In [2]:
```
# libraries
import numpy as np
import pandas as pd
import altair as alt
from sklearn.decomposition import PCA
alt.data_transformers.disable_max_rows()
```

Out[2]:
```
DataTransformerRegistry.enable('default')
```

PSTAT 100 Project plan report

This is a guide to preparing your project plan. It functions both as a guide to the work you'll need to do and as a guide to preparing the deliverable. You can use it as a template to draft the plan report; if so, please remove the text explanations of each section.

While you may find it useful initially to follow the outline given, you do not need to adhere to it exactly -- you're free to organize your submission in the way that seems most natural to you. However, please do keep the high-level sections, so that your report includes the following headers:

1. Background
2. Data description
3. Initial exporations
4. Planned work

Your report does not need to be long. It should be about 2-4 pages, and may not be much longer than this template once you replace the guiding text with your own work.

Group information

Group members: Alyssa Keeha , Jasmine Kwok, Jordan Tran

Contributions:

1. Alyssa wrote up the background and explored the relationship between State Frequencies and Police Budgets.
2. Jasmine wrote up the initial explorations part and provided the variable summaries.
3. Jordan wrote up the Data Descriptions and figured out a way to geocode the coordinates in the dataset and output and map of all the locations of Police Killings in the United States.

---

0. Background

This section should introduce your reader to the general topic you're engaging with in your project and explain any specialized knowledge that they may need to understand your dataset and why it's interesting. It doesn't need to be long, but should touch on the following points:

- Introduce the topic of your project.
- What area or areas of study are you in dialogue with for your project?
- What is your data about, broadly?
- What is the motivation for collecting the kind of data you're working with, and what sorts of things could you potentially learn?

You can look to the background sections in the homework assignments for examples. (There you can also see how to include images in your notebook.) The background sections of the homeworks are usually short and focused paragraphs intended to orient you to what you'll do in the assignment. They don't go into a lot of detail -- just enough to (hopefully) convince you that the data are interesting and explain any terminology or general information you may not know.

You may find it useful to write up the data description first, think about what the reader should know before they peek at your dataset, and then come back to the background section. I often write the background sections of your assignments last, once I have a sense of what kind of information would be most useful going into the assignment.

For our topic, we would like to perform an in-depth analysis on the factors involving the thousands of fatal police shootings in recent years. In light of the recent events circling the wrongful police killings in the past year, we wanted to create a meaningful and informative project that is relevant in our society right now.

The most obvious area of study we are engaging in with our project are civic and political issues. Civics is the study of people and their rights as citizens and politics is our country's way of making decisions for the people. The police are essentially supposed to be enforcing rules while protecting the citizens. Our goal is to find systematic similarities between each of these incidents and bring those findings to light. In addition, we are thinking of also working in geography into our analysis by geocoding the coordinates onto a map using Altair's geoshape feature.

Our data is an up to date log of Police Killings in the United States from the Washington Post for the past five years. It contains general notes about the event and even information about the victim, police and station. The information comes from several different news sources, social media posts and police reports. The data started being logged after the incident in 2014 where Michael Brown, an unarmed black man was killed by the police. A post investigation of this incident showed that the FBI severely undercounted the number of police-caused fatalities and the reason being many departments failed to require reports of this kind.

We want to look at this data for the same reason we explained before in that it is a relevant and important topic in our society today. The frequent and lawful police killings in just the last year have contributed to some of the hardest months of the pandemic. The connection between police killings and race has become an even bigger issue and has caused mass protesting and looting all over the country in the middle of a global pandemic. Some things we hope to learn with this research is the specific variables that contribute to fatalities and potentially ways to prevent them from happening. We want our project to be an educational and interesting read that may be helpful for further research in updating future police funding and practices.

---

1. Data description

This section should introduce your dataset in detail. It should reflect your having gone through the collect/acquaint/tidy stages of the lifecycle. Below I've provided you with an outline. You do not need to adhere to this strictly -- in fact, it would be more natural to divide the items among a few short paragraphs -- but you should touch on each item in a format that suits your project.

Basic information

Help your reader understand what your data is, where it came from, and how it can be used. Provide the following.

General description: provide a one- or two-sentence description of the data right at the beginning. For instance, "The data are diatom counts sampled from evenly-spaced depths in a sediment core from the gulf of California." Nothing too complicated, just something to give your reader a sense of the 'what' right off the bat.

Source: indicate where your data came from. Provide a verbal description -- who collected it as part of what project and where -- and either a citation or a hyperlink.

