DSA210-Term-Project-Final-Report

Project Overview

I am a Computer Science student from Sabanci University, and this is my Data Science Term Project. During the term, I am planning to analyse the relationship between Space Debris and the global leadership competition among various countries.

During the project this hypothesis will be tested:

- The global leadership race contributes to Space Debris.

Objectives

1- **Understanding which Countries are Competing for Global Leadership**

The Global Leadership race involves several key nations that draw attention because of their crucial technological, economic and military capabilities.

2- **Understanding which Countries are Contributing to Space Debris**

With the increasing number of satellite launches and space missions, Space Debris has become a significant problem. Furthermore, despite the fact that private companies started to engage in space missions, countries still have the largest share of Space Debris.

3- **Applying Data Science Skills**

Utilize the data science skills acquired in the DSA 210 course to implement real-world applications, which will deepen my knowledge of data analysis and visualization.

Motivation

1- Personal Curiosity

I am interested in space and in the future, I want to work on this area.

2- An Area that Needs Attention

Space Debris has become a crucial problem that is difficult to solve. Also, when it is compared with other areas related to space, it seems that there are low number of research.

3- The Long Term Impact

Finding will prove not only the relationship between Space Debris and global leadership but also pushing everyone to think on the solution.

4- Scientific Approach

Data which are going to be taken from reliable sources will help to evaluate the project reliably and scientifically.

Dataset

The dataset of this project consists of several records.

- Space Debris Share
- GDP and Economic Growth
- Military Spending
- Trade Statistics
- Educational Influence (Human Development Index)

Tools and Technologies

Following tools and technologies will be used during the project.

- Python
- Pandas
- NumPy

Analysis Plan

1. Data Collection

Import several records into a Pandas DataFrame and preprocess the data by handling missing values.

2. Visualization

Use scatter plots, heatmaps, and time series plots to explore relationships between variables.

3. Testing Hypothesis

The hypothesis testing will be structured as follows:

H0: The global leadership race does not contribute to Space Debris.

H1: The global leadership race contributes to Space Debris.

Conclusion

At the end of the project, I expect to answer the following questions:

- Which countries are in the Global Leadership Race?
- Which countries contributes to the Space Debris?
- Is there any relationship between the Space Debris and Global Leadership Race?

Data Collection

In order to analyze the global leadership race and its relationship with space debris, we collected data from multiple reliable sources. Each dataset was processed using Python and Pandas to ensure consistency and accuracy. Below is a summary of how each dataset was handled:

GDP and Economic Growth

GDP data was imported from an Excel file and reshaped using the melt function to create a long-format table with country and year values. Column names were standardized, and only the selected year range was retained. This dataset reflects the economic size of each country.

Trade Volume

Trade statistics were collected from a structured Excel file. After selecting valid entries and renaming columns, the data was reshaped into long format. Year filtering was applied for consistency with other datasets.

Human Development Index

HDI values were loaded from an Excel file and cleaned by renaming columns and removing formatting inconsistencies. Since the dataset did not include yearly values, the available index scores were used directly to indicate development levels.

Military Data

Military personnel data was extracted from an Excel file. Column names were cleaned, and the dataset was filtered to include only the years of interest. This data provides a quantitative measure of military capacity.

Space Debris

Space debris data was collected from a CSV file containing debris counts by country. Non-country entities such as the European Space Agency were excluded. A log transformation was applied to improve visual readability due to large value disparities.

Handling Missing Values

Missing values were handled during the preprocessing stage to maintain consistency across datasets. In most cases, rows with missing country or year information were removed. Debris count were filled with zero only when the absence of data indicated no recorded activity. This approach minimized bias while allowing for complete comparisons across countries.

Selection of Top 20 Countries

To explore the relationship between space debris and the global leadership race, the analysis focused on countries that are most likely to compete in this race. These countries were identified based on their rankings in four key indicators: GDP, trade volume, military size, and HDI. The top 20 performers in each category were selected to represent the group of potential global leaders.

Data Normalization

To improve comparability across variables with different scales, normalization techniques were applied where necessary. In particular, a $\log10(1 + x)$ transformation was used for the space debris dataset to reduce skewness caused by extreme values. This allowed for clearer visualizations and more balanced interpretations without distorting the underlying data.

