Life Expectancy Data

Sai Priya Muvvala

INFO 5709.002 – Data Visualization

Final Project

Introduction:

Population and health data at the global, regional, and national levels are essential for monitoring progress in development and health for allocating resources. WHO (World Health Organisation) Global Health Estimates, provides a list of causes of death each year with comprehensive and comparable data. The primary goal of WHO includes the monitoring of global health trends, conditions, and factors.

Globally, life expectancy has increased in the population annually. GHO provides a brief report every year regarding the trends and changes occurring in health issues in every country and state. The issued report summarizes all the information regarding diseases prevailing in the associated year, demographic changes, women, health, and the burden of specific diseases and also the statistics of the data pertaining to it.

The annual compilation of the data also provides the developmental goals and their associated targets. The exploration of data also serves as a retrospective analysis to predict the challenges that affect health in further coming years. The analytical report provides the outlines of the trends, challenges, strategic plans, and achievements made in a particular era in improvising health in various areas.

GHO data gives the inputs of data ranging from cancer to air pollution to road traffic injuries. There are numerous causes for mortality which may be briefly categorized into injuries and communicable and non-communicable diseases. The global data demonstrates key health indicators and current scenarios of health issues that intern lead to factors responsible for mortality and longevity of life. Monitoring the annual death rate aids in addressing their causes and modifying health care systems and effectively responding to the evoking reactions from various sectors such as in the healthy diet of people which affects the lifestyle diseases like diabetes and hypertension, providing mental health support, peer support and also in addressing the road traffic accidents and methods to prevent its occurrence.

Understanding the causes of death can aid in understanding how people live, improving health services, and lowering preventable deaths globally while effectively adapting to the change in epidemiological conditions.

Dataset

The dataset is collected from the source Kaggle. It is the collection of health data from the year 2000 to 2019 obtained from the Global Health Observatory (GHO) resource. The data includes different variables responsible for the life expectancy of a human and causes of mortality. Disease burden depicts the loss of the equivalent of one year of full health, which is calculated using disability-adjusted life year (DALY). It counts in different numbers for different diseases. The data is huge containing 147 columns and 1464 rows. The columns represent various attributes that serve as the factors responsible for affecting life expectancy.

The source link for the data used in the project is provided below:

https://www.kaggle.com/datasets/adamsmith852/life-expectancy-data-gho

Attributes:

19 attributes were used to describe the life expectancy data. The remaining attributes were removed from the data by performing EDA.

The attributes used in this project were described briefly as follows:

Country: Name of the countries globally

Year: The data collected from the years 2000 to 2019

Gender: Represents the human sex(identity) as Male or Female

BMI: Represents Body Mass Index

Alcohol: Consumption of alcohol in the respective year

Tuberculosis: Represents the proportion of people affected by the disease Tuberculosis

Syphilis: Represents the proportion of people affected by the disease Syphilis

HIV: Represents the proportion of people affected by the disease AIDS (HIV)

Self-Harm: Represents the proportion of people who died by committing suicides

Stroke: Represents the proportion of people affected by cardiac stroke

Natural Disasters: Represents the proportion of people affected by natural calamities

Epilepsy: Represents the proportion of people affected by the disease Epilepsy

Parkinson: Represents the proportion of people affected by the disease Parkinson

Alzheimer's: Represents the proportion of people affected by the disease Alzheimer's

Alcohol use disorders: Represents the proportion of people whose death occurred due to alcohol overuse

Diabetes: Represents the proportion of people affected by the disease Diabetes

Other nutritional deficiencies: Represents the proportion of people affected by nutritional deficiencies

Preterm birth complications: Represents the proportion of death of people affected by complications during preterm birth

Road injury: Represents the proportion people death that occurred due to road accidents

Tools:

- Python
- Tableau

Exploratory Data Analysis (EDA):

- EDA is one of the technical approaches to data analysis that highlights the key aspects of data sets.
- The pivotal role of Exploratory data analysis is to obtain the characteristic insights of the data.
- It majorly aids in dealing with large data sets to understand the relationship with unknown data.
- It identifies and removes duplicate values in the data
- Identifies and removes the unnecessary data
- It avoids the typo errors present in the data
- Identifies the null values or missing values in data
- Obtains the insights through visuals (charts)
- Removes the data outliers
- Identifies the correlated variables
- EDA identifies the various patterns which are useful for further data processing and modeling
- EDA also helps in creating a hypothesis for further analysis

  Thus Exploratory data analysis provides relevant data by data processing, cleaning, transforming, and analyzing the data to understand the relevant and real-time situations of every aspect of the data. The cleaned data is further assessed by suitable statistical methods, visualization methods, and also by data modelling.

