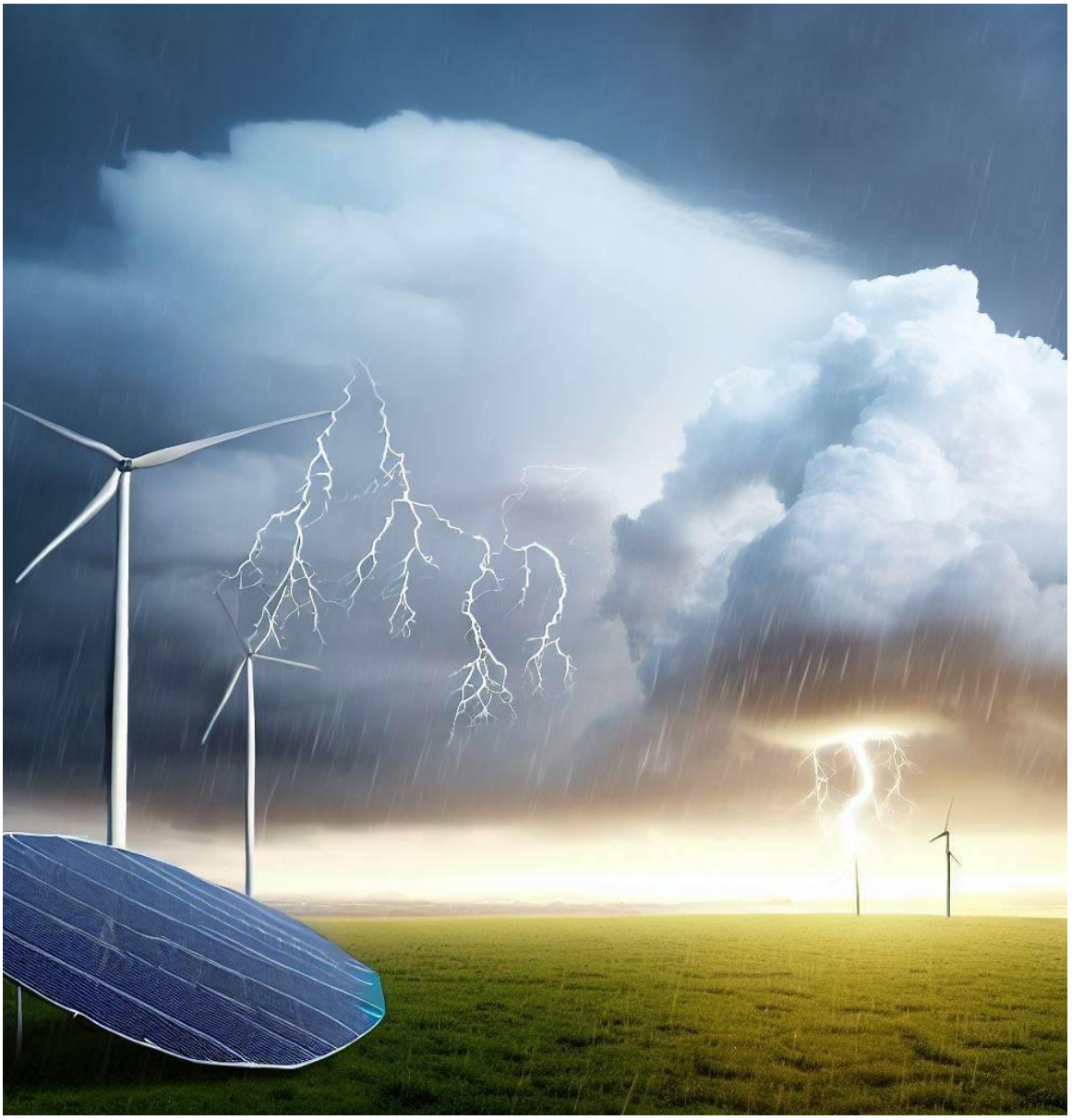
METEOROLOGIA I ENERGIES RENOVABLES



Nil Farrés Soler NIU: 1635864

Resum

Aquest treball presenta una anàlisi detallada de l'ús de dades meteorològiques per predir la màxima producció d'energia renovable diària. Utilitzant un model de regressió lineal múltiple, juntament amb tècniques avançades com l'ús de criteris AIC per a la selecció de models i mètodes bootstrap per obtenir estimacions més robustes, es demostra una capacitat predictiva significativa. L'estudi subratlla la importància de l'anàlisi de dades en el camp de l'energia renovable i destaca les oportunitats per millorar la gestió i la planificació de la producció d'energia. Amb resultats prometedors i recomanacions per a futures recerques, aquest treball ofereix una lectura essencial per a tots els interessats en l'àmbit de les energies renovables, la meteorologia i l'analítica de dades. Es vol aprofundir en la comprensió de com les dades meteorològiques poden ser utilitzades per pronosticar la producció d'energia, millorant així l'eficiència i la previsibilitat de l'aprofitament de les energies renovables.

ÍNDEX

1. Introducció

En el context actual de canvi climàtic, la necessitat de transitar cap a fonts d'energia més sostenibles mai no ha estat més urgent. L'augment de les temperatures globals, l'increment del nivell del mar i l'intensificació dels fenòmens meteorològics extrems posen de manifest que les conseqüències del canvi climàtic ja estan aquí i són inajornables. Això accentua la importància de la recerca en energies renovables, no només per a reduir les emissions de gasos d'efecte hivernacle, sinó també per a assegurar una provisió d'energia constant i sostenible per a les futures generacions.

L'estudi de l'energia renovable és un camp de recerca ric i divers, amb molts àmbits a explorar. Una àrea d'interès particular és la relació entre les condicions meteorològiques i la generació d'energia renovable. Aquesta relació és especialment rellevant per a la generació d'energia solar i eòlica, que depenen directament de les condicions climàtiques. Per exemple, com afecten les variacions en la radiació solar o en la velocitat del vent a la producció d'energia renovable? Com influeixen factors com la temperatura, la pressió atmosfèrica, la humitat o les precipitacions?

Aquest treball s'endinsa en aquestes preguntes explorant un conjunt de dades amb informació sobre la producció d'energia renovable i diverses variables meteorològiques.

L'**objectiu** principal és aprofundir en la comprensió de la relació entre les condicions meteorològiques i la generació d'energia renovable, prestant especial atenció a com les diferents variables meteorològiques afecten la producció d'energia. Per a això, s'utilitzen mètodes bootstrap paramètrics i no paramètrics, proporcionant una visió més completa i robusta d'aquesta relació.

L'esperança és que, amb aquesta anàlisi, no només es proporcioni una millor comprensió de la complexa interacció entre el clima i la generació d'energia renovable, sinó que també es predeixi la producció màxima d'energia renovable diària basada en les variables meteorològiques significatives. Alhora, planteja noves preguntes i possibilitats per a la recerca futura. Per exemple, com podríen les tendències de canvi climàtic afectar la producció d'energia renovable a llarg termini? Quines estratègies es poden desenvolupar per a mitigar aquests efectes i assegurar una provisió d'energia estable i sostenible?

2. Descripció del conjunt de dades

Per fer aquest treball, s'ha utilitzat el conjunt de dades "Renewable Energy and Weather Conditions". Aquest conté informació detallada sobre la producció d'energia renovable i una varietat de paràmetres meteorològics. Aquestes dades es recullen en intervals de 15 minuts, des de l'any 2017 fins l'any 2022, proporcionant una visió granular de com la producció d'energia i les condicions meteorològiques canvien al llarg del dia.

Els paràmetres meteorològics inclosos són la GHI (Irradiància horitzontal global en W/m² mesurada per un piranòmetre, que és la quantitat de radiació solar rebuda per una superfície horitzontal), la temperatura (en °C), la pressió atmosfèrica (en hPa), la humitat relativa (en %), la velocitat del vent (en m/s), la precipitació (pluja i neu en mm) i la cobertura de núvols. També s'inclouen dades sobre la durada de la llum del dia, el temps de llum solar disponible i una indicació binària de la presència de sol. En termes d'energia, el conjunt de dades proporciona la "Delta d'energia [Wh]", la quantitat d'energia renovable produïda en watts-hora (Wh). El conjunt de dades també inclou informació sobre l'hora i el mes de cada registre, permetent un anàlisi temporal de les dades.

El dataset s'ha obtingut de la pàgina web Kaggle, una plataforma reconeguda per proporcionar conjunts de dades de gran qualitat per a l'anàlisi i l'aprenentatge automàtic. La referència completa a aquesta font es pot trobar a la secció de bibliografia.

```
> colnames(data)
 [1] "Time"                "Energy.delta.Wh."    "GHI"                 "temp"                "pressure"
 [6] "humidity"            "wind_speed"          "rain_1h"             "snow_1h"             "clouds_all"
[11] "isSun"               "sunlightTime"        "dayLength"           "SunlightTime.daylength" "weather_type"
[16] "hour"                "month"
```

Amb l'objectiu de facilitar l'anàlisi, s'han realitzat alguns passos de pre-processament de les dades. En primer lloc, s'ha creat una columna de dates en el conjunt de dades original a partir de l'atribut de temps. Després, s'ha seleccionat el registre amb la màxima producció d'energia per cada dia. D'aquesta manera, només analitzarem i farem prediccions sobre les dades obtingudes quan l'energia renovable produïda és la màxima diària. Finalment, s'han eliminat les columnes que no es consideraven rellevants per a aquest anàlisi, incloent "isSun", "SunlightTime.daylength" i "weather_type". Més endavant també veurem que no ens interesserà tenir la variable GHI al model final.

Per tant, de moment, ens queden les dades a analitzar de la següent forma:

```
> print(data_max)
# A tibble: 2,070 × 14
# Groups:   date [2,050]
   date       Energy.delta.Wh.   GHI  temp pressure humidity wind_speed rain_1h snow_1h clouds_all sunlightTime dayLength  hour month
   <date>                <int> <dbl> <dbl>    <int>    <int>      <dbl>   <dbl>   <dbl>      <int>        <int>     <int> <int> <int>
 1 2017-01-01              161   4.7   3.8     1015       93        5.6       0       0         94          195       450    10     1
 2 2017-01-02             2115  45.1   2.1     1012       90        4.8       0       0         48          315       450    12     1
 3 2017-01-03              155   5.3   1.1     1012       97        5.7       0    0.23        100          210       450    10     1
 4 2017-01-04              458   9.2   2.7      991       83        9.8    0.71       0         83          255       465    11     1
 5 2017-01-10             2706  57.4  -2.2     1016       81        5.5       0       0         10          285       465    11     1
 6 2017-01-11             3198  51.1  -8.4     1009       76        8.2       0       0         54          165       465     9     1
 7 2017-01-12             1548   5.5   2.8     1001       85        7.7       0       0         72          255       465    11     1
 8 2017-01-13              168   5.7   0.1      985       98        3.1       0    0.77         97          195       465    10     1
 9 2017-01-14              244   6    -2.1     1000       98        1.9       0       0        100          135       480     9     1
10 2017-01-15              137   8.1  -0.4     1011       96        2.2       0       0         81          225       480    10     1
```

3. Descripció de les tècniques utilitzades

Per analitzar aquest conjunt de dades, utilitzarem una combinació de tècniques estadístiques i de ciència de dades.

