DSGP Proposal: “Forecasting Political Outcomes in Contemporary Government”

UNIVERSITY OF BIRMINGHAM

Group 18:

Connor Boyd-Lyon, Evelina Ivanova, Matthew Randall, Yarkın Yorulmaz, Yeni Jung, Zehn-Ul-Abideen Sharif

Table of Contents

Section 1: The Research Question

This project aims to explore using machine learning to predict election outcomes, instead of relying on traditional polling. It focuses on OECD countries, where data on social, political, and economic factors will be used to build predictive models. Through exploratory data analysis, we will identify key features and create classifiers that forecast whether a country is likely to experience a change in government. The success of the project will be measured by how accurately the models predict election results compared to traditional polling methods. If successful, this approach could have wide-ranging applications in politics, law, and economics.

“Can Machine Learning Be Applied to Predict Election Outcomes?”

Sub questions:

- *“Which features (economic, demographic, political, social etc.) are most relevant for election prediction?”*
- *“Could machine learning models be generalized across multiple countries, or do they need to be country-specific?”*
- *“Is there any significant difference between countries?”*
- *“How does machine learning differ from polling when predicting an election?”*

This project aims to investigate whether general election outcomes can be predicted using machine learning techniques rather than polling. Polling is a common election outcome predictor used by pollsters (groups that monitor voting preference) where people are asked their voting preferences. Whilst this is an effective method there are intrinsic inefficiencies; an automated prediction method would be preferable. By using machine learning to predict election outcomes, pollsters could focus less on *what* the outcome is and instead on the effects of that outcome.

The observation group chosen for study is The Organisation for Economic Co-operation and Development (OECD). The OECD (2025) describes itself as “a forum and knowledge hub for data, analysis and best practices in public policy”. There are 38 OECD member countries as of writing; these will be the focus of the initial analyses, due to the large volume of data that’s publicly available for these countries. EDA will be performed on these countries; based on the EDA a much smaller subset of statistically different countries will be selected for modeling. For this subset, we aim to create a series of predictive models that learn features that contribute significantly to election outcomes for those countries. The models will be classifiers, which take a country’s current ‘state’ as inputs, measured by index values, and output whether or not a country should expect to change governments. To achieve this, the group will:

- Find a database of election outcomes over time for the OECD countries
- Extract key features from social, political, and economic indexes for these countries
- Merge and clean these datasets to make a master data frame.
- Perform EDA, including a cluster analysis, to form relations between countries
- Choose a subset of countries from the cluster analysis.
- Develop a classification model trained on past election data to establish relationships between these factors and electoral results. The output will be whether a change in government is expected or not.
- Evaluate model performance by testing its accuracy against historical outcomes and comparing results with traditional polling-based predictions.

The project will be considered a success if significant correlation is detected between input features and election outcomes. If a model is made that successfully predicts election outcomes, it would have applications in the political, legal and economic domains.

Section 2: Data Sources and Evidence of Permission for Use

This section contains the sources of each of the datasets that are intended for use in this analysis. For each dataset, the appropriate license is provided.

- **Election Results:** The Parliaments and Governments database (ParlGov) // License *Creative Commons 0 v1.0* License: CC0 1.0.

- **Political Factors**
    - Judicial Independence and Media Freedom // License Creative Commons Attribution-NoDerivs 4.0 Global State of Democracies (1990-2022) by Institute for Democracy and Electoral Assistance is licensed under CC-BYND 4.0)
    - Human Development Index // License Creative Commons Attribution-ShareAlike 3.0 IGO
    - Corruption Perceptions Index // License Creative Commons Attribution-NoDerivs 4.0 (Corruption Perceptions Index (1990-2023) by Transparency International is licensed under CC-BYND 4.0)

- **Democracy and Electoral Integrity Factors**
    - Liberal democracy index // License CC BY 4.0
    - Countries where armed conflicts took place // License CC BY 4.0

- **Economic Factors**
    - Annual growth of GDP per capita // License Creative Commons BY license
    - Economic Inequality // License Creative Commons BY license
    - Exchange Rates // License CC BY-4.0
    - Inflation // License CC BY-4.0
    - Unemployment // License CC BY-4.0
    - VAT, Income Tax and Corporate Income tax: License CC BY-4.0

- **Social Factors**
    - **Ethnic or religious diversity**
        - EPR Core - ETH Zurich
        - ACD2EPR - ETH Zurich
        - The Ethnicity of Refugees (ER) - ETH Zurich
          (The following citation is for datasets from ETH Zurich above; *Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp Hunziker, and Luc Girardin. 2015. "Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Data Set Family." Journal of Conflict Resolution 59(7): 1327–42.*)
        - Historical Index of Ethnic Fractionalization Dataset (HIEF) - Harvard Dataverse // License CC0 1.0

- **Level of Education**
  - Government expenditure on education - UNESCO // License Creative Commons Attribution-ShareAlike 3.0 IGO
- **Population**
  - Urban population growth (annual %) - The World Bank // License CC BY-4.0
- **Healthcare**
  - Healthcare provider resources - OECD
  - Healthcare coverage - OECD

    (The source of the datasets above does not specify the license, but the following statement is specified: *"...Except where additional restrictions apply as stated above, you can extract from, download, copy, adapt, print, distribute, share and embed Data for any purpose, even for commercial use. You must give appropriate credit to the OECD by using the citation associated with the relevant Data, or, if no specific citation is available, you must cite the source information using the following format: OECD (year), (dataset name),(data source) DOI or URL (accessed on (date)). When sharing or licensing work created using the Data, you agree to include the same acknowledgment requirement in any sub-licenses that you grant, along with the requirement that any further sub-licensees do the same."*)

Section 3: A Datasheet For the Datasets

This section provides the combined datasheet for the datasets used in this project.

Parliaments and Governments (ParlGov) Database

ParlGov is a data infrastructure designed for political science research. It contains comprehensive information on political parties, elections, and government cabinets across all EU countries and most OECD democracies (totaling 37 countries) from 1945 to the present. The dataset exists to support political scientists in performing analyses.

Creators & Funding

Holger Döring and Philip Manow (The University of Bremen) are the project leaders for this dataset. There's a long list of editors, contributors, and funding bodies available on the readme.

