Life Expectancy Analysis

Jaiden Atterbury, Tanner Huck

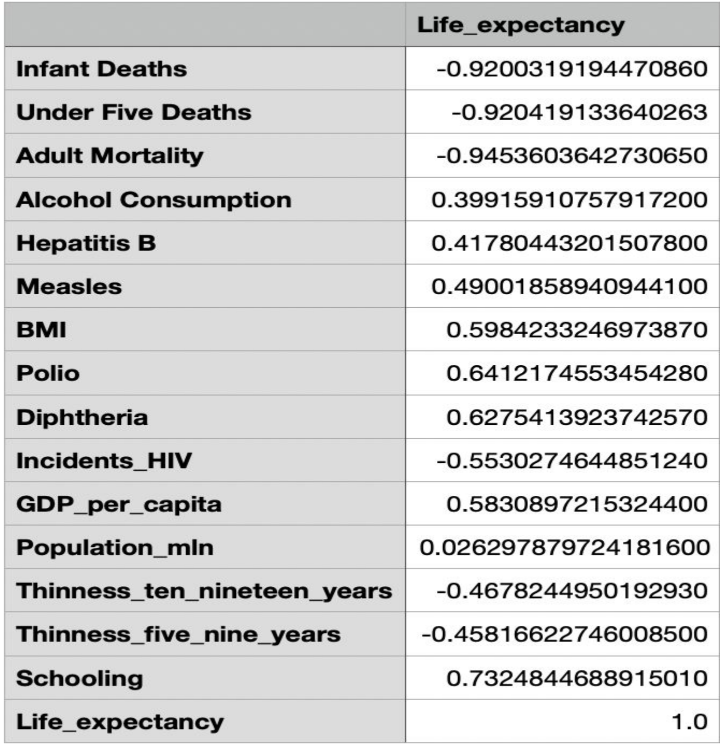
Motivation and background

Importance:

- Measure of number of years an individual is expected to life.
- Implications for public health and overall social well-being.
- Understanding geographical data and factors that impact life expectancy may help us understand how to increase life expectancy and give aid to nations with low expectancy.

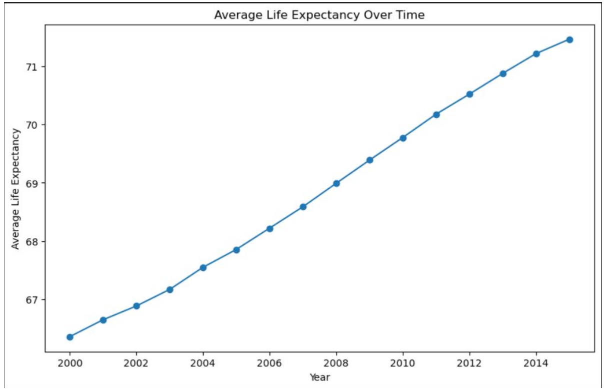
Data Exploration and Preprocessing

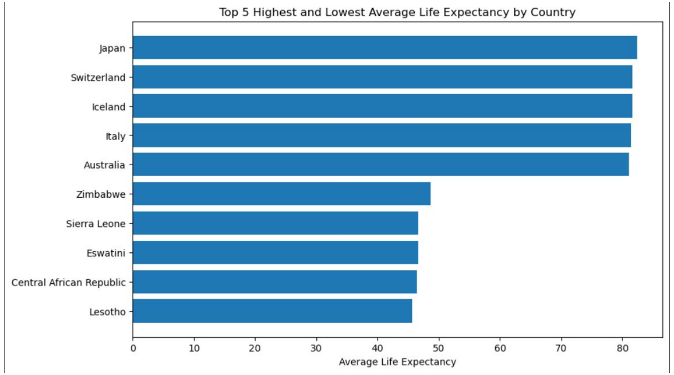
- Clean the data and handle any of the found missing values through the process of imputation or complete removal
- Must be careful when choosing to remove or fix any of the data set – missing data may provide valuable insights or tell a story about the data
- Gain insights into the distribution and relationships among certain variables of interest
- Make histograms, scatter plots, or other types of graphs to visually understand these relationships

