Práctica Oracle

Self Service Data Preparation

Presentado Por:

Jose David Angulo

Arnau Recio

Albert Ripoll

Profesor

Ramon Caihuelas | ramon.caihuelas@salle.url.edu | Tecnologías BI 2

Máster Universitario en Ingeniería de Datos Masivos (Big Data)

Universidad La Salle - Ramón Llull

TABLA DE CONTENIDO

Selección de las fuentes de datos “Data Set, BD o Ficheros”

Para los datos a analizar hemos seleccionado la fuente de “United Nations Population Division, New York, World Population Prospects: The 2022 Revision, last accessed July 2022”. De dicha fuente, se extrajeron 5 bases de datos históricas (discontinuas) que describen características de la mayoría de países del mundo, y fueron:

- Población, superficie y densidad
- PIB y PIB per cápita
- Gasto público en educación
- Crecimiento demográfico, fertilidad, esperanza de vida y mortalidad
- Fuerza laboral y desempleo

Cada base de datos contiene los siguientes campos:

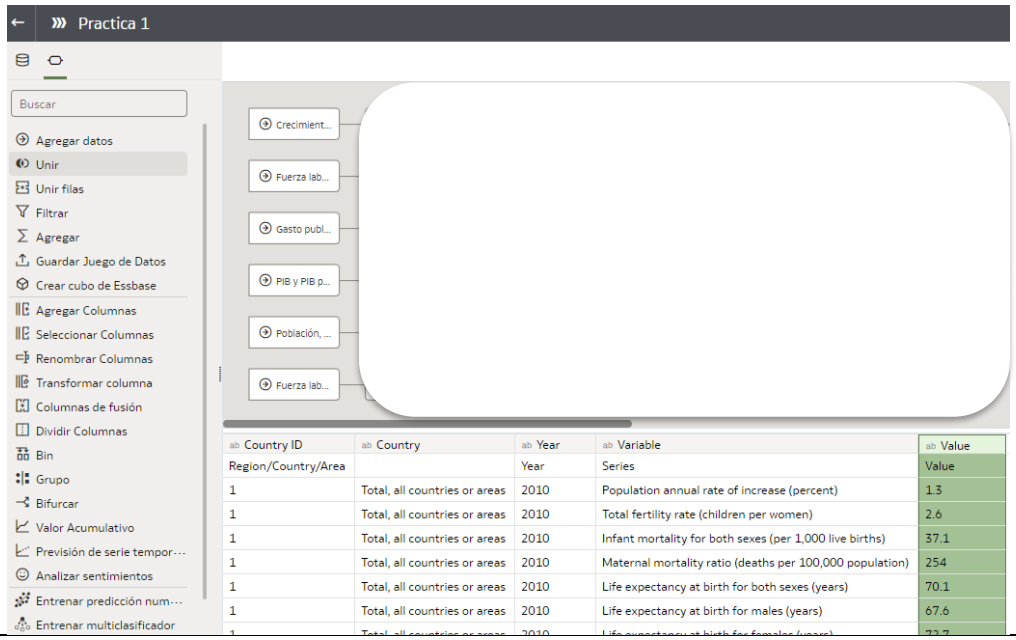
- **Region/Country/Area number (Country ID):** Tiene el código perteneciente al país.
- **Region/Country/Area name (Country):** El nombre del país.
- **Year (Year):** Serie de tiempo (año) con el cual se ha medido la “Serie” o variable.
- **Series (Variable):** Variable que se está midiendo con relación al año.
- **Value (Value):** Valor de la serie que se ha medido para ese año.
- **Footnotes:** Anotaciones sobre la medición (no se usará)
- **Source:** Fuente de la medición (no se usará)

Viendo los tipos de series que tenemos, seleccionamos las siguientes variables de interés:

- Life expectancy at birth for both sexes (years)
- Labour force participation - Total
- Public expenditure on education (% of GDP)
- GDP per capita (US dollars)
- Population density
- Unemployment rate - Total

