**Task 1**

**Title: Multivariate Forecasting of Scotland’s Monthly Birth Rates**

**Datascience Internship (TEAM BETA)**

**1. Introduction**

Accurate forecasting of birth rates is crucial for healthcare planning, educational infrastructure, social

services, and economic policy. While univariate models based solely on historical birth counts can

capture trends and seasonality, they often miss important drivers—such as economic conditions,

public health indicators, and sociocultural factors—that influence family-planning decisions. This

proposal outlines a multivariate time-series forecasting approach that integrates additional

predictors to enhance the accuracy and interpretability of birth-rate forecasts for Scotland.

**2. Problem Statement**

Existing forecasts for Scotland’s monthly birth volumes rely primarily on past birth data (univariate

models). However, fertility behavior is also shaped by:

• Economic indicators (unemployment rate, consumer price index)

• Health factors (maternal access to prenatal care, average maternal age)

• Education and socio-demographic variables (female tertiary-education enrollment, average

household income)

Gap: Without these covariates, forecasts may under-react to sudden economic shocks (e.g.,

recessions) or public-health events (e.g., pandemics).

Goal: Develop a multivariate forecasting framework that leverages multiple data sources to produce

more robust, real-time birth-rate predictions.

**Step 1: Collection of datasets**

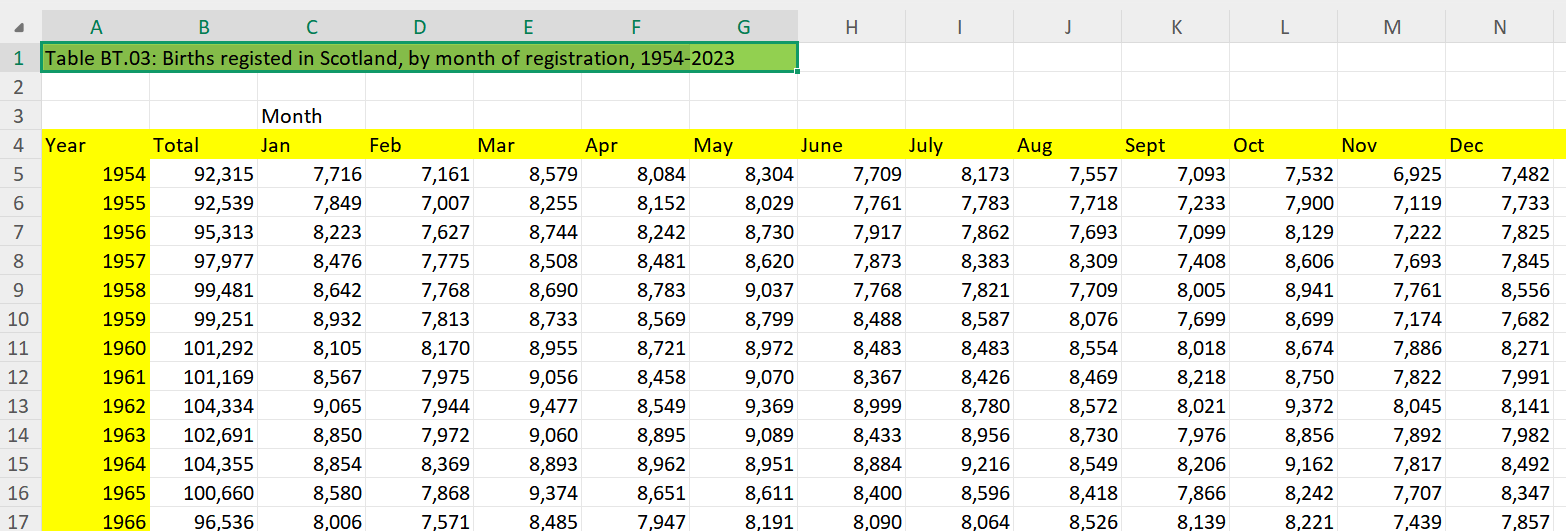
To construct the multivariate time-series forecasting model for Scotland’s monthly birth volumes, relevant datasets were identified, sourced, and documented. Each dataset was selected based on its potential explanatory power for fertility trends.

