

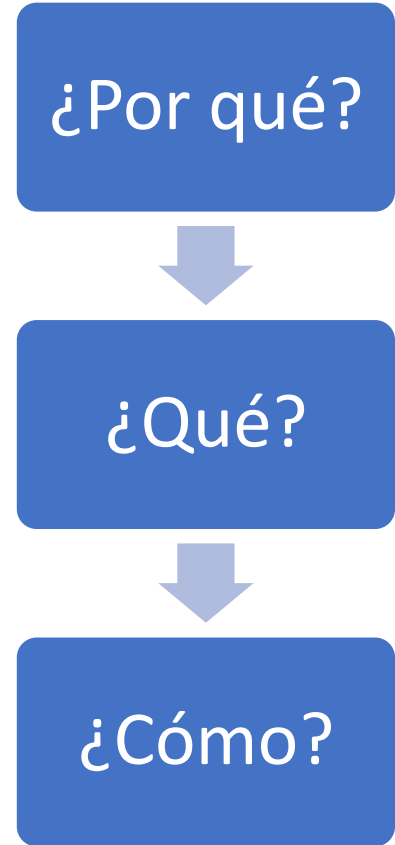
ITERACIÓN DE MODELOS EN R CON TIDYVERSE

JULIO VIALES



Agenda

1. Presentación
2. ¿Por qué iterar sobre modelos?
3. List Column Workflow: los conceptos
4. List Column Workflow: ejemplo con código en R
5. Preguntas y Respuestas
6. Ejemplo de clustering con List Column Workflow



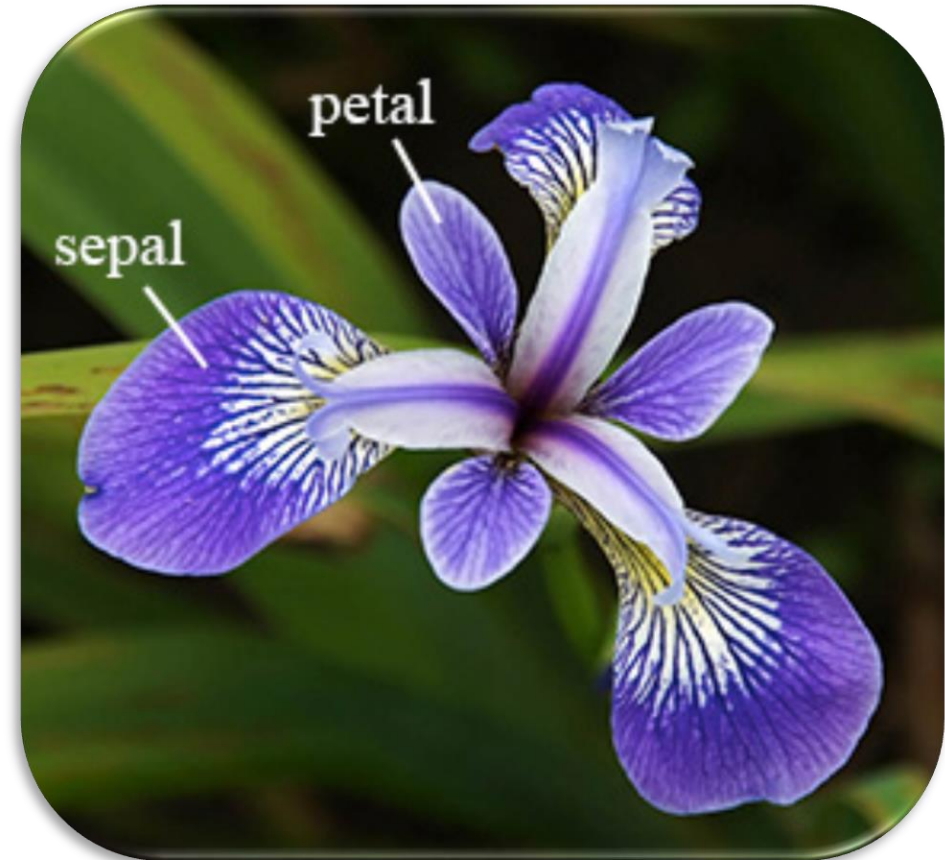
Sobre Julio...

