

DATA SCIENCE

Recursos:

- ML Interviews: <https://huyenchip.com/ml-interviews-book/> (<https://huyenchip.com/ml-interviews-book/>)
- Es tracta d'un esdeveniment per fomentar la participació en projectes de codi obert. És ideal per a gent que està començant a programar per aprendre i alhora contribuir en projectes. Hi ha diverses tecnologies per posar en pràctica: HTML, CSS, JavaScript, React, Python, Django, etc. <https://hacktoberfestes.dev/> (<https://hacktoberfestes.dev/>)
- DATETIME PYTHON CHEAT SHEET: <https://strftime.org/> (<https://strftime.org/>)
- Seleccionar virtual environments VSCode: <https://code.visualstudio.com/docs/python/environments> (<https://code.visualstudio.com/docs/python/environments>)

Web con buenas explicaciones. SciPy, Hypothesis testing, linear equations, etc.

<https://www.webpages.uidaho.edu/~mlowry/Teaching/Analysis&Design/build/html/index.html>
(<https://www.webpages.uidaho.edu/~mlowry/Teaching/Analysis&Design/build/html/index.html>)

msg discord

y a todos los colegas, os dejo aqui un recurso alternativo para el ejercicio de la geolocalizacion del ip, ya que ip2geotools me empezó a dar problemas de 'invalid request error': la paginaweb - <https://ipinfo.io/> (<https://ipinfo.io/>) y la documentacion - <https://github.com/ipinfo/python/blob/master/README.md> (<https://github.com/ipinfo/python/blob/master/README.md>) . Espero que os sea util! Hace falta login para tener acceso a un token personal, pero puedes hacerlo automaticamente usando su cuenta de github. Es super practico!

Buenas tardes. Me gustaría compartir un problema que acabo de resolver gracias a este enlace. Se trata del método GridSearchCV al intentar optimizar ROC_AUC. <https://stackoverflow.com/questions/49061575/why-when-i-use-gridsearchcv-with-roc-auc-scoring-the-score-is-different-for-gri> (<https://stackoverflow.com/questions/49061575/why-when-i-use-gridsearchcv-with-roc-auc-scoring-the-score-is-different-for-gri>)

El problema era que no me coincidía el AUC calculado por el método GridSearchCV y el `metrics.roc_auc_score` porque usan argumentos distintos

Para información Sprint19, Windows da problemas para extraer el archivo zip de kibana. Solución: descargar 7-zip.

Hola! Si algú està estancat en entendre Pipeline, m'ha anat molt bé llegir la documentació després dels recursos proveïts al curs <https://scikit-learn.org/stable/modules/compose.html#combining-estimators> (<https://scikit-learn.org/stable/modules/compose.html#combining-estimators>).

- Hola. A la tasca de webscraping de la web bolsamadrid.es quines pàgines heu "escrapejat"?
 - Jo vaig utilitzar el Selenium per navegar fins a les accions de l'Ibex-35 i vaig descarregar la taula. També vaig mostrar com poder descarregar altres dades (d'empreses). Entec que la idea es aprendre l'us bàsic de les diferents eines, així que tampoc vaig explorar extensivament la web
-

- Bon dia. Algú sap si es pot fer això en SQL ? Vull crear a una taula una columna de text que només admeti aquest patró: any-Wsetmana e.g. "2005-W14", "2013-W01", etc. I que no deixi insertar un registre si el valor d'aquesta columna no té aquest format. Sabeu si es pot fer ? I si es pot fer com es fa ? No trobo res a google... Merci
 - Si vols fer-ho "embedded" a la base de dades ho faria amb un trigger
 - Mira això si et pot ser d'utilitat <https://cvuorinen.net/2013/05/validating-data-with-triggers-in-mysql/> (<https://cvuorinen.net/2013/05/validating-data-with-triggers-in-mysql/>)
 - I aquí tens un thread a stackoverflow que parla d'això <https://stackoverflow.com/questions/17032420/doing-a-pattern-match-mysql-side-before-data-is-inserted-into-a-table> (<https://stackoverflow.com/questions/17032420/doing-a-pattern-match-mysql-side-before-data-is-inserted-into-a-table>)
-

Debugger: <https://usedevbook.com/> (<https://usedevbook.com/>)

- Bon dia! Si voleu fer debug amb Python podeu utilitzar això! <https://blog.jupyter.org/a-visual-debugger-for-jupyter-914e61716559> (<https://blog.jupyter.org/a-visual-debugger-for-jupyter-914e61716559>)

07/04/2021

Para qué sirve Numpy? Con numpy transformamos los **literales** (enero, febrero, masculino, rojo, etc.) en **categoricos**, esto es, en números. Por ejemplo, tenemos dos literales (hombre y mujer) y lo transformamos en una secuencia binaria (0, 1).

Conceptos a investigar:

- Masterclass de los notebooks sobre pre-procesamiento.
- dummy variables
- broadcasting https://www.tutorialspoint.com/numpy/numpy_broadcasting.htm
(https://www.tutorialspoint.com/numpy/numpy_broadcasting.htm)
- mask

NUMPY.MEDIAN()

Ejercicio 2 de la tasca 2A del sprint 3

<https://numpy.org/doc/stable/reference/generated/numpy.median.html> (<https://numpy.org/doc/stable/reference/generated/numpy.median.html>)

Sirve para calcular la media de un array: `numpy.median(a, axis, out, overwrite_input, keepdims)`.

Parámetros:

- a: el array
- axis: **nota**: revisar el tema axis en los ejemplos de la documentación porque no está claro.

8/04/21

Antes el sprint 6 mirar el inbalance (sprint 14). El imbalance afecta a problemas de clasificación. Sprint 9 (normalización (polinomios, dos variables) // standarización (raíces cuadradas)).

Sprint 6 y 7 de cara al lunes.

Paradigma big data (presentación. Esquema del flujo: ingesta, visualización, análisis, etc.). Vamos a montar un Checklist.

Tenemos una estructura y vamos a checklist de orden. Qué hago en la ingesta, en el almacenaje (falta el procesamiento), luego vendría la transformación. Normalización (transformación del dato de la base de datos SQL). Transformarlo para que pueda añadirlo al algoritmo.

Normalizar es transformar los datos en una distribución normal, a partir de la cual es más fácil hacer inferencia. Cuanto menos centrada esté, más variación (ruido) hay, y se hace más complicado inferir. Pese a que la normalización no sea perfecta, nosotros también podemos calcular el margen de error (ruido).

-kde mejor que hist (historiograma)

(Vuelos de aviones, retrasos, etc.)

Si funcionara perfectamente, tendríamos una distribución normal. Lo que nos interesa saber son

los patrones de lo que no funciona normal, eventos atípicos, los **outlayers**.

Anomalía (errores, colocando validación puedes resolver) vs **outlayer** (eventos que se pueden repetir, a qué es debido? Por ejemplo lo que hacía mi padre con el excel y mirar a ver si alguien se estaba bebiendo una coca cola de más o alguien está robando, etc. todo lo que se sale de lo normal y que tiene probabilidad de que vuelva a ocurrir)

- Definición de probabilidad en jupyter 2 de probabilidad y estadística.

