**soproni egyetem**

**simonyi károly műszaki, faanyagtudományi és művészeti kar**

**informatikai és gazdasági intézet**



**új nemzeti kiválóság program**

**ÚNKP-18-1-I-SOE-27**

**dokumentáció**

**csanaki richárd**

**dr. pödör zoltán, dr. jereb lászló**

**kzeadr**

**sopron,**

**2019. április 10.**

**Ösztöndíjas időszak**

2018. november 1. – 2019. március 31.

**Vállalások**

1. Kutatási tervben megfogalmazott célok teljesítése
2. Havonta legalább egy magyar vagy idegen nyelvű szakmai anyag feldolgozása
3. Egyetemi ÚNKP rendezvényen való részvétel

**Feladat megfogalmazás**

A Soproni Egyetem SKK Informatikai és Gazdasági Intézménye által működtetett SensorHUB [1] infrastruktúrában tárolt szenzoradatok automatizált elemzése adatbányászati, Big Data megoldásokkal [2]. Valamint informatív adatvizualizáció az adathalmazokból kinyert ténylegesen új információk szakmai szempontú megjelenítésére [3]. Ezeken az információkon alapuló döntéshozatal támogatása, majd a döntéshozatali ciklusból a humánfaktor részvételének minimalizálása, gépi tanulás (machine learning), öntanuló algoritmusok és az emberi agy (intelligencia forrása, neuronhálózatok rendszere) modellezési lehetőségeinek vizsgálata, a lehetséges megoldások illesztése az adott feladathoz.

**Feldolgozott irodalom, anyagok, kurzusok**

1. Life 3.0 – Max Tegmark
2. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython – Wes McKinney
3. Python in 24 Hours – Katie Cunningham
4. Tidy Data – Hadley Wickham
5. Hands-On Machine Learning with Scikit-Learn & TensorFlow
6. Think Stats (Exploratory Data Analysis in Python) – Allen B. Downey
7. Pandas for Data Analysis, SciPy 2017 Tutorial – Daniel Chen
8. Introduction to Probability (Harvard Statistics 110) – Joseph K. Blitzstein

**Felhasznált technológiák**

* Python
  + Interpretált programozási nyelv, Big Data és ML területen standard
* Anaconda
  + Könyvtár és virtuális környezet csomag Python nyelven való fejlesztéshez
* CDS API
  + Copernicus adatbázisok lekérdező felülete
* SensorHUB
  + Intézeti adattároló környezet
* ecCodes
  + Terminál eszköztár GRIB és NetCDF formátumú fájlok kezelésére
* CMake
  + C alapú build eszköztár
* QGIS
  + GRIB, netCDF és további formátumok megjelenítésére képes open source alkalmazás
* Pandas
  + Tabuláris adatelemző könyvtár
* Matplotlib
  + Tudományos adatvizualizációs könyvtár
* NumPy
  + Nagy teljesítményű mátrix könyvtár
* Scikit-Learn
  + ML eszköztár
* Jupyter Notebook
  + Interaktív Python interpreter (Ipython Shell)
* XArray
  + N-D dimenziójú tabuláris adatszerkezet
* CDO
  + Climate Data Operator – Parancssoros eszköztár netCDF fájlok kezelésére
  + Ugyanilyen névvel Python könyvtár hasonló funkciókkal
* GitHub és GitHub Desktop
  + Felhőalapú verziókövető szolgáltatás

**Felhasznált technológiák – bővebben(\*)**

***Python***

A Python egy általános célú, high-level interpretált programozási nyelv. Főbb ismérvei a könnyű olvashatóság és az élénk fejlesztői közösség. Webfejlesztés, adatelemzés és gépi tanulás területeken standardnak számít.

***Anaconda***

Data Science területen standard Python disztribúció, könyvtár gyűjtemény. Több, mint 1000 open source csomagot tartalmaz, többek között a Pandas, NumPy, Matplotlib, Scikit-Learn és TensorFlow könyvtárakat.

***Pandas***

Python könyvtár könnyen használható, nagy teljesítményű adatstruktúrákkal, adatelemző eszközökkel.

***NumPy***

Python könyvtár N-dimenziós tömbökkel és ezek transzformációs eszközeivel. A hagyományos adatszerkezeteknél könnyebb és gyorsabb feldolgozhatóságot biztosít tetszőleg adattípusokkal.

***Scikit-Learn***

Klasszikus machine learning Python könyvtár, regressziós, klaszterezős és klasszifikációs modulokkal. Támogatja az ún. Supervised és Unsupervised gépi tanulás módszereket is.

***Jupyter Notebook***

IPython Shell alapú böngészős Python fejlesztői környezet, mely több integrált könyvtár segítségével és megoldással könnyíti meg a projektek dokumentálását: Markdown markup nyelv támogatása, Matplotlib grafikonok kirajzolása, adatszerkezetek automatikus megjelenítése.

***SensorHUB [1]***

Az Informatikai és Gazdasági Intézet által működtetett adattároló technológia, az országszerte kihelyezett 6 mérőállomás (Püspökladány, Sárvár, Kecel, Napkor, Karád, Pilisszentlélek) által küldött adatok tárolási helye, az elkészült általános Gateway alkalmazás [2] átmeneti futtató környezete.

***CDS API***

Copernicus Climate Data Store lekérdező felülete, Python környezetből elérhető, az egész Földre kiterjedő meteorológiai adatokat tárolja, könnyen hozzáférhető programozói interfészt biztosítva.

***ecCodes***

Az ECMWF által fejlesztett parancssori eszköztár, melynek segítségével a GRIB és BUFR kiterjesztésű fájlok feldolgozhatóak, átalakíthatóak más programok által is feldolgozható formátumba. A klasszikus GRIB-API bővített verziója.

***CMake***

C alapú build alkalmazás, hordozható tudományos alkalmazások telepíthetőek vele.

***QGIS***

Grafikus felhasználói felülettel rendelkező alkalmazás mely automatikusan képes GRIB és netCDF kiterjesztésű fájlokat megjeleníteni, elemezni.

***GitHub és GitHub Desktop***

Web alapú, a git verziókövető rendszert használó szolgáltatás, melynek nem csak parancsoros eszközei vannak, hanem grafikus felhasználói felületet támogató alkalmazásai is. Ilyen alkalmazás a GitHub Desktop is, melynek segítségével a forráskódban és a dokumentációban eszközölt változások követhetők figyelemmel.

A projekt forráskódját és dokumentációját tartalmazó GitHub repository itt található meg:

< <https://github.com/csana23/UNKP> >

**(\*)**

A felsorolt alkalmazások, programok között nem szerepel a Microsoft által fejlesztett Power BI, attól függetlenül, hogy a kutatási tervben ezt az alkalmazást jelöltem meg az elsőszámú adatelemző technológiaként. A tervhez képest azért nem ezt használtam, hiszen nagyvállalati környezetben, folyamatos és standardizált dataflow-k elemzésére kiváló eszköz, azonban mi, ezen projekt folyamán nem ilyen adatforrásokkal dolgoztunk, hanem sokrétű, egymástól független és különböző szerkezetű adathalmazokkal. Így egy ilyen eszköz használata nem volt célravezető, továbbá a meteorológiai jellegű adatok elemzéséhez standard eszközök mind open source szoftverek, melyek szépen el vannak látva Python könyvtárakkal, API-kal. Természetesen ennek a megközelítésnek is megvannak a hátrányai: mindegyik fájl formátumhoz külön parancssori eszköz telepítése, új API-k kezelésének elsajátítása...

***A projektben részt vevő személyek***

Attól függetlenül, hogy a pályázati időszakban még úgy tűnt, hogy a munka során egyedül Dr. Pödör Zoltán szakmai konzulensemre számíthatok, rajta keresztül több, a szakmai kérdésekben rendkívül sok segítséget nyújtó személlyel dolgozhattam együtt.

Dr. Pödör Zoltán, szakmai konzulens

Az ÚNKP projekt koordinátora, adatbányászati technikák és ezek elsajátításukhoz szükséges anyagok forrása.

Dr. Gálos Borbála

Az erdészeti kapcsolatunk, az ő irányadásával nyílt új fejezet a projekt folyásában (indexek, modellek összehasonlítása, új adatforrások).

Molnár András

Doktorandusz hallgató, vele együtt fedeztük fel az új adatforrásokat és dolgoztunk a projekt második felén.

**Adatok**

Az elemzések első és legfontosabb lépése a sok és jó minőségű adat összegyűjtése.

