Group 10 – Final Project

Duleep Panthagani

Problem Description

- The primary goal of our analysis is to utilize a multivariate linear regression model to understand and predict life expectancy based on an array of demographic, economic, and health-related factors. Life expectancy is a crucial indicator of a population's overall health, quality of life, and development status, and dissecting its contributing elements can provide valuable insights for policy-making and health interventions.
- Our dataset comprises 17 independent variables ranging from economic indicators such as GDP and Total Expenditure to health metrics like HIV/AIDS prevalence, Hepatitis B, and BMI. Socio-demographic variables like country status (Developed or Developing), population, and years of schooling are also included. The dependent variable is Life Expectancy, which we aim to predict.

Data Summary

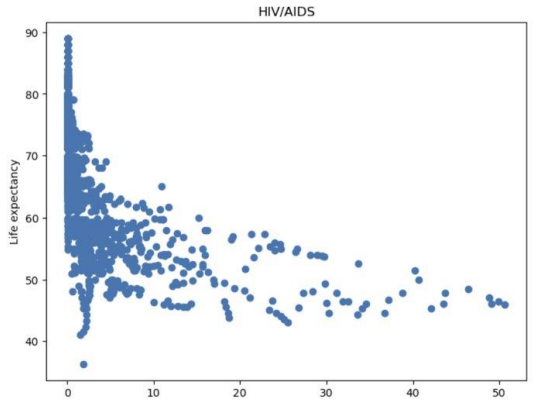
1. **Dataset Overview**: 2848 entries with 18 attributes.
2. **Attributes**: Country, Year, Status, Population, Hepatitis B, Measles, Polio, Diphtheria, HIV/AIDS, infant deaths, under-five deaths, Total expenditure, GDP, BMI, thinness 1-19 years, Alcohol, Schooling, Life expectancy.
3. **Missing Data**: Presence of null values in some columns including Population, Hepatitis B, Polio, Diphtheria, Total expenditure, GDP, BMI, thinness 1-19 years, Alcohol, Schooling.
4. **Duplicates**: No duplicate entries found.
5. **Preprocessing Steps**:
    1. Handle missing data with imputation (mean and median).
    2. Replace categorical field 'Status' with 0 (Developing) and 1 (Developed).
    3. Drop 'Country' column.

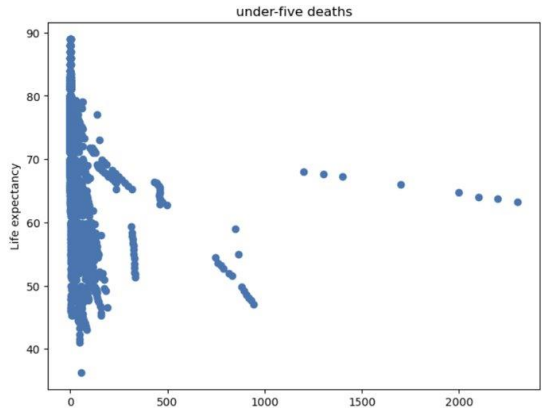
Data Visualization

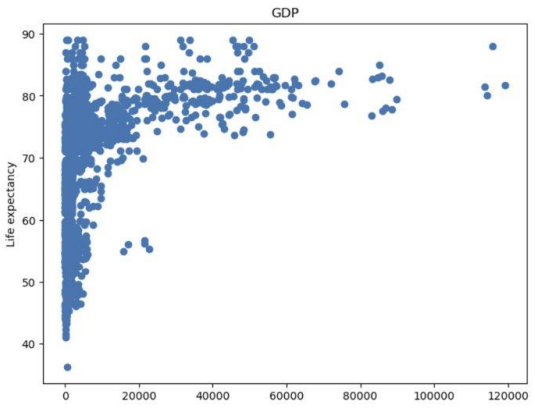
- In our analysis, we leveraged comprehensive data visualizations to understand the underlying structure of our dataset and reveal key insights. Three types of plots were generated: scatter plots, box plots, and histograms, which allowed us to study the relationships between life expectancy and various features, the influence of categorical variables, and the distribution of our dataset respectively.

