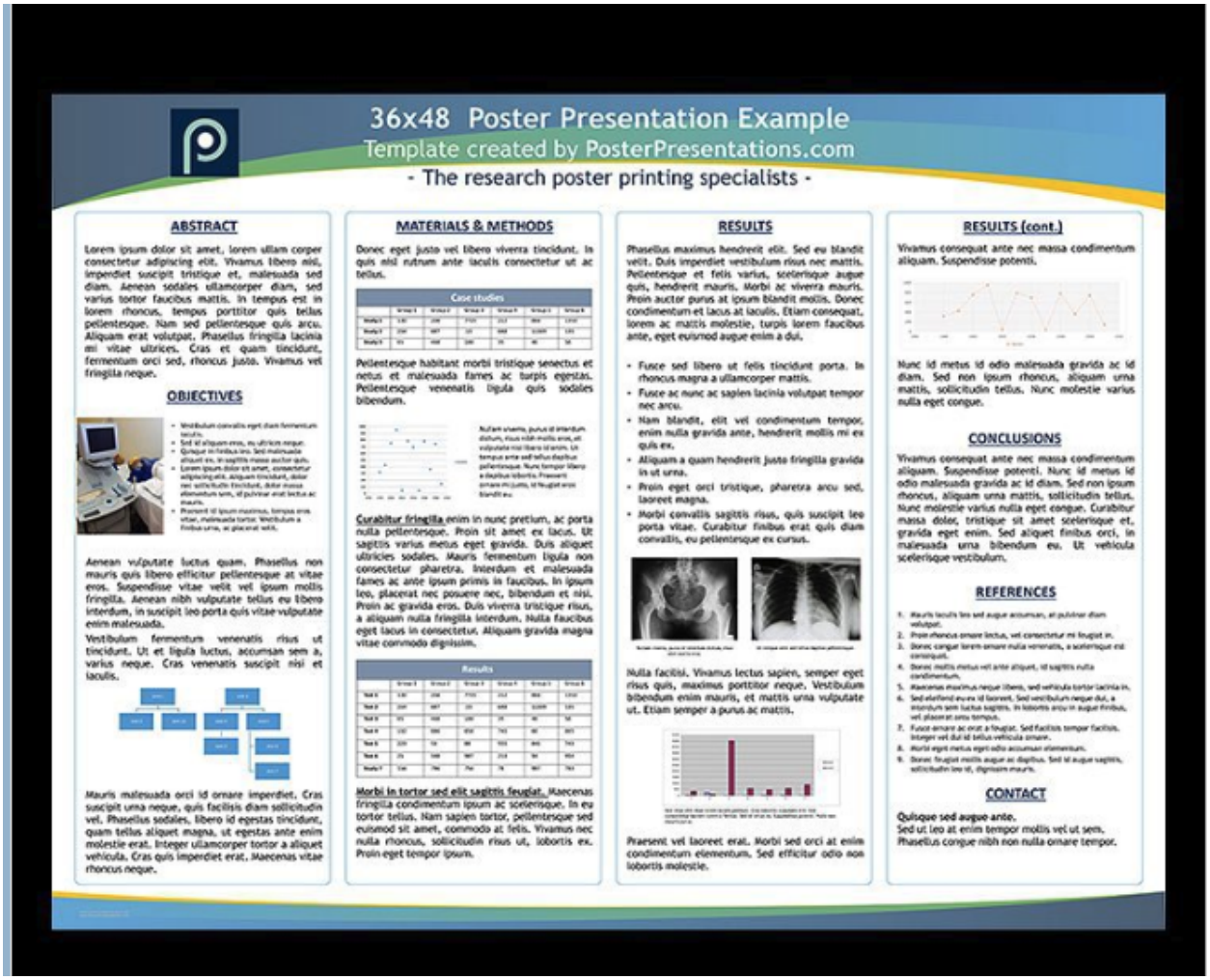
# 海报

尺寸：48” 英尺宽 and 36” 英尺高 = 121.92 cm宽 91.44 cm高

呈现字体尽量清晰 让读者方便看清

海报类型：




内容：（以下中文的部分不包含在海报中，只是方便理解，蓝色部分字体都需要加在海报里，我的内容中加粗的部分在海报中也要都加粗）：**红色部分是改动的部分,另外所有的图片尽可能的放大**

- Title(标题加粗加大)：**Exploring the impact of CO2 emissions, GDP, and health expenditure on individual life expectancy**
- Authors and group number(作者和组号）：

**Team number：(确定好了告诉你再添加） Authors: Xinyu Chang, Junhan Zhang, Xinlong Chen**

- **Motivation and background（这部分的标题放大，研究动机以及背景介绍，这个单词加大加粗）：**

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.

- **Research questions（这部分的标题放大，研究问题，这个单词加大加粗,整个部分都大一点）：**
- **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?**
- **Does an individual's health expenditure have an effect on their individual’s life expectancy?**

- **How do personal GDP, GDP level and CO2 emissions impact each person's 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?**
- **Methodology（这部分的标题放大，研究方法，这个单词加大加粗）：**

**Cleaning Data:** Our team cleaned and merged data from multiple CSV files, creating a Pandas DataFrame with country name, year, life expectancy, GDP, health expenditure, and CO2 emissions. Follow these steps: Prepare the environment, download the CSV files, save the script as a Python file, and call the clean_data() function that loads and merges data.

