ENTREGA INFORME PROYECTO FINAL STORYTELLING EN UN WEBSITE

Visualización y Storytelling – Team 13

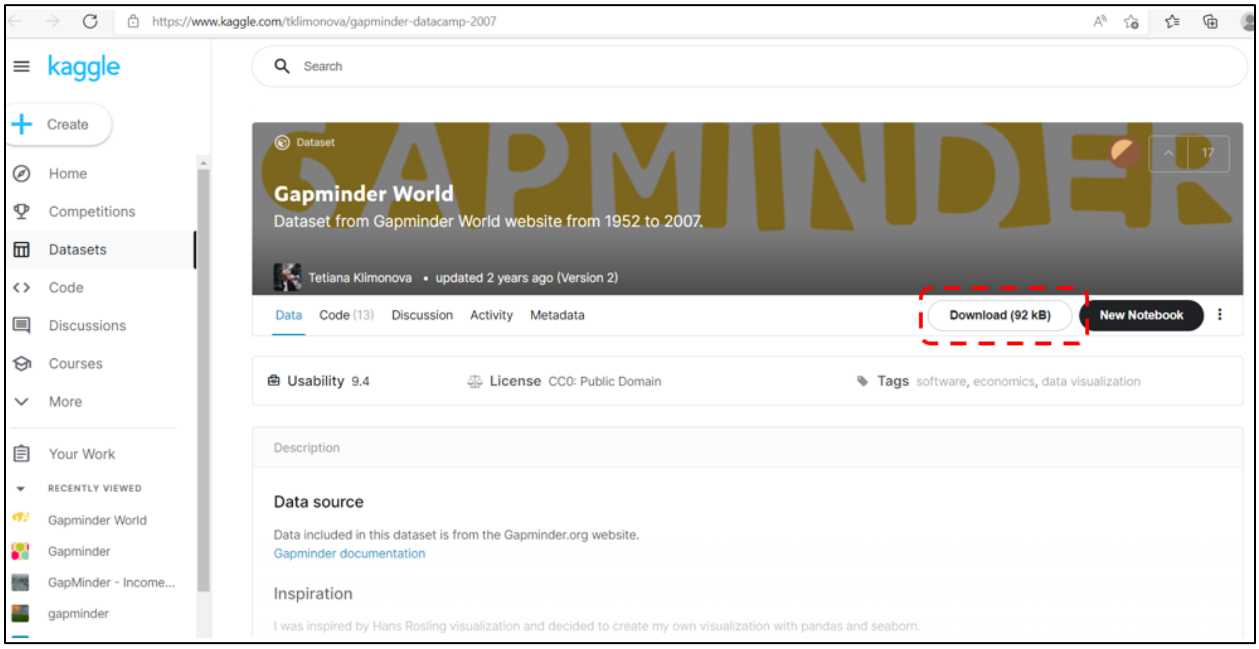
01- Contexto del problema:

La desigualdad social ha sido uno de los temas más relevantes a nivel mundial que mayor protagonismo ha tomado a lo largo del tiempo. La jerarquización de las diferencias ha impuesto múltiples condiciones a nivel mundial por las cuales las personas tienen acceso inequitativo a recursos valorados e importantes, por lo cual, todas las sociedades tienen un cierto nivel de desigualdad en un momento dado. En este orden de ideas, en este proyecto se identificarán dichas diferencias a través de distintos indicadores socioeconómicos que ilustran el nivel de desarrollo en las regiones y países de todo el mundo.

02- Modelamiento y perfilamiento de los datos

▪ Extracción de los datos

El dataset empleado para este proyecto corresponde a un archivo de datos de extensión "csv": ***'gapminder.csv'*** con 10545 observaciones, el cual tiene un total de 9 variables de interés; estos datos fueron tomados de la página de **"Kaggle"** disponible en el siguiente enlace: Gapminder World | Kaggle

