Dataset Search and Description

Dataset Information :

- **Dataset Name:** *Life Expectancy and GDP per Capita*
- **Source:** Our World in Data
- **License:** Open source (CC BY)
- **Observations:** Over 100 countries with recent GDP and life expectancy data
- **Variables Selected:**
    - **X (Independent Variable):** *GDP per capita* (in international dollars, inflation-adjusted)
    - **Y (Dependent Variable):** *Life Expectancy at Birth* (in years)

Dataset Context :

This dataset, compiled by Our World in Data, combines economic and demographic statistics from the World Bank and United Nations to explore how a country's wealth relates to the average lifespan of its citizens. It provides a real-world example of socioeconomic development and public health outcomes across nations.

For this project, we focus on the relationship between a country's GDP per capita and its Life Expectancy at birth. This relationship is widely known to be nonlinear and saturating, making it ideal for testing nonparametric smoothing methods.

Variable Descriptions:

The dataset contains several socio-economic and demographic indicators for multiple countries over different years. Below is a detailed description of each variable:

Variable Name	Description	Type
Entity	Name of the country or region (e.g., India, United States, Japan). Represents the observational unit for which the data is recorded.	Categorical
Code	Three-letter ISO country code corresponding to each Entity (e.g., IND, USA, JPN). Useful for data merging and referencing.	Categorical
Year	The calendar year corresponding to the observation. Allows tracking changes over time and conducting temporal analysis.	Integer
Period life expectancy at birth	The average number of years a newborn is expected to live if current mortality rates continue throughout their lifetime. This is the **dependent variable (Y)** in the project.	Continuous
GDP per capita	The average economic output (gross domestic product) per person, measured in international dollars adjusted for inflation and purchasing power parity (PPP). This is the **independent variable (X)**.	Continuous
900793-annotations	Metadata column containing additional notes or references for some records (e.g., data source, quality flags). Not used in analysis.	Text / Categorical
Population (historical)	Total population of the corresponding entity and year, based on historical records or UN estimates. Useful for contextual analysis but not directly used in this project.	Continuous
World regions according to OWID	Geographic or economic region classification assigned by *Our World in Data* (e.g., Asia, Europe, Sub-Saharan Africa). Useful for grouping or stratified analysis.	Categorical

Why This Dataset Was Chosen :

I selected this dataset because it captures an important and interpretable real-world phenomenon — how economic prosperity affects human well-being. The data is clean, continuous, and publicly available, making it well-suited for regression modelling. Moreover, the nonlinear pattern between wealth and health provides a clear opportunity to apply smoothing techniques.

Expected Relationship :

The relationship between GDP and Life Expectancy is **nonlinear**. At lower income levels, small increases in GDP are associated with **large gains in life expectancy** (due to improvements in nutrition, healthcare, and sanitation). However, as income rises further, these gains **diminish**, leading to a **saturating or logarithmic curve**.

This shape suggests diminishing returns — a perfect case for nonparametric regression.

Suspected Functional Form :

The relationship is expected to follow a **saturating nonlinear pattern**, approximated by a **logarithmic** or **exponential decay** type curve:

$$\text{Life Expectancy} = a + b \cdot \log(\text{GDP per capita}) + \varepsilon$$

Nonparametric smoothers (e.g., kernel, LOESS) can capture this shape **without assuming a specific functional form**.

Data Cleaning and Exploration

Data Selection and Subsetting :

From the original dataset of 2,855 rows and 8 columns, only two continuous variables relevant to this analysis were selected:

- **Independent Variable (X): GDP per capita**
- **Dependent Variable (Y): Life expectancy at birth**

The selected subset was renamed for clarity as GDP_per_capita and Life_expectancy.

Handling Missing Values :

