Manejo de Datos Abiertos en la Agricultura y Nutrición

Este curso de aprendizaje digital (e-learning) es el resultado de una colaboración entre socios de GODAN Action, incluyendo a ***Investigaciones Ambientales Wageininen (WUR)***, ***AgroKnow***, ***AidData***, la ***Organización de las Naciones Unidas para la Alimentación y la Agricultura*** (FAO por sus siglas en Ingles), ***El Foro Global sobre Investigaciones de Agricultura*** (GFAR), y el ***Instituto de los Estudios del Desarrollo*** (IDS), ***the Land Portal***, ***el Instituto de Datos Abiertos*** (IDI) y el ***Centro Técnico de Agricultura y cooperación Rural*** (CTA).

GODAN Action es un proyecto de tres años [por] el Departamento del Desarrollo Internacional del Reino Unido para capacitar a los que usan, producen, e intermediarios de datos para conectarse efectivamente con datos abiertos y maximizar la potencial por su impacto en los sectores de agricultura y nutrición.

GODAN ACTION
Global Open Data for Agriculture & Nutrition

UKaid
from the British people

Unidad 2: Usando los datos abiertos

Lección 2.3: Analisis de Datos y Visualizacion

Foto por Sebastián Sikora licenciado bajo CC BY 2.0

Objetivos y metas de aprendizaje

Se puede utilizar a datos abiertos cuando

Después de estudiar esta lección debes saber cómo:

- Explicar por qué los datos se deben analizar
- Preparar los datos para análisis
- Aplicar un numero de técnicas para analizar los datos
- Analizar los riesgos de analizar diferentes tipos de datos
- Explicar el propósito de visualización de datos
- Escoger una visualización de datos adecuado
- Evaluar la efectividad de un número de visualizaciones diferentes de datos

Contenido

Lista de gráficas

Lista de tablas

Introducción

La lección 2.1 introdujo métodos para hallar datos que están “en” y “dentro de” la web; la meta de esa lección era para introducir las técnicas para encontrar datos de la fuente original. La lección 2.2 la siguió con un análisis de la calidad y procedencia de los datos y los pasos necesarios para limpiar y preparar a los datos para análisis.

Esta lección mira a la próxima etapa del análisis de datos y visualización. Como otras lecciones, el conocimiento necesario variará mucho dependiendo del objeto exacto del análisis y visualización. Esta lección intenta tomar una perspectiva pragmática a los dos temas y ofrecer una teoría amplia de análisis que conduce a visualizaciones que generan un impacto. Los ejemplos prácticos se enfocan en datos cuantitativos en una hoja de cálculo y la extracción de características basado en el análisis de datos cualitativos.

El propósito de análisis de datos

Datos brutos sin procesar suelen estar desordenados y todavía no disponibles para visualizar. Esta sección mira a un número de técnicas que se puede utilizar para convertir a los datos en información incluso a:

- Derivación y extracción de características
- Combinación de conjuntos de datos
- Enriquecimiento de conjuntos de datos

1. Derivación y extracción de características

La derivación y extracción de características son similares en el sentido que son diseñados para añadir a datos existentes sin necesitar conjuntos de datos externos. Esto resulta en columnas adicionales (o características) que se añade basados en los datos únicamente.

La derivación

Un elemento de dato derivado es un elemento derivado de otros elementos de datos realizando una transformación matemática, lógica, u otro tipo (por ej. Aritmética, de fórmula, composición, o agregación).

Por ejemplo, datos de fuente original puede contener una serie de columnas de gastos mensuales. Se podría agregar una columna de totales para todos los meses. Así, esta columna es derivada de los otros

Extracción de característica

La extracción de característica es muy similar a los datos derivados, pero no necesariamente necesita involucrar una función.

Por ejemplo, se puede extraer la ciudad de una lista de datos no estructuradas de direcciones, haciendo que la ciudad sea una característica distinta del conjunto de datos. De igual manera, el color del pixel central de una imagen puede ser extraída sin el uso de una función matemática.

Un tutorial breve de extracción de característica disponible para Excel se encuentra aquí[1].

1.1. Combinando a conjuntos de datos

Aquí no referimos a adjuntar más registros al final de un conjunto de datos existente (un proceso conocido como consolidación), sino un proceso de añadir a los datos existentes.

Si estas usando una hoja de cálculos, el resultado de combinar dos conjuntos de datos con 10 columnas cada uno será un conjunto de datos con 19 columnas. Combinar datos sólo se puede hacer así si hay una columna en cada conjunto de datos con un valor en común. Esta columna se convierte en la clave por lo cual se puede realizar la combinación.

Por ejemplo, considera un conjunto de datos de calles y las cantidades de carros durante cada hora del día. Cada periodo de una hora es una columna y cada calle es una fila. Si hubieran varias fuentes de los datos, como fuentes con diferentes horas en todos recopilados individualmente, entonces podría usar el nombre de la calle como clave y combinar a todos los datos en un solo conjunto de datos.

Descubre como combinar dos conjuntos de datos en Excel[2] o en *OpenRefine*[3].

1.2. Enriqueciendo a datos geográficos

Muy similar a combinar datos, donde dos conjuntos de datos se pueden combinar basados en una característica en común (por ej. Nombre de calle), lo mismos se puede hacer con datos geográficos. Una gran diferencia es que es posible combinar datos geográficos basado en la ubicación y lugar de un punto geográfico dentro de límites. Este proceso se conoce como *spatial join* (unión espacial).

La unión espacial puede ser útil cuando estas mirando a características de diferentes edificios o servicios y mapeándolos según una jurisdicción regional para ver si algunas tendencias emergen. Por ejemplo, esto puede ser utilizado para ver si las escuelas en áreas controladas por concilios diferentes tienen un desempeño significativamente diferente.

Aprenda más sobre enriquecer y unir datos geográficos usando las guías excelentes de CartoDB[4].

Otra etapa esencial en preparar datos geográficos para análisis es el de geocodificación. La geocodificación es el proceso de tomar cualquier referencia o descripción de una ubicación física y agregar las coordenadas actuales del lugar a los datos. La geocodificación inversa por lo tanto es el opuesto, donde se extrae la descripción (por ej. Londres) de las coordenadas. La geocodificación también puede hacer referencia al proceso de convertir a los datos de un sistema de representación de coordenadas (por ej. “easting-northing”) a otro (por ej. “Latitud-longitud”). La geocodificación es

[1] https://www.extendoffice.com/documents/excel/3639-excel-extract-part-of-string.html
[2] https://superuser.com/questions/366647/merge-two-excel-files-using-a-common-column
[3] https://blog.ouseful.info/2011/05/06/merging-datesets-with-common-columns-in-googlerefine/
[4] https://carto.com/learn/guides

esencial cuando uno intenta analizar datos geográficos y realizar otras operaciones como uniones espaciales.