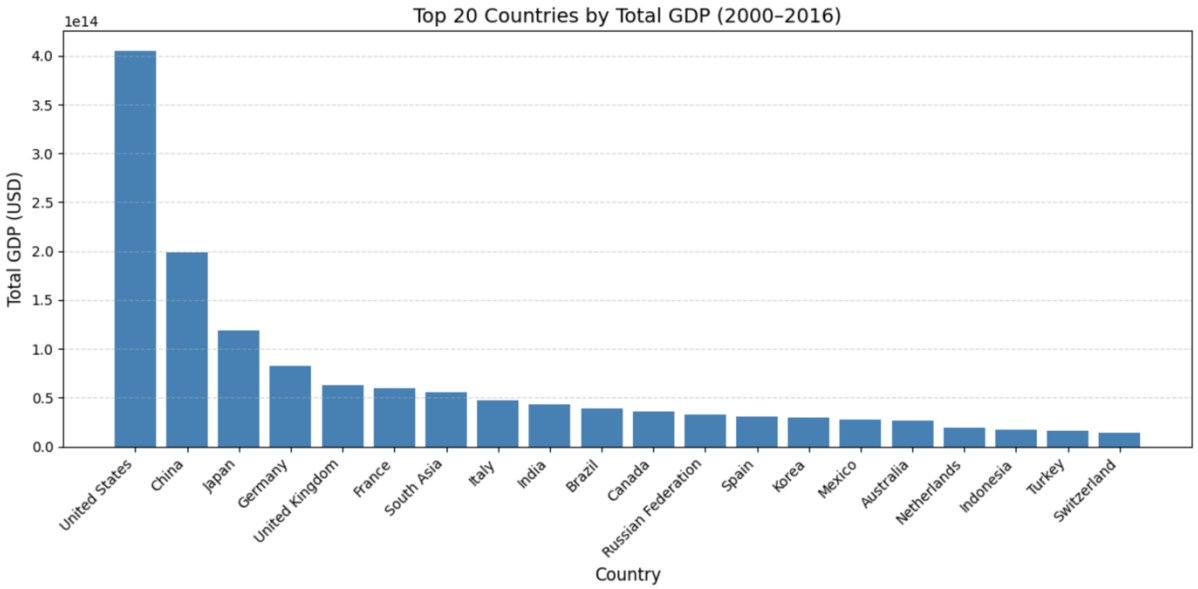
Normalization in Correlation Analysis

Although no normalization was applied during the individual visualizations of GDP, trade volume, and military size, normalization was necessary when constructing the correlation matrix. This ensured that variables with large numeric ranges did not disproportionately influence the results. Standard scaling or appropriate transformations were used to place all variables on a comparable scale, allowing for more accurate and interpretable correlation values.

Visualization

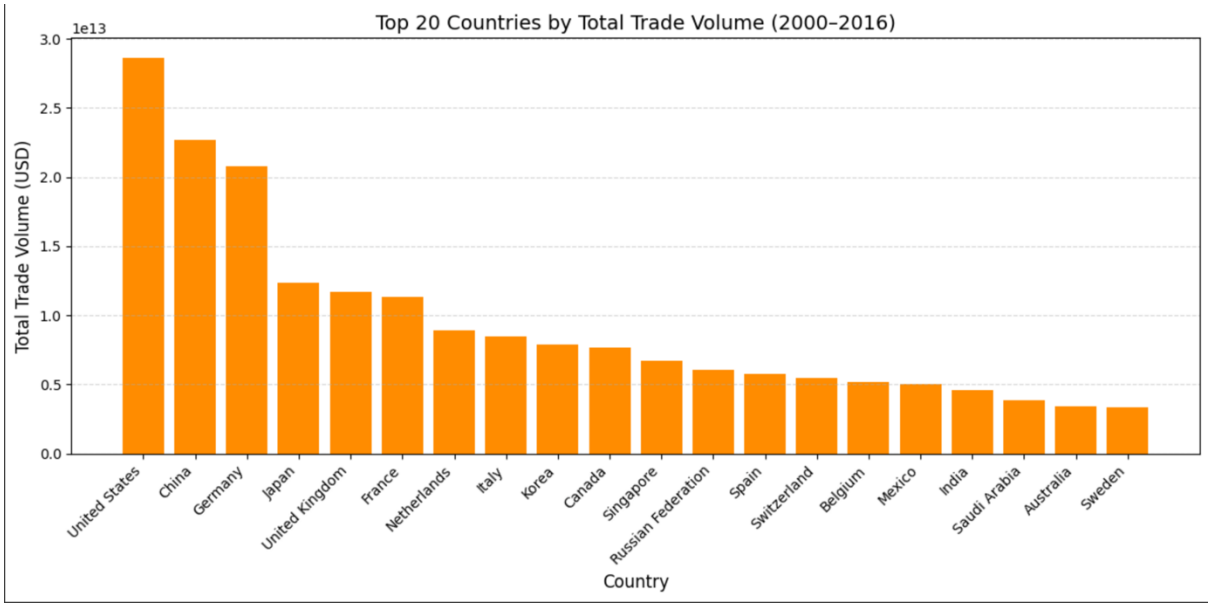
GDP and Economic Growth

The bar chart illustrates the top 20 countries by total Gross Domestic Product (GDP) accumulated between 2000 and 2016. The values were calculated by summing up the annual GDP figures for each country within the selected period. Countries classified as regions or income groups (e.g., "High income", "Sub-Saharan Africa") were excluded to ensure that the analysis includes only individual sovereign states. The GDP values are presented on a scientific scale (sci notation) to enhance readability due to the large numeric range.