Ley de los grandes números: cuanto más dato más probabilidad tengo de tener la probabilidad adecuada. Si tiro el dado 4 veces las probabilidades no son proporcionales, si lo tiro 40000 veces, se acercará más a la probabilidad que efectivamente es, esto es, se **estabilizará**. Qué está pasando? Que cuanto más datos tenemos más se estabiliza la variabilidad, menos variabilidad hay.

Cuanto menos dato tengo, la media será más dispersa (hay que compararla con el resto de observaciones), cuanto más dato, menos dispersa será.

Disjoint: (min 32, video 2). Disjoint es básicamente si dos eventos o variables son independientes. Esto es interesante porque cuando te enfrentas a una base de datos lo que te interesa saber es si los outliers tienen una causa, esto es, si **depende de otra variable o evento o si es independiente (disjoint) y no tiene que ver con otra cosa determinada**.

If A and B are any two events, *disjoint or not*, then the probability that at least one of them will occur is :

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

where $P(A \text{ and } B)$ is the probability that both events occur.

Para saber la probabilidad de $P(A)$ (dado 1) o $P(B)$ (dado 2) es la suma de la probabilidad de cada evento **menos (-) la probabilidad de que ocurran a la vez (en el caso de los dados es 0, por ello son dos eventos/variables disjoint)**

(A partir de min 40 no he escuchado y creo que es importante)

Sample space

El sample space de un dado es (1,2,3,4,5,6).

Shift + tabulador te da info para usar la función (min 50 aprox, no está muy accurate)

Sampling methods

La hipótesis define el tipo de muestra que quiero sacar.

BROADCASTING

<https://numpy.org/doc/stable/user/basics.broadcasting.html> ([https://numpy.org/doc/stable](https://numpy.org/doc/stable/user/basics.broadcasting.html)
[/user/basics.broadcasting.html](https://numpy.org/doc/stable/user/basics.broadcasting.html)) // https://www.tutorialspoint.com/numpy/numpy_broadcasting.htm

(https://www.tutorialspoint.com/numpy/numpy_broadcasting.htm)

El término broadcasting se refiere a la capacidad de numpy de tratar arrays de diferente o igual forma en operaciones aritméticas

Normalmente, las operaciones aritméticas se realizan sobre arrays con elementos que se corresponden, esto es, sobre arrays con la misma forma:

```
In [5]: import numpy as np

x = np.array([[3, 4, 200, 3443], [4, 3, 5, 6] ])
print('Forma de x:', np.shape(x))
y = np.array([[4, -33, -90, 90], [3, 0, 43, 6]])
print('Forma de y:', np.shape(y))

z = x * y

Forma de x: (2, 4)
Forma de y: (2, 4)
[[ 12 -132 -18000 309870]
 [ 12    0    215    36]]
```

A continuación, podemos observar que si los arrays son de diferente forma nos avienta un error:

```
In [8]: d = np.array([[3, 4, 5, 3], [3, 23, 2, 6], [5, 4, 3, 2]])
print(d.shape, x.shape)

(3, 4) (2, 4)
```

```
-----
-----
ValueError                                Traceback (most recent call
last)
<ipython-input-8-db120f83506c> in <module>
      2 print(d.shape, x.shape)
      3
----> 4 a = x + d

ValueError: operands could not be broadcast together with shapes (2,4)
(3,4)
```

Sin embargo, las matrices se pueden transmitir / broadcast si sus dimensiones coinciden pero también si una de las matrices tiene una dimensión de 1.

A set of arrays is said to be broadcastable if **one of the following is true**:

- Arrays have exactly the same shape.
- Arrays have the same number of dimensions and the length of each dimension is either a common length or 1.
- Array having too few dimensions can have its shape prepended with a dimension of length 1, so that the above stated property is true.

Traducción:

- Los arrays tienen exactamente la misma forma.
- Los arrays tienen el mismo número de dimensiones y la extensión de cada dimensión es común o 1.
- Los arrays con demasiado pocas dimensiones pueden "prepend" (transformar?) su forma con una dimensión de extensión 1, de forma que la condición de arriba es verdadera.

In [10]: `import numpy as np`

```
#si funciona porque g es de la misma extensión y de dimensión 1
x = np.array([[0, 0, 0, 0], [0, 0, 0, 0] ])
g = np.array([1, 2, 3, 4])
print(g+x)
```

```
#error, no son de la misma extensión
y = np.array([3, 4])
```

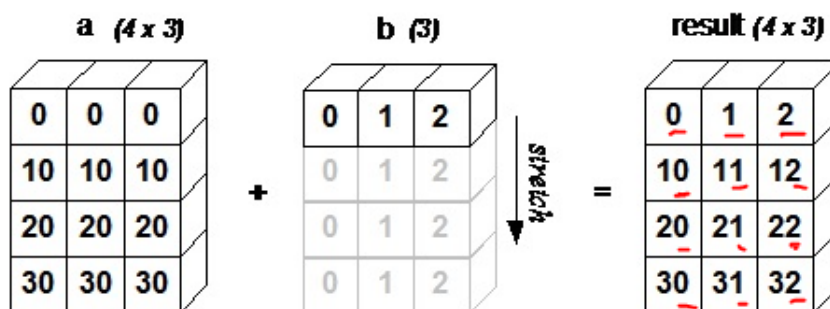
```
[[1 2 3 4]
 [1 2 3 4]]
```

ValueError Traceback (most recent call last)

```
<ipython-input-10-d7f75a9f570a> in <module>
      8 #error, no son de la misma extensión
      9 y = np.array([3, 4])
----> 10 print(x+y)
```

ValueError: operands could not be broadcast together with shapes (2,4) (2,)

The following figure demonstrates how array **b** is broadcast to become compatible with **a**.



In [12]: `import numpy as np`

```
x = np.array([[54, 23], [23, -90], [25, -43]])
```

```
y = np.array([[1, 2], [3, 4], [5, 6]])
```

```
print('x i y es poden transmetre o fer broadcast: ', x * y)
```

```

z = np.array([50, 100])

print('x i z es poden transmetre o fer broadcast:', x - z)

a = np.array([[500], [400]])
a_nou = a.reshape(2,)
print('a_nou i y es poden transmetre gràcies al reshape, si no no podrien

b = np.array([54, 23, 23, -90, 25, -43])
b_nou = b.reshape(3, 2)
print('b_nou i x es poden transmetre gràcies al reshape, si no no podrien

print('A continuació surtirà un error degut a que les dos matrius no comp
print(x*b)
x i y es poden transmetre o fer broadcast: [[ 54  46]
[ 69 -360]
[ 125 -258]]
x i z es poden transmetre o fer broadcast: [[  4 -77]
[ -27 -190]
[ -25 -143]]
a_nou i y es poden transmetre gràcies al reshape, si no no podrien:
[[501 402]
[503 404]
[505 406]]
b_nou i x es poden transmetre gràcies al reshape, si no no podrien: [[
108  46]
[  46 -180]
[  50 -86]]
A continuació surtirà un error degut a que les dos matrius no comparte
ixen característiques que els permeti fer broadcast:

```

```

-----
-----
ValueError                                Traceback (most recent call
last)
<ipython-input-12-0d4c48f6e4eb> in <module>
    20
    21 print('A continuació surtirà un error degut a que les dos matr
ius no comparteixen característiques que els permeti fer broadcast:')
--> 22 print(x*b)

ValueError: operands could not be broadcast together with shapes (3,2)
(6,)
```

