Project 2 - **Investigate a Dataset**

Step1 – Choose your data set. Downloaded "All TB deaths" from the data

URl used : https://docs.google.com/document/d/e/2PACX-1vTlVmknRRnfy_4eTrjw5hYGaiQim5ctr9naaRd4V9du2B5bxpd8FEH3KtDgp8qVekw7Cj1GLk1IXdZi/pub?embedded=True

Step2 – Get organized: Cleansed Data

Renamed a row name from India to Bharat, Changed the column name data format from numeric to string

No duplicates, nulls are there in the data

Step3 – Analyze Data: Came up with questions

1.Top 10 countries in TB deaths

2.Bottom 10 countries with TB deaths

3.What is the death rate over the years for India

4.Average Deaths per country

5.Top5 Death rates by mean and the country with lowest TB deaths

6.Chane the name of the country India to Bharat

7.Average TB deaths in the year 2007

Step4-Share your findings: Draw Bar charts and Pie Chart to communicate my analysis

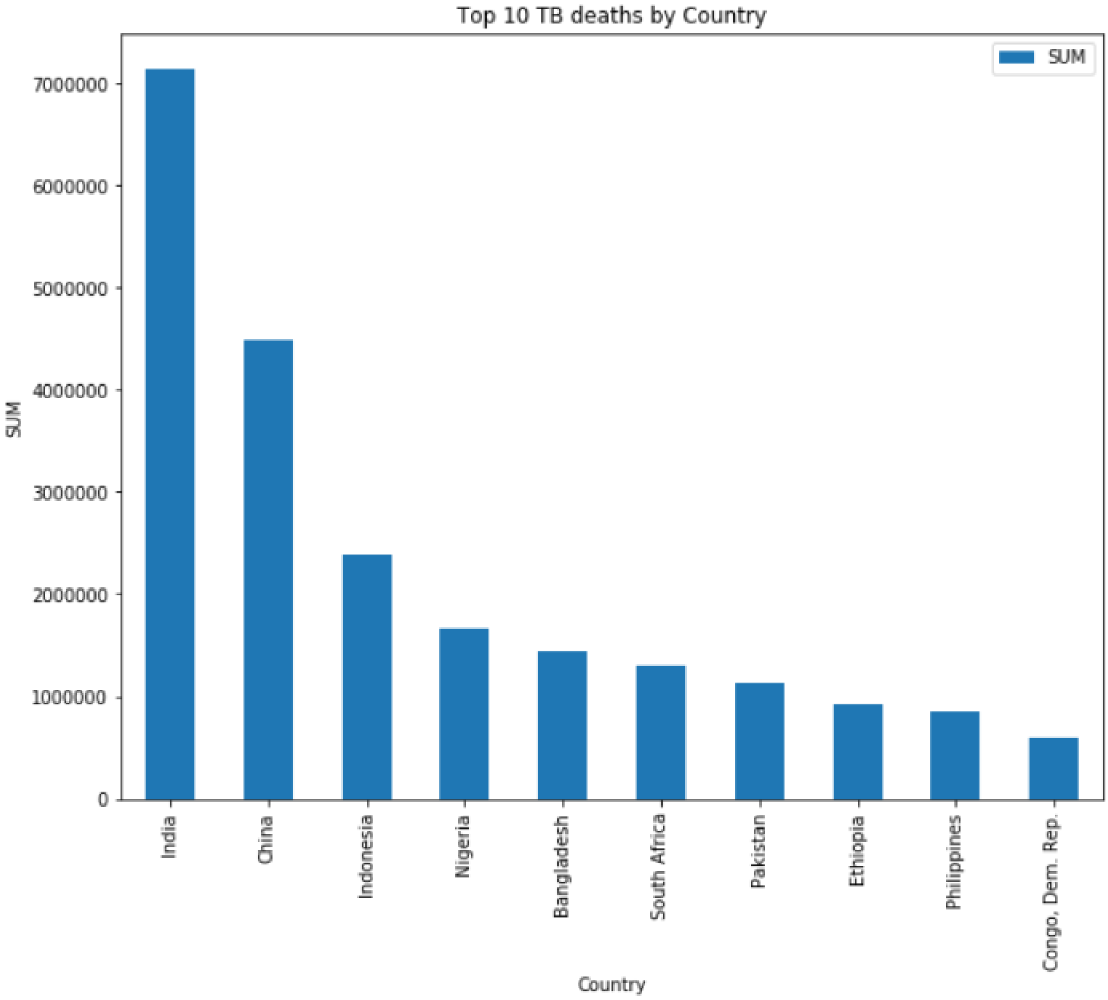
Conclusion:

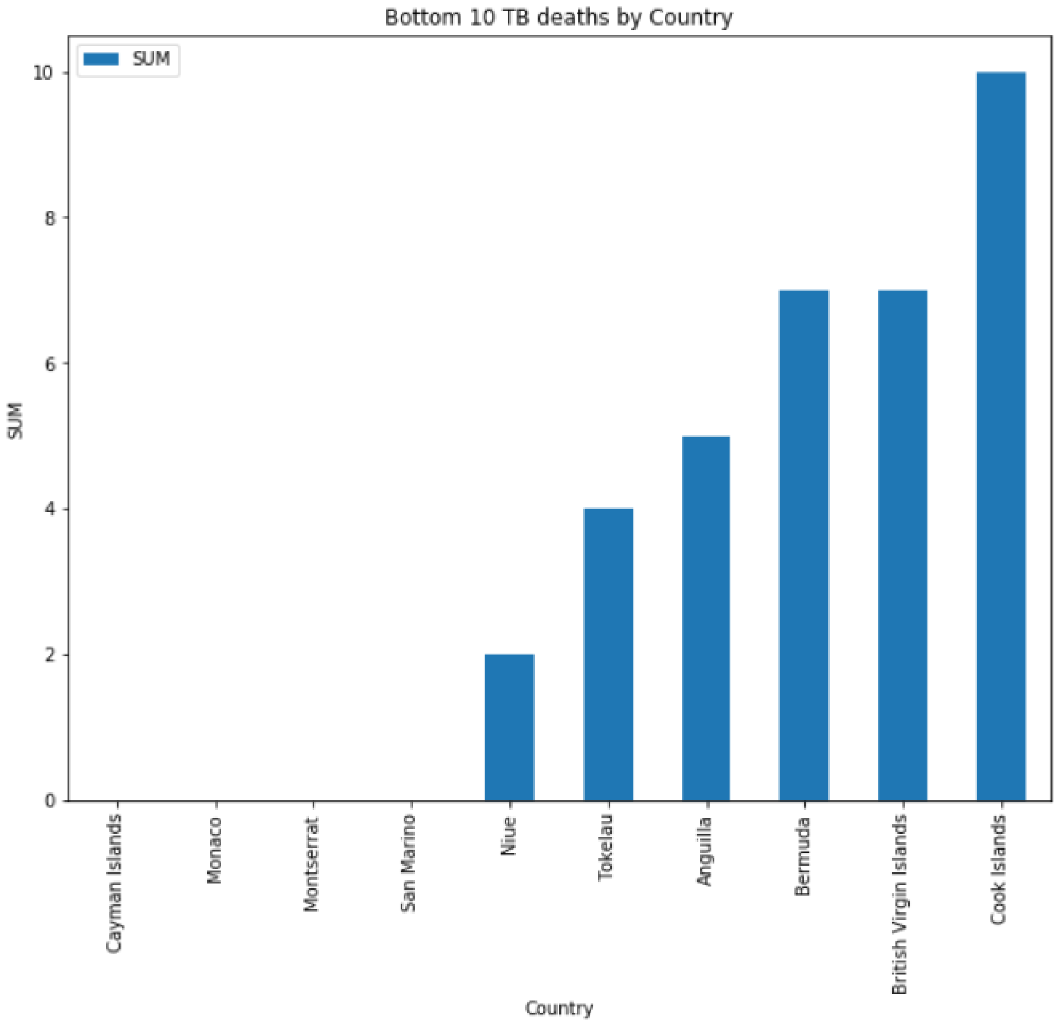
1. **India** is the country with largest number of TB deaths during the time period 1910 to 2007 with 41.7% of the total deaths, followed by China and Indonesia in the Secord and third places. This is based on the sum of deaths for all the years for a country
2. Cayman Islands is the country with lowest number of TB deaths during the time period 1910 to 2007. Before this we have Niue and Tokelau from the bottom
3. Avg. Death rate in the 2007 for all countries: **8554.599033816425**
4. Pie chart is used just to depict the contribution of countries from the top 5 list.
5. Histogram is used to describe the overall deaths on the particular years
6. Sample size is quite enough (17 years of data) and I am hoping it can be generalizable based on the numbers…

7. Data set only have information of deaths per year per country, this is the limitation we have. More details like Types of TBs, diagnosis details and areas much effected could have given more chances to improve Analysis
8. As we do not have nulls or duplicates in the data, the reports generated are good enough