2. Análisis de datos cualitativos y cuantitativos

Hay dos tipos principales de análisis estadístico que se menciona en referencia de los datos: cualitativo y cuantitativo. Simplemente, investigación cualitativa se trata de datos sin límites fijos, mayormente en texto, mientras que la investigación cuantitativa intenta enfocarse en datos medibles y objetivos en la forma de números u otros tipos de datos estructurados. La Tabla 1 (de la Universidad Abierta) muestra algunas de las diferencias principales entre los dos tipos de investigación.

Tabla 1: ***investigación cualitativa vs cuantitativa (Fuente: Universidad Abierta)***

	investigación cualitativa	Investigación cuantitativa
Tipo de conocimiento	subjetivo	Objetivo
Objetivo	Explorativo y observacional	Experimental y universal
Características	Flexible, mostrado en contexto, perspectiva dinámica y continuo de cambio	Controlado y fijo, variables independientes y dependientes, medición de cambio de antes y después.
Selección de muestra	Con propósito	Aleatorio
Colección de datos	Semi-estructurado o no estructurado	Estructurado
Naturaleza de datos	Narrativas, citas, descripciones, valores únicos y particulares	Números, estadísticas, replicación
Análisis	temáticos	estadístico

Aunque la tabla muestra que la investigación cualitativa y cuantitativa como distinto y opuesto, en práctica suelen de estar combinados o usar elementos el uno del otro, por ejemplo, una pregunta de encuesta se puede contestar en una escala de 1-10 pero la pregunta puede ser sujeto a una parcialidad personal.

Aun si algo tan simple como contar ovejas puede ser peligroso si el campo también tiene corderos, ¿estos se pueden considerar ovejas? ¿Cuando se conviertan en ovejas?

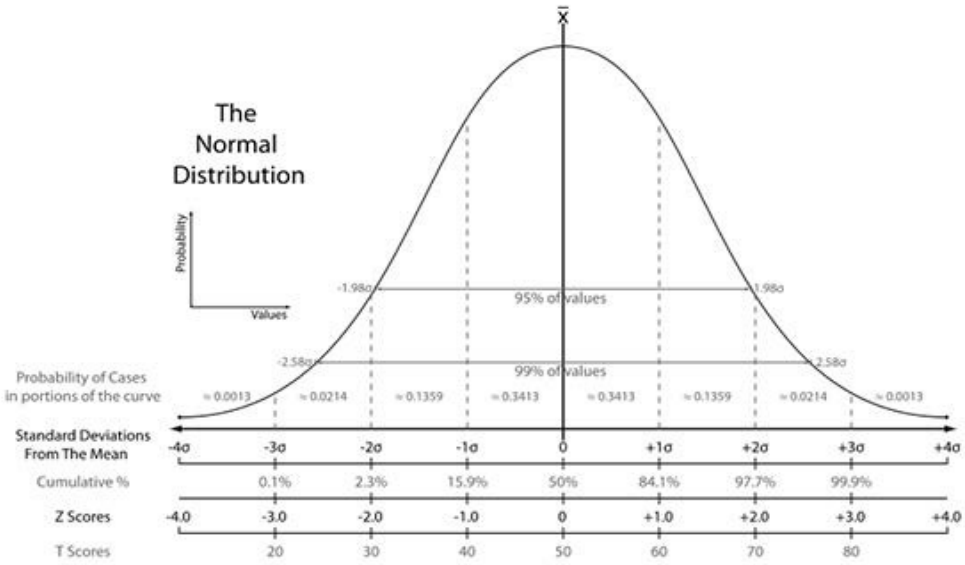
2.1. Analizando datos cuantitativos

Investigadores cuantitativos buenos buscan mantener un nivel de control de las diferentes variables y cuidadosamente definen a la vez el alcance y tamaño de la muestra considerada. También van a intentar eliminar o aceptar la influencia de otros factores en la muestra y resaltar esto claramente en la investigación.

Uno de los aspectos más importantes en la investigación es el de obtener un resultado estadístico significante. La significancia estadística es básicamente una medida que dice que hay menos de un chance de 5% que el resultado del análisis puede ser resultado a la casualidad.

Esto se puede explicar mejor con el lanzamiento de una moneda y evaluando si una moneda esta sesgada, que se puede probar con un experimento simple de hipótesis nula. La hipótesis nula es una declaración sobre el mundo que puede plausiblemente justificar un resultado observado, por ejemplo "la moneda es justa". Entonces podrías lanzar la moneda 100 veces y si la cara solamente sale una vez, se puede determinar que se puede rechazar a la hipótesis nula.

Pero ¿qué tal si 51 de las 100 veces que la lanzas sale de cara? O si lanzas la moneda 100,000 veces y 51,000 veces sale de cara. ¿Puede ser que es el resultado de la casualidad? o ¿está la moneda sesgada?



- *[la distribución normal]*
- *[eje y – probabilidad]*
- *[eje x – valores]*
- *[probabilidad de casos en porciones de la curva]*
- *[deviaciones estándares del promedio]*
- *[% cumulativo]*
- *[z-scores]*
- *[t-scores]*

Gráfica 1: ***Distribución normal y medidas de significancia***

Calculando si un resultado es significativo se puede hacer en dos maneras, o con un cálculo de significancia estadística o un cálculo de *z-score*. De cualquier forma, estamos tratando de calcular si el

resultado cae fuera de los 95% de observaciones donde la hipótesis nula es verdad, así refutándolo. Como la mayoría de los datos cuantitativos se puede mapear a una distribución normal, es esto que dicta que un resultado sea significativo estadísticamente. En la siguiente gráfica se puede ver que 95% de los valores siempre caen dentro de 1.98 deviaciones estándares del promedio ($\bar{x}$)

Un nivel de confianza de 95% significa que rechazamos la hipótesis nula si el cálculo de significancia estadística cae fuera de 95% del área de la curva normal. El *z-score* es un cálculo de cuantas deviaciones estándares de distancia esta la muestra.