---> cómo funciona newaxis + broadcast

- newaxis, como podemos ver en el siguiente ejemplo, hace que un array 1-d de x extensión se convierta en un array x-d de extensión 1.
- Al multiplicarlo con otro array de extensión "y", logramos broadcastearlo con el de extensión 1 y x dimensiones gracias a la ley del broadcast (lo único que sucede aquí es que ocurre en otro eje al que hemos visto anteriormente).
- El resultado es un array de x dimensiones + "y" de extensión.

```
In [23]: a = np.array([0.0, 10.0, 20.0, 30.0])
print(a.shape)
b = np.array([1.0, 2.0, 3.0])
z = a[:, np.newaxis] + b
print(z.shape)
print(a)
print(z)
print("")

u = np.arange(2)
y = b[:, np.newaxis] * u

(4,)
(4, 3)
[ 0. 10. 20. 30.]
[[ 1.  2.  3.]
 [11. 12. 13.]
 [21. 22. 23.]
 [31. 32. 33.]]

[[0. 1.]
 [0. 2.]
 [0. 3.]]
```

OPERACIONES VECTORIZADAS

https://www.pythonlikeyoumeanit.com/Module3_IntroducingNumpy/VectorizedOperations.html
[\(https://www.pythonlikeyoumeanit.com/Module3_IntroducingNumpy/VectorizedOperations.html\)](https://www.pythonlikeyoumeanit.com/Module3_IntroducingNumpy/VectorizedOperations.html)

En Numpy un array es homogéneo, es decir, contiene valores de un mismo tipo de dato, a diferencia de las listas y las tuplas de python. Esto significa que **no es necesario saber qué tipo de dato es cada valor del array**, puesto que deben ser todos el mismo.

Numpy es fantástico porque contiene un gran número de **funciones vectorizadas**, como por ejemplo `np.sum()`, la cual itera a lo largo de un array pero **sin necesidad de verificar qué dato es cada valor, cosa que reduce x50+ el tiempo de ejecución con respecto a una iteración de un array en python**.

Dicho esto, podemos entender que la **vectorización** hace referencia al **uso de código pre-compilado escrito en un low-level lenguaje de programación**. ((Por ejemplo: en machine learning pasamos **literales** como rojo, azul, masculino, etc. a **categoricos** para poder **manipular los datos como vectores y no ya como dato crudo**)).

- Unary functions $f(x) \rightarrow$ p. ej. `np.sqrt(x)`
- Binary functions $f(x, y) \rightarrow$ p. ej. `np.multiply(x, y)`
- Sequential Function: $f(\{x_i\}_{i=0}^n) \rightarrow$ p. ej. `np.median`

MASK

<https://numpy.org/doc/stable/reference/maskedarray.generic.html#module-numpy.ma>
<https://numpy.org/doc/stable/reference/maskedarray.generic.html#module-numpy.ma> //
<https://towardsdatascience.com/the-concept-of-masks-in-python-50fd65e64707>
<https://towardsdatascience.com/the-concept-of-masks-in-python-50fd65e64707>

Mask hace referencia a un proceso mediante el cual nos deshacemos del data que no queremos utilizar de un array. Lo que hace el masking es básicamente aplicar un array de booleans sobre nuestro array para filtrar.

Numpy tiene un módulo específico para hacer masking llamado **numpy.ma**. Veamos diferentes usos de este módulo:

```
In [24]: #importar
import numpy as np
import numpy.ma as ma

y = ma.array([1, 2, 3], mask = [0, 1, 0]) #To create an array with the se
z = ma.masked_values([1.0, 1.e20, 3.0, 4.0], 1.e20) #To create a masked a
print(y, z)
```

```
[1 -- 3] [1.0 -- 3.0 4.0]
```

```
In [25]: #less than (or less than equal to) a number
arr = np.array([1, 2, 3, 4, 5, 6, 7, 8])
ma_arr = ma.masked_less(arr, 4) #también se puede usar masked_less_equal(
print(ma_arr)
```

```
[-- -- -- 4 5 6 7 8]
```

```
In [27]: #Greater than (or greater than equal to) a number
arr = np.array([1, 2, 3, 4, 5, 6, 7, 8])
ma_arr = ma.masked_greater(arr, 4) #or masked_greater_equal()
```

```
[1 2 3 4 -- -- -- --]
```

```
In [28]: #Within a given range
arr = np.array([1, 2, 3, 4, 5, 6, 7, 8])
ma_arr = ma.masked_inside(arr, 4, 6)
```

```
[1 2 3 -- -- -- 7 8]
```

```
In [29]: #Outside a given range
arr = np.array([1, 2, 3, 4, 5, 6, 7, 8])
ma_arr = ma.masked_outside(arr, 4, 6)
```

```
[-- -- -- 4 5 6 -- --]
```

Neglecting NaN and/or infinite values during arithmetic operations

This is a cool feature! Often a realistic dataset has lots of missing values (NaNs) or some weird, infinity values. Such values create problems in computations and, therefore, are either neglected or imputed. You can easily exclude the NaN and infinite values using `masked_invalid()` that will exclude these values from the computations.

```
In [31]: arr = np.array([1, 2, 3, np.nan, 5, 6, np.inf, 8])
print(arr.sum())
ma_arr = ma.masked_invalid(arr)
print(ma_arr)

nan
[1.0 2.0 3.0 -- 5.0 6.0 -- 8.0]
25.0
```

Let's say you want to impute or fill these NaNs or inf values with the mean of the remaining, valid values. You can do this easily using **filled()**

```
In [33]: filled_arr = ma_arr.filled(ma_arr.mean())

[1.          2.          3.          4.16666667  5.          6.
 4.16666667  8.          ]
```

```
In [39]: print(ma.getmask(ma_arr)) #devuelve la máscara del array maskeado
print(ma.getmask(arr)) #devuelve false si el array no ha sido maskeado

[False False False  True False False  True False]
False
```

Nota al margen: "However, because you want to swap the True and False values, you can use the tilde operator `~` to reverse the Booleans".

```
"""Using Tilde operator to reverse the Boolean"""
ma_arr = ma.masked_array(arr, mask=[~((a<4) or (a>6)) for a in arr])
```

Oju:

- Exercici 6 Mask la matriu anterior, realitzeu un càlcul booleà vectoritzat, agafant cada element i comprovant si es divideix uniformement per quatre.

Això retorna una matriu de mask de la mateixa forma amb els resultats elementals del càlcul.

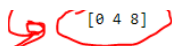
```
: import numpy.ma as ma
print('Matriu original:\n', matr_x)

ma_x = ma.masked_where(matr_x % 4 == 0, matr_x).mask
print('Mask:\n', ma_x)

Matriu original:
[[0 1 2]
 [3 4 5]
 [6 7 8]]
Mask:
[[ True False False]
 [False  True False]
 [False False  True]]
```

- Exercici 7 A continuació, utilitzeu aquesta màscara per indexar a la matriu de números original. Això fa que la matriu perdi la seva forma original, reduint-la a una dimensió, però encara obteniu les dades que esteu cercant.

```
: print(matr_x[ma_x])
```



[0 4 8]

13/04/2021 (FOUNDATIONS FOR INFERENCE)

Cómo documentar que la muestra que hemos extraído está bien para inferir -qué hacer pa quedarse tranquilo.

