Victor Baena Chamizo
Arnau Muns Orenga

Tipologia i cicle de vida de les dades

Pràctica 2 - Neteja i anàlisi de les dades

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El conjunt de dades que hem decidit estudiar és el que vam crear a la pràctica 1. Aquest dataset porta per títol "característiques_pisos_igualada.csv". Tal i com dóna a entendre el títol, aquest conjunt de dades està format per les diferents característiques dels pisos de la ciutat d'Igualada. Aquestes característiques inclouen, entre d'altres, el preu o els m2 del pis o casa concret. Aquestes característiques es detallen en les següents preguntes.

En aquest conjunt de dades cada observació fa referència a un pis o casa diferent, presentant-se per columnes totes les característiques prèviament esmentades. Aquestes presenten unes unitats de mesura que varien en funció de cada variable.

Considerem que aquest conjunt de dades és molt interessant per gent jove amb un perfil com el nostre, que estiguin buscant pis en el moment actual. El motiu d'això és que aquest data set aporta informació de totes les cases i pisos en venta en un moment concret del temps en una ciutat determinada, en aquest cas Igualada.

Alguns exemples de preguntes a les quals es pretén donar resposta amb aquest conjunt de dades són les següents:

<u>Preguntes relacionades amb els pisos en si</u>

- Quins són els 3 pisos més barats/assequibles de tota Igualada?
- Quins són els pisos més barats en €/m2 de tota Igualada?
- Quin és el pis més car i quin el més barat?
- Quin és el pis amb més habitacions? amb més lavabos? etc
- Quin és el pis amb una major ràtio d'equipament/€?

<u>Preguntes relacionades amb la zona</u>

- Quina zona té més anuncis de pisos en venda?
- Quina és la mitjana de preu dels pisos segons les diferents zones de la ciutat?
- Quina és la zona més barata i quina la més cara (en mitjana)?
- Quina és la zona amb més dispersió de preus?

<u>Preguntes generals</u>

- Quin seria el millor pis per comprar en base a les seves característiques si es disposa d'un pressupost limitat de X €?
- Quina seria la recomanació de compra per diferents pressupostos?
- Quina seria la recomanació de compra per un pressupost infinit segons la zona?

En general la idea que tenim de projecte per aquest conjunt de dades és la de poder construir un model predictiu del preu de venda d'un pis en base a les seves característiques per tal de poder predir, per un pis nou que es posa a la venda, quin hauria de ser el seu preu en base a altres pisos semblants que ja estiguin ofertats per internet. Aquest model, evidentment, vindrà precedit per un anàlisi descriptiu que pot incloure preguntes com les que s'han exemplificat anteriorment.

2. Integració i selecció de les dades d'interès a analitzar. Pot ser el resultat d'addicionar diferents datasets o una subselecció útil de les dades originals, en base a l'objectiu que es vulgui aconseguir.

En aquest cas no hem considerat cap procés d'integració ni de selecció de les dades.
Quant al procés d'integració, aquest no s'ha considerat perquè pensem que el conjunt de dades que es va construir a la pràctica 1 és suficientment complet pel que respecta a la cobertura dels pisos ofertats per internet a Igualada. Tanmateix si haguéssim volgut fer algún procés d'integració de dades pensem que el millor hagués estat extreure informació dels pisos en venda a Igualada d'altres pàgines web com podria ser Habitaclia amb l'objectiu d'unificar tots dos conjunts de dades en un mateix per cobrir els pisos que estan en una pàgina web però no en l'altre.
Quant al procés de selecció de dades no considerem que sigui necessàri ja que totes les observacions (pisos) del nostre conjunt de dades són rellevants per l'anàlisi final. Tanmateix si s'hagués volgut considerar algún procés de selecció de dades considerem que el més adient hagués estat filtrar les dades d'algún barri concret d'Igualada en el cas que haguessim volgut estudiar i construir un model només pels pisos d'aquell barri en concret.

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos.

En primer lloc al estudiar els elements buits o dades faltants amb la funció isnull de python es pot observar el següent:

```
pisos.isnull().sum()

id                      0
construidos_m           0
utiles_m                0
habitaciones            0
baños                   0
terraza                 0
balcon                  0
estado                  0
orientacion           161
garaje                  0
año_construido          0
calefaccion             0
planta                  0
ascensor                0
aire_acondicionado      0
zonas_verdes_o_jardin   0
ciudad                  0
zona                    0
etiqueta_energetica     0
precio                  0
precio_anterior         0
titulo_anuncio          0
dtype: int64
```

Sembla que la única variable que presenta valors missing o NaN és la variable *orientacion*, en total 161 registres. Això significa que durant el procés d'scraping la informació sobre l'orientació del pis no es va trobar per 161 pisos. Pensem que eliminar tots els registres que no presenten un valor per aquesta variable seria massa radical ja que pràcticament estariem eliminant la meitat de les observacions del nostre conjunt de dades. Per tant, al considerar-se una variable categòrica politòmica, hem decidit transformar els valors NaN de la variable *orientacion* en un altre nivell d'aquesta variable (*no_definido).* D'aquesta manera podem mantenir els registres assumint que l'orientació d'aquests 161 pisos no està definida en l'anunci. La variable final queda de la següent manera:

```
no_definido    161
sur             59
norte           24
este            18
este,           15
norte,          15
oeste            8
sur,             7
Name: orientacion, dtype: int64
```

En aquest punt es pot anticipar un altre problema en aquesta variable relacionat amb els valors duplicats a causa de les comes que es tractarà posteriorment al final d'aquest apartat de neteja.

Després de netejar la única variable que a priori sembla que presenti valors missing, sabem que cal netejar altres valors faltants en les variables numèriques que vam codificar com a -1. Aquests valors de -1 representen un valor faltant en les variables on apareguin quan el valor pel pis concret no apareixia en l'anunci.
Comencem per mostrar un resum amb els principals estadístics descriptius per cada una de les variables numèriques excepte l'id del pis:

	construidos_m	utiles_m	habitaciones	baños	año_construido	planta	precio	precio_anterior
count	307.000000	307.000000	307.000000	307.000000	307.000000	307.000000	307.000000	307.000000
mean	134.498371	111.172638	3.133550	1.807818	705.351792	1.247557	187059.377850	187779.263844
std	99.092265	84.176516	1.755965	0.935259	1870.794547	1.814237	121220.583121	121343.428545
min	-1.000000	-1.000000	-1.000000	-1.000000	-2022.000000	-1.000000	43000.000000	43000.000000
25%	80.000000	70.500000	3.000000	1.000000	-2022.000000	-1.000000	115100.000000	115100.000000
50%	101.000000	87.000000	3.000000	2.000000	1970.000000	1.000000	149000.000000	149000.000000
75%	128.000000	110.500000	4.000000	2.000000	2002.500000	3.000000	235000.000000	235000.000000
max	794.000000	450.000000	14.000000	7.000000	2023.000000	8.000000	850000.000000	850000.000000

Es pot observar que totes les variables numèriques excepte el preu i el preu anterior presenten algun valor codificat que representa dades faltants. El cas de la variable *planta* és un cas a part perquè en realitat pensem que aquesta variable s'hauria de tractar com una variable categòrica i no com a variable numèrica. Per tant els valors faltants de la variable *planta* es tractaran posteriorment durant la pràctica de la mateixa manera que la variable *orientacion*.

