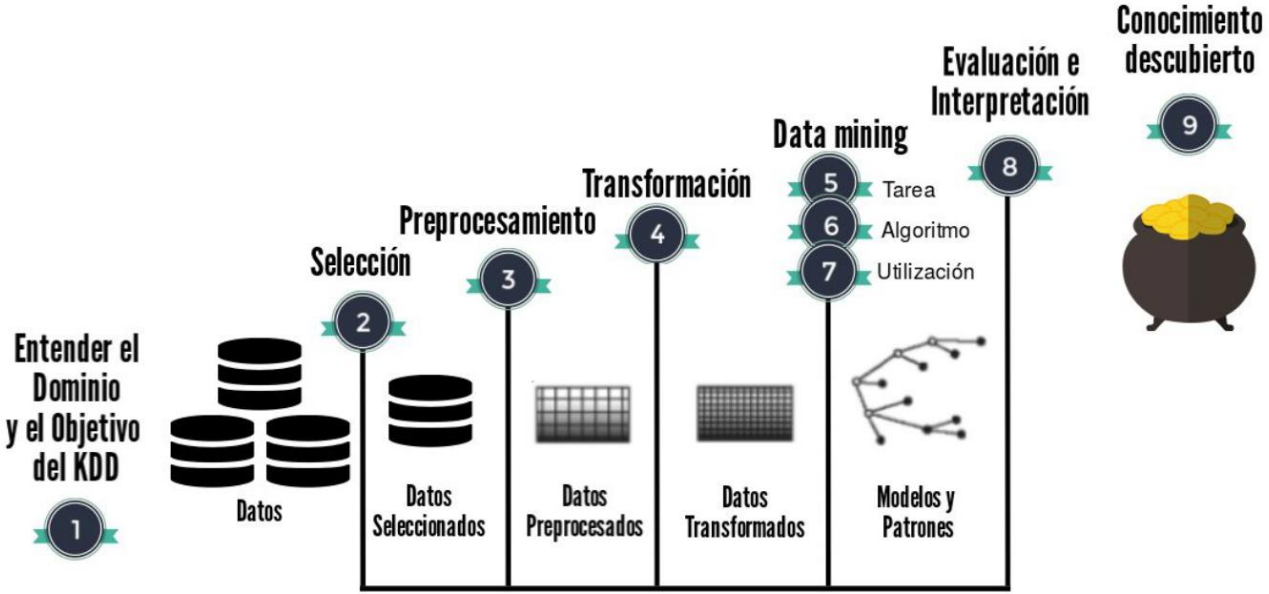
Preprocesamiento
&
Transformación de datos

75.06 / 95.58 Organización de Datos
Cat. Ing. Juan M. Rodríguez

Contenidos seleccionados del curso de Posgrado DM- UBA

Proceso de KDD (Knowledge Discovery in Databases)



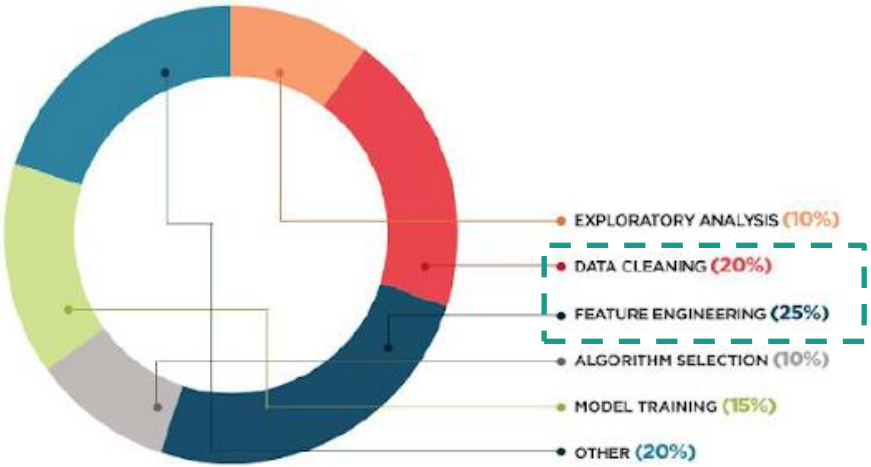
Hoy nos vamos a enfocar en las etapas de Preprocesamiento y Transformación

Preprocesamiento & Transformación

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
- ***Transformación de datos***: Normalizaciones, generación de jerarquías conceptuales, etc. (Feature Engineering)

