BUSINFO 702 2025Q3 Project Report

Project Title

Portrait of the Female Scientist: Global Drivers of Women's Participation in STEM

Project Group

Group number: 9

Group members and roles:

Name	SID	Responsible for
Griffin Chau	128235564	ERD, Schema Design
Shruti Chavan	138477436	SQL Methodology and Implementation
Leo Chung	669027346	Research, Motivation, Schema Design
Quoc Nguyen (Addy)	356205885	Dataset, Project Coordination, Editorial

Executive Summary

This project investigates the global drivers of women's participation in STEM fields between 2005 and 2021, using comprehensive data from six countries: Australia, Canada, China, Germany, India and the United States. The analysis integrates four datasets from Kaggle: "Women's Representation in Global STEM Education" (women_stem), "World Happiness Report- 2024" (world_happiness), the gender inequality index (gender_inequality), and "Share of the population using the Internet" (internet_usage). The analysis provides an intricate vision of how sociopolitical, economic, and technological factors influence female engagement in STEM education.

Drawing from the four integrated datasets, the analysis addressed four guiding questions:

- To what extent does digital access (internet usage) shape women's opportunities in STEM?
- How do year-on-year changes in gender inequality (GII) relate to the growth of female STEM representation?
- What is the relationship between freedom of expression and female STEM enrolment rates over time in Asia and America, and how do patterns differ between the two continents?

Following the implementation of analytical processes, the analysis yielded the following findings:

- In emerging economies, rising internet access coincided with notable gains in women's STEM participation, whereas in developed countries with long-standing connectivity, female enrolment and graduation did not consistently increase.
- Year-on-year changes in gender inequality showed weak and inconsistent relationships with female STEM enrolment and graduation. The correlations were near zero in countries like Australia and Canada, moderately positive in the United States and China, and slightly negative in Germany and India.
- STEM representation shows a steady uptrend in Asia where freedom of expression is developing, but relatively stable in America where freedom of expression is well-established.

Research Questions

RQ1: To what extent does digital access (internet usage) shape women's opportunities in STEM?

Current perspectives highlight digital access as a crucial driver in expanding women's participation in STEM by facilitating entry into education, professional development, and innovation, while addressing long-standing obstacles such as insufficient exposure, ingrained biases, and resource limitations. In countries with widespread internet connectivity like the United States, Canada, and Germany, women benefit from abundant online learning resources, global professional networks, and interactive skill-building platforms. This question therefore aims to quantify whether internet access and its benefits collectively foster a stronger environment for engagement and success in STEM fields.

RQ2: Between 2005 and 2021, how do year-on-year changes in gender inequality (GII) relate to the growth rates of female enrolment and graduation in STEM across Australia, Canada, China, Germany, India, and the USA?

The Gender Inequality Index (GII) captures disparities in health, empowerment, and labor market participation. In theory, reductions in inequality should encourage more women to enroll and graduate in STEM fields, while increases may create additional barriers. Developed countries like Australia, Canada, Germany, and the United States may show weaker associations given their relatively stable gender equality. In contrast, emerging economies like China and India may exhibit stronger links due to ongoing social and educational transitions. This research question tests whether yearly changes in GII would lead to measurable impacts on female STEM enrolment and graduation across different national contexts.

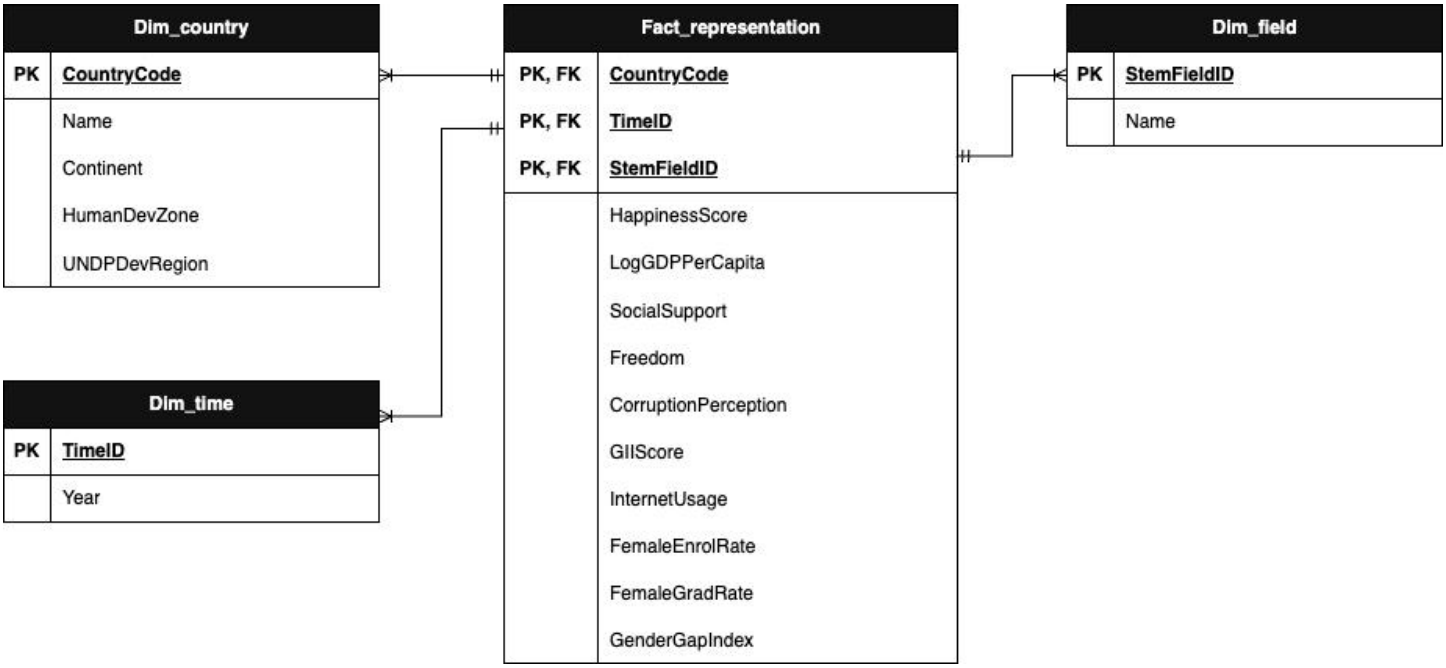
RQ3: What is the relationship between freedom of expression and female STEM enrolment rate in Asia and America over time? What are the differences between the two continents?

