Guided Project: Analyzing Forest Fire Data

Vamos analisar uma base de dados sobre incêndios florestais através de visualizações para encontrar tendências.

A base de dados pode ser obtida através **deste link**.

Bibliotecas Utilizadas

```
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
```

Coleta da Base

Temos uma base com 517 registros e 13 varíáveis a serem exploradas que são informações meteorológicas, índices/métricas usados por cientístas.

```
base <- read_csv(paste0("https://archive.ics.uci.edu/ml/",
                        "machine-learning-databases/forest-fires/forestfires.csv"))
```

```
## Rows: 517 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): month, day
## dbl (11): X, Y, FFMC, DMC, DC, ISI, temp, RH, wind, rain, area
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(base)
```

```
## Rows: 517
## Columns: 13
## $ X     <dbl> 7, 7, 7, 8, 8, 8, 8, 8, 8, 7, 7, 7, 6, 6, 6, 6, 5, 8, 6, 6, 6, 5~
## $ Y     <dbl> 5, 4, 4, 6, 6, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4~
## $ month <chr> "mar", "oct", "oct", "mar", "mar", "aug", "aug", "aug", "sep", "~
## $ day   <chr> "fri", "tue", "sat", "fri", "sun", "sun", "mon", "mon", "tue", "~
## $ FFMC  <dbl> 86.2, 90.6, 90.6, 91.7, 89.3, 92.3, 92.3, 91.5, 91.0, 92.5, 92.5~
## $ DMC   <dbl> 26.2, 35.4, 43.7, 33.3, 51.3, 85.3, 88.9, 145.4, 129.5, 88.0, 88~
## $ DC    <dbl> 94.3, 669.1, 686.9, 77.5, 102.2, 488.0, 495.6, 608.2, 692.6, 698~
## $ ISI   <dbl> 5.1, 6.7, 6.7, 9.0, 9.6, 14.7, 8.5, 10.7, 7.0, 7.1, 7.1, 22.6, 0~
## $ temp  <dbl> 8.2, 18.0, 14.6, 8.3, 11.4, 22.2, 24.1, 8.0, 13.1, 22.8, 17.8, 1~
## $ RH    <dbl> 51, 33, 33, 97, 99, 29, 27, 86, 63, 40, 51, 38, 72, 42, 21, 44, ~
## $ wind  <dbl> 6.7, 0.9, 1.3, 4.0, 1.8, 5.4, 3.1, 2.2, 5.4, 4.0, 7.2, 4.0, 6.7,~
## $ rain  <dbl> 0.0, 0.0, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,~
## $ area  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Dicionário disponibilizado pela Dataquest para o exercício

Campo	Significado
X	Coordenada espacial do eixo X no mapa do parque de Montesinho: 1 a 9
Y	Coordenada espacial do eixo Y no mapa do parque de Montesinho: 2 a 9
month	Mês do ano: ‘jan’ a ‘dec’
day	Dia da semana: mon’ a ‘sun’
FFMC	Índice Fine Fuel Moisture Code do Sistema FWI: 18.7 a 96.20
DMC	Índice Duff Moisture Code do Sistema FWI: 1.1 a 291.3
DC	Índice Drought Code do Sistema FWI: 7.9 to 860.6
ISI	Índice Initial Spread Index do Sistema FWI: 0.0 to 56.10
temp	Temperatura em graus Celsius: 2.2 to 33.30
RH	Umidade Relativa em percentual: 15.0 a 100
wind	Velocidade do vento em km/h: 0.40 to 9.40
rain	Volume de chuva em mm/m2 : 0.0 a 6.4
area	Área queimada da floresta (em hectares): 0.00 to 1090.84

A base é composta por registros onde cada registro representa um momento onde pode ou não ter ocorrido um incêndio, apresentando os valores medidos/calculados para cada métrica naquele período e local. A data da ocorrência não é clara, só é possível ver as informações consolidadas no mês ou dia da semana.

