CSE163 Final Deliverable Report
Winter 2023

Exploring the impact of CO2 emissions, GDP, and health expenditure on individual life expectancy

Xinyu Chang, University of Washington,

Xinlong Chen, University of Washington,

Junhan Zhang, University of Washington,

Email: xchang3@uw.edu, xc78@uw.edu, junhaz2@uw.edu

Summary of questions and results

What are the trends and patterns in the distribution, and outliers of individual GDP/CO2 Emissions/Life Expectancy/Health Expenditure datasets across 261 countries from 2000 to 2019?

After analyzing the cleaned data, it was found that:

- Life expectancy, individual GDP, individual health expenditure, and individual CO2 emissions belong to skewed distributions.
- Outliers were explored using box plots and frequency distribution plots for life expectancy and personal health expenditures. Outliers for life expectancy were mainly in countries with poor medical and health conditions. Outliers for health expenditures were due to the high cost of living in developed countries with high GDP. After removing outliers, the most common life expectancy was in the 70s, and health expenditure was mainly concentrated in the 0-400 USD range.

Does an individual's health expenditure have an effect on their individual's life expectancy?

The linear regression model suggests that:

- The individual's health expenditure have an effect on their individual's life expectancy.There is a positive correlation between an individual's health expenditure and that individual's life expectancy.
- The coefficient of the linear regression is 0.0031, which indicates that an increase of one unit(current US) of health expenditure will make the individual have 0.0031 years of improvement in life expectancy.
- The intercept 66.8389 suggests that when people do not make any expenditure on their health condition, they are expected to live to about 66 years old.

How do personal GDP, GDP level and CO2 emissions impact each person's life expectancy?

The linear regression model suggests that:

- There is no clear relationship between life expectancy and lower GDP levels.
- A model built up with life expectancy and GDP suggests that there is a positive relationship between GDP, CO2 emissions, and life expectancy.
- The coefficient of GDP in this model is 0.0002, suggesting that an increase in one unit of GDP will increase 0.0002 years of life expectancy.

- The coefficient of CO2 emission is 0.4079, suggesting that an increase in one unit of CO2 emission will increase 0.4079 years of life expectancy.
- The intercept 65.1501 suggests that when the CO2 emissions and GDP are zero, people are expected to live to 65 years old.
- Higher levels of CO2 emissions and GDP are associated with increased life expectancy.
- Based on our models, health expenditure, carbon dioxide emissions, and GDP have a positive effect on life expectancy. Health expenditure is the main factor affecting life expectancy. Carbon dioxide is the second main factor affecting life expectancy. GDP has the least impact on life expectancy.

Which independent variables(GDP, CO2, and Health expenditure) has the greatest impact on an individual's life expectancy? How does changing the max depth of the decision tree affect its accuracy and feature importance?

The decision tree model suggests that:

- Individual health expenditure has the greatest impact on individual life expectancy, followed by individual CO2 emissions, and individual GDP has the least impact.
- Increasing the max depth of the decision tree model can improve its accuracy and feature importance, and the feature importance of each variable tends to be stable after a certain level of max depth.

Motivation

Understanding the relationship between an individual's health expenditure and their Gross Domestic Product (GDP) level can have important implications for both individual and national-level policy decisions. The relationship between economic growth, life expectancy, and CO2 emissions is of increasing interest, as concerns about environmental sustainability and public health continue to grow. By examining the relationship between these factors, we can gain insights into the impact of economic development and environmental sustainability on public health, and inform policy decisions that aim to promote both economic growth and improved health outcomes.

Predicting life expectancy based on factors such as CO2 emissions, GDP, and health expenditure is important for both individual and population health planning. By examining the relationship between these factors and life expectancy, we can better understand the impact of various policy interventions aimed at improving health outcomes, and make more informed decisions about resource allocation for health promotion and disease prevention initiatives. The use of the decision model to make these predictions can provide valuable insights into the interplay between these factors, and help us to develop more effective strategies for promoting health and improving quality of life.

Dataset

The following datasets were collected separately from the "The World Bank":

- **GDP per Capita(current US$):** https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

The "GDP per Capita (current US$)" dataset contains data from 1960 to 2021 on the gross domestic product (GDP) per capita in 266 different nations. The GDP per Capita (current US$) dataset offers details on a nation's economic health, which can have an effect on the population's health and life expectancy.

- **CO2 emissions(metrics tons per capita):** https://data.worldbank.org/indicator/EN.ATM.CO2E.PC

The "CO2 Emissions (metric tons per capita)" dataset offers details on the per-capita carbon dioxide emissions of 266 nations between 1990 and 2019. The CO2 Emissions (metric tons per capita) dataset offers details about an individual's level of greenhouse gas emissions, which may have an effect on the environment and general welfare.

- **Current health expenditure per capita(current US$):** https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD

The "Current Health Expenditure per Capita (current US$)" dataset provides information about the current health expenditure per capita of 266 countries in current US dollars from 2000 to 2019. The Current Health Expenditure per Capita (current US) dataset offers details on the funding devoted to healthcare at an individual level, which may have an effect on the residents' life expectancy.

- **Life expectancy at birth per capita, total (years)** https://data.worldbank.org/indicator/SP.DYN.LE00.IN

The "Life Expectancy at Birth per Capita, Total (years)" dataset provides information about the life expectancy at birth in 266 countries from 1960 to 2021. The outcome of interest in this research topic is the Life Expectancy at Birth per Capita, Total (years) dataset, which offers details on the typical number of years a person may expect to live at birth. By comparing the life expectancy statistics with the data from the other datasets, it is possible to assess the link between CO2 emissions, GDP, health expenditures, and life expectancy.

Method

- **Cleaning Data(See file1)**

  For cleaning the data, our team read and cleaned data from multiple CSV files, merged them, and then saved the cleaned data to a new CSV file. The cleaned data is structured as a Pandas DataFrame with columns for country name, year, life expectancy, GDP, health expenditure, and CO2 emissions.

To clean the data, we would need to follow these steps:

- Prepare the environment with the required dependencies, such as Pandas.
- Download the CSV files containing the data related to life expectancy, GDP, health expenditure, and CO2 emissions.
- Save the script as a Python file and run it in the command-line interface.
- The script will call the clean_data() function, which consists of two helper functions: load_data() and merge_data().
- The load_data() function takes the name of the CSV file and the name of the column containing the data values as input. It loads the data from the CSV file, reshapes it using Pandas, and returns a cleaned Pandas DataFrame object.
- The merge_data() function takes a list of Pandas DataFrame objects as input and merges them into a single DataFrame by joining on the country name, country code, and year columns.
- The clean_data() function calls the load_data() function four times with the CSV files containing the data related to life expectancy, GDP, health expenditure, and CO2 emissions. It then calls the merge_data() function to merge the resulting four DataFrames into a single DataFrame. Finally, it saves the cleaned data to a new CSV file called "clean_data.csv".
- After running the script, we can open the "clean_data.csv" file and analyze the cleaned data. We can use various Pandas functions to manipulate and summarize the data, such as describe(), groupby(), and pivot_table().

- **Analyzing Data(See file 2)**

For analyzing the data, our team is loading data from a CSV file and calling the necessary functions to plot histograms, normal distributions, boxplots, and histograms with outliers removed for specified columns in the Pandas DataFrame.

To analyze the data, we would need to follow these steps:

- Prepare the environment with the required dependencies, such as Pandas, NumPy, Matplotlib, and SciPy.
- Download the CSV file containing the cleaned data.
- Save the script as a Python file and run it in the command-line interface.
- The script will call the main() function, which loads the cleaned data from the CSV file into a Pandas DataFrame.

  - The plot_distributions() function loops over the specified columns in the DataFrame, plots histograms and normal distributions for each column and shows the resulting plots.
  - The no_outliers_plot_distribution() function takes a Pandas DataFrame and the name of a column as input. It removes the outliers from the column, calculates the mean and standard deviation of the filtered data, plots a histogram and normal distribution of the filtered data, and shows the resulting plots.
  - The plot_boxplot() function takes a Pandas DataFrame and the name of a column as input. It plots a boxplot of the column and shows the resulting plot.
  - After running the script, we can analyze the plots to answer various research questions related to the challenge goals of promoting global health and sustainable development.

- **Linear regression model(See file 3)**
  To create the linear regression model, we used the Statsmodel package, ols(Ordinary Least Squared) regression model to estimate the relationship between each dependent variable and independent variable. We also predicted the life expectancy under each dependent variable condition and calculated the root mean square error (RMSE) as well as adjusted root mean square error of each model to test our model. We used the test_train_split function in sklearn package to train and test our model.

  To create the regression models to make predictions and estimations, the following steps were implemented:
  - Load in cleaned data and read the data in each regression model
  - Create a new column to store the prior GDP and calculate the GDP growth rate for each country. Take a GDP growth rate higher than 0.02 to be a high GDP level annotated as 1 and a low GDP level(lower than 0.02) annotated as 0.
  - Train the linear regression models based on the independent variables and dependent variables stated in the question.
  - Predict each country's life expectancy based on dependent variables or variables
  - Use sklearn train_test_split function to split the dataset into 20% and 80% models.
  - Calculate and print the root mean squared error and adjusted root mean square error based on train and test data.

- **Decision Tree model(See file4)**
  To create the decision tree model, we used the Scikit-learn (sklearn) package, (Decision tree) classification model to estimate the relationship between each dependent variable and independent variable. We involve training and testing a decision tree model to

classify life expectancy into different age groups based on several independent variables, such as GDP, CO2 emissions, and health expenditure.

To create the decision tree models to make predictions and estimations, the following steps were implemented:

- Load the cleaned dataset from a CSV file using pandas.
- Classify the life expectancy column into different age groups using the classify_lifeexp function. This will create a new DataFrame with the LifeExp column classified into different age groups.
- Filter the data to include only years after 2010 using DataFrame filtering.
- Call the decision_tree function to train and test a decision tree model with the following arguments:
    a. The DataFrame contains the classified LifeExp column and independent variables.
    b. A list of independent variable names (GDP, CO2, and Health_Expenditure).
    c. The dependent variable name (LifeExp).
- The decision_tree function will perform the following steps:
    a. Split the data into training and test sets using the train_test_split function.
    b. Create two decision tree classifiers, one with a max depth of 3 and one without.
    c. Train the classifiers on the training data.
    d. Use the trained classifiers to predict the labels of the test data.
    e. Print the accuracy score of the predictions.
    f. Get the feature names and target names for plotting the decision trees and feature importance graphs.
    g. Plot the decision tree of the classifier with a max depth of 3 using the plot_tree function.
    h. Calculate and print the feature importance of both classifiers.
    i. Calculate the accuracy scores and feature importances of classifiers with different max depths.
    j. Plot the accuracy scores and feature importances using the plot_depth_score and plot_feature_importances functions, respectively.

