Data Normalization & Classification: Police Violence in North America

June Skeeter

Learning Outcomes:

1) Investigate how data normalization impacts the way we perceive patterns in a dataset

2) Look at different data classification methods and how they impact the way we perceive patterns in a dataset

A) Revisit measurement scales: how are they related to classification methods?

B) Choropleth mapping: displaying ratio data

Content Warning:

This lecture deals with a difficult and painful subject that may be triggering to some people. The datasets we're using today describe incidents of police killings in Canada and the United States. My aim first and foremost is to use this data to raise awareness about a serious issue. Secondarily, I am to use this data to highlight the importance of data normalization and emphasize why it is vital to think critically about the information we are presented with.

Pre-Lecture Poll Questions:

1) Which country has a higher frequency of police violence?

   A) Canada

   B) The United States

   C) They're about equal

2) Which country has a greater racial disparity in incidents of police violence?

   A) Canada

   B) The United States

   C) They're about equal

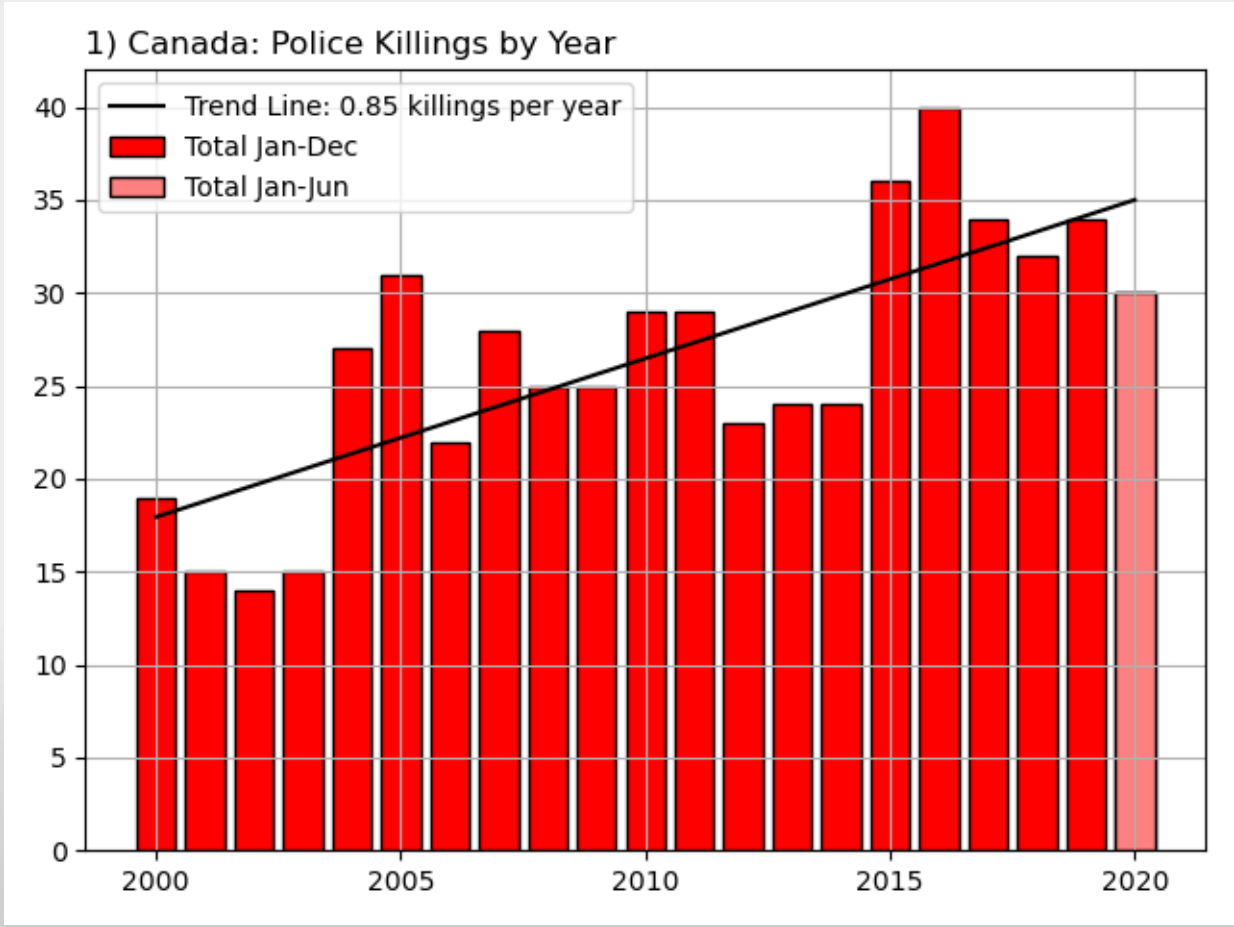
Part 1) The Police Violence Data

Canadian Police Violence Data

- This data was collected by the CBC and is available for download here: https://newsinteractives.cbc.ca/fatalpoliceencounters/
  - "There is no government database listing deaths at the hands of the police available to the public in Canada, so CBC News created its own. The CBC's research librarians have collected detailed information on each case, such as ethnicity, the role of mental illness or substance abuse, the type of weapon used and the police service involved, to create a picture of who is dying in police encounters. " - CBC
- This is not an official count.
  - This dataset is a collection of second hand information in the form of press releases, news articles, etc.
  - Some records are incomplete, and the total number of incidents is likely higher than detailed here.

1) Police killings by year

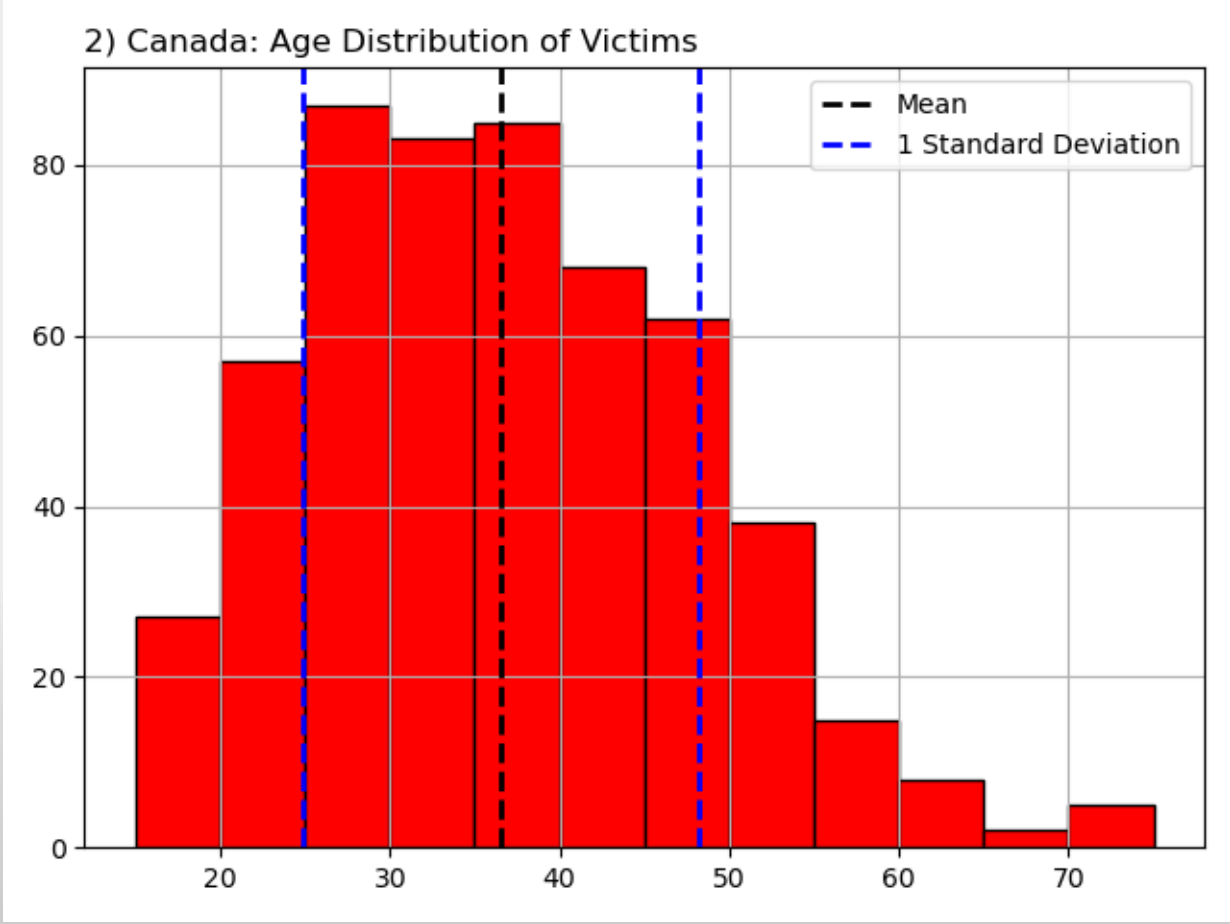
- There were 556 killings between January 2000 - June 2020
  - Increasing trend: 0.85 killings/year.
  - 2020 is on pace to be a record breaking year.



2) Age distribution of victims

Histograms show the shape and spread of a dataset.

- Here we see the age distribution of victims in 5 year increments.
  - The youngest was 15 and the oldest was 77
  - The mean age is 35.6, the standard deviation is 11.6
- The histogram shows us that the age is slightly skewed towards older ages
  - The distribution has a tail

