The Data Science Life Cycle: A Case Study on Police Effectiveness and Citizen Engagement.

By John W Preston Jr.

Patriarchal systems pose a significant challenge to gender equality, particularly in the realm of law enforcement. In some societies where traditional gender norms dominate, women often encounter barriers when seeking justice for gender-based violence and other crimes. Research studies, such as those by Sukhtankar, highlight innovative approaches to police reform. Women's Help Desk (WHD) and gender-sensitization training for police aim to improve responsiveness and effectiveness toward those found victims of gender-based violence. My report examines two critical questions through the lens of the data science life cycle.

1. What factors influence perceptions of police effectiveness in handling cases related to women?
2. What factors predict the likelihood of a citizen visiting a police station based on survey data?

The importance of these questions lies in uncovering biases and systemic challenges that hinder access to justice and building trust within marginalized communities. Insights into these types of questions can guide us to establish reforms for more than just police reform. Understanding how the implementation of WHD or targeted training can affect police stations can also give insight into how it would affect the public's perception of police and their effectiveness surrounding gender-based violence.

My analysis incorporates datasets, variables, and the idea of statistical techniques to uncover trends and predictors critical to improving police effectiveness. Additionally, over this semester, we discussed significant cases that focused on the ethical frameworks of Utilitarianism, Kantian deontology, and Aristotle's virtue ethics, which have been utilized in this report to ensure that each analysis and approach not only was conducted responsibly and equitably but also with social justice principles. From a Utilitarianism approach, the original body of research aimed to maximize benefits by identifying reforms that would improve police services for the greatest number of people, particularly marginalized women. As applied to the data, focusing on the maxim overall benefit to society serves not only the marginalized women, but also the community at large. Prioritizing changes that would enhance police effectiveness, accessibility, and trust for the greatest number of people disproportionately affected by gender-based violence was the focus. By using data driven insights to inform targeted interventions, my research analysis could create a widespread societal improvement. Utilitarianism emphasizes outcomes and ensures that the proposed reforms deliver tangible benefits to the most vulnerable populations.

Data privacy and respect for individual rights were prioritized throughout the data collection process. Kantian deontology helps us to remember that informed consent and anonymization of surveys protects the sensitive information and identities of the respondent. Adhering to the moral principle throughout the research ensures the handling procedures were designed to protect the citizens dignity and autonomy. Having measures in place to protect participants implemented a secure process to prevent the misuse of sensitive data. Unlike utilitarianism which focuses on outcomes, Kant's

perspective requires each action taken during the data collection to be ethical. By standing firm on Kant's belief helped me to respect the data collected by the respondents and respect their participant action as ends in themselves and not merely means to an end. By focusing on this approach, the ethical standards were never compromised for my personal research gain.

Since the beginning of this course, I have enjoyed learning about Aristotle's virtue ethics. By addressing systemic biases, my study promoted integrity and equity in policing. My research aimed to promote virtues of justice, fairness, and trust as a moral virtue. Police reforms instill a deep root to targeted training all while holding communities accountable to their actions. Reflecting on the concept of virtue ethics, it helps us prioritize the development of institutions and individuals. Virtue ethics stresses the importance of character and intentions. Finding the golden mean between two vices produces results but also upholds ethical excellence.

One important part of my research project started with the creation of questions followed by the dataset choices and variables utilized to complete my project. The two primary datasets I used were: Police_baseline data and Police_full data. Both of these datasets provided me with information on the police officer demographics and public perception of police effectiveness. I also utilized Citizen_full data which gave insight to the survey responses from households regarding their engagement with police and perceptions of their effectiveness. These datasets helped me understand the connection between law enforcement and the public, specifically in contexts of women.

Key variables I used in my research from the police datasets include baseline perceptions of effectiveness, endline perceptions of effectiveness (b_effective,

e_effective), and helpfulness (b_helpful). These variables shed light on how gender demographics influenced public perceptions of police performance. From the Citizen_full dataset, I utilized the critical variables (b_visit, e_visit) and perceptions of police handling of cases involving crimes against women (b_pol_handling, e_pol_handling), along with respondent gender (member_gender).

Data cleaning and processing were essential to addressing missing and inconsistent entries. I will state that I did not do the best at this stage as I am a novice with python. I do believe that having the techniques to do robust cleaning, merging, and carefully evaluating the data to maintain consistency is vital to the outcome of any project. After my presentation I reviewed the notes provided to me from Professor Emma. I first removed the columns from each dataset that I was not utilizing for this project. Afterwards I do believe I should have left some valuable data; however, I was not able to undo the cleaning process of the removed columns. Last, I decided to merge the police_full_df with the police_baseline_df datasets to create a new police_df dataset. This allowed me to focus only on the columns necessary for my questions. The merger was based on the similar columns.

Before my presentation I did not take the time to complete any bivariate analysis due to time constraints. I did however review my questions afterwards to determine some small yet useful approaches to build a correlation between variables. As I worked to answer my first question to determine if there was any real correlation between variables(Treatment group and b_effective), I realized based on my visualization one of two things. 1. The implementation of a WHD versus a help desk ran normally had no major difference. 2. I tainted my data somewhere while merging or replacing missing

values. I think it's common to be unsure whether variables are as they should be. That is why we should periodically check our data to ensure there are not errors. I decided to create a scatter plot with the same two variables (Treatment and b_effective) to see if there was any difference in my outcome; however, I received a similar visual. Knowing that b_effective variable was focused on perceptions of police before the survey timeline, I decided to create a new boxplot to test if I would be able to get a different outcome. The outcome shifted slightly showing that those who completed the survey made definite choices on the effectiveness regarding the implementation of WHD. One difference I noticed from the baseline to endline results was that at the endline, respondents did not choose that the implementation of WHD was 'neither effective nor ineffective". The respondents made a clear choice which is a positive that could be reviewed further.