- Licenciado en Ingeniería en Producción Industrial del TEC
- Programador autodidacta
- 4 años en el mundo de los datos
- Rol actual: Customer Behavior Analytics Lead en WU
- Hobbies: paseos en la montaña, lectura, boxeo



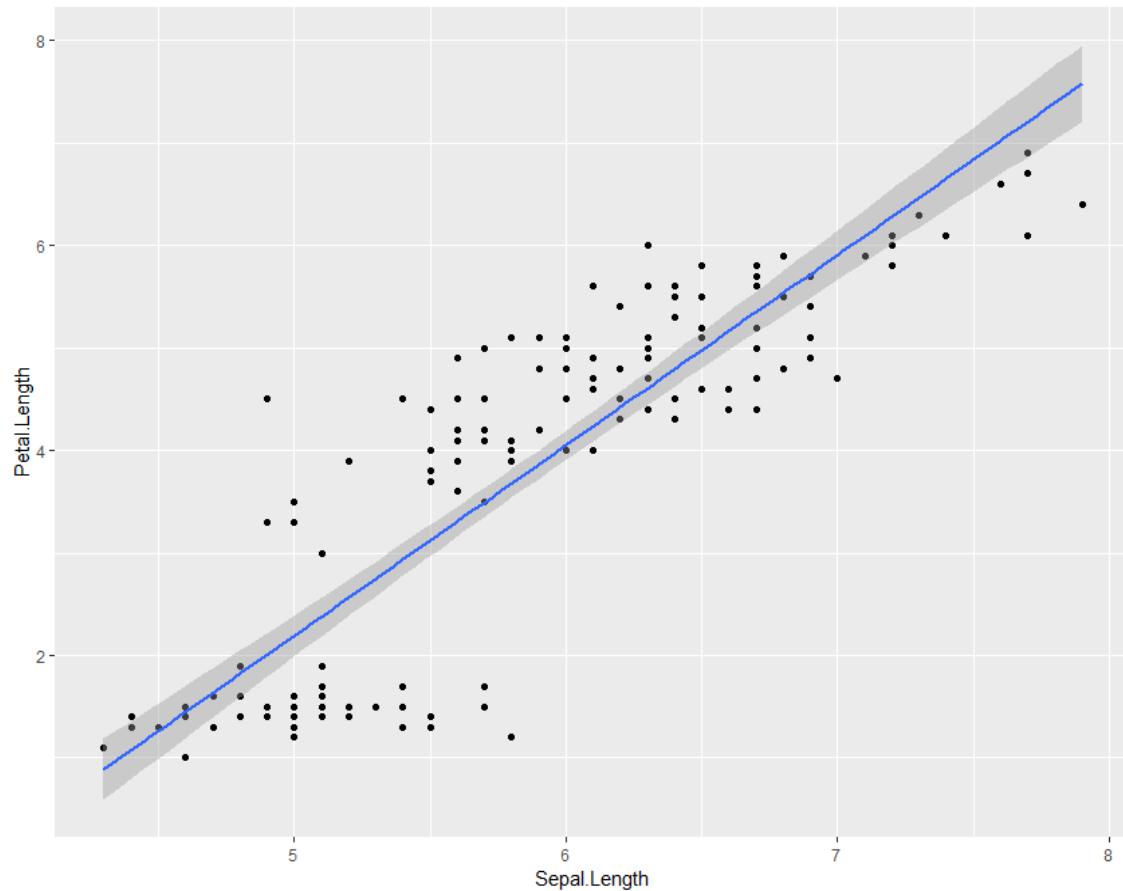
¿En qué casos podemos iterar modelos?