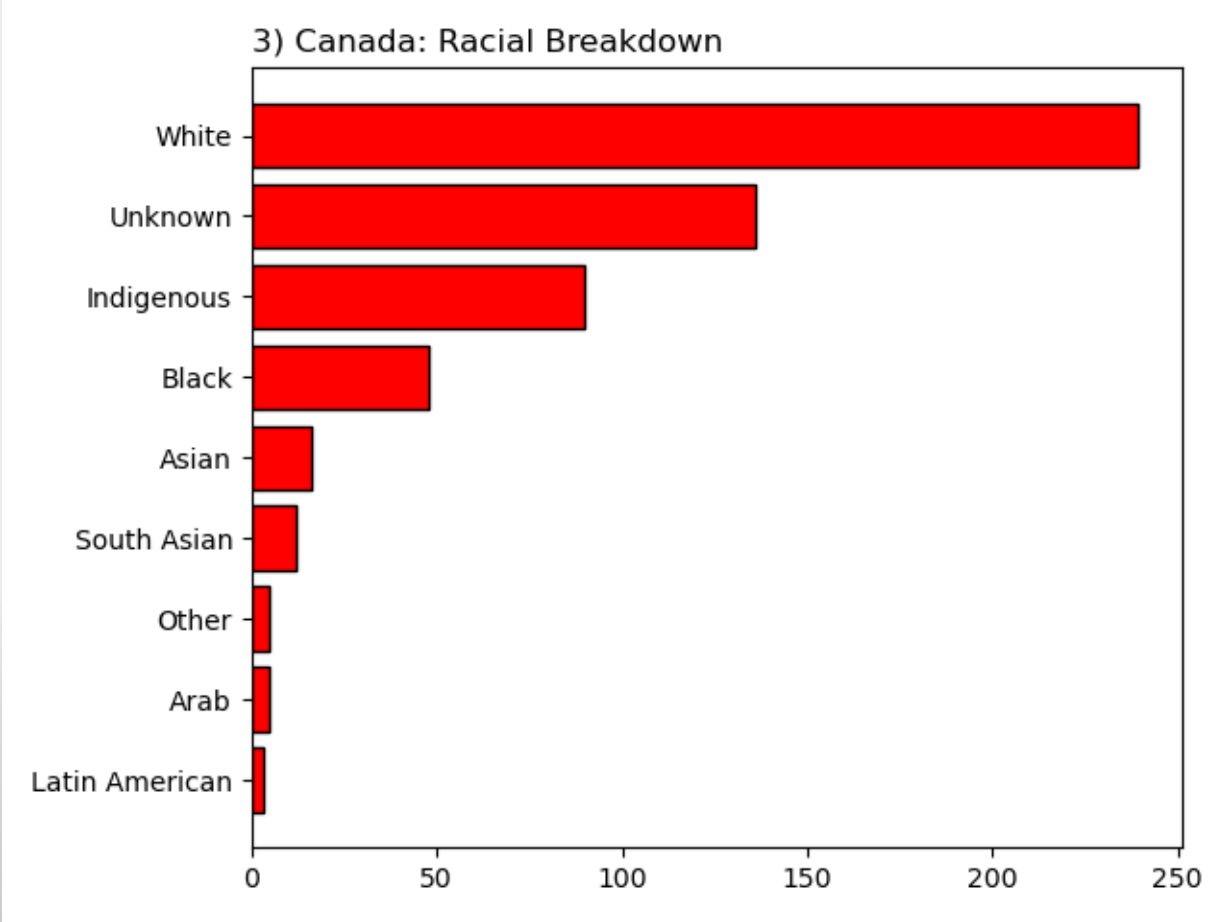


3) The racial breakdown of victims

The majority of police killings are white people

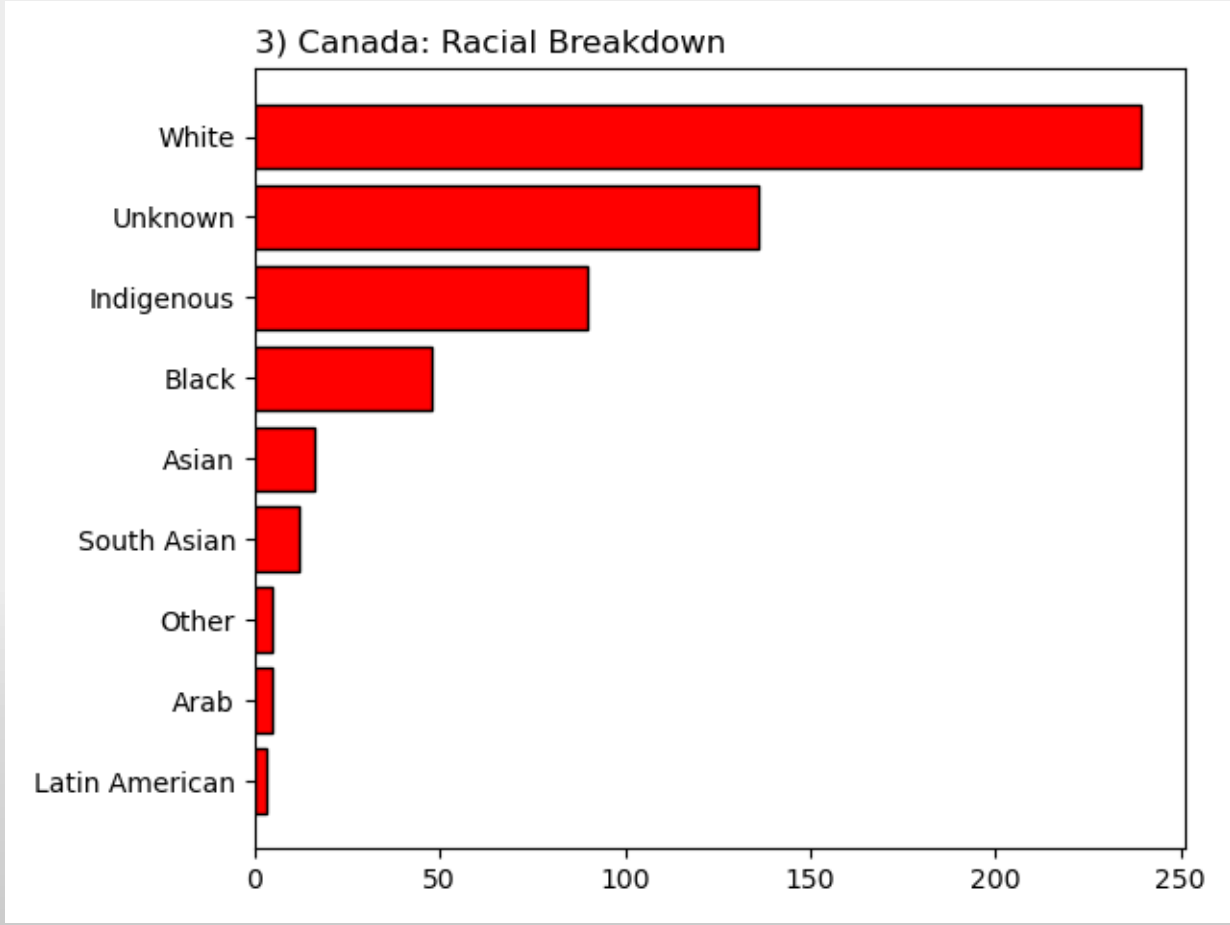
- The second largest demographic is "Unknown“
  - In most cases it means the this information was not recorded by the police.

- Demographic groups are not evenly represented in the populations
  - Canada is about 73.4% White but only 4.7% Indigenous and 3.4% Black



3) The racial breakdown of victims

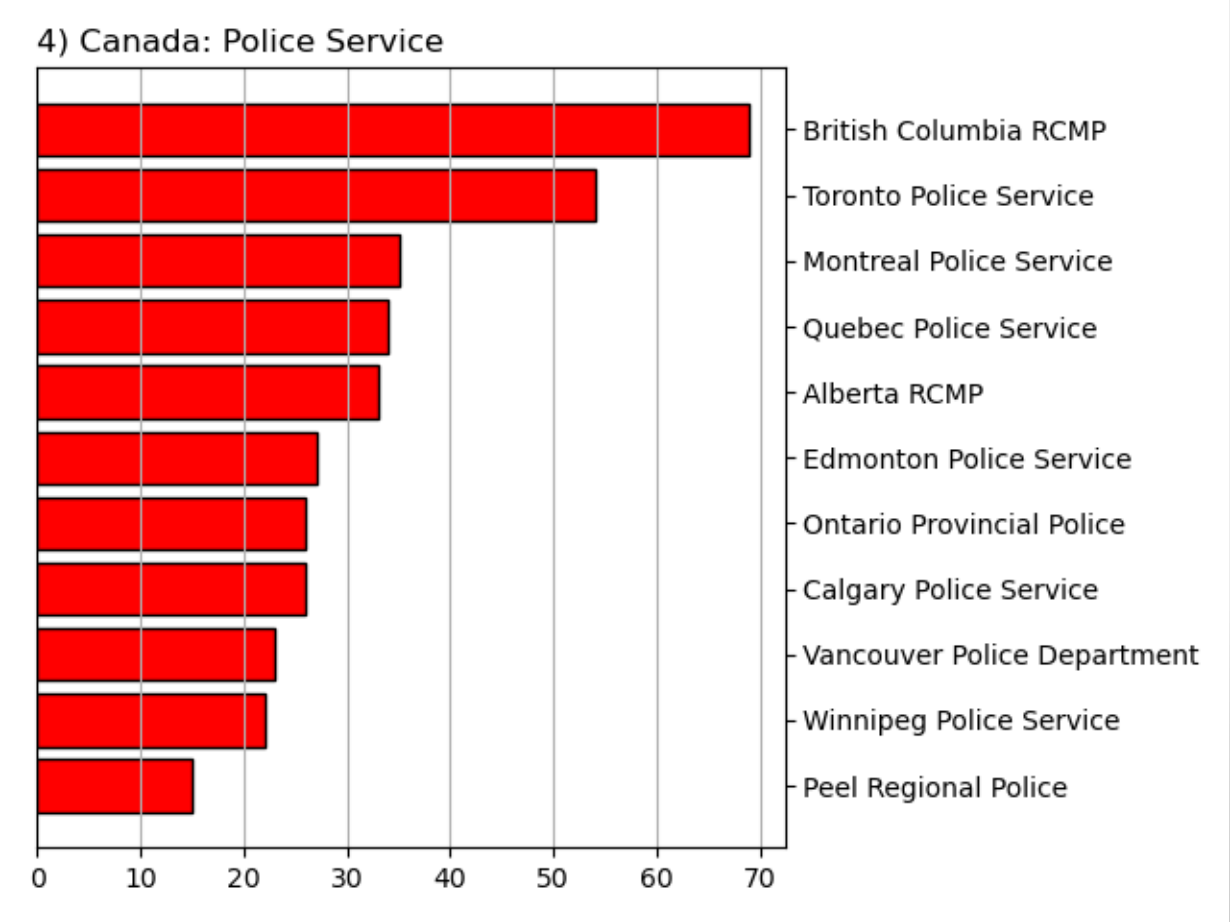
- Think about how comparing the total killings for population groups of very different sizes might impact the way you perceive patterns.
- Using this chart, what demographic group do you think is most likely to be killed by the police in Canada?



4) Which police departments are responsible?

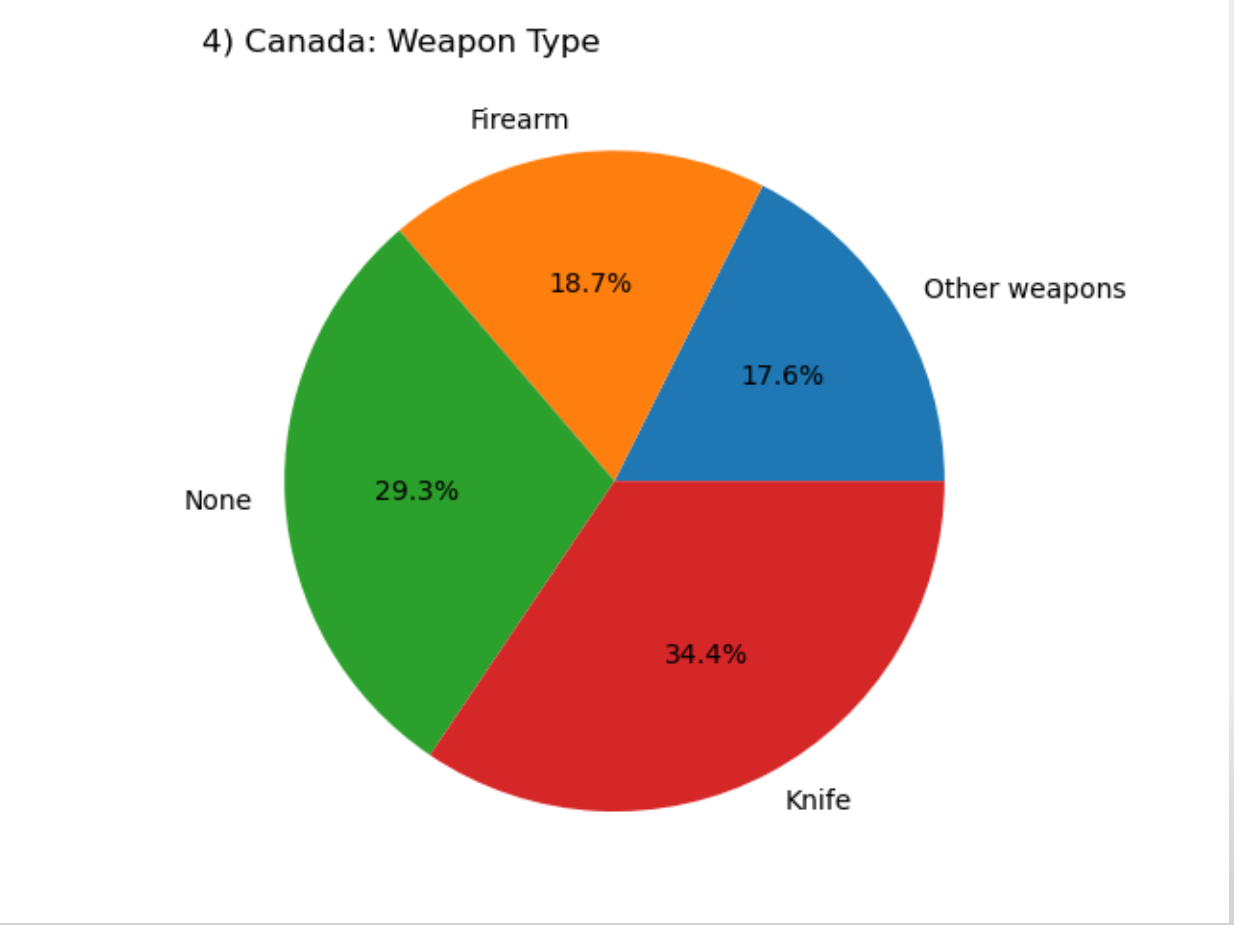
Here are all departments which have killed at least ten people in the last 20 years.

