Preprocesamiento & Transformación de datos

Limpieza de datos

- Datos faltantes
  - Missing completely at random MCAR
  - Missing Not At Random MNAR
  - Missing At Random MAR
- Estrategias para trabajar con datos faltantes
- Imputación de datos

Transformación de datos - Feature Engineering

- Normalización
  - Min-Max
  - Z score
  - Decimal scaling
- Discretización
  - Binning
  - Variables Dummies – One Hot Encoding
- Imaginación → Generación de nuevas variables

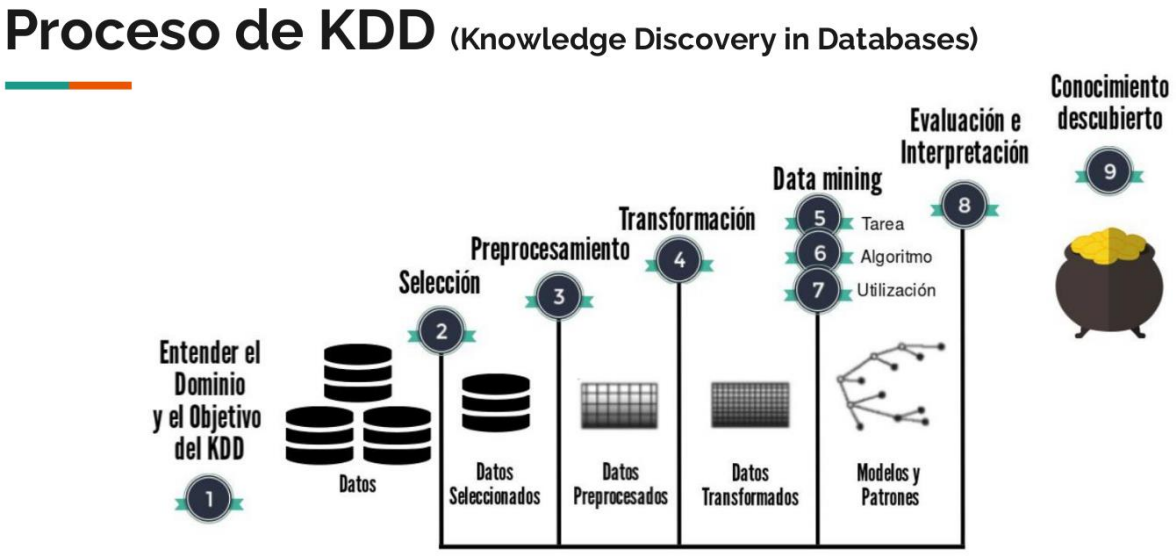
Valores atípicos

- Univariado
  - Z-Score
  - Z-Score modificado
  - Análisis de Box-Plot
- Multivariado
  - Distancia de Mahalanobis
  - LOF – Local Outlier Factor
  - Isolation Forest

Que es?

- Limpiar el dataset
- Amoldar, ajustar los datos para lo que se busca modelar
- Dividir datos de entrada para distintos usos
- Hacer algo con los NaN
- …

La idea de la transformación de datos es que forma parte de un proceso de descubrimiento del conocimiento en la ciencia de datos



Algunas tareas de estas etapas son:

● Integración de datos: Integración de múltiples bases de datos, archivos, etc.

● **Limpieza de datos:** Completar valores faltantes, eliminación de ruido, identificar o eliminar valores atípicos y corregir incoherencias

● Reducción de datos: Reducción de dimensionalidad, Reducción de Numerosidad Ej: tengo un dataset con muchas columnas que es dificil de manipular ⇒ puedo resumir la informacion en menos columnas sin perder esa info, varianza o datos relevantes

● **Transformación de datos:** Normalizaciones, generación de jerarquías conceptuales, etc. (Feature Engineering). Hacer conversiones en los datos. A veces son necesarias para usar algunos algoritmos. Muchas veces se mejora una performance en los modelos al hacer una normalizacion o estandarizacion de los datos

Limpieza de datos

Mismo dato representado de formas distintas:
1.5 / 1,5 → Inconsistencia

1m / 100cm Representa lo mismo pero esta en diferentes escalas

Type '/' for commands

Datos faltantes

No son solo Nans

Existen diferentes mecanismos de faltantes, los estándares son:

- Missing completely at random MCAR
- Missing Not At Random MNAR
- Missing At Random MAR

Missing completely at random MCAR

Los datos faltan completamente por azar

En este caso la razón de la falta de datos es ajena a los datos mismos.

No existen relaciones con la variable misma donde se encuentran los datos faltantes, o con las restantes variables en el dataset que expliquen porque faltan.

Missing Not At Random MNAR

La razón por la cual faltan los datos depende precisamente de los mismos datos que hemos recolectado (está relacionado con la razón por la que falta, con el valor de esa variable, el valor que tiene que tener esa variable)

Ej: Cada vez que una variable debería tener un valor entre 10 y 20, el mismo no se encuentra registrado (independientemente de los valores que tomen las variables restantes)
Como falta siempre, yo puedo asumir que hubo un problema cuando esa variable tenia que registrar esa medicion.

Missing At Random MAR

Punto intermedio entre los dos anteriores.

La causa de los datos faltantes no depende de estos mismos datos faltantes, pero puede estar relacionada con otras variables del dataset.

Por ejemplo: encuestas mal diseñadas
Si preguntan algun factor de mi vida que no coincide con mi realidad ⇒ puede que no lo conteste.
Si me pregunta si tengo hijos y contesto que no, y luego la siguiente pregunta es que hacen tus hijos, seguramente no va a tener respuesta.

Estrategias para trabajar con datos faltantes

Eliminar registros o variables

Si la eliminación de un subconjunto disminuye significativamente la utilidad de los datos, la eliminación del caso puede no ser efectiva (No se recomienda en situaciones que no sean MCAR)

Imputar datos

Utilizar métodos de relleno de faltantes.

