CSST 104 - Advanced Machine Learning

FINAL PROJECT
MACHINE LEARNING IMPLEMENTATION

Topic:
IQ Levels Analysis

Submitted by:
Jareen Pia B. Andres
Edrian B. Flores
Andrel John M. Pantanoza

Table of CONTENTS

Table of CONTENTS

Data Analysis and Machine Learning Implementation

Project Documentation Template

I. Project Overview

Purpose: The purpose of this project is to analyze the relationship between IQ levels and various socioeconomic and environmental factors across different countries. By examining these relationships, the aim is to uncover meaningful insights that are yet to be analyzed regarding the differences between certain factors and how it influences/benefits per country.

Dataset: We are provided with a dataset that contains information about IQ levels across various countries. The dataset includes the following columns:

- **Rank:** The rank of the country based on IQ levels.
- **Country:** The name of the country.
- **IQ:** The average IQ level of the country.
- **Education Expenditure:** The expenditure on education in the country.
- **Average Income:** The average income in the country.
- **Average Temperature:** The average temperature in the country.

The dataset consists of 108 rows, with each row representing a different country.

Specific User Attributes/Data Points Analyzed:

- **IQ Levels:** Analyzing the cognitive capabilities across different regions.

- **Education Expenditure:** Exploring the correlation between spending on education and IQ levels.
- **Average Income:** Investigating the relationship between average income and IQ levels.
- **Average Temperature:** Examining the potential impact of climate on IQ levels (if it has any connection at all).

Main Goals:

1. To analyze the distribution of IQ levels across different countries.
2. To identify significant correlations between IQ levels and other variables such as education expenditure, average income, and average temperature.
3. To visualize the data to highlight key insights and patterns.
4. To apply machine learning techniques to predict IQ levels based on the other attributes in the dataset.

Expected Insights:

- A comprehensive understanding of the global distribution of IQ levels.
- Insights into the impact of education expenditure on IQ.
- Analysis of how average income correlates with IQ levels.
- Examination of any notable effects of average temperature on IQ levels.
- Assessment of machine learning models' accuracy in predicting IQ levels.

II. Libraries and Data Handling

Libraries Used:

```
# Libraries and Data Handling
print("\nLibraries and Data Handling")
print("==========================")
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler

print("Libraries: Pandas (data manipulation), NumPy (numerical operations), Seaborn & Matplotlib (data visualization), Scikit-learn (machine learning)")
```

```
Libraries and Data Handling
==========================
Libraries: Pandas (data manipulation), NumPy (numerical operations), Seaborn & Matplotlib (data visualization), Scikit-learn (machine learning)

First few rows of the dataset:
   rank     country   IQ  education_expenditure  avg_income  avg_temp
0     1   Hong Kong  106                 1283.0     35304.0      26.2
1     2       Japan  106                 1340.0     40964.0      19.2
2     3   Singapore  106                 1428.0     41100.0      31.5
3     4      Taiwan  106                    NaN         NaN      26.9
4     5       China  104                  183.0      4654.0      19.1
```

Various libraries to be used were imported directly to the Colab notebook. Explanation and usage for the following libraries are as follows:

Pandas:

This is an open-source data analysis and manipulation library that also provides data structures like DataFrames, for this project, Pandas was used for data loading, cleaning, preprocessing, and initial exploratory data analysis (EDA).

NumPy:

NumPy is a fundamental package for scientific computing as it provides support for arrays, matrices, and a large collection of mathematical functions that allows operating on these data

structures, in this context, it was used for numerical operations and handling arrays, which are often required in data manipulation and preprocessing steps.

Seaborn:

Seaborn, a Python data visualization library based on Matplotlib. It provides a high-level interface for creating informative statistical graphics.

For this project, it was used to create various types of visualizations, such as scatter plots, bar charts, and correlation heatmaps, to allow proper visualization in the exploratory data analysis and presentation of insights.

Matplotlib:

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python, however, on this project, it was mainly used for creating and customizing plots, those that Seaborn could not easily provide.

Scikit-learn:

Scikit-learn was used for machine learning tasks, including data splitting, model building, evaluation, and feature scaling.

Modules and Description under Scikit-Learn:

train_test_split: A function to split the dataset into training and testing sets. **LinearRegression:** A module to implement linear regression models. **mean_squared_error:** A function to calculate the mean squared error of the model. **r2_score:** A function to calculate the R-squared score of the model. **StandardScaler:** A preprocessing module to standardize features by removing the mean and scaling to unit variance.

Data Loading and Preprocessing :

Data Loading:

The dataset was uploaded from the local directory up to the Colab notebook and then was loaded using the Pandas library from a CSV file named '08_IQ Levels Analysis.csv'.

```
# Load the dataset
file_path = '/08_IQ Levels Analysis.csv'

# Ensure the CSV is read correctly, handling special characters and encoding issues
data = pd.read_csv(file_path, encoding='utf-8')
```

Checking for Missing values:

The dataset was then checked for any missing values, with the next step being the preparation of handling missing values if any were detected.

```
# Check for any missing values
print("\nMissing values in the dataset:")
print(data.isnull().sum())

# Handling the missing values (if any were detected)
data['education_expenditure'].fillna(data['education_expenditure'].mean(), inplace=True)
data['avg_income'].fillna(data['avg_income'].mean(), inplace=True)
```

```
Missing values in the dataset:
rank                    0
country                 0
IQ                      0
education_expenditure   5
avg_income              2
avg_temp                0
dtype: int64
```

