

Pràctica 2: Com realitzar la neteja i l'anàlisi de dades?

Document d'entrega

Esther Manzano i Albert Queraltó

9 de enero 2023

Llibreries

Primer de tot, carregarem algunes llibreries que podríem necessitar per crear els diferents models, analitzar-los i generar gràfics.

```
# Instal·lem i/o carreguem les llibreries utilitzades per fer l'anàlisi exploratòria de les dades
packages <- c("ggplot2", "plyr", "dplyr", "stats", "lmtest", "caret", "corrplot",
           "Hmisc", "rpart", "rpart.plot", "useful", "ROCR", "precrec", "knitr",
           "lubridate", "stringr", "psych", "pROC", "zoo", "reshape2",
           "gridExtra", "ggpubr", "rstatix", "lattice", "ResourceSelection",
           "car", "quantreg", "rsq", "nortest", "MASS")  
  
# install.packages(packages) # Comentar si no cal instal·lar les llibreries
lapply(packages, require, character.only = TRUE)
```

Enllaç al vídeo

El vídeo explicatiu de la pràctica es pot trobar a:

https://drive.google.com/file/d/1-Yeq2jDGPJJzfE3Md-4nxC9DBUALn-_1/view?usp=share_link

1. Descripció del dataset

En els últims mesos, l'evolució del preu de l'energia elèctrica ha augmentat considerablement degut a la pujada dels preus dels combustibles fòssils, a la guerra d'Ucraïna i a l'especulació dels mercats. [1, 2] Això ha provocat una situació d'inflació insostenible per a moltes famílies i comerços, i ha posat de nou les energies renovables en el centre de la política energètica amb l'objectiu de reduir la dependència dels combustibles fòssils i redreçar la situació. [1-3] El preu de l'energia d'origen renovable ha disminuït considerablement en els últims anys si es compara amb altres fonts d'energia no renovables, com el gas natural o el carbó. [4] A més, hi ha una necessitat creixent de reduir la producció d'energia a partir de combustibles fòssils per a reduir les emissions de gasos d'efecte hivernacle a l'atmosfera i frenar el canvi climàtic. [5]

Per aquest motiu i amb l'objectiu de promoure la producció i utilització de l'energia renovable, es vol analitzar l'evolució del preu de l'energia elèctrica a Espanya i la producció general d'energies renovables en el període de temps comprès entre l'1 de Novembre del 2020 al 31 d'Octubre del 2022 en la mateixa zona geogràfica.

Aquesta informació es va recopilar en la primera pràctica mitjançant la tècnica de webscraping a partir de les dades obtingudes de la web <https://www.esios.ree.es/es/> (*dataset ESIOS*), per a la qual es va demanar permís explícit per a tractar i publicar el resultat obtingut amb aquestes dades. El resultat de la primera pràctica es pot trobar publicat a: <https://github.com/albert-queralto/scraping-energy-prices-spain> [6-7]

D'altra banda, cal mencionar també com la meteorologia influeix significativament en la producció de les energies renovables. Per exemple, es preveu que els dies assolellats tinguin una elevada producció d'origen fotovoltaic, així com l'energia eòlica serà representativa en els dies ventosos o la quantitat de precipitació contribuirà a omplir els embassaments que permeten generar energia hidroelèctrica.

Així doncs, per completar l'anàlisi, també s'ha decidit utilitzar dades climàtiques diàries obtingudes de la web <https://opendata.aemet.es/> (*dataset AEMET*). Entre altres, hi ha dades de temperatures, precipitació, vent, insolació, etc., en el mateix rang de dates que el dataset d'ESIOS. La combinació d'ambdós datasets ens permetrà assolir els objectius següents marcats per a aquesta pràctica:

- Analitzar l'evolució del preu de l'energia elèctrica a Espanya en el període de temps mencionat.
- Entendre la relació entre el preu de l'energia elèctrica, la producció d'energies renovables i la influència de la meteorologia.
- Discernir quina font d'energia renovable té un impacte més rellevant en el preu de l'energia.

2. Integració i selecció

Degut a l'estensió del preprocessat, aquest s'ha dut a terme en un fitxer a part (*PRA2_preprocessing.Rmd*). En aquest document, només es mostra la integració i selecció de les dades.

2.1. Càrrega i selecció del dataset ESIOS

2.1.1. Càrrega de les dades i anàlisi a alt nivell

Carreguem les dades del *dataset ESIOS* creat a la pràctica 1 i preprocessat (veure *PRA2_preprocessing.Rmd* i pdf associat). Aquestes dades són les obtingudes a partir de la web <https://www.esios.ree.es/es/> [6-7] i contenen la informació de l'evolució del preu de l'energia elèctrica a Espanya i la producció d'energies renovables en el període de temps comprès entre l'1 de Novembre del 2020 al 31 d'Octubre del 2022. Aquest dataset tenia una granularitat per hores. Tanmateix, el *dataset AEMET* contené dades per dia, per això, s'han calculat els promitjos diaris de les dades del *dataset ESIOS*. Per fer aquest càlcul, s'han hagut de tractar també els valors nuls, ja què si no, el promig diari no es podia calcular correctament. La imputació de valors nuls, tot i què aquests eren menys d'un 0.1% de les dades, s'ha fet interpolant les dades agrupades per la variable *energy_source*.

```
# Carreguem els datasets en un data frame
esios_dataset <- read.csv('data/esios_dataset_preprocessed.csv', header = TRUE,
                         sep=",", stringsAsFactors = TRUE)

# Transformem la columna date al format correcte
esios_dataset$date <- as.Date(esios_dataset$date, format = "%Y-%m-%d")

# Taula resum
str(esios_dataset)

## 'data.frame': 4380 obs. of 13 variables:
## $ date : Date, format: "2020-11-01" "2020-11-01" ...
## $ energy_source : Factor w/ 6 levels "hydroelectric",...: 1 2 3 4 5 6 1 2 3 4 ...
## $ avg_total_price_eur_mwh : num 31.6 31.6 31.6 31.6 31.6 ...
## $ avg_price_free_market_eur_mwh: num 31.3 31.3 31.3 31.3 31.3 ...
## $ avg_price_ref_market_eur_mwh : num 34.1 34.1 34.1 34.1 34.1 ...
## $ energy_total_mwh : num 22639 22639 22639 22639 22639 ...
## $ energy_free_market_mwh : num 19670 19670 19670 19670 19670 ...
## $ energy_ref_market_mwh : num 2969 2969 2969 2969 2969 ...
## $ free_market_share_perc : num 87.1 87.1 87.1 87.1 87.1 ...
## $ ref_market_share_perc : num 12.9 12.9 12.9 12.9 12.9 ...
## $ renewable_generation_perc : num 6.6 26.83 7.68 2.5 66.59 ...
## $ renewable_generation_kw : num 1494 6074 1738 565 15074 ...
## $ renewable_generation_perc_bin: Factor w/ 4 levels "0-35","36-75",...: 1 1 1 1 2 1 1 1 1 1 ...
```

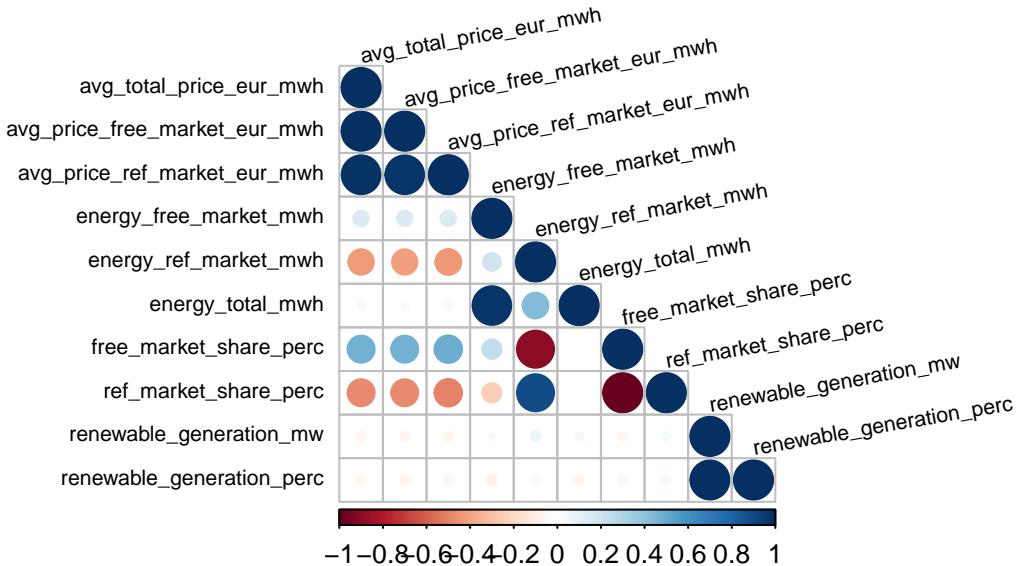
El dataset conté dades estructurades en 13 columnes i 4380 files. Les columnes *date*, *energy_source* i *renewable_generation_perc_bin* eren de tipus *character* i s'han transformat al tipus correcte (*Date* i *factor*).

2.1.2. Filtrat de variables numèriques

Filtrem les variables numèriques fent servir una matriu de correlació:

```
# Calculem la matriu de correlació
corr_vars_esios <- c("avg_total_price_eur_mwh", "avg_price_free_market_eur_mwh",
                      "avg_price_ref_market_eur_mwh", "energy_free_market_mwh",
                      "energy_ref_market_mwh", "energy_total_mwh",
                      "free_market_share_perc", "ref_market_share_perc",
                      "renewable_generation_kw", "renewable_generation_perc")
corr_matrix_esios <- cor(esios_dataset[, corr_vars_esios], use='complete.obs',
                           method=c("pearson"))

# Representem gràficament la matriu de correlació
corrplot(corr_matrix_esios, method = "circle", type = "lower",
          tl.col = "black", tl.srt = 10, tl.cex = 0.7)
```



Les variables *avg_price_free_market_eur_mwh*, *avg_price_ref_market_eur_mwh* i *avg_total_price_eur_mwh* tenen una elevada correlació, per això només conservarem la darrera. Pel que fa a les variables *free_market_share_perc*, *ref_market_share_perc*, *energy_free_market_mwh* i *energy_ref_market_mwh*, s'ha decidit eliminar-les ja que s'ha considerat que no aporten informació rellevant per a l'estudi proposat. Les variables *renewable_generation_kw* i *renewable_generation_perc* també estan fortament correlacionades i només conservarem la segona. Així doncs, les columnes que es volen conservar són: *date*, *avg_total_price_eur_mwh*, *energy_total_mwh*, *energy_source*, *renewable_generation_perc* i *renewable_generation_perc_bin*.

```
# Filtrem les columnes que no ens interessin
esios_dataset <- esios_dataset %>%
  dplyr::select(date, avg_total_price_eur_mwh, energy_total_mwh, energy_source,
               renewable_generation_perc, renewable_generation_perc_bin)

# Mostrem les primeres files del dataset
head(esios_dataset)

## #> #>   date avg_total_price_eur_mwh energy_total_mwh   energy_source
## #> 1 2020-11-01           31.60917     22638.71 hydroelectric
## #> 2 2020-11-01           31.60917     22638.71      nuclear
## #> 3 2020-11-01           31.60917     22638.71       solar
## #> 4 2020-11-01           31.60917     22638.71 thermorenewable
## #> 5 2020-11-01           31.60917     22638.71        total
## #> 6 2020-11-01           31.60917     22638.71       wind
## #>   renewable_generation_perc renewable_generation_perc_bin
## #> 1                 6.600605            0-35
## #> 2                26.829416            0-35
## #> 3                 7.678957            0-35
## #> 4                2.496646            0-35
## #> 5               66.585991            36-75
## #> 6                22.980367            0-35
```