- Various tools such as **Python, R, MATLAB**, etc. are used to perform Exploratory data analysis

- In this project, Python and pandas are used as a tool and library respectively to perform various tasks in EDA which are described as follows:

1.

```
In [33]: import pandas as pd
         data= pd.read_csv('Life Expectancy Data.csv')
         data
```

Out[33]:

	Country	Year	Gender	Life Expectancy at birth	BMI	Alcohol	Tuberculosis	Syphilis	Chlamydia
0	Afghanistan	2019	Male	63.29	NaN	0.003	4.454469	0.050986	0.000000
1	Afghanistan	2019	Female	63.16	NaN	0.022	5.384610	0.043190	0.001424
2	Afghanistan	2015	Male	61.04	22.5	0.002	6.109258	0.056666	0.000000
3	Afghanistan	2015	Female	62.35	24.0	0.014	7.384937	0.047379	0.001201
4	Afghanistan	2010	Male	59.60	22.1	0.006	5.652315	0.051922	0.000000
...	...	...	...	...	...	...	...	...	..
1459	Zimbabwe	2015	Female	60.96	25.3	9.290	0.457023	0.055791	0.004304
1460	Zimbabwe	2010	Male	49.58	22.0	1.470	0.711036	0.089442	0.000000
1461	Zimbabwe	2010	Female	53.21	25.1	7.150	0.464125	0.065319	0.006029
1462	Zimbabwe	2000	Male	45.15	21.7	0.880	2.530362	0.066511	0.000000
1463	Zimbabwe	2000	Female	48.12	24.7	4.220	1.337442	0.049303	0.005999

1464 rows × 147 columns

- The above is used to import the data into the Pandas library, to read the text. The obtained output conforms to the raw data consisting of 1464 rows and 147 columns

2.

```
In [34]: print(data.head(10))
         print()
         print(data.shape)
```

```
       Country  Year  Gender  Life Expectancy at birth   BMI  Alcohol
\
0  Afghanistan  2019    Male                     63.29   NaN    0.003
1  Afghanistan  2019  Female                     63.16   NaN    0.022
2  Afghanistan  2015    Male                     61.04  22.5    0.002
3  Afghanistan  2015  Female                     62.35  24.0    0.014
4  Afghanistan  2010    Male                     59.60  22.1    0.006
5  Afghanistan  2010  Female                     60.30  23.4    0.042
6  Afghanistan  2000    Male                     54.57  21.3    0.009
7  Afghanistan  2000  Female                     55.42  22.1    0.060
8      Albania  2019    Male                     76.25   NaN   11.020
9      Albania  2019  Female                     79.91   NaN    2.530

   Tuberculosis  Syphilis  Chlamydia  Gonorrhoea  ...  Poisonings
Falls  \

0      4.454469  0.050986   0.000000     0.000321  ...    0.057880  0.
620751
1      5.384610  0.043190   0.001424     0.004201  ...    0.325711  0.
284562
2      6.109258  0.056666   0.000000     0.000277  ...    3.980983  0.
056828
3      7.384937  0.047379   0.001201     0.003568  ...    0.310311  0.
322669
4      5.652315  0.051922   0.000000     0.000243  ...    0.087785  0.
697883
5      6.681426  0.044084   0.001053     0.003083  ...    0.417174  0.
619262
6      6.370347  0.046050   0.000000     0.000156  ...    0.077868  0.
519646
7      7.700545  0.039459   0.000651     0.001905  ...    0.358929  0.
407585
8      0.005686  0.000397   0.000000     0.000013  ...    0.005705  0.
064579
9      0.002742  0.000422   0.000120     0.000341  ...    0.003247  0.
060402

   Fire, heat and hot substances  Drowning  Exposure to mechanical fo
rces  \
0                       0.151339  0.801665                       1.54
5577
1                       0.196666  0.194389                       0.05
6229
2                       0.570412  0.151665                       0.76
9096
3                       0.183147  0.251741                       0.05
2141
4                       0.235376  1.370172                       1.61
1014
5                       0.233198  0.727940                       0.07
0492
6                       0.201763  1.081965                       1.16
5209
7                       0.180512  0.622627                       0.04
7653
8                       0.008443  0.019981                       0.03
4043
9                       0.011215  0.005423                       0.00
8541
```

```
   Natural  disasters  Other unintentional  injuries   Self-harm  \
0            0.067079                      2.008284    0.904954
1            0.067360                      1.233210    0.667653
2            1.382456                      0.286633    0.768236
3            0.172981                      1.203843    0.597401
4            0.219533                      2.513913    0.692336


5            0.137334                      1.827513    0.552652
6            0.000000                      2.034185    0.537209
7            0.000000                      1.529516    0.482996
8            0.033273                      0.070418    0.087084
9            0.020578                      0.035055    0.037885


   Interpersonal  violence  Collective violence and legal  intervention
0                 2.595521                                    12.843526
1                 0.621160                                    12.776039
2                 2.553344                                    16.771404
3                 0.576237                                     7.570893
4                 2.233730                                     5.684718
5                 0.404689                                     2.777301
6                 1.786525                                     4.001296
7                 0.245428                                     1.902804
8                 0.072844                                     0.000188
9                 0.029722                                     0.000153

[10 rows x 147 columns]

(1464, 147)
```

