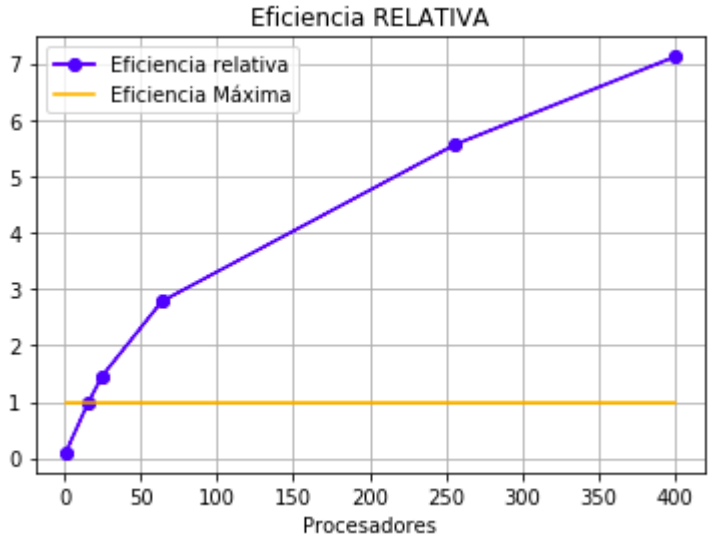
Martiñón Luna Jonathan José
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
Licenciatura en Ciencia de Datos

Con base en la tabla de datos mostrada en la diapositiva 4 de la presentación Introducción a la Visualización, realizar lo siguiente:

1. Encontrar el valor atípico y explicar por qué se considera un valor extraño.

El valor atípico es el tiempo inicial (`29278`), se considerará un valor extraño, puesto que al compararlo con los demás valores obtenidos sobre sale del patrón y afecta nuestros resultados.

Me parece que se puede apreciar de mejor manera en la siguiente gráfica:



Como podemos observar, es el primer valor el que se encuentra fuera del rango de los demás valores (0 – 7)

2. Crear un archivo de esos datos en el formato CSV y leerlo en un Dataframe de Pandas.

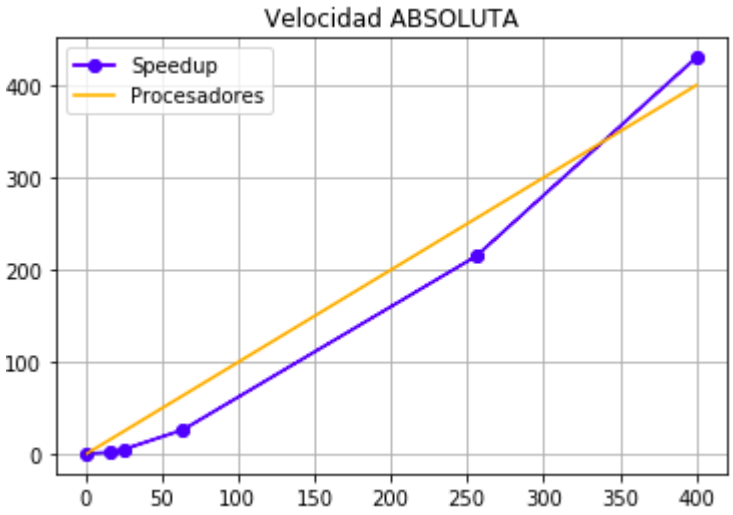
3. Completar las columnas de Speeup, Efficiency y Cost usando las fórmulas mostradas en la diapositiva 6, tanto para el caso absoluto como para el caso relativo con p' = 16. Agregar estos datos al dataframe antes creado y realizar los gráficos correspondientes usando Matplotlib.

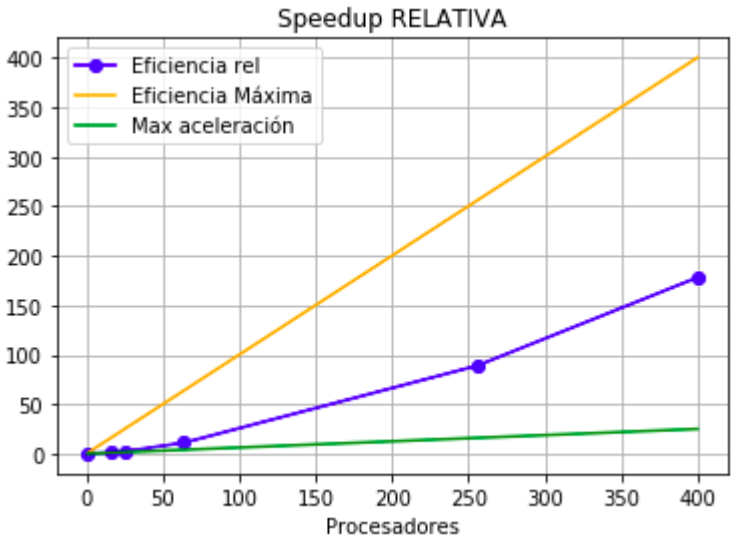
4. Generar información similar al punto 2, sustituyendo el valor atípico por el valor mostrado en la transparencia 9. Agregar estos datos al dataframe antes creado y realizar los gráficos correspondientes usando Matplotlib.