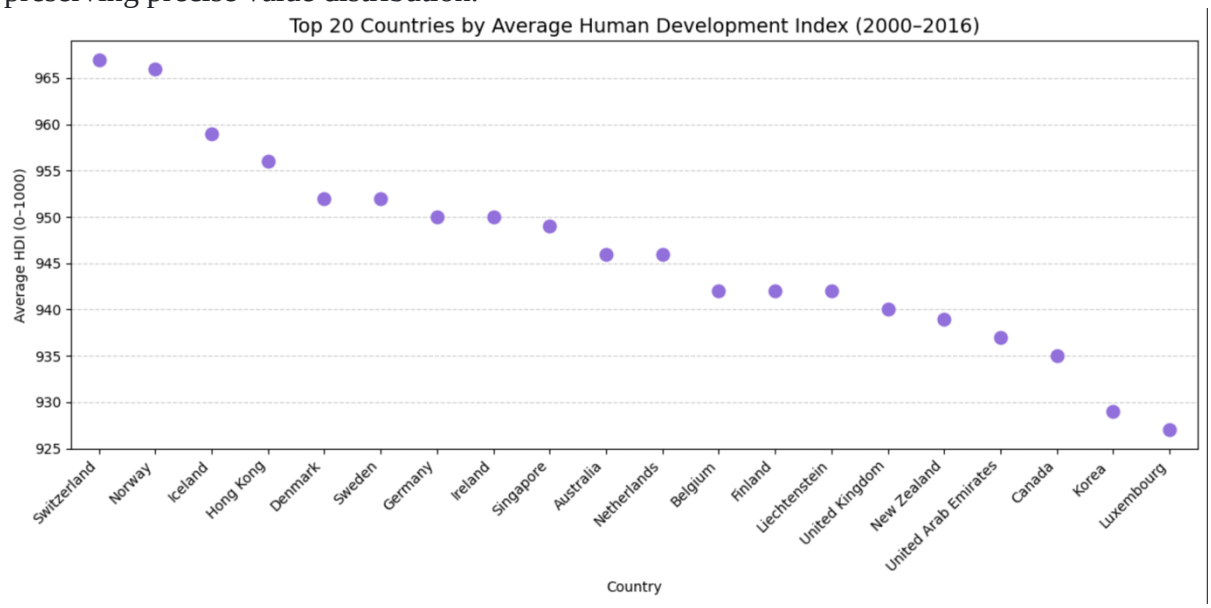
Trade Volume

The bar chart displays the top 20 countries by total trade volume between 2000 and 2016. The data was preprocessed by removing region-based or non-country entities to ensure that only sovereign countries were included. Total trade values were obtained by summing yearly trade data for each country. The chart uses scientific notation on the y-axis to accommodate the wide range of trade values, improving clarity and comparability across countries.



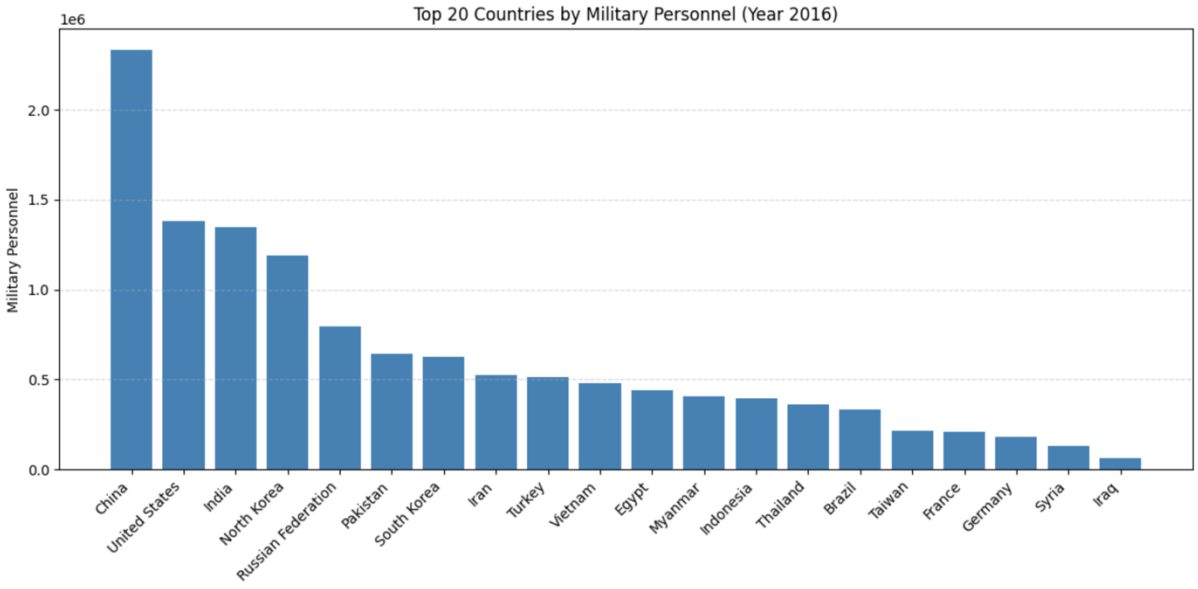
Human Development Index

The strip plot presents the top 20 countries with the highest average Human Development Index (HDI) scores between 2000 and 2016. Countries were ranked based on their overall HDI values, and only the top performers were included. The plot uses dot markers to visualize country-level differences in human development on a scale from 0 to 1000. This visual allows for quick comparison of development levels among leading nations while preserving precise value distribution.



