**SDS MINI PROJECT**

**PES UNIVERSITY, ECC**

**DATA ANALYSIS ON POLICE BRUTALITY IN USA**

**(LAST 5 YEARS)**

Mayuravarsha P PES2UG19CS225 D

Mohan R PES2UG19CS240 D

Mettupalle Saketh Kumar Reddy PES2UG19CS231 D

Mooli Raviteja Reddy PES2UG19CS242 D

**ABSTRACT**

With the death of George Floyd, there was a movement all across the globe for equality and against the discrimination of black people. So, we decided to extract a dataset from Kaggle which contains data about police brutality in US for the past 5 years and perform a detailed analysis on it. The black community strongly believe that they are vulnerable to the society and the cops as they are targeted more often than not. So, we carried out a detailed analysis of whether this is a myth or a fact by comparing the data of different races of victims of such similar brutalities. These data elements are studied w.r.t time and other factors like age and gender. Various types of Statistical and Graphical analysis are being carried out on the dataset like Hypothesis testing, Normalization & Standardization and Correlation to come up with a meaningful insight.

**INTRODUCTION**

We carried out a detailed analysis of whether the black community is targeted more often or if it's just and myth and we also analysed how the number of such cases are changing with time. We were able to do this by comparing the data of different races of victims of such similar brutalities. The extracted data had certain elements which were missing or ‘unknown’ or outliers which had to be removed or sorted out using various web scraping methods where a complete row is deleted or the empty blocks are assigned a value based on the other values in that column.

From the dataset, we wanted to infer to various things like the number of deaths from each race and the to study the correlation of number of cases as days progressed. We used various visual representation of the analysed data to provide a better overview on the dataset. We also tried to find various inferences such as the age groups which were targeted the most and its relationship with the number of cases.

**DATASET**

This dataset was picked from Kaggle ( https://www.kaggle.com/ahsen1330/us-police-shootings ). This dataset had all types of variables like discrete, categorical and continuous consisting for both numerical and non-numerical values. This dataset initially had 4895 rows and 15 columns out of which there were some discrepancies like invalid or null(NaN) data and some duplicate entries which had to be taken care of by Data Cleaning (approximately 4%). The dataset also consisted of inconsistent capitalization and outliers. It consists of various columns like name, age, race, flee, signs\_of\_mental\_illness , manner\_of\_death, and total\_days (difference between day of that incident and the first case in the dataset)

**DATA CLEANING**

The dataset which we picked had 3 columns with invalid entries, hence, it was converted to null(NaN) and either assigned a value based on other values in the column or that entire row was simply deleted from the dataset as we felt that we were lacking the minimum requirements for that row to give us meaningful insights from the available data. All these 3 columns which had these invalid entries were of type string. Then we checked if we had enough data in each row. The rows which had more than 2 null values were simply deleted from the dataset as we felt that it was too many null values for a row.

All the 3 columns which had invalid entries had to be updated to some meaningful data entry. Since all these 3 columns were of type string, i.e., non-numeric, we couldn’t fill in the missing values by performing mathematical or statistical operations like mean, median or mode. So, we decided to fill in these empty places with columns with elements which were adjacent to it, i.e., the ones above it or below it. It wouldn’t make sense to check for duplicate values in any other column apart from names as we assumed that no two victims had the same full names and that they were simply repeated and were deleted from the dataset. For numerical columns, we checked for outliers and deleted rows which contained these values.

