Unit 4

Linear Regression with Scikit-Learn

Dataset – Fuel Consumption:

I began the analysis by uploading the *FuelConsumption.csv* dataset into the correlation regression demo. To understand the variables, I first inspected the dataset and computed a correlation matrix, which provided insights into relationships between features. A heatmap and graphs effectively visualised these correlations, revealing both positive and negative relationships.

Exploratory Data Analysis:

- **Fuel consumption vs CO2 emission**: Plotted to identify trends.
- **Engine size vs CO2 emission**: Analysed to uncover the relationship between engine size and emissions.
- **Cylinders vs CO2 emission**: Explored to determine the impact of the number of cylinders on emissions.

Building the Machine Learning Model:

The dataset was split into training and testing subsets. Using the training data, I developed a linear regression model with Scikit-Learn. The model was then evaluated to assess its predictive capabilities.

Model Evaluation:

- **Mean Absolute Error (MAE)**: 9,843.99, suggesting significant prediction errors.
- **Mean Squared Error (MSE)**: 102,476,097.63, confirming large deviations from the actual data.
- **R^2 Score**: -16.47, indicating a poor fit, where the model performs worse than predicting the mean value for all outcomes.

Observations:

- The metrics suggest that the model fails to capture the relationship between features and the target variable effectively.
- Potential issues include non-linear relationships, noise, outliers, or insufficient preprocessing (e.g., normalisation).

Recommendations for Improvement:

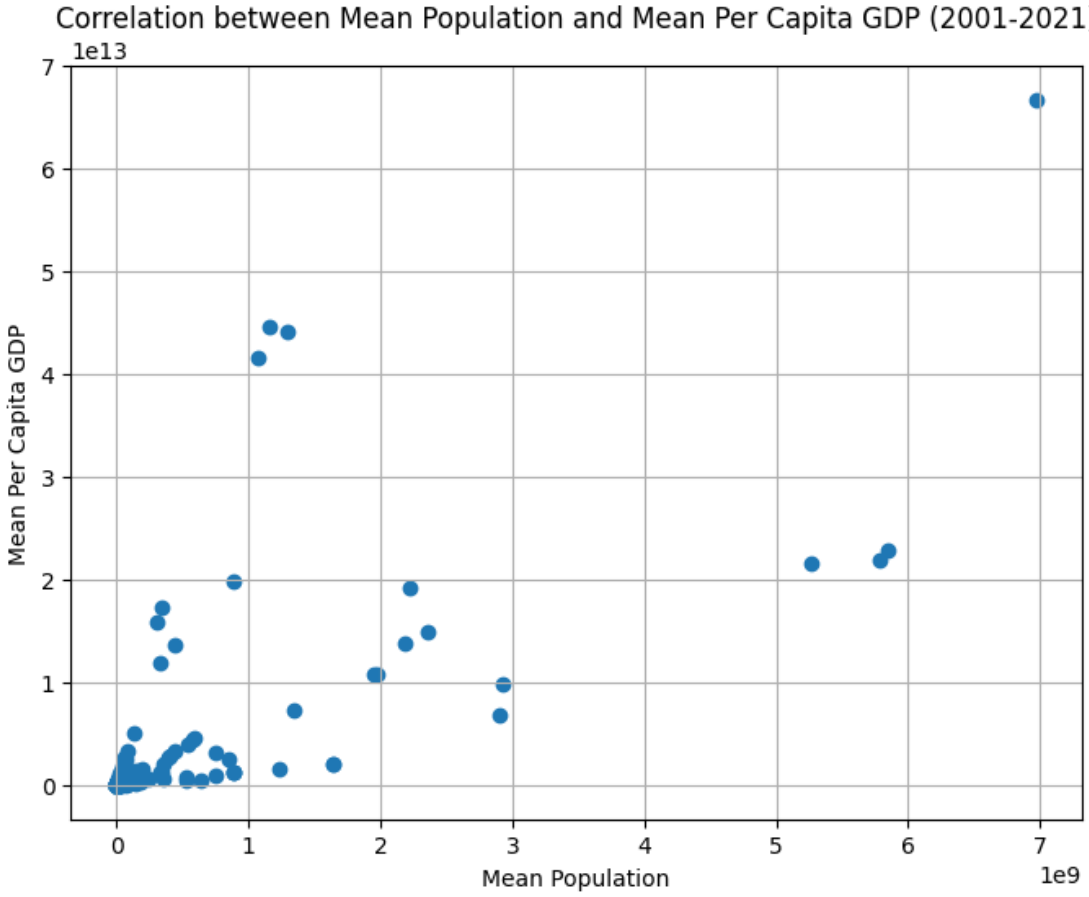
1. **Refine Features**: Reassess and transform features to better reflect underlying patterns.
2. **Address Data Issues**: Remove outliers and scale data for consistency.
3. **Explore Non-Linear Models**: Given the poor fit, consider using non-linear regression techniques.

