Life Expectancy: Potential Factors that may Affect Lifespan

Georgia Institute of Technology
ISYE 6414 – Regression Analysis
Semester Project – Nov 27, 2023

Group 8:
Ayushi Goyal
Divij Mishra
Aravind Rajeev Nair
Krishna Raj
Zhibo Zhang

Georgia Tech

Table of Content

1.Data Introduction

2.Project Flow*

- Data observation
- Data deduction
- Model
- Evaluation

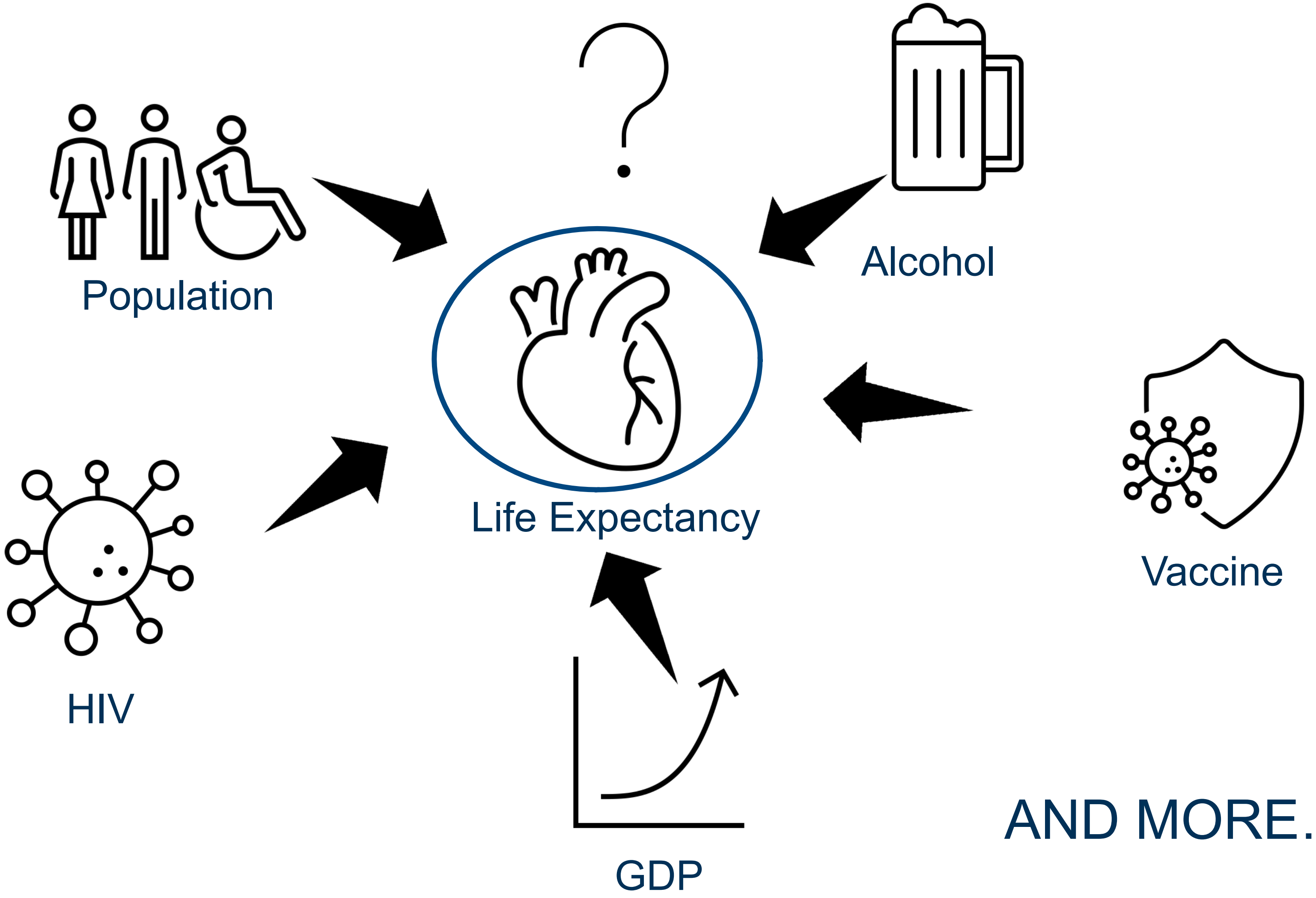
3.Findings and Interpretation*

4.Suggestions and Applications

5.Future Work

1. Data introduction

Project Question: What factors are impacting lifespan across regions and time?



AND MORE…

1. Data introduction

Data set contains:

Life expectancy, health, immunization, economic, demographic, etc.

About **179 countries from 2000-2015 years**

The primary source of this data is Kaggle which consolidates information from the World Health Organization (WHO) and the World Bank.

Quantitative variables (17):

- Infant deaths
- Under-five deaths
- Adult mortality
- Alcohol consumption
- Hepatitis B
- Measles
- BMI
- HIV Incidents
- Schooling
- GDP per capita
- Population
- Thinness 10-19 years
- Thinness 5-9 years
- Life expectancy
- Polio
- Diphtheria
- Year

Qualitative Variables (4):

- Country
- Region
- Economy status Developed
- Economy status Developing

Response Variables:

- Life expectancy

1. Data introduction

Project Goals

□ Insight into Health Metrics:
- Significant factors
- Highlight insight into public health interventions

