Autores:

- Alfredo Rubio Navarro
- Gabriel Loja Rodas

1. Contexto. Explicar en qué contexto se ha recolectado la información.
   Explique por qué el sitio web elegido proporciona dicha información.

   La idea es obtener una visión general de la evolución que ha tenido el cáncer a lo largo del tiempo en cuanto a nuevos casos diagnosticados y las muertes derivadas separando tanto por sexo como por raza (blanca/negra).
   El hecho de que los datos obtenidos sean localizados en EEUU hace que la distinción por raza sea relevante.
   Los datos han sido obtenidos de la página del SEER (El Programa de Vigilancia, Epidemiología y Resultados Finales).
   El SEER del NCI (National Cancer Institute) es una colección de registros centrales de cáncer en los Estados Unidos que recogen datos sobre la incidencia, la mortalidad y la supervivencia de los distintos tipos de cáncer.
   SEER es una fuente autorizada de estadísticas de cáncer en los Estados Unidos.
   La página principal que recoge toda esta información es https://seer.cancer.gov/.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

   El nombre elegido para el dataset es “evolucionCancer”.

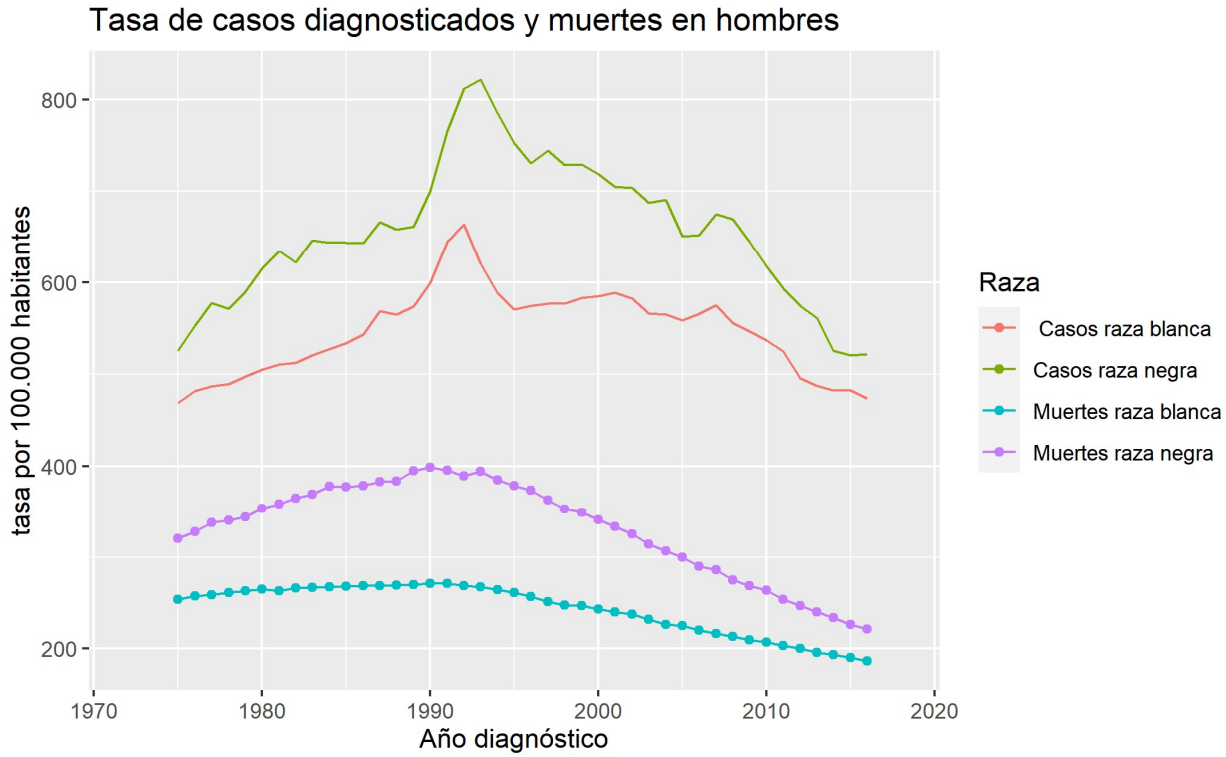
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

   Evolución histórica de los nuevos casos y la mortalidad para cada tipo de cáncer desglosado por sexo y raza.

