# INIA Uruguay 33

## Taller de Big Data de arroz

Everett Wetchler

BAYES IMPACT

BAYES IMPACT

# Misión: avanzar causas sociales con datos, algoritmos, y software.

- ONG, fundado en 2014 en San Francisco, California
- Ingenieros de sistema, estadísticos, y cientifícos de datos.
- Trabajamos con otros ONGs, gobiernos, y científicos, como consultorio o colegos.
- Nos enfoquecemos en problemes con gran impacto para seres humano.

# Objetivos

- Mostrar metodos y herramientas mas de hacer modelos de machine learning.
- Python vs R
- Metodos de limpieza y exploracion de datos
- Correlaciones basicos
- Hacer un poco de "feature engineering" -- la ingenería de variables

# iPython Notebook

- Editor interactivo de browser, originalmente para ciencia de datos de python
- Incluye codigos y graficos en linea
- Naturalmente documenta su trabajo, como tranformo los datos, cuales codigos hicieron los graficos que estamos viendo, etc.
- Organizacion del projecto, documenta pasos.
- **Se puede usar con R!**

Home | explo | limpia | crear | busca | Base | Inbox (5) | Prese | Work | Goog | Agror | Home | apren

localhost:8889/notebooks/aprender_r.ipynb#

7 Lifetimes | WorkFlowy | Linda Tirado on the r | StoryPilot | d3 Learning | R | Philosophy Bro | RNN | Ag | Other B

jupyter aprender_r (unsaved changes)

File Edit View Insert Cell Kernel Help

Code | Cell Toolbar: None

In [5]:
```
2 + 4
```

Out[5]: 6

In [1]:
```
library(ggplot2)
```

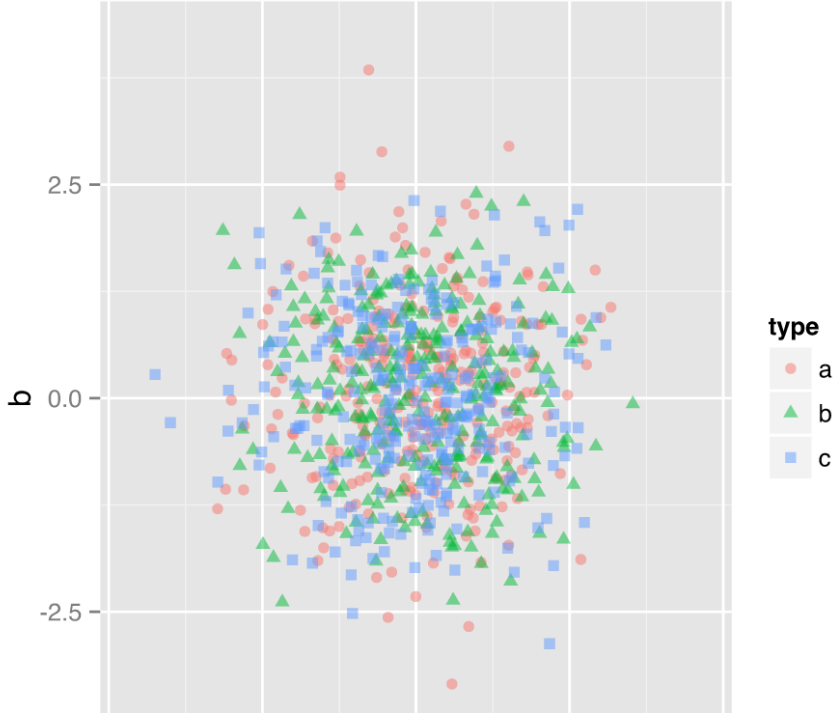
In [4]:
```
points = 300
types = 3
N = points * types
df <- data.frame(type=rep(letters[1:types], each = points),
                 a=rnorm(N), b=rnorm(N))
```

