Hyperiondev

TASK

Exploratory Data Analysis on the Number of Deaths by Risk Factor Dataset

Visit our website

Introduction

In this exploratory data analysis report we will be working with the number of deaths by risk factor dataset from our world in data. This dataset contains the number of deaths due to certain risk factors for different countries, regions, and the world for a period from 1990 to 2019. The dataset has 6840 rows, and 31 columns.

	Entity	Code	Year	Deaths - Cause: All causes - Risk: Outdoor air pollution - OWID - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: High systolic blood pressure - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Diet high in sodium - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Diet low in whole grains - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Alcohol use - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Diet low in fruits - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Unsafe water source - Sex: Both - Age: All Ages (Number)	...	Deaths - Cause: All causes - Risk: High body-mass index - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Unsafe sanitation - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: No access to handwashing facility - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Drug use - Sex: Both - Age: All Ages (Number)
0	Afghanistan	AFG	1990	3169	25633	1045	7077	356	3185	3702	...	9518	2798	4825	174
1	Afghanistan	AFG	1991	3222	25872	1055	7149	364	3248	4309	...	9489	3254	5127	188
2	Afghanistan	AFG	1992	3395	26309	1075	7297	376	3351	5356	...	9528	4042	5889	211

Before we can explore the dataset, it needs to be cleansed by removing any unwanted data and handling all missing data. All of this is done in the body of this report.

DATA CLEANSING

In this part of the analysis, I will look for duplicate rows, empty rows, and anything that I may need to remove from the dataframe before we begin working with the dataset. From the figure below it can be seen that the dataset contains no duplicates.

```
# removing duplicate rows
death_df.drop_duplicates(keep='first')

death_df.shape
```

```
(6840, 31)
```

There are no empty columns, but the headers of the columns are very messy and contain a lot of unnecessary information that we can do without.

	Entity	Code	Year	Deaths - Cause: All causes - Risk: Outdoor air pollution - OWID - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: High systolic blood pressure - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Diet high in sodium - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Diet low in whole grains - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Alcohol use - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Diet low in fruits - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Unsafe water source - Sex: Both - Age: All Ages (Number)	...	Deaths - Cause: All causes - Risk: High body-mass index - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Unsafe sanitation - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: No access to handwashing facility - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Drug use - Sex: Both - Age: All Ages (Number)
0	Afghanistan	AFG	1990	3169	25633	1045	7077	356	3185	3702	...	9518	2798	4825	174
1	Afghanistan	AFG	1991	3222	25872	1055	7149	364	3248	4309	...	9489	3254	5127	188
2	Afghanistan	AFG	1992	3395	26309	1075	7297	376	3351	5356	...	9528	4042	5889	211

For example, 'Deaths - Cause: All causes - Risk: Outdoor air pollution - OWID - Sex: Both - Age: All Ages (Number)', can be reduced to just 'Outdoor air pollution - OWID' since we know that the dataset

is about the number of deaths by risk factor, the column can just be the name of the risk factor. Below is the dataframe with the cleansed column headers.

	Entity	Code	Year	Outdoor air pollution - OWID	High systolic blood pressure	Diet high in sodium	Diet low in whole grains	Alcohol use	Diet low in fruits	Unsafe water source	Secondhand smoke	Low birth weight	Child wasting	Unsafe sex	Diet low in nuts and seeds	Household air pollution from solid fuels
0	Afghanistan	AFG	1990	3169	25633	1045	7077	356	3185	3702	4794	16135	19546	351	2319	34372

