Práctica 1: tipología y ciclo de vida de los datos

Descripción de la Práctica a realizar:
El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

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.

Buscando los datos de las encuestas de EEUU encontramos la página web RealClearPolitics (RCP). RCP es un sitio de noticias políticas estadounidense y un agregador de datos de encuestas creado en el año 2000. El sitio 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 tanto en el 2016 como en el 2020 y ver cuál fue la tendencia de los votos del candidato Donald Trump vs sus oponentes en ambas elecciones.

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

Resultados encuestas de las campañas electorales para presidente de los Estados Unidos.

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 extraído 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, etc) u organismos.
No todas las encuestas tienen la misma duración en el tiempo ni se encuesta a la misma cantidad de gente.
Una modificación interesante del primer dataset sería calcular la media (ponderada por la cantidad de gente encuestada) de los resultados de cada mes para ver la evolución por meses.

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

RStudio

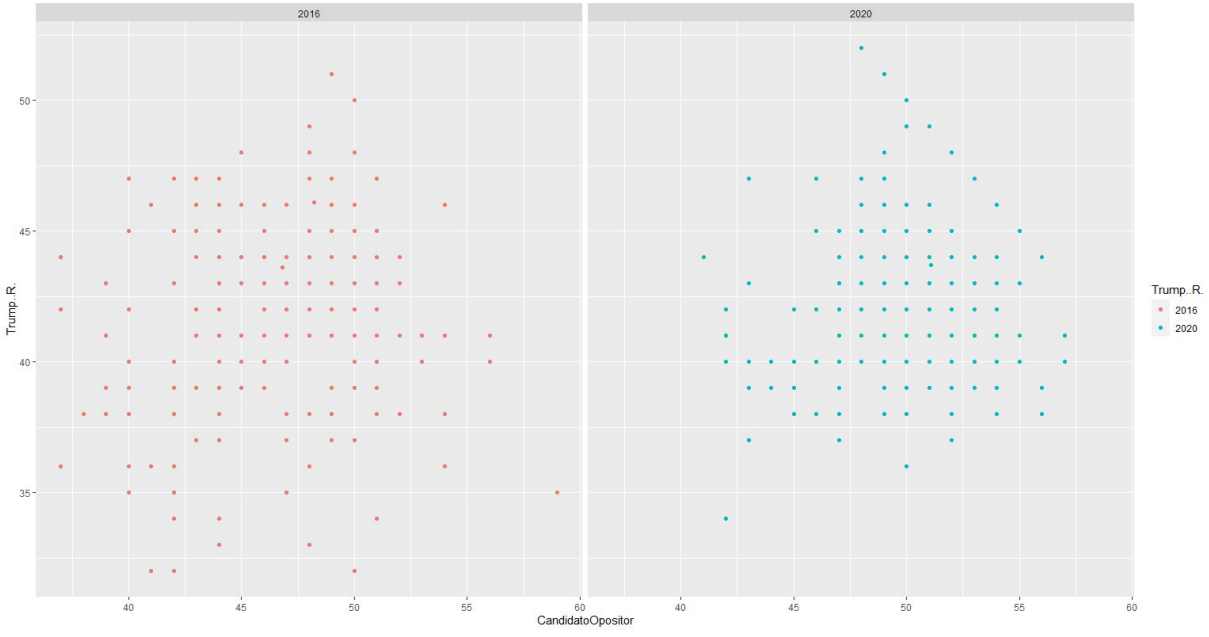
File Edit Code View Plots Session Build Debug Profile Tools Help