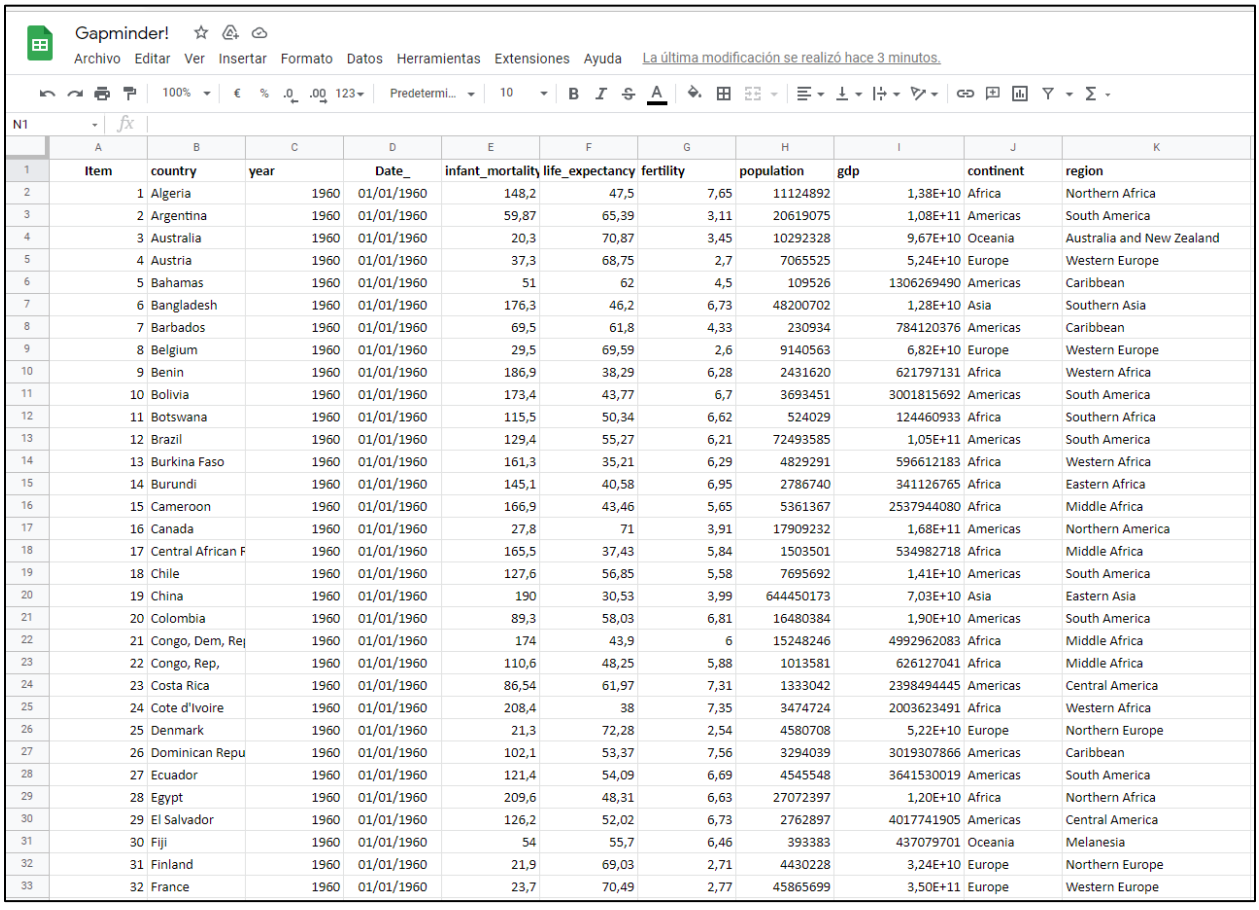


Los datos son de origen público y han sido cargados en esta web por "**Gapminder"** que es una sociedad sin ánimo de lucro fundada en la ciudad de Estocolmo, Suecia; los cuales tratan de dar una visión del mundo basada en los hechos. En este caso en particular; los datos relacionan variables tales como la tasa de mortalidad de infantes, expectativa de vida, tasa de fertilidad; entre otras más con el fin de explicar el problema de desigualdad social a nivel mundial el cual será tema de interés en el proyecto.

A continuación, se presenta una descripción de cada una de las variables del dataset:

VARIABLE	DESCRIPCIÓN
Country	*País donde se analiza la desigualdad social*
Year	*Año en el que se analiza la desigualdad social*
Infant Mortality	*Número de niños muertos por periodo de tiempo*
Life expectancy	*Expectativa de vida de la población*
Fertility	*Número de niños nacidos por periodo de tiempo*
Population	*Población donde se analiza la desigualdad social*
GDP	*Producto Interno Bruto de una region en particular*
Continent	*Continente donde se analiza la desigualdad social*
Region	*Región donde se analiza la desigualdad social*

Una vez descargado el dataset del repositorio oficial, este fue cargado en una hoja de Google Sheets tal como se muestra a continuación, con el fin de conocer y analizar los datos que se estaban trabajando para posteriormente ser importados a Google Data Studio, que es la herramienta utilizada para diseñar el Website del proyecto.



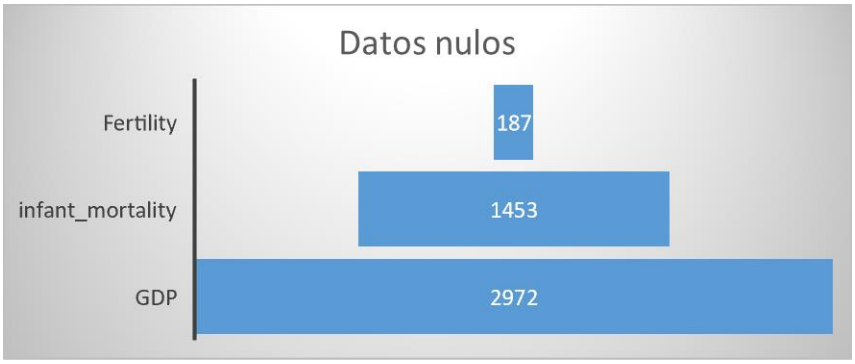
Gapminder!

Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones Ayuda La última modificación se realizó hace 3 minutos.

Item	country	year	Date_	infant_mortality	life_expectancy	fertility	population	gdp	continent	region
1	Algeria	1960	01/01/1960	148,2	47,5	7,65	11124892	1,38E+10	Africa	Northern Africa
2	Argentina	1960	01/01/1960	59,87	65,39	3,11	20619075	1,08E+11	Americas	South America
3	Australia	1960	01/01/1960	20,3	70,87	3,45	10292328	9,67E+10	Oceania	Australia and New Zealand
4	Austria	1960	01/01/1960	37,3	68,75	2,7	7065525	5,24E+10	Europe	Western Europe
5	Bahamas	1960	01/01/1960	51	62	4,5	109526	1306269490	Americas	Caribbean
6	Bangladesh	1960	01/01/1960	176,3	46,2	6,73	48200702	1,28E+10	Asia	Southern Asia
7	Barbados	1960	01/01/1960	69,5	61,8	4,33	230934	784120376	Americas	Caribbean
8	Belgium	1960	01/01/1960	29,5	69,59	2,6	9140563	6,82E+10	Europe	Western Europe
9	Benin	1960	01/01/1960	186,9	38,29	6,28	2431620	621797131	Africa	Western Africa
10	Bolivia	1960	01/01/1960	173,4	43,77	6,7	3693451	3001815692	Americas	South America
11	Botswana	1960	01/01/1960	115,5	50,34	6,62	524029	124460933	Africa	Southern Africa
12	Brazil	1960	01/01/1960	129,4	55,27	6,21	72493585	1,05E+11	Americas	South America
13	Burkina Faso	1960	01/01/1960	161,3	35,21	6,29	4829291	596612183	Africa	Western Africa
14	Burundi	1960	01/01/1960	145,1	40,58	6,95	2786740	341126765	Africa	Eastern Africa
15	Cameroon	1960	01/01/1960	166,9	43,46	5,65	5361367	2537944080	Africa	Middle Africa
16	Canada	1960	01/01/1960	27,8	71	3,91	17909232	1,68E+11	Americas	Northern America
17	Central African F	1960	01/01/1960	165,5	37,43	5,84	1503501	534982718	Africa	Middle Africa
18	Chile	1960	01/01/1960	127,6	56,85	5,58	7695692	1,41E+10	Americas	South America
19	China	1960	01/01/1960	190	30,53	3,99	644450173	7,03E+10	Asia	Eastern Asia
20	Colombia	1960	01/01/1960	89,3	58,03	6,81	16480384	1,90E+10	Americas	South America
21	Congo, Dem, Rep	1960	01/01/1960	174	43,9	6	15248246	4992962083	Africa	Middle Africa
22	Congo, Rep,	1960	01/01/1960	110,6	48,25	5,88	1013581	626127041	Africa	Middle Africa
23	Costa Rica	1960	01/01/1960	86,54	61,97	7,31	1333042	2398494445	Americas	Central America
24	Cote d'Ivoire	1960	01/01/1960	208,4	38	7,35	3474724	2003623491	Africa	Western Africa
25	Denmark	1960	01/01/1960	21,3	72,28	2,54	4580708	5,22E+10	Europe	Northern Europe
26	Dominican Repu	1960	01/01/1960	102,1	53,37	7,56	3294039	3019307866	Americas	Caribbean
27	Ecuador	1960	01/01/1960	121,4	54,09	6,69	4545548	3641530019	Americas	South America
28	Egypt	1960	01/01/1960	209,6	48,31	6,63	27072397	1,20E+10	Africa	Northern Africa
29	El Salvador	1960	01/01/1960	126,2	52,02	6,73	2762897	4017741905	Americas	Central America
30	Fiji	1960	01/01/1960	54	55,7	6,46	393383	437079701	Oceania	Melanesia
31	Finland	1960	01/01/1960	21,9	69,03	2,71	4430228	3,24E+10	Europe	Northern Europe
32	France	1960	01/01/1960	23,7	70,49	2,77	45865699	3,50E+11	Europe	Western Europe

- **Limpieza de Datos**

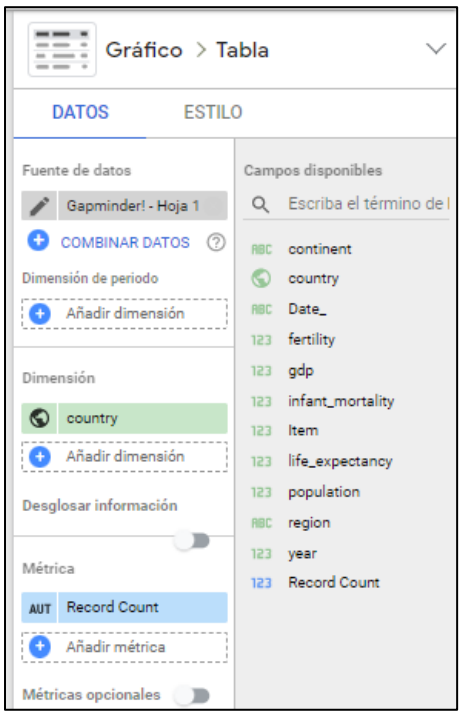
Uno de los objetivos principales fue limpiar los datos nulos que aparecían presentes en la data, donde se encontraron los siguientes hallazgos:



En la imagen anterior se puede apreciar que 3 variables presentaron datos nulos, pero en particular la variable que más sucia estaba era la del “Producto Interno Bruto (GDP)”, en este orden de ideas y considerando que esta variable era de las más representativas para la historia que se deseaba contar, las observaciones nulas de esta variable fueron eliminadas, y por consiguiente; también se eliminaron de manera directa las observaciones nulas de la variable “Fertility”. La decisión anterior se argumentó bajo los hechos de que la variable en mención representaba un índice socioeconómico muy influyente en la desigualdad social a nivel mundial, por ende; su imputación o reemplazo de valores nulos por algún parámetro como la media de los datos, podría generar sesgos en la historia a contar.

Por último, fueron eliminadas también las 434 observaciones de la variable “infant_mortality” ya que se consideraba un porcentaje muy pequeño comparado con el tamaño del dataset inicial, en este orden de ideas; el dataset final quedo con un total de 7139 observaciones para cada variable.

Ahora bien, al importar el data set a "Google Data Studio" desde “Google Sheets” y después de haber identificado en la extracción de los datos que el nombre de las variables estaba en inglés tal como se muestra a continuación, se decidió cambiar el nombre de las variables:

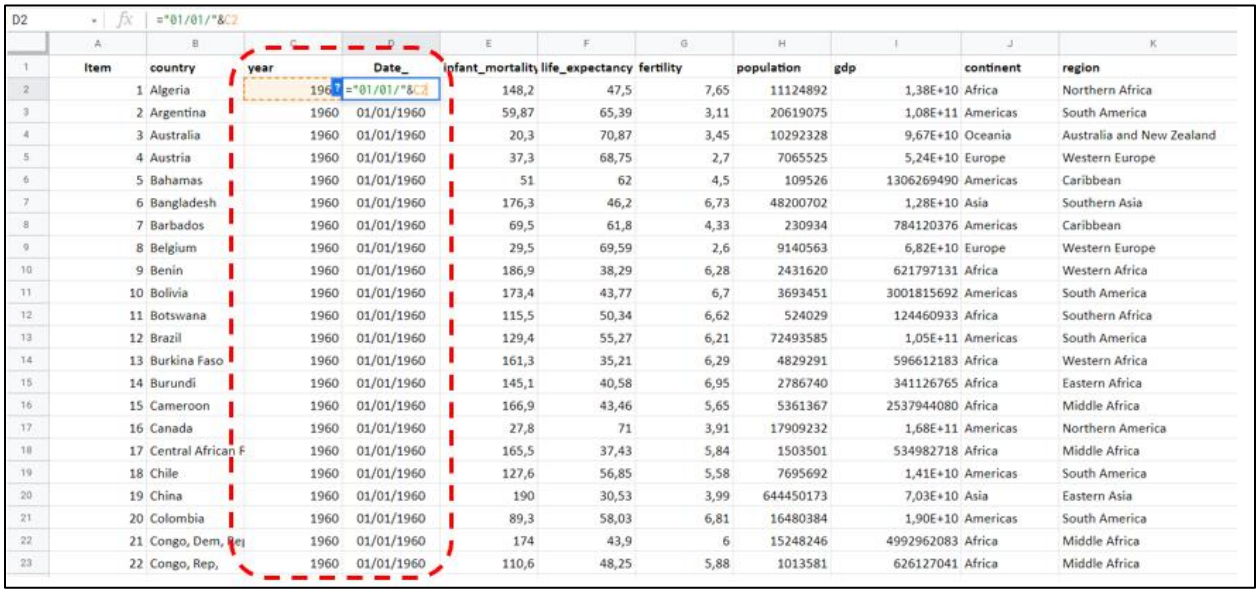


De la imagen anterior se puede apreciar que las variables son tipo numéricas y de texto, para este caso; se realizó el cambio de las variables "continent" y "country" a variables geográficas, ya que son estos países y continentes los que se presentan en el mapa. Por otro lado, las variables tienen información relevante de las diferentes regiones tales como GDP, mortalidad, fertilidad y expectativa de vida, las cuales serán variables a utilizar para minar el data set más adelante de este informe.

