**SMU Data Analytics Bootcamp | Project 4 Group 1: Write-Up**

**Title: -** Homicides More than 52,000 criminal homicides over the past decade in 50 US cities

**Contributors / Project Team Student Members:**

* Carlos Delarosa
* Raj Agrawal
* Ann Ly
* John Banowsky

**Faculty**

* Alex Booth - Instructure
* Sherhone Grant - TA
* Sean Fleming - SSM

**Homicide - Definition**

Homicide is the killing of a person by another with intent to cause death or serious injury, by any means. It excludes death due to legal intervention and operations of war.

In addition, [1] A homicide requires only a volitional act or omission that causes the death of another, and thus a homicide may result from accidental, reckless, or negligent acts even if there is no intent to cause harm.[2] Homicides can be divided into many overlapping legal categories, such as murder, manslaughter, justifiable homicide, assassination, killing in war (either following the laws of war or as a war crime), euthanasia, and capital punishment, depending on the circumstances of the death. These different types of homicides are often treated very differently in human societies; some are considered crimes, while others are permitted or even ordered by the legal system.

**Studies of homicide consequences**

Beyond its direct impact, homicide has serious negative effects on the lives of surviving family members, particularly children. Psychological effects include anxiety, depression, post-traumatic stress disorder, aggression, guilt and a heightened sense of vulnerability. Socio-occupational effects include problems in school and at work. Homicide may also lead families to incur expenses they can ill afford, such as funeral costs and lawyer fees. If the victim was a breadwinner, families may no longer be able to cope financially. Homicide can generate a sense of insecurity in society, and when high rates occur in countries with weak, inefficient and corrupt criminal justice systems, can contribute to undermining social and economic development.

**Studies of homicide risk factors**

Homicide is caused by mix of factors at the individual, relationship, community and societal levels. Demographic structure is a well-established risk factor for homicide. Societies where young people – particularly young males – make up a greater share of the population tend to have higher homicide rates. Transitions in political regimes may also be associated with increased homicide. The fall of Communism in Eastern Europe, apartheid in South Africa, and dictatorship in Brazil all saw rapid increases in homicide rates. Homicide rates may increase in the absence of good governance and effective rule of law. Homicide rates tend to be lower where states have legitimacy in the eyes of citizens, can deliver key political goods such as justice based on the rule of law, and have low levels of corruption. Poverty, economic inequality, ethnic fractionalization, and the availability of guns and alcohol are also risk factors for homicide.

**Examples of strategies and interventions**

This section contains examples of strategies that is required to minimize or eliminate

1. Community-based awareness
2. Safety educations

**DATA**

**Sources –**

1. *homicides\_data.csv* from “Homicides” – Dataset from Kaggle <https://www.kaggle.com/datasets/joebeachcapital/homicides>

The data set used from Kaggle is the main data set for this project. This dataset was base based on a Washington Post article from 2018. The data set contains details of over 52,000 homicides in 50 U.S. cities from 2007-2017. Further details of the original dataset and the Washington Post articles details can be found from the provided link.

1. *SubEst2010.csv* & *SubEst2020.csv* for Population Data – Datasets from US Census

* <https://www.census.gov/data/datasets/time-series/demo/popest/intercensal-2000-2010-cities-and-towns.html>
* <https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-cities-and-towns-total.html>

To add population analysis to the original dataset, we pulled predicted population data from the US Census.

**Reasoning –**

The homicide data from the Kaggle dataset contained demographic data, location data, time data and, most importantly, disposition data which shows whether the cases were open or closed and if an arrest was made. Apart from basic demographic visualizations, our intent is to create a predictive model to determine whether a specific homicide victim’s case would end in an arrest or not.

Since the data is from 50 different US cities with different population sizes, we will us Census data from the timeframe of the data set to incorporate the number of homicides per city by population size. This combination should improve our ability to observe patterns based on city size as well as incorporate population as a variable in our predictive model.

**Data Cleaning –**

We used Jupyter Notebook and pandas, pathlib, and numpy libraries.

*homicides\_data.csv*

The cleaning was composed of 3 parts: rewriting or removing data, binning data, and priming data for merger.

