Declaració de treball original (no plagi) de l'estudiant

Jo, Xavi Rubio, declaro que per realitzar aquest lliurament m'he inspirat, he mirat, i m'he informat però no he copiat ni plagiat ningún contingut ni un document d'un company ja sigui actual o passat.

Introducció a la ciència de dades

PAC1: Ciència a les dades?

Pregunta 1 (20% puntuació)

1- Com podem observar en aquesta notícia ens parla de la temperatura, compara les temperatures que hem tingut en relació amb la temperatura mitjana històrica. Aquestes dades són dades compostes, ja que encara que siguin simples (la temperatura expressada en graus Celsius) van amb relació a una altra dada, la ciutat, la temperatura dóna molta informació combinada, en aquest cas la combinen amb una ciutat per així saber la temperatura en aquella ciutat en concret. Ens diu que les dades han sigut recopilades per l'Agència Estatal de Meteorología (AEMET), així que podem dir que sí que són fiables.

2- N'hi ha dades de ciutats (els noms), de temperatures (graus Celsius) i anys, mesos i dies.

3- Els comentaris encara que no tenen cap estructura, ja que cada persona pot comentar i no segueix cap ordre, els comentaris serien dades estructurades, ja que cada comentari l'ha escrit una persona, i cada comentari està compost d'un usuari, de la data de publicació, del text de la persona, en aquest cas també n'hi ha *likes* i de les respostes si n'hi ha, també poden tenir moltes més dades que de primera mà no veiem però estan presents.

4- Aquest article vol respondre a **Quanta calor fa a la les províncies d'Espanya?**, **Fa més calor de lo normal?**, **Quina província registra la temperatura mes alta i quina es la temperatura mes alta de les províncies?**, i **N'hi ha molta diferencia respecte a les temperatures d'anys anteriors?**, aquestes són les preguntes que intenta respondre. L'article utilitza molts mètodes de visualització: mapes, gràfics, taules i gràfics de línies i barres. Amb tantes formes de visualització podem dir que sí que estan aprofitant les possibilitats de visualització principals, i, ho fan bé perquè són fàcil d'entendre-les.

Pregunta 2 (30% puntuació)

1- Com podem apreciar al mapa dels moviments sísmics, tenim punts blaus els quals ens indiquen on ha estat el moviment sísmic en concret. Primer de tot volem saber els moviments sísmics del volca que van començar el 19 de setembre de 2021, ja que ens agradaria analitzar el problema actual, ja que aquesta base de dades té registres des de 2017. Respondre a si les zones on hi ha registres estaven habitades o no. Veiem en el següent mapa i ho comparem amb l'anterior veiem que els moviments sísmics han impactat en unes zones amb molta població.

2- En primer lloc, per saber la informació dels moviments sísmics utilitzaria les dades proporcionades de l'exercici, també necessitaria dades de densitat de població per zona, també podria extreure les edats per zona afectada i així poder saber com podria afectar a la població d'aquella zona / illa, i, per trobar tot això aniria a intentar buscar una base de dades fiables, és a dir, d'entitats oficials ja sigui el govern d'un país, la NASA... En cas de no trobar res buscaria de diferents fonts (a ser possibles oficials) i comparar-les entre elles per veure la semblança entre elles.

3- Segons la meva experiència com a desenvolupador d'aplicacions, emmagatzemaria les dades en bases de dades relacionals, ja que podem fer relacions entre les dades, com per exemple si tenim la coordenada d'un moviment sísmic ho podem relacionar amb la zona de coordenades del poble on està situat. I per poder fer servir les dades les hem de processar, netejar d'impureses (*manual curation*) i donar-li el format que nosaltres utilitzarem com per exemple la data que a tots llocs estigui com a dd/mm/yyyy, no hi hagi dades en 'nul', etç.

4- Tenint en compte que principalment tenim dues dades principals (moviments sísmics i densitat de població per zona) creem una relació entre elles, aquesta relació és el lloc. En la base de dades proporcionada n'hi ha 639 registres, dels quals tenim tota la informació, i per l'altra la densitat de població per zona. Sabent això les organitzaria per coordenades i ensenyaria les magnituds de cada moviment sísmic diferenciant-les amb colors, vermell la magnitud més alta i groc per la més baixa així si superposem i ajuntem els dos mapes podríem veure on han estat els moviments sísmics més potents.

5- Ara que sé la magnitud de cada moviment sísmic i la densitat de població per zona tot relacionat, utilitzaria dos tipus de dades diferents, una amb les coordenades exactes dels moviments sísmics amb els reus respectius colors o un mapa de calor que seria tan fàcil com fer la mitjana de les magnituds per zona. Faria servir una pàgina web interactiva per una API de manera que les dades es puguin preservar i disseminar. Aquesta informació seria només informativa i estaria destinada a gent majoritàriament de la península, ja que els de l'illa tenen aquesta informació de primera mà.