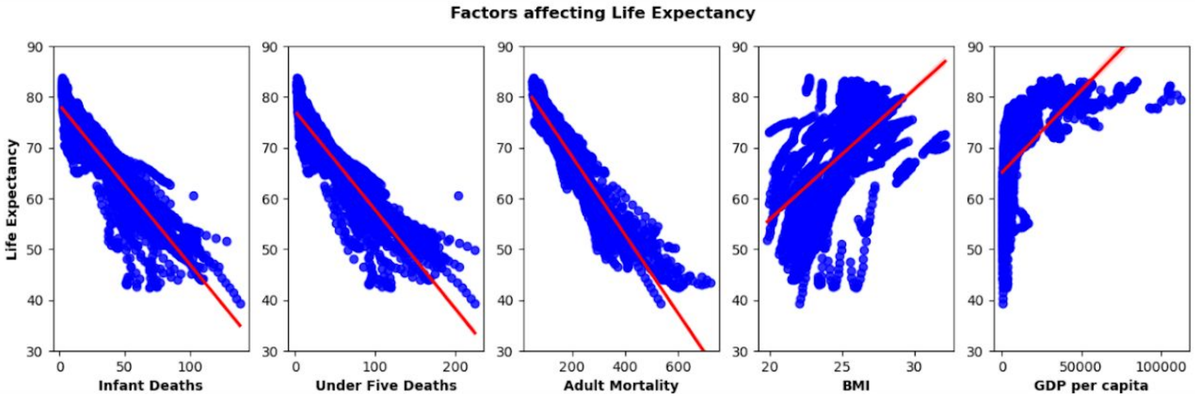


	Life_expectancy
Infant Deaths	-0.9200319194470860
Under Five Deaths	-0.920419133640263
Adult Mortality	-0.945360364273065O
Alcohol Consumption	0.399159107579172OO
Hepatitis B	0.41780443201507800
Measles	0.4900185894094410O
BMI	0.598423324697387O
Polio	0.641217455345428O
Diphtheria	0.627541392374257O
Incidents_HIV	-0.553027464485124O
GDP_per_capita	0.583089721532440O
Population_mln	0.0262978797241816OO
Thinness_ten_nineteen_years	-0.467824495019293O
Thinness_five_nine_years	-0.458166227460085OO
Schooling	0.732484468891501O
Life_expectancy	1.0

Data Exploration and Preprocessing





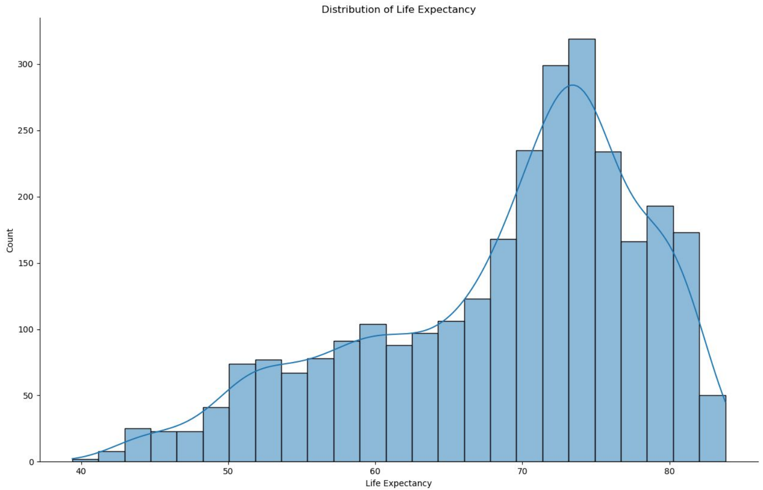


Research Question 1: Predicting Life Expectancy

- Which combination of variables creates the most significant/best model for predicting the life expectancy of a given country?

