Laboratorio 7: Integración de Datos y Extracción de Insights

Integrantes

- Abby Donis - 22440
- Mathew Cordero - 22982
- Josué Say - 220801

Ejercicio 1 - Integración de datos con una herramienta para procesos de ETL

Usamos Postgres (SQL), Knime (Gestor de conexiones) y MongoCompasDB (NoSQL) para esta parte del ejercicio y la información de como ejecutar el programa se puede ver en el siguiente enlace del **repositorio**.

1.1 Ingeste los datos que se encuentran en la base de datos relacional. Revise si es necesario realizar algún tipo de limpieza sobre los datos

Limpieza Para ello si se realizo limpieza con los scripts en la carpeta

lab7/

```
  DatosSQL/                 # CSVs de población y envejecimiento
  DatosNoSQL/               # JSONs con datos turísticos y Big Mac
  Parte1/                   # Código ETL e imágenes de resultados
     sql_clenaer.py              # Script de limpieza NOSQL
```

Se ejecuta con python y debe de cambiarse las propiedades

- Cambio de variables

```
DB_HOST = "localhost"  # Cambia si el servidor no está en tu máquina
DB_PORT = "5432"       # Puerto por defecto de PostgreSQL
DB_NAME = "nombre_db"
DB_USER = "usuario_db"
DB_PASSWORD = "contraseña"
TABLE_NAME = "nombre_tabla"
CSV_FILE = "./path_del_csv"
```

- Ejecucion

```
python sql_clenaer.py
```

- Se limpian los csv y se ingresan de una vez a postgres tambien

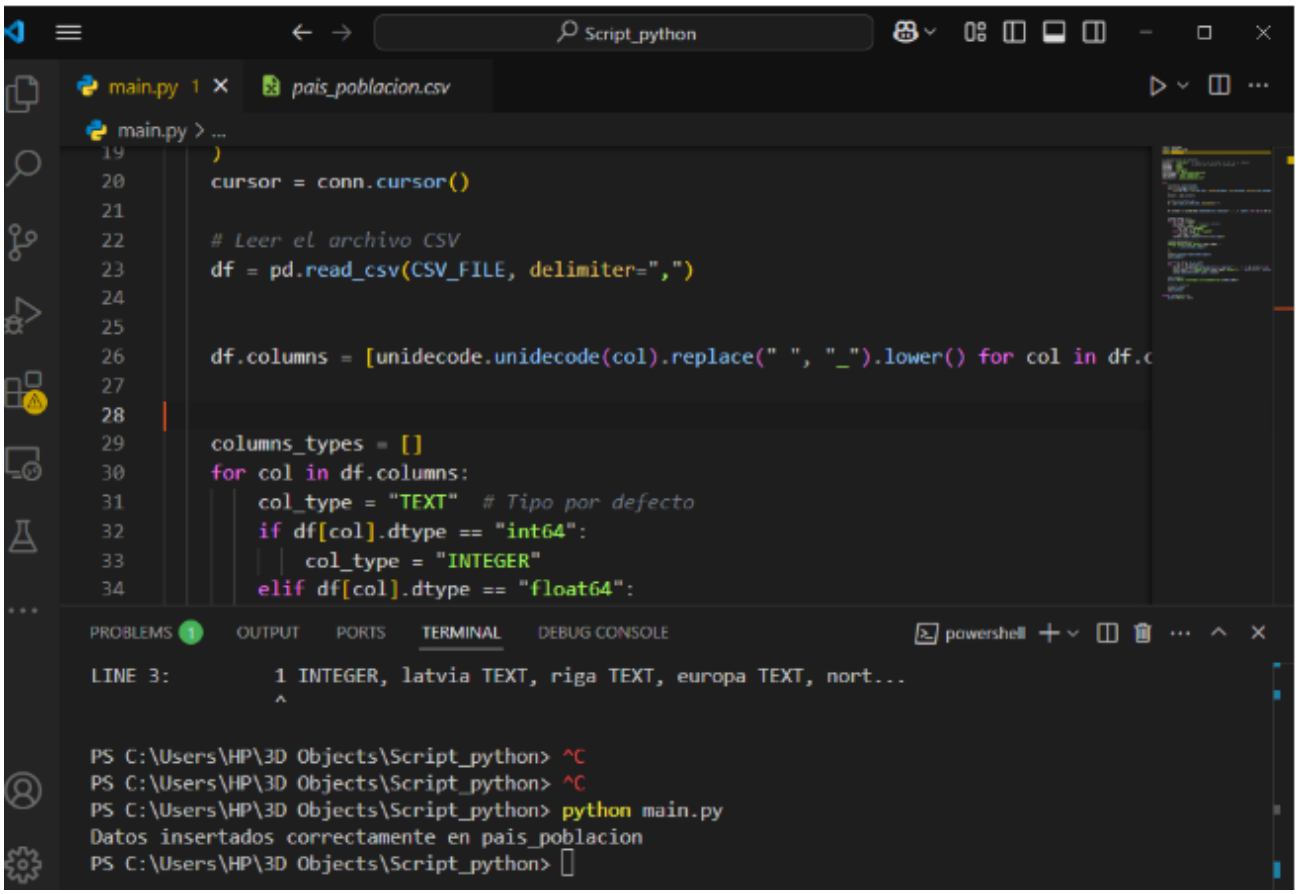


Figure 1: Limpieza

- Luego se ingreso en DDL de PgAdmin

1.2 Ingeste los datos que se encuentran en la base de datos no relacional. Revise si es necesario realizar algún tipo de limpieza sobre los datos

Limpieza de datos Esta en

lab7/

```
DatosSQL/                # CSVs de población y envejecimiento
DatosNoSQL/              # JSONs con datos turísticos y Big Mac
Parte1/                  # Código ETL e imágenes de resultados
    json_cleaner.py          # Script de limpieza NOSQL
```