Pel que fa a les altres variables numèriques es procedeix a transformar els valors de -1 en valors NaN per tal que es capturin amb la funció isnull de python. Això ens dona una idea més clara de quants valors faltants tenim en les variables numèriques:

```
id                      0
construidos_m           4
utiles_m                4
habitaciones           22
baños                   1
terraza                 0
balcon                  0
estado                  0
orientacion             0
garaje                  0
año_construido         98
calefaccion             0
planta                 93
ascensor                0
aire_acondicionado      0
zonas_verdes_o_jardin   0
ciudad                  0
zona                    0
etiqueta_energetica     0
precio                  0
precio_anterior         0
titulo_anuncio          0
dtype: int64
```

Tractament dels valors missing

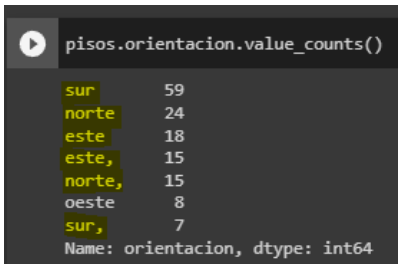
Per l'elaboració d'aquesta pràctica hem decidit tractar els valors faltants de les variables numèriques imputant aquests valors per la mitjana de les variables. Hem decidit optar per aquesta alternativa i no per eliminar els registres que contenen dades faltants per dos motius. En primer lloc cal tenir en compte que no tots els valors faltants per totes les variables són a la mateixa fila i, per tant, si els haguéssim decidit eliminat estaríem perdent més observacions que el nombre màxim de valors missing. En segon lloc hem decidit imputar els valors missing i no eliminar els registres perquè estem tractant amb un conjunt de dades relativament petit (307 observacions) i volíem poder aprofitar totes les observacions disponibles.

Es pot observar que després d'aplicar la funció *fillna* de python amb la mitjana no es troben més valors missing en cap variable:

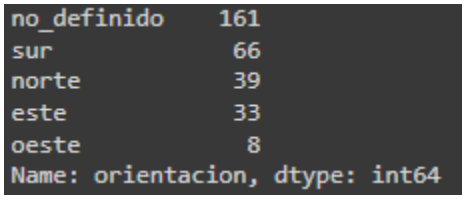
```
#imputar per la mitjana tots els valors null de les variables numèriques
pisos.fillna(pisos.mean(), inplace = True)
pisos.isnull().sum()

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: FutureWar

id                      0
construidos_m           0
utiles_m                0
habitaciones            0
baños                   0
terraza                 0
balcon                  0
estado                  0
orientacion             0
garaje                  0
año_construido          0
calefaccion             0
planta                  0
ascensor                0
aire_acondicionado      0
zonas_verdes_o_jardin   0
ciudad                  0
zona                    0
etiqueta_energetica     0
precio                  0
precio_anterior         0
titulo_anuncio          0
dtype: int64
```

Altres neteges requerides

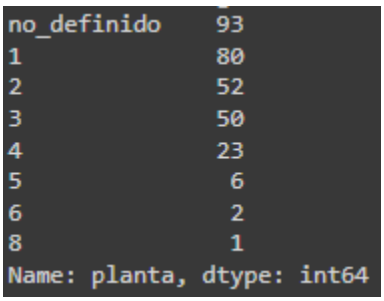
- Cal netejar la variable orientación perquè alguns valors estan duplicats a causa d'una coma extra. La següent imatge mostra el nombre de files per cada valor de la variable orientacion abans de la neteja



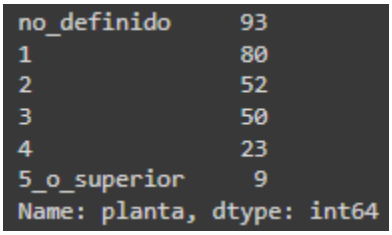
La següent imatge mostra el nombre de files per cada valor de la variable orientacion després de la neteja. Es pot observar que s'han eliminat els valors duplicats a causa de la coma.



- Es considera la transformació de la variable planta en una variable categòrica ja que en la majoria dels casos els pisos tenen entre 1 i 4 plantes. També aquesta variable presenta valors NA que es transformen en una categoria extra "no_definido".



S'ha decidit agregar les categories que representen menys d'un 5% dels casos en una categoria anomenada 5_o_superior. La variable netejada finalment queda de la següent manera



3.2. Identifica i gestiona els valors extrems.

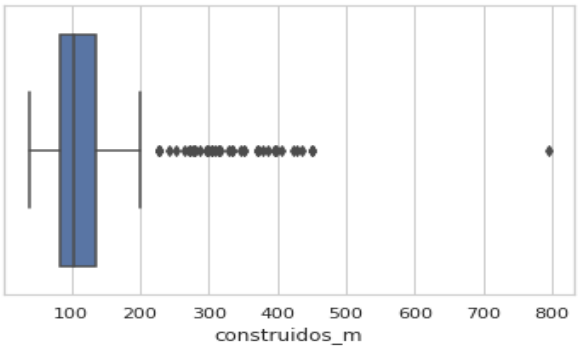
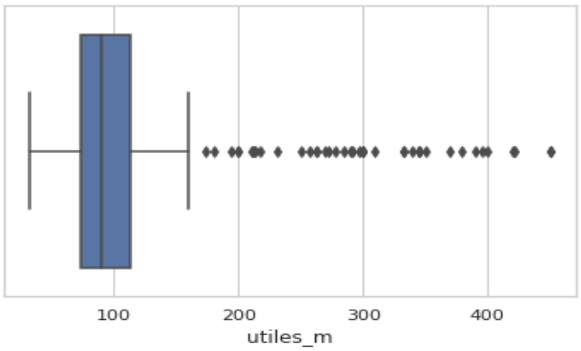
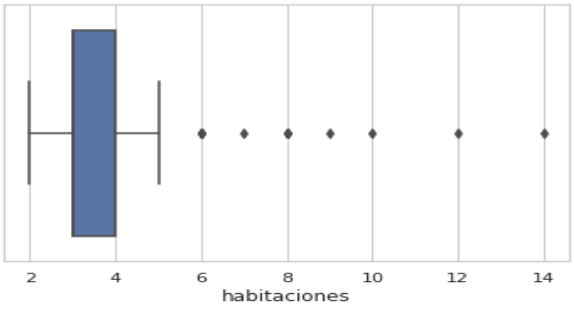
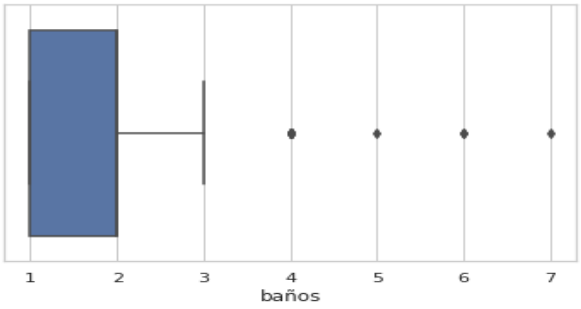
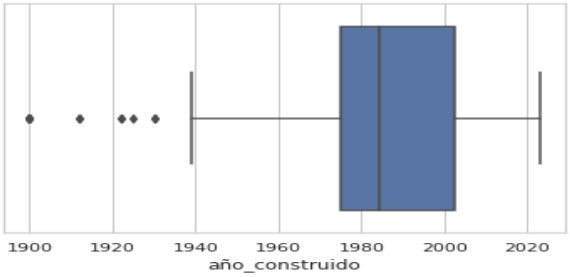
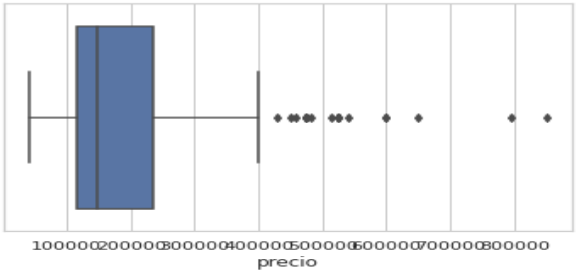
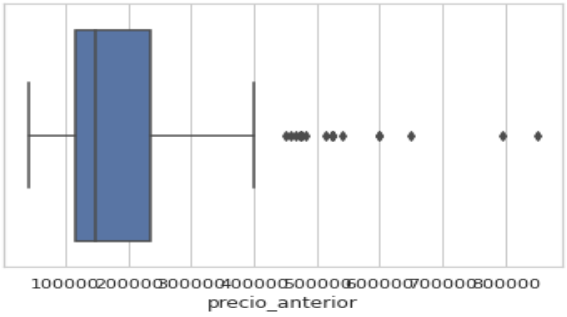
En primer lloc ens hem adonat que la variable *utiles_m* que fa referència als metres quadrats construits presenta un valor molt extrem que es captura com a mínim:

	construidos_m	utiles_m	habitaciones	baños	año_construido	precio	precio_anterior
count	307.000000	307.000000	307.000000	307.000000	307.000000	307.000000	307.000000
mean	136.287129	112.653465	3.452632	1.816993	1984.205742	187059.377850	187779.263844
std	97.857607	83.180735	1.326712	0.921337	24.275660	121220.583121	121343.428545
min	40.000000	1.000000	2.000000	1.000000	1900.000000	43000.000000	43000.000000
25%	83.500000	71.500000	3.000000	1.000000	1975.000000	115100.000000	115100.000000
50%	104.000000	88.000000	3.000000	2.000000	1984.205742	149000.000000	149000.000000
75%	136.287129	112.653465	4.000000	2.000000	2002.500000	235000.000000	235000.000000
max	794.000000	450.000000	14.000000	7.000000	2023.000000	850000.000000	850000.000000

Per aquest motiu es decideix tractar aquest valor d'1 en els registres que el contenen com un valor missing ja que 1 metre quadrat útil no és una dada que sembli realista. Com en la resta de dades missing es decideix imputar aquest valor per la mitjana de la variable. Una vegada aplicada la imputació els estadístics bàsics de la variable són els següents:

```
pisos.utiles_m.describe()

count    307.000000
mean     115.563002
std       81.145672
min       33.000000
25%       74.000000
50%       90.000000
75%      112.653465
max      450.000000
```

Posteriorment per analtizar els outliers de totes les variables numèriques es procedeix a realitzar el boxplot corresponent a cadascuna d'elles. Aquest gràfic ens permet analitzar d'una manera visual i ràpida la presència d'outliers en les nostres variables numèriques.

construidos_m	utiles_m
	
habitaciones	**baños**
	
año_construido	**precio**
	
precio_anterior	
	

Com es pot observar, en general, totes les variable numèriques presenten outliers. La majoria d'elles presenten outliers en el rang elevat de valors, això és, es detecten en els boxplot perquè aquestes observacions es troben per sobre del quartil 3 1.5 vegades el rang interquartílic. Tanmateix la única variable que presenta, a priori, outliers per sota del quartil 1 és la variable año_construido.

Parlant de la variable año_construido, en general considerem que els outliers s'han de mantenir en el dataset perquè pertanyen a edificis històrics construits a Igualada abans dels anys 40 del segle XX. S'ha de considerar que Igualada, amb llarga tradició industrial en el sector del cuir, hi han edificis en venda que van ser construits fa aproximadament entre 80 i 100 anys.

En general pensem que tota la resta d'outlier en les variables numèriques també s'haurien de mantenir perquè probablement fan referència a finques senceres que estan en venda i, evidentment, com la majoria dels pisos en el nostre conjunt de dades són només això, pisos, les finques senceres apareixen com a grans outliers. Això es pot observar en les variables construidos_m i utiles_m on tots els outliers són de finques amb molts metres quadrats. També es pot apreciar en les variables *habitaciones* i *baños* ja que tots els outliers són de finques amb més de 6 habitacions i més de 4 banys. Evidentment un pis normal normalment no té ni tantes habitacions ni tants lavabos i això és perquè aquestes observacions pertanyen a finques i no a pisos com ja s'ha comentat. A més a més també es pot observar el mateix fet en les variables *precio* i *precio_anterior*. Els outliers molt probablement pertanyen a les mateixes finques/cases que podem veure en els metres quadrats ja que és evident que una casa més gran costarà més diners. En la pàgina d'idealista no es va decidir filtrar només per pisos sino que també vam decidir incorporar les finques (tots els pisos o finques disponibles, de fet) i és per això que ara els podem observar en aquest anàlisi.

En definitiva hem decidit no tractar aquests outliers perquè pensem que són valors completament legítims a tenir en el nostre conjunt de dades ja que probablement pertanyen a finques o cases en comptes de pisos exclusivament. A més, com l'objectiu final és el d'elaborar un model predictiu del preu en base a les característiques dels pisos/cases, pensem que no és necessàri eliminar o tractar aquestes dades ja que aquestes observacions també aporten informació rellevant alhora d'estimar el preu de venda d'un pis o una casa a Igualada.

Arribats a aquest punt hem decit que aquest és el nostre conjunt de dades final. Cal remarcar que no s'ha escollit aplicar el procés d'enginyeria de característiques o feature engineering abans de guardar la taula final sino que aquest procés s'aplicarà posteriorment. Per tant s'ha procedit a guardar-lo amb el nom "caracteristiques_pisos_igualada_clean.csv".

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?).

