Práctica 1:

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

Esta práctica se realiza al mismo tiempo que tienen lugar los últimos debates y las elecciones de Estados Unidos de 2020. Nos parece interesante automatizar una forma de conseguir datos de las encuestas para ver cómo se van posicionando los candidatos a lo largo de la campaña electoral.
También nos puede servir para coger de forma automática los datos de las encuestas de todas las elecciones de las que dispone la página web, y de esta manera disponer de los datos históricos.

Buscando los datos de las encuestas de EEUU encontramos la página web RealClearPolitics (RCP), el cual es un sitio de noticias políticas estadounidense y un agregador de datos de encuestas. Real Clear Politics publica datos de todas las encuestas realizadas por los distintos medios estadounidenses durante las temporadas electorales.

Mediante los datos obtenidos se pueden hacer análisis interesantes como la comparación de la intención de votos en los diferentes procesos de elecciones.

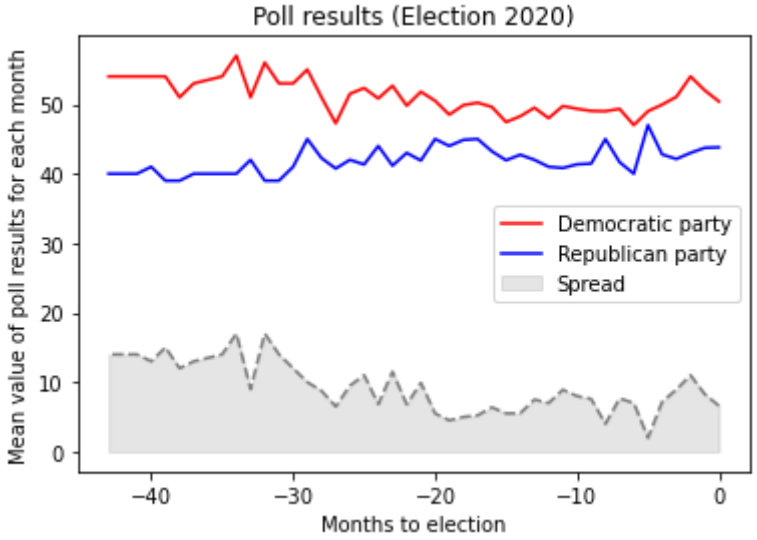
2. Definir un título para el dataset. Elegir un título que sea descriptivo.

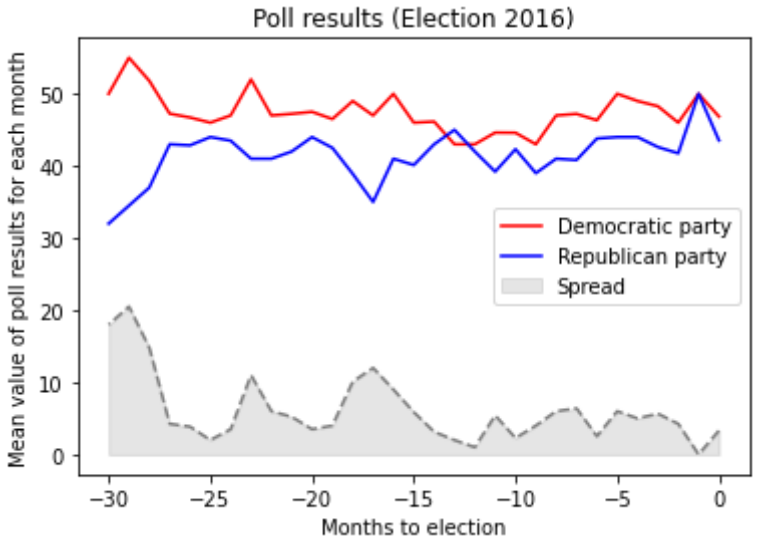
Resultado encuestas de las campañas electorales para presidente de los Estados Unidos (2020-2004).

