Introduction

We received a complete dataset of shootings in the United States by Police Officers and the topic of racism came into play where people thought they were shot or killed because of the race of the victims. In this project, we will be analysing the entire dataset and its columns to determine if race, age, gender, whether the victim was armed or not and how they were killed. This will also show if the poilice officers were strapped with body cameras when the incident happened.

Data Cleaning

This was the first process we went through before perfoming the Exploratory Data Analysis. This process is where we go through the entire dataset to make sure there are no empty rows, columns or cells. We also make sure there are no null values. The dataset must be complete.

Since the entire dataset consisted of four thousand, eight hundred and ninety-five rows, goin through it row by row would be very difficult. We therefore wrote a code to run through the entire dataset to make sure there were no missing values. The function of the code was to drop rows that were empty and check if there were null values in the dataset. Furthermore, the code checks if there are duplicated entries and drops one of them when it is found.

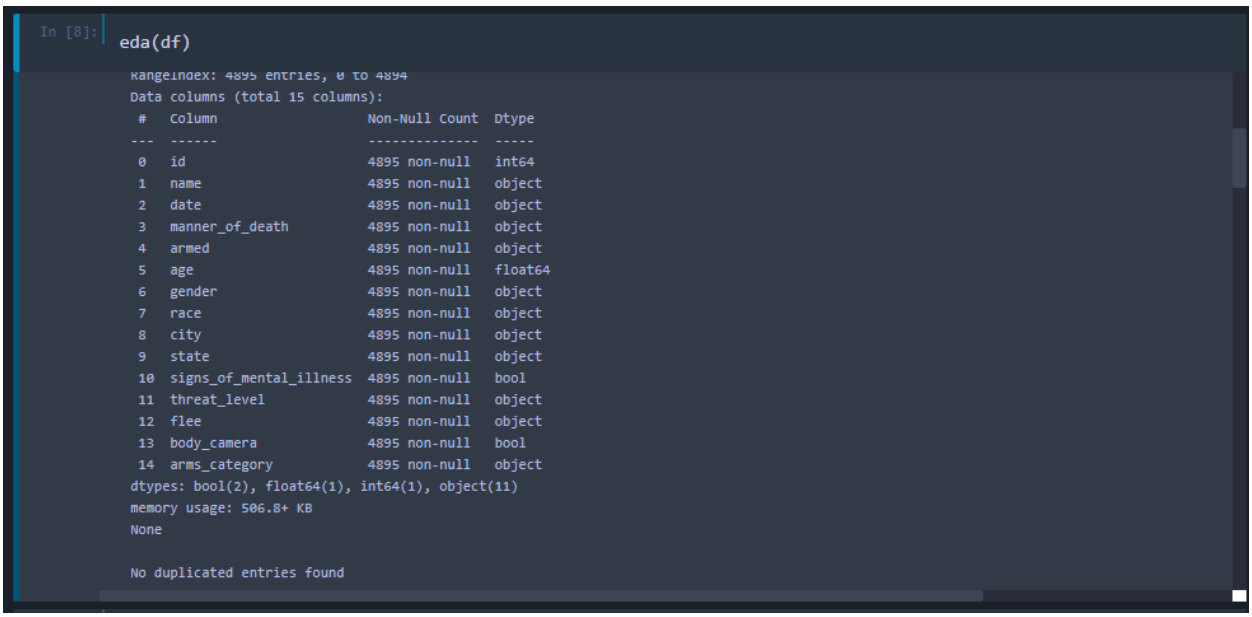


Fig. 1

```
# replace field that's entirely space (or empty) with NaN
df = df.replace(r'^\s*$', np.nan, regex=True)

print("\nTo check: \n (1) Total number of entries \n (2) Column types \n (3) Any null values\n")
print(df.info())

# generate preview of entries with null values
if df.isnull().any(axis=None):
    print("\nPreview of data with null values:")
    display(df[df.isnull().any(axis=1)].head(3))
    missingno.matrix(df)
    plt.show()

# generate count statistics of duplicate entries
if len(df[df.duplicated()]) > 0:
    print("\n***Number of duplicated entries: ", len(df[df.duplicated()]))
    display(df[df.duplicated(keep=False)].sort_values(by=list(df.columns)).head())
else:
    print("\nNo duplicated entries found")

# Drop duplicated entries if true
df.drop_duplicates(inplace=True)
```

Fig. 2

In addition, we changed the values of all the ages to integers and not floating numbers because no one has a floating number as their age.

