UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

BSD2333 DATA WRANGLING GROUP PROJECT

LECTURER: MOHD KHAIRUL BAZLI BIN MOHD AZIZ

TITLE: WORLDWIDE DEATHS BY COUNTRY

GROUP NAME: PLOTLY

MATRIC ID	NAME	SECTION
SD22015	LIM JON VINCE	01G
SD22031	ANIS AQILAH BINTI MOHD ASRI	01G
SD22059	LOW ANN GIE	02G
SD22029	ALENA NG PEI YEEN	02G
SD22061	ANG MEI YING	02G

Table of Contents

1.0 Project Synopsis

1.1 Description of the selected project

The study of mortality rates over the years is crucial to understanding the overall health challenges faced by populations worldwide. It provides insight into the health status of some populations. In this project, we aim to analyse the cause of deaths by various risk factors globally, with the goal of identifying key patterns and trends that could help in reduce these risks.

The dataset used for this project is sourced from Kaggle. The idea of analysing global death by risk was chosen since the dataset providing a rich source of information encompassing a wide range of demographics and geographic locations. By examining the patterns and trends in mortality data, we seek to identify the leading cause of death and monitor mortality trends over historical data. Additionally, we seek to understand the impact of different risk factors on public health and provide insight that could be used not only by public health policies but also impact other sectors such as the insurance sector, economic planning sector, social services sector, and urban development.

Furthermore, our project aims to contribute to the broader field of public health research by providing actionable findings that can inform policy decisions and interventions. By understanding the leading causes of death and their underlying risk factors, policymakers and healthcare professionals can develop targeted strategies to improve public health outcomes and reduce mortality rates.

In summary, our project represents a collaborative effort to utilise the knowledge that we learn in class into real world hands-on project to address global cause of death. By combining diverse skill sets and expertise, we aim to generate valuable insight than can lead to positive changes and improve well-being of population worldwide.

1.2 Problem to be solved

The dataset is obtained from reliable source, Kaggle and it covers all geographical regions and time. The database used requires a wrangling process as there are massive outliers and null values and to smooth our analysis, we have dropped a few columns as the columns are not being utilized. We have set some parameters which will be considered outliers since there are numerous outliers. Furthermore, we have chosen a few causes of death in different category to identify significant risk factors that increase the mortality rates. We used Jupyter Notebook and Python skills to analyze trends and patterns in mortality rates. Additionally, we want to examine the underlying factors contributing to the leading cause of death and how these causes spike to mortality rates.

1.3 Question to be answered

1) What is the risk factor that contributes most significantly to the number of deaths?

2) What is the trend of number of deaths caused by various risk factors has changed over time?

3) Which is the country contributing the highest number of deaths caused by risk factors?

1.4 Objectives

1) To determine the risk factor that contributes most significantly to the number of deaths

2) To explore the trend of number of deaths caused by various risk factors changing over time

3) To identify the country that contributes to highest number of deaths caused by risk factors

1.5 Basic description of the data

No	Attribute	Data Types
1	Country	String
2	Year	Integer
3	Unsafe water source	Float
4	Unsafe sanitation	Float
5	No access to handwashing facility	Float
6	Household air pollution from solid fuels	Float
7	Non-exclusive breastfeeding	Float
8	Discontinued breastfeeding	Float
9	Child wasting	Float
10	Child stunting	Float
11	Low birth weight for gestation	Float
12	Secondhand smoke	Float
13	Alcohol use	Float
14	Drug use	Float
15	Diet low in fruits	Float
16	Diet low in vegetables	Float
17	Unsafe sex	Float
18	Low physical activity	Float
19	High fasting plasma glucose	Float
20	High total cholesterol	Float
21	High body-mass index	Float
22	High systolic blood pressure	Float
23	Smoking	Float
24	Iron deficiency	Float
25	Vitamin A deficiency	Float
26	Low bone mineral density	Float
27	Air pollution	Float
28	Outdoor air pollution	Float

29	Diet high in sodium	Float
30	Diet low in whole grains	Float
31	Diet low in nuts and seeds	Float

2.0 Packages Required

In this project, we are going to clean the data and develop interactive data visualizations using Jupyter Notebook. Before we start, there are a few libraries that need to be installed into Jupyter Notebook for later use. The list of libraries is as following:

1. Pandas

Pandas is used to import data from external files such as .csv, .xls or .sql into python dataframe.

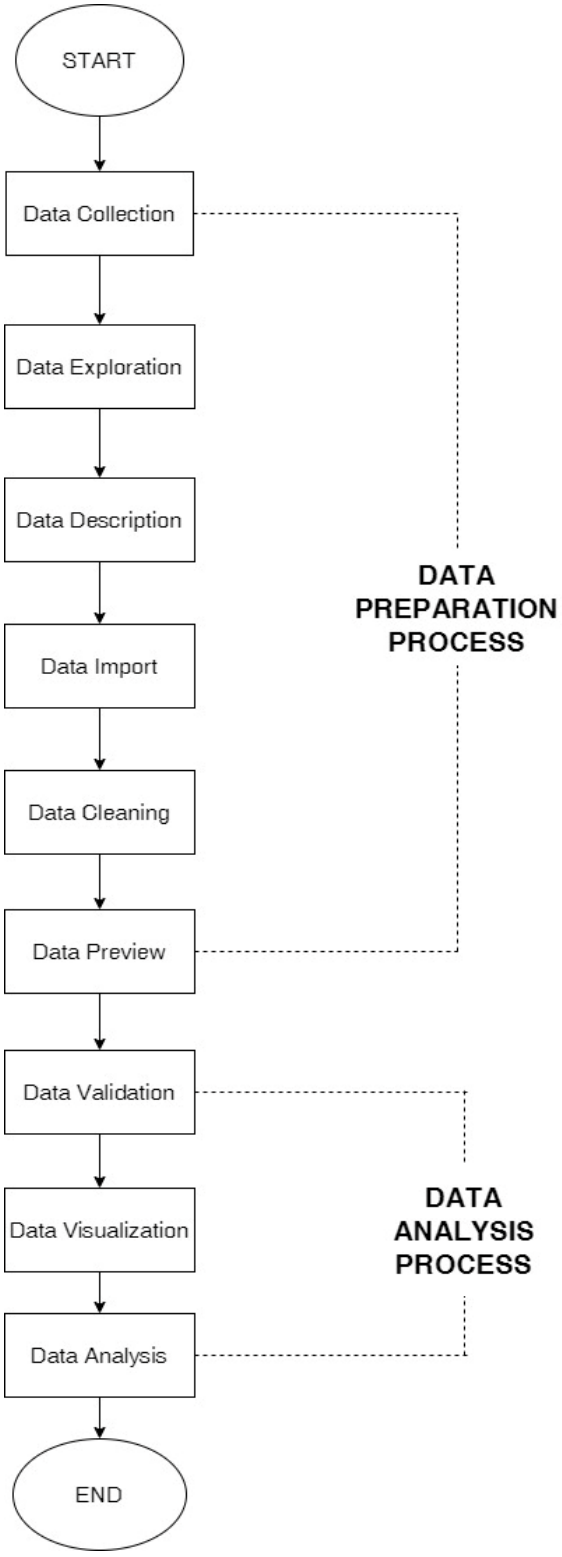
2. Matplotlib

Matplotlib is a comprehensive library for creating static, animated and interactive visualization. The matplotlib was originally developed by John D. Hunter in 2003. It provides customization options to control every aspect of the plot such as colour, label, line styles and others. In this project, the matplotlib library is used to draw the boxplot and histogram for identifying outliers.

3. Plotly

Plotly is an open-source library that can be used to create interactive charts and plots. It supports various types of plots like line charts, scatter plots, histograms, box plots and others. In this project, we used plotly.express to create choropleth map, treemap, sunburst chart, area chart and bubble plot. Plotly Express is a built-in part of the plotly library. It is easy to use, high-level interface to Plotly, which operates on a variety of types of data and produces easy-to-style figures. The plotly library allows users to explore the data by zooming in and out and panning across the data.

3.0 Data Preparation



We start this project by obtaining a dataset from Kaggle about causes of death. It has 200,539 data with 31 columns and 6469 rows. This data consists of one qualitative data which is country and 30 quantitative data such as “Unsafe Water Source”, “Unsafe Sanitation”, “Household Air Pollution”, “Smoking”, “Alcohol Use”, and others. In data exploration phrase, we explore our dataset by understanding the data structure, attributes and identifying the data types and relationships among variables. Then, we have listed all the variables included in the dataset and a brief description of what each variable represents. After that, we import the dataset from a CSV file into Jupyter Notebook and use the Pandas library to load the CSV files into data frame to be read. Once the data is imported, we can proceed to the next step which is data cleaning. After importing the raw data in Jupyter, we start the cleaning process by identifying the missing value, null value, and duplicate data. Besides, we also identify the outlier by using visualization such as boxplot and histogram. The outliers are cleaned by using flooring and capping method. Besides, we use the “head()” function to observe the top rows of the datasets. After the cleaning process, we use “isnull()” and “.duplicated()” functions to recheck for the missing value, null value, duplicate and outliers for the cleaned data. This process is to ensure that the data is consistent and accurate. In data visualisation phrase, we create various graphs to visualise the data and explore the relationship between different variables. Once the graph is created, we examine the patterns, trends, and interactions between the variables. After data visualisation, we analyse and extract meaningful insights from the created graph. Based on our observation, we summarise our findings in the interpretation section.

3.1 Data Import

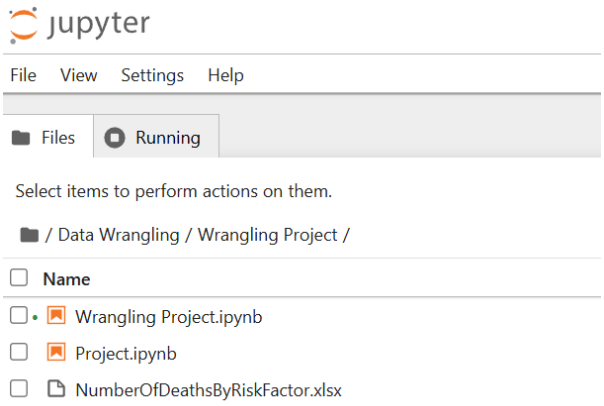


Figure 1: Import dataset into Jupyter

Firstly, we upload "NumberOfDeathsByRiskFactor.xlsx" excel file into Jupyter notebook to ensure the dataset can be read without copy the file path.

```
[1]: import pandas as pd

[2]: df = pd.read_excel("NumberOfDeathsByRiskFactor.xlsx")
```

Figure 2: Import Pandas library & Read the excel file

Next, we import "pandas" library and use "read_excel" function to import "NumberOfDeathsByRiskFactor" excel file. The "df" variable is assigned as data frame and used for storing the data.

```
[3]: df.head()
```

[3]:

	Country	Year	Unsafe water source	Unsafe sanitation	No access to handwashing facility	Household air pollution from solid fuels	Non-exclusive breastfeeding	Discontinued breastfeeding	Child wasting	Child stunting	...	High systolic blood pressure	Smoking	defi
0	Afghanistan	1990	7554.049543	5887.747628	5412.314513	22388.49723	3221.138842	156.097553	22778.84925	10408.43885	...	28183.98335	6393.667372	726.
1	Afghanistan	1991	7359.676749	5732.770160	5287.891103	22128.75821	3150.559597	151.539851	22292.69111	10271.97643	...	28435.39751	6429.253320	739.
2	Afghanistan	1992	7650.437822	5954.804987	5506.657363	22873.76879	3331.349048	156.609194	23102.19794	10618.87978	...	29173.61120	6561.054957	873.
3	Afghanistan	1993	10270.731380	7986.736613	7104.620351	25599.75628	4477.006100	206.834451	27902.66996	12260.09384	...	30074.76091	6731.972560	1040.
4	Afghanistan	1994	11409.177110	8863.010065	8051.515953	28013.16720	5102.622054	233.930571	32929.00593	14197.94796	...	30809.49117	6889.328118	1101.

5 rows × 31 columns

Figure 3: head() function

After that, we preview the top 5 rows of data by using “head()” function. Figure 3 shows the top 5 rows of data and we can see that the data from “NumberOfDeathsByRiskFactor” excel file is imported successfully.

3.2 Data Cleaning

```
[4]: df.shape

[4]: (6468, 31)
```

Figure 4: Check the number of rows and columns

Before starting the cleaning process, we use "shape" function to check the number of rows and columns of the data. Figure 4 shows that there are 6468 rows and 31 columns successfully uploaded to "df" data frame and no data is lost during importing process.

```
[5]: df.isna().sum()

[5]: Country                                    0
     Year                                       0
     Unsafe water source                        0
     Unsafe sanitation                          0
     No access to handwashing facility          0
     Household air pollution from solid fuels   0
     Non-exclusive breastfeeding                0
     Discontinued breastfeeding                 0
     Child wasting                              0
     Child stunting                             0
     Low birth weight for gestation             0
     Secondhand smoke                           0
     Alcohol use                                0
     Drug use                                   0
     Diet low in fruits                         0
     Diet low in vegetables                     0
     Unsafe sex                                 0
     Low physical activity                      0
     High fasting plasma glucose                0
     High total cholesterol                  4907
     High body-mass index                       0
     High systolic blood pressure               0
     Smoking                                    0
     Iron deficiency                            0
     Vitamin A deficiency                       0
     Low bone mineral density                   0
     Air pollution                              0
     Outdoor air pollution                      1
     Diet high in sodium                        0
     Diet low in whole grains                   0
     Diet low in nuts and seeds                 0
     dtype: int64
```

Figure 5: Check the missing value or null value

Then, we use "isna().sum()" function to check the total number of missing value or null value in each column. Based on Figure 5, there are 4907 null values in column "High total cholesterol" and 1 null value in column "Outdoor air pollution".

```
[6]: df.duplicated().sum()

[6]: 0
```

Figure 6: Check the number of duplicate data

Additionally, “duplicated().sum()” function is used to check the duplicated value in the data. Since the result shown in Figure 6 is equal to 0, it means no duplicated value exists inside the data.

```
[7]: df.drop('High total cholesterol', axis='columns', inplace=True)
     df
```

[7]:

	Country	Year	Unsafe water source	Unsafe sanitation	No access to handwashing facility	Household air pollution from solid fuels	Non-exclusive breastfeeding	Discontinued breastfeeding	Child wasting	Child stunting	...	High systolic blood pressure	Smoking
0	Afghanistan	1990	7554.049543	5887.747628	5412.314513	22388.497230	3221.138842	156.097553	22778.849250	10408.438850	...	28183.98335	6393.667372
1	Afghanistan	1991	7359.676749	5732.770160	5287.891103	22128.758210	3150.559597	151.539851	22292.691110	10271.976430	...	28435.39751	6429.253320
2	Afghanistan	1992	7650.437822	5954.804987	5506.657363	22873.768790	3331.349048	156.609194	23102.197940	10618.879780	...	29173.61120	6561.054957
3	Afghanistan	1993	10270.731380	7986.736613	7104.620351	25599.756280	4477.006100	206.834451	27902.669960	12260.093840	...	30074.76091	6731.972560
4	Afghanistan	1994	11409.177110	8863.010065	8051.515953	28013.167200	5102.622054	233.930571	32929.005930	14197.947960	...	30809.49117	6889.328118
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6463	Zimbabwe	2013	4254.282075	2977.649750	3913.210510	7613.561005	1037.968042	59.150493	7703.062474	1317.296056	...	11077.32708	9099.552194
6464	Zimbabwe	2014	4098.769691	2856.426187	3809.245683	7429.446352	972.886327	54.334796	7401.059382	1259.989023	...	11015.12167	8902.223776
6465	Zimbabwe	2015	3921.291358	2717.735794	3688.442102	7267.029297	912.248164	50.255551	7100.476546	1205.589945	...	11005.40982	8818.570004
6466	Zimbabwe	2016	3802.257512	2624.315858	3603.179799	7134.595677	875.706009	47.719473	6823.766727	1099.871279	...	11096.18244	8758.486720
6467	Zimbabwe	2017	3796.070615	2612.122560	3579.352078	6982.337249	866.902012	46.816760	6609.236886	1021.437703	...	11243.08932	8714.714332

6468 rows × 30 columns

Figure 7: Drop column “High total cholesterol”

Based on result from Figure 5, we decide use “.drop()” function to drop column “High total cholesterol”. The “inplace=True” means the dataset will be updated and the column will be permanently deleted. The column is removed because there are 75.88% null values in the column (4907 out of 6467 in rows) which is more than 50%. The huge number of null values in the column making it difficult to be used for visualisation and analysing. Figure 7 shows that we have 6468 rows and 30 columns after dropping the “High total cholesterol” column.

