Pràctica 1

Components:

Xavier Sellart Aldomà i Mireia Gutiérrez Lozano

***1. Context**. Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació.*

Els preus dels pisos a Barcelona varien segons diversos característiques com poden ser el barri, la mida del pis o el nombre d'habitacions. S'ha fet web scrapping del portal de venta de pisos "Habitaclia" per tal de poder extreure informació dels diferents pisos que hi ha actualment a la venta. Aquest conjunt de dades ofereix una bona visió general dels preus dels pisos de Barcelona en funció de nombroses variables.

***2. Títol.** Definir un títol que sigui descriptiu pel dataset.*

Dades dels pisos en venta a Barcelona

***3. Descripció del dataset**. Desenvolupar una descripció breu del conjunt de dades que s'ha extret. És necessari que aquesta descripció tingui sentit amb el títol escollit.*

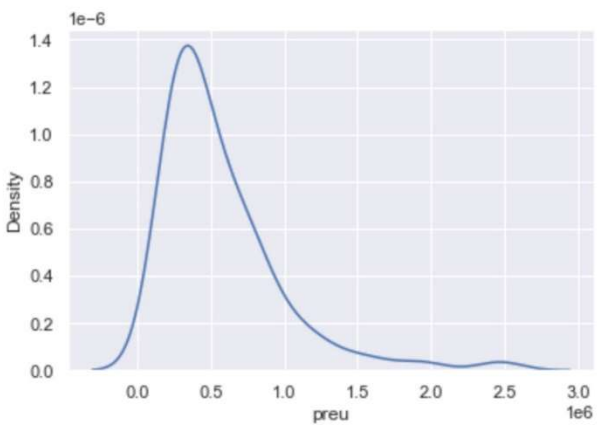
El conjunt de dades conté informació sobre múltiples característiques i propietats dels pisos que es trobaven disponibles a la web d'Habitaclia, com ara la superfície del pis, el preu de compra, el barri on està ubicat el pis, nombre d'habitacions i lavabos, classificació energètica... També s'inclouen dos atributs més generals, que són "Característiques generals" i "Equipament comunitaris" en les que es llisten les característiques que no estan incloses en la resta d'atributs del joc de dades.

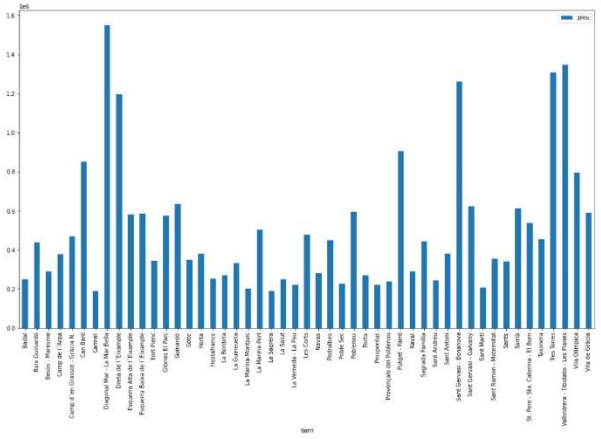
Les dades del dataset encara no han passat per un procés de preprocessat o neteja, per això existeixen inconsistències i el format no és necessàriament el més adequat. Un exemple de inconsistència es pot veure en l'atribut dels m² dels pisos: alguns pisos no tenen un valor numèric, sinó que tenen la paraula "desde", ja que l'anunci del qual s'ha fet l'scrapping anuncia un conjunt de pisos i indica la superfície del més petit (per exemple: desde 100m²).

Un exemple de dades que no tenen el format més adequat podria ser el de les columnes de característiques generals o equipament comunitari. Les diferents característiques o equipaments s'inclouen en únic camp, separant cada característica o equipament amb el símbol "+". En la fase de preprocessament es podria modificar aquests camps aplicant la tècnica de One-Hot encoding, més adient per aquests casos.

4. Representació gràfica. *Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.*

Les dues imatges escollides per identificar el dataset i el projecte ens donen informació sobre els preus del pisos a Barcelona. La primera és un "density plot" que indica la distribució de preus del dataset, i la segona un "barplot" que indica la mitjana dels preus dels pisos a cada barri de Barcelona.





5. Contingut. *Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.*

Aquest conjunt de dades conté dades dels pisos que havia en venta a la web d'Habitaclia el dia 10/04/2022, que es quan es va fer web scrapping de la pàgina web per tal d'obtenir el conjunt de dades.