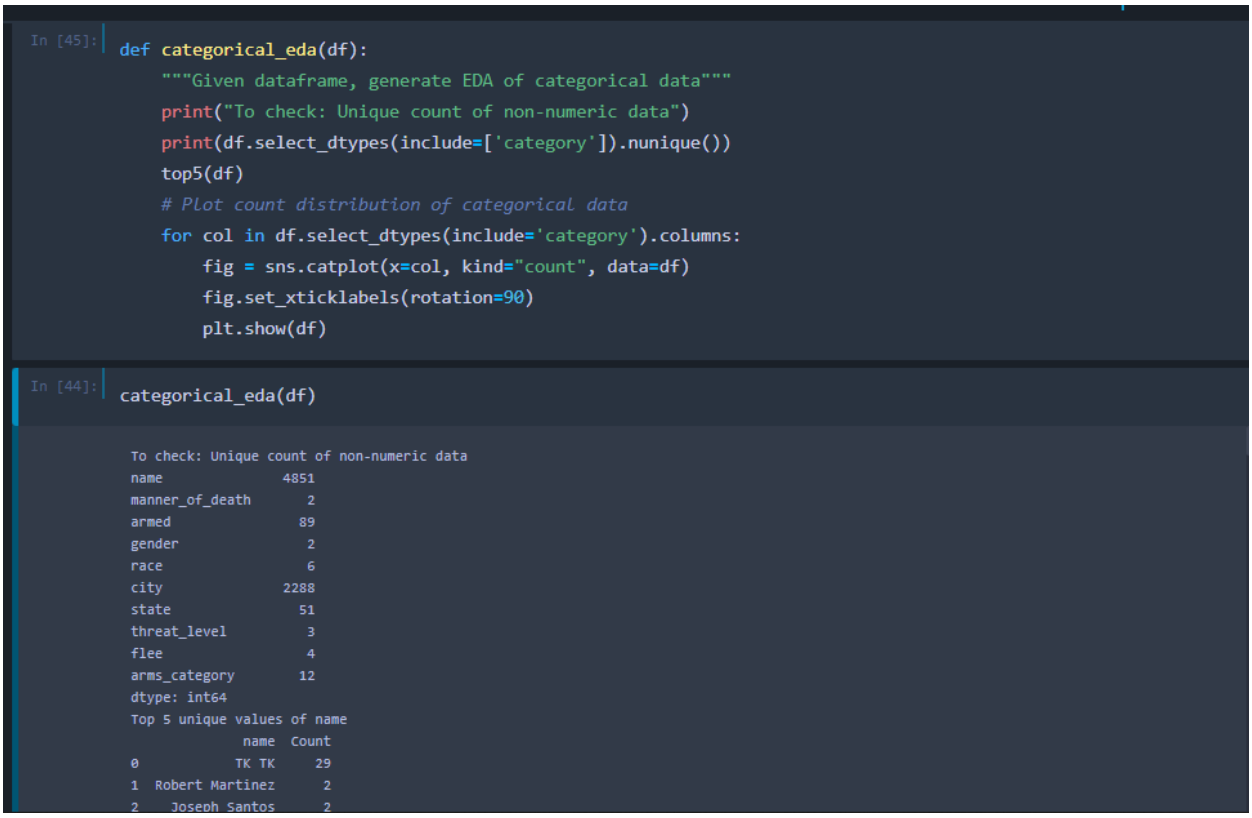
```
In [12]: #change the age value to an integer because no one has a floating age
         df.age = pd.Series(data=df.age, dtype='int')

In [13]:
```

Fig. 3

Exploratory Data Analysis

Exploring the data, we tried to find out the unique value or columns and their corresponding count or frequency of the non-numeric data. This is what is termed as categorical data.



```
In [45]:
def categorical_eda(df):
    """Given dataframe, generate EDA of categorical data"""
    print("To check: Unique count of non-numeric data")
    print(df.select_dtypes(include=['category']).nunique())
    top5(df)
    # Plot count distribution of categorical data
    for col in df.select_dtypes(include='category').columns:
        fig = sns.catplot(x=col, kind="count", data=df)
        fig.set_xticklabels(rotation=90)
        plt.show(df)

In [44]:
categorical_eda(df)

To check: Unique count of non-numeric data
name              4851
manner_of_death      2
armed               89
gender               2
race                 6
city              2288
state               51
threat_level         3
flee                 4
arms_category       12
dtype: int64
Top 5 unique values of name
              name  Count
0           TK TK     29
1  Robert Martinez      2
2    Joseph Santos      2
```

Fig. 4

This code and its output shows the unique non-numerical data from the dataset. The age column from the dataset contains numberical data but the name column from the dataset is made up on non-numerical data. The above figure shows how it is represented.

