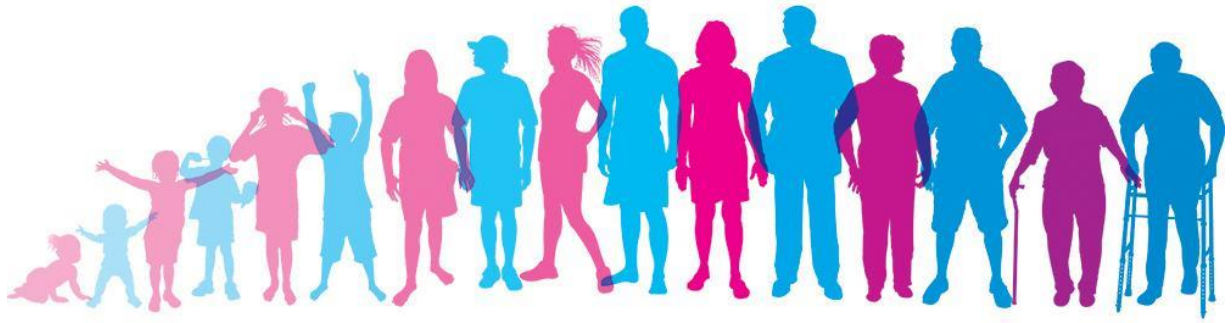
Life Expectancy Prediction

A PROJECT BY SOUMYADIP KUNDU

SPRINGBOARD DATA SCIENCES CAREER TRACK

Background

- Key Statistic for predicting population health
- Regularly monitored by WHO and health departments of all countries
- Various global characteristics affect its trend.
- Changed significantly in the past 20 years dur to increased growth in health sector



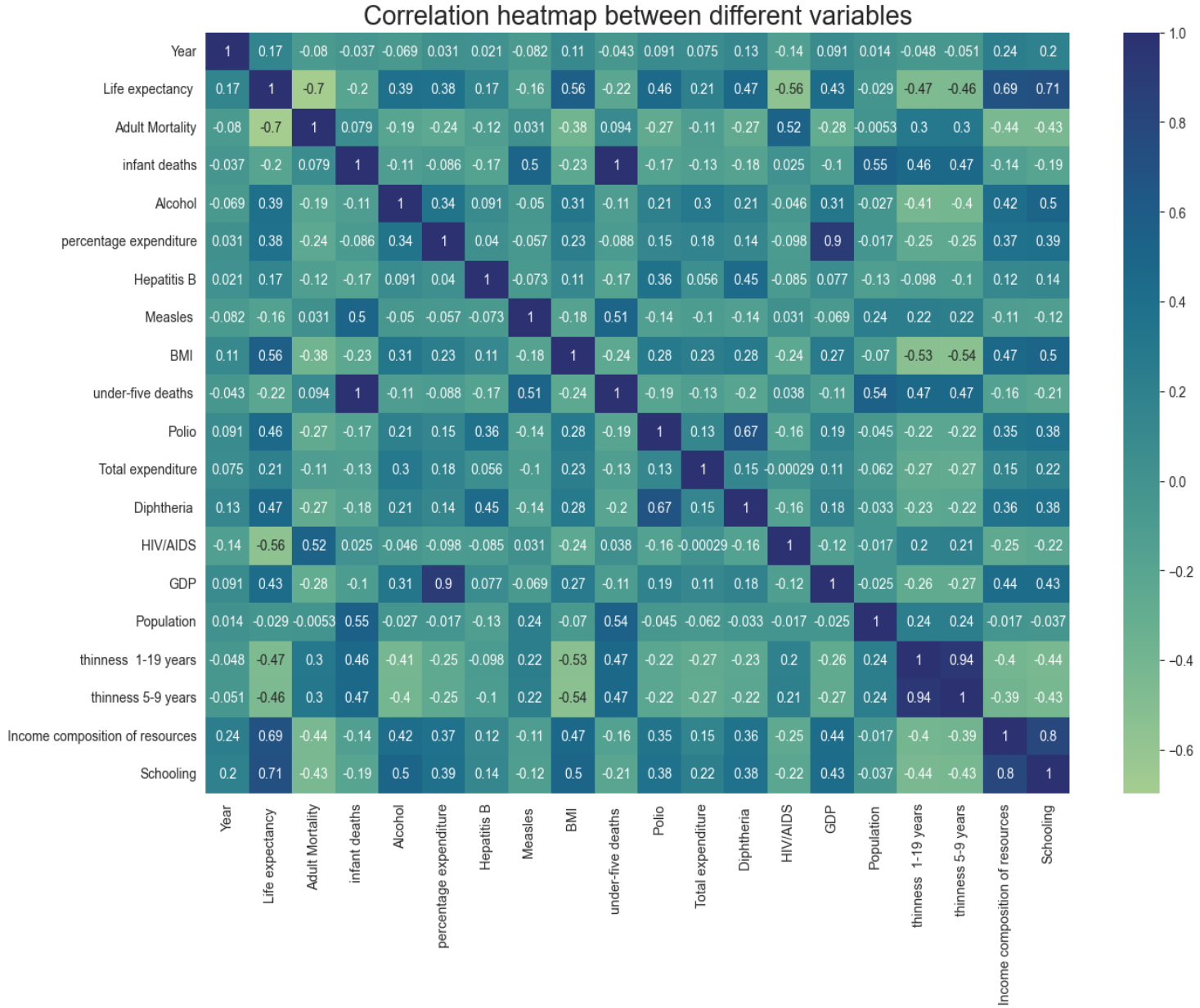
Problem Statement

Assessing the relationship between the various predictive factors that affect an individual's life expectancy in 193 countries.

Dataset

- Generated by The World Health Organization's Global Health Repository
- Downloaded through Kaggle
- 2938 rows and 22 columns


Exploratory Data Analysis
170
160
150
140
130
120
110
Sun
Mon
Tue
Wed
Thu
Fri
Sat



Correlation Heatmap

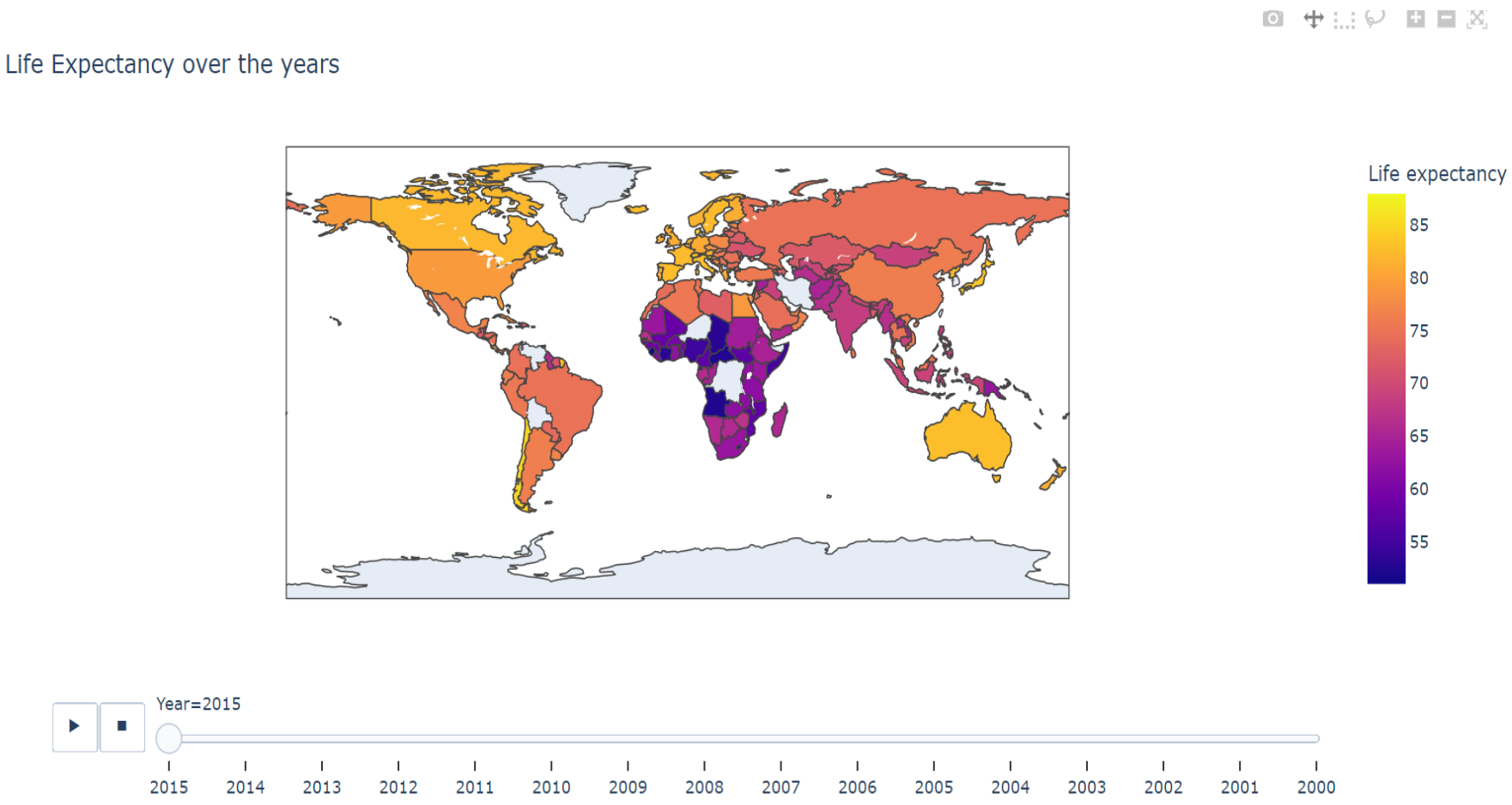
Top Positive correlations:

- Schooling
- Income

Top Negative correlations:

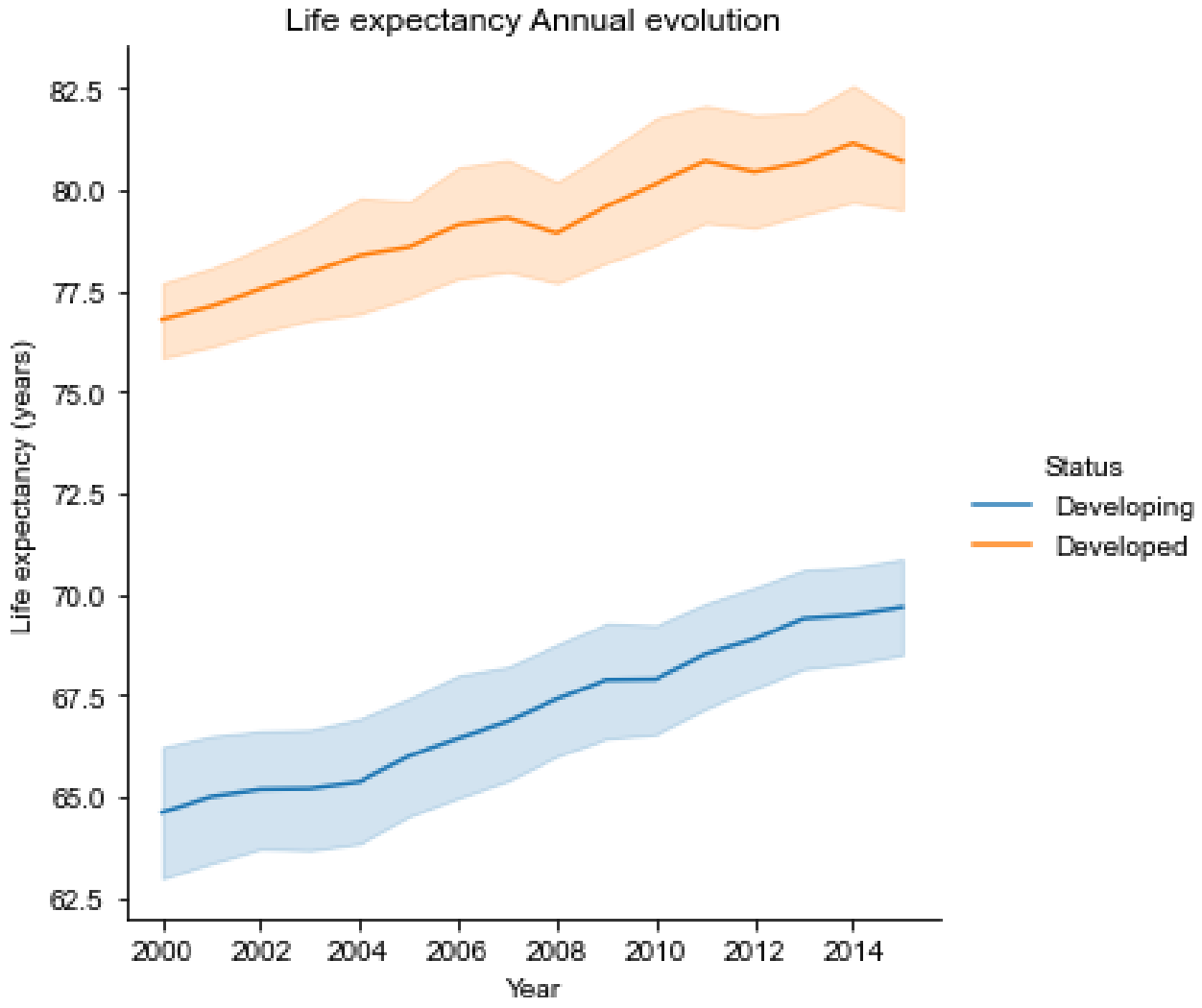
- Adult Mortality
- HIV/AIDS

Chloropleth map of Life Expectancy

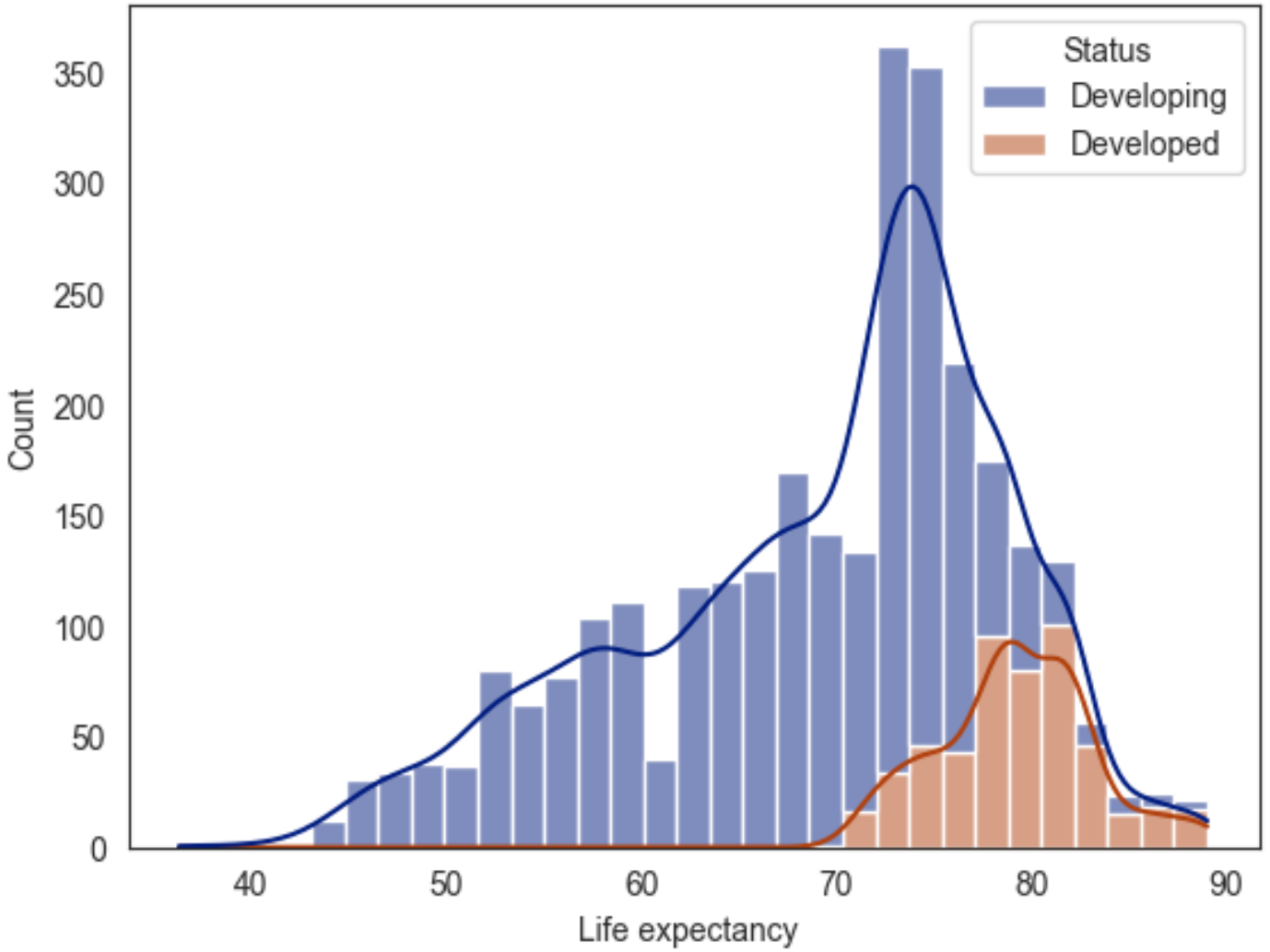