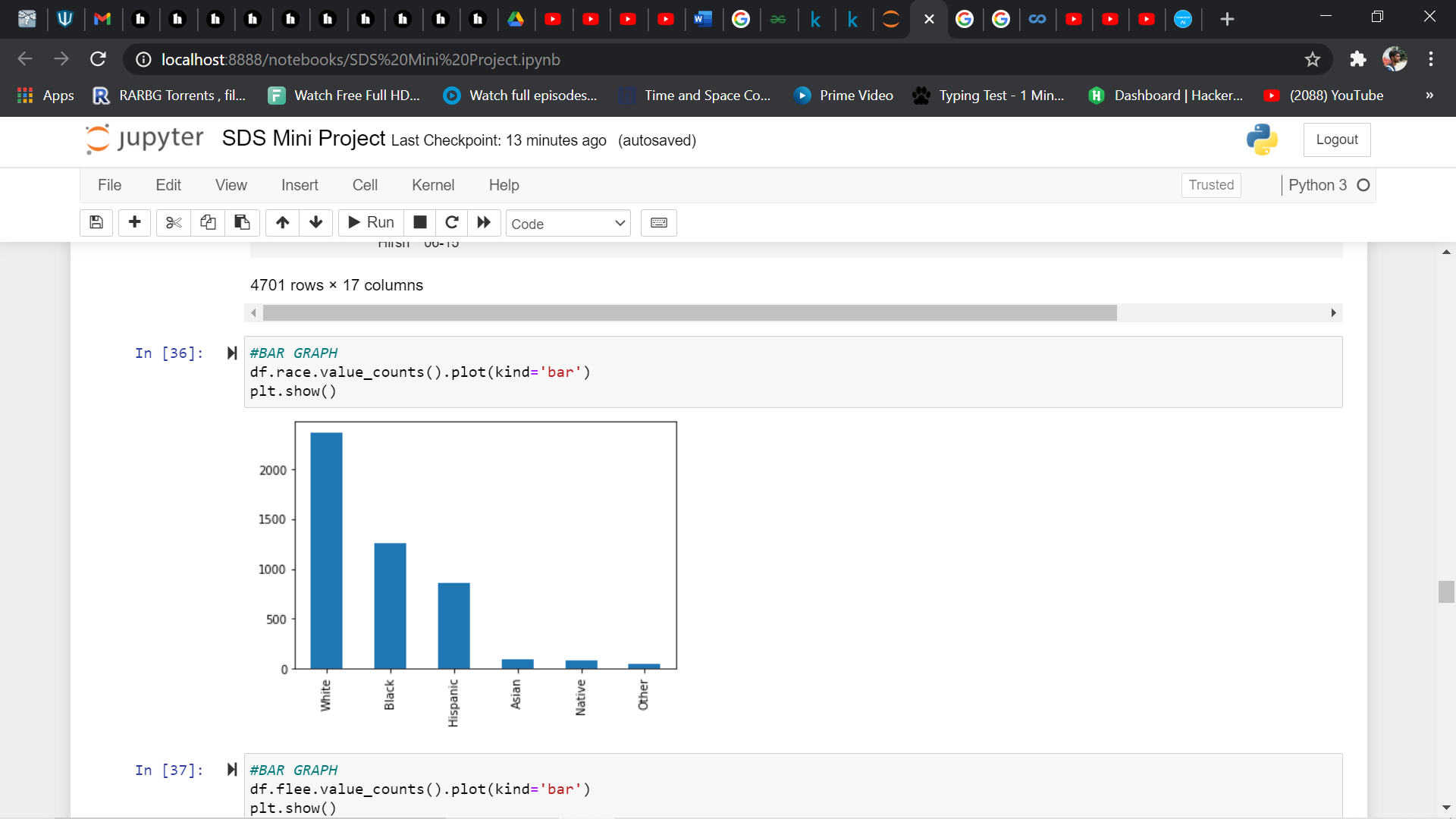
This process is extremely important while analysing the dataset as we depend on the data for analysis and irregularities must first be taken care of before we can actually implement various mathematical and statistical operations on them. Hence, cleaning of a dataset is very crucial for accurate analysis of the dataset and hence for the predictions and conclusions too. If not, the conclusions can appear to be very faulty and that can be very costly and offensive in this case.

