Números más repetidos en los sorteos de Euromillones.

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS
Práctica 1

DIEGO GONZÁLEZ MARTÍNEZ

ÍNDICE

Índice

1- Contexto

Como preámbulo, explicaré de forma breve en que consiste este juego de azar- Los juegos de azar y como referencia desde el 13 de febrero de 2004, tenemos el juego del Euromillones, tal y como indica su propio nombre se trata de un sorteo en el que participan varios países del entorno Europeo, al ser uno de los juegos que mayores premios reparte debido a la dificultad de obtener una combinación de 5 números entre el 1 y el 50 más dos estrellas del 1 al 12, la probabilidad de acertar esta combinación es de 1 entre casi 140 millones, por tanto, se trata de uno de los juegos de azar más difíciles de acertar y en el que se acumulan los botes más grandes hasta un límite de 190 millones de euros. Aunque inicialmente los sorteos se celebraban sólo los viernes, desde el 10 de mayo de 2011 se celebran dos sorteos semanales, martes y viernes, por tanto, desde esa fecha dispondremos de dos sorteos semanales.

La curiosidad de ver que números que más veces han resultado partícipes de la combinación ganadora, me ha motivado para realizar esta práctica sobre estos datos, buceando en internet, existe muchas webs que muestran la información de los sorteos y las combinaciones ganadoras, entre ellas he seleccionado la web *"https://www.combinacionganadora.com/"* para intentar extraer la información a través del web scraping.

Esta web dispone de información de varios juegos de azar, por tanto, en primer lugar debemos seleccionar el juego de Euromillones, tras acceder a esta sección, vemos que nos muestra el resultado del último sorteo y los botones de comandos, entre otros vemos que puede mostrarnos los sorteos anteriores, al clicar sobre este botón, vemos que selecciona la página *"https://www.combinacionganadora.com/euromillones/2019/10/25"*, por tanto, tenemos que tener en cuenta que la variación entre sorteos será la fecha de cada sorteo pero en un formato de año/mes/día. Analizamos la información que nos muestra el navegador tras inspeccionar la web, vemos como identifica la información que tratamos de conseguir[1]

Para conseguir nuestro dataset, crearemos un bucle while que en función de la fecha y si es día de sorteo, consultará la web y capturará los datos que necesitamos, posteriormente los guardaremos en un DataFrame para finalmente generar un fichero tipo texto ".csv", mostraremos las repeticiones de cada número, las estrellas y un resumen de la combinación de los números/estrellas más repetidas.

2- Título del Dataset

Como título del dataset, he determinado que será "Números más repetidos en los sorteos de Euromillones", ya que al final se trata de ver la combinación de números que más veces han salido en los sorteos de Eurmillones.

[1] durante las pruebas comprobamos que el webmaster cambió el nombre de la clase que identificaba los resultados de los sorteos, ver imagen.

3- Descripción del Dataset

Necesitamos capturar los datos relativos a los sorteos, pero sólo seleccionaremos los cinco números de cada sorteo y los dos números de las estrellas, la fecha del sorteo la calculamos a partir de una fecha inicial y vamos descontado días ya que formará parte del sistema de rastreo, posteriormente comprobamos si esa fecha es un viernes o un martes, teniendo en cuenta que los sorteos de los martes se iniciaron el 10 de mayo de 2011, y añadimos esa fecha en formato (yyyy/mm/dd) a la url base de la página web que vamos a rastrear, *"https://www.combinacionganadora.com/euromillones/"* .

He limitado la captura al 14 de Mayo 2004 ya que es la fecha del primer sorteo de este juego de azar que esta web tiene registrada, aunque la fecha del primer sorteo es del 13 de Febrero de 2004, para ver la historia de este juego he consultado la web "https://www.euro-millions.com/es/historia" donde podemos ver todos los detalles.

El dataset se compone de nueve atributos y un total de 1249 registros, tal y como sigue

- Fecha del sorteo en formato (dd/mm/yyyy). Variable tipo Date, realizamos una conversión ya que la url requiere la fecha en formato (yyyy/mm/dd)
- dia_sorteo, variable tipo string que nos indica si el sorteo es en viernes o martes

- Los atributos de numero1, numero2, numero3, numero4 y numero5 corresponde cada uno de los números aparecidos en el sorteo correspondiente, estas variables son de tipo integer.

- Los atributos estrella1 y estrella 2 corresponde a cada una de las estrellas aparecidas en el sorteo correspondiente, estas variables son de tipo integer.

