**Taller 1**

**Autores:** Juan Sebastián Vásquez Acevedo y Walter Leonardo Sanchez Salazar

**Fecha inicio:** 11/06/2022

Lo primero que se realizo fue la extracción de los datos. En este proceso se empleo un scraping de la pagina web https://ignaciomsarmiento. github.io/GEIH2018 sample/, en el cual no se presento ninguna restricción en el acceso a la información. Para poder realizar este proceso se hizo uso del Software R y de las librerías: tidyverse y rvest.

En primera medida, se incluye en un objeto la URL del problem set, seguido se logra definir esta como un HTML. En el análisis de la pagina web se identifica el “XPATH” el cual brindará la ruta de los datos. Es así como se logra encontrar los nodos de los cuales se extraen los links que contienen las tablas que conforman la base de datos. Por medio de un tibble se unen tanto los nombres de los chunks como los links de interés para la extracción de los datos. En este proceso logramos identificar que el primer link de los chunks no contiene la tabla de la base de datos, por lo cual se hace necesario encontrar el link “verdadero” con el cual lograremos extraer la tabla que contiene los datos de interés. Luego de una exploración del código fuente de la pagina web se logra encontrar el link “verdadero” en el cual se identifica una particularidad, y es que en este link solo cambia el número o indicador del chunk, este número va del 1 al 10 y hace referencia al número de tablas que conforman la base de datos. Con este link “verdadero” logramos encontrar la puerta de entrada al proceso de extracción de los datos. Para lograr este proceso de extracción se hizo uso de un “loop”.

El “loop” se encarga de entrar a la pagina y obtener el link “verdadero” para cada una de las tablas que conforman la base de datos que como se menciono anteriormente va del 1 al 10. Luego de ello extrae las tablas de las bases de datos y las convierte en un data frame, almacenándolas por medio una unión vertical en un data frame vacío previamente definido.