



# COMMUNITY DAY

---

MÉXICO

14 de junio de 2025 | Ciudad de México

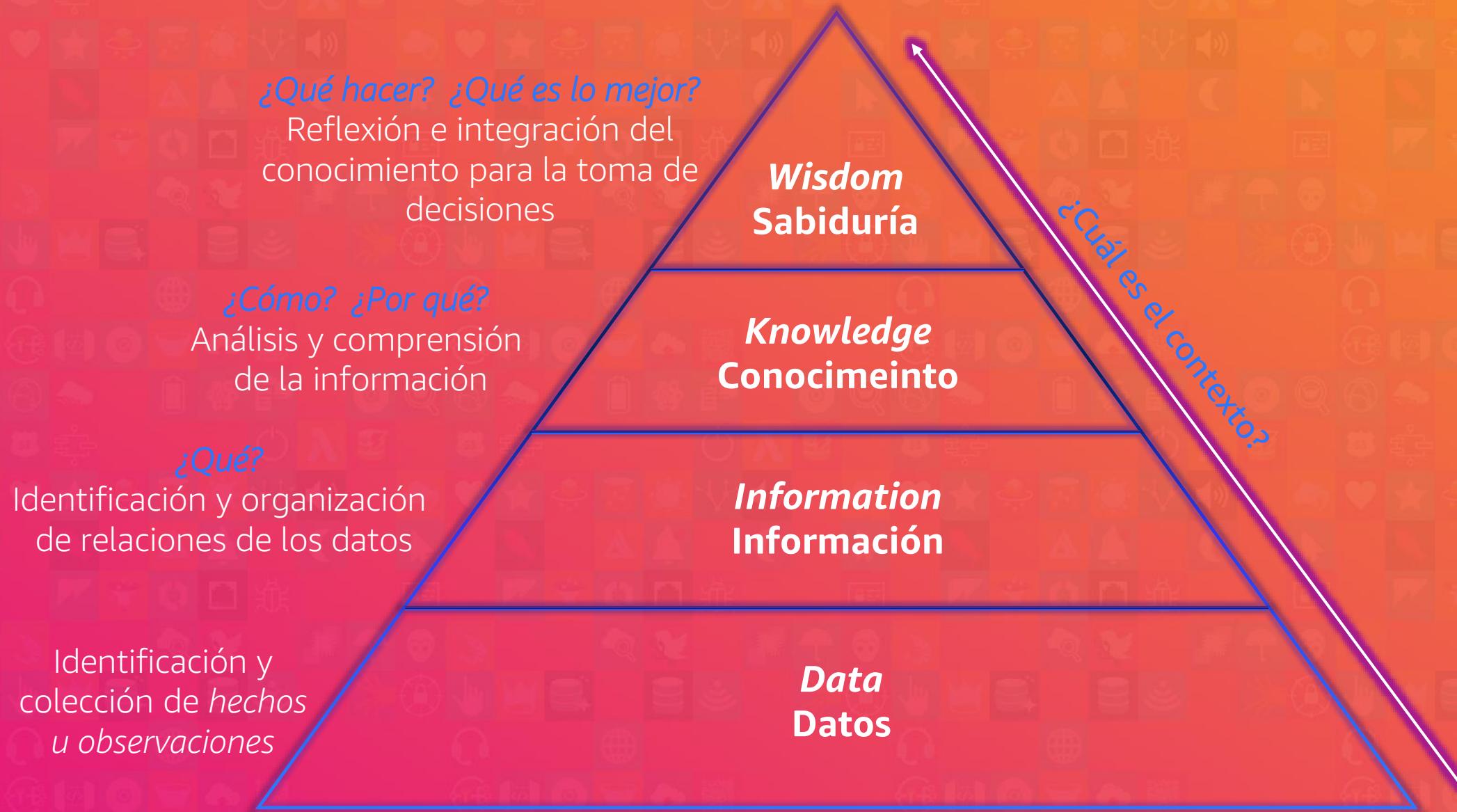
# Transformando datos y realidades: tu primera experiencia con AWS Glue

Zaira Chavarín

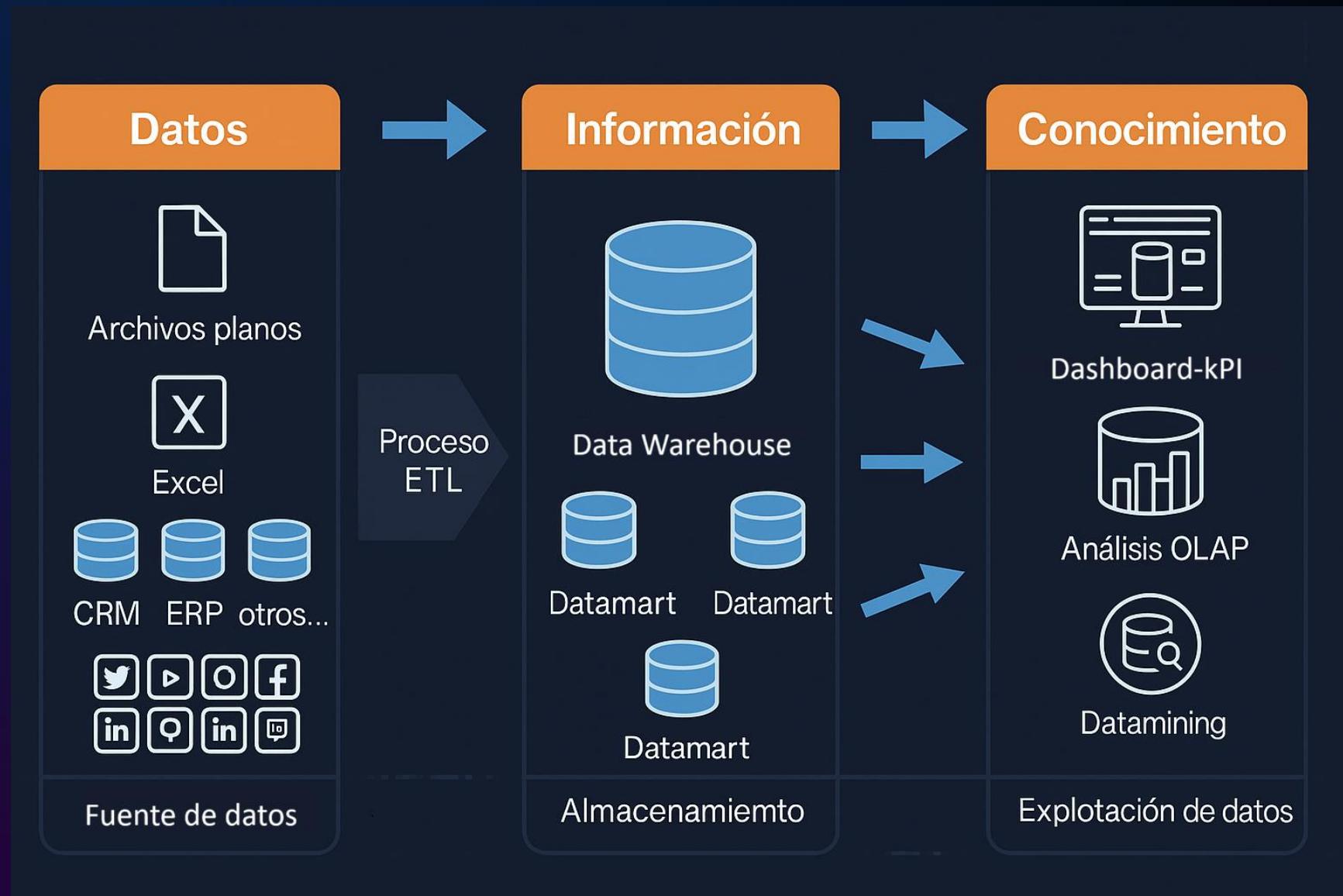
Data Analyst Engineer

- 01 ¿Cómo generamos conocimiento accionable?
- 02 Arquitectura
- 03 Selección y exploración de datos
- 04 ¡Manos a la obra: mi primer pipeline en AWS!
- 05 Conozcamos nuestros resultados en AWS QuickSight
- 06 Contextualizando la problemática
- 07 ¿Qué podemos hacer?

# La Pirámide del conocimiento (DIKW)



# Ciclo de vida de los datos



# ¿Qué es el proceso de ETL?



## Extracción

Es el proceso de conexión con alguna fuente de datos para su lectura y guardar los datos crudos de manera temporal en el área de preparación (*staging area*).

