# Synthetic Data Creation

## Expectations

I wanted to create data bases on a statistical method. I decided to make the Units Sold and the Price Sold variable with the inflation rate. As such, I do know that Inflation Rate should affect the Units Sold either positively or negatively. If I had left it that way, there should be an R Squared of 1 on the data for the prediction on Inflation Rate/Store ID.

I decided to force a difference based on a little bit of randomisation. The Units Sold were later randomised slightly by -3 to 3, to imitate a little bit of free will.

## Method

This data set evolved over 2 days.

I started by typing out my requirements into a single sheet and pasting it in. The problems there were that the Analysis crashed most of the time, or that when it did succeed, the dataset output did not include all of the specifications.

Eventually, I took the approach of building the data set up statement by statement. This was quicker to analyse and rarely failed. I exported the values in csv at every stage in case the connection reset. That made it possible to upload the data set and continue from that point.

Chat-GPT4 kept changing my numbers. There were times when I thought it was resolved, but Chat-GPT recalculated my Price Sold data, making prices into the 10s of thousands.

## Data Creation via Chat-GPT 4

User

**You**

I need to create some data. I need 45 unique Product IDs. 1 to 15 are for Product Category [‘Mens’] and 16 to 45 are for Product Category [‘Womens’]’ Product IDs must be numeric. I need 3 unique Store IDs for each of these Countries ['UK', 'Ireland', 'France']. The store ID should be numeric. So the Store IDs 1- 3 would be Ireland, 4-6 would be UK, 7-9 France, etc... I would also like to include Inflation Rates by country by month and GDP Growth Rate calculations by country by month. These can be set with random decimal numbers between -1 and 5. Inflation Rate and GDP Growth Rate should be consistent for each country and month combination and be displayed on each row. GDP Growth Rate can be randomised between -2 and 4 and should be consistent for each Month and Country. It should be rounded to 2 decimal points. We'll need Monthly Data from 2023-01 to 2024-06

ChatGPT

**ChatGPT**

I've created a data set with the following components:

1. **Product IDs and Categories**:
   * 45 unique Product IDs: 1 to 15 for "Mens" and 16 to 45 for "Womens".
2. **Store IDs and Countries**:
   * 9 unique Store IDs:
     + 1 to 3 for Ireland
     + 4 to 6 for the UK
     + 7 to 9 for France
3. **Monthly Data**:
   * Covers from January 2023 to June 2024.
   * Each entry includes:
     + **Inflation Rate**: Random values between -1 and 5, consistent per country and month.
     + **GDP Growth Rate**: Random values between -2 and 4, consistent per country and month, rounded to two decimal places.

The data structure combines these elements so that each product at each store is paired with the monthly economic indicators for each country. Here are the first 10 rows of this data set for demonstration:

| **Product ID** | **Category** | **Store ID** | **Date** | **Country** | **Inflation Rate** | **GDP Growth Rate** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-02-01 | Ireland | 2.70 | 2.00 |
| 1 | Mens | 1 | 2023-03-01 | Ireland | 4.66 | -1.21 |
| 1 | Mens | 1 | 2023-04-01 | Ireland | 3.09 | 2.30 |
| 1 | Mens | 1 | 2023-05-01 | Ireland | 1.16 | -0.26 |
| 1 | Mens | 1 | 2023-06-01 | Ireland | 1.62 | -0.90 |
| 1 | Mens | 1 | 2023-07-01 | Ireland | 3.19 | 1.52 |
| 1 | Mens | 1 | 2023-08-01 | Ireland | -0.64 | -1.88 |
| 1 | Mens | 1 | 2023-09-01 | Ireland | 3.00 | 2.97 |
| 1 | Mens | 1 | 2023-10-01 | Ireland | 3.02 | -1.97 |

Each product at each store for each country and month has a specific set of economic indicators assigned to it, ensuring that the data reflects both product-specific and economic elements. If you need further customizations or additional data, feel free to let me know! ​​

User

**You**

add a datepoint for every date in the month and name that 'Date of Sale'

ChatGPT

**ChatGPT**

It looks like I encountered an error while trying to expand the time-series data to include a "Date of Sale" for every day within the specified month. I'll correct this and generate the data correctly. Let me adjust the code and try again. ​​

The corrected data now includes a "Date of Sale" for every day within the specified months, from January 2023 to June 2024. Each product and store ID combination is now paired with daily sales dates, along with the consistent monthly economic indicators for each country. Here are the first 10 rows showing this structure:

| **Product ID** | **Category** | **Store ID** | **Date** | **Date of Sale** | **Country** | **Inflation Rate** | **GDP Growth Rate** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-01 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-02 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-03 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-04 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-05 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-06 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-07 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-08 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-09 | Ireland | 2.67 | -0.09 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-10 | Ireland | 2.67 | -0.09 |

Each row now uniquely identifies a sale date for a specific product and store, along with the inflation rate and GDP growth rate for the month, all categorized by country. If you need this data in a specific format or more details added, just let me know! ​​

User

**You**

add a column for ['Price Sold']. Start this as a random number between 20 and 40 on the first date. Round it to 2 decimal places.

