Final Project Paper

1. Introduction:

During the past years, ESG has turned into an important criterion of measurement regarding sustainability and ethical performance of companies. Investors, governments, and companies use ESG measures as a way to define the long-term sustainability of a company and its impact upon society. Complementing traditional finance metrics in your analysis, these factors allow insight into the behaviors and risks of a company. Tunguz's dataset on Kaggle is quite broad, with over 1,000 companies across the globe, containing a large number of different ESG metrics. It includes data on environmental responsibility, such as carbon footprint, energy use, and many other factors; social factors, such as employee treatment and diversity; and governance practices, including board diversity and executive compensation. This dataset would become really important for analysts, researchers, and investors looking to judge corporate performance through the lens of sustainability.

Key attributes of the dataset are:

- **Environmental Factors:** Indicators for the company regarding its effects on natural resources, levels of pollution, and sustainability development.
- **Social Factors:** Information on how companies are managing the relationship with employees, suppliers, customers, and the communities where their operations are present.
- **Governance Factors:** Data on executive management, compensation, audits and internal controls, and shareholder rights.

2. Dataset Description:

The dataset, titled "ESGCountry.csv," available under ESG Data on Kaggle, comes with the comprehensiveness of country-level ESG data. Actually, it is part of a larger dataset, which concentrates on environmental, social, and governance metrics; as such, the dataset is critical when it comes to determining the sustainability and ethical considerations for countries and regions. This dataset has 239 rows and 31 columns. Some of the key parameters in the dataset include:

1. **Country codes and names**: Unique identifiers that explain each nation, used for easy referencing and classification.

2. **Environmental Data**: Indicators of environmental performance of the nation, such as carbon footprints, energy usage, and concomitant sustainability efforts.

3. **Social Indicators**: Population, education level, income distribution, and health data are indicators of the social facets of sustainability.

4. **Governance**: Indicators that define Government effectiveness, political stability, regulatory quality, and transparency.

5. **Economic Data**: This indicates trade, GDP, income level, and industrial activity-clear insight into the performance of each country.

It would also be helpful for making analysis regarding country performance in diverse dimensions of ESG; useful to investors, policymakers, and researchers concerned with sustainability and global development.

The major issue in this dataset is the presence of a large number of null entries over most of the attributes. For example, in Special Notes, there were 148 nulls; in Country Unit, there were 46 nulls; in Lending categories, there were 96 nulls, and many more.

Summary of null and missing values:

Column	Null Values	Missing Values
Country Code	0	0
Short Name	0	0
Table Name	0	0
Long Name	0	0
2-alpha code	1	1
Currency Unit	46	46
Special Notes	148	148
Region	46	46
Income Group	46	46
WB-2 code	1	1
National accounts base year	47	47
National accounts reference year	173	173
SNA price valuation	47	47
Lending category	96	96
Other groups	180	180
System of National Accounts	49	49
Alternative conversion factor	192	192
PPP survey year	239	239
Balance of Payments Manual in use	51	51
External debt Reporting status	121	121
System of trade	53	53
Government Accounting concept	83	83
...		
Vital registration complete	141	141
Latest agricultural census	118	118
Latest industrial data	96	96
Latest trade data	7	7

Additionally, attributes such as Special Notes, Alternative Conversion Factor, PPP Survey Year, and External Debt Reporting Status are unrelated to environmental and social issues; hence much more importance is given to financial aspects. These parameters should not be considered, since including them would prevent a more focused analysis of sustainability measures.

3. Data Cleaning and Normalization:

In the cleaning process, we came to notice that the presence of missing values could heavily affect the veracity of our results. We also found some columns that did not match our focus on environmental and social governance. To handle this, we have meticulously cleaned and standardized the dataset. We removed columns containing more than 50% null values in our cleaning approach along with the unnecessary ones, and renamed some columns for feasibility (like renamed “Table Name” as “Country”). We preprocessed the rest of the columns by replacing missing values with mean or median and other methods that allowed us to keep our dataset intact. Further, we normalized these numeric columns by applying the Min-Max scaler algorithm, which allowed us to scale all numeric features into one range. Finally, we removed all extra columns that had zero contribution to our analysis.

