PRACTICE PROJECT · DATA SCIENCE & ECONOMICS

Assessing the Returns on Investment: Research & Development to Growth

Does R&D expenditure translate to total factor productivity?

Abir Hossain

April 07, 2026

Keywords: computational modeling · data science · R&D · patents · economics

Contents

1. Project Summary

Economic growth is the engine of national development — it funds infrastructure, drives socioeconomic progress, and improves quality of life. Growth fundamentally relies on mobilizing intellectual potential through strategic R&D investment to transform innovative ideas into measurable outcomes.

Economic leadership correlates directly with persistent investment in technical research and higher education. Conversely, underdevelopment is defined by chronic underinvestment in these sectors. Data from high-performing and emerging economies historically confirms this trajectory: growth consistently mirrors the intensity of R&D commitment.

This study attempts to formalize that intuition through computational analysis of historical data. The analytical chain under examination is:

R&D Expenditure → Patent Output → Economic Growth

As a link between investment and actualized innovation, patent production is viewed as a quantifiable stand-in for the useful results of research activities. The data spans allowed range by sources and practical scope of this study and was sourced from three internationally recognized repositories: the **World Bank**, the **OECD**, and the **USPTO**.

1.1. Primary Research Questions

→ Does national R&D spending correlate with patent output?
→ Does patent output correlate with GDP growth?
→ Is there a detectable lag effect between innovation activity and economic outcomes?

1.2. Secondary Research Questions

→ Which countries convert R&D investment into patents most efficiently?
→ Which countries convert patent output into GDP growth most efficiently?
→ Is innovation productivity saturating in high-income countries?
→ Are emerging economies more innovation-efficient per dollar spent?
→ Does patent volume plateau beyond certain R&D spending thresholds?

> Further details on data sources, indicators, and processing are documented in the **Methodology** section of this report.

2. Purpose of This Project

This project is the fifth part in an effort to practice data science, data engineering, and computational modeling toolsets and workflows.

The subject — R&D investment and its connection to economic growth — also relates to a wider area of personal interest because it stimulates a natural intuition through fundamental standards of reasoning and presents an intriguing challenge of experimentation in the hopes of obtaining practical answers to practical questions in finance and economics.

Beyond the research interests, this project served a specific technical objective: developing competency in **ETL pipeline design and implementation**. In an earlier project on proteins, the ETL component was left ambiguous. This project addressed that gap directly, using Jupyter notebooks to test and validate each pipeline stage before committing to production-level code.

Several large raw datasets were chosen intentionally to handle in memory via standard pandas loading. This project introduced **chunked data loading** and **Parquet files** as a practical solution — processing data in manageable portions to avoid exhausting system RAM.

The result is a workflow that is reproducible, falsifiable, and structured in a manner consistent with standards for credible empirical research.

3. Methodology

3.1. Data and Sources

Data for this study was obtained from three publicly available sources, each contributing a distinct category of indicators. The goal was to assemble a longitudinal, cross-national dataset covering as many countries and years as the sources permitted, with filtering applied downstream during the join and subset stages.

3.1.1. World Bank

The World Bank's open data API was queried to retrieve the following macroeconomic indicators:

Indicator	Description
`R&D expenditure (% of GDP)`	National spending on R&D as a share of GDP
`GDP growth (annual %)`	Year-on-year GDP growth rate
`Population`	Total population, used as a scaling variable
`Tertiary enrollment (% gross)`	Gross enrollment ratio in tertiary education

3.1.2. OECD

The OECD's **Main Science and Technology Indicators (MSTI)** dataset was accessed via the OECD Data Explorer API. The following indicators were extracted:

Indicator	Description
`GERD`	Gross domestic expenditure on R&D
`BERD`	Business enterprise expenditure on R&D
`GOVERD`	Government expenditure on R&D
`Researchers`	Full-time equivalent researchers per country per year

3.1.3. USPTO

Two core datasets were obtained from the United States Patent and Trademark Office's public bulk data repository, hosted on Amazon S3:

File	Description
`g_patent.tsv`	Patent-level records including application year, grant year, and patent type
`g_inventor_disambiguated.tsv`	Disambiguated inventor records linked to patents and location identifiers
`g_location_disambiguated.tsv`	Mapping of `location_id` values to country names and country codes

3.2. ETL Pipeline

3.2.1. Extraction

In order to consistently retrieve and validate a variety of datasets from the World Bank, OECD, and USPTO, this stage required building particular APIs and chunked-loading protocols. The procedure made sure that large-scale data extraction was reliable and repeatable by checking for endpoint accuracy and RAM efficiency.

3.2.2. Transformation

This stage involved validating data types and null coverage for World Bank and OECD sources while performing memory-efficient joins and ISO-standardization on millions of USPTO records, ultimately developing a clear strategy to transform the data into a usable format for analysis.