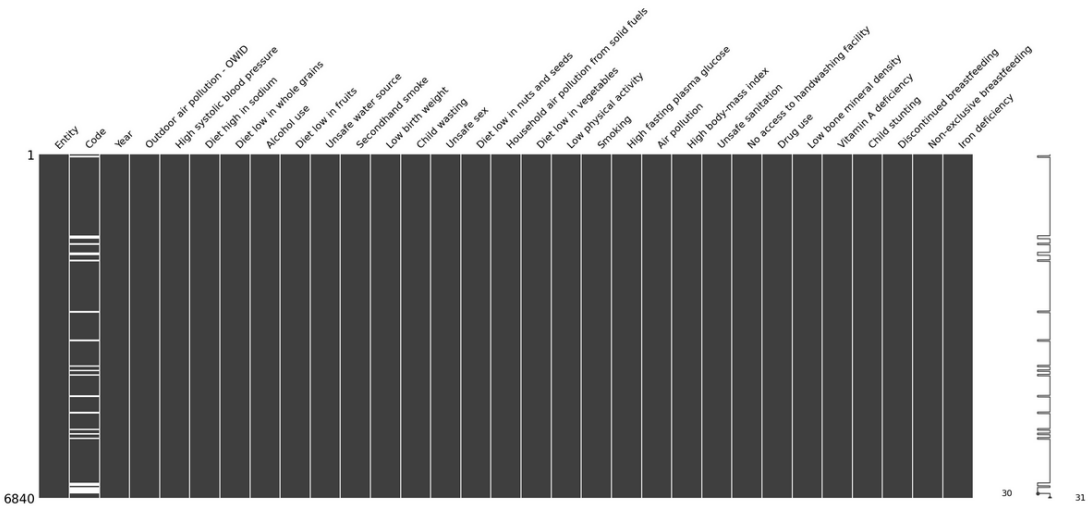
From below we can see that the datatypes for the columns are correct and only one column seems to have missing or null values that will be dealt with later. And of course, we can see that the names of the columns have been cleansed.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6840 entries, 0 to 6839
Data columns (total 31 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Entity                                 6840 non-null   object
 1   Code                                   6150 non-null   object
 2   Year                                   6840 non-null   int64
 3   Outdoor air pollution - OWID           6840 non-null   int64
 4   High systolic blood pressure           6840 non-null   int64
 5   Diet high in sodium                    6840 non-null   int64
 6   Diet low in whole grains               6840 non-null   int64
 7   Alcohol use                            6840 non-null   int64
 8   Diet low in fruits                     6840 non-null   int64
 9   Unsafe water source                    6840 non-null   int64
 10  Secondhand smoke                       6840 non-null   int64
 11  Low birth weight                       6840 non-null   int64
 12  Child wasting                          6840 non-null   int64
 13  Unsafe sex                             6840 non-null   int64
 14  Diet low in nuts and seeds             6840 non-null   int64
 15  Household air pollution from solid fuels  6840 non-null   int64
 16  Diet low in vegetables                 6840 non-null   int64
 17  Low physical activity                  6840 non-null   int64
 18  Smoking                                6840 non-null   int64
 19  High fasting plasma glucose            6840 non-null   int64
 20  Air pollution                          6840 non-null   int64
 21  High body-mass index                   6840 non-null   int64
 22  Unsafe sanitation                      6840 non-null   int64
 23  No access to handwashing facility      6840 non-null   int64
 24  Drug use                               6840 non-null   int64
 25  Low bone mineral density               6840 non-null   int64
 26  Vitamin A deficiency                   6840 non-null   int64
 27  Child stunting                         6840 non-null   int64
 28  Discontinued breastfeeding             6840 non-null   int64
 29  Non-exclusive breastfeeding            6840 non-null   int64
 30  Iron deficiency                        6840 non-null   int64
dtypes: int64(29), object(2)
```

Now that the data cleansing is done, the dataframe has no duplicate rows, but it does have missing datapoints in one column. The next step is to analyse this missing data and how to handle it.

MISSING DATA

First thing to do is to locate the missing data in the dataframe, and a visualisation of the dataframe will make it is easier to get a quick sense of the spread of the missing data.



From the plot above we can see that it confirms only one column has missing values or datapoints. This column contains information about the country code for each of the recorded entities. Below I will be looking at the total number of missing data and what percentage it makes up.

```
# getting the number of missing data points per column
missing_datapoints = death_df.isnull().sum()

# number of missing data points in the columns
missing_datapoints[missing_datapoints != 0]
```

```
Code    690
dtype: int64
```

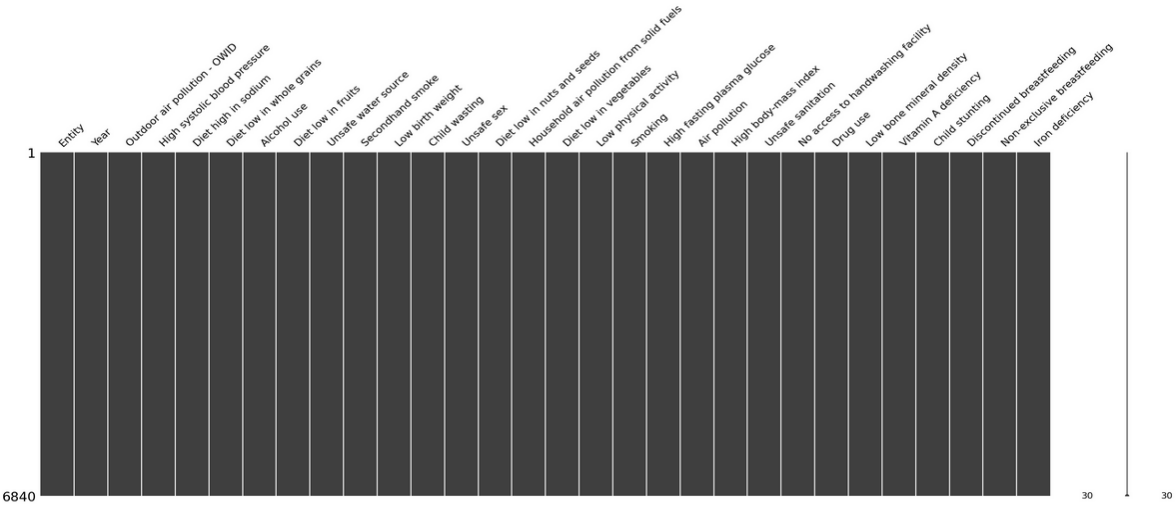
```
# getting the total number of cells
total = np.product(death_df.shape)

# gettng total number of missing data points
missing_data_total = missing_datapoints.sum()

# checking the percentage of missing data
percent = f"{round((missing_data_total/total)*100,2)}%"

print("Percentage of missing datapoints is",percent)
```

```
Percentage of missing datapoints is 0.33%
```

The missing datapoints make up a very small amount of the data at 0.33%, and the data in the column overall makes up only 3.1% of the total dataset. For this column imputation is not necessary as we do not need this column since we have the names of the entities. Therefore, I will be dropping this column since deleting the null datapoint rows will result in information loss.

