Tipologia i cicle de vida de les dades

Pràctica 1

L'objectiu d'aquesta activitat serà la creació d'un dataset a partir de les dades contingudes a un lloc web.

Respostes als exercicis plantejats

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

Hem decidit enfocar la pràctica a extreure les dades dels accidents mortals a Catalunya a partir de les notes de premsa del Servei Català del Trànsit (SCT) que publica cada vegada que es produeix un accident. Aquestes notes de premsa es poden trobar al següent enllaç:

> *https://transit.gencat.cat/ca/el_servei/premsa_i_comunicacio/comunicats_d_accidents_mortals/*

En aquest camp, les notes de premsa són una font fiable i actualitzada de dades perquè es publiquen sempre amb el mateix format, només amb unes hores de retard i sense excepcions després de cada accident. Tot i que el portal es va estrenar el 2011, ens hem centrat en els anys 2014-2022 ja que entre el 2011 i el 2013 el SCT publicava les notes en format de document pdf enllaçat, mentre que a partir de llavors ho fa en format XML.

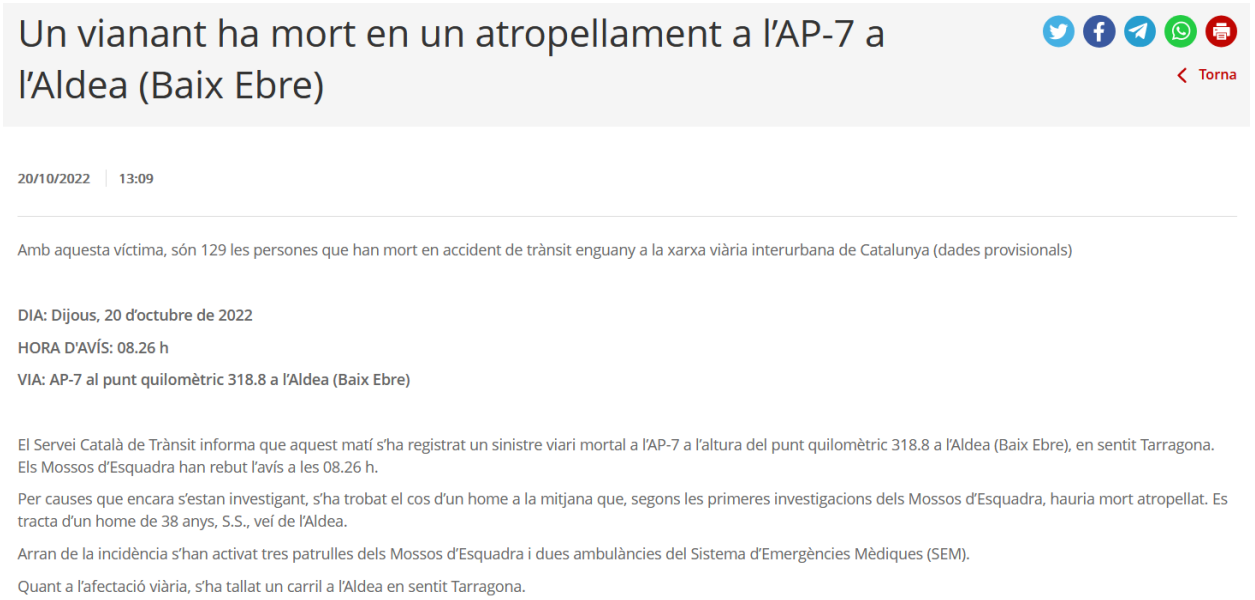
2. **Títol.** Definir un títol que sigui descriptiu pel dataset.

Accidents mortals a Catalunya entre el 2014 i el 2022.

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 s'ha extret de les notes de premsa fetes pel Servei Català del Trànsit des del 2014 al 2022. Aquest conjunt conté variables que permeten registrar i identificar cada accident mortal de trànsit de forma inequívoca.

