

DATA-ANALYYSI PERUSTEET

Osaamispaketti – AIMLearning-projekti

Teemu Voutilainen

TKI-asiantuntija, Centria-ammattikorkeakoulu



**Euroopan unionin
osarahoittama**



Elinvoimakeskus

Adobe Stock

DATA-ANALYYSI – PERUSTEET

Itseopiskeluopas vasta-alkajille.

Tämä opas on tuotettu suuren kielimallin avulla.

Ennen kuin aloitat

Tämä opas opettaa data-analyysin ydinajatuksia alusta alkaen. Matematiikkaa ei tarvita koulun peruslaskutoimituksia enempää, eikä seuraaminen vaadi mitään ohjelmistoa – taulukkolaskentaohjelmalla (Excel, Google Sheets tai LibreOffice) voit kuitenkin kokeilla asioita itse, mitä suositellaan vahvasti.

Käy luvut läpi järjestyksessä; jokainen rakentuu edellisen varaan. Oppaassa on pieniä **Harjoitus**-laatikoita. Tee ne – analyysistä lukeminen on kuin uimisesta lukemista. Vastaukset ovat oppaan lopussa.

Käytämme läpi koko oppaan yhtä esimerkkiä: pientä kioskia, joka myy juomia ja välipaloja puistossa. Tässä sen kymmenen päivän myynnit:

Päivä	Viikonpäivä	Myydyt juomat	Lämpötila (°C)	Sade
1	maanantai	40	15	ei
2	tiistai	48	16	ei
3	keskiviikko	56	18	ei
4	torstai	44	16	kyllä
5	perjantai	60	19	ei
6	lauantai	82	20	ei
7	sunnuntai	74	19	ei
8	maanantai	30	13	kyllä
9	tiistai	46	16	ei
10	keskiviikko	180	17	ei

Pidä sormi tällä taulukolla – palaamme siihen monta kertaa. (Kyllä, päivä 10 näyttää oudolta. Se on tarkoitus.)

1. MITÄ DATA-ANALYYSI OIKEASTAAN ON

Data-analyysi on **säännönmukaisuuksien etsimistä muistiin merkityistä havainnoista ja niiden käyttämistä parempien päätösten tekemiseen**. Siinä on koko ala yhdessä virkkeessä. Kaikki muu – työkalut, tilastotiede, kaaviot, koneoppiminen – on koneistoa tuon virkkeen palveluksessa.

Huomaa, mitä virke ei sano. Se ei mainitse tietokoneita, ohjelmointia eikä matematiikkaa sen enempää kuin jo osaat. Ihmiset analysoivat dataa tuhansia vuosia ennen kuin mitään niistä oli olemassa. Muinaisen Egyptin papit huomasivat, että Sirius-tähti nousi näkyviin juuri ennen aamunkoittoa samaan aikaan, kun Niili alkoi tulvia – havainto, joka muistiin merkittynä ja sukupolvien yli toistettuna antoi heille kyvyn ennustaa koko sivilisaatiota ruokkinut tulva (Encyclopaedia Britannica). Babylonian tähtitieteilijät jalostivat vuosisatojen taivaanseurannan hämmästyttävän tarkoiksi kalente-

reiksi ja tähtitieteellisiksi taulukoiksi (Neugebauer 1969, 97–144), ja mayat rakensivat kalenterijärjestelmän, jolla saattoi laskea päiviä kauas menneisyyteen ja tulevaisuuteen ja ennustaa tähtitieteellisiä tapahtumia, kuten Venuksen liikkeitä (Fraknoi, Morrison & Wolff 2022, luku 4.4). Metsästäjä, joka tietää, missä laaksossa peurat viihtyvät ensimmäisten pakkasten jälkeen, tekee data-analyysiä. Niin teet sinäkin, kun huomaat bussisi olevan perjantaisin aina myöhässä ja lähdet kotoa aikaisemmin.

Kaikki nämä noudattavat samaa silmukkaa, ja niin noudattaa kaikki tässä oppaassa:

havainnoi → merkitse muistiin → etsi säännönmukaisuus → ennusta → päättä

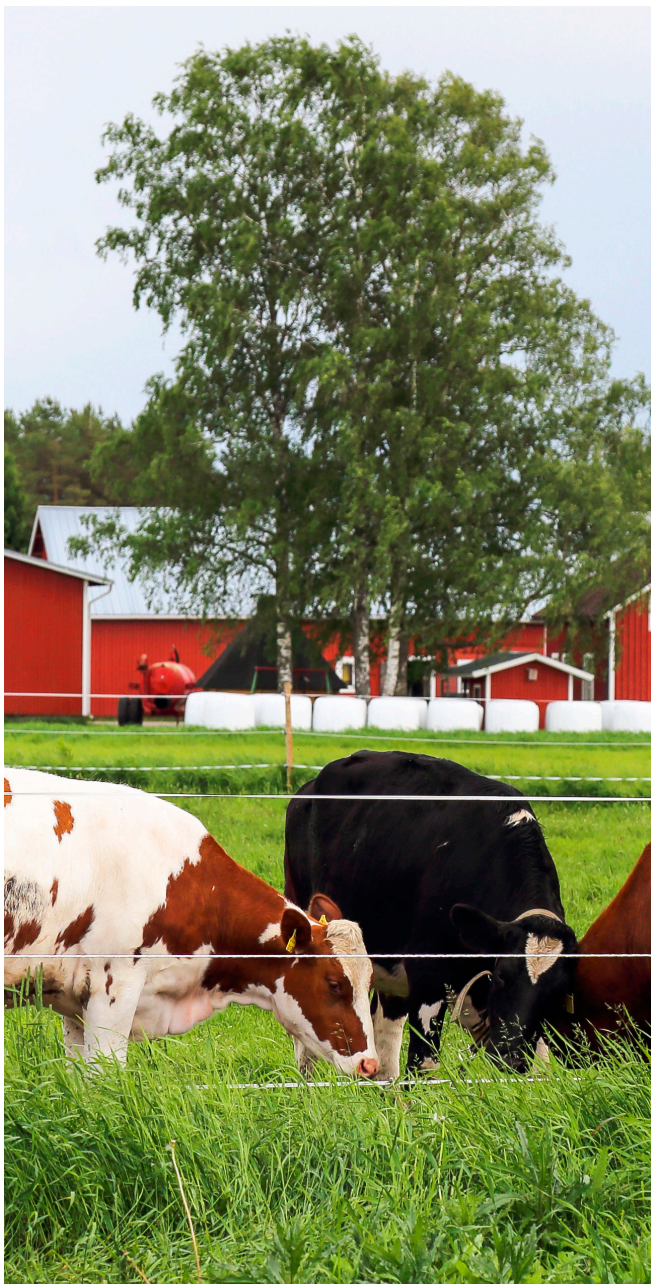
Miksi siitä tuli ammatti

Jos ihmiset ovat aina tehneet tätä, miksi ”data-analyytikko” on yhtäkkiä ammattinimike? Kaksi asiaa muuttui. Ensinnäkin tallennetun datan määrä räjähti: jokainen ostos, klikkaus, anturilukema ja bussimatka jättää nyt jäljen. Toiseksi tietokoneet

tekivät sen käsittelystä halpaa. Seitsemäntoista tuhannen rivin aineisto – nykymittapuulla vähäpätöinen – on jo kaukana siitä, minkä kukaan jaksaa silmäillä läpi.