```
# dropping the Code column
death_df.drop('Code',axis=1, inplace=True)
```

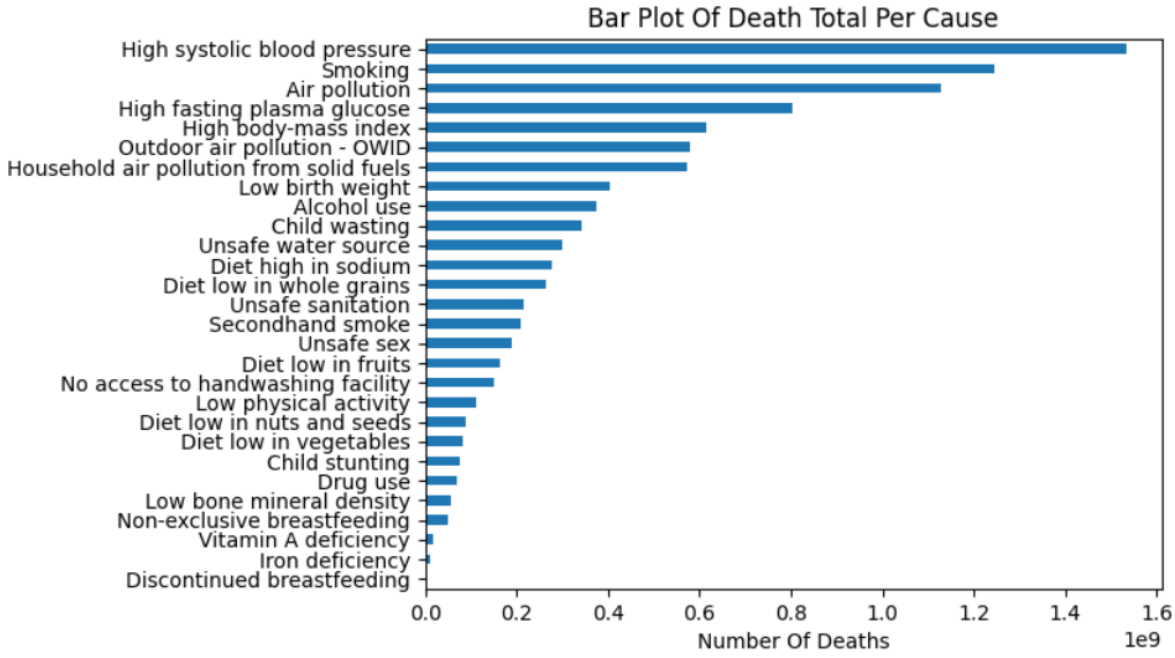
Below is the plot of the missing data after the Code column has been dropped, and it can be seen that there are no more null values.



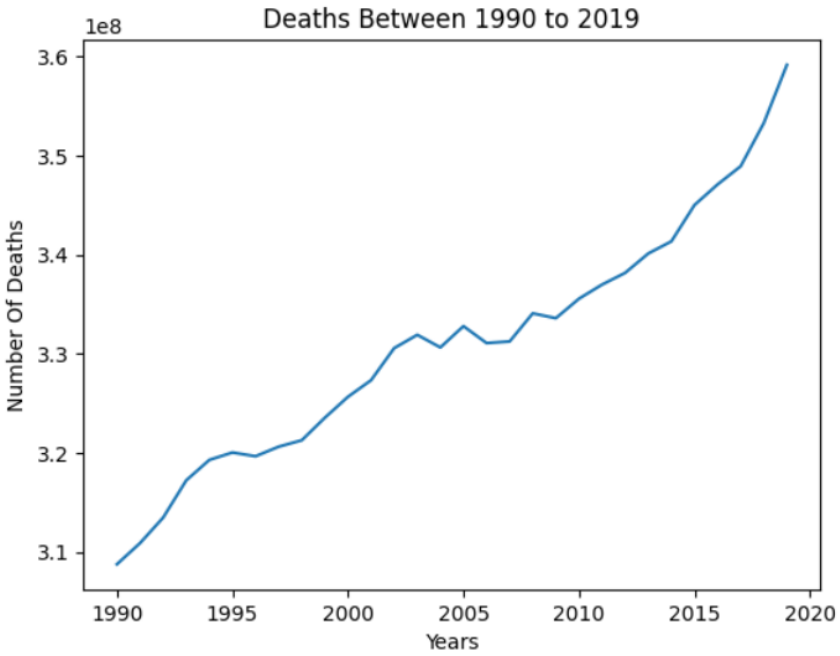
Now that the missing datapoints have been handled, we can move on to the next step which is the exploratory data analysis. From this we can hope to get some insights about this dataset.

DATA STORIES AND VISUALISATIONS

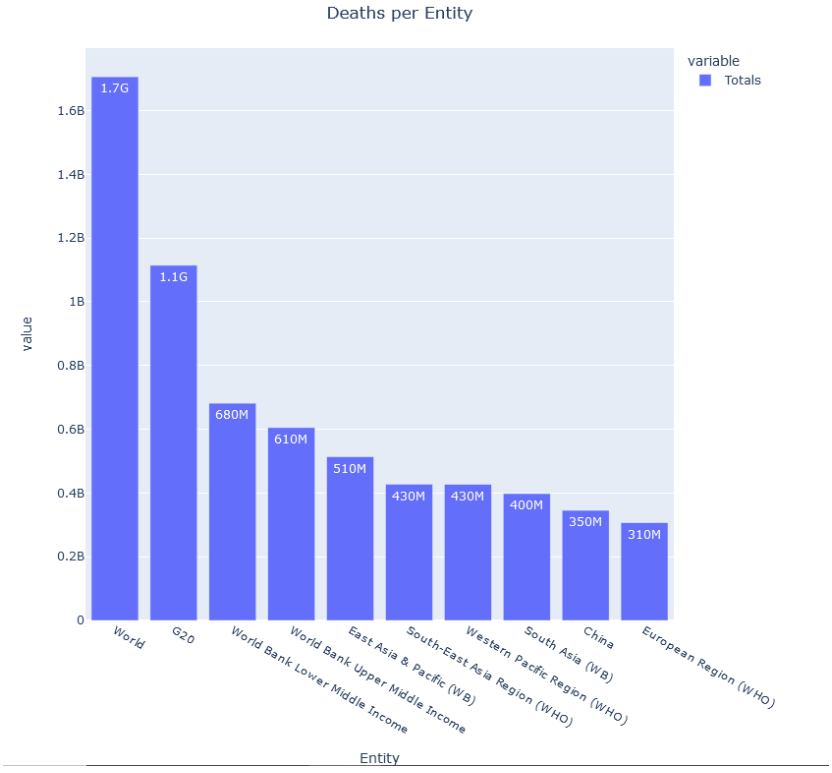
The first step is to see which of these risk factors has contributed the most to deaths in the years from 1990 to 2019. The bar plot below indicates that the top 3 factors that contribute to most deaths are high systolic blood pressure, smoking and air pollution. Of course, it is no surprise to see high body-mass index in the top 5 as obesity is known for contributing to the development of many chronic diseases and cancers.



