Fixing data

Data Driven Results Given

Avance De Proyecto

Multivitaminico 60+

SPRINT # 2

Agenda de Hoy...

01 Introducción

02 Roles del Sprint

03 Plan del Proyecto

04 Eda Exhaustivo

05 Imputaciones

06 DER - Modelado BBDD

07 Arquitectura

08 Pipeline

09 MVP DashBoad

10 Conclusión

Introducción

- Etapa de Data Engineering
- Sprint #2
- Foco en Escalabilidad y reducción de costos de infraestructura

Fixing data

Data Driven Results Given

Acerca de Fixing Data

Somos una joven y dinámica consultora de datos especializada en proyectos end-to-end. Nuestro equipo aborda proyectos que cubren todo el ciclo de vida de los datos. Con un enfoque multidisciplinario, ofrecemos soluciones integrales y valiosas para nuestros clientes.

Ing. Fernando G. Cofone
PM & Data Engineer

Ing. Paula Perosio
Data Analyst & Analitycs

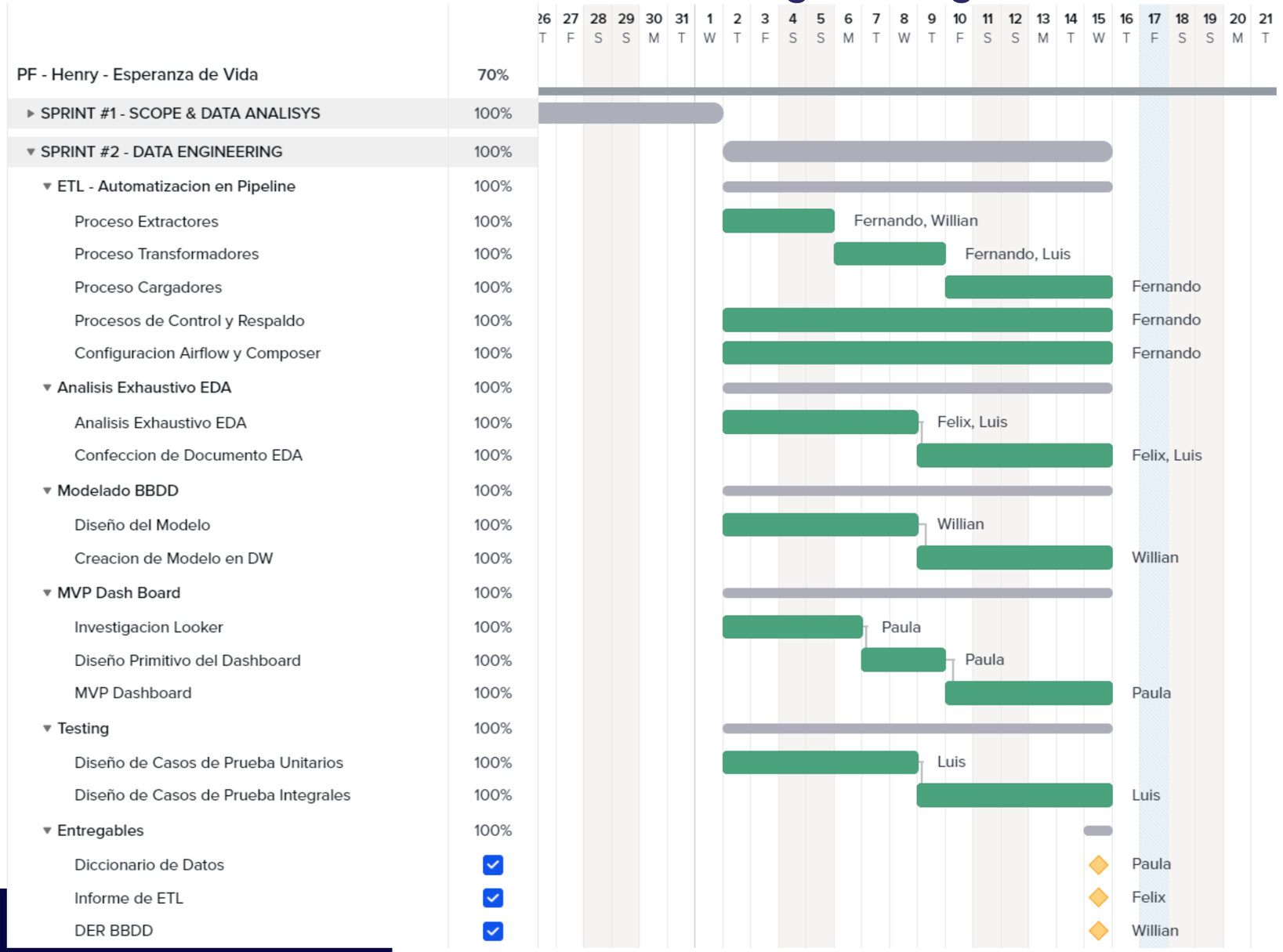
Ing. Willian Jose Suarez
Data Engineer

Felix Contreras
Data Analyst

Ing. Luis David Mendoza
Data Analyst

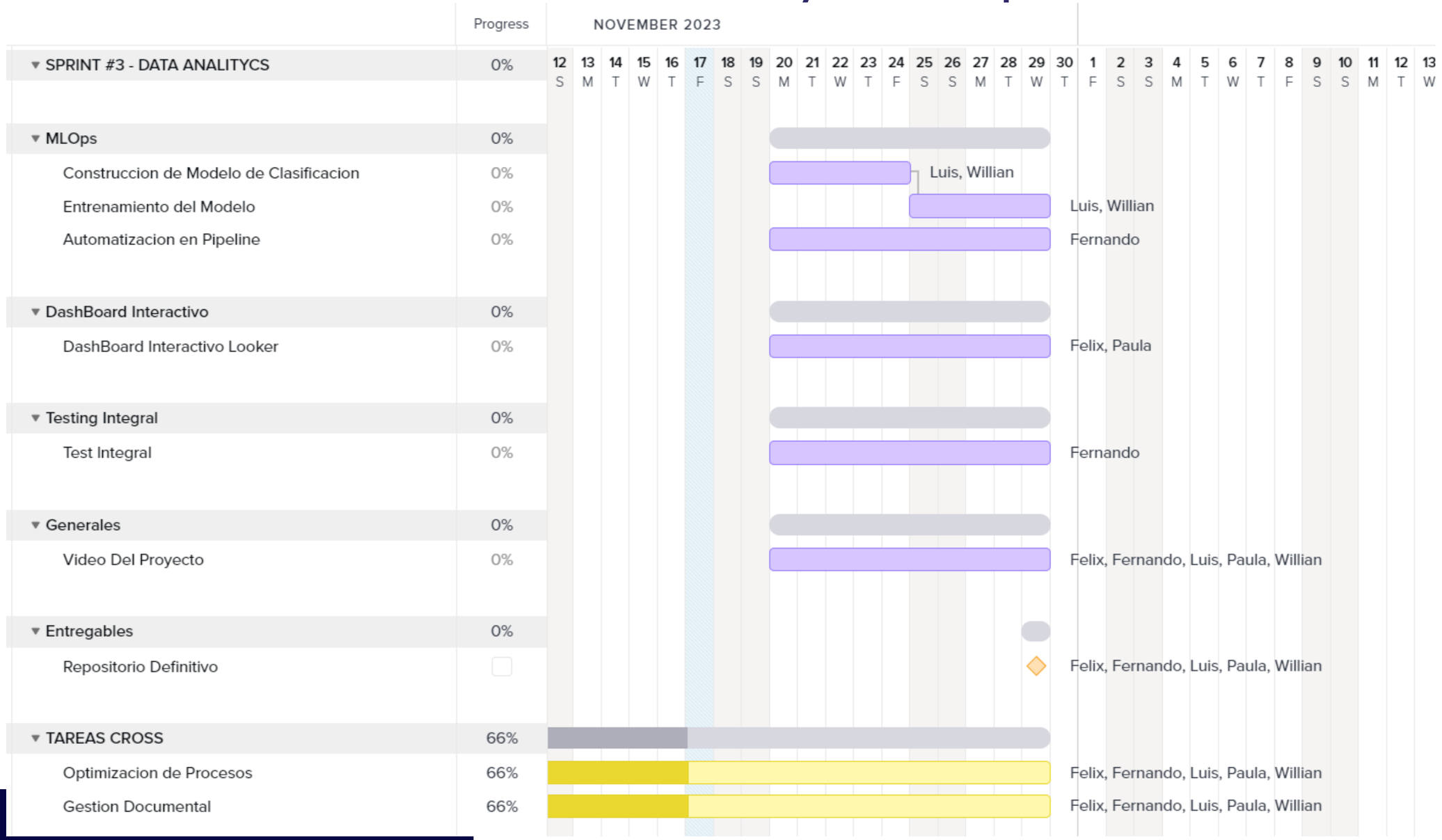
Plan Del Proyecto

SPRINT # 2 - Data Engineering



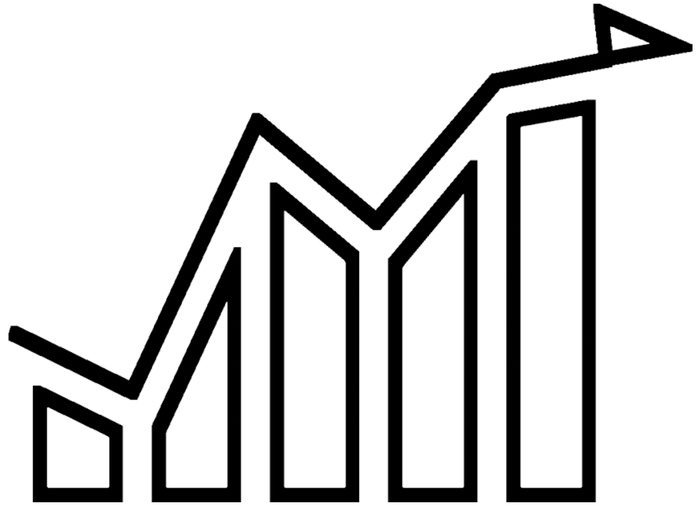
Plan Del Proyecto

SPRINT # 3 - Data Analytics & MLOps

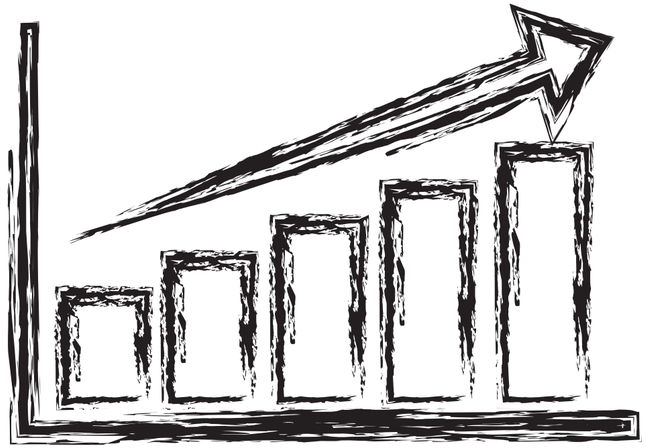


EDA EXHAUSTIVO

La fuente de datos principal es pagina oficial del Banco Mundial
https://databank.worldbank.org/databases

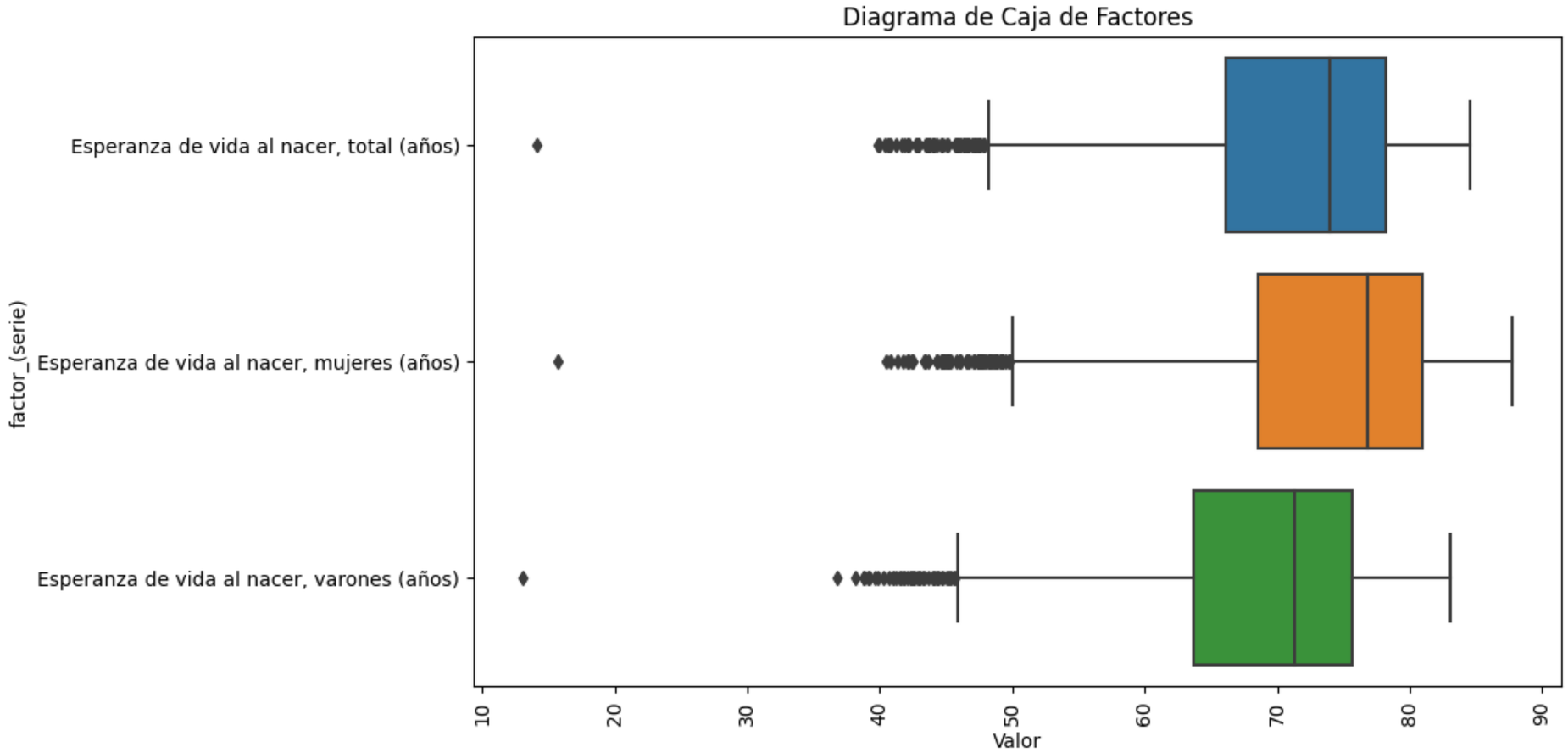


- El conjunto de países en estudio refleja una diversidad geográfica significativa, representando un total de 44 naciones.

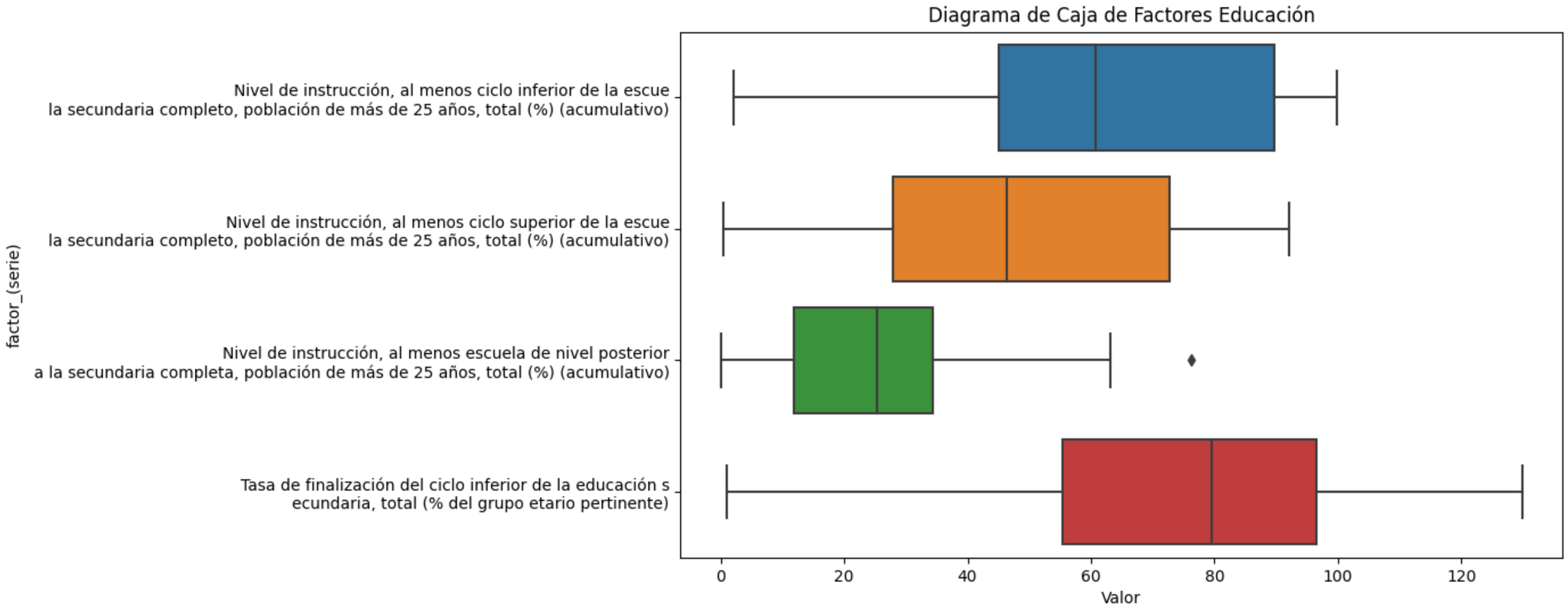


- Con respecto a lapso de tiempo en estudio comprende un extenso periodo, desde 1987 hasta 2022, con un rango variable de observaciones.

