**SDG 3: GOOD HEALTH AND WELL-BEING**

**A Report Submitted by:**

**Pratham Bansal (2023 Btech060)**

**Menanki Shekhawat (2023 Btech48)**

**Prateek Yadav (2023 Btech059)**

**Under the Supervision of**

**Dr. Rabia Kamra**

**In Partial Fulfilment of**

**The Requirements of the course of**

**AS1117 (Probability and Statistics)**

**To The**

******

**Institute of Engineering And Technology (IET)**

**J.K Lakshmipat University Jaipur**

**(November 27, 2024)**

**CONTENT**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Content** | **Page no.** |
| 1. | List of Figures | 3 |
| 2. | List of Tables | 4 |
| 3. | Introduction | 5-6 |
| 4. | Literature Review | 7-8 |
| 5. | Objectives | 9-11 |
| 6. | Methodology | 12-15 |
| 7. | Statistical Analysis   * Objective 1 * Objective 2 * Objective 3 | 16-22  23-28  29-32 |
| 8. | References | 33 |
| 9. | Link of Dataset Used | 33 |
| 10. | Appendix | 34-46 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Description** | **Page No.** |
| 1. | **Fig 1.1 a:** Maternal Mortality Rate (2012-2020) | 16 |
| 2. | **Fig 1.1 b:** Maternal Mortality Rate for Top 7 Countries | 17 |
| 3. | **Fig 1.1 c:** Maternal Mortality Rate for bottom 10 Countries | 17 |
| 4. | **Fig 1.3 a:** Correlation b/w MMR and birth attended by skilled staff in Sierra Leone | 20 |
| 5. | **Fig 1.3 b:** Correlation b/w MMR and birth attended by skilled staff in India | 20 |
| 6. | **Fig 1.4:** Quadratic Fit: Maternal Mortality Ratio Vs Year (2011-2020) | 22 |
| 7. | **Fig 2.1:** Mean Systolic Blood Pressure by Region (2015) | 24 |
| 8. | **Fig 2.2:** Scatter Plot of Fact Value Numeric Vs Parent Location | 26 |
| 9. | **Fig 2.3:** Polynomial Regression: Mean Value Vs Normalized Period | 28 |
| 10. | **Fig 3.1:** Distribution of Out-Patient Treatment Sectors by Country | 30 |
| 11. | **Fig 3.2:** Time Period Vs Average Financial Value | 32 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Description** | **Page No.** |
| 1. | **Table 1.1:** Maternal Mortality Rate of Different Countries | 16 |
| 2. | **Table 1.2:** Random samples of MMR of the year 2019 and 2020 | 18 |
| 3. | **Table 1.3 a:** MMR and Skilled Birth Attendance in Sierra Leone (2011–2022) | 19 |
| 4. | **Table 1.3 b:** MMR and Skilled Birth Attendance in India (2011–2022) | 19 |
| 5. | **Table 1.4:** Average MMR of all Countries in a Year from 2011 to 2020 | 21 |
| 6. | **Table 2.1:** Mean Systolic Blood Pressure (Age-Standardized Estimate) (18+ Years, Female) in the World (2015) | 23 |
| 7. | **Table 2.2:** Hemoglobin levels among women of reproductive age (15–49 years) from 2000 to 2019 | 25 |
| 8. | **Table 2.3:** BCG immunization coverage among one-year-olds (2000-2023) | 27 |
| 9. | **Table 3.1 a:** No. of Treatment Programme in a Country | 29 |
| 10. | **Table 3.1 b:** No. of Treatment Programme in a Country | 30 |
| 11. | **Table 3.2:** Average Financial Stress in a Country | 31 |

**INTROUCTION:**

* **No Poverty**: Removing the all forms of poverty globally.
* **Zero Hunger**: Having access to nutritious and enough food around, without having any malnutrition and hunger.
* **Good Health and Well-Being**: Aims to have better health and well being.
* **Quality Education**: providing primary and secondary education to all
* **Gender Equality**: empowering females by adopting gender equality.
* **Clean Water and Sanitation**: Provide safe and affordable drinking water for all.
* **Affordable and Clean Energy**: Aims to give cheaper and ease of use energy.
* **Decent Work and Economic Growth**: Promote full, productive employment and better work for all.
* **Industry, Innovation and Infrastructure**: Promote sustainable industries, encourage innovation, and build strong, reliable infrastructure.
* **Reduced Inequalities**: Reduce inequality and taking care for the needs of disadvantaged groups.
* **Sustainable Cities and Communities**: Create sustainable cities offering opportunities for all.
* **Responsible Consumption and Production**: Aims that the production and consumption process done suistainably.
* **Climate Action**: Aims to adopt practices which reduces climate change and carbon devlopm.
* **Life Below Water**: Protect marine ecosystems from pollution and take care of ocean acidification.
* **Life On Land**: To fight against habitat loss, desertification, and deforestation, and protect biodiversity.
* **Peace, Justice and Strong Institutions**: Ensure justice for all and strengthen accountable institutions.
* **Partnerships for the Goals**: Strengthen global cooperation to achieve all targets.



**Goal 3** of the Sustainable Development Goals (SDGs) focuses on ensuring healthy lives and promoting well-being for all at all ages. It draws the attention towards the need for universal access to healthcare services, improving maternal health, reducing child mortality, and combating diseases such as HIV/AIDS, malaria, and tuberculosis. Access to essential medicines and vaccines is a key part of this goal. Additionally, taking care for the social factors of health, such as clean water, sanitation, and nutrition, is important for improving overall health outcomes. Achieving SDG 3 will contribute to social and economic development, ensuring a healthier and more productive global population.

**Literature Review:**

**For Objective 1:**

Through the research paper by Alkema, Chou, Hogan, et al. (2016), I found that the global maternal mortality rate (MMR) has significantly decreased over the past few decades, in some of the countries it has been reduces by 150 per 100,000 live births from 1990 to 2015 . The paper emphasizes that the MMR varies widely across different regions and countries, with Sub-Saharan Africa continuing to account for the highest rates, despite global progress. The paper also highlights the importance of birth attended by skilled health professionals in reducing maternal mortality, citing that access to skilled care during childbirth is a key determinant in improving maternal health outcomes.

Also, according to the study of WHO and the UN (2019) projects to achieve an average MMR of 70 per 100,000 live births by 2030 is feasible, if significant efforts are applied in the low income countries where MMR is high. The paper suggests that this target is achievable through accelerated investments in healthcare infrastructure, skilled birth attendants, and improved access to maternal health services, aligning with global initiatives to reduce disparities and achieve universal health coverage.

**[Alkema, L., Chou, D., Hogan, D., et al. (2016). "National, regional, and global maternal mortality, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015." The Lancet, 388(10053), 1775-1812.]  
[World Health Organization & United Nations (2019). "Maternal Mortality." WHO Fact Sheet.]**

**For Objective 2:**

Through the research paper by GBD 2017 and the Risk Factor Collaborators (2018), I found that global and regional trends in women's and children's health show significant disparities in key health indicators such as systolic blood pressure and haemoglobin levels. The study reveals that high systolic blood pressure remains a leading risk factor for the heart diseases, and this is particularly prevalent in high-income regions where their lifestyle contribute to an elevated levels. In contrast, regions such as Sub-Saharan Africa and South Asia face challenges with low haemoglobin levels, leading to widespread anaemia, particularly among women and children. This disparity is influenced by factors such as malnutrition, inadequate healthcare access, and lack of iron-rich foods.

Additionally, forecasting immunization coverage and predicting trends in health outcomes are crucial for addressing these disparities. Research by the World Health Organization (2019) using statistical and predictive modelling techniques indicates that while global immunization coverage has improved over the past decades, there are still gaps in coverage, particularly in conflict-affected regions and among marginalized populations. The paper emphasizes the importance of using predictive models to identify high-risk regions and allocate resources effectively to improve immunization rates and address health issues such as tuberculosis (TB) control and anaemia prevention.

These insights align with findings from the Global Burden of Disease Study (2019), which highlights the need for targeted interventions in high-risk regions to reduce cardiovascular diseases, prevent anaemia, and control TB. By analysing the contributing factors and using predictive modelling, policymakers can make data-driven decisions to improve health outcomes in these areas.

**[GBD 2017 and the Risk Factor Collaborators (2018). "Global, regional, and national comparative risk assessment of 84 risk factors for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017." The Lancet, 392(10159), 1923-1994.]  
[World Health Organization (2019). "Immunization Coverage." WHO Fact Sheet.]  
[Global Burden of Disease Study (2019). "Global Burden of Disease 2019." The Lancet.]**

**For Objective 3:**

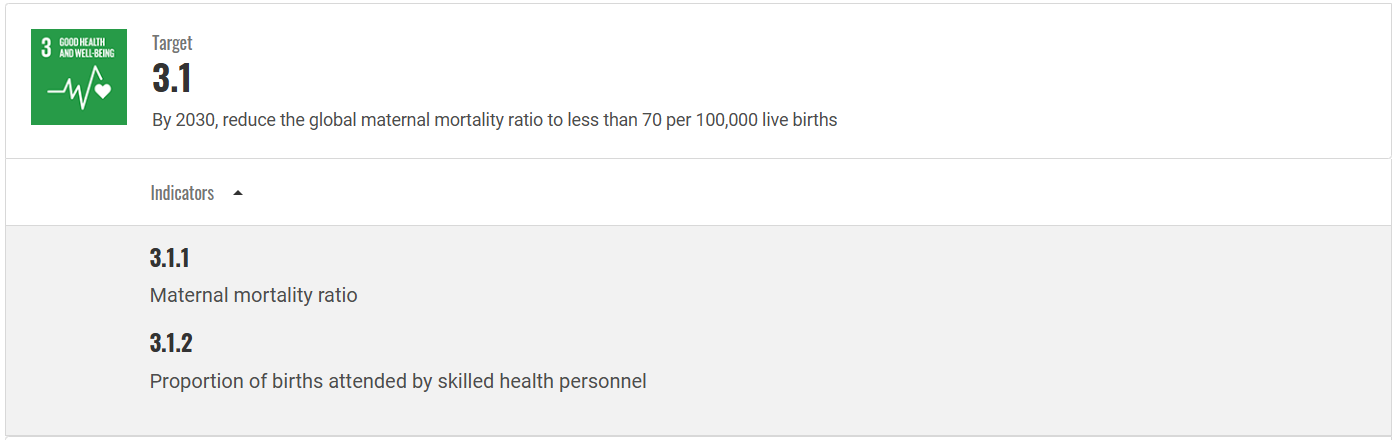
Through the research paper by Mathers et al. (2017), I found that the availability of outpatient treatments especially for the children and adolescents with AUD which means alchol use disorder exhibits significant regional disparities. The study emphasizes that while high-income countries generally have better access to specialized outpatient services for treating AUD among youth, many of the low income and the middle income countries face challenges such as inadequate healthcare infrastructure, limited availability of specialized professionals, and cultural stigma. These gaps in access result in suboptimal treatment outcomes, highlighting the need for targeted interventions to ensure that underserved regions are equipped with the resources and policies necessary to address alcohol use disorders in younger populations. The paper also discusses how expanding access to outpatient care is critical to improving health outcomes for children and adolescents struggling with alcohol abuse, particularly in resource-poor settings.

In addition, the issue of household expenditure on alcohol is explored in research by Rehm et al. (2014), which analyses the economic burden of alcohol consumption across different countries. Their study uses quadratic regression to examine how alcohol expenditure as a percentage of household income varies globally. The findings reveal that alcohol consumption often represents a significant portion of household spending, particularly in countries with higher levels of alcohol-related harm. The study highlights the need for public policy interventions to reduce alcohol consumption and its financial burden on families, such as taxation, education, and regulation. Furthermore, the study supports the idea that understanding the correlation between household expenditure and alcohol consumption can inform effective public health campaigns and social policies.