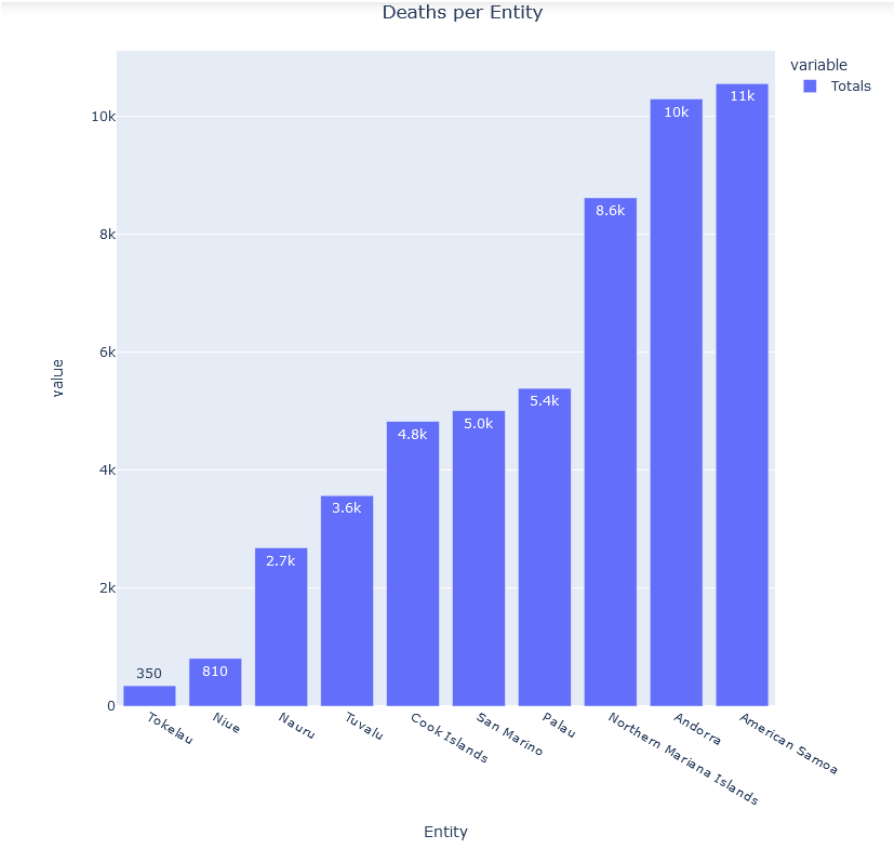
Below is a line plot of the number of deaths over the years and it can be seen that the numbers increase in each year and the prediction is that the trend will continue in this upward trajectory. But of course, this could be expected as the population of the world is increasing due to better living conditions and industrialisation, and with this also comes many diseases and unhealthy lifestyles.



Below is a bar plot showing the top ten entities that have the highest deaths in the 29-year period, the data contains numbers for individual countries, regions, and the whole world. It can be seen that in this period 1.7 billion people died in the whole world due to these risk factors. Note that China appears in the top ten even though it is a country and not a region made up of multiple countries, this could be due to it having a large population.