```r
#conferindo duplicidades ao agrupar por local e data
base %>%
  group_by(X,Y, month,day) %>%
  filter(n()>1) %>%
  summarise(Qtd = n()) %>%
  head()
```

```
## `summarise()` has grouped output by 'X', 'Y', 'month'. You can override using
## the `.groups` argument.
```

```
## # A tibble: 6 x 5
## # Groups:   X, Y, month [2]
##       X     Y month day     Qtd
##   <dbl> <dbl> <chr> <chr> <int>
## 1     1     2 aug   fri       3
## 2     1     2 aug   sun       2
## 3     1     2 aug   thu       2
## 4     1     2 aug   tue       2
## 5     1     2 aug   wed       2
## 6     1     2 sep   thu       3
```

```r
#checando duplicidades, 513 de 517 linhas são de fato distintas
base %>% distinct() %>% nrow()
```

```
## [1] 513
```

Resumo da distribuição das variáveis numéricas

```r
base %>% select(-month,-day) %>% summary()
```

```
##        X               Y              FFMC             DMC              DC
##  Min.   :1.000   Min.   :2.0   Min.   :18.70   Min.   :  1.1   Min.   :  7.9
##  1st Qu.:3.000   1st Qu.:4.0   1st Qu.:90.20   1st Qu.: 68.6   1st Qu.:437.7
##  Median :4.000   Median :4.0   Median :91.60   Median :108.3   Median :664.2
##  Mean   :4.669   Mean   :4.3   Mean   :90.64   Mean   :110.9   Mean   :547.9
##  3rd Qu.:7.000   3rd Qu.:5.0   3rd Qu.:92.90   3rd Qu.:142.4   3rd Qu.:713.9
##  Max.   :9.000   Max.   :9.0   Max.   :96.20   Max.   :291.3   Max.   :860.6
##       ISI              temp             RH              wind
##  Min.   : 0.000   Min.   : 2.20   Min.   : 15.00   Min.   :0.400
##  1st Qu.: 6.500   1st Qu.:15.50   1st Qu.: 33.00   1st Qu.:2.700
##  Median : 8.400   Median :19.30   Median : 42.00   Median :4.000
```

```
##  Mean   : 9.022   Mean   :18.89   Mean   : 44.29   Mean   :4.018
##  3rd Qu.:10.800   3rd Qu.:22.80   3rd Qu.: 53.00   3rd Qu.:4.900
##  Max.   :56.100   Max.   :33.30   Max.   :100.00   Max.   :9.400
##       rain              area
##  Min.   :0.00000   Min.   :   0.00
##  1st Qu.:0.00000   1st Qu.:   0.00
##  Median :0.00000   Median :   0.52
##  Mean   :0.02166   Mean   :  12.85
##  3rd Qu.:0.00000   3rd Qu.:   6.57
##  Max.   :6.40000   Max.   :1090.84
```

Muitas das informações acima são muito específicas para cientistas que trabalham com esses dados e a primeira vista não nos dizem muita coisa. No entanto **neste link** é possível entender a construção desses índices e vai esclarecer muito da relação entre as variáveis que vamos explorar a seguir.

Transformando informações categóricas

As informações como Mês e Dia da Senana vão ser úteis nos gráficos, mas por serem categóricas precisam ser configuradas para aparecerem na ordem certa. Caso contrário vão aparecer em ordem alfabética e não fará muito sentido nas análises.

