PRÀCTICA 2 – Creació de la visualització i lliurament del projecte

Aquest document explica tot el procés d'exploració, creació de visualitzacions i extracció de conclusions seguit al igual de com i perquè s'ha creat tot. Tot el codi, dades i accés a les visualitzacions es pot trobar al següent respositori GitHub:

https://github.com/Andreufb/PRACTICA2-VISUALITZACIO

Al README d'aquest s'explica l'estructura però crec que la millor forma d'avaluar aquesta pràctica es llegir i seguir aquest document ja que explica amb detalls to el que he fet i el seu perquè.

És complicat indicar des del principi les preguntes a respondre, he triat aquestes dades per fer un "estudi científic" en crear primer simples visualitzacions i anar analitzant el que m'estic trobant per veure que podria extreure d'interès o que és el que realment pot aportar valor a aquestes dades.

Encara així, les visualitzacions van dirigides a la comunitat científica europea interessada en l'estat i evolució actual d'una part del mediterrani i, més concretament, als científics d'EMSO (empresa per a la que treballo) que tenen en actiu projectes en els que es crucial entendre millor que està passant i que ha passat al Mediterrani; en especial, al mar Ligurian.

També m'agradaria que aquestes visualitzacions tinguessin un enfoc mediambiental centrat en el canvi climàtic ja que, encara que tingui poques dades, m'encantaria que fossin comprensibles per la població no científica ni tècnica per a, d'aquesta manera, conscienciar-los sobre aquest tema tant crucial.

Potser vull complir objectius massa diferents i arribar a un públic massa diferent però crec que alguna cosa puc aconseguir creant visualitzacions amb explicacions fàcils per a que el públic general ho entengui, sense limitar massa les relacions i informació de la visualització per a que un científic interessat en la zona pugui usar-les per avançar en la seva recerca, reflexionar sobre que està veient i passant o, al menys, entendre millor el problema de recerca que té davant.

Per tant, crec que es podran contestar preguntes amb prou precisió per a que un científic es doni per satisfet o, al menys, per a que ho pugui usar en conjunt amb altres dades i visualitzacions. Encara així, intentaré usar totes les dades de la font de dades, sense més fonts de dades externes per a que el científic o complementi amb les seves dades si li és necessari i que el ciutadà no s'agobiï amb un projecte massa gran/complex.

D'aquesta manera, com crear una història amb les diferents visualitzacions faria tot el treball més interessant i comprensible per al públic, he de generar alguna explicació i visualització que posi les dades en context; començant amb una introducció a les dades amb visualitzacions simples i acabant amb l'extracció de relacions i conclusions amb visualitzacions complexes.

Per explorar les dades, seguint amb el pensament de reproductibilitat mantenint la simplicitat, he arribat a la conclusió de que generar un contenidor de Docker és la millor idea amb diferència ja que m'asseguro que qui vulgui explorar les dades podrà seguir els meus passos i arribar a les mateixes solucions.

Pel que fa a aquest contenidor, en genero la imatge amb un Dockerfile i utilitzo Docker Compose per crear el contenidor amb aquesta imatge. Llavors, aquesta imatge és un Ubuntu 20.04 amb CUDA per a que es pugui utilitzar la GPU en el processat de les dades, diferents utilitats bàsiques com Python, pip, vim i git i les llibreries de Python per excel·lència (entre d'altres) a l'hora de processar i visualitzar les dades: Pandas, Math, Numpy, Scikit, Matplotlib i Seaborn.

També he decidit instal·lar les llibreries de Jupyter necessàries per poder usar Jupyter Notebook al contenidor creat i, d'aquesta manera, assegurar que les visualitzacions generades siguin fàcilment reproduïbles i modificables ja que els científics necessiten aquesta reproductibilitat per als articles, al igual que poder fer modificacions per acabar de perfilar les visualitzacions als seus estils i necessitats.

D'aquesta manera uso Jupyter Notebook i faig tota l'exploració, generació de visualitzacions simples i no tant simples i preparació de les dades per a visualitzacions més complexes amb aquest ja que, a part de l'exposat, així m'asseguro de poder entregar una guia a seguir fàcil i visual per arribar als mateixos resultats que jo i poder modificar el processat i visualitzacions a gust de l'usuari.