Composition

- **Instances & Representation:** The dataset consists of multiple tables, each representing different aspects of political systems, such as elections, cabinets, and party structures. The total number of instances varies across tables.
- **Coverage:** ParlGov aims to be comprehensive, covering all relevant instances from EU and OECD democracies. It's a curated dataset representing the historical and contemporary political landscape of these countries.
- **Data Structure:** Each instance contains structured numerical and categorical attributes detailing political events, party structures, election results, and government compositions.
- **Target Variables:** There are no explicit labels, but various attributes can be used as targets in analytical models, such as seats won in an election.
- **Missing Data:** Some missing data exists due to unavailable historical records.
- Explicit Relationships: Relationships between parties, governments, and elections are explicitly linked through unique identifiers.
- **Data Splits:** No predefined splits are provided.
- **Errors & Noise:** Issues arise when; a major party changes its name; a country changes its name; A voting system changes; etc.
- **Self-Containment:** The database is entirely self-contained. Tables within the database relate to each other.
- **Confidentiality & Sensitivity:** No personal or confidential data is included. However, it contains political information that may be sensitive in certain contexts.

Collection Process

- **Data Acquisition**: The data was gathered from official sources, academic publications, and government reports. It was manually curated and cross-validated.
- **Collection Mechanisms**: Manual data collection, academic sources, government databases, and structured validation methods.
- **Timeframe**: The dataset covers political data from 1945 to the present.

- **Ethical Review**: Since the dataset does not contain personal data, no formal ethical reviews were necessary.

Preprocessing, Cleaning, and Labeling

- **Preprocessing/Labeling**: The data is processed and labeled already for ease of access purposes. Slight calculations may be required; to find the winners of an election, the party with the highest number of seats could be found.
- **Raw Data Storage**: During the project, raw data will be maintained in the Deepnote workspace.

Uses

- **Previous Uses**: The dataset has been widely used in political science research, including studies on party systems, government stability, and election forecasting.
- **Potential Uses**: Predicting election outcomes; Analyzing government stability; Studying political party evolution.
- **Risks & Limitations**: Political data can be used to support biased narratives.
- **Restricted Uses**: It should not be used for personally identifiable predictions or manipulative political purposes.

Distribution

- **Availability**: The dataset is publicly available for research purposes.
- **Distribution**: ParlGov 2024 Release - Harvard Dataverse
- **Licensing**: CC0 1.0

Maintenance

- **Maintainers:** Holger Döring and Philip Manow at the University of Bremen.
- **Contact Information:** Contact details are provided in the dataset documentation.
- **Updates:** The dataset is periodically updated to reflect new political developments.
- **Version Control:** Previous versions remain available for reference.
- **Contributions:** Contributions are accepted through a structured submission process.

Political Factors:

The Global State of Democracy Indices

Overview

The Global State of Democracy (GSoD) Indices dataset provides systematic and nuanced data on democratic trends across countries and regions from **1975 to 2022**. It was developed by **International IDEA (International Institute for Democracy and Electoral Assistance)** to assess **democracy at the global, regional, and national levels**. The dataset contains **159 indicators** grouped under four core attributes of democracy.

Key Features

- **Temporal Coverage: 1975–2022**
- **Geographical Coverage: 174 countries** (excluding microstates with populations below 250,000)
- **Instances: Country-year level data** (~8,000+ instances)
- **Indicators:** 159 democracy-related variables, categorized into four main attributes:
    1. **Representation** (e.g., free elections, elected government, political party freedom)
    2. **Rights** (e.g., civil liberties, political equality, access to justice)
    3. **Rule of Law** (e.g., judicial independence, corruption, personal security)
    4. **Participation** (e.g., voter turnout, civic engagement)

Data Collection Methodology

The dataset is built from a **wide range of sources**, including:

- **Expert Surveys (ES):** Assessments by political scientists and country experts (e.g., V-Dem, ICRG)
- **Standards-Based In-House Coding (IC):** Data coded by researchers using reports and academic sources (e.g., Freedom House, Polity IV)
- **Observational Data (OD):** Measurable indicators like voter turnout, education levels (e.g., UNESCO, FAO)
- **Composite Measures (CM):** Aggregated indices combining multiple datasets (e.g., V-Dem's electoral democracy index)

Validation & Reliability:

- Data undergoes **cross-validation**, ensuring **consistency across sources**.

- **Uncertainty estimates** are included for most indices to indicate confidence levels.

Dataset Structure

- **Unit of Observation:** Country-year
    - **Variables:**
    - Example Indicators:
        - **Credible Elections Score** (0–1 scale)
        - **Freedom of Expression** (0–1 scale)
        - **Judicial Independence** (0–1 scale)
    - **Uncertainty Estimates** accompany most indicators.

Data Relationships:

- No explicit network structure, but **country-year trends** allow for time-series analysis.

Preprocessing & Updates

- **Normalization:** All indicators are scaled between **0 (low democracy) and 1 (high democracy)**.
- **Missing Data Handling:** Bayesian estimation is applied for some sub attributes.
- **Changes in Version 7 (2023):**
    - **Additional indicators** are included.
    - **A refined conceptual framework** based on expert feedback.

Usage & Applications

Existing Uses:

- International IDEA's Global State of Democracy Reports
- Government policy assessments
- Academic research on democratic trends
- NGO advocacy & risk analysis

Potential Uses:

- Predictive modeling of democratic decline
- Comparative political studies

- Democratic resilience assessments

Limitations & Risks:

- **Not designed for direct country rankings** due to its multi-dimensional nature.
- **Short-term variations should be interpreted cautiously** (uncertainty estimates available).

Distribution & Licensing

- **Available via International IDEA**: GSoD Indices Portal
- **Licensing:** Open access, except for a few proprietary indicators (e.g., ICRG data).

Maintenance & Future Updates

- **Maintained by: International IDEA**
- **Update Frequency:** Periodic (latest: Version 7, 2023)
- **Contact:** info@idea.int

Citation

International IDEA. 1990-**2023**. *The Global State of Democracy Indices, Version 7*. Stockholm: International IDEA.

Human Development Index (HDI)

Overview

The **Human Development Index (HDI)** is a composite measure developed by the United Nations Development Programme (UNDP) to assess and compare levels of human development across countries. It provides a broad understanding of well-being by combining three key dimensions: **health, education, and standard of living**. The HDI is widely used in policy-making, research, and global development analysis.

Key Features

- **Temporal Coverage:** 1990–present (updated annually)

- **Geographical Coverage:** Nearly all UN member states and territories
- **Data Points:** HDI scores and rankings for countries, along with disaggregated data for each of the three dimensions

Components and Methodology

The HDI is calculated using the following three **core dimensions**:

1. **Health:** Measured by **life expectancy at birth**, which reflects overall well-being and access to healthcare.
2. **Education:** Assessed through **mean years of schooling** (for adults aged 25 and older) and **expected years of schooling** (for children entering the education system).
3. **Standard of Living:** Evaluated using **Gross National Income (GNI) per capita** (adjusted for purchasing power parity, PPP).

Each of these dimensions is normalized on a scale of 0 to 1 and combined using a geometric mean to produce the final HDI score, ranging from 0 (lowest development) to 1 (highest development).