Conventional wisdom suggests that better freedom of expression might encourage female students to enroll in STEM fields against entrenched societal prejudices. The American continent, which includes the United States and Canada in our dataset, historically leads in terms of political and sociological freedom. In contrast, the Asia continent, which includes India and China, is still developing in that regard. Hypothetically, the American continent should show higher overall or faster-growing female enrolment over time. This research question seeks to verify whether having more freedom of expression would promote female participation in STEM and would therefore be an essential metric in policymaking.

Star Schema

Figure 1.

Star Schema of The Project's Data Warehouse



In the fact table and dim_field, the fields "STEM_Field", "FemaleEnrolRate", "FemaleGradRate", and "GenderGapScore" originate from the women_stem dataset, acting as the central metric of our report. While all datasets share the country and year attributes, we are concerned with the six countries available in the women_stem dataset. For the time dimension, we determined the mutual duration among all datasets to ensure as much data is available for all countries, from 2005 to 2021.

The "HappinessScore", "SocialSupport", "CorruptionPerception", "Freedom" and "LogGDPPerCapita" attributes come from the world_happiness dataset, describing these metrics in a given country in a specific year. Similarly, "InternetUsage" comes from the internet_usage dataset, and "GIIScore" comes from the gender_inequality dataset. This provides the sociopolitical backdrop to analyze trends in female STEM enrolment.

In the dim_country table, other attributes are included for every given country, including "Continent", "HumanDevelopmentGroups", and "UNDPDevelopmentRegions". "Continent" acts as a higher order in the country hierarchy, allowing for analyses on the continental level.

ETL Implementation in SQLite

Extract

The datasets used in this project were downloaded as CSV files from Kaggle and other publicly available sources. We used four datasets: women_stem, world_happiness, internet_usage, and gender_inequality. At this stage, no changes were made to the data itself; all records, formats, and headers were kept exactly as they appeared in the source files. This ensured transparency and that the raw data remained intact before any transformations.

Load

The raw staging tables from the four datasets were created with matching numbers and names of column headers. Subsequently, each dataset was imported to the corresponding raw staging table via the DB Browser CSV dataset import function. An import command is available for import operation from SQLite CLI.

```
/* 1) Creating a table for Women_in_STEM Dataset */
DROP TABLE IF EXISTS stg_women_stem_raw;
CREATE TABLE stg_women_stem_raw (
  Country TEXT,
  Year INTEGER,
  "Female Enrollment (%)" REAL,
  "Female Graduation Rate (%)" REAL,
  "STEM Fields" TEXT,
  "Gender Gap Index" REAL
);

-- import women_stem.csv into this table via DB Browser import function or execute the
command below in SQLiteCLI
-- .import --csv --skip 1 women_stem.csv stg_women_stem_raw

/* 2) Creating a table for the World_Happiness Dataset */
DROP TABLE IF EXISTS stg_world_happiness_raw;
CREATE TABLE stg_world_happiness_raw (
 "Country name"           TEXT,
  year                  INTEGER,
  "Life Ladder"            REAL,
  "Log GDP per capita"          REAL,
  "Social support"           REAL,
  "Healthy life expectancy at birth"  REAL,
  "Freedom to make life choices"    REAL,
  "Generosity"             REAL,
  "Perceptions of corruption"     REAL,
  "Positive affect"          REAL,
  "Negative affect"          REAL
);

-- import world_happiness.csv into this table via DB Browser import function or execute
the command below in SQLiteCLI
```

```sql
-- .import --csv --skip 1 world_happiness.csv stg_world_happiness_raw

/* 3) Creating a table for the Gender_Inequality_Index Dataset */
DROP TABLE IF EXISTS stg_gender_inequality_raw;
CREATE TABLE stg_gender_inequality_raw (
  ISO3 TEXT,
  Country TEXT,
  Continent TEXT,
  Hemisphere TEXT,
  "Human Development Groups" TEXT,
  "UNDP Developing Regions" TEXT,
  "HDI Rank (2021)" INTEGER,
  "GII Rank (2021)" INTEGER,
  "Gender Inequality Index (1990)" REAL,
  "Gender Inequality Index (1991)" REAL,
  "Gender Inequality Index (1992)" REAL,
  "Gender Inequality Index (1993)" REAL,
  "Gender Inequality Index (1994)" REAL,
  "Gender Inequality Index (1995)" REAL,
  "Gender Inequality Index (1996)" REAL,
  "Gender Inequality Index (1997)" REAL,
  "Gender Inequality Index (1998)" REAL,
  "Gender Inequality Index (1999)" REAL,
  "Gender Inequality Index (2000)" REAL,
  "Gender Inequality Index (2001)" REAL,
  "Gender Inequality Index (2002)" REAL,
  "Gender Inequality Index (2003)" REAL,
  "Gender Inequality Index (2004)" REAL,
  "Gender Inequality Index (2005)" REAL,
  "Gender Inequality Index (2006)" REAL,
  "Gender Inequality Index (2007)" REAL,
  "Gender Inequality Index (2008)" REAL,
  "Gender Inequality Index (2009)" REAL,
  "Gender Inequality Index (2010)" REAL,
  "Gender Inequality Index (2011)" REAL,
  "Gender Inequality Index (2012)" REAL,
  "Gender Inequality Index (2013)" REAL,
  "Gender Inequality Index (2014)" REAL,
  "Gender Inequality Index (2015)" REAL,
  "Gender Inequality Index (2016)" REAL,
  "Gender Inequality Index (2017)" REAL,
  "Gender Inequality Index (2018)" REAL,
  "Gender Inequality Index (2019)" REAL,
  "Gender Inequality Index (2020)" REAL,
  "Gender Inequality Index (2021)" REAL
  );
```

```
-- import gender_inequality.csv into this table via DB Browser import function or execute
the command below in SQLiteCLI
-- .import --csv --skip 1 gender_inequality.csv stg_gender_inequality_raw

/* 4) Creating a table for the Internet_Usage Dataset */
DROP TABLE IF EXISTS stg_internet_usage_raw;
CREATE TABLE stg_internet_usage_raw (
  Entity TEXT,           -- country name
  Code   TEXT,           -- ISO3
  Year   INTEGER,
  "Individuals using the Internet (% of population)" REAL
);

-- import internet_usage.csv into this table via DB Browser import function or execute
the command below in SQLiteCLI
-- .import --csv --skip 1 internet_usage.csv stg_internet_usage_raw
```

