Analysis of Death Risk Factors

By Modika Ishwarya

Abstract:

In the world, there are various causes of death. People are dying due to several risk factors and combinations of risk factors. It is crucial to understand the effect of the risk factors on people and prevent them. The main objective of this analysis is to understand the deaths that occurred due to the risk factors and their combinations.

This analysis can also help to understand the leading risks for each country and territory. So, it helps countries take effective measures to prevent the future risk of these factors on the people.

Data Collection:

The dataset is collected from www.kaggle.com. The data consists of information about deaths due to 29 risk factors from 1990 to 2017 in 230 countries and including the world.

Approach:

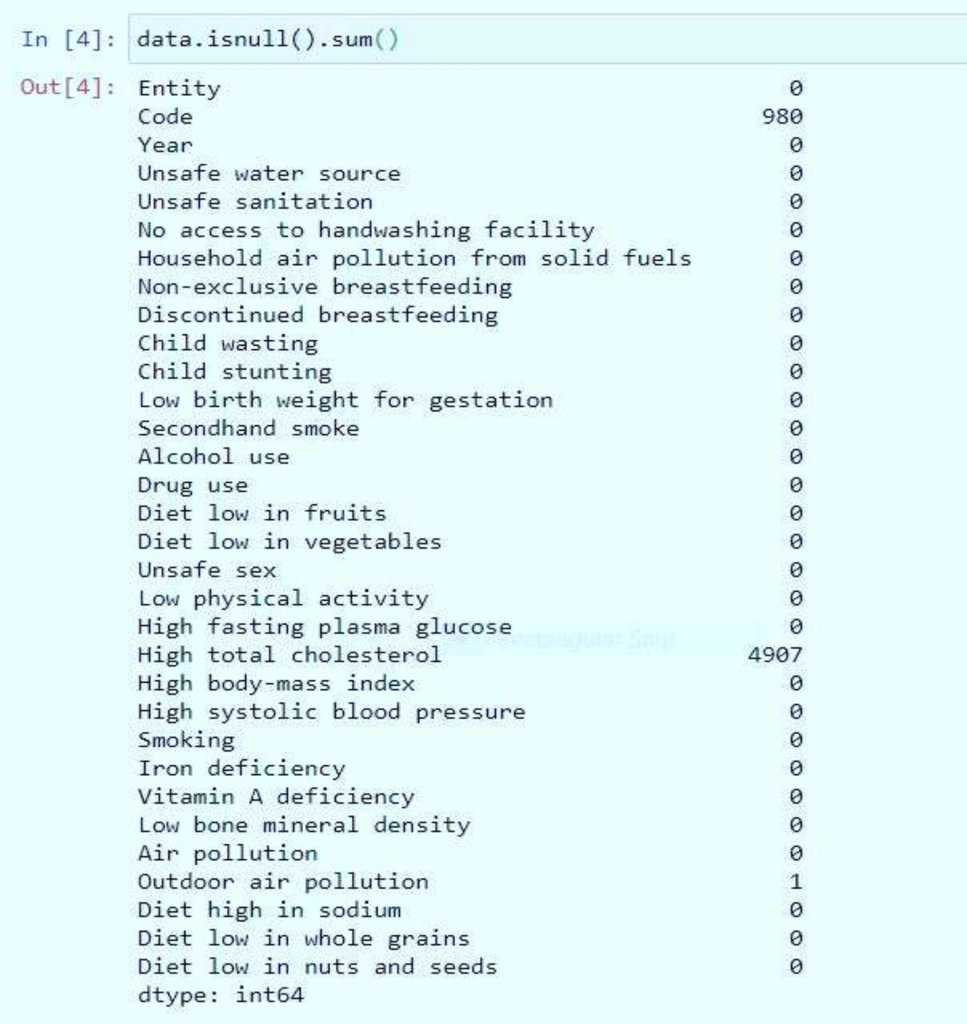
- ❖ The python libraries used are Pandas, Matplotlib, Seaborn, Numpy, Scikit-learn.
  - ➢ Numpy is an external library in Python. It helps to work with arrays. It has an array object ndarray. The use of Numpy is to perform mathematical operations on arrays.
  - ➢ Pandas is used for data cleaning and data analysis purpose.
  - ➢ Matplotlib is used for data visualization.
  - ➢ Seaborn is a data visualization library formed based on matplotlib. It is used to visualize random distributions.
  - ➢ Scikit-learn provides a selection of efficient tools for machine learning and statistical modeling.
- ❖ As part of data analysis, the data cleaning is performed to remove noise or missing values present if any.
- ❖ The statistical measures considered for analysis of data are mean, standard deviation.

- ❖ Regression analysis is considered for prediction.

Data Analysis & Visualization:

1. **Checking for missing values**
   - **Purpose:**

- As part of data cleaning, it is important to check whether data is having any missing values as the presence of missing values affects the data analysis and may lead to wrong conclusions.



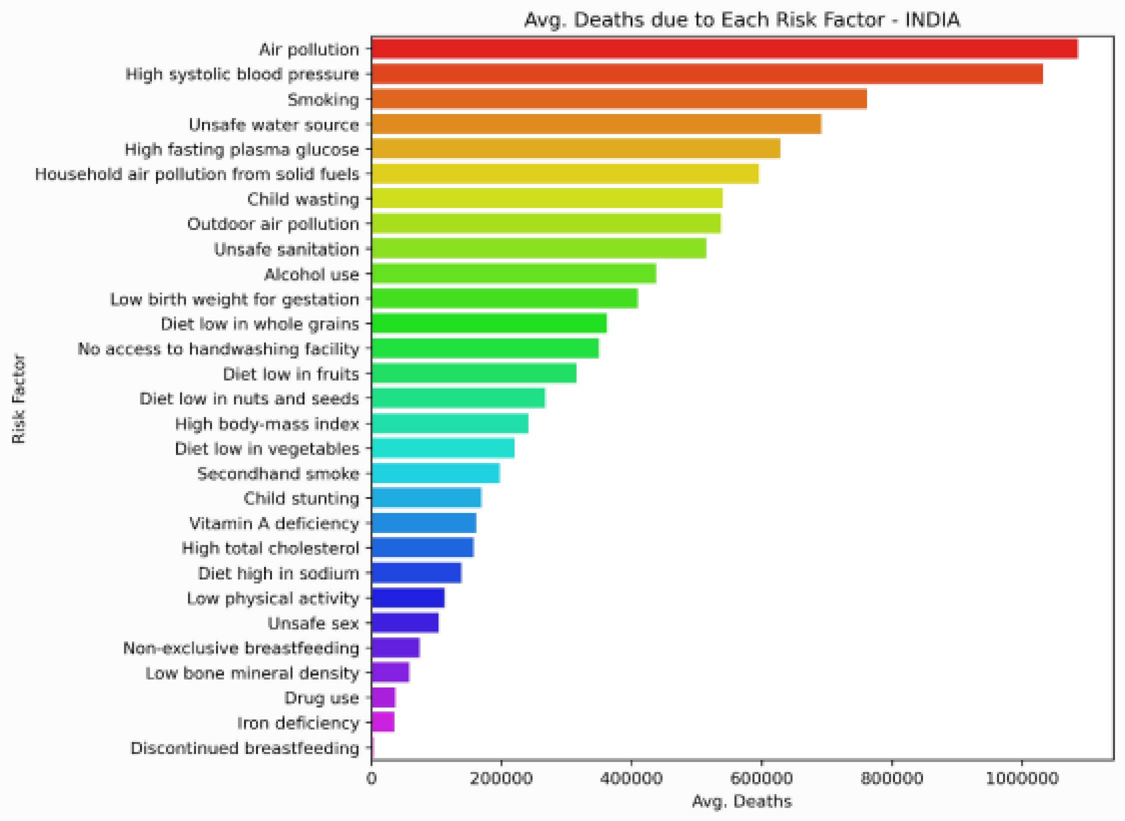
```
In [4]:  data.isnull().sum()

Out[4]:  Entity                                        0
         Code                                        980
         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
```

- **Inference:**
  - There are missing values in the data.
  - They are to be filled or removed based on the data.

