Life Expectancy Data:

Analysis and Evaluation

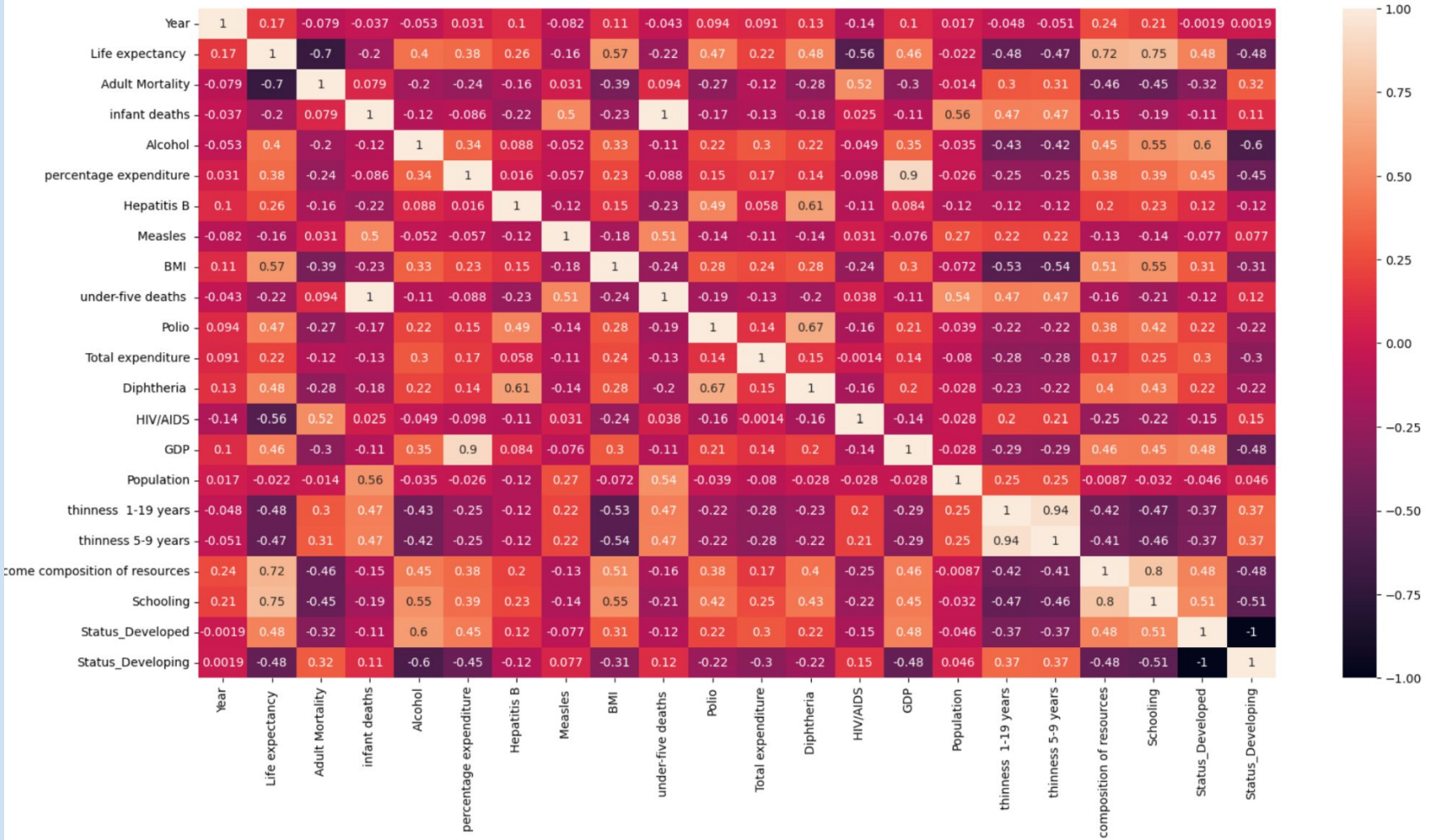
Harsh Panwar, Carlos Bonilla, Talha Choudhry, Nishtha Dandriyal
CS301

