FEB
23

SPRINT 1 - ENCUENTRO 7

Seaborn

Unirme

El encuentro comenzó hace
01:07:28

El arte de transformar

> *"Effective data visualization is both an art as well as a science." – John Tukey* **Estadístico.**

Agrandamos la caja de herramientas

En la bitácora anterior empezamos a conocer las primeras herramientas de visualización donde vimos que, lejos de ser un complemento a nuestro trabajo, son una parte fundamental. Las visualizaciones bien logradas nos dan la posibilidad de chequear los datasets con los que estamos trabajando, conocerlos mejor, al tiempo que nos permiten detectar errores y comunicar de una manera efectiva nuestras hipótesis a partir de la correcta lectura de los datos.

Conocimos también una primera librería que nos facilita realizar los gráficos: Matplotlib. En esta bitácora, nuestros esfuerzos van a estar enfocados en ampliar nuestro conocimiento para hacer más eficiente el uso de los datos y nuestro tiempo interactuando con ellos. Para ello, vamos a adentrarnos en **Seaborn.**

Esta nueva librería se basa en Matplotlib (cuando es así, se dice que Seaborn *corre* sobre Matplotlib) lo que hace que sea muy fácil combinarlas. Si sabes usar Matplotlib, es sencillo aprender Seaborn dado que trabaja con objetos definidos en esa librería, como por ejemplo figuras y ejes. ¿Cuáles son las ventajas de usar esta nueva librería?

- Facilita el trabajo con DataFrames de Pandas.
- Mejora automáticamente la estética de los gráficos.
- Permite una sintaxis sencilla para algunos gráficos complejos.

Aunque también tiene algunas desventajas:

- Precisa la instalación de una librería adicional (esto puede ser perjudicial en algunos contextos)
- Es menos flexible, o más compleja de configurar que Matplotlib

---

Tip: ¿Cómo decidimos cuándo utilizar una o la otra? Si quieres obtener gráficos rápidos, eligiendo dentro de una paleta predefinida y estéticamente agradables, es mejor usar Seaborn. Si quieres total control sobre tu gráfico, Matplotlib es la herramienta correcta.

Sumemos variedades

Durante la Bitácora y el encuentro anterior vimos algunos tipos de gráficos, principalmente orientados a visualizar relaciones entre **variables numéricas**. Hoy ampliaremos la variedad, agregando algunos tipos orientados a **variables numéricas** (diagrama de cajas) y otro a **variables categóricas** (gráficos de barra y de torta). Recordemos que esta última es una variable que puede tomar valores de un conjunto limitado, y en general fijo, de posibles valores. Seaborn incluye una interesante variedad de gráficos categóricos (categorical plots), por lo que veremos algunas características generales de ellos para poder aprovecharlos mejor.

La gran mayoría de los conjuntos de datos que usamos en Data Science combinan variables numéricas y variables categóricas. Veamos a continuación las formas de representar a cada una de ellas.

Representaciones para variables numéricas

Si queremos estudiar cómo está distribuida una variable numérica, vimos anteriormente que una muy buena opción son los histogramas, junto con algunos estadísticos que vimos en la bitácora 4.

En esta bitácora presentaremos otro gráfico: los diagramas de cajas o box plots. Ambos podemos usarlos para representar la distribución de una **variable numérica**, pero ¿por qué utilizaríamos un diagrama de caja en lugar de un histograma? Los diagramas de caja suelen ser más útiles para comparar la distribución de diferentes variables a la vez. De esta forma, obtenemos un gráfico más conciso que si hiciéramos un histograma para cada variable. Sin embargo, contiene menos información que un histograma y dos variables con distinta distribución pueden tener el mismo diagrama de caja. Al final de The *Datasaurus Dozen*, pueden encontrar ejemplos de este caso. De paso, puedes mirar qué es un gráfico de Violín.