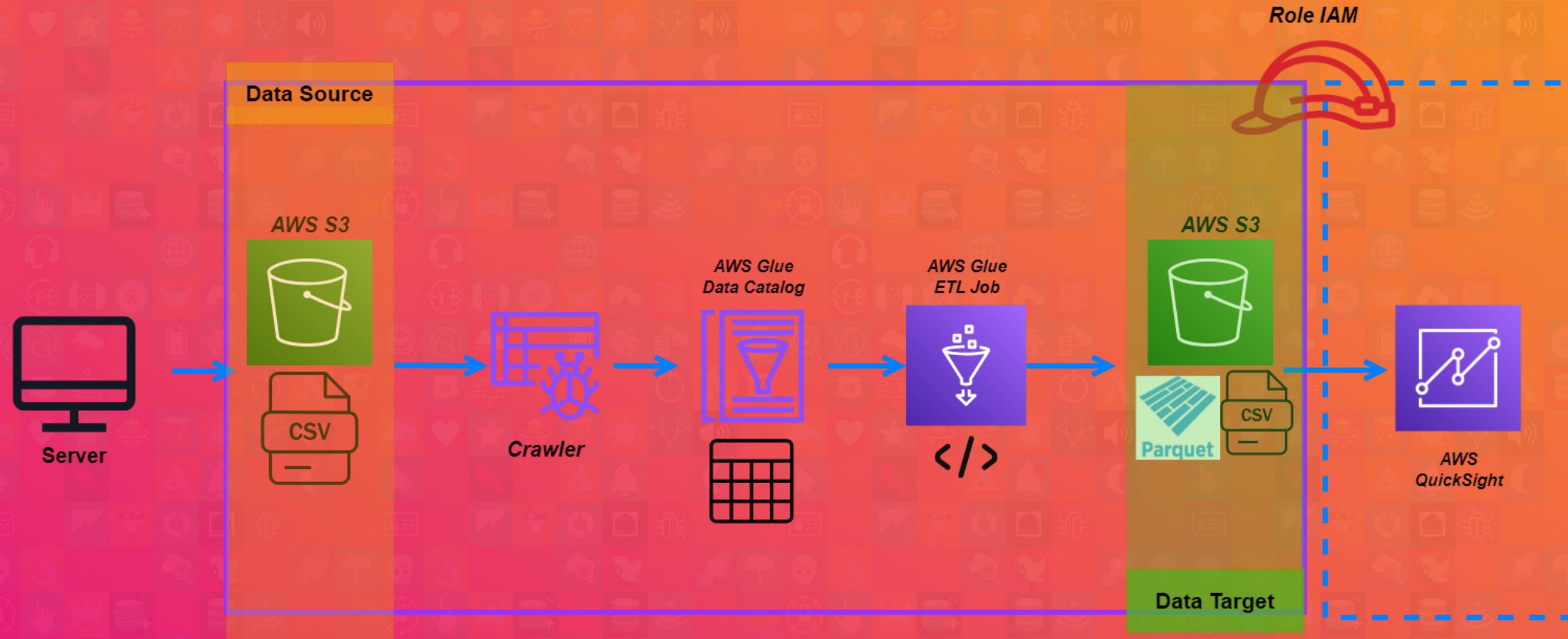
## Transformación

Normalización, limpieza, estructuración, modelado, validación de calidad y aplicación de reglas específicas a los datos.

## Carga

Integraremos los datos transformados en el *Data Warehouse*, se realiza una capa semántica donde traducimos los elementos técnicos, así como la optimización y particiones necesarias para su lectura.

# Arquitectura



# Objetivo

Visibilizar la situación de las mujeres latinoamericanas en la educación y desarrollo profesional dentro de la Ciencia, tecnología, ingeniería y matemáticas, en el último siglo por medio de estadísticas oficiales



- Porcentaje de personal académico por sexo 2013-2022
- Porcentaje de estudiantes en la educación superior por sexo 2013-2022



- Female share of graduates in other fields than Science, Technology, Engineering and Mathematics programmes, tertiary (%) [2000-2018]

1

## Servicios

**IAM**  
Administrar el acceso a los recursos de AWS

**Características principales**

[Grupos](#) [Usuarios](#) [Roles](#) [Políticas](#) [Analizador de acceso](#)

2

IAM > Roles > Crear rol

Paso 1

Seleccionar entidad de confianza

Agregar permisos

Asignar nombre, revisar y crear

### Seleccionar entidad de confianza Información

#### Tipo de entidad de confianza

Servicio de AWS

Permita que servicios de AWS como EC2, Lambda u otros realicen acciones en esta cuenta.

Cuenta de AWS

Permitir a las entidades de otras cuentas de AWS que le pertenezcan a usted o a un tercero realizar acciones en esta cuenta.

Identidad web

Permite a las personas federadas por el proveedor de identidad web externo especificado asumir este rol para realizar acciones en esta cuenta.

Federación SAML 2.0

Permitir que las personas federadas con SAML 2.0 a partir de un directorio corporativo realicen acciones en esta cuenta.

Política de confianza personalizada

Cree una política de confianza personalizada para permitir que otras personas realicen acciones en esta cuenta.

3

## Roles (5) Información

Un rol de IAM es una identidad que se puede crear y que tiene permisos específicos con credenciales que son válidas por períodos cortos. Los roles pueden ser a confianza.

<input type="checkbox"/>	Nombre del rol	Entidades de confianza	Última actividad
<input type="checkbox"/>	<a href="#">admid</a>	Cuenta: 717279701937	-
<input type="checkbox"/>	<a href="#">AWSServiceRoleForElasticLoadBalancing</a>	Servicio de AWS: elasticloadbalancin	Hace 101 días
<input type="checkbox"/>	<a href="#">AWSServiceRoleForSupport</a>	Servicio de AWS: support (Rol vincul:	-
<input type="checkbox"/>	<a href="#">AWSServiceRoleForTrustedAdvisor</a>	Servicio de AWS: trustedadvisor (Rol	-
<input type="checkbox"/>	<a href="#">glue-engineer</a>	Servicio de AWS: glue	Ayer

Ingrésa al servicio IAM desde tu consola para crear un rol de confianza con permisos para S3 y Glue: **AWSGlueServiceRole**

1

The screenshot shows the AWS Services console with the S3 service selected. The card displays the following information:

- S3** ★ Scalable Storage in the Cloud
- Top features**: Buckets, Access points, Batch Operations
- See all 7 results ▶**

2

The screenshot shows the Amazon S3 Buckets view for the 'woman-stem-latam' bucket. The interface includes:

- Header: Buscar, Estados Unidos (Norte de Virginia), demo-glue @ 7172-7970-1937
- Breadcrumbs: Amazon S3 > Buckets > woman-stem-latam
- Tab navigation: Objetos, Metadatos, Propiedades, Permisos, Métricas, Administración, Puntos de acceso
- Section: Objetos (0) with buttons: Copiar URI de S3, Copiar URL, Descargar, Abrir, Eliminar, Acciones, Crear carpeta, Cargar (highlighted with a yellow box)
- Note: Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el inventario de Amazon S3 para obtener una lista de todos los objetos de su bucket.
- Search bar: Buscar objetos por prefijo
- Table headers: Nombre, Tipo, Última modificación, Tamaño, Clase de almacenamiento
- Message: No hay objetos, No tiene objetos en este bucket.
- Buttons: Cargar (highlighted with a blue box)

3

The screenshot shows the 'Cargar' (Upload) page for the 'woman-stem-latam' bucket. The interface includes:

- Header: Buscar, Estados Unidos (Norte de Virginia), demo-glue @ 7172-7970-1937
- Breadcrumbs: Amazon S3 > Buckets > woman-stem-latam > Cargar
- Section: Cargar (highlighted with a blue box)
- Note: Agregue los archivos y las carpetas que desea cargar en S3. Para cargar un archivo de más de 160 GB, utilice la CLI de AWS, los SDK de AWS o la API REST de Amazon S3.
- Text: Arrastre y suelte aquí los archivos y carpetas que deseé cargar, o seleccione Add files (Agregar archivos) o Add folder (Agregar carpeta).
- Section: Archivos y carpetas (2 total, 6.9 KB)
 

Nombre	Carpeta	Tipo	Tamaño
personal_académico_STEM_clean.csv	-	text/csv	3.2 KB
estudiantes_STEM_clean.csv	-	text/csv	3.7 KB
- Buttons: Eliminar, Agregar archivos (highlighted with a yellow box), Agregar carpeta
- Search bar: Buscar por nombre

Ingresa a S3 para crear tu primer *bucket Source* y *bucket Target*, carga tus objetos en el primero.

1

The screenshot shows the AWS Glue console interface. On the left, there's a navigation sidebar with the following sections:

- AWS Glue** (selected)
- Getting started**
- ETL jobs
  - Visual ETL
  - Notebooks
  - Job run monitoring
- Data Catalog tables
- Data connections
- Workflows (orchestration)
- Zero-ETL integrations **New**
- Data Catalog**
  - Databases
  - Tables
  - Stream schema registries
  - Schemas
  - Connections
  - Crawlers** (highlighted with a yellow box)

2

The screenshot shows the 'Crawlers' page in the AWS Glue console. At the top, it says 'Crawlers (0) Info' and 'Last updated (UTC) April 29, 2025 at 01:21:17'. There are buttons for 'Action ▾', 'Run', and 'Create crawler' (which is highlighted with an orange box). Below this is a search bar labeled 'Filter crawlers' and a table header with columns: Name, State, Schedule, Last run, Last run ..., Log, and Table cha...'. A message below the table says 'No resources' and 'No resources to display.'