**Analyzing Data:** Our team analyzed cleaned data by plotting histograms, normal distributions, boxplots, and histograms with outliers removed. Follow these steps: Prepare the environment, download the CSV file, save the script as a Python file, and call the main() function that plots histograms and normal distributions for specified columns in the DataFrame.

**Linear Regression Model:** We used Statsmodel and sklearn packages to create linear regression models. Follow these steps: Load cleaned data, create a new column to store the prior GDP and calculate the GDP growth rate, train the linear regression models, predict life expectancy, and calculate the root mean squared error.

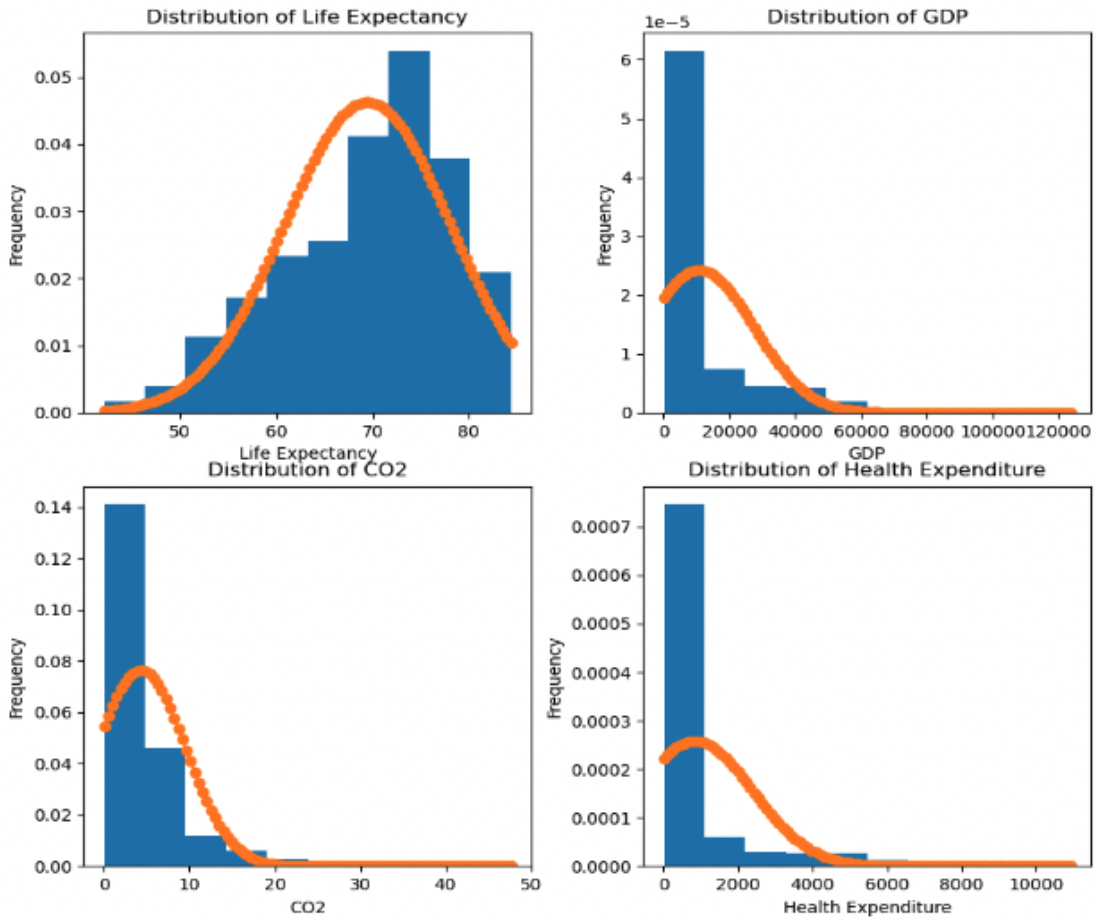
**Decision Tree Model:** We used Scikit-learn to create decision tree models for classifying life expectancy into different age groups based on independent variables. Follow these steps: Load cleaned data, classify the life expectancy column, filter the data, train and test the decision tree model, and plot the decision tree and feature importance graphs.

**Data Visualization:** We used Pandas, Plotly, Seaborn, and Matplotlib to visualize correlations between columns in the cleaned data set. Follow these steps: Prepare the environment, load the data, create scatter plots and heatmaps to visualize correlations.

