Introducción a R

Email: [delval@decsai.ugr.es](mailto:delval@decsai.ugr.es)

1. Exploratory Data Analysis
   1. *Ejemplo 1,* USAirlines *dataset*

\* Descargate el dataset USAirlines desde PRADO

\* Una vez descargado comprueba la dimensión y los nombres de las columnas del dataset. ¿Qué dimensión tiene? ¿qué tipo de datos alberga? (e.g. 4 variables numéricas continuas)

\* Calcula las tendencias centrales de todos los datos del dataset (mean, media) Tip: puedes usar la función apply()o tidyverse

\* Haz lo mismo para las medidas de dispersión mínimo y máximo. ¿Seria posible hacerlo con un único comando?. ¿Que hace la función range()?

\* Sin embargo las medidas mas populares de dispersión son la varianza (var()), su desviación standard (sd()) y la desviación absoluta de la mediana mad(). Calcula estas medidas para los valores de la variable RA

\* Imagina que quieres calcular dos de estos valores de una sola vez. ¿Te serviría este código?

f = function(x) c(median(x), mad(x))

f(USAirlines [,3])

\* Implementa estos cálculos para cost usando dplyr de tidyverse

\* ¿Cuál sería el resultado de aplicar apply(USAirlines,2,f)?

\* Vamos a medir la dispersión de la muestra utilizando el concepto de cuartiles. El percentil 90 es aquel dato que excede en un 10% a todos los demás datos. El cuartil (quantile) es el mismo concento, solo que habla de proporciones en vez de porcentajes. De forma que el percentil 90 es lo mismo que el cuartil 0.90. La mediana “median” de un dataset es el valor más central, en otras palabras exactamente la mitad del dataset excede la media. Calcula el cuartil .10 y .50 para la columna cost del dataset hip. Sugerencia: quantile()

\* Los cuantiles 0.25 y 0.75 se conocen como el “first quartile” o Q1 y el “third quartile” o Q3, respectivamente. Calcula los cuatro cuartiles para cost con un único comando.

\* Otra medida de dispersion es la diferencia entre el primer y el tercer cuartil conocida como rango intercuartil (IQR) Inter Quantile Range. ¿Obtienes ese valor con la función summary()?

\*Crea un scatterplot que te compare los valores de cost y output. Representa los puntos con el símbolo ‘.’ Y que estos puntos sean de color rojo si year el igual o mayor de 1980. Sugerencia: puedes usar dplyr/tidyverse o Rbase ifelse()

¿Ves algún patrón?

\* En vez de crear los plots por separado para cada par de columnas, hazlos con un solo comando con el scatterplot matrix

*Ejemplo , InsectSprays dataset*

El dataset “InsectSprays” está incluido en la libreria de R “datasets”. Contiene el conteo de insectos extraidos de diferentes zonas agrícolas tratadas de forma experimental con diferentes insecticidas. Haz un boxplot para determinar que insecticida parece ser el más efectivo.

*Ejemplo , Carseats*

Instala la library(ISLR), vamos a trabajar con el dataset Carseats. Si vas a usar dplyr puedes inspeccionar el paquete “dlookr”

1. Encuentra que variables tienen skewness
2. Genera dos listas, una de variables con skewness a la derecha y otra con skewness a la izquierda
3. Averigua que variables no están distribuidas de forma normal, crea gráficos que lo prueben
4. Encuentra que variables están correlaccionadas positiva y cuales negativamente. Crea el gráfico apropiado. ESTE NO

*Ejemplo , iris dataset*

\* Vamos a utilizar el ejemplo del dataset iris que está incluido en la distribución de R. Este dataset fue creado por Douglas Fisher. Consta de tres clases y tipos de 3 clases de tipos de flores:

* \_setosa\_
* \_virginica\_
* \_versicolor\_

Cada una de ellas con cuatro atributos:

* sepal width
* sepal length
* petal width
* petal length

\* Inspecciona las primeras filas del dataset y calcula el summary() del mismo con cada atributo del dataset