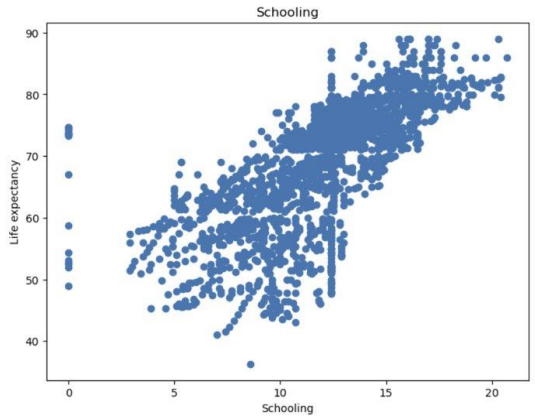
Scatter Plots

- We created scatter plots for each of the numerical features against life expectancy. Scatter plots provide a graphical view of the correlation between two variables. Some observations from these plots include:
- "Schooling" and "GDP" show a positive correlation with "Life expectancy." This indicates that higher levels of schooling and GDP generally correspond to higher life expectancy.
- "HIV/AIDS" and "Under-five deaths" show a negative correlation with "Life expectancy." This suggests that countries with higher HIV/AIDS prevalence or under-five death rates tend to have lower life expectancy



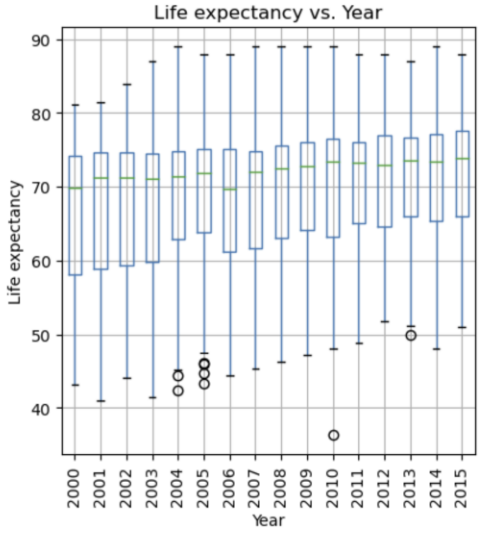


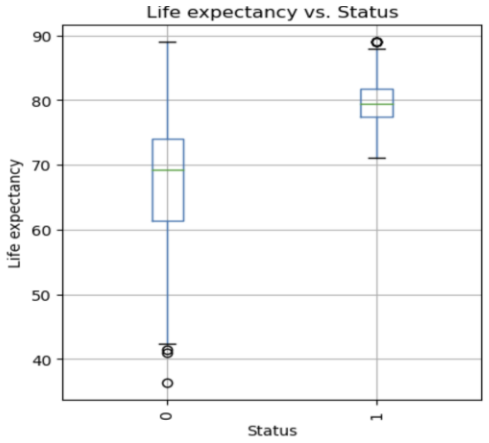




Box Plots

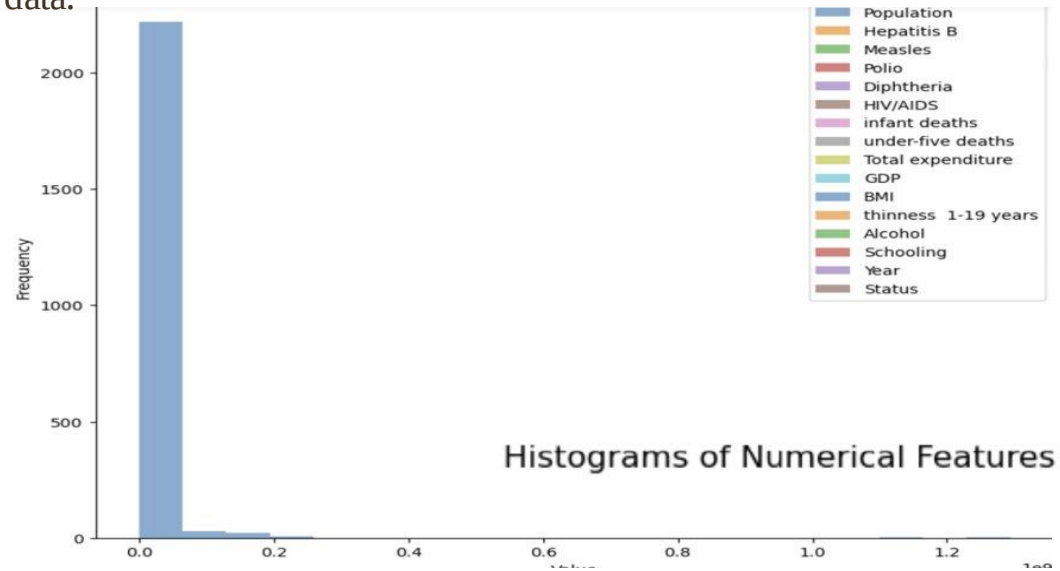
- We plotted box plots of life expectancy against the categorical features "Year" and "Status." Box plots provide a view of the data distribution and can help identify outliers.
- The box plot for "Status" shows that developed countries generally have a higher life expectancy than developing countries.
- Over the years, there seems to be an overall increase in life expectancy, as indicated by the "Year" box plot.

