KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Live Long and Prosper

An Exploratory Analysis of Life Expectancy Indicators

Analyst: Trinity Johnson *(Expected Graduation 2026)*
Advisor: Nicole Carder

INTRODUCTION

Using data available through Kaggle, we decided to examine the most influential factors when trying to determine a countries average life expectancy. The data that was originally collected by the World Health Organization (WHO) and included 22 variables from categories including immunization factors, mortality factors, economic factors, and social factors. The full dataset includes 193 developed and developing countries across the globe and spanned from years 2000-2015.

After analyzing the full set, the data was reduced to focus on just the 5 most recent years, 2010 – 2015. After cleaning the data, the final subset contains complete observations for 171 countries (29 developed and 142 developing) and spans a five-year time period. Countries that were removed were a result of extreme missingness in the observational units.

Through this analysis, we hope to uncover the most important indicators of low life expectancy, to better understand how we can serve these countries and help them live long and fruitful lives.

METHODS

Data Understanding: First, we created a data dictionary for the life expectancy data to understand what the indicator variables were measuring and the appropriate ways to investigate them.

Diagnosing and Cleaning: When diagnosis the data, we noticed that many of the variables had obvious error that appear to be related to an extraction error. Value that ended in 0 were truncated, drastically underrepresenting the true value. To address these errors, in addition to true missing values, we imputed any missing values and errors with averages based on similar time periods within the specific country in questions. We also decided to reduce the years variable from 2000-2015 to 2010-2015 because there were many error from the earlier years and the years between 2010-2015 would be more relevant to the current situation.

Exploring Relationships: After cleaning the data, we created a correlation matrix to find the most influential variables for life expectancy to use in the study and analyze further. From the correlation matrix we decided that Income composition of resources, Schooling, HIV/AIDS, and BMI were the variables that appeared to be the most deterministic of life expectancy, in addition to some moderately correlated features.

Analysis and Modeling: After determine the best indicators, we created a categorical representation of Life Expectancy to better understand trends in the quantitative factors, by comparing median measure across the categories. Furthermore, we aimed to build a parsimonious model to try to predict a countries life expectancy based on the identified factors.