Así que, con un lanzamiento de moneda con una hipótesis nula que la moneda esta justa, basado en la cantidad de tiros y resultados tenemos lo siguiente:

Moneda	Tiros	Porcentaje cara	z-score	Sesgado?
1	100	51%	0.2	No
2	100	60%	2.00	Si
3	50	60%	1.41	No significante
4	50	65%	2.12	Si

Aquí se ve que cuando el tamaño de la muestra incrementa también se puede observar a la significancia con el mismo porcentaje de tiros que salen de cara.

El tamaño de la muestra es un factor clave y ayuda eliminar la posibilidad de que otros factores tengan un efecto, como que la moneda siempre se tira de la misma manera, la misma fuerza y de la misma posición. No obstante, algunos experimentos tal vez necesitaran este tipo de control absoluto cuando la muestra es pequeña, por ejemplo, comparando a la resistencia a quebrar de las pantallas de teléfonos caros[5].

Con la mayoría de los análisis cuantitativos enfocados en los promedios, el tamaño de la muestra es importante de remover el efecto de valores atípicos en los datos. Adicionalmente es importante buscar el promedio más representativo para el conjunto de datos, como el promedio no siempre es una representación precisa de los datos si hay valores atípicos. Por ejemplo, en el discurso del estado de la unión del presidente Bush en 2008 atacó a su oposición, diciendo que sus planes para los impuestos significarían un incremento de sus impuestos de $1800 por persona. Pero esto era el valor promedio y los ingresos no están distribuidos de manera igual; las propuestas en realidad impactaron a los ganadores fuertemente, mientras que regresó dinero en los bolsillos de los que lo necesitaban.

No solamente es el tamaño de la muestra que afecta al resultado, sin embargo, hay otros factores que necesitan ser eliminados antes que un resultado se declara significante:

- **fluctuación**
  Los eventos aleatorios forman grupos, y así la fluctuación también es importante tener en mente. Al tirar la moneda, sacar una cantidad de “caras” seguidas no significa que vendrán más (a menos que la moneda esta sesgada). Lo mismo es el caso para accidentes del tránsito.

[5] http://bgr.com/2017/09/25/iphone-8-plus-vs-galaxy-note-8-drop-test-glass/

Instalando cámaras para dar multas de velocidad alta en puntos peligrosos no necesariamente esta correlacionada con una reducción subsecuente de accidentes.

- **Metas**
  Aunque la mayoría de los datos están distribuidos normalmente, las metas tienen un efecto profundo en el comportamiento de la gente. Una vez que se introduce una meta, la gente influencian su comportamiento para conformar su meta. Por ejemplo, un hospital en los EEUU operaba solamente en pacientes saludables para que pudieran obtener fondos para ser el hospital más exitoso.
- **Correlación**
  Sólo porque aparece una correlación no significa que hay una. Un ejemplo famoso de correlaciones dudosas era que el precio de los bananos correlacionaba fuertemente con la cantidad de gente que murieron por estar enredados en sus sábanas. O la correlación de 99% entre los gastos de los EEUU en la ciencia, tecnología y espacio y con los suicidios por ahorco.
- **Los Porcentajes**
  Los porcentajes pueden hacer que números pequeños se vean grandes y frecuentemente se usa así. Por ejemplo, el año pasado hubo un incremento de casos de cáncer relacionados con teléfonos móviles: creció de 1 caso a 2. El otro problema con los porcentajes es que muchos se confunde *puntos de porcentajes* con los porcentajes. Por ejemplo, considera que cuando el Impuesto de Valor Agregado en el Reino Unido subió 2.5% (2.5 puntos de porcentaje) de 17.5% a 20%; un incremento de 2.5% realmente (desde 17.5%) hubiera sido 17.9375%.
- **Datos no-normalizados**
  Otro error común es no ajustar a las cifras por variables flotantes; entonces decir que se predice que la incidencia de cáncer global se incrementaría 75 porciento por el año 2030[6] es muy alarmante. Pero eso será hasta que te des cuenta de que se calculó este porcentaje de los números brutos de personas y no ha sido ajustado por el crecimiento de la población. Otro titular común es decir que el gasto en un sector o servicio ha incrementado cada año, pero sin normalizarlo a la inflación y así revelar que el valor real en realidad se está cayendo.

No importa la técnica que se usa para analizar datos cuantitativos, ser verdaderamente objetivo puede requerir un esfuerzo grande, aun cuando tu sujeto de estudios es una simple moneda.

2.2. Analizando datos cualitativos

No todos los datos vienen en la forma de tablas estructuradas o datos geográficos ubicados con precisión. Con frecuencia los datos cuantitativos pueden ser los más difíciles para analizar y utilizar. Datos cualitativos son información que no se puede medir y como resultado están muy subjetivos (hasta color puede ser subjetivo)[7].

El objetivo principal del análisis de datos cualitativos es para reducir y hacer relevantes cantidades vastas de información, muchas veces de fuentes diferentes. El resultado de tales análisis es ofrecer una explicación, interpretación, o resumen temático de los datos. Los input en análisis cualitativo puede

[6] http://www.cancerresearchuk.org/about-us/cancer-news/news-report/2012-05-31-globalcancer-incidence-predicted-to-increase-by-75-per-cent-by-2030
[7] https://en.wikipedia.org/wiki/The_dress

tomar muchas formas incluso a transcritos de entrevistas, documentos, los blogs, encuestas, fotos, videos, etc.

El análisis de datos cualitativos es un proceso más natural para humanos quien por naturaleza intentan destilar los input a categorías y resultados claves, especialmente en reuniones o grupos de enfoque. La gente suele usar el mapeo-mental o "notas" basadas en mapas de pensamientos para ayudar en agrupar y categorizar una discusión amplia en temas claves.

El análisis de datos cualitativos debe de prestar atención a la palabra escrita en contexto, la consistencia y contradicción en perspectivas, frecuencia e intensidad de comentarios, y su especificidad igual a temas y tendencias emergentes.

Hay dos maneras principales de analizar datos cualitativos, el análisis de marco y análisis temático de redes.

