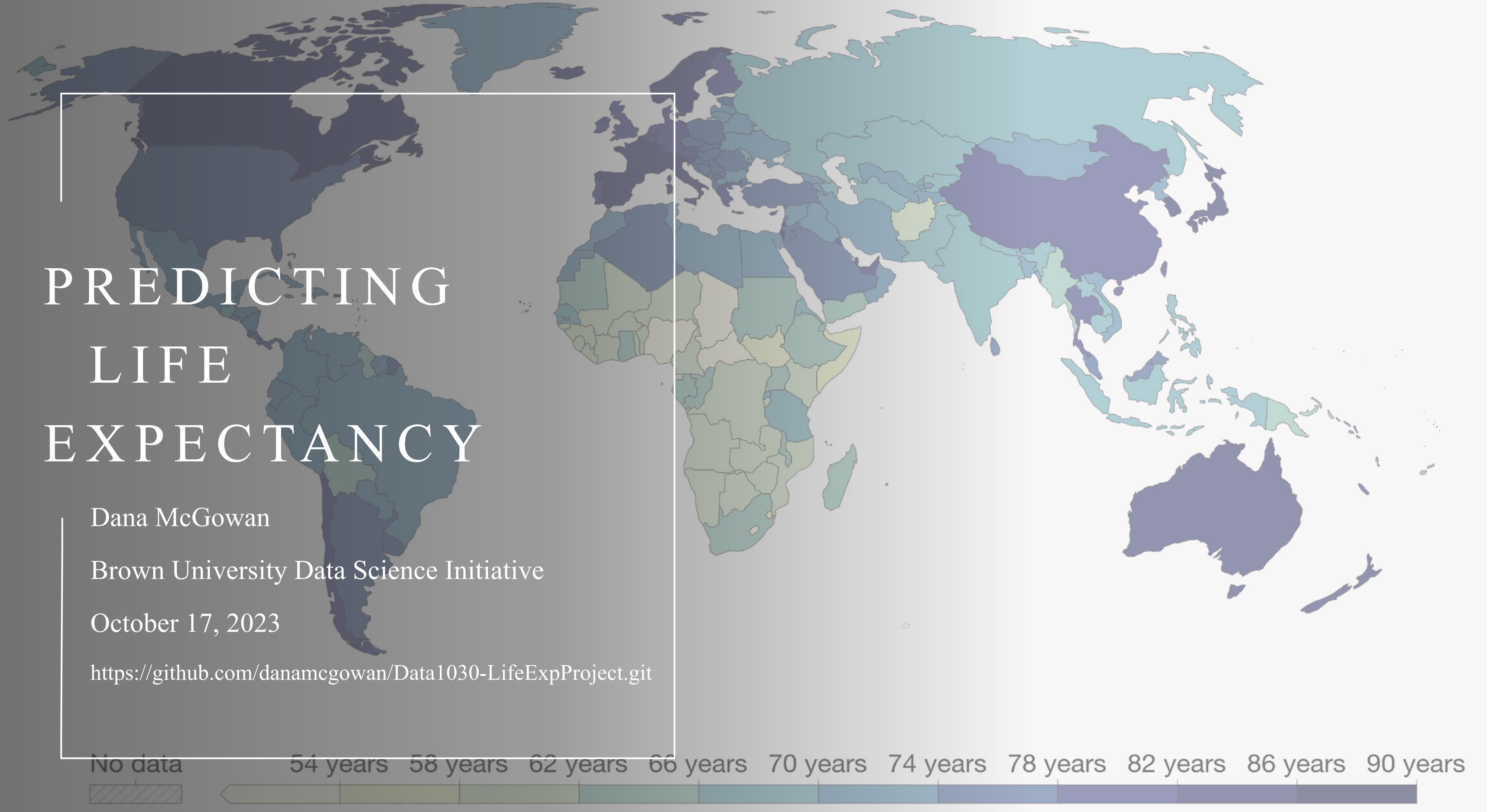
PREDICTING LIFE EXPECTANCY

Dana McGowan

Brown University Data Science Initiative

October 17, 2023

https://github.com/danamcgowan/Data1030-LifeExpProject.git



Source: UN WPP (2022); Zijdeman et al. (2015); Riley (2005)