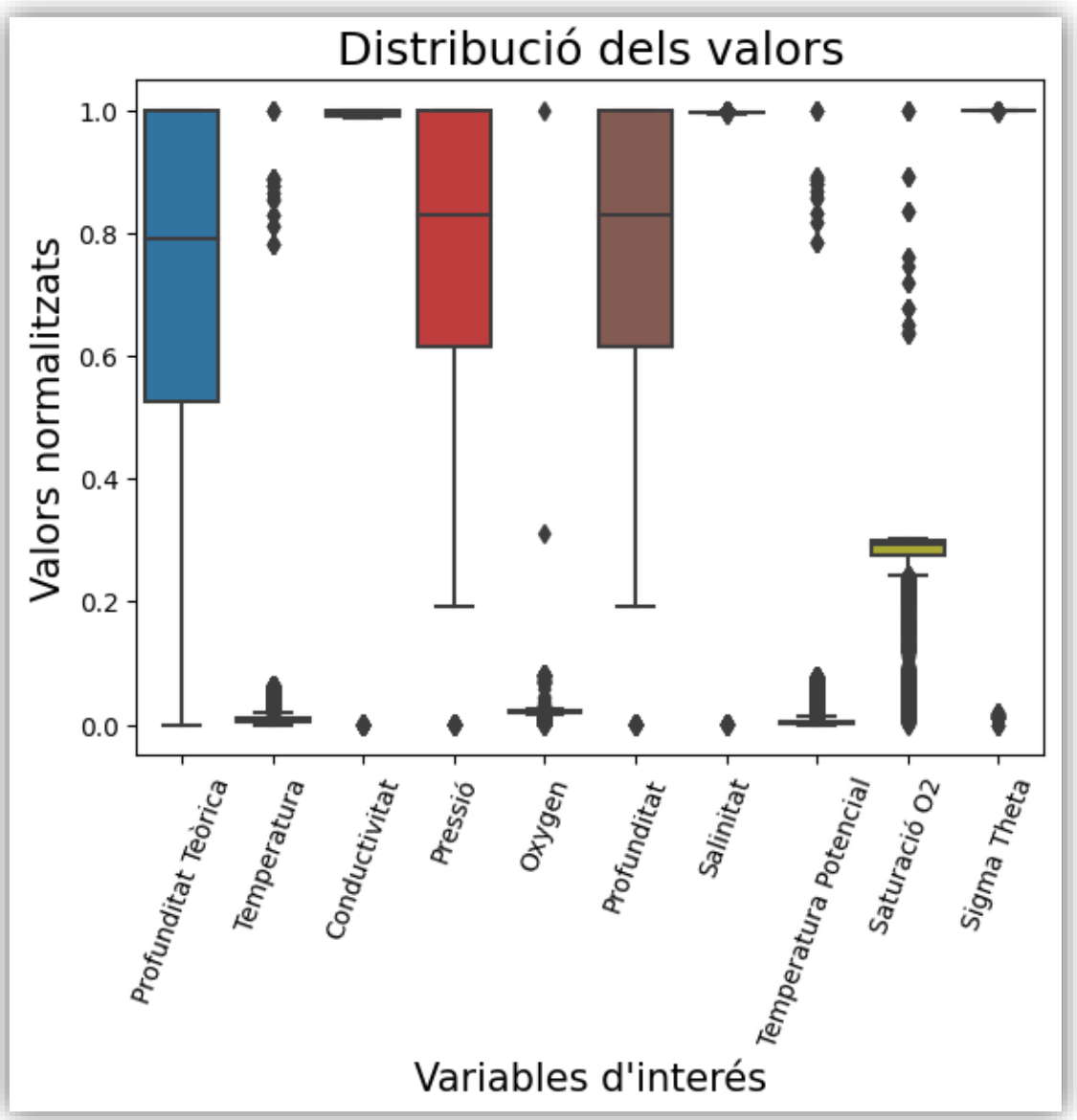
Un cop explicat com exploraré les dades per poder respondre a la necessitat plantejada de crear visualitzacions útils tant per a científics com per a ciutadans interessats en el tema mediambiental, començo amb aquesta exploració.

Durant l'explicació de l'exploració no mostraré algunes execucions de les que he extret part de l'explicació però es podran trobar al Jupyter Notebook del repositori GitHub i al HTML que generaré amb aquest per a una visualització més fàcil.

És crucial entendre les dades, assegurar-se de que tenen el format correcte i eliminar o arreglar els valors erronis i inexistents. Començo carregant les dades com un Dataframe de Pandas i observant les 10 primeres files; veig que s'han carregat bé però, en mirar-ne el tipus, s'observa com s'ha de modificar la columna de temps per a que es tracti com a tal.

Un cop he aconseguit tractar el temps com a tal sense perdre res d'informació (no ha estat fàcil), procedeixo a mostrar els valors mínims i màxims de cada columna mitjançant la implementació i execució d'una funció, obtenint així el que considero la primera visualització, només observable al Jupyter Notebook i HTML entregats. Però, ja que estic mirant-ne la forma, normalitzo les dades que un cop vist el seu tipus i valor mínim i màxim considero interessants a un nou set de dades, i uso aquest nou set per crear una visualització Box-Plot senzilla que pugui extreure del Notebook. Abans canvio el nom de les columnes pel bé de les visualitzacions i elimino les files que contenen valors vuits.

