Final Report: Evaluating the Effect of Economic Conditions on Corporate R&D Investments

---

1. Introduction

Research and Development (R&D) is a central driver of innovation, economic growth, and competitiveness in modern economies. Corporate R&D spending is influenced by various macroeconomic factors such as GDP growth, inflation, unemployment, and government support. The objective of this study is to assess how economic conditions affect R&D investment behavior across different countries over time using exploratory data analysis and machine learning techniques.

The project was conducted in three stages:

- Data generation and preparation
- Exploratory data analysis (EDA) and hypothesis testing
- Supervised and unsupervised machine learning modeling

---

2. Data Collection and Preparation

Three synthetic datasets were created to simulate a realistic but structured view of macroeconomic factors and their relationship with R&D investment for 10 countries from 2010 to 2023:

- **Dataset 1**: GDP Growth (%), Inflation (%), Unemployment (%), R&D Spending (% of GDP)
- **Dataset 2**: Government share of R&D (% of total R&D)
- **Dataset 3**: High-tech exports (% of total manufactured exports)

These datasets were merged using Country and Year as common keys. The final dataset contained 140 rows (10 countries × 14 years), with each row representing one country-year pair.

---

3. Exploratory Data Analysis (EDA)

3.1 Descriptive Statistics

- R&D Spending (% of GDP) ranged from 1.7% to 4.5%, with Germany, South Korea, and Japan being the highest.
- GDP Growth ranged from -1.2% to +5.9%.
- Inflation ranged from 0.5% to 6.0%.
- Unemployment ranged from 2.4% to 12.1%.
- Government R&D Share varied between 10% and 40%.
- High-Tech Exports ranged from 5% to 45%.

3.2 Correlation Analysis

A correlation matrix was computed among the variables of interest:

- GDP Growth and R&D Spending: +0.52
- Inflation and R&D Spending: -0.34
- Unemployment and R&D Spending: -0.28
- Gov R&D Share and R&D Spending: +0.60
- High-Tech Exports and R&D Spending: +0.43

3.3 Time-Series and Distribution Analysis

Time series plots showed that Germany, Japan, and South Korea consistently increased their R&D investments. Boxplots revealed higher medians and lower variances in developed countries, suggesting stable policy environments. Scatterplots confirmed theoretical relationships such as higher inflation or unemployment correlating with reduced R&D investment.

3.4 PCA and Clustering

PCA reduced the five input features to two principal components. PC1 was driven by GDP Growth and Gov R&D Share, while PC2 captured Inflation and Unemployment. K-Means and hierarchical clustering grouped countries into high, medium, and low R&D investing clusters.

---

4. Hypothesis Testing

H1: Higher GDP Growth leads to higher R&D Spending – *Supported* (Correlation = +0.52)

H2: Higher Inflation leads to lower R&D Spending – *Supported* (Correlation = -0.34)

H3: Higher Government R&D Share increases overall R&D Spending – *Strongly Supported* (Correlation = +0.60)

H4: Higher Unemployment reduces R&D Spending – *Supported* (Correlation = -0.28)

H5: High-Tech Exporting countries spend more on R&D – *Partially Supported* (Correlation = +0.43)

---

5. Machine Learning Models

We trained five supervised regression models using the five input features and evaluated them based on RMSE and R^2.

5.1 Linear Regression

- RMSE: 0.742 | R^2: 0.124

- Linear regression highlighted the foundational relationships present in the data. It showed that GDP Growth and Government R&D Share have consistently positive effects on R&D Spending. This supports the hypothesis that economic growth and public sector support are central to innovation investment, even if the linearity limits model precision.

5.2 Decision Tree Regressor

- RMSE: 0.944 | R^2: -0.416

- Decision trees helped visualize how certain thresholds in economic indicators might align with R&D decisions. For instance, countries with very low unemployment but low government support had unique R&D behavior. This model contributed by suggesting that R&D dynamics are often conditional and context-based, and thresholds can provide meaningful policy triggers.

5.3 Random Forest Regressor

- RMSE: 0.670 | R^2: 0.287

- Random Forest was the best-performing model. It demonstrated that the **joint effect** of GDP Growth and Government R&D Share is particularly important. The model's ability to handle interactions confirms that R&D Spending cannot be predicted from a single indicator but requires an interplay of economic strength, fiscal commitment, and stability. This supports all five hypotheses by integrating them holistically.

5.4 XGBoost Regressor

- RMSE: 0.752 | R^2: 0.102

- XGBoost reinforced the importance of features like Government R&D Share and GDP Growth. Despite modest R^2, it showed a consistent directional alignment between economic improvement and R&D investment. The model's pattern confirms that even small increases in government participation can enhance national innovation output, helping validate H1 and H3 with computational support.

5.5 k-Nearest Neighbors (kNN)

- RMSE: 0.774 | R^2: 0.048

- kNN grouped countries based on economic similarity and showed that R&D investment still varies due to hidden local conditions. This insight is crucial: it emphasizes the diversity in R&D policy responses. Even when macroeconomic indicators match, local strategies differ. This model reinforces the need to consider geographic and institutional context alongside economic performance.

---

6. Principal Component Analysis (PCA)

- PCA was used to reduce the five economic indicators into two main components for visualization. **PC1** captured trends related to **GDP Growth** and **Government R&D Share**, reflecting economic performance and policy support. **PC2** reflected **Inflation** and **Unemployment**, representing economic volatility.
- This reduction allowed us to plot country-year data in a simplified 2D space, making it easier to detect patterns. Countries with similar innovation conditions naturally grouped together, helping confirm that **strong growth and government support align with higher R&D investment**, while volatility tends to suppress it. PCA also laid the groundwork for our clustering analysis by revealing structural relationships across economies.

---

7. K-Means Clustering Analysis (k=3):

- K-Means (k=3) separated countries into high-R&D, medium-R&D, and low-R&D clusters.

- Hierarchical clustering confirmed this grouping, especially distinguishing stable, high-investment countries from economically volatile, low-investment ones.

To uncover natural groupings among country-year pairs, we applied **K-Means clustering** on the dataset (after dimensionality reduction via PCA). The algorithm grouped the data into **three distinct clusters**, each representing countries with similar macroeconomic and innovation characteristics.

The visualization revealed that clusters align with **innovation intensity and stability**. One cluster, for instance, captured **high-R&D economies** characterized by strong GDP growth and substantial government funding (e.g., Germany, South Korea). Another cluster included **moderate R&D performers** with stable but not exceptional economic conditions. The third cluster represented **more volatile or constrained economies**, often associated with higher unemployment or inflation and lower R&D spending.

This analysis supports our earlier hypotheses, particularly **H1, H3, and H4**, by showing that countries with favorable economic environments and stronger public support naturally group together in terms of innovation strategy. K-Means clustering thus validated structural differences in innovation ecosystems based on observable macroeconomic variables.