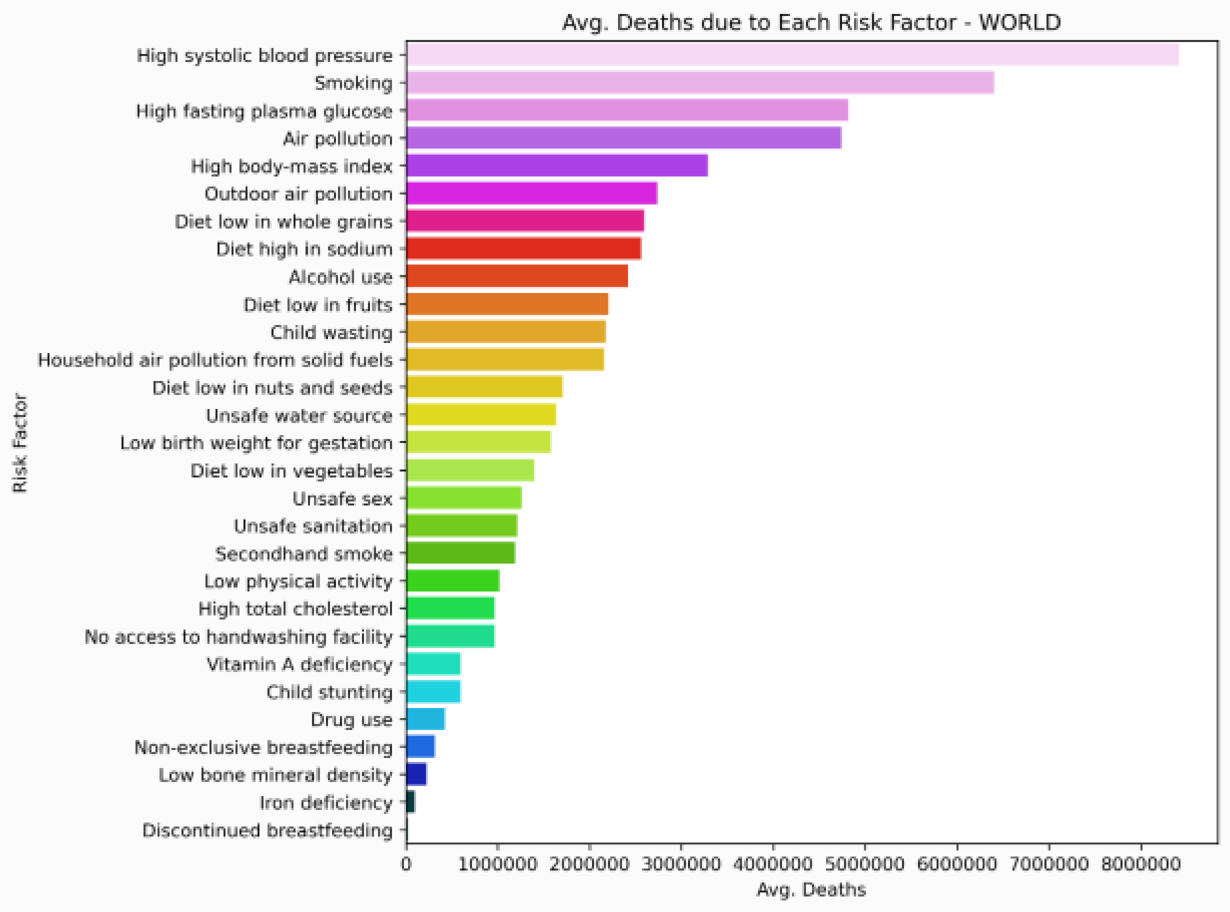
2. **Filling Missing Values:**
   - The missing values in the column High total cholesterol and Outdoor air pollution are filled with the mean of the respective columns.
   - The missing values in the Code column are filled with the first three characters of the Entity column.

3. **Average Deaths due to Each Risk Factor in India & World:**
   - **Purpose:**

- To obtain the details regarding the deaths due to each factor in India as well as the world from 1990 to 2017.
- To understand the effect of risk factors in India & the World.



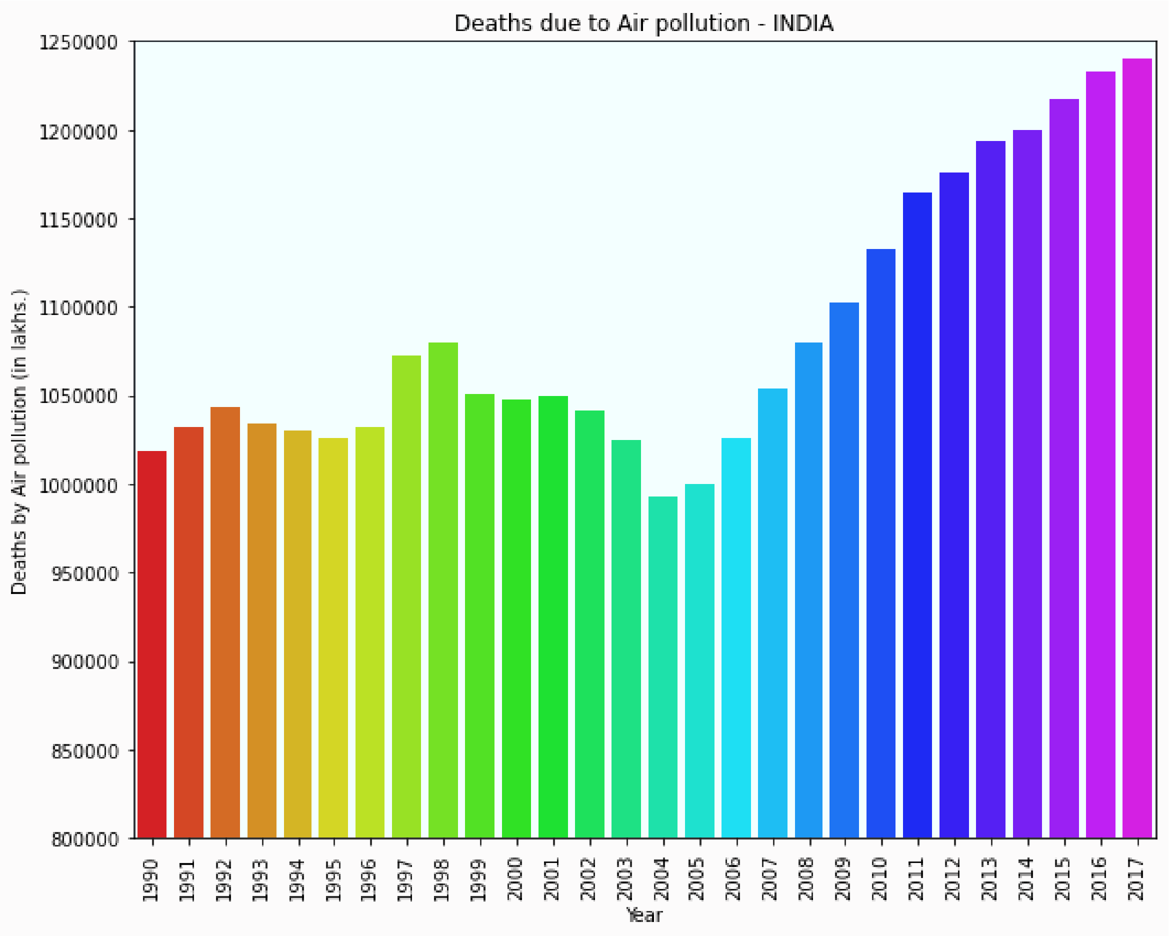
- **Inference:**
  - The top five major risk factors for deaths in India are Air pollution, High systolic blood pressure, Smoking, Unsafe water source, High fasting plasma glucose.
  - The highest deaths in India are recorded due to are Air pollution, High systolic blood pressure, Smoking.
  - The least deaths are recorded due to Discontinued breastfeeding.