- To interpret the data for better understanding, all 147 attributes (columns) are further categorized into a set of 10 columns as shown above.

In [36]:
```
df = data[[ "Country", "Year", "Gender", "BMI", "Alcohol", "Tuberculos
```

In [37]:
```
print(df.columns)
```

```
Index(['Country', 'Year', 'Gender', 'BMI', 'Alcohol', 'Tuberculosis',
       'Syphilis', 'HIV/AIDS', 'Self-harm', 'Stroke', 'Natural disast
ers',


s',

 'Diabetes mellitus', 'Other nutritional deficiencies', 'Preterm birth
 complications', 'Road injury'],
dtype='object')
```

- As the data is huge with 147 columns, it is further narrowed down to 20 selected columns to analyze the data on life expectancy

```
In [16]: df
```

Out[16]:

	Country	Year	Gender	BMI	Alcohol	Tuberculosis	Life Expectancy at birth	Syphilis	HIV/AIDS
0	Afghanistan	2019	Male	NaN	0.003	4.454469	63.29	0.050986	0.315000
1	Afghanistan	2019	Female	NaN	0.022	5.384610	63.16	0.043190	0.148926
2	Afghanistan	2015	Male	22.5	0.002	6.109258	61.04	0.056666	0.238000
3	Afghanistan	2015	Female	24.0	0.014	7.384937	62.35	0.047379	0.105000
4	Afghanistan	2010	Male	22.1	0.006	5.652315	59.60	0.051922	0.156000
...	...	...	...	...	...	...	...	...	...
1459	Zimbabwe	2015	Female	25.3	9.290	0.457023	60.96	0.055791	12.271000
1460	Zimbabwe	2010	Male	22.0	1.470	0.711036	49.58	0.089442	23.781000
1461	Zimbabwe	2010	Female	25.1	7.150	0.464125	53.21	0.065319	27.397000
1462	Zimbabwe	2000	Male	21.7	0.880	2.530362	45.15	0.066511	53.707495
1463	Zimbabwe	2000	Female	24.7	4.220	1.337442	48.12	0.049303	55.155806

1464 rows × 20 columns

- Hence, the data frame with selected 20 columns of attributes was prepared as shown above
- The obtained data set is further pre-processed to check the null values

```
In [17]: df.isnull().sum()

Out[17]: Country                                 0
         Year                                    0
         Gender                                  0
         BMI                                   384
         Alcohol                                20
         Tuberculosis                            0
         Life Expectancy at birth               16
         Syphilis                                0
         HIV/AIDS                                0
         Self-harm                               0
         Stroke                                  0
         Natural disasters                       0
         Epilepsy                                0
         Parkinson disease                       0
         Alzheimer disease and other dementias   0
         Alcohol use disorders                   0
         Diabetes mellitus                       0
         Other nutritional deficiencies          0
         Preterm birth complications             0
         Road injury                             0
         dtype: int64
```