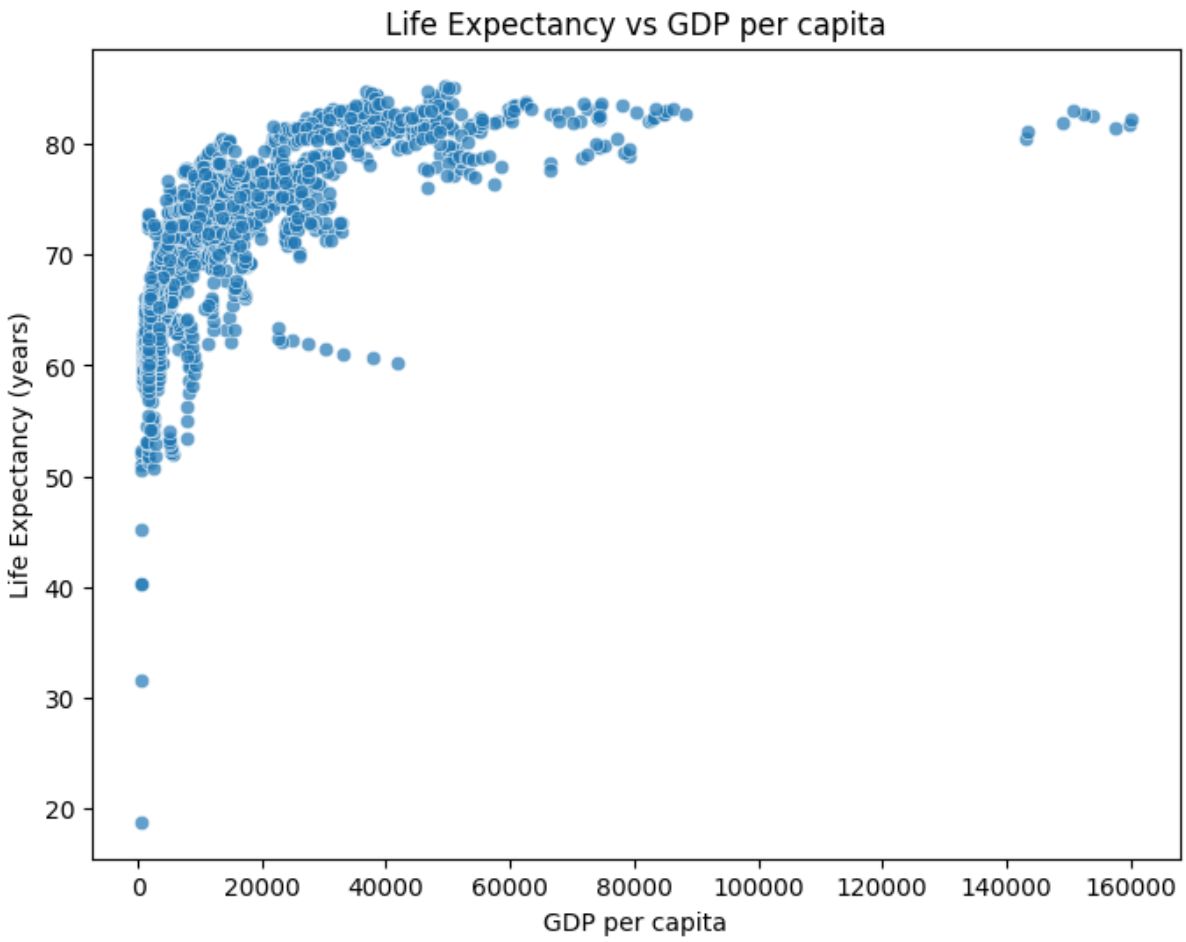
A check for missing values revealed:

- **Life_expectancy: 245 missing values**
- **GDP_per_capita: 1,262 missing values**

After removing rows with missing entries in either variable, the dataset was reduced to 1,493 valid observations, ensuring clean and consistent data for further analysis.

Plotting The Data :

After handling the missing values, the scatterplot of the two variables looks like this :



Outlier Detection and Removal :

Outliers were detected using the **Interquartile Range (IQR) method**:

$$IQR = Q3 - Q1 \quad \text{,and values outside } [Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$$

were considered outliers and removed for both variables (GDP_per_capita and Life_expectancy).

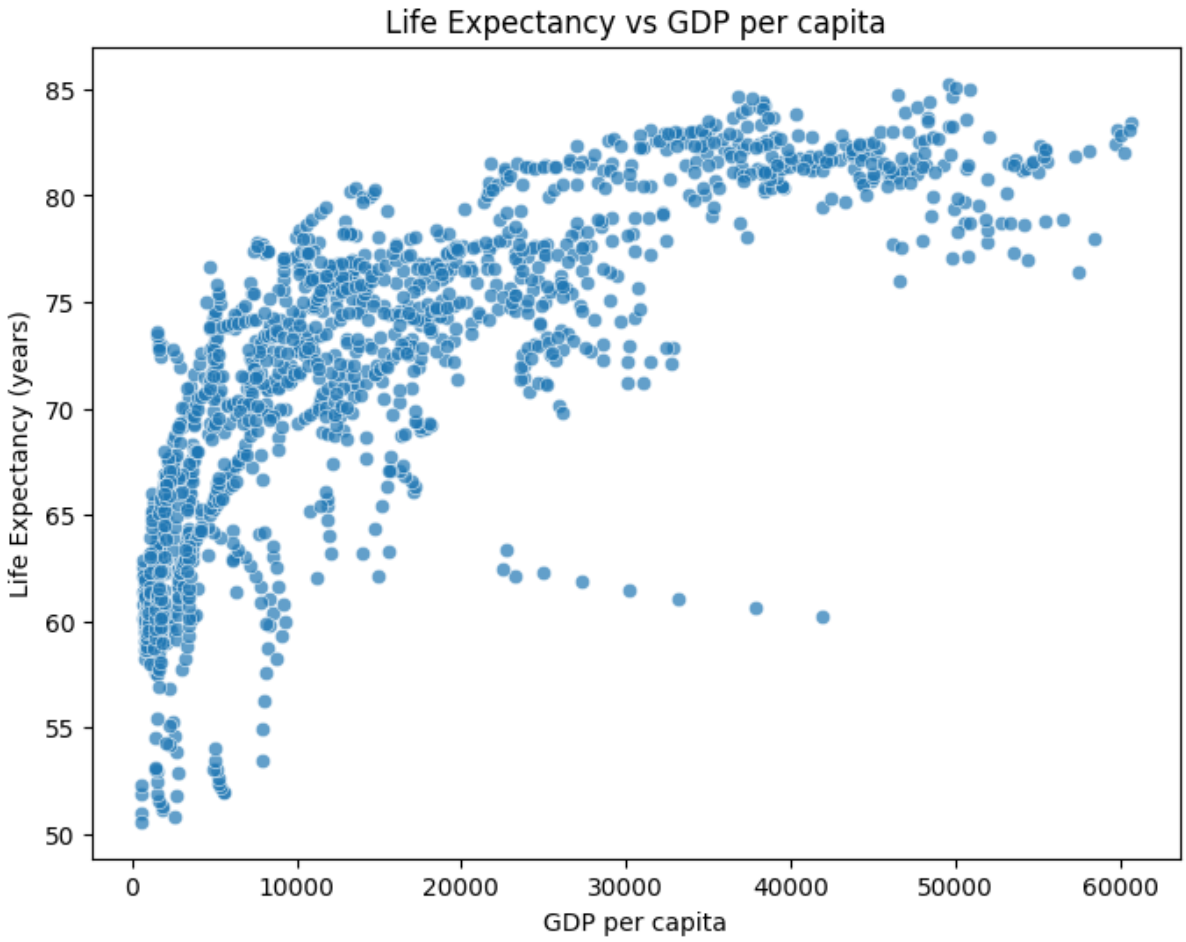
After removing outliers, the number of rows decreased from **1,493 to 1,438**, indicating that **55 extreme values** were identified and filtered out.