Imputación de datos

Sustitución de casos

Se reemplaza con valores no observados.
Debería ser realizado por un experto en esos datos

Sustitución por Media o Mediana

Se reemplaza utilizando la medida calculada de los valores presentes.

Algunas desventajas:
- La varianza estimada de la nueva variable no es válida porque está atenuada por los valores repetidos
- Se distorsiona la distribución
- Las correlaciones que se observen estarán deprimidas debido a la repetición de un solo valor constante

Imputación Cold Deck

Selecciona valores o usa relaciones obtenidas de fuentes distintas de la base de datos actual

	lat	lon	ciudad	barrio
	-34.607269	-58.375478	Capital Federal	NaN
	-34.607269	-58.375478	Capital Federal	NaN
	-34.625486	-58.455938	Capital Federal	NaN
	-39.261539	-68.779088	Capital Federal	NaN
	-38.827752	-68.144293	Capital Federal	NaN

Fuente de datos original: Properati

Fuente para imputación: capas geográficas provistas por GCBA

Falta el dato del barrio, pero tenemos las coordenadas de latitud y longitud. Entonces, a partir de dos variables que están en el dataset podríamos imputar la que me falta. Para eso necesitamos alguna fuente mas de datos.

Imputación Hot Deck

Se reemplazan los faltantes con valores obtenidos de registros que son los más similares.

(Hay que definir que es similar, K vecinos más cercanos puede servir)

	rooms	bathrooms
214	3.0	NaN
253	3.0	NaN
383	5.0	NaN
486	3.0	NaN
927	1.0	NaN

Podría pensar en definir registros similares como otros inmuebles con la misma cantidad de ambientes (rooms), y completar la variable bathrooms con el valor más probable en ellos

Imputación por regresión

El dato faltante es reemplazado con el valor predicho por un modelo de regresión

bedrooms	surface_total	property_type	rooms
7.0	640.0	Casa	NaN
7.0	1309.0	Casa	NaN
1.0	45.0	Casa	NaN
8.0	320.0	Casa	NaN
2.0	230.0	Casa	NaN

Podría pensar en predecir la variable rooms en función de la cantidad de habitaciones, la superficie total del inmueble y el tipo de inmueble

Si queremos imputar un valor que es una categoria ⇒ Regresion Logistica

Si es un rango continuo ⇒ Regresion Lineal, que puede ser simple o multiple

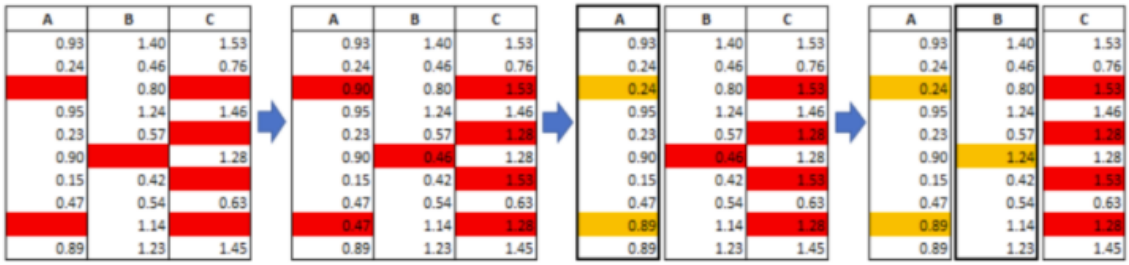
MICE - Multivariate Imputation by Chained Equations

Trabaja bajo el supuesto de que el origen de los faltantes es Missing At Random (MAR)

Es un proceso de imputación de datos faltantes iterativo, en el cual, en cada iteración cada valor faltante de cada variable se predice en función de las variables restantes.

Esta iteración se repite hasta que se encuentre convergencia en los valores.
Por lo general 10 iteraciones es suficiente.

⇒ En cada iteración genera un dataset



A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45

El método inicia el dataset original, con los datos faltantes

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.90	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.47	1.14	1.28
0.89	1.23	1.45

En cada dato faltante se imputa un valor utilizando por ej alguno de los métodos vistos antes

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45

Con las variables B y C, se genera un modelo para redecir los faltantes originales en la variable A

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	1.24	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45

Con las variables A y C, se genera un modelo para redecir los faltantes originales en la variable B

Continúa…

Transformación de datos - Feature Engineering

Esta etapa incluye cualquier proceso de modificación de la forma de los datos
(es común que los datos sufran algún tipo de transformación)

El objetivo principal de esta etapa es **mejorar el rendimiento de los modelos creados** mediante la transformación de los datos que utilizan

Algunas técnicas son:

- Normalización
- Discretización
- Lograr normalidad → A veces queremos que las variables tengan distribucion normal
- Imaginación → Generación de nuevas variables

Normalización

Estandarizar un valor

Se aplica sobre valores numéricos

Consiste en **escalar los features** de manera que puedan ser mapeados a un **rango más pequeño**.

Por ejemplo: 0 a 1 o -1 a 1

Es principalmente utilizada cuando:

- Las unidades de medidas dificultan la comparación de features.
- Se quiere evitar que atributos con mayores magnitudes tengan pesos muy diferentes al resto

Normalización - Min Max

Funciona al ver cuánto más grande es el valor actual del valor mínimo del feature y escala esta diferencia por el rango.

Consiste en tomar cada valor de la variable, restarle el minimo de toda la columna y dividirlo por el rango que es la diferencia entre valor max y min.

$$X^*_{\text{mm}} = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Los valores de normalización min-max van de 0 a 1

Normalización - Z score

Los valores para un atributo se normalizan en base a su media y desvío estándar

$$Z-score = \frac{X-mean(X)}{sd(X)}$$

Es útil cuando el verdadero mínimo y máximo del atributo no son conocidos, o cuando hay valores atípicos que dominan la normalización min-max.

