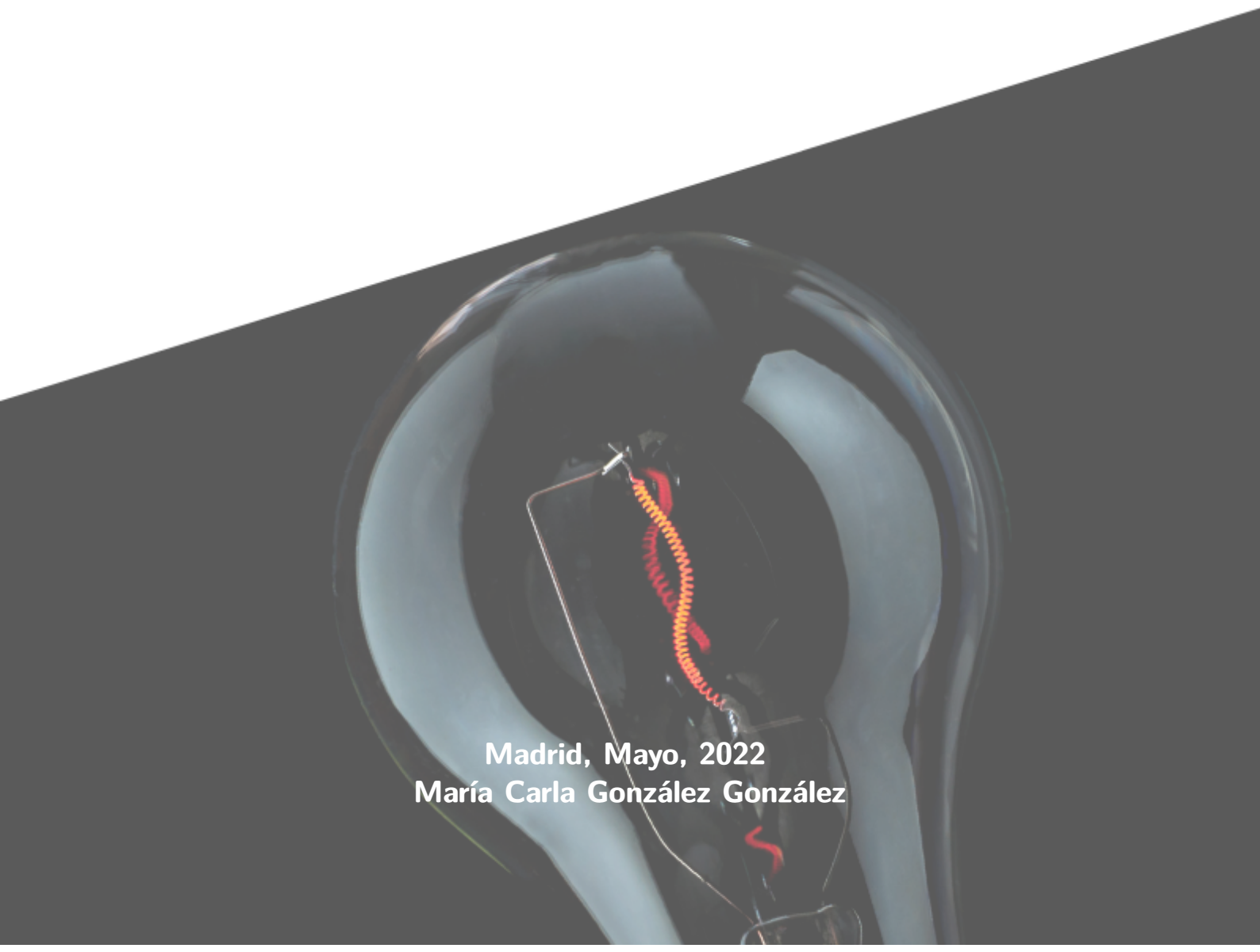
THE BRIDGE

Análisis exploratorio de datos (EDA): España y sus fuentes de energía.

Data Science

THE BRIDGE DIGITAL TALENT ACCELERATOR



Madrid, Mayo, 2022
María Carla González González

Índice

1 | Introducción

Este proyecto consiste en realizar un análisis exploratorio de datos (EDA) relacionado con la generación de energía en España en el periodo de tiempo comprendido entre el 1 de enero de 2015 y el 31 de diciembre de 2018.

Se plantea el mismo de forma que pueda ser ampliable y extrapolable a un periodo de tiempo mayor.

1.1 | Contexto

Se cuenta en un comienzo con dos bases de datos en formato .csv:

- `spn_energy_data.csv`: datos de generación de energía en el periodo comentado.
- `spn_weather_features.csv`: datos climáticos de cinco grandes ciudades de España en el periodo comentado.

Estas bases de datos son las que se van a tratar y acondicionar para su posterior análisis. Sobretodo este estudio se centra en los datos de la generación de energía, el archivo del clima es un .csv complementario que puede o no proporcionar algunas respuestas a las preguntas que se plantean.

1.2 | Hipótesis

Antes de abordar el acondicionamiento de estas bases de datos y el posterior análisis de los mismos, se hace necesario el plantamiento de una serie de preguntas, hipótesis y suposiciones que marquen la dirección de este proyecto.

La primera pregunta y más general que se pretende responder es:

¿Cuál es la energía o energías idónea(s) para invertir?