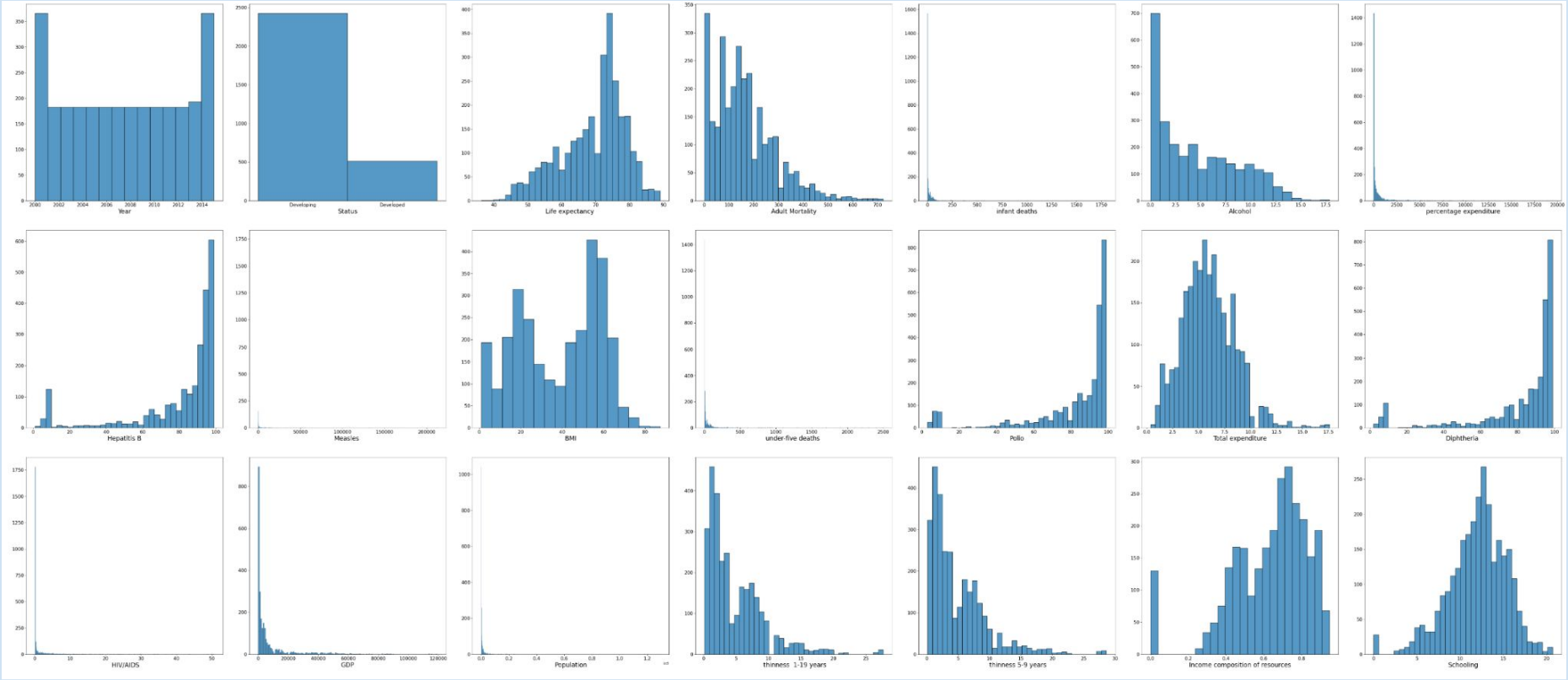
Data Introduction + Pre-processing

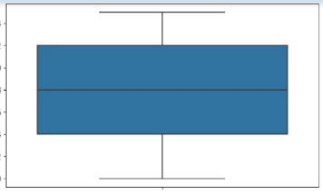
- Kaggle — sourced from World Health Organization
- 22 features
  - 20 Numerical
  - 2 Categorical
- One hot encoding for categorical features
- Dropped rows with any null values prior to training

