**M2.859 - Visualización de datos aula 2**

Aula

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A9: Creación de la visualización y entrega del proyecto (Práctica)** | | | |  |
|  | | | |
|  | | | |  |
| Inicio: | Entrega: | Calificación: | Dedicación: |  |
| **10/05/21**  **00:00h** | **09/06/21**  **24:00h** | **23/06/21** | **80 %** |  |

Descripción del enunciado

El estudiante debe desarrollar una visualización de datos que demuestre sus conocimientos del ámbito así como el uso de diferentes herramientas y técnicas. Además tendrá que librar un informe que describa el proceso de creación seguido y las decisiones de diseño que se hayan tomado a lo largo del desarrollo, así como una pequeña presentación de la visualización, el conjunto de datos que se visualiza, las preguntas que se responden y los elementos interactivos disponibles como parte de la misma visualización:

La visualización debe incluir los elementos siguientes:

1) Título de la visualización donde se presentan la visualización realizada. URL de la visualización y del código. Y descripción corta del documento y del que se presenta.

1.1) El código es tiene que publicar en un repositorio público de GitHub con una licencia de código abierto. Tiene que contenga los archivos necesarios para correr la aplicación en local con la ayuda de un servidor web si fuera necesario.

1.2) La URL de la visualización tiene que ser pública y poderse explorar sin login. Por ejemplo GitHub pages (https://pages.github.com/), surge.sh o subir la aplicación a un espacio web propio accesible o compartir un url accesible sin usuario y contraseña.

2) Explicar razonadamente qué preguntas responde la visualización presentada y qué uso puede tener por un usuario tipo.

3) Descripción técnica del proyecto: lenguajes, librerías, licencias, descripción técnica del proyecto.

4) La/s visualizaciones realizadas.

# Tópicos de salud y estilo de vida de mujeres alrededor del mundo.

A través del tiempo se han producido cambios en las condiciones de vida de mujeres en el mundo, siendo hasta hoy la desigualdad de género una problemática que afecta en los distintos ámbitos a nivel social y cultural, lo cual lleva preguntarnos ¿qué condujo a esta desigualdad? o ¿Qué aspectos deben tomarse en cuenta para acabar con estos problemas?

La visualización presentada utiliza una recopilación de 4 tipos sobre aspectos relacionados a la participación de la mujer en la salud así como también en el hogar, trabajos y quehaceres desde al 1960 a 2020.

La Url o enlace de interés que se ha publicado mediante Tableau es la siguiente:

<https://public.tableau.com/app/profile/christian.alejandro.zamora.flores/viz/Gender_16232680901670/Dashboard>

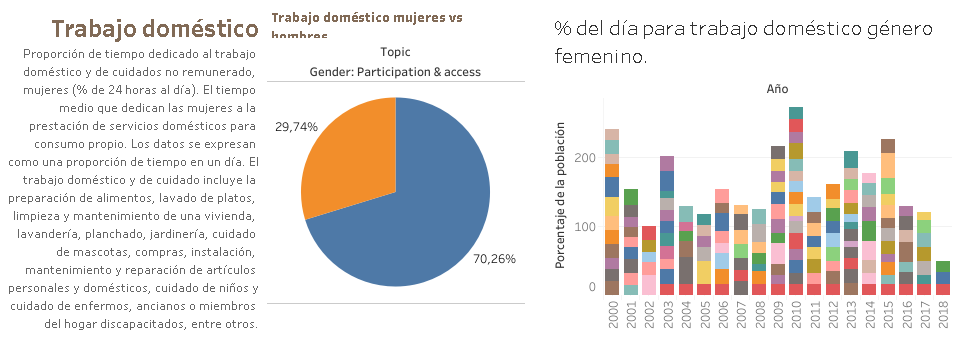
<https://public.tableau.com/shared/PWFQY3S9C?:display_count=n&:origin=viz_share_link>



A continuación se define los aspectos del trabajo:

**Trabajo doméstico**

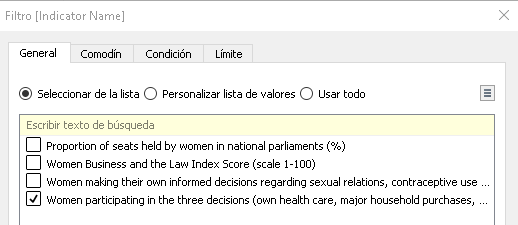
Para realizar esta visualización utilizamos un conjunto de datos en donde se filtra información referente al Tópico: “Participation (DataParticipation)”, cuyos indicadores definen la participación de las mujeres en el hogar. Se presenta un pastel para evidenciar la proporción diferencia de hombres y mujeres, mientras que por otra parte se aprecia un gráfico de barras para comparar la participación femenina en todos los países.