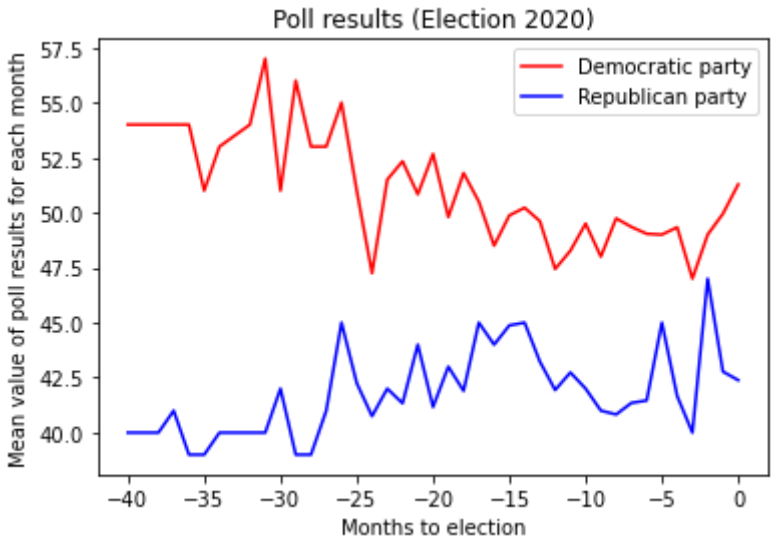
tipologia de datos.R × | base_elecciones ×