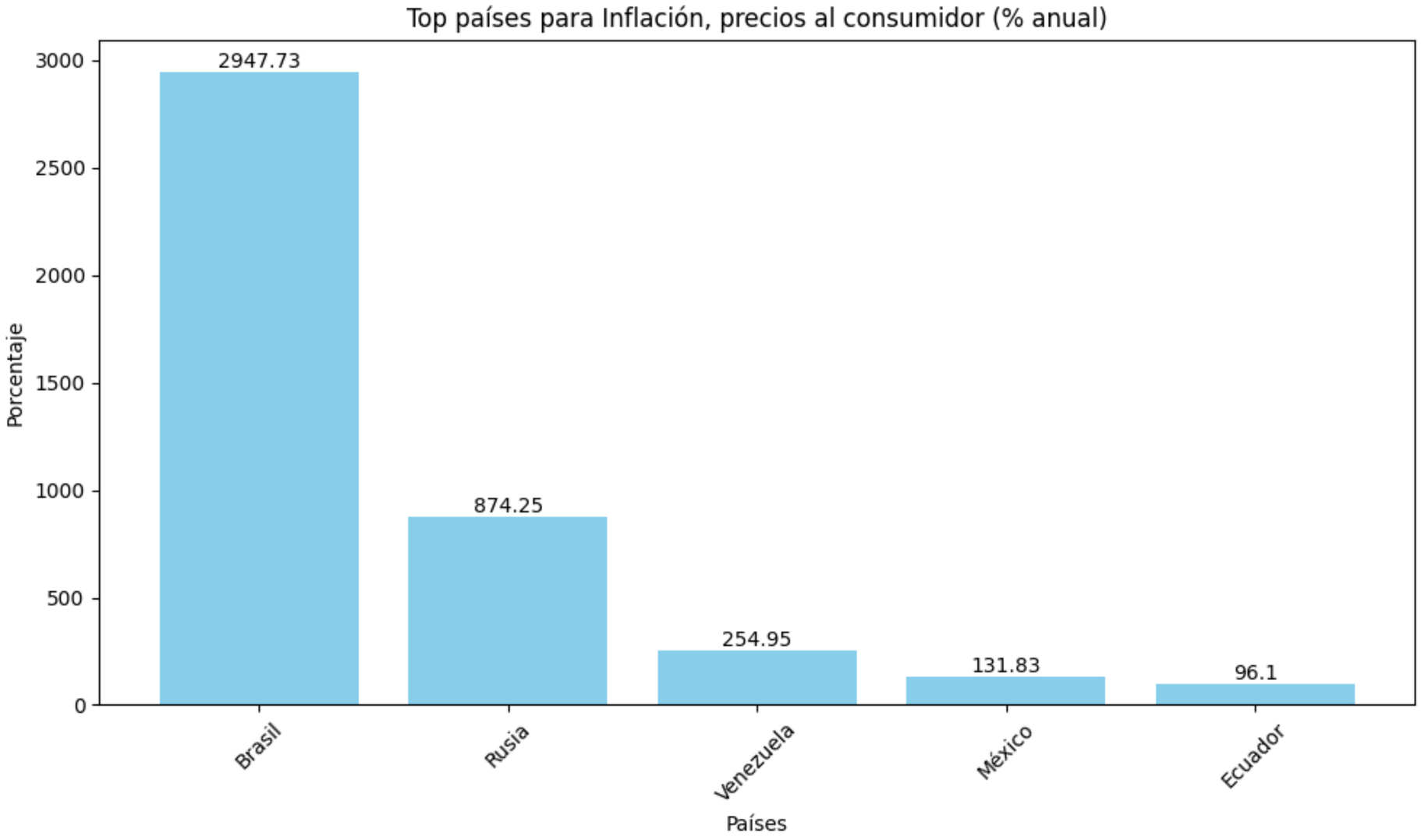
EDA EXHAUSTIVO



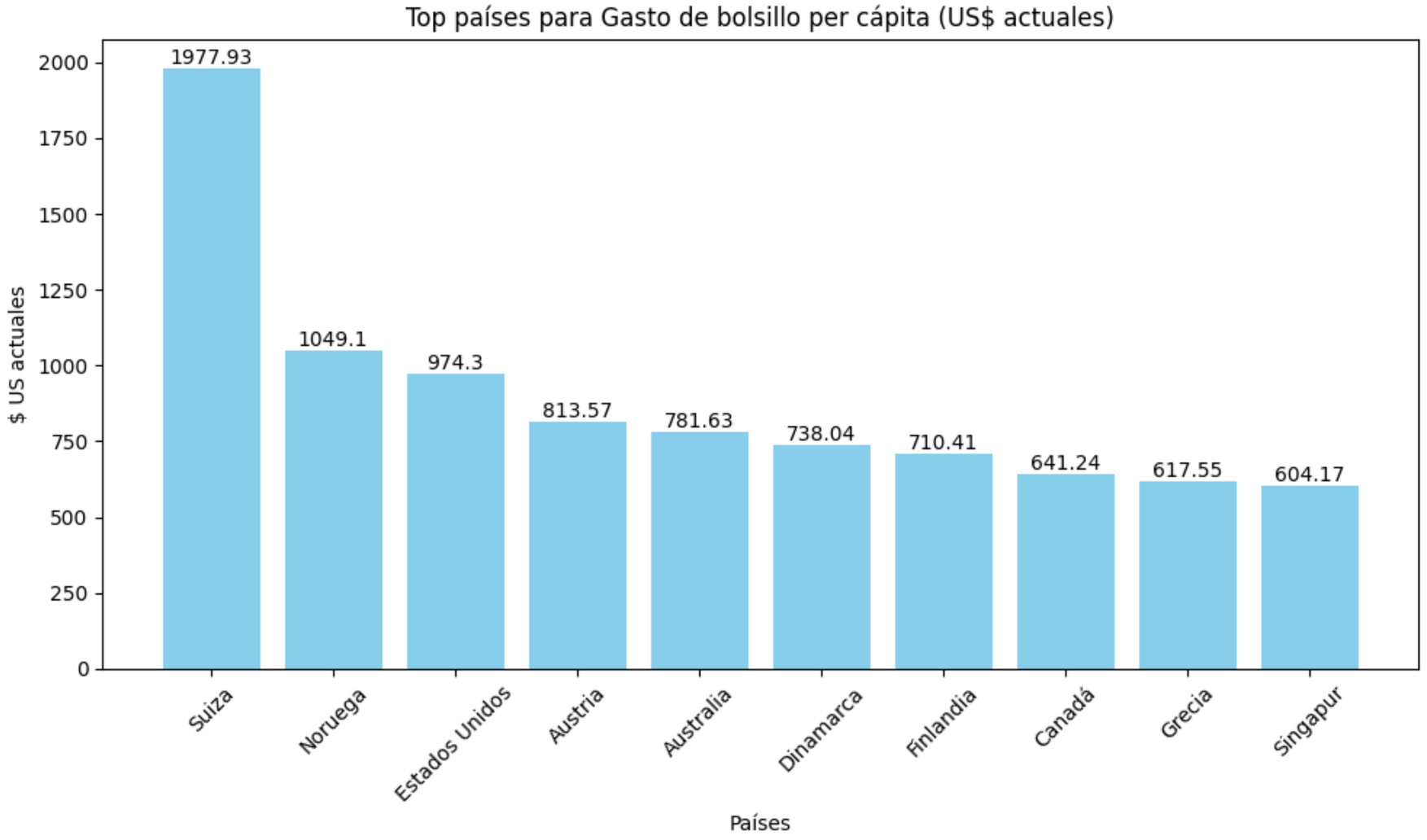
EDA EXHAUSTIVO



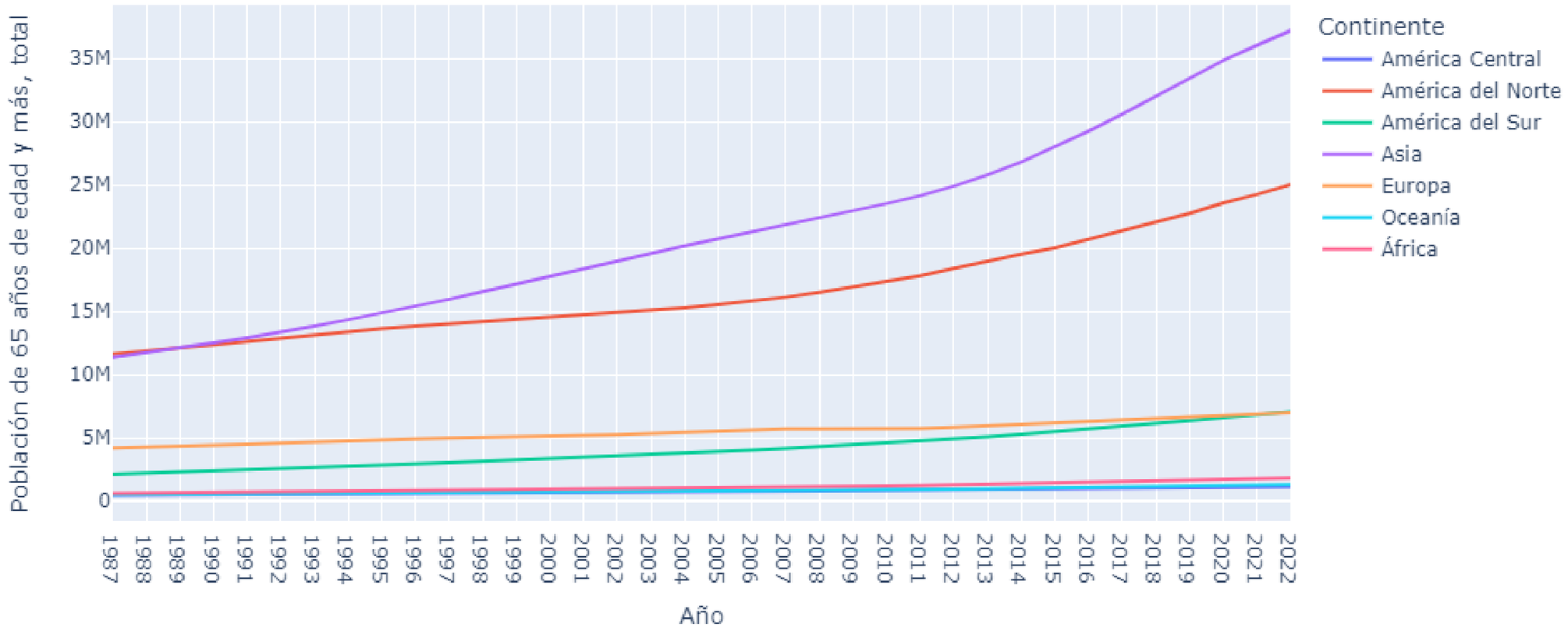
EDA EXHAUSTIVO



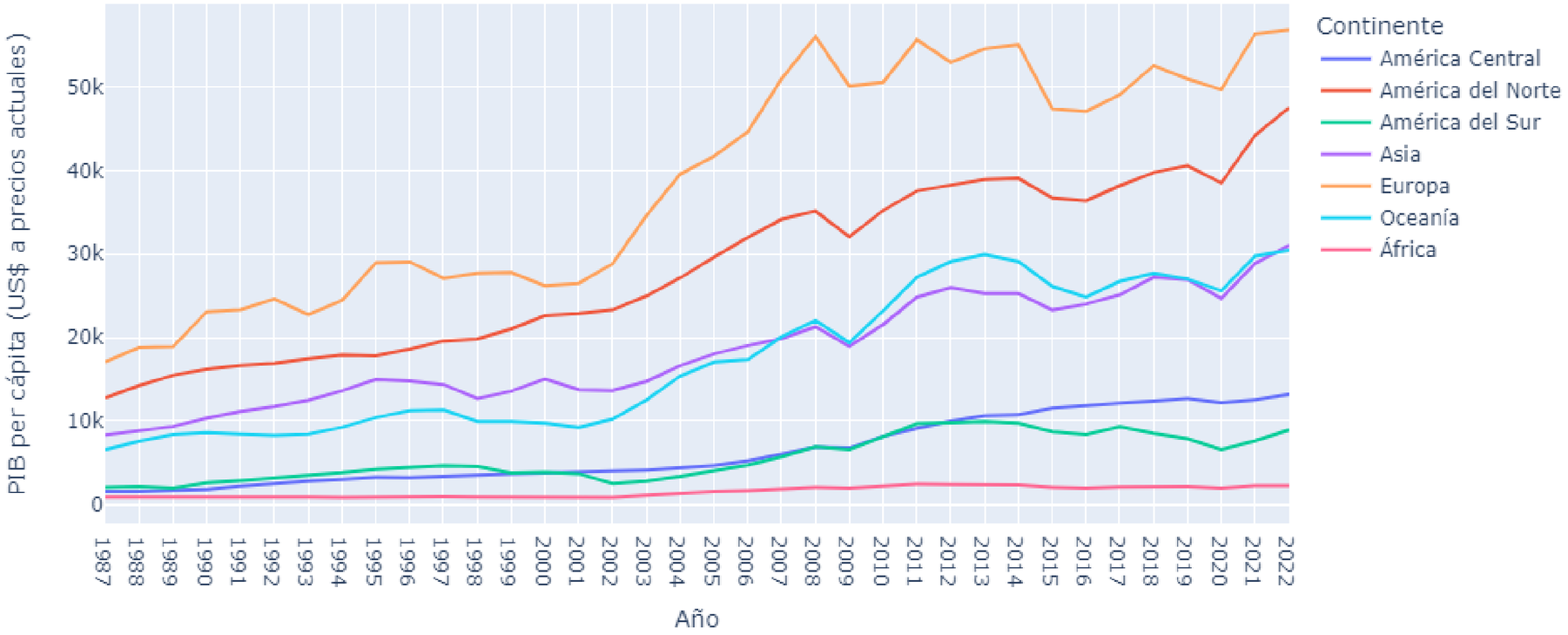
EDA EXHAUSTIVO



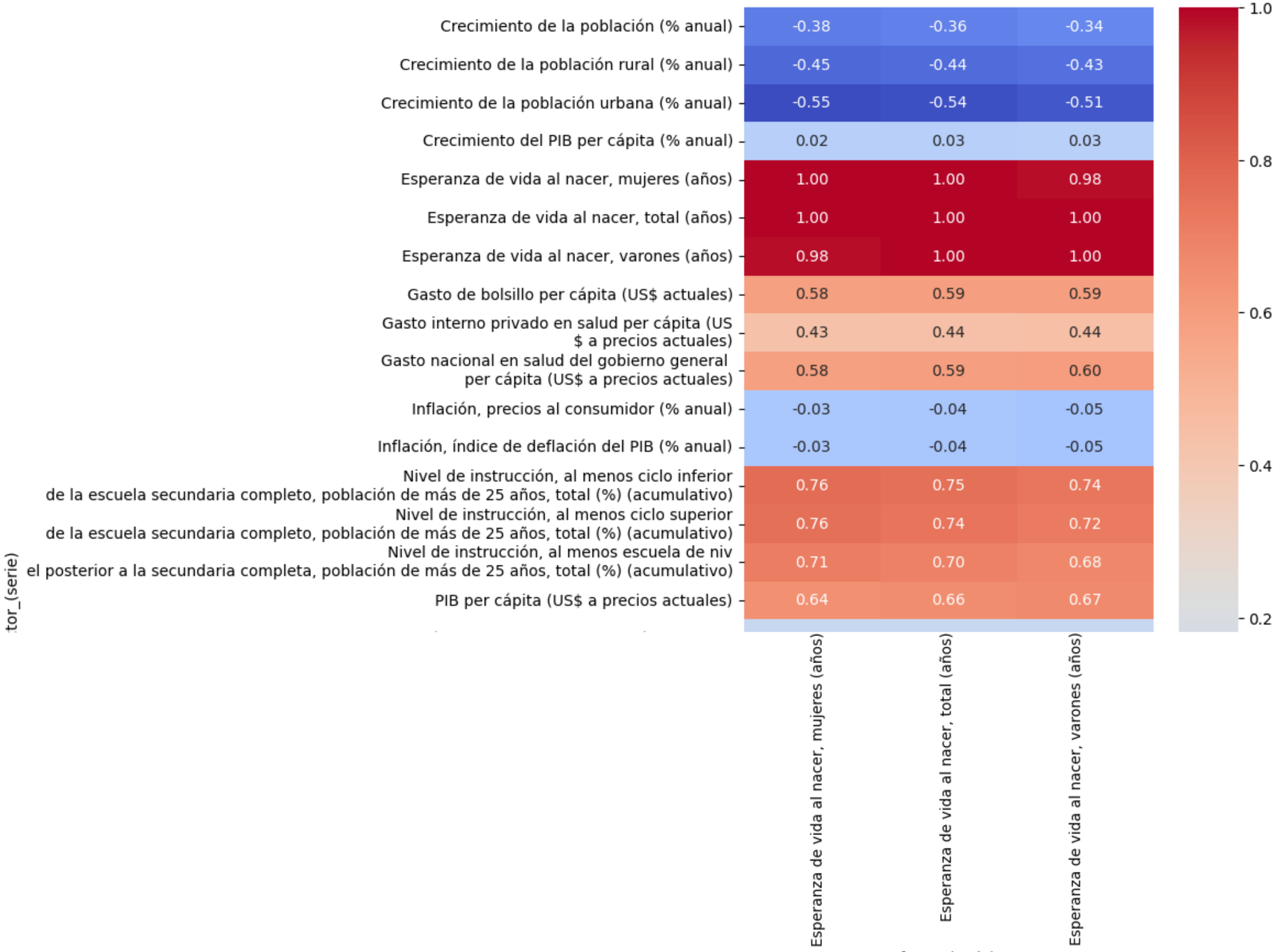
EDA EXHAUSTIVO



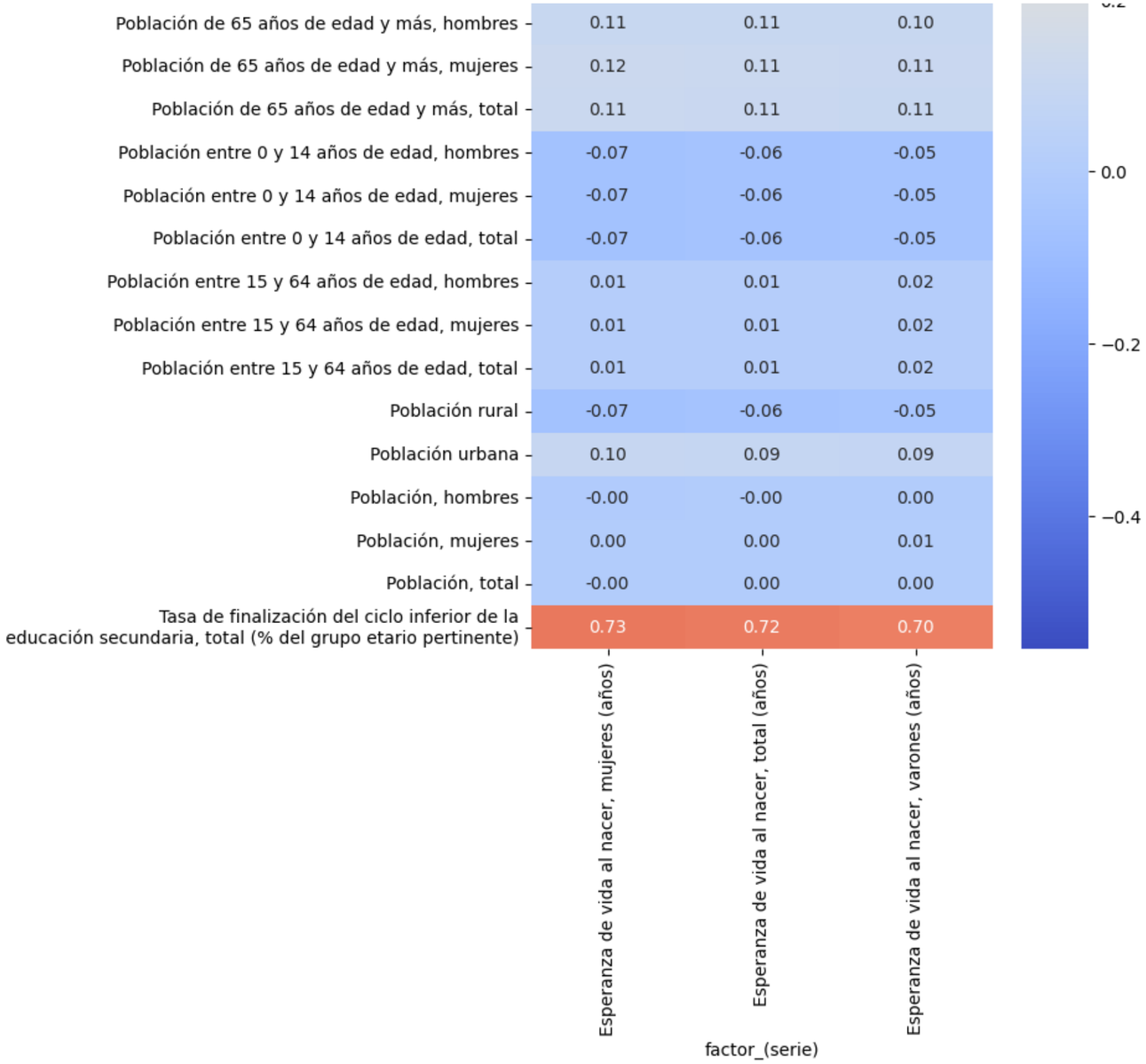
EDA EXHAUSTIVO



EDA EXHAUSTIVO



EDA EXHAUSTIVO



TRANSFORMACIONES

Selección de indicadores

Indicador_name	Total
Population ages 0-14, male	1584
Population ages 0-14, female	1584
Urban population	1584
Rural population	1584
Population, total	1584
Population, male	1584
Population, female	1584
Population ages 65 and above, total	1584
Population ages 65 and above, male	1584
Population ages 65 and above, female	1584
Population ages 15-64, total	1584
Population ages 15-64, male	1584
Population ages 15-64, female	1584
Population ages 0-14, total	1584
Urban population growth (annual %)	1584
Population growth (annual %)	1583
