decent improvements. But **Figure 3.7** tells a less promising story. The scatter plot displays the number of female staff by the average foot traffic, colored by control and treatment groups. The trend line indicates that there is no trend (it is almost completely horizontal) and treatment also does not seem to have any impact. The number of officers seems to have no relationship with the number of women coming to the station. Another very important fact to point out on this graphic is that the highest number of female staff members in a station is **15**. Also, these are *female staff*, which means they are in almost any role at the station and below rank Assistant sub-inspector (ASI). That is not a ton of women, and it seems impractical to think a handful of women will enact major societal change.

Table 3.5

	dayofweek	Fri	Mon	Sat	Sun	Thu	Tue	Wed
bavg_women	**mean**	13.59	13.54	13.58	12.24	12.61	13.74	13.75
	std	13.67	13.67	13.75	12.14	12.62	13.64	13.07
eavg_women	**mean**	14.96	16.14	14.08	14.12	14.16	14.06	15.31
	std	13.64	15.02	13.39	15.55	12.66	11.59	14.61

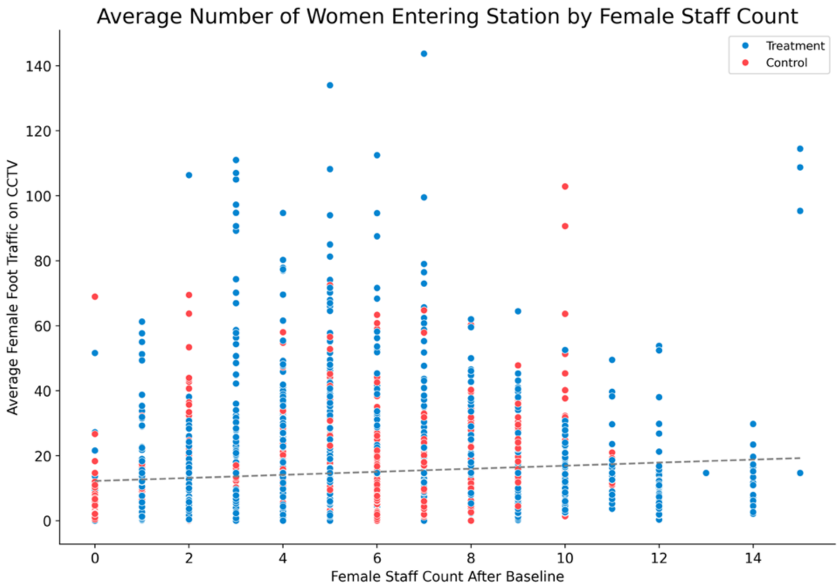


Figure 3.7

4. Modelling

4.1 Question 1

First, we examined the question of if there is a correlation between the police survey responses and the number of women at a given station via building a linear regression model. The predicted value would be the ratio of women staff to total staff at the respondent's station, while the predictive variables included the respondent's gender, their indexed responses to each survey question, and station-level heterogeneity data including population served and urban/non-urban location. A linear regression model seemed appropriate when looking to model a continuous variable, such as the fraction of women on staff at each station. The results of this model are shown in **Figure 4.1**, with blue dots representing the predicted (y) vs. actual (x) fraction of women at a respondent's station, and the red dashed line representing the ideal model x = y. Visually it is clear that there is little to no correlation found; indeed, the R^2 value was found to be 0.0312.

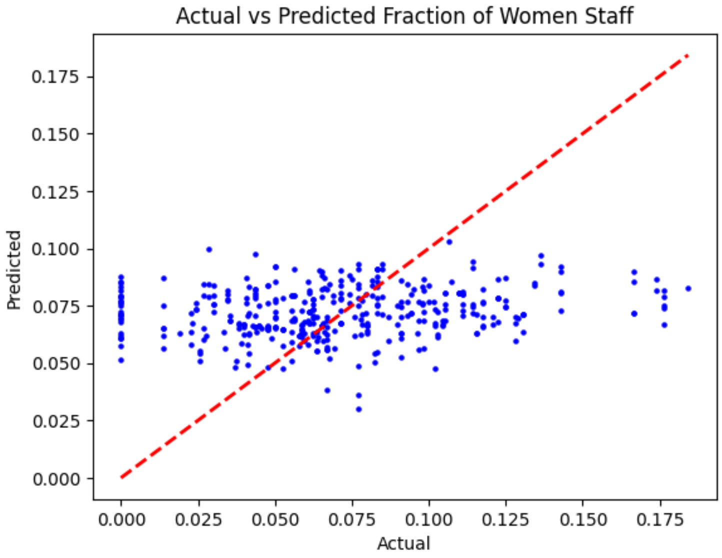


Figure 4.1

To explore these data from a different angle, a logistic regression was also developed, this time to try to use survey results and heterogeneity data to predict the respondent's gender. It was found that the accuracy of this model was 0.675, while its F[1] score was 0.279, suggesting above-random accuracy but very poor recall. Observing the breakdown of guesses vs. true values in **Figure 4.2**, the model correctly predicted the majority of male respondents as male (233 correct of 250 total instances). However, it also incorrectly predicted the majority of female respondents to be male (97 incorrect of 121 total instances). The model was frequently correct only in that it guessed male the majority of the time, and the majority of the respondents were male.

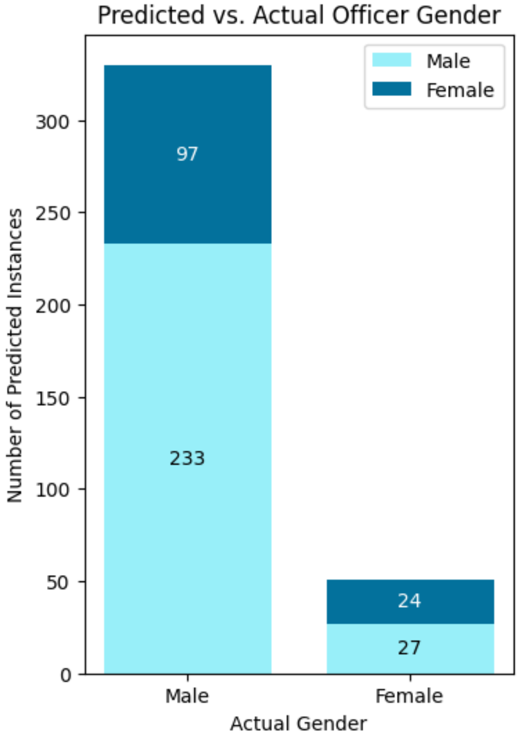


Figure 4.2

This leads to one of the weaknesses of this data set, which appears across multiple facets of our analysis: the police survey data set is heavily unbalanced towards male respondents over female respondents, in reflection of the fact that there are many more men than women working at the stations in this study. Of respondents to the police survey, only 225 of 1904 respondents (11.8%) were female. **Table 4.1** provides the breakdown of the fraction of women staff working at each of the 180 police stations in the study; note that the maximum value is below 20%, or less than 1 in 5 staff. It may be possible that the presence of women in a police station may improve the opinions of police officers regarding the importance of cases involving women *when there are a sufficient number of women present*, but that the data does not hold up this hypothesis may very well be due to the fact that there are so few women involved in the police stations of this study.

Table 4.1

	Fraction Female Staff
count	180
mean	0.070280
std	0.039593
min	0.000000
25%	0.043478
50%	0.065396
75%	0.094564
max	0.184211

4.2 Question 2

In this case, the question at hand regarded whether a station receiving treatment resulted in a noticeable change in the number of dial100 calls to that station between the baseline and endline periods. As such, it seemed reasonable to create a logistic regression model to attempt to predict the treatment arm, using the dial100 call count and heterogeneity data as inputs. As shown above (see section 3.2), there is not a statistically significant difference in the mean change in dial100 counts between the treatment and control branches, so there was little confidence in the efficacy of the logistic regression model. These expectations were indeed borne out, with an accuracy of 0.556 and an F1 score of 0.636. **Figure 4.3** demonstrates the resulting distribution, illustrating that the model performed, as the accuracy score would suggest, just barely above guessing.

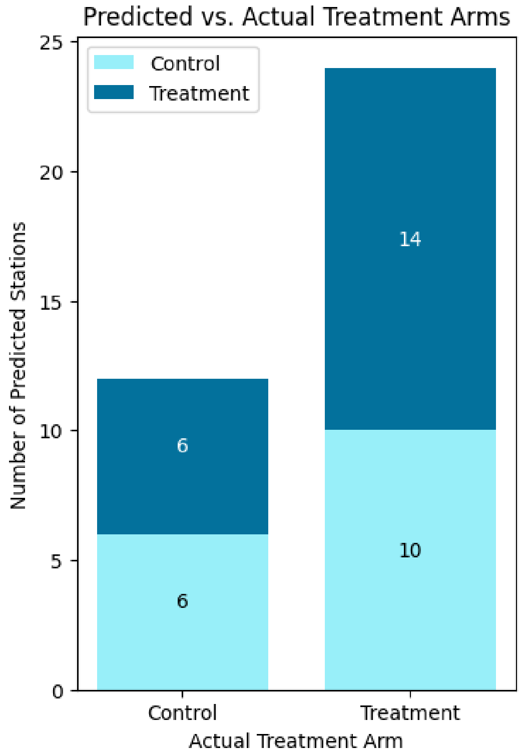


Figure 4.3

However, it is worth noting that 180 stations make up a relatively small data set to try to train and test a model; the confusion graph above only contains 36 test points. To see if this scarcity of data could be overcome, the data was reprocessed to have, rather than one average call difference per station, 12 data points per station, with one for each month, making up a data set of 1980 rows. The results of this model are shown in **Figure 4.4.** There was no improvement in performance found by granulating the data, and in fact it performed marginally worse, with an accuracy of 0.480 and an F1 of 0.542. These results are not surprising, given all the analysis we have performed thus far. According to this data, the change in dial100 calls each station experienced between the baseline and endline periods was not statistically significantly different between the treatment groups and the control group.