Animated version: https://skundu01.github.io/

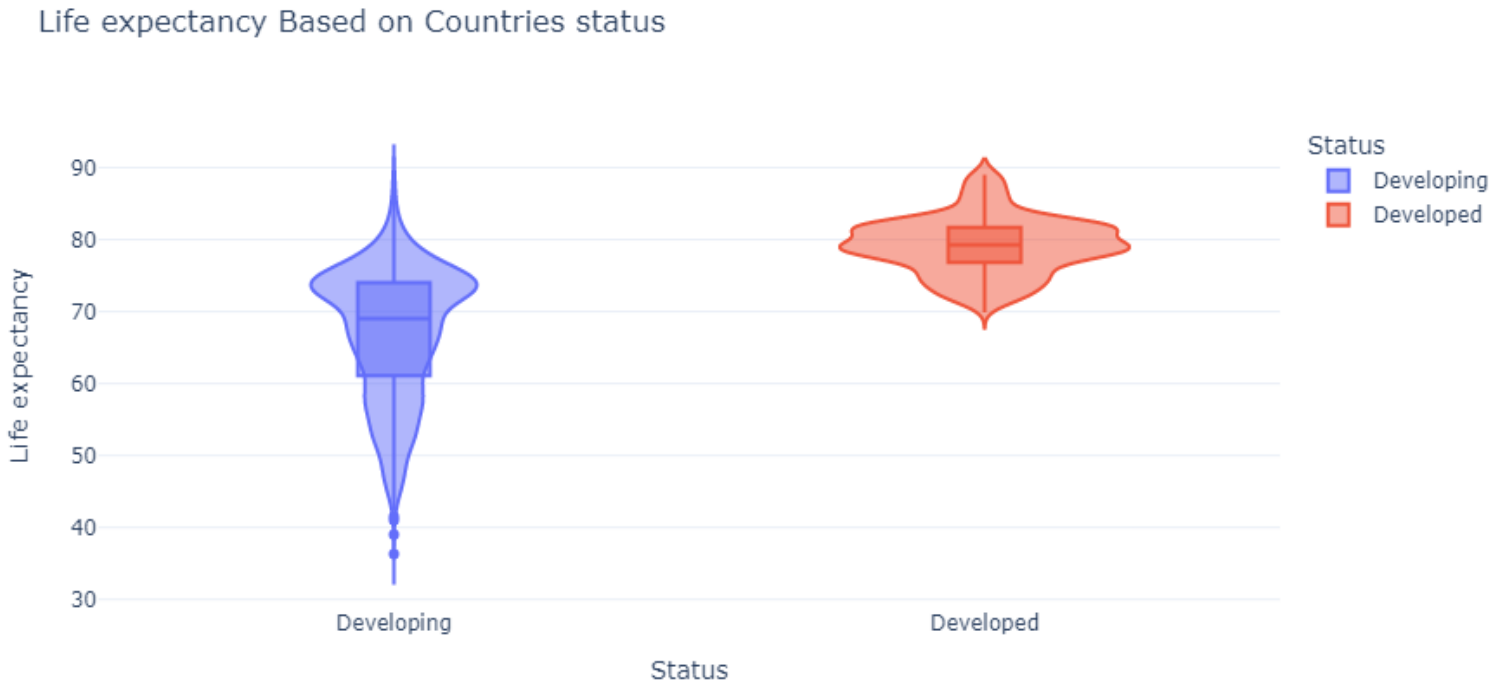
Annual Evolution of Life Expectancy



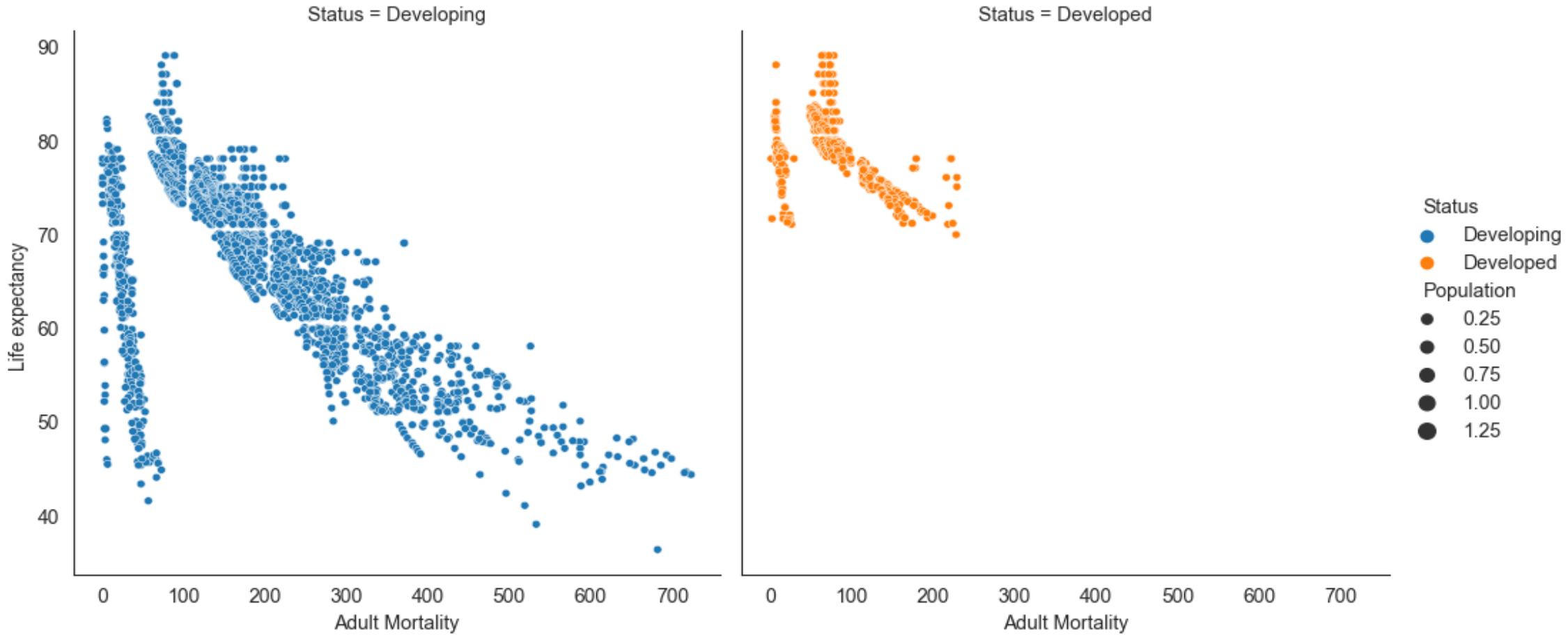
Distribution of Life Expectancy in Developing vs Developed Countries