Cargue de los datos

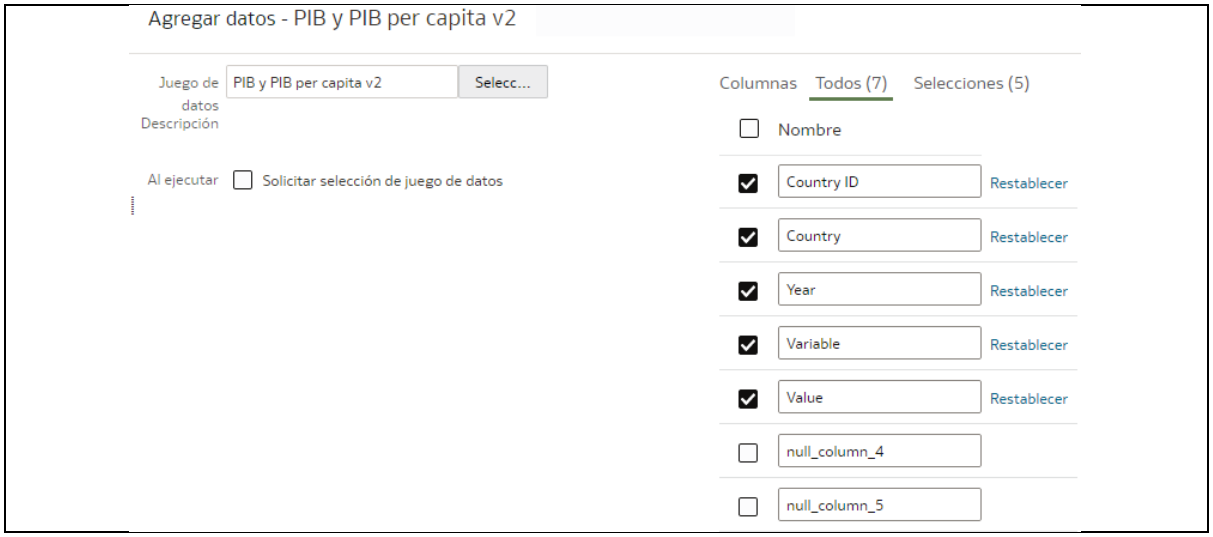
Se realiza el proceso de cargue o lectura de los diferentes Data Set (ficheros, BD) que se escogieron, así:



Country ID	Country	Year	Variable	Value
Region/Country/Area		Year	Series	Value
1	Total, all countries or areas	2010	Population annual rate of increase (percent)	1.3
1	Total, all countries or areas	2010	Total fertility rate (children per women)	2.6
1	Total, all countries or areas	2010	Infant mortality for both sexes (per 1,000 live births)	37.1
1	Total, all countries or areas	2010	Maternal mortality ratio (deaths per 100,000 population)	254
1	Total, all countries or areas	2010	Life expectancy at birth for both sexes (years)	70.1
1	Total, all countries or areas	2010	Life expectancy at birth for males (years)	67.6

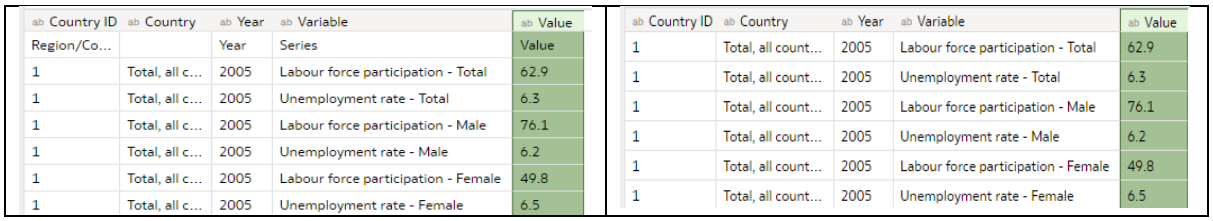
De todos los data set se excluyeron del cargue las últimas dos columnas (4 y 5) porque no aportaban valor para los casos de estudio posteriores. Ejemplo 1: "Data Set Fuerza Laboral"

De todos los data set se excluyeron del cargue las últimas dos columnas (4 y 5) porque no aportaban valor para los casos de estudio posteriores. Ejemplo 2: "Data Set PIB per cápita"