- **Data Visualization(See file5)**

For the data visualization part, our team loads a clean data set from a CSV file and uses different functions to create scatter plots, country-specific scatter plots, and heatmaps to visualize correlations between columns in the pandas DataFrame.

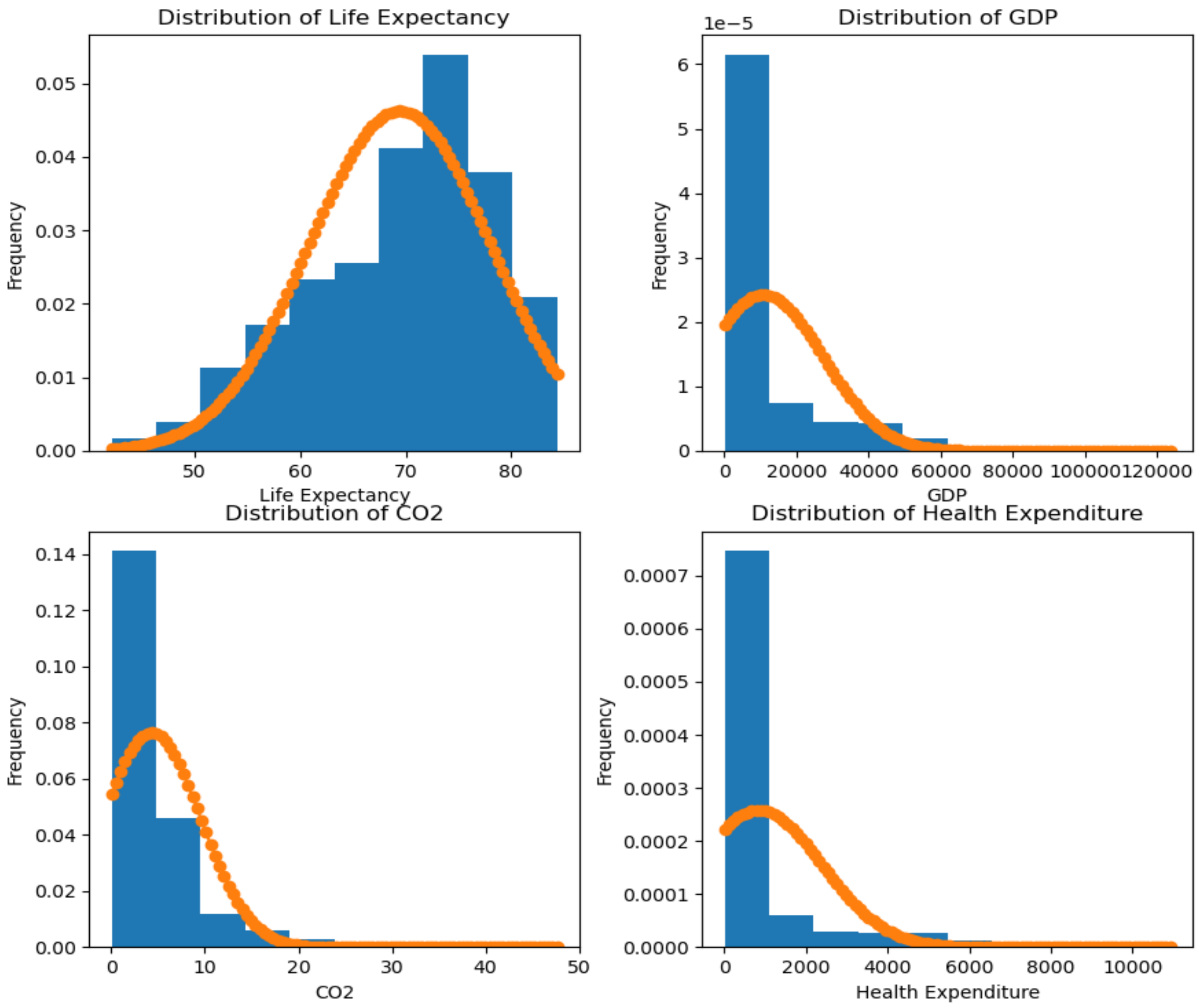
To visualize the data, we would need to follow these steps:

- Prepare the environment with the required dependencies, such as Pandas, Plotly, Seaborn, and Matplotlib.
- Load the data from the "clean_data.csv" file into a pandas DataFrame using the pd.read_csv() function.
- Call the create_scatter_plot() function with the DataFrame object and the x and y columns as arguments to create a scatter plot with a trendline for the specified x and y columns. The function uses the px. scatter() function from the plotly.express library to create the plot.
- Call the create_country_scatter_plot_to() function with the DataFrame object as an argument to create a scatter plot with trendlines for the Life Expectancy vs Health Expenditure relationship for four countries (United States, China, Germany, and South Africa) in the DataFrame. The function filters the DataFrame using the isin() method to select the rows corresponding to the four countries, and then uses the px.scatter() function from the plotly.express library to create the plot.
- Call the plot_corr_heatmap() function with the DataFrame object, a list of columns to drop (in this case, only the "Year" column), a flag indicating whether to display the correlation coefficients on the heatmap, the name of the matplotlib colormap to use, and the title of the heatmap as arguments to create a heatmap to visualize the correlation matrix of the specified columns in the DataFrame. The function uses the drop() method to drop the specified columns from the DataFrame, calculates the correlation matrix using the corr() method, and then uses the sns.heatmap() function from the seaborn library to create the heatmap.

<u>Results</u>

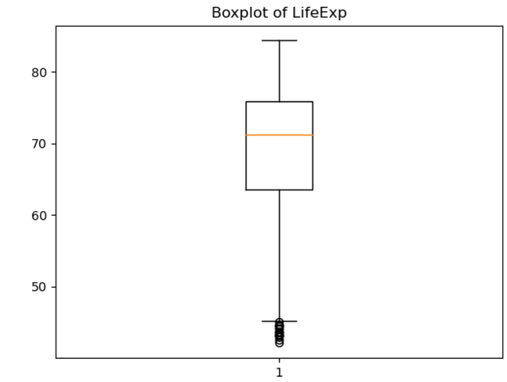
Research problem 1:What are the trends and patterns in the distribution and outliers of GDP/CO2/Life Expectancy/Health Expenditure datasets across 261 countries from 2000 to 2019?

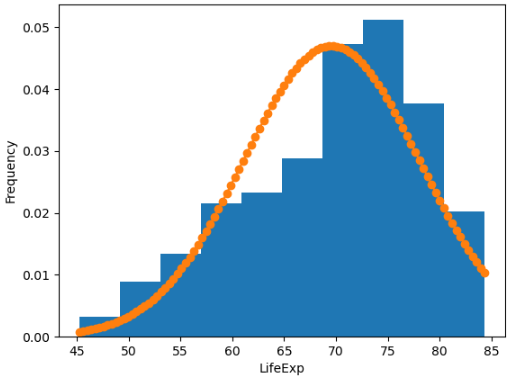
1.1 By using the frequency distribution graph, explore the distribution of individual life expectancy, individual GDP, individual health expenditure, and individual CO2 emissions data.



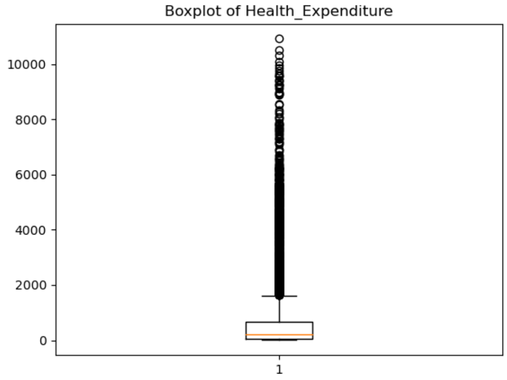
By analyzing the data after cleaning, we were surprised to find that life expectancy, individual GDP, individual health expenditure, and individual CO2 emissions all belong to skewed distribution. The data on life expectancy belongs to the left-skewed distribution, while individual GDP, individual health expenditure, and individual CO2 emission belong to the right-skewed distribution. According to the data on the world's average life expectancy (Earth-Place Explorer, n.d.), the world's average life expectancy is 72.27 years old. Compared with the data obtained through our study of data distribution, it shows that the world's average life expectancy has increased.

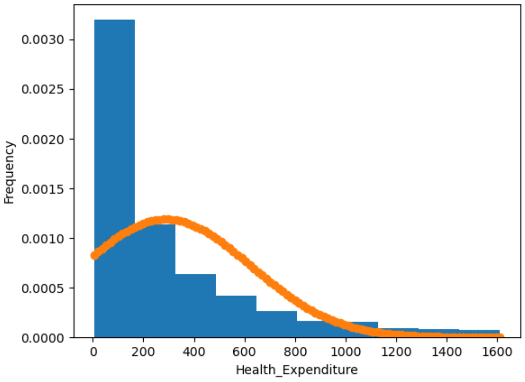
1.2 By using box plots and frequency distribution plots, explore outliers in life expectancy, personal health expenditures, and data distributions after outliers are removed.





For the research on life expectancy and health expenditure in the second question, we counted outliers in the data on the life cycle and health expenditure and analyzed the impact of outliers. Through the boxplot of life expectancy data, we found that the life expectancy data are mainly concentrated in the early 60s to the end of the 70s, and the outliers are concentrated in those under 50 years old. According to the research (Roser et al., n.d.), the backwardness of the country's development and poor medical and health conditions lead to low life expectancy. These outliers correspond to the low life expectancy of relatively backward countries in our database. By removing outliers and counting the distribution of life expectancy data, we found that life expectancy ranged from 45 to 85 years old, with the most common life expectancy in the 70s.



