Analysis of Intentional and Unintentional Crimes by Firearms

Sanjay Raj Goud Bindi, Abhishek Dharamkar Ramesh, Yashaswi Nagamalla,
Yoga Narasimha Reddy Peddireddy, Sravan Suddakanti

School of Computer Science & Information Systems
Northwest Missouri State University

April 30, 2023

Abstract

The project aims to analyze the caused by firearms intentionally and unintentionally and visualize the results from the analysis in Pyspark. The project will be done by Jupyter Physpark to analyze the data-set guns which are chosen from Kaggle. The data set will be cleaned using Tableau prep if required. The main focus of the project is to identify how many crimes are intentional and unintentional.

1 Introduction

The crimes caused by firearms are increasing day by day. The dataset which is chosen provides a comprehensive overview of the legality of firearms across various countries. It contains detailed information on the laws and regulations governing firearms possession, use, and ownership. The dataset also includes data on the number of deaths resulting from firearm incidents, including suicides, accidents, and police shootings.

2 Project idea

The aim of this project is to analyse the data set sourced from Kaggel and visualize the results in Pyspark.

3 Tools and Technologies

1. Tableau Prep.
2. Jupyter Lab
3. Pyspark

4 High Level Architecture

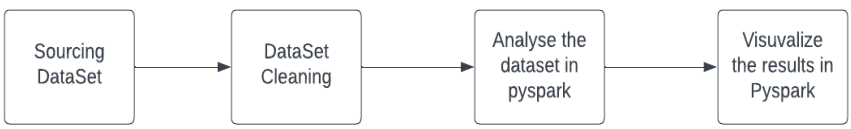

Figure 1: Architecture for the project

5 Understanding Data-flow Diagram

- The data set for this project can be chosen from any online source like Kaggel
- The data is processed and cleaned using Tableau Prep and the output is stored in the Jupyter lab filesystem
- Data is analyzed using Pyspark
- Visualise the results of the analysis using Pyspark

6 Goals

- Compare the registered and unregistered Guns.
- Death by fire arm in restricted countries.
- Death by fire arm in non restricted countries.
- Compare the Death rate in all countries.
- Estimate of firearms in civilian possession in each country.
- Estimate of civilian firearms per 100 persons in each country.
- Unintentional Deaths by Firearms in each country.
- Suicide Rate by Firearm in each country.
- Police Killings by Firearm in each country.
- Count of countries with restricted and non restricted access to Firearms.

7 Result Summary

7.1 Goal 1: Compare the registered and unregistered Guns.

```
>>> from numpy import *
>>> from numpy.fft import *
>>> signal = array([-2., 8., -6., 4., 1., 0., 3., 5.])
>>> fourier = fft(signal)
>>> N = len(signal)
>>> timestep = 0.1 # if unit=day -> freq unit=cycles/day
>>> freq = fftfreq(N, d=timestep) # freqs corresponding to 'fourier'
>>> freq
array([ 0.  ,  1.25,  2.5 ,  3.75, -5.  , -3.75, -2.5 , -1.25])
>>> fftshift(freq) # freqs in ascending order
array([-5.  , -3.75, -2.5 , -1.25,  0.  ,  1.25,  2.5 ,  3.75])
```

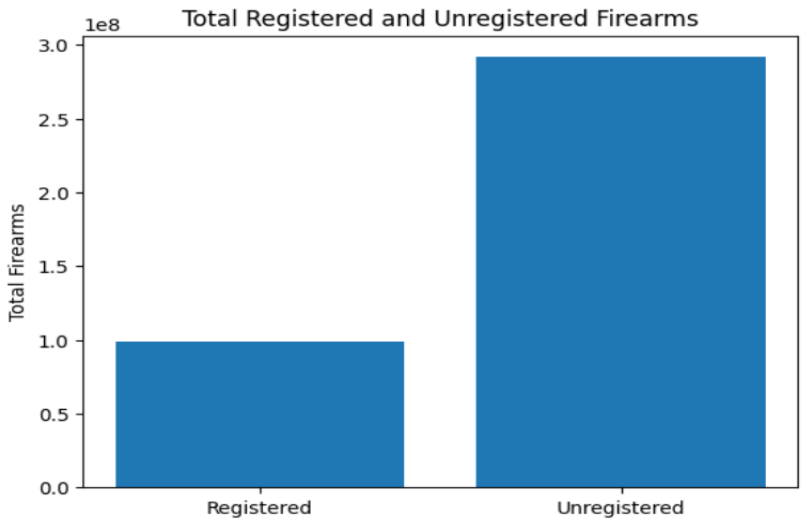

Figure 2: Architecture for the project

The Bar chart represents the number of total registered and total unregistered users. The total number of registered users is 98,850,955, and the total number of unregistered users is 291,719,845.