### ****1.1 Birth Rate Data****

**Source**: National Records of Scotland (NRS) / Public Health Scotland (PHS)

**Coverage**: Monthly data, 1998–2022

**Variables**: csv file named monthly-brth-registrations contains the births registered in Scotland



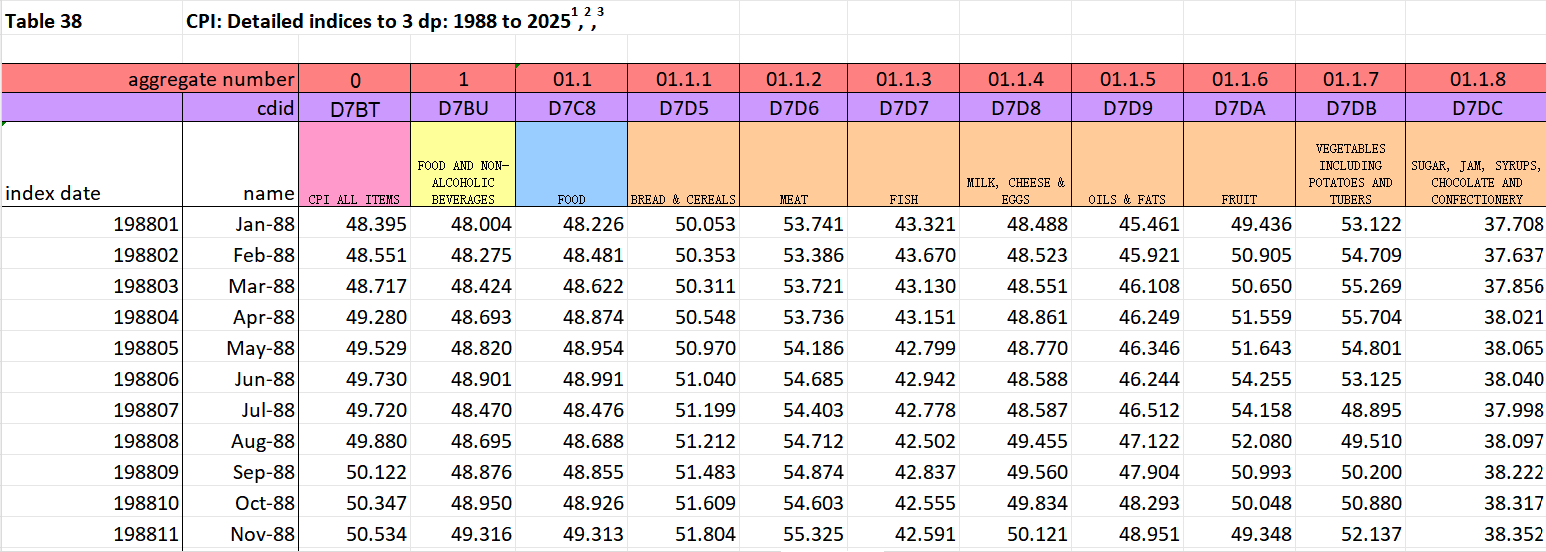
**Link: [BT.3 2023 Births Time Series Csv](https://www.nrscotland.gov.uk/media/hlvnkq1l/bt3-2023-births-time-series-csv.csv) monthly birth registrations (nrs=national records of scotland)**

### ****1.2 Consumer Price Index (CPI)****

**Source**: Office for National Statistics (ONS) / OECD / IMF

**Coverage**: Monthly, 1998–2022

Link: [Consumer price inflation tables - Office for National Statistics](https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/consumerpriceinflation/current)