```
base$month %>% table()
```

```
## .
## apr aug dec feb jan jul jun mar may nov oct sep
##   9 184   9  20   2  32  17  54   2   1  15 172
```

```
base$day %>% table()
```

```
## .
## fri mon sat sun thu tue wed
##  85  74  84  95  61  64  54
```

```
base <- base %>%
  mutate(month = factor(month,
                        levels = c("jan","feb","mar","apr","may","jun",
                                   "jul","aug","sep","oct","nov","dec")),
         day   = factor(day,
                        levels = c("sun","mon","tue","wed","thu","fri","sat")))
```

```
base %>% pull(month) %>% levels()
```

```
##  [1] "jan" "feb" "mar" "apr" "may" "jun" "jul" "aug" "sep" "oct" "nov" "dec"
```

```
base %>% pull(day) %>% levels()
```

```
## [1] "sun" "mon" "tue" "wed" "thu" "fri" "sat"
```

Analisando a frequência dos incêncios ao longo do tempo

Aqui o intuito é entender a sazonalidade dos eventos, se existem épocas do ano ou até dias da semana mais propícios para esse evento ocorrer.

Criando tabela frequência

```
month <- base %>%
  group_by(month) %>%
  summarise(Qtd = n())

month
```

```
## # A tibble: 12 x 2
##    month   Qtd
##    <fct> <int>
##  1 jan       2
##  2 feb      20
##  3 mar      54
##  4 apr       9
##  5 may       2
##  6 jun      17
##  7 jul      32
##  8 aug     184
##  9 sep     172
## 10 oct      15
## 11 nov       1
## 12 dec       9
```

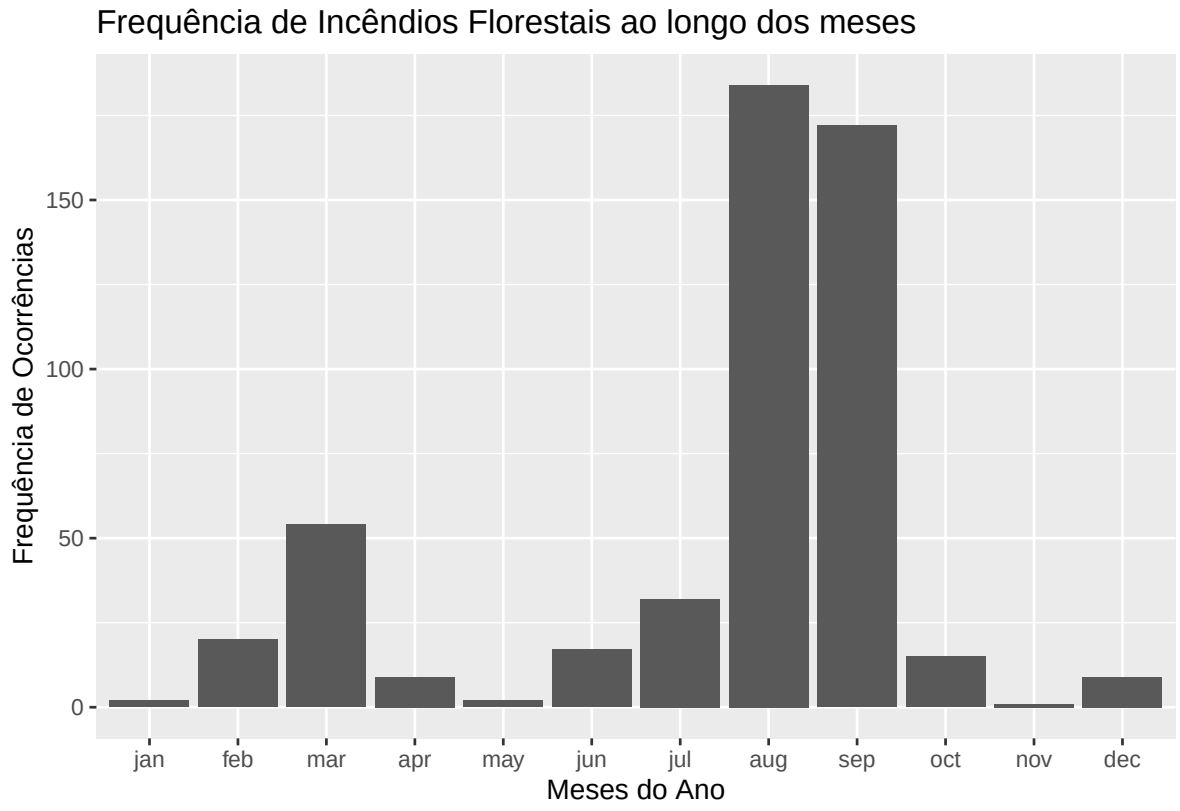
```
day <- base %>%
  group_by(day) %>%
  summarise(Qtd = n())