Normalización - Decimal scaling

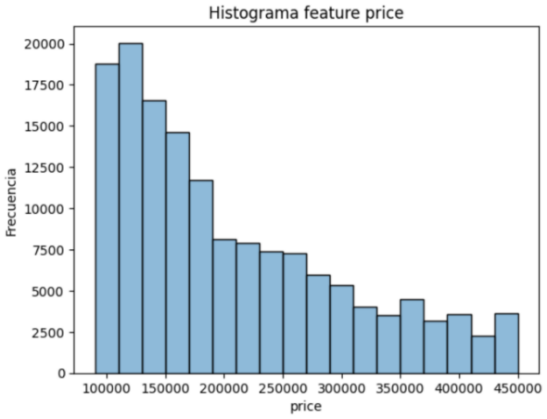
Asegura que cada valor normalizado se encuentra entre - 1 y 1.

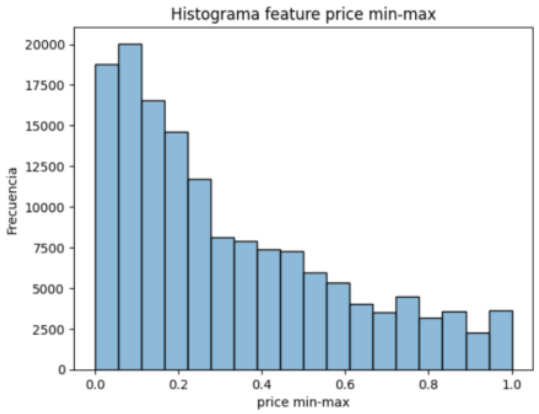
$$X_{decimal} = \frac{X}{10^d}$$

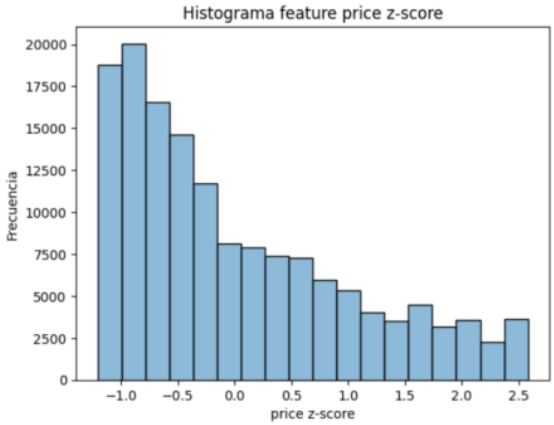
d representa el número de dígitos en los valores de la variable con el valor absoluto más grande

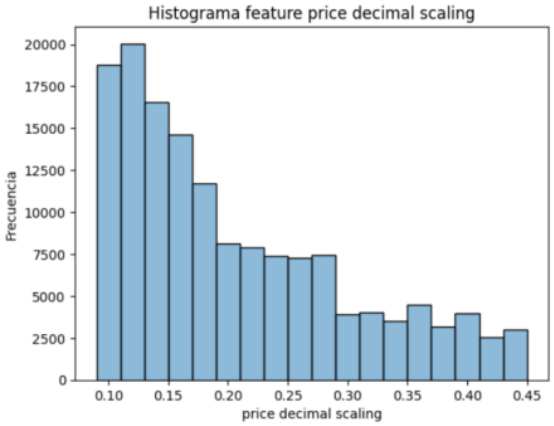
Entoncs, tomamos la variable, buscamos la que tiene el valor absoluto mas grande y la cant de digitos que tenga la usamos como potencia para el escalado decimal.

Ejemplo: propiedades y precios -> En el eje x cambio el rango.









Discretización

Es una técnica que permite dividir el rango de una variable continua en intervalos o rango discreto.

Se reducen los valores de una variable continua a un número reducido de etiquetas o categorias

Ej: edad esta en rango continuo, porque uno puede pensar en años, meses, segundos de vida. Puedo querer armar categoria: de 0 a 5 años, de 5 a 10, de 10 a ...

Llevo una variable de rango continuo a rango discreto

Discretización - Binning

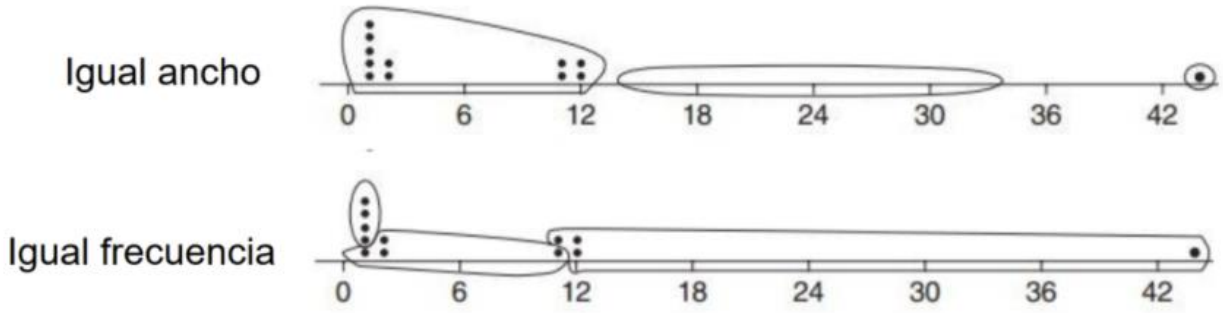
Se divide a la variable en un número específico de bins

Los criterios de agrupamiento pueden ser por ejemplo:

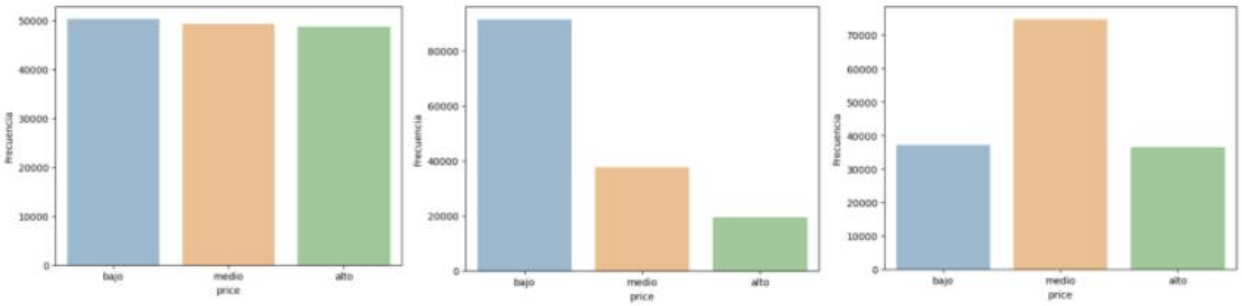
- Igual Frecuencia: La misma cantidad de observaciones en un bin
- Igual Ancho: Definimos rangos o intervalos de clases para cada bin
- Cuantiles: Separar en intervalos utilizando Mediana, Cuantiles, Percentiles.

A su vez para cada uno de los agrupamientos podemos hacer:

- Reemplazo por media o mediana
- Reemplazo por una etiqueta o valor entero



Ejemplos en el dataset de Properati



Igual frecuencia | Igual ancho de intervalo | Cuartiles

Tenemos la variable precio en un rango continuo y la queremos categorizar en tres categorias: precio bajo, medio y alto.

Igual ancho → Ej: de 0 a 30, 30 a 60, 60 a 90. Limites de los intervalos

Cuartiles → hasta 25, hasta 50, etc

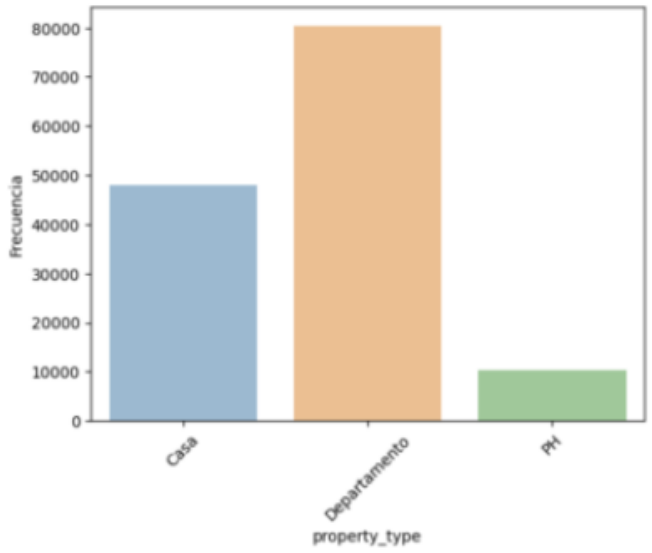
Dependiendo de la tecnica que se utilice, van a ser distintas la cantidad que haya de observaciones en cada una de las categorias

Variables Dummies – One Hot Encoding

Algunos métodos analíticos requieren que las variables predictoras sean numéricas

Cuando tenemos categóricos, podemos recodificar la variable categórica en una o más variables Dummies

Genero tantas variables nuevas como categorias tenga mi variable



nombre	casa	depto	ph
carlos	1	0	0
juan	0	1	0
romi	0	1	0

nombre	casa	depto
carlos	1	0
juan	0	1
mica	0	0

Imaginación → Generación de nuevas variables

Feature Engineering también es crear variables nuevas

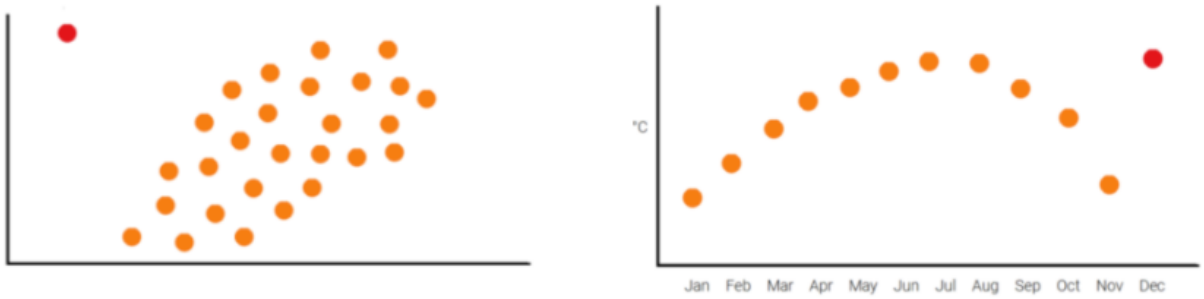
Por ejemplo: Sumar fuentes de información para calcular la distancia desde un inmueble en venta al espacio verde más cercano. Quiero probar si la cercania de espacios verdes tiene algun tipo de influencia en el precio de la propiedad.

Análisis de valores atípicos

Los outliers

- Valores que se alejan mucho de la media
- Un valor que a priori no parece estar cercano a lo que medimos
- Valores que no se distribuyen de manera similar a los demás
- Cosas que no siguen el patron de los datos

Un outlier es una observación que se desvía tanto de las otras observaciones como para despertar sospechas que fue generado por un mecanismo diferente



● Es un concepto subjetivo al problema.
Depende de lo que yo estoy estudiando, del dominio del problema, el area en la que estoy trabajando

● Son observaciones distantes del resto de los datos

● Pueden deberse a un error de medición, aleatoriedad, que esa instancia pertenezca a una familia distinta del resto, etc

Metodos univariados

Miramos en una variable o una dimension.