Data Collection Methodology

The HDI relies on data compiled from multiple international organizations, including:

- **The World Bank** (for income and economic indicators)
- **UNESCO Institute for Statistics** (for education data)
- **World Health Organization (WHO)** (for life expectancy data)

The UNDP verifies, standardizes, and aggregates this data annually to ensure comparability across countries.

Changes in Recent Versions

- **Inequality-Adjusted HDI (IHDI):** Introduced to reflect disparities in human development within countries.
- **Planetary Adjusted HDI (PHDI):** A recent experimental metric incorporating environmental sustainability.
- **Gender Development Index (GDI):** A disaggregated HDI measure assessing gender gaps in human development.

Citation

When using HDI data in research, please cite:
United Nations Development Programme (UNDP). Human Development Report [1990-2023]. New York: UNDP. Available at http://hdr.undp.org.

Corruption Perceptions Index (CPI)

Overview

The **Corruption Perceptions Index (CPI)** is an annual ranking developed by **Transparency International** to measure perceived levels of public sector corruption worldwide. The index assesses how corrupt a country's government and institutions are perceived to be, based on expert assessments and surveys. It is widely used by policymakers, researchers, and organizations to track corruption trends and advocate for transparency reforms.

Key Features

- **Temporal Coverage:** 1995–present (updated annually)
- **Geographical Coverage:** Over 180 countries and territories
- **Data Points:** Corruption perception scores and rankings, with country-specific analysis

Methodology

The CPI is compiled using **data from multiple independent sources**, including **surveys and expert assessments** conducted by institutions such as:

- **World Bank**
- **World Economic Forum**
- **African Development Bank**
- **Bertelsmann Foundation**
- **Economist Intelligence Unit (EIU)**

Each country receives a **CPI score on a scale of 0 to 100**, where:

- **0** = Highly corrupt
- **100** = Very clean (low corruption perception)

The **final CPI score** for each country is calculated by aggregating and standardizing data from at least three independent sources, ensuring comparability and reliability.

Data Collection and Reliability

- The CPI is based **only on perception-based data** rather than direct corruption measurements, as corruption is often hidden and difficult to quantify.
- Each year, Transparency International refines the methodology to improve consistency and account for **regional and economic differences** in corruption reporting.

Recent Changes and Trends

- **Sector-Specific Analysis:** Recent reports have highlighted corruption in **judicial systems, law enforcement, and procurement processes**.
- **Focus on Global Anti-Corruption Efforts:** New indicators track the effectiveness of **anti-corruption policies** in different regions.
- **Regional Disparities:** Scandinavian countries consistently rank as the least corrupt, while regions affected by conflict or weak governance score lower.

Citation

When using CPI data in research, please cite:
Transparency International. Corruption Perceptions Index [1990-2023]. Berlin: Transparency International. Available at https://www.transparency.org/en/cpi.

Economic Factors:

Exchange Rates Datasheet

General Information:

- **Indicator:** Official Exchange Rate (LCU per US$, Period Average)
- **Definition:** The exchange rate determined by national authorities or in the legally sanctioned exchange market, calculated as an annual average based on monthly averages.
- **Formula:** Market-based calculation
- **ID:** PA.NUS.FCRF
- **Source:** International Monetary Fund, International Financial Statistics
- **License:** CC BY-4.0

Aggregation & Methodology:

- **Development Relevance:**
    - Influences household, producer, and government resource allocation decisions
    - Reflects the interaction of economic agents within a country and globally
- **Limitations & Exceptions:**
    - Market rates may not reflect real purchasing power across different economies
    - Preferred alternative: Purchasing Power Parity (PPP) conversion factor
- **Statistical Concept & Methodology:**
    - Exchange rates are determined by governments or market forces
    - Countries with multiple exchange rate arrangements may have principal, secondary, and tertiary rates

Periodicity & Topic:

- **Periodicity:** Annual
- **Topic:** Financial Sector - Exchange Rates & Prices

Data Accessibility:

- Available online via the International Monetary Fund (IMF) database
- Metadata includes country-specific methodologies and data caveats

Inflation Statistics Datasheet

General Information:

- **Indicator**: Inflation, Consumer Prices (Annual %)
- **Definition**: Annual percentage change in the cost to the average consumer of acquiring a basket of goods and services
- **Formula**: Laspeyres formula
- **ID**: FP.CPI.TOTL.ZG
- **Source**: International Monetary Fund, International Financial Statistics and data files
- **License**: CC BY-4.0

Aggregation & Methodology:

- **Aggregation Metho**d: Median
- **Periodicity**: Annual
- **Topic**: Financial Sector - Exchange Rates & Prices
- **Long Definition**: Inflation as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. The Laspeyres formula is generally used.

Data Accessibility:

- Available online via the International Monetary Fund (IMF) database
- Metadata includes country-specific methodologies and data caveats

---

Unemployment Statistics Datasheet

General Information:

- **Indicator**: Unemployment, Total (% of Total Labor Force) (Modeled ILO Estimate)
- **Definition**: The share of the labor force that is without work but available for and seeking employment.
- **Formula**: Standardized measurement based on labor force surveys
- **ID**: SL.UEM.TOTL.ZS

- **Source**: International Labour Organization, "ILO Modelled Estimates and Projections Database (ILOEST)"
- **License**: CC BY-4.0

Aggregation & Methodology:

- **Aggregation Method**: Weighted Average
- **Development Relevance**:
    - Low unemployment rates can coexist with substantial poverty, while high unemployment can occur in economically developed nations.
    - Youth unemployment is a critical policy issue affecting economic and social stability.
    - Widespread youth unemployment hinders innovation and competitive human capital investment.
    - Unemployment is a key indicator for monitoring progress towards Sustainable Development Goal (SDG) 8.5.2.
- **Limitations & Exceptions:**
    - Definitions of "seeking work" vary by country.
    - Measuring employment in agriculture and informal sectors is challenging.
    - Hidden unemployment and discouraged workers may not be reflected in official counts.
    - Women face additional structural and social barriers affecting unemployment measurement.

Periodicity & Topic:

- **Periodicity: Annual**
- **Topic: Social Protection & Labor - Unemployment**

Statistical Concept & Methodology:

- Unemployment includes individuals actively seeking work and available for employment.
- Some unemployed individuals may not actively seek work but have arrangements for future employment.
- Labor force surveys are the primary data source, supplemented by household surveys and censuses.
- Imputed observations for missing data help track regional and global trends but should not be used for country comparisons.
- The ILO applies multiple models for estimation and projections.

