

Escuela de Arquitectura, Ingeniería y Diseño

PRÁCTICA. EDA

Nombre: Sofía Corral Caballero

Número de expediente: 21846911

Asignatura: Lenguajes de programación estadística

Profesor: Christian Vladimi Sucuzhanay Arevalo

INTRODUCCIÓN

En este trabajo voy a realizar el EDA(análisis exploratorio de los datos) de cuatro dataSets. Para así, poder conocer mejor mi negocio. Es importante saber que cuando hablamos de análisis exploratorio de los datos nos referimos a poder conocer mejor los datos y la información que nos transmiten. Aparte de conocer mejor los datos mediante herramientas de visualización de datos como puede ser el diagrama de dispersión, el histograma o los gráficos de sectores. También forman parte del EDA cálculos que me indiquen el máximo, mínimo, que me seleccionen una única columna.... En definitiva, todas aquellas operaciones que me ayuden a conocer mejor los datos y a poder sacar la mayor información posible.

Usuario de gitHub: sofiaCorral

PRIMER DATASET (POTENCIA TOTAL INSTALADA POR PAÍS)

Este dataSet me indica el número total de potencia que hay instalada en cada país. La finalidad principal que tengo con este dataSets es descubrir la evolución de España en relación a la potencia instalada.

Los datos los he obtenido haciendo scraping sobre la página web en la que aparecían los datos, he importado dichos datos a un excel y posteriormente los he puesto como públicos. Además, es importante tener en cuenta que el url generado trabaja en Streaming, es decir, los datos se van modificando en función se vayan modificando en el conjunto de datos.

Dentro de este conjunto de datos me he encontrado con algún que otro problema, y es que al hacer scraping de la página, esta te cogía todo, incluido los hipervínculos y otros símbolos. Haciendo que el dataSet estuviera sucio como aparece en la siguiente imagen:

L	M	N	O	P	Q	R	S
Total2012[240]	Total2013	Total2014	Total2015	Total2016	Total2017	Total2018	Total2019
102 024[241]	138 900[242]	177 000[125]	230 000	~306 500	~401 500	~512 000	~647 000 (est.)
8043	19 800[242]	28 100[125]	43 000[128]	77 000[127]	127 000[243]	~171 000	~205 700
68 640	78 970[244]	86 674[245]	95 200[128]	~101 500[127]	~109 000[243]	~120 000	~131 900
7665	12 100[246]	18 600[247]	28 400[128]	40 610[147]	51 000[243]	~61 400	~68 200
6704	13 600[242]	23 300[125]	33 300[128]	42 410[147]	49 000[248]	55 600[249]	~61 800
32 411	35 600[250]	38 128[251]	39 550[252]	40 600[253]	42 900[254]	~46 000	~49 900
1839	2180[255]	3382	5130[256]	10 000[257]	~20 000[243]	~28 000	~34 800
16 987	18 400[242]	18 500[125]	18 800[258]	19 160[147]	19 400	19 700	~20 900
2291	3100[259]	4100[260]	4728[261]	5440[261]	7203	11 085	15 900
1831	2706[262]	5000[263]	8437[264]	11 460[265]	12 790[266]	13 098	13 300
1006	1448[242]	2384	3200[267]	5000[268]	5835	7850	~10 500
3843	4598[269]	5700[125]	6800[270]	7170	8044[271]	~9000	~10 500
4537	4651	4656[272]	4667[212]	4686	4688	4714	8711

Para poder limpiar esto de una manera rápida, he recurrido a añadir en mi xpath '/td' que sirve para separarlo por celda y así obtener un mejor resultado que mediante normalización he conseguido que estuviera limpio.

Las búsquedas que he realizado de dicho dataSet y que me van a ayudar a conocer mejor mi negocio son:

- Miro si hay algún valor nulo en el conjunto de datos. El resultado ha sido:

```
> Nulos <- sum(is.na(potencia))  
> Nulos  
[1] 525
```

Es decir, si tenemos valores nulos. Observando el dataSet podemos ver como es cierto que los valores que no están rellenos deberían estarlo. Como en nuestro caso, el país más importante es España, puesto que mi proyecto está centrado solo en España, decido centrarme más en él.