day
```

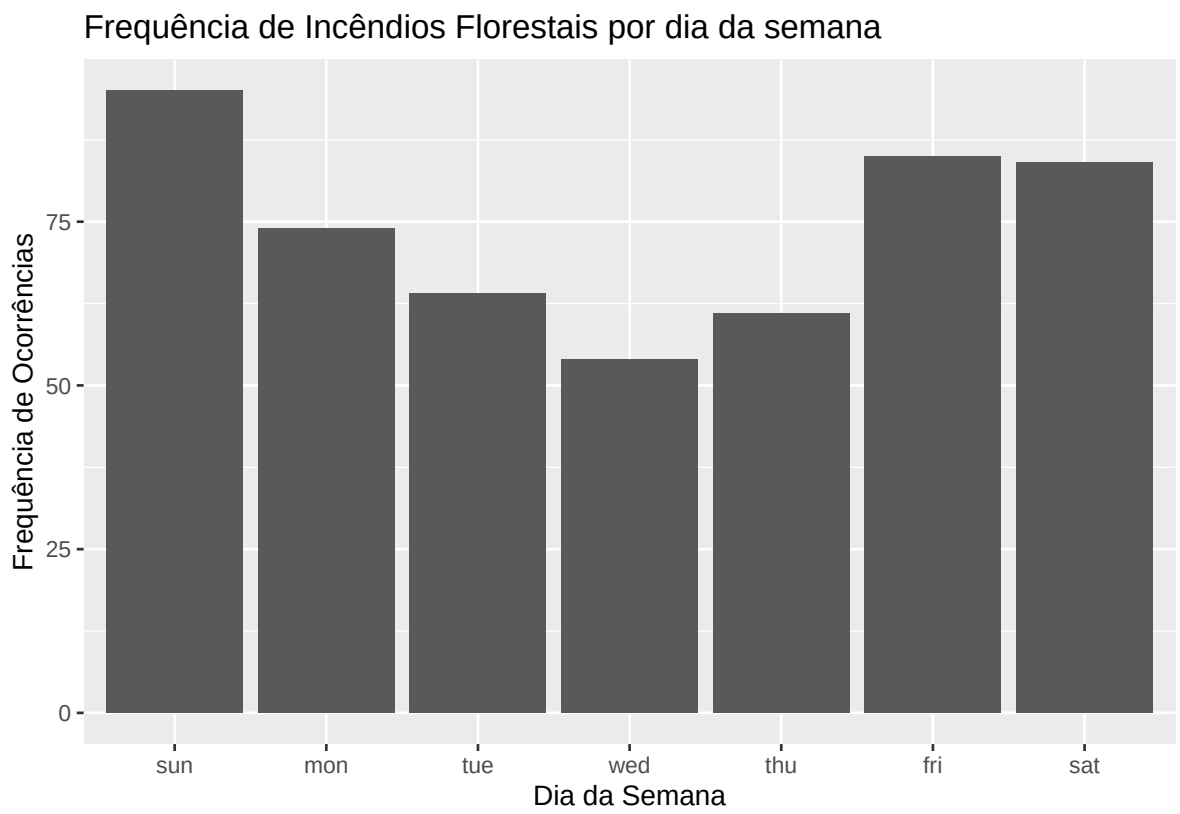
```
## # A tibble: 7 x 2
##   day     Qtd
##   <fct> <int>
## 1 sun      95
## 2 mon      74
## 3 tue      64
## 4 wed      54
## 5 thu      61
## 6 fri      85
## 7 sat      84
```

Visualizando resultado