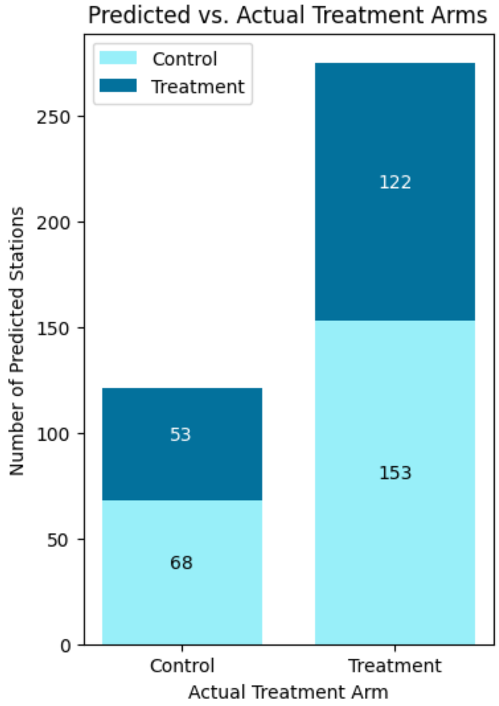


Figure 4.4

This finding aligns with those of Sukhtankar et al. (2022): as noted by the authors, while the number of CAW FIRs filed increased between baseline and endline at women-run WHD stations, "increases in registration of CAW cases do not come at the cost of reductions in other cases, with no discernible spillover effect on the overall number of FIRs or on other kinds of police reports" (p. 194). Further, the authors claim that any difference between the number of CAW cases filed is a result of changes in police behavior, but not by any observable difference in the rate of women reporting crimes to the police. If this is so, it would be highly unlikely to observe a difference in the rate of calls to stations regardless of treatment. The treatment targeted police behavior, not the behaviors of the public.

4.3 Question 3

Community perception metrics were naturally the most important pieces of data for understanding how treatment affects how the community views the police. Therefore, we built a logistic regression model to predict a community perception variable to see if treatment, or particular reforms, had strong relationships with it. User's *resolution* variable, which measures people's perception that their case will be resolved seemed like a great option to predict since it gets to the root of the question's motivation; people seem to not report cases because they do not trust the police, so if treatment is reversing that disposition, it is worthwhile to implement.

For the logistic regression to work, some decisions had to be made about how to transform the available data. We first needed to drop the other community perception metrics (*visitsats, comfort, respect,* and *fclitysats*), since they are fairly correlated with *resolution* (much more than any other varibales) (pictured in **Figure 4.5**). More importantly than the correlations though, they are products of the survey, along with *resolution*, and are not related to the treatments or what should make people feel better or worse about resolving their case. Station and user IDs were also taken out since the model should not attempt to pick up trends on particular people or stations, but rather generalize to understand trends within *resolution* and the User dataset. These are the variables that ended up being used in the model: *consented, gender, urban, population, training_score, implement_quality, comm_outreach_strength, regular_whd,* and *women_whd.* The variables describe something about the treatment, the context in which the station exists, or something about the user (which all seem relevant to how they may answer the survey). Many of these variables are categorical so they needed to be transformed into binary values for the model. For the ordinal target *resolution*, rather than use an ordinal logistic regression, the four categories were combined into two, confident or unconfident. An important note here is the distribution of *resolution*. Seen in **Table 4.2**, there are 2824 confident records to the 354 unconfident (88% of people are confident).

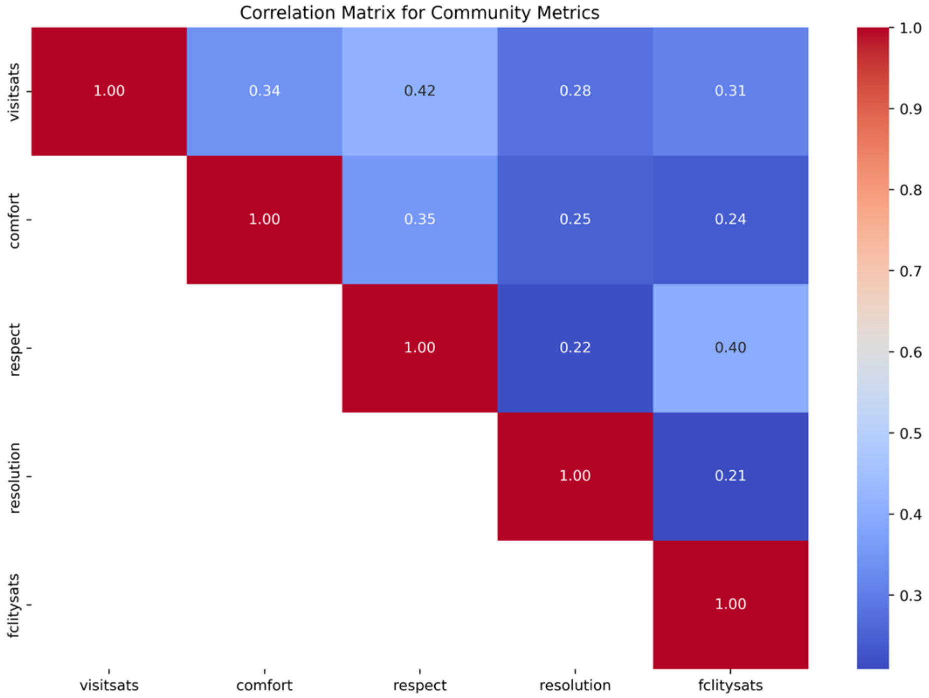


Figure 4.5

Table 4.2

resolution	***count***
Confident	2824
Unconfident	354

Then the model needed to be built. An important feature of the model was that it attempted to *balance* the categories it predicted, which forced the model to learn about the small amount of unconfident people and the large amount of confident people. After normalization and fitting, the regression was 58% accurate (moderately better than guessing) and has a F^1 score of 0.71 (which indicates the model is decently precise). But, seen clearly in **Figure 4.6**, this model clearly did not pick up any trends. It can predict "Confident" often (335 out of 380), but almost never predicts "Unconfident" correctly (31 out of 256). That indicates that it has almost no understanding of what makes someone unconfident in the police resolving their case, and mostly guesses it is confident. This makes sense though because there are so many more "Confident" entries. The model was built without the *balancing* before, but it never predicted the anyone to be "Unconfident" since it was extremely accurate when it guessed "Confident" every single time. But with the balancing, it shows that there are no strong trends between *resolution* and any of the predictors. **But** part of the issue was that the community perceptions were already good. The data, again, shows that people have a positive disposition towards the police, which directly opposes Sukhtankar and her co-authors claims.

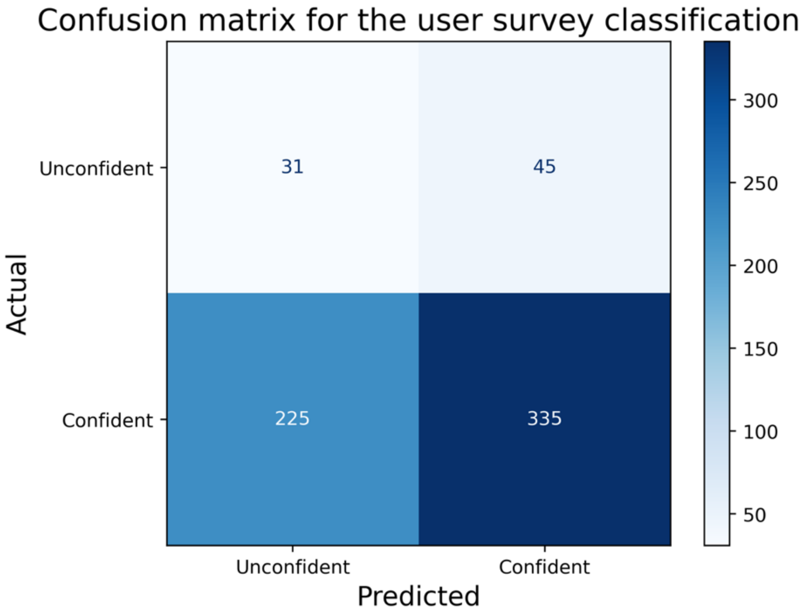


Figure 4.6

Based on the data, the answer to the question is a firm no. Treatments, while they seem to have some positive relationship with community perceptions, do not seem to consistently alter people's perceptions of the police, which the authors would agree with: "Citizens' attitudes toward the police, … did not shift, although this may reflect the short duration of the intervention" (196, Sukhtankar). But, based on this analysis, the claim that people mistrust the police is invalid. The data clearly shows that people are overwhelmingly positive about the police. Mostly positive feedback on police handling from the public (Citizen) and over 85% visitor satisfaction (User) do not exactly insight concern. But the author's intuition makes sense. There is precedent for police not doing well with CAW, especially in India. Again, based on the Indian government's National Family Health Survey, "77 percent have never sought any help nor told anyone about the violence they experienced" (648, International Institute for Population Sciences). Therefore, this analysis likely supports that this data contains bias. The surveys seem to have brought on a group of people that do not reflect society adequately. Most of the people that fill out surveys must be people that like the police, which makes sense. If one likes something, they will take the time to show their gratitude via the survey. But the majority, those who do not interact with the police often or had a subpar experience with them, may not give up their time. That idea is speculative offers a reasonable explanation about why the data does not match the societal consensus. The survey data fails to illuminate the opinion of the silent majority and, hence, seems unequipped to accurately explain how the community perceives the police.