3.2.3. Load

Cleaned data was stored in `.csv` and `.parquet` formats, with Parquet prioritized for its superior columnar storage efficiency and query performance during analysis.

The following analytical subsets were produced:

Subset	Contents
`subset_A.parquet`	Patent counts with GDP growth and population; no R&D expenditure requirement
`subset_B.parquet`	Patent counts with R&D expenditure, GDP growth, population, and tertiary enrollment
`subset_C.parquet`	A stricter filtered subset with fewer null values in some indicators
`patent_counts.parquet`	Aggregated patent counts by country and year
`oecd_clean.parquet`	Cleaned OECD MSTI data
`wb_clean.parquet`	Cleaned World Bank data

> The actual ETL pipeline was designed using notebooks `extraction_validation.ipynb`, `transformation_validation.ipynb`, and `load.ipynb`. This ensured all ideas, methods, and logic were thoroughly tested to find any vulnerabilities or inconsistencies that might arise during implementation.

> → All notebooks can be found in the dedicated `notebooks` directory on GitHub.

→ ETL implementation scripts: `extract.py`, `transform.py`, `load.py` — found in the `ETL` directory.
→ The analytical models were implemented via `model.py` in the `models` directory.

4. Notebooks

4.1. Overview

Jupyter notebooks served as a validation layer for the whole pipeline in this project, ensuring that every logical step was verified separately before being included in production scripts. In addition to ensuring that the final scripts were based on previously validated logic and results, this dual-track method produced a documented audit trail of the development process.

4.2. ETL Notebooks

4.2.1. Extraction — `extraction_validation.ipynb`

This notebook covers the design and testing of the data extraction pipeline.

World Bank: Before a dependable implementation was developed, several iterations of the API request were tried. At first, testing was restricted to two indicators for a shorter range of years. The entire indicator set and all accessible years were added to the extraction template after the pattern was verified.

OECD: The OECD Data Explorer's URL generator was used to produce a valid API endpoint. The trial-and-error process and the final URL structure are documented and explained within the notebook.

USPTO: The patent and inventor files are substantial in size — roughly 2.1 GB combined, with the inventor dataset alone containing approximately 23.7 million rows. Both files were downloaded in chunked batches from the public S3 source. Chunk size was determined empirically: sizes were tested against local RAM utilization, and a value of **500,000 rows per chunk** was selected as the safe upper bound for the used local system.

4.2.2. Transformation — `transformation_validation.ipynb`

This notebook documents all data quality checks, format corrections, and join logic applied to the raw datasets.

World Bank: Initial checks covered data types, null patterns, and country code formatting.

OECD: The full MSTI dataset was inspected and filtered to retain only the four indicators selected for this study.

USPTO: All processing was done in segments. The processed row counts: 23,753,556 rows from `g_inventor_disambiguated.tsv` and 9,361,444 rows from `g_patent.tsv`.

A data type discrepancy between the patent and inventor files required fixing before merging. Another significant problem discovered was that the inventor file only had a `location_id` field rather than a nation or country code column. In order to resolve location identifiers to country names, a third file, `g_location_disambiguated.tsv`, had to be downloaded. Country codes originally in **ISO alpha-2** format were converted to **alpha-3** format for consistency.

4.2.3. Load — `load.ipynb`

The load stage required no complex transformation. Its primary function was converting the cleaned and merged `.csv` files into `.parquet` format for efficient downstream querying. Both `.csv` and `.parquet` versions of all datasets were saved in the local `processed/` directory.

4.3. Analysis Notebook — `analysis_test.ipynb`

The project's analytical center is this notebook. It documents all exploratory concepts, statistical tests, interim results, and interpretation choices during the analysis stage. Each of the eight sections corresponds to one of the primary or secondary research questions.

> **Note on a data quality issue discovered late:** GDP or GDP per capita was left out of the initial indicator extraction. This indicator would have been required to independently categorize nations by income level. The workaround used in secondary analysis is described in the Analysis and Results section. The reasons behind this decision are explained in the Discussion section.

5. Analysis and Results

The analysis follows this funnel structure:

- Raw correlations (Pearson, Spearman) determine whether relationships exist and whether they are linear or monotonic.
- Log-transformations and OLS regression quantify the functional form and control for confounders like population.
- Lag analysis determines whether the relationships are time-displaced rather than concurrent.
- Income-group stratification determines whether aggregate signals hide divergent behavior across high-income and emerging economies.
- Efficiency measures are derived formulas that normalize patent and GDP production by investment and workforce size, allowing cross-country comparisons.
- The plateau question was answered and confirmed by combining binned aggregation with logarithmic curve fitting.