RESULTS

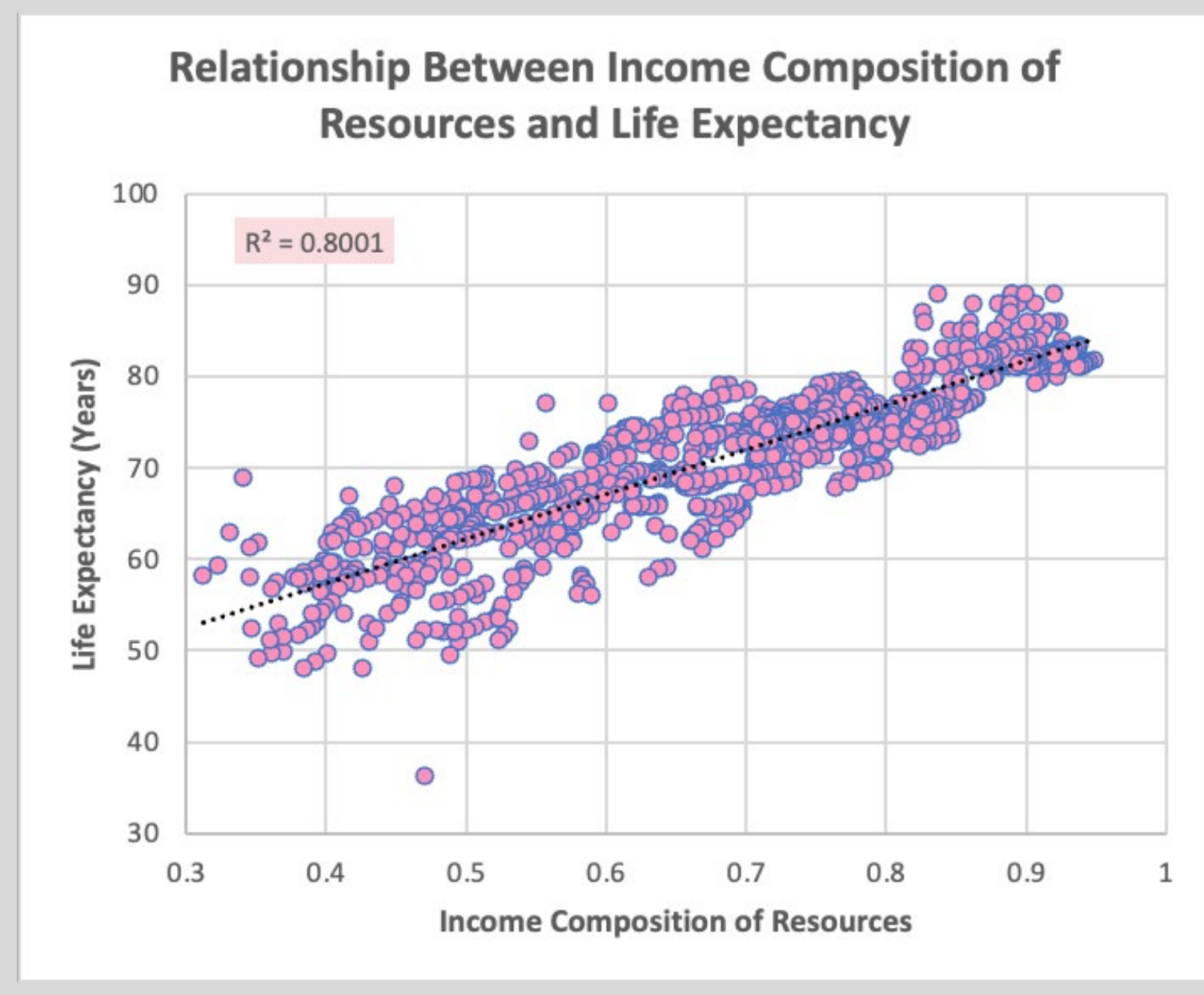
Focused Variable Set

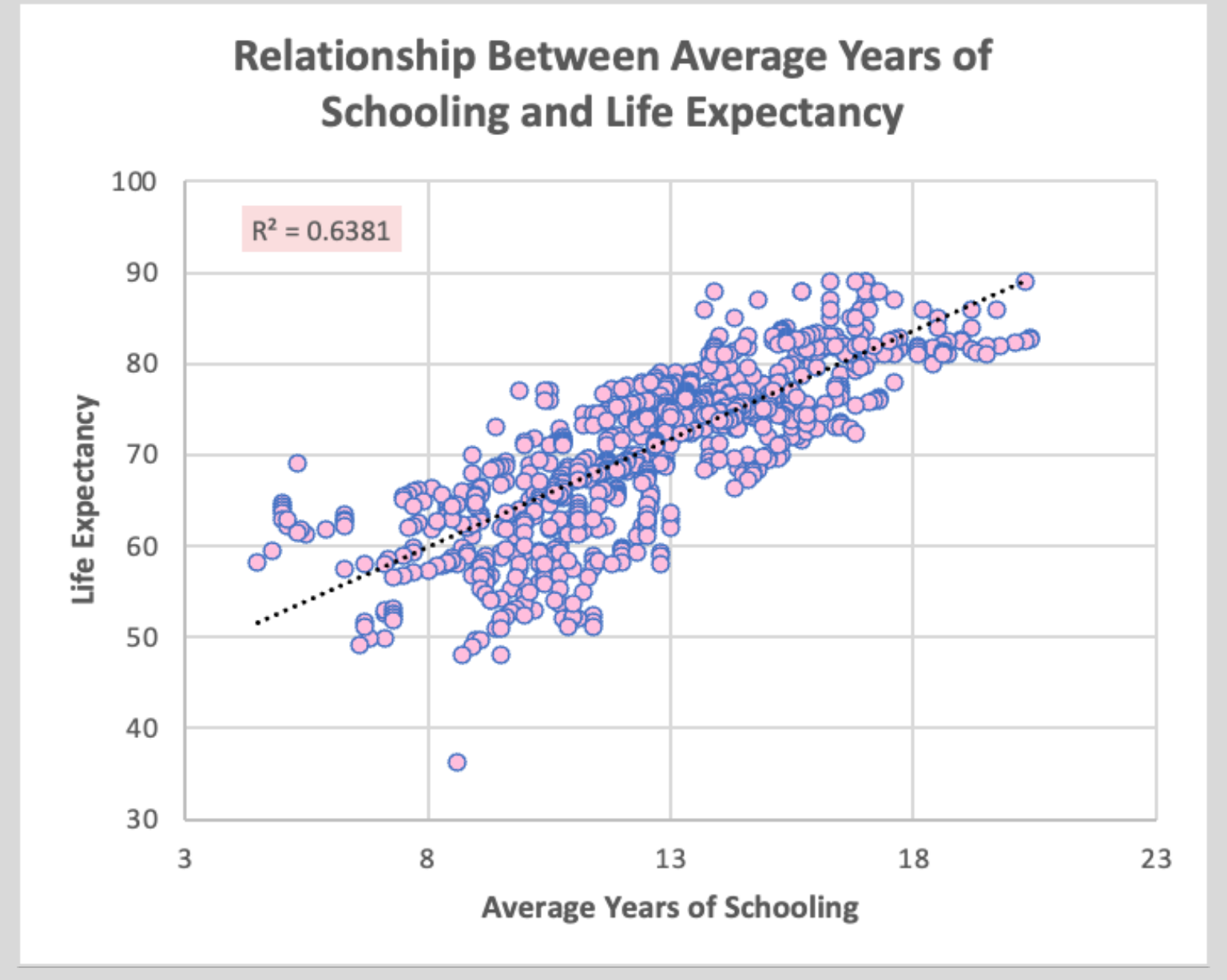
Variable Name	Variable Description
Adult Mortality	*Probability of dying between 15 and 60 years*
Income Composition of Resources	*Human development index in terms of income composition of resources (index ranging from 0 to 1)*
Schooling	*Number of years of schooling*
HIV/AIDS	*Deaths per 1000 live births HIV/AIDS (0-4 years)*
BMI	*Average body mass index of entire population (average female + average male)*
Under 5 Deaths	*Number of deaths under 5 years old per 1000 population*
Polio	*Polio (Pol3) immunization coverage among 1-year-olds (%)*
Infant Deaths	*Number of infant deaths per 1000 population*
Diphtheria	*Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)*
Thinness 5-9 years	*Prevalence of thinness among children for Age 5 to 9(%)*
Thinness 10-19 years	*Prevalence of thinness among children and adolescents for Age 10 to 19 (%)*
GDP	*Gross domestic product per capita (in USD)*
Life expectancy	*Life Expectancy in age*

Bivariate Explorations

Ranked Correlations with Life Expectancy

Rank	Indicator Name	Correlation
0	*Adult Mortality (PROXY)*	-0.9146
1	*Income Composition of Resources*	0.8732
2	*Schooling*	0.7988
3	*HIV/AIDS*	-0.7736
4	*BMI*	0.6985
5	*Under 5 Deaths (Transformed)*	-0.5759
6	*Polio (Transformed)*	0.5538
7	*Infant Deaths (Transformed)*	-0.5493
8	*Diptheria (Transformed)*	0.5461
9	*Thinness 5-9 (Transformed)*	-0.5406
10	*Thinness 10-19 (Transformed)*	-0.5372
11	*GDP*	0.4594