- **Anàlisi descriptiva**

  Aquesta és la fase inicial de l'anàlisi de dades, on es resumeixen les principals característiques de les dades. Aquest anàlisi inclou càlculs de mesures de tendència central (com la mitjana i la mediana), mesures de dispersió (com la desviació estàndard i el rang interquartil, que indica com estan distribuïdes les dades al voltant de la mitjana) i la creació de gràfics descriptius (histogrames, gràfics de barres, “boxplots”, etc). L'objectiu és obtenir una comprensió inicial de la distribució i les relacions entre les diferents variables que hi intervenen.

  El nostre dataset té una gran varietat de variables, des de la temperatura fins a la humitat, la radiació solar i la generació d'energia. L'anàlisi descriptiva pot ajudar a comprendre la distribució de cadascuna d'aquestes variables, així com les relacions potencials entre elles.

- **Anàlisi de correlació**

  Aquesta tècnica es fa servir per mesurar la força i la direcció de la relació entre dues variables. El coeficient de correlació varia entre -1 i 1. Un valor de -1 indica una correlació negativa perfecta (a mesura que una variable augmenta, l'altra disminueix). En canvi, un valor de 1 indica una correlació positiva perfecta (a mesura que una variable augmenta, l'altra també ho fa). Finalment, un valor de 0 indica que no hi ha correlació. Si trobem una correlació forta, això podria indicar que aquestes dues variables estan relacionades d'alguna manera.

  En el nostre cas, aquest anàlisi ens pot ser útil, per exemple, per entendre si hi ha una correlació positiva entre la radiació solar (GHI) i la generació d'energia.

- **Anàlisi de regressió**

  Aquesta tècnica es fa servir per modelar la relació entre una variable dependent (o variable de resposta) i una o més variables independents (o variables predictores). L'objectiu és utilitzar aquestes relacions per predir valors futurs de la variable dependent.

  En el context del nostre estudi, la variable dependent és la generació d'energia, mentre que les variables independents són les diverses condicions meteorològiques, com la radiació solar, la temperatura, la pressió, la humitat, etc. L'objectiu principal és utilitzar aquestes relacions per predir la producció d'energia a partir de les condicions meteorològiques donades.

  Les tècniques de selecció de models com “forward selection” i “backward selection” són útils quan es treballa amb anàlisi de regressió, especialment quan es té un gran nombre de variables predictores.

  - **“Forward selection”:** aquesta tècnica comença amb un model sense variables predictores i afegeix una a una les variables que millor millorin la bondat d'ajust del model. El procés continua fins que quan afegim noves variables no millorem significativament la bondat d'ajust.

  - **“Backward selection”:** aquesta tècnica comença amb un model que inclou totes les variables predictores i elimina una a una les variables que menys contribueixen a la bondat d'ajust del model. El procés continua fins que l'eliminació de més variables empitjoraria significativament la bondat d'ajust.

  Aquestes tècniques ens permeten seleccionar les variables meteorològiques més significatives per a la predicció de la generació d'energia.

- **Mètodes bootstrap**

  El bootstrap és una tècnica de reamostreig (obté múltiples mostres a partir d'un conjunt de dades original) que es fa servir per estimar l'incertesa (l'error estàndard o els intervals de confiança) de les estimacions estadístiques. El bootstrap paramètric suposa que les dades segueixen una distribució coneguda, mentre que el bootstrap no paramètric no fa aquesta suposició. En ambdós casos, el procés bàsic és el mateix: es generen múltiples mostres de reamostreig de les dades originals (amb reposició, la qual cosa significa que cada observació pot ser seleccionada més d'una vegada en cada mostra), es calcula l'estadística d'interès en cada mostra de reamostreig, i es fa servir la distribució d'aquestes estadístiques de reamostreig per estimar l'incertesa de l'estimació original. En el nostre cas, s’ha aplicat aquest mètode per quantificar la incertesa associada a les nostres prediccions de la generació d'energia.

4. Resultats de l'anàlisi de les dades

Aquesta secció del treball es dedica a presentar els resultats de l'anàlisi de les dades. Aquí, es presenten les estadístiques descriptives, els resultats de l'anàlisi de correlació, els resultats de l'anàlisi de regressió i els resultats de l'anàlisi bootstrap.

Anàlisi descriptiva:

Primer de tot, s'ha obtingut un resum de les principals estadístiques de cada variable.

```
> summary(data_max)
      date            Energy.delta.Wh.      GHI              temp           pressure       humidity       wind_speed       rain_1h
 Min.   :2017-01-01   Min.   :  11     Min.   :  3.40   Min.   :-10.60   Min.   : 982   Min.   : 24.00   Min.   : 0.100   Min.   :0.00000
 1st Qu.:2018-06-04   1st Qu.:1280     1st Qu.: 44.33   1st Qu.:  5.00   1st Qu.:1010   1st Qu.: 59.00   1st Qu.: 3.000   1st Qu.:0.00000
 Median :2019-10-28   Median :3380     Median :105.20   Median : 11.75   Median :1016   Median : 73.00   Median : 4.200   Median :0.00000
 Mean   :2019-10-29   Mean   :2757     Mean   :105.49   Mean   : 11.99   Mean   :1016   Mean   : 71.72   Mean   : 4.358   Mean   :0.05753
 3rd Qu.:2021-03-20   3rd Qu.:4031     3rd Qu.:162.18   3rd Qu.: 18.80   3rd Qu.:1022   3rd Qu.: 86.00   3rd Qu.: 5.600   3rd Qu.:0.00000
 Max.   :2022-08-31   Max.   :5020     Max.   :229.10   Max.   : 33.60   Max.   :1046   Max.   :100.00   Max.   :12.000   Max.   :4.28000
    snow_1h           clouds_all      sunlightTime     dayLength          hour           month
 Min.   :0.000000   Min.   :  0.00   Min.   : 60.0   Min.   : 450.0   Min.   : 4.00   Min.   : 1.000
 1st Qu.:0.000000   1st Qu.: 39.00   1st Qu.:285.0   1st Qu.: 570.0   1st Qu.:10.00   1st Qu.: 3.000
 Median :0.000000   Median : 79.00   Median :375.0   Median : 765.0   Median :10.00   Median : 6.000
 Mean   :0.006589   Mean   : 66.97   Mean   :377.3   Mean   : 747.8   Mean   :10.43   Mean   : 6.293
 3rd Qu.:0.000000   3rd Qu.: 99.00   3rd Qu.:465.0   3rd Qu.: 930.0   3rd Qu.:11.00   3rd Qu.: 9.000
 Max.   :2.660000   Max.   :100.00   Max.   :780.0   Max.   :1020.0   Max.   :16.00   Max.   :12.000
```

Tinguem sempre en compte que les dades meteorològiques pertanyen al moment diària en el qual s'ha registrat la màxima energia renovable produïda.

Per exemple, podem observar el següent:

- **Energy.delta.Wh.**: Aquesta és la quantitat d'energia generada cada dia. La generació d'energia varia de 11 a 5020 Wh/dia, amb una mitjana de 2757 Wh/dia. El 50% de les dades (mediana) estan al voltant de 3380 Wh/dia.

- **GHI**: Aquesta és la Irradiància horitzontal global. El valor mínim és de 3.40W/m², mentre que el valor màxim és de 229.10W/m², amb una mitjana de 105.49W/m².

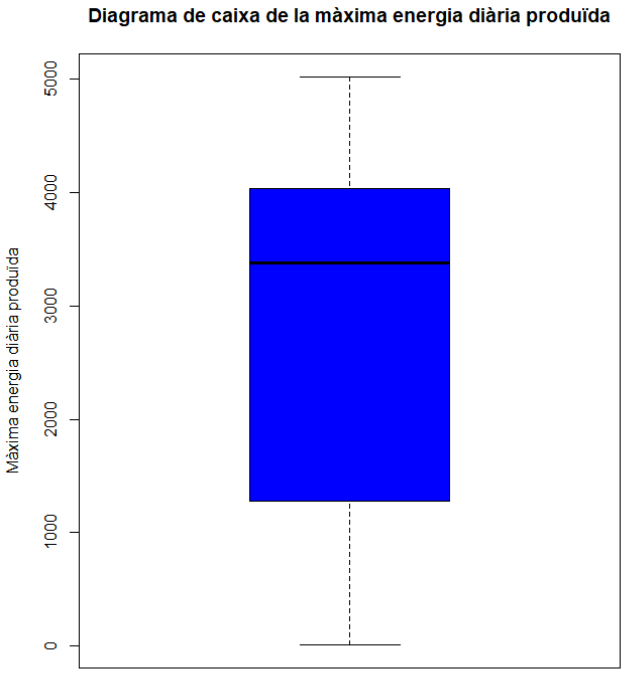
- **temp**: La temperatura va des de -10.6 fins a 33.6 graus Celsius, amb una mitjana d'11.99 graus.

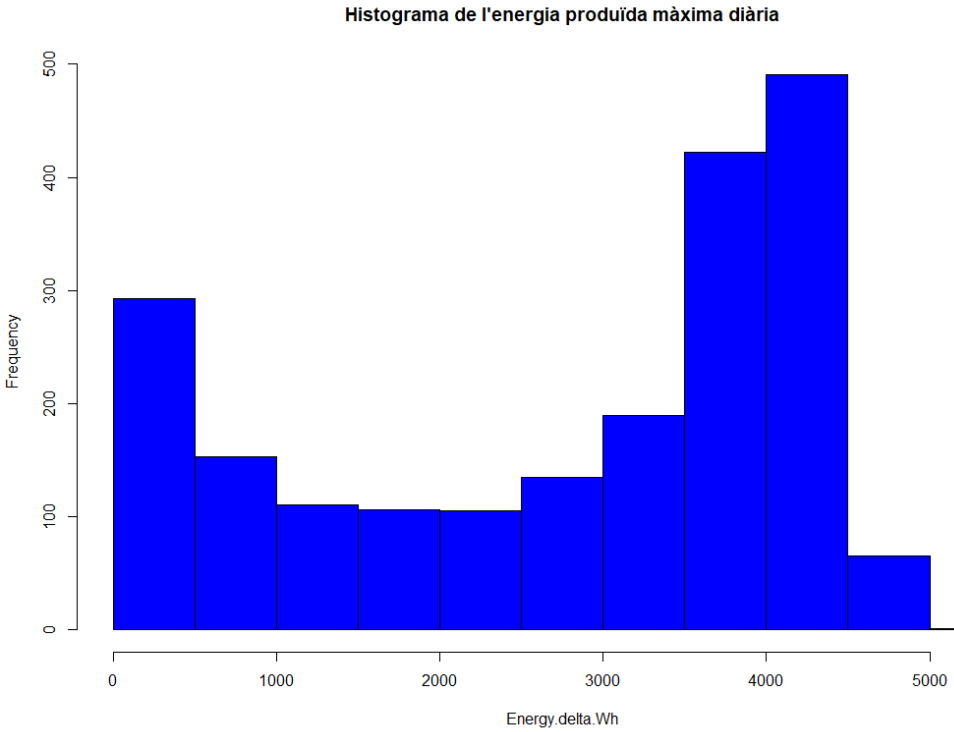
- **pressure**: La pressió atmosfèrica varia entre 982 i 1046 hPa, amb una mitjana de 1016 hPa.