Transform

Step 1: Data Cleaning

a) Women in STEM

The women_stem dataset originally reported female enrolment and graduation rates in percentages (0–100). To make the data easier to compare and compute, we converted these percentages into decimal values ranging from 0–1. We also trimmed country names for consistency, ensured years were integers, and standardised the STEM field names. The Gender Gap Index column was retained but also converted into a decimal scale for consistency.

```
-- Women in STEM → tidy & normalize (% → 0–1)
DROP TABLE IF EXISTS stg_women_stem;
CREATE TABLE stg_women_stem AS
SELECT
  TRIM("Country") AS country,
  CAST("Year" AS INT) AS year,
  TRIM("STEM Fields") AS field,
  CAST("Female Enrollment (%)"      AS REAL)/100.0 AS female_enrol_rate,
  CAST("Female Graduation Rate (%)" AS REAL)/100.0 AS female_grad_rate,
  CAST("Gender Gap Index" AS REAL)/100.0 AS gender_gap_index
FROM stg_women_stem_raw;
```

b) World Happiness

The world_happiness dataset contained many indicators, but we selected only a relevant few:

- **Life Ladder (Happiness Score)**
- **Social Support**
- **Perceptions of Corruption**
- **Freedom to Make Life Choices**
- **Log GDP per Capita**

Other variables like generosity, positive affect, and negative affect were excluded to keep the schema clean. Column names were also simplified to align with SQL naming conventions.

```
-- World Happiness (2024) → keep only required columns
DROP TABLE IF EXISTS stg_world_happiness;
CREATE TABLE stg_world_happiness AS
SELECT
  TRIM("Country name") AS country,
  CAST(year AS INT) AS year,
  CAST("Life Ladder"              AS REAL) AS happiness_score,
  CAST("Social support"             AS REAL) AS social_support,
  CAST("Perceptions of corruption"    AS REAL) AS corruption_perception,
  CAST("Freedom to make life choices"   AS REAL) AS freedom,
  CAST("Log GDP per capita"           AS REAL) AS log_gdp_per_capita
FROM stg_world_happiness_raw;
```

c) Internet Usage

The internet_usage dataset reported the percentage of individuals using the internet by country and year. Similar to Women in STEM, the percentage values were converted into decimals (0–1). To ensure consistency with our study period, we kept only the years 2005–2021. ISO3 codes were used to maintain alignment with other datasets.

```
-- Internet Usage (Kaggle) → tidy (% → 0-1)
DROP TABLE IF EXISTS stg_internet_usage;
CREATE TABLE stg_internet_usage AS
SELECT
  TRIM(Code)   AS iso3,
  TRIM(Entity) AS country_name,
  CAST(Year AS INT) AS year,
  CAST("Individuals using the Internet (% of population)" AS REAL)/100.0 AS
internet_usage
FROM stg_internet_usage_raw
WHERE Year BETWEEN 2005 AND 2021;
```

d) Gender Inequality Index (GII)

The gender_inequality dataset was initially in **wide format**, where each year (1990–2021) was a separate column. For analysis, this structure is not practical. We transformed it into a **long format**, where each row contains:

- Country and ISO3 code
- Year (2005–2021)
- The GII score for that year

This pivoting and transposition allowed us to align GII data with the other datasets (2005–2021) on a year-by-year basis.

```
-- GII (wide) → long for 2005-2021
DROP TABLE IF EXISTS stg_gender_inequality_long;
CREATE TABLE stg_gender_inequality_long AS
```

```sql
SELECT TRIM(Country) AS country, TRIM(ISO3) AS iso3, TRIM(Continent) AS continent,
       TRIM("Human Development Groups") AS human_dev_zone,
       TRIM("UNDP Developing Regions")  AS undp_dev_region,
       2005 AS year, CAST("Gender Inequality Index (2005)" AS REAL) AS gii_score
FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2006,CAST("Gender Inequality Index (2006)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2007,CAST("Gender Inequality Index (2007)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2008,CAST("Gender Inequality Index (2008)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2009,CAST("Gender Inequality Index (2009)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2010,CAST("Gender Inequality Index (2010)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2011,CAST("Gender Inequality Index (2011)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2012,CAST("Gender Inequality Index (2012)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2013,CAST("Gender Inequality Index (2013)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2014,CAST("Gender Inequality Index (2014)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2015,CAST("Gender Inequality Index (2015)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2016,CAST("Gender Inequality Index (2016)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2017,CAST("Gender Inequality Index (2017)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2018,CAST("Gender Inequality Index (2018)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2019,CAST("Gender Inequality Index (2019)" AS
REAL) FROM stg_gender_inequality_raw
```

```
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2020,CAST("Gender Inequality Index (2020)" AS
REAL) FROM stg_gender_inequality_raw
UNION ALL SELECT TRIM(Country),TRIM(ISO3),TRIM(Continent),TRIM("Human Development
Groups"),TRIM("UNDP Developing Regions"),2021,CAST("Gender Inequality Index (2021)" AS
REAL) FROM stg_gender_inequality_raw;
```