[8]:
```
df.dropna(inplace=True)
df.reset_index(drop=True, inplace=True)
df
```

[8]:

	Country	Year	Unsafe water source	Unsafe sanitation	No access to handwashing facility	Household air pollution from solid fuels	Non-exclusive breastfeeding	Discontinued breastfeeding	Child wasting	Child stunting	...	High systolic blood pressure	Smoking
0	Afghanistan	1990	7554.049543	5887.747628	5412.314513	22388.497230	3221.138842	156.097553	22778.849250	10408.438850	...	28183.98335	6393.667372
1	Afghanistan	1991	7359.676749	5732.770160	5287.891103	22128.758210	3150.559597	151.539851	22292.691110	10271.976430	...	28435.39751	6429.253320
2	Afghanistan	1992	7650.437822	5954.804987	5506.657363	22873.768790	3331.349048	156.609194	23102.197940	10618.879780	...	29173.61120	6561.054957
3	Afghanistan	1993	10270.731380	7986.736613	7104.620351	25599.756280	4477.006100	206.834451	27902.669960	12260.093840	...	30074.76091	6731.972560
4	Afghanistan	1994	11409.177110	8863.010065	8051.515953	28013.167200	5102.622054	233.930571	32929.005930	14197.947960	...	30809.49117	6889.328118
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6462	Zimbabwe	2013	4254.282075	2977.649750	3913.210510	7613.561005	1037.968042	59.150493	7703.062474	1317.296056	...	11077.32708	9099.552194
6463	Zimbabwe	2014	4098.769691	2856.426187	3809.245683	7429.446352	972.886327	54.334796	7401.059382	1259.989023	...	11015.12167	8902.223776
6464	Zimbabwe	2015	3921.291358	2717.735794	3688.442102	7267.029297	912.248164	50.255551	7100.476546	1205.589945	...	11005.40982	8818.570004
6465	Zimbabwe	2016	3802.257512	2624.315858	3603.179799	7134.595677	875.706009	47.719473	6823.766727	1099.871279	...	11096.18244	8758.486720
6466	Zimbabwe	2017	3796.070615	2612.122560	3579.352078	6982.337249	866.902012	46.816760	6609.236886	1021.437703	...	11243.08932	8714.714332

6467 rows × 30 columns

Figure 8: Drop the null value in a row

In addition, we also drop one row which contains a null value in column “Outdoor air pollution” by using “.dropna()” function. The “reset_index()” function is used to reset the index of data frame. Since out of 6468 columns only one row contains a null value, we decided to delete it as it would not affect the overall analysis. Based on Figure 8, the dataset remains 6467 rows and 30 columns after dropping the row containing null value.

```python
# Dictionary mapping country names to country codes
country_to_code = {

    'Afghanistan' : 'AFG',
    'Albania' : 'ALB',
    'Algeria' : 'DZA',
    'American Samoa' : 'ASM',
    'Andorra' : 'AND',
    'Angola' : 'AGO',
    'Antigua and Barbuda' : 'ATG',
    'Argentina': 'ARG',
    'Armenia' : 'ARM',
    'Australia' : 'AUS',
    'Austria': 'AUT',
    'Azerbaijan' : 'AZE',
    'Bahamas' : 'BHS',
    'Bahrain' : 'BHR',
    'Bangladesh' : 'BGD',
    'Barbados' :  'BRB',
    'Belarus' : 'BLR',
    'Belgium' : 'BEL',
    'Belize' : 'BLZ',
    'Benin' : 'BEN',
    'Bermuda' : 'BMU',
    'Bhutan' : 'BTN',
    'Bolivia': 'BOL',
    'Bosnia and Herzegovina' : 'BIH',
    'Botswana': 'BWA',
    'Brazil': 'BRA',
    'Brunei': 'BRN',
    'Bulgaria': 'BGR',
    'Burkina Faso': 'BFA',
    'Burundi': 'BDI',
    'Cambodia' : 'KHM',
    'Cameroon': 'CMR',
    'Canada': 'CAN',
    'Cape Verde' : 'CPV',
    'Central African Republic' : 'CAF',
    'Chad': 'TCD',
    'Chile' : 'CHL',
    'China' : 'CHN',
    'Colombia' : 'COL',
    'Comoros' : 'COM',
    'Congo': 'COG',
    'Costa Rica' : 'CRI',
    "Cote d'Ivoire" : 'CIV',
    'Croatia': 'HRV',
    'Cuba': 'CUB',
```

Figure 9

```python
    'Cyprus': 'CYP',
    'Czechia': 'CZE',
    'Democratic Republic of Congo': 'COD',
    'Denmark': 'DNK',
    'Djibouti': 'DJI',
    'Dominica': 'DMA',
    'Dominican Republic' : 'DOM',
    'Ecuador' : 'ECU',
    'Egypt' : 'EGY',
    'El Salvador' : 'SLV',
    'Equatorial Guinea': 'GNQ',
    'Eritrea' : 'ERI',
    'Estonia' : 'EST',
    'Eswatini' : 'SWZ',
    'Ethiopia' : 'ETH',
    'Fiji' : 'FJI',
    'Finland' : 'FIN',
    'France' : 'FRA',
    'Gabon' : 'GAB',
    'Gambia' : 'GMB',
    'Georgia' : 'GEO',
    'Germany' : 'DEU',
    'Ghana' : 'GHA',
    'Greece' : 'GRC',
    'Greenland': 'GRL',
    'Grenada' : 'GRD',
    'Guam' : 'GUM',
    'Guatemala' : 'GTM',
    'Guinea' : 'GIN',
    'Guinea-Bissau' : 'GNB',
    'Guyana' : 'GUY',
    'Haiti': 'HTI',
    'Honduras' : 'HND',
    'Hungary' : 'HUN',
    'Iceland' : 'ISL',
    'India' : 'IND',
    'Indonesia' : 'IDN',
    'Iran' : 'IRN',
    'Iraq' : 'IRQ',
    'Ireland': 'IRL',
    'Israel' : 'ISR',
    'Italy' : 'ITA',
    'Jamaica' : 'JAM',
    'Japan' : 'JPN',
    'Jordan' : 'JOR',
    'Kazakhstan' : 'KAZ',
    'Kenya' : 'KEN',
    'Kiribati' : 'KIR',
    'Kuwait' : 'KWT',
```

Figure 10

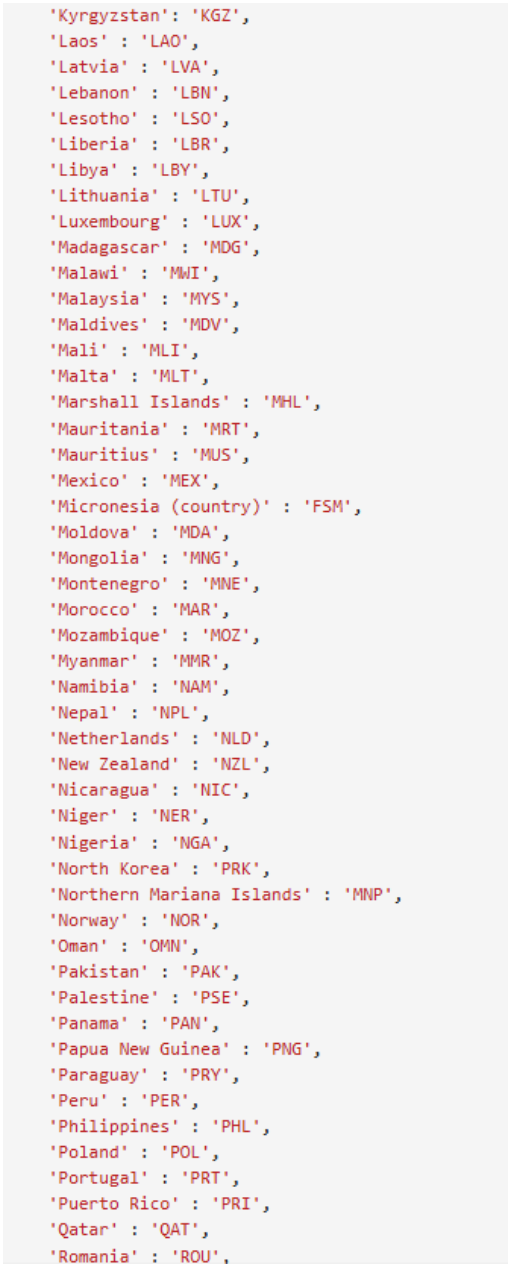


Figure 11

Figure 12

Figure 9, 10, 11 & 12: The column “Country Code” is added

```
    'Zimbabwe' : 'ZWE'

}

# Add a new column 'Country Code' based on the mapping
df['Country Code'] = df['Country'].map(country_to_code)
df
```

[9]:

ontinued tfeeding	Child wasting	Child stunting	...	Smoking	Iron deficiency	Vitamin A deficiency	Low bone mineral density	Air pollution	Outdoor air pollution	Diet high in sodium	Diet low in whole grains	Diet low in nuts and seeds	Country Code
6.097553	22778.849250	10408.438850	...	6393.667372	726.431294	9344.131952	374.844056	26598.006730	4383.83	2737.197934	11381.377350	7299.867330	AF
1.539851	22292.691110	10271.976430	...	6429.253320	739.245799	9330.182378	379.854237	26379.532220	4426.36	2741.184956	11487.832390	7386.764303	AF
6.609194	23102.197940	10618.879780	...	6561.054957	873.485341	9769.844533	388.130434	27263.127910	4568.91	2798.560245	11866.235570	7640.628526	AF
6.834451	27902.669960	12260.093840	...	6731.972560	1040.047422	11433.769490	405.577931	30495.561500	5080.29	2853.301679	12335.961680	7968.311853	AF
3.930571	32929.005930	14197.947960	...	6889.328118	1101.764645	12936.955860	415.349195	33323.161400	5499.23	2880.025765	12672.950190	8244.368430	AF
...	...	...	...	...	...	...	...	...	...	...	...	...	.
9.150493	7703.062474	1317.296056	...	9099.552194	382.544583	1130.714398	238.297856	9593.033931	2053.58	1018.389001	2687.636261	2409.930182	ZW
4.334796	7401.059382	1259.989023	...	8902.223776	353.386096	1094.267123	237.534426	9387.193480	2030.92	1016.407438	2654.381923	2399.261581	ZW
0.255551	7100.476546	1205.589945	...	8818.570004	332.355373	1068.810953	240.663191	9189.336702	1994.91	1019.971539	2635.950107	2398.525219	ZW
7.719473	6823.766727	1099.871279	...	8758.486720	319.692576	950.215259	244.719399	9092.577378	2030.88	1032.181216	2641.376815	2417.422521	ZW
6.816760	6609.236886	1021.437703	...	8714.714332	310.669683	882.986667	250.269185	9020.941349	2112.19	1049.402363	2664.132572	2449.546229	ZW

Figure 13: Output shows column “Country Code” is added

```
[10]: df.shape

[10]: (6467, 31)
```

Figure 14: Check number of rows and columns

Based on Figure 9-13, we added a new column into the “df” data frame named “Country Code” by using “.map()” function to ensure that the mapping related graph is able to read the country code and display easily. The country code column is added because the visualization for the choropleth map requires each country’s code to plot the country's location. Based on Figure 14, we have 6467 rows and 31 columns after adding column “Country Code”.

```python
# Dictionary mapping country names to country codes
country_to_region = {

    'Afghanistan' : 'Asia',
    'Albania' : 'Europe',
    'Algeria' : 'Africa',
    'American Samoa' : 'Oceania',
    'Andorra' : 'Europe',
    'Angola' : 'Africa',
    'Antigua and Barbuda' : 'North America',
    'Argentina' : 'South America',
    'Armenia' : 'Asia',
    'Australia' : 'Oceania',
    'Austria' : 'Europe',
    'Azerbaijan' : 'Asia',
    'Bahamas' : 'North America',
    'Bahrain' : 'Asia',
    'Bangladesh' : 'Asia',
    'Barbados' :  'North America',
    'Belarus' : 'Europe',
    'Belgium' : 'Europe',
    'Belize' : 'North America',
    'Benin' : 'Africa',
    'Bermuda' : 'North America',
    'Bhutan' : 'Asia',
    'Bolivia': 'North America',
    'Bosnia and Herzegovina' : 'Europe',
    'Botswana': 'Africa',
    'Brazil': 'South America',
    'Brunei': 'Asia',
    'Bulgaria': 'Europe',
    'Burkina Faso': 'Africa',
    'Burundi': 'Africa',
    'Cambodia' : 'Asia',
    'Cameroon': 'Africa',
    'Canada': 'North America',
    'Cape Verde' : 'Africa',
    'Central African Republic' : 'Africa',
    'Chad': 'Africa',
    'Chile': 'South America',
    'China' : 'Asia',
    'Colombia' : 'Asia',
    'Comoros' : 'Africa',
    'Congo': 'Africa',
    'Costa Rica' : 'North America',
    "Cote d'Ivoire" : 'Africa',
    'Croatia': 'Europe',
    'Cuba': 'North America',
```