Rewriting or Removing Data:

* In identifying the states found, we changed the state of Wisconsin from ‘wI’ to ‘WI’.
* We found a city called Tulsa, AL which was meant to be Tulsa, OK.
* We dropped “victim\_last” and “victim\_first” columns as it was not relevant to demographic information.
* We found errors in in “reported\_date” that included extra 1’s in the dates.
* We used a df.column.replace to change the dispositions options to a binary outcome from “Closed without arrest” and “Open/No arrest” to “No Arrest” and “Closed by arrest” to “Arrest Made”
* We used the datetime features to take ‘reported\_date” into year, month, and weekday.

Binning Data:

* We binned victim\_age to '0-17', '18-29', '30-44', '45-64', and '65+' for use in Tableau.
* Using a function to create a season column from the month’s column.

Merging Data:

* We made an unique id for each city by creating a ‘LOCATION’ column from ‘city’ & ‘state’.
* We created 2 paths for this data: Machine Learning and Tableau.
  + Machine Learning include us removing rows that had missing demographic data. This included 3 entire cities: Dallas, Phoenix and Kansas City. It dropped the data set from 52179 to 47478 counts of usable data.
  + Tableau did not drop these cities but focused on being usable for total counts.

*SubEst2010.csv* & *SubEst2020.csv*

The population data needed to be cleaned to identify the actual cities and identifying the specific years needed.

* We changed state names from the full name to the 2-letter abbreviation to merge with the homicide data.
* We edited the city columns to clean up the extra designators attached to the cities we needed.
* We created the same ‘LOCATION’ unique id column from ‘NAME’ & ‘STNAME’
* We reduced the data down to only the tables we needed.
* We then merged the 2010 and 2020 datasets.
* We then only kept the 50 (47) cities by using the LOCATIONS saved from the homicide data.
* We used a melt function to create a row for each year per city.

*Merged Data*

We merged the homicide and population data sets using the LOCATION to include population data in the homicide data.

We continued to create a dataset with homicide per 100,000 by city and year.

**Cleaned Data Files –**

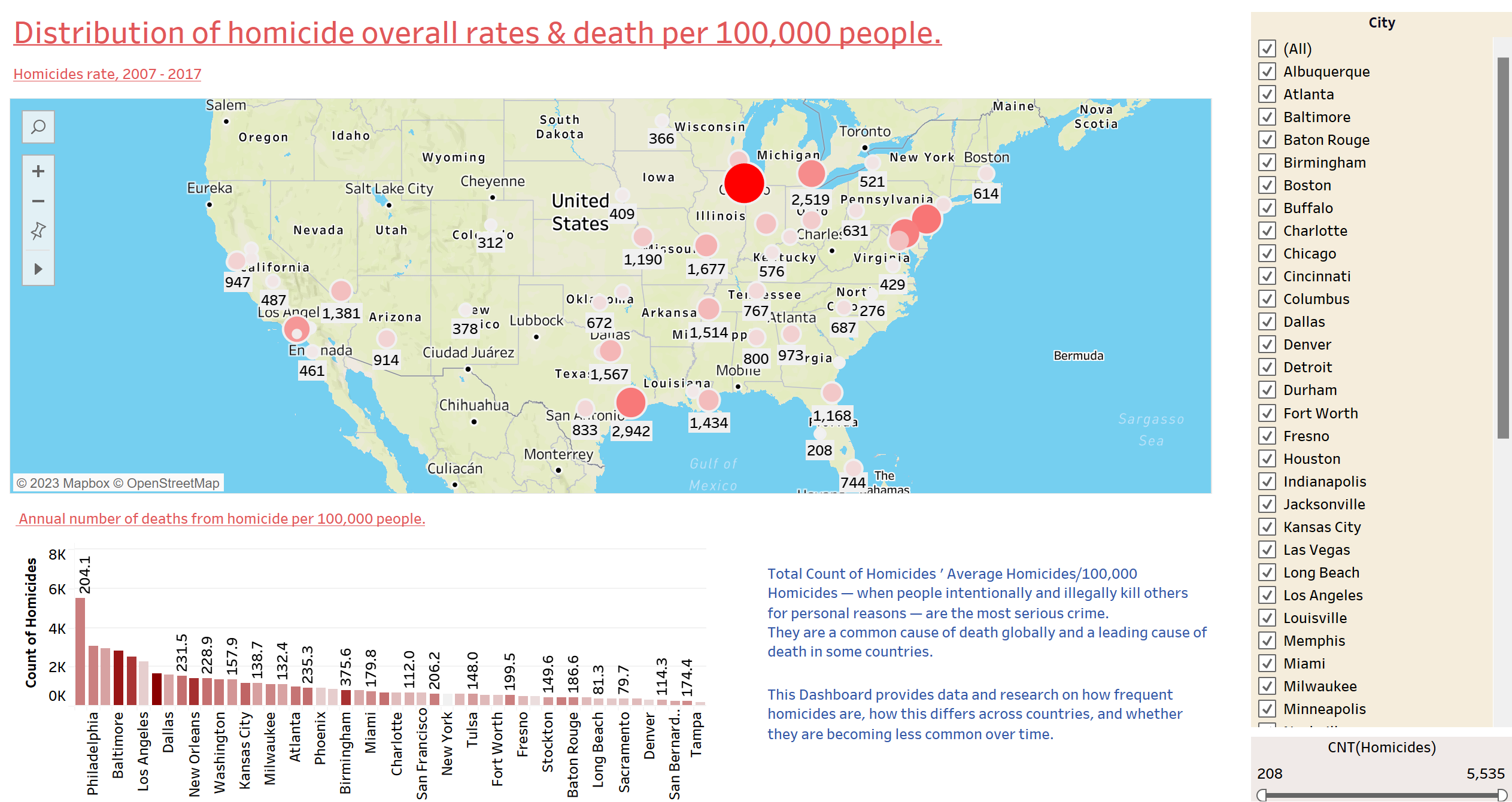
* ml\_clean\_homicide\_data.csv
* tbl\_no\_drop\_homicide\_data.csv
* tbl\_homicide\_count\_per\_100000.csv