This step ensured that the remaining data represented the main global trend without being distorted by a few unusually high or low values.

Re-Plot and Pattern Description :

The cleaned scatterplot of Life Expectancy vs GDP per capita showed a **smooth, nonlinear increasing trend**:

- At low GDP levels, life expectancy increases sharply.
- Beyond middle-income levels, the growth slows, and the curve begins to flatten.
  This suggests a **saturating relationship** between wealth and longevity — consistent with known global health-economic patterns.

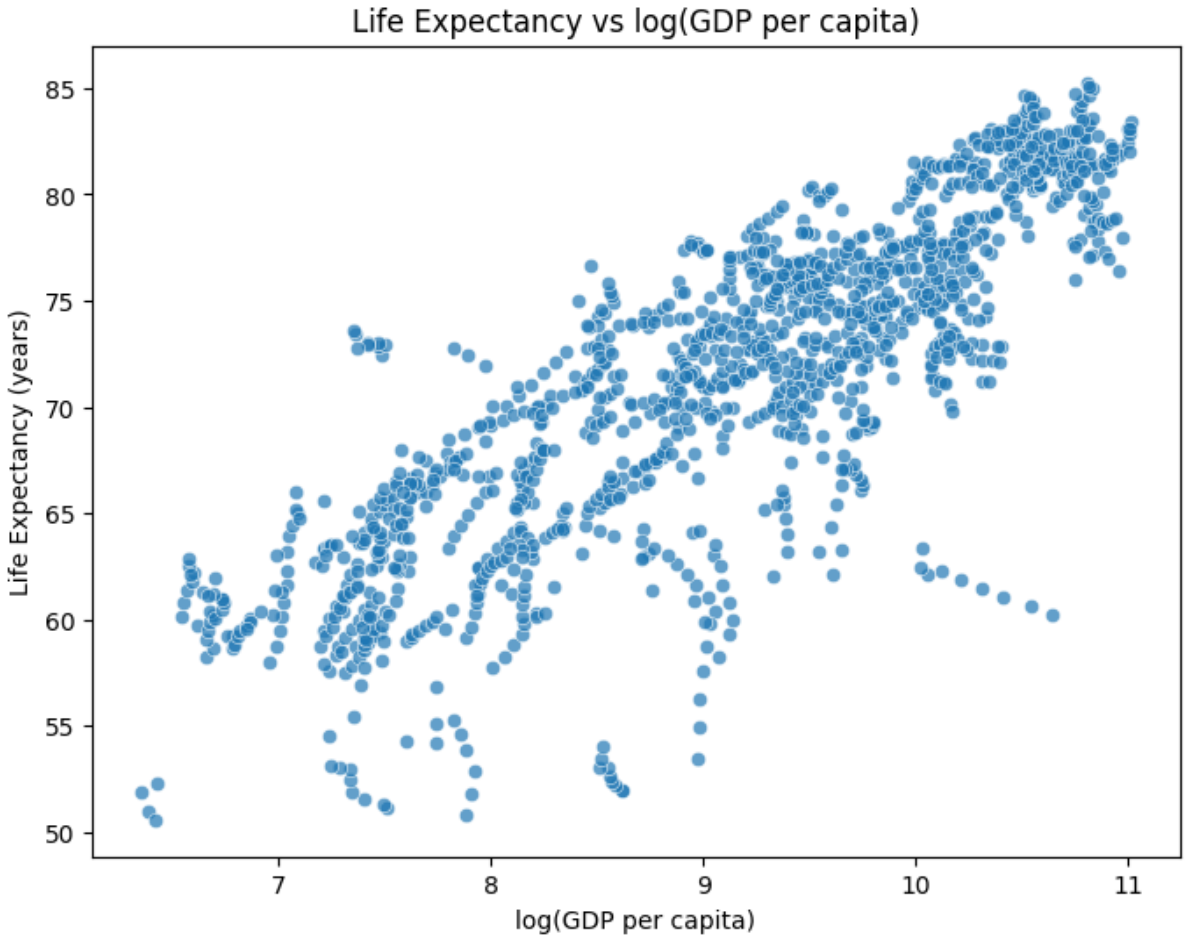


Model Building and Cross-Validation

A simple linear model would not fit well because the relationship is clearly nonlinear and asymptotic.

Although transformations like log(GDP) could approximate linearity, nonparametric smoothing methods (Bin, KNN, LOESS, Kernel) are better suited to flexibly model this gradual saturation pattern without assuming a specific functional form.

First, we log transformed the GDP values for linearity, then we plotted the data.



Train-Test Split :

The cleaned dataset contained **1,438 observations** after removing missing values and outliers.
It was randomly divided into:

- **Training Set (80%)**: 1,150 observations
- **Test Set (20%)**: 288 observations