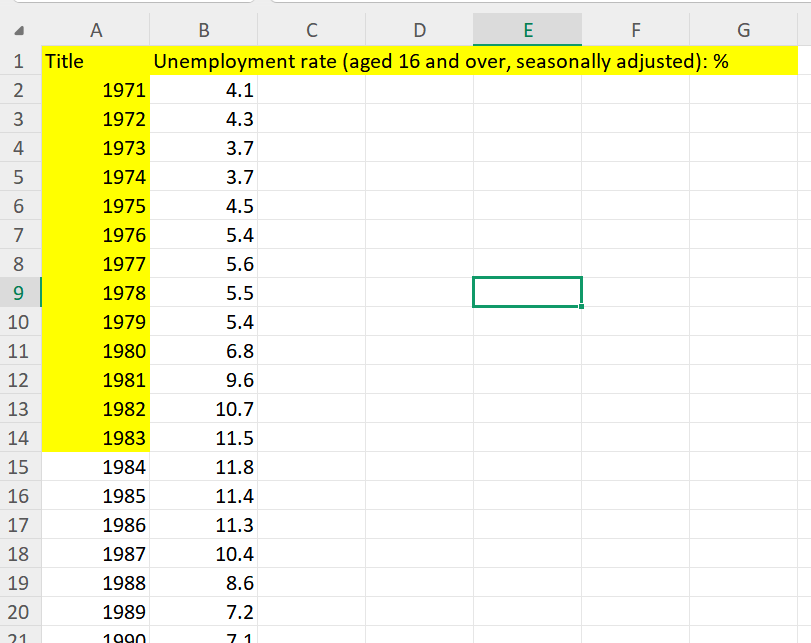


### ****1.3 Unemployment Rate (CPI)****

**Source**: Office for National Statistics (ONS)

**Coverage**: Monthly, 1998–2022

Link: **<https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/timeseries/mgsx/lms>**



### ****1.4 M****aternal health indicators

The SMR02 dataset captures comprehensive maternity records for all births in Scotland, including mother and child characteristics, pregnancy risk factors, and birth outcomes. It is the primary source for monitoring maternal health indicators across NHS boards.

For this study, annual aggregated data (1998–2022) was extracted from **Tables 1–8** of the SMR02 publication.

**Source**: Births in Scotland SMR02

**Coverage**: Monthly, 1998–2022

Link: [Births in Scotland - Year ending 31 March 2023 - Births in Scotland - Publications - Public Health Scotland](https://publichealthscotland.scot/publications/births-in-scotland/births-in-scotland-year-ending-31-march-2023/)

**Data files :**

**Table 1: maternities xlx**

**Indicator:** (Total no of maternity per year)

**Table 2: live births xlx**

**Indicator:** (total live births per year)

**Table 3: BMI xlx**

**Indicator:** (obeseity % per year)

**Table 4: diabetes xlx**

**Indicator:** (diabetes% per year)

**Table 5: induction xlx**

**Indicator:** (induction%per year)

**Table 6: method xlx**

**Indicator:** (method of birth cesearian normal+emergency % per year)11.2+15.3=26.5

**Table 7: gestation xlx**

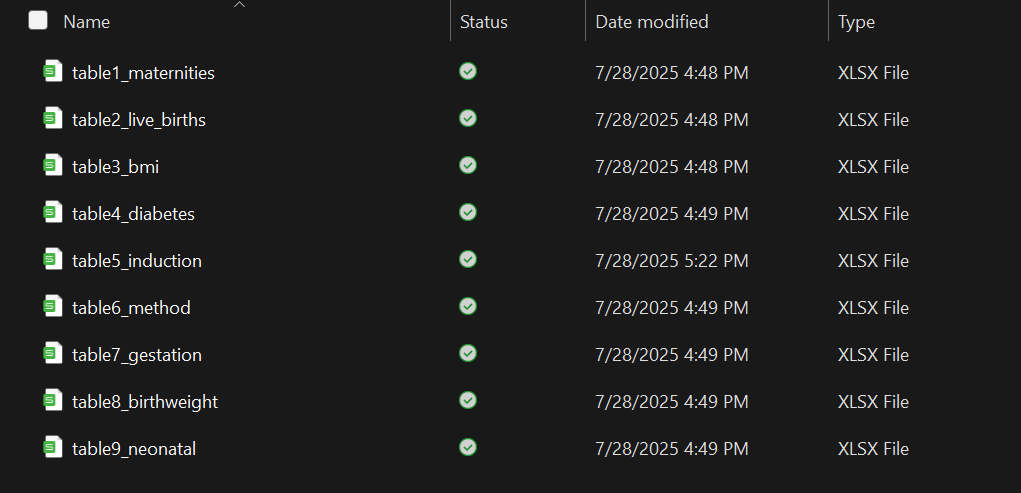
**Indicator:** (<37 =premature, >37 weeks= full) so we will write their percentages