- **Inference:**
  - ➢ The top five major risk factors for deaths in World are High systolic blood pressure, Smoking, High fasting plasma glucose, Air pollution, High body-mass index.
  - ➢ The highest number of deaths in the World are recorded due to are High systolic blood pressure, Smoking, High fasting plasma glucose.
  - ➢ The least deaths are recorded due to Discontinued breastfeeding.

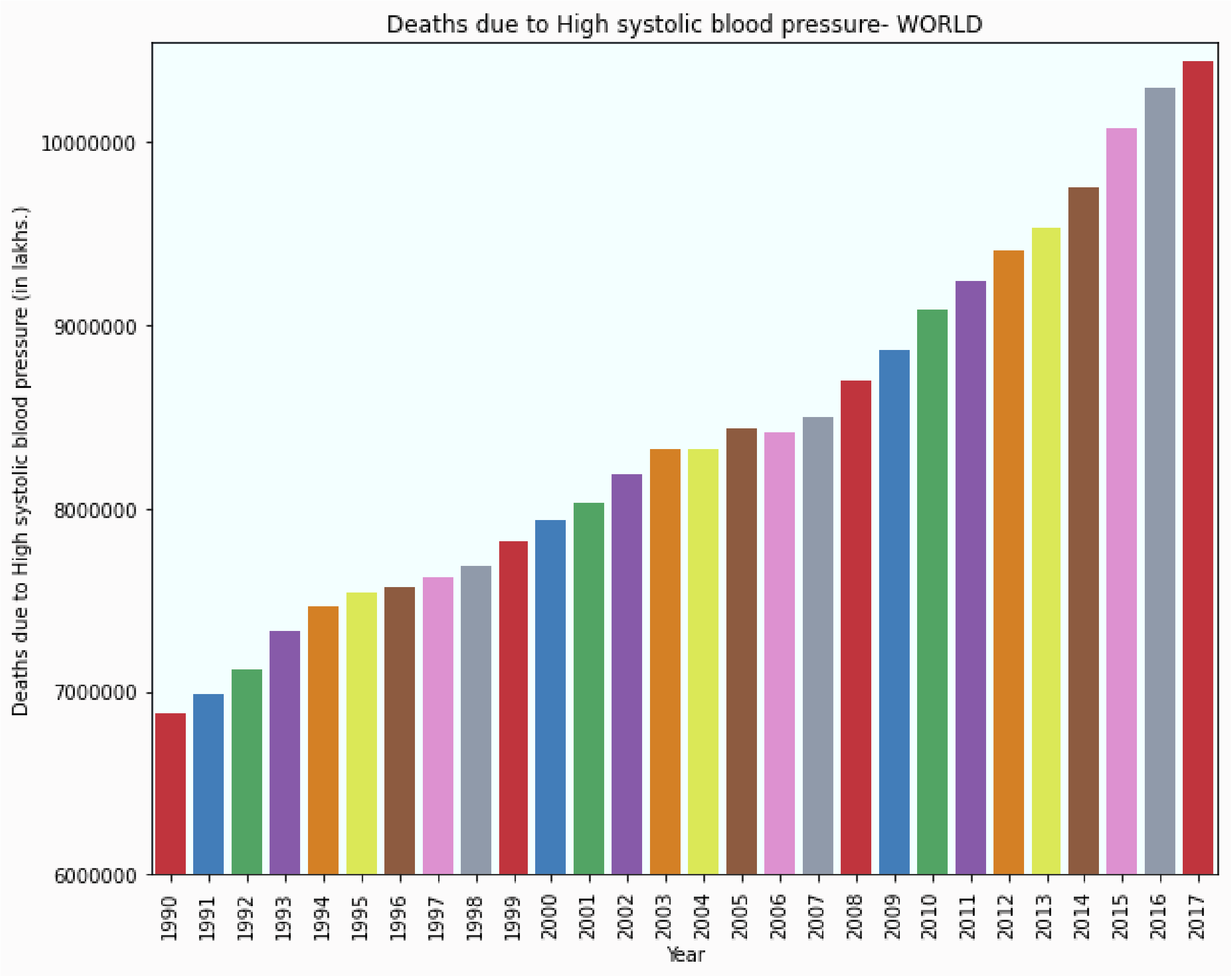
4. **The number of deaths due to Air pollution in India:**

- **purpose:**
  - ➢ As more deaths are caused due to Air pollution in India, it is necessary to understand how its effect was from 1990 to 2017.