GDP per capita (current US$)	1551
Inflation, GDP deflator (annual %)	1548
GDP per capita growth (annual %)	1548
Rural population growth (annual %)	1548
Life expectancy at birth, total (years)	1540
Life expectancy at birth, male (years)	1540
Life expectancy at birth, female (years)	1540
Inflation, consumer prices (annual %)	1405
Domestic private health expenditure per capita (current US$)	924
Out-of-pocket expenditure per capita (current US$)	924
Domestic general government health expenditure per capita (current US$)	924
Lower secondary completion rate, total (% of relevant age group)	739
Educational attainment, at least completed upper secondary, population 25+, total (%) (cumulative)	404
Educational attainment, at least completed lower secondary, population 25+, total (%) (cumulative)	370
Educational attainment, at least completed post-secondary, population 25+, total (%) (cumulative)	233

Nos quedaremos con aquellos que tengan más de 1000 valores no nulos

TRANSFORMACIONES

Manejo de valores nulos

Pais	Año	Indicador	Valor	Indicador_name
SGP	1987	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1988	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1989	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1990	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1991	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1992	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1993	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1994	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1995	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1996	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1997	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1998	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)
SGP	1999	SP.RUR.TOTL.ZG	0.0	Rural population growth (annual %)

Singapur no cuenta con población rural, por lo que reemplazamos los valores NaN por 0

