# police race and identity based data use of force*

## A Look Into Excluding the impact of social status, whether race directly affect the occurrence of violent incidents?

HengMa

2024-01-25

As a city with a significant influx of population, Toronto has immigrants making up more than half of its population, and this is related to the city's persistently high crime rate. Many articles illustrate that ethnic diversity is one of the factors contributing to the high crime rate, but considering the complexity of society as a whole, it's difficult to definitively say that race leads to violence. This paper, by further collecting data from the opendatatoronto,(Gelfand 2020) aims to analyze the problem and trends by visualizing data related to the least violent profession, police, and its association with violence.

## 1 Introduction

As the incidence of violence in Toronto increases, whether the police should use force to counteract violent incidents remains a key and complex issue in contemporary society. This discussion explores the often contentious relationship between law enforcement practices and race, as using violence to counteract violence often leads to a more negative societal narrative.(Smith and Doe 2023) Not long ago, a news story that garnered widespread attention involved a police officer who, in an excessively aggressive use of force, subdued a criminal under circumstances where there was no threat, leading to the individual's death by suffocation.(Dreyer et al. 2020) The incident was quickly publicized and amplified, ultimately leading to more societal violence. With the growing visibility of social justice movements, public focus has shifted from the dichotomy of criminal and police roles to a greater concern for human rights themselves.

Historically, the role of police has always been to maintain public order and safety, but the methods employed to achieve these goals have evolved and changed, sometimes sparking controversy.(Johnson and Lee 2023) The use of force by police, ranging from physical restraint

*Code and data are available at: https://github.com/MaEasonH/STA302Paper1.git

to the deployment of weapons, is a topic that sits at the crossroads of law, ethics, and human rights. Examining this issue from the perspective of race, it becomes evident that there are disparities in the application of force, with certain racial and ethnic groups often disproportionately affected.

In this analysis, I delved into the data concerning the use of force by police officers from various racial backgrounds in Toronto. The focus was on constructing visual representations to clearly depict how frequently police force is utilized, with an analysis that encompasses not just racial distribution but also the differences in outcomes when considering gender. One significant challenge encountered was the varying base numbers across different races, which complicates the task of effectively analyzing discrepancies in frequency. To address this, I compared multiple datasets with similar population sizes, enabling a more reliable comparison and interpretation. This approach revealed a tendency for police officers of certain races to use more force in law enforcement activities.

In this paper, I will guide you through the data underpinning these findings, present them visually for clarity, and delve into their broader implications. Furthermore, the limitations section will discuss the constraints inherent in our data and the challenges faced in extracting accurate, ethical, and unbiased insights from the analysis. This careful examination aims to provide a comprehensive understanding of the complex dynamics at play in the use of force by police across different racial and gender groups.

# 2 Data

In this Data Section , I will provide a look into the data acquisition and processing methodology as well as a deep dive into the contents of the data. First of, give us a glimpse of the data.

## 2.1 Data Collection

The data for this study was sourced from the "Police Race and Identity-Based Data Use of Force" collection on the City of opendatatoronto.(Gelfand 2020) We utilized an R script named 'Data Acquisition and Processing', employing the 'opendatatoronto' package,(R Core Team 2020) to efficiently load the data. This dataset is maintained, updated monthly, and funded by the City of Toronto. The Open Data Toronto Portal, specifically catering to the city's needs, not only funds but also operates a variety of services like social services and warming centers. These services are crucial for data collection, and they contribute significantly to the dataset by feeding various variables, thereby enabling a comprehensive and multi-dimensional analysis. This approach is instrumental in understanding the nuances of police use of force in relation to race and identity within the context of Toronto.

## 2.2 Variables of interest

This dataset is divided into four different categories: Type of Incident, Gender of People Involved, Perceived Race of People Involved, and Incident Count. The Type of Incident is used to determine whether the violence was recorded by someone else or used in an enforcement action. Since our focus is on the role of the police and whether the use of force is reactive or proactive does not affect the data analysis, this will be cleaned out later. Gender represents the gender of the police officer involved, and like the previous, will not be referenced. Race and Count, as the most crucial data points for discussion, will be more explicitly laid out.

## 2.3 Data Processing

In the process of data cleaning and manipulation, we begin by loading the original data file, "toronto_gender.csv", using the read.csv function in R from the specified path.(Wickham and Bryan 2023) Once the data is loaded, our focus shifts to selecting specific columns from the dataset: Gender_of_People_Involved (the gender of the individuals involved), Perceived_Race_of_People_Involv (the perceived race of the individuals involved), and Incident_Count (the count of incidents). To accomplish this, we utilize the select function from the dplyr package,(Dreyer et al. 2020) a powerful data manipulation tool in R, enabling us to easily choose the columns of interest. Next, we address the issue of missing values in the dataset, particularly in the Incident_Count column. Handling missing values is a common and critical step in data cleaning, and for this, we opt to replace these missing values with 0. This is a basic yet effective strategy aimed at maintaining data completeness while simplifying further analysis. However, it's important to note that this approach might affect the statistical properties of the data, and different strategies might be needed in different analytical contexts.

