Proyecto - Curso R

Juan M. Pujol

2023-03-07

Intro

Voy a trabajar con un conjunto de datos que consiste en medidas climáticas en los bosques de dos regiones de Algeria, y si hubo o no incendios forestales en ese día (https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++).

Primero llamo las librerías que voy a usar y leo los datos:

- Agrego la región como un factor
- Paso también de formato día-mes a un solo "día".

```
library("tidyverse")
library("gridExtra")

forest_data_bej <- read_csv("data/Algerian_forest_fires.csv", skip = 1, n_max = 122) %>%
  mutate(region = "Bejaia")%>%
  rowid_to_column("day_abs")
forest_data_sid <- read_csv("data/Algerian_forest_fires.csv", skip = 126)%>%
  mutate(region = "Sidi-Bel Abbes")%>%
  rowid_to_column("day_abs")
forest_data <- rbind(forest_data_bej, forest_data_sid) %>%
  select(-year)
forest_data$FWI <- as.double(forest_data$FWI)
forest_data$month <- as.integer(forest_data$month)
```

Ahora miramos las primeras filas para ver de qué se trata:

```
head(forest_data)
```

```
## # A tibble: 6 x 15
##   day_abs day   month Temperat~1    RH    Ws  Rain  FFMC   DMC    DC   ISI   BUI
##     <int> <chr> <int>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1       1 01        6         29    57    18   0    65.7   3.4   7.6   1.3   3.4
## 2       2 02        6         29    61    13   1.3  64.4   4.1   7.6   1     3.9
## 3       3 03        6         26    82    22  13.1  47.1   2.5   7.1   0.3   2.7
## 4       4 04        6         25    89    13   2.5  28.6   1.3   6.9   0     1.7
## 5       5 05        6         27    77    16   0    64.8   3    14.2   1.2   3.9
## 6       6 06        6         31    67    14   0    82.6   5.8  22.2   3.1   7
## # ... with 3 more variables: FWI <dbl>, Classes <chr>, region <chr>, and
## #   abbreviated variable name 1: Temperature
```

Vemos que, además de fecha, se tiene: temperatura, velocidad de viento (Ws), lluvia, y una serie de índices particulares (FFMC, DMC, DC, ISI, BUI) que forman parte del cálculo del índice general, el Fire Weather Index (FWI). Por último, la variable "Classes" indica si hubo o no incendio, y "region" es una de las dos posibles regiones.

Ahora paso a ver como se caracterizan estas variables si las agrupo, primero, por "Classes", es decir, si hubo o no incendio

```
forest_data %>%
  group_by(Classes) %>%
  summarize_at(vars(Temperature, Rain, FFMC, DMC, DC, ISI, BUI, FWI), list(avg = mean))
```

```
## # A tibble: 2 x 9
##   Classes  Temperature_~1 Rain_~2 FFMC_~3 DMC_avg DC_avg ISI_avg BUI_avg FWI_avg
##   <chr>             <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 fire               33.8  0.0971    87.5   21.0    70.4    7.43   23.9   11.7
## 2 not fire           30.0  1.62      65.3    6.45   21.8    1.28    7.22   0.964
## # ... with abbreviated variable names 1: Temperature_avg, 2: Rain_avg,
## #   3: FFMC_avg
```

Se ve que, como era de esperar, los días **con** incendio tienen valores en promedio mayores de temperatura y FWI, y valores mucho menores de lluvia.

Hago lo mismo, pero separando por región

```
forest_data %>%
  group_by(region) %>%
  summarize_at(vars(Temperature, Rain, FFMC, DMC, DC, ISI, BUI, FWI), list(avg = mean))
```

```
## # A tibble: 2 x 9
##   region         Temper~1 Rain_~2 FFMC_~3 DMC_avg DC_avg ISI_avg BUI_avg FWI_avg
##   <chr>             <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 Bejaia             31.2   0.843    74.7    12.3   53.2    3.66    15.4    5.58
## 2 Sidi-Bel Abbes     33.2   0.679    81.1    17.0   45.4    5.86    17.9    8.52
## # ... with abbreviated variable names 1: Temperature_avg, 2: Rain_avg,
## #   3: FFMC_avg
```

De esto se puede intuir que la región de Sidi-Bel Abbes es más propensa a incendios.