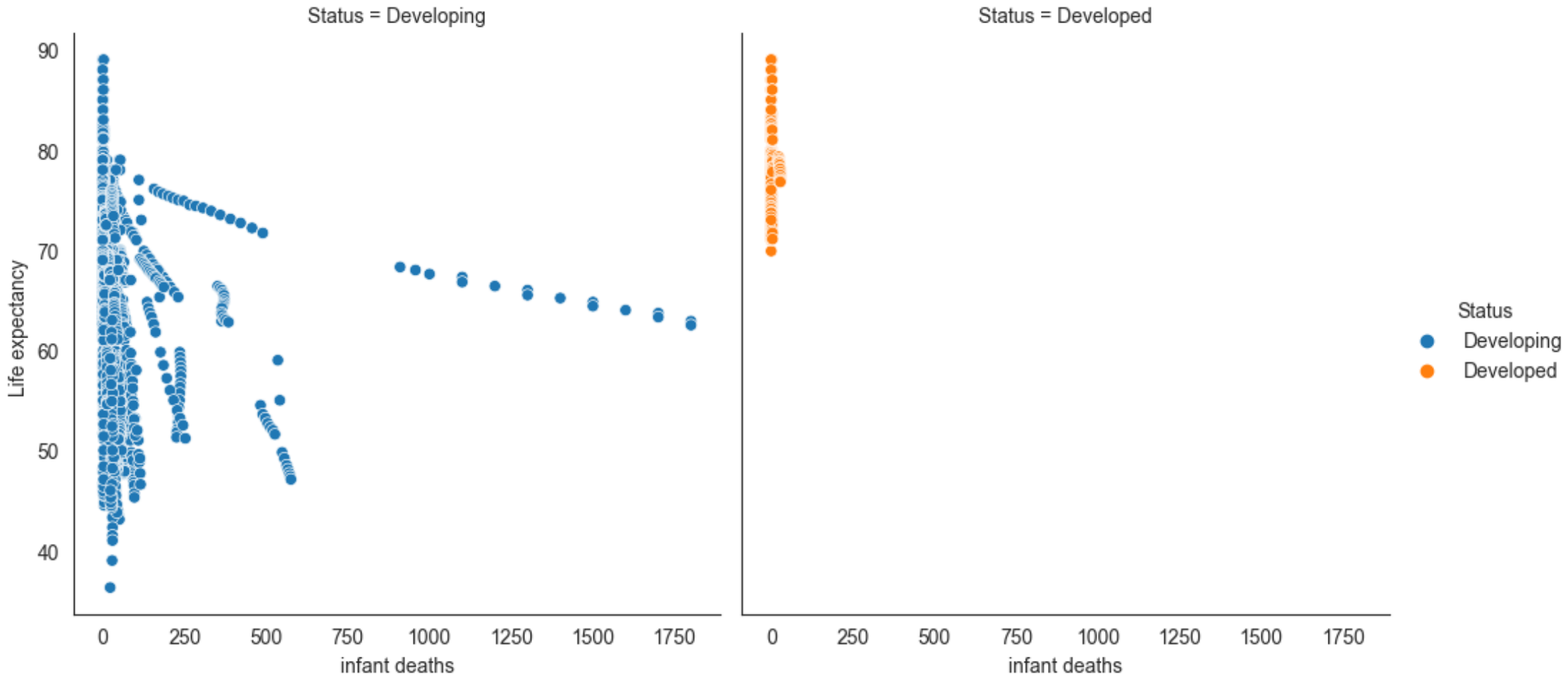
Violin plot of Life expectancy based on status