**Data limitations –**

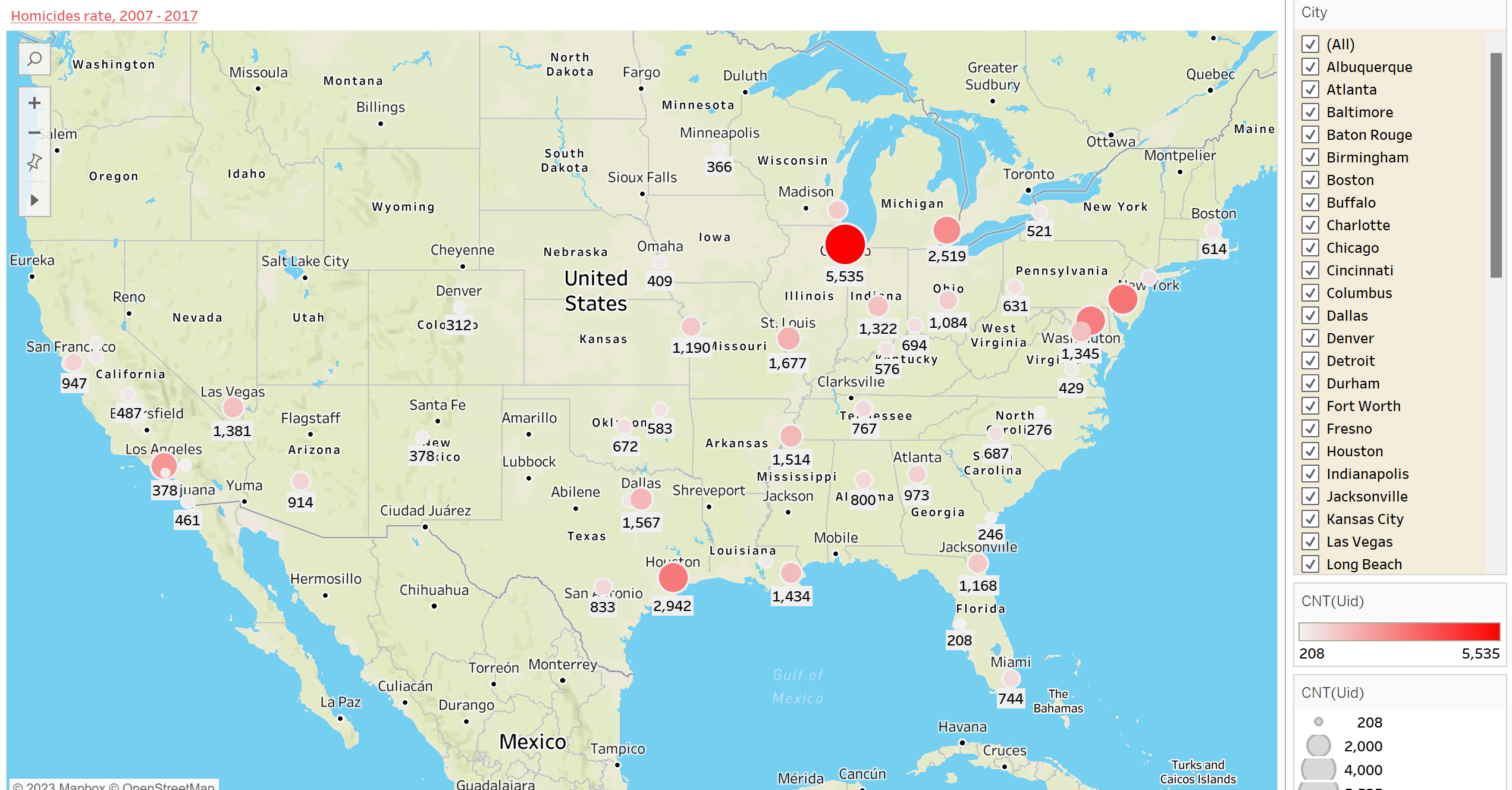
* The data is limited to only 50 US cities.
* The data does not include specific details of the homicide such as number of victims involved, time of day, weapon, specific location type, etc. These details could help determine whether an arrest could be made.
* The unknown data dropped our usable machine learning data substantially.