Adobe Stock



Adobe Stock

Tämä siirsi asiantuntemuksen paikkaa. Egyptin pappi oli Niilin asiantuntija, joka sattui käyttämään dataa. Nykyanalyytikko on menetelmien asiantuntija, joka voi liikkua kaupunkipyörien, sairaalajonon ja pelitalouksien välillä, koska menetelmät

ovat kaikkialla samat. Aihepiirin tuntemus (**substanssiosaaminen**) on yhä valtavan tärkeää – näemme luvussa 6 tarkalleen miksi – mutta nykyään se jakautuu analyytikon ja hänen yhteistyökumppaniensa kesken.

Minkä tahansa analyysin viisi vaihetta

Käytännössä analyysiprojekti etenee viiden vaiheen läpi. Teet ne kaikki pienoiskoossa tässä opissa.

1. **Kerää** – hanki data: mittaa se, lataa se, vie se ulos jostakin järjestelmästä.
2. **Siivoa** – korjaa sotkut: lyöntivirheet, puuttuvat arvot, kaksoiskappaleet, mahdottomat luvut. Analyytikot vitsailevat, että tämä on 80 % työstä (Wickham 2014, 1 [Dasu & Johnson 2003]). Se on vain puoliksi vitsi.
3. **Tutki** – laske tiivistelmiä, piirrä kaavioita, metsästä säännönmukaisuuksia.
4. **Päättele** – testaa, ovatko löydökset todellisia, ja päättä, mitä ne merkitsevät.
5. **Viesti** – selitä löydös niin, että joku voi toimia sen pohjalta. Analyysi, jota kukaan ei ymmärrä, ei muuta mitään.

2. DATAN YMMÄRTÄMINEN

Ennen kuin dataa voi analysoida, sitä on osattava lukea. Lähes kaikki data, jota tulet kohtaamaan, on järjestetty taulukoksi, ja taulukoilla on kielioppinsa.

Jokainen **rivi** on yksi havainto: yksi päivä, yksi asiakas, yksi pelimatsi, yksi mittaus. Jokainen **sarake** on yksi muuttuja: jotakin, joka on mitattu jokaisesta havainnosta. Kioskitaulukossa jokainen rivi on päivä, ja neljä muuttujaa ovat viikonpäivä, myydyt juomat, lämpötila ja sade. Tätä ”rivit ovat asioita, sarakkeet asioiden ominaisuuksia” -järjestystä kutsutaan **siistiksi dataksi** (engl. tidy data), ja sotkuisen datan saattaminen tähän muotoon on suuri osa todellista analyysityötä (Wickham 2014, 4–5).



Adobe Stock

Muuttujien tyypit

Kaikki sarakkeet eivät ole samanlaisia, ja tyyppi määrää, mitä sarakkeella voi mielekkäästi tehdä.

Numeeriset muuttujat ovat määriä: lämpötila, myydyt juomat, hinta, ikä. Niitä voi laskea yhteen ja niistä voi laskea keskiarvon. Osa on jatkuvia (lämpötila voi olla 16,4) ja osa diskreettejä lukumääriä (16,4 juomaa ei voi myydä).

Kategoriset muuttujat ovat nimilappuja: viikonpäivä, kaupunki, lempigenre. Niiden keskiarvo on merkityksetön – ”keskimääräistä viikonpäivää” ei ole – mutta niitä voi laskea ja niiden mukaan voi ryhmitellä, mikä osoittautuu yhdeksi analyysin voimakkaimmista liikkeistä.

Järjestysasteikolliset muuttujat ovat kategorioita, joilla on järjestys: kyselyvastaukset huono / kohtalainen / hyvä tai vaatekoot S / M / L. Järjestys kantaa tietoa, mutta askelten välit eivät välttämättä ole yhtä suuria.

Totuusarvoiset (kyllä/ei) muuttujat, kuten Sade-sarakkeemme, ovat kategorioita, joilla on tasan kaksi arvoa. Ne kirjoitetaan usein muodossa 1 ja 0, mikä mahdollistaa näppärän tempun: 1/0-sarakkeen keskiarvo on kyllä-vastausten osuus. Jos kymmenestä päivästä kolme sateista on koodattu ykkösiksi, keskiarvo on 0,3 – eli 30 % päivistä oli sateisia.

Päivämäärät ja kellonajat ansaitsevat kunnioitusta: ne näyttävät yksinkertaisilta mutta kätkevät sisäänsä rakennetta (viikonpäivä, kuukausi, vuodenaika, pyhäpäivä), joka usein selittää enemmän kuin mikään muu muuttuja.

Vielä yksi erottelu, jonka kuulet jatkuvasti: **perusjoukko** on kaikki, mistä olet kiinnostunut (kaikki päivät, joina kioskia on koskaan auki); **otos** on se osa, joka sinulla on (nämä kymmenen päivää). Kaikki, mitä laskemme, kuvaa otosta, ja hiljaa toivomme sen muistuttavan perusjoukkoa. Mitä suurempi ja edustavampi otos, sitä turvallisempi tuo toivo. Kymmenen päivää on hyvin pieni otos – todellisia johtopäätöksiä varten haluttaisiin kokonainen sesonki tai enemmän.

HARJOITUS 1.

Luokittele kioskitaulukon neljä muuttujaa (viikonpäivä, myydyt juomat, lämpötila, sade) tyyppin mukaan. Ja vielä: voitko laskea ”keskimääräisen viikonpäivän”? Mitä viikonpäiväsarakeella sen sijaan voi tehdä?

3. DATAN KUVAILEMINEN: REHELLINEN TIIVISTELMÄ

Kenellekään ei voi näyttää 17 000:ta riviä, ja jo 10 rivin taulukkomme sisäistäminen vaatii ponnisteluja. Analyysin ensimmäinen tehtävä on **tiivistää**: pelkistää monta lukua muutamaan, jotka kertovat

tarinan. Klassiset tiivistelmät vastaavat kahteen kysymykseen – missä on keskikohta ja miten hajal-
laan arvot ovat?

Keskikohta: keskiarvo, mediaani ja moodi

Keskiarvo on summa jaettuna lukumäärällä. Myydylle juomille: $40+48+56+44+60+82+74+30+46+180 = 660$, ja $660 \div 10 = 66$ juomaa päivässä.

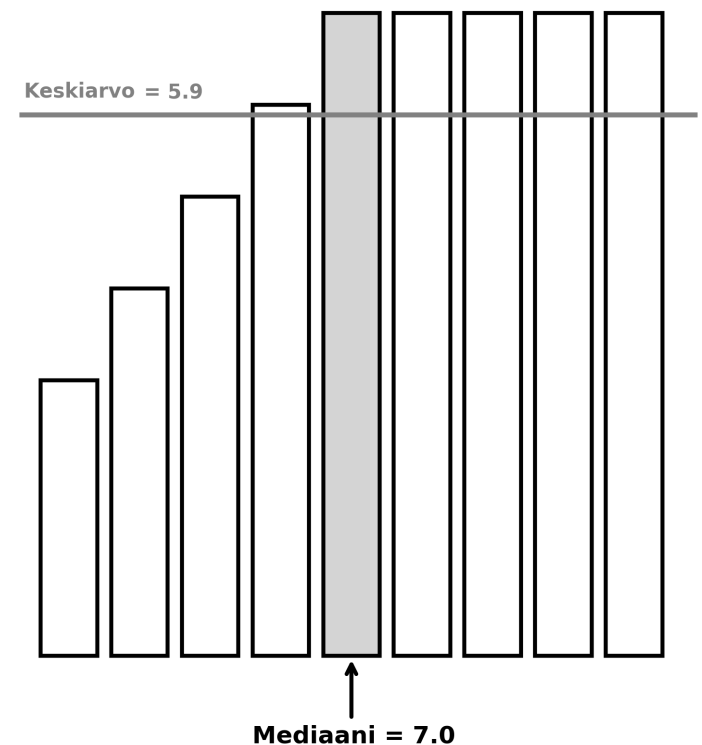
Pysähdy ja vertaa lukua taulukkoon. Kuusikymmentäkuusi on enemmän kuin kioskki myi kahdeksana päivänä kymmenestä. ”Tyypillisenä päivänä” se tuntuu väärältä – ja onkin, koska päivän 10 järjetön 180 veti keskiarvoa ylöspäin. Keskiarvo on demokraattinen huonoimmalla mahdollisella tavalla: jokainen arvo äänestää, ja ääriarvot huu-
tavat.