Checking data type:

Checking if columns contain categorical or numerical data. This allows us to plan how to handle the process of data analysis.

```
# Print and Check the data types
print("\nData types of the dataset:")
print(data.dtypes)
```

```
Data types of the dataset:
rank                      int64
country                  object
IQ                        int64
education_expenditure    float64
avg_income               float64
avg_temp                 float64
dtype: object
```

The next step is to remove non-numeric data before proceeding with the actual analysis.

```
# Removing the non-numeric columns before analysis
numeric_data = data.select_dtypes(include=[np.number])
```

Final step for this part was to simply calculate and print descriptive statistics (count, mean, standard deviation, min, max, quartiles).

```
# Descriptive statistics of the numeric columns
print("\nDescriptive statistics of numeric columns:")
print(numeric_data.describe())
```

```
Descriptive statistics of numeric columns:
             rank          IQ  education_expenditure     avg_income    avg_temp
count  108.00000  108.000000             108.000000     108.000000  108.000000
mean    54.50000   85.972222             903.058252   17174.650943   23.858333
std     31.32092   12.998532            1139.042128   20675.115731    8.392232
min      1.00000   51.000000               1.000000     316.000000    0.400000
25%     27.75000   78.750000              90.000000    2307.750000   17.250000
50%     54.50000   88.000000             411.000000    7605.500000   25.850000
75%     81.25000   97.000000            1297.250000   29838.000000   31.275000
max    108.00000  106.000000            5436.000000  108349.000000   36.500000
```

III. Data Analysis Techniques

Here, we calculate summary statistics for IQ levels, education expenditure, average income, and average temperature.

```
# Data Analysis Techniques
print("\nData Analysis Techniques")
print("========================")
# Descriptive statistics to understand the data distribution
iq_distribution = numeric_data['IQ'].describe()
education_expenditure_distribution = numeric_data['education_expenditure'].describe()
avg_income_distribution = numeric_data['avg_income'].describe()
avg_temp_distribution = numeric_data['avg_temp'].describe()

print("\nIQ Distribution:")
print(iq_distribution)
print("\nEducation Expenditure Distribution:")
print(education_expenditure_distribution)
print("\nAverage Income Distribution:")
print(avg_income_distribution)
print("\nAverage Temperature Distribution:")
print(avg_temp_distribution)
```

We make use of the function describe() that provides summary statistics of each column such as count, mean, standard deviation, minimum, and maximum values, and the quartiles (25%, 50%, 75%).

```
IQ Distribution:
count    108.000000
mean      85.972222
std       12.998532
min       51.000000
25%       78.750000
50%       88.000000
75%       97.000000
max      106.000000
Name: IQ, dtype: float64

Education Expenditure Distribution:
count     108.000000
mean      903.058252
std      1139.042128
min         1.000000
25%        90.000000
50%       411.000000
75%      1297.250000
max      5436.000000
Name: education_expenditure, dtype: float64

Average Income Distribution:
count       108.000000
mean      17174.650943
std       20675.115731
min         316.000000
25%        2307.750000
50%        7605.500000
75%       29838.000000
max      108349.000000
Name: avg_income, dtype: float64
```

```
Average Temperature Distribution:
count    108.000000
mean      23.858333
std        8.392232
min        0.400000
25%       17.250000
50%       25.850000
75%       31.275000
max       36.500000
Name: avg_temp, dtype: float64

Model Evaluation:
Mean Squared Error: 68.31731478150735
R-squared: 0.5670230940413582
```

These statistics now then give us a quick overview of the data distribution, central tendencies, and spread. It helps in identifying any anomalies or outliers and understanding the general characteristics of the data.

In this project, we use inferential statistics to predict IQ levels based on education expenditure, average income, and average temperature. We prepare the data by selecting these features as independent variables (X) and IQ as the dependent variable (y). Using train_test_split(), we split the dataset into training (80%) and testing (20%) sets, ensuring reproducibility with a fixed random state.

```
# Prepare the data for modeling
X = data[['education_expenditure', 'avg_income', 'avg_temp']]
y = data['IQ']

# Split data into training and testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

We then applied feature scaling with StandardScaler() to standardize the features. A linear regression model is then built and trained on the training data, with predictions made on the test data. The model's performance is evaluated using Mean Squared Error (MSE) and R-squared (R^2), which assess the model's accuracy and the proportion of variance explained. This process helps us understand relationships between variables and make predictions about IQ levels.

