Sección 3 - YAML y opciones locales*

Lección 2 - Diseño del documento

Adrián Cidre

18 jun 2023

*Este documento ha sido realizado gracias a Quarto

En este documento se explican los principios básicos de YAML

Contenidos

Lista de figuritas

Listado de Tablas

1 Introducción

Quarto es un paquete de software que se incluye en el entorno de desarrollo integrado (IDE) RStudio. Este paquete permite a los usuarios crear informes dinámicos y reproducibles utilizando una variedad de lenguajes de programación, incluyendo R, Python y SQL.

Quarto proporciona herramientas para la creación de informes interactivos que pueden incluir gráficos, tablas, imágenes y otros elementos. Además, los informes pueden ser generados en diferentes formatos, como HTML, PDF y Microsoft Word, lo que los hace muy versátiles y fáciles de compartir con otras personas.

Quarto también incluye características avanzadas, como la capacidad de crear informes que se actualizan automáticamente en función de los cambios en los datos subyacentes y la capacidad de colaborar en equipo en la creación de informes.

En resumen, Quarto es un paquete de software muy útil para cualquier persona que trabaje con datos y necesite crear informes dinámicos, reproducibles e interactivos.

En este documento vamos a ver varias de las opciones más básicas para crear un documento académico.

2 Dataset de Iris

El dataset de iris consiste en una base de datos de 150 observaciones de flores pertenecientes a las tres especies siguientes:

- *Iris setosa*
- *Iris versicolor*
- *Iris virginica*

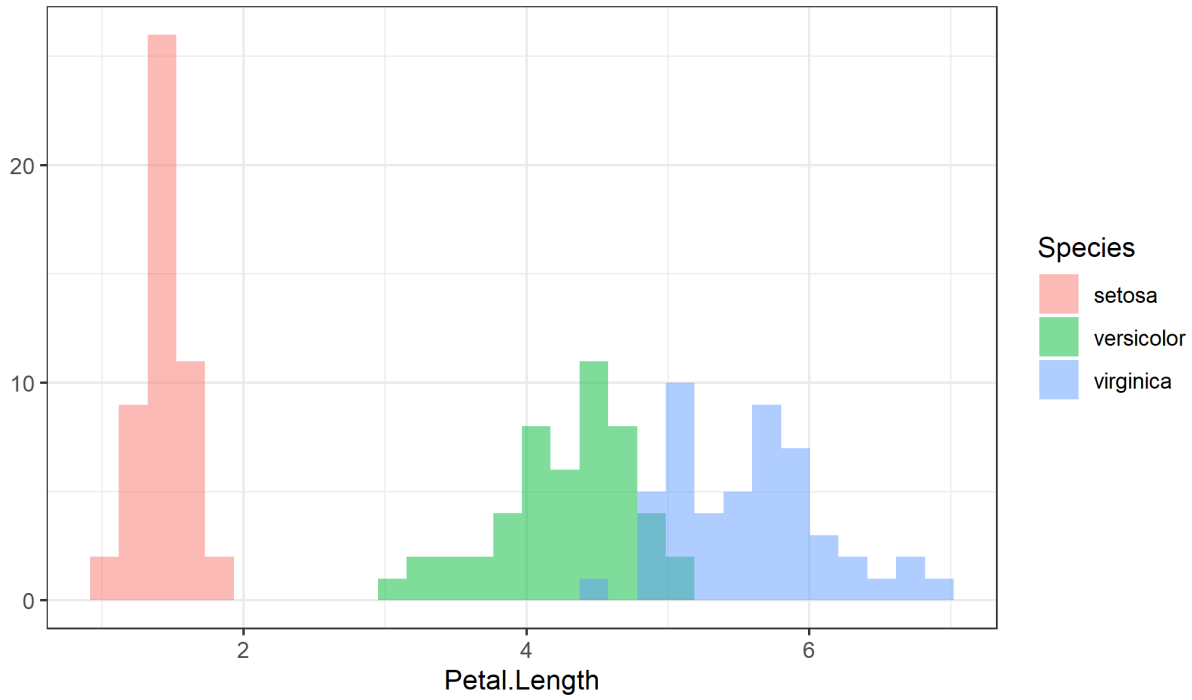
Estas especies pueden visualizarse en la Fig. 2.1.

Para cada una de las observaciones se han tomado 5 variables: longitud del pétalo, longitud del sépalo, anchura del pétalo, anchura del sépalo y especie. En las siguientes secciones se va a analizar esta base de datos.

2.1 Análisis descriptivo

En la Fig. 2.2 podemos ver la distribución de frecuencias de la longitud del pétalo según la especie. En este gráfico vemos que la longitud de los pétalos de *Iris setosa* es mucho menor que

Figurita 2.1: Imágenes de las especies de Iris



Figurita 2.2: Histograma de la longitud del pétalo de 3 especies de Iris

de las otras dos especies, siendo *I. virginica* la que tiene una mayor longitud.

