Visualització de dades

PAC 2: Projecte de visualització de dades (A5)

Jordi Boldú Millà

Títol de la visualització presentada.

Acords de pau signats durant el període 1990 - 2019

Descripció

En aquest document es realitza un breu anàlisi de l'evolució dels diferents acords de pau (*Agreement*) que s'han signat arreu del món durant el període 1990 – 2019, tenint en compte l'àmbit del conflicte (entre estats, local, ...) i des d'una perspectiva **territorial** i **temporal.**

Per dur a terme aquest anàlisi sobre el conjunt de dades de mostra (*dataset*) i obtenir un major nombre d'*insights*, s'han emprat les següents tècniques:

- Càlcul de mètriques, indicadors estàndards (mitjana, desviació estendard, etc.) i distribució dels diferents atributs del dataset
- Elaboració de representacions visuals amb finalitats exploratòries d'on extreure coneixement

L'anàlisi s'acompanya d'un enllaç web a un *dashboard* públic fet amb l'eina Tableau per consultar les de manera interactiva els indicadors i aspectes que he considerat més rellevants

Nota: tots els scripts i processos intermedis que s'han fet servir per dur a terme aquest anàlisi s'adjunten amb el lliurament d'aquest informe.

Origen de les dades i criteri de cerca aplicat

Lloc web	*Peace Agreements Database*
Enllaç	https://www.peaceagreements.org/
Descripció	Lloc web que conté una base de dades de documents sobre acords de pau a tot el món, des del 1990 fins a avui, creada i mantinguda per la Law School de la University of Edinburgh. Per a més informació, consultar l'enllaç https://www.peaceagreements.org/about
Llicència	Creative Commons Attribution- NonCommercialShareAlike 4.0 International Licence (CC BY-NC-SA 4.0)
Formats	CSV/Excel/PDF
Cita autor(s)	Bell, Christine, Sanja Badanjak, Juline Beujouan, Robert Forster, Tim Epple, Astrid Jamar, Kevin McNicholl, Sean Molloy, Kathryn Nash, Jan Pospisil, Robert Wilson, Laura Wise (2020). *PA-X Codebook, Version 3*. Political Settlements Research Programme, University of Edinburgh, Edinburgh. www.peaceagreements.org Bell, C. and Badanjak, S. (2019) 'Introducing PA-X: A new peace agreement database and dataset', *Journal of Peace Research*, 56 (3).

Donat el desconeixement previ sobre la temàtica tractada i veient que el nombre total d'elements existents suposa un volum assequible i manejable de dades, he decidit **no aplicar cap criteri de cerca** i descarregar **tot el contingut** de la base de dades que fa referència a la informació dels acords de pau i llurs metadades **sense incloure'n** el corpus.

El format escollit per a la descàrrega ha estat l'*Excel*

Descripció de les dades

Format i estructura de de les dades

Les metadades dels acord de pau consten de ***265 atributs*** que es poden diferenciar en dos grups:

- Les metadades **bàsiques** que representen el conjunt mínim d'informació que permet descriure un acord de pau (*entitats o països en conflicte, àmbit territorial, tipus de conflicte, procés de pau dins el què s'inclou el tractat, data de signatura ...*). Aquest grup està format per **25 atributs**.
- Metadades per indicar **si es fa menció** de determinats tòpics (*col·lectius, grups, categories o aspectes socials, econòmics, religiosos, ...*) i **quin tipus de menció** se'n fa. Aquest grup està format per **240 atributs.**

Per a obtenir informació més detallada, consultar el manual PEACE AGREEMENTS DATABASE AND DATASET V3 - codebook

Perfilat de les dades de la mostra treballada

Per dur a terme l'anàlisi de les dades del fitxer que he descarregat (*pax_all_agreements_data.xlsx*) he fet un servir el llenguatge de programació *Python* sobre *Jupyter Notebook* i **després d'analitzar les dades i fer els canvis oportuns**, he utilitzat la llibreria Pandas Profiling per a generar un informe.

