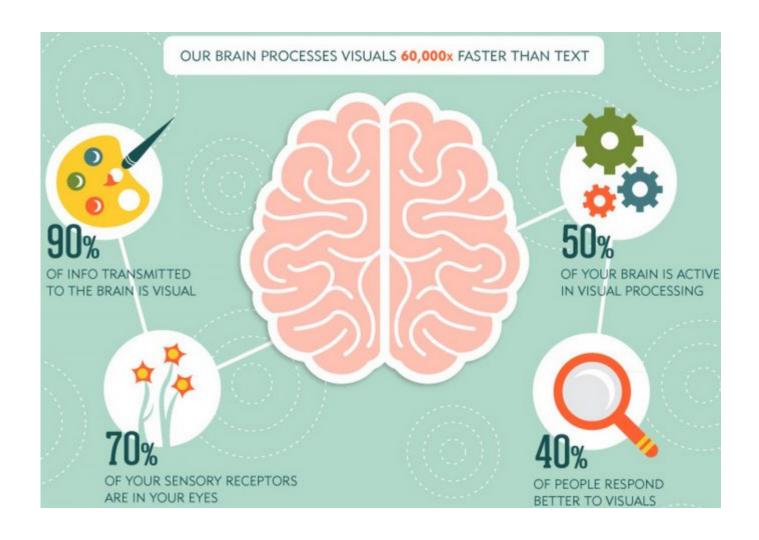
Visualización en Data Science

<u>Diplomatura CDAAyA 2018</u>

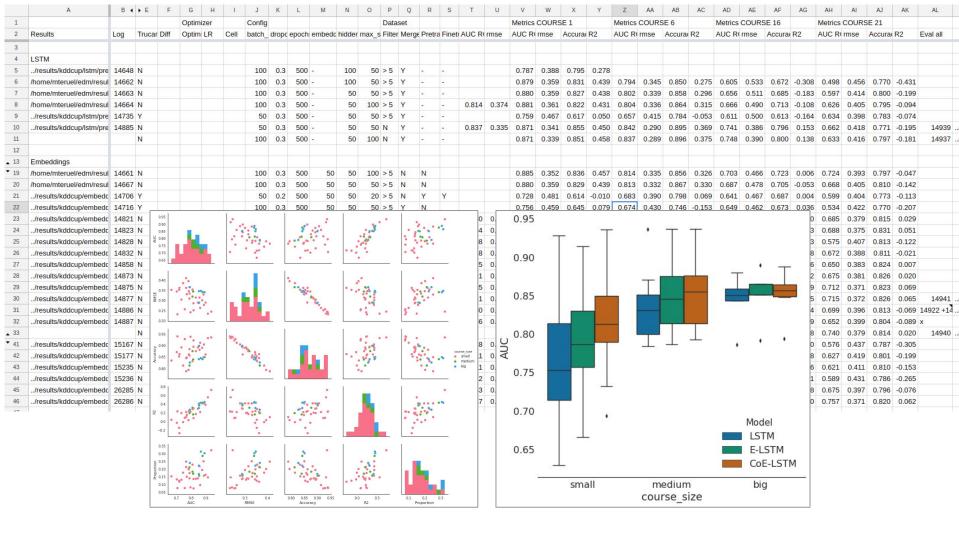


- Our brain process visuals 60000 faster than text
- 90% of the information transmitted to the brain is visual
- 70% of your sensory receptors are in your eyes
- 50% of your brain is active in visual processing
- 40% of people respond better to visuales