OurWorldInData.org/life-expectancy • CC BY

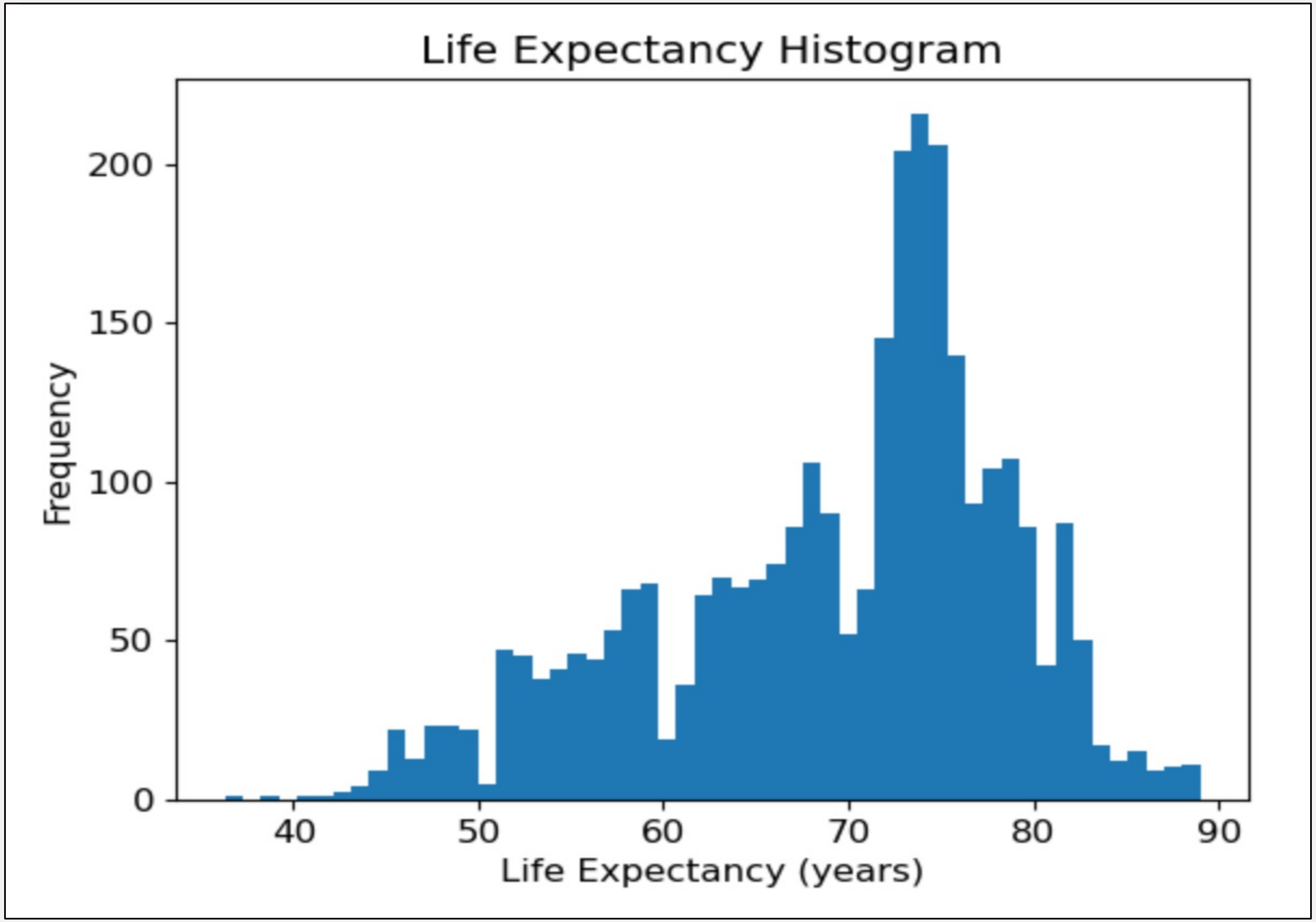
PROJECT INTRODUCTION

- The problem: Life expectancy has been studied immensely, but not heavily as it relates to factors that governments can affect.
  - Can the average life expectancy of a previously unseen country be predicted using immunization and human development index data?