Para poder realizar un diagrama de cajas tenemos que:

1. Crear una "caja" con los cuartiles Q1 y Q3 (25% y 75%). El centro de la caja es la mediana (50%).
2. Calcular el rango intercuartílico con los cuartiles Q1 y Q3 (IQR = Q3 - Q1) para crear los "bigotes", que corresponden a Q3 (+ 1.5xIQR) y Q1 (- 1.5xIQR).
3. Graficar aquellos puntos que estén por fuera de los bigotes.

En la siguiente figura podemos ver a qué región de una distribución normal corresponde cada componente:

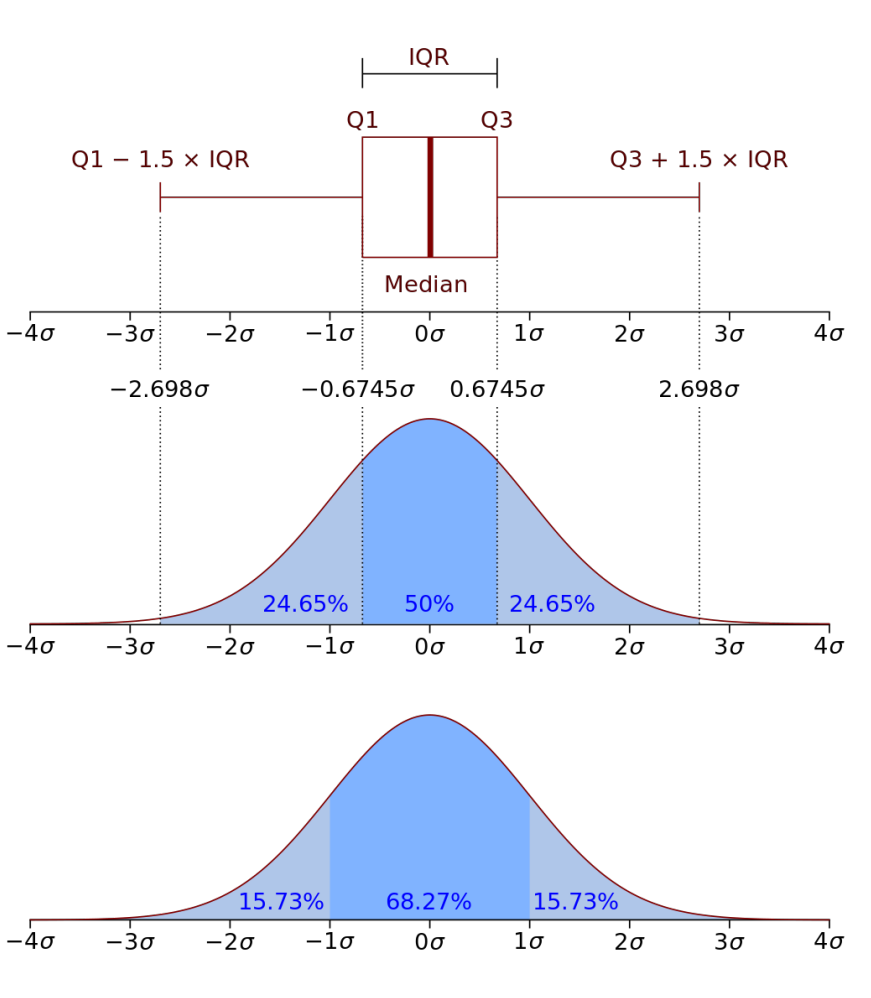


Figura 1: Componentes de un diagrama de cajas y su equivalencia en una distribución normal. Es importante aclarar que la caja siempre abarca un 50% de la distribución, ya que corresponde al rango de valores entre Q1 y Q3. Pero para otras distribuciones distintas a la normal no es necesariamente cierto que entre los extremas de la caja y los bigotes abarque el 24.65% de la distribución. Fuente: Wikipedia.

Con estos datos, tenemos suficiente información para darnos una idea cómo está distribuida nuestra variable, si la distribución es asimétrica o no. También identificar si hay valores atípicos (outliers), que por lo general están por fuera de los bigotes. En el segundo bloque de la carrera veremos cómo se utilizan las técnicas para detectarlos.