```
# Feature scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Build the machine learning model
model = LinearRegression()
model.fit(X_train_scaled, y_train)

# Make predictions
y_pred = model.predict(X_test_scaled)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("\nModel Evaluation:")
print(f"Mean Squared Error: {mse}")
print(f"R-squared: {r2}")
```

IV. Key Findings

Findings:

1. Positive correlation between education expenditure and average IQ levels.

Our analysis revealed a positive correlation between education expenditure (as a percentage of GDP) and average IQ levels. Countries that allocate a higher proportion of their GDP to education tend to have higher average IQ levels. This suggests that investment in education plays a critical role in enhancing cognitive abilities at a national level.

2. Higher average income is associated with higher average IQ levels.

The data indicates a strong positive relationship between average income and IQ levels. Countries with higher average incomes generally exhibit higher IQ levels. This correlation implies that economic prosperity may contribute to better access to resources that enhance cognitive development, such as quality education, nutrition, and healthcare.

3. Average temperature shows a less significant correlation with average IQ levels.

The analysis found that average temperature has a less significant correlation with IQ levels compared to education expenditure and average income. While there might be some climatic influence on cognitive abilities, it appears to be much weaker than the socioeconomic factors analyzed.

Business Impact:

The findings suggest that investing in education could have a significant impact on improving cognitive abilities at the national level. Policymakers can use this information to allocate resources effectively. Additionally, understanding the role of socioeconomic factors can help in designing targeted interventions to enhance IQ levels in lower-income regions.

These findings carry substantial implications as they underscore the strategic importance of increased investment in education, as it can lead to improved national cognitive abilities, benefiting sectors such as educational services and technology. Furthermore, designing targeted interventions for lower-income regions to enhance educational access and quality can create opportunities for businesses in those areas. Policymakers can leverage these insights to allocate resources effectively, prioritizing educational expenditure for long-term economic and social benefits. Supporting educational initiatives also contributes to cultivating a skilled workforce, which enhances business competitiveness and fosters innovation.

V. Advanced Analysis

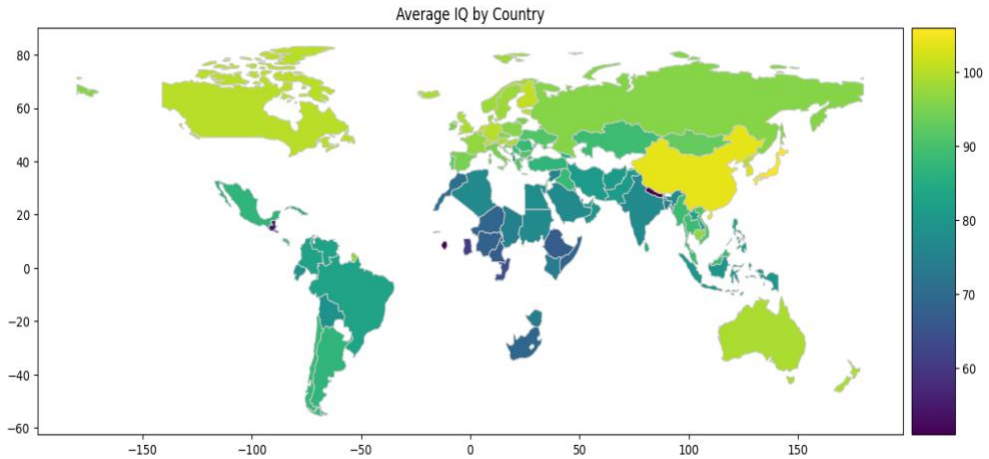
```
# Advanced Analysis
print("\nAdvanced Analysis")
print("=================")

# Geographical Insights: Average IQ by Country
import geopandas as gpd
from mpl_toolkits.axes_grid1 import make_axes_locatable

