CSP 571 Data Preparation and Analysis

Group Name: The A-Team

Project Title: The Impact of Environmental Factors on Socioeconomic Outcomes

Primary Datasets:

1. NASA SEDAC – Light-based Geospatial Income Inequality (LGII)
2. Kaggle – Geospatial Environmental and Socioeconomic Data

1. Overview

For this stage, we expand our project to integrate two complementary datasets:

- The Kaggle environmental dataset providing spatial climate, pollution, vegetation, and hazard indicators.
- The NASA SEDAC LGII dataset, which provides raw, light-derived Gini coefficients and socioeconomic metrics (population grids, density, openness, census counts) from 1992–2013.

Using both creates a richer dataset, introduces real data cleaning work, and differentiates our project from similar Kaggle-only analyses. Our goal remains to study how environmental conditions relate to socioeconomic inequality across countries and over time.

2. Dataset Integration Plan

LGII Dataset (Socioeconomic / Inequality)

- Rows: ~4,500 (country × year)
- Years: 1992–2013
- Variables: multiple weighted Gini measures, population metrics, economic openness, density, census grids
- Cleaning Needs: missing values, inconsistent types, redundant Gini columns, country name/ISO alignment

Kaggle Dataset (Environmental)

- Raster + tabular environmental variables at ~1km resolution (NDVI, temperature, precipitation, PM2.5, hazards).
- Requires CRS alignment, raster aggregation, missing-value handling.

Merging Plan

- Aggregate Kaggle raster layers to country level.
- Standardize both datasets to ISO codes.
- Merge LGII yearly data with environmental summaries for overlapping years.
- Build final dataset containing environmental predictors + inequality metrics + socioeconomic indicators.

3. Research Questions & Hypotheses

1. **Do environmental conditions correlate with income inequality?**
   Hypothesis: Higher pollution/lower vegetation → higher Gini.
2. **Does environmental change predict inequality trends over time?**
   Hypothesis: Temperature/NDVI shifts correspond to inequality increases.

3. **Are poorer regions more exposed to environmental stress?**
   Hypothesis: Lower-income countries show higher hazard and pollution exposure.

4. Early Exploratory Insights

Initial checks show:

- LGII captures meaningful inequality variation across years.
- NDVI, PM2.5, and GDP strongly correlate with development patterns in the Kaggle dataset.
- Countries with high population density show more extreme inequality shifts.

These patterns indicate that linking the two datasets is feasible and informative.

5. Planned Methods

Data Transformation

- Clean LGII, drop redundant fields, handle missing values
- Aggregate environmental rasters
- Normalize variables and create derived indices (Environmental Stress, Socioeconomic Vulnerability)

Modeling & Analysis

- Linear / multiple regression
- Random forest for non-linear effects
- Temporal analysis using inequality change (ΔGini)
- Spatial analysis (Moran's I)
- Clustering of countries by combined environmental–socioeconomic profiles

Visualization

- Choropleth maps
- Scatter plots of environment vs. inequality
- Time-series graphs for country-level trends

6. Expected Outcomes

- A merged, cleaned dataset combining environment + inequality + socioeconomic metrics
- Quantitative evidence of how environmental conditions relate to socioeconomic outcomes
- Identification of high-risk countries
- A final Climate–Socioeconomic Vulnerability Index (CSVI) and global map summarizing combined risk

7. Members: Atharv Rathore (A20595125), Samarth Rajput (A20586237), Gaurav Acharya (A20583612), Jenil Panchal (A20598955)

8. Dataset Links:

Geospatial Environmental and Socioeconomic Data – Kaggle

earthdata.nasa.gov