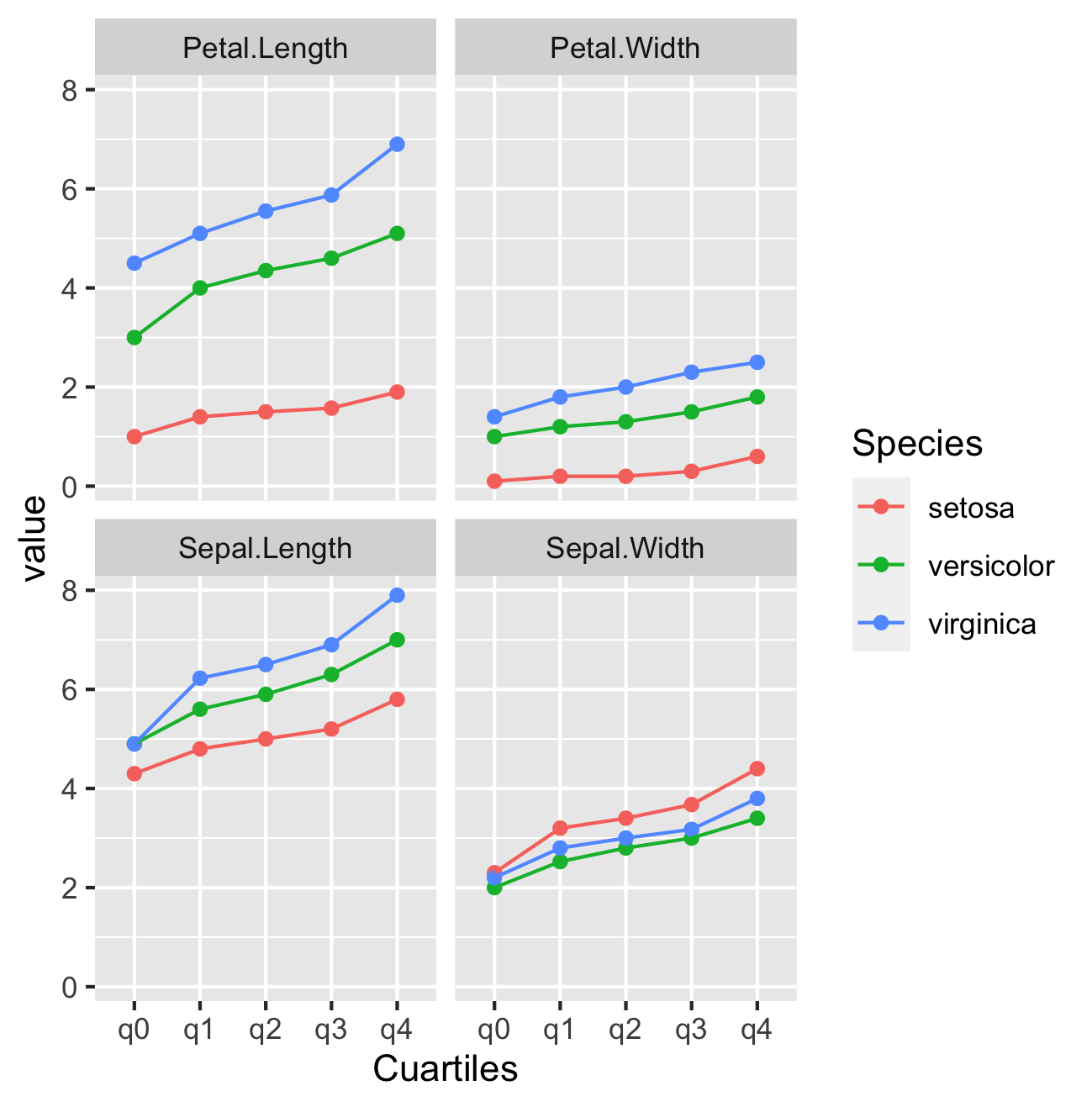
\* Crea un histograma de petal.width , teniendo en cuenta que el numero de bins es variable fija este a 9. Añádele color y nombres al eje x "Petal Width"y al gráfico dale el nombre de "Histogram of Petal Width". Crea un histograma para cada variable

