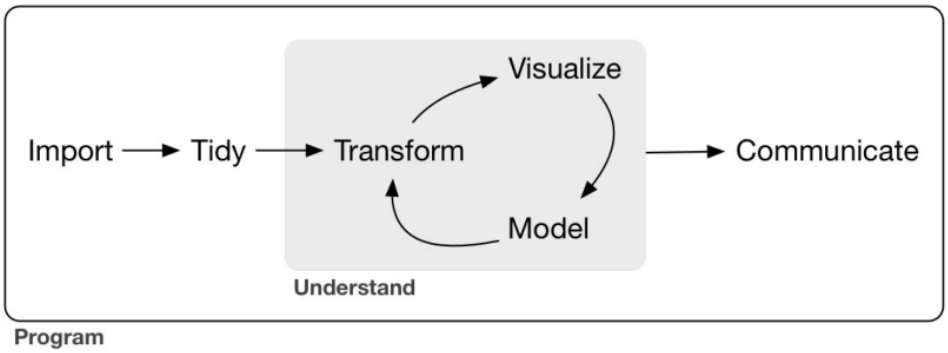
ICO 187 ANÁLISIS DE DATOS

Clase 07: ¿Qué es análisis de datos?

Año: 2021
Profesor: Sebastián Egaña

1. Introducción

Recordar que estabamos revisando los distintos pasos del proceso de análisis de datos.



Fuente: Wickham & Grolemound (2017)

En este sentido, falto analizar el apartado de visualización, debido a que en este curso no entraremos en lo relacionado con Modelación.

Por otra parte, veremos un apartado que se encuentra dentro del proceso de ordenar los datos, pero que por su complejidad debe verse de manera separada.

2. Visualización

La visualización, es en parte un arte y en parte una ciencia (Wilke, 2019).

La relevancia de la visualización se genera como una aproximación visual resumida a los datos. El problema, es que no todas las visualizaciones son buenas, en otras palabras no todos los gráficos son construidos de manera correcta.

Un buen gráfico no responde necesariamente sobre el ¿cómo se ve?, sino que debe responder a la necesidad de ¿quién lo ve? y ¿por qué lo ve?

2.1. Lo que no se debe hacer y lo que se debe hacer.

Veamos un primer ejemplo:

Fuente: Healy (2018)

- ¿Qué le parece a usted? ¿es un buen o mal gráfico?

Otro ejemplo:

Figure 1.6: "Monstrous Costs" by Nigel Holmes (1982). Also a classic of its kind.

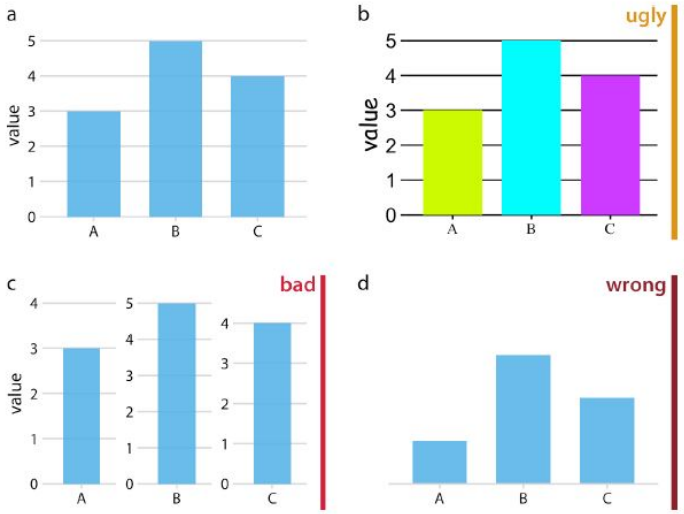
Fuente: Healy (2018)

2.2. Feo, malo y malísimo (Wilke, 2019).

Un gráfico puede tener tres errores:

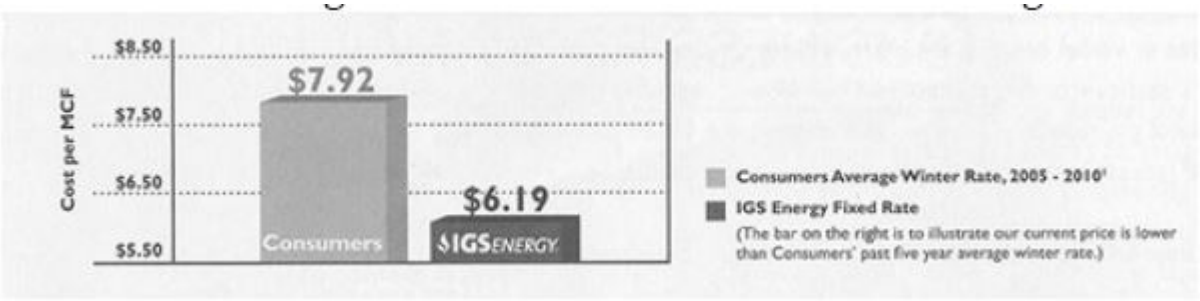
- Feo (Ugly): Esto corresponde al caso cuando falla la estética, pudiendo ser claro e informativo.
- Malo (Bad): Se da cuando existen problemas de percepción en el gráfico, pudiendo ser poco claro, consufo o engañoso.
- Malísimo (Wrong): Cuándo existen errores matemáticos detrás del gráfico.

