PRA1 – Selecció del conjunt de dades

Nom: Albert Salvador Yuste

Descripció i enunciat

Aquesta activitat, primera part de la pràctica final, consisteix en la selecció per part de l'estudiant d'un conjunt de dades del seu interès que serà usat en el projecte de creació de la visualització de dades, d'acord amb uns criteris establerts.

Dataset utilitzat

Per la elaboració d'aquesta pràctica he escollit un segui de datasets. Les fonts dels datasets escollits són les següents:

- Determinants of Airbnb prices in European cities: A spatial econometrics approach. Zenodo. (https://zenodo.org/record/4446043#.ZF4fRHZByUk)
  - Llicència: https://creativecommons.org/licenses/by/4.0/legalcode
- Transit system of the world. Kaggle (https://www.kaggle.com/datasets/citylines/city-lines)
  - Llicència: https://opendatacommons.org/licenses/odbl/1-0/

En el primer enllaç es poden observar un seguit de datasets de diferents ciutats europees (Amsterdam, Atenes, Barcelona, Berlín, Budapest, Lisboa, Londres, París, Roma i Vienna) en què es llisten els preus de diferents anuncis de Airbnb, així com altres característiques. Disposem de dos datasets per a cada ciutat, un per als preus durant els dies laborables i un altre amb els dels caps de setmana (resultant un total de 20 datasets).

El segon enllaç conté també un seguit de datasets amb informació sobre línies i estacions de transport públic, bàsicament de metros, tramvies, etc., de moltes ciutats arreu del món. Dels datasets que aquí hi ha inicialment se'n plantegen fer servir tan sols dos: un amb informació de les ciutats que conté el dataset (cities.csv) i un altre amb la localització de les estacions de transport públic (stations.csv). Si durant el transcurs de la pràctica es considera adequat o necessari incorporar-ne més informació entre la que hi ha disponible, potencialment es poden incorporar més datasets d'aquest mateix enllaç.

Qüestió 1 [10%]

Justifiqueu breument la vostra selecció, sigui per motius personals o professionals.

Tot i que s'han seleccionat dos conjunts de datasets, el principal objectiu és proporcionar un element més d'anàlisi sobre els preus que tenen els pisos, apartaments, cases, etc., disponibles a Airbnb, tot procurant conèixer més bé quins són els elements principals que influeixen en el preu d'aquests, tant des d'un punt de vista de característiques de l'inmoble com de la seva localització. Així, la motivació d'incorporar informació addicional de transport públic al conjunt de datasets originals preveu millorar-lo i aportar dades i informació que pot ser d'interès.

El meu interès no obstant no ve des d'un punt de vista ni personal ni professional, però sí que considero que pot ser una font d'informació prou interessant de com està l'estat

d'aquest tipus de negocis en algunes de les principals ciutats europees, algunes d'elles especialment afectades per les conseqüències que sovint se'n deriven d'aquests negocis.

Qüestió 2 [10%]

La rellevància del conjunt de dades en llur context. Són dades actuals? Tracten un tema important per algun col·lectiu concret? S'ha tingut en compte la perspectiva de gènere?

Les dades de les característiques dels immobles s'han extret de l'enllaç anteriorment indicat, el qual indica que es va publicar el 13 de gener de 2021, així que suposem que són anteriors a aquesta data. De fet aquestes dades van ser utilitzades en la elaboració d'un article (https://www.sciencedirect.com/science/article/pii/S0261517721000388), i en cap moment aquest menciona res del COVID-19, per lo que no seria agosarat arribar a concloure que aquestes dades són prèvies a la pandèmia.

Al respecte del segon dataset amb informació de transport públic, la informació proporcionada indica que no s'actualitzen les dades des de fa quelcom més de 4 anys, així que podem estimar que deuen estar desactualitzades des de la segona meitat de 2018 i primera meitat de 2019. Com que en general no s'inauguren moltes estacions de metro de manera freqüent en els últims anys, alhora ja ens pot interessar tenir dades sutilment desactualitzades, de tal manera que creuar-les amb les anteriors pugui ser més correcte.