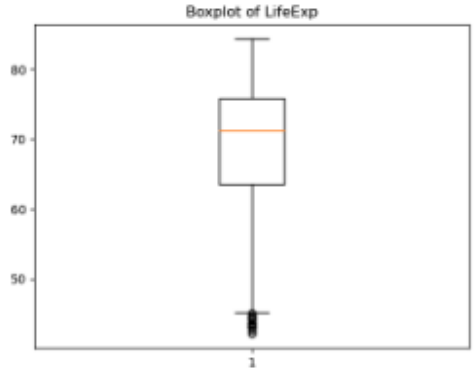
- **Results, visualizations, and key takeaways（这部分的标题放大，研究结果，可视化以及主要的发现，这句话加大加粗）：**

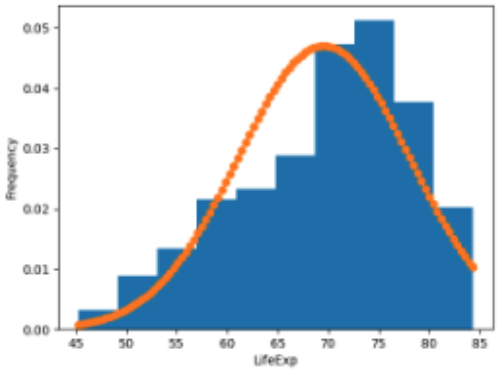
**1.After analyzing the cleaned data, it was found that（这句话加大加粗）：**

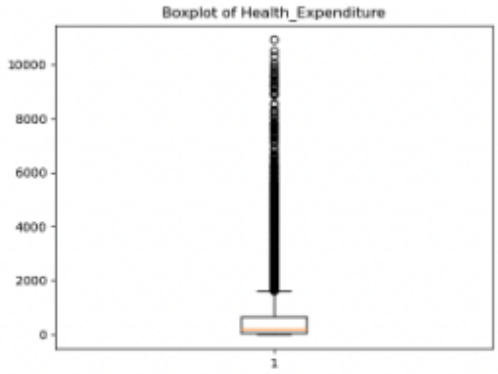
- Life expectancy, individual GDP, individual health expenditure, and individual CO2 emissions belong to skewed distributions.(对应下面的图）

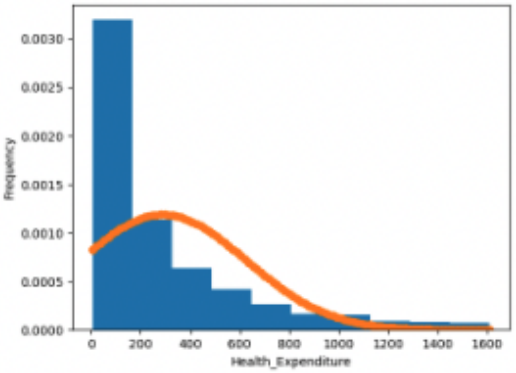



- Outliers were explored using box plots and frequency distribution plots for life expectancy and health expenditures.
- Life expectancy outliers were mainly in countries with poor medical and health conditions.
- Health expenditure outliers were due to the high cost of living in developed countries 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.（对应下面的图）