- **Methodology:**

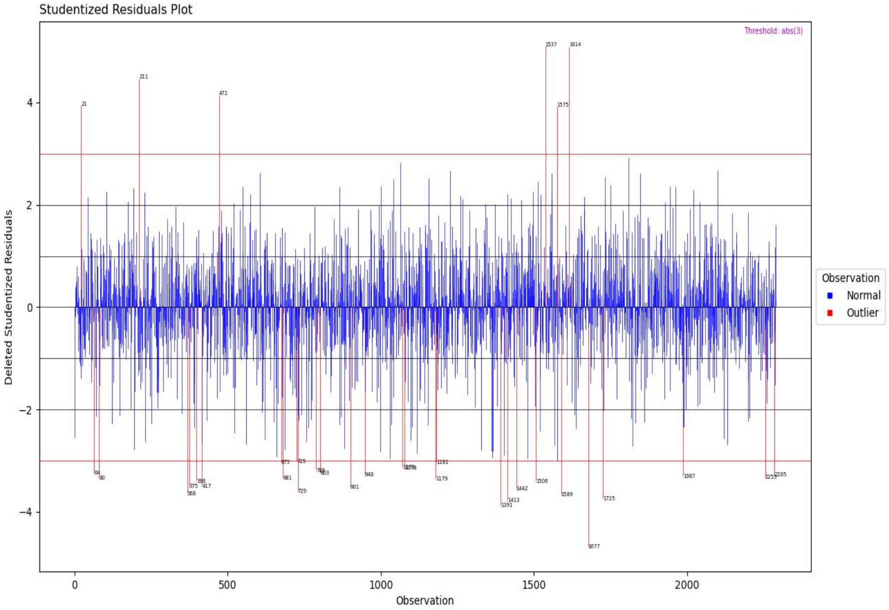
1. Find the distribution of life expectancy.
2. Make necessary transformations, based on 1.
3. Fit initial model to find significant predictors.
4. Check the VIF of the significant predictors.
5. Split data into training and test set, fit a new model.
6. Check model assumptions (linearity, normality, etc.).
7. Check the model accuracy.



Research Question 1: Predicting Life Expectancy

Results:

1. Distribution of life expectancy wasn't normal.
2. No transformation did the trick.
3. Significant predictors: Infant Deaths, Under Five Deaths, Adult Mortality, Alcohol Consumption, Hepatitis B, BMI, Incidents_HIV, GDP_per_capita, Thinness_ten_nineteen_years, and Schooling.
4. Calculating VIF dropped these down to 5 for the final model.
5. Final equation shown below.
6. 3 out of the 6 assumptions were violated.
7. Training set mean squared error was 5.222004318670649, and the testing mean squared error was 5.489100109185608.



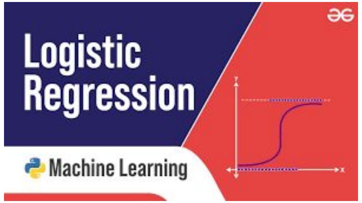
$$\boxed{\text{Life }\widehat{\text{Expectancy}} = 83.1 - 0.1 \cdot \text{Adult Mortality} + 0.3 \cdot \text{Alcohol Consumption} + 0.7 \cdot \text{Incidents\_HIV} + 0.00001 \cdot \text{GDP\_per\_capita} - 0.1 \cdot \text{Thinness\_ten nineteen\_years}}$$

Research Question 2: Classifying Development Status

- Which combination of variables creates the most significant/best model for classifying if a nation is developing or developed? Furthermore, in terms of accuracy, how does a logistic regression compare to a decision tree classifier when fit onto this data set?

- **Methodology:**

1. Build a logistic regression model to find significant predictors.
2. Drop all insignificant predictors and split the data into a training and test set.
3. Refit the logistic regression model onto the training set.
4. Find training and testing accuracy.
5. Fit a DecisionTreeClassifier onto the same data.
6. Find training and testing accuracy.
7. Compare the results.



Research Question 2: Classifying Development Status

- **Results:**

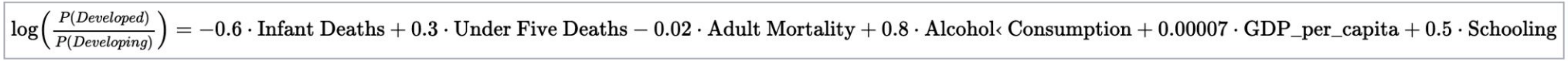
1. The significant predictors of the developmental status variable are: infant deaths, under five deaths, adult mortality, alcohol consumption, HIV incidents, GDP per capita, “thinness” of individuals aged 10 to 19 years old, and lastly schooling

2. Model equation shown below.

3. The logistic regression model has a training accuracy score of 97.4% and a testing accuracy score of 97.6%, while the DecisionTreeClassifier has a training accuracy score of 100% and a testing accuracy score of 98.1%.