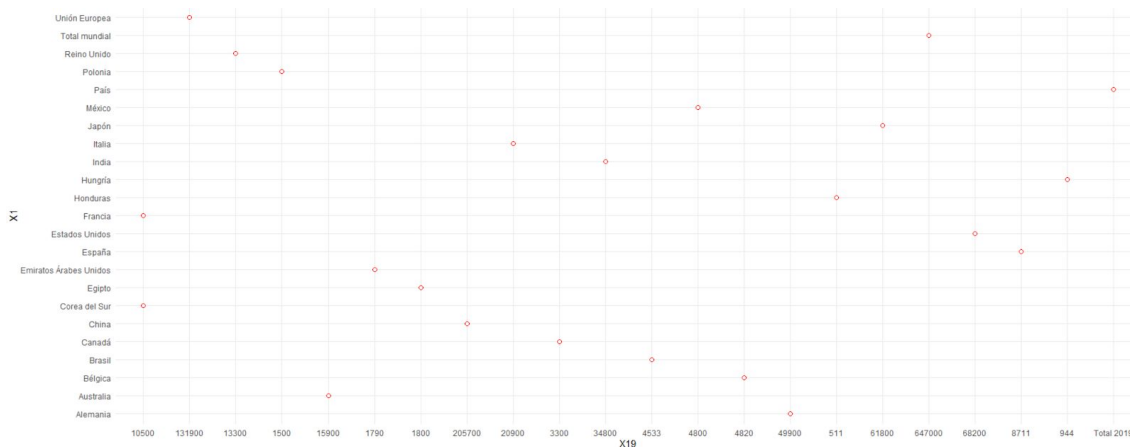
- Pese a que mi dataSet me daba información de todos los países, mi negocio solo estaba centrado en España. Por lo que he decidido en primer lugar imprimir solo la fila correspondiente a los valores de España mediante un filtro. Dicha columna es:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18
1 España	7	12	23	48	145	693	3354	3438	3892	4214	4537	4651	4656	4667	4686	4688	4714	

- Visto que en el primer cálculo, había muchos valores nulos, decidí calcular si había algún nulo en la fila correspondiente con España, en este caso el resultado fue 0, no hay valores nulos:

```
> NulosEspania <- sum(is.na(Espania))  
> NulosEspania  
[1] 0
```

- Como el conjunto de datos era de tipo character, no era posible realizar ningún diagrama de líneas o de barras, decidí hacer un diagrama de dispersión que eso sí es posible. En él, represento de todos los países el los Kw instalados en 2019. El resultado fue:



Para que el resultado quedará bien, he decidido eliminar antes los valores nulos, es decir, únicamente están representados aquellos países que no tenían valores nulos, ya que un dato nulo no aporta información útil al respecto.

SEGUNDO DATASET (EPDATA EVOLUCIÓN ENERGÍA CONSUMIDA EN ESPAÑA)

Este dataSet me indica la evolución que ha habido en España en función al consumo de energías. La finalidad de analizar el dataSet es conocer principalmente la evolución de la energía consumida en España.

El principal problema que me he encontrado con este dataSet es que no me permitía conseguir el url correspondiente al conjunto de datos, debido a la gran seguridad que tiene dicha plataforma para no permitirnos hacer scraping. EpData funciona mediante una API privada y de la que no puedo conseguir su Query. Por lo tanto, como permite descargarnos el dataSet, esta vez me lo he descargado y para cargarlo más fácil en R he usado 'file.choose'. El dataSet también estará subido en el repositorio de GitHub.