Collection methods: How were the data values obtained? Provide a simple description of how measurements were taken (using scientific equipment? web scraping? surveys?).

Sampling design and scope of inference: Indicate the relevant population. If identifiable from data documentation, state the sampling frame and sampling mechanism and indicate the scope of inference. If no information is available about the sampling design, indicate this instead, and discuss the extent to which having no scope of inference is a limitation for the particular topic you're investigating.

In recent years, negligent fatal shootings by police officers have been shoved to the public spotlight after various notorious incidents. However, officials of the FBI and the Centers for Disease Control and Prevention have confirmed that their data log for fatal shootings by police is incomplete, resulting in inaccurate analysis of our potential policing issue. Therefore, at the start of 2015, The Washington Post started a project to begin recording all future fatal shootings by on-duty police officers in the United States and various descriptors of the incident. Their database is located here: https://github.com/washingtonpost/data-police-shootings.

The Washington Post accumulates their data by culling local news reports, law enforcement websites, social media, and from our databases such as Killed by Police and Fatal Encounters. The Post filed open-records requests with police departments in order to collect additional information for each fatal shooting and also requests comment per incident. In 2015, the FBI and the Centers for Disease Control and Prevention documented less than half of the amount of fatal shootings by police than the Washington Post. The Washington Post's database is kept up-to-date as fatal shootings and facts emerge from the Post's collective methods.

Our first dataset we are working with is Washington Post's data of police fatal shooting incidents from Jan. 1, 2015. The population of our data is all fatal shootings in the United States by a police officer in the line of duty. The population does not include deaths of people in custody, fatal shootings by off-duty officers, or non-shooting deaths. The sampling frame consists of all entities in the population that were publicly reported in any form of way since Jan. 1, 2015. The sampling mechanism is a census and therefore our sample consists of all entities in our sampling frame. The sample consists of 6280 observations of fatal police shootings from Jan 2, 2015 to May 9, 2021. Our data is administrative and has no scope of inference.

Data semantics and structure

Units and observations: State the observational units.

Variable descriptions: Provide a table of variable descriptions. If your dataset is large and you'll only work with a subset of the total available variables, limit your attention to the variables that you'll work with. Here's a template you can work with:

Name	Variable description	Type	Units of measurement
id	a unique identifier for each victim	Numeric	Calendar year

Example rows: Print a few example rows of your dataset in tidy format. Please don't include the codes you used to manipulate the raw data. Do that in a separate notebook and export the result to a .csv file -- `data.to_csv('tidy-data.csv')` -- to load directly into the cell below.

The observational units are incidents of a fatal shooting in the United States by a police officer in the line of duty.

Name	Variable description	Type	Units of measurement
id	unique identifier for each victim	Nominal	None
name	name of the victim	Nominal	None
date	date of the fatal shooting	Nominal	YYYY-MM-DD
manner_of_death	manner of victim's death	Nominal	None
armed	what the victim was armed with	Nominal	None

Name	Variable description	Type	Units of measurement
age	age of the victim	Numeric	Years
gender	gender of the victim	Nominal	None
race	race of the victim	Nominal	None
city	city where the fatal shooting took place	Nominal	None
state	state where the fatal shooting took place	Nominal	None
signs_of_mental_illness	indicator if the victim had a history of mental health issues, expressed suicidal intentions or was experiencingmental distress at the time of the shooting	Nominal	None
threat_level	threat level of victim to the police	Nominal	None
flee	indicator if the victim was moving away from officers	Nominal	None
body_camera	indicator if an officer had a body camera on at the incident	Nominal	None
longitude	longitude location of the shooting expressed as WGS84 coordinates, geocoded from addresses	Numeric	Degrees
latitude	latitude location of the shooting expressed as WGS84 coordinates, geocoded from addresses	Numeric	Degrees
is_geocoding_exact	indicator of the accuracy of the coordinates	Nominal	None
Budget_Per_Capita	police budget per capita of the state where incident occured	Numeric	Dollars

In [34]:
```
# load tidied data and print rows
police_killings = pd.read_csv('complete_police.csv')
police_killings.head(5)
```

Out[34]:

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_mental_illness	threat_level	flee	body_can
0	3	Tim Elliot	2015-01-02	shot	gun	53.0	M	A	Shelton	WA	True	attack	Not fleeing	F
1	4	Lewis Lee Lembke	2015-01-02	shot	gun	47.0	M	W	Aloha	OR	False	attack	Not fleeing	F
2	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23.0	M	H	Wichita	KS	False	other	Not fleeing	F
3	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32.0	M	W	San Francisco	CA	True	attack	Not fleeing	F
4	9	Michael Rodriguez	2015-01-04	shot	nail gun	39.0	M	H	Evans	CO	False	attack	Not fleeing	F

2. Initial explorations

At this stage, you may spend most of your effort on the computing side tidying up the data. You're not expected to complete a thorough exploratory analysis, and if your dataset was especially messy to start with, you may not even begin your exploratory analysis by the time you prepare this report. You have the option to leave exploration for the next stage of work and simply report basic properties of the dataset, but you should at minimum address the items in the 'basic properties' section below.

Basic properties of the dataset

Help the reader get acquainted with your dataset on a simple level by identifying characteristics of the dataset and variable summaries. Some amount of code is fine here, but try to use code cells sparingly.

Dimensions: state the dimensions of the data (in tidy format, of course).

Missing values: Are there missing values? If so, why are they missing?

Variable summaries: Provide simple variable summaries for the most important variables in your dataset. Preferably, you'll do this for all variables, but if you have a large number, you might need to prioritize and focus on the ones most of interest. What exactly you do is a little case-specific, but think of things like means and variances, min/max, number of levels and observation counts for categorical variables, etc.

There are 6280 observations and 18 columns. This dataset consists of one integer variable (id), four float variables (age, longitude, latitude, budget_per_capita), three boolean variables (signs_of_mental_illness,body_camera,is_geocoding_exact) and the rest of the variables have an object data type which indicates that it is a string.

There are a total of 2325 missing values in the dataset. There are 233 missing names, 208 missing armed indications, 281 missing age, 1 missing gender, 583 missing race, 388 missing indications on how the victims were moving away from officers, 307 missing longitude,307 missing latitude and 17 missing budget_per_capita.

The details of each police killing tracked by the Washington Post were gathered from law enforcement websites, local news reports, social media, and through monitoring independent databases. It is possible that the missing values on name, age, and race were due to unreported deceased details we missing by chance on external websites or intentionally excluded to protect the confidentiality of the victims and their

families. The missing values for Budget_Per_Capita were all for Washington, DC. The Budget_Per_Capita variable was merged from another dataset and the information on Budget_Per_Capita DC was already missing in the other dataset. There was 1 missing gender for this dataset which was the row with id 2956. Gender is classified in a binary form of male and female for this dataset, however, in reality gender is not limited to these two categories and can be more fluid. Searching on the internet, the gender of the deceased with id 2956 does not fit into the binary category which might be a reason for it to be missing.

In [40]:
```
# dimension of the data set
police_killings.shape #(6280, 18)
# variable data types
police_killings.dtypes
# total number of missing values in the dataset
police_killings.isna().sum().sum()
# missing values in each variable
police_killings.isna().sum()
```

Out[40]:
```
id                          0
name                      233
date                        0
manner_of_death             0
armed                     208
age                       281
gender                      1
race                      583
city                        0
state                       0
signs_of_mental_illness     0
threat_level                0
flee                      388
body_camera                 0
longitude                 307
latitude                  307
is_geocoding_exact          0
Budget_Per_Capita          17
dtype: int64
```

basic statistical details of the data set

	age	longitude	latitude	Budget_Per_Capita
count	5999.000000	5973.000000	5973.000000	6263.000000
mean	37.104017	-97.210832	36.654752	343.246527
std	13.010878	16.635403	5.387437	87.101575
min	6.000000	-158.137000	19.498000	186.000000
25%	27.000000	-112.117000	33.470000	277.000000
50%	35.000000	-94.371000	36.093000	327.000000
75%	45.500000	-83.089000	39.995000	406.000000
max	91.000000	-67.867000	71.301000	530.000000