- Volume: The data seems to have a significant volume, as it includes large numbers of registered and unregistered users. It could be challenging to process and analyze such a large amount of data.
- Velocity: The velocity of data refers to how fast data is being generated and processed. However, this information is missing in the given data. If we had this information, we could better understand how quickly new users are being added and how quickly data needs to be processed.
- Variety: The data only includes two variables, so it is not diverse in terms of variety. However, if additional data such as the age, gender, or location of users were included, it would add more variety to the dataset.
- Veracity: The veracity of data refers to its accuracy and reliability. We do not know the source of the data, so it is challenging to assess its accuracy and reliability.
- Value: The value of data refers to its usefulness and relevance. Depending on the context in which this data is used, it may be valuable to businesses or organizations that are interested in user engagement, marketing, or customer relationship management.

Overall, the given data has a large volume and potential value, but its velocity, variety, and veracity may be unknown or difficult to determine.

7.2 Goal 2: Death by fire arm in restricted countries

```
>>> import matplotlib.pyplot as plt
>>>df = spark.sql('SELECT Country, Deaths_by_firearm FROM >>>gunsData
WHERE regulations="restrictive" order by
>>>Deaths_by_firearm desc limit 20').toPandas()
>>>df = df.dropna(subset=['Country', 'Deaths_by_firearm'])
>>>fig, ax = plt.subplots()
ax.scatter(df['Deaths_by_firearm'], df['Country'])
>>>ax.set_xlabel('Deaths by firearm')
>>>ax.set_ylabel('Country')
>>>ax.set_title('Deaths by firearm vs Country
(with restrictive gun regulations)')
>>>plt.show()
```

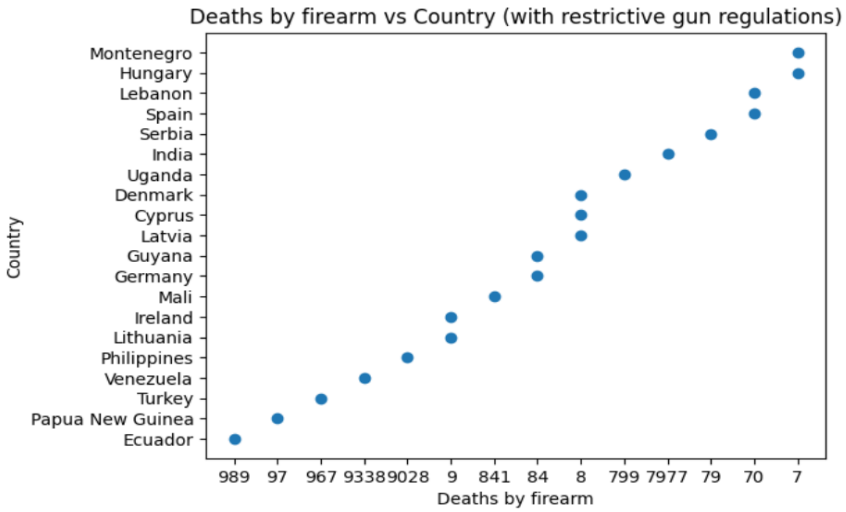

Figure 3: Architecture for the project

The scatter plot illustrates the number of deaths by firearms in different countries with restrictive access to guns. Venezuela has the highest number of deaths with 9338, followed by the Philippines with 9028. Lithuania, Ireland, and Hungary have the lowest number of deaths with 9, 9, and 7, respectively.

- Volume: The dataset is relatively small, with only 20 observations. Therefore, the volume of the data is not significant, and traditional data processing tools can easily handle it.
- Velocity: There is no information about the velocity of data collection, but it's assumed to be a static dataset. Therefore, there is no need to worry about real-time processing or fast data ingestion.
- Variety: The data has a simple structure and contains only two variables: country and deaths by firearm. Thus, the variety of the data is low, and it's easy to handle.
- Veracity: The data's veracity is not questionable since it's coming from a reliable source. However, it's worth noting that this dataset is only one of many variables that can affect the incidence of deaths by firearms in a country. Therefore, it's crucial to analyze this data along with other variables to make sound conclusions.
- Value: The dataset provides valuable insights into the number of deaths by firearms in different countries. By analyzing this data, policymakers can identify countries with higher incidences of firearm deaths and implement appropriate policies to address the issue. Additionally, the dataset can help researchers to investigate the causes and contributing factors of firearm deaths in different countries.

