DMDW Project

Racial Disparities in Fatal Police Shootings

Rushabh Maru -201070008
Priyansh Agarwal - 201070018
Pranav Kute - 201070020
Aayush Mehta - 201070030

Purpose of the Project

The main goal of this project is to understand more about people who were killed by the police in the United States during the year 2015. We want to figure out how and why these incidents happened. The United States has higher rates of police-involved deaths compared to similar Western countries, and we want to explore if there's a specific focus on factors like race and ethnicity in these cases.

Abstract

Unjust police shootings have sparked national outrage in the United States and have started international social movements, such as Black Lives Matter that protest against incidents of racially motivated police violence. The media spotlight has been shined on numerous infamous cases of potentially prejudiced police shooting victims, like the shooting of Breonna Taylor. Our project aims to explore whether police shootings are racially biased and if certain racial groups are targeted more. We also aim to look specifically at California which is the State that has the most police shootings and to see if there are any racial disparities in its victims. We find that there is a difference between the racial distribution of police shooting victims and the racial distribution of the population and that African Americans are overly represented in the population of police shooting victims. We determine that our county demographic data on education and income have no correlated variables with black shooting victim population and that California reflects our previous findings and also that Hispanics are overly represented in the State's population of police shooting victims as well

Introduction

The frequent and lawful police killings in just the last year have contributed to some of the hardest months of the pandemic. With the media coverage of the Black Lives Matter Movements, it seems that so many Police Shooting Victims are African American. However, we noticed that in total there are more White victims than Black in police killings. The connection between police killings and race, especially, has become an even larger issue and has led to mass protesting and looting all over the country. In light of these events circling the wrongful police killings, we intend to create an educational, meaningful, and informative project that is relevant to the current society and helpful for further research in updating future police funding and practices.

Our goal is to perform an in-depth analysis on factors, like race, that were accounted for in thousands of fatal police shootings in recent years. With this research, we hope to learn the specific variables that contribute to racially biased fatalities and explore potential solutions to prevent them from occurring.

Why this Project Matters:

Understanding why police-involved deaths happen is crucial for improving law enforcement practices and ensuring the safety of communities.

By exploring the role of race and ethnicity, you aim to shed light on potential disparities and work towards a fair and just system.

Datasets:

https://www.kaggle.com/datasets/kwullum/fatal-police-shootings-in-the-us

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 5 datafiles are:

1. MedianHouseholdIncome2015.csv
2. PercentOver25CompletedHighSchool.csv
3. PercentagePeopleBelowPovertyLevel.csv
4. PoliceKillingsUS.csv
5. ShareRaceByCity.csv

https://github.com/washingtonpost/data-police-shootings

Here's an example of the main dataset:

id	name	date	manner_c	armed	age	gender	race	city	state	signs_of_r	threat_lev	flee	body_cam
3	Tim Elliot	########	shot	gun	53	M	A	Shelton	WA	TRUE	attack	Not fleein	FALSE
4	Lewis Lee	########	shot	gun	47	M	W	Aloha	OR	FALSE	attack	Not fleein	FALSE
5	John Paul	########	shot and T	unarmed	23	M	H	Wichita	KS	FALSE	other	Not fleein	FALSE
8	Matthew I	########	shot	toy weapc	32	M	W	San Franci	CA	TRUE	attack	Not fleein	FALSE
9	Michael R	########	shot	nail gun	39	M	H	Evans	CO	FALSE	attack	Not fleein	FALSE
11	Kenneth J	########	shot	gun	18	M	W	Guthrie	OK	FALSE	attack	Not fleein	FALSE
13	Kenneth A	########	shot	gun	22	M	H	Chandler	AZ	FALSE	attack	Car	FALSE
15	Brock Nich	########	shot	gun	35	M	W	Assaria	KS	FALSE	attack	Not fleein	FALSE
16	Autumn S	########	shot	unarmed	34	F	W	Burlington	IA	FALSE	other	Not fleein	TRUE
17	Leslie Sap	########	shot	toy weapc	47	M	B	Knoxville	PA	FALSE	attack	Not fleein	FALSE
19	Patrick We	########	shot and T	knife	25	M	W	Stockton	CA	FALSE	attack	Not fleein	FALSE
21	Ron Sneec	########	shot	gun	31	M	B	Freeport	TX	FALSE	attack	Not fleein	FALSE
22	Hashim Ha	########	shot	knife	41	M	B	Columbus	OH	TRUE	other	Not fleein	FALSE
25	Nicholas R	########	shot	gun	30	M	W	Des Moine	IA	FALSE	attack	Car	FALSE
27	Omarr Juli	########	shot	gun	37	M	B	New Orlea	LA	FALSE	attack	Foot	TRUE
29	Loren Sim	########	shot	vehicle	28	M	W	Huntley	MT	FALSE	undeterm	Not fleein	FALSE
32	James Duc	########	shot	shovel	42	M	W	Salt Lake C	UT	FALSE	attack	Not fleein	TRUE
36	Artago Da	########	shot	unarmed	36	M	B	Strong	AR	FALSE	attack	Not fleein	FALSE