Figure 15

```python
    'Cyprus': 'Europe',
    'Czechia': 'Europe',
    'Democratic Republic of Congo': 'Africa',
    'Denmark': 'Europe',
    'Djibouti': 'Africa',
    'Dominica': 'North America',
    'Dominican Republic' : 'North America',
    'Ecuador' : 'South America',
    'Egypt' : 'Africa',
    'El Salvador' : 'North America',
    'Equatorial Guinea': 'Africa',
    'Eritrea' : 'Africa',
    'Estonia' : 'Europe',
    'Eswatini' : 'Africa',
    'Ethiopia' : 'Africa',
    'Fiji' : 'Oceania',
    'Finland' : 'Europe',
    'France' : 'Europe',
    'Gabon' : 'Africa',
    'Gambia' : 'Africa',
    'Georgia' : 'Asia',
    'Germany' : 'Europe',
    'Ghana' : 'Africa',
    'Greece' : 'Europe',
    'Greenland': 'Europe',
    'Grenada' : 'North America',
    'Guam' : 'Oceania',
    'Guatemala' : 'North America',
    'Guinea' : 'Africa',
    'Guinea-Bissau' : 'Africa',
    'Guyana' : 'South America',
    'Haiti': 'North America',
    'Honduras' : 'North America',
    'Hungary' : 'Europe',
    'Iceland' : 'Europe',
    'India' : 'Asia',
    'Indonesia' : 'Asia',
    'Iran' : 'Asia',
    'Iraq' : 'Asia',
    'Ireland': 'Europe',
    'Israel' : 'Asia',
    'Italy' : 'Europe',
    'Jamaica' : 'North America',
    'Japan' : 'Asia',
    'Jordan' : 'Asia',
    'Kazakhstan' : 'Asia',
    'Kenya' : 'Africa',
    'Kiribati' : 'Oceania',
    'Kuwait' : 'Asia',
```

Figure 16

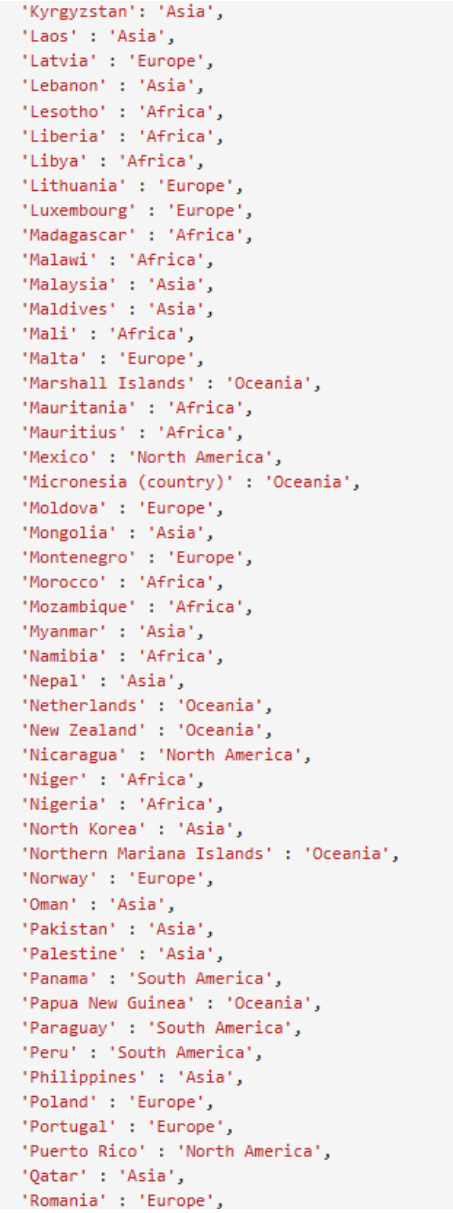


Figure 17

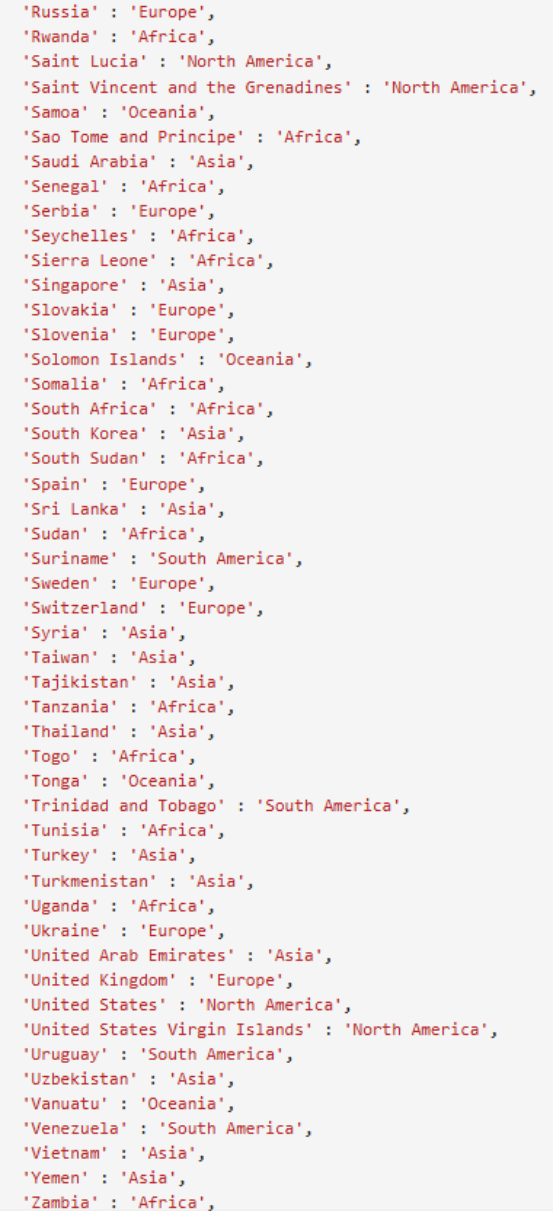


Figure 18

Figure 15,16,17&18: The column “Region” is added

```
    'Zimbabwe' : 'Africa'

}

# Add a new column 'Country Code' based on the mapping
df['Region'] = df['Country'].map(country_to_region)
df
```

[11]:

Discontinued breastfeeding	Child wasting	Child stunting	...	Iron deficiency	Vitamin A deficiency	Low bone mineral density	Air pollution	Outdoor air pollution	Diet high in sodium	Diet low in whole grains	Diet low in nuts and seeds	Country Code	Region
156.097553	22778.849250	10408.438850	...	726.431294	9344.131952	374.844056	26598.006730	4383.83	2737.197934	11381.377350	7299.867330	AFG	Asia
151.539851	22292.691110	10271.976430	...	739.245799	9330.182378	379.854237	26379.532220	4426.36	2741.184956	11487.832390	7386.764303	AFG	Asia
156.609194	23102.197940	10618.879780	...	873.485341	9769.844533	388.130434	27263.127910	4568.91	2798.560245	11866.235570	7640.628526	AFG	Asia
206.834451	27902.669960	12260.093840	...	1040.047422	11433.769490	405.577931	30495.561500	5080.29	2853.301679	12335.961680	7968.311853	AFG	Asia
233.930571	32929.005930	14197.947960	...	1101.764645	12936.955860	415.349195	33323.161400	5499.23	2880.025765	12672.950190	8244.368430	AFG	Asia
...	...	...	...	...	...	...	...	...	...	...	...	...	...
59.150493	7703.062474	1317.296056	...	382.544583	1130.714398	238.297856	9593.033931	2053.58	1018.389001	2687.636261	2409.930182	ZWE	Africa
54.334796	7401.059382	1259.989023	...	353.386096	1094.267123	237.534426	9387.193480	2030.92	1016.407438	2654.381923	2399.261581	ZWE	Africa
50.255551	7100.476546	1205.589945	...	332.355373	1068.810953	240.663191	9189.336702	1994.91	1019.971539	2635.950107	2398.525219	ZWE	Africa
47.719473	6823.766727	1099.871279	...	319.692576	950.215259	244.719399	9092.577378	2030.88	1032.181216	2641.376815	2417.422521	ZWE	Africa
46.816760	6609.236886	1021.437703	...	310.669683	882.986667	250.269185	9020.941349	2112.19	1049.402363	2664.132572	2449.546229	ZWE	Africa

Figure 19: Output shows column "Region" is added

```
[12]: df.shape

[12]: (6467, 32)
```

Figure 20: Check the number of rows and columns

Based on Figure 15-19, we also add a new column into the "df" data frame named "Region" by using the ".map()" function to group each country to the corresponding regions. Since the total number of countries in the dataset is more than 200, it is crucial to separate the countries by region to analyze the dataset easily during visualization section. Based on Figure 20, it shows that our dataset has 6467 rows and 32 columns after adding column "Region".

```
[13]: df.isna().sum()
```

```
[13]: Country                                 0
      Year                                    0
      Unsafe water source                     0
      Unsafe sanitation                       0
      No access to handwashing facility       0
      Household air pollution from solid fuels  0
      Non-exclusive breastfeeding             0
      Discontinued breastfeeding              0
      Child wasting                           0
      Child stunting                          0
      Low birth weight for gestation          0
      Secondhand smoke                        0
      Alcohol use                             0
      Drug use                                0
      Diet low in fruits                      0
      Diet low in vegetables                  0
      Unsafe sex                              0
      Low physical activity                   0
      High fasting plasma glucose             0
      High body-mass index                    0
      High systolic blood pressure            0
      Smoking                                 0
      Iron deficiency                         0
      Vitamin A deficiency                    0
      Low bone mineral density                0
      Air pollution                           0
      Outdoor air pollution                   0
      Diet high in sodium                     0
      Diet low in whole grains                0
      Diet low in nuts and seeds              0
      Country Code                         1063
      Region                               1063
      dtype: int64
```

Figure 21: Check the missing value or null value

```
[14]: df.dropna(subset = ['Region'], inplace=True)
      df.reset_index(drop=True, inplace=True)
      df
```

[14]:

	Country	Year	Unsafe water source	Unsafe sanitation	No access to handwashing facility	Household air pollution from solid fuels	Non-exclusive breastfeeding	Discontinued breastfeeding	Child wasting	Child stunting	...	Iron deficiency	Vitamin A deficienc
0	Afghanistan	1990	7554.049543	5887.747628	5412.314513	22388.497230	3221.138842	156.097553	22778.849250	10408.438850	...	726.431294	9344.13195
1	Afghanistan	1991	7359.676749	5732.770160	5287.891103	22128.758210	3150.559597	151.539851	22292.691110	10271.976430	...	739.245799	9330.18237
2	Afghanistan	1992	7650.437822	5954.804987	5506.657363	22873.768790	3331.349048	156.609194	23102.197940	10618.879780	...	873.485341	9769.84453
3	Afghanistan	1993	10270.731380	7986.736613	7104.620351	25599.756280	4477.006100	206.834451	27902.669960	12260.093840	...	1040.047422	11433.76949
4	Afghanistan	1994	11409.177110	8863.010065	8051.515953	28013.167200	5102.622054	233.930571	32929.005930	14197.947960	...	1101.764645	12936.95586
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5399	Zimbabwe	2013	4254.282075	2977.649750	3913.210510	7613.561005	1037.968042	59.150493	7703.062474	1317.296056	...	382.544583	1130.71439
5400	Zimbabwe	2014	4098.769691	2856.426187	3809.245683	7429.446352	972.886327	54.334796	7401.059382	1259.989023	...	353.386096	1094.26712
5401	Zimbabwe	2015	3921.291358	2717.735794	3688.442102	7267.029297	912.248164	50.255551	7100.476546	1205.589945	...	332.355373	1068.81095
5402	Zimbabwe	2016	3802.257512	2624.315858	3603.179799	7134.595677	875.706009	47.719473	6823.766727	1099.871279	...	319.692576	950.21525
5403	Zimbabwe	2017	3796.070615	2612.122560	3579.352078	6982.337249	866.902012	46.816760	6609.236886	1021.437703	...	310.669683	882.98666

5404 rows × 32 columns

Figure 22: Drop the null value in columns "Region" and "Country Code"

Based on Figure 21, we have detected 1063 null values in both columns "Country Code" and "Region". This is due to inaccurate country data which cannot be recognized as country, and it cannot be grouped by regions. Therefore, we used "dropna()" function to drop the rows. Then, we also used "reset_index()" function to reset again the index of data frame. Based on Figure 22, we remain 5404 rows and 32 columns after deleting the null values.

1. Andean Latin America	20. North Africa and Middle East
2. Australasia, Caribbean	21. North America
3. Central Asia	22. North Macedonia
4. Central Europe	23. Europe
5. Central Europe, Eastern Europe, Central Asia	24. Northern Ireland
6. Central Latin America	25. Oceania
7. Central Sub-Saharan Africa	26. Scotland
8. East Asia	27. South Asia
9. Eastern Europe	28. Southeast Asia
10. Eastern Sub-Saharan Africa	29. Southeast Asia, East Asia, and Oceania
11. England	30. Southern Latin America
12. High SDI	31. Southern Sub-Saharan Africa
13. High-income	32. Sub-Saharan Africa
14. High-income Asia Pacific	33. Timor
15. High-middle SDI	34. Tropical Latin America
16. Latin America and Caribbean	35. Wales
17. Low SDI	36. Western Europe
18. Low-middle SDI	37. Western Sub-Saharan Africa
19. Middle SDI	38. World

Table 1: List of countries that deleted

Table 1 shows the list of countries that were deleted. We delete the countries because some countries' names are invalid and some are inaccurate, as the country name is a region.

```
[16]: df.isna().sum()

[16]: Country                                  0
      Year                                     0
      Unsafe water source                      0
      Unsafe sanitation                        0
      No access to handwashing facility        0
      Household air pollution from solid fuels 0
      Non-exclusive breastfeeding              0
      Discontinued breastfeeding               0
      Child wasting                            0
      Child stunting                           0
      Low birth weight for gestation           0
      Secondhand smoke                         0
      Alcohol use                              0
      Drug use                                 0
      Diet low in fruits                       0
      Diet low in vegetables                   0
      Unsafe sex                               0
      Low physical activity                    0
      High fasting plasma glucose              0
      High body-mass index                     0
      High systolic blood pressure             0
      Smoking                                  0
      Iron deficiency                          0
      Vitamin A deficiency                     0
      Low bone mineral density                 0
      Air pollution                            0
      Outdoor air pollution                    0
      Diet high in sodium                      0
      Diet low in whole grains                 0
      Diet low in nuts and seeds               0
      Country Code                             0
      Region                                   0
      dtype: int64
```

Figure 23: Recheck the missing value and null value

Based on Figure 23, we recheck again the null values inside the dataset using “isna().sum()” function to ensure that all null values are totally clean at the end of the cleaning process.

```
[15]: df.shape

[15]: (5404, 32)
```

Figure 24: Recheck the number of row and column after the cleaning process

Based on Figure 24, it shows that our dataset has 5404 rows and 32 columns after the cleaning process is completed.