Las diferentes consultas que he realizado con este dataSet son:

- Calcular el valor máximo de energía solar consumida: El resultado obtenido es el siguiente, que como podemos observar no es muy elevado.

```
> max(epData$Solar)
[1] "3,2"
```

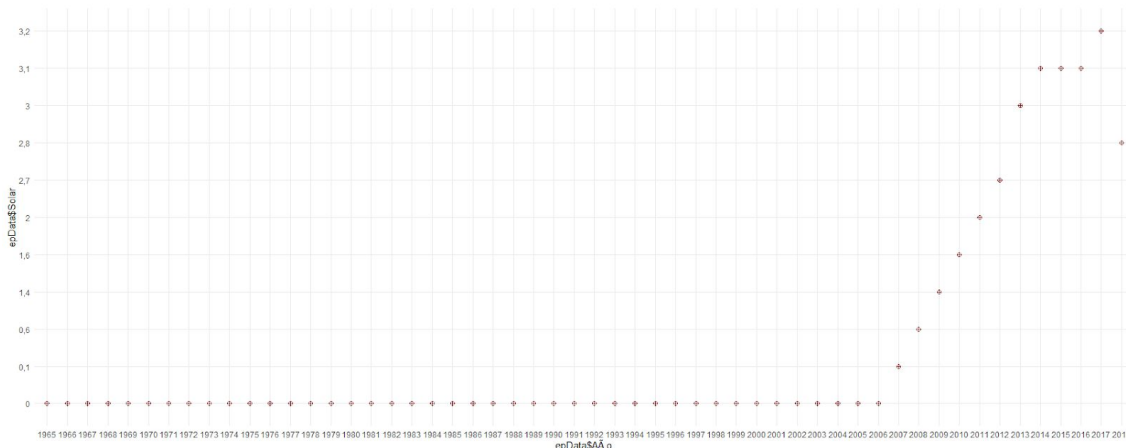
- También he mirado si existe algún valor nulo y, en este caso la respuesta es bastante favorecedora, ya que el dataSet no tiene ningún valor nulo.

```
> NulosepData <- sum(is.na(epData))
> NulosepData
[1] 0
```

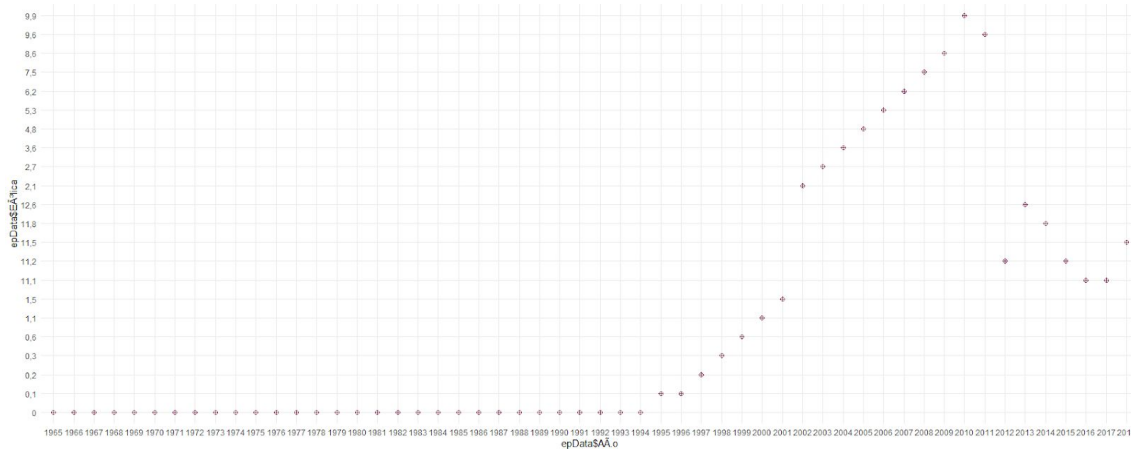
- Como de todos los valores, el que más me interesa es el más actual. He decidido imprimir la fila correspondiente al año 2018, que es el más actual que recoge mi dataSet. El resultado ha sido:

```
> epData2018 <- epData[epData$Año == "2018", ]
> epData2018
  Año Periodo Petr leo Gas Carb n Nuclear Hidroel ctrica
54 2018      A o      66,6 27,1   11,1   12,6           8
  Geotermal..biomasa.y.otra Solar E lica
54                1,7   2,8   11,5
```