Herramienta para la comunicación

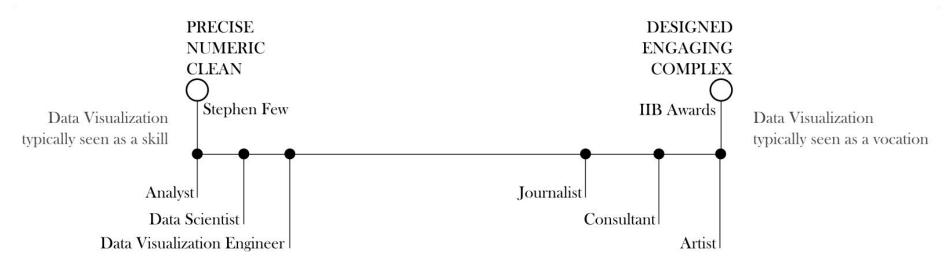
Herramienta para la compresión

//stm/pre 14648 N 100 0.3 !	ooch: embedc hidder	Dataset max_s Filter Merge Pretra Fineto		COURSE 1	Metrics COURSE 6	Metrics COURSE 16				
//stm/pre 14648 N 100 0.3 !	ooch: embedc hidder	max_s Filter Merge Pretra Fineto	AUC Remse AUC Rem	mse Accura R2		Metrics Course 16		Metrics COU	RSE 21	
					AUC R(rmse Accura(R)	2 AUC Remse Accur	racR2	AUC Ri rmse	Accurac R2	Eval all
	500 - 100	50 > 5 Y	0.787	0.388 0.795 0.27	'8					
dm/resul 14662 N 100 0.3 5	500 - 100	50 > 5 Y	0.879	0.359 0.831 0.43	9 0.794 0.345 0.850 0	0.275 0.605 0.533 0.67	2 -0.308	0.498 0.45	6 0.770 -0.43	31
dm/resul 14663 N 100 0.3	500 - 50	50 > 5 Y	0.880	0.359 0.827 0.43	8 0.802 0.339 0.858 0	0.296 0.656 0.511 0.68	35 -0.183	0.597 0.41	4 0.800 -0.19	99
	500 - 50	100 > 5 Y	0.814 0.374 0.881		1 0.804 0.336 0.864 (5 0.795 -0.09	
/lstm/pre 14735 Y 50 0.3 5	500 - 50	50 > 5 Y	0.759	0.467 0.617 0.050	0 0.657 0.415 0.784 -0					
	500 - 50	50 N Y			0 0.842 0.290 0.895				8 0.771 -0.19	
	500 - 50	100 N Y			8 0.837 0.289 0.896 0					
dm/resul 14661 N 100 0.3 5	500 50 50	100 > 5 N N	0.885	0.352 0.836 0.45	7 0.814 0.335 0.856 (0.326 0.703 0.466 0.72	3 0.006	0.724 0.39	3 0.797 -0.04	17
	500 50 50	50 > 5 N N			9 0.813 0.332 0.867 (05 -0.053		5 0.810 -0.14	
	500 50 50	20 > 5 N Y Y			.0 0.683 0.390 0.798 (
	500 50 50				9 0.674 0.430 0.746 -0					
	500 50 50				3 0.810 0.334 0.868 0					
	500 50 50				2 0.813 0.333 0.865 (5 0.831 0.05	
STREET, STREET	500 50 100			(212 22 July 2	7 0.801 0.344 0.846 (
	500 50 100				6 0.788 0.340 0.861 (
	500 50 100				0 0.806 0.338 0.854 0				3 0.824 0.00	
	500 20 100				3 0.815 0.335 0.857 (2 0.122		1 0.826 0.00	
	500 20 100				2 0.815 0.340 0.846 (1 0.823 0.06	
CONTRACTOR OF THE PROPERTY OF	500 20 50				3 0.819 0.334 0.859 (2 0.826 0.06	
	500 50 50				1 0.850 0.291 0.895					
	500 20 100				8 0.843 0.289 0.896 (
A CONTRACTOR OF THE CONTRACTOR	500 20 50				6 0.843 0.290 0.892 (
	500 20 50				2 0.839 0.294 0.896 (
					6 0.823 0.304 0.890 (
	500 50 50			0.345 0.841 0.43						
/embedc 26286 N 100 0.3 5	500 20 20	300 N Y Y Y	0.857 0.322 0.883	0.334 0.857 0.47	4 0.845 0.288 0.895 0	0.375 0.783 0.364 0.82	0.250	0.757 0.37	1 0.820 0.00	52
/embedc 26285 N 100 0.3 !	500 500 500	20 20	20 20 300 N Y N	20 20 300 N Y N 0.853 0.325 0.881	20 20 300 N Y N 0.853 0.325 0.881 0.335 0.854 0.46	20 20 300 N Y N 0.853 0.325 0.881 0.335 0.854 0.469 0.841 0.291 0.895 0	20 20 300 N Y N 0.853 0.325 0.881 0.335 0.854 0.469 0.841 0.291 0.895 0.362 0.790 0.369 0.81	20 20 300 N Y N 0 0.853 0.325 0.881 0.335 0.854 0.469 0.841 0.291 0.895 0.362 0.790 0.369 0.814 0.228	20 20 300 N Y N 0.853 0.325 0.881 0.335 0.854 0.469 0.841 0.291 0.895 0.362 0.790 0.369 0.814 0.228 0.675 0.39	20 20 300 N Y N 0.853 0.325 0.881 0.335 0.854 0.469 0.841 0.291 0.895 0.362 0.790 0.369 0.814 0.228 0.675 0.397 0.796 -0.076



