Práctica 1[1]

1. Título del Dataset

El juego de azar no depende de la suerte a nivel económico y social.

2.

2.1 Subtítulo del Dataset

Los participantes no necesitan ni habilidad ni inteligencia. El ganar o el perder depende exclusivamente de la suerte.

2.2 Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo

En la última década ha resurgido la creación en España de empresas cuya actividad más destacadas es el juego en todas sus modalidades: apuestas deportivas, loterías, etc. Algunos de estos formatos tienen una gran acogida a nivel social.

Por este motivo surgió, dependiendo de la Secretaria de Estado de Hacienda, la dirección General de Ordenación del juego. La cual se hace eco de un juego responsable entre la población. A este mismo organismo se le encargó un estudio para analizar los factores de riesgo en los trastornos derivados del juego donde la manifestación más clara son las adicciones.

En dicho estudio se identifican los trastornos del juego de una muestra de afectados de diferentes regiones geográficas a nivel nacional, procedentes de 28 centros de salud y con diferentes niveles socioeconómicos donde un 92,4% de la muestra son hombres con una media de edad de 43 años.

Dicho estudio señala que los factores predictores de una mayor gravedad en el trastorno son:

- Hombres: Empleados, solteros y con algún problema de salud.

[1] Nota: en este documento se recogen las respuestas a todas las preguntas y puntos de la práctica. Sin embargo, en puntos como el último (10), es necesario adjuntar físicamente ficheros. Asimismo, para apreciar otros detalles, como el código o la imagen adjunta, es necesario hacer una ampliación sobre el fichero original. Es necesario, asimismo, tener en cuenta las incidencias y problemáticas con las que el equipo se ha encontrado a lo largo del desarrollo de la práctica. Por ello, para el acceso a toda la documentación, nos remitimos en su defecto a los diferentes componentes (archivos, imágenes, código, etc.) incorporados en el sitio del equipo en Github del que se proporciona link a continuación: https://github.com/jestebango/uoc

- Mujeres: Sin empleo, sin pareja, con cierta edad y beneficios sociales.

Respecto a las preferencias de juego con mayor prevalencia y en orden decreciente han sido:

- Hombres: máquinas tragaperras, loterías, quinielas, salas de juego, loterías instantáneas y apuestas por Internet.
- Mujeres: máquinas tragaperras, loterías y bingos.

A partir de las conclusiones de este estudio se puede deducir que, si las actividades de las empresas prestadoras de estos servicios crecen, y llevan aparejado un crecimiento de beneficios, también se incrementarán las posibilidades de que los trastornos analizados derivados del juego aumenten en un futuro y puedan visualizarse en los próximos estudios.

En este sentido hay que tener en cuenta el último estudio realizado por la Dirección General de Ordenación del juego, en el que se recoge el hecho de que, durante el año 2016, se declararon como ingresos por las empresas del sector más de 35.000 millones de euros. De los cuales 10.000 millones de euros pertenecen al negocio online y 25.000 millones de euros al negocio tradicional. Ello permite poder realizar una comparativa entre los ingresos de este sector, y otros sectores de relieve como la banca o los seguros, implicando un peso destacado en el cómputo del PIB, de mucho mayor relieve que otros como el sector primario o el cultural.