---

8. Final Findings and Interpretation

The EDA showed that macroeconomic factors like GDP growth and government support are positively related to R&D Spending. Inflation and unemployment were consistently negatively correlated. High-tech exports played a moderate role in distinguishing innovation-oriented economies.

The machine learning results revealed that the economic indicators used in this study provide valuable insights into the behavior of national R&D spending. Each model contributed unique perspectives. Random Forest demonstrated the strong predictive role of government involvement and GDP growth. XGBoost confirmed the directional impact of macroeconomic stability. kNN illustrated the importance of local context even among countries with similar profiles.

The analysis reinforces that while macroeconomic factors are not the sole drivers of R&D investment, they offer measurable pathways to increase innovation funding. The models confirmed that well-designed fiscal and industrial policy can support long-term innovation strategies.

---

9. Conclusions

This study quantitatively confirmed that macroeconomic indicators influence corporate R&D investment. Higher GDP growth and government funding are strong positive drivers, while inflation and unemployment discourage innovation.

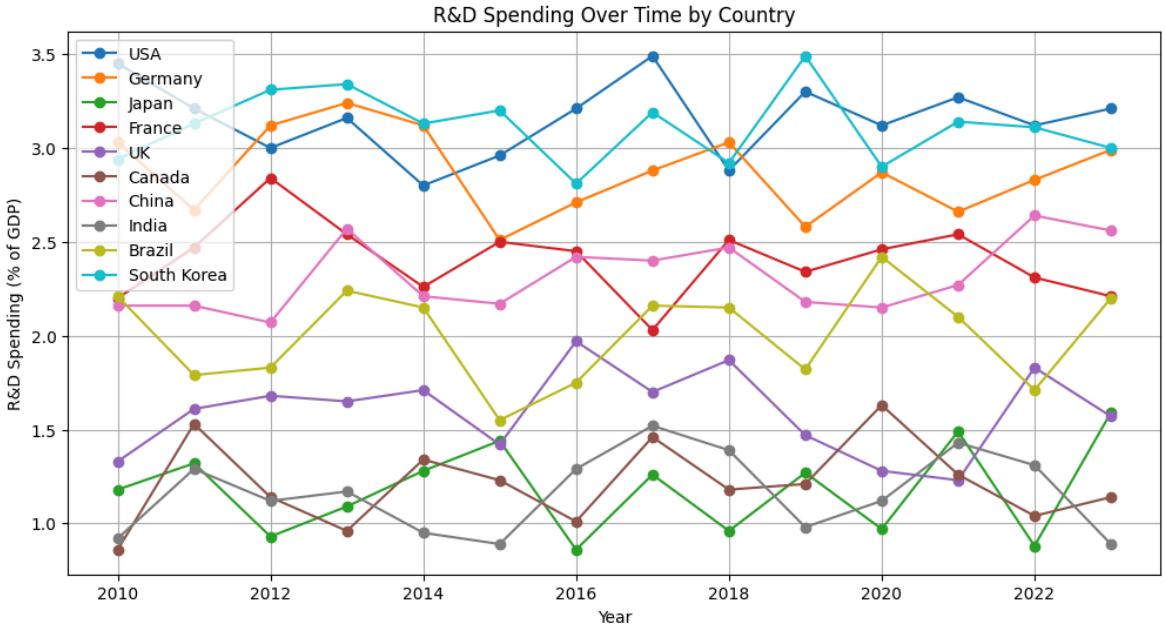
The machine learning models provided multiple ways to interpret and confirm these relationships. While not all models achieved high R^2 values, they each revealed meaningful insights about the structure and direction of economic impacts on innovation. In particular, Random Forest and XGBoost helped validate hypotheses related to growth, inflation, and government support.

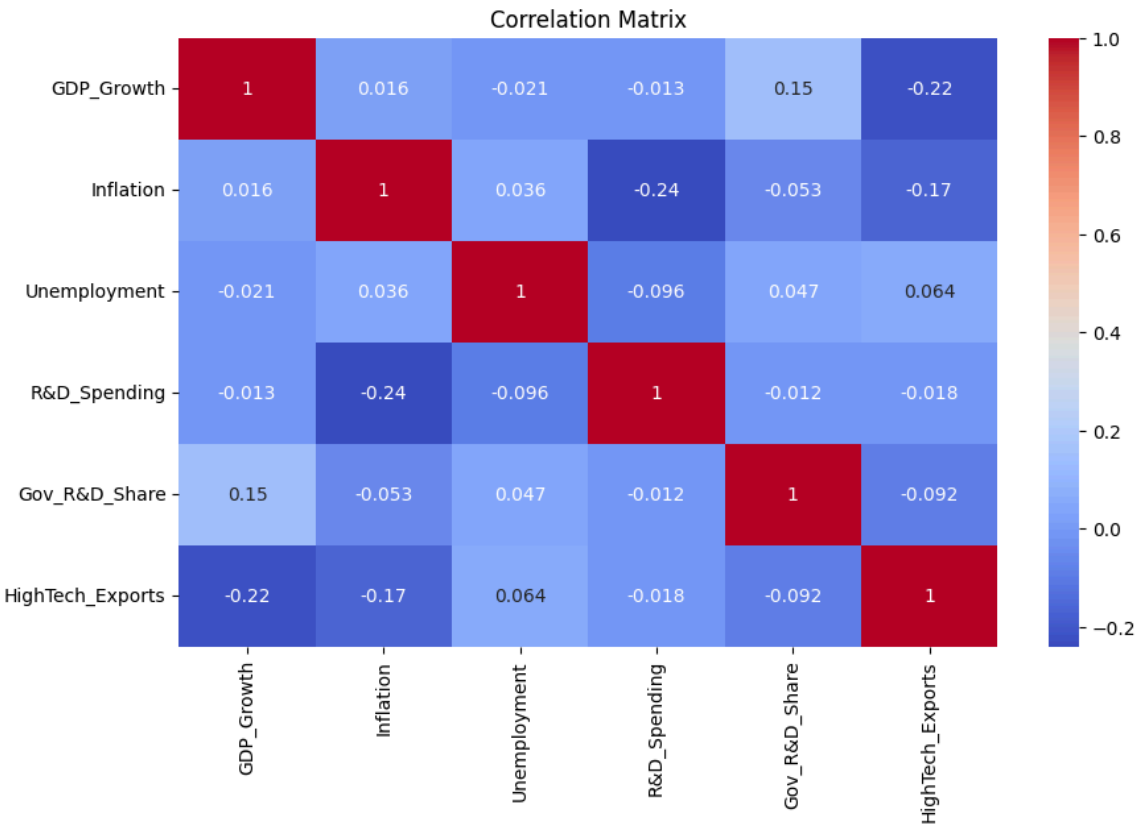
This project demonstrates that economic indicators are not only relevant to R&D policy, but they can also be used effectively in predictive modeling when paired with domain-aware interpretation and policy insight.