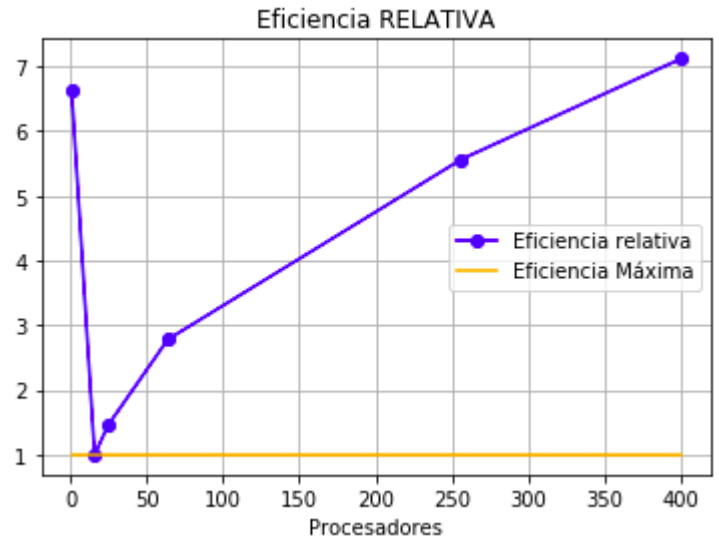
	Proc.	Time [s]
0	1	430
1	16	178
2	25	78
3	64	16
4	256	2
5	400	1

Nuestros nuevos datos a graficar

	Proc.	Time [s]	Speedup	Efficiency	Cost	Aceleracion_rel	Eficiencia_rel	Max_ac_rel
0	1	430	1.000000	1.000000	430	0.413953	6.623256	0.0625
1	16	178	2.415730	0.150983	2848	1.000000	1.000000	1.0000
2	25	78	5.512821	0.220513	1950	2.282051	1.460513	1.5625
3	64	16	26.875000	0.419922	1024	11.125000	2.781250	4.0000
4	256	2	215.000000	0.839844	512	89.000000	5.562500	16.0000
5	400	1	430.000000	1.075000	400	178.000000	7.120000	25.0000







5. Contar la historia del análisis de esta información usando gráficos que resalten lo realizado en los incisos 2 y 3.

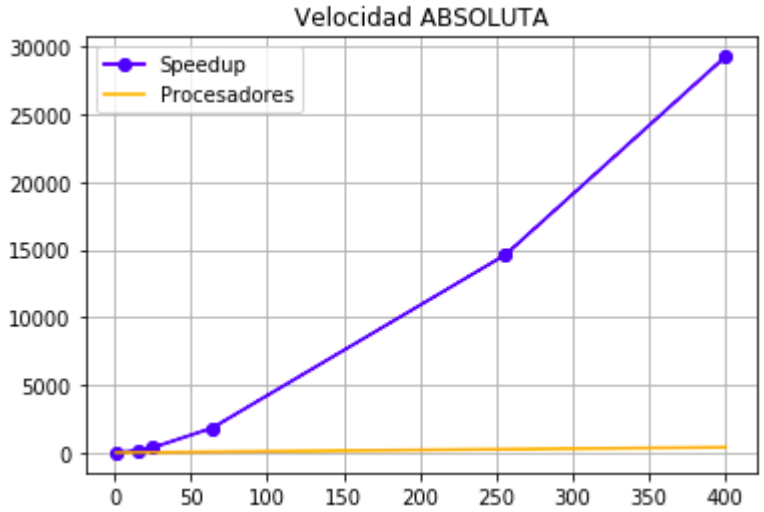
Se cuenta con cierta información acerca de la cantidad de procesadores con los que se está trabajando, así como el tiempo que se tardan en realizar la misma operación, de tal forma que podamos visualizar con qué cantidad de procesadores nos conviene realizar dichas operaciones.