4.4 Question 4

The main indicator used to address this question is the average foot traffic of women at endline (*eavg_women*). If the number of female officers caused more women to come into the station, that would show that community perceptions have increased, and more reports would have the potential to be filed. There is evidence from other questions that community perceptions for not change when treatments are implemented, but this is a different angle that could yield different results. Therefore, we decided to build a linear regression model to predict average foot traffic (*eavg_women*). The model will illustrate the relationship between female foot traffic at the stations and female the number of officers, along with many others.

Like logistic regression, linear regression requires some data preparation. The next step is to pick which variables the model would consider. As this is a joined dataset, there were a lot of columns (52 to be precise), so many of them could not be included in the model. The variables chosen were ones that seemed like they should influence the average number of women coming into the station: *dayofweek, timeofday, e_male_officers, e_female_officers, regular_whd, women_whd, population, training_score, implement_quality,* and *comm_outreach_strength*. These variables cover logistical things like timing of visits and number of people in the areas, but also the personnel and treatment factors. Variables such *dayOfWeek* and *timeofday* are ordinal, so they needed to be encoded as binary variables to work in the model (they were split into multiple binary columns). Together, those variables should be able to explain make makes people come to the station. The data also dropped all outliers before modelling; there were some that threw off the model considerably in earlier renditions. Then the data was split and normalized, and the model was built. The R^2 score was 0.1501, which means the variables in the model have little to no predictive power over *eavg_women*. In a more technical sense, the model is horrific. **Figure E** shows a scatter plot of the actual versus predicted values for the model and it shows how the model attempts to guess the mean to maximize accuracy since it cannot pick up any trends from the data.

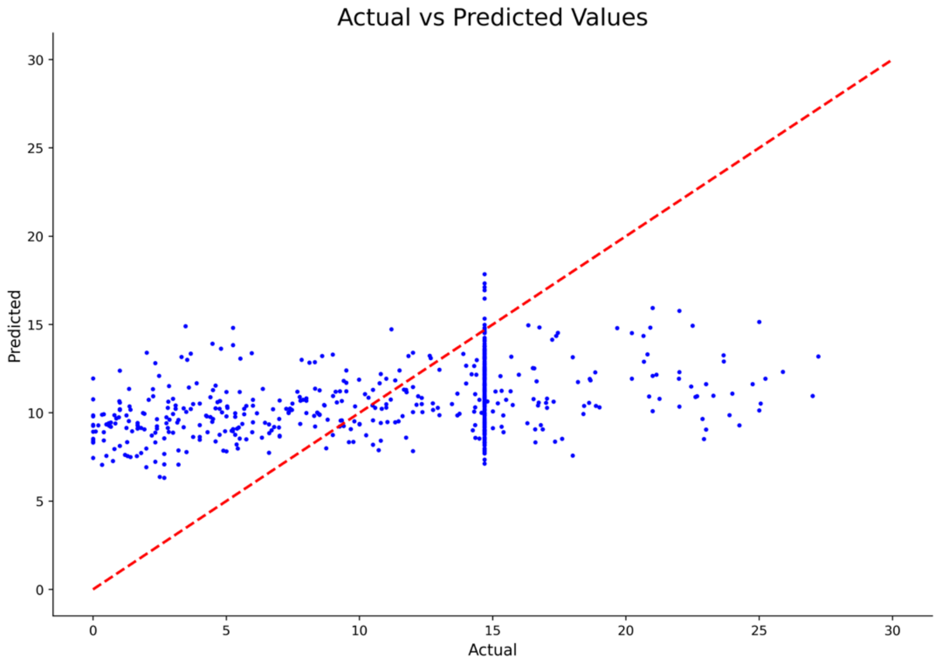


Figure 4.7

The presence of female officers does not increase the number of women that will come into the station. That also aligns with the lack of increase in FIRs seen in Admin during EDA. It seems like the the female officers did not make much of an impact at all. Perhaps it is due to the idea "that gender differentials in police behavior are diminished as female officers operate within male-dominated policing cultures, noting a lack of increased gender sensitivity among female officers who, like their male counterparts, often blame victims or dismiss their claims" (191, Sukhtankar). Female officers attempt to emulate the bravado their male counterparts endorse, leaving them no just as biased as the male officers. While there are counter arguments to that idea, it is unsurprising that the power of women in the stations is diminished in such a patriarchal culture. But that **is** the problem.

Figure 4.8 shows average female foot traffic by the number of female officers (women of rank ASI or higher). The graphic supports that there is no relationship between these variables and the most women a station will have been six. While there are not a ton of people in this officer category overall, women are still a police minority across the board. Women have no chance to affect societal changes in police perceptions because there are not enough of them. Sofia Amaral's article shows that WPS drastically improved reporting (29% increase). So it seems having a lot of women and women in leadership roles makes the difference. Only then society will see that women have made a difference in their local police and are willing and able to bring their unique skillset to GBV cases. That is when women will feel more apt to report their cases. Societal opinions will adapt when real and serious change occurs around it.

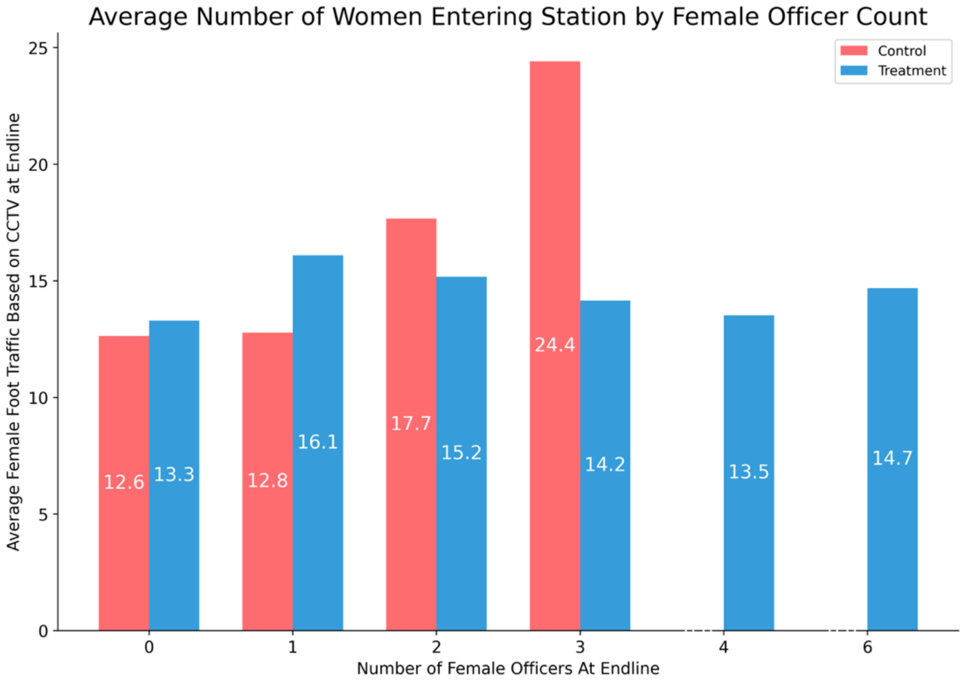


Figure 4.8

5. Conclusions and Considerations

Over the course of this project, we examined a variety of questions and possible correlations, and ran into several limitations of the data set itself. In looking for a correlation between the fraction of women staff at a station and the opinions of the officers at that station, confusions regarding reverse-coding surfaced. In attempting to model the relationship between the presence of female officers and community perceptions of the police, it was found that the data is skewed heavily toward positive attitudes, which potentially flags an issue in sampling methods or other contributors to bias. And in several of our lines of research, a major limitation of this data was found to be the extremely low presence of women staff or officers at the treatment and control police stations. This hinders attempts to identify the relationship between the presence of women officers and any other metric from this dataset; there is simply too little data on women officers or stations with high numbers of women officers to make meaningful correlations.

With these limitations in mind, we proceeded to attempt to model different relationships in this data set and found no correlation between any of the metrics we compared. This is, in of itself, still valuable information. We found that the presence of a WHD or a women-run WHD had no effect that we could find on: the number of dial100 calls made to the station; the number of women walking into the station; or public confidence in the police. As discussed above, these findings support those of Sukhtankar et al., in that we found no behavioral or attitudinal difference in the public that might account for the observed increases in DIRs and FIRs filed at stations with a WHD. The absence of these other factors adds support to the authors' claim that the changes seen are all on the part of the officers' behavior, thanks to exposure and additional training on the importance of addressing CAW in their communities.

