Descriptive Analysis

Eric Hausken

Table of contents

Red Eléctrica electricity generation

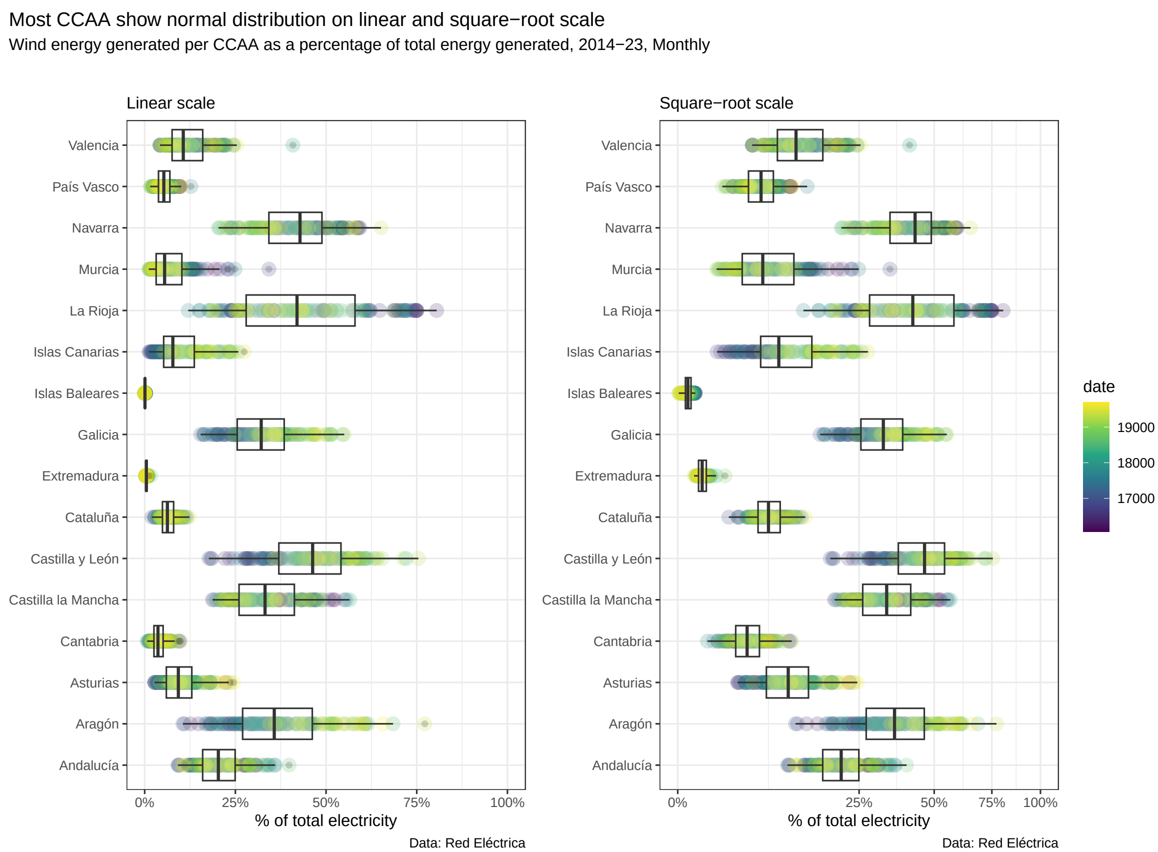
I have already pulled the amount of wind energy generated as a percentage of all electricity for each autonomous community of Spain since 2014. You can see the summary below:

```
red_data <- read_csv("red_data.csv")
wind_data <- red_data |>
  filter(name == "Eólica")

summary(factor(red_data$name) )
```

```
       Eólica Generación total
         1856             2280
```

```
library(patchwork)

plta <- wind_data |>
  ggplot(aes(x = percentage, y =ccaa, color = date)) +
  geom_point(size =4, alpha = .2) +
  geom_boxplot(alpha = .2) +
  scale_x_continuous(labels = scales::label_percent(),
                     limits = c(0,1),
  ) +
 scale_color_continuous(type = "viridis") +
  theme_bw() +
  labs(
    x = "% of total electricity",
    y = NULL,
    subtitle = "Linear scale",
    title = "",
    caption = "Data: Red Eléctrica"
  )

pltb <- wind_data |>
  ggplot(aes(x = percentage, y =ccaa, color = date)) +
  geom_point(size =4, alpha = .2) +
  geom_boxplot(alpha = .2) +
  scale_x_continuous(labels = scales::label_percent(),
                     limits = c(0,1),
                     trans = "sqrt"
  ) +
```

```
  scale_color_continuous(type = "viridis") +
  theme_bw() +
  labs(
    x = "% of total electricity",
    y = NULL,
    subtitle = "Square-root scale",
    title = "",
    caption = "Data: Red Eléctrica"
  )

plta + pltb + plot_layout(guides = "collect") +
  plot_annotation(subtitle = "Wind energy generated per CCAA as a percentage of total ener
                  title = "Most CCAA show normal distribution on linear and square-root sc
```



ANOVA tests

```r
aov1 <- aov(data = wind_data, percentage ~ ccaa + factor(date) )
summary(aov1)
```

```
               Df Sum Sq Mean Sq F value Pr(>F)    
ccaa           15  47.69   3.179 550.732 <2e-16 ***
factor(date)  119   3.41   0.029   4.969 <2e-16 ***
Residuals    1721   9.93   0.006                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
aov2 <- aov(data = wind_data, percentage ~ ccaa + date )
summary(aov2)
```

```
              Df Sum Sq Mean Sq F value  Pr(>F)    
ccaa          15  47.69   3.179   444.1 < 2e-16 ***
date           1   0.18   0.184    25.7 4.4e-07 ***
Residuals   1839  13.16   0.007                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
aov3 <- aov(data = wind_data, percentage ~ ccaa + lubridate::year(date) )
summary(aov3)
```

```
                        Df Sum Sq Mean Sq F value   Pr(>F)    
ccaa                    15  47.69   3.179  444.82  < 2e-16 ***
lubridate::year(date)    1   0.20   0.205   28.62 9.92e-08 ***
Residuals             1839  13.14   0.007                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
aov4 <- aov(data = wind_data, percentage ~ ccaa + factor(lubridate::year(date)) )
summary(aov4)
```

```
                                 Df Sum Sq Mean Sq F value  Pr(>F)    
ccaa                             15  47.69   3.179 451.238 < 2e-16 ***
factor(lubridate::year(date))    10   0.45   0.045   6.458 7.5e-10 ***
Residuals                      1830  12.89   0.007                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
aov5 <- aov(data = wind_data, percentage ~ ccaa +
              factor(lubridate::year(date)) + factor(lubridate::month(date)) )
summary(aov5)
```

```
                                  Df Sum Sq Mean Sq F value   Pr(>F)    
ccaa                              15  47.69   3.179 518.413  < 2e-16 ***
factor(lubridate::year(date))     10   0.45   0.045   7.419 1.23e-11 ***
factor(lubridate::month(date))    11   1.74   0.158  25.766  < 2e-16 ***
Residuals                       1819  11.15   0.006                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

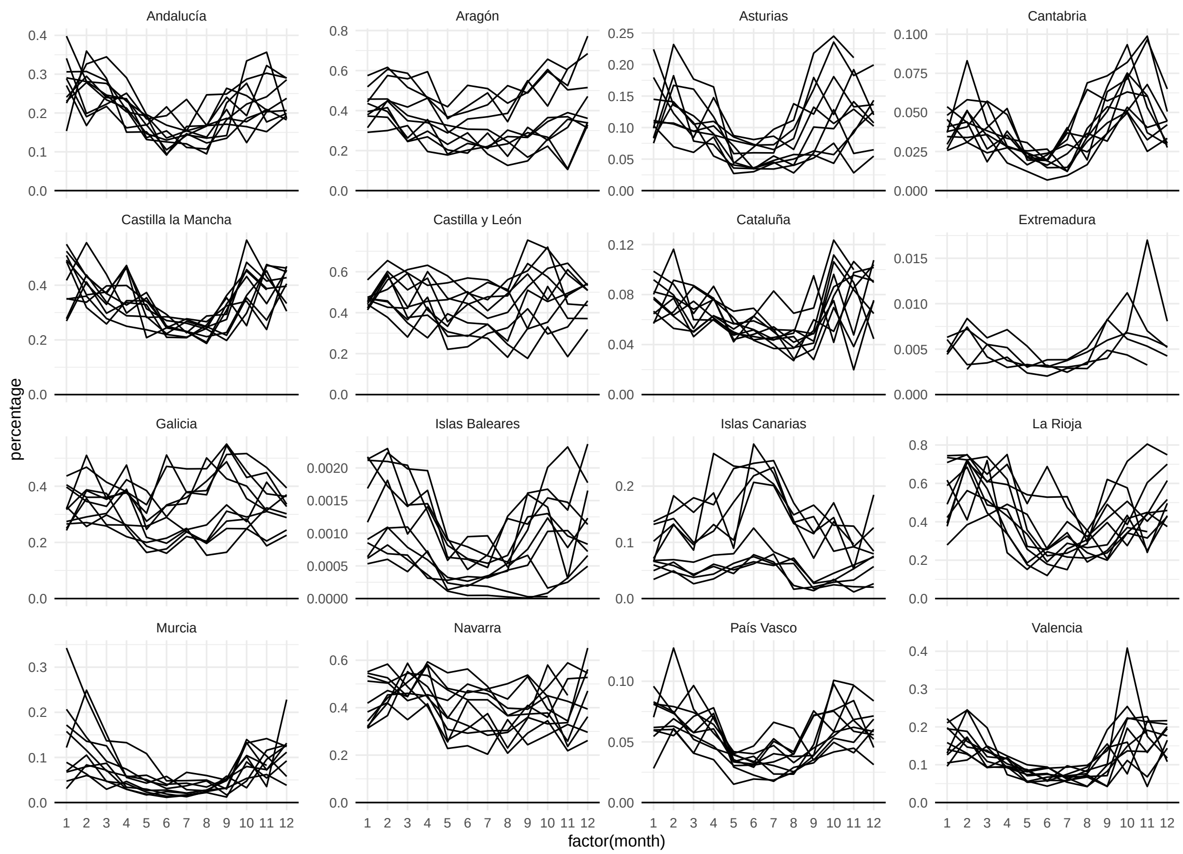
```r
anova(aov1, aov2, aov3, aov4, aov5)
```

```
Analysis of Variance Table

Model 1: percentage ~ ccaa + factor(date)
Model 2: percentage ~ ccaa + date
Model 3: percentage ~ ccaa + lubridate::year(date)
Model 4: percentage ~ ccaa + factor(lubridate::year(date))
Model 5: percentage ~ ccaa + factor(lubridate::year(date)) + factor(lubridate::month(date))
  Res.Df     RSS   Df Sum of Sq       F    Pr(>F)    
1   1721  9.9342                                     
2   1839 13.1635 -118   -3.2293  4.7411 < 2.2e-16 ***
3   1839 13.1429    0    0.0206                      
4   1830 12.8925    9    0.2504  4.8203 2.219e-06 ***
5   1819 11.1544   11    1.7380 27.3726 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The best model in terms of lowest *Residual Sum of Sq* is Model5, which includes two factor variables for the date, split up into `year` and `month`. That way we can get seasonal affects and year-over-year changes.

Let's plot how it looks for each monthly average:

```
wind_data |>
  mutate(month = lubridate::month(date),
         year_ccaa = paste(lubridate::year(date), ccaa)
         ) |>
  ggplot(aes(x = factor(month), y = percentage, group = year_ccaa)) +
  geom_line() +
  geom_hline(yintercept = 0) +
  theme_minimal() +
  facet_wrap(~ccaa, scales = "free_y")
```

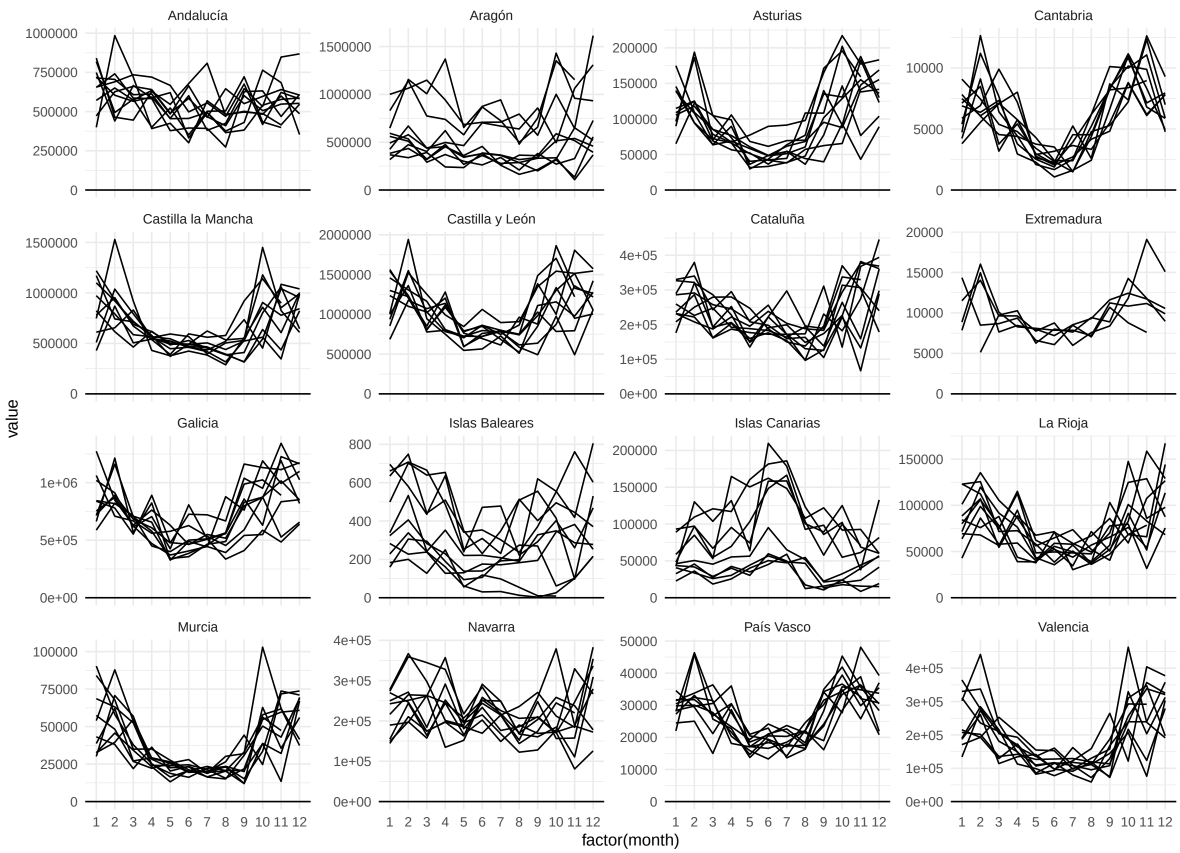


This plot above is very messy, but it it clear that many comunidades have a dip in `percentage` during the summer months. There are several exceptions, including Canarias, Galicia and Navarra.

But that is the percentage of the total electricity generated, as opposed to the absolute value.

Let’s see if there is a change during the summer:

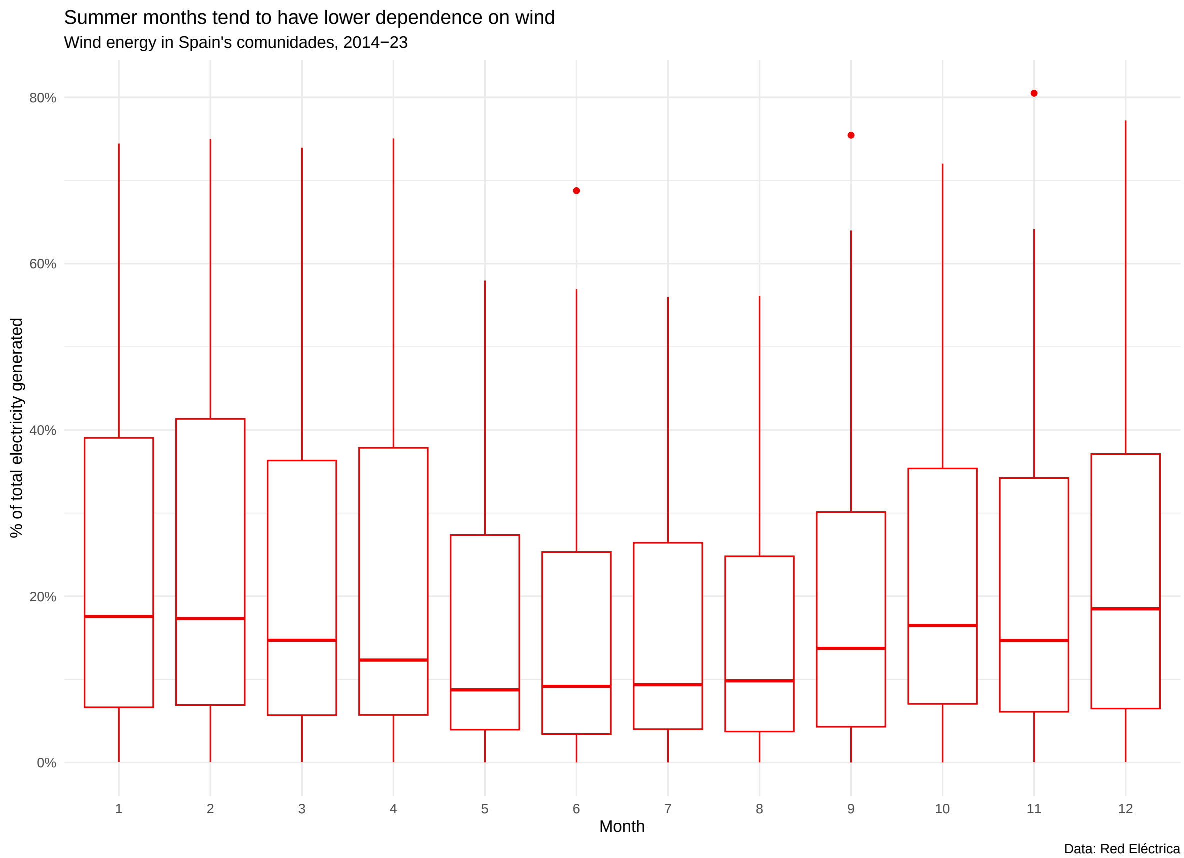
```
wind_data |>
  mutate(month = lubridate::month(date),
         year_ccaa = paste(lubridate::year(date), ccaa)
         ) |>
  ggplot(aes(x = factor(month), y = value, group = year_ccaa)) +
  geom_line() +
  geom_hline(yintercept = 0) +
  theme_minimal() +
  facet_wrap(~ccaa, scales = "free_y")
```



Interesting! We see a dip in the summer also for the actual generated value. That would suggest that the dip in percentage is not due to other sources increasing during the summer. For example, we can assume that solar power increases during the summer because of more sunshine hours. So that leaves us with a question, *Why do summer months generate less wind energy than do winter months?*

Distribution for each month

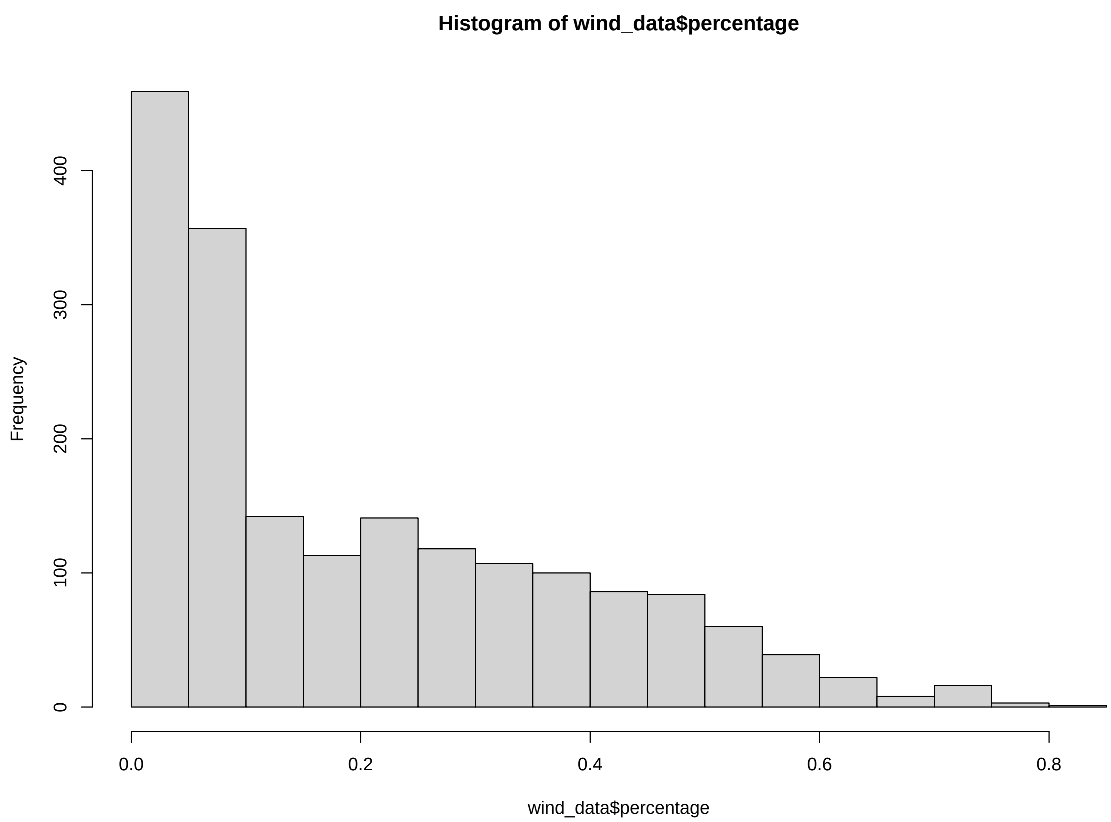
```
wind_data |>
  mutate(month = lubridate::month(date),
         year_ccaa = paste(lubridate::year(date), ccaa)
  ) |>
  ggplot(aes(x = factor(month), y = percentage)) +
  geom_boxplot(color = "red2", ) +
  theme_minimal() +
  labs( title = "Summer months tend to have lower dependence on wind",
    subtitle = "Wind energy in Spain's comunidades, 2014-23",
    caption = "Data: Red Eléctrica",
    x = "Month",
    y = "% of total electricity generated"
  ) +
  scale_y_continuous(labels = scales::label_percent())
```