Preprocesamiento & Transformación



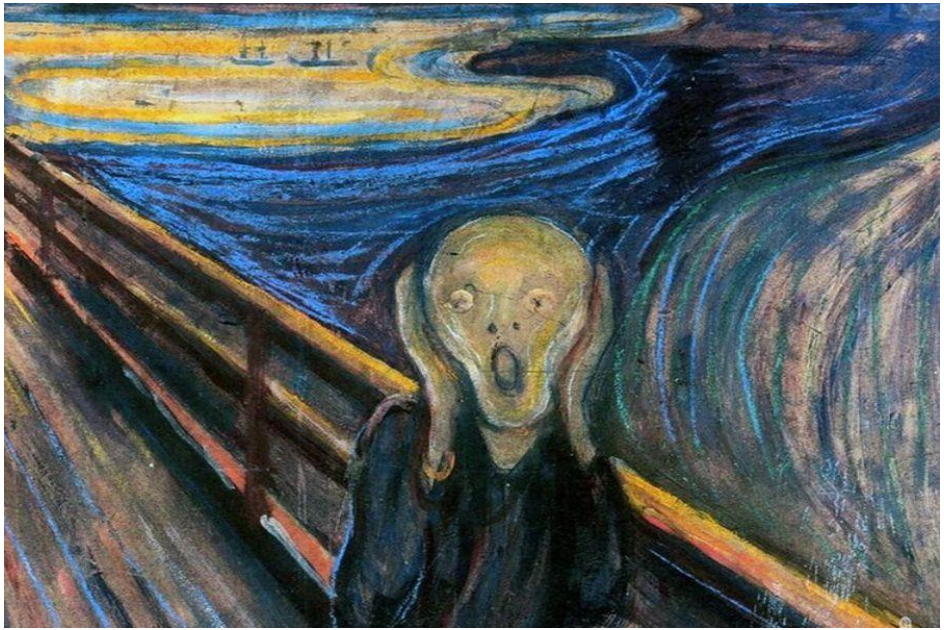
Hoy nos vamos a enfocar en las etapas de Preprocesamiento y Transformación

Limpieza de datos

Limpieza de datos

Y una lista infinita…

Limpieza de datos



Limpieza de datos

Datos faltantes

Tipos de datos faltantes

No sólo Nans!

Existen diferentes mecanismos de faltantes, los estándares son:

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

Tipos de datos faltantes

Missing completely at random MCAR

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.

Tipos de datos faltantes

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)

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)

Tipos de datos faltantes

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

Estratégias 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:

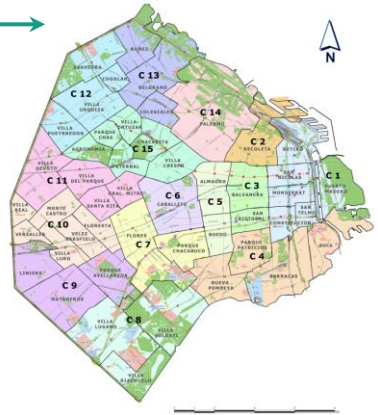
- La varianza estimada de la nueva variable no es válida porque está atenuada por los valores repetidos
- Se distorsiona la distribución
- Las correlaciones que se observen estarán deprimidas debido a la repetición de un solo valor constante

Imputación de datos

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

Imputación 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

Fuente de datos original: Properati

Imputación de datos

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

Imputación de datos

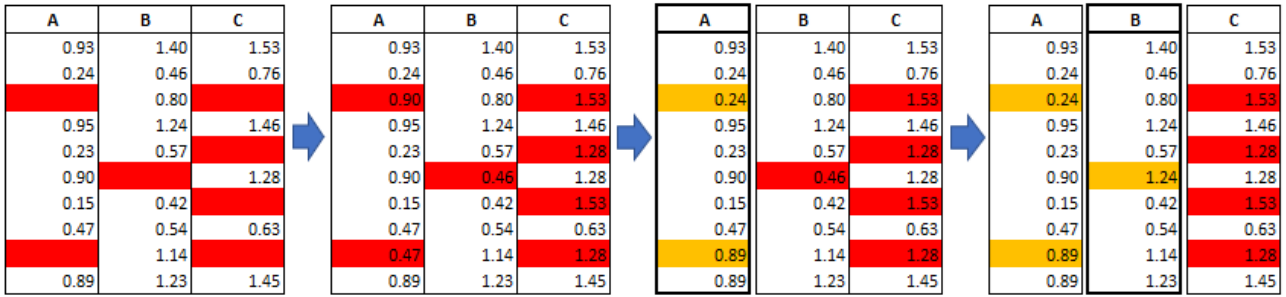
MICE - Multivariate Imputation by Chained Equations

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)

Imputación de datos

MICE - Multivariate Imputation by Chained Equations



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

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

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

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

El método inicia el dataset original, con los datos faltantes

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

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

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

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

Feature Engineering

Algunas técnicas son:

- Normalización
- Discretización
- Lograr normalidad
- Imaginación (Generación de nuevas variables)

Normalización

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

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

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.