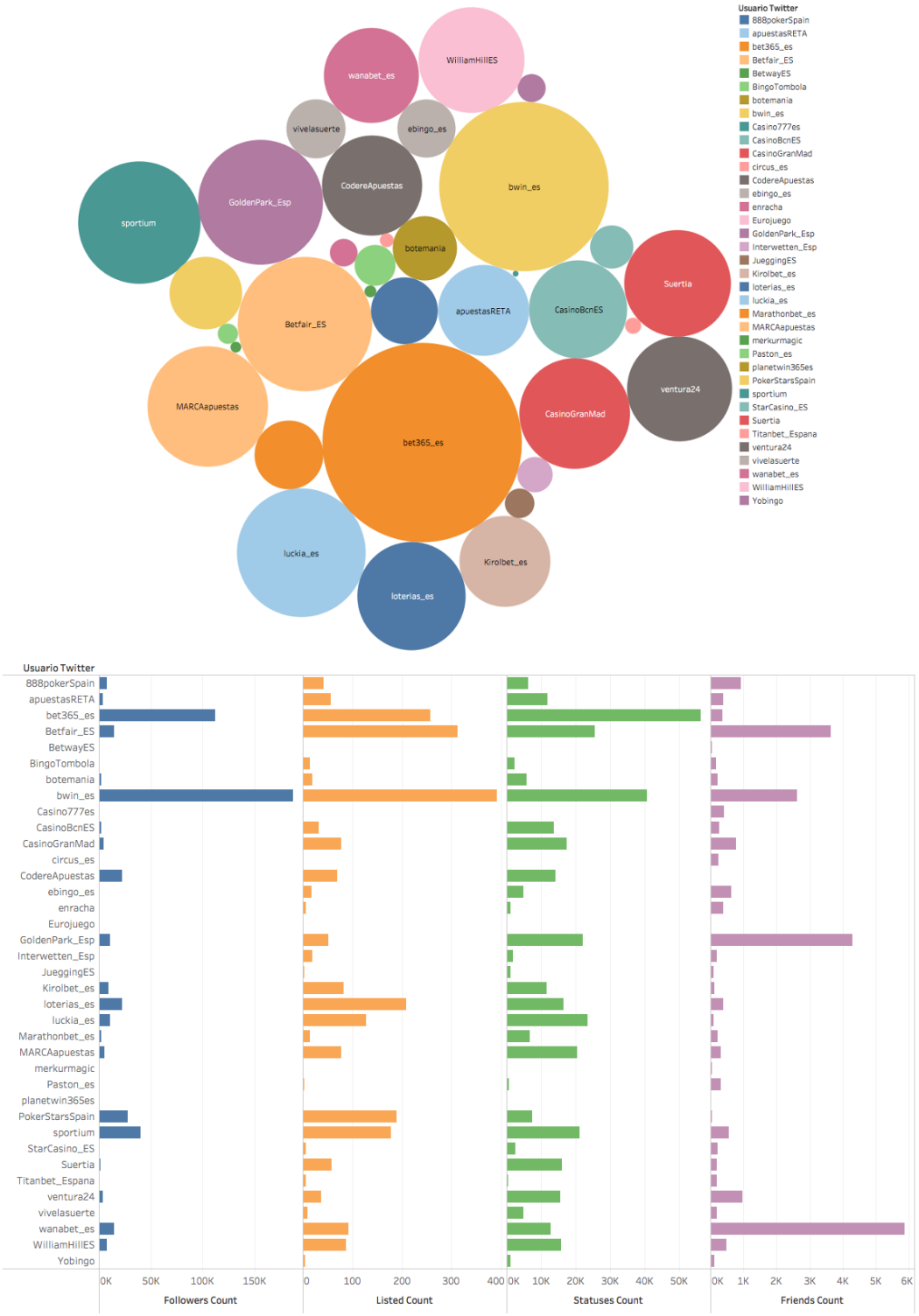
Si se tienen en cuenta los ingresos netos, queda claro que el relacionado con el negocio online mejora de forma fulminante, aunque aún es superado notablemente por el negocio tradicional. Sin lugar a dudas la parte más importante de este negocio tradicional corresponde a las loterías. Parte de estos ingresos corresponden a los 3 millones de ludópatas esporádicos.

El aumento del negocio online se centra principalmente en la rapidez con que un jugador obtiene un premio de forma casi inmediata en lugar de una espera a cambio de un premio superior. El jugador asociado a este estilo de juego corresponde a varones con dos intervalos de edad de 23 a 44 años y un segundo intervalo mayores de 56 cuyo estatus laboral es fijo. Aunque según se produce un ascenso en la ludopatía debemos incluir también a personas desempleadas.

Pero, asimismo, no se debe perder de vista el potencial mercado que se abre a las empresas de juego a través de los nuevos canales electrónicos, y el alcance a un mayor grupo de personas a través de las redes sociales. La Sociedad de la Información trae consigo un abanico de posibilidades para las empresas de juego que, desde el márketing, hasta nuevos modelos de negocio, tiene un potencial impacto en un mayor número de personas.

3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente.

Empresas de juego en España: actividad en Twitter



4. Contexto. ¿Cuál es la materia del conjunto de datos?

La materia de este conjunto no es otro que las empresas de juego de azar, con/sin actividad en el negocio on-line, cada vez es mayor su presencia y actividad en las redes sociales como Twitter. Las empresas orientan su actividad a un público objetivo específico. En este caso, las empresas de juego en España tienen, cada vez más, un terreno abonado en las redes sociales, dado que es allí donde su potencial público objetivo se encuentra también hoy día. Ello es más relevante, si cabe, dado que, con la globalización y el uso de internet, el juego online es una creciente fuente de ingresos de estas empresas, con un volumen sumamente importante. Es por ello, por lo que hemos considerado fundamental explorar la actividad de estas empresas en las redes sociales y más concretamente en Twitter.