A continuació es mostra una captura de pantalla d'un exemple de nota de premsa:



Un vianant ha mort en un atropellament a l'AP-7 a l'Aldea (Baix Ebre)

Torna

20/10/2022 | 13:09

Amb aquesta víctima, són 129 les persones que han mort en accident de trànsit enguany a la xarxa viària interurbana de Catalunya (dades provisionals)

DIA: Dijous, 20 d'octubre de 2022

HORA D'AVÍS: 08.26 h

VIA: AP-7 al punt quilomètric 318.8 a l'Aldea (Baix Ebre)

El Servei Català de Trànsit informa que aquest matí s'ha registrat un sinistre viari mortal a l'AP-7 a l'altura del punt quilomètric 318.8 a l'Aldea (Baix Ebre), en sentit Tarragona. Els Mossos d'Esquadra han rebut l'avís a les 08.26 h.

Per causes que encara s'estan investigant, s'ha trobat el cos d'un home a la mitjana que, segons les primeres investigacions dels Mossos d'Esquadra, hauria mort atropellat. Es tracta d'un home de 38 anys, S.S., veí de l'Aldea.

Arran de la incidència s'han activat tres patrulles dels Mossos d'Esquadra i dues ambulàncies del Sistema d'Emergències Mèdiques (SEM).

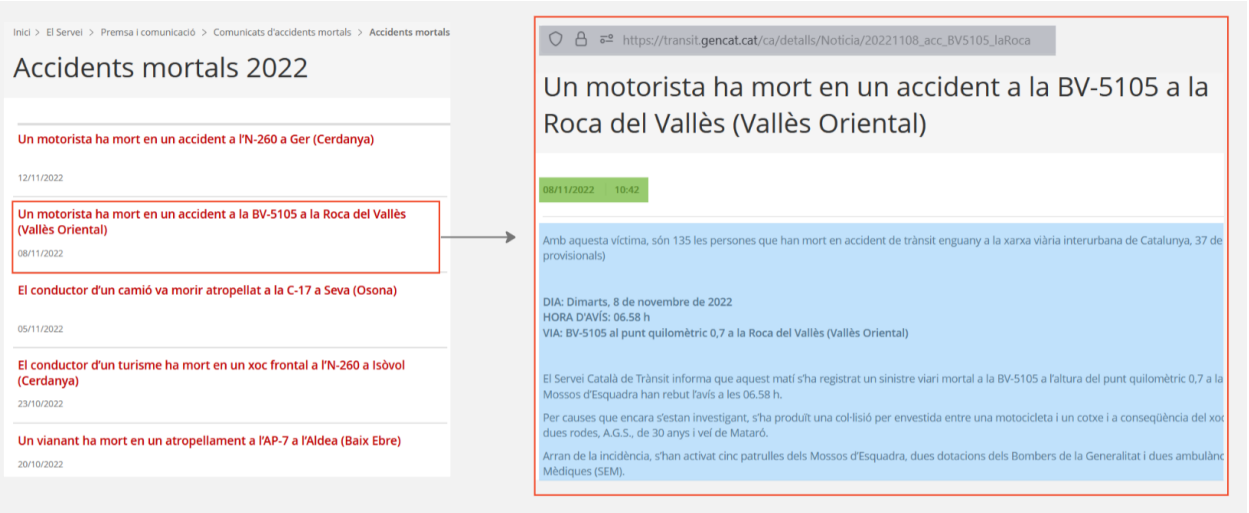
Quant a l'afectació viària, s'ha tallat un carril a l'Aldea en sentit Tarragona.