In [3]:
```
xabsmax = max(sapply(df$a, abs))*1.1
yabsmax = max(sapply(df$b, abs))*1.1
ggplot(df) +
aes(x = a, y = b,
    xmin=-xabsmax, xmax=xabsmax,
    ymin=-yabsmax, ymax=yabsmax,
    shape=type,
   color=type) +
geom_point(size=2, alpha=0.5)
```

```
Error in if (args[[1]]$name == "C title" && !is.null(args[[2]])) {: missing value where TRUE/FALSE needed
```

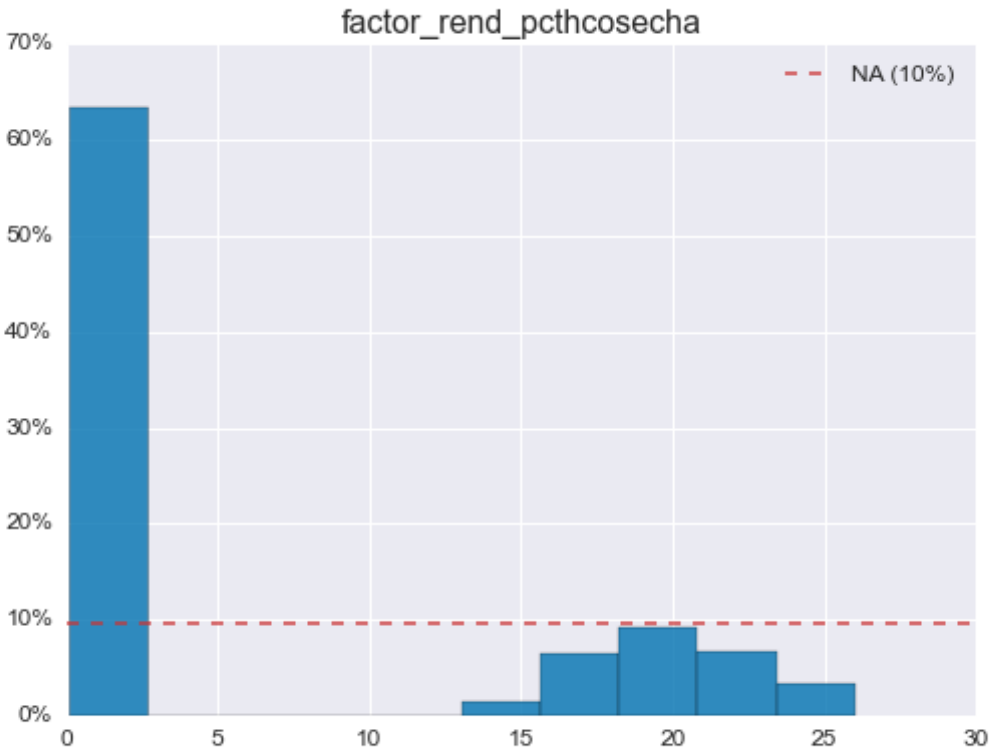
In [7]:
```
ggplot(df) +
aes(x = a, y = b,
    xmin=-xabsmax, xmax=xabsmax, ymin=-yabsmax, ymax=yabsmax, shape=type, color=type) +
geom_point(size=2, alpha=0.5)
```

```
Error in if (args[[1]]$name == "C_title" && !is.null(args[[2]])) {: missing value where TRUE/FALSE needed
```