5. Contenido

5.1 ¿Qué campos incluye?

Tabla de los campos de las variables

VARIABLE	CARACTERÍSTICA	EJEMPLO
Nombre	Nombre de la empresa (puede tener múltiples delegaciones / franquicias) con las que comparte las licencias	Codere Online, SAU
Licencias	Licencias singulares obtenidas para cada tipología. Puede estar vacío (‘Sin licencia vigente’), lo que indica que está extinguida	Ruleta, Black Jack,...
Dominios	Sitios web donde opera la empresa ofreciendo el servicio para el que tiene licencia. Puede no tener ningún dominio o varios (0,n)	Www.codere.es,...
Usuario Twitter	El usuario @...	CodereApuestas
statusesCount	Actividad del usuario, incluyendo retweets	14213
FollowersCount	Seguidores del usuario	22267
friendsCount	Usuarios a los que sigue	0
Created	Fecha de creación del usuario	2012-11-19 13:00:41

ListedCount	Listas públicas del que el usuario es parte	69

Los campos que se incluye en dataset son:
-Nombre. Nombre de las empresas.
-Licencias. Licencias obtenidas sobre el tipo de juego de azar que gestiona
-Dominios. En caso de tener licencia para servicio web, sería el dominio sobre el cual opera.
-Usuario Twitter. Usuarios de la red social Twitter
-Statusses Count. Actividad del usuario, incluyendo retweets
-Followers Count. Seguidores del usuario

5.2 ¿Cuál es el periodo de tiempo de lo tiempo de los datos y cómo se ha recogido?

Las variables Nombre, Licencias y Dominios sale directamente de la Dirección General de Ordenación del Juego (DGOJ). Esta institución es un órgano de la Secretaria de Estado de Hacienda que regula, licencia, vigila, controla e incluso puede llegar a multar todo lo relacionado con el juego de azar, independientemente de la plataforma en la cual se desarrolle.
Las variables Usuario Twitter, Statusses Count, Followers Count vienen de la red social Twitter. En la que se mide la presencia de las empresas de juego de azar mediante las variables anteriormente descritas.

La temporalidad para estos datos corresponde al cierre del año (2016).
El scraping se ha realizado directamente desde la web de la Dirección General de Ordenación del Juego (DGOJ).
En el caso de la búsqueda en Twitter, no se impone limitación de tiempo, aunque se recoge la fecha de inicio o creación del usuario, ya que lo que nos interesa es el volumen acumulado de dicha actividad a fecha de hoy.

Ahora detallaremos la forma de extracción de los datos, comenzando por el scraping de la Dirección General de Ordenación del Juego. En la página web http://www.ordenacionjuego.es/es/operadores/buscar aparece la información de forma estructurada, cada tipo de juego de azar requiere una licencia con este hecho se realiza la consulta hasta obtener 53 empresas dedicadas al juego de azar con su correspondiente licencia y plataforma on-line en caso de existir.

De este grupo de empresas extraemos la información de Twitter anteriormente especificada. Para la recogida de los datos hemos creado en primer lugar una API (Juego) a través del sitio para desarrolladores de Twitter, como instrumento a utilizar en la exploración y extracción dentro de Twitter.

6. Agradecimientos. ¿Quién es el propietario del conjunto de datos?

El propietario del Origen de datos es la Dirección General de Ordenación del juego (esta cita tiene un carácter obligatorio por las condiciones generales de uso de la propia página) que recordemos que depende del Ministerio de Hacienda y teniendo en cuenta la extracción realizada que se ha llevado a cabo a través de Twitter, todas las cuentas son abiertas, y no tienen restricciones a la recopilación de esta información, es decir, es público, en el caso de la extracción de Twitter no existe propietario de los datos. No obstante, el hecho de analizar las empresas, nos lleva a agradecer desde aquí, indirectamente, a dichas empresas la posibilidad de examinar su usuario y actividad en Twitter.

Hay investigaciones realizadas anteriormente que dependen de la Federación Española de jugadores de Azar Rehabilitados Fejar.org/ludopatía/

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder a la comunidad?

