Analysis of On-duty Police Officers Deaths in the United States

Problem Statement and Background:

The purpose of the project is to Analyse the deaths of on-duty police officers in the United States from 1900s through 2020.

Every year since 1900s, the number of police officers who lost their lives has risen gradually. There were no comprehensive statistics available on the cause of police deaths.
As a result, we wanted to gain insights and dig deep into the factors causing their death to derive analysis. So, this built the interest in us to take up this project.

We intended to acquire reliable insights by carefully assessing and analysing the data. We intended for this project to be used in the future so that all future data could be appended to the existing data to efficiently generate valuable insights from the data, in a way that it could be used in the future as well. Since delivering the finest analytics was one of our main goals, consistency was one of the important success indicators. The data should be consistent and up to date.

Raw Data Structure:

The data contains 4 fields and includes the officer's names, the department they worked for, the state they were assigned to, the End of Watch (EOW), and the reason of death.

EOW: used as an abbreviation when referring to the date of death of a police officer killed in the line of duty.

The US government, Police Security and Authority address this problem. From the insights gained from this problem, to some extent police officials could be protected and analysing the root causes if any, thereby addressing the issue and make this world a safer place for each and everyone.

Methods:

The raw data that we obtained 4 key fields.

1. Police Official's Name
2. Department they belonged to, along with the state
3. End of Watch (EOW)
4. Cause of Death

1.Data Cleaning and Organization:

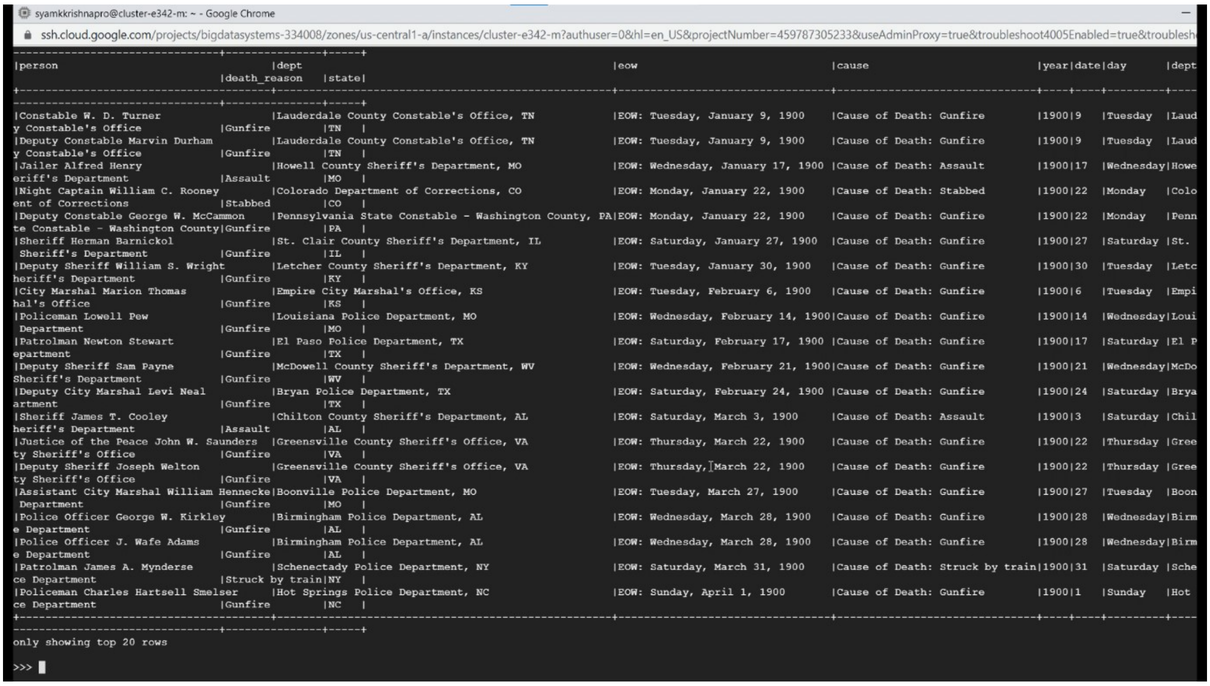
We pulled the raw data in the CSV file format into the jupyterhub notebook and used pandas library to store the raw CSV data into a dataframe. We created six additional columns from the exiting above colums, they are:

- death_reason (extracted from 4)
- date (modified and extracted from 3)
- month (also extracted from 3)
- year (extracted from 3)
- dept_name (extracted from 2)

- state (extracted from 2)

we have filtered data by using both spark and pandas. Here are the screen prints

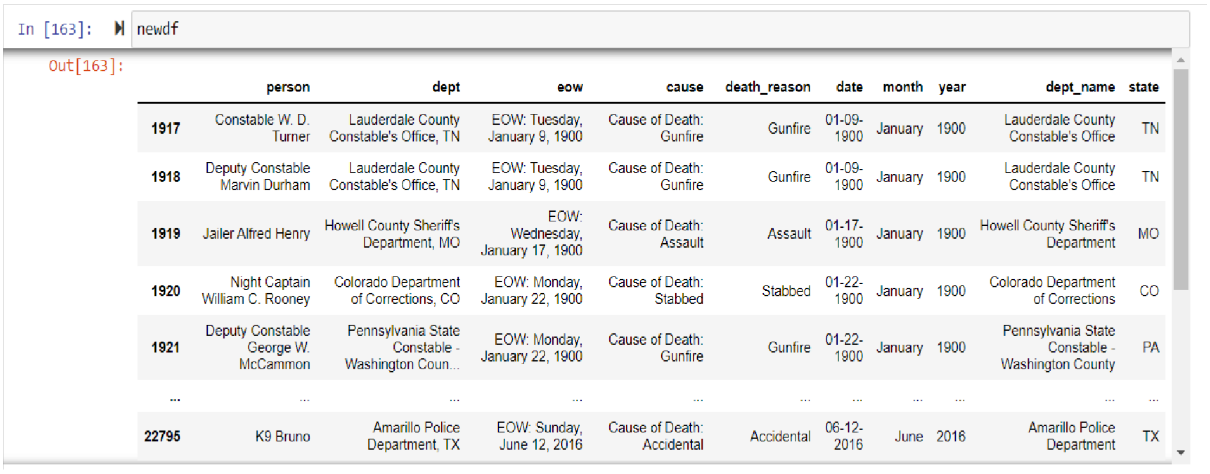
Data filtered by spark:




```
-------------------------------+--------------+-----+
|person                                  |dept                                                 |eow                                |cause                          |year|date|day      |dept
                               |death_reason  |state|
+----------------------------------------+-----------------------------------------------------+-----------------------------------+-------------------------------+----+----+---------+----
-------------------------------+--------------+-----+
|Constable W. D. Turner                  |Lauderdale County Constable's Office, TN             |EOW: Tuesday, January 9, 1900      |Cause of Death: Gunfire        |1900|9   |Tuesday  |Laud
y Constable's Office           |Gunfire       |TN   |
|Deputy Constable Marvin Durham          |Lauderdale County Constable's Office, TN             |EOW: Tuesday, January 9, 1900      |Cause of Death: Gunfire        |1900|9   |Tuesday  |Laud
y Constable's Office           |Gunfire       |TN   |
|Jailer Alfred Henry                     |Howell County Sheriff's Department, MO               |EOW: Wednesday, January 17, 1900   |Cause of Death: Assault        |1900|17  |Wednesday|Howe
eriff's Department             |Assault       |MO   |
|Night Captain William C. Rooney         |Colorado Department of Corrections, CO               |EOW: Monday, January 22, 1900      |Cause of Death: Stabbed        |1900|22  |Monday   |Colo
ent of Corrections             |Stabbed       |CO   |
|Deputy Constable George W. McCammon     |Pennsylvania State Constable - Washington County, PA|EOW: Monday, January 22, 1900      |Cause of Death: Gunfire        |1900|22  |Monday   |Penn
te Constable - Washington County|Gunfire       |PA   |
|Sheriff Herman Barnickol                |St. Clair County Sheriff's Department, IL            |EOW: Saturday, January 27, 1900    |Cause of Death: Gunfire        |1900|27  |Saturday |St.
 Sheriff's Department          |Gunfire       |IL   |
|Deputy Sheriff William S. Wright        |Letcher County Sheriff's Department, KY              |EOW: Tuesday, January 30, 1900     |Cause of Death: Gunfire        |1900|30  |Tuesday  |Letc
heriff's Department            |Gunfire       |KY   |
|City Marshal Marion Thomas              |Empire City Marshal's Office, KS                     |EOW: Tuesday, February 6, 1900     |Cause of Death: Gunfire        |1900|6   |Tuesday  |Empi
hal's Office                   |Gunfire       |KS   |
|Policeman Lowell Pew                    |Louisiana Police Department, MO                      |EOW: Wednesday, February 14, 1900|Cause of Death: Gunfire        |1900|14  |Wednesday|Loui