4. Thus, the DecisionTreeClassifier is only slightly better than the logistic regression model for this specific split of training and testing data.

5. It is important to note that all models in this report would change depending on which data points made the training set versus the testing set and vice versa.

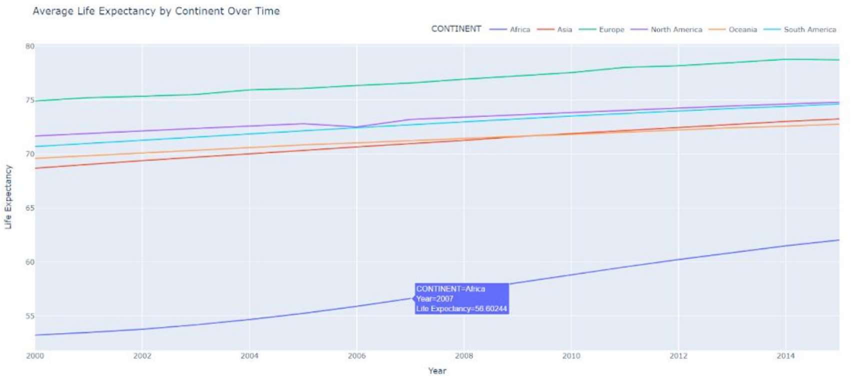


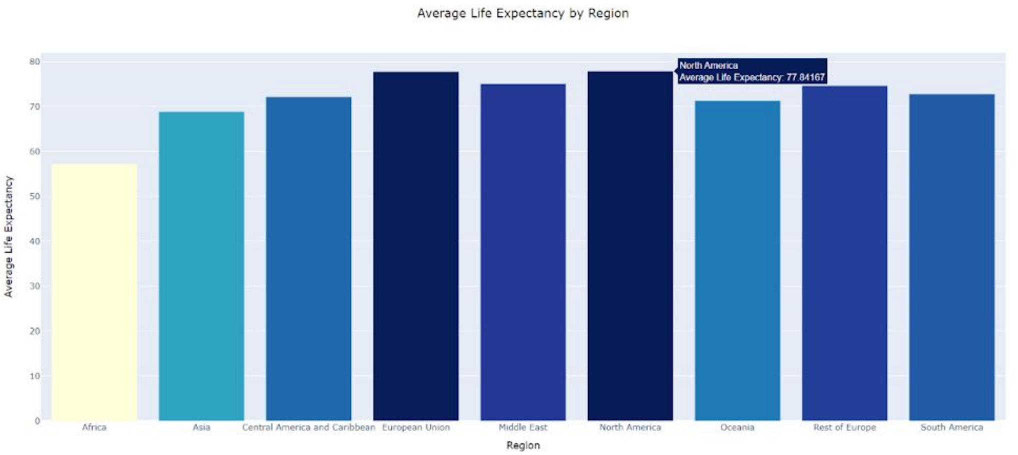
$$\log\left(\frac{P(Developed)}{P(Developing)}\right) = -0.6 \cdot \text{Infant Deaths} + 0.3 \cdot \text{Under Five Deaths} - 0.02 \cdot \text{Adult Mortality} + 0.8 \cdot \text{Alcohol‹ Consumption} + 0.00007 \cdot \text{GDP\_per\_capita} + 0.5 \cdot \text{Schooling}$$