Para dar respuesta a esta, se crean subretos o subpreguntas:

- ¿Cuál ha sido la evolución de las distintas fuentes de energía?
  - Sería lógico pensar en un posible aumento en renovables y disminución en combustibles fósiles.
  - Es posible que haya estacionalidad, debido a los factores climáticos, sobretodo en renovables.
- ¿Se ha generado tanto como se ha demandado?
- ¿Cuán eficiente es cada fuente de energía? Teniendo en cuenta efectos adversos.
  - Se puede pensar que la nuclear es la más eficiente.
- ¿Hay una clara influencia de las condiciones climáticas?

Se hacen algunos comentarios de hipótesis o pensamientos que se dan en a priori,que serán o no ciertos tras el estudio.

Además de estas, muchas otras se plantean, que también pueden ser relevantes para estudios posteriores como el factor económico: *¿Cuánto cuesta tanto producir la energía como las instalaciones?¿Rentabilidad?*

2 | Acondicionamiento de datos

Se acondicionan las bases de datos comentadas: `spn_energy_data.csv` y `spn_weather_features.csv`. Sobretodo este apartado está orientado a la primera, que es la utilizada durante todo el EDA.

Nociones previas:
Número de filas y columnas:

- `spn_energy_data.csv`:(35064,29)
- `spn_weather_features.csv`:(178398,17)

La variable temporal 'time': comprendida desde 1 de enero 2015 00:00 hasta 31 de diciembre de 2018 23.59. En intervalos de tiempo de una hora siendo tipo str.

2.1 | Limpieza y exploración

1. Se examina qué variables están en el dataset.
2. Se comprueban si hay alguna de la que se prefiera prescindir para el análisis. Y en tal caso se elimina dicha columna.
3. Se comprueban las columnas y filas que sean NaN completamente y se eliminan.
4. Se comprueban las columnas que sean totalmente cero, es decir, no aportan información. Y también se eliminan.
5. Tratamiento de los NaN restantes, en este caso se sustituyen por el valor anterior para `spn_energy_data.csv` e interpolan para `spn_weather_features.csv`
6. Comprobación de los outliers. En este caso se deciden dejar para `spn_energy_data.csv` y eliminar para `spn_weather_features.csv`.

Completada esta fase el dataset estaría limpio y listo para realizar las operaciones pertinentes.

2.2 | Agrupación

1. Se crean las variables correspondientes a las agrupaciones de columnas:
   - Renovables.
   - No renovables.
   - Conjunto de energías fósiles.
   - Conjunto de energías hidráulicas.
2. Se modifica realiza una copia del dataset, para no perder el limpio si hay errores.
3. Se altera el formato del tiempo y se realizan nuevos dataset en función de la variable temporal.
   - Día.
   - Mes.
   - Año.
4. Divisiones concreta y alteraciones que fueron siendo necesarias para las visualizaciones.

En el caso de `spn_weather_features.csv` se agrupó por fecha y se realizó la media de todas las ciudades.

3 | Análisis de los datos

El análisis de los datos consta de dos partes:

- Análisis visual.
- Análisis estadístico.

En los siguientes apartados se pretende responder a las preguntas plantedas en el apartado 1.2.

3.1 | Evolución de la generación de energía en España

En cuanto a la progresión de la generación de las distintas fuentes de energía se analizan:

- Las posibles tendencias.
- Las porcentajes de generación de cada fuente.
- La posibilidad de estacionalidad de las fuentes.

La mayoría de los análisis se realizó de manera visual.

Tendencia:
En un primer momento se puede pensar en un aumento de las energías renovables, dado que se incrementó en esos años la inversión en dicha tecnología. Sin embargo, no se puede detectar esa tendencia creciente en el intervalo evaluado, como se observa en la figura A.2. Esto puede ser debido a factores como:

- La influencia del clima.
- A pesar del crecimiento en las instalaciones renovables, la tecnología no ha madurado lo suficiente.

También se ha visto la progresión año por año, habiendo picos a la alza y a la baja pero no pudiendo hacer una afirmación en cuanto a tendencia de cara a los siguientes años. Para ello será necesario ampliar el intervalo de tiempo.

Porcentaje de generación:
En las gráficas A.3, A.4 y A.5 se observan los porcentajes de generación totales del periodo. Y la división entre renovables y no renovables.

La fuente que en su conjunto produjo una mayor generación de energía son las fósiles, sin embargo, la nuclear y la eólica tienen también unos porcentajes altos y las siguientes a comentar serían el conjunto de las hidráulicas y la solar.

Estacionalidad:
Otra característica a comprobar de las distintas fuentes es la estacionalidad, es decir, si hay relacion entre la época/mes del año y el aumento o disminución de la producción de energía.

Se responde afirmativamente a esta pregunta tanto por el análisis visual como el estadístico. Visualmente se puede ver en A.6 que, sobretodo la solar y la eólica dependen del periodo del año, que puede tener relacion con las condiciones climáticas, para un aumento o disminución en su generación. Añadir, que se observa inversa la producción entre la eólica y la solar. La nuclear, por su parte, se mantiene bastante constante a pesar de haber alguna época más baja, que podría ser circunstacial por el cierre de alguna central nuclear.

Tras realizar un estudio de la progresión de las energías y ver que no hay una tendencia clara que permita tomar una determinación, ni responder a las preguntas que en un principio se plantean. Entonces, se decide buscar nuevas fuentes que den más luz en este proyecto. En distintos artículos se nombra como fuente el organismo ree: red eléctrica de España y se decide extraer más datos e información a partir de su api: https://www.ree.es/es/apidatos.

3.2 | Energía generada vs energía demandada

Se hace una comparación entre la energía generada y la demandada en este periodo. A priori no se podría saber si España es importadora o exportadora, ni si genera la suficiente energía para autoabastecerse. El autoabastecimiento entendido desde el punto de vista de generar al menos lo mismo que demanda, ya que la idea de autoabastecimiento se ve afectada por el mercado global que hay alrededor de la energía y que afecta el balance económico, además de otros factores como podrías ser la regulación de emisiones y las penalizaciones por las mismas.

