

Visualización en R

Ciencia de Datos para Economía y Negocios

Prof. Nicolás Sidicaro (UBA)

2025-10-31

Tipos de Gráficos según Variables

Criterios de decisión para visualizaciones efectivas

Objetivos del encuentro

- Desarrollar criterios para seleccionar el gráfico apropiado según los datos
- Analizar ventajas y limitaciones de cada tipo de visualización
- Aplicar la taxonomía "From Data to Viz" como herramienta de decisión
- Construir un marco conceptual para la toma de decisiones visuales

Marco Conceptual: From Data to Viz

La taxonomía "From Data to Viz"

Principio fundamental

La elección del gráfico debe basarse en el tipo de datos, no en preferencias estéticas

Clasificación por variables

- Categóricas (nominal, ordinal)
- Numéricas (continuas, discretas)
- Temporales (fechas, series)
- Espaciales (coordenadas, mapas)

Combinaciones comunes

- Una variable categórica
- Una variable numérica
- Categórica + Numérica
- Numérica + Numérica
- Series temporales

Preguntas clave antes de elegir

1. ¿Qué quiero comunicar?

- Comparación entre categorías
- Distribución de una variable
- Relación entre variables
- Evolución temporal
- Composición o partes de un todo

2. ¿Cuáles son mis limitaciones?

- Cantidad de categorías/observaciones
- Audiencia (técnica vs. general)
- Medio de presentación (impreso, digital, presentación)
- Tiempo disponible para interpretar

Variables Categóricas

Gráfico de barras: cuándo y por qué

- Comparar cantidades entre 2-15 categorías
- Ranking o jerarquización es importante
- Precisión en la lectura de valores
- Audiencia amplia (muy intuitivo)

Ventajas y limitaciones:








-  Comparación fácil entre categorías
-  Lectura precisa de valores
-  Escalable (horizontal para nombres largos)
-  Funciona bien en B&N
-  Solo valores positivos (barras tradicionales)
-  Espacio limitado para muchas categorías
-  No muestra proporciones del total automáticamente

Gráfico de barras: variaciones estratégicas

Barras horizontales vs. verticales

- Horizontales: nombres de categorías largos, muchas categorías, ranking natural
- Verticales: pocas categorías, nombres cortos, evolución temporal

Ordenamiento de barras

- Descendente: cuando el ranking importa
- Ascendente: para mostrar progresión o crecimiento
- Alfabético: cuando las categorías tienen orden lógico
- Por frecuencia: para destacar lo más común

Consideración clave

El orden importa más de lo que creemos: puede cambiar completamente el mensaje

Gráficos de torta: controversia y criterios

- Máximo 4-5 categorías
- Una categoría domina claramente (+50%)
- Concepto de "todo" es relevante
- Audiencia no técnica y contexto informal

Problemas fundamentales:




- ✗ Difícil comparar segmentos similares
- ✗ Ángulos son difíciles de interpretar para humanos
- ✗ Segmentos pequeños (<5%) son imperceptibles
- ✗ No funciona con valores negativos
- ✗ Ocupa mucho espacio para poca información

Alternativas casi siempre mejores:

- Barras horizontales ordenadas
- Donas (si necesitás el concepto de "todo")
- Waffle charts para proporciones

Gráficos de donas: evolución de la torta




Ventajas sobre la torta:

-  Espacio central para información clave (total, porcentaje principal)
-  Menos dependiente del área visual
-  Más moderno estéticamente

Cuándo usar:

- Cuando la torta sería apropiada, pero querés destacar el total
- Dashboard donde el espacio es limitado
- Una métrica principal con descomposición simple

Mantiene las limitaciones:









-  Mismos problemas de comparación que la torta
-  Máximo 4-5 segmentos
-  Requiere etiquetas externas para claridad

Variables Numéricas

Histogramas: explorando distribuciones

- Explorar distribución de una variable continua
- Identificar patrones: normalidad, sesgo, multimodalidad
- Detectar outliers o valores atípicos
- Análisis exploratorio de datos

Ventajas y limitaciones:

-  Información completa de la distribución
-  Identifica patrones no evidentes en resúmenes
-  Base para decidir análisis estadísticos posteriores
-  Fácil de interpretar con contexto adecuado
-  Sensible al número de bins: puede cambiar completamente la interpretación
-  Difícil comparar múltiples grupos
-  Requiere muchos datos (>50 observaciones idealmente)
-  No funciona con datos discretos con pocos valores

Número de bins: decisión crucial

Reglas generales:

- Pocos bins (5-10): patrones generales, presentaciones ejecutivas
- Muchos bins (20-50): análisis detallado, identificar múltiples modas
- Raíz cuadrada de n : regla conservadora general
- Prueba múltiples opciones: no existe "el número correcto"

Señales de alerta:









- Bins muy anchos: pueden ocultar patrones importantes
- Bins muy angostos: ruido que distrae del patrón general
- Un bin domina: considerar escala logarítmica
- Múltiples picos: ¿hay subgrupos en los datos?