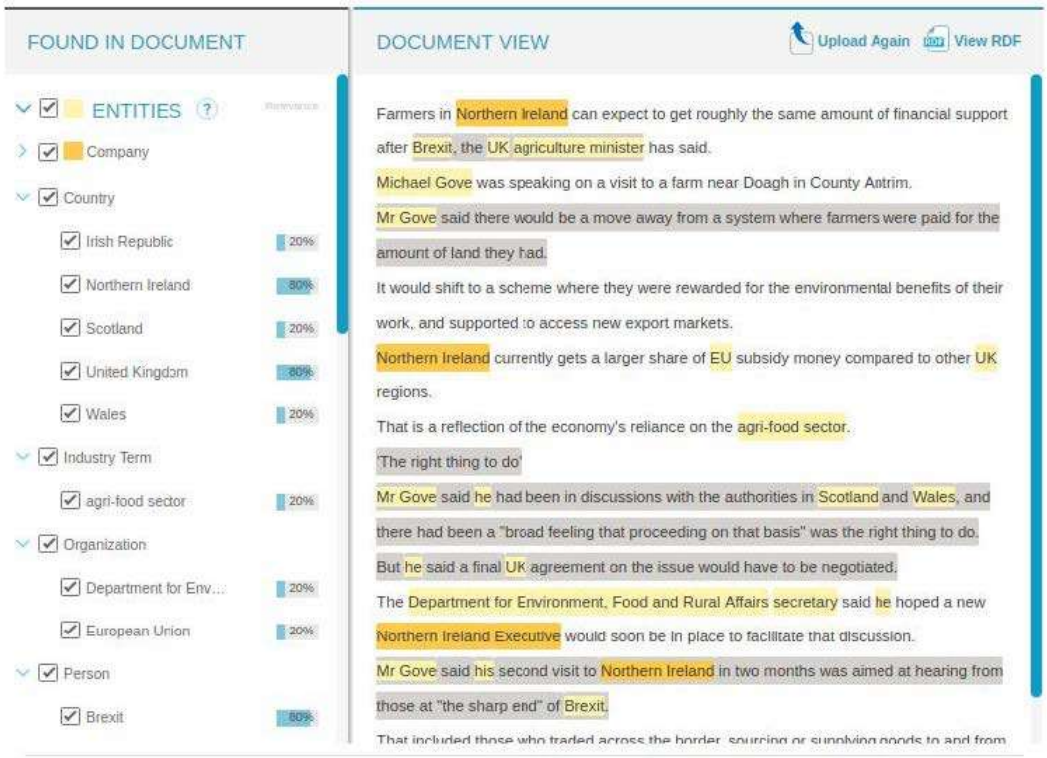
El análisis de marco se trata de crear un conjunto de criterios predefinidos que claramente reflejan a sus objetivos, metas, e intereses. Usando este conjunto de criterios, se puede extraer a las piezas relevantes de información y compararlas con otros input dentro del marco. Utilizando un marco permite que varios investigadores hagan la extracción y minimiza las oportunidades de parcialidad en el análisis cualitativo. Se puede introducir una parcialidad en la etapa de diseño del marco que puede resultar en la falta de información clave.

Una manera alternativa es no construir un marco, sino aplicando un análisis temático de redes. *El análisis temático de redes* es una estrategia más exploratoria que anima el análisis de todos los datos del input que puede influenciar la forma del resultado en direcciones inesperadas. En realidad, la mayoría del análisis de datos cualitativos va a involucrar una combinación de los dos métodos.

Sin importar el método que seleccione, el primer paso en cualquier análisis de datos cualitativos se trata con conocer bien los datos; leyendo y releyendo las respuestas. A la misma vez es una buena idea empezar a codificar los datos por anotar palabras claves y temas que tratan de filtrar e interpretar los datos. El resultado del proceso de codificación puede ser el análisis temático de redes o un marco donde todas las respuestas necesitan ser codificados. De todos modos, la codificación puede ser un proceso largo y repetitivo, pero hay un número de herramientas que pueden ayudar con el análisis temático de redes.

2.2.1. Herramientas de análisis temático de redes

Herramientas para el reconocimiento de entidades provee un ejemplo de una técnica que puede ayudar a analizar y enriquecer los datos cualitativos. Básicamente el reconocimiento de entidades intenta ubicar y clasificar entidades nombrados en el texto en categorías predefinidas como nombres de individuos, organizaciones, lugares, expresiones de tiempo, cantidades, valores en dinero, porcentajes, etc.



Gráfica 2: ***El etiquetado Calais de Noticias BBC: NI Granjas "pueden esperar mismo apoyo" después del Brexit – Gove***

Calais[8] (vea la gráfica 2 arriba) de Thomson Reuters es un ejemplo de un motor de reconocimiento de entidades. Mas que reconocimiento de entidades, el servicio Calais vincula a esas entidades a registros de datos que existen dentro de la base de datos permID de Thomson Reuters de entidades y registros financieros.

De manera más general, el reconocimiento de entidades se usa por servicios como TheyWorkForYou[9] para registrar las actividades de los políticos y proveerlo en una forma fácil para el público.

Calais es un buen ejemplo de una herramienta en línea que puede realizar un reconocimiento de entidades y proveer enlaces a datos adicionales para un tema. Otras técnicas pueden ser más sencillas pero igual de efectivo, como generadores de nubes de palabras[10] (Word clouds).

[8] Open Calais - http://www.opencalais.com
[9] https://www.theyworkforyou.com
[10] https://www.jasondavies.com/wordcloud/

3. Visualización de datos

Otra manera de rápidamente interpretar datos es para visualizarlos. El cerebro humano es mucho más adepto a consumir y entender datos presentados en una visualización que en texto.

La mayoría de las gráficas que se usa en la visualización de datos se deriva de los diseños originales de William Playfair (1759-1823), un economista político. Playfair inventó varios tipos de gráficas: en el año 1786 la gráfica de línea, de área y de barra de datos económicos; y en 1801 la gráfica circular (*Pie Chart*), utilizado para mostrar relaciones entre partes del total.

