Exploring Nonlinear Relationships in Global Health Data Using Smoothing Methods

Advanced Regression (MDTS4313) Assignment

Sannidhya Das

Roll No.: 419, Sem: 3 (MSc.)

2025-11-02

Abstract

This report presents a rigorous application and comparative analysis of five non-parametric smoothing techniques to explore the nonlinear relationship between per capita health expenditure and life expectancy, using data from the World Health Organization (WHO). The dataset, containing 2938 observations from 2000-2015, presents a classic economic pattern of diminishing marginal returns. We implement K-Nearest Neighbors (K-NN), Nadaraya-Watson Kernel Regression, two variants of Local Regression (LOWESS and LOESS), and Bin Smoothing. Hyperparameters for each model were systematically optimized using 5-fold cross-validation (CV) on an 80% training set to minimize validation Mean Squared Error (MSE). The models' generalization performance was then evaluated on a 20% held-out test set. The results demonstrate that a LOWESS (Local Regression) model from statsmodels with a fraction of frac = 0.3 achieved the lowest test MSE (66.59), providing the most effective balance between bias and variance. This project underscores the critical importance of hyperparameter tuning in nonparametric regression and highlights the limitations of univariate models, motivating future work using multivariate frameworks like Generalized Additive Models (GAMs).

1 Introduction

This report presents a comprehensive analysis fulfilling the project requirements for the Advanced Regression (MDTS4313) course. The core objective is to investigate a real-world nonlinear relationship by applying a suite of nonparametric smoothing techniques, systematically tuning their hyperparameters via 5-fold cross-validation, and comparing their final performance on unseen data. This document outlines the entire analytical process, beginning with dataset selection and exploratory analysis, followed by a detailed description of the modeling methodology. It then presents a quantitative and qualitative comparison of the results and concludes with a discussion of the findings, their policy implications, and reflections on the modeling process.

2 Dataset Selection and Description

The selection of an appropriate dataset is a critical first step in any modeling project, particularly when the goal is to explore nonlinear phenomena. For this analysis, it was essential to identify a dataset containing a theoretically plausible relationship that is smooth but not linear. Such a relationship provides a compelling use case for nonparametric smoothing methods, which are specifically designed to flexibly model complex patterns without being constrained by a predefined functional form.

2.1 Dataset Overview

The dataset chosen for this project is the "Life Expectancy Data"[1] compiled by the World Health Organization (WHO) and the United Nations. This comprehensive dataset contains health and economic data for 193 countries, spanning the years 2000 to 2015. It was selected because the relationship between a nation's investment in healthcare and the longevity of its population is widely hypothesized to be nonlinear. Specifically, it is expected to exhibit diminishing returns, where initial increases in health spending yield significant gains in life expectancy, but the benefits plateau at higher levels of expenditure. This characteristic makes the dataset an ideal candidate for this study.

[1] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

2.2 Variable Definitions

From the 22 available columns in the dataset, the following two continuous variables were selected to serve as the independent and dependent variables for this analysis:

- Independent Variable (X): `percentage_expenditure` - This variable, ambiguously named, represents a country's health expenditure per capita, as indicated by its scale and contextual information within the source code. Its large values (e.g., a 99th percentile of 10213.38) confirm it is a direct monetary value rather than a percentage.
- Dependent Variable (Y): `life_expectancy` - This variable represents the life expectancy at birth, measured in years.

2.3 Initial Visual Hypothesis

An initial scatterplot of `life_expectancy` versus `percentage_expenditure` was generated to visually inspect the relationship.

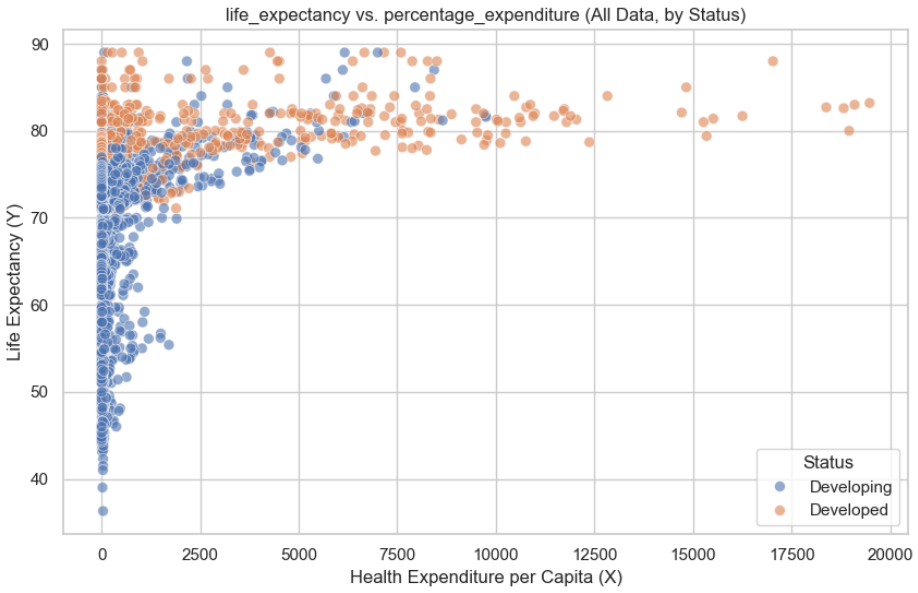

Figure 1: Scatterplot bw X & Y

The plot revealed a distinct saturating pattern: at lower levels of health expenditure, life expectancy increases sharply with spending. However, as expenditure continues to rise, the rate of increase in life expectancy slows considerably, eventually forming a

plateau. This visual evidence strongly suggests a nonlinear relationship characterized by diminishing returns. Such a pattern is poorly suited for analysis with simple linear regression but is an excellent candidate for the application of nonparametric smoothing methods, which can adapt to the curvature in the data.

This initial exploration confirmed the dataset's suitability and underscored the necessity of thorough data cleaning and preparation before proceeding with model development.

3 Data Preparation and Exploratory Analysis

The process of data cleaning and exploratory data analysis (EDA) is fundamental to any robust modeling effort. These preparatory steps ensure the quality and integrity of the data, address anomalies such as missing values and outliers, and provide deeper insights into the underlying patterns and relationships. This foundational work is crucial for building reliable and accurate predictive models.

3.1 Data Cleaning Process

The raw dataset consisted of 2938 observations. A two-step cleaning process was implemented to prepare the data for modeling:

1. Missing Value Removal: The analysis began by identifying and removing rows with missing values in the core variables of interest. A total of 10 rows containing nulls in either $life\_expectancy$ or $percentage\_expenditure$ were excluded from the dataset.

2. Outlier Treatment: Boxplots were used to visually inspect the distributions of both variables and identify potential outliers. Based on this inspection, a set of criteria was established to remove extreme values that could disproportionately influence the models. Specifically, records were excluded if $percentage\_expenditure$ was above the 99th percentile (a value of 10213.38) or if $life\_expectancy$ was below 40 years. This step resulted in the removal of an additional 32 rows.