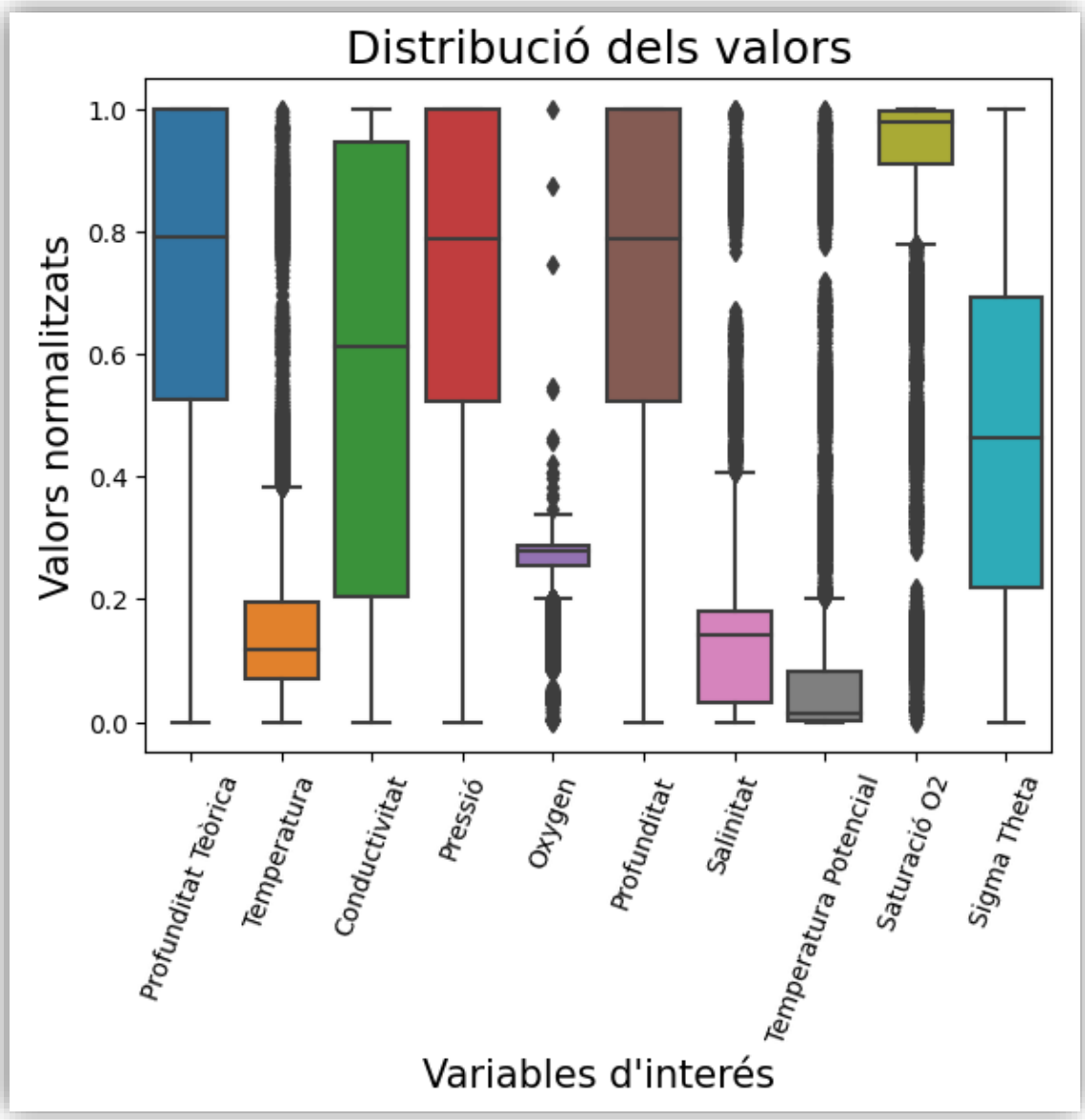
D'aquesta manera, la primera visualització guardada dins la carpeta visualitzacions és *1-BoxPlotExploratori.png*:



En aquest Box Plot de les variables d'interès normalitzades s'observa com a part de la profunditat i la pressió (molt relacionades), les altres variables tenen moltíssims valors atípics i, per tant s'ha de fer un processat d'aquestes.

Procedeixo a eliminar els valors atípics de les dades. Per fer-ho, per cada columna d'interès, extrec el primer valor del quartil 99 i del quartil 1 i filtro el dataset eliminant les files on la columna té un valor major al del quartil 99 o menor al del quartil 1.