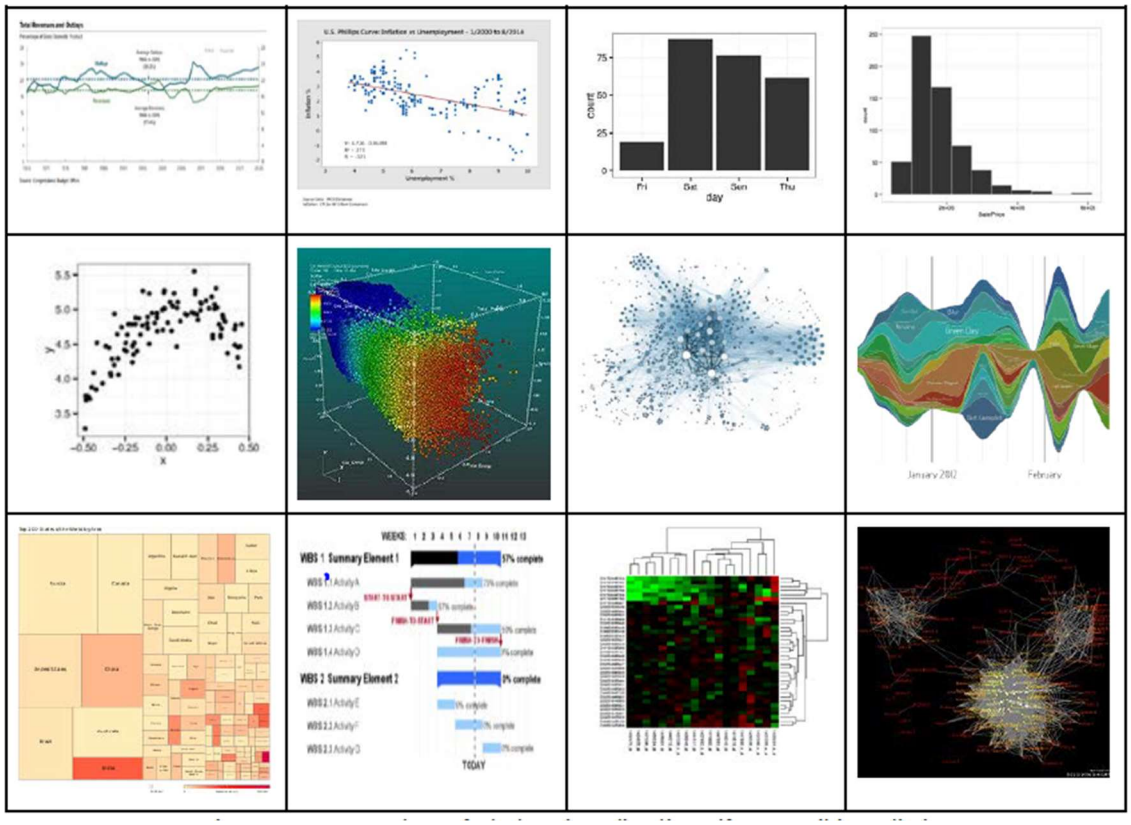
La selección de cual técnica de visualización usar depende en el objetivo de la visualización y el tipo de datos para visualizar. Esta sección explora los diferentes objetivos y los tipos de visualizaciones adecuados para cada uno con ejemplos.

3.1. Las metas en visualizar datos

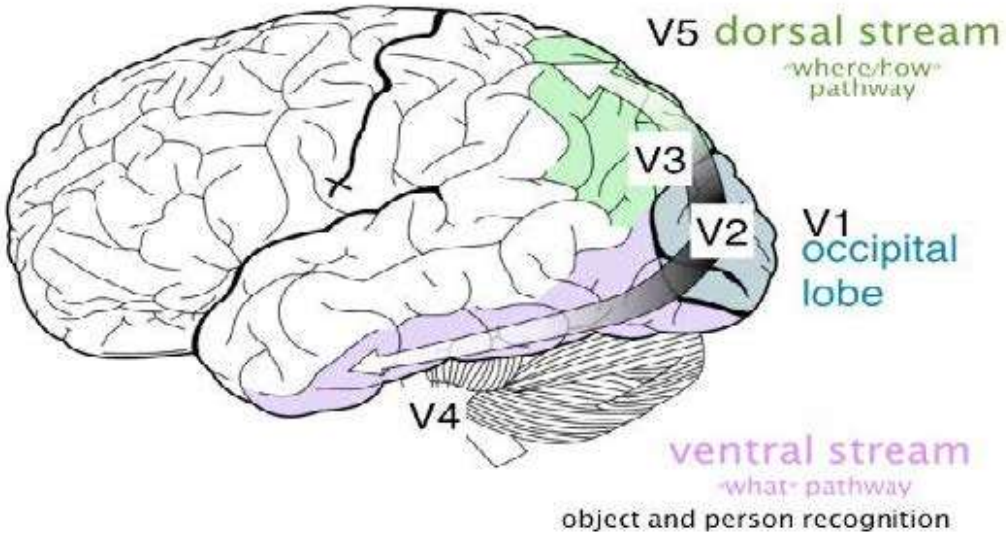
La meta de visualización es para comunicar información clara y eficientemente a usuarios.

Según Friedman (2008) el "objetivo principal de la visualización de datos es comunicar a información clara y efectivamente por manera gráfica. No quiere decir que la visualización debe de ser aburrido para ser útil o muy sofisticado para ser bonito. Para transmitir las ideas efectivamente, ambos la forma estética tiene que ir de la mano con la funcionalidad, proveyendo ideas de un conjunto complejo de datos amplios y dispersos y comunicar sus aspectos claves de una manera más intuitiva. Pero los diseñadores muchas veces fallan en lograr realizar un balance entre la forma y la funcionalidad, creando visualizaciones hermosas de datos que no sirven en su propósito principal – para comunicar información".

Uno de los problemas principales con la visualización de datos es que es un tema amplio para cubrir una gama de visualizaciones diferentes que están diseñados para propósitos diferentes. La comunicación es solamente un objetivo de una visualización de datos. Las visualizaciones de datos también se pueden utilizar durante la etapa de análisis de datos para poder entender a los datos y guiar al análisis.



Gráfica 3: ***Ejemplos de visualización (de Wikipedia)***



Gráfica 4: ***Caminos en la corteza***

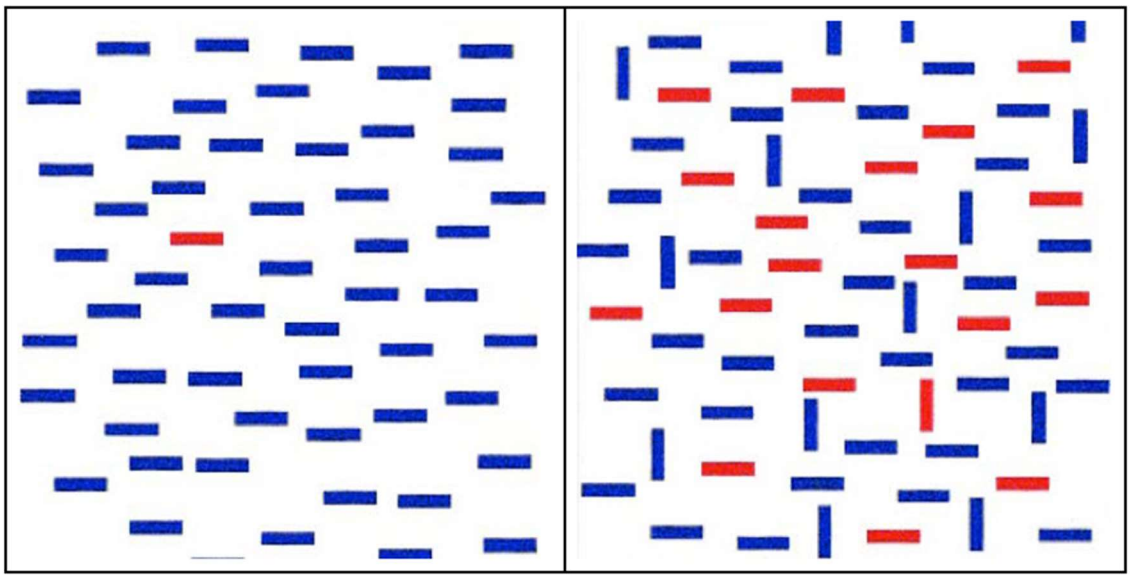
Considera los siguientes ejemplos de la página de visualización de datos de Wikipedia: posiblemente, sólo las gráficas de barra (arriba a la derecha y treemap (mapa de árbol, abajo a la izquierda) son buenas visualizaciones para comunicación instantánea de datos. Esto es porque ambos de estas gráficas usan una técnica de visualización que se llama *"pop out"* (sobresalir).