Data cleaning results and normalization are as follows:
No. of columns – 23

No. of rows – 59
Enumerating significant attributes includes the following:
Country Code, Country, Long Name, 2-alpha Code, Currency Unit, Region, Income Group, WB-2 Code, National Accounts Base Year, SNA Price Valuation, Lending Category, System of National Accounts, Balance of Payments Manual in Use, System of Trade, Government Accounting Concept, IMF Data Dissemination Standard, Latest Population Census, Latest Household Survey, Source of Most Recent Income and Expenditure Data, Latest Agricultural Census, Latest Industrial Data, and Latest Trade Data.

Numerical Columns Descriptive Statistics:

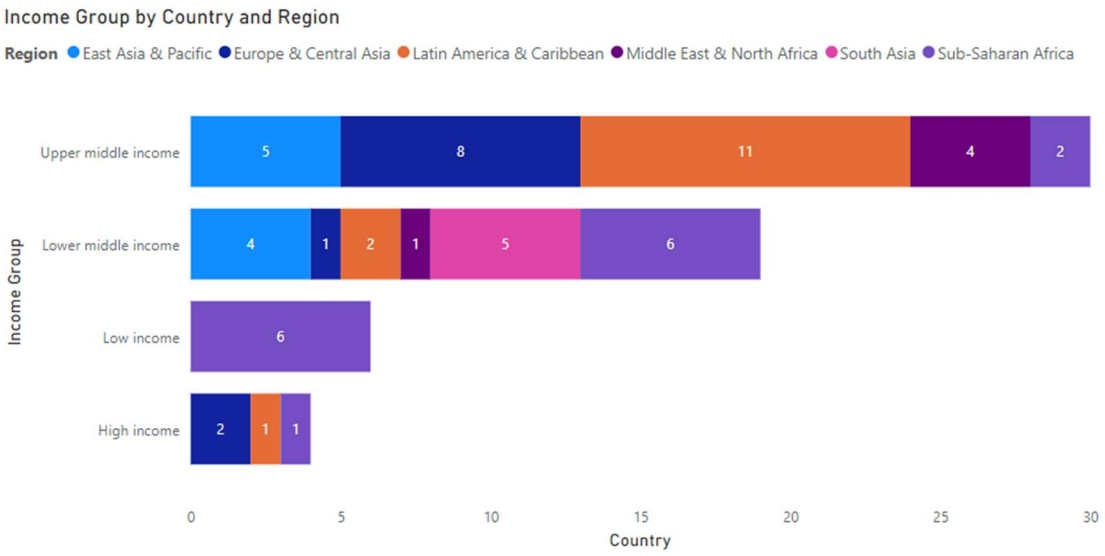
	count	mean	std	min	\
National accounts base year	37.0	2010.756757	4.929533	2000.0	
Latest population census	58.0	2010.913793	9.995235	1943.0	
Latest agricultural census	39.0	2011.282051	2.973114	2006.0	
Latest industrial data	59.0	0.796610	0.246979	0.0	
Latest trade data	59.0	0.977401	0.136519	0.0	

	25%	50%	75%	max
National accounts base year	2007.000000	2011.000000	2015.00000	2018.0
Latest population census	2010.000000	2011.000000	2015.00000	2020.0
Latest agricultural census	2009.000000	2011.000000	2014.00000	2016.0
Latest industrial data	0.646341	0.926829	0.97561	1.0
Latest trade data	1.000000	1.000000	1.00000	1.0

4. Descriptive Analytics:

Descriptive analytics is the first step in data analysis, enabling organizations to gain deep insight into historical data and identify patterns or trends within the data. The ESG Country dataset contains various dimensions at the country level, including population, income groups, and trade data. Descriptive analytics in this dataset plays a crucial role in summarizing and interpreting these dimensions.