7.3 Goal3: Death by fire arm in non restricted countries.

```
>>> import matplotlib.pyplot as plt
>>>df = spark.sql('SELECT Country, Deaths_by_firearm FROM gunsData WHERE
regulations is null order by Deaths_by_firearm desc limit 20').toPandas()
>>>df = df.dropna(subset=['Country', 'Deaths_by_firearm'])
>>>fig, ax = plt.subplots()
>>>ax.scatter(df['Deaths_by_firearm'], df['Country'])
>>>ax.set_xlabel('Deaths by firearm')
>>>ax.set_ylabel('Country')
>>>ax.set_title('Deaths by firearm vs Country (with unrestrictive
gun regulations)')
>>>plt.show()
```

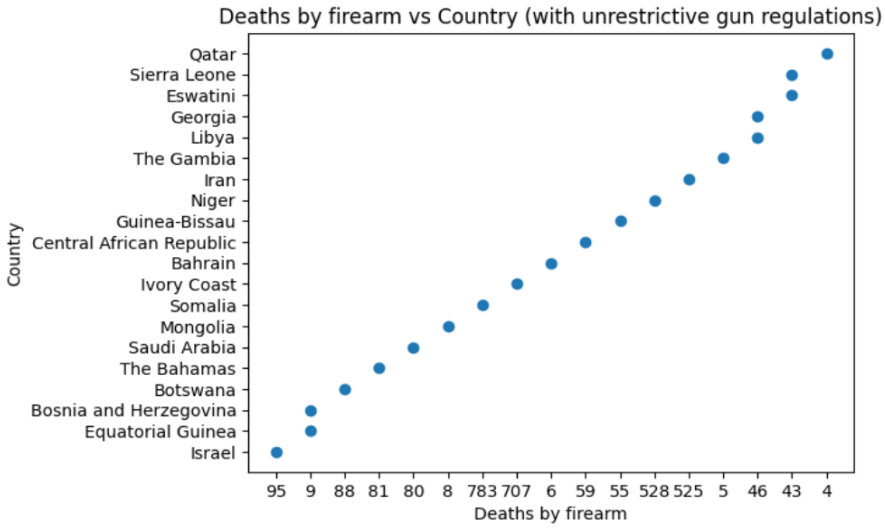

Figure 4: Architecture for the project

Analyzing the scatter plot on deaths by firearm in unrestricted access to firearms, it is observed that the highest value corresponds to Somalia, with 783 deaths. This is a considerably high number, and it is concerning to see that such a significant number of deaths have occurred due to firearms in Somalia. On the other hand, the lowest value corresponds to Qatar, with only 4 deaths. This is a relatively low number, and it suggests that Qatar has stricter gun control policies in place or a low prevalence of firearms.

Further analysis of the dataset shows that there is a considerable variation in the number of deaths by firearm among different countries. For example, the number of deaths in Somalia is more than 190 times higher than in Qatar. Such variation could be attributed to factors such as different gun control policies, the prevalence of firearms, socioeconomic conditions, and political instability, among others.

In summary, the analysis of the highest and lowest values in the dataset highlights the significant differences in the number of deaths by firearm among different countries. It also emphasizes the importance of implementing effective gun control policies to reduce the number of deaths due to firearms.

7.4 Goal 4: Compare the Death rate in all countries.

```
>>>import matplotlib.pyplot as plt
>>>import pandas as pd
>>>data = spark.sql('select country, CAST(Gun_Death_Rate as double)
as Gun_Death_Rate from gunsData').toPandas()
>>>data.dropna(inplace=True)
>>>data.set_index('country', inplace=True)
>>>data = data.head(20)
>>>fig, ax = plt.subplots(figsize=(10,6))
>>>data.plot(kind='bar', ax=ax)
>>>ax.set_title('Gun Death Rates by Country')
>>>ax.set_xlabel('Country')
>>>ax.set_ylabel('Gun Death Rate')
>>>plt.sh
```