Transformación de los datos

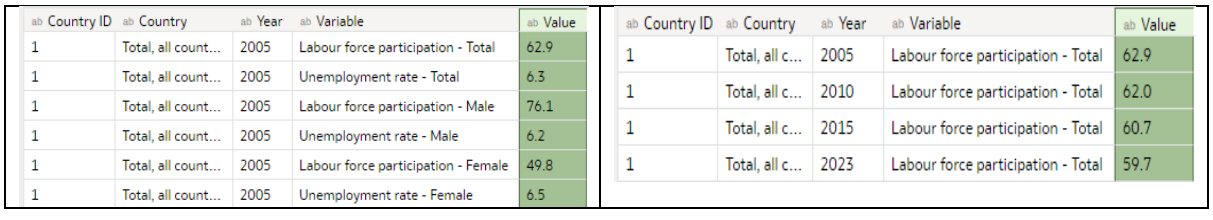
En cada data set hay dos filas iniciales como encabezados. Oracle lee la primera fila como encabezado y la segunda fila como dato. Para que detecte la segunda fila como encabezado y se elimine la primera, aplicamos un filtro. Seleccionamos el country ID e indicamos al filtro que excluya la que tenga como nombre Región/Country/Area.



Country ID	Country	Year	Variable	Value
Region/Co...		Year	Series	Value
1	Total, all c...	2005	Labour force participation - Total	62.9
1	Total, all c...	2005	Unemployment rate - Total	6.3
1	Total, all c...	2005	Labour force participation - Male	76.1
1	Total, all c...	2005	Unemployment rate - Male	6.2
1	Total, all c...	2005	Labour force participation - Female	49.8
1	Total, all c...	2005	Unemployment rate - Female	6.5

Country ID	Country	Year	Variable	Value
1	Total, all count...	2005	Labour force participation - Total	62.9
1	Total, all count...	2005	Unemployment rate - Total	6.3
1	Total, all count...	2005	Labour force participation - Male	76.1
1	Total, all count...	2005	Unemployment rate - Male	6.2
1	Total, all count...	2005	Labour force participation - Female	49.8
1	Total, all count...	2005	Unemployment rate - Female	6.5

Para seleccionar las variables que queremos, aplicamos un nuevo filtro que incluya las variables queridas dentro de cada data set (base de datos). También observamos que no es posible seleccionar más de una variable por cada data set, ya que queda en forma de fila (caso data set de Fuerza laboral y desempleo). Entonces primero hicimos una transformación que incluía un filtro que nos seleccionó las variables deseadas y nos eliminaba las sobrantes.



Country ID	Country	Year	Variable	Value
1	Total, all count...	2005	Labour force participation - Total	62.9
1	Total, all count...	2005	Unemployment rate - Total	6.3
1	Total, all count...	2005	Labour force participation - Male	76.1
1	Total, all count...	2005	Unemployment rate - Male	6.2
1	Total, all count...	2005	Labour force participation - Female	49.8
1	Total, all count...	2005	Unemployment rate - Female	6.5

Country ID	Country	Year	Variable	Value
1	Total, all c...	2005	Labour force participation - Total	62.9
1	Total, all c...	2010	Labour force participation - Total	62.0
1	Total, all c...	2015	Labour force participation - Total	60.7
1	Total, all c...	2023	Labour force participation - Total	59.7

Luego tuvimos que cargar el data set “Fuerza laboral y desempleo” por segunda vez puesto que se necesitaba extraer dos variables y no una; *Labour force participation* y *Unemployement rate*. En los otros casos (demás data set), finalmente usamos solo una variable.

Data Set "Fuerza laboral y desempleo_copia 1**"**

Country ID	Country	Year	Variable	Value
1	Total, all c...	2005	Labour force participation - Total	62.9
1	Total, all c...	2010	Labour force participation - Total	62.0
1	Total, all c...	2015	Labour force participation - Total	60.7
1	Total, all c...	2023	Labour force participation - Total	59.7

Data Set "Fuerza laboral y desempleo_copia 2**"**

Country ID	Country	Year	Variable	Value
1	Total, all co...	2005	Unemployment rate - Total	6.3
1	Total, all co...	2010	Unemployment rate - Total	6.3
1	Total, all co...	2015	Unemployment rate - Total	6.0
1	Total, all co...	2023	Unemployment rate - Total	5.8