Después de la extracción del método. Saber que la muestra que he cogido sirve para inferir sobre la población.

"How sure are we that the estimated mean, \bar{x} , is near the true population mean, μ ?"

Intentamos estimar la media de la población a partir de la muestra. A veces no tengo el total de la población.

- A la media de la población le asignamos μ
- A la media de la muestra le asignamos \bar{x}

Hacer una exploración (mirar los datos, las formas de los gráficos, etc.) antes del pre-procesado, así podemos comparar con el resultado de la eliminación de outliers, etc.

Si se desplaza la gráfica a la derecha o a la izquierda significa que hay outliers o anomalías.

Running mean

Sacar una serie de observaciones e ir sacando medias, hacer la media entre las medias que van saliendo del for loop para acercarse a la media real. Esto es el running mean.

Sirve para saber qué cantidad de muestra me interesa coger. Obviamente cuanto más dato mejor, pero no siempre es posible.

Error estándar de la media

El error estándar de la muestra es la desviación estándar de la población σ entre la raíz cuadrada de n observaciones independientes

$$SE = \frac{\sigma}{\sqrt{n}}$$

A reliable method to sample observations is to conduct a simple random sample consisting of less than 10% of the population.

Se trata en ver el margen de error entre la media

La probability QQ plot, si está recta, nos demuestra que el dato es correcto y la muestra adecuada. **La cuestión es saber en qué momento debo aplicar el QQ plot.**

Intervalo de confianza

El punto de estimación provee un valor plausible para un parámetro (el parámetro en cuestión equivalente a la población; la media de la población - la media de la muestra; cuando estamos en el ámbito de la población ya no es una estimación sino un parámetro). Sin embargo, en un punto de estimación normalmente conlleva cierto error en la estimación. En lugar de proveer solo un punto de estimación, sería mejor plantear un **rango plausible de valores para un parámetro**.

P-value y alfa

El p-value y el alfa me va a determinar si la hipótesis es válida o más vale rechazarla.

19/04 Machine learning

- Aprendizaje supervisado: conocemos las variables y usamos el algoritmo para predecir comportamientos.
- Aprendizaje no supervisado: no tenemos las variables y se analizan los clusters para buscar dichas variables. El algoritmo busca patrones y clasifica los datos por mí.
- Aprendizaje reforzado.

REGRESIÓN LINEAR

27/04

BIG DATA SCOPE

El Big Data nace cuando el dato está sobredimensionado y excede los métodos tradicionales de analizar los datos.

La hoja de cálculo controla los problemas que pueden haber en el RP (gestión de control de toda la cadena de valor desde que entra el pedido hasta que sale). Automatizar el proceso de las hojas de cálculo. Transferencia horizontal de info entre departamentos.

Data quality manegement. Ingesta de los datos masivos que están fuera del RP. Intentar racionalizarlas: a partir de API-rest. Data lake (entorno donde trabajo como DS. Arquitectura del "lago de datos", de las bases de datos).

Dato estructurado (SQL) vs no estructurado (imágenes, etc.). Archivo raw (crudo): txt, csv, etc. Conexiones a bases de datos para transformarlo y guardar el dato crítico, racionalizado. Es necesario automatizarlo porque el flujo de dato es constante (datos dinámicos). Asegurar que la calidad del dato es persistente. Control de gestión: dashbord para monitorizar la mejora del dato.

Velocidad del proceso de ingesta y procesamiento para el uso es vital.

Cuatro Vs:

- volumen
- velocidad: dependiendo de las necesidades empresariales.
- variedad: data lake (dato estructurado y no estructurado)
- veracidad: garbage in garbage out. Sin dato de calidad no hacemos nada bueno.

La parte más cara (coste de computación) es mover el dato. Dato semiestructurado: transporte, transmisión, etc. del dato (XML, JSON).

Empezar por analizar la arquitectura de la empresa, sus capacidades, antes de realizar un proyecto de Big Data.

Dato predictivo (predecir y el ser humano decide que hacer) vs prescriptivo (la máquina realiza directamente).

PARADIGMA (EL ECOSISTEMA BIG DATA)

Extracción (de las diversas fuentes) --> Load a un data lake (ingesta) --> Transformación (Pre-procesado (s3), Exploratoria (s4), Muestreo (s5))

El dato es dinámico: **pre-procesado antes de la transformación** propiamente (datos categóricos (clasificación y cluster) a dummy variables). El pre-procesado es la preparación del dato histórico. Se almacena y se inyecta a una función que solo admite números (int, float o binarios). El pre-procesado va junto a la exploración (qué co.. es este dataset, cómo es?). El pre-procesado es quitar null values, convertir ciertos datos a otros tipos legibles (**definir el dato; si es int, float, categórico, etc.**).

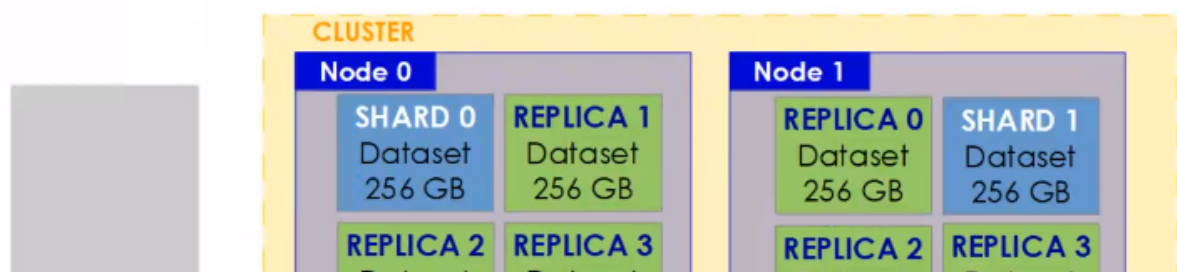
--str solo te ordena alfabéticamente, categórico te clasifica en niveles (bueno<medio<malo)--

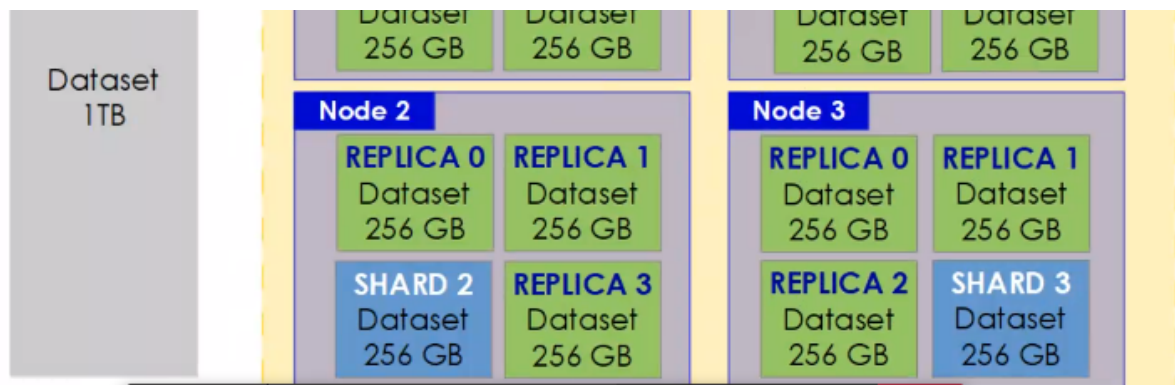
(Limitaciones agrupación de recursos Big Data. Formas de escalar el procesado almacenaje y RAM (almacenamiento) (cluster):

- Vertical scaling
- Horizontal scaling (+ recursos para trabajar en paralelo). Gestor que administra la pertenencia de un dato al cluster. Coordinar) Tomar en cuenta el crecimiento del dato en la empresa.