Mediaani on keskimäinen arvo, kun data järjestetään suuruusjärjestykseen. Järjestettynä juomamme ovat 30, 40, 44, 46, 48, 56, 60, 74, 82, 180; kymmenellä arvolla mediaani on 5. ja 6. arvon keskiarvo, $(48+56)/2 = 52$. Huomaa, miten tyyni mediaani on: vaihda 180:n tilalle 1 800, eikä

se liikahtaa lainkaan. Mediaani on puolustukseksi ääriarvoja vastaan.

Moodi eli tyyppi-arvo on yleisin arvo – hyödyllisin kategorioille (”yleisin sää oli: ei sadetta”).

Keskiarvon ja mediaanin ero on itsessään informaatiota. Kun luet, että jonkin ryhmän ”keskipalkka” on yllättävän korkea, tarkista mediaani: kourallinen valtavia palkkoja vetää keskiarvon ylös mediaanin pysyessä paikallaan. Vinon datan (palkat, asuntojen hinnat, somen seuraajamäärät) rehellinen kuvaus kertoo mediaanin.



Keskiarvo: Kuinka korkeita pylväät ovat keskimäärin?
Mediaani: Mikä pylväs on keskimäinen, ja kuinka korkea se on?

Hajonta: vaihteluväli ja keskihajonta

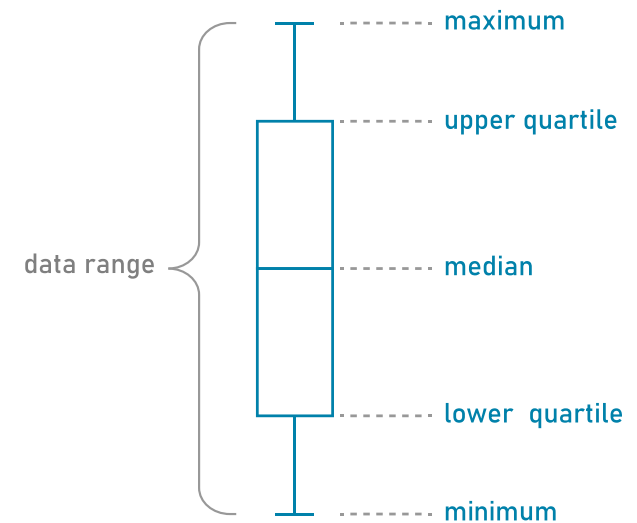
Kaksi kioskia voi kumpikin myydä keskimäärin 66 juomaa – toinen myy joka päivä 60–70, toinen heiluu 10:n ja 150:n välillä. Sama keskikohta, täysin erilainen liiketoiminta. Hajonnalla on väliä.

Vaihteluväli (maksimi – minimi) on karkein mittari: $180 - 30 = 150$ kioskillamme. Yksinkertainen, mutta riippuu kokonaan äärimmäisestä arvosta.

Keskihajonta mittaa karkeasti sitä, kuinka kaukana tyypillinen päivä on keskiarvosta. Sitä lasketaan harvoin käsin (jokaisessa työkalussa se on valmiina: Excelissä `=KESKHAJONTA.S(...)`), mutta sitä pitää osata lukea: keskiarvo 66 ja keskihajonta 43

(kioskimme) kuvaa rajusti vaihtelevaa liiketoimintaa; keskiarvo 66 ja keskihajonta 5 tylsän tasaista. Nyrkkisääntönä monissa aineistoissa noin kaksi kolmasosaa arvoista osuu yhden keskihajonnan päähän keskiarvosta.

Persentiilit yleistävät mediaanin: 25. persentiili on arvo, jonka alapuolelle jää neljäsosa datasta, 75. persentiilin alapuolelle kolme neljäsosaa. Mediaani on 50. persentiili. Yhdessä 25.–75. persentiilin väli (**kvartiiliväli**) kertoo, missä datan keskimäinen puolikas asuu – kioskillamme suunnilleen 45–70 juomaa.



Adobe Stock

Laatikko-jana -kaavio, joka esittää aineiston viiden luvun yhteenvedon: minimiarvon, alakvartiilin (Q1), mediaanin, yläkvartiilin (Q3) ja maksimiarvon.

HARJOITUS 2.

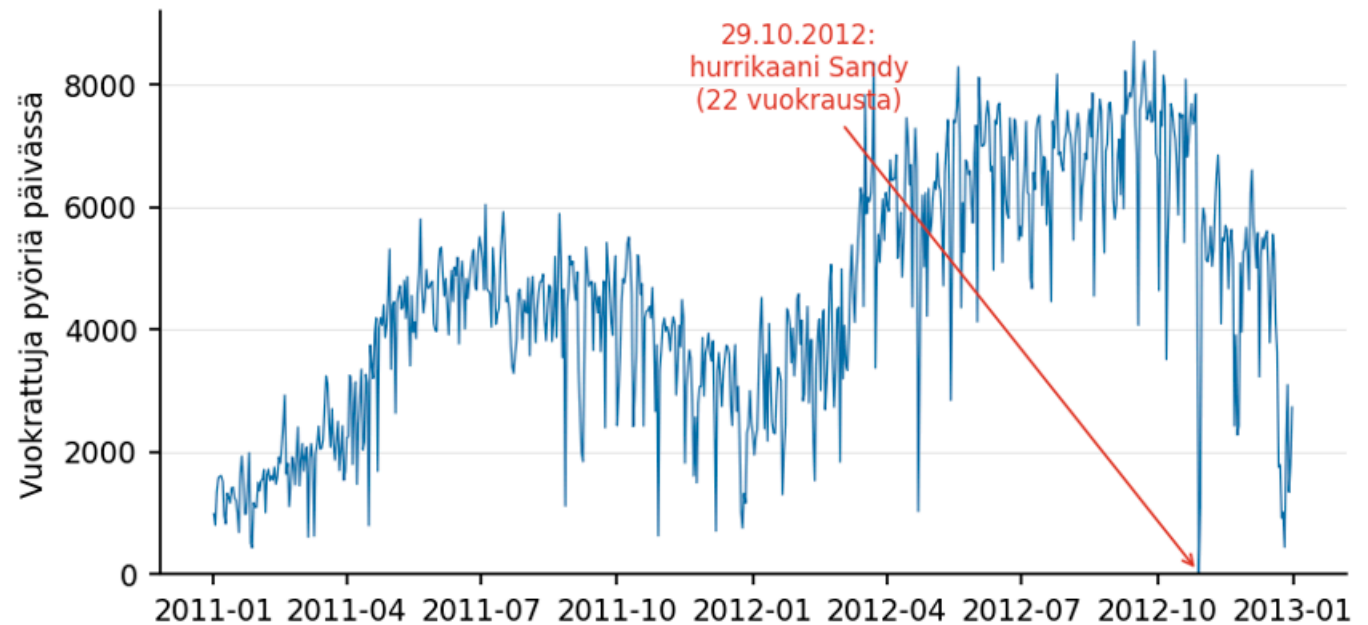
Laske kymmenen päivän lämpötilojen keskiarvo ja mediaani. Ovatko ne lähellä toisiaan vai kaukana toisistaan? Mitä se kertoo, kun muistat, mitä juuri luit päivän 10 myynnistä?