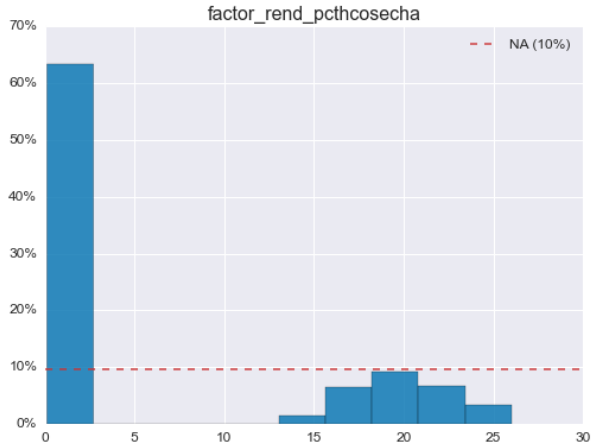
# Limpieza de datos

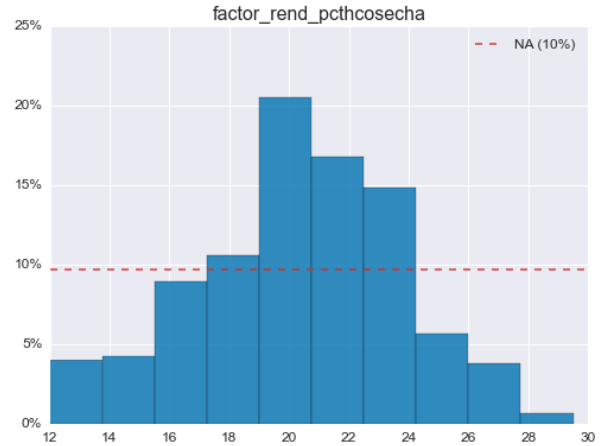
- No cambio los datos originales!
- Ver visualmente los datos en Excel… pero quiero data frame
- Codigos para la limpeza
  - input: datos originales (csv, excel, …)
  - output: csv limpio
- Codigos que automaticamenta hacen histograms y bars para explorar todos los columnos visualmente
- Proceso iterativo: Encontrar problema con datos -> crear codigos para limpiarlos -> averiguar que mejore el problema -> encontrar otro problema….
- Ejemplo vivo


factor_rend_pcthcosecha
NA (10%)
70%
60%
50%
40%
30%
20%
10%
0%
0
5
10
15
20
25
30

| AM | AN | AO |
|---|---|---|
| Fenologia_Emerg-Cosecha_Dias | Factor_Rend_%Hcosecha | Rendimiento_seco_kg/ha |
| 0 | 41 | 0 |
| 136 | 21% | 7877.25 |
| 131 | 23% | 8597.19 |
| 131 | 18% | 8503.07 |
| 138 | 19% | 7815.19 |
| 139 | 19% | 8945.95 |
| 132 | 18% | 8503.07 |
| 148 | 14% | 4953.83 |
| 148 | 14% | 5073.54 |
| 147 | 14% | 4834.69 |
| 120 | 20% | 7063.19 |
| 147 | 14% | 4444.51 |
| 127 | 21% | 6236.61 |
| 127 | 21% | 6241.53 |
| 143 | 1650.0% | 7524.60 |
| 143 | | 7524.60 |
| 143 | 1500.0% | 8178.81 |
| 143 | | 8178.81 |
| 142 | 1550.0% | 7205.30 |
| 141 | 1550.0% | 7205.30 |
| 141 | 1500.0% | 7314.24 |
| 121 | 2050.0% | 8386.82 |
| 121 | | 8386.82 |
| 120 | 2400.0% | 7973.65 |
| 131 | 2580.0% | 10905.34 |