And below is a bar plot representation of the countries with the lowest number of deaths in the 29-year period, this is to be expected since these are islands which have very small populations. For instance, Tokelau has only 350 deaths and its population as of today is only 1411 people.



Now having analysed the data and gained some insights, I will try to gain death insights on the following risk factors.

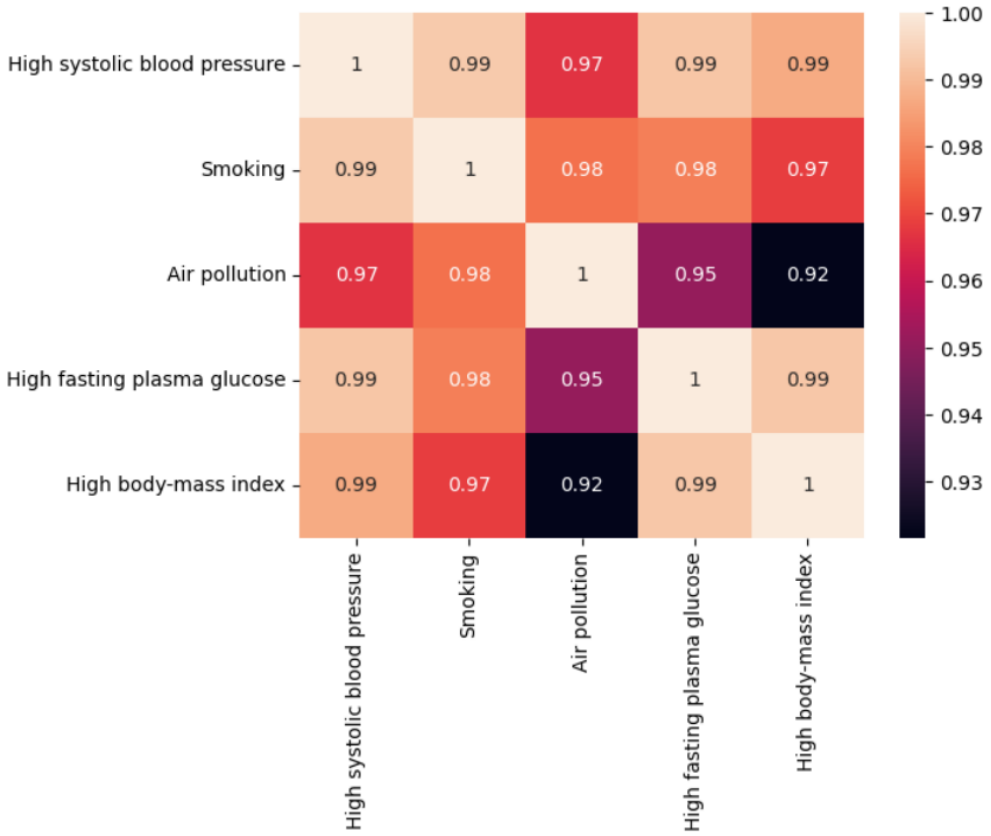
- Relationship between the top 5 risk factors
- Deaths due to air pollution over the years
- Relationship between unsafe sex, alcohol use and drug use deaths.
- Deaths due to unsafe sex over the years.
- Relationship of high body-max index and High systolic blood pressure.
- Deaths due to unsafe sanitation
- Deaths due to unsafe water source

But first I will remove the rows that contain data for the whole world, regions, and any grouped entities data. The focus of the analysis will be on data collected based on individual countries.