```
month %>%
  ggplot(aes(x=month, y=Qtd)) +
  geom_col() +
  labs(title = "Frequência de Incêndios Florestais ao longo dos meses",
       x     = "Meses do Ano",
       y     = "Frequência de Ocorrências")
```



```
day %>%
  ggplot(aes(x=day, y=Qtd)) +
  geom_col()+
  labs(title = "Frequência de Incêndios Florestais por dia da semana",
       x     = "Dia da Semana",
       y     = "Frequência de Ocorrências")
```



Com os gráficos acima foi possível notar que os incendios floretais que estamos analisando costumam acontecer com mais frequência entre Agosto e Setembro e existe uma tendência de acentuar no final de semana.

Encontrando informações que expliquem as tendências

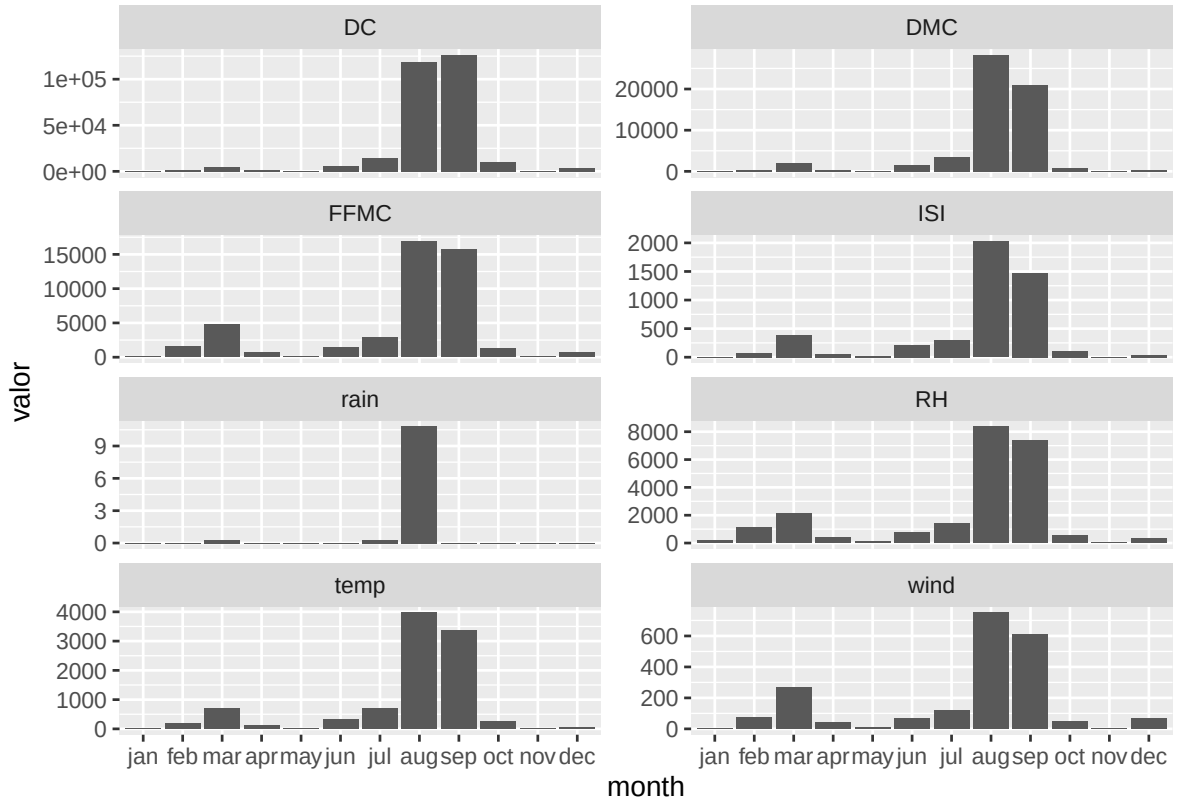
Vamos agora criar gráficos com as informações meteorológicas que temos para descobrir quais tem maior influência nessa tendência ao longo do tempo que enxergamos nos gráficos acima.

Para conseguir visualizações melhores, começamos pivotando as informações da base.