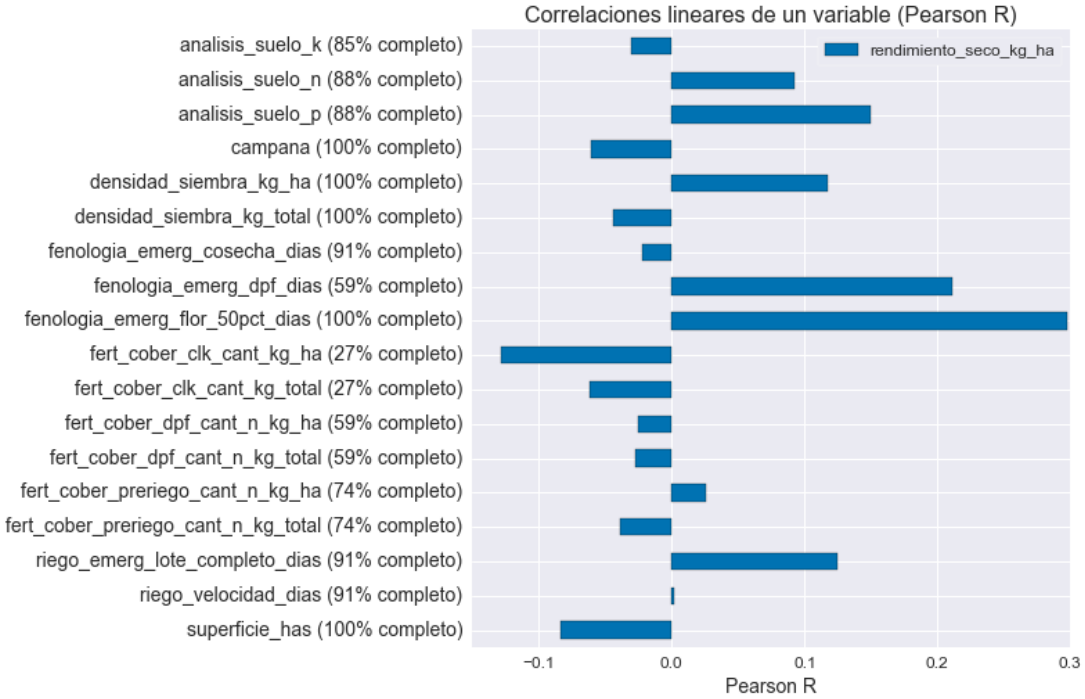
# Limpieza


factor_rend_pcthcosecha
NA (10%)
70%
60%
50%
40%
30%
20%
10%
0%
0
5
10
15
20
25
30