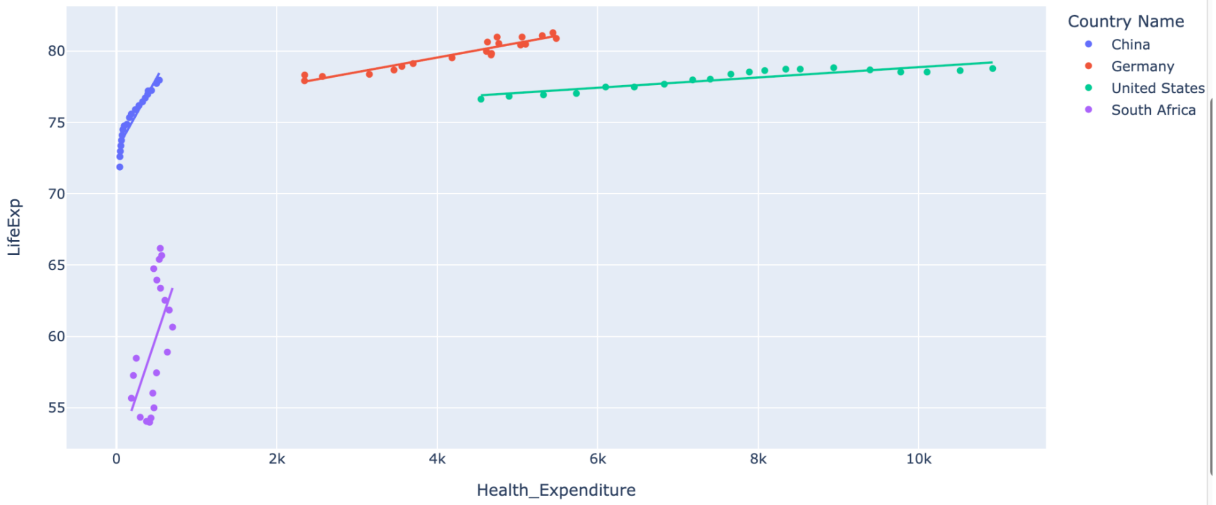


Through the box plot of the health expenditure data, we found that the health expenditure data is mainly concentrated in the range of 0 USD to 2000 USD, and a large number of outliers are above 2000 USD to 10000 USD. According to research (*Health at a Glance 2021: OECD Indicators*, 2021, #), national GDP is related to health expenditure, and the cost of living in developed countries with high GDP will far exceed that of underdeveloped countries. These outliers come from the high cost of living in the developed countries of the world, which reveals the inhomogeneity of world health development.

By removing outliers and counting the distribution of life expectancy data, we found that life expectancy ranges from 0 USD to 2000 USD and is mainly concentrated in the 0-400 USD range.

Research problem 2:Does an individual's health expenditure have an effect on their individual’s life expectancy?

2.1 By using the scatterplot, explore the relationship between health expenditure and life expectancy for four countries: the United States, China, Germany, and South Africa from 2000 to 2019.

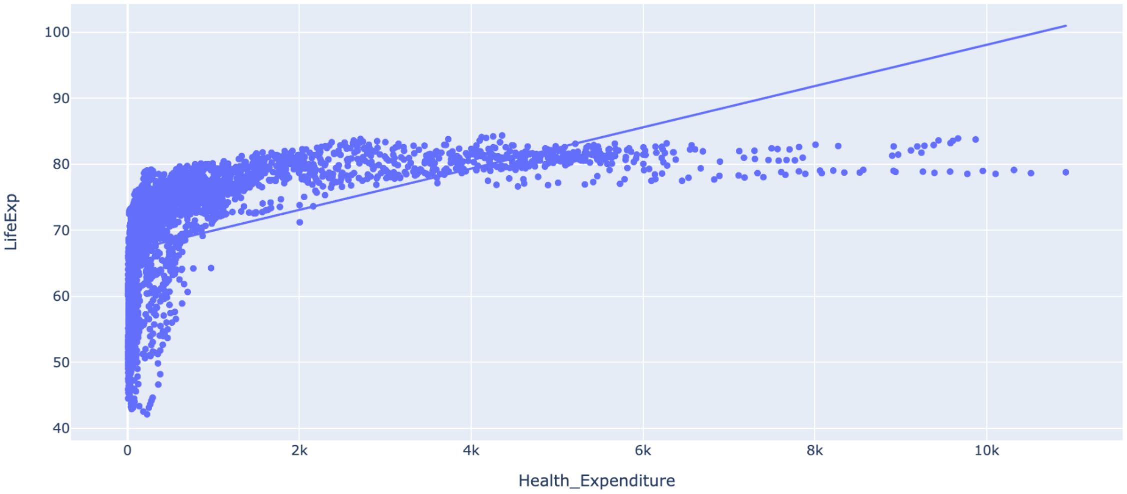


The relationship between health expenditure and life expectancy is a topic of interest to researchers, policymakers, and public health officials. The United States, China, Germany, and South Africa are all countries with different health systems and levels of economic development, making them potentially interesting case studies for investigating this relationship. A above scatterplot depicting the relationship between health expenditure and life expectancy for four representative nations (United States, China, Germany, and South Africa). A scatterplot of four countries on one graph enables analysis and comparison of the relationship between health expenditure and life expectancy in the four nations.

In all four nations, the scatter plot of health spending and life expectancy reveals a substantial positive association. This conclusion is consistent with prior research demonstrating a positive correlation between health expenditures and health outcomes, including life expectancy (Frenk et al., 2019). This conclusion emphasizes the necessity of prioritizing investments in healthcare to enhance health outcomes and extend life expectancy. In comparison to other nations, Germany has the greatest healthcare expenditures and life expectancy(about 76-83 years old), while South

Africa has the lowest(about 55-63 years old). Nevertheless, it is important to notice that scatterplots may not give adequate information to draw conclusions about the relationship between variables; further research, such as linear regression modeling, is necessary to validate and quantify the association.

2.2 By using the scatterplot, explore the relationship between health expenditure and life expectancy for 261 countries from 2000 to 2019.

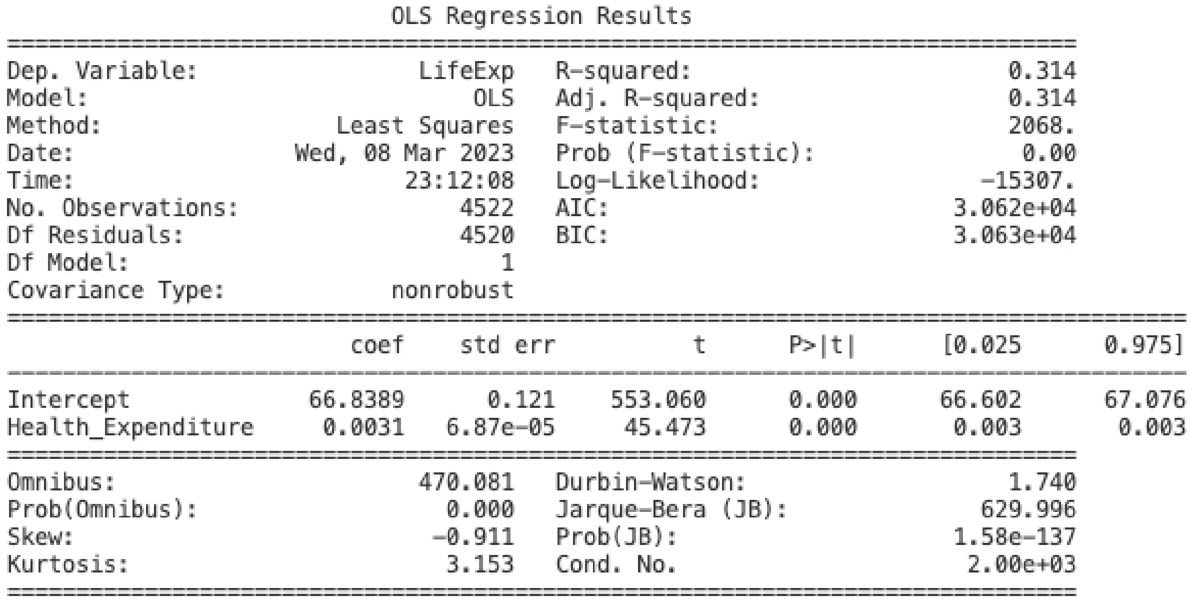


The above scatterplot between personal health expenditure and life expectancy demonstrates that there is no clear linear relationship between the two variables, a finding that differs from the data visualization in section 2.1. The main reason why this phenomenon may occur is that the research scope we have chosen is different. In 2.1, we focus on these four typical countries as research objects. However, in Section 2.2 we extended the scope of our research objects to a range of 261 countries. In addition, according to the previous analysis(See 1.2), we can conclude that there are some outliers in our data (The health expenditure of the developed countries is much higher than that of other developing countries), which causes the uneven distribution of the scatterplot. The scatterplot in this part indicates that there is no discernible pattern or trend based only on the scatterplot. There may be a correlation between personal health expenditures and life expectancy, but it may not be straightforward or clear. Also, the regression line is deviated and unrealistic. In the figure, when the life cost exceeds 10k, the life expectancy is as high as 100 years old, which is very unrealistic and rare in reality.

To investigate if an individual's health spending has a major impact on his or her life expectancy. We can estimate the degree and direction of the association between personal life expenditures and life expectancy using a linear regression model.

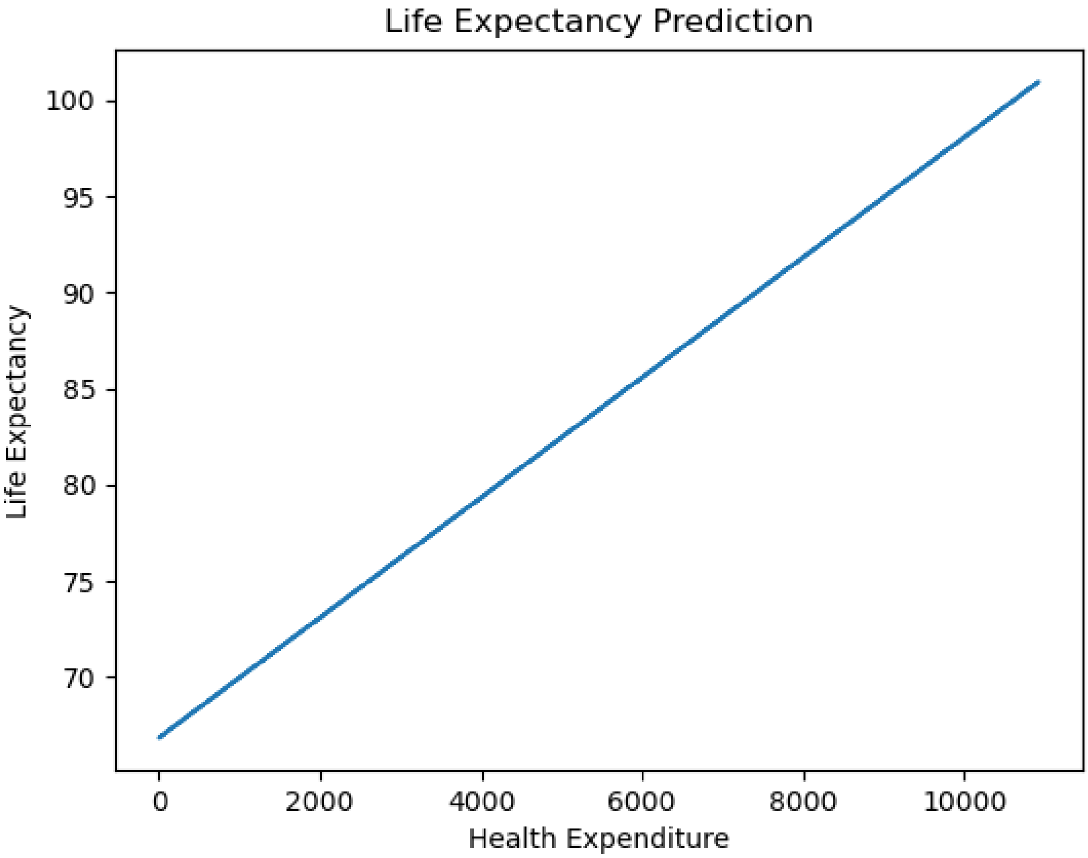
2.3 By using a linear regression model, explore the relationship between health expenditure and life expectancy for 261 countries from 2000 to 2019.