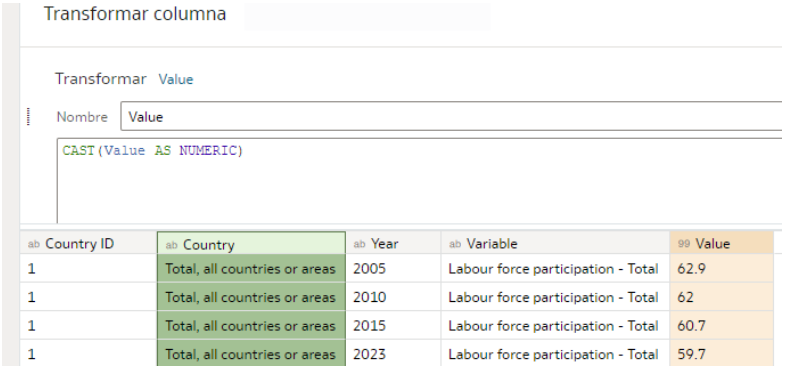
Data Set "Gasto público en educación**"**

Country ID	Country	Year	Variable	Value
4	Afghanistan	2010	Public expenditure on education (% of GDP)	3.5
4	Afghanistan	2015	Public expenditure on education (% of GDP)	3.3
4	Afghanistan	2019	Public expenditure on education (% of GDP)	3.2
4	Afghanistan	2020	Public expenditure on education (% of GDP)	2.9

Data Set "PIB y PIB per cápita**"**

Country ID	Country	Year	Variable	Value
1	Total, all co...	1995	GDP per capita (US dollars)	5,446
1	Total, all co...	2005	GDP per capita (US dollars)	7,287
1	Total, all co...	2010	GDP per capita (US dollars)	9,533
1	Total, all co...	2015	GDP per capita (US dollars)	10,140

Ahora necesitábamos convertir los VALUE en datos numéricos porque posteriormente operarían como cantidades. En este caso para la mayoría de data set se realizó la misma operación de reemplazo de carácter, exceptuando el data set de "PIB y PIB per cápita".



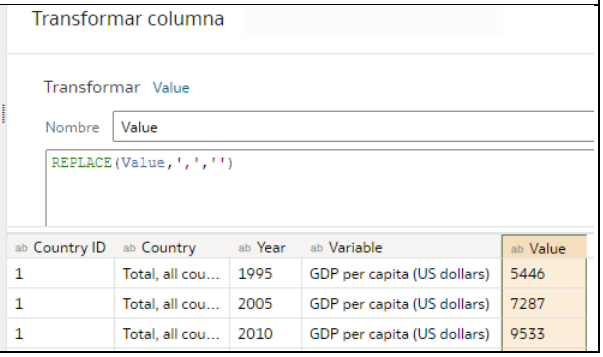
Country ID	Country	Year	Variable	Value
1	Total, all countries or areas	2005	Labour force participation - Total	62.9
1	Total, all countries or areas	2010	Labour force participation - Total	62
1	Total, all countries or areas	2015	Labour force participation - Total	60.7
1	Total, all countries or areas	2023	Labour force participation - Total	59.7

En el caso del data set "PIB y PIB per cápita" como estábamos hablando de cantidades en miles y no porcentuales, tuvimos que eliminar primero las comas y posterior convertir el campo en número, de lo contrario las cifras dejaban de reflejar miles y se convertirían en porcentuales, lo que ocasionan error en las cifras reales.

Value con separador COMAS

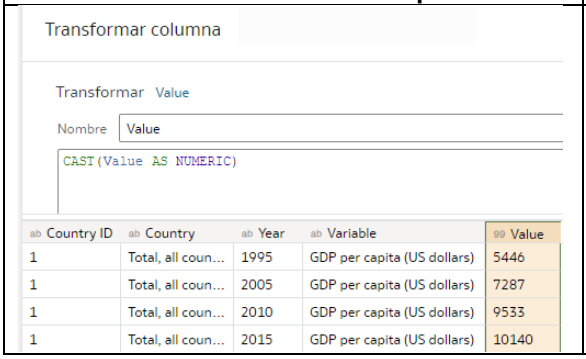
Country ID	Country	Year	Variable	Value
1	Total, all co...	1995	GDP per capita (US dollars)	5,446
1	Total, all co...	2005	GDP per capita (US dollars)	7,287
1	Total, all co...	2010	GDP per capita (US dollars)	9,533
1	Total, all co...	2015	GDP per capita (US dollars)	10,140
1	Total, all co...	2019	GDP per capita (US dollars)	11,301