Species	Petal Length	Sepal Length
Setosa	1.4	5.1
Setosa	1.4	4.9
Setosa	1.3	4.7
Versicolor	4.7	7.0
Versicolor	4.5	6.4
Versicolor	4.9	6.9
Virginica	6.0	6.3
Virginica	5.1	5.8
Virginica	5.9	7.1

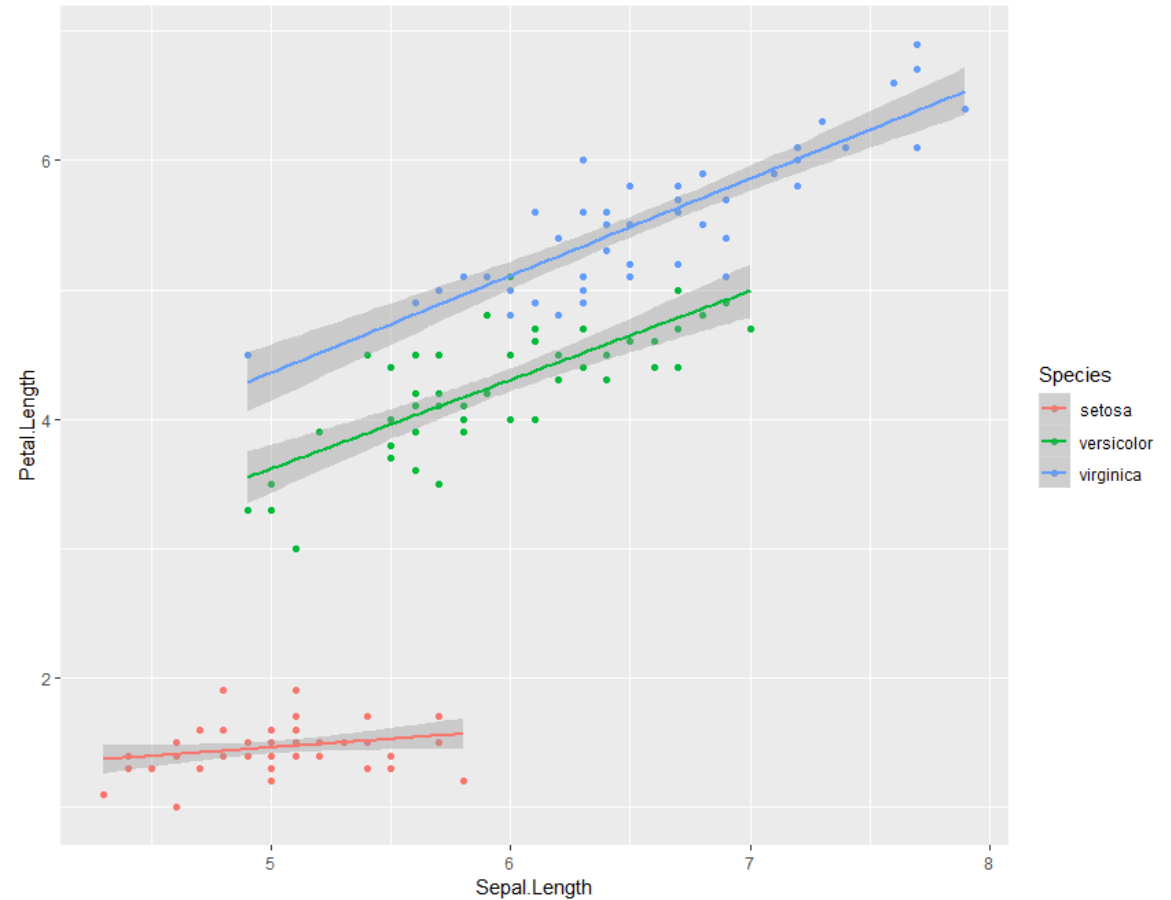


En aquellos casos en que posible dividir nuestros datos con base en una variable categórica (país, género, especie, etc...) y crear un modelo para cada categoría.