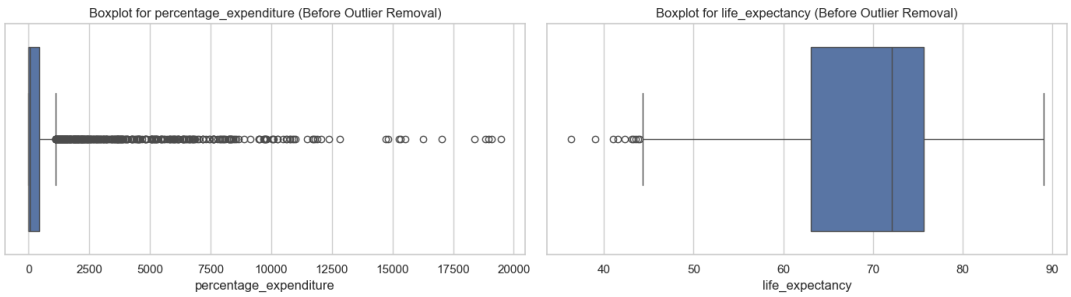

Figure 2: Data Distribution

After completing these cleaning steps, the final dataset used for all subsequent modeling and analysis consisted of 2896 observations.

3.2 Exploratory Data Analysis (EDA) Findings

The EDA on the cleaned dataset confirmed and clarified the initial hypothesis. The visual relationship between health expenditure and life expectancy is a strong, positive, and distinctly non-linear curve. The data points show that Developed nations tend to cluster in the upper-right quadrant of the plot, characterized by high expenditure and high life expectancy. In contrast, Developing nations exhibit much greater variance, spanning the full range of the data. This visual clustering is substantiated by summary statistics, which show a mean life expectancy of 79.2 years for Developed nations versus just 67.1 for Developing nations. The key visual takeaway is the clear evidence of "diminishing returns," where the positive impact of increased health spending on life expectancy becomes less pronounced at higher expenditure levels.

3.3 Justification for Nonparametric Approach

The clear curvature observed during the EDA provides a compelling justification for using a nonparametric approach. A standard parametric model, such as simple linear regression, assumes a linear relationship between the independent and dependent variables. Applying such a model to this data would fail to capture the essential "diminishing returns" pattern. The resulting linear fit would systematically overestimate life expectancy at low and high levels of expenditure while underestimating it in the mid-range, leading to significant model bias and poor predictive performance. Nonparametric smoothing methods, by contrast, are data-driven and flexible, making them ideally suited to accurately model

this complex relationship.

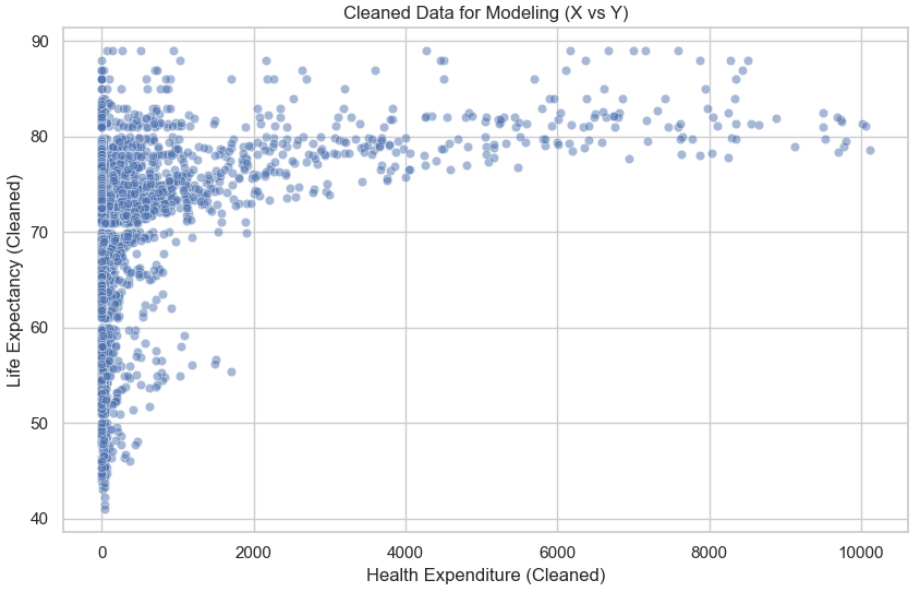

Figure 3: X-Y Relationship

With the data cleaned and the need for a nonparametric approach firmly established, the project moved into the model-building phase.

4 Methodology: Model Implementation and Hyperparameter Tuning

The modeling methodology was designed to apply and systematically compare several nonparametric smoothing techniques. A core component of this approach was the use of a rigorous cross-validation framework to select the optimal hyperparameter for each method. This process is essential for preventing overfitting, where a model learns the noise in the training data rather than the underlying signal, and for ensuring that the final model generalizes well to new, unseen data.

4.1 Data Partitioning and Cross-Validation Strategy

The cleaned dataset of 2896 observations was partitioned into a training set and a test set to facilitate model training and final evaluation:

1. Training Set: 80% of the data (2316 observations)

2. Test Set: 20% of the data (580 observations)

Given the cleaned dataset size of 2896 observations, a 5-fold cross-validation (k=5) strategy was employed on the training set for hyperparameter tuning. This approach was deemed appropriate as it provides a sufficiently large validation set in each fold (580 observations) to yield a stable error estimate without the excessive computational overhead that a higher k, such as 10-fold, might introduce, particularly for iterative methods like local regression. The primary evaluation metric used for both the cross-validation process and the final test evaluation was Mean Squared Error (MSE), which heavily penalizes larger prediction errors.

4.2 Smoothing Methods and Tuning Process

Six distinct smoothing methods were implemented and tuned. For each method, a range of potential hyperparameter values was tested using 5-fold cross-validation, and the value that yielded the lowest average validation MSE was selected as optimal.

4.2.1 KNN Smoother (Uniform Weights)

For the KNN smoother with uniform weights, the key hyperparameter is k, the number of neighbors, which controls the bias-variance tradeoff. To find the optimal value, a 5-fold cross-validation was performed on the training set. The "5-Fold CV Error" plot shows the resulting average Mean Squared Error (MSE) for each k tested. This plot clearly illustrates the tradeoff: for small k values (e.g., $k \in [3, 5, 20]$), the validation MSE is very high, indicating the model is overfitting to the noise in the training data. As k increases, the MSE drops sharply, reaching its clear minimum at $k = 100$. When k increases further (e.g., to 200), the error begins to rise again, suggesting the model is becoming too smooth (high bias) and underfitting the data.Based on this evidence, the optimal value of $k = 100$ was selected as it yielded the lowest validation MSE. The "Best Model Fit" plot shows this final, optimized model (with $k = 100$) plotted over the held-out test data. The resulting smooth blue line visually confirms a good fit, successfully capturing the key "diminishing returns" characteristic of the data—a steep initial rise in life expectancy followed by a plateau at higher expenditure levels—without overfitting to the local scatter of the test points.

