**Project Summary: Life Expectancy Analysis:**

**Ivan Chavez, Jessica McPhaul, Rafia Mizra**

# Introduction and Objective

Analysis of life expectancy using WHO data to model country life expectancy based on economic and health factors using regression models.

* + This project focuses on analyzing a dataset to build regression models. The dataset used is detailed, encompassing various economic and health factors to model a country's life expectancy. The primary aim is to explore key relationships using regression analysis and interpret these in a structured manner.

# Data Description and Processing

Data from WHO, covering health and economic indicators. Involves data cleaning, preprocessing, and exploratory data analysis (EDA) to identify key patterns and correlations.

* + The dataset utilized in this analysis originates from the World Health Organization (WHO) and is publicly available on Kaggle. It encompasses an extensive collection of 2,848 records, each representing annual health and economic indicators for various countries. The dataset includes a range of variables such as Life Expectancy, Adult Mortality, Alcohol Consumption, GDP per Capita, among others. These variables offer a comprehensive view of the health status and economic conditions across different nations.
  + The preprocessing of the dataset involved several critical steps to ensure its reliability and validity for regression analysis. The primary focus was on handling missing values and refining the dataset structure. For numerical variables, missing values were substituted with the median of the respective columns, while categorical variables were imputed using the mode. This approach aimed to preserve the dataset's integrity without introducing significant bias.
  + In addition to imputation, columns with a substantial number of missing values were eliminated from the dataset to maintain the quality and consistency of the analysis. Furthermore, to enhance clarity and facilitate data handling, column names were systematically standardized.
  + In addition to imputation, columns with a substantial number of missing values were eliminated from the dataset to maintain the quality and consistency of the analysis. Furthermore, to enhance clarity and facilitate data handling, column names were systematically standardized.

# Exploratory Data Analysis (EDA):

Data EDA revealed interesting patterns and correlations. Key visualizations include box plots showing life expectancy variation by economy status and time, and correlation matrices highlighting relationships between variables. Scatter plots of selected variables against life expectancy underscored specific trends. These analyses provided foundational insights for model building, emphasizing the relationship between health indicators and life expectancy.

# Objective 1: Regression Model Development and Analysis

Enhanced linear regression model development with variable selection based on statistical significance. Interpretation of coefficients to understand variable impact on life expectancy.

1. Model Fitting Approach:
   * We utilized an enhanced linear model, emphasizing variables that significantly contribute to predicting life expectancy. This iterative refinement process involved assessing statistical significance and model diagnostics.
2. Variable Selection:
   * Variables were selected based on their impact on life expectancy, as evaluated through statistical significance and exploratory data analysis (EDA). Notable predictors included Adult Mortality, BMI, and GDP per Capita.
3. Feature Selection Summary:
   * We employed both manual selection and statistical techniques to refine the model, ensuring it included only significant predictors. This approach balanced model complexity and interpretability.
4. Final Regression Model Definition:
   * Our final model incorporated various predictors, such as Adult Mortality, Alcohol Consumption, GDP per Capita, and other health and economic indicators. Each predictor's coefficient was calculated to understand its unique contribution to life expectancy.
5. Coefficients Summary Table:
   * The analysis included a comprehensive table of coefficients, presenting estimates, standard errors, and significance levels for each predictor. This table is essential for interpreting the influence of each variable on life expectancy.
6. Regression Coefficient Interpretation:
   * Interpretations for key coefficients, like Adult Mortality and Alcohol Consumption, were provided, focusing on their magnitude and significance. These interpretations helped understand how changes in these variables impact life expectancy.
7. Model Evaluation and Comparison:
   * The model's performance was evaluated using visualizations and statistical analyses like ANOVA.
   * These evaluations helped compare different model iterations and validate the final model choice

# Objective 2: Comprehensive Summary of Our Enhanced Regression Modeling, Further Model Evaluation and Comparative Analysis of Our Enhanced Regression Modeling for Life Expectancy Prediction

Evaluation and comparison of different regression models, focusing on predictive accuracy and suitability for life expectancy prediction.

In this crucial phase, we focused on enhancing the complexity of our initial regression model for improved predictive performance, guided by our comprehensive exploratory data analysis (EDA) and iterative model development strategies.

1. **Advanced Model Development and Iterative EDA:**
   * We significantly increased the complexity of our regression model by incorporating sophisticated techniques such as variable interactions, non-linear relationships, and polynomial terms where appropriate.
   * Our iterative EDA, both initial and extended, was pivotal in uncovering intricate and especially non-linear relationships and interactions between variables, guiding our model development choices.
2. **In-Depth Model Refinement:**
   * Utilizing the insights from our ongoing EDA, we continuously integrated complex terms into our model. This ensured that our model evolved dynamically, reflecting new insights and patterns discovered during the analysis.

# Methodology and Decision Making

Iterative approach balancing model complexity and interpretability. Emphasis on data-driven decision-making in the modeling process.

1. **Comprehensive Model Comparison and Inclusion of Nonparametric Models:**
   * Our comparative analysis spanned multiple models, including enhanced linear models with added complexities and at least one nonparametric model, such as Random Forest or KNN.
   * This diverse strategy was crucial in assessing each model's capability to capture complex relationships and enhance predictive performance.
2. **Extended Model Evaluation Metrics:**
   * To compare and evaluate these models effectively, we employed a variety of metrics including Mean Squared Error (MSE), R-squared/Adjusted R-squared, and additional metrics like AIC or BIC.
   * These metrics were instrumental in providing insights into each model's accuracy, explanatory power, and overall suitability for predicting life expectancy.
3. **Detailed Comparative Analysis and Final Recommendations:**
   * We conducted a detailed comparative analysis of the models based on their performance, focusing on their strengths and limitations in the context of life expectancy prediction.
   * Our final recommendations were made by selecting the most suitable model(s) based on a balance between model complexity, interpretability, and predictive accuracy.
4. **Key Focuses of Our Approach:**
   * **Iterative Process**: We prioritized an iterative approach in both EDA and model refinement, allowing for continuous adjustments and improvements.
   * **Balance and Flexibility**: We maintained a balance between complexity and interpretability and stayed flexible to adapt our models as new insights were gained.
   * **Data-Driven Decision Making**: Every step, from the initial EDA to the final model selection, was guided by data-driven insights.

This consolidated approach represents our commitment to leveraging a structured methodology for clear and directed progression, while also embracing an iterative and exploratory process for depth and adaptability in modeling life expectancy

# Conclusion

Model\_1 was the best model for our analysis. RFE, VIF, AIC/BIC/MSE/RMSE/MAE/ANOVA/MANOVA processes were all used to train, test and validate the models.