Name: Varshil Shah
UID: 2022301013
Batch: D

Aim: To create Advanced Charts using Python on socio-economic dataset

Theory:

Dataset: https://www.kaggle.com/datasets/ashydv/country-socioeconomic-data?select=Country-data.csv

Dataset Description:
This dataset contains socioeconomic and health indicators for 167 countries. The data provides insights into various factors that affect the overall development and well-being of these nations. The dataset includes metrics related to child mortality, economic performance, health expenditures, life expectancy, fertility rates, and income levels, among others. These variables are crucial for understanding the disparities in development across different regions and can be used for various analytical purposes, including identifying patterns, making comparisons, and predicting future trends.

Column Descriptions:

1. **Country** - The name of the country.
2. **Child_mort** - Number of deaths of children under the age of 5 per 1,000 live births.
3. **Exports** - Exports as a percentage of the country's Gross Domestic Product (GDP).
4. **Health** - Health expenditure as a percentage of GDP.
5. **Imports** - Imports as a percentage of the country's Gross Domestic Product (GDP).
6. **Income** - Average income per person, represented in monetary units.
7. **Inflation** - Annual inflation rate, reflecting the percentage increase in prices over a year.
8. **Life_expec** - Average life expectancy at birth in years.
9. **Total_fer** - Total fertility rate, representing the average number of children born to a woman over her lifetime.
10. **Gdpp** - Gross Domestic Product per capita, reflecting the average economic output per person.

Charts:

1. Word Cloud

Theory:

A word cloud is a visual representation of text data, where the size of each word indicates its frequency or importance within the dataset. Words that appear more frequently are displayed in larger, bolder fonts, while less frequent words are smaller. Word clouds are often used for quickly identifying the most prominent terms in a large corpus of text.

Strengths:

- **Visual Impact**: Easily highlights the most important words or terms in a dataset.
- **Simplicity**: Quickly communicates the main themes or topics without requiring complex analysis.
- **Customizability**: Colors, fonts, and layout can be customized to fit the context or enhance readability.

Limitations:

- **Lack of Context**: Word clouds do not provide context or show relationships between words.
- **Quantitative Precision**: The exact frequency of each word is not easily discernible from a word cloud.
- **Overemphasis on Common Words**: Frequent but potentially unimportant words might dominate the visualization.

2. Box and Whisker Plot

Theory:

A box and whisker plot (or box plot) is a graphical representation of the distribution of a dataset. It displays the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum values. The "box" represents the interquartile range (IQR), containing the middle 50% of the data, while the "whiskers" extend to the minimum and maximum values within 1.5 times the IQR. Outliers are often shown as individual points beyond the whiskers.

Strengths:

- **Summary of Data**: Provides a concise summary of the dataset's distribution, central tendency, and variability.
- **Outlier Detection**: Easily identifies outliers and extreme values in the dataset.
- **Comparison**: Useful for comparing distributions across different categories or groups.

Limitations:

- **Not Detailed**: Doesn't show the exact distribution or frequency of individual data points.
- **Less Effective for Small Datasets**: With small datasets, the box plot may not accurately represent the underlying distribution.

- **Potential Misinterpretation**: Misleading if the data distribution is heavily skewed, as the mean and median can be quite different.

3. Violin Plot

Theory:

A violin plot combines elements of a box plot and a kernel density plot. It shows the distribution of the data across different values by depicting the density of the data at different values along the y-axis. The width of the "violin" at different y-values indicates the density of the data, while the inner box plot shows the interquartile range and median.

Strengths:

- **Detailed Distribution**: Provides a richer understanding of data distribution than a box plot alone.
- **Comparison**: Effective for comparing distributions between different groups or categories.
- **Handles Complex Distributions**: Can represent multimodal distributions (distributions with multiple peaks).

Limitations:

- **Complexity**: Can be harder to interpret than a standard box plot, especially for non-experts.
- **Overplotting**: In cases of very small datasets, the plot may become overly complex or cluttered.
- **Subjective Interpretation**: The shape of the violin can sometimes lead to subjective interpretations of the data’s density.

4. Regression Plot - Linear Regression

Theory:

A linear regression plot is used to visualize the relationship between two continuous variables and to fit a linear model that describes this relationship. The plot typically shows data points scattered on a graph, with a line of best fit that minimizes the sum of the squared differences between the observed and predicted values.

Strengths:

- **Simplicity**: Linear regression is easy to understand and implement.
- **Interpretability**: Provides clear insights into the relationship between two variables, including the direction and strength of the relationship.

- **Predictive Power**: Useful for making predictions if the relationship between variables is approximately linear.

Limitations:

- **Assumption of Linearity**: Assumes a linear relationship between variables, which may not hold true in all cases.
- **Sensitivity to Outliers**: Outliers can significantly affect the model, leading to biased or inaccurate predictions.
- **Limited Flexibility**: Not suitable for modeling complex, non-linear relationships between variables.

5. Regression Plot - Logistic Regression

Theory:

Logistic regression is used to model the probability of a binary outcome (0 or 1, true or false) as a function of one or more independent variables. The logistic regression plot typically shows the probability of the outcome on the y-axis and the independent variable on the x-axis, with a sigmoid curve representing the probability function.