While working with data this complex and sensitive, it is important to consider how the public nature of this data could impact those whose data is contained in it. A main area of concern is the privacy and safety of the women reporting their experiences of CAW, but all participants deserve the same level of respect and care. This includes making sure that data such as survey responses cannot be traced back to an individual. For example: in the police attitudinal survey, the only demographic information of each respondent that was taken was their gender. However, also included in each entry was the station code for their station. This is obviously necessary for using these data in further analyses, but it is worth noting that the table of station-level information includes the breakdown of women and men staff and officers at each station, measured at baseline and endline. It is thus a trivial matter to merge the tables and identify all rows for a female respondent from a station with only one female staff member: in fact, there are 19 such rows in this dataset, using baseline measurements. Along with population served and urban/non-urban status, there is very likely enough information about the size and officer/staff gender breakdown in this data set to hunt down, with a little searching, the exact female police staff member who answered the survey. It would be even easier for a member of the police force to identify a given station, and thus a given respondent.

While in this case, the data gathered about each police respondent was only about their opinions on how their station handles CAW, there is still a possibility of harm if a coworker or superior were to identify a female staff member's response and decide they reflected poorly on the station, and punish her socially or economically. And this is a good model for how easy it is to let identifiable information slip into the public, which could have much more harmful effects if the data set were to contain, for example, medical or credit information. Subsequently, it is good practice to either further anonymize or aggregate data when the sample size is so small (such as a single female officer at a station).

There is much to be learned by exploring this data set, both from the data itself and from the experience of processing, examining, modeling and visualizing it. We have learned much about the pitfalls of small sample sizes and biased samples, how to handle null results, and how to look for potential ethical problems in places we might not expect. There is much more still to learn, both from this data and from our future studies – but this has been a very educational start.

6. Works Cited

Amaral, S., Bhalotra, S., & Prakash, N. (2021). Gender, Crime and Punishment: Evidence from Women Police Stations in India. *IZA – Institute of Labor Economics.* docs.iza.org/dp14250.pdf

International Institute for Population Sciences (IIPS) and ICF. (2022). National family health survey (NFHS-5), 2019-21: India: Volume I. IIPS, Mumbai.

Morrison, A., Ellsberg, M., & Bott, S. (2007). Addressing Gender-Based Violence: A Critical Review of Interventions. *The World Bank Research Observer, 22*(1), pp. 25–51. https://www.jstor.org/stable/40282335

Sukhtankar, S., Kruks-Wisner, G., & Mangla, A. (2022). Policing in patriarchy: An experimental evaluation of reforms to improve police responsiveness to women in India. *Science*, *377*, pp. 191- 198. https://doi.org/10.1126/science.abm7387

7. Appendix

7.1 Tables

Table 3.1

b_add_officer	
count	1009
mean	4.429138
std	0.632442
min	1
25%	4
50%	4
75%	5
max	5

b_add_female	
count	1009
mean	4.555996
std	0.652442
min	1
25%	4
50%	5
75%	5
max	5

Table 3.2

	e_gender	***Female***	***Male***
b_pol_handling	**mean**	0.35	0.41
	std	0.47	0.46
e_pol_handling	**mean**	0.54	0.50
	std	0.44	0.50
b_safety	**mean**	3.32	3.51
	std	0.47	0.44
e_safety	**mean**	3.19	3.23
	std	0.50	0.47

Table 3.3

	group	**control**	**regular mhd**	**women officers**
dir_count	**mean**	0.02	0.79	0.71
	std	0.35	2.75	2.27
	min	0.00	0.00	0.00
	max	8.00	33.00	32.00

Table 3.4

	group	control	regular mhd	women officers
fir_overall_count	**mean**	33.59	33.36	33.03
	std	24.34	23.60	21.28
	min	0.00	2.00	0.00
	max	271.00	419.00	235.00

Table 3.5

	dayofweek	Fri	Mon	Sat	Sun	Thu	Tue	Wed
bavg_women	**mean**	13.59	13.54	13.58	12.24	12.61	13.74	13.75
	std	13.67	13.67	13.75	12.14	12.62	13.64	13.07
eavg_women	**mean**	14.96	16.14	14.08	14.12	14.16	14.06	15.31
	std	13.64	15.02	13.39	15.55	12.66	11.59	14.61

Table 4.1

	Fraction Female Staff
count	180
mean	0.070280
std	0.039593
min	0.000000
25%	0.043478
50%	0.065396
75%	0.094564
max	0.184211