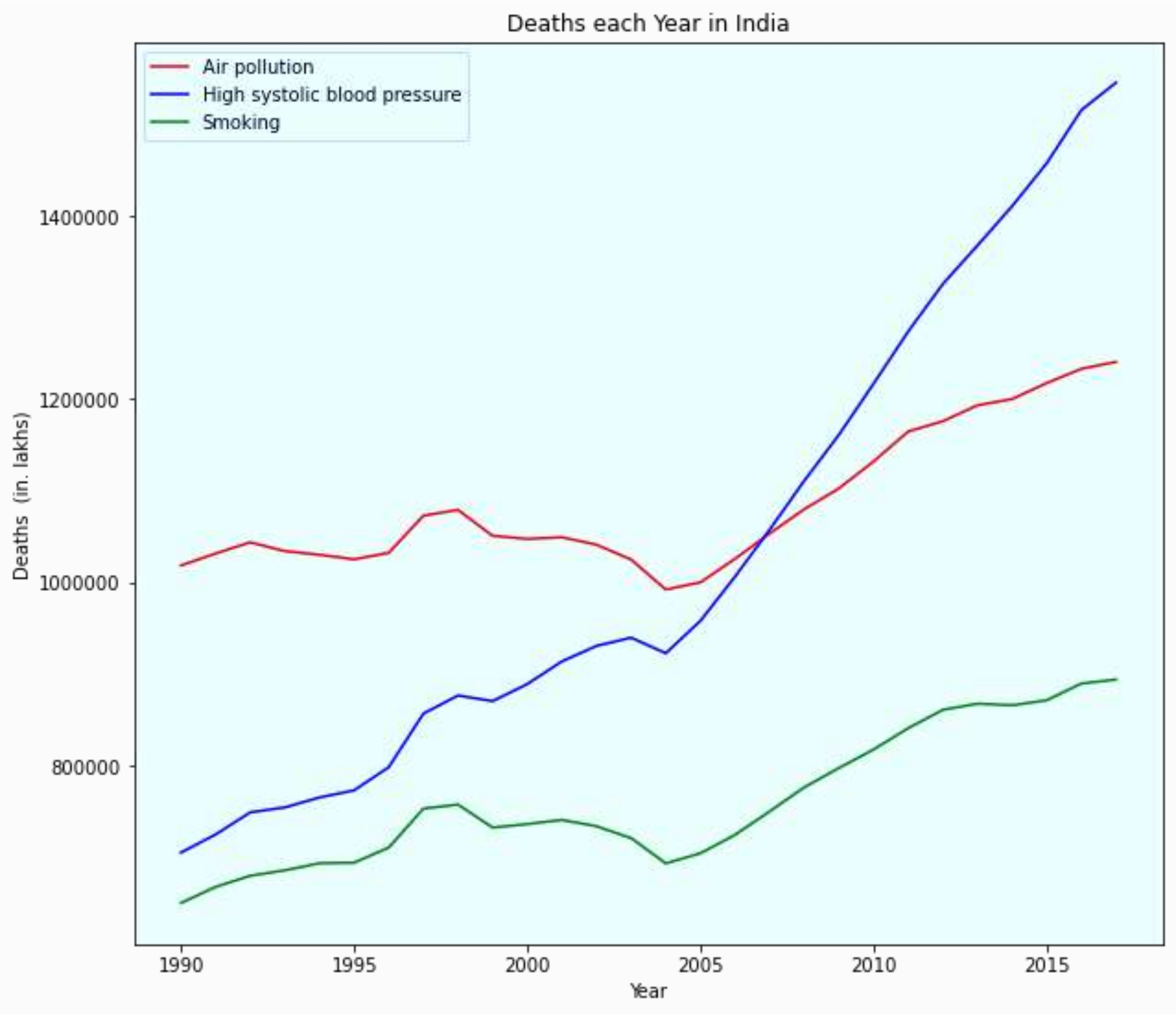
- **Inference:**
  - In India, the highest number of Deaths due to Air pollution was in 2017 and the lowest was in 2004.

5. **The number of deaths due to High systolic blood pressure in the World:**
   - **purpose:**
     - As the first dreadful factor in the World is High systolic blood pressure, it is crucial to get an insight into the effect of this risk factor on the World from 1990 t0 2017.



- **Inference:**
  - In World, the highest number of Deaths due to High systolic blood pressure was in 2017 and the lowest was in 1990.

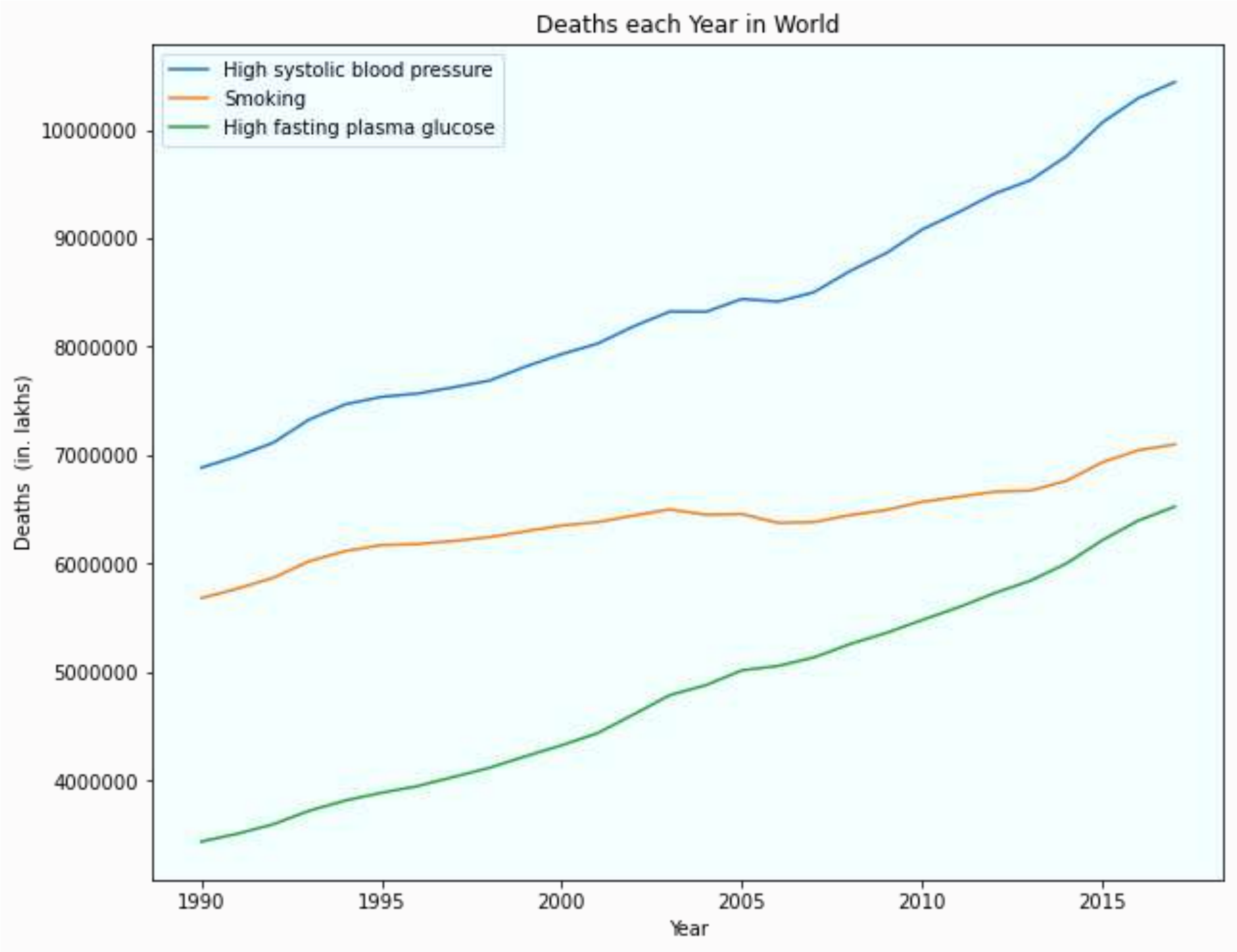
6. **The number of deaths in India due to Air pollution, High systolic blood pressure, Smoking:**
   - **Purpose:**
     - The top three major risk factors for deaths in India are Air pollution, High systolic blood pressure, Smoking.
     - So, It is important is compare and understand their effect from 1990 to 2017.



- **Inference:**
  - In India, the effect of Air Pollution is consistent in comparison with the effect of Smoking & High systolic blood pressure from 1990 to 2017.