2.1.10. Representem les variables resultants

Representem les variables numèriques del dataset resultant en diferents histogrames i les categòriques en gràfics de barres:

```
# Histograma de les variables numèriques
total_price_hist <- ggplot(esios_dataset, aes(x = avg_total_price_eur_mwh)) +
  geom_histogram(bins = 50, fill = c("#33CCFF"), alpha = 0.5) +
  labs(title = "avg_total_price_eur_mwh", x = "Avg total price (euro/MWh)",
       y = "Freqüència") + theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

energy_total_hist <- ggplot(esios_dataset, aes(x = energy_total_mwh)) +
  geom_histogram(bins = 50, fill = c("#33CCFF"), alpha = 0.5) +
  labs(title = "energy_total_mwh", x = "Energy total (MWh)", y = "Freqüència") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

renew_perc_hist <- ggplot(esios_dataset, aes(x = renewable_generation_perc)) +
  geom_histogram(bins = 50, fill = c("#33CCFF"), alpha = 0.5) +
  labs(title = "renewable_generation_perc", x = "Renewable generation (%)",
       y = "Freqüència") + theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

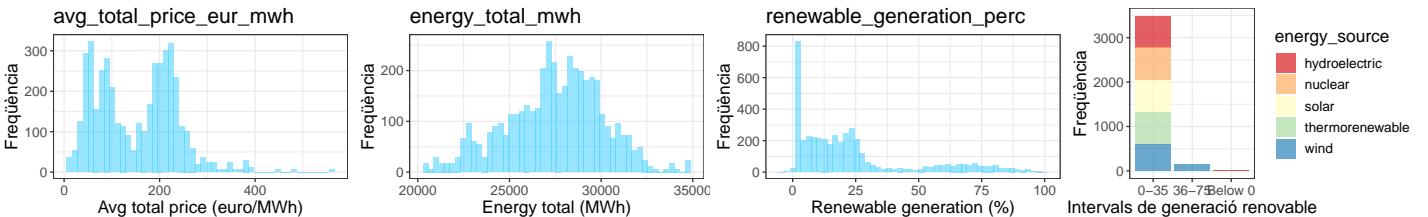
# Barplot per renewable_generation_perc_bin en funció de energy_source
```

```

energy_source_data <- esios_dataset %>% filter(energy_source != "total")
renew_perc_bin_bar <- ggplot(energy_source_data,
  aes(x = renewable_generation_perc_bin,
      fill = energy_source)) + geom_bar(alpha=0.7) +
  scale_fill_brewer(palette="Spectral") +
  labs(x = "Intervals de generació renovable", y = "Frequència") +
  theme_bw() + theme(text = element_text(size = 14))

grid.arrange(total_price_hist, energy_total_hist, renew_perc_hist,
            renew_perc_bin_bar, ncol = 4)

```



Observem que la variable *renewable_generation_perc* té freqüència més alta al voltant del 5% (~800 MW), mentre què aquesta freqüència disminueix a mesura que augmenta la potència renovable, essent gairebé nul · la en valors del 100% (~30000 MW). Això indicaria que la producció renovable és normalment baixa, entre el 0 i el 25%. L'*avg_total_price_eur_mwh* presenta dues distribucions, la primera amb un màxim entorn a 100 MW i la segona amb un màxim al voltant de 200 MW. Això indicaria la presència de períodes amb dos rangs de preus molt diferents. La variable *energy_total_mwh* presenta una distribució lleugerament esbiaixada cap a l'esquerra amb un màxim de freqüència entorn els 28000 MWh. Pel que fa al gràfic de barres d'*renewable_generation_perc_bin*, la majoria d'observacions tenen un percentatge de generació entre el 0% i el 35% amb una representació bastant semblant per les diferents fonts d'energia. Per tant, es pot concloure que la generació renovable es troba normalment per sota del 35%. Pel que fa a l'energia eòlica, s'observen també algunes observacions en el rang del 36% al 75%, essent la font d'energia que més contribueix a la producció renovable. D'altra banda, l'energia hidroelèctrica consumeix energia en alguns casos.

Representem la sèrie temporal de les mateixes variables numèriques:

```

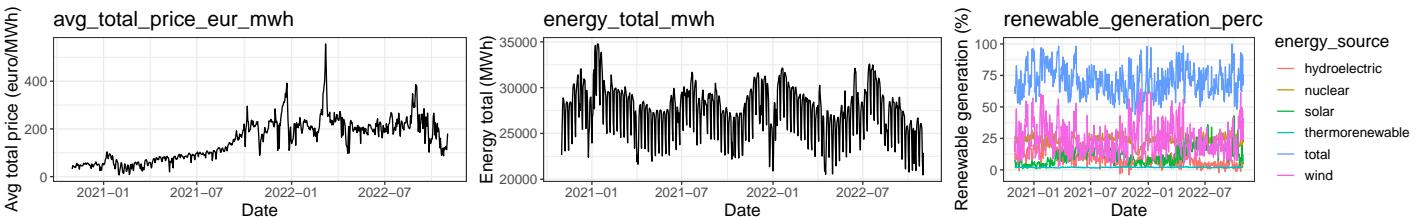
# Serie temporal de les variables avg_total_price_eur_mwh, energy_total_mwh,
#renewable_generation_mw i renewable_generation_perc
total_price_plot <- ggplot(esios_dataset,
  aes(x = date, y = avg_total_price_eur_mwh)) +
  geom_line() + labs(title = "avg_total_price_eur_mwh", x = "Date",
                     y = "Avg total price (euro/MWh)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

energy_total_plot <- ggplot(esios_dataset, aes(x = date, y = energy_total_mwh)) +
  geom_line() + labs(title = "energy_total_mwh", x = "Date",
                     y = "Energy total (MWh)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

renewable_generation_perc_plot <- ggplot(esios_dataset, aes(x = date,
  y = renewable_generation_perc, color = energy_source)) +
  geom_line() + labs(title = "renewable_generation_perc", x = "Date",
                     y = "Renewable generation (%)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14
    )))

grid.arrange(total_price_plot, energy_total_plot, renewable_generation_perc_plot,
            ncol = 3)

```



La presència de les dues distribucions per *avg_total_price_eur_mwh* es veu clarament ja que els preus eren més propers als 100 €/MWh per l'any 2021, mentre què aquests van pujar entorn els 200 €/MWh al 2022. L'evolució de *energy_total_mwh* mostra que la producció a Espanya es situa entre els 22000 i els 32000 MW, aproximadament. Pel que fa a les energies lliures de CO₂, l'energia eòlica i la nuclear predominen en la producció. Malgrat això, la solar també a vist un agument en els darrers mesos del 2022. La disminució de la producció hidroelèctrica pot estar ocasionada per la sequera extrema que

estem vivint actualment, mentre què la producció termorenovable no té una representació gaire significativa.

2.2. Càrrega i selecció del dataset AEMET

2.2.1. Càrrega de les dades i anàlisi a alt nivell

Carreguem les dades del *dataset AEMET* preprocessat (veure *PRA2_preprocessing.Rmd* i pdf associat). Aquestes dades s'han obtingut de la web <https://opendata.aemet.es/> utilitzant l'API proporcionada. La informació continguda al dataset va des de l'1 d'Octubre del 2020 al 30 de Novembre del 2022. Aquest estava format per la data d'adquisició dels registres, el número identificador de l'estació meteorològica, el seu nom, la província, l'altitud sobre el nivell del mar (en metres), la temperatura mitjana diària ($^{\circ}\text{C}$), la precipitació diària (en mm), la temperatura mínima ($^{\circ}\text{C}$), l'hora i minuts d'adquisició de la temperatura mínima, la temperatura màxima ($^{\circ}\text{C}$), l'hora i minuts d'adquisició de la temperatura màxima, la direcció de la ratxa de vent màxima, la velocitat mitjana del vent (m/s), l'hora de la ratxa màxima, l'insolació (en hores), la pressió màxima (hPa), l'hora de la pressió màxima, la pressió mínima (hPa), l'hora de la pressió mínima.

La presència de diverses estacions meteorològiques per província ha requerit el càlcul dels promitjos diares per província per tal de simplificar les dades. Per fer aquest càlcul, s'han hagut de tractar també els valors nuls durant el preprocessat, ja què si no, el promig diari no es podia calcular correctament. La imputació de valors nuls s'ha fet interpolant les dades agrupades per la variable *province*. A més, també s'ha decidit eliminar les columnes *climate_indicator*, *station_name*, *hour_min_temp*, *hour_max_temp*, *hour_max_wind_speed*, *hour_max_pressure* i *hour_min_pressure* durant el preprocessat ja que la informació aportada no era rellevant pel nostre anàlisi. També, s'ha creat la variable *avg_pressure* a partir de les variables *max_pressure* i *min_pressure*.

```
# Carreguem les dades del dataset AEMET preprocessat
aemet_dataset <- read.csv("data/aemet_dataset_preprocessed.csv", sep = ",",
                           header = TRUE, stringsAsFactors = TRUE)

# Transformem la columna date al format correcte
aemet_dataset$date <- as.Date(aemet_dataset$date, format = "%Y-%m-%d")

# Taula resum
str(aemet_dataset)

## 'data.frame': 37960 obs. of 16 variables:
## $ date : Date, format: "2020-11-01" "2020-11-01" ...
## $ province : Factor w/ 52 levels "ACORUÑA","ALBACETE",...: 1 2 3 4 5 6 7 8 9 10 ...
## $ height_above_sea : num 154 715 178 240 513 ...
## $ avg_daily_temp : num 17.9 16.2 17.4 17.2 17.4 ...
## $ daily_precipitation: num 7.66 0 0.05 0 0 ...
## $ min_daily_temp : num 17.1 8.6 10.2 11.9 13.5 ...
## $ max_daily_temp : num 18.9 23.8 24.6 22.5 21.2 ...
## $ max_wind_direction : num 22.8 29.8 59.5 37.2 25 ...
## $ avg_wind_speed : num 7.55 1.4 1.68 1.68 5.8 ...
## $ max_wind_speed : num 15.91 5.8 5.55 7.72 10.8 ...
## $ insolation : num 0.181 8.988 9.381 7.565 7 ...
## $ max_pressure : num 998 944 1009 998 965 ...
## $ min_pressure : num 996 941 1006 995 962 ...
## $ avg_pressure : num 997 943 1008 996 964 ...
## $ avg_pressure_bin : Factor w/ 3 levels "Anticyclone",...: 2 2 2 2 2 2 2 2 3 ...
## $ insolation_bin : Factor w/ 3 levels "0-5","10+","5-10": 3 2 2 2 2 3 2 2 3 ...
```

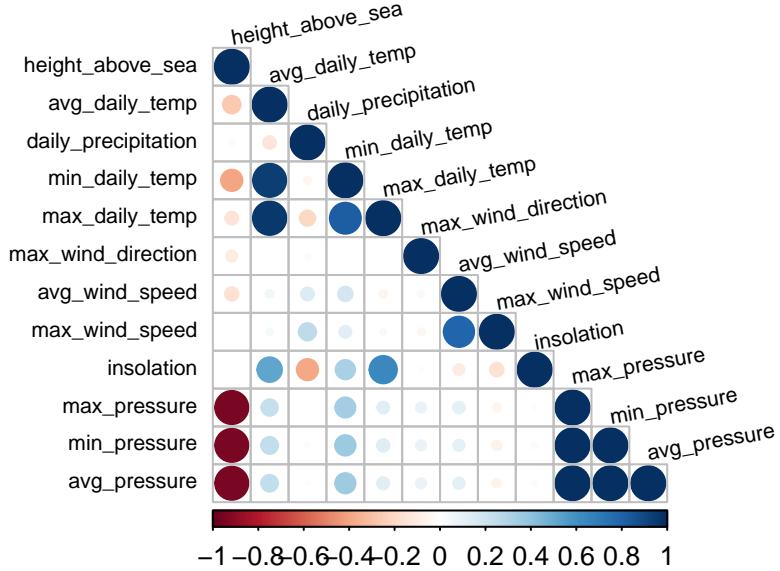
El dataset conté 16 columnes i 37960 files. Les columnes *date*, *province*, *avg_pressure_bin* i *insolation_bin* eren de tipus *character* i s'han transformat al tipus correcte (*Date* i *factor*).