Este conjunto de dato es interesante ya que debido a la inclusión del negocio de azar de forma digital. Se produce unos incrementos considerados en el número de ludópatas, debido principalmente a la facilidad de acceso a este tipo de juego y a la obtención inmediata del premio, aunque la cuantía es aún muy inferior a la de otros juegos de azar desde otras plataformas tradicionales. Recordemos que, para el jugador con trastornos ludópatas, el premio no tiene tanta importancia como el propio hecho de jugar.

Desde el punto de vista económico estas nuevas plataformas online para juegos de azar están generando ingresos muy atractivos para las empresas de este sector, las cuales, por supuesto, no dejan de invertir en esta nueva modalidad que les produce tantos beneficios.

Una de las preguntas que le resultaría interesante a la comunidad podría ser
¿Dónde se encuentra el equilibrio para la sociedad entre los beneficios y el aumento de usuarios con posibles trastornos ludópatas?
¿Qué patrones siguen usuarios que finalmente sufren trastornos ludópatas?
¿Qué valor debería tener el premio para reducir el número de jugadas?
¿Qué coste social sanitario tendrá el aumento de miembros de la comunidad con trastornos ludópatas?
¿Las plataformas online propagan la deslocalización de la problemática, su reducción o su aumento?

Aunque estas preguntas resulten difíciles de responder, no olvidemos que la transformación digital de las empresas permite tener disponible multitud de datos tanto de tipo estructurados como no estructurados, por tanto, el estudio de patrones de

comportamiento de los usuarios es viable y obtener respuestas de forma analítica que nos de conocimiento es posible.

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado.

La licencia considerada es Released Under CC BY-NC-SA 4.0 License.

Detallamos el motivo por el cual ha sido elegido esta licencia para ello debemos tener en cuenta un pequeño análisis de las licencias indicadas que luego nos servirá como base a la elección del tipo de licencia.
Hay que tener en cuenta que todas las licencias CC(Creative Commons) permiten derechos de autor sin restricciones pero no puede existir ningún tipo de intencionalidad comercial ni realización de modificaciones.
BY permite la utilización de la información tanto para copiar, transmitir, utilizar tanto de manera visual como origen de nuevos trabajos solamente si se da constancia del autor de la información.
SA Permite que sea origen de nuevos trabajos, pero estos deben tener la misma licencia que el trabajo original.
NC Permite copiar, transmitir, utilizar y realizar trabajos que deriven de este, pero con una finalidad no comercial.
ND Permite copiar, transmitir y utilizar, pero no realizar trabajos que deriven de este.
4.0 Corresponde a la última versión de CC, estas son utilizadas en gran medida por las jurisdicciones.
Teniendo en cuenta las licencias de cada una de las páginas web que hemos accedido se considera como las más apropiada la licencia Released Under CC BY-NC-SA 4.0 License.
En la página de la Dirección General de Ordenación del Juego se establece que las condiciones para la reutilización de la información que ofrece deben ser la siguiente:
1.No se debe modificar el sentido de la información
2.Debe citarse la fuente de los documentos objetos de la reutilización.
3.Se debe indicar la fecha en la que se han actualizado por última vez los datos.
4.No se debe de ninguna manera indicar que la Dirección General de la Ordenación del Juego tiene algo que ver como beneficio propio la reutilización de la información.
5.Hay que indicar en todo momento las condiciones de reutilización de los datos.

En el caso de Twitter no ha existido restricción, ya que las cuentas no están protegidas y la actividad de los usuarios en este sentido es pública. Pero debemos tener en cuenta que la información extraída por esta página no debe tener ningún fin abusivo, de engaño o incluso que causa un mal físico o moral a ninguna persona.