Ejemplo: Buscamos outliers en la columna edad → Una persona de 120 años de edad

Metodos multivariados

Vemos dos o mas dimensiones de nuestros datos.

Ejemplo 1: peso y altura en conjunto, y vemos si esa combinacion de valores es un outlier, veo si esta fuera del rango de lo que yo asumo como normal en mis datos

Ejemplo 2: Una persona de 4 años que mide 1.80mts
Esto lo vemos combinando las dos columnas, si haciamos analisis univariado no lo ibamos a detectar este dato.

with outlier	without outlier
Mean: 20.08	Mean: 12.72
Median: 14.0	Median: 13.0
Mode: 15	Mode: 15
Variance: 614.74	Variance: 21.28
Std dev: 24.79	Std dev: 4.61

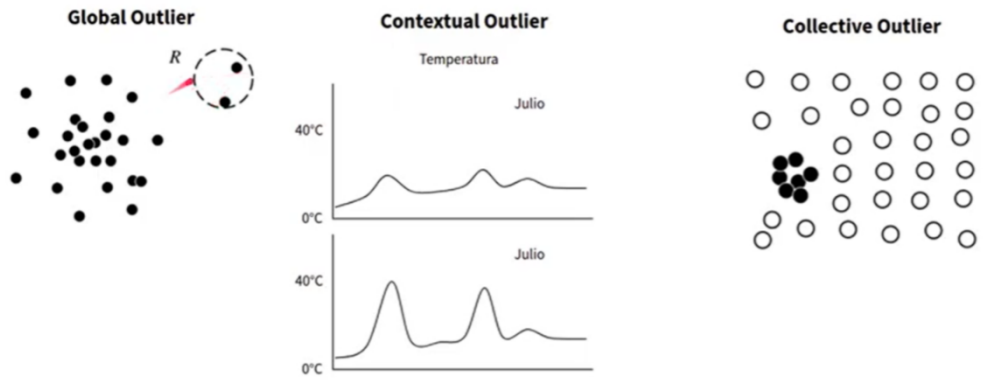
La detección de outliers es importante → su presencia puede influenciar los resultados de un análisis estadístico clásico.

Si queremos calcular media, correlacion ⇒ los valores que son atipicos o estan muy desviados me afectan la medicion o el analisis estadistico

¿Es necesario eliminarlos?

Uno lo que quiere es identificar los outliers, no eliminarlos. Los buscamos para entenderlos. Después, dependiendo de lo que yo entienda de los datos puedo borrarlos o puedo estudiarlos en profundidad

- Deben ser cuidadosamente inspeccionados
- Pueden estar alertando anomalías, en algunas situaciones nuestra tarea de interés será encontrarlos:
  - Detección de Fraudes
  - Detección de Fallas
  - Patologías Médicas



Tipos de outlier

Univariado

- Son valores atípicos que podemos encontrar en una simple variable.
- El problema de los enfoques univariados es que son buenos para detección de extremos pero no en otros casos.

Multivariado

- Los valores atípicos multivariados se pueden encontrar en un espacio n-dimensional.
- Para detectar valores atípicos en espacios n-dimensionales es necesario ajustar un modelo.

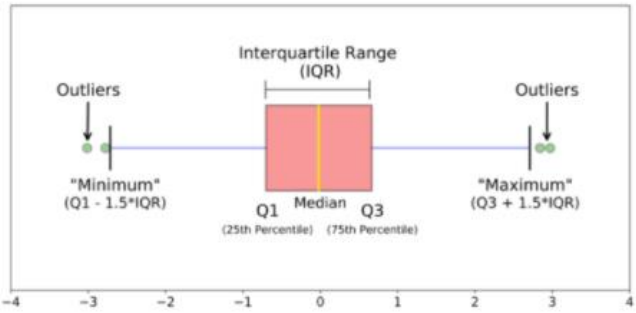
En grandes volúmenes de datos la detección de outliers resulta más eficiente estudiando todas las variables.

Los outliers, en casos multivariados, pueden provocar dos tipos de efectos:

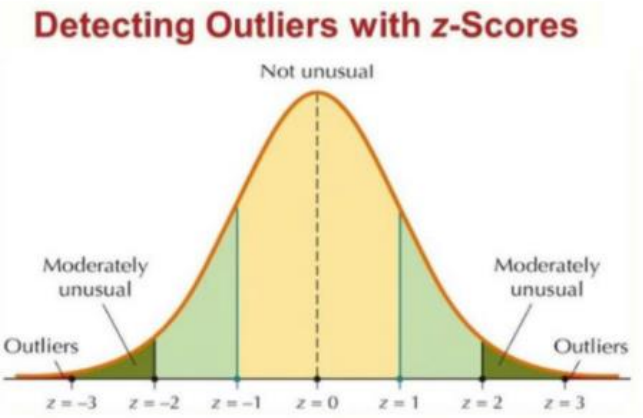
- El **efecto de enmascaramiento** se produce cuando un grupo de outliers esconden a otro/s. Es decir, los outliers enmascarados se harán visibles cuando se elimine/n el o los outliers que los esconden.
- El **efecto de inundación** ocurre cuando una observación sólo es outlier en presencia de otra/s observación/es. Si se quitara/n la/s última/s, la primera dejaría de ser outlier.
  Ej del niño de 4 años que mide 1,80mts → si saco la edad el 1,80 no va a ser outlier

Métodos Univariados

- IQR: Analizar los valores que están por fuera del IRQ



- Z-score y Z-score Modificado

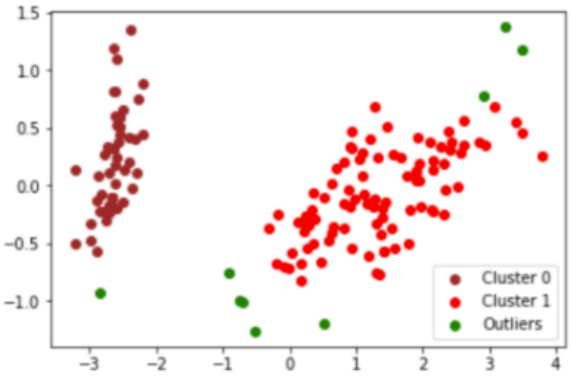


- Identificar valores extremos a partir de 1, 2 o 3 desvíos de la media.