- The selected data contains null values in BMI, life expectancy at birth and alcohol columns. Null values are removed to prevent misinterpretation of data and to obtain and analyze the visuals efficiently. hence the null values are replaced with the mean of BMI and alcohol respectively

```
In [20]: df["BMI"].mean()

Out[20]: 25.05333333333333

In [21]: df["Alcohol"].mean()

Out[21]: 5.883121191135733

In [22]: df["Life Expectancy at birth"].mean()

Out[22]: 70.19941988950276

In [23]: mean_BMI= df['BMI'].mean()
         df['BMI'].fillna(mean_BMI,  inplace=True)
         print(df)

In [24]: mean_Alcohol= df['Alcohol'].mean()
         df['Alcohol'].fillna(mean_Alcohol, inplace=True)
         print(df)

In [25]: mean_BMI= df['Life Expectancy at birth'].mean()
         df['Life Expectancy at birth'].fillna(mean_BMI,  inplace=True)
         print(df)

In [26]: df.to_csv('df.csv',  index=False)
```

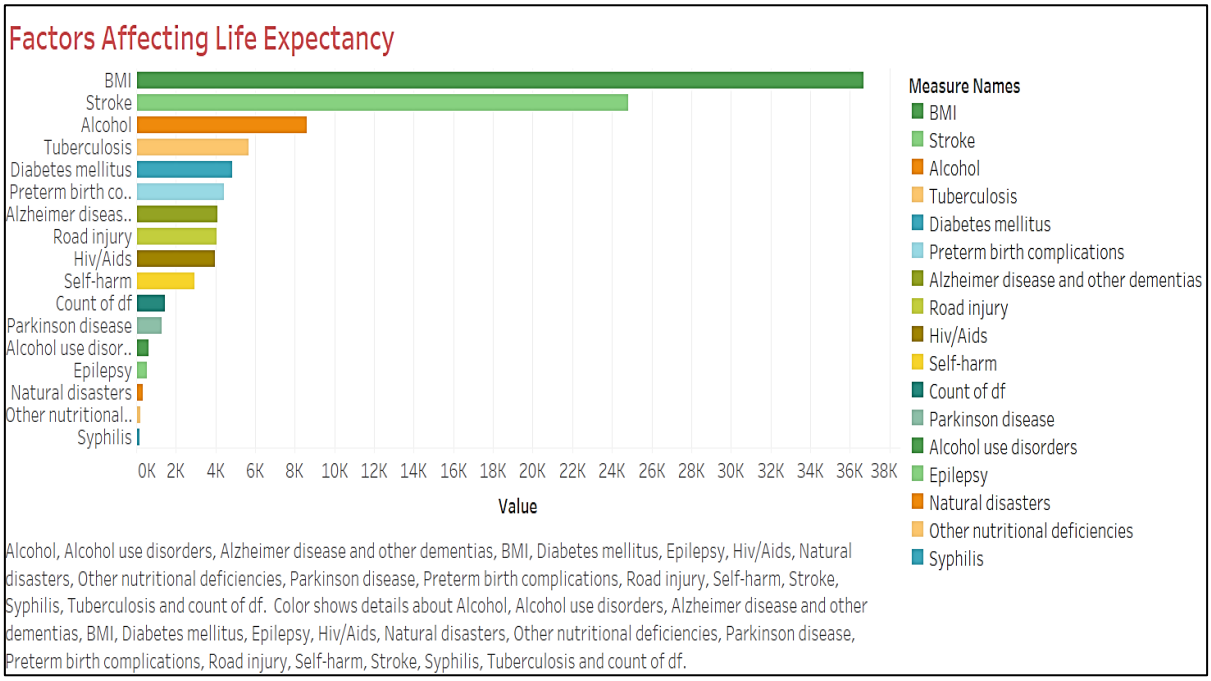
- The above codes are used to replace the null values with the mean of the respective columns. After performing the EDA, the data obtained is cleaned and can be used for further analysis and interpretation of data

Hypothesis

1. Identify the factors affecting life expectancy. Interpret which factors contribute the most and least to the mortality
2. Estimate the mortality in various countries and analyze how life expectancy at birth and preterm birth complications are relating to the life expectancy of people
3. Analyse the wide distribution and proportion of factors affecting the expectancy of life globally in subsequent years and deduce the common factors responsible for mortality with respect to gender