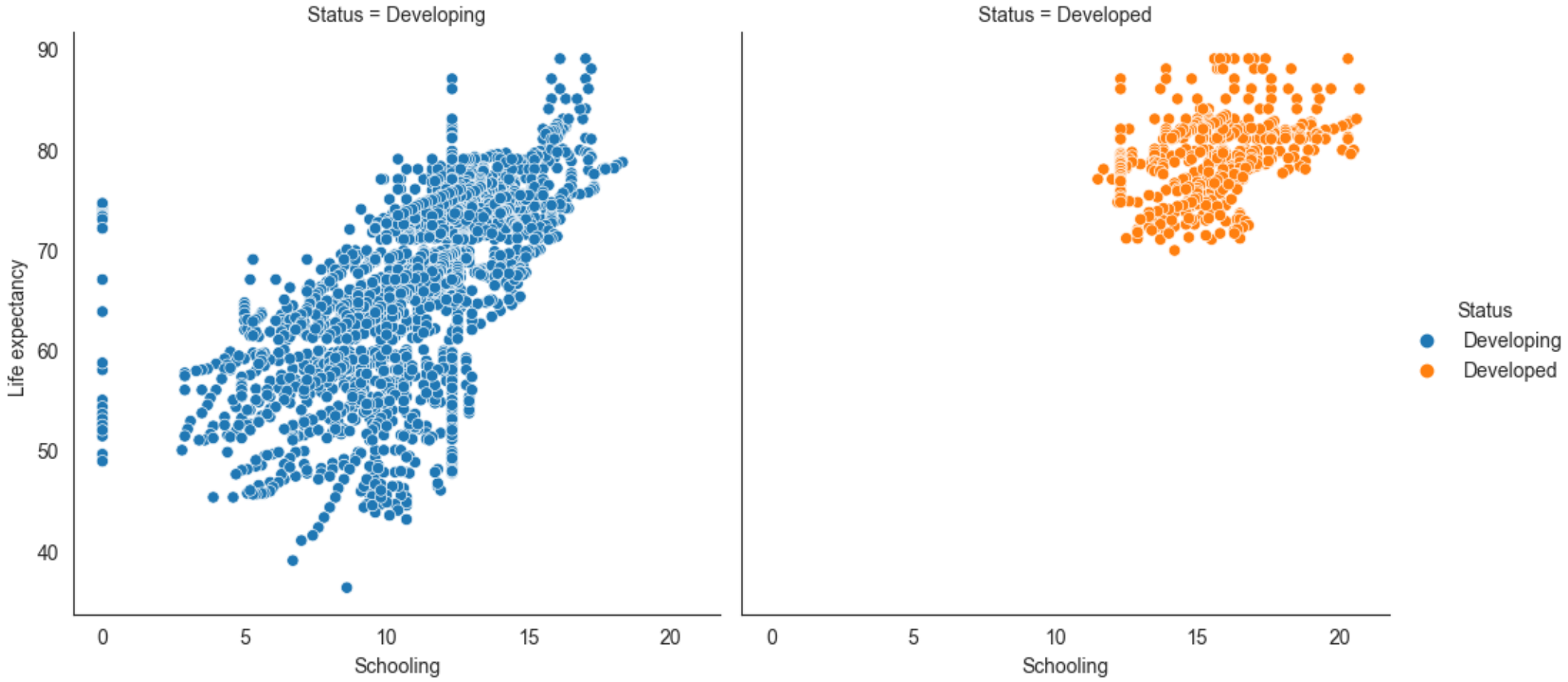
Life Expectancy vs Adult Mortality



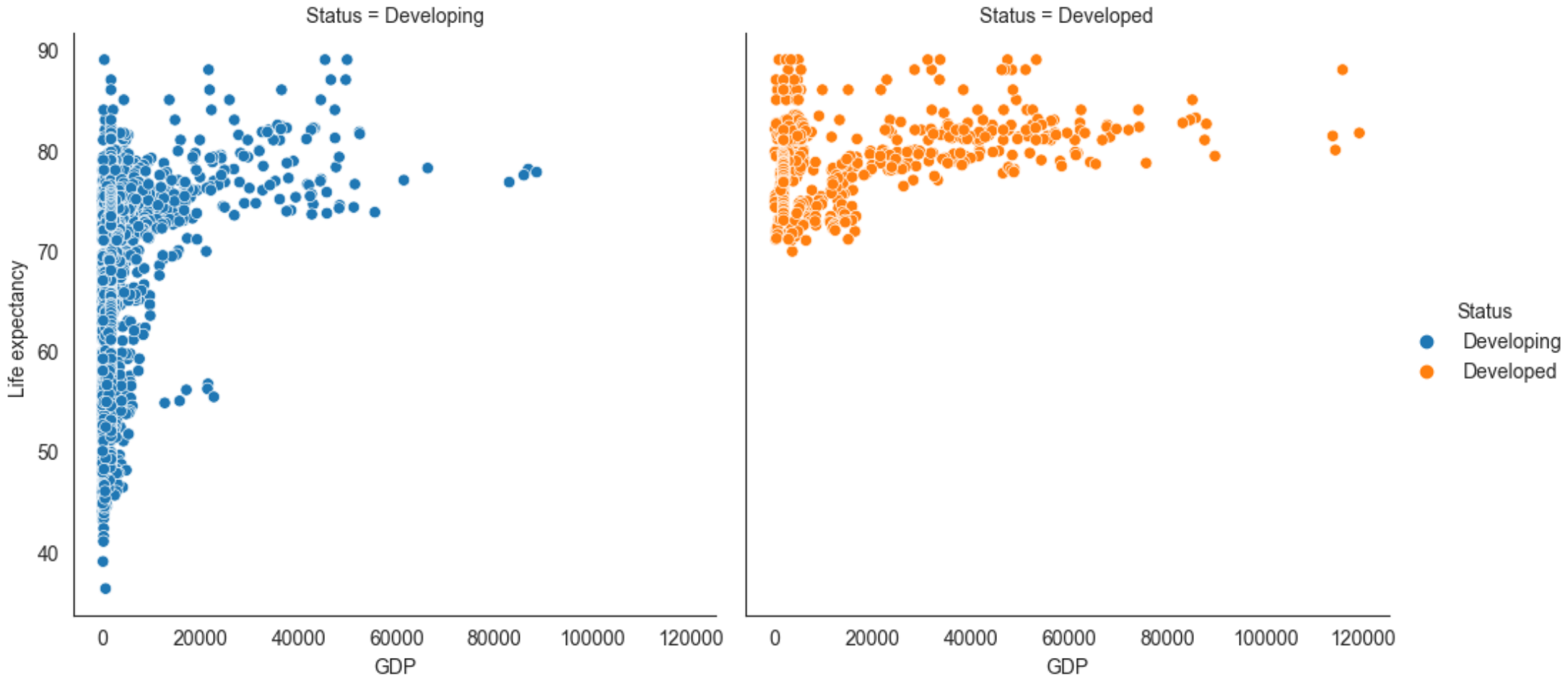
Life Expectancy vs Infant Deaths



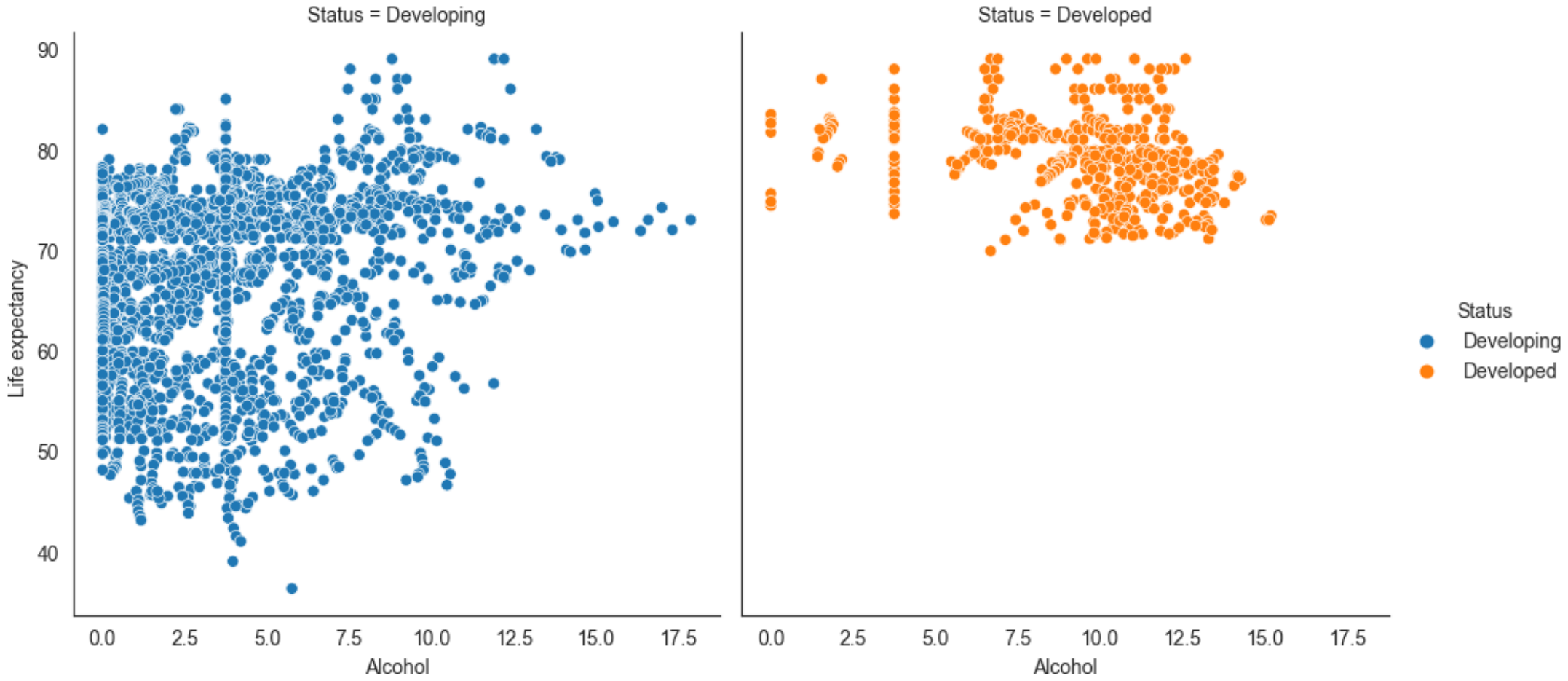
Life Expectancy vs Schooling



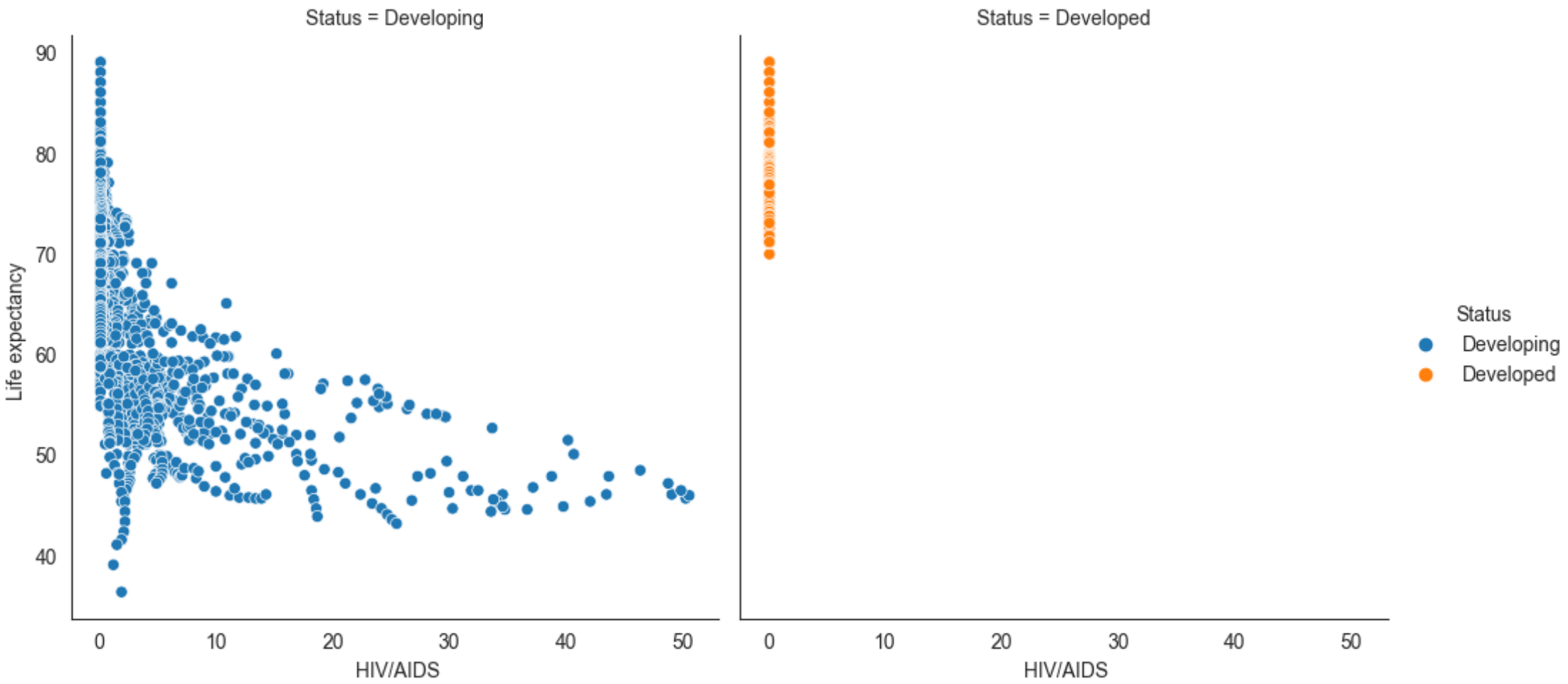
Life Expectancy vs GDP



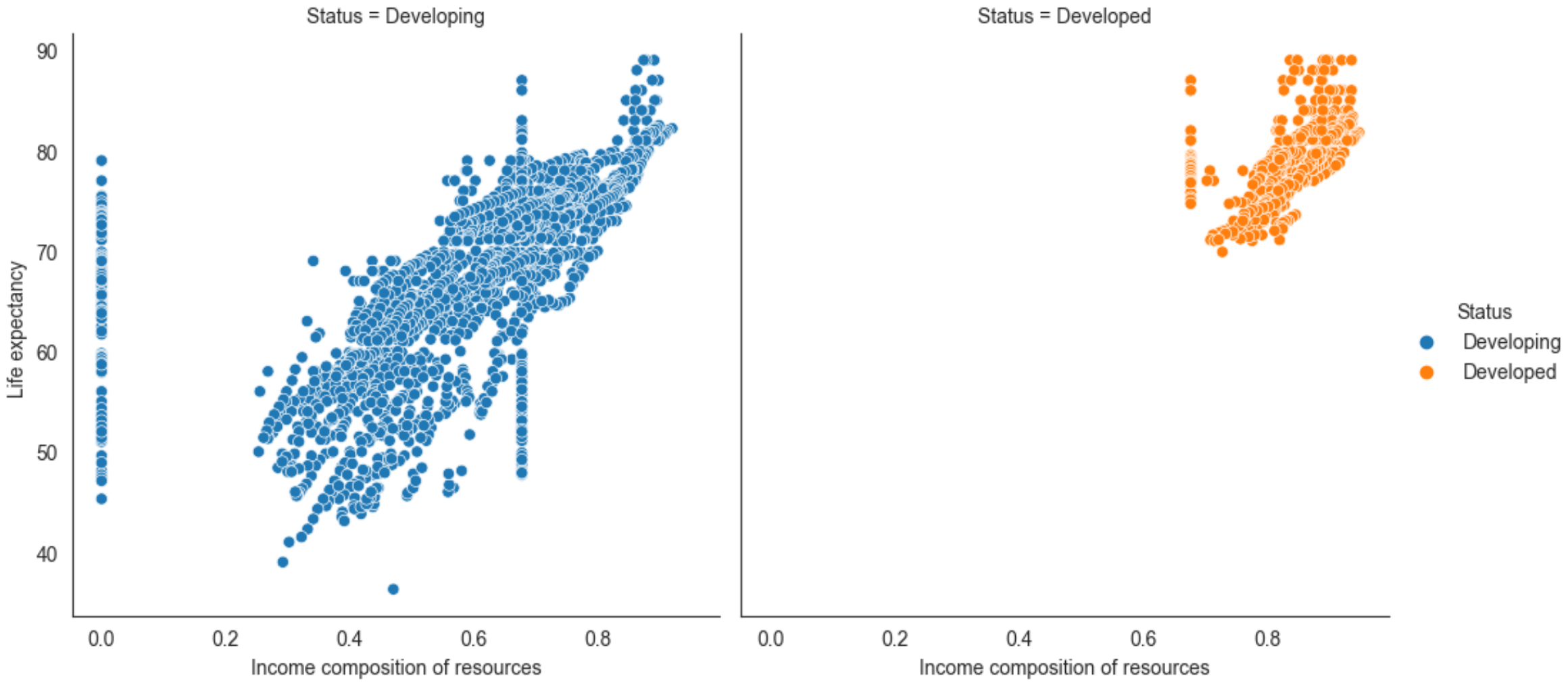
Life expectancy vs Alcohol Consumption



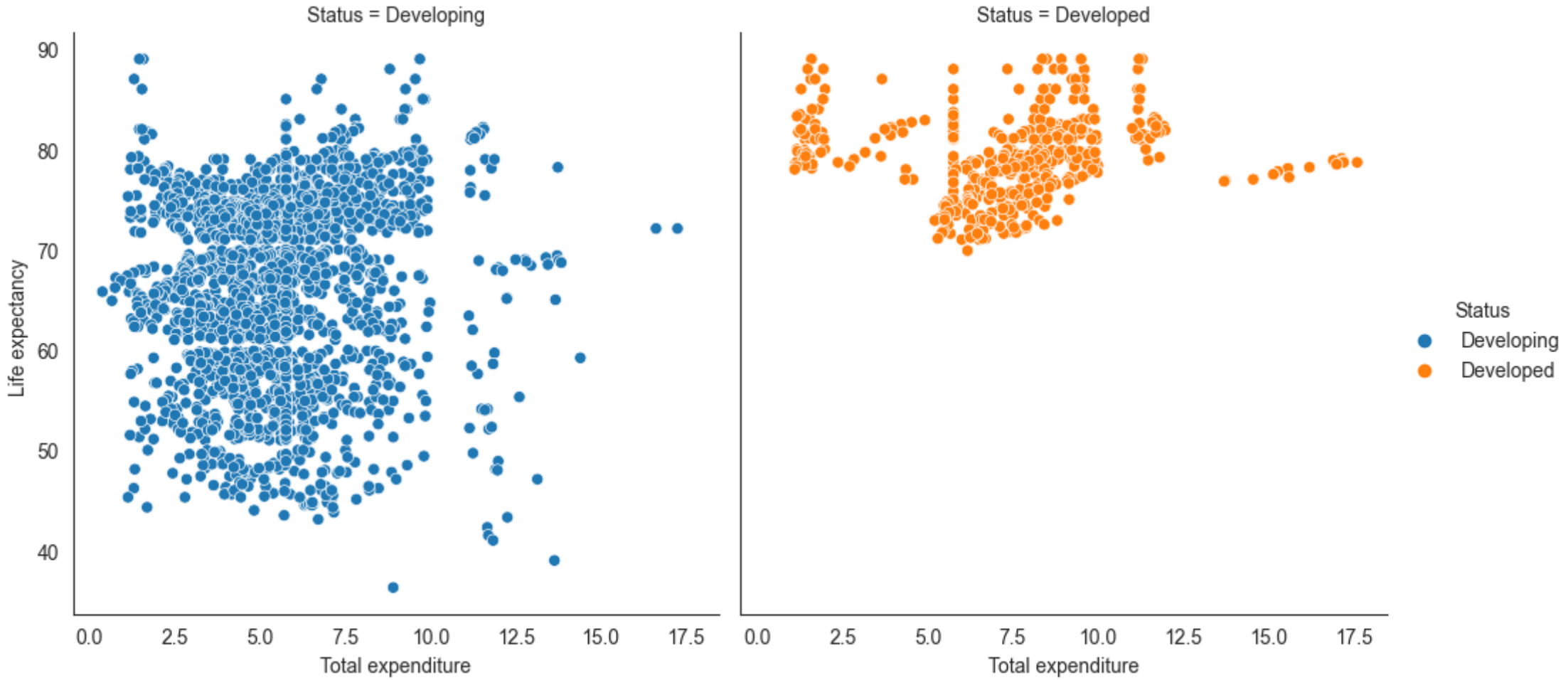
Life Expectancy vs HIV/AIDS



Life Expectation vs Total Income



Life Expectancy vs Total Expenditure



Data Cleaning and Preprocessing

- 3 columns contained about 25% missing data
- Simple Imputer used to impute values with the median values
- Numerical values normalized by MinMaxScaler
- Categorical values encoded by Label Encoder
- 80-20 train-test split created with shuffle