jupyter PAC2_jboldum Last Checkpoint: hace 24 minutos (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3

Run Code

```
# Informació general sobre la mostra de dades
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1832 entries, 0 to 1831
Columns: 265 entries, Con to ImSrc
dtypes: float64(5), int64(244), object(16)
memory usage: 3.7+ MB
```

In [2]:

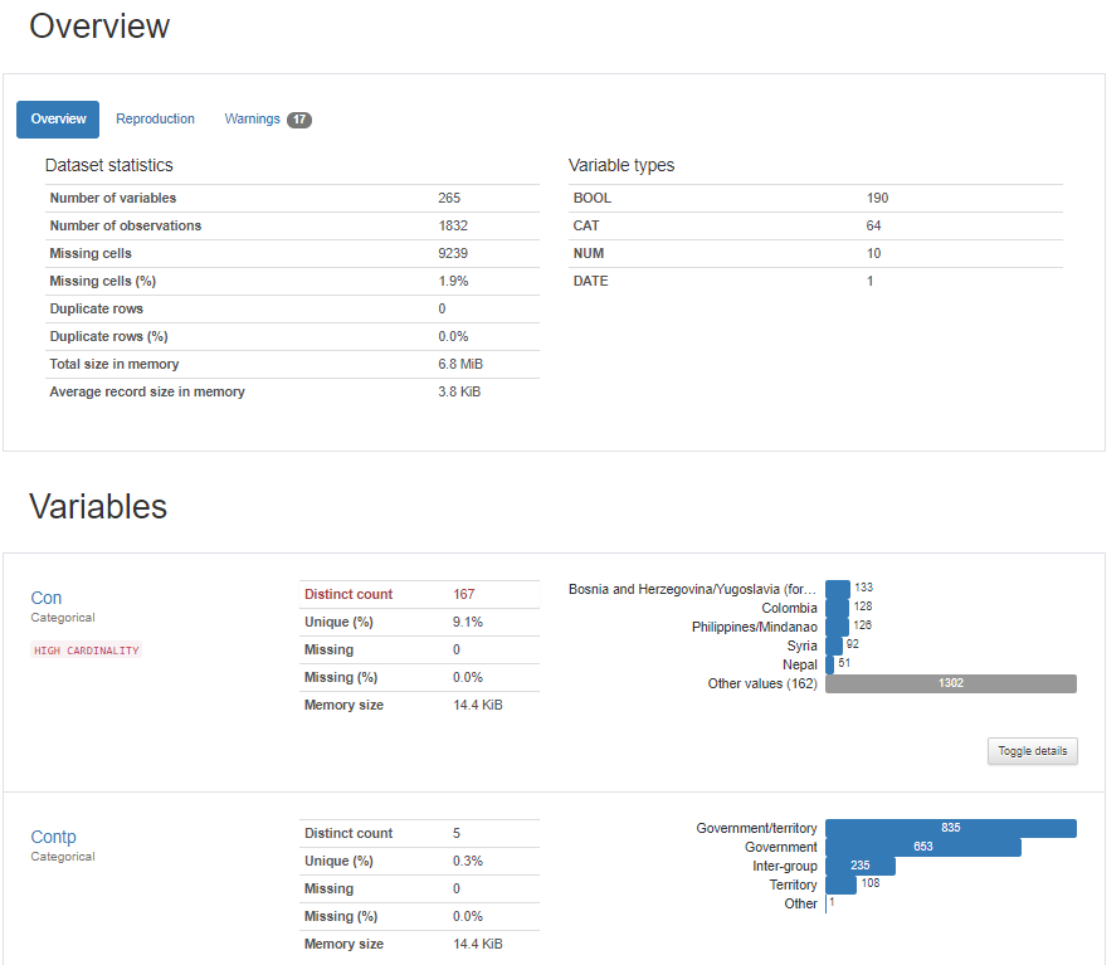
```
# Establim un valor fix (seed) per poder reproduir l'aleatorietat
seed = 20200413

# Exemple de 10 elements de la mostra
df.sample(10, random_state=seed)
```