9. Código. Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python.

Para el scraping de la página web de la Dirección General de Ordenación del juego se utiliza Python. El script utilizado es el siguiente

```
import urllib
import ssl
import pandas as pd
from bs4 import BeautifulSoup

# Función para devolver una cadena única separando los elementos por comas a partir de los elementos de una lista
def lista_cadena(lista):
    enCadena=""

    if (len(lista)==1):
        enCadena=enCadena+lista[0]

    else:
        for elto in lista:

            enCadena=enCadena+elto
            if (elto!=lic[len(lic)-1]):
                enCadena=enCadena+','

    return enCadena


# Con archivos locales guardados con el navegador
file1 = "file:///F:\PEC1WEBSCRAPING\Buscador de Operadores _ Dirección General de Ordenación del Juego.html"
file2 = "file:///F:\PEC1WEBSCRAPING\Buscador de Operadores _ Dirección General de Ordenación del Juego2.html"
file3 = "file:///F:\PEC1WEBSCRAPING\Buscador de Operadores _ Dirección General de Ordenación del Juego3.html"

# Creación de contexto SSL
ctx = ssl.create_default_context()
ctx.check_hostname = False
ctx.verify_mode = ssl.CERT_NONE

# URL remotas
# CONCURSOS
url1 = "https://www.ordenacionjuego.es/es/operadores/buscar?field_got_tid=All&field_gat_tid=All&field_gct_tid=999993&field_dominio="
# APUESTAS
url2 = "https://www.ordenacionjuego.es/es/operadores/buscar?field_got_tid=All&field_gat_tid=999992&field_gct_tid=All&field_dominio="
# OTROS JUEGOS
url3 = "https://www.ordenacionjuego.es/es/operadores/buscar?field_got_tid=999991&field_gat_tid=All&field_gct_tid=All&field_dominio="
```

```
urls = (url1,url2,url3)
#urls = {file1,file2,file3}


# CON URL REMOTA

# Dos alternativas:
#   - CASO 1: No almacenar el tipo de licencia singular (concursos, apuestas u otros)
#   - CASO 2: Almacenar el tipo de licencia
# NOTA: Se ha implementado sólo el CASO 1, para el caso 2 habría que crear un campo adicional para almacenar el tipo de
# licencia; en cada iteración un tipo diferente por lo que habría empresas repetidas con mismo nombre, url pero distintas
# licencias.
# Almacenar el tipo de licencia podría ser útil por ejemplo si se desea realizar un estudio más específico por categorías

# CASO 1 (sólo es necesario un diccionario para las tres iteraciones, el diccionario no admite duplicados en el índice)
# ----------------------------------------------------------------------------------------------------------------------
# datURL almacena los nombres y urls

datURL = dict()

for url in urls:

    pagina = urllib.request.urlopen(url,context=ctx)
    soup = BeautifulSoup(pagina,"lxml")
    contenidos = soup.find("div",class_="view-content").find_all("a")

    for empresa in contenidos:
        nombre = empresa.text.strip()
        link = "https://www.ordenacionjuego.es" + empresa.get("href").strip()
        datURL[nombre]=link
        print(nombre,link)

#print(datURL)

# Web scraping de empresas individuales
# Con los resultados almacenados en el diccionario datURL, se realiza scraping sobre cada página individual de cada empresa

datasetFinal = []

#url1 = "file:///F:\PEC1WEBSCRAPING\Bluesblock, SA _ Dirección General de Ordenación del Juego.html"

#url1="https://www.ordenacionjuego.es/es/op-antena3juegos"


#if url1 != '':
for url in datURL.values():

    #pagina = urllib.request.urlopen(url)
    pagina = urllib.request.urlopen(url, context=ctx)
    soup = BeautifulSoup(pagina,"lxml")
    contenidos = soup.find(attrs={'id':'operatorContent'})
```

```
    # Campo NOMBRE
    nombre = contenidos.find(attrs={'id':'operatorTitle'}).text.strip() # Strip para eliminar posibles espacios

    # Campo LICENCIAS(SINGULARES)
    licencias = contenidos.find(attrs={'id':'operatorSingular'}).find_all('li')
    lic=[]

    for licencia in licencias:
        lic.append(licencia.text)

    # Se procesa para que se almacenen con la forma Concursos,Lotería,... en una cadena de texto única
    # Se podrían haber almacenado como variables binarias (Concursos=Sí/No, Loterías=Sí/No,...)

    campoLicencias = lista_cadena(lic)

    # Campo DOMINIOS
    dominios = contenidos.find(attrs={'id':'opetatorBody'}).find_all('li')
    #print(dominios)
    dom=[]
    #if not dominios:
    #    dom = 'Sin_dominio'

    #else:
    #    for dominio in dominios:
    #            if dominio.find('a'):
    #            dom.append(dominio.find('a').text)

    if dominios:
        for dominio in dominios:
        if dominio.find('a'):
                dom.append(dominio.find('a').text)

    # Se procesan los dominios recuperados para dejarlos en una única cadena

    campoDominios = lista_cadena(dom)

    # Se añaden al total de observaciones

    datasetFinal.append((nombre,campoLicencias,campoDominios))

# Peculiaridades de los datos almacenados
#   1. Licencias: puede haber operadores que al buscar en el formulario aparezcan como poseedores de licencias, pero que
#      al acceder a su URL específica no muestren ninguna licencia (ej: antena3juegos). Si se consulta la tabla se puede ver
#      que tiene la licencia extinguida. Se especifica como 'Sin licencia vigente'
#   2. Dominios: en el caso de operadores que no tengan dominio, se especifica como 'Sin dominio'

# ALMACENAMIENTO EN DATAFRAME
#df['date'] = pd.to_datetime(df['date']) # yyyy-mm-dd

df = pd.DataFrame(datasetFinal, columns=['Nombre','Licencias','Dominios'])
#df.head()
#df.tail()
print(df)
df.to_csv('F:\\operadoresDGOJ.csv',index=False, encoding='utf-8')
#df = pd.read_csv('operadores.csv',encoding='utf-8')
```

```
#for valores in zip(datURL.keys(),datURL.values()):
#    print(valores)
```