Métodos Multivariados

- Análisis globales: Clustering

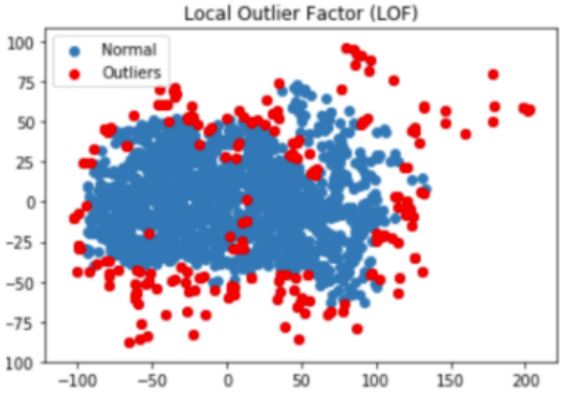
Utilizando medidas de distancia como Mahalanobis.
Los valores similares son agrupados y los que quedan aislados pueden ser considerados outliers.



- Local Outlier Factor (LOF)

Es un método de detección de outliers basado en distancias.

Calcula un score de outlier a partir de una distancia que se normaliza por densidad.



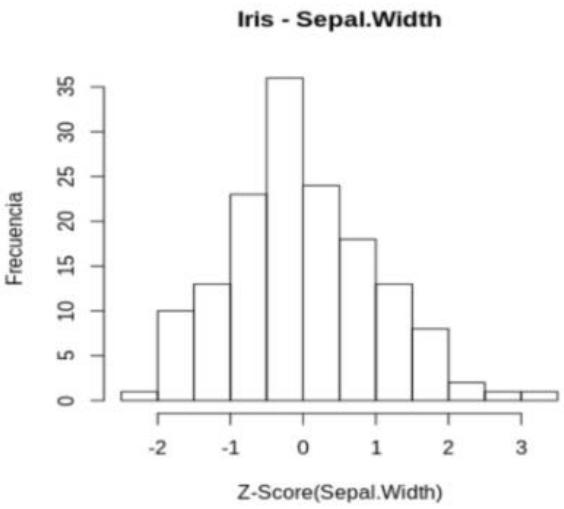
- Métodos basados en árboles de búsqueda: IsolationForest

Métodos Univariados

Z-Score

Z-Score es una métrica que indica cuántas desviaciones estándar tiene una observación de la media muestral, asumiendo una distribución gaussiana.

$$z_i = \frac{x_i - \mu}{\sigma}$$



Cuando calculamos Z-Score para cada muestra debemos fijar un umbral:

- "**regla de oro**" todo los valores Z > 3 y Z < -3 son outliers

Z-Score Modificado

La media de la muestra y la desviación estándar de la muestra, pueden verse afectados por los valores extremos presentes en los datos

Es una medida mas robusta

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

Median Absolute Deviation

$$MAD = median\{|x_i - \tilde{x}|\}$$

Regla de oro:
valores mayores a 3.5 son considerados outliers

Es la mediana de los desvíos absolutos respecto de la mediana.

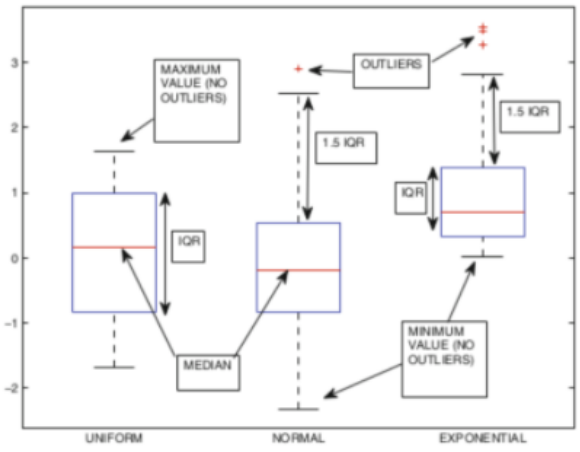
Para hacer MAD comparable con la desviación estándar, se normaliza por 0.6745

Análisis de Box-Plot

Los Box-Plots permiten visualizar valores extremos univariados.

Las estadísticas de una distribución univariada se resumen en términos de cinco cantidades:

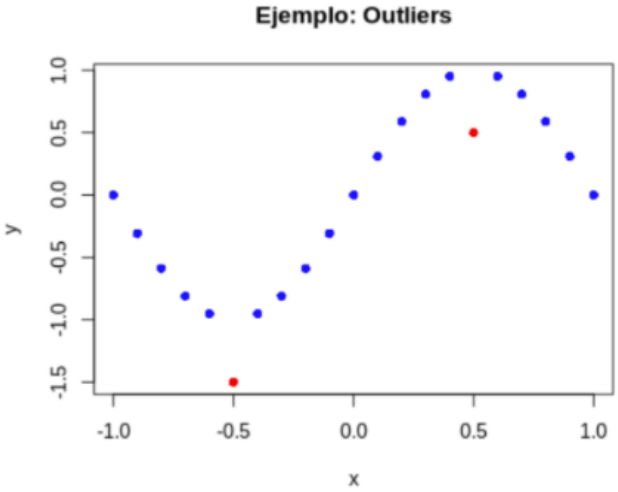
- Mínimo/máximo (bigotes)
- Primer y tercer cuantil (caja)
- Mediana (línea media de la caja)
- IQR = Q3 - Q1



Generalmente la regla de decisión:

+/- 1.5 * IQR **Outliers moderados**

+/- 3 * IQR **Outliers severos**



En el scatter se observan dos valores atípicos.