5.1. Primary Questions

5.1.1. Q1. Does national R&D spending correlate with patent output?

The relationship's shape can be seen in an initial scatter plot of raw R&D spending (% of GDP) against the total number of patents. The distribution is heavily right-skewed, with most countries clustered near the origin and a small number of high-investment nations producing disproportionately large patent volumes.

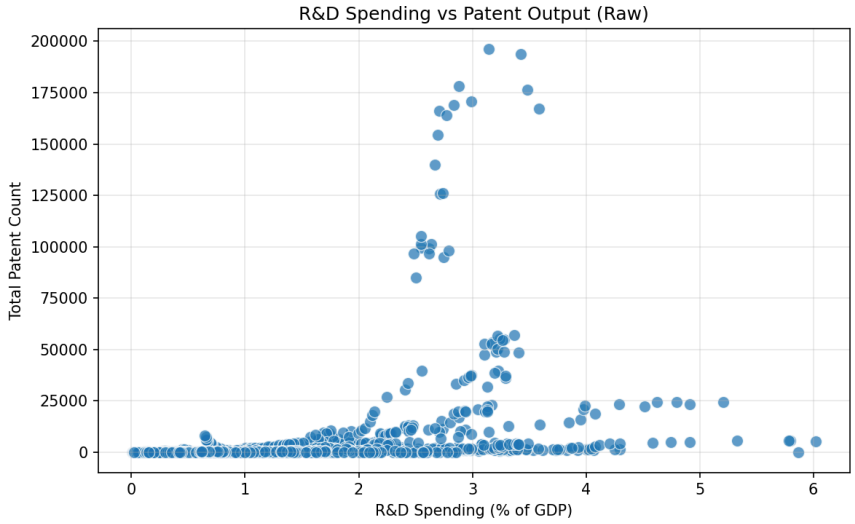

Figure 1: R&D spending (% of GDP) vs. raw patent count across all country-year observations. The distribution is heavily right-skewed, with most countries clustered near the origin and a small number of high-investment nations producing disproportionately large patent volumes.

The Pearson correlation on raw values yields a moderate positive association:

$$r = 0.355, \quad p < 0.001$$

However, this understates the true strength of the relationship. The Spearman rank correlation — robust to outliers and distributional skew — is substantially higher:

$$\rho = 0.808, \quad p < 0.001$$

The large gap between the two statistics ($\Delta = 0.453$) indicates that the relationship is **monotonic but non-linear**: as R&D spending increases, patent output increases at an accelerating rather than a constant rate.

Applying a log-transformation to patent counts ($\log(\text{patents} + 1)$) substantially improves the linear fit:

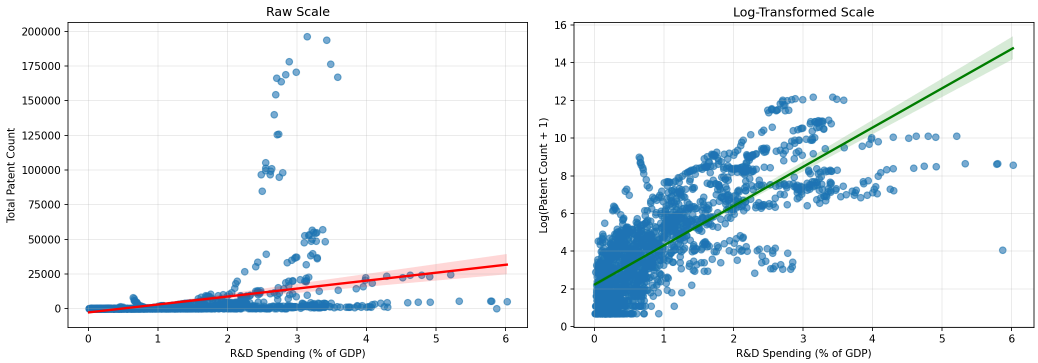

Figure 2: Regression plots of R&D spending vs. patent output. Left: raw scale (Pearson r = 0.355). Right: log-transformed patent count (Pearson r = 0.791). The log scale linearises the relationship and confirms an exponential pattern.

$$r_{\log} = 0.791, \quad p < 0.001$$

A given increase in R&D spending (% of GDP) is associated with a **multiplicative** increase in patent output rather than an additive one.

OLS regression with log(patents) as the outcome and both `rd_gdp` and log(population) as predictors yielded:

Predictor	Coefficient
Intercept	−7.76
R&D spending (% GDP)	+2.00
log(Population)	+0.61

Holding population constant, each additional percentage point of GDP directed toward R&D is associated with a **doubling** of expected patent output on the log scale.

