Exploratory Data Analysis of ***'Cause of Death'*** dataset

In this assignment, we'll work on the ***Cause of Death*** dataset. We'll import the required libraries and prepare the data for further processing. **The objective of this analysis is** to find patterns within the dataset to gain further understanding of the data to get better insights of the dataset. In this Dataset, we have Historical Data of different cause of deaths for all ages around the World for various countries.

Step 1 : Importing the required Libraries

Pandas, Numpy, Matplotlib, Pyplot, Seaborn are some of the requird libraries we imported.

Step 2 : Loading dataset in the python

We start by loading the data into a data frame using Pandas read.csv() function

Step 3 : Viewing Dimensions of the dataset

After loading the dataset we'll first look at the dimensions of the dataset, using df.shape function. The data has 6120 records and 34 features.

Step 4 : Viewing few rows of the dataset

With the help of the df.head () and df.tail () functions of the Pandas library, we can easily check out the first and last lines of your Data Frame, respectively.

At first glance, the data set is consist of the country names, years and Number of deaths caused due to different disease.

Step 5 : Listing the variables in dataset :

Next, we'll look at what columns are present in our data set. The data is larger (with 34 of features), getting an idea of features looking at the first few records seems to be not possible. We will list down all the columns headers with the help of df.columns function.

There are first three columns named as `'Country/Territory', 'Code', 'Year'`

And there are 31 columns contains number of deaths due different diseases as follows :

```
'Meningitis','Alzheimer's Disease and Other Dementias', 'Parkinson's Disease',
'Nutritional Deficiencies', 'Malaria', 'Drowning',
'Interpersonal Violence', 'Maternal Disorders', 'HIV/AIDS',
'Drug Use Disorders', 'Tuberculosis', 'Cardiovascular Diseases',
'Lower Respiratory Infections', 'Neonatal Disorders',
'Alcohol Use Disorders', 'Self-harm', 'Exposure to Forces of Nature',
'Diarrheal Diseases', 'Environmental Heat and Cold Exposure',
'Neoplasms', 'Conflict and Terrorism', 'Diabetes Mellitus',
'Chronic Kidney Disease', 'Poisonings', 'Protein-Energy Malnutrition',
'Road Injuries', 'Chronic Respiratory Diseases',
'Cirrhosis and Other Chronic Liver Diseases', 'Digestive Diseases',
'Fire, Heat, and Hot Substances', 'Acute Hepatitis'
```

Step 6 : Checking for any null values in dataset :

Looking at the missing values in each feature using df.isna().sum() function.

No feature has null values or missing values at all.

Step 7 : Viewing Statistical summary of the columns :

Statistical summary of the features can be useful in inspecting the feature distribution and anomalies, if any using df.describe()

Mean and Median are not close to each other for almost all the input features, which might be due to outliers.

Standard deviation is also comparatively high, and thus we can say that data is highly disperse.

Step 8 : Unique values in datasets :

nunique() function can be used to identify how many countries are included in data set, total number years for which the data is established.

204 Countries are included in the data set for each 30 years, from 1990 to 2019.

Step 9 : Feature Engineering

i. Introducing new column 'Total Deaths'

In order to get broader perspective, it is not possible to look at every single row individually. Instead, we will sum the total number of deaths happened due to any cause in that particular year in particular country.

We used df['Total Deaths'] =df.iloc[:,3:].sum(axis=1) query to sum up the data.

ii. Top 20 records based on total deaths

In the new dataset, we already introduced Total Deaths column recently. Based on descending order, we will observe which records comes at the top of the table.

We used df.sort_values function to redorder the dataset based data in Total Deaths column by='Total Deaths', and in descending sequence using ascending=False.