7. **The number of deaths in World due to Air pollution, High systolic blood pressure, Smoking:**

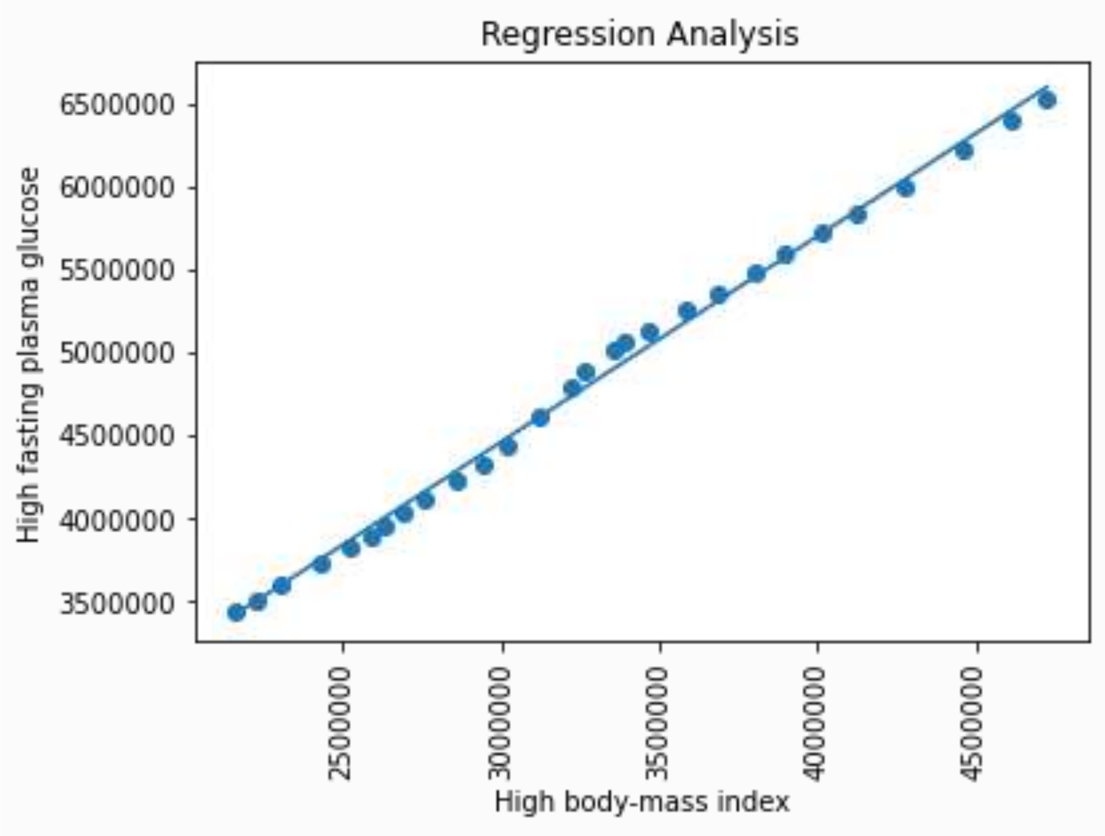
- **Purpose:**
  - The top three major risk factors for deaths are High systolic blood pressure, Smoking, High fasting plasma glucose.
  - So, It is important is compare and understand their effect from 1990 to 2017.



- **Inference:**
  - ➢ In World, the number of deaths due to High systolic blood pressure consistently increased from 1990 to 2017.
  - ➢ This means the effect of the risk factor(High systolic blood pressure) kept increasing from 1990 to 201

8. **Linear Regression Analysis of High body-mass index & High fasting plasma glucose(World data):**

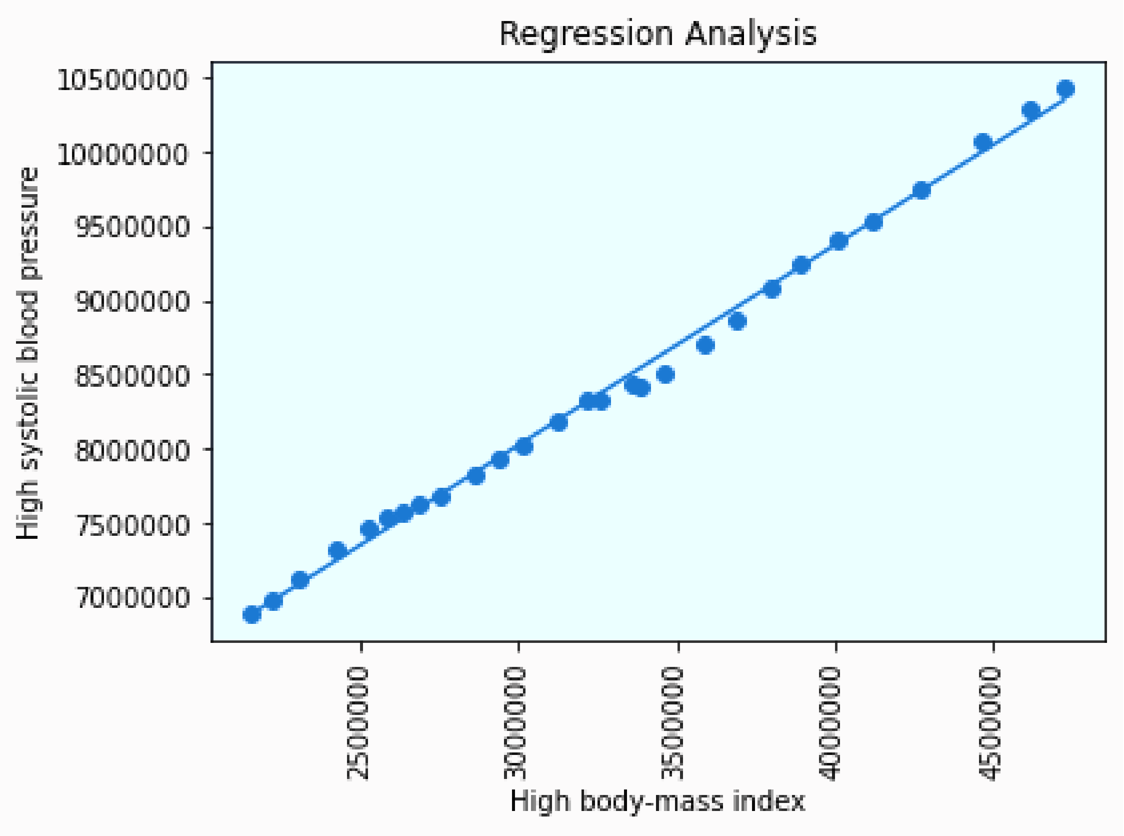
- **Purpose:**
  - ➢ To understand the relationship between the risk factors High body mass index & High fasting plasma glucose.
  - ➢ To predict the dependent variable based on the values of the independent variable.



- **Inference:**
  - ➢ The regression coefficient is 1.24195538
  - ➢ There is a positive correlation between High body mass index & High fasting plasma glucose.
  - ➢ The coefficient of regression is 1.243 which indicates that if the effect of the High body-mass index on deaths increases by 1 unit then the mean effect of High fasting plasma glucose on deaths increases by 1.243 units.
  - ➢ R-squared is a goodness-of-fit measure for linear regression models.

- ➢ R-squared measures the strength of the relationship between your model and the dependent variable.
- ➢ The r2_score value is 0.9959, which indicates that the model is 99% accurate that it fits well to the data and the level of correlation is high.

9. **Linear Regression Analysis of High body-mass index & High systolic blood pressure(World data):**
   - **Purpose:**
     - ➢ To understand the relationship between the risk factors High body-mass index & High systolic blood pressure.
     - ➢ To predict the dependent variable based on the values of the independent variable.