¿Por qué iterar modelos?



VS



Diferentes relaciones entre "x" y "y" para cada subconjunto.

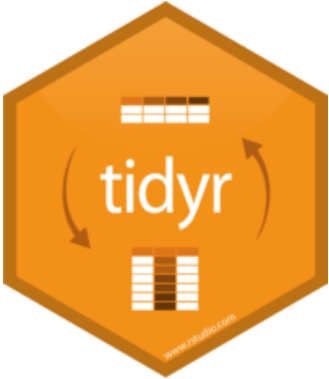
List Column Workflow

Es un flujo de trabajo para iterar modelos sobre subconjuntos de un set de datos utilizando varios paquetes del tidyverse:



Haremos uso de una de las características particulares de los **tibbles**: su capacidad de almacenar elementos complejos en las columnas, en forma de listas.

Etapas del List-Column Workflow



Anidar los datos

`nest()`



Iterar modelo(s)

`map()`



Extraer información
de los modelos

`glance()`
`tidy()`


1) Anidar los datos con nest()

Species	Petal Length	Sepal Length
Setosa	1.4	5.1
Setosa	1.4	4.9
Setosa	1.3	4.7
Versicolor	4.7	7.0
Versicolor	4.5	6.4
Versicolor	4.9	6.9
Virginica	6.0	6.3
Virginica	5.1	5.8
Virginica	5.9	7.1

Species	Data
Setosa	<tibble [3 x 2]>
Versicolor	<tibble [3 x 2]>
Virginica	<tibble [3 x 2]>

```
> iris_nested$data[[1]]
# A tibble: 50 x 4
  Sepal.Length Sepal.width Petal.Length Petal.width
      <dbl>      <dbl>      <dbl>      <dbl>
1         5.1         3.5         1.4         0.2
2         4.9         3         1.4         0.2
3         4.7         3.2         1.3         0.2
4         4.6         3.1         1.5         0.2
5         5         3.6         1.4         0.2
6         5.4         3.9         1.7         0.4
7         4.6         3.4         1.4         0.3
8         5         3.4         1.5         0.2
9         4.4         2.9         1.4         0.2
10        4.9         3.1         1.5         0.1
# ... with 40 more rows
```

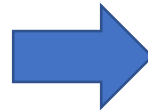

2) Iterar modelos con map()

Species	Data		Species	Data	Model
Setosa	<tibble [3 x 2]>		Setosa	<tibble [3 x 2]>	<S3: lm>
Versicolor	<tibble [3 x 2]>		Versicolor	<tibble [3 x 2]>	<S3: lm>
Virginica	<tibble [3 x 2]>		Virginica	<tibble [3 x 2]>	<S3: lm>

```
> iris_nested$mod[[1]]  
Call:  
lm(formula = Petal.Length ~ Sepal.Length, data = .x)  
Coefficients:  
 (Intercept)  Sepal.Length  
      0.8031      0.1316
```

2) Extraer información de los modelos con glance() y tidy()

Species	Data	Model
Setosa	<tibble [3 x 2]>	<S3: lm>
Versicolor	<tibble [3 x 2]>	<S3: lm>
Virginica	<tibble [3 x 2]>	<S3: lm>



Species	Data	Model	R_sq	Coef
Setosa	<tibble [3 x 2]>	<S3: lm>	0.07	0.132
Versicolor	<tibble [3 x 2]>	<S3: lm>	0.57	0.686
Virginica	<tibble [3 x 2]>	<S3: lm>	0.75	0.750

```
> iris_nested_clean
# A tibble: 3 x 5
  Species    data      mod      r_sq  coef
  <fct>    <list>   <list>   <dbl> <dbl>
1 setosa  <tibble [50 x 4]> <S3: lm> 0.0714 0.132
2 versicolor <tibble [50 x 4]> <S3: lm> 0.569 0.686
3 virginica <tibble [50 x 4]> <S3: lm> 0.747 0.750
>
```

Hora de verlo en acción!