Reflection:

This activity solidified my understanding of linear regression and its application. Despite the poor model performance, applying the theory and interpreting the results was rewarding, ensuring alignment with the learning objectives from this week's unit. The exercise demonstrated the importance of data preparation and model selection in predictive analysis.

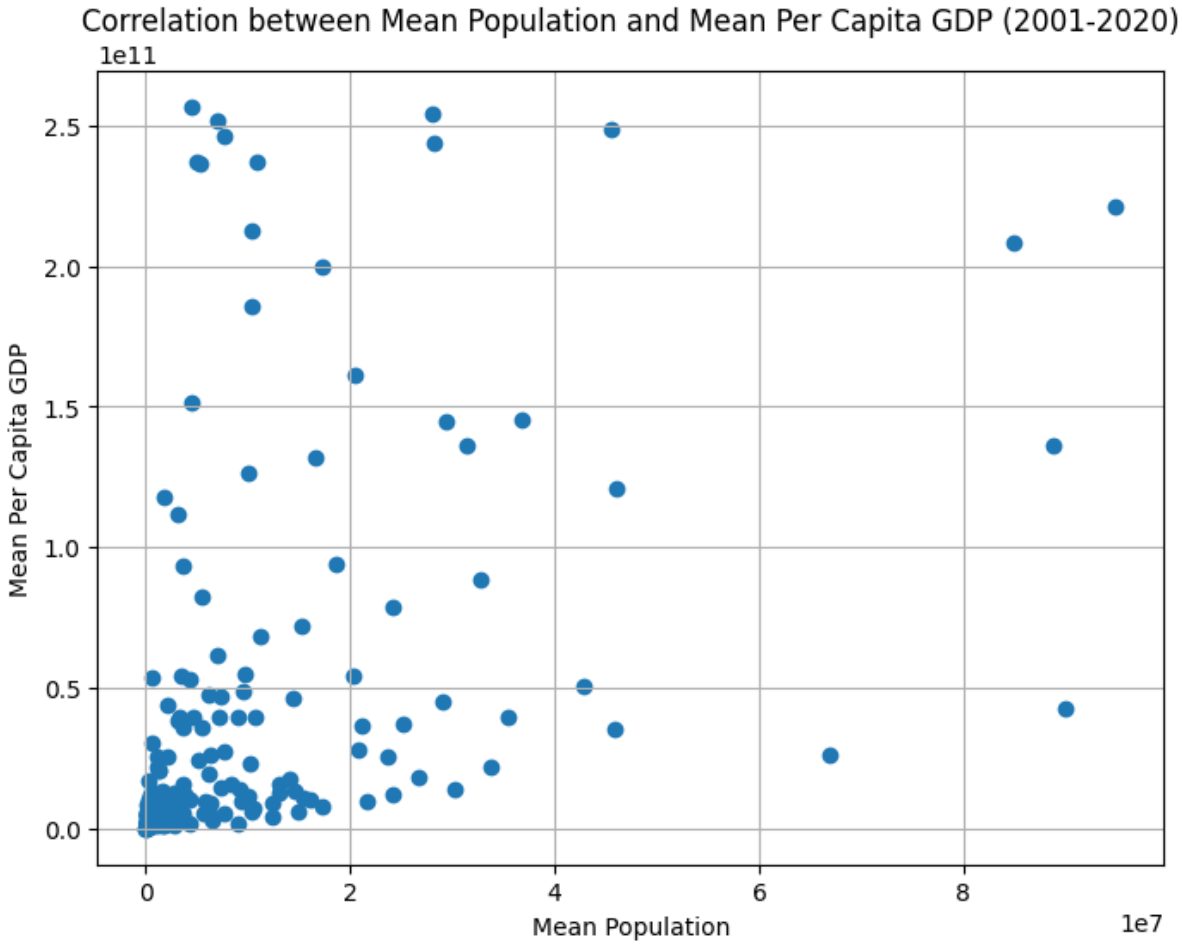
Task A:

For this task, I calculated the mean population and mean per capita GDP for each country from 2001 to 2021. Upon reviewing the datasets, I found that the years did not fully align. To address this, I limited the data to the overlapping period of 1961–2020 and merged the datasets. Subsequently, I plotted the mean population against the mean per capita GDP, resulting in the following graph.



Correlation Between Mean Population and Mean Per Capita GDP (2001–2021): 0.7213.

This indicates a strong positive correlation between the two variables. However, the data revealed several outliers, prompting their removal and the creation of a revised plot for more accurate analysis.



Correlation Between Mean Population and Mean Per Capita GDP (2001–2020): 0.4299.

This weaker positive correlation suggests that removing outliers significantly impacts the relationship between the variables, highlighting the importance of these outliers in the dataset.