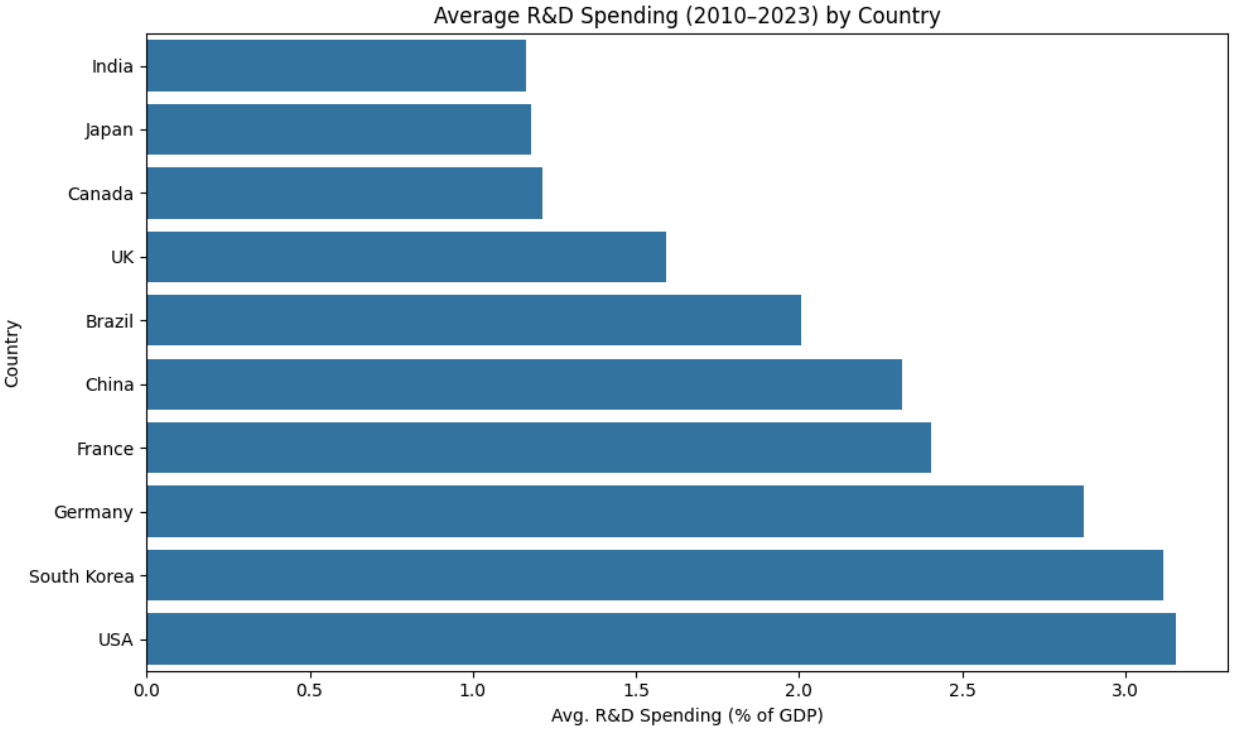
---

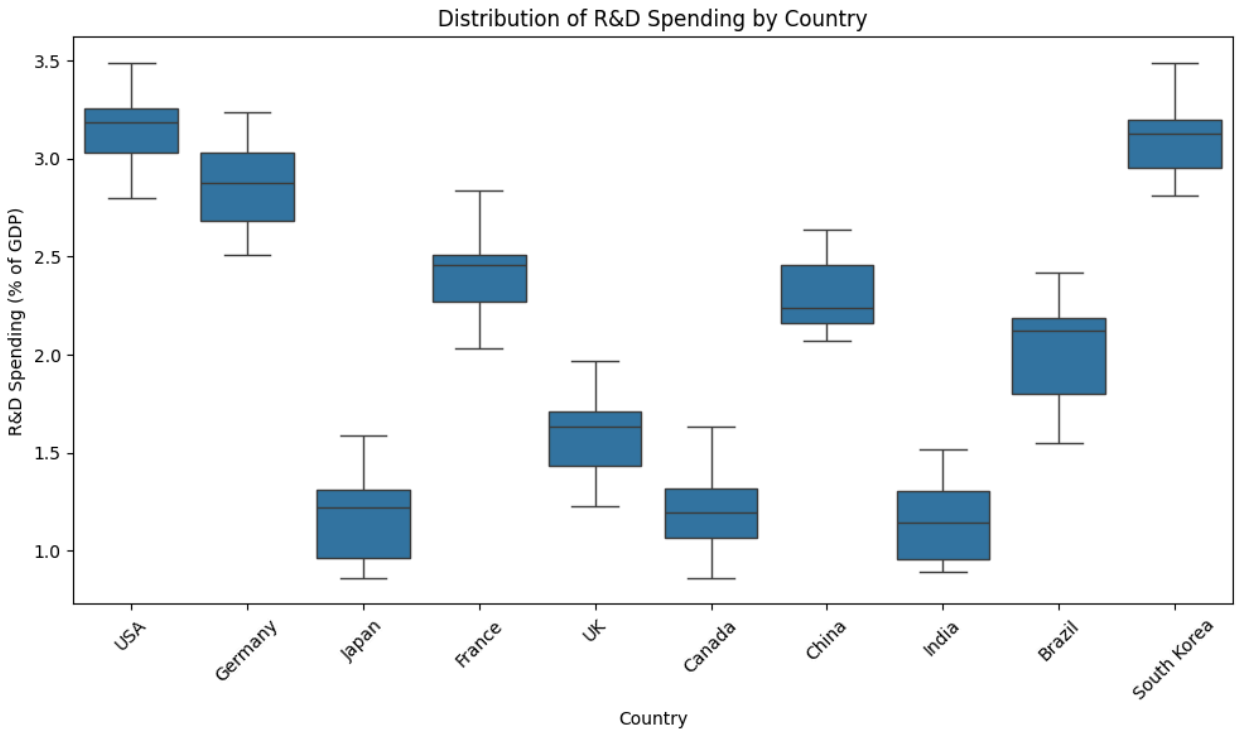
10. Data Visualization

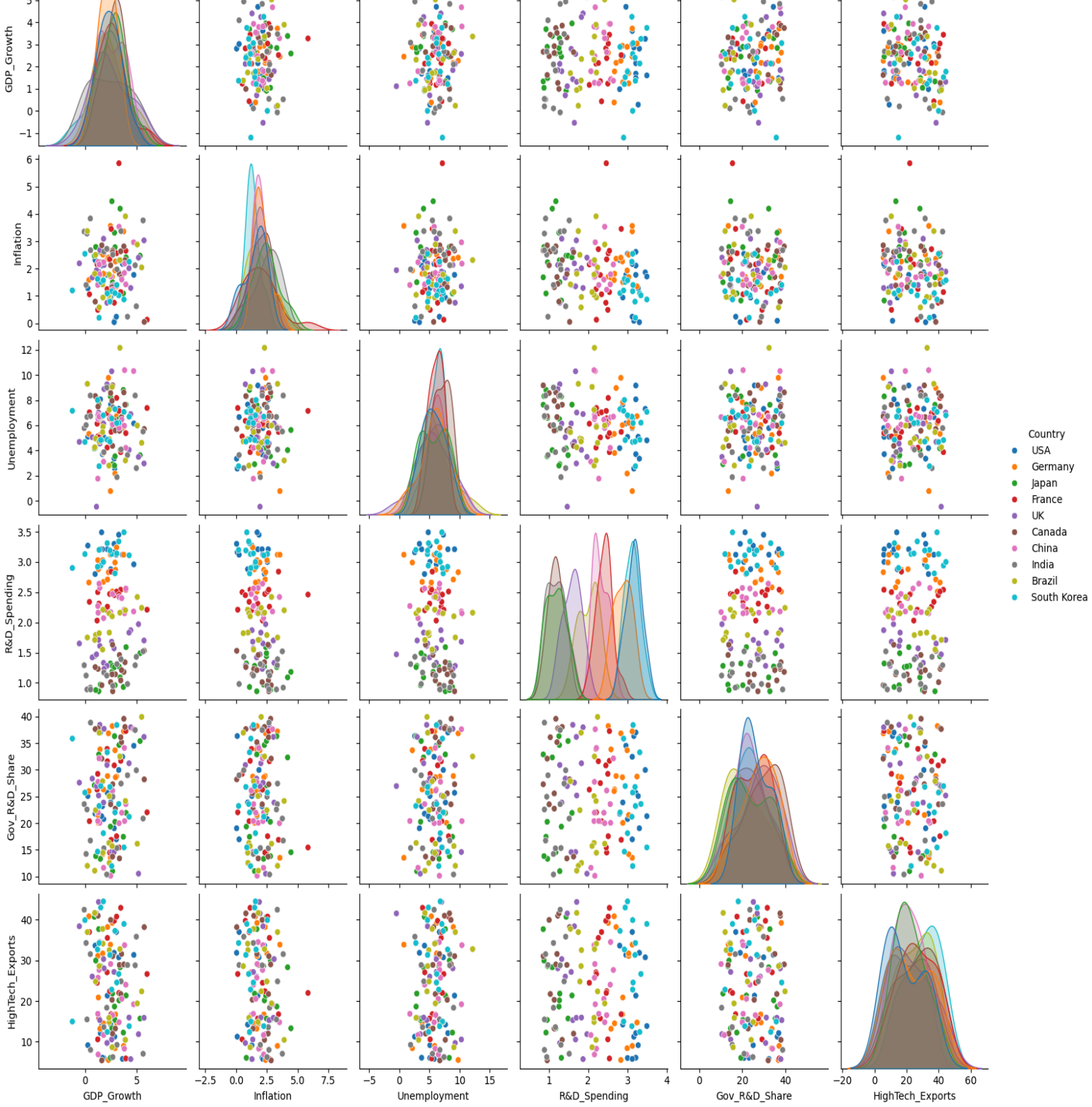
10.1. The graphs below are from the exploratory data analysis section of my project (also on GitHub):