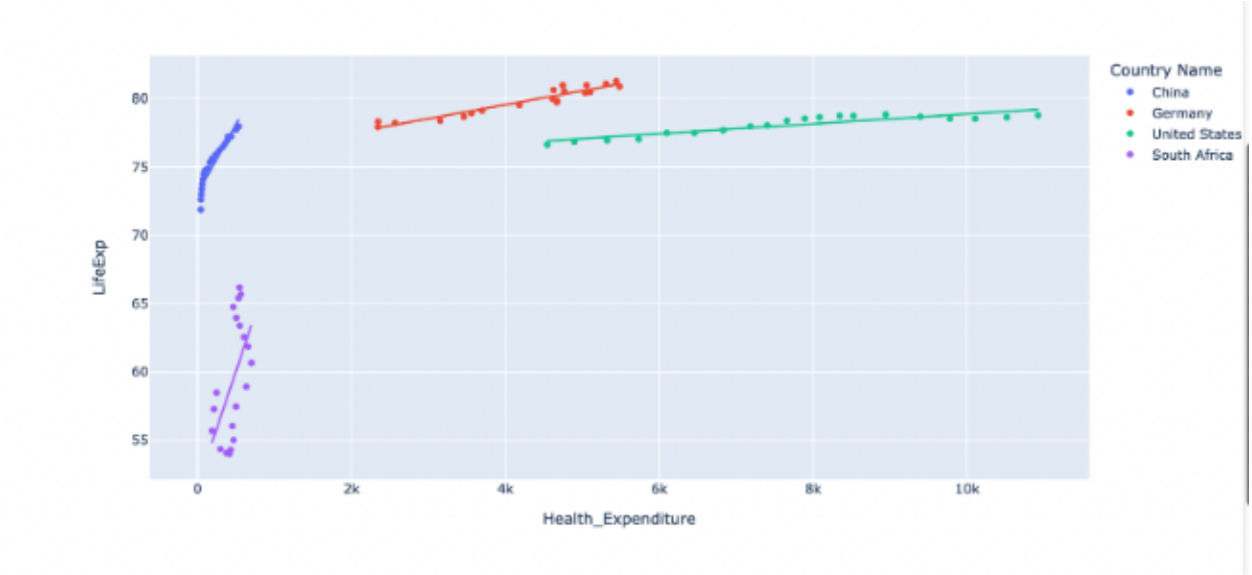



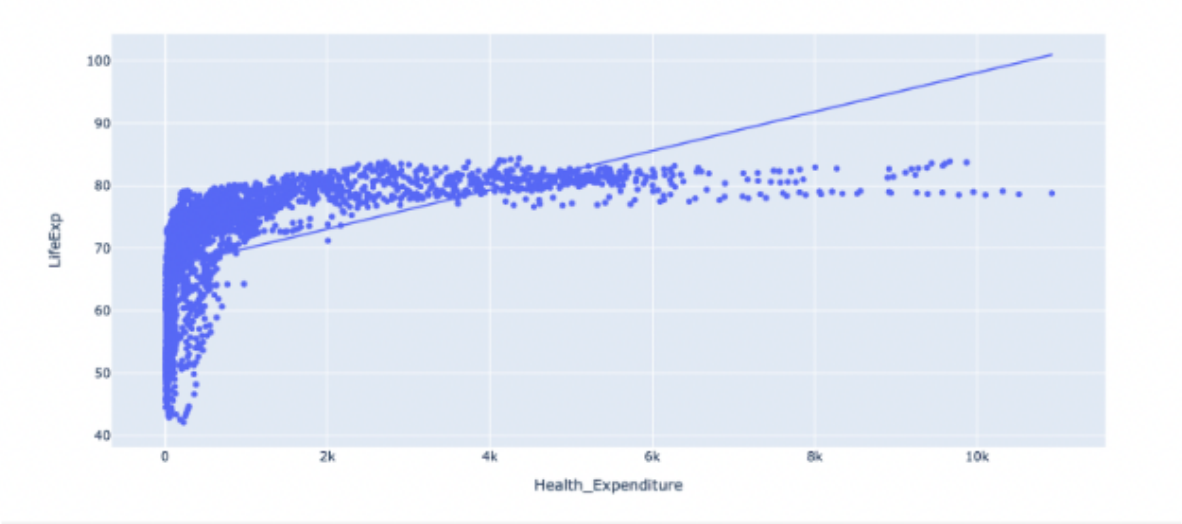

**2.1 After using the scatterplot, explore the relationship between health expenditure and life expectancy for four countries: the United States, China, Germany, and South Africa from 2000 to 2019, it found that（这句话加大加粗）:**

- Health spending and life expectancy show a positive correlation in all four nations.
- Prior research supports the positive correlation between health expenditures and health outcomes.
- Investing in healthcare is necessary to improve health outcomes and increase life expectancy.
- Germany has the highest healthcare expenditures and life expectancy, while South Africa has the lowest.
- Scatterplots may not provide enough information to draw conclusions, and further research is necessary to validate and quantify the association.(对应下面的图）




**2.2 After using the scatterplot, explore the relationship between health expenditure and life expectancy for 261 countries from 2000 to 2019, it found that（这句话加大加粗):**

- There may be a correlation between personal health expenditures and life expectancy, but it may not be straightforward or clear.
- Outliers, such as high health expenditures in developed countries, may have caused the uneven distribution of the scatterplot.
- The regression line is deviated and unrealistic, with unrealistic predictions of life expectancy beyond 100 years.(对应下面的图）




**After using a linear regression model, explore the relationship between health expenditure and life expectancy for 261 countries from 2000 to 2019, it found that（这句话加大加粗):**