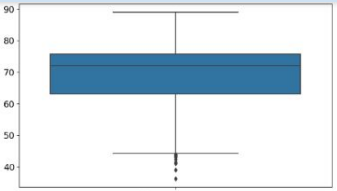
Dataset Analysis

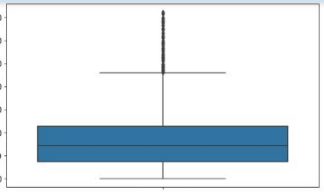
Important Features	Significance
Adult Mortality	Probability of dying between 15 and 60 years per 1000 population
Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
BMI	Average Body Mass Index of the entire population
HIV/AIDS	Deaths per 1,000 live births HIV/AIDS (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Income	Average amount of money earned
Schooling	Sum of the age specific enrollment rates for levels of education
Development status	Developed or developing?
Thinness	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)

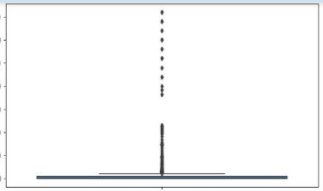

Year 1 0.17 -0.079 -0.037 -0.053 0.031 0.1 -0.082 0.11 -0.043 0.094 0.091 0.13 -0.14 0.1 0.017 -0.048 -0.051 0.24 0.21 -0.0019 0.0019
Life expectancy 0.17 1 -0.7 -0.2 0.4 0.38 0.26 -0.16 0.57 -0.22 0.47 0.22 0.48 -0.56 0.46 -0.022 -0.48 -0.47 0.72 0.75 0.48 -0.48
Adult Mortality -0.079 -0.7 1 0.079 -0.2 -0.24 -0.16 0.031 -0.39 0.094 -0.27 -0.12 -0.28 0.52 -0.3 -0.014 0.3 0.31 -0.46 -0.45 -0.32 0.32
infant deaths -0.037 -0.2 0.079 1 -0.12 -0.086 -0.22 0.5 -0.23 1 -0.17 -0.13 -0.18 0.025 -0.11 0.56 0.47 0.47 -0.15 -0.19 -0.11 0.11
Alcohol -0.053 0.4 -0.2 -0.12 1 0.34 0.088 -0.052 0.33 -0.11 0.22 0.3 0.22 -0.049 0.35 -0.035 -0.43 -0.42 0.45 0.55 0.6 -0.6
percentage expenditure 0.031 0.38 -0.24 -0.086 0.34 1 0.016 -0.057 0.23 -0.088 0.15 0.17 0.14 -0.098 0.9 -0.026 -0.25 -0.25 0.38 0.39 0.45 -0.45
Hepatitis B 0.1 0.26 -0.16 -0.22 0.088 0.016 1 -0.12 0.15 -0.23 0.49 0.058 0.61 -0.11 0.084 -0.12 -0.12 -0.12 0.2 0.23 0.12 -0.12
Measles -0.082 -0.16 0.031 0.5 -0.052 -0.057 -0.12 1 -0.18 0.51 -0.14 -0.11 -0.14 0.031 -0.076 0.27 0.22 0.22 -0.13 -0.14 -0.077 0.077
BMI 0.11 0.57 -0.39 -0.23 0.33 0.23 0.15 -0.18 1 -0.24 0.28 0.24 0.28 -0.24 0.3 -0.072 -0.53 -0.54 0.51 0.55 0.31 -0.31
under-five deaths -0.043 -0.22 0.094 1 -0.11 -0.088 -0.23 0.51 -0.24 1 -0.19 -0.13 -0.2 0.038 -0.11 0.54 0.47 0.47 -0.16 -0.21 -0.12 0.12
Polio 0.094 0.47 -0.27 -0.17 0.22 0.15 0.49 -0.14 0.28 -0.19 1 0.14 0.67 -0.16 0.21 -0.039 -0.22 -0.22 0.38 0.42 0.22 -0.22