```
pivot <- base %>%
  pivot_longer(c("FFMC","DMC","DC","ISI","temp","RH","wind","rain"),
               names_to = "indicador",
               values_to = "valor")
pivot %>% head(3)
```

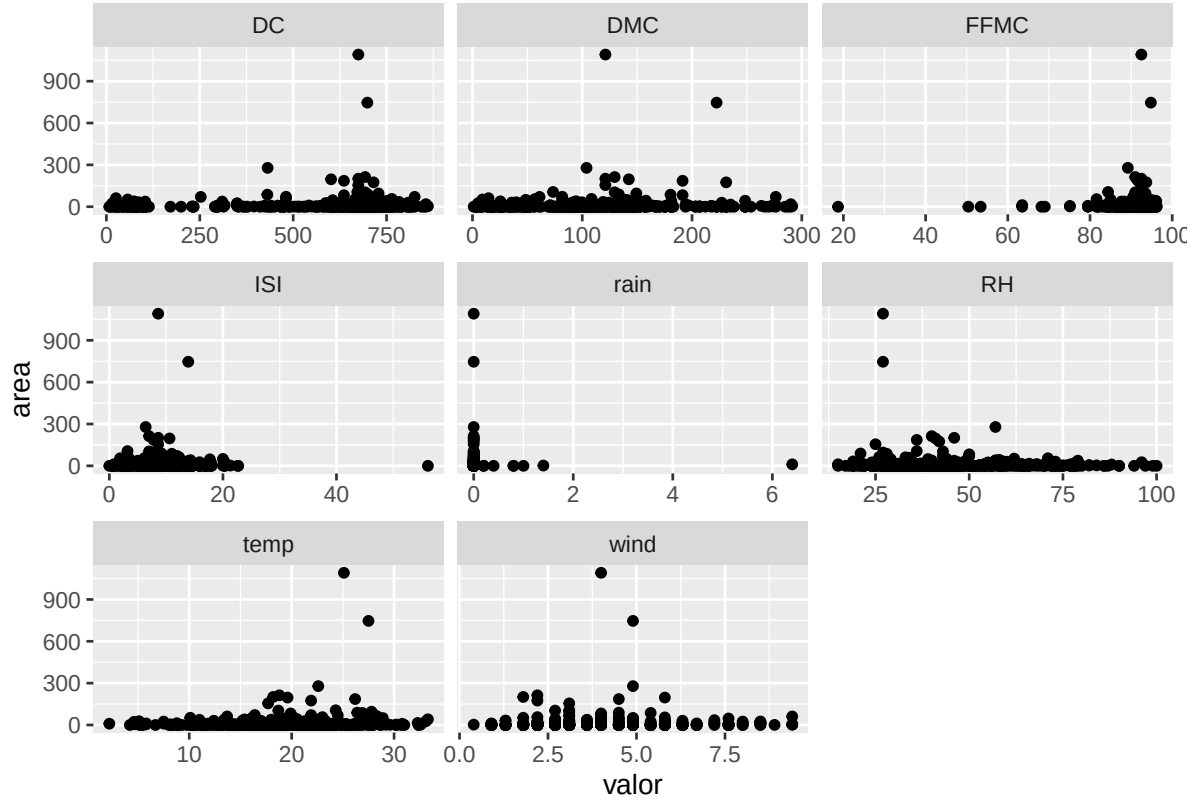
```
## # A tibble: 3 x 7
##       X     Y month day    area indicador valor
##   <dbl> <dbl> <fct> <fct> <dbl> <chr>     <dbl>
## 1     7     5 mar   fri       0 FFMC       86.2
## 2     7     5 mar   fri       0 DMC        26.2
## 3     7     5 mar   fri       0 DC         94.3
```