3

The screenshot shows the 'Set crawler properties' wizard. It's Step 1: Set crawler properties. The steps are listed on the left: Step 1 (radio button selected), Step 2 (radio button selected), Step 3, Step 4, Step 5. The main area is titled 'Set crawler properties' and contains 'Crawler details' with a 'Name' field set to 'women-stem-latam' (highlighted with a blue box). There are also sections for 'Description - optional' (with a placeholder 'Enter a description') and 'Tags - optional' (with a placeholder 'Use tags to organize and identify your resources.').

En AWS Glue creamos y nombramos nuestros primer *Crawlaer*. Esto nos permitirá rastrear el esquema y estructura de nuestros datos



4

Add crawler

- Set crawler properties
- Step 2 Choose data sources and classifiers**
- Step 3 Configure security settings
- Step 4 Set output and scheduling
- Step 5 Review and create

### CHOOSE DATA SOURCES AND CLASSIFIERS

#### Data source configuration

Is your data already mapped to Glue tables?

Not yet Select one or more data sources to be crawled.

Yes Select existing tables from your Glue Data Catalog.

**Data sources (0)** Info Edit Remove Add a data source

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
You don't have any data sources.		
<span>Add a data source</span>		

6

### Choose data sources and classifiers

#### Data source configuration

Is your data already mapped to Glue tables?

Not yet Select one or more data sources to be crawled.

Yes Select existing tables from your Glue Data Catalog.

**Data sources (2)** Info Edit Remove Add a data source

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
<input type="radio"/> S3	s3://woman-stem-latam/estudiante...	Recrawl all
<input type="radio"/> S3	s3://woman-stem-latam/personal_a...	Recrawl all

5

### Add data source

#### Data source

Choose the source of data to be crawled.

**S3**

#### Network connection - optional

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Clear selection Add new connection

#### Location of S3 data

In this account  
 In a different account

#### S3 path

Browse for or enter an existing S3 path.

s3://bucket/prefix/object View Browse S3

All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

#### Subsequent crawler runs

This field is a global field that affects all S3 data sources.

Crawl all sub-folders  
Crawl all folders again with every subsequent crawl.

Crawl new sub-folders only  
Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.

Crawl based on events  
Rely on Amazon S3 events to control what folders to crawl.

Cancel Add an S3 data source

Agrega la(s) ruta(s) de tu bucket para rastrear los datos

7

## Configure security settings

IAM role [Info](#)

Existing IAM role

glue-engineer glue-engineer C View

Create new IAM role Update chosen IAM role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

8

## Set output and scheduling

### Output configuration [Info](#)

#### Target database

Choose a database women-stem-latam C

Clear selection Add database

9

## Create a database

Create a database in the AWS Glue Data Catalog.

### Database details

Name women-stem-latam

Database name is required, in lowercase characters, and no longer than 200 characters.

### Description - optional

Enter text Enter text

Descriptions can be up to 2048 characters long.

### Database settings

Location - optional  
Set the URI location for use by clients of the Data Catalog.

10

## Databases (1)

Last updated (UTC)  
April 29, 2025 at 01:57:35 C Edit Delete Add database

A database is a set of associated table definitions, organized into a logical group.

<input type="text"/> Filter databases			▲   Name	▲   Description	▼   Location URI	▼   Created on (UTC)
<input type="checkbox"/>	women-stem-latam	-	-	-	-	April 29, 2025 at 01:57:31
<input type="checkbox"/>						

11

## Set output and scheduling

### Output configuration [Info](#)

#### Target database

women-stem-latam C

Clear selection Add database

#### Table name prefix - optional

Type a prefix added to table names

Agregamos permisos para el rol IAM previamente hecho y creamos una database en AWS Glue Data Catalog

12

## Review and create

### Step 1: Set crawler properties

#### Set crawler properties

Name  
woman-stem-latam

Tags  
-

### Step 2: Choose data sources and classifiers

#### Data sources (1) Info

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://woman-stem-latam	Recrawl all

### Step 3: Configure security settings

#### Configure security settings

IAM role  
AWSGlueServiceRole-EngineerDataAnalyst

Lake Formation configuration  
-

### Step 4: Set output and scheduling

#### Set output and scheduling

Database  
woman-stem-latam

Maximum table threshold - optional  
-

Table prefix - optional  
-

Schedule  
On demand

13

⌚ One crawler successfully created

The following crawler is now created: "woman-stem-latam"

## Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

### Crawlers (1) Info

Last updated (UTC)  
April 29, 2025 at 04:26:21



Action ▾

Run

Create crawler

View and manage all available crawlers.

Filter crawlers

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run ...	Log	Table cha...
<input type="checkbox"/>	woman-ste...	Ready	-	-	-	-	-

Verificamos las configuraciones del rastreador y finalizamos la creación del crawler

1

AWS Glue

**Getting started**

- ETL jobs
  - Visual ETL
  - Notebooks
  - Job run monitoring
- Data Catalog tables
- Data connections
- Workflows (orchestration)
- Zero-ETL integrations [New](#)

**▼ Data Catalog**

- Databases** (highlighted with a yellow box)
- Tables
- Stream schema registries
- Schemas
- Connections
- Crawlers

2

db-woman-stem-latam

Last updated (UTC) April 30, 2025 at 23:07:17 [Edit](#) [Delete](#)

**Database properties**

Name	db-woman-stem-latam	Description	-	Location	-	Created on (UTC)	April 30, 2025 at 23:03:17
------	---------------------	-------------	---	----------	---	------------------	----------------------------

**Tables (2)**

Last updated (UTC) April 30, 2025 at 23:07:19 [Edit](#) [Delete](#) [Add tables using crawler](#) [Add table](#)

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column stats...
estudiantes_stem_c	db-woman-stem-lat	s3://woman-stem-l-	CSV	-	Table data	View data quality	View statistics
personal_acad_micc	db-woman-stem-lat	s3://woman-stem-l-	CSV	-	Table data	View data quality	View statistics

3

AWS Glue > Tables > woman\_stem\_latam

**Schema** (13) [Edit schema as JSON](#) [Edit schema](#)

**Schema (13)**

View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	col0	string	-	-
2	col1	string	-	-
3	col2	string	-	-
4	col3	double	-	-
5	col4	double	-	-
6	col5	double	-	-
7	col6	double	-	-
8	col7	double	-	-
9	col8	double	-	-
10	col9	double	-	-
11	col10	double	-	-
12	col11	double	-	-
13	col12	double	-	-

Entramos a Databases en Data Catalog para revisar el esquema de nuestras bases de datos

14

AWS Glue

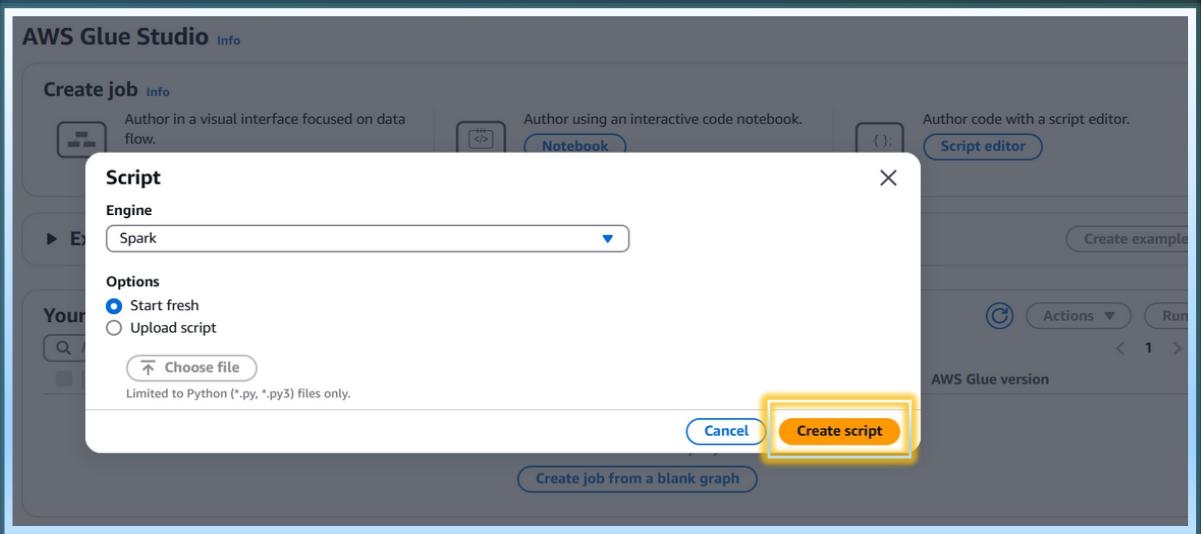
**Getting started**

- ETL jobs**
- Visual ETL
- Notebooks
- Job run monitoring
- Data Catalog tables
- Data connections
- Workflows (orchestration)
- Zero-ETL integrations **New**

**Data Catalog**

- Databases
- Tables
- Stream schema registries
- Schemas
- Connections
- Crawlers