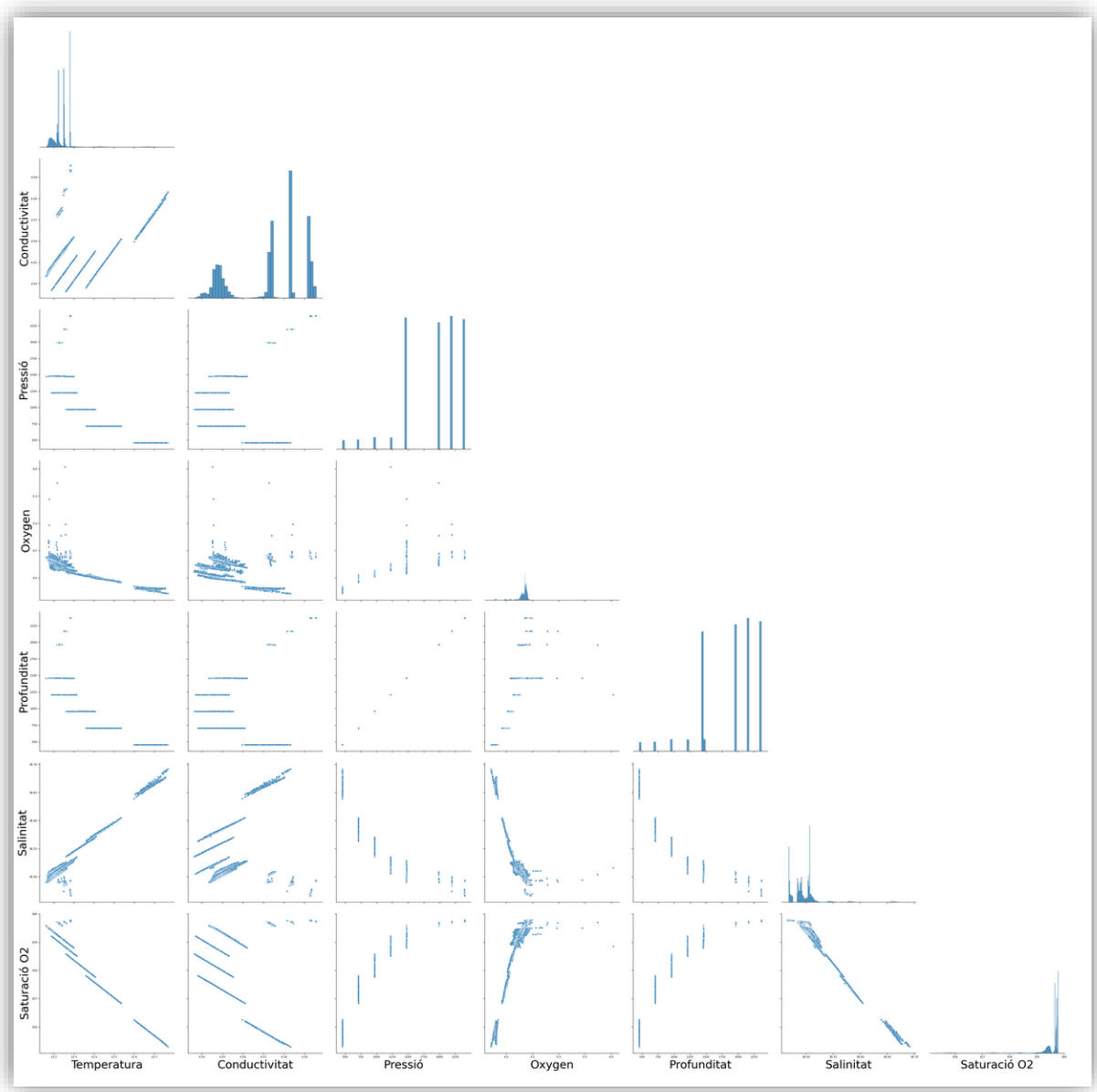
Ara, per fer un altre filtrat, calculo la desviació estàndard i elimino tots els valors que se'n vagin 2 desviacions estàndards de la mitja. Finalment, torno a normalitzar les dades i genero un Box Plot com l'anterior per veure la diferència tot mostrant-lo i guardant-lo com *2-BoxPlotDadesNetes.png*:



Ha hagut en canvi molt notable amb la distribució dels valors de les diferents variables on, encara que es segueixin marcant valors atípics, ja és més probable que es deguin a com Seaborn defineix els valors atípics a que realment siguin certs.