- 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 66 years old.(对应下面的图）

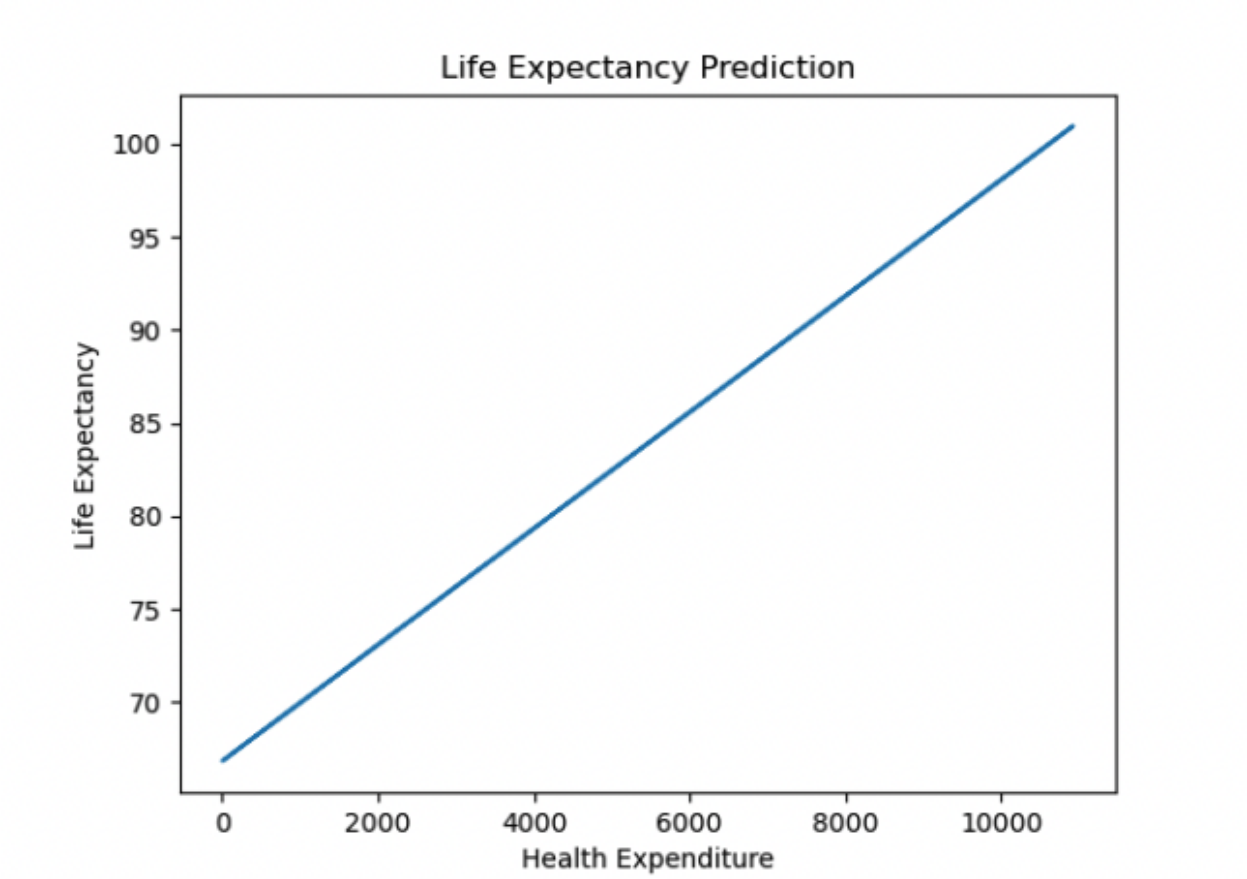
OLS Regression Results

| | | | |
|---|---|---|---|
| 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 |

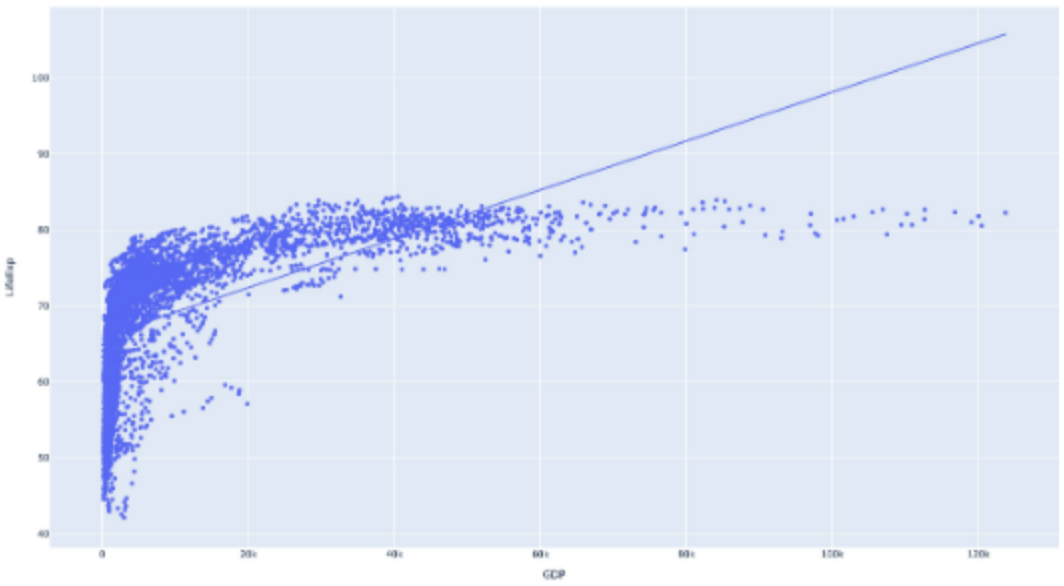
| | | | |
|---|---|---|---|
| Omnibus: | 470.081 | Durbin-Watson: | 1.740 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 629.996 |
| Skew: | -0.911 | Prob(JB): | 1.58e-137 |
| Kurtosis: | 3.153 | Cond. No. | 2.00e+03 |