15



16

```

import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
job.commit()

```

Actions [Save](#) [Run](#)

Script [Info](#)

untitled job

Script | Job details | Runs | Data quality | Schedules | Version Control

Python Ln 1, Col 1 Errors: 0 Warnings: 0

Ingresamos a ETL Jobs para crear nuestro script en Sapark



# Exploración de datos

País	Género	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Argentina	Femenino	0.4898	0.4895	0.4949	0.497	0.4975	0.497	0.4993	0.5017	0.504	0.5093
Argentina	Masculino	0.5102	0.5105	0.5051	0.503	0.5025	0.503	0.5007	0.4983	0.496	0.4907
Brasil	Femenino	0.4567	0.4584	0.4598	0.4601	0.4631	0.4652	0.4677	0.4682	0.4699	0.4724
Brasil	Masculino	0.5433	0.5416	0.5402	0.5399	0.5369	0.5348	0.5323	0.5318	0.5301	0.5276
Chile	Femenino	0.4257	0.4272	0.4299	0.4332	0.4354	0.4402	0.4447	0.4452	0.4483	0.453
Chile	Masculino	0.5743	0.5728	0.5701	0.5668	0.5646	0.5598	0.5553	0.5548	0.5517	0.547
Colombia	Femenino	0.3593	0.3646	0.3679	0.3705	0.3737	0.3828	0.3843	0.3868	0.3983	0.4034
Colombia	Masculino	0.6407	0.6354	0.6321	0.6295	0.6263	0.6172	0.6157	0.6132	0.6017	0.5966
Costa Rica	Femenino				0.438	0.4438	0.4411	0.4403	0.4432	0.4395	0.4488
Costa Rica	Masculino				0.2079	0.5556	0.5589	0.5597	0.5568	0.5605	0.5512
Cuba	Femenino	0.5394	0.422	0.5696	0.5789	0.586	0.59	0.5913	0.5913	0.591	0.6026
Cuba	Masculino	0.4606	0.578	0.4304	0.4211	0.414	0.41	0.4087	0.4087	0.409	0.3974
Ecuador	Femenino	0.359	0.3724	0.3821	0.3909	0.3949	0.3997	0.4028	0.4063	0.4153	0.4218
Ecuador	Masculino	0.641	0.6276	0.6179	0.6091	0.6051	0.6003	0.5972	0.5937	0.5847	0.5782
El Salvador	Femenino	0.3681	0.3657	0.3727	0.3766	0.3832	0.3839	0.385	0.3799	0.3979	0.4041
El Salvador	Masculino	0.6319	0.6343	0.6273	0.6234	0.6168	0.6161	0.615	0.6201	0.6021	0.5959
España	Femenino	0.4069	0.4171	0.4247	0.4288	0.4344	0.4386	0.4443	0.4494	0.4542	0.4576
España	Masculino	0.5931	0.5829	0.5753	0.5712	0.5656	0.5614	0.5557	0.5506	0.5458	0.5424
Honduras	Femenino	0.4132	0.3879	0.408							
Honduras	Masculino	0.5868	0.6121	0.591							
Méjico	Femenino	0	0.4095	0.4118	0.4145	0.4174	0.421	0.4264	0.5727	0.5644	0.5559
Méjico	Masculino	0	0.5905	0.5882	0.5855	0.5826	0.579	0.5736	0.4273	0.4356	0.4441
Panamá	Femenino	0.4735	0.471	0.4803	0.4984	0.4823	0.488	0.4904	0.4871	0.4804	0.5001
Panamá	Masculino	0.5265	0.529	0.5197	0.5016	0.5177	0.512	0.5096	0.5129	0.5196	0.4999
Perú	Femenino					0.3325	0.3292	0.3379	0.3387	0.352	0.3579
Perú	Masculino					0.6675	0.6708	0.6621	0.6613	0.648	0.6421
Portugal	Femenino	0.4404	0.4398	0.444	0.4445	0.4428	0.4476	0.4511	0.458	0.458	0.4621
Portugal	Masculino	0.5596	0.5602	0.556	0.5555	0.5572	0.5524	0.5489	0.542	0.542	0.5379

País	Género	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Brasil	Femenino	0.5713	0.5729	0.5706	0.5708	0.5684	0.569	0.5729	0.5773	0.5827	0.5878
Brasil	Masculino	0.4287	0.4271	0.4294	0.4292	0.4316	0.431	0.4271	0.4227	0.4173	0.4122
Chile	Femenino	0.5193	0.52	0.5193	0.5225	0.5264	0.53	0.5304	0.534	0.5416	0.5381
Chile	Masculino	0.4807	0.48	0.4807	0.4775	0.4736	0.47	0.4696	0.466	0.4584	0.4619
Colombia	Femenino	0.5271	0.5276	0.5293	0.5288	0.5292	0.5298	0.5269	0.5297	0.534	0.5344
Colombia	Masculino	0.4729	0.4724	0.4707	0.4712	0.4708	0.4702	0.4731	0.4703	0.466	0.4656
Costa Rica	Femenino	0.5399	0.5429	0.5565	0.5424	0.5474	0.5381	0.5422	0.5492	0.5579	0.559
Costa Rica	Masculino	0.4601	0.4571	0.4435	0.4576	0.4526	0.4619	0.4578	0.4508	0.4421	0.441
Cuba	Femenino	0.5945	0.5638	0.5679	0.6237	0.5988	0.6102	0.6337	0.6297	0.6365	0.6414
Cuba	Masculino	0.4055	0.4362	0.4321	0.3763	0.4012	0.3898	0.3663	0.3703	0.3635	0.3586
Ecuador	Femenino	0.5493	0.5425	0.538	0.5316	0.5184	0.5241	0.5247	0.5319	0.4656	0.4637
Ecuador	Masculino	0.4507	0.4575	0.462	0.4684	0.4767	0.4759	0.4753	0.4681	0.3875	0.3821
El Salvador	Femenino	0.5335	0.5347	0.5359	0.5376	0.536	0.5381	0.54	0.5393	0.5515	0.4391
El Salvador	Masculino	0.4665	0.4653	0.4641	0.4624	0.464	0.4619	0.46	0.4607	0.4485	0.5609
España	Femenino	0.5355	0.5334	0.5314	0.5326	0.5332	0.5357	0.5366	0.5407	0.5421	0.5443
España	Masculino	0.4645	0.4666	0.4686	0.4674	0.4668	0.4643	0.4634	0.4593	0.4579	0.4557
Honduras	Femenino	0.5726	0.57	0.57	0.5722	0.5682	0.5688	0.5725	0.5627	0.5694	0.5855
Honduras	Masculino	0.4274	0.43	0.43	0.4278	0.4318	0.4312	0.4275	0.4373	0.4306	0.4145
Méjico	Femenino	0.4932	0.4935	0.493	0.4986	0.5016	0.506	0.5098	0.5154	0.5252	0.5355
Méjico	Masculino	0.5068	0.5065	0.507	0.5014	0.4984	0.494	0.4902	0.4846	0.4748	0.4645
Panamá	Femenino	0.5917	0.6066	0.6046	0.6043	0.6062	0.6071	0.5973	0.6003	0.6054	0.6091
Panamá	Masculino	0.4083	0.3934	0.3954	0.3957	0.3938	0.3929	0.4027	0.3997	0.3946	0.3909
Paraguay	Femenino							0.5328	0.5586	0.5708	0.5647
Paraguay	Masculino							0.4672	0.4414	0.4292	0.4353
Perú	Femenino					0.5174	0.519	0.5224	0.5234	0.5366	0.5383
Perú	Masculino					0.4826	0.481	0.4776	0.4766	0.4634	0.4617
Portugal	Femenino	0.5303	0.5328	0.5327	0.5306	0.5323	0.5348	0.5379	0.5379	0.5356	0.536
Portugal	Masculino	0.4697	0.4672	0.4673	0.4694	0.4677	0.4652	0.4621	0.4621	0.4644	0.464
Puerto Rico	Femenino	0.5824	0.5804	0.5787	0.5814	0.584	0.5881	0.5928	0.608	0.6071	

Country Name	Country Code	Year	Value	Disaggregation
Argentina	ARG	2010	0.43463	female, Science, Technology, Engineering and Mathematics (STEM)
Argentina	ARG	2009	0.44164	female, Science, Technology, Engineering and Mathematics (STEM)
Argentina	ARG	2008	0.43302	female, Science, Technology, Engineering and Mathematics (STEM)
Argentina	ARG	2007	0.47421	female, Science, Technology, Engineering and Mathematics (STEM)
Argentina	ARG	2006	0.44815	female, Science, Technology, Engineering and Mathematics (STEM)
Belize	BLZ	2015	0.4183	female, Science, Technology, Engineering and Mathematics (STEM)
Brazil	BRA	2017	0.36644	female, Science, Technology, Engineering and Mathematics (STEM)
Brazil	BRA	2016	0.352	female, Science, Technology, Engineering and Mathematics (STEM)
Brazil	BRA	2015	0.34447	female, Science, Technology, Engineering and Mathematics (STEM)
Brazil	BRA	2014	0.34879	female, Science, Technology, Engineering and Mathematics (STEM)
Brazil	BRA	2012	0.3233	female, Science, Technology, Engineering and Mathematics (STEM)
Brazil	BRA	2011	0.31625	female, Science, Technology, Engineering and Mathematics (STEM)
Brazil	BRA	2010	0.30608	female, Science, Technology, Engineering and Mathematics (STEM)
Brazil	BRA	2009	0.32158	female, Science, Technology, Engineering and Mathematics (STEM)

- Número de columnas y filas
- Tipo de datos
- Datos nulos y NaN
- Datos duplicados
- Valores atípicos
- Estadísticas descriptivas
- Consistencia y formato

# Tratamiento de datos nulos según su tipo

## Missing completely at random (MCAR)

La pérdida no está relacionada con las características observadas. Todas las variables y **observaciones** tienen la misma probabilidad de faltar.