Total expenditure 0.091 0.22 -0.12 -0.13 0.3 0.17 0.058 -0.11 0.24 -0.13 0.14 1 0.15 -0.0014 0.14 -0.08 -0.28 -0.28 0.17 0.25 0.3 -0.3
Diphtheria 0.13 0.48 -0.28 -0.18 0.22 0.14 0.61 -0.14 0.28 -0.2 0.67 0.15 1 -0.16 0.2 -0.028 -0.23 -0.22 0.4 0.43 0.22 -0.22
HIV/AIDS -0.14 -0.56 0.52 0.025 -0.049 -0.098 -0.11 0.031 -0.24 0.038 -0.16 -0.0014 -0.16 1 -0.14 -0.028 0.2 0.21 -0.25 -0.22 -0.15 0.15
GDP 0.1 0.46 -0.3 -0.11 0.35 0.9 0.084 -0.076 0.3 -0.11 0.21 0.14 0.2 -0.14 1 -0.028 -0.29 -0.29 0.46 0.45 0.48 -0.48
Population 0.017 -0.022 -0.014 0.56 -0.035 -0.026 -0.12 0.27 -0.072 0.54 -0.039 -0.08 -0.028 -0.028 -0.028 1 0.25 0.25 -0.0087 -0.032 -0.046 0.046
thinness 1-19 years -0.048 -0.48 0.3 0.47 -0.43 -0.25 -0.12 0.22 -0.53 0.47 -0.22 -0.28 -0.23 0.2 -0.29 0.25 1 0.94 -0.42 -0.47 -0.37 0.37
thinness 5-9 years -0.051 -0.47 0.31 0.47 -0.42 -0.25 -0.12 0.22 -0.54 0.47 -0.22 -0.28 -0.22 0.21 -0.29 0.25 0.94 1 -0.41 -0.46 -0.37 0.37
come composition of resources 0.24 0.72 -0.46 -0.15 0.45 0.38 0.2 -0.13 0.51 -0.16 0.38 0.17 0.4 -0.25 0.46 -0.0087 -0.42 -0.41 1 0.8 0.48 -0.48
Schooling 0.21 0.75 -0.45 -0.19 0.55 0.39 0.23 -0.14 0.55 -0.21 0.42 0.25 0.43 -0.22 0.45 -0.032 -0.47 -0.46 0.8 1 0.51 -0.51
Status_Developed -0.0019 0.48 -0.32 -0.11 0.6 0.45 0.12 -0.077 0.31 -0.12 0.22 0.3 0.22 -0.15 0.48 -0.046 -0.37 -0.37 0.48 0.51 1 -1
Status_Developing 0.0019 -0.48 0.32 0.11 -0.6 -0.45 -0.12 0.077 -0.31 0.12 -0.22 -0.3 -0.22 0.15 -0.48 0.046 0.37 0.37 -0.48 -0.51 -1 1
Year
Life expectancy
Adult Mortality
infant deaths
Alcohol
percentage expenditure
Hepatitis B
Measles
BMI
under-five deaths
Polio
Total expenditure
Diphtheria
HIV/AIDS
GDP
Population
thinness 1-19 years
thinness 5-9 years
composition of resources
Schooling
Status_Developed
Status_Developing
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Year
Status
Developing
Developed
Life expectancy
Adult Mortality
infant deaths
Alcohol
percentage expenditure
Hepatitis B
Measles
BMI
under-five deaths
Polio
Total expenditure
Diphtheria
HIV/AIDS
GDP
Population
thinness 1-19 years
thinness 5-9 years
Income composition of resources
Schooling