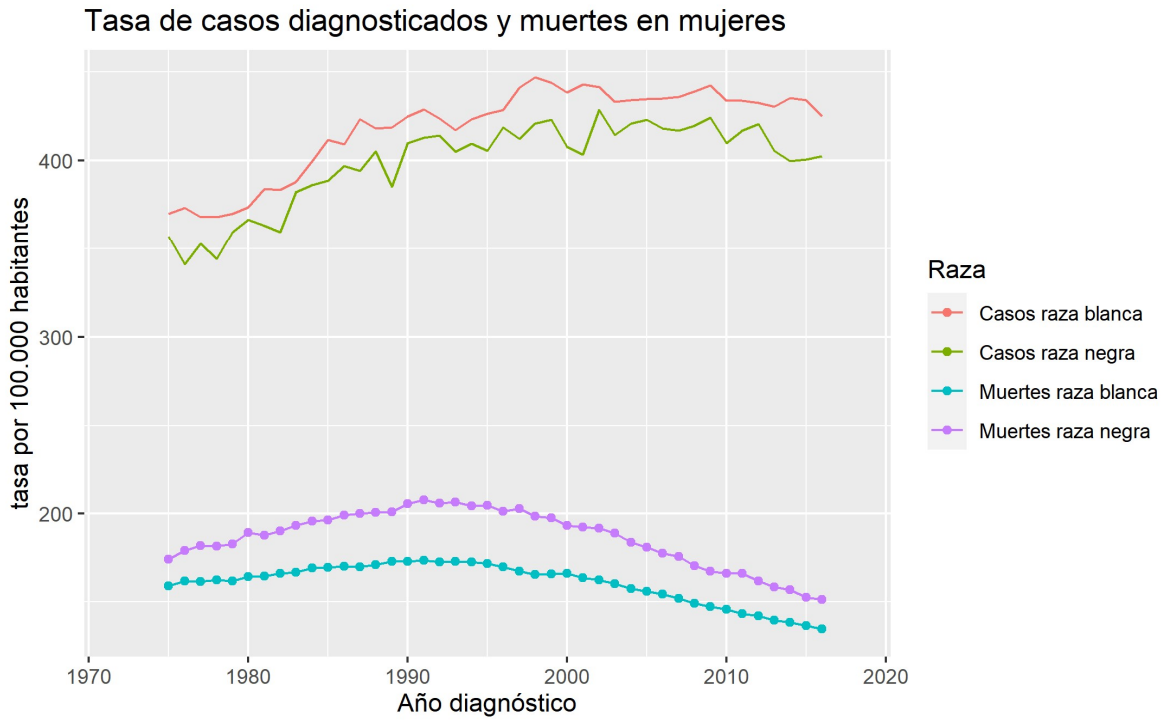
4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.

El dataset principal es muy amplio, ya que recoge los tipos de cáncer principales (42 tipos en total). Para hacer una representación abreviada de los datos, hemos basado las siguientes gráficas en los valores agregados de todos los tipos de cáncer.
Hemos diseñado la aplicación en Python para que nos permita también extraer los datos agregados.

Por un lado se muestran los datos relativos a los hombres separados por raza:



Por otro lado las mujeres, también separadas por raza:



5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

El programa de recogida de datos permite seleccionar distintas opciones:
- Obtener los totales de todos los tipos de cáncer o desglosado por cada tipo.
- Obtener los datos de un año en concreto dentro de los disponibles o todos los años.
- Desglosar el dataset por sexo y/o raza.

Para nuestro caso, nos interesa conocer los totales de casos y muertes de todos los tipos de cáncer para todos los años disponibles (1975 a 2016) y desglosado por sexo y raza, de forma que el comando a emplear sería: `python3 query.py –s –r`

De esta forma, los campos que compondrán el dataset son:

Campo	Descripción
tipo_cancer	El nombre del tipo de cáncer
año	El año al que corresponden los datos: año de diagnóstico o año del fallecimiento
casos_hombre_raza_blanca	Nuevos casos diagnosticados en hombres de raza blanca
casos_mujer_raza_blanca	Nuevos casos mujeres de raza blanca
casos_hombre_raza_negra	Nuevos casos hombres de raza negra
casos_mujer_raza_negra	Nuevos casos mujeres de rana negra
muertes_hombre_raza_blanca	Fallecimientos en hombres de raza blanca
muertes_mujer_raza_blanca	Fallecimientos en mujeres de raza blanca
muertes_hombre_raza_negra	Fallecimientos en hombres de raza negra
muertes_mujer_raza_negra	Fallecimientos en mujeres de raza negra

Observaciones:
- Los valores de nuevos casos están expresados como el número de cánceres por cada 100.000 habitantes en situación de riesgo.
- El número de nuevos cánceres puede incluir múltiples cánceres primarios que se producen en un paciente (El sitio primario notificado es el sitio de origen y no el sitio metastásico).
- El número de muertes por cáncer viene expresado como el número de cánceres por cada 100.000 habitantes en riesgo.
- Para los tipos de cáncer que se presentan en un solo sexo, se utiliza la población específica del sexo (por ejemplo, las mujeres para el cáncer de cuello uterino).
- La población utilizada para el estudio está ajustada por edad, lo que permite comparaciones de poblaciones ya que tiene en cuenta las diferencias de distribución de edades entre las poblaciones.
  El ajuste por edad toma la distribución de la población de los Estados Unidos en el año 2000 y la aplica a otros períodos de tiempo que se están considerando. Esto asegura que tales tasas no reflejen ningún cambio en la distribución de edad de la población.