3.2.1 Outliers

In this section, we use boxplot, histogram and interquartile range to identify the outliers of data for eight different variables. The eight variables include unsafe water source, unsafe sanitation, drug use, low physical activity, smoking, air pollution, outdoor air pollution and diet low in nuts and seeds.

```
plt.subplot(2, 2, 1)
plt.hist(df['Unsafe water source'], bins=30, edgecolor='k', alpha=0.7)
plt.title('Histogram of Unsafe water source')
plt.xlabel('Unsafe water source')
plt.ylabel('Frequency')
```

Figure 25: Outliers Identification using Histogram

```
plt.subplot(2, 2, 3)
plt.boxplot(df['Unsafe water source'], vert=False)
plt.title('Boxplot of Unsafe water source')
plt.xlabel('Unsafe water source')
```

Figure 26: Outliers Identification using Boxplot

Based on figures 25 and 26, the histogram and boxplot are plotted to identify the outliers in column “Unsafe water source”. The “bins=30” means the range of data is divided into 30 equal parts. The “edgecolor=k” indicates that the edges of bars in the histogram are set to black colour. The “alpha=0.7” is used to control the transparency of the bars in the histogram. For boxplot, the “vert=False” means the boxplot is placed horizontally.

```
Q1 = df['Unsafe water source'].quantile(0.25)
Q3 = df['Unsafe water source'].quantile(0.75)

IQR = Q3 - Q1
IQR

lower = np.where(df['Unsafe water source'] <= (Q1 - 1.5 * IQR))
upper = np.where(df['Unsafe water source'] >= (Q3 + 1.5 * IQR))

print("Lower Outlier: ", lower)
print("Upper Outlier: ", upper)
```

Figure 27: Outliers Identification using Interquartile Range

Figure 27 shows the use of the interquartile range method to identify the outliers. Firstly, it calculates the 25th percentile and 75th percentile. Then, the interquartile range is calculated. Besides, the location for the lower boundary and upper boundary of outliers are calculated and displayed.

```
Lower Outlier:  (array([], dtype=int64),)
Upper Outlier:  (array([   3,    4,    5,    6,    7,    8,    9,   10,
          14,   15,   16,   17,  140,  141,  142,  143,  144,  145,  146,
         147,  148,  149,  150,  151,  152,  153,  154,  155,  156,  157,
         158,  159,  160,  161,  162,  163,  164,  165,  166,  167,  392,
         393,  394,  395,  396,  397,  398,  399,  400,  401,  402,  403,
         404,  405,  406,  407,  408,  409,  410,  411,  412,  413,  414,
         415,  416,  417,  418,  419,  700,  701,  702,  703,  704,  705,
         706,  707,  708,  709,  710,  711,  712,  713,  784,  785,  786,
         787,  788,  789,  790,  791,  792,  793,  794,  795,  796,  797,
         798,  799,  800,  801,  802,  803,  804,  805,  806,  807,  808,
         809,  810,  811,  868,  869,  870,  871,  872,  873,  874,  875,
         876,  877,  878,  879,  880,  881,  882,  883,  884,  885,  886,
         887,  888,  889,  890,  891,  892,  893,  894,  895,  980,  981,
         982,  983,  984,  985,  986,  987,  988,  989,  990,  991,  992,
         993,  994,  995,  996,  997,  998,  999, 1000, 1001, 1002, 1003,
        1004, 1005, 1006, 1007, 1036, 1037, 1038, 1039, 1040, 1041, 1042,
        1043, 1044, 1045, 1046, 1047, 1048, 1049, 1050, 1051, 1052, 1053,
        1054, 1176, 1177, 1178, 1179, 1180, 1181, 1182, 1183, 1184, 1185,
        1186, 1187, 1188, 1189, 1190, 1191, 1192, 1193, 1194, 1195, 1196,
        1197, 1198, 1199, 1200, 1201, 1202, 1203, 1316, 1317, 1318, 1319,
        1320, 1321, 1322, 1323, 1324, 1325, 1326, 1327, 1328, 1329, 1330,
        1331, 1332, 1333, 1334, 1335, 1336, 1337, 1338, 1339, 1340, 1341,
        1342, 1343, 1484, 1485, 1486, 1487, 1488, 1489, 1490, 1491, 1492,
        1493, 1494, 1495, 1496, 1652, 1653, 1654, 1655, 1656, 1657, 1658,
        1659, 1660, 1661, 1662, 1663, 1664, 1665, 1666, 1667, 1668, 1669,
        1670, 1671, 1672, 1673, 1674, 1675, 1676, 1677, 1678, 1679, 1876,
        1877, 1878, 1879, 1880, 1881, 1882, 1883, 1884, 1885, 1886, 1887,
        2016, 2017, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053,
        2054, 2055, 2128, 2129, 2130, 2131, 2240, 2241, 2242, 2243, 2244,
        2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255,
        2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266,
        2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277,
        2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288,
        2289, 2290, 2291, 2292, 2293, 2294, 2295, 2548, 2549, 2550, 2551,
        2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562,
        2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573,
        2574, 2575, 2884, 2885, 2886, 2887, 2888, 2889, 2890, 2891, 2892,
        2893, 2894, 2895, 2896, 2897, 2898, 2899, 2900, 2901, 2902, 2903,
```

Figure 28: Output for Lower and Upper Outliers

Figure 28 shows the lower outliers and upper outliers. It does not exist any lower outliers. However, there are many upper outliers. This also indicates that the distribution of unsafe water source is right skewed.

```
Q1 = df['Unsafe water source'].quantile(0.25)
Q3 = df['Unsafe water source'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

filtered_df = df[(df['Unsafe water source'] >= lower_bound) & (df['Unsafe water source'] <= upper_bound)]
```

Figure 29: Process of cleaning outliers

Based on figure 29, the outliers are removed using flooring and capping methods. The flooring indicates the lower values whereas the capping indicates the higher values. The ".quantile(0.25)" is used to calculate the 25th percentile while the ".quantile(0.75)" is used to calculate the 75th percentile. Then, the interquartile range (IQR) is calculated. The lower boundary and upper boundary are calculated to determine which data points are considered outliers. If the data points below the lower boundary or above the upper boundary, it is considered as outliers.

```
plt.subplot(2, 2, 2)
plt.hist(filtered_df['Unsafe water source'], bins=30, edgecolor='k', alpha=0.7)
plt.title('Histogram of Unsafe water source (Outliers Removed)')
plt.xlabel('Unsafe water source')
plt.ylabel('Frequency')
```

Figure 30: The histogram is plotted after removing the outliers

```
plt.subplot(2, 2, 4)
plt.boxplot(filtered_df['Unsafe water source'], vert=False)
plt.title('Boxplot of Unsafe water source (Outliers Removed)')
plt.xlabel('Unsafe water source')
```

Figure 31: The boxplot is plotted after removing the outliers

Based on figures 30 and 31, the histogram and boxplot are plotted after removing the outliers in columns "Unsafe water source".

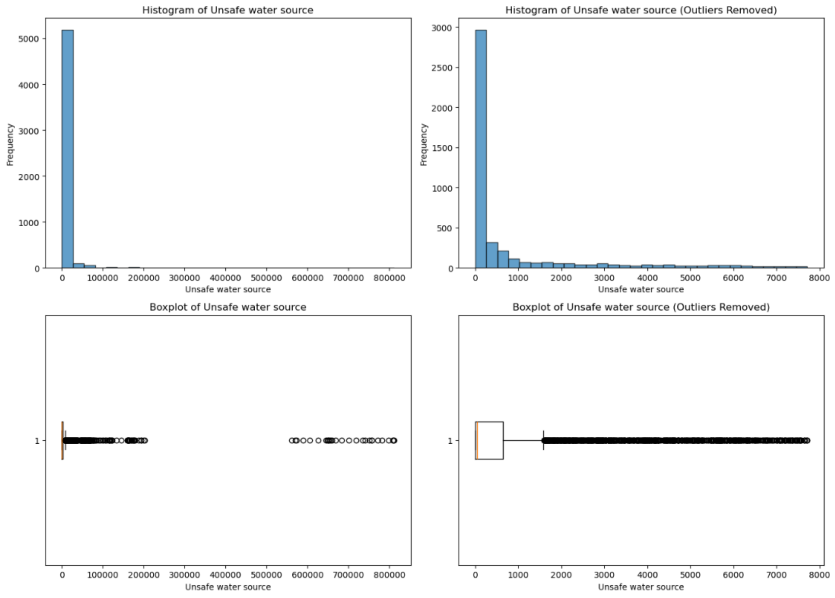


Figure 32: Identifying and cleaning outliers for column "Unsafe water source"

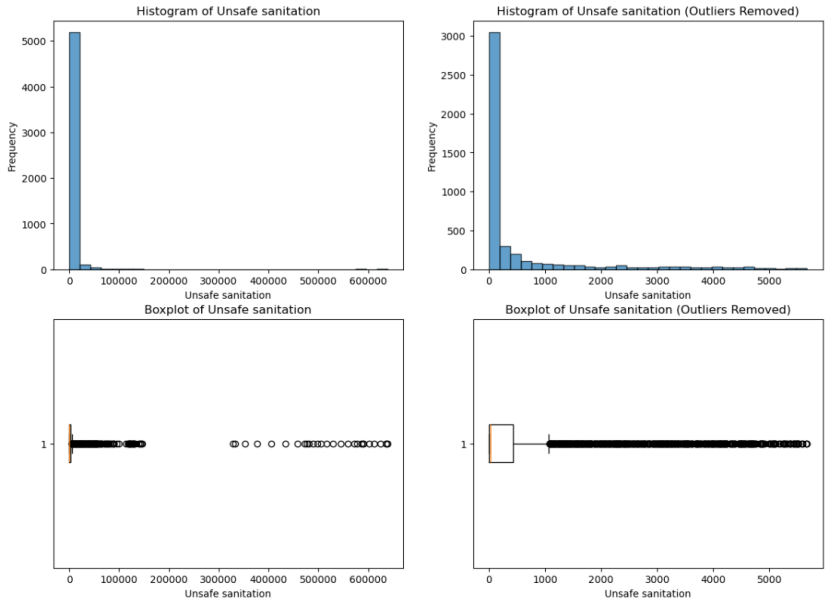


Figure 33: Identifying and cleaning outliers for column "Unsafe sanitation"

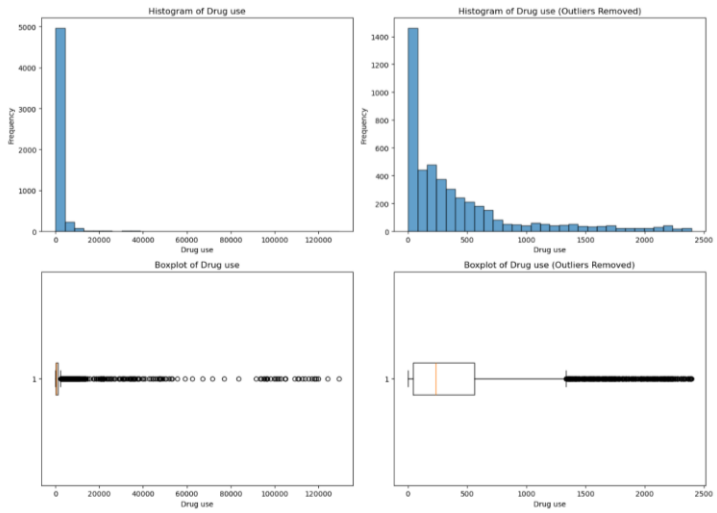


Figure 34: Identifying and cleaning outliers for column "Drug use"

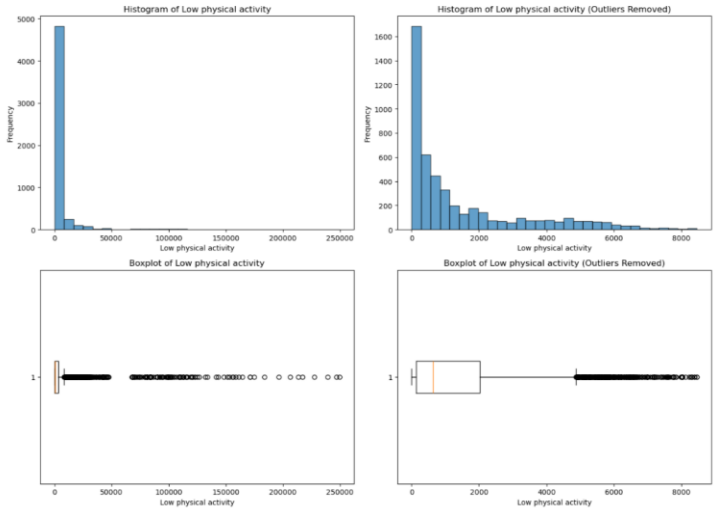


Figure 35: Identifying and cleaning outliers for column "Low physical activity"

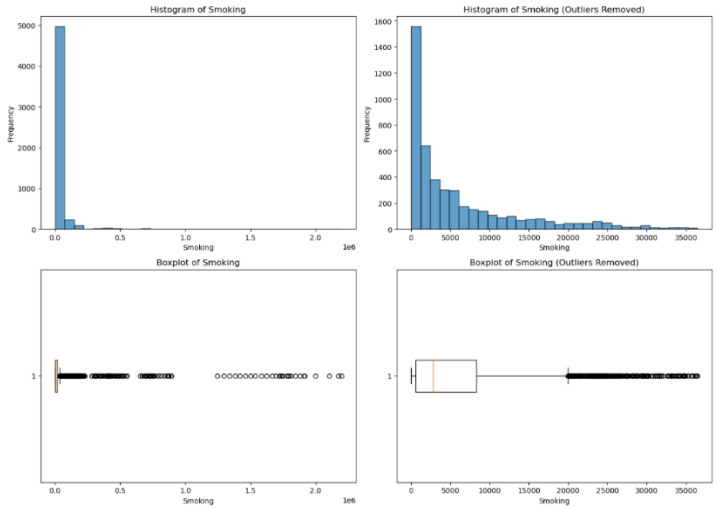


Figure 36: Identifying and cleaning outliers for column "Smoking"

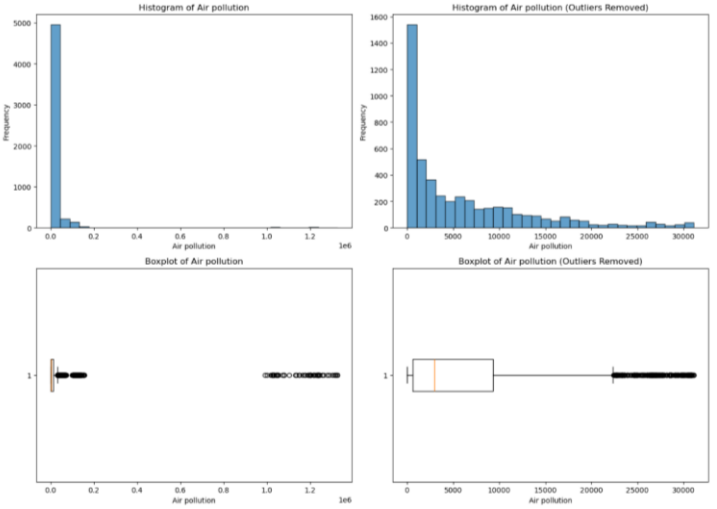


Figure 37: Identifying and cleaning outliers for column "Air pollution"

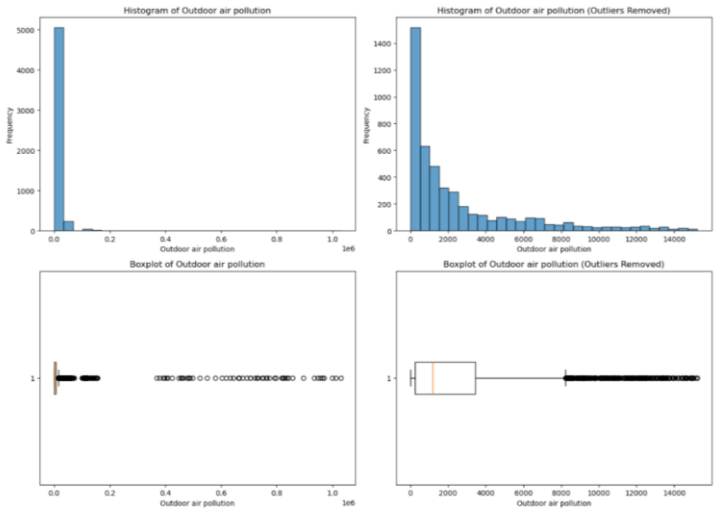


Figure 38: Identifying and cleaning outliers for column "Outdoor air pollution"

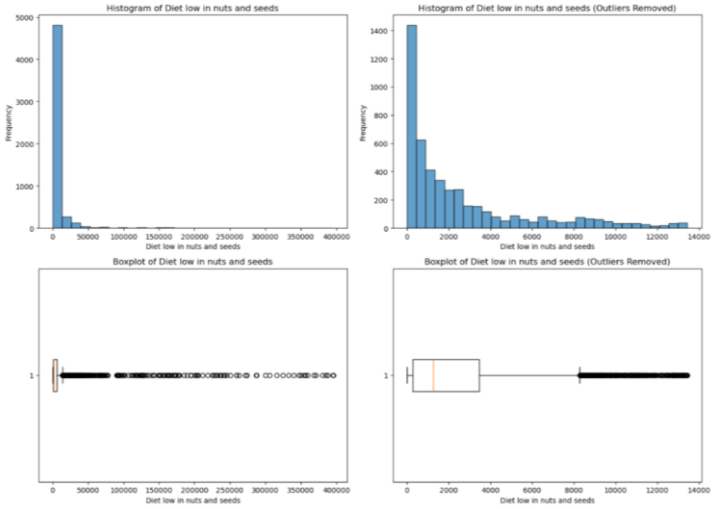


Figure 39: Identifying and cleaning outliers for column "Diet low in nuts and seeds"

The figures 32, 33, 34, 35, 36, 37, 38 and 39 show the before and after removing several outliers using boxplot and histogram. However, we will consider all the outliers in this analysis. This is because after removing the outliers, several countries will disappear as they have extreme values. Besides, in this project we only focus on observing the relationship between the variables instead of predicting the outcome. Therefore, we do not remove the outliers in the analysis.

