Tipologia y ciclo de vida de los datos

Alumna: Mabel Arroyo Adeba

PRA1 Web Scraping

Fecha: Noviembre 2020

Indice

1. Contexto

La ludopatía es una adicción que sufre un 1% de la población en España. En los últimos años, además, se ha detectado que esta enfermedad mental a aumentado entre los jóvenes de 18 a 25 años.

Basándonos en el cálculo combinatorio, en el sorteo de la Primitiva, la probabilidad de acertar los 6 números y el reintegro es de 139.838.160. Tal vez este dato no sea suficiente para que las personas entiendan que este sorteo no es más que un juego.

En la página de loterías y apuestas del estado, es posible recoger los números de los sorteos que se han celebrado desde el 20/06/1991.

Mediante técnicas de Web Scraping se capturarán estos datos, para crear un dataset cuya descripción veremos en los puntos siguientes.

Este dataset, además servirá para hacer estadísticas sobre frecuencia de números aparecidos, importe de premios según categoría, etc.

2. Título del dataset

El nombre del dataset será: **Historico_Primitiva.csv.** Este nombre es bastante descriptivo ya que nos informa del carácter acumulativo de los datos durante un

largo periodo de tiempo. Con la palabra "Primitiva" se entiende que se refiere al sorteo de la lotería Primitiva.

3. Descripción del dataset

El dataset Historico_Primitiva contendrá tantas líneas como sorteos se hayan celebrado entre las fechas que se indiquen. En cada línea aparecerá la fecha del sorteo, los números de las 6 bolas de la combinación en orden ascendente y la del número complementario. Además se recogerá la recaudación, el bote y el número de acertantes y premios según categoría. Los datos referidos a la población donde ha sido agraciado y los datos correspondientes al número del Joker, no son capturados. Se deja para una fase posterior del proyecto.

4. Representación gráfica

5. Contenido

Los datos se recogen de la página de loterías y apuestas del estado. En un principio se pensó hacer la captura desde:

https://www.loteriasyapuestas.es/es/resultados/primitiva

pero debido a problemas técnicos, se optó finalmente por utilizar una página un poco más sencilla que contenía la misma información:

https://www.loteriasyapuestas.es/es/buscador

, incluyendo parámetros en la petición, tal y como se puede ver en el script.

Se capturan sorteos celebrados desde el 20/06/1991 hasta 31/10/2020. Las fechas pueden parametrizarse fácilmente cambiando el valor de las variables y dando la posibilidad de capturar otro rango de datos. Durante la ejecución del script, el usuario puede ver el progreso ya que se va mostrando la fecha de cada sorteo procesado.
El fichero de salida será un fichero plano .csv. Como separador se ha escogido ";" para no interferir en el símbolo decimal "," de los datos numéricos.

Los campos del fichero Historico_Primitiva.csv son:

Nombre Campo	Tipo	Descripción	Ejemplo
INDICE	Numérico	Número de línea	1
FEC_LARGA	Texto	Fecha descriptiva	Resultado del sábado 31 de octubre de 2020
FEC_SORTEO	Fecha	Fecha del sorteo en formato AAAMMDD	20201031
B1	Numérico	Valor Bola 1	2
B2	Numérico	Valor Bola 2	13
B3	Numérico	Valor Bola 3	21
B4	Numérico	Valor Bola 4	32
B5	Numérico	Valor Bola 5	41
B6	Numérico	Valor Bola 6	43
C	Numérico	Complementario	40

R	Numérico	Reintegro	9
RECAUDACION	Numérico	Importe recaudado	11.420.804,00 €
BOTE	Numérico	Importe del bote	7.500.000,00 €
ACE	Numérico	Acertantes de la categoría especial (6 aciertos + R)	0
PCE	Numérico	Premio de la categoría especial	0,00 €
AC1	Numérico	Acertantes de la categoría 1ª (6 aciertos)	1
PC1	Numérico	Premio categoría 1ª	1.450.282,32 €
AC2	Numérico	Acertantes de la categoría 2ª (5 aciertos + C)	9
PC2	Numérico	Premio categoría 2ª	24.171,37 €
AC3	Numérico	Acertantes de la categoría 3ª (5 aciertos)	206
PC3	Numérico	Premio categoría 3ª	2.288,07 €
AC4	Numérico	Acertantes de la categoría 4ª (4 aciertos)	10.098
PC4	Numérico	Premio categoría 4ª	75,40 €
AC5	Numérico	Acertantes de la categoría 5ª (3 aciertos)	189.207
PC5	Numérico	Premio categoría 5ª	8,00 €
ACR	Numérico	Acertantes del Reintegro	1.100.069
PCR	Numérico	Premio Reintegro	1,00 €

6. Agradecimientos

- El propietario de los datos es la Sociedad Estatal Loterías y Apuestas del Estado, S.M.E., S.A. (SELAE), domiciliada en la C/ Poeta Joan Maragall 53, 28020 Madrid. Su NIF es A86171964, inscrita en el Registro Mercantil de Madrid al tomo 28078, folio 202, sección 8ª, hoja M-505970, inscripción 1ª