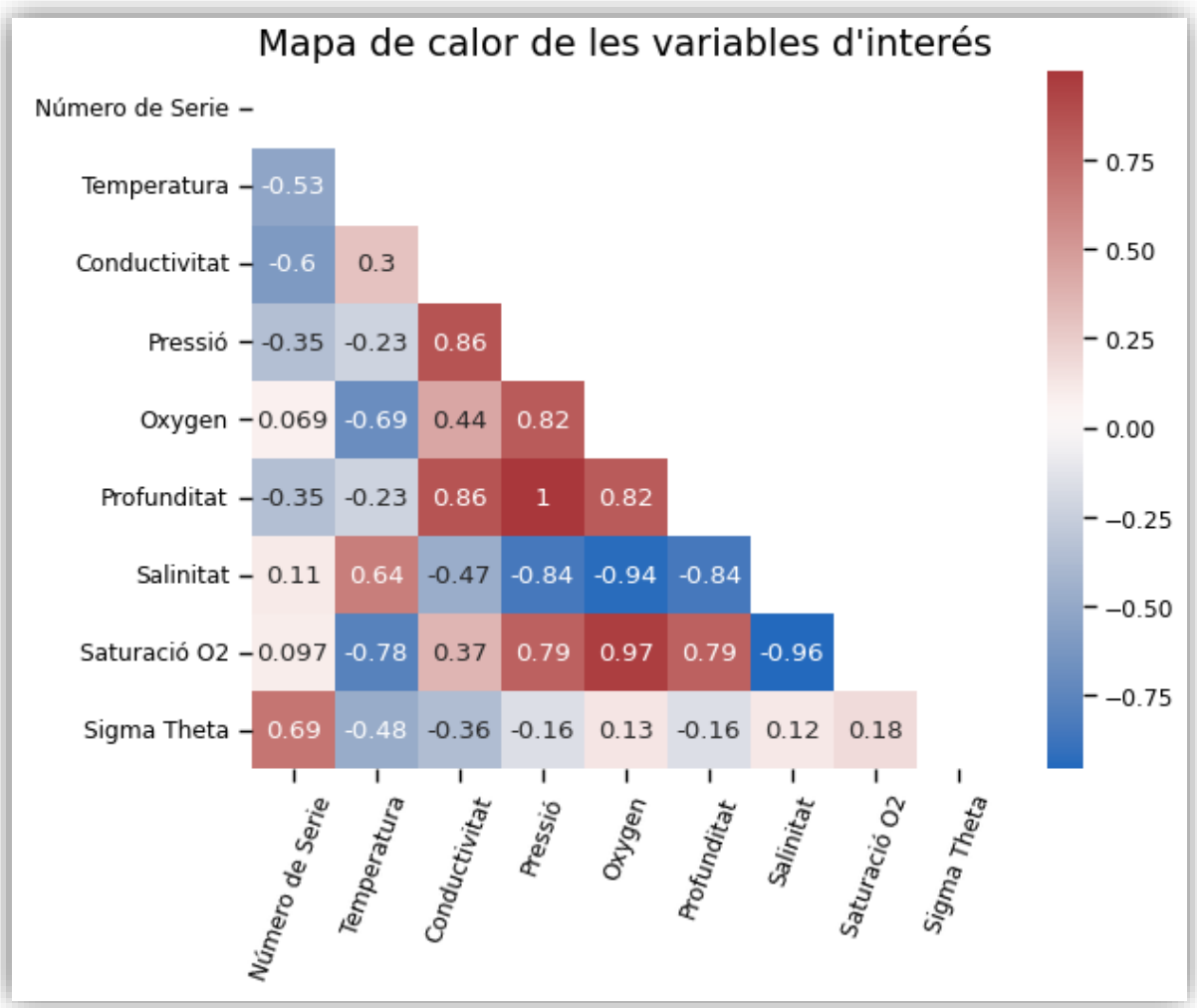
Veig que les parelles de variables profunditat teòrica i mesurada i, temperatura i temperatura potencial aporten la mateixa informació. Per tant, elimino les teòriques per quedar-me amb els valors adquirits pels sensors.

Ara, veient i coneixent les variables, genero una visualització nomenada *3-PairPlot.png* que compara les diferents variables d'interès entre sí amb gràfics de punts tot mostrant un histograma de les diferents variables a la vertical:



He anat modificant aquesta visualització per veure amb claredat quines variables són les que s'estan comparant però, encara així, s'ha d'obrir fora d'aquest document per poder observar-la amb claredat. He eliminat el triangle superior de comparacions i he adaptat la mida de les lletres i de la visualització en sí.

Amb aquesta visualització de 21 gràfics de punts i 7 histogrames es pot extreure molta informació però, abans de posar-m'hi, genero una altra visualització nomenada *4-MapaDeCalor.png*; un mapa de calor per veure quines variables estan relacionades i de quina manera per a, d'aquesta manera, poder extreure informació i generar hipòtesis més fàcilment:



He creat aquest mapa de calor calculant les correlacions, generant una màscara per eliminar el triangle superior, adaptant els eixos, mostrant a cada quadre la correlació de les dues variables en qüestió i pintant cada un d'aquest quadres seguint una paleta que va de blau (-1) a roig (1) passant per blanc (0) per veure fàcilment les correlacions més fortes ja siguin positives o negatives.

Per seguir avançant, em poso en la pell d'un científic per arribar a conclusions sobre les dades i crear les hipòtesis que dictaran quines visualitzacions crearé a partir d'aquí. D'aquesta manera, procedeixo a exposar el que extrec d'aquí:

1) Les parelles de variables profunditat-pressió i saturació en oxigen-oxigen, estan correlacionades al 100 i 97% indicant que amb el valor d'una es pot saber el de l'altra i, per tant, no calen 4 sensors sinó 2 (un per cada parella).