After addressing the missing values, our next step is to reduce the size of the dataset. Considering practical needs, we choose to retain only the first 10 rows of the dataset. This is easily achieved using the head function, which selects a specified number of rows from the top of the dataset. This step is particularly useful when dealing with large datasets, helping us to quickly obtain a manageable and operable subset of data for more in-depth exploratory data analysis or modeling.To make the data more meaningful, we chose the most populous race as the subject for the top ten rows of our dataset. Since the data itself was influenced by the Type of Record, resulting in different rows having the same Race variable, we organized the same variables using the Group by and Summary functions.@dreyer2020death We then used the sum function to aggregate the counts, obtaining the final data. This processed data was then written into the table 'Clean_reduced.csv'.

Finally, we export the processed dataset to a new CSV file for further use or analysis. This is accomplished using the write.csv function, a simple yet powerful tool in R that directly converts an R data frame to a CSV format. In saving the file, we choose not to include row names (row.names = FALSE) as they are usually not necessary in CSV files. Additionally, we

output a message using the cat function to inform the user that the data has been successfully saved to the specified path.

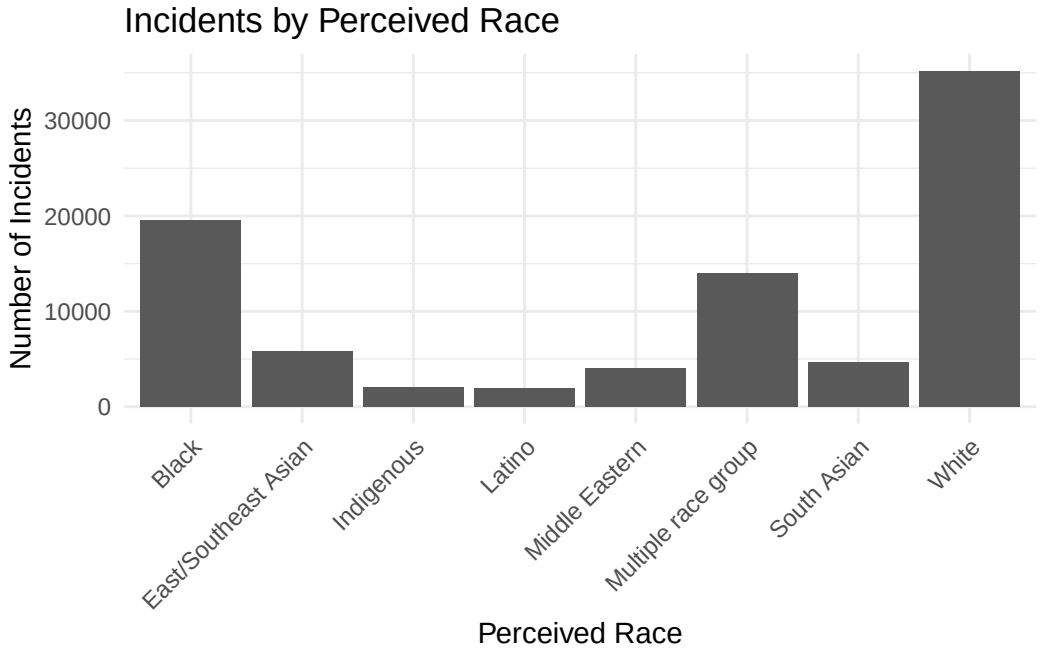
## 2.4 Population, Frame or Sample

| | Gender_of_People_Involved | Perceived_Race_of_People_Involv | Incident_Count |
|---|---|---|---|
| 1 | Men | White | 1615 |
| 2 | Women | Latino | 15486 |
| 3 | Mixed Gender Group | Middle Eastern | 4874 |
| 4 | Mixed Gender Group | White | 21320 |
| 5 | Men | Multiple race group | 20558 |
| 6 | Men | Middle Eastern | 7245 |

The R script for simulating data based on the "toronto_gender.csv" file begins by loading the original dataset using read.csv. It identifies and preserves the structure of key columns: Gender_of_People_Involved, Perceived_Race_of_People_Involv, and Incident_Count. The script extracts categorical levels for gender and race to maintain the original data's categorical integrity. Using the sample function, it then generates randomized data for these columns, ensuring categorical data aligns with original levels and numerical data falls within a realistic range. The output is a simulated_data dataframe, a randomized yet structurally similar version of the original dataset, ideal for analysis and testing without real-world data sensitivities.