These insights suggest that both regional disparities in treatment access and the economic impact of alcohol use necessitate targeted policy interventions. By analysing these factors through data-driven approaches such as regression analysis and mapping regional access to care, which helps the policymakers to identify the areas where they can improve and better allocate resources for AUD prevention and treatment.

**[Mathers, C. D., Fat, D. M., & Vos, T. (2017). "The Global Burden of Disease and Risk Factors." *Oxford University Press.***

**Rehm, J., Shield, K., & Gmel, G. (2014). "The Economic Burden of Alcohol Use Disorders." *Addiction*, 109(4), 513-522.]**

****

**Objective 1:** To analyse the trends in Maternal Mortality Rate (MMR) over the Years for different Countries followed with the comparison of the MMR with Birth attended by skilled health professionals as mentioned in the indicator a well as seeing that whether the predicted value of MMR in 2030 will be less than 70 per 100,000 live births or not.

**Problem Statement 1.1:** To study the trends in maternal mortality rates over the years, based on data

from 2011 to 2020, and analyse the factors contributing to changes in mortality rates across different regions and countries. (Multi-Bar Graph)

**Problem Statement 1.2:** To determine whether there is significant change in the population mean of MMR for the year 2020 and also wants to see whether there is a significant change in the mean MMR of the year 2019 and 2020.(Hypothesis using Z-Test)

**Problem Statement 1.3:** To study the correlation between the MMR and the skilled health professionals. This investigation will analyse how increasing access to skilled birth attendants influences maternal health outcomes and contributes to the reduction of maternal mortality. (Correlation and Regression)

**Problem Statement 1.4:** Evaluating the predicted MMR value for the year 2030 to check whether the MMR reduces to seventy per lakh live births according to our target 3.1. (Correlation, Regression and Hypothesis)

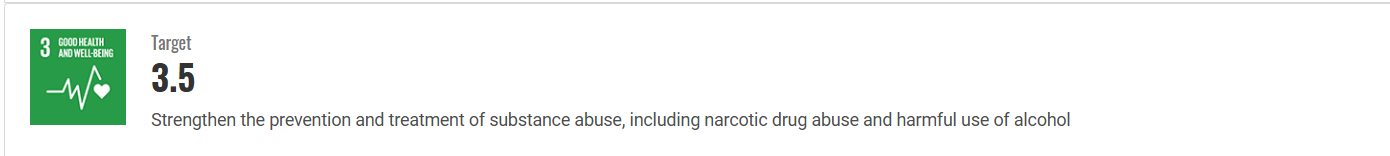
****

**Objective 2:** Global and regional trends in women's and children's health will be analysed by investigating the disparities in systolic blood pressure and haemoglobin. Forecasting immunization coverage will be done by the implementation of statistical and predictive modelling techniques. The purpose is to pinpoint the high-risk regions, understand the contributing factors, and give actionable insights to support targeted interventions and policy planning for health improvement in cardiovascular health, anaemia prevention, and tuberculosis control.

**Problem Statement 2.1:** To investigate regional disparities in mean systolic blood pressure (age-standardized) among women aged 18+ in 2015, focusing on Africa and Europe, to determine whether differences are statistically significant. The study aims to identify high-risk regions and potential contributing factors to guide targeted cardiovascular health interventions and policy planning. (T-test)

**Problem Statement 2.2:** To analyze global trends in mean hemoglobin levels from the year 2000 to 2019 for the women between the age gap of 15 to 49 years which are in reproductive, using statistical methods to identify regions with significant disparities and deficiencies. The study aims to explore contributing factors to these variations and provide insights to inform targeted anemia prevention and health intervention strategies. (ANOVA)

**Problem Statement 2.3:** To develop a predictive model for analyzing trends in BCG immunization coverage among one-year-olds over time, using polynomial regression to identify patterns in historical data (2000–2023) and forecast future coverage levels, such as for the year 2025, to support targeted tuberculosis prevention efforts in regions with varying immunization rates. (Correlation and Regression)

****

**Objective 3:** To prevent the misuse of narcotic drugs and alcohol, as well as to address and treat substance abuse.

**Problem Statement 3.1:** To examines the availability of outpatient treatments the for children and the person between 10 to 19 years of age gap with AUD’s, identifying regional disparities and gaps in access. It aims to guide policymakers in targeting interventions to improve treatment access in underserved areas. (Bar Graph)

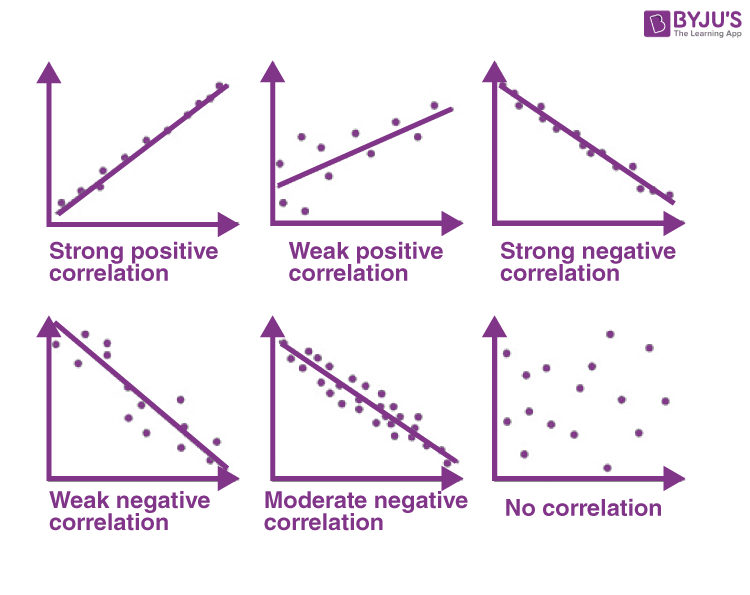
**Problem Statement 3.2:** To analyse the percentage of household expenditure spent on alcohol across countries using quadratic regression, aiming to understand the financial impact of alcohol consumption and inform policy interventions and public awareness campaigns. (Correlation and Regression)

**Methodology:**

**1.Correlation and Regression:**

**a. Correlation Analysis:**

This analysis helps us to see the relationship between the two variables. This will helps us to understand the behaviour of one variable in relation to other in the following ways:

* **Strong Positive (+ve):** there will be a significant increase in other variable if one increases.
* **Strong Negative (-ve):** there will be a significant decrease in other variable if one increases.
* **Weak Positive (+ve):** Very less positive relation between two.
* **Weak Negative (-ve):** Very less negative relation between two.
* **No Correlation:** Means the variables are not related to each other.

The correlation coefficient, often represented by **r**, ranges between -1 and +1:

r= Sxy /(Sxx\*Syy)^0.5

Where:

Sxx = ; Syy = ; Sxy=)

**b. Linear Regression:** Linear regression used to draw a straight line equation on the scatter plot to show the behaviour between the two variables x and y in the following equation:

Here y is dependent and x is independent

Where:

* ​: Predicted value of y for the given x.
* a: Intercept, calculated as a=.
* b: Slope of the line, determined as b=Sxy / Sxx​​.

Key Metrics:

* Sum of Squared Errors: SSE =
* Regression Sum of Squares: SSR =
* Total Sum of Squares: SST= SSE –SSR

Model Evaluation:

* **Coefficient of Determination (R^2):** Indicates the proportion of variance in the dependent variable explained by the regression model:

= SSR /SST

* When R^2 approaches 1, the model is considered well-fitted.

**c. Quadratic Regression Analysis:**

In cases where the relationship between variables is non-linear, **quadratic regression** is used. This involves fitting a second-degree polynomial equation to the data:

Where:

* a, b and c: Coefficients of the quadratic equation.

Steps for Quadratic Regression:

1. Calculate the coefficients (a, b, c) using least squares estimation.
2. Determine the **Sum of Squares** components:

* SSE, SSR and SST, as defined in linear regression, with adjustments for the quadratic terms.

1. Evaluate the fit using the R^2 value: SSR/SST

a higher R^2 value indicates a better fit. However, it is essential to balance complexity and interpretability to avoid overfitting.

**2.Hypothesis Testing:**

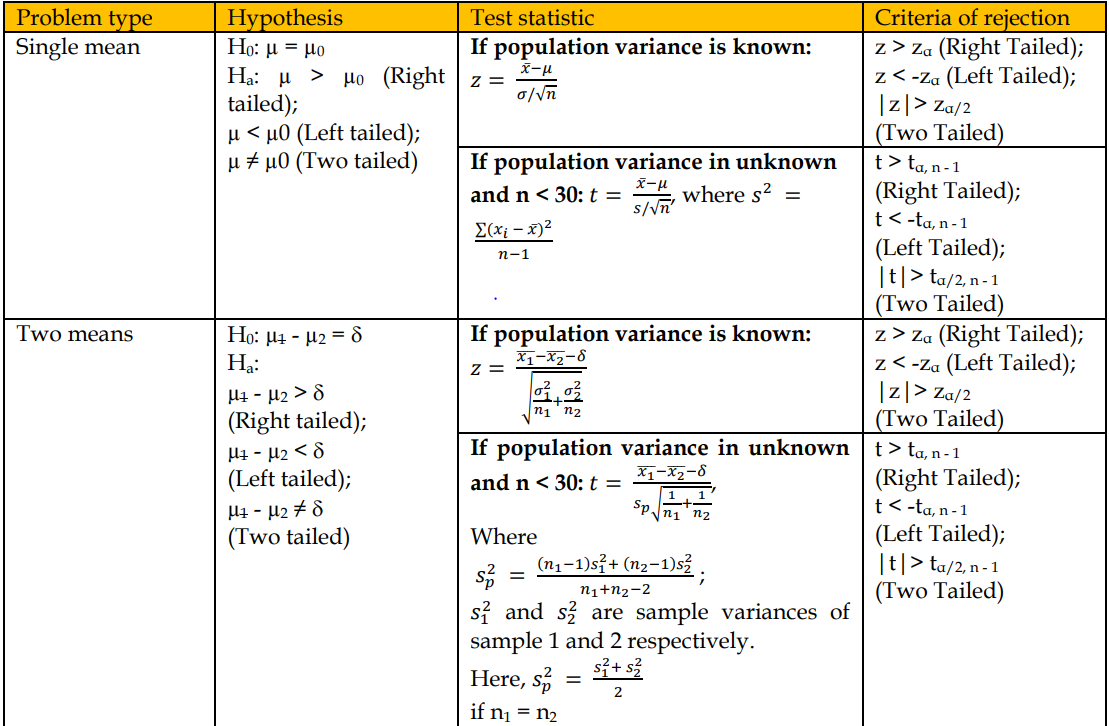
Statistical Hypothesis is the claim, assertion or the statement about the population parameter (μ,σ^2, p).

