Material Imprimible

Curso de Análisis de datos con R

## Módulo V

**Contenidos:**

- Visualización en R con ggplot2
- gráficos de barra simples o agrupados
- Gráficos de línea
- histogramas
- Gráficos de dispersión
- Gráficos de distribución

## Visualización en R con ggplot2

La visualización de datos es esencial para comprender patrones, tendencias y relaciones en conjuntos de datos. En R existe un paquete que permite realizar de manera simple múltiples visualizaciones llamado ggplot2. Esta librería se basa en la filosofía de "Grammar of Graphics", proporcionando una sintaxis coherente y flexible para construir gráficos de manera intuitiva.

## Principales Características de ggplot2:

**Capas:** organiza los gráficos en capas, donde cada capa representa un componente del gráfico, como puntos, líneas o etiquetas. Esto permite construir gráficos complejos agregando capas de manera incremental, yendo de lo más simple a lo más complejo en solo algunas líneas de código.
**Aesthetics:** Las estéticas en ggplot2 definen cómo se mapean los datos a propiedades visuales del gráfico, como color, forma y tamaño. Se especifican mediante la función aes().
**Geometrías:** Las geometrías determinan el tipo de gráfico que se está creando, como puntos (geom_point()), líneas (geom_line()), barras (geom_bar()), entre otros.
**Facetas:** ggplot2 permite dividir los gráficos en facetas, creando paneles separados para diferentes categorías o niveles de una variable.

Antes de empezar a usar esta librería, los invitamos a conocer la documentación oficial y su hoja de trucos:
https://rstudio.github.io/cheatsheets/html/data-visualization.html

Para instalar, utilizar lo siguiente:
install.packages("ggplot2")
En el script:
library(ggplot2)

## Gráficos de barra simples o agrupados

Los gráficos de barras representan la frecuencia o cantidad de una variable categórica en un conjunto de datos. Los gráficos de barras simples muestran barras independientes

para cada categoría, mientras que los gráficos de barras agrupados colocan varias categorías en la misma barra, permitiendo comparaciones directas entre ellas.

Utilizan la geometría geom_bar() dentro de ggplot

```
ggplot(datos_barras, aes(x = Categoria, y = Valor)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Gráfico de Barras Simple", x = "Categoría", y = "Valor")
```

## Gráficos de línea

Los gráficos de línea conectan puntos de datos con líneas, mostrando la tendencia o cambio en una variable a lo largo de un eje continuo. Típicamente, en el eje X utilizan una variable de tiempo como el periodo, mes o año.

Usan la figura geom_line()

```
ggplot(datos_linea, aes(x = Tiempo, y = Valor)) +
  geom_line(color = "blue") +
  labs(title = "Gráfico de Línea", x = "Tiempo", y = "Valor")
```

## Histograma

Los histogramas representan la distribución de una variable numérica dividiéndola en intervalos y mostrando la frecuencia de observaciones en cada uno.

```
ggplot() +
  geom_histogram(aes(x = datos_histograma), bins = 30, fill = "green", color = "black") +
  labs(title = "Histograma", x = "Valor", y = "Frecuencia")
```

## Gráficos de dispersión

Los gráficos de dispersión muestran la relación entre dos variables numéricas mediante puntos en un plano cartesiano. Es interesante este tipo de visualización para poder entender cómo se distribuyen los puntos entre dos variables continuas, y permite validar visualmente la correlación entre las mismas.

```
ggplot(datos_dispersion, aes(x = Variable1, y = Variable2)) +
  geom_point(color = "red") +
  labs(title = "Gráfico de Dispersión", x = "Variable1", y = "Variable2")
```

## Gráficos de distribución

Los gráficos de distribución representan la distribución de una variable numérica y proporcionan información sobre su forma y concentración. Al ser una unica variable, el grafico de torta (Pie Chart) utiliza una modificación del geom_bar()

A continuación, se expresa un ejemplo:

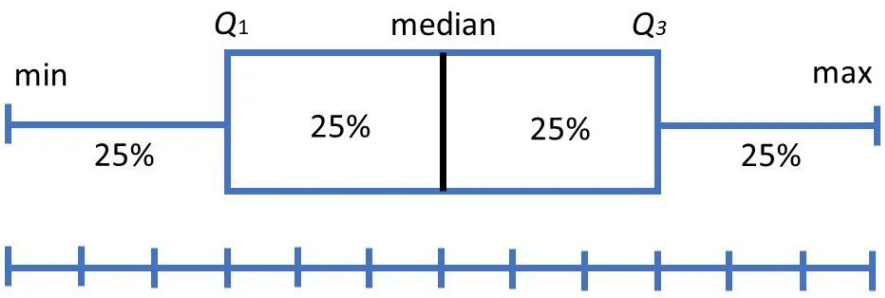
```
ggplot(datos_torta, aes(x = "", y = Valor, fill = Categoria)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  labs(title = "Gráfico de Torta", fill = "Categoría") +
  theme_void()
```

Se prescinde del eje Y y se eliminan líneas divisorias así el gráfico circular queda estéticamente como un gráfico de torta. Es importante que el eje X esté declarado como vacío (X = “”)

## Caja y Bigote / Boxplot

Un gráfico de caja o boxplot es una representación visual de la distribución estadística de un conjunto de datos. Proporciona una visión rápida de la mediana, los cuartiles, y los valores atípicos (outliers) del conjunto de datos.

```
ggplot(datos_boxplot, aes(x = Grupo, y = Valor)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.7) +
  labs(title = "Boxplot", x = "Grupo", y = "Valor")
```



El boxplot muestra las siguientes características:

- Caja: Representa el rango intercuartílico (IQR), es decir, el rango entre el primer cuartil (Q1) y el tercer cuartil (Q3). La línea en el centro de la caja es la mediana.
- Bigotes: Extienden hasta los valores más extremos dentro de 1.5 veces el IQR desde los cuartiles.