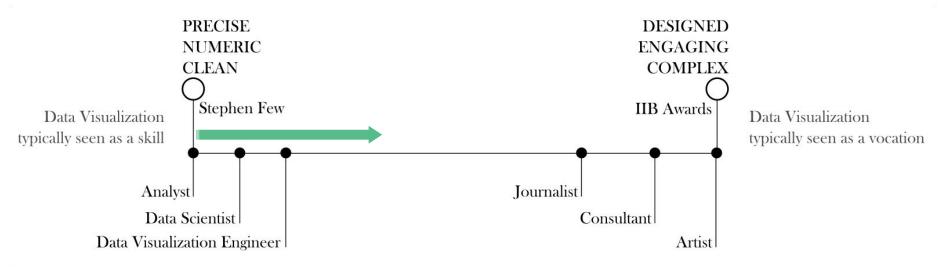
```
1,9Uee7oEuuMmgPx2IzPfFkWgkHZyPbWr0,DPnLzkJJg00PRJfBxIHbQEERiYHu5ila
3,1qXC7Fjbwp66GPQc6pHLfEu08WKozxG4,7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
4, FIHlppZyoq8muPbdVxS44gfvceX9zvU7, DPnLzkJJq00PRJfBxIHbQEERiYHu5ila
5,p1Mp7WkVfzUijX0peVQKSHbgd5pXyl4c,7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
6, dpK33RH9yepUAnyoywRwBt1AJzxGlaja, AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd
7, I1KwJ6EdCZnEPLfC8Q7yWpIkL0Hn7h02, 7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
9, J1oRHoSJ0InehnrxVdh32dK7QnDuCJWo, DPnLzkJJq00PRJfBxIHbQEERiYHu5ila
12,9tsGjrRgtMZ6V7yrA0yf0QPZHaltDHAp,DPnLzkJJq00PRJfBxIHbQEERiYHu5ila
13, hDbSkVrFRj9Ryk3c5E1JYJQLyxm4jLRb, 5X6FeZozNMgE2VRi3MJYjkkFK8SETtu2
14,X0hIczT5nEe052jMq1vN7QziDk8L2jnI,DPnLzkJJq00PRJfBxIHbQEERiYHu5ila
16,mPSPvu82Gr17tV9GJ95bDC7exvsVnwDE,DPnLzkJJg00PRJfBxIHb0EERiYHu5ila
18, b0Hk5D3sJulvyuC4JEm5kvAv0LAxswgQ, DPnLzkJJq00PRJfBxIHbQEERiYHu5ila
20, BoK7CAUaCFgnLgmWLxeOHg8YkXUSeCtc, DPnLzkJJgOOPRJfBxIHbQEERiYHu5ila
22,dPBUV0FPFjTZZK079rPAeq0WXhW4DUkF,7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx
23, BoK7CAUaCFqnLgmWLxeOHg8YkXUSeCtc, AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd
26, vcAiZWU2sfUK00mnfjDwm0iTzACrKr78, DPnLzkJJq00PRJfBxIHbQEERiYHu5ila
28, BoK7CAUaCFgnLgmWLxeOHg8YkXUSeCtc, TAYxxh39I2LZnftBpL0LfF2NxzrCKpkx
30, JPkczY0xyoDZBjwZAAQHmjpSvnPQzwV0, DPnLzkJJq00PRJfBxIHbQEERiYHu5ila
```

enrollment id, username, course id



Credit: <u>Medium Article</u>

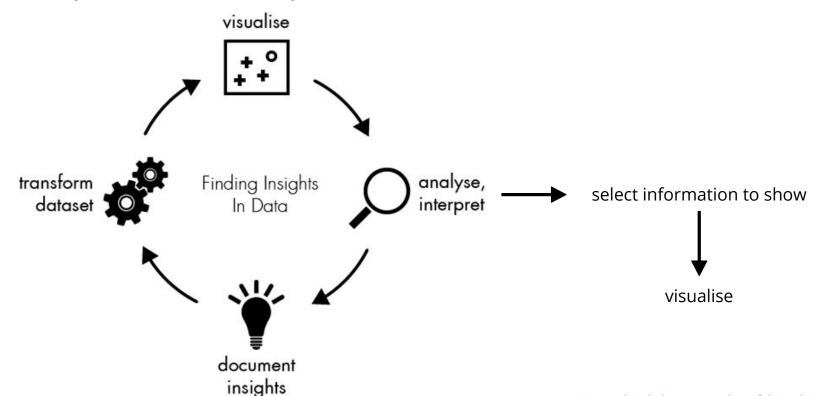
¿Podés identificar tu rol en esta línea? ¿A dónde querés llegar?



Credit: Medium Article

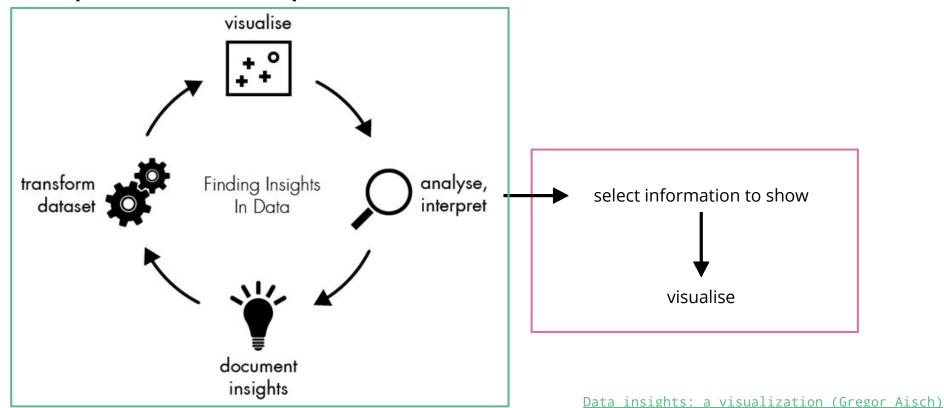
¿Podés identificar tu rol en esta línea? ¿A dónde querés llegar?