Els grups de dades que es volen analitzar són els següents tenint en compte que es vol analitzar quins elements afecten al preu de venda d'un pis:

Precio - construidos_m
Precio - utiles_m
Precio - habitaciones
Precio - baños
Precio - año_construido
Precio - planta

Precio - estado
Precio - orientación
Precio - ascensor
Precio - zona
Construidos_m - zona

Descriptius bàsics del conjunt de dades

	id	construidos_m	utiles_m	habitaciones	baños	año_construido	precio	precio_anterior
count	307.000000	307.000000	307.000000	307.000000	307.000000	307.000000	307.000000	307.000000
mean	153.000000	136.287129	115.563002	3.452632	1.816993	1984.205742	187059.377850	187779.263844
std	88.767487	97.857607	81.145672	1.326712	0.921337	24.275660	121220.583121	121343.428545
min	0.000000	40.000000	33.000000	2.000000	1.000000	1900.000000	43000.000000	43000.000000
25%	76.500000	83.500000	74.000000	3.000000	1.000000	1975.000000	115100.000000	115100.000000
50%	153.000000	104.000000	90.000000	3.000000	2.000000	1984.205742	149000.000000	149000.000000
75%	229.500000	136.287129	112.653465	4.000000	2.000000	2002.500000	235000.000000	235000.000000
max	306.000000	794.000000	450.000000	14.000000	7.000000	2023.000000	850000.000000	850000.000000

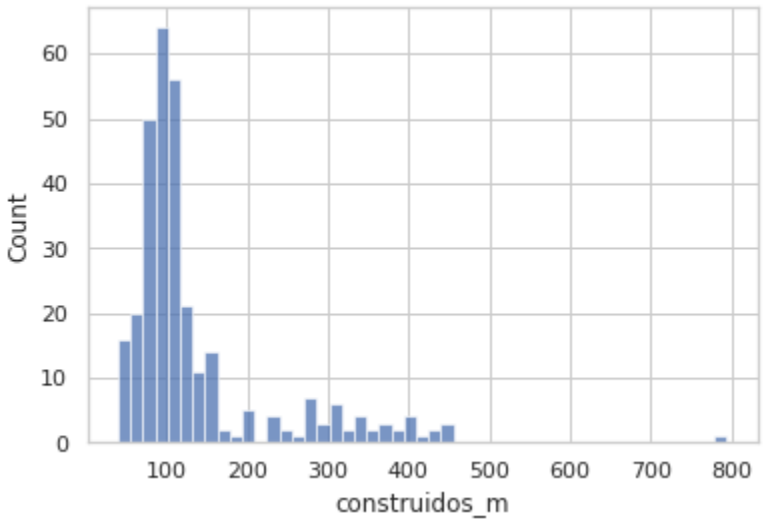
	terraza	balcon	estado	orientacion	garaje	calefaccion	planta	ascensor	aire_acondicionado	zonas_verdes_o_jardin	ciudad	zona	etiqueta_energetica	titulo_anuncio
count	307	307	307	307	307	307	307	307	307	307	307	307	307	307
unique	2	2	2	5	2	2	5	2	2	3	1	7	8	138
top	NO	NO	Buen estado	no_definido	NO	NO	2	NO	NO	NO	Igualada	Igualada	e	Piso en venta en Pla de Sant Magi
freq	160	206	280	161	231	163	145	158	243	281	307	188	103	39

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

En aquest apartat es procedeix a comprovar la normalitat de totes les dades numèriques del nostre conjunt de dades utilitzant el test de Shapiro-Wilk. Aquest test estadístic es considera un dels mètodes més potents per contrastar la normalitat.

```
Normalitat de la variable: construidos_m
(0.7027856707572937, 5.848039829480856e-23)

Normalitat de la variable: utiles_m
(0.6830496788024902, 1.2965091482653805e-23)

Normalitat de la variable: habitaciones
(0.6640591621398926, 3.2520903230415206e-24)

Normalitat de la variable: baños
(0.757533609867096, 5.943277353115732e-21)

Normalitat de la variable: año_construido
(0.868304967880249, 1.5932724677962464e-15)

Normalitat de la variable: precio
(0.7999213337898254, 3.7851159275088303e-19)

Normalitat de la variable: precio_anterior
(0.8010028600692749, 4.2432499097713247e-19)
```

Els valors resaltats en groc són el p-valor del test de Shapiro-Wilk. Com es pot observar els p-valors dels contrastos per totes les variable numèriques són molt menors a 0 es conclou que es rebutgen totes les hipòtesis nul·les. Això significa que es tenen evidències estadísticament significatives de que aquestes variables no provenen d'una distribució normal. Això no ens hauria d'extranyar ja que per exemple si analitzem gràficament la distribució de la variable construidos_m es pot veure clarament que, a causa d'incloure les finques i cases en el conjunt de dades, les variables numèriques no són normals:



Seguidament es comprova l'homogeneïtat de la variança utilitzant el test de Fligner-Killeen. Aquest test es tracta de l'alternativa no paramètrica que s'utilitza quan les dades no compleixen amb la condició de normalitat, com és el cas. Les variables per les quals es comprova l'homoscedasticitat són:

Precio - construidos_m, Precio - habitaciones, Precio - baños, Precio - año_construido

```
Precio - construidos_m
FlignerResult(statistic=370.7154963369166, pvalue=1.3074008933542813e-82)

Precio - utiles_m
FlignerResult(statistic=370.9741310812382, pvalue=1.1484082910393395e-82)

Precio - habitaciones
FlignerResult(statistic=376.12401021392384, pvalue=8.686331053079116e-84)

Precio - baños
FlignerResult(statistic=377.8496307244461, pvalue=3.657075083277913e-84)

Precio - año_construido
FlignerResult(statistic=372.8486800256504, pvalue=4.486983828809817e-83)

Homogeneitat de variances entre totes les variables
FlignerResult(statistic=1119.7155101168573, pvalue=4.033843608605797e-241)
```

Com es pot apreciar a partir dels valors resaltats en groc en la figura anterior cap de les variables presenta homogeneïtat en la variança en relació amb la variable preu. Al ser el p-valor molt inferior a 0.05 per tots els contrastos realitzats es conclou que hi ha evidències estadísticament significatives per rebutjar la hipòtesi nul·la de que les variances de les diferents variables són iguals respecte a la variable preu.

Després d'estudiar la normalitat i la homoscedasticitat de les variables del nostre conjunt de dades arribem a la conclusió de que caldrà realitzar contrastos d'hipòtesi no paramètrics ja que no es compleixen ni la normalitat ni la homoscedasticitat en les dades.

4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

CORRELACIONS ENTRE VARIABLES NUMÈRIQUES

Amb l'objectiu d'estudiar la correlació entre les diferents variables numèriques del nostre conjunt de dades es procedeix a calcular la matriu de correlacions. Aquesta eina ens permet analitzar la correlació, positiva o negativa, entre la variable objectiu del posterior model de predicció (el preu) i la resta de variables numèriques del nostre conjunt de dades. També ens permet veure si algunes de les variables regressores estan correlacionades entre elles, de manera que ens permet anticipar problemes de colinearitat o multicolinearitat de cara a la creació del model predictiu.