Eliminación, media mediana o moda, Hot-deck o nearest-neighbor , Maximum Likelihood (ML)

## Mising at random (MAR)

La pérdida **está relacionada solo con las características observadas**; con el valor de la variable o de otras variables del conjunto de datos

Imputación Múltiple (MICE), EM (Expectation–Maximization), Full Information Maximum Likelihood (FIML), Inverse Probability Weighting (IPW)

## Not missing at random (NMAR)

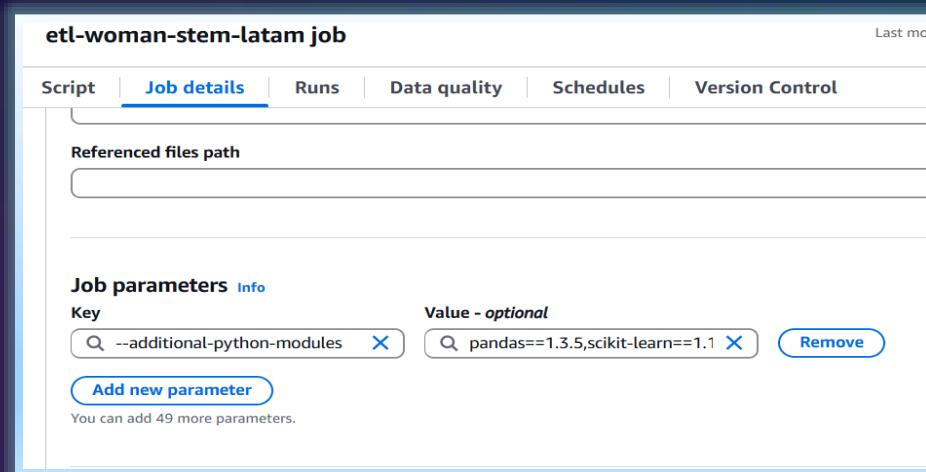
La pérdida está relacionada con características no observadas y quizás con características observadas

Modelos de selección (Heckman, Diggle–Kenward), Patrón-mezcla (Pattern-mixture), Modelado bayesiano con “missingness mechanism”

# Importación de librerías y parámetros de entrada

## #Importación de librerias

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
import pandas as pd
from sklearn.experimental import
enable_iterative_imputer # habilita
IterativeImputer
from sklearn.impute import IterativeImputer
```



```
## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
job.commit()

# Parámetros de entrada
DB_NAME      = "db-woman-stem-latam"
TABLE_STUD   = "estudiantes_stem_clean_aws_csv"
TABLE_PERS    =
"personal_acad_mico_stem_clean_aws_csv"
TARGET_BUCKET = "s3://target-woman-stem-latam/"
year_cols     =
['2013','2014','2015','2016','2017','2018','2019','2020
','2021','2022']
```

# Conexión con catálogo y partición

```
# 1. Leer desde el Catálogo
# Creamos DynamicFrames y los convertimos a Spark
DataFrames
df_students =
glueContext.create_dynamic_frame.from_catalog(
    database=DB_NAME, table_name=TABLE_STUD
).toDF()

df_personal =
glueContext.create_dynamic_frame.from_catalog(
    database=DB_NAME, table_name=TABLE_PERS
).toDF()

# 2. Asegurarnos de que las columnas de año sean
numéricas
for c in year_cols:
    df_students = df_students.withColumn(c,
    df_students[c].cast("double"))
    df_personal = df_personal.withColumn(c,
    df_personal[c].cast("double"))
```

```
# 3. Función para aislar solo los % de cada año en
Spark
def aislar_porcentajes(df):
    """Selecciona únicamente las columnas de año."""
    return df.select(*year_cols)

df_est_students = aislar_porcentajes(df_students)
df_est_personal = aislar_porcentajes(df_personal)
```

# Aplicación de MICE

```
# 5. Convertir a pandas para aplicar IterativeImputer  
(MICE)
```

```
pdf_students = df_est_students.toPandas()  
pdf_personal = df_est_personal.toPandas()
```

```
# 6. Función genérica de imputación MICE
```

```
def imputar_mice(df_pdf):
```

```
    """
```

Toma un DataFrame pandas con las columnas year\_cols, aplica IterativeImputer y devuelve un pandas.DataFrame con los mismos nombres de columna.

```
    """
```

```
    imputer = IterativeImputer(random_state=0)  
    arr = imputer.fit_transform(df_pdf)  
    return pd.DataFrame(arr,  
columns=df_pdf.columns)
```

```
# 7. Ejecutamos la imputación
```

```
students_mice_pdf =  
imputar_mice(pdf_students)  
personal_mice_pdf =  
imputar_mice(pdf_personal)
```

# Reconstrucción de DataFrames

```
# 8. Reconstruir Spark DataFrames  
juntando IDs + columnas imputadas
```

```
id_cols_students =  
df_students.columns[:2]  
id_cols_personal =  
df_personal.columns[:2]
```

```
# 9. Convertir pandas imputado de vuelta a Spark  
DataFrame  
spark_students_mice = spark.createDataFrame(  
  
pd.concat([df_students.select(*id_cols_students).  
toPandas(), students_mice_pdf], axis=1)  
)  
  
spark_personal_mice = spark.createDataFrame(  
  
pd.concat([df_personal.select(*id_cols_personal).  
toPandas(), personal_mice_pdf], axis=1)  
)
```

# Convertir datos a formato CSV o Parquet y guardarlos en el *Bucket Target*

```
# 10. Escribir los resultados en Parquet en el bucket de destino
spark_students_mice.write.mode("overwrite")
\

.parquet(f"{TARGET_BUCKET}estudiantes_mice/")
)

spark_personal_mice.write.mode("overwrite")
\

.parquet(f"{TARGET_BUCKET}personal_mice/")
```

```
# 10. opción con csv para Quicksight
spark_students_mice.write.mode("overwrite")
\ .option("header", "true") \
.csv(f"{TARGET_BUCKET}estudiantes_mice_csv/")
)

spark_personal_mice.write.mode("overwrite")
\ .option("header", "true") \
.csv(f"{TARGET_BUCKET}personal_mice_csv/")
```