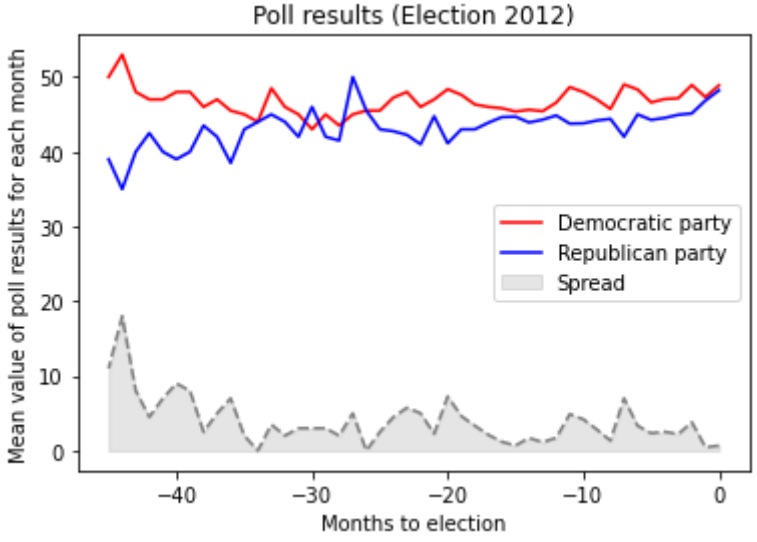
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

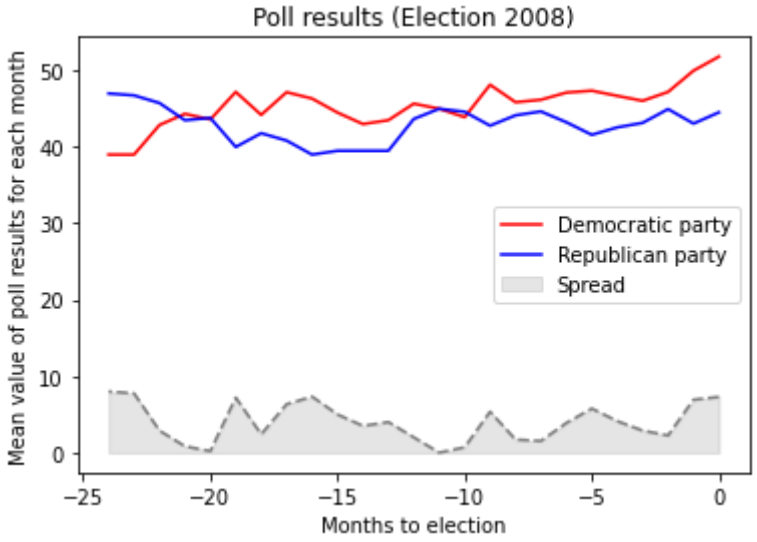
El dataset consta de los datos de las encuestas para la presidencia de Estados Unidos realizadas en Estados Unidos. Las encuestas están realizadas por distintos medios de comunicación (BBC, Fox News, etc...) u otros organismos cómo firmas internacionales de estudios de mercado.

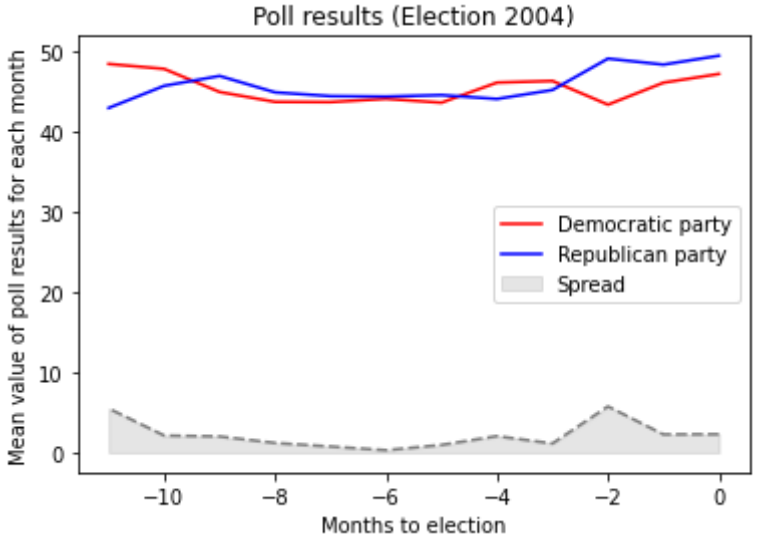
4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente











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

El dataset incluye los datos de las encuestas presidenciales de los años 2020, 2016, 2012, 2008 y 2004.