En primer lloc es calcula la matriu de correlacions:

	construidos_m	utiles_m	habitaciones	baños	año_construido	precio	precio_anterior
construidos_m	1.000000	0.924165	0.546782	0.649674	-0.172114	0.745067	0.744925
utiles_m	0.924165	1.000000	0.528213	0.601398	-0.178971	0.650930	0.650623
habitaciones	0.546782	0.528213	1.000000	0.636832	-0.156845	0.402457	0.401227
baños	0.649674	0.601398	0.636832	1.000000	0.028590	0.661630	0.661120
año_construido	-0.172114	-0.178971	-0.156845	0.028590	1.000000	0.168977	0.169556
precio	0.745067	0.650930	0.402457	0.661630	0.168977	1.000000	0.999565
precio_anterior	0.744925	0.650623	0.401227	0.661120	0.169556	0.999565	1.000000

Es poden observar moltes coses en l'anterior taula. En primer lloc destaquem que hi ha una correlació positiva molt elevada entre la variable construidos_m i la variable utiles_m amb un valor de 0.92. Això significa que haurem d'eliminar la variable utiles_m del model final de predicció del preu per evitar problemes de colinearitat.

També s'observa una correlació lleugerament positiva entre la variable construidos_m i la variable habitaciones. Això significa que a més metres quadrats construits sembla que es tenen més habitacions (0.54). El mateix s'observa en la relació entre construidos_m i el número de banys. A més metres quadrats construits més banys.

D'altra banda també es pot observar que sembla existir una lleugera correlació positiva entre habitaciones i baños, cosa que significa que quantes més habitacions es tenen sembla ser que més lavabos es tenen en una casa (0.63).

Quant a les correlacions llegurament negatives es detecten entre les variables construidos_m i año_construido i entre habitacions i año_construido. Tot i ser molt lleugera, aquesta correlació es pot interpretar com a que hi ha una relació inversa entre els metres quadrats i els anys construits. És a dir que les cases més velles solen ser les que tenen més metres quadrats i les noves en tenen menys. El

mateix pel nombre d'habitacions: Com més vella és la casa sembla que hi han més habitacions, o el que és el mateix, com més nova és la casa sembla que es tenen menys habitacions.

Quant al preu, la nostra variable resposta, es pot observar que en general presenta correlacions positives significatives amb la majoria de variables excepte amb el año construido, que al ser propera a 0 es pràcticament una correlació no existent. És a dir que no sembla que l'any de construcció estigui positivament correlacionat amb el preu d'una manera forta. Pel que fa a la resta de variables sí que s'observa correlació positiva significativa. Amb els metres quadrats construits, cosa que significa que a més metres quadrats construits més preu. També amb el nombre d'habitacions i lavabos, cosa que significa que a més preu més habitacions i lavabos es tenen. Evidentment la correlació és pràcticament 1 amb la variable precio anterior per la qual cosa caldrà eliminar la variable precio_anterior del conjunt de dades i no incloure-la en el model.

TEST DE DIFERÈNCIA DE MEDIANES

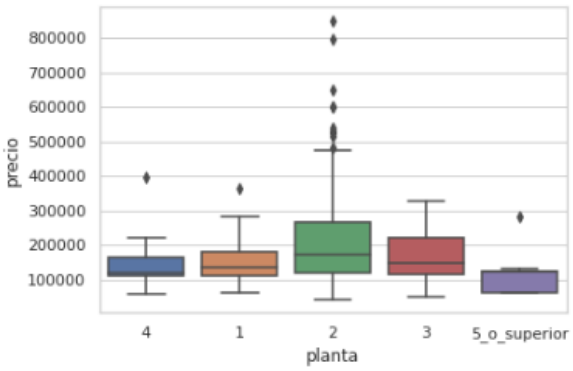
Una vegada analitzada la correlació entre les nostres variables numèriques es volen estudiar les relacions entre la variable preu i les variables categòriques. Per fer-ho es proposa l'estudi dels tests de diferència de mitjanes que ens permeten veure si hi han diferències estadísticament significatives entre les mitjanes dels preus pels diferents grups de les variables categòriques.
Cal remarcar que en les opcions en les que es vol comparar diferents grups (més de dos nivells) s'escull l'alternativa no paramètrica de l'ANOVA (test de Kruskal Wallis) perquè les mostres són independents.

Precio - planta

Per analitzar si hi ha diferències en mediana en el preu per totes les plantes s'utilitza el test de kruskal wallis per comparar més de dos grups atès que no es pot aplicar l'alternativa paramètrica (ANOVA) al no ser les dades normals:

```
Precio - planta
KruskalResult(statistic=6.851269616330961, pvalue=0.008857830450562525)
```

Al ser el p-valor del test de kruskal wallis inferior a 0.05 es conclou que es tenen evidències significatives per rebutjar la hipòtesis nul·la d'igualtat de medianes en el preu per les diferents plantes. És a dir, hi han evidències estadístiques per afirmar que hi ha diferència en el preu depenent de la planta on es trobi el pis. Tanmateix rebutjar la hipòtesis nul·la d'aquest test no ens indica quin dels grups és el que presenta diferències en la mitjana. Per això es decideix graficar la relació entre el preu i la planta:



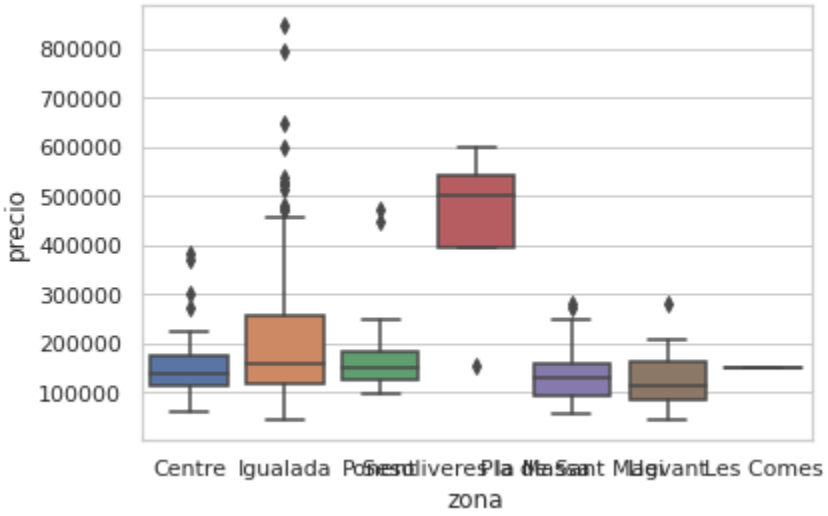
Com es pot observar en el box-plot anterior, sembla que la 2ª planta presenta diferències estadísticament significatives en el preu respecte la resta de plantes, sent el preu més elevat quan el pis té 2 plantes. També sembla que el preu tender a ser més baix en mediana per pisos en plantes ubicades en la 5ª o superior.