The training data was used for model fitting and hyperparameter tuning, while the test data was reserved exclusively for evaluating final model performance.

Cross-Validation Setup :

A **5-fold cross-validation** approach was used on the training data for all smoothers.
This choice balances computational efficiency and statistical reliability:

- With 1,150 data points, 5 folds ensure sufficient data per fold for stable error estimates.
- Higher folds (like 10) would increase computation without substantial gain in bias-variance trade off.

Each smoother's key hyperparameter was tuned over a reasonable range, and **Mean Squared Error (MSE)** was used as the primary metric for model selection because Sensitive to large deviations, emphasizes major errors — ideal for model selection., with **Mean Absolute Error (MAE)** also tracked for interpretability.

Smoothing Methods and Hyperparameter Tuning :

1. Bin Smoother

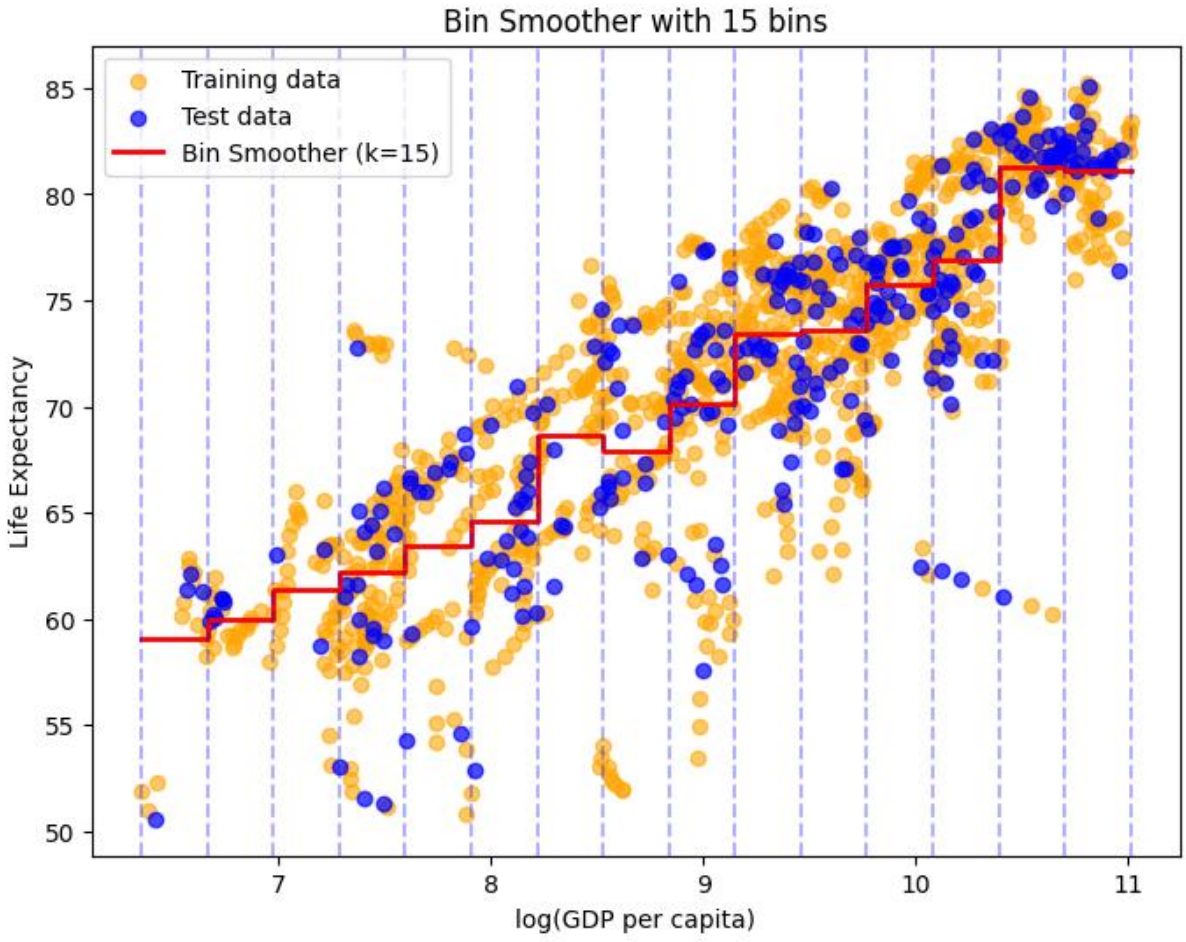
- **Hyperparameter tuned:** Number of bins (3 to 20)
- **Optimal Value:** 15 bins
- **Observation:**
  As the number of bins increased, CV MSE initially decreased — reflecting finer granularity — but beyond 15 bins, overfitting caused CV error to rise again.
- **Best CV MSE:** 21.41
- **Best CV MAE**: 3.07
- **Test MSE:** 16.36
- **Test MAE:** 2.89