Scatterplots are useful exploratory tools, but machine learning models are necessary for rigorous analysis. The above scatterplots provide limited information about the strength and direction of the relationship between the independence and dependence variables, we cannot determine the statistical significance or identify nonlinearities. Thus, we decided to use the regression model for a more in-depth and comprehensive analysis by quantifying the relationship and identifying interactions, so that we can make predictions based on the relationship.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                LifeExp   R-squared:                       0.314
Model:                            OLS   Adj. R-squared:                  0.314
Method:                 Least Squares   F-statistic:                     2068.
Date:                Wed, 08 Mar 2023   Prob (F-statistic):               0.00
Time:                        23:12:08   Log-Likelihood:                -15307.
No. Observations:                4522   AIC:                         3.062e+04
Df Residuals:                    4520   BIC:                         3.063e+04
Df Model:                           1
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept             66.8389      0.121    553.060      0.000      66.602      67.076
Health_Expenditure     0.0031   6.87e-05     45.473      0.000       0.003       0.003
==============================================================================
Omnibus:                      470.081   Durbin-Watson:                   1.740
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              629.996
Skew:                          -0.911   Prob(JB):                    1.58e-137
Kurtosis:                       3.153   Cond. No.                     2.00e+03
==============================================================================
```

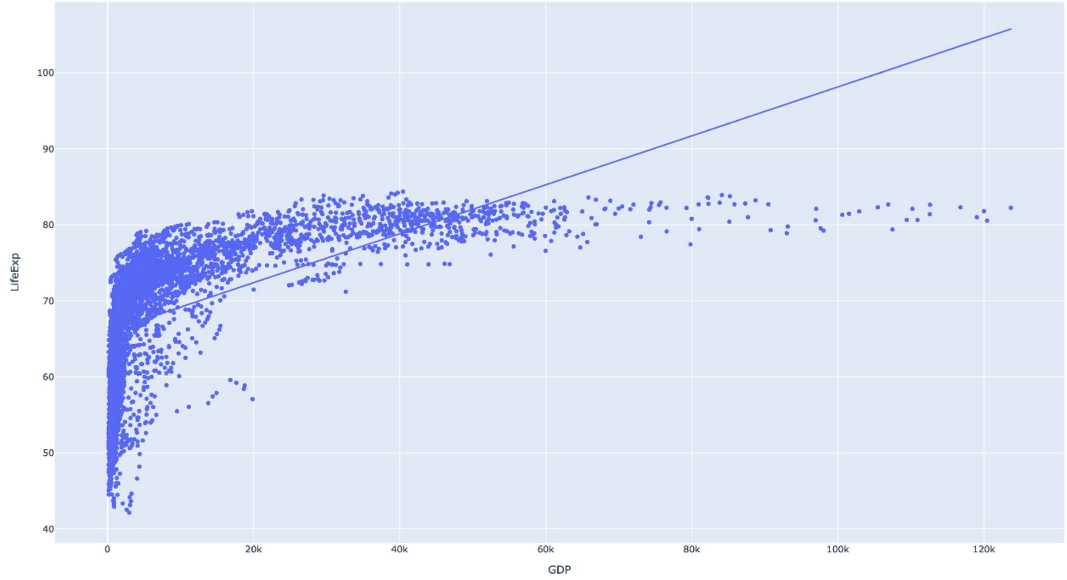
From the linear regression model above, we can conclude that there is a positive correlation between the individual health expenditure and individual life expectancy. Thus, an individual's health expenditure has a positive effect on their individual’s life expectancy.One unit increase in life expenditure will lead to 0.0031 year increase of life expectancy. When people do not spend any money on their health, they are expected to live to about 66 years old. The statistical significance of the results was assessed using a p-value and a t-test. The results of this model have statistical significance, since the p value is smaller than the alpha we used in this model. The t test for this model is 2.581 based on the degree of freedom 4520 and alpha 0.05. The t value for health expenditure is 45.473. The t value of the intercept is 553.06. Both coefficient and intercept t values are bigger than the t test value, suggesting that both of them are statistically significant. The null hypothesis can be rejected based on these results.

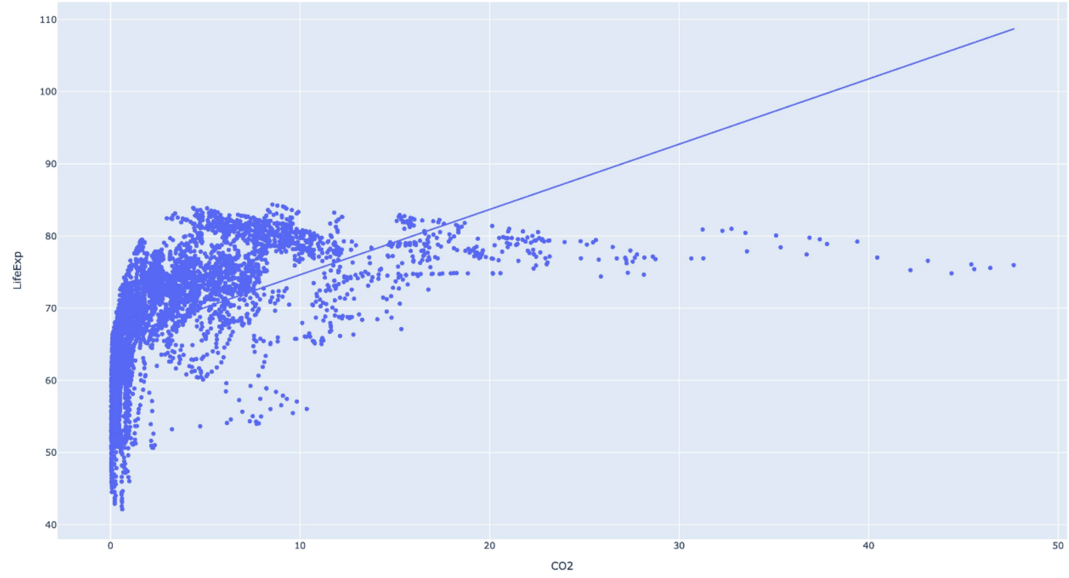


Based on the linear regression model, we create a line graph. According to the above line graph, there is a positive correlation between health expenditures and life expectancy. In particular, the graph demonstrates that spending around $1000 on health is related to an average life expectancy of 69 years. When health expenditures rise, so does life expectancy, showing that people who are ready to spend more on their health may attain a longer life span. Notably, the graph depicts a maximum life expectancy of 100 years, which is an expected constraint given that no amount of health expenditure can guarantee an indefinite longevity.We splitted the dataset into eighty and twenty percent. The adjusted root mean square error of this model is 0.169112, indicating that our model can well predict life expectancy based on health expenditure.The root mean square root errors for train and test data are 7.072511261847639 and 7.413254432192041. The difference between the two root mean square errors is not significant, so our model's accuracy is reliable enough to predict life expectancy. Although the model reveals an association between health spending and life expectancy, it does not account for other factors that may impact life expectancy, such as genetics, lifestyle factors, and environmental factors, which may limit the generalizability of the results.

Research problem 3: How do personal GDP, GDP level and CO2 emissions impact each person's life expectancy?

3.1 By using the scatterplots, explore the impact of individual CO2 emissions and individual GDP on the life expectancy of 261 countries from 2000 to 2019.





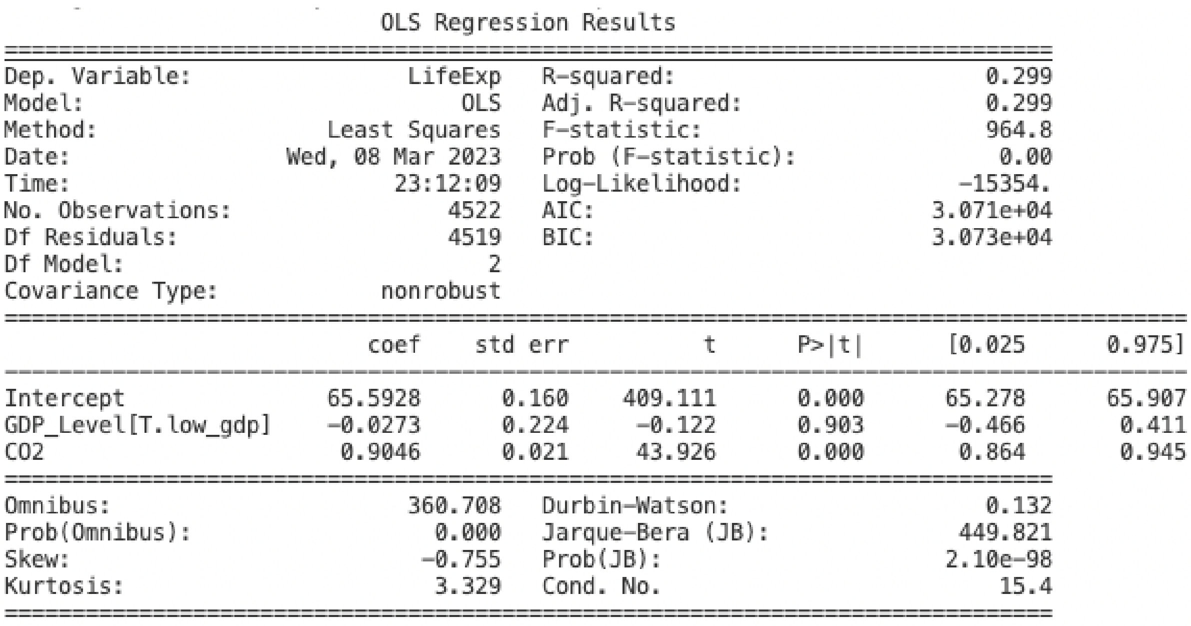
Personal GDP and CO2 emissions are also two important factors that may impact each person's life expectancy. To explore the relationship between these variables, we can use scatterplots to visually represent the data. The scatterplot between personal GDP and life expectancy shows that there is no obvious linear relationship between the two variables. Also, as we mentioned before(2.2), the regression line of these scatterplots is deviated and unrealistic. In the first figure, when the GDP value exceeds 80k, the life expectancy is as high as 90 years old, which is very unrealistic and rare in reality. This means that there is no clear pattern or trend that can be seen by just looking at the scatterplot. However, there may still be a relationship between personal GDP and life expectancy, but it may not be a simple or direct relationship. Similarly, the scatterplot between personal CO2 emissions and life expectancy also shows no obvious linear relationship between the two variables. Most importantly, when we study the impact of two variables (GDP and CO2) on life expectancy, we cannot just consider the impact of one variable(GDP or CO2) on life expectancy unilaterally. We need to consider the two variables together, and how these two variables together affect life expectancy.

To explore the question of how personal GDP and CO2 emissions impact each person's life expectancy, we can use a linear regression model. This model will allow us to determine the strength and direction of the relationship between personal GDP, CO2 emissions, and life expectancy.

3.2 By using the linear regression model, explore the impact of individual CO2 emissions and individual GDP on the life expectancy of 261 countries from 2000 to 2019.

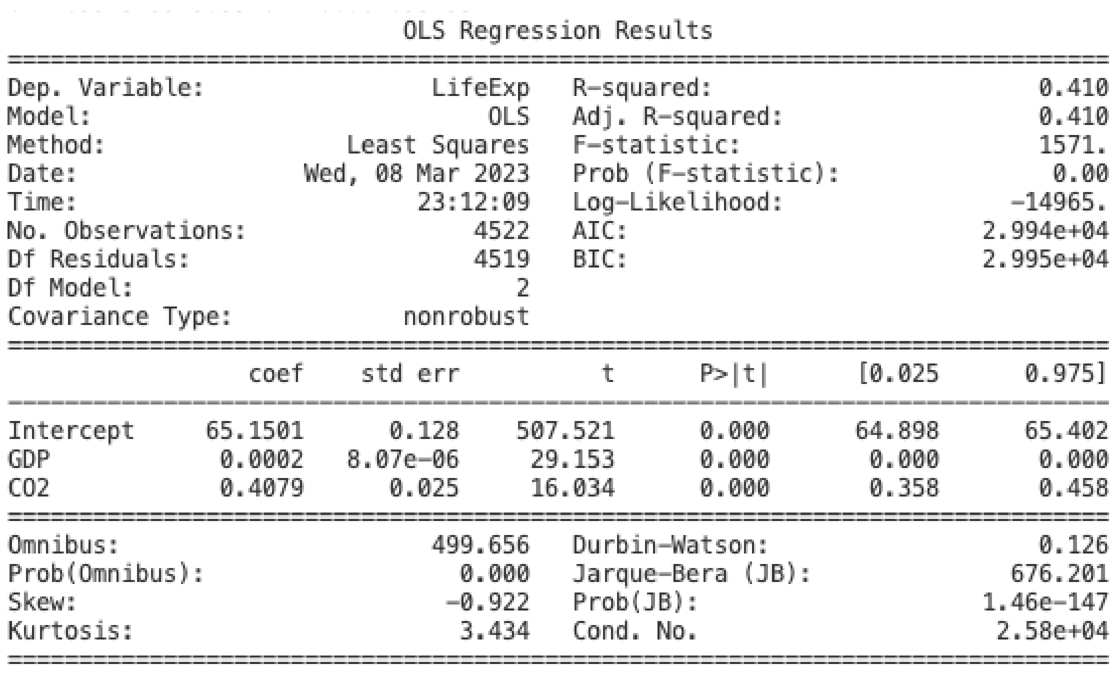
The scatterplots are useful exploratory tools, but machine learning models are necessary for rigorous analysis.Although above scatterplots provide limited information about the strength and direction of the relationship between the independence and dependence variables, we cannot determine statistical significance or identify nonlinearities. Thus, we decided to use the regression model for a more in-depth and comprehensive analysis by quantifying the relationship and identifying interactions, so that we can make predictions based on the relationship.

There are two measures to analyze if the GDP is high or low in one region. The first one is to convert GDP per capita to GDP level. That is creating GDP growth rates and taking the growth rate higher than 2 percent as high GDP region and the growth rate lower than 2 percent as low GDP region. The other method is directly looking at the GDP per capita, and the higher the GDP per capita is, the better the GDP at that region.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                LifeExp   R-squared:                       0.299
Model:                            OLS   Adj. R-squared:                  0.299
Method:                 Least Squares   F-statistic:                     964.8
Date:                Wed, 08 Mar 2023   Prob (F-statistic):               0.00
Time:                        23:12:09   Log-Likelihood:                -15354.
No. Observations:                4522   AIC:                         3.071e+04
Df Residuals:                    4519   BIC:                         3.073e+04
Df Model:                           2
Covariance Type:            nonrobust
==========================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept             65.5928      0.160    409.111      0.000      65.278      65.907
GDP_Level[T.low_gdp]  -0.0273      0.224     -0.122      0.903      -0.466       0.411
CO2                    0.9046      0.021     43.926      0.000       0.864       0.945
==============================================================================
Omnibus:                      360.708   Durbin-Watson:                   0.132
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              449.821
Skew:                          -0.755   Prob(JB):                     2.10e-98
Kurtosis:                       3.329   Cond. No.                         15.4
==============================================================================
```