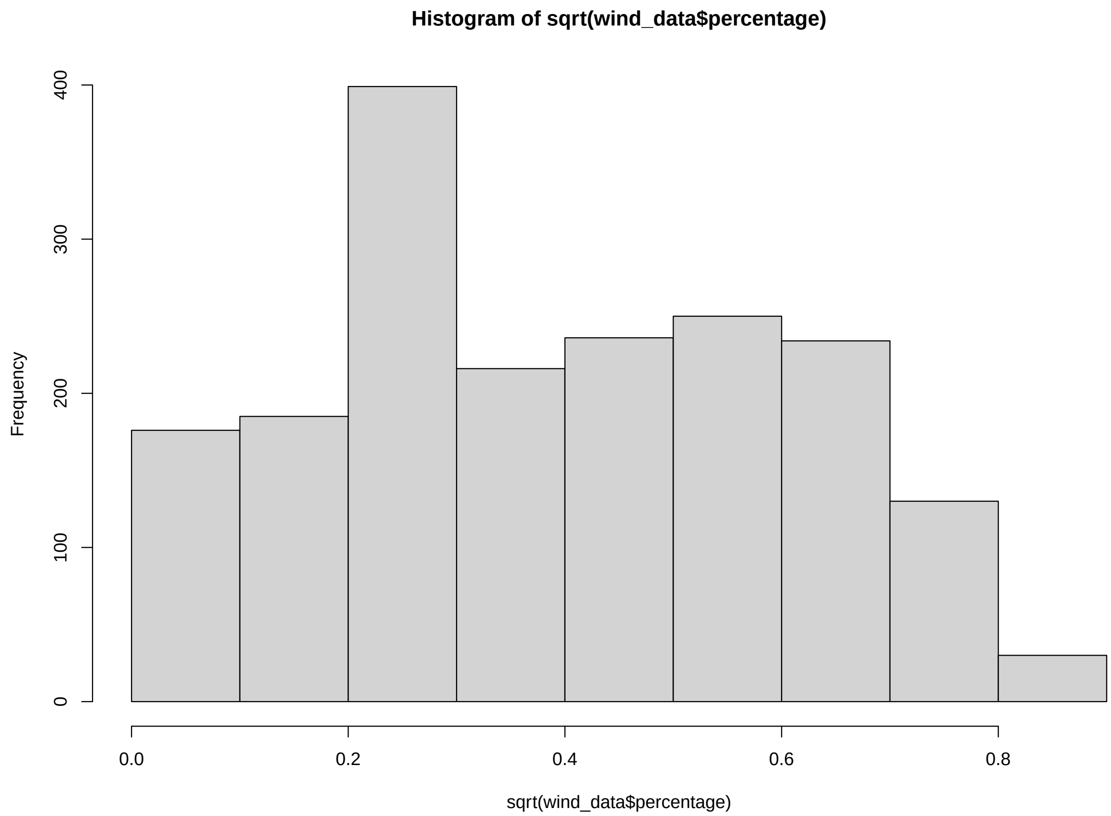
Now this plot is clearly shows that the summer months have a lower median percentage than the winter months. There appears to be a seasonal affect on wind energy generation, likely due to climate.

Distributions of `percentage`

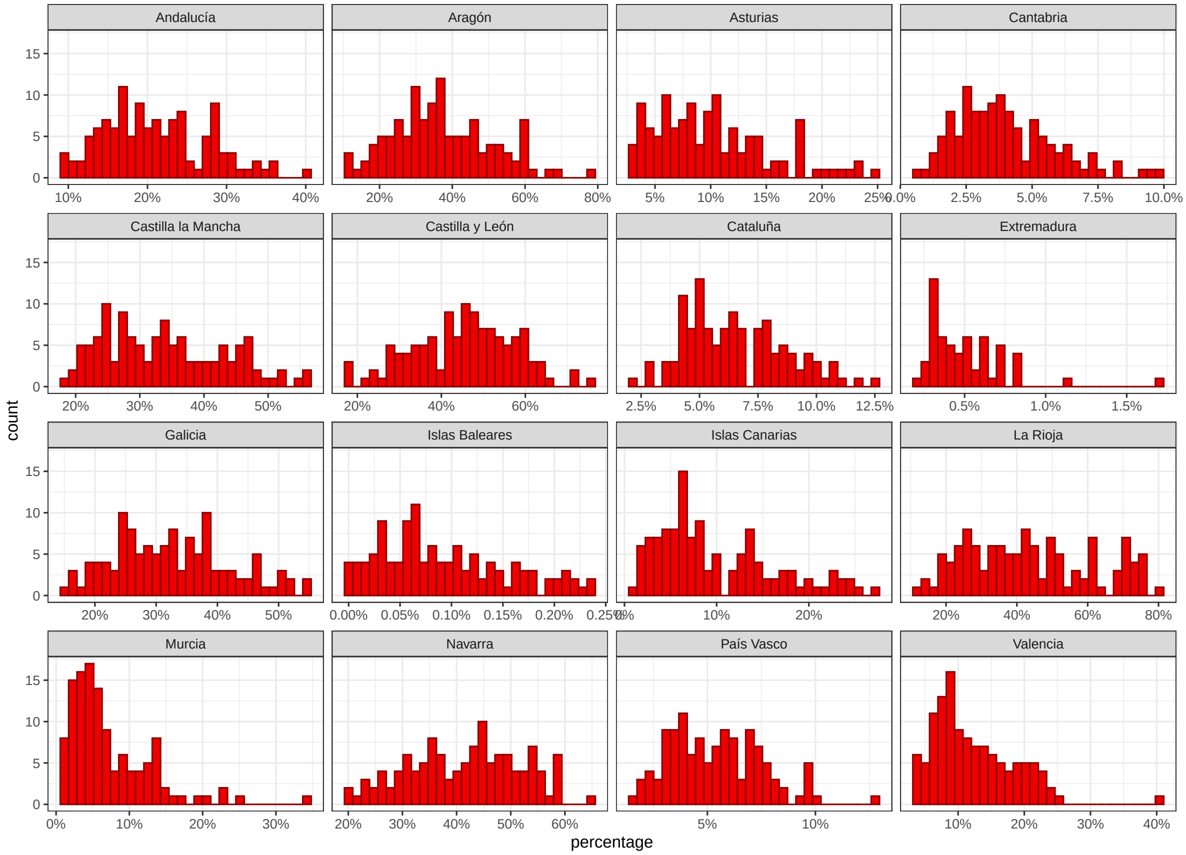
```r
# distribution of percentage of wind energy
hist(wind_data$percentage)
```




```r
hist(sqrt(wind_data$percentage))
```




```r
# but is it a bell curve for each CCAA?
wind_data |>
  ggplot(aes(percentage)) +
  geom_histogram(fill = "red2", color = "red4") +
  theme_bw() +
  facet_wrap(~ccaa, scales = "free_x") +
  scale_x_continuous(labels = scales::label_percent())
```



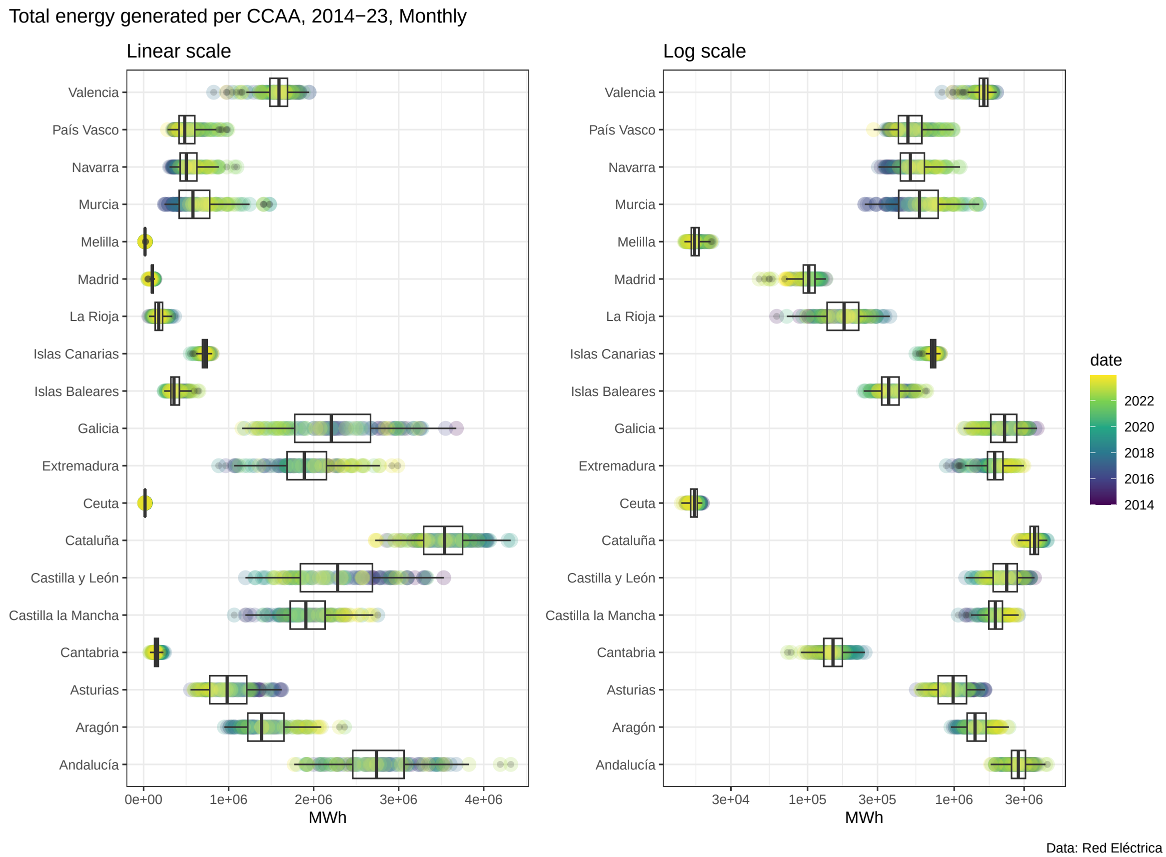
After splitting up the data into 16 facets, one for each comunidad, we can see that the distribution of `percentage` is less skewed than when compiled together. There is still a positive skew for **Murcia** and **Valencia**, but most comunidades do not have a clear distribution.

Total generated

Now let's do the same thing for total generation and see the differences:

```
library(patchwork)
plta <- red_data |>
  filter(name == "Generación total") |>
  ggplot(aes(x = value, y =ccaa, color = date)) +
  geom_point(size =4, alpha = .2) +
  geom_boxplot(alpha = .2) +
  scale_color_continuous(type = "viridis", trans = "date") +
  # scale_x_log10() +
```

```r
  theme_bw() +
  labs(
    x = "MWh",
    y = NULL,
    subtitle= NULL,
    title = "Linear scale",
  )

pltb <- red_data |>
  filter(name == "Generación total") |>
  ggplot(aes(x = value, y =ccaa, color = date)) +
  geom_point(size =4, alpha = .2) +
  geom_boxplot(alpha = .2) +
  scale_color_continuous(type = "viridis", trans = "date") +
  scale_x_log10() +
  theme_bw() +
  labs(
    x = "MWh",
    y = NULL,
    subtitle = NULL,
    title = "Log scale",
  )

plta + pltb +plot_layout(guides = "collect") +
  plot_annotation(title = "Total energy generated per CCAA, 2014-23, Monthly",
                  caption = "Data: Red Eléctrica")
```



ANOVA tests for Total

Null hypothesis -> Do all the CCAA have the same mean monthly wind generation?

```
aov1 <- aov(data = red_data[red_data$name == "Generación total",], value ~ ccaa + factor(d
summary(aov1)
```

```
                 Df    Sum Sq   Mean Sq  F value   Pr(>F)
ccaa             18 2.345e+15 1.303e+14 1552.898 < 2e-16 ***
factor(date)    119 1.379e+13 1.159e+11    1.382 0.00485 **
Residuals      2142 1.797e+14 8.390e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
aov2 <- aov(data = red_data[red_data$name == "Generación total",], value ~ ccaa + date )
summary(aov2)
```

```
              Df    Sum Sq   Mean Sq  F value Pr(>F)    
ccaa          18 2.345e+15 1.303e+14 1521.691 <2e-16 ***
date           1 4.407e+09 4.407e+09    0.051  0.821    
Residuals   2260 1.935e+14 8.562e+10                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
aov3 <- aov(data = red_data[red_data$name == "Generación total",], value ~ ccaa + lubridat
summary(aov3)
```

```
                        Df    Sum Sq   Mean Sq  F value Pr(>F)    
ccaa                    18 2.345e+15 1.303e+14 1521.685 <2e-16 ***
lubridate::year(date)    1 3.573e+09 3.573e+09    0.042  0.838    
Residuals             2260 1.935e+14 8.562e+10                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
aov4 <- aov(data = red_data[red_data$name == "Generación total",], value ~ ccaa + factor(l
summary(aov4)
```

```
                                Df    Sum Sq   Mean Sq  F value Pr(>F)    
ccaa                            18 2.345e+15 1.303e+14 1528.879 <2e-16 ***
factor(lubridate::year(date))   10 1.681e+12 1.681e+11    1.973 0.0325 *  
Residuals                     2251 1.918e+14 8.522e+10                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
aov5 <- aov(data = red_data[red_data$name == "Generación total",], value ~ ccaa +
              factor(lubridate::year(date)) + factor(lubridate::month(date)) )
summary(aov5)
```

```
                                 Df    Sum Sq   Mean Sq  F value   Pr(>F)    
ccaa                             18 2.345e+15 1.303e+14 1581.169  < 2e-16 ***
factor(lubridate::year(date))    10 1.681e+12 1.681e+11    2.040   0.0261 *  
```

```
factor(lubridate::month(date))   11 7.250e+12 6.591e+11    7.999 8.15e-14 ***
Residuals                      2240 1.846e+14 8.240e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

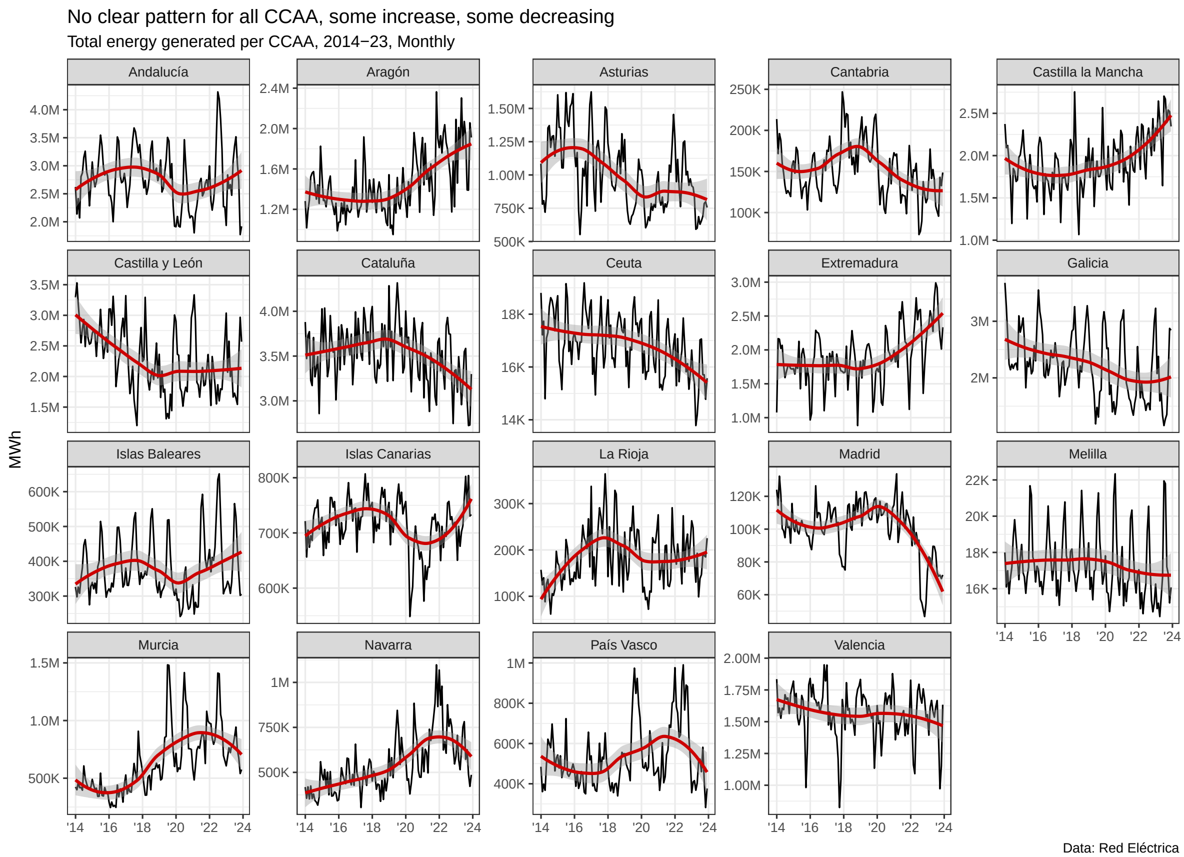
```
anova(aov1, aov2, aov3, aov4, aov5)
```

```
Analysis of Variance Table

Model 1: value ~ ccaa + factor(date)
Model 2: value ~ ccaa + date
Model 3: value ~ ccaa + lubridate::year(date)
Model 4: value ~ ccaa + factor(lubridate::year(date))
Model 5: value ~ ccaa + factor(lubridate::year(date)) + factor(lubridate::month(date))
  Res.Df        RSS   Df   Sum of Sq      F    Pr(>F)
1   2142 1.7971e+14
2   2260 1.9350e+14 -118 -1.3789e+13 1.3928  0.004106 **
3   2260 1.9350e+14    0 -8.3377e+08
4   2251 1.9182e+14    9  1.6775e+12 2.2216  0.018324 *
5   2240 1.8457e+14   11  7.2500e+12 7.8558 1.658e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Once again, different `ccaa` clearly have different mean total generation. In contrast with the ANOVA tests on percentage of wind energy, `date` does not have a strong affect on the outcome. One possible explanation is that some comunidades decreased generation over the last 10 years, while other increased.

```
red_data |>
  filter(name == "Generación total") |>
  ggplot(aes(x = as.Date(datetime), y = value)) +
  geom_line() +
  geom_smooth(color = "red3") +
  facet_wrap(~ccaa, scales = "free_y") +
  theme_bw() +
  scale_y_continuous(
    labels = scales::label_number(scale_cut = scales::cut_short_scale())
  ) +
  scale_x_date(date_labels = "'%y", minor_breaks = NULL) +
  labs(
    y = "MWh",
```

```r
    x = NULL,
    subtitle = "Total energy generated per CCAA, 2014-23, Monthly",
    title = "No clear pattern for all CCAA, some increase, some decreasing",
    caption = "Data: Red Eléctrica"
  )
```



