Version: Jun 1, 2022

Assessment

Data Science Fellow | CZ Research + Data

Pre-work. Please provide a work sample: a Python or R script you wrote in the past for data cleaning and/or engineering. Include the file with your responses below.

Overview

Please spend no more than 2 hours on this take-home assessment. Please send a copy of this document, your responses, any code you wrote for this, and the files we ask that you create in **Part 3** of this assignment in a .zip file to CZ's Lead Data Scientist, Andrew Zaharia at andrew@campaignzero.org.

You are free to use any resources, whether they be books or internet resources, to assist you. For any sources used, please cite them where appropriate in your response. All work for this assessment must be your original work, and not include previous work you have done. If there are parts of the assessment you won't be able to fully complete, don't sweat it – you can write bullet points about what you'd do. We want to see how you manage and prioritize your time.

Please indicate in your response how much time you spent on each part.

A press request for Mapping Police Violence

We occasionally get press requests which require quick analyses of our data. Here is an example of a real request we got: "I am working on a story about police traffic stops. We were wondering if Mapping Police Violence happens to have data on the number of civilian deaths that occur during traffic stops each year."

Mapping Police Violence (MPV) is a public dataset that CZ runs that tracks lives lost during police encounters in the US and compiles information around the circumstances of these incidents.

Trigger warning: this data contains demographic information about the civilians killed and information about the manner in which they were killed. The `circumstances` field contains brief descriptions of the incidents, some of which are graphic in detail.

Part 1

The **police_killings.csv** file provided to you contains a snapshot of this data. Each row corresponds to one victim from one incident. This file includes both the finalized data that we've made public, and internal-only, incomplete data that is in the process of being populated before we publish it. See **codebook.pdf** for a comprehensive list of each field and its explanation.

How many traffic stop-involved police killings (TSPKs) are there in total? What proportion of all police killings do TSPKs comprise? Briefly describe how you arrived at these numbers and explain any decisions you made in the process.

Part 2

What is the total number of incidents each year? How do these incidents break down by race? What are the top 3 agencies responsible for these incidents, and how many TSPKs are each responsible for?

Part 3

To prepare to launch a campaign on TSPKs, the front-end developers need 2 tables:

1. For every state, the absolute number of TSPKs, broken down by race as well as the total for all races.
2. The same table, but with TSPKs per capita instead of absolute numbers.

Create the two tables as **tspk_abs.csv** and **tspk_percap.csv**. Which state has the highest per capita rate of TSPKs overall and what is it? Which state has the lowest per capita rate, and what is it? How do these numbers compare to the national per capita rate?

Use **race_eth_by_state_2020_census.csv** (provided) to compute TSPKs per capita.

Part 4

Do you have any concerns about the data used to generate the results in **Parts 1-3**? If so, what are they?

Part 5

We just noticed some `agency_responsible` fields were not input correctly. Some entries have trailing spaces, and some fields have multiple agencies entered, separated by a comma. This may have impacted the answers we got in **Part 2**.

Write a script that re-computes the top 3 agencies and their TSPK counts, correcting for these data entry issues, and attach it with the rest of your code and answers.

Does making this correction change the answer in **Part 2**?