EXPLORATORY DATA ANALYSIS
Life Expectancy Dataset
Comprehensive Summary Report

Dataset	WHO Life Expectancy — 152 Countries, 2000–2015
Records	2,204 rows × 5 columns
Status	Complete — No Missing Values
Prepared by	Automated EDA Pipeline
Date	February 27, 2026

1. Dataset Overview

The Life Expectancy dataset contains records for 152 countries spanning 16 years (2000–2015), sourced from the World Health Organization (WHO). The dataset is compact — just 5 columns — but provides rich information linking infant mortality to life expectancy outcomes across developing nations.

1.1 Structure & Memory

Total Records	2,204
Total Features	5
Numerical Features	3 (Year, LifeExpectancy, InfantMortality)
Categorical Features	2 (Country, Status)
Memory Usage	0.33 MB
Year Range	2000 – 2015 (16 years)
Countries Covered	152

1.2 Column Inventory

Column	Type	Non-Null	Unique Values	Notes
Country	object	2,204	152	High cardinality — all developing
Year	int64	2,204	16	2000 to 2015
LifeExpectancy	float64	2,204	340	**Target variable**
InfantMortality	float64	2,204	69	Scaled / normalised values
Status	object	2,204	1	All rows = 'Developing'

2. Data Quality Assessment

✓ **Zero Missing Values**

All 2,204 records are fully populated across all 5 columns. No imputation or row removal is required, which is exceptional for a real-world WHO dataset spanning 16 years.

2.1 Missing Value Analysis

Overall Missing Data %	0.00%
Columns with Missing Data	0 of 5
Rows Requiring Removal	0

Imputation Required	None

2.2 Data Quality Flags

While the dataset has no missing values, two data quality observations are noteworthy:

> **⚠ InfantMortality — Appears Standardised**
>
> The InfantMortality column has values ranging from -1.25 to 1.80, with a mean of 0.00 and std of 1.00. This is characteristic of z-score normalisation applied before the EDA. The column is not in its original unit (deaths per 1,000 live births). Coefficient of Variation is mathematically undefined (mean ≈ 0).

> **⚠ Status Column — Zero Variance**
>
> The Status column contains only one unique value: 'Developing' (100% of records). This column carries no predictive information and should be dropped before model training to avoid noise.

3. Target Variable Analysis — Life Expectancy

Life Expectancy is the primary target variable for regression modelling. It represents the average number of years a newborn is expected to live, averaged at a country-year level.

3.1 Descriptive Statistics

Mean	67.85 years
Median	69.90 years
Std Dev	8.72 years
Minimum	42.30 years
Maximum	89.00 years
Range	46.70 years
IQR (Q1–Q3)	62.30 – 74.20 years (IQR = 11.90)

3.2 Distribution Shape

Skewness	-0.710 (moderate negative skew)
Kurtosis	-0.175 (platykurtic — flatter than normal)
Shapiro-Wilk	Statistic = 0.9430 | p-value = 0.0000
Normality	Rejected at $p < 0.05$ — not perfectly normal

The negative skew indicates that while most countries cluster in the 62–74 year range (middle-to-high life expectancy), a tail of lower-performing countries pulls the mean (67.85) below the median (69.90). The Shapiro-Wilk result confirms the distribution is not perfectly Gaussian, though the deviation is moderate and unlikely to severely impact regression model performance.

3.3 Interpretation

- The 46.70-year spread (42.3 → 89.0) reflects profound global inequality in health outcomes between developing nations.
- The IQR of 11.90 years indicates that the middle 50% of countries have life expectancy between 62.3 and 74.2 years.
- 8 outliers were detected via IQR — all within the plausible human range (44.45 – 92.05 bounds), suggesting no data errors.
- Zero Z-score outliers ($|z| > 3$) confirms no extreme anomalies in the target variable.

4. Feature Analysis

4.1 Year

Range	2000 – 2015
Mean	2007.62
Std Dev	4.61
Skewness	-0.031 (nearly symmetric)
Kurtosis	-1.212 (uniform-like distribution)

Year is nearly uniformly distributed — as expected for a balanced panel dataset. Most countries contribute 16 observations (one per year), though 10 countries have fewer records due to data availability issues (e.g. Philippines: 15, Kenya: 10, Brazil: 9, Afghanistan: 3).