2.2.2. Filtrat de variables numèriques

Es calcula la matriu de correlació entre les variables numèriques per veure aquelles que estan altament correlacionades i conservar-ne només una:

```
# Calculem la matriu de correlació
corr_matrix_aemet <- cor(aemet_dataset[, c("height_above_sea", "avg_daily_temp",
                                             "daily_precipitation", "min_daily_temp", "max_daily_temp",
                                             "max_wind_direction", "avg_wind_speed", "max_wind_speed",
                                             "insolation", "max_pressure", "min_pressure",
                                             "avg_pressure")], use = "complete.obs", method=c("pearson"))

# Representem gràficament la matriu de correlació
corrplot(corr_matrix_aemet, method = "circle", type = "lower", tl.col = "black",
          tl.srt = 10, tl.cex = 0.7)
```



Els resultats mostren com les variables `min_daily_temp` i `max_daily_temp` estan altament correlacionades amb `avg_daily_temp`, per tant, mantindrem aquesta última. Pel que fa a les variables `height_above_sea`, `max_pressure`, `min_pressure` i `avg_pressure`, aquestes també tenen una forta correlació. Per tant, mantindrem només `avg_pressure`. Finalment, també eliminarem la direcció del vent, ja què hem considerat que no és útil per la nostra anàlisis. Així doncs, les columnes que es volen conservar són: `date`, `province`, `avg_daily_temp`, `daily_precipitation`, `avg_wind_speed`, `max_wind_speed`, `insolation` i `avg_pressure`:

```
# Filtem les columnes que no ens interessin
aemet_dataset <- aemet_dataset %>% dplyr::select(date, province, avg_daily_temp,
                                                 daily_precipitation, avg_wind_speed,
                                                 max_wind_speed, insolation, avg_pressure,
                                                 avg_pressure_bin, insolation_bin)
```

2.2.3. Representem les variables resultants

Representem les variables numèriques del dataset resultant en diferents histogrames i les categòriques en gràfics de barres:

```
# Histograma de les variables numèriques
avg_temp_hist <- ggplot(aemet_dataset, aes(x = avg_daily_temp)) +
  geom_histogram(bins = 50, fill = c("#33CCFF"), alpha = 0.5) +
  labs(title = "avg_daily_temp", x = "Temperatura (°C)", y = "Freqüència") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

precipitation_hist <- ggplot(aemet_dataset, aes(x = daily_precipitation)) +
  geom_histogram(bins = 50, fill = c("#33CCFF"), alpha = 0.5) +
  labs(title = "daily_precipitation", x = "Precipitació (mm)", y = "Freqüència") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

avg_wind_hist <- ggplot(aemet_dataset, aes(x = avg_wind_speed)) +
  geom_histogram(bins = 50, fill = c("#33CCFF"), alpha = 0.5) +
  labs(title = "avg_wind_speed", x = "Avg. wind speed (km/h)", y = "Freqüència") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

max_wind_hist <- ggplot(aemet_dataset, aes(x = max_wind_speed)) +
  geom_histogram(bins = 50, fill = c("#33CCFF"), alpha = 0.5) +
  labs(title = "max_wind_speed", x = "Max. wind speed (km/h)", y = "Freqüència") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

avg_pressure_hist <- ggplot(aemet_dataset, aes(x = avg_pressure)) +
  geom_histogram(bins = 50, fill = c("#33CCFF"), alpha = 0.5) +
  labs(title = "avg_pressure", x = "Avg. pressure (hPa)", y = "Freqüència") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

insolation_hist <- ggplot(aemet_dataset, aes(x = insolation)) +
  geom_histogram(bins = 50, fill = c("#33CCFF"), alpha = 0.5) +
  labs(title = "insolation", x = "Insolation (h)", y = "Freqüència") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

# Barplot per avg_pressure_bin i insolation_bin
```

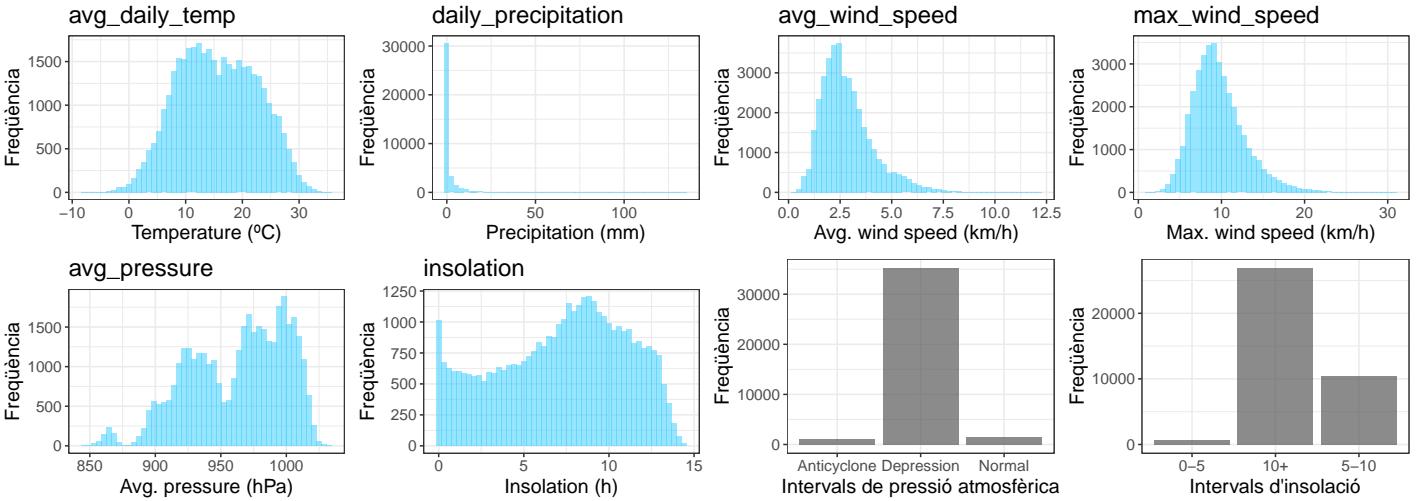
```

avg_pressure_bin_bar <- ggplot(aemet_dataset, aes(x = avg_pressure_bin)) +
  geom_bar(alpha=0.7) + scale_fill_brewer(palette="Spectral") +
  labs(x = "Intervals de pressió atmosfèrica", y = "Freqüència") +
  theme_bw() + theme(text = element_text(size = 14))

insolation_bin_bar <- ggplot(aemet_dataset, aes(x = insolation_bin)) +
  geom_bar(alpha=0.7) + scale_fill_brewer(palette="Spectral") +
  labs(x = "Intervals d'insolació", y = "Freqüència") +
  theme_bw() + theme(text = element_text(size = 14))

grid.arrange(avg_temp_hist, precipitation_hist, avg_wind_hist, max_wind_hist,
            avg_pressure_hist, insolation_hist, avg_pressure_bin_bar,
            insolation_bin_bar, ncol = 4)

```



La variable *avg_daily_temp* presenta una distribució semblant a una gaussiana bastant ample. Així, la temperatura diària mitjana a Espanya en el període analitzat és de 15°C. Fins i tot, es podria separar en dues distribucions, de mitjanes properes als 10 i 20 °C, respectivament. Això, tindria el seu origen en províncies del nord on acostuma a fer més fred, mentre què en les del sud, fa més calor. La *daily_precipitation* té una distribució molt esbiaixada cap a la dreta ja que la majoria de valors són propers a 0 mm. Per tant, no hi ha hagut gaire precipitació en el període estudiat. Pel que va a les variables *avg_wind_speed* i *max_wind_speed*, tenen distribucions prou normals amb un cert biaix cap a la dreta. Les mitjanes són de 2.5 i 10 km/h, aproximadament. Sorprèn que l'energia eòlica sigui la font renovable amb més representació, malgrat estem tractant amb valors promitjats que poden introduir cert biaix en l'anàlisi. L'*avg_pressure* presenta quatre distribucions amb valors promitjats al voltant de 870, 920, 970 i 1000 hPa, aproximadament. La classificació feta a *avg_pressure_bin* indicaria que la majoria de pressions atmosfèriques estan associades amb depressió, valors per sota de 1013 hPa. Tanmateix, això no estarà sempre associat amb precipitació o mal temps tot i ser més probable, ja què la pressió atmosfèrica només mesura el pes de l'aire en una ubicació. Finalment, la *insolation* mostra una distribució una mica uniforme, tot i què s'observa un increment d'observacions a partir de les 5 hores amb un màxim al voltant de 8h. La classificació feta a *insolation_bin* indica que predominen els dies assolellats, ja que les hores d'insolació es troben en els intervals 5-10 hores o 10+ hores.

Representem la sèrie temporal de les mateixes variables numèriques:

```

# Serie temporal de les variables avg_daily_temp, daily_precipitation,
# avg_wind_speed, max_wind_speed, insolation i avg_pressure
avg_temp_plot <- ggplot(aemet_dataset, aes(x = date, y = avg_daily_temp)) +
  geom_line(color = 'orange', alpha = 0.5) +
  geom_line(data = aggregate(avg_daily_temp ~ date, aemet_dataset, mean),
            aes(x = date, y = avg_daily_temp), color = 'blue', alpha = 0.7) +
  labs(title = "avg_daily_temp", x = "Date", y = "Temperature (°C)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

precipitation_plot <- ggplot(aemet_dataset, aes(x = date, y = daily_precipitation)) +
  geom_line(color = 'orange', alpha = 0.5) +
  geom_line(data = aggregate(daily_precipitation ~ date, aemet_dataset, mean),
            aes(x = date, y = daily_precipitation), color = 'blue', alpha = 0.7) +
  labs(title = "daily_precipitation", x = "Date", y = "Precipitation (mm)") +
  theme(plot.title = element_text(hjust = 0.5)) + theme_bw() +
  theme(text = element_text(size = 14))

avg_wind_plot <- ggplot(aemet_dataset, aes(x = date, y = avg_wind_speed)) +
  geom_line(color = 'orange', alpha = 0.5) +
  geom_line(data = aggregate(avg_wind_speed ~ date, aemet_dataset, mean),

```

```

aes(x = date, y = avg_wind_speed), color = 'blue', alpha = 0.7) +
  labs(title = "avg_wind_speed", x = "Date", y = "Avg. wind speed (km/h)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() + theme(text = element_text(size = 14))

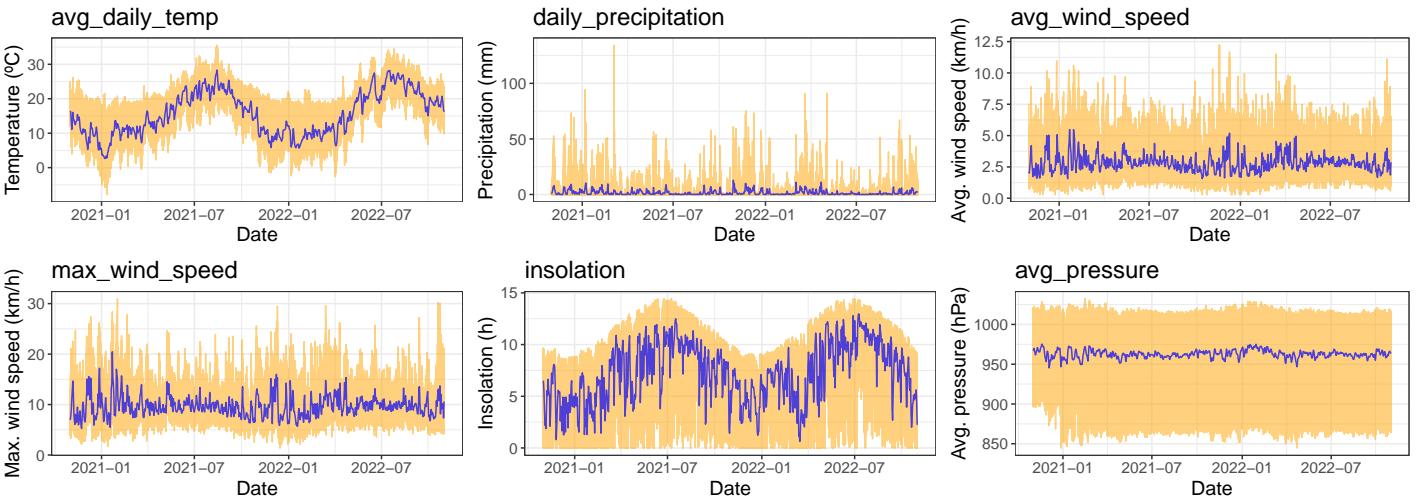
max_wind_plot <- ggplot(aemet_dataset, aes(x = date, y = max_wind_speed)) +
  geom_line(color = 'orange', alpha = 0.5) +
  geom_line(data = aggregate(max_wind_speed ~ date, aemet_dataset, mean),
            aes(x = date, y = max_wind_speed), color = 'blue', alpha = 0.7) +
  labs(title = "max_wind_speed", x = "Date", y = "Max. wind speed (km/h)") +
  theme(plot.title = element_text(hjust = 0.5)) + theme_bw() +
  theme(text = element_text(size = 14))

insolation_plot <- ggplot(aemet_dataset, aes(x = date, y = insolation)) +
  geom_line(color = 'orange', alpha = 0.5) +
  geom_line(data = aggregate(insolation ~ date, aemet_dataset, mean),
            aes(x = date, y = insolation), color = 'blue', alpha = 0.7) +
  labs(title = "insolation", x = "Date", y = "Insolation (h)") +
  theme(plot.title = element_text(hjust = 0.5)) + theme_bw() +
  theme(text = element_text(size = 14))

avg_pressure_plot <- ggplot(aemet_dataset, aes(x = date, y = avg_pressure)) +
  geom_line(color = 'orange', alpha = 0.5) +
  geom_line(data = aggregate(avg_pressure ~ date, aemet_dataset, mean),
            aes(x = date, y = avg_pressure), color = 'blue', alpha = 0.7) +
  labs(title = "avg_pressure", x = "Date", y = "Avg. pressure (hPa)") +
  theme(plot.title = element_text(hjust = 0.5)) + theme_bw() +
  theme(text = element_text(size = 14))

grid.arrange(avg_temp_plot, precipitation_plot, avg_wind_plot,
             max_wind_plot,
             insolation_plot, avg_pressure_plot, ncol = 3)