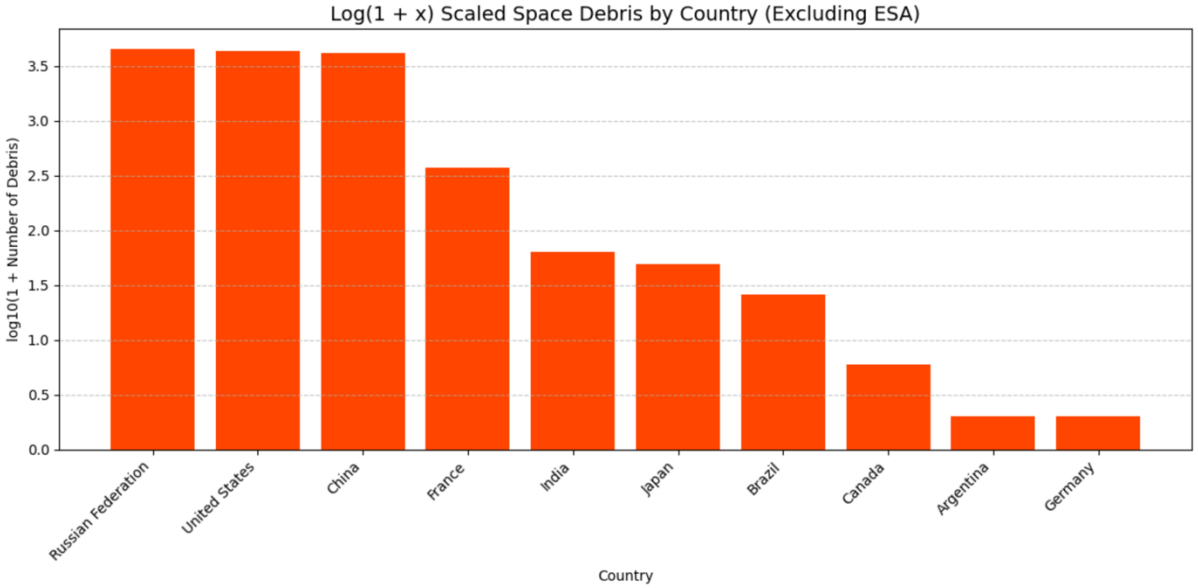
Military Data

The bar chart visualizes the top 20 countries by total active military personnel for the most recent year available in the dataset (2016). The data was filtered to include only the latest year and countries with the highest troop numbers. Countries were then sorted in descending order based on personnel count. This representation highlights the distribution of military power across nations and is useful for comparing defense capacity in the context of global influence.



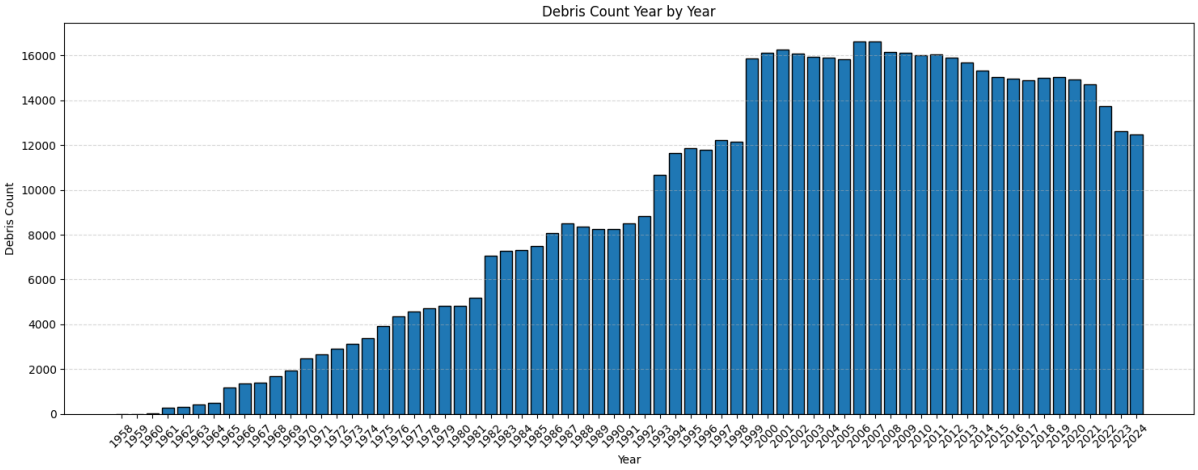
Space Debris

The bar chart illustrates the number of space debris contributions per country, scaled using the log10(1 + x) transformation to improve visibility of lower values. The dataset was filtered to exclude the European Space Agency, as it is not a sovereign state. Countries were sorted by total recorded debris, and the y-axis represents the transformed values to allow clearer comparison across a wide numerical range.