We initially included GDP level, CO2 emissions, and life expectancy in our statistical model and obtained the table above. However, the p-value for the lower level GDP coefficient was found to be 0.903, which is greater than the alpha value of 0.05. This suggests that we cannot reject the null hypothesis, and that the coefficient is not statistically significant. Consequently, this model cannot be used to predict life expectancy, and the relationship between GDP, CO2 emissions, and life expectancy should be approached with caution. However, it is important to note that this model is not entirely useless. It indicates that there is no significant relationship between GDP level and life expectancy, which may be due to the fact that a good GDP growth rate does not necessarily translate to high income (GDP) for individuals in a given region. Thus, it may be more reasonable to use GDP per capita values directly in the model to analyze this relationship.



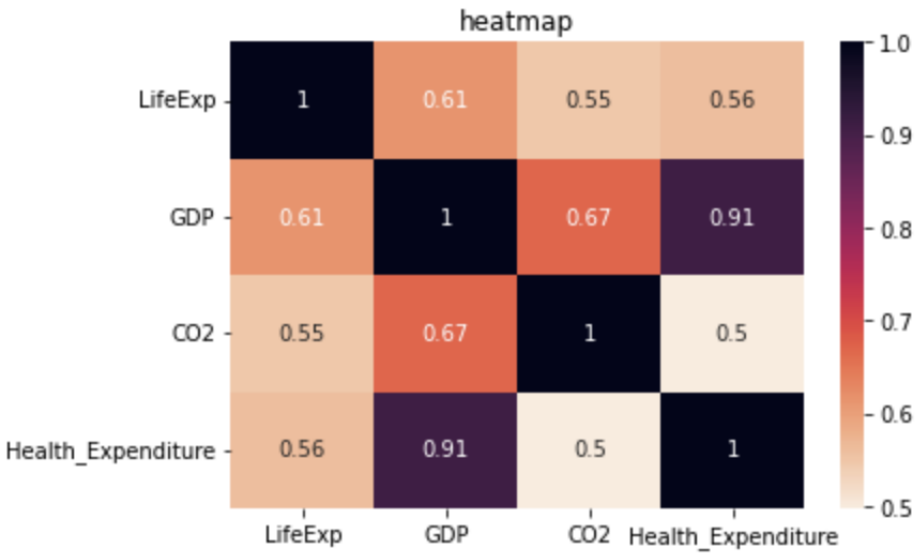
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                LifeExp   R-squared:                       0.410
Model:                            OLS   Adj. R-squared:                  0.410
Method:                 Least Squares   F-statistic:                     1571.
Date:                Wed, 08 Mar 2023   Prob (F-statistic):               0.00
Time:                        23:12:09   Log-Likelihood:                -14965.
No. Observations:                4522   AIC:                         2.994e+04
Df Residuals:                    4519   BIC:                         2.995e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     65.1501      0.128    507.521      0.000      64.898      65.402
GDP            0.0002   8.07e-06     29.153      0.000       0.000       0.000
CO2            0.4079      0.025     16.034      0.000       0.358       0.458
==============================================================================
Omnibus:                      499.656   Durbin-Watson:                   0.126
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              676.201
Skew:                          -0.922   Prob(JB):                    1.46e-147
Kurtosis:                       3.434   Cond. No.                     2.58e+04
==============================================================================
```

Based on this model, we can conclude that there is a positive relationship between CO2 emissions, GDP per capita, and life expectancy. The CO2 emissions coefficient of 0.4079 suggests that a one-unit increase in CO2 emissions is associated with a 0.4079 increase in life expectancy. Similarly, the GDP per capita coefficient of 0.0002 suggests that a one-dollar increase in GDP per capita is associated with a 0.0002 increase in life expectancy. The intercept of the regression at 65.1501 suggests that in the absence of CO2 emissions and GDP, people are expected to live up to 65 years. The p value is zero for every coefficient and interception. The t test for this model is 2.581 based on the degree of freedom 4520 and alpha 0.05. The t value for CO2 emissions is 16.034. The t value for GDP is 29.153. The t value of the intercept is 507.521. Since the p value is less than the alpha we use in this model and the t value for the coefficient and intercept is bigger than the t test value. The values of this model are statistically significant. We splitted the dataset into eighty and twenty percent. The adjusted root mean square error of this model is 0.15679917, indicating that our model can well predict life expectancy based on health expenditure. The root mean square root errors for train and test data are 6.639395957097121 and 6.56691455560923. The difference between the two root mean square errors is not significant, so our model's accuracy is reliable enough to predict life expectancy.

However, it is important to note that this model may have limitations. For example, it only considers the relationship between CO2 emissions, GDP per capita, and life expectancy, and does not account for other factors that may affect life expectancy, such as healthcare, education, and social policies. Additionally, the model assumes a linear relationship between the variables, which may not always hold in reality. Therefore, the conclusions drawn from this model should be interpreted in the context of its limitations.

Research problem 4: Which independent variables(GDP, CO2, and Health expenditure) has the greatest impact on an individual's life expectancy? How does changing the max depth of the decision tree affect its accuracy and feature importance?

4.1 By using the heat map, explore the relationship between these variables(individual GDP, individual CO2 emissions, and individual health expenditure) and the individual health expectancy of 261 countries from 2000 to 2019.



	LifeExp	GDP	CO2	Health_Expenditure
LifeExp	1	0.61	0.55	0.56
GDP	0.61	1	0.67	0.91
CO2	0.55	0.67	1	0.5
Health_Expenditure	0.56	0.91	0.5	1

This analysis utilized a heat map with life expectancy, GDP, CO2 emissions, and health expenditure as its x and y axes, forming a 4x4 data matrix. This heat map reveals that the variable having the largest influence on life expectancy is individual GDP, with a value of 0.61. Hence, there is a substantial positive correlation between GDP and life expectancy; as GDP rises, so does life expectancy. The other factors in the heat map have a lesser effect on life expectancy. Surprisingly, there is also a positive correlation between CO2 emissions and life expectancy(0.55), although it is smaller than the correlation between GDP and life expectancy. Health expenditure has a value of 0.56, indicating a positive association between health expenditure and life expectancy; however, this relationship is weaker than the one between GDP and life expectancy.

Although a heat map can offer an overview of the association between factors such as individual life expectancy, individual GDP, individual CO2 emissions, and individual health expenditure, it may not provide an exact or precise assessment of the relationship between all these variables and an individual's life expectancy. In other words, the heat map can only provide a basic sense of the degree of the relationship between these variables; it cannot jointly explore the impact of three variables on life expectancy. Hence, a decision tree model may provide a more precise estimate of a person's life expectancy depending on their CO2 emissions, GDP, and health expenditure. The decision tree approach may identify the most significant factors and their interactions, allowing for a more accurate estimation of an individual's life expectancy. In addition, the decision tree model handles missing data more efficiently than a heat map or other graphical depiction, which is particularly important when working with actual data.

4.2 By using the decision tree model, explore the impact of the three independent variables of individual CO2 emissions, individual GDP, and individual health expenditure on the dependent variable of individual life expectancy.

The above scatterplots are useful exploratory tools, but machine learning models are necessary for rigorous analysis. Although the above scatterplots provide limited information about the strength and direction of the relationship between the independence and dependence variables, we cannot determine the statistical significance or identify nonlinearities. Thus, we decided to use the decision tree model for a more in-depth and comprehensive analysis by quantifying the relationship and identifying interactions, so that we can make predictions or classifications based on the relationship.

First of all, the first step in creating a decision tree model is to determine the max depth of the model and convert our continuous variable into a discrete variable for classification. For example, in the question we explored, we categorized the dependent continuous variable life expectancy into the following groups:

- younger than 50 years old
- 50-60 years old
- 60-70 years old
- 70-80 years old
- older than 80 years old

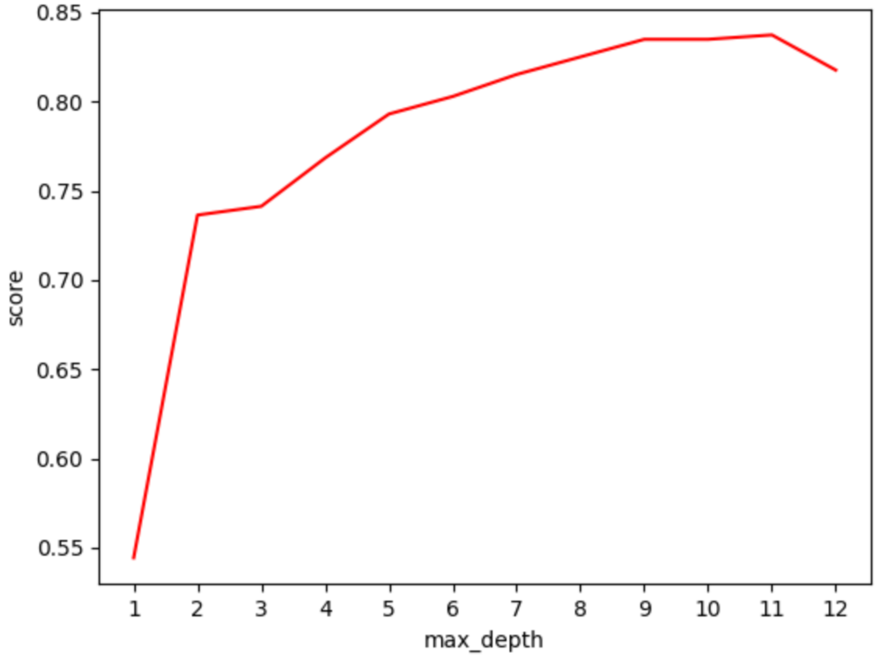
When we determine the max depth of the model, we first create two decision tree classifiers, one with a max depth of 3(with an accuracy score is 74.14%) and one without limiting the max depth(with an accuracy score is 83.49%). But we found that when the max depth of the model is set to 3, the features importance of GDP, CO2, and Health expenditure are 0.0, 0.37, and 0.63 respectively. It can be found that the feature importance of GDP is only 0.0. On the other hand, when the max depth of the model we created is not set, the feature importance of GDP, CO2, and Health expenditure is 0.14, 0.37, and 0.49 respectively. Through this interesting phenomenon, We realize that the possible reasons for this phenomenon are:

- The impact of GDP itself on life expectancy is very small, causing the result got is only 0.0.

- The max depth of our initially established model is only 3 layers, and the accuracy of 3 layers(with an accuracy score is 74.14%) is lower than the accuracy of the model without limiting the max depth(with an accuracy score is 83.49%).

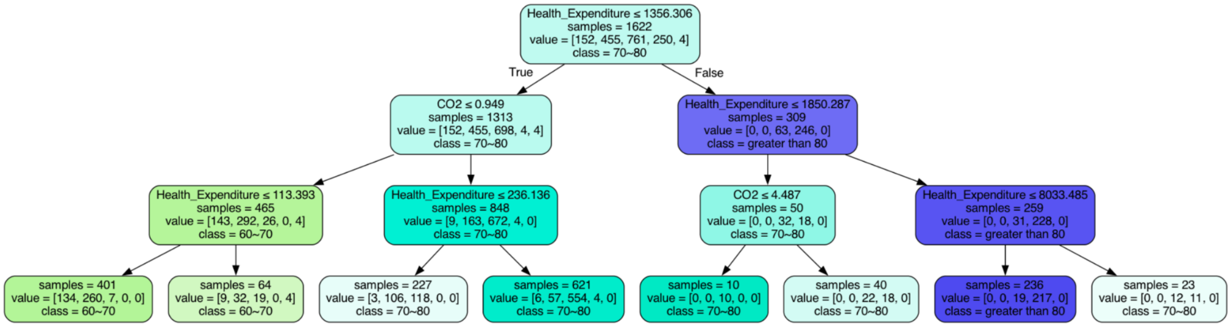
Therefore, based on this conjecture, we conducted further verification. We set the max depth of the model to “12” to ensure that:

- The max depth of the model set is interpreted without caution, so it avoids overfitting the training data.
- At the same time, our model can eliminate the deviation because the number of max depth layers is high enough. Therefore, we used this idea to do the visualization to explore the impact of the max depth of the model for the features importance of variables and what value the max depth of the model should use to ensure that the accuracy of our model is as high as possible.



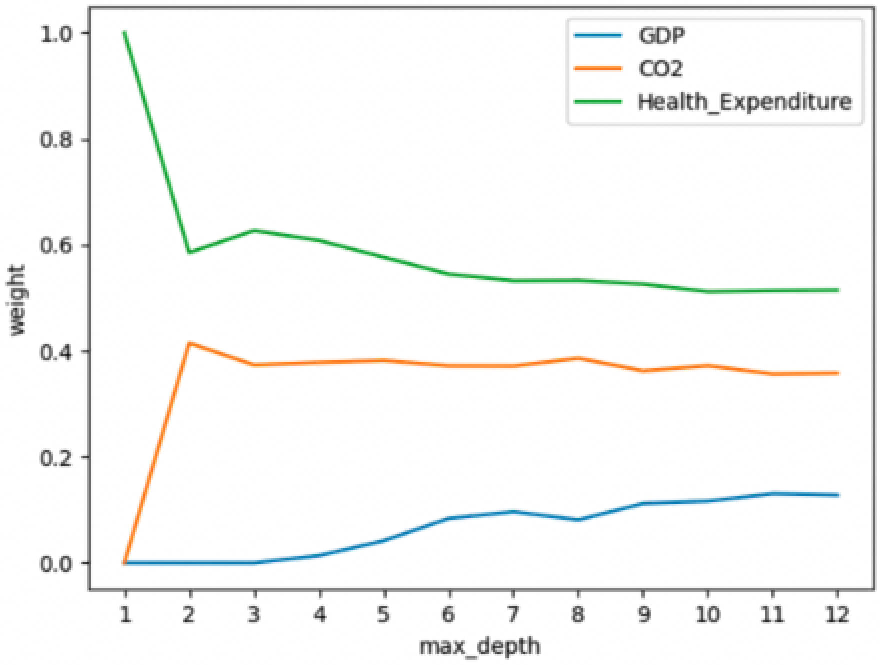
In order to better express the relationship that max depth of the decision tree model has a significant impact on the the accuracy of the model, we show a linear graph above representing the accuracy score of the model accuracy changes with max-depth changes.

Through the above image of the model, we can see the relationship between the max depth of the model and the accuracy of the model. When the max depth of the model is 1, the corresponding model accuracy is about 55%. And when the max-depth of the model changes from 1 layer to 2-3 layers, the accuracy of the model has a significant increase, rising to about 74%. But it is worth noting that the accuracy of the model rises to about 83% when the model has about 9-10 layers.