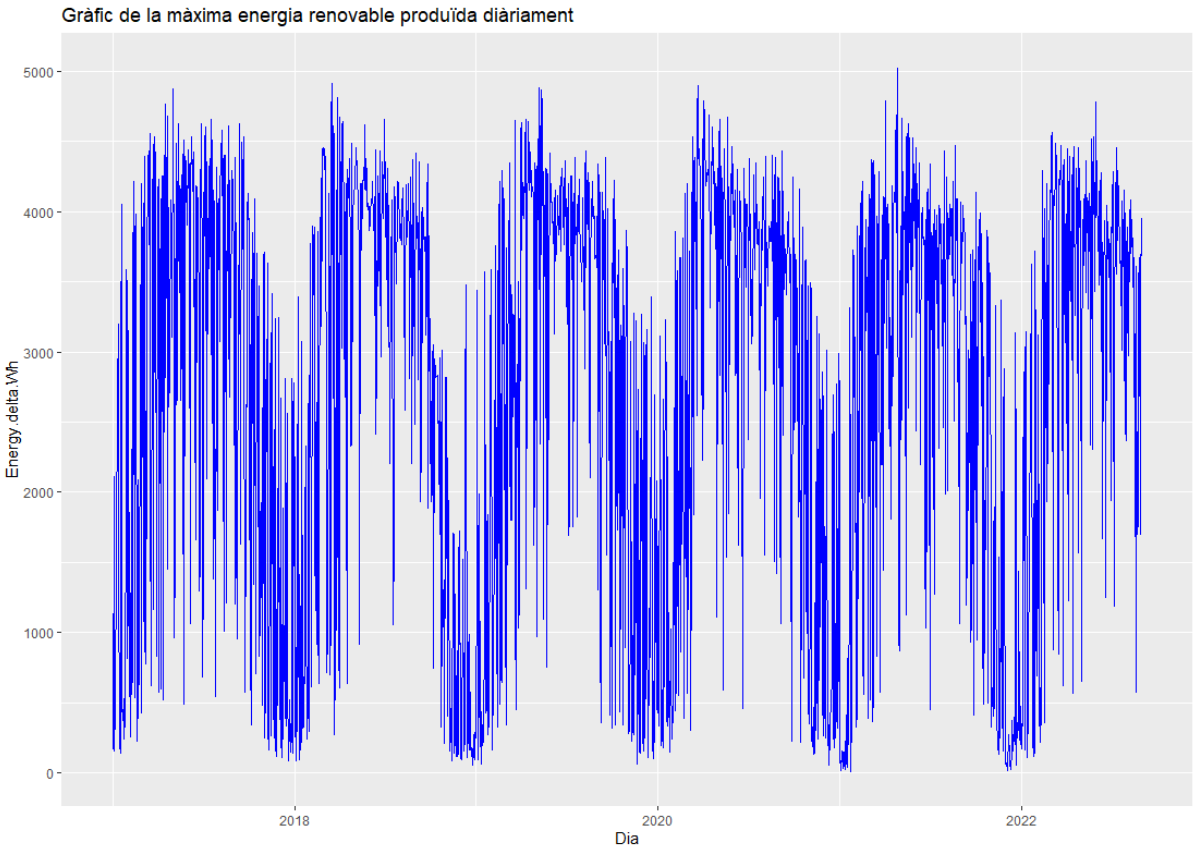
- **humidity**: El percentatge d'humitat varia entre el 24% i el 100%, amb una mitjana del 71.72%.

- **wind_speed**: La velocitat del vent va de 0.1 a 12 m/s, amb una mitjana de 4.358 m/s, que equivaldria a uns 15.7km/h.

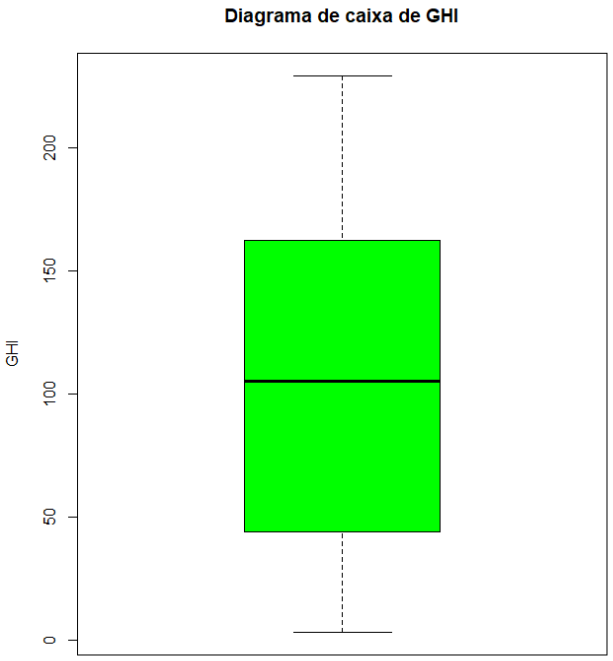
A continuació, s'ha realitzat diversos gràfics, histogrames i boxplots sobre les diferents variables. Analitzem els de més rellevància.