```
# list of entities to removes
combined_entities = [

    'African Region (WHO)','East Asia & Pacific (WB)','Eastern Mediterranean Region (WHO)',
    'Europe & Central Asia (WB)','European Region (WHO)','Latin America & Caribbean (WB)',
    'Middle East & North Africa (WB)','North America (WB)','OECD Countries','Region of the Americas (WHO)',
    'South Asia (WB)','South-East Asia Region (WHO)','Sub-Saharan Africa (WB)','United States Virgin Islands',
    'Western Pacific Region (WHO)', 'World','World Bank High Income', 'World Bank Low Income',
    'World Bank Lower Middle Income', 'World Bank Upper Middle Income'
]

# iterating through list of entities to remove
for entity in combined_entities:

    # removing entity from dataframe
    death_df = death_df[death_df.Entity != entity]

death_df.shape
```

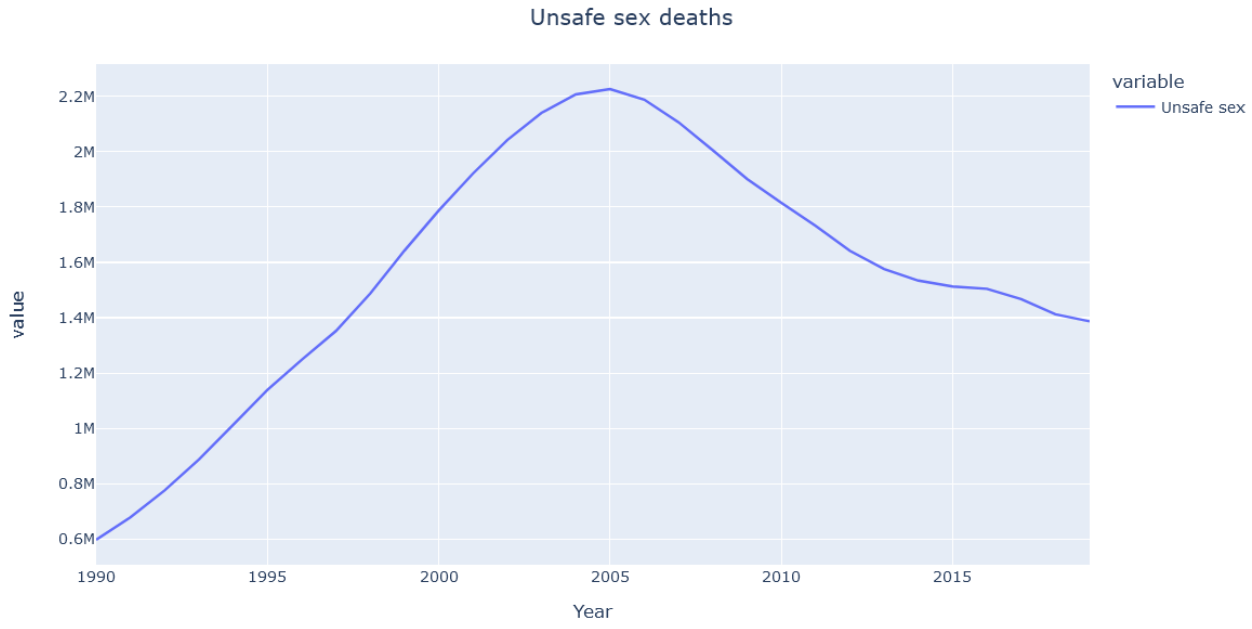
```
(6240, 31)
```

Relationship between the top 5 risk factors



There is a strong positive correlation between all the deaths due to these risk factors. This indicates that most people are likely to die due to one or multiple factors at the same time deteriorating their state of health . This also indicates that some of these factors may develop due to other factors putting one at a very high risk of dying.

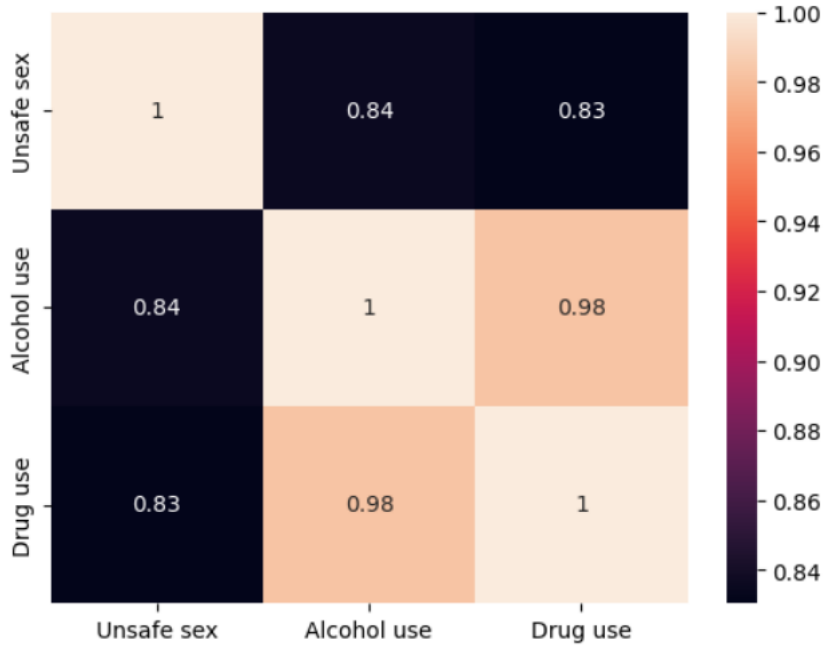
Deaths due to unsafe sex over the years



In this period the number of deaths due to unsafe sex have been on a steady incline from 1990 and reaching a peak in 2005 with deaths of at least 2.2 million in that year, where a steady decline seems to have occurred since.