Covariate data

```r
covariates.raw <- read_csv("./covariate_data.csv")

glimpse(covariates.raw)
```

```
Rows: 6,422
Columns: 21
```

```
$ DATE        <date> 1996-01-01, 1996-01-01, 1996-01-01, 1996-01-01, 1996-01-0~
$ price.index <dbl> 67.21, 67.21, 67.21, 67.21, 67.21, 67.21, 67.21, 67.21, 67~
$ month       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2~
$ year        <dbl> 1996, 1996, 1996, 1996, 1996, 1996, 1996, 1996, 1996, 1996~
$ CCAAnombre  <chr> "Asturias", "Islas Baleares", "Cantabria", "Ceuta", "La Ri~
$ mes         <chr> "enero", "enero", "enero", "enero", "enero", "enero", "ene~
$ prec        <dbl> 1182, 532, 1257, 855, 435, 453, 556, 254, 858, 413, 688, 3~
$ tmin        <dbl> 1.9, 5.7, 2.1, 10.5, -0.1, 0.0, 9.9, 3.3, 0.8, 11.4, 3.4, ~
$ tmax        <dbl> 105, 143, 103, 161, 79, 94, 168, 138, 86, 179, 134, 87, 10~
$ tmed        <dbl> 62, 100, 62, 133, 39, 47, 134, 86, 47, 146, 84, 41, 49, 33~
$ id          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ PIB         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ CCAA        <dbl> 3, 4, 6, 18, 17, 13, 19, 14, 15, 5, 1, 2, 7, 8, 9, 11, 12,~
$ response_p  <dbl> 0.3789474, 0.3814433, 0.3750000, 0.4285714, 0.3928571, 0.4~
$ total.x     <dbl> 95, 97, 56, 7, 28, 586, 6, 128, 62, 180, 746, 108, 190, 22~
$ response_sd <dbl> 0.04977277, 0.04931953, 0.06469365, 0.18704391, 0.09229619~
$ mas         <dbl> 0.5000000, 0.4845361, 0.6545455, 0.7142857, 0.5000000, 0.5~
$ menos       <dbl> 0.2127660, 0.2164948, 0.1272727, 0.1428571, 0.1785714, 0.2~
$ igual       <dbl> 0.2872340, 0.2989691, 0.2181818, 0.0000000, 0.3214286, 0.2~
$ ns          <dbl> 0.000000000, 0.000000000, 0.000000000, 0.142857143, 0.0000~
$ total.y     <dbl> 94, 97, 55, 7, 28, 585, 6, 128, 61, 180, 745, 108, 190, 22~
```

1. Electricity prices index

```
spain_electricity_index |>
  ggplot(aes(x = DATE, y = price.index)) +
  geom_point(alpha = 0.5) +
  geom_line() +
  theme_bw() +
  geom_smooth(se = T, color = "peru", span = 0.25, method = "loess" ) +
  labs(
    y = "Price index (100 = JAN'15)",
    x = NULL
  ) +
  theme(
  )
```

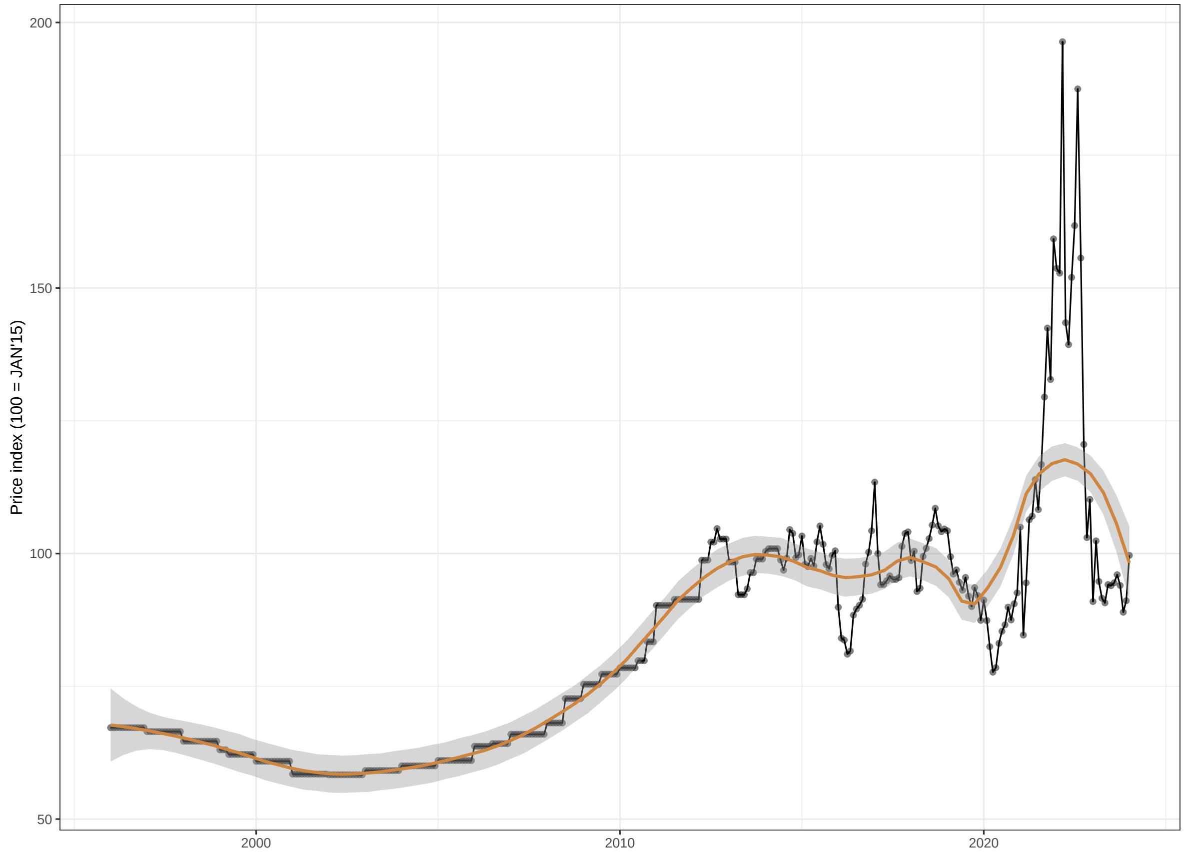
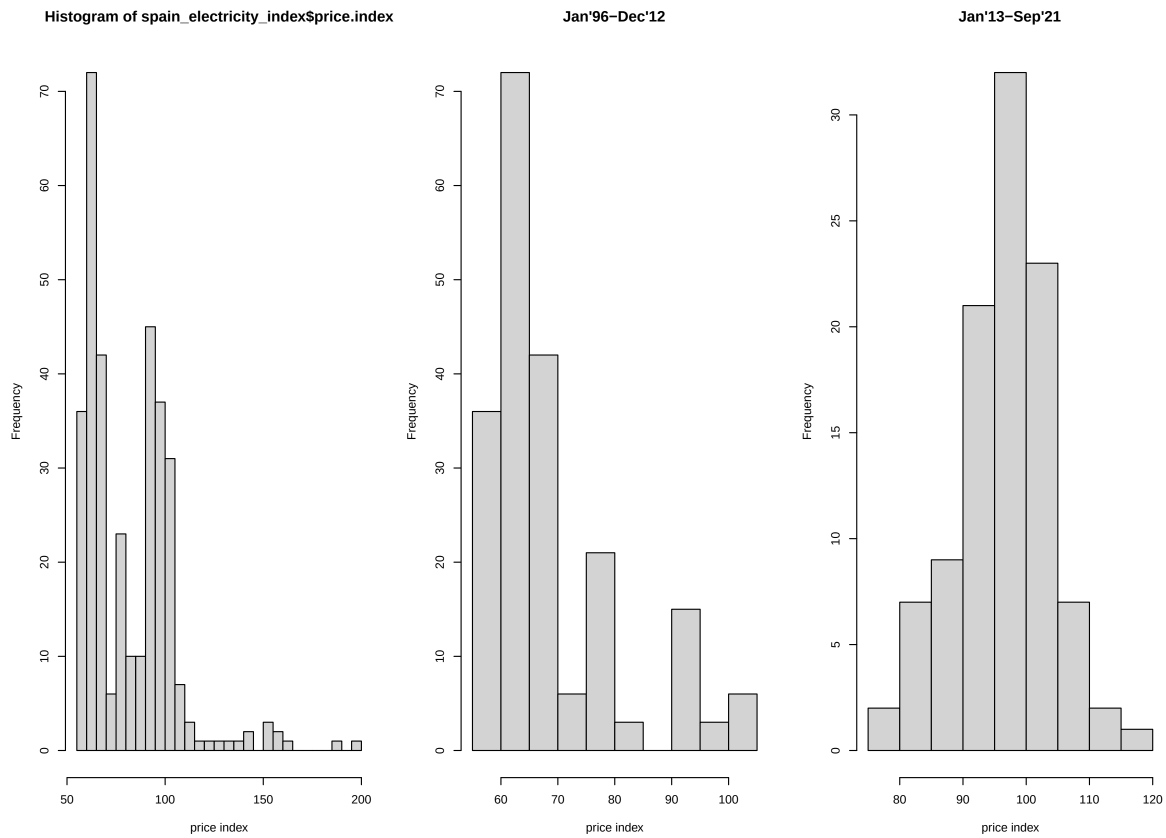


Figure 1: Timelapse chart

Looking at the histograms below, it appears that the distribution is bi-modal. There are a few outliers that occur between 2021 to 2023 when prices spiked enormously. Other than those outliers, the timelapse chart above shows two phases to electricity prices in Spain. First, from 1996 to 2010 the prices varied very little from month to month. Prices started gradually increasing in 2001 until 2013. After 2013, prices varied much more month to month, but overall prices stayed about the same price. After the spike of 2021-23, prices returned to the average from 2013-2020. The two phases I see are between 1996-2013 and 2013-today.

```
par(mfrow = c(1,3))
hist(spain_electricity_index$price.index, breaks = 20, xlab = "price index")

phase_1 <- spain_electricity_index$price.index[spain_electricity_index$DATE < as.Date("201
hist(phase_1, breaks = 10, main = "Jan'96-Dec'12", xlab = "price index")

phase_2 <- spain_electricity_index$price.index[spain_electricity_index$DATE >= as.Date("20
```

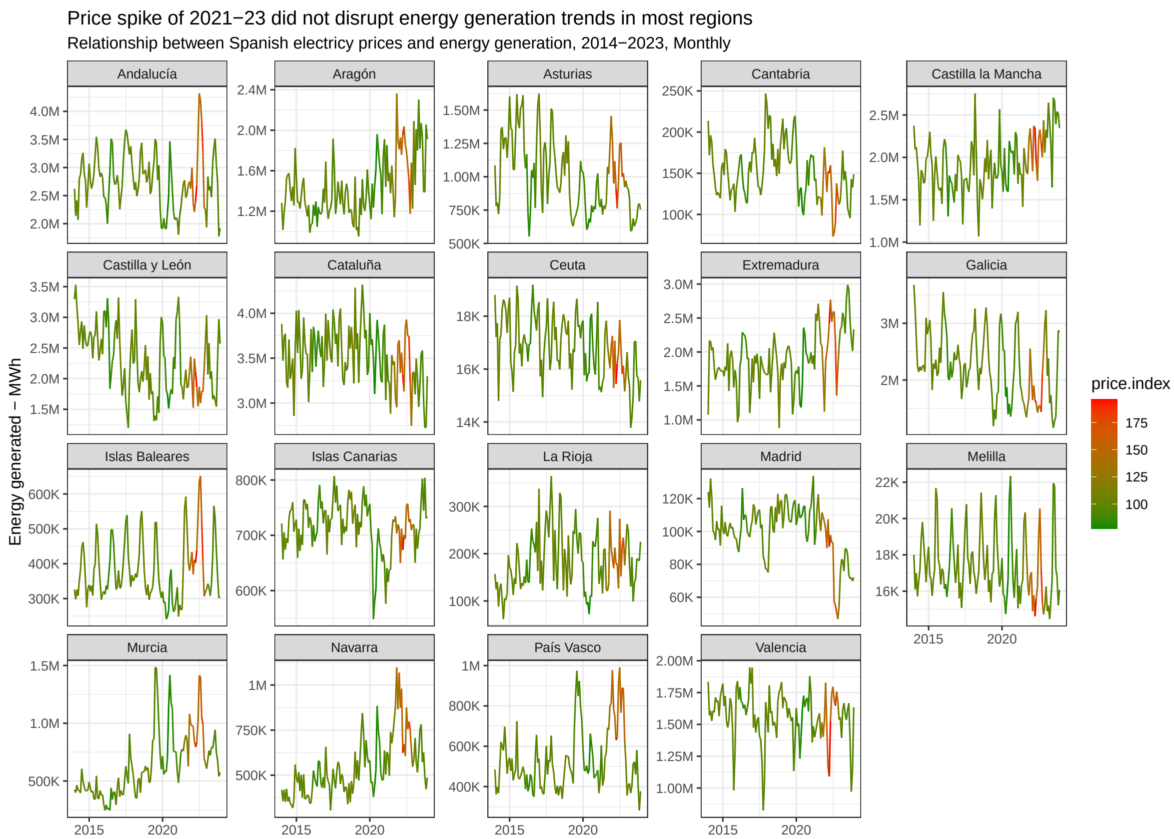
```r
hist(phase_2, breaks = 10, main = "Jan'13-Sep'21", xlab = "price index")
```



Generating energy per CCAA and Spanish electricity prices

```r
red_data |>
  mutate(month = lubridate::month(date),
         year = lubridate::year(date)) |>
  left_join(spain_electricity_index, by = c("year", "month")) |>
  filter(name == "Generación total") |>
  ggplot(aes(date, y = value, color = price.index)) +
 # geom_point() +
  geom_line() +
  scale_y_continuous(trans = "identity",
                     labels = scales::label_number( scale_cut = scales::cut_short_scale())
                     ) +
```

```
scale_x_continuous(
    trans = "date", n.breaks = 4
    ) +
  scale_color_gradient(
    low = "green4", high = "red"
    ) +
  facet_wrap(~ccaa, scales = "free_y") +
  theme_bw() +
  labs(
    x = "",
    y = "Energy generated - MWh",
    subtitle = "Relationship between Spanish electricy prices and energy generation, 2014-
   title = "Price spike of 2021-23 did not disrupt energy generation trends in most region
  )
```



From the chart above, you can see that in

2. Climate data

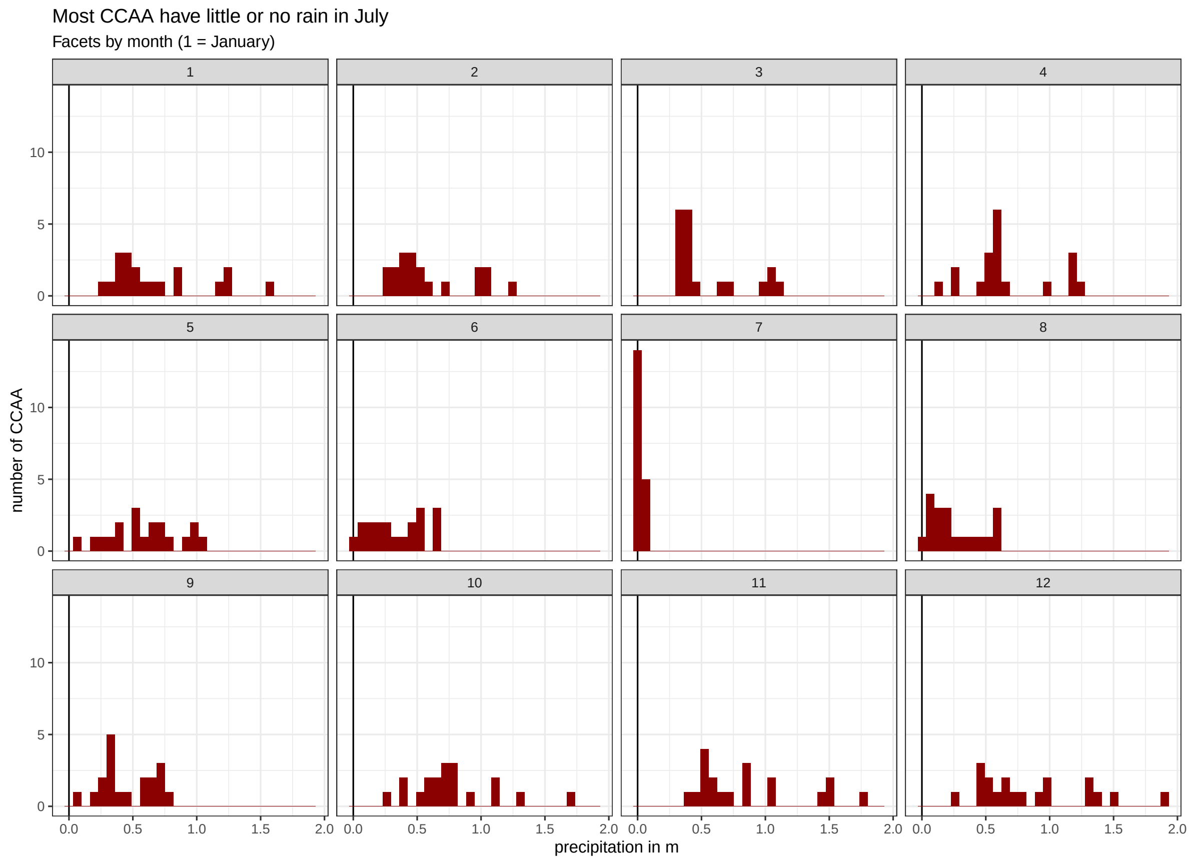
The climate data shows the average precipitation and temperature for each month for each comunidad. The values are taken from averages between 1981-2010. There is also a value for the average annual precipatation and temperature for each CCAA.

```r
clima <- climate_data |>
  pivot_wider(names_from = parametro, values_from = value)
glimpse(clima)
```

```
Rows: 247
Columns: 7
$ CCAAnombre <chr> "Asturias", "Asturias", "Asturias", "Asturias", "Asturias",~
$ mes        <chr> "enero", "febrero", "marzo", "abril", "mayo", "junio", "jul~
$ mes_num    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, NA, 1, 2, 3, 4, 5, 6~
$ prec       <dbl> 1182.0, 1073.0, 1022.0, 1222.0, 997.0, 651.0, 47.3, 562.0, ~
$ tmin       <dbl> 1.9, 2.0, 37.0, 48.0, 75.0, 106.0, 125.0, 126.0, 106.0, 80.~
$ tmax       <dbl> 105, 116, 139, 146, 174, 208, 230, 235, 217, 177, 134, 109,~
$ tmed       <dbl> 62, 68, 88, 97, 124, 157, 178, 181, 162, 128, 91, 68, 117, ~
```

Let’s see the distributions of average precipitation for each CCAA, split up by each month in one small multiple. We can clearly see that July has the smallest variance *and* smallest mean precipitation across regions. All of Spain has very small–if any– amounts of rain in July. The largest variance and mean appears to be in the winter months. I will test this out below.

```r
clima |>
  filter(mes!= "anual") |>
  ggplot(aes(prec)) +
  geom_vline(xintercept = 0, color = "black") +
  geom_histogram(fill = "red4") +
  facet_wrap(~mes_num, scales = "fixed") +
  scale_x_continuous(
    labels = scales::label_number(scale = 0.001) # scale to 'meters' instead of 'mm'
  ) +
  labs(
    y = "number of CCAA",
    x = "precipitation in m",
    subtitle = "Facets by month (1 = January)",
    title = "Most CCAA have little or no rain in July "
  ) +
  theme_bw()
```

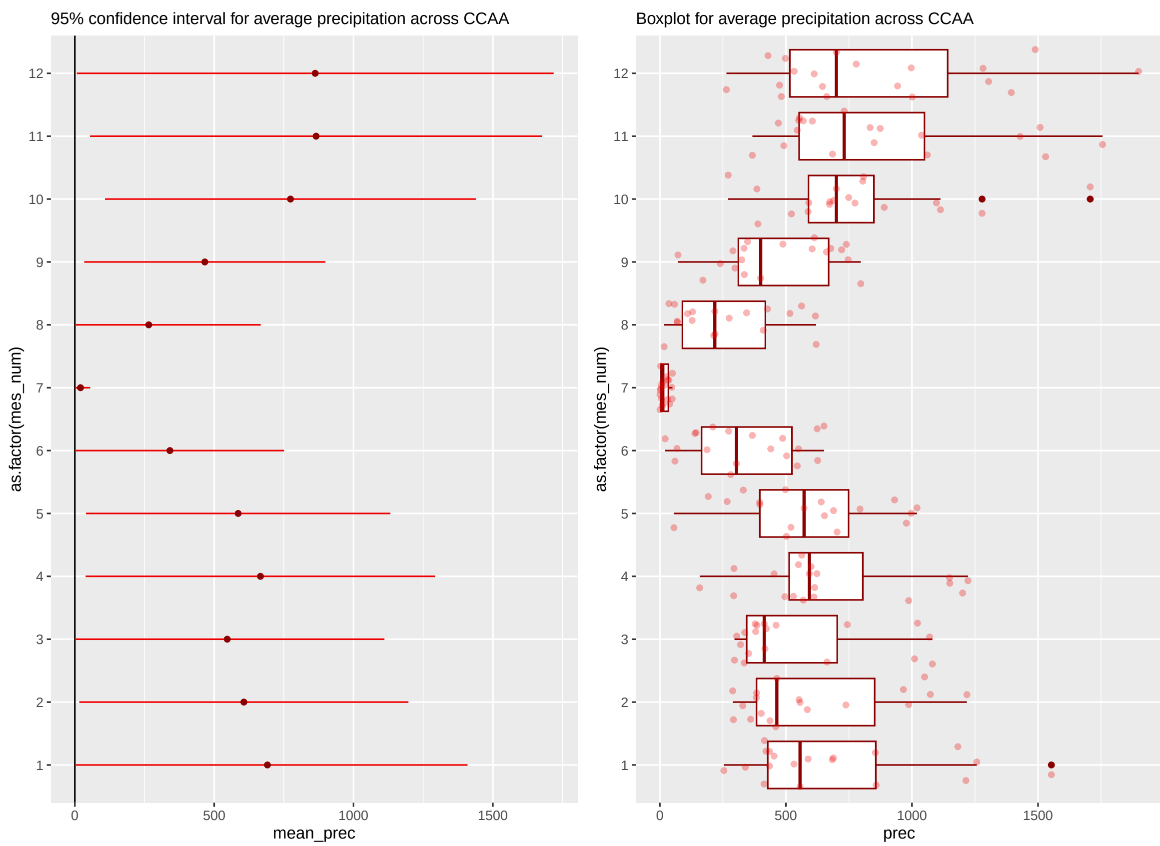



```r
mes_clima <- clima |> filter(mes != "anual")

aov <- aov(data = mes_clima, prec ~ mes + CCAAnombre)
summary(aov)
```

```
             Df   Sum Sq Mean Sq F value Pr(>F)
mes          11 13242197 1203836   41.30 <2e-16 ***
CCAAnombre   18 14032181  779566   26.74 <2e-16 ***
Residuals   198  5771562   29149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
z <- 1.96
pltA <- mes_clima |>
  group_by(mes_num) |>
  summarise(
```

```r
    sd_prec = sd(prec),
    mean_prec = mean(prec)
  ) |>
  ggplot(aes(x = mean_prec, y = as.factor(mes_num) ,
             xmin = mean_prec - z*sd_prec,
             xmax = mean_prec + z*sd_prec)) +
  geom_vline(xintercept = 0, color = "black") +
  geom_linerange(color = "red2")+
  geom_point(color = "red4")+
  scale_x_continuous(
   limits = c(0, NA),
   oob = scales::squish
  ) +
  labs(
    subtitle = "95% confidence interval for average precipitation across CCAA"
  )

pltB <- mes_clima |>
  ggplot(aes(x = prec, y = as.factor(mes_num)) ) +
#   geom_vline(xintercept = 0, color = "black") +
  geom_boxplot(color = "red4")+
  geom_jitter(color = "red2", alpha = 0.3)+
  scale_x_continuous(
   limits = c(0, NA),
   oob = scales::squish
  ) +
  labs(
    subtitle = "Boxplot for average precipitation across CCAA"
  )

library(patchwork)
pltA + pltB
```



The chart with 95% confidence intervals shows precipitation has a high amount of variance all year except in July. Although the *mean* count of precipitation changes a lot by season (summer months have fewer rain), there is a lot of variation across region. Meanwhile, the boxplot chart shows that the distribution is skewed right for the rainier winter months due to some outliers with high amounts of precipitation. Looking at the boxplots, you can see that the variation is probably not as wide as the 95% confidence interval based on *mean*.