Exploración vs presentación

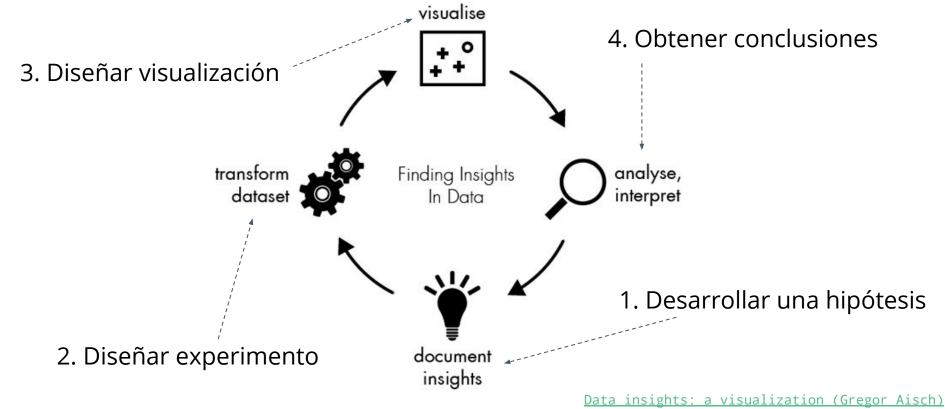


Data insights: a visualization (Gregor Aisch)

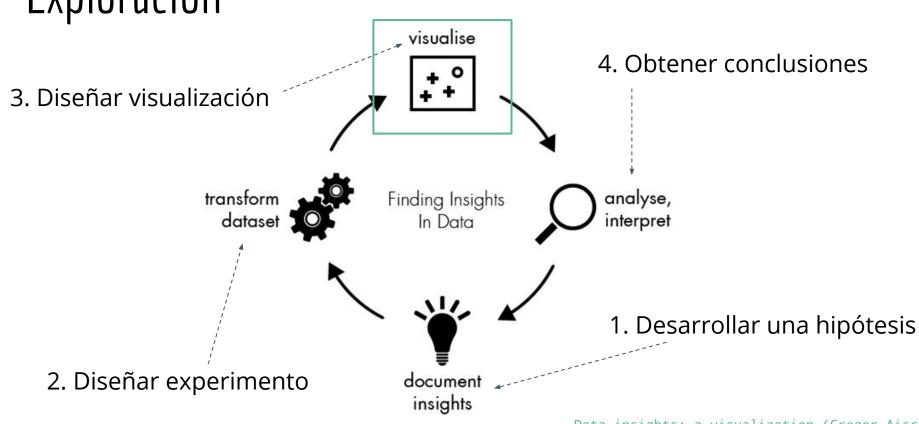
Exploración vs presentación



Exploración



Exploración



Data insights: a visualization (Gregor Aisch)

Configurar la notebook

Tipos básicos de gráficos

Tablas

variables

			Provincia	Población 2001	Población 2010	Variación absoluta	Variación relativa (%)
•	Muestran cantidades exactas	0 Ciuc	ıdad de Buenos Aires	2.776.138	2.890.151	114.013	4,1
i maestran eantradaes		1	Buenos Aires	13.827.203	15.625.084	1.797.881	13,0
•	Representan cualquier tipo de	2	Catamarca	334.568	367.828	33.260	9,9
	·	3	Chaco	984.446	1.055.259	70.813	7,2
	datos	4	Chubut	413.237	509.108	95.871	23,2
		5	Córdoba	3.066.801	3.308.876	242.075	7,9
•	Son de acceso universal	6	Corrientes	930.991	992.595	61.604	6,6
		7	Entre Rios	1.158.147	1.235.994	77.847	6,7
•	Son fáciles de crear	8	Formosa	486.559	530.162	43.603	9,0
•	Permiten comprar muchas	9	Jujuy	611.888	673.307	61.419	10,0

Tablas

	Ticker	Name	Value	Change
Las tablas en general no son		Dow Jones	15,988.08	↓ -2.39%
buenas resaltando patrones ,		S&P 500	1,880.33	↓ -2.16%
padicios,		Technology		1 2.10%
pero con un poco de formato se	IBM	IBM	130.00	↓ -2.19%
vuelven más legibles (y	AAPL	Apple	97.05	↓ -2.48%
	MSFT	Microsoft	50.99	↓ -3.99%
cautivantes)	https://www	.r-bloggers.com/for	-matting-table-out	out-in-r/