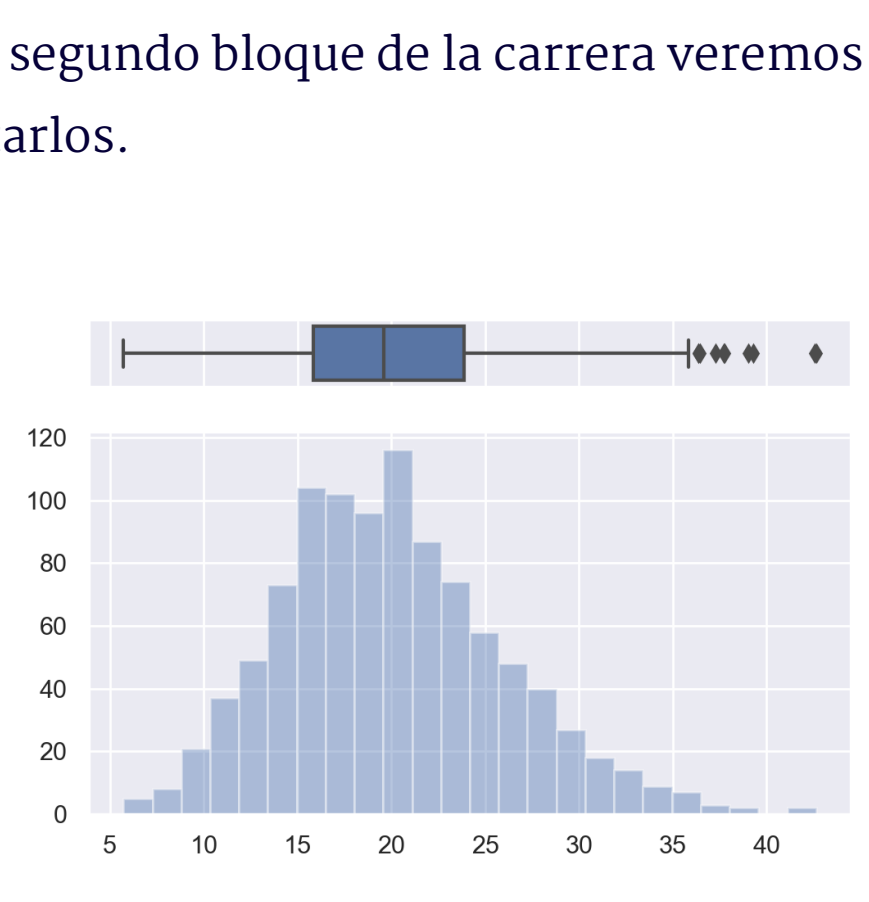


Figura 2: Histograma de una muestra de datos y su diagrama de cajas.

Representaciones para variables categóricas

¿Y si queremos estudiar una variable categórica? En general, esto suele ser más sencillo que el caso de variables numéricas.

En el caso de una sola variable, no podemos hacer mucho más que contar cuántas categorías hay, cuántas instancias pertenecen a cada categoría y, además, qué porcentaje representan. Esto se puede hacer con los **gráficos de barras.** Se podría argumentar que son lo mismo que un histograma, pero para una variable categórica. Esa persona tendría mucha razón. Muchas veces son indistinguibles de un histograma y, de hecho, un histograma es un tipo de gráfico de barras. Pero no está mal reservar la palabra histograma para una variable numérica y gráfico de barras para el caso de una variable categórica (¡Aunque tampoco es necesario! ¡Qué confuso!).

Tenemos que tener en cuenta que los histogramas son un tipo de gráfico de barras, por eso a veces son difíciles de distinguir y hasta se los nombran de la misma manera. Pero es preferible mantener ese término para las visualizaciones que representan variables numéricas, y usar gráfico de barras para las que utilizan variables categóricas.

Para los próximos ejemplos utilizaremos un famoso dataset: el del Titanic. Allí encontramos información sobre pasajeros del transatlántico: edad, género, clase en la que viajaba, si sobrevivió o no, y mucho más. Pueden ver más sobre este dataset en la siguiente competencia de Kaggle: Titanic: Machine Learning from Disaster.