Implementation:

We have implemented a jupyter notebook, which performs data cleaning and exploratory data analysis. Along with that we have used various models like Random Forest, Logistic Regression and K nearest neighbours which predicts the race depending on the other data in the dataset.

Next we built a website and a tableau dashboard which helps stakeholders to understand the real picture and make the easily understand the concept we are trying to explain.

Notebook:

Steps performed:

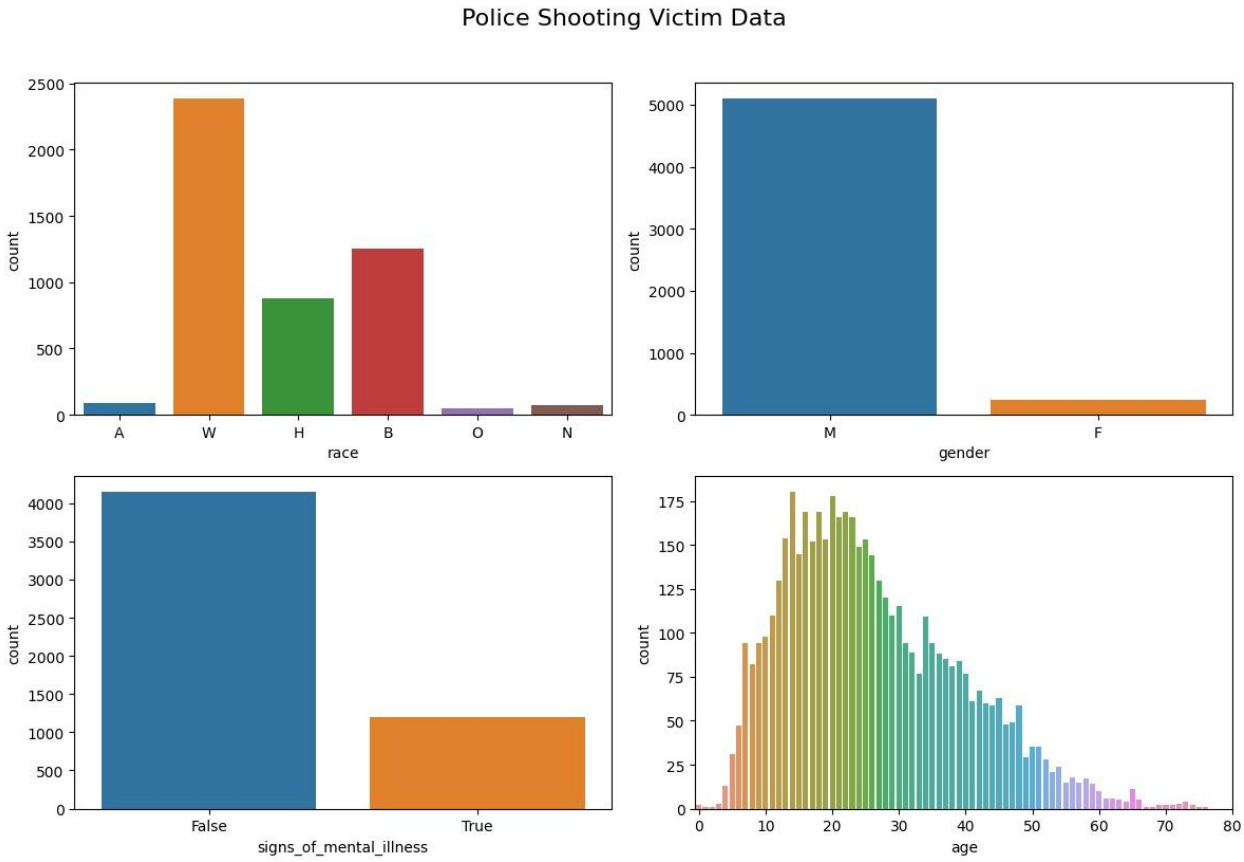
1. Data understanding
2. Data cleaning
3. Exploratory data analysis