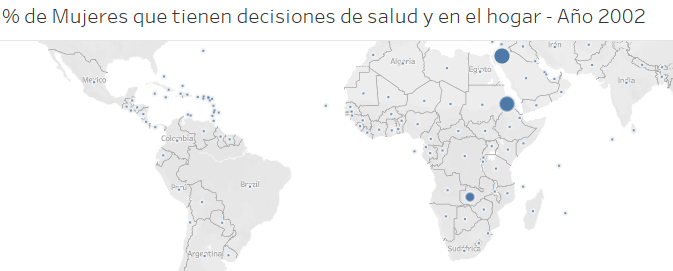
**Decisión en la salud y el hogar.**

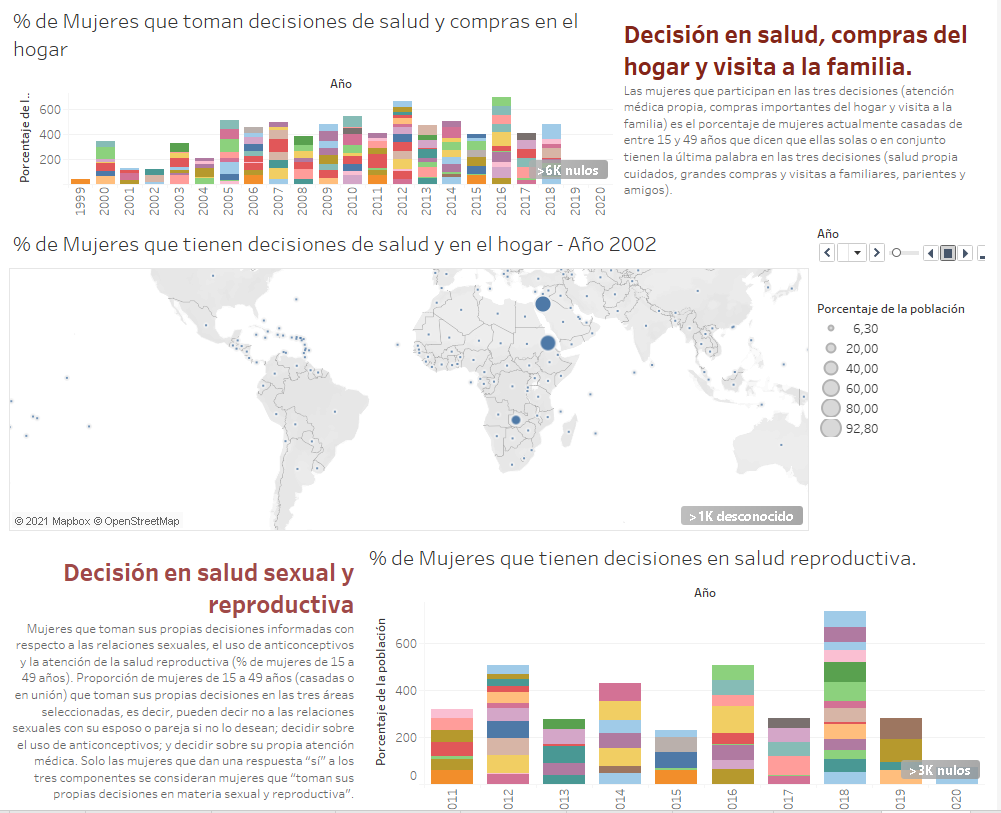
Utilizamos un conjunto de datos filtrando el tema: “LifePublic (DataLifePublic)”, este a su vez tiene varios indicadores y porcentajes relacionados según el tema. A continuación se citan algunos:

* Decisión en salud, compras del hogar y visita a la familia.
* Decisión en salud sexual y reproductiva.
* Participación política de la mujer.
* Emprendimiento.

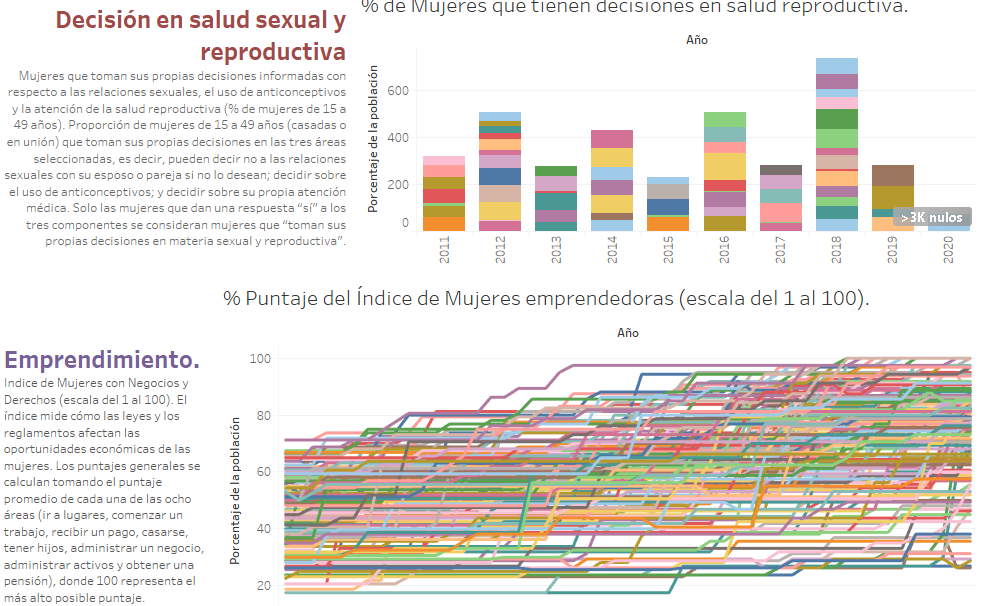


Además a esta visualización se agrega un mapa, el cual describe los porcentajes del indicador a través del tiempo, expresado en los años en que se obtiene la información. En este caso el indicador visualizado es sobre la participación de la mujer en las decisiones del hogar.