El *'pop out'* ayuda a dirigir el ojo al área correcta instantáneamente y el ojo humano esta atraído a los colores más brillantes, cosas más grandes, y que sobresalen por una diferencia. Es una característica que está programado en la corteza visual. Dentro de la corteza visual hay dos corrientes, el corriente ventral (que) y la corriente dorsal (dónde/cómo). Es la corriente dorsal que procesa la información del ojo sobre nuestro ambiente en tiempo actual para que podamos reaccionar instantáneamente en situaciones de riesgo (por ejemplo, un objeto tirado hacia ti).

Es el corriente dorsal que busca donde están las cosas y como relacionan con otras cosas que hace que el '*pop out*' funciona tan rápido. Al contrario, la corriente ventral es responsable por descifrar "qué" es la cosa. Este es un proceso mucho más lento y es la razón porque tal vez reconoces la cara de una persona, pero no recordar el nombre.

Las mejores visualizaciones de datos para la comunicación apelan a la corriente dorsal y hacen que la información sobresalga. Si una visualización se requiere el uso de la corriente ventral para ayudar a contextualizar la información, entonces hay una probabilidad alta de que individuos lo interpreten diferente.

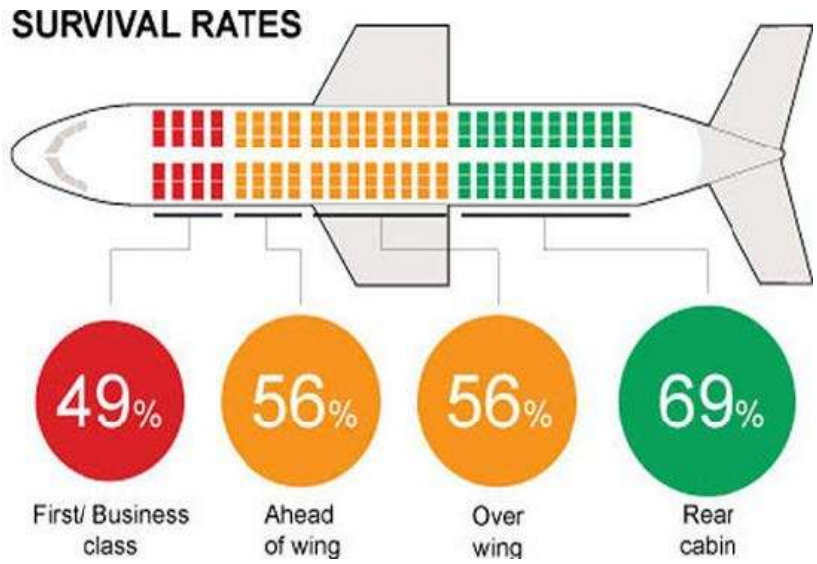
Inténtalo tú con el juego del '*pop out*' abajo (Gráfica 5). Lo único que tienes que hacer es encontrar el atípico en las dos imágenes abajo.



Gráfica 5: *Encontrando el atípico*

¿Cuál era más fácil encontrar?

Otro aspecto importante de la visualización es el uso de color. En adición de hacer que la información "sobresalga", el color puede también transmitir significado. Por ejemplo, el rojo, amarillo, y verde representan peligro, cautela, y seguridad, como en un semáforo. (vea Gráfica 6).



[Tasas de sobrevivencia]

[clase primera/ejecutiva] [delante de las alas] [sobre las alas] [cabina trasera]

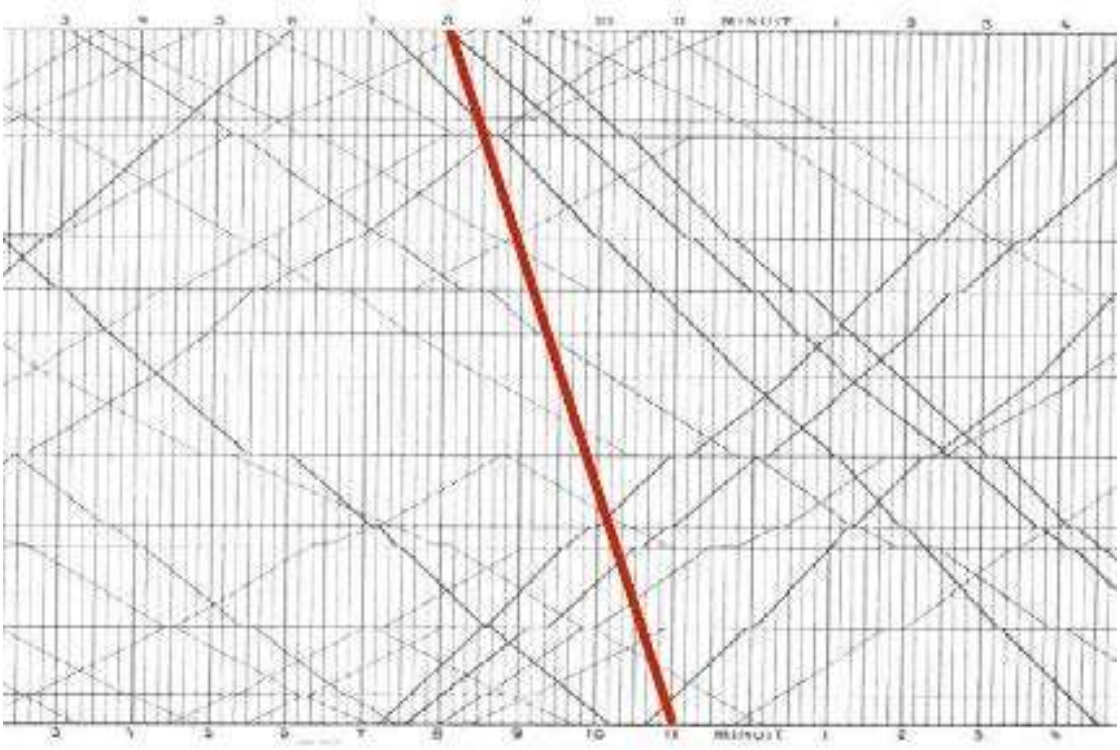
Gráfica 6: ***Colores del semáforo indicando seguridad***