Se procedió entonces a realizar el cambio de nombre de las variables a español, esto con el fin de trabajar cómodamente con la data, a continuación; se presenta cada una de las variables con su nuevo nombre:

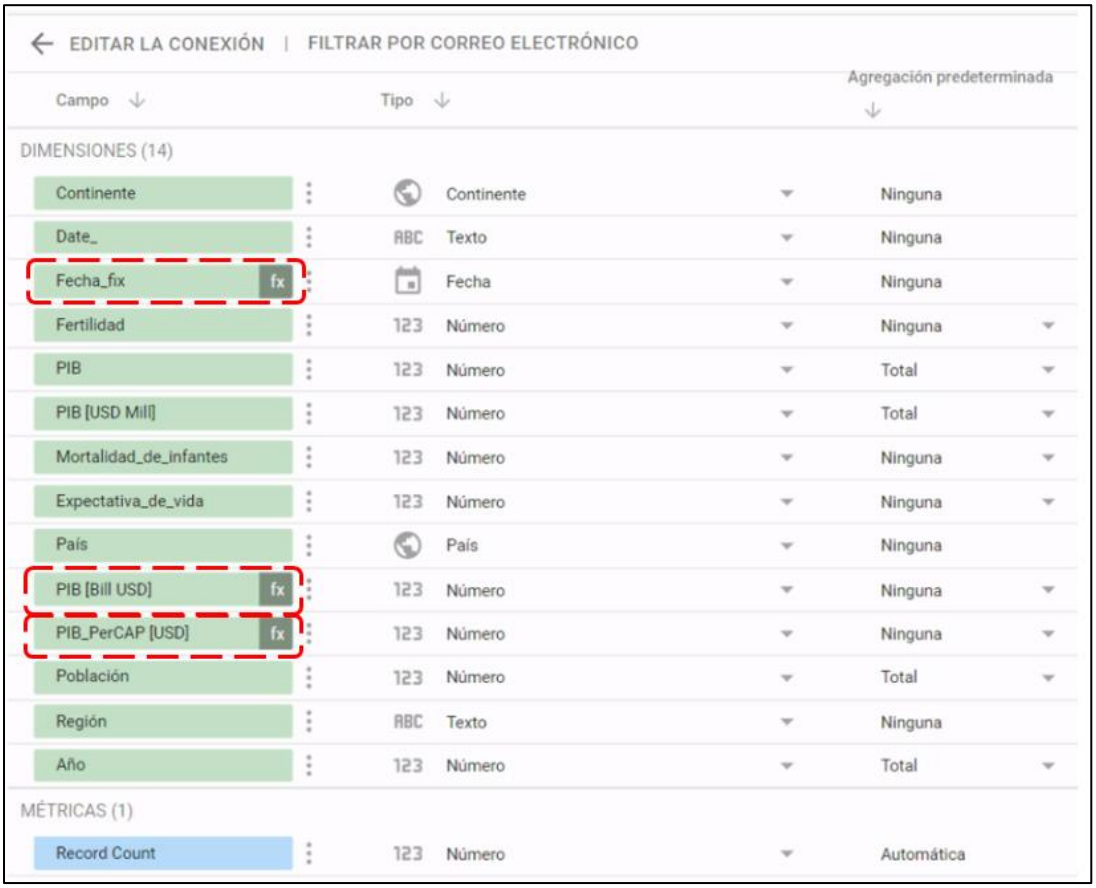
Ítem	Variable en inglés	Variable en español
1	Country	País
2	Year	Año
3	Infant Mortality	Mortalidad de infantes
4	Life expectancy	Expectativa de vida
5	Fertility	Fertilidad
6	Population	Población
7	GDP	PIB
8	Continent	Continente
9	Region	Región

Cabe mencionar que antes de que el dataset fuera cargado a Google Data Studio, se creó una nueva variable **"date"** en la hoja de Google Sheets con el fin de obtener un formato de fecha completo compuesto por día, mes y año; lo anterior debido a que el dataset solo contenía el número del año, para esto se concateno este año con una cadena de texto representativa por el día y mes de la fecha, tal como se muestra a continuación:

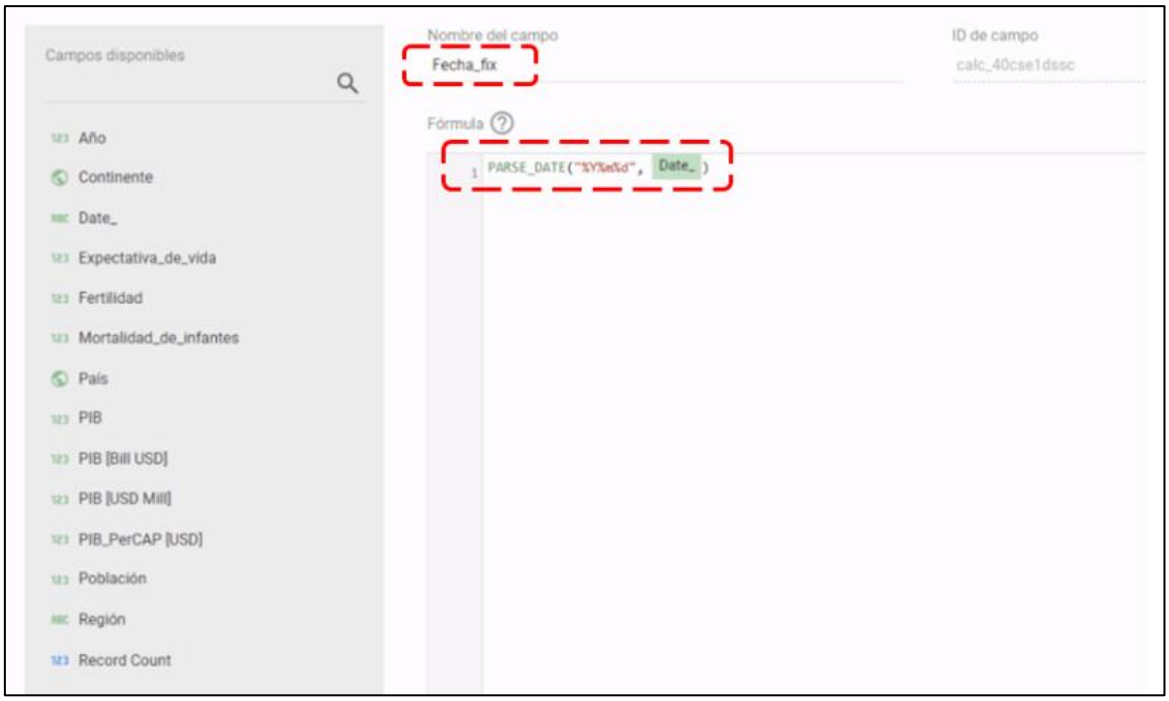


	A	B	C	D	E	F	G	H	I	J	K
1	Item	country	year	Date_	infant_mortality	life_expectancy	fertility	population	gdp	continent	region
2	1	Algeria	196	="01/01/"&C2	148,2	47,5	7,65	11124892	1,38E+10	Africa	Northern Africa
3	2	Argentina	1960	01/01/1960	59,87	65,39	3,11	20619075	1,08E+11	Americas	South America
4	3	Australia	1960	01/01/1960	20,3	70,87	3,45	10292328	9,67E+10	Oceania	Australia and New Zealand
5	4	Austria	1960	01/01/1960	37,3	68,75	2,7	7065525	5,24E+10	Europe	Western Europe
6	5	Bahamas	1960	01/01/1960	51	62	4,5	109526	1306269490	Americas	Caribbean
7	6	Bangladesh	1960	01/01/1960	176,3	46,2	6,73	48200702	1,28E+10	Asia	Southern Asia
8	7	Barbados	1960	01/01/1960	69,5	61,8	4,33	230934	784120376	Americas	Caribbean
9	8	Belgium	1960	01/01/1960	29,5	69,59	2,6	9140563	6,82E+10	Europe	Western Europe
10	9	Benin	1960	01/01/1960	186,9	38,29	6,28	2431620	621797131	Africa	Western Africa
11	10	Bolivia	1960	01/01/1960	173,4	43,77	6,7	3693451	3001815692	Americas	South America
12	11	Botswana	1960	01/01/1960	115,5	50,34	6,62	524029	124460933	Africa	Southern Africa
13	12	Brazil	1960	01/01/1960	129,4	55,27	6,21	72493585	1,05E+11	Americas	South America
14	13	Burkina Faso	1960	01/01/1960	161,3	35,21	6,29	4829291	596612183	Africa	Western Africa
15	14	Burundi	1960	01/01/1960	145,1	40,58	6,95	2786740	341126765	Africa	Eastern Africa
16	15	Cameroon	1960	01/01/1960	166,9	43,46	5,65	5361367	2537944080	Africa	Middle Africa
17	16	Canada	1960	01/01/1960	27,8	71	3,91	17909232	1,68E+11	Americas	Northern America
18	17	Central African F	1960	01/01/1960	165,5	37,43	5,84	1503501	534982718	Africa	Middle Africa
19	18	Chile	1960	01/01/1960	127,6	56,85	5,58	7695692	1,41E+10	Americas	South America
20	19	China	1960	01/01/1960	190	30,53	3,99	644450173	7,03E+10	Asia	Eastern Asia
21	20	Colombia	1960	01/01/1960	89,3	58,03	6,81	16480384	1,90E+10	Americas	South America
22	21	Congo, Dem, Rep	1960	01/01/1960	174	43,9	6	15248246	4992962083	Africa	Middle Africa
23	22	Congo, Rep,	1960	01/01/1960	110,6	48,25	5,88	1013581	626127041	Africa	Middle Africa