This could be due to better education about sex and a reduced stigma against STD's, STI's and HIV/AIDS. Treatment and prevention for these has been improving over the years, so it is expected that the numbers would drop.

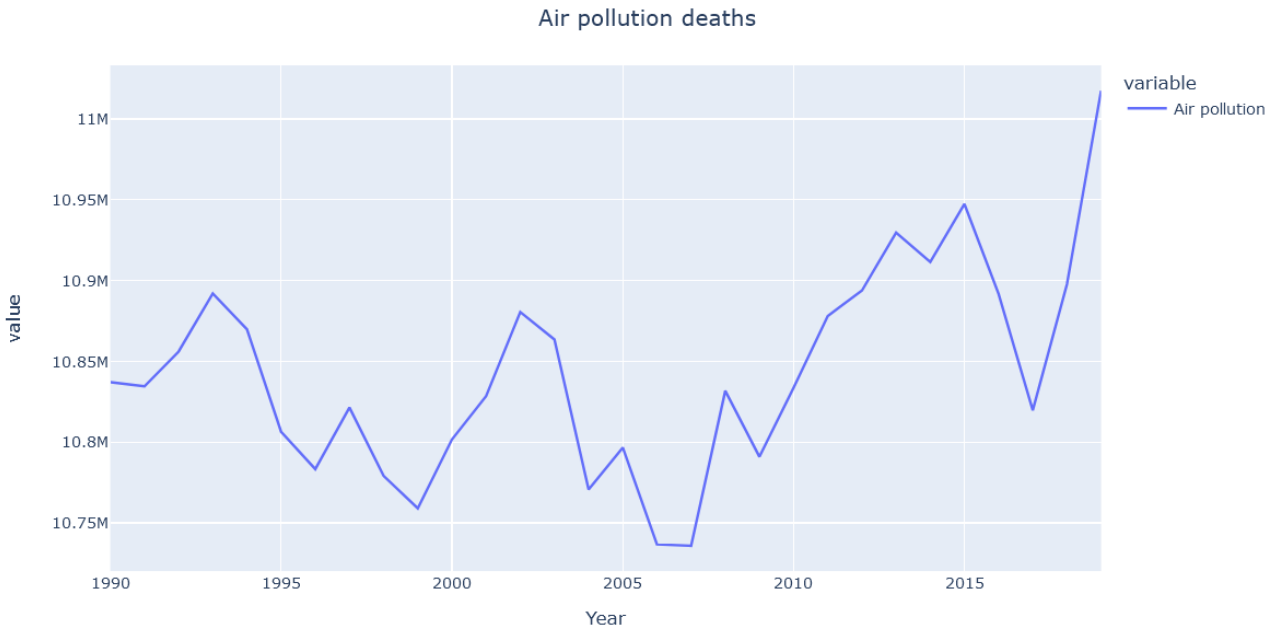
Relationship between unsafe sex, alcohol use and drug use deaths



The correlation indicates that there is a strong chance that if you use alcohol you might engage in unsafe sex and thus lead to your death resulting from unsafe sex and the same is true when engaging in drug use.

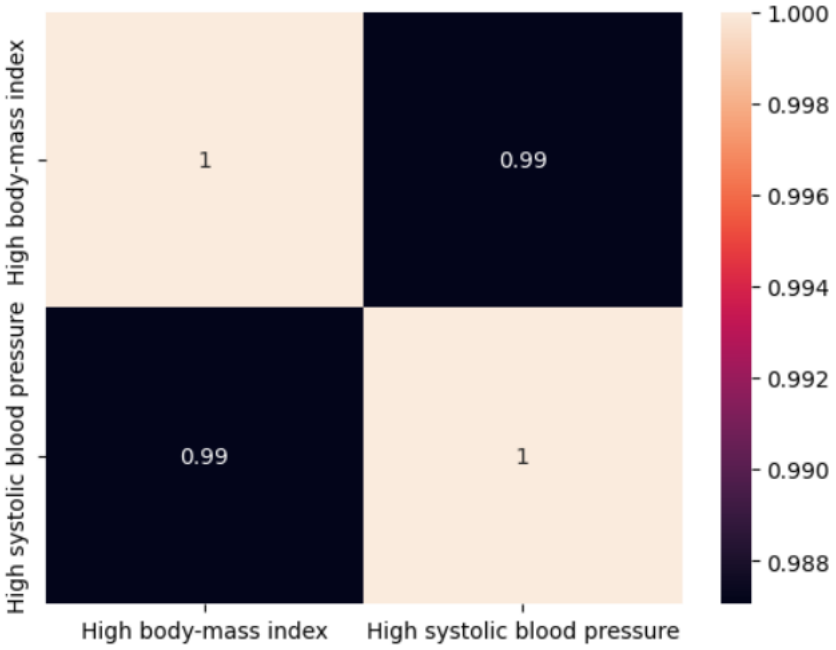
The correlation also indicates that there is a very strong chance that if you engage in alcohol use, you will also engage in drug use and thus these will lead to your death simultaneously.

Deaths due to air pollution over the years



Air pollution death has been fluctuating over the years and this could be due to the awareness of climate change and environmental activism reducing air pollution from time to time. But the plot above shows that after a certain period air pollution rises and then drops and the cycle repeats. This could be due to changes in sovereign and international politics from time to time.

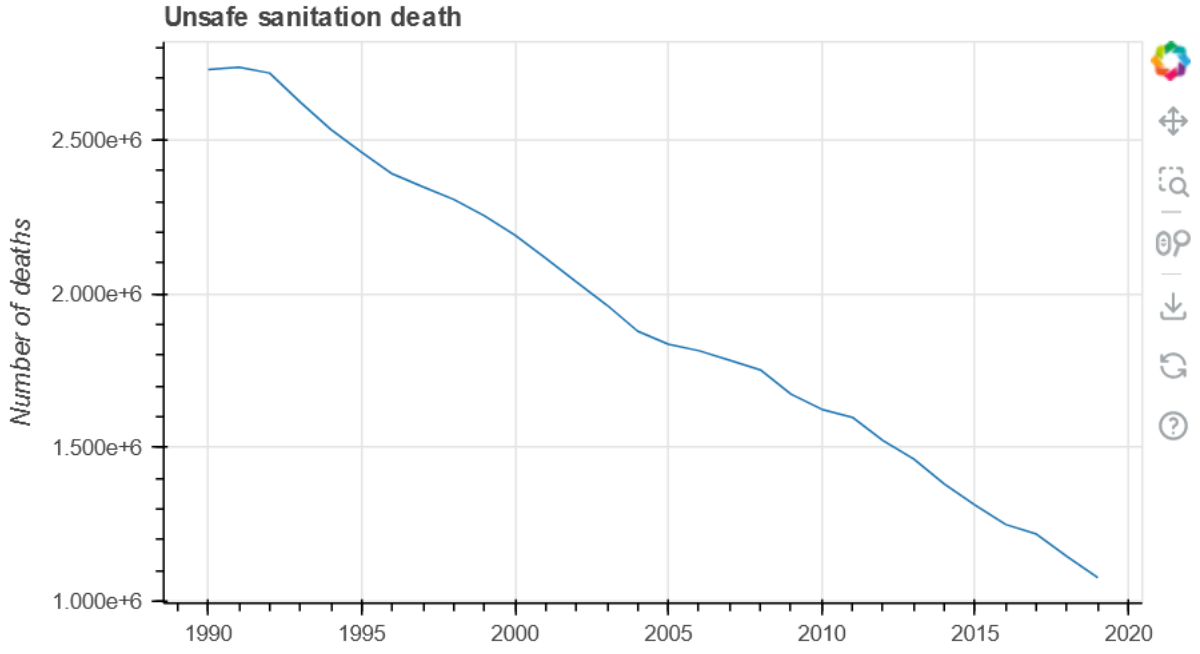
Relationship of high body-max index and High systolic blood pressure.



The correlation between these variables is not surprising as it is known that High body-mass index leads to insulin resistance which contributes to the development of chronic diseases. So, with a high body-mass index, there is high likelihood of high systolic blood pressure and therefore the relationship between deaths due to these factors is highly likely.