- Cambio de variables

```
json_path = "./json_path"
```

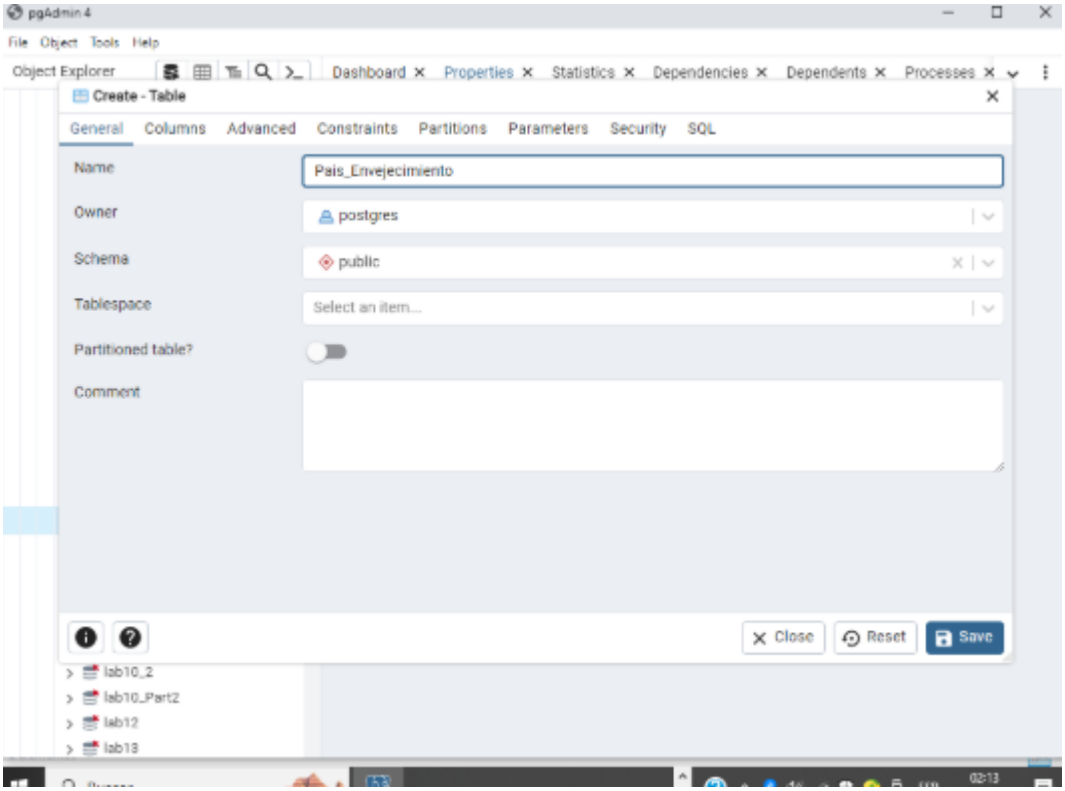


Figure 2: Carga de datos

- Ejecucion

  ```
  python json_cleaner.py
  ```

1.3 Integre ambas fuentes de datos por medio de la herramienta de procesos de ETL

- Para esto se uso Knime, con un conector de Postgres y MongoDB
- Este fue el diagrama hecho, en knime juntando los json en una sola db y haciendo uso de json to table
- Luego se hizo un concatenate y un join con las otras base de datos SQL
- Aqui la salida de la db completamente limpia y junta