For example, October (`mes` = 10) has a *mean* of 773 and *median* a bit lower at 700 mm. The middle 50%, shown in the "box" of the boxplot, is between 590 ad 859, which appears to be symmetrical around the median. The "whiskers" also appear symmetrical and equidistant from the median, reaching to the minimum of 271. However, there are two major outliers that skew the distribution. Those two outliers are **Asturias** and **Galicia**. Also note that the middle two quartiles are relatively small compared to the other months, except for July.

```
print(mes_clima |>
  filter(mes_num == 10 ) |>
  summary()
```

```
)
```

```
  CCAAnombre             mes               mes_num         prec       
 Length:19          Length:19          Min.   :10   Min.   : 271.0  
 Class :character   Class :character   1st Qu.:10   1st Qu.: 589.5  
 Mode  :character   Mode  :character   Median :10   Median : 700.0  
                                       Mean   :10   Mean   : 773.8  
                                       3rd Qu.:10   3rd Qu.: 849.0  
                                       Max.   :10   Max.   :1707.0  
      tmin             tmax            tmed      
 Min.   : 59.0   Min.   :174.0   Min.   :117.0  
 1st Qu.: 80.5   1st Qu.:181.5   1st Qu.:131.5  
 Median : 87.0   Median :195.0   Median :141.0  
 Mean   :104.0   Mean   :203.1   Mean   :153.6  
 3rd Qu.:117.0   3rd Qu.:226.0   3rd Qu.:172.5  
 Max.   :172.0   Max.   :239.0   Max.   :205.0  
```

```
print(mes_clima |>
  filter(mes_num == 10 &
           prec > 1250 ) # get the outliers greater than the whiskers
)
```

```
# A tibble: 2 x 7
  CCAAnombre mes     mes_num  prec  tmin  tmax  tmed
  <chr>      <chr>     <dbl> <dbl> <dbl> <dbl> <dbl>
1 Asturias   octubre      10  1278    80   177   128
2 Galicia    octubre      10  1707    81   179   130
```

3. Survey responses

The survey data is broken up into two questions.

`p_energy` is the proportion of responsdents, for each CCAA, that included an issue related to the climate, environment or energy in the question about what are the most critical issues this decade. These are the topics included from the CIS survey:

1. "La destrucción de la naturaleza y de la biodiversidad, la deforestación"
2. "La falta de recursos naturales, la escasez, de materias primas"
3. "El cambio climático. el calentamionto global"
4. "La energía (encarecimento, escasez, dependencia)"

`mas`, `menos`, `igual`, and `ns` are the percentage of respondents who answered with that response in the question about how bad the environment in Spain will worsen.

```
survey <- covariates.raw |>
  distinct(CCAAnombre, response_p, response_sd, total.y, mas, menos, igual, ns)

glimpse(survey)
```

```
Rows: 19
Columns: 8
$ CCAAnombre  <chr> "Asturias", "Islas Baleares", "Cantabria", "Ceuta", "La Ri~
$ response_p  <dbl> 0.3789474, 0.3814433, 0.3750000, 0.4285714, 0.3928571, 0.4~
$ response_sd <dbl> 0.04977277, 0.04931953, 0.06469365, 0.18704391, 0.09229619~
$ total.y     <dbl> 94, 97, 55, 7, 28, 585, 6, 128, 61, 180, 745, 108, 190, 22~
$ mas         <dbl> 0.5000000, 0.4845361, 0.6545455, 0.7142857, 0.5000000, 0.5~
$ menos       <dbl> 0.2127660, 0.2164948, 0.1272727, 0.1428571, 0.1785714, 0.2~
$ igual       <dbl> 0.2872340, 0.2989691, 0.2181818, 0.0000000, 0.3214286, 0.2~
$ ns          <dbl> 0.000000000, 0.000000000, 0.000000000, 0.142857143, 0.0000~
```

First question: most critical issues

```
z = 1.96
survey |>
  ggplot(aes(y = reorder(CCAAnombre, response_p), x = response_p,
             xmax = z * response_sd + response_p,
             xmin = response_p - z * response_sd)) +
  geom_linerange(color = "red4") +
  geom_point(color = "red2") +
  labs( title = "",
    subtitle = "95% C.I. for mean proportion of respondents that said energy or climate ch
    x = NULL,
    y = NULL
  ) +
  scale_x_continuous(
    labels = scales::label_percent(),
    limits = c(0,1),
    oob = scales::squish
  )  +
  theme_bw() +
  theme(
```

```
    axis.text.y = element_text(size = 12, face = "bold")
  ) +
  annotate(
    "text", x = 0.7, y = 1.2, label =
      expression(paste(bold("Ceuta"), " and ", bold("Melila"), " have very low sample size
     )
```

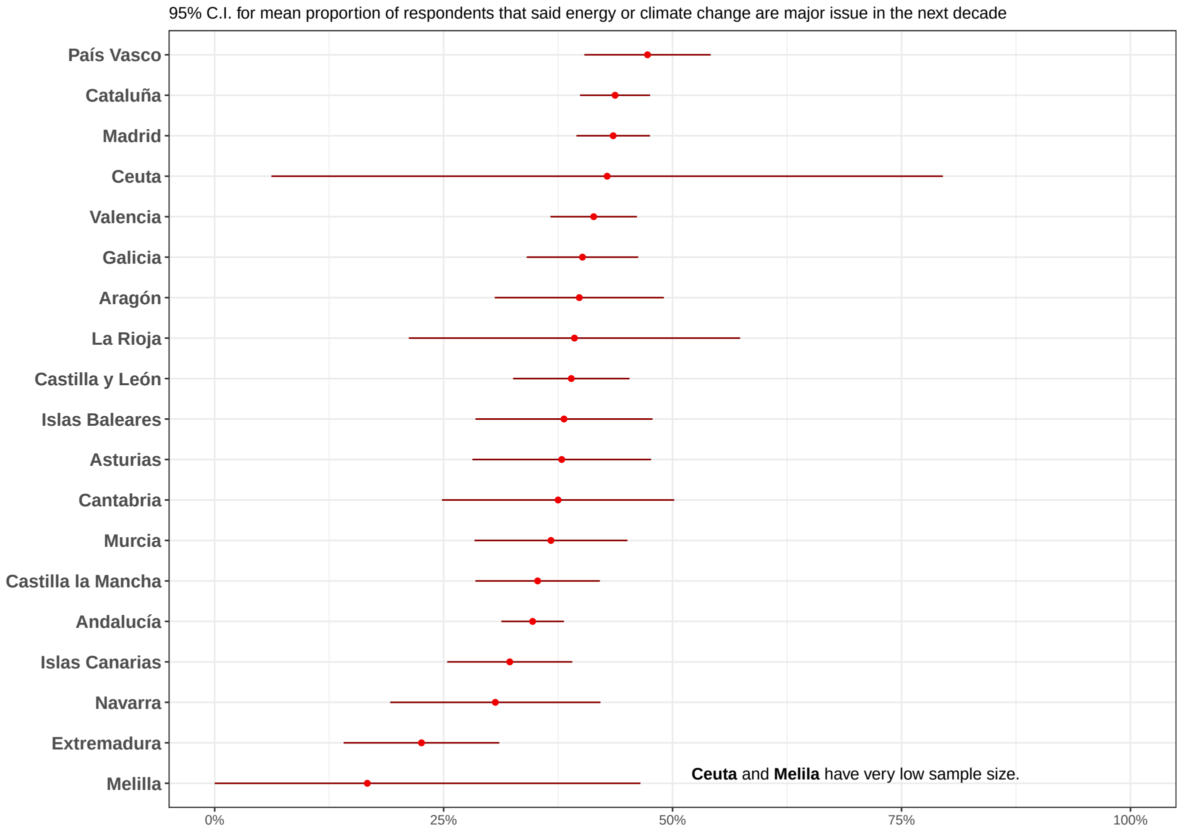


Figure 2: Figure A

We can say, at a 95% confidence level, that **Extremadura** likely has the lowest proportion of respondents that think energy and natural resources are the most important issues today. Although **Melilla** has the smallest proportion, there were only 6 respondents from that CCAA in this survey. **Ceuta** also had a very small sample size of only 7 respondents. The top three CCAA with the highest proportion were **País Vaso**, **Cataluña** and **Madrid**.

The chart suggests that Spanish CCAA are different in a statistically significant way. The confidence interval shows the estimated true proportion for each region with 95% likelihood and they do not all overlap. Interestingly, **Andalucía** is significantly different from the top three.

Second question: environmental destruction

The sample sizes are basically the same for this question.

```
z <-  1.96
survey |>
  select(CCAAnombre, mas, menos, igual, total.y) |>
  mutate(moe_mas = z * sqrt(mas * (1 - mas) / total.y)) |> # calculate margin of error wit
    ggplot(aes(y = reorder(CCAAnombre, mas), x = mas,
             xmax = z * moe_mas + mas,
             xmin = mas - z * moe_mas)) +
  geom_linerange(color = "red4") +
  geom_point(color = "red2") +
  labs( title = "All confidence intervals overlap",
    subtitle = "95% C.I. for mean proportion of respondents that said the environmental de
    x = NULL,
    y = NULL
  ) +
  scale_x_continuous(
    labels = scales::label_percent(),
    limits = c(0,1),
    oob = scales::squish
  )  +
  theme_bw() +
  theme(
    axis.text.y = element_text(size = 12, face = "bold")
  ) +
  annotate(
    "text", x = 0.2, y = 18.5, label =
      expression(paste(bold("Ceuta"), " and ", bold("Melila"), " have very low sample size
     )
```

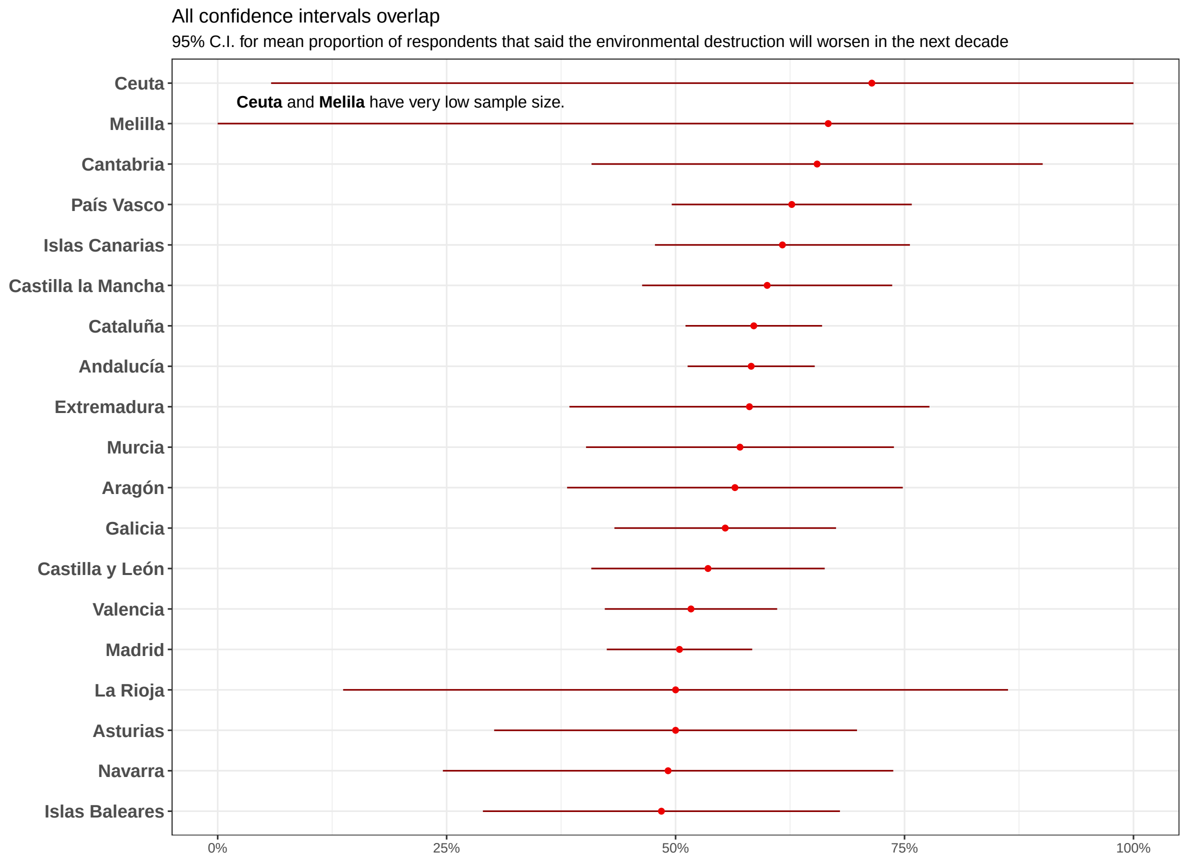


Figure 3: Figure B

Unlike the first question about the most important issues, the responses to this question appear to have similar results across Spanish comunidades. You can see that **Madrid** had one of the lowest proportions, with 50% saying that environmental degradation will get worse. Its confidence interval overlaps with every other CCAA's interval.

> Define margin of error: If the survey was replicated many times with similar respondents, 95% of those surveys' confidence intervals would include the true population mean. In other words, out of 20 surveys done in the same way, 19 of them would have a confidence interval that includes the true population estimate.

Despite the large confidence intervals due to small sample sizes, there could still be a small relationship between region and opinions about environmental issues. I tested this out with the Chi-squared test and Cramer's V measure. You can see in the results below that the Chi-squared test indicates evidence of an association, but not in the Cramer's V. Therefore, I conclude that there is not a strong relationship between region and public opinion on this issue.

```r
cont.table <- with(CIS.data2,  table(CCAA, medio_ambiente) )
cont.table
```

```
    medio_ambiente
CCAA   1   2   3   8
  1  434 125 180   6
  2   61  28  19   0
  3   47  20  27   0
  4   47  21  29   0
  5  111  31  37   1
  6   36   7  12   0
  7  114  23  53   0
  8  121  39  65   1
  9  377 118 145   4
  10 215  92 108   1
  11  54  17  22   0
  12 138  53  55   3
  13 295 147 139   4
  14  73  21  34   0
  15  30  16  15   0
  16 126  32  42   1
  17  14   5   9   0
  18   5   1   0   1
  19   4   0   2   0
```

```r
chisq.test(cont.table, simulate.p.value = T)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  cont.table
X-squared = 90.447, df = NA, p-value = 0.009995
```

```r
cramersV <- lsr::cramersV(cont.table, simulate.p.value = T)
```

Cramer's V: 0.0856165

The Chi-sq test and cramer's V for the first question indicated a stronger association but still not that strong.

```r
cont.table <- with(CIS.data2,  table(CCAA, response_flag) )
cont.table
```

```
    response_flag
CCAA   0   1
  1  487 259
  2   65  43
  3   59  36
  4   60  37
  5  122  58
  6   35  21
  7  123  67
  8  138  88
  9  363 282
  10 245 173
  11  72  21
  12 149 100
  13 331 255
  14  81  47
  15  43  19
  16 106  95
  17  17  11
  18   4   3
  19   5   1
```

```r
chisq.test(cont.table, simulate.p.value = T)
```

```
	Pearson's Chi-squared test with simulated p-value (based on 2000
	replicates)

data:  cont.table
X-squared = 42.257, df = NA, p-value = 0.0009995
```

```r
lsr::cramersV(cont.table, simulate.p.value = T)
```

```
[1] 0.1012627
```

Logistic regression

The simple logistic regression summary below shows evidence that different regions are statistically different in their opinions about the major issues of this decade in Spain. See that Cataluña (, Madrid and País Vasco are significantly different from Anadalucía. These results are in line with the data visualization of the confidence intervals (**Figure B**).

```
logit_q1 <- glm(response_flag ~ CCAA -1, data = CIS.data2, family = binomial)
summary(logit_q1)
```

```
Call:
glm(formula = response_flag ~ CCAA - 1, family = binomial, data = CIS.data2)

Coefficients:
       Estimate Std. Error z value Pr(>|z|)
CCAA1  -0.63144    0.07691  -8.211  < 2e-16 ***
CCAA2  -0.41319    0.19657  -2.102 0.035556 *
CCAA3  -0.49402    0.21149  -2.336 0.019495 *
CCAA4  -0.48343    0.20903  -2.313 0.020739 *
CCAA5  -0.74358    0.15949  -4.662 3.13e-06 ***
CCAA6  -0.51083    0.27603  -1.851 0.064221 .
CCAA7  -0.60749    0.15184  -4.001 6.31e-05 ***
CCAA8  -0.44992    0.13642  -3.298 0.000974 ***
CCAA9  -0.25250    0.07938  -3.181 0.001468 **
CCAA10 -0.34797    0.09931  -3.504 0.000458 ***
CCAA11 -1.23214    0.24801  -4.968 6.76e-07 ***
CCAA12 -0.39878    0.12927  -3.085 0.002037 **
CCAA13 -0.26085    0.08332  -3.131 0.001744 **
CCAA14 -0.54430    0.18336  -2.968 0.002993 **
CCAA15 -0.81676    0.27548  -2.965 0.003028 **
CCAA16 -0.10956    0.14128  -0.775 0.438049
CCAA17 -0.43532    0.38695  -1.125 0.260593
CCAA18 -0.28768    0.76376  -0.377 0.706423
CCAA19 -1.60944    1.09545  -1.469 0.141776
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5712.9  on 4121  degrees of freedom
Residual deviance: 5476.3  on 4102  degrees of freedom
AIC: 5514.3
```

```
Number of Fisher Scoring iterations: 4
```

```r
knitr::kable(cis_regions, type = "text")
```

CCAA	CCAAnombre
1	Andalucía
2	Aragón
3	Asturias
4	Islas Baleares
5	Islas Canarias
6	Cantabria
7	Castilla la Mancha
8	Castilla y León
9	Cataluña
10	Valencia
11	Extremadura
12	Galicia
13	Madrid
14	Murcia
15	Navarra
16	País Vasco
17	La Rioja
18	Ceuta
19	Melilla