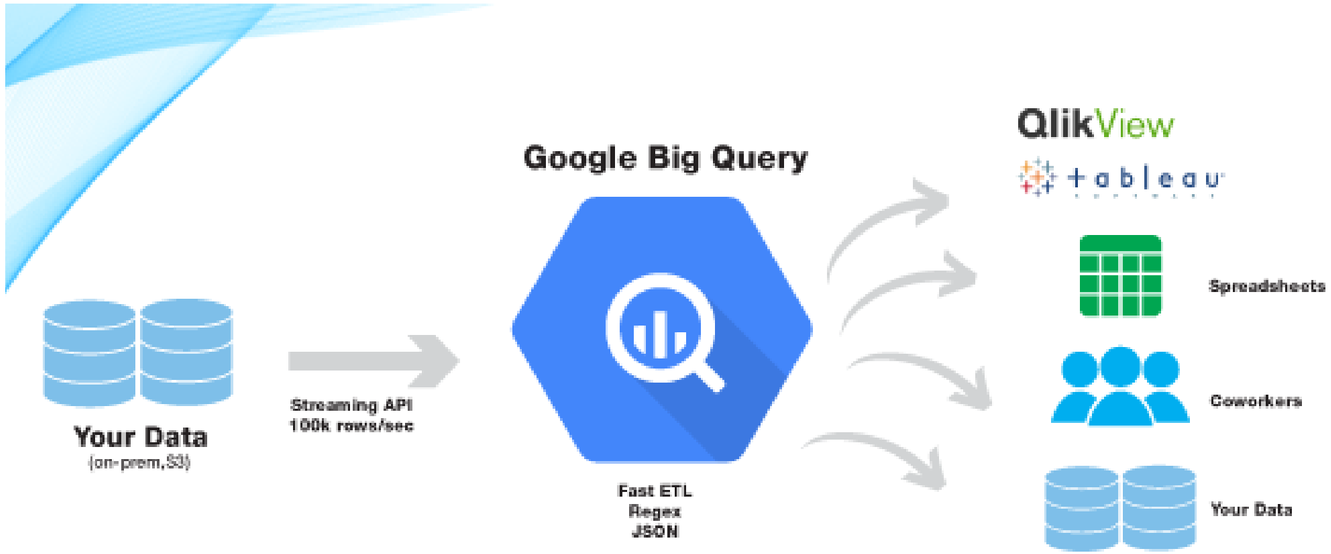
TRANSFORMACIONES

Pais	Año	Indicador	Valor	Indicador_name
AFG	1987	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1988	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1989	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1990	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1991	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1992	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1993	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1994	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1995	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1996	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1997	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1998	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	1999	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	2000	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	2001	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	2002	NY.GDP.PCAP.KD.ZG	NaN	GDP per capita growth (annual %)
AFG	2003	NY.GDP.PCAP.KD.ZG	0.927029	GDP per capita growth (annual %)
AFG	2004	NY.GDP.PCAP.KD.ZG	-2.497255	GDP per capita growth (annual %)

Para ciertas combinaciones de país - indicador existen demasiados valores nulos, reemplazarlos todos por la media no sería bueno por lo que eliminamos esos registros

BigQuery

- Almacenamiento y análisis de grandes cantidades de datos
- Integración con aprendizaje automático
- Visualización de datos
- BigQuery no es una base de datos relacional convencional



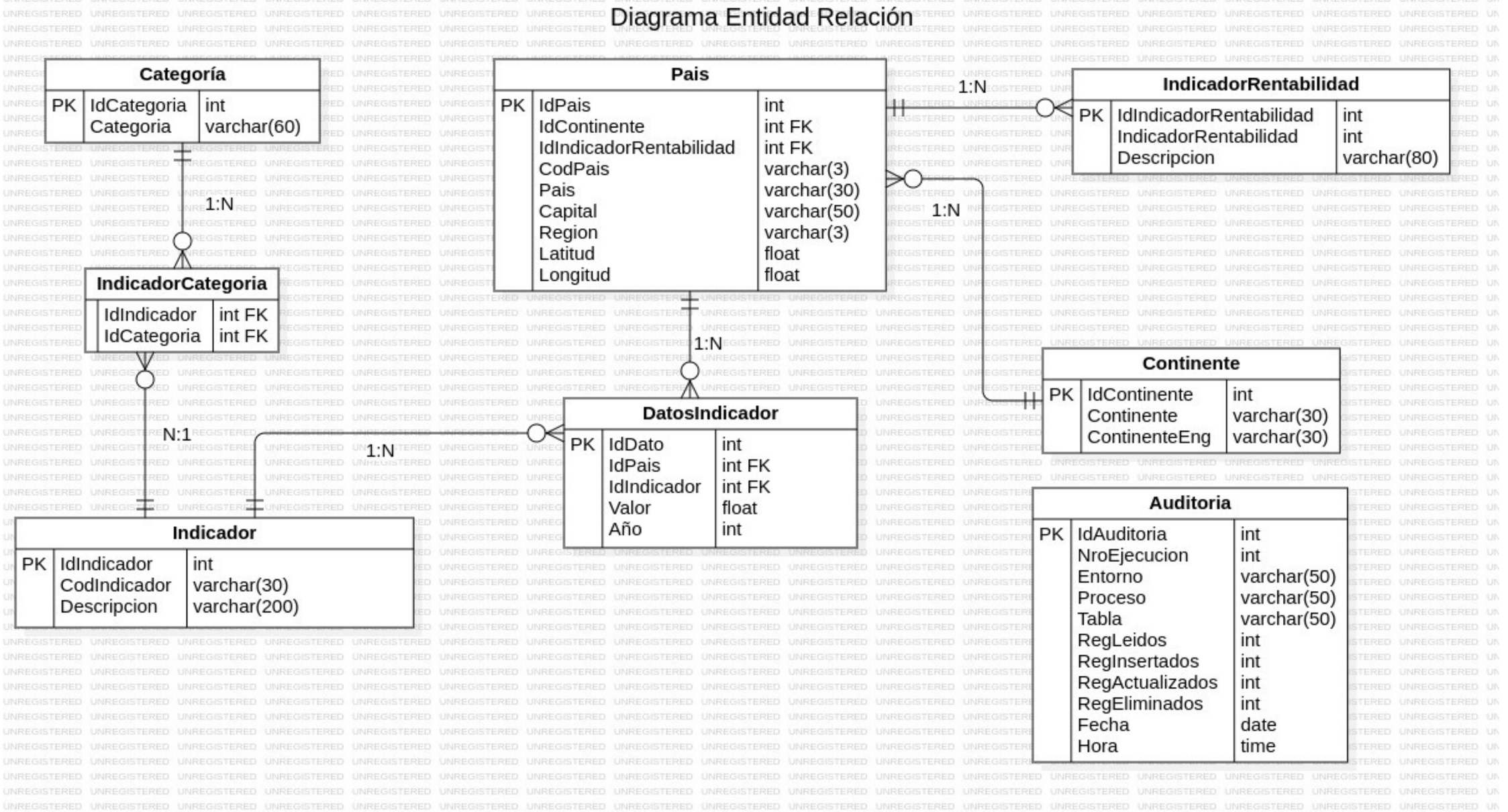
What is BigQuery?