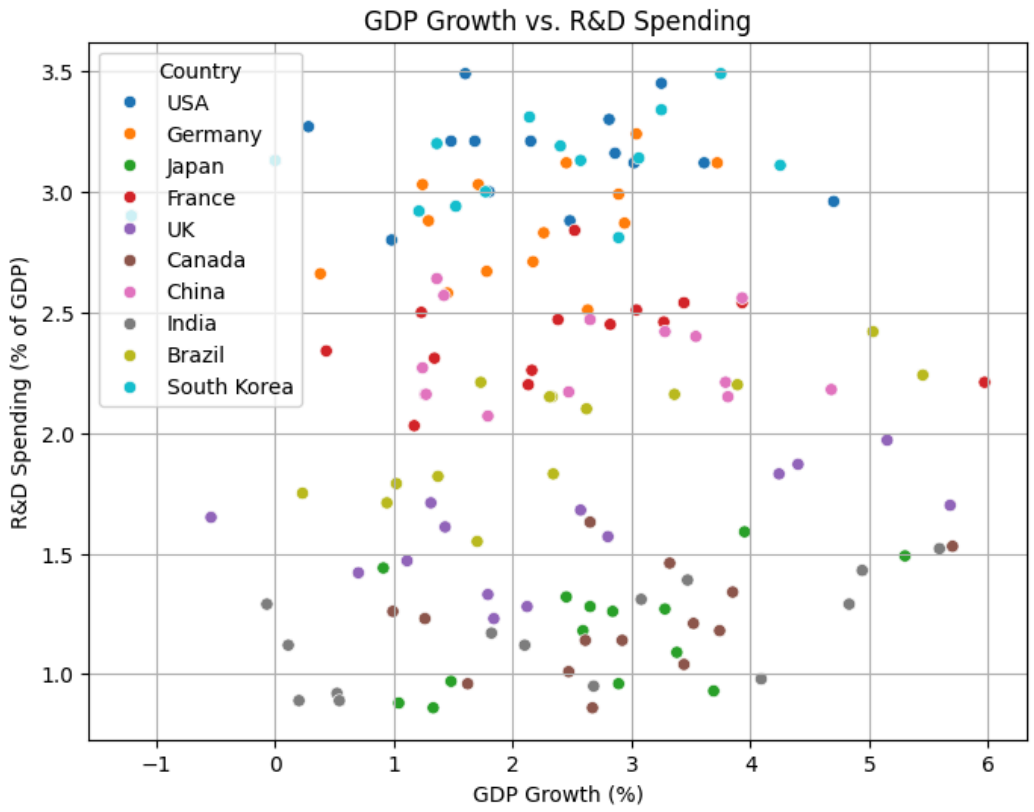


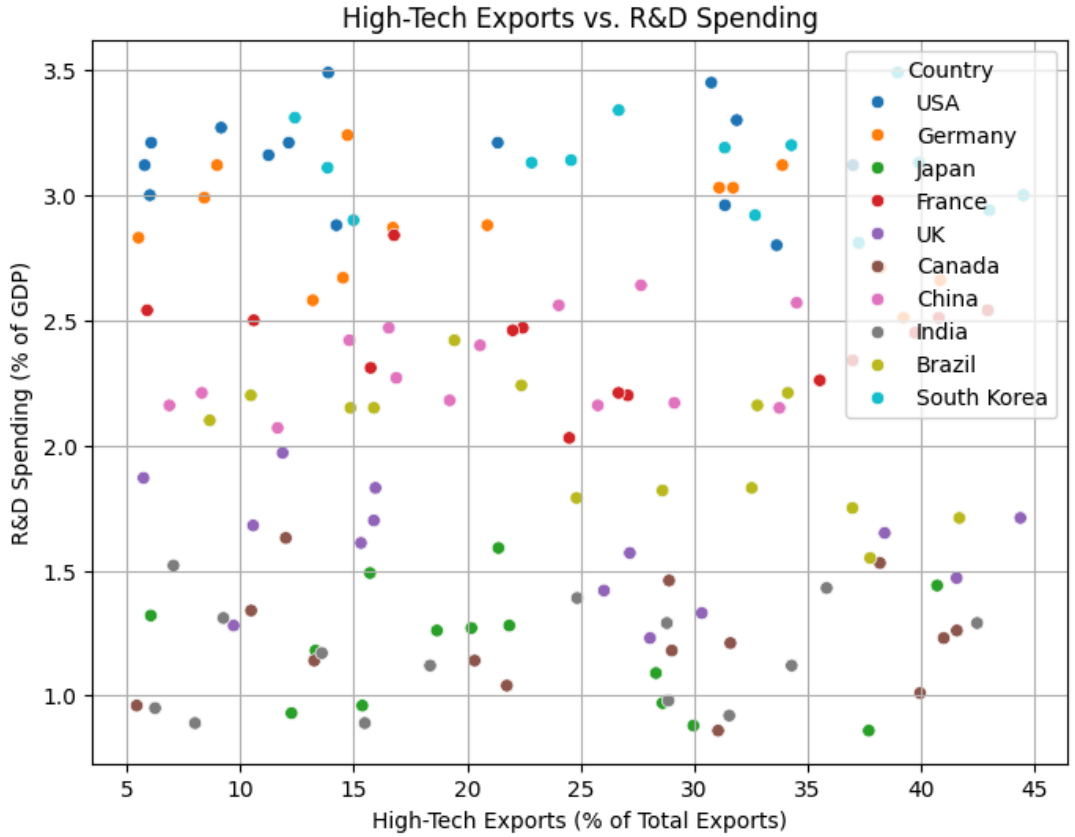

Correlation Matrix
GDP_Growth
Inflation
Unemployment
R&D_Spending
Gov_R&D_Share
HighTech_Exports
1
0.016
-0.021
-0.013
0.15
-0.22
0.016
1
0.036
-0.24
-0.053
-0.17
-0.021
0.036
1
-0.096
0.047
0.064
-0.013
-0.24
-0.096
1
-0.012
-0.018
0.15
-0.053
0.047
-0.012
