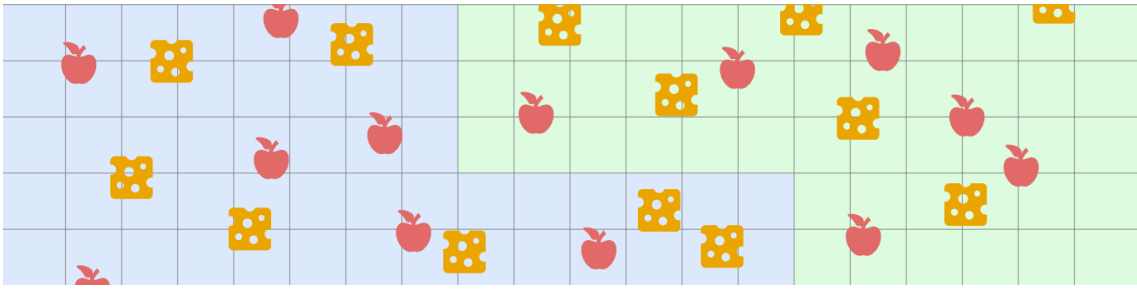
cooking skills

comparando lo incomparable

sobre cómo comparar datos para una visualización

ph.d. franklin hernández-castro

instituto tecnológico de costa rica



v. 2S.2023

introducción

este documento resume algunos casos típicos que se encuentran cuando un diseñador desea comparar datos para una visualización.

esta hecho para mis estudiantes del curso de visualización de datos de la escuela de ingeniería en diseño industrial, del tecnológico de costa rica.

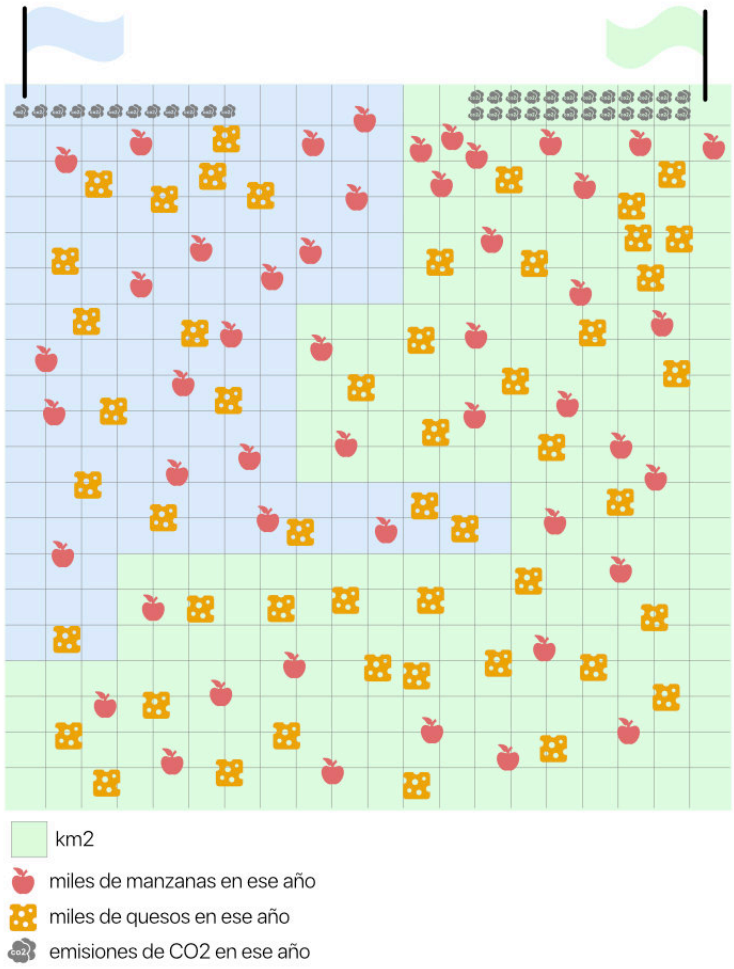
generalidades

cuando un diseñador desea hacer una visualización con el fin de sacar información de los datos a menudo se encuentra comparando datos que vienen de muchas fuentes distintas, por supuesto en formatos y unidades distintas.

los casos más comunes son solucionados a través del típico parsing de datos (más información en [1]). sin embargo, hay muchos casos que van más allá de hacer el parsing para saber qué datos se tienen y cuáles no, sino que tiene que ver más con comparar "mangos con chayotes" :), veamos.

caso de estudio

como es habitual voy a tratar de explicar todo partiendo de un ejemplo. en este ejemplo tenemos dos países (el celeste y el verde).



estos dos países tiene un mapa que se ve en la figura y producen manzanas, queso, carne, huevos y papas. así que uno podría empezar a preguntarse cosas acerca de estos dos países y sus industrias.