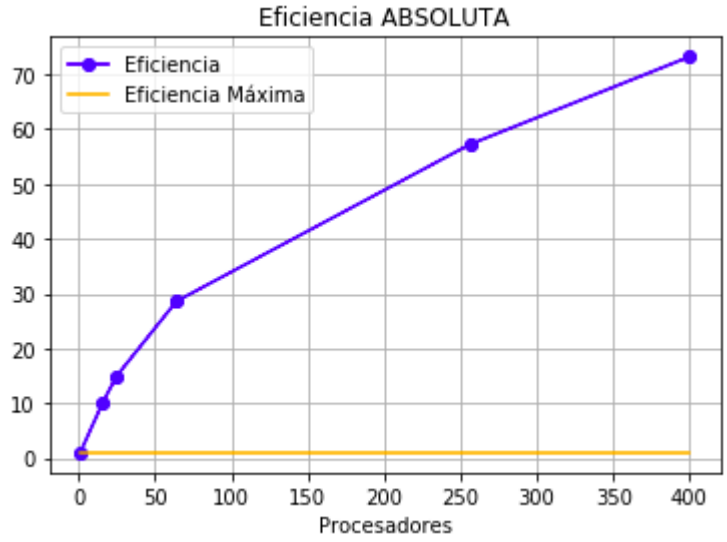
	Proc.	Time [s]
0	1	29278
1	16	178
2	25	78
3	64	16
4	256	2
5	400	1

Sin embargo, se logran extender los datos para obtener ya no sólo los tiempos, sino también diversas cualidades, tales como eficiencia, speedup y costo, así como la aceleración, eficiencia y máxima aceleración todas relativas.

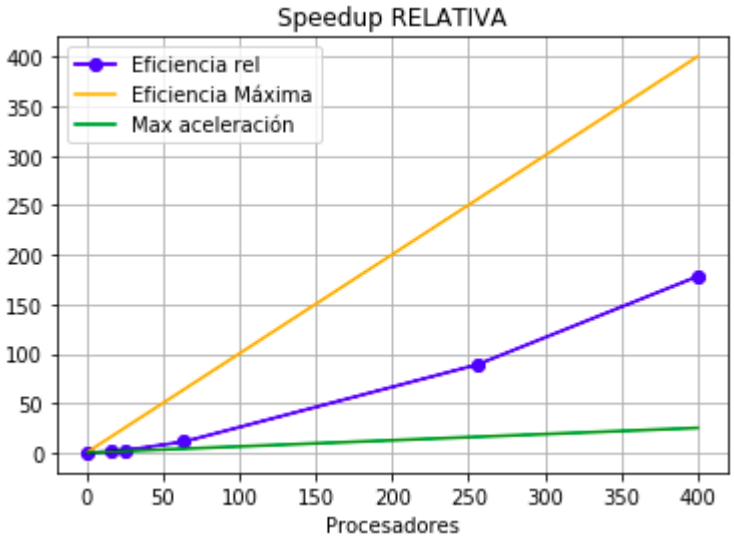
	Proc.	Time [s]	Speedup	Efficiency	Cost	Aceleracion_rel	Eficiencia_rel	Max_ac_rel
0	1	29278	1.000000	1.000000	29278	0.006080	0.097274	0.0625
1	16	178	164.483146	10.280197	2848	1.000000	1.000000	1.0000
2	25	78	375.358974	15.014359	1950	2.282051	1.460513	1.5625
3	64	16	1829.875000	28.591797	1024	11.125000	2.781250	4.0000
4	256	2	14639.000000	57.183594	512	89.000000	5.562500	16.0000
5	400	1	29278.000000	73.195000	400	178.000000	7.120000	25.0000

A simple vista los datos no reflejan mucha información, están ahí, pero cuesta interpretarlos a simple vista, es por ello que nos ayudamos de diversas gráficas para poder llevar a cabo un mejor análisis

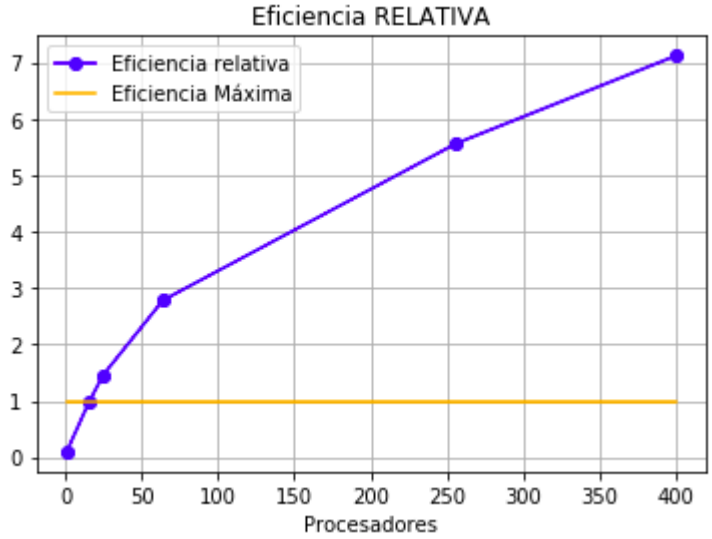




Como podemos observar en dichas gráficas, los datos aparentan estar por encima del máximo esperado, algo que no debería suceder, es decir, suena impresionante una eficiencia por encima del 100% (1), así como una velocidad por encima de la máxima posible.



