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

2024-01-23

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 Open Data Portal of the City of Toronto, 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. 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. 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. The use of force by police, ranging from physical restraint to the deployment of weapons, is a topic that sits at the crossroads of law, ethics, and human rights. Examining

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

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 examined data on the use of force by police officers of different races in Toronto. My focus was on creating visual representations of the frequency of force used by the police, analyzing the distribution of data not only by race but also comparing differences in outcomes based on gender. Generally speaking, due to varying base numbers across different races, it's challenging to analyze frequency differences effectively. Therefore, I chose to compare a few datasets with similar population sizes to draw conclusions, finding that police officers of certain races are more inclined to use force in law enforcement. In this paper, I will guide you through the data used to arrive at these findings, present them visually, and discuss their implications. Additionally, in the limitations section, we will address the constraints of our data and the challenges in extracting true, ethical, and unbiased information from the implications of these findings.

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

All the data used in this paper comes from the City of Toronto Open Data Portal, titled "police race and identity-based data use of force". We loaded the data in an R script named 'Data Acquisition and Processing' using the R package opendatatoronto. This data is uploaded and funded by the City of Toronto and is updated monthly. As explained on the Open Data Toronto Portal website, "The City of Toronto funds and operates services, specifically for the city of Toronto." They go on to explain that the services they provide, such as emergency shelters, social services, warming centers, etc., all use the Shelter Management Information System (SMIS) to track and record any party receiving or executing services. These data are collected and even fed into different variables to measure the data.

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. 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, 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.

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

```r
set.seed(1)

original_data <- read.csv('/cloud/project/toronto_gender.csv')

gender_levels <- levels(factor(original_data$Gender_of_People_Involved))
race_levels <- levels(factor(original_data$Perceived_Race_of_People_Involv))

n <- nrow(original_data)

set.seed(2)
```

```r
simulated_data <- data.frame(
  Gender_of_People_Involved = sample(gender_levels, n, replace = TRUE),
  Perceived_Race_of_People_Involv = sample(race_levels, n, replace = TRUE),
  Incident_Count = sample(0:max(original_data$Incident_Count, na.rm = TRUE), n, replace =
)
head(simulated_data)
```

```
  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

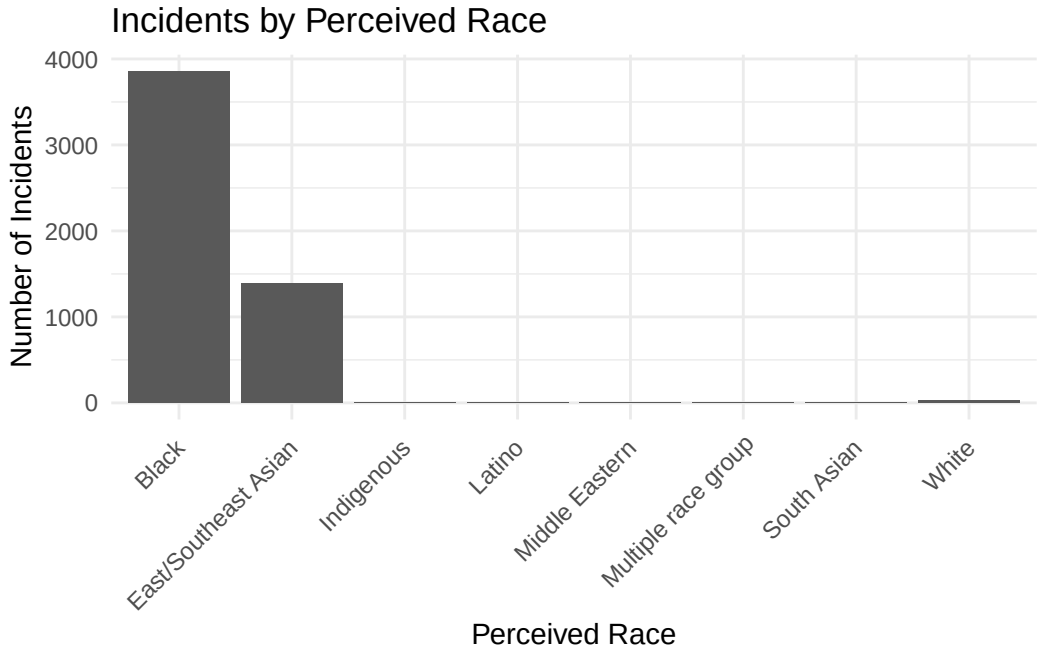
```r
install.packages("ggplot2")
```

```
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
(as 'lib' is unspecified)
```

```r
library(ggplot2)

clean_reduced <- read.csv("/cloud/project/input/data/clean_reduced.csv")
```

```r
ggplot(clean_reduced, aes( y = Incident_Count, x = Perceived_Race_of_People_Involv)) +
  geom_bar(position = "dodge", stat = "identity") +
  theme_minimal() +
  labs(title = "Incidents by Perceived Race", x = "Perceived Race", y = "Number of Inciden
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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, 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.

4 Limitations

4.1 Uneven Bin width for Age Groups.

4.2 What about the others?

5 Next Steps

6 References