Se elimina la COMA a los Value



Country ID	Country	Year	Variable	Value
1	Total, all cou...	1995	GDP per capita (US dollars)	5446
1	Total, all cou...	2005	GDP per capita (US dollars)	7287
1	Total, all cou...	2010	GDP per capita (US dollars)	9533

Se convierte a número el campo VALUE



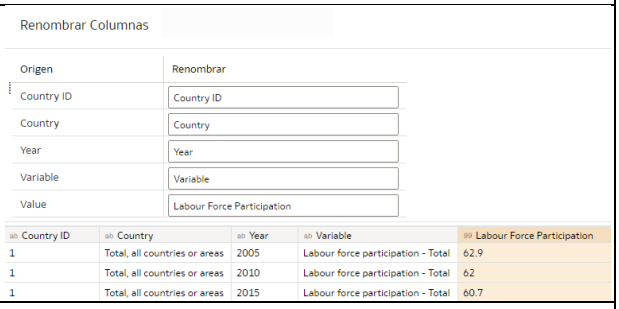
Country ID	Country	Year	Variable	Value
1	Total, all coun...	1995	GDP per capita (US dollars)	5446
1	Total, all coun...	2005	GDP per capita (US dollars)	7287
1	Total, all coun...	2010	GDP per capita (US dollars)	9533
1	Total, all coun...	2015	GDP per capita (US dollars)	10140

Posterior a la depuración de variables por cada data set, encontramos que se requería pasar cada variable como una columna individual con sus respectivos valores. Por tanto, tuvimos que transponer todas las filas. De esta forma después pudimos eliminar la columna con el nombre de “Variable” y entonces la columna con el valor (Value) se renombro como la variable inicial. Y al final pudimos unir todos los data set (*INNER JOIN*.) por los campos “Country - Year “(llaves primarias) para tener una sola base de datos enriquecida o base de datos para analítica (OLAP) con seis variables.

La Variable en filas

Country ID	Country	Year	Variable	Value
1	Total, all co...	2005	Labour force participation - Total	62.9
1	Total, all co...	2010	Labour force participation - Total	62
1	Total, all co...	2015	Labour force participation - Total	60.7
1	Total, all co...	2023	Labour force participation - Total	59.7
2	Africa	2005	Labour force participation - Total	64.4

Se renombre Value como la Variable

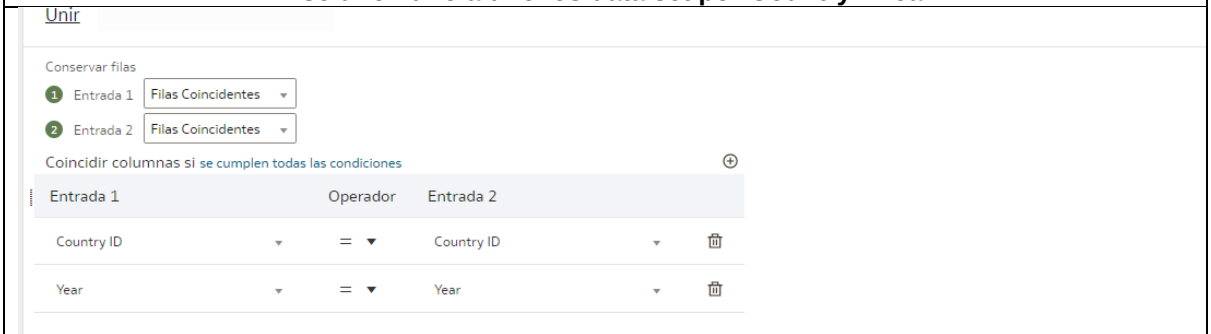


Country ID	Country	Year	Variable	Labour Force Participation
1	Total, all countries or areas	2005	Labour force participation - Total	62.9
1	Total, all countries or areas	2010	Labour force participation - Total	62
1	Total, all countries or areas	2015	Labour force participation - Total	60.7

Se excluye del modelo la columna “Variable”