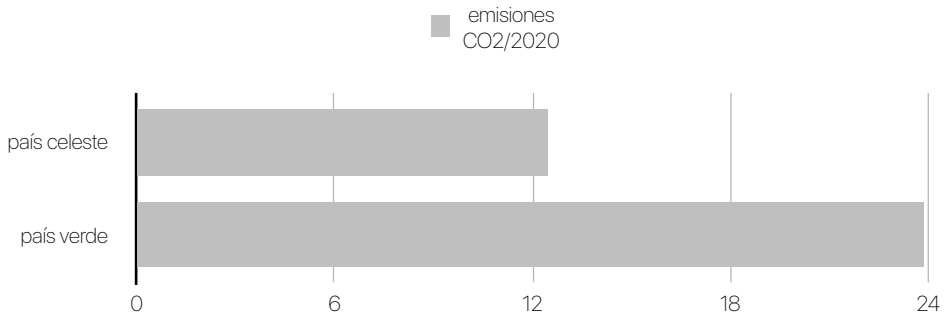
1 datos absolutos

las primeras preguntas básicas que nos podemos hacer serían cosas como: quién produce más manzanas?, quién produce más quesos?, o quién produce más emisiones de CO2?

a juzgar por lo que nos dice el mapa tendríamos los siguientes datos:

	manzanas (M)	queso (M)	emisiones CO2/2020
país celeste	18	16	12.5
país verde	31	34	23.9

de donde queda claro que el país verde produce casi el doble de productos que los celestes. por otro lado, los verdes también son el país más contaminante en emisiones de CO2 pues también producen casi el doble de este gas que los celestes; éstas afirmaciones son datos no información, pues se extraen directamente del gráfico con solo "contar".



sin embargo, alguien podría argumentar que la comparación no es justa pues el tamaño (área geográfica) de los países es muy distinta, así como su población, y por lo tanto era obvio que uno iba a contaminar más que otro y que si se quiere ser justo deberíamos de tomar en cuenta esas diferencias obvias. veamos entonces otros datos al respecto:

	área Km2	población (M)
país celeste	143	32.00
país verde	257	75.00

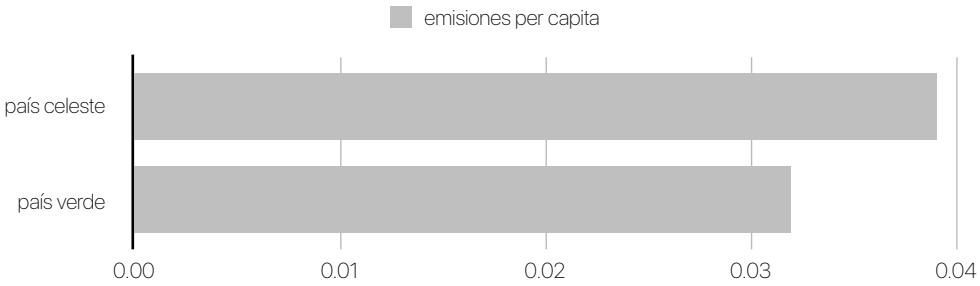
ah, ahora sabemos que el país verde tiene casi el doble de kilómetros cuadrados que el celeste y que lo mismo sucede con la cantidad de cantidad de habitantes.