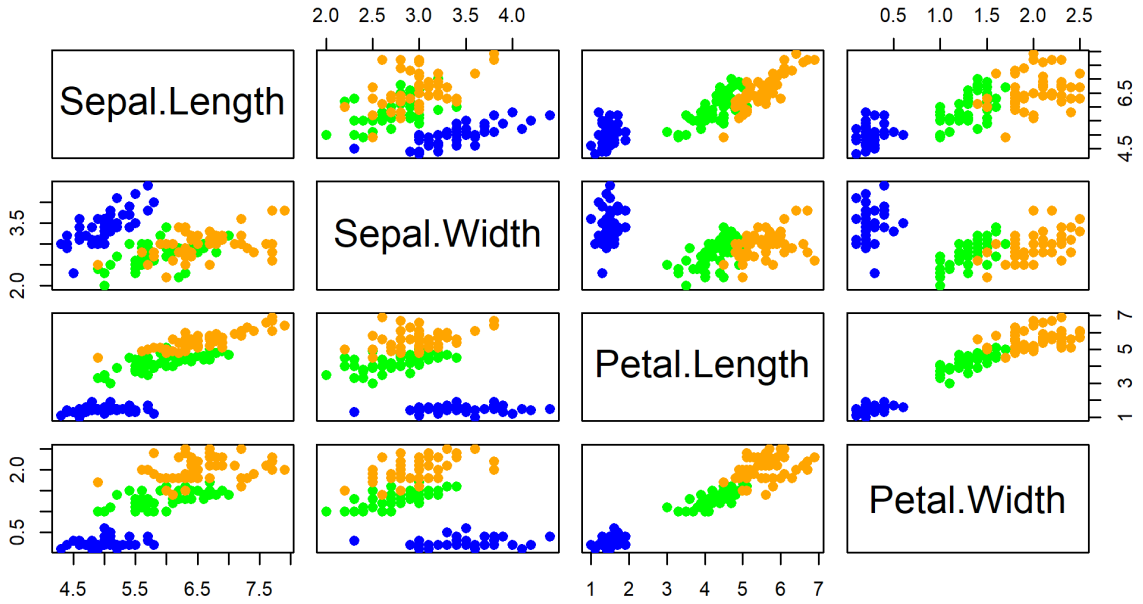
Para evaluar si los cambios eran significativos se utilizó un test ANOVA seguido del *Tukey HSD pos hoc* test utilizando el software R . En la Tablita 2.1 se muestran los resultados, donde podemos ver que las diferencias entre especies son significativas.

Tabla 2.1: ANOVA para la longitud del pétalo de 3 especies de Iris

	diff	lwr	upr	p.adj
versicolor-setosa	2.798	2.59422	3.00178	0
virginica-setosa	4.090	3.88622	4.29378	0
virginica-versicolor	1.292	1.08822	1.49578	0

Para complementar, se muestra la relación entre cada par de variables mediante *scatterplots* en la Fig. 2.3 Todas las variables muestran una buena separación de las tres especies, especialmente la especie *I. setosa*.

Finalmente, en la Fig. 2.4 se muestra la matriz de correlaciones que se ha calculado utilizando el paquete *corrplot* en el software R (Wei y Simko 2021) . En esta matriz podemos ver que tres



Figurita 2.3: Scatter plots de las especies de Iris

de las variables (excepto anchura del sépalo) muestran una alta correlación.

2.2 Modelo de clasificación

Para predecir las especies de *Iris* se propone un modelo en el que solamente se utilizan las medidas de ancho de sépalo y largo de pétalo, ya que como se vio en la Sec. 2.1 parecen ser las variables más importantes para clasificar las tres especies.

Para ello, se ha utilizado un modelo de regresión logística multinomial a través del paquete de R *nnet* (Venables y Ripley 2002) El modelo tiene una exactitud global de . El modelo multinomial ha tenido un error por omisión de *I. versicolor* por *I. virginica* y otro error por omisión de *I. virginica* por *I. versicolor*.

	setosa	versicolor	virginica
setosa	15	0	0
versicolor	0	14	1
virginica	0	0	15

Figurita 2.4: Matriz de correlación del dataset iris

2.3 Modelo de regresión

Finalmente se creó un modelo de regresión para estimar la longitud del pétalo a través de la anchura del sépalo y la anchura del pétalo.

El modelo se entrenó con un 70% de los datos y se evaluó en los datos restantes. Los estadísticos de bondad de ajuste han dado unos excelentes resultados, con un R^2 de y el *rmse* un valor de . La Eq. 2.1 muestra la ecuación de regresión resultante.

$$PetalLength = 1.93 \ + \ 0.26 \ SW \ + \ 2.22 \ PW \tag{2.1}$$

Bibliografía

Venables, W. N. y B. D. Ripley (2002). “Modern Applied Statistics with S”. En: url: https://www.stats.ox.ac.uk/pub/MASS4/.

Wei, Taiyun y Viliam Simko (2021). “R package ’corrplot’: Visualization of a Correlation Matrix”. En: url: https://github.com/taiyun/corrplot.