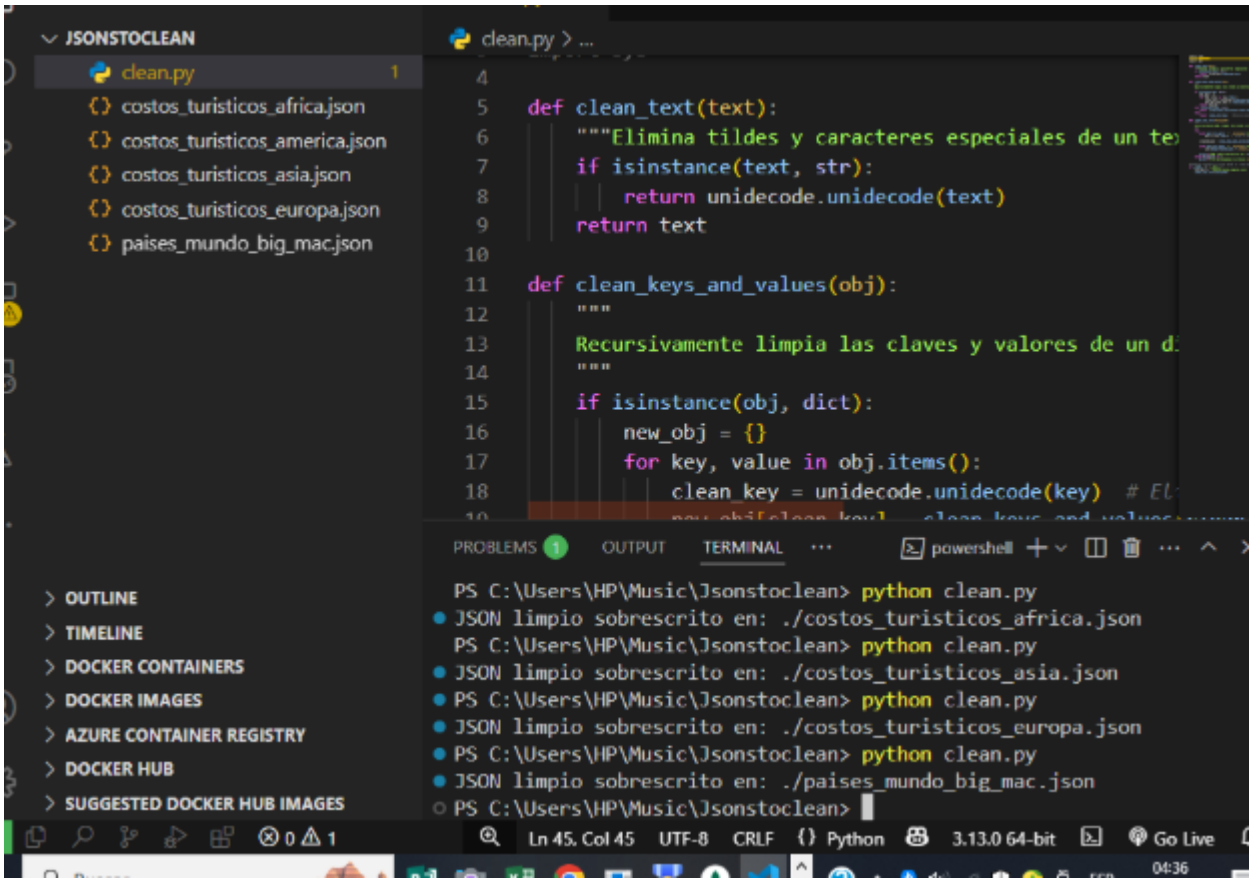


Figure 3: Limpieza

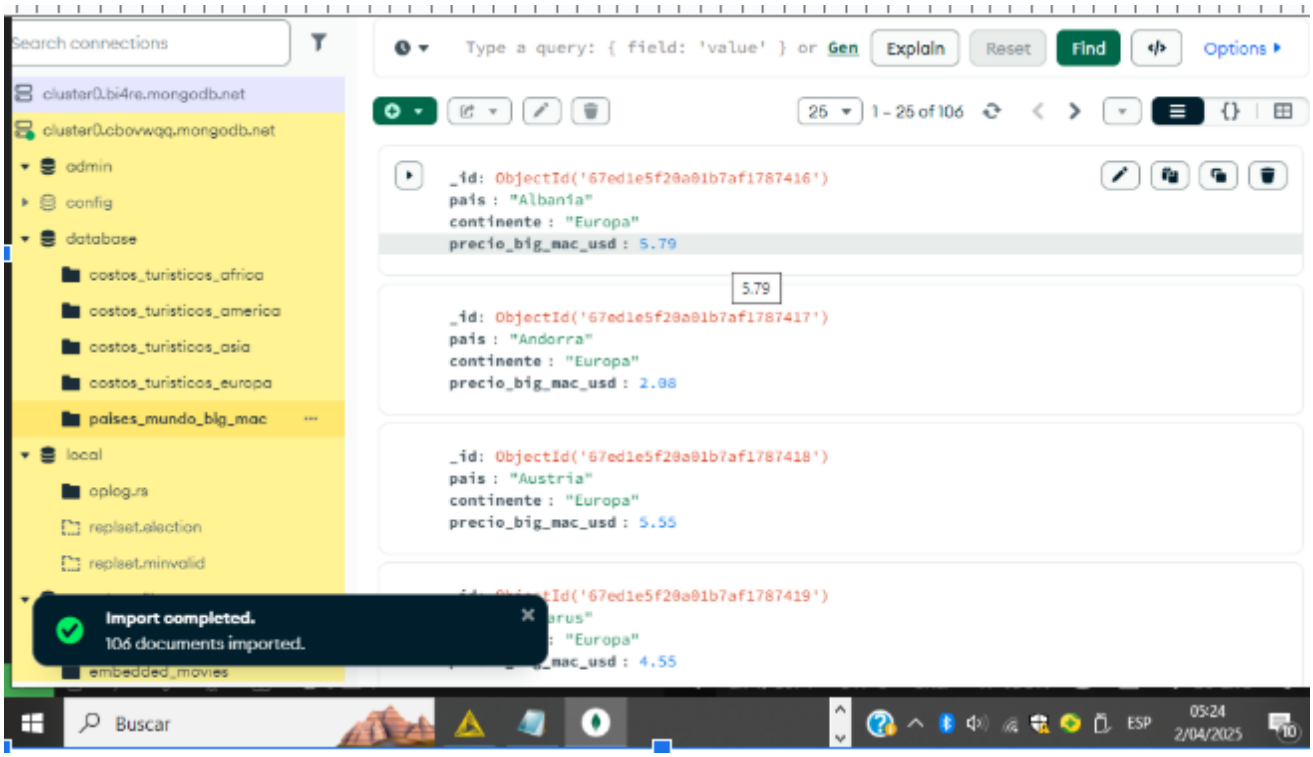


Figure 4: Carga de Datos