Veamos esto de manera aplicada:



Fuente: Wilke (2019)

Un ejemplo más realista de lo antes mencionado. Este gráfico, ¿está feo, malo o malísimo?

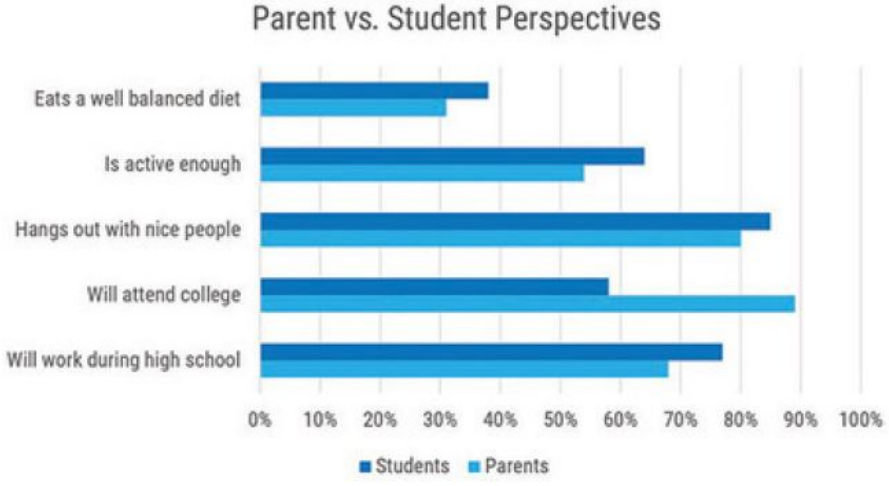


Fuente: Evergreen (2019)

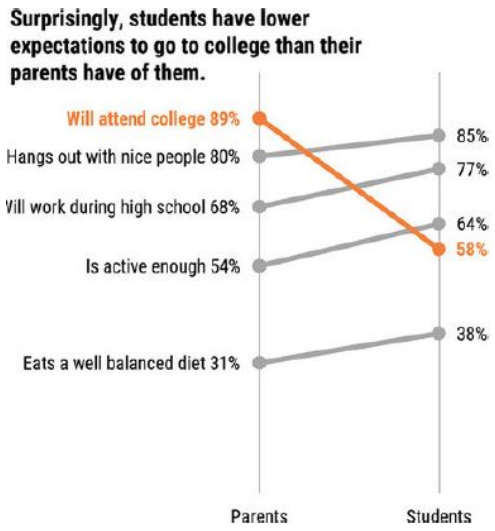
2.3. La intencionalidad detrás del gráfico.

Dos maneras distintas de graficar lo mismo: una mejor que otra.

Figure 1.1 Traditional clustered bar graphs can cloud the point.



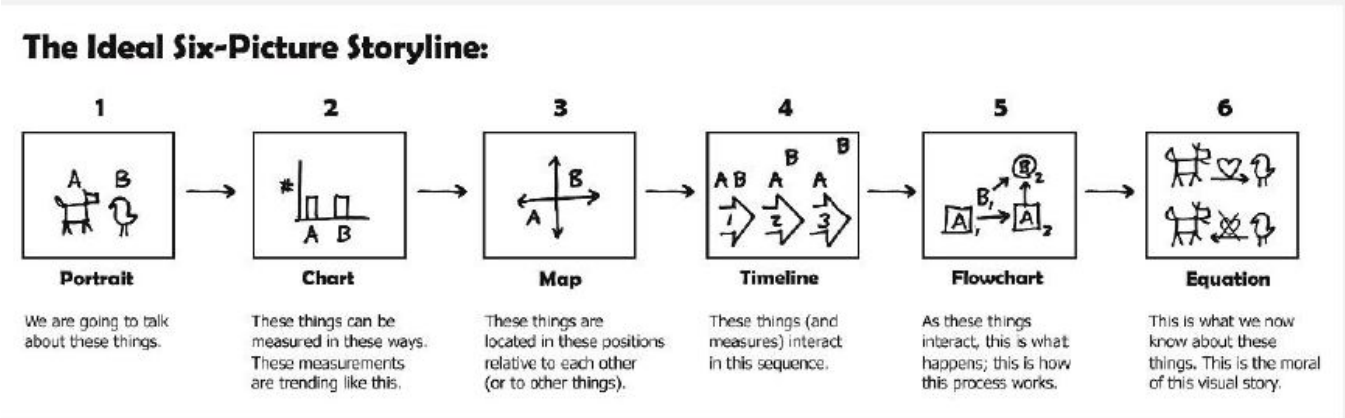
Fuente: Evergreen (2019)