 Department                    |Gunfire       |MO   |
|Patrolman Newton Stewart                |El Paso Police Department, TX                        |EOW: Saturday, February 17, 1900  |Cause of Death: Gunfire        |1900|17  |Saturday |El P
epartment                      |Gunfire       |TX   |
|Deputy Sheriff Sam Payne                |McDowell County Sheriff's Department, WV             |EOW: Wednesday, February 21, 1900|Cause of Death: Gunfire        |1900|21  |Wednesday|McDo
Sheriff's Department           |Gunfire       |WV   |
|Deputy City Marshal Levi Neal           |Bryan Police Department, TX                          |EOW: Saturday, February 24, 1900  |Cause of Death: Gunfire        |1900|24  |Saturday |Brya
artment                        |Gunfire       |TX   |
|Sheriff James T. Cooley                 |Chilton County Sheriff's Department, AL              |EOW: Saturday, March 3, 1900       |Cause of Death: Assault        |1900|3   |Saturday |Chil
heriff's Department            |Assault       |AL   |
|Justice of the Peace John W. Saunders   |Greensville County Sheriff's Office, VA              |EOW: Thursday, March 22, 1900      |Cause of Death: Gunfire        |1900|22  |Thursday |Gree
ty Sheriff's Office            |Gunfire       |VA   |
|Deputy Sheriff Joseph Welton            |Greensville County Sheriff's Office, VA              |EOW: Thursday, March 22, 1900      |Cause of Death: Gunfire        |1900|22  |Thursday |Gree
ty Sheriff's Office            |Gunfire       |VA   |
|Assistant City Marshal William Hennecke|Boonville Police Department, MO                       |EOW: Tuesday, March 27, 1900       |Cause of Death: Gunfire        |1900|27  |Tuesday  |Boon
 Department                    |Gunfire       |MO   |
|Police Officer George W. Kirkley        |Birmingham Police Department, AL                     |EOW: Wednesday, March 28, 1900     |Cause of Death: Gunfire        |1900|28  |Wednesday|Birm
e Department                   |Gunfire       |AL   |
|Police Officer J. Wafe Adams            |Birmingham Police Department, AL                     |EOW: Wednesday, March 28, 1900     |Cause of Death: Gunfire        |1900|28  |Wednesday|Birm
e Department                   |Gunfire       |AL   |
|Patrolman James A. Mynderse             |Schenectady Police Department, NY                    |EOW: Saturday, March 31, 1900      |Cause of Death: Struck by train|1900|31  |Saturday |Sche
ce Department                  |Struck by train|NY   |
|Policeman Charles Hartsell Smelser      |Hot Springs Police Department, NC                    |EOW: Sunday, April 1, 1900         |Cause of Death: Gunfire        |1900|1   |Sunday   |Hot
ce Department                  |Gunfire       |NC   |
+----------------------------------------+-----------------------------------------------------+-----------------------------------+-------------------------------+----+----+---------+----
-------------------------------+--------------+-----+
only showing top 20 rows

>>>
```

Data filtered by Pandas:



In [163]: newdf

Out[163]:

	person	dept	eow	cause	death_reason	date	month	year	dept_name	state
1917	Constable W. D. Turner	Lauderdale County Constable's Office, TN	EOW: Tuesday, January 9, 1900	Cause of Death: Gunfire	Gunfire	01-09-1900	January	1900	Lauderdale County Constable's Office	TN
1918	Deputy Constable Marvin Durham	Lauderdale County Constable's Office, TN	EOW: Tuesday, January 9, 1900	Cause of Death: Gunfire	Gunfire	01-09-1900	January	1900	Lauderdale County Constable's Office	TN
1919	Jailer Alfred Henry	Howell County Sheriff's Department, MO	EOW: Wednesday, January 17, 1900	Cause of Death: Assault	Assault	01-17-1900	January	1900	Howell County Sheriff's Department	MO
1920	Night Captain William C. Rooney	Colorado Department of Corrections, CO	EOW: Monday, January 22, 1900	Cause of Death: Stabbed	Stabbed	01-22-1900	January	1900	Colorado Department of Corrections	CO
1921	Deputy Constable George W. McCammon	Pennsylvania State Constable - Washington Coun...	EOW: Monday, January 22, 1900	Cause of Death: Gunfire	Gunfire	01-22-1900	January	1900	Pennsylvania State Constable - Washington County	PA
...	...	...	...	...	...	...	...	...	...	...
22795	K9 Bruno	Amarillo Police Department, TX	EOW: Sunday, June 12, 2016	Cause of Death: Accidental	Accidental	06-12-2016	June	2016	Amarillo Police Department	TX

We attempted to identify null values from the data and filter them out. We also filtered out other kinds of values such as ‘N/a’, ‘na’ and ‘np.nan’ from the raw data. After getting rid of all the null values and invalid fields, we began appending new columns to the dataset.

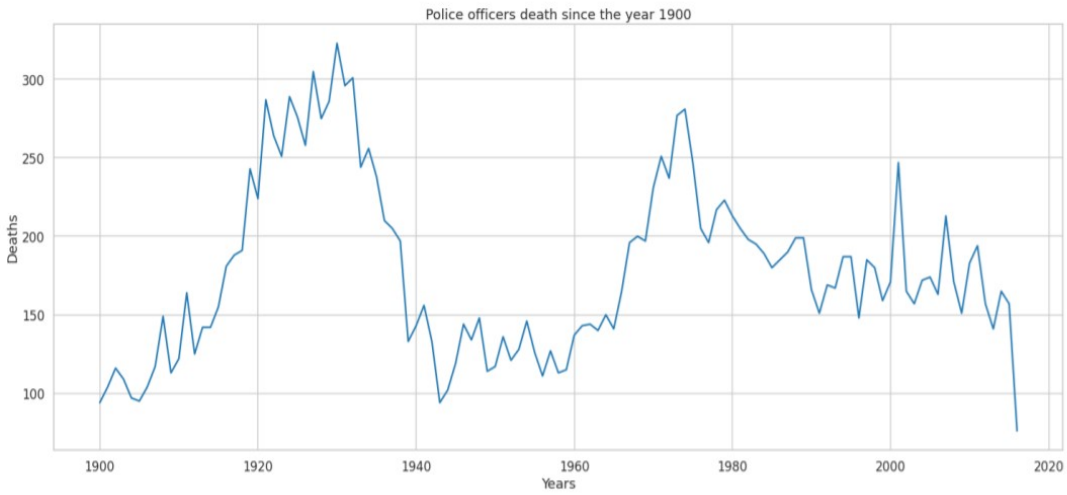
The additional columns have been created in order to make it easy for analysis and processing of data, so that data is in structured format. Data being in structured format makes it easier for any person to understand the data and no further explanation would be required.