I expect to see that the coefficient estimates will not be statistically significant for the second question. In fact, the results of the multinomial model turned out be very bad at prediction. This model still shows that different CCAA have different mix of opinions.

```r
## partition data first
library(caret)
index <- createDataPartition(CIS.data2$medio_ambiente, p = .7, list = FALSE)
train.data <- CIS.data2[index,]
test.data <- CIS.data2[-index,]

## run multinomial
multi_q2 <- nnet::multinom(medio_ambiente ~ CCAA, data = train.data)
```

```
# weights:  80 (57 variable)
```

```
initial  value 3995.300349
iter  10 value 2909.856165
iter  20 value 2889.891817
iter  30 value 2883.937094
iter  40 value 2883.365344
iter  50 value 2882.920625
iter  60 value 2882.866027
final  value 2882.863690
converged
```

```
summary(multi_q2)
```

```
Call:
nnet::multinom(formula = medio_ambiente ~ CCAA, data = train.data)

Coefficients:
  (Intercept)       CCAA2        CCAA3       CCAA4        CCAA5       CCAA6
2  -1.2992812   0.6526559   0.39641166   0.4883583   0.04651848  -0.3101759
3  -0.8033859  -0.1010681   0.02568183   0.2643991  -0.20821230  -0.1129107
8  -3.9019747 -11.9384790 -12.85476592 -12.3377248 -23.17945198 -19.3387312
        CCAA7       CCAA8       CCAA9      CCAA10      CCAA11        CCAA12
2  -0.35894817  0.3007523  0.07837971   0.4697232   0.3502057  3.150798e-01
3  -0.01651068  0.3846784 -0.25988544   0.1814681  -0.3280136 -7.816088e-02
8 -28.96562935 -0.4287663 -1.03251337 -15.2390869 -16.0826032 -9.872828e-06
       CCAA13      CCAA14       CCAA15     CCAA16      CCAA17     CCAA18
2  0.68617814  -0.1047163   0.77318765 -0.1205374   0.2876676 -21.033860
3  0.06768139  -0.1486215   0.01492774 -0.5294180   0.1972604 -21.584996
8 -0.25692475 -11.9525869 -16.67308250 -0.6088926 -20.3112017   2.292615
       CCAA19
2 -21.3377193
3   0.3979127
8 -17.4942869

Std. Errors:
  (Intercept)        CCAA2        CCAA3        CCAA4        CCAA5        CCAA6
2   0.1253501 2.915076e-01 3.307664e-01 3.255616e-01 2.723128e-01 5.056833e-01
3   0.1043351 3.058085e-01 3.110246e-01 2.937389e-01 2.441046e-01 3.884403e-01
8   0.4123510 3.395210e-06 1.196625e-06 1.952607e-06 8.166090e-11 1.236762e-09
         CCAA7     CCAA8     CCAA9       CCAA10       CCAA11    CCAA12
2 3.001958e-01 0.2541355 0.1774962 1.944445e-01 3.623587e-01 0.2298710
3 2.231985e-01 0.2098659 0.1577961 1.734855e-01 3.783435e-01 0.2130189
```

```
8 2.734530e-13 1.0877499 0.8207544 4.436855e-07 3.976070e-08 0.8247063
     CCAA13       CCAA14       CCAA15    CCAA16       CCAA17       CCAA18
2 0.1747466 3.235511e-01 3.716050e-01 0.2686192 5.971810e-01 2.455839e-10
3 0.1642063 2.718164e-01 3.953990e-01 0.2520742 5.181312e-01 2.304922e-10
8 0.7131485 4.542052e-06 1.564986e-08 1.0867522 2.077662e-10 1.170458e+00
        CCAA19
2 1.390243e-10
3 9.188147e-01
8 1.052969e-09

Residual Deviance: 5765.727
AIC: 5879.727
```

```
# confusionMatrix(predict(multi_q2, type = "probs", newdata = test.data), test.data$medio_
```

Let's try the same thing but with a logistic model on just the first answer: *environmental damage will get worse*. You can see from the results below that only five CCAA have coefficient estimates that are significant at a 99.0% confidence level. There is a better than 50-50 chance that a respondent from **Andalucía**, **Islas Canarias**, **Castilla la Mancha**, **Cataluña**, and **País Vasco** believe that the environment will get *worse* in the next 10 years. For all the other regions, we cannot say with certainty that the probability that a respondent will say *worse* any different from 50% chance. However, the second model shown below sets `Andalucía` as the reference and you can see that two CCAA are statistically different at 99% confidence level: **Madrid** and **Valencia**. Therefore, I will conclude that there are three groups :

1) CCAA with better than 50-50 chance that a respondent thinks environmental damage will get worse

   - Andalucía
   - Islas Canarias
   - Castilla la Mancha
   - Cataluña
   - País Vasco

2) CCAA with about 50% chance

   - Madrid
   - Valencia

3) CCAA without sufficient data

   - The rest

```r
mas <- ifelse(CIS.data2$medio_ambiente == 1, 1, 0)

logit_q2 <- glm(mas ~ CIS.data2$CCAA -1, family = binomial)
logit_q2_intercept <- glm(mas ~ CIS.data2$CCAA, family = binomial)
stargazer::stargazer(logit_q2, logit_q2_intercept, type = "text")
```

```
==============================================
                      Dependent variable:
                  ----------------------------
                              mas
                      (1)            (2)
----------------------------------------------
CCAA1               0.333***
                    (0.074)

CCAA2                0.261         -0.073
                    (0.194)        (0.208)

CCAA3                0.000         -0.333
                    (0.206)        (0.219)

CCAA4               -0.062         -0.395*
                    (0.203)        (0.216)

CCAA5               0.475***        0.142
                    (0.153)        (0.170)

CCAA6               0.639**         0.306
                    (0.284)        (0.293)

CCAA7               0.405***        0.072
                    (0.148)        (0.166)

CCAA8                0.142         -0.191
                    (0.133)        (0.153)

CCAA9               0.345***        0.012
                    (0.080)        (0.109)

CCAA10               0.067         -0.266**
                    (0.098)        (0.123)
```

CCAA11	0.325	-0.008
	(0.210)	(0.223)
CCAA12	0.218*	-0.116
	(0.127)	(0.148)
CCAA13	0.017	-0.316***
	(0.083)	(0.111)
CCAA14	0.283	-0.050
	(0.179)	(0.193)
CCAA15	-0.033	-0.366
	(0.256)	(0.267)
CCAA16	0.519***	0.186
	(0.146)	(0.164)
CCAA17	0.000	-0.333
	(0.378)	(0.385)
CCAA18	0.916	0.583
	(0.837)	(0.840)
CCAA19	0.693	0.360
	(0.866)	(0.869)
Constant		0.333***
		(0.074)
Observations	4,113	4,113
Log Likelihood	-2,806.546	-2,806.546
Akaike Inf. Crit.	5,651.092	5,651.092
Note:	*p<0.1; **p<0.05; ***p<0.01	

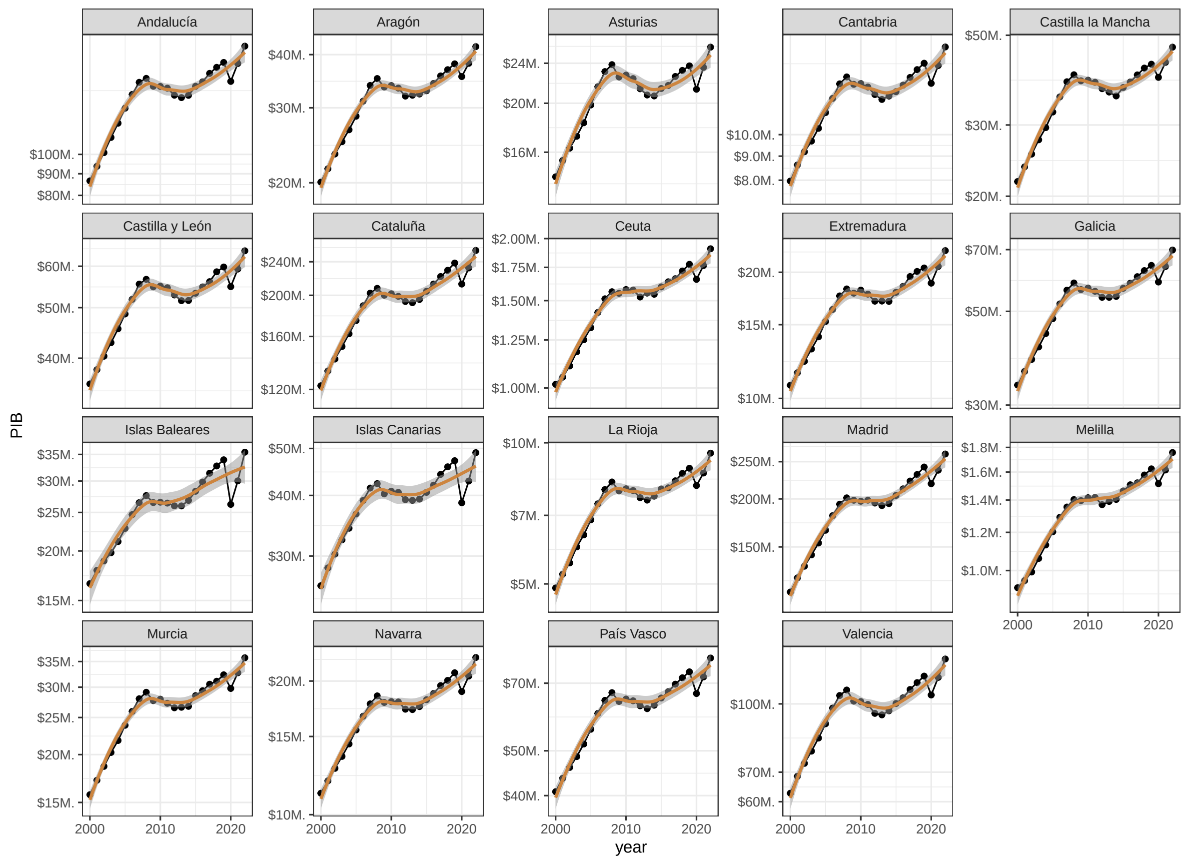
4. GDP by CCAA

```
glimpse(CCAA_PIB_yearly)
```

```
Rows: 437
Columns: 4
$ id        <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01", "01",~
$ PIB       <dbl> 180224284, 164003435, 148779089, 164929489, 160586830, 1553~
$ year      <dbl> 2022, 2021, 2020, 2019, 2018, 2017, 2016, 2015, 2014, 2013,~
$ CCAAnombre <chr> "Andalucía", "Andalucía", "Andalucía", "Andalucía", "Andalu~
```

See from the small multiples chart below that all CCAA follow the same trend for GDP at differing levels.

```
CCAA_PIB_yearly |>
  ggplot(aes(x = year, y = PIB)) +
  geom_line() +
  geom_point() +
  geom_smooth(color = "peru", alpha = .5,) +
  facet_wrap(~CCAAnombre, scales = "free_y") +
  scale_y_continuous(
    trans = "log10",
    labels = scales::label_dollar(scale_cut = scales::cut_short_scale(), suffix = "€")
  ) +
  scale_x_continuous(breaks = c(2000, 2010, 2020)) +
  theme_bw()
```

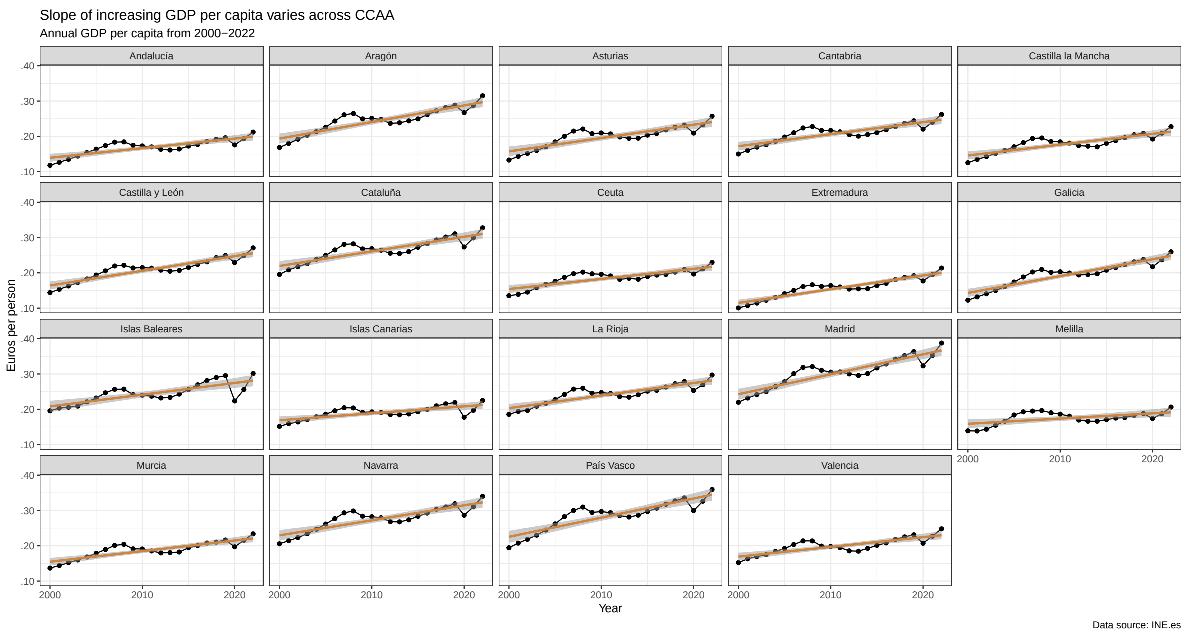



```
# suffix = " Mil€", scale = (1/1000000), largest_with_cents = 0
```

Now with GDP per capita, you can see that each comunidad is at a different level of economic production and rate of change from 2000 to 2022.

```
combined_data_pop_pib |>
  ggplot(aes(x = year, y = PIB/pop)) +
  geom_line() +
  geom_point() +
  geom_smooth(color = "peru", alpha = .5, method = "lm") +
  facet_wrap(~CCAAnombre, scales = "fixed", ncol = 5) +
  scale_y_continuous(
    labels = scales::label_dollar(prefix = "€")
  ) +
  scale_x_continuous(breaks = c(2000, 2010, 2020)) +
  theme_bw() +
```

```r
  labs( title = "Slope of increasing GDP per capita varies across CCAA",
    subtitle = "Annual GDP per capita from 2000-2022",
    x = "Year",
    y = "Euros per person",
    caption = "Data source: INE.es"
  )
```




```r
linear_reg <- lm(PIB/pop ~ CCAAnombre * I(year-2000) , data = combined_data_pop_pib)
summary(linear_reg)
```

```
Call:
lm(formula = PIB/pop ~ CCAAnombre * I(year - 2000), data = combined_data_pop_pib)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0961 -1.1014 -0.1437  1.0027  4.0837 

Coefficients:
                                             Estimate Std. Error t value
(Intercept)                                 13.935827   0.621013  22.440
CCAAnombreAragón                             5.434159   0.878245   6.188
CCAAnombreAsturias                           1.810230   0.878245   2.061
```

```
CCAAnombreCantabria                        3.276963   0.878245   3.731
CCAAnombreCastilla la Mancha               0.654986   0.878245   0.746
CCAAnombreCastilla y León                  2.472338   0.878245   2.815
CCAAnombreCataluña                         8.002060   0.878245   9.111
CCAAnombreCeuta                            1.472413   0.878245   1.677
CCAAnombreExtremadura                     -2.464400   0.878245  -2.806
CCAAnombreGalicia                          0.370347   0.878245   0.422
CCAAnombreIslas Baleares                   6.876370   0.878245   7.830
CCAAnombreIslas Canarias                   2.955236   0.878245   3.365
CCAAnombreLa Rioja                         6.395076   0.878245   7.282
CCAAnombreMadrid                          10.331408   0.878245  11.764
CCAAnombreMelilla                          1.957660   0.878245   2.229
CCAAnombreMurcia                           1.535543   0.878245   1.748
CCAAnombreNavarra                          9.014110   0.878245  10.264
CCAAnombrePaís Vasco                       8.598578   0.878245   9.791
CCAAnombreValencia                         2.971792   0.878245   3.384
I(year - 2000)                             0.272050   0.048346   5.627
CCAAnombreAragón:I(year - 2000)            0.197368   0.068371   2.887
CCAAnombreAsturias:I(year - 2000)          0.105571   0.068371   1.544
CCAAnombreCantabria:I(year - 2000)         0.068428   0.068371   1.001
CCAAnombreCastilla la Mancha:I(year - 2000) 0.034041  0.068371   0.498
CCAAnombreCastilla y León:I(year - 2000)   0.144126   0.068371   2.108
CCAAnombreCataluña:I(year - 2000)          0.141730   0.068371   2.073
CCAAnombreCeuta:I(year - 2000)             0.013833   0.068371   0.202
CCAAnombreExtremadura:I(year - 2000)       0.116633   0.068371   1.706
CCAAnombreGalicia:I(year - 2000)           0.205032   0.068371   2.999
CCAAnombreIslas Baleares:I(year - 2000)    0.060249   0.068371   0.881
CCAAnombreIslas Canarias:I(year - 2000)   -0.075057   0.068371  -1.098
CCAAnombreLa Rioja:I(year - 2000)          0.083702   0.068371   1.224
CCAAnombreMadrid:I(year - 2000)            0.292859   0.068371   4.283
CCAAnombreMelilla:I(year - 2000)          -0.124166   0.068371  -1.816
CCAAnombreMurcia:I(year - 2000)            0.028592   0.068371   0.418
CCAAnombreNavarra:I(year - 2000)           0.152970   0.068371   2.237
CCAAnombrePaís Vasco:I(year - 2000)        0.270412   0.068371   3.955
CCAAnombreValencia:I(year - 2000)          0.004054   0.068371   0.059
                                          Pr(>|t|)
(Intercept)                                < 2e-16 ***
CCAAnombreAragón                          1.52e-09 ***
CCAAnombreAsturias                        0.039932 *
CCAAnombreCantabria                       0.000218 ***
CCAAnombreCastilla la Mancha              0.456233
CCAAnombreCastilla y León                 0.005118 **
CCAAnombreCataluña                         < 2e-16 ***
```

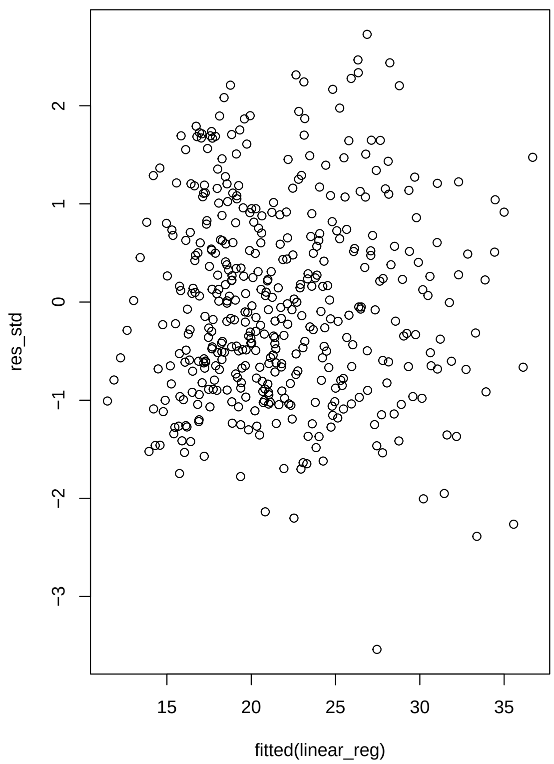
```
CCAAnombreCeuta                             0.094415 .
CCAAnombreExtremadura                       0.005261 **
CCAAnombreGalicia                           0.673478
CCAAnombreIslas Baleares                    4.46e-14 ***
CCAAnombreIslas Canarias                    0.000840 ***
CCAAnombreLa Rioja                          1.77e-12 ***
CCAAnombreMadrid                             < 2e-16 ***
CCAAnombreMelilla                           0.026366 *
CCAAnombreMurcia                            0.081160 .
CCAAnombreNavarra                            < 2e-16 ***
CCAAnombrePaís Vasco                         < 2e-16 ***
CCAAnombreValencia                          0.000786 ***
I(year - 2000)                              3.46e-08 ***
CCAAnombreAragón:I(year - 2000)             0.004105 **
CCAAnombreAsturias:I(year - 2000)           0.123360
CCAAnombreCantabria:I(year - 2000)          0.317516
CCAAnombreCastilla la Mancha:I(year - 2000) 0.618842
CCAAnombreCastilla y León:I(year - 2000)    0.035655 *
CCAAnombreCataluña:I(year - 2000)           0.038818 *
CCAAnombreCeuta:I(year - 2000)              0.839767
CCAAnombreExtremadura:I(year - 2000)        0.088810 .
CCAAnombreGalicia:I(year - 2000)            0.002880 **
CCAAnombreIslas Baleares:I(year - 2000)     0.378734
CCAAnombreIslas Canarias:I(year - 2000)     0.272962
CCAAnombreLa Rioja:I(year - 2000)           0.221591
CCAAnombreMadrid:I(year - 2000)             2.31e-05 ***
CCAAnombreMelilla:I(year - 2000)            0.070112 .
CCAAnombreMurcia:I(year - 2000)             0.676036
CCAAnombreNavarra:I(year - 2000)            0.025816 *
CCAAnombrePaís Vasco:I(year - 2000)         9.05e-05 ***
CCAAnombreValencia:I(year - 2000)           0.952751
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.538 on 399 degrees of freedom
Multiple R-squared:  0.9164,    Adjusted R-squared:  0.9087
F-statistic: 118.2 on 37 and 399 DF,  p-value: < 2.2e-16
```

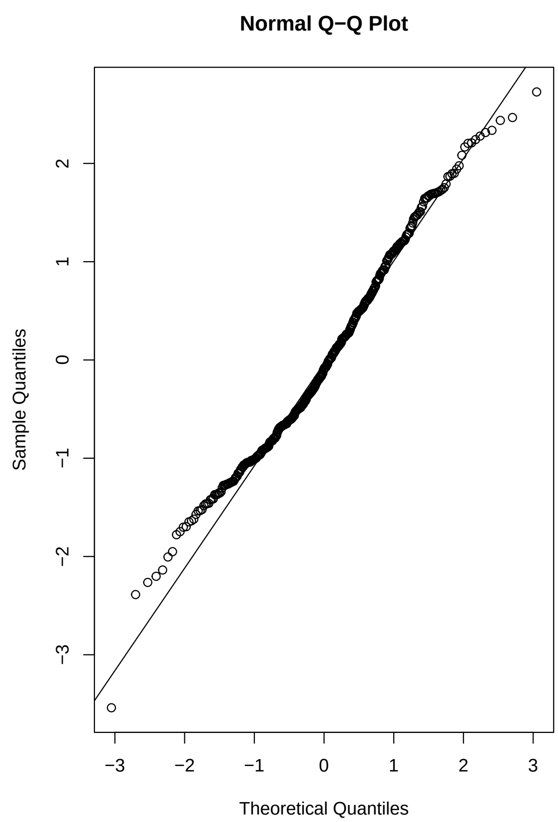
I notice a several things here:

- **Madrid** appears to have the highest level of per capita production. The linear regression results show that it also had the largest annual rate of change since 2000. **País Vaso** also had very high rate of change.