Data Accessibility:

- Available online via the ILOSTAT database
- Metadata includes country-specific methodologies and data caveats

---

Inflation Statistics Datasheet

General Information:

- **Indicator**: Inflation, Consumer Prices (Annual %)
- **Definition**: Annual percentage change in the cost to the average consumer of acquiring a basket of goods and services
- **Formula**: Laspeyres formula
- **ID**: FP.CPI.TOTL.ZG
- **Source**: International Monetary Fund, International Financial Statistics and data files
- **License**: CC BY-4.0

Aggregation & Methodology:

- **Aggregation Method**: Median
- **Periodicity**: Annual
- **Topic**: Financial Sector - Exchange Rates & Prices
- **Long Definition**: Inflation as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. The Laspeyres formula is generally used.

Data Accessibility:

- Available online via the International Monetary Fund (IMF) database
- Metadata includes country-specific methodologies and data caveats

---

Global Tax Revenue Statistics Datasheet

General Information:

- **Document Title:** Constructing the Global Revenue Statistics Database
- **Publication Date:** June 2018
- **Coverage Period:** 1965-2015 (varies by region)
- **Number of Countries Covered:** 80
- **Data Sources:** OECD, National Ministries of Finance, Tax Administrations, National Statistics Offices

Key Indicators:

- **Tax-to-GDP Ratio:**
    - Measures total tax revenues as a percentage of GDP
    - Available for each country and regional averages
- **Tax Structure (% of Total Tax Revenue):**
    - Classification by tax category:
        - Income, profits, and capital gains (1000)
        - Social security contributions (2000)
        - Payroll and workforce (3000)
        - Property taxes (4000)
        - Taxes on goods and services (5000)
        - Other taxes (6000)
- **Revenue Data:**
    - Revenue in national currency
    - Revenue in USD
    - Revenue as % of GDP

Regional Classification:

- OECD Countries (35 countries) - Data from 1965 onwards
- Latin America & Caribbean (25 countries) - Data from 1990 onwards
- Africa (16 countries) - Data from 2000 onwards
- Asia & Pacific (7 countries) - Data from 1990 onwards

OECD Tax Classification System:

- **1000 - Taxes on Income, Profits, and Capital Gains**
    - Personal Income Tax (1100)
    - Corporate Tax (1200)

- **2000 - Social Security Contributions**
    - Employee (2100), Employer (2200), Self-employed (2300)
- **3000 - Payroll and Workforce Taxes**
- **4000 - Property Taxes**
- **5000 - Goods and Services Taxes**
    - VAT (5111), Sales Tax (5112), Excises (5121), Customs Duties (5123)
- **6000 - Other Taxes**

Methodology:

- Tax revenue figures are submitted annually by national authorities
- GDP data sourced from OECD, IMF, and national accounts
- Data classified using the OECD Interpretative Guide
- Tax revenue reported on a cash or accrual basis

Data Accessibility:

- Available online at the OECD Tax Statistics Database
- Metadata includes country-specific methodologies and data caveats

Full Data Sheet available here; link.

GDP Growth Dataset

Overview:

The GDP Growth Dataset provides annual data on the growth of GDP per capita based on constant local currency, using currency exchange rates, without any adjustment for price differences between countries. It includes standardized measures of economic growth based on internationally recognized sources, allowing for cross-country comparisons and analysis of economic trends.

Key Features:

- **Temporal Coverage:** 1961–2023
- **Geographical Coverage:** Global
- **Data Points:** Annual growth rates of GDP per capita
- **Variables:**
    - GDP per capita growth (annual %)
    - Country code
    - Country
    - Year

Data Collection Methodology:

This dataset was compiled from multiple official sources, primarily the **World Bank** and **OECD – World Development Indicators (WDI)**. The data has been processed by **Our World in Data** to ensure consistency across different reporting standards and methodologies.

- The **World Development Indicators (WDI)** is the primary World Bank collection of development indicators, compiled from officially recognized international sources. It presents the most current and accurate global development data available and includes national, regional, and global estimates.
- Standardized adjustments were applied, such as:
    - Ensuring consistent country names and regional definitions
    - Converting values to a unified format
    - Adding necessary metadata for interpretation

Changes in the Latest Version:

- **Extended Coverage:** Updated data up to 2023
- **Clarifications:** Improved definitions for world region aggregation methodology

Citation:

When using this dataset in research, please cite: the World Bank and OECD (2025). "World Development Indicators – GDP per capita growth dataset. Processed by Our World in Data."

Source:

- **World Bank and OECD – World Development Indicators**
- Retrieved on: **January 24, 2025**
- Retrieved from: World Bank Data Catalog

About the Data:

The dataset aggregates economic growth data from officially recognized institutions. Users must adhere to the sources' licensing terms when reusing this data.

Statistical Concept and Methodology:

The **annual percentage growth rate of GDP per capita** is based on constant local currency. GDP per capita is defined as the **gross domestic product (GDP) divided by midyear population**. It accounts for total economic output adjusted for population size, without deductions for depreciation or depletion of resources.

For additional technical details, refer to the metadata for:

- **Constant U.S. dollar GDP (NY.GDP.MKTP.KD)**
- **Total population (SP.POP.TOTL)**

This dataset provides valuable insights into economic growth trends across countries, supporting research on economic policy, development, and financial forecasting.

Economic Inequality Dataset

Overview:

The Economic Inequality Dataset provides annual data on income disparities within and between countries. This dataset aggregates multiple measures of economic inequality, including pre-tax and post-tax income distribution indicators.

Key Features:

- **Temporal Coverage:** Varies by source, typically covering decades of economic inequality trends.
- **Geographical Coverage:** Global
- **Data Points:** Annual economic inequality metrics per country

Variables:

- **Gini Coefficient (0-1):** Measures income inequality, where higher values indicate greater inequality. Depending on the country and year, this data relates to either pre-tax or post-tax income.
- **Country**

- **Year**

Data Collection Methodology:

This dataset is compiled from various reputable sources, including:

- **World Inequality Database (WID):** Provides pre-tax income inequality data, giving insights into income distribution before fiscal interventions.
- **World Bank:** Offers data on post-tax and post-benefit income distribution, highlighting the effects of taxation and government transfers.

To ensure consistency, Our World in Data applied standardized processing, including:

- Standardizing country names and world region definitions.
- Adjusting for inflation to maintain comparability over time.
- Harmonizing income definitions across different data sources.

Changes in the Latest Version:

- **Expanded Data Sources:** Updated to include recent WID and World Bank data.
- **Enhanced Comparability:** Adjusted methodology for cross-country comparisons.

Citation:

When using this dataset in research, please cite:
World Inequality Database (WID); World Bank – "Economic Inequality Dataset." Processed by Our World in Data.

Source:

- World Inequality Database (WID)
    - Retrieved from: https://wid.world
- World Bank
    - Retrieved from: https://data.worldbank.org

About the Data:

The dataset aggregates economic inequality data from officially recognized sources. Users must adhere to the sources' licensing terms when reusing this data.