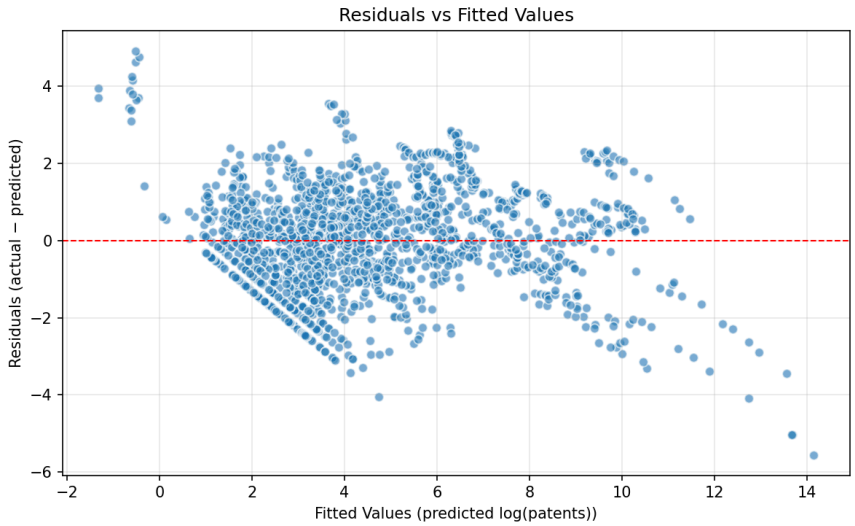

Figure 3: Residuals vs. fitted values from the OLS model (outcome: log patent count; predictors: R&D % GDP and log population). Residuals are centred near zero (mean = −0.0000, SD = 1.30) with no obvious structure, indicating a reasonable model fit.

Summary: R&D spending is a strong positive predictor of patent output. The relationship is monotonic and non-linear — consistent with an exponential scaling pattern — and is highly statistically significant across all tested specifications.

5.1.2. Q2. Does patent output correlate with GDP growth?

The raw-scale and log-scale scatter plots are shown side by side below.

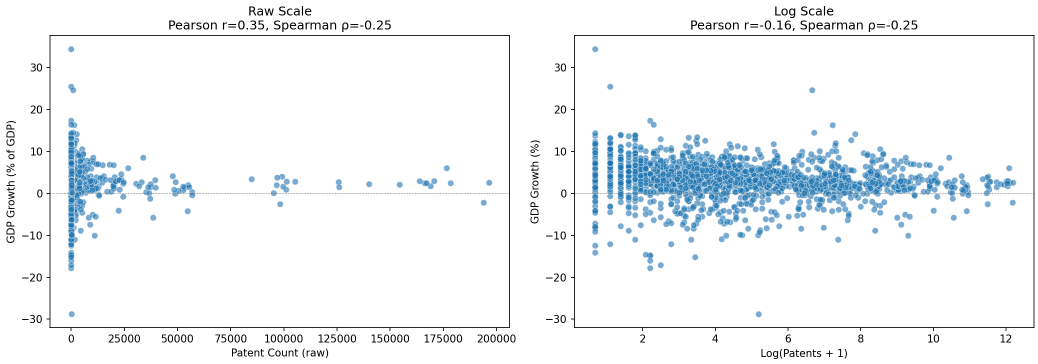

Figure 4: GDP growth (% of GDP) as a function of patent count. Left: raw patent count (Pearson r = −0.055, Spearman ϱ = −0.248). Right: log-transformed patent count (Pearson r = −0.162, Spearman ϱ = −0.248). The data likely follows a non-linear or skewed downward trend, where the rank-order relationship is stronger than the direct linear association.

The near-zero Pearson correlation ($r = -0.055$) confirms the absence of a linear relationship, while the Spearman correlation ($\rho = -0.248$, $p < 0.001$) reveals a persistent, non-linear downward trend.

After log-transforming patent counts, both measures align:

$$r_{\log} = -0.162, \quad \rho_{\log} = -0.248, \quad p < 0.001$$

Countries with higher patent output do not systematically experience higher GDP growth.

5.1.3. Q3. Is there a lag effect?

Given that the economic payoff from research and innovation would realistically take years to materialise, a multi-year lag analysis was conducted for both directions of the relationship.

R&D spending → Patent output (lags 0–5 years):

Lag (years)	Spearman ϱ	p-value	n
0	0.808	< 0.001	1,801
1	0.806	< 0.001	1,672
2	0.806	< 0.001	1,556
3	0.806	< 0.001	1,449
4	0.805	< 0.001	1,347
5	0.804	< 0.001	1,250

The R&D → patent association is virtually time-invariant across all lag lengths, suggesting that patent output responds to R&D investment without a detectable multi-year delay at annual resolution.