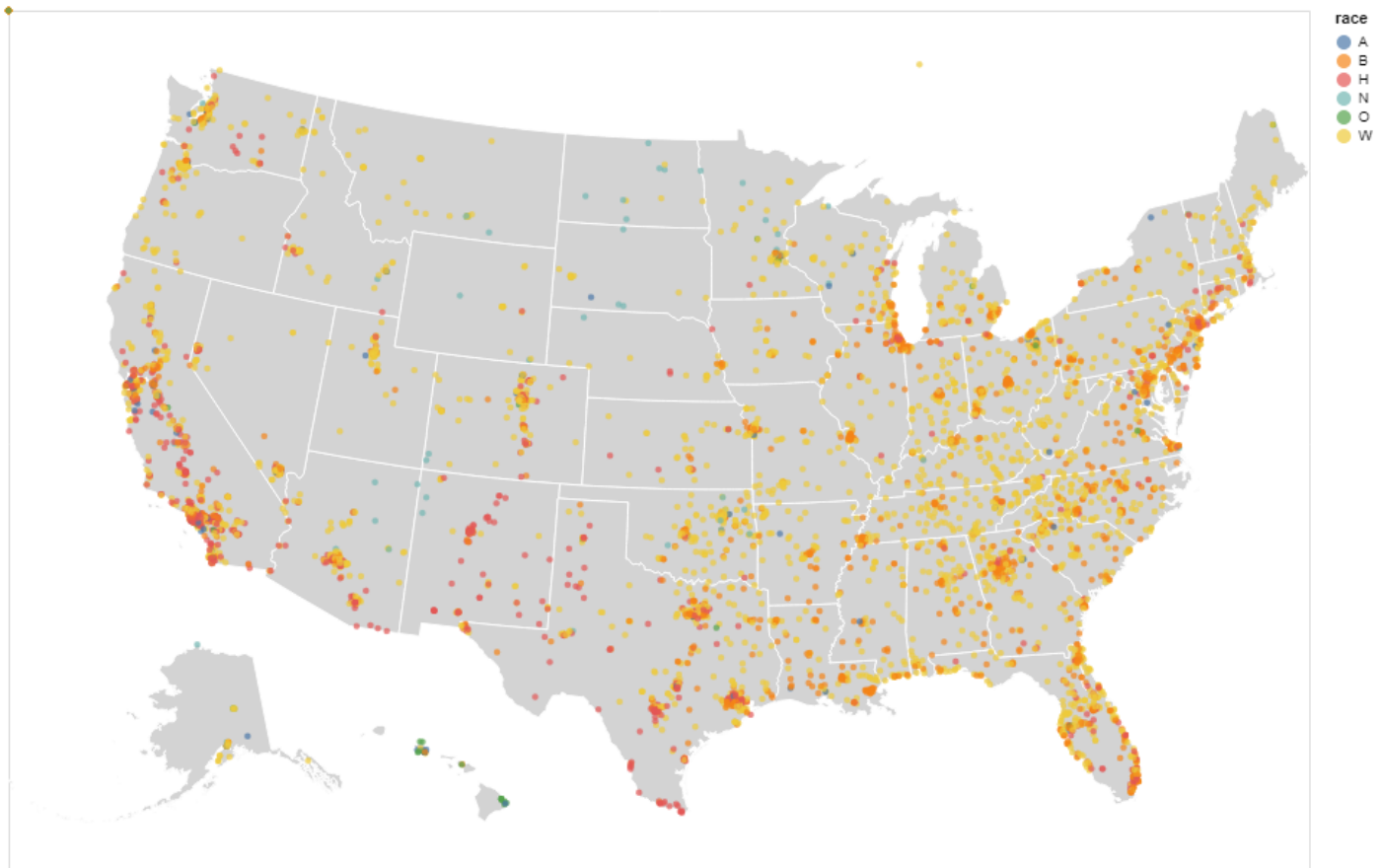
Exploratory analysis

If you were lucky and your dataset was neat, you should aim to include a few exploratory plots or tables here -- they don't need to be polished at this stage, but you should select plots that are informative (rather than including all plots you may have looked at).

If you do include exploratory graphics or tables, please explain in a sentence or two what each one shows. Try to include a minimum of code. Consider saving your plots as images and inputting images into markdown cells instead of generating them anew via code cells.