Se extrae de la api del ree información sobre la demanda energética entre [2015,2019) y se compara con la generación total en la gráfica 3.2. Observando que la demanda en ese intervalo es mayor que la producción. Esto según algunos artículos puede deberse a varios factores:

- La crisis económica, junto con el impuesto del CO2 hicieron que fuese más barato comprar la energía, es decir importarla que producirla para el Estado Español.

España hasta 2016 había sido exportadora en computo final, sin embargo en este periodo se ha invertido la ecuación.

3.3 | Eficiencia energética vs efectos adversos

Hay mucho debate sobre las ventajas y desventajas de una fuente y otra. Entre ellos:

- El coste de la producción de energía, habiendo energías más y otras menos rentables generación-coste.
- El impacto medioambiental: visto de todo punto de vista.
- Los problemas socioeconómicos que puedan provocar apostar por una u otra fuente de generación.

Como algo si parece estar más claro y es lo perjudicial de las emisiones, en nuestro caso, se decide eliminar las energías fosiles como posible inversión. Como guiño, no propondría su eliminación sino su reconversión dado que evita otros problemas y puede causar beneficios, por ejemplo como centro de almacenaje de energía cuando avance dicha tecnología.

Por otro lado, también se toma la decisión de no hacer énfasis en las hidráulicas por su limitación a las condiciones geográficas. A pesar de ser una opción que en concreto en este periodo tuvo un aporte fuerte y se intuye el ratio generación/potencia instalada bastante positivo.

Entoces, nos centramos en:

- No renovable: nuclear.
- Renovable: eólica y solar.

Con los datos extraídos, mediante la api del ree, de la potencia instalada en ese periodo para dichas fuentes de energías se obtiene la gráfica A.9. En este se observa una significativa diferencia entre las renovables y la nuclear. Este resultado era esperable, dado que desde 1983 España cuenta con 7 centrales nucleares de las cuales mínimo en ese periodo estaban activas 5 de ellas y estas producen una gran cantidad de energía de forma constante. Entre la solar y la eólica, la solar es la que tiene mayor eficiencia y se comprueba en esta gráfica. Sin embargo, sorprende que en cómputo total sea la eólica la primera en generación de las renovables. Esta apuesta por la eólica puede haber estado motivado por cuestiones climáticas, dado que la solar es más sensible a estos factores. Se comprueba en el siguiente apartado.

3.4 | Influencia del clima en la generación de energía

La influencia o no del clima en estas tres fuentes de energía que se están evaluando, se realiza mediante el .csv que se ha tratado con anterioridad, donde se evalua si la media de las variables meteorológicas tiene una influencia en la generación de energía. Este análisis se hace estadisticamente y visualmente mediante un mapa de calor que las correlaciona. Este mapa de calor es el de la figura 3.4. El análisis estadístico se rearliza utilizando el test χ^2.

En ambos casos se obtiene el mismo resultado:

- No hay una dependencia clara: se podría añadir, que no es concluyente este estudio dado que se trata de una muestra de las condiciones climáticas en España. Habría que ampliar el estudio, hacer otras comprobaciones para confirmar que la muestra sea suficiente y contrastar si es cierta esta independencia climática con otros estadísticos si corresponde.
- Sin embargo, visualmente se puede ver las correlaciones y aunque débiles todas. Las renovables son las que tienen alguna relación de dependencia, sobretodo la solar que se ve afectada de forma inversamente proporcional a las nubes y a la humedad. La eólica por su parte tiene cierta correlación con el viento.

4 | Conclusiones

En este primer análisis de los datos se puede concluir que:

- En la evolución de la generación de energía en el periodo 2015-2019 no hay una tendencia clara de crecimiento o decrecimiento significativo en ninguna de las fuentes.
- Aunque la generación de las energías fósiles en su conjunto son las que más han generado en esos años, no están muy atrás la nuclear y la eólica, es más, si se divide dicho conjunto toman el primer y segundo puesto, respectivamente, en generación total.
- Teniendo en cuenta las emisiones de CO2, se decide descartar las fósiles como opción de inversión. Al igual que la hidroélectrica por su fuerte dependencia y limitación de los factores geográficos.
- Opciones a destacar:
  - No renovables: energía nuclear.
  - Renovables: energía eólica y solar.
- La nuclear es la que se mantiene más constante en su generación a lo largo del periodo.
- La eólica y la solar tienen mayor fluctuación, gráficamente y estadísticamente los datos nos indican cierta estacionalidad (la nuclear también pero menor). Se ve una estacionalidad inversa entre eólica y solar.
- En cuanto a la eficiencia, la nuclear es la lider, seguida de la solar y después la eólica. Es un resultado curioso, dado que la solar aún siendo más eficiente que la eólica genera menos. Es posible que venga motivado por la mayor dependencia de la solar de las condiciones climáticas que la eólica y por ello se ha invertido menos en infraestructura.
- Como se ha comentado gráficamente se puede ver una dependencia, aunque débil de factores climáticos y las renovables.
- España en este periodo se confirma como importadora de energía en balance tras la evaluación generación-demanda de energía.