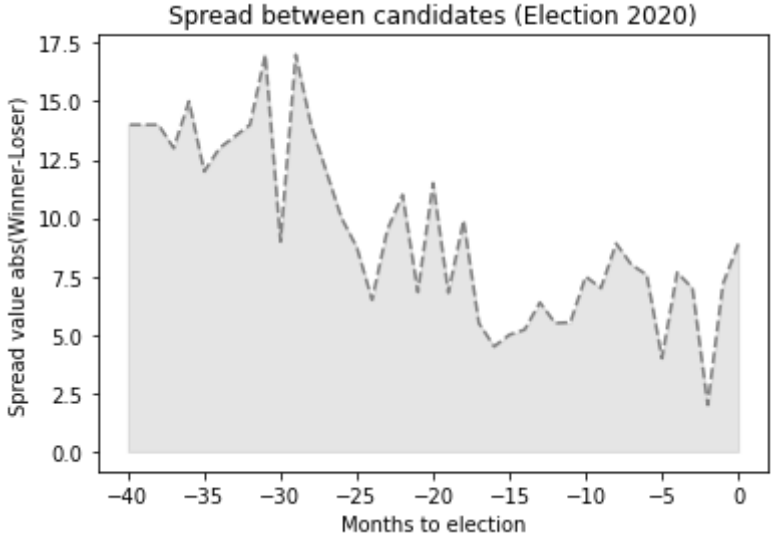
Filter

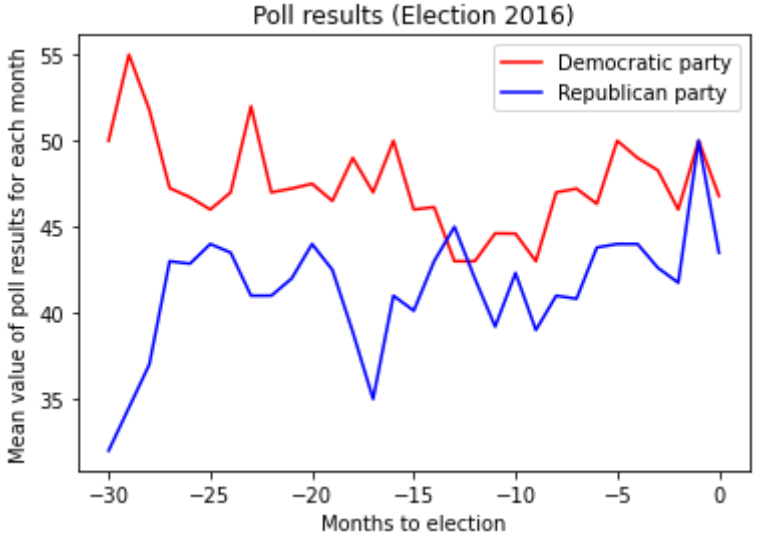
	Poll	Date	Sample	MoE	CandidatoOpositor	Trump..R.	Spread	año
1	Final Results	--	--	--	48.2	46.1	Clinton +2.1	2016
2	RCP Average	11/1 - 11/7	--	--	46.8	43.6	Clinton +3.2	2016
3	BloombergBloomberg	11/4 - 11/6	799 LV	3.5	46.0	43.0	Clinton +3	2016
4	IBD/TIPP TrackingIBD/TIPP Tracking	11/4 - 11/7	1107 LV	3.1	43.0	42.0	Clinton +1	2016
5	Economist/YouGovYouGov	11/4 - 11/7	3669 LV	--	49.0	45.0	Clinton +4	2016
6	LA Times/USC TrackingLA Times	11/1 - 11/7	2935 LV	4.5	44.0	47.0	Trump +3	2016
7	ABC/Wash Post TrackingABC/WP Tracking	11/3 - 11/6	2220 LV	2.5	49.0	46.0	Clinton +3	2016
8	FOX NewsFOX News	11/3 - 11/6	1295 LV	2.5	48.0	44.0	Clinton +4	2016
9	MonmouthMonmouth	11/3 - 11/6	748 LV	3.6	50.0	44.0	Clinton +6	2016
10	NBC News/Wall St. JrnlNBC/WSJ	11/3 - 11/5	1282 LV	2.7	48.0	43.0	Clinton +5	2016
11	CBS NewsCBS News	11/2 - 11/6	1426 LV	3.0	47.0	43.0	Clinton +4	2016
12	Reuters/IpsosReuters	11/2 - 11/6	2196 LV	2.3	44.0	39.0	Clinton +5	2016
13	McClatchy/MaristMcClatchy	11/1 - 11/3	940 LV	3.2	46.0	44.0	Clinton +2	2016

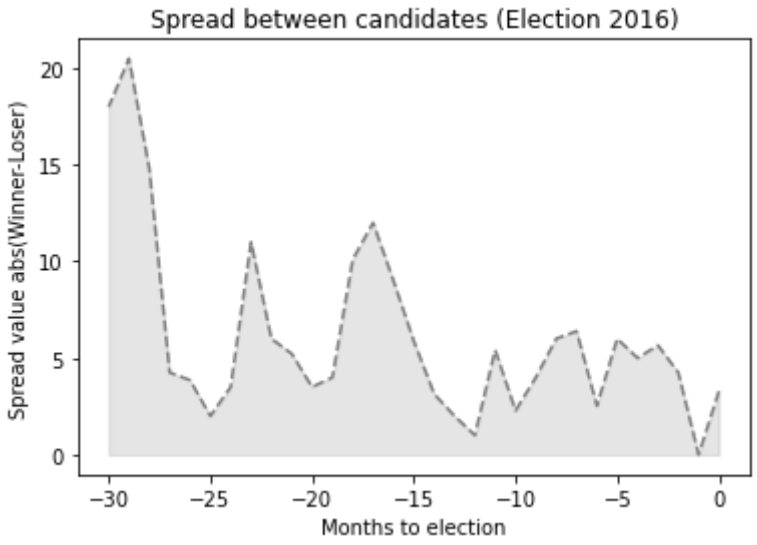
Showing 1 to 15 of 545 entries, 8 total columns




Poll results (Election 2020)
Democratic party
Republican party
Mean value of poll results for each month
Months to election


Spread between candidates (Election 2020)
Spread value abs(Winner-Loser)
Months to election


Poll results (Election 2016)
Democratic party
Republican party
Mean value of poll results for each month
Months to election



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 2016 y 2020. Para cada año hay dos tablas: la tabla con los datos desagregados por encuestas y los datos agregados de las encuestas por mes.