En el siguiente gráfico, podemos ver cuántos pasajeros había en cada clase del Titanic según este dataset:

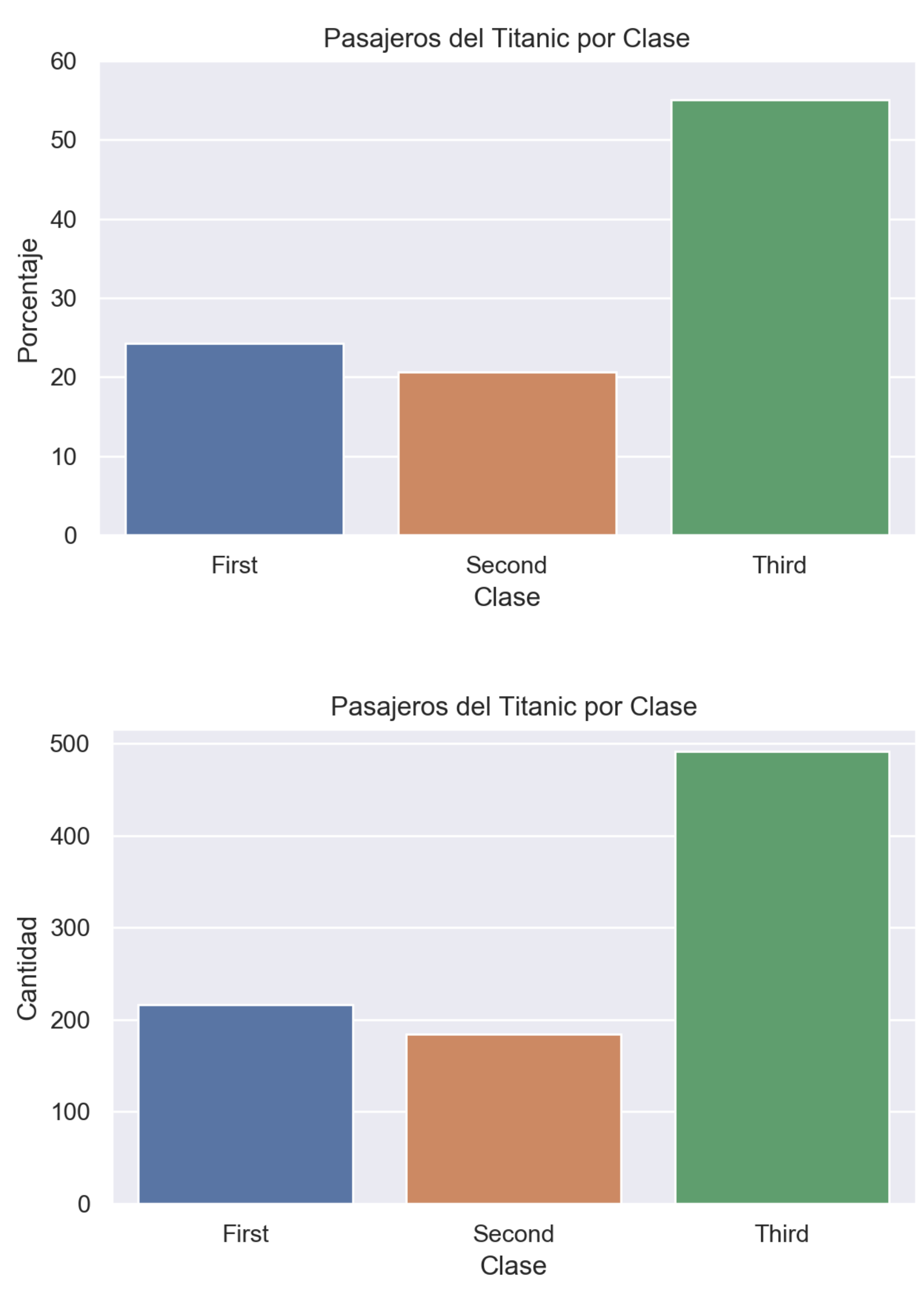


Figura 1.