**Tableau screen shots---**

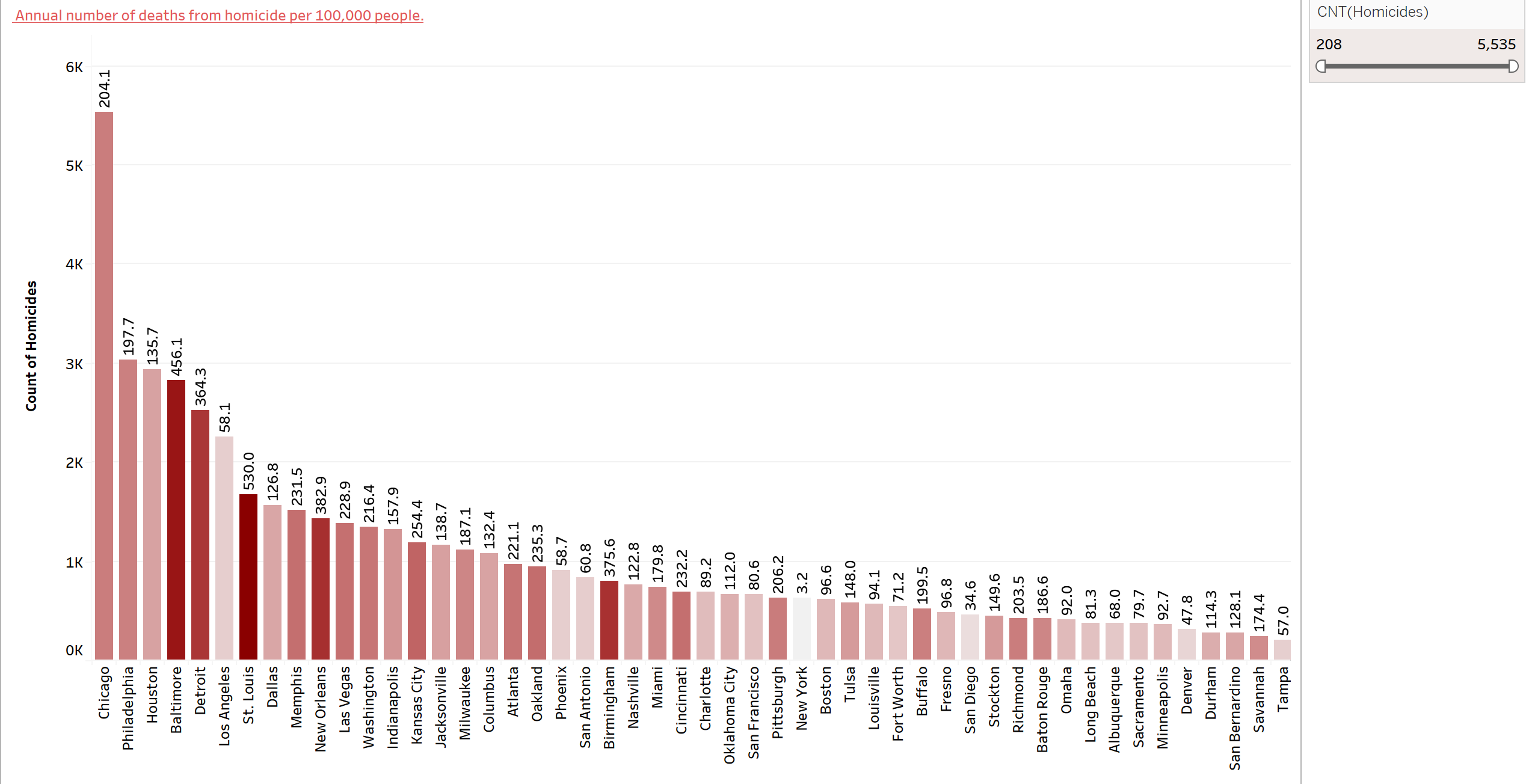
**SMU\_Proj4\_Group1\_homicides\_TD1**



**Tableau dashboard – main page explains about distributions of the homicides across US cities. The dashboard has filters based on Cities & a range of count of homicides**