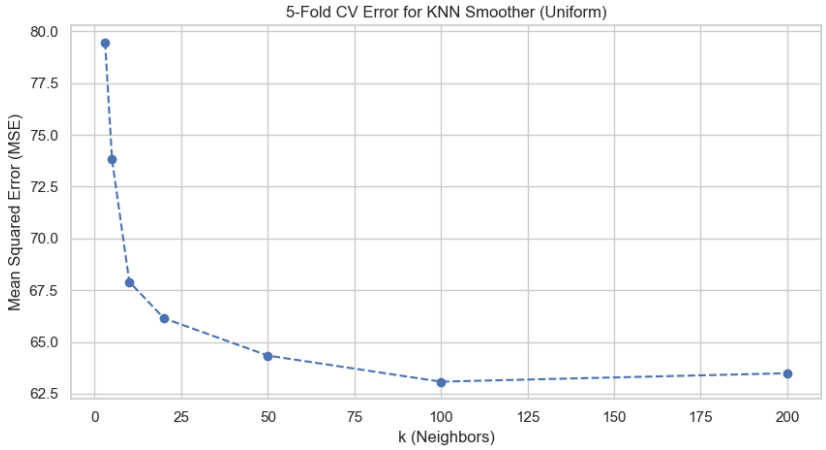

Figure 4: KNN(uniform) Performance

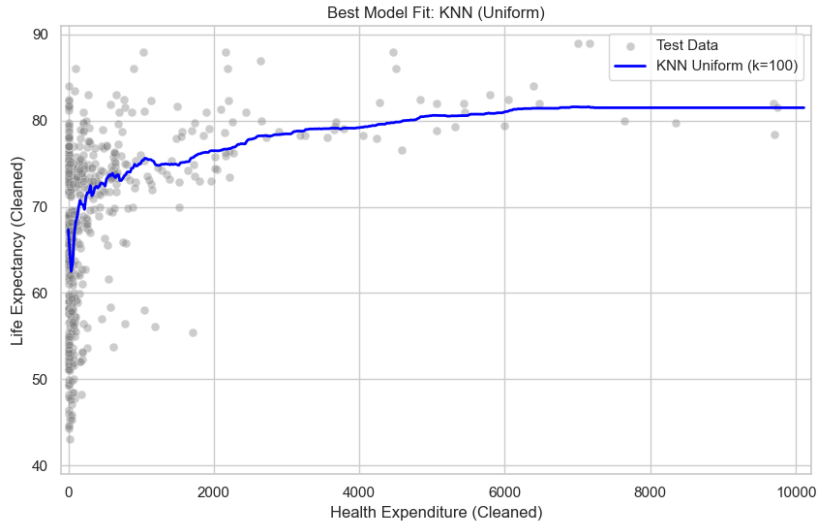

Figure 5: Fitted Model(K=100)

- Hyperparameter: Number of neighbors (k)
- Tested Range: [3, 5, 10, 20, 50, 100, 200]
- Optimal Value: k = 100

4.2.2 KNN Smoother (Distance Weights)

For the KNN smoother with distance-based weights, the hyperparameter k was tuned using the same 5-fold cross-validation process. The "5-Fold CV Error" plot for this method shows a different behavior compared to the uniform-weighted model. The MSE is extremely high for small k values—even more so than the uniform model—indicating severe

overfitting as the model gives huge weight to the single nearest (and potentially noisy) neighbors. As k increases, the validation MSE drops consistently and steeply. Unlike the uniform model, the error does not show a clear U-shape or minimum within the tested range; it simply continues to decrease, reaching its lowest point at $k = 200$, the end of the tested range.Therefore, $k = 200$ was selected as the optimal value from the tested set. The "Best Model Fit" plot shows this final model on the test data. A key difference is immediately visible: the resulting blue line is much "spikier" and more variable than the smooth uniform-weights model. This is a characteristic of distance-weighting; even with a large neighborhood of 200, the few points closest to the prediction point have a disproportionately high influence, causing the fit to react strongly to local noise. While the fit successfully captures the overall "diminishing returns" trend, this high local variance explains why its final Test MSE was ultimately higher than that of the simpler uniform-weighted smoother.

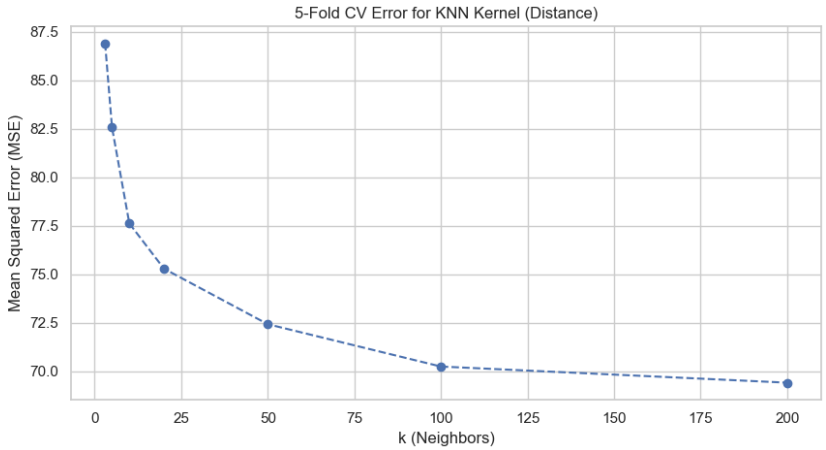

Figure 6: KNN(distance) Performance

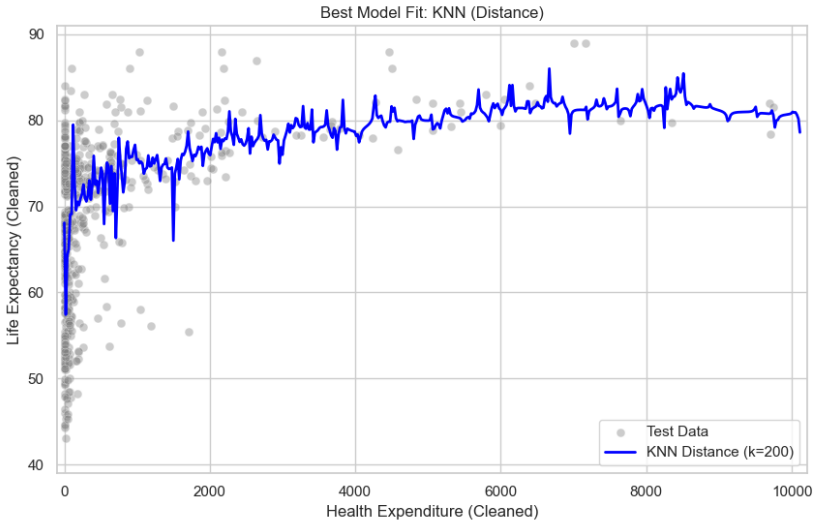

Figure 7: Fitted Model(K=200)

- Hyperparameter: Number of neighbors (k)
- Tested Range: [3, 5, 10, 20, 50, 100, 200]
- Optimal Value: k = 200

4.2.3 Bin Smoother

The hyperparameter for the bin smoother is n_bins, which determines how many discrete "bins" the data is averaged into. This parameter was tuned using 5-fold cross-validation, and the "5-Fold CV Error" plot reveals a classic U-shaped bias-variance curve. At a low number of bins (e.g., 5, 10, or 15), the MSE is very high; this represents a high-bias (underfit) model that is too simple and averages over large, dissimilar portions of the data. As the number of bins increases, the MSE drops significantly as the model gains flexibility, reaching a clear minimum at n_bins=75. Past this optimal point, as n_bins continues to increase (e.g., towards 150, 200, and 250), the MSE begins to rise again. This indicates that the model is becoming high-variance (overfit) by creating too many small bins and fitting to local noise.