- Provincial police services and large municipal police departments are responsible for the most deaths.
- The RCMP is the provincial police force in eight provinces and the territories.
    - All together, the RCMP is responsible for 34% of deaths.



5) What type of weapon did the victim have?

- Nearly 30% of victims were unarmed.
  - Note - Being armed does not justify any individual police killing.
- In aggregate, a higher number of killings of unarmed people can indicate a predisposition towards excessive use of force.



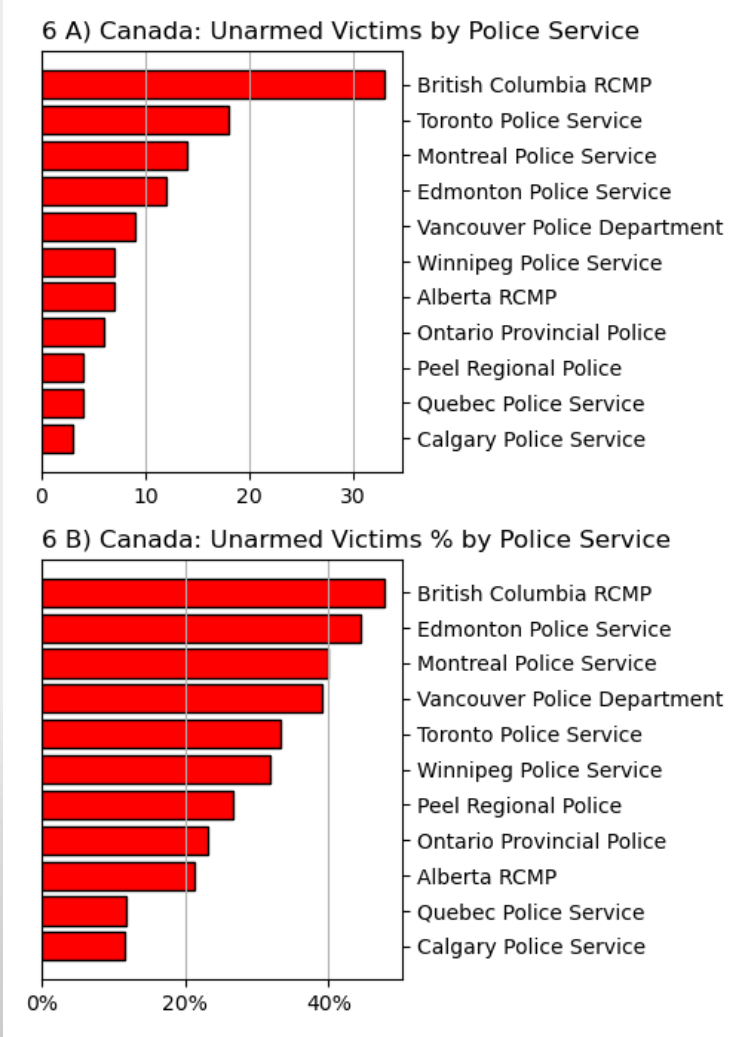
Data Normalization

- Normalization, is the process of scaling (aka. Normalizing) one number by another.
- For example, we can as the question:
  - Which police departments are most likely to kill an unarmed person?
- We need two pieces of information for each police department
  A. The total unarmed victims
  B. The total victims
- We can divide A by B, this will tell us what percentage of each department's victims were unarmed.
- So our normalization calculation would look like:
  - $\%Victims\ Unamred = \frac{\text{Unarmed Victims}}{\text{Total Victims}} * 100$

6) Unarmed killings by department

We see different patterns in the data when using raw counts vs. normalized data.

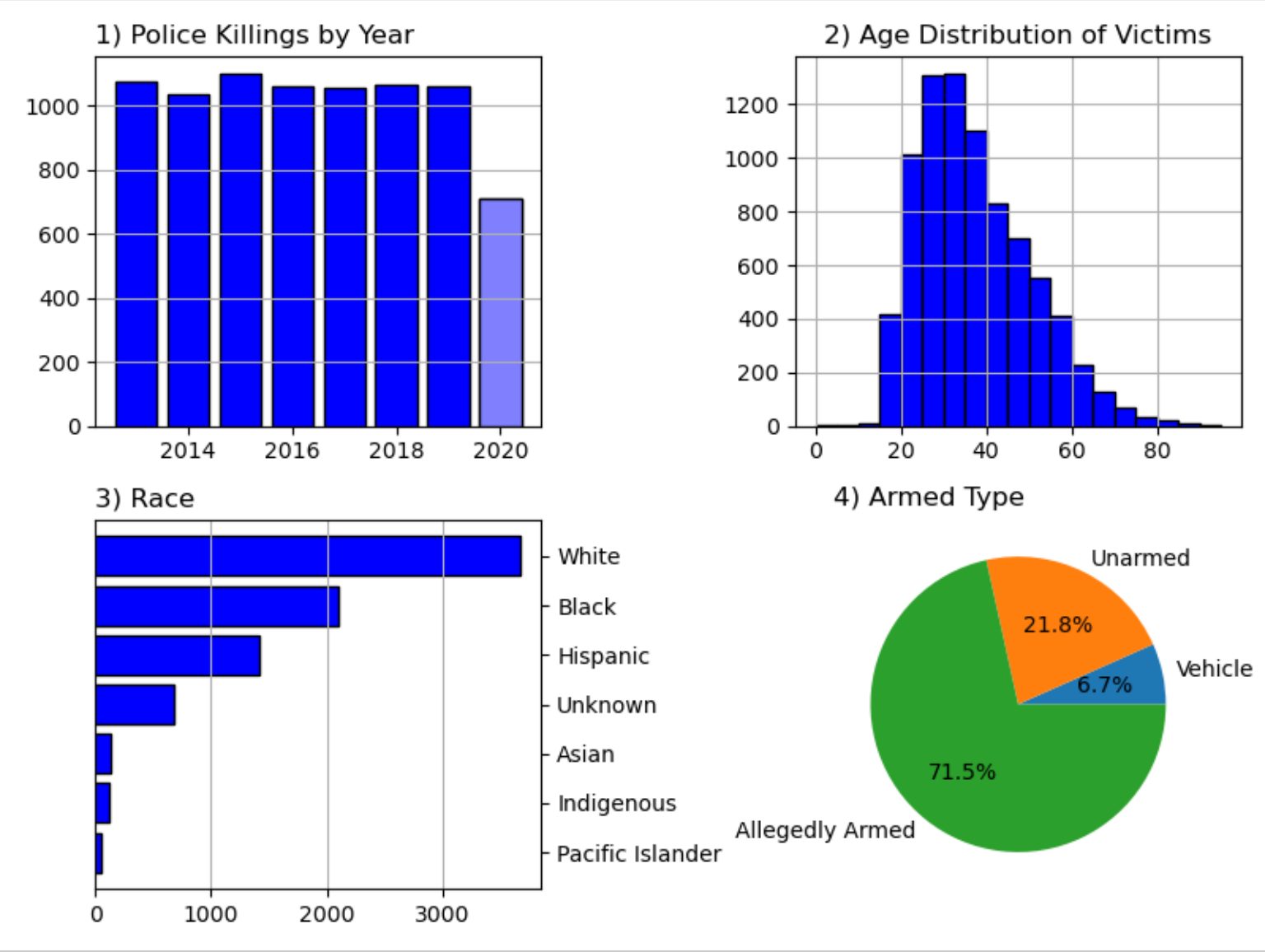
- BC RCMP are kill the most unarmed people.
  - Nearly half the people killed by BC RCMP did not have a weapon.
- Note, the rank/order change when we normalize.
  - Vancouver & Edmonton go up, Toronto goes down.
- This information should be widely available.
  - The RCMP and other Police Services across Canada need to be held accountable.



United States Police Violence Data

- The United States data is collected by a collaboration of researchers and data scientists and is available for download here: https://mappingpoliceviolence.org/
  - "… the data represented on this site is the most comprehensive accounting of people killed by police since 2013… our database includes additional incidents such as cases where police kill someone through use of a chokehold, baton, taser or other means as well as cases such as killings by off-duty police."
- This is not an official count.
  - This dataset is a collection of second hand information in the form of press releases, news articles, etc.
  - Some records are incomplete, and the total number of incidents is likely higher than detailed here.

United States Police Violence Data



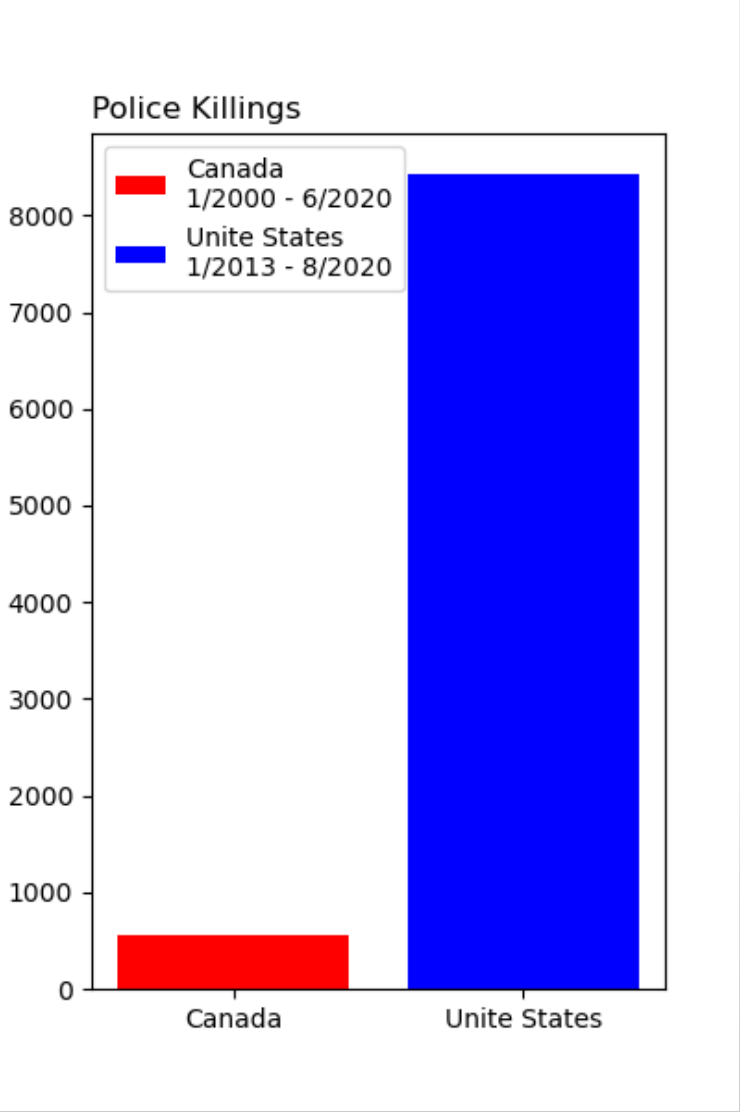
Questions

- What are some other applications for data normalization?
  - Which country has a higher proportion of unarmed victims of police killings?
- What metric(s) might you want to consider when looking at total number of electric cars in each province to gauge electric car adoption?
  - A) Kilometers driven
  - B) Cars per family
  - C) Median income
  - D) Total population
  - E) Average car price

Part 2) Comparing to the United States

Comparing the Two Countries

- What factors do we need to look at to compare police killings between Canada and the United States?
  - The United States has ten times the population of Canada. If we don't account for that, our comparison won’t make any sense.
  - The two datasets have different periods of record.