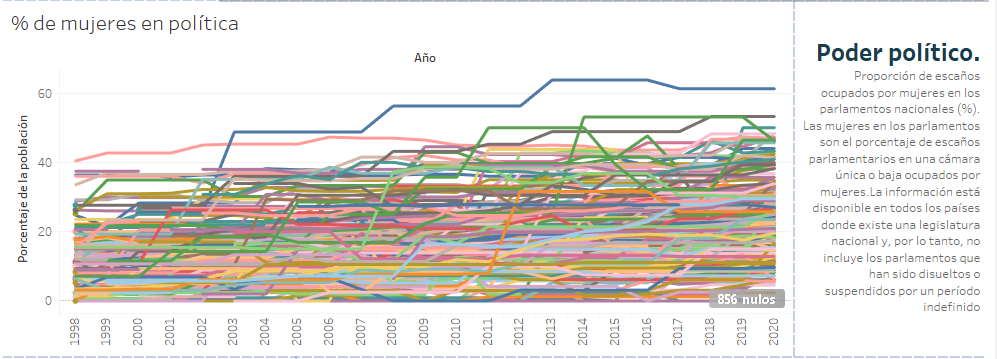




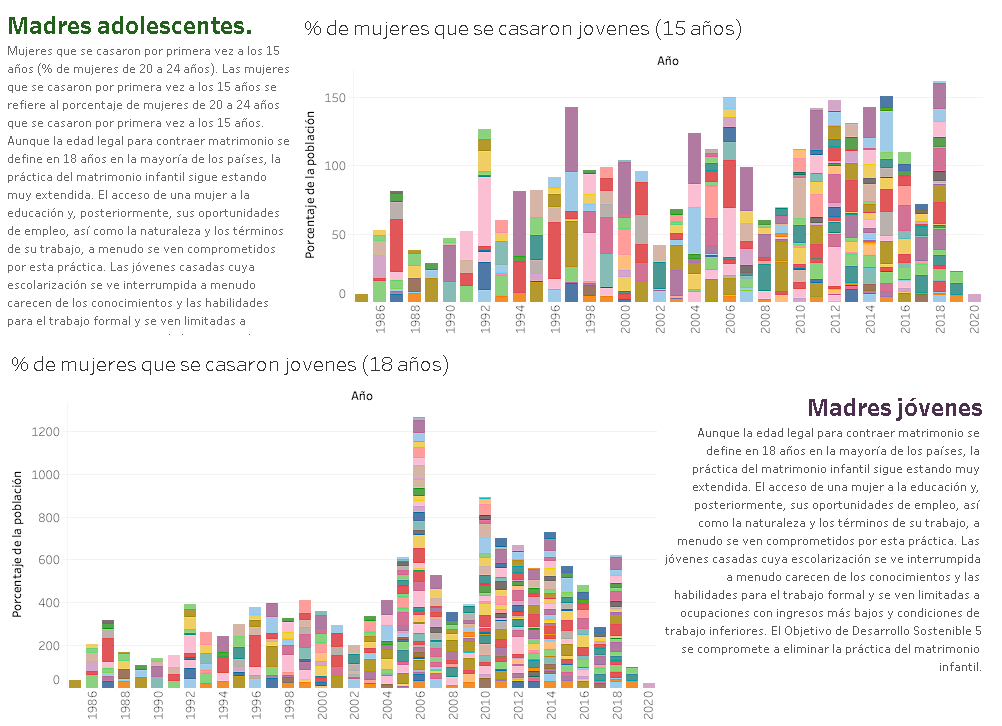
Luego para cada indicador se realiza la respectiva visualización gráfica junto con la descripción. Para diferenciar el país en donde se presenta la muestra, se aplica una paleta de colores.



Debido a que en ciertas ocasiones la gráfica de barras se torna más segmentada, la técnica de visualización exige simplificar para mostrar al usuario una secuencia escalonada a través del tiempo. Para esto se usan líneas escaladas, en donde la altura define el porcentaje en que ocurre una muestra.

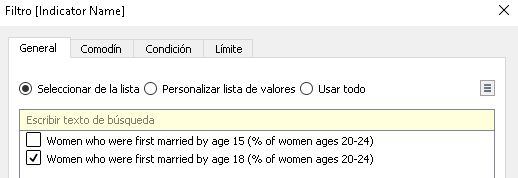


**Mujeres que son madres.**



Utilizamos el conjunto de datos: “Agencia”, este a su vez contiene 2 indicadores y porcentajes relacionados. A continuación se citan estos:

* Madres adolescentes.
* Madres jóvenes.