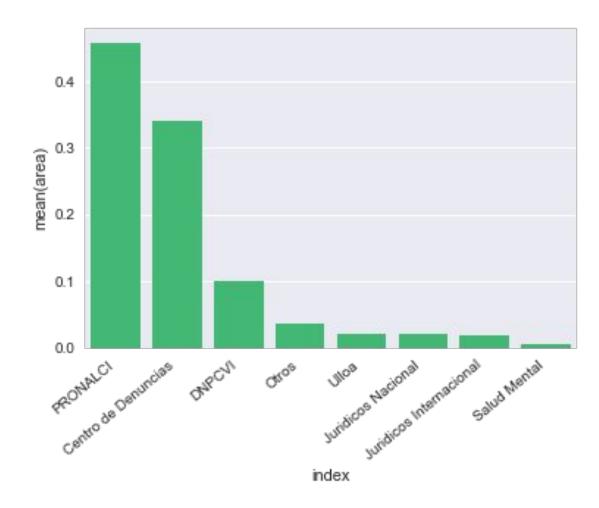
Tablas

Table 1. Grading rubric for writing assignments. Lists categories, evaluation criteria, and point values for each criterion.

Possible grade	Length	Topic	Argument	Mechanics	Citations
Α	The paper meets the page length requirement and is formatted correctly.	Topic fits the scope of the project, makes a clear argument.	Project includes in-depth discussion and elaboration in all sections.	No spelling and/or grammar mistakes.	Cites all information from out of class discussion sources. APA citation style is used in both text and bibliography.
	10 points	20 points	20 points	5 points	10 points
В	The paper meets the length requirement but has inconsistent citation formatting.	The paper is focused but does not make a clear argument.	Project includes in-depth discussion and elaboration in most sections. Minimal spelling and/or grammar mistakes.		Cites most information obtained from other sources.
	8.5 points	17 points	17 points	4.25 points	8.5 points
С	The paper is up to 1 page too short or too long or is incorrectly formatted.	Topic is either too broad or too narrow.	Project has omissions of content or content runs-on excessively. Paper relies heavily on quotations for content.	Several spelling and grammar mistakes.	Cites some information from other sources. Citation style is either inconsistent or incorrect.
	7.5 points	15 points	15 points	3.75 points	7.5 points
D	The paper is more than 1 page longer or shorter than assigned.	Paper does not stay on topic.	Project has cursory discussion in all the sections of the paper or brief discussions in only a few sections.	Many spelling and grammar mistakes that make the paper hard to understand.	Does not cite sources.
	6.5 points	13 points	13 points	3.25 points	6.5 points

categóricas entre variables

Son uno de los encodings más fieles y fáciles de percibir

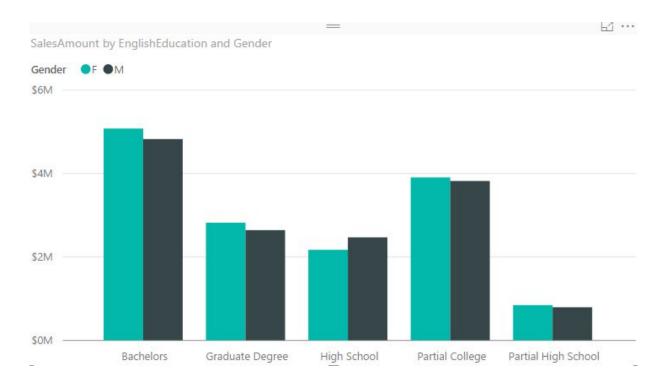


Permiten comparar cantidades

en **grupos**

Grouped vs stacked

Stacked at 100%



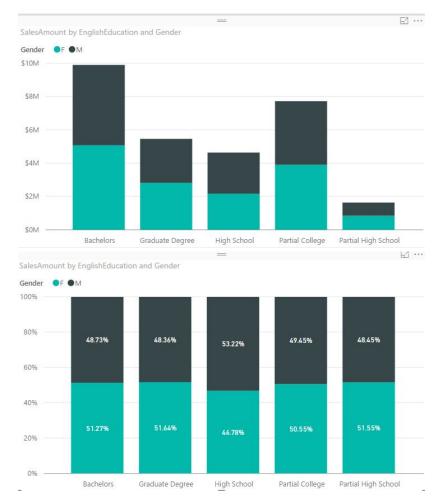
http://radacad.com/stacked-chart-or-clustered-which-one-is-the-best