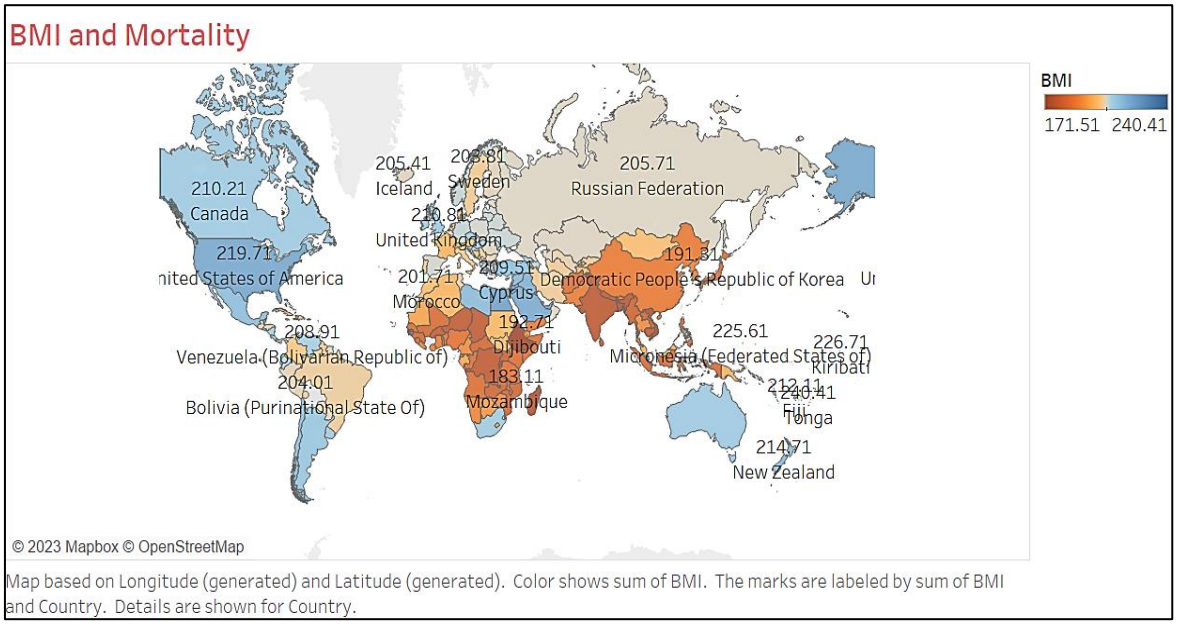
1. **Horizontal bars:** In the given data, there are various factors involved that contribute to the life expectancy of humans. Horizontal bars represent the contribution of individual factors which are further sorted in descending order and coloured to distinguish and identify the leading factors.

- BMI (Body mass index) and syphilis rank first and least respectively contribute to mortality.



2. As we came to know from the above visualization the leading factor for mortality is BMI. Furthermore, we analyze geographically the proportion of its role in different states.

- The geographical map depicts the contribution of BMI in all states globally. From the graph, we can predict that BMI is showing more mortality in the USA and least in Niger. The gradient of colour represents the increasing rate of BMI for mortality. Marks are labeled to represent the total amount of BMI.