- Como visualizaci n de datos, lo primero que he realizado ha sido un gr fico que me indique en funci n de los a os la evoluci n de la energ a solar. Gracias a el resultado, podemos ver como ha ido en aumento, aunque en 2018 bajo, quiz  porque como los datos solo llegan hasta esa fecha no est  al completo el resultado. Tambi n podemos ver como el uso de la energ a solar es algo nuevo. El gr fico obtenido es el siguiente:



- Otro gr fico importante es el de la evoluci n e lica. Cuando comenc  con las ideas del proyecto no sab a si centrarme m s en energ a solar o e lica. Investigando, descub  que lo que en la actualidad est  m s en auge era la solar. Gracias al siguiente gr fico puedo corroborar dicha informaci n. Adem s, podemos ver c mo en funci n va creciendo la energ a solar, disminuye la e lica. El gr fico resultante sobre la evoluci n de la energ a e lica en Espa a es el siguiente:




TERCER DATASET (ANIMALES COGIDO DE ESPECIES)

Este dataSet es uno de los más importantes, ya que me da información sobre todos los tipos de especies que hay y donde están distribuidos. Conocer las especies que hay en España, es muy importante de cara a poder detectar un lugar adecuado y que cumpla con todas las normas legales.

Este dataSet, al igual que el anterior tiene un gran sistema de seguridad, ya que no permite poder hacer scraping, sino que es necesario descargar todo el conjunto de datos para poder hacer el análisis de ellos. El conjunto de datos está subido a gitHub al igual que el anterior y, para poder cargarlo fácilmente en R he usado 'file.choose'.

Lo primero que he hecho ha sido determinar cuántas filas o ejemplos tiene mi dataSet. Para ello, no es necesario aplicar ningún código.

 datosAnimales	38588 obs. of 26 variables
--	-----------------------------------

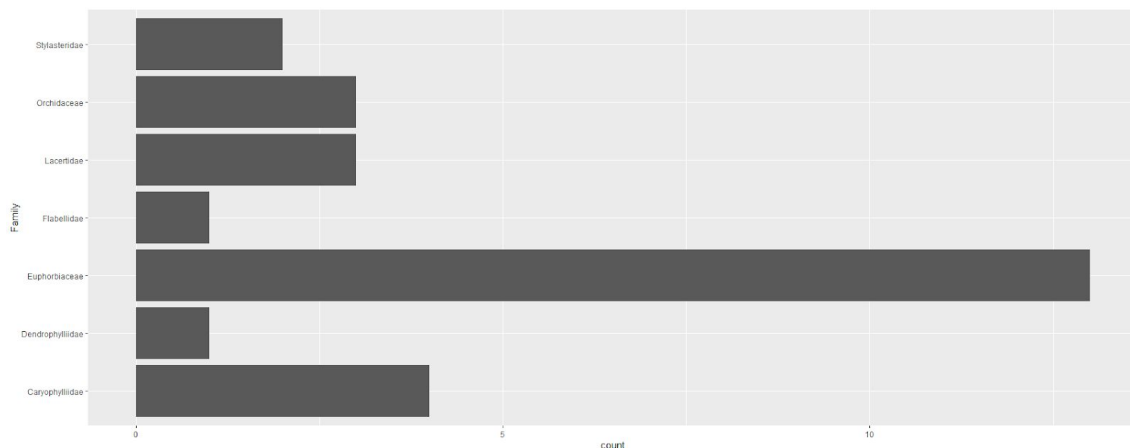
La operación importante de dicho dataSet es poder descubrir el tipo de especies que hay en España, para ello, he usado expresiones regulares:

```
#especies solo espania####
soloEspania <- datosAnimales %>% filter(All_DistributionFullNames=='Spain')
soloEspania
```

