Exercise 1 | **Interviewing Police brutality dataset**

Author: Ludek Cizinsky

Date: 07. 02. 2022

TASK 1 | **EDA**

EDA | **Exploration and brainstorming 1**

Properties of the dataset

- There are no missing values in the first sheet
- There are in total 60 rows representing info about region/city corresponding to certain police department
- The features include population of the given region, police killing rate in that region etc. In addition, many features are splitted according to color skin
- The data in certain columns come from different years
- There is also a disparity column which I do not know how to interpret even after googling
- In the sheet with incident level data, there is a detailed description of crimes which can be used to get low level details about incidents

Transformation of data

- We might want to divide the data into different categories by for example state they come from - e.g. southern states
- We might want to just look at columns which are normalized by the population of the given region, otherwise certain metrics are not comparable
- Perhaps to look at the detailed description of incidents, we could join the first sheet with the other sheet using the columns of police department

EDA | **Exploration and brainstorming 2**

Questions

- The fact that there is a split according to skin color might lead to certain types of questions based on skin color, for example:
  - Is there a higher killing rate in regions with higher black population? Of course this does not necessarily imply racism, but would indicate that something wrong might be going on
- More general question using the high level data:
  - Is there a difference in killing rate between southern and east coast states
  - What are states/regions with highest/lowest police killing rate? And what could be the reasons?
  - What are the states/regions with highest/lowest violent crime rate? And what could be the reasons?
- More general questions using the low level data:
  - What are the most common words occurring within description of incidents?

TASK 2 | **Concepting**

Concepting | **Plot selection**

Question 1: Is there a higher police killing rate in states with higher population of black people?

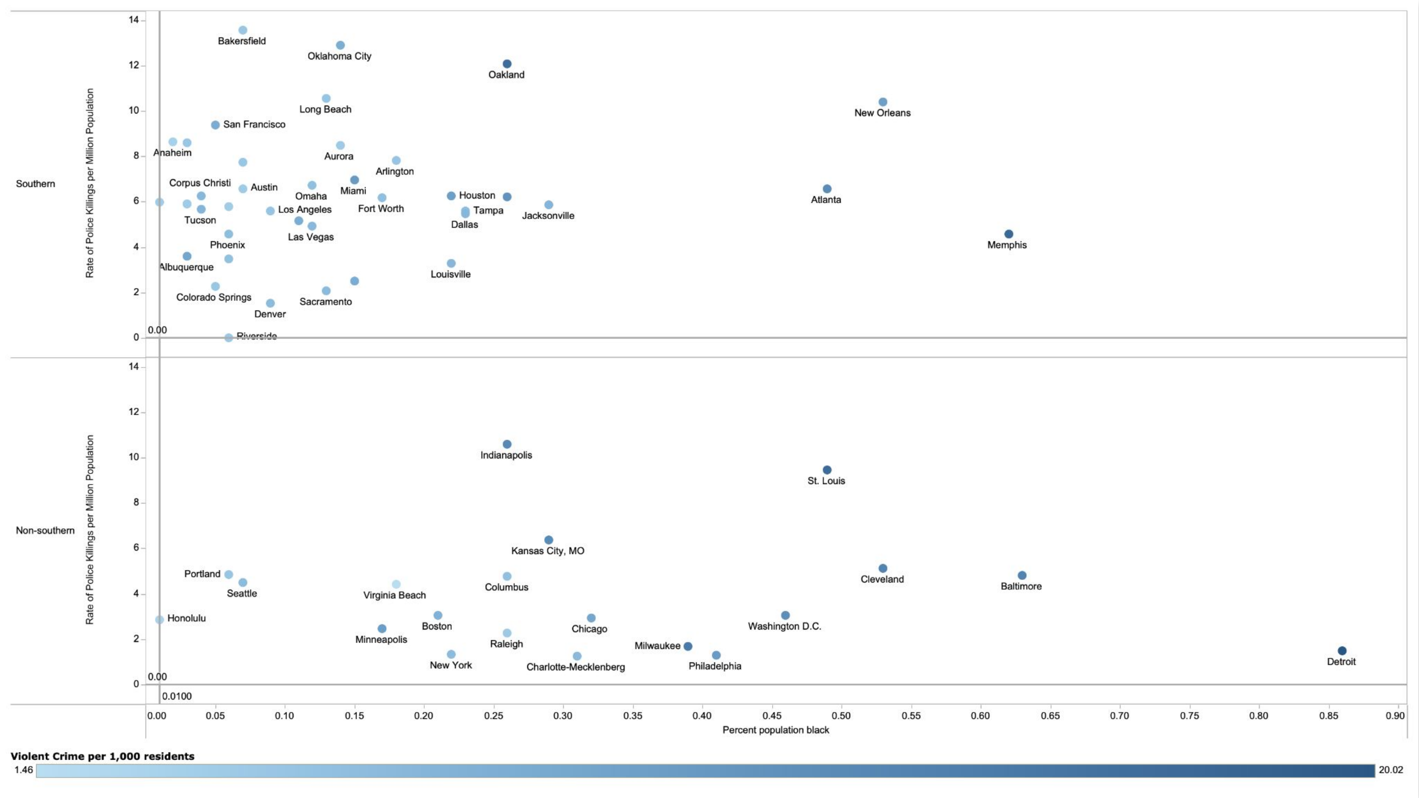
- One possible chart to use would be a **scatter plot**. I would plot on x-axis the percentage of black population and then on y-axis I would put police killing rate per million population
- The scatter plot could be further extended with labels of police departments
- In addition, a third dimension could be added representing the state in which the given region is. This could be done for example using different colors or shapes. This would give us an idea if the problem occurs not only on regional level but also at a state level

Question 2: What are the most common words occurring within description of incidents?