```



Com era d'esperar, la variable *avg_daily_temp* mostra una fluctuació associada amb les estacions de l'any on l'hivern té les temperatures més baixes i l'estiu les més altes. Pel que fa a la *daily_precipitation*, s'observen alguns dies amb pluviometries elevades sobretot entre els mesos de Novembre i Març, en menor mesura en els mesos d'Abril a Juny. Tot i això, s'observa un descens de la precipitació des de Novembre del 2020 fins a l'Octubre del 2022 amb menys pics que sobrepassen els 50 mm. Pel que fa a les velocitats del vent, no s'observen variacions significatives en els valors promitjats en funció de la data, indicant que l'època de l'any no afectaria a l'electricitat produïda amb energia eòlica. La *insolation* fluctua amb la data com era d'esperar ja que a l'hivern hi ha menys dies de llum en comparació amb l'estiu on el dia dura més hores. Finalment, l'*avg_pressure* no mostra una tendència clara amb la data, demostrant clarament que no és estrictament necessari tenir baixes pressures perquè hi hagi precipitació i vent.

2.3 Unió dels dos datasets

Unim els dos datasets per tal de poder fer l'anàlisi de les dades. La columna que utilitzarem per fer la unió és la data:

```

# Unim els dos datasets
final_dataset <- left_join(esios_dataset, aemet_dataset, by = c("date"))

# Resum dataset
str(final_dataset)

## 'data.frame': 227760 obs. of 15 variables:
## $ date                  : Date, format: "2020-11-01" "2020-11-01" ...
## $ avg_total_price_eur_mwh : num 31.6 31.6 31.6 31.6 31.6 ...

```

```

## $ energy_total_mwh      : num  22639 22639 22639 22639 22639 ...
## $ energy_source         : Factor w/ 6 levels "hydroelectric",...
## $ renewable_generation_perc : num  6.6 6.6 6.6 6.6 6.6 ...
## $ renewable_generation_perc_bin: Factor w/ 4 levels "0-35","36-75",...
## $ province               : Factor w/ 52 levels "ACORUÑA","ALBACETE",...
## $ avg_daily_temp          : num  17.9 16.2 17.4 17.2 17.4 ...
## $ daily_precipitation     : num  7.66 0 0.05 0 0 ...
## $ avg_wind_speed          : num  7.55 1.4 1.68 1.68 5.8 ...
## $ max_wind_speed          : num  15.91 5.8 5.55 7.72 10.8 ...
## $ insolation              : num  0.181 8.988 9.381 7.565 7 ...
## $ avg_pressure             : num  997 943 1008 996 964 ...
## $ avg_pressure_bin         : Factor w/ 3 levels "Anticyclone",...
## $ insolation_bin           : Factor w/ 3 levels "0-5","10+","5-10": 3 2 2 2 2 3 2 2 2 3 ...

```

El dataset final té gairebé 228.000 files i 15 columnnes. D'aquestes, tenim una variable tipus *date*, 9 variables numèriques i 5 categòriques.

2.4 Guardar dataset final en un fitxer csv

Guardem el dataset final en un fitxer csv.

```
# Guardem el dataset final en un fitxer csv
write.csv(final_dataset, file = "data/final_dataset.csv", row.names = FALSE)
```

3. Neteja de les dades

3.1. Gestió dels zeros o elements buits

Degut a la necessitat de reduir la quantitat de dades com també assegurar que els datasets ESIOS i AEMET tenien la mateixa granularitat per poder-los unir, s'ha decidit tractar els valors nuls identificats durant el preprocessat (veure fitxer *PRA2_preprocessing.Rmd* i pdf associat). El tractament que s'ha fet, ha estat l'imputació de valors a partir de l'interpolació lineal, ja què estem treballant amb dades temporals. Per tant, el dataset en aquest punt ja no conté valors buits:

```
# Resum dels elements buits
summary(is.na(final_dataset))

##      date      avg_total_price_eur_mwh energy_total_mwh energy_source
## Mode :logical  Mode :logical      Mode :logical  Mode :logical
## FALSE:227760  FALSE:227760      FALSE:227760  FALSE:227760
## renewable_generation_perc renewable_generation_perc_bin province
## Mode :logical      Mode :logical      Mode :logical
## FALSE:227760      FALSE:227760      FALSE:227760
## avg_daily_temp daily_precipitation avg_wind_speed max_wind_speed
## Mode :logical  Mode :logical      Mode :logical  Mode :logical
## FALSE:227760  FALSE:227760      FALSE:227760  FALSE:227760
## insolation avg_pressure avg_pressure_bin insolation_bin
## Mode :logical  Mode :logical      Mode :logical  Mode :logical
## FALSE:227760  FALSE:227760      FALSE:227760  FALSE:227760
```

3.2. Identificació i gestió de valors extrems

Generalment, es consideren outliers tots aquells valors que es troben molt lluny dels valors de la resta de la mostra. En una distribució normal, això equival a $\pm 3\sigma$ respecte el valor promig. Per això, podem estandarditzar les columnnes numèriques i representar els boxplots:

```
# Creem variable amb columnes numèriques
num_vars <- c("avg_total_price_eur_mwh", "energy_total_mwh", "avg_daily_temp",
            "renewable_generation_perc", "daily_precipitation",
            "avg_wind_speed", "max_wind_speed", "insolation", "avg_pressure")

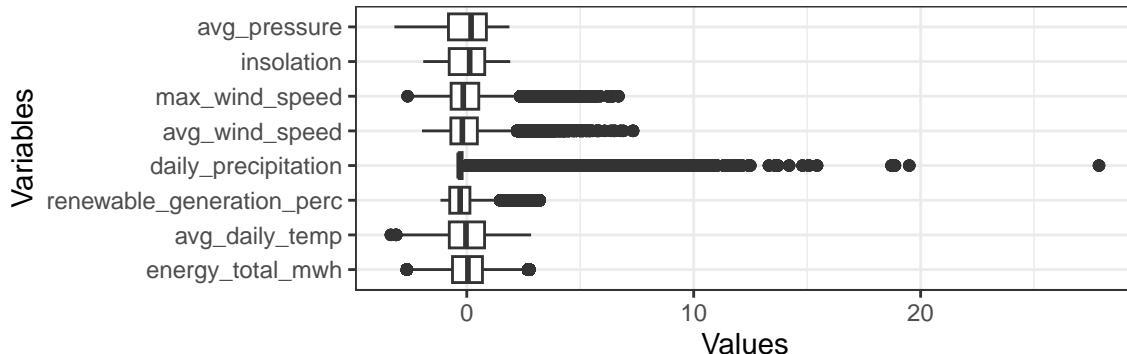
# Subset del dataset amb les columnnes numèriques sense la columna date
subset_final <- final_dataset %>% dplyr::select(all_of(num_vars))
subset_final <- subset_final[,-1]

# Estandarditzem les columnnes numèriques (entre -1 i 1)
normalized_data <- scale(subset_final, center = TRUE, scale = TRUE)

# Transposem les dades, eliminem la columna Var1 i renombrem les columnes
normalized_data <- melt(normalized_data)
normalized_data <- normalized_data[,-1]
colnames(normalized_data) <- c("Variables", "Values")

# Boxplots de les variables estandarditzades
ggplot(normalized_data, aes(y = Variables, x = Values)) +
  geom_boxplot() +
```

```
theme_bw() +
theme(text = element_text(size = 11))
```



A partir dels boxplots, podem veure que hi ha diferent nombre d'outliers en les variables numèriques. Les variables *avg_pressure* i *insolation* no tenen outliers, mentre què *avg_daily_temp*, *renewable_generation_perc*, *energy_total_mwh* i *avg_total_price_eur_mwh* en presenten alguns. Finalment, *max_wind_speed*, *avg_wind_speed* i *daily_precipitation* tenen un nombre molt elevat d'outliers, essent la darrera, aquella amb més valors atípics. Tanmateix, aquestes són dades que provenen de fonts fiables. A més, les dades d'origen meteorològic estan sotmeses a una gran aleatorietat deguda a la naturalesa caòtica d'aquest tipus de sistemes. D'altra banda, la fluctuació de les variables energètiques té també un alt component aleatori ja que estan influenciades per factors externs com la meteorologia, la demanda, l'especulació dels mercats, etc. Per tant, la presència d'outliers no està ocasionada per registres incorrectes sinó per la naturalesa de les dades i no els eliminarem de les nostres analisis.

