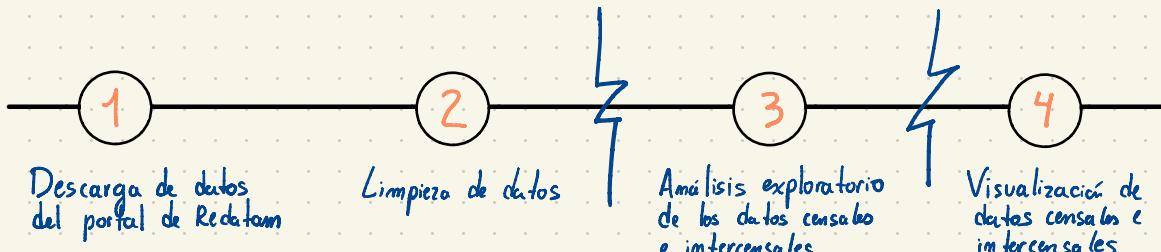


Flujo de trabajo de Redapay



- Descarga del excel
- Scrapping → Redapay - query
- } Redapay . Frecuencia
- } Redapay . tabla cruzada
- Continuous = True
- Pivot = True

- Datos univariados → Funciones de limpieza (Frecuencias)

- Var categórica
- Var escala raro

- Multivariados
(Cruce de variables)

→ Funciones de limpieza

- Var categórica
- Var escala raro
- Var categórica

- Descriptivos (Suma, desv std, min, max, cuartiles, quintiles)

- Series de tiempo intercensal

- Tasas de crecimiento intercensal

- Gráficos de barras

- Mapas de frecuencias (objetos / vectores)



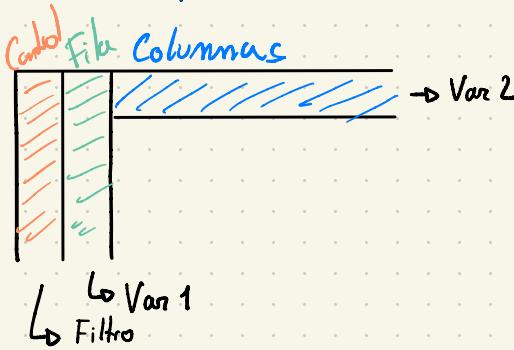
↓
Esta sección debe apuntar a recibir retroalimentación de los usuarios del código

Apuntes Reunión 19/08/22

- Ajustar el código de extracción Redatum Univariable para extraer varias variables univariadas al mismo tiempo → Opción B, hacen un loop (posiblemente poco eficiente)
- Crear el código para extraer tablas de cruce de variables
Adicional: es posible crear un cruce de más de 2 variables usando el lenguaje Redatum?
- Probar los descriptivos de Pandas para evitar scrapear la sección de descriptivos del Redatum.
- Buscar info abierta espacial para los mapas
- Crear las funciones de limpieza faltantes

Cuestiones pendientes

- ¿Cuándo hacer la limpieza de caracteres? → Mayúsculas, caracteres especiales, etc.
- ¿Scraping debe tener otro módulo que se importa por separado? → From readpy import Query
- ¿Cuáles son todos los inputs del scraping?
 - área, var1, var2, selection, filter_a, otros?
- Filtros → Dónde es más eficiente realizar el filtro?
 - Redapy - query → En el pedido de información al procesador de Redatam
 - Redapy → ¿En el multimedex o en el pivot?
 - Durante la limpieza?
- Falta hacer la prueba con una tabla cruzada con variable de control.



- Es necesario documentar funciones → Español o Inglés?
- Visualización de datos
 - Gráficos de frecuencias
 - Mapas

¿Qué librería se va a utilizar?

 - Matplotlib
 - Seaborn
 - Plotly
- Desarrollar readme → Estructura, ejemplos, gráficos
- ¿Cómo lo publicamos? → Reseña, artículo, etc?

Librerías

- Python 3.8

Redapaj - cleaning

- Pandas → Se utiliza para la limpieza y manejo de data frames
- Numpy → Se instala al instalar pandas. Ayuda a realizar operaciones de forma más eficiente

Redapaj - query

- lxml → read_html usa por default "lxml", en segundo lugar usa "bs4" + "html5lib".
A pesar de haber instalado beautifulsoup, read_html no jala "bs4"
"lxml", se instala usando pip install
- BeautifulSoup → No queda claro donde se usa esta librería.
- Selenium → Requiere driver de Chrome o firefox → especificar la ruta del .exe
→ Se necesita actualizar la versión y cambiar un comando
- Requests → -
- import_ipynb → Esta librería no debería ser necesaria si las funciones (query y make_query)
son convertidas a ".py"

Visualización

Matplotlib → Librería base para visualizar datos.

Geopandas → Permite trabajar con datos vectoriales con sistemas de coordenadas geográficas