2) Hi ha una gran correlació de l'oxigen a l'aigua amb la temperatura, la pressió i, sobretot, amb la salinitat on, com més O2 hi ha, més profunditat i fred i menys salinitat hi ha. Aquest resultat indica que hi ha una comunitat de flora i fauna sana ja que l'oxigen de les capes superiors és usat per aquests i, a mesura que baixem en profunditat, no es gasta l'O2 en no haver tanta vida i la descomposició d'aquesta ajuda a l'augment de la concentració d'O2. Però, només amb aquesta informació no es sap del cert que segueixi aquest patró arribant a les preguntes de:

- Hi ha diferents comunitats d'organismes al llarg de la columna d'aigua?
- Les diferents possibles comunitats canvien segons l'època de l'any?

3) La variable Sigma Theta, densitat de l'aigua a una temperatura donada, té una lleugera correlació negativa amb la temperatura. Per tant, encara no descobrint res de nou, s'observa com a més densitat té l'aigua, menor és la temperatura d'aquesta.

4) La salinitat disminueix amb la profunditat, és un comportament anòmal ja que normalment com l'aigua més salada és més densa, va cap al fons, per tant, alguna característica de l'aigua o la zona fa que l'aigua més salada es trobi al fons. Només amb aquesta correlació no es pot veure més però es poden plantejar les preguntes:

- Com es distribueix l'aigua en profunditat segons la seva salinitat?
- En cas de disminuir la salinitat en profunditat, a que es deu aquest comportament inusual?

5) Mirant els histogrames de les diferents variables, pareix que:

- Les variables temperatura, oxigen i salinitat tenen valors atípics en la seva part superior.
- La variable saturació en oxigen té valors atípics en la seva part inferior.

No se a que es deuen aquests valors però; pot ser que els diferents sensors estan calibrats diferent fent que, en conjunt, algunes variables tinguin valors atípics?

Amb aquesta exploració he pogut extreure algunes conclusions i plantejar 5 preguntes a les que intentaré donar resposta en la visualització:

Hi ha diferents comunitats d'organismes al llarg de la columna d'aigua? Com es distribueix l'aigua en profunditat segons la seva salinitat? En cas de disminuir la salinitat en profunditat, a que es deu aquest comportament inusual?

Per donar resposta a aquestes tres preguntes, despès de donar-li voltes i provar diferents visualitzacions, he decidit preparar les dades per crear una carrera de línies animada amb la variació de les variables implicades (temperatura, oxigen i salinitat) en profunditat i, d'aquesta manera, poder donar resposta a tres de les preguntes plantejades amb una sola visualització.

Començo creant una copia de les dades normalitzades usades abans i quedant-me amb les columnes de temperatura, oxigen, salinitat i profunditat. Un cop ho tinc, he buscat per la interfície gràfica Flourish el gràfic desitjat i he mirat la forma que han de tenir les dades arribant a la conclusió de que el més fàcil és usar la columna de profunditat (sense normalitzar) com a primera fila del dataset i les altres variables com a noves files.

D'aquesta manera, afegeixo la columna de profunditat sense normalitzar i ordeno el Dataframe pel valor d'aquesta columna. Ara, genero una llista que vagi de la profunditat mínima a la màxima en salts d'1 i una altra que indiqui l'interval al que fa referència el valor per crear una variable nomenada "bins" que indiqui a quin interval fa referència la profunditat en qüestió. Un cop tinc aquesta variable, agrupo els valors del dataset per aquesta tot mostrant la mitja de les altres variables.

Per acabar amb el processat, elimino la columna de profunditat i visualització els possibles valors nuls, defineixo que la columna bins és l'índex del dataset i el transposo per tenir les files com a columnes (format que necessito per Flourish).

Per poder descarregar les dades fàcilment i incorporar-les a Flourish, genero una funció que deixa descarregar el Dataframe de Pandas directament al Notebook mitjançant la creació d'un enllaç. Aquest fitxer es pot descarregar del Notebook directament o extreure de la carpeta datasets. Així, l'executo, descarrego les dades i les incorporo a Flourish per generar visualització que ajudi a respondre la pregunta:

https://public.flourish.studio/visualisation/12502316/

En aquesta visualització generada, he afegit títol, subtítol, valors als eixos x i y i peu d'imatge per a donar un bon context a aquesta i que es pugui entendre per si sola.

He anat modificant els tipus de lletra, colors i mida per a millorar la visualització al igual que el tipus de línies, colors, mides i velocitats.

Es pot extreure molta informació d'aquesta visualització però, centrant-me amb les preguntes:

Hi ha sis comunitats diferents al llarg dels punts de la columna d'aigua en els que s'han adquirit dades ja que just la concentració d'oxigen disminueix en 6 punts indicant que s'usa per alguns organismes (disminueix), és expulsat d'aquests (augmenta) i torna a ser usat. S'hauria de comprovar amb mostres d'aigua de diferents profunditats per buscar restes orgàniques però, després de preguntar, són molt cares i no es solen adquirir aquest tipus de dades.

