Nom: Guillem Cortiada Rovira
Link Dataset 1: https://analisi.transparenciacatalunya.cat/Educaci-/Concerts-educatius-Unitat-alumnes-dotaci-de-planti/8spq-9nx7
Link Dataset 2: **https://analisi.transparenciacatalunya.cat/Sector-P-blic/Dades-generals-dels-ens-locals-de-Catalunya/6nei-4b44**

Aquesta activitat, primera part de la pràctica final, consisteix en l**a 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. Bàsicament, la temàtica és lliure, però es valoraran els aspectes següents:

[10%] Justifiqueu breument la vostra selecció, sigui per motius personals o professionals.

Degut a que han sorgit noves tecnologies d'intel·ligència artificial amb un gran potencial, el sistema educatiu hauria de veure's adaptat. Això fa, que els centres educatius siguin una eina important de l'educació i la cultura dels joves. Per això, aquest conjunt de dades pot ser útil per a investigadors que estan interessats en la despesa pública que hi ha actualment en els centres d'educació concertats, i també la quantitat de professors implicats.
En resum, aquest dataset podria proporcionar informació valuosa sobre els centres educatius, la seva dotació de plantilla, els costos associats i el nombre d'alumnes de cada centre.

[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? (10%)

El conjunt de dades sobre centres educatius es rellevant en el context d'entendre els diners que s'han destinat en els últims anys en l'educació concertada, per tal de poder fer estudis posteriors i adaptar els pressupostos i millorar el sistema educatiu públic i concertat, ja que tots els centres eduactius contribueixen en el desenvolupament personal i social dels alumnes.

El dataset, inclou dades desde el curs 2017-2018 fins al curs 2021-2022, ja que pel curs 2022-2023 encara no es tenen dades actualitzades. En concret, tracten un tema important per tal d'observar la despesa pública dels centres educatius concertats i que posteriorment, pot ajudar als investigadors a evaluar i millorar l'oferta educativa en funció de la dotació de plantilla i la gestió de despeses.
En aquest cas, no s'ha tingut en compte la prespectiva de gènere, ja que en dades sobre les despeses de centres educatius conccertats, no es pot diferenciar per gènere.

[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. (25%)

El dataset té una mida de 3479 fileres (milers de registres) i 23 columnes (desenses de variables), on les variables són del tipus categòric i numèric. En concret, les variables són les següents:
- Curs escolar: curs escolar de referència (categòrica).
- Codi_centre: Codi del centre educatiu (categòrica).
- Nom_centre: Nom del centre educatiu (categòrica).
- Municipi: Municipi on s'ubica el centre educatiu (categòrica).
- Titularitat: Titularitat del centre educatiu (categòrica).
- Unitats concertades – EINF: Nombre d'unitats concertades d'educació infantil (quantitativa).
- Unitats concertades – PRI: Nombre d'unitats concertades d'educació primària (quantitativa).

- Unitats concertades – ESO: Nombre d’unitats concertades d’ESO (quantitativa).
- Unitats concertades – EdE: Nombre d’unitats concertades d’educació especial (quantitativa).
- Unitats concertades – BATX: Nombre d’unitats concertades de batxillerat (quantitativa).
- Unitats concertades – CFPM: Nombre d’unitats concertades de cicles formatius de grau mitjà (quantitativa).
- Unitats concertades -CFPS: Nombre d’unitats concertades de cicles formatius de grau superior (quantitativa).
- Alumnes – EINF: Alumnes en educació infantil (quantitativa).
- Alumnes – PRI: Alumnes en educació primària (quantitativa).
- Alumnes – ESO: Alumnes d’ESO (quantitativa).
- Alumnes – EdE: Alumnes en educació especial (quantitativa).
- Alumnes – BATX: Alumnes de batxillerat (quantitativa).
- Alumnes – CFPM: Alumnes de cicles formatius de grau mitjà (quantitativa).
- Alumnes – CFPS: Alumnes de cicles formatius de grau superior (quantitativa).
- Dotació plantilla: Dotació de plantilla (quantitativa).
- Nòmina del personal del centre: Salari del conjunt del personal del centre (quantitativa).
- Seg.Social: Despesa en Seguretat social (quantitativa).
- Despeses de funcionament: Despeses de funcionament (per despeses corrents) (quantitativa).

A més, afegirem un altre dataset (Dades generals dels ens locals de Catalunya), per tal de poder tenir una categoria i subcategoria,i poder respondre preguntes agregades. En concret, es un dataset que a partir de la variable ‘Municipi’ d’aquest dataset es relacionara amb l’altre conjunt de dades (Dades generals dels ens locals de Catalunya), que conté la informació general referent als ens locals de Catalunya per l’àmbit terriotorial i per tant, ens permetrà obtenir despeses dels centres educatius concertats per municipi, comarca i provincia.
Aquest segon conjunt de dades, que esta relacionat amb les dades generals dels ens local de Catalunya, té uns dimensions diferents, on conté 62 columnes i 11800 registres, tot i que només utilitzarem els 3479 registres que coincideixin amb el municipi del primer dataset.
Les variables que utilitzarem del segon dataset seràn les següents:
- Municipi: Municipi de l’adreça postal (categòrica).
- Nom_Capital: Variable amb el nom del municipi original, variable que ens permetrà fusionar datasets (categòrica).
- Comarca: Nom de la comarca (categòrica).
- Provincia: Nom de la província (categòrica).
- Cens: Cens de la població (numèrica).