Step 2: Table creation & normalization

We aligned country names across sources and limited the scope to the six focus countries. Then we filtered all staging data to **2005–2021** and produced clean, aligned "normalized staging" tables ready for dimensional modeling.

```
/* Whitelist of 6 countries for the study */
DROP TABLE IF EXISTS country_whitelist;
CREATE TABLE country_whitelist(iso3 TEXT PRIMARY KEY, std_name TEXT);
INSERT OR REPLACE INTO country_whitelist VALUES
 ('AUS','Australia'),
 ('CAN','Canada'),
 ('CHN','China'),
 ('DEU','Germany'),
 ('IND','India'),
 ('USA','United States');

-- Map country name to ISO3 country codes from GII (plus common US variants)
DROP TABLE IF EXISTS country_map;
CREATE TABLE country_map AS
SELECT DISTINCT TRIM(country) AS raw_name,
       TRIM(iso3) AS iso3,
       CASE WHEN TRIM(iso3)='USA' THEN 'United States' ELSE TRIM(country) END AS std_name
FROM stg_gender_inequality_long;
INSERT OR IGNORE INTO country_map (raw_name, iso3, std_name) VALUES
 ('United States','USA','United States'),
 ('United States of America','USA','United States'),
 ('USA','USA','United States');

 -- Women in STEM
DROP TABLE IF EXISTS stg_women_stem_norm;
CREATE TABLE stg_women_stem_norm AS
SELECT m.iso3, m.std_name AS country_name,
       s.year, s.field,
       s.female_enrol_rate, s.female_grad_rate, s.gender_gap_index
FROM stg_women_stem s
JOIN country_map m       ON s.country = m.raw_name
JOIN country_whitelist w ON w.iso3    = m.iso3
WHERE s.year BETWEEN 2005 AND 2021;

-- World Happiness
```

```sql
DROP TABLE IF EXISTS stg_world_happiness_norm;
CREATE TABLE stg_world_happiness_norm AS
SELECT m.iso3, m.std_name AS country_name,
       h.year,
     h.happiness_score,
       h.social_support,
       h.corruption_perception,
       h.freedom,
       h.log_gdp_per_capita
FROM stg_world_happiness h
JOIN country_map m       ON h.country = m.raw_name
JOIN country_whitelist w ON w.iso3    = m.iso3
WHERE h.year BETWEEN 2005 AND 2021;

-- Internet usage
DROP TABLE IF EXISTS stg_internet_usage_norm;
CREATE TABLE stg_internet_usage_norm AS
SELECT iu.iso3, iu.country_name, iu.year, iu.internet_usage
FROM stg_internet_usage iu
JOIN country_whitelist w ON w.iso3 = iu.iso3
WHERE iu.year BETWEEN 2005 AND 2021;

-- GII (6 countries 2005–2021)
DROP TABLE IF EXISTS stg_gender_inequality_norm;
CREATE TABLE stg_gender_inequality_norm AS
SELECT g.iso3, g.country AS country_name,
       g.year, g.gii_score,
       g.continent, g.human_dev_zone, g.undp_dev_region
FROM stg_gender_inequality_long g
JOIN country_whitelist w ON w.iso3 = g.iso3
WHERE g.year BETWEEN 2005 AND 2021;
```

Step 3: Star Schema (Dimensional Model)

Our analytical model uses a classic **star schema**: three-dimensional tables that describe the “where/when/what”, and one central fact table that stores the measurements. This structure makes it easy to slice metrics by country, year, and STEM field while keeping queries fast and readable.

a) **dim_country — “Where (the place)”**

What it contains: One row per included country with stable descriptors.

Why it exists: Business users can filter and group results by country or regional groupings.

How we populate it: We take authoritative country descriptors from the **Gender Inequality (GII)** dataset (which carries ISO3 codes and regional breakdowns) and keep only the six study countries.

```sql
-- DDL
-- dim_country
```

```sql
DROP TABLE IF EXISTS dim_country;
CREATE TABLE dim_country (
  CountryCode   TEXT PRIMARY KEY,
  Name          TEXT,
  Continent     TEXT,
  HumanDevZone  TEXT,
  UNDPDevRegion TEXT
);

-- LOAD
-- dim_country from GII descriptors (authoritative ISO3 + geo)
INSERT OR IGNORE INTO dim_country (CountryCode, Name, Continent, HumanDevZone, UNDPDevRegion)
SELECT DISTINCT iso3, country_name, continent, human_dev_zone, undp_dev_region
FROM stg_gender_inequality_norm;
```

b) dim_time — "When (the year)"

What it contains: One row per year used in analysis.

Why it exists: Provides a clean "time spine" so every fact is tied to a precise year and we can trend over time.

How we populate it: We take distinct years from **Women in STEM** (our densest source) to ensure all study years are represented.

```sql
-- DDL
--dim_time
DROP TABLE IF EXISTS dim_time;
CREATE TABLE dim_time (
  TimeID INTEGER PRIMARY KEY,
  Year   INTEGER UNIQUE
);

-- LOAD
-- dim_time from years present across study (use women_stem as anchor)
INSERT OR IGNORE INTO dim_time (Year)
SELECT DISTINCT year
FROM stg_women_stem_norm
ORDER BY year;
```

c) dim_field — "What (the STEM discipline)"

What it contains: One row per STEM field (e.g., Engineering, Computer Science).

Why it exists: Lets us compare and filter by discipline while keeping field names consistent.

How we populate it: We take the set of unique fields from the **Women in STEM** dataset.

```sql

-- DDL
-- dim_field
DROP TABLE IF EXISTS dim_field;
CREATE TABLE dim_field (
```

```
  FieldID INTEGER PRIMARY KEY,
  Name    TEXT UNIQUE
);

-- LOAD
-- dim_field from STEM fields
INSERT OR IGNORE INTO dim_field (Name)
SELECT DISTINCT field
FROM stg_women_stem_norm
WHERE field IS NOT NULL;
```

d) fact_representation — "The measures"

What it contains: One row per Country × Year × Field with the actual KPIs we analyze:

- FemaleEnrolRate and FemaleGradRate, GenderGapIndex (from women_stem, scaled 0–1)
- HappinessScore (Life Ladder), SocialSupport, CorruptionPerception, Freedom, LogGDPPerCapita (from world_happiness)
- InternetUsage (from internet_usage, scaled 0–1)
- GIIScore (from gender_inequality, annual score)