What to Account For

A) Record Length

- The time periods of these datasets are different.
- We could only look at the time period when they overlap, but this would require us to ignore some of the data.
- Alternatively, we can calculate the average number of killings per year.
  - The data are not from the same periods, but they will be on the same time scale, and they will be as inclusive as possible.

B) Population

- Canada has about 35 million residents, the US has about 327 million.
- To make the datasets directly comparable, we need to normalize by the total population of each country. This will allow us to calculate the police killing rate.

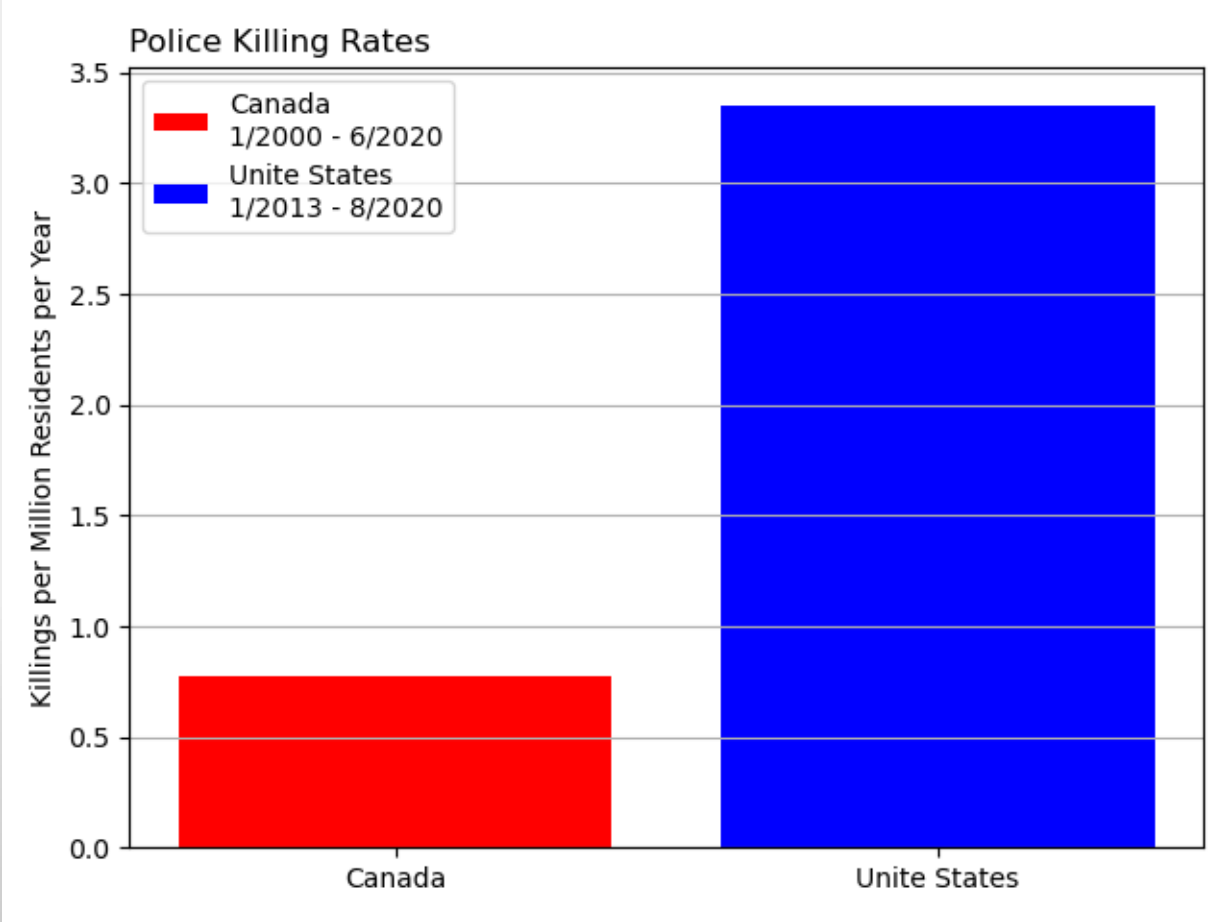
C) Scale

- Dividing by the population would give us the average number of police killings per person per year. This will be a very small decimal. Integers (round numbers) are easier to interpret. We can divide by the population in millions instead.

Police Killing Rates

By normalizing, we can more directly compare the patterns between geographic regions with different characteristics (population) and datasets of different lengths.

- The United States police killing rate is 4.3 times Canada's.
- On average, the US is more dangerous and police are more likely to kill someone there.

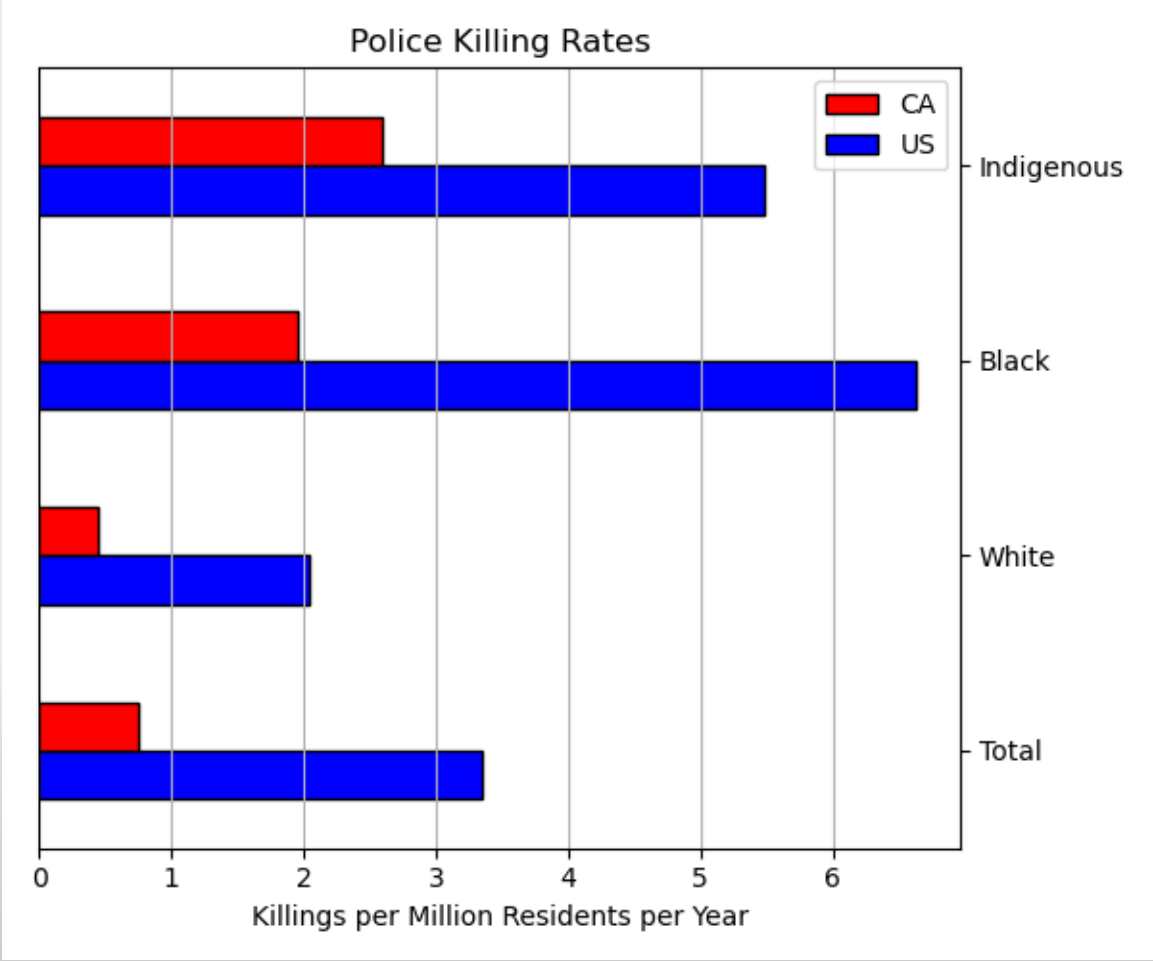


Racial Disparities

- Systemic racism is pervasive on both sides of the border.
- The police violence dataset and census for each country use different demographic groupings.
  - We'll compare the police killing rates of three demographic groups: White, Black, and Indigenous because they are in both datasets.
    - Whites are the majority in both countries, while black and indigenous people are disproportionately impacted by police killings on both sides of the border.
- One Caveat, the race of the victim is unknown for 24% of Canadian and 10% of United States.
  - This adds uncertainty to the comparison. It also means that the police killing rates by race are underestimated, especially for Canada.

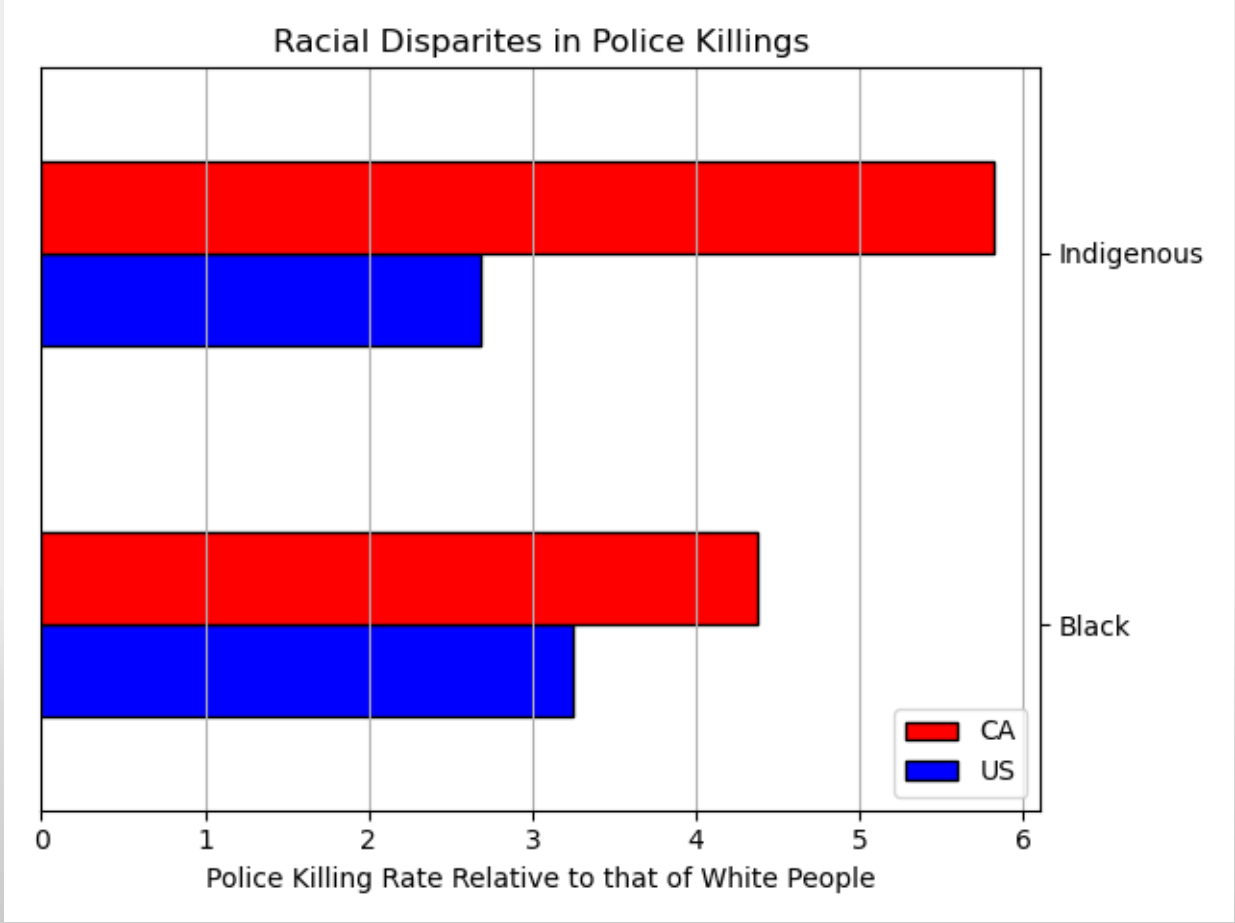
Systemic Racism in Policing

- Scaled to their respective populations, we can prove that Indigenous and Black people are much more likely to be killed by the police than white people in both Canada and the United States.
  - The overall rates for each race are higher in the US than Canada.
  - However the disparity between races is actually greater in Canada.