Country ID	Country	Year	Labour Force Participation
1	Total, all co...	2005	62.9
1	Total, all co...	2010	62
1	Total, all co...	2015	60.7
1	Total, all co...	2023	59.7
2	Africa	2005	64.4
2	Africa	2010	63.8
2	Africa	2015	62.7

Y se unen uno a uno los data set por Country - Year



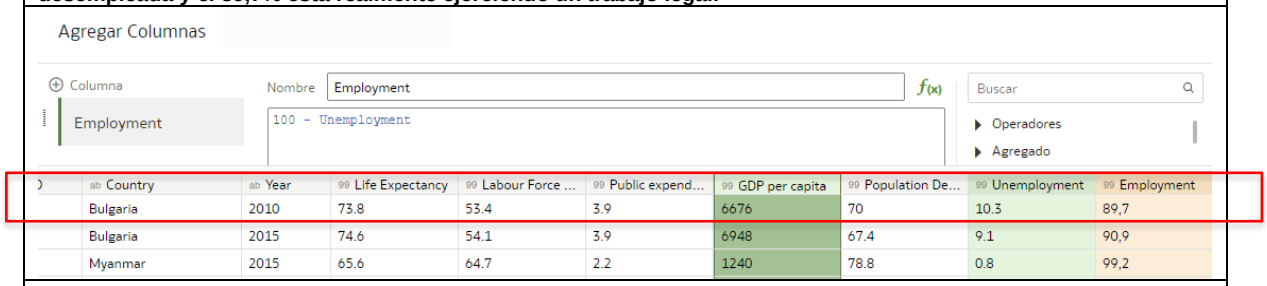
Country ID	Country	Year	Labour Force ...	Country ID_1	Country_1	Year_1	Public expend...	GDP per capita
100	Bulgaria	2010	53.4	100	Bulgaria	2010	3.9	6676
100	Bulgaria	2015	54.1	100	Bulgaria	2015	3.9	6948
104	Myanmar	2015	64.7	104	Myanmar	2015	2.2	1240
108	Burundi	2010	79.5	108	Burundi	2010	6.8	223

Este es el resultado final como data set total con las 6 variables

Country ID	Country	Year	Life Expectancy	Labour Force ...	Public expendit...	GDP per capita	Population Density	Unemployment
100	Bulgaria	2010	73.8	53.4	3.9	6676	70	10.3
100	Bulgaria	2015	74.6	54.1	3.9	6948	67.4	9.1
104	Myanmar	2015	65.6	64.7	2.2	1240	78.8	0.8
108	Burundi	2010	57.1	79.5	6.8	223	351.7	1.6

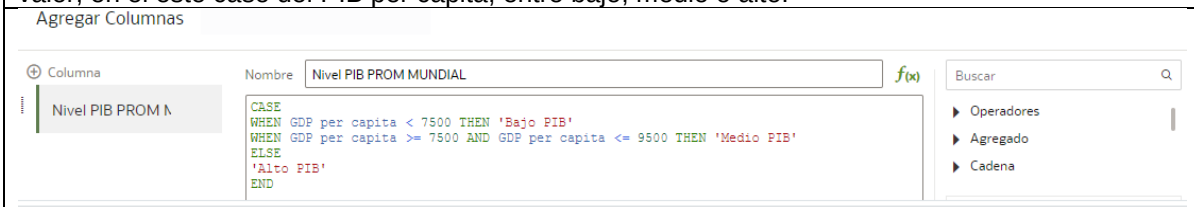
Los siguientes pasos fueron tratar de hacer operaciones y clasificaciones de los datos para enriquecerlos de tal forma que se pudieran agrupar y en un posterior análisis visual se presentara una visión global de lo que los datos quieren decir.

Se agrego una columna calculada que se encargue de conservar el valor de la empleabilidad real, es decir el valor de la fuerza laboral es mi 100% (53.4%) de personas aptas y en edad de trabajar, de esa población el 10,3% esta desempleada y el 89,7% está realmente ejerciendo un trabajo legal.



Country	Year	Life Expectancy	Labour Force ...	Public expend...	GDP per capita	Population De...	Unemployment	Employment
Bulgaria	2010	73.8	53.4	3.9	6676	70	10.3	89,7
Bulgaria	2015	74.6	54.1	3.9	6948	67.4	9.1	90,9
Myanmar	2015	65.6	64.7	2.2	1240	78.8	0.8	99,2

Luego se agregó otra columna que operara expresión de sentencia para categorizar o clasificar al valor, en el este caso del PIB per cápita; entre bajo, medio o alto.