4. Anàlisi de les dades

4.1. Anàlisi de la correlació entre variables numèriques

Selecció dels grups de dades i tipus d'anàlisi

Seleccionarem les variables numèriques *avg_total_price_eur_mwh*, *energy_total_mwh*, *renewable_generation_perc*, *avg_daily_temp*, *daily_precipitation*, *avg_wind_speed*, *max_wind_speed*, *insolation* i *avg_pressure* per crear una matriu de correlació lineal:

```
# Selecció dels grups de dades i tipus d'anàlisi
numeric_variables <- c("avg_total_price_eur_mwh", "energy_total_mwh", "renewable_generation_perc",
"avg_daily_temp", "daily_precipitation", "avg_wind_speed", "max_wind_speed",
"insolation", "avg_pressure")

# Creem un subset amb les variables numèriques
subset_numeric_variables <- final_dataset[, numeric_variables]
```

També estandarditzarem les variables per reduir la influència que poden tenir els diferents rangs de dades:

```
# Estandarditzem les variables numèriques
standard_num_vars <- scale(subset_numeric_variables)

# Unim el dataset standard_num_vars amb les columnes que mancan de final_dataset
standard_num_vars <- cbind(standard_num_vars,
final_dataset[,c("date", "energy_source", "avg_pressure_bin", "province",
"insolation_bin", "renewable_generation_perc_bin")])
```

```
summary(standard_num_vars)

##    avg_total_price_eur_mwh    energy_total_mwh    renewable_generation_perc
##  Min.   :-1.71877          Min.   :-2.63695          Min.   :-1.1580
##  1st Qu.:-0.86988          1st Qu.:-0.63006          1st Qu.: 0.7522
##  Median : 0.06963           Median : 0.05027           Median : 0.2918
##  Mean   : 0.00000           Mean   : 0.00000           Mean   : 0.0000
##  3rd Qu.: 0.75438           3rd Qu.: 0.69083           3rd Qu.: 0.1390
##  Max.   : 4.78197           Max.   : 2.75815           Max.   : 3.2055
##
##    avg_daily_temp    daily_precipitation    avg_wind_speed    max_wind_speed
##  Min.   :-3.34359          Min.   :-0.3306          Min.   :-1.9726          Min.   :-2.6083
##  1st Qu.:-0.77054          1st Qu.:-0.3306          1st Qu.:-0.7029          1st Qu.:-0.6895
##  Median : 0.03445           Median : 0.3306           Median : 0.1950           Median : 0.1548
##  Mean   : 0.00000           Mean   : 0.0000           Mean   : 0.0000           Mean   : 0.0000
##  3rd Qu.: 0.78427           3rd Qu.:-0.2046          3rd Qu.: 0.4652           3rd Qu.: 0.5293
##  Max.   : 2.83252           Max.   : 27.8596          Max.   : 7.3256           Max.   : 6.6866
##
##    insolation    avg_pressure        date
##  Min.   : 0.00000          Min.   : 0.00000          Min.   : 2010-01-01
##  1st Qu.: 0.00000          1st Qu.: 0.00000          1st Qu.: 2010-01-01
##  Median : 0.00000           Median : 0.00000          Median : 2010-01-01
##  Mean   : 0.00000           Mean   : 0.00000          Mean   : 2010-01-01
##  3rd Qu.: 0.00000           3rd Qu.: 0.00000          3rd Qu.: 2010-01-01
##  Max.   : 0.00000           Max.   : 0.00000          Max.   : 2010-01-01
```

```

## Min.   :-1.9182   Min.   :-3.1901   Min.   :2020-11-01
## 1st Qu.:-0.7746   1st Qu.:-0.8071   1st Qu.:2021-05-02
## Median : 0.1275   Median : 0.1891   Median :2021-10-31
## Mean   : 0.0000   Mean   : 0.0000   Mean   :2021-10-31
## 3rd Qu.: 0.7916   3rd Qu.: 0.8634   3rd Qu.:2022-05-02
## Max.   : 1.9141   Max.   : 1.8788   Max.   :2022-10-31
##
##          energy_source      avg_pressure_bin      province
## hydroelectric :37960  Anticyclone: 6864  A CORUÑA : 4380
## nuclear       :37960  Depression :211518  ALBACETE : 4380
## solar          :37960  Normal     : 9378  ALICANTE : 4380
## thermorenewable:37960                           ALMERIA  : 4380
## total          :37960                           ARABA/ALAVA: 4380
## wind           :37960                           ASTURIAS : 4380
##                                         (Other)    :201480
## insolation_bin renewable_generation_perc_bin
## 0-5   : 3780   0-35   :181532
## 10+  :161514   36-75  : 31980
## 5-10: 62466   76-100 : 13624
##                   Below 0:   624
##
##
```

Comprovació de la normalitat i homogeneïtat de la variància

Comprovem la normalitat de les variables numèriques meteorològiques *avg_daily_temp*, *daily_precipitation*, *avg_wind_speed*, *max_wind_speed*, *insolation* i *avg_pressure* fent servir gràfics Quantil-Quantil:

```

# Gràfic quantile-quantile
qq_adt <- ggplot(standard_num_vars, aes(sample = avg_daily_temp)) +
  stat_qq() + stat_qq_line() +
  labs(x="Teòrics", y="Reals",
       title="avg_daily_temp") + theme_bw() + theme(text = element_text(size = 14))

qq_dp <- ggplot(standard_num_vars, aes(sample = daily_precipitation)) +
  stat_qq() + stat_qq_line() +
  labs(x="Teòrics", y="Reals",
       title="daily_precipitation") + theme_bw() +
  theme(text = element_text(size = 14))

qq_aws <- ggplot(standard_num_vars, aes(sample = avg_wind_speed)) +
  stat_qq() + stat_qq_line() +
  labs(x="Teòrics", y="Reals",
       title="avg_wind_speed") + theme_bw() + theme(text = element_text(size = 14))

qq_mws <- ggplot(standard_num_vars, aes(sample = max_wind_speed)) +
  stat_qq() + stat_qq_line() +
  labs(x="Teòrics", y="Reals",
       title="max_wind_speed") + theme_bw() + theme(text = element_text(size = 14))

qq_i <- ggplot(standard_num_vars, aes(sample = insolation)) +
  stat_qq() + stat_qq_line() +
  labs(x="Teòrics", y="Reals",
       title="insolation") + theme_bw() + theme(text = element_text(size = 14))

qq_avg_p <- ggplot(standard_num_vars, aes(sample = avg_pressure)) +
  stat_qq() + stat_qq_line() +
  labs(x="Teòrics", y="Reals",
       title="avg_pressure") + theme_bw() + theme(text = element_text(size = 14))

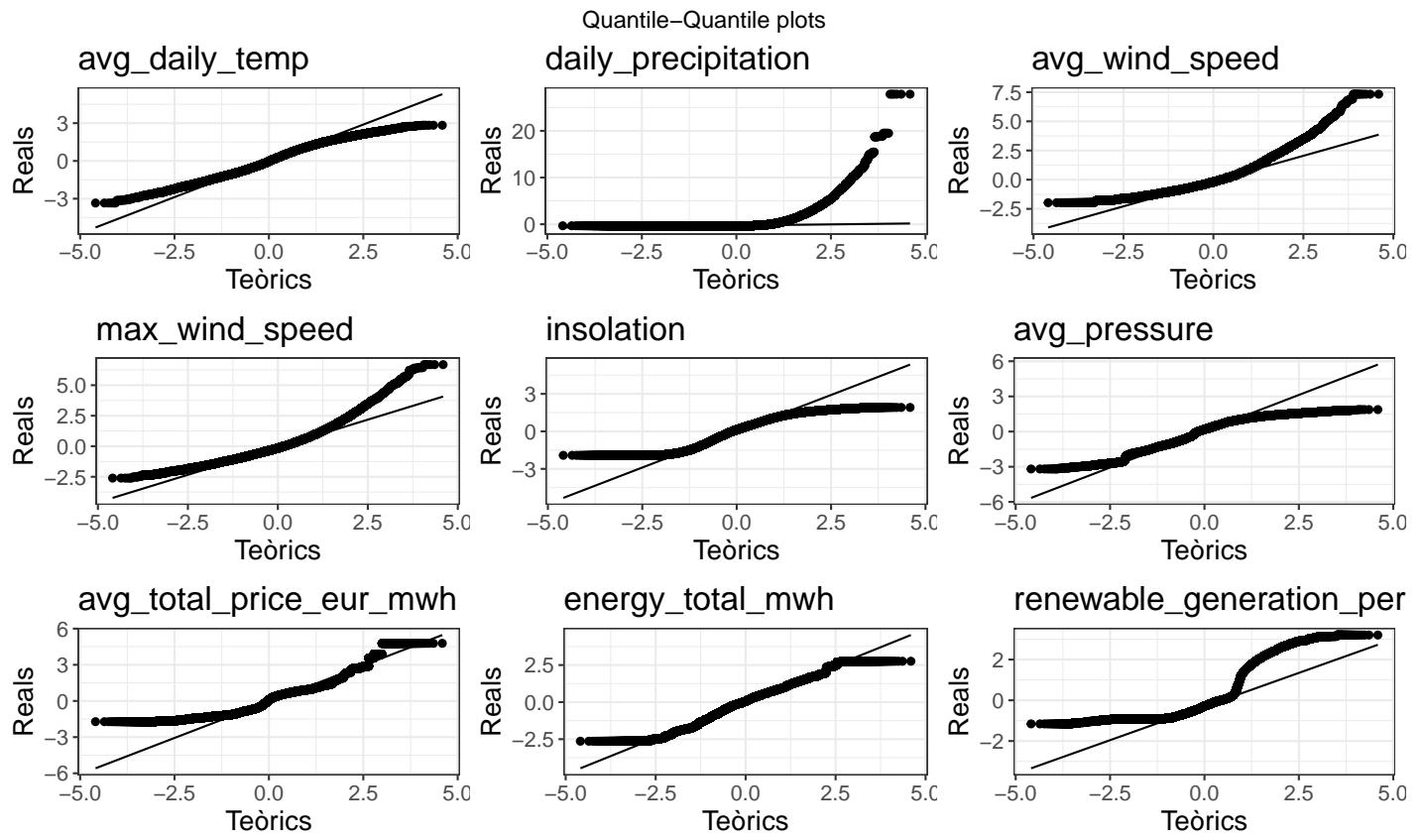
qq_avg_price <- ggplot(standard_num_vars, aes(sample = avg_total_price_eur_mwh)) +
  stat_qq() + stat_qq_line() +
  labs(x="Teòrics", y="Reals",
       title="avg_total_price_eur_mwh") + theme_bw() + theme(text = element_text(size = 14))

qq_et <- ggplot(standard_num_vars, aes(sample = energy_total_mwh)) +
  stat_qq() + stat_qq_line() +
  labs(x="Teòrics", y="Reals",
       title="energy_total_mwh") + theme_bw() + theme(text = element_text(size = 14))

qq_rgp <- ggplot(standard_num_vars, aes(sample = renewable_generation_perc)) +
  stat_qq() + stat_qq_line() +
  labs(x="Teòrics", y="Reals",
       title="renewable_generation_perc") + theme_bw() + theme(text = element_text(size = 14))

grid.arrange(qq_adt, qq_dp, qq_aws, qq_mws, qq_i, qq_avg_p, qq_avg_price, qq_et,
             qq_rgp, ncol=3, top="Quantile-Quantile plots")

```

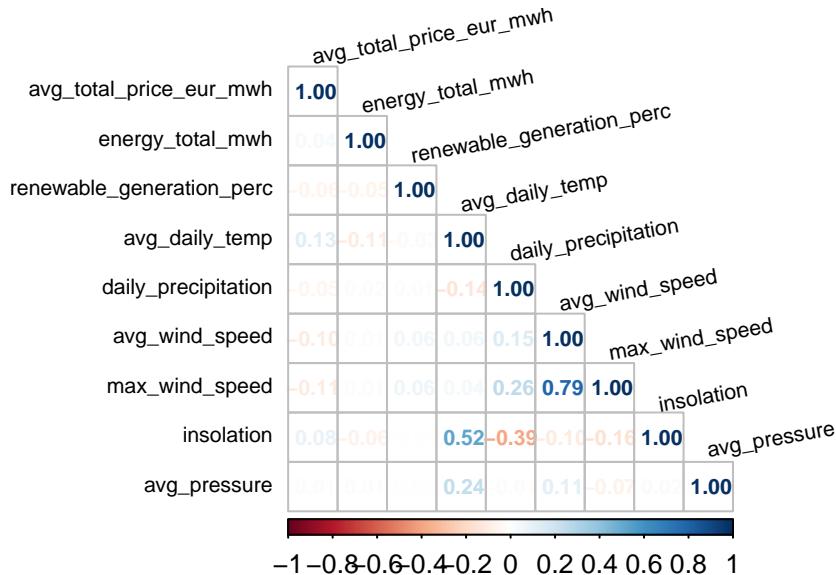


El *QQ-plot* permet observar com de propera és la distribució de les dades a la normal. És a dir, com més s'acosti la gràfica a la línia recta més normal serà la distribució. Per tant, podem dir què *avg_daily_temp* i *energy_total_mwh* segueixen una distribució pràcticament normal tot i què les dues tenen bastantes observacions (veure seccions 2.1.10 i 2.2.3). La resta de variables no presenten distribucions normals; l'*avg_total_price_eur_mwh* presenta dues distribucions, la primera amb un màxim entorn a 100 MW i la segona amb un màxim al voltant de 200 MW, mentre què *renewable_generation_perc* presenta una doble distribució, però aquesta té moltes més observacions en el rang de percentatges entre 0 i 35% (veure secció 2.1.10). La variable *insolation* té una distribució bastant uniforme amb un augment de les observacions en el rang de 5 a 8 hores, mentre que l'*avg_pressure* té quatre distribucions amb valors promitjats al voltant de 870, 920, 970 i 1000 hPa (veure secció 2.2.3). Les variables *daily_precipitation*, *avg_wind_speed* i *max_wind_speed* tenen distribucions esbiaxades cap a la dreta amb un efecte més pronunciat a *daily_precipitation* (veure secció 2.2.3).