Statistical Concept and Methodology:

The Economic Inequality Dataset is structured to facilitate cross-country comparisons. Different

measures of inequality allow for a more comprehensive analysis of income distribution and trends.

For additional technical details, refer to:

- Gini Coefficient Methodology
- Palma Ratio Calculations
- Income Share Definitions

This dataset provides valuable insights into global economic disparities and supports research on inequality trends, policy analysis, and socioeconomic development.

Social Factors:

Ethnic or Religious Diversity

The Ethnic Power Relations (EPR) Core Dataset:

The Ethnic Power Relations (EPR) Core Dataset 2021 identifies all politically relevant ethnic groups and their access to state power in every country of the world from 1946 to 2021. It provides annual data on over 800 groups, coding the degree to which their representatives held executive-level state power—from total control of the government to overt political discrimination.

Key Features:

- **Temporal Coverage**: 1946–2021
- **Geographical Coverage**: All countries with a population of at least 250,000 where ethnicity has been politicized
- **Data Points**: Annual data on over 800 ethnic groups
- **Variables**:
    - Ethnic group's access to executive state power
    - The relative size of each group as a share of the total population

Data Collection Methodology:

The dataset was compiled through an online expert survey involving nearly one hundred country and regional experts. These experts identified the ethnic categories most salient for national politics in each country, creating a list of all politically relevant ethnic groups, irrespective of their size. The codings were then reviewed by the EPR Management Committee and in region-specific workshops to ensure inter-coder reliability and global consistency.

Changes in Version 2021:

- **Temporal Extension**: The dataset's coverage has been extended by four years, now including data up to 2021.
- **Clarification**: The category "state collapse" has been defined more explicitly.

Citation:

When using this dataset in the research, please cite:

Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp Hunziker, and Luc Girardin. 2015. "Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Data Set Family." Journal of Conflict Resolution 59(7): 1327–42.

ACD2EPR Dataset:

The ACD2EPR dataset is an extension of the Ethnic Power Relations (EPR) dataset, linking ethnic groups to conflict actors from the Uppsala Conflict Data Program (UCDP). This linkage facilitates the analysis of conflicts involving ethnic groups by providing detailed information on claims, recruitment, and support related to these groups.

Key Variables:

- **Claim**
    - **Type:** Integer
    - **Value Range:** [-1, 2]
    - **Description:** Indicates whether a rebel group claims to represent an ethnic group.
    - **Coding Details:**
        - 0: No claim
        - 1: Direct evidence of claim
        - 2: Indirect evidence (e.g., group name)
        - -1: No information
- **Recruitment**
    - **Type:** Integer
    - **Value Range:** [-1, 2]
    - **Description:** Indicates whether a rebel group recruits members from an ethnic group.
    - **Coding Details:**
        - 0: No recruitment
        - 1: Recruitment present
        - 2: Both rebels and the government recruit from the group
        - -1: No information
- **Support**
    - **Type:** Integer
    - **Value Range:** [-1, 2]

- **Description:** Indicates whether a rebel group is supported by at least 50% of an ethnic group's members.
- **Coding Details:**
  - 0: No or little support
  - 1: Large support
  - 2: Group supports both rebels and government (in non-ethnic conflicts)
  - -1: No information

Recent Updates:

The latest version of the dataset links ethnic groups from the EPR dataset directly to UCDP conflict actors (version 20.1), enhancing the dataset's applicability for researchers analyzing ethnic dimensions of conflicts.

Citation:

For the main structure of the dataset and when referring to the original ACD2EPR v.1.2 dataset, please cite: Wucherpfennig, Julian, Nils W. Metternich, Lars-Erik Cederman, and Kristian Skrede Gleditsch. 2012. "Ethnicity, the State, and the Duration of Civil War." *World Politics* 64(1): 79-115.

Ethnicity of Refugees Dataset:

The Ethnicity of Refugees (ER) dataset provides detailed information on the ethnic composition of refugee populations between neighboring countries worldwide from 1975 to 2020. It identifies up to the three largest ethnic groups within each country-dyadic refugee population and indicates their respective shares of the total refugee stock. The dataset covers refugee groups consisting of at least 2,000 individuals in neighboring countries or countries in proximity (maximum distance between borders ≤ 950 km). Information on countries of asylum and origin, as well as refugee stocks, is based on data from the United Nations High Commissioner for Refugees (UNHCR) and the United Nations Relief and Works Agency for Palestine Refugees (UNRWA).

Key Variables:

- **CoA (Country of Asylum):** The country where refugees have sought asylum.
- **ccode_coa:** GWcode of the country of asylum.
- **CoO (Country of Origin):** The country from which refugees have fled.
- **ccode_coo:** GWcode of the country of origin.
- **year:** The year of observation.
- **totalrefugees:** Total number of refugees and people in refugee-like situations.

- **minimal_distance:** Minimal distance between the country of origin and country of asylum in kilometers.
- **groupname1, groupname2, groupname3:** Names of the first, second, and third largest ethnic refugee groups, respectively.
- **GWgroupid1, GWgroupid2, GWgroupid3:** GWgroupid of the first, second, and third ethnic refugee groups, respectively.
- **groupshare1_num, groupshare2_num, groupshare3_num:** Categorical size of the first, second, and third ethnic refugee groups, respectively, with values indicating "Dominant," "Majority," or "Minority."
- **group1share_multiplier, group2share_multiplier, group3share_multiplier:** Estimated share multipliers for the first, second, and third ethnic refugee groups, respectively.
- **group1_size, group2_size, group3_size:** Estimated number of refugees belonging to the first, second, and third ethnic refugee groups, respectively.

Data Collection Methodology:

The definition of "ethnicity" in the ER dataset is based on the Ethnic Power Relations (EPR-ETH) dataset, which identifies all politically relevant ethnic groups in a country and records the level of access to state power by their representatives. The definition includes ethno-linguistic, racial, and ethno-religious groups. Coders used the EPR-ETH dataset to identify ethnic groups living in refugee-sending countries. In some cases, refugees belonging to politically irrelevant ethnic groups not recorded in the EPR dataset were also coded in the ER dataset.

Citation:

When using this dataset in the research, please cite: Rüegger, Seraina, and Heidrun Bohnet. 2018. "The Ethnicity of Refugees (ER): A new dataset for understanding flight patterns." *Conflict Management and Peace Science* 35(1): 65-88.

Historical Index of Ethnic Fractionalization Dataset (HIEF):

The Historical Index of Ethnic Fractionalization Dataset (HIEF) provides an ethnic fractionalization index for 165 countries across all continents from 1945 to 2013. The dataset measures ethnic diversity by calculating the probability that two randomly drawn individuals within a country do not belong to the same ethnic group.