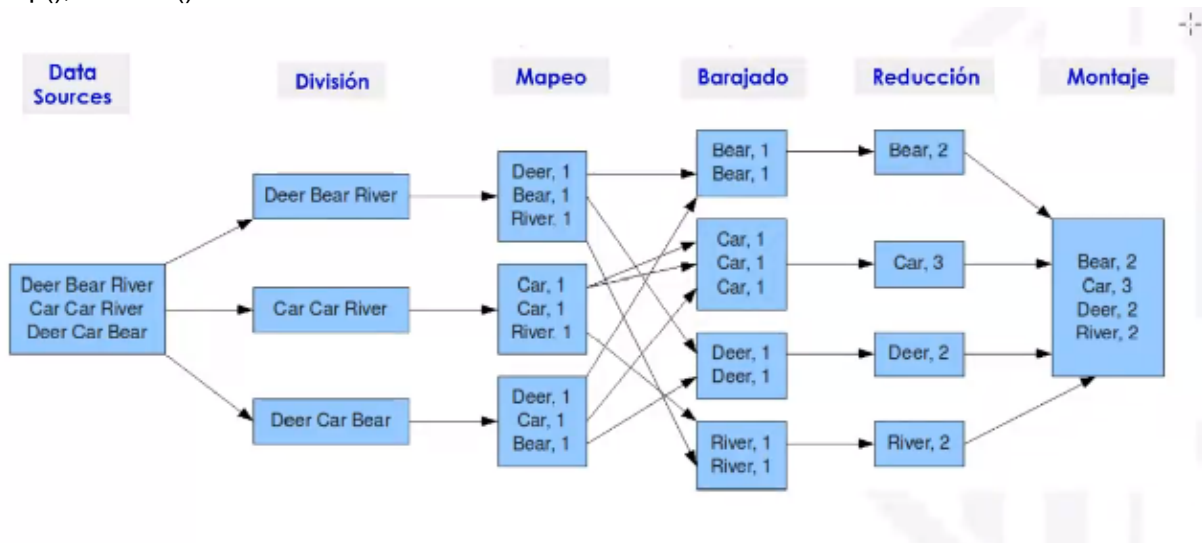
Data source. Dato almacenado vs dato transformado a tiempo real.

HDFS. Sistema de distribución de archivos open source Hadoop. Sistema distribuido de archivo (DFS), organizado a partir de nodos: NameNode (metadato de qué dato está en qué nodo) / DataNode (aquellos que guardan y albergan el dato). Tolerancia a fallos.





Fase Transformación: pasar el dato a una matriz. Normalización. Batchprocessing, MapReduce: Map(), Reduce().



Fase Análisis de datos: algoritmos.

Fase Visualización de resultados.



Para manejar todo este proceso es importante que exista alguien que lo gestione: el YARN (gestor de recursos).

MASTERCLASS ELASTICSEARCH (KIBANA)

Antes de visualizar con kibana, escribir en python (y sus librerías).

1. Instalar elasticsearch (se inicia automáticamente)
2. Instalar kibana (hay que bajar un archivo y ejecutar el kibana.bat)

Subir archivo csv. Máx 100MB. Escribir "archivo.index" para que el nombre vaya bien (importar) (minúsculas tb). En mapping cambiamos el tipo de variable (si ha detectado numérico pero sabemos que es categórico, lo podemos modificar ahí). Hacer limpieza en python antes de cargarlo.

Una vez uploaded los datos:

- Discover.
- Dashboard. Si aplico un filtro (click) en un gráfico, se actualizan todos los gráficos según ese filtro.

Cloud u otro servidor: servidor donde compartir el dashbord (si lo tienes en localhost no se puede compartir con un link).

MASTERCLASS SCRAPY

<https://docs.scrapy.org/en/latest/intro/overview.html> (<https://docs.scrapy.org/en/latest/intro/overview.html>)

Adquisición de datos de fuentes. Data estructurada a partir de una sopa HTML (texto) (dato no estructurado). Otras herramientas (además de scrapy): beautiful soup (proyectos pequeños, contenido acotado), selenium (mapeo del sitio, interactuar con los objetos web). Scrapy + Selenium (muy común). Selenium es un complemento ideal para scrapy y beautiful soup y emular el comportamiento humano. Ejemplo: comprobar cómo funciona la web ante 10000 reservas de viajes (imposible hacerlo manualmente).

Scrapy

Todo asíncrono (peticiones simultáneas a diferentes páginas web): no es secuencial. Puede ser usado para otros entornos más allá que el web (HTML). Podemos bajar los datos en archivos json o directamente a una base de datos (MongoDB p. ej.). No quiero toda la info del html, hay que sacar lo que nos interesa. Hay que estudiar un poco la estructura de la página. Ctrl + F: buscar cosis en el html. XPath: buscar algún tutorial o algo. Querys para llegar a los elementos.

Índice del taller:

1. Instalar Scrapy (dónde? En un entorno virtual (en anaconda). El entorno aísla las librerías resoecto a tu sistema operativo).
2. Crear un entorno virtual en Anaconda. Comando "shell" en la línea de comandos para practicar cosas.
3. Abrir jupyter o VSCode desde el entorno virtual (para abrir el VSCode utilizar el comando

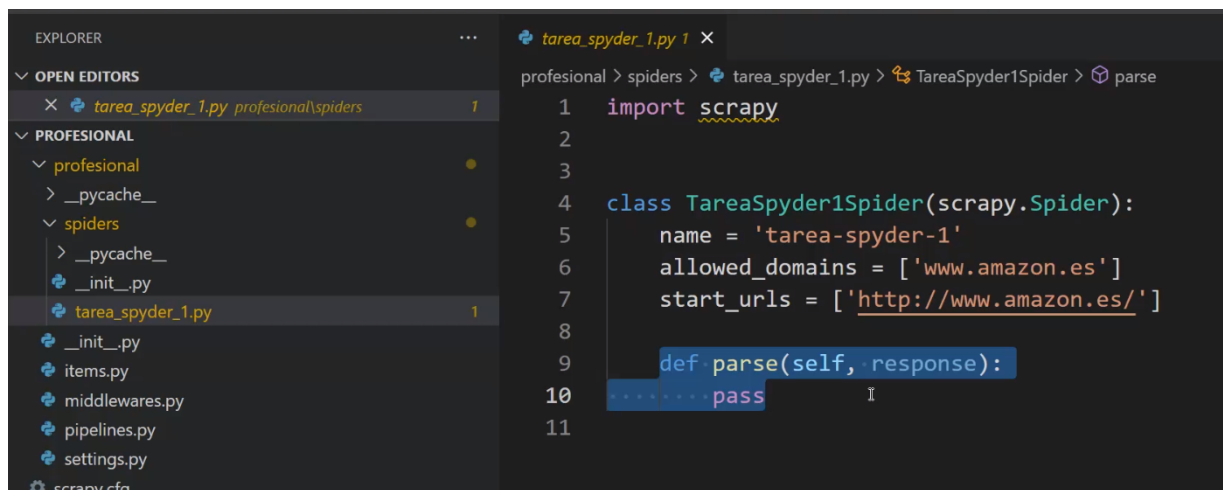
"code .").