Health_Expenditure ≤ 1356.306
samples = 1622
value = [152, 455, 761, 250, 4]
class = 70~80
True
False
CO2 ≤ 0.949
samples = 1313
value = [152, 455, 698, 4, 4]
class = 70~80
Health_Expenditure ≤ 1850.287
samples = 309
value = [0, 0, 63, 246, 0]
class = greater than 80
Health_Expenditure ≤ 113.393
samples = 465
value = [143, 292, 26, 0, 4]
class = 60~70
Health_Expenditure ≤ 236.136
samples = 848
value = [9, 163, 672, 4, 0]
class = 70~80
CO2 ≤ 4.487
samples = 50
value = [0, 0, 32, 18, 0]
class = 70~80
Health_Expenditure ≤ 8033.485
samples = 259
value = [0, 0, 31, 228, 0]
class = greater than 80
samples = 401
value = [134, 260, 7, 0, 0]
class = 60~70
samples = 64
value = [9, 32, 19, 0, 4]
class = 60~70
samples = 227
value = [3, 106, 118, 0, 0]
class = 70~80
samples = 621
value = [6, 57, 554, 4, 0]
class = 70~80
samples = 10
value = [0, 0, 10, 0, 0]
class = 70~80
samples = 40
value = [0, 0, 22, 18, 0]
class = 70~80
samples = 236
value = [0, 0, 19, 217, 0]
class = greater than 80
samples = 23
value = [0, 0, 12, 11, 0]
class = 70~80

From the standpoint that the report image can be readily navigated, the decision tree model above is the model when the model max depth is set to "3" so that the data in the model can be seen more clearly.The decision tree analysis in this code suggests that the most important factor impacting life expectancy is health expenditure. If health expenditure is less than or equal to 1356.306, then the tree splits further based on the CO2 emissions. If CO2 emissions are less than or equal to 0.949, the model then splits again based on the health expenditure variable. If health expenditure is less than or equal to 113.393, then the predicted life expectancy is in the 60-70 age range. If health expenditure is greater than 113.393, then the predicted life expectancy is in the 70-80 age range. On the other hand, if CO2 emissions are greater than 0.949, then the predicted life expectancy is in the 70-80 age range, regardless of health expenditure.

Interestingly, the decision tree model suggests that personal GDP does not have a significant impact on life expectancy. This is evident as the model does not split based on the personal GDP variable. Furthermore, the model suggests that increasing CO2 emissions may actually have a positive impact on life expectancy. However, it is important to note that the increase in life expectancy is marginal (0.4742 years for each unit increase in CO2 emissions), and other factors such as health expenditure still have a larger impact on life expectancy. Overall, this decision tree model can provide insights into the complex relationship between different variables and their impact on life expectancy.



Also, in order to better express the relationship that max depth of the decision tree model has a significant impact on the the feature importance of independent variables, we show a linear graph above representing the feature importance of the three independent variables changes with max depth changes.

Through the above image, we can find that the impact of individual GDP on the life expenditure of individuals is very small. This above line graph also verifies the conclusions obtained from the decision tree graph that we can hardly see the data on the impact of individual GDP on the life expenditure of individuals. Also, in the line graph, although we can clearly see that although GDP has always been stable in feature importance, compared with the other two variables, its feature importance has always been the smallest. When the max depth is 1 to 5, its feature importance is always less than 0.1. Until the max depth is 9, its feature importance will reach about 0.15. In contrast, the feature importance of CO2 emissions is higher than that of GDP. It is worth noting that when the model max-depth ranges from 1 to 2, its feature importance increases from 0.0 to 0.4, with a dramatic increase. When the model max-depth reaches 3, it gradually tends to be stable. Among all independent variables, health expenditure has the greatest impact on life expectancy. Unlike the other two independent variables(CO2 emissions and GDP), the feature importance of health expenditure shows a downward trend when the model max-depth is 1 to 2, decreasing from 1.0 to 0.6. However, with the increase of max depth, it also maintains the highest feature importance. It is noticeable that the feature importance of the three independent variables remains relatively stable after the max depth of the decision tree model is "9".

The decision tree model we established can help us understand the relationship between life expectancy and independent variables (GDP, CO2, and Health expenditure). Based on the model we have built, we can draw the following conclusions:

- Among the three independent variables, individual health expenditure is the factor that has the greatest impact on life expectancy, followed by individual CO2 emissions, and finally individual GDP, which also proves the conclusion that we got from the linear regression model in the previous problems.
- Moreover, we also found that the max depth of the decision tree model has a significant impact on the accuracy of the model and the feature importance of independent variables. The results suggest that when the max depth of the model is about 9-10, it is more accurate than the model with a max depth of 3, and the feature importance of GDP is higher than the result of the max depth of 3.

Despite the benefits of the decision tree model, it has some limitations:

- It requires a relatively large dataset, and the model may overfit the training data.
- In addition, decision trees are vulnerable to instability, and slight changes in the data can lead to completely different models.
- Moreover, decision trees are susceptible to bias towards dominant classes and may not be suitable for complex datasets.
- Finally, while the model can identify correlations between variables, it cannot prove causality. Therefore, it is essential to consider these limitations when using decision trees in real-world applications.

Impact and Limitations

Policymakers and public health professionals may be able to utilize the information to make educated decisions about health expenditures, GDP, and CO2 emissions as a result of the findings. Countries with high health spending outliers should modify their policies to make healthcare more accessible, and countries with high CO2 emissions could consider implementing more sustainable practices. In addition, the conclusion that individual health expenditure has the biggest influence on life expectancy shows the significance of healthcare access and the necessity for politicians to guarantee that individuals can afford to take care of their health.

The analysis might be beneficial to politicians, academics, and healthcare professionals, who could utilize the data to devise treatments that could improve health outcomes. However, the analysis may omit or injure individuals who do not belong to the investigated population, such as those in nations not included in the research or with distinctive health systems. Just like in

Zaman's (2017) study, the researchers mentioned that "Health system characteristics, such as the provider payment mechanisms and the degree of private provision of the services were not included in the study due to the lack of adequate time series data.”Our datasets may not cover all countries, such as small and remote countries. In addition, there are a few missing data in our original data sets. The account of the results may also be featured by the analysis's biases, such as missing data and outliers. At the same time, if we want to predict the effects of individual health expenditures, individual GDP, and CO2 emissions on individual life expectancy by models more accurately, we also need more years of data, such as exploring data before 2000, so that our data research can cover more comprehensive data. As Chawla's (2021) said, the researchers mentioned that "Researchers have demonstrated that massive data can lead to lower estimation variance and hence better predictive performance."

Due to the observational nature of the study, the analysis's shortcomings include an inability to demonstrate causality. In addition, the study only investigated four variables, but additional variables, such as education and access to clean water, might alter life expectancy. As Hathaway (2020) mentioned, “there were also notable differences in rates of death by education level. Approximately 13% of participants with a high school degree or less education died compared with only approximately 5% of college graduates.” In addition, due to the sampling strategy of the data set, the conclusions of the analysis might not apply to particular populations or regions. Policymakers should take the study's findings as a reference and also consider other aspects when making choices.

In conclusion, our research gives significant insights into the variables that affect life expectancy and how they differ between nations, regions, and years. The information might be used by policymakers to build targeted initiatives that could enhance health outcomes, but they need also to consider other aspects when making judgments. When adopting the study's results as a basis for their own work, researchers should be cognizant of the study's limitations.

Challenge Goals

Our challenge goals were expanded. Except for the "Machine Learning" and "Multiple Datasets" challenges mentioned in the proposal, we also encountered “New Library” and “ Result Validity” challenges for building the model and plotting. For example, when making a decision tree model, we need to use a new library"graphviz". In addition, in order to prove the credibility and accuracy of our model, we also added the part about test validity for our linear regression model and the decision tree model.

- **Multiple Datasets**