En el caso de la exploración de Twitter se aprovecha la potencia de las funciones ya existentes en las librerías:

- library(twitteR)
- library(ROAuth)

```
## Twitter API  (Juego-DataScience, para la práctica)
#Sitio de la API https://apps.twitter.com/app/14410452/show (con login twitter)


#Carga de librerías
library(twitteR)
library(ROAuth)

#Credenciales

clave_API <- "hashclaveAPI" # From dev.twitter.com
clave_API_Secreta <- "hashclavesecretaAPI" # De dev.twitter.com
token <- "hashtoken" # De dev.twitter.com
token_secreto <- "hashtokensecreto" # De dev.twitter.com

# Conexión con Twitter
setup_twitter_oauth (clave_API, clave_API_Secreta, token, token_secreto)

#Creamos lista completa para toda las empresas (tengan o no usuario de twitter)
listadeusuarios <- c('Bluesblock, SA','CodereApuestas','Concursos_multiplataformas', 'CasinoGranMad','Eurojuego','JS2015Games', 'loterias_es', 'Suertia',
'MARCAapuestas', '888pokerSpain', 'JueggingES', 'Starvegas SA', 'Betfair_ES', 'BetwayES', 'CasinoBcnES', 'sportium', 'ijuego',
'ebingo_es','apuestasRETA','bwin_es','Paston_es','EurobetInternational','circus_es','GoldenPark_Esp','GtechSpain','bet365_es','Interwetten_Esp','Titanbet_
Espana','KambiSpain','luckia_es','Marathonbet_es','merkurmagic','Paf
consulting','carcaj','StarCasino_ES','PokerStarsSpain','wanabet_es','planetwin365es','Kirolbet_es','WilliamHillES','Yobingo','Casino777es','botemania','Ci
gagameonline','HillsideEspa_aLeisure','NetEntGamming','PrimaNetworks','PtEntretenimientoonline','enracha','BingoTombola','ventura24','vivelasuerte')

#Lanzamos búsqueda datos sobre usuarios específicos que si tienen cuenta.
#Almacenamos los resultados en una lista
usodetwitter <- lookupUsers(c('CodereApuestas','CasinoGranMad','Eurojuego','loterias_es', 'Suertia', 'MARCAapuestas', '888pokerSpain', 'JueggingES',
'Betfair_ES', 'BetwayES', 'CasinoBcnES',
'sportium','ebingo_es','apuestasRETA','bwin_es','Paston_es','circus_es','GoldenPark_Esp','bet365_es','Interwetten_Esp','Titanbet_Espana','luckia_es','Mara
thonbet_es','merkurmagic','StarCasino_ES','PokerStarsSpain','wanabet_es','planetwin365es','Kirolbet_es','WilliamHillES','Yobingo','Casino777es','botemania
','enracha','BingoTombola','ventura24','vivelasuerte'), includeNA=TRUE)
```

Ilustración 1.- Contenido del script en R para la extracción de Twitter.

10. Dataset: Dataset en formato CSV

https://github.com/jestebango/uoc/blob/master/data/operadoresDGOJTwitter.csv

Componentes del grupo:

María Sonia Rodríguez Cepedano

Carlos E. Jiménez

José Luis Esteban González