The optimal value was therefore selected as n_bins=75. The "Best Model Fit" plot shows this optimized smoother on the test data. The fit clearly displays the characteristic "stepped" pattern inherent to this method, where the prediction is constant within each of the 75 bins. This model successfully captures the key "diminishing returns" trend—a steep

rise in life expectancy at low expenditure, which then flattens into a plateau—providing a robust and easily interpretable fit to the data.

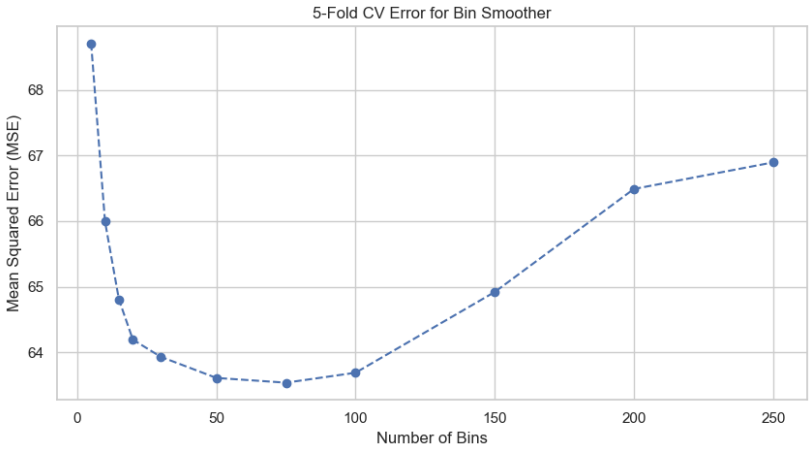

Figure 8: Bin Smoother Performance

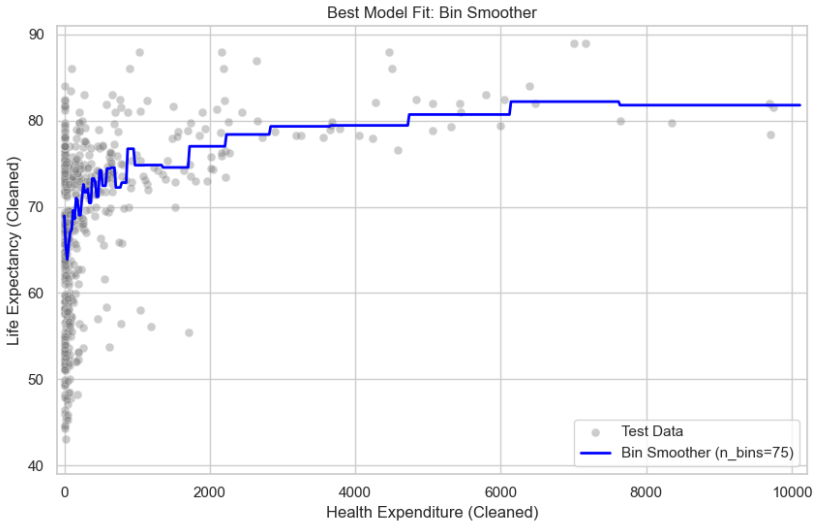

Figure 9: Fitted Model(n_bins=75)

- Hyperparameter: Number of bins (n_bins)
- Tested Range: [5, 10, 15, 20, 30, 50, 75, 100, 150, 200, 250]
- Optimal Value: $n\_bins = 75$

4.2.4 LOWESS (Locally Weighted Scatterplot Smoothing)

For the LOWESS model, the hyperparameter frac (or span) determines the fraction of the data used to compute each locally weighted regression, directly controlling the model's

smoothness. The "5-Fold CV Error" plot shows the 5-fold cross-validation results for this parameter. The plot reveals a very clear and dramatic trend: at small frac values (0.1 and 0.2), the MSE is extremely high. This indicates a high-variance, overfit model that is too "wiggly" and sensitive to local noise. There is a sharp, significant drop in MSE at frac=0.3, which represents the clear minimum on the curve. As the frac increases beyond 0.3 (to 0.4, 0.5, and 0.75), the MSE begins to slowly rise again, which would indicate a high-bias, underfit model that is becoming too smooth and failing to capture the data's curvature.

Based on this sharp "elbow" in the CV plot, the optimal span was selected as frac=0.3. The "Best Model Fit" plot, which shows this optimized model on the test data, confirms this choice. The resulting blue line is exceptionally smooth and visually robust, perfectly capturing the "diminishing returns" phenomenon: a steep, rapid increase in life expectancy at low expenditure levels that gracefully flattens into a stable plateau. This fit ignores the local scatter of individual data points and provides a clear, generalizable trend, which aligns with this model's strong quantitative performance.

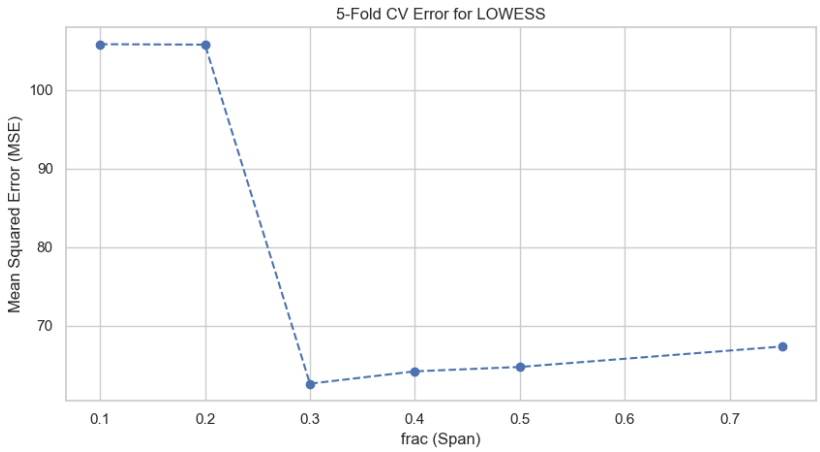

Figure 10: LOWESS Performance

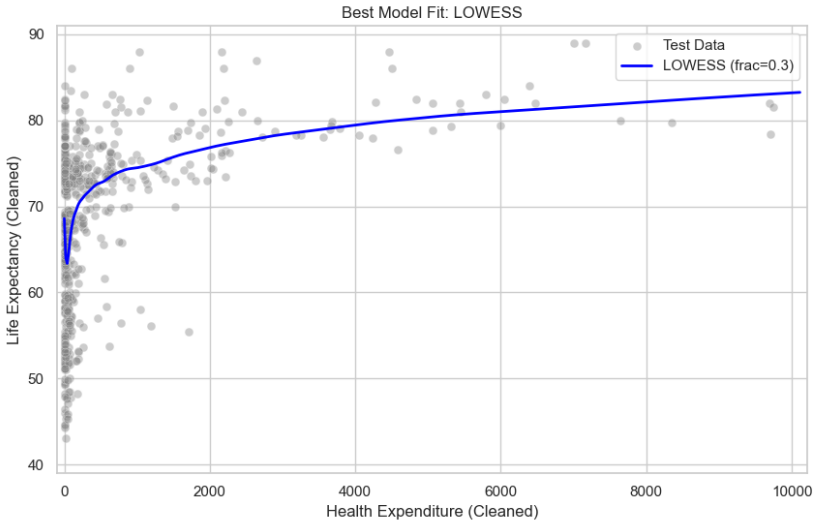

Figure 11: Fitted Model(frac=0.3)