\*Crea los cuartiles del dataset

\* Representa en un boxplot la variable de ancho de hoja dependiendo del tipo de hoja que tengan. Esta información está en la variable especie

\* (NO hacer a no ser que os queráis testear)

Crea un gráfico donde para cada variable numérica del dataset puedas ver los cuartiles por cada tipo de iris. Represéntalos en un plot de líneas donde cada especie tenga un color diferente. En este caso la gráfica tiene que tener esta forma:



Los pasos a seguir son los siguientes:

1. crea un subgrupo de datos para cada especie
2. calcula los cuartiles con la función quantile() de cada variable por especies
3. comprueba que tengas una columna con el nombre de la especie y otra con el nombre del cuartil
4. Uno los tres datasets, uno por especie
5. Intenta crear una tabla con esta forma. Puedes usar la función pivot\_longer() de tidyr. La explicaremos el tercer dia

Species Cuartiles variable value

<chr> <chr> <chr> <dbl>

1 setosa q0 Sepal.Length 4.3

2 setosa q0 Sepal.Width 2.3

3 setosa q0 Petal.Length 1

4 setosa q0 Petal.Width 0.1

1. Genera el gráfico

\* Crea los boxplot de la longitud del pétalo en función de la especie de Iris.

\* Compara con scatter plots las variables entre sí.

\* Crea una nueva columna llamada proporción que es el ratio entre Sepal.Length y Sepal.Width. Podeis hacerlo en R base o usando el paquete dplyr. Los que no lo conozcan tienen información al final del documento

*Ejemplo , swiss dataset (NO)*

\* El conjunto de datos “swiss” contiene una medida estandarizada de fecundidad y varios indicadores socioeconómicos para cada una de las 47 provincias francófonas de Suiza.

1. ¿Qué diagrama dibujaría para mostrar la distribución de todos los valores? ¿Qué conclusiones sacarías?

2. Dibuje gráficos para cada variable. ¿Qué puede concluir de las distribuciones con respecto a su forma y posibles valores atípicos?

3. Dibuja un diagrama de dispersión de Fertilidad frente a % Catholic. ¿Qué tipo de áreas tienen las tasas de fertilidad más bajas?

4.¿Qué tipo de relación existe entre las variables Educación y Agricultura?

\* El conjunto de datos de aceites de oliva es bien conocido y se puede encontrar en varios paquetes, por ejemplo, como aceitunas en **extracat.**. La fuente original de los datos es el artículo [Forina et al., 1983].

1. Dibuje un scatterplot de las ocho variables continuas. ¿Cuáles de los ácidos grasos están fuertemente asociados positivamente y cuáles fuertemente asociados negativamente?

2. ¿Hay valores atípicos u otras características que valga la pena mencionar?

*Ejemplo , HSAUR2 dataset (NO)*

\* El conjunto de datos se llama Lanza del paquete HSAUR2.

1. Se informan los datos de cuatro estudios. Dibuje un diagrama para mostrar si los cuatro estudios son igualmente grandes.

2. El resultado se mide por la clasificación de la variable con puntuaciones de 1 (mejor) a 5 (peor). ¿Cómo describirías la distribución?

\* El paquete vcdExtra incluye datos de un viejo estudio de cáncer de mama sobre la supervivencia o muerte de 474 pacientes.

1. Convierta los datos en un data frame y dibuje gráficos para comparar las tasas de supervivencia, primero, por grado de malignidad y, en segundo lugar, por centro de diagnóstico.

2. ¿Qué diagrama dibujaría para comparar las tasas de supervivencia tanto por grado de malignidad como por centro de diagnóstico? ¿Importa el orden de las variables explicativas?

\* Dataset Crabs (del paquete MASS) [Venables y Ripley, 2002]. Los autores inicialmente se transforman a una escala logarítmica y luego escriben que:

“The data are very highly correlated and scatterplot matrices and brush plots [i.e. interactive graphics] are none too revealing.”.

Utilizando gráficos generales, comente si la transformación logaritmica fue una buena idea y si está de acuerdo con su afirmación sobre las correlaciones.