Project Documentation: Exploratory Data Analysis using Python

# **Project Information**

|  |  |
| --- | --- |
| Title | LIFE EXPECTANCY DATA ANALYSIS |
| Name | INDHARAJITH |
| Course/Subject | DADS |
| Batch | DATA ANALYTICS & DATA SCIENCE-JUN-RE |
| Online/Offline | OFFLINE |
| Roll Number | 16625CBRE18 |

INDEX

|  |  |  |
| --- | --- | --- |
| S.NO | TABLE OF CONTENT | Pg.no |
| 1 | Introduction | 3 |
| 2 | Aim | 3 |
| 3 | Buisness Problem / Problem Statement | 3 |
| 4 | Project Workflow | 3-4 |
| 5 | Data Understanding | 4 |
| 6 | Data Cleaning | 4-5 |
| 7 | Obtaining Derived Metrics | 5 |
| 8 | Filtering Data for Analysis | 5 |
| 9 | Statistical Analysis | 5 |
| 10 | Exploratory Data Analysis (EDA) - Univariate Analysis | 5-6 |
| 11 | Bivariate Analysis | 6-7 |
| 12 | Multivariate Analysis | 7-8 |
| 13 | Overall Insights from Analysis | 8-9 |
| 14 | Conclusion | 9 |

**1. Introduction**

Life expectancy is one of the most important measures of public health and socioeconomic progress. It reflects the average number of years a person can expect to live based on prevailing health, education, and economic conditions. Understanding the factors that affect life expectancy helps policymakers, researchers, and governments design strategies to improve healthcare, reduce mortality, and enhance the quality of life.

This project uses global life expectancy data covering multiple years and countries, with variables such as GDP, Schooling, Status (Developed/Developing), Adult Mortality, Infant Mortality, HIV/AIDS prevalence, and Health Expenditure. By analyzing these indicators, we aim to uncover patterns, disparities, and actionable insights.

**2. Aim**

The main aim of this project is to:

* Analyze socioeconomic and health-related indicators affecting life expectancy.
* Compare Developed vs Developing countries.
* Identify the most influential factors linked with longevity.
* Provide recommendations for improving health outcomes globally.

**3. Problem Statement**

Global inequality in life expectancy persists. Developing countries often face high mortality rates, weak healthcare systems, and limited education opportunities, while developed countries show consistent improvements. The challenge lies in identifying which factors most strongly influence life expectancy, so that policymakers can target interventions effectively.

**4. Project Workflow**

* Data Collection: Dataset obtained from WHO and UN sources.
* Data Cleaning: Handling missing values, correcting inconsistencies, and filtering reliable observations.
* Exploratory Data Analysis (EDA): Univariate, bivariate, and multivariate analyses to understand distributions and relationships.
* Statistical Analysis: Hypothesis testing, correlation, and regression modeling.
* Visualization: Histograms, boxplots, violin plots, scatter plots, and 3D visualizations for insights.
* Conclusion & Recommendations: Actionable takeaways for policy and further research.

**5. Data Understanding**

* Rows: ~2,938
* Columns: 22
* Time Period: 2000–2015
* Key Variables:
  + Country – Nation name
  + Year – Year of observation
  + Status – Developed / Developing
  + Life expectancy – Target variable
  + GDP – Economic indicator
  + Schooling – Average years of education
  + Adult Mortality, Infant Deaths, HIV/AIDS – Health burden indicators
  + Health Expenditure – Share of GDP spent on healthcare

**6. Data Cleaning**

* Missing Values: Imputed using median for numerical values and mode for categorical.
* Outliers: Identified via boxplots and capped where necessary.
* Column Standardization: Removed leading/trailing spaces, renamed columns consistently.
* Filtered Years: Focused on 2000–2015 for consistency across nations.

# **OUTLIERS**[**¶**](https://nb.anaconda.com/jupyterhub/user/f5e974a8-0118-4e06-a626-00d02a003c87/lab/workspaces/auto-f/tree/Untitled1.ipynb?#OUTLIERS)

A graph with a blue line

AI-generated content may be incorrect.

**7. Derived Metrics**

* Life Expectancy Gap between Developed and Developing countries.
* GDP-to-Schooling Ratio: Proxy for education efficiency.
* Mortality-adjusted expectancy: Adjusted life expectancy after factoring mortality rates.

**8. Statistical Analysis**

* Descriptive Statistics: Mean life expectancy is ~78 years for Developed vs ~65 years for Developing.
* Hypothesis Testing: Independent t-test confirms a statistically significant difference (p < 0.01) in life expectancy between groups.
* Correlation: Positive correlation between GDP, Schooling, and Life Expectancy; negative correlation with mortality and HIV/AIDS prevalence.

**9. Exploratory Data Analysis**

**Univariate Analysis**

* Life expectancy in developing nations shows a right-skewed distribution.
* GDP distribution is highly skewed, with a few wealthy nations driving the upper tail.
* Schooling varies between 8–15 years, higher in Developed nations.

A graph with red lines

AI-generated content may be incorrect.

**Bivariate Analysis**

* GDP vs Life Expectancy: Strong positive relationship.
* Schooling vs Life Expectancy: Higher schooling corresponds to higher life expectancy.
* Status vs Life Expectancy: Developed nations consistently outperform developing.

A screen shot of a graph

AI-generated content may be incorrect.

**Multivariate Analysis**

* 3D Scatter of GDP, Schooling, and Life Expectancy shows clear clustering of developed vs developing nations.
* Regression Analysis: GDP, Schooling, and HIV/AIDS emerge as the strongest predictors of life expectancy.

A screenshot of a computer

AI-generated content may be incorrect.

**10. Overall Insights**

* GDP and Schooling are the most influential factors for longer life expectancy.
* Healthcare burden indicators (HIV/AIDS, Infant Mortality, Adult Mortality) strongly reduce life expectancy.
* The Developed vs Developing gap persists, even though both groups improved over time.
* Education acts as a multiplier: More schooling → better jobs → higher income → longer, healthier life.

**11. Conclusion & Recommendations**

**Conclusion**

Life expectancy is a multi-dimensional outcome influenced by economic, educational, and health factors. While global progress is visible, disparities remain between developed and developing countries. Education and GDP emerge as the strongest contributors, while health burdens continue to suppress life expectancy in low-income nations.

**Recommendations**

* Strengthen Education Systems: Ensure universal access to schooling to boost awareness and long-term health.
* Healthcare Investments: Prioritize infrastructure and access in low-GDP countries.
* Public Health Campaigns: Increase vaccination, sanitation, and nutrition awareness in developing regions.
* Economic Development Programs: Drive GDP growth through employment, trade, and innovation.
* Target Vulnerable Populations: Focus on rural and marginalized groups to reduce inequality.

**12. Future Work**

* Extend dataset with more recent years (post-2015) to capture modern trends, including pandemic impacts.
* Include environmental and lifestyle factors (pollution, obesity, climate change) for deeper analysis**.**
* Develop predictive models (e.g., machine learning regressors) to forecast life expectancy.
* Perform regional-level analysis for tailored policy recommendations.