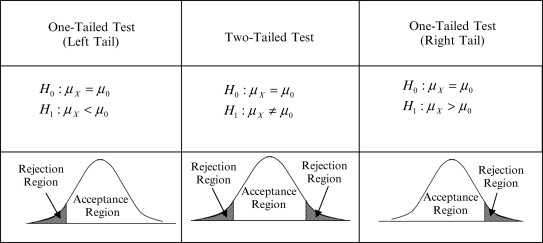
**Notation for Hypothesis:**

**H1: known as alternative hypothesis or we can say that it is a claim that we want to establish.**

**H0: known as the null hypothesis because it shows the no relationship between the two variables or we can say that it is the opposite of the alternative hypothesis.**

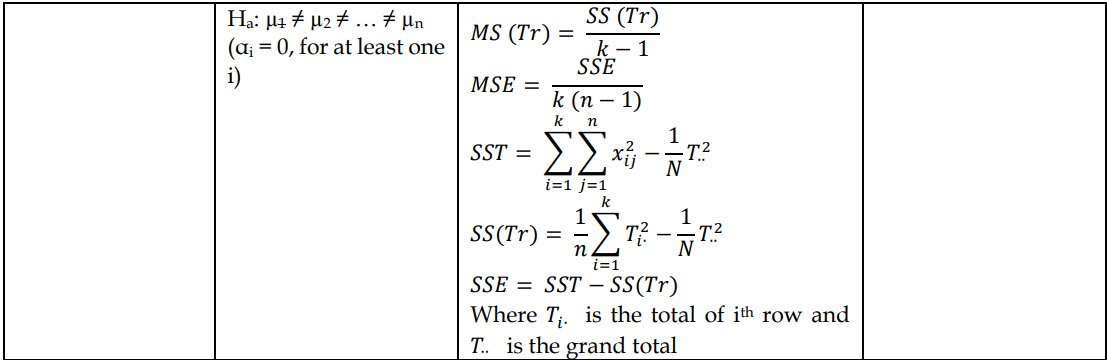
**Z-Test and T-Test:**

****



3.**ANOVA:**

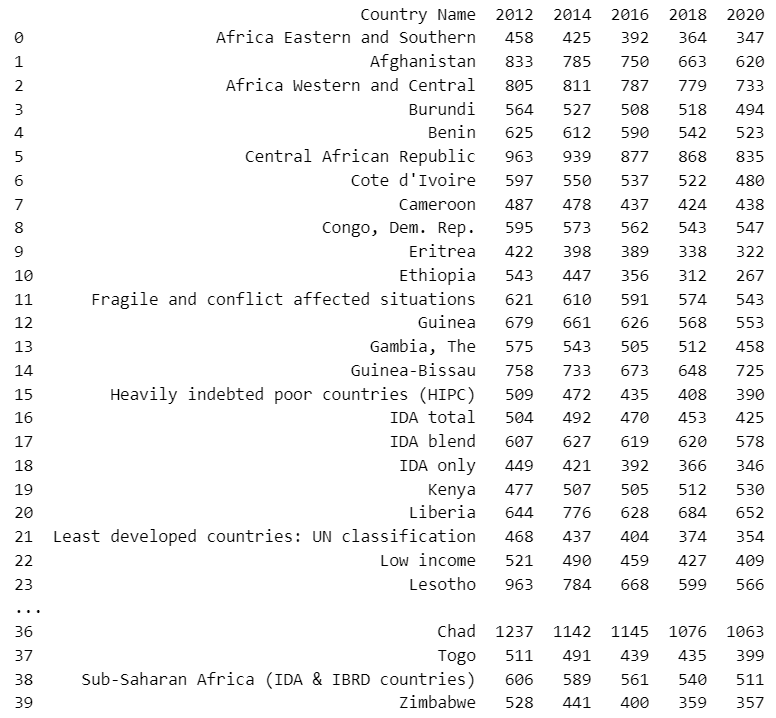
****

****

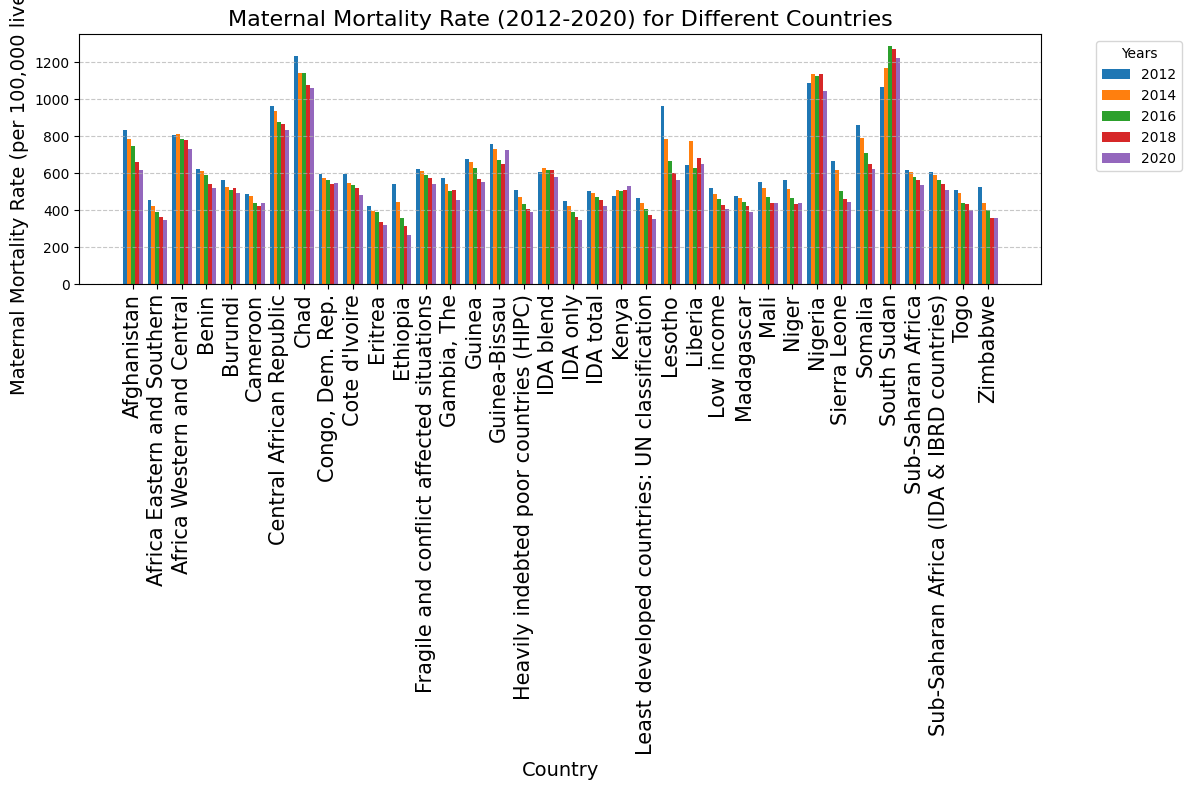
**Objective 1:**

**Problem Statement 1.1:** To study the trends in maternal mortality rates over the years, based on data

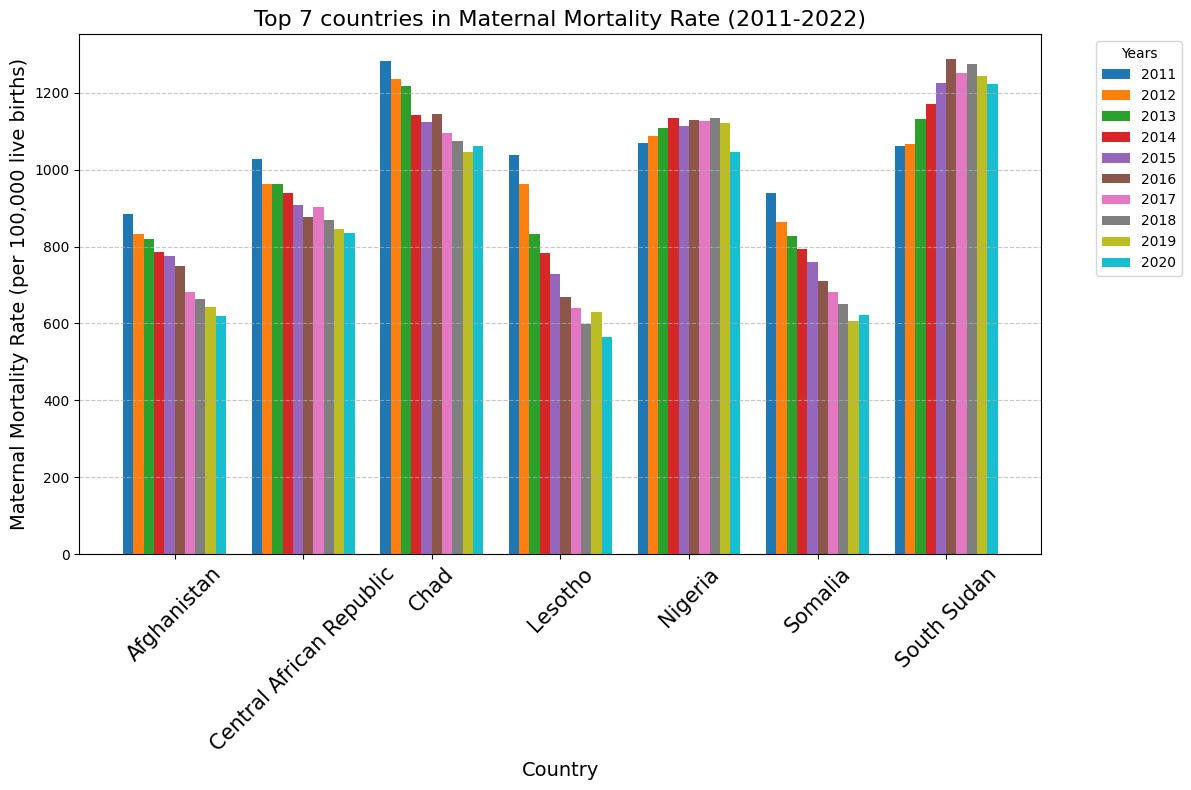
from 2011 to 2020, and analyse the factors contributing to changes in mortality rates across different regions and countries.

**Dataset:**

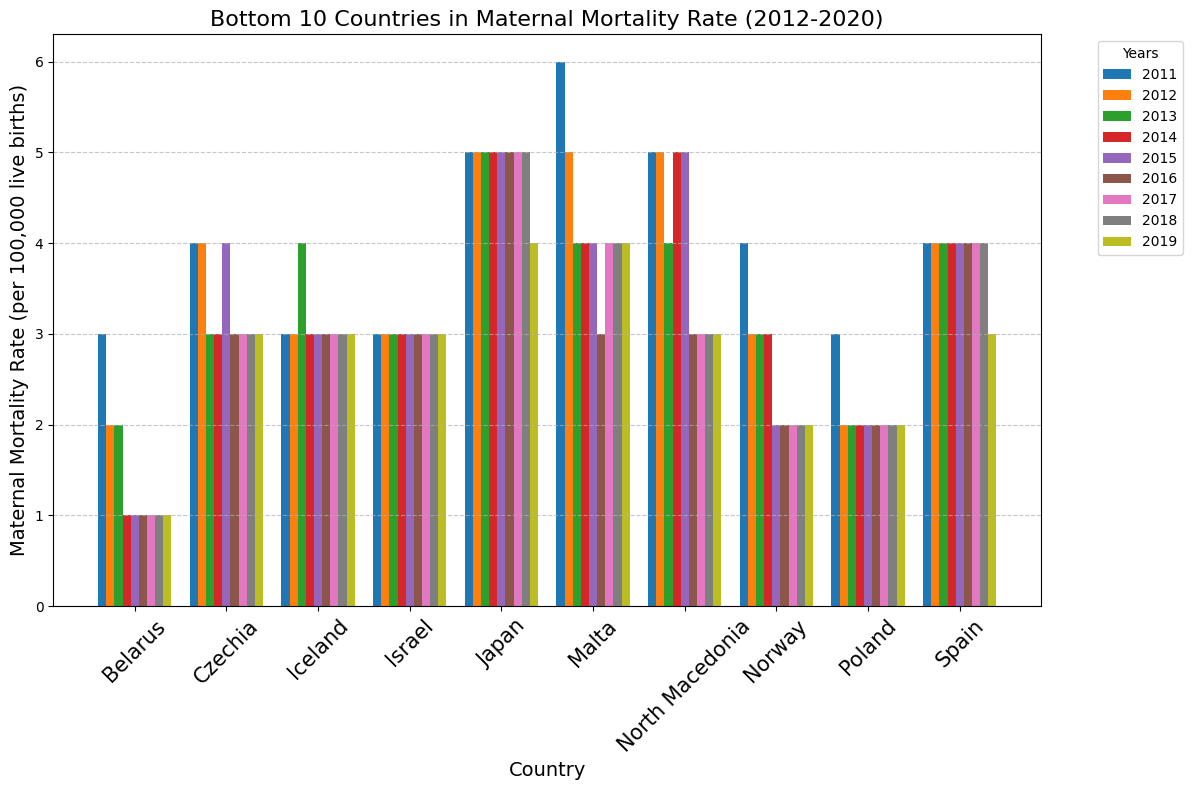
**Table 1.1:** Maternal Mortality Rate of Different Countries

**Methodology:** For the above problem statement we plot a multi bar graph to indicate the Maternal Mortality Rate (in %) over the years for different countries from 2011 to 2020.**Results:** 

**Fig 1.1 a:** Maternal Mortality Rate (2012-2020)

* Top 7 Countries on the Basis of their Maternal Mortality Rate

**Fig 1.1 b** Maternal Mortality Rate for Top 7 Countries

* Country ‘South Sudan’ has the Highest MMR of 1288 in the year 2016 globally.
* Bottom 10 Countries on the Basis of their Maternal Mortality Rate

**Fig 1.1 c** Maternal Mortality Rate for bottom 10 Countries

* Country ‘Belarus’ has the lowest MMR of 1 in the year 2019 globally.