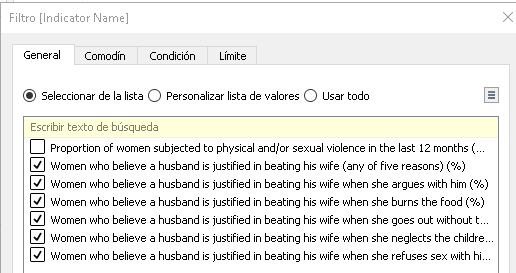
En este caso realizamos dos filtros para separar la temática de la gráfica. Es decir una visualización para datos filtrados de madres en edad de 15 años (adolescentes) y edad de 18 para madres jóvenes.

En este caso utilizamos gráficas de barras, estas a su vez están separadas en colores que identifican la muestra obtenida.

**Violencia sexual y física.**

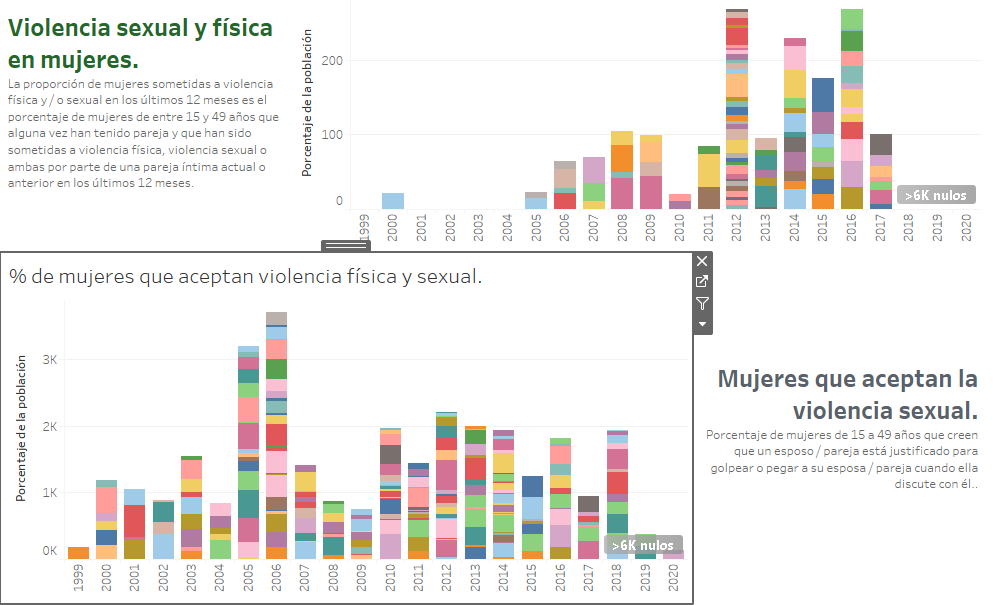
Para este último ejemplo utilizamos el conjunto de datos: “Health”, que a su vez contiene varios indicadores y porcentajes relacionados. Estos a su vez se dividen en dos grupos de indicadores descritos a continuación.

* Mujeres agredidas física y sexualmente en los últimos 12 meses.
* Mujeres que justifican la violencia de sus esposos.



Para este caso utilizamos gráficos de barras segmentados por países tal como se ha venido haciendo durante el transcurso de la práctica.

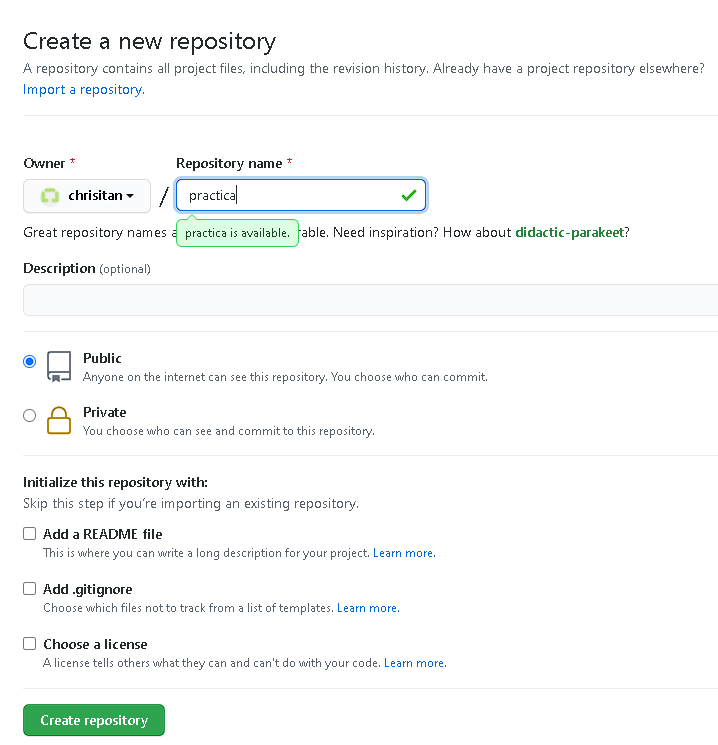
Finalmente esta información es consolidada en un Dashboard y subida al cloud de Tableau en la nube pública.