Table 4.2

resolution	***count***
Confident	2824
Unconfident	354

7.2 Figures

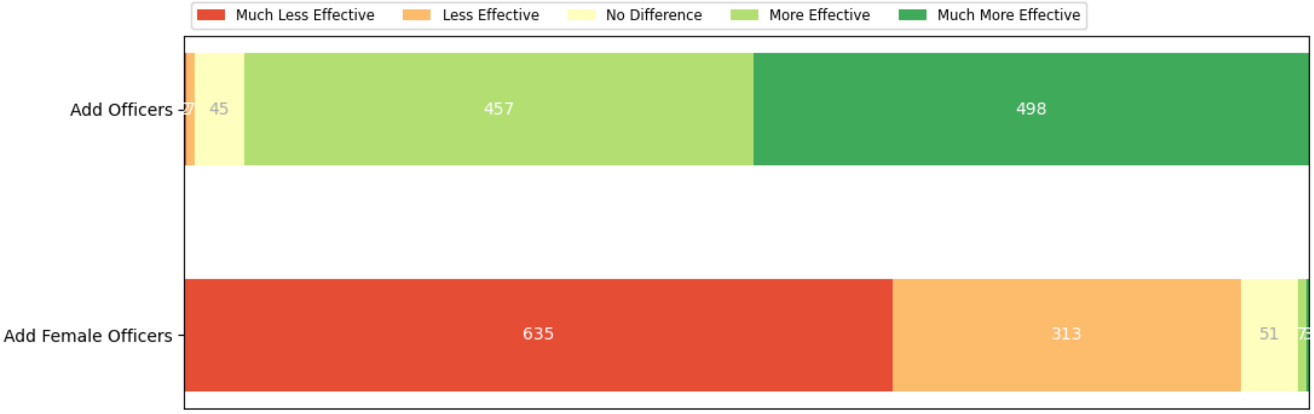


Figure 3.1

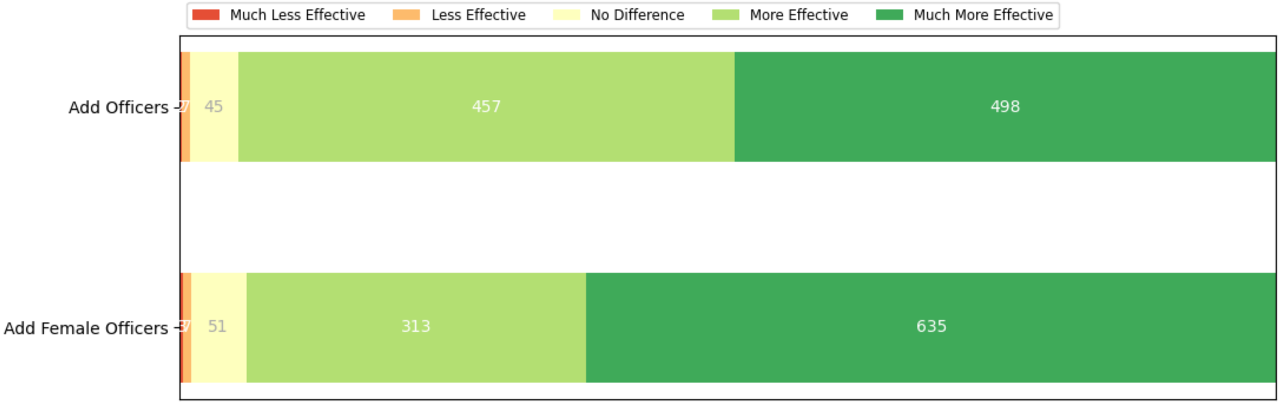


Figure 3.2

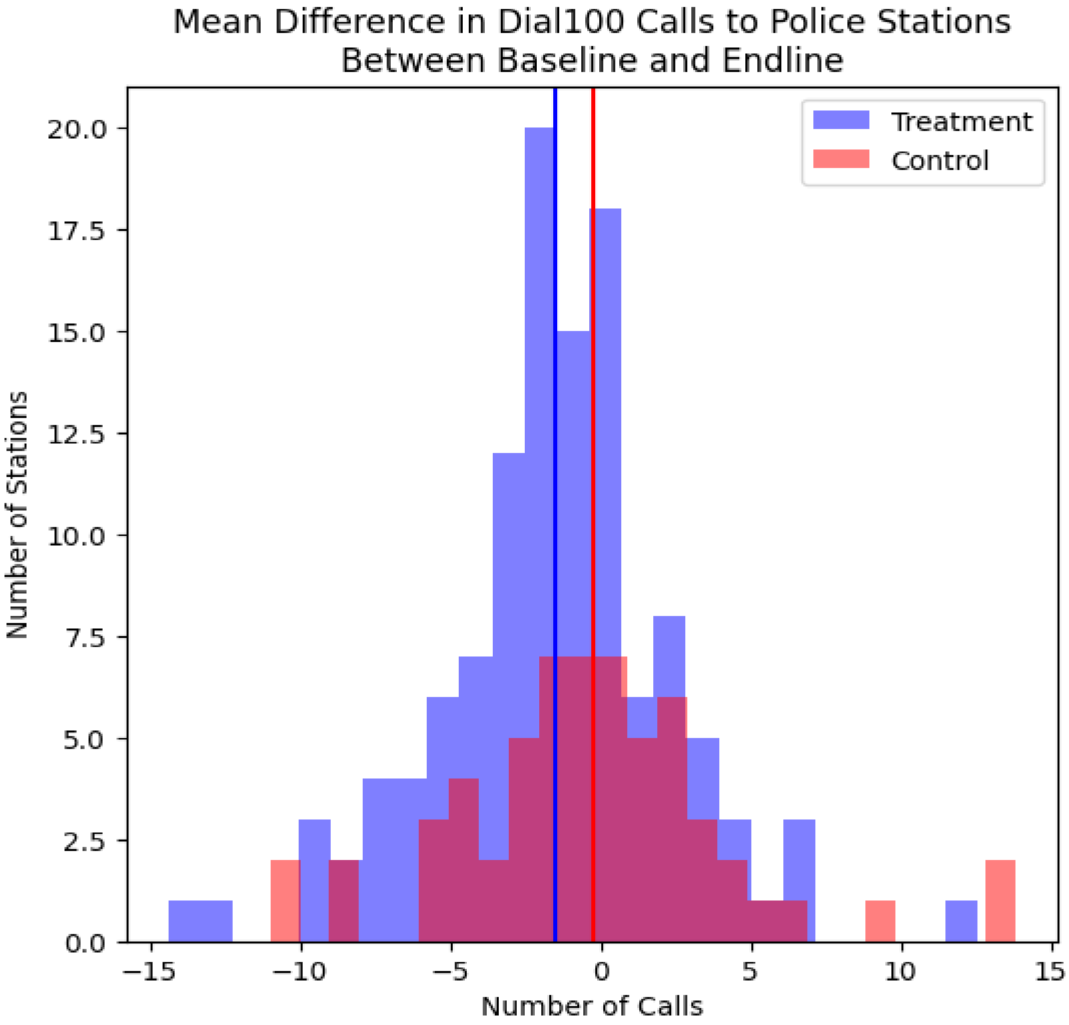

Mean Difference in Dial100 Calls to Police Stations
Between Baseline and Endline
Treatment
Control
Number of Stations
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
−15
−10
−5
0
5
10
15
Number of Calls