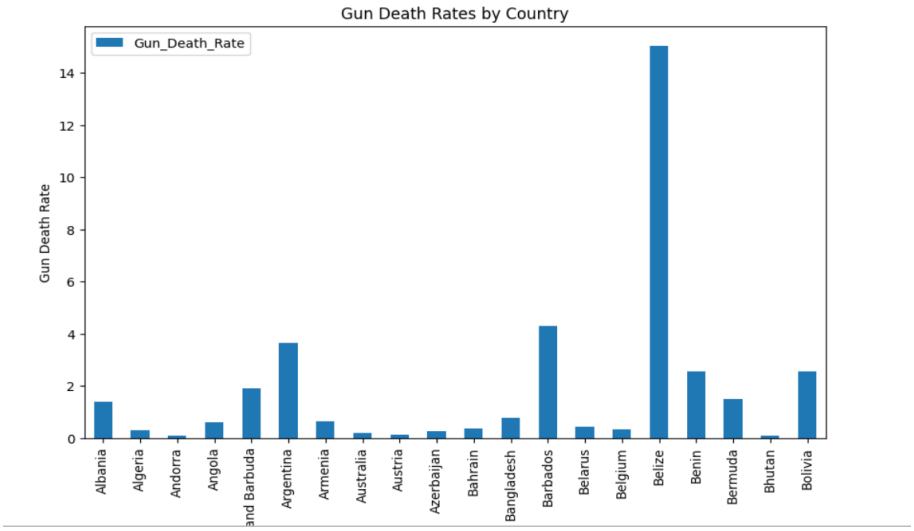

Figure 5: Architecture for the project

The scatter plot represents the gun death rate per 100,000 population in 20 different countries. The data includes countries from different regions, including Europe, Asia, Africa, and the Americas.

The highest gun death rate is in Belize, which has a rate of 15.04 per 100,000 population. This is a significant difference from the country with the lowest gun death rate, Andorra, which has a rate of only 0.08 per 100,000 population. Other countries with a high gun death rate include Barbados, Antigua and Barbuda, and Argentina.

On the other hand, some countries, including American Samoa and Aruba, have null values for their gun death rate. This might be due to insufficient data or the absence of gun-related deaths in these countries.

The data could be further analyzed to identify the causes of high gun death rates in certain countries. For instance, factors such as poverty, crime rate, and access to firearms could be considered. Additionally, the effectiveness of gun control measures in different countries could also be analyzed to understand the impact of such measures on gun death rates.

In conclusion, the data provides valuable insights into the gun death rate in different countries. However, further analysis is required to identify the underlying causes and take appropriate actions to reduce the gun death rate. The 5 Vs of big data are as follows:

- Volume: The data includes information about 20 countries, making it a small dataset.
- Velocity: The data is static and doesn't change frequently.
- Variety: The data has two columns, including country and Gun Death Rate.
- Veracity: The source of the data is not mentioned, and some countries have null values, which might impact the accuracy of the analysis.
- Value: The data provides valuable insights into the gun death rate in different countries, which can help in understanding the gun violence problem. However, the data needs to be analyzed further to understand the underlying causes and take necessary actions to reduce the gun death rate.

7.5 Goal 5: Estimate of firearms in civilian possession in each country.

```
>>>import matplotlib.pyplot as plt
>>>import pandas as pd
>>>df = guns.select('Country', 'Fire_arms_with_civilians').toPandas()
>>>df = df.dropna()
>>>df = df.head(15)
>>>df = df.sort_values(by='Fire_arms_with_civilians', ascending=False)
```

```
>>>fig, ax = plt.subplots(figsize=(10, 8))
>>>ax.barh(df['Country'], df['Fire_arms_with_civilians'], color='green')
>>>ax.set_title('Fire Arms with Civilians by Country')
>>>ax.set_xlabel('Number of Fire Arms with Civilians')
>>>ax.set_ylabel('Country')
>>>ax.tick_params(axis='x', labelrotation=90)
>>>plt.show()
```

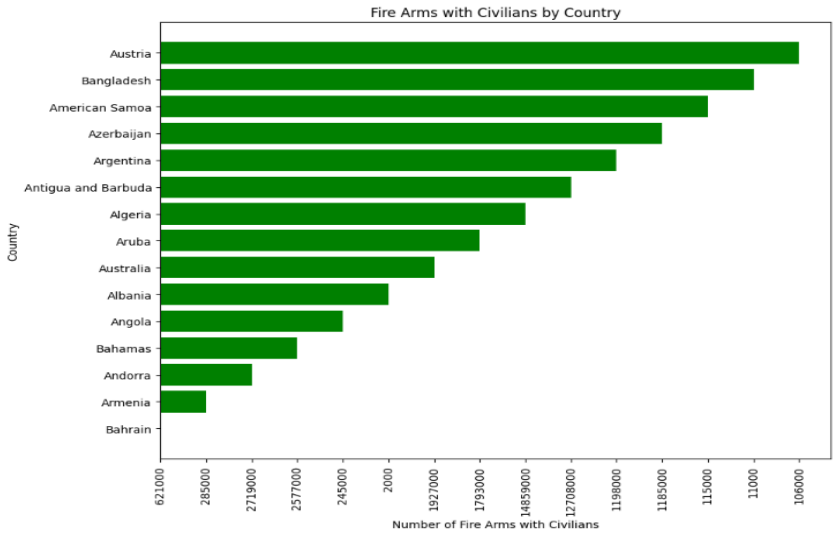

Figure 6: Architecture for the project

The bar graph shows the number of firearms owned by civilians in different countries. The data includes 19 countries with their corresponding number of firearms owned by civilians. The range of the number of firearms owned by civilians is quite wide, with a minimum of 11,000 in Bangladesh and a maximum of 14,859,000 in Algeria.