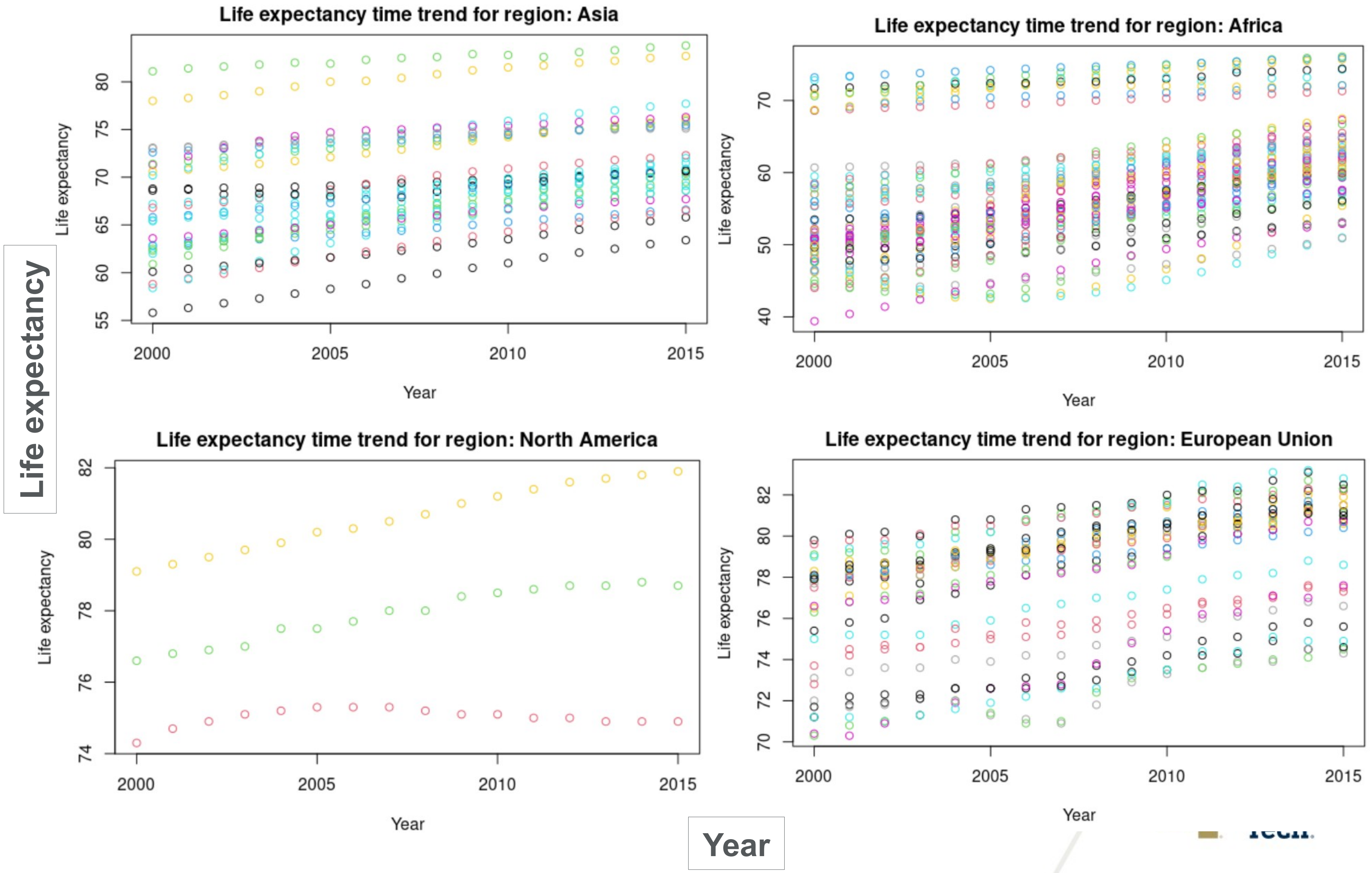
□ Understanding Life-Altering Factors:
- Elucidate potential life-altering risks

□ Model Utility and Application:
- Build a practical model - can be used in applications

2. Project Flow – Data observation

Time Series:

- Other regions are not shown here
- Increasing trend of life expectancy generally
- Not strictly linear
- Rate of increase varies significantly across different regions and time periods