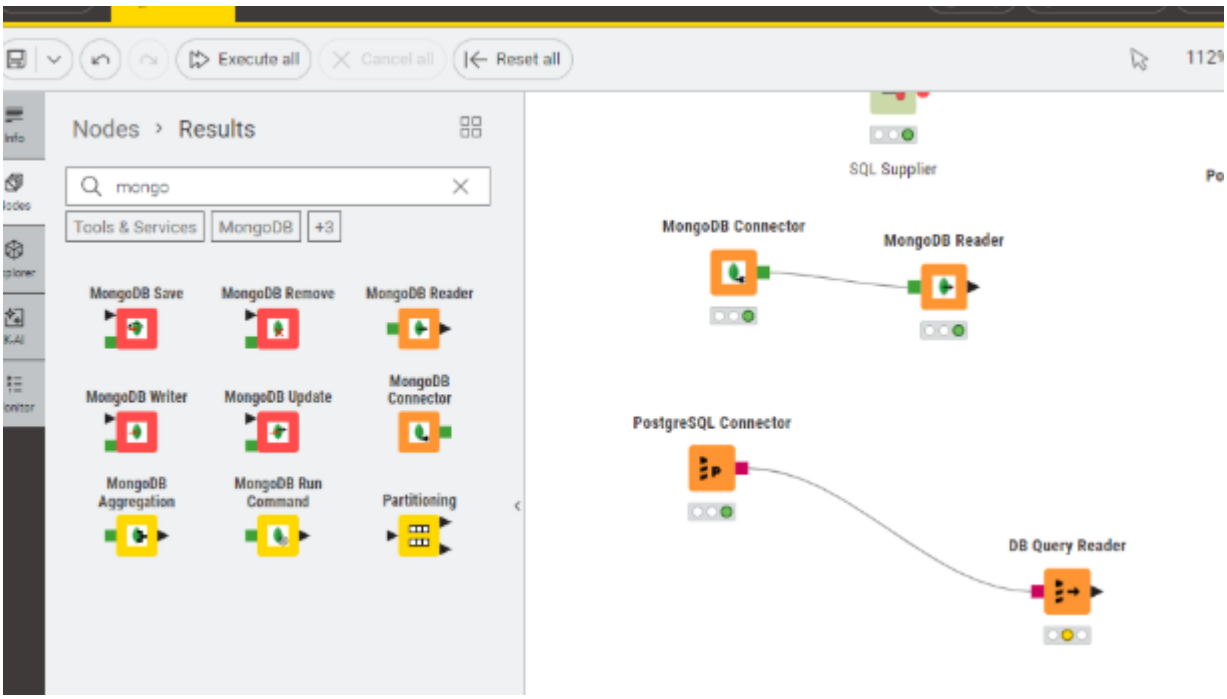


Figure 5: Uso de Knime

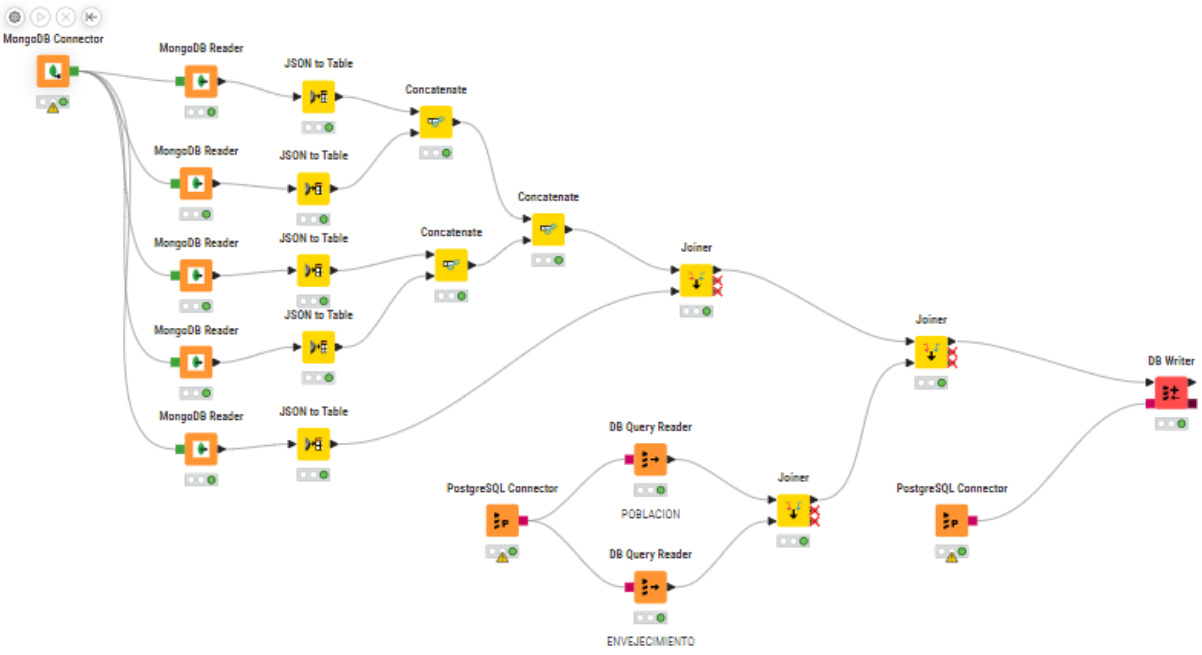
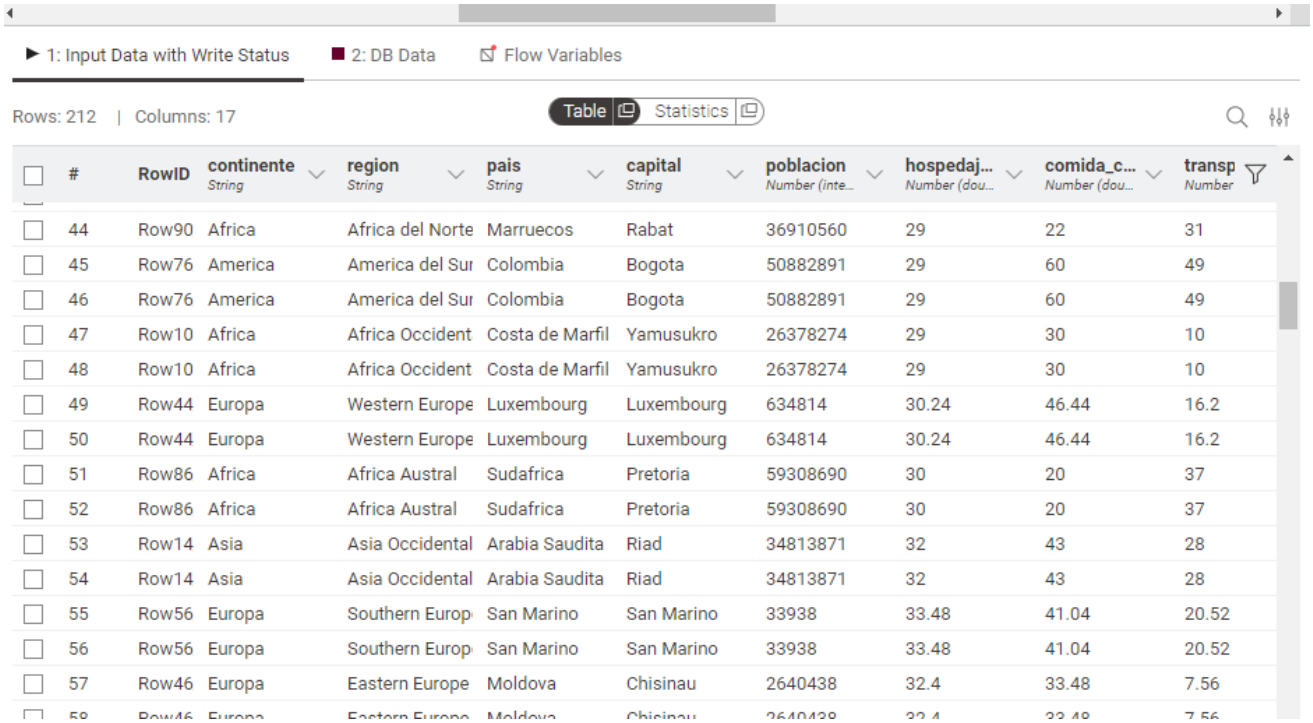