Visto eso, pasamos a los plots

Plots

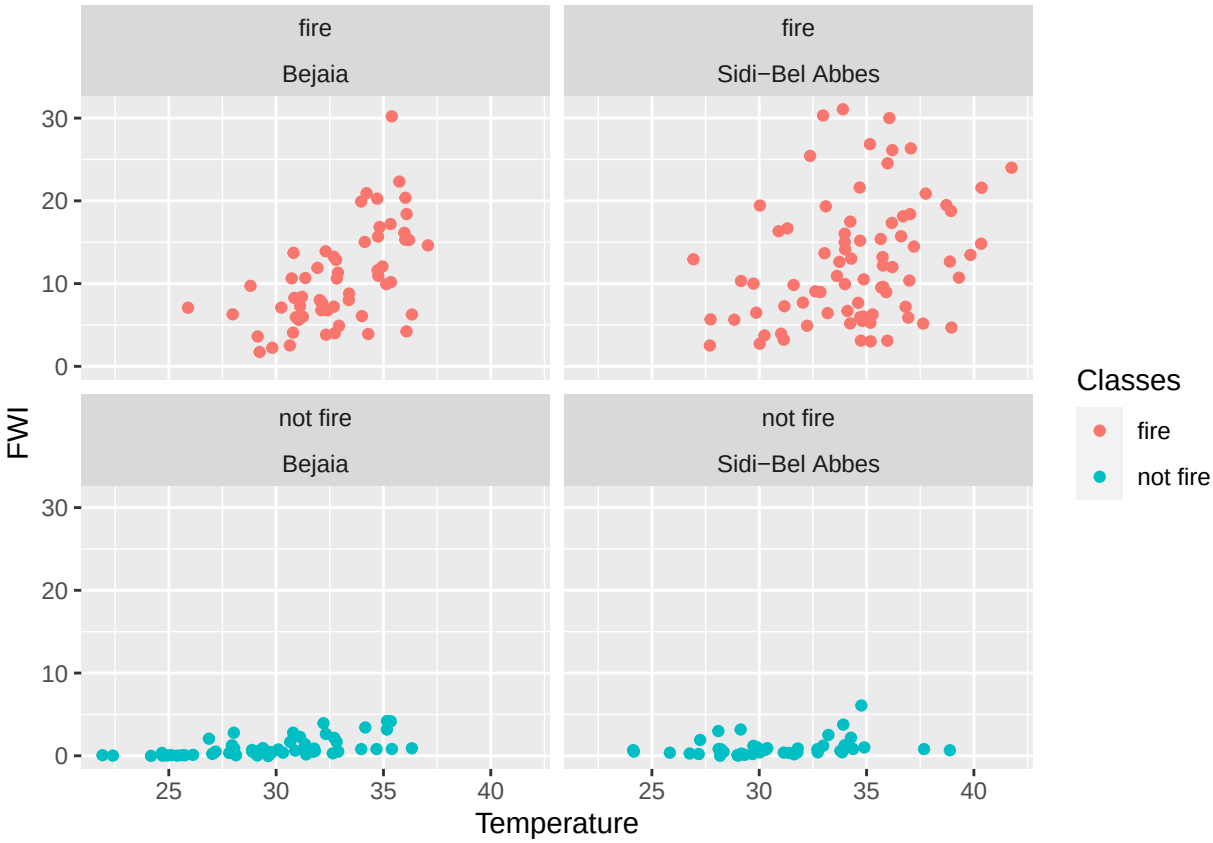
Para empezar, ploteamos las distintas variables numéricas para ver si se cumplen las relaciones esperadas.

```
p <- ggplot(forest_data)

p + aes(x = Temperature, y = FWI, colour = Classes) + geom_jitter()
```

Si bien hay una leve correlación positiva entre Temperatura y FWI, no es tan marcada, ya que la Temperatura es solo una de las muchas variables que determinan el índice FWI.

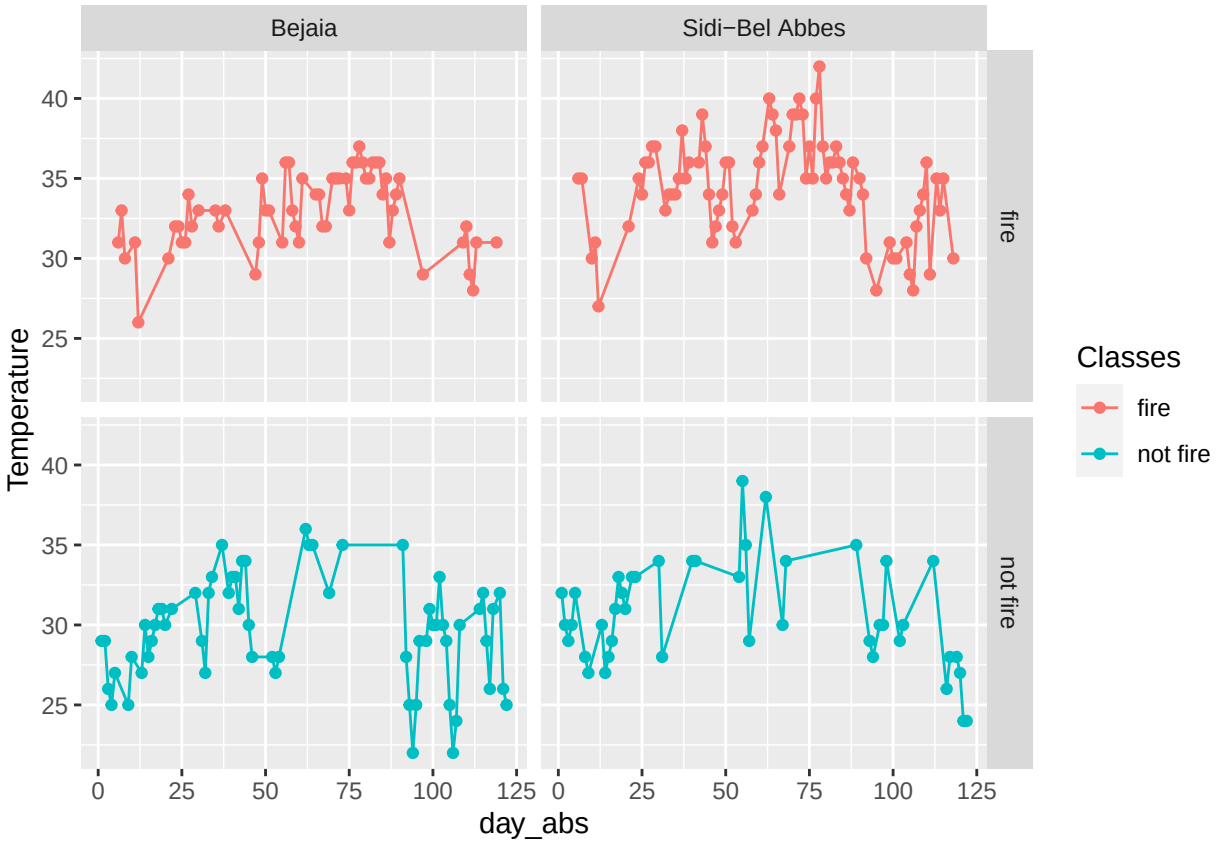
Probamos ahora dividiendo según región y clase:

```
p + aes(x = Temperature, y = FWI, colour = Classes) + geom_jitter() + facet_wrap(Classes~region)
```