```
pivot %>%
  ggplot(aes(x=month,y=valor)) +
  geom_col() +
  facet_wrap(facets = "indicador",
             scales = "free_y",
             nrow = 4, ncol = 2)
```



Aparentemente, praticamente todas as métricas em análise tiveram pico nos meses de agosto e setembro. No entanto aqui estamos olhando para a frequência de ocorrências. Um ponto que falta ser levado em consideração é a intensidade do incêncio. Uma forma de analisar isso é pela área que foi afetada, quanto maior a área, mais intenso o incêndio.

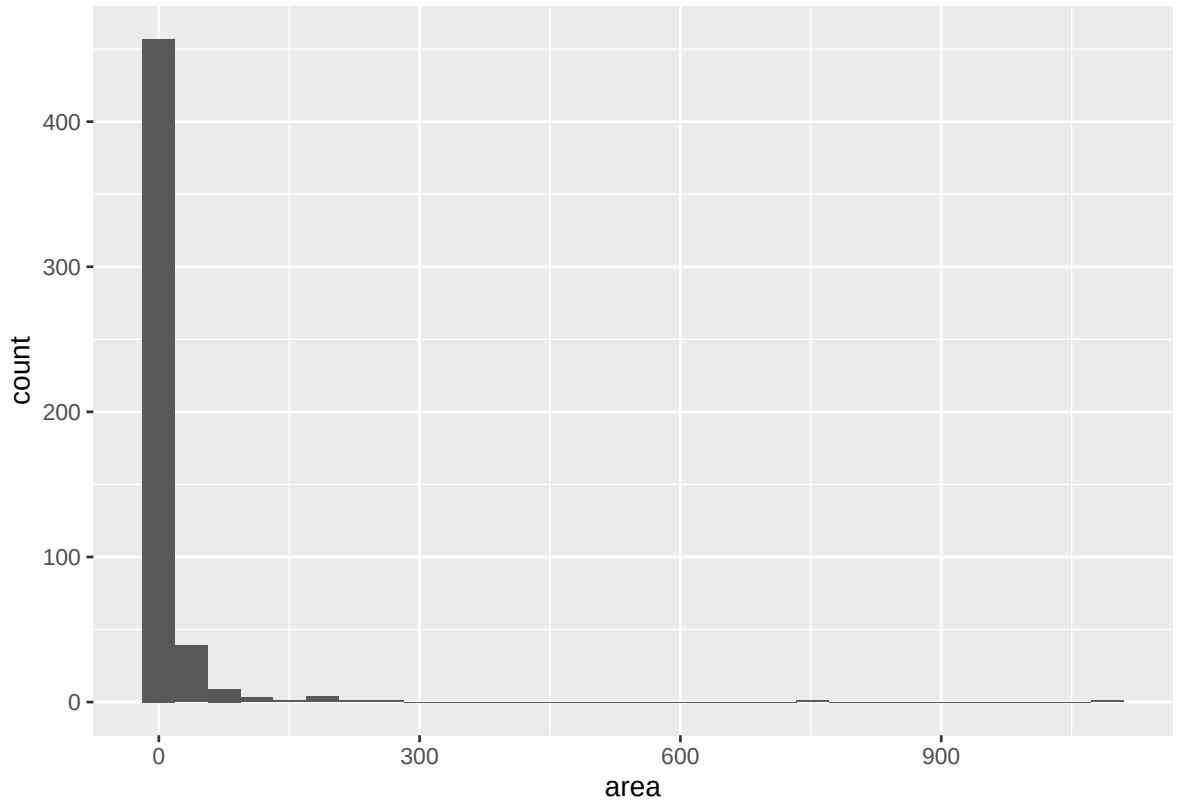
```
pivot %>%
  ggplot(aes(x=valor,y=area)) +
  geom_point() +
  facet_wrap(
    facets = "indicador",
    scales = "free_x",
    nrow = 3,
    ncol = 3)
```



Os gráficos acima não estão ajudando a chegar a uma conclusão, no entando olhando mais a fundo temos valores extremos (outliers) que estão prejudicando a visualização e impedindo de ter uma análise efetiva.