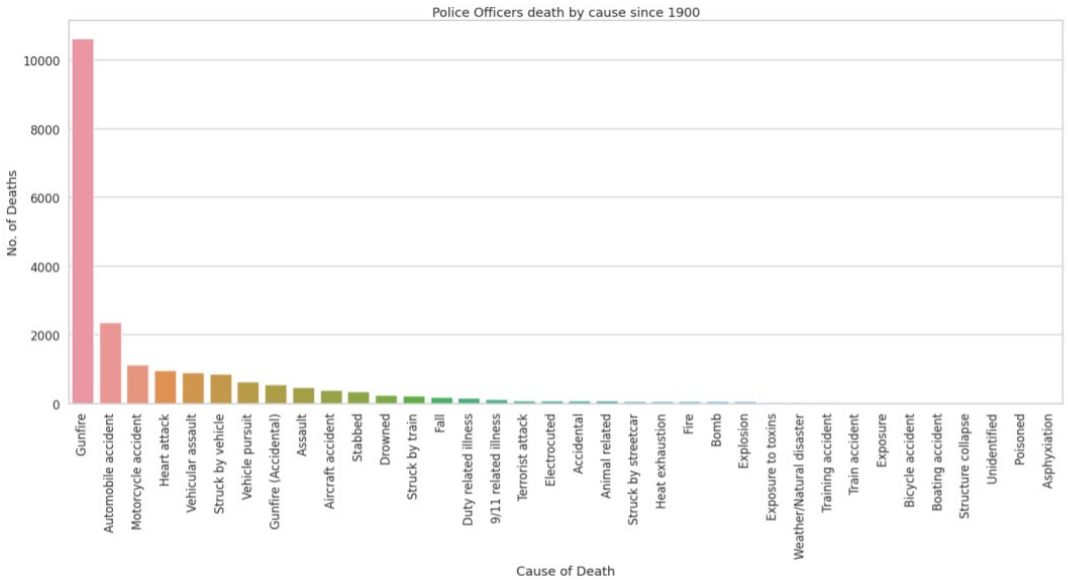
2.Data Visualization:

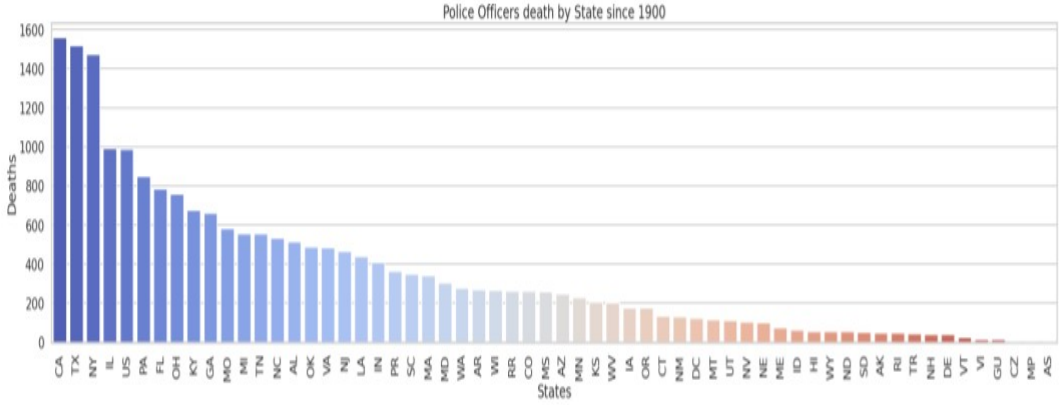
Matplotlib library is a primary data visualization package for Python and Numpy which runs on all platforms. It provides an open-source alternative to MATLAB.