2 datos relativos

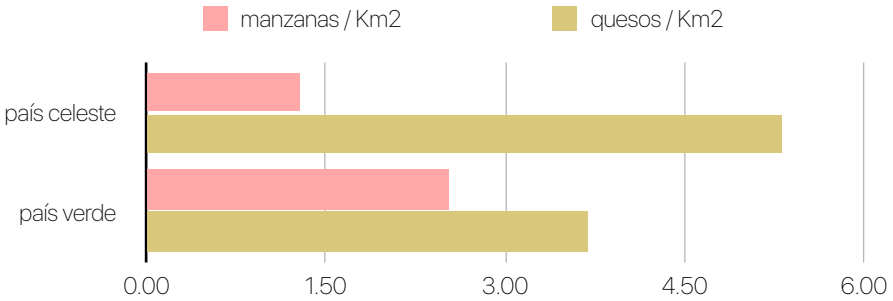
así que sería más justo saber cuántas emisiones se generan "per capita" en cada país o cuánto CO2 se genera por unidad de manzanas o quesos (que en nuestro ejemplos están en miles y se supone que todos los datos son del mismo año) se producen por kilómetro cuadrado. en estos casos pues simplemente se dividen unos datos entre los otros y listo. por ejemplo cantidad total de emisiones entre cantidad de habitantes, pues nos da la emisiones per capita, lo mismo podríamos hacer con las manzanas y/p quesos y los kilómetros cuadrados.

en el caso de las emisiones se maneja generalmente por población y dependiendo de las dimensiones de los datos se usa por ejemplo por cada 100,000 habitantes, esto con el fin de que los números no tengas tantos ceros decimales y sean más fáciles de comparar (no tengo que aclarar que todos estos datos no son más que fantasía y que no tiene relación con nada de la realidad)

	área Km2	población (M)	emisiones CO2	emisiones por 100k habitantes	manzana (M)	área cultivada manzana	manzana / Km2	queso (M)	área usada queso	quesos / Km2
país celeste	143	32.00	12.5	**0.04**	18	23	**1.28**	16	85	**5.31**
país verde	257	75.00	23.9	**0.03**	31	78	**2.52**	34	125	**3.68**



ahora el panorama ha cambiado, los habitantes celestes tiene más gasto de emisiones de CO2 que los verdes cosa que parecía entes lo contrario. este es un caso que no es atípico, siempre hay que tener mucho cuidado qué datos se comparan y en qué condiciones, todos los días veo en las noticias como la gente compara cosas de forma equivocada y llega a conclusiones igualmente erróneas, cosa que en una democracia es muy peligroso pues la gente termina convencida de cosas que no son ciertas.



con los nuevos datos también vemos que con respecto al uso del terreno, el cultivo de manzanas en los celestes es más intensivo que en los verdes, en el caso de la producción de quesos es el contrario.

de aquí empezamos a saber algo de información y de paso se abren nuevas preguntas: es mejor usar intensivamente el terreno?, de modo de usar menos área para la misma producción, o como podría ser en el caso de los quesos, es mejor usar más terreno pues de algún modo insinúa que los animales viven en mejores condiciones?

el panorama va quedando más claro, y es menos concluyente; lo que nos debería de parecer normal pues rara vez la realidad es blanco o negro, a menudo hay tonos de gris.

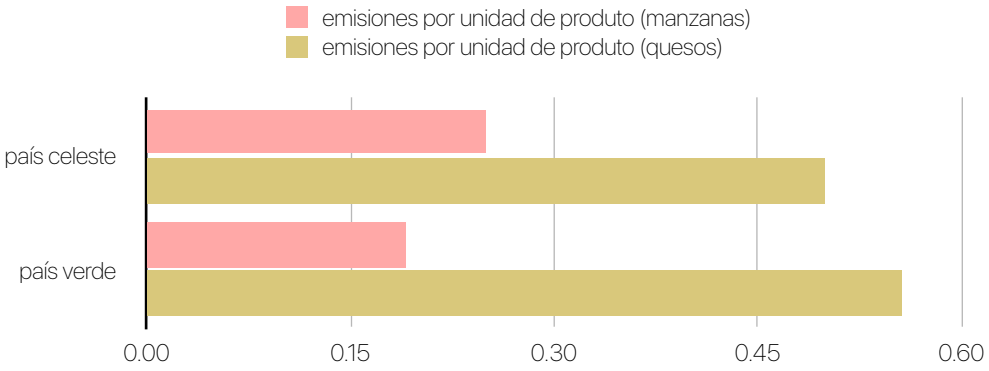
3 información

ahora pasemos a una etapa en la que tratamos de concluir más información, por ejemplo una profundización lógica sería tratar de saber cuánto es el "costo" en emisiones que tiene cada producto, es decir cuánto CO2 se emite para producir mil manzanas, o mil quesos, en cada país, con esto podríamos ver cuál país tiene enfoques de producción más contaminantes en términos de ese gas; veamos:

	cantidad manzanas (M)	emisiones CO2/2020 manzanas	parte GDP 2020 manzanas	producción queso (M)	emisiones CO2/2020 quesos	parte GDP 2020 quesos
país celeste	18	4.5	45	16	8.0	76
país verde	31	5.9	62	34	18.9	146

teniendo datos como los de la tabla anterior (que en realidad no son difíciles de encontrar, cada país tiene los suyos) ya podemos tratar de ir más cerca y sacar un "costo" en emisiones por "unidad" de producción en cada producto.

	cantidad manzanas (M)	emisiones CO2 manzanas	emisiones por unidad de produto	parte GDP 2020	cantidad queso (M)	emisiones CO2	emisiones por unidad de produto	parte GDP 2020 quesos
país celeste	18	4.5	**0.25**	45	16	8.0	**0.50**	76
país verde	31	5.9	**0.19**	62	34	18.9	**0.56**	146



ahora tenemos más información que hemos venido extrayendo de los datos, por ejemplo, definitivamente la unidad de manzana (miles) es mucho menos costosa en términos de emisiones de CO2 que la misma cantidad en términos de quesos.

además vemos claramente que el país celeste logra producir quesos menos contaminantes por unidad que el verde, pero que en el caso de las manzanas es lo contrario, el país verde los hace con menos emisiones por unidad.

esta información extraída de los datos ya nos dice cosas que definitivamente no estaban claras al inicio y que bien puede motivar un estudio más profundo de estas conclusiones preliminares.

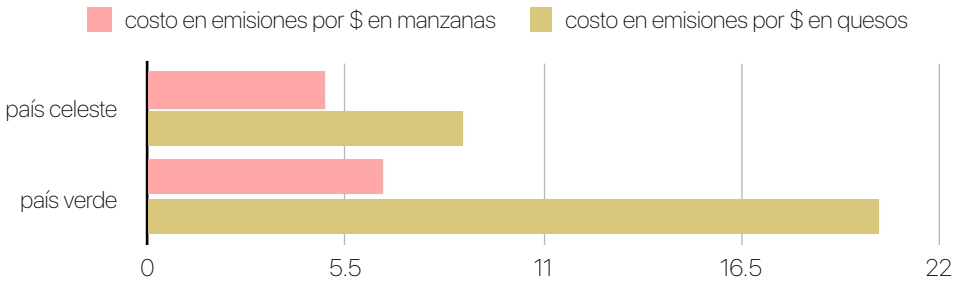
comparando lo incomparable

comparación en una unidad compartida

en nuestro ejemplo alguien podría argumentar que producir mil manzanas es muy distinto que producir mil quesos, así que no es justo compara una cosa con la otra como si fuera lo mismo, en estos casos sería bueno encontrar una unidad que los comprare en forma más imparcial, y como tenemos el aporte de cada producto al PIB podríamos usarlo.

para esta idea lo que haríamos es sacar el "costo en emisiones de CO2" de cada "dólar producido en cada tipo de producto", con este cálculo tendríamos una buena unidad sobre el costo ecológico de cada dólar que se produce por "industria". por supuesto para hacer esto debemos de dividir la cantidad de dólares producidos por sector entre las emisiones de CO2, veamos:

	emisiones CO2/2020 manzanas	parte GDP ($) 2020 manzanas	costo en emisiones por $ manzanas	emisiones CO2/2020 quesos	parte GDP ($) 2020 quesos	costo en emisiones por $ quesos
país celeste	4.5	45	**4.95**	8.0	76	**8.76**
país verde	5.9	62	**6.52**	18.9	146	**20.36**



esta es una unidad que nos acabamos de inventar claro, pero finalmente nos dice claramente cuál industria produce más CO2 por una unidad comparable entre ellas y una unidad que además es importante para las poblaciones: el dólar producido. de aquí podríamos concluir varias cosas:

- el país verde produce mucho más CO2 por dólar generado en todos los productos
- la producción de queso en los países es muy distinta pues su peso ecológico también varía mucho entre los dos países.
- quizás los celestes venda queso con "más valor agregado" pues por dólar logran hacerlo con mucho menos emisiones.

como se ve en este simple y ficticio ejemplo, las posibilidades de sacar información valiosa de datos aparentemente crudos son muchas y un análisis inteligente de los mismos puede ser muy útil.

comparación de unidades distintas

motivados por las conclusiones anteriores podría ser que quisiéramos comparar varios productos para saber cuál es su impacto en el medio ambiente.