**Table 8: birthweight xlx**

**Indicator:** (3columns)SGA, AGA, LGA

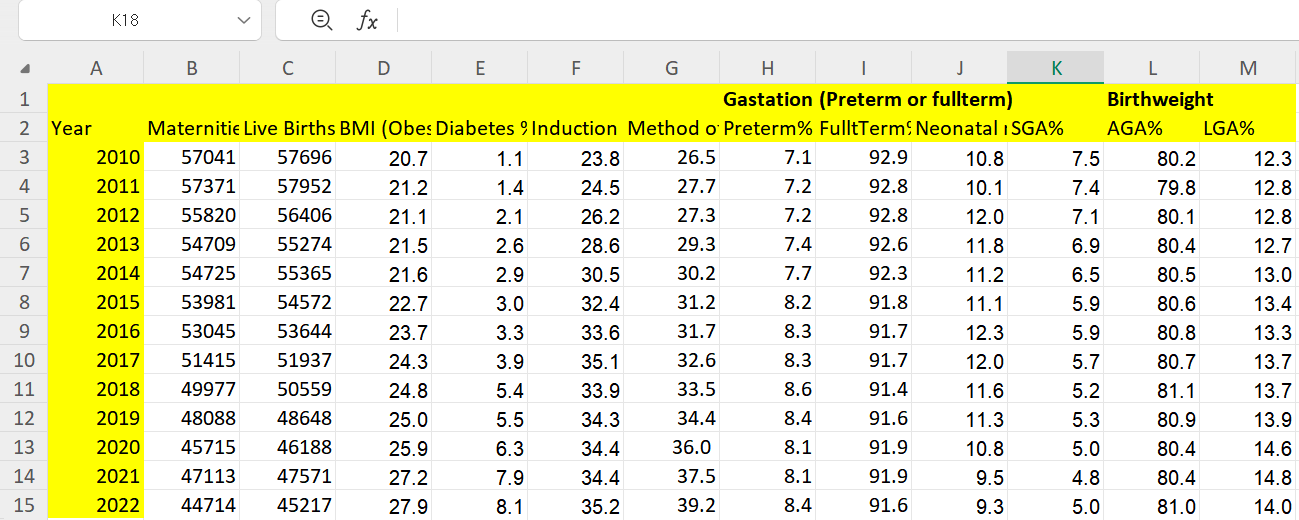
**SGA (Small for Gestational Age)**: Below the 10th percentile for weight at gestational age.

**AGA (Appropriate for Gestational Age)**: 10th–90th percentile.**LGA (Large for Gestational Age)**: Above the 90th percentile.Percentages indicate fetal growth health and are linked to maternal nutrition and health.



**Table 9: neonatal mortality xlx**

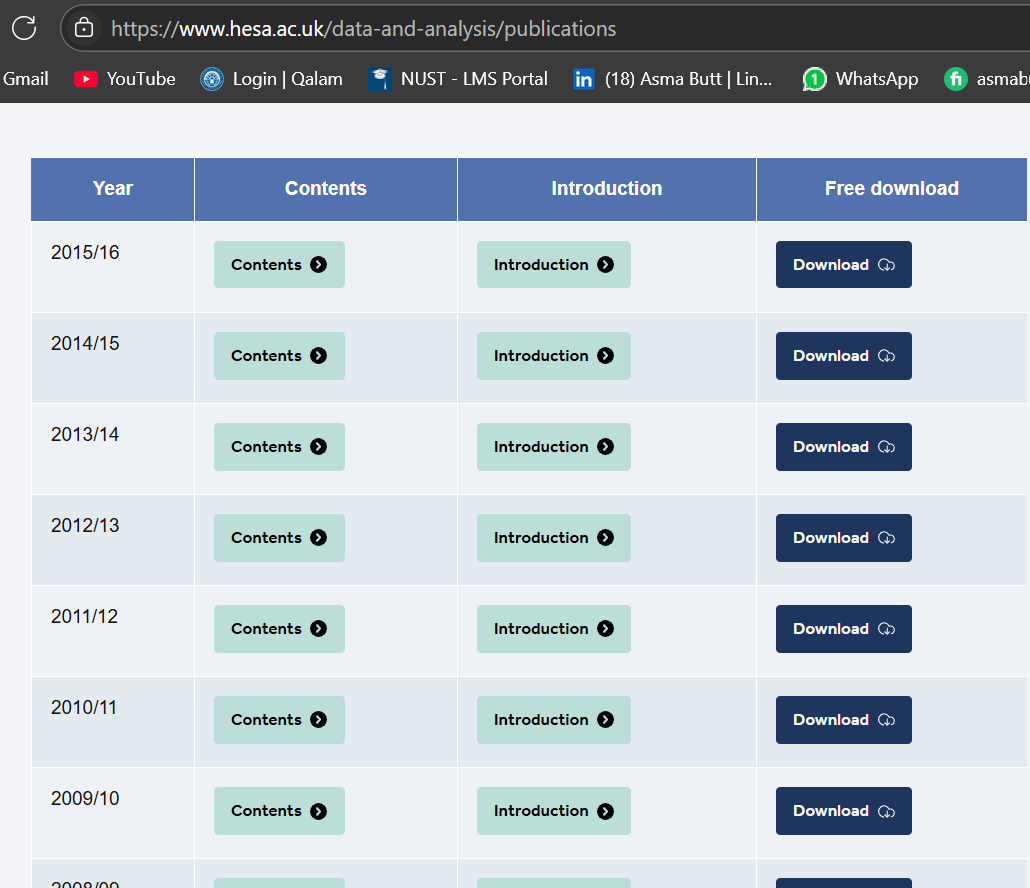
We have downloaded these tables for maternal health indicators table