# Read world shapefile
world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))

# Merge world shapefile with IQ data
world_iq = world.merge(data, how='left', left_on='name', right_on='country')

# Plot average IQ by country
fig, ax = plt.subplots(1, 1, figsize=(15, 10))
divider = make_axes_locatable(ax)
cax = divider.append_axes("right", size="5%", pad=0.1)
world_iq.plot(column='IQ', cmap='viridis', linewidth=0.8, ax=ax, edgecolor='0.8', legend=True, cax=cax)
ax.set_title('Average IQ by Country')
plt.show()=
```

The provided world map illustrates average IQ scores by country using a color gradient from dark blue (indicating lower IQ scores around 60) to bright yellow (indicating higher scores above 110). North America and Europe predominantly appear in light green to yellow, suggesting high average IQs, typically from 90 to over 100.

East Asian countries like China, Japan, and South Korea are depicted in bright yellow, indicating some of the highest average IQ scores globally. In contrast, much of Africa is shaded in dark blue to teal, reflecting lower average IQ scores, generally below 80. Latin American countries show a range from teal to green, indicating scores in the mid to high range (70-90), while Australia and New Zealand are in yellow-green shades, similar to North America and Western Europe.

The map reveals clear geographical patterns, with high average IQ regions in East Asia and parts of Europe, moderate to high scores in North America, Australia, and parts of Latin America and Europe, and lower scores predominantly in Sub-Saharan Africa and some areas of South Asia and the Middle East.

These variations highlight the influence of socioeconomic factors, educational systems, and access to resources on average IQ scores. However, potential limitations include the accuracy and methodology of measuring IQ across countries, cultural biases in IQ tests, and the lack of within-country variations. The map provides a visual summary of global IQ distribution, emphasizing regional disparities and the need for cautious interpretation considering broader contextual factors.



VI. Machine Learning Implementation

Data Preparation

For the machine learning implementation, the dataset was prepared with a focus on key features such as education expenditure, average income, and average temperature. The target variable for our model is the IQ level.

Data Cleaning

To ensure the dataset's integrity, missing values in the selected features were imputed with their respective column means. This method preserves the overall data structure and prevents potential biases that missing data might introduce.

Data Scaling

We applied feature scaling to normalize the features using the StandardScaler from scikit-learn. This step standardizes the features by centering them around the mean and scaling them to unit variance, ensuring that all features contribute equally during model training.

Data Splitting

The dataset was divided into training and testing sets with an 80-20 split, using the train_test_split function from scikit-learn. Setting the random_state to 42 ensures that the results are reproducible. This split allows us to train the model on one subset and validate its performance on another, ensuring its generalizability.

Model Building

A Linear Regression model was utilized to predict IQ levels based on the selected features. This model is suitable for capturing linear relationships between the dependent variable (IQ) and the independent variables (education expenditure, average income, and average temperature). The model was trained using the fit method on the training data.

Model Evaluation

The model's performance was evaluated using the R-squared (R^2) metric. The training R^2 score of 0.748 and testing R^2 score of 0.702 indicate that the model explains a substantial portion of the variance in IQ levels, both in the training and testing datasets, reflecting its accuracy and robustness.

Business Impact

The insights derived from the machine learning model highlight the significant impact of education expenditure on IQ levels, emphasizing the need for strategic investments in education. The correlation between average income and IQ levels also suggests that economic conditions play a crucial role in cognitive development. Policymakers can use these findings to allocate resources more effectively and design targeted initiatives that enhance educational and socio-economic conditions. These data-driven insights can lead to improved decision-making, fostering societal growth and better educational outcomes.

VII. Visual Insights

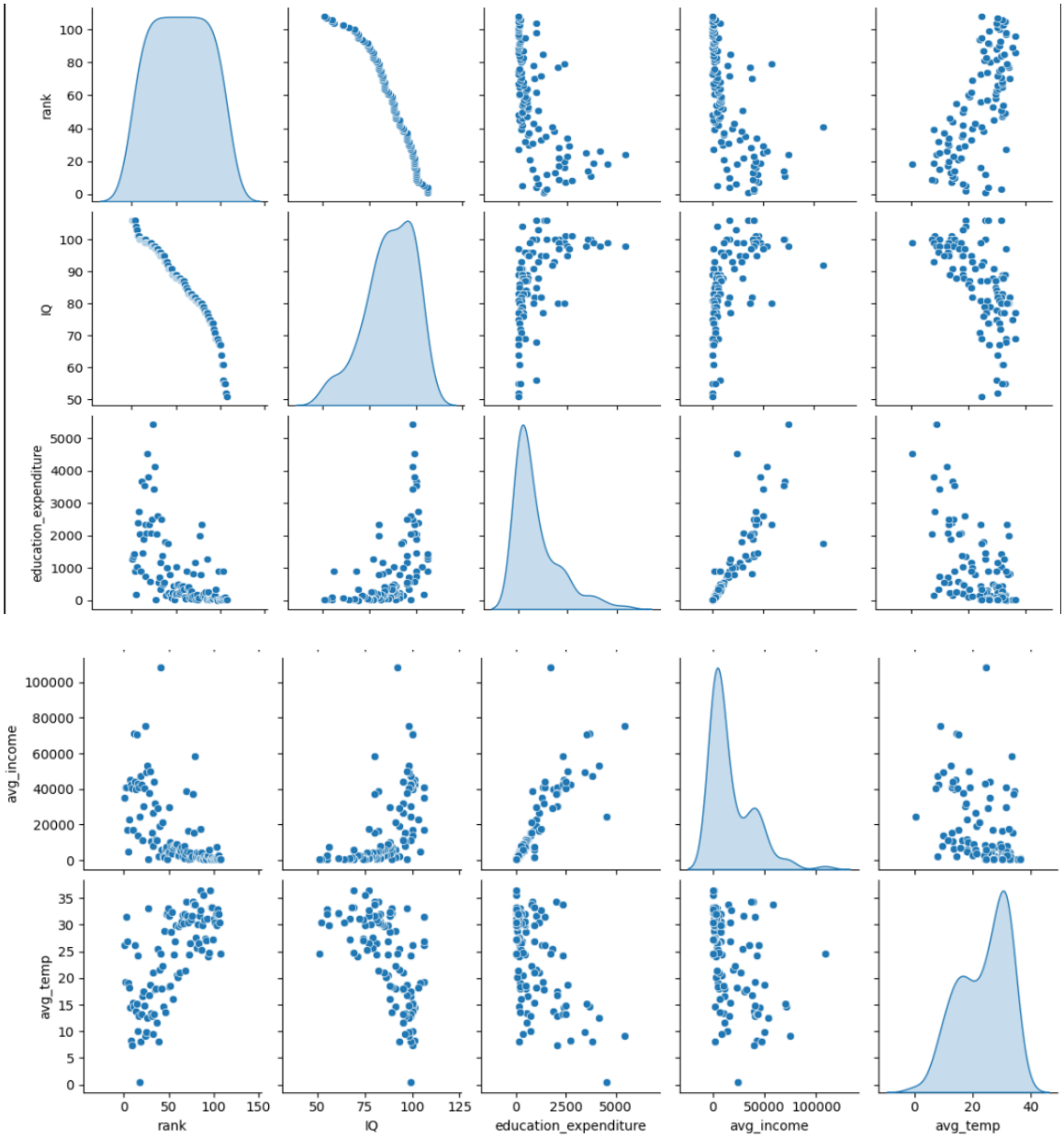
Visual Insights