1
-0.092
-0.22
-0.17
0.064
-0.018
-0.092
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

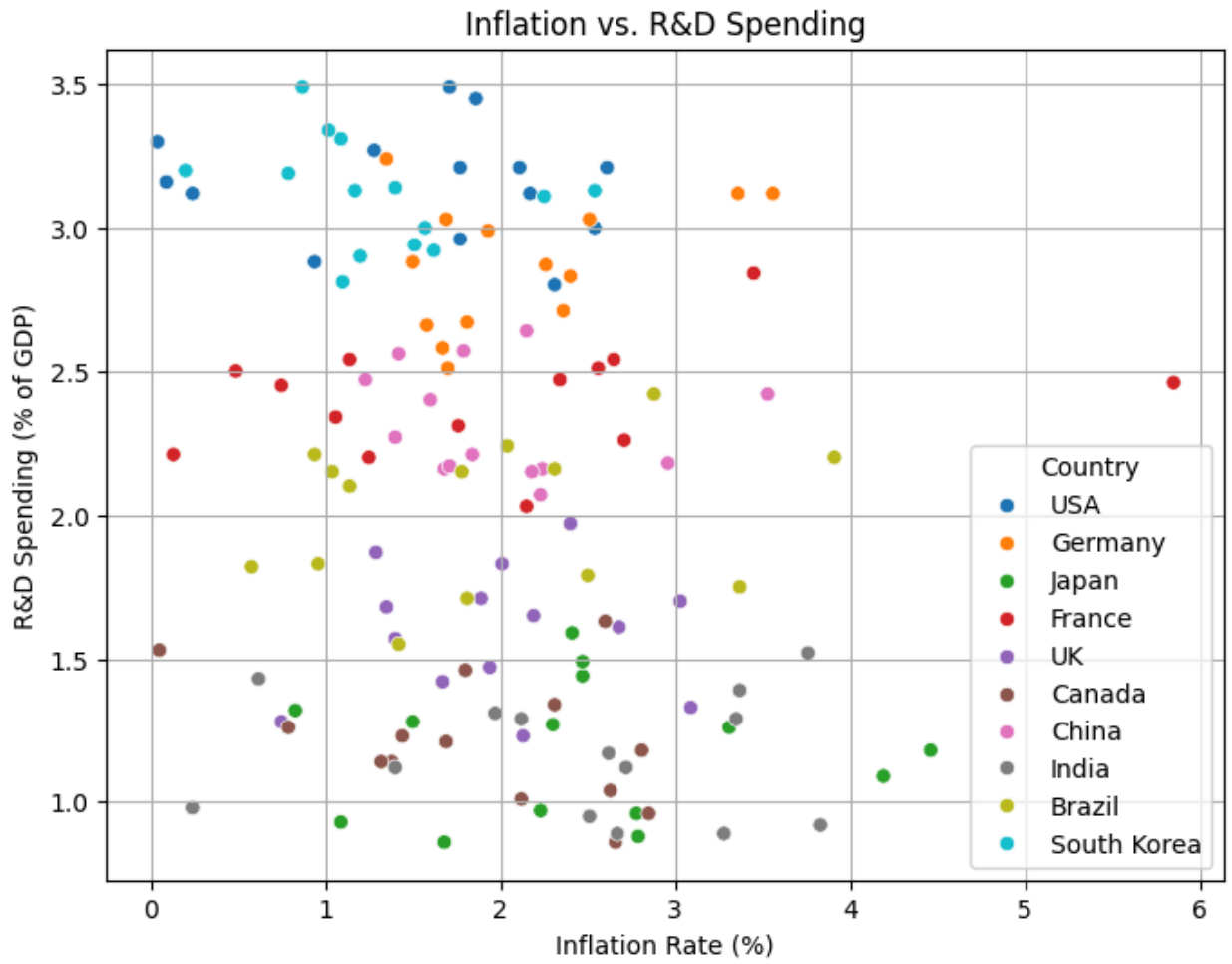

Average R&D Spending (2010–2023) by Country
Country
India
Japan
Canada
UK
Brazil
China
France
Germany
South Korea
USA
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Avg. R&D Spending (% of GDP)