2. Project Flow – Data observation

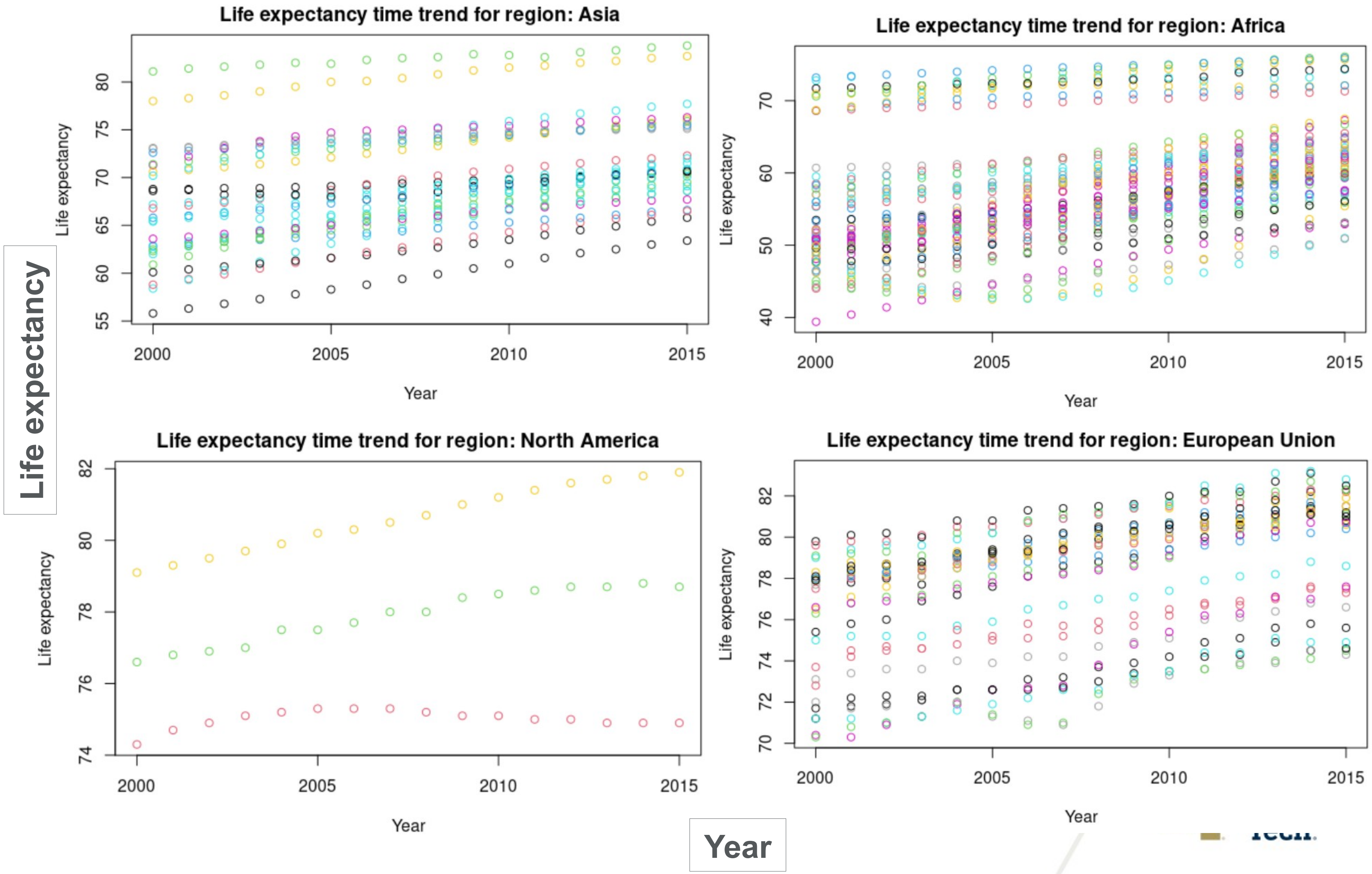
Uptrend Attribution:

- Economic development
- Development in Healthcare

Therefore

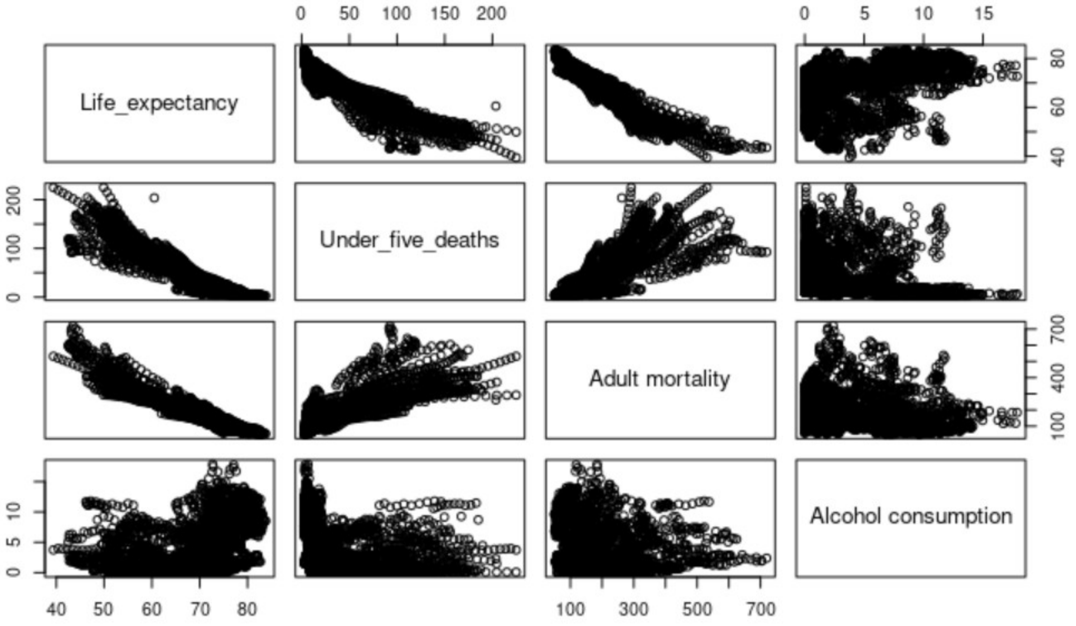
Analyze Non-Time Variables

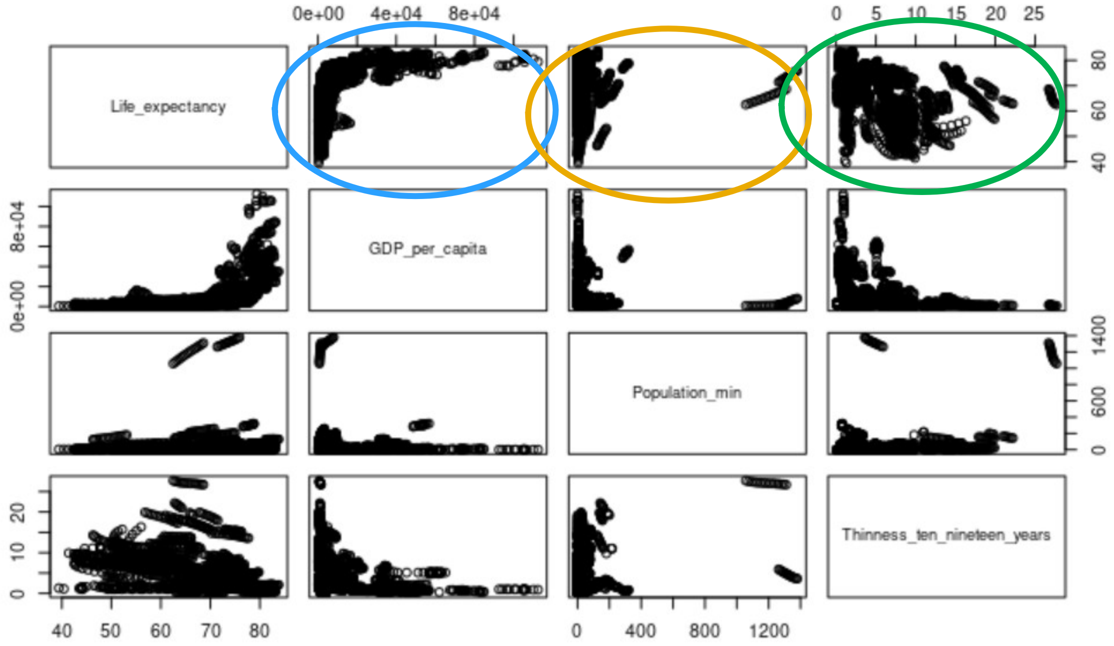
We want: The factors aside from the passage of time impact life expectancy