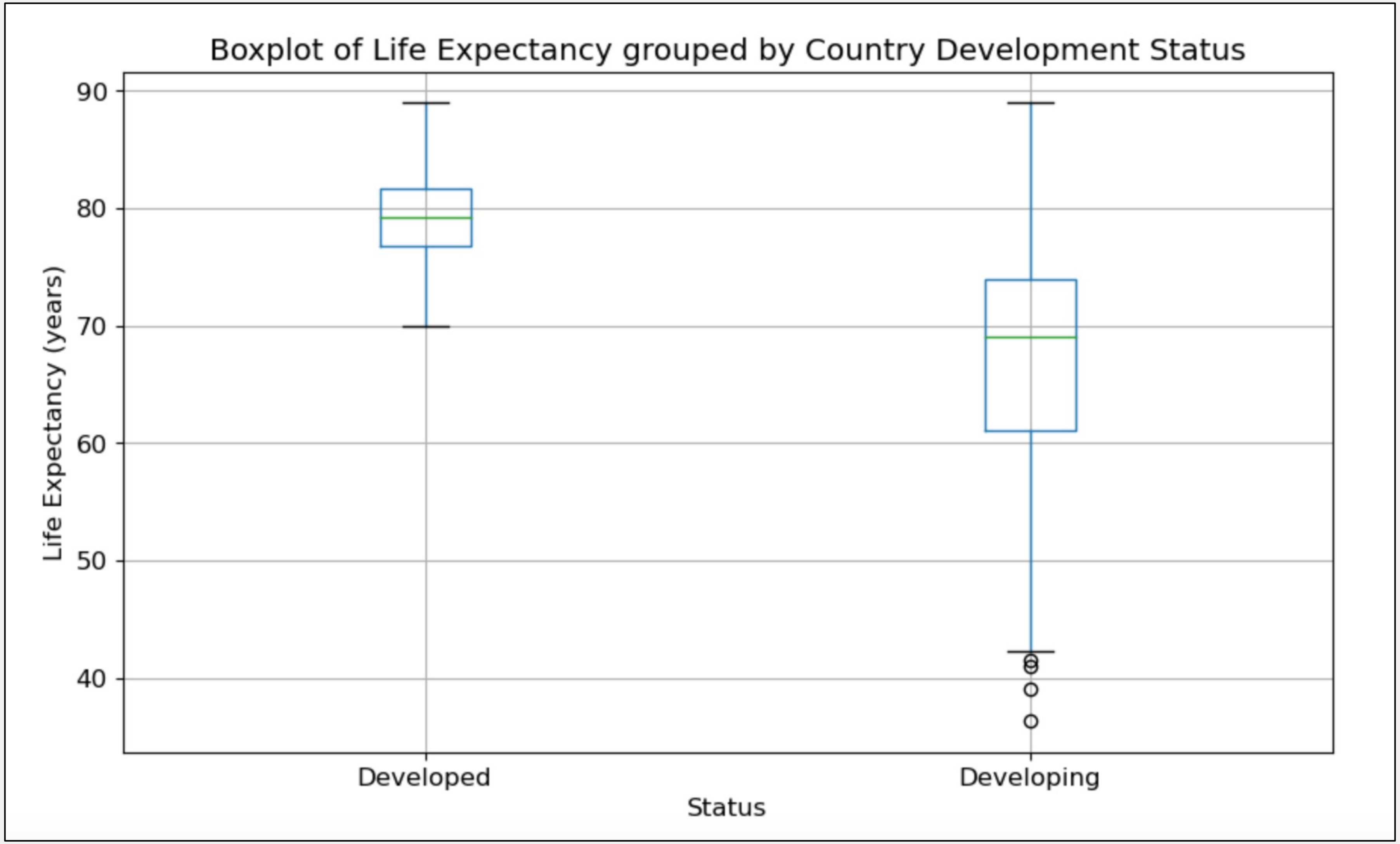
- Regression Problem

- Data
  - Pulled from Kaggle – Collected by World Health Organization (WHO) and United Nations Website
  - 2,938 data points and 22 features
    - 193 countries from 2000-2015