Permiten comparar cantidades

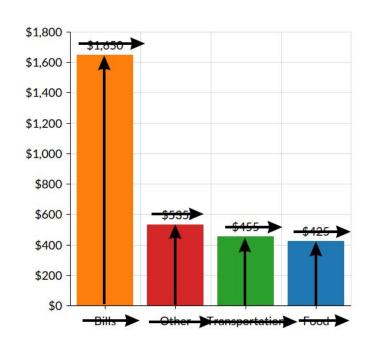
en **grupos**

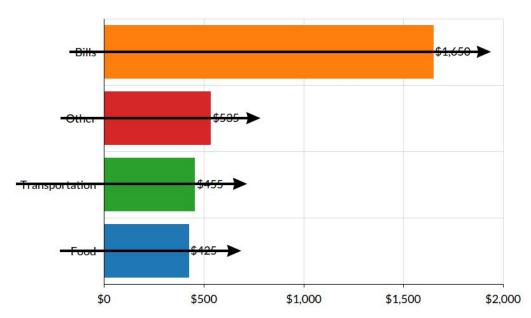
Grouped vs stacked

Stacked at 100%



http://radacad.com/stacked-chart-or-clustered-which-one-is-the-best





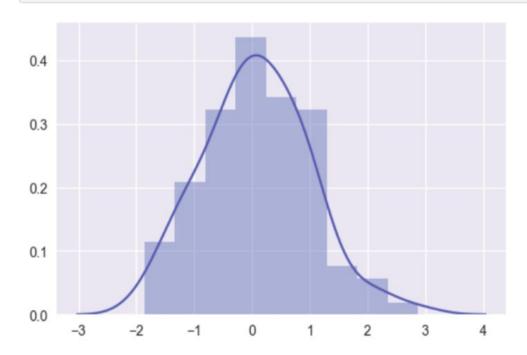
Histograma

- ¡No es lo mismo que un gráfico de barras!
- Muestra la distribuciónde una variablenumérica
- Divide los datos en bins

Plotting univariate distributions

The most convenient way to take a quick look at a univariate distribution in seaborn is the distplot(

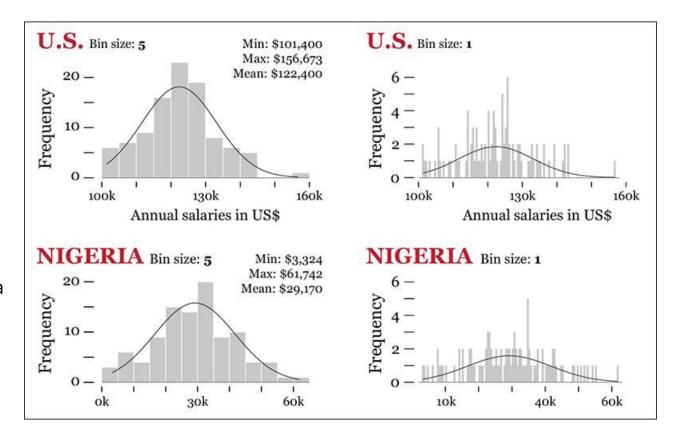
```
x = np.random.normal(size=100)
sns.distplot(x);
```



https://seaborn.pydata.org/tutorial/distributions.html

Histograma

La **binarización** es muy importante en el histograma resultante.



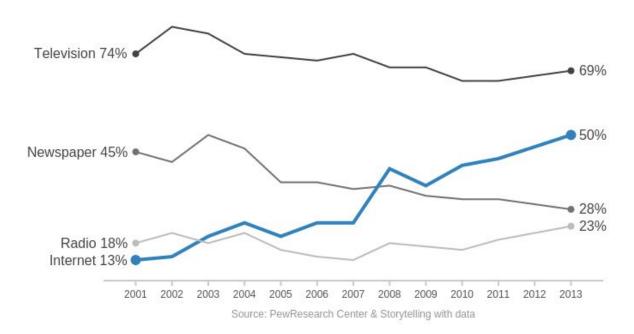
Gráficos de línea

Cada línea representa la variación de dos **variables**

numéricas

Múltiples líneas permiten comparar distintas categorías

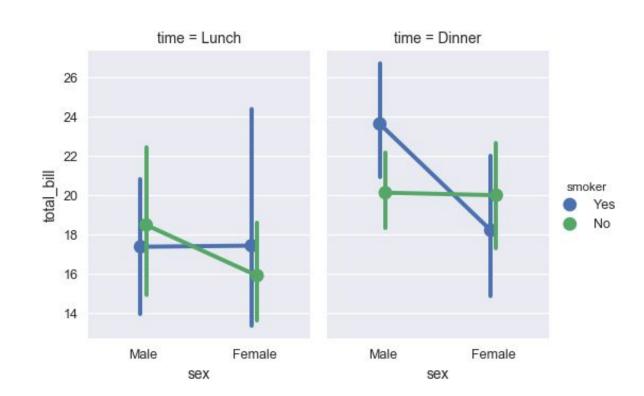
Main Source for News



Gráficos de línea

Son muy versátiles

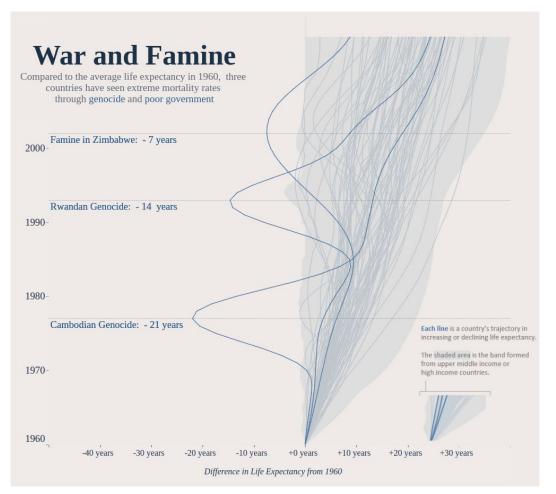
Generan gráficos visualmente simples, por lo que pueden contener mucha información