Els apartaments d'Airbnb i webs similars en els últims 10 anys han assolit una popularitat indiscutible, tant per lo que han aportat a molts viatgers/es que volien establir-se temporalment en immobles més tradicionalment turístics com ara hotels, hostals, albergs, etc., com per les conseqüències que han aportat a les ciutats i poblacions on aquests es localitzen, amb l'aparició de molts pisos turístics, els quals en alguns indrets a suposat un increment, sovint descontrolat, del volum de turisme.

Realitzar un anàlisi de la influència dels preus segons les característiques de l'immoble i la localització no és potser quelcom que pugui aportar solució a molts dels problemes que aquesta pràctica ha derivat, però sí que pot aportar un element comparador quan es pugui comparar amb els pisos i cases de lloguer tradicional, les quals estiguin en localitzacions similars o properes a immobles de lloguer temporal.

Qüestió 3 [25%]

La complexitat (mida, variables disponibles, tipus de dades, etc.). Té de l'ordre de centenars o milers de registres? Té de l'ordre de desenes de variables? Combina dades categòriques i quantitatives? Inclou altres tipus de dades? Eviteu els conjunts excessivament simples.

El conjunt de datasets original dels immobles de Airbnb (els quals són, juntament amb el segon conjunt de datasets d'informació de metros, en format CSV) tenen els següent atributs:

- realSum: el preu de l'immoble per a dues persones i dues nits (en €)
- room_type: el tipus d'immoble (categòric)
- room_shared: si té habitacions compartides (veritat o fals)
- room_private: si té habitacions privades (veritat o fals)
- person_capacity: el nombre màxim d'hostes (numèric)
- host_is_superhost: si el propietari té títol de "super-host" (veritat o fals)

- multi: si el propietari té també en propietat entre 2 i 4 ofertes (veritat o fals)
- biz: si el propietari té també en propietat més de 4 ofertes (veritat o fals)
- cleanliness_rating: puntuació de neteja (numèric)
- guest_satisfaction_overall: puntuació global de l'immoble (numèric)
- bedrooms: nombre d'habitacions (0 per estudis) (numèric)
- dist: distància des del centre de la ciutat (en km) (numèric)
- metro_dist: distància des del metro més proper (en km) (numèric)
- attr_index: índex corresponent a nombre d'elements d'interès (numèric)
- attr_index_norm: índex anterior normalitzat (0-100) (numèric)
- rest_index: índex corresponent a restaurant propers (numèric)
- attr_index_norm: índex anterior normalitzat (0-100) (numèric)
- lng: coordenada longitud de l'immoble (numèric)
- lat: coordenada latitud de l'immoble (numèric)

Així doncs, inicialment estem treballant amb un total de 19 variables, la majoria numèriques o booleanes, i alguna de categòrica. No obstants, algunes variables com puntuacions, índexs, nombre de persones, etc., es podrien considerar també com a categòriques si es considerés necessari i adient. A més a més d'aquestes, fàcilment en podem incloure dues més de categòriques: una d'elles corresponent a la ciutat del qual prové el dataset (de la llista anterior enumerada) i si el preu correspon o no a valors de cap de setmana (de divendres a diumenge) o no (de dilluns a dijous).

Alhora, s'ha de tenir en compte que s'incorporaran més variables per a cada registre, com per exemple el metro més proper, la localització d'aquest, quantes estacions de metro hi ha en una certa distància, el tipus de transport públic, etc. Aquestes variables s'aniran incorporant segons el què es consideri de més interès.

En quant el nombre de registres encara no els he ajuntat formalment en un dataset únic, però després d'un anàlisi preliminar he calculat que tenim un dataset global d'aproximadament 51.707 registres (tenint sempre present que un mateix immoble aparegui dues vegades segons si es presenta el preu de cap de setmana o no).