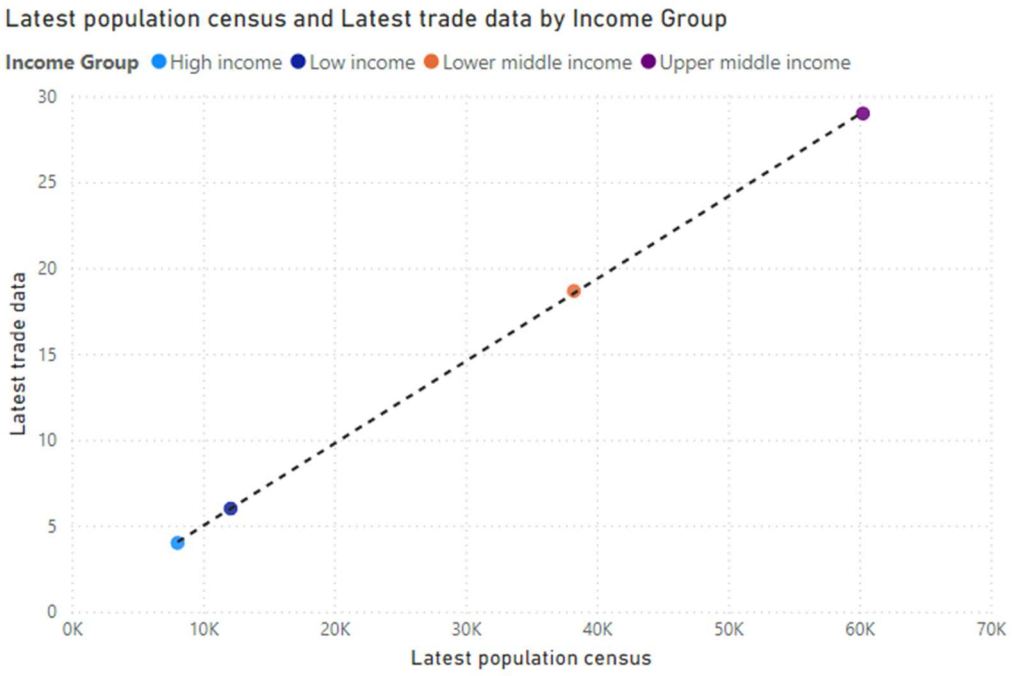
a) Bar Chart of Income Group by Country and Region:



This bar chart shows the distribution of countries in various income groups for each region. Each bar corresponds to one region, and within that bar, there are segments corresponding to low income, lower middle income, upper middle income, and high income. The width of the segment in each bar shows the number of countries belonging to that particular class of income in that region.

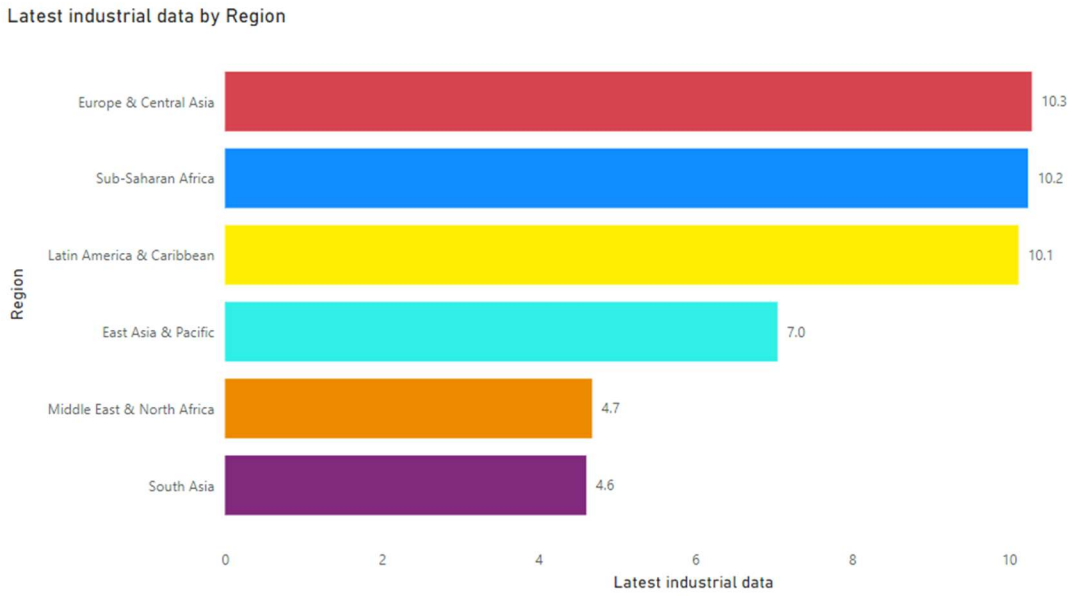
In this regard, the chart shows that the Sub-Saharan Africa region has a certain number of countries in each income group, with the highest numbers in the lower middle as well as low income groups. Similarly, other regions like Europe & Central Asia, Latin America & the Caribbean, and others have varying distributions of countries across the income groups. This visualization helps to understand the economic diversity within each region and how countries are spread across different income levels globally.