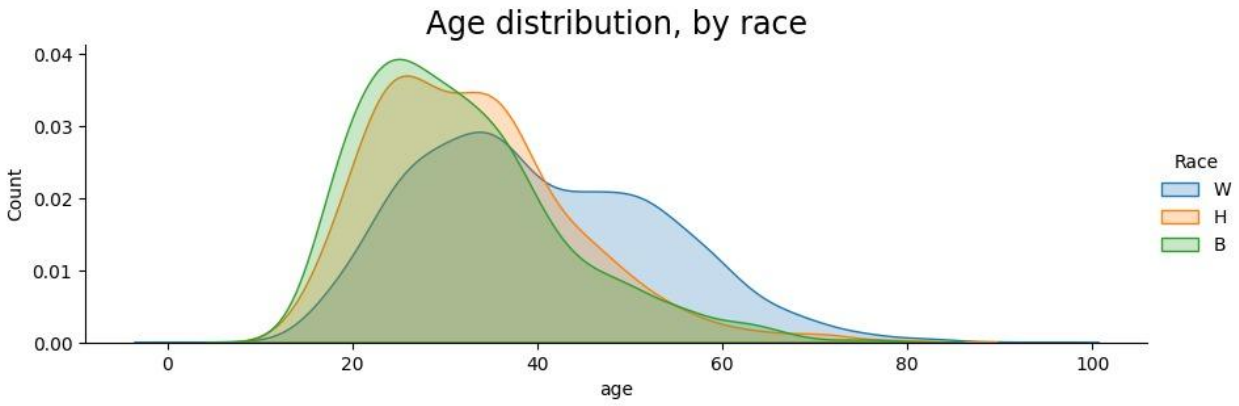
4. Applying the Random Forest model
5. Applying the Logistic Regression model
6. Applying the KNN model
7. Conclusion

Exploratory data analysis and data cleaning:

We understood the dataset and then cleaned the data set. For instance, we converted string type categorical values to numeric types so that models perform better.

The we used various graphs to perform exploratory data analysis:





Models used:

1. Random Forest:

   Random Forest is an ensemble machine learning algorithm that is widely used for both classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.

   Random Forest is known for its high accuracy, ease of use, and ability to handle complex relationships in data. It is particularly valuable when dealing with large datasets and offers a robust solution for predictive modeling tasks.

2. Logistic Regression

   Logistic Regression is a statistical model used for binary classification tasks, predicting the probability that an instance belongs to a particular class. Despite its name, it is primarily employed for classification rather than regression.