Figure 6: Uso de Knime 2



▶ 1: Input Data with Write Status ■ 2: DB Data Flow Variables

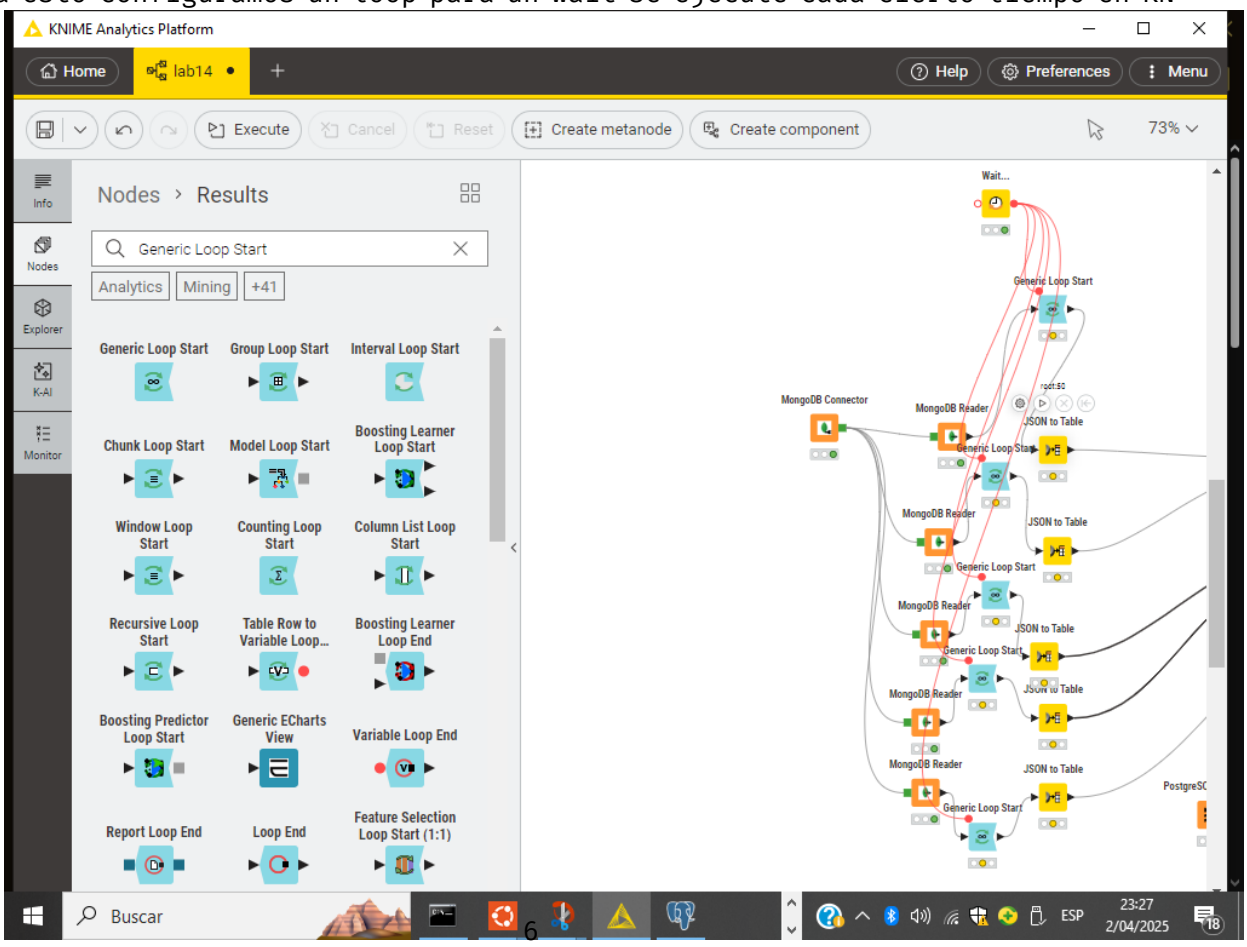
Rows: 212 | Columns: 17 Table Statistics

	#	RowID	continente String	region String	pais String	capital String	poblacion Number (inte...	hospedaj... Number (dou...	comida_c... Number (dou...	transp Number
☐	44	Row90	Africa	Africa del Norte	Marruecos	Rabat	36910560	29	22	31
☐	45	Row76	America	America del Sur	Colombia	Bogota	50882891	29	60	49
☐	46	Row76	America	America del Sur	Colombia	Bogota	50882891	29	60	49
☐	47	Row10	Africa	Africa Occident	Costa de Marfil	Yamusukro	26378274	29	30	10
☐	48	Row10	Africa	Africa Occident	Costa de Marfil	Yamusukro	26378274	29	30	10
☐	49	Row44	Europa	Western Europe	Luxembourg	Luxembourg	634814	30.24	46.44	16.2
☐	50	Row44	Europa	Western Europe	Luxembourg	Luxembourg	634814	30.24	46.44	16.2
☐	51	Row86	Africa	Africa Austral	Sudafrica	Pretoria	59308690	30	20	37
☐	52	Row86	Africa	Africa Austral	Sudafrica	Pretoria	59308690	30	20	37
☐	53	Row14	Asia	Asia Occidental	Arabia Saudita	Riad	34813871	32	43	28
☐	54	Row14	Asia	Asia Occidental	Arabia Saudita	Riad	34813871	32	43	28
☐	55	Row56	Europa	Southern Europ	San Marino	San Marino	33938	33.48	41.04	20.52
☐	56	Row56	Europa	Southern Europ	San Marino	San Marino	33938	33.48	41.04	20.52
☐	57	Row46	Europa	Eastern Europe	Moldova	Chisinau	2640438	32.4	33.48	7.56

Figure 7: Resultados Knime

1.4 Configure la herramienta para que el proceso de ETL se ejecute cada cierto tiempo (la frecuencia de ejecución queda a su criterio)

- Para esto configuramos un loop para un wait se ejecute cada cierto tiempo en KNIME



1.4 Los datos integrados se deberán cargar en la base de datos que hace las veces de data warehouse, sin que se necesite su intervención

Esto se valida y se termina el loop cuando variable condition osea los row de los insertdos son mayores a 0.