Systemic Racism in Policing

- To show this, we can divide the black and indigenous rates for each country by the white rate.
  - This will tell us how many times more likely a black or indigenous individual is to be killed by the police than a white individual in each country.
- By this metric, you could suggest that police in Canada are more racially biased.



Systemic Racism in Policing is a Canadian Problem

- This issue isn't restricted to America, it's pervasive in Canada as well and can not be overlooked.
  - The RCMP were created by Prime Minister John A. Macdonald. He got the idea for the Mounties from the Royal Irish Constabulary, a paramilitary police force the British created to keep the Irish under control. Initially called the "North West Mounted Rifles", their primary purpose was to clear Indigenous people off their land. The name was changed to "North-West Mounted Police" because officials in the United States raised concerns that an armed force along the border was a prelude to a military buildup. This organization was renamed the Royal Canadian Mounted Police in 1904.

Question

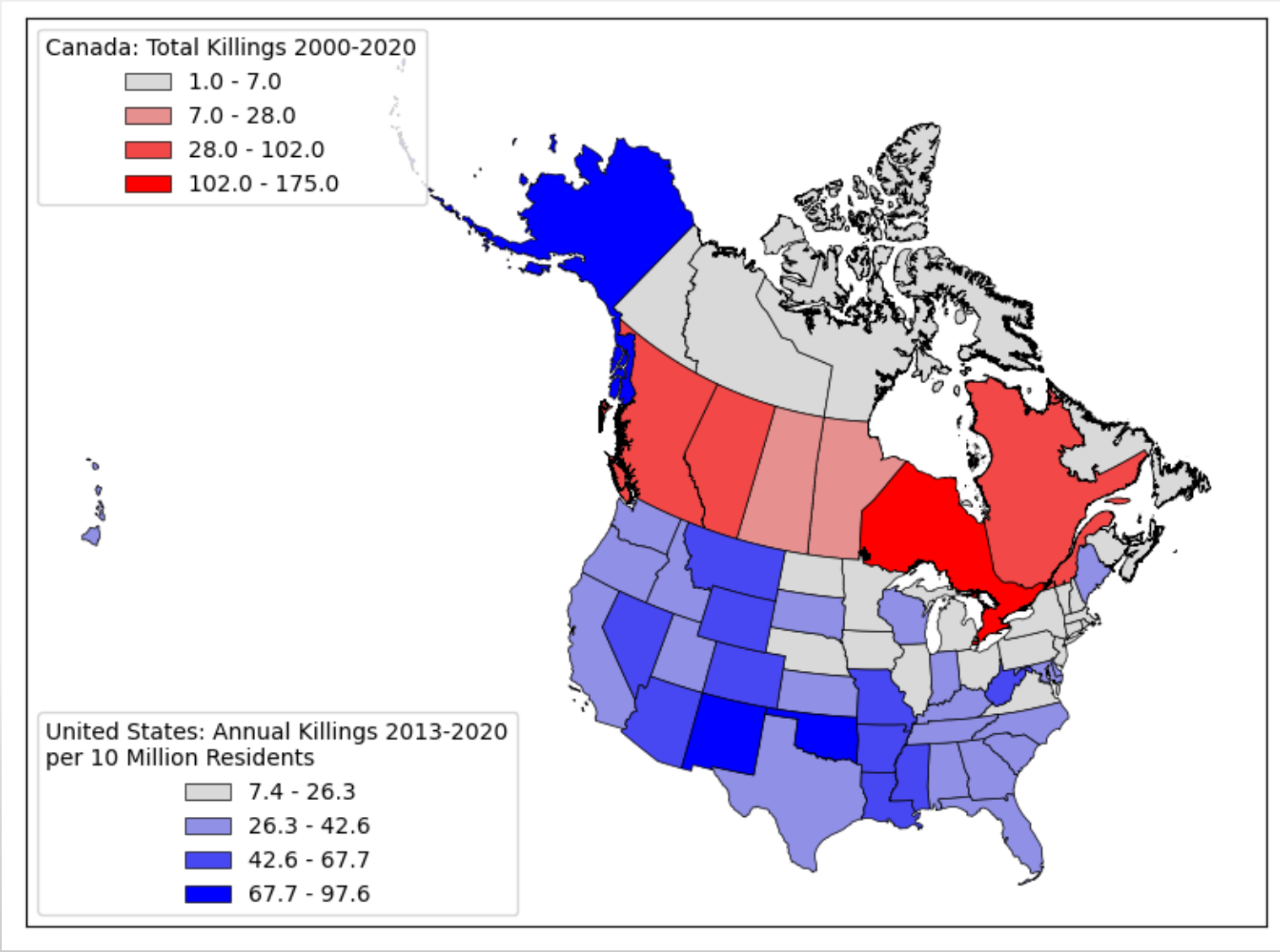
Which country's data is normalized on this map?

A) Canada

B) The United States

C) Both

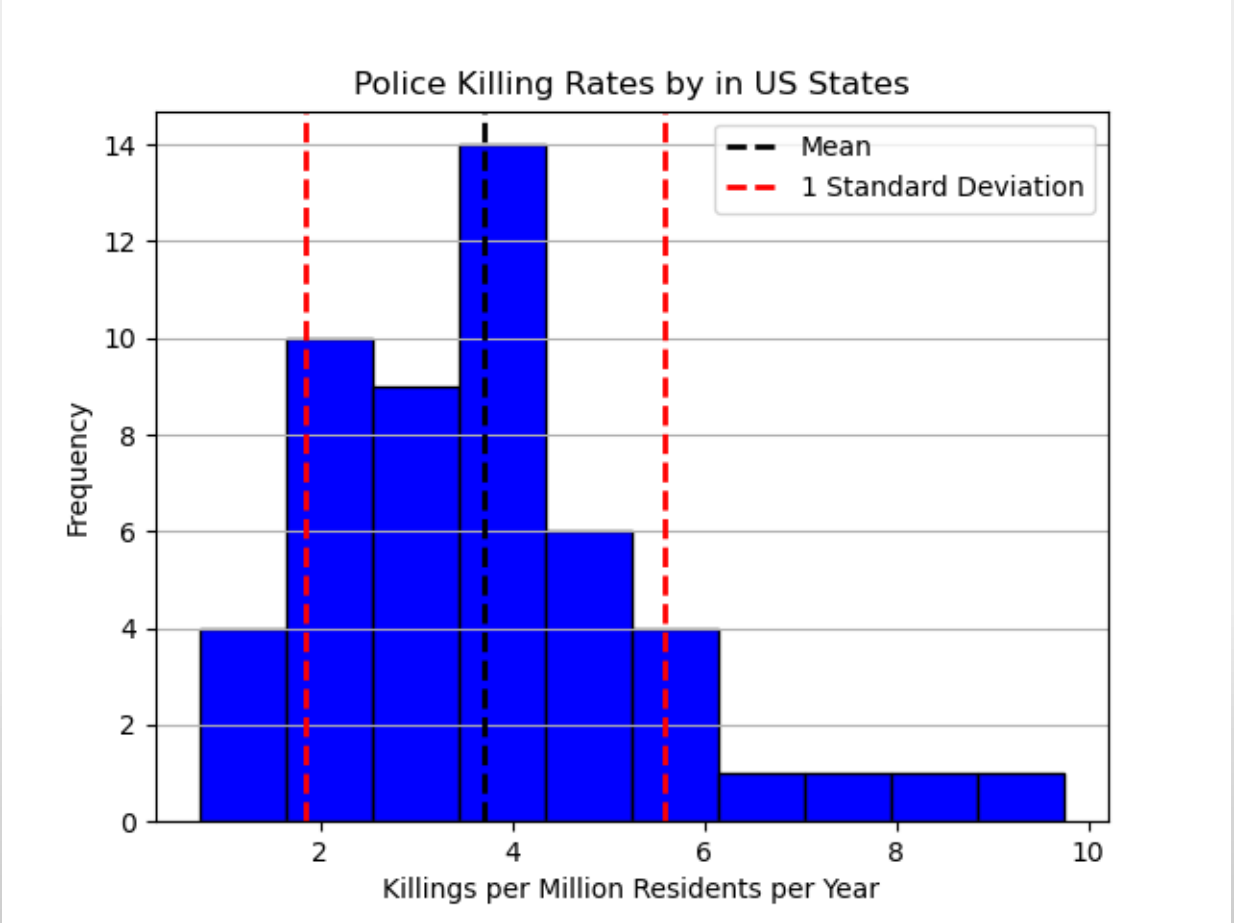
D) Neither



Part 3) Histograms and Data Classification

Rates by Province/State

- Police killing rates vary by administrative divisions, e.g. (State/Province).
  - If we want to compare rates, the first step is to look at histograms.
- A histogram shows us the frequency distribution of a given variable.
  - Data is grouped into a set of bins and counted.

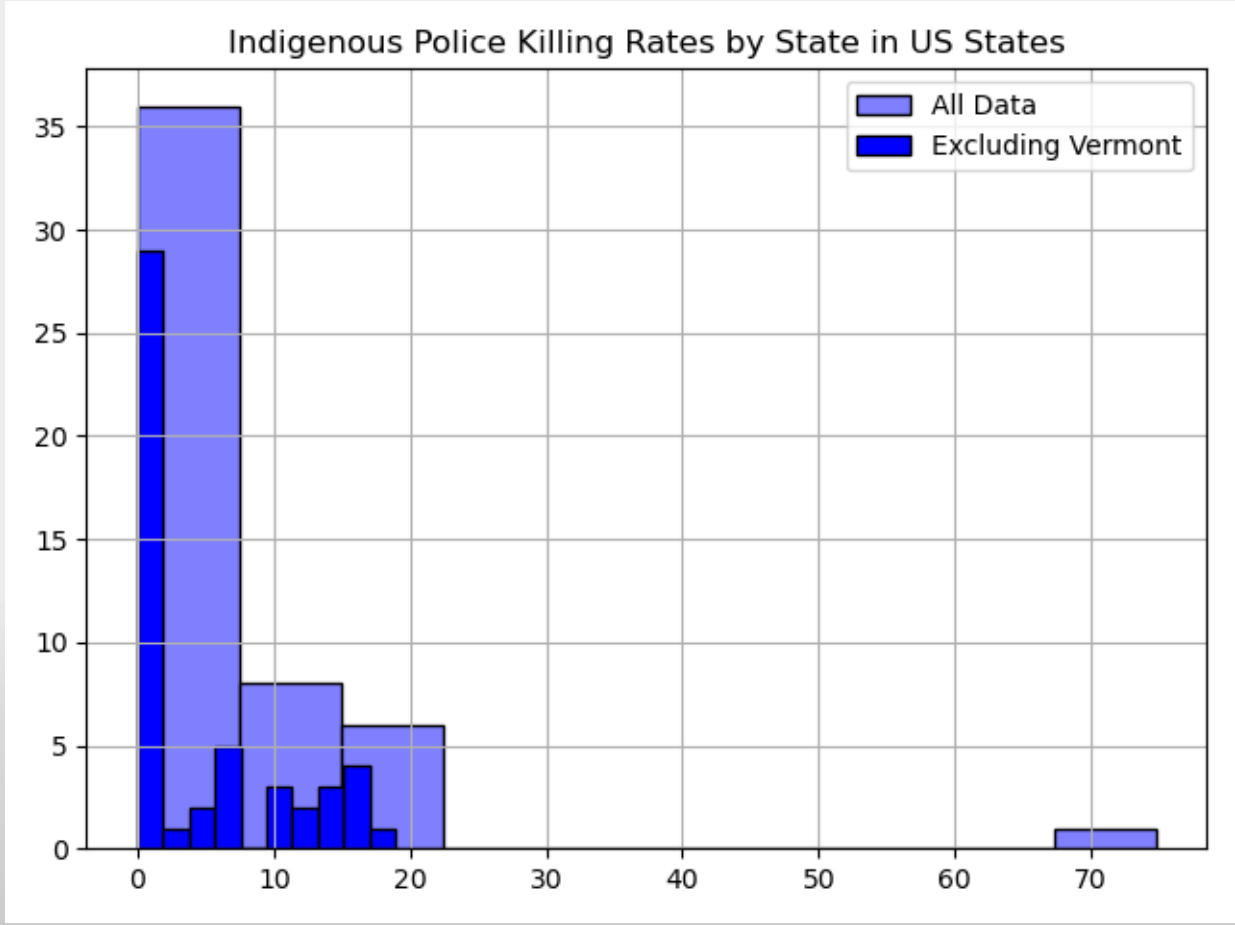


Outliers

Histograms can be useful for spotting outliers in a dataset.