To further investigate, I examined the influence of missing data by removing all NA values from the merged dataset. Recalculating the Pearson correlation yielded a stronger value of 0.721, emphasizing the role of complete data in producing more accurate correlations.

TASK B:

In this task, I developed a linear regression model using the two provided datasets.

Steps Taken:

1. **Data Preparation:**
   I started by ensuring the datasets were in the correct format for regression analysis. This involved cleaning and organizing the data before splitting it into training and test sets.

2. **Model Training:**
   The model was trained using the training data, and I extracted the following parameters:
   - **Coefficient (Slope):** 6023.90
   - **Intercept:** 477,219,583,395.84

Explanation:

- **Coefficient (Slope):** The coefficient represents the change in the dependent variable (mean per capita GDP) for each one-unit change in the independent variable (mean population). In this case, for every unit increase in the mean population, the per capita GDP is expected to increase by approximately 6023.90 units.
- **Intercept:** The intercept is the predicted value of the dependent variable when the independent variable is zero. Here, it suggests that when the mean population is zero, the model predicts a per capita GDP of 477,219,583,395.84, although this might not be meaningful due to the unrealistic scenario of a zero population.

These values are crucial for understanding the model's relationship between the variables. However, the next step would be to evaluate the model's performance using metrics such as R^2, Mean Squared Error (MSE), and Mean Absolute Error (MAE) to determine how well the model fits the data and makes predictions.

My linear regression model shows significant performance issues:

- **MAE (Mean Absolute Error):** 956 billion units, indicating large prediction errors.
- **MSE (Mean Squared Error):** Extremely high (5.15×10^{24}), confirming significant prediction discrepancies.
- **R^2 Score:** 0.201, suggesting the model explains only 20.1% of the variance in the data, showing a poor fit.

Likely Causes:

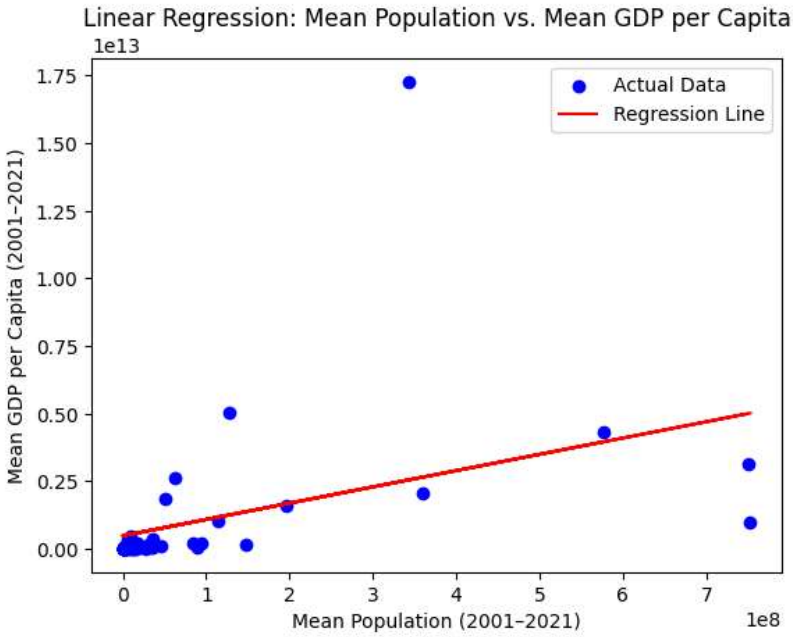
1. Weak or no linear relationship between population and GDP per capita.
2. Population alone may be insufficient to predict GDP; additional features are needed.
3. Data scale discrepancies and potential outliers might skew results.

Recommendations:

1. **Visualize and Examine Correlation:** Use scatterplots and calculate correlation coefficients.

2. **Feature Engineering:** Add relevant predictors (e.g., economic or demographic data).
3. **Preprocess Data:** Address scaling and outliers.
4. **Consider Non-Linear Models:** Explore polynomial regression or other methods.

This is the visualisation of the model:



Reflection:

I initially found Task A somewhat challenging, especially when it came to understanding the exact requirements of the question. My initial assumption was that there would be no significant correlation between the mean population and mean GDP for the countries in the dataset. I also mistakenly thought the question was asking about GDP trends over time, rather than the relationship between mean population and GDP.

This bias in my thinking led me to focus on a different form of analysis, which caused some confusion. After revisiting the task, I realised the focus was on understanding the correlation between the mean population and the mean GDP for each country between 2001 and 2021, not about GDP trends over time.

This shift in perspective was an important lesson for me, highlighting how easy it is to make assumptions based on prior knowledge and experiences. It reminded me that in data analysis, it is crucial to fully comprehend the task at hand before diving into calculations or drawing conclusions. This experience reinforced the importance of staying open-minded and being clear about the objectives of the analysis.