Precio - zona

Per analitzar si hi ha diferències en mediana en el preu per tots els valors de la variable *zona* s'utilitza el test de kruskal wallis per comparar més de dos grups atès que no es pot aplicar l'alternativa paramètrica (ANOVA) al no ser les dades normals:

```
Precio - zona
KruskalResult(statistic=473.60686860273904, pvalue=5.2588895511829585e-105)
```

Al ser el p-valor del test de kruskal wallis molt inferior a 0.05 es conclou que es tenen evidències significatives per rebutjar la hipòtesis nul·la d'igualtat de medianes en el preu per les diferents zones. És a dir, hi han evidències estadístiques per afirmar que hi ha diferència en el preu depenent de la zona de la ciutat on es trobi el pis. Això no és d'extranyar perquè ja és conegut que Igualada té zones molt més cares que d'altres en relació amb el preu de la vivenda.
Tanmateix rebutjar la hipòtesis nul·la d'aquest test no ens indica quin dels grups és el que presenta diferències en la mediana. Per això es decideix graficar la relació entre el preu i la zona:



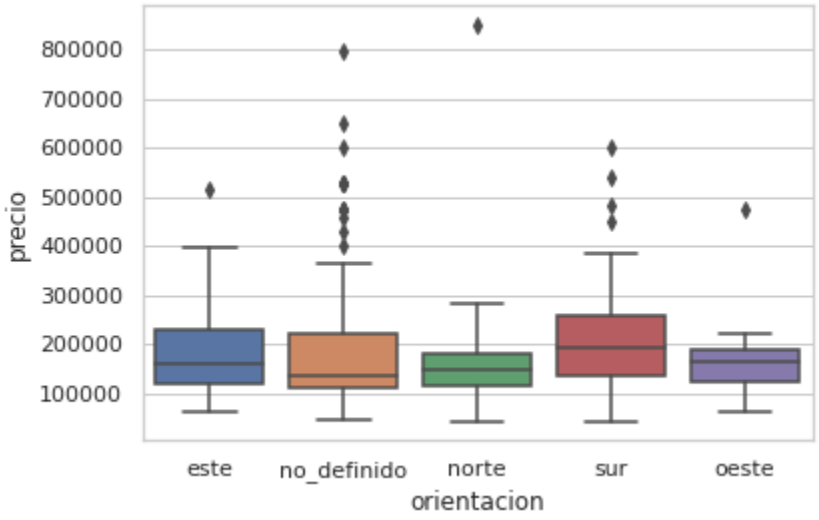
Com es pot observar en el box-plot anterior, sembla que la zona de Ses Oliveres presenta diferències estadísticament significatives en el preu respecte la resta de zones, sent el preu molt més elevat quan el pis o la casa està a Ses Oliveres (zona més rica de la ciutat).

Precio - orientación

Per analitzar si hi ha diferències en mediana en el preu per tots els valors de la variable *orientacion* s'utilitza el test de kruskal wallis per comparar més de dos grups atès que no es pot aplicar l'alternativa paramètrica (ANOVA) al no ser les dades normals:

```
Precio - orientación
KruskalResult(statistic=468.99046101926996, pvalue=5.3143760103939533e-104)
```

Al ser el p-valor del test de kruskal wallis molt inferior a 0.05, es rebutja la hipòtesi nul·la d'igualtat de medianesi es conclou que que es tenen evidències estadísticament significatives de que la mediana del preu no és la mateixa per les diferents orientacions cap on està encarat el pis o la casa. És a dir, hi han evidències estadístiques per afirmar que hi ha diferència en el preu depenent de la orientació del pis.
Tanmateix rebutjar la hipòtesis nul·la d'aquest test no ens indica quin dels grups és el que presenta diferències en la mediana. Per això es decideix graficar la relació entre el preu i la orientació:



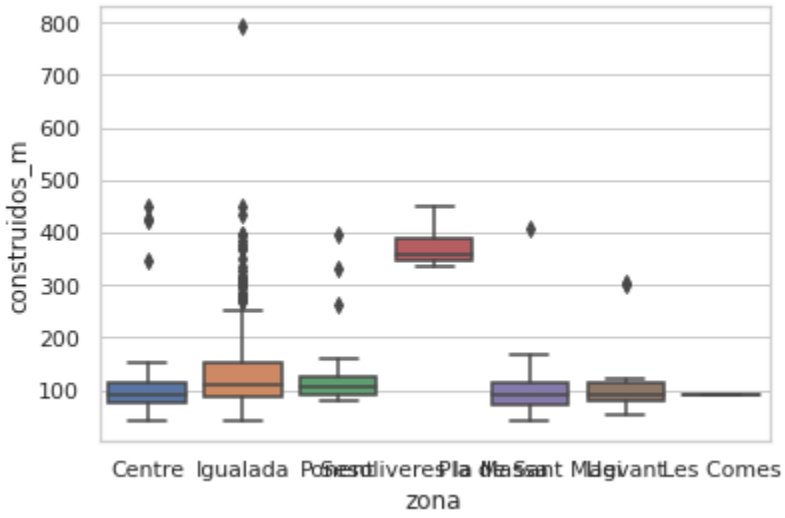
Com es pot apreciar en la figura anterior sembla que la mediana del preu pels pisos que estan orientats al sud és més gran que per la resta. Això no ens hauria d'extranyar ja que els pisos orientats al sud són els més buscats ja que són els que tenen més hores de llum solar directe comparats amb les altres orientacions.

Construidos_m - zona

Per analitzar si hi ha diferències en mitjana en els metres quadrats construïts per tots els valors de la variable *zona* s'utilitza el test de kruskal wallis per comparar més de dos grups atès que no es pot aplicar l'alternativa paramètrica (ANOVA) al no ser les dades normals:

```
Construidos_m - zona
KruskalResult(statistic=473.61994514474395, pvalue=5.224545843288039e-105)
```

Al ser el p-valor del test de kruskal wallis molt inferior a 0.05, es rebutja la hipòtesi nul·la d'igualtat de medianesi es conclou que que es tenen evidències estadísticament significatives per afirmar que la mediana dels metres quadrats construïts dels pisos i les cases és diferent en funció de la zona de la ciutat on es trobin. És a dir, hi han evidències estadístiques per afirmar que hi ha diferència en els metres quadrats construïts en funció de la zona de la ciutat.
Tanmateix rebutjar la hipòtesis nul·la d'aquest test no ens indica quina zona és la que presenta diferències en la mediana. Per això es decideix graficar la relació entre els metres quadrats i la zona:



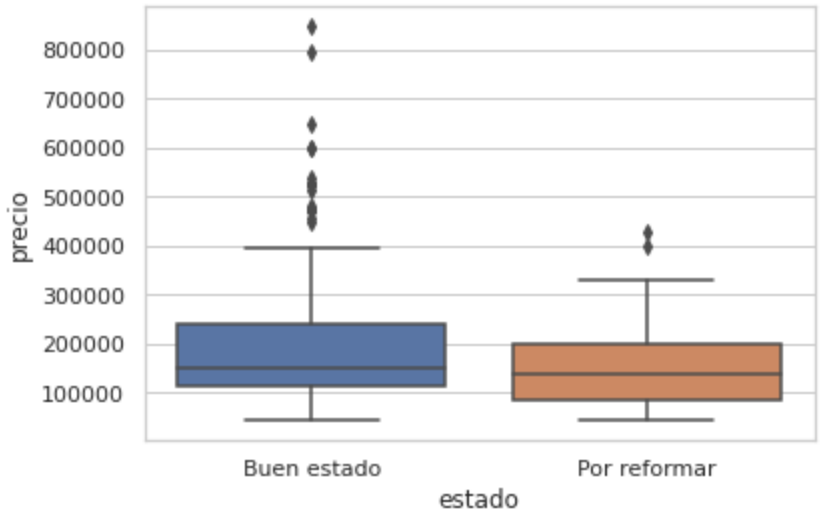
Com es pot apreciar en la figura anterior sembla que la mediana dels metres quadrats construïts és bastant més elevada en els pisos / cases que estan ubicats a la zona de Ses oliveres. Pensem que això té sentit ja que és la zona amb més renda per càpita d'Igualda (més rica) i també una de les més apartades i per tant aglomera la majoria de cases i pisos grans del nostre conjunt de dades.