2. Project Flow – Data observation

Exploratory Data Analysis - Scatter Plots





Relationships: life expectancy vs. independent variables

- A portion of the plots is shown
- Transformation may need
- Low Correlation Variables
- Potential Removal

2. Project Flow – Data deduction

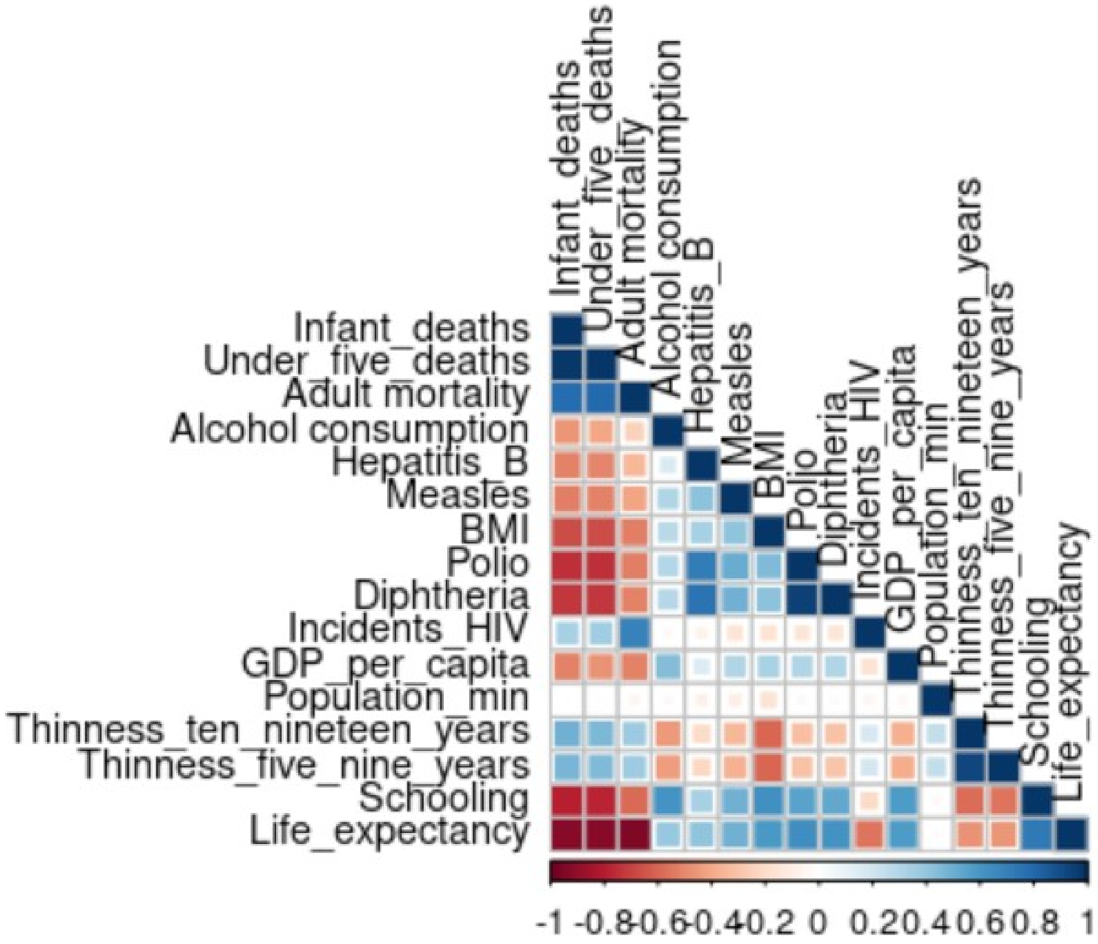
Multicollinearity:

There is a presence of multicollinearity, especially between various disease rates and mortality rates

Model Variables selection:

Choose a set of variables that are more strongly correlated with life expectancy:

- BMI
- GDP_per_capita
- Schooling
- Region
- Economy_status_Developed
- Infant_deaths
- Adult_mortality

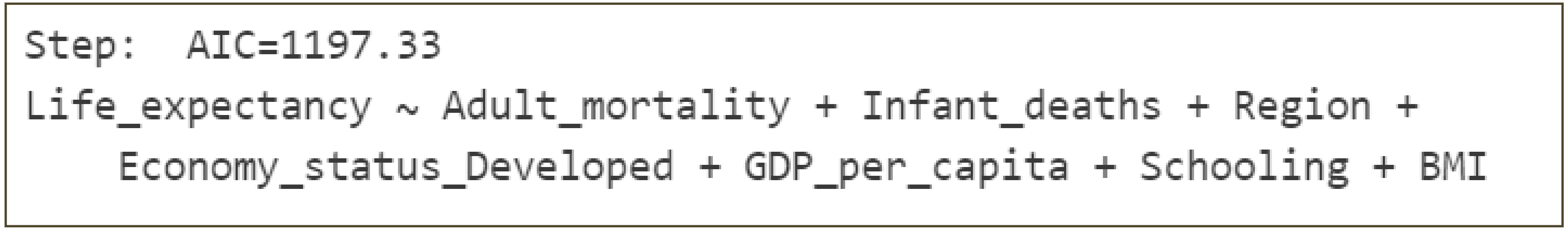


2. Project Flow – Main Model

Multiple Linear Regression (by R)