Out[2]:

	Con	Contp	PP	PPName	Reg	AgtId	Agt	Dat	Status	Lgt	N_characters	Agtp	Stage
1152	Pakistan/Taliban	Government/territory	86	Pakistan-Taliban process	Cross-regional	1531	Shakai Peace Agreement	2004-07-05	Multiparty signed/agreed	1	1961	Intra	SubPar
124	Bosnia and Herzegovina/Yugoslavia (former)	Government/territory	125	Bosnia peace process	Europe and Eurasia	1179	Action Plan of the European Union for the Form...	1994-02-28	Unilateral document	5	11903	InterIntra	Pre
657	Georgia/Russia/Ossetia	Government/territory	44	South Ossetia peace process	Europe and Eurasia	1031	Memorandum on Measures of Providing Safety and...	1996-05-16	Multiparty signed/agreed	2	4395	Intra	Pre
1594	Sri Lanka	Government/territory	114	Sri Lanka LTTE 2002 onward process	Asia and Pacific	1161	Accelerated Action on Resettlement and Humanit...	2003-01-09	Multiparty signed/agreed	2	7207	Intra	Pre
							Joint Statement -						

Captura de pantalla del codi *Python* en *Jupyter Notebook*[1]

[1] El codi font del Jupyter Notebook (*PAC2_jboldum.ipynb*) i la seva versió en format es poden trobar dins la carpeta *Altres_fitxers* d'aquest lliurament



Captura de pantalla de l'informe generat amb la llibreria *Pandas Profiling*[2]

Informació sobre les dades explorades

- El tamany del *dataset* és de 1832 files x 265 columnes
- Hi ha un total de 9239 cel·les amb valors *missing* (l'1,9% sobre el total)
- Tipus de variables detectades
    - 10 variables de tipus *numèric*
    - 64 de tipus *categòric*
    - 190 de *tipus booleà/binari* (valors entre 0 i 1)
    - 1 de tipus *data* (*variable **Dat** que s'ha convertit a data durant l'exploració de les dades amb Jupyter Notebook*)

Observacions sobre les dades explorades

Sobre les variables **categòriques** (64) podem distingir 2 tipologies:

- Aquelles que contenen un gran nombre de literals diferents i **que en dificulten el seu anàlisi.**
    - En el cas de la variable ***Con,*** per exemple, que representa els països/territoris/entitats involucrades en l'acord de pau, és una cadena de text amb els diferents **literals concatenats enlloc de ser variables o registres diferents**, un per entitat. A més, val a dir que conceptualment, s'hi **barregen entitats que no sempre són comparables** (països o territoris amb entitats com Les Nacions Unides o faccions polítiques). A més, mateixos conceptes **apareixen escrits de maneres diferents** (errors orto tipogràfics?)

[2] L'informe generat amb la llibreria Pandas Profile (*informe_dades_PAC2.html*) es pot trobar dins la carpeta *Altres_fitxers* d'aquest lliurament

- Les variables que fan referència al nom dels acords (**Agt**) o processos de pau (**PPName**) contenen text lliure difícilment tractable
- Algunes variables (**Part, ThrdPart, OthAgr**) contenen text lliure difícilment tractable

- Aquelles que contenen un nombre breu de literals que sempre apareixen escrits igual i que realment representen categories **(Contp, Status, Agtp, Stage, StageSub, Loc1Iso, Loc2Iso)** a més de totes aquelles variables que representen tipus de referències o mencions de determinats grups dins de l'acord de pau, sempre que aquestes mencions no siguin del tipus binari (no es menciona = 0 / es menciona = 1)

Sobre les variables **numèriques** (10)**,** podem distingir 2 tipologies:

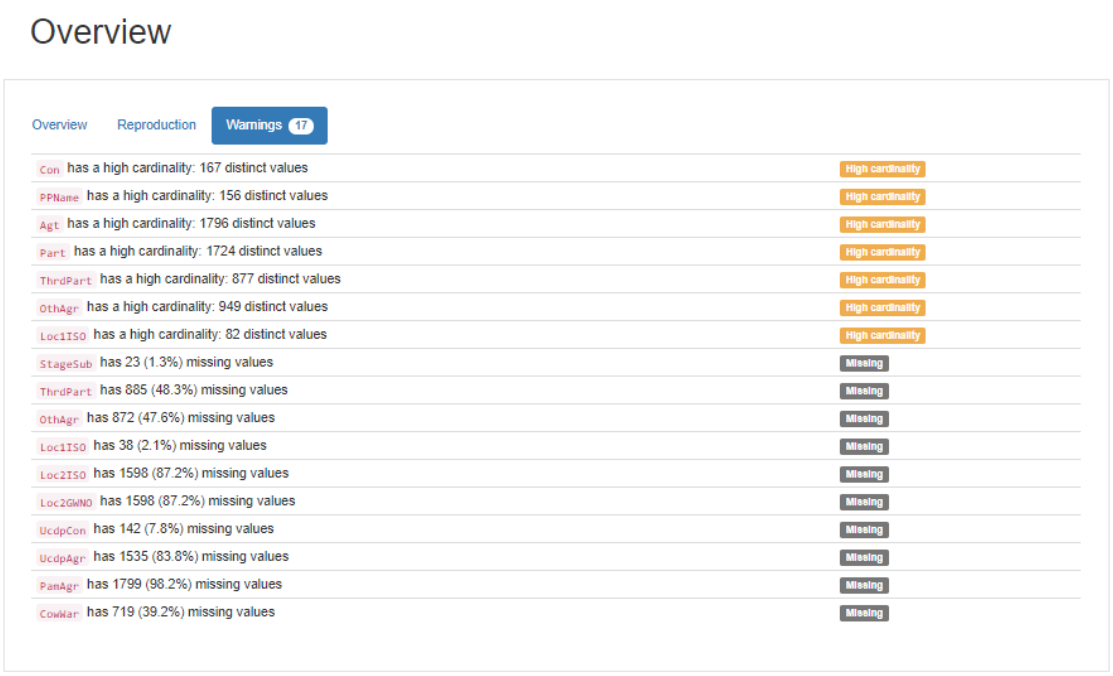
- Aquelles que representen codis identificadors interns o externs **(PP, AgtId, UcdpCon, UcdpAgr, PamAgr, CowWar, Loc1GWNO, Loc2GWNO, ...)**
- Aquelles que representen la dimensió del valor d'una variable (**Lgt, N_characters**, ...)

Sobre les variables **booleanes o binàries** (190) només comentar que prenen valors 0 o 1 en funció de si el grup/concepte que representen apareix o no mencionat dins de l'acord de pau. **En són la major part**

Només hi ha 1 variable de tipus **data** i representa la data de signatura de l'acord de pau i que ja he transformat en el *dataset* original

Altres aspectes a considerar

- Alt nombre d'elements diferents dins de les variables categòriques pels motius abans exposats (veure imatge adjunta)
- Distribució no uniforme de valors *missing* dins del *dataset* (veure imatge adjunta)



Captura de pantalla de l'informe generat amb la llibreria *Pandas Profiling*[3]

[3] L'informe generat amb la llibreria Pandas Profile (*informe_dades_PAC2.html*) es pot trobar dins la carpeta *Altres_fitxers* d'aquest lliurament

- Existència **d'*outliers*** a:
    - Les variables **Lgt** i **N_characters** que fan referència, respectivament, a l'extensió de pàgines i nombre de caràcters dels acords de pau (veure detall informe generat)
    - 2 valors de la variable **UcdpAgr** (id extern) que són números de 13 dígits i que semblen un error tipogràfic en la introducció de les dades