```
# 11. Finaliza el Job
job.commit()
```

1

The screenshot shows the AWS Glue Job details page for the 'etl-woman-stem-latam' job. It displays a table of job runs, with one entry for a successful run on May 1, 2025, at 03:01:09. The run details section provides information such as Start time (Local), End time (Local), Duration, Capacity, Worker type, and Glue version.

Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (DPU)	Worker type	Glue version
Succeeded	0	04/30/2025 20:59:06	04/30/2025 21:00:44	1 m 31 s	2 DPUs	G.1X	3.0

3

The screenshot shows the AWS S3 Buckets page. It lists three buckets under the 'Buckets de uso general' tab:

- aws-glue-assets-717279701937-us-east-1**: Located in US East (N. Virginia) us-east-1. Created on 30 Apr 2025 8:51:16 PM CST.
- target-woman-stem-latam**: Located in US East (N. Virginia) us-east-1. Created on 30 Apr 2025 8:40:02 PM CST.
- woman-stem-latam**: Located in US East (N. Virginia) us-east-1. Created on 28 Apr 2025 7:13:33 PM CST.

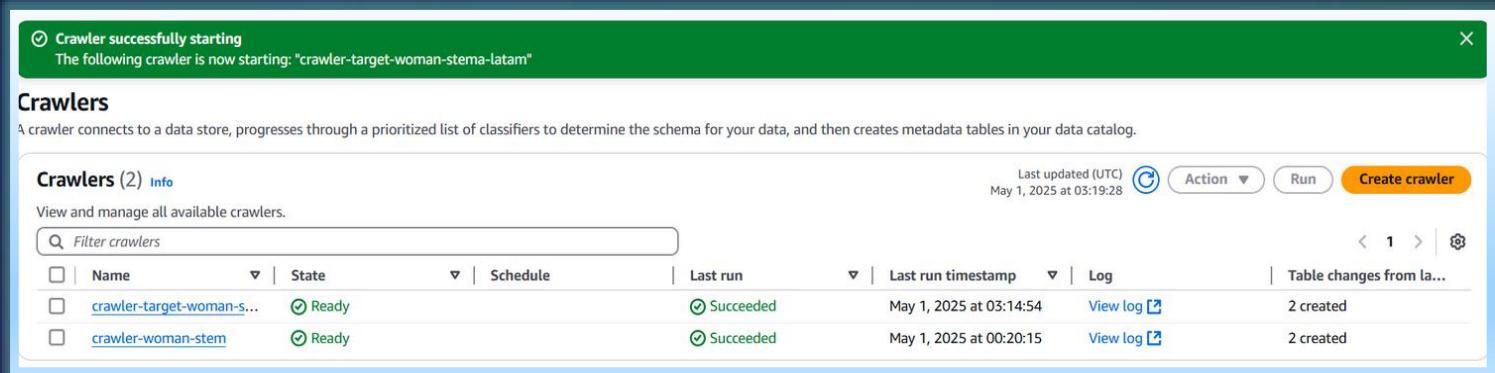
# Corremos nuestro Job y verificamos resultados en S3 dentro de nuestro *bucket Target*

4

The screenshot shows the AWS S3 Objects page for the 'target-woman-stem-latam' bucket. It lists several objects, all of which are parquet files named with a specific prefix. The objects were created on April 30, 2025, at various times between 9:00:32 PM and 9:03:27 PM CST.

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
part-00000fea249ce-4c19-49d3-80a5-66651b037c16-c000.snappy.parquet	parquet	30 Apr 2025 9:00:32 PM CST	3.7 KB	Estándar
part-00001fea249ce-4c19-49d3-80a5-66651b037c16-c000.snappy.parquet	parquet	30 Apr 2025 9:00:32 PM CST	3.7 KB	Estándar
part-00002fea249ce-4c19-49d3-80a5-66651b037c16-c000.snappy.parquet	parquet	30 Apr 2025 9:00:32 PM CST	3.7 KB	Estándar
part-00003fea249ce-4c19-49d3-80a5-66651b037c16-c000.snappy.parquet	parquet	30 Apr 2025 9:00:32 PM CST	3.8 KB	Estándar

1



Crawler successfully starting  
The following crawler is now starting: "crawler-target-woman-stem-latam"

**Crawlers**

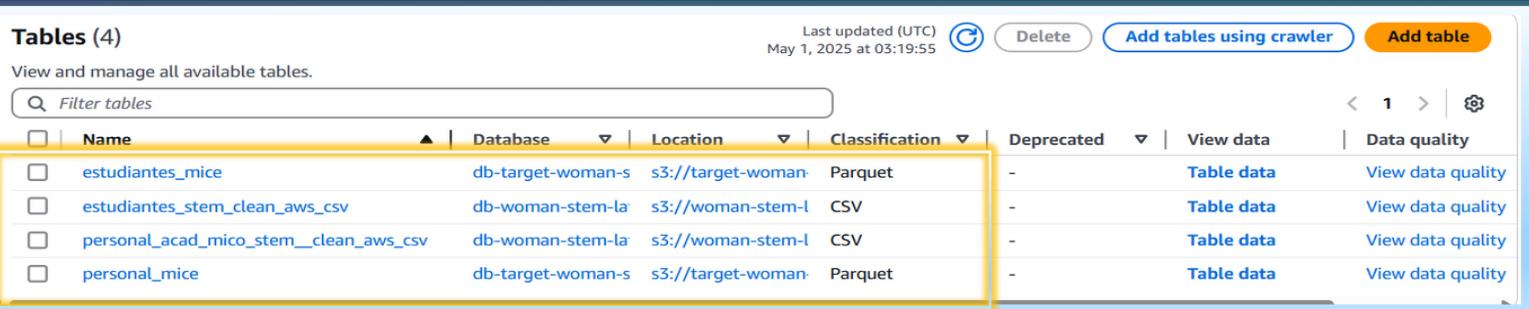
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

**Crawlers (2) Info**

Last updated (UTC) May 1, 2025 at 03:19:28

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from la...