La salinitat segueix un patró invers a la temperatura amb gran exactitud on, a mesura que augmenta la profunditat, augmenta poc a poc tal i com sol ser però, sobre els 2150 metres, disminueix d'una forma bestial a valors propis d'aigua dolça mentre la temperatura augmenta en la mateixa mesura. D'aquesta manera, el que pareix més lògic es que hi ha una termoclina generada per una gran massa d'aigua dolça provinent d'un riu important o d'una forta ploguda que per factors com vents o densitats s'enfonsa fins aquestes profunditats fent que en global pareix que la salinitat disminueix amb profunditat, però no és així, simplement hi ha una massa d'aigua dolça més càlida i menys salada. Aquesta conclusió té més sentit mirant l'O2 on, a aquest punt de 2150 metres, hi ha un pic de concentració d'oxigen, indicant així que aquesta aigua no ha estat molt en superfície sinó que ràpidament s'ha enfonsat mantenint l'oxigen que tenia quan estava en superfície.

Un cop contestades aquestes tres preguntes, he conclòs que hi ha diferents comunitats i, per tant, ja puc plantejar-me i resoldre la pregunta de:

Les diferents comunitats canvien segons l'època de l'any?

Aquesta pregunta es pot respondre veient l'evolució de la concentració d'oxigen al llarg de la columna d'aigua les diferents èpoques de l'any. D'aquesta manera, començo adquirint un dataset que només contingui el mes, concentració d'oxigen i profunditat.

Després de donar-li voltes a com mostrar aquesta informació d'una manera clara i bonica per permetre al públic entendre fàcilment que està passant, he decidit generar un gràfic de bombolles de Floursih. Així, el procés per aconseguir les dades, les quals estan disponibles a la carpeta datasets com *cummuinity_change_data.csv* i descarregables des del Notebook, ha estat:

- Adquirir les columnes d'interès: Oxigen, Profunditat, Mes de l'any.
- Compartimentar la profunditat en sectors i eliminar la profunditat com a tal.
- Eliminar les files amb valors nuls i ordenar el dataset segons el mes de l'any.

- Agrupar les dades segons aquests sectors de profunditat i segons el mes.
- Eliminar les files amb valors nuls i ordenar el dataset segons la profunditat.
- Normalitzar la columna Oxigen per mostrar millor el canvi de concentració d'oxigen en la mida de les bombolles representades.
- Crear un link de descarrega del dataframe generat i posar-lo al Notebook.

Ara, he agafat les dades del CSV generat (*cummuinity_change_data.csv*) i les he afegit a Floursih tot fent les modificacions necessàries per arribar a aquesta visualització:

https://public.flourish.studio/visualisation/12506990/

Tal i com s'observa he fet moltes modificacions; he afegit títol, subtítol, valors als eixos x i y i peu d'imatge per a donar un bon context a aquesta i que es pugui entendre per si sola. He anat modificant els tipus de lletra, colors i mida per a millorar la visualització al igual que el tipus de bombolles, colors, escales i informació que mostren al passar per sobre.

Es pot extreure molta informació d'aquesta visualització però, centrant-me amb la pregunta a respondre, primerament veig que no tots els mesos estan representats a totes les profunditats i, per tant, s'haurien d'adquirir més dades a les diferents èpoques de l'any ja que no existeixen a cap lloc.

Ara, centrant-me amb la informació que tinc, s'observa com per a la mateixa profunditat segons el mes de l'any hi ha una o altra concentració indicant així que no s'està fent ús de la mateixa quantitat d'oxigen i, per tant, encara que es podria deure a diferents corrents d'aigua segons l'època, pareix que hi ha diferents comunitats segons l'època de l'any ja que cada comunitat usa i deixa l'aigua amb una concentració d'oxigen diferent.

Aprofundint més en la visualització i gràcies al poder mostrar només els mesos d'interès, veig que encara faltant moltes dades per alguns dels mesos, passat l'estiu (Setembre) hi ha una elevadíssima concentració d'oxigen per la descomposició de tots els organismes que es creen a les capes superficials durant l'estiu, que ben segur fa canviar la comunitat d'organismes a les diferents profunditats i, llavors, poc a poc va disminuint aquesta concentració mes a mes (diferents comunitats) fins que al Febrer s'estabilitza (mateixa comunitat). Llavors, finalment, al Juliol comença a augmentar poc a poc fins l'Agost i, dallí es torna al màxim de Setembre.

Per tant, encara faltant dates, fent una bona inspecció de la visualització, arribo a la conclusió que realment hi ha diferents comunitats segons l'època de l'any encara que és molt complicat extreure'n el nombre.

Els diferents sensors estan calibrats diferent fent que, en conjunt, algunes variables tinguin valors atípics?

Per donar resposta a aquesta pregunta he fet una copia de les dades, he seleccionat les columnes conflictives (amb més valors atípics), he afegit una columna que indica el mes i una que compartimenta la profunditat.