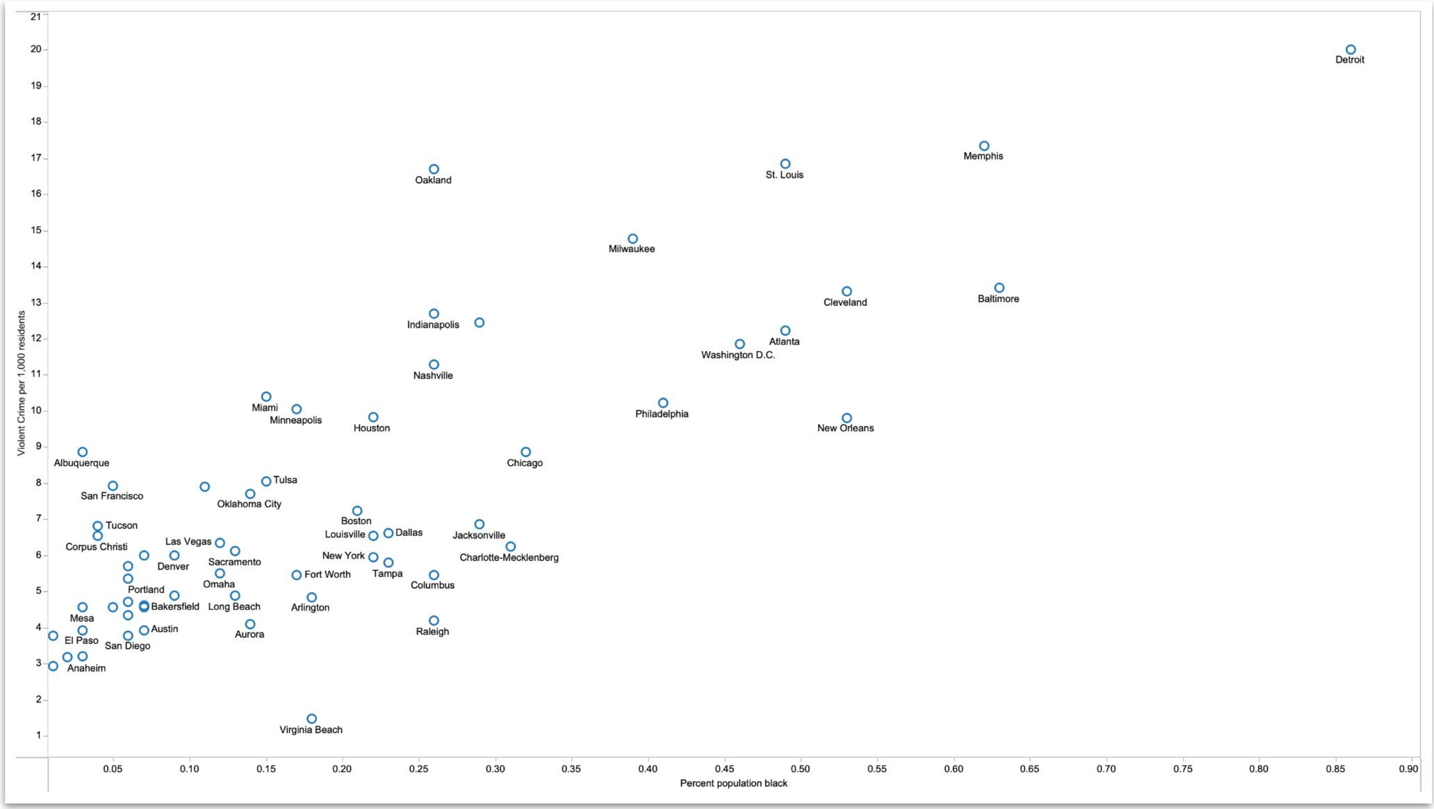
- Here I would use **word cloud** which would make words with high occurrence large and vice versa
- Looking at this plot, we could tell if there is some interesting pattern which we could further examine using quantitative data
- The description data might need to be transformed such that common words like “the” would be filtered out

TASK 3 | **Editorial brainstorming**

Editorial brainstorming | **Percent of black population vs Rate of police killing**



Editorial brainstorming | **Closer look at positive relationship between black population and violent crime rates**



Editorial brainstorming | **Notes to the previous visualizations**

- In the visualization, I splitted the departments according to state into southern and non-southern. It can be nicely seen that non-southern regions have on average larger black population.
- We can also observe that there is no clear trend that would indicate the higher the black population the higher the killing rate
- From the third dimensions, however, we can see that there is a positive relationship between black population and violent crime rates. This is further supported by the next visualization. However, it is important to emphasize that this does NOT imply causality.
- We could further extend the visualization by looking at the selected regions at detail. For example Detroit seems as a good candidate and compare it with for example Oklahoma city. By details, I precisely mean looking at historic data from more years and try to see if there are any patterns.