**CS-2704: Data Analytics Using Python**

**Instructor: Dr. Jong-Kyou Kim**

**Final Project Proposal**

**Team Members-** Jaspinder Singh(3770406),

Syed Owais Haider Kazmi,

Nomaan Imran Saiyed

**Dataset**

I will use two publicly available datasets from the World Bank:

* **GDP per capita (current US$)  
  Source:** <https://data.worldbank.org/indicator/NY.GDP.PCAP.CD>
* **Unemployment Rate (% of total labor force)  
  Source:** <https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS>

These datasets provide yearly economic indicators by country and will be merged using “Country Name” and “Year”.

**GitHub Repository**

Repository Link: <https://github.com/jaspindersingh919/CS2704>  
The repo will include:

* Raw dataset and cleaned dataset
* Source code for analysis
* This proposal
* Final report and slides

**Hypothesis**

There is a negative correlation between GDP per capita and unemployment rate across countries.

In other words, as the GDP per capita increases, the unemployment rate is expected to decrease.

**Plan for Testing the Hypothesis**

1. **Data Collection & Cleaning:**
   * Download CSVs from the World Bank
   * Merge them based on common columns (Country, Year)
   * Handle missing or invalid values
2. **Descriptive Analytics:**
   * Generate summary statistics
   * Visualize GDP vs. Unemployment with scatter plots
   * Create correlation matrix/heatmap
3. **Predictive Analytics:**
   * Apply simple linear regression (Unemployment as dependent variable, GDP per capita as independent)
   * Analyze regression output: coefficients, p-value
   * Discuss statistical significance
4. **Discussion:**
   * Reflect on findings and anomalies
   * Explore whether certain regions follow the trend more strongly than others

**Expected Output**

I expect to find a statistically significant negative correlation between GDP per capita and unemployment rates. However, this correlation may vary between regions and depend on additional socioeconomic factors.