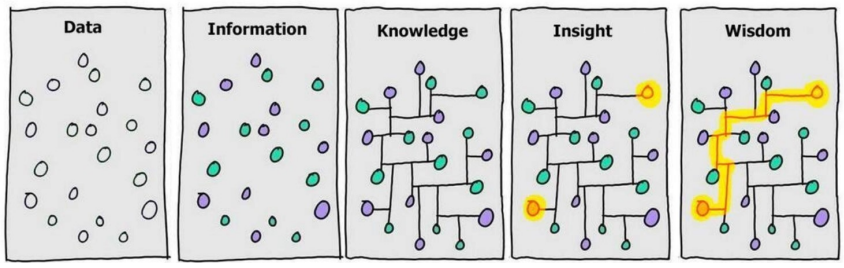
Pràctica 1
El poder de les dades: El cas Airbnb

- El termini d'entrega de la pràctica finalitza el **23 d'octubre de 2.024 a les 23:55**.
- L'entrevista de la pràctica es farà durant la sessió de laboratori del **24 d'octubre de 2.024**.
- Al Campus Virtual heu de penjar l'**informe PDF** explicant com heu resolt la pràctica i tots els **codis/scripts** utilitzats.
- La pràctica es realitza en grup de **dues o tres persones**.
- La nota d'aquesta pràctica equival a un **15%** de la nota global de l'assignatura.

Motivació i objectius

En l'era digital en què vivim, les dades representen un actiu fonamental per a les organitzacions en la consecució dels seus objectius i la seva competitivitat. No obstant, tindre les dades no és suficient, cal analitzar-les i interpretar-les de manera adient per tal de descobrir tendències o patrons, predir comportaments, anticipar-se a esdeveniments, detectar problemes precoçment i prendre decisions estratègiques informades. Així doncs, les organitzacions que saben aprofitar adequadament les dades es troben en una posició privilegiada per a adaptar-se al canvi, innovar i mantenir-se competitives en un entorn empresarial cada vegada més complex i dinàmic. Per aquest motiu, existeix el dit "la informació és poder". Tots aquests recursos s'utilitzen a dia d'avui dins l'àmbit del *data science*, una disciplina cada vegada més important dins les organitzacions[1].



En aquesta pràctica, treballarem habilitats pràctiques en l'anàlisi exploratòria de dades utilitzant conjunts de dades proporcionats per Airbnb. Més concretament, s'espera que els alumnes puguin:

- Comprendre l'estructura i el contingut de les dades d'Airbnb.
- Aplicar tècniques de neteja de dades per abordar problemes comuns com ara valors desconeguts i valors atípics (*outliers*).

[1] T. H. Davenport and D. J. Patil, "Is Data Scientist Still the Sexiest Job of the 21st Century?", in Harvard Business Review, July 2022, URL: `https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century`

- Realitzar un anàlisi descriptiu complet, utilitzant eines estadístiques i de visualització per explorar les característiques bàsiques de les dades.
- Identificar patrons, tendències i correlacions importants dins de les dades d'Airbnb.
- Interpretar i comunicar de manera efectiva els resultats de l'anàlisi exploratòria.

Conjunts de dades

Impulsats per la filosofia de les *dades obertes*, Airbnb s'ha unit a aquesta iniciativa amb el projecte *Inside Airbnb*, on proporcionen dades d'allotjaments de més d'un centenar de ciutats de tot el món. Aquestes dades són accessibles a través del següent enllaç: `https://insideairbnb.com/get-the-data`.

> En aquesta pràctica, heu d'analitzar les dades de, com a mínim, dues ciutats. L'elecció d'aquestes ciutats és decisió de cada grup.

Per cada ciutat, trobareu una sèrie de fitxers. Per a realitzar aquesta pràctica, us recomanem que us centreu en els següents:

- Fitxer *listings.csv.gz*: Fitxer CSV que conté els detalls dels allotjaments d'aquella ciutat. Cada allotjament té unes 70 dimensions/variables indicant les seves característiques (dimensions *name*, *description*, *property_type*, *room_type*, *bedrooms*, *bathrooms*, *amenities*,...), ubicació (dimensions *latitude*, *longitude*, *neighbourhood*, *neighbourhood_cleansed*,...), propietari (dimensions *host_**), disponibilitats (dimensions *price*[2], *minimum_nights*, *maximum_nights*, *availability_**, ...), mètriques de ressenyes i puntuacions (dimensions *number_of_reviews*, *review_scores_**, ...).
- Fitxer *neighbourhoods.geojson*: Fitxer GeoJSON[3] que conté la informació geogràfica dels barris d'aquella ciutat.
- Si teniu qualsevol dubte sobre el significat d'algun fitxer/dimensió, podeu consultar el recurs *Data Dictionary* de l'enllaç proporcionat.

Part 1: Anàlisi exploratòria de dades

Una anàlisi exploratòria (en anglès, *exploratory data analysis*) és un tractament estadístic utilitzat per explorar, descriure, resumir i entendre les dades amb les que es treballen. Aquestes anàlisis són extremadament útils abans de realitzar anàlisis més avançades com, per exemple, mineria de dades o filtratge col·laboratiu.

Per iniciar-vos, haureu d'estudiar les variables individualment per tal d'explorar les característiques bàsiques de les dades. En funció de si les variables són qualitatives o quantitatives, haureu d'aplicar les tècniques adients per tal d'interpretar aquelles dades.

[2] Tot i que en els fitxers sempre aparegui el símbol '$', el valor numèric està en funció de la moneda d'aquell país. Per exemple, a les ciutats de la UE i UK significa euros i lliures esterlines, respectivament.