Proposta de millora arran de les troballes realitzades

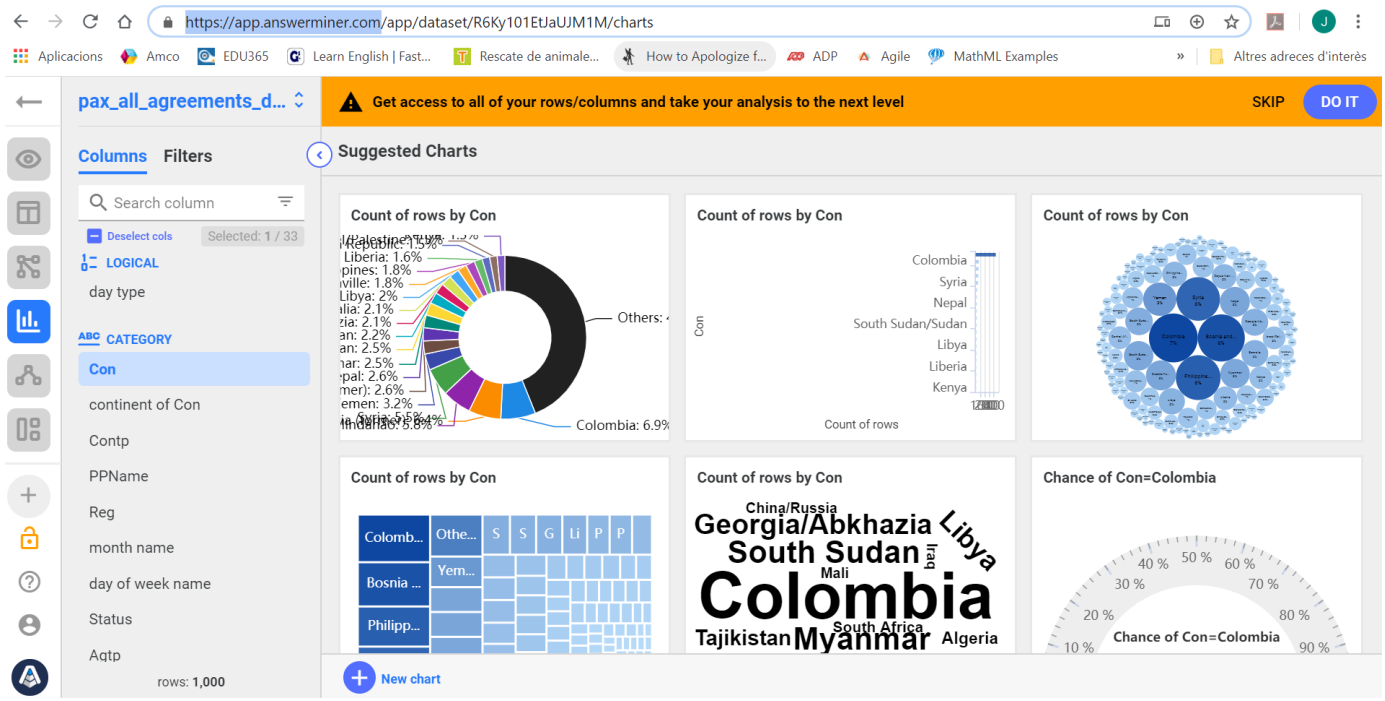
- Revisar i uniformitzar les diferents variacions que apareixen en les variables categòriques per una millor explotació de la informació
- Explorar nous formats de dades que permetin representar millor la naturalesa de la informació i la cardinalitat de la relació entre el seus actors (1 a N, N:M, 1 a 1, ...), com per exemple, **separar i pivotar** tots els possibles valors de la variable **Con** (*Aquesta prova l'he fet*