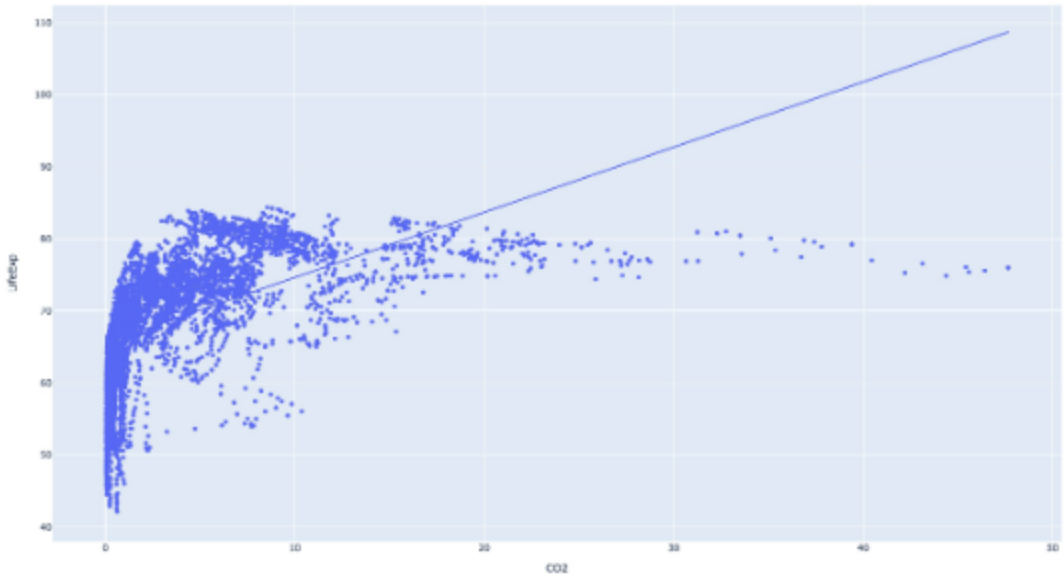
- The linear regression model shows a positive correlation between health expenditures and life expectancy.
- The line graph indicates that spending around $1000 on health is associated with an average life expectancy of 69 years.
- Higher health expenditures correspond to longer life expectancy, up to a maximum of 100 years.
- Other factors such as genetics and lifestyle may also impact life expectancy and were not accounted for in the model.
- The dataset was split into train and test data, with similar root mean square errors indicating reliable accuracy for life expectancy predictions.(对应下面的图)




**3.1 After using the scatterplots, explore the impact of individual CO2 emissions and individual GDP on the life expectancy of 261 countries from 2000 to 2019, we found that（这句话加大加粗):**

- Personal GDP and CO2 emissions are important factors that may impact life expectancy.
- The scatterplot between personal CO2 emissions and life expectancy also shows no obvious linear relationship.
- It is important to study how these variables together affect life expectancy.(对应下面的图）

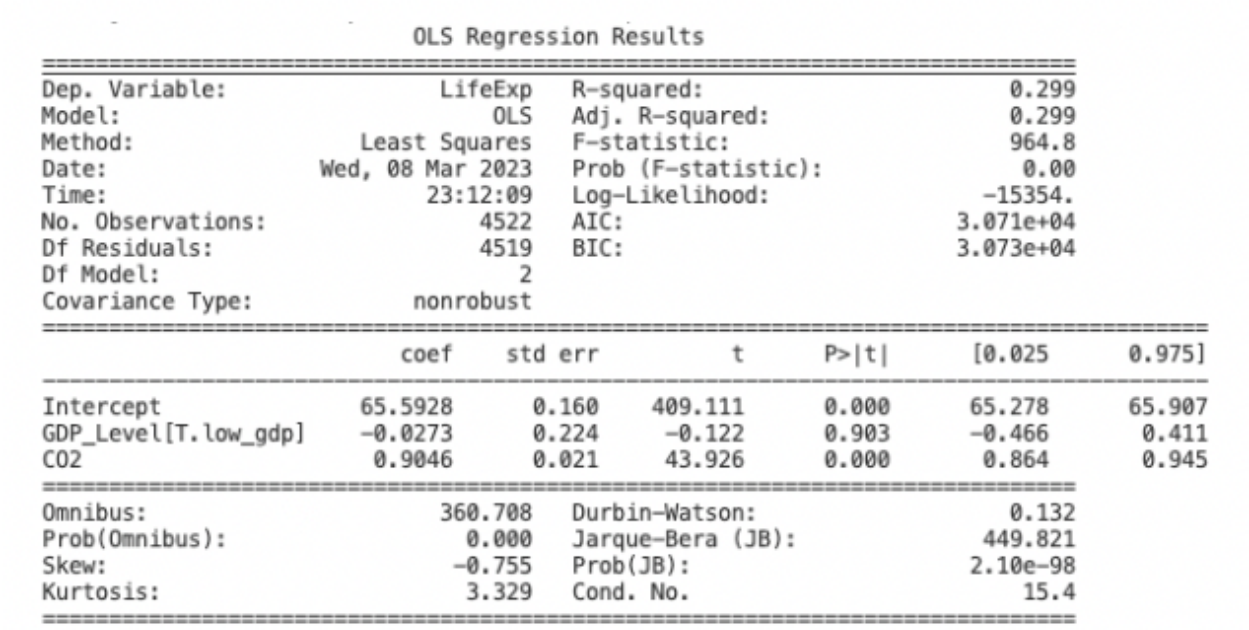



**3.2 After using the linear regression model, explore the impact of individual CO2 emissions and individual GDP on the life expectancy of 261 countries from 2000 to 2019, we found that（这句话加大加粗):**

- Two measures can be used to analyze the GDP level of a region: creating GDP growth rates or directly looking at GDP per capita.