For question number one I also reviewed whether there was any correlation between treatment groups and the perception of help received from the police. Police effectiveness in gender-based violence refers to how well law enforcement resolves cases, ensures justice, and protects victims, often measured by first information reports or convictions. In contrast, receiving help from the police focuses on the victim's experience of support during the process of the incident. I believe help is more personal and reflects the victim's perception with officers when faced with gender-based violence. It's important to point out that both help from the police and the effectiveness of support are both important to police reform surrounding gender-based violence.

After reviewing the outcome of perception of help from the police compared to the treatment group, I recognized there was a slightly higher count that the treatment

group had for being helpful; however, the count for unhelpful and very unhelpful were higher for treatment groups compared to the same with the control groups. From reviewing the box plots to reviewing the value counts, there actually seemed not to be a major difference surrounding the correlation of variables. I did try to run .describe to review the statistical data, and quickly was reminded that these variables are not numerical in nature. Although I was able to review the value counts for the perception of help from police split between treatment groups, I do not know how to change categorical data to numerical in order to calculate descriptive statistics without help from an online source. For this project I made the choice not to include any outside sources until I can learn python and other key elements first to build a more solid foundation.

Understanding the factors predicting the likelihood of a citizen visiting a police station provides insights into public engagement with law enforcement, particularly in addressing crimes against women (CAW). With question two, my focus was towards whether we could predict the likelihood of a citizen visiting a police station based on survey data. I started by reviewing the univariate data of the variables b_visit and e_visit to visualize whether a household member visited a police station before(baseline) or during (endline) the survey timeline. This helped me see any trends for visiting police stations. After reviewing bivariate analysis of member_gender with both b_pol_handling, e_pol_handling, it's a hard possibility that the respondent gender or the perception of police effectiveness has an influence on whether a household member would visit a police station. I decided to go back into the original dataset to review what variables were there to see if I could take a different route to find some correlation. To my surprise, factors such as age and income were removed from the data. I reviewed the

surveys from the MPP Science Package, and I recognized that age and salary questions were on the baseline citizen survey but not the endline. Although the information was gathered on the survey it was not included in the dataset. There were questions surrounding whether or not a respondent received a salary; however, no direct data was collected to explore the connection between age, income, effectiveness, or helpfulness surrounding police and gender-based violence.

The absence of demographic data, such as age and income raise some important ethical questions for me. How can police reform address systemic barriers without a comprehensive understanding of demographic factors. To really focus on socioeconomic status, it requires transparency up to the limit of sharing identifying information. I do believe that providing age and income in the survey data could have helped me determine whose voices were missing and how those gaps could skew the report. The utilitarian framework reminds us to always ask the question, who is the relevant population? Who are we seeking to help? To really find those answers, we need more consistent questions across surveys, and the inclusion of demographic data in the dataset. Another thing that came to mind while reviewing the surveys was that it asked for the address of the respondent. I would have loved to see that data turned into a categorical column that let us know whether the respondent lived in an urban or rural community. The question could have also been changed to simply ask if the respondent lived in an urban or rural area; however, the data could have been skewed if the respondent considered their urban area to be rural or not understand the question and guessed. To really sum it up, a true ethical question that would convict us would be can you base policy changes on incomplete data that may overlook not only marginalized

women, but also marginalized communities? These questions really highlight the need for consistent inclusive data collection practices to capture the diverse experiences of those impacted by gender-based violence.

I am strongly motivated by the potential insights my research questions can uncover through data correlations. However, I recognize that my current inexperience with Python presents a significant barrier to implementing advance statistical methods, modeling, and hypothesis testing. While this challenge has been humbling, it has also fueled my determination to overcome it. I plan to dedicate myself to mastering Python by continuously leaning and leveraging the resources available to strengthen my skills. By committing to these goals, I will be building a solid foundation that will not only support my research but also prepare me for a successful career as a data scientist.

Resources

1. Sukhtankar, S., Krukswisner, G., & Mangla, A. (2022). Policing in patriarchy: An experimental evaluation of reforms to improve police responsiveness to women in IndiaLinks to an external site.. *Science*, *377*, 191-198. https://doi.org/10.1126/science.abm7387
2. Manna, S., Singh, D., Barik, M. *et al.* Prevalence of intimate partner violence among Indian women and their determinants: a cross-sectional study from national family health survey – 5. *BMC Women's Health* **24**, 363 (2024). https://doi.org/10.1186/s12905-024-03204-x
3. World Bank. "Standing up to Fight Gender-Based Violence in South Asia." *World Bank Blogs*, 21 Nov. 2018, https://blogs.worldbank.org/en/endpovertyinsouthasia/standing-fight-gender-based-violence-south-asia.
4. Ferrari G, Torres-Rueda S, Chirwa E, et al. Prevention of violence against women and girls: A cost-effectiveness study across 6 low- and middle-income countries. *PLoS Med*. 2022;19(3):e1003827. Published 2022 Mar 24. doi:10.1371/journal.pmed.1003827 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8946747/