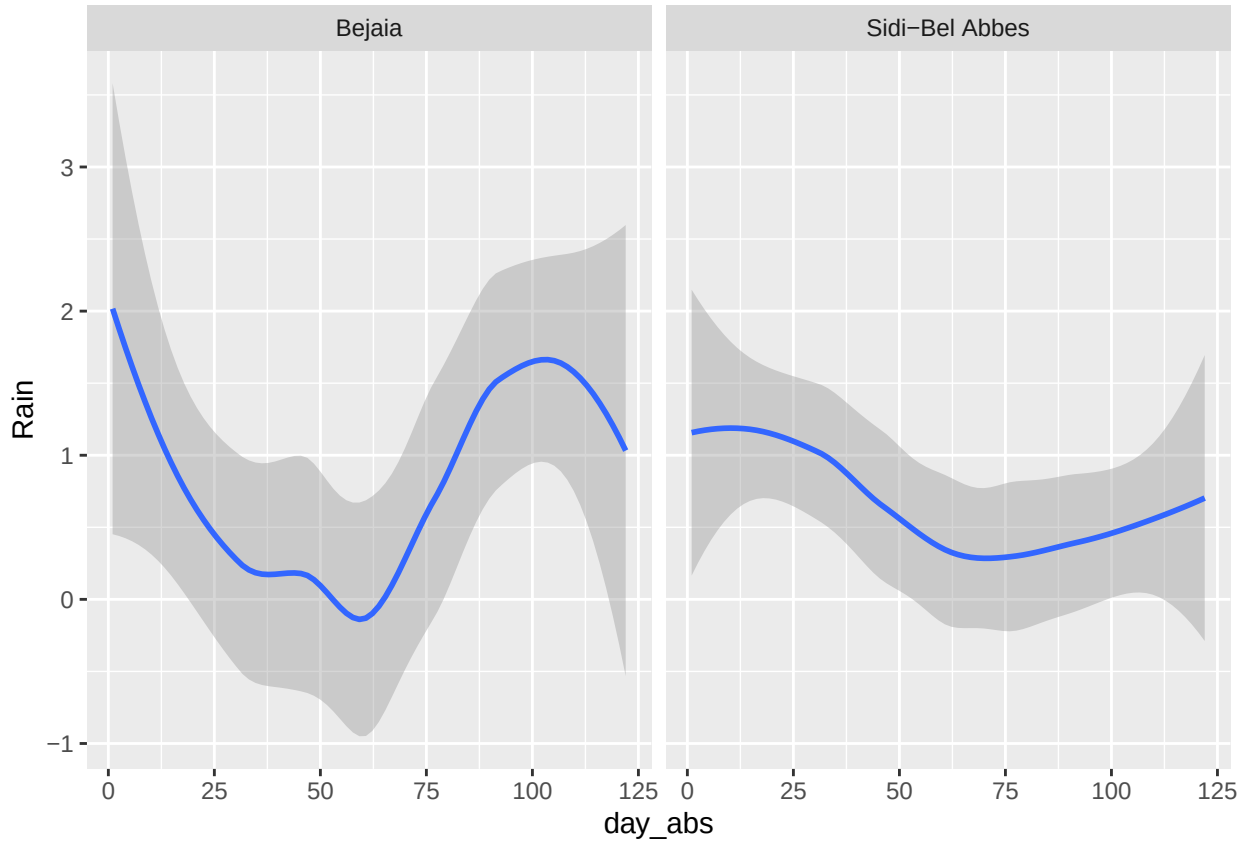


Ahora ploteo algunas variables en función del día con el fin de ver cómo cambian las variables a lo largo del tiempo (recordar que los datos van del 1/6 al 30/9)

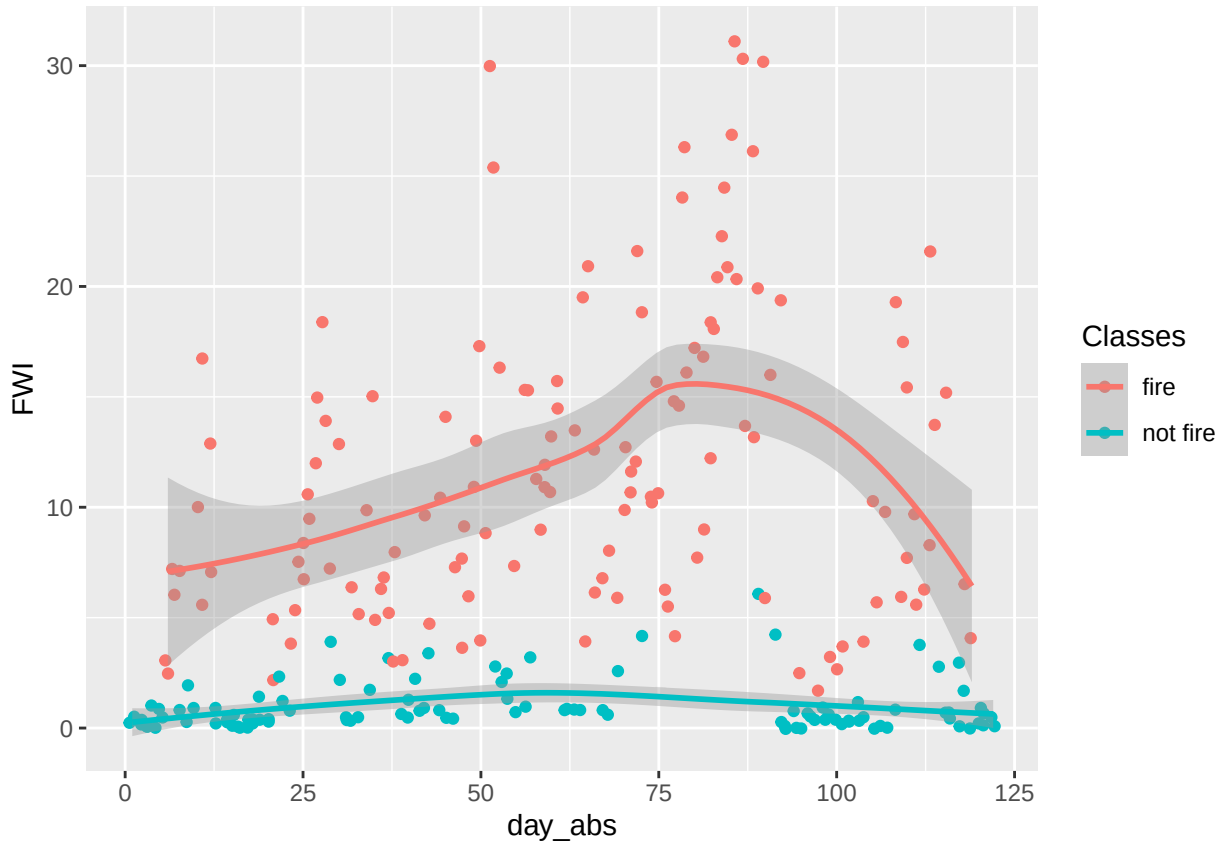
```
p + aes(x = day_abs, y = Temperature, colour = Classes) +geom_line() + geom_point()+ facet_grid(Classes
```




```
p + aes(x = day_abs, y = Rain) + geom_smooth()  + facet_grid(~region)
```




```
p + aes(x = day_abs, y = FWI, colour = Classes) + geom_jitter() + geom_smooth()
```