La tabla de los datos desagregados presenta las siguientes columnas:

- Poll: qué medio o organismo realiza la encuesta.
- Date: tiempo de inicio y finalización de la encuesta en el formato mm/dd - mm/dd.
- Sample: número de personas encuestadas y tipos de personas encuestadas (RG: registered voters/ LV: likely voters)
- Margin of Error: margen de error de la encuesta.
- Result for the Democratic Candidate: resultado del Candidato Democratico.
- Result for the Republican Candidate: resultado del Candidato Republicano.
- Spread: candidato ganador de la encuesta y diferencia en puntos del candidato ganador respecto del perdedor.
- Month: mes de la encuesta (valor del mes al finalizar la encuesta).
- Sample Size: número de personas encuestadas.
- Sample Type: tipos de encuestados (RG/LV).
- Winner: candidato ganador de la encuesta.
- Spread value: diferencia entre los dos candidatos (candidatos ganador - candidato perdedor)

La tabla de los datos agregados presentaría las siguientes columnas:

- Candidato A: media ponderada de los resultados de las encuestas durante el mes del candidato A.
- Candidato B: media ponderada de los resultados de las encuestas durante el mes del candidato B.
- Mes: mes realización de las encuestas.

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

(Todos los medios de comunicación u organismos que realizan las encuestas y a la página RCP)

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

Es interesante porque es la elección del presidente la primera potencia mundial 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**
○ **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**