**EXPLORATORY DATA ANALYSIS**

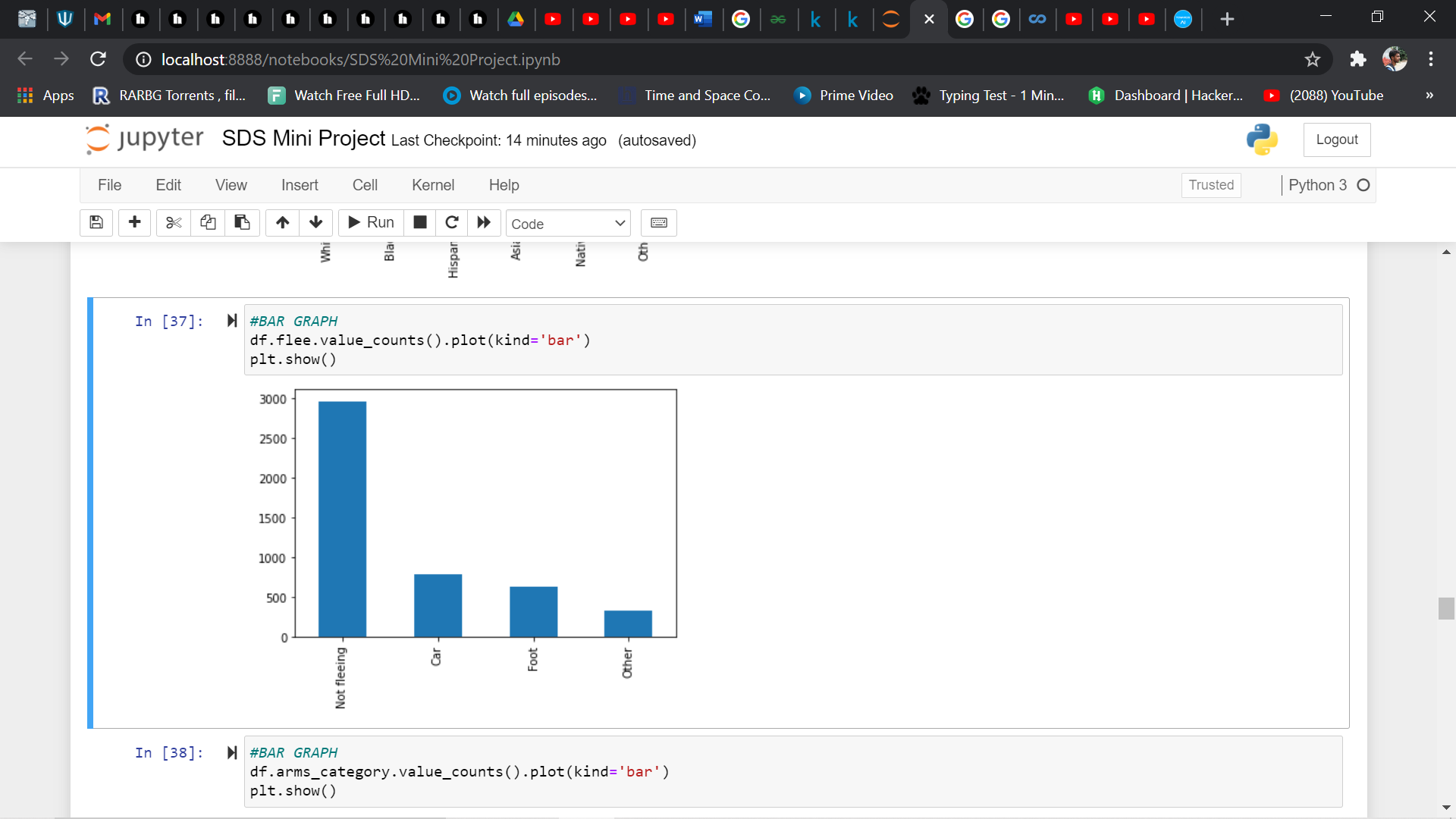
The dataset that we chose had 15 columns out of which only age and id were numerical. Hence, to carry out meaningful data analysis, we decided to add 2 more columns to this dataset from the already available column ‘date’. The columns added were ‘year’ and ‘total\_days’ (difference between day of that incident and the first case in the dataset). The id which was given as a column was very inconsistent as it had a lot of jumps and wasn’t very linear and it had to be reset.

Now coming to insights drawn from the Graphical analysis:

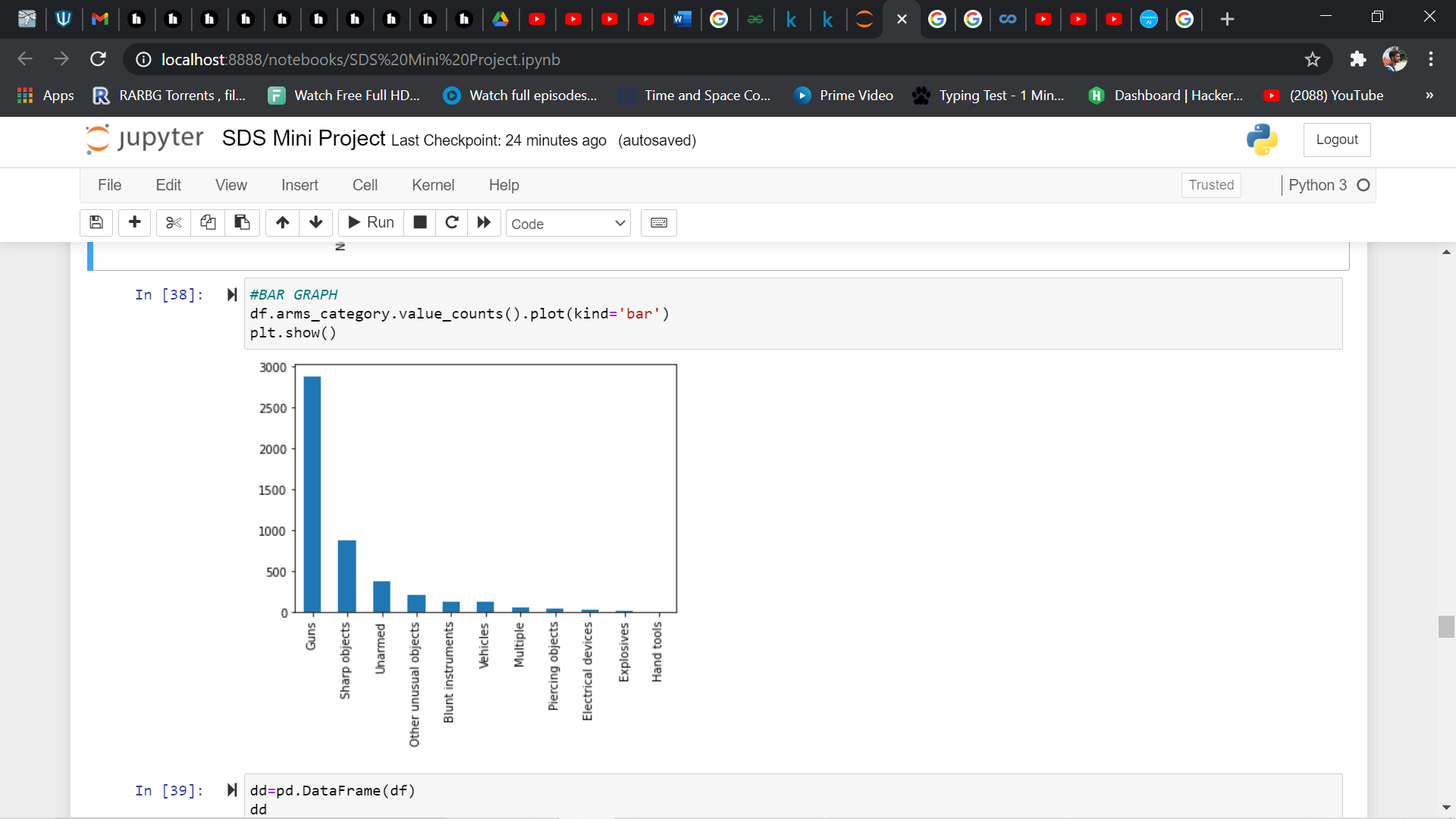
We want to see and compare the number of deaths for different races and categorize them. This can be done by a simple bar graph given below