De acá se ven varias cosas:

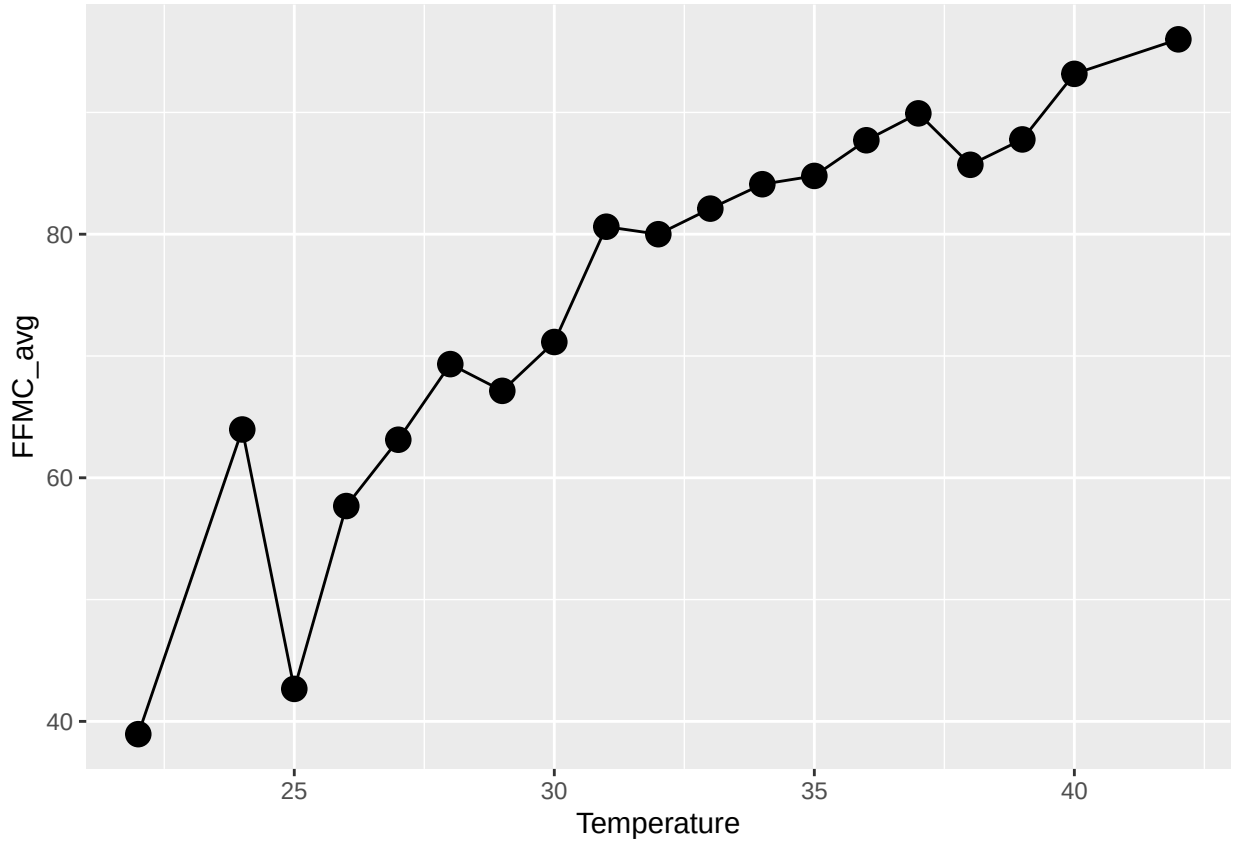
- Las temperaturas crecen hasta alrededor del día 70-80, y luego empiezan a decrecer.
- Las lluvias llegan a un mínimo entre los días 50-70.
- Como es de esperar por estas ocurrencias, el FWI crece hasta alrededor del día 75, donde alcanza un valor máximo, y vuelve a decrecer.

Ahora agrupo por temperatura para ver cosas en función de la temperatura. Para eso primero agrupo:

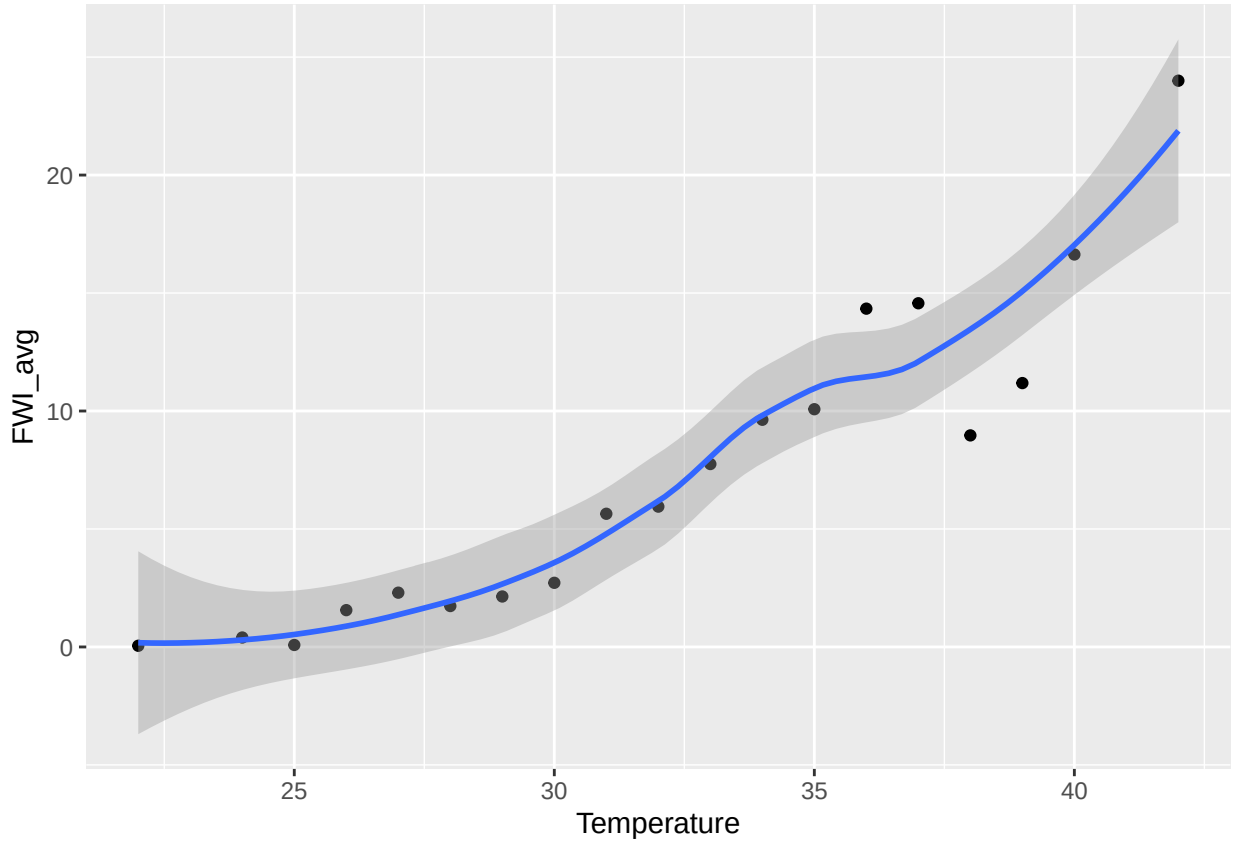
```
data_byT <- forest_data %>% group_by(Temperature) %>% summarize_at(vars(FFMC, DMC, DC, ISI, BUI, FWI),
data_byT_class <- forest_data %>% group_by(Temperature, Classes)  %>% summarize_at(vars(FFMC, DMC, DC,
```

Ahora ploteo para estudiar el comportamiento de algunas variables en función de la temperatura.

```
ggplot(data_byT) + aes(x = Temperature, y = FFMC_avg) + geom_line() + geom_point(size = 4)
```



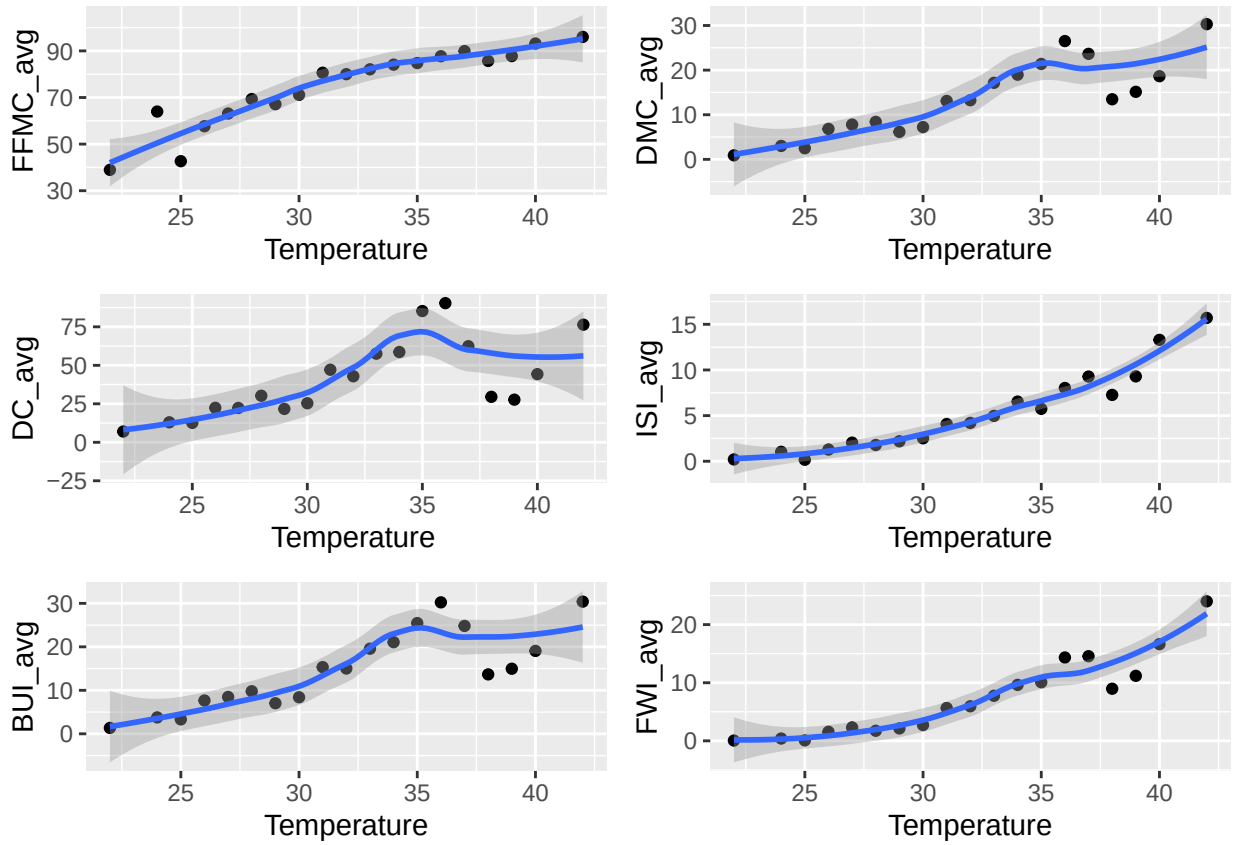

```
ggplot(data_byT) + aes(x = Temperature, y = FWI_avg) + geom_point() + geom_smooth()
```



Se ve que los índices de FFMC y de FWI tienden a aumentar con la temperatura.