Proves estadístiques

```
# Calculem la matrіu de correlació
corr_matrix_data <- cor(standard_num_vars[, numeric_variables],
                        use = "complete.obs", method=c("pearson"))

# Representem gràficament la matrіu de correlació
corrplot(corr_matrix_data, method = "number", type = "lower",
          tl.col = "black", tl.srt = 10, tl.cex = 0.7, number.cex = 0.7)
```



La matriu de correlació de les variables numèriques mostra una forta correlació lineal positiva entre les variables *max_wind_speed* i *avg_wind_speed*. Això és degut a què com més vent fa, major serà la velocitat màxima, a més, el càlcul de la velocitat promitja contindrà també els valors màxims. Per tant, en posteriors anàlisis es podrà descartar una de les dues. També s'observa una moderada dependència lineal positiva entre les variables *insolation* i *avg_daily_temp*. Això té sentit ja què els dies més càlids solen estar associats amb bon temps i aquests es troben normalment a l'estiu on la duració del dia és més llarga, augmentant les hores d'insolació. D'altra banda, els dies plujosos solen tenir una menor temperatura perquè sol ploure més a la primavera o tardor on les temperatures no són tan elevades com a l'estiu (lleugera correlació positiva entre *avg_daily_temp* i *daily_precipitation*). També s'observa una correlació negativa entre les variables *insolation* i *daily_precipitation*. Això és degut a què els dies plujosos, els núvols tapen el sol i, per tant, hi ha menys hores d'insolació. A més, també sol fer més vent, ja què hi ha certa correlació positiva entre *insolation* i *max_wind_speed*. Les variables *max_wind_speed* (i *avg_wind_speed*) i *daily_precipitation* tenen una correlació lleugerament positiva, ja què alguns dies plujosos sol fer vent, mentre què les variables *avg_pressure* i *avg_daily_temp* tenen una correlació lleugerament positiva, ja què els dies amb anticiclop solen ser dies més càlids, ja què els núvols bloquejen part de la radiació solar quan fa mal temps. També s'observa certa correlació positiva entre *avg_daily_temp* i *avg_total_price_eur_muh* indicant què el preu es pot veure lleugerament afectat per la temperatura. Això podria ser degut al fet què els dies més càlids la demanda sol incrementar per l'ús de l'aire acondicionat i, per tant, el preu del MWh augmenta. D'altra banda, s'observa certa correlació negativa entre les variables *max_wind_speed* i *avg_wind_speed* indicant que els dies ventosos disminueix el preu de l'electricitat. Això pot estar degut a què l'energia eòlica és renovable i té un preu de producció més barat.

4.2. Anàlisi de *renewable_generation_perc_bin* en funció de la meteorologia i *energy_source*

Selecció dels grups de dades i tipus d'anàlisi

Analitzarem la contribució de la meteorologia i el tipus de font d'energia en el percentatge de producció d'energia renovable. Per això, crearem un subset de les dades on s'exclourà el valor *energy_source=total*. D'altra banda, la variable *renewable_generation_perc_bin* conté 4 grups (veure gràfic de barres a la secció 2.1.10), però el grup *Below 0* es pot filtrar, ja què només afecta a l'energia hidroelèctrica i el grup *76-100* es correspon amb les files que tenen *energy_source=total*, ja què mai hi ha una font d'energia renovable que arribi a aquests percentatges per si sola. Per últim, descartarem la variable *avg_wind_speed* en el model ja què té una forta correlació amb la variable *max_wind_speed* (veure secció 4.1).

```
# Filtem les dades per energy_source != total i
#renewable_generation_perc_bin != Below 0 i 75-100
subset_energy_source_type <- standard_num_vars
subset_energy_source_type$energy_source <- factor(
  subset_energy_source_type$energy_source,
  levels = c("hydroelectric", "wind", "solar", "nuclear",
            "thermorenewable"))
subset_energy_source_type$renewable_generation_perc_bin <- factor(
  subset_energy_source_type$renewable_generation_perc_bin,
  levels = c("0-35", "36-75"))

# Eliminem nuls ja que l'eliminació dels factors només els transforma a NA,
# però les observacions segueixen presents
```

```

subset_energy_source_type <- subset_energy_source_type %>% drop_na()

summary(subset_energy_source_type)

##   avg_total_price_eur_mwh energy_total_mwh   renewable_generation_perc
##   Min.   :-1.718768      Min.   :-2.636953      Min.   :-0.9876
##   1st Qu.:-0.869876      1st Qu.:-0.626411     1st Qu.:-0.8256
##   Median : 0.066950      Median : 0.051348      Median : 0.4605
##   Mean    : 0.000132      Mean    : 0.005302      Mean    : -0.3935
##   3rd Qu.: 0.754381      3rd Qu.: 0.691982      3rd Qu.:-0.0651
##   Max.    : 4.781967      Max.    : 2.758153      Max.    : 1.7051
##
##   avg_daily_temp   daily_precipitation avg_wind_speed
##   Min.   :-3.343591      Min.   :-0.330619      Min.   :-1.972602
##   1st Qu.:-0.772590      1st Qu.:-0.330619      1st Qu.:-0.702891
##   Median :-0.038018      Median :-0.330619      Median :-0.195007
##   Mean   :-0.001025      Mean   :-0.000469      Mean   :-0.000921
##   3rd Qu.: 0.784274      3rd Qu.: 0.204582      3rd Qu.: 0.461336
##   Max.   : 2.832516      Max.   : 27.859627      Max.   : 7.325589
##
##   max_wind_speed   insolation   avg_pressure
##   Min.   :-2.608274      Min.   :-1.9181696     Min.   :-3.190125
##   1st Qu.:-0.689530      1st Qu.:-0.7724794     1st Qu.:-0.807095
##   Median :-0.154798      Median : 0.1274685     Median : 0.189384
##   Mean   :-0.001149      Mean   : 0.0007532     Mean   : 0.000159
##   3rd Qu.: 0.528222      3rd Qu.: 0.7927436     3rd Qu.: 0.863562
##   Max.   : 6.686628      Max.   : 1.9140810     Max.   : 1.878834
##
##   date           energy_source   avg_pressure_bin
##   Min.   :2020-11-01  hydroelectric :37336  Anticyclone: 5709
##   1st Qu.:2021-05-02  wind        :37960  Depression :175677
##   Median :2021-10-31  solar       :37960  Normal     : 7790
##   Mean   :2021-10-31  nuclear     :37960
##   3rd Qu.:2022-05-01  thermorenewable:37960
##   Max.   :2022-10-31
##
##   province   insolation_bin renewable_generation_perc_bin
##   ACORUÑA    : 3638  0-5 : 3140  0-35 :181532
##   ALBACETE   : 3638  10+ :134193 36-75:  7644
##   ALICANTE   : 3638  5-10: 51843
##   ALMERIA    : 3638
##   ARABA/ALAVA: 3638
##   ASTURIAS   : 3638
##   (Other)    :167348

```

L'anàlisi que aplicarem serà una regressió logística multivariable on *renewable_generation_perc_bin* serà la variable objectiu i contindrà dos grups (0-35 i 36-75).

Comprovació de la normalitat i homogeneïtat de la variància

La normalitat de les variables numèriques s'ha comprovat en l'apartat 4.1. Per tant, no la tornarem a comprovar. En aquest cas, ens centrarem en les variables categòriques: *energy_source* té una distribució de comptes uniforme per les diferents categories tal i com mostra el resum de més amunt, mentre què *renewable_generation_perc_bin* està desbalancejada amb la majoria d'observacions (~96%) en el rang 0-35. Per això, aplicarem la tècnica de *downsampling*, seleccionant un subset de les dades per tal de contrarestar aquest desbalanç:

```

# Apliquem el downsampling al dataset
downsampled_data <- downSample(x = subset_energy_source_type,
  !colnames(subset_energy_source_type) %in% "renewable_generation_perc_bin"),
y = subset_energy_source_type$renewable_generation_perc_bin,
yname = "renewable_generation_perc_bin")

# Mostrem resultats del downsampling per la variable renewable_generation_perc_bin
table(downsampled_data$renewable_generation_perc_bin)

```

```

## 
## 0-35 36-75
## 7644 7644

```

Comprovem l'homogeneitat de les variàncies de les variables numèriques respecte la variable *renewable_generation_perc_bin*:

```

# Comprovem la homogeneitat de les variàncies
leveneTest(downsampled_data$avg_daily_temp ~ downsampled_data$renewable_generation_perc_bin)

```

```

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     1 599.23 < 2.2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
leveneTest(downsampled_data$daily_precipitation ~ downsampled_data$renewable_generation_perc_bin)

```

```

## Levene's Test for Homogeneity of Variance (center = median)

```

```

##          Df F value    Pr(>F)
## group      1 292.01 < 2.2e-16 ***
##        15286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
leveneTest(downscaled_data$max_wind_speed ~ downsampled_data$renewable_generation_perc_bin)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group      1 409.17 < 2.2e-16 ***
##        15286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
leveneTest(downscaled_data$insolation ~ downsampled_data$renewable_generation_perc_bin)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group      1 54.773 1.423e-13 ***
##        15286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
leveneTest(downscaled_data$avg_pressure ~ downsampled_data$renewable_generation_perc_bin)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group      1 2.0015 0.1572
##        15286

```

El test de Levene rebutja la hipòtesi nul·la per les variables numèriques que volem analitzar respecte la variable *renewable_generation_perc_bin*, per tant, les distribucions no provenen de mostres amb variàncies uniformes. Tot i els resultats obtinguts que mostren que les distribucions no són normals i la no uniformitat de la variància, la regressió logística és menys sensible a les condicions de normalitat i homoscedasticitat, per tant, procedirem amb la seva implementació. [8]

Proves estadístiques

```

# Calculem la regressió logística entre les variables
logistic_renew_perc <- glm(renewable_generation_perc_bin ~ avg_daily_temp +
                           daily_precipitation + max_wind_speed +
                           insolation + avg_pressure + energy_source,
                           data = downsampled_data,
                           family = binomial(link = logit))

summary(logistic_renew_perc)

##
## Call:
## glm(formula = renewable_generation_perc_bin ~ avg_daily_temp +
##       daily_precipitation + max_wind_speed + insolation + avg_pressure +
##       energy_source, family = binomial(link = logit), data = downsampled_data)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q     Max 
## -3.0164 -0.0001  0.0275  0.4285  3.2602 
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -20.06791  247.82280 -0.081  0.93546  
## avg_daily_temp -0.66026   0.03864 -17.089 < 2e-16 ***
## daily_precipitation -0.07890   0.03056 -2.582  0.00983 ** 
## max_wind_speed  0.71370   0.03365  21.207 < 2e-16 ***
## insolation    -0.38134   0.03828 -9.962 < 2e-16 ***
## avg_pressure   0.28850   0.03340  8.637 < 2e-16 ***
## energy_sourcewind 21.29445  247.82280  0.086  0.93153  
## energy_sourcesolar 17.30911  247.82281  0.070  0.94432  
## energy_sourcenuclear -0.03050  349.84261  0.000  0.99993  
## energy_sourcethermorenewable -0.05999  348.93531  0.000  0.99986 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 21193.7 on 15287 degrees of freedom
## Residual deviance: 7098.2 on 15278 degrees of freedom
## AIC: 7118.2
##
## Number of Fisher Scoring iterations: 18

```

El model obtingut mostra que les variables amb més pes són, en primer lloc, l'energia eòlica seguida de la solar ja que tenen coeficients alts, respecte l'energia hidroelèctrica. L'energia nuclear i la termorenewable no tenen un pes significatiu amb coeficients propers a 0. Tot i això, cal tenir en compte que aquestes variables tenen valors p propers a 1, fet que indicaria que la relació entre aquestes variables i *renewable_generation_perc_bin* no és significativa. Pel que fa a les variables numèriques,

l'`avg_daily_temp`, la `max_wind_speed`, l'`insolation` i l'`avg_pressure` semblen tenir un pes relativament alt a l'hora de determinar el valor de `renewable_generation_perc_bin`. Els seus coeficients tenen valors p inferiors a 0.05, per tant, la relació és significativa. Finalment, la variable `daily_precipitation` té poc pes en la predició de `renewable_generation_perc_bin`.