A veces nos interesa ver cómo está distribuida una variable *desagregada* por otra variable, es decir, separada por casos donde por lo general se ven afectadas por variables categóricas. En la Figura 2, podemos observar cuántas personas había en cada clase (variable categórica) desagregadas por supervivencia (otra variable categórica):

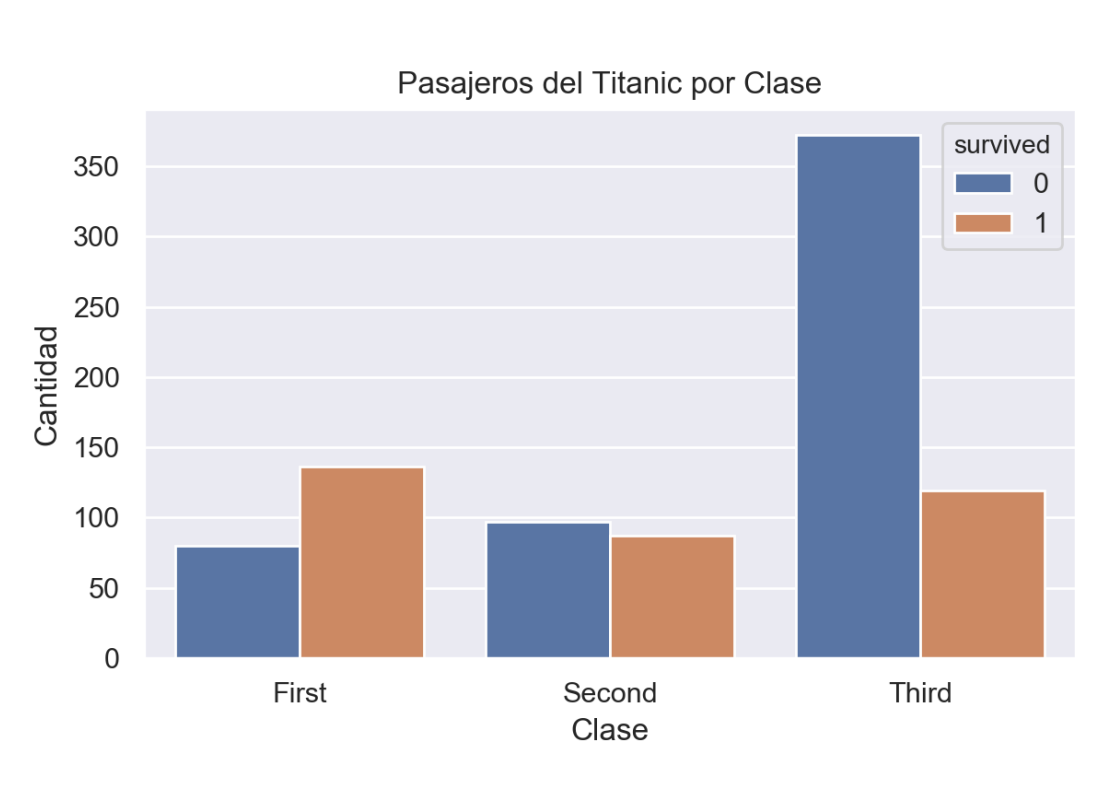
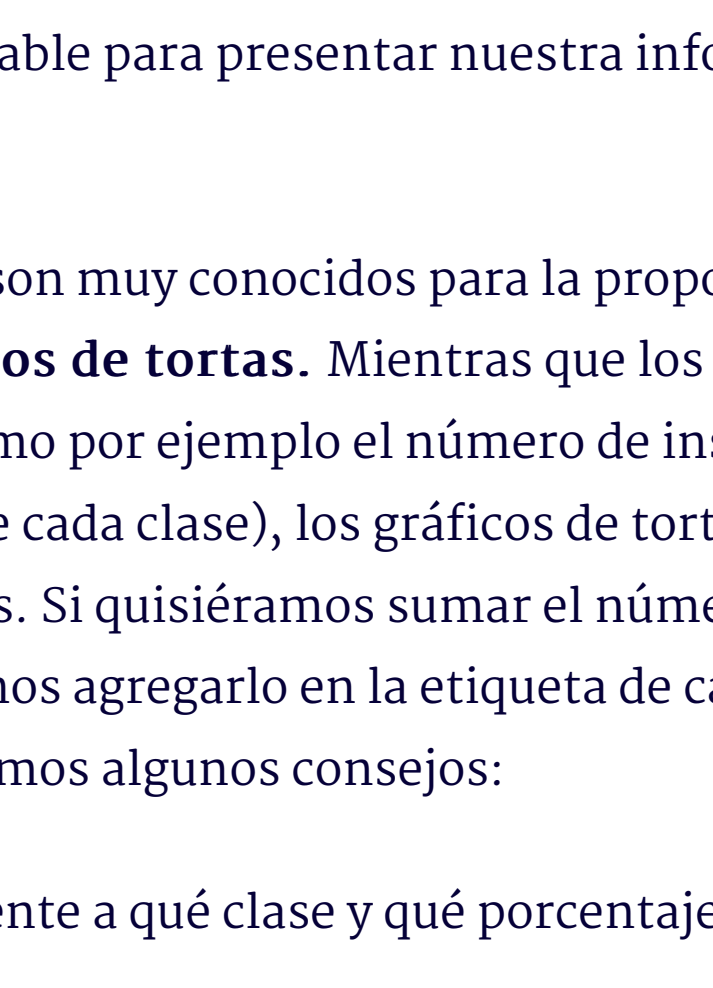