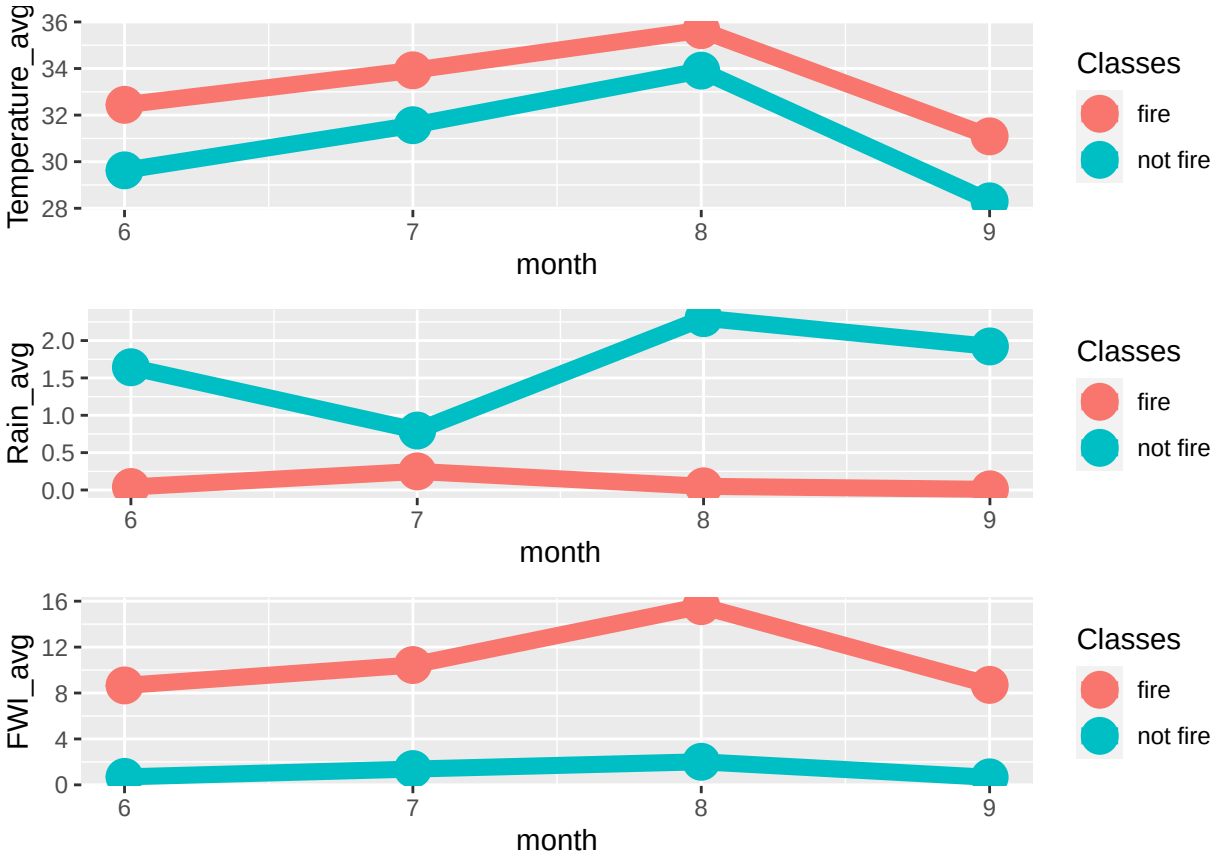
Tomo ahora todos los factores que afectan el FWI y los ploteo en un grid en función de la temperatura:

```
p1 <- ggplot(data_byT) + aes(x = Temperature, y = FFMC_avg) + geom_point() + geom_smooth()
p2 <- ggplot(data_byT) + aes(x = Temperature, y = DMC_avg) + geom_point() + geom_smooth()
p3 <- ggplot(data_byT) + aes(x = Temperature, y = DC_avg) + geom_point() + geom_smooth()
p4 <- ggplot(data_byT) + aes(x = Temperature, y = ISI_avg) + geom_point() + geom_smooth()
p5 <- ggplot(data_byT) +aes(x = Temperature, y = BUI_avg) + geom_point() + geom_smooth()
p6 <- ggplot(data_byT) + aes(x = Temperature, y = FWI_avg) + geom_point() + geom_smooth()

grid.arrange(p1, p2, p3, p4, p5, p6, ncol = 2)
```