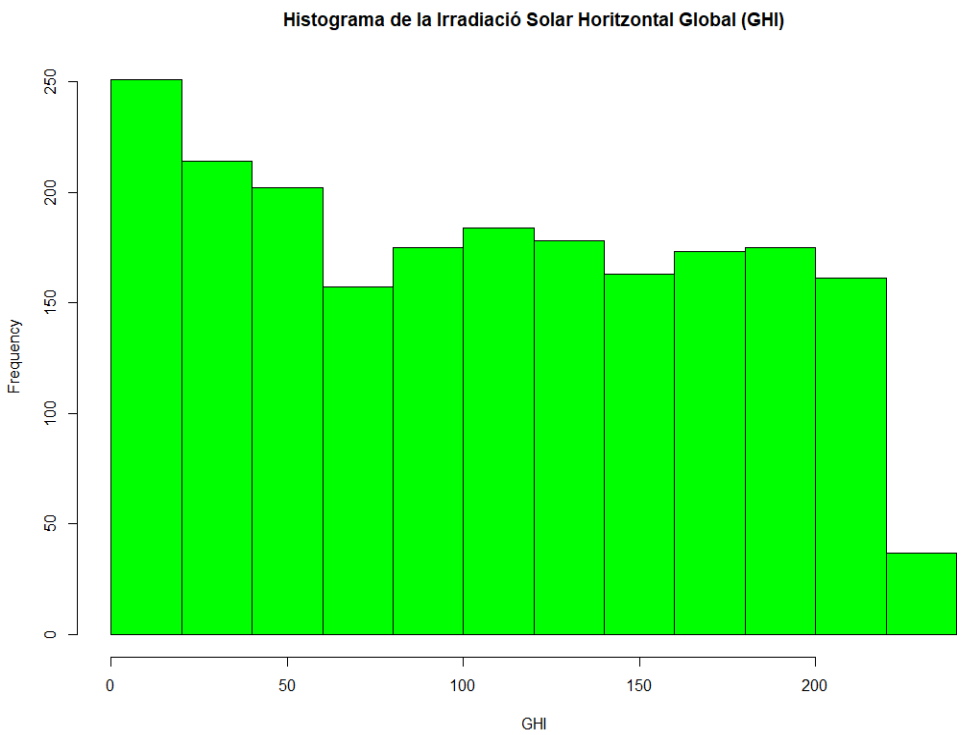


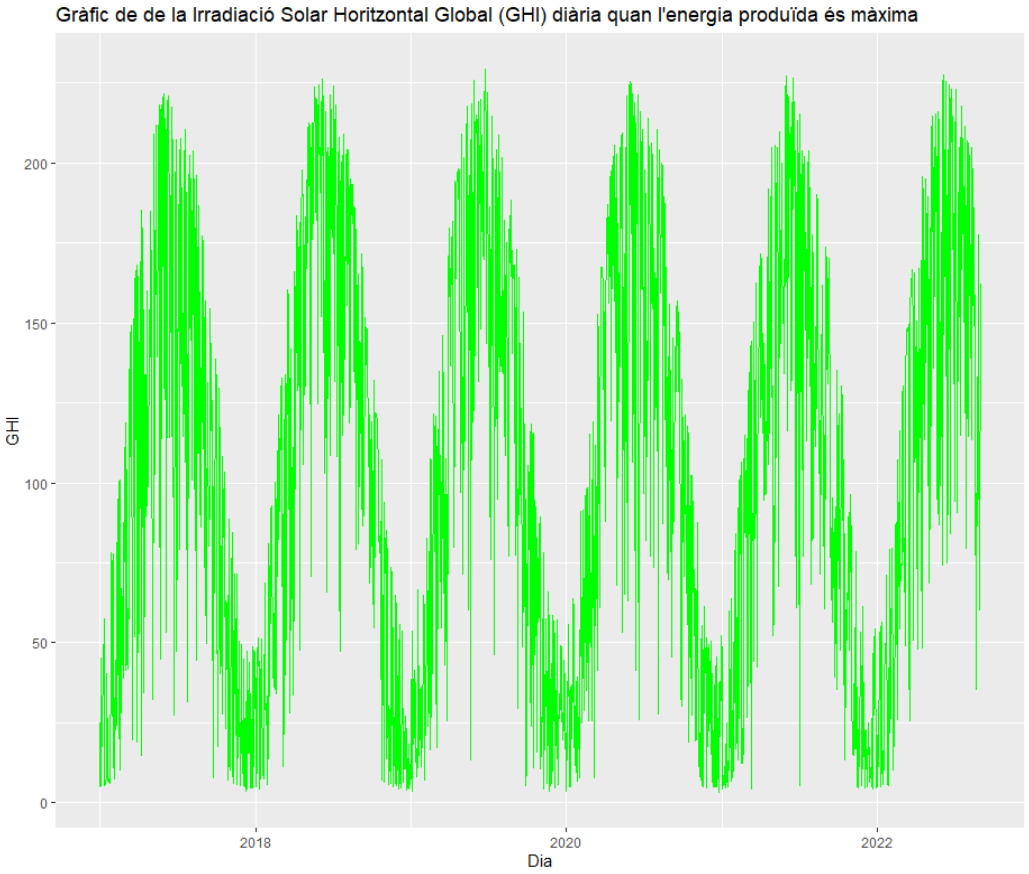




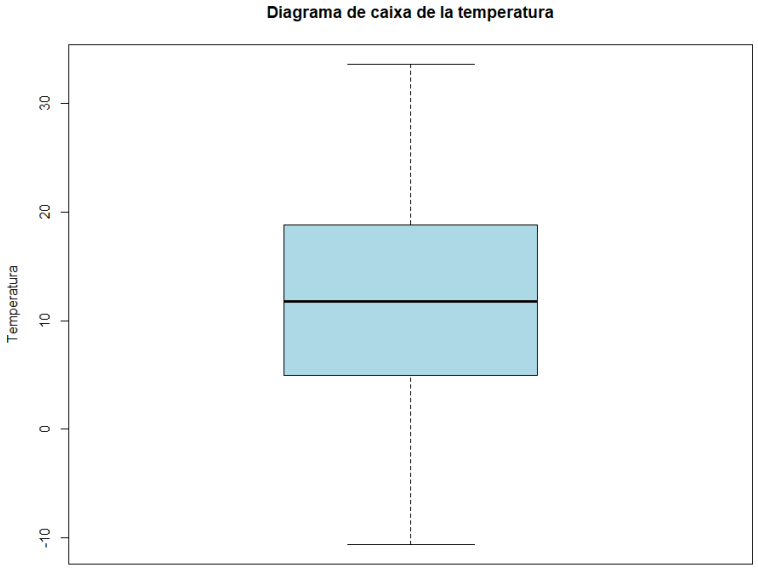
En aquests gràfics podem observar com, clarament, durant els mesos de primavera i estiu l'energia produïda creix notablement (arribant a màxims que s'acosten i superen en algun cas els 5000 Wh/dia) i decreix en els mesos de tardor i hivern (arribant, en alguns casos, a no produir energia durant tot el dia).

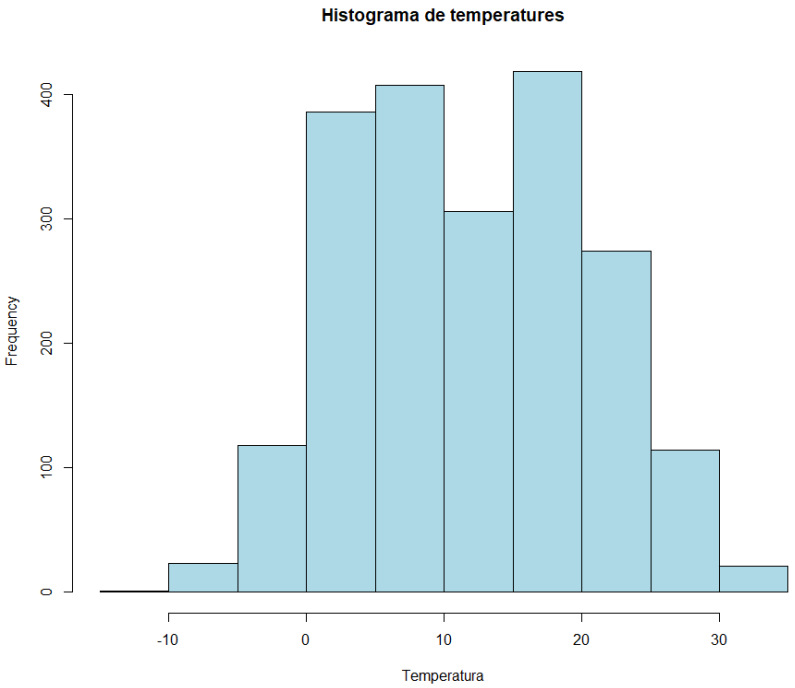


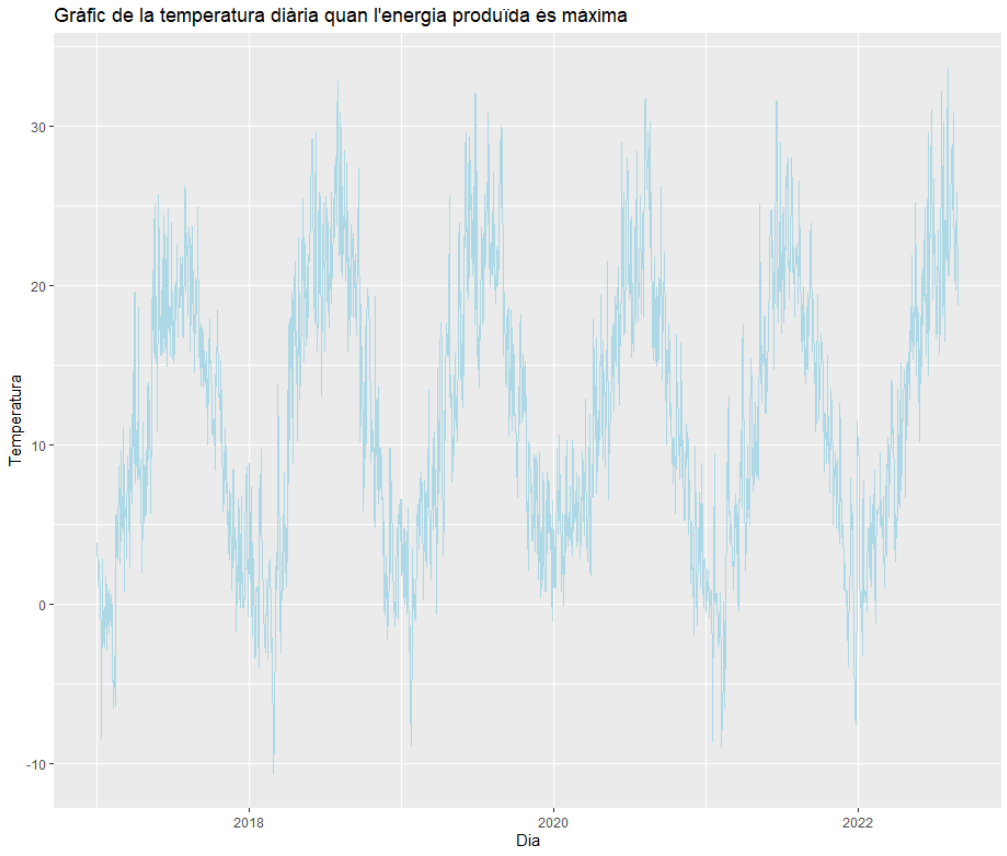




Podem observar, que la irradiació solar evoluciona de manera semblant a l'energia (en els mesos d'estiu creix a causa de la durada del dia). Per tant, segurament l'energia produïda depèn d'aquesta variable.

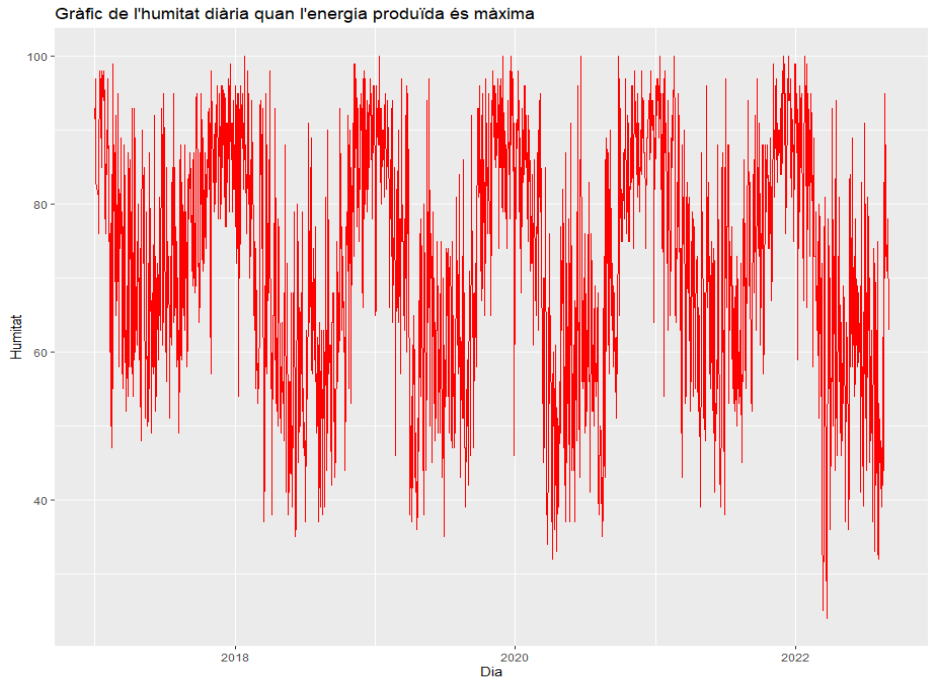






Com era d'esperar, la temperatura també es comporta similar que ens casos anteriors, cosa que fa pensar que hi ha una relació forta entre les variables temp, GHI i Energia.delta.Wh. De fet, això es pot observar en un gràfic de dispersió de les variables GHI i Energia.delta.Wh. Es veu clarament que, a mesura que GHI augmenta, hi ha una clara tendència de l'energia màxima produïda diària a augmentar també.

Un altre gràfic interessant podria ser el de la humitat relativa.



Es pot veure com la humitat relativa disminueix en els mesos més calorosos i augmenta a la tardor arribant als seus màxims a l'hivern. Per tant, sembla que quan l'energia màxima augmenta, la humitat disminueix.

Altres gràfics, com els de la velocitat del vent o de la pressió atmosfèrica, que acostumen a ser menys visualment clars es poden executar amb el codi de l'Script en R de l'Annex.