- **Andalucia**, the reference dummy variable in the model above, had one of the lowest levels of GDP per-capita and annual rate of change.
- The model looks very good, with R^2 of 91%. The standard residuals are normally distributed– with a few outliers as expected. I suspect that the large outlier are from 2020, when the world stopped because of COVID-19 and supply chain shocks.

```
par(mfrow = c(1,2))
res_std <- rstandard(linear_reg)
plot(fitted(linear_reg), res_std)
qqnorm(res_std)
qqline(res_std)
```

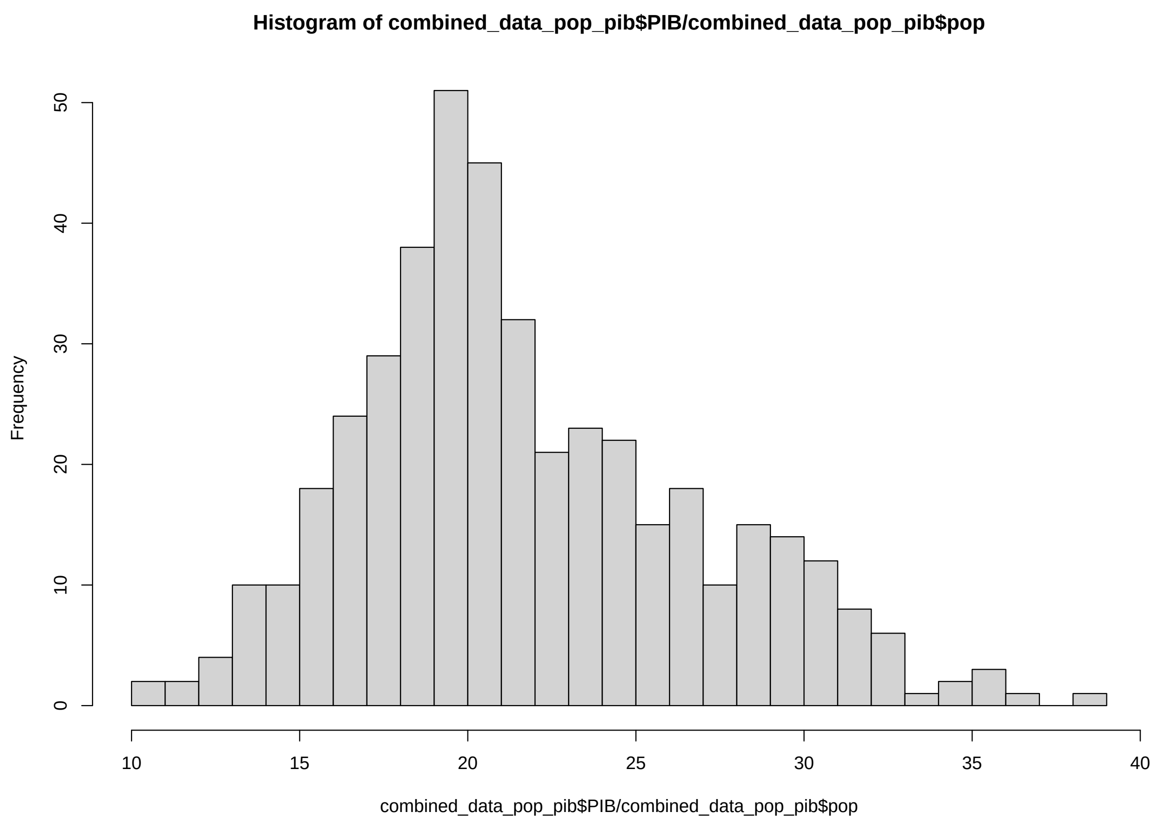




- There are once again different stages to the timelapse of annual economic production. The first stage (2000-2007) has a steep and positve slope. The second stage (2008-2013) has a slight downward slope. The third phase (2014-today) has a positive slope, yet not as steep as the first stage. Important to note is that 2020 was an outlier due to the pandemic and supply chain shocks around the world. It appears that the following years (2021-22) rebounded from the recession and caught up with the third stage trend.

- The target variable GDP/population in the model is on a bell curve, skewed to the right

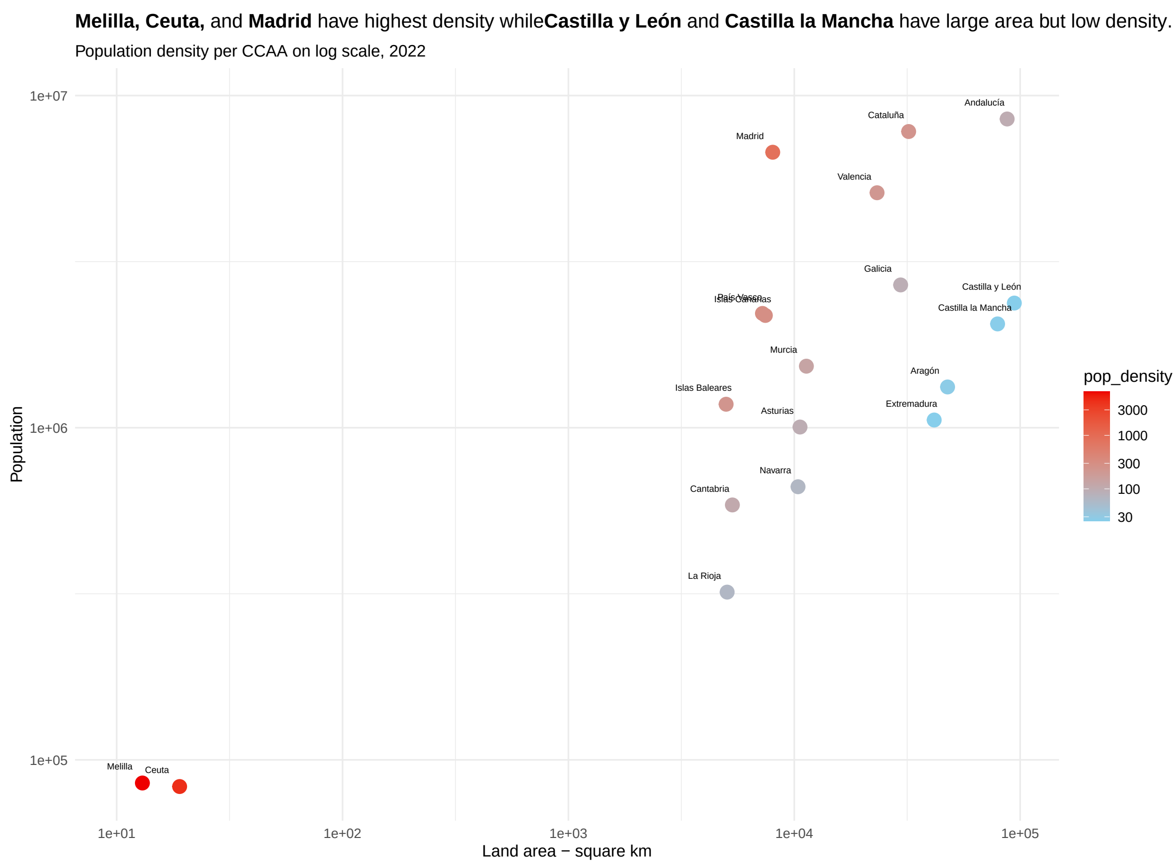
```r
hist(combined_data_pop_pib$PIB/combined_data_pop_pib$pop, breaks = 30)
```



5. Land area

```r
land_area |>
  left_join(combined_data_pop_pib, by = "CCAAnombre") |>
  mutate(pop_density = pop / Superf) |>
  filter(year == 2022) |>
  ggplot(aes(x = Superf, y = pop, group = CCAAnombre, color = pop_density )) +
  geom_point(size = 4) +
  geom_text(aes(label = CCAAnombre),
```

```
                check_overlap = F,
                nudge_x = -0.1,
                size = 2,
                color = "black",
                nudge_y = .05
                ) +
    labs(
      x = "Land area - square km",
      y = "Population",
      subtitle = "Population density per CCAA on log scale, 2022",
      title = expression(paste(bold("Melilla, Ceuta,"), " and ", bold("Madrid"), " have high
    ) +
    scale_y_log10() + scale_x_log10() +
    scale_color_gradient(trans = "log10", low = "skyblue", high = "red2") +
    theme_minimal()
```



Four of the five largest CCAA have the lowest density in Spain. **Andalucía** has the second-largest area but is populated enough to be in the middle. **Melilla** and **Ceuta** are densely populated, but take up less than 20 km^2 each. Of the rest, **Madrid** is the most densely populated comunidad.

Let's see if this matches up with the economic and energy data.

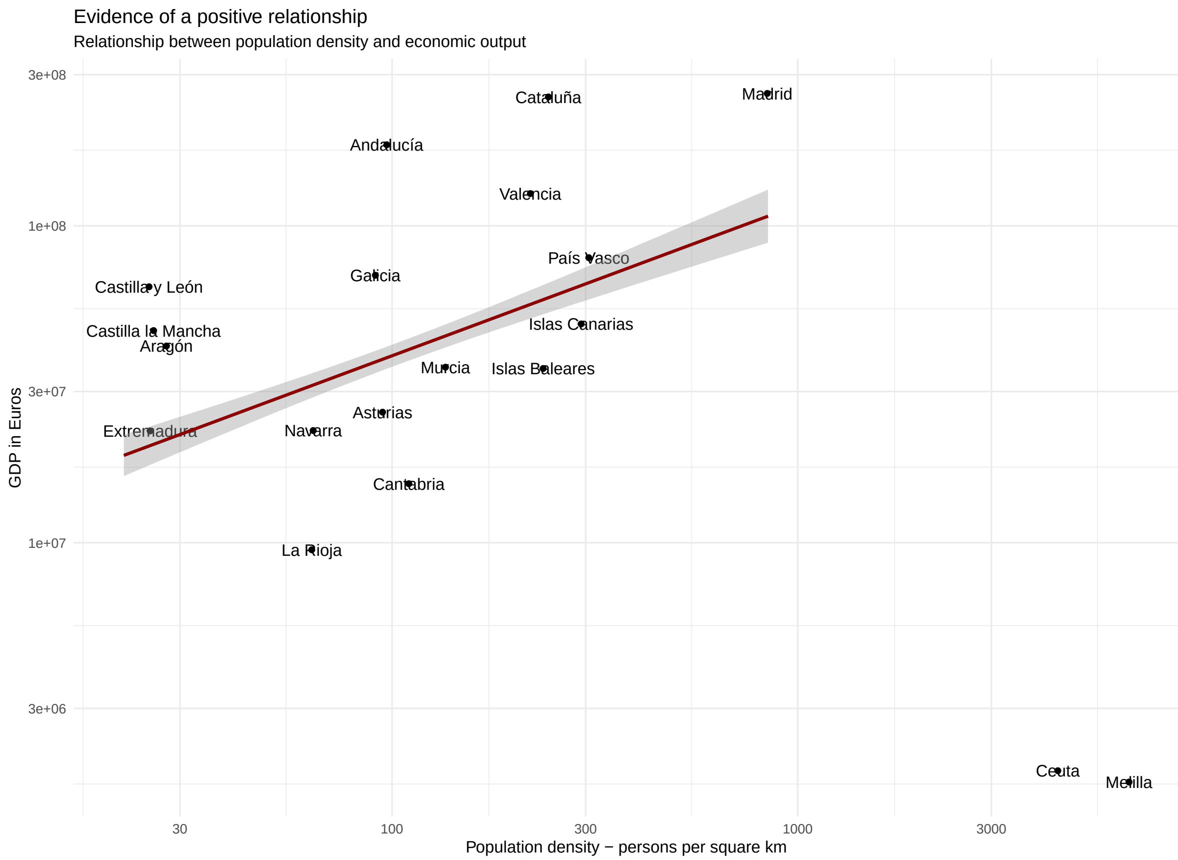
Compare pop density with GDP & Energy output

```
combined_data_pop_pib_land <- land_area |>
  select(CCAAnombre, Superf) |>
  left_join(combined_data_pop_pib, by = "CCAAnombre") |>
  mutate(pop.density = pop / Superf)

subset_wOut_CM_2022 <- combined_data_pop_pib_land |>
  filter( ! CCAAnombre %in% c("Ceuta","Melilla") ) |>
  filter(year == 2022)

cor.test(log10(subset_wOut_CM_2022$pop.density), log10(subset_wOut_CM_2022$PIB))
```

```
	Pearson's product-moment correlation

data:  log10(subset_wOut_CM_2022$pop.density) and log10(subset_wOut_CM_2022$PIB)
t = 2.2483, df = 15, p-value = 0.04002
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02821832 0.79166276
sample estimates:
      cor
0.5020537
```

```
combined_data_pop_pib_land |>
 filter(year == 2022) |>
  ggplot(aes(pop.density, PIB)) +
  geom_text(aes(label = CCAAnombre)) +
  geom_point() +
  scale_x_log10() + scale_y_log10() +
  geom_smooth(
    method = "lm", data = combined_data_pop_pib_land[combined_data_pop_pib_land$pop.densit
    color = "red4", ) +
  theme_minimal() +
```

```
labs(
  x = "Population density - persons per square km",
  y = "GDP in Euros",
  subtitle = "Relationship between population density and economic output",
  title = "Evidence of a positive relationship"
)
```



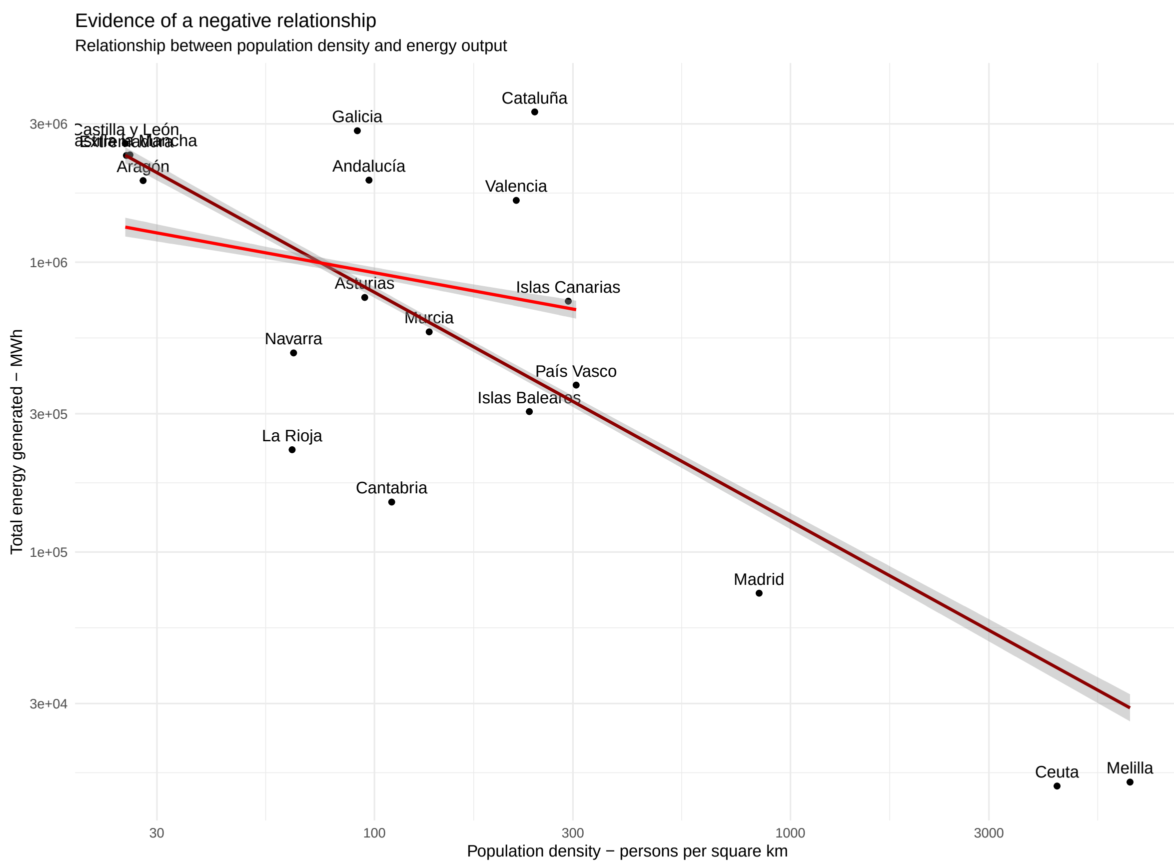
Leaving out the the clear outliers, you can see that there is evidence of a positive relationship between population density and economic output. The Pearson's product moment correlation test shows evidence that there is a relationship. But when you include **Melilla** and **Ceuta** the linear relationship breaks. I think I need another variable to take into account those two comunidades located in the African continent for a linear relationship to work for all the data.

```r
combined_data_pop_pib_land_energy <- combined_data_pop_pib_land |>
  filter(year == 2022) |>
  full_join(red_data, by = join_by("CCAAnombre" == "ccaa"))

subset_total <- combined_data_pop_pib_land_energy |>
  filter( name == "Generación total" )

cor.test(log10(subset_total$pop.density),
         log10(subset_total$value ))
```

```
	Pearson's product-moment correlation

data:  log10(subset_total$pop.density) and log10(subset_total$value)
t = -58.163, df = 2278, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7890531 -0.7559816
sample estimates:
       cor
-0.7730421
```

```r
subset_total |>
 filter(date == as.Date("2023-11-30")) |>
  ggplot(aes(pop.density, value)) +
  geom_text(aes(label = CCAAnombre), nudge_y = 0.05) +
  geom_point() +
  scale_x_log10() +
  scale_y_continuous(
      trans = "log10",
     # labels = scales::label_number( scale_cut = scales::cut_long_scale())
      ) +
  geom_smooth(
    method = "lm",
    color = "red4", data = subset_total) +
   geom_smooth(
    method = "lm",
    color = "red", data = subset_total[subset_total$pop.density < 400,]) +
  theme_minimal() +
  labs(
    x = "Population density - persons per square km",
    y = "Total energy generated - MWh",
```

```
    subtitle = "Relationship between population density and energy output",
    title = "Evidence of a negative relationship"
  )
```



Although there appears to be a relationship between population density and energy generation, the CCAAs with the most generated vary in land area and population. See the top 5 here:

```
subset_total |>
  filter(date == as.Date("2023-11-30")) |>
  slice_max(value, n = 5)
```

```
# A tibble: 5 x 13
  CCAAnombre          Superf id.x       PIB  year id.y      pop pop.density  value
  <chr>                <dbl> <chr>    <dbl> <dbl> <chr>   <dbl>       <dbl>  <dbl>
1 Cataluña             32091 09      2.55e8  2022 09     7.79e6       243.  3.30e6
2 Galicia              29575 12      6.98e7  2022 12     2.69e6        91.0 2.84e6
```