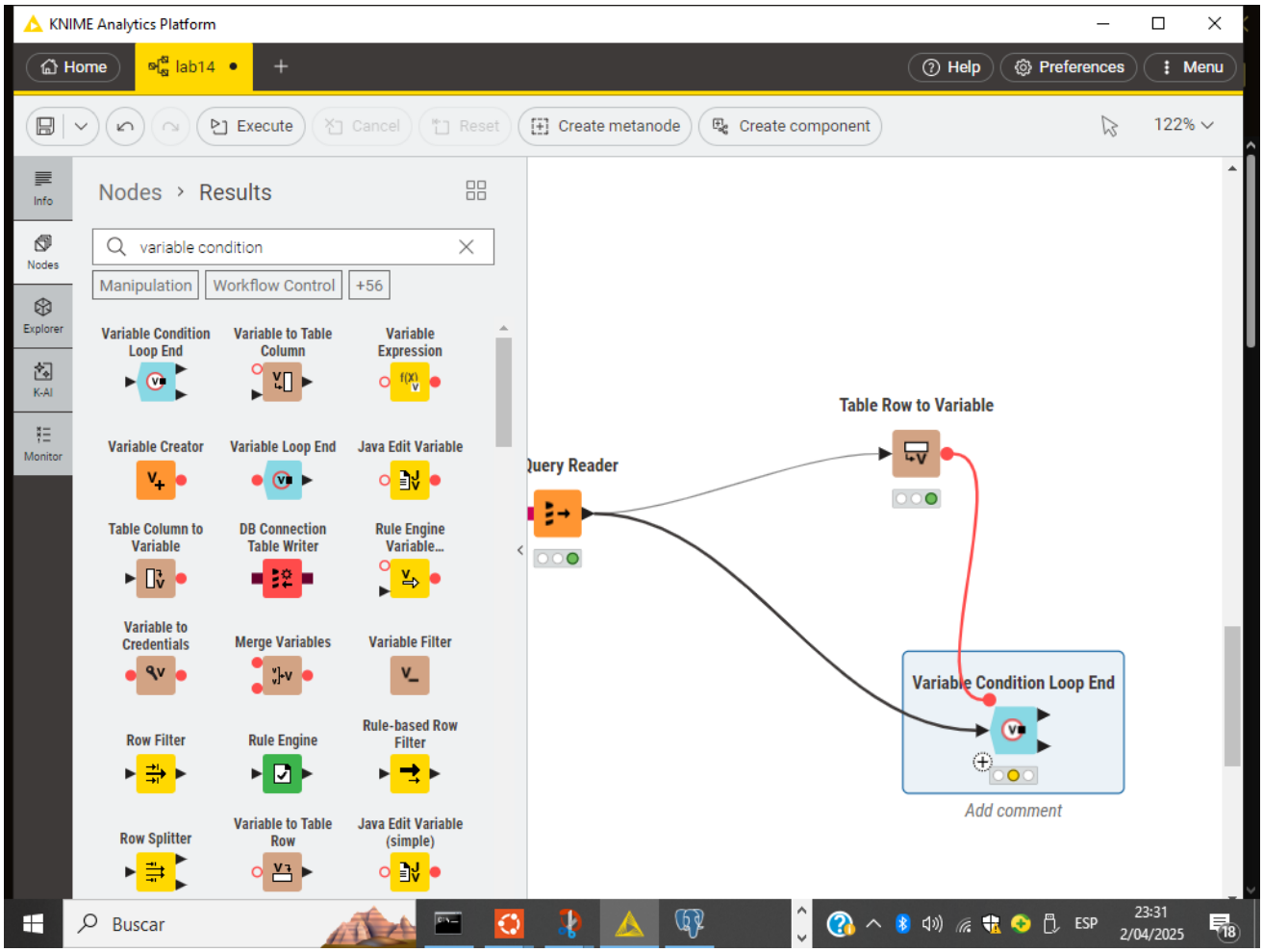
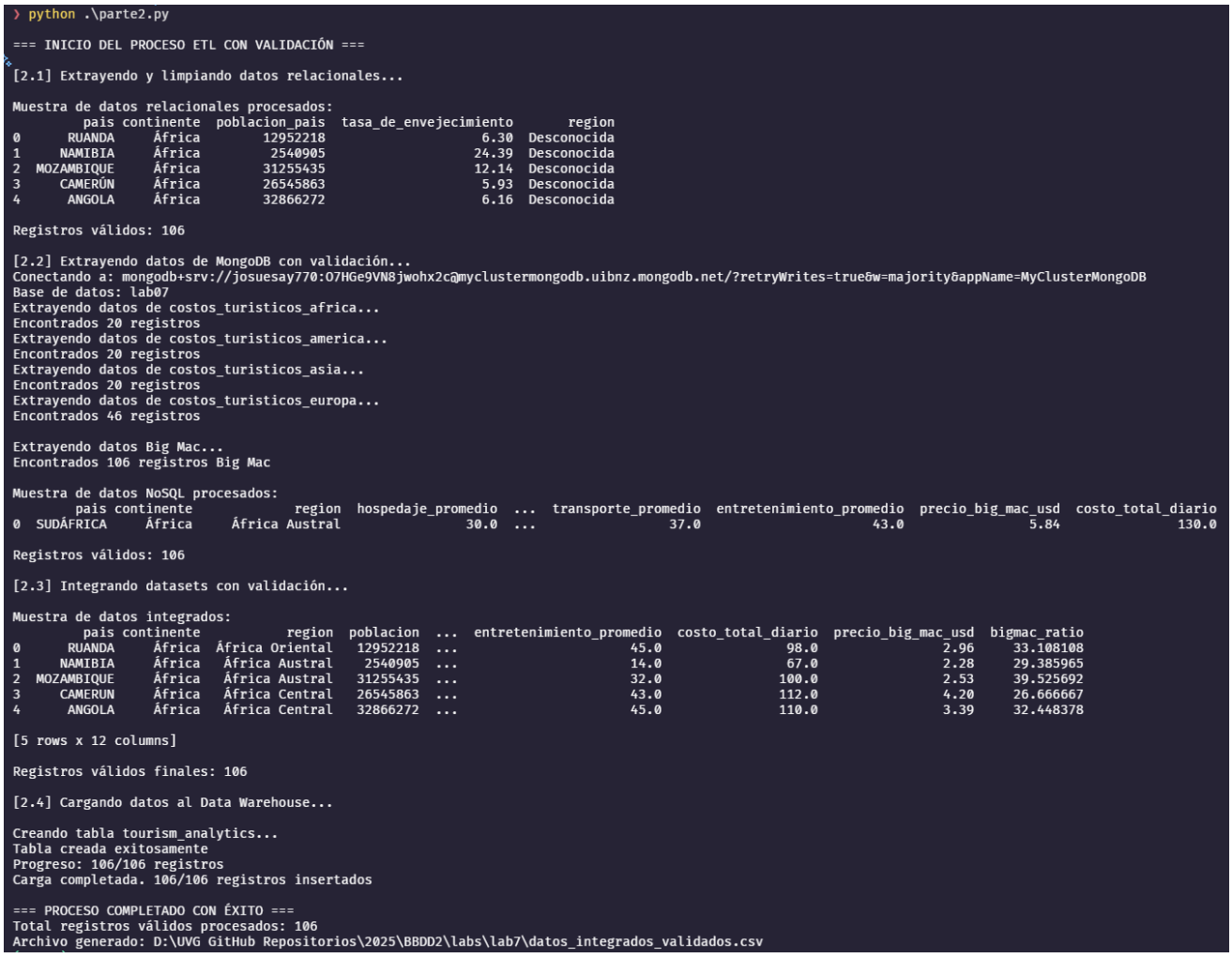