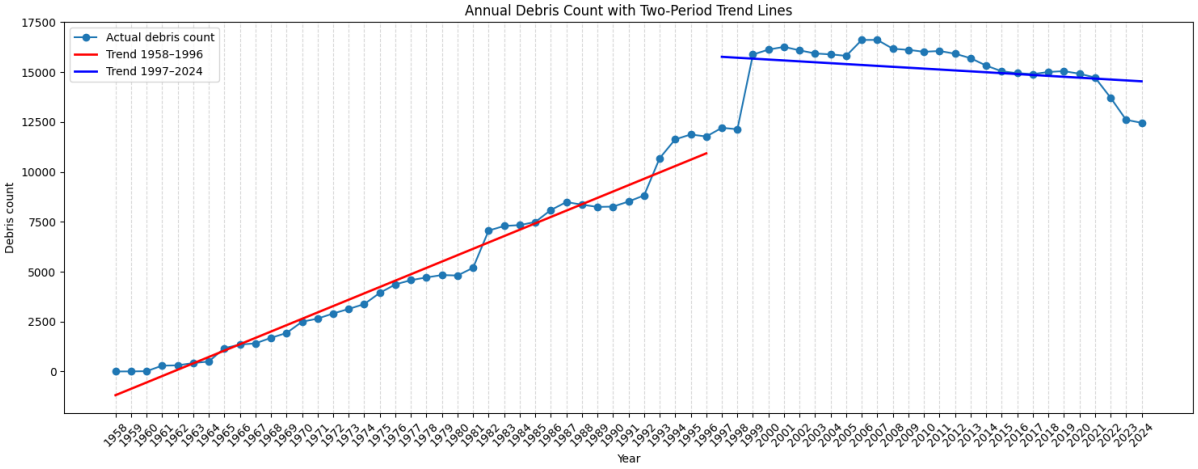


Total Number of Space Debris

This bar chart displays the annual space debris count from 1958 to 2024. The x-axis shows each year and the y-axis shows the number of debris objects tracked. Each bar’s height represents the total debris count for that specific year.

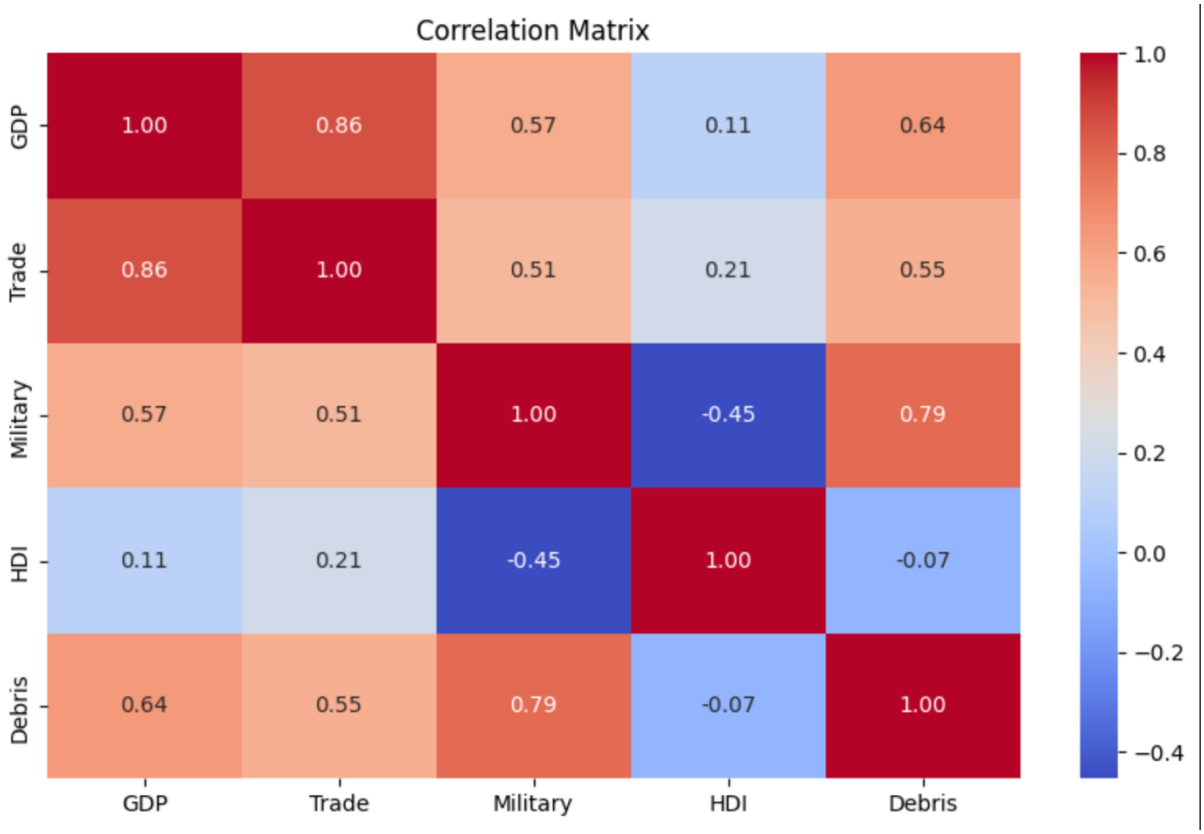


The chart below displays the annual number of debris objects tracked in low Earth orbit from 1958 through 2024. Two linear trend lines are overlaid: the red line fits the period 1958–1996 and the blue line covers 1997–2024. This dual-phase approach highlights how debris accumulation accelerated rapidly in the first period and then continued to grow at a slower rate afterward. By comparing these trends, we can clearly see a change in growth dynamics around 1996.