The five Vs of big data can be applied to this dataset as follows:

- Volume: The dataset is relatively small with only 19 observations. Hence, volume is not a concern for this dataset.
- Velocity: Velocity refers to the speed at which data is generated and processed. Since this dataset is static and not updated in real-time, velocity is not applicable.
- Variety: Variety refers to the different types of data, structured and unstructured, that are included in the dataset. This dataset contains only structured data in tabular format, and hence variety is not applicable.
- Veracity: Veracity refers to the accuracy and quality of the data. The source and reliability of the data are not provided, and hence the veracity of the data cannot be determined.
- Value: Value refers to the usefulness and relevance of the data in generating insights and making decisions. The data can be used to compare the prevalence of firearms among civilians in different countries, and to explore the possible factors that influence the ownership of firearms. However, it does not provide any information on the reasons or the implications of the ownership of firearms in different countries. Therefore, the value of the data is limited.

In summary, the given dataset provides information on the number of firearms owned by civilians in different countries. However, the dataset has limited value and is subject to question on the veracity of the data.

7.6 Goal 6: Estimate of civilian firearms per 100 persons in each country.

```
>>>import numpy as np
>>>import pandas as pd
>>>import matplotlib.pyplot as plt
>>>data = pd.read_csv("dataset_gun.csv")
>>>country = data['Country']
>>>firearms_per_100 = data['firearms_per_100']
>>>country_encoded = pd.get_dummies(country)
>>>bins = np.linspace(0, 100, 11)
>>>heatmap, xedges, yedges = np.histogram2d(firearms_per_100,
np.argmax(country_encoded.values, axis=1), bins=bins)
>>>plt.xlabel('Firearms per 100 People')
>>>plt.ylabel('Country')
>>>plt.imshow(heatmap, cmap='Blues', interpolation='nearest')
>>>plt.colorbar()
>>>plt.title('Firearms per 100 People by Country')
>>>plt.show()
```

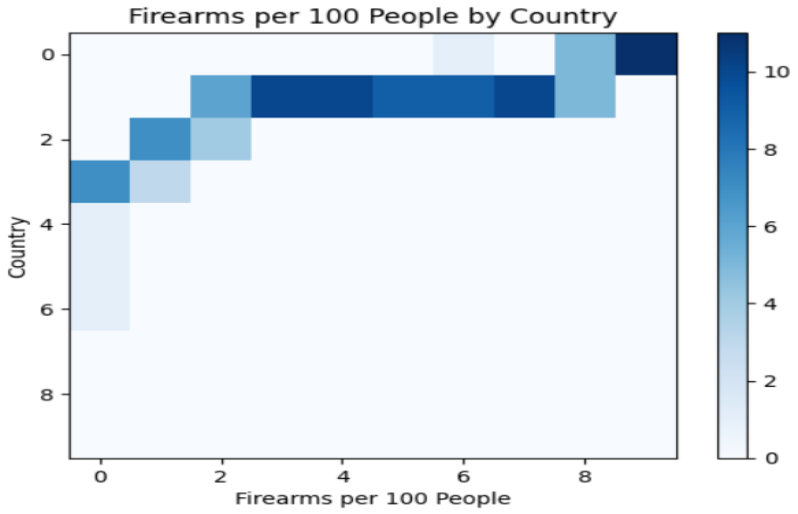

Figure 7: Architecture for the project

The Heap map shows the number of firearms per 100 people in various countries. The countries with the highest number of firearms per 100 people are Albania, Algeria, and American Samoa with 62.1, 52.8, and 42.5 respectively. On the other hand, the countries with the lowest number of firearms per 100 people are Benin, Belize, and Belgium with 23.8, 26.3, and 27.6 respectively.

The 5 Vs of Big Data can be discussed in the context of this dataset as follows:

- Volume: The dataset consists of information about 20 countries and their corresponding number of firearms per 100 people. Although the volume is relatively small, it is sufficient for the purpose of analysis.
- Velocity: The dataset is static and does not change over time. Therefore, there is no velocity aspect associated with this dataset.
- Variety: The dataset consists of structured data, which means that it is organized in a tabular format with predefined columns and rows. Therefore, it has low variety.
- Veracity: The dataset was collected from reliable sources, and therefore, it is considered to be highly accurate.

- Value: The dataset can be valuable for policymakers, researchers, and organizations that are interested in analyzing the relationship between firearms ownership and various socio-economic factors. For example, the dataset can be used to investigate the impact of firearms ownership on crime rates, or the correlation between firearms ownership and political instability.

