CS 2704 Final Report

Project Title: Analyzing the Relationship Between GDP per Capita and Unemployment Rate

Team Name: Data Grinders

Team Members: Jaspinder Singh, Syed Owais Haider Kazmi, Nomaan Imran Saiyed

# 1. Introduction and Background

Unemployment and GDP per capita are two fundamental indicators of a country's economic health. GDP per capita reflects the average economic output per person, while unemployment rate shows the percentage of people actively seeking jobs. Many economists hypothesize that wealthier nations tend to maintain lower unemployment rates due to stronger labor markets and economic stability. Our project investigates whether there is a statistically significant relationship between these two variables on a global scale. By analyzing data across countries and applying regression and machine learning models, we aimed to explore both the strength and direction of this potential relationship.

# 2. The Hypothesis

Null Hypothesis (H₀): There is no relationship between GDP per capita and unemployment rate.

Alternative Hypothesis (H₁): There is a significant relationship between GDP per capita and unemployment rate.

Our expectation was a negative correlation, implying that countries with higher GDP per capita generally have lower unemployment. To test this, we used multiple regression approaches—linear, polynomial, and random forest—and analyzed correlation coefficients and p-values from statistical tests.

# 3. Analysis and Implication

a) Data Source & Description:  
- GDP per Capita: World Bank (NY.GDP.PCAP.CD)  
- Unemployment Rate: World Bank dataset  
- Merged Dataset Columns: Country, Year, GDP\_per\_capita, Unemployment\_Rate, log\_GDP\_per\_capita

b) Descriptive Analytics:  
- Basic stats revealed a wide GDP range from under $1,000 to over $100,000.  
- Unemployment ranged from below 1% to over 30%.  
- A correlation heatmap showed weak correlation between GDP and Unemployment.

c) Visualizations:  
- Scatter Plot: Indicated a slight downward trend.  
- Regression Plot (Log GDP): Clearer trend but not strong.  
- Boxplot: Medium-GDP countries had the highest unemployment variability.  
- Top/Bottom 10 Plots: Showed inequality and outliers.

d) Predictive Analytics & Model Evaluation:  
- Linear Regression: Model showed poor fit. High p-value → relationship is not statistically significant.  
- Polynomial Regression: Slightly better curve fit but still weak prediction.  
- Random Forest: Captured more variance, but still poor performance overall.  
- Hypothesis Test: Pearson correlation and p-value from OLS regression support failure to reject the null.

# 4. Conclusion

Despite economic theories suggesting a negative relationship, our analysis found no statistically significant correlation between GDP per capita and unemployment rate across countries. While certain trends were visible in scatter and regression plots, statistical testing showed high p-values and weak predictive power. This suggests that unemployment is influenced by more complex and country-specific factors, including labor policies, education, demographics, and government intervention. Our findings highlight the importance of multidimensional analysis when exploring global economic indicators.

# 5. References

- World Bank Data: https://data.worldbank.org  
- Dataset Merged from: GDP per capita (NY.GDP.PCAP.CD), Unemployment (% of total labor force)  
- Python Libraries: pandas, matplotlib, seaborn, scikit-learn, statsmodels

# 6. Team Contributions

This project was a collaborative effort by Team Data Grinders:  
- Jaspinder Singh: Data acquisition, cleaning, visualization  
- Syed Owais Haider Kazmi: Hypothesis testing, regression modeling, interpretation  
- Nomaan Imran Saiyed: Random forest regression, insights, presentation development  
All members jointly contributed to planning, discussion, and reviewing final deliverables.

# 7. Limitations and Future Work

Limitations:  
- Data limitations: Different years and inconsistencies  
- Variable complexity: Unemployment is influenced by more than GDP  
- Model limitations: Random Forest did not perform significantly better  
  
Future Work:  
- Include more socio-economic indicators  
- Explore time series analysis for trends  
- Use clustering for regional trends  
- Test causation using advanced statistical models