Gráficos de densidad: suavizando la información

Cuándo preferir sobre histogramas:

- Comparar distribuciones de múltiples grupos
- Presentación más elegante (menos "escalones")
- Overlay de grupos fácil de interpretar
- Enfoque en la forma general más que en frecuencias exactas

Ventajas y Limitaciones:

-  Comparación múltiple sencilla
-  Independiente de bins
-  Estéticamente superior para presentaciones
-  Funciona bien con datos de diferente tamaño
-  Menos intuitivo para audiencia no técnica
-  Pierde información de frecuencias absolutas
-  Puede suavizar demasiado patrones importantes
-  Parámetros de suavizado afectan interpretación

Box plots: el resumen compacto

Información que proporciona:

- Mediana (línea central): valor "típico"
- Cuartiles (caja): rango del 50% central
- Rango intercuartil: variabilidad central
- Whiskers: extensión hasta $\sim 1.5 \times \text{IQR}$
- Outliers: puntos individuales fuera de whiskers

Cuándo es ideal:

- Comparar múltiples grupos (3 o más)
- Identificar outliers rápidamente
- Audiencia técnica que entiende los conceptos
- Espacio limitado pero necesitas información de distribución
- Datos con outliers que distorsionarían histogramas

Box plots: el resumen compacto

Limitaciones importantes:

- ✗ Pierde información de la forma de distribución
- ✗ Puede ocultar distribuciones bimodales
- ✗ Menos intuitivo para público general
- ✗ No adecuado para distribuciones muy asimétricas

Categórico + Numérico

Barras agrupadas vs. apiladas: decisión estratégica

Barras agrupadas - Cuándo usar:

- Comparar valores entre grupos y categorías
- Diferencias entre grupos son el foco principal
- Valores similares entre grupos (fácil comparación)
- Máximo 3-4 grupos por categoría

Barras apiladas - Cuándo usar:

- Total por categoría es relevante
- Composición dentro de cada categoría importa
- Muchos grupos (>4) hacen agrupadas ilegibles
- Proporciones relativas son más importantes que valores absolutos

Barras apiladas 100% - Caso especial:

- Cuando solo las proporciones importan
- Totales muy diferentes entre categorías
- Composición relativa es el mensaje principal

Box plots por grupos: comparación de distribuciones





Cuándo es la mejor opción:

- 3 o más grupos a comparar
- Distribuciones completas importan, no solo promedios
- Outliers pueden ser informativos
- Audiencia técnica o contexto analítico
- Sospechas de diferente variabilidad entre grupos

Cuándo evitar:

- ❌ Audiencia no técnica sin contexto estadístico
- ❌ Distribuciones muy asimétricas o bimodales
- ❌ Pocos datos por grupo (<20 observaciones)
- ❌ Presentación ejecutiva donde simplicidad es clave

Ventajas específicas:

-  Información completa en espacio compacto
-  Outliers visibles por grupo
-  Comparación de variabilidad entre grupos
-  Escalable a muchos grupos




Violin plots: lo mejor de dos mundos

Combinan box plot (resumen) + densidad (forma) en una visualización




Cuándo usar:

- Audiencia técnica que aprecia el detalle
- Forma de distribución es relevante para la decisión
- Comparar grupos con diferentes tipos de distribución
- Investigación o análisis donde el detalle importa

Ventajas únicas:

-  Información más completa que box plots
-  Detecta bimodalidad u otros patrones
-  Estéticamente atractivo para presentaciones técnicas

Limitaciones:








-  Complejidad visual puede confundir
-  Requiere explicación para audiencia general
-  Menos establecido que box plots

Cleveland dot plots: la alternativa elegante

Cuándo preferir sobre barras:

- Ranking o comparación de valores es clave
- Muchas categorías (>8) hacen barras problemáticas
- Precisión en lectura es importante
- Estética minimalista es preferida

