Cápsula 1: Organización de datos tabulares

Hola, bienvenidxs a una cápsula del curso Visualización de Información. En esta hablaré sobre organización de datos tabulares.

Por **organización de datos tabulares**, me refiero a una decisión de diseño que cubre todos los aspectos del uso de **canales espaciales como codificaciones** visuales. Como vimos en las cápsulas de percepción, son canales espaciales bidimensionales los más efectivos para atributos tanto ordenados como categóricos.

Por eso, el organizar espacialmente es la decisión de codificación visual **más importante** a realizar en una visualización, porque crea el mapa mental del usuario de cómo entenderá un *dataset*.

Particularmente lo hablamos en el contexto de *datasets* tabulares, porque estos **no necesariamente tiene ya una representación espacial intrínseca**, y al transformarlos a algo visual, debemos darles una forma espacial. Para el caso de *datasets* espaciales y de redes, los veremos de forma separada ya que tienen sus propias particularidades.

Si recordamos, los *datasets* tabulares se entienden como simplemente un grupo de ítems que tienen valores según ciertos atributos. Dichos atributos pueden ser de tipo categórico, ordinal o cuantitativo.

También, la distinción entre atributos de **llave y de valor** es muy relevante cuando queremos organizar espacialmente a un conjunto de datos. Revisaremos distintas formas de dividir el espacio dependiendo de la cantidad de atributos llave y de valor presentes. Por lo general, se utilizan llaves para definir regiones de espacio para cada ítem, en donde se muestran uno o más valores.

Por ejemplo, un gráfico de barras es un *idiom* que muestra un atributo llave y un atributo valor. Los mapas de calor son *idioms* que usan dos atributos que forman una llave y un tercer atributo de valor. También pueden no haber llaves presentes, como en el caso de los gráficos de dispersión o *scatter plots*.

En las siguientes cápsulas, dividiremos la decisión de organización espacial de datos tabulares en tres sub-decisiones: expresión de valores cuantitativos; definición de regiones categóricas; y orientación de ejes.

También iré mencionando distintos *idioms* recurrentes en las distintas decisiones y los clasificaremos utilizando el modelo anidado.

Comenzaremos por la decisión de **expresión de valores cuantitativos**. Aún refiriéndonos al uso del espacio como codificación, en este caso, de atributos cuantitativos. En este caso el

uso es bastante directo, ya que es posible hacer una correspondencia entre los valores de un atributo a una posición espacial a lo largo de un eje.

El caso más simple es un atributo a lo largo de un único eje, marcando su posición con alguna marca. Más atributos pueden codificarse al mismo tiempo utilizando canales no espaciales como el color o tamaño. Utilizar al mismo tiempo más canales espaciales es posible, pero limitado. Para eso, se puede utilizar más ejes, o eventualmente dentro de glifos usados como marcas.

El **gráfico de dispersión o *scatter plot*** es un buen ejemplo de uso del espacio para codificar valores cuantitativos. Codifican dos atributos mediante posición horizontal y vertical de forma simultánea, y utiliza la marca de punto.

Suelen expandirse utilizando color para codificar un atributo extra, e incluso se utiliza el tamaño de la marca para un cuarto atributo. Aunque generalmente en este último caso se conoce como gráfico de burbujas o *bubble plot*.

Los gráficos de dispersión son efectivos para tareas que involucren proveer un resumen o caracterizar distribuciones, como también encontrar *outliers* o valores extremos.

Son aún más efectivos para la tarea abstracta de **juzgar correlación** entre dos atributos, ya que perceptualmente se traduce en identificar si los puntos se alinean en una diagonal. Mientras más cercanos a la diagonal, más estrecha es la correlación. Muchas veces también se hace uso del dato derivado correspondiente a la línea de regresión correspondiente para buscar esta correlación.

Lo que podría limitar a un *scatter plot* es su escalabilidad cuando es necesario identificar puntos entre ellos, dejando espacio para docenas o cientos de elementos. En pantalla puedes ver un resumen de lo mencionado utilizando el *framework*, y explicaremos esto para las distintas codificaciones que vayamos revisando.

Con eso termina el contenido de esta cápsula. Recuerda que si tienes preguntas, puedes dejarlas en los comentarios del video para responderlas en la sesión en vivo de esta temática. ¡Chao!