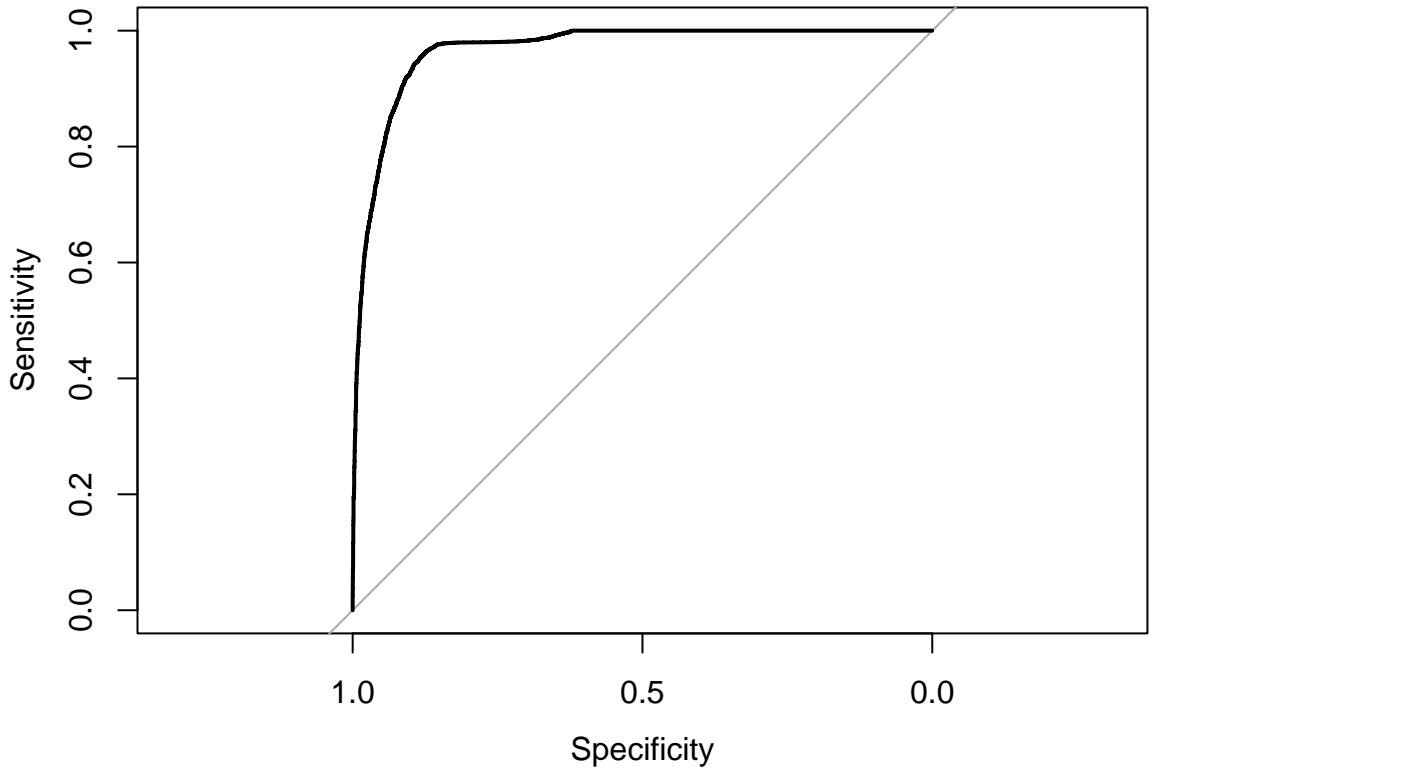
Calculem la matriu de confusió i la corba ROC per avaluar els resultats del model:

```
# Generem predictions per poder calcular la matriu de confusió
predictions_log <- predict(logistic_renew_perc, data = downsampled_data,
                           type = 'response')

# Establim probabilitat per assignar els dos grups i assignem els noms originals
# a les prediccions
predictions_binary <- as.factor(predictions_log>0.5)
predicition_names <- levels(
  downsampled_data$renewable_generation_perc_bin)[predictions_binary]

# Calculem la matriu de confusió
confusionMatrix(as.factor(predicition_names),
                 downsampled_data$renewable_generation_perc_bin)

## Confusion Matrix and Statistics
##
##          Reference
## Prediction 0-35 36-75
##      0-35   6556   200
##      36-75  1088  7444
##
##              Accuracy : 0.9158
##              95% CI : (0.9112, 0.9201)
##      No Information Rate : 0.5
##      P-Value [Acc > NIR] : < 2.2e-16
##
##              Kappa : 0.8315
##
## McNemar's Test P-Value : < 2.2e-16
##
##              Sensitivity : 0.8577
##              Specificity : 0.9738
##      Pos Pred Value : 0.9704
##      Neg Pred Value : 0.8725
##              Prevalence : 0.5000
##      Detection Rate : 0.4288
##      Detection Prevalence : 0.4419
##      Balanced Accuracy : 0.9158
##
##      'Positive' Class : 0-35
##
# Càcul de la corba ROC i de la AUC
r_log = roc(downscaled_data$renewable_generation_perc_bin,
             as.numeric(predictions_log),
             data = downsampled_data, auc = TRUE, plot = TRUE)
```



r_log

```
## 
## Call:
## roc.default(response = downsampled_data$renewable_generation_perc_bin,      predictor = as.numeric(predictions_log), auc = TRUE, plot = TRUE,
## 
## Data: as.numeric(predictions_log) in 7644 controls (downsampled_data$renewable_generation_perc_bin 0-35) < 7644 cases (downsampled_data$renewable_generation_perc_bin 36-75)
## Area under the curve: 0.9673
```

La matriu de confusió mostra com el model de regressió logística és capaç de discriminar correctament els valors de classe, ja que té una *accuracy* del 91.5% i un interval de confiança de [0.91, 0.92]. La *sensitivity* ens indica la proporció de valors positius (36-75 en aquest cas) classificats correctament, mentre que la *specificity* ens indica la proporció de valors negatius (0-35 en el nostre cas) que han estat classificats correctament. Si ens fixem en els valors de sensibilitat (0.86) i la especificitat (0.98), veiem que el model classifica millor la classe 0-35 tot i haver utilitzat un *downsampling* de la classe majoritària. L'àrea sota la corba té un valor de 0.92, indicant que el model seria capaç de classificar correctament entre les diferents categories de *renewable_generation_perc_bin*.

4.3. Anàlisi i contrast d'hipòtesis de *avg_total_price_eur_mwh* respecte els grups de *insolation_bin* i *avg_pressure_bin*

Selecció dels grups de dades i tipus d'anàlisi

Analitzarem el preu mitjà total en eur/mwh *avg_total_price_eur_mwh* sobre les variables de *insolation_bin* i *avg_pressure_bin*. La variable *insolation_bin* determinarà la quantitat de radiació solar comptabilitzada en una àrea determinada, mentre que la variable *avg_pressure_bin* determinarà la pressió atmosfèrica capturada en la mateixa àrea. Amb aquesta anàlisis es vol analitzar si una major pressió atmosfèrica i radiació solar equivalen a un menor preu de l'energia.

Aplicarem un contrast d'hipòtesi per tal d'analitzar les diferències potencials en la variable *avg_total_price_eur_mwh* respecte els diferents grups de les variables *insolation_bin* i *avg_pressure_bin*.

Comprovació de la normalitat i homogeneïtat de la variància

La normalitat de les variables numèriques es pot comprovar en l'apartat 4.1 on s'ha vist que la variable *avg_total_price_eur_mwh* no tenia una distribució normal. Tanmateix, en aquest cas, es comprovarà la normalitat per a cadascun dels grups de les variables *insolation_bin* i *avg_pressure_bin*. Per això, dividirem la variable *avg_total_price_eur_mwh* per grups:

```
# Dividim avg_total_price_eur_mwh en grups de avg_pressure_bin
pressure_groups <- split(standard_num_vars, standard_num_vars$avg_pressure_bin)

# Dividim avg_total_price_eur_mwh en grups de insolation_bin
insolation_groups <- split(standard_num_vars, standard_num_vars$insolation_bin)
```

Apliquem el test de normalitat *Anderson-Darling* respecte *avg_pressure_bin*. Aquest test és aplicable a mostres grans, ja que altres tests com *Shapiro* són més conservadors i poden rebutjar la hipòtesi nul · la malgrat la distribució de les dades sigui normal.

```
# Anderson-Darling test de avg_total_price_eur_mwh respecte avg_pressure_bin
for (i in 1:length(pressure_groups)) {
  print(ad.test(pressure_groups[[i]]$avg_total_price_eur_mwh))
}

## 
## Anderson-Darling normality test
##
## data: pressure_groups[[i]]$avg_total_price_eur_mwh
## A = 232.94, p-value < 2.2e-16
##
##
## Anderson-Darling normality test
##
## data: pressure_groups[[i]]$avg_total_price_eur_mwh
## A = 3571.6, p-value < 2.2e-16
##
##
## Anderson-Darling normality test
##
## data: pressure_groups[[i]]$avg_total_price_eur_mwh
## A = 173.21, p-value < 2.2e-16
```

Els resultats mostren com el valor p és inferior a 0.05, per tant, rebutjant la hipòtesi nul · la que confirmaria la normalitat de les distribucions. És a dir, les dades no estan distribuïdes normalment.

Apliquem el test de normalitat *Anderson-Darling* respecte *insolation_bin*:

```
# Anderson-Darling test de avg_total_price_eur_mwh respecte insolation_bin
for (i in 1:length(insolation_groups)) {
  print(ad.test(insolation_groups[[i]]$avg_total_price_eur_mwh))
}

## 
## Anderson-Darling normality test
##
## data: insolation_groups[[i]]$avg_total_price_eur_mwh
## A = 158.77, p-value < 2.2e-16
##
##
## Anderson-Darling normality test
##
## data: insolation_groups[[i]]$avg_total_price_eur_mwh
## A = 2715.2, p-value < 2.2e-16
##
##
## Anderson-Darling normality test
##
## data: insolation_groups[[i]]$avg_total_price_eur_mwh
## A = 1632.7, p-value < 2.2e-16
```

Altra vegada, es rebutja la hipòtesi nul · la, indicant que les distribucions dels diferents grups no són normals.