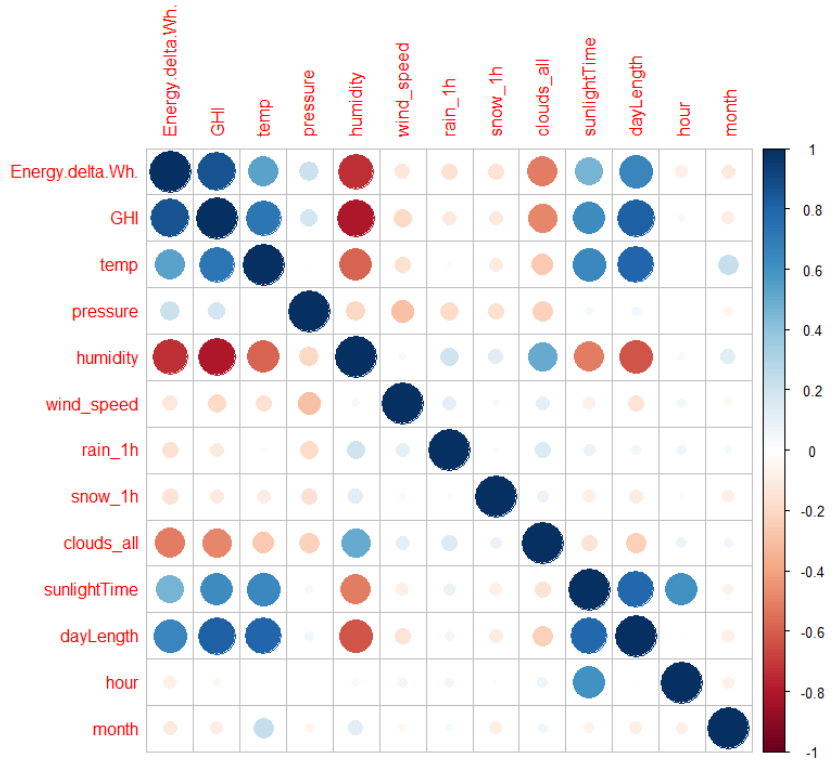
Anàlisi de correlació:

Per fer aquest anàlisi, s'ha calculat, amb la funció *cor*, la correlació entre totes les parelles de columnes de les nostres dades numèriques. El resultat és una matriu que conté els coeficients de correlació per a cada parella de columnes. A més, s'ha utilitzat la funció *round* per arrodonir cada coeficient de correlació a dos decimals per tal de facilitar-ne la lectura.

La matriu de correlació obtinguda és la següent:

```
> #Agafem només les dades numèriques.
> numeric_data <- data_max[, sapply(data_max, is.numeric)]
> #Construim la matriu de correlació
> correlation_matrix <- cor(numeric_data)
> # Arrodonim la matriu de correlació a dos decimals
> correlation_matrix <- round(correlation_matrix, 2)
> # Mostra la matriu de correlació
> print(correlation_matrix)
                Energy.delta.Wh.   GHI  temp pressure humidity wind_speed rain_1h snow_1h clouds_all sunlightTime dayLength  hour month
Energy.delta.Wh.             1.00  0.86  0.53     0.21    -0.74      -0.13   -0.16   -0.15      -0.52         0.46      0.65 -0.09 -0.12
GHI                          0.86  1.00  0.72     0.18    -0.80      -0.20   -0.12   -0.12      -0.49         0.62      0.82 -0.04 -0.10
temp                         0.53  0.72  1.00    -0.01    -0.59      -0.16    0.02   -0.11      -0.27         0.64      0.79 -0.01  0.23
pressure                     0.21  0.18 -0.01     1.00    -0.21      -0.30   -0.19   -0.16      -0.23         0.04      0.05 -0.01 -0.06
humidity                    -0.74 -0.80 -0.59    -0.21     1.00       0.04    0.20    0.13       0.50        -0.52     -0.63 -0.04  0.13
wind_speed                  -0.13 -0.20 -0.16    -0.30     0.04       1.00    0.11    0.03       0.11        -0.09     -0.15  0.05 -0.04
rain_1h                     -0.16 -0.12  0.02    -0.19     0.20       0.11    1.00   -0.02       0.15         0.08      0.05  0.06  0.04
snow_1h                     -0.15 -0.12 -0.11    -0.16     0.13       0.03   -0.02    1.00       0.08        -0.09     -0.11  0.01 -0.09
clouds_all                  -0.52 -0.49 -0.27    -0.23     0.50       0.11    0.15    0.08       1.00        -0.15     -0.24  0.07  0.06
sunlightTime                 0.46  0.62  0.64     0.04    -0.52      -0.09    0.08   -0.09      -0.15         1.00      0.78  0.60 -0.07
dayLength                    0.65  0.82  0.79     0.05    -0.63      -0.15    0.05   -0.11      -0.24         0.78      1.00 -0.01 -0.09
hour                        -0.09 -0.04 -0.01    -0.01    -0.04       0.05    0.06    0.01       0.07         0.60     -0.01  1.00 -0.08
month                       -0.12 -0.10  0.23    -0.06     0.13      -0.04    0.04   -0.09       0.06        -0.07     -0.09 -0.08  1.00
```

Finalment, s'ha utilitzat la llibreria *corrplot* per tal de visualitzar la matriu de correlació a través d'un gràfic on cada cel·la de la matriu és representada com un cercle, amb àrea i color representant el valor de cada coeficient de correlació (com més blau fosc, correlació més alta).



Podem observar que les correlacions més significatives són:

- **Energy.delta.Wh. i GHI:** Tenen una correlació alta, el que indica que quan l'índex de radiació solar global augmenta, l'energia produïda també ho fa. Això té sentit, ja que més radiació solar generalment significa més energia solar produïda.

- **Energy.delta.Wh. i humidity:** Tenen una correlació negativa alta, el que indica que quan la humitat augmenta, l'energia produïda disminueix. Això podria ser degut al fet que els dies amb alta humitat solen estar més ennuvolats, cosa que fa que la producció d'energia solar disminueixi.

- **Energy.delta.Wh. i dayLength:** Tenen una correlació moderada, el que indica que quan la longitud del dia augmenta, l'energia produïda també ho fa. Això té sentit, ja que un dia més llarg proporciona més hores de llum solar per a la producció d'energia.

- **Energy.delta.Wh. i temp:** Tenen una correlació moderada, el que indica que quan la temperatura augmenta, l'energia produïda també ho fa. Això podria ser degut a que un dia més calent pot correspondre a un dia assolellat, cosa que augmentaria la producció d'energia solar.

- **GHI i temp:** Tenen una correlació alta, el que indica que quan l'índex de radiació solar global augmenta, la temperatura també ho fa. Això té sentit, ja que un índex de radiació solar més alt generalment significa més calor.

- **GHI i humidity:** Tenen una correlació negativa alta, el que indica que quan l'índex de radiació solar global augmenta, la humitat disminueix. Això podria ser degut al fet que en dies assolellats la humitat acostuma a ser més baixa.

S'observa que GHI té una correlació molt alta amb l'energia renovable produïda. Per aquest motiu, no tindrem en compte la variable GHI en el model final i ens quedarem amb la variable de temperatura, que també també té una correlació alta i afecta de manera semblant a la màxima energia produïda que GHI.

Anàlisi de regressió:

Primer de tot, s'ha carregat la llibreria *MASS*. Després, amb la funció lm (Linear Model), s'ha creat un model de regressió lineal que utilitza totes les variables disponibles en les dades (*Energy.delta.Wh. ~ .*). S'utilitza *summary* per obtenir una descripció detallada del model. Seguidament, s'utilitza la funció *stepAIC* per realitzar tres tècniques de selecció de models:

- ***Backward selection:*** es comença amb un model que conté totes les variables i, en cada pas, s'elimina la variable que menys contribueix al model (la que té el valor *AIC* més petit una vegada eliminada).

- ***Forward selection:*** es comença amb el model nul (que no conté cap predictor) i, a cada pas, s'afegeix la variable que més contribueix al model (la que té el valor *AIC* més petit una vegada afegida).

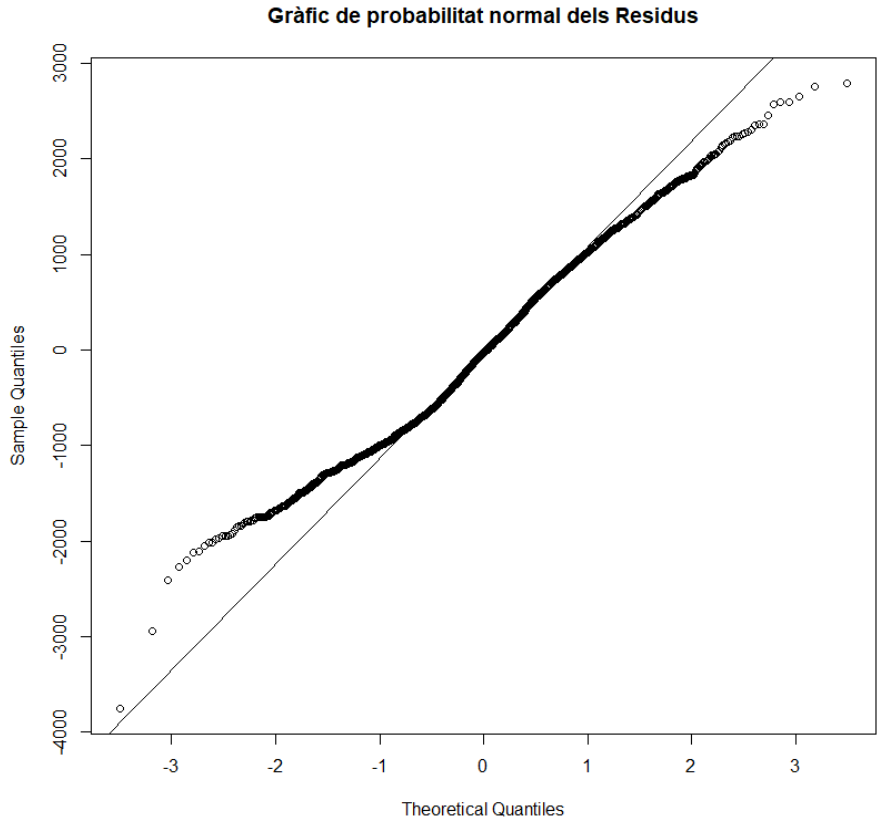
- ***Stepwise selection:*** és una combinació de *forward i backward selection*. Es comença amb el model complet i, a cada pas, s'afegeix o s'elimina la variable que millora més el model.