Para finalizar y con vista futuros análisis relacionados, o amplicación de este:

- Es necesario seguir indagando para hacer una elección más certera. Sugerencias:
  - Ampliar el periodo de evaluación, para ver tendencias, etc.
  - Factores económicos.
- También se plantea la pregunta del propósito, es decir, el peso eficiencia-efectos adversos.
- En cuanto a eficiencia la nuclear estaría por delante, pero si se pretende invertir en la optimización de las fuentes renovables sugeriría la eólica o una combinación de ambas dada su estacionalidad inversa.

A | Apéndice A: Representaciones gráficas

Se introducen las gráficas que se han considerado más relevantes para la explicación del proyecto y las cuales han sido fundamentales para la argumentación. El resto de análisis se encuentran en los archivos notebook que se han ido realizando a lo largo del análisis.

A.1 | Evolución de la generación de energía en España

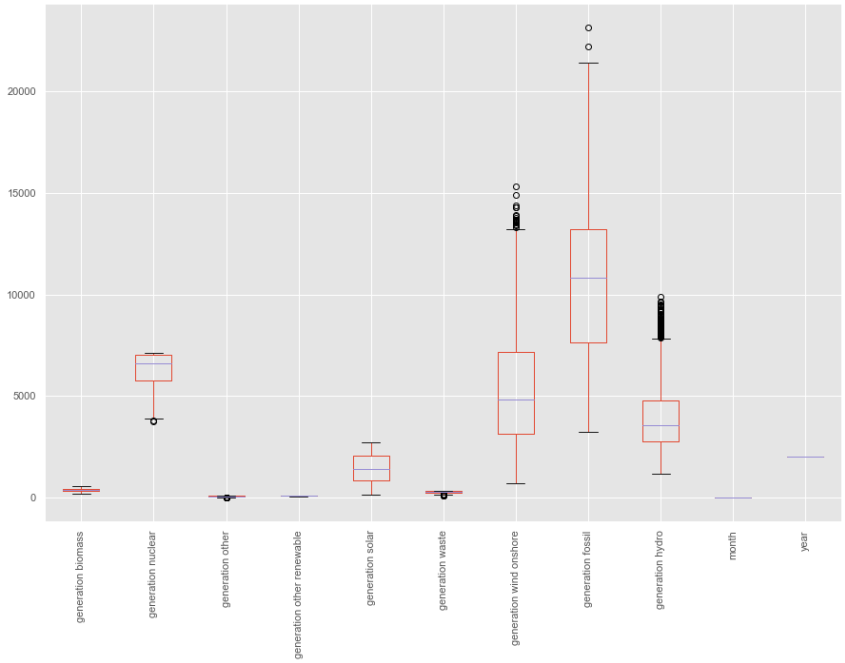


Figura A.1: Boxplot fuentes de energía.

Volver al apartado 3.1.

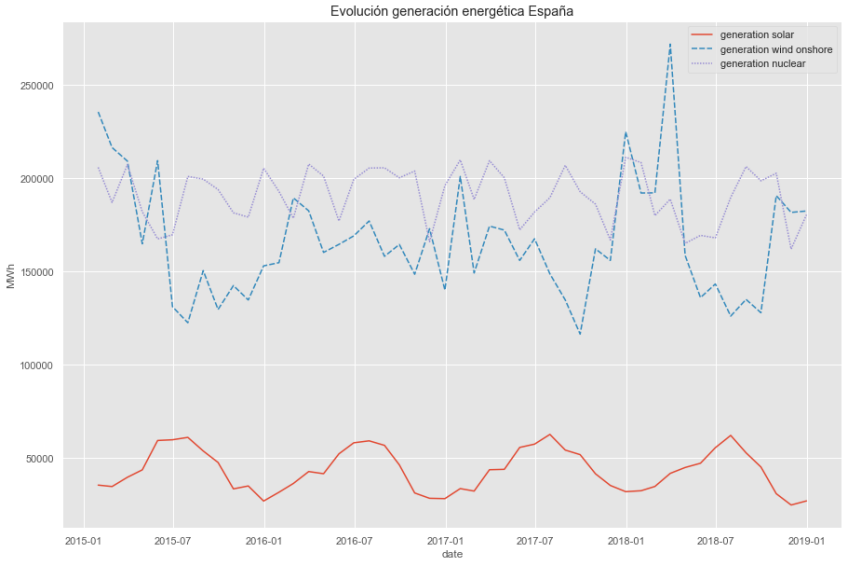


Figura A.2: Tendencia fuentes de energía.

Volver al apartado 3.1.

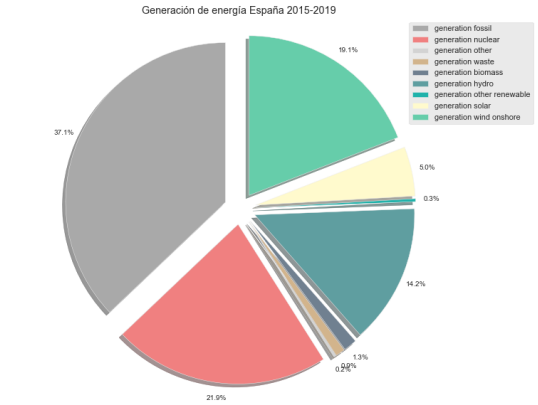


Figura A.3: Porcentaje de generación fuentes de energía 2015-2019.

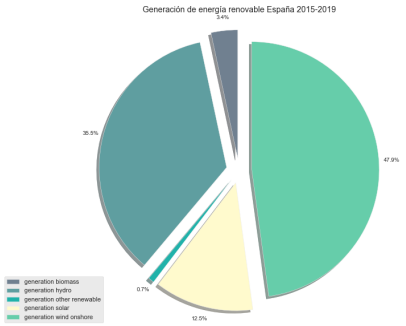


Figura A.4: Porcentaje de generación fuentes de energía no renovables 2015-2019.