4. KUVAT VOITTAVAT TAULUKOT: VISUALISOINTI

Ihmissilmä on hahmontunnistuskone, mutta vain kun tieto on muotoiltu sille sopivaksi. Numerosarake kätkee tarinansa; oikea kaavio tekee tarinasta mahdollittoman ohittaa. Jokainen peruskaaviotyyppi vastaa eri kysymykseen, ja oikean valitseminen on puolet taidosta.

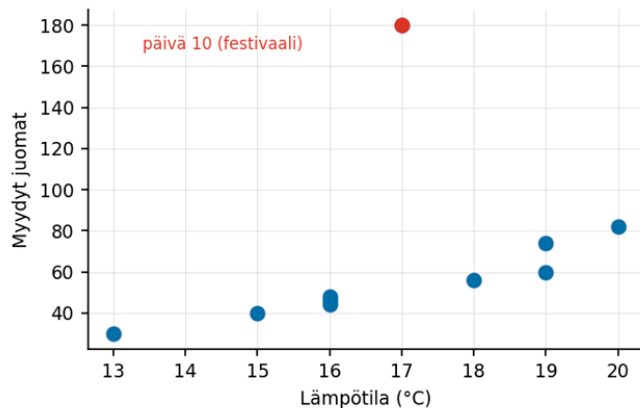
Pylväskaavio vastaa kysymykseen *”miten kategoriat vertautuvat toisiinsa?”* Kategoriat yhdelle akselille, numeerinen tiivistelmä toiselle: keskimyynti viikonpäivittäin, pelaajamäärät genreittäin. Pylväät kutsuvat vertailemaan, ja siksi niiden pitäisi normaalisti alkaa nolasta – lisää aiheesta tuonnempana.

Viivakaavio vastaa kysymykseen *”miten asia muuttui ajassa?”* Aika virtaa vasemmalta oikealle; viivan muoto paljastaa trendit, kausivaihtelut ja äkilliset katkokset. Kaksi vuotta päivittäistä kaupunkipyörädataa viivana piirrettyinä näyttää välittömästi kesä–talvi-aallon, jota mikään taulukko ei paljastaisi (Fanaee-T 2013).



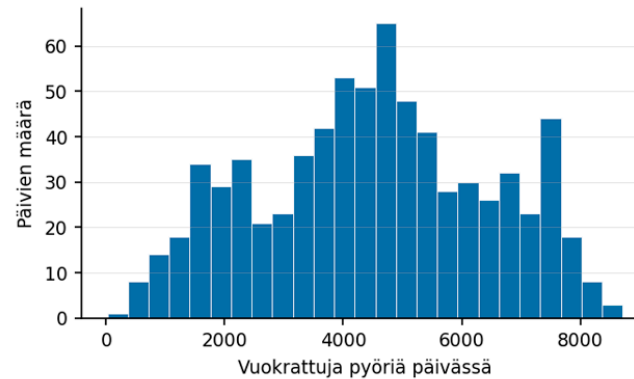
KUVIO 1. Päivittäiset pyörävuokraukset Washington D.C:ssä 2011–2012 viivakaaviona: kesä–talvi-aalto näkyy heti, ja yksi päivä romahtaa 22 vuokraukseen – hurrikaani Sandy, josta lisää luvussa 7 (aineisto: Fanaee-T 2013).

Hajontakuvi vastaa kysymykseen ”*liittyvätkö nämä kaksi lukua toisiinsa?*” Jokaisesta havainnosta tulee yksi piste, jonka paikan määräävät sen kaksi arvoa – kioskillä lämpötila vaakakselille, myydyt juomat pystyakselille. Jos pisteet ajautuvat ylöspäin vasemmalta oikealle, lämpiminä päivinä myydään enemmän. Hajontakuvi on analyytikon lempityökalu, koska se näyttää *jokaisen yksittäisen havainnon* ilman mitään tiivistämistä: poikkeamat, rypäät, kaaret ja aukot ovat kaikki näkyvissä yhtä aikaa.



KUVIO 2. Kioskidatan hajontakuvi: lämpötila ja myydyt juomat. Nouseva suunta erottuu, ja päivä 10 poikkeaa muiden muodostamasta kuviosta.

Histogrammi vastaa kysymykseen ”*miten yhden muuttujan arvot jakautuvat?*” Se pilkkoo vaihteluvälin lokeroihin ja laskee havainnot lokeroittain paljastaen, kasautuvatko arvot yhteen kekoon vai kahteen, vai laahaako perässä pitkä häntä. Juuri tuo pitkä häntä on sitä vinoutta, joka tekee keskiarvoista harhaanjohtavia – tulojen histogrammi näyttää välittömästi, miksi taloustieteilijät suosivat mediaania.



KUVIO 3. Samojen 731 päivän pyörävuokrausten histogrammi kuin KUVIOSSA 1: useimmat päivät kasautuvat keskivaiheille, ja hännät ulottuvat hiljaisiin ja vilkkaisiin päiviin (aineisto: Fanaae-T 2013).

Piirrä ensin, tiivistä sitten. Kuuluisa opetusaineisto, Anscomben kvartetti, sisältää neljä aineistoa, joilla on *identtiset* keskiarvot ja korrelaatiot mutta jotka näyttävät piirrettyinä täysin erilaisilta – yksi on siisti suora, yksi kaari, yhden pilaa yksittäinen poikkeama (Anscombe 1973, 19–20). Tiivistelmäluvut voivat valehdella vaikenemalla; hyvä kuva valehtelee harvoin.

Miten kaaviot valehtelevat

Koska kaaviot ovat vakuuttavia, ne ovat myös harhaanjohtajien lempiase (Huff 1954). Kolme klassikkoa, joita kannattaa varoa sekä muiden kaavioissa että omissasi:

Katkaistu akseli. Pylväskaavio, jonka pystyakseli alkaa 95:stä eikä nolasta, saa 96:n ja 99:n eron näyttämään valtavalta. Pylväät ilmaisevat määrää pituudellaan, joten akselin katkaiseminen on visuaalista valehtelua – valitettavan yleistä uutisgrafiikassa ja myyntikalvoissa.

Kirsikanpaimintaikkuna. ”Hinnat ovat laskeneet tammikuusta!” – kaaviossa, joka sopivasti alkaa tammikuusta, epätavallisen korkealta huipulta. Kysy aina, mitä tapahtui ennen vasenta reunaa.

HARJOITUS 3.

Nimeä kuhunkin kysymykseen paras kaaviotyyppi: (a) Myikö kioski viikonloppuisin enemmän kuin arkisin? (b) Liittyykö lämpötila myyntiin? (c) Miltä tyypillisen päivän myynti näyttää, ja onko joukossa outoja päiviä? (d) Miten myynti kehittyi kymmenen päivän aikana?

Kaksois-y-akseli. Kaksi viivaa, kaksi eri asteikkoa, venytettyinä kunnes ne näyttävät kulkevan yhtä jalkaa. Kun kummankin akselin saa skaalata vapaasti, melkein mitkä tahansa kaksi muuttujaa saadaan näyttämään toisiinsa liittyviltä.

Oman kaavion testi on yksinkertainen: kestäisikö vaikutelma, jos lukija näkisi koko rehellisen kuvan? Jos ei, korjaa kaavio, älä kuvatekstiä.