Pregunta 3 (30% puntuació)

- En aquesta redacció m'agradaria ometre els exemples posposats i capficar-me en una altra empresa la qual he tingut experiència treballant amb ells i ara que estic entrant en

el món del Big data m'agradaria analitzar. Aquesta empresa es Amazon, una de les grans tecnològiques al món actual. Amazon és coneguda sobretot per al seu e-Commerce encara que dona més serveis, per això analitzaré la botiga d'Amazon, ja que tenen moltes més dades i ho fan d'una forma curiosa comparades amb la resta d'empreses. Amazon és l'empresa més gran en e-Commerce, tothom ha sentit a parlar-hi i inclús l'hem utilitzat almenys un cop, i ven una enorme quantitat de productes els quals té un sistema de recomanació ajustat a cada persona on en aquest cas cada compte; i això és el que m'agradaria analitzar, com ho fan, l'impacte que té sobre l'empresa i tots els beneficis que li pot generar només amb simplement una idea que a primera vista es veu fàcil. Amazon és líder en l'ús exhaustiu del CFE (Collaborative Filtering Engine). Aquest és un motor el qual fa ús de filtres per a poder recomanar de forma quasi precisa la següent cerca que l'usuari farà. Aquest motor es basa en un nou concepte que no coneixia que és "Behavioural Analytics", en poques paraules aquesta manera d'analitzar no només et diu coses com: d'on s'ha connectat l'usuari, quant de temps hi ha estat connectat... Si no que és una anàlisi de comportament, és a dir, analitza com l'usuari es comporta en la pàgina, un exemple molt senzill, t'agraden els ordinadors, ets un entusiasta d'ells i en compte de comprar-te un el vols fer tu mateix a peces, imagina't que ets una persona jove i no tens suficients diners per comprar tot l'ordinador així que et compres la memòria RAM i la CPU perquè estaven en oferta, aquí és on entra l'anàlisi de comportament, al següent cop que entris veuràs que et recomana altres components, i inclús ja no et recomanarà ni RAM, ni CPU, ara les seves recomanacions seran la gran majoria de motherboards, ja que és on es connecten els components que vas comprar i segons als components que entris a mirar, perquè sí, no només és comprant si entres a mirar també analitza el que estàs fent, i així es pot veure com Amazon ja sap que el que vols és fer-te un ordinador i no només comparar components, ja que pot ser que ja no funcionin o vulguis actualitzar-los. Aquesta manera d'analitzar les dades és el que ha fet que Amazon sigui la gegant tecnològica que és avui dia. D'aquesta manera es poden impulsar les vendes, i gràcies a només aquesta manera d'analitzar les dades compromet un 35% de les vendes anuals. Aquesta metodologia d'analitzar dades és una forma molt eficient i tot e-Commerce hauria de fer-ho, ja que no només ajuda a l'empresa a poder fer bones recomanacions, sinó que també ajuda al client, ja que no fa falta que busqui per una web de 12 milions de productes, on encara que tinguis filtres segueixen sent molts productes els quals filtrar, a més a més que com a vegades passa potser no saps concretament el producte que vols i no tens manera de fer una cerca tan precisa. Com a empresa jo utilitzaria aquesta anàlisi de la conducta sense cap dubte.

Dayton Emily Bigcommerce [Online] // Amazon Statistics You Should Know: Opportunities to Make the Most of America’s Top Online Marketplace. - https://www.bigcommerce.com/blog/amazon-statistics/#amazon-everything-to-everybody.
Pathak Ritesh Analytics steps [Online] // How Amazon uses Big Data?. - Nov 03, 2020. - https://www.analyticssteps.com/blogs/how-amazon-uses-big-data.
Rangaiah Mallika Analytics steps [Online] // Customer Behavioral Analytics - An Overview. - Apr 08, 2020. - https://www.analyticssteps.com/blogs/customer-behavioral-analytics-overview.

Pregunta 4 (20% puntuació)

1- Les dades obertes són dades a les quals qualsevol pot accedir, utilitzar i compartir. Sigui governs, empresa o simplement una persona com qualsevol altra poden fer servir les dades obertes per obtenir beneficis socials, econòmics i ambientals. Es posen a la disposició dels usuaris multitud de continguts oberts perquè puguin fer amb les dades els que ells vulguin sense cap mena d'impediment, sigui consultar, analitzar, descarregar o generar aplicacions i serveis.

2- Té dues formes d'accedir-hi, es pot fer cercant per categories (economia, transport, ciència, salut...) i per sectors (aigües, energia, educació...). També té un input on escrius el que vols i et retorna una llista de dades segons la teva cerca. I un altre forma amb la que pots trobar també dades és amb un apartat de dades destacades. Aquestes dades poden ser utilitzades per qui sigui i pel que sigui segons posa a la seva pàgina Web, però n'hi ha llicències per poder crear o elaborar anàlisis d'aquestes amb la mateixa eina de cerca.

3- Tens diferents formats de dades, com són: en forma de document; de taules; o amb un mapa interactiu, depenent del tipus de dades que vulguis obtindràs un format o un altre. Et pots descarregar les dades en diferents formats segons les hagin pujat o segons el que més t'interessi. Aquests formats són: CSV, KML, Shapefile i GeoJSON.

4- Podem trobar dades de qualsevol classe com he mencionat anteriorment, però no només tenen dades seves, sinó que també utilitzen les dades d'altres portals d'Open Data com menciona a la seva pàgina Web. Aquestes dades només són capturades, organitzades per taules i les emmagatzemen en aquest cas en el servidor de base de dades d'aquesta empresa.

Clase10 [Online] // LAS 4 FASES DE LA GESTIÓN DEL CICLO DE VIDA DE LOS DATOS. - https://www.clase10.com/las-4-fases-de-la-gestion-del-ciclo-de-vida-de-los-datos/.