Why it exists: This central table brings together all metrics in the same grain so we can answer questions like: How do enrolment and graduation trends vary by field across countries over time, and how are they associated with national context variables (happiness, freedom, internet usage, GII, etc.)?

```
-- DDL
-- fact_representation
DROP TABLE IF EXISTS fact_representation;
CREATE TABLE fact_representation (
  CountryCode     TEXT    NOT NULL,
  TimeID          INTEGER NOT NULL,
  FieldID         INTEGER NOT NULL,
  FemaleEnrolRate REAL,
  FemaleGradRate  REAL,
  GenderGapIndex  REAL,
  HappinessScore  REAL,
  InternetUsage         REAL,
  SocialSupport         REAL,
  CorruptionPerception  REAL,
  Freedom               REAL,
  LogGDPPerCapita          REAL,
  GIIScore              REAL,
  PRIMARY KEY (CountryCode, TimeID, FieldID),
  FOREIGN KEY (CountryCode)    REFERENCES dim_country(CountryCode),
  FOREIGN KEY (TimeID)         REFERENCES dim_time(TimeID),
  FOREIGN KEY (FieldID)        REFERENCES dim_field(FieldID)
);

-- LOAD FACT
INSERT OR IGNORE INTO fact_representation (
```

```
  CountryCode, TimeID, FieldID,
  FemaleEnrolRate, FemaleGradRate,
  GenderGapIndex, HappinessScore,
  InternetUsage, SocialSupport, CorruptionPerception, Freedom,
  LogGDPPerCapita, GIIScore
)
SELECT DISTINCT
  c.CountryCode,
  tt.TimeID,
  f.FieldID,
  ws.female_enrol_rate,
  ws.female_grad_rate,
  ws.gender_gap_index,
  wh.happiness_score,
  iu.internet_usage           AS InternetUsage,
  wh.social_support           AS SocialSupport,
  wh.corruption_perception    AS CorruptionPerception,
  wh.freedom                  AS Freedom,
  wh.log_gdp_per_capita       AS LogGDPPerCapita,
  g.gii_score                 AS GIIScore
FROM stg_women_stem_norm ws
JOIN dim_country c ON c.CountryCode = ws.iso3
JOIN dim_time    tt ON tt.Year      = ws.year
JOIN dim_field   f  ON f.Name       = ws.field
LEFT JOIN stg_gender_inequality_norm g
       ON g.iso3 = ws.iso3      AND g.year      = ws.year
LEFT JOIN stg_world_happiness_norm wh
    ON wh.iso3 = ws.iso3 AND wh.year = ws.year
LEFT JOIN stg_internet_usage_norm iu
    ON iu.iso3 = ws.iso3 AND iu.year = ws.year;
```

Step 4: Quality Assurance and Finalization

After the schema was fully populated, quality assurance checks were conducted to confirm the integrity of the database. We verified that every record in the fact_representation table had valid foreign key links to its corresponding country, year, and STEM field in the dimension tables. Additional checks ensured there were no duplicate country–year combinations and that the year range covered 2005–2021 consistently across all datasets.

Once these validations were complete, we dropped the temporary staging and helper tables (such as raw imports and mapping tables). This step ensured the database contained only the clean star schema tables required for analysis, thereby improving clarity, efficiency, and storage usage.

```
-- Expect only the 6 countries
SELECT COUNT(DISTINCT CountryCode) AS countries_in_dim_country FROM dim_country;

-- Year range present
SELECT MIN(Year) AS min_year, MAX(Year) AS max_year FROM dim_time;

-- Foreign key orphan checks (should all be zero rows)
```

```sql
SELECT 'country_fk_missing' AS issue, COUNT(*) AS n
FROM fact_representation f
LEFT JOIN dim_country d ON d.CountryCode = f.CountryCode
WHERE d.CountryCode IS NULL

UNION ALL
SELECT 'time_fk_missing', COUNT(*)
FROM fact_representation f
LEFT JOIN dim_time t ON t.TimeID = f.TimeID
WHERE t.TimeID IS NULL

UNION ALL
SELECT 'field_fk_missing', COUNT(*)
FROM fact_representation f
LEFT JOIN dim_field x ON x.FieldID = f.FieldID
WHERE x.FieldID IS NULL;

-- Drop raw staging tables
DROP TABLE IF EXISTS stg_women_stem_raw;
DROP TABLE IF EXISTS stg_world_happiness_raw;
DROP TABLE IF EXISTS stg_gender_inequality_raw;
DROP TABLE IF EXISTS stg_internet_usage_raw;

-- Drop intermediate cleaned staging
DROP TABLE IF EXISTS stg_women_stem;
DROP TABLE IF EXISTS stg_world_happiness;
DROP TABLE IF EXISTS stg_internet_usage;
DROP TABLE IF EXISTS stg_gender_inequality_long;

-- Drop normalized staging
DROP TABLE IF EXISTS stg_women_stem_norm;
DROP TABLE IF EXISTS stg_world_happiness_norm;
DROP TABLE IF EXISTS stg_internet_usage_norm;
DROP TABLE IF EXISTS stg_gender_inequality_norm;

-- Drop helper maps
DROP TABLE IF EXISTS country_map;
DROP TABLE IF EXISTS country_whitelist;
```

SQL for Business Analytics

RQ1: How have changes in national internet usage between 2005 and 2021 affected year-on-year trends in female enrolment and graduation rates in STEM fields across Australia, Canada, Germany, and the United States, and to what extent does greater technological access shape women's participation in STEM within these countries?

```
-- 1. The agg CTE (Common Table Expression)

WITH agg AS (
  SELECT CountryCode, TimeID,
         AVG(FemaleEnrolRate) AS avg_enrol, -- Calculating average female enrol rate
         AVG(FemaleGradRate) AS avg_grad -- Calculating average female graduation rate
  FROM fact_representation
  GROUP BY CountryCode, TimeID
) -- This simplifies the dataset so downstream queries only require one row per country per
year instead of one row per STEM Field.

-- 2. Main SELECT Query

SELECT
  dc.Name AS country,
  dt.Year AS year, -- dc.Name / dt.Year extracts the country name and the year from the
dimension tables
ROUND(MAX(s.InternetUsage)*100, 2) AS InternetUsage,
-- Takes the maximum InternetUsage value from dim_score for that country/year.
ROUND (AVG (agg.avg_enrol) *100, 2)   AS AvgFemaleEnrolment,
ROUND (AVG (agg.avg_grad) *100, 2)    AS AvgFemaleGrad -- Multiplies by 100 and rounds to 2
decimal places
FROM agg
JOIN dim_country dc USING (CountryCode) -- Maps CountryCode to Name (India as 'IND')
JOIN dim_time    dt USING (TimeID) -- maps TimeID to Year (Year is 2005)
LEFT JOIN fact_representation s
       ON s.CountryCode = dc.CountryCode
      AND s.TimeID      = dt.TimeID  -- LEFT JOIN ensures missing rows won't disrupt the
function.
GROUP BY dc.CountryCode, dt.Year -- ensures one row per (Country,Year) in the final table.
ORDER BY country, year; -- arranges by year sequentially
```