Following is the output :

Sr. No	Country/ Territory	Year	Total Deaths
1	China	2019	10442561
2	China	2018	10163943
3	China	2017	9978653
4	China	2016	9814213
5	China	2015	9591222
6	China	2014	9503904
7	China	2013	9411928
8	China	2011	9366974
9	China	2012	9364587
10	China	2010	9284664

Sr. No.	Country/Territory	Year	Total Deaths
11	China	2009	9074833
12	China	2005	8982702
13	China	2008	8972670
14	China	2004	8960684
15	India	2019	8812747
16	China	2006	8794396
17	China	2007	8755201
18	China	2003	8750361
19	India	2018	8698039
20	China	2002	8610956

Here we can see that, among the top 20 records there are only two countries that have witnessed the largest number of deaths due to all collective causes. There are only 2 records found for India, and rest belongs to China alone.

iii. Top 30 countries based on total deaths in 30 Years:

In earlier step, we have seen that China occupies most of the records in overall death numbers in last 30 years. Which have given idea to arrange the dataset based on the Countries based on their last 30 years total number of deaths.

To do this, we used groupby('Country/Territory').sum() function. Following are the results of the top 20 countries ordered by total deaths in 30 years.

Sr . No	Country/Territory	Total Deaths	Sr . No	Country/Territory	Total Deaths
1	**China**	265408106	**11**	**Bangladesh**	24803502
2	**India**	238158165	**12**	**Ukraine**	21245451
3	**United States**	71197802	**13**	**Ethiopia**	20880668
4	**Russia**	59591155	**14**	**Democratic Republic of Congo**	17446538
5	**Indonesia**	44046941	**15**	**United Kingdom**	17281600
6	**Nigeria**	43670014	**16**	**Italy**	16779302
7	**Pakistan**	38151878	**17**	**South Africa**	15807129
8	**Brazil**	32674112	**18**	**Mexico**	15720801
9	**Japan**	31922807	**19**	**France**	15093782
10	**Germany**	25559667	**20**	**Egypt**	14878359

By doing this, we have observed that topmost countries need to be taken in priority for data analysis.

iv. Years which have experienced Most deaths:

For checking the trend between years and total deaths in that year in any country, we will sort data based on the years. This can be done using groupby('Year').sum() and then sorting the data based on total death column data . Following are some major insights :

Sr. No.	Year	Total Deaths	Sr, No.	Year	Total Deaths
1	**2019**	54362920	**16**	**2004**	49330171
2	**2018**	53545244	**17**	**2003**	49123952
3	**2017**	52789758	**18**	**2002**	48897031
4	**2016**	52337435	**19**	**2001**	48385692
5	**2015**	51856393	**20**	**2000**	48050317
6	**2014**	51268375	**21**	**1999**	47652090
7	**2013**	50931550	**22**	**1998**	47066088
8	**2012**	50597654	**23**	**1997**	46672370
9	**2010**	50422775	**24**	**1996**	46320827
10	**2011**	50413303	**25**	**1994**	46182613
11	**2008**	50115740	**26**	**1995**	46177018
12	**2009**	49900666	**27**	**1993**	45185713
13	**2005**	49591909	**28**	**1992**	44459130
14	**2007**	49495216	**29**	**1991**	44059729
15	**2006**	49424521	**30**	**1990**	43518516

Number of total deaths are on increasing trend as years getting increased

v. Records of maximum deaths caused due to every cause :

There are 32 different causes present in data set. We will refine the data set in order to find which country and which year have experienced deaths due to every particular cause.

To do this, we will use idxmax() function to locate maximum value in every column, and then return the whole record corresponding to it.

Sr. No.	Cause Name	Death Count	Country	Year
1	Cardiovascular Diseases	4584273	China	2019
2	Neoplasms	2716551	China	2019
3	Chronic Respiratory Diseases	1366039	China	1994
4	Diarrheal Diseases	1119477	India	1992
5	Neonatal Disorders	852761	India	1990
6	Lower Respiratory Infections	690913	India	1990
7	Tuberculosis	657515	India	1992
8	Conflict and Terrorism	503532	Rwanda	1994
9	Digestive Diseases	464914	India	2019
10	Road Injuries	329237	China	2009