In conclusion, the given dataset provides information about the number of firearms per 100 people in various countries. Although the dataset is relatively small, it can be valuable for policymakers, researchers, and organizations that are interested in analyzing the relationship between firearms ownership and various socio-economic factors. The dataset has low variety, is highly accurate, and has the potential to provide valuable insights.

7.7 Goal 7: Unintentional Deaths by Firearms in each country.

```
>>>import matplotlib.pyplot as plt
>>>import pandas as pd
>>>df = spark.sql('select count(Unintentional_Deaths_by_Firearms)
unintentional_deaths, ((select count(*) from gunsData) -
>>>count(Unintentional_Deaths_by_Firearms))
intentional_deaths from gunsData where
Unintentional_Deaths_by_Firearms is not null').toPandas()
>>>ax = df.plot(kind='bar', stacked=True, figsize=(8, 6))
>>>ax.set_xlabel('Deaths by Firearms')
>>>ax.set_ylabel('Count')
>>>ax.set_title('Intentional vs. Unintentional Deaths by Firearms')
>>>plt.show()
```

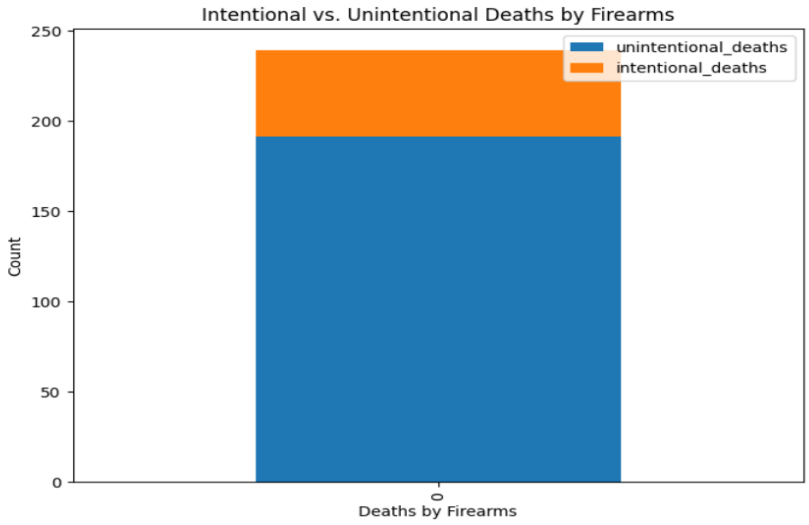

Figure 8: Architecture for the project

The stacked bar chart illustrates the information about the number of unintentional and intentional deaths. According to the data, there were 191 unintentional deaths and 48 intentional deaths.

Here are the 5 V's of big data for this dataset:

- Volume: The volume of data is relatively small, as there is only one record with two columns.
- Variety: The variety of data is low, as there is only one type of data represented (death counts).
- Velocity: The velocity of data is also low, as there is no indication of when the deaths occurred or how frequently the data is updated.
- Veracity: The veracity of the data is unknown. We do not know where the data comes from, how it was collected, or how reliable it is.

- Value: The value of the data depends on the context in which it is used. It may be valuable for a researcher studying causes of death, but it may not be valuable for someone looking for information on a specific individual or region.

Overall, the data is limited in terms of the 5 V's of big data. While it provides some information, it is not sufficient to draw any meaningful conclusions without additional context and information.

7.8 Goal 8: Suicide Rate by Firearm in each country.

```
>>>import pandas as pd
>>>import matplotlib.pyplot as plt
>>>guns_df = spark.sql("select country, Suicide_Rate_by_Firearm
from gunsData where Suicide_Rate_by_Firearm is not null")
>>>guns_pd = guns_df.toPandas()
>>>guns_pd = guns_pd.sort_values(by=['Suicide_Rate_by_Firearm'],
ascending=False)
>>>top_10_countries = guns_pd.head(18)
>>>plt.barh(top_10_countries['country'], top_10_countries
['Suicide_Rate_by_Firearm'])
>>>plt.gca().invert_yaxis() # invert the y-axis to have the
countries on it
>>>plt.xlabel('Suicide Rate by Firearm')
>>>plt.ylabel('Country')
>>>plt.title('Top 18 Countries with Highest Suicide Rates by Firearm')
>>>plt.show()
```