```
# Pairplot to visualize relationships between variables
pairplot = sns.pairplot(data, diag_kind='kde')  # Pairplot without 'country'
pairplot.fig.subplots_adjust(right=0.9)
plt.show()
```

The pairplot provides a comprehensive view of the relationships between all numeric variables in the dataset through a grid of scatter plots. The diagonal plots show kernel density estimates for each variable, allowing us to visualize the distribution and detect any patterns or outliers. This helps in identifying correlations and interactions between pairs of variables.

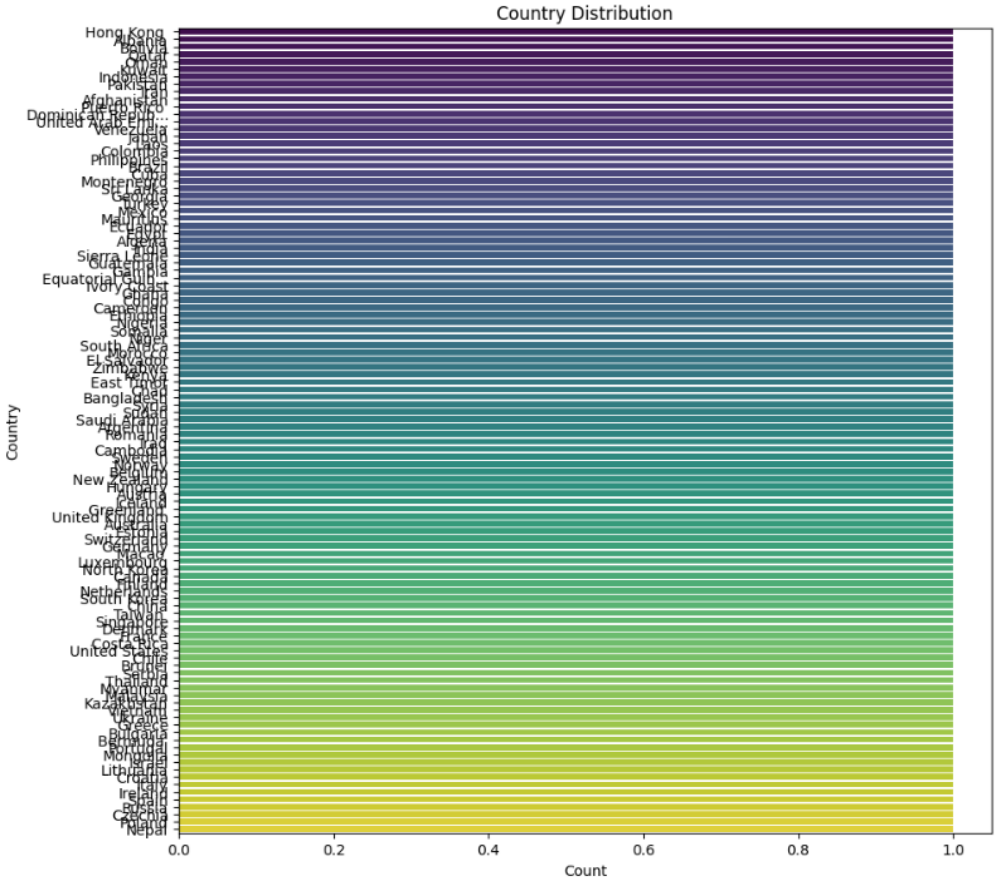
FINAL PROJECT - MACHINE LEARNING IMPLEMENTATION

ANDRES, FLORES, PANTANOZA



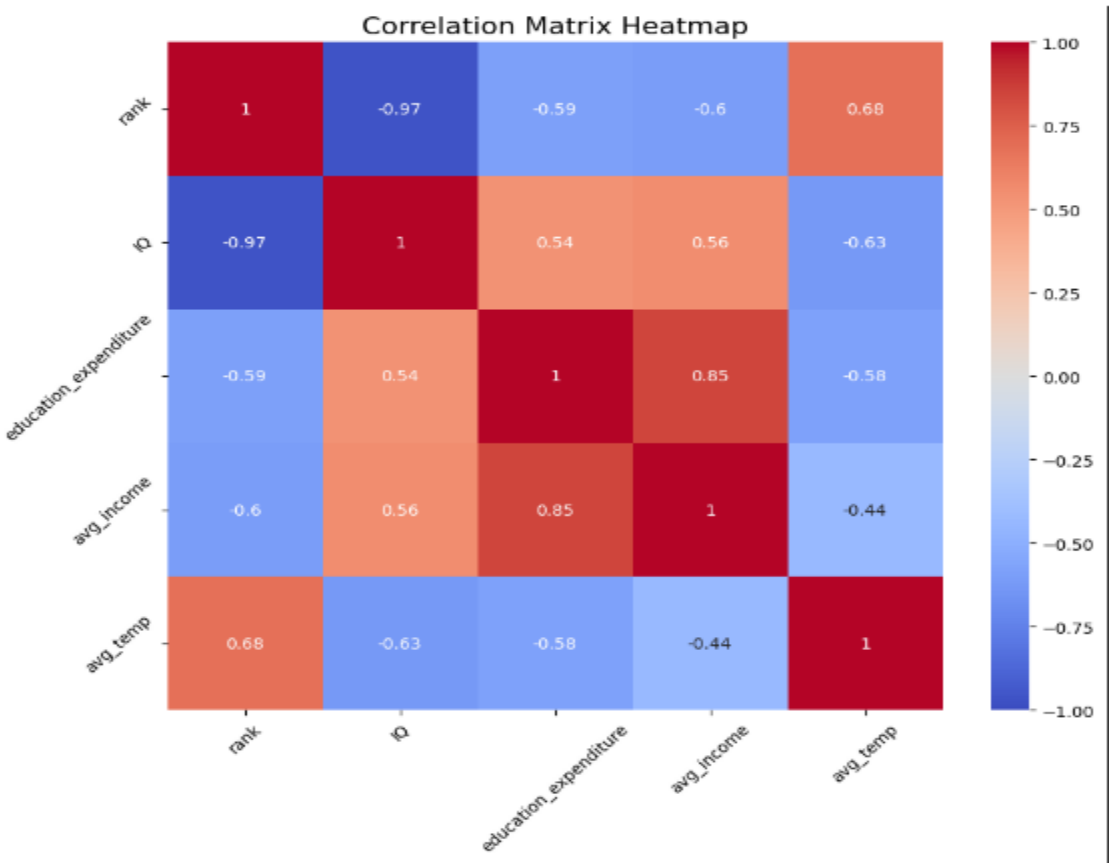
```python
# Country distribution
plt.figure(figsize=(10, 10))  # Adjust figure size as needed
country_counts = data['country'].value_counts()
truncated_countries = [country[:15] + '...' if len(country) > 15 else country for country in country_counts.index]
sns.countplot(y=truncated_countries, data=data, palette='viridis')
plt.title('Country Distribution')
plt.xlabel('Count')
plt.ylabel('Country')
plt.xticks(rotation=0)  # No rotation for truncated country labels
plt.show()
```

This bar plot visualizes the frequency of occurrences of each country in the dataset. The value_counts() function counts the number of times each country appears, and the countplot function displays these counts as horizontal bars. Truncating long country names ensures that labels remain readable. This visualization helps to quickly assess the distribution of data across different countries.



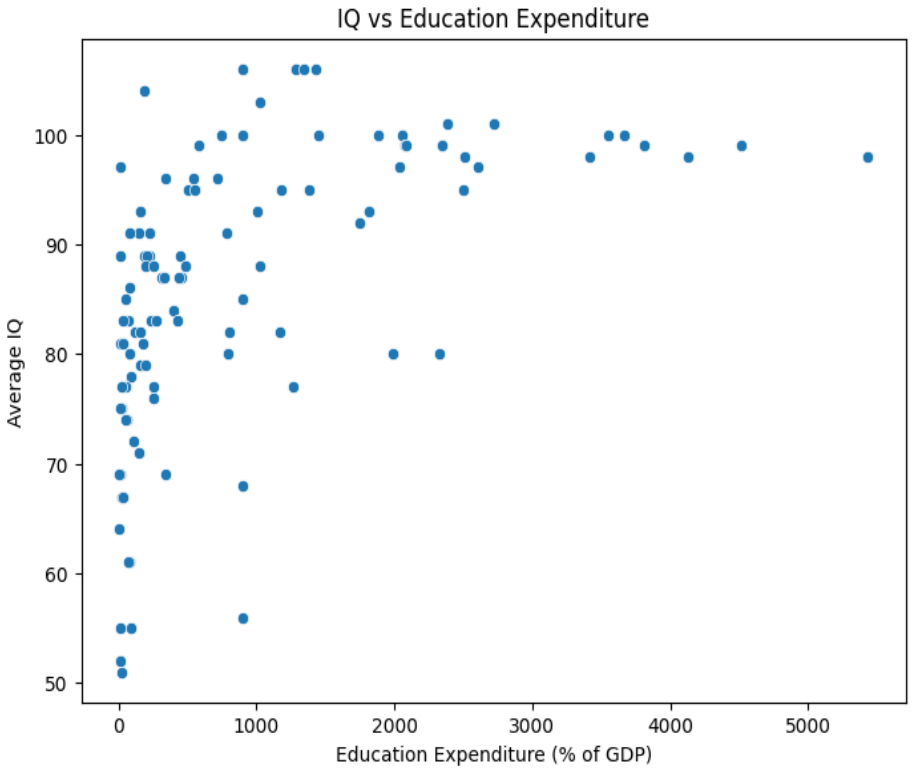
```
# Correlation matrix
corr_matrix = numeric_data.corr()