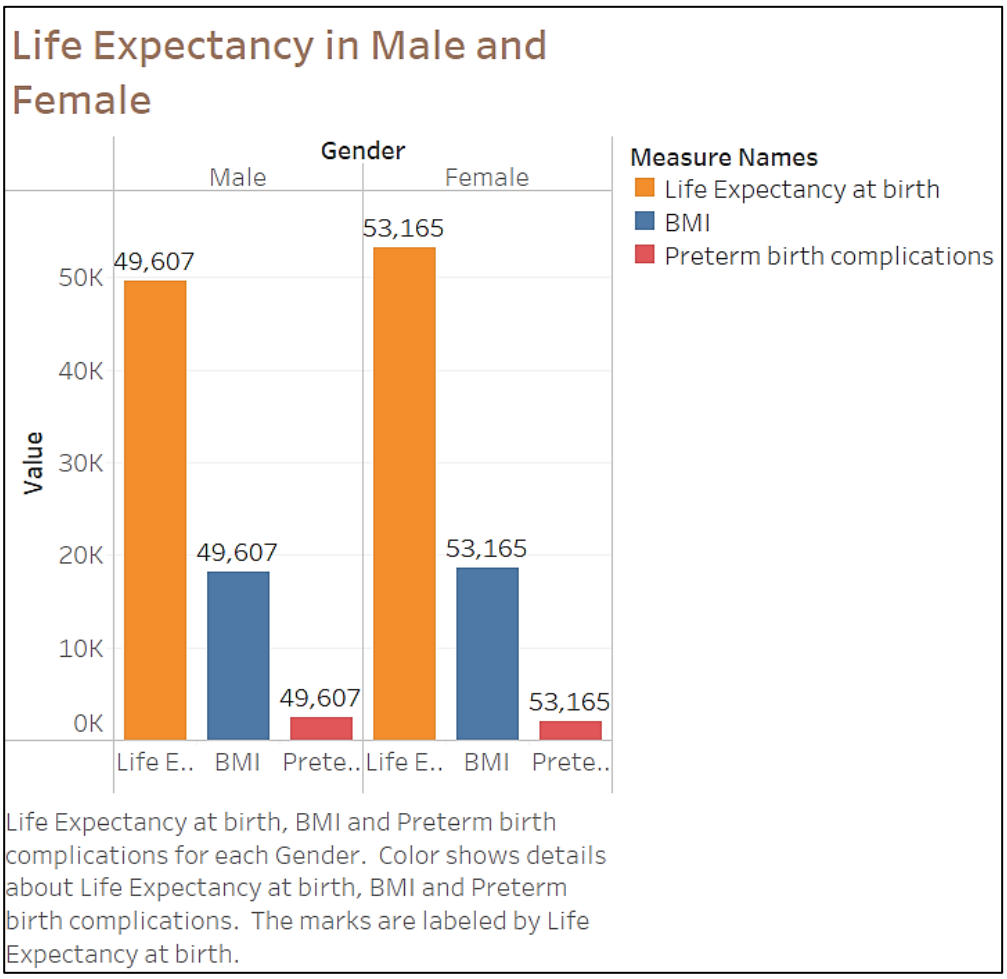


3. Factors such as preterm birth defects, BMI, and life expectancy at birth are some of the major factors contributing to mortality in both males and females.

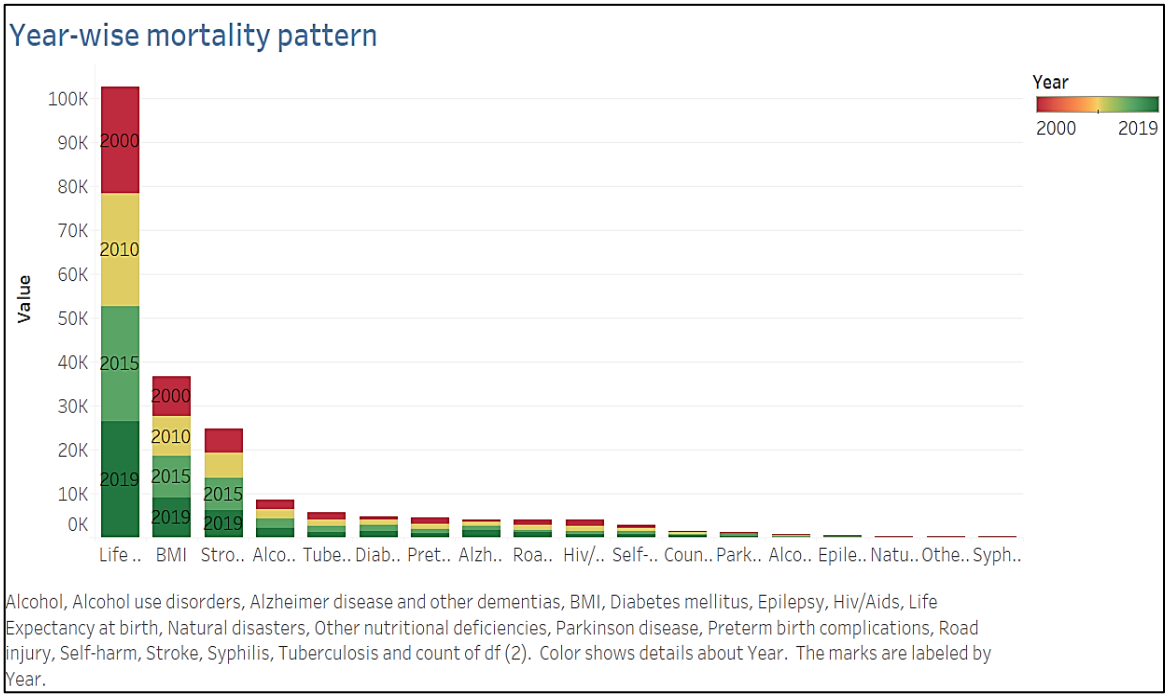
- The bar graph represents and compares the contribution of these factors in both genders. The below visual depicts that females are more prone than males. i.e. the mortality is more in females than in males due to the above factors.