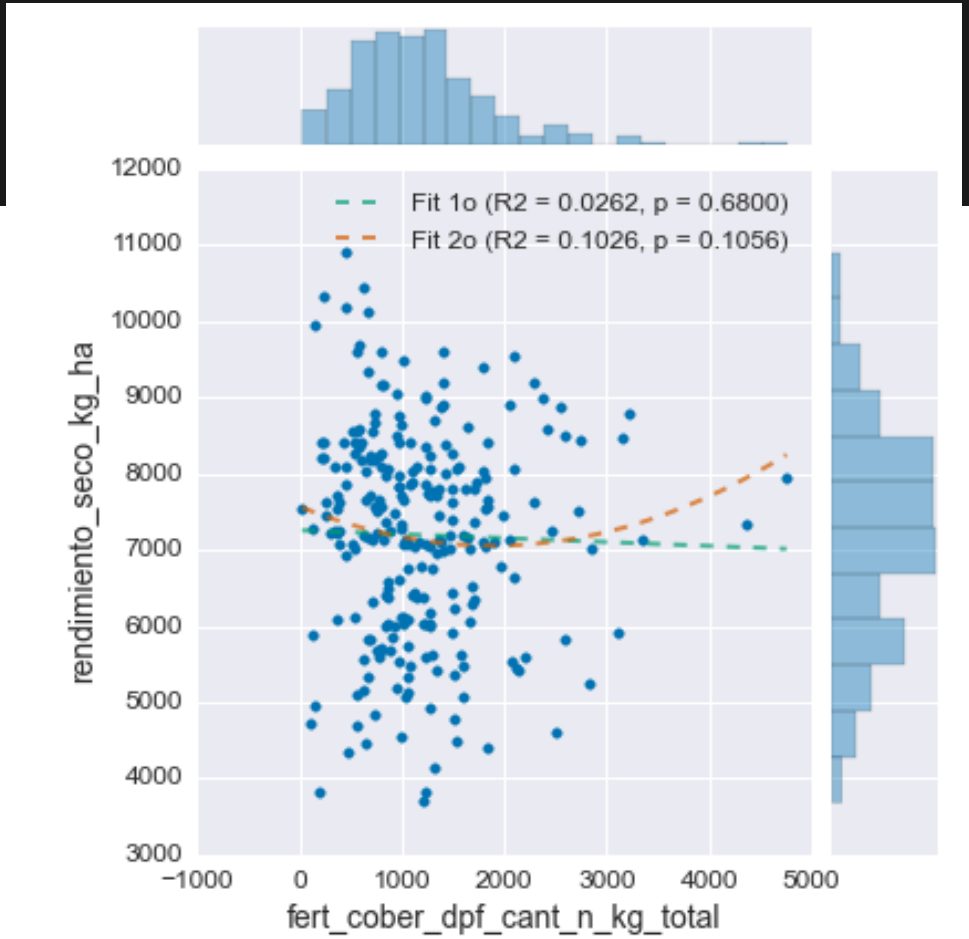

factor_rend_pcthcosecha
NA (10%)
25%
20%
15%
10%
5%
0%
12
14
16
18
20
22
24
26
28
30

# Correlaciones basicos

- Ya estan bastante limpios los datos
- DataFrame.corr() -- todos los correlaciones lineales entre variables (Pearson o qualquiera)
- Bar para visualizarlos
- Mira: siempre incluyo mediciones de datos vacios
- El paquete “seaborn” hace graficos guapos, similar a ggplot. Viene de estudiante PhD de Stanford.
- seaborn.jointplot()
- “numpy” (otro paquete) para hacer lineas y cuervas de fit

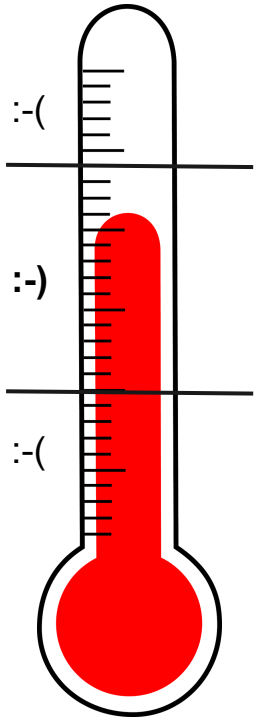

Correlaciones lineares de un variable (Pearson R)
rendimiento_seco_kg_ha
analisis_suelo_k (85% completo)
analisis_suelo_n (88% completo)
analisis_suelo_p (88% completo)
campana (100% completo)
densidad_siembra_kg_ha (100% completo)
densidad_siembra_kg_total (100% completo)
fenologia_emerg_cosecha_dias (91% completo)
fenologia_emerg_dpf_dias (59% completo)
fenologia_emerg_flor_50pct_dias (100% completo)
fert_cober_clk_cant_kg_ha (27% completo)
fert_cober_clk_cant_kg_total (27% completo)
fert_cober_dpf_cant_n_kg_ha (59% completo)
fert_cober_dpf_cant_n_kg_total (59% completo)
fert_cober_preriego_cant_n_kg_ha (74% completo)
fert_cober_preriego_cant_n_kg_total (74% completo)
riego_emerg_lote_completo_dias (91% completo)
riego_velocidad_dias (91% completo)
superficie_has (100% completo)
−0.1
0.0
0.1
0.2
0.3
Pearson R

# Ejemplo


Fit 1o (R2 = 0.0262, p = 0.6800)
Fit 2o (R2 = 0.1026, p = 0.1056)
rendimiento_seco_kg_ha
fert_cober_dpf_cant_n_kg_total
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
−1000
0
1000
2000
3000
4000
5000

