INICIACIÓN A LA PROGRAMACIÓN ESTADÍSTICA CON R

Actividad integradora

Introducción

La presente actividad de programación en R tiene como objetivo introducir y aplicar diversas técnicas de análisis y visualización de datos utilizando las bibliotecas ggplot2, plotly y dplyr. Estas herramientas son fundamentales en el campo del análisis de datos, ya que permiten manipular, explorar y representar gráficamente conjuntos de datos de manera efectiva.

En la primera parte de la actividad, se utiliza la biblioteca ggplot2 para crear gráficos de dispersión y de caja, permitiendo visualizar la relación entre variables y la distribución de los datos. Además, se explora la creación de gráficos de densidad y se muestra cómo personalizar los gráficos mediante colores y estilos. La segunda parte de la actividad introduce la biblioteca plotly, que ofrece la capacidad de crear gráficos interactivos. Esto facilita la exploración más detallada de los datos, permitiendo zoom, rotación y otras interacciones, lo cual resulta especialmente útil en el análisis exploratorio de conjuntos de datos complejos. La tercera parte de la actividad se centra en el manejo de valores faltantes o NA en R. Se presentan técnicas como la eliminación de filas con NA y la imputación de valores faltantes utilizando el método de promedio múltiple (pmm) de la biblioteca mice. Finalmente, en la cuarta parte de la actividad se utiliza la biblioteca dplyr para realizar operaciones de agrupación y cálculo estadístico. Se calcula la media de una variable después de la imputación de valores faltantes y se muestra cómo realizar un suavizado de datos y realizar agrupaciones por categorías.

Esta actividad es relevante y justificada porque proporciona a los participantes las habilidades necesarias para realizar análisis de datos efectivos utilizando herramientas populares y ampliamente utilizadas en el campo del análisis de datos. Las técnicas y conceptos presentados son fundamentales en la exploración, manipulación y visualización de datos, y pueden ser aplicados en diversos contextos, como el análisis empresarial, científico o de salud.

Al finalizar la actividad, los participantes estarán familiarizados con las bibliotecas ggplot2, plotly y dplyr, y podrán utilizar estas herramientas para realizar análisis de datos y generar visualizaciones informativas y atractivas en R. Esto les

permitirá extraer conocimientos significativos de los conjuntos de datos y respaldar la toma de decisiones basada en datos en diversas áreas profesionales.