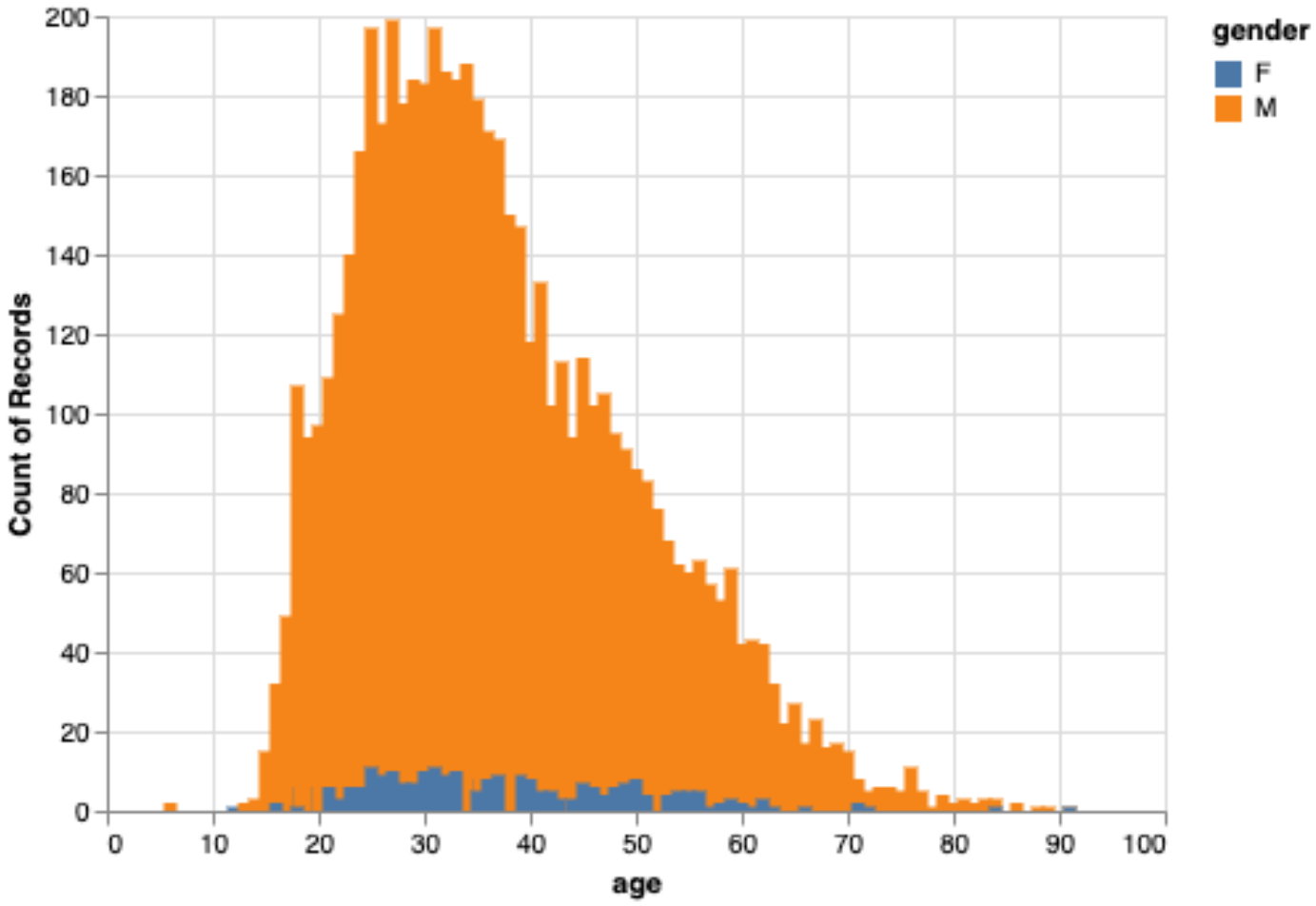
Map

The following map was generated by importing the vega dataset and mark_geoshape() to make the map and mark_circle() to plot the corrdinates.