- Initial statistical model including GDP level, CO2 emissions, and life expectancy had a non-significant lower level GDP coefficient (p-value = 0.903).
- The model cannot be used to predict life expectancy, and the relationship between GDP, CO2 emissions, and life expectancy should be approached with caution.
- The model suggests that there may be no significant relationship between GDP level and life expectancy, and that using GDP per capita values directly may be more reasonable to analyze this relationship.(对应下面的图)



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 |

- 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.(对应下面的图)

OLS Regression Results

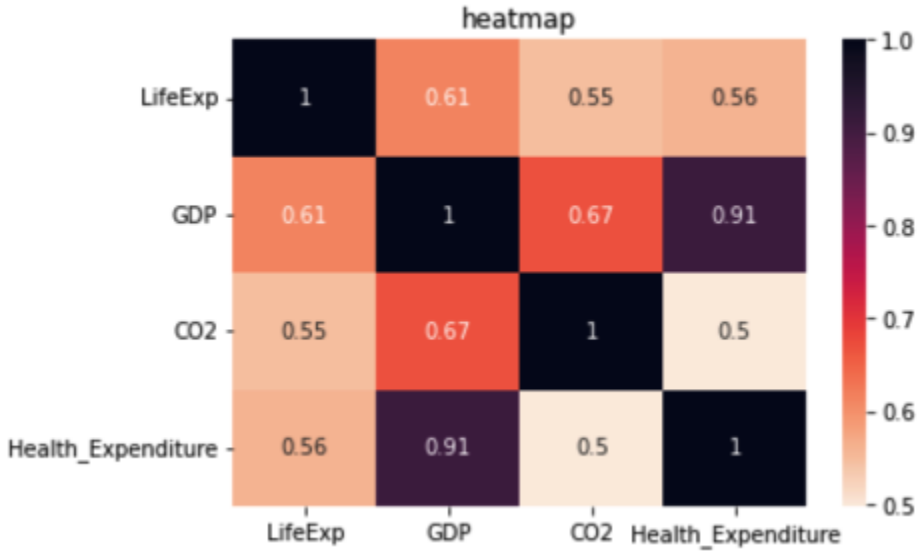
| | | | |
|---|---|---|---|
| Dep. Variable: | LifeExp | R-squared: | 0.410 |
| Model: | OLS | Adj. R-squared: | 0.410 |
| Method: | Least Squares | F-statistic: | 1571. |
| Date: | Wed, 08 Mar 2023 | Prob (F-statistic): | 0.00 |
| Time: | 23:12:09 | Log-Likelihood: | -14965. |
| No. Observations: | 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 |

**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（这句话加大加粗):**

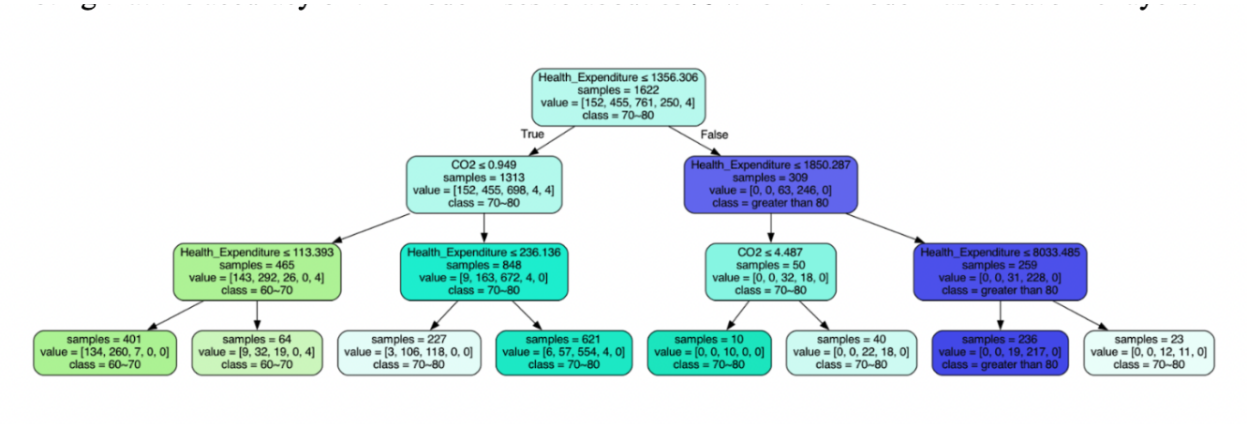
- A heat map was used to analyze the influence of life expectancy, GDP, CO2 emissions, and health expenditure on each other.
- Individual GDP has the largest influence on life expectancy, with a value of 0.61.
- CO2 emissions also have a positive correlation with life expectancy (0.55).
- Health expenditure has a weaker positive association with life expectancy (0.56) compared to GDP.(对应下面的图）