- Furthermore, among the three factors life expectancy at birth contributes more followed by BMI and preterm birth defects.

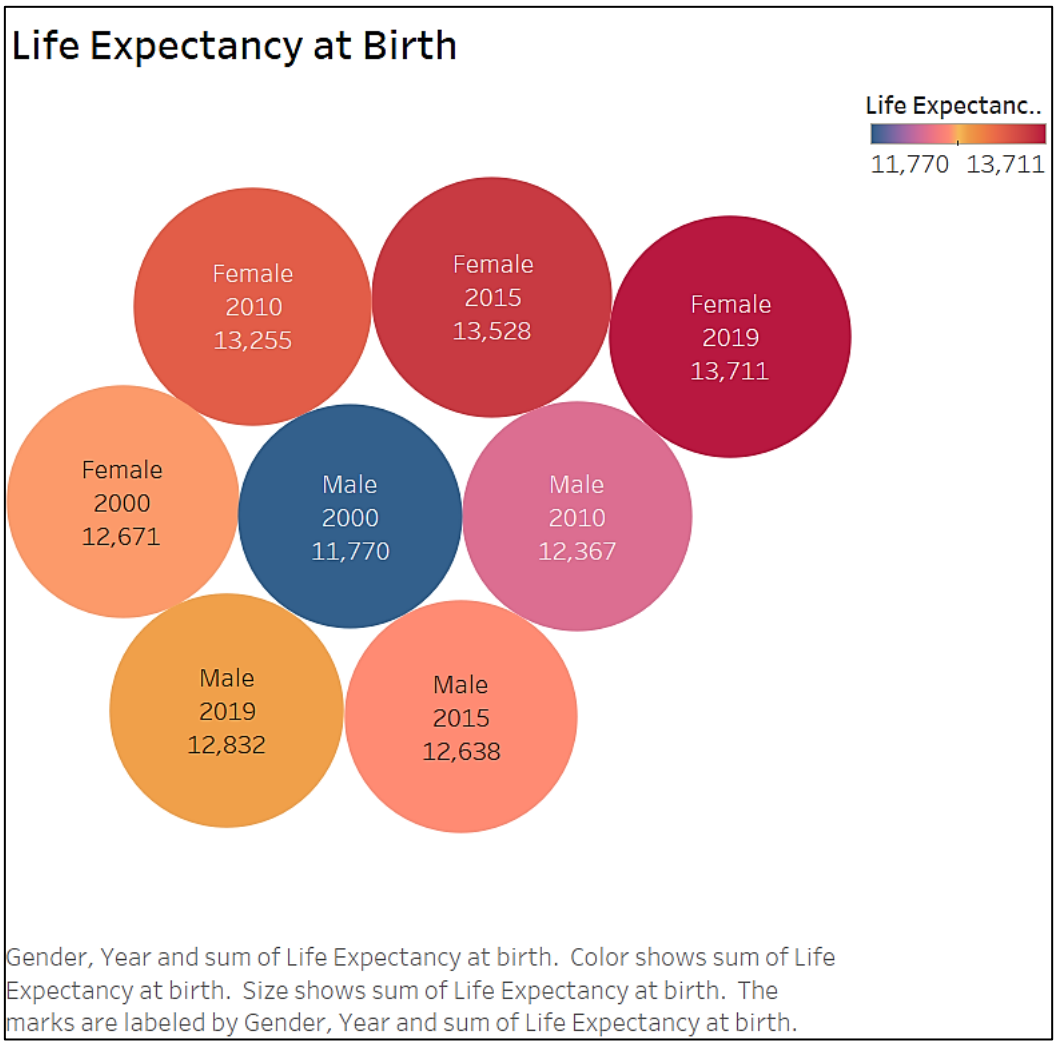
- The color shows the discrete data and differentiates visually and pictorially the major factors. Marks are further labelled by life expectancy at birth as it is the most contributed factor with respect to gender



4. The below graph depicts the year-wise pattern of the mortality rate for years 2000, 2010, 2015, and 2019. Colour represents various years visually. Factors are sorted in ascending order which explains the leading factors to mortality.

