Cápsula 2: Agregación

Hola, bienvenidxs a una cápsula del curso Visualización de Información. En esta hablaré sobre agregación, otra de las opciones de decisión de diseño de reducción de datos.

Aquí la intención es juntar un grupo de elementos, y reemplazarlos por un único elemento que actúa como representante del grupo. De forma similar a la filtración, es posible agrupar ítems individuales de un *dataset*, como también agrupar atributos.

Usualmente la agregación implica usar algún tipo de dato derivado. Casos simples de operaciones de agregación son el promedio, el conteo, la suma, mínimo y máximo. Pero como bien sabemos, usar este tipo de medidas implican pérdida de información. Esta es la fuerza que hay que contrarrestar cuando trabajamos con agregación.

De todas formas es muy poderosa, sobre todo en herramientas interactivas donde el nivel de agregación puede cambiarse según la demanda del usuario.

La forma más directa de aplicar agregación es mediante *idioms* estáticos. El histograma es probablemente el *idiom* que utiliza agregación más conocido. Estos muestran la distribución de elementos según un atributo, usualmente del *dataset* original, separado en algún número de contenedores.

Los canales visuales utilizados son los mismos que el gráfico de barras, pero usualmente los datos que codifican originalmente son distintos. El histograma no muestra ítems individuales directamente, si no que genera una agregación de ellos a modo de mostrar una versión resumida y más concisa de los datos originales.

El número de contenedores usado puede ser independiente del número original de ítems que resume, y cuántos son afecta directamente a cómo se aprecian los datos. Una opción es escoger ese número en base a características del *dataset*, o permitir al usuario escogerlo mediante interacción.

Para notar esas diferencias podemos ver el *idiom* de *sand dance*, que vemos en pantalla. Esta herramienta específica la desarrolló Microsoft para visualizar datos de pasajeros del Titanic. Cada cuadrado pequeño individual representa un ítem de este dataset, y al agruparlos genera regiones de forma similar al histograma.

Permite también cambiar el número de contenedores, con lo que podemos apreciar el nivel de detalle que se percibe a mayor número. Pero al mismo tiempo, requiere de más espacio para mostrar la misma cantidad de datos. Este mismo tipo de interacción se puede traspasar a histogramas, y generar agregación interactivamente.

El *sand dance* es un caso interesante y especial, ya que no realiza agrupación como lo propusimos. No representa directamente grupo de ítems del *dataset* original con un solo elemento. Todos los ítems están presentes en la visualización, solo se ordenan de forma distinta.

Son las figuras que generan, los bloques generados por proximidad, lo que podemos pensar como elementos agregados. Este es un buen ejemplo de un punto medio entre aplicación de *facet* y reducción de datos mediante agrupación.

Otro ejemplo de *idiom* que usa agregación es el gráfico de dispersión continuo. Los gráficos de dispersión corrientes pueden presentar problemas de oclusión cuando el número de elementos es muy alto, dificultando la apreciación de cuántos elementos hay realmente debido a la superposición de muchos puntos.

Los gráficos de dispersión continuos buscan arreglar ese problema, graficando valores agregados de cantidad de elementos presentes en cada pixel, en vez de efectivamente mostrar cada elemento por separado. En este caso el atributo codificado mediante color es derivado de la densidad de posicionamiento, por lo que sería sólo calculable una vez que se intenta mostrar visualmente.

Otro ejemplo común es el diagrama de caja, que muestra mediante un glifo un resumen estadístico agregado de los valores que presenta un atributo cuantitativo. Presenta cinco puntos de interés: la mediana, los cuartiles inferior y superior, y los límites inferior y superior. Los cuartiles y mediana marcan puntos de cómo se distribuyen los valores al ordenarse: el 25%, el 75% y el 50%, respectivamente. Se utiliza una caja para marcar esa región y enfatizar el grueso central de la distribución de valores.

Por otro lado los límites se ubican cerca de los extremos del rango de valores, y todo lo que esté más allá de un límite se considera un dato atípico. Estos se codifican mediante líneas, conocidas como bigotes, que marcan el rango completo de la distribución. Mediante este glifo es posible codificar de forma muy condensada visualmente la distribución de un atributo.

También hay variaciones de diagramas de caja mediante vasijas, que representan mediante otra dimensión espacial la densidad dentro de la caja en la sección entre cuartiles. Eso permite apreciar si hay múltiples modas dentro de una distribución, por ejemplo. Claro que esta opción requiere de un poco más de espacio, y es un ejemplo normal de *trade-off*s de codificaciones.

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!