Data Modeling

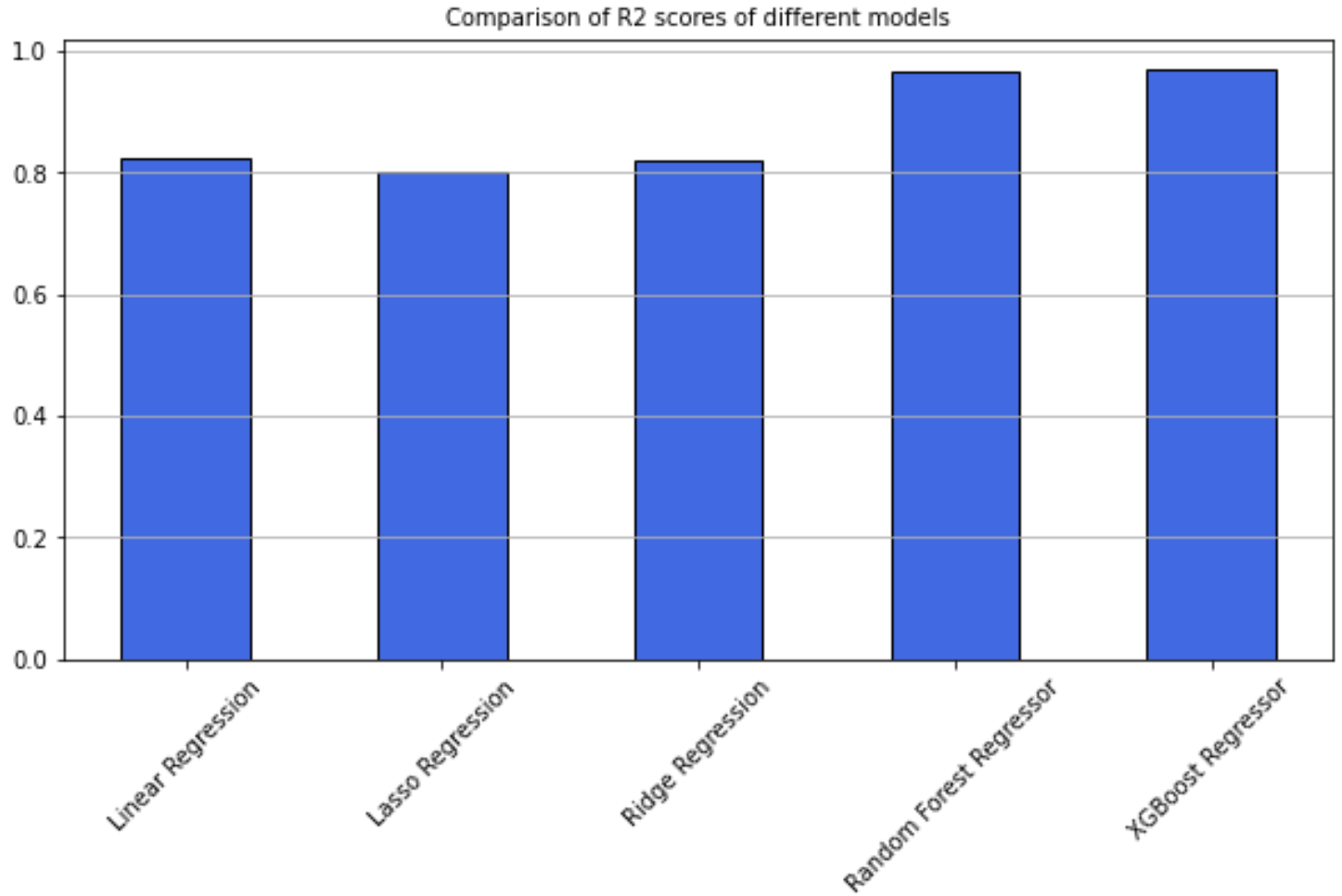
Models used

- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest Regressor
- XG Boost Regressor

Model Performance Comparison

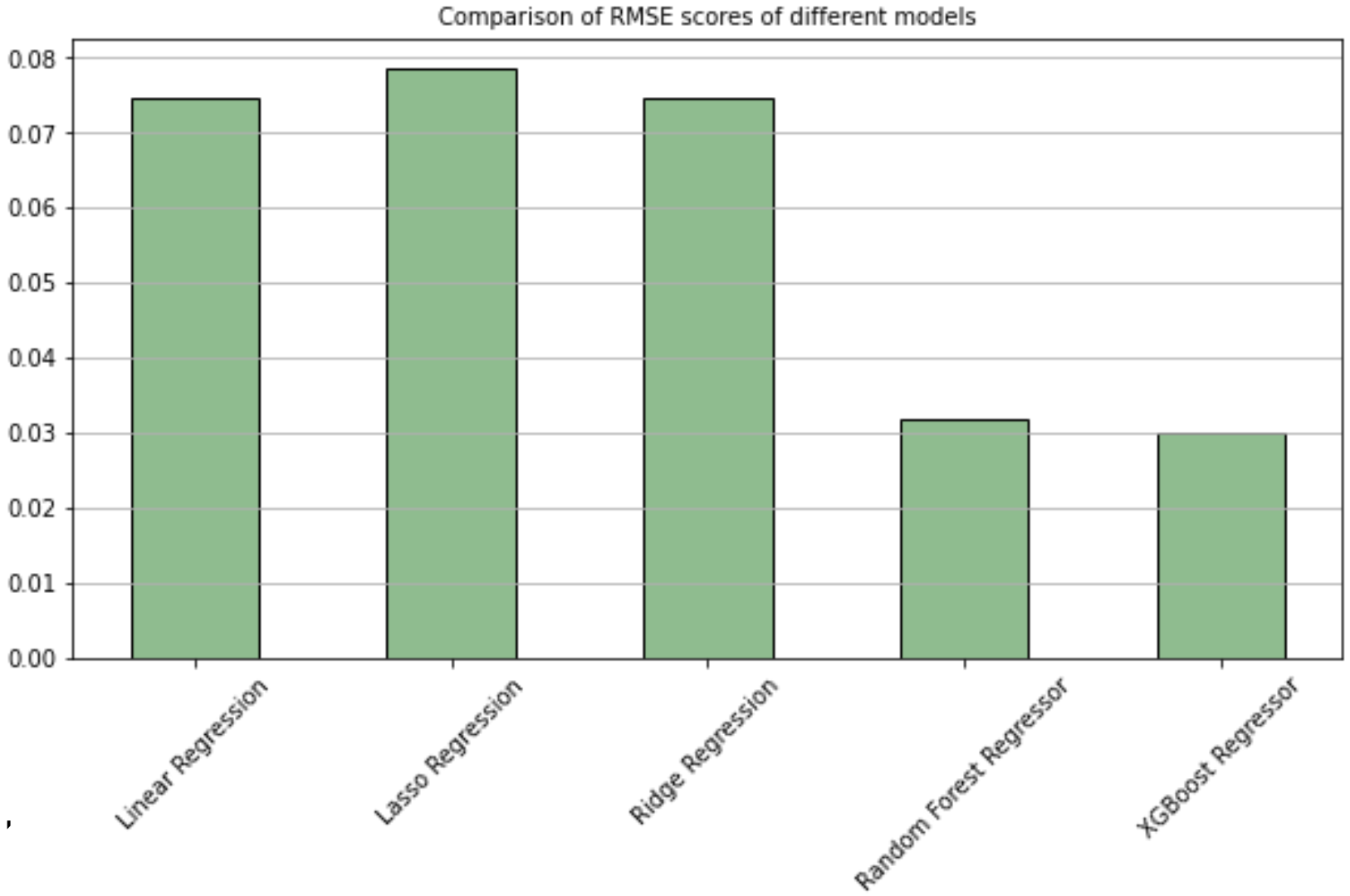
Model Name	RMSE	R2
Linear Regression	0.074506	0.822111
Lasso Regression	0.078636	0.801845
Ridge Regression	0.074777	0.820814
Random Forest Regression	0.031716	0.967766
XG Boost Regression	0.030095	0.970976