Research Question 3: Temporal and Geographical Analysis

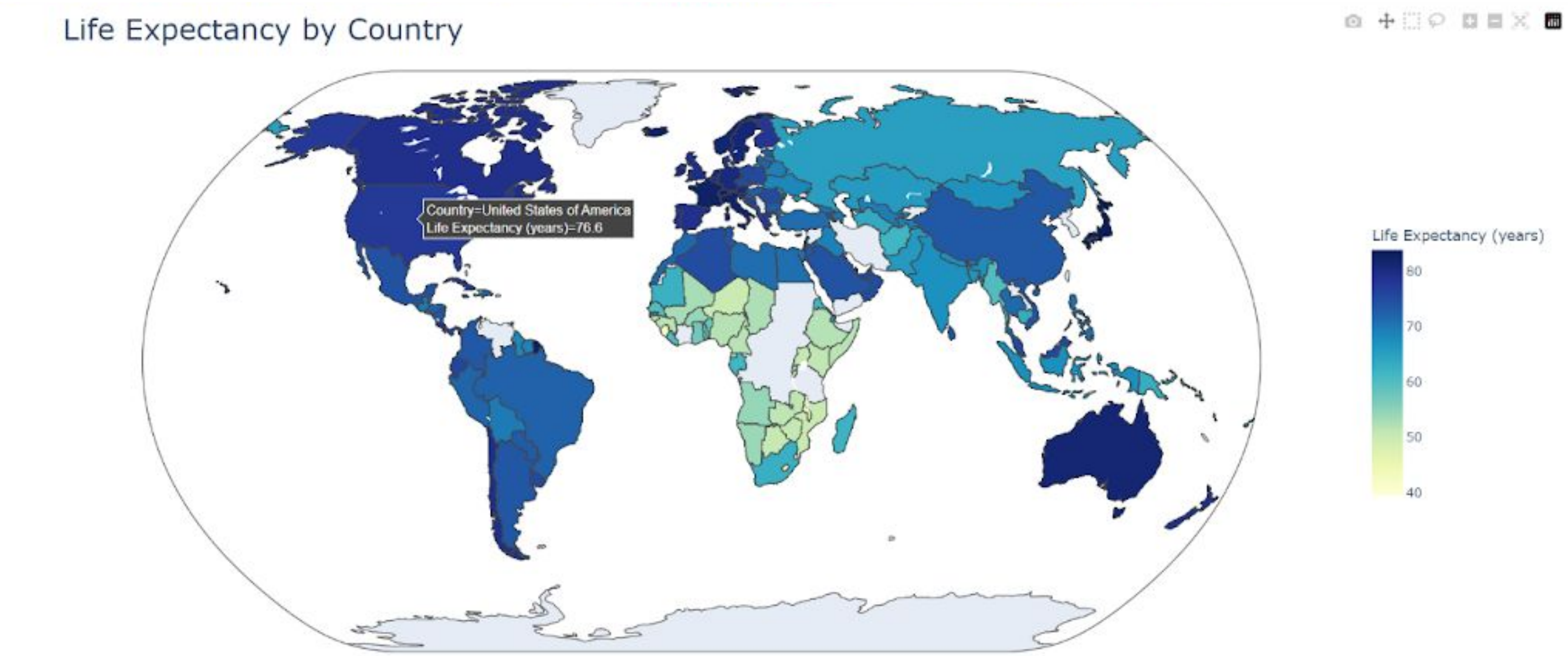
How does life expectancy vary across different areas of the world?

- Using different visualizations, we will analyze how life expectancy changes for different locations and identify regions where life expectancy and disease prevalence are most and least prevalent.





Research Question 3: Temporal and Geographical Analysis


Life Expectancy by Country
Country=United States of America
Life Expectancy (years)=76.6
Life Expectancy (years)
80
70
60
50
40

Impacts and Future work

- An important impact and one of our main motivations were the possible public health interventions that can be made with the aid of our analysis.
- Develop targeted public health interventions that may improve life expectancy in the areas that need help.
- Governments, healthcare organizations, and policymakers may use our findings to design strategies to improve health and extend life expectancy.
- Advocacy and awareness of life expectancy. This project may raise attention to different factors that reduce life expectancy and help people avoid them.

Works cited:

Here is a list of resources we used to write the code for this report, as well as some of the inspiration for the project in general:

- Plotly documentation: https://plotly.com/python/
- Plotly express documentation: https://plotly.com/python/plotly-express/
- Inspiration for making interactive cloropleth map using plotly express: https://stackoverflow.com/questions/75980836/i-made-a-plotly-express-choropleth-mapbox-of-us-zip-codes-can-i-add-a-choroplet
- Inspiration for making multiple scatter plots in a single figure: https://stackoverflow.com/questions/55126088/scatter-plot-grid-faceted-by-columns-in-matplotlib-or-seaborn
- Matplotlib documentation: https://matplotlib.org/stable/index.html
- Seaborn distplot documentation: https://seaborn.pydata.org/generated/seaborn.displot.html\
- Statsmodel documentation: https://www.statsmodels.org/stable/index.html
- Inspiration for making studentized residual plot: https://rpubs.com/pfr088883/1033107
- Rule of thumb for the Durbin-Watson test: https://www.statology.org/durbin-watson-test-python/