b) Scatter Plot of Latest population census and Latest trade data by Income Group:



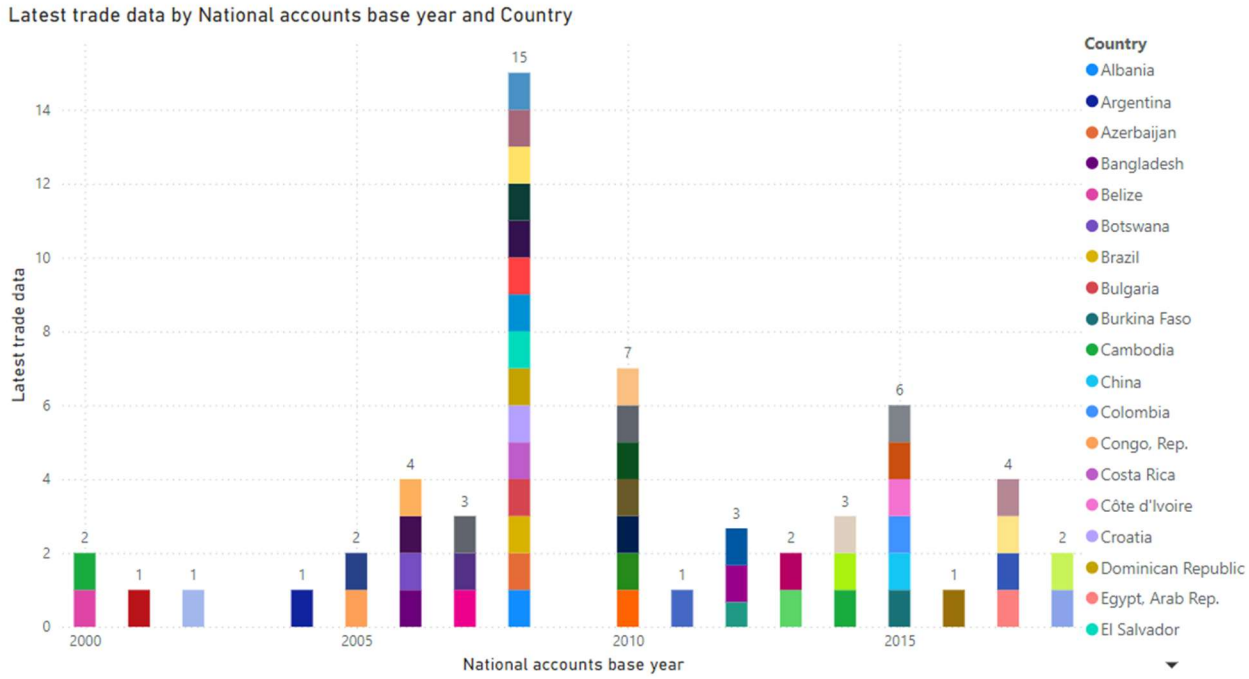
The scatter plot shows a positive correlation between "Latest trade data" and "Latest population census." There are four data points on the graph, each representing different income groups (e.g., low income, lower-middle income, upper-middle income, and high income). A black dashed line on the plot indicates a positive trend, suggesting that as the population census increases, the trade data also tends to increase across different income groups. This visualization helps to understand the relationship between population census and trade data among various income groups globally.