Strengths:

- **Probabilistic Interpretation**: Outputs probabilities, providing a measure of certainty about predictions.
- **Binary Classification**: Well-suited for binary classification problems, such as spam detection, disease diagnosis, etc.
- **Interpretable Coefficients**: The coefficients can be interpreted in terms of odds ratios, making the model’s predictions more interpretable.

Limitations:

- **Assumes a Linear Decision Boundary**: Assumes a linear relationship between the independent variables and the log-odds of the outcome, which may not be appropriate for all datasets.
- **Sensitivity to Outliers**: Like linear regression, logistic regression can be affected by outliers.
- **Not Suitable for Multi-Class Classification**: Logistic regression is inherently designed for binary outcomes, making it less suitable for multi-class classification without modification.

6. 3D Scatter Plot

Theory: A 3D scatter plot displays data points on three axes in a three-dimensional space. Each point represents an observation with three variables, typically represented by its x, y, and z coordinates.

Strengths:

- Visualizes relationships between three variables simultaneously.
- Can reveal patterns or clusters that might not be apparent in 2D visualizations.
- Useful for exploring complex datasets with multiple variables.

Limitations:

- Can be difficult to interpret on a 2D screen.
- May suffer from overplotting with large datasets.
- Challenging to read exact values from the plot.

Implementation: We used matplotlib's 3D plotting capabilities to create a scatter plot of GDP per capita, life expectancy, and total fertility rate.

7. Jitter Plot

Theory: A jitter plot is a variation of a scatter plot where a small amount of random noise is added to the data points to reduce overplotting.

Strengths:

- Reduces overplotting in datasets with many similar values.
- Reveals the density of data points more clearly than a standard scatter plot.
- Useful for visualizing the distribution of data along an axis.

Limitations:

- The added noise can slightly distort the perception of the true data points.
- May not be suitable for presentations where exact point locations are crucial.

Implementation: We used seaborn's stripplot function to create a jitter plot of child mortality vs income.

8. Line Plot

Theory: A line plot connects data points with lines, typically used to show trends over time or some other continuous variable.

Strengths:

- Excellent for showing trends and patterns over a continuous variable.
- Easy to compare multiple series on the same plot.

- Intuitive to read and interpret.

Limitations:

- Can become cluttered with too many lines.
- May not be suitable for categorical data or non-continuous variables.

Implementation: We created a line plot showing various metrics for the top 5 GDP countries.

9. Donut Chart

Theory: A donut chart is a variant of a pie chart with a hole in the center, used to show proportions of a whole.

Strengths:

- Visually appealing and space-efficient.
- Good for showing proportional data.
- The central space can be used for additional information.

Limitations:

- Difficult to compare slice sizes accurately.
- Not suitable for many categories or time-series data.

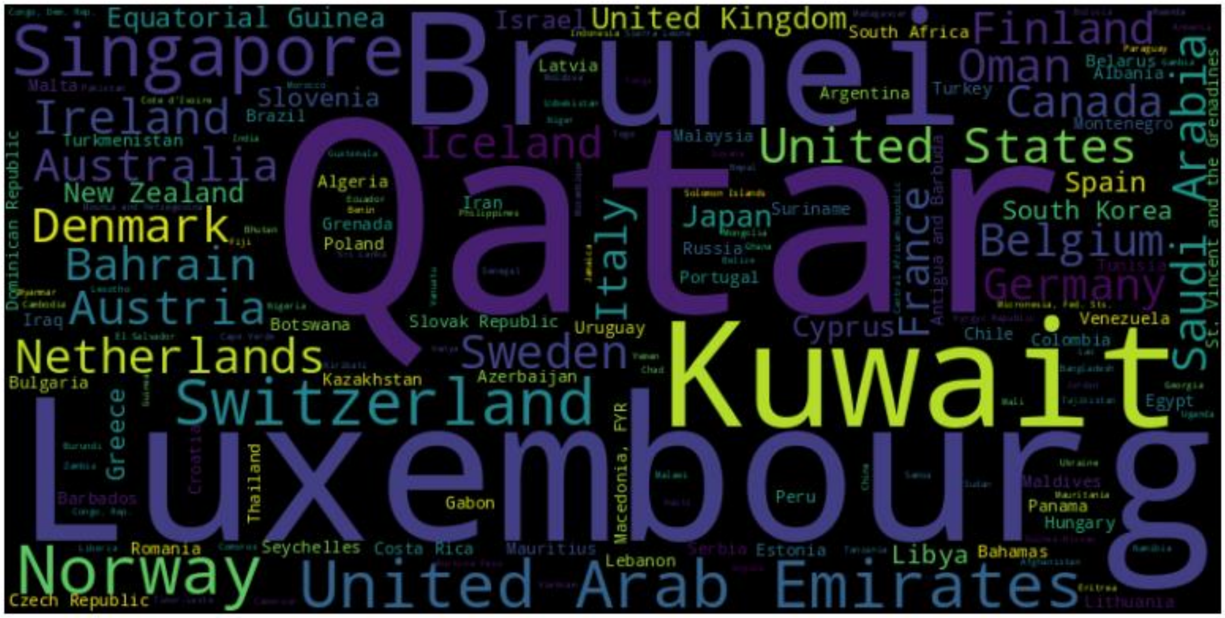
Implementation: We created a donut chart to show the distribution of countries across income groups.