Looker
PART OF Google Cloud

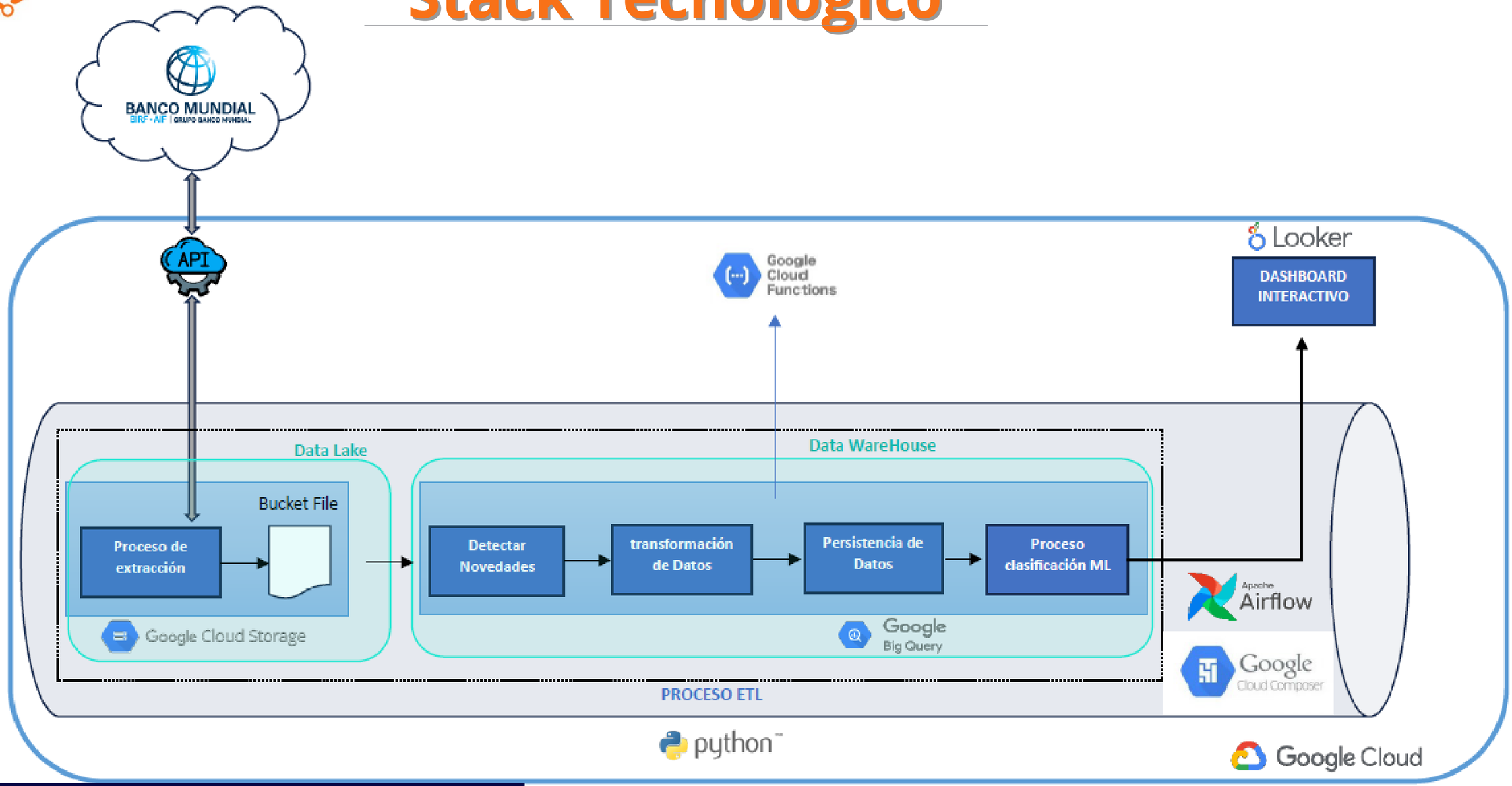
Google BigQuery

Google Cloud Storage

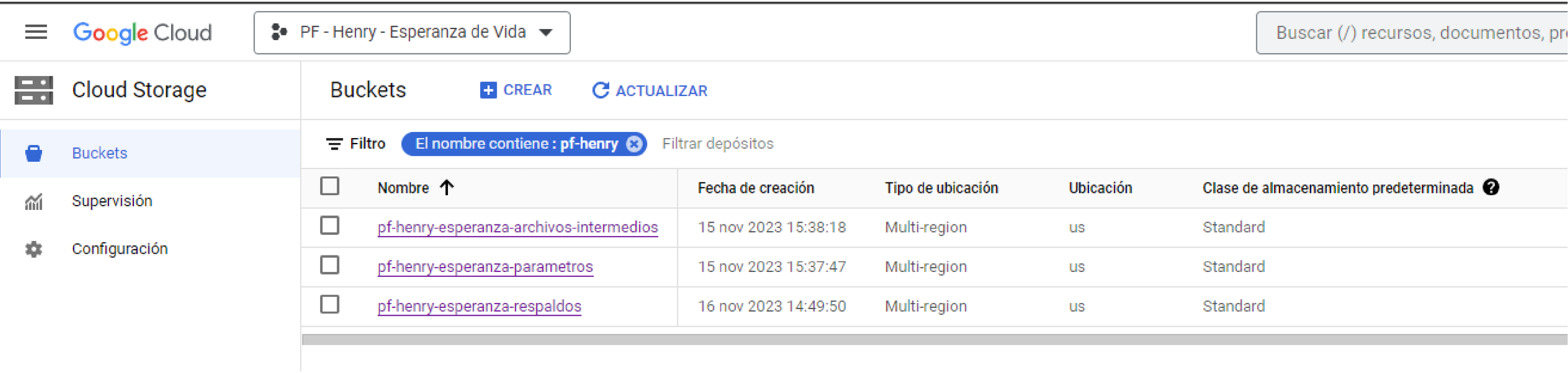
Diagrama Entidad Relación



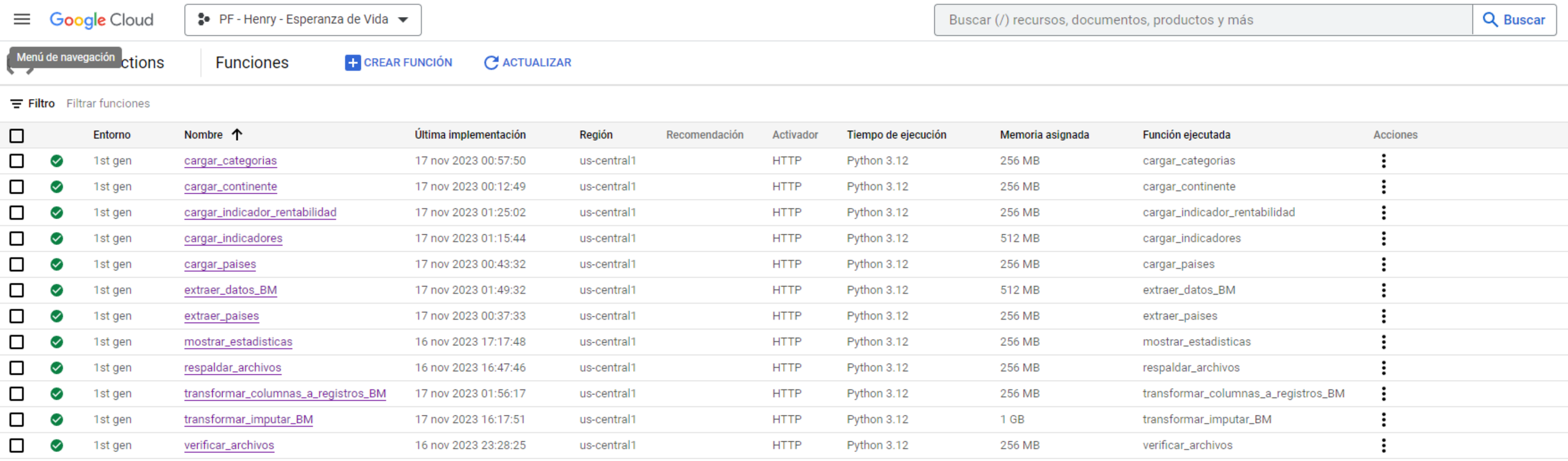
Stack Tecnológico



Buckets


Google Cloud
PF - Henry - Esperanza de Vida
Buscar (/) recursos, documentos, pr
Cloud Storage
Buckets
Supervisión
Configuración
Buckets
CREAR
ACTUALIZAR
Filtro
El nombre contiene : pf-henry
Filtrar depósitos
Nombre
Fecha de creación
Tipo de ubicación
Ubicación
Clase de almacenamiento predeterminada
pf-henry-esperanza-archivos-intermedios
15 nov 2023 15:38:18
Multi-region
us
Standard
pf-henry-esperanza-parametros
15 nov 2023 15:37:47
Multi-region
us
Standard
pf-henry-esperanza-respaldos
16 nov 2023 14:49:50
Multi-region
us
Standard

Cloud Functions



Entorno	Nombre	Última implementación	Región	Recomendación	Activador	Tiempo de ejecución	Memoria asignada	Función ejecutada	Acciones
1st gen	cargar_categorias	17 nov 2023 00:57:50	us-central1		HTTP	Python 3.12	256 MB	cargar_categorias	
1st gen	cargar_continente	17 nov 2023 00:12:49	us-central1		HTTP	Python 3.12	256 MB	cargar_continente	
1st gen	cargar_indicador_rentabilidad	17 nov 2023 01:25:02	us-central1		HTTP	Python 3.12	256 MB	cargar_indicador_rentabilidad	
1st gen	cargar_indicadores	17 nov 2023 01:15:44	us-central1		HTTP	Python 3.12	512 MB	cargar_indicadores	
1st gen	cargar_paises	17 nov 2023 00:43:32	us-central1		HTTP	Python 3.12	256 MB	cargar_paises	
1st gen	extraer_datos_BM	17 nov 2023 01:49:32	us-central1		HTTP	Python 3.12	512 MB	extraer_datos_BM	
1st gen	extraer_paises	17 nov 2023 00:37:33	us-central1		HTTP	Python 3.12	256 MB	extraer_paises	
1st gen	mostrar_estadisticas	16 nov 2023 17:17:48	us-central1		HTTP	Python 3.12	256 MB	mostrar_estadisticas	
1st gen	respaldar_archivos	16 nov 2023 16:47:46	us-central1		HTTP	Python 3.12	256 MB	respaldar_archivos	
1st gen	transformar_columnas_a_registros_BM	17 nov 2023 01:56:17	us-central1		HTTP	Python 3.12	256 MB	transformar_columnas_a_registros_BM	
1st gen	transformar_imputar_BM	17 nov 2023 16:17:51	us-central1		HTTP	Python 3.12	1 GB	transformar_imputar_BM	
1st gen	verificar_archivos	16 nov 2023 23:28:25	us-central1		HTTP	Python 3.12	256 MB	verificar_archivos	