*Milyen mennyiségtől tekintünk egy adathalmazt Big Data-nak?*

Tudományos körökön kívül, egyéni felhasználóként akár az 1-től 3 gigabájt (109 bájt) méretig terjedő adathalmazok már hatalmasnak számítanak, főleg a személyi számítógépek drága és limitált memóriakapacitása miatt. Az elemzéshez használt adatokat azért kell a memóriában tárolni, hogy az egyes Data Science könyvtárak által alkalmazott algoritmusok lefutása viszonylag gyorsan lefuthasson, pár GB-nyi adattal (1-5 évig terjedő adathalmaz) PC-ken is pillanatok alatt képesek ezek a programok eredményt generálni, főleg azért, mert ezen méreteknél nagyságrendekkel nagyobb mennyiségű adat feldolgozására lettek felkészítve.

Viszonytásképp a szuperszámítógépekkel felszerelt intézetek és vállalatok napi szinten petabájt (1015 bájt) nagyságrendű adatbázisokkal dolgoznak. Ebből következik, hogy maga a „Big Data” elnevezés erősen szituáció függő.

*Mitől lesz jó minőségű az adat?*

Az adatok összegyűjtése, azok szűrése és manipulálása külön tudományággá vált. Erről szól Hadley Wickham tanulmánya is, mely kiinduló standarddá vált adat elemező körökben. Az ebben az anyagban megfogalmazott alapelvek:

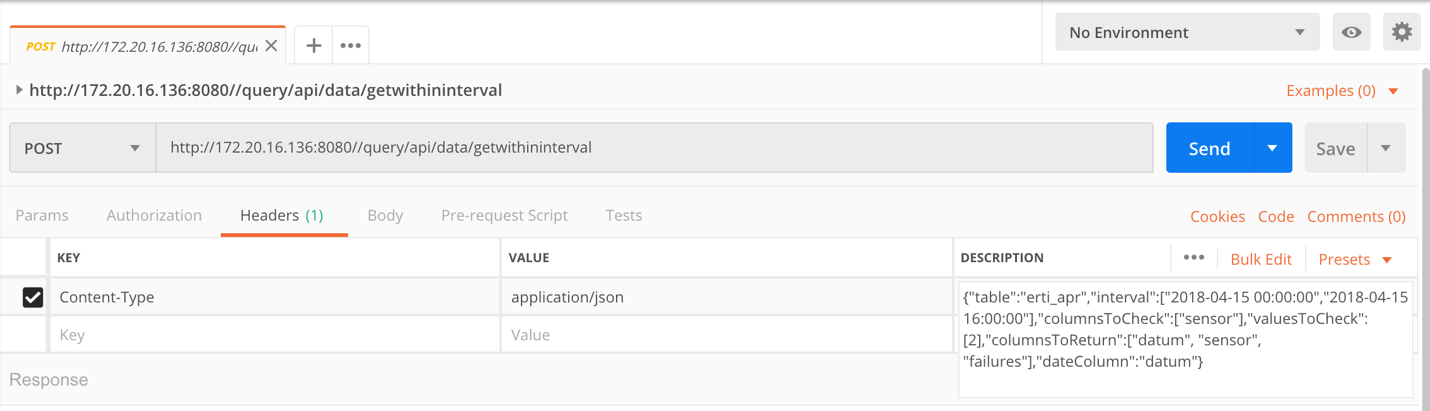
1. Az egyes jelenségeket leíró adathalmazok feloszthatók tulajdonságokra és megvalósulásokra, azaz attribútumokra és ezek alapján rekordokra.
2. Az attribútumok a jelenség elemi tulajdonságait jelentik, így elemi adattípusokkal leírhatók.
3. A jelenséget leíró adathalmazok nem feltétlenül tabulárisak (táblázatos felépítésűek), bár a különböző adatbázis elmező szoftverek a tabuláris adatbázisok elemzésire már jól optimalizáltak.
4. Ne féljünk belenyúlni a megkapott adathalmazba! Ez az úgynevezett „data wrangling” folyamat, melyet jó közelítéssel adatmanipulációnak nevezhetünk. Ezen lépés teszi ki az adatelemzői procedúra oroszlánrészét, nem véletlenül: a hibás adatok (rossz mérések, duplikátumok, stb.) használata rossz és elvi hibás következtetések levonását eredményezik.
5. Ügyeljünk arra, hogy minimalizáljuk az adatmanipuláció során elvesztett adat mennyiségét, ugyanis a hiányzó adat sokszor hiányzó információt jelent.

***A projektben használt adatforrások***

* Az intézeti SensorHUB környezet 2015 év szeptemberéig visszamenőleg tárol adatokat, melyek forrása az Erdészeti Tudományos Intézet (ERTI) által működtetett 6 mérőállomás (Püspökladány, Sárvár, Kecel, Napkor, Karád, Pilisszentlélek). Itt a Query API segítségével JSON formátumban kérhetők le az adatok.
* Copernicus Climate Data Store – bárki által elérhető klímaadatok online tárháza, mely a bemutatott CDS API-val kezelhető. Főleg az ERA5 adatbázissal dolgoztunk (egész Földre kiterjedő mérési adatok), mely GRIB formátumban tárolja az eredményeket.
* ESGF Node – A CDS-hez hasonlóan ez is Európai Úniós kezdeményezés, innen netCDF formátumú fájlok érhetőek el, szintén globális lefedettséggel.

***Az adatok összegyűjtése***

* SensorHUB

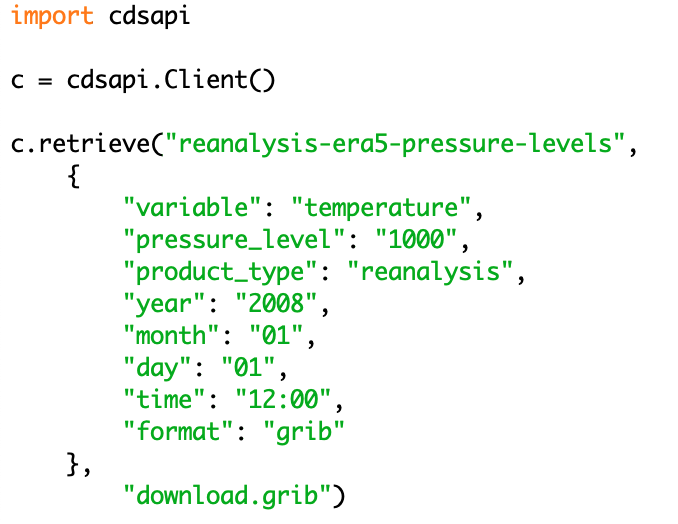


*Egyszerű HTTP lekérdezéssel*



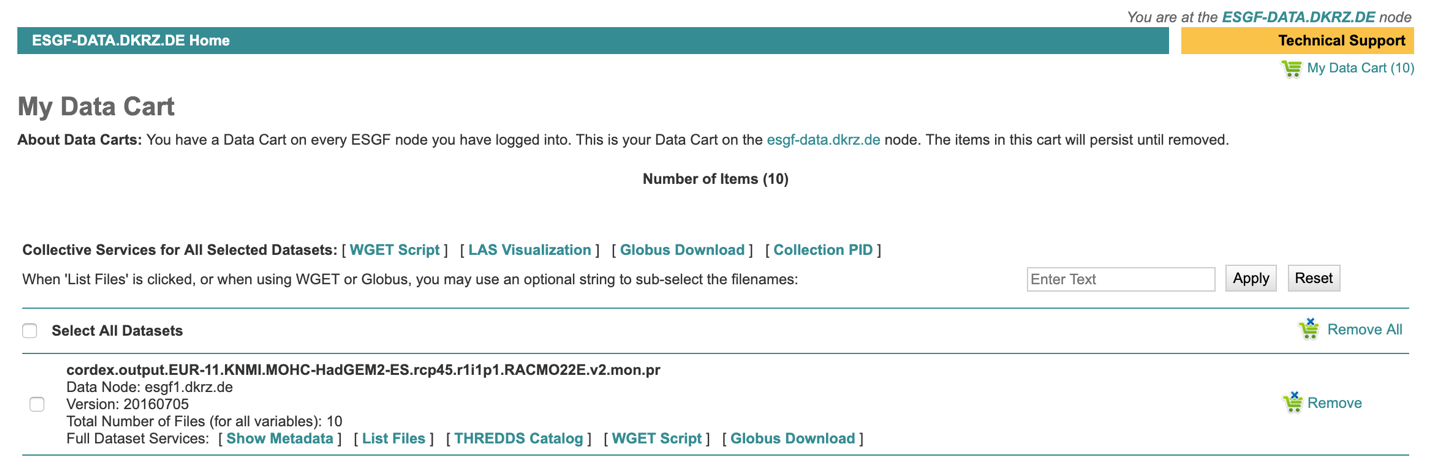
*Pythonból a cassandra-driver könyvtár függvényeivel*