- The Indigenous Police Killing Rate has a significant outlier.
  - Vermont has a rate many times higher than the nearest value.
  - Over half the states have zero indigenous killings.

	STUSPS	Indigenous_Rate	Indigenous_Killings	Indigenous	Indigenous_Fraction
0	VT	74.707	1.0	1743	0.278
1	ND	18.931	6.0	41270	5.430
2	ID	16.965	3.0	23026	1.313
3	WY	16.091	2.0	16185	2.801
4	WA	16.054	12.0	97329	1.292



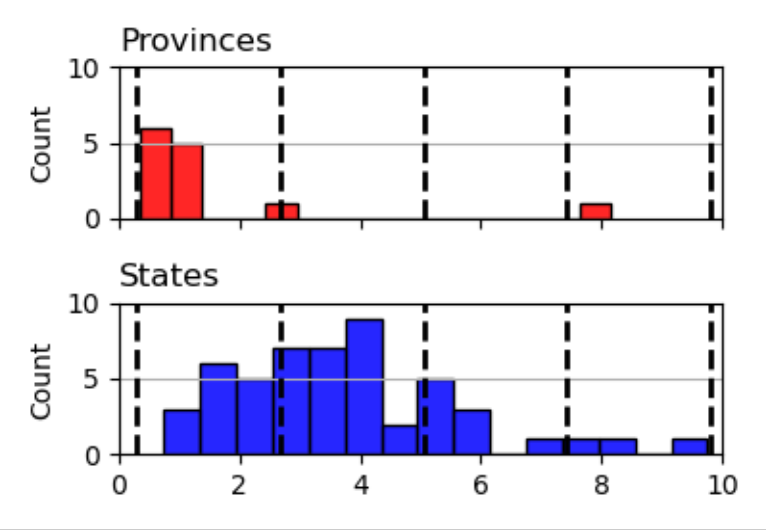
Classification Methods

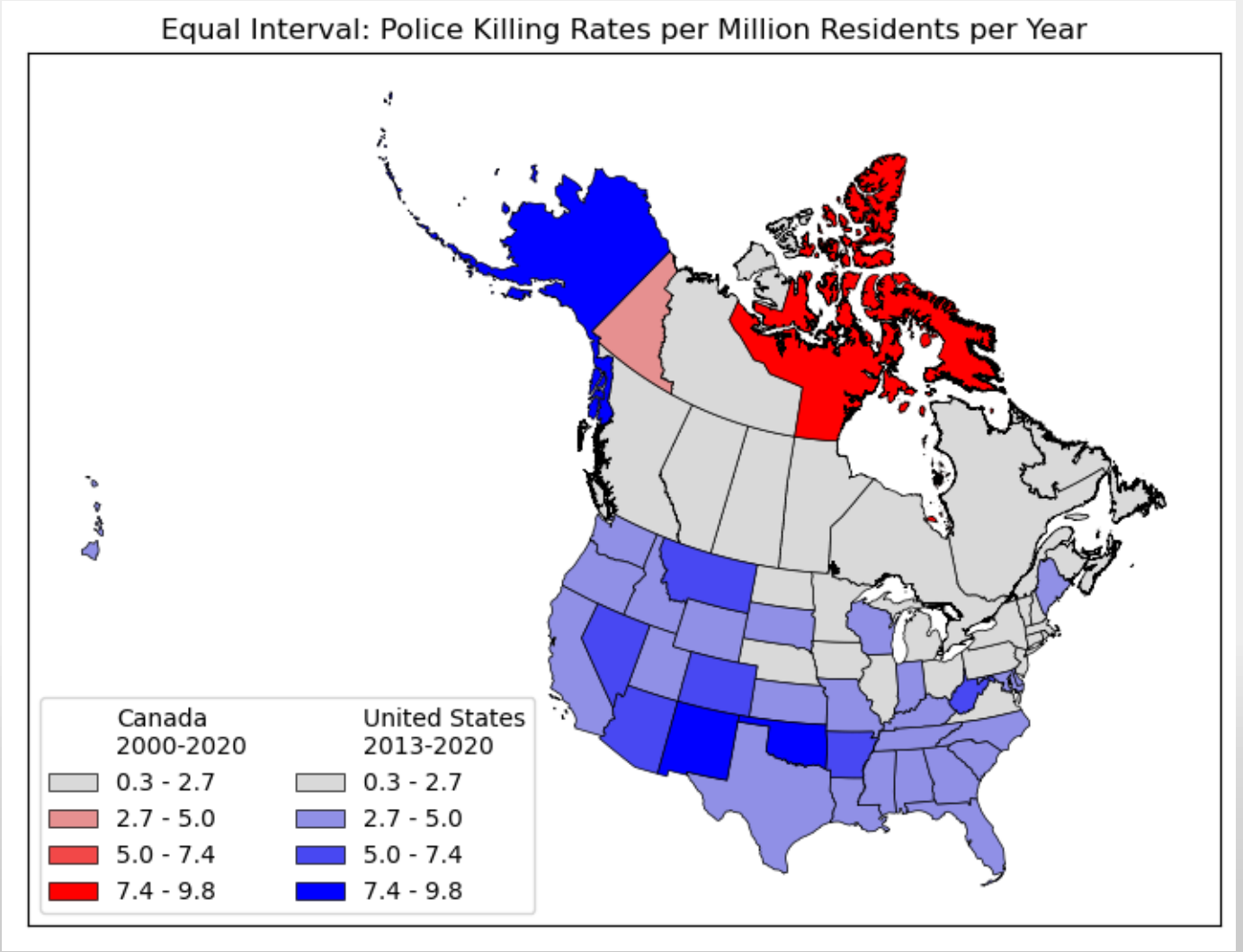
We'll cover five classification methods

1) Equal Interval
   - Data is split into bins of equal width regardless of distribution
2) Quantiles
   - Data is split by percentiles
3) Natural Breaks
   - Data is split using the Jenks algorithm
4) Standard Deviation
   - Data is split into bins based on distance from the mean
5) Manual Breaks
   - Data is split into bins of equal width regardless of distribution

Equal Interval

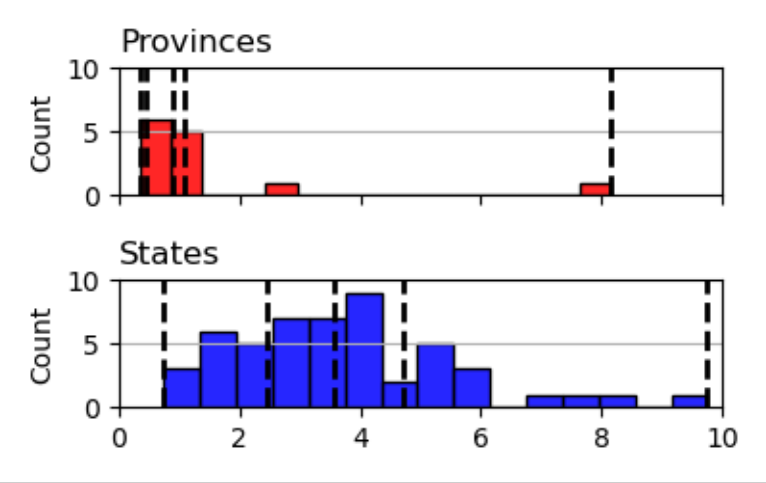
- The simplest classification scheme is to just break the data into classes of equal sizes
  - The minimum is .3 (PEI) and the maximum is 9.8 (NM)
    - Split into four bins 2.4 units wide

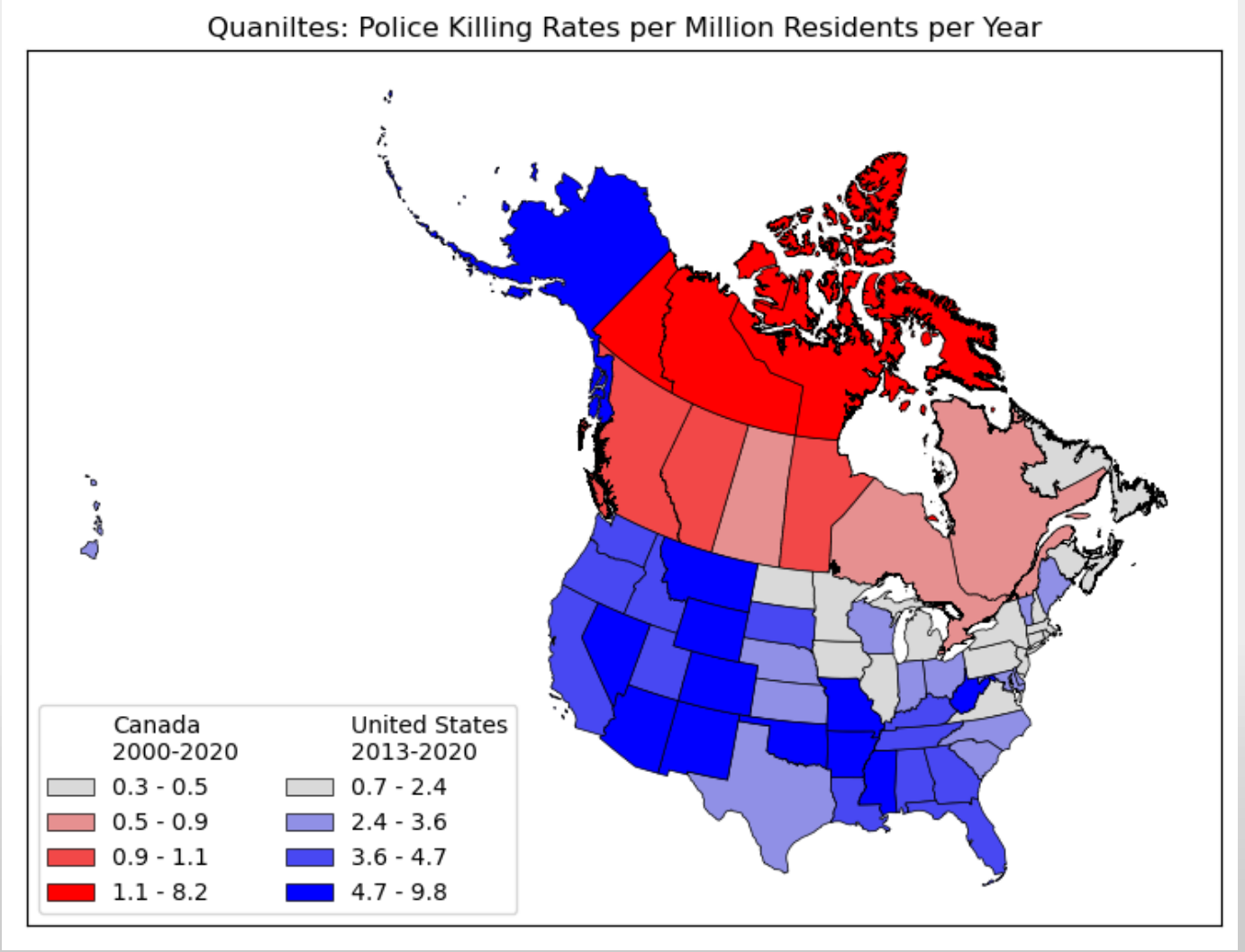




Quantiles

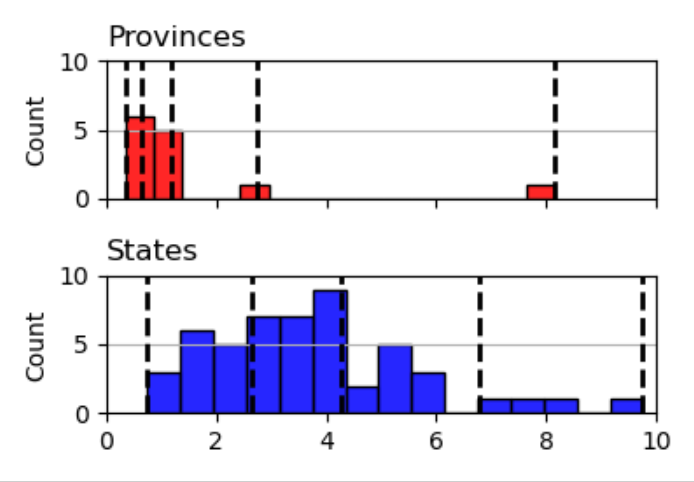
- Simple scheme based on the distribution of the data
  - All bins have about the same number of members
    - Canada, each bin has 3 members, except 0.3-0.5 has 4