Eines emprades

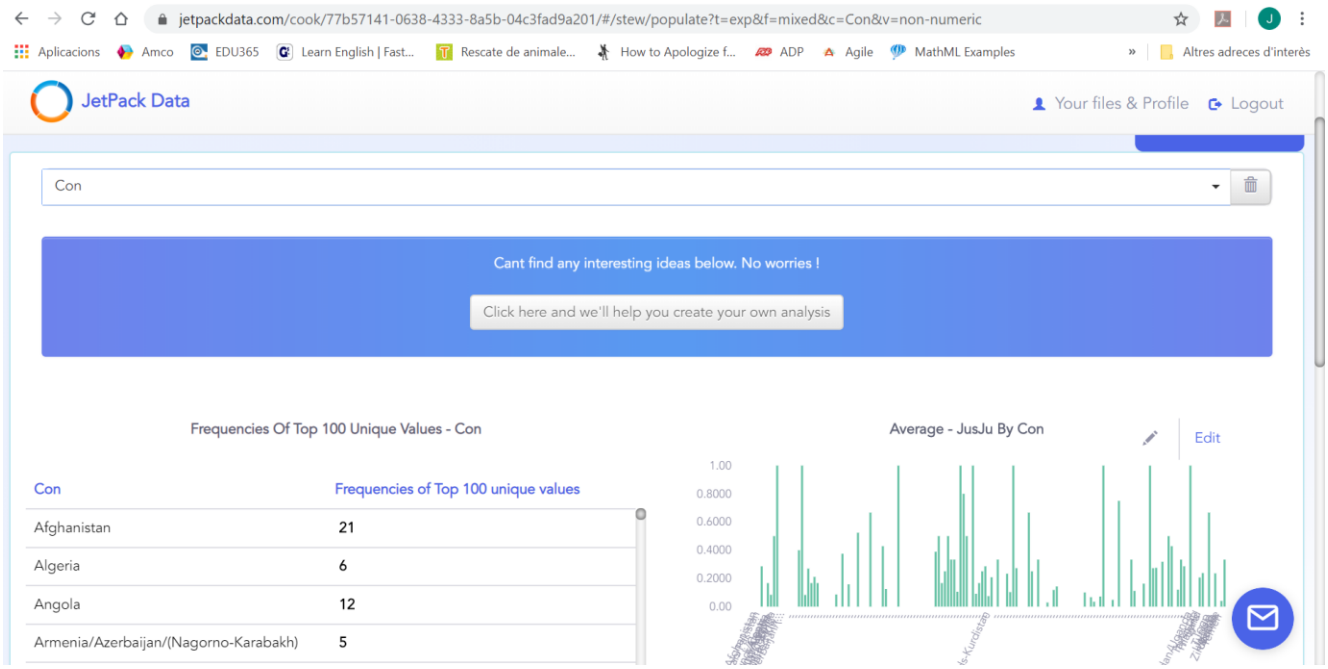
A més de fer servir *Python* i la llibreria *Pandas Profile* per a generar aquest informe, també he aprofitat per explorar diferents eines existents, gratuïtes o de pagament però amb un període de temps d'ús gratuït per fer-me una idea del què hi ha al mercat, més enllà de les ja conegudes **Tableau, Power BI, Qlikview** o el mateix **Excel.**

Apart de les referenciades a la pràctica, he "descobert" i provat les següents eines de les que adjunto captures de pantalla sobre el *dataset* de dades descarregat