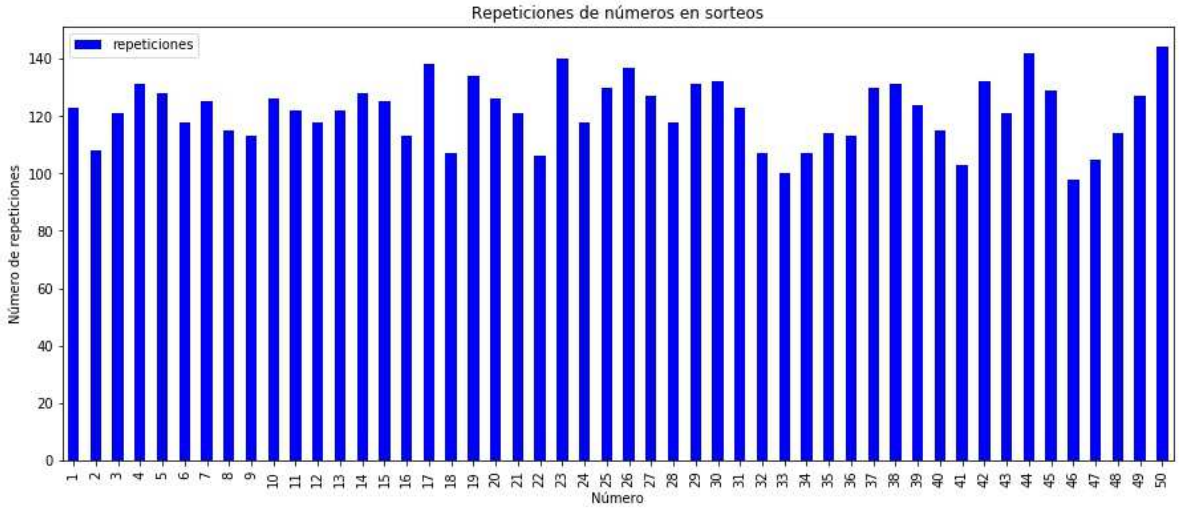
Para comprobar que números son los más repetidos entre todos los sorteos y teniendo en cuenta que no existe una posición determinada para que aparezcan en una variable u otra, sumamos las apariciones de cada número del sorteo y de cada estrella en todas las variables de forma que obtenemos un total de repeticiones de cada número indistintamente de la posición que ocupara en cada sorteo.

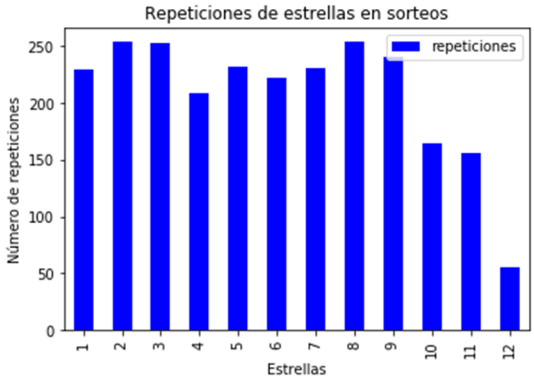
Realizamos dos gráficos por las repeticiones de cada número y observamos que números son los más repetidos de todos los sorteos, hacemos lo mismo con las estrellas. Finalmente ordenamos los dos dataset obtenidos por el número de repeticiones y obtenemos los cinco números más repetidos y las dos estrellas más repetidas en todos los sorteos.

4- Representación gráfica

Como ya he mencionado en el punto anterior, a partir de los datos de todos los sorteos, sumamos las veces que cada uno de los 50 números (del 1 al 50) indistintamente de la posición que ocupen, es obvio que el número 1 sólo puede aparecer en la variable 'numero1' y que el número 50 sólo puede aparecer en la variable 'numero5', lo mismo pasa con las estrellas, la estrella 1 solo puede aparecer en la variable 'estrella1' y la 12 sólo puede aparecer en la variable 'estrella2', ya que la secuencia de números y estrellas de cada sorteo aparece ordenadas, aunque no es relevante para nuestro proceso de captura y cálculo ya que no lo hemos tenido en cuenta.

Una vez disponemos las repeticiones de los números y estrellas, creamos dos nuevos datasets, uno con las dos variables del tipo integer con las variables 'numero' y 'repeticiones' y el segundo dataset con las variables 'estrella' y 'repeticiones'. A partir de estos nuevos datasets generamos dos gráficos para ver las repeticiones de cada uno de los números y cada una de las estrellas:





Por último, podemos ver los 5 números y las dos estrellas que más veces se han repetido en los sorteos:

	numero	repeticiones
49	50	147
43	44	147
22	23	143
25	26	139
16	17	138

	estrella	repeticiones
1	2	254
7	8	254

5- Contenido

Como ya he detallado en el punto 3, disponemos de un total de 7 campos en el dataset principal y de dos campos de los datasets auxiliares que he utilizado para generar las gráficas y los números más repetidos a lo largo de los sorteos.

El periodo de captura de los datos va desde el 10 de Mayo de 2004 y hasta el último sorteo realizado el pasado 25 de Octubre de 2019.