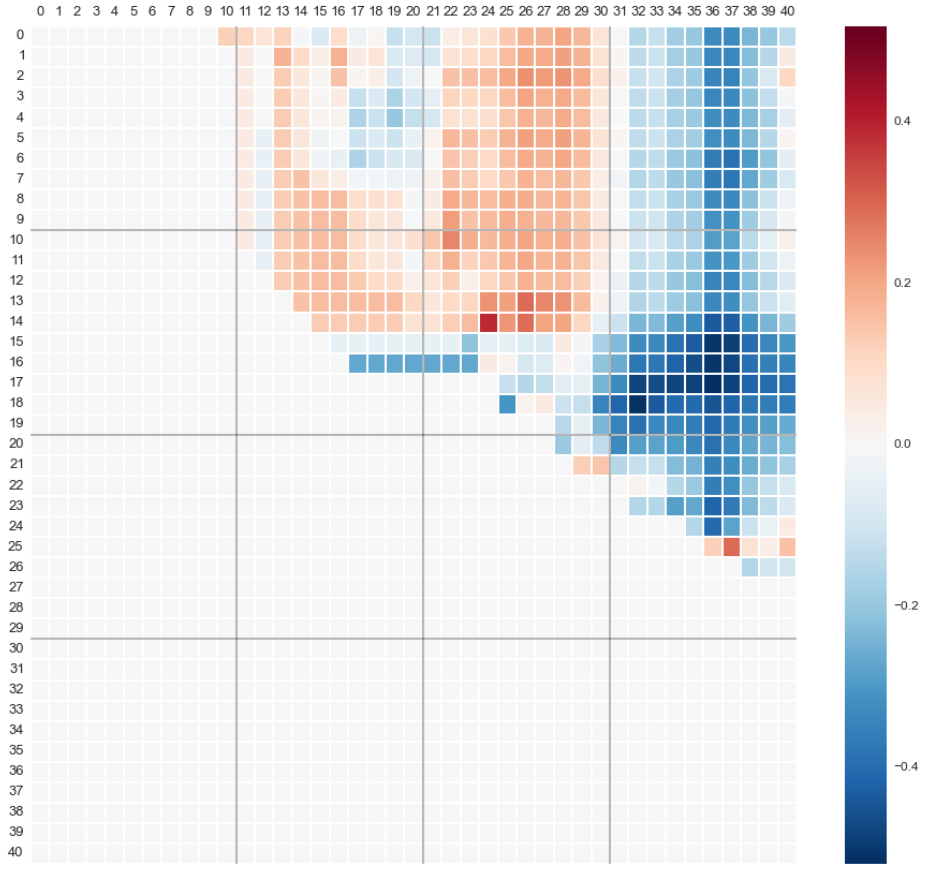
# Ingeneria de variables

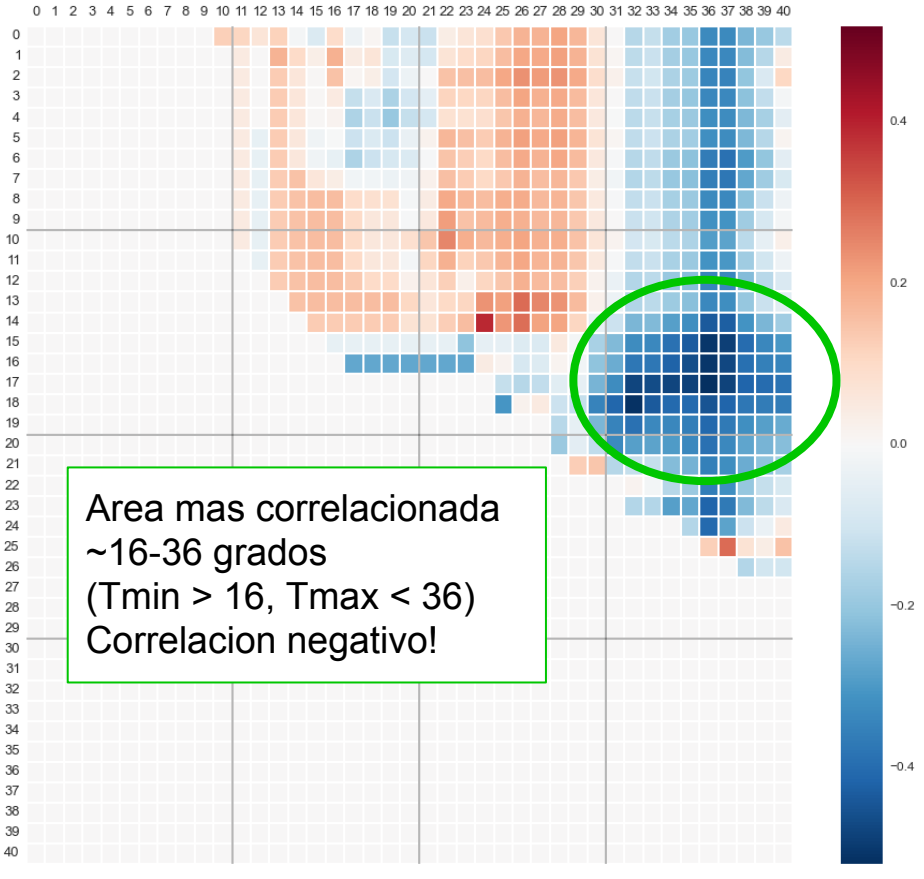
- El mejor variable que ya tenemos explica unos ~30% de la variabilidad.
- No incluye datos de clima.
- Cuales variables climas debemos hacer? “Dias con Tmax < 35”? < 34? Tavg > 20?
- Teoria: hay un alcance critica que predice el exito del arroz. Mas dias durante el cutivo entre este alcance, mejor rienda.
  - 20-30? 20-32? 18-35?
  - Buscamos el mejor alcance



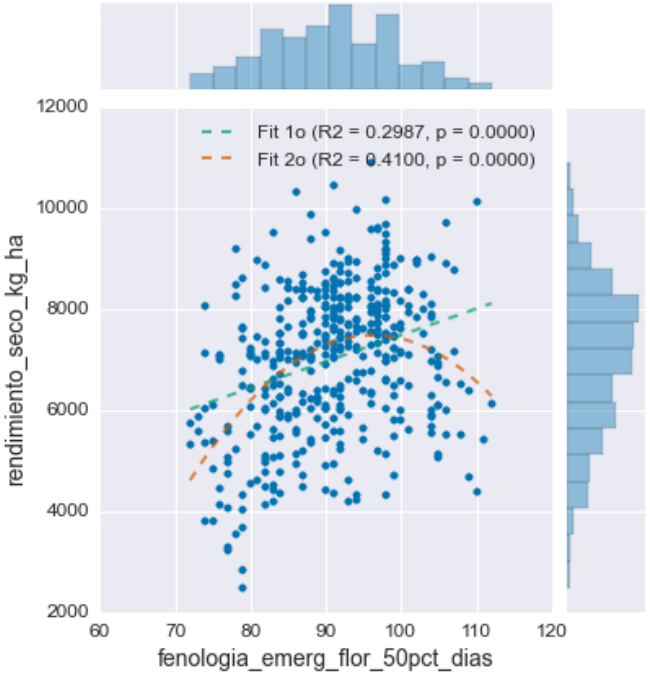
# Solucion

- Probar todos posibilidades ([0,1], [0,2], …,)
- Por cada alcance, calcula, para cada chacra, la porcentaje de los dias cultivas cuando la clima se mantenaba entre ellos todo el dia.
  - e.g. [20,30], % dais con tmax < 30 y tmin > 20
- Medir cuanta variabilidad explicaria un variable asi (R^2)
- Resultado:


0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
0.4
0.2
0.0
−0.2
−0.4


0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
0.4
0.2
0.0
−0.2
−0.4
Area mas correlacionada
~16-36 grados
(Tmin > 16, Tmax < 36)
Correlacion negativo!

# Otra vista




**Mejor explicador**

**<- Antes**

**Despues ->**