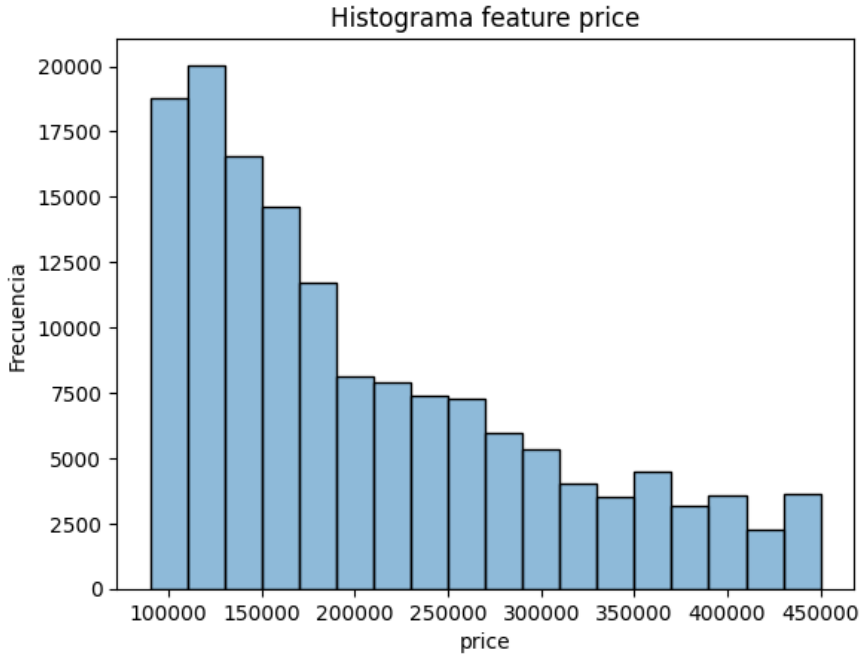
Normalización - Decimal scaling

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

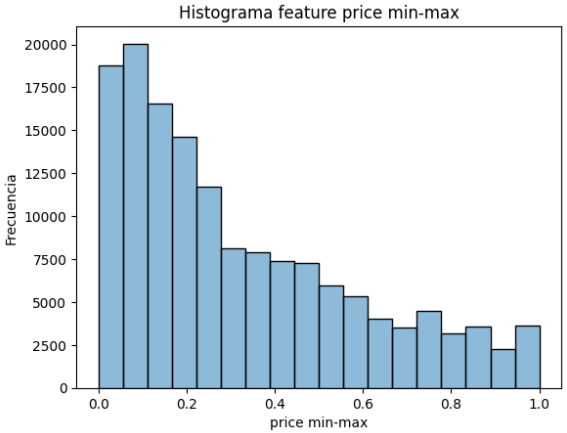
$$X_{decimal} = \frac{X}{10^d}$$

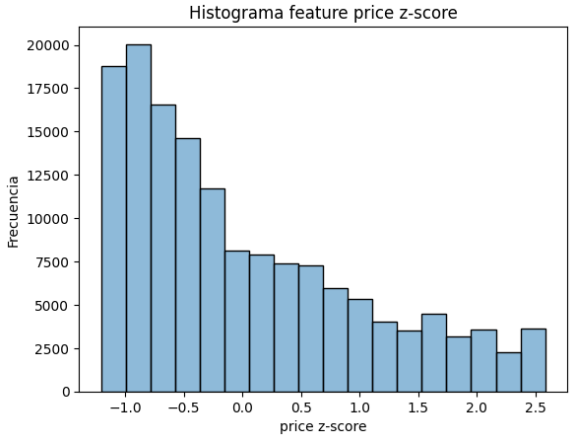
Donde **d** representa el número de dígitos en los valores de la variable con el valor absoluto más grande