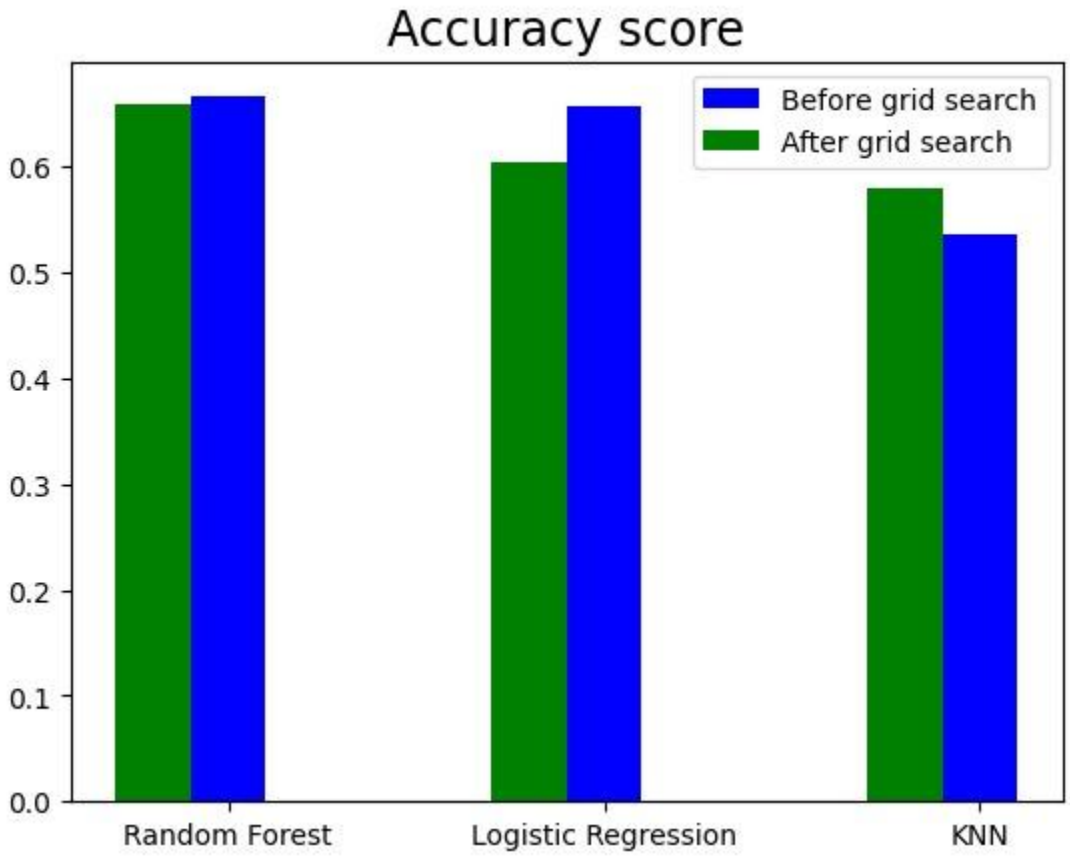
   Logistic Regression is widely used in various fields, including medical diagnosis, finance, and social sciences. It is especially suitable when the relationship between the input features and the binary outcome is assumed to be approximately linear.

3. KNN:

   K-Nearest Neighbors is a simple and intuitive machine learning algorithm used for both classification and regression tasks. It makes predictions based on the majority class (for classification) or the average of the nearest neighbors' values (for regression).

   K-Nearest Neighbors is used in various applications such as pattern recognition, image processing, and recommendation systems. It is particularly effective when the decision boundary is nonlinear, and the data is not well-separated.

Here, we have tried to find the best method for k in the knn algorithm and then we have compared these different models:



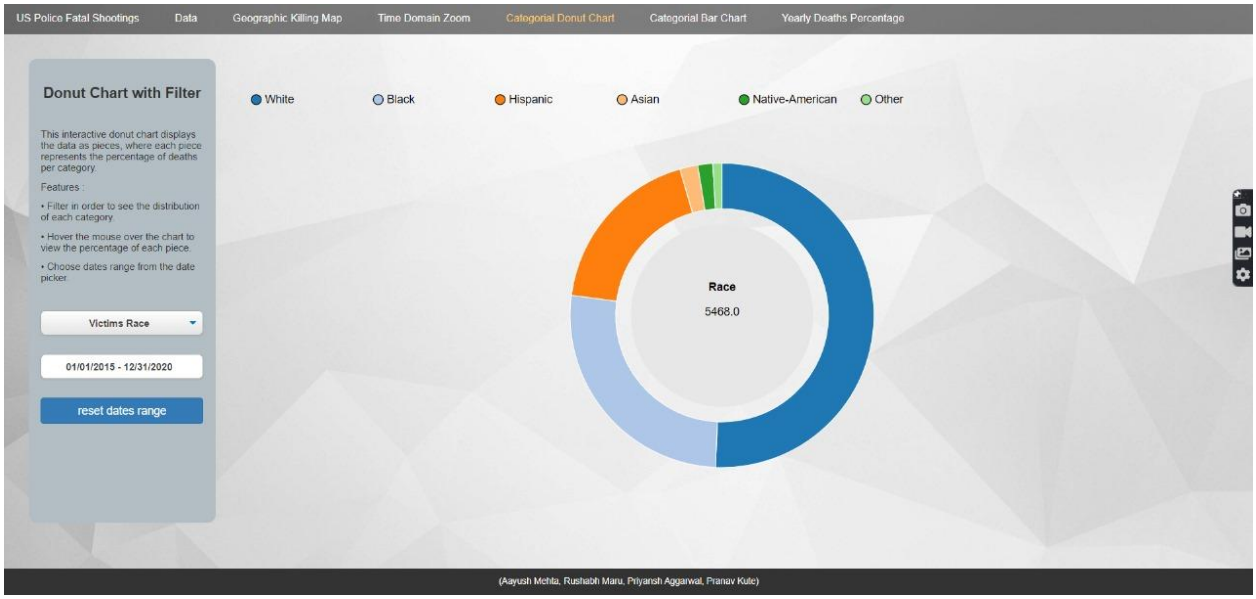
Website:

We then created a simple website using html, css and javascript to display the findings and create simple charts.