Figura 2.

Hay más de una manera de presentar las barras. Pueden ser verticales u horizontales, o pueden estar apiladas (stacked bar chart), etc. Así que debemos siempre prestar atención a qué representa cada barra y, si creemos que hay un formato que es más favorable para presentar nuestra información, ¡buscar cómo hacerlo!

Otros gráficos útiles que son muy conocidos para la proporción entre variables categóricas son los **gráficos de tortas.** Mientras que los gráficos de barras pueden ser absolutos (como por ejemplo el número de instancias por clase) o relativos (el porcentaje de cada clase), los gráficos de tortas solamente representan proporciones. Si quisiéramos sumar el número de instancias a un gráfico de torta, deberíamos agregarlo en la etiqueta de cada clase, lo que podría saturar el gráfico. Te dejamos algunos consejos:

- Etiqueta correctamente a qué clase y qué porcentaje corresponde cada porción de la torta.
- Si hay algunas clases con muy bajo porcentaje que saturan una zona del gráfico con el texto de sus etiquetas, agrúpalos en una nueva categoría, *Otros.*
- No abuses de los gráficos de torta y nunca los hagas en perspectiva o tridimensionales. De esa forma son más difíciles de leer y se modifican las proporciones.



Hasta acá vimos algunas formas más de visualización, ¿hay alguna que no hayamos mencionado aún? ¿alguna otra que sea utilizada con mayor frecuencia?

Empieza a trabajar en el notebook

Instala Seaborn (recuerda activar el ambiente y usar conda) y trabaja en el siguiente notebook, ya sea en Colab o en tu computadora. Para el encuentro, lleva hecha las secciones 1 y 2, el resto del notebook lo trabajaremos en clase.

Notebook: DS_Bitácora_07_Seaborn

¡Prepárate para el próximo encuentro!

Profundiza
Te invitamos a conocer más sobre el tema de esta bitácora.

Comunidad
Sumérgete en la red donde los profesionales de DS interactúan.

Rincón Matemático
¡¡¿¿Matemática??!! No te asustes... aquí tienes herramientas para revisar el contenido necesario para sobrevivir.

Challenge
Te proponemos el siguiente desafío, ¿te animas?

Lista de recursos

- Comunidad
- Profundiza
- Rincón Matemático
- Challenge
- Notebook: DS_Bitácora_07_Seaborn