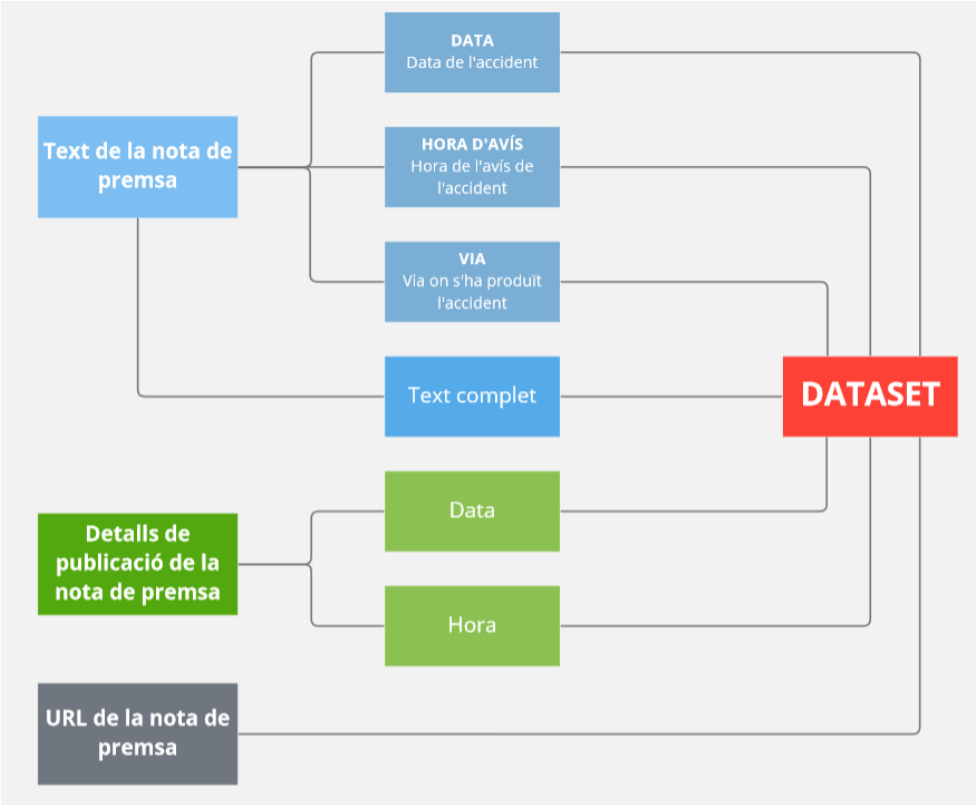
De cada accident es guarda en la mateixa estructura la data, l'hora i la via on ha ocorregut l'accident. Aquestes dades es troben en format de text pla, majoritàriament ressaltades en negreta, però no sempre, i a continuació del títol directament o després d'un subtitol. La selecció d'aquestes dades s'ha fet mitjançant l'ús de seqüències Regex.

Com que l'entrada de la informació es realitza de forma manual per part d'algun treballador del Servei Català del Trànsit, aquestes poden contenir errors d'estructura concrets que no permeten la correcta selecció per part de les seqüències Regex entrades, malgrat que les hem fet el més generals possibles. Per aquesta raó, en cada accident hem inclòs també el text sencer de la nota de premsa i la URL específica de l'accident per facilitar la recuperació de les dades durant la fase de la neteja de dades. A més a més, a partir d'aquestes dades es poden extreure, mitjançant neteja i ús de regex, informacions més específiques sobre cada accident com el tipus d'usuari que ha perdut la vida (ciclista, motorista, conductor de turisme, etc), el nombre de morts en l'accident o la població de la via, per exemple.

Finalment, també guardem el dia i l'hora en què es va comunicar la nota de premsa, els quals es troben en cel·les separades immediatament a sota del títol de la nota de premsa.

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





5. **Contingut.** Explicar els camps que inclou el dataset i el període de temps de les dades.

El dataset està format per un total de 7 columnes i 1169 registres. Les següents columnes o atributs són:

- **Data_accident**: Data quan es va provocar l'accident amb el format *Dia de la setmana, dia de mes de any.* Exemple: ***Dimecres, 31 de desembre de 2014***
- **Hora_avis**: Hora en la que es va notificar de l'accident. Exemple: ***2.4***
- **Via**: Carretera on ha passat l'accident amb el quilòmetre on ha ocorregut així com la població. Exemple: ***C-53, al punt quilomÃ¨tric 127 a la Fuliola (Urgell)***
- **Data_publicacio**: Data quan es va publicar la notícia al portal en format DD/MM/YYYY. Exemple: ***31/12/2014***
- **Hora_publicacio**: Hora quan es va publicar la notícia al portal en format HH:MM. Exemple: ***12:35***
- **Text complet**: Cos de la notícia. Es guarda per si cal recuperar la informació manualment o en cas que s'estimin dades extres a les estructurades.
- **URL**: Link complert a la notícia. Exemple: ***https://transit.gencat.cat/ca/detalls/Noticia/20141231_acc_c53***

6. **Propietari.** 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 del conjunt de dades és la Generalitat de Catalunya, la qual utilitza una llicència CC0 1.0 de Creative Commons per la difusió de les dades. Dins de la Generalitat, les dades són proporcionades pel Servei Català de Trànsit.