crawler-target-woman-s...	Ready		Succeeded	May 1, 2025 at 03:14:54	View log	2 created
crawler-woman-stem	Ready		Succeeded	May 1, 2025 at 00:20:15	View log	2 created

2



Tables (4)

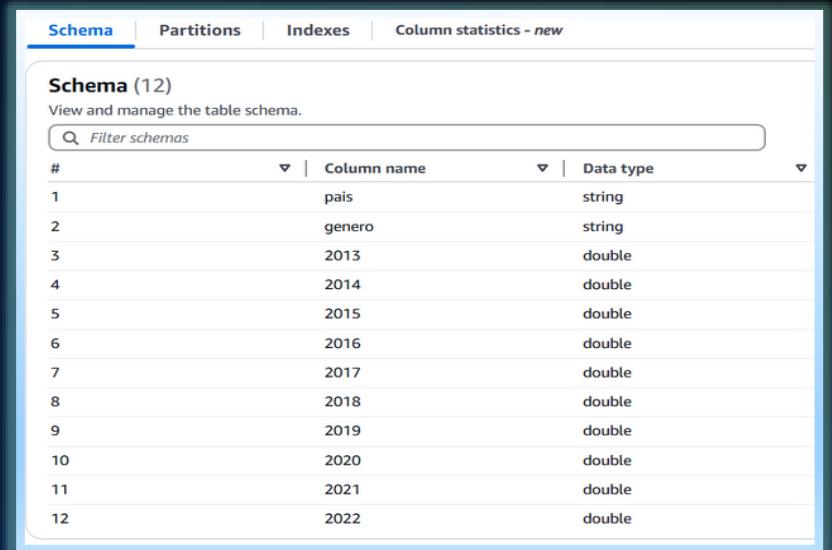
Last updated (UTC) May 1, 2025 at 03:19:55

**Tables (4)**

Add tables using crawler Add table

Name	Database	Location	Classification	Deprecated	View data	Data quality
estudiantes_mice	db-target-woman-s	s3://target-woman-	Parquet	-	Table data	View data quality
estudiantes_stem_clean_aws_csv	db-woman-stem-la	s3://woman-stem-l	CSV	-	Table data	View data quality
personal_acad_mico_stem_clean_aws_csv	db-woman-stem-la	s3://woman-stem-l	CSV	-	Table data	View data quality
personal_mice	db-target-woman-s	s3://target-woman-	Parquet	-	Table data	View data quality

3



**Schema** | Partitions | Indexes | Column statistics - new

**Schema (12)**

View and manage the table schema.

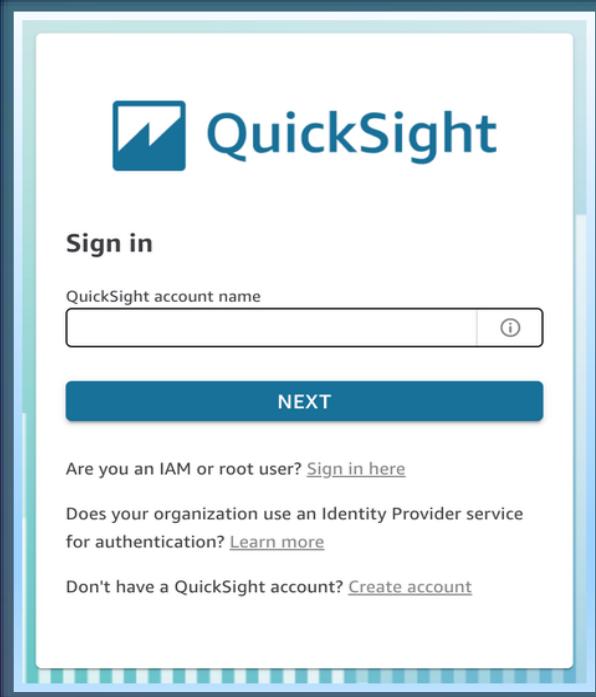
Filter schemas

#	Column name	Data type
1	pais	string
2	genero	string
3	2013	double
4	2014	double
5	2015	double
6	2016	double
7	2017	double
8	2018	double
9	2019	double
10	2020	double
11	2021	double
12	2022	double

**Creamos un nuevo *crawler* para verificar el esquema de nuestras nuevas tablas**

# Visualización con QuickSight

1

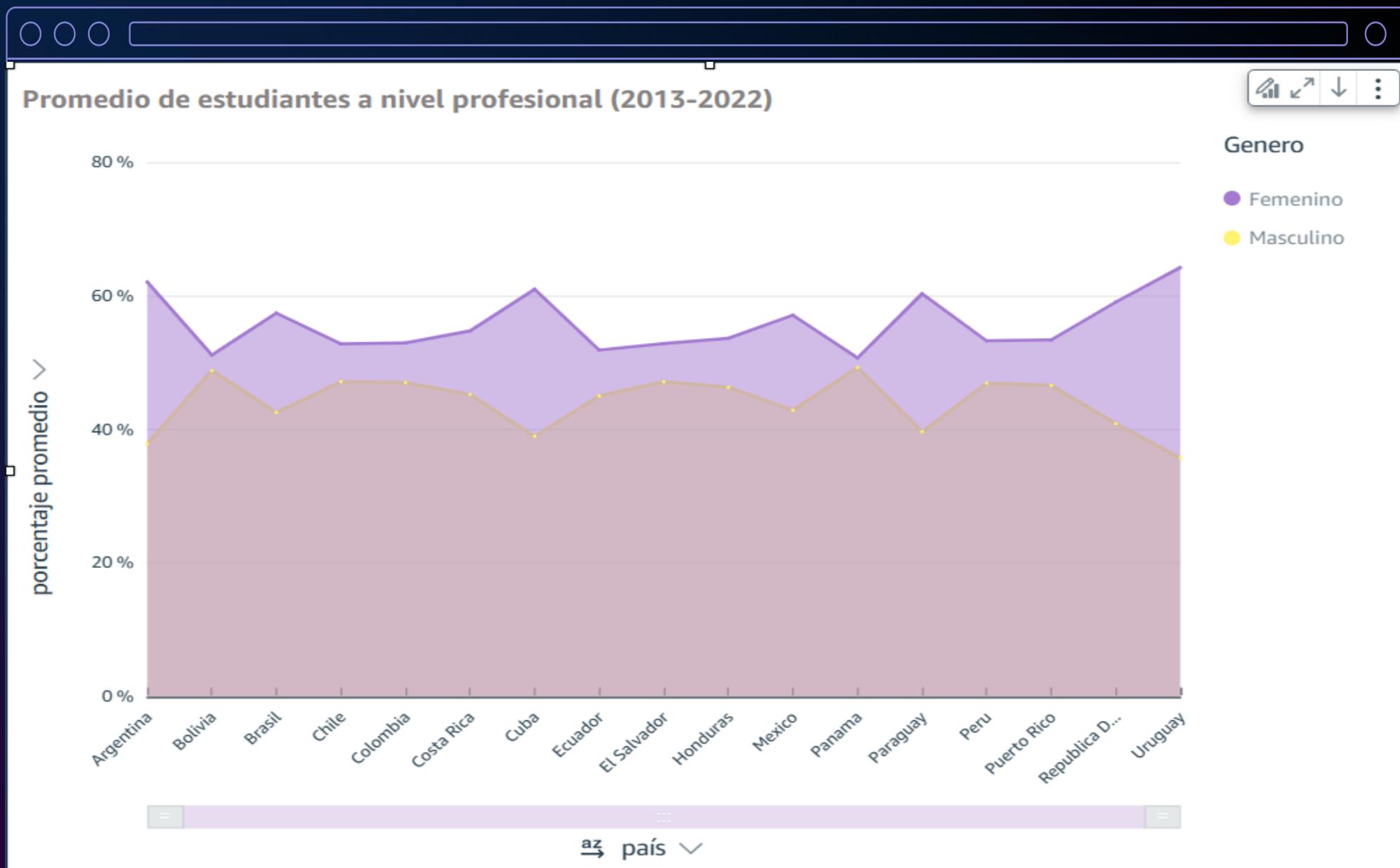


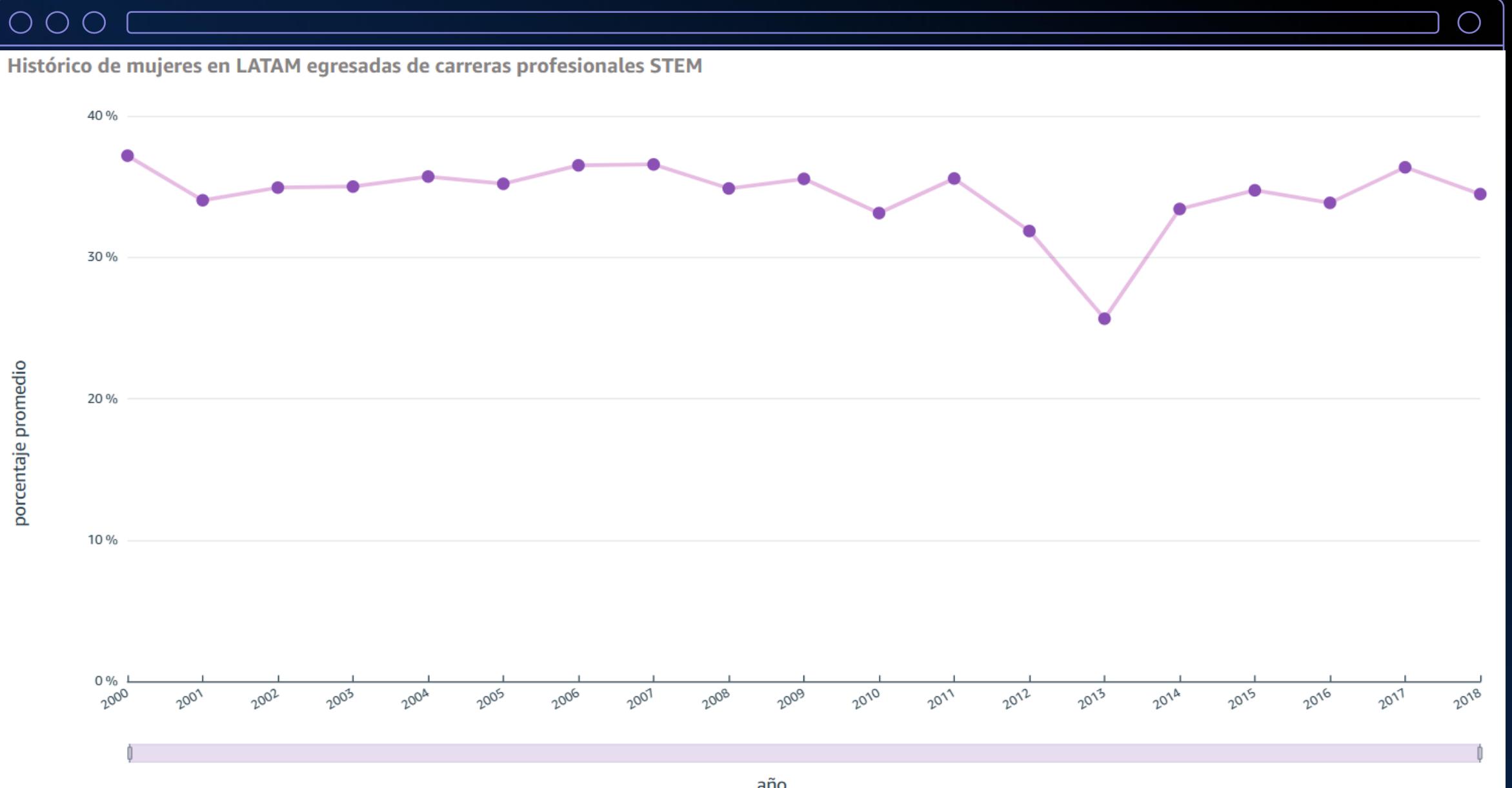
2

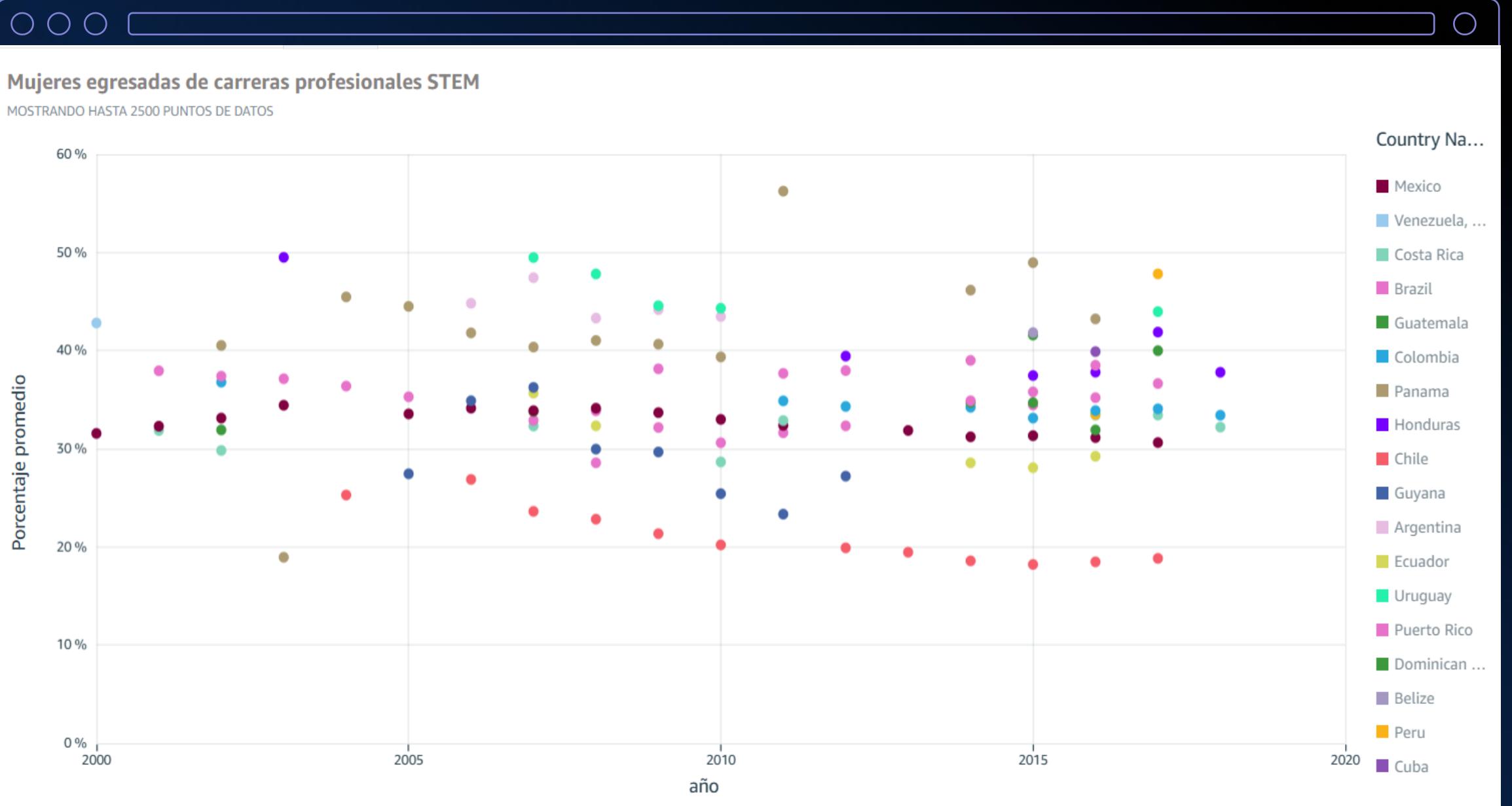
```
{  
  "fileLocations": [  
    {  
      "URIprefixes": [  
        "s3://target-woman-stem-  
        latam/estudiantes_mice/"  
      ]  
    },  
    {  
      "URIprefixes": [  
        "s3://target-woman-stem-  
        latam/personal_mice/"  
      ]  
    }  
  "globalUploadSettings": {  
    "format": "CSV"  
  }  
}
```

3

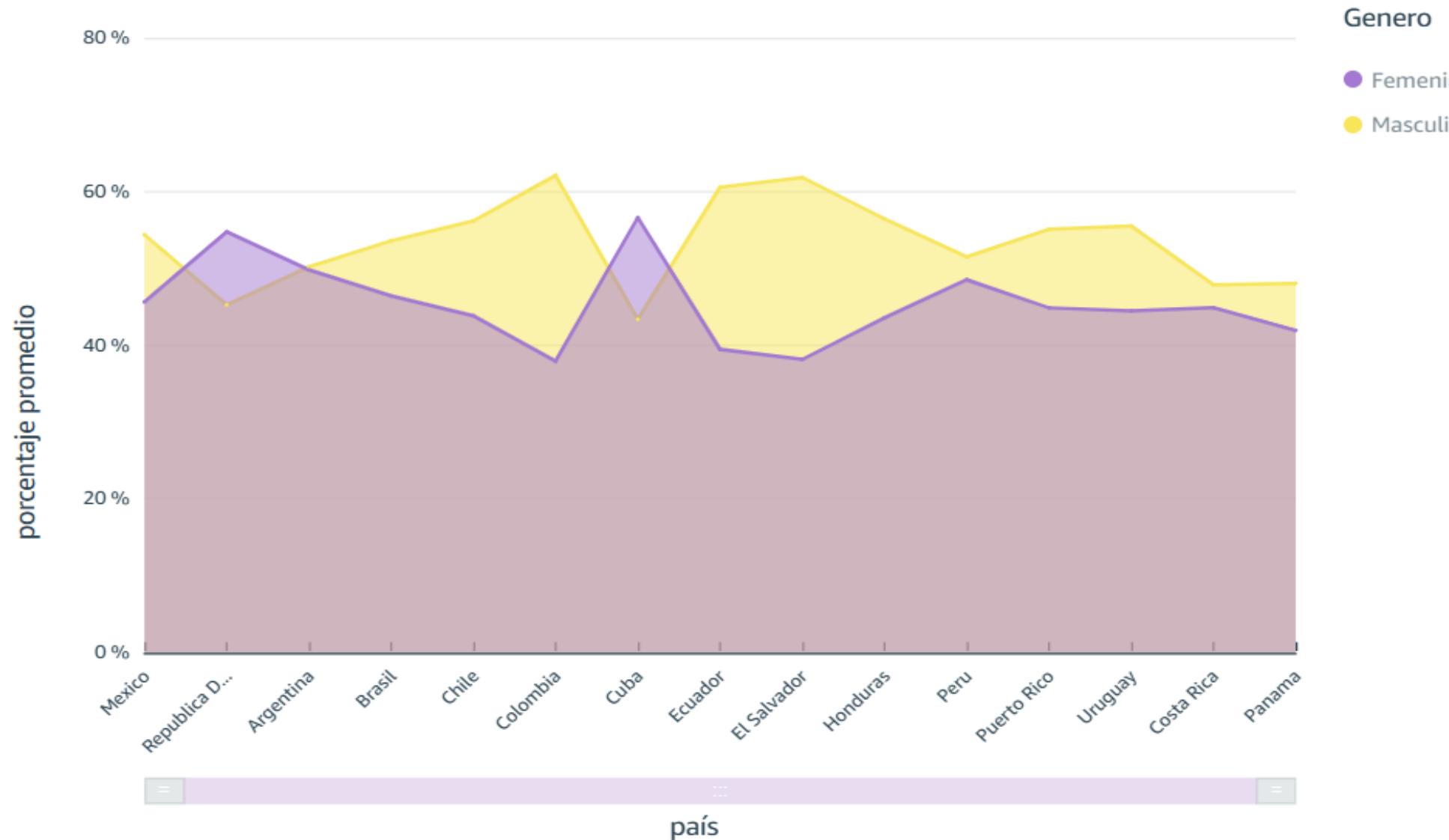








## Promedio de personal académico STEM (2013-2022)



# Conclusiones

A pesar de que las mujeres profesionales han aumentado de manera constante, la proporción de egresadas en carreras STEM se ha mantenido por debajo del 40 % durante casi veinte años.

- Solo países como Argentina, Panamá y Uruguay han logrado superar puntualmente esa barrera de manera constante: de hecho, Panamá alcanzó un récord del 56 % en 2011.
- En contraste, Chile presenta el mayor rezago, con tasas de graduación inferiores al 30 % y, en algunos años, incluso por debajo del 20 %.