3.3 Data Preview

After the data cleaning process, we show the top 5 rows and last 5 rows in the data frame to preview the structure of data.

```
In [57]: df.reset_index(drop=True).head(5)
```

Out[57]:

	Country	Year	Unsafe water source	Unsafe sanitation	No access to handwashing facility	Household air pollution from solid fuels	Non-exclusive breastfeeding	Discontinued breastfeeding	Child wasting	Child stunting	...	Iron deficiency	Vitamin deficier
0	Afghanistan	1990	7554.049543	5887.747628	5412.314513	22388.49723	3221.138842	156.097553	22778.84925	10408.43885	...	726.431294	9344.1319
1	Afghanistan	1991	7359.676749	5732.770160	5287.891103	22128.75821	3150.559597	151.539851	22292.69111	10271.97643	...	739.245799	9330.1823
2	Afghanistan	1992	7650.437822	5954.804987	5506.657363	22873.76879	3331.349048	156.609194	23102.19794	10618.87978	...	873.485341	9769.8445
3	Afghanistan	1993	10270.731380	7986.736613	7104.620351	25599.75628	4477.006100	206.834451	27902.66996	12260.09384	...	1040.047422	11433.7694
4	Afghanistan	1994	11409.177110	8863.010065	8051.515953	28013.16720	5102.622054	233.930571	32929.00593	14197.94796	...	1101.764645	12936.9558

5 rows × 32 columns

Figure 40: Display top 5 rows

```
In [58]: df.reset_index(drop=True).tail(5)
```

Out[58]:

	Country	Year	Unsafe water source	Unsafe sanitation	No access to handwashing facility	Household air pollution from solid fuels	Non-exclusive breastfeeding	Discontinued breastfeeding	Child wasting	Child stunting	...	Iron deficiency	Vitamin deficienc
5371	Zimbabwe	2013	4254.282075	2977.649750	3913.210510	7613.561005	1037.968042	59.150493	7703.062474	1317.296056	...	382.544583	1130.71439
5372	Zimbabwe	2014	4098.769691	2856.426187	3809.245683	7429.446352	972.886327	54.334796	7401.059382	1259.989023	...	353.386096	1094.26712
5373	Zimbabwe	2015	3921.291358	2717.735794	3688.442102	7267.029297	912.248164	50.255551	7100.476546	1205.589945	...	332.355373	1068.81095
5374	Zimbabwe	2016	3802.257512	2624.315858	3603.179799	7134.595677	875.706009	47.719473	6823.766727	1099.871279	...	319.692576	950.21525
5375	Zimbabwe	2017	3796.070615	2612.122560	3579.352078	6982.337249	866.902012	46.816760	6609.236886	1021.437703	...	310.669683	882.98666

5 rows × 32 columns

Figure 41: Display last 5 rows

3.4 Data Description

No	Attribute	Data Description
1	Country	The country where data was collected
2	Year	The year when the data was collected
3	Unsafe water source	Drinking or using contaminated water for daily activities
4	Unsafe sanitation	Lack of proper sanitation facilities
5	No access to handwashing facility	No access to handwashing with soap and water
6	Household air pollution from solid fuels	Air pollution from burning solid fuels in households
7	Non-exclusive breastfeeding	Babies fed substances other than breastmilk during first 6 months
8	Discontinued breastfeeding	Stopping breastfeeding earlier than recommended
9	Child wasting	Children under 5 years old underweight for their height, often indicating acute malnutrition
10	Child stunting	Children under 5 years old shorter than expected for their age
11	Low birth weight for gestation	Babies born lighter than expected for their gestational age
12	Secondhand smoke	Exposure to tobacco smoke from others
13	Alcohol use	Consumption of alcohol
14	Drug use	Consumption of substances that can alter body functions
15	Diet low in fruits	Diet lacking fruit
16	Diet low in vegetables	Diet lacking vegetables
17	Unsafe sex	Sexual activities without protection
18	Low physical activity	Not enough exercise

19	High fasting plasma glucose	High blood sugar levels after fasting
20	High total cholesterol	High cholesterol levels in the blood
21	High body-mass index	High body-mass index, indicating potential obesity
22	High systolic blood pressure	High blood pressure
23	Smoking	smoke cigarettes
24	Iron deficiency	Not enough iron in the diet
25	Vitamin A deficiency	Not enough vitamin A
26	Low bone mineral density	Person’s bone has less mineral content
27	Air pollution	Presence of harmful substances that are present in the air
28	Outdoor air pollution	Presence of harmful particles that are present in the air outside
29	Diet high in sodium	Consume food that contain a lot of salt
30	Diet low in whole grains	Consume insufficient amounts of whole grains
31	Diet low in nuts and seeds	Consume insufficient amounts of nuts and seeds

4.0 Exploratory Data Analysis

4.1 Bubble Plot: Causes of Death due to 'Drug Use' across the World

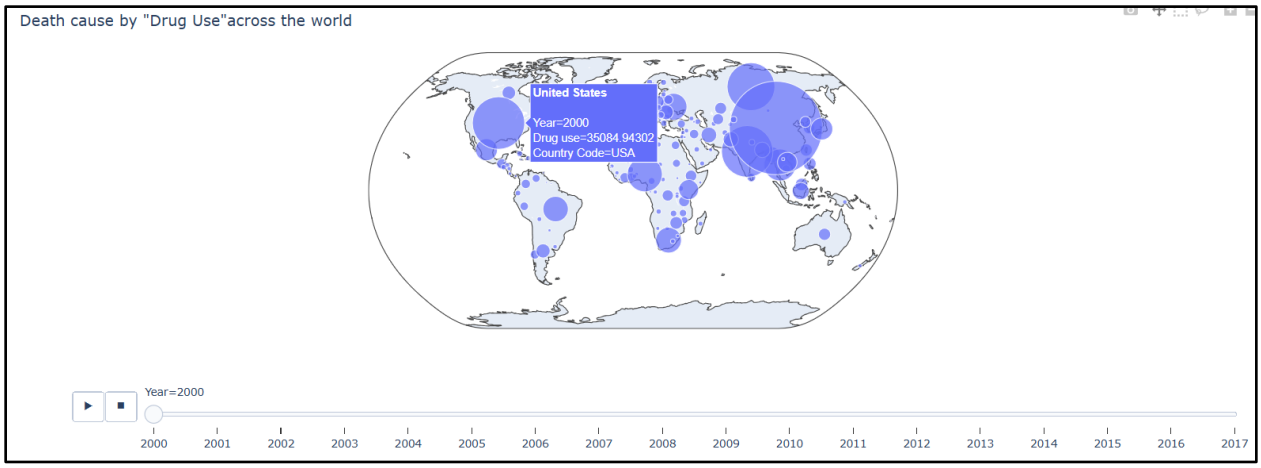


Figure 42: Causes of Death due to 'Drug Use' across the World in 2000

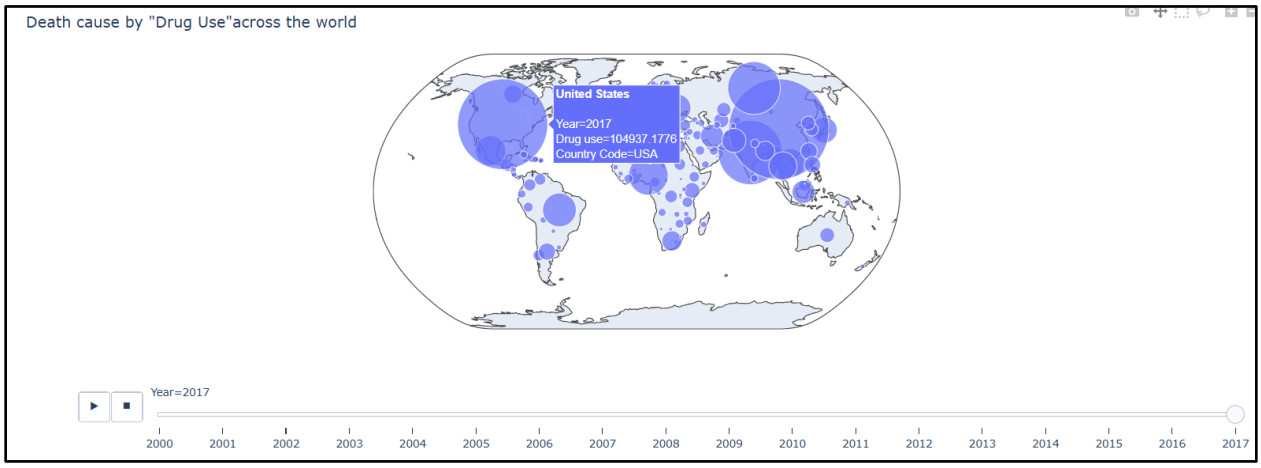


Figure 43: Causes of Death due to 'Drug Use' across the World in 2017

The visualization above is used to analyze the cause of death due to drug use across the nation between year 2000 and 2017. The size of bubble represents the amount of death that associated with drug use in different regions. Larger bubbles indicate a higher number of deaths. In contrast the smaller bubble, indicate a lower amount of death due to drug use. The slider positioned under the globe consists of the year from 2000 until 2017. It allows users to choose the specific year they want to observe. In figure 42 and 43, we can observe that amount of death related to drug use in United State increase aggressively from 35,084 cases in year 2000 to 10,4937 cases

in year 2017. This significant increase may be due to several factors including drug availability, lack of awareness from government organizations and may be due to weak drug policies.

In addition to the increasing number of death related drug, the visualization above also shows that China consistently had the highest number of deaths cases from year 2000 to 2017. The number of deaths due to drug use in China increase from 11,0660 cases in year 2000 to 129,306 cases in year 2017. This data shows a slight increase each year without any significant differences. The factors that contributing to this slightly increase might be due to China's large population, which leads to a higher total number of cases even if the rate per person is lower. Other than that, the increase in drug-related deaths can be influenced by a country's drug laws and policies. Strict laws might scare people from seeking help, lead to more overdoses and contributed to death.

By examining the changes over time and across different regions, related organization such as policymakers, health professionals, and the academician can identify the countries with the highest need for intervention and allocate resources to them more effectively. This insight highlights the urgency of addressing drug use and implementing comprehensive public health strategies. For instance, improving access to addiction treatment and increasing the prevention programs to spread awareness.

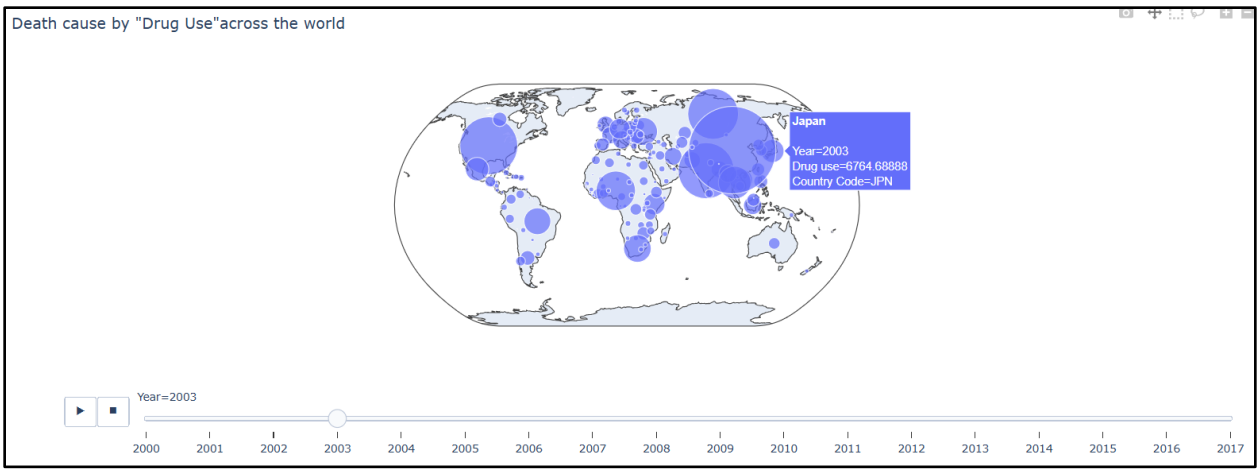


Figure 44: Hover function is used to show information for specific data points.

As shown in figure 44, the visualization provides an interactive layer of information with its hover function. When the cursor is moved over a specific location on the map, it displays detailed information about that location. This includes the name of the country, the year selected, the number of drug-related deaths, and the ISO code for that country. For example, hovering over Japan in the year 2003 would reveal details such as the country name (JAPAN), the year (2003), the number of drug-related deaths (6764.68888), and the country code (JPN). This interactivity enables users to easily access specific data points and gain deeper insights into the drug-related death statistics for different countries and years.