Existeixen diversos anàlisis i estructuració del conjunt de dades com ara:

- Dades obertes: *https://analisi.transparenciacatalunya.cat/Transport/Accidents-de-tr-nsit-amb-morts-o-ferits-greus-a-Ca/rmgc-ncpb*.
- Informe periòdic dels accidents de trànsit (2021): *https://govern.cat/govern/docs/2022/01/27/13/29/ecdc0f94-80a1-452d-b8c7-9d0155a4b9b7.pdf*

Malgrat existir certs anàlisis com els presentats, aquests presenten limitacions temporals que són explicades a l'exercici 7.

Per tal de seguir amb els principis ètics i legals exposats a l'assignatura, els dos estudiants han llegit l'avís legal de la Generalitat de Catalunya sobre l'ús d'informació. Aquest especifica que "*les dades es troben obertes per a to el món i sense cap mena de limitació temporal ni restricció en els termes establerts* ***per la la Llicència oberta d'ús d'informació – Catalunya o per l'equivalent instrument legal CC0 de Creative Commons*** *d'acord amb les condicions i règim establert a l'article 17.1 de la Llei 19/2014, de 29 de desembre, de transparència, accés a la informació pública i bon govern i més enllà de les condicions bàsiques establertes en l'article 8 de la Llei 37/2007 sobre la reutilització de la informació del sector públic (citació de la font, no alteració ni desnaturalització de la informació i especificació de la data d'última actualització), i sempre que no es contradigui amb la llicència o avís que pugui tenir una obra i que és la que preval* ".

En aquest projecte, les dades simplement han estat recollides sense cap tipus de modificació i les que es pretenen fer en fases posteriors són fases de neteja per tal de millor la qualitat de la dada, per la qual considerem que no s'està infringint cap legalitat ni principi ètic.

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.

Els accidents de trànsit provoquen cada any centenars de ferits i morts a les ciutats, pobles i carreteres del país. Només el 2021, a Catalunya, hi va haver 179 accidents en els quals van morir 183 persones segons l'últim informe del Servei Català del Trànsit. Sovint, aquests accidents passen en els mateixos punts de la xarxa viària, en els anomenats punts negres. Tot i així, per la forma en com queden registrats els accidents, fa difícil que els treballadors del Servei Català del Trànsit, la ciutadania o periodistes puguin demostrar objectivament la perillositat d'una via amb l'objectiu d'exigir millores sobre la mateixa.

El departament de Trànsit té un registre obert d'aquestes dades al portat de dades obertes de la Generalitat:

> *https://analisi.transparenciacatalunya.cat/Transport/Accidents-de-tr-nsit-amb-morts-o-ferits-greus-a-Ca/rmgc-ncpb*.

El problema d'aquest registre és que s'actualitza manualment periòdicament, sent a dia d'avui l'últim registre del 15/03/2021, és a dir, de fa més d'un any.

Per altra banda, també existeixen registres dels accidents i se'n realitzen informes periòdics, però aquests tenen una periodicitat molt alta i es troben en dades no estructurades de forma que la seva explotació pot ser realment complicada. L'últim informe és de 2021 i compara només la sinistralitat del 2021 vs 2019 vs 2010, es pot trobar al següent enllaç:

> *https://govern.cat/govern/docs/2022/01/27/13/29/ecdc0f94-80a1-452d-b8c7-9d0155a4b9b7.pdf*

Per totes aquestes raons, considerem que a dia d'avui la font més fiable a temps real de les

dades dels accidents mortals a Catalunya són les notes de premsa que fa el Servei Català del Trànsit cada vegada que es produeix un accident. Aquestes notes són una font fiable perquè es publiquen immediatament i sense excepció després de cada accident que hi ha a la xarxa viària catalana, només amb unes hores de retard.