Bins	CV_MSE	CV_MAE	Interpretation
3	28.999	3.986	Model too coarse (underfitting) — error is high due to few bins.
5	22.917	3.296	Improvement — finer granularity reduces error.
8	21.932	3.124	Further improvement — bins are better capturing variability.
10	22.215	3.210	Slight overfitting or noise; error increases again.

Bins	CV_MSE	CV_MAE	Interpretation
15	21.409	3.073	Lowest errors — best generalization among all. ✅
20	21.448	3.086	Nearly same as 15 — no real gain, possible over-smoothing.

The Bin Smoother captured broad trends effectively but lost precision at higher GDP values due to its stepwise nature.

Bin Smoother for the optimal number of bins :



---

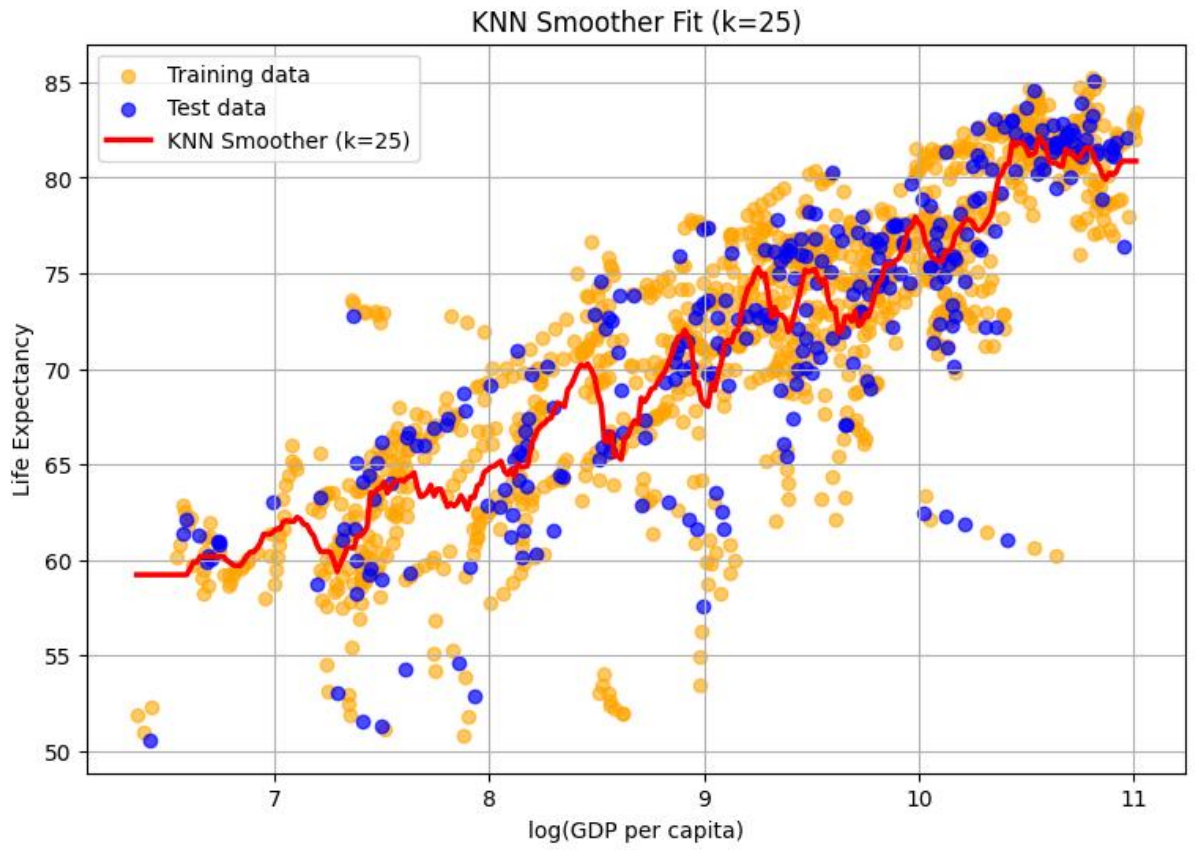
2. K-Nearest Neighbors (KNN) Smoother

- **Hyperparameter tuned:** Number of neighbors (*k* = 3 to 35)
- **Optimal Value:** *k* = 25
- **Observation:**
  CV error decreased as *k* increased from small values (reducing noise), but very large *k* led to oversmoothing and rising error.
- **Best CV MSE:** 16.26
- **Best CV MAE**: 2.93
- **Test MSE:** 16.88
- **Test MAE:** 2.96

k (Neighbors)	**CV_MSE**	**CV_MAE**	**Interpretation**