4.2 Bubble Plot: Risk of Death due to Smoking in Asia

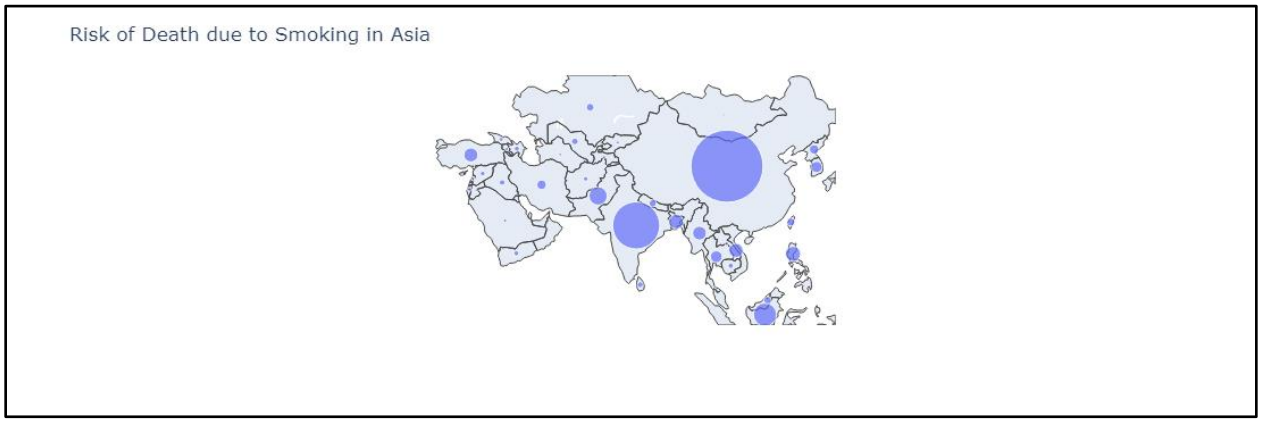


Figure 45: Risk of Death Due to Smoking in Asia

The bubble plot visualization above is used to analyse the risk of death due to smoking focusing on Asia Region. It highlights various countries with circles of different sizes. The size of the circle represents the relative risk of death due to smoking in that country. Larger circles indicate a higher relative risk of death due to smoking. This is often caused by factors such as high smoking prevalence, weak tobacco control laws, and limited access to healthcare and smoking cessation programs. Country with large size of bubbles highlights the necessity for targeted public health interventions such as implementing stricter regulations and enhancing access to smoking cessation programs. While on the other hands, low-risk countries, represented by smaller circles, offer valuable insights into successful strategies for reducing smoking-related death such as cultural norms that discourage smoking, and comprehensive smoking cessation support, which can be replicated in other regions.

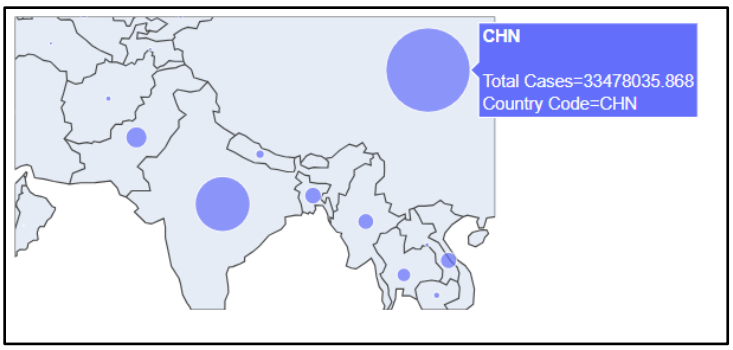


Figure 46: Hover function is used to show information for each country

This visualization utilizes a hover function, providing detailed information about each country when the cursor is moved over specific locations. This includes the country code and the total number of smoking-related deaths. For instance, China, represented by the largest bubble, indicates the highest number of deaths related to smoking among all Asian countries, as shown by the figure above "Total Cases=33 478 035.868" and "country code=CHN". This visualization could provide valuable insights into the specific health impacts of smoking in each country, highlighting the urgency for targeted public health interventions in high-risk areas like China. The difference between bubble sizes makes it easy to see which countries are more severely affected bay smoking-related death. Make it a crucial tool to help the policymakers quickly identify where immediate action and resources are needed.

4.3 Choropleth Map: Cause of Death due to 'Unsafe Water Source' across the World

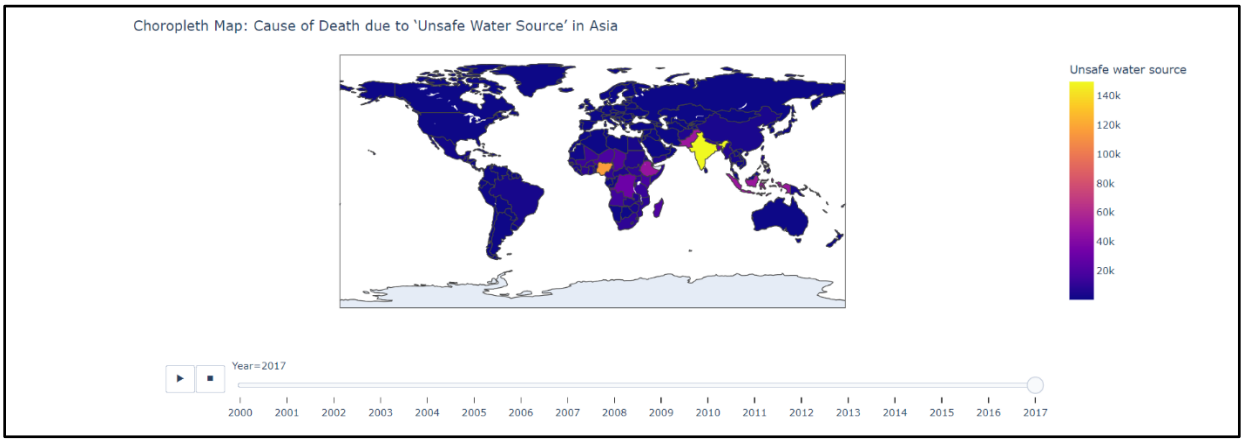


Figure 47: Cause of Death due to 'Unsafe Water Source' across the World

The figure above is used to analyse the cause of death due to unsafe water source globally between year 2000 to 2017. The lighter shades typically indicate a higher number of deaths due to unsafe water sources in a specific country while the darker shades indicate lower number. The slider below represents the year of this study which by having this slider, we can gain insight and analyze the cause in a specific year. In 2017, countries in Asia have lighter shade which refers to serious deaths caused by unsafe water source compared to other regions. Furthermore, among the countries in Asia, India has the highest death rates. Water pollution in India is a major environmental issue and the largest source is untreated sewage followed by agricultural runoff and unregulated small-scale industry. Due to industries, untreated sewage and solid waste, most of the rivers, lakes, and surface water in India are polluted. These have led to pollution of rivers and groundwater sources and caused unsafe water sources.

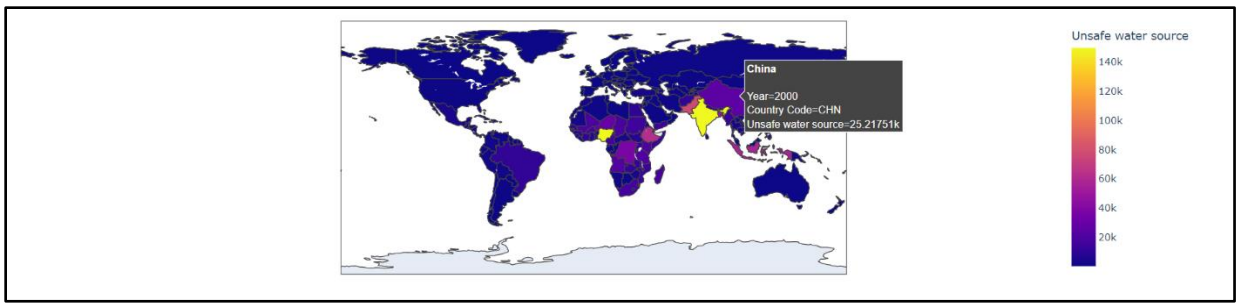


Figure 48: Utilization of Hover function

All the information of the country will be shown such as the country name, year, country code, and the death rates when a hover function is utilized. In figure 47, the cursor is moved to China, and thus, all the information we can obtain from it with detail.

4.4 Treemap: Cause of Death due to 'Unsafe Sanitation' across the World

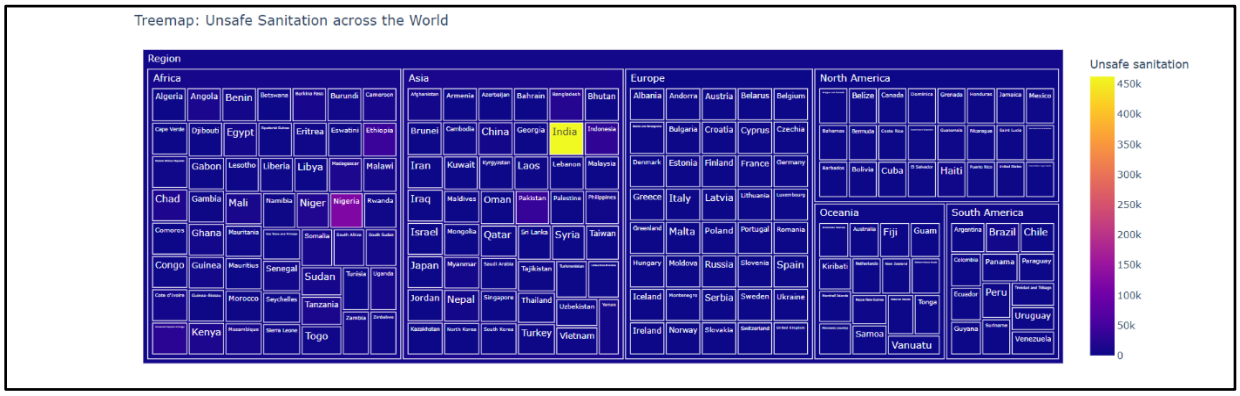


Figure 49: Cause of Death due to 'Unsafe Sanitation' across the World

The figure above is used to analyse the cause of death due to unsafe sanitation globally between year 2000 to 2017. In the treemap above, the lighter shades typically indicate a higher number of deaths due to unsafe water sources in a specific country while the darker shades indicate lower number of deaths due to unsafe water sources. It shows in different categories levels and each rectangle represents a different region. Based on the figure, India contributed the highest death rates among other countries in Asia, followed by Nigeria in Affrica. According to the report Water, Sanitation and Hygiene, n.d., almost half of the people in India suffered the indignity of defecating in fields, forests, bodies of water, or other public spaces due to lack of access to toilets. Furthermore, not washing hands regularly and water contamination in homes and communities increase the risk of spreading diarrheal and waterborne diseases. Additionally, inadequate water, sanitation and hygiene (WASH) services in India's health facilities contributes to high neonatal mortality rate. To mitigate the problem, we all need to always ensure sustained use of toilets and hygiene practices to avoid from unsafe sanitation that can cause to death.

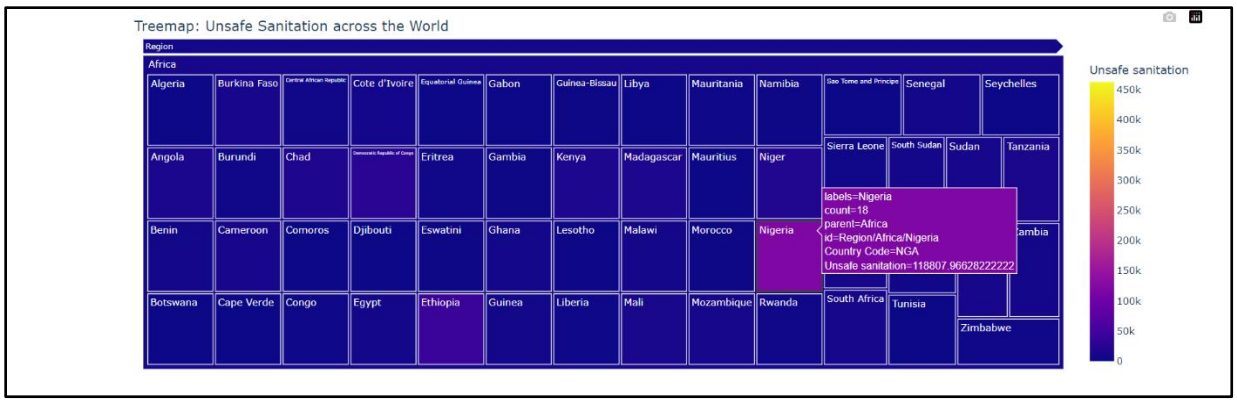


Figure 50: Drill Down into Africa

Treemap has a hover function which by moving the cursor to a specific country and we may obtain all the information of the country. The unique feature of a treemap is it can drill down into subcategories to understand specific aspects of the problem. This feature enhances and speeds up the analysis phase. Therefore, we can conduct additional analysis by selecting specific regions and countries of interest.

4.5 Sunburst Chart: Cause of Death due to Low Physical Activity across the World

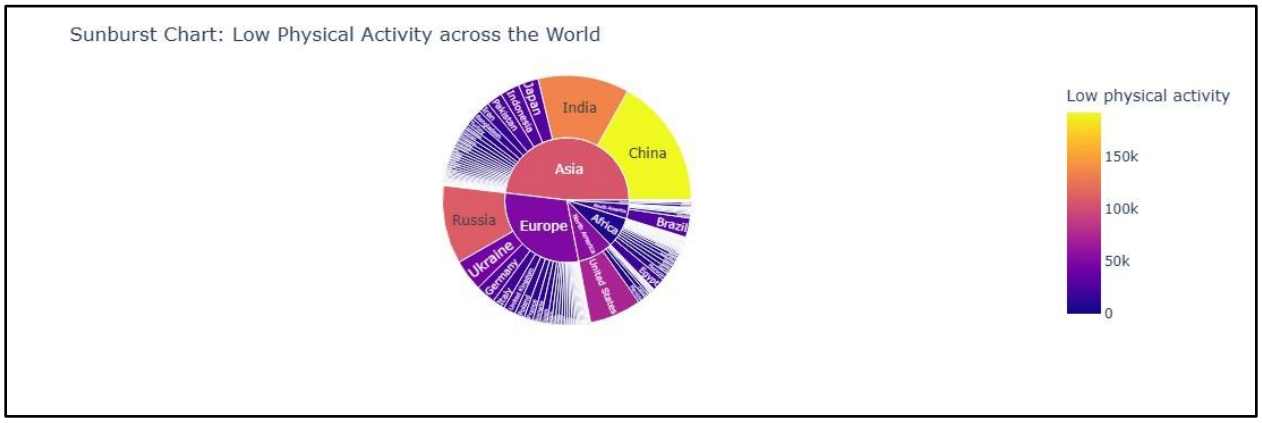


Figure 51: Cause of Death due to Low Physical Activity across the World

Low physical activity (LPA) is linked to several major communicable diseases (NCDs) which includes heart disease, diabetes, colon and breast cancers, mental health disorders and premature mortality (Xu et al., 2022). The figure above shows the cause of death due to low physical activity globally. Overall, China have obtained highest number of death cause by low physical activity. Among the six regions, Asia contributes highest number of death due to low physical activity while Oceania has obtained lowest number of death due to low physical activity. The countries in Asia region like China and India have large population and high population density in urban area which causes limited space for physical activity. In contrast, the Oceania region has a small population over large area so people living there can have more space to carry out physical activity. Besides, people live in Oceania region like Australia and New Zealand more emphasis in engaging in physical activity. Asia region more focus on development of industry and technology so people living in Asia less engaged physical activity in daily work whereas Oceania region more concentrate on agriculture, tourism, and sport fields so they engaged more physical activity during working.