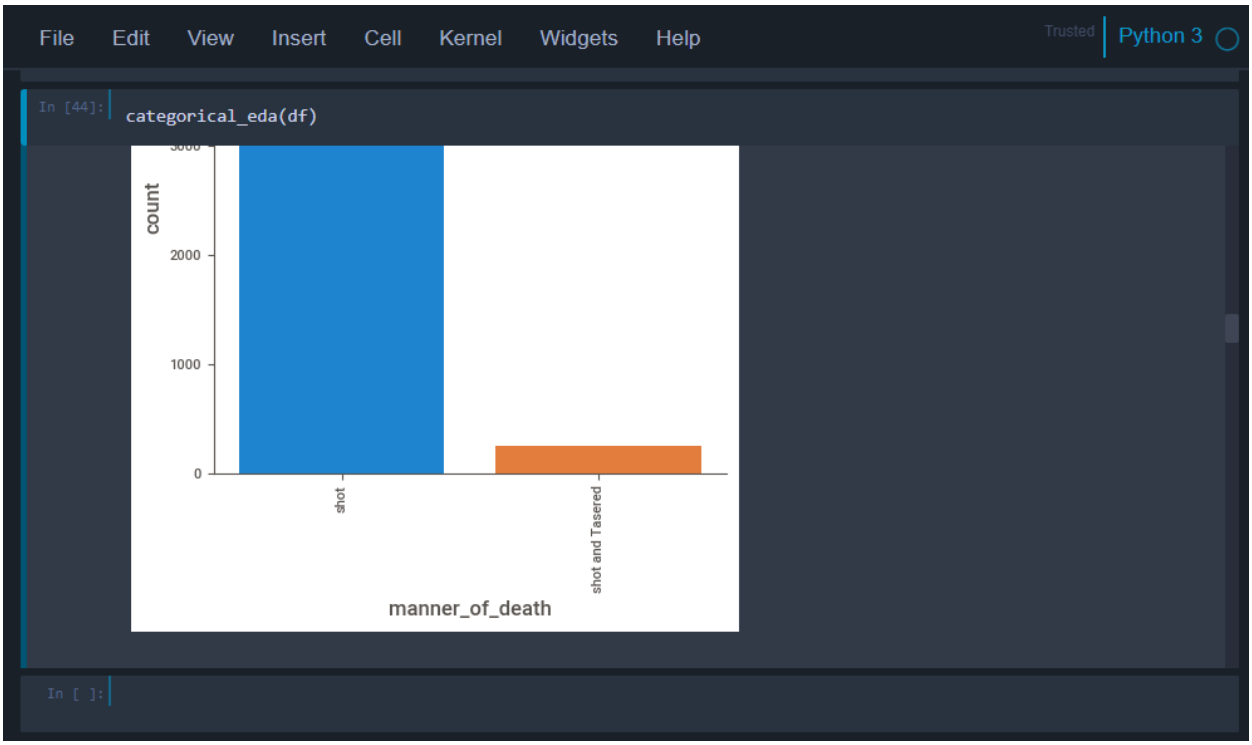


Fig. 5

Fig. 5 displays the manner of death and how the many people were killed in that manner. We can clearly see from the graph that the most of the people were killed with shot more. The other manner in which they died was being shot and tased at the same time.

Fig. 6

In this graph, we can tell from it that Males were targeted the most and killed as compared to females.