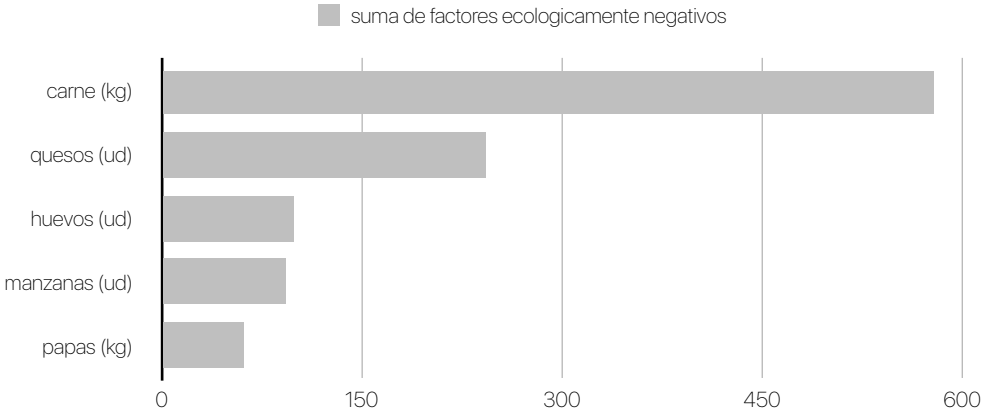
para esto nos hemos propuesto tomar en cuenta cinco factores y compararlos según su "gasto ecológico" por dólar producido:

- emisiones de CO2
- huella hídrica
- terreno necesario
- cantidad de fertilizantes usados
- cantidad de insecticidas usados

después de varios cálculos parecidos a los de la sección anterior tendríamos una nueva tabla como esta:

	CO2 (Kg)	huella hídrica (lit)	terreno necesario m2	fertilizantes gr.	insecticidas gr.	suma
manzanas (ud)	3.54	87	0.23	0.5	2.35	**93.62**
quesos (ud)	7.4	178	56	0.12	0.87	**242.39**
carne (kg)	20.45	324	234	0.56	0.45	**579.46**
huevos (ud)	4.39	60.56	34.8	0.02	0	**99.77**
papas (kg)	1.56	56	1.23	1.12	1.62	**61.53**

la tabla se ve muy bien y nos da una suma de los factores usados en cada tipo de producto.



según este enfoque la carne es la peor seguida de los quesos y mucho más abajo estarían los huevos y los vegetales.

normalización de datos

sin embargo, hay un problema, sin bien el uso de fertilizantes e insecticidas son clave en el impacto ecológico de la producción de cada artículo, su peso relativo en la suma de factores es muy bajo.

esto sucede por el uso de escalas que son tan diferentes, por ejemplo, la huella hídrica va de 56 a 324 de "litros de agua usados para producir un dólar" en ese artículo, mientras que los fertilizantes están en un rango de 0.02 a 0.56 pues estos están en "gramos por dólar".

así que si usáramos 10 veces más fertilizantes o insecticidas de los necesarios esto no haría mella alguna en la clasificación de "amigable ecológicamente" pues este cambio al sumarse con escalas tan grandes sería despreciable; a pesar que en realidad el aumento en el daño ecológico sería considerable.

en estos casos lo que se hace, en primera instancia, es usar una normalización de datos. lo que en palabras simples es convertir cada escala de mediciones en mediciones entre cero y uno. por ejemplo, los valores de CO2 van desde un valor mínimo de 1.56 a un máximo de 20.45 así que se convierten por una simple relación de razones o proporciones (regla de tres para los de la vieja escuela) que se conoce como "normalización máximo-mínimo):