Pero en el boxplot estan dentro de los valores esperados.

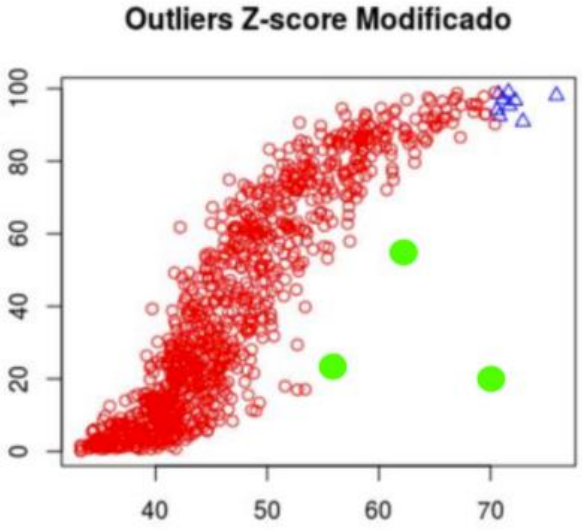
Problema

Una forma de tratar valores atípicos es eliminar los valores más altos y más bajos de una variable.

Esto puede funcionar bastante bien, pero no tiene en cuenta las combinaciones de variables.

Por ej: Los puntos verdes estan dentro de los rangos esperados de la variable, pero no siguen este mecanismo de generacion, la relacion que hay entre ellas. Son outliers, pero no son valores extremos.

No es el caso de edad 120, sino el del niño de 4 años como 1.80 mts



Métodos Multivariados

Distancia de Mahalanobis

Es una medida de distancia entre el punto $\vec{x} = (x_1, x_2, x_3, \ldots, x_N)^T$

y un conjunto de observaciones con media $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$

y una matriz de covarianza S.

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$

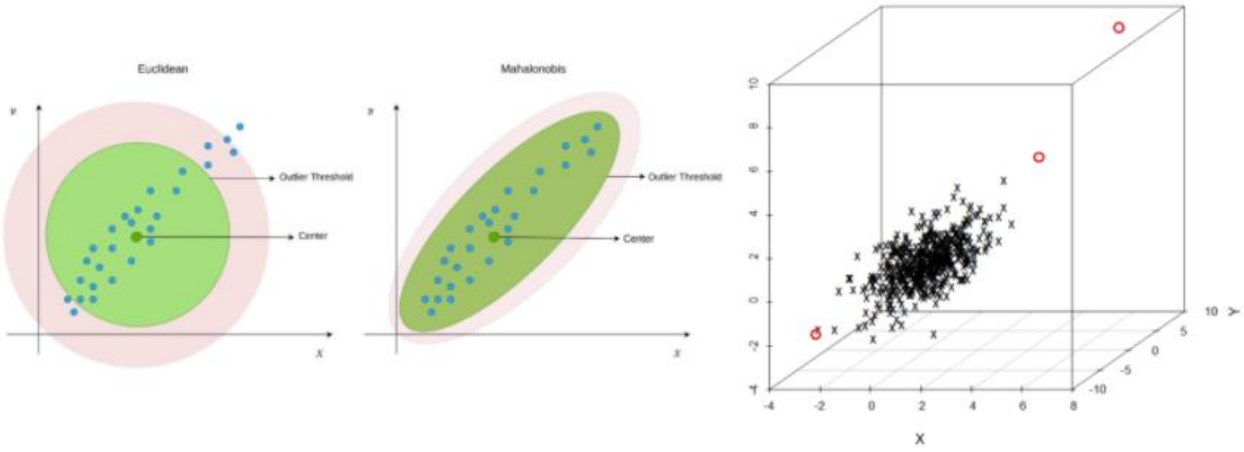
Matriz de distancias con respecto a la media

Inversa de la matriz de covarianzas

Vector x es una fila del dataset, y los x1,..,xn son los valores que tienen las columnas, las variables.

Vector u tiene la media de cada una de las columnas

Ej: en x tengo (edad, peso, altura) y en u las medias de eso



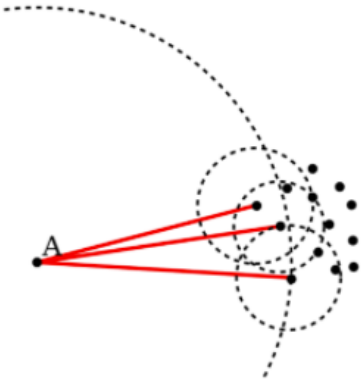
- En la Euclidea la distancia es en **circunferencia**
- En Mahalanobis la distancia es en **elipse**

LOF – Local Outlier Factor

El método LOF valora puntos en un conjunto de datos multivariados.