3	21.349	3.328	Very low (k): model fits noise → **high variance** (overfitting).
5	18.636	3.155	Smoother predictions → variance decreases, accuracy improves.
10	17.112	3.023	Balanced bias-variance trade-off — good performance.
15	16.835	2.998	Slightly better smoothing — less noise impact.

k (Neighbors)	CV_MSE	CV_MAE	Interpretation
20	16.483	2.955	Continues to improve — smoother generalization.
25	16.261	2.933	**Best overall — lowest MSE & MAE. ✅ Optimal k.**
30	16.377	2.940	Plateau — no significant gain; possibly mild bias increase.
35	16.335	2.936	Stable performance — model fully smoothed.

KNN Smoother for best value of k :



The KNN Smoother provided a good bias-variance trade off, capturing smooth transitions without excessive noise.

3. LOESS (Locally Estimated Scatterplot Smoothing)

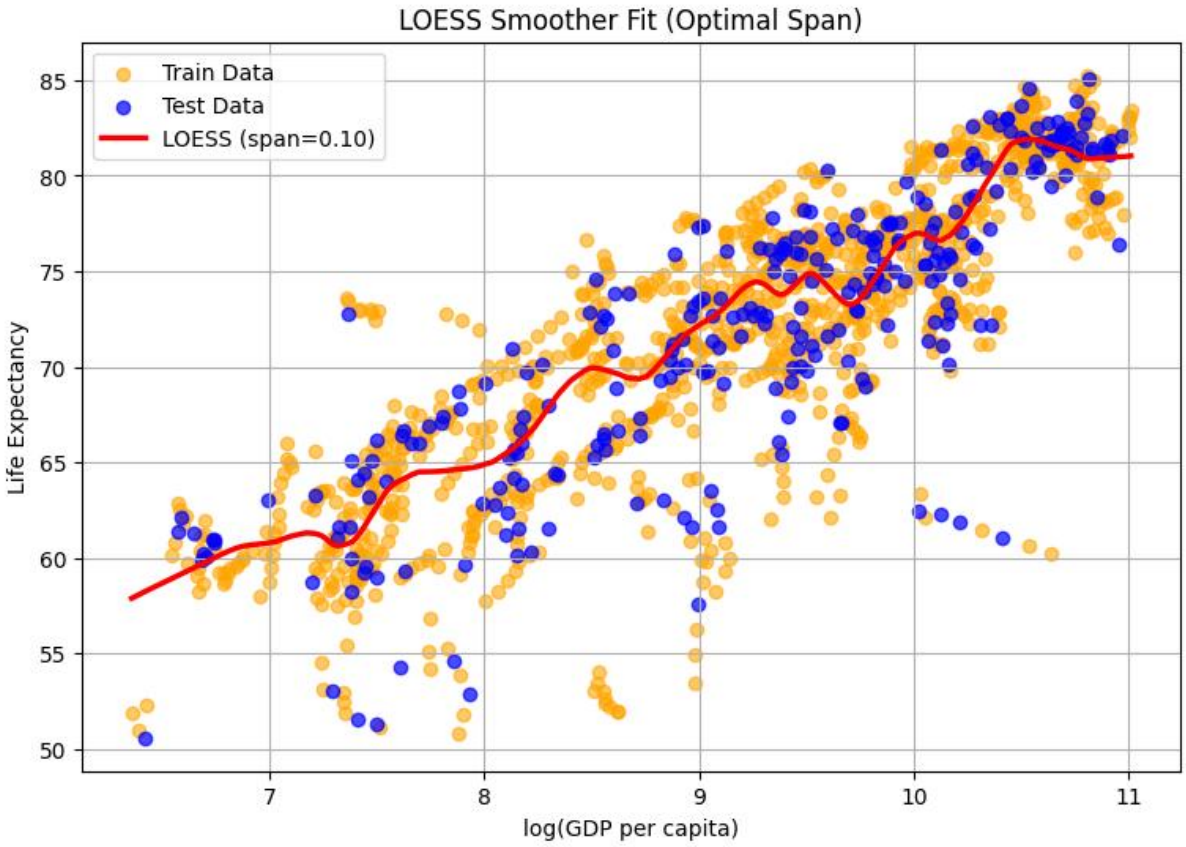
- **Hyperparameter tuned:** Span (0.05 to 0.95)
- **Optimal Value:** Span = 0.1
- **Observation:**
  Smaller spans gave lower bias but slightly higher variance. The best performance was achieved at a small span (0.1), balancing flexibility and generalization.
- **Best CV MSE:** 17.15
- **Best CV MAE:** 2.88
- **Test MSE:** 16.64
- **Test MAE:** 2.83