- Taking a 90:10 split of Training and Validation data sets
- Apply Forward Stepwise Regression (AIC)
- Apply Backward Stepwise Regression (AIC)

Both directions generate the same result:



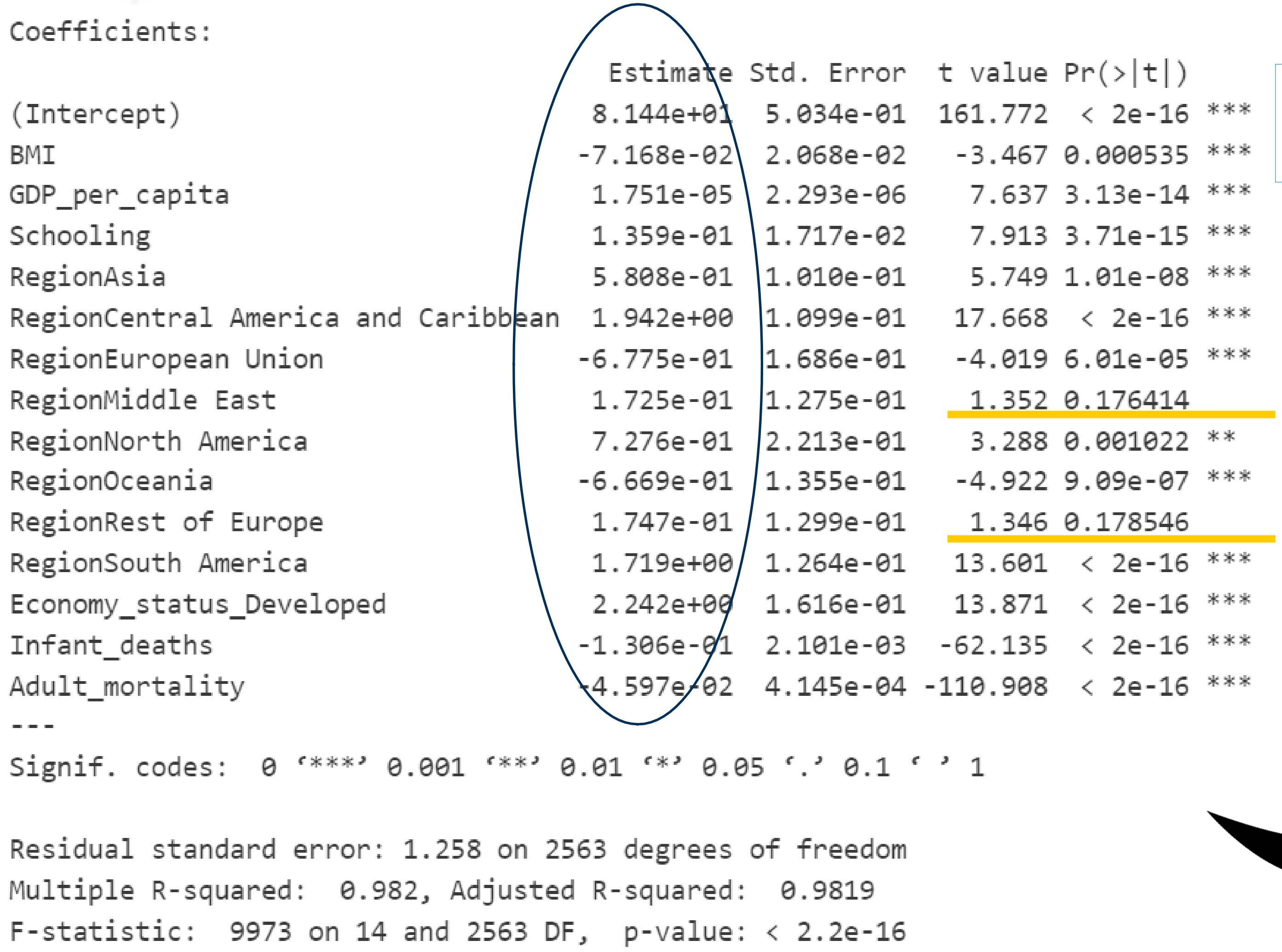
```
Step:  AIC=1197.33
Life_expectancy ~ Adult_mortality + Infant_deaths + Region +
    Economy_status_Developed + GDP_per_capita + Schooling + BMI
```

Therefore, all selected variables are included under this criteria

Note: Region is a Qualitative Variable

2. Project Flow – Main Model

Model Summary:



```
Coefficients:
                                     Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                         8.144e+01  5.034e-01  161.772  < 2e-16 ***
BMI                                -7.168e-02  2.068e-02   -3.467 0.000535 ***
GDP_per_capita                      1.751e-05  2.293e-06    7.637 3.13e-14 ***
Schooling                           1.359e-01  1.717e-02    7.913 3.71e-15 ***
RegionAsia                          5.808e-01  1.010e-01    5.749 1.01e-08 ***
RegionCentral America and Caribbean 1.942e+00  1.099e-01   17.668  < 2e-16 ***
RegionEuropean Union               -6.775e-01  1.686e-01   -4.019 6.01e-05 ***
RegionMiddle East                   1.725e-01  1.275e-01    1.352 0.176414    
RegionNorth America                 7.276e-01  2.213e-01    3.288 0.001022 ** 
RegionOceania                      -6.669e-01  1.355e-01   -4.922 9.09e-07 ***
RegionRest of Europe                1.747e-01  1.299e-01    1.346 0.178546    
RegionSouth America                 1.719e+00  1.264e-01   13.601  < 2e-16 ***
Economy_status_Developed            2.242e+00  1.616e-01   13.871  < 2e-16 ***
Infant_deaths                      -1.306e-01  2.101e-03  -62.135  < 2e-16 ***
Adult_mortality                    -4.597e-02  4.145e-04 -110.908  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.258 on 2563 degrees of freedom
Multiple R-squared:  0.982, Adjusted R-squared:  0.9819 
F-statistic:  9973 on 14 and 2563 DF,  p-value: < 2.2e-16
```

Respectively,

…

Evaluation!