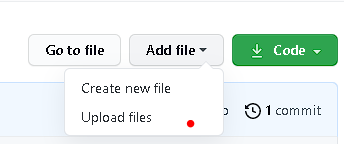
**Upload en GITHUB.**

Ahora realizamos la carga de datos del proyecto en GITHUB, para esto nos registramos con una cuenta en GITHUB utilizando nuestro correo electrónico.

Creamos un nuevo proyecto de repositorio. Se llamará práctica 2. No es necesario configurar otros aspectos, simplemente damos click en el botón “Create Repository”

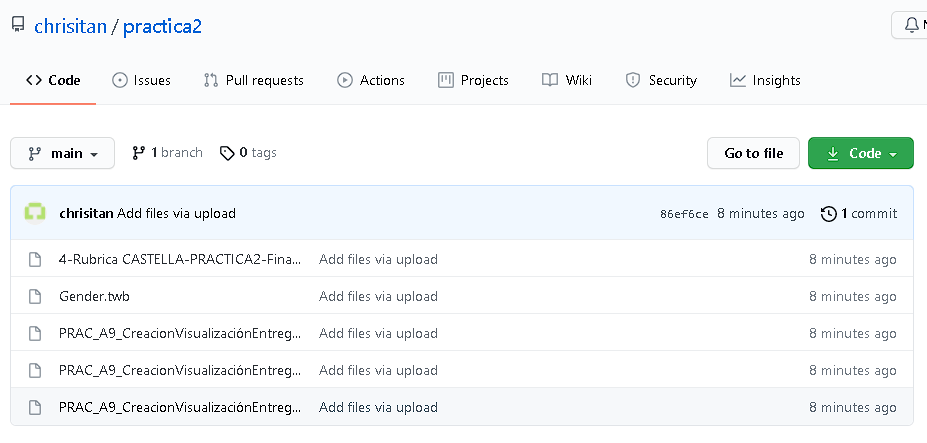


Una vez creado el repositorio utilizamos la función de añadir archivos para subir toda la información del computador al repositorio.



A continuación se cargarán todos los datos en la nube. Luego de estos podremos observar la carpeta compartida del repositorio. La url o dirección es la siguiente:

[GitHub - chrisitan/practica2: Selección y extracción de datos para utilizarlos en procesos de visualización](https://github.com/chrisitan/practica2)



A continuación se describe la codificación que contiene las funciones y librerías que se utilizan para la obtención y selección del conjunto de datos antes de proceder a visualizarlo.

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

# Función para plotear gráficas.

def group\_by\_plot(dataset, by, ascending=None, head=None, kind="bar", figsize=(20,5), title=None, \*args, \*\*kwargs):

    tmp = dataset.groupby(by=by).size()

    if ascending is not None:

        tmp = tmp.sort\_values(ascending=ascending)

    return tmp.head(head).plot(kind = kind, figsize = figsize, title = title , \*args, \*\*kwargs)

# Formatea codificación del documento cuando se extrae

encoding='utf-8'

encoding = "cp1252"

encoding = "ISO-8859-1"

# Extraemos la data, pero al estar en formato excel, debemos instalar la funcionalidad adicional en el paquete xlrd, para esto ejecutamos el comando conda install -c anaconda xlrd, o pip install xlrd, de esta forma se invoca a la función que extraerá la información de las hojas de cálculo.

DataDE = pd.read\_excel("Data\_Exploration/WDIEXCEL.xlsx",sheet\_name="Data")

CountryDE = pd.read\_excel("Data\_Exploration/WDIEXCEL.xlsx",sheet\_name="Country")

SeriesDE = pd.read\_excel("Data\_Exploration/WDIEXCEL.xlsx",sheet\_name="Series")

# Verificamos la complejidad de la data estructura de la data.

DataDE.head(5)

# Información de la estructura del dataset (dataDE)

print("Número de columnas:", len((DataDE.columns)))

print("Número de filas:", np.shape(DataDE)[0])

# Para ver cuántos países están representados, ejecute la siguiente comando:

paises = DataDE.loc[:,'Country Name'].unique()

len(paises)

# Para ver cuántos indicadores están representados, ejecute la siguiente comando:

indicadores = DataDE.loc[:,'Indicator Code'].unique()

len(indicadores)

# Para mostrar los diferentes tipos de cálculo que uno podría estar interesado en realizar, consideremos un solo país, por ejemplo, Brasil.

# Supongamos que estamos interesados en la información sobre la perspectiva de género que es el tópico al que hace referencia la práctica.

# Primero, seleccionamos las filas cuya columna Pais coincida con Brasil, usando la siguiente línea de comando:

DataDE\_br = DataDE.loc[DataDE.loc[:,'Country Name']=='Brazil',:]

# En el código que precede a la línea de comando, considere la siguiente expresión:

# Ahora seleccionamos todas las filas y columnas de la tabla del resultado cuando se refiere a los datos de la perspectiva de género.