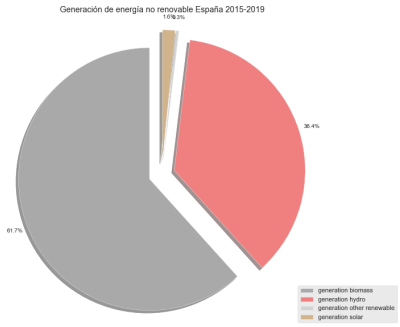


Figura A.5: Porcentaje de generación fuentes de energía renovables 2015-2019.

Volver al apartado 3.1.

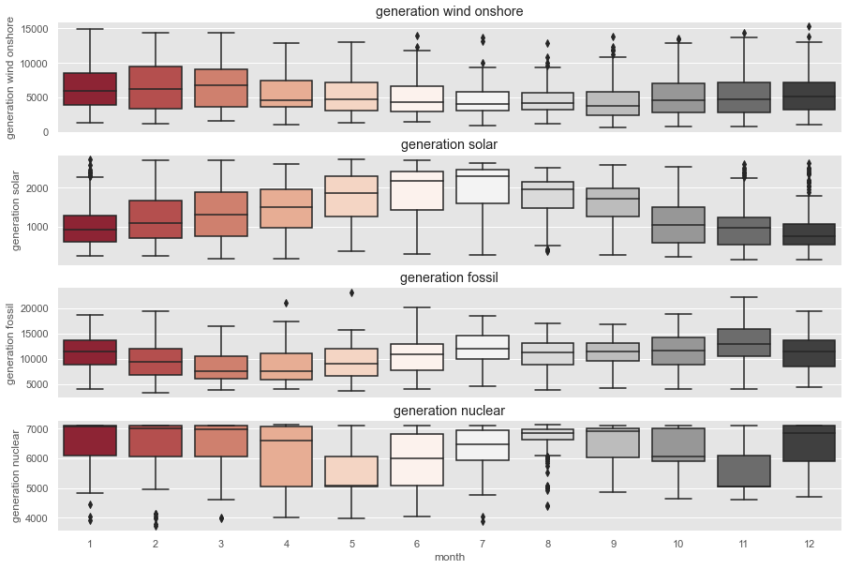


Figura A.6: Estacionalidad y generación de energía 2015-2019.

Volver al apartado 3.1.

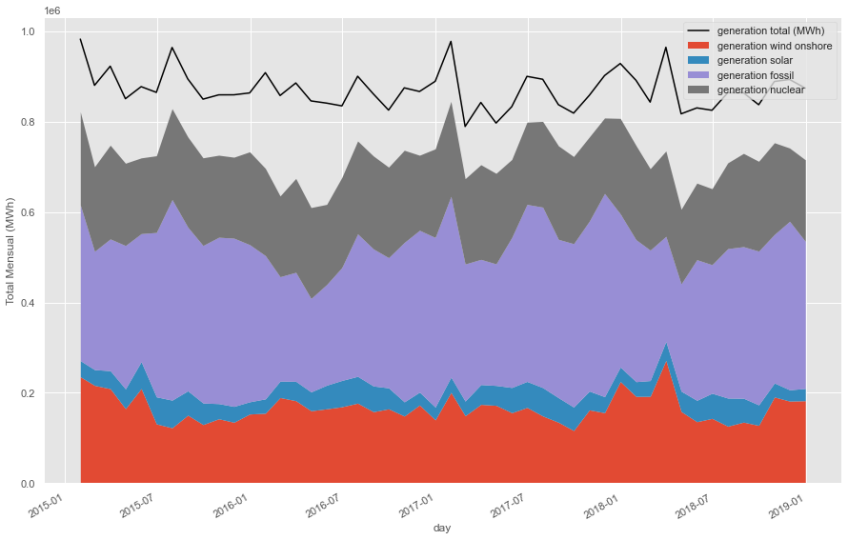


Figura A.7: Comparación generación total y el aporte acumulado de cada fuente de energía.

Volver al apartado 3.1.

A.2 | Energía generada vs energía demandada

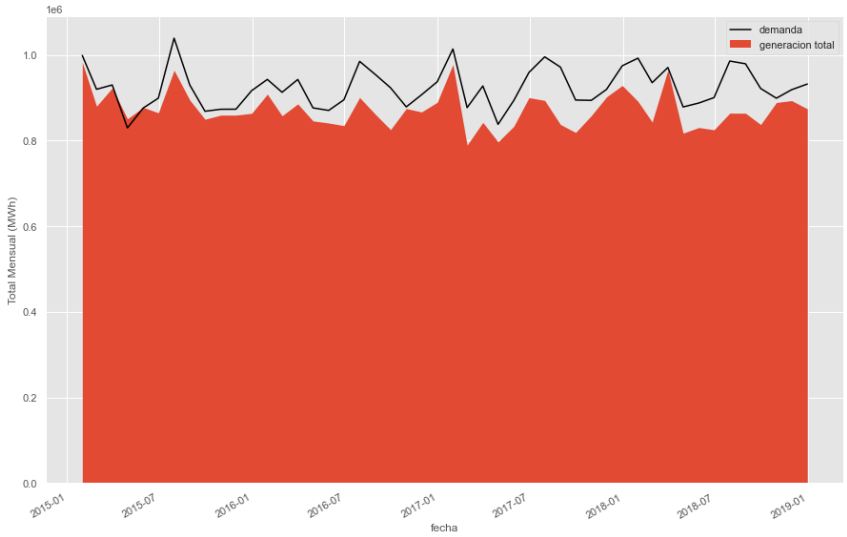


Figura A.8: Demanda vs Generación.