Figure 3.3

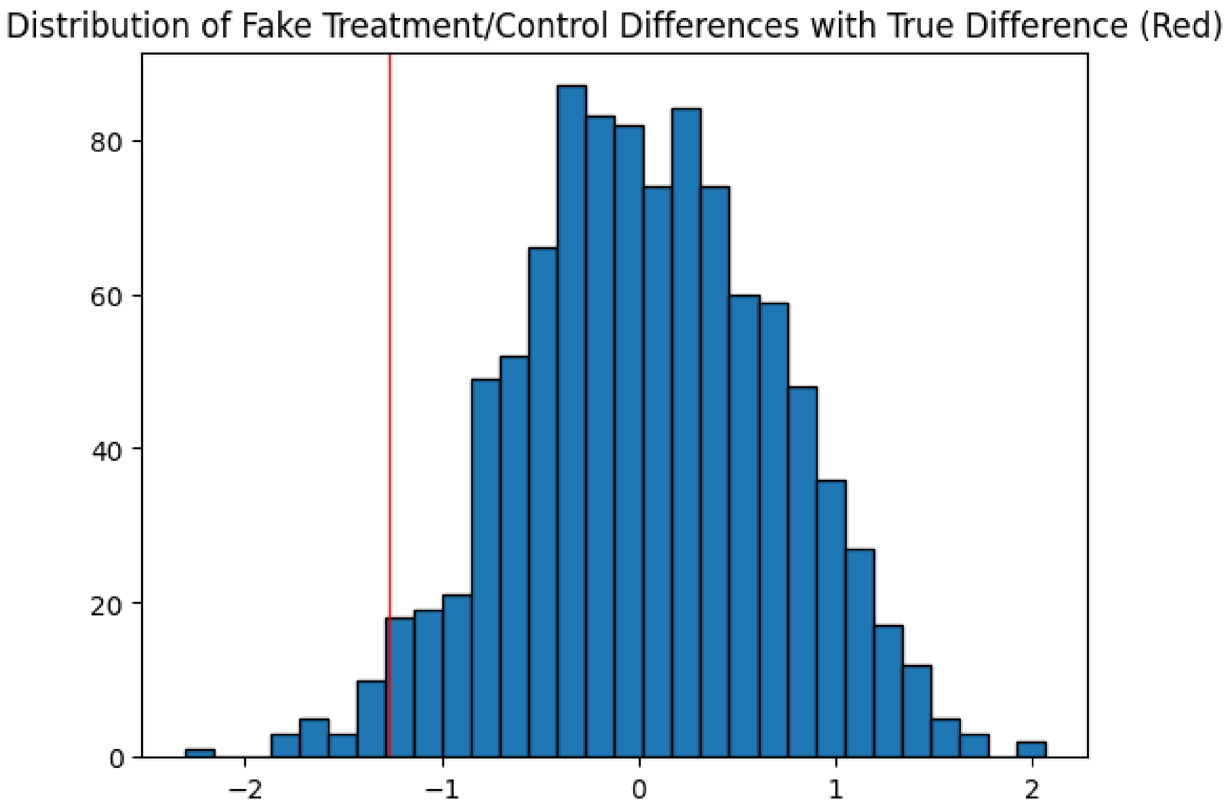


Figure 3.4

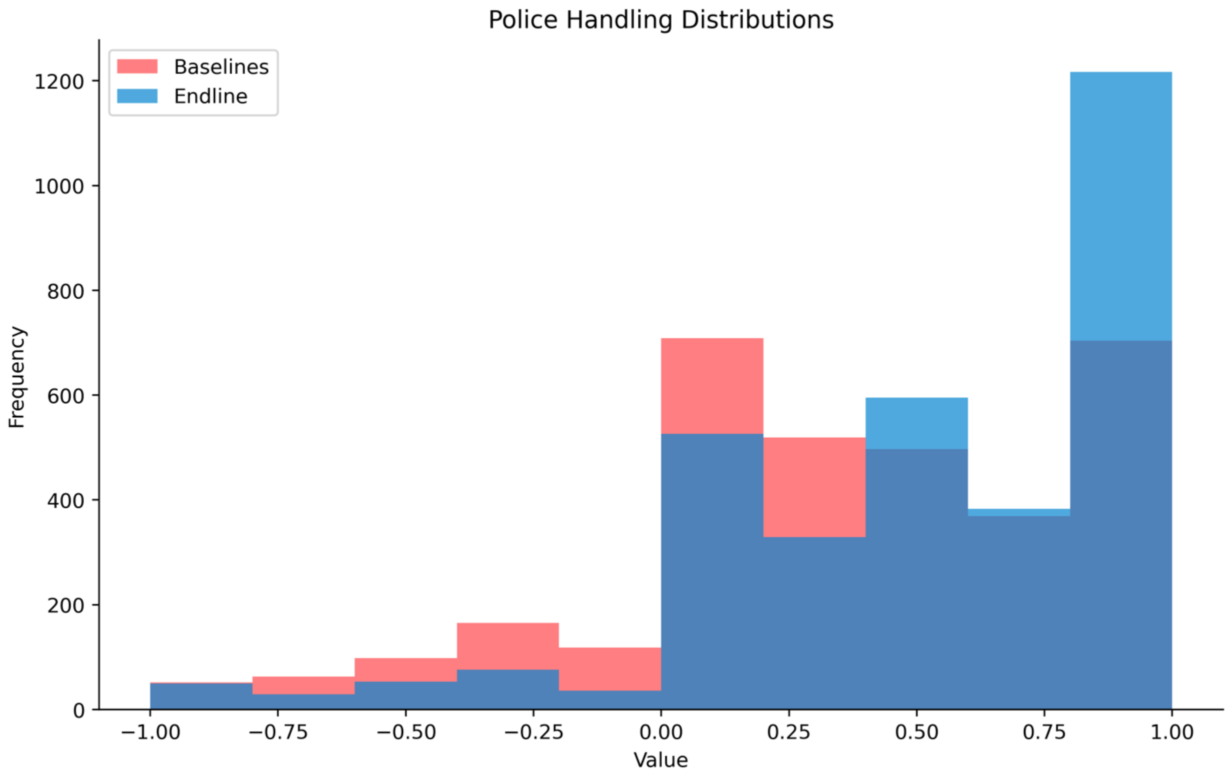


Figure 3.5

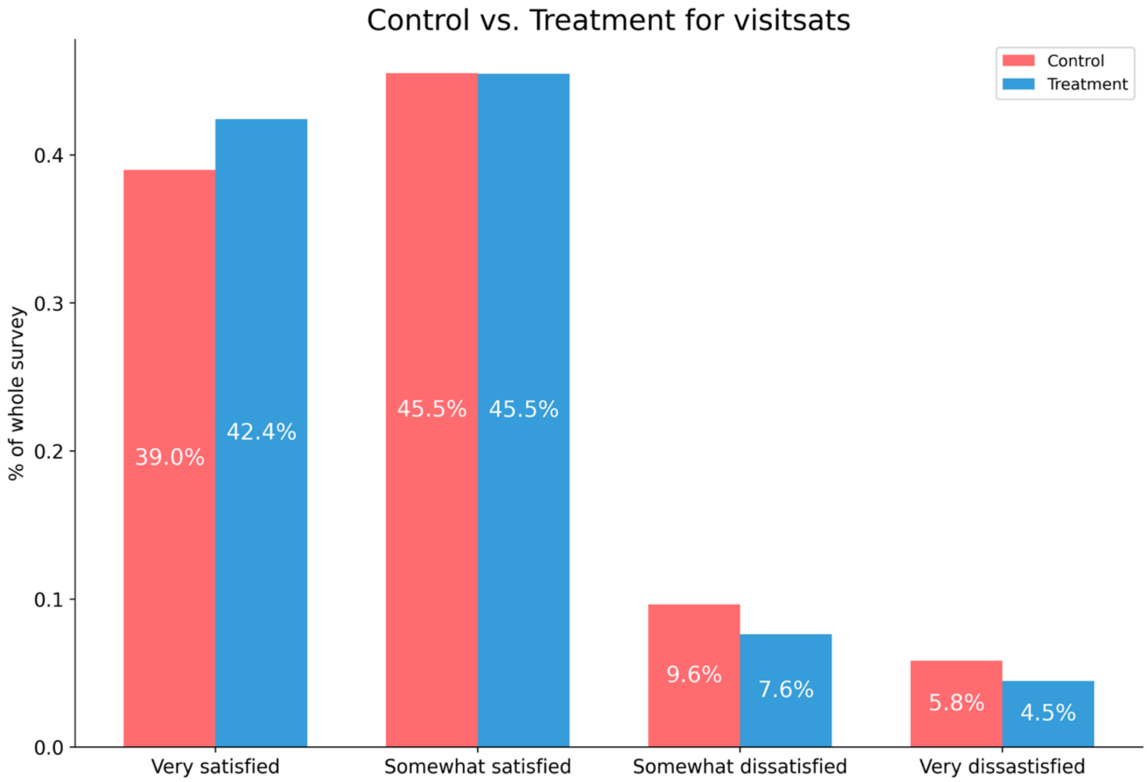


Figure 3.6

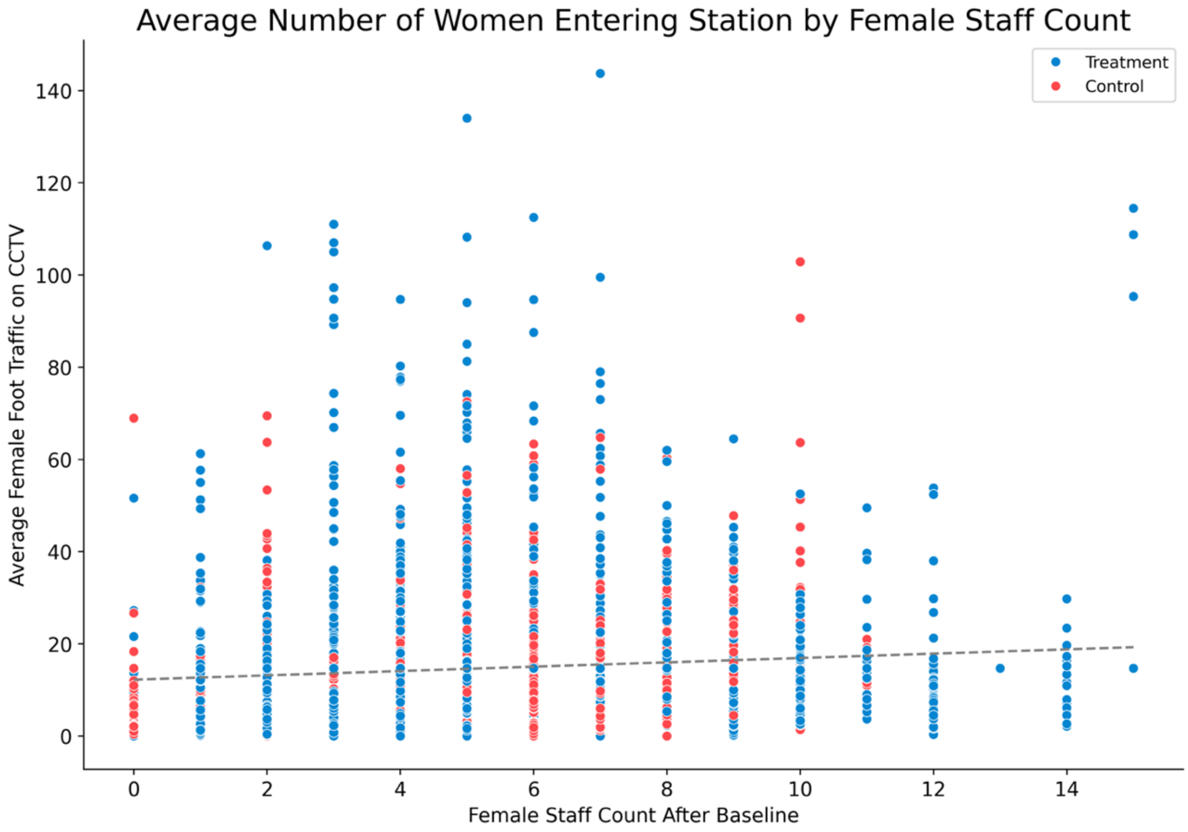


Figure 3.7

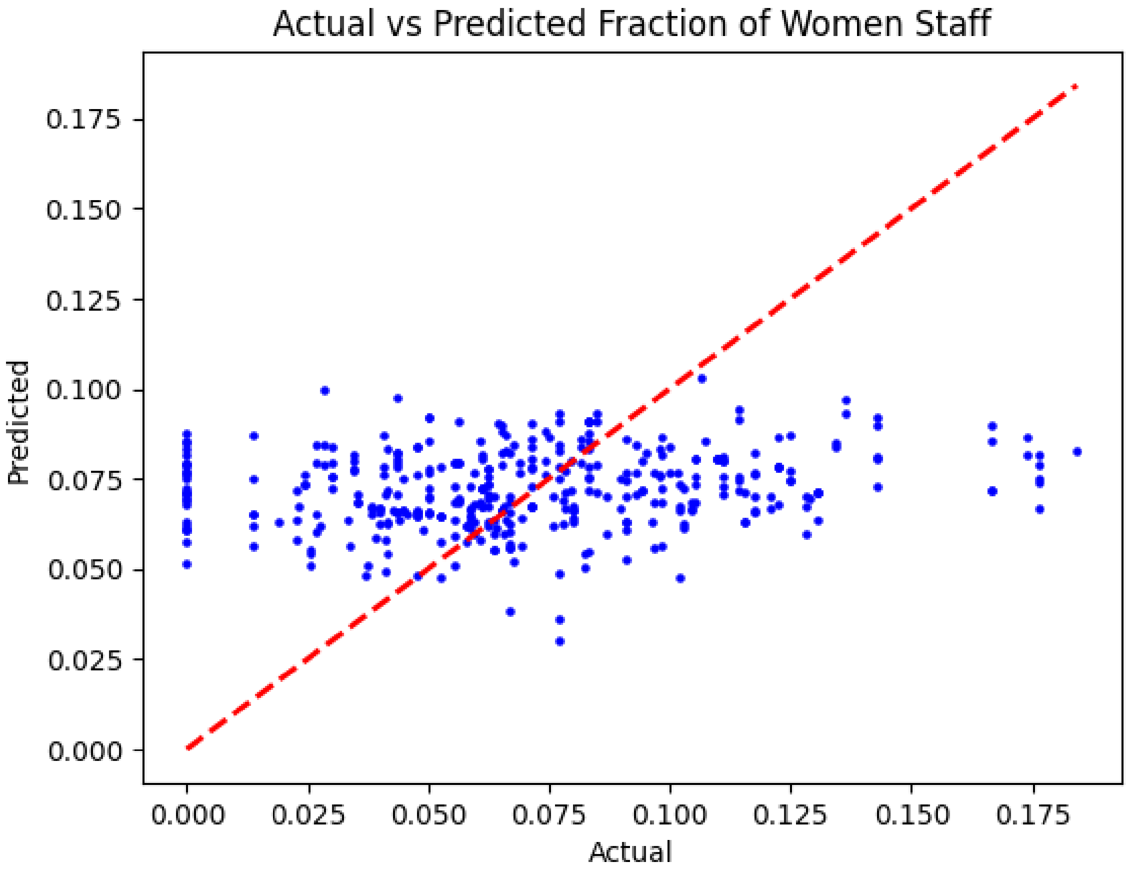

Actual vs Predicted Fraction of Women Staff
Predicted
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Actual