Gráficos de línea

Son muy versátiles

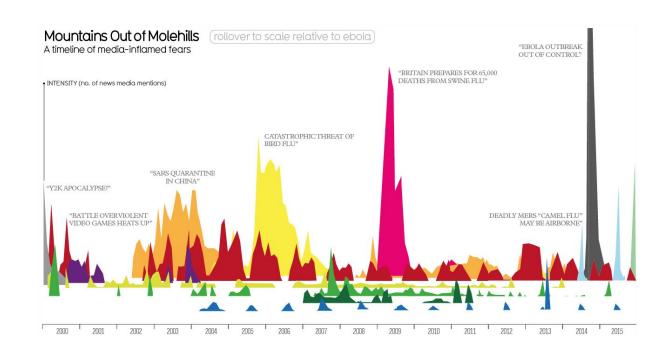
Generan gráficos visualmente simples, por lo que pueden contener mucha información



War and Famine

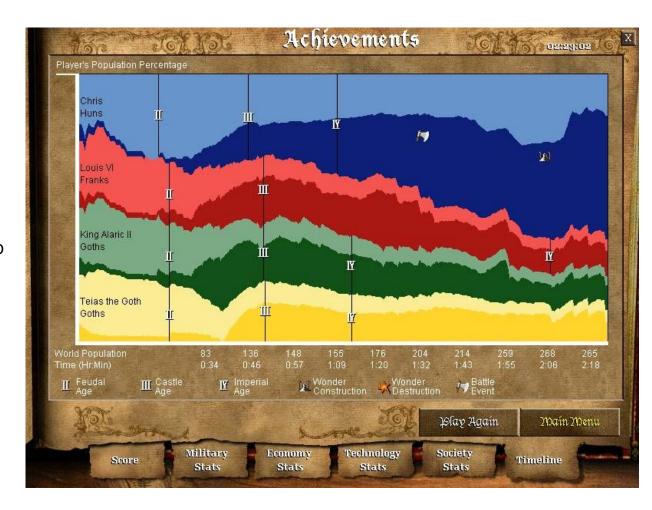
Gráficos de área

Iguales a los gráficos de línea, pero tienen más **impacto visual**.



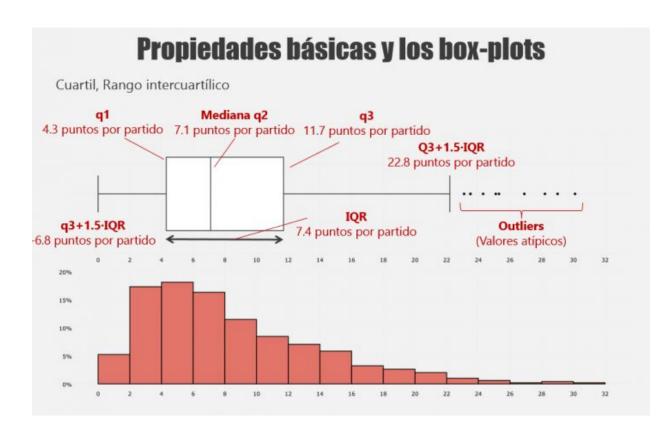
Gráficos de área

Pueden ser **apilados**, tomando propiedades similares a los gráficos de barra apilados

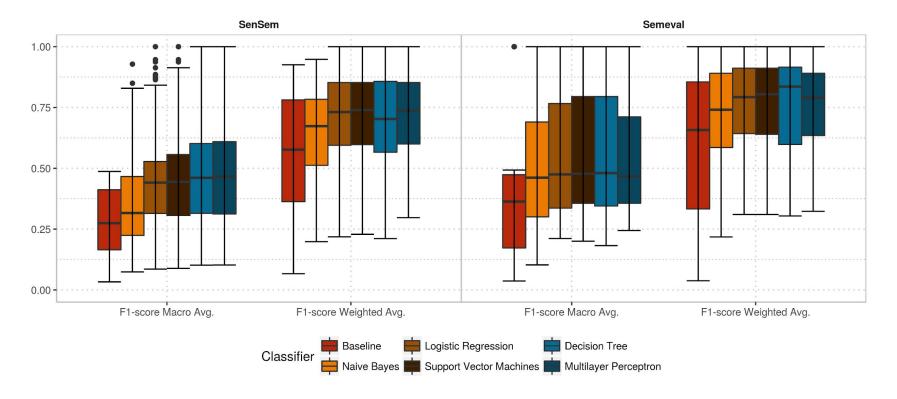


Gráficos de cajas

- Muestra la distribución
 de una variable
 numérica continua.
- Muestra información de forma más condensada que un histograma.