Key Information:

- **Time Coverage**: **1945–2013** (annually)
- **Geographical Coverage**: **165 countries** across all continents

- **Data Source**:
    - The dataset originates from the **Composition of Religious and Ethnic Groups (CREG) project** at the **Cline Center for Democracy**, University of Illinois at Urbana-Champaign.
    - The **original dataset contained inconsistencies** (e.g., duplicated ethnic group data), which were **checked and corrected** before compiling HIEF.

Ethnic Fractionalization Index (EF Index) Calculation:

The Ethnic Fractionalization Index (EF Index) is computed using the widely applied Herfindahl concentration index transformation:

$$EF_c = 1 - \sum_{i=1}^{n} S_i^2$$

where:

- EF_c = Ethnic fractionalization index for country c.
- S_i = Share of ethnic group i in the total population of country c.
- n = Number of ethnic groups in a country c.

Interpretation of EF Index:

- **0.00** → No ethnic diversity (all individuals belong to the same ethnic group).
- **1.00** → Maximum ethnic diversity (each individual belongs to a different ethnic group).

The unit of observation in the dataset is country-year.

Key variables:

- **Country**: Name of the country
- **Year**: Corresponding year (1945–2013); for some countries, data is available only after independence.
- **EFindex**: Ethnic fractionalization index for that country-year.

This dataset is particularly useful for analyzing patterns of ethnic diversity over time and across different countries.

Population

Urban population (% of total population) Dataset:

The Urban Population (% of total population) Indicator provides insights into the annual percentage of the urban population across various countries and regions. This indicator is essential for understanding urbanization trends.

Key Details:

- **Indicator Code**: SP.URB.TOTL.IN.ZS
- **Indicator Name**: Urban population (% of total population)
- **Definition**: This indicator measures the size of the urban population as a percentage of the total population, which includes individuals residing in areas defined as urban by national statistical offices. The data are collected and smoothed by the United Nations Population Division.
- **Source**: United Nations Population Division. World Urbanization Prospects: 2018 Revision.
- **Topic**: Environment: Density & Urbanization
- **Periodicity**: Annual
- **Aggregation Method**: Weighted average
- **Statistical Concept and Methodology**: The urban population is people living in urban areas as defined by national statistical offices. The indicator is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects. The urban percentages are the numbers of persons residing in an area defined as "urban" per 100 total population. Countries differ in the way they classify populations as "urban" or "rural."

This indicator is part of the World Development Indicators database, which offers a comprehensive collection of development data covering various aspects of global development.

Healthcare

Healthcare human resources Dataset:

The OECD Healthcare Resources dataset provides comprehensive statistics on healthcare employment, physical and technical resources, and medical technology across OECD member and partner countries. It covers various indicators related to the healthcare workforce, hospital infrastructure, and medical technology availability.

Overview:

- **Dataset Name**: OECD Healthcare Resources
- **Version**: OECD Health Statistics 2023
- **Publication Year**: July 2023
- **Maintaining Institution**: Organisation for Economic Co-operation and Development (OECD)
- **Data Source**: OECD Health Statistics
- **Geographical Coverage**: OECD member and partner countries
- **Temporal Coverage**: Varies by indicator
- **Data Format**: Structured country-year dataset
- **Unit of Observation**: Country-Year

Key Variables and Categories:

- **Healthcare Employment and Education**
    - **Total health and social employment**
    - **Physicians**
        - Practicing physicians
        - Professionally active physicians
        - Physicians licensed to practice
    - **Physicians by age group and gender**
    - **Physicians by category**
    - **Midwives**
        - Practicing midwives
        - Professionally active midwives
        - Midwives licensed to practice
        - **Nurses**
            - Practicing nurses
            - Professionally active nurses
            - Nurses licensed to practice
            - Professional nurses
            - Associate professional nurses
        - **Caring personnel (Personal care workers)**
            - Practicing caring personnel
            - Professionally active caring personnel
        - **Dentists**
            - Practicing dentists
            - Professionally active dentists
            - Dentists licensed to practice

        - **Pharmacists**
            - Practicing pharmacists
            - Professionally active pharmacists
            - Pharmacists licensed to practice
        - **Physiotherapists**

- **Physical and Technical Resources**
    - **Hospital employment**
    - **Graduates**
        - Medical graduates
        - Dentists graduates
        - Pharmacists graduates
        - Midwives graduates
        - Nursing graduates
        - Professional nursing graduates
        - Associate professional nursing graduates
    - **Remuneration of health professionals**
        - General practitioners
        - Specialists
        - Hospital nurses

- **Hospital Infrastructure**
    - **Hospitals**
        - Publicly-owned hospitals
        - Not-for-profit privately owned hospitals
        - For-profit privately owned hospitals
        - General hospitals
    - **Hospital beds**
        - Hospital beds by sector (public, private, not-for-profit)
        - Hospital beds by function (curative, rehabilitative, psychiatric, long-term)
    - **Intensive Care Unit (ICU) beds and usage**

- **Medical Technology**
    - **Medical technology equipment**
        - Computed Tomography (CT) scanners

    - Magnetic Resonance Imaging (MRI) units
    - Positron Emission Tomography (PET) scanners
    - Gamma cameras
    - Mammographs
    - Radiation therapy equipment

Data Collection and Methodology:

- The dataset is compiled from OECD Health Statistics, which gathers healthcare resource data from national statistical offices, ministries of health, and international health organizations.
- Data definitions and methodology are standardized across OECD member countries to ensure comparability.
- Differences in national classifications and reporting standards may exist.

Applications and Use Cases:

The dataset is valuable for:

- Healthcare workforce planning and policy-making
- Hospital and healthcare infrastructure assessments
- Comparative healthcare analysis across OECD countries
- Trends in medical technology and resource allocation

Citation and Attribution:

If using the OECD Healthcare Resources Dataset, please cite it as follows: OECD (2023). *OECD Health Statistics 2023: Healthcare Resources Dataset*. Organization for Economic Co-operation and Development.

Healthcare Coverage Dataset:

The OECD Health Protection dataset provides comprehensive statistics on health insurance coverage, social protection in healthcare, and voluntary health insurance across OECD member and partner countries. It covers three major components:

- Total Public and Primary Voluntary Health Insurance (HF.1 & HF.2.1)
- Government/Compulsory Health Insurance (HF.1)

- Voluntary Health Insurance (HF.2.1)

These datasets track the coverage and financing mechanisms for healthcare services, differentiating between publicly funded and privately funded health insurance systems.