Sin embargo, no todos los colores transmiten fácilmente un significado fácil de reconocer, y el uso de demasiados colores se puede confundir. De esta manera, trate de minimizar el uso de colores.

"pop-out" funciona mejor cuando se usa una combinación de pistas visuales en combinación, como por ejemplo el color, tamaño, y grosor.

En la Gráfica 7 una combinación de color y grosor conduce el ojo a la línea más empinada. Esta línea representa el primer tren de alta-velocidad desde Paris a Lyon. Los escarpados de la línea indica velocidad.

En cualquier visualización, es esencial remover cualquier información extraña que detrae el mensaje que intenta transmitir.



Gráfica 7: ***'Pop-out' y otras pistas visuales***

3.2. Escogiendo la visualización adecuada para datos

Escoger la visualización adecuada para los datos depende de dos aspectos claves.

1. El tipo de datos
2. El mensaje que quiere transmitir

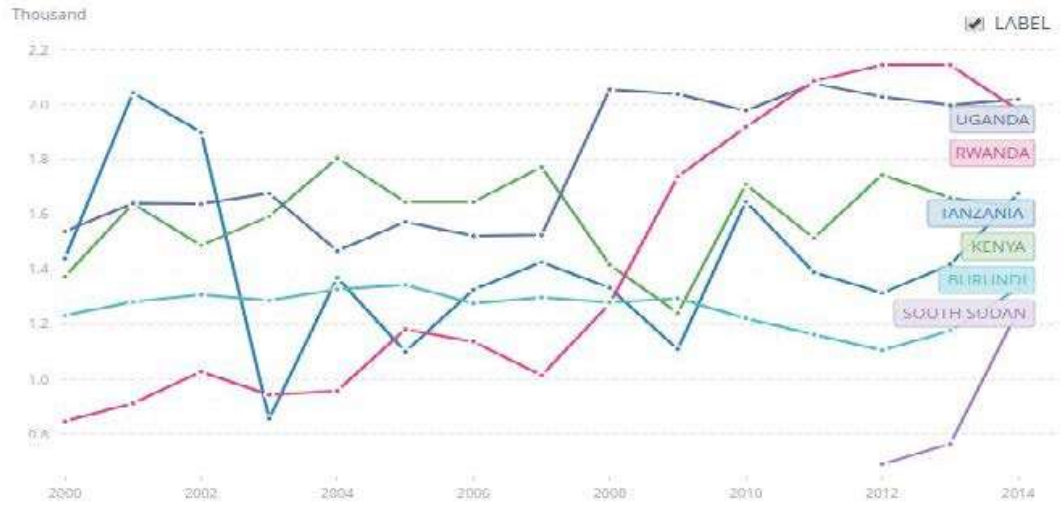
Hay tres tipos da datos principales destacados en la Tabla 2.

Tabla 2: ***Tipos de datos***

Tipo de datos	descripción	Ejemplo de visualización
Serie de tiempo	Observaciones de los mismos objetos en el tiempo	Gráfica de línea, de movimiento, diagrama de área polar, Gráfica Gantt, Gráfica de barra
Población	Observaciones de diferentes objetos de un punto en tiempo singular	Gráfica de barra, mapa, treemap, Gráfica círculo (pie chart)
multivariado	Observaciones de diferentes objetos en puntos de tiempo diferente	Gráfica de movimiento multidimensional, Gráfica de barra, treemap

Considera por ejemplo datos de rendimiento de cereales en la Comunidad Africana del Este de Tanzania, Rwanda, Uganda, Kenya, Burundi, y Sudan del Sur, que está disponible como datos abiertos del Banco Mundial[11]. Se puede visualizar a los datos en varias maneras por su naturaleza multivariado.

Los datos disponibles del Banco Mundial están disponibles para cada país desde el año 1961. Si los datos se trataban de un sólo país tendríamos un conjunto de datos de serie de tiempo. Si los datos eran solamente para un año entonces sería un conjunto de datos de población, pero los dos están disponibles, así que es un conjunto de datos multivariados.

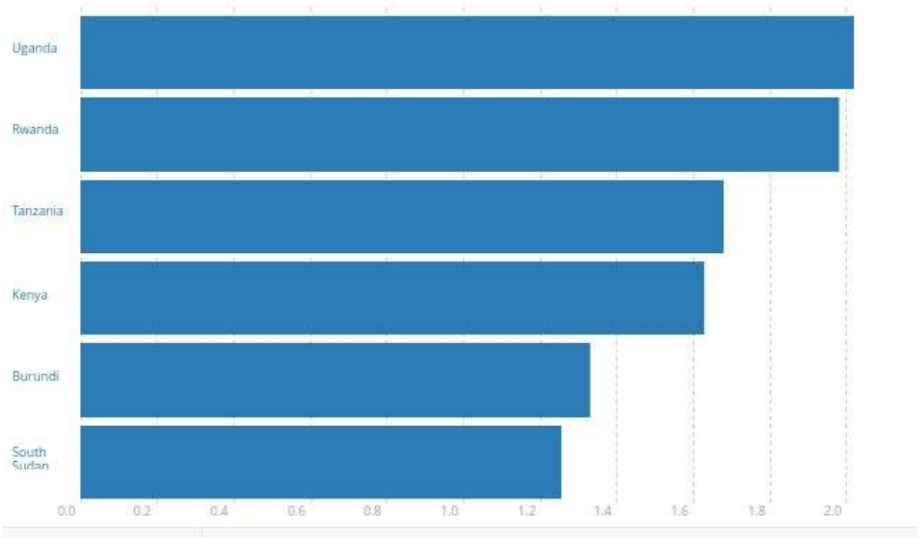


Gráfica 8: ***Visualización de serie de tiempo del rendimiento de cereales en la Comunidad Africana del Este desde el año 2000***

La gráfica de línea en la gráfica 8 arriba es naturalmente adecuado para mostrar una serie de tiempo para los que leen de la izquierda a derecha. También permite una comparación relativamente fácil entre países en diferentes puntos en tiempo. Nota como el rosado más oscuro para Rwanda sobresale, mientras que los tonos más claros de azul se mezclan. El azul es un color muy neutral, donde los rojos y amarillos son mucho más adecuado para hacer que la información sobresalga. Significativamente, esto guía al ojo para ver que Rwanda ha mejorado en comparación a los otros países de la CAE desde el año 2000.