Link to Notebook:
https://colab.research.google.com/drive/1CsTf0Xyqc27uVD6WXjlshsDqyiOtFGyO?usp=sharing

Graphs and Observations:

1. **Word Cloud**

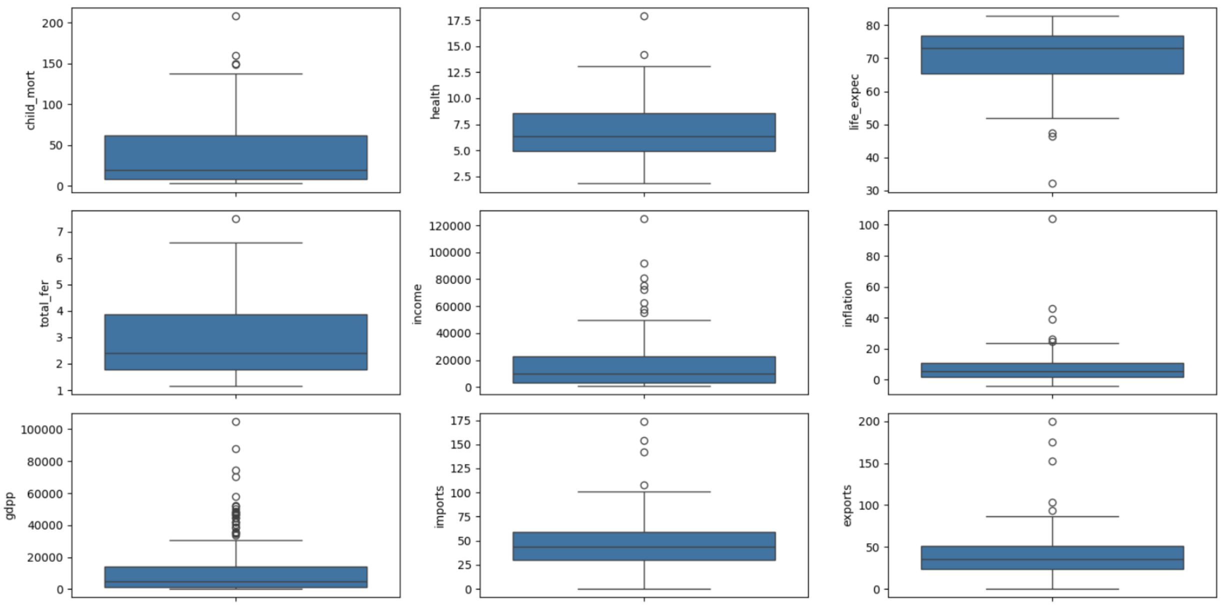
a. Chart:



b. Observations:
The word cloud highlights global income disparities, with countries like Qatar, Luxembourg, Kuwait, and Brunei standing out due to their exceptionally high income levels, as indicated by their large text size. Other economically strong nations like the United States, Japan, and Germany also appear prominently, though with slightly smaller text, reflecting substantial but not top-tier income levels. In contrast, smaller text sizes represent lower-income countries such as Equatorial Guinea and Malawi, showcasing the significant variation in economic strength across different regions.

2. **Box and Whisker Plot**:

a. Chart:



b. Observations:

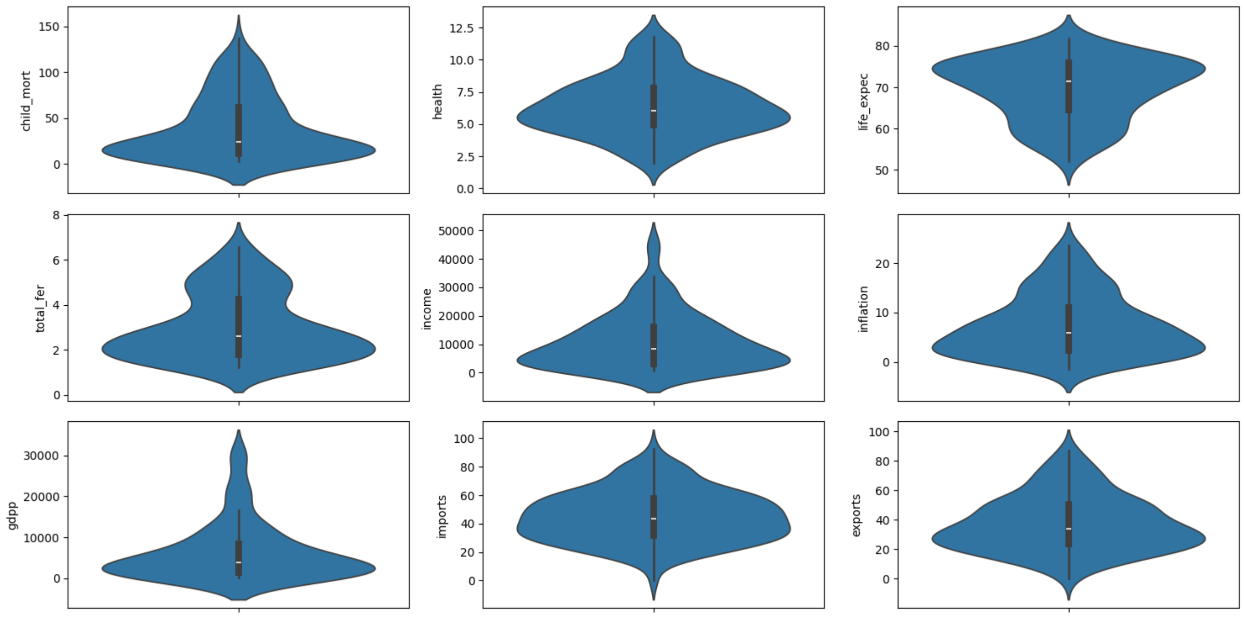
The boxplots illustrate the distribution and spread of various economic and social indicators across different countries. Key observations include:

1. **Child Mortality and Total Fertility**: These indicators show a wide range, with several outliers indicating countries with exceptionally high child mortality and fertility rates. However, most countries cluster around lower values for both.
2. **Health Expenditure and Life Expectancy**: Both indicators show a relatively normal distribution with some outliers. Most countries have moderate health expenditures, and life expectancy tends to be high across the majority of countries.
3. **Income and GDP per Capita (GDPP)**: These indicators exhibit significant variability with notable outliers, indicating that a few countries have exceptionally high income levels and GDP per capita, while most have much lower values.
4. **Inflation, Imports, and Exports**: These economic indicators have a narrower range for most countries, but with a few notable outliers, suggesting that while inflation, imports, and exports are stable in most countries, some experience extreme values in these areas.

Overall, the boxplots reveal that while most countries tend to have similar median values across these indicators, there are significant outliers that indicate substantial variability in economic and social conditions globally.

3. Violin Plot:

a. Chart:



b. Observations:

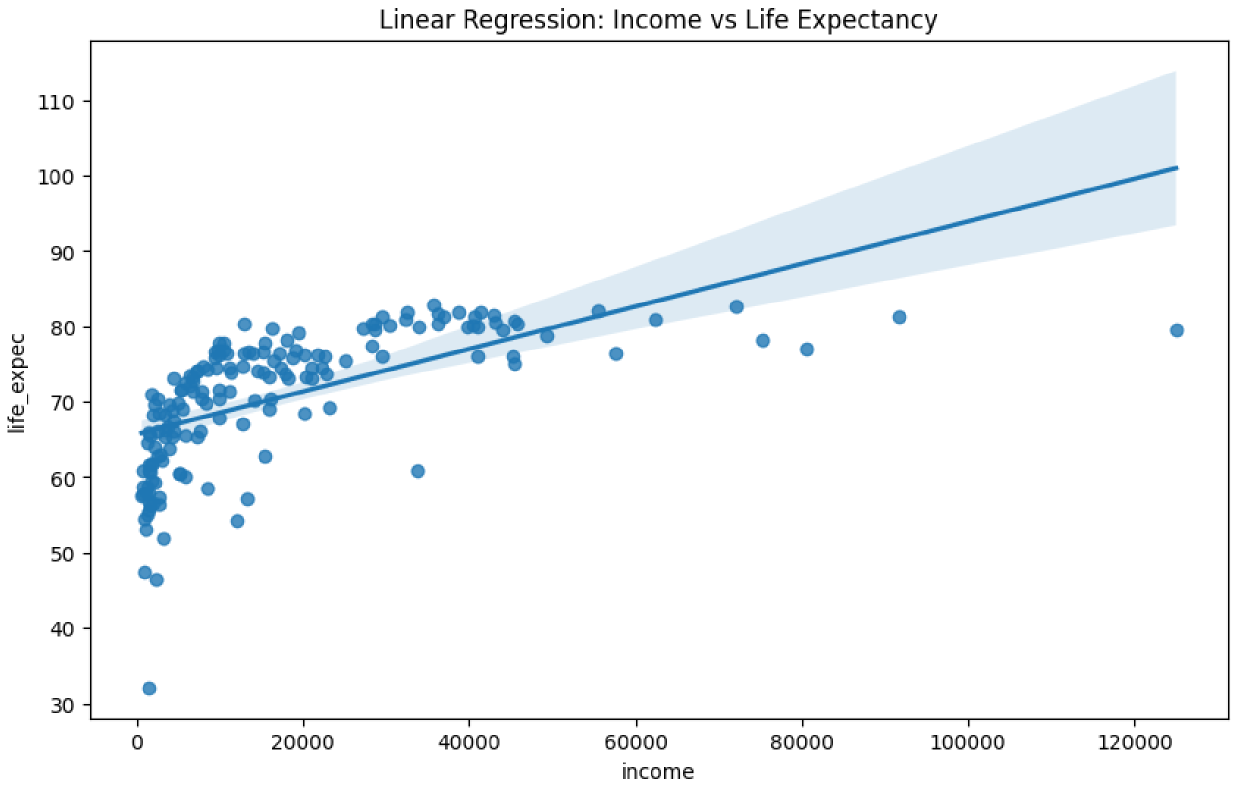
The violin plots provide a detailed view of the distribution of various economic and social indicators, excluding outliers.