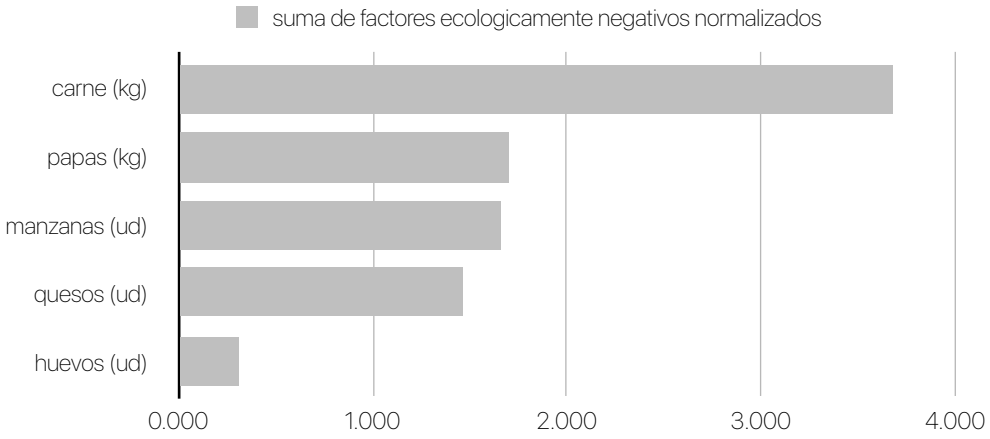
$$v = \frac{va - vmin}{vmax - vmin}$$

donde:

- v es el valor buscado
- va es el valor que se desea normalizar
- vmin es el valor mínimo en los datos
- vmax es el valor máximo en los datos

usando la formula en cada una de las cinco escalas obtenemos una tabla muy distinta a la anterior:

	CO2 (Kg)		huella hídrica (lit)		terreno necesario m2		fertilizantes gr.		insecticidas gr.		suma
carne (kg)	20.45	**1.000**	324	**1.000**	234	1	0.56	**0.491**	0.45	**0.191**	**3.682**
papas (kg)	1.56	**0.000**	56	**0.000**	1.23	0.004	1.12	**1.000**	1.62	**0.689**	**1.694**
manzanas (ud)	3.54	**0.105**	87	**0.116**	0.23	0.000	0.5	**0.436**	2.35	**1.000**	**1.657**
quesos (ud)	7.4	**0.309**	178	**0.455**	56	0.239	0.12	**0.091**	0.87	**0.370**	**1.464**
huevos (ud)	4.39	**0.150**	60.56	**0.017**	34.8	0.148	0.02	**0.000**	0	**0.000**	**0.315**



como vemos ahora todos las escalas está definidas entre 0 y 1 (o sea están normalizadas) y de ahí que el peso en la suma de factores sea el mismo entre ellas; es decir, un incremento desmesurado en "insecticidas" en un producto en específico con respecto a los otros datos o dimensiones se va a notar inmediatamente en la "evaluación" del mismo en frente de los otros.

de hecho como vemos el orden cambió, y si bien la carne sigue liderando los artículos menos amigables ecológicamente hablando, ahora lo siguen las papas y manzanas (pues el uso de agroquímicos es mayor en esos productos).

ahora todos los factores pesan los mismo y los podemos manejar como queramos, es decir, aun alguien podría decir algo como que la huella hídrica o las emisiones de CO2 son más importantes que el "terreno usado" o los "agroquímicos" y la solución sería ponderar los factores, es decir, dándoles más peso a aquellos que se creen más importantes, por ejemplo multiplicándolos por algún factor (digamos 1.5 o 2). pero a partir de una comparación equitativa de factores y no como antes a partir de escalas completamente distintas que de antemano distorsionaban enormemente el peso de uno con respecto al otro. ese ejercicio se lo dejamos a los expertos en huella ecológica que de todos modos encontrarán mi ejemplo risible por sus datos y enfoque técnico absolutamente inventado; la idea de este ejemplo es explicar la relación entre los datos, jamás hablar sobre la huella ecológica de los productos mencionados.

references

[1] Hernández-Castro, Franklin. "Dashboard design cookbook: metodología para el diseño de visualizaciones de datos para la toma de decisiones." (2021). Disponible: https://repositoriotec.tec.ac.cr/handle/2238/13281