Precio - estado

Per analitzar si hi ha diferències en mediana en el preu pels dos valors de la variable *estado* s'utilitza el test de Mann-Whitney per comparar dos grups atès que no es pot aplicar l'alternativa paramètrica (test de la T-Student) al no ser les dades normals:

```
Precio - estado
MannwhitneyuResult(statistic=0.0, pvalue=8.786526520427694e-113)
```

Al ser el p-valor del test U de Mann-Whitney molt inferior a 0.05, es rebutja la hipòtesi nul·la d'igualtat de distribucions i es conclou que que es tenen evidències estadísticament significatives per afirmar que la distribució del preu és diferent en funció del estat del pis o casa. És a dir, hi han evidències estadístiques per afirmar que hi ha diferència en el preu en relació a l'estat de conservació de la casa.
Seguidament es presenta el boxplot del preu en funció dels dos nivells de la variable estado per estudiar quin estat és el que a priori presenta una mediana del preu més alt.



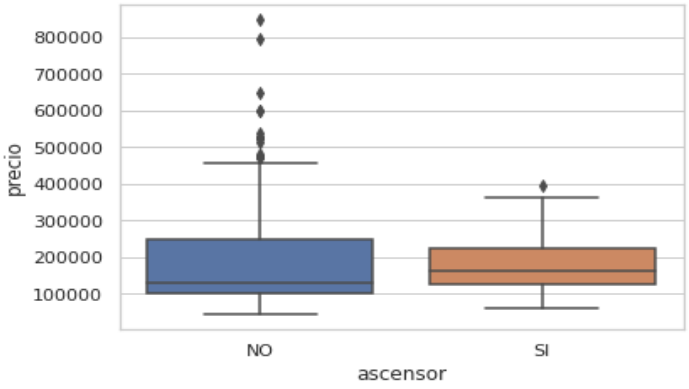
Com es pot apreciar gràcies a la figura anterior, clarament el preu sembla ser major en mediana quan l'estat és "buen estado". Això no és d'extranyar ja que s'espera que el preu sigui més alt per un pis ben conservat. En canvi s'esperaria que el preu fos en mediana més baix quan el pis està per reformar.

Precio - ascensor

Per analitzar si hi ha diferències en mitjana en el preu per tots els valors de la variable *ascensor* s'utilitza el test de Mann-Whitney per comparar dos grups atès que no es pot aplicar l'alternativa paramètrica (test de la T-Student) al no ser les dades normals:

```
Precio - ascensor
MannwhitneyuResult(statistic=0.0, pvalue=1.5835778647746657e-105)
```

Al ser el p-valor del test U de Mann-Whitney molt inferior a 0.05, es rebutja la hipòtesi nul·la d'igualtat de distribucions i es conclou que que es tenen evidències estadísticament significatives per afirmar que la distribució del preu és diferent en funció de si el pis o la casa té ascensor o no. És a dir, hi han evidències estadístiques per afirmar que hi ha diferència en el preu si el pis presenta o no ascensor.
Seguidament es presenta el boxplot del preu en funció dels dos nivells de la variable ascensor per estudiar quin estat és el que a priori presenta una mediana del preu més alt.



Com es pot observar a la figura anterior, clarament el preu sembla ser major en mediana quant el bloc de pisos on està el pis o la casa té ascensor. Altra vegada, això no és d'extranyar ja que el fet de tenir ascensor és quelcom molt ben valorat pels compradors (com és evident, per no haver de pujar escales) i per tant és d'esperar que el preu sigui més elevat si el bloc té ascensor.

MODEL DE PREDICCIÓ DEL PREU

En aquest apartat l'objectiu és el de crear un model de predicció del preu de venda dels pisos i cases en base a les seves característiques. Al tractar-se el preu d'una variable numèrica s'ha optat per construir un model de regressió usant el model Random Forest. Les variables explicatives que s'han afegit al model són les següents:

```
features

['construidos_m',
 'habitaciones',
 'baños',
 'terraza',
 'balcon',
 'estado',
 'orientacion',
 'garaje',
 'año_construido',
 'calefaccion',
 'planta',
 'ascensor',
 'aire_acondicionado',
 'zonas_verdes_o_jardin',
 'zona',
 'etiqueta_energetica']
```

Seguidament es procedeix a transformar totes les variables categòriques en variables dummy (1 o 0) per cadascún dels nivells de cada una de les variables categòriques. Aquest és un pas normal en el procés de construcció d'un model de Machine Learning ja que al codificar les variables d'aquesta manera ajuda a millorar el rendiment del model. Una vegada codificades les variables categòriques es procedeix a realitzar la partició del conjunt de dades en el subconjunts d'entrenament i prova de manera aleatoria. El percentatge d'observacions que es dediquen a la mostra de prova és del 5% del total d'observacions (unes 15).

Posteriorment s'ajusta un model base amb els paràmetres del Random Forest per defecte. Els resultats són els següents:

```
Model Performance
Average Error: 32600.9744 degrees.
Accuracy = 80.33%.
```

S'obté un rendiment del model amb una accuracy del 80.33%. Val a dir que aquests resultats, per ser un model base, són bastant prometedors sobre la mostra de prova. L'accuracy en aquest cas es defineix com 100 - MAPE on MAPE és el Mean Absolute Percentage Error.

Tanmateix es decideix intentar millorar la capacitat predictiva del model mitjançant l'optimització dels hyperparàmetres a través de cerca aleatòria. La cerca aleatòria és un concepte utilitzat freqüentment al món del machine learning quan l'espai d'hyperparàmetres on s'ha de buscar la millor combinació és

relativament gran. Consisteix en provar només una mostra triada aleatòriament entre tot el conjunt d'hyperparàmetres a provar. Cal destacar que, en aquest punt, més iteracions cobririen un espai més ampli dins el conjunt de cerca però incrementar-les augmenta molt el temps de computació. En aquest cas la millor combinació d'hyperparàmetres tobada per aquestes dades és la següent:

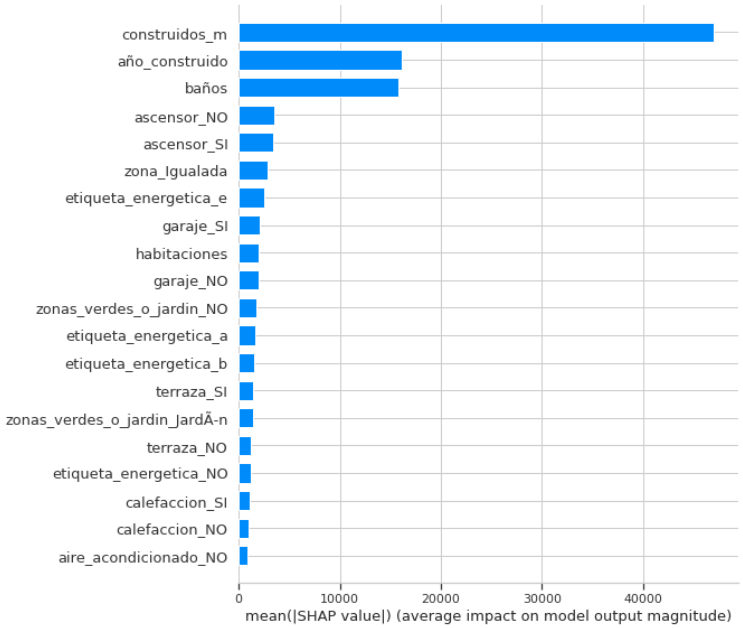
```
Fitting 3 folds for each of 100 candidates, totalling 300 fits
{'bootstrap': True,
 'max_depth': 70,
 'max_features': 'auto',
 'min_samples_leaf': 2,
 'min_samples_split': 2,
 'n_estimators': 1666}
```

Aplicant la millor combinació d'hyperparàmetres trobada s'ajusta un model amb el següent rendiment:

```
best_random = rf_random.best_estimator_
random_accuracy = evaluate(best_random, X_test, y_test)

Model Performance
Average Error: 31471.4298 degrees.
Accuracy = 80.39%.
```

Com es pot observar després d'optimitzar els hyper paràmetres del model Random Forest s'obté un model amb un lleuger increment del rendiment, arribant l'accuracy a un valor de 30.39%

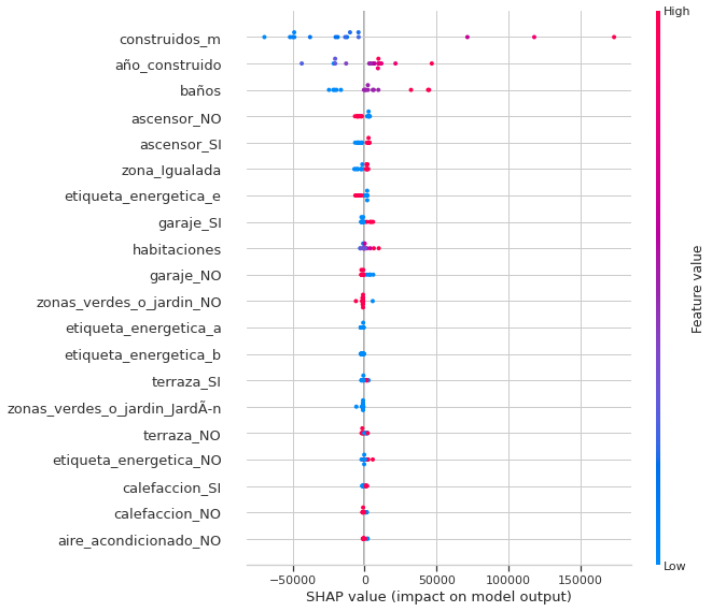
Una vegada obtingut el millor model després d'optimitzar els hyper paràmetres cal interpretar aquest model. Per aquest motiu es grafica a continuació l'importància de la variable utilitzant els valors de SHAP. Aquesta tècnica permet interpretar fàcilment els models de machine learning i ens permet entendre quins són els principals elements que afecten al preu dels pisos en aquest model.



Com es pot apreciar en la figura anterior, la variable més important per definir el preu de venda d'un pis sembla ser els metres quadrats construïts, seguida per l'any de construcció i el nombre de lavabos. El fet

que els metres quadrats construïts sigui la variable més important és quelcom que ja preveiem perquè ja s'ha observat una forta correlació entre aquestes variables.

La següent gràfica presenta els valors de SHAP i l'impacte que tenen cadascuna de les variables en el model (positiu o negatiu)



Com es pot apreciar en la gràfica anterior, les variables més importants per determinar el preu d'un pis són els metres quadrats construïts i després l'any de construcció i el nombre de banys. La resta de variables tenen un efecte marginal en el preu de venda.
Quant als metres quadrats construïts, l'efecte que aquests tenen en el preu de venda és el següent. Valors baixos de construidos_m fan reduïr el preu de venda. És a dir, pisos amb pocs metres quadrats veuen el seu preu de venda reduit. En canvi valors elevats d'aquesta variable fan que que el preu augmenti, el que bàsicament significa que pisos o cases més grans són més cars.
Quant a l'any de construcció sembla que passa una mica el mateix. Valors baixos d'aquesta variable fan que el preu de venda baixi, mentre que els pisos o les cases més nous tendeixen a ser venuts per un preu més elevat.
Quant a la variable baños la interpretació va en línea amb el sentit comú. Pisos que tenen menys lavabos emputxen el preu de venda cap abaix. Pisos que tenen més lavabos fan que el preu de venda sigui superior.

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Els resultats obtinguts a partir de l'anàlisi de les dades sí permeten respondre al problema, que era el d'analitzar quins són els factors que afecten al preu de venda dels pisos i les cases a Igualada.

Una de les conclusions que es desprenen de l'anàlisi és que el preu depèn principalment dels metres quadrats, també del nº d'habitacions i de lavabos (més lavabos que habitacions). L'any de construcció del pis també és un factor important per determinar el preu ja que com més nous són els pisos més cars costen.

També s'ha arribat a la conclusió de que com més metres quadrats tingui un pis o una casa, més habitacions i lavabos tindrà. Aquesta és una conclusió molt en línea amb el sentit comú ja que no tindria massa sentit fer un pis amb molts metres quadrats que tingués només 1 habitació o lavabo.

Una altra conclusió a la que s'arriba després de l'anàlisi és que tant el preu com els metres quadrats depenen molt de la zona de la ciutat en la que es troben. La zona més rica de la ciutat, Ses Oliveres, és que la que concentra els pisos més cars així com els més grans en metres quadrats construïts.

Pel que respecta a l'orientació s'ha pogut comprovar que, en mediana, els pisos orientats al sud són més cars que els que estan orientats a qualsevol altre direcció.

Finalment ens agradaria destacar que s'ha pogut construir satisfactòriament un model de predicció del preu de venda dels pisos utilitzant un model Random Forest i que s'ha obtingut un rendiment gens menyspreable del voltant del 80% d'accuracy. Gràcies a aquest model es pot arribar a la conclusió, i quantificar, quin és l'impacte que té en el preu els metres quadrats construïts, que com ja s'ha discutit és la variable més important del conjunt de dades.

Taula de contribucions

Contribucions	**Signatura**
Investigació prèvia	A.M.O | V.B.C
Redacció de les respostes	A.M.O | V.B.C
Desenvolupament del codi	A.M.O | V.B.C