

# Tehtävä: Data-analytiikan tehtävä 1

---

Ohessa ensimmäisen tehtävän Colab-notebook. Saat sen auki seuraavasti:

- lataa tiedosto ao. linkistä koneellesi
- avaa selaimessa Google Drive ([drive.google.com](https://drive.google.com))
- etsi hakemisto "colab notebooks"
- lataa tiedosto sinne joko raahaamalla ja tiputtamalla tai menemällä ensin kansioon ja sitten New > File upload
- klikkaa tiedostoa ja valitse "open with colab" ylhäältä

Jos on ongelmia, niin katsotaan yhdessä ohjaus/opetuskerralla. Voit laittaa myös mailia tai viestiä.

## TEHTÄVÄ

Tehtävänä on siis tarkastella eri parametrien korrelaatioita retentio-muuttujaa vastaan:

- Kopioi viimeinen koodisolu
- Muuta ensimmäistä koodiriviä `gr = data['GRAMMAGE']` esim. näin; `pol = data['POLYMER_FLOW']`. Ole tarkkana kirjoittaessasi muuttujanimeä hakasulkujen sisään, nimen ympärillä täytyy olla hipsut ja nimen täytyy olla kirjoitettu täsmälleen samoin kuin se datan otsikkorivillä on.
- Muuta rivillä `plt.scatter(gr, ret)` muuttujan `gr` paikalle muuttuja `pol`
- Aja koodi, kuva pitäisi tulla näkyviin. Huom! Jotta viimeisen koodisolun ajo toimisi, sinun täytyy ajaa kaikki aiemmat solut ainakin kerran. Koodi ajetaan painamalla solun vasemmassa laidassa olevaa painiketta.
- Toista tämä samalla tavalla kymmenelle itse valitsemallesi muuttujalle
- Arvioi lopuksi (tekstisoluun), missä muuttujissa näit prosessihäiriön vaikutuksen.

## Tehtävä: Data-analytiikan tehtävä 2

---

Tehtävänkuvaus löytyy opetusvideosta 23.3. (dokumentit > linkki).

Tehtävän tarkoitus on opetella korrelaatiokartan käyttämistä. Tehtävässä vertaillaan paperikoneen normaalitilannetta ja ongelmatilannetta.

Tehtävän suoritus:

Vaihda dataksi joko normaalitilanne tai ongelmatilanne. Saat koodirivin pois käytöstä kirjoittamalla risuaidan # koodirivin eteen. Alla olevassa koodissa ongelmatilanteen data on poistettu käytöstä.

```
dat = data_normal #normaalitilanteen data
```

```
#dat = data_problem #ongelmatilanteen data
```

Aja korrelaatiokarttojen koodit ja tarkastele, miten korrelaatiot muuttuvat. Isommasta korrelaatiokartasta löydät helpommin muuttuvat parametrit, kun ajat ensin normaalitilanteen, otat kuvakaappauksen ja vertaat sitten ongelmatilanteen korrelaatiokarttaan. Jos löydät vahvoja/keskivahvoja korrelaatioita, tarkastele niitä lähemmin alemman korrelaatiokartan avulla.

Viimeisessä koodisolussa (sns.pairplot...) voit vaihtaa muuttujia kemiantila-, annostelu- tai ajettavuusparametreihin, ja ajaa koodin sitten normaalidatalla ja ongelmadatalla. Voit käyttää ylempää karttaa löytääksesi mielenkiintoisia riippuvuuksia. Voit vaikka kopioida koodisolun, niin alkuperäinen säilyy mallina.

Liitä selostukseen vertailukuvat ja lyhyesti kommentit siitä, mitä havaitsit.

## Tehtävä: Data-analytiikan tehtävä 4

---

Moi!

Tässä jälleen data-analytiikkaa teille. Ymmärrän, että teillä saattaa olla työkuormaa paljon, mutta kuten luennolla kerroin, jokainen harjoitus edistää kurssisuoritustanne merkittävästi.

Tehtävän kuvaus on kirjoitettu colab-dokumenttiin, joten lukekaa se tarkkaan. Kysykää, jos jokin on epäselvää.