- Comando "fetch("escribir dirección html")". Robots web que te dicen lo que puedes y no scrapear/hacer. Petición-respuesta:

```
In [1]: fetch("http://quotes.toscrape.com/page/1")
2021-07-15 17:40:54 [scrapy.core.engine] INFO: Spider opened
2021-07-15 17:40:55 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (308) to <GET http://quotes.toscrape.co
m/page/1/> from <GET http://quotes.toscrape.com/page/1/>
2021-07-15 17:40:55 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://quotes.toscrape.com/page/1/> (referer: None)
```

```
In [6]: response.css("title::text").getall()
Out[6]: ['Quotes to Scrape']
```

```
(scrapy-web) D:\Masterclass>scrapy startproject profesional
```



In []:

REGISTROS LOG

- Common log format: https://en.wikipedia.org/wiki/Common_Log_Format (https://en.wikipedia.org/wiki/Common_Log_Format)
- User agent: <https://developer.mozilla.org/es/docs/Web/HTTP/Headers/User-Agent> (<https://developer.mozilla.org/es/docs/Web/HTTP/Headers/User-Agent>)
- Apache log format: <https://www.sumologic.com/blog/apache-access-log/> (<https://www.sumologic.com/blog/apache-access-log/>)

GEOIP2

Cosas aprendidas

- Para buscar la documentación hay que referirse a la "API".
- Escribiendo help() y dentro de los paréntesis el nombre del paquete te da lo que necesitas saber.
-

NaN vs None

NaN can be used as a numerical value on mathematical operations, while None cannot (or at least shouldn't).

NaN is a numeric value, as defined in IEEE 754 floating-point standard. None is an internal Python type (NoneType) and would be more like "inexistent" or "empty" than "numerically invalid" in this context.

The main "symptom" of that is that, if you perform, say, an average or a sum on an array containing NaN, even a single one, you get NaN as a result...

In the other hand, you cannot perform mathematical operations using None as operand.

So, depending on the case, you could use None as a way to tell your algorithm not to consider invalid or inexistent values on computations. That would mean the algorithm should test each value to see if it is None.

Numpy has some functions to avoid NaN values to contaminate your results, such as `nansum` and `nan_to_num` for example.

FUENTE: <https://stackoverflow.com/questions/17534106/what-is-the-difference-between-nan-and-none> (<https://stackoverflow.com/questions/17534106/what-is-the-difference-between-nan-and-none>)

PARA EXPANDIR: https://pandas-docs.github.io/pandas-docs-travis/user_guide/missing_data.html (https://pandas-docs.github.io/pandas-docs-travis/user_guide/missing_data.html)

ENTORNO VIRTUAL

- Intro al entorno virtual python: <https://www.geeksforgeeks.org/python-virtual-environment/> (<https://www.geeksforgeeks.org/python-virtual-environment/>)
- Manejar entorno virtual conda: <https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html> (<https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html>)
- Jupyter + entorno virtual: <https://medium.com/@eleroy/jupyter-notebook-in-a-virtual-environment-virtualenv-8f3c3448247> (<https://medium.com/@eleroy/jupyter-notebook-in-a-virtual-environment-virtualenv-8f3c3448247>)

In []:

In []:

In []:

CONCEPTOS DE PROBABILIDAD Y ESTADÍSTICA

https://machinelearningknowledge.ai/tutorial-numpy-mean-numpy-median-numpy-mode-numpy-standard-deviation-in-python/#Numpy_Mode (https://machinelearningknowledge.ai/tutorial-numpy-mean-numpy-median-numpy-mode-numpy-standard-deviation-in-python/#Numpy_Mode)

Moda: la moda es el valor con mayor frecuencia en una de las distribuciones de datos.

Desviación típica/estándar: En estadística, la desviación típica es una medida que se utiliza para cuantificar la variación o la dispersión de un conjunto de datos numéricos. Una desviación estándar baja indica que la mayor parte de los datos de una muestra tienden a estar agrupados cerca de su media, mientras que una desviación estándar alta indica que los datos se extienden sobre un rango de valores más amplio. **Fórmula:** La primera es elevando al cuadrado las desviaciones, dividir entre el número total de observaciones y por último hacer la raíz cuadrada para deshacer el elevado al cuadrado, tal que:

$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{N}}$$

Rango: El Rango es el intervalo entre el valor máximo y el valor mínimo.

Cuartiles: Los cuartiles son valores que dividen una muestra de datos en cuatro partes iguales: 25% / 50% / 75% / 100%. <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/graphs/how-to/boxplot/interpret-the-results/quartiles/> (<https://support.minitab.com/es-mx/minitab/18/help-and-how-to/graphs/how-to/boxplot/interpret-the-results/quartiles/>)

Valor P: El valor p es una medida de probabilidad empleada para hacer pruebas de hipótesis. El objetivo de una prueba de hipótesis es determinar si hay evidencia suficiente para apoyar una determinada hipótesis sobre los datos. De hecho, formulamos dos hipótesis: la hipótesis nula y la hipótesis alternativa. En el análisis de correlación, usualmente, la hipótesis nula expresa que la relación observada entre las variables es producto del mero azar (esto es, que el coeficiente de correlación en realidad es cero y no hay una relación lineal). La hipótesis alternativa expresa que la correlación que hemos medido está legítimamente presente en nuestros datos (esto es, que el coeficiente de correlación es distinto a cero).

El valor p es la probabilidad de observar un coeficiente de correlación distinto a cero en los datos de nuestra muestra cuando en realidad la hipótesis nula es verdadera. Un valor p bajo nos lleva a rechazar la hipótesis nula. Un umbral típico para rechazar la hipótesis nula es un valor p de 0,05. Esto es, si el valor p es inferior a 0,05, rechazaríamos la hipótesis nula en favor de la hipótesis alternativa: que el coeficiente de correlación es diferente a cero.

Coeficiente de correlación (https://www.jmp.com/es_mx/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html#404f1893-ae56-43ed-b84c-f6c99f313eca)

(https://www.jmp.com/es_mx/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html#404f1893-ae56-43ed-b84c-f6c99f313eca): El coeficiente de correlación r es un valor sin unidades **entre -1 y 1**. La significancia estadística se indica con un valor p . Por lo tanto, usualmente las correlaciones se escriben con dos números clave: $r = p =$.

- Cuanto más se aproxima r a cero, más débil es la relación lineal.
- Los valores de r positivos indican una correlación positiva, en la que los valores de ambas variables tienden a incrementarse juntos.
- Los valores de r negativos indican una correlación negativa, en la que los valores de una variable tienden a incrementarse mientras que los valores de la otra variable descienden.
- Los valores 1 y -1 representan una correlación "perfecta" positiva y negativa, respectivamente. Dos variables perfectamente correlacionadas cambian conjuntamente a una tasa fija. Decimos que tienen una relación lineal; cuando representados en un gráfico de dispersión, todos los puntos correspondientes a los datos pueden conectarse con una misma línea recta.
- El valor p nos ayuda a determinar si podemos o no concluir de manera significativa que el coeficiente de correlación de la población es diferente a cero, basándonos en lo que observamos en la muestra.