- Analicé el fichero robots.txt y la página que yo necesitaba, estaba permitida.

```
User-Agent: *
###
Disallow: /portal/site/
Disallow: /es/paginas-informativas/trabaja-con-nosotros*
Disallow: /*.json$
Disallow: /*.formatoRSS$
Disallow: /*.corporativa
Disallow: /*.info
Disallow: /*.info2
Disallow: /index.php/
Disallow: /mod.detalle/
Disallow: /mod.documentos/
Disallow: /mod.faqs/
Disallow: /mod.geo/
Disallow: /mod.global/
Disallow: /mod.imagenes/
Disallow: /mod.indice/
Disallow: /mod.multimedia/
Disallow: /mod.noticias/
Disallow: /mod.pags/
Disallow: /mod.resultados/
Disallow: /mod.sorteos/
Disallow: /mod.widgets/
Disallow: /*/red-comercial
Disallow: /*/botes
Disallow: /*/escrutinios
Disallow: /f/loterias/documentos/mig/
Disallow: /vgn-ext-templating/
Disallow: /*/*/sorteos/19
Disallow: /*/*/sorteos/200
Allow: /*/*/sorteos/2019
Disallow:/*/*/sorteos/201
Disallow: /*/buscador
# Otras con parametros problemáticos
Disallow: /*?*%
Disallow: /*?./lang
###

User-agent: MJ12bot
Disallow: /

Sitemap: https://www.loteriasyapuestas.es/sitemap.xml
Sitemap: https://juegos.loteriasyapuestas.es/sitemap.xml


#
# Version robots.txt:13/06/2019
```

- Solicité permiso de acceso el día 28/10/2020 enviando un correo a la dirección cau@selae.es. Mientras esperaba contestación, continué con el proyecto. El día 04/11/2020 recibí respuesta con la autorización para la extracción de los datos.

7. Inspiración

El objetivo de la recopilación de los datos es la de hacer una aplicación en la que el usuario pueda introducir una combinación de 6 números y comprobar si en estos últimos años, esa combinación ha sido ganadora y con qué premio. Si los números elegidos han aparecido en algún momento, la probabilidad de que esa combinación vuelva a salir es prácticamente nula por lo que es absurdo seguir apostando por ella. Si, por el contrario, la combinación nunca se ha dado en este tiempo, es una manera de hacer entender que, si en tantos años no hemos acertado, mejor hubiésemos ido guardando el dinero de la apuesta en una hucha, y hoy en día, el dinero acumulado superaría el posible premio.
En conclusión: lo que se pretende es dar a las personas una visión con datos, de la realidad de los juegos de azar. Hacer entender que no hay mucho más allá de la ilusión de que algún día te toque la Primitiva. Por eso, lo que hay que ver es que no es más que un juego para entretenerse. Y tal vez con ello, alguien que lo necesita, vea que jugar de forma compulsiva es absurdo.

Además de estas razones de carácter subjetivo y objetivo social, este dataset puede responder a preguntas estadísticas tales como:
¿Cuántas veces ha salido tal número?
¿Cuál ha sido el premio más alto conseguido?
¿Qué probabilidad condicional hay de que salga una combinación que elijamos?
Etc.

8. Licencia

Antes de elegir qué licencia es la más adecuada para el dataset que se construye, repasemos el significado de cada una de ellas tal y como se describe en https://creativecommons.org/licenses y https://opendatacommons.org/licenses/:

- **Released Under CC0: Public Domain License:** Es la opción "sin derechos de autor reservados" es el conjunto de herramientas de Creativa Commons – significa efectivamente renunciar a todos los derechos de autor y derechos similares que posee sobre una obra y dedicar esos derechos al dominio público.

- **Released Under CC BY-NC-SA 4.0 License:** Esta licencia permite a otros remezclar, adaptar y construir sobre su trabajo de forma no comercial, siempre que le otorguen crédito y se licencien sus nuevas creaciones bajo los mismos términos.

- **Released Under CC BY-SA 4.0 License:** Esta licencia permite a otros remezclar, adaptar y construir sobre su trabajo incluso con fines comerciales, siempre que le otorguen crédito y se licencien sus nuevas creaciones bajo los mismos términos.

- **Database released under Open Database License, individual contents under Database Contents License:** Son dos licencias. La primera se refiere a la base de datos en conjunto y permite a los usuarios compartir, modificar o usar la base de datos manteniendo esa misma licencia en la redistribución e indicando atribución. La segunda se refiere al contenido. Es una licencia simple en caso de que el propietario posea los derechos de licencia de los contenidos.

- **Other(specified above):** Otra licencia no especificada en los apartados anteriores.

- **Unknow license:** Licencia desconocida

Una vez entendidas, quería optar por asignar a mi dataset la licencia **CC0: Public Domain License.** El motivo de esta elección es que pienso que si a alguien le puede ser útil mi trabajo, que lo aproveche. Me parece bien que sea de dominio público.
Sin embargo al dar de alta el dataset en Zenodo, no me dio esta opción así que finalmente quedó grabado bajo la licencia **Creative Commons Attribution 4.0 International (CC BY-SA 4.0),** que como se ha indicado en el resumen anterior, permite a otros remezclar, adaptar y construir sobre su trabajo incluso con fines comerciales, siempre que le otorguen crédito y se licencien sus nuevas creaciones bajo los mismos términos.

9. Código

La codificación del proyecto se ha realizado con el lenguaje de programación Python utilizando la plataforma Jupiter Notebooks. Para la generación del dataset se ha utilizado técnicas de web scraping utilizando principalmente la librería BeautifulSoup. Hay que anotar que durante el desarrollo del script se han detectado falta de datos en algunos campos para algunos registros. Esto se ha solucionado rellenando con valor "None". Antes de explotar el dataset se ha de pensar cómo solucionar esta carencia en la información. El fichero con el script es Historico_Primitiva.ipynb y está disponible junto a este documento en el repositorio:
https://github.com/mabelarroyoadeba/Historico_Primitiva

10. Obtención del DOI

Después de darme de alta en Zenodo y una vez creado el dataset, lo subí al repositorio.
El DOI asignado ha sido: **10.5281/zenodo.4244093** y está accesible en el enlace:
https://zenodo.org/record/4244093#.X6KFKdt7njA

11. Firma de participantes:

Contribuciones	Firma
Investigación previa	MAA
Redacción de las respuestas	MAA
Desarrollo código	MAA