* Copernicus Climate Data Store



*ERA5 adatok lekérdezése CDS API segítségével*

* ESGF Node

**

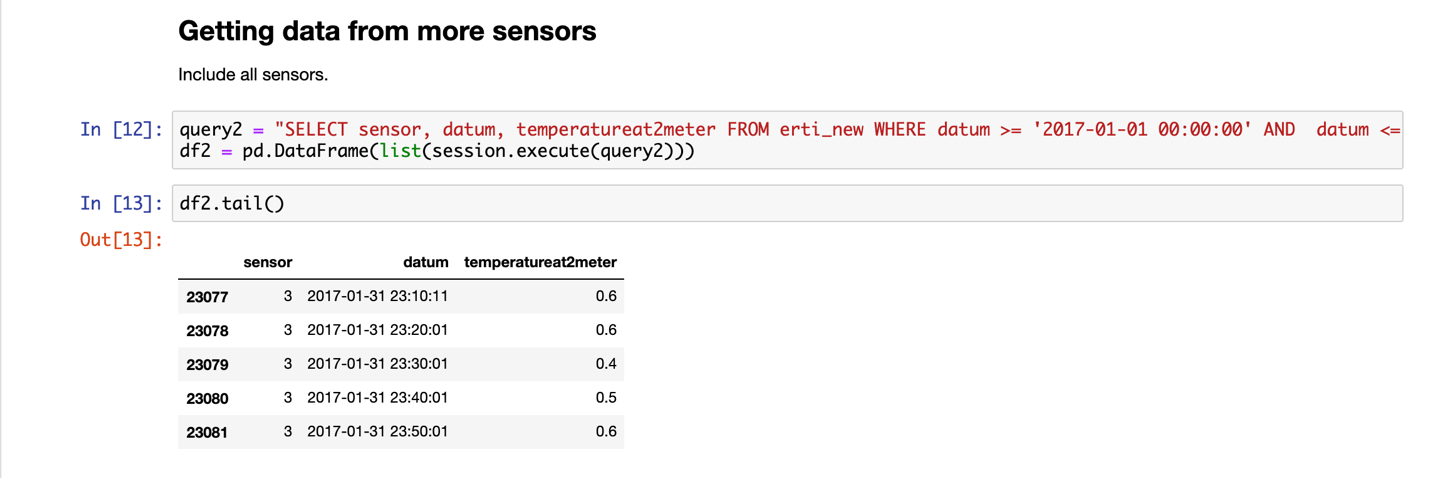
*ESGF Node Data Cart egyik eleme*

**

*Data Cart kicsomagolása wget parancssori eszközzel és OpenID-val*

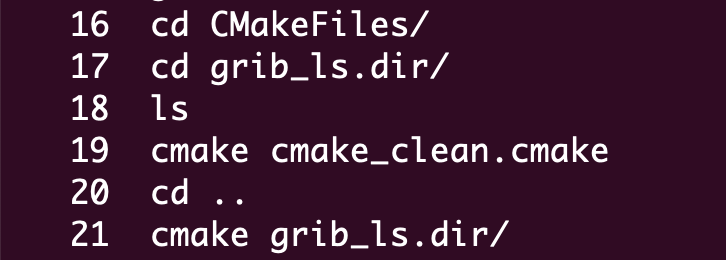
***Az adatok feldolgozása, előkészítése elemzésre***

* SensorHUB

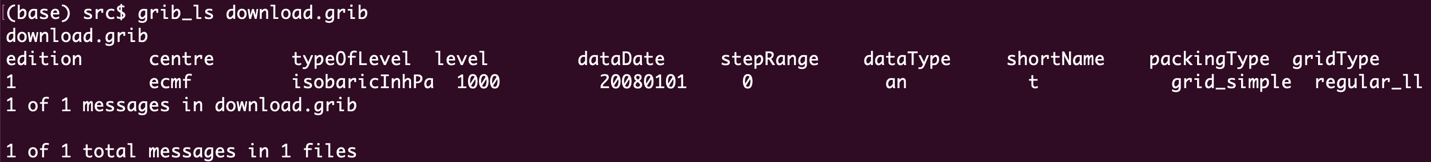


*A lekérdezés eredményét egy tabuláris adatszerkezetté konvertáljuk, majd a hibás adatokat NaN-al jelöljük*

* Copernicus Climate Data Store
  + GRIB fájlok értelmezése ecCodes parancssori eszközzel



*ecCodes telepítése CMake parancsori eszközzel*



*ecCodes grib\_ls parancs a GRIB fájl metaadatainak listázására*

* + GRIB fájl átkonvertálása CSV formátumba ecCodes segítségével



*ecCodes parancs CSV formátumba konvertáláshoz*

* + Kapott CSV fájl formázása Python-nal



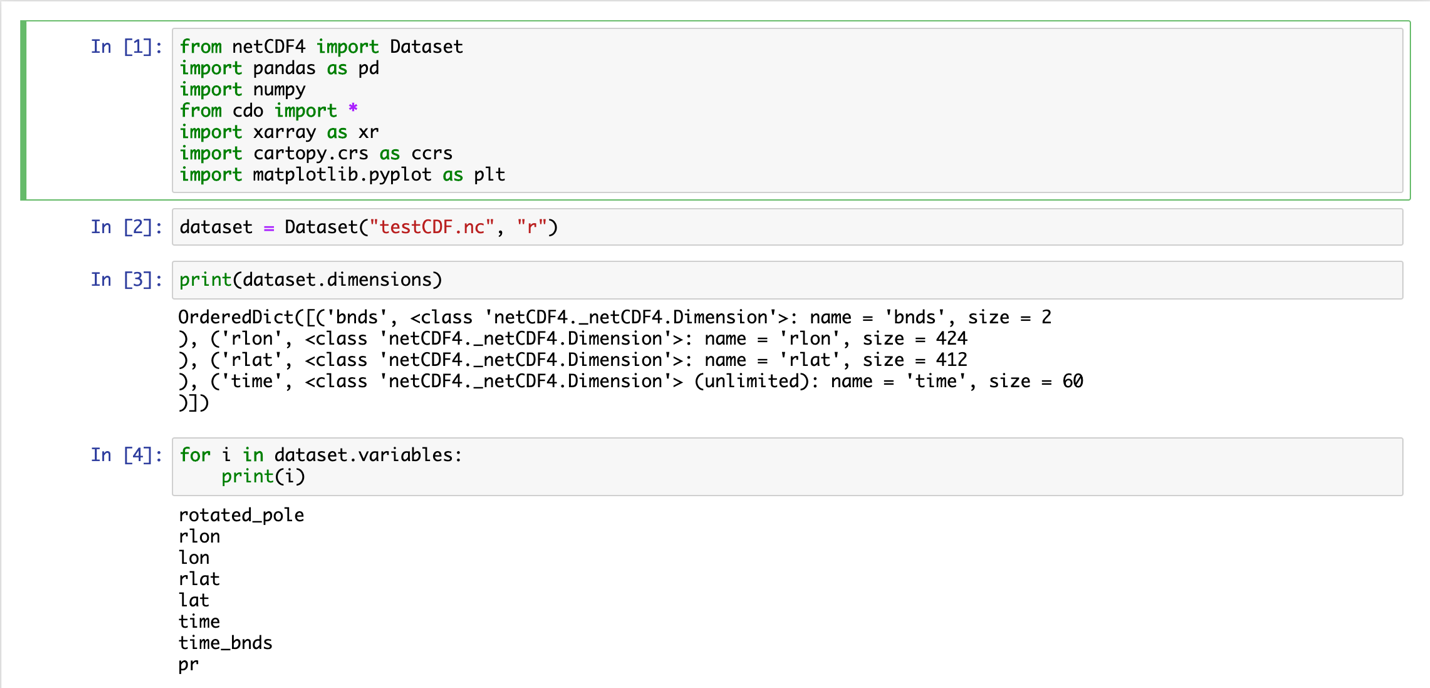
*Utolsó simítások a CSV fájlon*

* + CSV fájl betöltése pandas DataFrame-be (tabuláris adataszerkezet)



*CSV fájl (GRIB-ből átalakítva) betöltése memóriába Jupyter Notebook-ban*

* ESGF Node



*Az ESGF Node Data Cart-ból letöltött és kicsomagolt netCDF fájlok betöltése memóriába Jupyter notebook-ban*