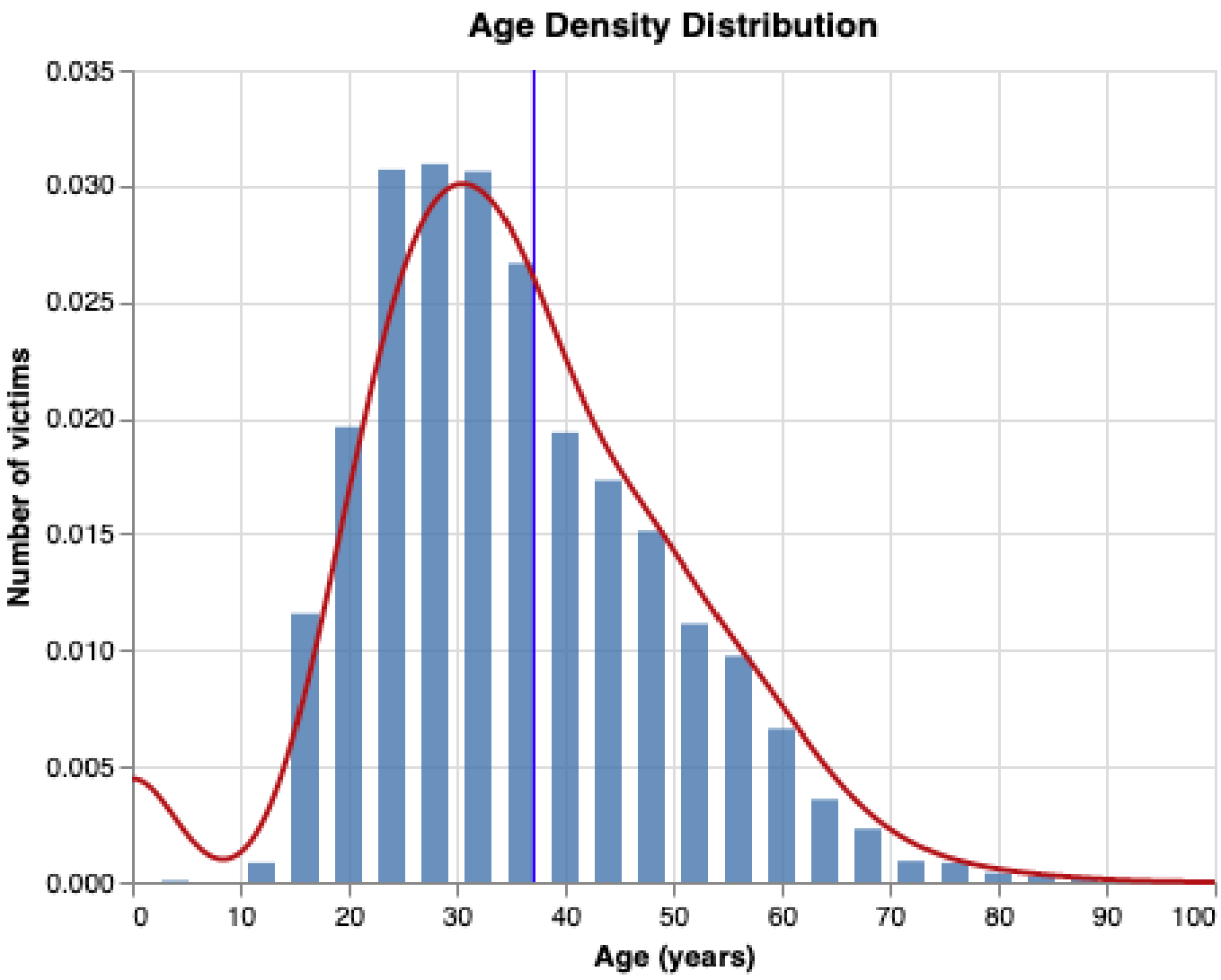
It displays the locations of every single reported fatal Police Shooting from 2015 and depicts the race of the victim by color.

Below is a histogram of the Ages of Fatal Police Killings separated by gender with color.



We see a large range for women, but less frequencies, while for men a strong mean around an age between 25-35 with high frequencies.

Below is another histogram plotting the density of the ages of each police fatal shooting.



---

3. Planned work

Here you should indicate your tentative ideas for your analysis. Don't worry, these aren't final -- you can always change your mind later or shift gears if they don't pan out. The objective is to have you start thinking ahead about what you'll do.

Questions

Please propose three focused questions that you plan to explore.

1. *Is there any indication of Mental health issues of victims having a relationship with the threat level felt by the police officer?*
2. *Do economic factors, such as a city's poverty rate, have an association with the number of fatal shootings?*
3. *How does the racial distribution of a particular area relate to the possibility of a fatal shooting occur?*

Proposed approaches

For each question, please describe an idea or two about how you might approach the question.

1. *Create a bar chart and facet by the factors for the mental_health_issues variable to depict the counts for each threat level.*
2. *Finding economic data on cities and merging with our data. We can then create histograms to show the distribution of various economic factors across the cities where a fatal shootings occured.*
3. *Identify areas with exceptionally high or low number of police shootings through the heat map and select several cities or state to look into more depth. Then using this information, subset the selected cities or state and plot individual scatterplots by race.*

Submission Checklist

1. Save file to confirm all changes are on disk
2. Run *Kernel > Restart & Run All* to execute all code from top to bottom
3. Save file again to write any new output to disk
4. Select *File > Download as > HTML*.
5. Open in Google Chrome and print to PDF on A3 paper in portrait orientation.
6. Submit to Gradescope