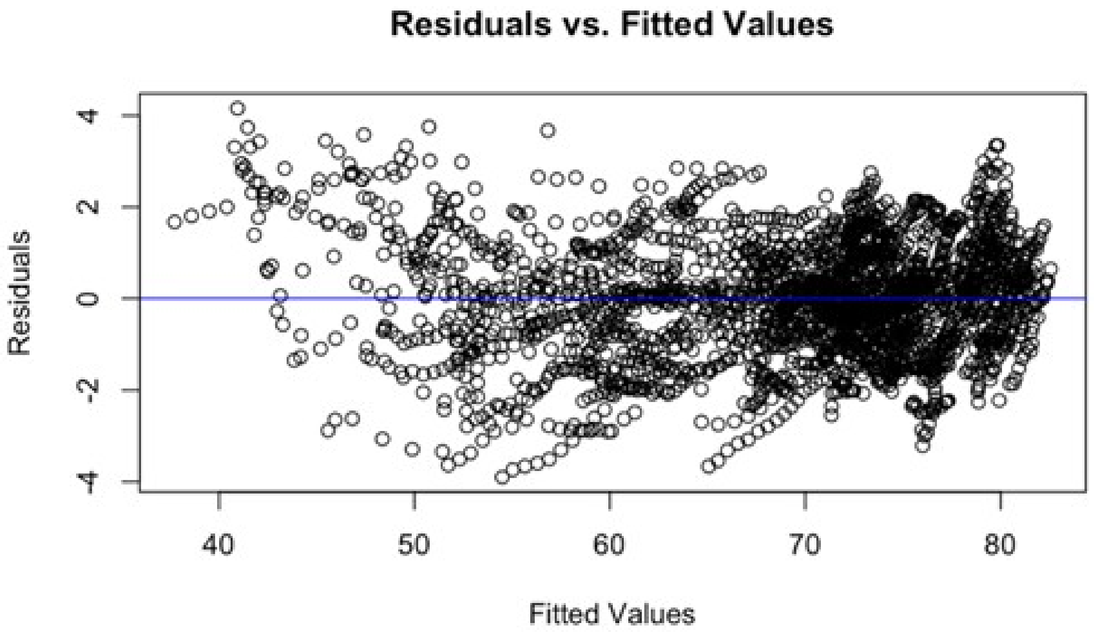
2. Project Flow – Evaluation

Residual and Diagnostics Analysis:

Residual

- Mean of the residuals: 5.383161e-17 ≈ 0

- Plot of Residuals vs. Fitted Values:
  - Absence of patterns
  - Good model fit
  - Constant variance

