Análisis descriptivo e Iteración de Modelos en R

Dr. Misael Erikson Maguiña Palma

PREPARA TUS DATOS PARA EL ANÁLISIS

- Limpieza de datos,
- Eliminar datos duplicados,
- Identificación de datos perdidos,
- Imputación de datos perdidos.

ELIGE LA MEJOR HERRAMIENTA

- **GRÁFICOS**
  - *Un gráfico todo en uno*
  - *Gráfico de barras para variables categóricas*
    - Gráfico de barras para 2 variables
  - *Histograma para variables numéricas*
  - *Diagrama de cajas para variables numéricas*
    - Cómo interpretar el Diagrama de cajas
    - Cómo interpretar la simetría de nuestros datos
    - Cómo combinar varias variables en un solo gráfico

ELIGE LA MEJOR HERRAMIENTA

TABLAS DE FRECUENCIA

- *Tablas simples*
- *Tablas de contingencia*

RESUMEN NUMÉRICO

- *Estadísticos de Tendencia central, de Dispersión o variabilidad, de Posición y de forma*
- *Descripción básica*
- *Descripción detallada*
- *Descripción por grupo*

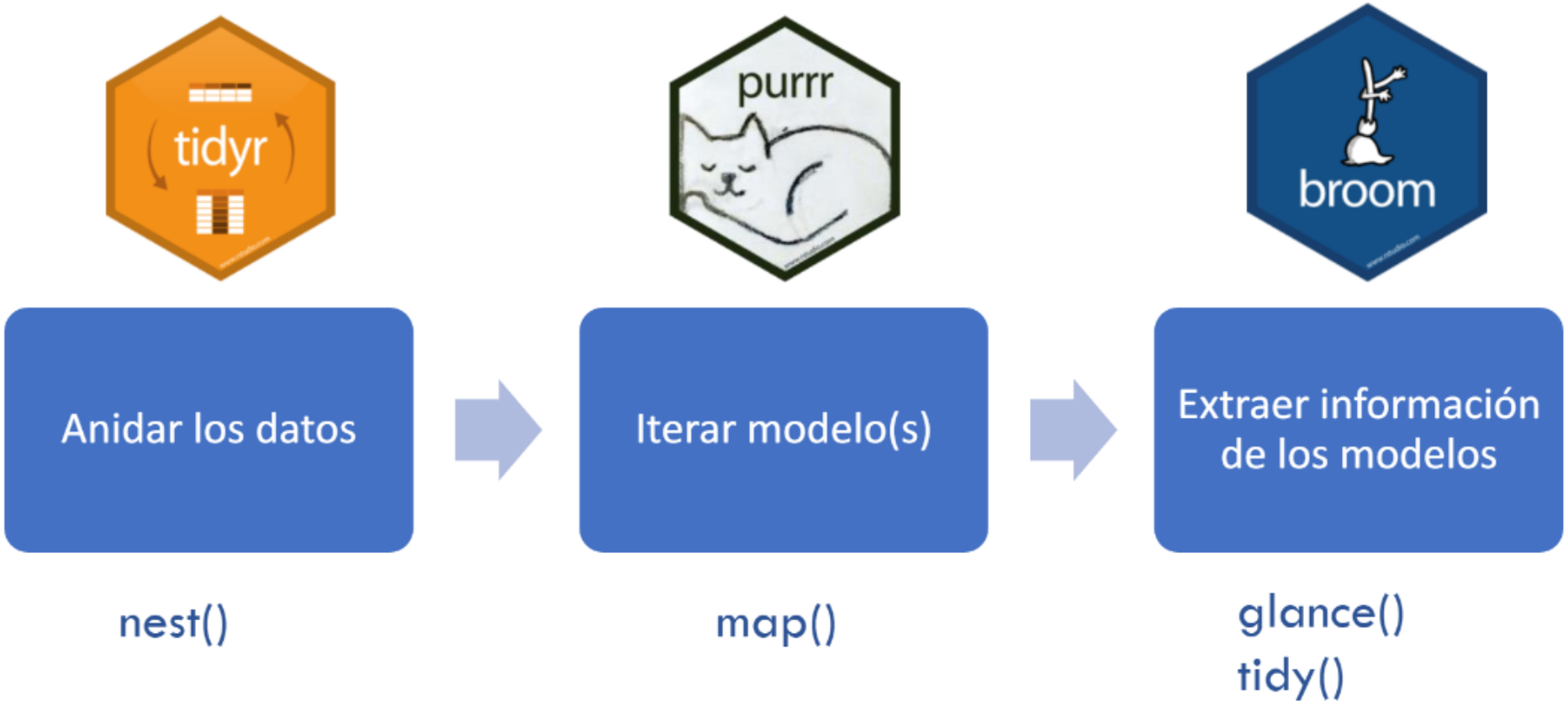
Prestigio de las ocupaciones canadienses

En este caso práctico trabajaremos con el set de datos de **“Prestige”** disponible en la librería **car** del paquete **carData**.

Descripción

- El dataset Prestige esta compuesto por 102 observaciones y 6 columnas. Las observaciones son ocupaciones. Este dataset contiene las siguientes columnas:
- **education:** Educación media de los titulares ocupacionales.
- **income:** Ingreso promedio en dólares.
- **women:** Porcentaje de mujeres por ocupación.
- **Prestige:** Prestigio de la ocupación, resultado de una encuesta social realizada a mediados de la década de 1960.
- **census:** Código ocupacional del censo canadiense.
- **type:** Tipo de ocupación. Un factor con niveles: bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar

Etapas de listas a Columnas



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
```

Iterar modelos con map()

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

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

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
>
```

adj.r.squared: Estadístico R cuadrado ajustado, que es como el estadístico R cuadrado, excepto que se tienen en cuenta los grados de libertad.
AIC: Criterio de información de Akaike para el modelo.
BIC: Criterio de información bayesiano para el modelo.
Deviance: Desviación del modelo.
df.residual: Grados de libertad residuales.
logLik: La probabilidad logarítmica del modelo. [stats :: logLik ()] puede ser una referencia útil.
nobs: Número de observaciones utilizadas.
p.value: Valor p correspondiente al estadístico de prueba.
r.squared: Estadístico R cuadrado, o el porcentaje de variación explicado por el modelo. También conocido como coeficiente de determinación.
sigma: Error estándar estimado de los residuos.
statistic: Estadística de prueba.
df: Los grados de libertad del numerador del estadístico F general. Esto es nuevo en Broom 0.7.0. Anteriormente, esto informaba el rango de la matriz de diseño, que es uno más que los grados de libertad del numerador del estadístico F general.