IT ACADEMY - PROYECTO DATA SCIENCE

1. DESCRIPCIÓN DEL DATASET

El dataset que se utilizará para el trabajo se ha extraído del conjunto de datos **Spanish Houses**, correspondiente a una práctica de Web Scraping y que *puede encontrarse en el siguiente repositorio: https://github.com/trueuoc/spa_housing_crawler*)

En dicha práctica se implementó un rastreador web, con el cual se consiguieron extraer 100.000 anuncios de diferentes áreas en relación a viviendas en venta y alquiler del portal *idealistas.com*.

A partir de la información obtenida por provincias se ha trabajado para generar una nueva base de datos con información adicional a la original y que contiene los campos siguientes:

df_casas.shape: `(100000, 44)`

`df_casas.info():`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 40 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   provincia                100000 non-null  object
 1   ad_description           95426 non-null   object
 2   ad_last_update           100000 non-null  object
 3   air_conditioner          100000 non-null  object
 4   balcony                  100000 non-null  object
 5   bath_num                 100000 non-null  object
 6   built_in_wardrobe        100000 non-null  object
 7   chimney                  100000 non-null  object
 8   condition                86059 non-null   object
 9   construct_date           32059 non-null   object
 10  energetic_certif         74691 non-null   object
 11  floor                    79693 non-null   object
 12  garage                   40811 non-null   object
 13  garden                   100000 non-null  object
 14  heating                  25714 non-null   object
 15  house_id                 100000 non-null  object
 16  house_type               100000 non-null  object
 17  kitchen                  2212 non-null    object
 18  lift                     58965 non-null   object
 19  loc_city                 100000 non-null  object
 20  loc_district             86253 non-null   object
 21  loc_full                 100000 non-null  object
 22  loc_neigh                43690 non-null   object
 23  loc_street               14314 non-null   object
 24  loc_zone                 100000 non-null  object
 25  m2_real                  100000 non-null  object
 26  m2_useful                52844 non-null   object
 27  obtention_date           100000 non-null  object
```