L'objectiu final del conjunt de dades és aconseguir una base de dades que permeti l'anàlisi evolutiu dels morts a Catalunya per descobrir quines vies i punt quilomètric que poden arribar a tenir més accidents, a quines hores es produeixen i quin és l'usuari més vulnerable.

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

La llicència feta servir per la Generalitat de Catalunya és CC0 1.0 la qual entrega el document al domini públic renunciant als drets de propietat intel·lectual. Dona't que l'origen de les dades és d'aquesta llicència i nosaltres simplement hem fet una recopilació a les dades, afegit a que no ens farem responsables de l'ús i de la qualitat d'aquest més enllà del que sigui exigit en el nostre màster, creiem ètic i responsable fer servir la mateixa llicència.

9. **Codi.** Codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

Per poder realitzar web scraping de forma correcte, s'ha hagut de navegar a diferents nivells de la pàgina web.

1. El nivell superior és la pàgina que presenta links a la col·lecció d'accidents per cada any (link). La forma en que es codifica la següent pàgina canvia segons l'any. Fins el 2017 al link principal s'hi afegia *accidents-mortals-* + *str(year)* ; mentres que del 2018 fins a dia d'avui s'utilitza *accidents_mortals_* + *str(year).* Un cop es té en compte aquesta particularitat es pot entrar al següent nivell.

2. El següent nivell és la col·lecció de notes de premsa per l'any seleccionat. Cada any hi ha més de 100 accidents mortals, fet pel qual, no tots els accidents apareixen dins de la mateixa URL, si no un màxim de 15 resultats. El següent accident, es guarda en una altra pàgina, per la qual s'hi pot navegar fent click al botó que hi diu *Següent.* El nostre codi és capaç de captar aquest botó a cada iteració i a fer web scraping al contingut que té la URL d'aquest.

Aquest botó anomenat *Següent* presenta un problema i és que si s'arriba a l'última pàgina segueix generant URLs però que són còpies de l'última pàgina d'aquell mateix any. Per això mateix, agafem el botó anomenat *Last* i que permet anar a l'última pàgina. A cada iteració comprovem si l'URL de la pàgina en la que ens trobem és igual a la del botó *Last.* Si és el cas, aquesta és l'última iteració que realitza sobre l'any explorat.

3. I l'últim nivell de navegació és la informació de cada notícia. A cada pàgina de l'any de hi ha entre 1 i 15 noticies que corresponen a cada accident i a les quals s'hi entra per agafar la informació final del dataset. En aquest punt es realitzen les accions detallades a l'apartat 3 i 4.

10. **Dataset.** Publicar el dataset obtingut en format CSV a Zenodo, incloent-hi una breu descripció. Obtenir i adjuntar l'enllaç del DOI del dataset (https://doi.org/…). El dataset també haurà d'incloure's a la carpeta **/dataset** del repositori.

El dataset obtingut d'aquesta pràctica es pot trobar en el següent enllaç:

https://doi.org/10.5281/zenodo.7316989

11. **Vídeo**. Realitzar un breu vídeo explicatiu de la pràctica **(màxim 10 minuts)**, que haurà de comptar amb la participació dels dos integrants del grup. Al vídeo s'haurà de realitzar una presentació del projecte, destacant els punts més rellevants, tant de les respostes als apartats com del codi utilitzat per a extreure les dades. Indicar l'enllaç del vídeo (https://drive.google.com/…), que haurà d'estar al Google Drive de la UOC.

El video es pot trobar en el següent enllaç:

https://drive.google.com/file/d/1lHt0-xhv4BM76hn4VZWuxVWjPl08oeda/view?usp=share_link

Contribució de cada membre en cada apartat

Contribucions	Signatura
Investigació prèvia	CGA, SCN
Redacció de les respostes	CGA, SCN
Desenvolupament del codi	CGA, SCN
Participació al vídeo	CGA, SCN