4.2 InfantMortality (Standardised)

Range	-1.25 to 1.80
Mean	0.00 (z-score centred)
Std Dev	1.00 (z-score scaled)
Skewness	0.222 (slight positive skew)
Kurtosis	-1.291 (flatter than normal)

InfantMortality has been standardised to a z-score before this EDA. Higher (more positive) values indicate higher-than-average infant mortality; lower (more negative) values indicate better-than-average infant survival. Zero outliers were detected under both IQR and Z-score methods, confirming clean standardisation.

5. Categorical Variable Analysis

5.1 Country (152 Unique Values)

Country is a high-cardinality identifier column with 152 unique values. The dataset is largely balanced at the country level:

Most frequent countries	16 observations each (majority)
Countries with 16 obs	~140 of 152
Partially represented	Philippines (15), Sierra Leone (13), Sudan (12), Myanmar (12), Kenya (10), Brazil (9)
Sparse countries	Afghanistan (3), Uganda (2), Angola (2)
Single-year countries	8 countries with only 1 observation

Modelling Note — Country Variable

Country should NOT be used as a raw feature in regression models due to its 152 unique string values. Instead, consider: (1) One-hot encoding for tree-based models, (2) Target encoding (mean life expectancy per country), or (3) Dropping Country and relying on numerical features only. Countries with fewer than 5 observations may introduce noise and could be grouped or excluded.

5.2 Status (Zero Variance — Drop)

Unique Values	1
Only Value	Developing (100% of records)
Predictive Value	None — zero variance
Recommendation	Drop before modelling

⚠ Action Required

The Status column is a constant ('Developing') and contains absolutely no information for model training. Including it after encoding would introduce a collinear dummy variable. It must be dropped during the preprocessing step.

6. Correlation Analysis

6.1 Pearson Correlation Matrix

Variable	Year	LifeExpectancy	InfantMortality
Year	**1.000**	0.166	-0.028
LifeExpectancy	0.166	1.000	-0.538

Variable	Year	LifeExpectancy	InfantMortality
InfantMortality	-0.028	-0.538	1.000

6.2 Feature Correlations with Life Expectancy

Feature	Pearson r	Interpretation
Year	+0.166	Weak positive — slight upward global trend
InfantMortality	-0.538	Moderate-strong negative — key predictor

InfantMortality is the strongest predictor of Life Expectancy (r = -0.538), confirming the well-established epidemiological link between child survival and overall population longevity. Year shows a modest positive trend, suggesting marginal global health improvements over the 2000–2015 period in developing countries.

6.3 Multicollinearity Assessment

✓ No Multicollinearity Detected

No feature pair exceeds the |r| > 0.80 threshold. Year vs InfantMortality (r = -0.028) is essentially uncorrelated. All features can safely be used together in a multiple regression model without VIF concerns.

7. Outlier Detection

7.1 Summary Table

Feature	IQR Outliers	IQR %	Z-Score Outliers	Bounds (IQR)
LifeExpectancy	8	0.36%	0	44.45 – 92.05
Year	0	0.00%	0	1992 – 2024
InfantMortality	0	0.00%	0	-3.24 – 3.40

7.2 Interpretation

- LifeExpectancy has 8 mild IQR outliers (0.36%) — all within biologically plausible bounds (44–89 years). No action required.
- Year has zero outliers, confirming the panel covers 2000–2015 without gaps or errors.
- InfantMortality has zero outliers under both methods, consistent with clean z-score standardisation.
- No Z-score extreme outliers (|z| > 3) across any feature — the dataset is clean and ready for modelling.

8. Temporal Analysis (2000–2015)

The dataset spans 16 years of health data for developing countries. Temporal analysis reveals the following macro-trends:

8.1 Life Expectancy Trend

- Global average life expectancy showed a positive upward trend from 2000 to 2015.
- Year correlates at r = +0.166 with Life Expectancy — modest but consistent improvement.
- The trend reflects healthcare advances, vaccination programs, and improved sanitation in developing nations over this period.