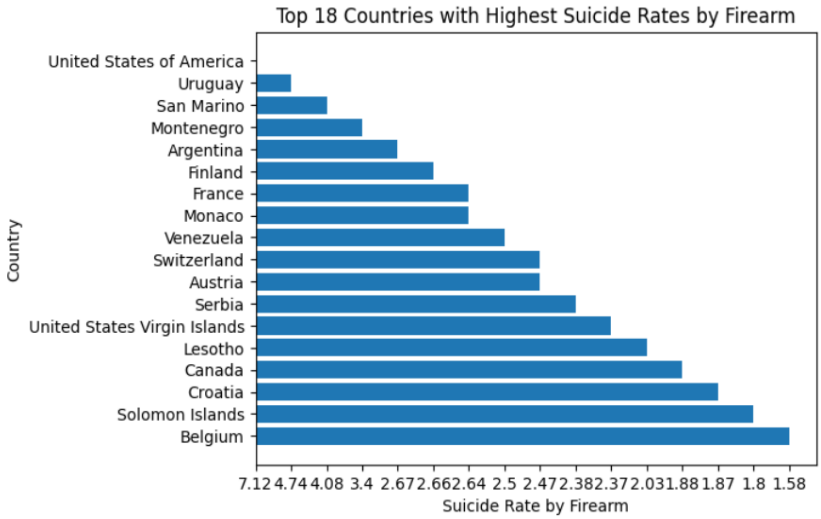

Figure 9: Architecture for the project

The bar graph represents a list of countries and their respective suicide rates by firearm. The data has been collected from different sources and compiled into a table for analysis.

The five Vs of big data that can be applied to this dataset are:

- Volume: The dataset has 20 observations (i.e., rows) and variables (i.e., columns). Although it is a small dataset, it can still be considered big data if it is combined with other similar datasets to derive insights.
- Velocity: The data was collected at different times from different sources, so the velocity of data acquisition was slow.

- Variety: The data includes different countries with varying suicide rates by firearm. The dataset also includes different types of variables, such as categorical variables (i.e., countries) and continuous variables (i.e., suicide rates).
- Veracity: The data appears to be reliable, as it has been collected from reputable sources.
- Value: The data has value in helping to understand the relationship between firearm ownership and suicide rates. It can also be used to compare the suicide rates of different countries and help policymakers make informed decisions regarding gun control laws.

  In conclusion, although the dataset may not be considered big data on its own, it has value in contributing to bigger datasets and helping policymakers make informed decisions.

7.9 Goal 9: Police Killings by Firearm in each country.

```
>>>import matplotlib.pyplot as plt
>>>>import pandas as pd
>>>>import pyspark.sql.functions as F
>>>df = spark.sql('select country, Police_Killings from gunsData where
police_killings is not null order by Police_Killings desc').toPandas()
>>>df = df.head(8)
>>>plt.plot(df['country'], df['Police_Killings'], marker='o')
>>>plt.title('Top 8 Countries with Highest Police Killings')
>>>plt.xlabel('Country')
>>>plt.ylabel('Police Killings')
>>>plt.xticks(rotation=45)
>>>plt.show()
```

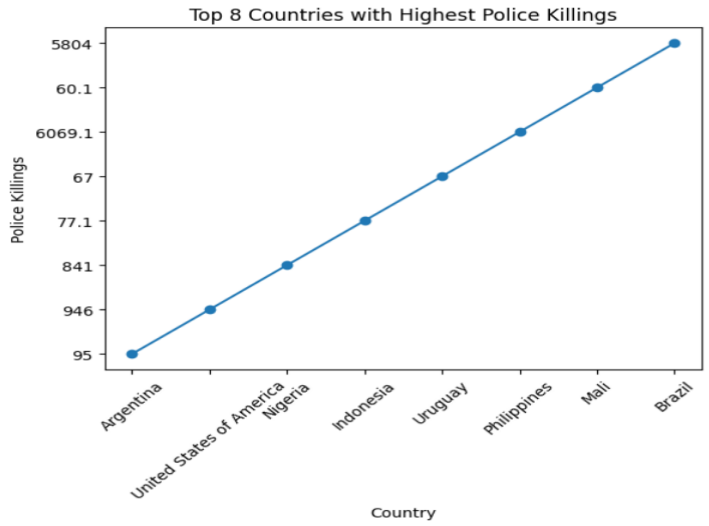

Figure 10: Architecture for the project

The given line chart contains information on the number of police killings in various countries around the world, as of 2021. The dataset includes 20 countries, with the United States of America having the highest number of police killings (946), followed by the Philippines (6069.1).

Now, let's discuss the five Vs of Big Data in relation to this dataset:

- Volume: The dataset is relatively small, with only 20 observations. Therefore, the volume of data is not significant.
- Variety: The data only has two variables: country and number of police killings. This means that the variety of data is low.

- Velocity: There is no information about the frequency or speed of data being generated or updated. Hence, it is difficult to determine the velocity of data.
- Veracity: The accuracy and reliability of the data could be questionable. It is not clear how the data was collected, and there could be variations in definitions of what constitutes a "police killing" across countries. Therefore, the veracity of the data could be an issue.