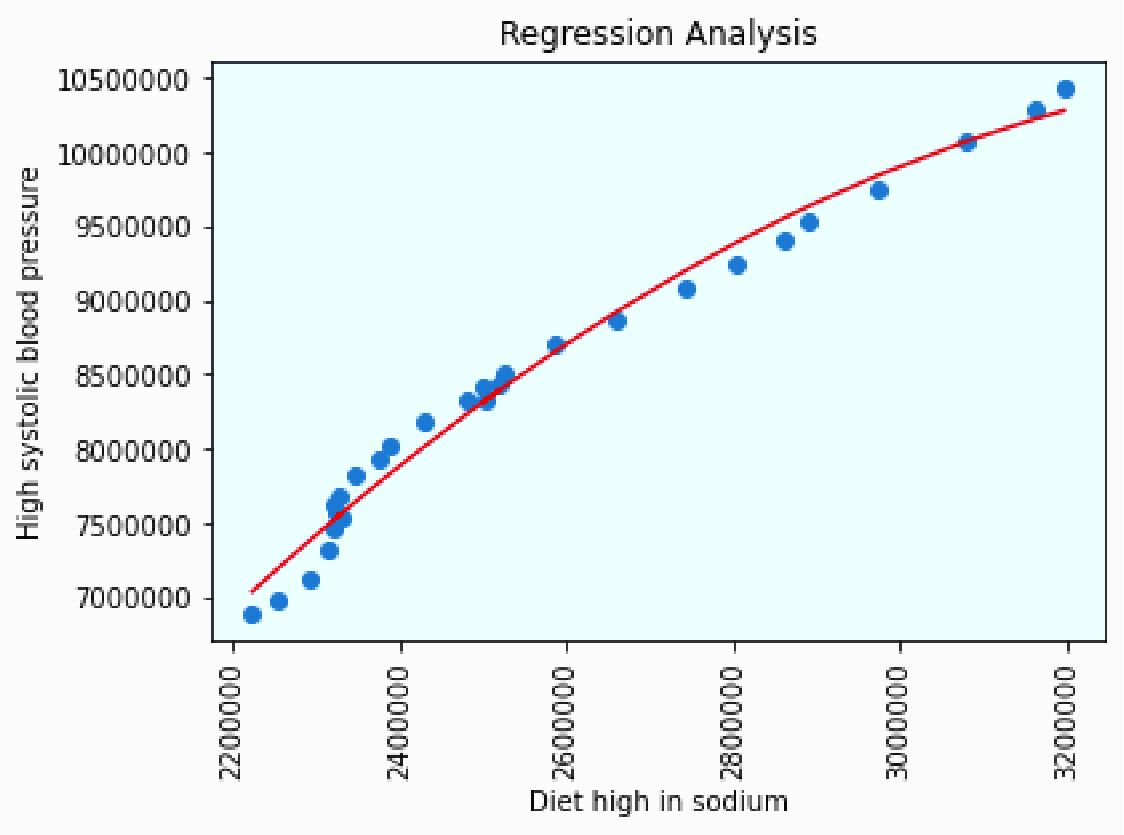
   - **Inference:**
     - ➢ The regression coefficient is 1.34933723.
     - ➢ There is a positive correlation between High body mass index & High systolic blood pressure.
     - ➢ The coefficient of regression is 1.349 which indicates that if the effect of the High body-mass index on deaths increases by 1 unit then the

mean effect of High systolic blood pressure on deaths increases by 1.349 units.

- R-squared is a goodness-of-fit measure for linear regression models.
- R-squared measures the strength of the relationship between your model and the dependent variable.
- The r2_score value is 0.99582, which indicates that the model is 99% accurate that it fits well to the data and the level of correlation is high.

10. **Polynomial Regression for Diet high in sodium & High systolic blood pressure(World data)**:

- **Purpose**:
  - The data is curvilinear.
  - To model a non-linear relationship between the independent(Diet high in sodium) and dependent variable(High systolic blood pressure).

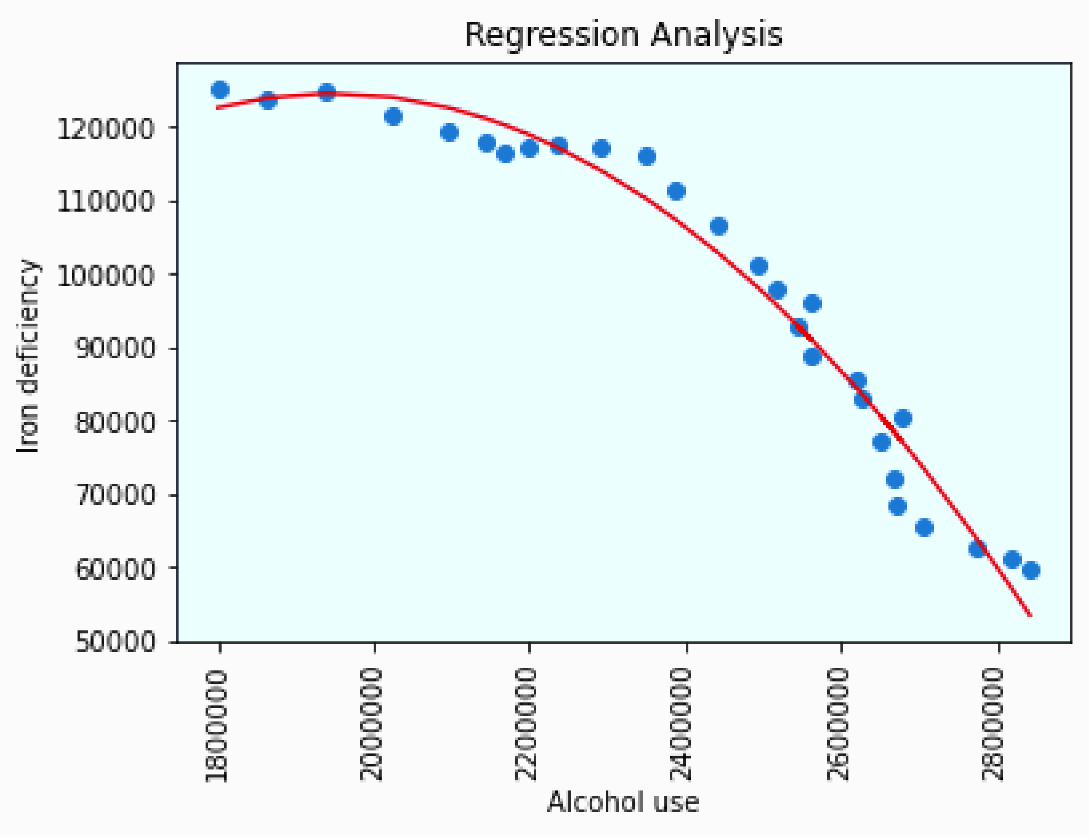


- **Inference:**
  - R-squared is a goodness-of-fit measure for regression models.
  - R-squared measures the strength of the relationship between your model and the dependent variable.

- ➢ The r2_score value is 0.983096, which indicates that the model is 98% accurate that it fits well to the data and the level of correlation is high.