Overview:

- **Dataset Name**: OECD Health Protection
- **Version**: OECD Health Statistics 2024
- **Publication Year**: November 2024
- **Maintaining Institution**: Organisation for Economic Co-operation and Development (OECD)
- **Data Source**: OECD Health Statistics
- **Geographical Coverage**: OECD member and partner countries
- **Temporal Coverage**: Varies by dataset and indicator
- **Data Format**: Structured country-year dataset
- **Unit of Observation**: Country-Year

Key Variables and Categories:

- **Government and Compulsory Health Insurance (HF.1)**
- **Definition**: The proportion of the population eligible for a defined set of basic healthcare goods and services under government-financed schemes, social health insurance, compulsory private insurance, or medical savings accounts.
- **Eligibility Criteria**:
    - Population covered under **social health insurance (SHI)**
    - **Universal healthcare systems** (e.g., UK NHS, Canada's Medicare)
    - **Compulsory private insurance** (e.g., Switzerland, Netherlands)
    - **State-subsidized insurance programs**
- **Source of Funding**:
    - **General taxation**
    - **Payroll-based contributions**
    - **Compulsory employer contributions**

- **Voluntary Health Insurance (HF.2.1)**
- **Definition**: Private health insurance schemes financed through private health premiums, where purchase is not mandated by law.
- **Types of Voluntary Health Insurance**:
    - **Primary VHI**: Used as the **only** source of healthcare coverage (e.g., Chile, US uninsured market).

    - **Duplicate VHI**: Provides **f**aster access or expanded provider choice for services already covered under public insurance (e.g., Australia, UK).
    - **Complementary VHI**: Covers co-payments, cost-sharing, or non-covered services (e.g., France, Belgium).
    - **Supplementary VHI**: Covers additional services not provided by public insurance, such as dental, vision, and alternative medicine.

- **Examples of Funding Mechanisms**:
    - Employer-sponsored health benefits
    - Self-funded private insurance schemes
    - Government-subsidized voluntary health insurance (e.g., Medicaid expansion in the US)

- **Social Protection in Healthcare**
- **Definition**: Broad policy measures aimed at ensuring financial protection in accessing healthcare services.
- **Components**:
    - **Universal Health Coverage (UHC) implementation**
    - **Out-of-pocket expenditure tracking**
    - **Healthcare affordability and equity metrics**
    - **Social safety nets for vulnerable populations**

Data Collection and Methodology:

The dataset is compiled from OECD Health Statistics, which gathers healthcare financing and coverage data from:

- National Ministries of Health
- Social Security Institutions
- Insurance Regulatory Authorities
- International Health Organizations (WHO, World Bank, Eurostat)

Standardized OECD SHA Classification (System of Health Accounts) ensures comparability across countries.

Applications and Use Cases:

This dataset is essential for:

- Comparing health insurance coverage models across OECD countries
- Evaluating the effectiveness of public vs. private health financing
- Tracking trends in voluntary health insurance uptake

- Assessing financial barriers to healthcare access
- Monitoring social protection policies in healthcare

Citation and Attribution:

If using the OECD Health Protection Dataset, please cite it as follows: OECD (2024). *OECD Health Statistics 2024: Health Protection Dataset*. Organization for Economic Co-operation and Development.

Level of Education

Government Expenditure on Education, total % (of GDP) Dataset

Indicator Description:

- **Definition**: This indicator measures general government expenditure on education (current, capital, and transfer spending) as a percentage of GDP. It includes spending funded by transfers from international sources to the government. General government typically refers to local, regional, and central governments.
- **Calculation Method**: The total general government expenditure on all levels of education is divided by GDP and multiplied by 100 to express it as a percentage.
- **Data Collection**:
- **Education data**: Collected by the UNESCO Institute for Statistics (UIS) through annual education surveys.
- **GDP data**: Obtained from the World Bank.
- **Education levels**: Aligned with the **International Standard Classification of Education (ISCED)** framework to ensure international comparability.
- **Academic year reference**: The reference year corresponds to the **end year of the academic cycle** (e.g., if the academic year runs from September 2010 to June 2011, the reference year is 2011).

Key Variables:

- **Country**: Name of the country
- **Year**: The year of data observation
- **SE.XPD.TOTL.GD.ZS**: Government expenditure on education, total (% of GDP)

Development Relevance:

- The percentage of GDP allocated to education is a useful metric for comparing government prioritization of education across countries.
- Higher percentages suggest a greater policy emphasis on education and stronger public revenue capacity.
- Countries with significant private sector or household contributions to education financing may have lower government expenditure shares.

Limitations and Considerations:

- Data may only reflect expenditures by the Ministry of Education, excluding spending by other government departments.
- Variations in definitions and reporting methods exist across countries.
- Some countries do not provide annual updates, leading to gaps in data availability.

Democratic Factors:

Liberal Democracy Index Dataset

Overview:

The Liberal Democracy Index Dataset provides annual data on the extent to which the principles of liberal democracy are upheld globally. It measures protections for individual and minority rights, the independence of the judiciary, and the presence of effective checks and balances on executive power. This dataset also incorporates elements of electoral democracy.

Key Features:

- Temporal Coverage: 1789–2023
- Geographical Coverage: Global
- Data Points: Annual liberal democracy scores per country

Variables:

- Liberal Democracy Index (0-1): Higher values indicate stronger protections for civil liberties, independent judiciary, and government constraints.
- Country Code
- Country
- Year

Data Collection Methodology:

This dataset is compiled from the Varieties of Democracy (V-Dem) Country-Year Dataset. V-Dem relies on expert assessments from approximately 3,500 country specialists to evaluate multiple dimensions of democracy. The dataset is managed by the V-Dem Institute at the University of Gothenburg, Sweden.

To ensure consistency across time and geographical regions, Our World in Data applied standardized processing, including:

- Standardizing country names and regional classifications
- Averaging country values to generate regional aggregates
- Enhancing methodological consistency for comparability

Changes in the Latest Version:

- Extended Coverage: Updated data up to 2023
- Clarifications: Improved definitions for regional aggregation methods

Citation:

When using this dataset in research, please cite:
V-Dem (2024) – "Liberal Democracy Index Dataset." Processed by Our World in Data.

Source:

Varieties of Democracy (V-Dem) – Country-Year Dataset
Retrieved on: March 18, 2024
Retrieved from: http://v-dem.net/vdemds.html

About the Data:

The dataset aggregates democracy scores from officially recognized sources. Users must adhere to the sources' licensing terms when reusing this data.

Statistical Concept and Methodology:

The Liberal Democracy Index is based on an interval scale (0-1), where higher values indicate stronger democratic protections. It evaluates constitutional constraints, executive oversight, and civil liberties. The dataset is structured to facilitate cross-country comparisons and trend analysis in governance and political stability.

This dataset provides valuable insights into the evolution of liberal democracy worldwide and supports research on governance, political participation, and institutional development.