En els tres casos, el model final al qual arribem després d'haver eliminat la variable GHI pel que hem comentat abans, és el següent (tenint en compte el valor *AIC* i no el *p-value*):

```
> #Creem un model final amb les variables seleccionades
> model_final <- lm(Energy.delta.Wh. ~ date + temp + pressure +
+                     humidity + wind_speed + rain_1h + snow_1h + clouds_all,
+                   data = data_max)
> # Resum del model
> summary(model_final)

Call:
lm(formula = Energy.delta.Wh. ~ date + temp + pressure + humidity +
    wind_speed + rain_1h + snow_1h + clouds_all, data = data_max)

Residuals:
    Min      1Q  Median      3Q     Max
-3753.7  -771.4   -27.8   726.7  2794.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 2536.49257 2602.54785   0.975  0.32986
date          -0.10561    0.03555  -2.970  0.00301 **
temp          27.40917    3.20578   8.550  < 2e-16 ***
pressure       5.96454    2.43952   2.445  0.01457 *
humidity     -48.97644    1.85771 -26.364  < 2e-16 ***
wind_speed   -37.55283   12.00574  -3.128  0.00179 **
rain_1h      -73.27124   91.65726  -0.799  0.42415
snow_1h     -758.00580  280.30845  -2.704  0.00690 **
clouds_all    -8.34211    0.71095 -11.734  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 952.7 on 2061 degrees of freedom
Multiple R-squared:  0.5985,    Adjusted R-squared:  0.597
F-statistic: 384.1 on 8 and 2061 DF,  p-value: < 2.2e-16
```

Podem veure, per exemple, que per cada unitat d'increment en *temp*, *Energy.delta.Wh.* augmenta en 27.4 unitats, mantenint constant la resta de variables. De manera similar, per cada unitat d'increment en *humidity*, *Energy.delta.Wh.* disminueix en 48.9 unitats, mantenint constant la resta de variables. El valor R-quadrat multiple és de 0.597, això vol dir que aproximadament el 59.7% de la variabilitat de Energy.delta.Wh. es pot explicar per aquest model de regressió lineal. Per acabar, el valor de l'estadístic F és de 384.1 amb un p-valor extremadament petit (< 2.2e-16). Això indica que almenys una de les variables independents té un efecte significatiu sobre Energy.delta.Wh.

A més a més, també s'han calculat els residus del model (diferència entre els valors observats i els valors predits) amb la funció *resid( )* i s'ha generat un gràfic de probabilitat normal de residus amb la funció *qqnorm( )*. En aquest gràfic, els residus es representen en l'eix vertical i els quantils d'una distribució normal estàndard es representen en l'eix horitzontal. També s'ha utilitzat la funció *qqline( )* per afegir una línia al gràfic de probabilitat normal que representa el lloc on caurien els punts si els residus segueixen exactament una distribució normal.



Com que els punts cauen, majoritàriament, a prop de la línia, podem concloure que els residus segueixen aproximadament una distribució normal.

Tot el codi en R utilitzat es pot visualitzar a l'Annex.

Mètodes bootstrap:

Primer de tot, s'han generat valors ajustats (es generen a partir del model de regressió final ajustat a les dades originals) i residus (diferència entre les observacions originals i els valors ajustats). Després s'ha generat una nova resposta (o variable dependent) afegint un residu re-mostrejat aleatòriament als valors ajustats, s'han creat noves dades substituint la variable dependent original per la nova resposta generada i s'ha fet l'ajust d'un model de regressió a aquestes noves dades.

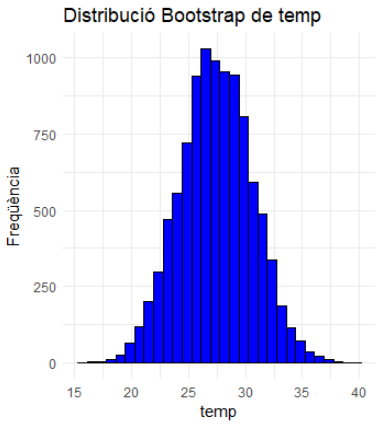
- **Bootstrap paramètric:** es crea un bucle on es repeteix el procés de generació de noves respostes, creació de noves dades i ajustament del model 10000 vegades. Per a cada iteració, s'emmagatzemen les estimacions dels paràmetres del model ajustat a les noves dades.

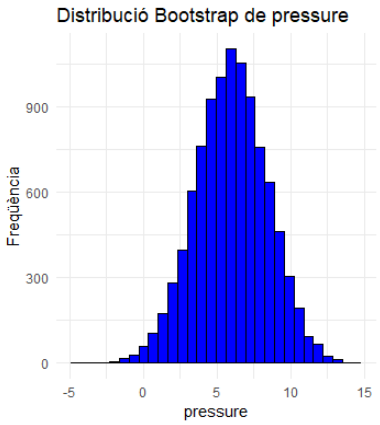
```
> # Inicialitzem una matriu per emmagatzemar les estimacions dels paràmetres
> param_estimates_bparam <- matrix(nrow = 10000, ncol = length(coef(model_final)))
> # Bucle Bootstrap paramètric
> for(i in 1:10000) {
+
+   # Generem una nova resposta
+   new_response <- fitted_values + sample(resids, replace = TRUE)
+
+
+   # Actualitzem les dades
+   new_data$Energy.delta.wh. <- new_response
+
+   # Ajustem el model a les noves dades
+   new_fit <- lm(Energy.delta.wh. ~ date + temp + pressure +
+                   humidity + wind_speed + rain_1h + snow_1h + clouds_all,
+                 data = new_data)
+
+   # Emmagatzemem les estimacions dels paràmetres
+   param_estimates_bparam[i, ] <- coef(new_fit)
+
+ }
```

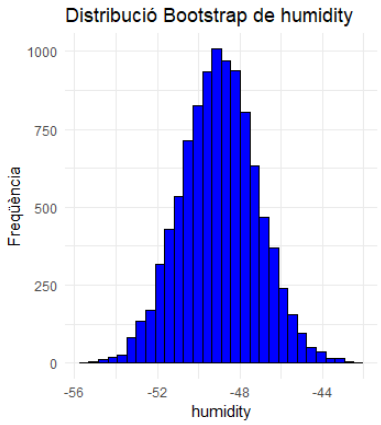
- **Bootstrap no paramètric:** es crea un bucle on, en lloc de generar una nova resposta, es re-mostreja directament les observacions originals. Això es fa seleccionant aleatòriament files de les dades originals amb reposició. A continuació, es fa l'ajust d'un model de regressió a aquestes dades re-mostrejades i s'emmagatzemen les estimacions dels paràmetres.

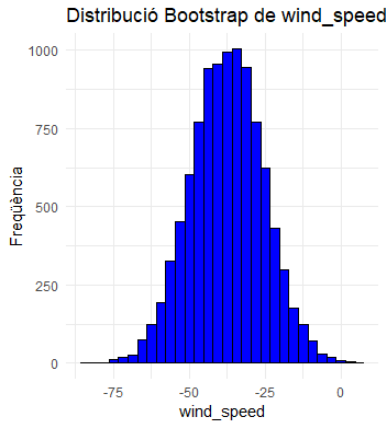
```
> # Initialize a matrix to store parameter estimates
> param_estimates_bnoparam <- matrix(nrow = 10000,
+                                    ncol = length(coef(model_final)))
> # Bootstrap loop
> for(i in 1:10000) {
+
+   # Resample data
+   new_data <- data_max[sample(nrow(data_max), replace = TRUE), ]
+
+   # Fit the model to the new data
+   new_fit <- lm(Energy.delta.wh. ~ date + temp + pressure +
+                   humidity + wind_speed + rain_1h + snow_1h + clouds_all,
+                 data = new_data)
+
+   # Store parameter estimates
+   param_estimates_bnoparam[i, ] <- coef(new_fit)
+
+ }
```

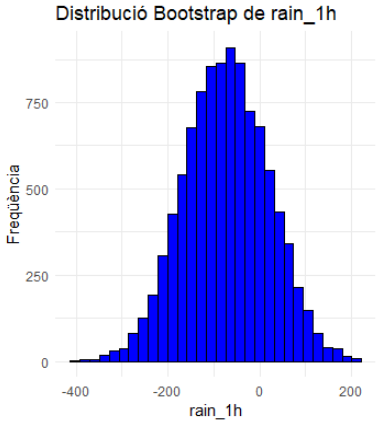
Ara ja es poden generar histogrames per a cada paràmetre estimat de les iteracions bootstrap. El propòsit d'aquests histogrames és visualitzar la distribució de les estimacions de paràmetres obtingudes de la simulació bootstrap. En altres paraules, aquests histogrames mostren com variarien les estimacions de paràmetres si repetissim l'anàlisi amb múltiples mostres de dades. Podem veure, per exemple, que la temperatura varia entre 15 i 37 ºC, o que la pressió atmosfèrica varia entre -3 i 15 hPa. La humitat, en canvi, varia entre valors negatius ja que, com hem vist abans, el seu coeficient en el model final és negatiu (si la humitat augmenta, la producció d'energia renovable disminueix).