***Adatvizualizáció, az adathalmazok felfedezése, „data wrangling”***

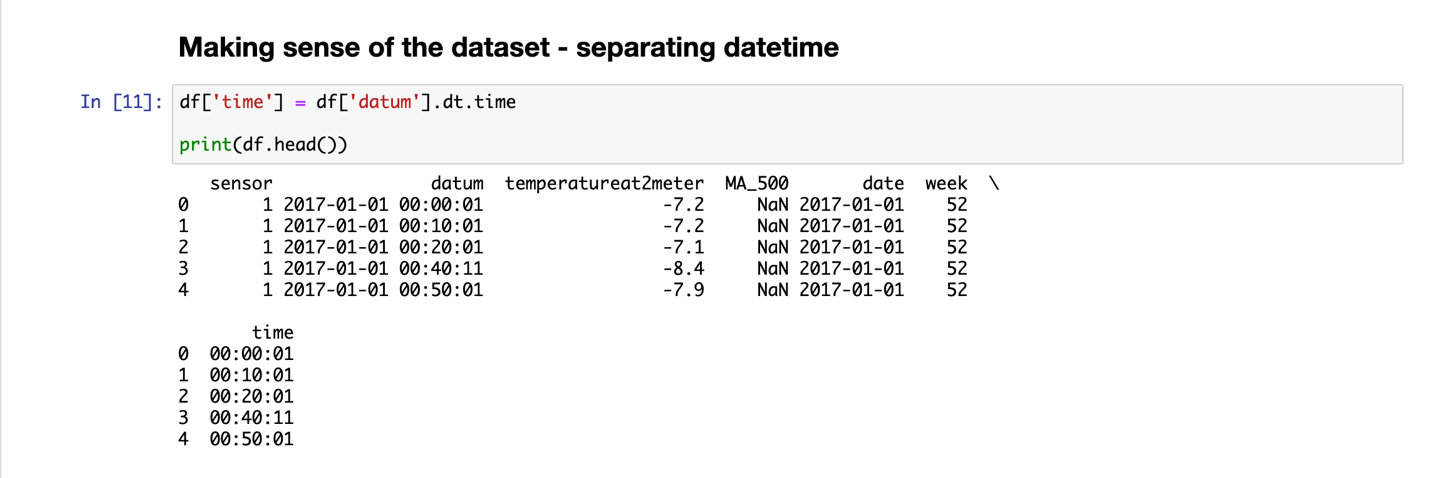
* SensorHUB



*A memóriába betöltött SensorHUB adatok megjelenítése*

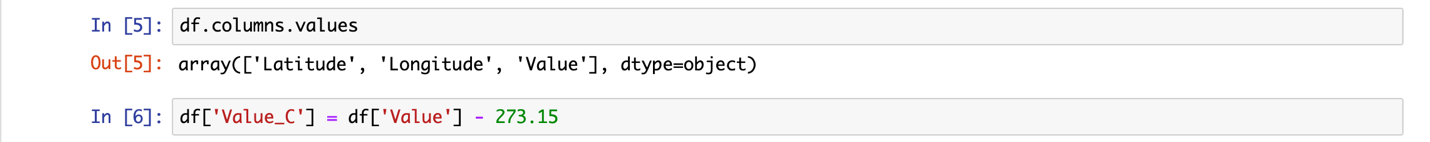
**

*A hibás adatok megjelölése NaN (null) értékkel*

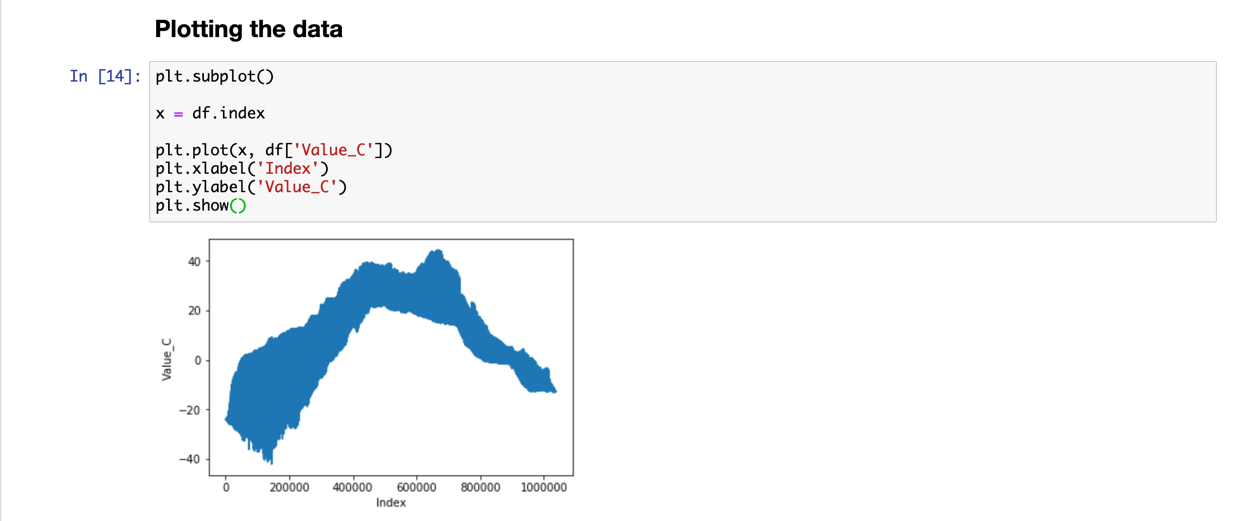
**

*Attribútum halmaz bővítése az idő adatok felbontásával*