c) Bar Chart of Latest industrial data by Region:



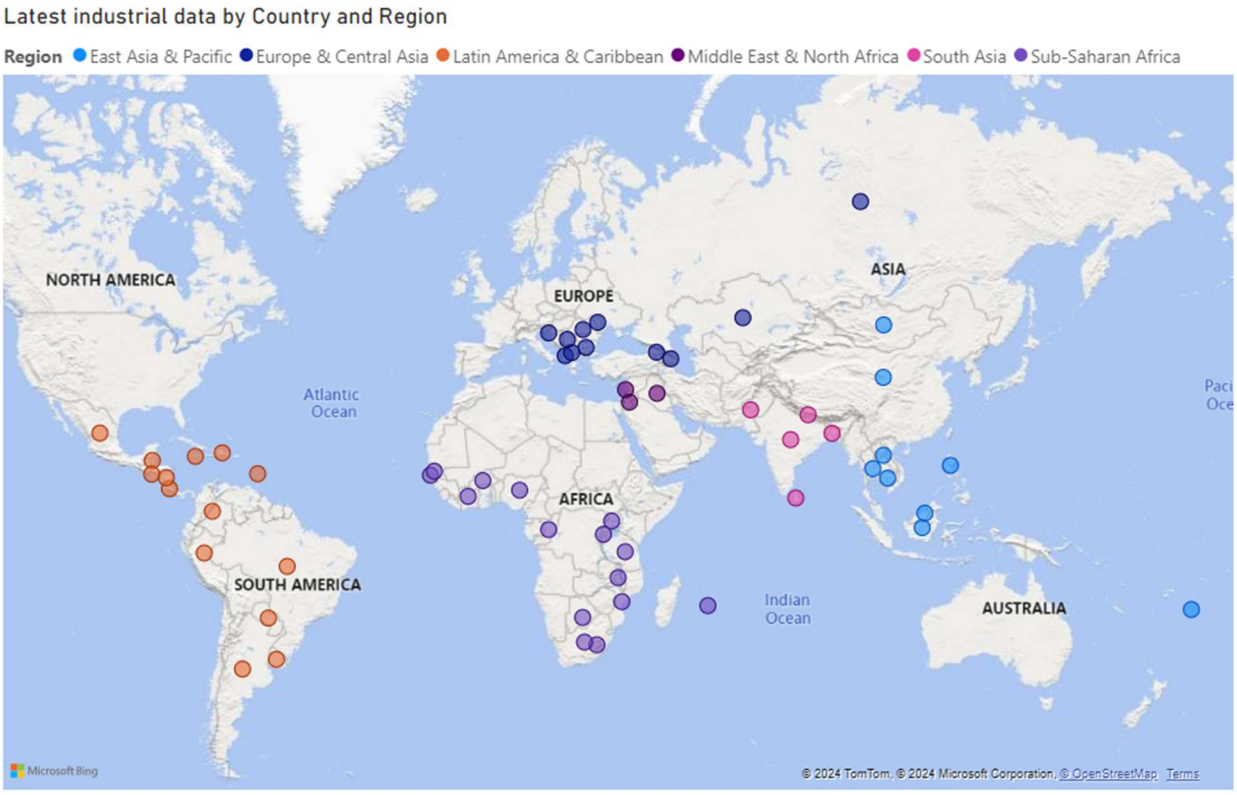
The bar chart offers a comparison of recent industrial data from various regions around the world. The most recent industrial data is shown on the x-axis, with the regions listed on the y-axis. Europe & Central Asia has the highest industrial data value of 10.3, with Sub-Saharan Africa closely following at 10.2 and Latin America & the Caribbean at 10.1. East Asia & Pacific sees a notable drop to 7.0, showing a significant contrast with the top three regions. The industrial data values of 4.7 and 4.6 are reported for Middle East & North Africa and South Asia, respectively, placing them at the bottom of the chart. In general, this chart effectively shows the different levels of industrial activity in various regions, with Europe & Central Asia at the top and South Asia at the bottom in terms of data values. Variations in industrial success can be ascribed to discrepancies in economic circumstances, technological progress, and local regulations.