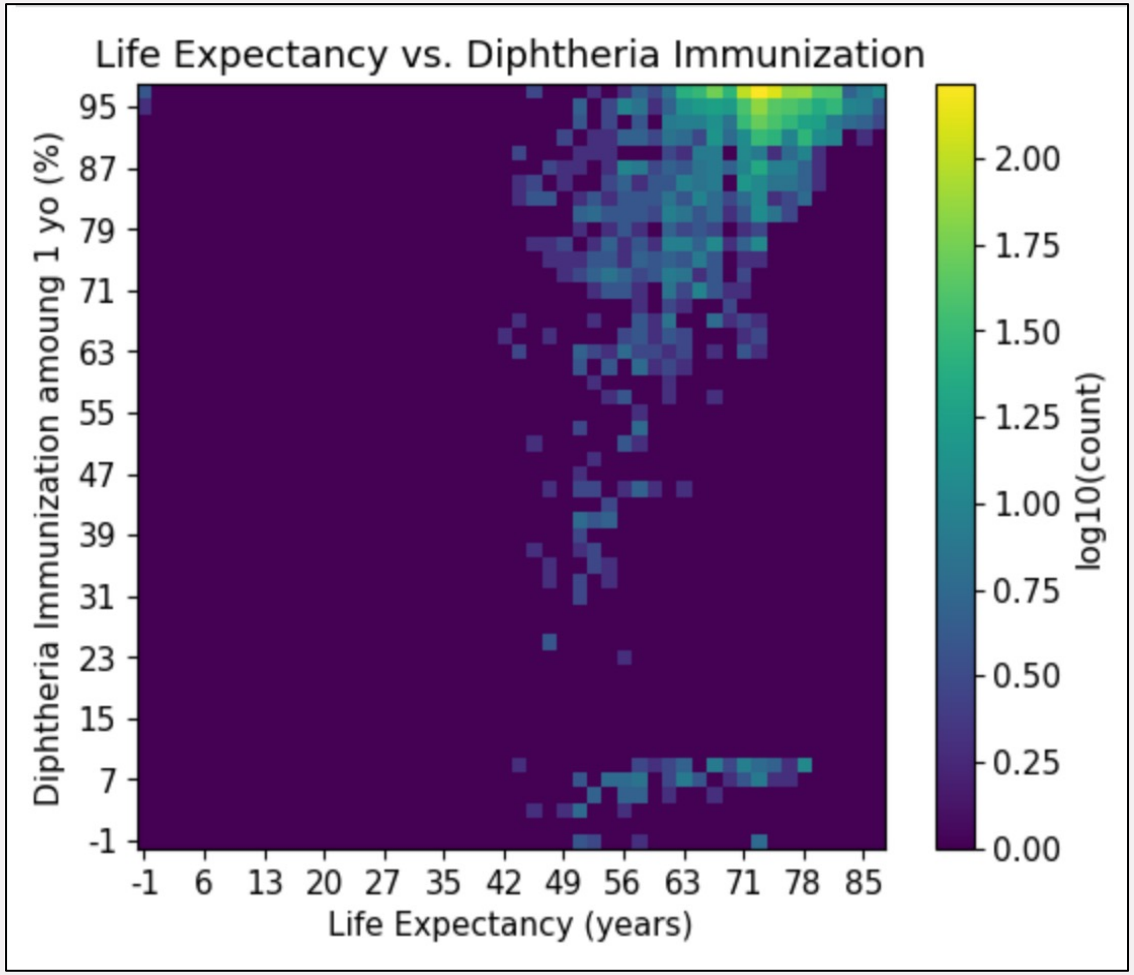
EXPLORATORY DATA ANALYSIS

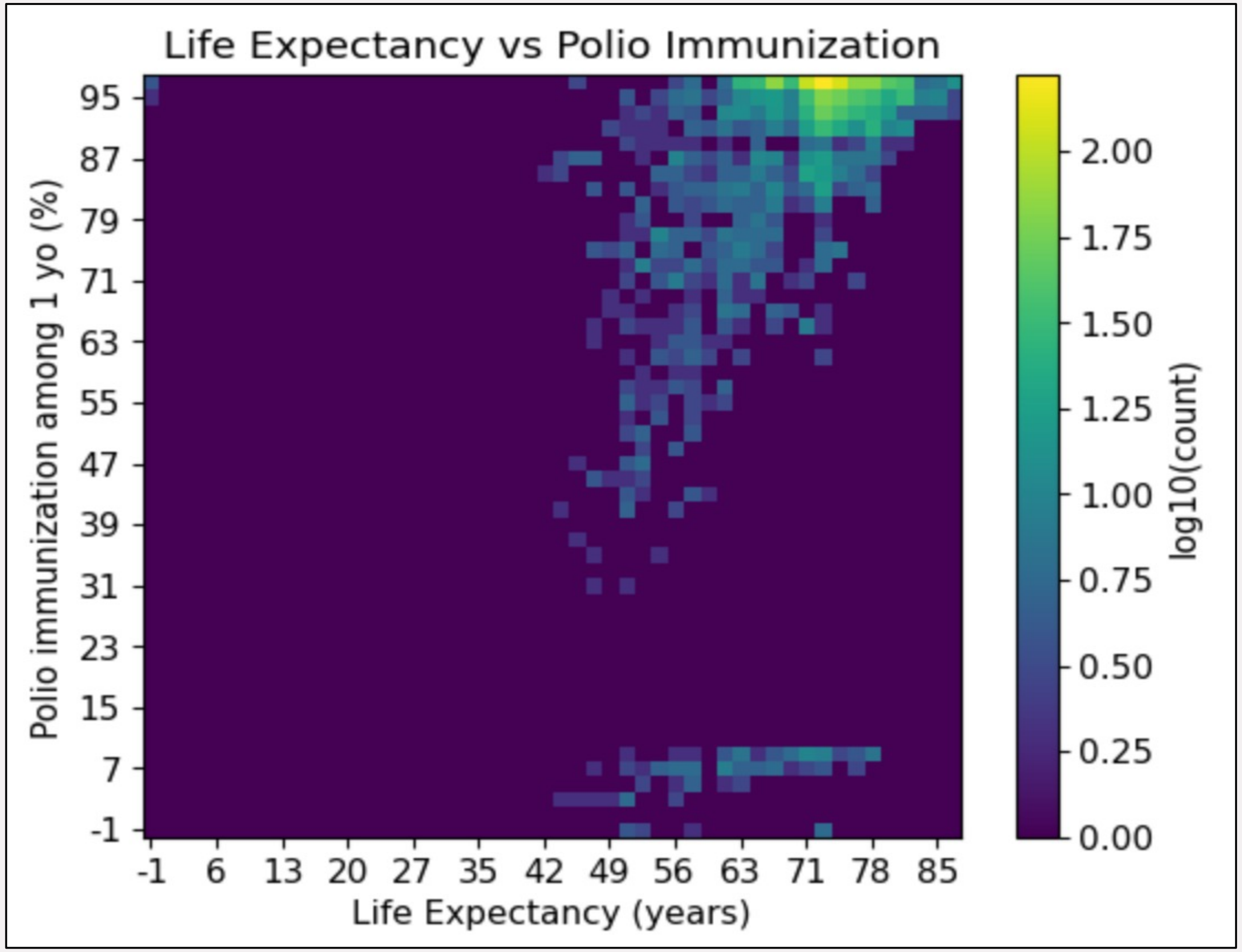


EXPLORATORY DATA ANALYSIS



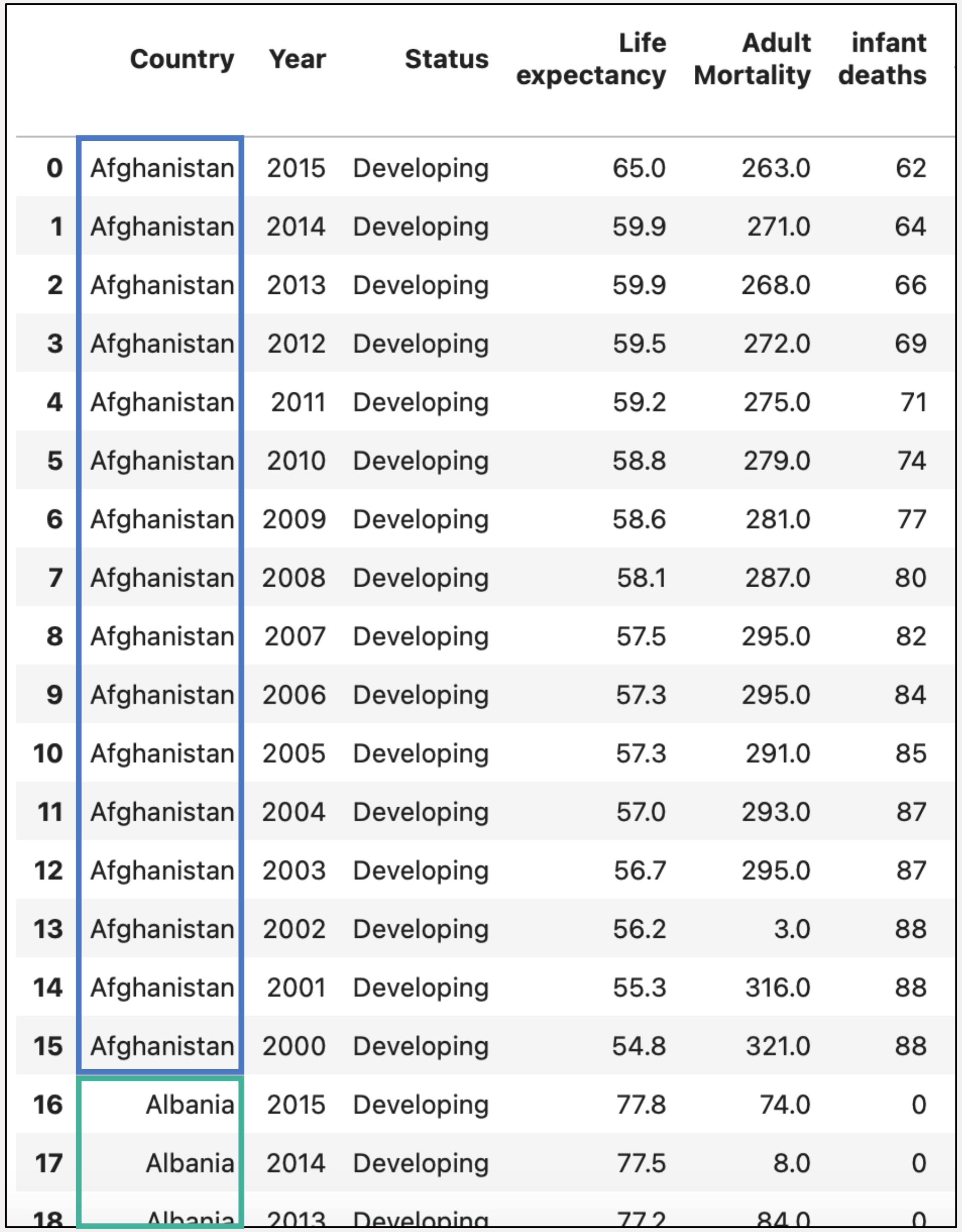
EXPLORATORY DATA ANALYSIS





SPLITTING DATA

- Group Shuffle Split
  - Training size set to 0.8
- Group K Fold
  - K = 4
- Final Groups:
  - Training Set – 1,764 rows and 21 columns
  - Validation Set – 580 rows and 21 columns
  - Test Set – 594 rows and 21 columns



	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths
0	Afghanistan	2015	Developing	65.0	263.0	62
1	Afghanistan	2014	Developing	59.9	271.0	64
2	Afghanistan	2013	Developing	59.9	268.0	66
3	Afghanistan	2012	Developing	59.5	272.0	69
4	Afghanistan	2011	Developing	59.2	275.0	71
5	Afghanistan	2010	Developing	58.8	279.0	74
6	Afghanistan	2009	Developing	58.6	281.0	77
7	Afghanistan	2008	Developing	58.1	287.0	80
8	Afghanistan	2007	Developing	57.5	295.0	82
9	Afghanistan	2006	Developing	57.3	295.0	84
10	Afghanistan	2005	Developing	57.3	291.0	85
11	Afghanistan	2004	Developing	57.0	293.0	87
12	Afghanistan	2003	Developing	56.7	295.0	87
13	Afghanistan	2002	Developing	56.2	3.0	88
14	Afghanistan	2001	Developing	55.3	316.0	88
15	Afghanistan	2000	Developing	54.8	321.0	88
16	Albania	2015	Developing	77.8	74.0	0
17	Albania	2014	Developing	77.5	8.0	0
18	Albania	2013	Developing	77.2	84.0	0