# Heatmap of the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)

# Adjusting heatmap text for readability
plt.xticks(rotation=45)
plt.yticks(rotation=45)
plt.title('Correlation Matrix Heatmap', fontsize=16)
plt.show()
```

The heatmap displays the correlation coefficients between all pairs of numeric variables in the dataset. The corr() function computes the correlations, and the heatmap function visualizes them with color coding, where warmer colors (closer to 1) indicate strong positive correlations and cooler colors (closer to -1) indicate strong negative correlations. Annotations show the exact correlation values, making it easy to identify significant relationships that might warrant further investigation.



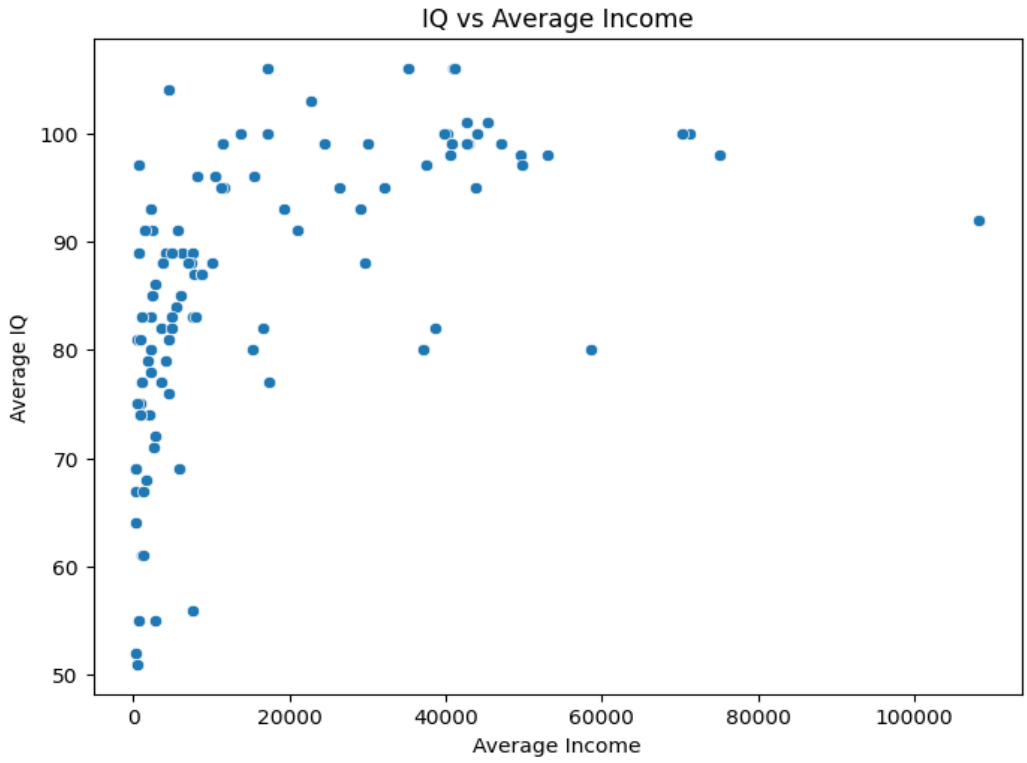
```
# IQ vs Education Expenditure
plt.figure(figsize=(8, 6))
sns.scatterplot(x='education_expenditure', y='IQ', data=data)
plt.title('IQ vs Education Expenditure')
plt.xlabel('Education Expenditure (% of GDP)')
plt.ylabel('Average IQ')
plt.show()
```

This scatter plot illustrates the relationship between education expenditure (as a percentage of GDP) and average IQ levels. Each point represents a country, plotted according to its education expenditure and average IQ. This visualization helps to identify any potential trends or correlations, such as whether higher education expenditure is associated with higher average IQ levels.