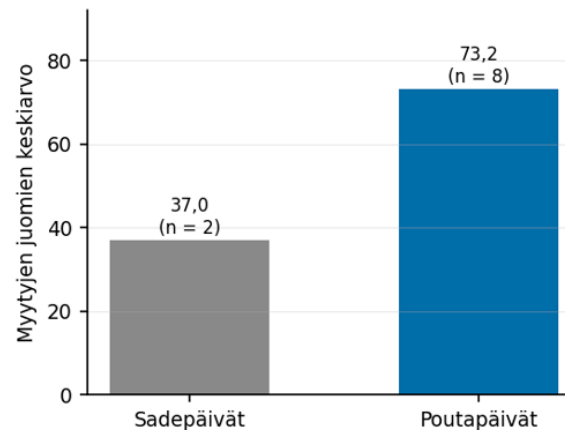


Adobe Stock

5. RYHMIEN VERTAILU

Tässä kohtaa analyysi alkaa tuottaa todellisia vastauksia. Useimmat käytännön kysymykset ovat vertailuja valepuvussa: *Haittaavatko sadepäivät myyntiä? Kuluttavatko pelaajat enemmän päivituksen jälkeen? Kumpi myymäläjärjestys myy enemmän?* Liike on aina sama – **jaa rivit ryhmiin kategorisen muuttujan mukaan ja laske sama tiivistelmä jokaiselle ryhmälle.**

Kioskille: kahtena sadepäivänä myynti oli 44 ja 30, keskiarvo **37**. Kahdeksan poutapäivän keskiarvo on $(40+48+56+60+82+74+46+180)/8 = \mathbf{73,25}$. Poutapäivät näyttävät myyvän tuplasti. Taulukkolaskennassa tämä on yksi funktio: `=KESKIARVO.JOS(sade_sarake; "kyllä"; myynti_sarake)` – suodatus ja keskiarvo yhdellä liikkeellä. Analyytikot tekevät tätä niin usein, että se on joka työkalussa sisäänrakennettuna (Pythonissa sen nimi on `groupby`).



KUVIO 4. Pylväskaavio kioskin ryhmävertailusta: myytyjen juomien keskiarvo sade- ja poutapäivinä. Ryhmäkoot kuuluvat kuvaan – kaksi sadepäivää on hyvin ohut pohja johtopäätökselle.

Kolme varoitusta tekee tästä yksinkertaisesta liikkeestä luotettavan.

Tarkista ryhmäkoot. Sateisten ryhmässämme on **kaksi** päivää. Kaksi! Kahden arvon keskiarvoon pitää suhtautua syvällä epäluulolla – yksi outo päivä kääntää sen kokonaan. Kerro aina, montako havaintoa kussakin ryhmässä on; ”n = 2” on terveysvaroitusta minkä tahansa johtopäätöksen kyljessä.

Tarkista, mikä muu eroaa ryhmien välillä. Sadepäivämme olivat myös viileitä päiviä (13 °C ja 16 °C, vaihteluvälin kylmässä päässä). Haittaako myyntiä siis sade vai kylmyys? Tällä datalla emme voi tietää – selitykset ovat kietoutuneet yhteen. Kietoutumisella on nimi, **sekoittuminen**, ja se on seuraavan luvun pääkonna.

Varo keskiarvojen keskiarvoja. Jos lauantain keskiarvo perustuu 50 havaintoon ja maanantain kolmeen, niiden laskeminen yhteen tasavertaisina on väärin. Kun ryhmät ovat kovin erikokoisia, yhdistetyt tulokset voivat jopa kääntyä päinvastaisiksi – ilmiö tunnetaan nimellä Simpsonin paradoksi (Simpson 1951). Sitä ei tarvitse vielä hallita; riittää, kun muistat, että *ryhmäkoolla on väliä*.

HARJOITUS 4.

Laske kioskitaulukosta viikonlopun (lauantai + sunnuntai) ja arkipäivien keskimyynti. Ennen kuin luotat tulokseen: montako havaintoa kummassakin ryhmässä on, ja eroaako ryhmien välillä jokin muukin kuin viikonloppuisuus?

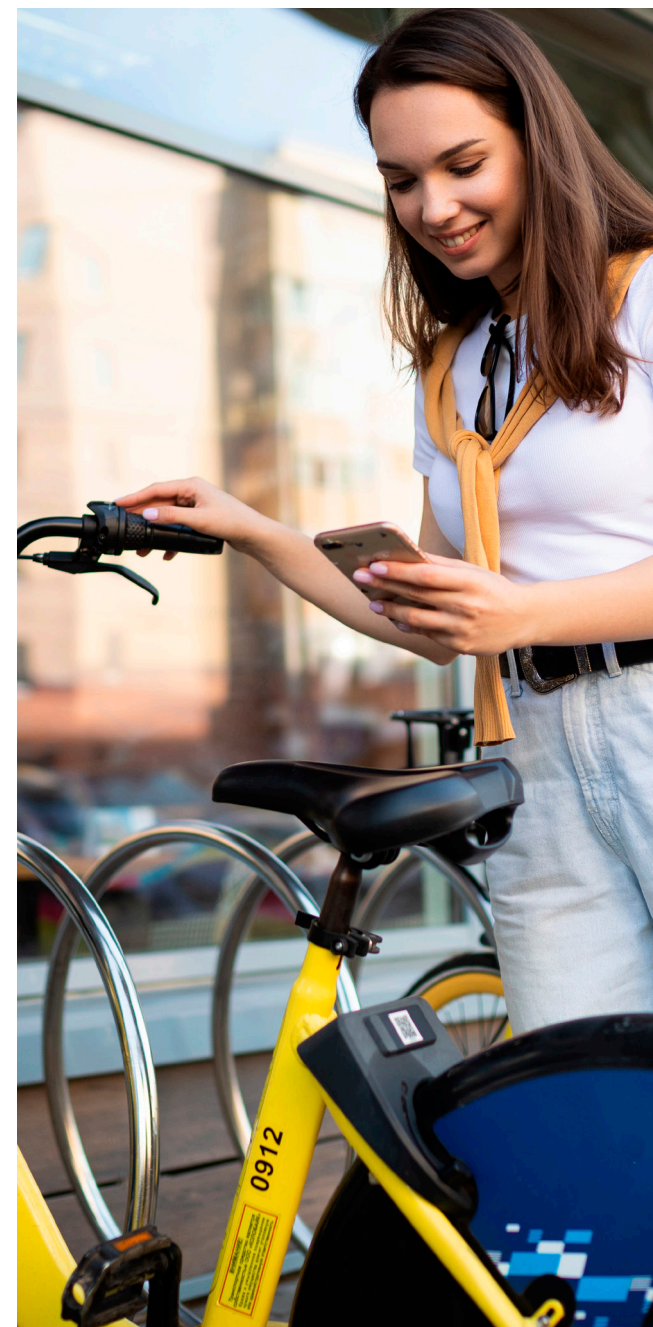
6. RIIPPUVUUDET: KORRELAATIO JA SEN KUULUISA ANSA

Ryhmävertailu toimii, kun toinen muuttuja on kategorinen. Kun *molemmat* muuttujat ovat numeerisia – lämpötila ja myynti, pelitunnit ja rahankäyttö – kysymme sen sijaan: *liikkuvatko ne yhdessä?* Tämä on **korrelaatio**.