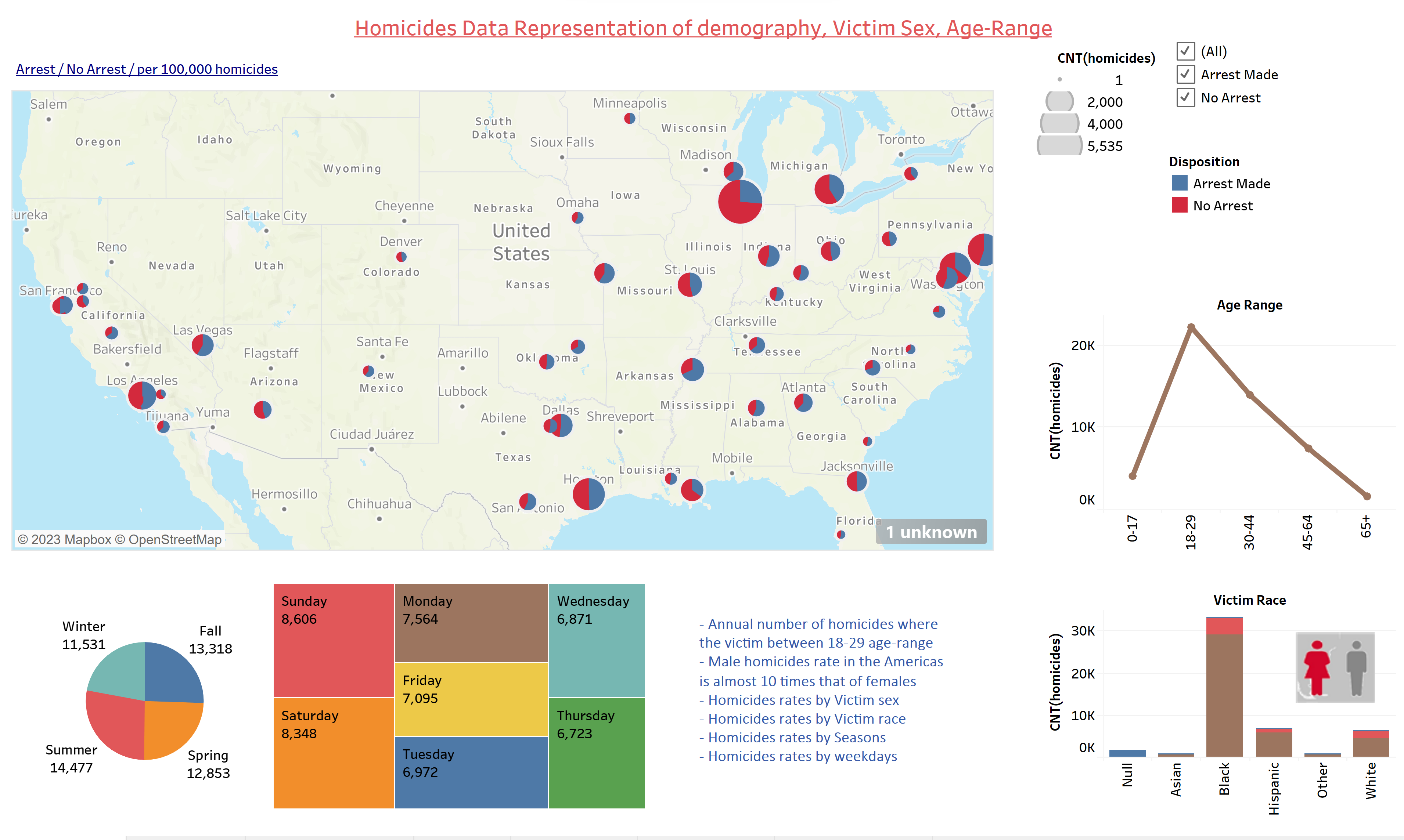


**Tableau workbook – part of dashboard**

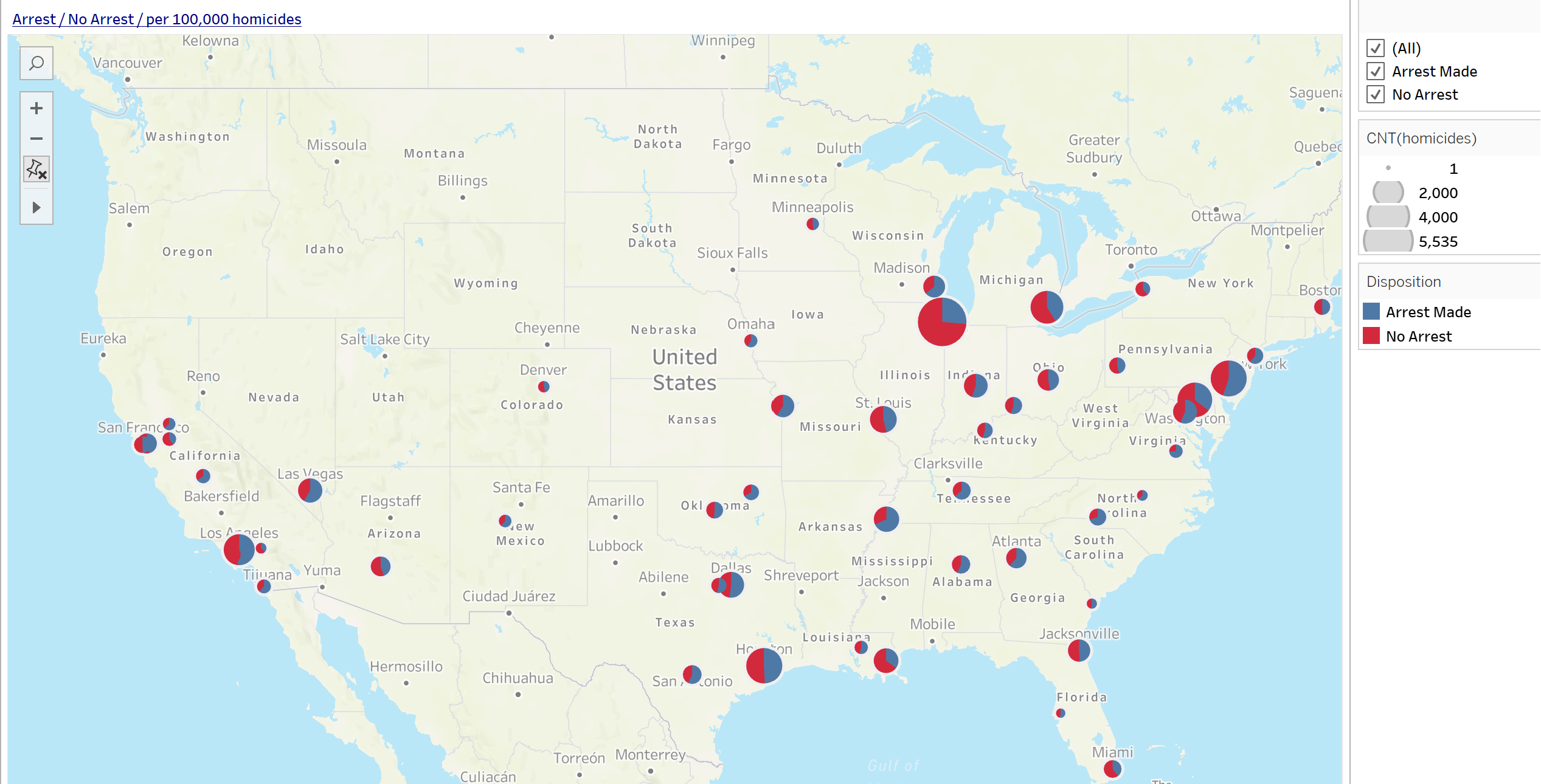


**Tableau workbook – part of dashboard**