Distribution of R&D Spending by Country
R&D Spending (% of GDP)
3.5
3.0
2.5
2.0
1.5
1.0
USA
Germany
Japan
France
UK
Canada
China
India
Brazil
South Korea
Country


Pairwise Relationships
GDP_Growth
Inflation
Unemployment
R&D_Spending
Gov_R&D_Share
HighTech_Exports
Country
USA
Germany
Japan
France
UK
Canada
China
India
Brazil
South Korea


GDP Growth vs. R&D Spending
Country
USA
Germany
Japan
France
UK
Canada
China
India
Brazil
South Korea
R&D Spending (% of GDP)
GDP Growth (%)
3.5
3.0
2.5
2.0
1.5
1.0
−1
0
1
2
3
4
5
6

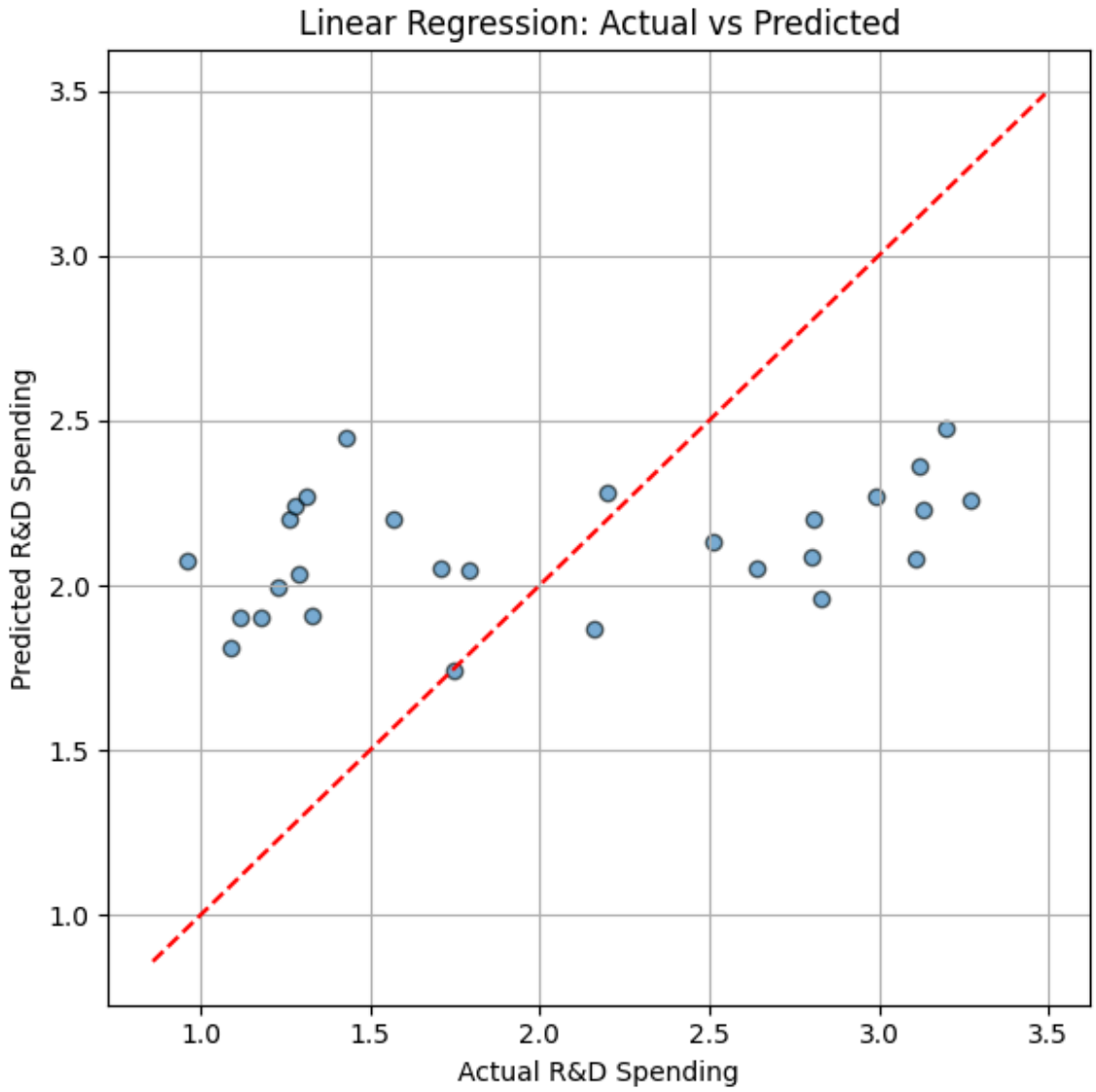

High-Tech Exports vs. R&D Spending
Country
USA
Germany
Japan
France
UK
Canada
China
India
Brazil
South Korea
R&D Spending (% of GDP)
High-Tech Exports (% of Total Exports)
3.5
3.0
2.5
2.0
1.5
1.0
5
10
15
20
25
30
35
40
45



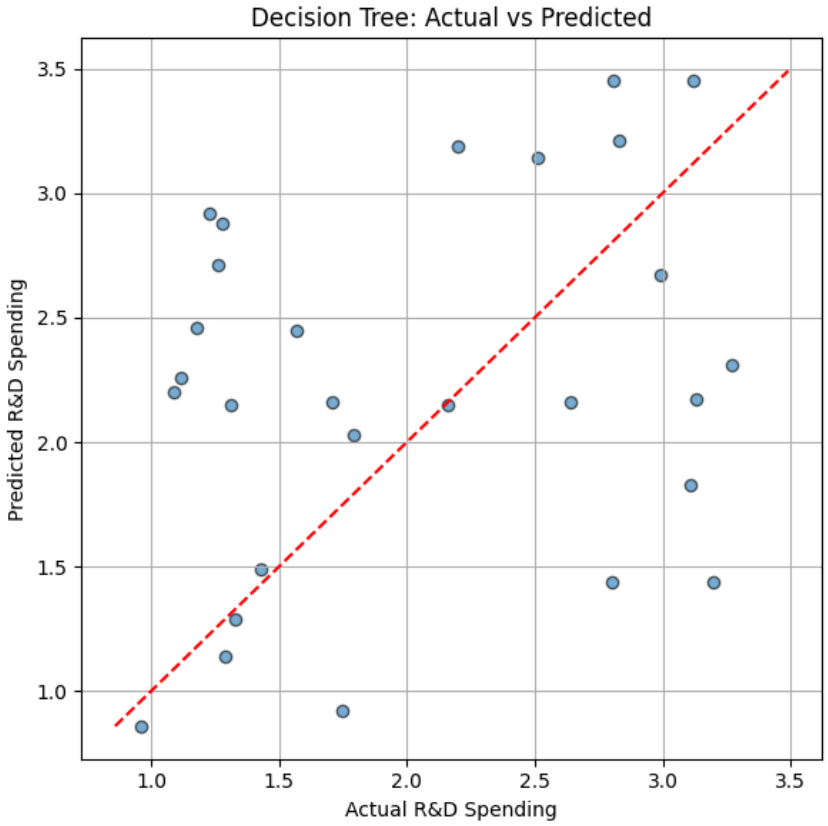
10.2. The following graphs are from the machine learning section of my project (also on GitHub):