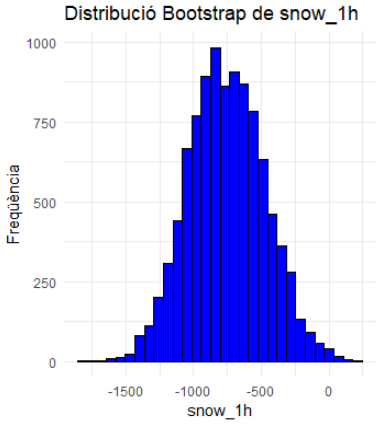


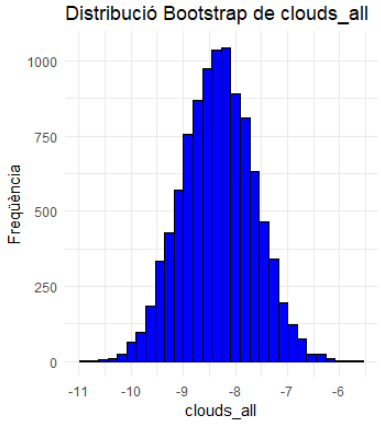




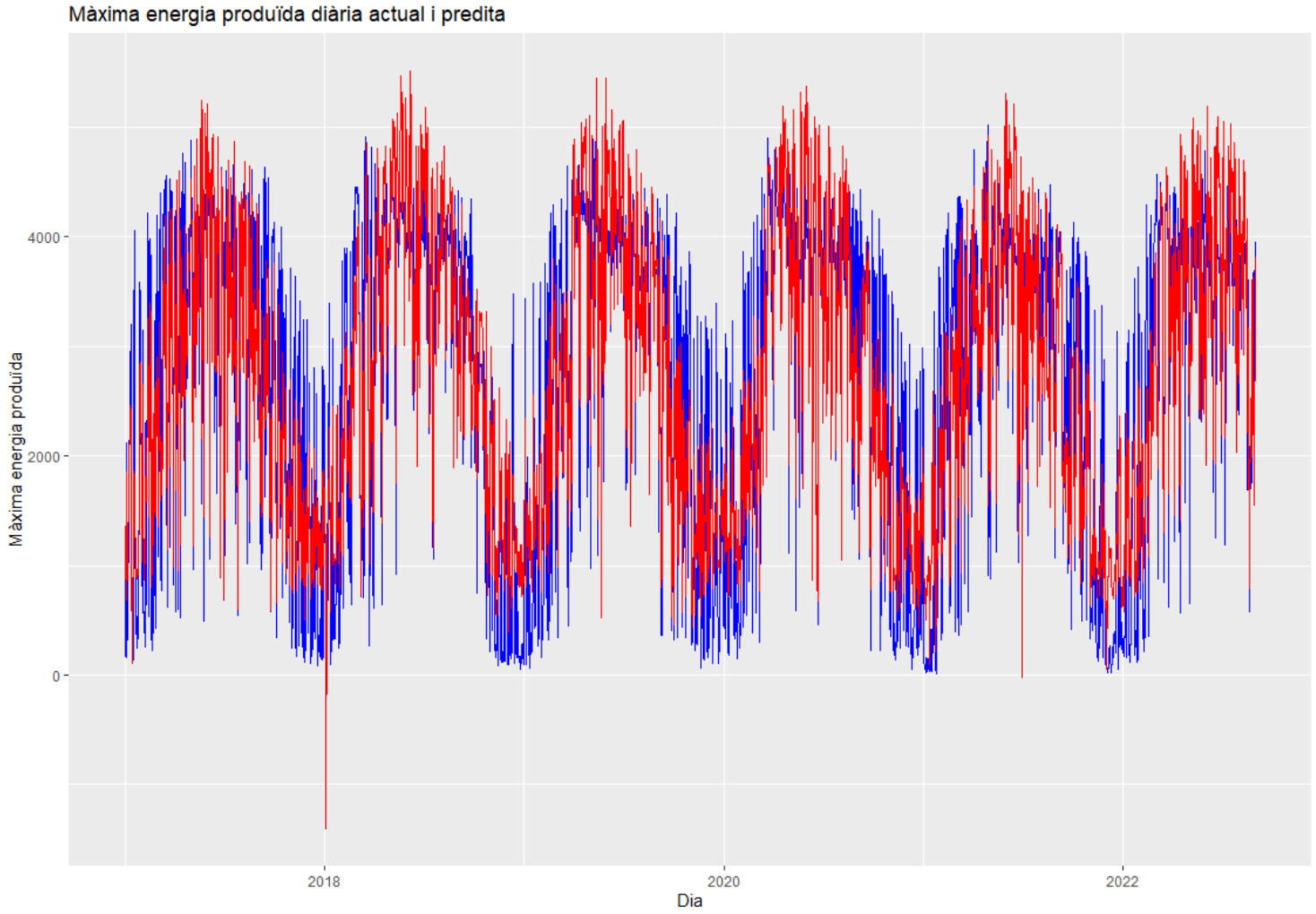








Finalment, s'ha generat un gràfic que mostra tant les dades originals de la producció d'energia (en color blau) com les prediccions del model per la producció d'energia al llarg del temps (en color vermell). Per fer-ho, s'ha creat un nou dataframe que conté tant les dades originals com les prediccions del model. El propòsit d'aquest gràfic és visualitzar la precisió del model en la predicció de la producció d'energia.



Podem observar que la predicció generada és força precisa ja que la proximitat entre la línia vermella (predicció) i la línia blava (dades reals) indica que el model està fent un bon treball a l'hora de predir la producció d'energia. A més, el model és capaç de captar les tendències i patrons de les dades (en els mesos de primavera i estiu la predicció de màxima energia produïda diària augmenta i, en canvi, en els mesos de tardor i hivern disminueix). És cert, però, que a l'inici del 2018, s'ha generat una predicció incorrecta, ja que l'energia produïda és negativa, un valor impossible. També cal dir que la predicció acostuma a exagerar una mica els pics d'energia màxima produïda als mesos d'estiu.

5. Conclusions

Exploració de les Dades:
En l'estudi inicial de les dades, s'ha realitzat un anàlisi descriptiu i s'ha estudiat la correlació entre les diferents variables. S'han identificar diverses variables que presenten una correlació significativa amb la variable objectiu, "Energy.delta.Wh". La Radiació Horitzontal Global (GHI), la temperatura i la longitud del dia tenen una correlació positiva, és a dir, que si alguna d'aquestes augmenta, la màxima energia produïda diària també ho fa. En canvi, la humitat té una correlació negativa alta (quan augmenta, la màxima energia produïda diària disminueix).

Anàlisi de Regressió:
S'ha implementat un model de regressió lineal múltiple per predir la variable "Energy.delta.Wh" a partir de les altres variables del conjunt de dades. El model de regressió s'ha construït utilitzant una combinació de selecció cap enrere i cap endavant basada en el criteri AIC, en lloc de simplement utilitzar el p-valor. Això ha permès la construcció d'un model amb una major capacitat predictiva.

Validació del Model:
El model resultant ha obtingut un R-Quadrat ajustat de 0.597, indicant que el model explica aproximadament el 59.7% de la variabilitat en la producció d'energia.

Mètodes Bootstrap:
S'han aplicat tècniques de Bootstrap paramètric i no paramètric per a obtenir una estimació més robusta dels coeficients del model. Això ha permès obtenir un interval de confiança per a cada coeficient, aportant una millor comprensió de l'incertesa associada a cada estimació.

Anàlisi de Residus:
Els residus del model final han sigut inspeccionats mitjançant un gràfic de probabilitat normal per a assegurar la suposició de normalitat dels errors. Aquesta inspecció no ha revelat cap violació significativa de la suposició.

Precisió de la Predicció:
El model final ha proporcionat prediccions força precises de la màxima producció d'energia diària a partir de diverses variables meteorològiques, captant les tendències i patrons de les dades. Tot i així, també hi ha hagut petits errors en les prediccions. Per tant, encara hi ha espai per a la millora, i futures recerques podrien explorar l'ús de models més avançats, com ara regressió polinòmica, models d'aprenentatge automàtic o xarxes neuronals.

6. Bibliografia

- Ramón Pérez, E. Guijarro, J. Lorente (2016). The use of weather forecasts in the energy sector. State Meteorological Agency (AEMET). Bulletin of the World Meteorological Organization, 60: 2-4. Available at: https://repositorio.aemet.es/bitstream/20.500.11765/3716/1/BolOMM%2060_2-4.pdf

- World Meteorological Organization (2018). Meteorology and the energy sector - WMO perspective. Public WMO INT. Available at: https://public.wmo.int/en/bulletin/meteorology-and-energy-sector-wmo-perspective

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

- Fox, J., & Weisberg, S. (2018). An R Companion to Applied Regression. Sage Publications.

- International Energy Agency (2021). World Energy Outlook. Available at: https://www.iea.org/reports/world-energy-outlook-2021

- European Commission (2020). The European Green Deal. Available at: https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

- Samaneh Manavi (2023). Renewable Energy and Weather Conditions Dataset. Kaggle. Available at: https://www.kaggle.com/datasets/samanemami/renewable-energy-and-weather-conditions