Figure 4.1

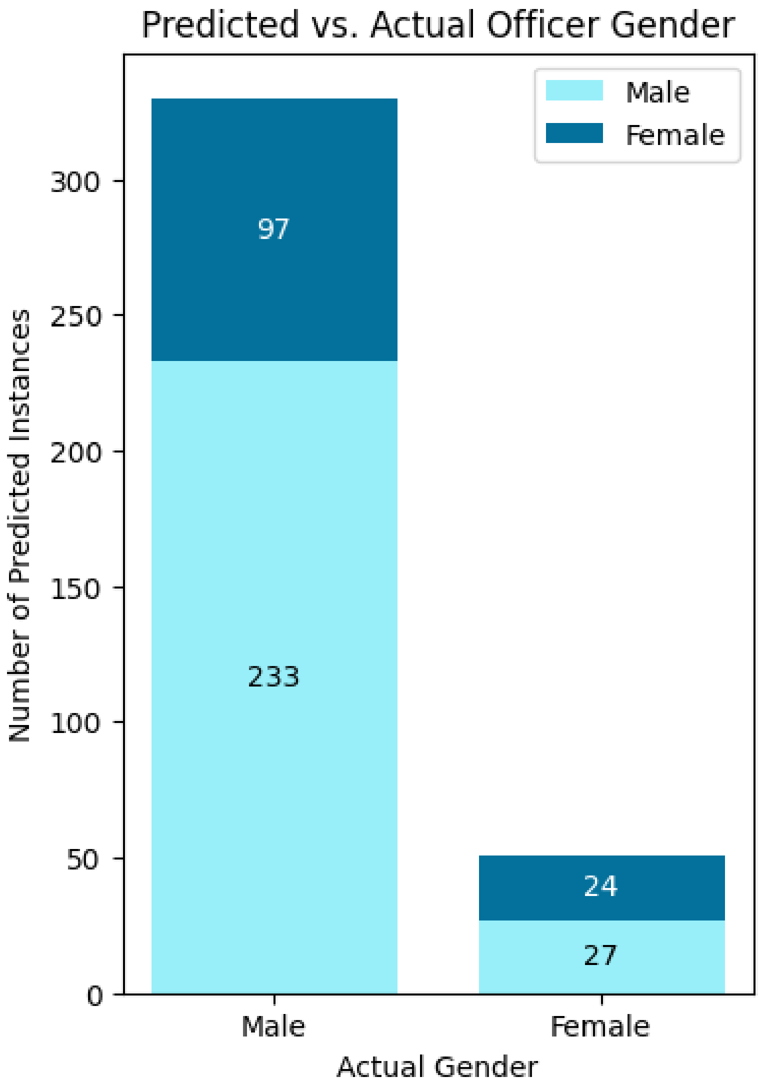

Predicted vs. Actual Officer Gender
Male
Female
300
250
200
150
100
50
0
Number of Predicted Instances
97
233
24
27
Male
Female
Actual Gender

Figure 4.2

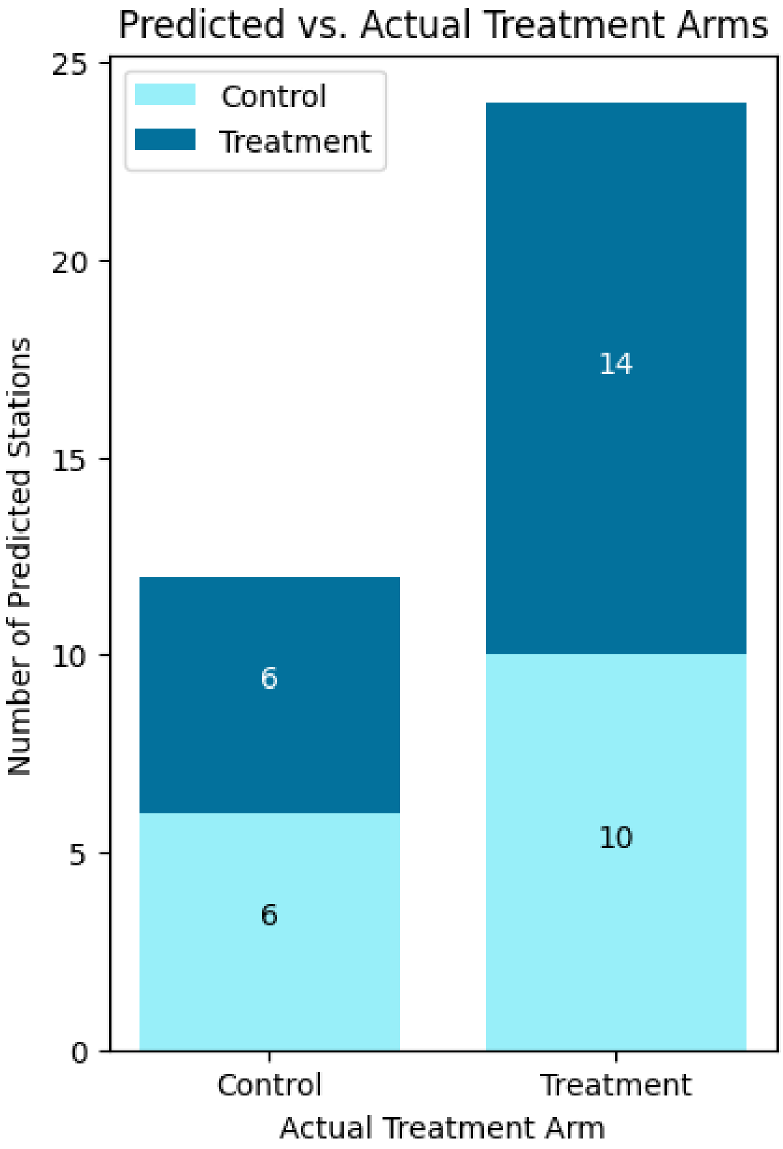

Predicted vs. Actual Treatment Arms
Control
Treatment
Number of Predicted Stations
0
5
10
15
20
25
6
6
10
14
Control
Treatment
Actual Treatment Arm

Figure 4.3

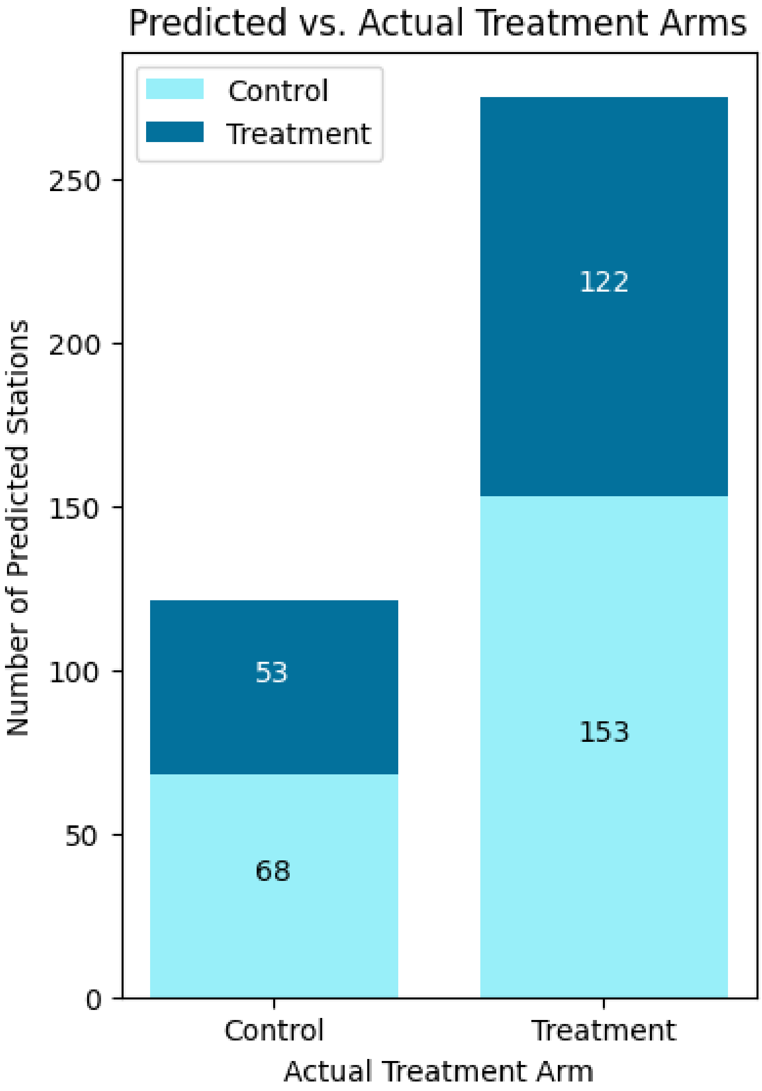

Predicted vs. Actual Treatment Arms
Control
Treatment
250
200
150
100
50
0
Number of Predicted Stations
122
153
53
68
Control
Treatment
Actual Treatment Arm