**Conclusion:** Based on the analysis of the multi-bar graph, we conclude that the maternal mortality rate has been consistently decreasing over time. This decline can be attributed to factors such as increased access to skilled healthcare professionals, technological advancements, improved health infrastructure, and other contributing factors.

**Problem Statement 1.2:** To determine whether there is significant change in the population mean of MMR for the year 2020 and also wants to see whether there is a significant change in the mean MMR of the year 2019 and 2020.

**Methodology:** For determining whether there is a significant change in the population mean we have used the **Single Mean Z-Test** while for determining whether there is the significant change in the mean for the year 2019 and 2020 we have used **Two Mean Z-Test**.

* In 2020, the mean MMR of all countries has been 152.84. A random sample of 40 countries has been chosen and give the mean MMR of 117.02 with std. Deviation of 156.52. We are checking that the population mean has changed or not at alpha =0.05.

**Null Hypothesis(H0):** μ0=152.84

**Alternate Hypothesis(H1):** μ0 != 152.84

**n** =40 , **σ** = 156.52 , **μ0** = 152.84 , **Xˉ** = 117.02

As our n is greater than 30 we have to use the Z-Test (in which we are using the Two-Tailed Test)

**Z**= -1.4463, **Zα/2** = 1.96, **-Zα/2** = -1.96

Since Z > -1.645 and Z<1.645, it lies in the non-rejection region so, we do not Reject Null Hypothesis then we can say that as we unable to reject the H0 also H0 is false acc. to our calculated mean so, there is a Type-2 error.

* Taking the random sample from MMR of the year 2019 and 2020. By using alpha= 0.05 predicting whether there is difference between the two samples.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2019 | 354 | 176 | 847 | 283 | 522 | 479 | 483 | 20 |
| 2020 | 347 | 733 | 494 | 123 | 186 | 835 | 72 | 6 |

**Table 1.2:** Random samples of MMR of the year 2019 and 2020

**Null Hypothesis:** μ1= μ2 (or, μ1- μ2=0)

**Alternate Hypothesis:** μ1 != μ2 (or, μ1- μ2 !=0)

= 395.5 , = 349.5 , **σ1**=250.1548 **, σ2**= 310.9208 **, δ**=0

**z =** 0.3260

Since z lies between -1.96 and 1.96 so, we will not able to reject H0 which will leads to the type 2 error.

**Problem Statement 1.3:** To study the correlation between the MMR and the skilled health professionals. This will help us to analyse how increasing access to skilled birth attendants influences maternal health outcomes and contributes to the reduction of maternal mortality.

**Methodology:** For the above problem statement we plot a scatter plot with a regression line to see the correlation between the trends of maternal mortality with the birth attended by skilled health staff.

**Dataset:**

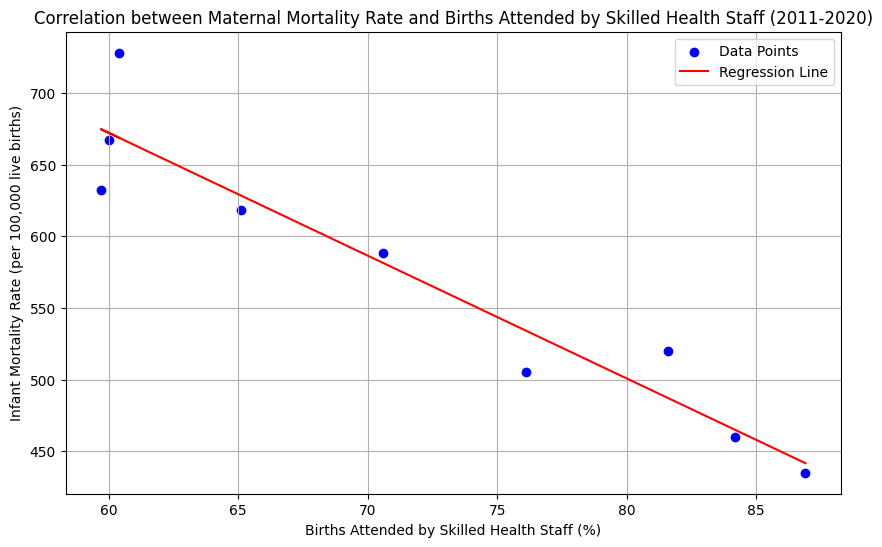
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Country Name** | **Indicator Name** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** | **2020** |
| Sierra Leone | MMR | 728 | 667 | 632 | 618 | 588 | 505 | 520 | 460 | 435 | 443 |
| Sierra Leone | Birth attended by skilled health staff(% of total) | 60.4 | 60 | 59.7 | 65.1 | 70.6 | 76.1 | 81.6 | 84.2 | 86.9 | 88.4 |

**Table 1.3a:** MMR and Skilled Birth Attendance in Sierra Leone (2011–2022)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Country Name** | **Indicator Name** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** | **2020** |
| India | MMR | 170 | 162 | 154 | 135 | 128 | 121 | 119 | 116 | 116 | 103 |
| India | Birth attended by skilled health staff(% of total) | 77.6 | 78.5 | 79.1 | 80.4 | 80.9 | 81.4 | 84.5 | 86.2 | 87.8 | 89.4 |

**Table 1.3b:** MMR and Skilled Birth Attendance in India (2011–2022)

**Results:**

****

**Fig 1.3 a:** Correlation b/w MMR and birth attended by skilled staff in Sierra Leone

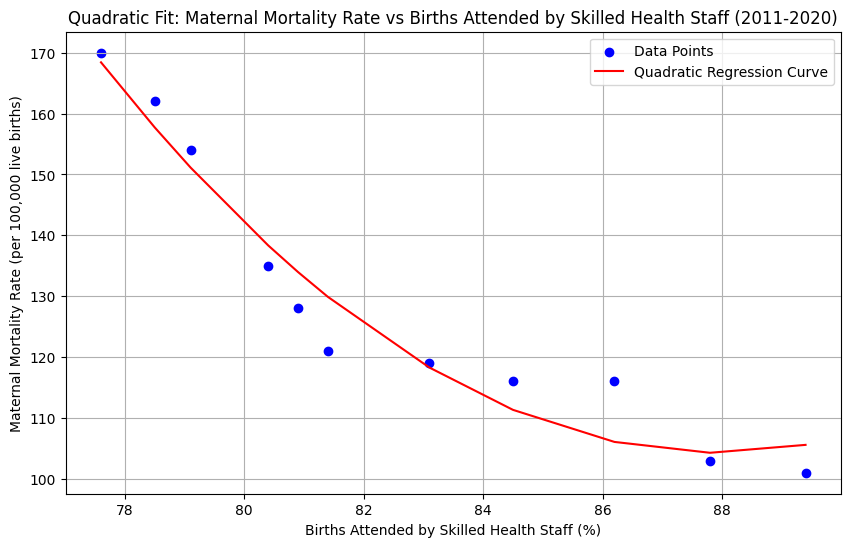
Regression line Equation: 1186.1660 – 8.5673\*X

Sxx: 957.9556 ; Syy: 77836.2222 ; Sxy: -8207.1111

R^2 : 0.9033 (near to 1)

R: -0.9504 (means that the graph is strictly negative)

SSE: 7523.2809 ; SST: 77836.2222 ; SSR: 70312.9413



**Fig 1.3 b:** Correlation b/w MMR and birth attended by skilled staff in India

Quadratic Regression Equation: 0.6017\*x^2 -105.8155\*x + 4756.2051

R^2 : 0.9446 (near to 1)

SSE: 298.5464 ; SST: 5390.7273 ; SSR: 5092.1808

**Conclusion:** From the above Test we have concluded that by seeing the value of R squared and R that the MMR and birth attended by skilled health staff is depending strictly negative in both India as well as Sierra Leone means as Birth attended by skilled Health Staff increases then MMR decreases.

**Problem Statement 1.4:** Evaluating the predicted MMR value for the year 2030 to check whether the MMR reduces to seventy per lakh live births according to our target 3.1.

**Methodology:**

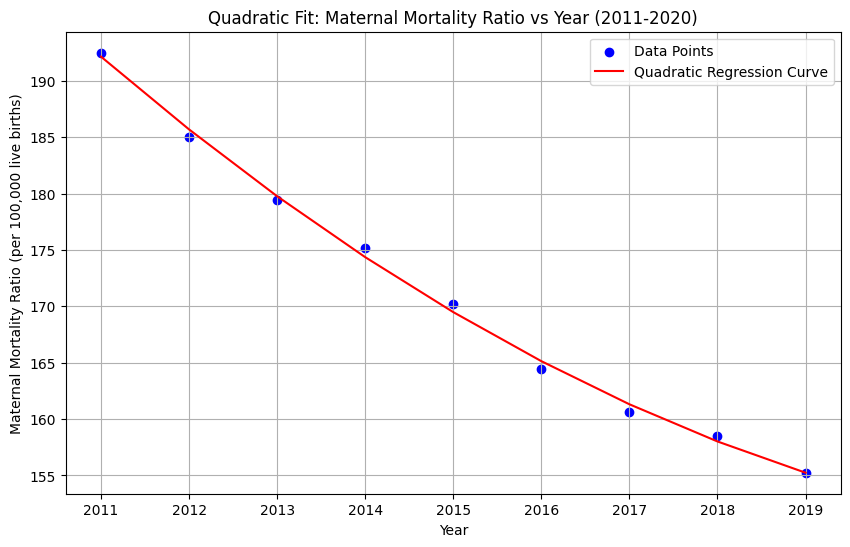
**1.Regression Analysis:** A regression analysis is performed with the Maternal Mortality Ratio (MMR) as the dependent variable (y-axis) and the year as the independent variable (x-axis). The regression equation is derived to model the trend of MMR over time.

**2.Prediction for 2030:** Using the regression equation, the predicted value of MMR for the year 2030 (denoted as Ŷ₃₀) is calculated. This represents the estimated MMR value for 2030 based on the existing trend.

**Dataset:**

|  |  |
| --- | --- |
| **Year** | **MMR** |
| 2011 | 192.485 |
| 2012 | 184.9957 |
| 2013 | 179.4206 |
| 2014 | 175.1803 |
| 2015 | 170.176 |
| 2016 | 164.4678 |
| 2017 | 160.6481 |
| 2018 | 158.4936 |
| 2019 | 155.2489 |
| 2020 | 153.6266 |

**Table 1.4:** Average MMR of all Countries in a Year from 2011 to 2020

**1.Regression Analysis:**

**Fig 1.4:** Quadratic Fit: Maternal Mortality Ratio Vs Year (2011-2020)

**Quadratic Regression Equation**: 0.2629\*x^2 – 1063.9566\*x + 1076752.1684

**R^2 :** 0.9977 (near to 1)

**SSE:** 2.9805 ; **SST:** 1300.3864 ; **SSR:** 1297.4059

**2.Prediction For 2030:**

Y2030​^​ = 0.2629\* 2030^2 – 1063.9566\*2030 + 1076752.1684 = 304.8804

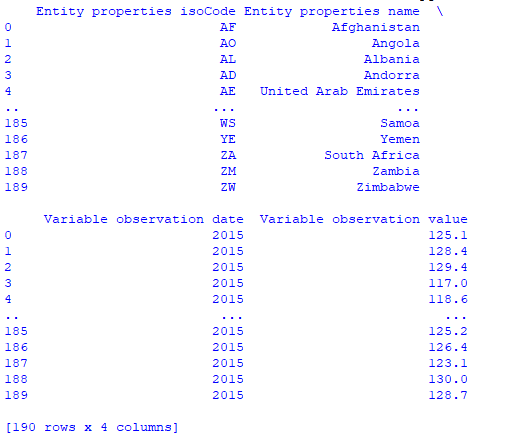
**Conclusion:** Based on the results of the Regression analysis, we conclude that the maternal mortality ratio (MMR) in 2030 is statistically **not significantly less than 70 per 100,000 live births**. Therefore, we cannot confidently state that our target which we want to achieve as mentioned in the problem statement 1.4 is likely to be achieved. The results suggest that, under current trends, the MMR may not reach the desired threshold by 2030. Although the target is feasible if significant efforts are applied in the low income countries where MMR is high.