Qüestió 4 [25%]

L'originalitat. No repetiu els conjunts de dades clàssics. Podeu, però, combinar-ne o millorar visualitzacions existents. Així, hi ha altres visualitzacions basades en aquest conjunt de dades? És una evolució o actualització d'un conjunt anterior? Heu enriquit un conjunt de dades ja existent?

No he sapigut identificar que el dataset original en qüestió sigui un conjunt de dades clàssic, doncs és relativament nou. Alhora, he trobat alguns anàlisis que altres usuaris, per exemple en Kaggle, han realitzat sobretot basant-se en el conjunt de datasets originals, és a dir, que han utilitzat la mateixa font que estic pretenent també fer servir jo en aquesta pràctica.

Alguns enllaços d'anàlisis d'altres usuaris:

- https://www.kaggle.com/code/micaeld/airbnb-eda-and-prediction
- https://www.kaggle.com/code/antonisraptakis/airbnb-a-data-science-mini-project

Algunes de les visualitzacions que realitzen segurament seran potencials de ser incorporades en el meu estudi, i tot i que potser algunes no són susceptibles de ser actualitzades, es procuraran millorar i aportar-ne de noves.

També tot tenint present que, tot i que el dataset original amb els preus dels pisos conté un paràmetre de distància des del l'estació de metro més propera, potser es podria enriquir aquest dataset tot proporcionant quin metro dels disponibles en el segon conjunt de datasets és el més proper, com ja s'ha comentat anteriorment, així com per exemple es podria indicar un nombre d'estacions de metro que estiguin disponibles a menys de X quilòmetres, entre d'altres, tot aportant nous paràmetres que ens permetin enriquir les dades originals, aportants nous punts de vista i visualitzacions fins ara no aportades en cap altre estudi.

Qüestió 5 [30%]

Les qüestions que respondreu amb la visualització de dades, tenen en compte els punts anteriors? Estan ben plantejades? Són adequades pel conjunt de dades triat?

En aquest apartat llistarem algunes de les preguntes que poden resultar d'interès resoldre, així com d'altres que poden sorgir durant el desenvolupament de la pràctica. S'anirà valorant la seva adequació de representació o no, així com el mètode, segons l'evolució de la pràctica.

- Totes les ciutats presenten nombre similar d'anuncis?
- Elements com el nombre màxim de possibles d'hostes, el nombre d'habitacions (potencialment relacionats amb el tamany de l'immoble) repercuteix en el preu?
- El tipus d'immoble té algun impacte en el preu?
- Com influeix en el preu el fet que sigui cap de setmana o no? És similar per a totes les ciutats?
- La puntuació de l'immoble influeix en el preu?
- La puntuació presenta diferències segons la regió de cada ciutat?
- Les proporcions entre tipus d'immoble ofertats (caldrà detectar possibles elements duplicats) són similars en totes les ciutats? Presenten diferències segons la regió de cada ciutat?
- Les distàncies amb les estacions de metro són similars entre ciutats?
- Les distàncies respecte al centre o al metro repercuteix en el preu? En totes les ciutats?
- Els immobles tenen moltes estacions de metro properes? Influeix el tipus d'immoble?
- Existeixen algunes estacions de metro en què se'n destaqui un gran nombre d'anuncis al voltant?
- Hi ha diferències en el preu segons si el propietari té l'etiqueta de "super-host"?
- Tenir una puntuació alta de restaurants propers presenta alguna influència? I dels elements turístics d'interès?
- Quina distribució presenten els dos anteriors índexs en cada ciutat?
- Hi ha ciutats amb un major nombre d'anuncis amb habitacions compartides que d'altres?

Aquestes i d'altres preguntes que aniran sorgint durant l'evolució de la pràctica s'intentaran respondre i representar en la següent part de la pràctica.