Neljä pistettä on mahdollista saada tästä harjoituksesta.

# Tehtävä: Data-analytiikan tehtävä 5

---

Oheisessa Colab-tiedostossa on vesidataa kaivosprosessista. Data sisältää:

- jatkuvaa dataa (pitoisuudet, juoksutus, sadanta tms)
- luokittelumuuttujia, joiden otsikon perässä on "indeksi". Luokittelut on tehty siten, että pH, johtokyky, sadanta ja sulfaatti ovat jaettu korkeaan ja matalaan arvoväliin ko. jatkuvan muuttujan mediaanin perusteella. Siten esim. pH\_indeksin arvo 1 kuvaa matalaa pH:ta (pienempi kuin pH-mediaani) ja arvo kaksi kuvaa mediaania korkeampaa pH:ta.

Voit käyttää luokittelumuuttujia apuna tarkastellessasi pH:n, johtokyvyn, sadannan ja sulfaattipitoisuuden vaikutuksia esim. veden ainepitoisuuksiin.

Epäselvien termien selityksiä:

- pH on veden happamuus
- johtokyky on veden sähkönjohtavuus, joka kuvaa liuenneiden suolojen kokonaispitoisuutta
- kiintoaine on liukenematonta ainetta (sameus)
- kokonaiskovuus on veden kyky "puskuroida" pH-muutoksia; "kovalla" vedellä kyky on vähäinen ja "pehmeällä" suuri.
- juoksutus on vesimäärä, joka altaasta juoksutetaan vesistöön.

Datasta on tehty kaksi osadataa: dat sisältää kaiken datan, kun taas df-osajoukkoon voit sisällyttää haluamasi määrän sarakkeita otsikkojen perusteella. Käytä df-dataa kun teet esim. korrelaatiokartan sns.pairplotilla.

Sinun tulisi käyttää ainakin seuraavia menetelmiä (esimerkkikuvat liitteenä):

- Korrelaatiokartta (sns.heatmap) kaikelle datalle, alustava analyysi
- Box-plot ja klusterointi indeksoitujen muuttujien perusteella, tarkempi analyysi
- sns.pairplot klusteroimalla dataa indeksoitujen muuttujien perusteella

Tavoittena on siis selvittää, miten indeksoidut muuttujat vaikuttavat pitoisuuksiin ja sitä kautta kaivoksen ympäristökuormitukseen. Tarkastele myös, miten sadanta vaikuttaa juoksutukseen. Käyn tämän vielä läpi ensi maanantaina, niin ei jää epäselvyyksiä.

Kaikki vaadittavat menetelmät ja koodinpätkät voi kopioida aiemmista tehtävistä. Joudut kuitenkin muuttamaan datasarakkeiden nimiä.

Tästä voi saada paljon pisteitä, mutta maksimipisteet saadaksesi tee huolella ottaen huomioon kaikki yllä mainitut asiat.

# Tehtävä: Data-analytiikan tehtävä 6

Ehdin sittenkin tehdä aikasarja-analytiikan tehtävän tälle viikolle...

En kuitenkaan tehnyt tähän ARIMA-mallin esimerkkiä, sen jätän tuleviin vapaaehtoiisiin tehtäviin. Nyt on tarjolla FFT-analyyseja, joissa on riittävästi tehtävää.

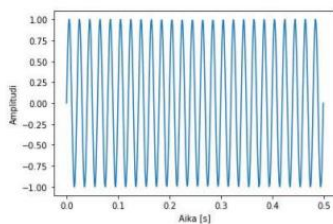
Liitteenä olevassa kuvassa on erilaisia aaltomuotoja. Kaikki ovat systeemitekniiikan kannalta relevantteja erilaisten sinimuotoisten signaalien yhdistelmiä.

Analysoi aaltomuodot FFT:llä, ja päättelä, mikä FFT-spektrikuvaaja vastaa mitäkin aikatasossa esitettyä aaltomuotoa. Koodissa on esitetty, missä voit vaihtaa kulloinkin analysoitavan aaltomuodon.