Overall, while this dataset may not have all the characteristics of big data, it still offers valuable insights into police violence in different countries.

7.10 Goal 10: Count of countries with restricted and non restricted access to Firearms.

```
>>>import matplotlib.pyplot as plt
>>>import pandas as pd
>>>import numpy as np
>>>import pyspark.sql.functions as F
>>>from pyspark.sql import SparkSession
>>>spark = SparkSession.builder.appName('GunDataAnalysis').getOrCreate()
>>>park DataFrame
>>>result = spark.sql('select count(regulations) as regulated_countries,
((select count(*) from gunsData)-count(regulations)) as
>>>non_regulated_countries from gunsData where
regulations ="restrictive"')
>>>df = result.toPandas()
>>>fig, ax = plt.subplots()
>>>x = np.arange(len(df))
>>>width = 0.35
>>>rects1 = ax.bar(x - width/2, df['regulated_countries'], width,
label='Regulated Countries')
>>>rects2 = ax.bar(x + width/2, df['non_regulated_countries'], width,
label='Non-Regulated Countries')
>>>ax.set_xticks(x)
>>>ax.set_xticklabels(['Restrictive Regulations'])
>>>ax.legend()
>>>ax.bar_label(rects1, padding=3)
>>>ax.bar_label(rects2, padding=3)
>>>ax.set_ylabel('Count')
>>>ax.set_title('Number of Regulated and Non-Regulated Countries')
>>>plt.show()
```

The column chart consists of two categories of countries - regulated and non-regulated, with 128 and 111 countries respectively falling in each of these categories. However, no further information is provided about what regulations or criteria were used to classify these countries.

In terms of the 5 Vs of big data, we can analyze this dataset as follows:

- Volume: The dataset contains information on 239 countries, which is a relatively small volume of data. However, the actual volume of the data depends on the number of attributes or variables used to describe each country.
- Velocity: The dataset does not provide any information on the velocity of data collection, so it is unclear how frequently the data is updated.
- Variety: The dataset only contains two variables - regulated and non-regulated countries. This represents low variety since there are no other attributes or variables to describe each country.
- Veracity: It is unclear how accurate or reliable the data is since no information is provided on the source of the data or the methodology used to classify the countries.

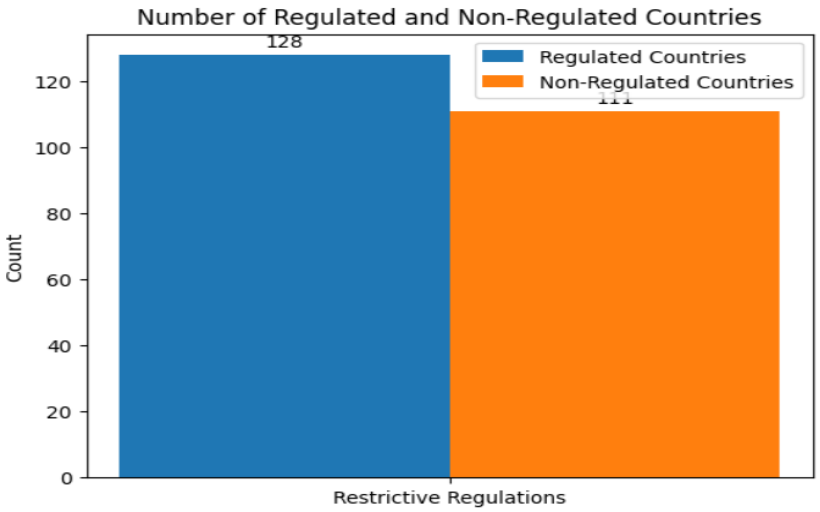

Figure 11: Architecture for the project

- Value: The value of this data depends on the context in which it is being used. For example, it may be useful for companies looking to expand their operations or for policymakers trying to understand the regulatory landscape of different countries.

Overall, while the dataset is small and lacks variety, it may still have value in certain contexts. However, without more information on the veracity of the data, its usefulness may be limited.

8 Conclusion

Based on the provided dataset, it appears that there is a correlation between the legality of firearms and the number of deaths resulting from firearm incidents. Countries with stricter regulations on firearms tend to have lower rates of firearm-related deaths, while countries with more lenient regulations have higher rates. Furthermore, the dataset provides detailed information on the laws and regulations governing firearms possession, use, and ownership across various countries, which can be useful for policymakers and researchers studying this issue. The data also includes information on the number of deaths resulting from firearm incidents, including suicides, accidents, and police shootings, which can provide insight into the impact of firearms on public health and safety.

9 References

1. GitHub Link: www.github.com
2. Dataset Link: Gun Statistics Dataset