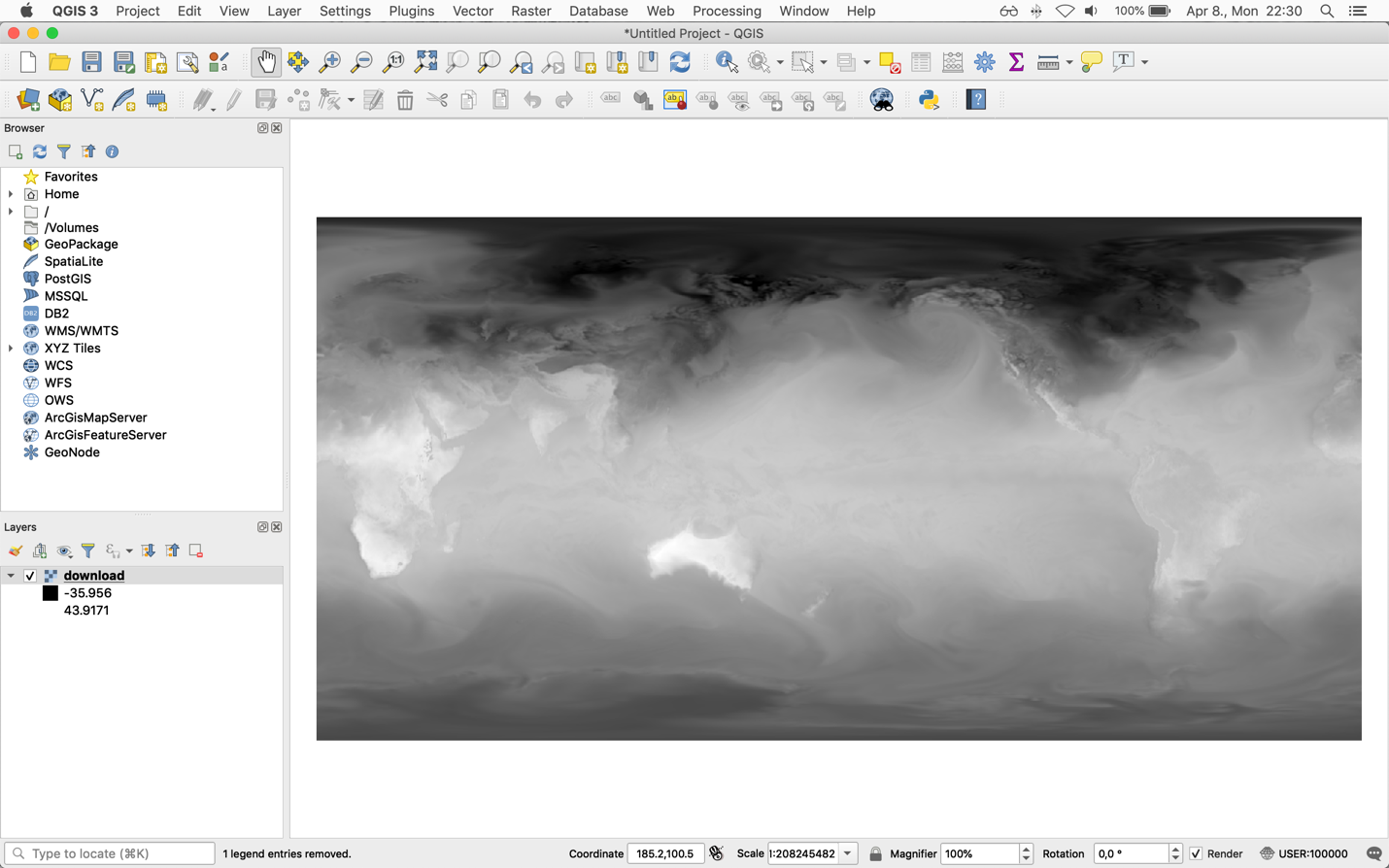
* Copernicus Climate Data Store



*A GRIB fájl hőmérséklet adatainak átváltása Kelvinből Celsiusba*

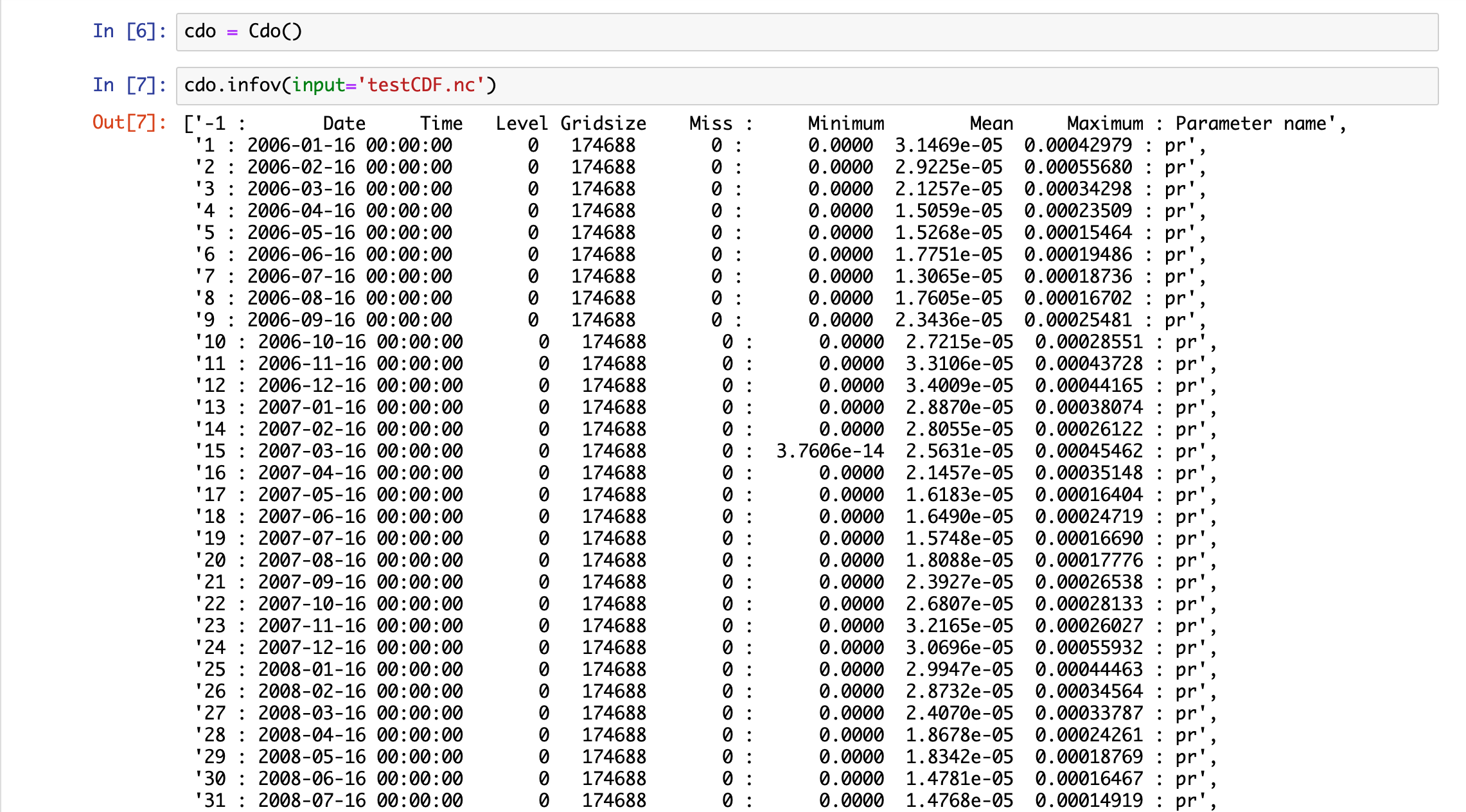
**

*A fájl tartalmának ábrázolása*

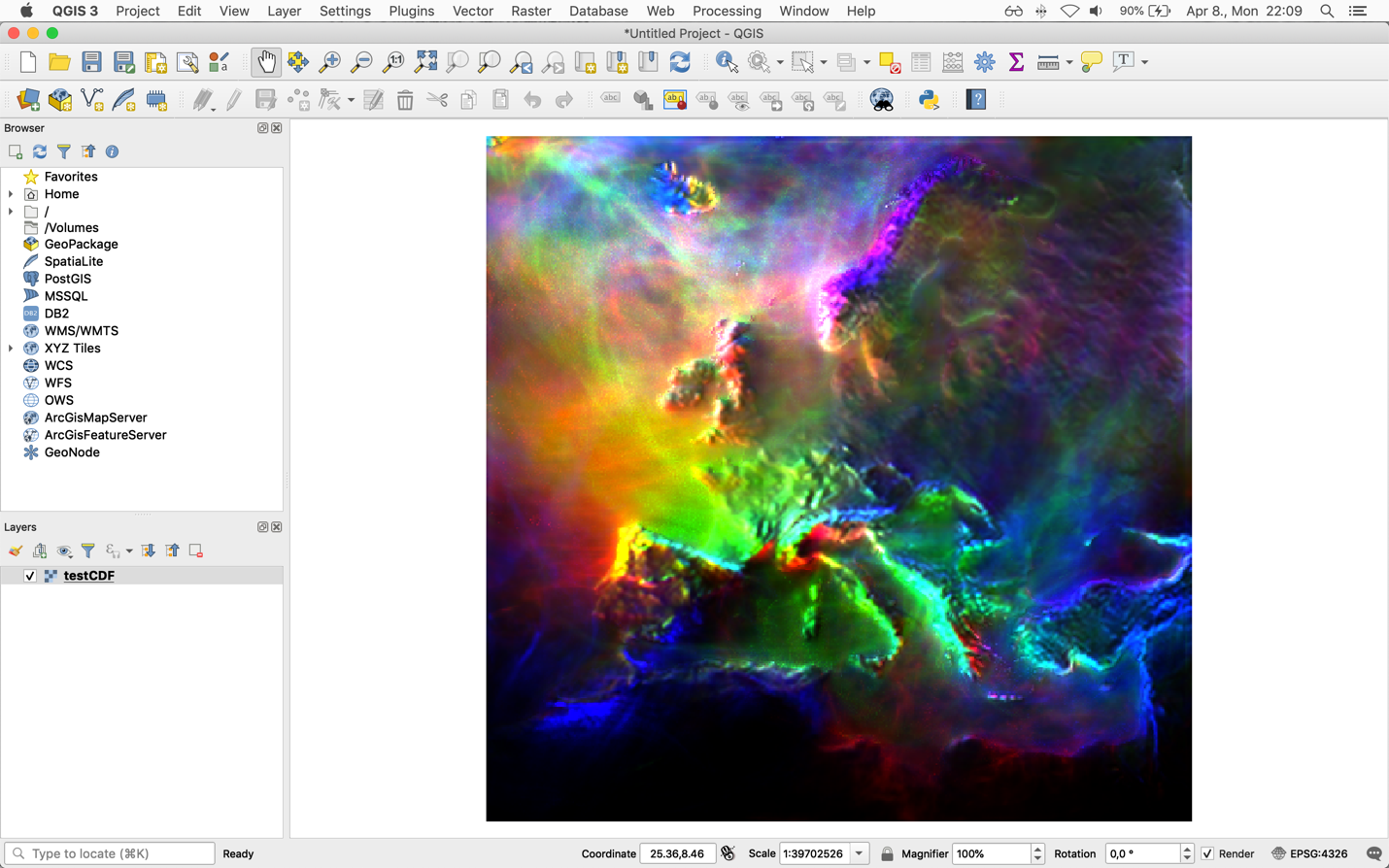
**

*GRIB fájl megjelenítése a QGIS alkalmazásban*

* ESGF Node



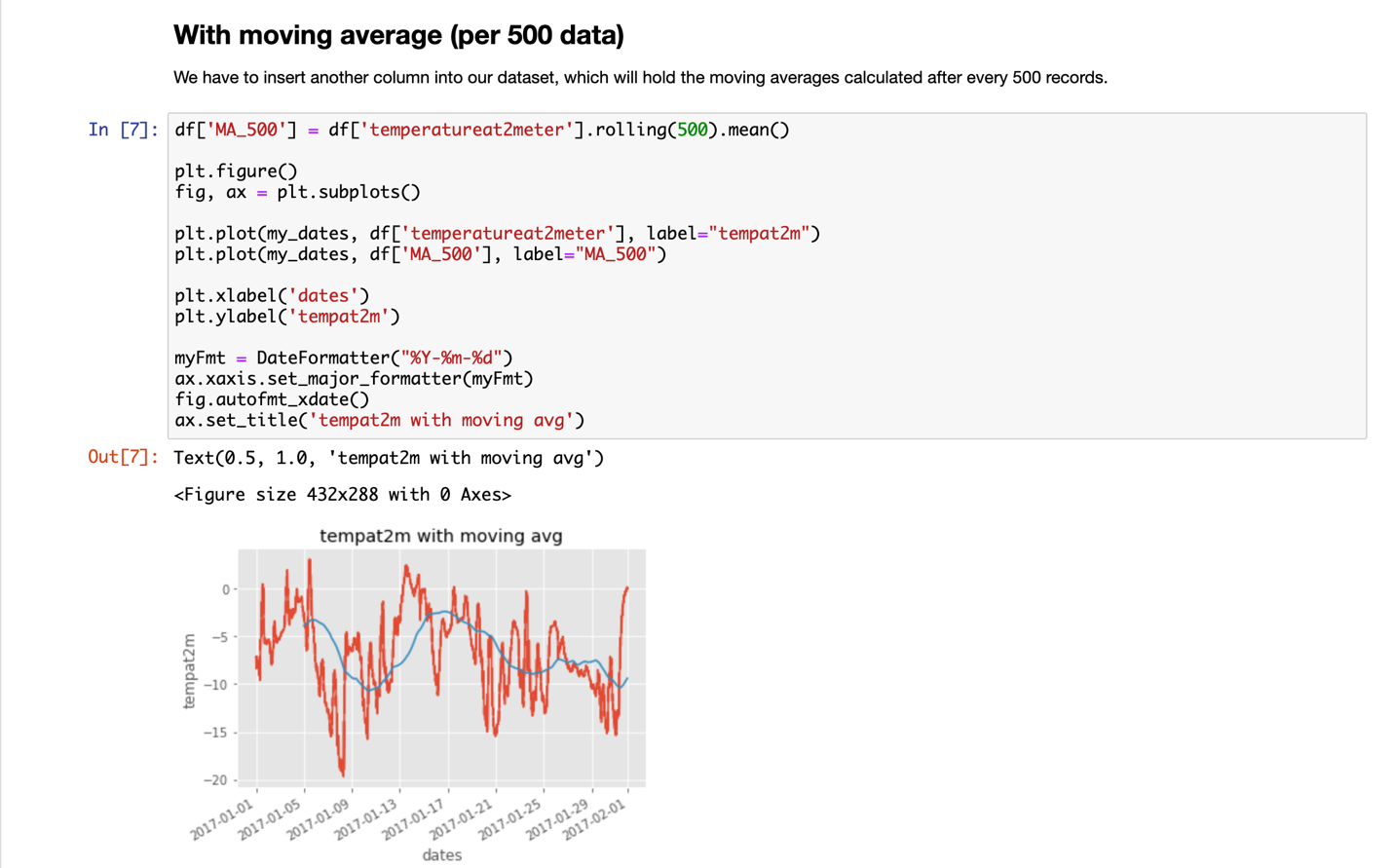