1. **Child Mortality and Total Fertility**: These plots show that both indicators have a skewed distribution, with most countries clustering towards lower child mortality and fertility rates, but with a tail that suggests a few countries still have higher values.
2. **Health Expenditure and Life Expectancy**: Both indicators exhibit a relatively symmetrical distribution. Health expenditure is centered around moderate values, while life expectancy is predominantly high across most countries, with less variability.
3. **Income and GDP per Capita (GDPP)**: The income and GDP per capita plots reveal a more concentrated distribution after removing outliers, with most countries having low to moderate levels, and a peak indicating a higher concentration of countries at these levels.
4. **Inflation, Imports, and Exports**: These economic indicators show that most countries have stable values, with inflation showing a tighter distribution around lower values, while imports and exports have slightly wider distributions, indicating more variability among countries.

Overall, after removing the outliers, the violin plots reveal more centralized and less dispersed distributions for most indicators, suggesting that while there is still variability, the majority of countries fall within a narrower range of values for these economic and social measures.

4. Regression Plot - Linear Regression:

a. Chart:

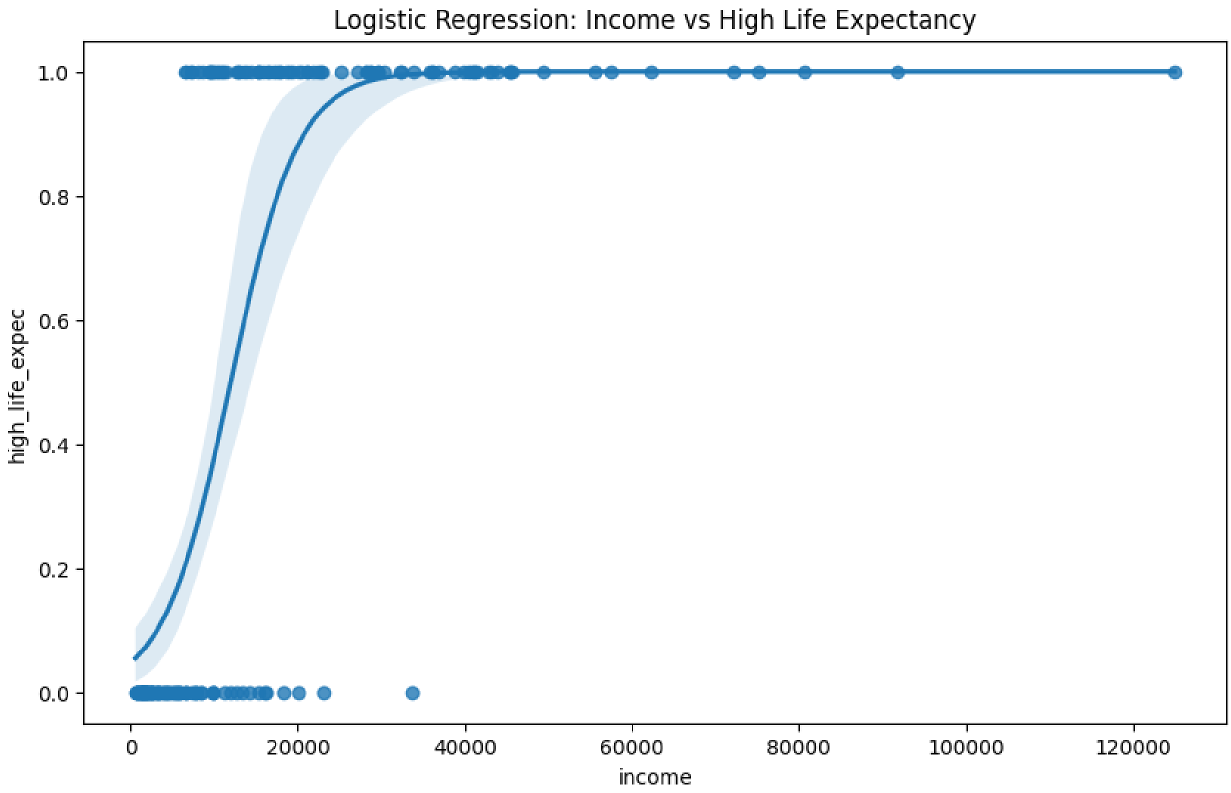


b. Observations:

This regression plot demonstrates the relationship between income and life expectancy across countries. The blue line represents the linear regression model, showing a positive correlation between the two variables. The shaded area around the line indicates the 95% confidence interval, which widens at the extremes where data is sparser. While the overall trend is positive, the scatter of data points suggests a non-linear relationship, with life expectancy increasing more rapidly at lower income levels before leveling off. The plot reveals considerable variation in life expectancy at similar income levels, particularly in the lower to middle income ranges, indicating that factors beyond income also play significant roles in determining life expectancy.