Tot seguit, comprovem l'homogeneitat de les variàncies de *avg_total_price_eur_mwh* respecte els grups de les variables *insolation_bin* i *avg_pressure_bin*. Com hem vist, la variable no té una distribució normal, per tant, haurem d'aplicar el test de Levene que no es veu afectat pel tipus de distribució:

```
# Levene test de avg_total_price_eur_mwh respecte avg_pressure_bin
leveneTest(avg_total_price_eur_mwh ~ avg_pressure_bin, data = standard_num_vars)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group      2  67.889 < 2.2e-16 ***
##           227757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Levene test de avg_total_price_eur_mwh respecte insolation_bin
leveneTest(avg_total_price_eur_mwh ~ insolation_bin, data = standard_num_vars)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group      2 1989.4 < 2.2e-16 ***
##           227757
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Els resultats retornen valors p menors a 0.05, rebutjant la hipòtesi nul·la. Això vol dir què les variàncies de *avg_total_price_eur_mwh* respecte els diferents grups de *insolation_bin* i *avg_pressure_bin* no són homogènies.

Proves estadístiques

Volem comprovar si les distribucions dels diferents grups de *avg_total_price_eur_mwh* en funció de *insolation_bin* i *avg_pressure_bin* són iguals o no per analitzar la influència d'aquestes variables en el preu total de l'electricitat. Com que la distribució de *avg_total_price_eur_mwh* no és normal i les variàncies no són uniformes, aplicarem el test de *Kruskal-Wallis* que és un test paramètric que compara les medianes de les distribucions per dos o més grups. Per tant, és adequat per aquest tipus de dades:

```
kruskal.test(avg_total_price_eur_mwh ~ insolation_bin, data = standard_num_vars)
```

```
## 
## Kruskal-Wallis rank sum test
##
## data: avg_total_price_eur_mwh by insolation_bin
## Kruskal-Wallis chi-squared = 2145.8, df = 2, p-value < 2.2e-16
kruskal.test(avg_total_price_eur_mwh ~ avg_pressure_bin, data = standard_num_vars)

## 
## Kruskal-Wallis rank sum test
##
## data: avg_total_price_eur_mwh by avg_pressure_bin
## Kruskal-Wallis chi-squared = 104.42, df = 2, p-value < 2.2e-16
```

Observem doncs que els valors p dels tests respecte les variables *insolation_bin* i *avg_pressure_bin* són inferiors a 0.05. Per tant, hi ha una diferència prou significativa entre les medianes dels diferents grups, indicant que *insolation_bin* i *avg_pressure_bin* tenen una influència a l'hora de determinar el preu de l'electricitat.

4.4. Anàlisi de *avg_total_price_eur_mwh* respecte *renewable_generation_perc* i *energy_source*

Analitzarem la contribució del tipus de font d'energia renovable en el preu de l'electricitat. Per això, crearem un subset de les dades on s'exclourà el valor *energy_source*=*total*.

```
# Filtrem les dades per energy_source != total
subset_energy_source_type_price <- dplyr::select(standard_num_vars, energy_source,
                                                renewable_generation_perc,
                                                avg_total_price_eur_mwh)
subset_energy_source_type_price$energy_source <- factor(
  subset_energy_source_type_price$energy_source,
  levels = c("hydroelectric", "wind", "solar", "nuclear",
            "thermorenewable"))

# Eliminem nuls ja que l'eliminació dels factors només els transforma a NA,
# però les observacions segueixen presents
subset_energy_source_type_price <- subset_energy_source_type_price %>% drop_na()
summary(subset_energy_source_type_price)

##          energy_source  renewable_generation_perc avg_total_price_eur_mwh
## hydroelectric    :37960   Min.   :-1.15801      Min.   :-1.71877
## wind           :37960   1st Qu.: -0.82792      1st Qu.: -0.86988
## solar          :37960   Median : -0.46225      Median :  0.06963
## nuclear         :37960   Mean   : -0.39568      Mean   :  0.00000
## thermorenewable:37960   3rd Qu.: -0.06574      3rd Qu.:  0.75438
##                  Max.   :  1.70511      Max.   :  4.78197
```

Comprovació de la normalitat i homogeneïtat de la variància

La normalitat de les variables numèriques es pot comprovar en l'apartat 4.1 on s'ha vist que la variable *avg_total_price_eur_mwh* no tenia una distribució normal. Tanmateix, en aquest cas, es comprovarà la normalitat per a cadascun dels grups de *energy_source*:

```
# Dividim avg_total_price_eur_mwh en grups de energy_source
energy_groups <- split(subset_energy_source_type_price,
                        subset_energy_source_type_price$energy_source)
```

Apliquem el test de normalitat *Anderson-Darling* respecte *energy_source*, tal i com s'ha fet en l'apartat 4.3:

```
# Anderson-Darling test de avg_total_price_eur_mwh respecte energy_source
for (i in 1:length(energy_groups)) {
  print(ad.test(energy_groups[[i]]$avg_total_price_eur_mwh))
}
```

```
##  
## Anderson-Darling normality test  
##  
## data: energy_groups[[i]]$avg_total_price_eur_mwh  
## A = 654.84, p-value < 2.2e-16  
##  
##  
## Anderson-Darling normality test  
##  
## data: energy_groups[[i]]$avg_total_price_eur_mwh  
## A = 654.84, p-value < 2.2e-16  
##  
##  
## Anderson-Darling normality test  
##  
## data: energy_groups[[i]]$avg_total_price_eur_mwh  
## A = 654.84, p-value < 2.2e-16  
##  
##  
## Anderson-Darling normality test  
##  
## data: energy_groups[[i]]$avg_total_price_eur_mwh  
## A = 654.84, p-value < 2.2e-16  
##  
##  
## Anderson-Darling normality test  
##  
## data: energy_groups[[i]]$avg_total_price_eur_mwh  
## A = 654.84, p-value < 2.2e-16
```

Els resultats retornen valors p menors a 0.05, rebutjant la hipòtesi nul·la i indicant que les distribucions de *avg_total_price_eur_mwh* respecte *energy_source* no segueixen una distribució normal.

Tot seguit, comprovem l'homogeneitat de les variàncies de *avg_total_price_eur_mwh* respecte els grups de *energy_source*. Com que les dades no segueixen una distribució normal, aplicarem el test de Levene:

```
# Levene test de avg_total_price_eur_mwh respecte energy_source
leveneTest(avg_total_price_eur_mwh ~ energy_source,
           data = subset_energy_source_type_price)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group      4      0     1
##             189795
```

Els resultats retornen un valor p major a 0.05, acceptant la hipòtesi nul·la. Això vol dir que les variàncies de *avg_total_price_eur_mwh* respecte els diferents grups de *energy_source* són homogènies.

Proves estadístiques

Volem comprovar analitzar la influència de la font d'energia renovable en el preu del MWh. Com que la distribució de les dades no és normal, aplicarem un mètode robust de regressió, *Least Absolute Deviations* que és adequat per distribucions no normals i amb valors atípics:

```

# Regressió LAD de energy_source i renewable_generation_perc respecte
# avg_total_price_eur_mwh
lad <- rlm(avg_total_price_eur_mwh ~ renewable_generation_perc + energy_source
            data = subset_energy_source_type_price)

summary(lad)

##
## Call: rlm(formula = avg_total_price_eur_mwh ~ renewable_generation_perc +
##           energy_source, data = subset_energy_source_type_price)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8355 -0.8466  0.1083  0.7704  4.8352
##
## Coefficients:
##                               Value     Std. Error t value
## (Intercept)             -0.2668    0.0067  -39.6933
## renewable_generation_perc -0.3894    0.0076  -51.3144
## energy_sourcewind        0.2511    0.0085  29.5608
## energy_sourcesolar       0.0440    0.0070   6.2588
## energy_sourcenuclear     0.2284    0.0083  27.6009
## energy_sourcethermorenewable -0.1181   0.0073  -16.0706
##
## Residual standard error: 1.199 on 189794 degrees of freedom

```

Els resultats del model indiquen que el percentatge d'energia renovable contribueix a reduir el preu del MWh. De les diferents fonts, l'energia èolica i la nuclear tenen una influència més elevada en el preu de l'electricitat respecte el valor de referència (l'energia hidroelèctrica), ja que tenen coeficients positius i amb valors més grans que la resta de fonts d'energia.

Aquests coeficients positius semblen indicar que el preu augmentaria. Això podria ser degut a la fluctuació en la producció eòlica i al preu més car de la producció d'energia nuclear respecte la resta de renovables. També podria haver-hi una influència d'algún factor que no s'ha tingut en compte com la producció d'energia en les centrals de cicle combinat que consumeixen gas natural.

Resolució del problema: Conclusions

A l'inici de la pràctica ens plantejàvem les següents hipòtesis a resoldre:

- Analitzar l'evolució del preu de l'energia elèctrica a Espanya en el període de temps mencionat.
- Entendre la relació entre el preu de l'energia elèctrica, la producció d'energies renovables i la influència de la meteorologia.
- Discernir quina font d'energia renovable té un impacte més rellevant en el preu de l'energia.

Dels resultats obtinguts al llarg de la pràctica podem determinar les següents conclusions tenint en compte les hipòtesis anteriors:

- L'evolució del preu de l'energia presenta una tendència a l'alça en el període de temps analitzat en aquesta pràctica. S'observa un augment progressiu, essent el valor mitjà del preu de l'energia (eur/mwh) el més baix capturat el dia 31 de Gener de 2021 a 6.9 eur/mwh i el més alt el dia 8 de Març de 2022 a 556.07 eur/mwh.
- S'observa una forta correlació entre la influència de la meteorologia i la producció d'energies renovables. Per exemple, les províncies que presenten temperatures mitjanes més elevades, un menor percentatge de precipitació i un major nombre en la variable insolació, tendiran a tenir una producció d'energia solar més elevada. Així mateix, les províncies amb una mitjana o màxima de velocitat del vent més elevades tendiran a tenir una producció d'energia eòlica superior.
- Veiem com l'eòlica i la solar són les energies renovables amb un impacte més rellevant sobre la producció d'energia renovables respecte l'energia hidroelèctrica, nuclear i termorenovable.
- També hem observat que la meteorologia té una certa influència en el preu del MWh, en particular la insolació i la pressió promig, tal i com mostra el test de Kruskal-Wallis.
- El percentatge d'energia renovable contribueix a reduir el preu del MWh. De les diferents fonts, l'energia eòlica i la nuclear tenen una influència més elevada en el preu de l'electricitat degut a la fluctuació en la producció eòlica i al preu més car de la producció d'energia nuclear respecte la resta de renovables.
- Finalment, en aquest anàlisi podria haver-hi una influència d'algún factor que no s'ha tingut en compte com la producció d'energia en les centrals de cicle combinat que consumeixen gas natural, el preu del qual s'ha disparat en els últims mesos o a l'especulació dels mercats.

Contribucions

```
df <- data.frame(Contribucions=rep(c('Investigació prèvia',
                                      'Redacció de les respostes',
                                      'Desenvolupament del codi',
                                      'Participació al vídeo'), each=1),
                  Signatura=rep(c('Albert Queraltó, Esther Manzano',
                                  'Albert Queraltó, Esther Manzano',
                                  'Albert Queraltó, Esther Manzano',
                                  'Albert Queraltó, Esther Manzano'), times=1))

df

##           Contribucions          Signatura
## 1   Investigació prèvia Albert Queraltó, Esther Manzano
## 2 Redacció de les respostes Albert Queraltó, Esther Manzano
## 3 Desenvolupament del codi Albert Queraltó, Esther Manzano
## 4   Participació al vídeo Albert Queraltó, Esther Manzano
```

Referències

[1] Alexis Romero, “La guerra a Ucraïna i la creixent inflació ressusciten el debat sobre la intervenció del mercat elèctric” (Público, 01/03/2022). <https://www.publico.es/public/encariment-preus-guerra-ucraina-i-creixent-inflacio-ressusciten-debat-intervencio-mercat-electric.html>

- [2] Vicenç Batalla, “La UE frenarà la pujada de l'electricitat per la guerra” (El Punt Avui, 12/03/2022). <https://www.elpuntavui.cat/politica/article/17-politica/2111814-la-ue-frenara-la-pujada-de-l-electricitat-per-la-guerra.html>
- [3] elnacional.cat, “La situació actual de les energies renovables a Catalunya: què tenim i cap on anem?” (El Nacional, 06/06/2022). https://www.elnacional.cat/ca/branded/energies-renewables-catalunya-cap-anem_685709_102.html
- [4] Ourworldindata.org, “La caiguda del preu de l'electricitat a partir de fonts renovables” (Ourworldindata.org, 2022). <https://ca.dsnsolar.com/info/the-price-decline-of-electricity-from-renewabl-70291585.html>
- [5] naciodigital.cat, “Tindrem prou energia només amb les renovables?” (Nació Digital, 2022). <https://www.naciodigital.cat/noticia/230097/canvi-climatic-tindrem-prou-energia-nomes-renovables>
- [6] Esther Manzano i Albert Queraltó, “Webscraping of Energy Prices and Renewable Energy Production” (GitHub, 2022). <https://github.com/albert-queralto/scraping-energy-prices-spain>
- [7] Esther Manzano i Albert Queraltó, “Evolution of the price of electricity and the renewable energy production in Spain” (Zenodo, 2022). <https://doi.org/10.5281/zenodo.7335618>
- [8] Montserrat Guillén Estany i María Teresa Alonso Alonso, “Models de regressió logística”, UOC (2020).