Span (frac)	CV_MSE	CV_MAE	Interpretation
0.05	17.2069	2.8793	Very small span — local fit captures fine details, some noise present (slightly high variance).
0.10	**17.1506**	**2.8780**	✅ **Best performance** — optimal bias-variance balance.
0.15–0.25	17.3–17.4	2.90–2.95	Gradual smoothing, small loss in local accuracy.

Span (frac)	CV_MSE	CV_MAE	Interpretation
0.30–0.60	17.48–17.70	2.96–3.02	Increasing span leads to **over-smoothing** (bias ↑).
0.65–0.95	17.71–17.72	3.03–3.04	Curve becomes too smooth — stable but less adaptive to local variation.

LOESS effectively modelled the nonlinear, saturating pattern between GDP and life expectancy, showing robust generalization on the test data.

LOESS Smoother for best fraction value :



4. Kernel Smoother

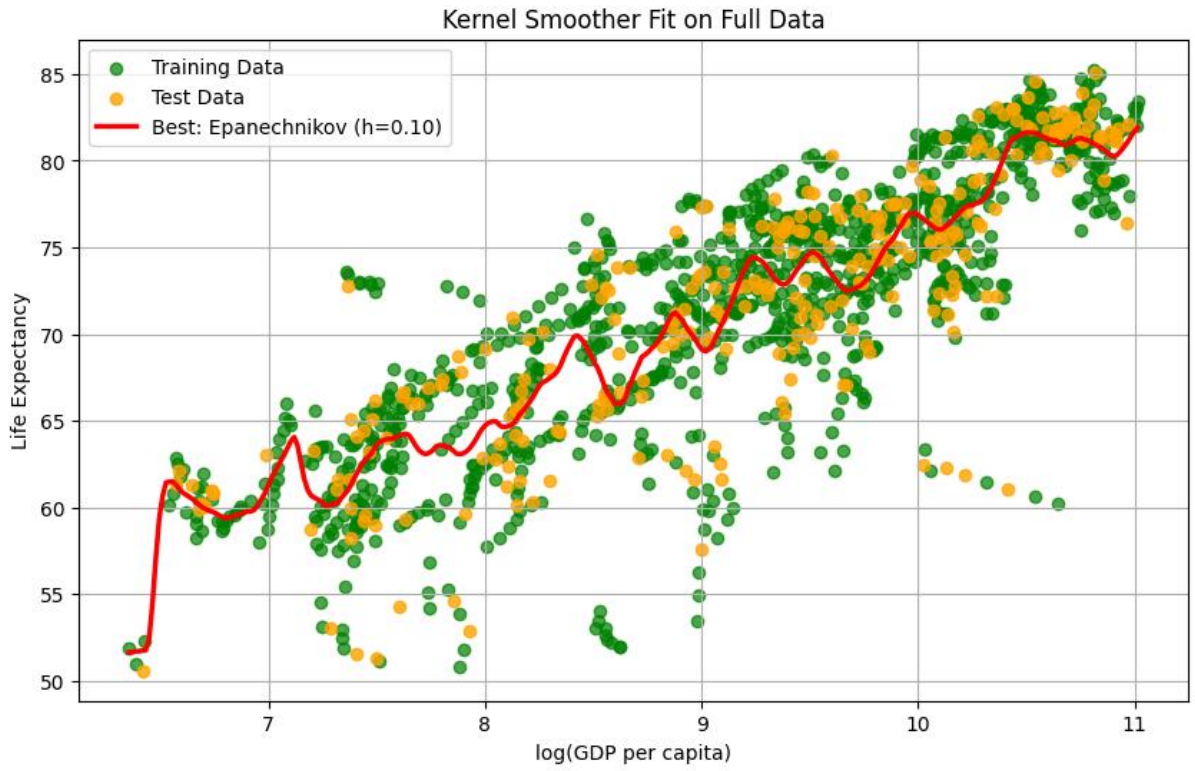
- **Hyperparameter tuned:** Kernel type (Uniform,Quartic, Epanechnikov,Gaussian,Triweight) and bandwidth (0.1–1.0)
- **Optimal Value:** Epanechnikov kernel with bandwidth **h = 0.10**
- **Observation:**
  Validation error followed a U-shape pattern — too small *h* caused high variance, and too large *h* oversmoothed the curve.
  The Epanechnikov kernel with moderate bandwidth achieved the best trade off.
- **Best CV MSE:** 15.94
- **Test MSE:** 15.98
- **Test MAE:** 2.85

Kernel	**Best Bandwidth (h)**	**Best CV_MSE**	**Interpretation**
Gaussian	0.1	16.3047	Error increases sharply as bandwidth grows → model highly sensitive to local structure; best with very local smoothing.
Epanechnikov	**0.1**	**15.9428**	✅ **Lowest overall MSE** — best performing kernel. Excellent bias–variance balance.