d) Ribbon chart of Latest trade data by National accounts base year and Country:



The ribbon chart provides a visual representation of the latest trade data by national accounts base year and country, with each bar corresponding to a different year from 2000 to 2018. The height of each bar indicates the number of countries that have up-to-date trade data for that specific year. The chart reveals a trend of increasing numbers of countries with the latest trade data over time. For instance, starting from the year 2000, only 2 countries had the most recent trade data, this number raised to 4 countries by 2006, and surged to 15 countries by 2008. The chart also indicates that the distribution of countries with the latest trade data is not even across the years. Overall, this trend suggests that the number of countries with current trade data is likely to continue rising, as more countries engage in data collection and reporting of trade-related information.

e) Map visualization of Latest industrial data by Country and Region:

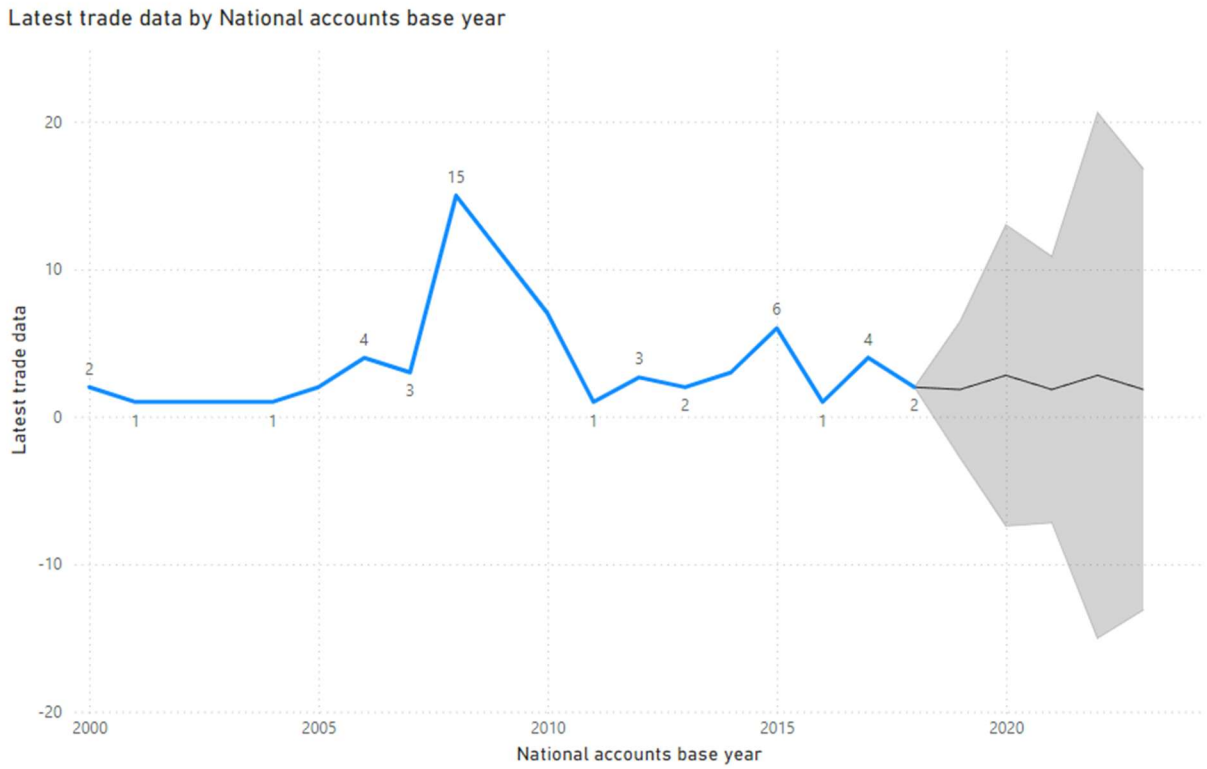