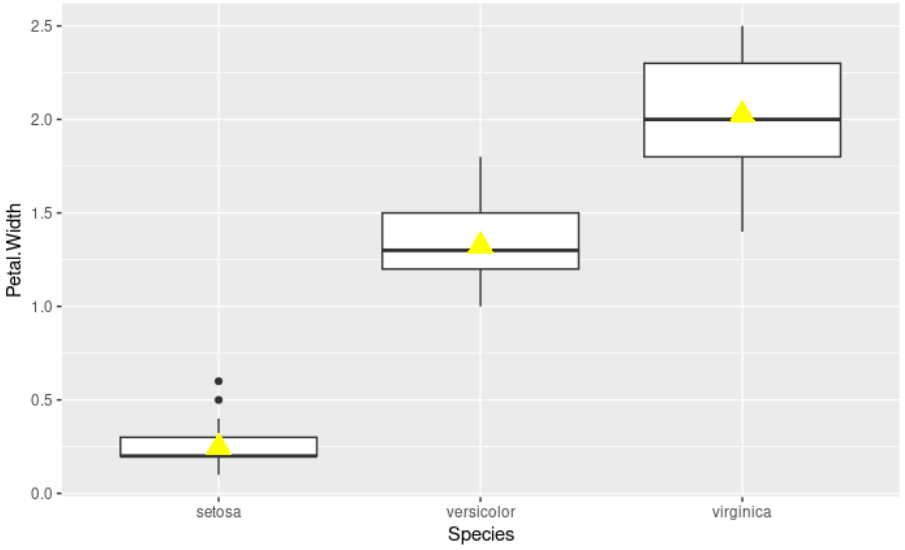
Desarrollo

Parte 1: Carga y exploración de datos

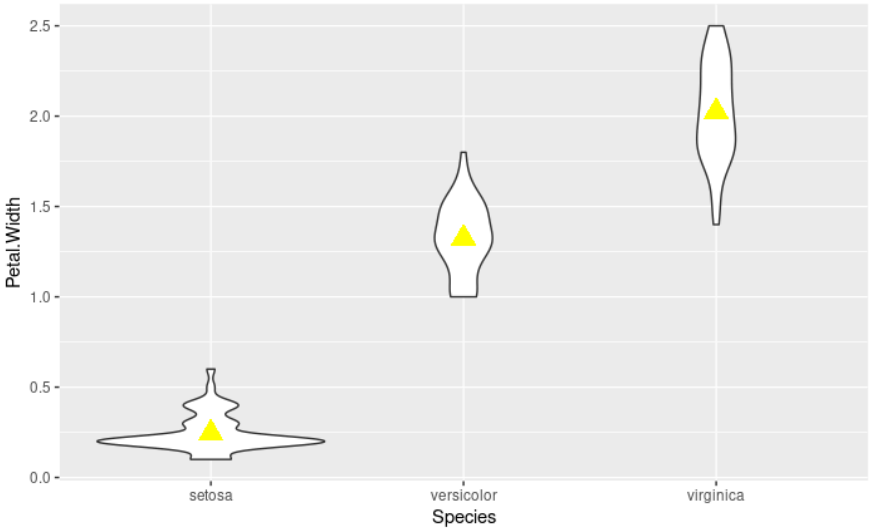
1. **Carga el conjunto de datos iris en un dataframe llamado iris.**
2. **Explora la estructura y un resumen estadístico del dataframe con funciones como str() y summary().**

Parte 2: Visualización con ggplot2

3. **Utiliza ggplot2 para crear un gráfico de caja que muestre la distribución de Petal.Width por cada categoría de Species, pero usa variantes de los argumentos situados dentro de stat_summary() para que el color del punto que señala el promedio sea amarillo, sea de tamaño 5 y tenga un shape personalizado por el usuario.**

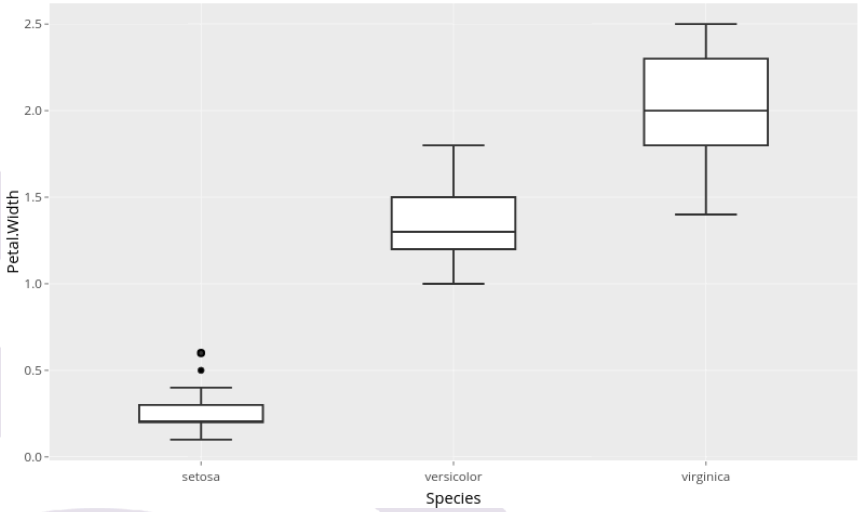


4. **Ahora genera un gráfico de violín similar, que situé el promedio con un punto también amarillo.**



Parte 3: Gráficos interactivos con plotly

5. Utiliza plotly para crear un gráfico de caja que al igual que el punto 3, muestre la distribución de Petal.Width por cada categoría de Species. Debes aprender a generar el mismo tipo de gráfico con ambas librerías.



Parte 4: Limpieza y transformación de datos

5. Elimina las filas con valores faltantes en tu dataframe ya sea a través de la eliminación de filas con NA y la imputación de valores faltantes utilizando el método de promedio múltiple (pmm) de la biblioteca mice, o con la función complete.cases (queda a tu libre elección usar cualquiera de los dos métodos). Para poder ahondar mas al respecto, por favor consulta la página 8 del Libro presente en el módulo de Visualización de datos

Parte 5: Análisis estadístico y síntesis con dplyr

6. **Agrupa los datos por Species utilizando dplyr y calcula el rango intercuartílico (IQR) de Sepal.Width.**
7. **Crea un gráfico de barras apiladas ya sea en ggplot2 o plotly que muestre la proporción de observaciones por cada categoría presente en Species.**

El código RMarkdown se puede encontrar en el siguiente enlace: notebook.