Kernel	Best Bandwidth (h)	Best CV_MSE	Interpretation
Uniform	0.1	16.0650	Performs well at small bandwidth but less efficient than Epanechnikov (discontinuous weights).
Triweight	0.1	15.9884	Very close to Epanechnikov — smooth weighting function gives stable performance.
Quartic	0.1	15.9506	Also competitive; slightly higher bias than Epanechnikov.

The Kernel Smoother produced the **lowest test error**, capturing the nonlinear and saturating nature of the relationship most accurately.

Kernel Smoother for the best kernel and best bandwidth :



Model Comparison and Discussion

Summary of Smoother Performance :

All four smoothing methods — Bin Smoother, KNN Smoother, LOESS, and Kernel Smoother — were trained on the cleaned GDP vs. Life Expectancy dataset using 5-fold cross-validation for hyperparameter tuning. Their best configurations and performance metrics are summarized below**.**

Method	**Best Hyperparameter**	**CV MSE**	**Test MSE**	**Test MAE**
Kernel Smoother	Epanechnikov, h = 0.10	**15.94**	**15.98**	**2.85**
Bin Smoother	15 bins	21.41	16.36	2.89
LOESS	Span = 0.1	17.15	16.64	**2.83**
KNN Smoother	k = 25	16.26	16.88	2.96

From the results:

- The **Kernel Smoother** achieved the **lowest validation and test MSE**, indicating the best overall predictive accuracy.
- The **LOESS smoother** had slightly higher test MSE but the lowest MAE, suggesting strong robustness and interpretability.

- The **Bin smoother** performed reasonably but was less flexible, and the **KNN smoother** tended to over smooth for larger *k*.

Visual Comparison of Fitted Curves :

When all fitted curves from the best-performing models were plotted together on the original scatterplot:

- The **Kernel Smoother (Epanechnikov, h=0.10)** produced a **smooth, saturating curve**, capturing the diminishing returns effect — life expectancy rises sharply with GDP at low-income levels, then levels off for higher GDPs.
- The **LOESS smoother** followed a similar shape but exhibited slightly more local variability.
- The **KNN smoother** yielded a flatter trend at high GDPs, missing subtle curvature.
- The **Bin smoother** created a stepped pattern — informative for trend direction but less continuous.

Overall, the kernel-based approach most faithfully represented the expected nonlinear and saturating pattern, while maintaining visual smoothness and low error.

Key Takeaways:

The **Kernel Smoother (Epanechnikov kernel, bandwidth = 0.10)** achieved the lowest **test MSE (15.98)**, confirming it as the best-performing model.

The Kernel Smoother produced a visually smooth, continuous curve without abrupt transitions.
However, it was slightly smoother than LOESS, which allowed more local flexibility.
This aligns with the expected behaviour: smaller bandwidths or spans capture more detail but risk overfitting, whereas moderate smoothing yields a more stable, visually appealing curve.

- **LOESS** could outperform the kernel smoother when the underlying relationship has **highly localized variations**, since LOESS adapts polynomial fits to local regions.
- **KNN smoothing** might work better in **noisy data** with unevenly spaced X-values, as it adjusts neighbourhood size adaptively.
- **Bin smoothing** could be preferable for **interpretability** or when explaining average group trends (e.g., in socioeconomic studies).

Thus, while the kernel method is optimal here, no single smoother universally dominates — performance depends on data structure and noise characteristics.

This study highlights that **hyperparameter tuning is crucial** for achieving balance between bias and variance in nonparametric regression:

- Undersmoothing (too small bandwidth or span) can cause noisy, wiggly fits.
- Oversmoothing (too large bandwidth or *k*) can hide meaningful patterns.
- Proper tuning using **cross-validation** ensures that the smoother generalizes well to unseen data.

The project thus emphasizes that nonparametric models are powerful for uncovering nonlinear relationships, but their success depends critically on systematic hyperparameter optimization.

Conclusion :

The analysis demonstrated that the **relationship between GDP per capita and life expectancy is nonlinear and saturating** — consistent with economic theory (the Preston curve).
Among the tested methods, the **Kernel Smoother** with Epanechnikov kernel and bandwidth 0.10 provided the **best fit**, lowest error, and most interpretable visual pattern.
Overall, smoothing techniques proved highly effective in capturing complex real-world trends without assuming a rigid functional form.