```
 28  orientation                39415 non-null   object
 29  price                      100000 non-null  object
 30  reduced_mobility           100000 non-null  object
 31  room_num                   100000 non-null  object
 32  storage_room               100000 non-null  object
 33  swimming_pool              100000 non-null  object
 34  terrace                    100000 non-null  object
 35  unfurnished                646 non-null     object
 36  source                     100000 non-null  object
 37  number_of_companies_prov   100000 non-null  int64
 38  population_prov            100000 non-null  int64
 39  Valor_catastral_medio      100000 non-null  float64
dtypes: float64(1), int64(2), object(37)
memory usage: 30.5+ MB
```

2. OBJETIVOS DEL ESTUDIO

El presente proyecto se establecen los siguientes objetivos principales.

A) Integrar y limpiar el conjunto de datos que se pretende estudiar.

Las variables extraídas del Web Scraping se han descrito, evaluado y asignado una tipología y se han ordenada por conceptos similares:

Propiedades del Anuncio

- **house_id**: Identificador numérico único del anuncio de la vivienda (coincide con el path del anuncio) {num}
- **ad_last_update**: Fecha de la última actualización del anuncio {text}
- [**ad_description**: Descripción de texto del anuncio de la venta/alquiler de la vivienda {text}

Propiedades de la vivienda

- **price**: Precio de alquiler/venta de la vivienda {num}
- **bath_num**: Número de baños de la vivienda {num}
- [**condition**: Estado en el que se encuentra la vivienda {cat}
- **construct_date**: Año de construcción de la vivienda {date}
- **energetic_certif**: Certificado energético de la vivienda {cat}
- [**floor**: Número y/o ubicación de la(s) planta(s) de la vivienda {cat}
- **ground_size**: Metros cuadrados del terreno donde se ubica la vivienda {num}
- **heating**: Sistema de calefacción de la vivienda {cat}
- **house_type**: Tipo de vivienda {cat}
- **m2_real**: Metros cuadrados totales de la vivienda {num}
- **m2_useful**: Metros cuadrados útiles de la vivienda {num}
- **orientation**: Orientación de la vivienda {cat}
- **room_num**: Número de habitaciones de la vivienda {num}

Equipamiento adicional de la vivienda

- **air_conditioner**: Indica si la vivienda posee aire acondicionado {cat binary}
- **balcony**: Indica si la vivienda tiene balcones {cat binary}
- **built_in_wardrobe**: Indica si la vivienda consta de armarios empotrados {cat binary
- [19] **chimney**: Indica si la vivienda consta de chimenea {cat binary}
- **garage**: Indica el precio del garaje (en caso de que tenga) {num}
- **garden**: Indica si la vivienda tiene jardín {cat binary}
- **kitchen**: Indica si la vivienda está ya equipada con cocina {cat binary}
- **lift**: Indica si la vivienda consta de ascensor {cat binary}
- **reduced_mobility**: Indica si la vivienda está adaptada para personas con movilidad reducida {cat binary}
- **storage_room**: Indica si la vivienda consta de trastero {cat binary}
- **swimming_pool**: Indica si la vivienda tiene piscina {cat binary}
- **terrace**: Indica si la vivienda tiene terraza {cat binary}
- **unfurnished**: Indica si la vivienda está sin amueblar {cat binary}

Ubicación de la vivienda

- **loc_full**: Dirección completa de la vivienda {text}
- **loc_zone**: Provincia en la que se ubica la vivienda {text}
- **loc_district**: Área de la provincia en la que se ubica la vivienda {text}
- **loc_city**: Localidad en la que se ubica la vivienda {text}
- **loc_neigh**: Vecindario en la que se ubica la vivienda {text}
- **loc_street**: Calle en la que se ubica la vivienda {text}

Metadatos

- **obtention_date**: Fecha de obtención de los datos de la vivienda {date}

DATOS ADICIONALES

En este apartado se han añadido a nuestro conjunto de datos principal información demográfica y socio-económica, en concreto las variables siguientes:

- Number_of_companies_prov: Número de compañías por provincia {num}
- population_prov: Población por provincias {num}
- Valor_catastral_medio: Valor catastral medio por provincias {num}

B) Análisis Exploratorio

Se realizará un análisis exploratorio de los datos que nos permita extraer conocimiento, sobre todo en el estudio de relaciones entre las diferentes características de las viviendas y su precio.

Existen varias actividades al hacer un análisis exploratorio de datos, pero en cuanto a la minería de datos los puntos clave que se deben realizar son:

B.1) Descripción de la estructura de los datos.
B.2) Identificación de datos faltantes.
B.3) Detección de valores atípicos.
B.4) Correlación entre pares de variables.

B.1) Descripción de la Estructura de datos:

Distribución de inmuebles en venta por provincias

Provincia	
balears	24.822
bizkaia	21.515
coruna	8.311
gipuzkoa	7.030
madrid	5.248
sevilla	5.072
albacete	4.454
alava	3.801
zamora	3.221
alicante	1.944
ciudad_real	1.816
girona	1.808
segovia	1.526
valencia	1.401
soria	1.261
santa_cruz_de_tenerife	1.218
barcelona	885
cadiz	787
huelva	657
tarragona	219
valladolid	210
Total	97.206

Como se puede observar la base de datos cuenta con un gran número de datos para dos comunidades en concreto como son Balears y el País Vasco (Biskaia, Guipuzkoa y Alava-Arava), pero para las siguientes comunidades/provincias se han extraído un número menor de anuncios.

Se ha optado por trabajar con la base de datos completa y no intentar homogeneizar las muestras por provincia, ya que los modelos que se van a utilizar son de aprendizaje supervisado de regresión y nos interesa tener el mayor número de inputs sobre las variables que inciden en el precio de venta de los inmuebles que es nuestra variable objetivo.

Propiedades por tipologías

house_type	
Piso	53964
Casa o chalet independiente	13592
Casa o chalet	6868
Chalet adosado	6441
Ático	3345
Casa de pueblo	3001
Dúplex	2577
Chalet pareado	2482