La tabla de datos presenta las siguientes columnas (excepto para las encuestas de las elecciones de 2004 que no está disponible la columna MoE):
- *Poll*: medio u organismo que realiza la encuesta.
- *Date*: tiempo de inicio y finalización de la encuesta en formato mm/dd - mm/dd.
- Sample: número de personas encuestadas y tipos de personas encuestadas (RG: registered voters/ LV: likely voters /A: adults / V: voters)
- *MoE*: margen de error de la encuesta.
- *Democratic Party*: resultado del Candidato Democratico.
- *Republican Party*: resultado del Candidato Republicano.
- *Spread*: candidato ganador de la encuesta y diferencia en puntos del candidato ganador respecto del candidato perdedor.

- *Election Year*: año de las elecciones presidenciales, no necesariamente de la encuesta ya que hay encuestas realizadas en año previo a las elecciones.

El periodo de tiempo del dataset va desde las encuestas realizadas en 2003 para las elecciones presidenciales de 2004 hasta el 11/2/2020 cuando se realizó la última encuesta para las elecciones presidenciales de 2020.

Cada medio de comunicación u organismo realiza su encuesta y los resultados de estas encuestas RCP los recoge y los muestra en tablas en la página web.

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

Entendemos que el propietario de los datos es el sitio web RealClearPolitics qué es el que recoge los resultados de todas las encuestas realizadas durante la campaña electoral.
Agradecer especialmente a todos los medios de información u otros organismos que se encargan de realizar las encuestas de manera periódica durante la campaña electoral para la presidencia de Estados Unidos.

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

Este conjunto de datos es interesante ya que es la elección del presidente la primera potencia mundial y de esto dependerá el futuro de la economía del mundo.

Está en juego ámbitos de toda índole en cuestiones políticas entre sus primeras instancias, y es interesante poder modelar los datos de estas elecciones de tal manera que se pueda analizar y contestar no solo la incógnita planteada sino encontrar patrones de intención de votos y poder estimar mediante modelos estadísticos y de aprendizaje automático el próximo ganador de la elecciones .

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

- Released Under CC0: Public Domain License

Las licencias Public Domain License se utilizan para hacer que las obras protegidas por derechos de autor puedan ser utilizadas por cualquier persona sin condiciones, al tiempo que se evitan las complejidades de la atribución o la compatibilidad de licencias que se producen con otras licencias.
Hemos elegido esta licencia ya que los datos del dataset son accesibles a cualquier persona que se conecte a la web de RealClearPolitics, por tanto, nos parecía que lo más correcto era utilizar Public Domain License.

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

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait as wait
from selenium.webdriver.support import expected_conditions as EC
import requests
from bs4 import BeautifulSoup
import pandas as pd
from datetime import datetime

driver = webdriver.Chrome('/Users/ester/Downloads/chromedriver')
driver.get("https://www.realclearpolitics.com/epolls/latest_polls/")
wait(driver, 10).until(EC.element_to_be_clickable((By.XPATH, "//span[text()='Poll/Map']"))).click()
elems = driver.find_elements_by_xpath("//a[contains(@href, '/president/us/general_election')][not(@href = following::a/@href)]")
list3 = []
for elem in elems:
    if elem.get_attribute("href").count("vs") == 1: list3.append(elem.get_attribute("href"))

datasets = []
for i in list3:
    page = requests.get(i)
    soup = BeautifulSoup(page.content, 'html.parser')
    heading = [th.get_text() for th in soup.find("tr").find_all("th")] + ['Election Year']
    datasets.append(heading)
    for row in soup.find_all("tr")[1:]:
        dataset = list(td.get_text() for td in row.find_all("td"))
        election = dataset.append(i.split('/')[4])
        if len(dataset) > 6:
            datasets.append(dataset)
        df = pd.DataFrame(datasets,columns=['Poll', 'Date', 'Sample', 'MoE', 'Democratic Party', 'Republican Party', 'Spread', 'Election Year'])
        df = df.drop_duplicates()

now = datetime.now()
date_time = now.strftime("%m_%d_%Y_%H_%M_%S")
df.to_csv(str(date_time + 'eleccionesUSA.csv'), index = False, encoding='utf-8')
```

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

DOI publicación del dataset en Zenodo: 10.5281/zenodo.4263188