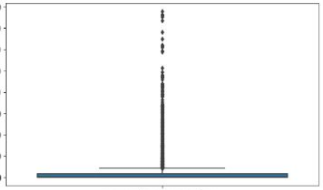

Year

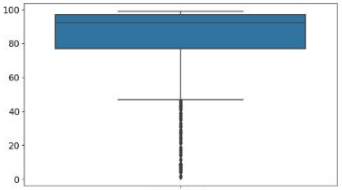

Life expectancy

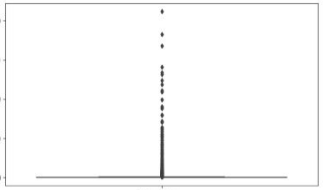

Adult Mortality

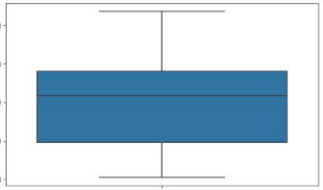

infant deaths

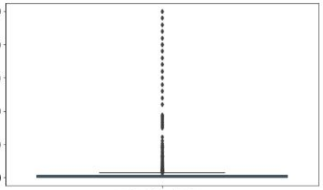

Alcohol

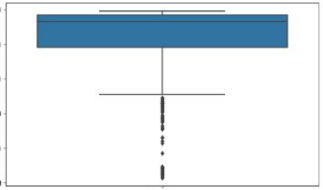

percentage expenditure

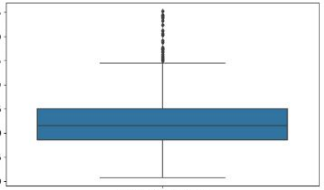

Hepatitis B

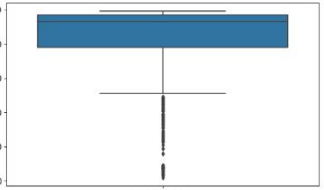

Measles

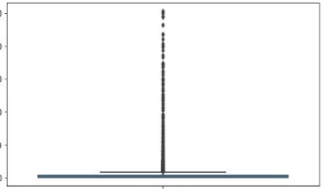

BMI

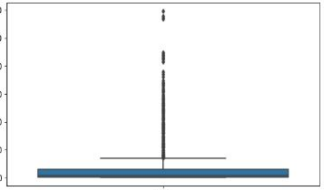

under-five deaths

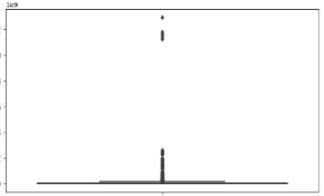

Polio

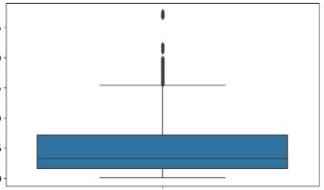

Total expenditure

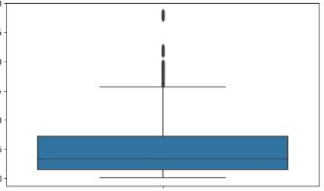

Diphtheria

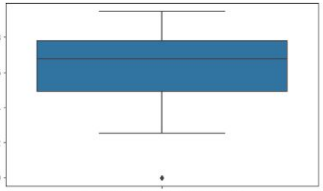

HIV/AIDS

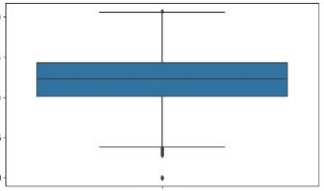

GDP


Population


thinness 1-19 years


thinness 5-9 years


Income composition of resources

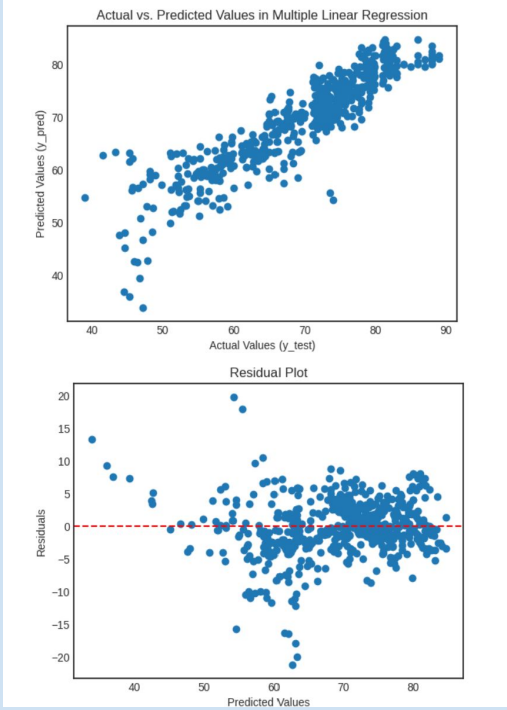

Schooling

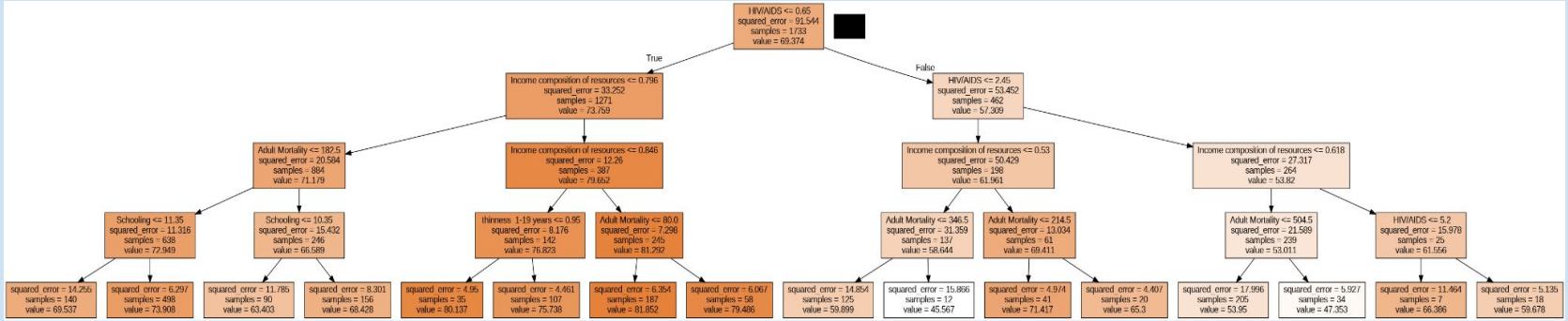
Training Models

- Multiple Linear Regression
  - Optimization: Gradient Descent