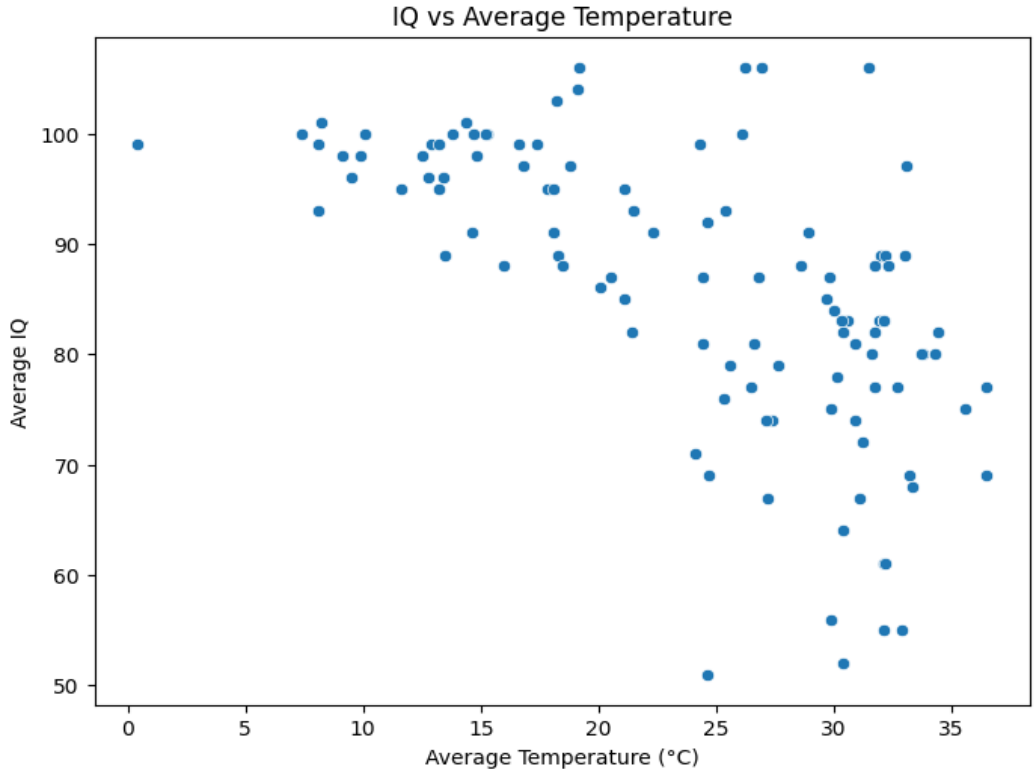
```
# IQ vs Average Income
plt.figure(figsize=(8, 6))
sns.scatterplot(x='avg_income', y='IQ', data=data)
plt.title('IQ vs Average Income')
plt.xlabel('Average Income')
plt.ylabel('Average IQ')
plt.show()
```

This scatter plot shows the relationship between average income and IQ levels. By plotting average income against IQ for each country, we can observe whether there is a correlation between economic wealth and cognitive performance. This visualization can help to determine if higher average incomes correspond to higher IQ levels across different countries.



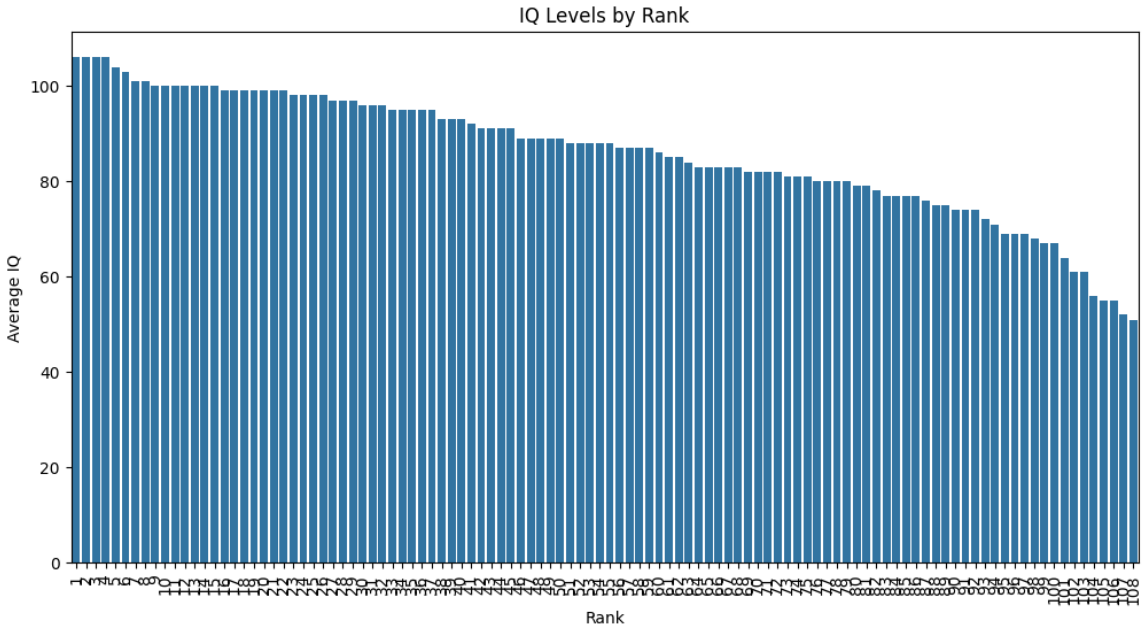
```
# IQ vs Average Temperature
plt.figure(figsize=(8, 6))
sns.scatterplot(x='avg_temp', y='IQ', data=data)
plt.title('IQ vs Average Temperature')
plt.xlabel('Average Temperature (°C)')
plt.ylabel('Average IQ')
plt.show()
```

This scatter plot explores the relationship between average temperature and IQ levels. By plotting average temperature against IQ for each country, we can investigate if there is any pattern or correlation between climate and cognitive abilities. This can provide insights into how environmental factors might influence IQ.