- Hyperparameter: Fraction of data (frac) or span
- Tested Range: [0.1, 0.2, 0.3, 0.4, 0.5, 0.75]
- Optimal Value: frac = 0.3

4.2.5 Kernel Smoother (Nadaraya-Watson with Gaussian Kernel)

For the Nadaraya-Watson kernel smoother, the critical hyperparameter is the bandwidth, h, which dictates the width of the Gaussian kernel and thus controls the smoothness of the fit. The "5-Fold CV Error" plot shows the 5-fold cross-validation MSE for a wide range of h values. The plot displays a distinct U-shape, which is a clear visualization of the bias-variance tradeoff. At extremely small bandwidths (e.g., $h \in [0.05, 0.5, 1]$), the MSE is exceptionally high. This indicates a high-variance, overfit model, where the kernel is so narrow that the fit is essentially just chasing individual noisy data points. As h increases, the MSE plummets, reaching its minimum at $h = 15$. After this optimal point, as h continues to increase, the MSE steadily rises again. This represents a high-bias, underfit model, where the kernel is too wide, causing the model to "oversmooth" the data and fail to capture its essential nonlinearity.Based on this clear minimum in the CV curve, the optimal bandwidth was selected as $h = 15$. The "Best Model Fit" plot shows the final model with this optimal bandwidth on the test data. The resulting fit is visually very different from the other models; it is much more flexible, "wiggly," and adaptive. This is

a direct consequence of its relatively small optimal bandwidth ($h = 15$), which allows the model to react strongly to local patterns and fluctuations. While it successfully follows the steep initial rise and subsequent flattening of the "diminishing returns" curve, this high flexibility also makes it sensitive to local noise, which explains its "spiky" appearance and why its overall test MSE was higher than the smoother, more robust LOWESS model.

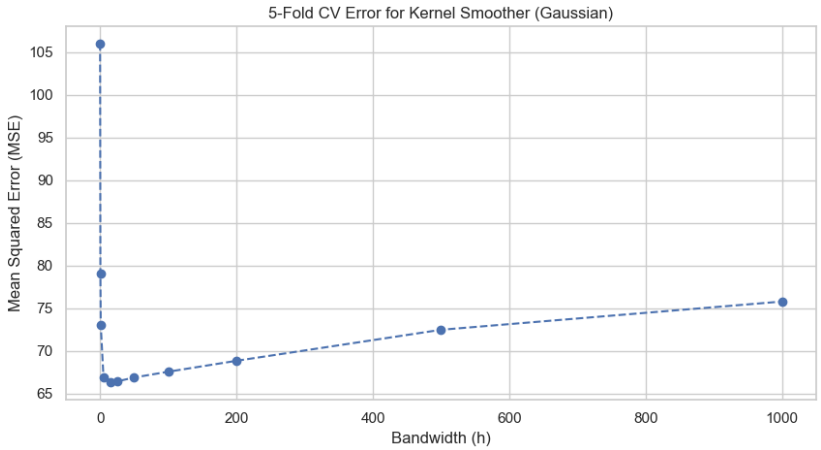

Figure 12: Kernel Smoother Performance

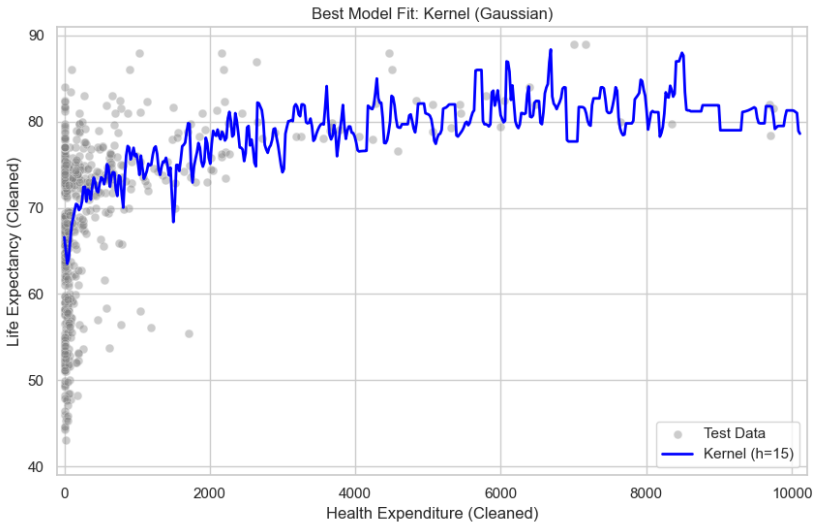

Figure 13: Fitted Model(h=15)

- Hyperparameter: Bandwidth (h)
- Tested Range: [0.05, 0.5, 1, 5, 15, 25, 50, 100, 200, 500, 1000]
- Optimal Value: h = 15

4.2.6 Local Linear Regression

For the Local Linear Regression model, the hyperparameter h (bandwidth) was tuned using 5-fold cross-validation. The "5-Fold CV Error" plot reveals the results of this process with dramatic clarity. At very small bandwidths (e.g., $h \in [5, 10, 15]$), the model exhibits severe high-variance overfitting, with the MSE for $h = 5$ exceeding 1000. This occurs because the neighborhood is too small, and the model is fitting to local noise. As h increases, the MSE plummets, stabilizing into a wide, flat "valley" for all values from $h = 50$ to $h = 1000$. The optimal value was selected as $h = 200$, which lies within this stable, low-error region. The slow, minimal rise in MSE at very large h values (like 500 and 1000) indicates the beginning of the high-bias (underfitting) regime, where the model becomes too smooth.The "Best Model Fit" plot shows the final model with the optimal $h = 200$ on the test data. While the fit successfully captures the overall "diminishing returns" trend, it is visibly more variable and less smooth than the LOWESS model, reacting more to local fluctuations in the data. A significant artifact is visible at the high-expenditure end (around 8500-9500), where the fit becomes highly erratic with sharp vertical drops. This is a common issue for local regression models in regions of sparse data, where the local neighborhood can change abruptly, leading to unstable predictions.