Otra consulta que he realizado es, contar cuantos ejemplos o animales distintos hay en España, el resultado en este caso se puede ver sin utilizar código:

 soloEspania	27 obs. of 26 variables
--	--------------------------------

Por último, he realizado un gráfico de barras que me indica las distintas familias de especies que hay. El resultado es el siguiente:



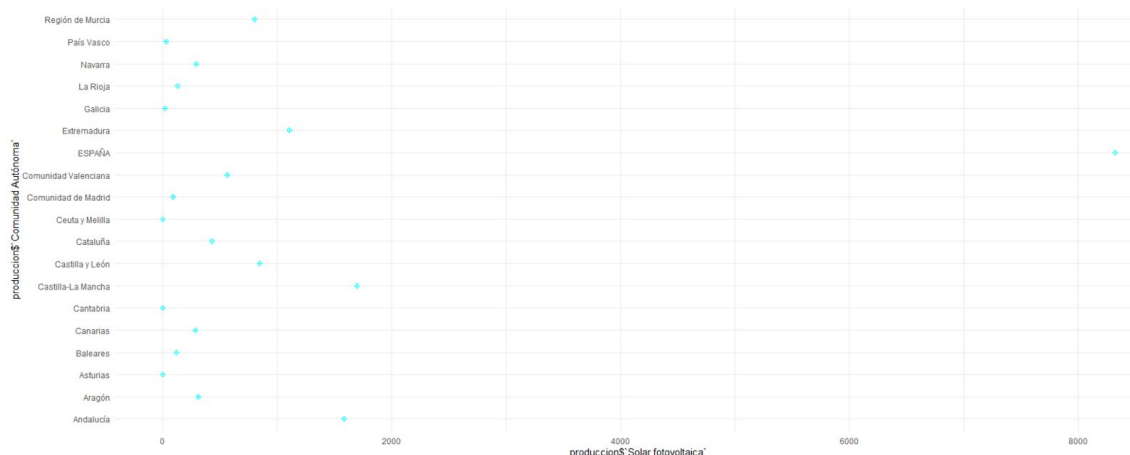
CUARTO DATASET (PRODUCCIÓN DE ENERGÍA POR COMUNIDAD)

Este dataSet me indica por comunidad autónoma la producción de kw de cada tipo de energía que hay (hidráulica, eólica, solar...). La finalidad principal que tengo con este dataSets es conocer mejor mi negocio y la zona en la que más energía solar fotovoltaica se produce.

En primer lugar, los datos los he obtenido haciendo scraping de la página e importarlos en un excel que se va a ir modificando si lo hace la página (streaming). Una vez que tenía los datos, los he cargado en R y he comenzado a explorar mi dataset.

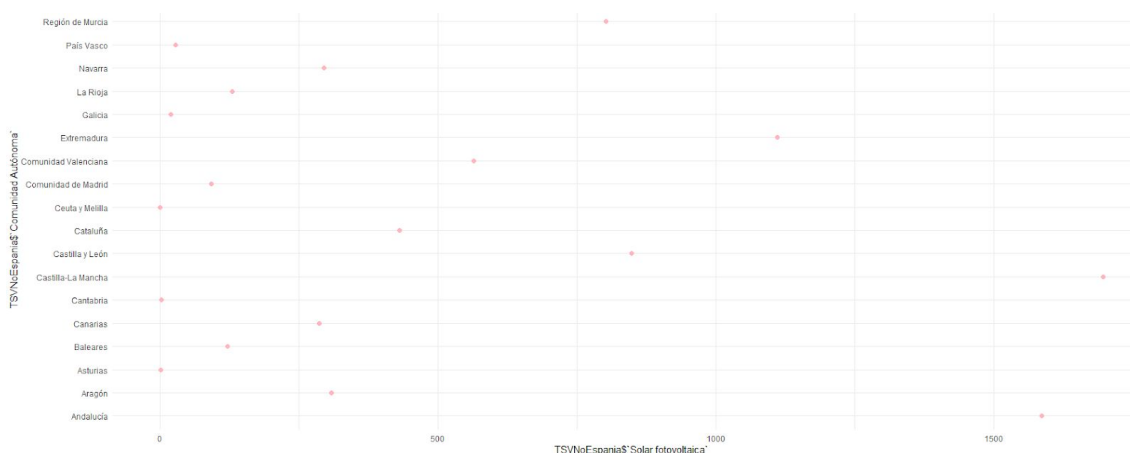
Sabiendo que mi dataSet tiene el mismo número de filas o ejemplos que comunidades autónomas hay en España más una fila que me indica la producción total de España, he procedido a ver diversos datos que me pudieran parecer relevantes. La información que he sacado de este dataset es:

- Visión general de la producción de energía fotovoltaica: En este caso, he generado un diagrama de dispersión donde me indica de cada comunidad autónoma y de España, el número total de Kw que se han producido de energía fotovoltaica. El resultado es el siguiente:



Con esta gráfica podemos darnos cuenta cómo al estar España, va a hacer que no podamos ver con claridad la producción entre Comunidades Autónomas.

- Crear un nuevo dataset con solo los valores de las Comunidades autónomas: El objetivo de esto es poder ver a una mayor escala la diferencia que hay de producción de energía fotovoltaica en las diversas comunidades.
- Determinar cuál es la Comunidad Autónoma que más energía fotovoltaica produce mediante un diagrama: En este caso, he decidido representarlo con un diagrama de dispersión al igual que la anterior, ya que es con la que mejor vamos a poder ver la diferencia. Para este caso, hemos utilizado el dataset creado sin España y el resultado es el siguiente:



Gracias a este diagrama de dispersión podemos observar cómo la Comunidad autónoma que más produce energía fotovoltaica es Castilla La Mancha. No obstante, la Comunidad Autónoma de Andalucía no está muy lejos.

- Ver numéricamente la Comunidad Autónoma que más energía produce: Para este caso, como solo queremos saber el valor máximo de Kw que ha producido la Comunidad Autónoma. Solo imprimimos dicho valor, en este caso el valor es:

```
`solar fotovoltaica`  
      <dbl>  
      1697
```

- Ver la Comunidad autónoma que menos energía fotovoltaica produce: En este caso, lo que quiero saber es qué Comunidad Autónoma es la que menos energía fotovoltaica produce. Para ello, solo imprimimos el nombre de la Comunidad Autónoma, que en este caso es:

```
`Comunidad Autónoma`  
      <chr>  
1 Ceuta y Melilla
```

Como podemos ver, este conjunto de datos no tiene ningún valor nulo, es decir, no presenta ninguna anomalía. A parte, los resultados obtenidos tienen relación con la realidad. Por lo tanto, podemos decir que el conjunto de datos está correcto. Para confirmar que el dataSet no tiene ningún valor nulo, podemos ver la siguiente imagen:

```
> Nulos <- sum(is.na(produccion))  
> Nulos  
[1] 0
```

CONCLUSIÓN

Con la realización de este trabajo, he podido ampliar mis conocimientos. Me ha ayudado a conocer todas las posibles soluciones que hay de conseguir datos y que, no siempre es posible hacer scraping, como puede ser en el caso de epData y del dataSet de animales.

Además, he podido poner en práctica todos los conocimientos adquiridos por las presentaciones realizadas en clase en relación a la visualización de datos y comprender que gráficas son las posibles para realizar en función el tipo de datos.

Algo bastante importante a tener en cuenta es que, la información adquirida no siempre tiene que ser mediante un gráfico o diagrama.

Por último y relacionado con las conclusiones que he podido sacar de cada dataSet en rasgos generales es que, la producción de potencia total instalada en España está aumentando, la Comunidad Autónoma con mayor producción de energía es Castilla La Mancha. La evolución de la energía solar está en aumento pero aún le queda un gran camino por delante y, hay un total de 6.284 especies diferentes en España.

En definitiva, un trabajo bastante interesante e importante a desarrollar de cara a la realización de mi proyecto final.