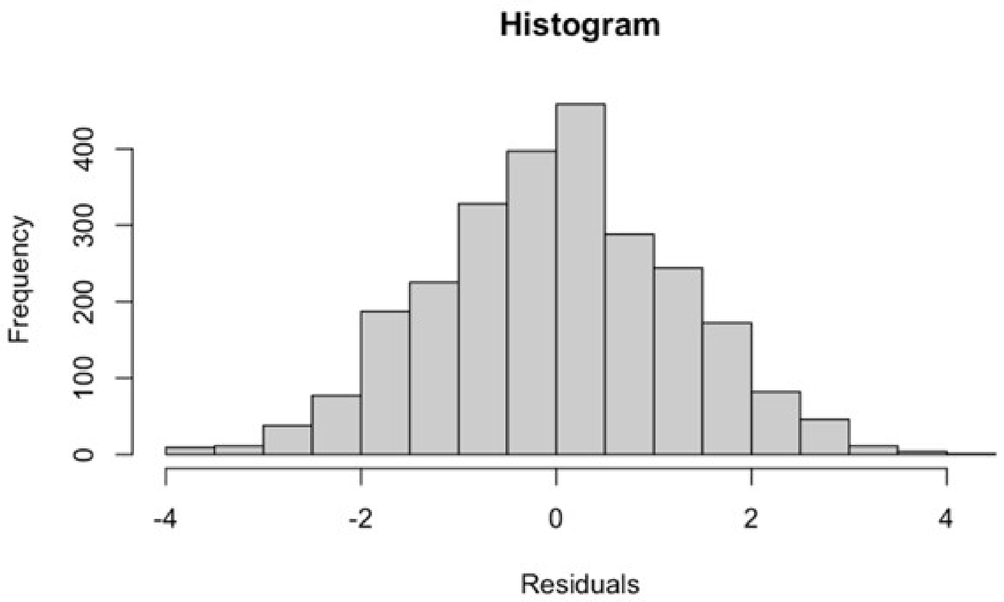


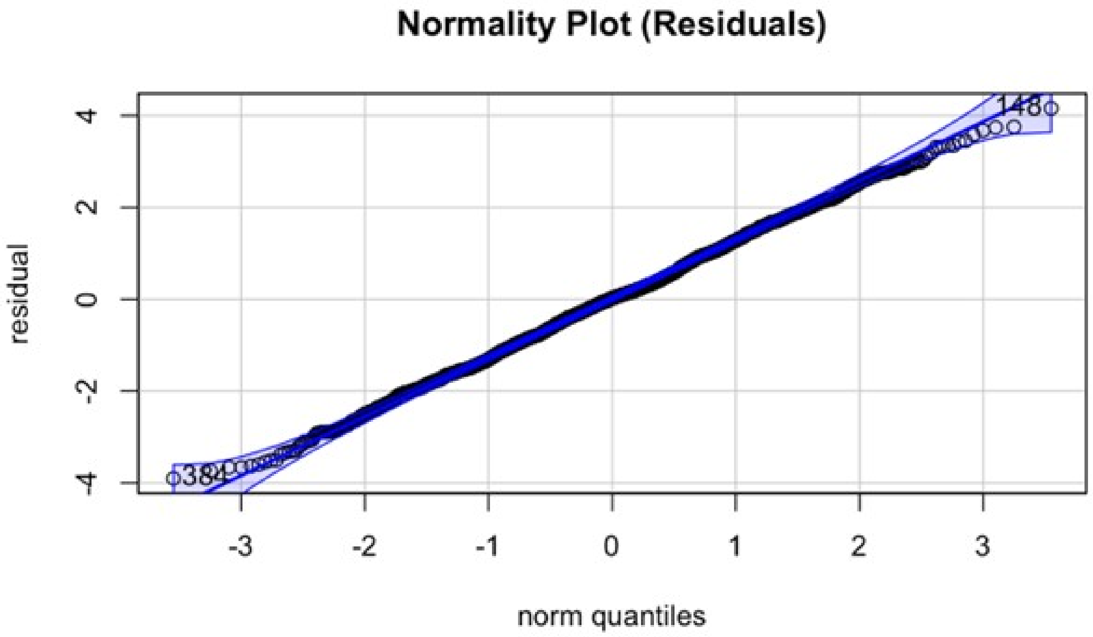
2. Project Flow – Evaluation

Residual and **Diagnostics** Analysis:

Normality

- Histogram and QQ plots of residuals **suggested normality – Linearity Assumption**