**Summary of Use:**  
These eight indicators—maternities, live births, BMI obesity rates, diabetes prevalence, induction rates, caesarean rates, preterm/full-term percentages, and birthweight categories—are combined into a **Maternal Health Indicator dataset** for Scotland (2010–2022). This serves as a health-related explanatory variable in the multivariate birth-rate forecasting model.

### ****1.5 Female Tertiary Enrollment****

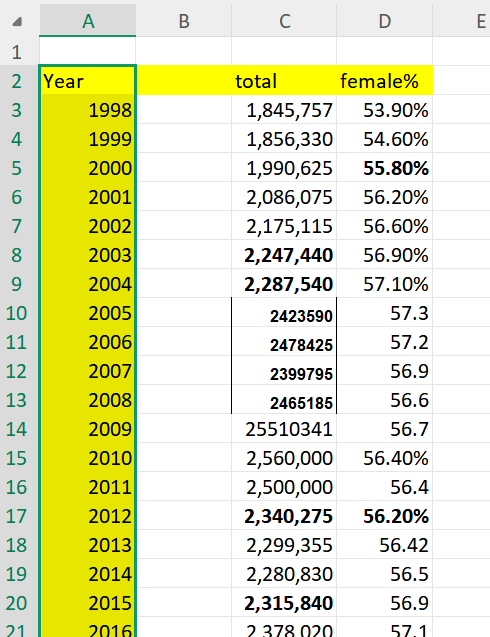
**Source**: Higher Education Statistics Agency (HESA)



**Coverage**: Annual 1998 – 2022

**Variables:**

1. Female Tertiary Enrollment (%) – % of female population enrolled in higher education
2. Total Female Students – Number of female students enrolled



Link: [Consumer price inflation tables - Office for National Statistics](https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/consumerpriceinflation/current)

Data was found from HESA archives and was documented manually

**Step 2: Merging Datasets :**

In this step, all datasets from different sources were transformed into a **consistent format** and then merged into one comprehensive dataset.

**1.Monthly Birth Data**

* The monthly birth registrations file was read using pandas.read\_csv().
* The dataset, which was initially in **wide format** (columns for each month), was converted into **long format** using pandas.melt() so that each row contains a Year, Month, and Births value.
* The months were ordered chronologically using pd.Categorical() to maintain the correct sequence.

**2.Monthly CPI (Consumer Price Index):**

* The CPI dataset was read from Excel using pandas.read\_excel().
* The year column was standardized:
* If stored as a datetime, the month and year were extracted directly.
* If stored as text (e.g., "Jan 98"), the month abbreviation and two-digit year were parsed. Two-digit years were converted to four-digit years using a cutoff (values above 50 treated as 19xx, otherwise 20xx).
* A month mapping dictionary ensured that CPI month names matched the birth dataset format.

**3.Annual Datasets (Unemployment, Maternal Health, Education)**

* Three annual datasets (unemployment rate, maternal health indicators, and female tertiary enrollment) were loaded.
* The Year column in each dataset was cleaned to ensure a consistent **four-digit integer format**.

**4.Data Type Consistency**

* All Year columns across datasets were converted to integers to avoid merge errors caused by mismatched types (string vs. int).

**5.Dataset Merging**

* First, **monthly birth data** was merged with **monthly CPI data** using Year and Month as keys.
* Next, **annual datasets** (unemployment rate, maternal health, education) were merged with the monthly dataset on Year using a **left join** to retain all birth-month records.

**6.Final Output**

* The combined dataset was saved as final\_merged\_dataset.csv for further analysis.