Figure 8: Integración Knime

Ejercicio 2 - Integración de datos con un lenguaje de programación

Se usó python para esta parte del ejercicio y la información de como ejecutar el programa se puede ver en el siguiente enlace del **repositorio**. Dicho ejercicio obtiene este resultado:

También la data csv retornada por el código adjuntada en la entrega o al ejecutar el código de esta parte del ejercicio.



```
❯ python .\parte2.py

=== INICIO DEL PROCESO ETL CON VALIDACIÓN ===

[2.1] Extrayendo y limpiando datos relacionales...

Muestra de datos relacionales procesados:
        pais continente  poblacion_pais  tasa_de_envejecimiento       region
0      RUANDA     África        12952218                    6.30  Desconocida
1     NAMIBIA     África         2540905                   24.39  Desconocida
2  MOZAMBIQUE     África        31255435                   12.14  Desconocida
3     CAMERÚN     África        26545863                    5.93  Desconocida
4      ANGOLA     África        32866272                    6.16  Desconocida

Registros válidos: 106

[2.2] Extrayendo datos de MongoDB con validación...
Conectando a: mongodb+srv://josuesay770:O7HGe9VN8jwohx2c@myclustermongodb.uibnz.mongodb.net/?retryWrites=true&w=majority&appName=MyClusterMongoDB
Base de datos: lab07
Extrayendo datos de costos_turisticos_africa...
Encontrados 20 registros
Extrayendo datos de costos_turisticos_america...
Encontrados 20 registros
Extrayendo datos de costos_turisticos_asia...
Encontrados 20 registros
Extrayendo datos de costos_turisticos_europa...
Encontrados 46 registros

Extrayendo datos Big Mac...
Encontrados 106 registros Big Mac

Muestra de datos NoSQL procesados:
       pais continente           region  hospedaje_promedio  ...  transporte_promedio  entretenimiento_promedio  precio_big_mac_usd  costo_total_diario
0  SUDÁFRICA     África     África Austral                30.0  ...                 37.0                      43.0                5.84               130.0

Registros válidos: 106

[2.3] Integrando datasets con validación...

Muestra de datos integrados:
        pais continente          region  poblacion  ...  entretenimiento_promedio  costo_total_diario  precio_big_mac_usd  bigmac_ratio
0      RUANDA     África  África Oriental   12952218  ...                      45.0                98.0                2.96     33.108108
1     NAMIBIA     África   África Austral    2540905  ...                      14.0                67.0                2.28     29.385965
2  MOZAMBIQUE     África   África Austral   31255435  ...                      32.0               100.0                2.53     39.525692
3     CAMERUN     África   África Central   26545863  ...                      43.0               112.0                4.20     26.666667
4      ANGOLA     África   África Central   32866272  ...                      45.0               110.0                3.39     32.448378

[5 rows x 12 columns]

Registros válidos finales: 106

[2.4] Cargando datos al Data Warehouse...

Creando tabla tourism_analytics...
Tabla creada exitosamente
Progreso: 106/106 registros
Carga completada. 106/106 registros insertados

=== PROCESO COMPLETADO CON ÉXITO ===
Total registros válidos procesados: 106
Archivo generado: D:\UVG GitHub Repositorios\2025\BBDD2\labs\lab7\datos_integrados_validados.csv
```

Figure 9: Resultado Parte 2

Ejercicio 3 - Insights sobre los datos

Una vez integrada la información proveniente de la base relacional y no relacional, se deben extraer conocimientos útiles que impulsen la acción. La información de como ejecutar el programa se puede ver en el siguiente enlace del **repositorio**

Insight 1: Países con alta tasa de envejecimiento no necesariamente tienen altos costos turísticos

- **Evidencia encontrada:**
  Al analizar los 10 países con mayor **tasa de envejecimiento**, se observa que el **costo promedio de hospedaje y comida** varía significativamente. Por ejemplo:
  - Namibia (24.39% de envejecimiento) tiene un **costo promedio muy bajo** (USD 18.50).
  - Montenegro (24.31%) tiene un **costo alto** (USD 54.54).
  - Honduras (23.15%) muestra un **costo más alto** que EE. UU. con menor tasa.

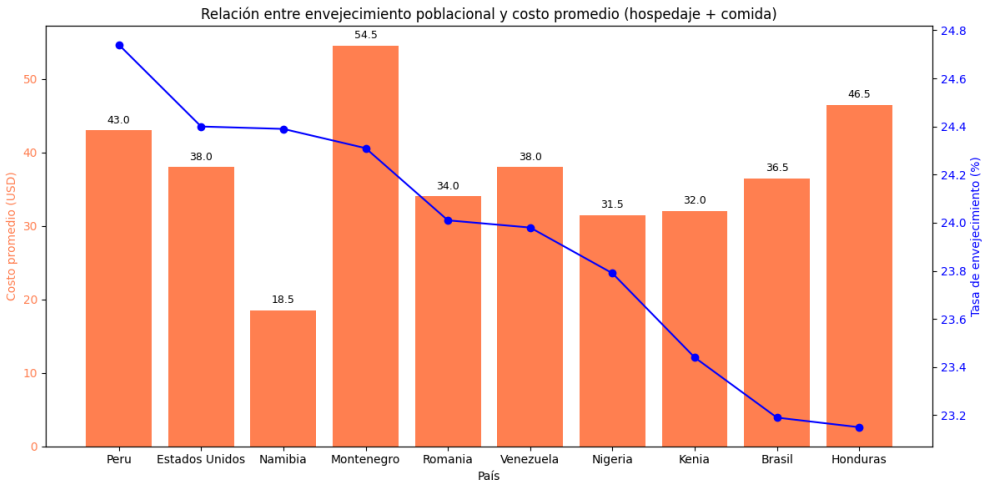


Figure 10: Insight 1

- **Relevancia:**
  Este hallazgo **rompe la suposición** de que los países más envejecidos tienen mayor costo de vida o turismo. Esto es clave para estrategias de turismo inclusivo y accesible, especialmente si se piensa en atraer adultos mayores a destinos con infraestructura adecuada pero menor costo.
- **Recomendación:**
  Se recomienda **segmentar campañas turísticas dirigidas a adultos mayores** hacia países como **Namibia o Venezuela**, que combinan alta tasa de envejecimiento y costos bajos. Esto puede ser atractivo para turistas que buscan destinos tranquilos, accesibles y adecuados a sus necesidades.