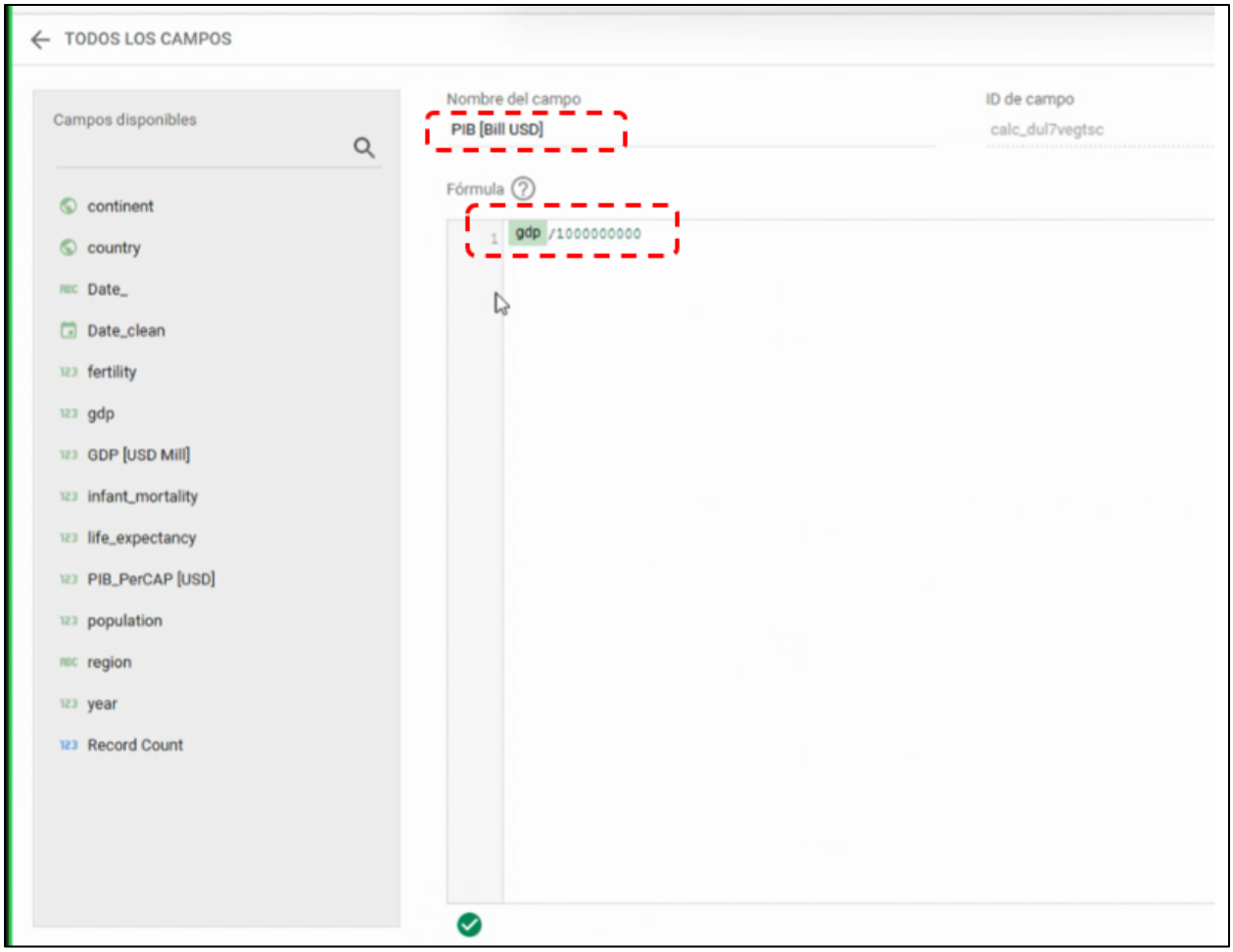
Una vez realizado el paso anterior, se garantizó el formato completo de la variable "date" a Google Data Studio. Ahora bien, después de haber realizado el cambio de nombre de las variables y la creación de la variable "date", se procedió a realizar en Google Data Studio la creación de nuevas variables, con el fin de minar el dataset y de esta manera lograr análisis más limpios y más enfocados en lo que se desea narrar al público objetivo. Las nuevas variables se presentan a continuación:



Debido a que los datos son estáticos, se utilizaron scripts que ayudaron con la creación de estas nuevas variables, la primera de ellas; **"Fecha_fix"** surge debido a la necesidad de obtener la variable "date" en formato tipo fecha, ya que cuando esta fue creada en la hoja de Google Sheets esta variable quedo en formato de texto, en este orden de ideas se realizó un casteo de la variable para realizar el cambio de formato y de esta manera crear una nueva variable tal como se muestra a continuación:

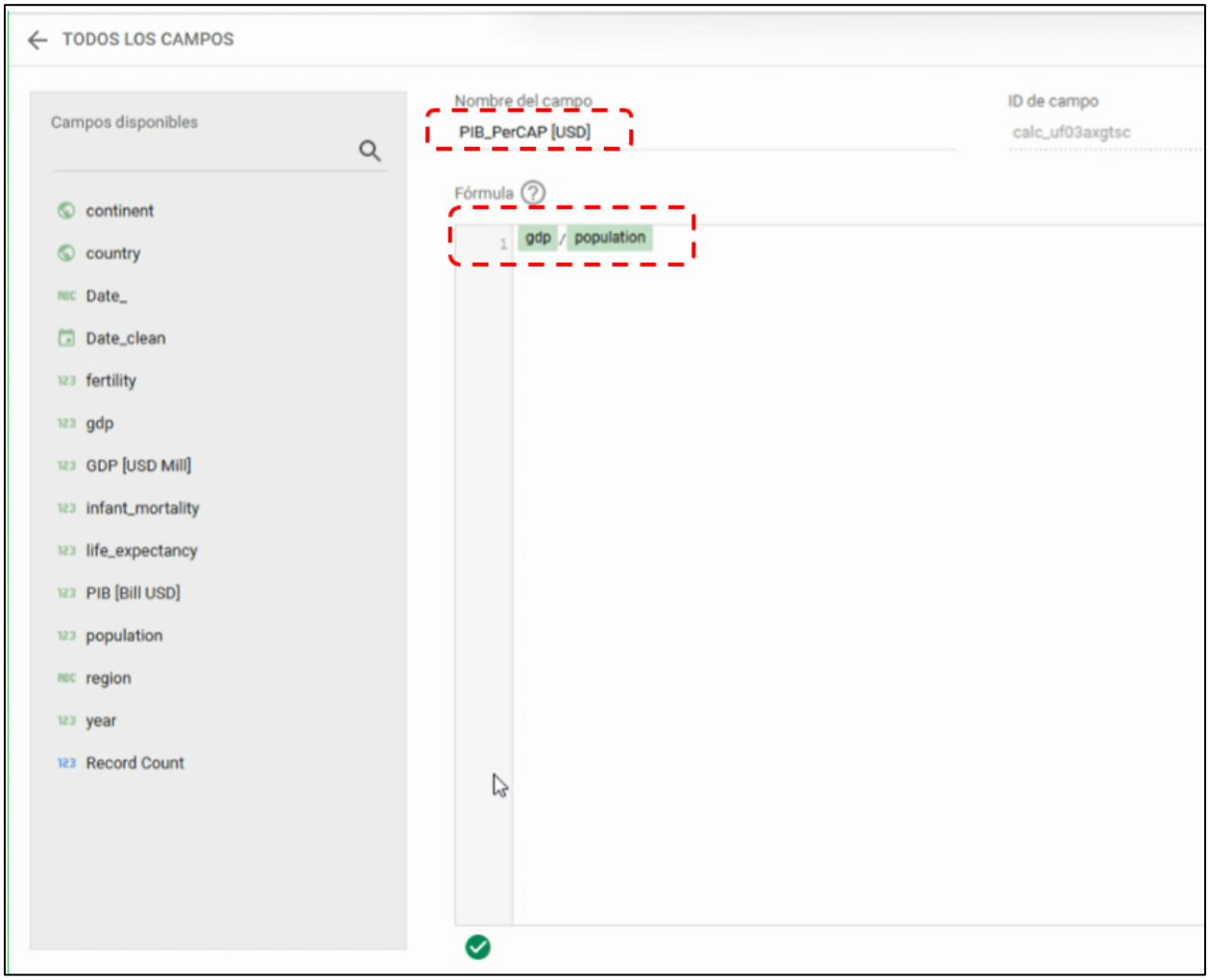


Con el script anterior se garantizó el formato adecuado para la linea de tiempo de la historia que se desea contar al publico en general. Ahora bien, la segunda variable que se genero fue la de **"PIB (Bill USD)",** debido a que los valores del campo de "PIB" del dataset eran demasiado grandes y esto podria conllevar a una saturación en las visualizaciones a implementar en la página web, por lo anterior; se decidió por manejar esta variable en millones de dolares, tal como se muestra a continuación:



En el script anterior, al dividir la variable PIB por mil millones, se logra la creación de la nueva variable en mención. Por último, la tercera variable que se creó fue la de Producto Interno Bruto per capital **"PIB_PerCAP (USD)",** lo anterior teniendo en cuenta que esta nueva variable juega un papel demasiado importante a la hora de analizar y explicar la desigualdad social a nivel mundial, debido a que esta explica y relaciona la riqueza promedio por cada habitante y se obtiene al realizar el cociente entre el PIB y el número de población. También es un excelente indicador de la varianza que puede presentar cada continente con respecto a su desigualdad social, ya que por ejemplo el continente americano presenta un mayor PIB a nivel mundial, pero a la hora de analizar que países son los que más contribuyen con este valor de PIB, se puede apreciar claramente que son solo algunas naciones las mayores contribuyentes a este elevado índice socioeconómico, y esto es un indicador de varianza y por ende de desigualdad social.

El script para obtener esta nueva variable se presenta a continuación:



- **Visualización de los datos**

La visualización de los datos se presenta en la página web creada en Google Data Studio disponible para su lectura en el siguiente link:

https://datastudio.google.com/reporting/2a740d2d-b430-4cbf-b12a-53ac727b5f84