*CDO Python könyvtár használata netCDF fájl részleteinek feltárására*

**

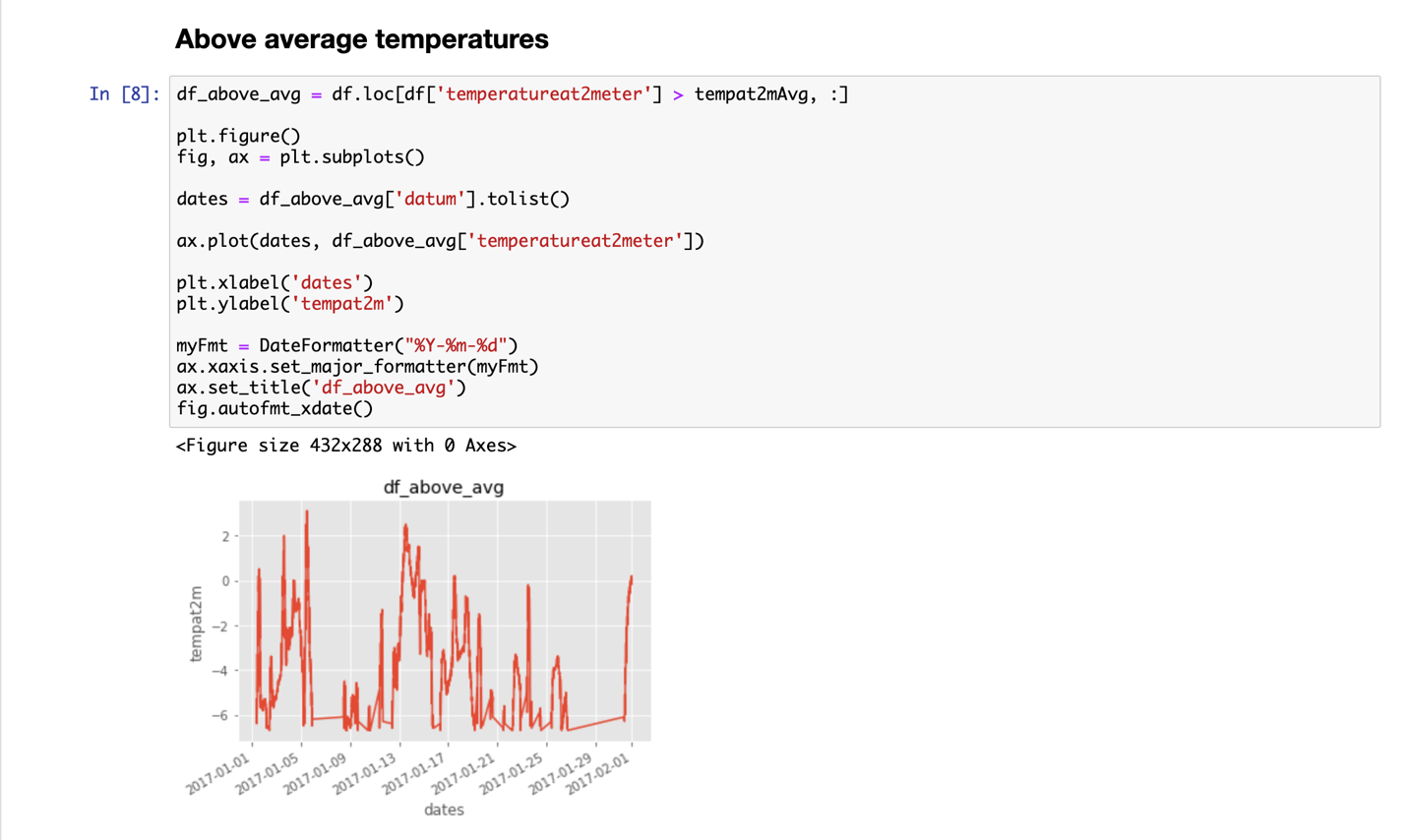
*netCDF fájl megjelenítése a QGIS programban*

***Adatelemzés, értelmezés***

* SensorHUB
  + Itt első körben egy 2017 januárjában, az első szenzor által mért hőmérsékleti adatokat tartalmazó halmazt vizsgálunk. Az adathalmaz relatíve kicsi, az eredmények gyors kinyerése (és értelmezhetőségének) érdekében.