These include:

1. Time series chart
2. Categorical donut chart
3. Gender based categorical bar chart
4. USA map chart


US Police Fatal Shootings
Data
Geographic Killing Map
Time Domain Zoom
Categorial Donut Chart
Categorial Bar Chart
Yearly Deaths Percentage
Time-series chart
The D3 chart currently displayed on metacat UI where the data is shown in a zoomable time range.
Features :
• Mini 'brush' chart shows the context of the data in focus in the main chart.
• Y-axis resizes automatically to match the maximum y-value of the data in focus.
• X-axis zooming and panning is limited to the range of available data.
• Text at top changes to indicate the range of data in focus.
• When no data is available, displays text instead.
• Buttons zoom focus to an interval equal to one year (if sufficient data), one month, or the range of the data.
Month
Year
Data
Showing data from : Jan 2015 - Jul 2020
Monthly deaths
(Aayush Mehta, Rushabh Maru, Priyansh Aggarwal, Pranav Kute)


US Police Fatal Shootings
Data
Geographic Killing Map
Time Domain Zoom
Categorial Donut Chart
Categorial Bar Chart
Yearly Deaths Percentage
Donut Chart with Filter
This interactive donut chart displays the data as pieces, where each piece represents the percentage of deaths per category.
Features :
• Filter in order to see the distribution of each category.
• Hover the mouse over the chart to view the percentage of each piece.
• Choose dates range from the date picker.
Victims Race
01/01/2015 - 12/31/2020
reset dates range
White
Black
Hispanic
Asian
Native-American
Other
Race
5468.0
(Aayush Mehta, Rushabh Maru, Priyansh Aggarwal, Pranav Kute)

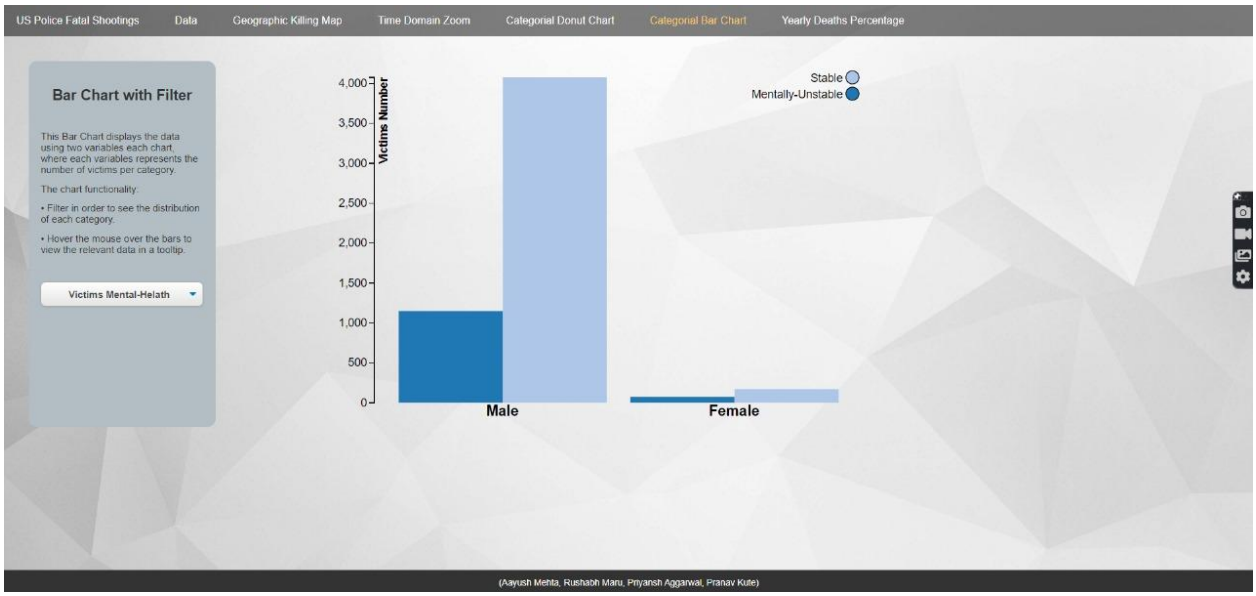
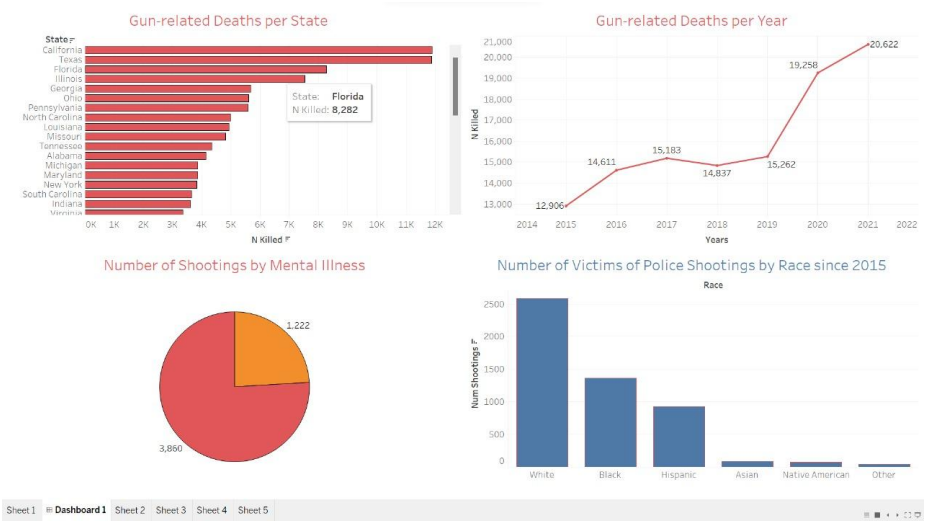

