Tabla de contenido

Proceso

Extracción

Web Scraping

IDH (web-scraping-idh.R)

Para la obtención del índice de desarrollo humano histórico de México de 1990 a 2010, se siguieron los siguientes pasos:

1. Descargar la página web, como estrategia para facilitar su extracción ya que la tabla se carga de manera dinámica vía llamadas JavaScript.
2. Se realizó una extracción de los datos a través de la libraría **rvtest**
3. Se extrajo la fila correspondiente a México
4. Se separaron los datos y se realizó limpieza de los mismo
5. Se guardó en un variable para su posterior uso.

Obtención de listado de servicios de mapa de comunidades indígenas (web-scraping-mapas-services.R)

En este caso el proceso realizado es más complejo, ya que proviene de dos cuentas distintas de CARTO (Pueblos Indígenas y AtlasMX).

Pueblos Indígenas

1. Se recorren las 6 páginas de datasetets obteniendo los nombres de todos
2. Se descartan aquellos que no contenga el sufijo “doc”, ya que se comprobó que no se usan.
3. Se consulta la URL de cada uno de los dataset, obteniendo el nombre del mapa en el que se está usando que al final será el nombre del pueblo indígena que representa.
4. Se construye un dataframe con el nombre de todos los pueblos y su la URL que representa el servicio que consultaremos.

AtlasMX

1. Se recorre las cinco páginas de mapas.
2. Se extra la URL de cada mapa.
3. Por cada URL consultada se extrae el nombre del datase en uso y el nombre del mapa
4. Se guarda en el dataframe el nombre del mapa junto con el URL del servicio por cada set en uso en cada mapa.

Una vez realizados ambos procesos, se mezclan ambos dataframes y se guarda el valor de la variable.

Archivos

En el caso de los archivos, solo se realizaron los siguientes pasos:

1. Se descargaron los archivos.
2. Se renombraron por la siguiente nomenclatura:
   [tipo_poblacion]_[segregación] _[dato_representado]_[años]
   tipo_población:

   a. **pi** = Población Indígena
   b. **pg** = Población General

   segregación:

   c. **sex** = Sexo
   d. **edo** = Por entidad fedrativa

   dato representado:

   a. **idh** = Índice de desarrollo humano
   b. **gm** = Grado de marginación

*También se incluye un archivo GEOJSON que contiene los polígonos de los estados de México. (estados-de-mexico.geojson). Dichos documentos se encuentran en el directorio "/Preprocesado/Files").

Servicios

En el caso de uso de servicios (81 servicios obtenidos del web-scraping-mapas-services.R), se consultaron todos los servicios y se organizó la información en un solo dataframe. (Consultar archivo /Preporcesado/Services/services-map.R)

Se realizó la homologación de los pueblos indígenas al catálogo de lenguas indígenas, ya que la información recabada no tiene información que permitiera su filtrado y comparación con el resto de los datos.

También se realizó limpieza del HTML que se mostrará al hacer clic sobre el punto en el mapa.

Tratamiento y preparación

Para el tratamiento y preparación se retoman las variables guardadas en el proceso de extracción, y se aplicaron distintitos y variados proceso de transformación ("/Preprocesado/preparation.R").

Se creó un dataframe nuevo para cada gráfico y mapa, necesario y se guardaron a través de la función **save** (ver el directorio "/Preprocesado/charts-data")**,** para su recuperación en los archivos de visualización.

Las acciones más comunes fueron la mezcla de dataframes a través de la función **merge** y **cbind**, el método **gather** para convertir a formato llave-valor distintas columnas, se ordenaron o agruparon datos con **arrange**, **order** y **group_by,** entre otras cosas como, búsqueda de patrones a través de la función **gsub, grepl**, y el renombramiento de columnas.

Las aplicaciones de visualización estática y dinámica se encuentran en las carpetas con el mismo nombre (visualizacion-estatica.R y visualizacion-dinamica.R), logrando así una organización más entendible y que permita la separación adecuada del proceso de recolección, proceso y visualización de nuestra aplicación.

Justificación

Las gráficas que se muestran tanto en la visualización dinámica como estática fueron elegidas con respecto al tipo de datos tratados.

En principio significó una limitación considerable el tipo de información seleccionada, así como el grado de fragmentación y poca homologación entre ellos, sin embargo aun cuando la información no es rica como una serie temporal o datos de contexto geográfico más complejo, la selección y construcción de gráficos se realizó con el criterio de poder contar una historia desde un contexto histórico llegando a la actualidad, desde un contexto nacional, estatal y llegando a un contexto de comunidad-lengua indígena, así como desde un contexto de poblacional general y su comparativo indígena, asociando en todo momento índices de calidad de vida y marginación (IDH y GM respectivamente).

Por la naturaleza de los datos se seleccionaron gráfico como el de barras, pirámide poblacional y mapas, ya que son los que representaban mejor el objetivo comparativo antes descrito.

Conclusiones (Insights)

De manera general, fue interesante encontrar que gráficos muy distintos pueden contar la misma información, cada uno fortaleciendo o enriqueciendo una narrativa distinta.

El **StoryTelling** es una herramienta sólida para guiar al lector a través de los insights encontrados en la exploración y entendimiento de los datos, es una forma más agradable de mantener la atención del lector y nos puede servir para centrar su atención en el objetivo o mensaje que queremos brindarle de forma concreta. Sin embargo, el problema al que me enfrenté fue la necesidad de acotarlo, de hacerlo breve y perder la descripción de otros insights encontrados que no eran directamente relacionados a la temática que se intentaba describir.

En cambio, el **dashboard** me permitió que de manera implícita estos insights perdidos en el storytelling quedaran disponibles para nuevos usuarios o lectores, sin embargo, queda al interés, entendimiento y conocimiento del usuario la interpretación de los datos.