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

```
library(rvest)
library(purrr)
library(tibble)
library(ggplot2)
base_url
<-'https://www.realclearpolitics.com/epolls/2020/president/us/general_election_trump_vs_biden-6247.html#polls'
base_url2<-'https://www.realclearpolitics.com/epolls/2016/president/us/general_election_trump_vs_clinton-5491.html'
pg <- read_html(base_url)
pg2 <- read_html(base_url2)


target2 <- pg2 %>%
  html_nodes(xpath='//*[@id="polling-data-full"]/table')%>%html_table()

target <- pg %>%
```

```
  html_nodes(xpath='//*[@id="polling-data-full"]/table')%>%html_table()

target2=data.frame(target2)
target2$año='2016'
target=data.frame(target)
target$año='2020'

colnames(target2)[5]='CandidatoOpositor'
colnames(target)[5]='CandidatoOpositor'

base_elecciones=rbind(target2,target)

View(base_elecciones)

base_elecciones%>%ggplot(aes(x=CandidatoOpositor,y=Trump..R.,col=Trump..R.))+geom_
point(aes(col=año))+facet_wrap(~año)
```

```
#ELECCIONES 2016
page =
requests.get("https://www.realclearpolitics.com/epolls/2016/president/us/general_election_tr
ump_vs_clinton-5491.html")
soup = BeautifulSoup(page.content, 'html.parser')
headings = [th.get_text() for th in soup.find("tr").find_all("th")]
datasets = []
for row in soup.find_all("tr")[1:]:
    dataset = list(td.get_text() for td in row.find_all("td"))
    datasets.append(dataset)

#CREATING THE DATAFRAME
df = pd.DataFrame(datasets, columns=headings)
df = df.drop_duplicates()
df1 = df.dropna()
df1['Month'] = df1['Date'].str.extract(r'\- (.*?)\/')
df1[['Sample Size','Sample Type']] = df.Sample.str.split(" ",expand=True,)
df1[['Winner','Spread Value']] = df.Spread.str.split(" +",expand=True,)

df2 = df1[['Clinton (D)', 'Trump (R)', 'Month']]
df2 =df2.drop([0,1], axis=0)
type(df2['Month'])
df2['Clinton (D)'] = df2['Clinton (D)'].astype('int')
df2['Trump (R)'] = df2['Trump (R)'].astype('int')
df2['Month'] = df2['Month'].astype('category')

df3 = df2.groupby((df2.Month!=df2.Month.shift()).cumsum()).mean().reset_index(drop=False)
df3['Months to election'] = -df3['Month']
df3['Spread'] = abs(df3['Clinton (D)'] - df3['Trump (R)'])
```

```
#GRÁFICOS
plt.plot(df3['Months to election'],df3['Clinton (D)'], color='red', label = 'Democratic party')
plt.plot(df3['Months to election'],df3['Trump (R)'], color='blue', label = 'Republican party' )
plt.xlabel('Months to election')
plt.ylabel('Mean value of poll results for each month')
plt.title('Poll results (Election 2020)')
plt.legend(loc=1)
plt.show()

plt.plot(df3['Months to election'],df3['Spread'], color='grey',linestyle='dashed')
plt.fill_between(df3['Months to election'],df3['Spread'], 0,
                 facecolor="grey", # The fill color
                 color='grey',       # The outline color
                 alpha=0.2)          # Transparency of the fill
plt.xlabel('Months to election')
plt.ylabel('Spread value abs(Winner-Loser)')
plt.title('Spread between candidates (Election 2020)')
plt.show()

#FICHEROS CSV
from datetime import datetime
now = datetime.now() # current date and time
date_time = now.strftime("%m_%d_%Y_%H_%M_%S")
df1.to_csv(str(date_time + 'elecciones2016.csv'), index = False, encoding='utf-8')
df3.to_csv(str(date_time + 'AGREG_elecciones2016.csv'), index = False, encoding='utf-8')


#ELECCIONES 2020
page =
requests.get("https://www.realclearpolitics.com/epolls/2020/president/us/general_election_tr
ump_vs_biden-6247.html")
soup = BeautifulSoup(page.content, 'html.parser')
headings = [th.get_text() for th in soup.find("tr").find_all("th")]
datasets = []
for row in soup.find_all("tr")[1:]:
    dataset = list(td.get_text() for td in row.find_all("td"))
    datasets.append(dataset)

#CREATING THE DATAFRAME
df = pd.DataFrame(datasets, columns=headings)
df = df.drop_duplicates()
df1 = df.dropna()
df1['Month'] = df1['Date'].str.extract(r'\- (.*?)\/')
df1[['Sample Size','Sample Type']] = df.Sample.str.split(" ",expand=True,)
df1[['Winner','Spread Value']] = df.Spread.str.split(" +",expand=True,)

df2 = df1[['Biden (D)', 'Trump (R)', 'Month']]
df2 =df2.drop([0,1], axis=0)
```

```
type(df2['Month'])
df2['Biden (D)'] = df2['Biden (D)'].astype('int')
df2['Trump (R)'] = df2['Trump (R)'].astype('int')
df2['Month'] = df2['Month'].astype('category')

df3 = df2.groupby((df2.Month!=df2.Month.shift()).cumsum()).mean().reset_index(drop=False)
df3['Months to election'] = -df3['Month']
df3['Spread'] = abs(df3['Biden (D)'] - df3['Trump (R)'])

#GRÁFICOS
plt.plot(df3['Months to election'],df3['Biden (D)'], color='red', label = 'Democratic party')
plt.plot(df3['Months to election'],df3['Trump (R)'], color='blue', label = 'Republican party' )
plt.xlabel('Months to election')
plt.ylabel('Mean value of poll results for each month')
plt.title('Poll results (Election 2020)')
plt.legend(loc=1)
plt.show()

plt.plot(df3['Months to election'],df3['Spread'], color='grey',linestyle='dashed')
plt.fill_between(df3['Months to election'],df3['Spread'], 0,
                 facecolor="grey", # The fill color
                 color='grey',       # The outline color
                 alpha=0.2)          # Transparency of the fill
plt.xlabel('Months to election')
plt.ylabel('Spread value abs(Winner-Loser)')
plt.title('Spread between candidates (Election 2020)')
plt.show()

#FICHEROS CSV
from datetime import datetime
now = datetime.now() # current date and time
date_time = now.strftime("%m_%d_%Y_%H_%M_%S")
df1.to_csv(str(date_time + 'elecciones2020.csv'), index = False, encoding='utf-8')
df3.to_csv(str(date_time + 'AGREG_elecciones2020.csv'), index = False, encoding='utf-8')
```

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