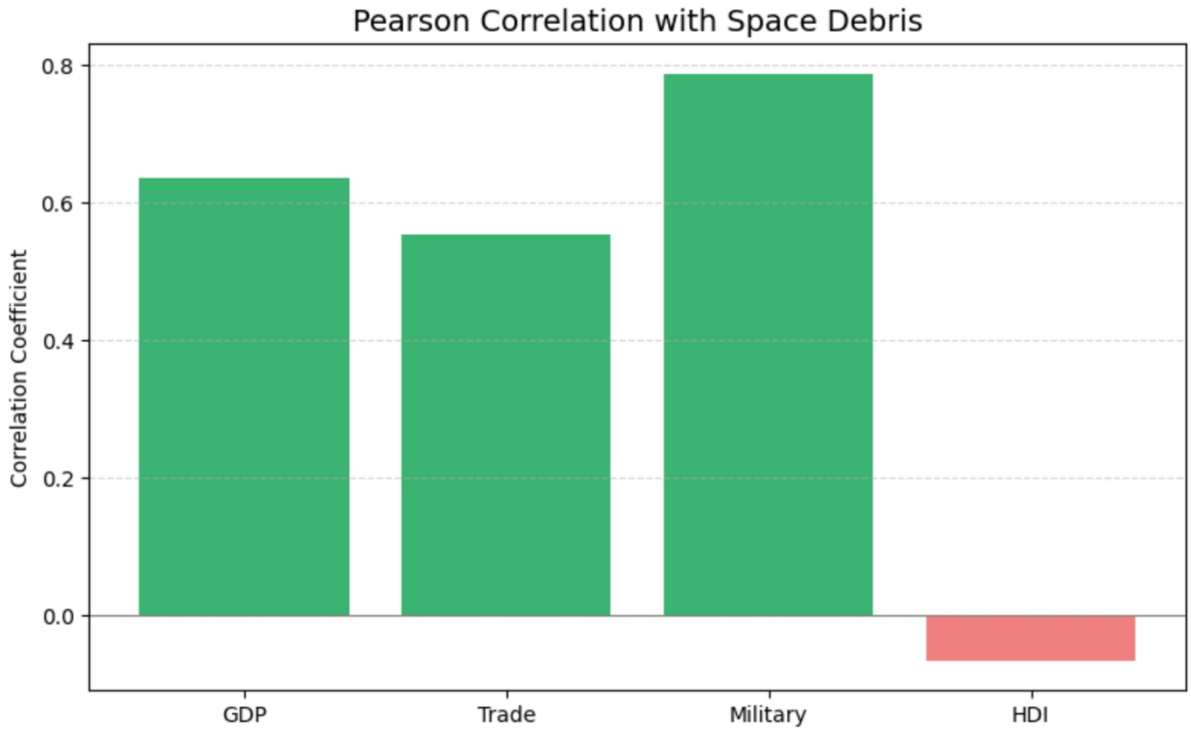


Correlation Matrix

The heatmap visualizes the Pearson correlation coefficients between key variables such as GDP, trade volume, military size, HDI, and space debris. Only numeric columns were included in the correlation matrix to ensure accurate analysis. High positive values (e.g., GDP–Trade: 0.86) suggest strong associations, while negative values (e.g., Military–HDI: –0.45) indicate inverse relationships. This matrix helps identify patterns and dependencies between indicators relevant to global leadership and space activity.



Hypothesis Testing



Analysis of GDP vs Space Debris

Null Hypothesis (H_0): There is no correlation between a country's GDP and the amount of space debris it contributes.
Alternative Hypothesis (H_1): There is a significant correlation between GDP and space debris contribution. A two-sided Spearman correlation test was used to evaluate the hypothesis at a 0.05 significance level.

Results:
Pearson Correlation: 0.6356
P-Value: < 0.01
Reject the null hypothesis. There is a statistically significant moderate positive correlation between GDP and space debris.

Analysis of Trade Volume vs Space Debris

Null Hypothesis (H_0): Trade volume is not associated with the amount of space debris.
Alternative Hypothesis (H_1): Countries with higher trade volume tend to produce more space debris. Spearman correlation analysis was conducted.

Results:
Pearson Correlation: 0.5531
P-Value: < 0.01
Reject the null hypothesis. Trade volume shows a statistically significant positive correlation with space debris.

Analysis of Military Size vs Space Debris

Null Hypothesis (H_0): There is no link between military personnel size and space debris.
Alternative Hypothesis (H_1): Countries with larger military personnel contribute more to space debris. Tested with Spearman correlation.

Results:
Pearson Correlation: 0.7882
P-Value: < 0.01
Reject the null hypothesis. The test indicates a strong and statistically significant positive correlation between military size and space debris.

Analysis of HDI vs Space Debris

Null Hypothesis (H_0): Human Development Index is unrelated to space debris contribution.
Alternative Hypothesis (H_1): There is a significant association between HDI and space debris. Spearman correlation was applied.