Patent output → GDP growth (lags 1–12 years):

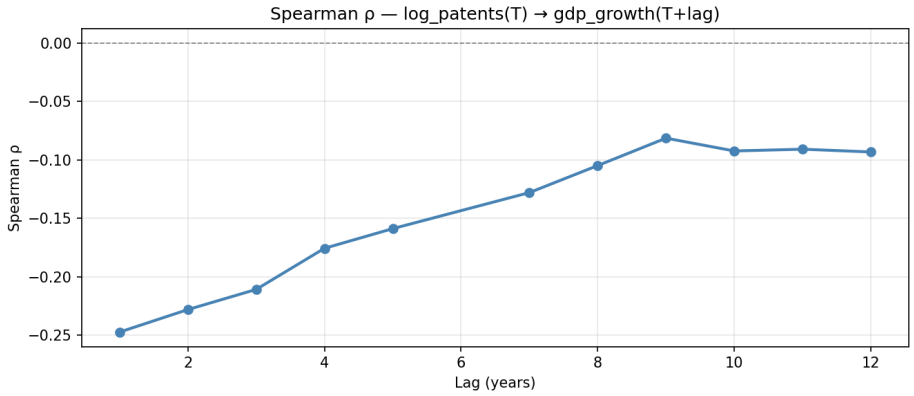

Figure 5: Spearman ρ between lagged log(patent count) and GDP growth, for lags 1 through 12 years. The correlation remains negative throughout, weakening gradually from $\rho = -0.247$ at lag 1 to $\rho \approx -0.081$ at lag 9 before levelling off. No positive reversal is observed at any lag length tested.

Lag (years)	Spearman ρ	p-value
1	−0.247	< 0.001
2	−0.228	< 0.001
3	−0.211	< 0.001
4	−0.176	< 0.001
5	−0.159	< 0.001
7	−0.128	< 0.001
8	−0.105	0.001
9	−0.081	0.016
10	−0.092	0.009
11	−0.091	0.015
12	−0.093	0.018

Testing for reverse causality: lagging GDP growth by one year and correlating with current patent output yielded $\rho = -0.259$ ($p < 0.001$) — also negative. This rules out the possibility that GDP growth is itself driving patent activity.

> **Summary:** The most plausible interpretation is a **short-term resource cost** — patenting is expensive and R&D-intensive, generating a drag on measured GDP growth in the near term. Any long-term growth payoff, if it exists, appears too diffuse or conditional to emerge in a simple bivariate correlation at annual resolution.

5.2. Secondary Questions

5.2.1. Q4. R&D → Patent Efficiency by Country

Efficiency metric:

$$E_1 = \frac{\text{patent\_count}}{\text{rd\_gdp} \times \text{inventor\_count}} \times 10^6$$

This formulation normalises patent output by both the scale of R&D investment and the size of the inventor workforce. Minimum thresholds were applied: at least 500 patents, 2,000 inventors, 1.0% R&D/GDP, and 5 million population — reducing the dataset to **324 observations across 20 countries**.

Top 10 countries by mean R&D → patent efficiency:

Country	Mean E_1	Total Patents	Mean R&D (% GDP)
Italy (ITA)	600,573	68,816	1.24%
Spain (ESP)	432,166	12,205	1.29%
China (CHN)	406,881	244,159	1.96%
United Kingdom (GBR)	402,119	156,642	1.95%
Australia (AUS)	400,311	17,759	2.04%
Canada (CAN)	399,072	158,514	1.84%
Netherlands (NLD)	376,873	64,011	1.95%
Singapore (SGP)	365,001	6,239	1.94%
France (FRA)	303,343	148,026	2.17%
Belgium (BEL)	276,552	19,515	2.85%

5.2.2. Q5. Patent → GDP Growth Efficiency by Country

Efficiency metric:

$$E_2 = \frac{\text{gdp\_growth}}{\text{patent\_count} \;/\; \text{population}} \times 10^6$$

Top 10 countries by mean patent → GDP growth efficiency:

Country	Mean E_2	Total Patents	Mean R&D (% GDP)
China (CHN)	2.01×10^{12}	244,159	1.96%
Spain (ESP)	7.45×10^{10}	12,205	1.29%
South Korea (KOR)	2.68×10^{10}	319,187	3.51%
Australia (AUS)	2.15×10^{10}	17,759	2.04%
United Kingdom (GBR)	1.65×10^{10}	156,642	1.95%
France (FRA)	1.60×10^{10}	148,026	2.17%
Canada (CAN)	1.20×10^{10}	158,514	1.84%
Netherlands (NLD)	1.16×10^{10}	64,011	1.95%

Israel (ISR)	1.15×10^{10}	63,317	4.56%
Italy (ITA)	1.11×10^{10}	68,816	1.24%

China's dominant position in E_2 reflects its high GDP growth rates during the study period combined with its large population, which keeps the per-capita patent denominator low even as absolute patent volume grows.