The following map uses geographic markers to display industrial data across various regions. Each region is color-coded according to the region, corresponding with the treemap for continuity. For instance, light blue dots marking industrial centers in countries like China and Southeast Asia comes under East Asia and Pacific region; dark blue dots across Europe, showing a dense concentration of industrial activity falls under Europe and Central Asia region and so on. This map allows for a geographical interpretation of industrial data, showing the spread and concentration of industry across the world. Europe has a high concentration of industrial centers, as do parts of East Asia and the Americas. Africa and the Middle East have a more dispersed industrial presence.

5. Predictive Analytics:

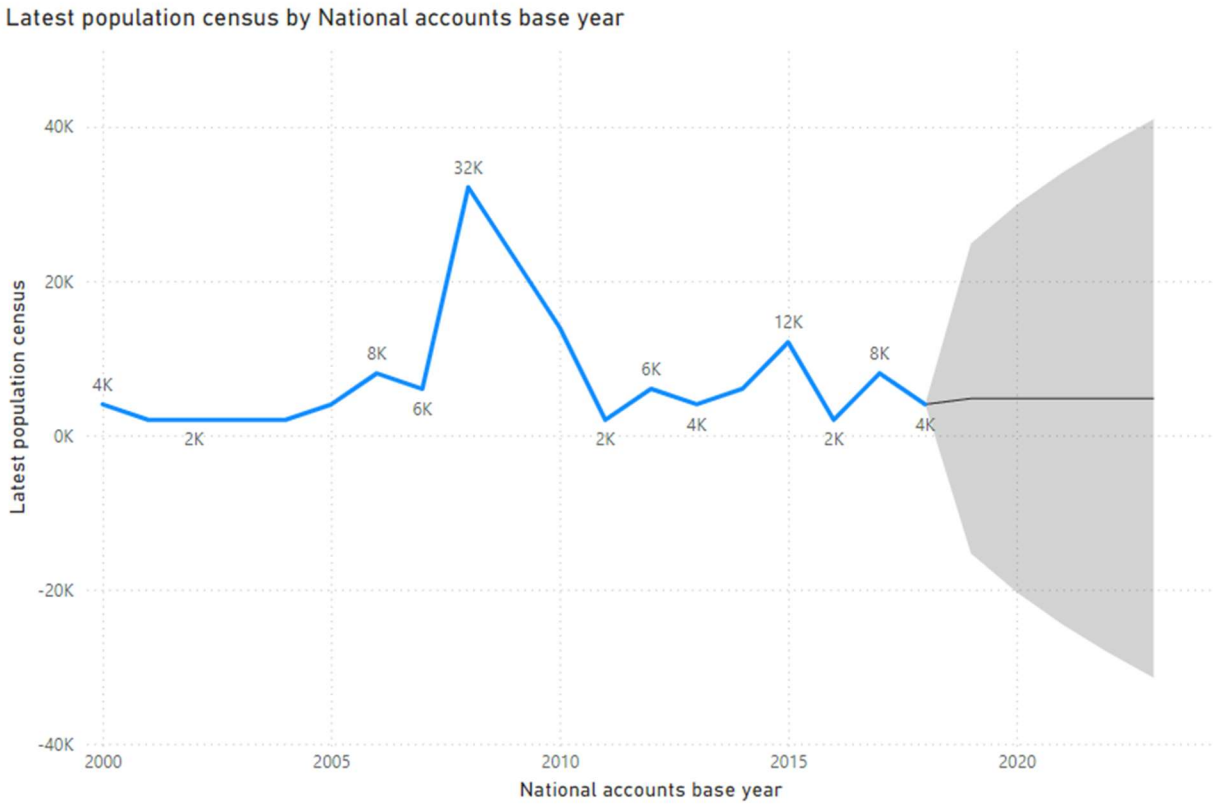
Predictive analytics comprises of historical data and statistical models to forecast future outcomes, thus helping organizations in anticipating trends and making proactive decisions. In the ESG Country dataset, which includes key metrics like population data, income groups, and trade information, predictive analytics can be applied to forecast future changes in population growth, economic shifts, or trade volumes. By leveraging techniques such as time series forecasting, we can predict future population census trends, trade performance, and how different regions or income groups might evolve over time.