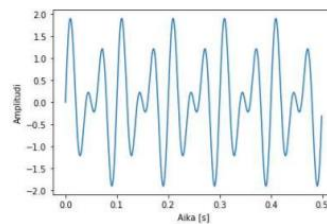
Voit myös itse koodata koodirivin, jolla voit esittää aaltomuodot kuvaajana. Helpottaa päättelyä, eikä ole aiempien esimerkkien perusteella vaikeaa.

Vastaa selkkarissa myös seuraaviin kysymyksiin:

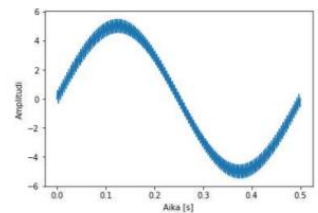
1. Mitä taajuuksia jokaisesta aaltomuodosta löytyy (yksi tai useampia)?
2. Yksi aaltomuodoista on ns. chirp-signaali, jossa taajuus muuttuu lineaarisesti alkutaajuudesta lopputaajuuteen. Mitkä ovat kyseisen signaalin alku- ja lopputaajuudet FFT:n perusteella?
3. Aaltomuodoista löytyy kaksi signaalia, joissa on sama taajuus, mutta toisessa amplitudi muuttuu lineaarisesti. Miten tämä näkyy FFT-spektrissä verrattuna signaaliin, jossa on vakioamplitudi?



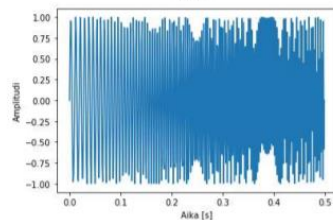
1



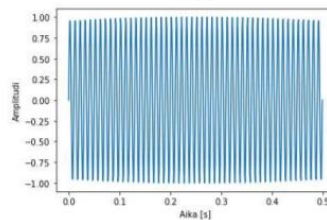
4



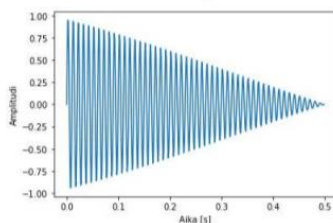
7



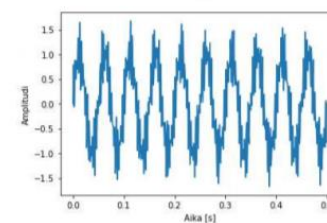
2



5



3



6



# Tehtävä: Data-analytiikan tehtävä 7

---

Tehtävän esittely luentotallenteessa 22.4.

Tämä tehtävä on vapaaehtoinen, mutta tästä saa tenttipisteitä.

Ohjeet;

Tarkasteltava kohdemuuttuja on kastepiste (DP) ja tarkoituksena on siis tarkastella, miten ilmankosteus ja lämpötilamittaukset (huoneilma ja LVI-järjestelmä) vaikuttavat kastepisteeseen.

Täydennä **korrelaatiokartta** kaikilla käytettävissä olevilla parametreilla ja katso, mitkä muuttujat korreloivat parhaiten kastepisteen kanssa.

Täydennä myös **regressiomalliin** Y-muuttujat (kaikki paitsi DP) ta tarkastele, mitkä muuttujat selittävät regressiomallia parhaiten.

Muuta myös tietokantahakua siten, että lisäät siihen hakuehtoja (Y-muuttujille) ja/tai rivimäärille siten, että saat siivottua datasta äkillisiä muutoksia pois. Vertaa tuloksia analyysiin, jossa kaikki data on mukana. Näkökö muutoksia korrelaatioissa ja regressiomallissa?

Voisitte palauttaa pdf-tulosteen, josta lähinnä näen, että olette mainitulla tavalla tehneet. Vaihtoehtoisesti voitte palauttaa koko tehtävän vasta sitten, kun olen julkaissut ja esitellyt tehtävän jatko-osan. Jatketaan siis tehtävää vielä ensi maanantaina siten, että lisään tähän koneoppimisesimerkin ja aikasarjamallin.