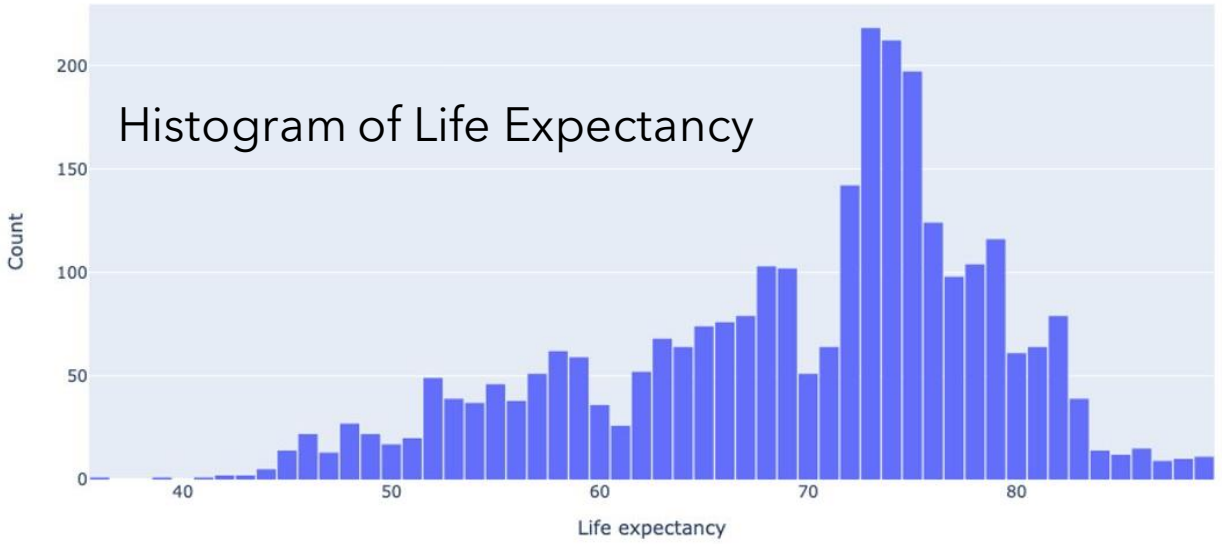




Histograms

We generated histograms for all the numerical features in our dataset. Histograms illustrate the underlying frequency distribution (shape) of the set of data.





Model Performance

Linear Regression:

- R-squared: 0.7465144327251217
- Mean Squared Error (MSE): 23.04193690827356
- Root Mean Squared Error (RMSE): 4.800201757038297

•Decision Tree Regressor (Best Performing Model):

- R-squared: 0.86044584247504
- Mean Squared Error (MSE): 12.685527336120177
- Root Mean Squared Error (RMSE): 3.56167479370593

- KNeighbors Regressor:
  - R-squared: 0.8461237513674368
  - Mean Squared Error (MSE): 13.987411002491127
  - Root Mean Squared Error (RMSE): 3.7399747328680077

- Expertise in selecting a suitable model:

In selecting the best-performing model, we typically look for the model with the highest R-squared value and the lowest values for Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The R-squared value represents how well the model fits the data, with higher values indicating a better fit. On the other hand, MSE and RMSE measure the accuracy of the predictions, where lower values indicate better accuracy.

- After evaluating the three models, we can compare their respective R-squared, MSE, and RMSE values. The model with the highest R-squared and the lowest MSE and RMSE is considered the best-performing model. Therefore, the best model among the three is the one that shows the highest R-squared value and the lowest MSE and RMSE values in the evaluation results

Results Visualization

- **Final Model Performance:**
  - R-squared: 0.9292499305114745
    - **~93% of the variation in life expectancy can be attributed to the input features used in our model.**
  - Root Mean Squared Error (RMSE): 2.616575702392872
    - **Average prediction error in life expectancy units.**
- **Distribution of Real and Predicted Values:**
  - The histogram demonstrates that the distribution of predicted values closely resemble that of actual values.
  - This suggests that our model captures the underlying patterns in the data and produces predictions that align well with the real-life expectancy values.
- **Residual Plot:**
  - In the plot, most points are scattered around the horizontal line at y = 0, indicating that our model's predictions have relatively small errors and are centered on the correct values.
  - The lack of any noticeable pattern in the residuals suggests that our model is unbiased and captures the life expectancy variations effectively.
- **Actual Values vs. Predicted Values:**
  - The scatterplot provides a direct assessment of how well our model's predictions align with the true values for each data point.
- **Predicted vs. True Line Plot:**
  - The plot shows a strong alignment between the true-life expectancy and the model's predictions.
  - The points closely follow the trend of the ideal line, indicating accurate predictions. The regression line further supports the model's effectiveness in approximating the true values
- **Feature Importance for Final Model:**
  - The bar chart highlights the relative importance of each feature in predicting life expectancy.
  - Features with higher coefficient values are more influential in determining life expectancy variations.