a. Line Chart of Latest trade data by National accounts base year (Time series forecasting):



The line chart illustrates the latest trade data by national accounts base year with trade data forecast. The trade data depicts moderate fluctuations between 2000 and 2005, hovering around 2 to 4 units, with a significant peak around 2008, where the trade data spikes dramatically to 15 units. This could indicate a period of high trade activity or economic boom. The forecast extends beyond 2018 with a relatively flat trend. The trade data is predicted to sustain around 2 units, indicating potential stagnation or stability in future trade activity. The confidence interval (the shaded area) grows wider as we move further into the future. This increasing uncertainty suggests that future trade could vary greatly ranging from significant declines to potential recovery and based on external factors like global trade policies, market demand, or geopolitical events.

b. Line Chart of Latest population census by National accounts base year (Forecasting Population Growth):



The line chart depicts the latest population census by national accounts base year with population growth forecast. The population census data shows steady but low growth between 2000 and 2005, remaining around 4K to 6K units. A massive spike is observed around 2008, where the population census data reaches 32K units, indicating a sudden surge in population. The population forecast shows a relatively flat trend, predicting no significant population growth or decline in the near future, with values around 4K units. Similar to the trade data, the confidence interval widens as we move further into the future. This indicates a wide range of possible outcomes, from large population gains to substantial declines, highlighting uncertainty in population dynamics.

6. Conclusion:

The analysis of the ESG Country dataset reveals significant trends and insights related to industrial activities, trade, and populations in various global regions. Europe & Central Asia, Sub-Saharan Africa, and Latin America & the Caribbean stand out as areas with the highest industrial activity, whereas South Asia and the Middle East & North Africa exhibit comparatively lower levels. These differences can be attributed to diverse economic structures, industrial policies, and levels of technological advancement among the regions. Regarding trade, there has been a marked increase in the availability of recent trade data, signifying those nations are prioritizing data collection and reporting more than before. The positive relationship between population size and trade volume, as demonstrated by the scatter plot, indicates that countries with larger populations generally experience higher trade activities. This finding is particularly valuable for policymakers, as it underscores the necessity to invest in infrastructure and regulatory enhancements in areas with growing populations to stimulate trade expansion. Furthermore, predictive analysis offers significant insights into prospective trends, suggesting that while population growth may stabilize, trade activities might fluctuate due to external influences like geopolitical events or changes in the global economy. These insights could form a basis for future policy initiatives aimed at promoting sustainable growth, especially by boosting industrial productivity in underachieving regions such as South Asia. In summary, the results stress the necessity of customizing sustainability, economic, and industrial policies to the distinct characteristics and challenges of each region to encourage balanced and inclusive global development.

7. References:

- Tunguz, B. (2021, April 17). Environment, social and Governance Data. Kaggle. https://www.kaggle.com/datasets/tunguz/environment-social-and-governance-data/data
- Ibm. (2024a, June 7). Topics. IBM. https://www.ibm.com/cloud/learn/descriptive-analytics
- Predictive analytics: What it is and why it matters. SAS. (n.d.). https://www.sas.com/en_us/insights/analytics/predictive-analytics.html
- MiguelMyersMS. (n.d.). Visualization types in power bi - power bi. Power BI | Microsoft Learn. https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-types-for-reports-and-q-and-a
- Medium. (n.d.). https://towardsdatascience.com/introduction-to-time-series-forecasting-6c3e2d2ceb5e

DONE BY –
ABHIJIT BAJRANG MORE