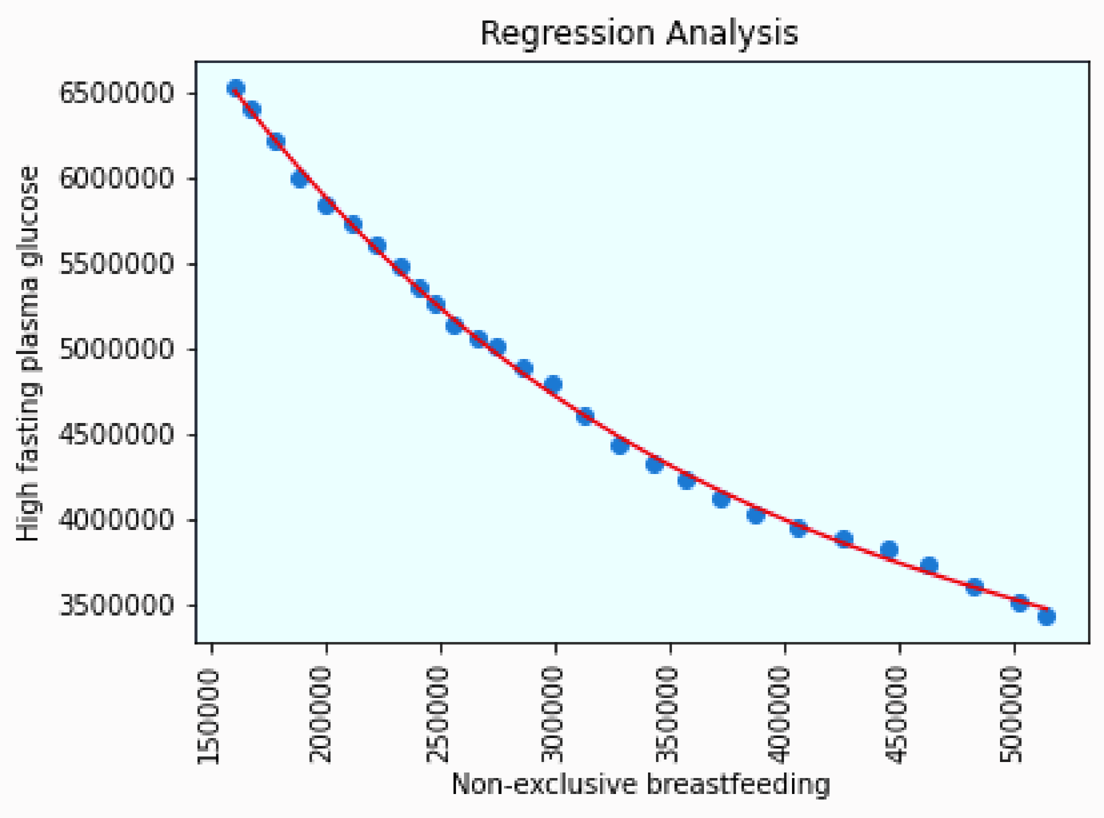
11. **Polynomial Regression for Alcohol use & Iron deficiency(World data):**

- **Purpose**:
  - ➢ The data is curvilinear.
  - ➢ To model a non-linear relationship between the independent(Alcohol use) and dependent variable(Iron deficiency).



- **Inference:**
  - ➢ R-squared is a goodness-of-fit measure for regression models.
  - ➢ R-squared measures the strength of the relationship between your model and the dependent variable.
  - ➢ The r2_score value is 0.9668, which indicates that the model is 96% accurate that it fits well to the data and the level of correlation is high.

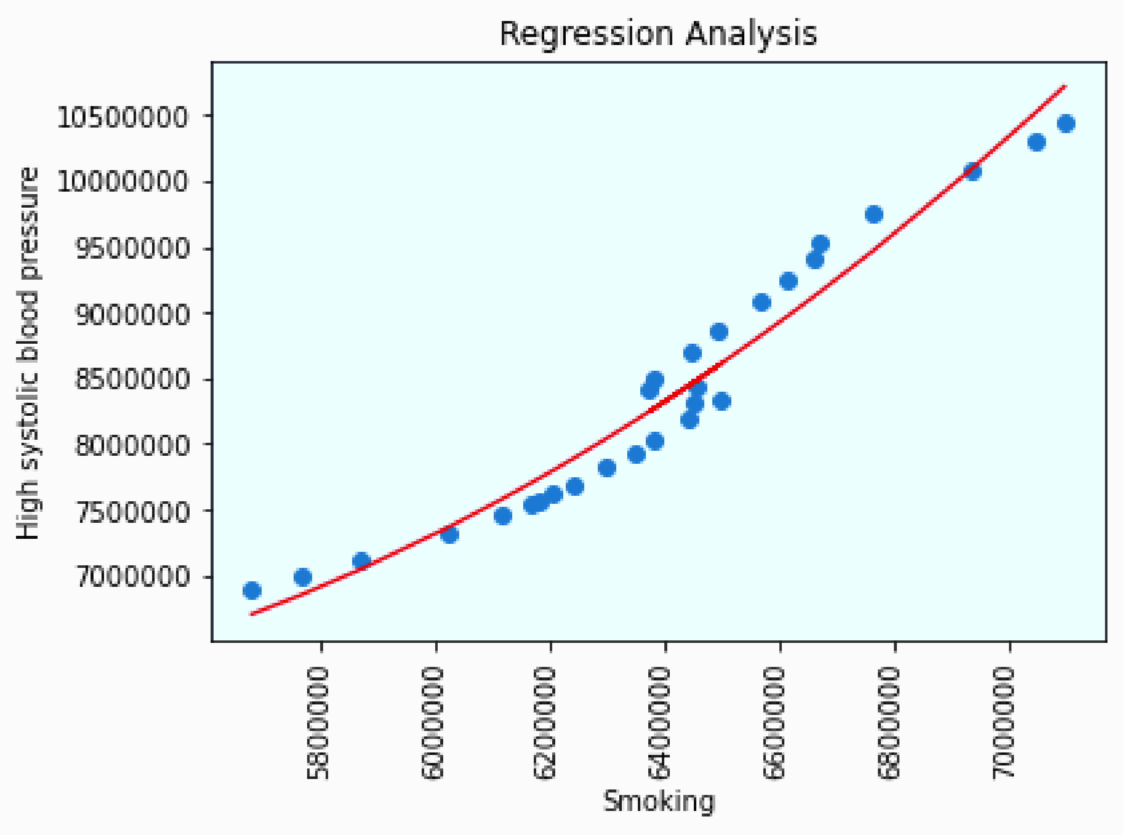
12. **Polynomial Regression for Non-exclusive breastfeeding, High fasting plasma glucose(World data):**
    - **Purpose**:
        - ➢ The data is curvilinear.
        - ➢ To model a non-linear relationship between the independent(Non-exclusive breastfeeding) and dependent variable(High fasting plasma glucose).



- **Inference:**
    - ➢ R-squared is a goodness-of-fit measure for regression models.
    - ➢ R-squared measures the strength of the relationship between your model and the dependent variable.
    - ➢ The r2_score value is 0.9986220451166413, which indicates that the model is 99% accurate that it fits well to the data and the level of correlation is high.

13. **Polynomial Regression for Smoking & High systolic blood pressure(World data):**
    - The data is curvilinear.

- To model a non-linear relationship between the independent(Smoking) and dependent variable(High systolic blood pressure).



- **Inference:**
  - ➢ R-squared is a goodness-of-fit measure for regression models.
  - ➢ R-squared measures the strength of the relationship between your model and the dependent variable.
  - ➢ The r2_score value is 0.95188, which indicates that the model is 95% accurate that it fits well to the data and the level of correlation is high.