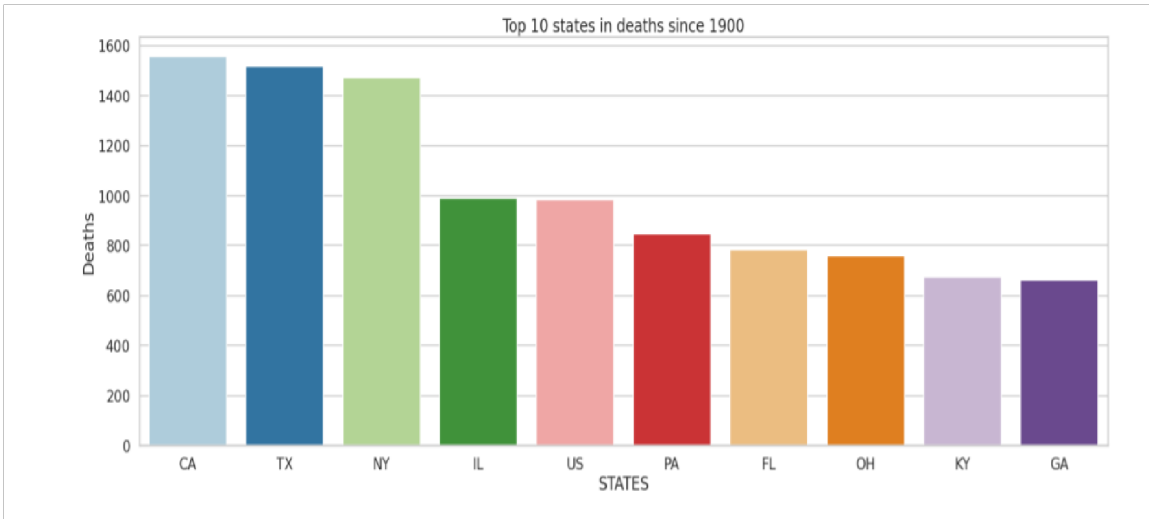
We imported the matplotlib and seaborn libraries from the Python libraries to create visuals that allow us to share our findings in a pictorial format that anyone can understand and benefit from.

- We created a line chart that depicts the number of police officers that died since the year 1900
- Created a barplot that depicts and categorizes the causes of deaths of the police officers since the year 1990.
- Created a barplot that depicts and categorizes into states differentiating between the state with highest number of deaths and leas number of deaths.
- Created a barplot which lists the top 10 states with highest number of police deaths.









3. ML Algorithms used:

Implemented two algorithms

1. Decision Tree Classifier

```
In [685]:
import sklearn
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
#from sklearn.metrics import acccuracy_score

#X=new2df.drop(columns=['person','dept','eow','cause','date','month','dept_name','state','death_reason'])
y=newdf['death_reason']
X_train, X_test, y_train, y_test =train_test_split(X,y,test_size=0.2)

model=DecisionTreeClassifier()
model.fit(X_train, y_train)
predictions=model.predict(X_test)
```

```
In [688]:
from sklearn.metrics import accuracy_score
accuracy_score(y_test,predictions)
```

```
81.24869400526694
```

2. K Nearest Neighbour (KNN)

When k=1

```
In [685]:
import sklearn
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
#from sklearn.metrics import acccuracy_score

#X=new2df.drop(columns=['person','dept','eow','cause','date','month','dept_name','state','death_reason'])
y=newdf['death_reason']
X_train, X_test, y_train, y_test =train_test_split(X,y,test_size=0.2)

model=DecisionTreeClassifier()
model.fit(X_train, y_train)
predictions=model.predict(X_test)

In [688]:
from sklearn.metrics import accuracy_score
accuracy_score(y_test,predictions)*100

81.24869400526694
```

When k=5

```
In [692]:
from sklearn import linear_model
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5)

In [693]:
knn.fit(X_train, y_train)

Out[693]:
KNeighborsClassifier()

In [694]:
knnpred=knn.predict(X_test)

In [695]:
accuracy_score(y_test,knnpred)*100

Out[695]:
47.091213789801294
```

When k=10

```
In [706]:
from sklearn import linear_model
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=10)

In [707]:
knn.fit(X_train, y_train)

Out[707]:
KNeighborsClassifier(n_neighbors=20)

In [708]:
knnpred=knn.predict(X_test)

In [709]:
accuracy_score(y_test,knnpred)*100

Out[709]:
51.089298539621744
```

Both the above algorithms can be used for solving regression and classification. But we implement classification in our project. In the learning step we provided the model 80% of our data and used the rest 20% of the data for testing the accuracy.

Result & Performance Metrics:

1.Spark vs pandas

The runtime was one of our key considerations and we tried to implement multiple nodes in the cluster while using Spark and the runtime for spark seemed to be faster compared to pandas. Since we used multiple nodes in Spark the runtime reduced as and when the numbers of nodes were increased.