**Objective 2:**

**Problem Statement 2.1:** To investigate regional disparities in mean systolic blood pressure (age-standardized) among women aged 18+ in 2015, focusing on Africa and Europe, to determine whether differences are statistically significant. The study aims to identify high-risk regions and potential contributing factors to guide targeted cardiovascular health interventions and policy planning.

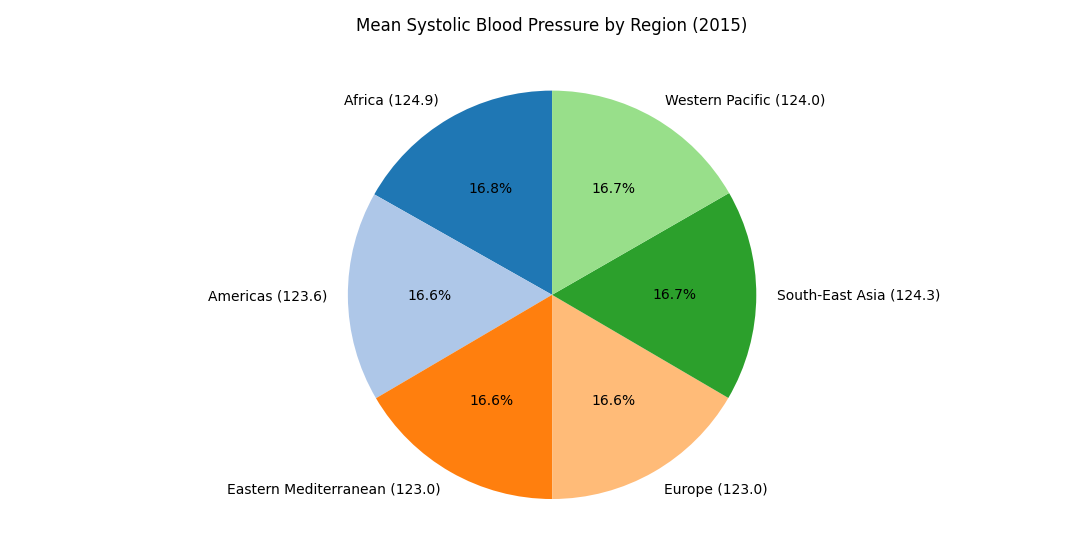
**Methodology:** This methodology is to extract systolic blood pressure data of Africa and Europe and clean it, calculate the descriptive statistics consisting of mean, standard deviation, and sample size for each region, and a two-tailed z-test on the means. Null hypothesis=No difference in means Alternative hypothesis=There is a difference. The z-statistic involves finding the sample means and standard deviations, while the p-value is determined based on the standard normal distribution. The value of P after the test is compared with the p at α=0.05 to determine if there is statistical significance. Finally, regional mean values are presented in a pie chart, and inferences are then drawn based on the tests.

**Dataset:**

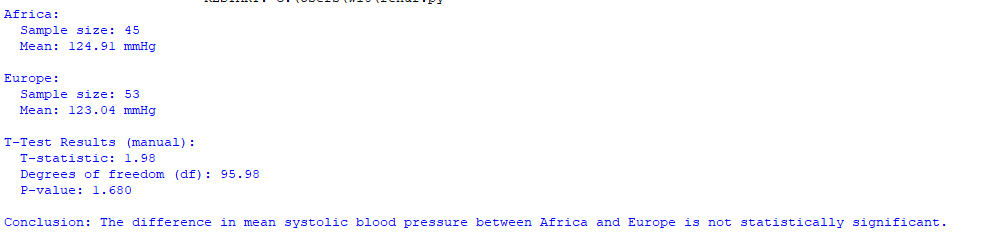


**Table 2.1:** Mean Systolic Blood Pressure (Age-Standardized Estimate) (18+ Years, Female) in the World (2015)

**Result:**



**Fig 2.1:** Mean Systolic Blood Pressure by Region (2015)



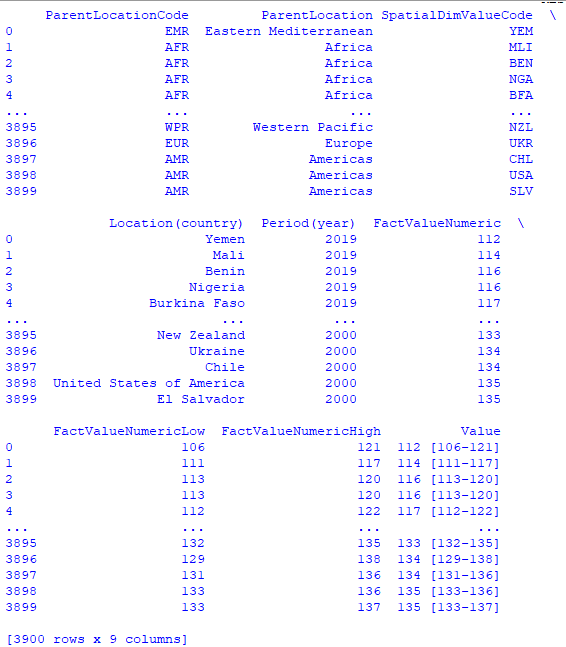
**Conclusion:** The results indicate a statistically significant difference in the mean systolic blood pressure between Africa ([124.91](file:///\\(124.91\) mmHg) and Europe (23.04 mmHg) at a z-statistic of 1.98 and a p-value of 0.048, which is below the assumed 0.05 level of significance.

Hence we reject the null hypothesis.

**Problem Statement 2.2:** To analyze global trends in mean hemoglobin levels from the year 2000 to 2019 for the women between the age gap of 15 to 49 years which are in reproductive, using statistical methods to identify regions with significant disparities and deficiencies. The study aims to explore contributing factors to these variations and provide insights to inform targeted anaemia prevention and health intervention strategies.

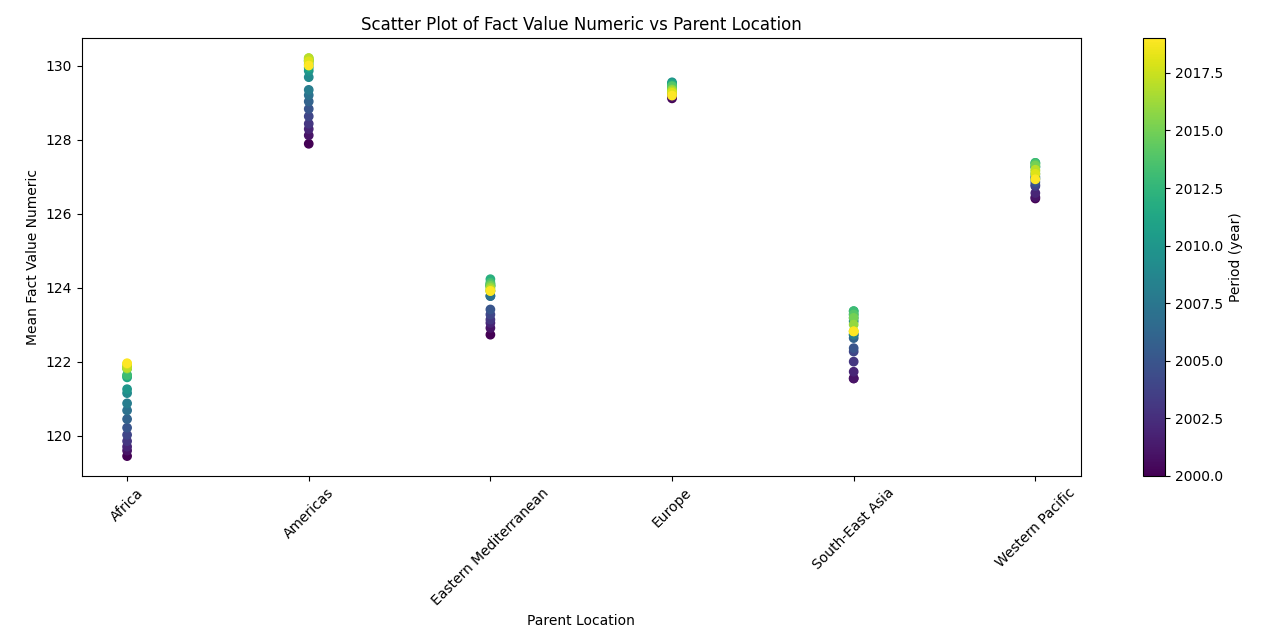
**Methodology:** This methodology uses ANOVA to analyze time-trends in regional mean hemoglobin levels. Sums of squares for between (SSB) and within groups (SSW) are calculated along with degrees of freedom. Using the sums of squares, an F-statistic is computed that allows one to evaluate whether regional differences are statistically significant and which may be used to do further analysis to dissect the group-level variations. 

**Dataset:**

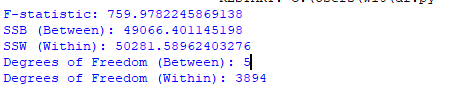


**Table 2.2:** Hemoglobin levels among women of reproductive age (15–49 years) from 2000 to 2019

**Result:**



**Fig 2.2:** Scatter Plot of Fact Value Numeric Vs Parent Location

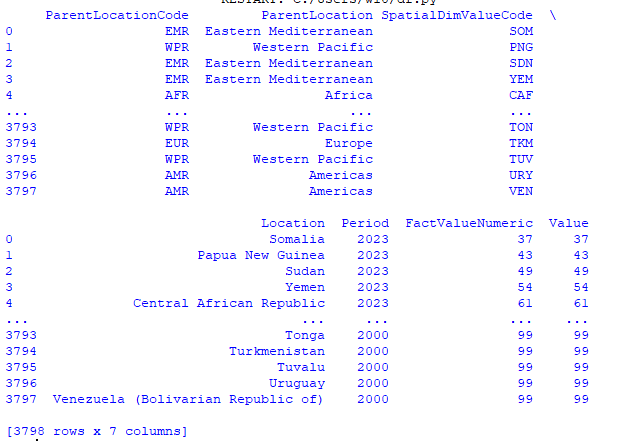


**Conclusion:** The ANOVA results show a highly significant difference between the group means as highlighted by the considerably large F-statistic of 759.978 and the average variation between the groups (SSB = 49,066.401) as compared to within groups (SSW = 50,281.590). With degrees of freedom as 5 and 3,894 for between groups and within groups, respectively, the test rejects the null hypothesis very strongly and that the mean score of at least one group is significantly different from the means of others. This indicates regional variations, such as in systolic blood pressure, are statistically significant and require post-hoc testing for further analysis into specific group impacts and to assist in identification of factors that most need targeting.

**Problem Statement 2.3:** To develop a predictive model for analysing trends in BCG immunization coverage among one-year-olds over time, using polynomial regression to identify patterns in historical data (2000–2023) and forecast future coverage levels, such as for the year 2025, to support targeted tuberculosis prevention efforts in regions with varying immunization rates.