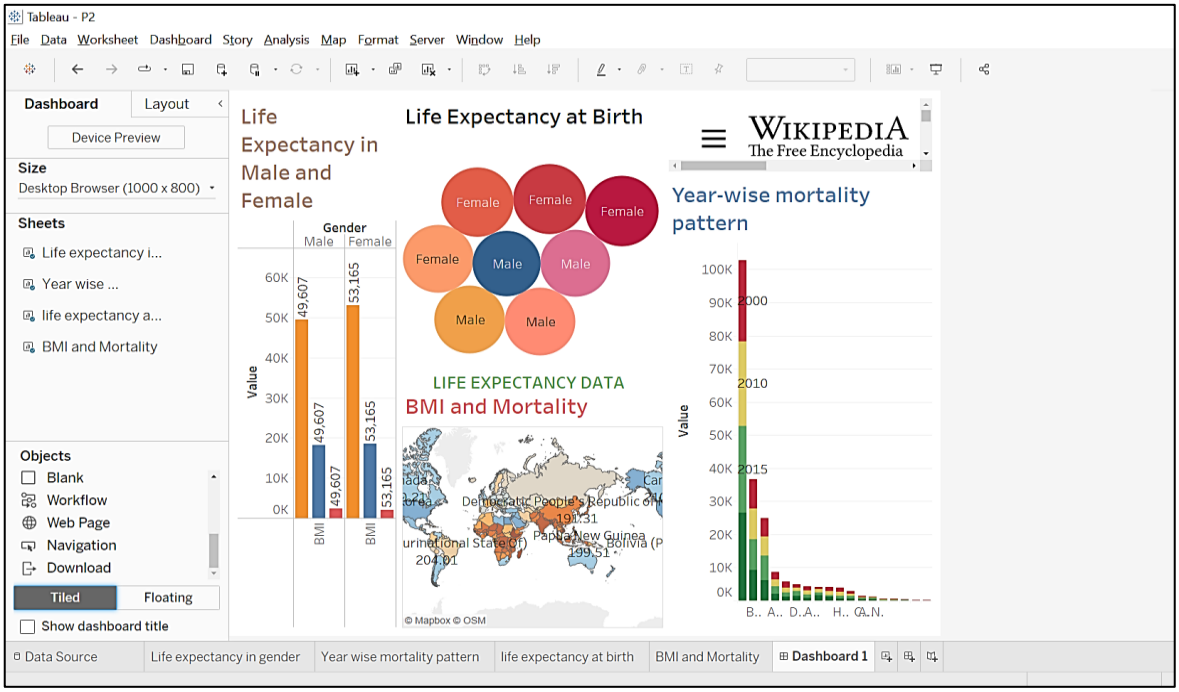


5. The above analysis is furthermore supported by bubble chart representation which depicts, mortality due to life expectancy is more in the year 2019 and further mortality is more in females than in males. In the below graph, the increase in the size of the circle represents the sum of life expectancy and the gradient of colour also represents the sum of life expectancy. To understand more marks are labelled with gender and year.



6. An interactive dashboard was created with the above graphs which vividly illustrates the various factors affecting life expectancy visually and clearly. A text box was inserted which represents the data and objects such as text box and weblink were used to understand more about mortality and Life expectancy

- The below image depicts the created dashboard:



Conclusion:

This project encapsulates the various factors affecting life expectancy in the years 2000, 2010, 2015, and 2019**.** EDA was conducted to concise the data accordingly to understand and analyze the data. The visuals depict the most affecting factors on mortality every year and also among the gender. The geographical distribution of these factors in all states was also observed. An interactive dashboard was created that summarises and analyses all the respective graphs. The data shows mortality in females was more and BMI, life expectancy at birth, and preterm birth defects are the most important factors affecting the life expectancy rate in humans.

References:

1. Global health estimates: Leading causes of death https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

2. GHE: Life expectancy and healthy life expectancy https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy

3. Global health estimates: Leading causes of DALYs https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/global-health-estimates-leading-causes-of-dalys.