PREPROCESSING

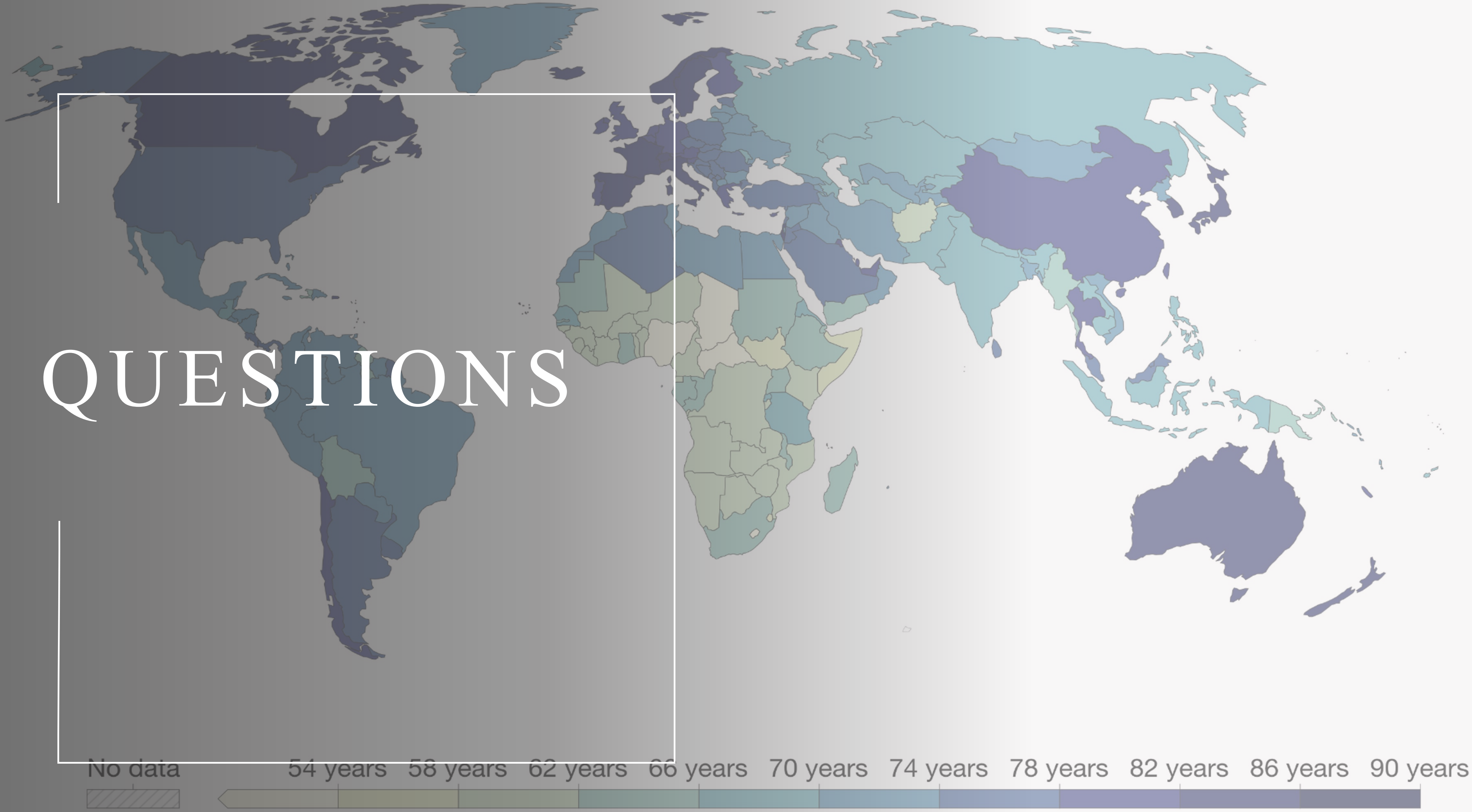
- Preprocessors
  - One Hot Encoder – Country and Status
  - Standard Scaler – all other features

- Shape of Data
  - Before Preprocessing:
    - Training Set: 1,764 rows and 21 columns
    - Validation Set - 580 rows and 21 columns
    - Test Set – 594 rows and 21 columns
  - After Preprocessing:
    - Training Set: 1,764 rows and 135 columns
    - Validation Set - 580 rows and 135 columns
    - Test Set – 594 rows and 135 columns

MISSING VALUES

- 43.9% of the points have missing values
- 63.6% of features have missing values. Features with missing values include:
    - Life expectancy
    - Adult Mortality
    - Alcohol
    - Hepatitis B
    - BMI
    - Polio
    - Total expenditure
    - Diphtheria
    - GDP
    - Population
    - Thinness 1-19 years
    - Thinness 5-9 years
    - Income composition of resources
    - Schooling

QUESTIONS



Source: UN WPP (2022); Zijdeman et al. (2015); Riley (2005)

OurWorldInData.org/life-expectancy • CC BY