Insight 2: Alta disparidad entre el costo de comida local y el precio del Big Mac

- **Evidencia encontrada:**
  Se identificaron países donde el costo promedio de la comida local supera los **USD 50 diarios**, pero el **precio del Big Mac está por debajo de los USD 3**, como por ejemplo:
  - **Guatemala:** comida local ≈ 72 USD / Big Mac: 2.83 USD
  - **Perú:** 63 USD / 2.81 USD
  - **El Salvador:** 60 USD / 2.32 USD
  - **Irán:** 53 USD / 1.62 USD

  La gráfica muestra una marcada diferencia entre las barras (costo local) y la línea (precio Big Mac), evidenciando la falta de correlación directa.

- **Relevancia:**
  Esta disparidad puede deberse a varios factores:
  - Diferencias entre el **mercado formal y el informal.**
  - **Subsidios o regulación de precios** para marcas globales.
  - **Estrategias comerciales** de McDonald's u otras cadenas para mantener precios competitivos.

  Es una señal de que el Big Mac no siempre refleja fielmente el costo de vida local, lo cual es clave al usarlo como indicador económico.

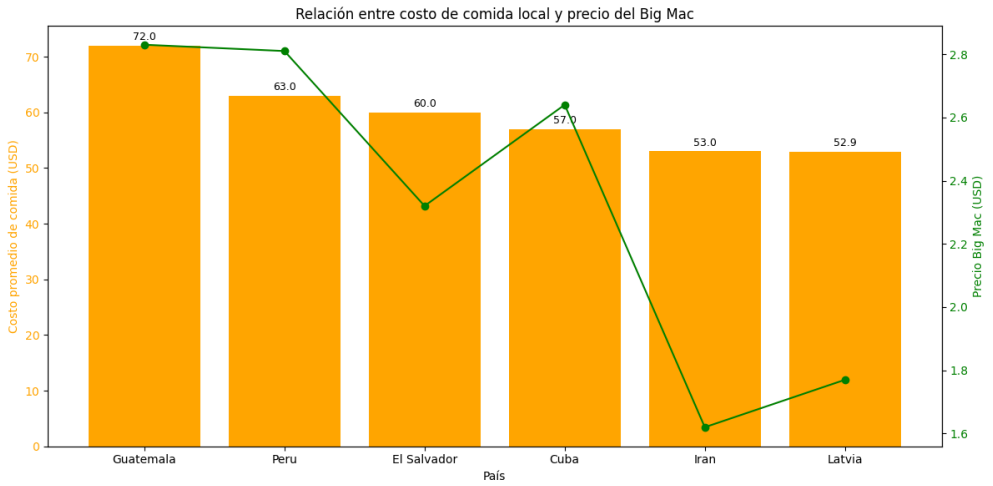


Figure 11: Insight 2

- **Recomendación:**
  Evitar utilizar el **Índice Big Mac** como única referencia para evaluar el costo de vida entre países. Aunque es una herramienta popular por su simplicidad, los resultados muestran que **el precio del Big Mac puede no reflejar fielmente la realidad económica local**, especialmente en países con:

- **Distorsiones de mercado**, como subsidios o regulaciones específicas.
- **Estrategias comerciales globales** que mantienen precios homogéneos a pesar del contexto económico.
- **Altos niveles de informalidad**, donde los costos de productos locales no se alinean con los de cadenas internacionales.

Esto limita su precisión y puede llevar a interpretaciones erróneas en análisis financieros, decisiones de inversión o estudios comparativos de poder adquisitivo.

Insight 3: Países con bajo envejecimiento poblacional y bajo costo de entretenimiento

- **Evidencia encontrada:**
  Al analizar los países con **tasa de envejecimiento menor al 10%** y **costos de entretenimiento menores a USD 25**, encontramos destinos como:
  - **Grecia** (8.92% envejecimiento / USD 10.80 entretenimiento)
  - **Hungría y Eslovaquia** (≈8.1% / USD 19.44)
  - **Paraguay** (5.51% / USD 21.00)

  La gráfica muestra cómo estos países combinan dos factores relevantes: una **población mayoritariamente joven** y **actividades recreativas accesibles.**

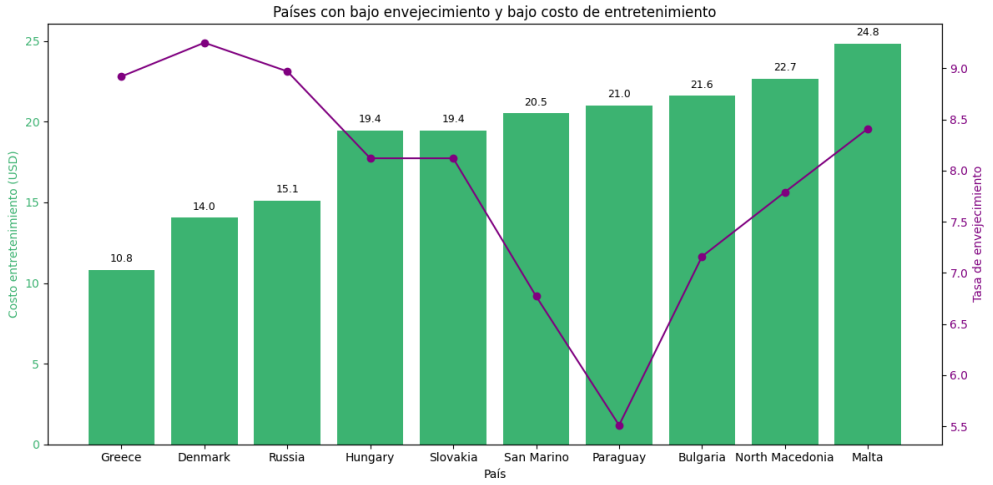


Figure 12: Insight 3

- **Relevancia:**
  Este grupo de países representa destinos potencialmente atractivos para:
  - Jóvenes turistas que buscan lugares con **ambiente dinámico y bajo costo de vida.**
  - Inversionistas o emprendedores del sector ocio/cultura que deseen ingresar

a mercados **menos saturados** pero con buena relación entre infraestructura y costos.

- **Recomendación:**
  Se recomienda enfocar **estrategias turísticas, culturales o tecnológicas** hacia estos países, aprovechando la combinación de:
  - Población joven (mayor adopción tecnológica y demanda de entretenimiento)
  - Costos accesibles (mejora de margen para productos/servicios)