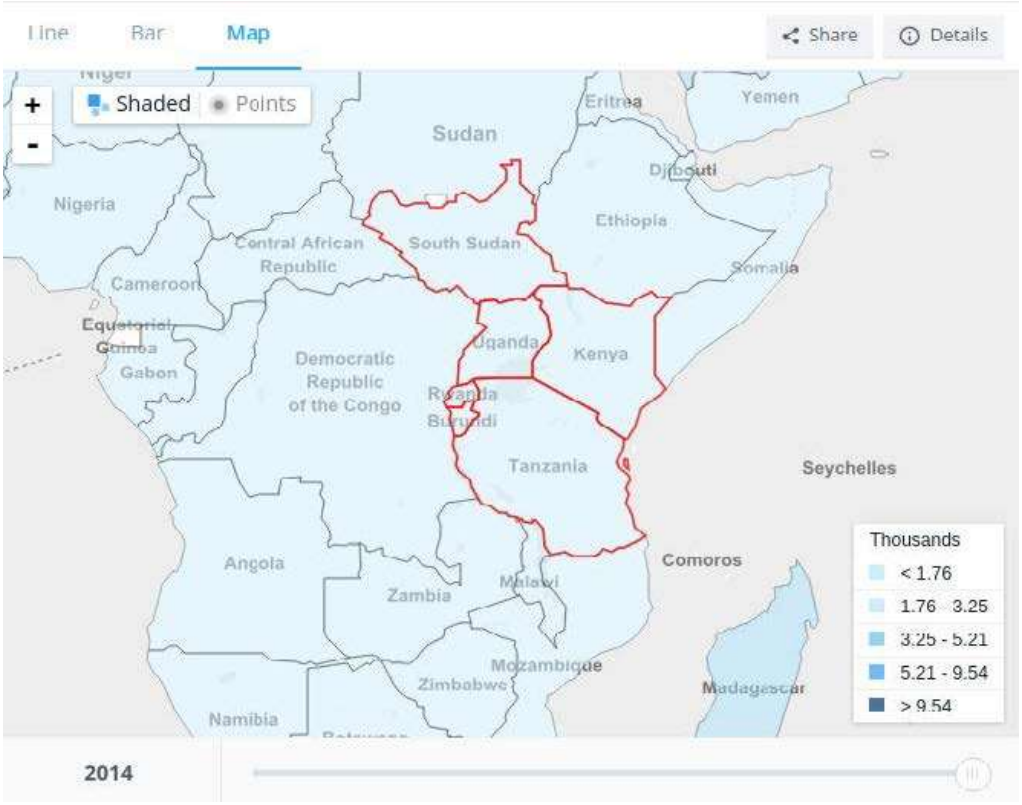
Al contrario, la gráfica de barra (gráfica 9) compara a los países en un punto particular en tiempo (2014 en este caso). Usando el mismo color aquí significa que el ojo no está atraído a un país en particular, pero está enfocado en el tamaño de las barras. Tal vez lo que sobresale aquí son los tres grupos de Uganda con Rwanda, Tanzania con Kenya, y Burundi con Sudan del Sur. Sin embargo, la gráfica de línea de serie de tiempo nos dice que estos grupos tal vez no son significantes en el tiempo.

[11] World Bank - Cereal yield (kg per hectare) - https://data.worldbank.org/indicator/AG.YLD.CREL.KG



Gráfica 9: ***Visualización de gráfica de barra del rendimiento de cereales en la Comunidad Africana del Este en 2014***

Para explorar las agrupaciones más profundas podríamos mirar otra visualización de un conjunto de datos, esta vez un mapa. Como tenemos países es lógico mostrarlos en un mapa (gráfica 10).



Gráfica 10: ***Visualización de conjunto de datos geográficos de población del rendimiento de cereales en la Comunidad Africana del Este en el año 2014***

Desafortunadamente la visualización del mapa del Banco Mundial no filtra a los países solamente a los que has seleccionado. En una escala global los países de la CAE son muy similares en la producción de cosechas, no como lo demuestra en las gráficas de barra y serie de tiempo.

Para revelar las diferencias sutiles de los datos de estos países se puede usar otra herramienta de mapeo (carto.com en este caso) para crear su propio mapa (como en la Gráfica 11).

Gráfica 11: ***Países de la Comunidad Africana del Este en el mapeo alternativo***

Este último grupo de visualizaciones demostró como remover datos puede cambiar el mensaje que se transmite. Idealmente la nueva visualización quitaría a los otros países, porque el color blanco se puede confundir y no necesitas un mapa del mundo entero.

Resumen

El análisis de datos y visualización es un tema difícil y complejo. Esto es especialmente el caso de datos de naturaleza multivariada y el hecho de que muchos conjuntos de datos ahora son “grandes”, con el significado de que los problemas de datos multivariados están multiplicados.

Escoger la técnica de análisis correcta está fuertemente ligada con la técnica de colección de datos para asegurar que el sesgo se limita donde pueda. Se debe dar una consideración cuidadosa a la colección de datos y como el análisis puede ser influenciado por factores externos como metas de política.

La visualización de datos debe evitar introducir el sesgo u otros aspectos que pueden engañar a la audiencia. Dado que el ojo procesará la visualización en una fracción de un segundo, la visualización debe estar diseñado cuidadosamente para asegurar que el mensaje clave se transmite correctamente. ¿Por qué 78 y 103 están mucho más cerca que 71 y 78?

Referencias

Friedman, V. 2008. 'Data Visualization and Infographics'. *Smashing Magazine*, 14 January, 2008. Available at: https://www.smashingmagazine.com/2008/01/monday-inspiration-datavisualization-and-infographics/#top

Huberman, M. and Miles, M.B., 2002. *The Qualitative Researcher's Companion*. Sage, Thousand Oaks, CA, USA. 410 pp.