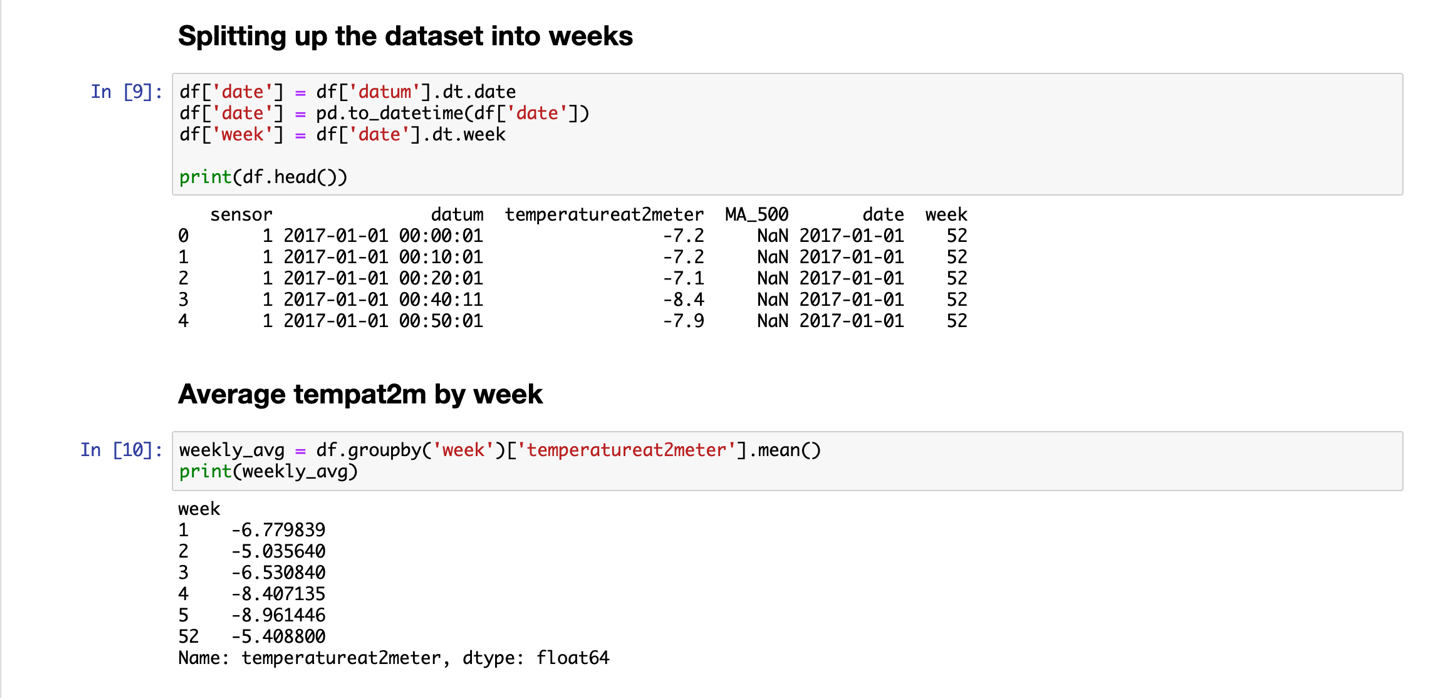


*Mozgóátlag számolása és megjelenítése*

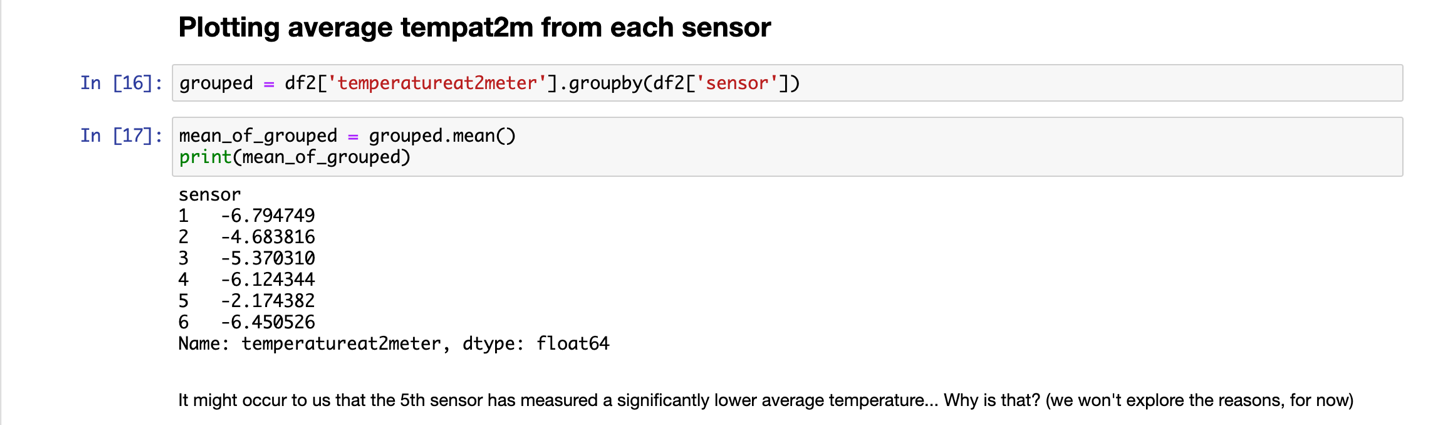
**

*A mintaátlag feletti értékek számolása és megjelenítése*

Innentől kezdve bármilyen statisztikai, adatelemző metódust, függvényt alkalmazva nyerhetünk ki információt az adott adathalmazból. A következőkben ezekre láthatunk néhány példát.



*Az adathalmaz attribútumainak bővítése egy heti bontással, továbbá a heti átlaghőmérsékletek kalkulálása*

**

*Bővítettük az adathalmazt az összes többi szenzor adataival és szenzorokra vetített bontásban számoltunk átlaghőmérsékletet – vegyük észre, hogy az 5. szenzorhoz tartozó érték különösen alacsony*

**

*Ezt egyszerűbben levonhatjuk, mint következtetés, hogyha grafikus módon jelenítjük meg a számolt értékeket. De mi lehet ennek az oka?*

Próbáljuk meg az előzőekhez hasonlóan vizualizálni az 5. szenzor által mért adatokat!

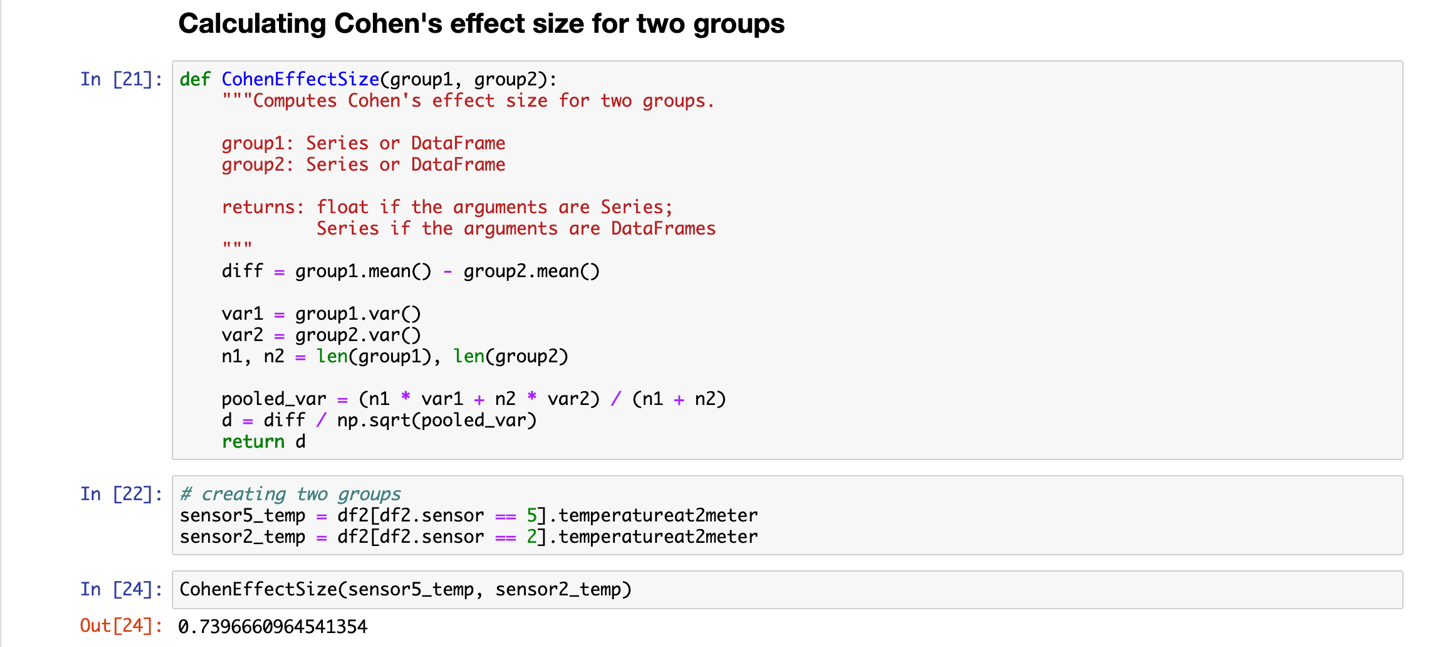


*Szépen kivehető, hogy az 5. szenzortól a vizsgált időszak egy jelentős hányadában nem érkeztek adatok!*

Így nem meglepő, hogy a számolt átlaghőmérséklet ennyire eltért a többi szenzor által mért adatokból számolt átlagoktól.

További elemzések:

* Cohen-féle hatásnagyság (Cohen effect size) – Két minta várható értékének különbsége szórásban kifejezve, megadja a standardizált különbséget két átlag között. 0,8-as érték felett a hatásnagyság nagy.