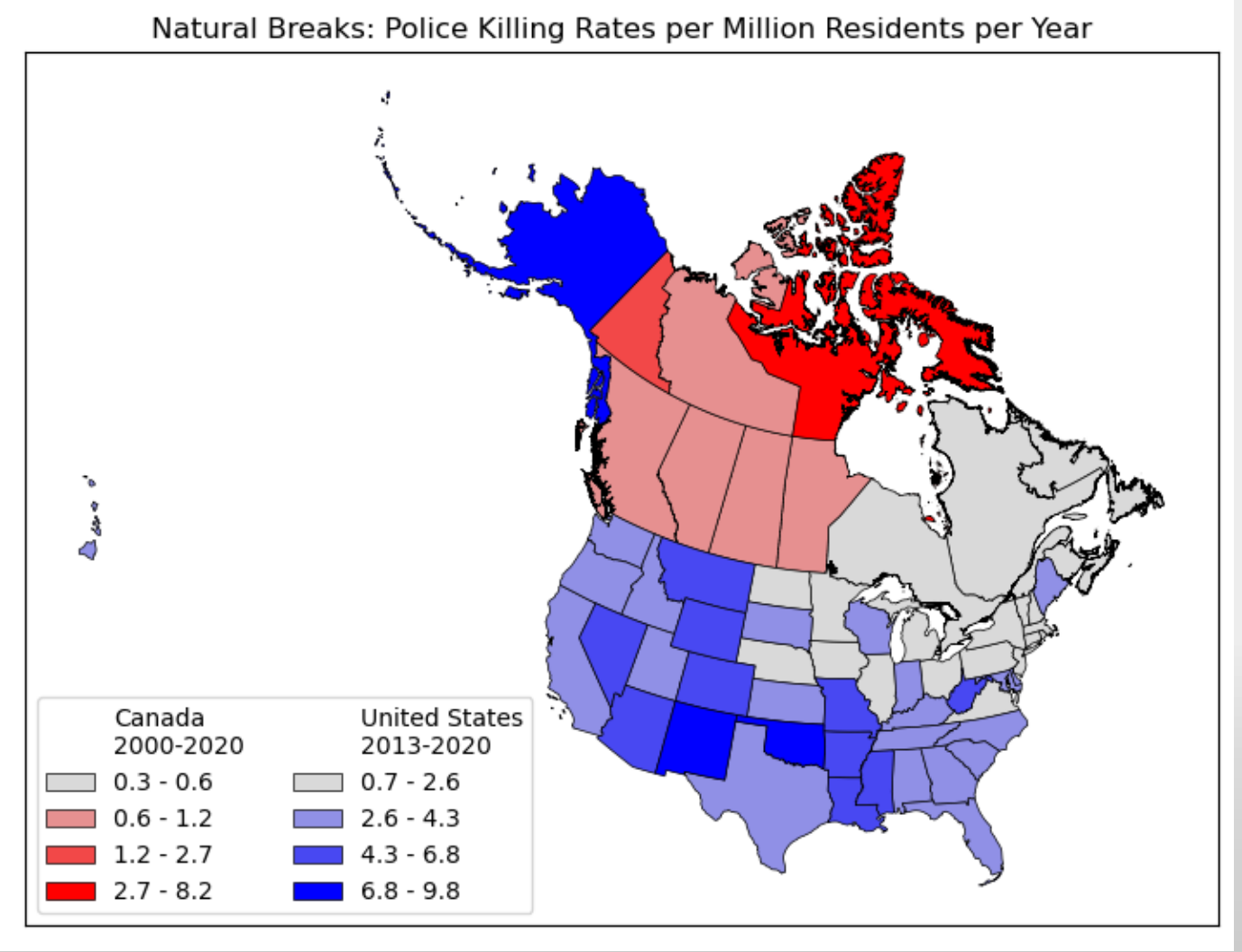




Natural Breaks

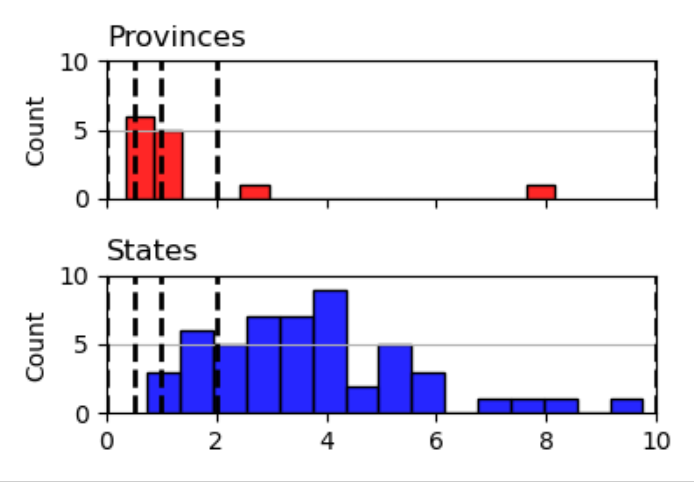
- Uses the Jenks algorithm to determine bins based on data distribution
  - Maximize within group similarity and between group dissimilarity

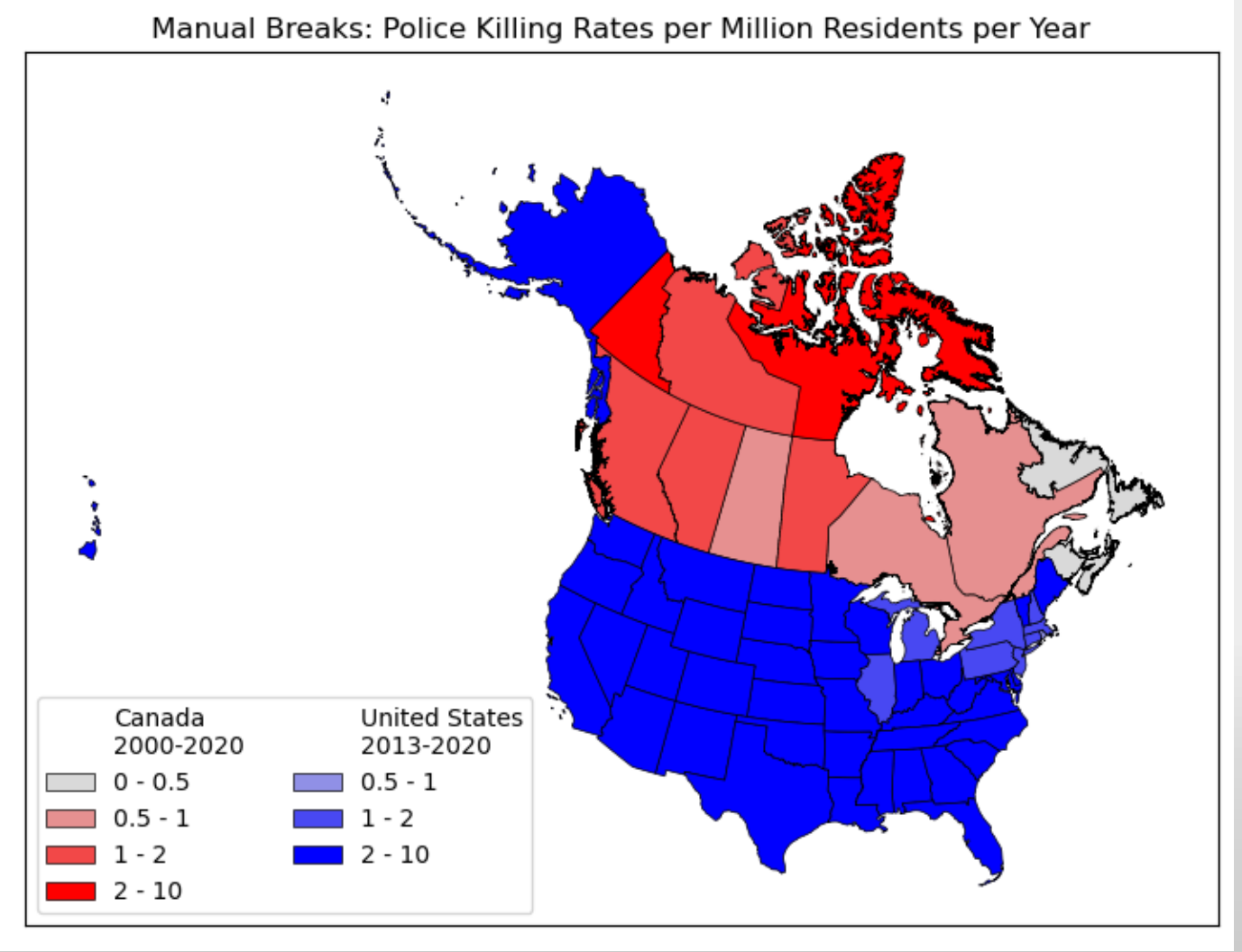




Manual Breaks

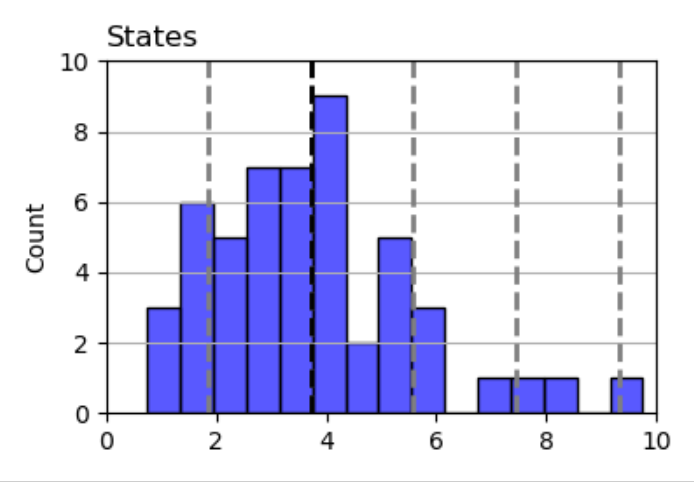
- We can define our own breaks
  - Best for comparisons
  - We can choose more “intuitive” break values
    - Whole number, halves, quarters etc.