5.2.3. Q6. Innovation Productivity Saturation in High-Income Countries

Data workaround applied here. Because GDP per capita was not extracted in the original pipeline, countries could not be classified by income level from the data alone. Income group labels were sourced externally from World Bank and IMF reference lists and applied as a static lookup. This classification has not been independently verified within the scope of this study.

The chart below tracks mean E_1 (R&D → patents) and E_2 (patents → GDP growth) over time for High Income and Emerging economy groups.

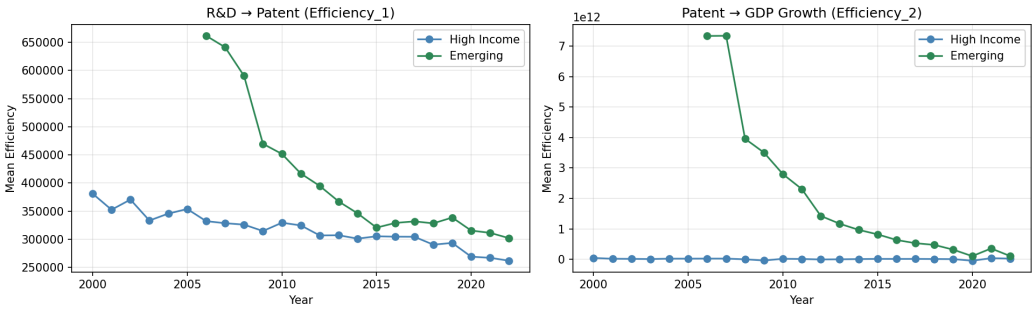

Figure 6: Mean R&D → patent efficiency (E_1, left) and patent → GDP growth efficiency (E_2, right) by year and income group (2000–2022). Both groups show declining E_1 over time. E_2 declines significantly for Emerging economies but shows no statistically significant trend for High Income countries.

Linear regression of mean E_1 over time by group:

Group	Slope (E_1 per year)	R^2	p-value
High Income	−4,374	0.907	< 0.001
Emerging	−20,644	0.777	< 0.001

Both groups show a statistically significant **decline** in R&D → patent efficiency over time, consistent with a diminishing returns hypothesis. The steeper decline in emerging economies may reflect rapid scaling of R&D programmes outpacing the growth of productive inventor capacity.

The scatter below shows where each group sits in R&D spend vs. patent output space, revealing the structural differences in innovation volume between the two groups.

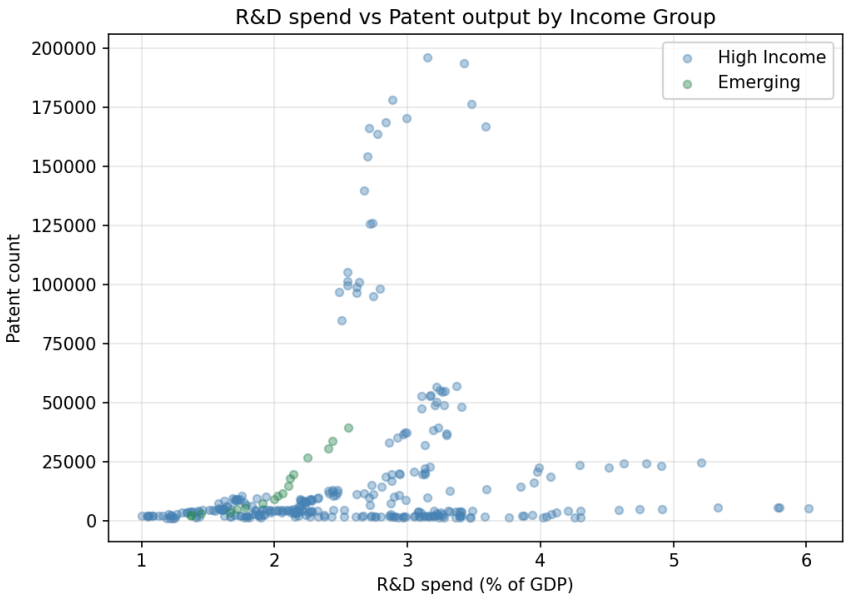

Figure 7: R&D spending (% of GDP) vs. patent count, coloured by income group. High Income countries span a wider range of R&D intensities and consistently produce higher absolute patent volumes. Emerging economies are concentrated at lower R&D spending levels with greater variance in patent output.

5.2.4. Q7. Are Emerging Economies More Innovation-Efficient Per Dollar?

A per-unit efficiency metric was computed: $E_{1,\text{norm}} = E_1 / \text{rd\_gdp}$

Mean normalised efficiency by group:

Group	Mean E_1 / R&D unit
High Income	155,885
Emerging	226,521

Emerging economies produce approximately **45% more patents per unit of R&D investment** than high-income countries on average.