RQ2: Between 2005 and 2021, how do year-on-year changes in gender inequality (GII) relate to the growth rates of female enrolment and graduation in STEM across Australia, Canada, China, Germany, India, and the USA?

```sql
-- 1. Calculate avg GII score, avg Female Enrol and Grad proportions per country per year
DROP TABLE IF EXISTS agg2;
CREATE TEMP TABLE agg2 AS
SELECT c.Name AS CountryName, --Take the Name in the dim_country
    t.Year AS 'Year', -- Take the Year in the dim_time
    AVG(f.GIIScore) AS GIIScore, -- Calculate average GIIScore for each country in each year
    AVG(f.FemaleEnrolRate) AS FemaleEnrolProp, -- Avg Female Enrol Proportion for each
country in each year
    AVG(f.FemaleGradRate) AS FemaleGradProp -- Avg Female Grad Proportion for each country in
each year
FROM fact_representation AS f -- Extract data from fact_representation
JOIN dim_country AS c -- Join with dim_country to take the Name
ON c.CountryCode = f.CountryCode
JOIN dim_time AS t -- Join with dim_time to take the Time
    ON t.TimeID = f.TimeID
WHERE t.Year BETWEEN 2005 AND 2021 -- Take years from 2005 to 2021
    AND c.Name IN ('Australia', 'Canada', 'China', 'Germany', 'India', 'United States') --
Focus on 6 target countries
GROUP BY c.Name, t.Year; -- Group by country and year

-- 2. Compute the differences in GII Score, Female Enrol Prop, and Female Grad Prop between
year t and year t-1
DROP TABLE IF EXISTS yoy;
CREATE TEMP TABLE yoy AS
SELECT CountryName, Year, GIIScore,
                GIIScore - LAG(GIIScore)
             OVER (PARTITION BY CountryName ORDER BY Year) AS DeltaGenderInequalityIndex, --
pulls the previous year'ss GII within the same country
         100.0 * (FemaleEnrolProp - LAG(FemaleEnrolProp)
             OVER (PARTITION BY CountryName ORDER BY Year)) AS DeltaFemaleEnrolProp, --
Subtract last year's enrol value to get the delta in proportion, then multiply by 100 to
convert to percentage
         100.0 * (FemaleGradProp - LAG(FemaleGradProp)
             OVER (PARTITION BY CountryName ORDER BY Year)) AS DeltaFemaleGradProp --
Subtract last year's grad value to get the delta in proportion, then multiply by 100 to
convert to percentage
FROM agg2;

-- 3. Remove NULL values
DROP TABLE IF EXISTS cleaned_yoy;
CREATE TEMP TABLE cleaned_yoy AS
SELECT *
FROM yoy
WHERE DeltaGenderInequalityIndex IS NOT NULL
  AND DeltaFemaleEnrolProp IS NOT NULL
  AND DeltaFemaleGradProp IS NOT NULL;
```

```sql
-- 4. Calculate Enrol Statistics
DROP TABLE IF EXISTS stats_enrol;
CREATE TEMP TABLE stats_enrol AS
SELECT CountryName,
  COUNT(*) AS n,
  SUM(DeltaGenderInequalityIndex) AS sumdgii, --Adds up all the yearly ΔGII values (year-on-year changes in GII) for that country
  SUM(DeltaFemaleEnrolProp) AS sumdenrol, -- Adds up all the yearly changes in female STEM enrolment proportion for that country.
  SUM((DeltaGenderInequalityIndex)*(DeltaGenderInequalityIndex)) AS sumdgii_sumdgii, -- Squares each year's ΔGII value and then sums them for variance
  SUM((DeltaFemaleEnrolProp)*(DeltaFemaleEnrolProp)) AS sumdenrol_sumdenrol, -- Squares each year's enrolment change and then sums them for variance
  SUM((DeltaGenderInequalityIndex)*(DeltaFemaleEnrolProp)) AS sumdgii_sumdenrol -- Calculate covariance numerator to show how changes in GII and enrolment
FROM cleaned_yoy
GROUP BY CountryName;

-- 5. Calculate Grad
DROP TABLE IF EXISTS stats_grad;
CREATE TEMP TABLE stats_grad AS
SELECT CountryName,
  COUNT(*) AS n,
  SUM(DeltaGenderInequalityIndex) AS sumdgii,
  SUM(DeltaFemaleGradProp) AS sumdgrad, -- Adds up all the yearly changes in female STEM graduation proportion for that country.
  SUM((DeltaGenderInequalityIndex)*(DeltaGenderInequalityIndex)) AS sumdgii_sumdgii,
  SUM((DeltaFemaleGradProp)*(DeltaFemaleGradProp)) AS sumdgrad_sumdgrad, -- Squares each year's graduation change and then sums them for variance
  SUM((DeltaGenderInequalityIndex)*(DeltaFemaleGradProp)) AS sumdgii_sumdgrad --Calculate covariance numerator to show how changes in GII and graduation
FROM cleaned_yoy
GROUP BY CountryName;

-- 6. Calculate correlation and slope
SELECT e.CountryName AS "Country Name",
  ROUND((e.n*e.sumdgii_sumdenrol - e.sumdgii*e.sumdenrol) /
        NULLIF(SQRT((e.n*e.sumdgii_sumdgii - e.sumdgii*e.sumdgii) *
(e.n*e.sumdenrol_sumdenrol - e.sumdenrol*e.sumdenrol)), 0), 3)
        AS "Correlation ΔGII vs ΔFemale Enrolment", -- Correlation for Female Enrolment
  ROUND((g.n*g.sumdgii_sumdgrad - g.sumdgii*g.sumdgrad) /
        NULLIF(SQRT((g.n*g.sumdgii_sumdgii - g.sumdgii*g.sumdgii) * (g.n*g.sumdgrad_sumdgrad - g.sumdgrad*g.sumdgrad)), 0), 3)
        AS "Correlation ΔGII vs ΔFemale Graduation" -- Correlation for Female Graduation
FROM stats_enrol e
JOIN stats_grad  g ON g.CountryName= e.CountryName
ORDER BY e.CountryName;
```

RQ3: What is the relationship between freedom of expression and female STEM enrolment rate in Asia and America over time? What are the differences between the two continents?