ChatGPT

**ChatGPT**

The data now includes a "Price Sold" column. I started the price for each product as a random number between 20 and 40, rounded to two decimal places, on the first date. This price is consistent across all dates for the same product and store combination. Here are the first 10 rows showing this:

| **Product ID** | **Category** | **Store ID** | **Date** | **Date of Sale** | **Country** | **Inflation Rate** | **GDP Growth Rate** | **Price Sold** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-01 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-02 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-03 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-04 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-05 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-06 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-07 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-08 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-09 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-10 | Ireland | 2.67 | -0.09 | 28.34 |

This setup ensures that each product has a specific starting price that remains constant throughout all recorded dates. If you require the price to vary over time or have any other adjustments in mind, please let me know! ​​

User

**You**

download to csv

ChatGPT

**ChatGPT**

The data has been saved to a CSV file. You can download it from the following link:

Download Product Sales Data CSV ​​

User

**You**

At every change in the month, set the ['Price Sold'] to be (['Price Sold'] \*( [Inflation Rate]/100) )+ [Price Sold]

ChatGPT

**ChatGPT**

The "Price Sold" for each product has been adjusted to reflect changes due to inflation at the start of each new month. Here are the first 10 rows with the initial adjustments:

| **Product ID** | **Category** | **Store ID** | **Date** | **Date of Sale** | **Country** | **Inflation Rate** | **GDP Growth Rate** | **Price Sold** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-01 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-02 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-03 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-04 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-05 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-06 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-07 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-08 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-09 | Ireland | 2.67 | -0.09 | 28.34 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-10 | Ireland | 2.67 | -0.09 | 28.34 |

The price for each product remains the same within the same month but will adjust with the inflation rate at the start of each new month. If you need this updated data saved as a CSV, or if there are any other changes or details to include, please let me know! ​​

User

**You**

download csv

ChatGPT

**ChatGPT**

The updated data has been saved to a CSV file. You can download it from the following link:

Download Updated Product Sales Data CSV ​​

User

**You**

Now we need to add a ['Units Sold'] column. This will start as a random integer between 10 and 30 on the first date. It will be the same for every date in the first month.

ChatGPT

**ChatGPT**

The "Units Sold" column has been successfully added to the dataset. The number of units sold starts as a random integer between 10 and 30 on the first date of each month and remains the same for every date in that month. Here are the first 10 rows with the "Units Sold" column:

| **Product ID** | **Category** | **Store ID** | **Date** | **Date of Sale** | **Country** | **Inflation Rate** | **GDP Growth Rate** | **Price Sold** | **Units Sold** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-01 | Ireland | 2.67 | -0.09 | 28.34 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-02 | Ireland | 2.67 | -0.09 | 28.34 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-03 | Ireland | 2.67 | -0.09 | 28.34 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-04 | Ireland | 2.67 | -0.09 | 28.34 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-05 | Ireland | 2.67 | -0.09 | 28.34 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-06 | Ireland | 2.67 | -0.09 | 28.34 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-07 | Ireland | 2.67 | -0.09 | 28.34 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-08 | Ireland | 2.67 | -0.09 | 28.34 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-09 | Ireland | 2.67 | -0.09 | 28.34 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-10 | Ireland | 2.67 | -0.09 | 28.34 | 18 |

If you need to save this updated data to a CSV file or make further adjustments, please let me know! ​​

User

**You**

Correction . Units Sold starts as a random integer between 10 and 40. The random integer is set on the first date for every product and store combination. The Units Sold remains the same for every date in that month for the store and product combination.

ChatGPT

**ChatGPT**

The "Units Sold" column has been updated to start as a random integer between 10 and 40 for each product and store combination on the first date of every month. This value remains consistent for all dates within that month for the specified store and product combination. Here are the first 10 rows with the corrected "Units Sold" column:

| **Product ID** | **Category** | **Store ID** | **Date** | **Date of Sale** | **Country** | **Inflation Rate** | **GDP Growth Rate** | **Price Sold** | **Units Sold** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-01 | Ireland | 2.67 | -0.09 | 28.34 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-02 | Ireland | 2.67 | -0.09 | 28.34 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-03 | Ireland | 2.67 | -0.09 | 28.34 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-04 | Ireland | 2.67 | -0.09 | 28.34 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-05 | Ireland | 2.67 | -0.09 | 28.34 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-06 | Ireland | 2.67 | -0.09 | 28.34 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-07 | Ireland | 2.67 | -0.09 | 28.34 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-08 | Ireland | 2.67 | -0.09 | 28.34 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-09 | Ireland | 2.67 | -0.09 | 28.34 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-10 | Ireland | 2.67 | -0.09 | 28.34 | 20 |

This setup ensures each combination maintains a consistent "Units Sold" amount for the entirety of each month. If you need this data saved as a CSV or any further modifications, just let me know! ​​