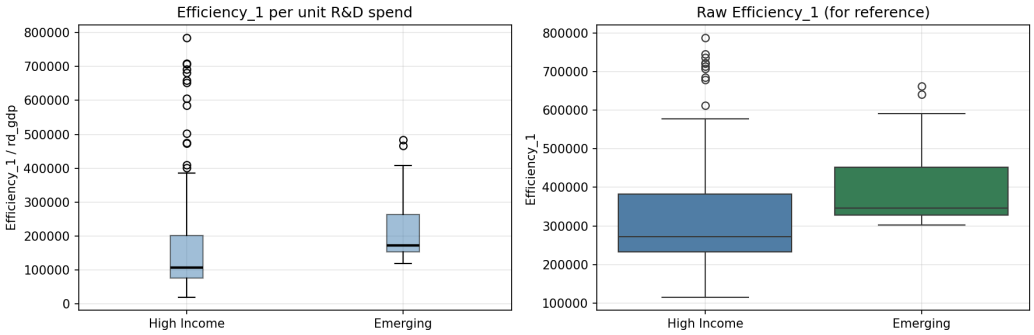

Figure 8: Left: Distribution of E_1 normalised by R&D spending (E_1 / rd_gdp) for High Income vs. Emerging economies. Emerging economies show a higher median and wider spread. Right: Raw E_1 for reference. The higher per-dollar efficiency of emerging economies is consistent across the distribution, not driven by outliers alone.

This is consistent with theoretical expectations: high-income economies direct R&D toward complex, incremental frontier innovation where the marginal cost per patent is higher, while emerging economies may capture more available patent space through technology adoption and applied engineering at lower marginal cost.

5.2.5. Q8. Does Patent Volume Plateau Beyond Certain R&D Spending Thresholds?

R&D spending was binned into deciles and mean patent output was calculated within each bin, separately for High Income and Emerging countries, to test for a saturation effect.

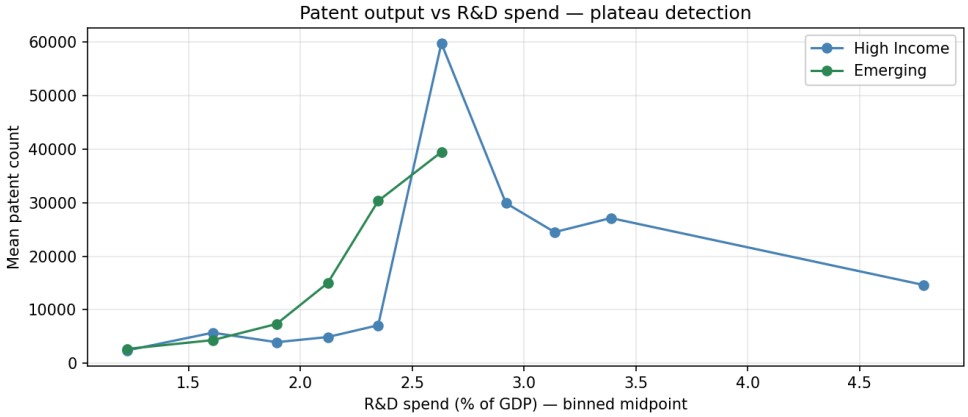

Figure 9: Mean patent count by R&D spending decile (bin midpoint), for High Income and Emerging economy groups. Both curves follow a broadly concave trajectory — rising steeply at lower R&D spending levels and flattening at higher spending — consistent with diminishing marginal returns to R&D investment.

A logarithmic curve was then fitted directly to High Income country observations to quantify the saturation shape:

$$\hat{y} = a \cdot \ln(x) + b$$

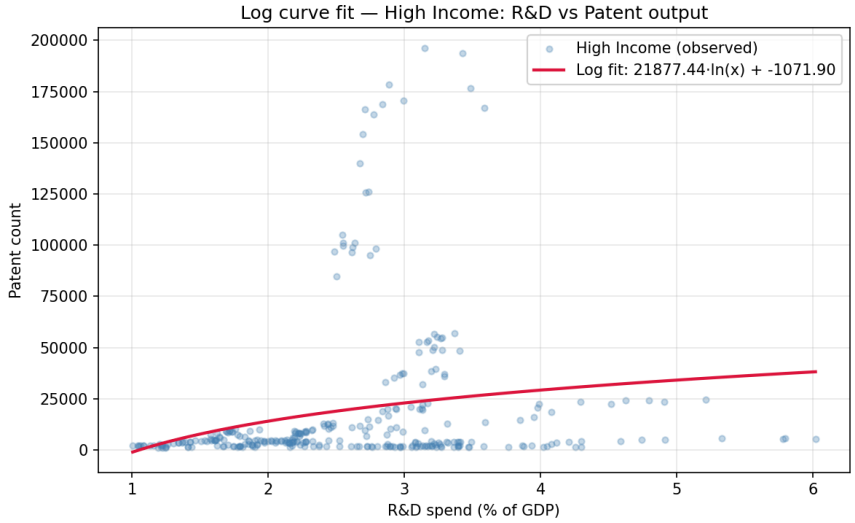

Figure 10: Logarithmic curve fit (crimson line) to High Income country observations of R&D spending vs. patent count. The log model captures the concave shape of the relationship: large patent gains at lower R&D spending levels, with returns flattening progressively as spending rises toward and beyond 3% of GDP.