Our research uses four datasets, which presents a problem in integrating and making sense of the varied information that each dataset offers. Although the CO2 emissions, life expectancy, health expenditure, and GDP of each nation are all significant markers of the health and growth of a society, there isn't always an obvious correlation between them. We will need to integrate the datasets in a meaningful way that enables us to discover insights that can't be derived from any one dataset alone in order to do more in-depth research. Dealing with missing or inconsistent data, matching observations across datasets based on shared factors like country or time, and dealing with various units or scales of measurement will likely be involved in joining or merging the datasets. The handling of outliers, missing data, and confounding variables that could be altering the correlations between the indicators we are researching will also need some decision-making on our part. Finally, in order to use the merged datasets to respond to our research questions, we will need to be inventive. To generate meaningful comparisons, we could need to establish new variables, merge already existing ones, or apply statistical techniques to account for confounding variables. Instead of just using more data for the sake of using more data, the datasets are combined to provide a deeper and more sophisticated analysis.

- **Machine Learning**

Our project's aim is to study the relationship between an individual’s CO2 emissions, individual’s life expectancy, individual’s health expenditure, and individual’s GDP of each nation while also making future predictions using machine learning. Our group chose to apply the linear regression model and the decision tree model to accomplish this objective. Preprocessing the data, dividing it into training and test sets, creating the model, assessing the model's performance, and interpreting the findings are some of the major tasks in the project. Our ultimate objective is to understand the relationships between the variables and make precise predictions about the future. This will be evaluated by assessing the model's performance on the test data and then interpreting the results to determine which variables are most crucial for predicting the life expectancy of an individual.

- **New Library**

    - plotly
        - The plotly library is used to create interactive and high-quality visualizations in the functions of file4(Data Visualization File). Specifically, the "plotly.express" module is used to create scatterplots and the "plotly.subplots" module is used to create subplots for the country-specific scatter plots. The library offers a wide range of customization options for the visualizations such as trendlines, colors, and titles.
        - In the "create_scatter_plot" function, the plotly.express module is used to create a scatter plot with a trendline for the specified x and y columns in a pandas DataFrame. The "show" method is then called on the resulting plot object to display the plot.

        - In the "create_country_scatter_plot_to" function, the "px.scatter" function is used again to create a scatter plot with a trendline, this time with all four countries plotted on the same graph. The "color" parameter is used to distinguish between the countries, and the resulting plot object is displayed using the "show" method.

    - Scipy
        - In file 2 (Data Analysis File), the line ‘from scipy.stats import norm’ imports the norm function from the scipy.stats module.
        - The norm function is a part of the scipy.stats module and is used to create a normal distribution object. The norm.pdf function is then called with the object, and the minimum and maximum values of the data to generate a normal distribution plot for the given data.
        - In the plot_distributions function, the norm.pdf function is used to plot the normal distribution curve for each column in the DataFrame. The fit variable is assigned to the output of the norm.pdf function called with a sequence of values generated using np.linspace, the mean and standard deviation of the data. This generates a sequence of y-values for the normal distribution curve for the given column.
        - The norm.pdf function is used in combination with the histogram plot to show the distribution of the data for each column in the DataFrame.

    - graphviz
        - The plot_tree function uses the export_graphviz function from the sklearn.tree module to create a Graphviz representation of the decision tree model. It takes in the trained decision tree model, feature names, and target label names as arguments.
        - The export_graphviz function converts the decision tree model into a Graphviz-compatible format, which is a plain text file that describes the tree's structure and properties. The resulting string is then passed to the Graphviz.Source function to create a Graphviz object.
        - Finally, the render function of the Graphviz object is called to save the Graphviz source code as a file named 'tree.gv'.

    - Statsmodels
        - The statsmodels library provides ordinary least squared modeling for creating regression models.
        - The smf.ols function in statsmodels.formula.api is used to fit ordinary least squares linear regression models.

        - The fit method is called on the OLS object to fit the model and return a RegressionResults object that contains the results of the fit.
        - The fitted model allows us to make predictions on the dependent variables based on the independent variables used in the model.

- **Result Validity**
    - Accuracy score
        - In our decision tree module(file4), result validity is applied through the use of the accuracy_score() function from the sklearn.metrics module. We include the cross-validation test involves splitting the dataset into training and testing sets and evaluating the model's performance on the testing set. The accuracy_score() function is used to calculate and print the accuracy of the predictions made by the decision tree classifiers with and without a max depth of 3. This provides a measure of how well the model is able to predict the target variable based on the input features.
        - Additionally, the code also uses a loop to train and test decision tree classifiers with different max depths and calculates the accuracy scores and feature importances for each classifier. This helps to verify the validity of the results and ensure that they are not likely to happen by chance. The accuracy scores and feature importance are plotted to provide a visual representation of the results. The use of the accuracy_score() function and the loop to train and test multiple decision tree classifiers with different max depths help to ensure the validity of the results obtained from the decision tree model.
    - Root mean square error (RMSE) and Adjusted root mean square error:
        - We used root mean square error to test the difference between our prediction and the real value. It provides data analysis of the variance between predicted and actual values. The value should be analyzed based on the research question.
        - To verify the validity of the results, our team uses the train-test split method to evaluate the model's performance on the test set. This technique splits the dataset into two subsets: the training set, which is used to fit the model, and the test set, which is used to evaluate the model's performance. The code calculates the RMSE based on the training set and the test set separately.
        - To test the accuracy of the prediction, our team uses the adjusted root mean square error to numericalize the accuracy. If the adjusted root mean is close to zero, the accuracy is promised in the model and vice versa

Work Plan Evaluation

Working on the assignment assigned to them, each team member will write code in their individual development environment. Each team member will test their individual pieces to ensure functionality before merging the code. In order to test the entire system and make sure it operates as intended, code integration will be used. Team members must work together, and frequent meetings will be conducted to examine the project's status, address any issues, and discuss future goals. If a task is challenging, team members will support one another with their knowledge and skills. It will also be determined that each member's roles are properly explained in order to provide help as necessary.

Our team thinks that the proposed time estimate for the work plan is not very precise. We spend far more time than anticipated. Our team thinks the following are the primary causes of this result:

- When processing data, the original year must be converted from numerous independent columns to a combined column. Due to the frequent occurrence of error messages during the merging process, we devote a great deal of work to this step.
- We invested a great deal of effort in the modeling process. Initially, we intended to model using a logistic regression model. However, the logistic regression model is sensitive to outliers, which assumes that the data follows a Gaussian distribution, and outliers can significantly affect the model. In contrast, the decision tree is not affected by outliers and can handle them using majority voting or averaging. So, we adopted a decision tree model for research problem 4.
- For the decision tree model, we also spent a considerable amount of time considering, for example, how to decide the hierarchical structure of the decision tree and the max depth of the decision tree. In addition, we spent a great deal of work determining the correctness of our model and incorporated several additional libraries.
- For the test validity section, we also spent a considerable amount of effort evaluating the correctness of the model for the computed root mean square error (RMSE), adjusted root mean square error and accuracy score, as well as determining if these two models can be applied to the real background to anticipate life expenditure.

Here is our work plan details:

Task 1: Set up a development environment and Collect Data (6 hours)

- Prepare the environment with the required dependencies, such as Pandas, NumPy, Matplotlib, and SciPy.
- Collect data on individual health expenditure, individual life expectancy, individual GDP, and individual CO2 emissions for various individuals.

- Ensure that the data is comprehensive and covers a large number of individuals and notions across a range of time.

Task 2: Data Cleaning and Combination (6 hours)

- Clean and combine the four datasets for analysis, only selecting the columns that our analysis needed.
- Ensure that all data is in the correct format and remove any outliers.