Fórmula:

$$r = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$

Fórmula anotada:

The diagram illustrates the components of the correlation coefficient formula with the following annotations:

- Sample Correlation Coefficient:** Points to the variable r .
- Summation: "Take The Sum Of":** Points to the summation symbol \sum .
- Value of X:** Points to the variable x_i .
- Mean of X Variable:** Points to the mean symbol \bar{x} .
- Value of Y:** Points to the variable y_i .
- Mean of Y Variable:** Points to the mean symbol \bar{y} .
- Sum of the squared deviations for X:** Points to the term $\sum (x_i - \bar{x})^2$.
- Sum of the squared deviations for Y:** Points to the term $\sum (y_i - \bar{y})^2$.

$$r = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$

Square Root

De hecho, es importante recordar que basarse exclusivamente en el coeficiente de correlación puede llevar a errores, especialmente en situaciones con relaciones curvilíneas o valores extremadamente atípicos. Un coeficiente de correlación nulo o cerca de cero no necesariamente implica que no haya relación entre las variables, solamente significa que **no hay una relación lineal**. De manera similar, observar un gráfico de dispersión puede aportarnos información sobre cómo los valores atípicos (observaciones poco habituales dentro de nuestros datos) pueden sesgar el coeficiente de correlación.

$$r = \frac{S_{XY}}{S_X S_Y}$$

¿De donde sacamos estos valores?

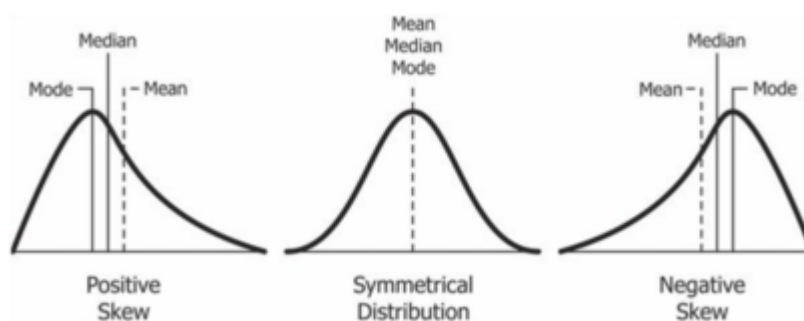
S_{XY} = Covarianza

$S_X S_Y$ = Desviación Estandar de X multiplicada por la Desviación Estandar de Y



<https://platzi.com/tutoriales/1269-probabilidad-estadistica/2308-coeficiente-de-correlacion-que-es-y-para-que-sirve/> (<https://platzi.com/tutoriales/1269-probabilidad-estadistica/2308-coeficiente-de-correlacion-que-es-y-para-que-sirve/>)

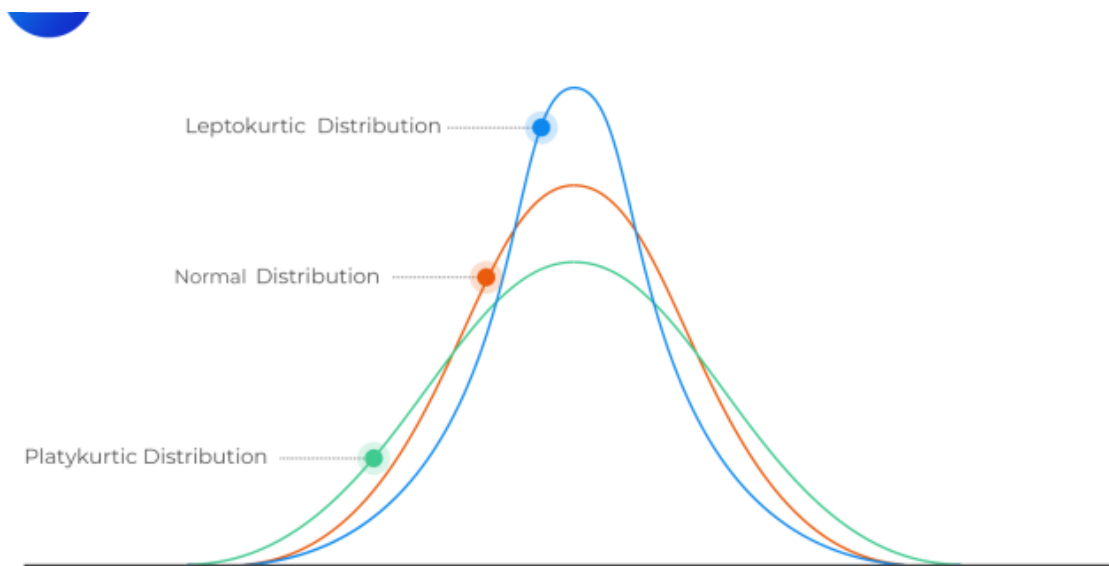
Skew:



Kurtosis:



Kurtosis



SPRINT PROBABILIDAD Y ESTADÍSTICA

<https://www.mathsisfun.com/data/> (<https://www.mathsisfun.com/data/>)

Estadística

Previo:

- Sigma:

The handy Sigma Notation says to sum up as many terms as we want:

start at this value → 1
go to this value → 4
what to sum → n

$$\sum_{n=1}^4 n = 1+2+3+4 = 10$$

Sigma Notation

Media

- Media (aritmética). La de toda la vida.
- Weighted mean: valores que tienen más o menos peso (importancia) que otros

$$\text{Weighted Mean} = \frac{\sum wx}{\sum w}$$

- Media geométrica:

$$\sqrt[n]{a_1 \times a_2 \times \dots \times a_n}$$

- Media armónica:

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \dots}$$

Media, mediana y moda en una tabla de frecuencias

For grouped data, we cannot find the exact Mean, Median and Mode, we can only give estimates.

- Media de una tabla de frecuencias

To estimate the **Mean** use the **midpoints** of the class intervals:

$$\text{Estimated Mean} = \frac{\text{Sum of (Midpoint} \times \text{Frequency)}}{\text{Sum of Frequency}}$$

- Mediana

To estimate the **Median** use:

$$\text{Estimated Median} = L + \frac{(n/2) - B}{G} \times w$$

where:

- **L** is the lower class boundary of the group containing the median
- **n** is the total number of data
- **B** is the cumulative frequency of the groups before the median group
- **G** is the frequency of the median group
- **w** is the group width

- Moda

To estimate the **Mode** use:

$$\text{Estimated Mode} = L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times w$$

where:

- **L** is the lower class boundary of the modal group
- f_{m-1} is the frequency of the group before the modal group
- f_m is the frequency of the modal group
- f_{m+1} is the frequency of the group after the modal group
- **w** is the group width

Difusión

- Cuartiles, Interquartil range, percentil, etc.
- Mean deviation: How far, on average, all values are from the middle.

$$\text{Mean Deviation} = \frac{\sum |x - \mu|}{N}$$

- Σ is Sigma, which means to sum up
- $||$ (the vertical bars) mean Absolute Value, basically to ignore minus signs
- x is each value (such as 3 or 16)
- μ is the mean (in our example $\mu = 9$)
- N is the number of values (in our example $N = 8$)
- Standard deviation: Deviation just means how far from the normal. The Standard Deviation is a measure of how spread out numbers are. Its symbol is σ (the greek letter sigma). It is the square root of the **Variance**.
- Variance: The average of the squared differences from the Mean. Se calcula diferente según se haga sobre la población o sobre una muestra:

- The Population: divide by N when calculating Variance (like we did)
- A Sample: divide by $N-1$ when calculating Variance

The "**Population** Standard Deviation":
$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

The "**Sample** Standard Deviation":
$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Looks complicated, but the important change is to divide by **$N-1$** (instead of **N**) when calculating a Sample Variance.

