Part I: Selecció del conjunt de dades

Autor: Josep Rau

Assignatura: M2.989 - Arquitectures de bases de dades no relacionals - Aula 1

Índex

Descripció del conjunt de dades

El conjunt de dades escollit per a la pràctica final és "**Canadian Anti-Fraud Centre Reporting Data**". Aquest conjunt de dades proporciona informació sobre els atacs cibernètics denuncias desde 2021 fins a 2025 a Canadà.

Nom	Canadian Anti-Fraud Centre Reporting Data
Files	328650
Columnes	21
Format	Csv
Font	Government of Canada
URL	https://open.canada.ca/data/en/dataset/6a09c998-cddb-4a22-beff-4dca67ab892f/resource/43c67af5-e598-4a9b-a484-fe1cb5d775b5
Idioma	Anglès i Francès
Tamany	70.8MB
Última actualització	01/04/2025

Justificació

M' he decantat per aquest conjunt de dades perquè tracta d'un tema actual i que afecta a tot el món, en aquest cas els atacs cibernètics. A tots ens ha arribat algun cop el missatge de la DGT reclamant una multa no pagada i exigint el seu pagament mitjançant un link, mentre que una gran majoria de persones pot identificar aquest missatge com a fraudulent, una altra part de la població desconeix les seves intencions i per tant són més vulnerables. Tot i no haver viscut de primera mà un scam, sempre se sent algun cas per les notícies o en el boca a boca. És per aquest motiu que m' interessa analitzar com han evolucionat durant els últims anys, quin origen tenen, i com es distribueixen en quant edats i gèneres, en aquest cas en el país de Canadà.

Rellevància

Amb l'auge de les tecnologies i la digitalització de la societat, qualsevol persona té un dispositiu amb el qual realitzar gestions, trucades, fer pagaments, etc. El fet de disposar d' un dispositiu (mòbil, ordinador, tablet, etc.) ja incrementa substancialment les probabilitats de ser víctima d' atac. I si a més es realitzen altres gestions incrementen encara més les probabilitats de clicar en links que no toquen, o comprar en llocs que no existeixen, o donar la contrasenya del banc, etc. És un fet que pot passar a qualsevol persona si no s'està atent i no hi ha un conscienciació prèvia. El conjunt de dades escollit ens proporciona dades suficients per analitzar aquests atacs a Canadà fins l' any 2025. Per tant, és un conjunt de dades actualitzat fa un mes i sobre un tema actual i rellevant per a les societats.

Complexitat

És un conjunt de dades que compte amb 328650 registres i 21 columnes, les que combinen atributs quantitatius i categòrics. A continuació es detalla la composició del conjunt de dades:

Atribut	Tipus	Descripció	Exemple
Numero d'identification / Number ID	Numèric identificatiu	Número identificador del cas	1
Date Received / Date recue	Dates	Data en la que es reb la denúncia	02/01/2021
Complaint Received Type	Categòric	Tipus de denúncia	CAFC Website
Type de plainte recue	Categòric	Tipus de denúncia en francès	CAFC site web
Country	Categòric	País on es reporta el frau	Canada
Pays	Categòric	País on es reporta el frau en francès	Canada
Province/State	Categòric	Estat on es reporta el frau	Saskatchewan
Province/Etat	Categòric	Estat on es reporta el frau en francès	Saskatchewan
Fraud and Cybercrime Thematic Categories	Categòric	Tipus d'atac	Merchandise
Categories thematiques sur la fraude et la cybercriminalite	Categòric	Tipus d'atac en francès	Marchandise
Solicitation Method	Categòric	Mitjà utilitzat per inciar el frau	Other/unknown
Methode de sollicitation	Categòric	Mitjà utilitzat per inciar el frau en francès	Autre/inconnu
Gender	Categòric	Génere de la víctima	Not Available
Genre	Categòric	Génere de la víctima en francès	non disponible
Language of Correspondence	Categòric	Idioma utilitzat en la comunicació.	Not Available
Langue de correspondance	Categòric	Idioma utilitzat en la comunicació en francès	non disponible
Victim Age Range / Tranche d'age des victimes	Categòric	Rang d'edat de la víctima	'Not Available / non disponible
Complaint Type	Categòric	Tipus d' estafa.	Attempt
Type de plainte	Categòric	Tipus d' estafa en francès	Tentative
Number of Victims / Nombre de victimes	Quantitatiu	Nombre de víctimes	0
Dollar Loss /pertes financieres	Quantitatiu	Cost econòmic del frau	$0.00

És un conjunt de dades senzill, però que aporta informació necessària per poder fer un anàlisi profund. Per una banda tenim les mètriques a estudiar, en aquest cas el nombre de víctimes i les pèrdues econòmiques associades al frau. I per altra banda tenim les dimensions o perspectives per les quals analitzar les mètriques, com poden ser País, Estat, tipus d'atac, mitjà utilitzat, etc.

Originalitat

És un conjunt de dades extret de la pàgina web del govern de Canadà i no he detectat cap pàgina popular com Kaggle o Tableau amb un report de les dades. No obstant, sobre temes similars com fraus bancàris, o atacs cibernètics a nivell d'empreses sí que apareixen en aquestes pàgines. És per això que el considero un dataset força original per al projecte.

Aquest conjunt de dades es podria enriquir creuant-lo amb un dataset que contingués dades demogràfiques de cada estat per tal de poder estudiar variables com fraus per habitant. Tot i això és un dataset senzill però força complet que ens permetrà analitzar les mètriques de cost econòmic i nombre de víctimes desde diferents perspectives.

Qüestions

Amb aquest conjunt de dades em plantejo respondre les preguntes següents:

- Quina és la tendència dels fraus en els últims anys ?
- Són fraus nacionals o hi ha alguna tendència geogràfica?
- Quin és el nombre mig d'afectats?
- Quin és el cost econòmic mitjà segons l'atac?
- Quin és el mitjà més comú segons el tipus d'atac?
- Analitzar com es distribueixen les víctimes segons gènere i edat.

La meva idea a dia d' avui és presentar tota aquesta informació en un format que permeti a l'usuari interactuar i extreure'n les seves pròpies conclusions. He mostrat algunes de les preguntes que es podran resoldre amb la visualització, però al fer-ho interactiu, aquestes dependran també de les intencions i creativitat de l'usuari.