Task 3: Data Modelling (30 hours)

- **Problem 1:** Analyze the trends and patterns in the distribution, range, and outliers of GDP/CO2/Life Expectancy/Health Expenditure datasets across 261 countries from 2000 to 2019 by using the methods of loading data from a CSV file and calling the necessary functions to plot histograms, normal distributions, box plots, and histograms with outliers removed for specified columns in the Pandas DataFrame.
- **Problem 2:** Analyze the relationship between health expenditure and life expectancy by using statistical methods of building the linear regression model in python to compute the correlation between these variables.Using the scatterplots to visualize the correlation and relations between health expenditure and life expectancy.
- **Problem 3:** Evaluate the impact of GDP and CO2 emissions on life expectancy by using statistical methods of building a linear regression model in python to compute the correlation between these variables.Using the scatterplots to visualize the correlation and relations of GDP and CO2 emissions on life expectancy.
- **Problem 4:** Develop a decision tree model to predict life expectancy based on the three factors (CO2 emissions, GDP, and health expenditure). Evaluate the accuracy of the model using the accuracy score method. Evaluate the performance of the decision tree model in predicting life expectancy. Determine if the model is robust and can generalize well to new data.

Task 4: Test (8 hours)

- To test the validity of the results, we can perform several statistical tests such as the p-value, the t-value, the R-Squareed test, and the cross-validation test.
- The p-value test and t-value test measures the significance of the relationship between the dependent and independent variables and checks whether the coefficients are significantly different from zero.
- The R-Squareed test measures value indicating a better fit.
- The cross-validation test involves splitting the dataset into training and testing sets and evaluating the model's performance on the testing set.

Task 5: Data analysis and interpretation (10 hours)

- By observing and interpreting the data analysis result, write a report summarizing the findings and recommendations of the research questions:
    a. What are the trends and patterns in the distribution and outliers of GDP/CO2/Life Expectancy/Health Expenditure datasets across 261 countries from 2000 to 2019?
    b. Does an individual's health expenditure have an effect on their individual's life expectancy?
    c. How do personal GDP and CO2 emissions impact each person's life expectancy?
    d. Which independent variables(GDP, CO2, and Health expenditure) has the geatest impact on an individual's life expectancy?How does changing the max depth of the decision tree affect its accuracy and feature importance?

- Consider the limitations of the data analysis and the facts of the research background information to analyze the data justice and the data ethics of the whole data analysis results.
- Prepare a presentation and a poster to show our report on the data science fair.

Testing

In our file3 (Linear Regression Model), we import the train_test_split function as well as smf.ols functions to test our model.

- For our regression models, the generated model summary chart is a resource of testing the model. The p value and t value give insights of whether the coefficients and intercept is statistically significant or not.

- The root mean square error makes it possible to test the variance between the predicted value and actual value. The root mean squared error tells us how far the predicted value is from the actual value.

- The adjusted root mean square value also gives insight in the difference between predicted value and actual value.

- The R squared value tells us how well the model fits the data. The larger the R squared value is, the better the model fits the data.

P value and t value in our summary chart suggest that our coefficients and intercept are statistically significant, so the null hypothesis can be rejected and our models can be used to predict life expectancy based on independent variables.

In our file4 (Decision tree module), we use various testing methods, such as train-test split and accuracy_score function, to ensure the decision tree model is accurate and reliable, while plot_depth_score and plot_feature_importances functions help to validate the results obtained:

- We use several testing methods to ensure the report is correct. Firstly, it uses a train-test split to split the data into training and testing datasets. It then trains the decision tree model on the training dataset and uses it to predict the labels of the testing dataset. The accuracy_score function from sklearn.metrics is used to calculate the accuracy of the predictions, which is printed for both classifiers.

- Additionally, we also test the decision tree model with different max depths to calculate the accuracy scores and feature importances. The accuracy scores are plotted against different max_depth values using the plot_depth_score function, while the feature importances are plotted against different max_depth values using the plot_feature_importances function.

- Overall, the testing methods used in the code help to ensure that the decision tree model is accurate and reliable and that the results obtained from the model are trustworthy.

In our file6 (Test file), we import specific functions from other files for testing:

- clean_data() from file1.py: A function that cleans the data and saves the cleaned data to a new CSV file.

- plot_distributions() and no_outliers_plot_distribution() from file2.py: Functions used to create distribution plots of the input data.

- create_scatter_plot() and create_country_scatter_plot_to() from file4.py: Functions used to create scatter plots of the input data.

The first test function, test_clean_data(), tests the clean_data() function in file1.py by verifying that it removes any NaN values from the dataset and saves the cleaned data to a new CSV file named clean_data.csv. The function then reads the saved file and checks that there are no NaN values in the data and that the shape of the data is correct. The test function uses assert statements to check that the data shape and the number of NaN values are as expected.

The second test function, test_functions(), tests the plot_distributions() and no_outliers_plot_distribution() functions in file2.py. The test function creates a test dataset with specific column names and distributions, then passes the dataset to each of the functions and verifies that the expected output is produced. The function uses assert statements to check the correctness of the output.

The third test function, test_create_scatter_plot(), tests the create_scatter_plot() function in file4.py. The test function creates a test dataset with specific values and column names, then calls the function with the dataset and verifies that the expected plot is produced. The function uses assert statements to check the correctness of the plot.

The fourth test function, test_create_country_scatter_plot_to(), tests the create_country_scatter_plot_to() function in file4.py. The test function reads a test dataset from a CSV file, then calls the function with the dataset and verifies that the expected plot is produced. The function uses assert statements to check the correctness of the plot.
All the test functions use assert statements to check the correctness of the output of the data. Using assert statements is a good way to ensure that the expected output is produced and that any unexpected behavior or bugs can be easily detected. The test data used in each function was carefully chosen to cover a wide range of scenarios, including edge cases, to ensure that the functions behave as expected under various conditions.

Overall, the test functions provide good coverage of the code and ensure that the functions work as expected. The test data and the use of assert statements increase the reliability of the tests, and the inclusion of the test files ensures that the tests can be easily repeated by others to verify the correctness of the results.

Collaboration

We consulted the following platforms for this project:

- Stack Overflow: Examine past posts to see how others have resolved the same problem as us.
- Reddit: Examine past posts to see how others have resolved the same problem as us.
- Quora: Examine past posts to see how others have resolved the same problem as us.

References

CO2 emissions (KT). Data. (2022). Retrieved February 11, 2023, from https://data.worldbank.org/indicator/EN.ATM.CO2E.KT?end=2019&start=1990&view=char

Current health expenditure per capita (current US$). Data. (2022). Retrieved February 12, 2023, from https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD

Chawla, V. (2021, June 9). *Is more data always better for building analytics models?* Analytics India Magazine. Retrieved March 7, 2023, from https://analyticsindiamag.com/is-more-data-always-better-for-building-analytics-models/

Earth - Place Explorer. (n.d.). Data Commons. Retrieved March 6, 2023, from https://datacommons.org/place/Earth?utm_medium=explore&mprop=lifeExpectancy&popt=Person&hl=en

Frenk, J., Chen, L., Bhutta, Z. A., Cohen, J., Crisp, N., Evans, T., ... & Zurayk, H. (2019). *Health professionals for a new century: transforming education to strengthen health systems in an interdependent world.* The Lancet, 376(9756), 1923-1958.

GDP (current US$). Data. (2022). Retrieved February 11, 2023, from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

Hathaway, B. (2020, February 20). *Want to live longer? Stay in school, study suggests*. YaleNews. Retrieved March 7, 2023, from https://news.yale.edu/2020/02/20/want-live-longer-stay-school-study-suggests

Health at a Glance 2021: OECD Indicators. (2021). OECD Publishing.

Life expectancy at birth, total (years). Data. (2022). Retrieved February 11, 2023, from https://data.worldbank.org/indicator/SP.DYN.LE00.IN

Roser, M., Ortiz, E., & Ritchie, H. (n.d.). *Life Expectancy*. Our World in Data. Retrieved March 6, 2023, from https://ourworldindata.org/life-expectancy#citation

Zaman, S. B., Hossain, N., Mehta, V., Sharmin, S., & Mahmood, S. A. I. (2017). *An Association of Total Health Expenditure with GDP and Life Expectancy*. Journal of Medical Research and Innovation, 1(2), AU7-AU12. https://doi.org/10.15419/jmri.72