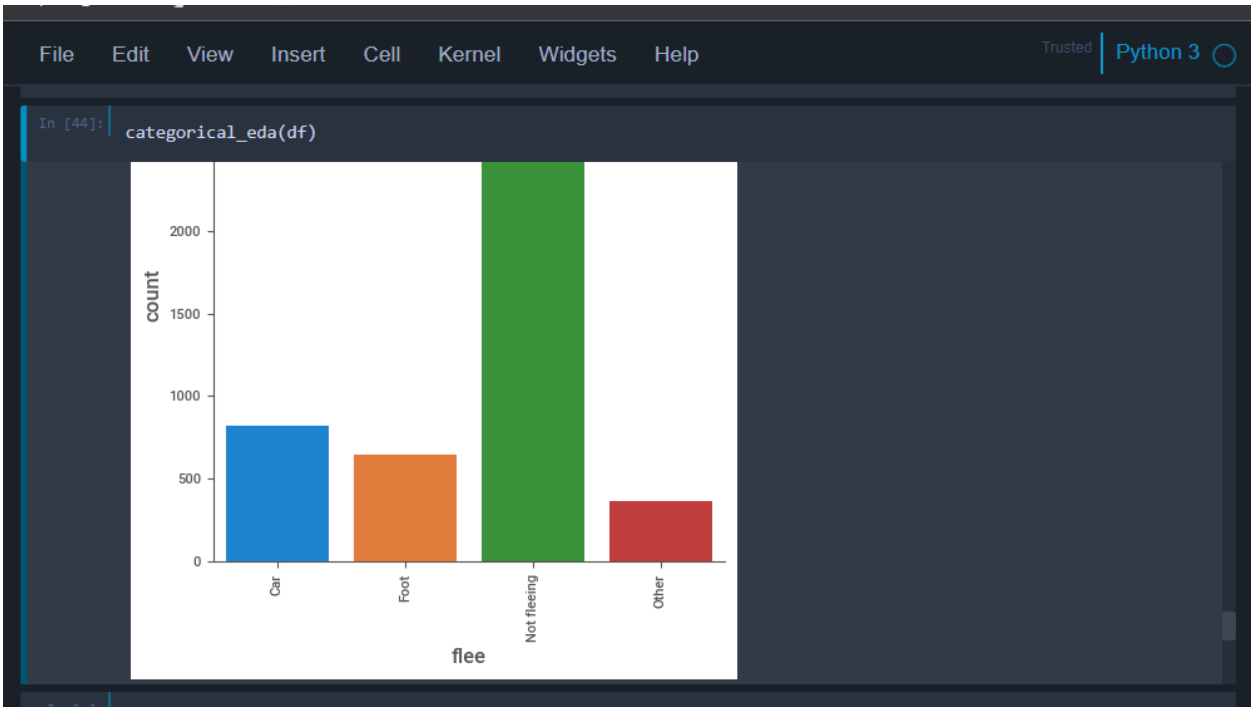


Fig. 7

This figure demonstrates the manner in which the victims were caught and killed. When we take a closer look at the graph, we notice that few of them were fleeing with the aid of something else, some of them were fleeing on foot when they were caught and killed and others were also fleeing in vehicles or cars. The most concerning aspect discovered was the fact that most of them were not fleeing when they were caught and killed.

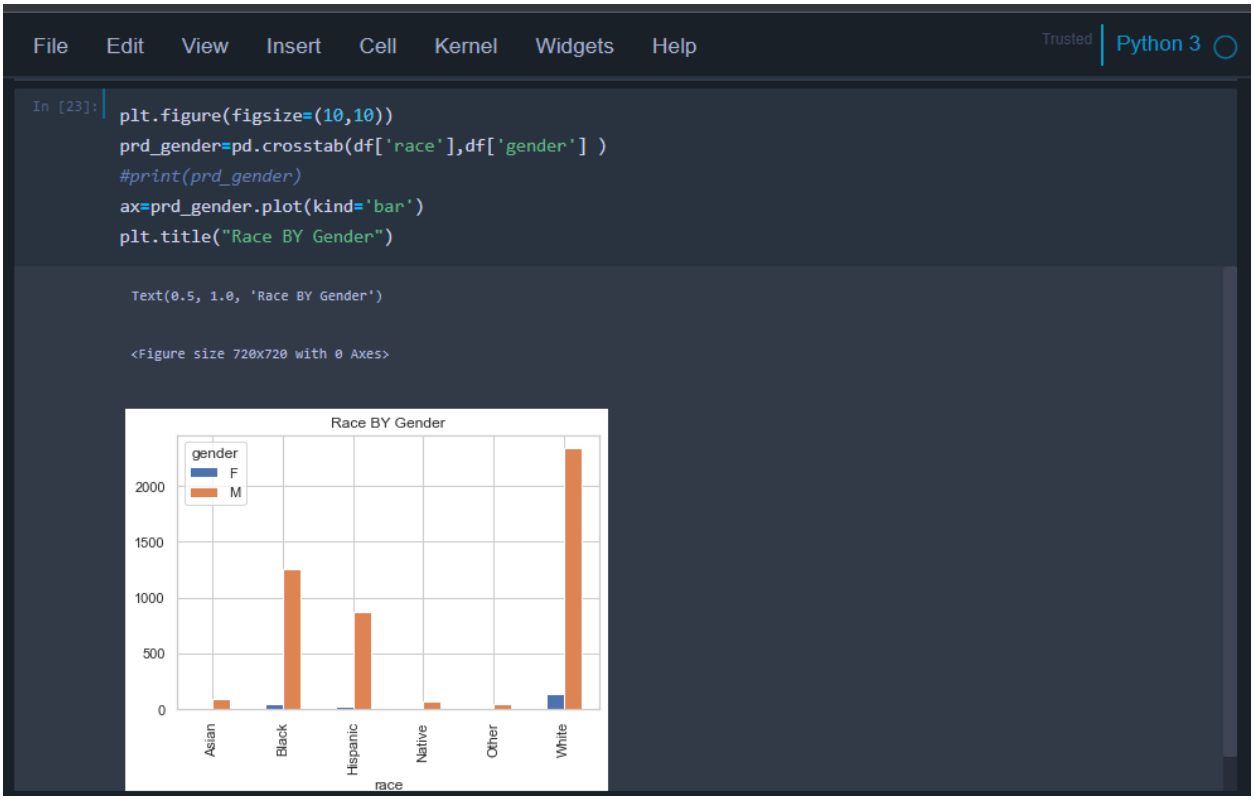


Fig. 8

This graph displays race by gender and the frequency with which they were killed. Most of those killed were males and they were also White. The next highest bar shows that male blacks are the most killed next to White males.