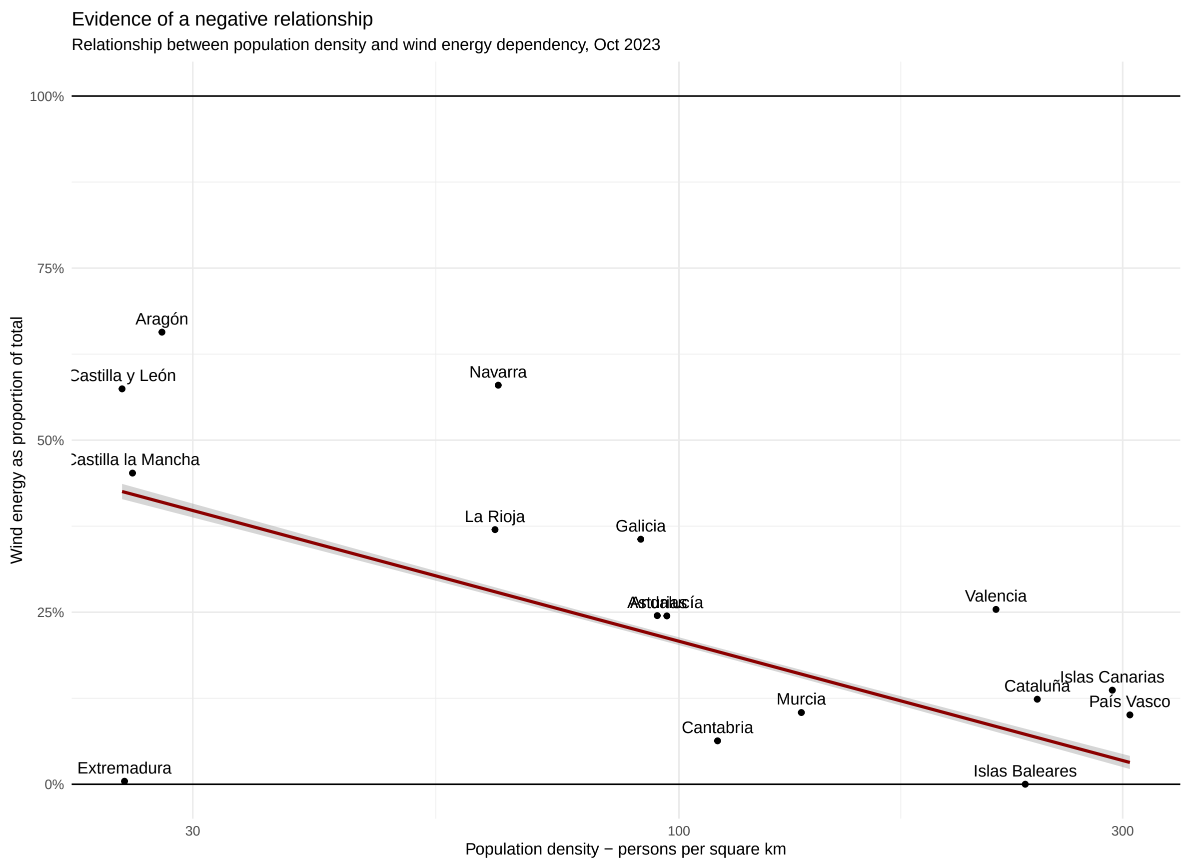
```
3 Castilla y León     94227 07      6.42e7  2022 07    2.37e6         25.2 2.57e6
4 Castilla la Mancha  79462 08      4.67e7  2022 08    2.05e6         25.8 2.35e6
5 Extremadura         41635 11      2.25e7  2022 11    1.05e6         25.3 2.33e6
# i 4 more variables: percentage <dbl>, datetime <dttm>, name <chr>,
#   date <date>
```

```r
subset_wind <- combined_data_pop_pib_land_energy |>
  filter( name == "Eólica" )

cor.test(log10(subset_wind$pop.density),
         log10(subset_wind$value ))
```

```

	Pearson's product-moment correlation

data:  log10(subset_wind$pop.density) and log10(subset_wind$value)
t = -22.484, df = 1854, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4978924 -0.4263582
sample estimates:
       cor 
-0.4628786 
```

```r
subset_wind |>
 filter(date == as.Date("2023-10-31")) |>
  ggplot(aes(pop.density, percentage)) +
  geom_hline(yintercept = c(0,1)) +
  geom_text(aes(label = CCAAnombre),
            nudge_y = 0.02) +
  geom_point() +
  scale_x_log10() +
  scale_y_continuous(
    labels = scales::label_percent()
  ) +
  geom_smooth(
    method = "lm", formula = y ~ x,
    color = "red4",
    data = subset_wind[subset_wind$CCAAnombre != "Extremadura",]) +
  theme_minimal() +
  labs(
    x = "Population density - persons per square km",
```

```
    y = "Wind energy as proportion of total",
    subtitle = "Relationship between population density and wind energy dependency, Oct 20
    title = "Evidence of a negative relationship "
  )
```



Energy generation and GDP

```
library(patchwork)

pltA <- subset_total |>
  filter(date == as.Date("2023-10-31")) |>
  ggplot(aes(PIB, value)) +
  geom_point() +
  geom_text(aes(label = CCAAnombre),
            nudge_y = 0.05,
```

```r
            color = "gray30", size = 2) +
  scale_x_continuous(
    trans = "log10",
   # labels = scales::label_number( scale_cut = scales::cut_long_scale())
    ) +
  scale_y_continuous(
    trans = "log10",
   # labels = scales::label_number( scale_cut = scales::cut_long_scale())
  ) +
  labs(
    y = "Total energy generated - MWh",
    x = "Annual GDP - €",
    subtitle = "Relationship on log scale, Oct 2023"
  ) +
  theme_bw()

pltB <- subset_total |>
  filter(date == as.Date("2023-10-31")) |>
  ggplot(aes(PIB, value)) +
  geom_abline(
    slope = c(1/100, 4/100),
    color = "red4",
    linetype = "dashed"
  ) +
  geom_point()  +
  geom_text(aes(label = CCAAnombre),
           nudge_y = 50000,
            color = "gray30", size = 2
          ) +
  scale_x_continuous(
   # labels = scales::label_number( scale_cut = scales::cut_long_scale())
    ) +
  scale_y_continuous(
   # labels = scales::label_number( scale_cut = scales::cut_long_scale())
  ) +
  labs(
    y = "Total energy generated - MWh",
    x = "Annual GDP - €",
    subtitle = "Relationship on linear scale, Oct 2023"
  ) +
  theme_bw()
```

```r
pltA + pltB + patchwork::plot_annotation(title = "Evidence of both log-log (A) and linear
```

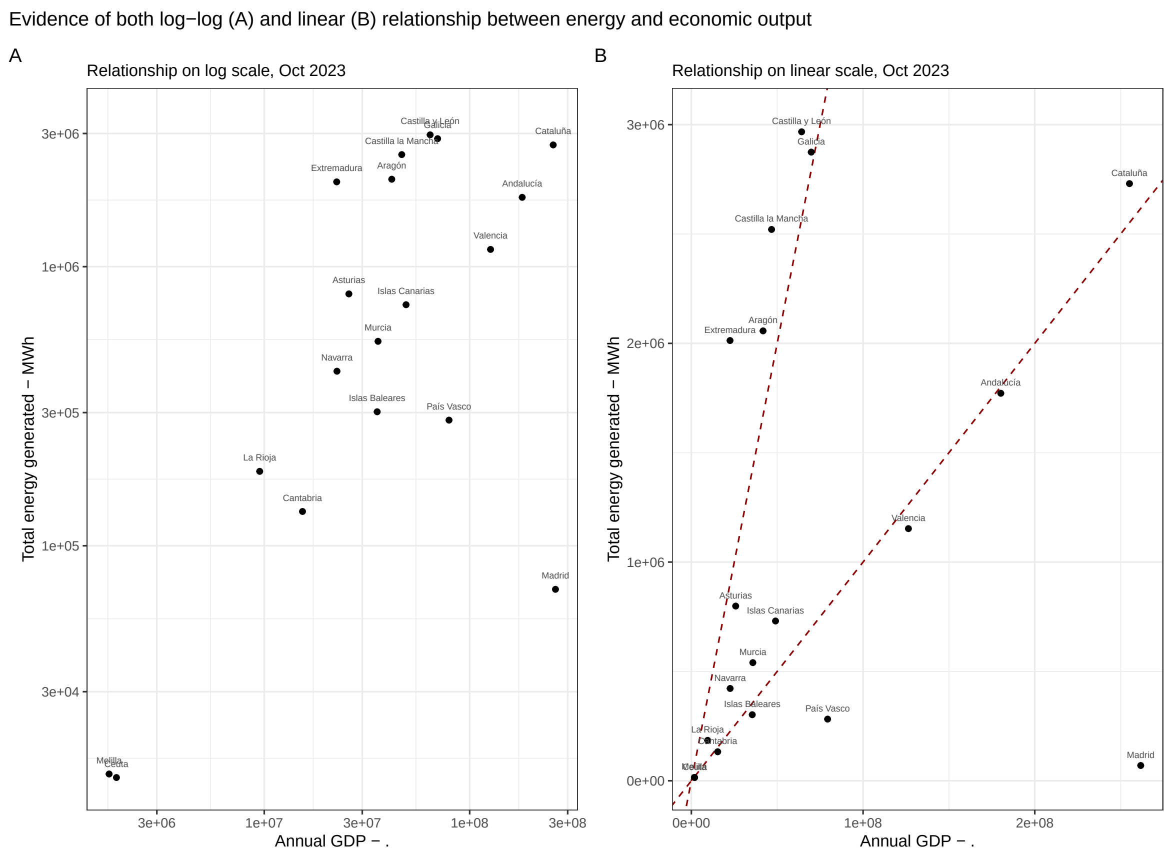


Figure 4: Energy Intensity

Here an interesting thing appears: the regions with the largest population density (**Mellilla**, **Ceuta**, **Madrid**) are not shown because they do not produce any wind energy. But there still appears to be a negative relationship between density and wind energy, as a proportion of total energy. Except for **Extremadura**– a clear outlier– the sparsely populated regions like **Aragón** and the two "Castilla" comunidades generated a lot of their energy from wind.

Extremadura, despite being one of the largest total energy generators, has basically zero wind energy.

Navarra and **La Rioja** generate a lot of their energy from wind, despite being on the low end of total energy.

Conclusions:

Here are key takeaways:

1. There are clear groups with similar energy intensity:
    1. **Galicia**, **Aragón**, **Extremadura** and the two **Castillas**
    2. **Melilla** and **Ceuta**
    3. **Madrid**
    4. The rest.
2. Population density is negatively correlated with proportion of energy generation sourced from wind. Densely populated regions like **Madrid**, **Cataluña**, and **Islas Baleares** have very low wind generation. **Extremadura** is an outlier because it is sparsely populated but has basically no wind generation. One explanation for it's high energy generation is that it has nuclear power plants.
3. Energy generation, measured by Watt-hours, is roughly normally distributed when split up by CCAA. As a whole, the data from years 2014-2023 does not follow a bell curve. But when we split up the data by region, you can see a bell curve and a few trends:
    1. The 2021 to 2023 price inflation crisis did not affect energy generation. The trends from before continued for almost all regions.
    2. There is no clear pattern across CCAA for time trends. Varying trends appear in the data. On the other hand, there is a clear pattern across CCAA for GDP. All the CCAA show the same trend in GDP over time, despite producing at varying levels.
4. Survey results show that different regions vary on opinions about the environment and energy. However, the data show that it is likely that most people, in most CCAA, believe that environmental degradation will get worse. It appears that the most densely populated CCAA have a higher likelihood of having more people believe that climate and energy-related issues are the most important in Spain nowadays.

Possible next steps:

a) PCA to find groups or clusters of CCAA

b) research more literature and discuss with experts

c) Regression models with all data

d) Check for autocorrelation across time

Notes with Prof Pablo

GDP per capita would be useful as covariate.

Share of industry in each region.

Structure of economy has big influence on energy intensity. Financial services are less energy intense, as apposed to industrial economy. Higher share of services will have lower energy intensity.

energy is consumed by individuals and by companies/production/industry.

electricity prices are same everywhere, but not oil, gas, petrol, prices. so it's not relevant unless you take into account inflation for each ccaa. use petrol prices instead.

NO ELECTRICITY TAXES per region

be careful that the explanatory is not being explained by target variable, instead of other away around.

More analysis

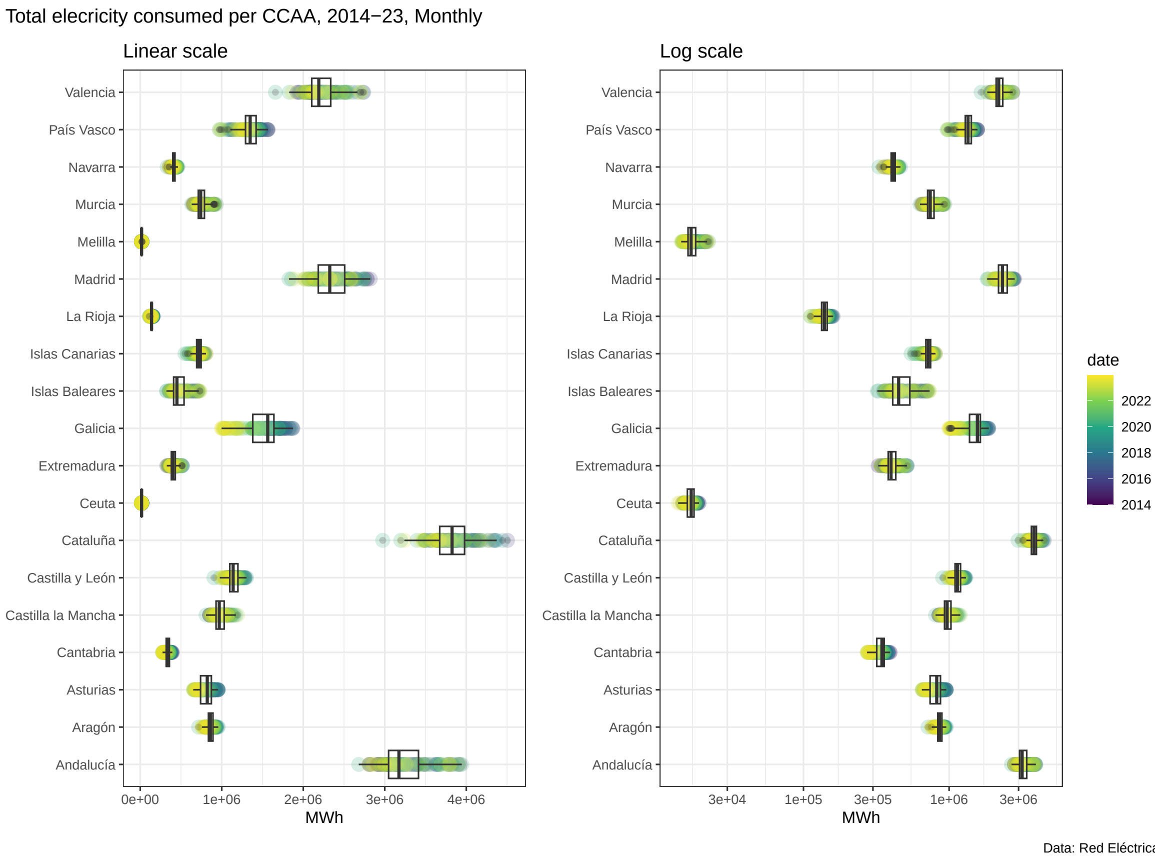
Electricity consumption

```
consumo_data <- read_csv("./consumo_data.csv") |>
  select(-percentage)
```

```
PIVOT_data_pop_pib_land_energy <- combined_data_pop_pib_land_energy |>
  left_join(consumo_data, by = join_by("CCAAnombre" == "ccaa", "datetime")) |>
  pivot_wider(names_from = c(name.x, name.x), values_from = c(value.x, percentage),values_
  mutate(
    month = lubridate::month(lubridate::days(1) + datetime)
  )
```

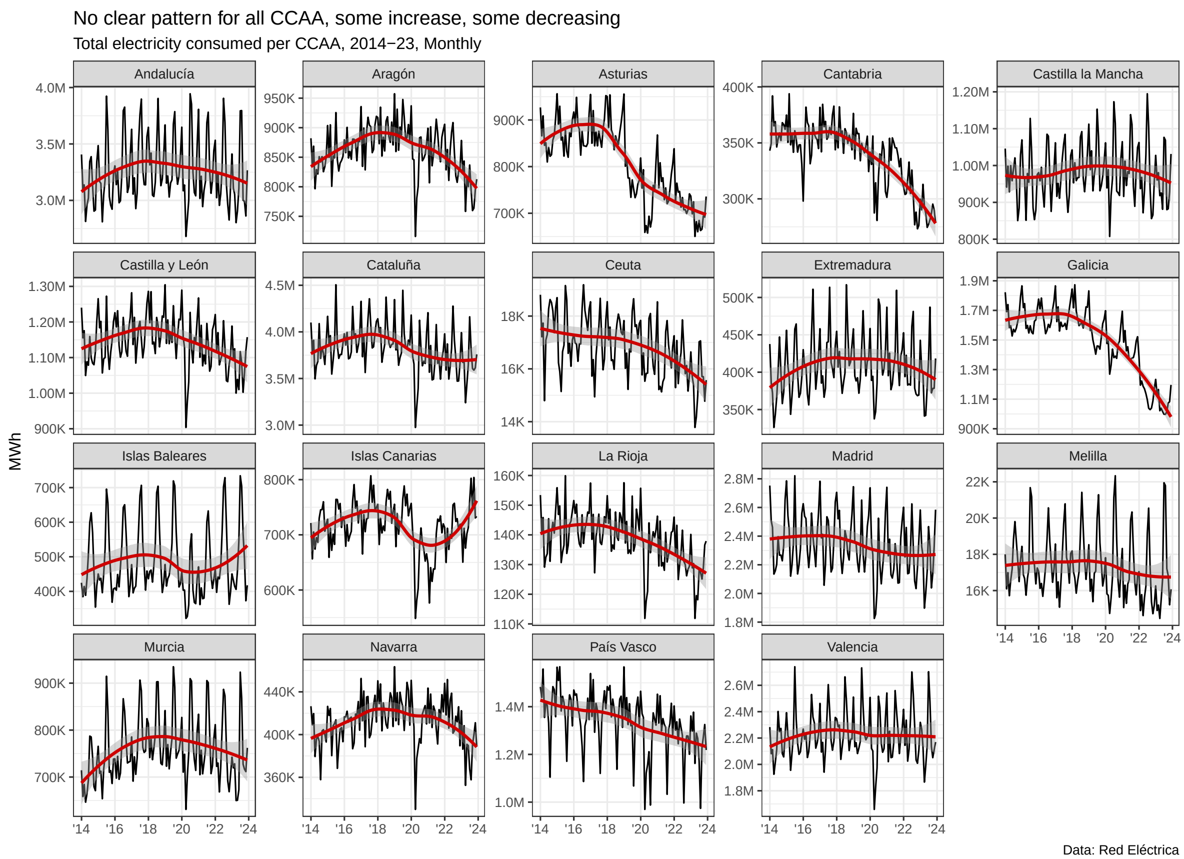
```
library(patchwork)
plta <- consumo_data |>
  ggplot(aes(x = value, y =ccaa, color = date)) +
  geom_point(size =4, alpha = .2) +
  geom_boxplot(alpha = .2) +
  scale_color_continuous(type = "viridis", trans = "date") +
  # scale_x_log10() +
  theme_bw() +
```

```
  labs(
    x = "MWh",
    y = NULL,
    subtitle= NULL,
    title = "Linear scale",
  )

pltb <- consumo_data |>
  ggplot(aes(x = value, y =ccaa, color = date)) +
  geom_point(size =4, alpha = .2) +
  geom_boxplot(alpha = .2) +
  scale_color_continuous(type = "viridis", trans = "date") +
  scale_x_log10() +
  theme_bw() +
  labs(
    x = "MWh",
    y = NULL,
    subtitle = NULL,
    title = "Log scale",
  )

plta + pltb +plot_layout(guides = "collect") +
  plot_annotation(title = "Total elecricity consumed per CCAA, 2014-23, Monthly",
                  caption = "Data: Red Eléctrica")
```




```r
consumo_data |>
  ggplot(aes(x = as.Date(datetime), y = value)) +
  geom_line() +
  geom_smooth(color = "red3") +
  facet_wrap(~ccaa, scales = "free_y") +
  theme_bw() +
  scale_y_continuous(
    labels = scales::label_number(scale_cut = scales::cut_short_scale())
  ) +
  scale_x_date(date_labels = "'%y", minor_breaks = NULL) +
  labs(
    y = "MWh",
    x = NULL,
    subtitle = "Total electricity consumed per CCAA, 2014-23, Monthly",
    title = "No clear pattern for all CCAA, some increase, some decreasing",
    caption = "Data: Red Eléctrica"
```

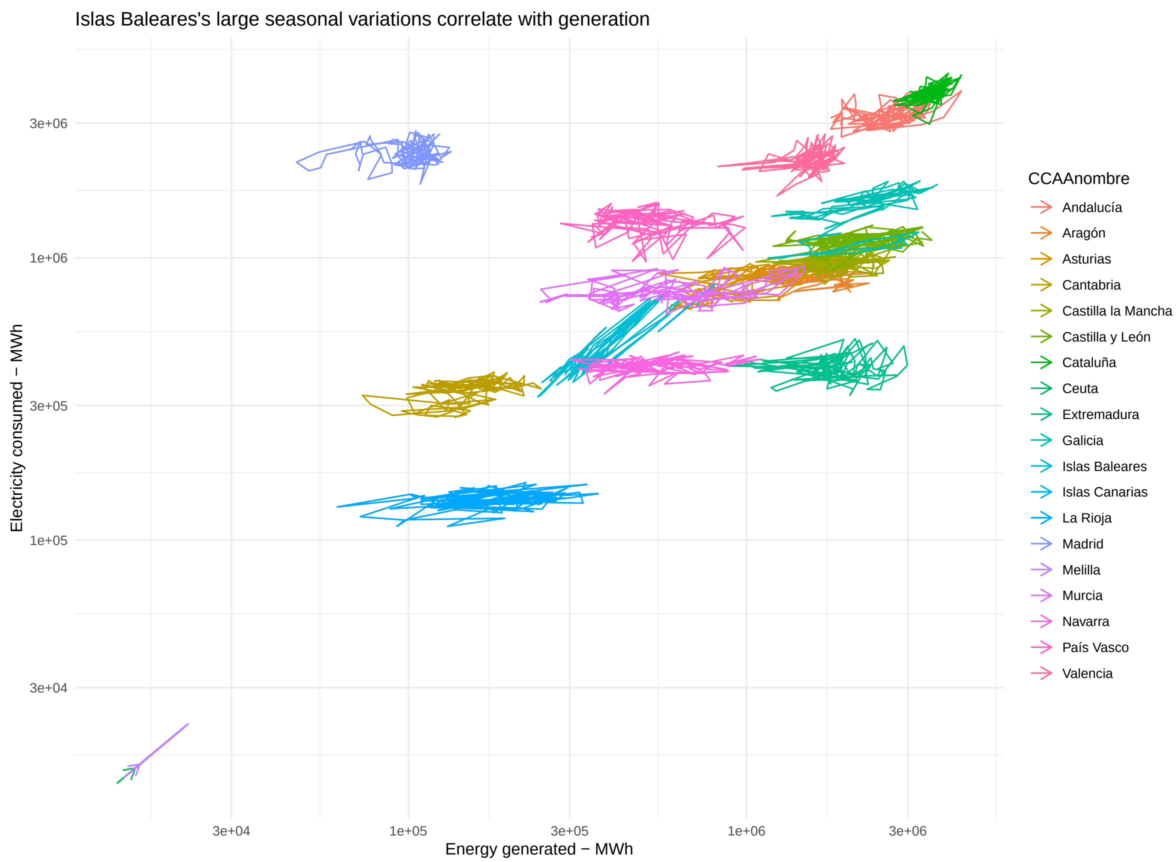
```
)
```



Relationship between consumed and generated

```
PIVOT_data_pop_pib_land_energy |>
  ggplot(aes(`value.x_Generación total`, value.y, color = CCAAnombre, label = CCAAnombre))
  geom_path(arrow = arrow(length = unit(0.3, "cm"))) +
  # geom_text() +
  scale_x_log10() +
  scale_y_log10() +
  labs(
    x = "Energy generated - MWh",
    y = "Electricity consumed - MWh",
    title = "Islas Baleares's large seasonal variations correlate with generation"
  ) +
```

```
  theme_minimal()
```



Islas Baleares does not have wind energy. But it does fluctuate a lot in both generation and consumption at the same times of the year.