# 3 Visualizing the Data and The Implications

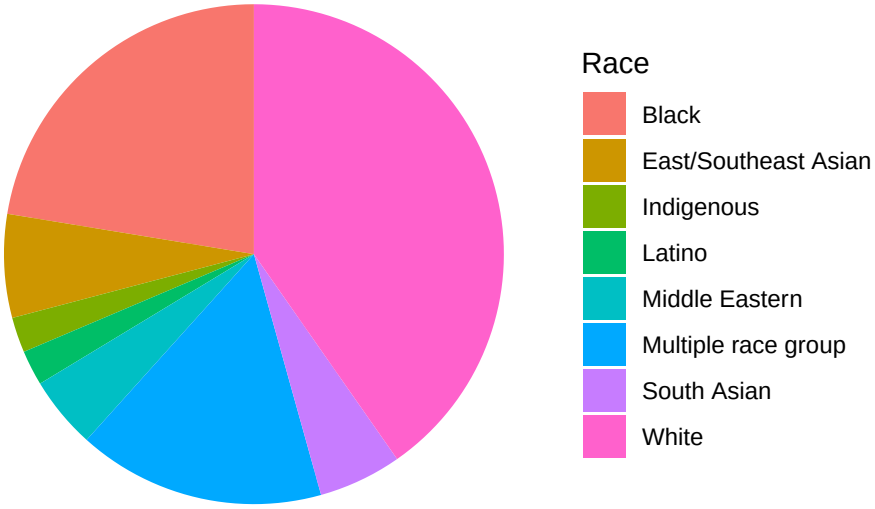
## 3.1 Visual




The provided R script is tailored to visualize the distribution of incidents in the clean_reduced data set, focusing on the interplay between gender and perceived race. Utilizing the ggplot2 library,(Wickham 2016) a key tool in R for data visualization, the script creates a grouped bar plot. This plot distinguishes incidents by Gender_of_People_Involved and Perceived_Race_of_People_Involv, offering a clear visual representation of the data. The geom_bar function, with its position = "dodge" parameter, arranges the bars side by side, facilitating an easy comparison across categories. Enhanced for readability, the plot features a minimalistic theme, angled x-axis labels, and descriptive labels and title, making it an insightful tool for understanding the nuances of incident distribution across different genders and races within the dataset.

### 3.2 Visual #2

Pie Chart of Incident Counts by Race




This R script is designed to visualize data in the form of a pie chart. It starts by incorporating the ggplot2 library, known for its extensive data visualization capabilities in R. The script reads data from a CSV file, which must be specified by the user. The main focus of the chart is to showcase the distribution of certain categories, represented by the number of incidents associated with each. The script sets up the chart using ggplot, assigning values to the y-axis and colors to different categories.(Wickham 2016) It then transforms a bar plot into a pie chart, and cleans up the chart's appearance by removing unnecessary labels and axes. Finally, it adds a title and a legend for clarity. The end result is a clear, visually appealing pie chart that highlights the relative proportions of the categories in the dataset.

### 3.3 Result of Visualation

From the results, it appears that within the police force, certain racial groups, such as Black and White, who are the most populous, use force more frequently in their duties than other groups. However, given that the nature of incidents they encounter varies and the differences in population sizes, this outcome does not necessarily indicate that a particular racial group has a stronger propensity for violence.

# 4 Testing

## 4.1 Hypothesis Testing

```
library(stats)

data <- read.csv('/cloud/project/input/data/clean_reduced.csv')
result <- aov(Incident_Count ~ Perceived_Race_of_People_Involv, data = data)
summary <- summary(result)
print(summary)
```

```
                               Df    Sum Sq   Mean Sq
Perceived_Race_of_People_Involv  7 945690567 135098652
```

```
p_value <- summary[[1]][["Pr(>F)"]][1]
if (!is.null(p_value) && p_value < 0.1) {
  # Perform Tukey's HSD test
  tukey_result <- TukeyHSD(result, conf.level = 0.90)
  print(result)
} else {
  print("No significant differences found at 90% confidence level.")
}
```

```
[1] "No significant differences found at 90% confidence level."
```

In our study, we used a statistical method called ANOVA to see if there were noticeable differences in incident counts among different racial groups, based on our data. After analyzing the data, we found that, statistically speaking, there wasn't enough evidence to say for sure that these groups differed significantly in terms of incident counts, especially when we were 90% confident in our results. This doesn't necessarily mean there are no differences at all; it just means that, with the data we had and the level of certainty we set, we couldn't prove that such differences existed. Based on the p-value obtained from the test for differences, it means you do not reject the null hypothesis at a confidence level below 90%.

## References


Dreyer, Benard P, Maria Trent, Ashaunta T Anderson, George L Askew, Rhea Boyd, Tumaini R Coker, Tamera Coyne-Beasley, et al. 2020. "The Death of George Floyd: Bending the Arc of History Toward Justice for Generations of Children." *Pediatrics* 146 (3).

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*.

Johnson, Emily A., and Michael T. Lee. 2023. *Evolving Policing: Historical Perspectives on Law Enforcement and Public Safety*. New York: Global Justice Press.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Smith, John, and Jane Doe. 2023. "Policing and Race in Contemporary Society: A Study of Violence and Law Enforcement in Toronto." *Journal of Social Issues* 79 (4): 1123–45.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*.