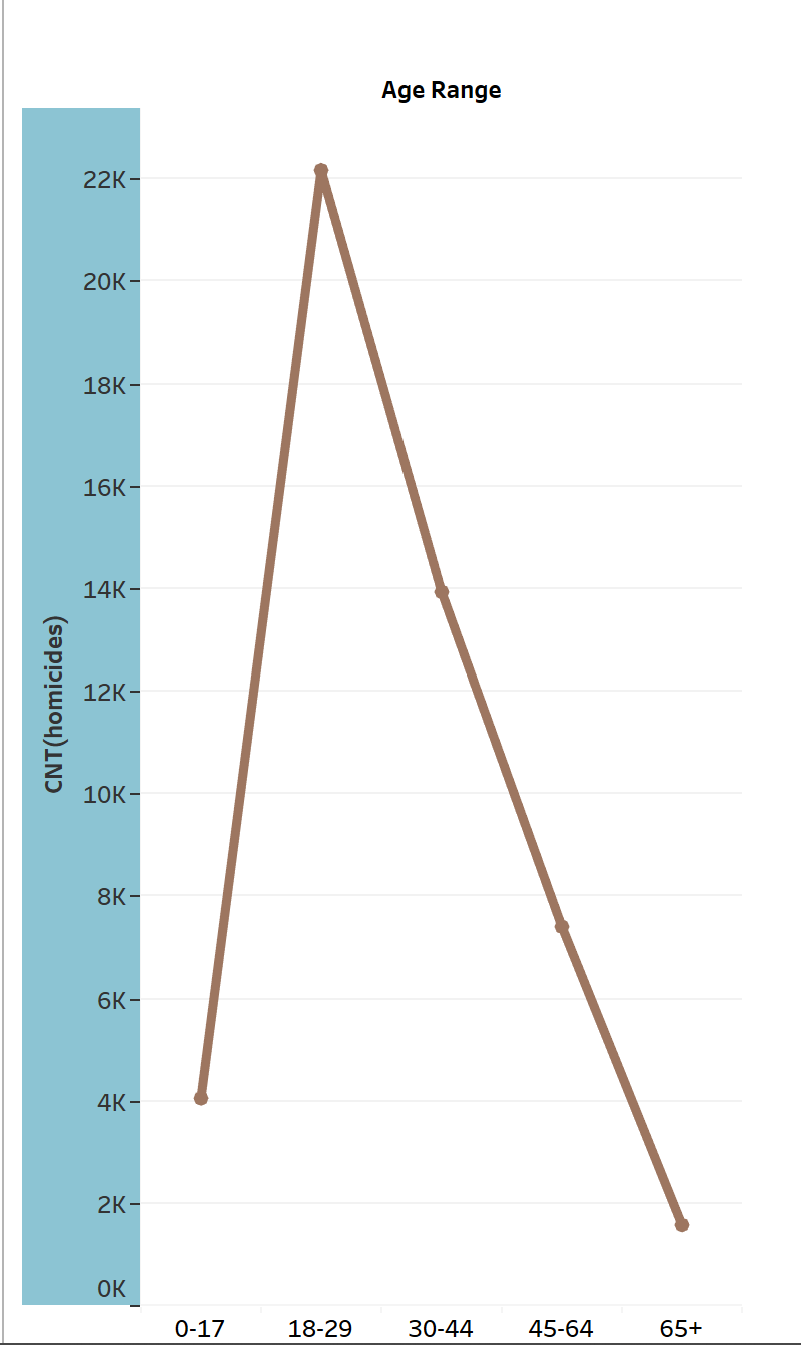
**SMU\_Proj4\_Group1\_homicides\_TD2**

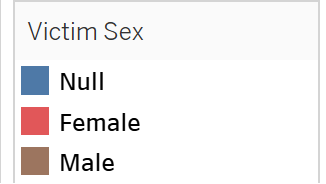


**Tableau dashboard – second page explains demography wise distributions of the homicides across US cities. The dashboard has filters based on “Arrest” “no arrest” & a range of count of homicides**