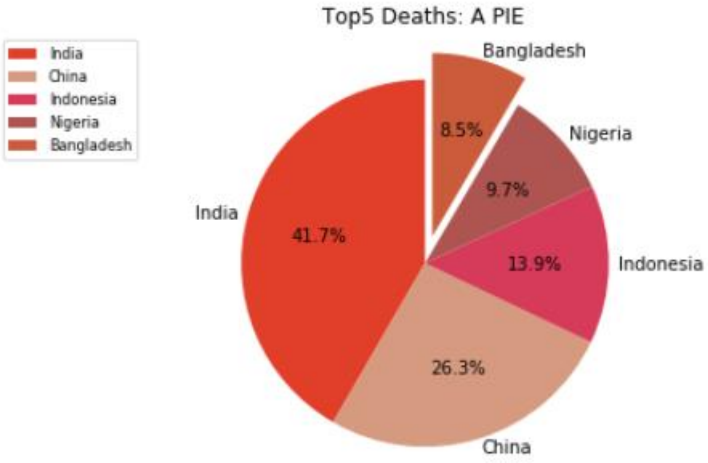
Given top10 and bottom 10 countries by total deaths

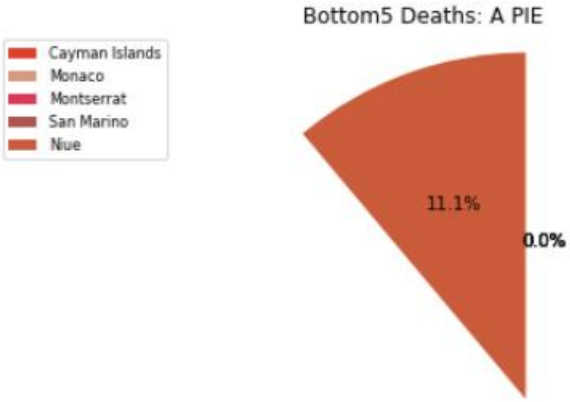
Bar Chart output


Top 10 TB deaths by Country
SUM
SUM
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
India
China
Indonesia
Nigeria
Bangladesh
South Africa
Pakistan
Ethiopia
Philippines
Congo, Dem. Rep.
Country