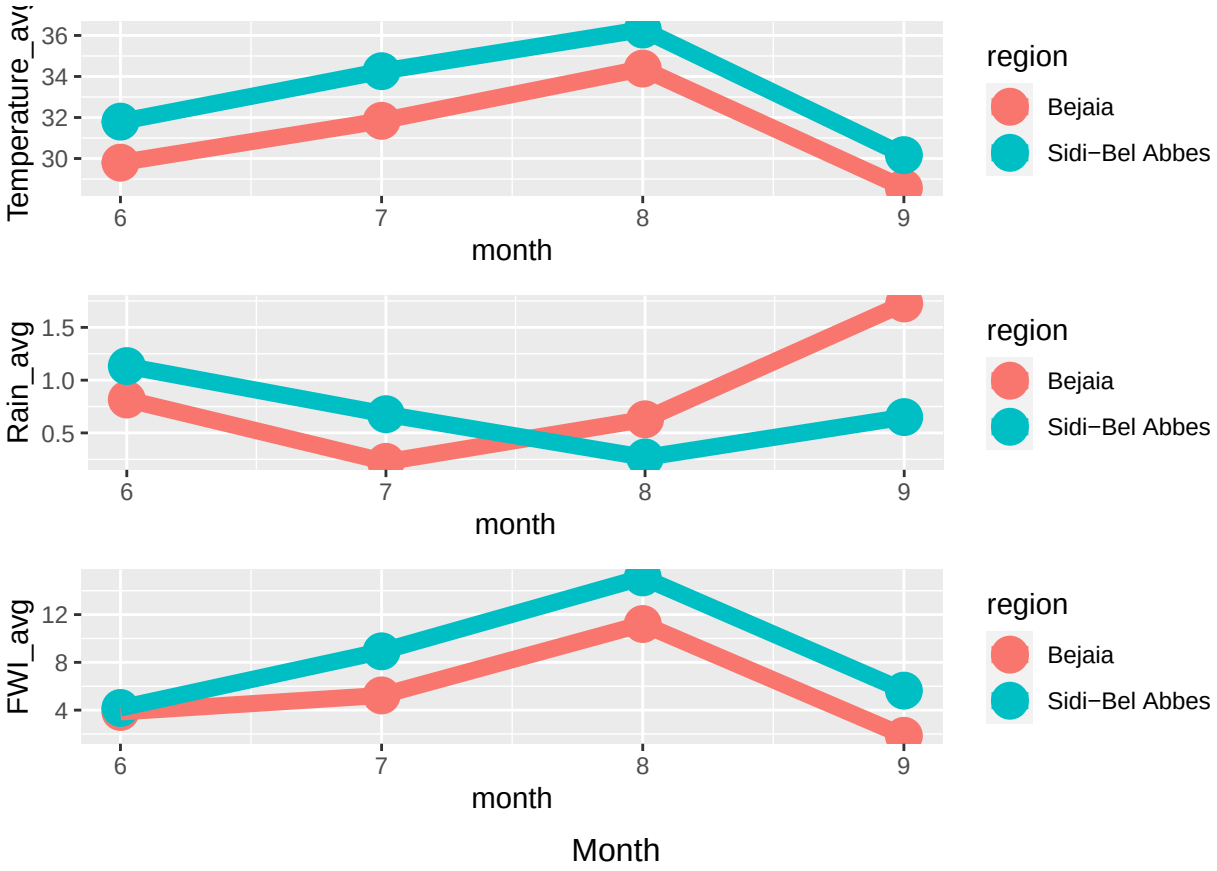
Agrupo por mes y grafico los valores promedios de temperatura, lluvia, y FWI, separado en si hubo o no incendio.

```
a <- forest_data %>% group_by(month, Classes) %>% summarize_at(vars(Temperature, Rain, FFMC, DMC, DC, IS
a1 <- ggplot(a) + aes(month, Temperature_avg, colour = Classes) + geom_point( size = 6)+ geom_line(linew
a2 <- ggplot(a) + aes(month, Rain_avg, colour = Classes) + geom_point( size = 6)+ geom_line(linewidth =
a3 <- ggplot(a) + aes(month, FWI_avg, colour = Classes) + geom_point( size = 6)+ geom_line(linewidth = 3

grid.arrange(a1, a2, a3, ncol = 1)
```



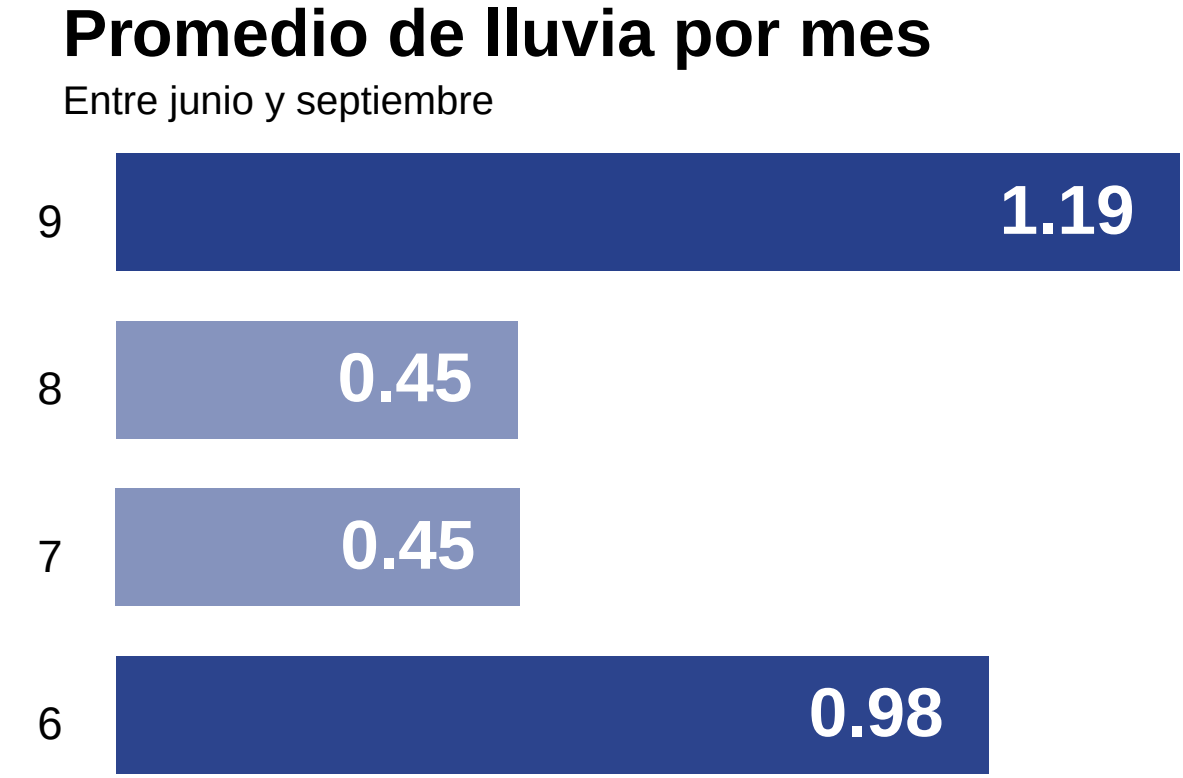
Ahora veo cómo difieren en esos valores ambas regiones_

```
b <- forest_data %>% group_by(month, region) %>% summarize_at(vars(Temperature, Rain, FFMC, DMC, DC, ISI
b1 <- ggplot(b) + aes(month, Temperature_avg, colour = region) + geom_point( size = 6)+ geom_line(linew
b2 <- ggplot(b) + aes(month, Rain_avg, colour = region) + geom_point( size = 6)+ geom_line(linewidth = 3
b3 <- ggplot(b) + aes(month, FWI_avg, colour = region) + geom_point( size = 6)+ geom_line(linewidth = 3)

grid.arrange(b1, b2, b3, ncol = 1, bottom  = "Month")
```