Volver al apartado 3.2.

A.3 | Eficiencia energética vs efectos adversos

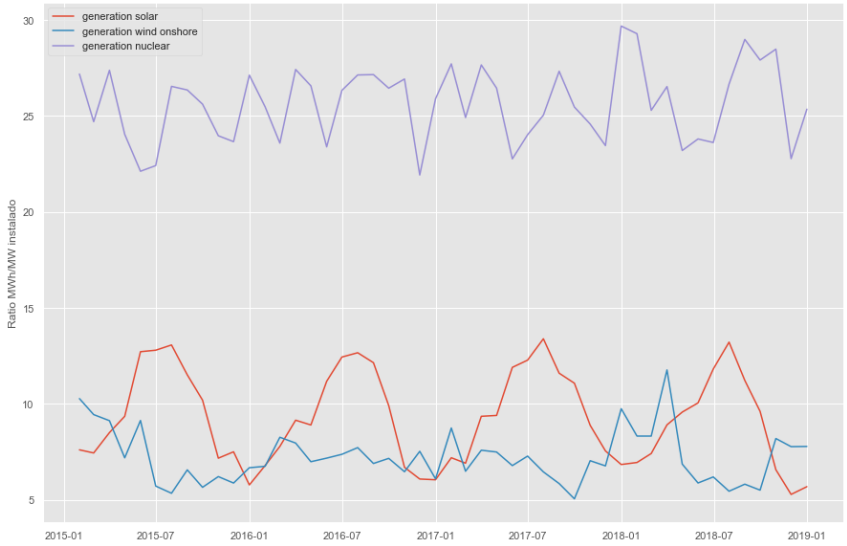


Figura A.9: Ratio MWh/ MW instalado.

Volver al apartado 3.3.

A.4 | Influencia del clima en la generación de energía

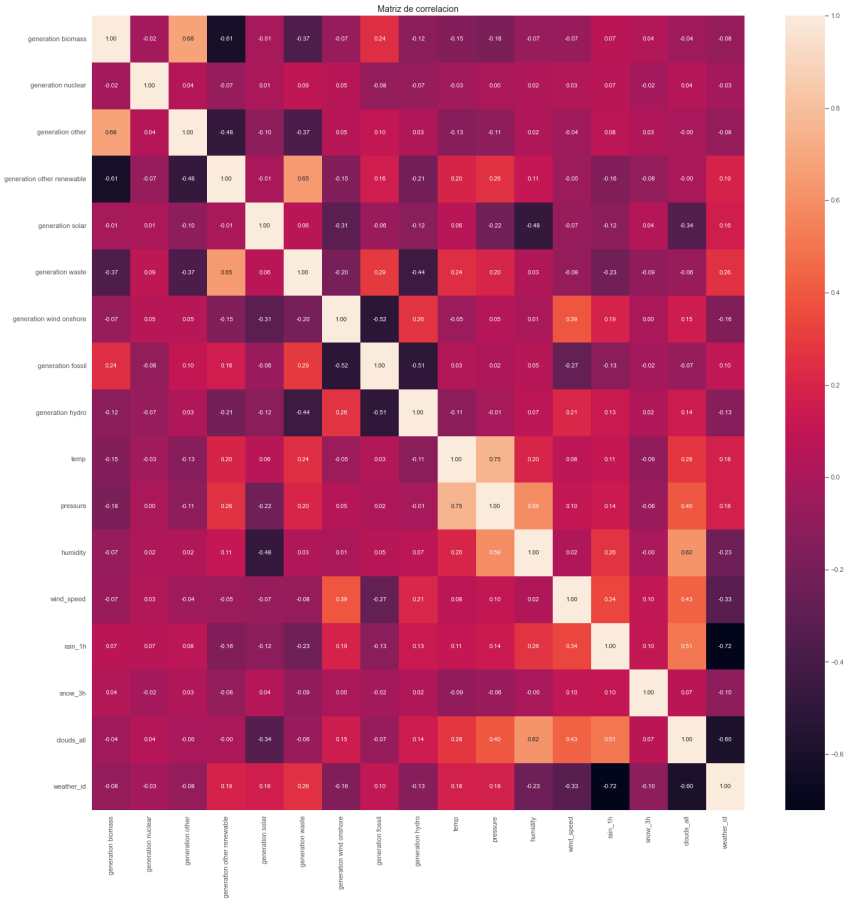


Figura A.10: Mapa de calor de correlación generación - variables climáticas.

Volver al apartado 3.4.

B | Apéndice B: Análisis estadístico

B.1 | Comprobaciones previas al análisis

Se hace necesaria la realización de un estudio previo para la posterior elección del estadístico más conveniente y que mejor se ajuste.

B.1.1 | Estudio de normalidad

Se pretende comprobar si la distribución que siguen las variables a evaluar con el estadístico correspondiente es una distribución normal. Lo más habitual es que no lo sea, como es el caso.

Se ha evaluado la normalidad mediante el test de D'Agostino's ya que es el que se emplea si la muestra no es pequeña.

El mayor problema de que no siga una distribución normal es que además se tengan pocos datos, dado que los resultados se ven afectados, pudiendo no ser un resultado fiable. Sin embargo, según el teorema central del límite, si la muestra de datos que se tiene es lo suficientemente grande los resultados no se ven afectados porque, aunque no se distribuyan como una normal, se asemejan a ella y los resultados no varían significativamente. Siendo esta último la situación de los datos de este proyecto.