The log curve fit confirms a **saturation pattern**: early increases in R&D spending yield proportionally large gains in patent output, but the marginal return diminishes as spending rises. The effect is most clearly demonstrated in High Income countries, which occupy a wider R&D spending range where the flattening of the curve is directly observable.

6. Discussion

6.1. Concluding Remarks

Across the eight analyses, the most robust finding is the **strong positive association between R&D spending and patent output** (Spearman $\rho \approx 0.81$), which is stable across lag windows, statistically highly significant, and consistent with a log-linear functional form. This relationship holds across income groups and is not an artifact of country size or population.

The **patent → GDP growth** relationship is weaker, consistently negative in bivariate tests, and does not resolve to a positive signal at any lag length tested up to twelve years. This most likely reflects the limitations of a bivariate framework that cannot account for institutional quality, human capital, sector composition, and other confounders that would be necessary for a credible causal claim.

The secondary findings — declining innovation efficiency over time, higher per-dollar efficiency in emerging economies, and logarithmic saturation in the R&D → patent curve — are coherent and mutually reinforcing. They suggest that the global innovation system is characterised by diminishing marginal returns, and that the productivity advantage of emerging economies, while real, is narrowing.

6.2. About This Study

This study offers a data-driven examination of how R&D investment connects to patent output and economic growth across nations. While the analytical chain is logically coherent and the methodology follows a defensible structure, the results should not be interpreted as definitive or as directly comparable to peer-reviewed international research standards. Several factors constrain the conclusions:

- Data acquisition was limited in size, temporal range, indicator depth, and geographic scope relative to what a fully resourced study would employ.
- The statistical methods, while appropriate for the questions posed, are not exhaustive. A study of this kind pursued to academic publication standard would incorporate panel regression, instrumental variables, and more rigorous causal identification.
- This is a solo, unfunded practice project completed over a short timeframe.
- The filtering thresholds applied in the efficiency calculations are reasonable but not derived from a standard benchmark. Different threshold choices would produce different country rankings.
- The primary objective was never to produce a novel research finding, but to practice data engineering workflows, ETL pipeline design, and analytical methods on a large, multi-source dataset.

With those constraints acknowledged, the study remains a credible starting point for anyone interested in approaching this topic quantitatively. It is designed to be reproducible, scalable, and falsifiable.

6.3. Limitation Acknowledgements

ETL design gap: The most consequential mistake in this project was proceeding to full pipeline implementation before completing the analysis design. Because the transformation and loading stages were built out ahead of time, the absence of a GDP per capita indicator was not discovered until the secondary analysis questions required country income classification. By that point, re-running the pipeline was not practical within the project time-limit. The lesson is clear: analysis requirements must be fully mapped before ETL design is locked.

Notebook structure: In hindsight, the decision to organise work into four strictly separated notebooks — one per pipeline stage — was not the most efficient approach. A more effective strategy might have been to begin from the analysis layer and work backwards toward the data, moving fluidly between layers while maintaining a clear record of each decision.

Documentation depth: The documentation does not achieve the level of mathematical rigour or explanatory detail that a formal research report would require. This was a deliberate trade-off: the focus was on tooling and workflow practice, not on presentation.

Version control: The codebase was not tracked with Git. Given a working familiarity with Git, this was an omission rather than a knowledge gap — one that would not be repeated in any collaborative or longer-term project.

6.4. Author's Note

Five projects of this kind completed within roughly 30–40 days have produced something tangible: a working fluency with the full workflow from raw data acquisition through ETL design, analysis, and documentation. The process is no longer unfamiliar territory.

This is the last of the practice projects. The purpose they served — building baseline competency in data engineering, statistical analysis, and reproducible research design — has been fulfilled to a degree sufficient to move forward with original work. Future projects will be given more time individually, allowing for greater depth, rigour, and quality at every stage.

The direction from here is toward statistically and mathematically deeper methods, professionally acknowledged frameworks, and eventually machine learning tools including GPU-accelerated computation, PyTorch, and deep learning architectures where the problem warrants them. Several original project ideas are already in early groundwork. The foundation is in place.