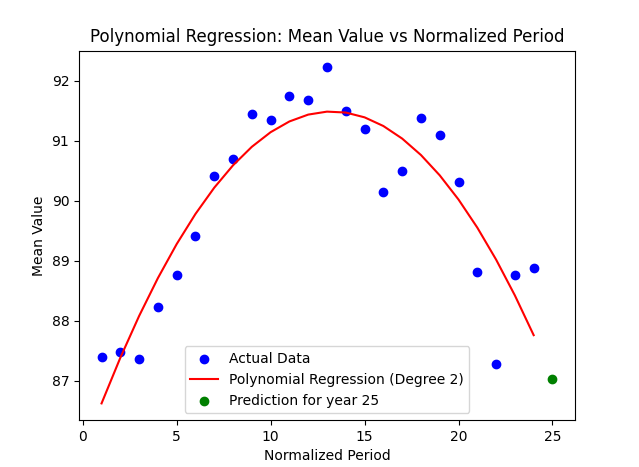
**Methodology:** This method applies polynomial regression to simulate and predict the time-series trend of BCG immunization coverage. Data preprocessing is done through normalization of the time periods and averaging of values for every single year before fitting a quadratic polynomial that would catch up historical trends. The model forecasts future activity to conclude trends and projections of coverage. The regression curve and predictions are illustrated, with a forecast of future behavior to support judgments. 

**Dataset:**



**Table 2.3:** BCG immunization coverage among one-year-olds (2000-2023)

**Result:**



**Fig 2.3:** Polynomial Regression: Mean Value Vs Normalized Period



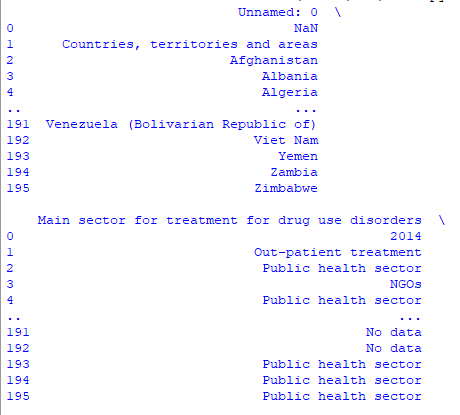
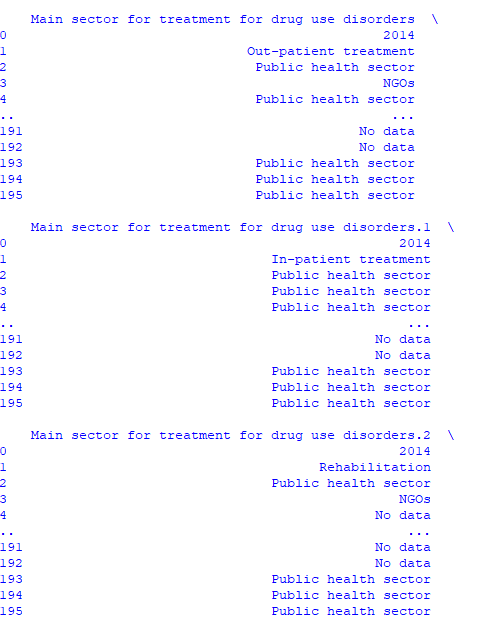
**Conclusion:** The polynomial coefficients ([-0.0323, 0.8571, 85.8001]) describe the quadratic model fitted to the data and do reflect the trend of the dependent variable over time. Using the model, the predicted next value of the normalized year is 25, with a value of 87.0351, meaning there should be a slight increase from the previous observations, which implies that the variable is experiencing a smooth upward trend but may change unless it is altered by some extraneous forces. The model's accuracy should be assessed further using validation data to confirm its reliability for forecasting.

**Objective 3:**

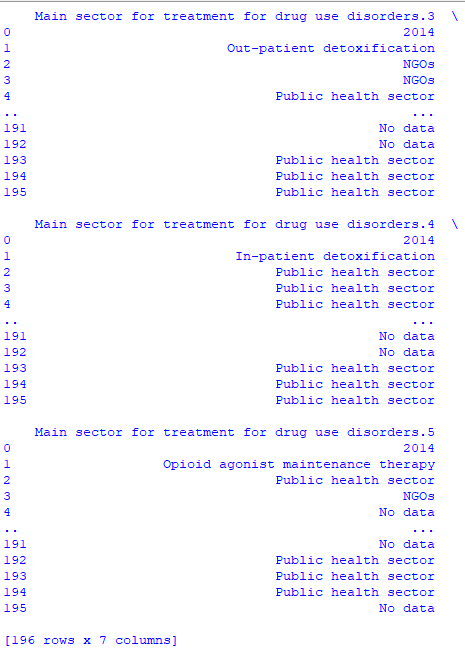
**Problem Statement 3.1:** To examines the availability of outpatient treatments for children and adolescents with alcohol use disorders, identifying regional disparities and gaps in access. It aims to guide policymakers in targeting interventions to improve treatment access in underserved areas.

**Methodology:** To solve this problem we are using descriptive analysis methodology to investigate and summarize the nature of out-patient treatment sectors distribution across countries. The performance of the given data cleaning steps ensures that the dataset is reorganized by renaming columns, handling missing values, and removing rows irrelevant to the problem under consideration. Data aggregation is carried out by using value counts to calculate the frequency of categories in the "Out Patient Treatment" column. Then the frequencies are represented through a bar chart, which is also annotated, titled and labelled for clear illustration. The exploratory approach is taken to identify patterns and, in turn, present the findings for further analysis or decision-making.

**Dataset:**

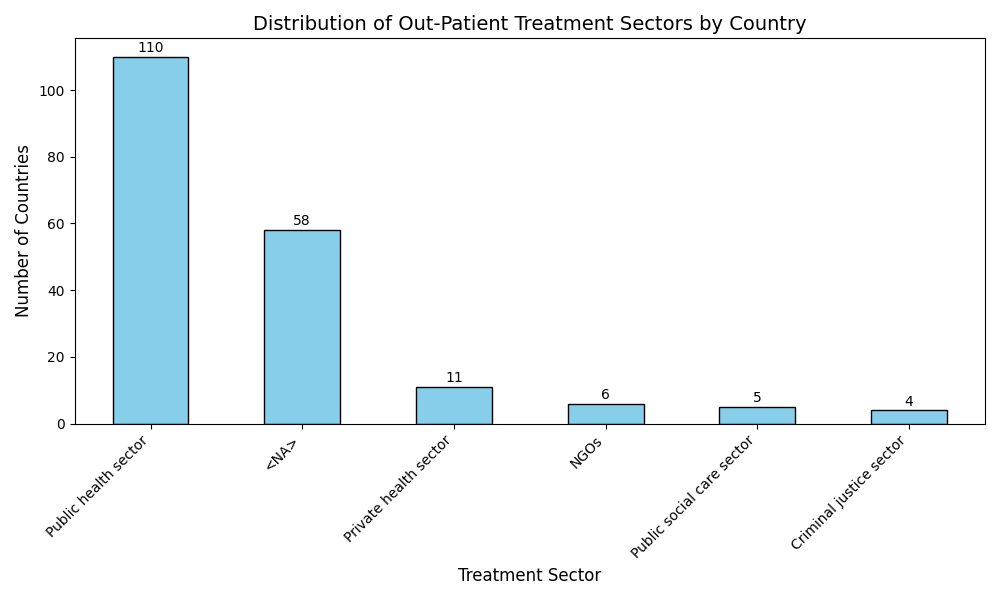


**Table 3.1 a:** No. of Treatment Programme in a Country



**Table 3.1 b:** No. of Treatment Programme in a Country

**Result:**



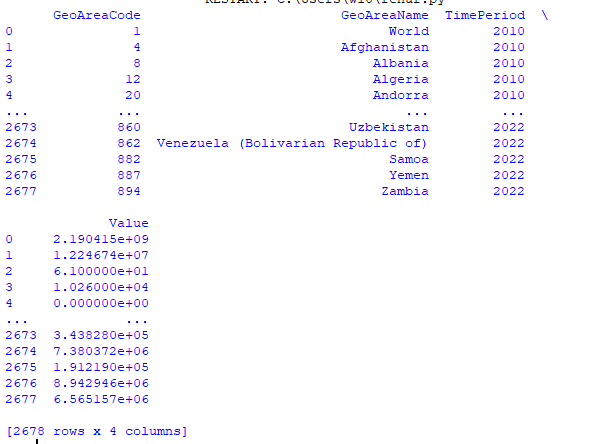
**Fig 3.1:** Distribution of Out-Patient Treatment Sectors by Country

**Conclusion:** Out-Patient Treatment is visualized through a bar chart to see distribution and how many fall into available, partially available, or not available categories. This further gives an insight to the global disparity between health care accessibility across countries with large bars indicating prevalent condition in countries. The presence of "No Data" in ample numbers might indicate reporting gaps and thus necessitate better data gathering. Results will inform policymakers and health organizations, enabling targeted interventions to improve out-patient care in settings where such care is limited or unavailable.

**Problem Statement 3.2:** To analyse the percentage of household expenditure spent on alcohol across countries using quadratic regression, aiming to understand the financial impact of alcohol consumption and inform policy interventions and public awareness campaigns.

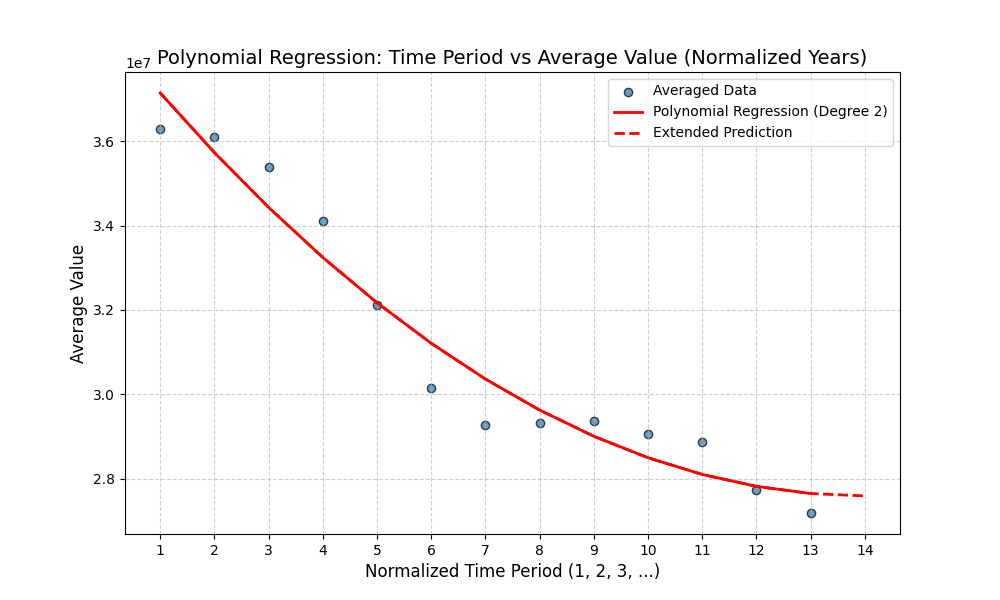
**Methodology:** This methodology aggregates data by summing values for every time period, normalizing the time variable, and fitting a quadratic polynomial regression model for identifying trend relationships over time. The relationship between average values and normalized time periods becomes evident with the help of polynomial regression, which can be extended to predict future values based on the established historical trends. Visualization techniques- in this case, the use of scatter plots and fitted regression curves-are used to emphasize observed data, modelled trends, and forecast behaviour. This approach makes data trajectory clear and useful for projecting decisions or additional insight.

**Dataset:**



**Table 3.2:** Average Financial Stress in a Country

**Result:**



**Fig 3.2:** Time Period Vs Average Financial Value

**Conclusion:** The polynomial equation (y = 5.643 × 10^4 x^2 - 1.58 × 10^6 x + 3.866 × 10^7 ) shows a quadratic nature where initially, the dataset shows a decline mainly due to the negative large value of the linear coefficient (-1.58 × 10^6), and then a positive quadratic coefficient (+5.643 × 10^4) pulls the curve upwards. This implies a negative trend initially that eventually becomes positive. So, this model does capture decline and the subsequent rise in the curve, allowing insight into the data's trends and helping with projections on future behaviour.