8.2 InfantMortality Trend

- Infant mortality showed a declining trend over the 16-year period (on the standardised scale, values shifted toward more negative, i.e. better outcomes).
- The correlation between Year and InfantMortality (r = -0.028) is near-zero, suggesting the overall standardised scale obscures within-country temporal variation.

Coverage Note

Some countries have incomplete time series. Philippines (15 years), Sierra Leone (13), Sudan and Myanmar (12 each), Kenya (10), Brazil (9) and several others have fewer than 16 data points. This imbalanced panel should be considered when building time-series or country-level models.

9. Regression Modelling Readiness

9.1 Readiness Checklist

Check	Result	Status
Target variable present (LifeExpectancy)	Yes	✓ PASS
Sufficient data points (≥ 100)	2,204	✓ PASS
Multiple numerical features available	3	✓ PASS
No excessive missing values (< 50%)	0.00%	✓ PASS
Target variable is continuous (float64)	Yes	✓ PASS

🎉 **Dataset is READY for Regression Modelling**

All five readiness criteria pass. The dataset can be fed directly into regression pipelines after the preprocessing steps outlined in Section 10.

9.2 Recommended Models

Model	Best Use Case	Key Consideration
Simple Linear Regression	Quick baseline using InfantMortality only	r = -0.538 gives decent R^2
Multiple Linear Regression	All numerical features combined	Low multicollinearity ✓
Polynomial Regression	Capture non-linear LE vs mortality curve	Try degree 2; watch overfitting

10. Preprocessing & Next Steps

10.1 Required Preprocessing Steps

1. Drop the Status column — zero variance, no predictive value.
2. Decide on Country encoding: target-encode, drop, or one-hot for tree models.
3. Confirm InfantMortality units — if already standardised, no further scaling needed for that column; otherwise apply StandardScaler.
4. Scale Year and LifeExpectancy for distance-based or gradient-descent models (MinMaxScaler or StandardScaler).
5. Train-test split — recommended 80/20 stratified by Year to avoid temporal leakage.
6. Handle sparse countries (< 5 observations) — exclude or group into 'Other'.

10.2 Modelling Roadmap

7. Build Simple Linear Regression baseline: InfantMortality → LifeExpectancy.

8. Build Multiple Linear Regression: all numerical features → LifeExpectancy.
9. Test Polynomial Regression (degree 2) to capture any non-linear mortality–longevity curve.
10. Evaluate all models on: R^2, RMSE, MAE on held-out test set.
11. Perform residual analysis — check for heteroscedasticity and normality of residuals.
12. Compare models and select the best-performing approach.
13. Optional: explore Ridge/Lasso regularisation if overfitting occurs.

10.3 Evaluation Metrics to Track

R^2 (Coefficient of Determination)	Variance explained — target > 0.70
RMSE (Root Mean Squared Error)	Penalises large errors — lower is better
MAE (Mean Absolute Error)	Average absolute error in years — interpretable
Adjusted R^2	Penalises extra features — use for MLR comparison

11. Key Findings & Insights

11.1 Top Findings

- **InfantMortality is the dominant predictor of Life Expectancy (r = -0.538) — countries with lower infant mortality rates consistently achieve higher life expectancy.**
- **The dataset is 100% clean with no missing values, which is uncommon for multi-country, multi-year WHO data and allows immediate modelling.**
- **Global life expectancy in developing countries improved marginally from 2000–2015, despite the dataset covering only the 'Developing' tier.**
- **Status column is a data quality issue — a single constant category provides zero information and must be removed.**
- **InfantMortality appears already z-score standardised, suggesting this is a pre-processed subset of the full WHO dataset.**

11.2 Limitations

- Only developing countries are represented — findings cannot be generalised to developed nations.
- With only 2 predictors available for modelling (Year, InfantMortality), model complexity is inherently limited.
- The standardised InfantMortality values make domain interpretation less intuitive (cannot say 'X deaths per 1,000 births').
- 10 countries have incomplete time series — panel is unbalanced.
- The dataset likely represents a filtered or pre-processed subset of the original WHO Life Expectancy dataset which contains 20+ features.

EDA Complete — Ready for Modelling

2,204 records | 0% missing | 152 countries | 2000–2015 | 5 features