El codi utilitzat per fer el scrapping començar a partir del següent link https://www.habitaclia.com/viviendas-barcelona.htm. En aquest enllaç i hi ha diversos enllaços a anuncis de pisos i el que fa el codi i és obtenir aquests enllaços per després accedir-hi i obtenir la informació de cada pis. Un cop ha acabat amb tots els pisos del primer enllaç avança al següent, que seria https://www.habitaclia.com/viviendas-barcelona-1.htm, i repeteix el mateix procediment d'obtenir tots els enllaços, descarregar la informació dels pisos i avançar a la següent pàgina. El codi segueix aquest procediment fins que no existeix una pàgina següent, el que vol dir que s'han inspeccionat tots els habitatges de Barcelona que hi ha a la web d'Habitaclia.

El conjunt de dades conté 15 columnes. A continuació es descriu la informació que conté cadascuna de les columnes:

- Títol: String que descriu de manera general el pis. Exemple: *"Piso modernista en venta situado en la rambla de catalunya en Barcelona".*
- Barri: String amb el nom del barri on està situat el pis. Exemple: *"Dreta de l´Eixample".*

- m²: Integer que indica la quantitat de metres quadrats del pis. Exemple*: "189".*
- Habitacions: Integer que indica el nombre d'habitacions que té el pis. Exemple: *"4".*
- Lavabos: Integer que indica el nombre de lavabos que té el pis. Exemple: *"3".*
- Preu/m²: Integer que indica el preu per metre quadrat del pis, en €/m². Exemple: *"5.794".*
- Preu: Integer que indica el preu del pis, en €. Exemple: *"1.095.000".*
- Immobiliària: String que indica quina immobiliària ha publicat l'anunci del pis. Exemple: *"APROPERTIES REAL ESTATE BARCELONA Nº Aicat 6388".*
- Característiques generals: Llista de diferents característiques que té el pis. Exemples d'algunes d'aquestes característiques són: si té calefacció, si té aire condicionat, si està proper a transport públic... Exemple: *"+ Calefacción + Amueblado + Aire acondicionado + Año construcción 1958".*
- Equipament comunitari: Llista de diferents equipaments comunitaris que pot tenir la comunitat de veïns del pis. Exemples d'alguns d'aquests equipaments són: si té ascensor, si té vigilància, quina és la quota de comunitat... Exemple: *"+ Ascensor + Piscina comunitaria".*
- Consum etiqueta: Character que indica la qualificació energètica del pis en funció del seu consum energètic. Els valors van de A (millor qualificació) a G (pitjor qualificació). Exemple: *"E".*
- Consum kW: Integer que indica els kW h m² / any consumits. Exemple: *"180".*
- Emissions etiqueta: Character que indica la qualificació energètica del pis en funció de les seves emissions. Els valors van de A (millor qualificació) a G (pitjor qualificació). Exemple: *"D".*
- Emissions kg: Integer que indica els kg CO_2 m² / any generats. Exemple: *"19".*
- Última modificació: String que indica en quina data es va produir l'última modificació de l'anunci, en format dd/mm/aaaa. Exemple: "21/03/2022"

6. ***Agraïments.*** *Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.*

El propietari de les dades que hem extret és Habitaclia, que va crear el domini «habitaclia.com» al Gener del 2002.

Habitaclia pertany al grup Adevinta, que és una de les majors empreses de publicacions online a occident mitjançant plataformes per ajudar a les persones a vendre i comprar tota mena

d'articles. A Espanya, té marketplaces com Fotocasa, habitaclia, InfoJobs, coches.net, motos.net i Milanuncios.

Respecte a anàlisi similars, s'han fet múltiples estudis per crear models que facin prediccions de preus a partir de variables. Per exemple, a Kaggle tenim un estudi realitzat per Burhan Y. Kiyakoglu, que utilitza varis mètodes de regressió per tal de predir preus de cases a King Country (USA).

https://www.kaggle.com/code/burhanykiyakoglu/predicting-house-prices

Un altra possibilitat és crear una base de dades per analitzar l'evolució del preu del m^2 al llarg del temps o en funció del barri o ciutat. Podem veure un exemple en el repositori de GitHub House-Price-evolution-in-Madrid.