Figure 4.4

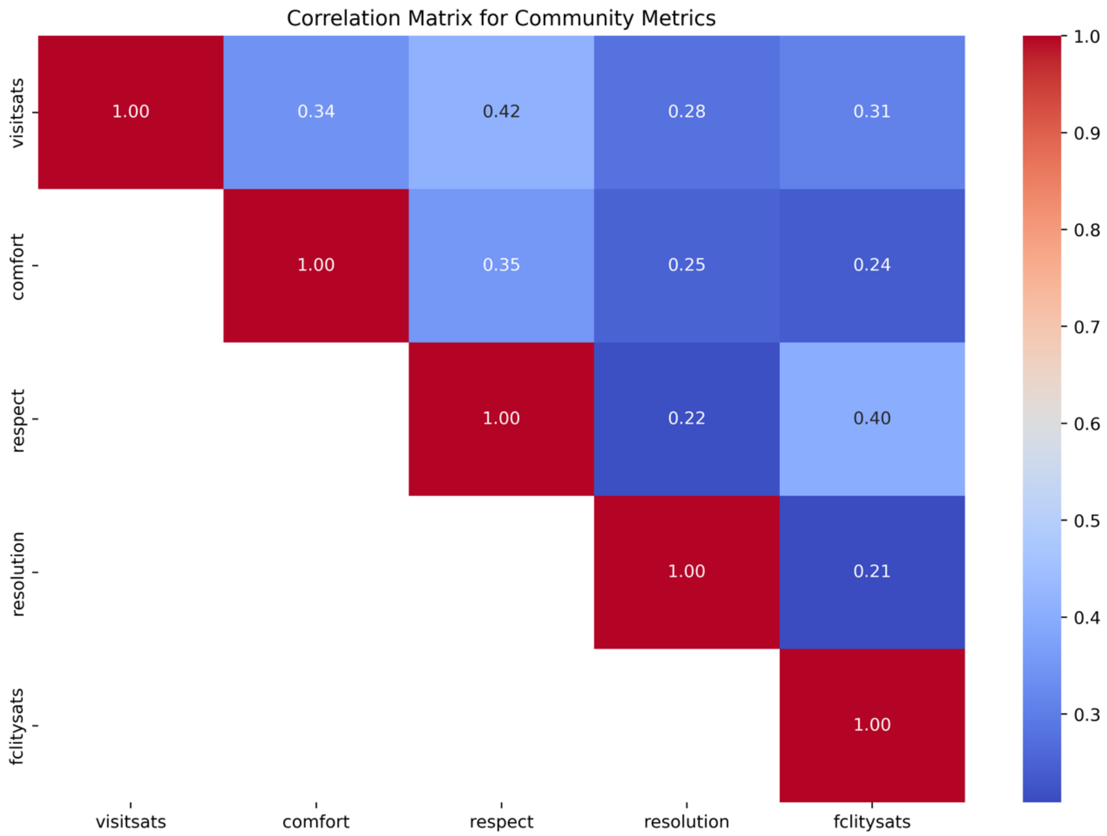


Figure 4.5

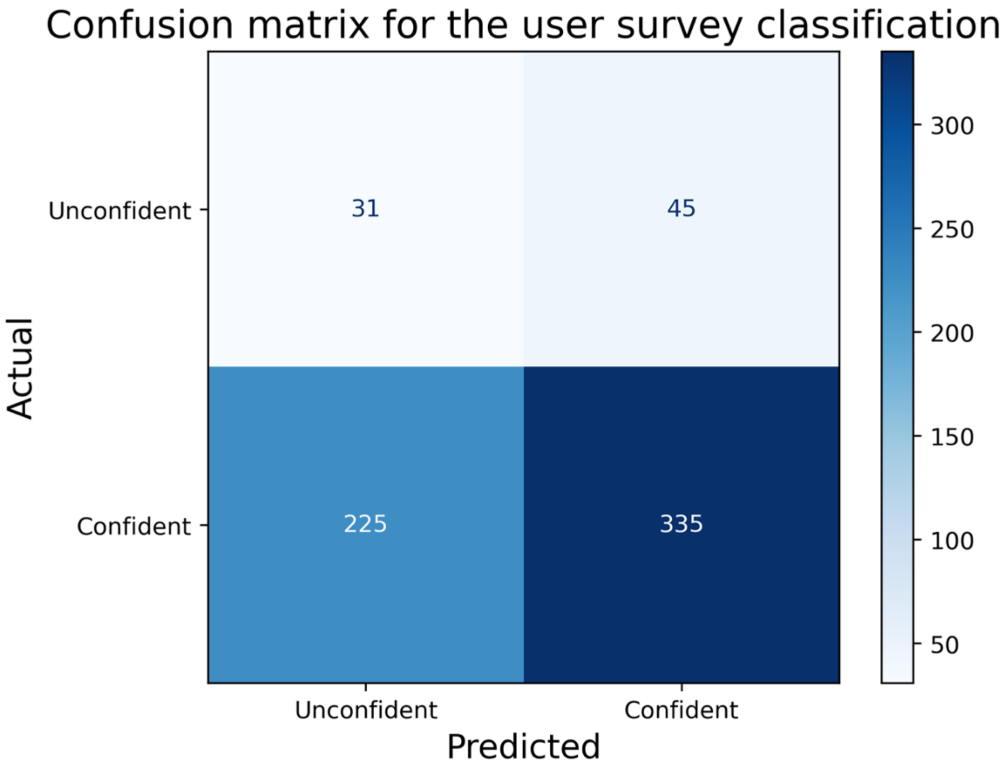


Figure 4.6

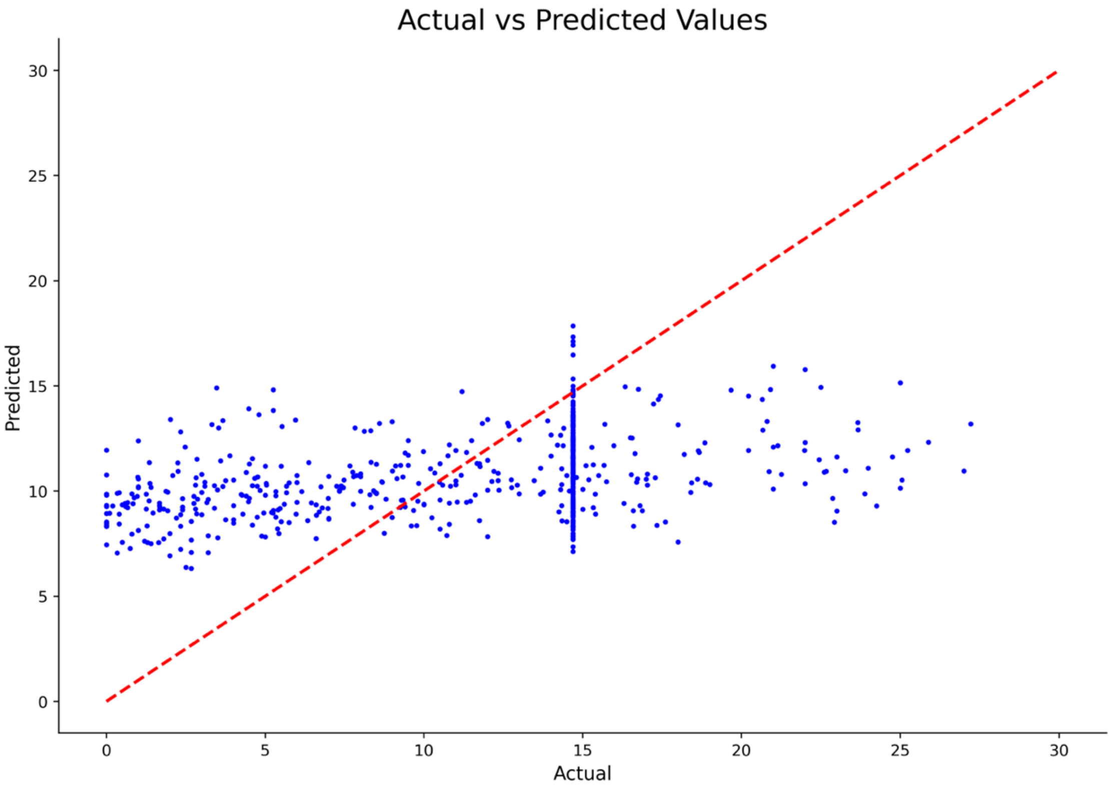


Figure 4.7

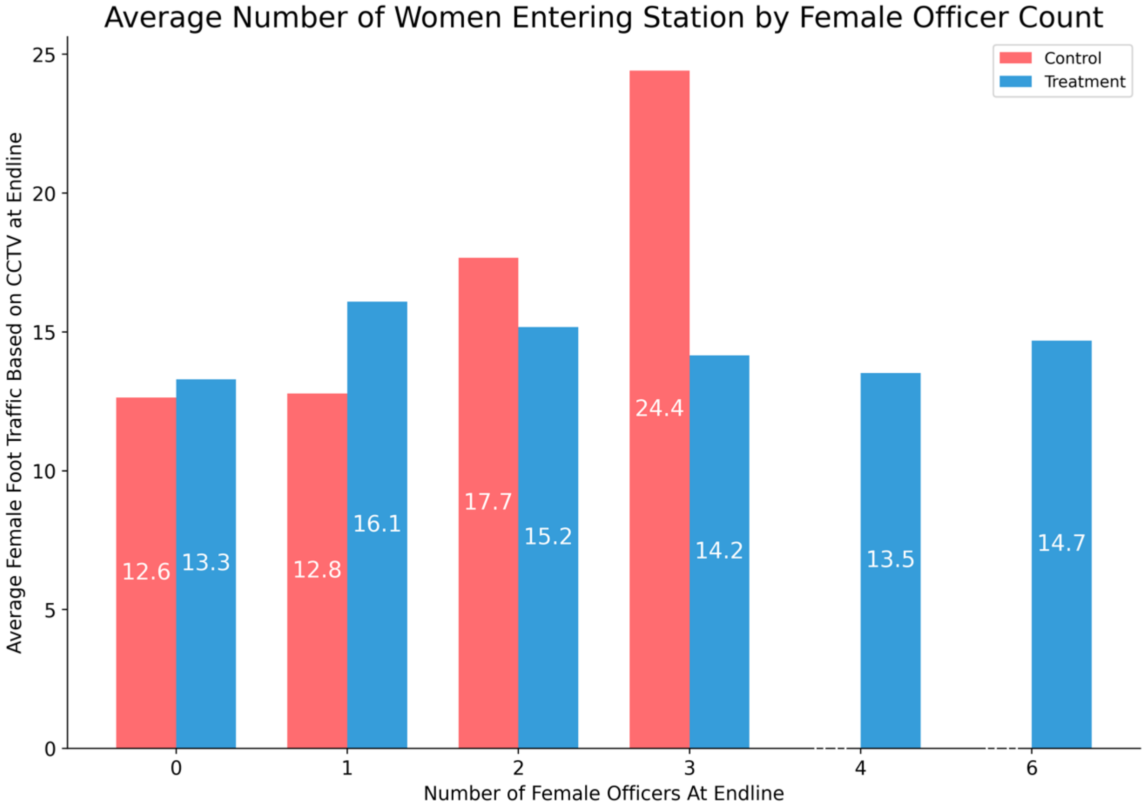


Figure 4.8