B.1.2 | Estudio de homocedasticidad

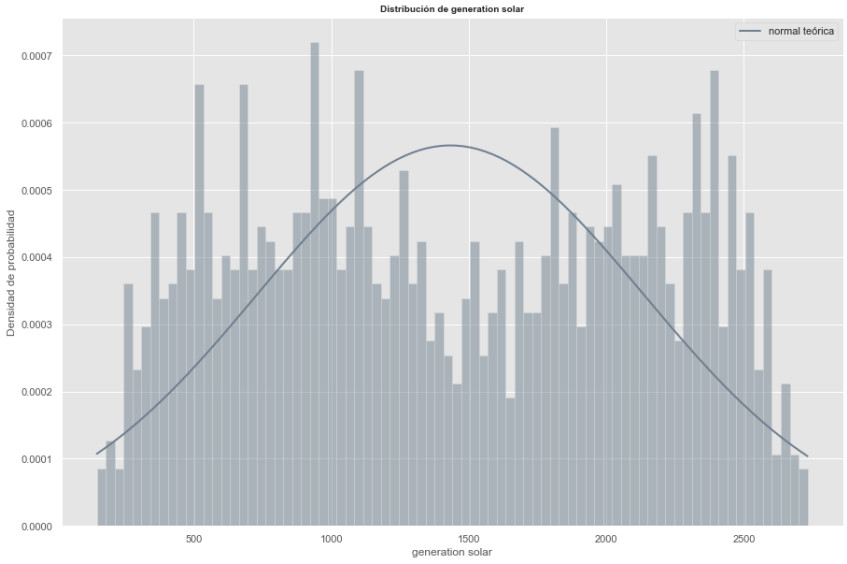
La homocedasticidad es que literalmente la dispersión es la misma, es decir, la varianza se mantiene. Se estudia la homocedasticidad con la fórmulas de Levene y Fligner. En ambos casos, el resultado de los datos comparados es heterocedácticos.

B.2 | Test estadísticos empleados

En el análisis estadístico realizado se han utilizado:

- Test ANOVA: que aunque entre sus condiciones está que los datos se distribuyan según una normal y la homocedasticidad, anota que si la muestra es lo suficientemente grande los resultados se aproximan bastante.
- Test KRUSKAL: es un test no paramétrico, recomendado para un caso como este.
- Test χ^2: ocurre algo similar a la ANOVA, en cuanto a sus condiciones de empleo.

B.3 | Gráficas



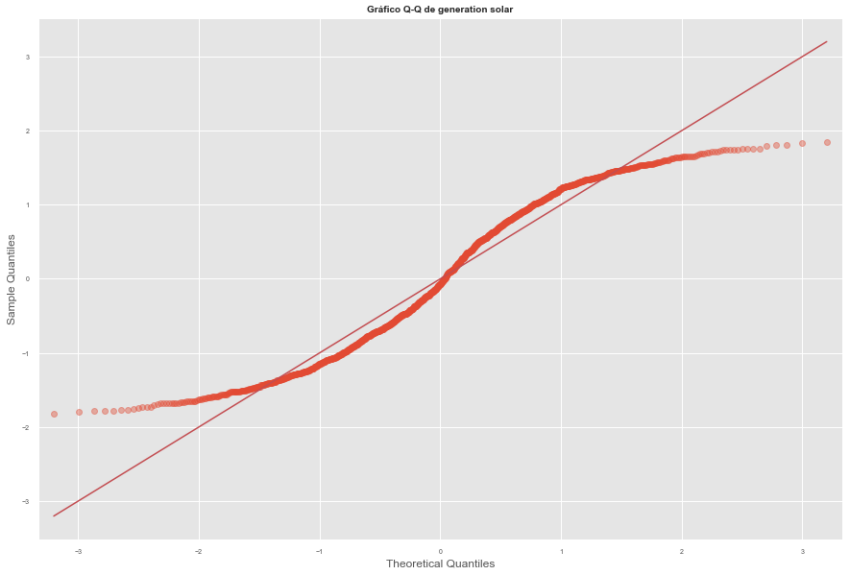
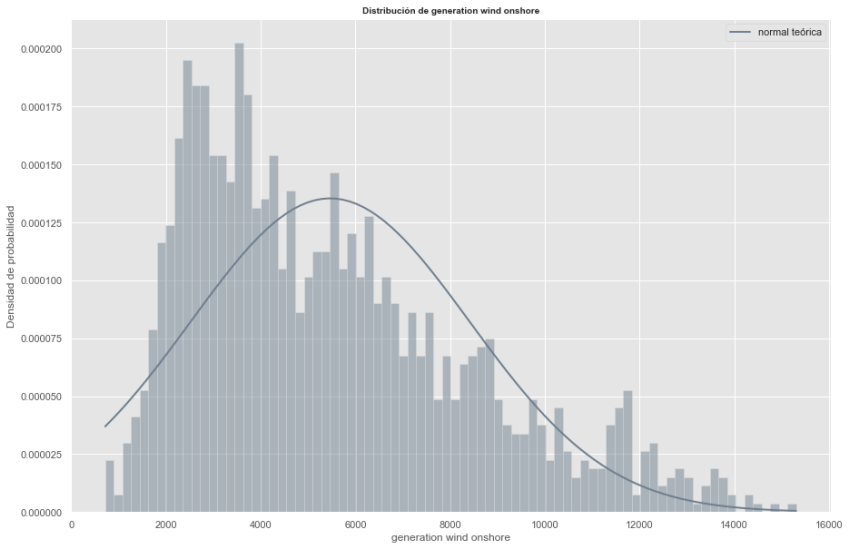