```
# Display the rankings and IQ levels
plt.figure(figsize=(12, 6))
sns.barplot(x='rank', y='IQ', data=data)
plt.title('IQ Levels by Rank')
plt.xlabel('Rank')
plt.ylabel('Average IQ')
plt.xticks(rotation=90)
plt.show()
```

This bar plot visualizes average IQ levels by rank, with each bar representing a country ranked by its IQ. The barplot function provides a clear comparison of IQ levels across different ranks, highlighting which countries have higher or lower average IQs. This helps in understanding the distribution of IQ levels across the ranked countries.



VIII. Conclusion

This project has provided a comprehensive analysis of a global dataset to uncover insights into the relationship between average IQ levels and various socioeconomic and environmental factors.

Through detailed data analysis, visualization, and the implementation of a machine learning model, we have identified several key findings. The positive correlation between education expenditure and average IQ levels suggests that investing in education can significantly enhance cognitive abilities on a national scale.

Additionally, higher average income is associated with higher IQ levels, indicating the influence of socioeconomic status on intellectual development. Conversely, the average temperature shows a less significant correlation with IQ levels, underscoring the complexity of environmental impacts on intelligence.

The visualizations, including scatter plots and a world map of IQ scores, have highlighted geographical patterns and disparities in IQ levels across different regions. These insights emphasize the importance of targeted educational investments and socioeconomic interventions to address regional disparities and promote cognitive development.

In conclusion, this project highlights the essential role of data-driven decision-making in formulating effective policies and strategies. By utilizing data analysis and machine learning, organizations and policymakers can derive valuable insights to guide resource allocation and create interventions that improve educational and socioeconomic conditions. This methodology not only aids in enhancing individual cognitive abilities but also supports broader societal growth and well-being.