11	Alzheimer's Disease and Other Dementias	320715	China	2019
12	HIV/AIDS	305491	South Africa	2006
13	Malaria	280604	Nigeria	2008
14	Diabetes Mellitus	273089	India	2019
15	Cirrhosis and Other Chronic Liver Diseases	270037	India	2019
16	Nutritional Deficiencies	268223	India	1990
17	Chronic Kidney Disease	222922	India	2019
18	Exposure to Forces of Nature	222641	Haiti	2010
19	Self-harm	220357	China	1995
20	Protein-Energy Malnutrition	202241	India	1990
21	Drowning	153773	China	1990
22	Maternal Disorders	107929	India	1992
23	Meningitis	98358	India	1990
24	Parkinson's Disease	76990	China	2019
25	Interpersonal Violence	69640	Brazil	2016
26	Drug Use Disorders	65717	United States	2019
27	Acute Hepatitis	64305	India	1993
28	Alcohol Use Disorders	55200	Russia	2003
29	Poisonings	30883	China	2011
30	Environmental Heat and Cold Exposure	29048	Russia	1994
31	Fire, Heat, and Hot Substances	25876	India	2019

Step 9 : EDA for Top 2 countries in the list of total deaths

China have witnessed highest most death numbers in all time, which is then followed by India. So we will focus on it.

i. Creating separate data frames of these two countries :

It will be good idea to create independent data frames for China and India. We used df['Country/Territory'] == 'India' to select records of particular countries.

ii. Use of correlation heatmap to find insights:

On the basis of pearson correlation we plotted Heatmap, with the help of MatplotLib. It is observed that, there are high numbers of data columns have pearson correlation coefficient more than 0.90. So, it is advisable to use any one column out of each pair.

From China DataFrame, following 23 data columns are highly correlated :

```
'Acute Hepatitis','Alcohol Use Disorders', "Alzheimer's Disease and Other Deme
ntias",'Cardiovascular Diseases','Chronic Kidney Disease','Diabetes Mellitus',
'Diarrheal Diseases','Drowning','Environmental Heat and Cold Exposure','Fire,
Heat, and Hot Substances','HIV/AIDS','Interpersonal Violence','Lower Respiraty
Infections','Maternal Disorders','Neonatal Disorders','Neoplasms','Nutritional
Deficiencies',"Parkinson's Disease",'Poisonings','Protein-Energy Malnutrition'
```

```
'Self-harm','Total Deaths','Tuberculosis'
```

From India DataFrame, following 24 data columns are highly correlated :

```
'Acute Hepatitis','Alcohol Use Disorders','Cardiovascular Diseases',
'Chronic Kidney Disease','Chronic Respiratory Diseases',
'Cirrhosis and Other Chronic Liver Diseases','Diabetes Mellitus',
'Diarrheal Diseases','Digestive Diseases','Drowning','Drug Use Disorders',
'Lower Respiratory Infections','Malaria','Maternal Disorders','Meningitis',
'Neonatal Disorders','Neoplasms','Nutritional Deficiencies',
"Parkinson's Disease", 'Poisonings','Protein-Energy Malnutrition',
'Road Injuries','Total Deaths','Tuberculosis'
```

Identifying the highly correlated columns will help you to reduce the dimensions of the data to be train, which will be ultimately results in easy computation for ML algorithm. Because, two highly correlated columns can be replaced with any one out of it.

iii. Insights of the Year Vs Causes of Deaths in China:

1. China have witnessed sudden drop in Meningitis from 1990 to 2006, and remain near to flat till 2016.
2. Alzheimer's Disease and Parkinson's Disease is spreading rapidly, and death numbers are on increasing trend. It is at a all-time high death count.
3. Number of deaths due to Nutritional Deficiencies was on decreasing trend till 2007-08. But there is again increasing trend afterwards till 2017, and then remain flat.
4. Deaths due to Drowning are on down trend.
5. Graph of total deaths remain flat till 1998, and then increase till 2005. Again there is a drop till 2007, afterwards it is on upward trend.

iv. Insights of the Year Vs Causes of Deaths in India:

1. India have witnessed drop in Meningitis from 1990 to 2017, and remain near to flat till 2019
2. Marginal growth observe in Numbers of deaths due to Alzheimer's Disease and at all time in 2019.
3. Number of deaths due to Nutritional Deficiencies is on decreasing trend.
4. Graph of total deaths do not experiencing linearity at all, and shown wavy pattern. 2019 year experienced all time high death count.