Després de pensar i provar diverses visualitzacions, he arribat a la conclusió de que he de veure per a cada profunditat i mes de l'any, quins sensors marquen quins valors d'una de les variables conflictives i anar creant diverses visualitzacions per a les diverses variables. Per tant, la visualització que millor em deixa representar aquestes dades, tot podent fer comparacions entre els valors dels sensors és un gràfic de punts dividit segons el sensor i que mostri per a un mes i una profunditat determinats quins valors marquen els sensors per a una de les variables conflictives.

D'aquesta manera, segueixo el processat eliminant la columna de profunditat i agrupant les dades restants segons el sensor, el mes i la profunditat compartimentada. Ara, he re-ordenat les files segons el valor de la profunditat i he creat un link de descarrega de les dades per a que es puguin descarregar des del Notebook. Així, les dades es troben a la carpeta datasets com *sensor_data.csv*.

Un cop tinc les dades, les he afegit a Floursih tot fent les modificacions necessàries per arribar a aquesta visualització:

https://public.flourish.studio/visualisation/12507921/

Tal i com s'observa he fet moltes modificacions; he afegit títol, subtítol, valors als eixos x i y i peu d'imatge per a donar un bon context a aquesta. He modificat els tipus de lletra, colors i mida per a millorar la visualització al igual que el tipus de punts (un per cada sensor), colors i informació que mostren al passar per sobre.

Ha estat una mica frustrant veure aquesta última visualització, m'he donat compte que no puc respondre la quinta pregunta amb la suficient precisió ja que cada sensor s'utilitza a una profunditat determinada i quasi mai s'usa el mateix sensor per adquirir les mateixes dades al mateix mes i profunditat fent que no es puguin comparar les dades adquirides pels diferents sensors i veure si realment estan calibrats de forma diferent.

Per tant, per contestar aquesta pregunta o s'usen varis sensors al mateix temps i profunditat o és fan probes al laboratori amb valors coneguts de diferents variables per veure si els diferents sensors marquen valors diferents. Encara així, hi ha 3 punts on el mateix mes i profunditat es van suar sensors per adquirir dades i, per tant, tinc algunes dades per extreure conclusions.

Centrant-me amb aquests 3 punts, s'observa com els valors indicats pels sensors són molt similars, tenen diferències a partir del tercer decimal i, per tant, amb la poca informació que tinc dedueixo que els sensors funcionen bé, estan ben calibrats i els

valors atípics es deuen més a una qüestió de com es marquen en Seaborn que no pas a que realment existeixin.

Un cop contestades les 5 preguntes proposades i donat per acabada el procés d'exploració i visualització, procedeixo a descriure els fets més interessants que he extret, incloent els resultats inesperats i que no he tingut en compte:

- S'ha augmentat el coneixement global sobre aquesta zona del mar Ligurian ja que ningú havia generat visualitzacions amb aquestes dades i menys explorar-les per veure realment que poden aportar.
- S'han donat resposta a preguntes amb prou precisió per a que un científic ho pugui usar en conjunt amb altres dades i visualitzacions per arribar a conclusions robustes sobre el que està passant ja que com he comentat es necessiten dades adquirides en futurs estudis/projectes per completar-se.
- Relacionat amb l'anterior punt, no he tingut en compte la completesa de les dades i que encara havent més de 80.000 files, per a un estudi científic es queda just ja que es necessita robustesa i completesa en els resultats.
- Hi ha explicacions i visualitzacions que posen les dades en context anant de lo més simple a lo més complex.
- S'ha generat un contenidor de Docker amb Jupyter Notebook per a fer tot el processat de les dades i algunes de les visualitzacions per entregar una guia fàcil i visual per arribar als mateixos resultats i poder fer les modificacions que l'usuari cregui necessàries.
- En el primer diagrama de caixes he vist la magnitud de la tragèdia en topar-me amb moltíssims valors atípics fent que haguí aplicat un processat doble:
    - Eliminar les dades amb valors superiors al quartil 99 i inferiors a l'1.
    - Eliminar les dades que se'n vagin 2 desviacions estàndards de la mitja.
- Després de crear la visualització 3 i 4 (28 gràfics comparatius i el mapa de calor) he pogut extreure molta informació per extreure conclusions i plantejar 5 preguntes a les que dono resposta amb les últimes 3 visualitzacions.
- La quinta visualització és una carrera de línies animada amb la variació de la temperatura, oxigen i salinitat en profunditat de la que he conclòs:
    - Hi ha 6 comunitats diferents al llarg dels punts de la columna d'aigua.
    - Pareix que la salinitat disminueix amb profunditat, però no és així, hi ha una termoclina generada per una gran massa d'aigua dolça provinent d'un riu important o d'una forta ploguda.
- La sexta visualització és un gràfic de punts complex en el que he pogut veure que hi ha diferents comunitats segons l'època de l'any encara que és molt complicat extreure'n el nombre.
- L'última visualització també és un gràfic de punts en el que he vist que cada sensor s'usa a una profunditat diferent i, per tant, pareix que els sensors funcionen bé, estan ben calibrats.