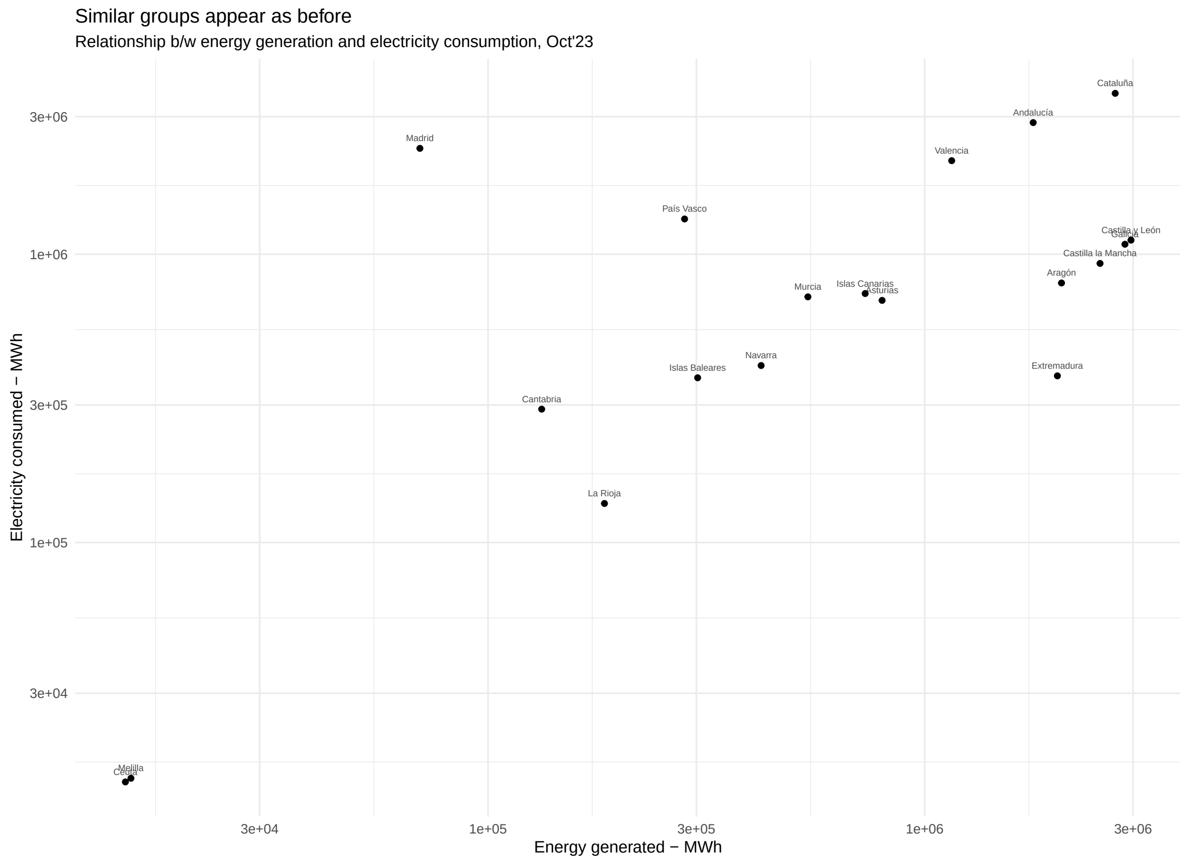
```
islasB <- PIVOT_data_pop_pib_land_energy |>
  filter(CCAAnombre == 'Islas Baleares')
cor.test(islasB$`value.x_Generación total`, islasB$value.y)
```

```

	Pearson's product-moment correlation

data:  islasB$`value.x_Generación total` and islasB$value.y
t = 26.386, df = 118, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8935675 0.9469871
```

```
sample estimates:
      cor
0.9247043
```

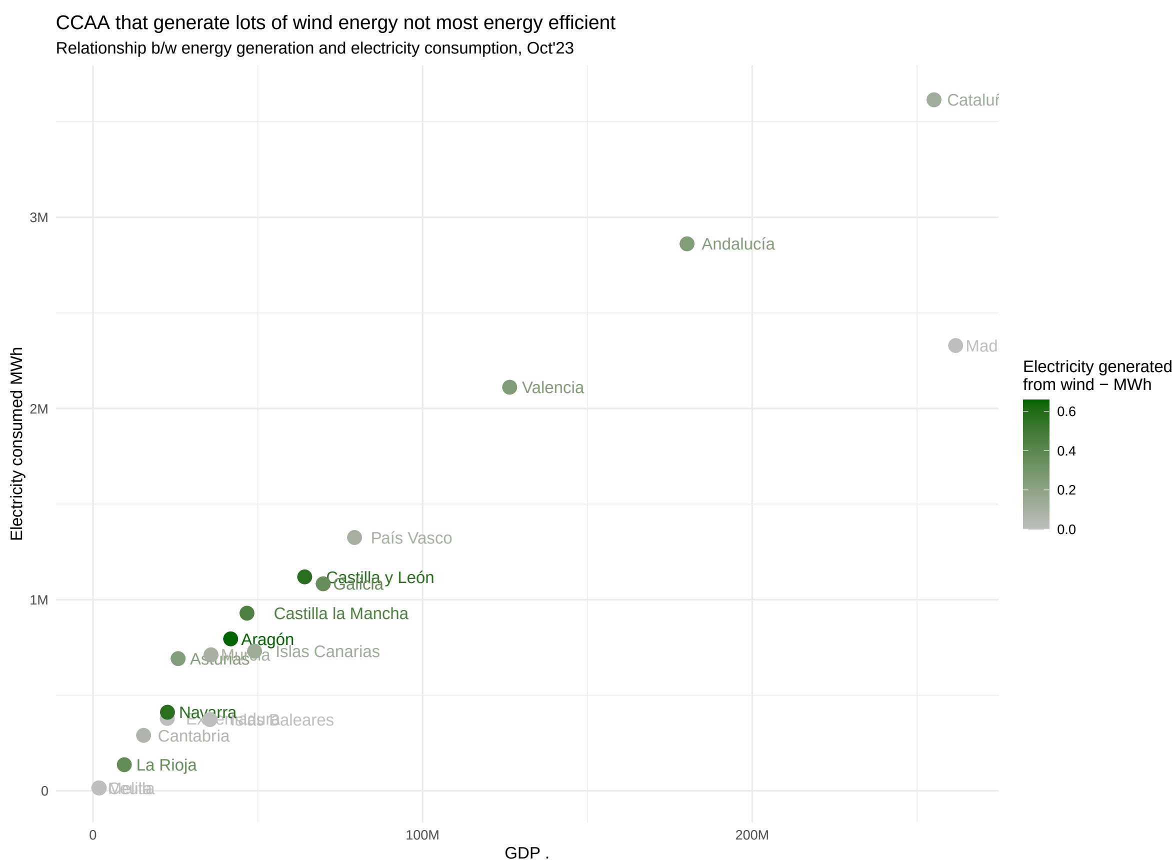
```r
DataExplorer::create_report(PIVOT_data_pop_pib_land_energy)
```

```r
PIVOT_data_pop_pib_land_energy |>
  filter(date.x == as.Date("2023-10-31")) |>
  ggplot(aes(x = `value.x_Generación total`, y = value.y, label = CCAAnombre)) +
  geom_text(
           vjust = -1.1,
            color = "gray30", size = 2
          ) +
  geom_point() +
  scale_x_continuous(
    trans = "log10",
   # labels = scales::label_number( scale_cut = scales::cut_long_scale())
  ) +
  scale_y_continuous(
    trans = "log10",
   # labels = scales::label_number( scale_cut = scales::cut_long_scale())
  ) +
  labs(
    x = "Energy generated - MWh",
    y = "Electricity consumed - MWh",
    title = "Similar groups appear as before",
    subtitle = "Relationship b/w energy generation and electricity consumption, Oct'23"
  ) +
  theme_minimal()
```




```
PIVOT_data_pop_pib_land_energy |>
 #  filter(pop > 100000) |>
  filter(date.x == as.Date("2023-10-31")) |>
  ggplot(aes(x = PIB, y = value.y, label = CCAAnombre,
             color = percentage_Eólica)) +
  geom_text(
          hjust = -.20, size = 4
         ) +
  geom_point(size = 4) +
  scale_x_continuous(
    # trans = "log10",
    labels = scales::label_number( scale_cut = scales::cut_long_scale())
  ) +
  scale_y_continuous(
    # trans = "log10",
    labels = scales::label_number( scale_cut = scales::cut_long_scale())
```

```
) +
scale_color_gradient(labels = scales::label_number( scale_cut = scales::cut_short_scale(
  low = "gray",
  high = "darkgreen"
) +
labs(
  x = "GDP €",
  y = "Electricity consumed MWh",
  title = "CCAA that generate lots of wind energy not most energy efficient",
  subtitle = "Relationship b/w energy generation and electricity consumption, Oct'23",
  color= "Electricity generated\nfrom wind - MWh"
) +
theme_minimal()
```



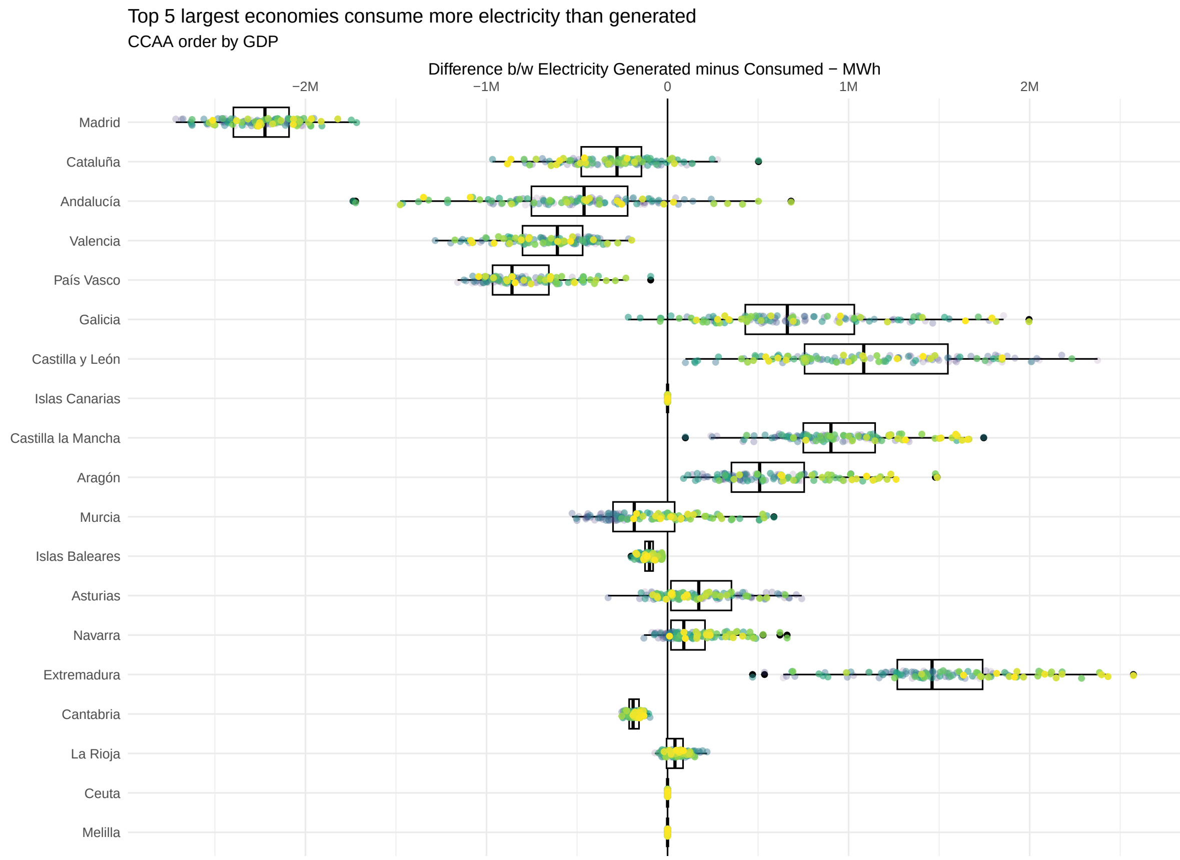
Energy consumption is strongly correlated to GDP

```r
cor.test(PIVOT_data_pop_pib_land_energy$PIB, PIVOT_data_pop_pib_land_energy$value.y)
```

```
	Pearson's product-moment correlation

data:  PIVOT_data_pop_pib_land_energy$PIB and PIVOT_data_pop_pib_land_energy$value.y
t = 113.74, df = 2278, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9157191 0.9280289
sample estimates:
     cor 
0.922107 
```

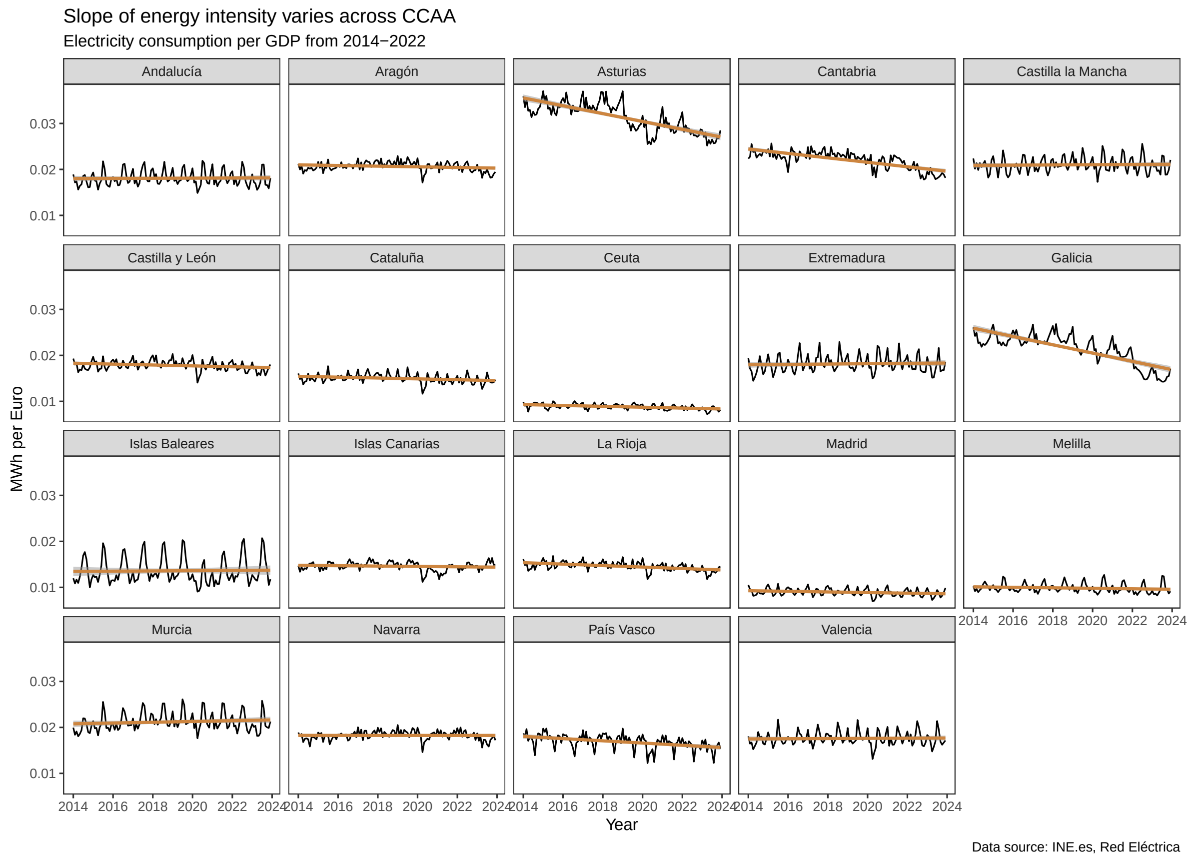
Difference between consumption and generated? And compared to GDP?

```r
PIVOT_data_pop_pib_land_energy |>
  ggplot(aes(x = `value.x_Generación total`- value.y,
             y = reorder(CCAAnombre, PIB, max),
             # color = 0 < `value.x_Generación total`- value.y,
             color = datetime,
             alpha = datetime)) +
  geom_vline(xintercept = 0) +
  geom_boxplot(color = "black", alpha = 1) +
  geom_jitter(height = .1) +
  theme_minimal() +
  labs(
    y = NULL,
    x = "Difference b/w Electricity Generated minus Consumed - MWh",
    subtitle = "CCAA order by GDP",
    title = "Top 5 largest economies consume more electricity than generated"
  ) +
  scale_x_continuous(
    labels = scales::label_number(scale_cut = scales::cut_long_scale()),
    position = "top"
  ) +
  scale_color_continuous(type = "viridis") +
  theme(
    legend.position = "none"
  )
```




```
PIVOT_data_pop_pib_land_energy |>
  ggplot(aes(y = value.y/PIB, x = datetime)) +
  geom_line() +
  geom_smooth(color = "peru", alpha = .5, method = "lm") +
  facet_wrap(~CCAAnombre, scales = "fixed", ncol = 5) +
  scale_y_continuous(
  ) +
#  scale_x_continuous(breaks = c(2000, 2010, 2020)) +
  theme_bw() +
  labs( title = "Slope of energy intensity varies across CCAA",
    subtitle = "Electricity consumption per GDP from 2014-2022",
    x = "Year",
    y = "MWh per Euro",
    caption = "Data source: INE.es, Red Eléctrica"
  ) +
  theme(
```

```
  panel.grid = element_blank()
)
```



Questions to narrrow down on?

1. **Extremadura** is an outlier in no wind. Why?
2. Aragón is

```
library(lme4)
lmm1 <- lme4::lmer(value.y ~ date.x + (1 | CCAAnombre) + PIB ,
                   data = PIVOT_data_pop_pib_land_energy)
summary(lmm1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: value.y ~ date.x + (1 | CCAAnombre) + PIB
   Data: PIVOT_data_pop_pib_land_energy

REML criterion at convergence: 60358.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.3506 -0.3295 -0.0190  0.2680  5.2786

Random effects:
 Groups     Name        Variance Std.Dev.
 CCAAnombre (Intercept) 1.65e+11 406257
 Residual               1.75e+10 132286
Number of obs: 2280, groups:  CCAAnombre, 19

Fixed effects:
              Estimate Std. Error t value
(Intercept)  6.443e+05  1.350e+05   4.772
date.x      -2.235e+01  2.628e+00  -8.505
PIB          1.258e-02  1.209e-03  10.404

Correlation of Fixed Effects:
       (Intr) date.x
date.x -0.348
PIB    -0.634  0.000
fit warnings:
Some predictor variables are on very different scales: consider rescaling
```

```r
save.image()
```