El proceso de captura de los datos se ha realizado a través del lenguaje de programación Python, utilizando el IDE Spyder de Anaconda. Esquemáticamente definimos el proceso de captura de los datos:

- Importamos las librerías necesarias
- Definimos e inicializamos las variables que necesitamos para la captura de los datos
- Creamos el agente que utilizaremos para el rastreo
- Definimos la url que utilizaremos en el proceso
- Definimos el periodo de captura a través de las variables 'fecha_ini' y 'fecha_fin'
- Transformamos la fecha a formato requerido por la url
- Definimos los días que restaremos de la fecha y otras variables que son necesarias para establecer si la fecha es de sorteo o no
- Definimos el DataFrame con las columnas que utilizaremos
- Hacemos un bucle 'While' hasta que fin sea igual 0 para determinar la url de cada sorteo que corresponderá con la url inicial añadiendo la fecha de cada sorteo con el formato ya definido:
    - Dentro de cada url del sorteo, procedemos a realizar dos bucles for para capturar los datos de los números y estrellas de cada sorteo que añadimos a un array.
    - Finalmente añadimos los datos del array al DataFrame
    - Restamos un día de la fecha del sorteo hasta que disponemos la fecha corresponde a "martes" o a "viernes"
- Una vez la fecha resultante es menor que la fecha_fin, finalizamos el bucle while, generemos el fichero "Euromillones.csv"
- Para generar las gráficas con las repeticiones de cada uno de los números y estrellas que participan en el sorteo, definimos las variables que contendrá las repeticiones de cada columna, para ello debemos convertir los valores del DataFrame en tipo integer, también definimos las variables totales para cada número y estrella
- Definimos los dos DataFrame que contendrá los datos de repeticiones
- Generamos un bucle for que recorra todos los números del sorteo y cuente las repeticiones de cada columna y para cada número, finalmente los añadimos al DataFrame.
- Generamos un bucle for que recorra todas las estrellas del sorteo y cuente las repeticiones de cada columna y para cada estrella, finalmente los añadimos al DataFrame.
- Generamos las gráficas, las guardamos en un fichero .jpg y las mostramos en la consola
- Para finalizar este proceso, ordenamos los DataFrame por el número de repeticiones de forma descendente y mostramos los 5 primeros para los números y los dos primeros para las estrellas.

6- Agradecimientos

A la web *"https://www.combinacionganadora.com/"* y a su webmaster por recolectar y mantener los resultados de todos los sorteos de Euromillones, así como facilitar su rastreo para poder realizar el web scraping necesario para poder obtener los datos mostrados en esta práctica.

7- Inspiración

Teniendo en cuenta los botes que se generan en estos sorteos, hasta un máximo de 190 millones de Euros, tenía curiosidad por saber los números que más veces han aparecidos en los sorteos de este juego, por supuesto no pretendo establecer un método para calcular o predecir las posibles combinaciones ya que la probabilidad de acierto es tan sumamente remota (1/140 millones aprox.).

Soy consciente que ya existen sitios webs que informan de estas estadísticas, pero me ha parecido oportuno rastrear y crear estas estadísticas por mi cuenta y aplicando los conocimientos vistos en los módulos de estudio, también como no practica el lenguaje Python que a pesar de que está considerado uno de los más sencillos para iniciarse en la programación, me resulta especialmente complejo por el sistema de identación o sangrado.

8- Licencia

Teniendo en cuenta que se trata de un código muy sencillo y que no veo ninguna necesidad de no compartirlo, he establecido que lo publicaré bajo licencia CC BY-NC-SA 4.0, ya que según https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es_ES presenta las siguientes características:

- Los datos, documentos, código, etc., están sujetos a compartir, copiar y redistribuir en cualquier medio o formato.
- Se pueden adaptar, mezclar y transformar a partir de este material

Para poder realizar uso de este material, se deberá considerar:

- Que es necesario dar crédito de manera adecuada, indicar la licencia y significar los cambios realizados en los materiales sujetos a esta licencia
- No se podrá hacer uso con fines comerciales
- El uso de estos materiales se debe distribuir bajo la misma licencia

9- Código

Ver fichero EuroMill.py en el siguiente enlace:

https://github.com/gonmard/UOC_PRAT1_WEB_SCRAPING/blob/master/code/EuroMill.py

Ver fichero EuroMill_Selenium.py en el siguiente enlace:

https://github.com/gonmard/UOC_PRAT1_WEB_SCRAPING/blob/master/code_selenium/EuroMill_Selenium.py

10- Dataset

El dataset generado “EuroMillones.csv”, consta de 10 columnas separadas por comas, la primera columna corresponde al índice del dataset, las dos siguientes corresponde a la fecha del sorteo y el día del sorteo, las cinco siguientes corresponde a la combinación del sorteo y las dos últimas a las estrellas, para ver el detalle del dataset, podemos consultar el siguiente enlace:

https://github.com/gonmard/UOC_PRAT1_WEB_SCRAPING/commit/6c6a27654f3411717f6d5cc8143a8ef95b07c5e1

12- Componentes del grupo

Esta práctica ha estado realizada por Diego González Martínez.

13- Complemento a la práctica

Tras la entrega previa, la profesora Laia Subirats me sugirió complementar la práctica con la aplicación de la tecnología de Selenium, siguiendo su recomendación, he modificado parcialmente el código original creando una versión que utiliza parcialmente la tecnología de Selenium.