Es un método basado en densidad que utiliza la búsqueda de vecinos más cercanos.
→ **Tomar los puntos y medir la densidad de sus vecinos**

- Se compara la densidad de cualquier punto de datos con la densidad de sus vecinos
- Parámetro k (cantidad de vecinos) y métrica de distancia

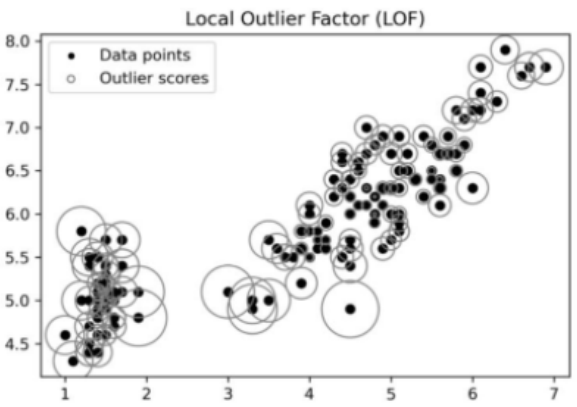


El método calcula los **scores para cada punto**, se debe definir un **umbral de corte** (depende del dominio)

- Si el score del punto X es 5, significa que la densidad promedio de los vecinos de X es 5 veces mayor que su densidad local

El radio de la circunferencia es el score del punto

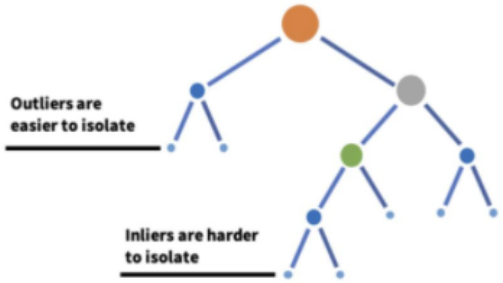
Las circunferencias mas grandes van a ser los outliers. Y eso no iene que ver con que los grupos esten separados, sino que el punto esta separada de alguna nube



Isolation Forest

Es un algoritmo no supervisado y no paramétrico basado en árboles de decisión.

- Idea Principal: los datos anómalos se pueden aislar de los datos normales mediante particiones recursivas del conjunto de datos.
  → Partir el valor de la variable y ver cuales quedan de un lado y cuales del otro



Tomar una muestra de los datos y construir un árbol de aislamiento:

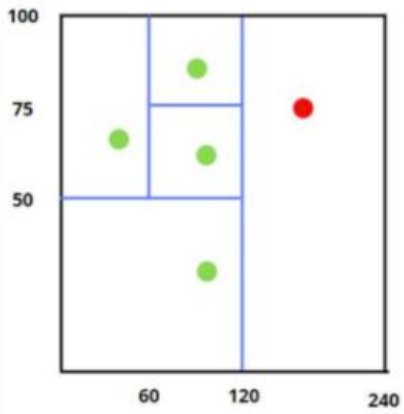
1. Seleccionar aleatoriamente n características .

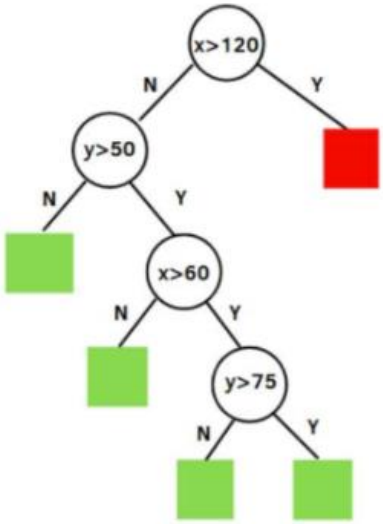
2. Dividir los puntos de datos seleccionando aleatoriamente un valor entre el mínimo y el máximo de las características seleccionadas.

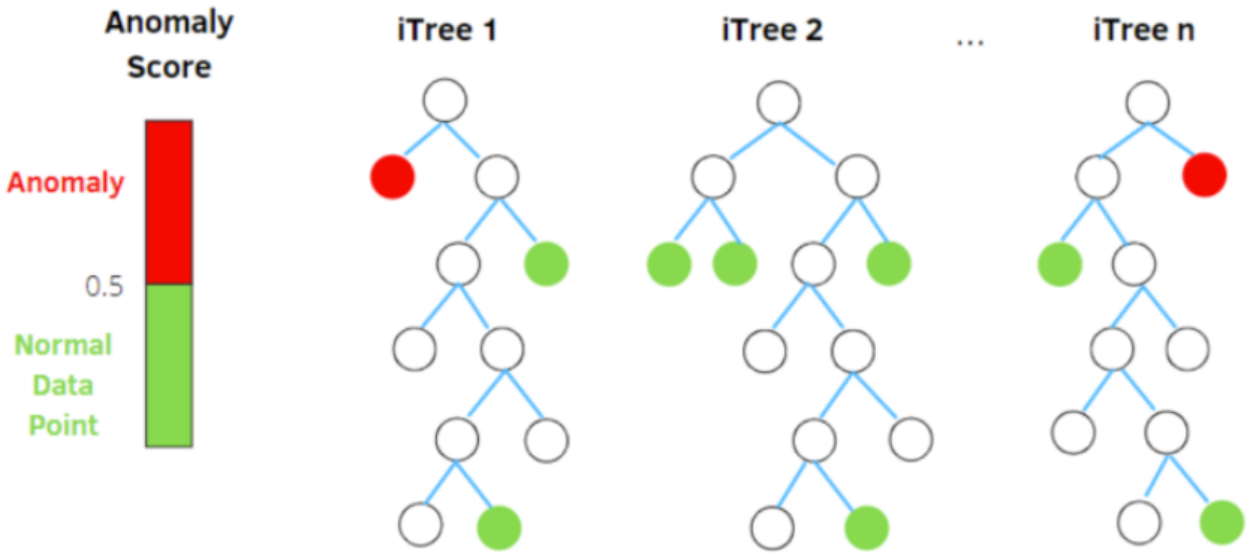
La partición de observaciones se repite recursivamente hasta que todas las observaciones estén aisladas.

Isolation Forest identifica anomalías como las observaciones con longitudes de ruta promedio cortas en los árboles de aislamiento.

• Utiliza la altura del árbol (cantidad de aristas)







Varias arboles considerando distintos pares de variables. Me fijo cuantas veces la observacion quedo aislada tempranamente en cada uno de los arboles.

Cuantas mas veces se encuentre en distintos arboles y cuantas mas veces este cerca de la raiz ⇒ es un outlier.