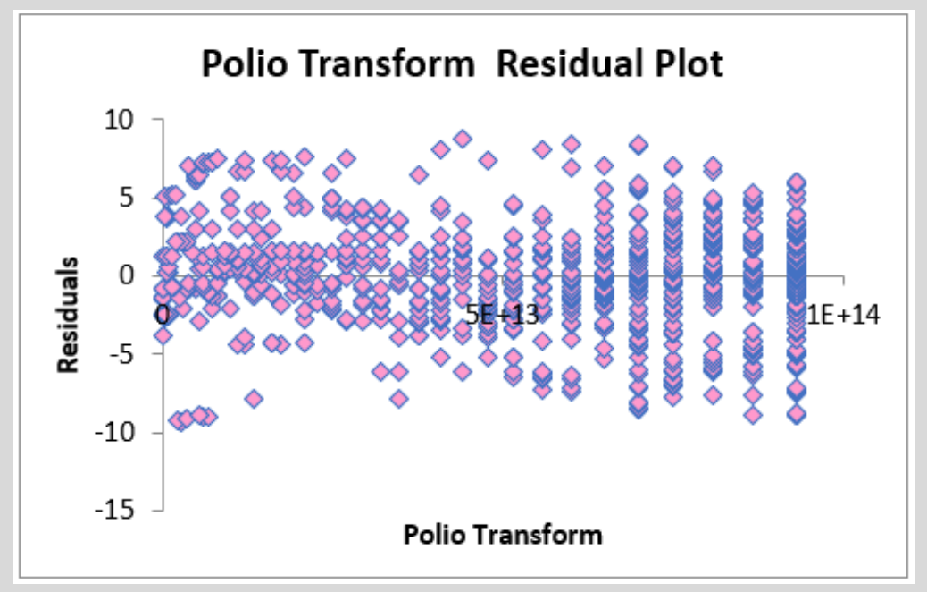


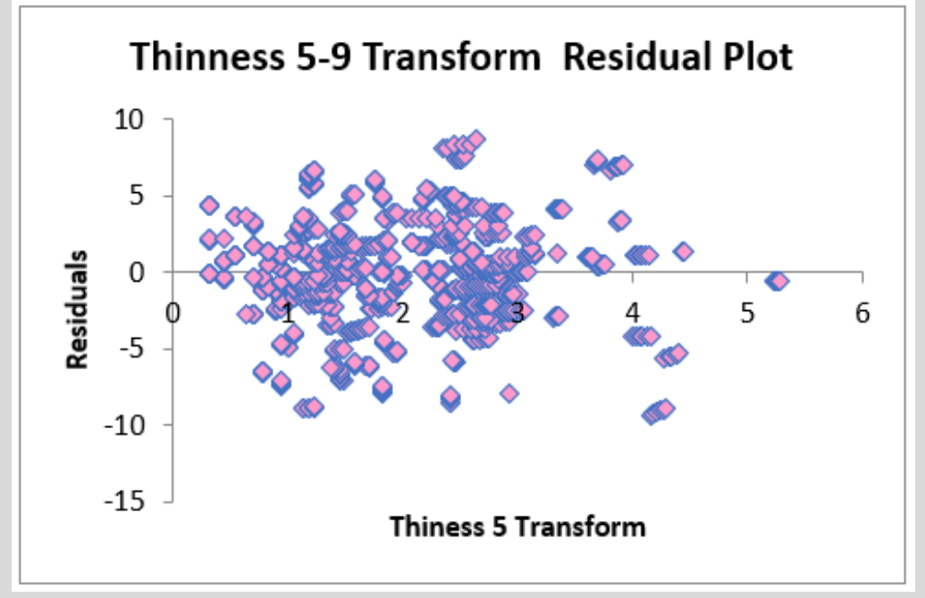
Median Comparisions

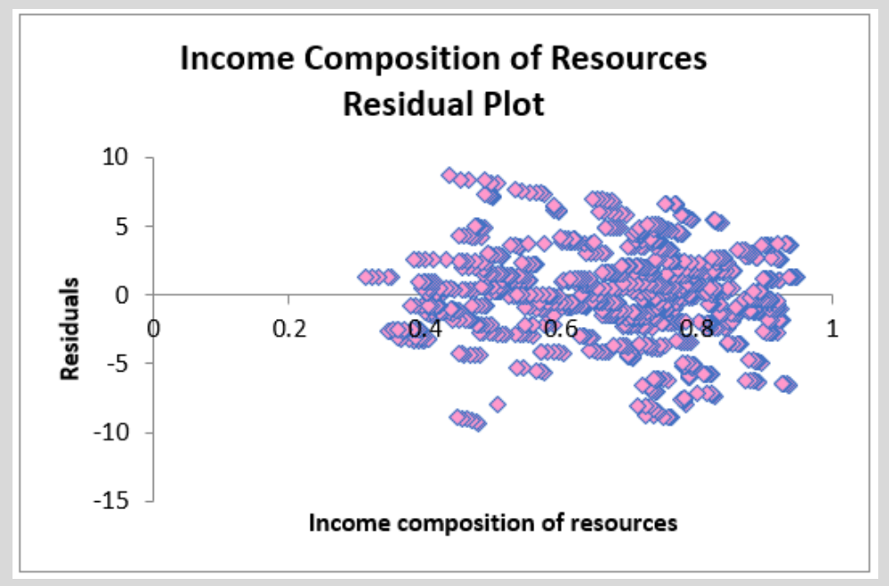
Life Expectancy Class	Life Expectancy	Adult Mortality Rate	Income composition of resources	Average Schooling (Years)	HIV/AIDS	BMI	Under 5 Deaths Per 1000	Polio % Vaccinated	Infant Deaths Per 1000	Diphtheria % Vaccinated	Thinness 5-9 years	Thinness 1-19 years	GDP
Very Low	57.3	35.7%	0.452	9.6	1.30	22.6	42.0	80%	27.5	80%	7.4%	7.4%	$ 604.78
Low	65.9	22.0%	0.562	10.9	0.79	27.7	12.0	87%	9.0	88%	6.1%	5.9%	$ 814.55
Average	74.8	12.9%	0.751	13.5	0.63	56.5	1.0	96%	1.0	95%	2.8%	2.7%	$ 3,961.06
High	82.3	6.4%	0.894	16.3	0.63	61.0	0.0	96%	0.0	96%	0.8%	0.9%	$ 22,486.47