La metodología para la obtención de los datos ha sido aprovechar una de las opciones que tiene la página de realizar una consulta parametrizada rellenando una serie de formularios que nos permiten seleccionar los datos que buscamos (https://seer.cancer.gov/canques/).
Tras completar los datos la página nos muestra los resultados en una URL con determinadas variables (parámetros HTML de tipo GET).

Basándonos en las variables que generan las distintas consultas identificamos las que son necesarias para confeccionar las consultas que nos interesan.
La página utiliza el acento circunflejo (^) para separar elementos dentro de cada variable

Las variables representativas en nuestro caso son:

- dir/db:
  'seer2016' / '1' para la base de datos con las tasas de nuevos casos
  'usmort2016' / '7' para la base de datos con las tasas de mortalidad
- sel: Es una lista con las distintas opciones de selección y filtrado de los datos a mostrar.
- x: variables para el desglose de las filas
- y: variables para el desglose de las columnas
- dec: el número de decimales a mostrar

El programa Python primero genera las 2 URLs basándose en los datos pasados por línea de comandos con las variables necesarias para obtener la información que buscamos.

Para cada una de las dos consultas, la de nuevos casos y la de mortalidad, procesa el código HTML resultante de la petición y extrae los datos individualmente de cada base de datos para luego unificarlos en un solo diccionario que luego volcaremos al fichero csv.

Las consultas están diseñadas para que sean los mismos tipos de cáncer y los mismos años de estudio en ambos casos, de lo contrario los valores obtenidos no serían comparables.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Antes de extraer los datos, hemos revisado el contenido del fichero robots.txt.
Por suerte el fichero existe pero está en blanco, lo cual nos indica que no hay partes restringidas.

Lo único que el sitio menciona en cuanto a la seguridad es que utilizan programas de software para monitorear el tráfico para identificar intentos no autorizados de cargar o cambiar información o causar daños, que no es nuestro caso.

De todas formas, a la hora de extraer la información hemos usado “Firefox” como “User Agent” para asegurarnos de que la página no nos bloquea al no tratarse de un navegador web.

Agradecemos al National Cancer Institute (agencia del gobierno de los Estados Unidos para la investigación del cáncer) y en concreto al programa SEER (Programa de Vigilancia, Epidemiología y Resultados Finales) la publicación sin derechos de autor de los datos necesarios para nuestro estudio sobre la evolución del cáncer según figura en su política de reutilización de la información (https://www.cancer.gov/policies/copyright-reuse).

Fuente de los datos: https://seer.cancer.gov/

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

   No cabe duda de que el cáncer se ha convertido en una de las causas de mortalidad más importante de nuestra época. Por desgracia todos tenemos algún caso en nuestro entorno más o menos cercano. Abordar el estudio de esta enfermedad desde cualquier punto de vista creemos que es importante de cara a poder erradicarla o curarla algún día.
   La pregunta que intenta responder el estudio es cómo afecta cada tipo de cáncer a las personas según su raza y su sexo.
   Por desgracia, los datos que hemos podido obtener no están muy actualizados (hasta 2016) y los únicos datos que son más recientes corresponden a estimaciones en el año 2019.
   A pesar de ello nos puede dar una idea de cual es la tendencia que siguen.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

   - Released Under CC0: Public Domain License
   - Released Under CC BY-NC-SA 4.0 License
   - Released Under CC BY-SA 4.0 License
   - Database released under Open Database License, individual contents under Database Contents License
   - Other (specified above)
   - Unknown License

   La fuente original del dataset está libre de derechos de autor y puede reutilizarse libremente según https://www.cancer.gov/policies/copyright-reuse.

   Ello significa que los datos originales son de dominio público y por lo tanto nuestro dataset también debería calificarse como tal.

   Las licencias CC no se aplican a material de dominio público (https://creativecommons.org/faq/).

   A la hora de elegir una licencia para el repositorio de GitHub, hemos escogido "The Unlicense", que corresponde a la licencia con menos restricciones, básicamente es de dominio público (https://choosealicense.com/licenses/unlicense/).

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

   https://github.com/alrubnaUOC/cancerEvolutionScraper/blob/master/src/query.py

10. Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

    https://doi.org/10.5281/zenodo.3749385

Contribuciones	Firma
Investigación previa	A.R.N y G.L.R
Redacción de las respuestas	A.R.N y G.L.R
Desarrollo código	A.R.N y G.L.R