Aquí podemos observar que no está tan mal nuestra eficiencia relativa, tiene cierto sentido, a pesar de que no tiene una gran eficiencia, sí suena más lógico.



Nuevamente nos encontramos con el problema de la eficiencia aún cuando se trata de la relativa.

6. Leer el artículo: Francis Anscombe, (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966. JSTOR 2682899. (LINK LOCAL). Poner especial atención a la sección: 3. An example.

7. Completar el notebook "Anscombe.ipynb" con lo siguiente: (A) Su punto de vista del artículo mencionado en el punto 1 (particularmente de la sección 3); (B) Realizar los cálculos descritos en la transparencia 11 de la presentación Intro. a la Vis. (página 11); (C) "Retocar" los gráficos para que se muestre la información que es necesaria (etiquetas en los ejes, títulos, leyendas, tipo de marcadores, colores, etc).

A) De acuerdo con las nociones presentadas en un inicio, cabe recalcar que existían diferentes tecnologías que limitaban los avances de las investigaciones (a mi parecer), motivo por el cual posiblemente en 1973 no fuese tan exacto el uso de gráficos, sin embargo, es algo que actualmente nos ayuda bastante a comprender diversos comportamientos de los datos de una manera más “amigable”. Tampoco comparto la idea de que exista sólo un camino para un correcto análisis, pues a mi parecer, existen diversas formas en que se pueden explotar y analizar los datos.

Referente al punto 3, me parece que expresa la idea de cómo es que los datos atípicos logran afectar nuestros análisis, sin embargo, hay que tener en cuenta que no siempre tendremos datos ajustados “perfectamente”, pues como se menciona, se elimina Alaska, ¿y luego? ¿Se continua eliminando datos infinitamente? Me parece que existen diversos métodos para trabajar con dichos datos, sin embargo, me pregunto si es lo adecuado, pues a fin de cuentas son datos obtenidos, y tendrán sus motivos, pienso que se debería investigar su origen y la importancia que tienen dentro del análisis.

B) Para la serie 1

```
La media de x = 9.0
La varianza de x =  10.0
La media de y = 7.500909090909091
La varianza de y = 3.752062809917356
```

```
Con coeficiente =  [[0.50009091]]

Para la serie 2
La media de x = 9.0
La varianza de x =  10.0
La media de y = 7.500909090909091
La varianza de y = 3.752390082644628
Con coeficiente =  [[0.5]]

Para la serie 3
La media de x = 9.0
La varianza de x =  10.0
La media de y = 7.5
La varianza de y = 3.747836363636364
Con coeficiente =  [[0.49972727]]

Para la serie 4
La media de x = 9.0
La varianza de x =  10.0
La media de y = 7.500909090909091
La varianza de y = 3.7484082644628103
Con coeficiente =  [[0.49990909]]
```

Los coeficientes de correlación son:

	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
x_1	1.000000	0.816421	1.000000	0.816237	1.000000	0.816287	-0.500000	-0.314047
y_1	0.816421	1.000000	0.816421	0.750005	0.816421	0.468717	-0.529093	-0.489116
x_2	1.000000	0.816421	1.000000	0.816237	1.000000	0.816287	-0.500000	-0.314047
y_2	0.816237	0.750005	0.816237	1.000000	0.816237	0.587919	-0.718437	-0.478095
x_3	1.000000	0.816421	1.000000	0.816237	1.000000	0.816287	-0.500000	-0.314047
y_3	0.816287	0.468717	0.816287	0.587919	0.816287	1.000000	-0.344661	-0.155472
x_4	-0.500000	-0.529093	-0.500000	-0.718437	-0.500000	-0.344661	1.000000	0.816521
y_4	-0.314047	-0.489116	-0.314047	-0.478095	-0.314047	-0.155472	0.816521	1.000000

8. Realizar todo lo anterior en Python 3.7 usando notebooks de Jupyter y subirlos a SU REPOSITORIO DEL CURSO EN GITHUB.
9. Indicar que la tarea está lista por este medio y poner el link de los notebooks correspondientes.