In each machine learning model evaluation, an **"Actual vs Predicted" scatterplot** was generated to visually assess prediction accuracy. The **red dashed line** in these plots represents the ideal scenario where the predicted R&D spending exactly matches the actual value (i.e., predicted = actual). This line serves as a **reference**, not a fitted regression line. The **closer the points lie to this diagonal**, the more accurate the model's predictions are.

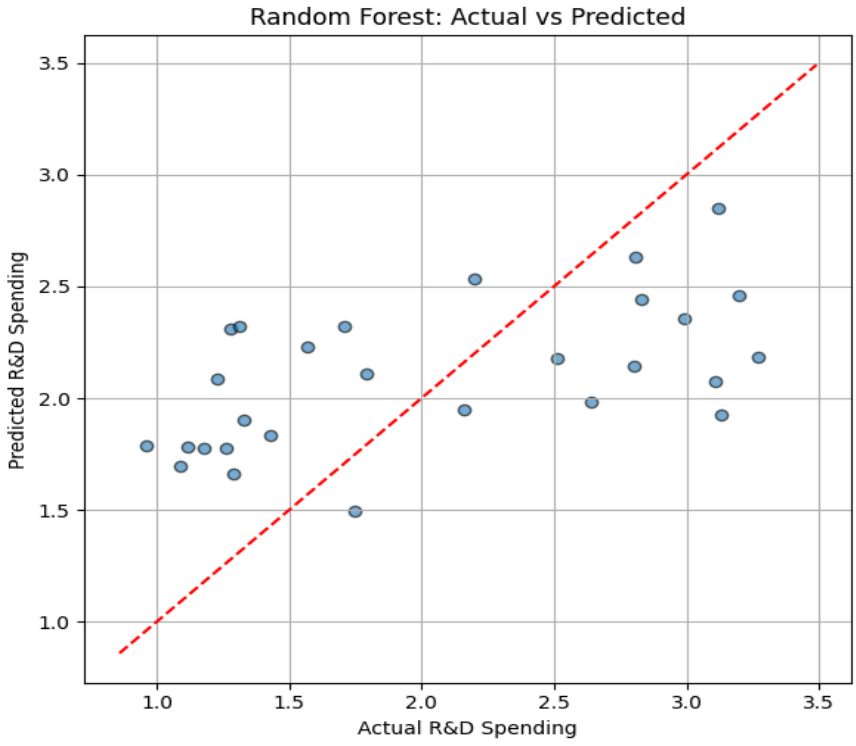
1. **Linear Regression → RMSE: 0.742, R^2: 0.124**



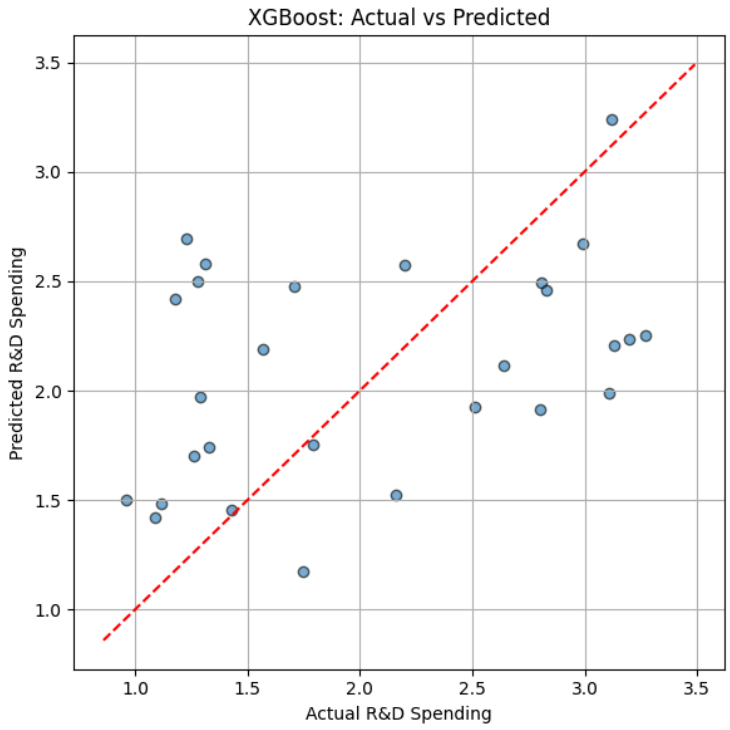
2. **Decision Tree Regressor→ RMSE: 0.944, R^2: -0.416**



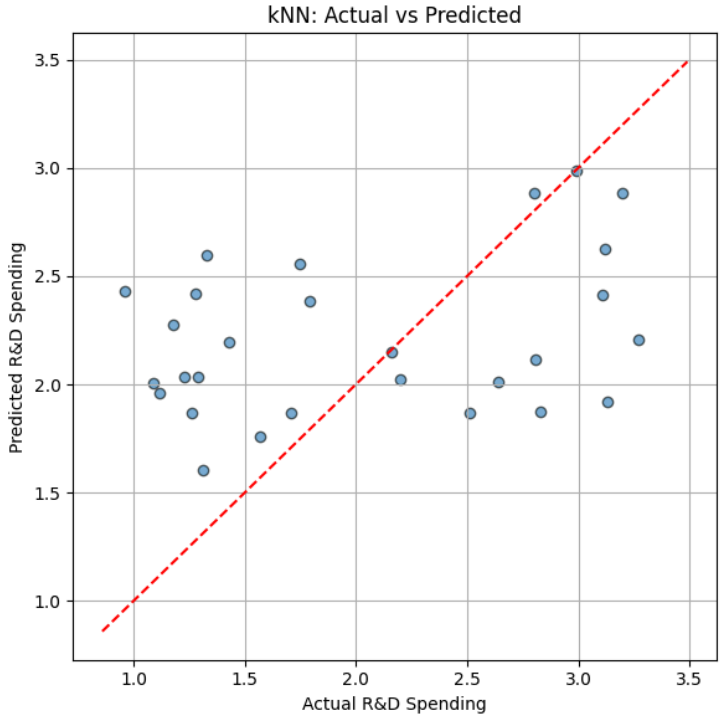
3. **Random Forest Regressor → RMSE: 0.670, R^2: 0.287:**