**4.2 After 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, it found that（这句话加大加粗):**

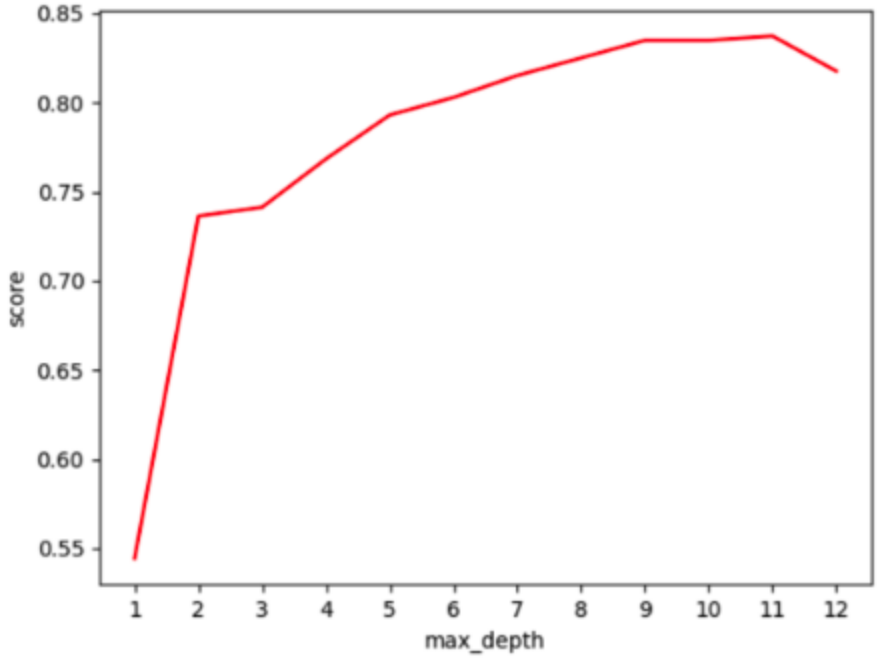
**Individual health expenditure has the greatest impact on individual life expectancy, followed by individual CO2 emissions, and individual GDP has the least impact.(这句话加大加粗，对应下面的图）**

- Decision tree analysis shows that health expenditure has the greatest impact on life expectancy.
- If health expenditure is <= 1356.306, the tree splits further based on CO2 emissions.
- Personal GDP does not have a significant impact on life expectancy.
- Increasing CO2 emissions may have a positive impact on life expectancy, but the increase is marginal (0.4742 years for each unit increase).

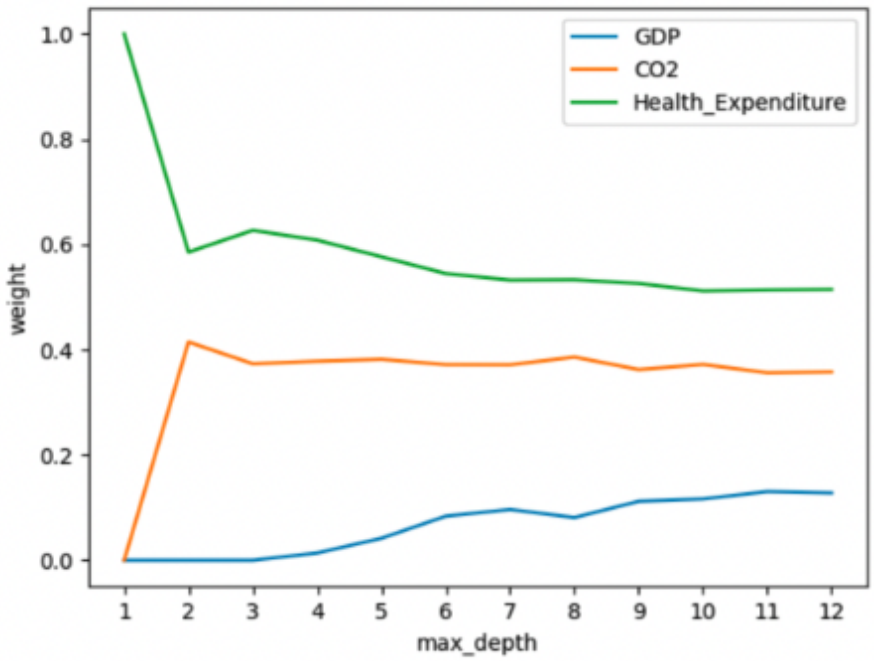



**Increasing the max depth of the decision tree model can improve its accuracy and feature importance, and the feature importance of each variable tends to be stable after a certain level of max depth.(这句话加大加粗，对应下面的图）**

- The linear graph shows the relationship between the max depth of the decision tree model and the accuracy of the model.
- At max depth 1, the accuracy is about 55%, but it significantly increases to about 74% at max depth 2-3.
- The accuracy of the model continues to increase up to about 83% at max depth 9-10.
- The graph demonstrates that the max depth of the model has a significant impact on the accuracy of the independent variables.(对应下面的图）




- Health expenditure has the greatest impact on life expectancy among all independent variables.
- Feature importance of health expenditure shows a downward trend at max depth 1-2, decreasing from 1.0 to 0.6, but maintains the highest importance with increasing max depth.
- Feature importance of all independent variables remain relatively stable after max depth 9 in the decision tree model.(对应下面的图）




- **Future work, or future directions for this project（这部分的标题放大，未来的工作，这句话加大加粗):**
- Findings can guide policymakers and public health professionals in making educated decisions about health expenditures, GDP, and CO2 emissions.
- Countries with high health spending outliers should modify policies to make healthcare more accessible.
- Countries with high CO2 emissions could consider implementing more sustainable practices.
- Individual health spending has the biggest influence on life expectancy, emphasizing the significance of healthcare access.
- The research provides valuable insights into the variables affecting life expectancy, but limitations should be considered when adopting the results for policymaking or research.