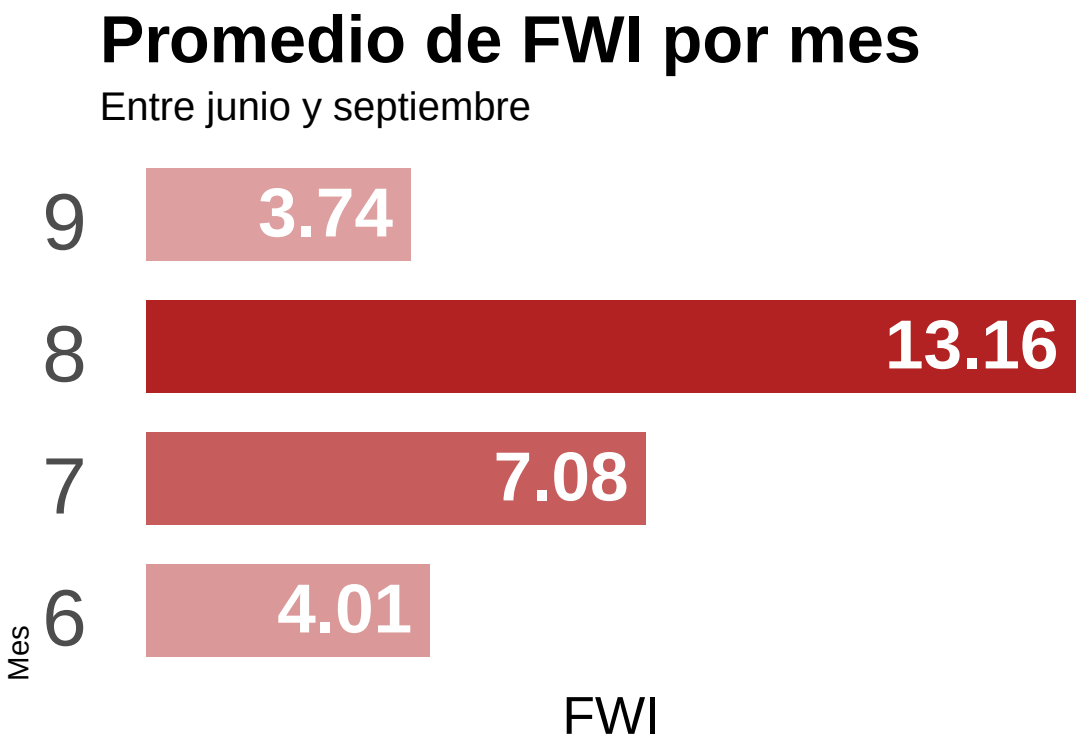


Para concluir, pruebo plots del promedio de lluvias por mes, y promedio de FWI por mes, probando distintos parámetros estéticos.

```
ggplot(forest_data %>% group_by(month) %>% summarise(rain_avg = mean(Rain))) +
  geom_col(aes(month, rain_avg, alpha = rain_avg), fill = "royalblue4",  width = 0.7, just = 0.43, show
  scale_alpha_continuous(range = c(0.2,1), limits = c(0, 1)) +
  geom_text(aes(x = month, y= rain_avg,  label = round(rain_avg, digits = 2)), hjust = 1, nudge_y = -0.
  coord_flip() +
  theme_void() +
  theme(plot.title = element_text(size = 25, hjust = 0, face = "bold"), plot.subtitle = element_text(si
  labs(title = "Promedio de lluvia por mes", subtitle = "Entre junio y septiembre")
```




```
ggplot(forest_data %>% group_by(month) %>% summarise(FWI_avg = mean(FWI))) +
  geom_col(aes(month, FWI_avg, alpha = FWI_avg), fill  = "firebrick", width = 0.7, just = 0.43, show.leg
  scale_alpha_continuous(range = c(0.1,1), limits = c(0, 10)) +
  coord_flip() +
  geom_text(aes(x = month, y= FWI_avg,  label = round(FWI_avg, digits = 2)), hjust = 1, nudge_y = -0.25
  theme(plot.title = element_text(size = 25, hjust = 0, face = "bold"),
        plot.subtitle = element_text(size = 15, hjust = 0),
        axis.text.y = element_text(size = 30, hjust = 1),
        axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        axis.text.x = element_blank(),
        axis.title.x = element_text(size = 20),
        axis.title.y = element_text(hjust=0),
        panel.grid = element_blank(),
        panel.background = element_blank(),
        plot.margin = margin(rep(25, 4))) +
  labs(title = "Promedio de FWI por mes", subtitle = "Entre junio y septiembre", x = "Mes", y = "FWI")
```


Promedio de FWI por mes
Entre junio y septiembre
9
8
7
6
3.74
13.16
7.08
4.01
Mes
FWI