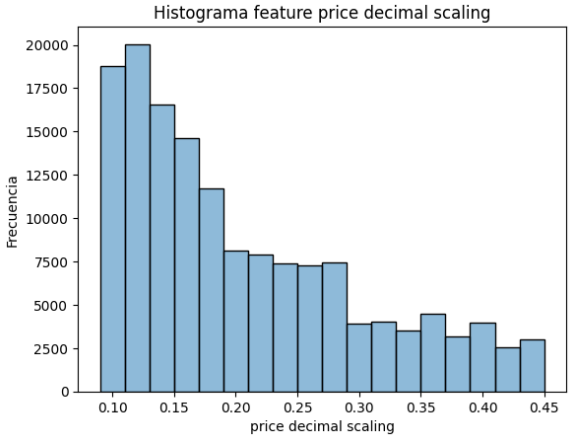
Normalización



Normalización







Transformaciones para lograr normalidad

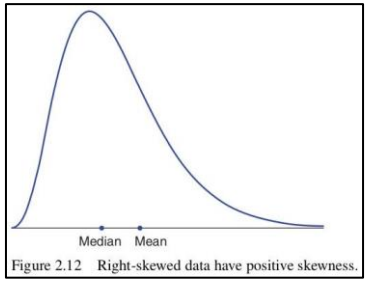


Figure 2.12 Right-skewed data have positive skewness.

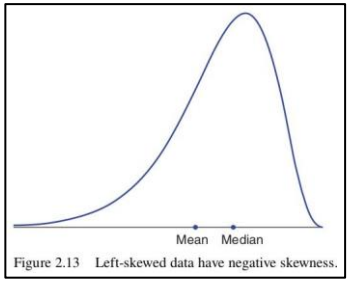


Figure 2.13 Left-skewed data have negative skewness.

Podemos reducir este sesgo a partir de transformaciones, por ejemplo:

- Raíz cuadrada
- Logaritmos
- Inversa de la raíz cuadrada
- Transformaciones de Box-Cox

Discretización

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

Se reducen los valores de una variable contínua a un número reducido de etiquetas

Discretización - Binning

Se divide a la variable en un número específico de **bins**

Los criterios de agrupamiento pueden ser por ejemplo:

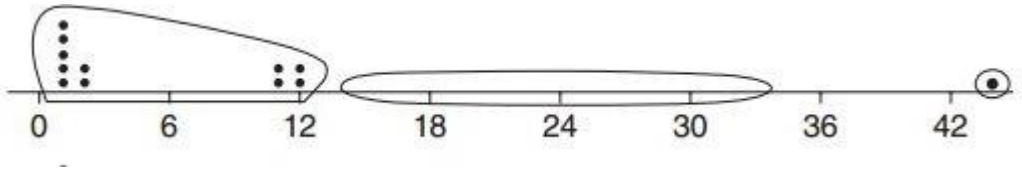
- Igual-Frecuencia: La misma cantidad de observaciones en un bin
- Igual-Ancho: Definimos rangos o intervalos de clases para cada bin
- Cuantiles: Separar en intervalos utilizando Mediana, Cuantiles, Percentiles.

Discretización - Binning

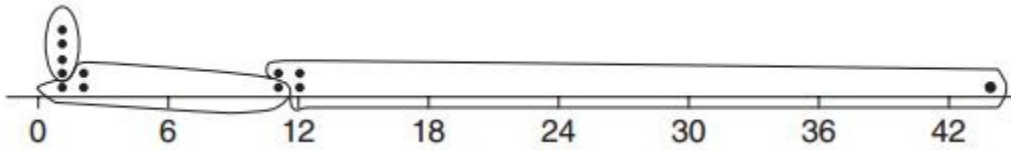
A su vez para cada uno de los agrupamientos podemos hacer:

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

Igual ancho

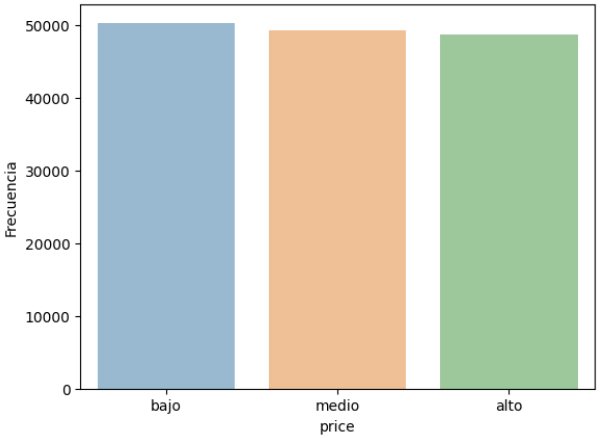


Igual frecuencia

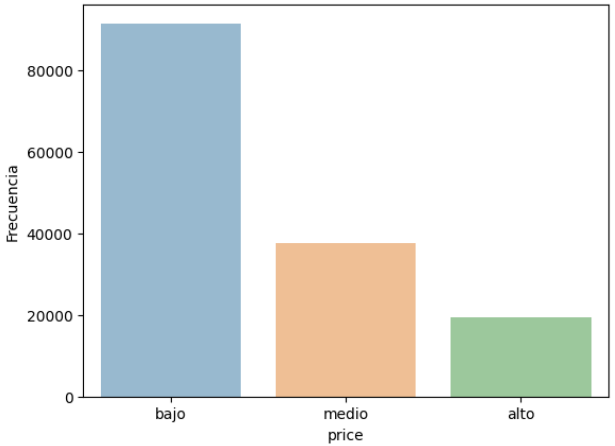


Discretización - Binning

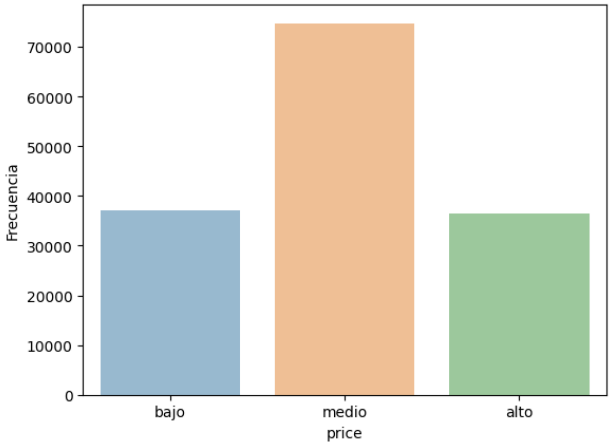
Ejemplos en el dataset de Properati



Igual frecuencia



Igual ancho de intervalo

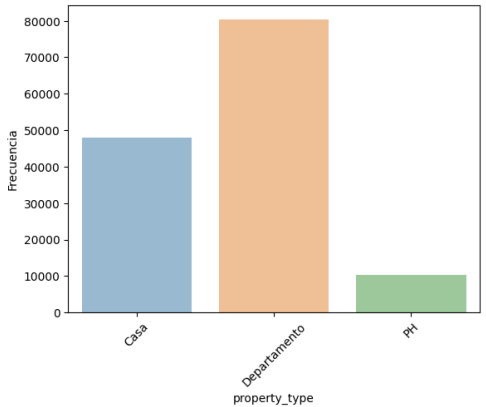


Cuartiles

Variables Dummies – One Hot Encoding

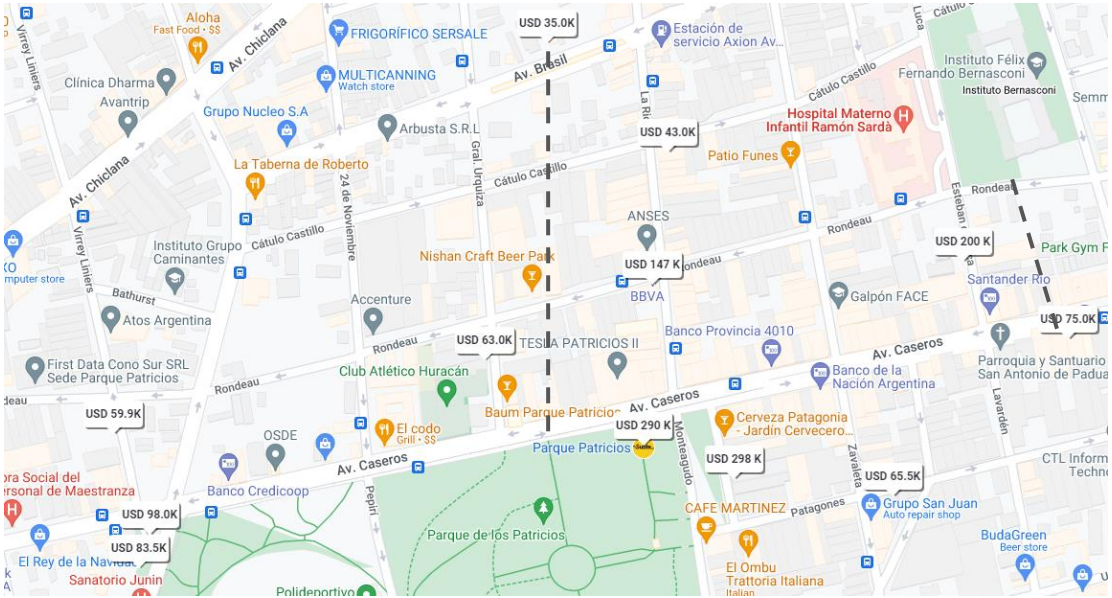
Algunos métodos analíticos requieren que las variables predictoras sean numéricas

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



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