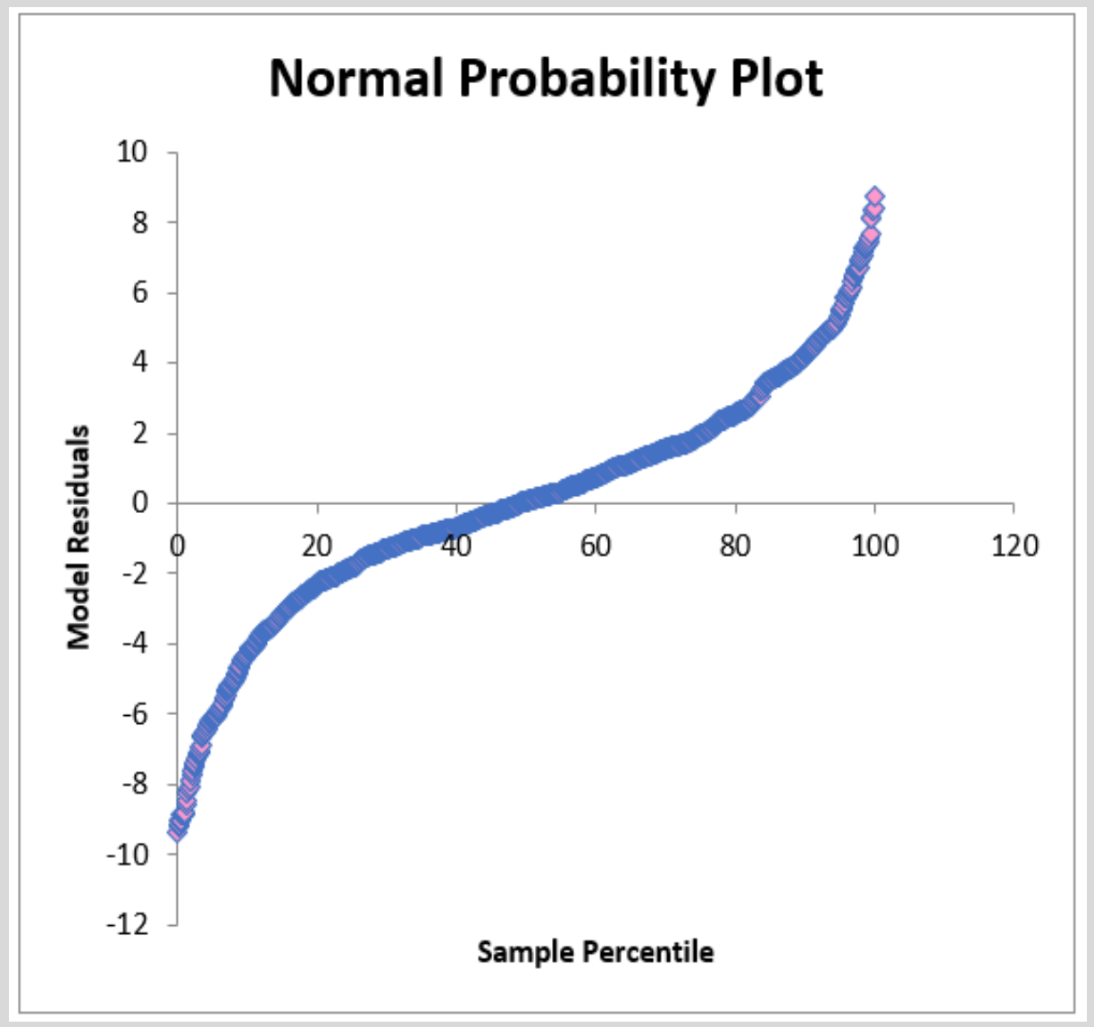
Preliminary Model Results

Regression Statistics	
Multiple R	0.9104
R Square	0.8289
Adjusted R Square	0.8284
Standard Error	3.3192
Significance F	0.0000









Model Parameters	*Coefficients*	*Std Error*	*t Stat*	*P-value*
Intercept	41.964	0.783	53.602	0.0000
Polio Transform	0.000	0.000	5.285	0.0000
Thinness 5 Transform	-0.642	0.140	-4.578	0.0000
Income composition of resources	43.033	0.984	43.713	0.0000

The preliminary model seems to perform well but with a minor violation of the model residuals. While the residual values are constantly varied, the distribution of them has heavy tailed residuals. Solutions for this could be to go back and further clean and edit the data, investigate other potential predictors, or choose different preparation and modeling techniques that can better handle this data.

CONCLUSIONS

Through doing this analysis, we found that Income composition of resources and schooling were the most influential variables when determining a countries average life expectancy. Many charity groups are geared toward providing food and first aid to countries in need and while those things are good, they aren't necessarily helping these countries moved forward. We propose that in addition to giving these countries food and first aid, we provide them with resources to better their education systems and teach them how to use their resources more efficiently. There is a famous quote that states ,"If you give a man a fish, you feed him for a day. If you teach a man to fish, you feed him for a lifetime." If people in these countries were taught how to use their resources in a way that generate long-lasting benefits, then they will not have to depend on more developed countries for help as they will be able to help themselves.