US Police Fatal Shootings
Data
Geographic Killing Map
Time Domain Zoom
Categorial Donut Chart
Categorial Bar Chart
Yearly Deaths Percentage
Bar Chart with Filter
This Bar Chart displays the data using two variables each chart, where each variables represents the number of victims per category.
The chart functionality:
• Filter in order to see the distribution of each category.
• Hover the mouse over the bars to view the relevant data in a tooltip.
Victims Mental-Helath
Victims Number
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Stable
Mentally-Unstable
Male
Female
(Aayush Mehta, Rushabh Maru, Priyansh Aggarwal, Pranav Kute)

Tableau Dashboard:

Tableau is a powerful data visualization tool that offers numerous advantages for creating informative and compelling dashboards. With its user-friendly interface, Tableau enables users to effortlessly connect to various data sources, transform raw data into meaningful insights, and generate interactive visualizations. One of its key strengths lies in its ability to handle large datasets and quickly process complex calculations, ensuring efficient and responsive dashboards. Tableau's drag-and-drop functionality allows users, even those without extensive technical expertise, to effortlessly design visually engaging and dynamic dashboards. Its robust set of visualization options, including charts, maps, and dashboards, facilitates a comprehensive understanding of data patterns and trends. Tableau promotes collaboration by offering easy sharing and embedding options, fostering a culture of data-driven decision-making across teams. Furthermore, its real-time data connectivity and interactive features empower users to explore data on the fly, uncovering insights that might otherwise remain hidden. Overall, Tableau simplifies the process of transforming data into actionable insights, making it an invaluable tool for data-driven decision-makers.



Conclusion:

FINDINGS - SUMMARY

- Blacks are 3 times more likely to become victims of police shootings than Whites.
- The average age of Black and Hispanic victims is lower (31 and 33 respectively) than that of White victims (40).
- California is the state with the most fatal police shootings, and Los Angeles is the most dangerous city.
- The most common way of being armed is by gun.

Now, we explore the US Police Shootings dataset, which has data on fatal shootings by on-duty police officers in the US between January 2015 and May 2020. Using this dataset we will explore several important questions which can be found below. For each question we will attempt to create at least one visualisation which helps to understand the data behind the answers by using the packages seaborn and plotly.

1. Who Are The Police Killing?
2. Are African-Americans Disproportionately Killed?
3. Where Do Shootings Happen?
4. Are Police Shooting Deaths Increasing?
5. Are Victims Armed or Unarmed?
6. Who Are The Unarmed Victims?