  - Cost Function: Mean Square Error
- Decision Tree
  - Max Depth = 4
  - Optimization: CART
  - Cost Function: Mean Square Error

Model Evaluation

- Multiple Linear Regression
  - Evaluation Metrics:
    - Mean Absolute Error: 3.1173210283190875
    - Mean Squared Error: 18.682813144063324
    - Root Mean Squared Error: 4.322361986699324
    - R-squared: 0.8166738752230451
- Decision Tree
  - Evaluation Metrics:
    - Mean Absolute Error: 2.58380725154973
    - Mean Squared Error: 13.366842793673236
    - Root Mean Squared Error: 3.656069309199873
    - R-squared: 0.8688371247428789





Comparison and Conclusion

- Overall, decision tree had better evaluation scores
- All metrics (RMSE, MAE, MSE,R^2) were smaller in Decision Tree
- Important note:
  - Decision Tree will be restricted to the data within this range
  - To predict values outside of the range, MLR will be a better choice

Model + Dataset

Model:
https://colab.research.google.com/drive/1nUCQD8mjUMs-TKhQU4J8cs2tPx_UDEAA?usp=sharing

Dataset:
https://drive.google.com/file/d/1clYx7i3dMytWbhYU--g1TUiIO9akIHHg/view?usp=sharing