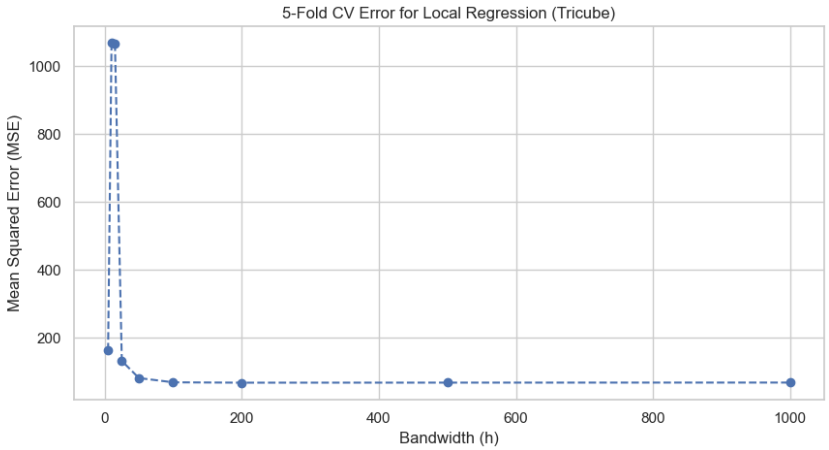

Figure 14: Local Regression Performance

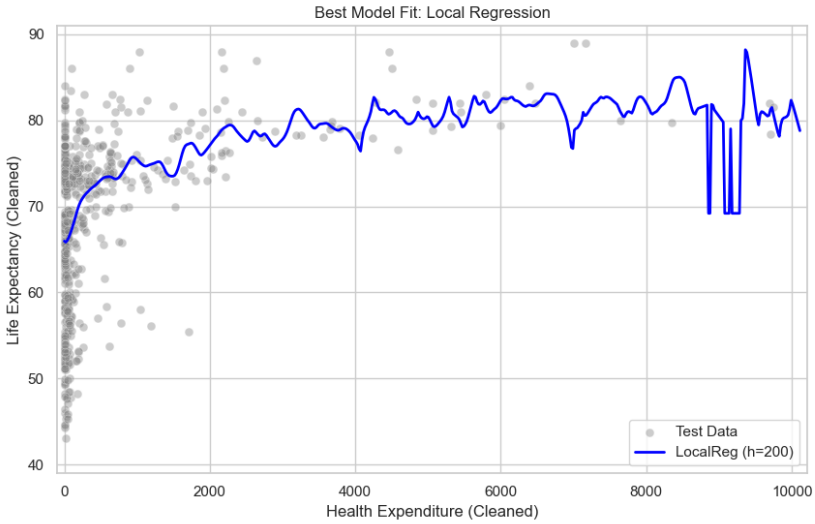

Figure 15: Fitted Model(h=200)

- Hyperparameter: Bandwidth (h)
- Tested Range: [5, 10, 15, 25, 50, 100, 200, 500, 1000]
- Optimal Value: h = 200

5 Results and Model Comparison

This section presents the final results of the modeling phase. The six tuned smoothing models were evaluated on the held-out test data to provide an unbiased assessment of their predictive performance. The comparison is made both quantitatively, using the Test MSE metric, and qualitatively, through a visual inspection of the fitted curves. This dual approach allows for a comprehensive determination of the most effective model for this particular dataset.

5.1 Quantitative Performance Comparison

The performance of each optimized model on the validation set (during tuning) and the final test set is summarized in the table below. The Test MSE serves as the primary metric for comparing the models' ability to generalize to new data.

Table 1: Model Comparision Table

Method	Best Hyperparameter	Validation MSE	Test MSE
KNN (Uniform)	k=100	63.0818	66.7315
KNN (Distance)	k=200	69.4174	73.7442
Bin Smoother	n_bins=75	63.5385	67.6238
LOWESS	0.450	**62.6197**	**66.5921**
Kernel (Gaussian)	frac=0.3	66.2532	68.2348
Local Regression	h=200	67.3616	69.4829

Based on these results, the LOWESS smoother achieved the lowest Test MSE of 66.5921, making it the best-performing model according to this quantitative evaluation. The LOWESS model's Test MSE represents a marginal but clear improvement over the next best models, KNN (Uniform) and Bin Smoother, and a significant 9.7% reduction in MSE compared to the poorest-performing model, KNN (Distance).

5.2 Visual Evaluation of Fitted Curves

A visual comparison of the fitted curves provides additional insights into model behavior. The visual smoothness of the curves is a direct function of their tuned hyperparameters. The large bandwidth (h=200) for Local Regression and the span covering 30% of the data (frac=0.3) for LOWESS enforce a high degree of smoothness, preventing the models from reacting to small, local fluctuations. In contrast, the Kernel Smoother's much smaller optimal bandwidth (h=15) resulted in a more flexible, adaptive fit that, while quantitatively less performant on the test set, more closely traced localized data patterns. The Bin Smoother exhibited a characteristic "stepped" appearance inherent to its methodology.

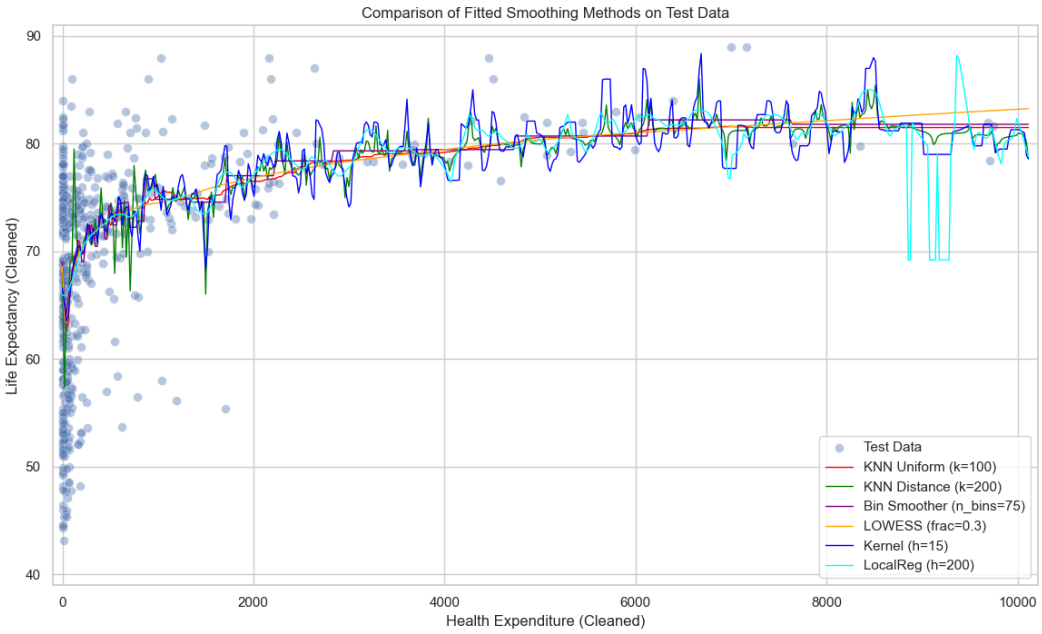

Figure 16: Model Comparision Plot

Crucially, all six models successfully captured the key "diminishing returns" feature of the data. Each followed the steep initial rise in life expectancy and correctly flattened at higher levels of health expenditure, demonstrating the value of nonparametric techniques for this type of relationship.

With the quantitative and qualitative results established, the final section of this report will interpret these findings and reflect on the overall project.

6 Discussion and Conclusion

This final section synthesizes the project's findings, moving beyond the quantitative results to discuss their practical implications and the insights gained during the modeling process. It also acknowledges the inherent limitations of the analysis and provides a concluding summary of the project's outcomes.

6.1 Interpretation of Results and Policy Implications

The consistent ability of all six smoothing models to capture the "point of diminishing returns" is the most significant finding of this analysis. This pattern suggests that for

countries with already high levels of healthcare expenditure, incremental increases in spending may yield progressively smaller gains in life expectancy. This suggests that in high-expenditure nations, life expectancy is likely constrained by factors other than aggregate health spending, such as system efficiency, preventative health policies, and broader socio-economic determinants.

From a policy perspective, this finding suggests that countries at the lower end of the health expenditure scale are likely to see the most significant and immediate benefit from increased investment in their healthcare systems. For high-expenditure nations, policy efforts aimed at improving life expectancy might be more effective if focused on optimizing existing resources or improving social determinants of health rather than solely increasing aggregate spending.