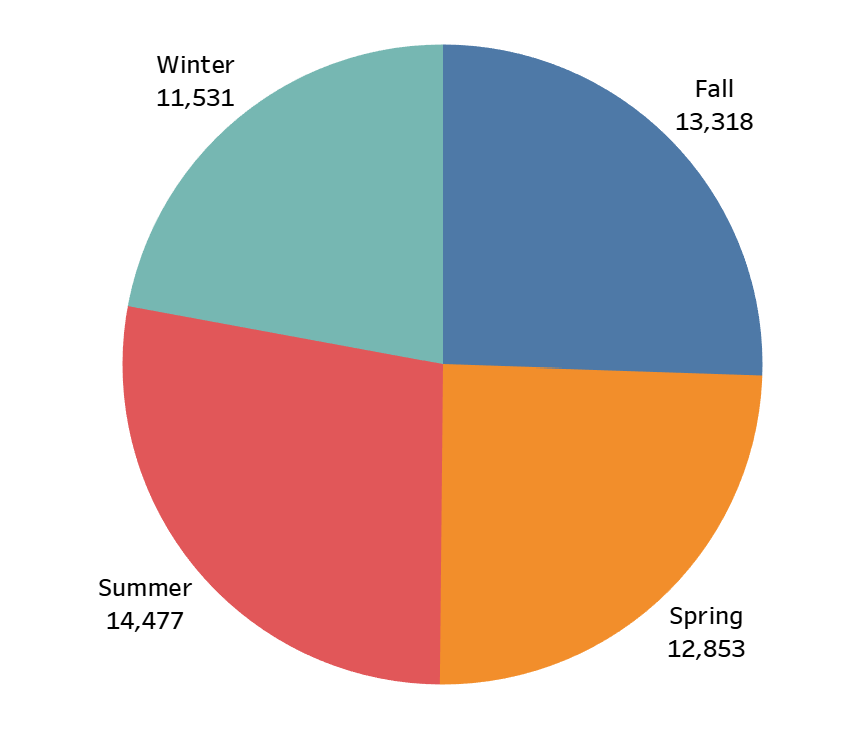


**Tableau workbook – part of dashboard**





**Tableau workbook – part of dashboard**



**Tableau workbook – part of dashboard**



**Tableau workbook – part of dashboard -CARLOS – WEBSITE writeups**

"The development of the homicide website is driven by the goal of offering a valuable online resource for individuals interested in gaining insights into homicide data. This website leverages a dataset spanning the years 2012 to 2017. Some of the highlight and key features and the technologies employed are described. The website's structure adheres to a standard navigation bar while emphasizing two primary features:

The Arrest Prediction Page: This page utilizes data from the years 2012 to 2017 to offer predictive insights regarding arrests related to homicide cases.

Dataset Overview Page: We have dedicated a page to provide an extensive overview of our dataset, incorporating the Ydata library. This dataset overview was originally created within Jupyter Notebooks and subsequently exported as an HTML file.

A dedicated 'Resources' page consolidates various materials that have contributed to the development of the website.

The website itself was constructed using Python and Flask. Its user interface is designed around a straightforward form that prompts users to input data, with clear formatting instructions provided at the top of the form.

The visualization of data is presented on two separate web pages, both developed using Tableau. These Tableau pages, originally designed using Tableau Public, have been integrated into the website through the Tableau API. Providing faster response time than embedding HTML code into the page.

Deployment of the website was accomplished through PythonAnywhere , mirroring the same website structure established during local development. While the deployment process encountered minimal technical challenges, one notable issue arose due to the absence of a required library for the Light GBM (LGBM) machine learning model on the PythonAnywhere platform. As a result, a transition was made to an AdaBoost (ADA) model to ensure functionality.

The website, constructed with Flask, encompasses a total of ten pages. The main page, designed to capture immediate attention, features a dynamic background—a looping video clip. This choice serves to create a visually engaging and thought-provoking backdrop, underscoring the sensitivity of the topic of homicide.

The website's color palette was intentionally selected from Bootswatch's 'Darkly' theme. This choice was made to maintain a tone of seriousness, respecting the gravity of the subject matter—homicide data representation. It underscores our commitment to ensuring that the significance of this data, which represents victims of homicide, is never overshadowed by colors that do not align with the gravity of the topic."

====================ANN=============================