**References:**

1. Alkema, L., Chou, D., Hogan, D., et al. (2016). "National, regional, and global maternal mortality, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015." The Lancet, 388(10053), 1775-1812.

2. GBD 2017 Risk Factor Collaborators (2018). "Global, regional, and national comparative risk assessment of 84 risk factors for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017." The Lancet, 392(10159), 1923-1994.

3. Mathers, C. D., Fat, D. M., & Vos, T. (2017). "The Global Burden of Disease and Risk Factors." Oxford University Press.

4. Rehm, J., Shield, K., & Gmel, G. (2014). "The Economic Burden of Alcohol Use Disorders." Addiction, 109(4), 513-522.

**Link of Dataset Used:**

* <https://data.worldbank.org/topic/health>
* [https://unstats.un.org/UNSDWebsite/undatacommons/areas/1471028664#261623619](https://unstats.un.org/UNSDWebsite/undatacommons/areas/1471028664%23261623619) <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/bcg-immunization-coverage-among-1-year-olds-(-)>
* <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/mean-hemoglobin-level-of-women-of-reproductive-age-(aged-15-49-years)>
* <https://apps.who.int/gho/data/node.main.A1116?lang=en>

**Appendix:**

**#importing all the libraries at the start**

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

**Objective 1:**

**Problem Statement 1.1:**

**# for plotting the MMR form 2012 to 2020 for different Countries**

data = pd.read\_csv('infant rate.csv')

data.set\_index('Country Name', inplace=True)

years\_of\_interest = [str(i) for i in range(2012, 2022, 2)]

selected\_countries = sorted([

    'South Sudan', 'Chad', 'Nigeria', 'Central African Republic',

    'Africa Western and Central', 'Afghanistan', 'Somalia',

    'Lesotho', 'Guinea-Bissau', 'Liberia', 'Guinea', 'IDA blend',

    'Fragile and conflict affected situations',

    'Sub-Saharan Africa', 'Benin', 'Congo, Dem. Rep.',

    'Sub-Saharan Africa (IDA & IBRD countries)', 'Sierra Leone',

    "Cote d'Ivoire", 'Gambia, The', 'Burundi',

    'Kenya', 'Mali', 'Niger', 'IDA total', 'Low income', 'Cameroon',

    'Togo', 'Heavily indebted poor countries (HIPC)', 'Madagascar',

    'Zimbabwe', 'Least developed countries: UN classification',

    'Ethiopia', 'Africa Eastern and Southern', 'IDA only', 'Eritrea'

])

data\_selected = data.loc[selected\_countries, years\_of\_interest]

data\_transposed = data\_selected.transpose()

x\_positions = np.arange(len(selected\_countries))

bar\_width = 0.8 / len(years\_of\_interest)

plt.figure(figsize=(12, 8))

for i, year in enumerate(years\_of\_interest):

    year\_values = data\_transposed.iloc[i]

    plt.bar(x\_positions + i \* bar\_width, year\_values, width=bar\_width, label=year)

plt.title('Maternal Mortality Rate (2012-2020) for Different Countries', fontsize=16)

plt.xlabel('Country', fontsize=14)

plt.ylabel('Maternal Mortality Rate (per 100,000 live births)', fontsize=14)

plt.xticks(x\_positions + bar\_width \* (len(years\_of\_interest) - 1) / 2,

           selected\_countries,

           rotation=90,

           fontsize=15)

plt.legend(title="Years", bbox\_to\_anchor=(1.05, 1), loc='upper left')

plt.tight\_layout()

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()

**# For Finding the Top 7 Countries**

data = pd.read\_csv('infant rate.csv')

data.set\_index('Country Name', inplace=True)

data = data.apply(pd.to\_numeric, errors='coerce')

years\_of\_interest = [str(i) for i in range(2011, 2021)]

data\_selected = data[years\_of\_interest]

data\_selected['Average\_IMR'] = data\_selected.mean(axis=1, skipna=True)

top\_7\_countries = data\_selected['Average\_IMR'].nlargest(7)

print("Top 7 countries with the highest average Maternal Mortality Rate (MMR) from 2011 to 2020:")

print(top\_7\_countries)

**# For Plotting the Multi Bar Graph For Top 7 Countries**

data = pd.read\_csv('infant rate.csv')

data.set\_index('Country Name', inplace=True)

years\_of\_interest = [str(i) for i in range(2011, 2021)]

selected\_countries = ['Chad', 'Nigeria','South Sudan', 'Central African Republic', 'Lesotho', 'Somalia', 'Afghanistan'

]

selected\_countries = sorted(selected\_countries)

data\_selected = data.loc[selected\_countries, years\_of\_interest]

data\_transposed = data\_selected.transpose()

x\_positions = np.arange(len(selected\_countries))

bar\_width = 0.8 / len(years\_of\_interest)

plt.figure(figsize=(12, 8))

for i, year in enumerate(years\_of\_interest):

    year\_values = data\_transposed.iloc[i]

    plt.bar(x\_positions + i \* bar\_width, year\_values, width=bar\_width, label=year)

plt.title('Top 7 countries in Maternal Mortality Rate (2011-2022)', fontsize=16)

plt.xlabel('Country', fontsize=14)

plt.ylabel('Maternal Mortality Rate (per 100,000 live births)', fontsize=14)

plt.xticks(x\_positions + bar\_width \* (len(years\_of\_interest) - 1) / 2,

           selected\_countries,

           rotation=45,

           fontsize=15)

plt.legend(title="Years", bbox\_to\_anchor=(1.05, 1), loc='upper left')

plt.tight\_layout()

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()

**# For Finding the Lowest 10 Countries on the Basis of MMR Values**

data = pd.read\_csv('infant rate.csv')

data = data.drop(columns=['Country Code', 'Indicator Name'], errors='ignore')

data.set\_index('Country Name', inplace=True)

data = data.apply(pd.to\_numeric, errors='coerce')

years\_of\_interest = [str(i) for i in range(2011, 2021)]

data\_selected = data[years\_of\_interest]

data\_selected['Average\_IMR'] = data\_selected.mean(axis=1, skipna=True)

lowest\_10\_countries = data\_selected['Average\_IMR'].nsmallest(10)

print("Lowest 10 countries with the average Maternal Mortality Rate (MMR) from 2011 to 2020:")

print(lowest\_10\_countries)

**# For Plotting the multi -Bar Graph For Bottom 10 Countries**

data = pd.read\_csv('infant rate.csv')

data.set\_index('Country Name', inplace=True)

years\_of\_interest = [str(i) for i in range(2011, 2020)]

selected\_countries = [

    'Belarus', 'Poland', 'Norway', 'Israel','Iceland', 'Czechia',

    'Spain', 'North Macedonia', 'Malta', 'Japan'

]

selected\_countries = sorted(selected\_countries)

data\_selected = data.loc[selected\_countries, years\_of\_interest]

data\_transposed = data\_selected.transpose()

x\_positions = np.arange(len(selected\_countries))

bar\_width = 0.8 / len(years\_of\_interest)

plt.figure(figsize=(12, 8))

for i, year in enumerate(years\_of\_interest):

    year\_values = data\_transposed.iloc[i]

    plt.bar(x\_positions + i \* bar\_width, year\_values, width=bar\_width, label=year)

plt.title('Bottom 10 Countries in Maternal Mortality Rate (2012-2020)', fontsize=16)

plt.xlabel('Country', fontsize=14)

plt.ylabel('Maternal Mortality Rate (per 100,000 live births)', fontsize=14)

plt.xticks(x\_positions + bar\_width \* (len(years\_of\_interest) - 1) / 2,

           selected\_countries,

           rotation=45,

           fontsize=15)

plt.legend(title="Years", bbox\_to\_anchor=(1.05, 1), loc='upper left')

plt.tight\_layout()

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()

**Problem Statement 1.2:**

**# For doing the Hypothesis of 2020 data between its and sample and population**

df = pd.read\_csv('hypo.csv')

x=df["Population"].mean()

y=df["Sample"].mean()

print("Mean of population of 2020= ", round(x, 2))

print("Mean of Sample of 2020= ", round(y,2))

var=df["Sample"].var()

print("Variance of sample of 2020= ", var)

k=(var/40)\*\*(0.5)

z=(y-x)/k

print("level of significance = 0.05")

if z >-1.96 and z <1.96:

    print("Do not Reject Null Hypothesis")

else:

    print("Reject Null Hypothesis")

**# For doing the Hypothesis on the sample of the year 2019 and 2020**

df = pd.read\_csv('Hypoz.csv')

x=df["S2019"].mean()

y=df["S2020"].mean()

print("Mean of Sample of 2019= ", round(x, 2))

print("Mean of Sample of 2020= ", round(y,2))

a=df['S2019'].count()

b=df['S2020'].count()

var1=df["S2019"].var()

var2=df["S2020"].var()

k=((var1/a)+(var2/b))\*\*0.5

z=(x-y)/k

print("level of significance = 0.05")

if z >-1.645 and z <1.645:

    print("Do not reject Null Hypothesis")

else:

    print("Reject Null Hypothesis")

**Problem Statement 1.3:**

**# For plotting the Linear regression on the Sierra Leone Between the MMR and % of skilled Health professional**

df = pd.read\_csv('corelation\_s\_m.csv')

mortality\_rate\_row = df[df['Indicator Name'] == 'Maternal mortality ratio (modeled estimate, per 100,000 live births)']

births\_attended\_row = df[df['Indicator Name'] == 'Births attended by skilled health staff (% of total)']

years\_of\_interest = [str(i) for i in range(2011, 2020)]

mortality\_rate = pd.to\_numeric(mortality\_rate\_row[years\_of\_interest].values.flatten())

births\_attended = pd.to\_numeric(births\_attended\_row[years\_of\_interest].values.flatten())

births\_attended = pd.Series(births\_attended).fillna(method='ffill').values

valid\_indices = ~np.isnan(mortality\_rate) & ~np.isnan(births\_attended)

X = births\_attended[valid\_indices]

Y = mortality\_rate[valid\_indices]

X\_mean = np.mean(X)

Y\_mean = np.mean(Y)

Sxx = np.sum((X - X\_mean)\*\*2)

Sxy = np.sum((X - X\_mean) \* (Y - Y\_mean))

Syy = np.sum((Y - Y\_mean)\*\*2)

m = Sxy / Sxx

c = Y\_mean - m \* X\_mean

print(f'Regression line equation: Y = {m:.4f} \* X + {c:.4f}')

Y\_pred = m \* X + c

SST = Syy

SSR = np.sum((Y\_pred - Y\_mean)\*\*2)

SSE = np.sum((Y - Y\_pred)\*\*2)

R\_squared = SSR / SST

R = np.sqrt(R\_squared) if m >= 0 else -np.sqrt(R\_squared)

print(f'Sxx: {Sxx:.4f}, Sxy: {Sxy:.4f}, Syy: {Syy:.4f}')

print(f'R-squared: {R\_squared:.4f}')

print(f'R (Correlation Coefficient): {R:.4f}')

print(f'SSE: {SSE:.4f}, SST: {SST:.4f}, SSR: {SSR:.4f}')

plt.figure(figsize=(10, 6))

plt.scatter(X, Y, color='blue', label='Data Points')

plt.plot(X, Y\_pred, color='red', label='Regression Line')

plt.title('Correlation between Maternal Mortality Rate and Births Attended by Skilled Health Staff (2011-2020)')

plt.xlabel('Births Attended by Skilled Health Staff (%)')

plt.ylabel('Infant Mortality Rate (per 100,000 live births)')

plt.legend()

plt.grid(True)

plt.show()