[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? (25%)

No es un dataset clàssic, i actualment, no hi visualitzacions basades en aquest conjunt de dades.
El dataset principal, no conté dades personals i conté la llicència igual que totes les dades del Govern obert de la Generalitat de Catalunya. Per tant, hem de citar aquestes fonts:

Generalitat de Catalunya. Departament d’Educació. Dades Obertes Catalunya, Concerts educatius: Unitat, alumnes, dotació de plantilla i despesa (Darrere Actualització: 6 de març de 2023).

https://governobert.gencat.cat/ca/dades_obertes/llicencia-oberta-informacio-catalunya/

Pel dataset secundari, hem de dur a terme el mateix:

Generalitat de Catalunya. Departament de Presidència. Dades Obertes Catalunya, Dades generals dels ens local de Catalunya (Darrere Actualització: 11 de maig de 2023).

[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? (30%)

La visualització va dirigida al públic en general o investigadors que vulguin observar l'evolució de la despesa en els centres educatius concertats en els últims anys i no fa falta un context previ per a la visualització de les dades.
En el nostre cas no durem a terme una infografia, ja que no fa falta ampliar la representació visual amb text per entendre la visualització. Llavors, ens decantarem per dur a terme una visualització de dades.

○ **Què vull respondre amb la visualització?**

- Evolució del nombre de professors de cada centre educatiu concertat a Catalunya (per any).
- Evolució dels gastos de cada centre educatiu concertat a Catalunya (per any). I Comarca? I Provincia? (Quines diferències hi ha en els gastos dels centres eductaius concertats segons la provincia?)
- Quina es la comarca amb més centres educatius concertats? I Provincia?
- En quin centre cobren més els professors? I Comarca? I Provincia?
- Quin es el nombre d'alumnes que cursen batxillerat a un centre educatiu concertat per comarca? Comparació amb ESO i Primaria.

○ **El dataset em permet contestar-les amb precisió?**

Si combinem el dataset principal amb el dataset secundari, podrem respondre les preguntes correctament, ja que ens permetra relacionar municipis amb comarques i provincies.

○ **Quines dades concretes del meu conjunt de dades vull explorar?**

En concret es volen explorar les despeses, el nombre de centres educatius concertats, professorat i alumnes d'aquests, i relacionar-ho amb les comarques i provincies on es localitza cada centre educatiu concertat.

○ **Necessito més fonts de dades per donar context al meu conjunt de dades?**

Si, en concret s'ha de complementar el dataset principal amb un secundari per tal de poder relacionar municipis amb comarca i provincia.

Preparació i Exploració de les Dades

Durant el procés de preparació de les dades realitzat mitjançant jupyter-notebook, hem seguit els següents punts:
- Llegir les dades mitjançant la llibrería pandas. S'han llegit el dataset principal I el secundari en format csv.
- S'ha realitzat un resum del dataset principal mitjançant
- S'han revisat els valors perduts:
    - S'han assignat tots els valors perduts numèrics a 0.
    - S'ha eliminat la filera (centre educatiu concertat) que contenia una dotació de plantilla nul·la.

- S'han comprovat si en alguna variable categòrica, s'havien d'homogenitzar les variables. No s'ha trobat cap cas a homogenitzar (com es pot observar en el jupyter-notebook).

- S'han detectat valors extrems en les dades numèriques, duen a terme un boxplot i un histograma per totes les variables numèriques. S'ha pogut observar que hi ha algunes variables que no ens poden aportar molta informació, com son les variables 'Unitats concertades - ', ja que en els boxplots i histogrames veiem com la majoria de valors son 0, cosa que segurament no s'han mesurat/assignat bé i per tant, no podrem tenir valors representatius per aquestes variables. D'altra banda, algunes de les variables de 'Alumnes -' també ens passa el mateix, tot i que pot ser normal degut a que els centre educatius concertat no tingui algún mòdul. Pero per les variables numèriques que volem centrar-nos que son 'Dotació plantilla', 'Nòmina del personal de centre', 'Seg.Social' i 'Despeses de funcionament', observem correctament les variables amb una distribució asimètrica positiva.

- Despres, s'ha realitzat una extracció de característiques, on principalmente hem dut a terme el següent:
    - Eliminar la variable 'Titularitat' del dataset principal, ja que no es rellevant per l'anàlisis que volem dur a terme.
    - Crear una nova variable 'Total Alumnes', on es sumin els alumnes de cada centre educatiu concertat, a partir de les variables de 'Alumnes – Primaria', 'Alumnes – Eso', etc.
    - Finalment, del dataset complementari, només ens hem quedat amb les variables rellevants, que serien les següents: 'MUNICIPI', 'NOM_CAPITAL', 'COMARCA', 'PROVINCIA', 'CENS'.

- Hem combinat els dos dataset (principal i complementari) mitjançant la funció merge de la llibreria pandas, juntant els dos datasets per les columns "Municipi" i "NOM_CAPITAL".

- Per acabar, hem dut a terme un anàlisi de correlació entre les variables a partir d'una matriu i a partir d'un gràfic de dispersió, i hem pogut observar que que les variables més correlaciones són dels despeses de funcionament, la seguretat social, les nòmines del personal del centrei el Total d'alumnes. També, hem pogut observar una correlació entre les 'Unitats concertades - *' amb el número d'alumnes de cada unitat 'Alumnes - *' (p. ex. 'Unitats Concertades – CFPM' i 'Alumnes CFPM').

Per afegir, hem dut a terme uns anàlisis bàsics per tenir un anàlisi previ. En concret, hem obtingut el total d'alumnes per comarca i el total d'alumnes que cursa batxillerat per comarca, pel curs 2021-2022.

Dataset Principal: DespesaCentresEducatiusConcertatsCatalunya.csv
Dataset Secundari: Dades_generals_dels_ens_locals_de_Catalunya.csv

Dataset Fusionat i Preprocessat: DespesaCentresEducatiusConcertatsCatalunya_Preprocessed.csv