DAG

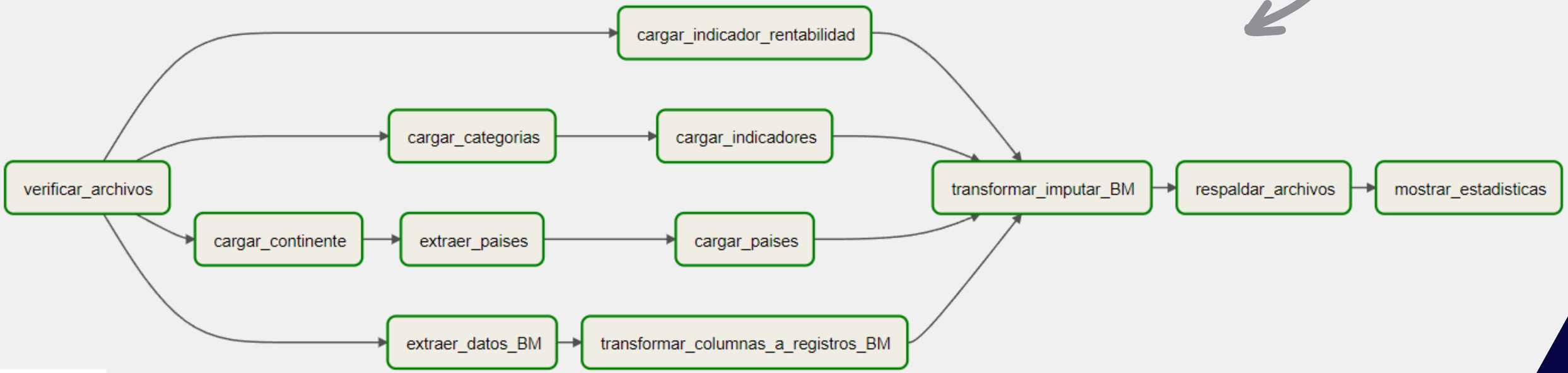
Compositor | Entornos | CREAR | BORRAR

Filtro Filter environments

Estado	Nombre	Ubicación	Versión de Composer	Versión de Airflow	Fecha y hora de creación	Hora de actualización	Webserver de Airflow	Lista de los DAG	Registros	Carpeta de DAG	Etiquetas
✓	composer-pf-henry-esperanza	us-central1	1.20.12	2.4.3	15/11/23, 09:16	15/11/23, 09:33	Airflow	DAG	Registros	DAG	Ninguno

Apache Airflow

Bash Operator



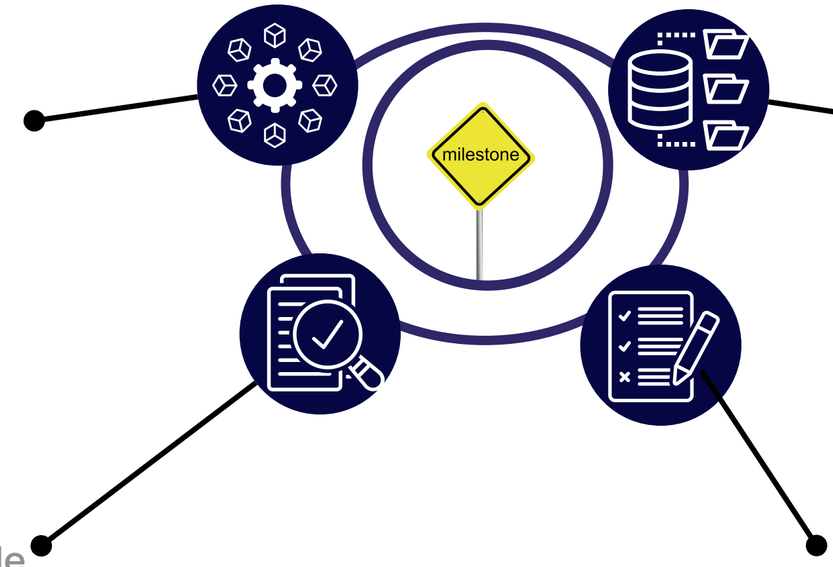
Google Cloud Composer

Hitos del SPRINT # 2

DATA ENGINEERING

Automatización de Pipeline

Automatizacion del proceso de ETL



Persistencia en DW

Definición del modelo de BBDD para persistir los datos en nuestro DW

EDA Exhaustivo

Informe resultante del análisis profundo de los datos.

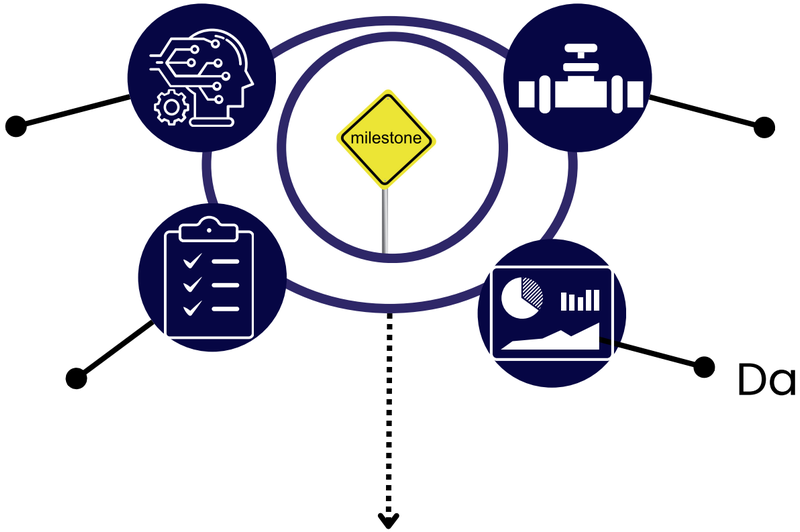
Optimización de Arquitectura

Se mejoraron los recursos utilizados en el Cloud

Próximos Pasos

DATA ENGINEERING

ML Operativo



MLOps en Pipeline

Apache Airflow

Google Cloud Composer

Test Integral

DashBoard Interactivo

Looker

Google Cloud

MVP Dashboard

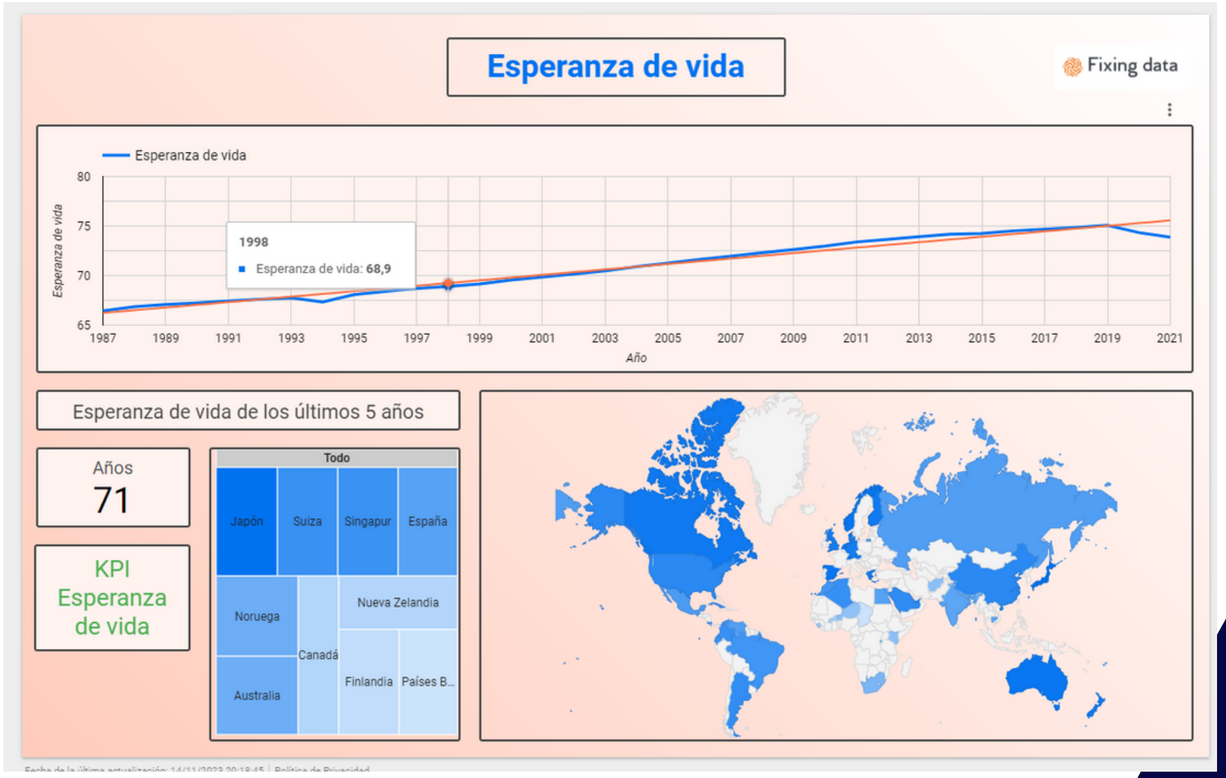
Looker Studio

01 Prototipo de dashboard

02 Panel interactivo

03 Datos provisorios

04 Análisis de una muestra



Conclusión

SPRINT #2

- Proceso escalable
- Saber fácilmente donde y que modificar
- Entregables de calidad profesional

SPRINT #3

- Implementación del modelo de ML
- Dasboard definitivo

Fixing data

Data Driven Results Given

MUCHAS GRACIAS

Por Tu Atención!

Fixing data

Data Driven Results Given

Preguntas?