Aloita kuvasta. Piirrä pisteet (hajontakuvio); jos ne ajautuvat yhdessä ylöspäin, korrelaatio on positiivinen; alaspäin, negatiivinen; muodoton pilvi tarkoittaa lähellä nollaa olevaa. Jos sitten haluat luvun, **korrelaatiokerroin r** puristaa tuon suunnan yhdeksi arvoksi välille $-1...+1$. Excelissä se on =KORRELAATIO(sarake1; sarake2); jokaisesta analyysityökalusta se löytyy.

r :n lukeminen vaatii kalibrointia, koska oppikirjaesimerkit kalibroivat ihmiset väärin. Todellinen data on kohinaista: $r \pm 0,9$ tai enemmän ei juuri koskaan esiinny fysiikan laboratoriodien ulkopuolella; $\pm 0,5$ – $\pm 0,7$ on vahva, arvokas signaali sotkuisessa inhimillisessä datassa (lämpötila ja pyörävuokraukset kahden vuoden yli: 0,63); $\pm 0,2$ – $\pm 0,4$ on heikko mutta joskus merkityksellinen; ja lähellä nollaa tarkoittaa, ettei *lineaarista* yhteyttä ole. Kioskin lämpötilalle ja myynnille $r \approx 0,36$ – mutta ilman outoa päivää 10 se hyppää noin 0,95:een (muistutus siitä, että yksittäiset poikkeamat voivat pilata tiivistelmäluvut; lisää luvussa 7).

Sana *lineaarinen* kätkee ansan: r mittaa vain suoraviivaisia yhteyksiä. Pyörävuokraukset kasvavat tasaisesti lämpötilan mukana – noin 30 asteeseen asti, jossa ne kääntyvät laskuun, koska polkemaan on liian kuuma. Todellinen yhteys on kaari, mutta r , jonka on pakko tiivistää kaari suoraksi, raportoi keskinkertaisen arvon ja piilottaa kiinnostavimman tosiasian. Lääke on sama kuin aina: **katso hajontakuvio ensin, laske vasta sitten**.



Adobe Stock



Adobe Stock

Korrelaatio ei ole kausaatiota

Nyt koko data-analyysin kuuluisin virke. Oletetaan, että huomaat jäätelönmyynnin korreloivan vahvasti hukkumiskuolemien kanssa kuukausi kuukaudelta. Aiheuttaako jäätelö hukkumisia? Pitäisikö se kieltää? Ei tietenkään: *kesä* aiheuttaa molemmat – jäätelöä syödään enemmän ja useampi ihminen ui. Kesä on **sekoittava tekijä**: kolmas muuttuja, joka ajaa kumpaakin mittaamaasi muuttujaa ja lavastaa niiden välille korrelaation.

Kun huomaat A:n korreloivan B:n kanssa, mahdollisuuksia on aina vähintään neljä: A aiheuttaa B:n; B aiheuttaa A:n (suunta on käänteinen – ehkä valmiiksi paljon myyville kauppoilla on varaa parempiin liikepaikkoihin, eikä liikepaikka aiheutakaan myyntiä); jokin C aiheuttaa molemmat; tai kyse on yhteensattumasta. Kun muuttujia on tarpeeksi, yhteensattumia tulee taatusti – kokonaiset verkkosivustot keräävät absurdeja ”valekorrelaatioita”, kuten juuston henkeä kohti lasketun kulutuksen ja lakanoihin kietoutumiskuolemien yhteisen tahdin ($r = 0,95$ vuosikymmenen yli, eikä se merkitse mitään) (Hrabač & Trkulja 2020, 293 [Vigen 2015]).

Korrelaatio on silti kallisarvoinen: se kertoo, *mistä kannattaa etsiä*. Se tuottaa epäilyjä, ei tuomioita. Kausaation osoittamiseen tarvitaan vahvempia työkaluja – kultainen standardi on **kontrolloitu koe**, jossa *muutat* yhtä asiaa satunnaisesti valitulle puolelle kohteista ja vertaat (verkkosivustot ajavat näitä jatkuvasti ”A/B-testeinä”: puolet käyttäjistä näkee uuden painikkeen, puolet vanhan, ja satunnaisuus takaa, ettei mikään sekoittava tekijä erota ryhmiä toisistaan).

Tässä substanssiosaaminen ansaitsee palkkansa. Data ei yksin voi kertoa, että kesä selittää jäätelö–hukkuminen-yhteyden – *sinun* on tunnettava maailma riittävän hyvin keksiäksesi sen. Analyytikko, joka osaa vain menetelmät muttei tunne kontekstia, raportoi itsevarmasti hölynpölyä.

HARJOITUS 5.

Tutkimus havaitsee, että lasten kengännumero korreloi vahvasti lukutaidon kanssa. Mikä on sekoittava tekijä? Ja vaikeampi: kioskisi myynti korreloi puiston laidalle pysäköityjen polkupyörien määrän kanssa. Luettele neljä mahdollista selitystä.

7. POIKKEAMAT, PUUTTUVA DATA JA SIIVOAMISEN TAITO

On aika käsitellä päivä 10 ja sen 180 juomaa – reilusti yli kaksi keskihajontaa keskiarvon yläpuolella. Arvo, joka on kaukana muun datan muodostamasta kuviosta, on **poikkeama** (engl. outlier), ja se, mitä sille tekee, on analyysin seurauksellisimpia harkintaratkaisuja.

Ensimmäinen sääntö: **poikkeama on kysymys, ei virhe**. Mahdollisuuksia on kolme, ja jokainen vaatii erilaisen vastauksen. Se voi olla *virhe* – lyönti-virhe (18:sta tuli 180), rikkinäinen anturi, kahteen kertaan laskettu kassa. Tutki; jos virhe varmistuu, korjaa tai poista se, ja kirjaa muistiin, että teit niin. Se voi olla *todellinen mutta epäolennainen* – yksittäinen festivaalipäivä, itsessään merkityksellinen mutta myrkkyä, jos tavoitteesi on kuvata normaalia toimintaa; sen voi analysoida erikseen ja jättää pois ”tyypillisen päivän” tiivistelmästä, jälleen valinta kirjaten. Tai se voi olla itse *löytö*. Kuuluisa todellinen esimerkki: kahden vuoden Washington D.C:n kaupunkipyörädatassa yksi lokakuun päivä näyttää 22 vuokrausta, kun normaali taso on noin 5 000. Tuo poikkeama on hurrikaani Sandy, joka tallentui vahingossa polkupyöräaineistoon. (Fanae-T & Gama 2014, luku 3.2.) Poista se ”datavirheenä”, ja poistat samalla datan kiinnostavimman tosiasian.

Sitä, mitä ei saa koskaan tehdä, on poikkeamien poistaminen hiljaa siksi, että ne pilaavat siistin johtopäätöksen. Se ei ole siivoamista; se on petosta lisävaihein.

Puuttuva data on toinen arkipäivän riesa: tyhjiä soluja, ”N/A”, tai – paljon pahempaa – oikeiksi arvoiksi naamioituneita puuttuvia arvoja (0 siellä, missä mitään ei kirjattu, tai paikkamerkkihölyn-pölyä kuten lämpötila –999). Naamioituneet puuttuvat arvot myrkyttävät hiljaa jokaisen niiden yli lasketun keskiarvon, minkä vuoksi ensimmäinen askel uuden aineiston kanssa on aina sama: laske jokaisen sarakkeen minimi ja maksimi ja tuijota kaikkea mahdotonta. Se, pudotetaanko vajaan rivit

vai paikataanko aukot arvioilla, riippuu siitä, *miksi* data puuttuu – satunnaisesti pettävä anturi on harmillinen; anturi, joka pettää juuri myrskyjen aikana, poistaa juuri kiinnostavimmat rivit ja vinouttaa kaiken sen jälkeen lasketun.