Comparing data

- outliers have the biggest effect on the mean, and not so much on the median or mode. Hint: calculate the median and mode when you have outliers.

PROBABILIDAD

General

- Sample Space: all the possible outcomes of an experiment. The Sample Space is made up of Sample Points.
- Outcome: A possible result of an experiment.
- Event: one or more outcomes of an experiment.
- Basic counting principle (only independent): When there are m ways to do one thing, and n ways to do another, then there are $m \times n$ ways of doing both. Example: you have 3 shirts and 4 pants. That means $3 \times 4 = 12$ different outfits.
- Relative Frequency: How often something happens divided by all outcomes.

Event

- Complement: lo contrario a un evento. Cuando el evento es cara, el complemento será cruz.
- Events can be: Independent , Dependent , Mutually Exclusive.

Combinations and permutations

- When the order doesn't matter, it is a Combination. When the order does matter it is a Permutation.
- Permutación con repetición:

Example: in the lock above, there are 10 numbers to choose from (0,1,2,3,4,5,6,7,8,9) and we choose 3 of them:

$$10 \times 10 \times \dots (3 \text{ times}) = 10^3 = 1.000 \text{ permutations}$$

- Permutación sin repetición: Without repetition our choices get reduced each time.

$$\frac{n!}{(n-r)!}$$

where ***n*** is the number of things to choose from,
and we choose ***r*** of them,
no repetitions,
order matters.

- Combinación sin repetición:

$$\frac{n!}{r!(n-r)!}$$

- Combinación con repetición:

$$\binom{r+n-1}{r} = \frac{(r+n-1)!}{r!(n-1)!}$$

where ***n*** is the number of things to choose from,
and we choose ***r*** of them
repetition allowed,
order doesn't matter.

- Teorema de Bayes:

$$P(A|B) = \frac{P(A) P(B|A)}{P(B)}$$

Which tells us: how often A happens *given that B happens*, written ***P(A|B)***,
When we know: how often B happens *given that A happens*, written ***P(B|A)***
and how likely A is on its own, written ***P(A)***
and how likely B is on its own, written ***P(B)***

$$P(B) P(A|B) = P(A) P(B|A)$$

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)}$$

We just saw "A" with two cases (A and not A), which we took care of in the bottom line.

When "A" has 3 or more cases we include them all in the bottom line:

$$P(A1|B) = \frac{P(A1)P(B|A1)}{P(A1)P(B|A1) + P(A2)P(B|A2) + P(A3)P(B|A3) + \dots \text{etc}}$$

- **Intervalo de confianza:** A Confidence Interval is a range of values we are fairly sure our true value lies in.

The Confidence Interval is based on Mean and Standard Deviation. Its formula is:

$$\bar{X} \pm Z \frac{s}{\sqrt{n}}$$

Where:

- \bar{X} is the mean
- Z is the Z-value from the table below
- s is the standard deviation
- n is the number of observations

Confidence Interval	Z
80%	1,282
85%	1,440
90%	1,645
95%	1,960
99%	2,576
99,5%	2,807
99,9%	3,291

Dependiendo del intervalo de confianza que deseemos, elegimos un Z-value determinado: para un 80% de I. de confianza hemos de utilizar un z-Value=1,282, etc. El z-value es el número de desviaciones estándar respecto a la media. Fórmula para el z-value:

$$z = \frac{x - \mu}{\sigma}$$

- z is the "z-score" (Standard Score)
 - x is the value to be standardized
 - μ ('mu') is the mean
 - σ ("sigma") is the standard deviation
- **P-Value:** "p" is the probability the variables are independent. Si el p-value es inferior a 0,05/0,01(para estar más seguros), significa que las dos variables son dependientes. Este método (Chi Test) solo sirve para datos categóricos. Para llevarlo a cabo, es necesario tener una **hipótesis** (es decir, una aserción que puede ser verdadera o falsa. P. ej: Gender and preference for cats or dogs are independent. / Gender and preference for cats or dogs are not independent). Se puede calcular el p-value mediante el chi-test, cuya formula es:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- Σ means to sum up (see [Sigma Notation](#))
- O = each **Observed** (actual) value
- E = each **Expected** value

Para calcular el expected value solo hace falta multiplicar el total de la fila x la columna de la observación en cuestión y dividirlo por la suma total de filas y columnas. A través de esta pág. web podemos saber el p-value una vez sepamos el chi y el degree of freedom:

<https://www.mathsisfun.com/data/chi-square-calculator.html> (<https://www.mathsisfun.com/data/chi-square-calculator.html>) El degree of freedom se calcula de la siguiente manera:

$$\text{Degree of Freedom} = (\text{rows} - 1) \times (\text{columns} - 1)$$

sample we have 2 rows and 2 columns:

$$DF = (2 - 1)(2 - 1) = 1 \times 1 = 1$$

- **Regresión lineal:** Muy sensible a outliers.

$$y = mx + b$$

Where:

- y = how far up

- **x** = now far along
- **m** = Slope or Gradient (how steep the line is)
- **b** = the Y Intercept (where the line crosses the Y axis)

$$m = \frac{N \sum(xy) - \sum x \sum y}{N \sum(x^2) - (\sum x)^2}$$

$$b = \frac{\sum y - m \sum x}{N}$$

Binomial Distribution: "Bi" means "two" (like a bicycle has two wheels) ... so this is about things with two results (heads or tails, yes or no, etc.). Más info si necesaria:

<https://www.mathsisfun.com/data/binomial-distribution.html> (<https://www.mathsisfun.com/data/binomial-distribution.html>)

Normal Distribution

- Estandarizar:

So to convert a value to a Standard Score ("z-score"):

- first subtract the mean,
- then divide by the Standard Deviation

And doing that is called "Standardizing":



CORRECCIONES ENTREGAS

S03 T03: Como comentario de mejora, para mirar si un array es simétrico, hubieses podido utilizar la **función reverse** de python. https://www.tutorialspoint.com/python/list_reverse.htm (https://www.tutorialspoint.com/python/list_reverse.htm).

S03 T05: Como comentario de mejora, en vez de tabla de correlaciones normalmente se dibuja un heatmap de correlaciones (y el **lower triangle del heatmap solo para no repetir información como indica este recurso** <https://stackoverflow.com/questions/57414771/how-to-plot-only-the-lower-triangle-of-a-seaborn-heatmap> (<https://stackoverflow.com/questions/57414771/how-to-plot-only-the-lower-triangle-of-a-seaborn-heatmap>)).