In Asia region, China, India, Japan, and Indonesia have contributed high number of death due to low physical activity. Since people living in China concentrate more on academic and career success, they do not have extra time for exercising and joining physical activity. Besides, the large population in India may contributes to high number of deaths caused by low physical activity. The working hour in Japan is long compared to other countries so people living in Japan do not have enough time and energy to exercise after working. Besides, high proportions of elderly people in Japan also leading to unable carry out physical activity. The infrastructure for physical activities in Indonesia may not be fully developed in urban areas leading to less engage in physical activities.

In Europe region, Russia contributes high number of death due to low physical activity. Since Russia has long and harsh winters, people living in Russia difficult to carry out outdoor activities. In North America region, United States has obtained highest number of death due to low physical activity. In United States, majority people work in the office which requires sitting for long time. Instead of engaging in physical activity during free hours, they spend most of their time on social media, watching television and playing online games.

In Africa region, Egypt has achieved highest number of death due to low physical activity. The hot climate especially summer in Egypt can discourage outdoor activities. Besides, cities in Egypt like Cairo, has experienced rapid urbanization and population growth. It is difficult to carry out physical activity like cycling and walking because the urban environment makes the traffic congestion.

In South America region, Brazil has obtained highest number of death due to low physical activity. Brazil struggles with significant income inequality. Lower-income populations may have limited access to safe and affordable places for physical activity, such as gyms or sports facilities. In addition, they do not have enough time to engage in physical exercise because of working and family obligations.

4.6 Treemap: Cause of Death due to Diet Low in Nuts and Seeds across the World

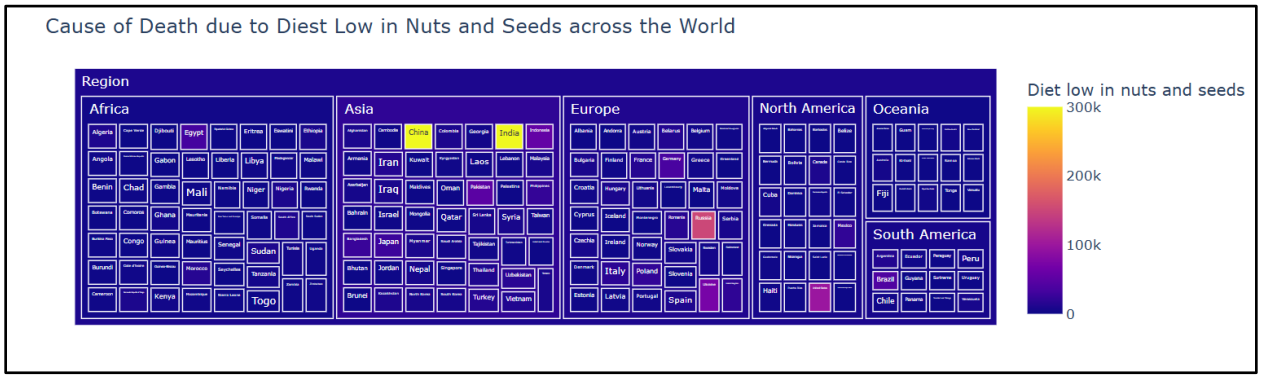


Figure 52: Cause of Death due to Diet Low in Nuts and Seeds across the World

Nuts and seed are the rich sources of essential nutrients such as protein, vitamin E and healthy fat. A study conducted by Arnesen et al. (2023) has shown that moderate intake of nuts and seeds is associated with a lower risk of death that caused by cardiovascular disease, cancer, diabetes and coronary heart disease. This is due to intake of nuts and seeds are helpful for us to lowering cholesterol, balancing blood sugar and reducing insulin resistance.

The figure 52 is used to analyze the cause of death due to diet low in nuts and seeds. The treemap analysis reveals that Asia has the highest number of deaths compared to other regions while South America records the lowest number of deaths caused by diet low in nuts and seeds. Across the world, each region has its own unique dietary traditions and preferences influenced by factors such as culture, history, geography, and agricultural practices. In many Asian countries like China, Japan, and Malaysia, people mainly eat rice, noodles, and soy-based products as their staple food, along with vegetables, seafood, and meat. Instead, nuts and seeds are commonly used as snacks, garnishes, ingredients or toppings in certain dishes. Compared to South America, nuts and seeds are widely used in sauces, bread, sandwiches, cereal, and yogurt. Therefore, South America has more consumption of nuts and seeds compared to Asia.

Upon closer examination for Asian region, we found that China records the highest number of deaths at 300762, closely followed by India with 300000 deaths. There is only a slightly difference of 762 deaths between China and India. Both China and India have the largest population in the world, so contribute to the highest number of deaths in Asia. India has a population of over 1.43 billion people while China has a population of just over 1.4 billion

inhabitants (Aaron O'Neill, 2024). This explains why the number of deaths between these two countries is only slightly different. A study from World Health Organization has shown that cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. Therefore, lack of nutritional education and awareness about the health benefits of nuts and seeds could indeed contribute to the higher number of deaths related to a low intake of these foods in China and India. If people are not informed about the importance of including nuts and seeds in their diet to reduce the risk of cardiovascular disease and diabetes, they may not prioritize consuming these foods.

Russia from Europe region ranks third with deaths of 151558, followed by United States from region North America with 98384 deaths. This may be due to nuts and seeds are relatively expensive compared to staple grains and vegetables. A study by M.Shahbandeh has shown that the wholesale price per kilogram for cashew nuts in United States is about $11.2 while Russia is approximately $7.3. These higher prices may make nuts and seeds less accessible to lower-income populations in Russia and eventually lead to lower consumption rates and potentially contributing to higher mortality rates from diet-related diseases including heart disease and diabetes.

Lastly, Ukraine ranks fifth with a total number of 71131 deaths. This may be due to the limited production of nuts and seeds that cannot meet demand. In 2022, China was the highest production of nuts with 3.8 million metric tons, followed by Vietnam, India, and the US (*Global Nuts Trends in 2022*, n.d.). Since Ukraine is not one of the top producers, therefore nuts and seeds may not be as readily available or affordable compared to other staple foods in Ukraine. Limited access to a variety of nuts and seeds at local markets or grocery stores could deter people from regularly incorporating them into their diet.

The top five countries indicate a concerning trend regarding deaths caused by low nut and seed intake. Since nuts and seeds are rich in essential nutrients and have been linked to various health benefits such as reducing the risk of cardiovascular diseases and diabetes, the high number of deaths in these countries could be partially attributed to inadequate consumption of these foods. This highlights the importance of promoting awareness about the health benefits of nuts and seeds and encouraging their inclusion in diets to improve overall health outcomes.

4.7 Sunburst Chart: Cause of Death due to Outdoor Air Pollution across the World in Year 2017

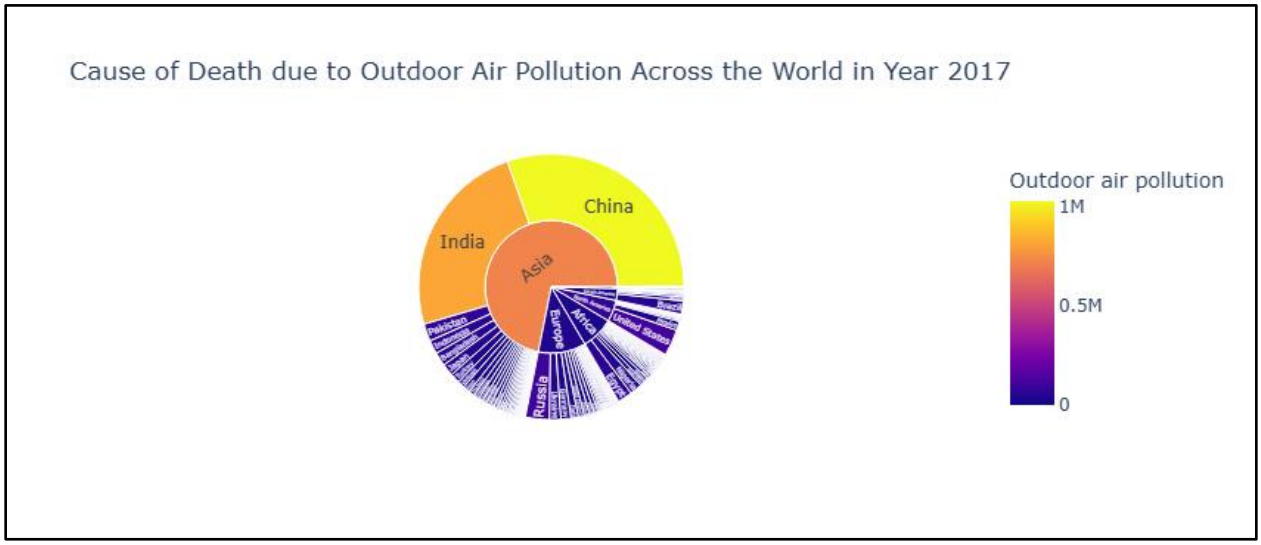


Figure 53: Sunburst Chart

Sunburst chart is a type of pie chart used to show categories and sub-categories in the dataset. This sunburst chart shows the cause of death due to outdoor air pollution across the world in year 2017. In this sunburst chart we can select the region by clicking with mouse to manipulate the visualization display in the sunburst chart.

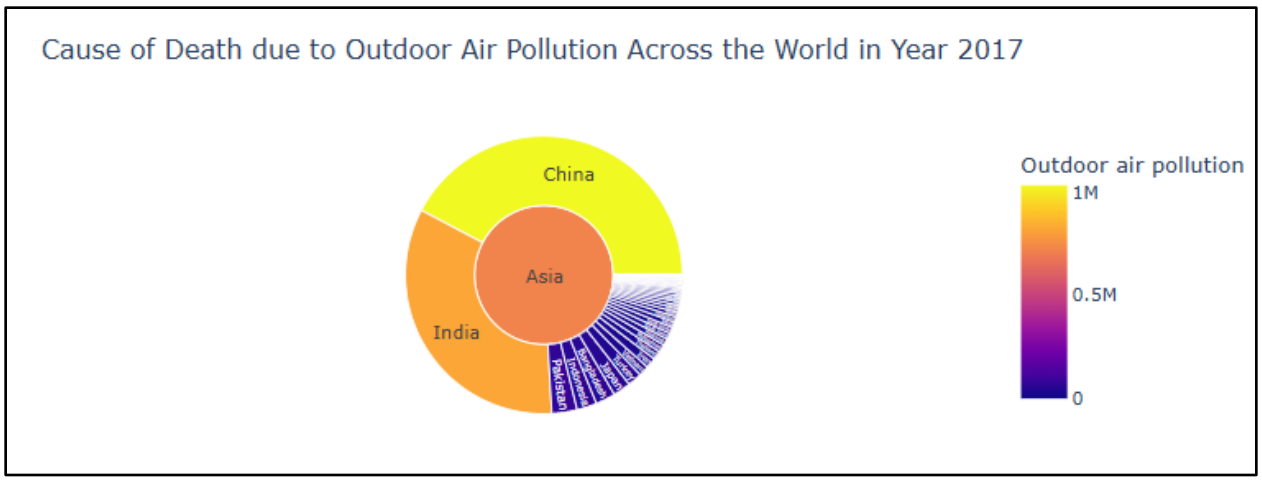


Figure 54: Select "Asia" Regions Sunburst Chart

Based on Figure 54, in Asia, the top 3 countries deaths cause by outdoor air pollution are China with 1.030M, India with 819k, and Pakistan with 71960.58. Compared to other countries in Asia, China and India have obtained significantly higher number of death due to outdoor air pollution. In 2017, the average values of fine particulate matter (PM2.5) in India is 407 µg/m³. The high number of vehicles on the roads and industrial emissions which released pollutants such as nitrogen oxides (NOx), carbon monoxide (CO), and particulate matter (PM) and significantly to air pollution. In China, they rely on coal-fired power plants as the main energy source especially in colder months. These fuel sources produce PM2.5 that rises into the air and can affect people's health when levels are high. This scenario may also cause by the large population in the China and India.

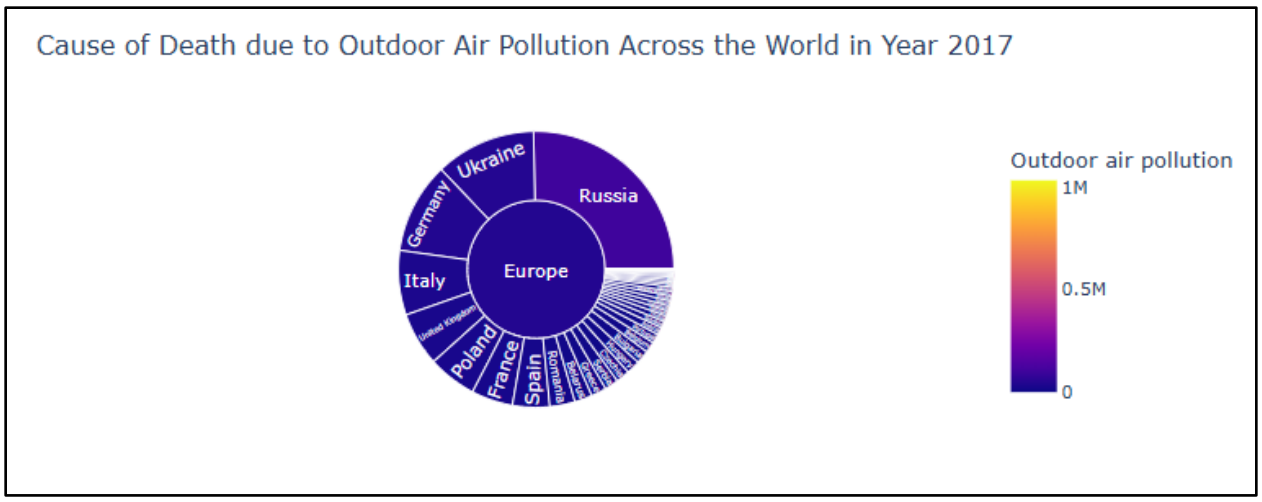


Figure 55: Select "Europe" Regions Sunburst Chart

Based on Figure 55, in Europe, the top 3 countries with deaths caused by outdoor air pollution are Russia with 99613.01, Ukraine with 46919.75, and Germany with 42060.97. In Europe region, Russia contributes to highest number of deaths due to outdoor air pollution. This may be due to it has a larger population compared to other European countries so it has the highest death toll from outdoor air pollution.