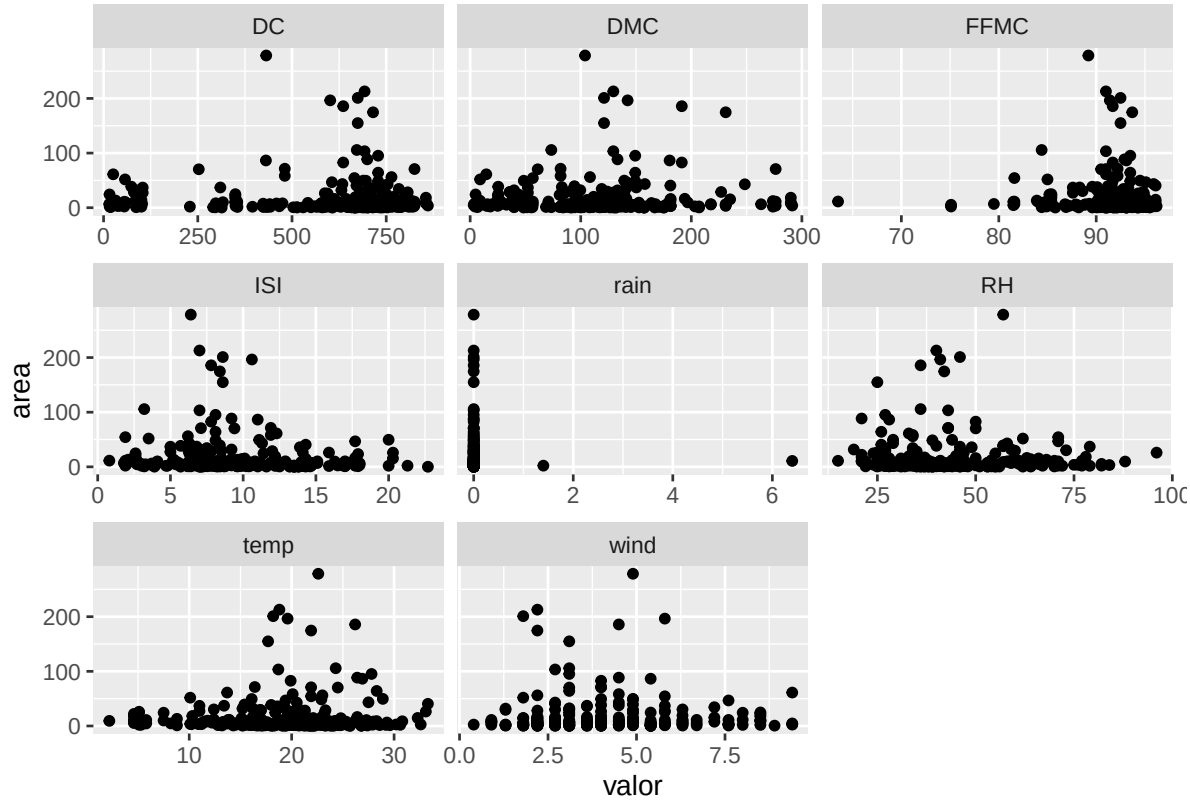
```
base %>%
  ggplot(aes(x=area)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Para isso vamos filtrar, removendo valores zerados, indicando que não houve área prejudicada, e valores extremos

```
pivot %>%
  filter(area >  0,
         area <= 300) %>%
  ggplot(aes(x=valor,y=area)) +
  geom_point() +
  facet_wrap(
    facets = "indicador",
    scales = "free_x",
    nrow = 3,
    ncol = 3)
```



Alguns indicadores que estão relacionados com uma área maior afetada pelos incêndios:

- quando a umidade relativa está baixa
- quando a temperatura está mais alta
- quando o FFMC e DC estão mais altos
- quando não há ocorrência de chuvas
- já os demais indicadores não trasmitem uma relação tão evidente