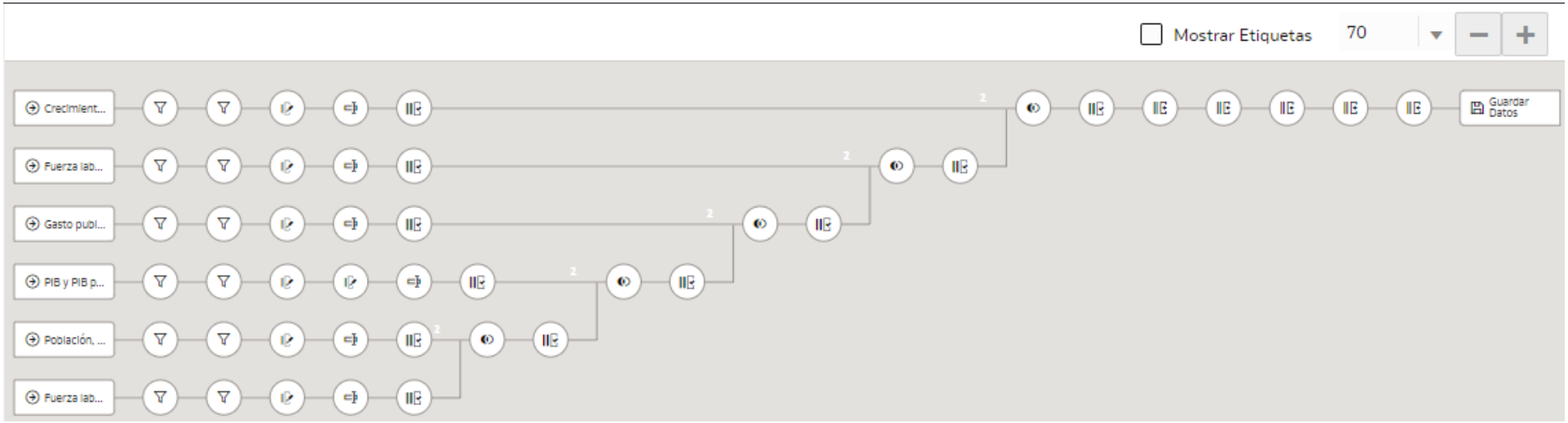
Year	Life Expectancy	Labour Force ...	Public expend...	GDP per capita	Population De...	Unemployment	Employment	Nivel PIB PRO...
2015	79.7	59.9	4.9	13650	23.7	6.5	93,5	Alto PIB
2010	75.6	71	3.8	4515	140.4	4.5	95,5	Bajo PIB
2015	77	69.3	3.8	7937	145.2	4.7	95,3	Medio PIB
2010	75	67.8	4.8	6394	39.9	11	89,0	Bajo PIB
2015	76.3	69.5	4.5	6228	41.9	8.3	91,7	Bajo PIB

Finalmente se siguió la misma lógica de enriquecimiento del data set, y se consiguió agregar cuatro columnas mas sobre categorización o clasificación; Nivel PIB, Nivel empleabilidad, Nivel inversión educación, Nivel esperanza de vida.

Public expend...	GDP per capita	Population De...	Unemployment	Employment	Nivel PIB PRO...	Nivel_Emplea...	Nivel_inversio...	Nivel_Life_Ex...
5.4	47618	3.7	8.1	91,9	Alto PIB	Medio	Alto	Alto
4.7	43550	3.9	6.9	93,1	Alto PIB	Medio	Alto	Alto
5.6	3193	129.3	10.7	89,3	Bajo PIB	Bajo	Alto	Medio
5.3	2891	137	11.8	88,2	Bajo PIB	Bajo	Alto	Medio
1.1	459	7.5	5.5	94,5	Bajo PIB	Medio	Bajo	Muy Bajo

NOTA: Se anexa un PDF llamada “DATA FLOW” para que se pueda observar el esquema simbólico de toda la transformación de los datos hasta su enriquecimiento.

DATA FLOW



GRUPO DE TRABAJO:

- José David Angulo
- Arnau Recio
- Albert Ripoll