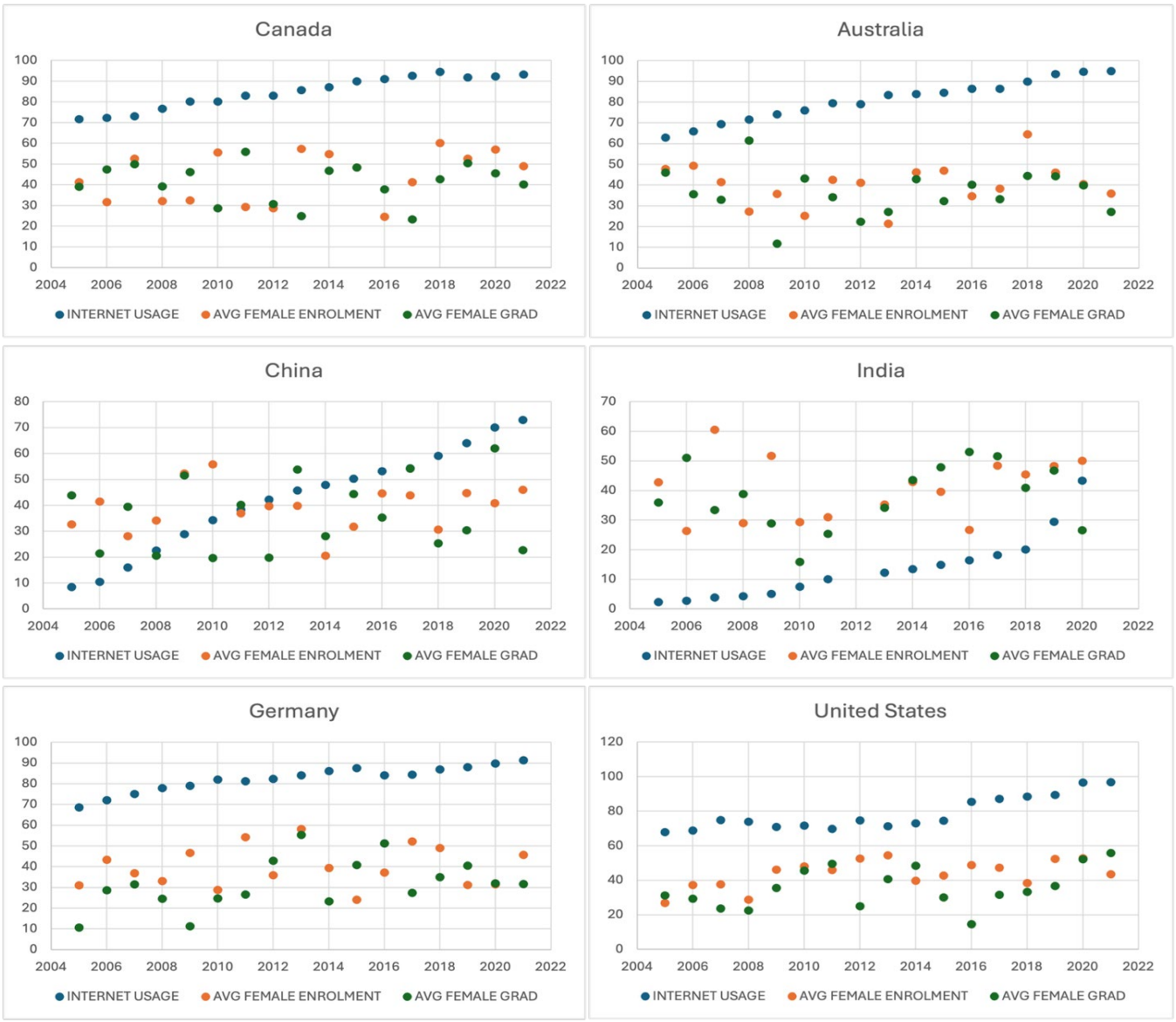
```
SELECT dc.Continent Continent, dt.Year Year, f.Freedom, ROUND(AVG(f.FemaleEnrolRate),2)
AvgFemaleEnrolment -- Select the relevant attributes to the question
FROM fact_representation f -- from fact_representation, joined with dim_country and dim_time
JOIN dim_country dc USING(CountryCode)
JOIN dim_time dt USING(TimeID)
WHERE UPPER (dc.Continent) IN ('AMERICA','ASIA') -- filtering by continent name (America and
Asia),
  AND f.Freedom IS NOT NULL -- exclude any missing data
GROUP BY Continent, Year -- grouping the results by Continent and Year
ORDER BY Continent, Year; --ordering by year in chronological order.
```

Insights and Future Work

Insights to RQ1: To what extent does digital access (internet usage) shape women's opportunities in STEM?

Figure 2.

Distribution of Internet Usage, Average Female Enrolment and Average Female Graduation Rate in Six Countries From 2005 to 2021



In this section, we discuss the results of our analysis to identify global drivers of female STEM representation. Firstly, we consider the technological dimension. As shown in Figure 2, internet access has been widespread in developed countries such as Germany, Canada, Australia, and the United States since 2005. While connectivity continues to expand, female STEM enrolment and graduation rates have not risen in parallel. This indicates that cultural norms, policy frameworks, and institutional support may play a more decisive role than digital access alone.

By contrast, emerging economies such as India and China began the period with relatively low levels of internet penetration but have experienced steady gains in both connectivity and female participation in STEM. This suggests that improved digital access can be especially influential in contexts where connectivity has historically been limited.

Digital access, therefore, appears to be a necessary but not sufficient condition for expanding women's opportunities in STEM. It facilitates participation when access is scarce, but once high penetration is achieved, further progress depends on addressing systemic barriers such as gender bias, institutional constraints, and cultural expectations. The significance to note is that the findings are reported nationally and may mask localized disparities and broader socioeconomic influences.

Insights to RQ2: Between 2005 and 2021, how do year-on-year changes in gender inequality (GII) relate to the growth rates of female enrolment and graduation in STEM across Australia, Canada, China, Germany, India, and the USA?