Whereas in pandas it uses only single node to perform similar operations as that of Spark, since our data was less than 10GB there was negligible difference between both Spark and Pandas. For much larger

data Spark would be preferable since it has the ability for parallel processing using multiple nodes in a cluster where pandas lack this feature.

2.Quality of data

Quality of data is very much important since it is the heart of the project. So, the quality of the data represents the quality of the output. In order to maintain the quality, we removed the anomalies in the complete dataset such as null values, N/a values, np.nan values and any other miscellaneous values.

3.Accuracy:

Accuracy was one of our key performance indicators since accuracy is used to determine, which model is good or better compared to other models and also to identify any relationships and patterns that are hidden between the variables in a dataset based on the training data provided.

The better the accuracy of a model is, the better the predictions and insights it can develop or produce.

In our approach we used two classifiers to perform predictions using our dataset.

1. The accuracy obtained by Decision Tree Classifier seemed to be 81.24869…%
2. The accuracy obtained by KNN Classifier was 34% when the value of K was 1. The accuracy improved upto 51% gradually until K was incremented by 1 until 10, and thereafter the accuracy didn't change much.

We could conclude that Decision Tree Classifier was more accurate than KNN classifier and it could perform better predictions than KNN.

4.Consistency:

Data should be consistent for each measurement of variables or columns in the dataset. And this consistency is very crucial since we are performing cleaning, organizing and filtering on the raw data and this should not lead to loss of any of the contents of the datasets hence preserving its originality.

The value to be gained from the data before and after applying any operations should not vary i.e there should be no loss of data after performing any of the operations on the dataset.

At any point in time the data should be constant in order to use it in different ways without modifying or removing its original structure

Different Variations in solutions:

We implemented data organizing, cleaning and pre-processing using both Spark and Pandas dataframes. The processing in Spark seemed to be a bit faster with multiple nodes in the cluster, but the overall results were same since our data was less than 10 GB so it didn't make much difference.

For machine learning algorithm, we tried classifying using different values of K, where 'K' is referred to the number of nearest neighbours. Every time we incremented the value of 'K' the accuracy seemed to improve gradually. The accuracy improved for every value of K starting from 1 incremented by 1 upto 10, and thereafter there didn't seem much increase in the accuracy.

Primary Solution Vs Baseline solution:

The primary baseline solution would be as a solution that would be implemented as if we didn't know any data science. The common baseline solution for the analysis would be, To summarize for each state, we can analyse:

- The number of deaths occurred in each state per year.
- The cause of death in each state per year.
- The states with highest number of deaths.
- Most frequently occurred types of deaths.
- And summarize all of them together to gain a deep insight of the above analysis.

Differences:

- Because we need to undertake analysis for each category and integrate them all together, baseline solutions would take a lot more time than the primary offered solution. For example, if we wanted to determine the states with the most deaths and their causes between 1900 and 2000, we'd have to first filter the states with the most deaths for each year, then filter them out depending on the cause of death, and finally add up all of the data.
- Instead of filtering several times and aggregating the results, we may use our proposed technique to apply various filters and acquire an overall picture.
- The primary solution would also include visualization graphs, which assist anyone viewing the graph in understanding it without the need for extra information.
- The baseline method can be applied to some extent when the data size is modest, but when the data size is large, often in the Terabytes or above, the baseline solution fails.
- Furthermore, baseline solutions do not allow us to classify data or use it for future predictions, whereas machine learning models allow us to use our data set to build a training model and implement a few algorithms, which we can then use for future predictions, which we cannot do with a baseline solution.
- We work on analysis to some extent with the baseline solution, but with the proposed use of data science, we may unleash a world of alternatives for performing analysis and coming up with a problem solution.

Big Data Systems and Tools The

tools that we used are:

PANDAS:

Pandas is a Python library that is used for working with data sets. It can be used for a variety of things.

Pandas is used to clean, explore, and manipulate the data.

We had raw data in the form of comma separated values (CSV), and since our dataset was not too vast (less than 10 GB), pandas was our first choice. We could observe the processing and output of any operations conducted on the data on the same screen.