Fuente: Evergreen (2019)

2.4. La visualización ideal en seis imagénes (Roam, 2016).

1. Who and what is involved (¿Quién o qué está involucrado?): Se debe iniciar con un resumen visual sobre lo que se hablará.
2. How many are involved (¿Qué cantidad está involucrada?): Se deben generar medidas cuantitativas de lo hablado. Cambios en los números también son relevantes.
3. Where the pieces are located (¿Dónde se ubica?): Presentar alguna relación visual entre lo hablado y su ubicación.
4. When things occur (¿Cuándo ocurre?): Mostrar algo relacionado con la temporalidad o la secuencia de los eventos en las que ocurren las interacciones relevantes.
5. How things impact each other (¿Cómo las cosas se relacionan?): Generar una visualización que presente la relación causa - efecto que afectan lo mostrado anteriormente.
6. WHy this matters (¿Por qué es importante?): Se debe concluir visualizado anteriormente.



Fuente: (Roam, 2016)

3. Unión de bases de datos (Joins).

Corresponde al intento de establecer relaciones entre distintas bases de datos o tablas. Esto permite el generar bases de datos más grandes, a tráves de la unión de dos relacionadas.

Las funciones que hemos revisados relacionadas con esto son las siguientes:

- Indice
- Coincidir
- Buscarv
- Buscarh
- Buscarx

A pesar de esto, nosotros buscamos aplicaciones más elaboradas de esto. Buscamos realizar uniones (joins) distintos de los anteriores, debido que esto nos obligan a añadir columnas o filas a tráves de relaciones. En el curso veremos como realizar esto utilizando Power Pivot y como realizarlo en Rstudio.

3.1. Uniones: llave primaria y foránea

En SQL, se habla principalmente de llaves en el sentido de columnas que nos permitan unir dos bases. Intentamos realizar un metodo que mantenga la integridad de la tabla a la cual añadiré información.

Una llave primaria corresponde a la columna de la base a la cual yo deseo unir otra columna, usando como punto de interacción una llave foránea.

Consideremos las dos bases, en donde A corresponde a la llave primaria y B corresponde a la base foránea de cada una de las bases:

A	B
1	3
2	4
3	5
4	6

¿Qué valores se repiten dentro de cada llave?

Una complejidad de las uniones en SQL, corresponde a los requisitos que se deben cumplir para poder realizar uniones entre bases. Por ejemplo, una particularidad es que la llave primaria no debe tener valores repetidos y esto se repite en programas como RStudio. Por otra parte, Excel al utilizar funciones de manera individual (celda por celda), no posee dicho problema.

Veamos ahora distintos tipos de uniones que pueden generarse.

- Inner Join: Refiere a la unión en base a intersecciones.

A	B
3	3
4	4

- Outer Join: Refiere a la unión de todas las columnas en A y todas las columnas en B. Las celdas sin correlativo en la otra base, generarán un valor nulo.

A	B
1	NULL
2	NULL
3	3
4	4
NULL	5
NULL	6

Esto tipo de unión, puede tener dos variaciones:

- Outer Join > Left Outer Join: Predominan las columnas en A, y las comunes en B.

A	B
1	NULL
2	NULL
3	3
4	4

En Rstudio solo se conoce como Left Join.

- Outer Join > Right Outer Join: Predominan las columnas en B, y las comunes en A.

A	B
3	3
4	4
NULL	5
NULL	6

En Rstudio solo se conoce como Right Join.

4. Avisos

Para la próxima clase debe tener activado el complemento de Power Pivot.

Revisa siempre cuál será su próxima evaluación.

5. Fechas Relevantes

Unidad	Evaluación	Ponderación	Fecha
Unidad I	Evaluación diagnóstica		25/03/2021
	Evaluación Individual Participación	(5 %)	05/04/2021
	Evaluación Grupal	(15 %)	27/04/2021 - 04/05/2021
	Evaluación Individual - Sumativa I	(30 %)	11/05/2021
Unidad II	Evaluación Formativa		13/05/2021
	Evaluación Individual Participación	(5 %)	27/05/2021
	Evaluación Grupal	(10 %)	08/06/2021 - 15/06/2021
	Evaluación Individual - Sumativa II	(15 %)	17/06/2021
Unidad III	Evaluación Formativa		22/06/2021
	Evaluación Individual Participación	(5 %)	24/06/2021
	Evaluación Individual Sesión I- Sumativa III	(15 %)	08/07/2021
	Evaluación Individual Sesión II- Sumativa III	(15 %)	13/07/2021