**Group Project Proposal**

**University of New Brunswick, Saint John**

**CS 2704: Data Analytics using Python**

**Professor Jong-Kyou Kim**

**Group Members:**

**Aryan Bansal**

**Aniket Shah**

**Dhruv Thumar**

### **Dataset**

I will use two publicly available datasets:

* **Electricity Consumption Data** (kWh per state per year)
  + Source: [U.S. Energy Information Administration (EIA)](https://www.eia.gov/electricity/data.php)
* **Temperature Data** (Average yearly temperature by state)
  + Source: [NOAA Climate Data](https://www.ncdc.noaa.gov/cdo-web/)

These datasets provide yearly electricity consumption and temperature data by U.S. state and will be merged using “State” and “Year.”

### **GitHub Repository**

**Repository Link**: [https://github.com/dhruvthumar/2704project](https://github.com/dhruvthumar/2704project?tab=readme-ov-file) The repo will include:

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

### **Hypothesis**

Higher average temperatures in a region correlate with increased electricity consumption due to air conditioning use.

**Plan for Testing the Hypothesis**

1. **Data Collection & Cleaning:**
   1. Download CSVs from EIA and NOAA
   2. Merge them using common columns (State, Year)
   3. Handle missing or invalid values
2. **Descriptive Analytics:**
   1. Compute summary statistics (e.g., average electricity consumption, average temperature)
   2. Calculate Pearson correlation coefficient between temperature and consumption
   3. Visualize data using scatter plots and correlation heatmaps
3. **Predictive Analytics:**
   1. Apply simple linear regression (Electricity consumption as dependent variable, temperature as independent variable)
   2. Extend analysis by introducing **Cooling Degree Days (CDD)**:
      1. CDD = Sum of (Daily Temperature - 65°F) when Temperature > 65°F
      2. Use CDD as an additional predictor in a multiple regression model
   3. Evaluate model using R-squared and p-values for statistical significance
4. **Discussion:**
   1. Interpret findings, including anomalies or variations across states
   2. Examine whether specific regions follow the trend more strongly

### **Expected Output**

I expect to find a statistically significant positive correlation between temperature and electricity consumption. However, additional factors like seasonal variations, energy efficiency policies, and income levels may influence the strength of this relationship.