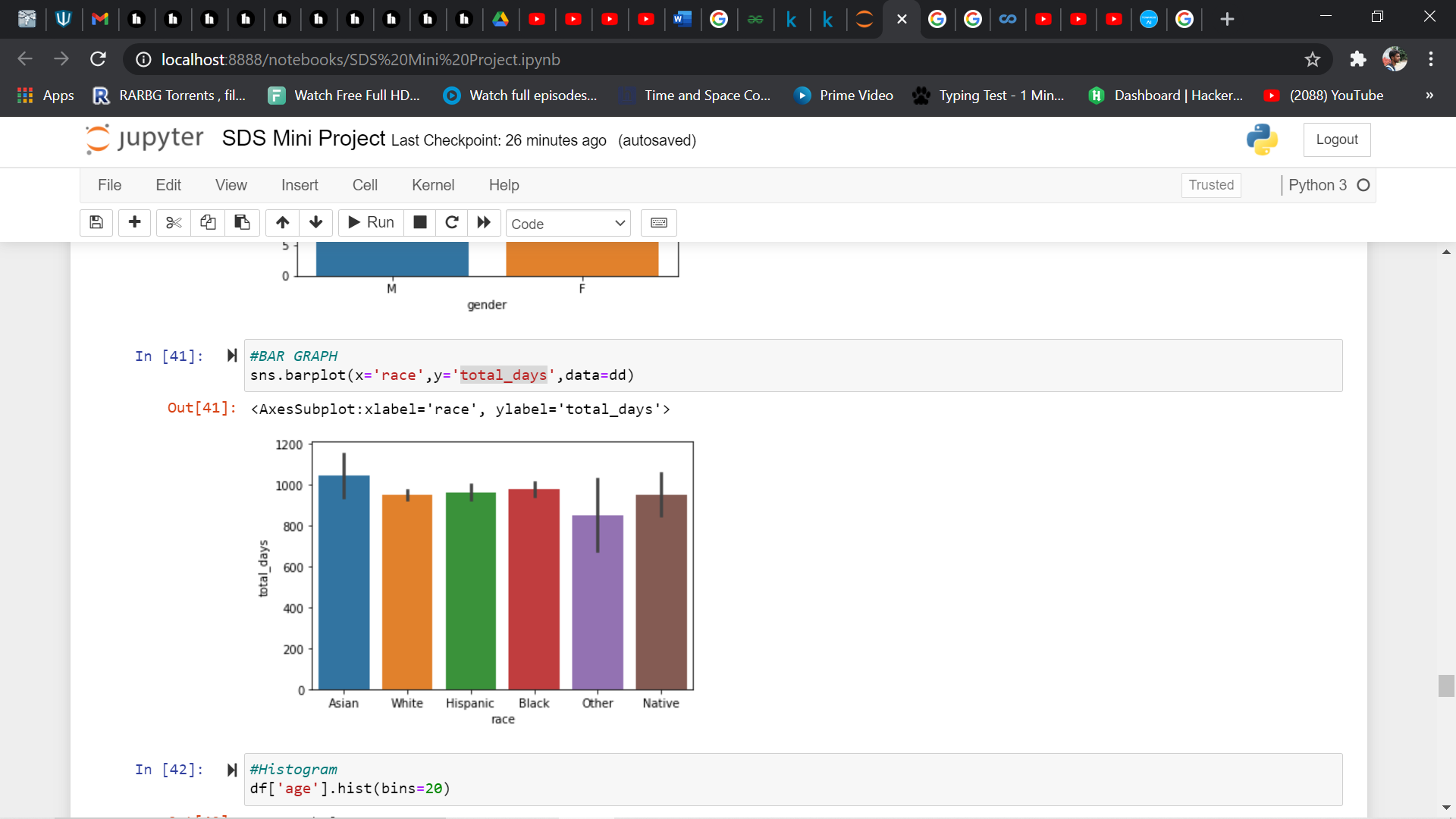
Then, we checked whether the victims were trying to flee away from the cops when they were confronted, if they did try to, then their means are mentioned:



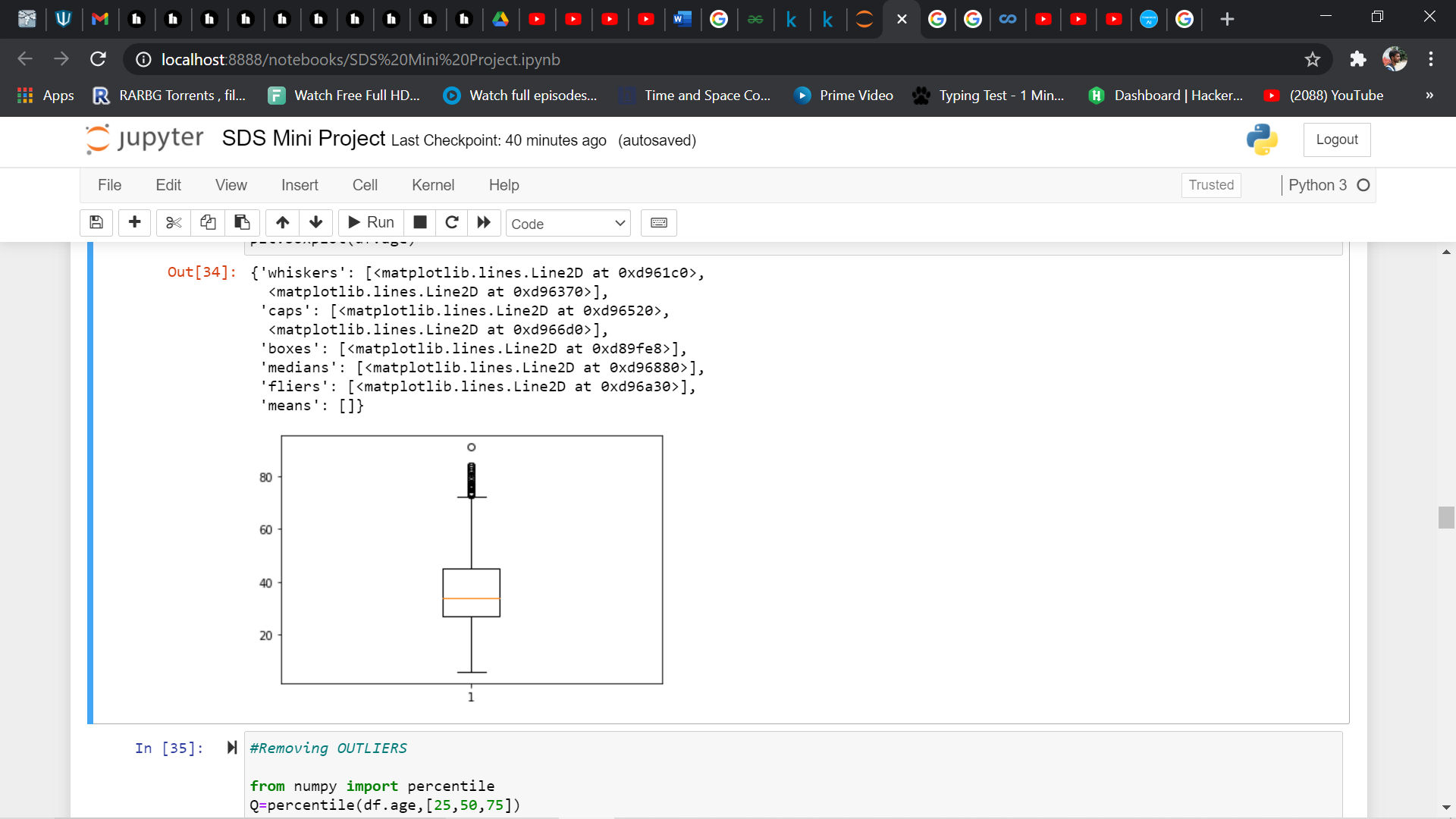
For a better understanding on the dataset, we also categorised the different types of arms (weapons):