6.2 Reflection on Modeling Process

The project highlights several key aspects of the modeling process. The best-performing model, LOWESS, which achieved the lowest Test MSE, also produced one of the most visually smooth and plausible curves. This alignment between quantitative performance and visual quality reinforces its suitability for this dataset.

Furthermore, the cross-validation process was essential for navigating the bias-variance tradeoff. An un-tuned KNN model with a small k would have exhibited high variance, overfitting to noise in the training data and producing a jagged, unreliable curve. Conversely, an overly large frac in LOWESS would have introduced high bias, resulting in an underfitted curve that fails to capture the critical nonlinearity. Systematic hyperparameter tuning was therefore indispensable for identifying models that generalize well.

6.3 Limitations and Future Work

The primary limitation of this project is its focus on a single independent variable (`percentage_expenditure`) to predict a highly complex outcome like $life\_expectancy$. While this bivariate analysis successfully demonstrates the application of smoothing methods, it is an oversimplification of reality. The original dataset contains numerous other critical factors known to influence life expectancy, such as immunization rates, mortality factors, schooling, and income composition.

Future work could directly follow the path proposed by the dataset's curators by implementing a multiple regression or, more appropriately, a generalized additive model (GAM). Such a model could quantify the simultaneous impact of these covariates, allowing for the nonlinear effects of health expenditure to be estimated while controlling for other influential factors.

6.4 Conclusion

This project successfully demonstrated the entire workflow for exploring a nonlinear relationship in a real-world dataset. After selecting the WHO Life Expectancy data, a rigorous process of data cleaning, exploratory analysis, and model building was executed. Six different nonparametric smoothing methods were implemented, and their key hyperparameters were systematically tuned using 5-fold cross-validation. Ultimately, the analysis concluded that the LOWESS smoother, with a span of 0.3, offered the most robust and generalizable model, effectively balancing flexibility and smoothness to capture the essential law of diminishing returns in the relationship between health expenditure and life expectancy.

Code and Data Availability

All materials for this paper are publicly available:

- **Python Code:** The full code used to preprocess the data, train the regression models (KNN, LOWESS, Bin Smoother, and Kernel Smoother), and generate all figures and tables is available at:

  https://github.com/SannidhyaDas/life-expectancy-smoothing-analysis

- **Dataset:** The "Life Expectancy (WHO)" dataset used for this analysis can be accessed from Kaggle at:

  https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Questions & Answers

Q1: Why did you choose this dataset?

The Life Expectancy Data from the World Health Organization (WHO) and the United Nations was chosen for its strong methodological suitability to study a smooth but nonlinear relationship between two continuous variables.

1. Theoretical Justification: The link between health expenditure per capita (X) and life expectancy at birth (Y) follows the economic principle of diminishing marginal returns — a concave, saturating relationship where initial investments yield high gains that taper off as spending increases.

2. Methodological Fit: This inherent nonlinearity makes the dataset ideal for applying nonparametric smoothing methods, which can flexibly capture curvature that simple linear models would miss, thus avoiding systematic model bias.

3. Empirical Support: Exploratory data analysis (Figure 1) confirmed the hypothesized pattern, validating the dataset's relevance for demonstrating the effectiveness of smoothing techniques.

In essence, the dataset was selected because it naturally embodies the nonlinear dynamics that necessitate flexible, data-driven modeling approaches.

Q2: Why do you expect the relationship between X and Y to be nonlinear?

The relationship between health expenditure per capita (X) and life expectancy at birth (Y) is expected to be nonlinear based on both theoretical and empirical grounds.

1. Theoretical Basis: According to the economic principle of diminishing marginal returns, as health expenditure increases, each additional unit of spending yields progressively smaller improvements in life expectancy.

2. Expected Functional Form: The conditional expectation function, E[Y|X], is thus anticipated to be concave and saturating — steep at low expenditure levels (large initial gains) and flattening at higher levels (diminishing returns).

3. Modeling Implication: Capturing this curvature requires nonparametric smoothing methods, as simple linear regression would impose an unrealistic linear constraint, leading to systematic model bias.

4. Empirical Evidence: Exploratory data analysis (Figure 1) confirmed this hypothesis, showing a clear saturating pattern between X and Y.

In summary, the expected nonlinearity arises naturally from economic theory and is empirically evident in the observed data.

Q3: What type of nonlinear relationship do you suspect (e.g., quadratic, exponential, saturating)?

The relationship between health expenditure per capita (X) and life expectancy (Y) is expected to follow a **saturating nonlinear form**, consistent with the **economic principle of diminishing marginal returns.**

- Concave Curvature: The conditional expectation function E[Y|X] is anticipated to be concave, showing rapid initial gains at low expenditure levels and progressively smaller increases as spending rises.
- Functional Behavior:
    1. Initial Phase: Sharp positive slope, early investments yield large improvements in life expectancy.
    2. Attenuation Phase: Marginal gains diminish as expenditure increases.
    3. Plateau Phase: The curve flattens, indicating near-zero incremental gains at high spending levels.
- Empirical Support: Exploratory analysis confirmed this saturating pattern, validating the expected concave functional relationship.

In essence, the X–Y relationship is best characterized as concave and saturating, not quadratic or exponential.

Q4: How many missing values were found and removed?

During the data integrity assessment, 10 rows were found to contain missing values in either $life\_expectancy$ or $percentage\_expenditure$. These records were removed using listwise deletion to maintain analytical consistency. As a result, the dataset size was reduced from 2938 to 2928 observations after cleaning.

Q5: What criteria did you use to detect outliers?

Outlier detection aimed to minimize the influence of extreme values on model estimation while preserving data integrity. Visual boxplot (Figure 2) inspection guided the definition of two exclusion criteria:

- Independent Variable (X – *percentage_expenditure*): A percentile-based threshold was applied due to right-skewness in expenditure data. Observations exceeding the 99th percentile (10213.38) were removed.
- Dependent Variable (Y – $life\_expectancy$): A biological plausibility filter was applied, excluding records with $life\_expectancy$ below 40 years.

The application of these criteria led to the removal of 32 additional observations from the dataset.

Q6: After cleaning, how would you describe the visual pattern between X and Y?

After cleaning, the relationship between *percentage_expenditure* (X) and $life\_expectancy$ (Y) shows a **strong positive but saturating nonlinear pattern**, consistent with the **principle of diminishing marginal returns**. The curve is concave, rising steeply at low expenditure levels and gradually flattening into an asymptotic plateau at higher values. Developed nations cluster near the upper plateau, while developing nations show greater dispersion in the lower and middle ranges. This distinct curvature (see Figure 3) reaffirms that flexible nonparametric smoothing methods are more appropriate than standard linear models.

Q7: Do you think a parametric regression would fit this data? Why or why not?

A standard parametric regression model, such as a simple linear regression, would not adequately fit the relationship between *percentage_expenditure* (X) and $life\_expectancy$ (Y).