Aquestes dades en permetrien realitzar millors inversions o analitzar les possibles causes d'aquestes evolucions per tal de prendre decisions amb fonament.

https://github.com/dlopezmaci001/House-Price-evolution-in-Madrid

També podríem analitzar quines característiques estan millor valorades, així com el barri o el temps que s'ha trigat en vendre un pis concret i si es veu afectat per una mala valoració del preu de venta o la ubicació.

Finalment, per tal d'actuar d'acord amb els principis ètics i legals, en primer lloc hem comprovat l'arxiu robots.txt d'Habitaclia, que limita algunes opcions, però no exclou als robots per tal de rastrejar.

A més a més, al estar recol·lectant dades publicades sense necessitat d'iniciar cap sessió ni acceptar condicions que ens prohibeixin explícitament el web scrapping, considerem que és informació pública.

Per últim, la base de dades creada no té cap finalitat comercial. S'ha generat per tal d'analitzar el sector immobiliari a Barcelona i es podria aplicar a altres municipis.

7. ***Inspiració.*** *Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.*

Aquest conjunt de dades és interessant ja que el sector immobiliari és molt important a nivell espanyol i es veu substancialment afectat per factors externs de la època en la que vivim. Per exemple, la pandèmia ha provocat canvis en las preferències dels compradors, que valoren més tenir terrassa o tenir uns casa espaiosa a un tuguri ben ubicat.

Aquesta variabilitat en el mercat immobiliari és la que suposa un repte alhora d'establir preus o determinar si el preu d'un pis està sobrevalorat o infravalorat. En el nostra cas, la base de dades generada ens permetria crear un model per fer prediccions de preus, tal com es va realitzar a l'estudi realitzat per Burhan Y. Kiyakoglu, però a la ciutat de Barcelona. També pot

ser interessant determinar quines són les característiques que influeixen més en el preu dels pisos, per tant, les més valorades pels compradors. Això ens permetria fer un estudi més sociològic sobre els impactes que tenen les circumstàncies actuals en el sector immobiliari de Barcelona i si es manté en el temps, ja que podríem tornar a extreure les dades anualment per veure la progressió, tal com es va fer en l'estudi House-Price-evolition-in.Madrid.

En el nostra cas, volem cobrir les necessitats de les persones que busquen pis per tal de que puguin establir en quins rangs de preus haurien d'estar els pisos en els que estan interessats.

*8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció:*

Llicència CC BY-NC-SA 4.0.

Aquesta llicència permet copiar i redistribuir la base de dades lliurament, així com transformar o adaptar la informació que aquesta conté a la conveniència de qui la vulgui utilitzar. Tot i això, requereix que es doni el crèdit pertinent al propietari del coneixement inicial i indicar els canvis realitzats en la nova distribució.

També exigeix que les noves distribucions que hagin realitzat transformacions o desenvolupat nou contingut a partir de la base de dades original estiguin sota la mateixa llicència CC BY-NC-SA 4.0.

Finalment, l'últim requeriment d'aquesta llicència és que no es faci un ús comercial de la informació continguda en la base de dades.

Hem decidit utilitzar aquesta llicència ja que nosaltres no tenim intenció de lucrar-nos amb aquest projecte. A part, tenint en compte les possibles conseqüències legals de comercialitzar informació extreta per web scrapping, aquesta llicència ens eximeix de infraccions relacionades amb els drets d'autor o marca registrada.

*9. **Codi.** Adjuntar al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.*

Aquí adjuntem l'enllaç al repositori de Github:

https://github.com/mirgutloz/Practica-1-Pisos-Barcelona

*10. **Dataset.** Publicar el dataset obtingut(*) en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.*

Aquí adjuntem l'enllaç al dataset penjat a Zenodo:

https://zenodo.org/record/6448642#.YlRejchByUk.

*11. **Vídeo**. S'ha de lliurar un vídeo explicatiu de la pràctica on cadascun dels integrants del grup expliqui amb les seves pròpies paraules tant les respostes del projecte com el codi utilitzat per a dur a terme l'extracció. El vídeo ha de ser enviat a través d'un enllaç a Google Drive que heu de proporcionar, juntament amb l'enllaç al repositori Git, al moment de lliurar la pràctica*

Aquí adjuntem l'enllaç vídeo:

https://drive.google.com/file/d/1Zh7t518p7nrcHUPPcdTgRT15fAe302f7/view?usp=sharing

*12. **Taula de contribucions al treball***

Contribucions	Signatura
Investigació prèvia	XSA, MGL
Redacció de les respostes	XSA, MGL
Desenvolupament del codi	XSA, MGL