Yleinen opetus: todellinen data on likaista, sen siivoaminen on suurin osa työstä, ja jokainen siivouspäätös on analyttinen päätös, joka voi muuttaa johtopäätöksen. Pidä raakadata koskemattomana, työskentele kopiolla ja pidä kirjaa jokaisesta muutoksesta – tuleva minäsi, joka toistaa analyysin kolmen kuukauden päästä, on kiitollinen.

HARJOITUS 6.

Oletetaan, että saat tietää päivän 10 olleen suuren puistofestivaalin päivä. Laske myytyjen juomien keskiarvo ja mediaani ilman päivää 10 ja vertaa niitä luvun 3 arvoihin (keskiarvo 66, mediaani 52). Kumpi tiivistelmä kesti poikkeaman paremmin? Pitäisikö päivä 10 poistaa aineistosta kokonaan?

8. SÄÄNNÖNMUKAISUUKSISTA ENNUSTEIKSI

Kaikki tähänastinen kuvaa menneisyyttä. Hyöty syntyy, kun sen suuntaa tulevaisuuteen: *huomenna on poutainen keskiviikko, 18 °C – montako juomaa kioskin kannattaa varata?* Sillä hetkellä, kun vastaat, olet rakentanut **mallin**: yksinkertaistetun, matemaattisen kuvauksen säännönmukaisuudesta, jota käytetään ennustamiseen.

Rakenna sellainen nyt, kolmessa tarkentuvassa askeleessa, pelkästään ideoilla, jotka sinulla jo on. Karkein ennuste on yleinen tyypillinen päivä: mediaani 52 (jätämme festivaalipäivän 10 pois kaikesta tästä – siivouspäätös, kirjattu muistiin). Tarkenna ryhmävertailulla: huomenna on poutaa, ja poutapäivät (ilman päivää 10) myyvät keskimäärin noin 58. Tarkenna vielä korrelaatiolla: 18 °C on vaihteluvälimme lämpimässä päässä, ja lämpiminä päivinä myydään enemmän, joten nosta hieman – ehkä 60–65. Tuo päättelyketju – lähde perustasosta, korjaa tunnetuilla olosuhteilla – on juuri sitä, mitä muodolliset menetelmät kuten *lineaarinen regressio* tekevät kurinalaisemmin: ne etsivät täsmällisen suoran lämpötila-myynti-hajontakuvion läpi ja lukevat ennusteen suoralta kohdasta 18 °C.

Kaksi oivallusta erottaa ihmiset, jotka käyttävät malleja hyvin, ihmisistä, joille mallit aiheuttavat palovammoja.

Ennusteet ovat aina epävarmoja, ja rehelliset ennusteet kertovat, kuinka epävarmoja. ”Noin 60” ja ”jossain 45:n ja 80:n välillä, todennäköisimmin lähellä 60:tä” ovat eri väitteitä; jälkimmäinen on hyödyllisempi juuri siksi, *että* se myöntää, mitä ei tiedä. Karkeakin haarukka voittaa väärän varmuuden. Sääennusteet oppivat tämän kauan sitten – ”70 % sateen todennäköisyys” on malli, joka on rehellinen.

Malli voi sopia menneisyyteen liian hyvin. Riittävän joustavalla mallilla voit selittää kymmenen päi-

vääsi *täydellisesti* – päivä 3 myi 56, *koska* oli poutainen keskiviikko ja 18 °C – ja ennustaa päivän 11 surkeasti, koska malli opetteli ulkoa kohinan eikä oppinut säännönmukaisuutta. Tämän epäonnistumisen nimi on **ylisovittaminen** (engl. overfitting), ja siksi vakavasti otettavat analyttikot testaavat mallia aina datalla, jota se ei ole koskaan nähnyt, ennen kuin arvioivat sitä. Jos joku näyttää sinulle mallin, jonka ennustehistoria on täydellinen, ensimmäinen kysymyksesi pitäisi olla: täydellinen silmä datalla, jolla malli opetettiin, vai aidosti uudella?

Koneoppiminen, kaikessa hohdokkuudessaan, on tämä sama silmukka – perustaso, säännönmukaisuus, ennuste, virheentarkistus – automatisoituina ja skaalattuna miljooniin riveihin ja tuhansiin muuttujiin. Jos ymmärrät kioskimallin, ymmärrät koko rakennelman luurangon.

HARJOITUS 7.

Ennusta samalla kolmen askeleen logiikalla myynti sateiselle maanantaille, 13 °C. Mitkä menneistä päivistä muistuttavat sitä eniten? Kuinka varma olet, ja miksi sinun ei pitäisi olla kovin varma?

9. TULOSTEN VIESTIMINEN

Analyysillä on merkitystä vain, jos joku toimii sen pohjalta, ja ihmiset toimivat sen pohjalta, minkä he ymmärtävät. Tämä viimeinen taito on aliarvostettu, koska se näyttää helpolta. Se ei ole.

Sano löydös ensin, yhdellä selkeällä virkkeellä, jonka ei-analyytikko voisi toistaa: *”Poutapäivät myyvät suunnilleen tuplasti sadepäiviin verrattuna, joten miehitys ja varasto kannattaa mitoittaa sääennusteen mukaan.”* Huomaa muoto: säännön mukaisuus, sitten seuraus. Aloita sillä – älä pakota yleisöä istumaan koko tutkimusmatkaasi läpi ennen kuin pointti tulee. Näytä sitten yksi rehellinen kaavio, joka tekee löydöksen näkyväksi (yleensä se, joka vakuutti *sinut*), kerro sen takana olevat luvut ja ole avoin rajoituksista: kymmenen päivää dataa, kaksi sadepäivää, sade kietoutunut kylmyyteen. Rajoitusten myöntäminen ei heikennä analyysiä – se tekee lopusta uskottavan, ja juuri se erottaa analyysin mainonnasta.

Käytännön tarkistuslista mille tahansa tulokselle, jota olet jakamassa: Voinko sanoa sen yhdellä virkkeellä? Selviääkö kaavio luvun 4 rehellisyysteististä? Kerroinko ryhmäkoot? Tarkistinko sekoittavat tekijät ja sanoinko, mitkä jäävät jäljelle? Löisinkö lounasrahani vetoa siitä, että tulos pätee myös uudessa datassa?



10. TYÖKALUPAKKISI JA MITÄ SEURAAVAKSI

Jokainen tämän oppaan idea vastaa yhtä komento-
riviä niissä kahdessa työkalussa, jotka todennäköi-
simmin kohtaat. Käytä tätä taulukkoa siltana aina,
kun istut harjoittelemaan.

(Taulukon Excel-funktiot ovat suomenkielisen
Excelin nimiä; englanninkielisessä Excelissä vas-
taavat funktiot ovat AVERAGE, MEDIAN, STDEV.S,
MAX, COUNT, AVERAGEIF ja CORREL.)

Idea (luku)	Excel / Sheets	Python (pandas)
Keskiarvo (3)	=KESKIARVO(B2:B11)	df['sales'].mean()
Mediaani (3)	=MEDIAANI(B2:B11)	df['sales'].median()
Hajonta (3)	=KESKIHAJONTA.S(B2:B11)	df['sales'].std()
Min / maks (7)	=MIN(...), =MAKS(...)	df['sales'].min(), .max()
Rivien määrä (2)	=LASKE(B2:B11)	len(df)
Ryhmäkeskiarvo (5)	=KESKIARVO.JOS(D2:D11;"kyllä";B2:B11)	df.groupby('rain')['sales'].mean()
Korrelaatio (6)	=KORRELAATIO(B2:B11;C2:C11)	df['sales'].corr(df['temp'])
Hajontakuvio (4)	Lisää → Pistekaavio	df.plot.scatter(x='temp',y='sales')
Lajittele / etsi poikkeamat (7)	Data → Lajittele	df.sort_values('sales')

Kukaan ei opettele näitä ulkoa – analyytikot tarkis-
tavat syntaksia jatkuvasti. Taito, jota olet tässä op-
paassa rakentanut, on tietää, *mitä kysyä*; syntaksi
on pelkkää oikeinkirjoitusta.