5. **Regression Plot - Logistic Regression**:

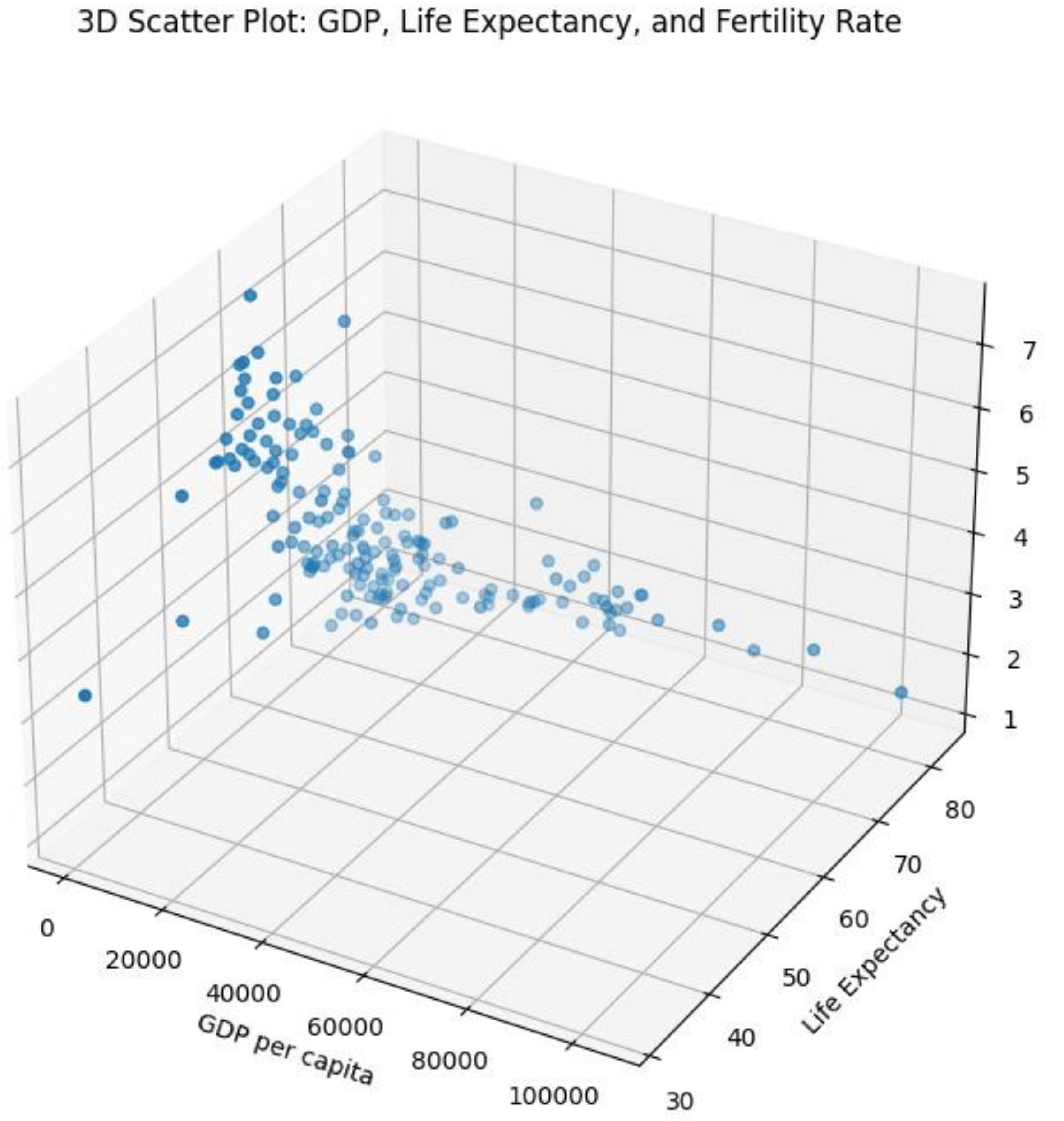
a. Chart:



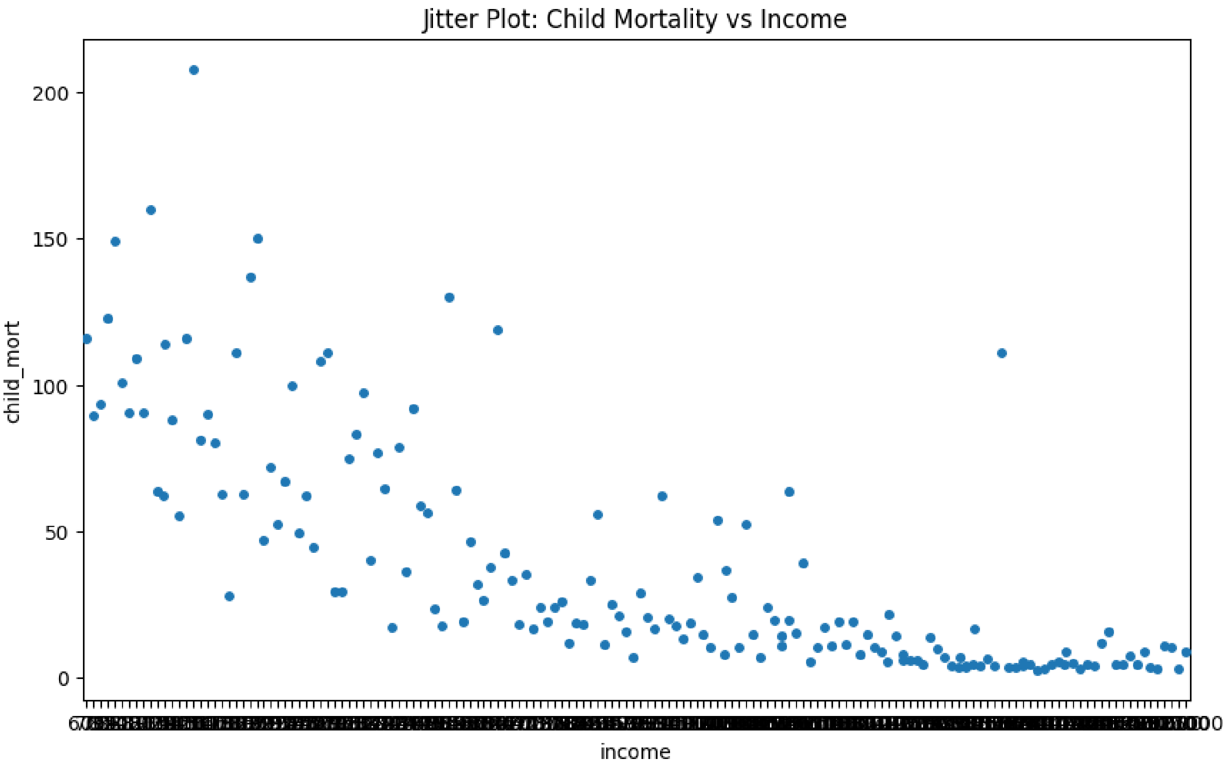
b. Observations:

This logistic regression plot illustrates the relationship between income and the probability of having high life expectancy. The S-shaped curve represents the logistic model, showing how the likelihood of high life expectancy changes with income. The plot demonstrates a clear threshold effect: at lower income levels, the probability of high life expectancy is near zero, but it rapidly increases around the middle-income range, eventually plateauing at higher incomes where the probability approaches 1. The shaded area represents the 95% confidence interval, which is wider in the middle income range where the transition occurs. Data points are clustered at 0 and 1 on the y-axis, indicating a binary classification of life expectancy (low/high). The steep slope in the middle suggests that moderate increases in income in this range are associated with significant improvements in the likelihood of high life expectancy, while further increases at high income levels yield diminishing returns.

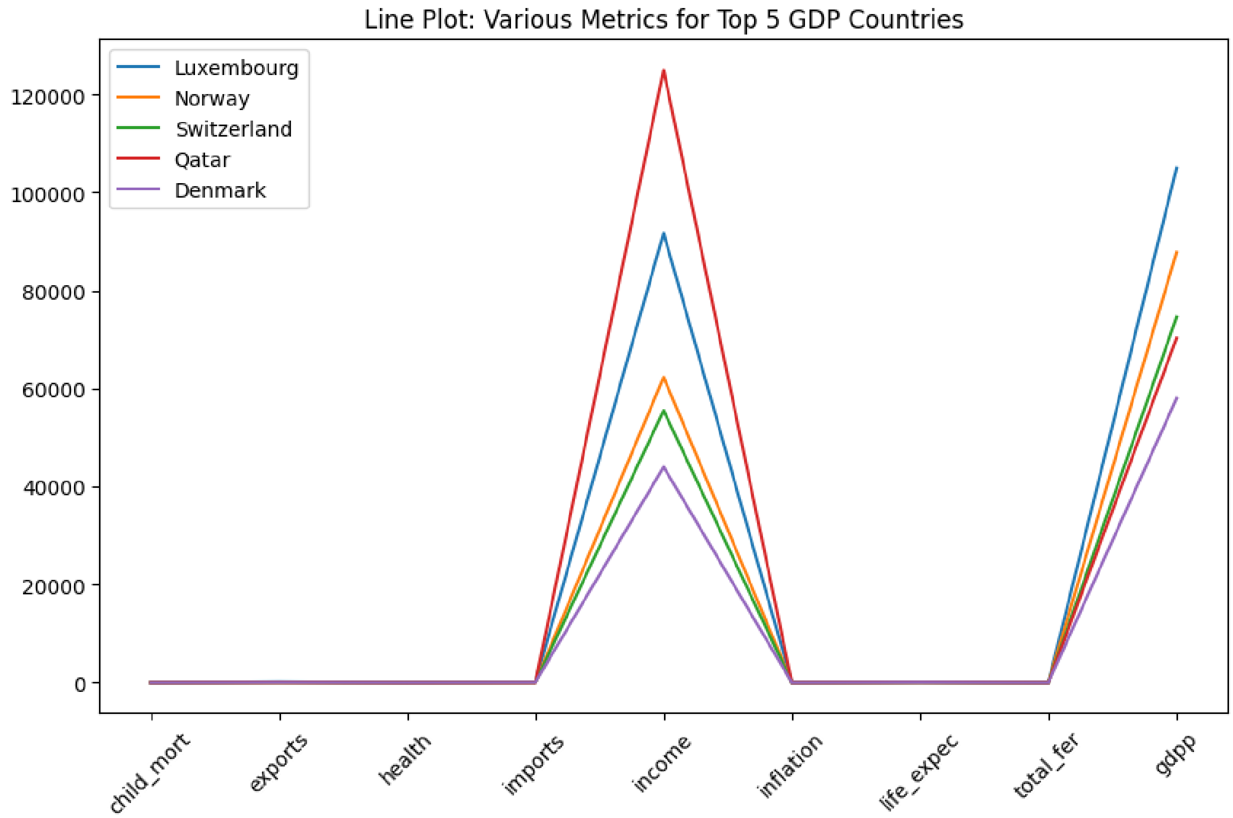
6. 3D Scatter plot -