Model Performance Metrics



HIGHER THE R2 SCORE, BETTER THE MODEL

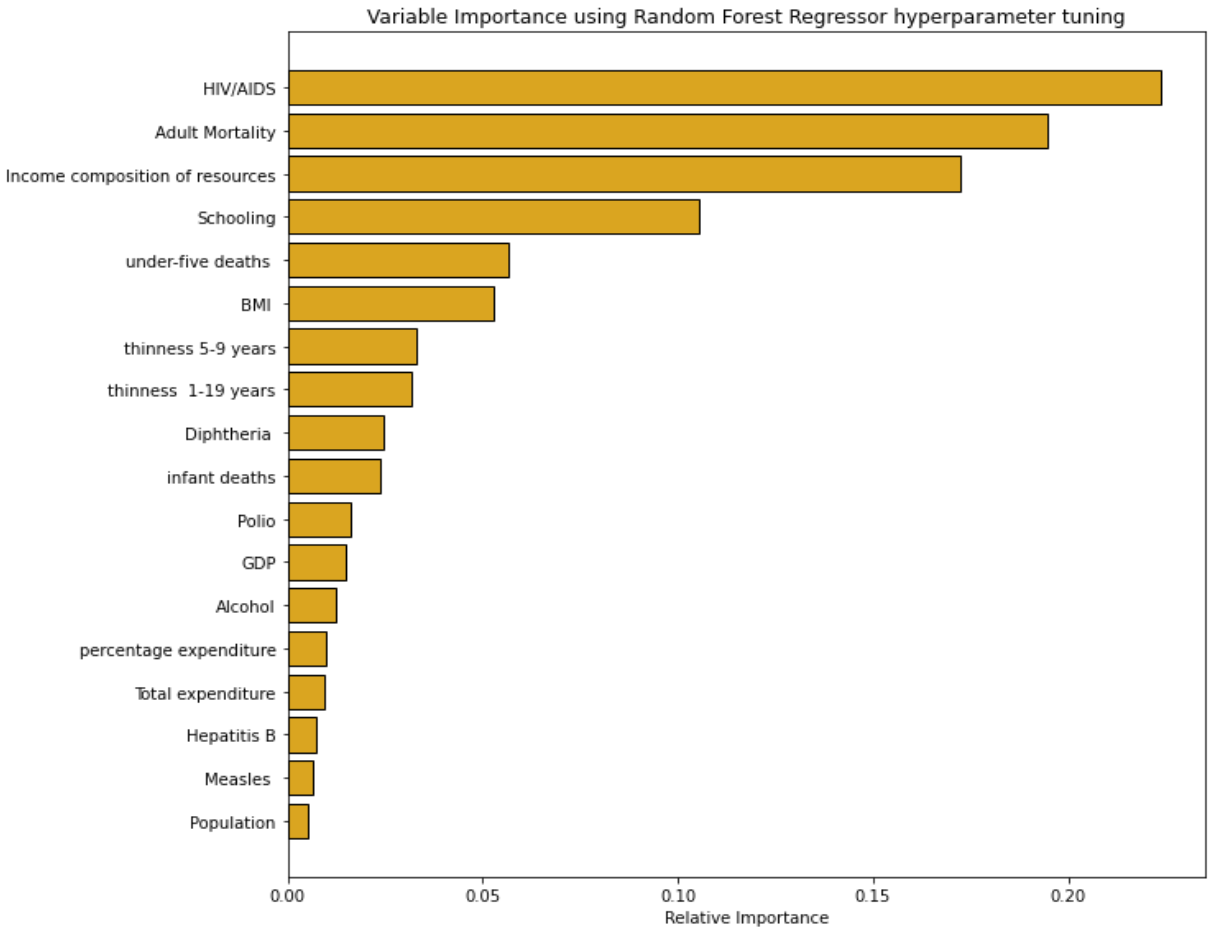
Model Performance Metrics contd...



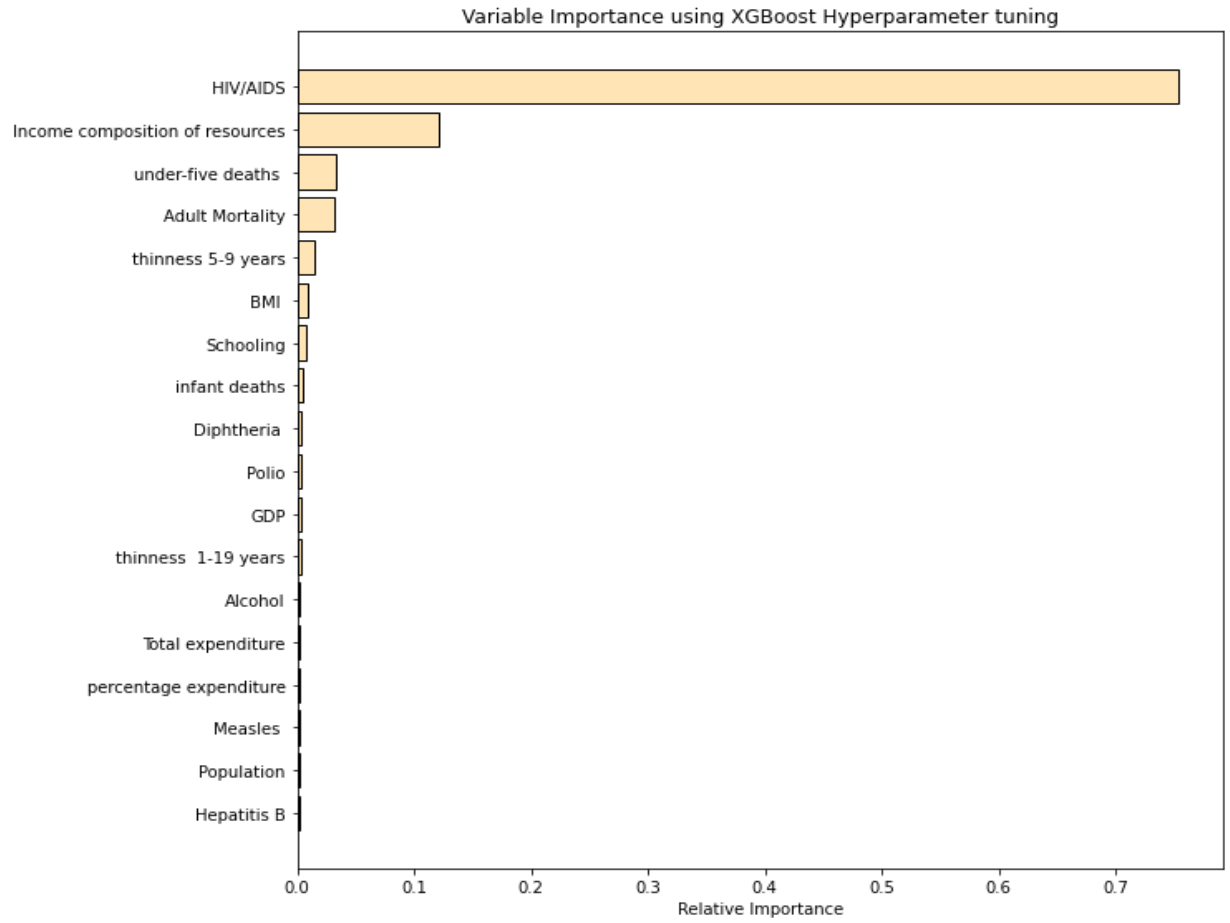
LOWER THE RMSE SCORE, BETTER THE MODEL

Hyperparameter tuning and Feature Selection

RANDOM FOREST



XG BOOST



Conclusion

Most important predictors according to Random Forest

- HIV/AIDS
- Adult Mortality
- Income
- Schooling

Most important predictors according to XG Boost

- HIV/AIDS
- Income
- Infant deaths
- Adult Mortality

Acknowledgements

- Kenneth Gil-Pasquel (Mentor)
- Springboard team
- World Health Organization
- Kaggle
- Cover images – Google images