7. Annex: codi en R

```
#-------------------------------------------------------------------------------
#------------------------------ PROJECTE FINAL ---------------------------------
#-------------------------------------------------------------------------------

# Establim el directori de treball
setwd("C:/Users/Nil Farrés/Documents/NIL FARRÉS SOLER/UAB/2n curs/2n Semestre/
      Anàlisi de Dades Complexes/Projecte Final")

# Llegim les dades a analitzar
data <- read.csv("solar_weather.csv")

# Mostrem les primeres files de les dades
head(data)

#Mirem els noms de les variables
colnames(data)

library(dplyr)

# Primer, creem la columna date en el data frame original
data$date <- as.Date(data$Time)

# Després, subconjuntem de nou el data frame per obtenir els màxims diaris
data_max <- data %>%
  group_by(date) %>%
  slice_max(Energy.delta.Wh.)

# Eliminem la columna Time
data_max <- data_max %>% select(-Time)

# Mou la columna date al principi
data_max <- data_max %>%
  relocate(date, .before = Energy.delta.Wh.)

# Eliminem les columnes isSun, SunlightTime.daylength i weather_type, ja que no
# les considero prou rellevants per l'anàlisi que vull fer.
data_max <- data_max %>%
  select(-isSun, -SunlightTime.daylength, -weather_type)

print(data_max)


#----------------------------- ANÀLISI DESCRIPTIVA -----------------------------

# Resum estadístic de les dades
summary(data_max)

# Càrrega de la llibreria
library(ggplot2)

#Histograma de l'energia produïda màxima diària
hist(data_max$Energy.delta.Wh., main = "Histograma de l'energia produïda màxima
     diària", xlab = "Energy.delta.Wh", col = "blue")

# Gràfic de la màxima energia renovable produïda diàriament
ggplot(data_max, aes(x = date, y = Energy.delta.Wh.)) +
  geom_line(color = "blue") +
  labs(title = "Gràfic de la màxima energia renovable produïda diàriament",
       x = "Dia", y = "Energy.delta.Wh")

# Histograma de la Irradiació Solar Horitzontal Global (GHI)
hist(data_max$GHI, main = "Histograma de la Irradiació Solar Horitzontal
     Global (GHI)", xlab = "GHI", col = "green")

# Gràfic de de la Irradiació Solar Horitzontal Global (GHI) diària quan
# l'energia produïda és màxima
ggplot(data_max, aes(x = date, y = GHI)) +
  geom_line(color = "green") +
  labs(title = "Gràfic de de la Irradiació Solar Horitzontal Global (GHI)
       diària quan l'energia produïda és màxima", x = "Dia", y = "GHI")
```

```r
# Histograma de temperatures
hist(data_max$temp, main = "Histograma de temperatures",
     xlab = "Temperatura", col = "lightblue")

# Gràfic de la temperatura diària quan l'energia produïda és màxima
ggplot(data_max, aes(x = date, y = temp)) +
  geom_line(color = "lightblue") +
  labs(title = "Gràfic de la temperatura diària quan l'energia produïda és
       màxima", x = "Dia", y = "Temperatura")

# Histograma de la humitat
hist(data_max$humidity, main = "Histograma de la humitat",
     xlab = "Humitat", col = "red")

# Gràfic de l'humitat diària quan l'energia produïda és màxima
ggplot(data_max, aes(x = date, y = humidity)) +
  geom_line(color = "red") +
  labs(title = "Gràfic de l'humitat diària quan l'energia produïda és màxima",
       x = "Dia", y = "Humitat")

# Histograma de la pressió atmosfèrica
hist(data_max$pressure, main = "Histograma de la pressió atmosfèrica",
     xlab = "Pressió atmosfèrica", col = "purple")

# Gràfic de la pressió atmosfèrica diària quan l'energia produïda és màxima
ggplot(data_max, aes(x = date, y = pressure)) +
  geom_line(color = "purple") +
  labs(title = "Gràfic de la pressió atmosfèrica diària quan l'energia produïda
       és màxima", x = "Dia", y = "Pressió atmosfèrica")

# Histograma de la velocitat del vent
hist(data_max$wind_speed, main = "Histograma de la velocitat del vent",
     xlab = "Velocitat del vent (m/s)", col = "yellow")

# Gràfic de la velocitat del vent diària quan l'energia produïda és màxima
ggplot(data_max, aes(x = date, y = wind_speed)) +
  geom_line(color = "yellow") +
  labs(title = "Gràfic de la velocitat del vent diària quan l'energia produïda
       és màxima", x = "Dia", y = "Velocitat del vent (m/s)")

# Histograma de pluja a les hores en què la producció d'energia renovable és
# màxima
hist(data_max$rain_1h, main = "Histograma de pluja a les hores en què la
     producció d'energia renovable és màxima", xlab = "Pluja (mm)",
     col = "orange")

# Gràfic de pluja a les hores en què la producció d'energia renovable és
# màxima en funció del temps
ggplot(data_max, aes(x = date, y = rain_1h)) +
  geom_line(color = "orange") +
  labs(title = "# Gràfic de pluja a les hores en què la producció d'energia
       renovable és màxima en funció del temps", x = "Dia", y = "Pluja (mm)")

# Histograma de neu a les hores en què la producció d'energia renovable és
# màxima
hist(data_max$snow_1h, main = "Histograma de neu a les hores en què la
     producció d'energia renovable és màxima", xlab = "Neu (mm)", col = "gray")

# Gràfic de neu a les hores en què la producció d'energia renovable és màxima
# en funció del temps
ggplot(data_max, aes(x = date, y = snow_1h)) +
  geom_line(color = "gray") +
  labs(title = "# Gràfic de neu a les hores en què la producció d'energia
       renovable és màxima en funció del temps", x = "Dia", y = "Neu (mm)")
```

```r
# Gràfic de dispersió de "temp" vs "pressure"
ggplot(data_max, aes(x = temp, y = pressure)) +
  geom_jitter(alpha = 0.5, size = 2, width = 0.3, height = 0.3) +
  labs(title = "temp vs pressure", x = "temp", y = "pressure")

# Gràfic de dispersió de "humidity" vs "wind_speed"
ggplot(data_max, aes(x = humidity, y = wind_speed)) +
  geom_jitter(alpha = 0.5, size = 2, width = 0.3, height = 0.3) +
  labs(title = "humidity vs wind_speed", x = "humidity", y = "wind_speed")

# Gràfic de dispersió de "Energy.delta.Wh." vs "wind_speed"
ggplot(data_max, aes(x = Energy.delta.Wh., y = wind_speed)) +
  geom_jitter(alpha = 0.5, size = 2, width = 0.3, height = 0.3) +
  labs(title = "Energy.delta.Wh. vs wind_speed", x = "Energy.delta.Wh.",
       y = "wind_speed")

# Gràfic de dispersió de "Energy.delta.Wh." vs "GHI"
ggplot(data_max, aes(x = Energy.delta.Wh., y = GHI)) +
  geom_jitter(alpha = 0.5, size = 2, width = 0.3, height = 0.3) +
  labs(title = "Energy.delta.Wh. vs GHI", x = "Energy.delta.Wh.", y = "GHI")

# Diagrama de caixa de la temperatura
boxplot(data_max$temp, data_max = data_max, xlab = "", ylab = "Temperatura",
        main = "Diagrama de caixa de la temperatura", col = "lightblue")

# Diagrama de caixa de la pressió atmosfèrica
boxplot(data_max$pressure, data_max = data_max, xlab = "",
        ylab = "Pressió atmosfèrica", main = "Diagrama de caixa de la pressió
        atmosfèrica", col = "purple")

# Diagrama de caixa de la humitat
boxplot(data_max$humidity, data_max = data_max, xlab = "", ylab = "Humitat",
        main = "Diagrama de caixa de la humitat", col = "red")

# Diagrama de caixa de la velocitat del vent
boxplot(data_max$wind_speed, data_max = data_max, xlab = "",
        ylab = "Velocitat del vent", main = "Diagrama de caixa de la
        velocitat del vent", col = "yellow")

# Diagrama de caixa de la màxima energia diària produïda
boxplot(data_max$Energy.delta.Wh., data_max = data_max, xlab = "",
        ylab = "Màxima energia diària produïda", main = "Diagrama de caixa
        de la màxima energia diària produïda", col = "blue")

# Diagrama de caixa de GHI
boxplot(data_max$GHI, data_max = data_max, xlab = "", ylab = "GHI",
        main = "Diagrama de caixa de GHI", col = "green")

#------------------------- ANÀLISI DE CORRELACIÓ -------------------------------

#Agafem només les dades numèriques.
numeric_data <- data_max[, sapply(data_max, is.numeric)]

#Construïm la matriu de correlació
correlation_matrix <- cor(numeric_data)

# Arrodonim la matriu de correlació a dos decimals
correlation_matrix <- round(correlation_matrix, 2)

# Mostra la matriu de correlació
print(correlation_matrix)

# Gràfic de correlació
library(corrplot)
corrplot(correlation_matrix)

# Matriu de dispersió
pairs(numeric_data)
```

```r
#------------------------ ANÀLISI DE REGRESSIÓ --------------------------------

# Carreguem el paquet MASS
library(MASS)

# Creem un model de regressió lineal complet amb totes les variables
model_full <- lm(Energy.delta.wh. ~ ., data = data_max)
summary(model_full)

# Realitzem la selecció de models cap enrere
model_backward <- stepAIC(model_full, trace = TRUE, direction = "backward")

# Creem el model nul (sense predictors)
null_model <- lm(Energy.delta.wh. ~ 1, data = data_max)

# Realitzem la selecció cap endavant
model_forward <- stepAIC(null_model, scope = list(lower = null_model,
                                                  upper = model_full),
                         direction = "forward", trace = TRUE)

# Apliquem stepwise selection
model_stepwise <- step(model_full)

# Creem un model final amb les variables seleccionades,
# eliminant la variable GHI, que no ens interessa.
model_final <- lm(Energy.delta.wh. ~ date + temp + pressure +
                    humidity + wind_speed + rain_1h + snow_1h + clouds_all,
                  data = data_max)

# Resum del model
summary(model_final)

# Calculem els residus del model
residus <- resid(model_final)

# Gràfic de probabilitat normal dels residus:
# Aquest gràfic es pot utilitzar per verificar la suposició de normalitat
# dels errors.
qqnorm(residus, main = "Gràfic de probabilitat normal dels Residus")
qqline(residus)

#---------------------------- MÈTODES BOOTSTRAP --------------------------------

# Generem valors ajustats (fitted values)
fitted_values <- predict(model_final)

# Generate residuals
resids <- residuals(model_final)

# Generem una nova resposta
new_response <- fitted_values + sample(resids, replace = TRUE)

# Creem new data
new_data <- data_max
new_data$Energy.delta.wh. <- new_response

# Ajustem el model a les noves dades
new_fit <- lm(Energy.delta.wh. ~ date + temp + pressure +
                humidity + wind_speed + rain_1h + snow_1h + clouds_all,
              data = new_data)
```

```r
# --- BOOTSTRAP PARAMÈTRIC ---

# Inicialitzem una matriu per emmagatzemar les estimacions dels paràmetres
param_estimates_bparam <- matrix(nrow = 10000, ncol = length(coef(model_final)))

# Bucle Bootstrap paramètric
for(i in 1:10000) {

  # Generem una nova resposta
  new_response <- fitted_values + sample(resids, replace = TRUE)


  # Actualitzem les dades
  new_data$Energy.delta.Wh. <- new_response

  # Ajustem el model a les noves dades
  new_fit <- lm(Energy.delta.Wh. ~ date + temp + pressure +
                  humidity + wind_speed + rain_1h + snow_1h + clouds_all,
                data = new_data)

  # Emmagatzemem les estimacions dels paràmetres
  param_estimates_bparam[i, ] <- coef(new_fit)

}


# --- BOOTSTRAP NO-PARAMÈTRIC ---

# Initialize a matrix to store parameter estimates
param_estimates_bnoparam <- matrix(nrow = 10000,
                                  ncol = length(coef(model_final)))

# Bootstrap loop
for(i in 1:10000) {

  # Resample data
  new_data <- data_max[sample(nrow(data_max), replace = TRUE), ]

  # Fit the model to the new data
  new_fit <- lm(Energy.delta.Wh. ~ date + temp + pressure +
                  humidity + wind_speed + rain_1h + snow_1h + clouds_all,
                data = new_data)

  # Store parameter estimates
  param_estimates_bnoparam[i, ] <- coef(new_fit)

}


# --- HISTOGRAMES PER A CADA PARÀMETRE ESTIMAT ---

library(ggpubr)
library(ggplot2)

# Creem una llista buida per emmagatzemar els histogrames
hist_list <- list()

# Convertim la matriu en un data.frame
df <- as.data.frame(param_estimates_bparam)

# Definim correctament els noms de les columnes
names(df) <- c("Intercept", "date", "temp", "pressure", "humidity",
               "wind_speed", "rain_1h", "snow_1h", "clouds_all")

# Creem un histograma per a cada paràmetre estimat i l'emmagatzema a la llista
for (i in 3:ncol(df)) {
  p <- ggplot(df, aes_string(x = names(df)[i])) +
    geom_histogram(bins = 30, fill = 'blue', color = 'black') +
    labs(title = paste("Distribució Bootstrap de", names(df)[i]),
         x = names(df)[i],
         y = "Freqüència") +
    theme_minimal()

  hist_list[[i-2]] <- p
}
```

```r
# Utilitzem ggarrange per combinar tots els histogrames en un sol gràfic
combined_plot <- ggarrange(plotlist = hist_list, ncol = 4, nrow = 2)

# Mostrem el gràfic combinat
print(combined_plot)


# --- GRÀFIC DE LES PREDICCIONS ---

# Creem un nou dataframe amb les dades originals i les prediccions
prediction_data <- data_max
prediction_data$predicted_energy <- predict(model_final, newdata = data_max)

# Utilitzem ggplot per a dibuixar les dades originals i les prediccions de la
# màxima energia produïda diàriament.
ggplot(prediction_data, aes(x = date)) +
  geom_line(aes(y = Energy.delta.Wh.), color = "blue") +
  geom_line(aes(y = predicted_energy), color = "red") +
  labs(x = "Dia", y = "Màxima energia produïda",
       title = "Màxima energia produïda diària actual i predita",
       color = "Source") +
  scale_color_manual(values = c("blue", "red"),
                     labels = c("Actual", "Predicted"))
```