User

**You**

download to csv

ChatGPT

**ChatGPT**

The data has been updated and saved to a CSV file. You can download it from the following link:

Download Final Updated Product Sales Data CSV ​​

User

**You**

for every row, set the Price Sold to round to 2 decimal places and the nearest .5

ChatGPT

**ChatGPT**

I've updated the "Price Sold" for every row to round to the nearest 0.5. Here's how the first 10 rows look now with the rounded prices:

| **Product ID** | **Category** | **Store ID** | **Date** | **Date of Sale** | **Country** | **Inflation Rate** | **GDP Growth Rate** | **Price Sold** | **Units Sold** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-01 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-02 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-03 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-04 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-05 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-06 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-07 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-08 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-09 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-10 | Ireland | 2.67 | -0.09 | 28.5 | 20 |

This ensures each "Price Sold" is now rounded and suitable for display or calculation. If you need to download this updated data to a CSV, just let me know! ​​

User

**You**

update the Units Sold field. At every change of the month, set the ['Units Sold'] to be (['Units Sold'] \*-1\*( [Inflation Rate]/100) )+ [Units Sold]

ChatGPT

**ChatGPT**

The "Units Sold" field has been updated. At every change in the month, the units sold are adjusted negatively based on the inflation rate. This reflects a reduction in the number of units sold due to the inflation impact. Here are the first 10 rows of the data with these adjustments:

| **Product ID** | **Category** | **Store ID** | **Date** | **Date of Sale** | **Country** | **Inflation Rate** | **GDP Growth Rate** | **Price Sold** | **Units Sold** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-01 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-02 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-03 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-04 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-05 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-06 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-07 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-08 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-09 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-10 | Ireland | 2.67 | -0.09 | 28.5 | 20.0 |

The number displayed in "Units Sold" reflects this approach, with the first month showing no change as there is no month change within this slice. If you'd like this updated data saved to a CSV or have further adjustments, let me know! ​​

User

**You**

download csv

ChatGPT

**ChatGPT**

The data with the adjusted "Units Sold" based on inflation has been saved to a CSV file. You can download it from the following link:

Download Final Revised Product Sales Data CSV ​​

User

**You**

Set the Units Sold to be an Integer.

ChatGPT

**ChatGPT**

I've updated the "Units Sold" column to be represented as integers. Here's how the first 10 rows look now:

| **Product ID** | **Category** | **Store ID** | **Date** | **Date of Sale** | **Country** | **Inflation Rate** | **GDP Growth Rate** | **Price Sold** | **Units Sold** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-01 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-02 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-03 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-04 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-05 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-06 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-07 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-08 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-09 | Ireland | 2.67 | -0.09 | 28.5 | 20 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-10 | Ireland | 2.67 | -0.09 | 28.5 | 20 |

If you need this updated data saved as a CSV or any further modifications, just let me know! ​​

User

**You**

download csv

ChatGPT

**ChatGPT**

The data with "Units Sold" as integers has been saved to a CSV file. You can download it from the following link:

Download Final Integer Units Sales Data CSV ​​

User

**You**

Update the Units Sold. For every row, Units Sold = Units Sold + a Random Integer between -3 and 3. The Units Sold value cannot be negative.

ChatGPT

**ChatGPT**

It seems there was an error in applying the **clip** function due to a syntax issue. I'll correct the method to ensure the 'Units Sold' value cannot go negative and reattempt the update. Let's fix that and try again. ​​

The "Units Sold" values have been successfully adjusted with random integers between -3 and 3 for every row, ensuring that the total does not drop below zero. Here are the first 10 rows showing these changes:

| **Product ID** | **Category** | **Store ID** | **Date** | **Date of Sale** | **Country** | **Inflation Rate** | **GDP Growth Rate** | **Price Sold** | **Units Sold** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-01 | Ireland | 2.67 | -0.09 | 28.5 | 19 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-02 | Ireland | 2.67 | -0.09 | 28.5 | 23 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-03 | Ireland | 2.67 | -0.09 | 28.5 | 22 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-04 | Ireland | 2.67 | -0.09 | 28.5 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-05 | Ireland | 2.67 | -0.09 | 28.5 | 17 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-06 | Ireland | 2.67 | -0.09 | 28.5 | 17 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-07 | Ireland | 2.67 | -0.09 | 28.5 | 19 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-08 | Ireland | 2.67 | -0.09 | 28.5 | 18 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-09 | Ireland | 2.67 | -0.09 | 28.5 | 19 |
| 1 | Mens | 1 | 2023-01-01 | 2023-01-10 | Ireland | 2.67 | -0.09 | 28.5 | 21 |

If you need this data saved to a CSV file, just let me know! ​​

User

**You**

download csv

ChatGPT

**ChatGPT**

The data with the adjusted "Units Sold" has been saved to a CSV file. You can download it from the following link:

Download Final Adjusted Units Sales Data CSV ​