[3] Podeu visualitzar el contingut d'aquest tipus de fitxers amb la eina on-line `https://geojson.io`

Això inclou calcular taules de freqüències i estadístics descriptius com medianes, mitjanes, quartils, desviacions estàndard i d'altres explicats a classe, així com generar-ne les visualitzacions oportunes amb gràfics de barres o de sectors, histogrames i diagrames de caixa, entre d'altres. Més endavant, us haureu d'endinsar a estudiar correlacions entre múltiples variables utilitzant models de regressió (lineals o no lineals), càlculs de coeficients de correlació, tests estadístics i gràfics de dispersió. Estudiar aquestes variables i correlacions us permetrà fer comparacions entre allotjaments de diverses ciutats i esbrinar si hi ha similitud entre allotjaments de diverses ciutats.

Per exemple, podeu estudiar si les característiques o la disponibilitat dels allotjaments són similars entre ciutats, si la distribució dels preus dels allotjaments varia entre ciutats (o ho fa en funció d'alguna altra variable), si hi ha propietaris que dominen el mercat, o què fa que els allotjaments tinguin millors puntuacions. Aquests són el tipus d'estudis que s'espera que estudieu en l'anàlisi exploratòria de dades. Sou lliures de plantejar-vos els estudis que creieu d'interès. Es valorarà, a més de la quantitat d'estudis realitzats, la qualitat, la dificultat, la variabilitat i l'originalitat dels mateixos.

> A l'informe, haureu indicar clarament quins estudis heu realitzat. Per cadascun d'ells, haureu d'explicar el procediment seguit per a la seva resolució: des de les variables utilitzades i les tècniques i estadístics emprats, fins a la presentació dels resultats mitjançant gràfiques i les conclusions que n'extraieu. És a dir, no us quedeu només amb la resposta o la visualització: argumenteu com heu plantejat i resolt cada problema.

Podeu utilitzar els llenguatges de programació (Python, R, Java,...) i les llibreries (pandas, NumPy, Matplotlib, seaborn, ggplot2,...) que preferiu.

Part 2: Visualitzacions geogràfiques

L'anàlisi de dades des d'una vessant geogràfica és essencial en sistemes com Airbnb, on la ubicació juga un paper crucial. Mitjançant l'anàlisi geogràfic, és possible comprendre millor la distribució espacial dels allotjaments, identificar patrons i tendències específiques de cada regió i entendre com els factors geogràfics influeixen en els preus i la demanda dels llocs d'allotjament.

Amb tot el coneixement adquirit durant la primera part de la pràctica, haureu de crear visualitzacions geogràfiques per representar les variables que creieu més interessants des d'una perspectiva geogràfica. Per això, haureu de crear mapes de coropletes (amb l'ajuda dels fitxers GeoJSON), categories, bombolla i/o calor (totes ells amb l'ajuda de la latitud i longitud dels allotjaments). Per exemple, podeu representar el nombre, valoració mitjana o preu mitjà dels allotjaments de cada barri de les ciutats estudiades, o la distribució d'un determinat tipus d'allotjaments a les ciutats. Això us permetrà veure si existeixen certs patrons espacials o variacions de certs factors en diferents àrees.

Enlloc de crear visualitzacions estàtiques, podeu afegir-hi funcionalitats per crear visualitzacions interactives i millorar l'experiència dels usuaris. Per exemple, podeu afegir

selectors de capes (*layers control*) que permeten sobreposar mapes, utilitzar diverses capes base accedint a diferents proveïdors TMS (Tile Map Service), personalitzar escales de color i implementar interaccions amb l'usuari (captura d'esdeveniments com mostrar informació addicional quan es fa click sobre un element del mapa).

> A l'informe, haureu indicar clarament quins estudis heu realitzat. Per cadascun d'ells, haureu d'explicar el procediment seguit per a la seva resolució: des de les variables i estructures de dades emprades, les tècniques dissenyades, fins a la presentació dels resultats mitjançant mapes i les conclusions que n'extraieu. És a dir, no us quedeu només amb el mapa: argumenteu com heu plantejat i resolt cada problema.

Podeu realitzar els mapes utilitzant qualsevol llenguatge de programació (Python, R, Javascript,...) i llibreria/framework (Leaflet, Folium, Plotly, Mapbox...) que preferiu. Es prohibeix l'ús de qualsevol client GIS, com Google Earth o QGIS, a l'hora de crear les visualitzacions geogràfiques.

Avaluació

Per valorar la pràctica es tindran en compte les següents ponderacions:

- Abast de la solució (quantitat, dificultat, variabilitat i originalitat dels estudis, nombre de ciutats i variables estudiades,...): 20%
- Qualitat de la part 1 (resolució analítica, tècniques utilitzades, ús d'estadístics, gràfiques, extracció de conclusions,...): 40%
- Qualitat de la part 2 (mapes, funcionalitats, interaccions,...): 30%
- Qualitat tècnica de l'informe i comunicació escrita: 10%

Lliurament

El lliurament de la pràctica es farà mitjançant la tasca habilitada al Campus Virtual fins el dia **23 d'octubre de 2.024 a les 23:55**. Haureu d'entregar un fitxer ZIP que contingui:

- L'informe PDF[4] explicant la resolució detallada de la pràctica.
- Els codis/scripts utilitzats.

Entrevista

L'entrevista de la pràctica es farà durant la sessió de laboratori del **24 d'octubre de 2.024**. En aquesta entrevista, el professor farà preguntes als membres del grup per

[4] Recomanem l'ús de LaTeX utilitzant la plataforma Overleaf: `www.overleaf.com`

validar l'autoria de la pràctica i avaluar el nivell de coneixements adquirit pels estudiants. **És obligatori realitzar l'entrevista; en cas contrari, la pràctica no s'avaluarà.**