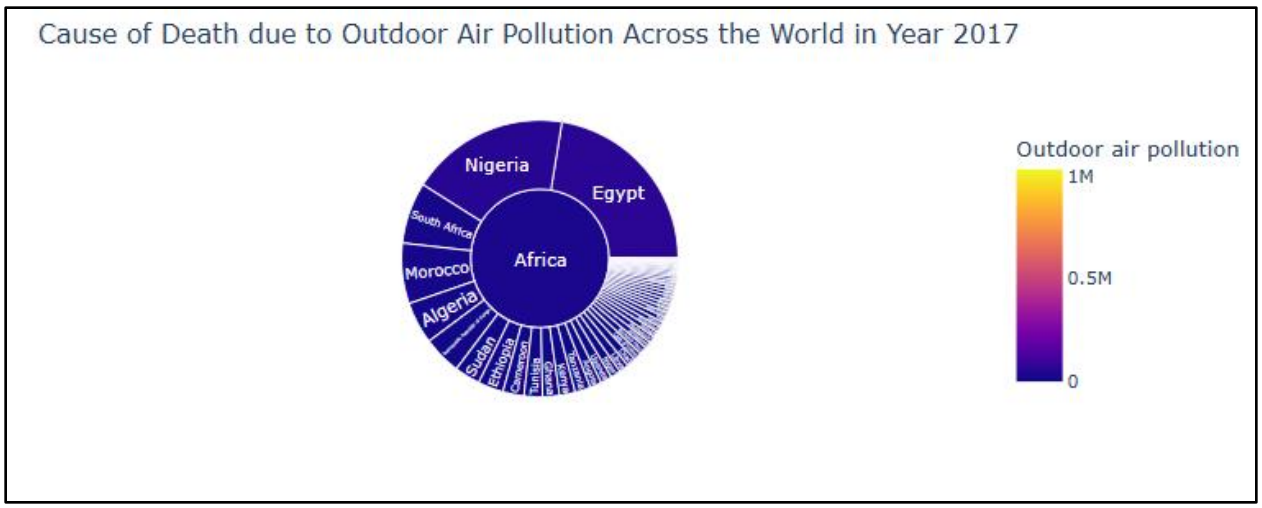


Figure 56: Select "Africa" Regions Sunburst Chart

From the results of sunburst chart filtered by Africa, Egypt, Nigeria and South Africa are the countries with the highest number of deaths from outdoor air pollution. The main pollutant in African countries should be the fossil fuel industry, as Africa has many fossil fuel power plants, smelters, and oil and gas infrastructure as its source of energy. As a result, Africa is home to some of the world's worst hotspots for harmful gases such as nitrogen dioxide, largely associated with its fossil fuel power plants. Then, exposure to air pollution in Africa has a huge impact on public health and is a leading cause of premature death.

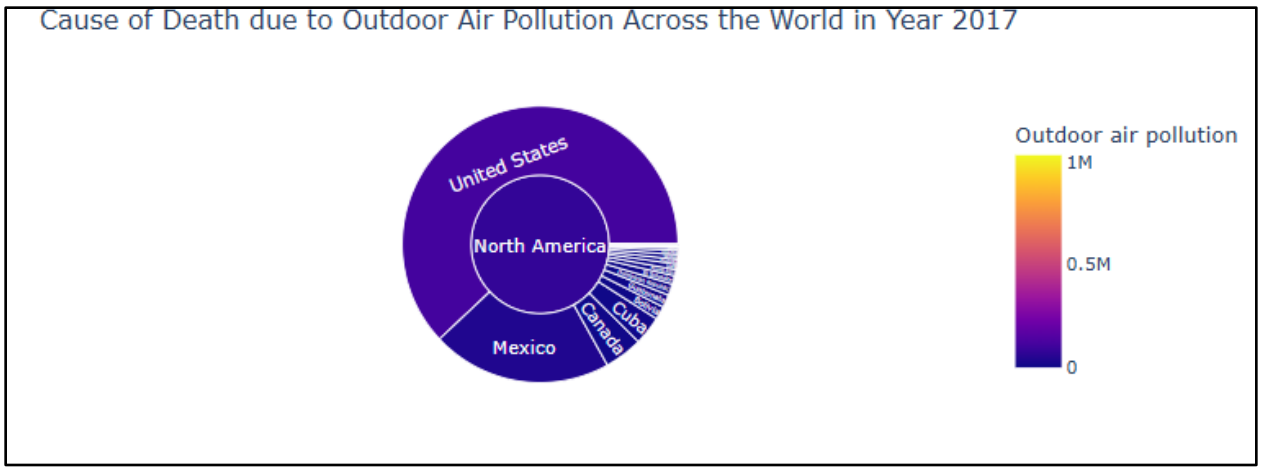


Figure 57: Select "North America" Regions Sunburst Chart

Based on figure 57, in North America, the top 3 countries deaths cause by outdoor air pollution are United States with 110k, Mexico with 37367.52, and Canada with 7937.37. The number of deaths caused by outdoor air pollution is significantly higher in the United States than in other countries in North America. This situation may be due to the large population of the United States and the fact that the United States is one of the high-income countries in the world. They more focus on developing economics and host various industrial facilities, power plants, and factories that contribute significantly to air pollution.

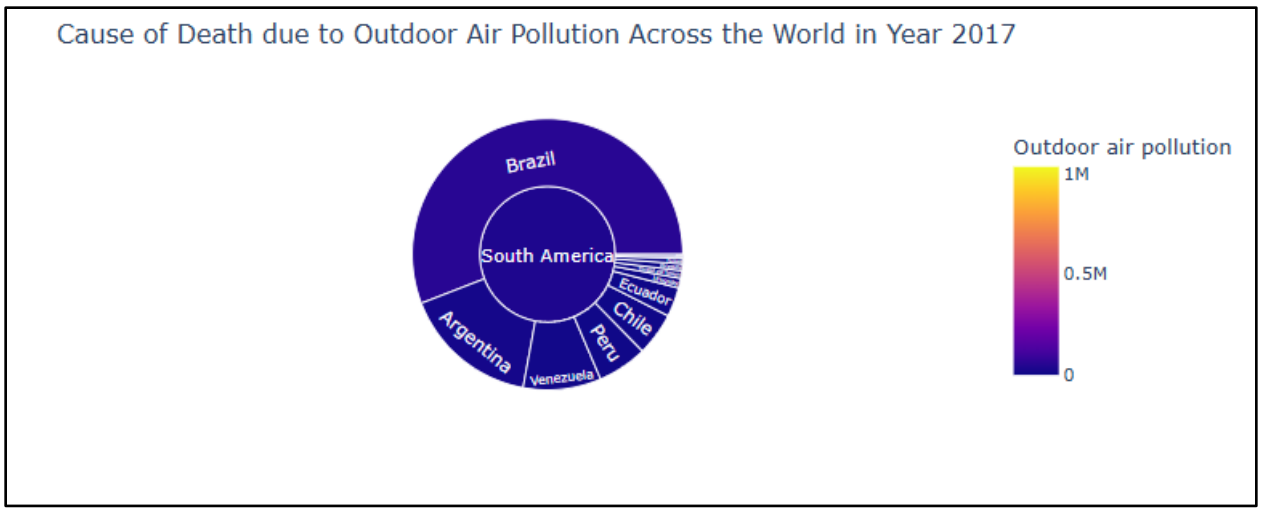


Figure 58: Select "South America" Regions Sunburst Chart

South America, the top 3 countries deaths cause by outdoor air pollution are Brazil with 54859.56, Argentina with 15930.98, and Venezuela with 9141.87. The number of deaths caused by outdoor air pollution is significantly higher in the Brazil than in other countries in South America. The reason for the pollution is Brazil tends to rely heavily on vehicles which cause by lack of public transport infrastructure. Hence, its year-round environmental pollution levels come from the exhaust and smog emitted by cars, with many heavy vehicles such as cars, motorcycles and trucks, buses and trucks on the road. Many of these vehicles are older models, especially in rural areas which run on fossil fuels, especially low-quality diesel, all of which release some pollutants.

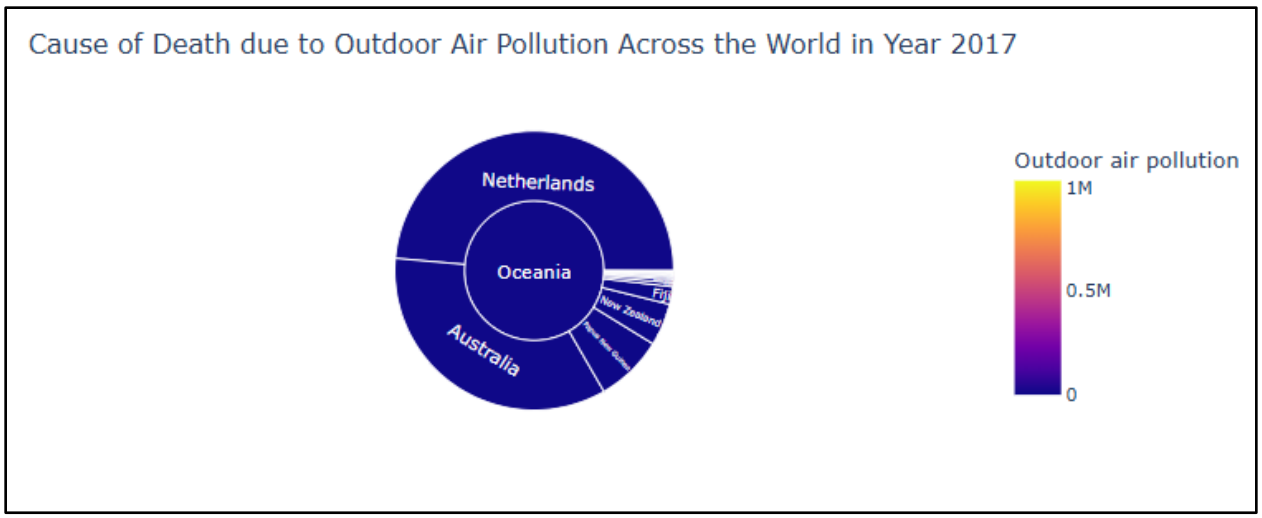


Figure 59: Select "Oceania" Regions Sunburst Chart

Based on Figure 59, in Oceania, the top 3 countries deaths cause by outdoor air pollution are Netherlands with 6824.72, Australia with 4880.68 and Papua New Guinea with 1105.87. The sunburst chart shows that Oceania countries pay attention to ecological protection and maintain good air quality over a long period of time.

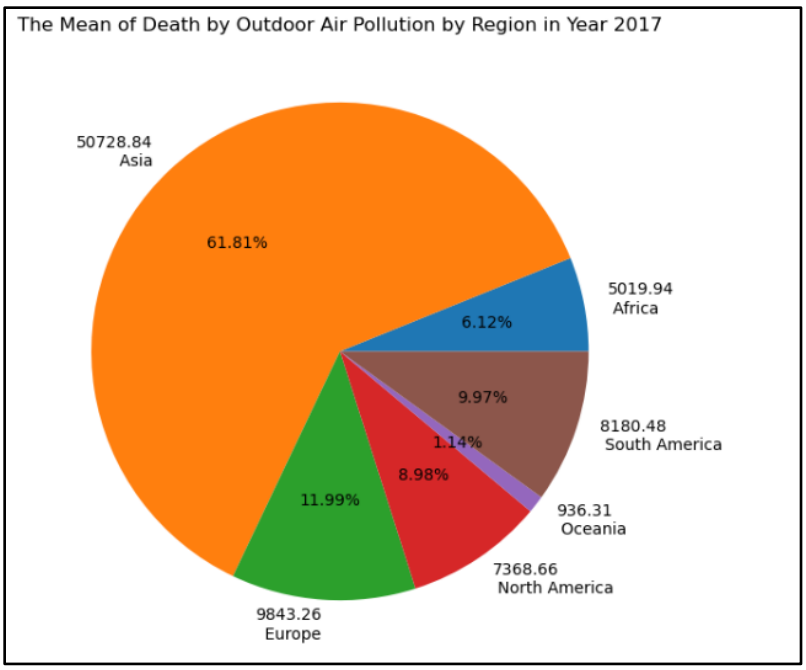


Figure 60: Pie Chart of the Mean of Death by Outdoor Air Pollution by Region in Year 2017

Among the 6 Regions, the overalls sunburst chart shown the highest death which cause by outdoor air pollution is in Asia and the lowest death which cause by outdoor air pollution is in Oceania. Since there are the most of countries in Asia, therefore we also analysis the mean of death by outdoor air pollution in pie chart. The results of the highest and lowest death values still in Asia with 50728.84 and Oceania with 936.31. Hence, we can interpret that the worst air quality is in Asia since there are a lot of developing countries which more focus on development in economic and industry, while Oceania has cleaner air quality than other regions since most of the countries in Oceania are focus on environmental protection such as Australia and New Zealand.

5.0 Summary

From 2000 to 2017, the global landscape of mortality has undergone significant changes. This analysis's main objective is to examine the causes of deaths worldwide during the period. The dataset is obtained from Kaggle website, and it covers the entire world. Based on the causes that have chosen, there are same country contributes more death rate in a different cause. In terms of drug use, North America, South America, parts of Europe, and Australia spiking up the death rates, while China has indirectly risen the death due to smoking. Furthermore, India leads in mortality rates associated with unsafe water sources and poor sanitation compared to other countries. Regarding low physical activity, China reports the highest death rates, and for diets low in nuts and seeds, both China and India have higher death rates compared to other countries. Additionally, China also has highest death rates from air pollution and outdoor air pollution compared to other nations. The trend of the factors that have been chosen is somehow unstable, as there are increasing and decreasing from year to year. Besides that, we obtained that China has more mortality rates than other countries as it contributed more death rates in various factors. This analysis will provide valuable findings into public health strategies and policymaking aimed at reducing deaths and improving population health.

6.0 Reference

Aaron O'Neill. (2024, May 22). *Twenty countries with the largest population in 2024 (in millions)*. Statista. https://www.statista.com/statistics/262879/countries-with-the-largest-population/

Arnesen, E. K., Thorisdottir, B., Bärebring, L., Söderlund, F., Nwaru, B. I., Spielau, U., Dierkes, J., Ramel, A., Lamberg-Allardt, C., & Åkesson, A. (2023). Nuts and seeds consumption and risk of cardiovascular disease, type 2 diabetes and their risk factors: a systematic review and meta-analysis. In *Food and Nutrition Research* (Vol. 67). https://doi.org/10.29219/fnr.v67.8961

Cardiovascular diseases. (n.d.). World Health Organization. Retrieved June 2, 2024, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

Global Nuts Trends in 2022. (n.d.). ReportLinker. Retrieved June 2, 2024, from https://www.reportlinker.com/clp/global/4050

M.Shahbandeh. (n.d.). *Cashew nuts: global wholesale price by country* . Statista. Retrieved June 2, 2024, from https://www.statista.com/statistics/967569/wholesale-prices-of-cashews-by-country-worldwide/

Xu, Y. Y., Xie, J., Yin, H., Yang, F. F., Ma, C. M., Yang, B. Y., Wan, R., Guo, B., Chen, L. D., & Li, S. L. (2022). The Global Burden of Disease attributable to low physical activity and its trends from 1990 to 2019: An analysis of the Global Burden of Disease study. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.1018866

Water, sanitation and hygiene. (n.d.). UNICEF India. https://www.unicef.org/india/what-we-do/water-sanitation hygiene#:~:text=In%202015%2C%20nearly%20half%20of,lack%20of%20access%20to%20toilets.

7.0 Appendix

The dataset link from Kaggle:

https://www.kaggle.com/datasets/varpit94/worldwide-deaths-by-risk-factors

The link for coding:

https://colab.research.google.com/drive/1vjXV5EqYw2_WE-h2fxJ3r9VNTb6wLQUO?usp=drive_link