Results:
Pearson Correlation: -0.0664
P-Value: 0.54

Fail to reject the null hypothesis. There is no statistically significant relationship between HDI and space debris.

Hypothesis Test Results: Correlation with Space Debris

Indicator	Pearson Correlation	Pearson p-value	Spearman Correlation	Spearman p-value	Sample Size
GDP	0.6356	2.10e-59	0.6434	2.83e-61	513
Trade	0.5531	1.95e-42	0.4838	1.84e-31	513
Military	0.7882	8.32e-110	0.5901	1.87e-49	513
HDI	-0.0664	1.33e-01	-0.3451	8.65e-16	513

Conclusion Summary

Among the four tested indicators, GDP, Trade, and especially Military size showed significant positive correlations with space debris, while HDI did not. These findings suggest that aspects of global leadership competition—particularly economic and military power—are likely contributors to increasing orbital debris.

Machine Learning Part

As part of this project, all supervised machine learning models introduced in the lecture slides were applied. These included:

k-Nearest Neighbors (kNN)
Linear Regression
Support Vector Regression (SVR)
Random Forest Regressor
Neural Network (MLPRegressor)

Each model was trained on post-1999 space debris data and evaluated using the R^2 score (coefficient of determination), which measures how well the model explains the variance in the target variable.

The table below presents the R^2 values and yearly predictions (2025–2034) for each model:

Model	R^2 Score	Notes
kNN (k=3)	0.9766	High R^2 but predicted the same debris value for every future year.
Linear Regression	0.6599	Low R^2, failed to capture overall trend.
SVR	-0.1819	Negative R^2, model performed worse than a horizontal line.
Random Forest	0.9936	High R^2 but returned a **single constant** value across all years.
Neural Network	-0.0920	Failed to converge, generated increasing but unrealistic predictions.

Although kNN and Random Forest had the highest R^2 scores, they were not used in the final model. The reason is that both predicted static (non-changing) debris values for 2025–2034, which is neither realistic nor useful for the research question.

On the other hand, Linear Regression, SVR, and MLP showed low or even negative R^2 scores, indicating that they were unable to effectively model the decreasing or changing trend in the data.

Final Decision

Due to these limitations, alternative modeling approaches were considered, including Polynomial Regression and Piecewise Linear Regression, both of which provided realistic and statistically robust forecasts.

These models not only offered interpretable trends but also enabled projections of country-level debris shares using proportional allocation.

After identifying the limitations of standard machine learning models, two alternative regression approaches were evaluated:

- Piecewise Linear Regression with a breakpoint at 1999
- Polynomial Regression with degree = 3

To compare these models, their R^2 scores were calculated based on the full historical dataset:

- **R^2 (Polynomial degree = 3): 0.7857**
- **R^2 (Piecewise at 1999): 0.7007**

Based on the R^2 results, Polynomial Regression (degree = 3) was selected as the final model. It offered:

- A strong balance between flexibility and interpretability
- Realistic forecasts for 2025–2034
- Compatibility with proportional country-level debris distribution

This model served as the foundation for all further analyses and visualizations in the project.

After predicting the total debris count for 2025–2034 using the selected Polynomial Regression model, the next goal was to estimate how much of that debris would be attributed to each country.

For this purpose, the SpaceDebrisStats.csv dataset was used, which contains each country's total space debris count. Each country's proportional share of total debris was calculated, assuming that this ratio remains constant over the next 10 years.

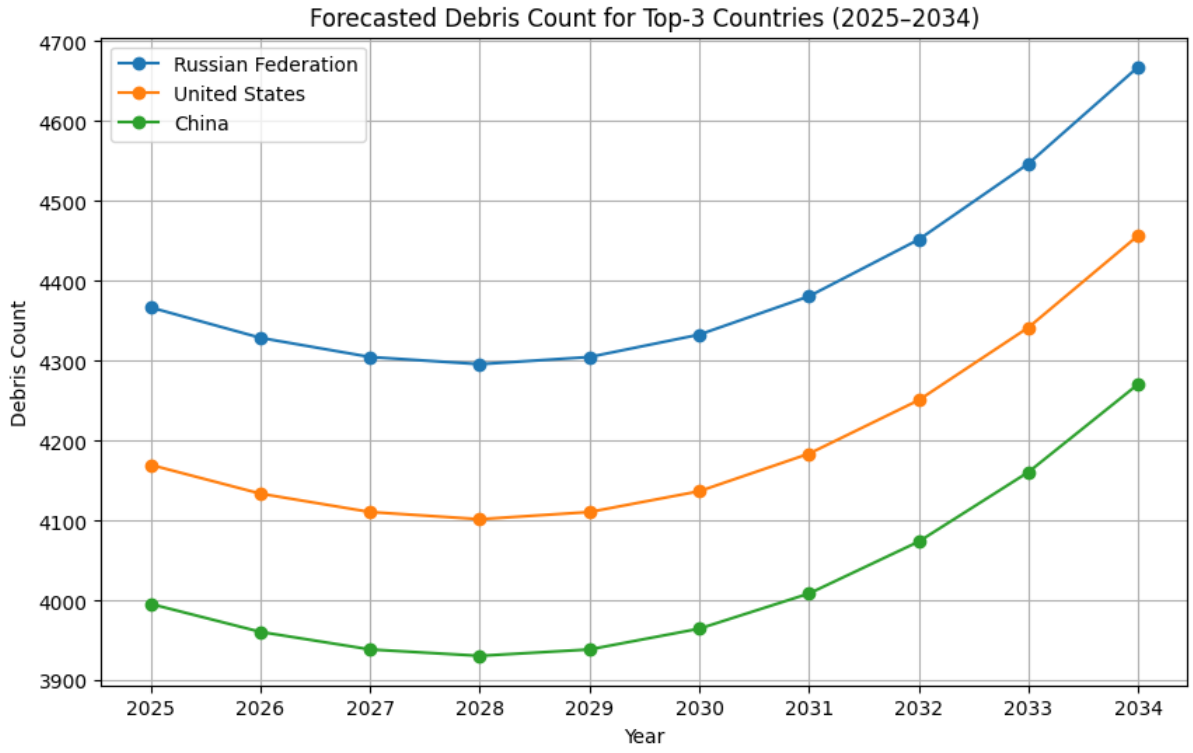
Then, for each year from 2025 to 2034:

- The forecasted global total was multiplied by each country's proportion
- Year-by-year debris predictions per country were generated
- A combined long-format table of yearly predictions for all countries was created

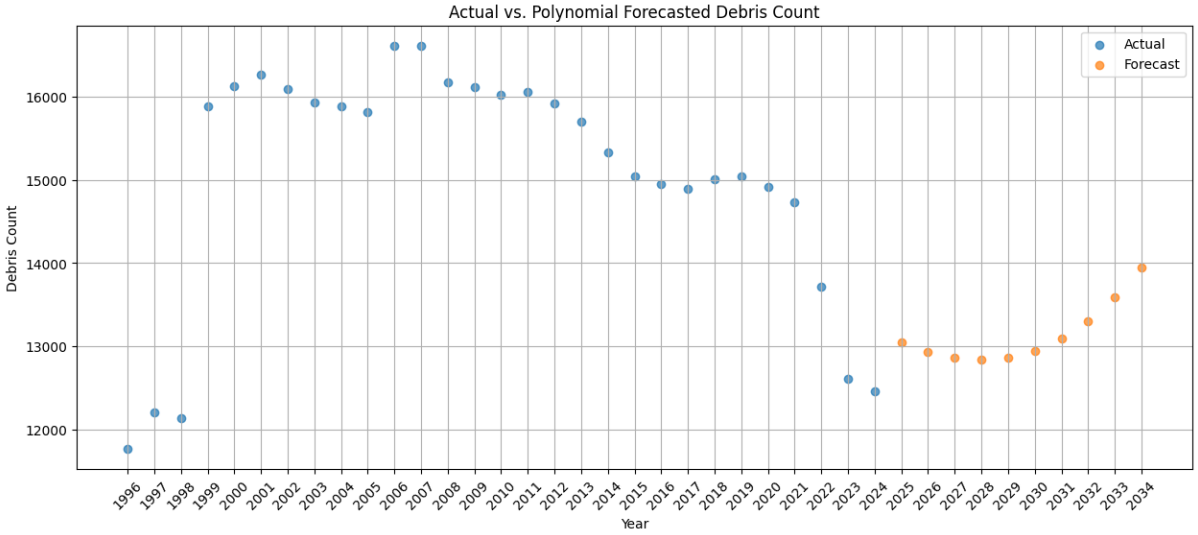
Finally, countries were ranked by total forecasted debris between 2025 and 2034, and the top 3 contributors were identified. The chart below visualizes the projected debris count for:

- Russian Federation
- United States
- China

These countries are expected to maintain the highest share of debris in orbit if current trends continue.



This visualization clearly illustrates how space debris is projected to evolve among the leading spacefaring nations.



This chart presents the actual number of space debris from 1958 to 2024 (blue) and the predicted values for 2025–2034 (orange), using a degree-3 Polynomial Regression model.

A decrease in debris is observed in recent years, but the model forecasts a new increase beginning after 2027, assuming current trends persist.

This visualization illustrates both the historical and projected global space debris trends.

Limitations and Future Work

Limitations

- The predictions are solely based on historical debris counts and a constant country share assumption, which does not account for potential future political, technological, or environmental shifts.
- The models exclude unexpected space-related events such as satellite explosions or major launch activities.
- Country-level debris estimates rely on fixed proportions derived from past data, overlooking possible changes in national space policies or program developments.
- The models presume clean and uninterrupted yearly trends. However, real-world data may contain irregularities, missing years, or anomalies that influence debris dynamics.

Future Work

- Incorporate additional features into the model, such as the number of launches, satellite failures, or international space activity metrics.
- Dynamically update country-level proportions based on forecasted indicators like military budgets, trade activity, or GDP, rather than relying on fixed historical shares.

- Explore the integration of collaborative datasets or real-time satellite tracking APIs to enhance prediction accuracy and enable continuous monitoring.