- Mientras que México, durante casi 20 años ha permanecido entre 31 y 34% de mujeres egresadas en carreras STEM

Estos factores son determinantes para que la participación de las mujeres que asumen la docencia, la investigación y el desarrollo tecnológico estén por debajo de del 50%

“

Más allá de la educación, la evidencia da cuenta de brechas de género aún más amplias al analizar el mercado laboral. Los datos sobre empleo en el sector STEM en ALC son limitados, pero por ejemplo para el caso del sector de Tecnologías de la Información y las Comunicaciones (TIC) se constata que sólo 3 de cada 10 empleados son mujeres, con variaciones significativas entre países

UNDP, 2024

“

**A los 6 años, las niñas ya asocian la inteligencia con los varones. Entre los 11 y 15 años, muchas pierden el interés en STEM por falta de modelos a seguir y estereotipos de género por lo que tan sólo el 0.5% de las adolescentes expresan interés en carreras tecnológicas y científicas**

Bian et al, 2017 y UNIFEC, 2020

# Lo conozco, decido y lo transformo

## Sector educativo

- Estimular el interés desde la niñez: talleres y clubes STEM.
- Incorporar ejemplos y referentes femeninos en áreas STEM.
- Programas que acompañen a niñas y jóvenes durante toda la formación, así como espacios en tecnología donde prioricen la participación femenina.

## Mundo laboral

- Incorporar taza laboral femenina, sailor equitativo y capacitación
- Cultura inclusiva, políticas de cero acoso y discriminación
- Horarios adaptables, permisos parentales equitativos.
- Redes y espacios de apoyo y comités de ética en pro de la diversidad

## Nivel social

- Autoconciencia y formación
- Si eres líder, ofrece mentoría a mujeres jóvenes
- Allyship activo: no normalices exclusión a mujeres ni chistes sexistas.
- Networking inclusivo y diverso
- Comparte recursos, conocimientos, oportunidades de manera equitativa
- Busca referentes de mujeres en STEM y reconócelas



[https://github.com/zai-zu/etl\\_aws\\_women\\_stem/](https://github.com/zai-zu/etl_aws_women_stem/)

## Repositorio en GitHub



# ¡Gracias!

Zaira Chavarín amiranda

 [zai-zu](mailto:zai-zu)

 [/in/zaira-chavarin-miranda/](https://in/zaira-chavarin-miranda/)



Por favor, completa la  
encuesta de la sesión.