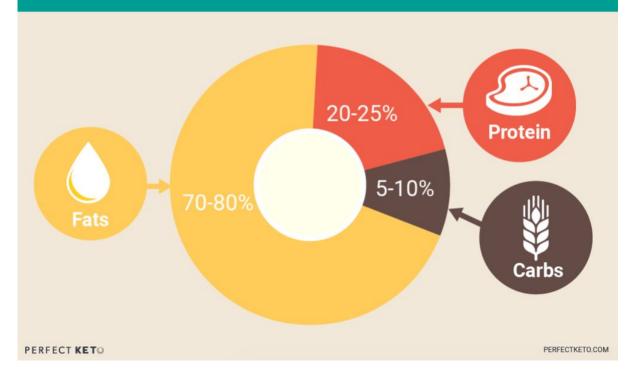
Gráficos de cajas



Gráficos de torta

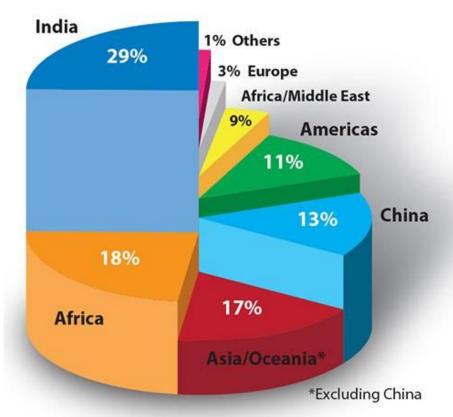
Ilustra la distribución de la población con respecto a una variable categórica.

KETOGENIC Diet BREAKDOWN



The horror of pie charts

Share of worldwide urban population growth 2010-2050



The Truthful Art: Data, Charts and Maps for Communication

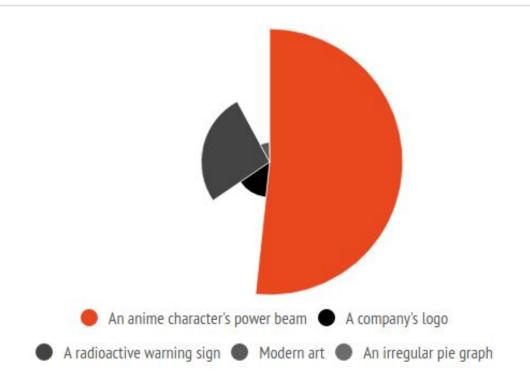
no los tiene!

Son tan malos, que seaborn

Gráficos de torta

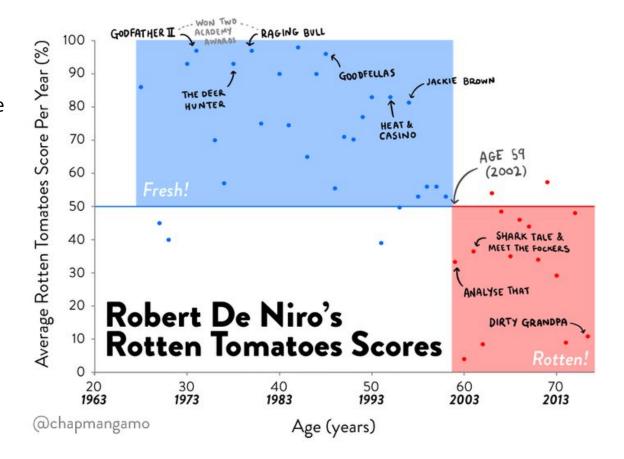
Elementos visuales adicionales facilitan la comparación entre elementos del gráfico

What people think this chart type is



Gráficos de puntos

- Muestra la relación entre
 2 o 3 variables
 numéricas continuas
- Puede usar color, forma
 de los puntos para
 variables categóricas, y
 tamaño para variables
 numéricas

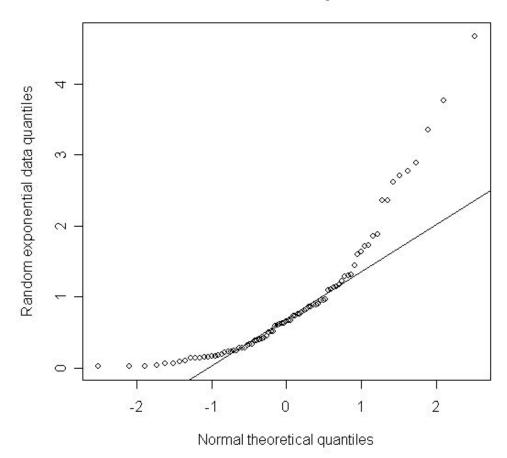


Scatter plot shows exactly when Robert de Niro stopped making good films

Normal Q-Q Plot with exponential data

Gráficos QQ

- Compara los cuartiles de dos muestras
- Sirve para ver que tan
 parecidas son las dos
 distribuciones de las que
 provienen las muestras



Tutoriales de Seaborn

- Visualizar datos categóricos:
 https://seaborn.pydata.org/tutorial/categorical.html
- Visualizar datos lineales
 https://seaborn.pydata.org/tutorial/relational.html#relational-tutorial
- Encontrar relaciones entre variables
 https://seaborn.pydata.org/tutorial/regression.html

¿preguntas?