```
Finca rústica                                                    2276
Alquiler de Piso                                                 2244
Casa rural                                                       1564
Estudio                                                           692
Caserón                                                           161
Alquiler de Ático                                                 101
Masía                                                             100
Alquiler de Dúplex                                                 72
Casa terrera                                                       69
Alquiler de Casa o chalet independiente                            61
Alquiler de Chalet adosado                                         54
Cortijo                                                            30
Alquiler de Estudio                                                23
Palacio                                                            18
Torre                                                              18
Alquiler de Casa o chalet                                          16
Alquiler de Chalet pareado                                         15
Alquiler de Casa de pueblo                                         11
Alquiler de Casa rural                                              9
Castillo                                                            8
Alquiler de Finca rústica                                           4
Alquiler de Caserón                                                 1
```

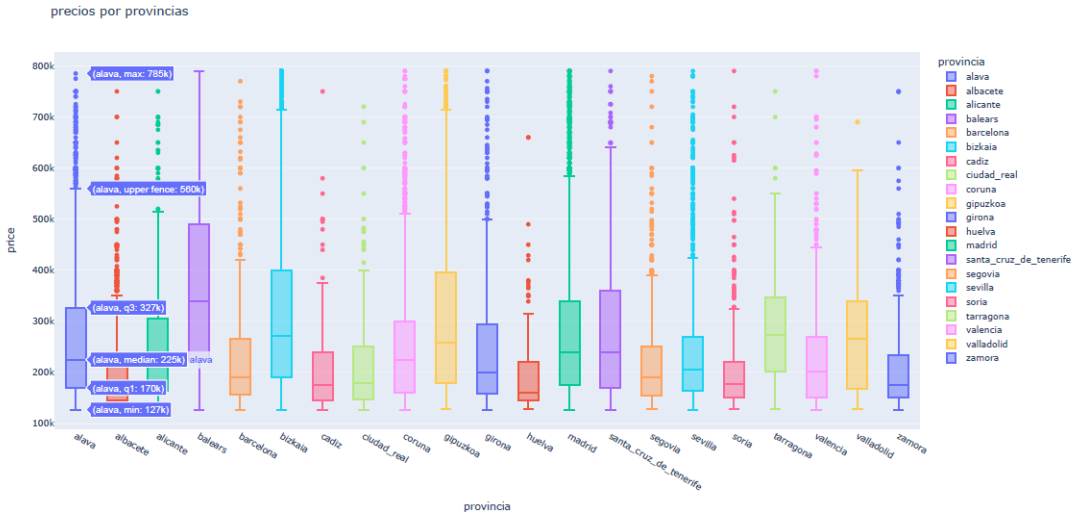
En cuanto a las tipologías, nos centraremos en las casas en venta exclusivamente, ya que deseamos estimar el precio final de los inmuebles sin tener que estimarlo a partir de las rentas mensuales de alquiler.

B.2) Identificación de los datos faltantes

Analizaremos las variables y estimaremos en la medida que sea posible resto de valores faltantes, los imputaremos utilizando, valores de la media, mediana, o mediante modelos, como puede ser K-Nearest-Neighbor.

B.3) Detección de Valores Atípicos

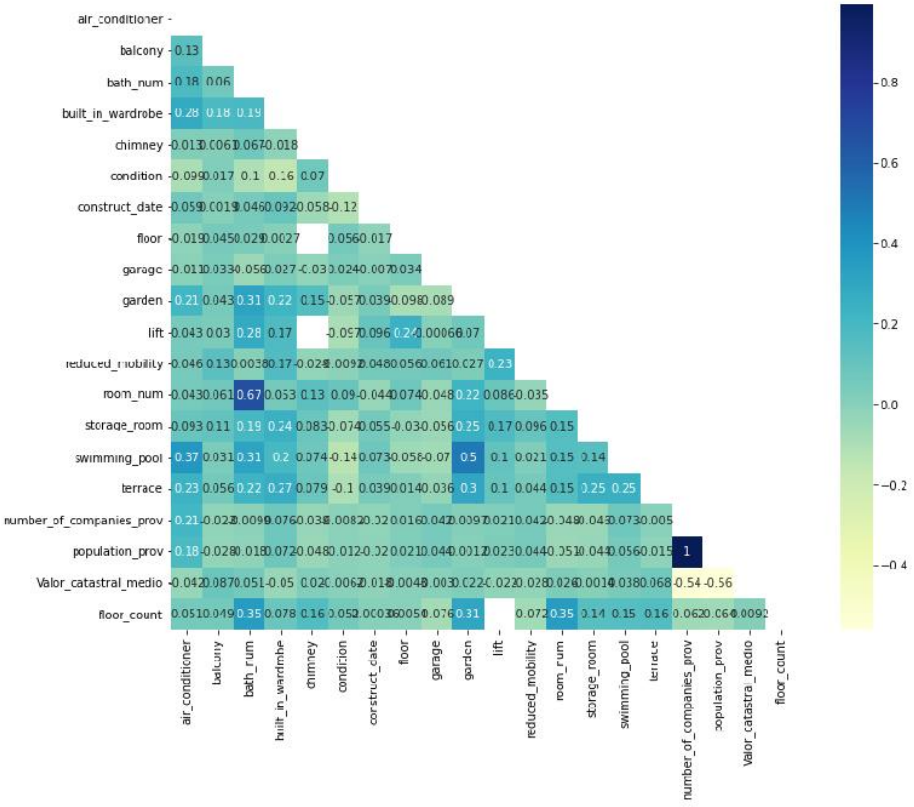
Precios por provincias



En el gráfico boxplot de precios de los inmuebles por provincias se observan como se distribuyen los precios por percentiles, en relación a la mediana. Los puntos que exceden del percentil 100%, corresponden a los outliers de la variable precio para cada provincia.

Se analizará el tratamiento más efectivo para reducir su influencia en el análisis final, con el objeto de maximizar el rendimiento de modelo final seleccionado.

B.4) Correlación entre variables



Se analizará la existencia de correlaciones entre pares de variables y se estudiará si es necesario reducir el número de las que sean redundantes y no aporten mayor precisión al modelo.

C) Implementar dos modelos de regresión

- Un modelo de regresión lineal que nos permitirá conocer si una vivienda se ajusta a nuestro presupuesto
- Modelo de aprendizaje supervisado que permita predecir el precio de las viviendas a partir de sus características (variables explicativas contenidas en el Dataset).

D) Crear un API que facilite el acceso al modelo de presupuesto.

Las API presentan una oportunidad única para que las empresas satisfagan las necesidades de sus clientes en diferentes plataformas. Por ejemplo, la API de mapas permite la integración de información de los mapas en sitios web, Android, iOS, etc. En nuestro caso queremos dar un acceso similar a la base de datos usadas en este trabajo mediante el uso de API.