Countries Where Armed Conflicts Took Place Dataset

Overview:

This dataset provides annual records of countries where at least one armed conflict event took place. An armed conflict is defined as a disagreement between organized groups, or between one organized group and civilians, resulting in at least 25 deaths during a given year. The dataset includes both combatant and civilian deaths due to fighting.

Key Features:

Temporal Coverage: 1989–2023
Geographical Coverage: Global
Data Points: Annual binary indicators of conflict presence per country

Variables:

- **Conflict Presence (0/1):**
    - **1:** At least one armed conflict event occurred in the given country during the year.
    - **0:** No recorded armed conflict event in the given country during the year.
- **Country Code**

- **Country**
- **Year**

Data Collection Methodology:

This dataset is compiled from the Uppsala Conflict Data Program's (UCDP) Georeferenced Event Dataset (GED). UCDP records events of organized violence occurring at a specific time and location, allowing for precise geo-coding of conflict events down to the level of individual villages and daily timeframes.

To ensure consistency, Our World in Data applied standardized processing, including:

- Standardizing country names and world region definitions.
- Mapping event coordinates to country borders using the Natural Earth dataset.
- Assigning conflict events falling outside country borders to the closest relevant country.

Changes in the Latest Version:

- **Extended Coverage:** Updated data up to 2023.
- **Improved Mapping:** Enhanced assignment of conflict events using geo-coded data.

Citation:

When using this dataset in research, please cite:
Uppsala Conflict Data Program (2024); Natural Earth (2022) – "Countries Where Armed Conflicts Took Place Dataset." Processed by Our World in Data.

Source:

- Uppsala Conflict Data Program – Georeferenced Event Dataset
    - Retrieved on: August 26, 2024
    - Retrieved from: https://ucdp.uu.se/downloads/index.html#ged_global
- Natural Earth – Large Scale Data (1:10m Cultural Vectors)
    - Retrieved on: November 28, 2023
    - Retrieved from: https://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-admin-0-countries/

About the Data:

The dataset aggregates conflict event data from officially recognized sources. Users must adhere to the sources' licensing terms when reusing this data.

Statistical Concept and Methodology:

The dataset is structured to facilitate cross-country comparisons of conflict incidence. Conflict event locations are mapped to country borders where possible, with exceptions made for disputed areas.

For additional technical details, refer to:

- UCDP Georeferenced Event Dataset Metadata
- Natural Earth Data Documentation

This dataset provides critical insights into the geography of armed conflicts and supports research on war, peace, and conflict resolution worldwide.

Section 4: Project Plan

Below is the Gantt chart representing the project plan over time. Each project phase is divided into tasks and the expected duration for each task is given. The mid-term demo and final submission dates are highlighted. The Gantt chart can be viewed in full here.

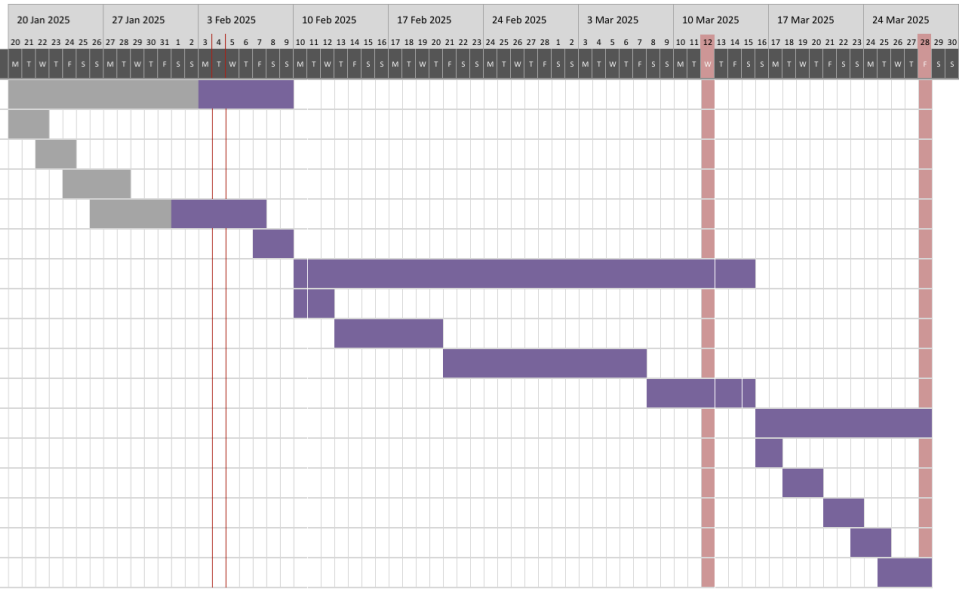
DSGS: “Forecasting Political Outcomes in Contemporary Government”

Connor Boyd-Lyon, Yeni Jung, Zehn-Ul-Abideen Sharif, Matthew Randall, Evelina Ivanova, Yarkin Yorulmaz

Project Start: Mon, 20/01/2025

Display Week: 1

TASK		ROLES	PROGRESS	START	END
Data Collection and Refining Scope			70%	20/1/25	9/2/25
Task 1	Choosing an area of interest	All	100%	20/1/25	22/1/25
Task 2	Researching the topic	All	100%	22/1/25	24/1/25
Task 3	Collecting data	All	100%	24/1/25	28/1/25
Task 4	Initial cleaning	All	50%	26/1/25	7/2/25
Task 5	Clustering to refine scope	Connor, Evelina	0%	7/2/25	9/2/25
Creating a Model to Predict Election Outcomes			0%	10/2/25	15/3/25
Task 1	Research Model Architecture	Yeni, Yarkin	0%	10/2/25	12/2/25
Task 2	Algorithm Selection & Baseline Model Setup	Zehn, Matthew	0%	13/2/25	20/2/25
Task 3	Hyperparameter Tuning & Optimization	All (should be done in c	0%	21/2/25	7/3/25
Task 4	Final Model Selection & Training	All (should be done in c	0%	8/3/25	15/3/25
Evaluate and Deploy an Election Prediction Model			0%	16/3/25	28/3/25
Task 1	Model Evaluation Metrics & Performance Analysis	Yeni, Evelina	0%	16/3/25	17/3/25
Task 2	Result Interpretation & Insights Extraction	Connor, Matthew	0%	18/3/25	20/3/25
Task 3	Bias & Fairness Analysis	Yarkin, Zehn	0%	21/3/25	23/3/25
Task 4	Data-Driven Decision Making & Storytelling	Yeni, Connor, Evelina	0%	23/3/25	25/3/25
Task 5	Documenting	All (should be unified)	0%	25/3/25	28/3/25



References

The Organisation for Economic Co-operation and Development (OECD), 2025. url: https://www.oecd.org/en.html (Accessed 03/02/2025).