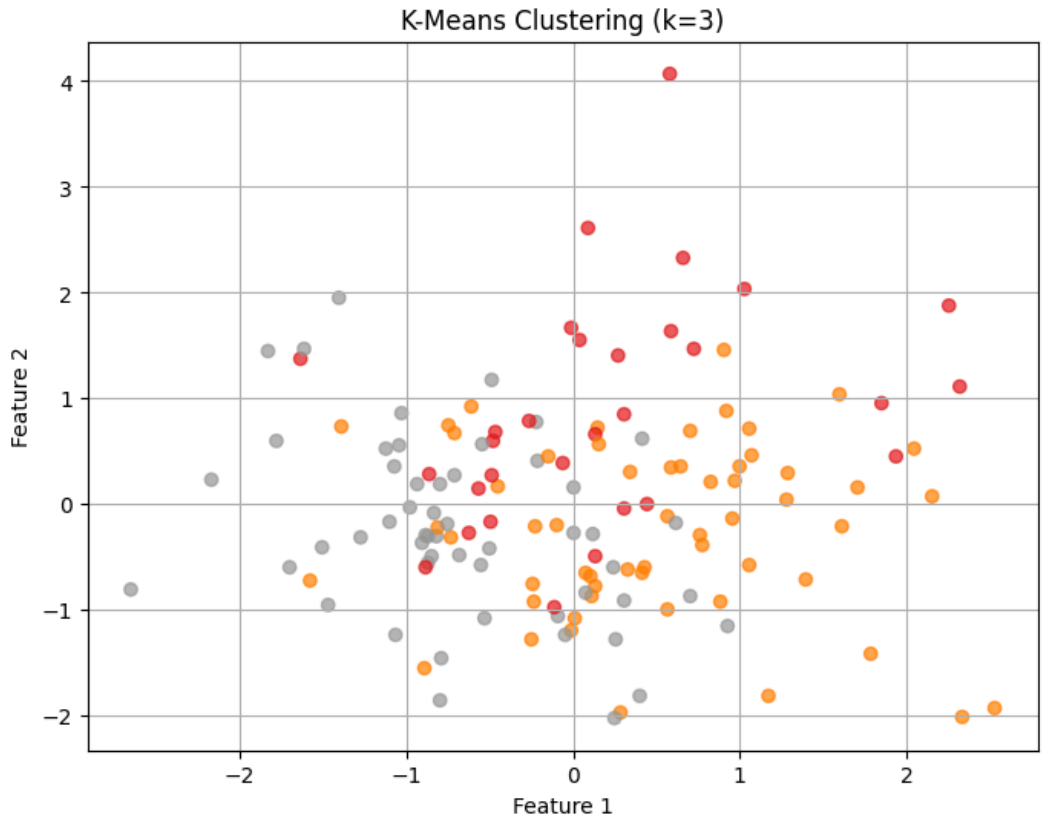
4. XGBoost Regressor → RMSE: 0.752, R²: 0.102:



5. kNN Regressor→ RMSE: 0.774, R²: 0.048:



6. **K-Means Clustering:**

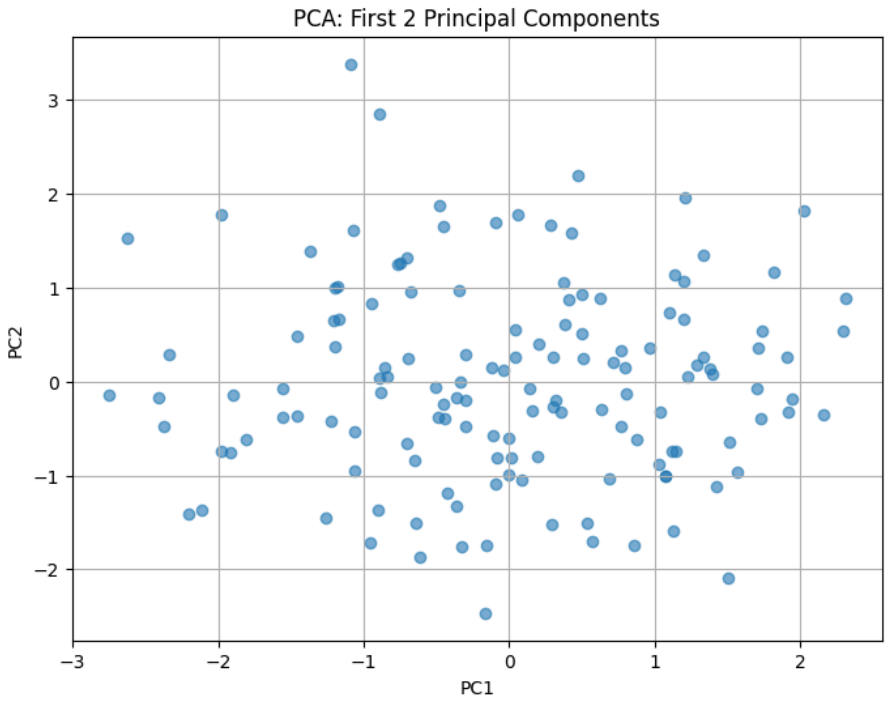


- One group (**red**) included **high-growth, low-unemployment economies** with strong government R&D support and high-tech exports — typically representing **high-investment innovation leaders** such as **Germany, South Korea, and Japan**.

- Another cluster (**gray**) reflected **economies with higher inflation and unemployment and lower government involvement**, corresponding to **resource-constrained or transitional economies** like **India, Brazil, and South Africa**.

- The third group (**orange**) showed **moderate macroeconomic conditions and R&D spending**, forming a **middle tier of innovation capacity** with countries like the **United Kingdom, China, and Turkey**.

7. PCA analysis



PCA Analysis Summary:

The PCA plot displays all country-year pairs in a reduced two-dimensional space based on their macroeconomic and innovation features. The horizontal axis (**PC1**) captures the tradeoff between **high-tech exports** (positive side) and **GDP growth, inflation control, and government R&D support** (negative side). For example, country-years like **Germany, South Korea, and Japan** are likely positioned on the **left side of PC1**, reflecting strong growth and high public R&D investment. Countries like **Brazil, India, and South Africa** likely appear on the **right side**, emphasizing export-driven performance with less fiscal support.

The vertical axis (**PC2**) is more influenced by **unemployment levels**, helping distinguish countries based on labor market stability. Country-years with relatively **higher unemployment** (e.g., **Greece, South Africa**) may be positioned **higher on the PC2 axis**, while those with **lower unemployment** (e.g., **Japan, Germany**) appear **lower**, reflecting more stable employment environments that favor sustained R&D investment.