# Definimos una función que, dada una cadena, determina si contiene la subcadena SG:

select\_fcn = lambda string: string.upper().find('SG.') >= 0

# Seleccionamos las filas en DataDE\_br que devuelven True cuando select\_fcn se aplica a la columna Código de indicador del SG.

criterio = DataDE\_br.loc[:,'Indicator Code'].map(select\_fcn)

DataDE\_br\_SG = DataDE\_br.loc[criterio,:]

len(DataDE\_br\_SG)

# Esto significa que hay 19 indicadores relacionados con la perspectiva de género para el país llamado Brasil.

# Ahora que existe una cantidad manejable de datos, podemos mostrar una tabla que ti ne

# los códigos del indicador y sus significados usando la siguiente línea de comando:

DataDE\_br\_SG.loc[:,['Indicator Code', 'Indicator Name']]

# Podemos suponer que nos interesa solo cuatro indicadores, es decir podemos continuar recortando los datos con el siguiente comando:

DataDE\_br\_SG = DataDE\_br\_SG.loc[[106772, 106773, 106774, 106775], :]

# Esto produce una tabla manejable de 4 filas y 58 columnas. Obteniendo en cada fila

# una serie de tiempo de los datos correspondientes sobre la perspectiva de género a partir del año 1960.

#

# La tabla esta presentada de forma transpuesta es decir las series de tiempo están en las filas de la tabla, en su lugar de estar abajo de las columnas. Entonces, todavía tenemos que trabajar un poco más con nuestra tabla.

# Queremos que los índices de nuestra tabla sean los años. También queremos tener una columna para cada indicador económico y desea utilizar los nombres de los indicadores económicos (no los códigos) como las etiquetas de las columnas. Así es como se puede hacer esto:

idx = DataDE\_br\_SG.loc[:,'1960':].columns

cols = DataDE\_br\_SG.loc[:,'Indicator Name']

data = DataDE\_br\_SG.loc[:,'1960':]#.as\_matrix()

# ## Descripción del conjunto de datos.

# En Cualquier conjunto de datos real, existe mucha información asociada con los datos intrínsecos. Estos suelen llamarse metadatos y describen información sobre el conjunto de datos principal, incluidos elementos como las etiquetas que se utilizan para las filas y/o columnas, brindando un detalle de la recopilación de datos y explicaciones sobre el significado de los datos.

#

# La hoja Country en el archivo es WDIEXCEL.xlsx, contiene información sobre el significado de las etiquetas de datos para las diversas series de tiempo contenidas en los datos.

#

# Los datos reales están en la hoja Data de WDIEXCEL.csv. Como este archivo contiene ainformación de metadatos, podremos hacer todo el trabajo utilizando solo este archivo.

#

# Es posible elegir una de las columnas del archivo de datos como índice pasando el parámetro index\_col al método read\_excel(), specificando su posición o etiqueta en el archivo.

#

# Se requiere algo de trabajo para poner en un formato adecuador para usarlo para la visualizar datos. Las filas del archivo contienen una serie temporal de datos anuales sobre un indicador económico.

#

# El archivo de data esta descargado en formato .xlsx:

#

# ### WDIEXCEL.xlsx

#

# Las hojas dentro del archivo se describen a continuación:

#

# #### Data

#

#     Country Name: Nombre del país.

#     Country Code: Código del país. Sirve para establecer una relación con otro conjunt de datos.

#     Indicator Name: Nombre del indicador.

#     Indicator Code: Código de indicador. Puede servir para establecer relacion con otro conjunto de datos.

#     Identificación para los años desde 1960 a 2016: una columna CADA UNO para el valor de la variable en cada año que estuvo disponible.