SPARK:

Spark is distributed processing system used for large-scale data processing. It can also used for a variety of thing.

Irrespective data size it optimizes the query execution for fast queries.

As mentioned above we have data in the form of CSV, to perform the filtering (query operations) on raw data we have used spark. What we observed that runtime of data load and querying is faster than the pandas.

MATPLOTLIB:

It's a Python-based two-dimensional plotting package.

We can create a variety of graphs for data sets, including maps, bar plots, histograms, and pie charts.

It is one of Python's most powerful visualization libraries.

Our ultimate goal was to extract useful information from the data.

Matplotlib assisted in the creation of the required visualization, which anyone may analyse after viewing it.

SCIKIT LEARN:

One of the most useful libraries in Python is Scikit-learn. It includes many statistical modelling and machine learning algorithms such as classification, regression, and clustering, among others.

We implemented two of the machine learning algorithms: Decision Tree Classifier and K-Nearest Neighbour Classifier (KNN)

Tools tried but didn't work

Initially, we planned to use beautiful soup library while data scraping, we faced few issues in the middle of the process. Exceptions that we faced HTML parser, XML parses and Key Error while scraping data from the ODMP (Officer Down Memorial Page). However, because everything did not go as expected, we pulled the identical raw data from the GitHub repository.

At the beginning, we tried to use K means which is a clustering algorithm but after profound analysis on dataset we realized that the k-means algorithm doesn't suit our requirement. So, we went ahead with KNN and Decision trees classification algorithms to train our models and perform predictions.

Lessons Learned and High-level summary

We began by addressing the anomalies in the dataset by cleaning and reorganizing the data in a structured format and creating the necessary fields required for the analysis.

After applying visualization tools like matplotlib, we gained deep insights like:

- The deaths of police officers have gradually risen beginning from the year 1900 and peaked between the mid 1920's and 1940's and thereby a decline in the deaths was seen up until 1960 and from there it varied.
- The highest number of deaths were recorded during the period 1920-1940.
- Most number of police officers lost their lives due to Gunfire which is over 10,000 between the 19th and 20th century, whilst automobile accident, motorcycle accidents, heart attack and so on were the reasons for death where the count for number of deaths for each category was below 2000. The least deaths were occurred by structure collapse, poisoned and asphyxiation.
- The highest number of deaths were recorded in the following states CA with more than 1500 deaths, TX with close to 1400 deaths, NY with around 1300 deaths and so on. The least number of deaths were by CZ, MP, AS below 50 since 19th century.

- The top 10 states that recorded the highest number of deaths are CA followed by TX followed by NY, IL, PA, FL, OH, KY and GA.

We implemented two machine learning models: Decision Tree Classifier and K-Nearest Neighbours.

From the training data the feature fields are ‘year’ and ‘statecode’ where each state was assigned a number and mapped into a column ‘statecode’ from ‘state’ column respectively and the label was ‘death_reason’. We could predict the type of death that a state could incur based on the input features year and statecode. The label is the death_reason which predicts the type of death for the given input features.

The accuracy for Decision tree classifier was around 81% and for KNN the accuracy was 34% with k value equals to 1, and the accuracy rose to 51% with k value equals to 10.

Here is the prediction for the following states CO, MO and TN for the years 2022,2025 and 2030 and the reasons for their death could be as predicted below:

```
In [744]:
import sklearn
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
#from sklearn.metrics import acccuracy_score

#X=new2df.drop(columns=['person','dept','eow','cause','date','month','dept_name','state','death_reason'])
y=newdf['death_reason']
X_train, X_test, y_train, y_test =train_test_split(X,y,test_size=0.2)

model=DecisionTreeClassifier()
model.fit(X_train, y_train)
predictions=model.predict(X_test)

In [756]:
model.predict([[2022,6],[2025,27],[2030,48]])

/opt/conda/lib/python3.7/site-packages/sklearn/base.py:446: UserWarning: X does not have valid feature names, but DecisionTr
eeClassifier was fitted with feature names
  "X does not have valid feature names, but"

Out[756]:
array([' Gunfire', ' Vehicular assault', ' Stabbed'], dtype=object)
```