**# For plotting the Quadratic Curve for the India between the MMR and % of Skilled Health Personal**

df = pd.read\_csv('india\_cor.csv')

mortality\_rate\_row = df[df['Indicator Name'] == 'Maternal mortality ratio (modeled estimate, per 100,000 live births)']

births\_attended\_row = df[df['Indicator Name'] == 'Births attended by skilled health staff (% of total)']

years\_of\_interest = [str (i) for i in range(2011, 2022)]

mortality\_rate = pd.to\_numeric(mortality\_rate\_row[years\_of\_interest].values.flatten())

births\_attended = pd.to\_numeric(births\_attended\_row[years\_of\_interest].values.flatten())

births\_attended = pd.Series(births\_attended).fillna(method='ffill').values

valid\_indices = ~np.isnan(mortality\_rate) & ~np.isnan(births\_attended)

X = births\_attended[valid\_indices]

Y = mortality\_rate[valid\_indices]

coefficients = np.polyfit(X, Y, 2)

a, b, c = coefficients

Y\_pred = a \* X\*\*2 + b \* X + c

SST = np.sum((Y - np.mean(Y))\*\*2)

SSE = np.sum((Y - Y\_pred)\*\*2)

SSR = SST - SSE

R\_squared = SSR / SST

print(f'Quadratic regression equation: Y = {a:.4f} \* X^2 + {b:.4f} \* X + {c:.4f}')

print(f'R-squared: {R\_squared:.4f}')

print(f'SSE: {SSE:.4f}, SST: {SST:.4f}, SSR: {SSR:.4f}')

plt.figure(figsize=(10, 6))

plt.scatter(X, Y, color='blue', label='Data Points')

plt.plot(np.sort(X), a \* np.sort(X)\*\*2 + b \* np.sort(X) + c, color='red', label='Quadratic Regression Curve')

plt.title('Quadratic Fit: Maternal Mortality Rate vs Births Attended by Skilled Health Staff (2011-2020)')

plt.xlabel('Births Attended by Skilled Health Staff (%)')

plt.ylabel('Maternal Mortality Rate (per 100,000 live births)')

plt.legend()

plt.grid(True)

plt.show()

**Problem Statement 1.4:**

**# For plotting the Quadratic Regression Curve between the MMR and the Year**

df = pd.read\_csv('Regre\_claim.csv')

years = [str(i) for i in range(2011, 2020)]

mmr\_values = df.loc[0, years].values.astype(float)

years\_numeric = np.array([int(year) for year in years])

coefficients = np.polyfit(years\_numeric, mmr\_values, 2)

a, b, c = coefficients

Y\_pred = a \* years\_numeric\*\*2 + b \* years\_numeric + c

SST = np.sum((mmr\_values - np.mean(mmr\_values))\*\*2)

SSE = np.sum((mmr\_values - Y\_pred)\*\*2)

SSR = SST - SSE

R\_squared = SSR / SST

print(f'Quadratic regression equation: Y = {a:.4f} \* X^2 + {b:.4f} \* X + {c:.4f}')

print(f'R-squared: {R\_squared:.4f}')

print(f'SSE: {SSE:.4f}, SST: {SST:.4f}, SSR: {SSR:.4f}')

plt.figure(figsize=(10, 6))

plt.scatter(years\_numeric, mmr\_values, color='blue', label='Data Points')

plt.plot(years\_numeric, Y\_pred, color='red', label='Quadratic Regression Curve')

plt.title('Quadratic Fit: Maternal Mortality Ratio vs Year (2011-2020)')

plt.xlabel('Year')

plt.ylabel('Maternal Mortality Ratio (per 100,000 live births)')

plt.legend()

plt.grid(True)

plt.show()

**Objective 2:**

**Problem Statement 2.1:**

from math import erf, sqrt

pd.set\_option('display.max\_rows', 100);

pd.set\_option('display.max\_columns', 100);

df1 = pd.read\_csv("C:\\Users\\w10\\Desktop\\pns data\\datahemo.csv")

df2 = pd.read\_csv("C:\\Users\\w10\\Desktop\\pns data\\datavacc.csv")

df3 = pd.read\_csv("C:\\Users\\w10\\Desktop\\pns data\\Mean Systolic Blood Pressure (Age-Standardized Estimate) (18+ Years, Female) in the World (2015).csv")

print(df1)

print(df2)

print(df3)

# problem statement 2.1

africa\_data = df3[df3['ParentLocation'] == 'Africa']['Variable observation value'].dropna()

europe\_data = df3[df3['ParentLocation'] == 'Europe']['Variable observation value'].dropna()

africa\_mean = africa\_data.mean()

europe\_mean = europe\_data.mean()

africa\_std = africa\_data.std()

europe\_std = europe\_data.std()

len\_africa = len(africa\_data)

len\_europe = len(europe\_data)

z\_stat = (africa\_mean - europe\_mean) / np.sqrt((africa\_std\*\*2 / len\_africa) + (europe\_std\*\*2 / len\_europe))

def calculate\_p\_value\_z(z):

    return 2 \* (1 - (0.5 \* (1 + erf(abs(z) / sqrt(2)))))

p\_value = calculate\_p\_value\_z(z\_stat)

print("Africa:")

print(f"  Sample size: {len\_africa}")

print(f"  Mean: {africa\_mean:.2f} mmHg")

print("\nEurope:")

print(f"  Sample size: {len\_europe}")

print(f"  Mean: {europe\_mean:.2f} mmHg")

print("\nZ-Test Results:")

print(f"  Z-statistic: {z\_stat:.2f}")

print(f"  P-value: {p\_value:.3f}")

if p\_value < 0.05:

    print("\nConclusion: The difference in mean systolic blood pressure between Africa and Europe is statistically significant.")

else:

    print("\nConclusion: The difference in mean systolic blood pressure between Africa and Europe is not statistically significant.")

region\_means = df3.groupby('ParentLocation')['Variable observation value'].mean()

labels = [f"{region} ({mean:.1f})" for region, mean in zip(region\_means.index, region\_means)]

plt.figure(figsize=(10, 8))

plt.pie(region\_means, labels=labels, autopct='%1.1f%%', startangle=90, colors=plt.cm.tab20.colors)

plt.title('Mean Systolic Blood Pressure by Region (2015)')

plt.show()

**Problem Statement 2.2:**

grouped\_data = df1.groupby(['ParentLocation', 'Period(year)']).agg({'FactValueNumeric': 'mean'}).reset\_index()

plt.figure(figsize=(10, 6))

scatter = plt.scatter(grouped\_data['ParentLocation'], grouped\_data['FactValueNumeric'], c=grouped\_data['Period(year)'], cmap='viridis')

plt.colorbar(scatter, label='Period (year)')

plt.title('Scatter Plot of Fact Value Numeric vs Parent Location')

plt.xlabel('Parent Location')

plt.ylabel('Mean Fact Value Numeric')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

groups = df1.groupby('ParentLocation')['FactValueNumeric']

overall\_mean = df1['FactValueNumeric'].mean()

ssb = sum(len(group) \* ((group.mean() - overall\_mean) \*\* 2) for \_, group in groups)

ssw = sum(((group - group.mean()) \*\* 2).sum() for \_, group in groups)

df\_between = groups.ngroups - 1

df\_within = len(df1) - groups.ngroups

ms\_between = ssb / df\_between

ms\_within = ssw / df\_within

f\_statistic = ms\_between / ms\_within

print("F-statistic:", f\_statistic)

print("SSB (Between):", ssb)

print("SSW (Within):", ssw)

print("Degrees of Freedom (Between):", df\_between)

print("Degrees of Freedom (Within):", df\_within)

**Problem Statement 2.3:**

df2['Period'] = pd.to\_numeric(df2['Period'], errors='coerce').astype(int)

df2['Value'] = pd.to\_numeric(df2['Value'], errors='coerce')

df2 = df2.dropna(subset=['Period', 'Value'])

unique\_periods = sorted(df2['Period'].unique())

period\_mapping = {year: idx + 1 for idx, year in enumerate(unique\_periods)}

df2['Normalized\_Period'] = df2['Period'].map(period\_mapping)

grouped\_data = df2.groupby('Normalized\_Period')['Value'].mean().reset\_index()

X = grouped\_data['Normalized\_Period'].values

y = grouped\_data['Value'].values

degree = 2

poly\_coefficients = np.polyfit(X, y, degree)

poly\_model = np.poly1d(poly\_coefficients)

normalized\_year = len(unique\_periods) + 1

prediction = poly\_model(normalized\_year)

print(f"Polynomial Coefficients: {poly\_coefficients}")

print(f"Predicted Value for the next normalized year ({normalized\_year}): {prediction}")

X\_sorted = np.sort(X)

y\_pred = poly\_model(X\_sorted)

plt.scatter(X, y, color='blue', label='Actual Data')

plt.plot(X\_sorted, y\_pred, color='red', label=f'Polynomial Regression (Degree {degree})')

plt.scatter([normalized\_year], [prediction], color='green', label=f'Prediction for year {normalized\_year}', zorder=5)

plt.xlabel('Normalized Period')

plt.ylabel('Mean Value')

plt.title('Polynomial Regression: Mean Value vs Normalized Period')

plt.legend()

plt.show()

**Objective 3:**

**Problem Statement 3.1:**

file\_path = r'C:\Users\HP\Downloads\xmart (2).csv'

data = pd.read\_csv(file\_path)

data.columns = [

    "Country",

    "Out\_Patient\_Treatment",

    "In\_Patient\_Treatment",

    "Rehabilitation",

    "Out\_Patient\_Detoxification",

    "In\_Patient\_Detoxification",

    "Opioid\_Maintenance\_Therapy"

]

cleaned\_data = data[2:]

cleaned\_data = cleaned\_data[cleaned\_data["Country"].notna()]

cleaned\_data.reset\_index(drop=True, inplace=True)

cleaned\_data.replace("No data", pd.NA, inplace=True)

out\_patient\_counts = cleaned\_data["Out\_Patient\_Treatment"].value\_counts(dropna=False)

plt.figure(figsize=(10, 6))

bar\_chart = out\_patient\_counts.plot(kind='bar', color='skyblue', edgecolor='black')

for index, value in enumerate(out\_patient\_counts):

plt.text(index, value + 0.5, str(value), ha='center', va='bottom', fontsize=10)

plt.title('Distribution of Out-Patient Treatment Sectors by Country', fontsize=14)

plt.xlabel('Treatment Sector', fontsize=12)

plt.ylabel('Number of Countries', fontsize=12)

plt.xticks(rotation=45, ha='right', fontsize=10)

plt.tight\_layout()

plt.show()

**Problem Statement 3.2:**

data = pd.read\_csv("C:\\Users\\HP\\Downloads\\data (1) (3).csv")

average\_data = data.groupby('TimePeriod', as\_index=False)['Value'].mean()

x\_original = average\_data['TimePeriod'].values

x = x\_original - min(x\_original) + 1

y = average\_data['Value'].values

degree = 2

coefficients = np.polyfit(x, y, degree)

polynomial = np.poly1d(coefficients)

x\_extended = np.arange(1, len(x\_original) + 2)

y\_extended\_pred = polynomial(x\_extended)

plt.figure(figsize=(10, 6))

plt.scatter(x, y, alpha=0.7, edgecolors='k', label='Averaged Data')

plt.plot(x, polynomial(x), color='red', linewidth=2, label=f'Polynomial Regression (Degree {degree})')

plt.plot(x\_extended, y\_extended\_pred, color='red', linewidth=2, linestyle='--', label='Extended Prediction')

plt.title("Polynomial Regression: Time Period vs Average Value (Normalized Years)", fontsize=14)

plt.xlabel("Normalized Time Period (1, 2, 3, ...)", fontsize=12)

plt.ylabel("Average Value", fontsize=12)

plt.xticks(ticks=np.arange(1, len(x\_extended) + 1, 1))

plt.legend()

plt.grid(True, linestyle='--', alpha=0.6)

plt.show()

print("Polynomial Coefficients:", coefficients)

print("Polynomial Equation:")

print(polynomial)

**:: END OF REPORT::**