#

# #### Country

#

#     Código de país

#     Nombre corto

#     Nombre de la tabla

#     Nombre largo

#     2-alphacode

#     Unidad monetaria

#     Notas especiales

#     Región

#     Grupo de ingresos

#     Código WB-2.

#     Año base de cuentas nacionales

#     Año.de.referencia.de.cuenta.nacional

#     SNA.price.valuation

#     Lending.category

#     Otros.grupos

#     Sistema.de.cuentas.Nacionales

#     Factor.conversión.alternativa

#     PPP.survey.year

#     Balance.de.pagos.Manual.en.uso.

#     External.debt.Reporting.status

#     Sistema.de.comercio

#     Concepto.de.contabilidad.de.gobierno

#     Norma.de.difusión.de.datos.del.MIF

#     Latest.population.census

#     Última.encuesta.de.hogares.

#     Fuente.de.los.datos.de.ingreso.y.gasto.más.recientes

#     Vital.registration.complete

#     Último.censo.agrícola

#     Latest.industrial.data

#     Latest.trade.data

#     Últimos.datos.de.extracción.de.agua.

#

# #### FootNote

#

#     Código de país

#     SeriesCode

#     Año

#     DESCRIPCIÓN

#

# ### Series-Time

#

#     SeriesCode

#     Año

#     DESCRIPCIÓN

#

# #### Series

#

#     Serie.Código

#     Tema

#     Indicator.Name

#     Definición corta

#     Definición larga

#     Unidad de medida

#     Periodicidad

#     Base.Period

#     Otras notas

#     Método de agregación

#     Limitaciones y excepciones

#     Notes.from.original.source

#     Comentarios generales

#     Fuente

#     Concepto.y.metodología.estadística

#     Relevancia.desarrollo

#     Vínculos.fuentes.relacionados

#     Otros.vínculos.web

#     Indicadores.relacionados

# ## Análisis del conjunto de datos.

#

# A nivel mundial la proporción de sexos en los seres humanos es de aproximadamente 1: 1 por lo que una hipótesis errónea podría ser que habría el mismo número de hombres y mujeres en todos los países del mundo.

#

# Pero no es el caso. Existen muchas razones por las que el equilibrio general de género de un país difiere del 50/50 de otro. Algunas de estas hipotesis son:

#

# #### Diferencias en la migración:

#

# Algunos países atraen o importan un gran número de trabajadores extranjeros para trabajos con desequilibrio de género, como la construcción o el cuidado de niños.

#

# #### Diferencias en las tasas de mortalidad:

#

# Estos pueden incluir:

#

# Mayores tasas de mortalidad masculina (por ejemplo, en países que han experimentado períodos prolongados de guerra (y donde es más probable que los hombres sean combatientes)).

#

# Mayores tasas de mortalidad femenina (por ejemplo, en países con altas tasas de muerte durante el parto, o donde los sistemas legales y sociales conducen a un aumento del infanticidio femenino)

#

# #### Discriminación de género

#

# Existe la hipótesis de que la discriminación de género obstaculiza el crecimiento económico.

#

# La relación negativa entre la discriminación basada en el género y el ingreso per cápita es alta, esto sugiere que el costo económico de la discriminación por motivos de género ocasiona pérdida de ingresos.

#

# Ahora que tenemos una idea de los fenómenos sociales que podríamos estar viendo, comenzamos a revisar los datos:

# Tipos de datos datos del dataset

DataDE.dtypes

# Consulta de valores nulos en las columnas del dataset.

# %%

l1 = []

for col in DataDE.columns:

    sum1 = DataDE[col].isnull().values.sum()

    if sum1 > 0:

        l1.append(col)

        # Comprobamos valores nulos

        print('%s valores nulos en la columna' % DataDE[col].isnull().values.sum(), col )

print(l1)

print(len(l1))

# Consulta de los tipos de columnas existentes (numéricas, clasificatorias, texto)

tipo\_text\_cat = ['Country Name','Country\_Code', 'Indicator Name', 'Indicator Code']

tipo\_num\_cat = []

tipo\_texto = ['Country Name','Country\_Code', 'Indicator Name', 'Indicator Code']

tipo\_num = [

'1960','1961','1962','1963','1964','1965','1966','1967','1968','1969','1970','1971','1972','1973','1974','1975','1976','1977','1978','1979','1980','1981','1982','1983','1984','1985','1986','1987','1988','1989','1990','1991','1992','1993','1994','1995','1996','1997','1998','1999','2000','2001','2002','2003','2004','2005','2006','2007','2008','2009','2010','2011','2012','2013','2014','2015','2016','2017','2018','2019','2020'

]

all\_type = []

all\_type.extend( tipo\_text\_cat )

all\_type.extend( tipo\_num\_cat )

all\_type.extend( tipo\_texto )

all\_type.extend( tipo\_num )

print( all\_type )

for name in DataDE.columns:

    if name not in all\_type:

        tipo\_num\_cat.append( DataDE )

print('Categorías Texto'    , len( tipo\_text\_cat ))

print('Categorías Numéricas', len( tipo\_num\_cat ))

print('Campos tipo Texto'   , len( tipo\_texto   ))

print('Campos tipo Número'  , len( tipo\_num     ))

# El resumen indica que existen 4 columnas con datos categóricos de tipo texto y 61 campos con valores numéricos.

# Existen también atributos categóricos, los cuales clasifican los acuerdos según características generales. Se describen a continuación:

#

# ### Country Name: Países/poblaciones donde se ha efectuado el estudio.

#

# ### Región:

#

# Región del mundo a la que pertenece cada pais siendo las sisguientes regiones:

#

#   África

#   Americas

#   Asia and Pacific

#   Europe and Eurasia

#   Middle East and North África

#   Cross-regional

#   Other

#

# ### Indicator Name: Nombre del indicador poblacional.

#

#   2005 PPP conversión factor, GDP (LCU per international $)

#     2005 PPP conversion factor, private consumption (LCU per international $)

#     Access to electricity (% of population)

#   Other: Otros indicadores

#

# ### Topic: Tema del indicador

#

#   Agricultura y producción

#   Land use

#   Economia

#   Otros

# ### Años desde 1960 a 2020

#

#   Valor que indica el año que se suscitó el indicador.

# Identifica los valores de las categorías de tipo textual.

for name in tipo\_text\_cat:

    values = DataDE[name].unique()

    print(name, len(values), values)

# Comenzaremos a analizar indicadores de género para algunos paises. Tomamos com ejemplo Zimbabwe, primero extraemos la información referente para este pais.