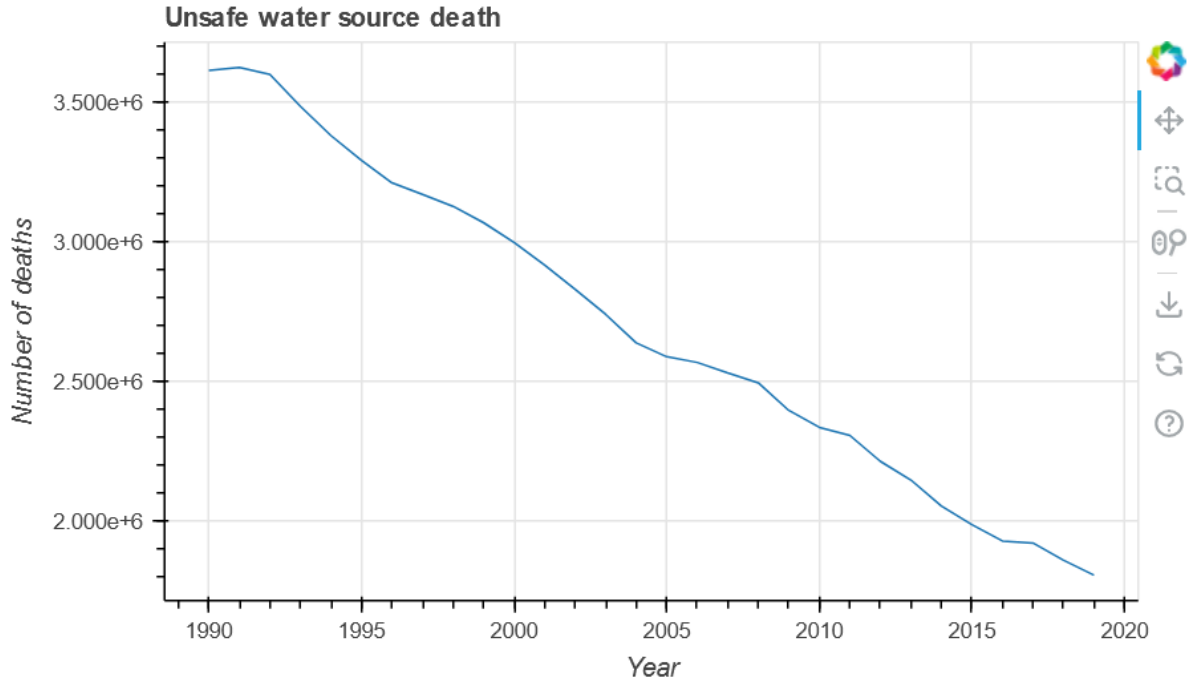
Deaths due to unsafe sanitation

The figure below shows that over the years death due to unsafe sanitation has been on a steady decline. This means over the years the majority of the human population has gained access to adequate facilities for human waste disposal and maintenance of hygienic conditions.



Deaths due to unsafe water source

The number of people dying from the lack of safe water sources has been reduced significantly over the years. This means the majority of the population global has access to clean water that is not contaminated by chemicals/pollution, parasites, germs etc.



Conclusion

There is a lot of risk factors in the world that face the health of the human race and these risk factors kill people in varying degrees. The top 5 factors that result in the death of the most population all have strong relationships indicating that people may not die from just one but multiple simultaneously.

But the data also shows that even though we still are battling some of these risk factors, the world has found preventions and cures for some and thus reduced their impact in the death of people over the years, this gives hope that the world will find solution to improve the life span and health of the population.

THIS REPORT WAS WRITTEN BY : Thabiso Maqhajana

DATA SOURCE : https://ourworldindata.org/obesity