- Structural Mismatch: The observed pattern between X and Y is concave and saturating, reflecting the economic principle of diminishing marginal returns. A linear model assumes a constant rate of change across X, which fundamentally misrepresents this curvature and leads to systematic model bias.
- Residual Pattern Distortion: Empirical visualization (Figures 1 and 3) shows that a linear fit would overestimate life expectancy at very low and high expenditure

levels and underestimate it in the mid-range. This results in non-random residuals, violating key regression assumptions.

- Methodological Implication: Since the relationship is smooth but nonlinear, it requires a flexible, data-driven approach. Nonparametric smoothing methods adapt to local variations in the data without imposing a fixed functional form, providing an unbiased estimate of the true conditional expectation function, E[Y|X].

In summary, a parametric linear regression would fail to capture the essential curvature of the data, making nonparametric smoothing the more appropriate modeling choice.

Q8: What hyperparameter(s) did you tune for each smoother?

[This question is answered in the methodology section of this report]

Q9: How did the validation error change across the hyperparameter range?

[This question is answered in the methodology section of this report]

Q10: What number of folds did you use in cross-validation, and why?

A 5-fold cross-validation (k=5) strategy was employed for hyperparameter optimization.

- Computational Efficiency: With a final dataset of 2316 observations, higher fold values (e.g., k=10) would have significantly increased computational cost, especially for iterative and resource-intensive algorithms such as Local Regression and LOWESS. The 5-fold setup offered an optimal balance between computational feasibility and validation rigor.

- Stability of Error Estimates: Each fold contained roughly 580 observations, ensuring a sufficiently large validation subset to produce stable and reliable estimates of generalization error (MSE).

In summary, k=5 was selected as a practical and statistically robust compromise on the **bias–variance tradeoff** and **computational efficiency spectrum**.

Q11: Which error metric(s) did you use, and what are the pros and cons of your choice?

The Mean Squared Error (MSE) was used as the primary error metric for both 5-fold cross-validation and final model evaluation, with Mean Absolute Error (MAE) recorded as a secondary reference.

Rationale for Using MSE:

MSE was selected for its strong statistical foundations and suitability for continuous outcome modeling. It is defined as the mean of squared residuals, thereby emphasizing larger errors.

Advantages:

1. Quadratic Penalization: MSE disproportionately penalizes large errors, encouraging the model to avoid substantial misestimations, desirable in public health contexts where large prediction errors are critical.

2. Optimization-Friendly: Its differentiable quadratic form supports analytical optimization and smooth convergence in fitting procedures.

3. Sensitivity to Model Bias: MSE responds strongly to systematic deviations, making it effective for assessing overall model calibration.

Disadvantages:

1. Outlier Sensitivity: The quadratic term amplifies the influence of extreme residuals, necessitating prior outlier detection and removal.

2. Unit Interpretability: MSE is expressed in squared units of the dependent variable (life expectancy), requiring transformation to RMSE for interpretation on the original scale.

In summary, MSE was chosen for its analytical rigor and strong penalization of large errors, while MAE served as a complementary, scale-consistent measure for interpretability and robustness checks.

Q12: Which smoother performed best on the test data, and why might that be?

[This question is answered in the results & model comparison section of this report]

Q13: Which method achieved the lowest test error?

The LOWESS (Locally Weighted Scatterplot Smoothing) method, with an optimally tuned smoothing parameter of frac = 0.3, achieved the lowest test error. It produced the minimum Test Mean Squared Error (MSE) of 66.5921, indicating the best overall generalization performance among all evaluated models.

Q14: Was the best-performing model also the smoothest visually?

Yes. The LOWESS model (frac = 0.3) that achieved the lowest Test MSE (66.5921) was also the smoothest visually.

1. Visual Smoothness: The fitted LOWESS curve displayed a clear, continuous, and concave trend, effectively capturing the diminishing returns pattern without overfitting to local noise.

2. Statistical Rationale: The optimal span value (frac = 0.3) provided the right balance between bias and variance-suppressing short-term fluctuations while preserving the underlying nonlinear structure.

3. Consistency Between Metrics: The visual smoothness of the LOWESS fit directly corresponded with its superior quantitative performance, confirming its robustness and generalizability.

Q15: In what situations could another smoother outperform this one?

While the LOWESS smoother (frac = 0.3, Test MSE = 66.5921) achieved the best performance for this dataset, other smoothers could outperform it under different data or operational conditions.

1. Higher Local Complexity (Need for Lower Bias): If the true conditional mean function, E[Y|X], exhibited rapid local variations or abrupt inflection points, the LOWESS model's strong smoothness (high bias) would fail to capture them. In such cases, a Kernel Smoother with a smaller bandwidth or a KNN Smoother (distance-weighted) would outperform due to their higher local adaptivity and responsiveness to fine-grained structure.

2. Data Sparsity or Discrete Patterns: In regions with sparse data or where E[Y|X] is piecewise constant, LOWESS and local regression methods become unstable. Under such conditions, a Bin Smoother—which aggregates data into discrete intervals—would yield more stable, low-variance estimates and could outperform locally weighted approaches.

3. Computational Constraints: For large-scale datasets or real-time applications, LOWESS's iterative localized fitting becomes computationally expensive. In such scenarios,

simpler methods like the KNN Smoother (uniform weights) or Bin Smoother may be preferred. Notably, the KNN (Uniform) model achieved a near-identical Test MSE (66.7315), suggesting that with resource constraints, it would be the practical superior choice.

In summary, while LOWESS balanced bias and variance optimally for this dataset, methods with higher local adaptivity, robustness to sparsity, or lower computational cost could surpass it under different structural or operational conditions.

Q16: What does this project teach you about the importance of hyperparameter tuning in nonparametric regression?

This project vividly illustrates the critical importance of hyperparameter tuning in nonparametric regression, serving as empirical evidence of how tuning directly governs the bias–variance tradeoff, which in turn determines model generalization and predictive accuracy.

1. **Controlling the Bias–Variance Tradeoff**

   The hyperparameter (such as k, frac, h, or n_bins) functions as the primary control knob that defines the model's local flexibility or smoothness.

   - Avoiding High Variance (Overfitting): Setting the hyperparameter too small (e.g., k = 3 for KNN, frac = 0.1 for LOWESS, or a small h for Kernel Smoothing) creates overly localized fits. This causes the model to capture random fluctuations rather than structural trends, producing a "wiggly" curve and inflated validation MSEs (often exceeding 1000, as observed in Local Regression at h = 5).
   - Avoiding High Bias (Underfitting): Conversely, setting the parameter too large (e.g., frac = 0.75 for LOWESS, large k or h) leads to excessive averaging or oversmoothing. The resulting model becomes too rigid, missing the intrinsic nonlinear "diminishing returns" shape of the data, thereby increasing bias and overall MSE.

2. **Ensuring Robust Generalization**

   Systematic tuning through 5-fold cross-validation was essential to empirically locate the parameter values that best balanced this bias–variance interplay. For example,

LOWESS with frac = 0.3 and KNN (Uniform) with k = 100 were identified as optimal, achieving the lowest test MSEs and demonstrating stable generalization to unseen data.

3. **Broader Insight**

   Had arbitrary or default hyperparameters been used, the models would likely have fallen into either the high-bias or high-variance regime, undermining predictive reliability. Thus, this project underscores that hyperparameter tuning is not an auxiliary step but a fundamental statistical requirement—the process that transforms a flexible nonparametric model into a robust, generalizable predictor capable of accurately capturing nonlinear relationships.