# Zimbabwer (indicador de género 2019)

ZWE = DataDE[(DataDE["Country Name"]=='Zimbabwe')].reset\_index(drop=True)

ZWE.head(2)

# Graficamos el resultado

color = plt.cm.Set1(np.linspace(0,1,20))

ZWE["Indicator Name"].value\_counts().sort\_values(ascending=False).head(20).plot.pie(y="2019",colors=color,autopct="%0.1f%%")

plt.title("Zimbabwe Indicators 2019")

plt.axis("off")

plt.show()

# Extraemos los países Arabes

ARB = DataDE[(DataDE["Country Name"]=='Arab World')].reset\_index(drop=True)

ARB.head(1)

# Graficamos datos referentes a los indicadores de género.

color = plt.cm.RdBu(np.linspace(0,1,20))

ARB["Indicator Name"].value\_counts().sort\_values(ascending=False).head(10).plot.pie(y="2020",colors=color,autopct="%0.1f%%")

plt.title("Arab World Indicators 2020")

plt.axis("off")

plt.show()

# Ahora vemos los indicadores para Brasil

BRA = DataDE[(DataDE['Country Name']=='Brazil')].reset\_index(drop=True)

BRA.head(1)

# Visualizamos indicadores para el año 1960

color = plt.cm.summer(np.linspace(0,1,20))

ARB["Indicator Name"].value\_counts().sort\_values(ascending=False).head(10).plot.pie(y="1960",colors=color,autopct="%0.1f%%")

plt.title("Brazilian Indicators 1960")

plt.axis("off")

plt.show()

# Ahora visualizamos indicadores para el año 2020

color = plt.cm.winter(np.linspace(0,1,20))

ARB["Indicator Name"].value\_counts().sort\_values(ascending=False).head(20).plot.pie(y="2020",colors=color,autopct="%0.1f%%")

plt.title("Brazilian Indicators 2020")

plt.axis("off")

plt.show()

# Ahora creamos una gráfica la cual describe la relacion entre los porcentajes y cantidades establecidas en el año 2020 para el indicador respectivo representado en el eje de las Y. El tipo de gráfica es "viridis", algo parecido a un mapa de calor

fig,ax = plt.subplots(1,1, figsize=(8,8))

degree = BRA.groupby(['Indicator Name', '2020']).size().unstack()#.fillna(0)

degree = degree.div(degree.sum(axis=1), axis=0)

sns.heatmap(degree, cmap='viridis')

SeriesDE = pd.read\_excel("Data\_Exploration/WDIEXCEL.xlsx",sheet\_name="Series")

DataDE\_mg = pd.merge(DataDE, CountryDE[['pkCountry\_Code','Region']], left\_on="Country\_Code", right\_on="pkCountry\_Code")

DataDE\_mg = pd.merge(DataDE\_mg, SeriesDE[['Series Code','Topic']], left\_on="Indicator Code", right\_on="Series Code")

DataDE\_mg.head(5)

DataDEdrop = DataDE\_mg.drop(['Country\_Code','Indicator Code','Region','pkCountry\_Code','Series Code' ],axis=1)

DataDEdrop.head(5)

DataDE\_filter = DataDEdrop.loc[DataDE\_mg['Topic'].str.contains("Gender", case=False)]

print(DataDE\_filter.shape)

g = DataDE\_filter.groupby('Indicator Name')

g.groups.keys()

g = DataDE\_filter.groupby('Topic')

g.groups.keys()

DataDE\_Agency = DataDE\_filter.loc[DataDE\_mg['Topic'].str.contains("Gender: Agency", case=False)]

DataDE\_Health = DataDE\_filter.loc[DataDE\_mg['Topic'].str.contains("Gender: Health", case=False)]

DataDE\_Participation = DataDE\_filter.loc[DataDE\_mg['Topic'].str.contains("Gender: Participation & access", case=False)]

DataDE\_LifePublic = DataDE\_filter.loc[DataDE\_mg['Topic'].str.contains("Gender: Public life & decision making", case=False)]

DataDE\_Agency.dropna(how='all', axis=1, inplace=True)

DataDE\_Health.dropna(how='all', axis=1, inplace=True)

DataDE\_Participation.dropna(how='all', axis=1, inplace=True)

DataDE\_LifePublic.dropna(how='all', axis=1, inplace=True)

# Exporta el nuevo dataframe para realizar la respectiva visualizacion de datos

DataDE\_Agency.to\_excel("DataAgency.xlsx", sheet\_name='Agency' , index=False)

DataDE\_Health.to\_excel("DataHealth.xlsx", sheet\_name='Health' , index=False)

DataDE\_Participation.to\_excel("DataParticipation.xlsx", sheet\_name='Participation' , index=False)

DataDE\_LifePublic.to\_excel("DataLifePublic.xlsx", sheet\_name='LifePublic' , index=False)