- https://www.answerminer.com/



- https://www.jetpackdata.com/



Troballes significatives i curiositats

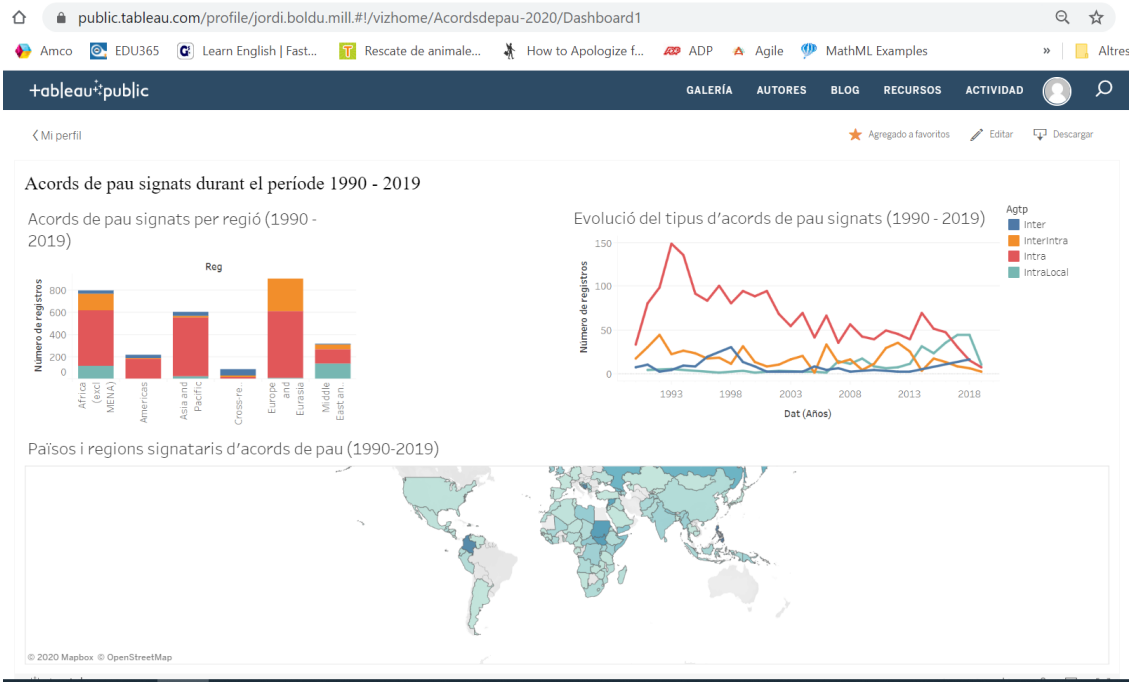
Després d'analitzar les dades amb diferents eines, puc fer una breu descripció de les troballes més significatives

- Els **processos de pau** sota els que s'han signat més *acords de pau*, amb diferència, són el **Philippines - Mindanao process** i el **Bosnia peace process**
- La majoria de conflictes han tingut lloc a les zones d'**Àfrica (excloent MENA)** i la zona d'**Europa i Euràsia**
- El **75%** dels acords de pau **estan redactats en 5 pàgines o menys**
- Prop del 70% dels acords de pau fan referència a conflictes interns (Intra)
- El **Consell de Seguretat de les Nacions Unides** és l'entitat que **ha signat més acords de pau**
- El **Royal Government of Norway** és el **segon organisme que ha participat més vegades com a observador** (Third Party) en acords de pau
- La majoria d'acords de pau reflecteixen conflictes que s'han donat als països **Filipines, Bòsnia Herzegovina, Colòmbia, Síria i el Sudan**
- Pràcticament, la **totalitat** dels acords de pau **no fa cap referència/menció** al col·lectiu LGTBI

Visualització sobre Tableau

La visualització es pot visualitzar en el següent enllaç:

https://public.tableau.com/profile/jordi.boldu.mill.#!/vizhome/Acordsdepau-2020/Dashboard1?publish=yes



Bibliografia

- Bell, Christine, Sanja Badanjak, Juline Beujouan, Robert Forster, Tim Epple, Astrid Jamar, Kevin McNicholl, Sean Molloy, Kathryn Nash, Jan Pospisil, Robert Wilson, Laura Wise (2020). *PA-X Codebook*, *Version 3*. Political Settlements Research Programme, University of Edinburgh, Edinburgh. www.peaceagreements.org

- Bell, C. and Badanjak, S. (2019) 'Introducing PA-X: A new peace agreement database and dataset', *Journal of Peace Research*, 56 (3).

- Material de l'assignatura *Visualització de dades*. UOC