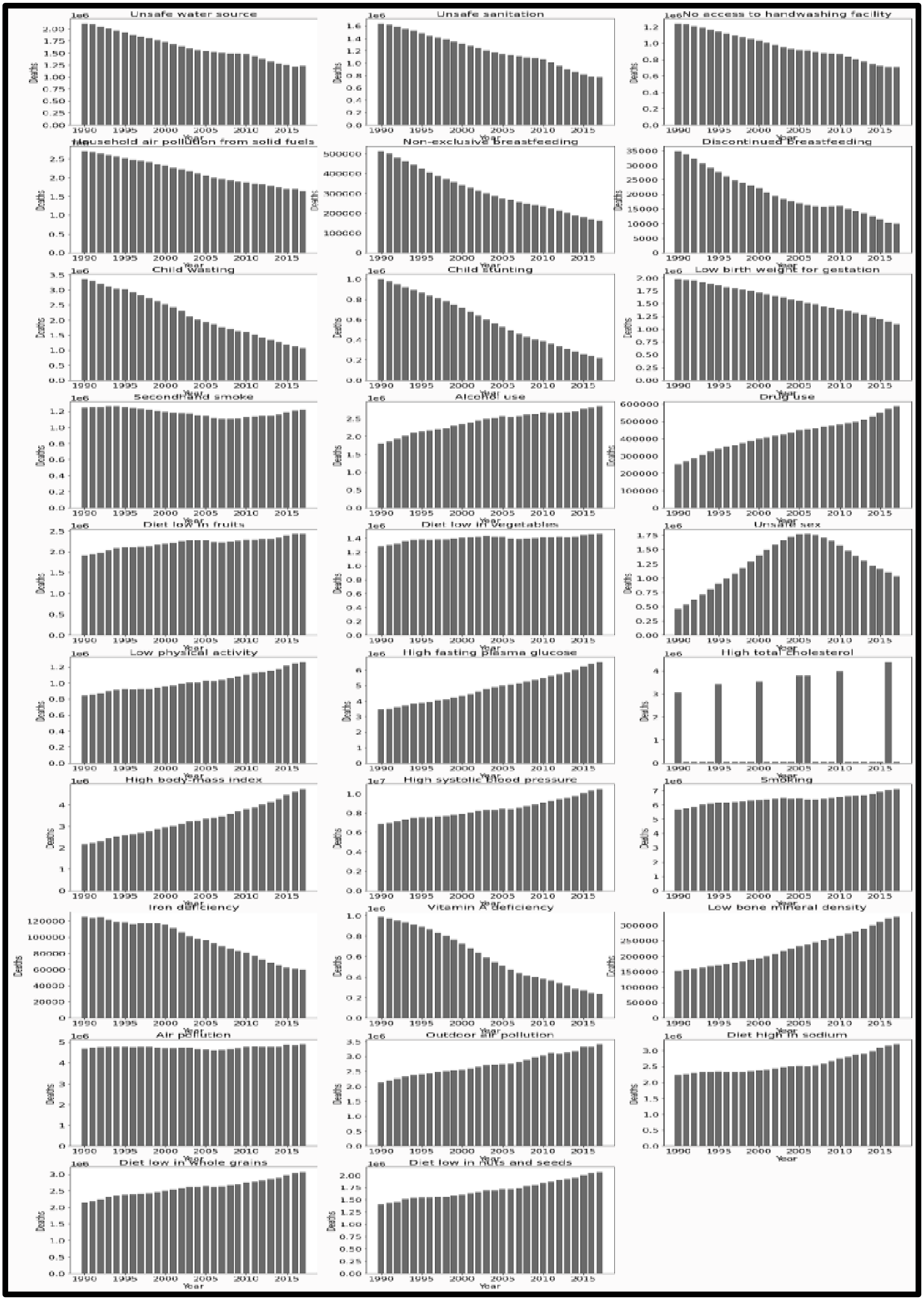
14. Multiple Regression Analysis of Independent(Outdoor air pollution, Household air pollution) and (dependent)Air pollution:

-
  - ➢ To have a better prediction from multiple predictors.
  - ➢ To understand the relationship between a single dependent variable and multiple independent variables.
- Inference:
  - ➢ There is a positive correlation between Outdoor air pollution and Air pollution. The coefficient of regression is 0.94338962, which

indicates that if the effect of Outdoor air pollution on deaths increases by 1 %
then the mean effect of Air pollution on deaths increases by 9.4%.

  - ➢ There is a positive correlation between Household air pollution from solid fuels and Air pollution. The coefficient of regression is 0.92075656, which indicates that if the effect of Outdoor air pollution on deaths increases by 1 % then the mean effect of Air pollution on deaths increases by 9.2%.
  - ➢ The r2_score value is 0.99, which indicates that the model is 99% accurate that it fits well to the data and the level of correlation is high.

15. **Multiple Regression Analysis between independent variables(Unsafe water source, No access to handwashing facility) dependent variable(Unsafe sanitation):**
    - **Purpose**:
      - ➢ To have a better prediction from multiple predictors.
      - ➢ To understand the relationship between a single dependent variable and multiple independent variables.
    - **Inference:**
      - ➢ There is a negative correlation between Unsafe water sources and Unsafe sanitation. The coefficient of regression is -0.5706703, which indicates that if the effect of Unsafe water sources on deaths increases by 1 % then the mean effect of Unsafe sanitation on deaths decreases by 5.7%.
      - ➢ There is a positive correlation between No access to handwashing facility and Unsafe sanitation. The coefficient of regression is 2.52404146, which indicates that if the effect of No access to handwashing facility on deaths increases by 1 % then the mean effect of Unsafe sanitation on deaths increases by 25%.
      - ➢ The r2_score value is 0.99, which indicates that the model is 99% accurate that it fits well to the data and the level of correlation is high.
16. **Deaths analysis in World due to each risk factor:**
    - **Purpose:**
      - ➢ To obtain an insight into the effect of each risk factor in world


Unsafe water source
1e6
Unsafe sanitation
1e6
No access to handwashing facility
1e6
Household air pollution from solid fuels
1e6
Non-exclusive breastfeeding
Discontinued breastfeeding
Child wasting
1e6
Child stunting
1e6
Low birth weight for gestation
1e6
Secondhand smoke
1e6
Alcohol use
1e6
Drug use
Diet low in fruits
1e6
Diet low in vegetables
1e6
Unsafe sex
1e6
Low physical activity
1e6
High fasting plasma glucose
1e6
High total cholesterol
1e6
High body-mass index
1e6
High systolic blood pressure
1e7
Smoking
1e6
Iron deficiency
Vitamin A deficiency
1e6
Low bone mineral density
Air pollution
1e6
Outdoor air pollution
1e6
Diet high in sodium
1e6
Diet low in whole grains
1e6
Diet low in nuts and seeds
1e6
Deaths
Year
1990
1995
2000
2005
2010
2015

- **Inference:**
    - ❖ Each risk factor has shown its effect on people in the world.
    - ❖ The High total cholesterol effect was not continuous.

Conclusion:

- ❖ The highest deaths in India are recorded due to are High systolic blood pressure, Smoking, High fasting plasma glucose.
- ❖ The least deaths are recorded due to Discontinued breastfeeding.
- ❖ The highest number of deaths in the World are recorded due to are High systolic blood pressure, Smoking, High fasting plasma glucose.
- ❖ The least deaths are recorded due to Discontinued breastfeeding.