Ventajas específicas y Limitaciones:

-  Menos tinta para la misma información
-  Maneja mejor muchas categorías
-  Enfoque en diferencias entre valores
-  Funciona bien con nombres largos
-  Menos familiar para audiencia general
-  No muestra "magnitud" tan claramente como barras
-  Puede requerir línea de referencia en cero






Numérico + Numérico

Scatter plots: explorando relaciones





Cuándo usar:

- Investigar relación entre dos variables continuas
- Buscar correlaciones o patrones no evidentes
- Identificar outliers en el contexto de dos variables
- Análisis exploratorio previo a modelado estadístico
- Validar supuestos de linealidad en análisis

Ventajas específicas:

-  Muestra relación directa entre variables
-  Identifica outliers bivariados (no visibles en análisis univariado)
-  Detecta no-linealidades y patrones complejos
-  Base para análisis de regresión y correlación
-  Funciona con cualquier cantidad de observaciones

Limitaciones:

-  Overplotting con muchas observaciones (puntos superpuestos)
-  Difícil interpretar sin contexto estadístico
-  Puede sugerir causalidad cuando solo hay correlación
-  Menos efectivo con relaciones muy débiles

Scatter plots: variaciones estratégicas

Por tamaño (bubble charts):

- Cuándo: tercera variable continua para representar
- Ejemplo: ingresos vs. educación, tamaño = población de la ciudad
- Cuidado: puede generar confusión si los tamaños son muy similares

Por color/forma (grupos):

- Cuándo: variable categórica adicional para segmentar
- Ejemplo: altura vs. peso, color = género
- Límite: máximo 5-6 grupos para mantener claridad

Con líneas de tendencia:

- Cuándo: relación es evidente y querés destacarla
- Tipos: lineal, polinomial, smooth (loess)
- Cuidado: no forzar líneas donde no hay relación clara

Cuándo NO usar scatter plots

Alternativas más apropiadas:

Variables discretas con pocos valores:

- Problema: puntos superpuestos ocultan frecuencias
- Solución: tablas de contingencia, heat maps

Una variable es categórica:

- Problema: ejes no son comparables
- Solución: box plots por grupo, violin plots

Relación temporal:

- Problema: conexión temporal se pierde
- Solución: líneas de tiempo, series temporales

Muchas observaciones (>10,000):

- Problema: overplotting severo
- Solución: hexbin plots, contour plots, sampling

Matrices de correlación: panorama completo

Cuándo usar:

- Múltiples variables numéricas (3 o más)
- Análisis exploratorio inicial
- Identificar colinealidad en datasets complejos
- Audiencia técnica familiarizada con correlaciones

Limitaciones:

- ✗ Solo relaciones lineales
- ✗ Pierde información sobre distribuciones
- ✗ Puede ser abrumadora con muchas variables
- ✗ Requiere interpretación estadística

Ventajas:

- ✓ Vista general de todas las relaciones
- ✓ Identifica patrones no evidentes
- ✓ Compacta mucha información
- ✓ Detecta redundancia entre variables

Heat maps de correlación: decisiones clave

Elementos de diseño críticos:

Escala de colores:

- Divergente (-1 a +1): destaca correlaciones positivas/negativas
- Secuencial (0 a 1): solo valores absolutos
- Evitar rainbow: confunde más que aclara

Ordenamiento de variables:

- Clustering jerárquico: agrupa variables similares
- Por importancia: variables clave primero
- Alfabético: cuando no hay orden natural

Información adicional:

- Valores numéricos: para precisión
- Significancia estadística: asteriscos o símbolos
- Tamaño de muestra: crucial para interpretación




Scatter plot matrices: herramienta exploratoria

Matriz de scatter plots para todas las combinaciones de variables





Cuándo usar:

- Análisis exploratorio profundo
- Dataset con 3-8 variables numéricas
- Búsqueda de relaciones no lineales
- Contexto de investigación o análisis técnico

Ventajas únicas:

-  Información más rica que matriz de correlación
-  Detecta no-linealidades y outliers
-  Histogramas en diagonal muestran distribuciones

Limitaciones:

-  Muy denso visualmente con más de 6 variables
-  Requiere pantalla grande o impresión de calidad
-  Tiempo de interpretación considerable
-  No apropiado para presentaciones generales