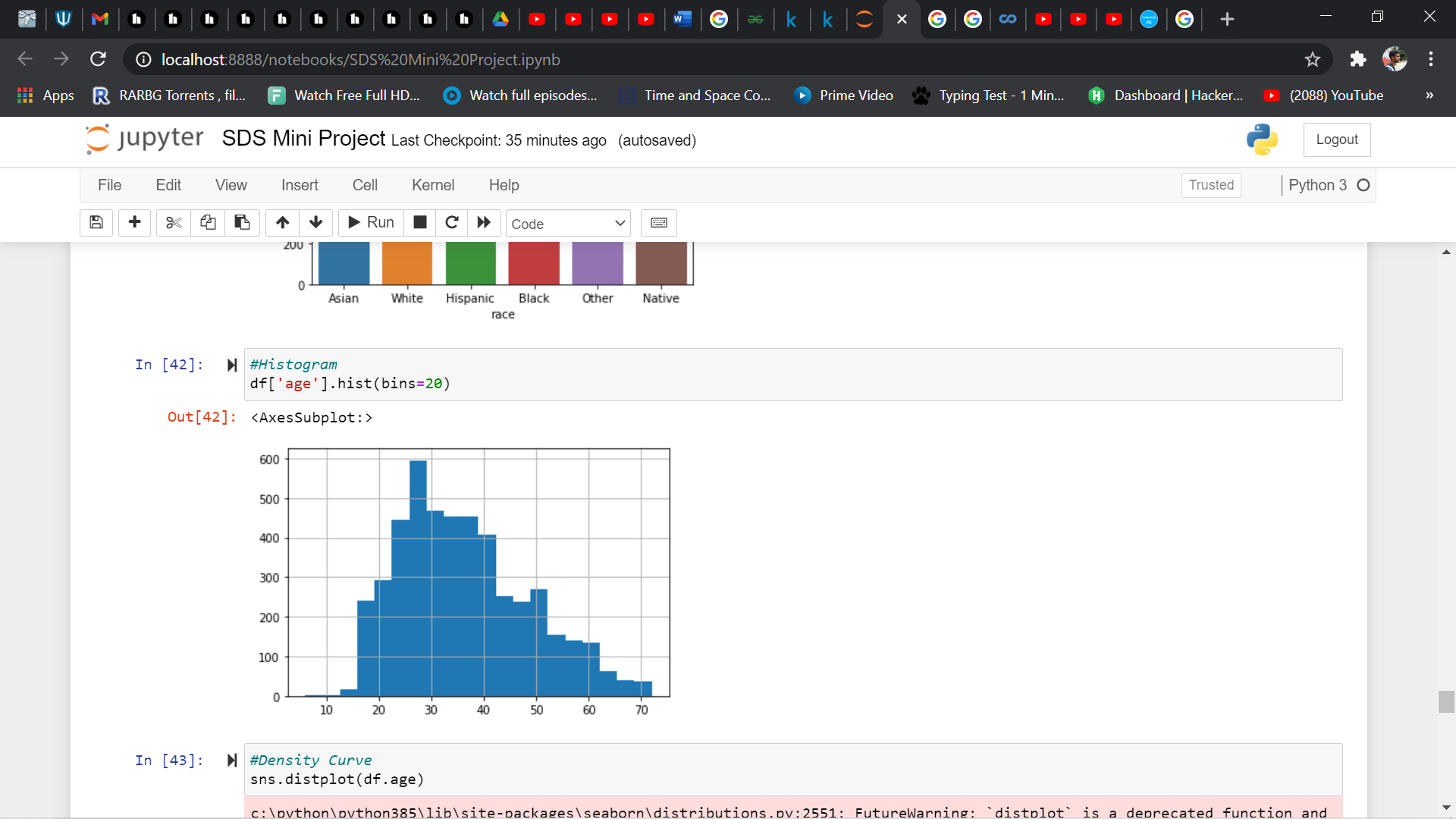
We tried to compare different races w.r.t total\_days (difference between day of that incident and the first case in the dataset). In this analysis, the total number of deaths from a particular community does not matter:



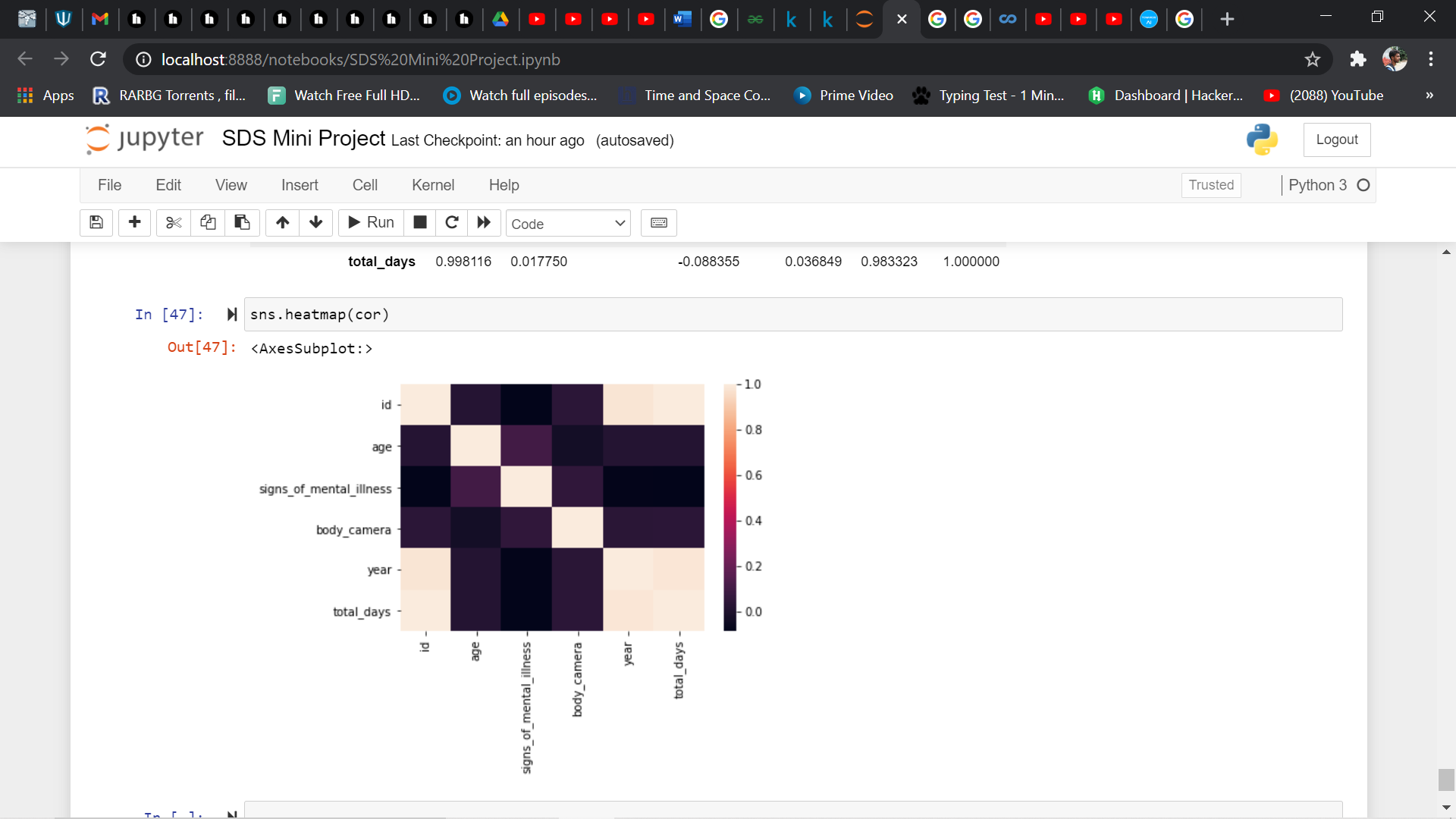
A box plot will tell us about the outliers that are present in the given dataset. It was later manually removed by applying certain algorithms and formulas. Here only the column ‘age’ had outliers:



A histogram of age will give us a rough idea about the age groups of victims:



Heatmap tells us whether there’s a correlation between different variables of the column, if it does exist, it tells us how strongly the 2 variables are related:



This data was Normalized too which only helped us increase the reliability of the data and made it easier for analysing the data more clearly. It also helped us perform certain operations on the dataset which would’ve been harder without normalizing it. The data thus obtained makes it easier for us to predict better and come to more accurate conclusions.