*Jelentős a hatásnagyság az egyes és a kettes szenzorok által mért adatok között*

* Pearson-féle korrelációs együttható

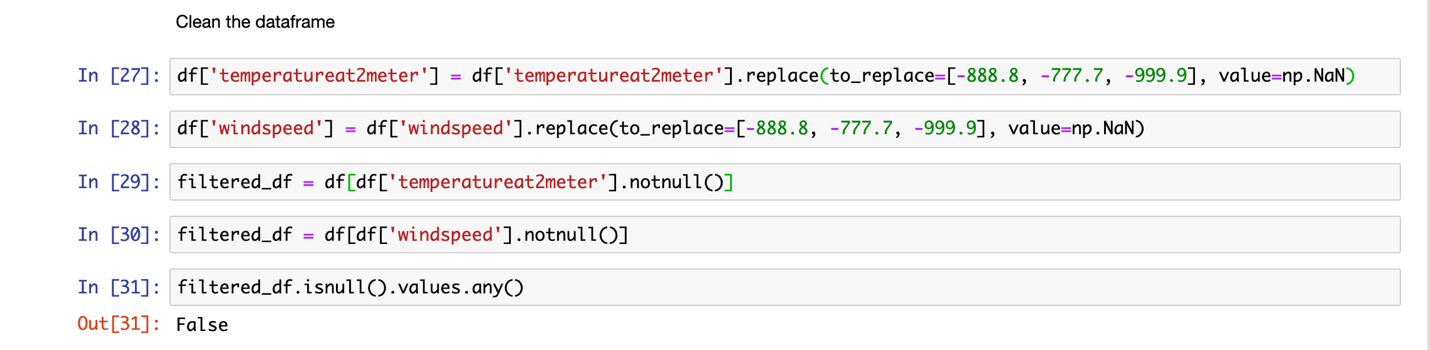


*Pearson-féle korrelációs együttható számolása új adathalmazra, több attribútum között. Úgy tűnik, hogy jó közelítéssel lineáris a kapcsolat a vizsgált attribútumok között.*

* Lineáris regresszió
  + Új adathalmaz: januári hőmérséklet adatok az első szenzortól
  + Elvek: adathalmaz felbontása tréning és teszt részhalmazokra (az adathalmaz egy részén „betanítjuk” a modellt, a másik részén – teszt – pedig ellenőrizzük a kész modellt), a kapott modell megjelenítése



*Adatok beszerzése a már ismert módon*

**

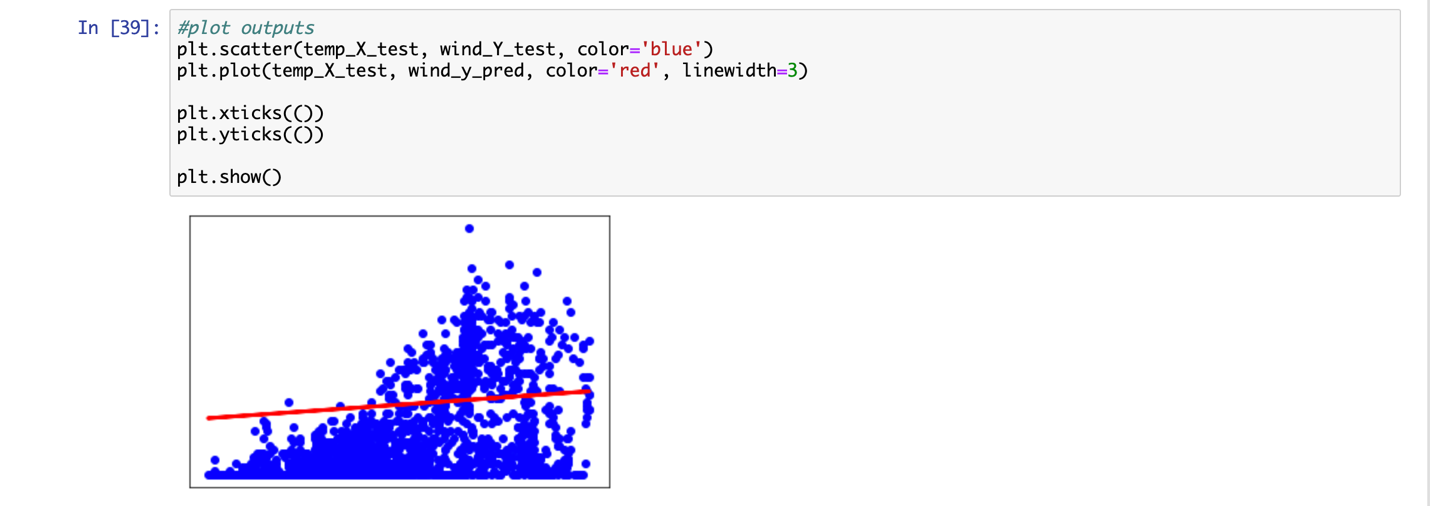
*Adathalmazból hibás adatok kiszűrése*

**

*Adathalmaz felosztása training és test részhalmazokra*

**

*Regressziós modell példányosítása, felkészítése a training részhalmazon és előrejelzések kalkulálása a test részhalmazon*

**

*A kapott regressziós megjelenítése az aktuális adatok felett*

Jól látható, hogy a regressziós egyenes nem pontosan becsüli meg a teszt részhalmaz értékeit, attól függetlenül, hogy az általános trendet jól „megtanulta”. Ennek több oka is van:

1. Kis méretű adathalmaz
2. Az adathalmaz rossz felosztása tréning és teszt részhalmazokra

Ezekből is következik, hogy a megfelelő adathalmaz kiválasztása és a részhalmazok kijelölése regressziós problémáknál kritikus.

* Copernicus Data Store, ESGF Node
  + Ezen masszív adatforrások további kutatást igényelnek
* Az elért eredmények:
  + GRIB és netCDF formátumú fájlok összegyűjtése
  + Értelmezése (parancssoros – ecCodes, CDO - vagy programozói eszközökkel)
  + Megjelenítése (Python, QGIS)
* Jelenleg fejlesztés alatt:
  + netCDF fájlokból indexek számolása a Kárpát-medencére: Módosított Ellenberg index, Forestry Aridity Index
  + Ezen indexek alapján az egyes klíma modellek összehasonlítása

**A gépi tanulás alapjai**

*„If it’s in Python then it’s probably machine learning. On the other hand, if it’s in PowerPoint then it’s AI.”*

A gépi tanulás definíciója: a számítógépek azon képessége, hogy anélkül képesek tanulni, hogy erre külön meg lettek volna tanítva. [2]

Különböző ML (Machine Learning) rendszerek típusai, kategorizálási lehetőségek:

* Felügyelt, nem felügyelt, részben felügyelt és megerősített tanulás
  + Felügyelt tanulás esetén a rendszerbe bevitt adatok tartalmazzák a kívánt eredményeket (labels). Például: osztályozás (classification).
  + Regressziós problémák: adott értékek becslése, előrejelzése. (Lásd: lineáris regresszió az adatelemző szekcióban). Fajták: neurális hálók, döntési fák, stb.
  + Nem felügyelt tanulás: nem címkézett adat, azaz a kezdeti adathalmaz nem tartalmazza a kívánt eredményeket. Fajták: Klaszterezés (clustering), dimenzió redukció (dimension reduction), anomális detektálás, stb.
  + Részben felügyelt rendszerek: felügyelt és nem felügyelt megoldások hibridjei
  + Megerősített tanulás (Reinforcement Learning): a rendszer megfigyeli a környezetét és ennek hatására javaslatokat tesz, ezen javaslatok vagy helyesek vagy nem, ennek függvényében a rendszert vagy jutalmazzuk vagy nem. Például: DeepMind AlphaGo
* Online és batch tanulás („on the fly” tanulás és előre elkészített, konstans adathalmazon végzett tanulás)
  + Batch tanulás: offline tanulás, a rendszer a működése közben képtelen új adatot magába fogadni, az induláskor az elérhető összes adatot be kell tölteni
  + Online tanulás: szekvenciális adatbetöltés lehetősége, korlátozott memóriával rendelkező rendszerek estében preferált megoldás
* Példány alapú és modell alapú tanulás
  + Példány alapú tanulás: adott adatok között hasonlóságokat fedez fel a rendszer és ezen hasonlóságok alapján dönt egy még nem látott adatról
  + Modell alapú tanulás: attribútumhalmazok egybefűzésével tár fel kapcsolatot az egyes attribútumok között

Természetesen a gyakorlatban használatos ML rendszerek kategorizálása nem ennyire egyértelmű, hiszen ezen rendszerek legtöbbször a felsorolt megoldások keverékei.

***Gépi tanulás eszközök***

Mára rengeteg ML könyvtár, megoldás és keretrendszer létezik, álljon itt egy felsorolás a legismertebbekről. Ezen eszközök mindegyike nyílt forráskódú.

* Scikit-Learn
  + Python könyvtár
  + Könnyű használhatóság (lásd: lineáris regresszió az adatelemző szekcióban)
* Google TensorFlow
  + Komplexebb könvytár, mely támogatja a párhuzamosított számításokat (akár videókártyás rendszerekben)
* Keras
  + Magas szintű API
  + Többek között TensorFlow támogatással
  + Deep Learning
* PyTorch
  + Python könyvtár
  + TensorFlow-hoz hasonlóan elosztott számítások támogatása