Criterios de Decisión Avanzados

Matriz de decisión: tipo de dato vs. objetivo

Tipo de Datos	Distribución	Comparación	Ranking	Composición	Relación
1 Categórica	Barras	Barras	Barras ordenadas	Torta/Dona	N/A
1 Numérica	Histograma/Densidad	Box plot	Cleveland dots	N/A	N/A
Cat + Num	Violin plots	Barras agrupadas	Cleveland dots	Barras apiladas	N/A
Num + Num	Scatter plots	N/A	N/A	N/A	Scatter/Correlación
Múltiples Num	Scatter matrix	Heat map correlación	N/A	N/A	Matriz correlación

Consideraciones adicionales:

- Audiencia: técnica vs. general
- Medio: impreso vs. digital vs. presentación
- Cantidad de datos: pocos vs. muchos puntos
- Tiempo de interpretación: rápido vs. análisis detallado

Factores contextuales críticos

1. Audiencia

- Ejecutiva: simplicidad, mensaje claro, barras y líneas
- Técnica: puede manejar box plots, violin plots, densidades
- Pública general: barras, líneas simples, evitar complejidad

2. Medio de presentación

- Impreso B&N: evitar dependencia del color, patrones claros
- Presentación oral: gráficos simples, texto grande
- Dashboard interactivo: puede ser más complejo
- Reporte técnico: información densa aceptable

3. Cantidad de datos

- Pocos puntos (<10): considera tablas en lugar de gráficos
- Muchos puntos (>1000): histogramas, densidades, agregaciones
- Datos faltantes: impacto en tipo de visualización

Errores comunes en la selección

1. Selección por estética, no por funcionalidad

- Problema: gráfico "bonito" pero inapropiado para los datos
- Solución: función primero, forma después

2. No considerar la cantidad de categorías

- Problema: 15 categorías en un gráfico de torta
- Solución: agrupar, filtrar, o cambiar tipo

3. Ignorar las limitaciones de la audiencia

- Problema: violin plot para presentación ejecutiva
- Solución: adaptar complejidad al contexto

4. Forzar todos los datos en un gráfico

Checklist de validación

Antes de finalizar tu elección:

1. ¿El tipo de gráfico coincide con el tipo de variables?
2. ¿La audiencia podrá interpretar este gráfico sin explicación extensa?
3. ¿El gráfico responde la pregunta principal de forma directa?
4. ¿Hay alternativas más simples que comuniquen lo mismo?
5. ¿Los datos son suficientes para justificar este tipo de gráfico?
6. ¿El medio de presentación es compatible con esta visualización?

Herramientas de Decisión

Algoritmo de decisión simplificado

1. ¿Cuántas variables tengo?

- └ Una → ¿Categórica o numérica?
- └ Dos o más → ¿Qué combinación?
 - └ Cat + Num → Comparar grupos
 - └ Num + Num → Buscar relaciones

2. ¿Cuál es mi objetivo principal?

- └ Distribución → Histograma/Box plot
- └ Comparación → Barras/Box plots múltiples
- └ Ranking → Barras ordenadas/Cleveland dots
- └ Composición → Barras apiladas/Dona
- └ Relación → Scatter plots/Correlación

3. ¿Mi audiencia maneja complejidad técnica?

- └ Sí → Violin plots, densidades, matrices OK
- └ No → Barras, líneas simples, scatter básicos

4. ¿Tengo limitaciones de espacio/medio?

- └ Sí → Priorizar simplicidad
- └ No → Puedo usar gráficos más informativos

Recursos de referencia rápida

Sitios web para consulta:

- From Data to Viz (data-to-viz.com): árbol de decisión interactivo
- R Graph Gallery: ejemplos por tipo de dato

Preguntas para auto-evaluación:

1. ¿Puedo explicar por qué elegí este gráfico en 15 segundos?
2. ¿Una persona nueva podría interpretar esto sin mi ayuda?
3. ¿Hay una alternativa más simple que comunique lo mismo?
4. ¿Los datos justifican esta complejidad visual?

Regla de oro:

Cuando dudes entre dos opciones, elegí la más simple

Construcción de criterio personal

Desarrollar intuición requiere:

1. **Práctica deliberada:** probar diferentes opciones con los mismos datos
2. **Análisis crítico:** evaluar gráficos que encuentres en medios y reportes
3. **Feedback:** mostrar opciones a colegas y observar sus reacciones
4. **Iteración:** raramente el primer intento es el óptimo

Ejercicio recomendado:

Con cualquier dataset, crear 3 **gráficos diferentes** para la misma pregunta y evaluar cuál comunica mejor según el contexto.