Table 1

Correlation and Slope Estimates Between Year-on-Year Changes in Gender Inequality (ΔGII) and Female STEM Enrolment/Graduation (2005–2021)

Country Name	Correlation ΔGII vs ΔFemale Enrolment	Correlation ΔGII vs ΔFemale Graduation
Australia	0.022	0.151
Canada	0.001	-0.145
China	0.286	-0.316
Germany	-0.212	-0.169
India	-0.184	-0.287
United States	0.513	-0.307

Table 1 highlights how changes in gender inequality (ΔGII) relate to changes in female STEM enrolment and graduation across six countries between 2005 and 2021. The results show that these year-on-year relationships are weak and vary considerably by context. For enrolment, Canada and Australia exhibit almost no correlation. At the same time, the United States and China show moderate positive associations, suggesting that inequality changes may coincide with increases in female enrolment in specific settings. In contrast, Germany and India display negative associations, where worsening inequality is linked to stagnation or decline in enrolment. On the other hand, graduation rates show a more consistent pattern: most countries (China, Germany, India, and the United States) exhibit negative correlations, indicating that rising inequality tends to coincide with declining graduation outcomes. However, the strength of these correlations is modest, indicating that ΔGII independently is not a dominant driver of STEM outcomes. This suggests that while improvements in gender inequality may create enabling conditions for female participation, they are insufficient. Cultural norms, institutional practices, and policy interventions still play a critical role in ensuring that increasing equality can lead to tangible improvements in STEM enrolment and graduation.

Insights to RQ3: What is the relationship between freedom of expression and female STEM enrolment rate in Asia and America over time? What are the differences between the two continents?

Figure 3.

Freedom of Expression Index and Female Enrolment Rate Progression in America and Asia from 2008 to 2021

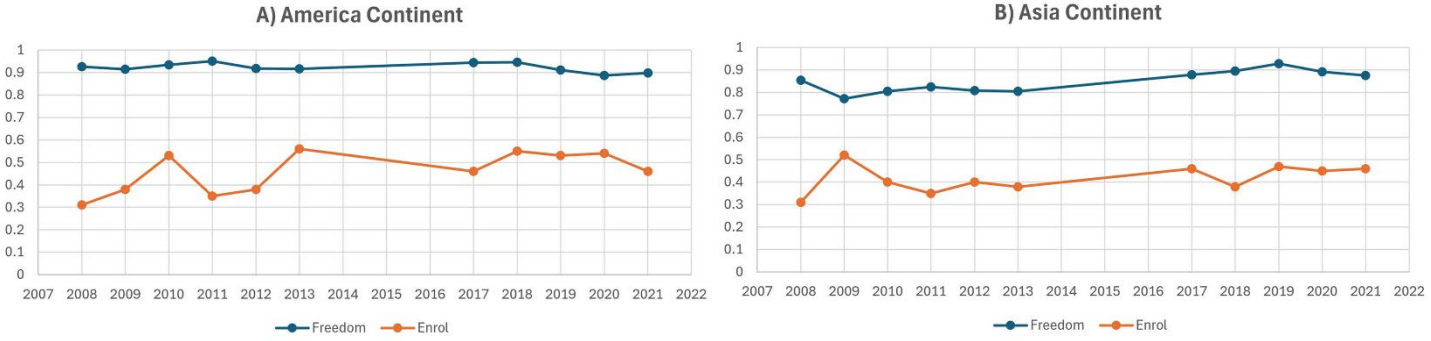


Figure 3 indicates that the American continent, with a higher baseline freedom of expression, has a stable female STEM enrolment rate from 2008 to 2021. In contrast, the Asia continent exhibited rising freedom of expression alongside an increasing female STEM enrolment rate in the same period. This suggests a positive correlation exists between the two dimensions, and that it is noteworthy for developing countries to improve their perceived freedom of expression through policymaking to encourage women's participation in the exact sciences.

Limitations

Despite these findings, this analysis is nonetheless subject to limitations. Namely, the world_happiness dataset lacked reported metrics for certain countries over several years, leading to omissions during the analytical process. Additionally, the women_stem dataset records enrolment rate but not the actual volume of students, thus presenting doubt as to whether one enrolment rate may be directly comparable to another. In this same dataset, the absence of the level of education (bachelor, master, or doctorate) precluded the element of barriers to entry and length of study. At the same time, the graduation rate metric carries an implicit time-lag that may not be readily applicable in a streamlined analysis.

Future Work

Future analysis should incorporate more socioeconomic factors, such as gender income disparity, national wealth metrics, and global education ranking, to increase dimensional complexity, while utilizing statistical methods to empirically evaluate the relationship between factors and the central variable of women STEM representation. Such factors may better capture the complex drivers behind women’s progress in STEM education and provide stronger evidence for more naunced evidence-based policy recommendations.

Appendix

Data Source

Dataset 1
Source name: Women's Representation in Global STEM Education
Alias: women_stem
Source type: CSV
Source link: https://www.kaggle.com/datasets/bismasajjad/womens-representation-in-global-stem-education

Dataset 2
Source name: Gender Inequality Index by Country
Alias: gender_inequality
Source type: CSV
Source link: https://www.kaggle.com/datasets/iamsouravbanerjee/gender-inequality-index-dataset

Dataset 3
Source name: World Happiness Report
Alias: world_happiness
Source type: CSV
Source link: https://www.kaggle.com/datasets/jainaru/world-happiness-report-2024-yearly-updated/data?select=World-happiness-report-updated_2024.csv

Dataset 4
Source name: Share of the population using the Internet
Alias: internet_usage
Source type: CSV
Source link: https://www.kaggle.com/datasets/chaudharisanika/share-of-the-population-using-the-internet?resource=download

Schema and Attributes

Fact: fact_representation

Name	Key	Unit/Format	Source
CountryCode	PK, FK	Text (3 characters)	women_stem
TimeID	PK, FK	Date (YYYY)	women_stem
StemFieldID	PK, FK	Integer	women_stem
FemaleEnrolRate		Real (0-1)	women_stem
FemaleGradRate		Real (0-1)	women_stem
GenderGapIndex		Real (0-1)	women_stem
HappinessScore		Real (0-1)	world_happiness
InternetUsage		Real (0-1)	world_happiness
SocialSupport		Real (0-1)	world_happiness
CorruptionPerception		Real (0-1)	world_happiness
Freedom		Real (0-1)	world_happiness
LogGDPPerCapita		Real (0-1)	world_happiness
GIIScore		Real (0-1)	gender_inequality

Dim1: dim_country

Name	Key	Unit/Format	Source
CountryCode	PK	Text (3 characters)	gender_inequality
Name		Text	gender_inequality
Continent		Text	gender_inequality
HumanDevZone		Text	gender_inequality
UNDPDevRegion		Text	gender_inequality

Dim2: dim_time

Name	Key	Unit/Format	Source
TimeID	PK	Integer	
Year		Date (YYYY)	women_stem

Dim3: dim_field

Name	Key	Unit/Format	Source
StemFieldID	PK	Integer	
Name		Text	women_stem