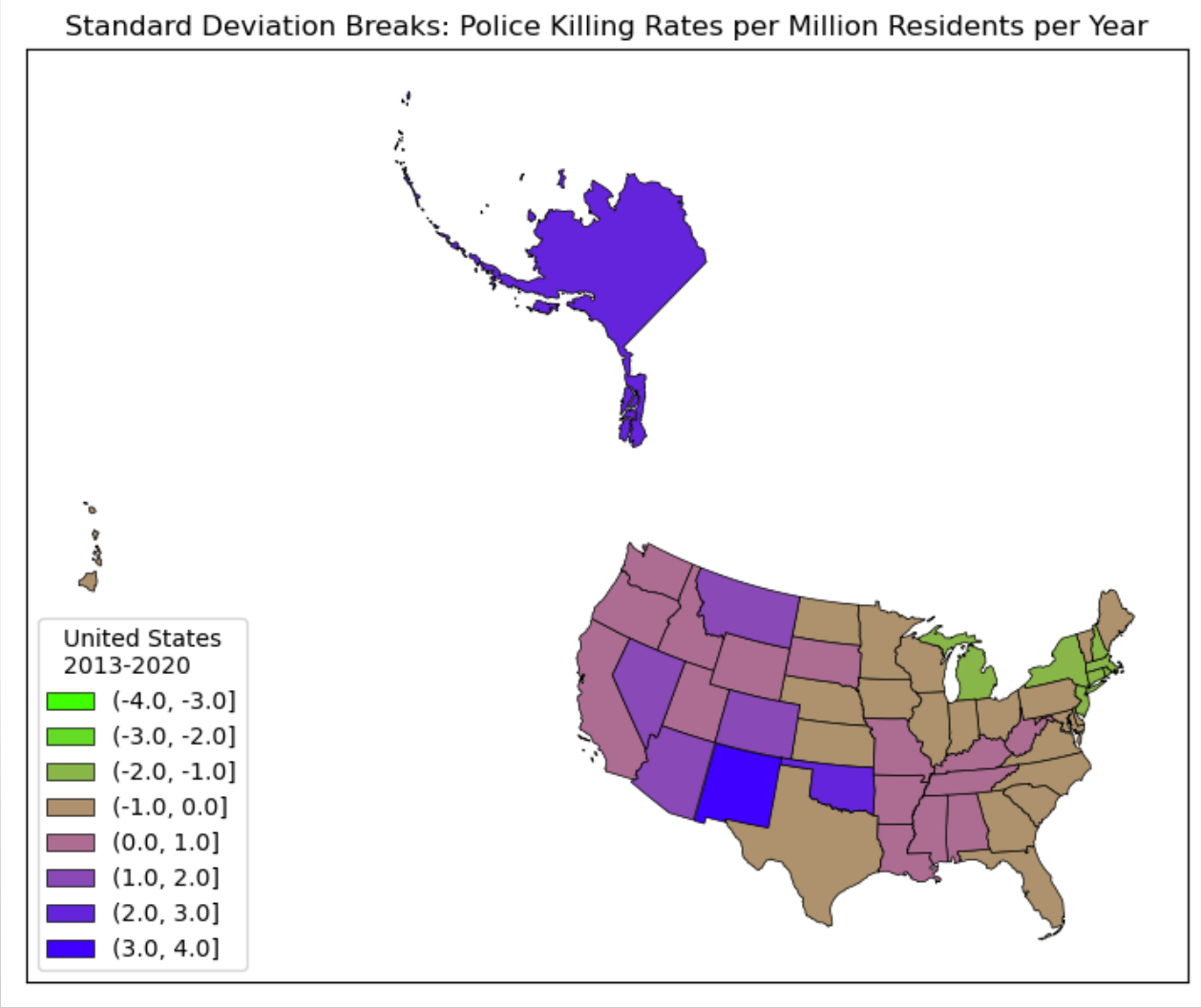




Standard Deviation

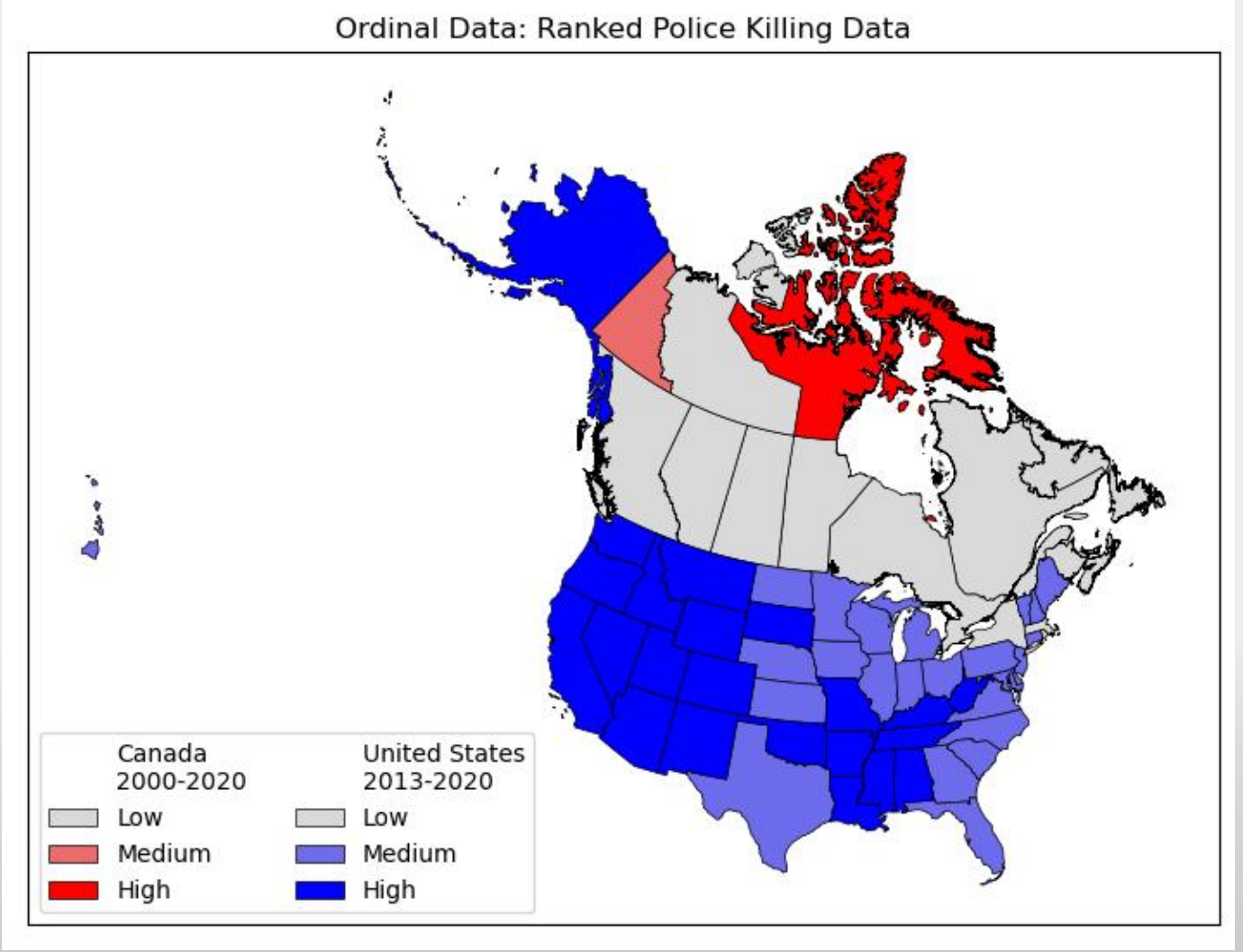
- Alternatively, we can classify data by distance from the mean
  - This will reveal different patterns in the data





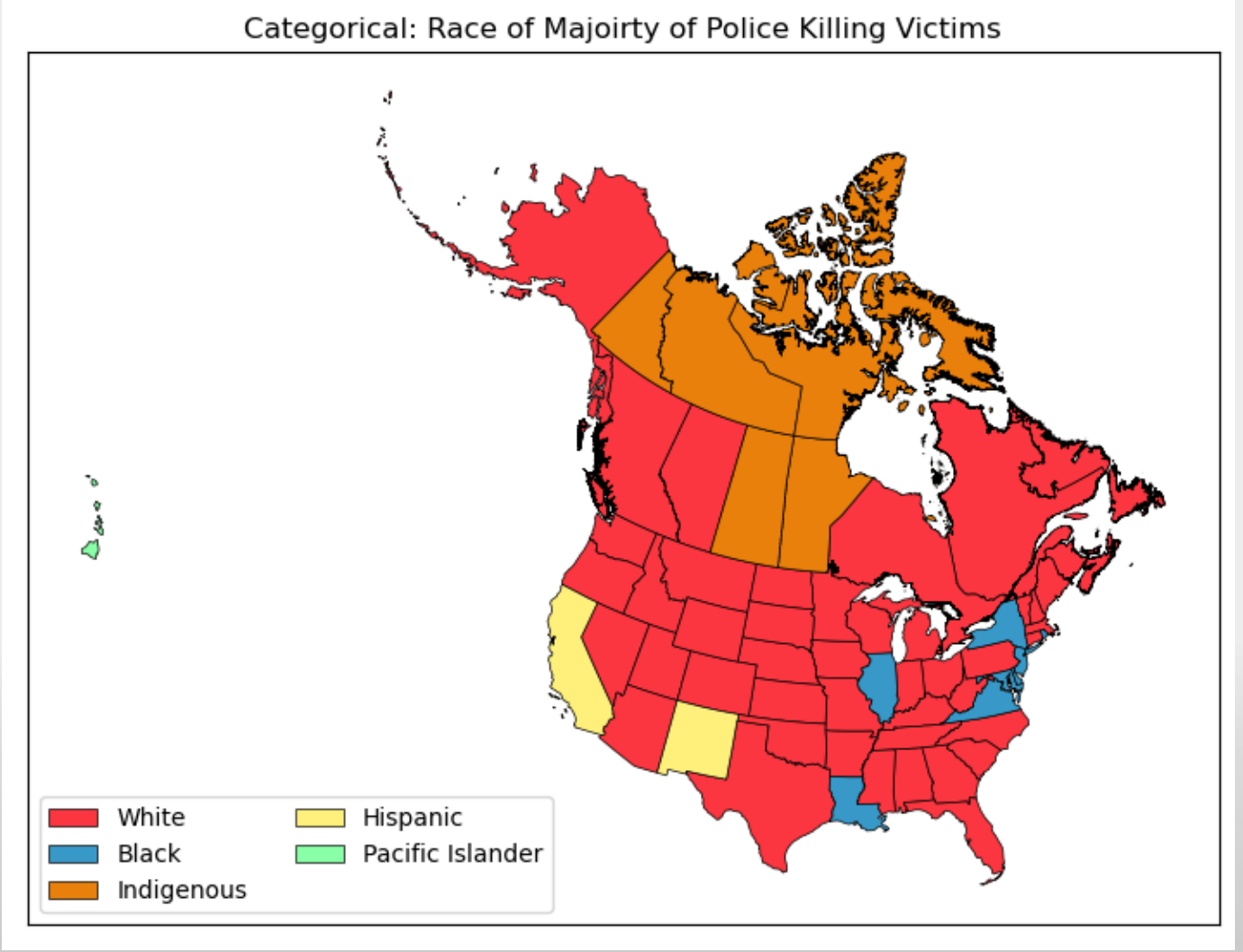
Ordinal Data

- Ranked (Ordinal) data can be derived from ratio or interval data
  - We can see the order, but what are the values based on?
- Ranked data can also be arbitrary (not derived from quantitative values)
  - Ex: spice level (mild, medium, hot)



Categorical Data

- Many data points have no rank/order
  - Displaying categorical data requires a different color scheme
    - If you use a graded scheme, you risk implying a rank/order that does not exits
- This map shows the racial majority for each state/provinces police killings



Questions

1) Which country has a higher frequency of police violence?

A) Canada

B) The United States

C) They're about equal

2) Which country has a greater racial disparity in incidents of police violence?

A) Canada

B) The United States

C) They're about equal

3) In which country are police more likely to kill an unarmed person?

A) Canada

B) The United States

C) They're about equal

4) Which classification schemes are best?

5) Think back to the Categorical Map. What is wrong with displaying the data this way? What might be a better way.