Gráfico Q-Q de generation solar
Sample Quantiles
Theoretical Quantiles


Distribución de generation wind onshore
normal teórica
Densidad de probabilidad
0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
0
2000
4000
6000
8000
10000
12000
14000
16000
generation wind onshore

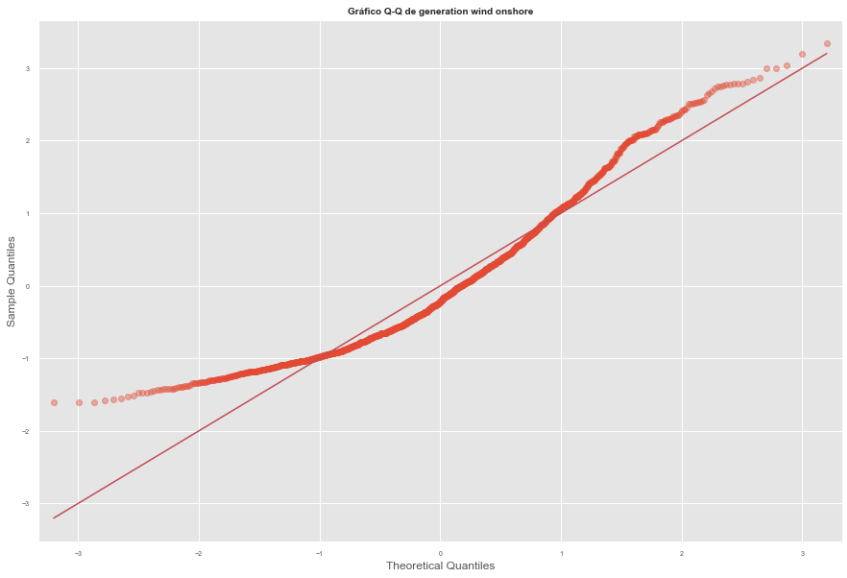
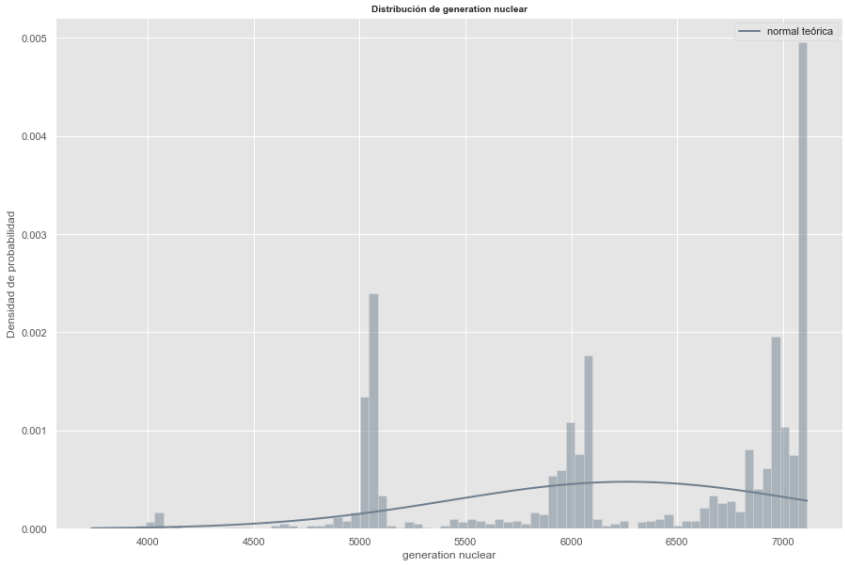


Gráfico Q-Q de generation wind onshore
Sample Quantiles
Theoretical Quantiles


Distribución de generation nuclear
normal teórica
Densidad de probabilidad
0.005
0.004
0.003
0.002
0.001
0.000
4000
4500
5000
5500
6000
6500
7000
generation nuclear

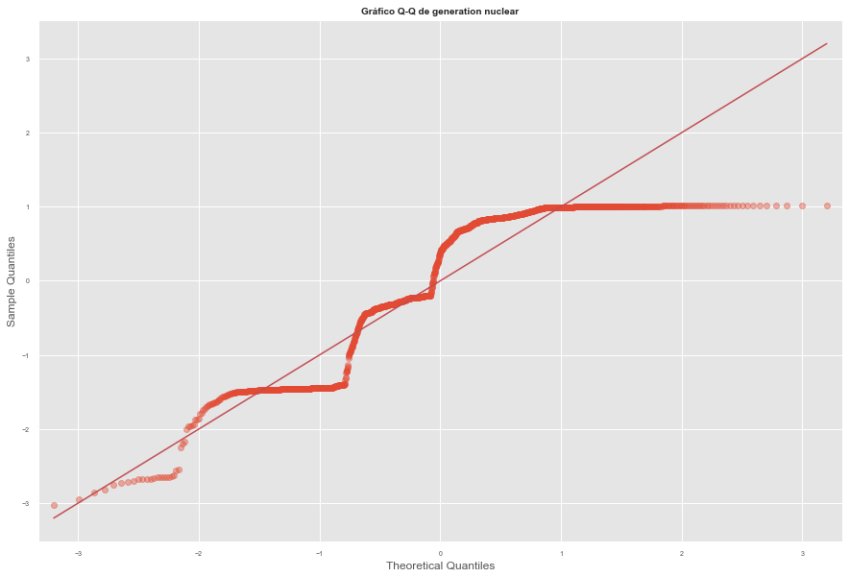


Gráfico Q-Q de generation nuclear
Sample Quantiles
Theoretical Quantiles