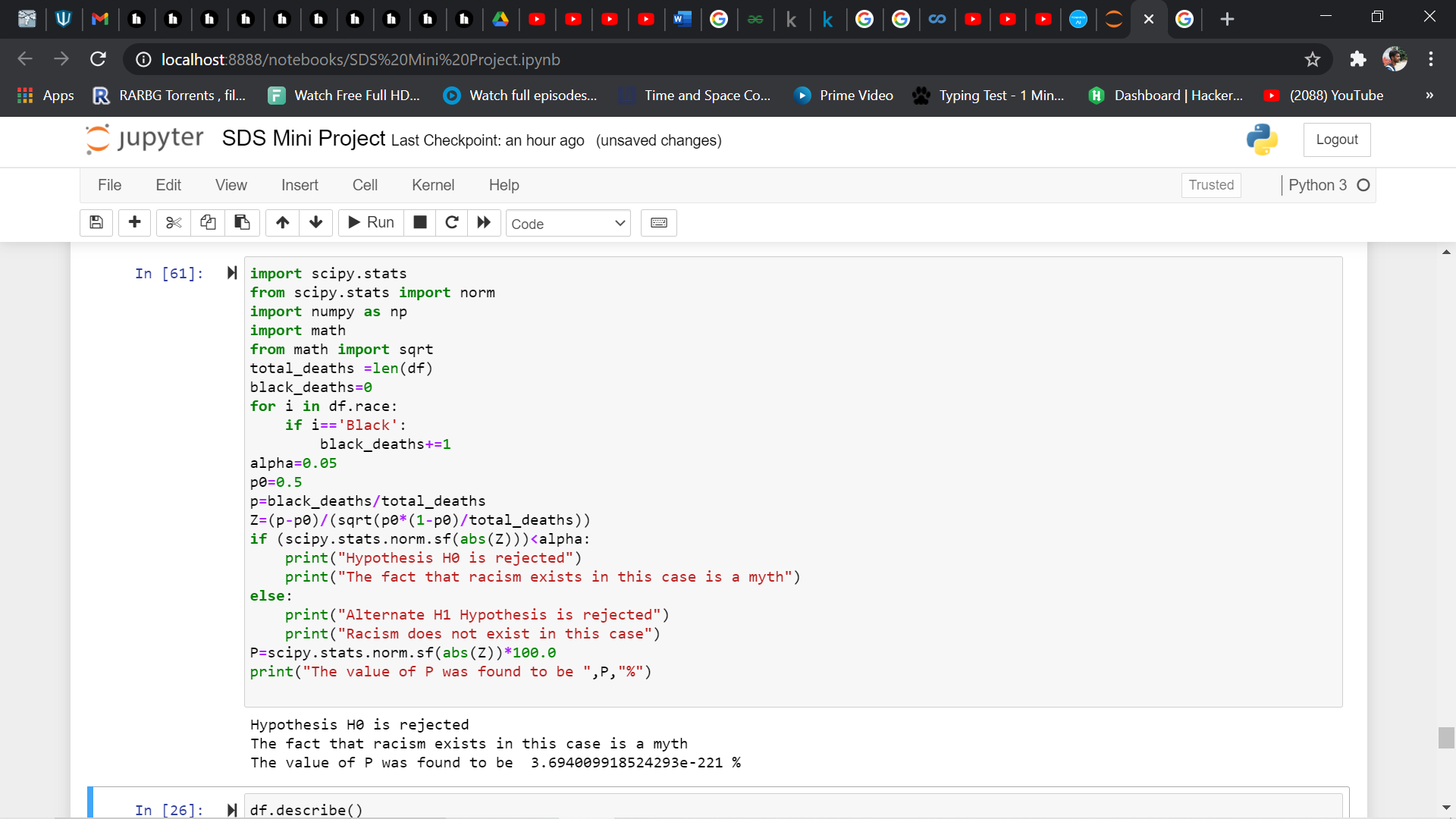
During normalization, all the numerical data columns are scaled down to values between 0 and 1.

When we standardize a data column, then mean will be equal to 0 and variance will be equal to 1 for that column.

**HYPOTHESIS TESTING:**

The black community in general feel that they are vulnerable to the current society and have always been unfairly targeted at.

So, in order to test their claim, we have used the data from the above dataset and tested it using Hypothesis testing. Null Hypothesis, H0, can be considered for the case where 50% or less than that of the total deaths were from the black community and the Alternate Hypothesis, H1, can be considered for the case where more than 50% of the total deaths were from the black community.



The value of P came out to be very small, hence, the initial assumption was incorrect

**RESULTS:**

From Hypothesis Testing, we can clearly conclude that racism when it comes to police brutality is purely a myth and that every race suffers from this unfortunate events. It is also important to note that most of these victims were not trying to flee away and hence the harsh actions that were taken on them seems to be very cruel. We can also see that most of the victims were middle aged and most of them were males and guns were mostly used. We can see that there is a very strong correlation (**≈ 1)** between id and total days. The column id basically tells us about the case number and total days tells us the number of days since the first incident in the dataset. Hence, we can say that as time passes by, these numbers are directly rising. We would like to conclude that there has to be stricter gun laws in the country in order to reduce these numbers significantly.