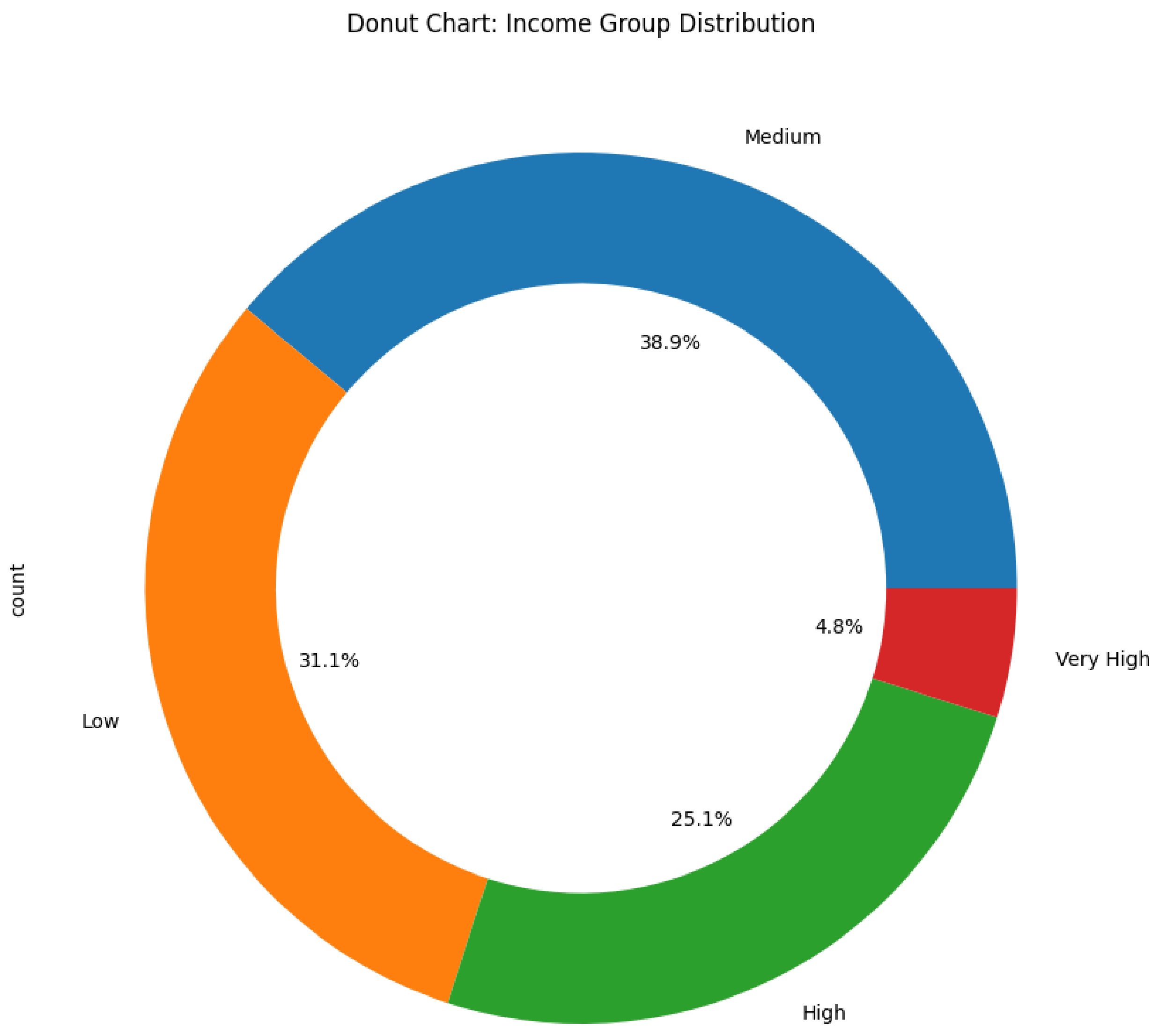
7. Jitter plot -


Jitter Plot: Child Mortality vs Income
child_mort
200
150
100
50
0
income

9. Line plot -



10. Donut chart -



Conclusion:

In this experiment, we utilized various Python visualization techniques to analyze a dataset of 167 countries, revealing key socioeconomic and health indicators. The word cloud effectively highlighted countries with higher incomes, providing a visual emphasis on wealth distribution. Box and whisker plots illustrated the range and variability of metrics like child mortality and health expenditure, highlighting both central tendencies and outliers. Violin plots offered a comprehensive view of the distribution and density of these variables, uncovering patterns and potential multimodal distributions. Finally, the linear regression plot showcased relationships between continuous variables, while logistic regression plots, although not executed here, would be ideal for exploring binary outcomes. Collectively, these visualizations enhanced our understanding of the data's structure and underlying trends.