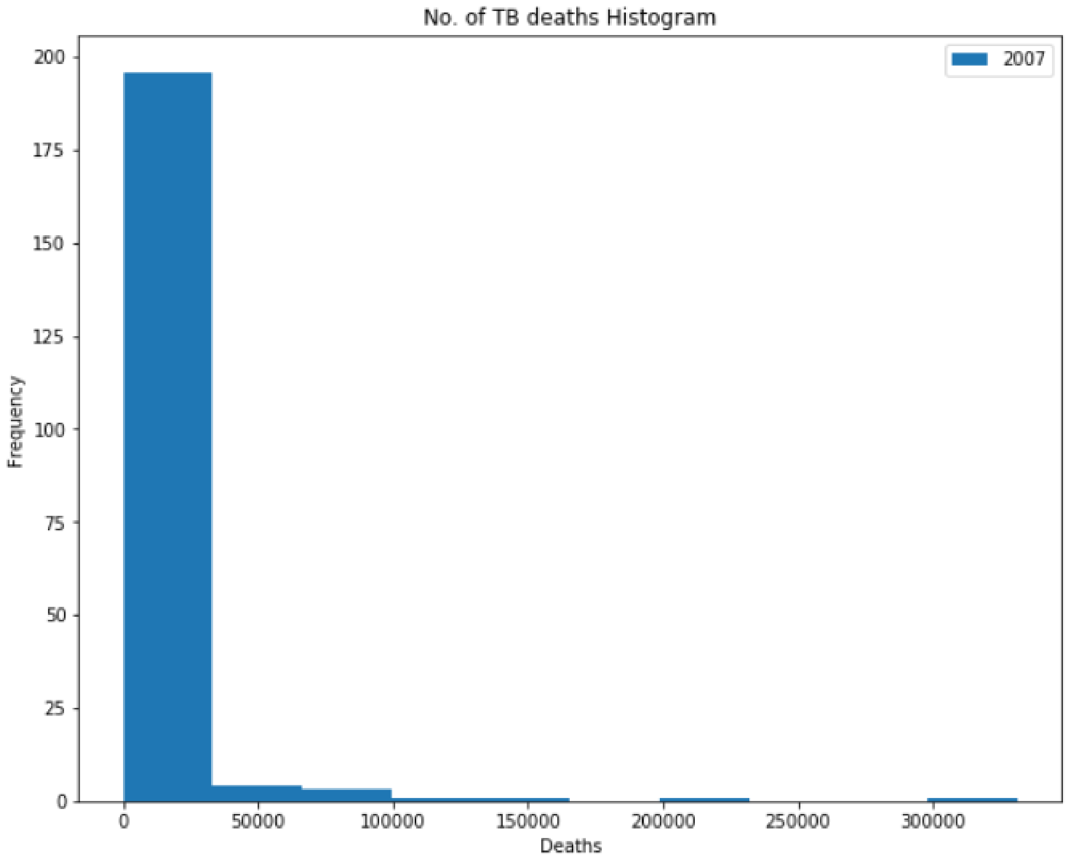


Pie Chart output


Top5 Deaths: A PIE
India
China
Indonesia
Nigeria
Bangladesh
Bangladesh
8.5%
Nigeria
9.7%
India
41.7%
13.9%
Indonesia
26.3%
China


Bottom5 Deaths: A PIE
Cayman Islands
Monaco
Montserrat
San Marino
Niue
11.1%
0.0%

Histogram



Code:

Data Cleansing

```
-------------------------

import pandas as pd

import numpy as np

import matplotlib as plt

df=pd.read_csv('All tb deaths.csv')

rows, columns = df.shape

## Print no. of Rows

print(rows)

##Print no. of columns

print(columns)

df.dtypes
df.apply(pd.to_numeric, errors='ignore')
df.columns = df.columns.astype(str)
## to clear spaces
df.columns=df.columns.str.strip()
##Know Data types
df.info()
##To display sample data
df.head()
## avg. death rate by country
df['MEAN'] = df.mean(axis=1)
## avg. death rate in the year 2007
AvgDeathTB2007=np.mean(df['2007'])
print(AvgDeathTB2007)
## Sum of deaths due to TB for Bharat
df['SUM'] = df[df.columns].sum(axis=1)
df['Country'] = df['Country'].replace({'India':'Bharat'})
df.loc[df['Country'] == 'Bharat']
## TB deaths by Top 10 countries
Top10=df.nlargest(10,"SUM")
##TB deaths by bottom 10 countries
Bottom10=df.nsmallest(10,"SUM")
```

```
# Create a bar chart for top 10 and Bottom10 coutries with TB deathsimport pandas as pd
---------------------------------------------------------------------------------------------------
import numpy as np
import matplotlib.pyplot as plt
df=pd.read_csv('All tb deaths.csv')
rows, columns = df.shape
```

```
print(rows)
print(columns)
df.dtypes
np.mean(df['2007'])
df['SUM'] = df[df.columns].sum(axis=1)
Top10=df.nlargest(10,"SUM")
Bottom10=df.nsmallest(10,"SUM")
Top10.plot(x='Country', y='SUM', kind='bar')
plt.show()
Bottom10.plot(x='Country', y='SUM', kind='bar')
plt.show()
```

Pie Chart Creation

```
-------------------------
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df=pd.read_csv('All tb deaths.csv')
df['MEAN'] = df.mean(axis=1)
df.head()
Top5_Mean=df.nlargest(5,"MEAN")
Top5_Mean['MEAN']

# Create a list of colors (from iWantHue)
colors = ["#E13F29", "#D69A80", "#D63B59", "#AE5552", "#CB5C3B"]
# Create a pie chart
plt.pie(
    # using data Mean TB deaths
```

```
  Top5_Mean['MEAN'],
  # with the labels being Country names
  labels=Top5_Mean['Country'],
  # with no shadows
  shadow=False,
  # with colors
  colors=colors,
  # with one slide exploded out
  explode=(0, 0, 0, 0, 0.15),
  # with the start angle at 90%
  startangle=90,
  # with the percent listed as a fraction
  autopct='%1.1f%%',
  )

# View the plot drop above
plt.axis('equal')

# View the plot
plt.tight_layout()
plt.show()
```

Histogram code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df=pd.read_csv('All tb deaths.csv')
rows, columns = df.shape
## Print no. of Rows
```

```
print(rows)
##Print no. of columns
print(columns)
df.describe()
df['1990'].hist(alpha=1,label='1990')
df['1998'].hist(alpha=1,label='1998')
df['2007'].hist(alpha=1,label='2007')
plt.legend();
plt.show()
```