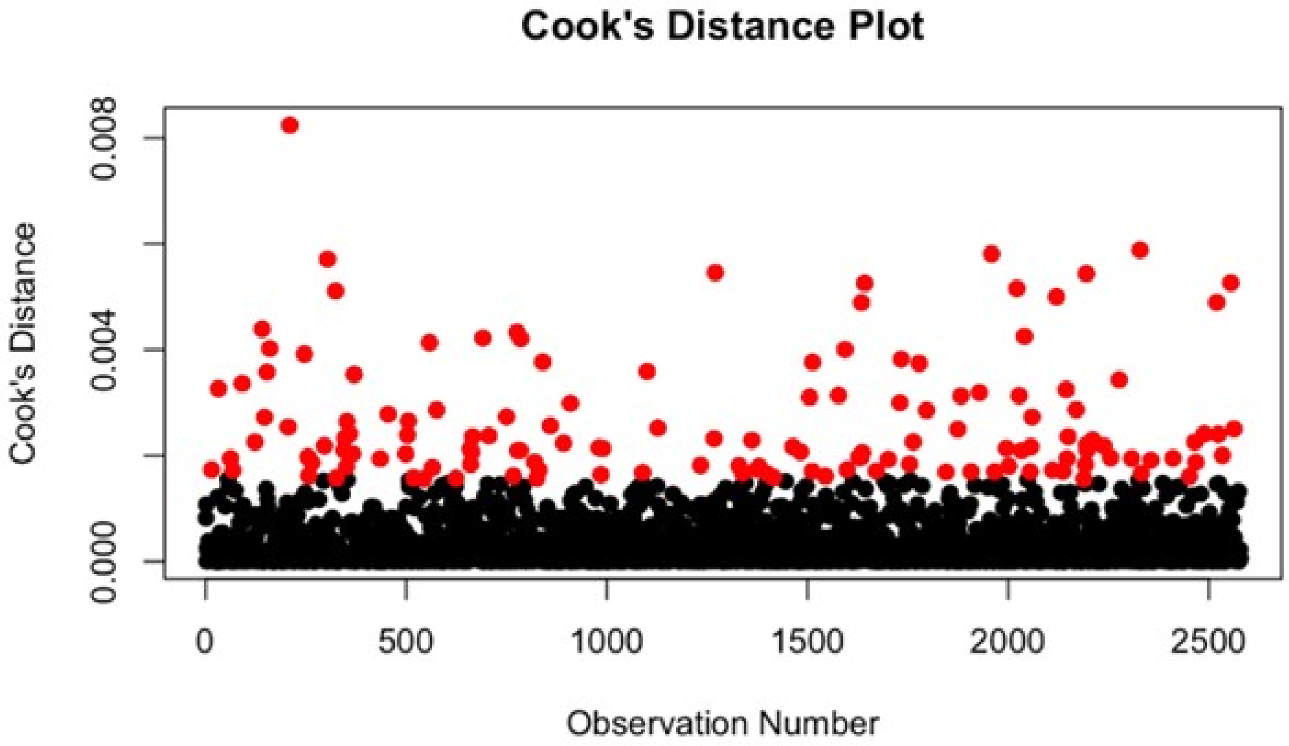


2. Project Flow – Evaluation

Outliers

- Cook's Distance identifies 147 outliers, primarily from the African region

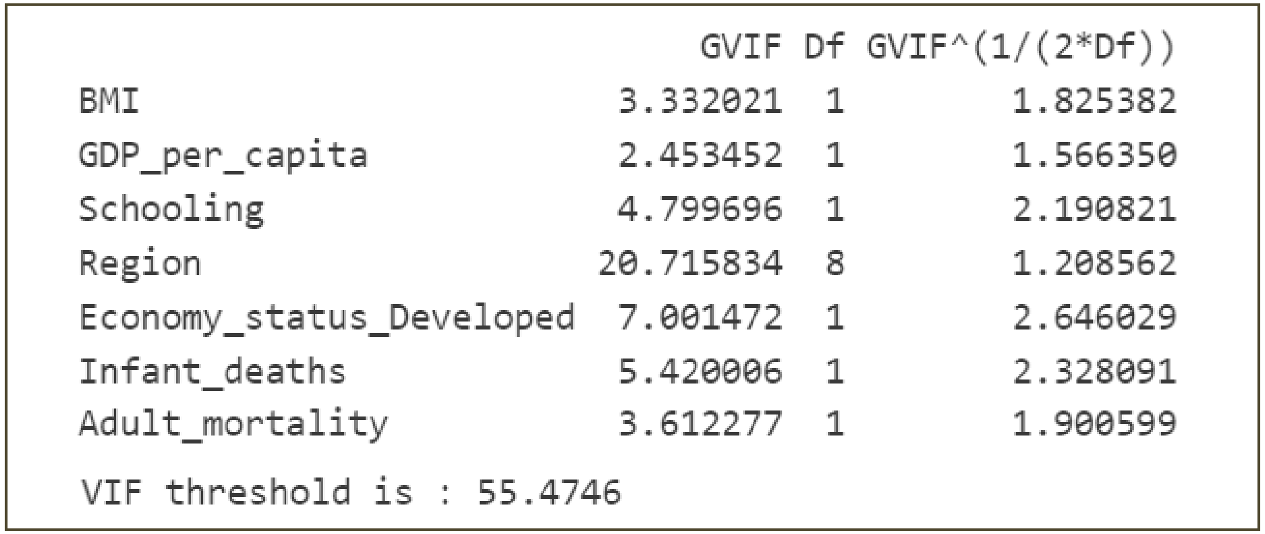
- However, these outliers were not removed
    - Interpret them as the life expectancy of underdeveloped countries
    - Are influential in understanding life expectancy variations
    - Remove them may lead to a bias in the result



2. Project Flow –Evaluation

Multicollinearity:

- VIF values
  - All predictors look good < 55.47
  - **No multicollinearity concerns**

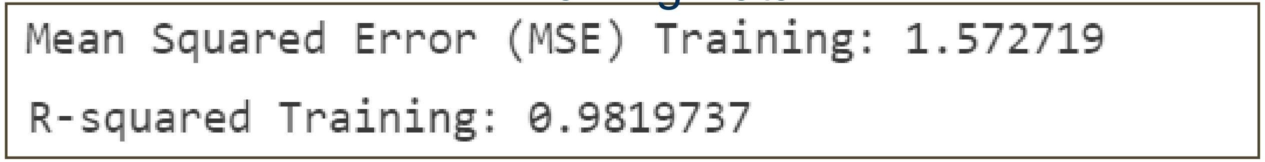


```
                                GVIF Df GVIF^(1/(2*Df))
BMI                         3.332021  1        1.825382
GDP_per_capita              2.453452  1        1.566350
Schooling                   4.799696  1        2.190821
Region                     20.715834  8        1.208562
Economy_status_Developed    7.001472  1        2.646029
Infant_deaths               5.420006  1        2.328091
Adult_mortality             3.612277  1        1.900599
VIF threshold is : 55.4746
```

Model Performance:

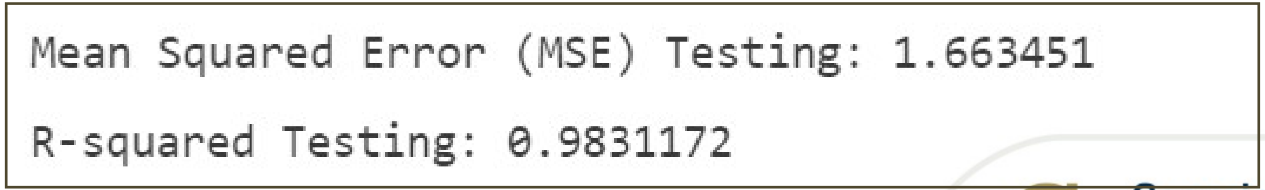
- Note: 90:10 split of Training and Validation(testing) data
- Low Mean Squared Error (MSE)
- High R-squared
- **The model is performing well**

Training Data



```
Mean Squared Error (MSE) Training: 1.572719
R-squared Training: 0.9819737
```

Testing Data



```
Mean Squared Error (MSE) Testing: 1.663451
R-squared Testing: 0.9831172
```

3. Findings and Interpretation

Results:

- Life expectancy is significantly influenced by a combination of factors including:
  - Adult Mortality
  - Infant deaths
  - Regional Differences
  - Economic Status
  - GDP_per_capita
  - Schooling Levels
  - BMI

- Model has a high R-squared value (0.982)
  - Suggests that these variables collectively offer a robust predictive power for life expectancy

Suggestions and Applications

Suggestions and Applications

Public Health Policies:

- Correlation between life expectancy and factors (adult mortality, infant deaths, and BMI) can inform public health strategies.

Educational Initiatives:

- Impact of schooling on life expectancy - investments in education could be a strategic approach to enhance public health outcomes.

Economic Development:

- Significant role of GDP per capita - economic growth and stability can positively affect life expectancy.

Suggestions and Applications

Suggestions and **Applications**

Regional Health Programs:

- Tailor health programs to achieve more effective outcomes based on area characteristics and needs identified in the analysis.

Targeted Interventions:

- Prioritize interventions focusing on reducing adult and infant mortality rates in regions with lower life expectancy.

Future Work

The validation results of the model on the test dataset (low MSE values and high R-squared values) confirm the **predictive accuracy and reliability of the model**.

However, future studies could explore the **inclusion of additional variables**, **such as environmental factors or genetic predispositions**, to further refine the understanding of the determinants of life expectancy, and to give valid recommendations and case applications that take more factors into account.

Thank you for listening!

Georgia Tech