Life Expectancy Analysis: A Data Science Approach

Subtitle: "Analyzing Factors Affecting Life Expectancy in Japan Using Data Science"
Date: January 2025
Name: Khalil Mosbah, Data Scientist

Executive Summary:

This project analyzes key factors affecting life expectancy in Japan. Using Linear regression, OLS regression and SHAP analysis we found that *education*, *healthcare access*, and *park availability* in a prefecture are the most significant predictors of life expectancy in our data set.

Data & Methodology:

- **Data:** Kaggle Data Set by **Gianina-Maria Petrașcu**

https://www.kaggle.com/datasets/gianinamariapetrascu/japan-life-expectancy

- **Methodology:**
  - **EDA Observations & Feature Engineeing:**

***Worst vs best life Expectancy:** Aomori 82.8 & Shiga 85.5 years
***Correlation Heatmap:**
Strong (+) : 'University_%' & 'Junior_col_%'
Moderate (-): 'Physician_100kP', 'Park_land_%' & 'Salary'
(-): 'Elementay_school_%' & 'Ambulances'
***VIF Analysis: Severe multicolinearity** 'University_%' & 'Salary'
***Normalization:** Applied to all numeric values to reduce skewness
*PCA: Combined 'Salary' & 'University_%' into a new feature called 'Socioeconomic_index' to avoid multicolinearity

  - **Model Evaluation & Feature Importance:**

***Linear Regression Model** trained on the features 'Physician_100kP', 'Junior_college_%', 'Socioeconomic_index', 'Ambulances_100kP', & Park_Land_% gave the best performance:

$$R^2 = 0.37 \ \& \ RMSE = 0.34$$

***SHAP Analysis:**
Key Predictors: Physician_100kP' & 'Junior_col_%'
Moderate Impact: 'Park_Land_%', 'Socioeconomic_index' & 'Ambulances_100kP'

  - **Model Experiments:**

***OLS Regression:** Education, healthcare access and Park access are statistically significant predictors of life expectancy while socioeconomic index and Ambulances_100kP were insignificant.

  - **Insights:**

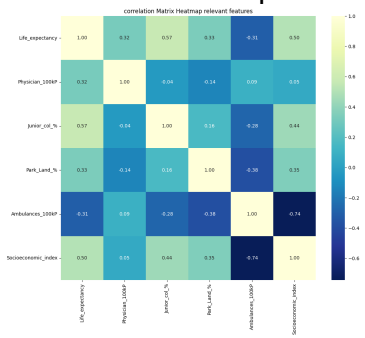
***Education:** 1% Jun_colege increase -> LifeE. increase 1.2y
***Healthcare access:** Additional physician/100kP ->LifeE. Increase 0.7y
***Access to Parks:** 1% increase park land-> LifeE. increase 0.56y

Recommendations:

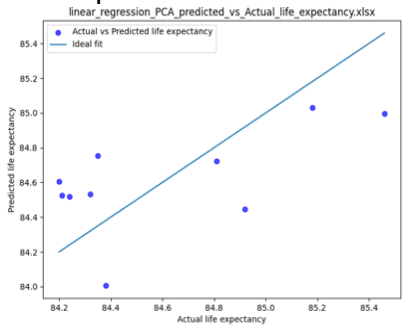
- Invest in education and healthcare**: These are the primary drivers of life expectancy.**
- Increase park availability**: More green spaces could significantly improve public health.**
- Policy implications**: Policymakers should focus on education, healthcare, and urban planning.**

Visual Elements:

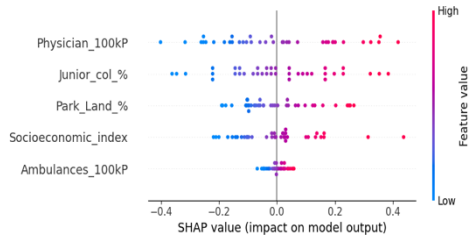
Correlation Heatmap



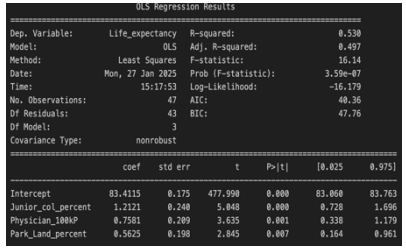
Model prediction:



SHAP Feature Visualization



OLS Regression Results



OLS Regression Results

Dep. Variable:	Life_expectancy	R-squared:	0.530
Model:	OLS	Adj. R-squared:	0.497
Method:	Least Squares	F-statistic:	16.14
Date:	Mon, 27 Jan 2025	Prob (F-statistic):	3.59e-07
Time:	15:17:53	Log-Likelihood:	-16.179
No. Observations:	47	AIC:	40.36
Df Residuals:	43	BIC:	47.76
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	83.4115	0.175	477.990	0.000	83.060	83.763
Junior_col_percent	1.2121	0.240	5.048	0.000	0.728	1.696
Physician_100kP	0.7581	0.209	3.635	0.001	0.338	1.179
Park_Land_percent	0.5625	0.198	2.845	0.007	0.164	0.961

LinkedIn: https://www.linkedin.com/in/khalil-mosbah-3174a41a1/
Email: khalil.mosbeh@yahoo.fr

Github project: https://github.com/khalil-hub/Life-expectancy-analysis-Japan