Jatkoa varten: harjoittele oikeilla aineistoilla osoit-
teessa **kaggle.com** (tuhansia ilmaisia aineistoja
jalkapallo-otteluista Spotify-kappaleisiin – valitse
aihe, josta jo valmiiksi välität, koska substanssi-
osaaminen tekee analyysistä dramaattisesti hel-
pompaa ja hauskeempaa). Aja Pythonia ilmaiseksi
selaimessa osoitteessa **colab.google** ilman mitään
asennuksia. Ja kun kohtaat varsinaisen tilastotie-
teen – hypoteesitestit, luottamusvälit, regression
– huomaat niiden olevan huolellisia, muodollisia
versioita kysymyksistä, jotka osaat jo esittää: *onko
tämä ero todellinen vai tuuria? kuinka varmoja
olemme? mitä hajontakuvion läpi kulkeva suora
ennustaa?*

PIENI SANASTO

Korrelaatio – se, missä määrin kaksi numeerista muuttujaa liikkuvat yhdessä, välillä $-1...+1$.

Sekoittava tekijä – piilossa oleva kolmas muuttuja, joka ajaa kahta muuta ja lavastaa niiden välille yhteyden.

Substanssiosaaminen – datan takana olevan todellisen aihepiirin ymmärrys.

Jakauma – muuttujan arvojen kokonaishahmo (ks. histogrammi).

Keskiarvo / mediaani / moodi – kolme versiota ”keskikohdasta”: aritmeettinen keskiarvo / järjestetyn datan keskimäinen arvo / yleisin arvo.

Malli – yksinkertaistettu kuvaus säännönmukaisuudesta, jota käytetään ennustamiseen.

Poikkeama – havainto kaukana muun datan muodostamasta kuviosta.

Ylisovittaminen – malli opettelee opetusdatansa ulkoa sen sijaan, että oppisi säännönmukaisuuden, ja epäonnistuu uudella datalla.

Otos vs. perusjoukko – data, joka sinulla on, vs. kaikki se, mistä haluaisit tietää.

Keskihajonta – arvojen tyypillinen etäisyys keskiarvostaan.

Siisti data – yksi havainto per rivi, yksi muuttuja per sarake.

HARJOITUSTEN VASTAUKSET

1. Viikonpäivä: kategorinen. Myydyt juomat: numeerinen (diskreetti lukumäärä). Lämpötila: numeerinen (jatkuva). Sade: totuusarvoinen/kategorinen. Keskimääräistä viikonpäivää ei ole – mutta voit ryhmitellä viikonpäivän mukaan ja laskea *myynnin* keskiarvon kussakin ryhmässä, mikä on koko luvun 5 temppu.

2. Lämpötilat: $15+16+18+16+19+20+19+13+16+17 = 169$, keskiarvo 16,9. Järjestetyn jonon keskimäinen pari on 16 ja 17 $\rightarrow$ mediaani 16,5. Keskiarvo ja mediaani ovat lähes samat, joten lämpötilat datassa ei ole villiä poikkeamaa – päivän 10 *myynti* oli äärimmäinen, mutta sen *lämpötila* (17 °C) oli tavallinen. Poikkeamat asuvat tietyissä sarakkeissa, eivät kokonaisissa riveissä.

3. (a) Pylväskaavio (kategorioiden vertailu). (b) Hajontakuvio (kaksi numeerista muuttujaa). (c) Histogrammi (jakauma ja oudot päivät). (d) Viivakaavio (muutos ajassa).

4. Viikonloppu: $(82+74)/2 = 78$. Arkipäivät: $(40+48+56+44+60+30+46+180)/8 = 63$. Viikonloppu näyttää korkeammalta – mutta viikonlopun $n = 2$, ja arkipäivien ryhmässä on festivaalipoikkeama, joka nostaa sen keskiarvoa; ilman päivää 10 arkipäivien keskiarvo putoaa noin 46,3:een, jolloin viikonloppuvaikutus näyttää paljon suuremmalta. Pienet ryhmät ja käsittelemättömät poikkeamat voivat kumpikin kääntää johtopäätöksen.

5. Kengännumero ja lukutaito: sekoittava tekijä on *ikä* – isommilla lapsilla on isommat jalat ja parempi lukutaito. Kioski ja polkupyörät: (1) pyörät tuovat asiakkaita ($A \rightarrow B$); (2) juomia ostamaan tulleet pysäköivät pyöränsä ($B \rightarrow A$); (3) aurinkoinen sää aiheuttaa sekä pyöräilyä että janoa (sekoittava tekijä C); (4) yhteensattuma vain kymmenen päivän jaksolla. Huomaa, että (1), (2) ja (3) voivat kaikki olla osittain totta yhtä aikaa.

6. Ilman päivää 10: summa 480, yhdeksän päivää, keskiarvo $\approx 53,3$ (oli 66 – liikkui melkein 13 yksikköä). Yhdeksän järjestetyn arvon mediaani on 48 (oli 52 – liikahti tuskin lainkaan). Mediaani oli selvästi kestävämpi. Päivää 10 *ei* pidä poistaa aineistosta kokonaan – se on todellinen ja kiinnostava (analysoi festivaalit erikseen!) – mutta se pitää jättää pois tiivistelmistä, jotka väittävät kuvaavansa *normaalia* päivää, ja poisjätto on kerrottava.

7. Samankaltaisin päivä on päivä 8 (sateinen maanantai, 13 °C): 30 juomaa. Toinen sadepäivä (44) ja yleinen ”sade suunnilleen puolittaa myyntiä” -sääntö tukevat ennustetta noin 30–40. Varmuuden pitää olla matala: datassa on tasan kaksi sadepäivää, sade on kietoutunut kylmyyteen, ja maanantaita on vähän. Leveä, rehellisesti ilmoitettu haarukka – ”luultavasti 25–45” – on ammattilaisen vastaus.

LÄHTEET

- Anscombe, F. J. 1973. Graphs in statistical analysis. *The American Statistician*, 27(1), 17–21. Saatavissa: <https://doi.org/10.1080/00031305.1973.10478966>. Viitattu 9.7.2026.
- Dasu, T. & Johnson, T. 2003. *Exploratory Data Mining and Data Cleaning*. Hoboken: John Wiley & Sons.
- Encyclopaedia Britannica. *Sirius*. Saatavissa: <https://www.britannica.com/place/Sirius-star>. Viitattu 9.7.2026.
- Fanaee-T, H. 2013. *Bike Sharing*. Aineisto. UCI Machine Learning Repository. Saatavissa: <https://doi.org/10.24432/C5W894>. Viitattu 9.7.2026.
- Fanaee-T, H. & Gama, J. 2014. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2–3), 113–127. Saatavissa: <https://doi.org/10.1007/s13748-013-0040-3>. Viitattu 9.7.2026.
- Fraknoi, A., Morrison, D. & Wolff, S. 2022. *Astronomy 2e*. Houston: OpenStax, Rice University. Saatavissa: <https://openstax.org/books/astronomy-2e/pages/4-4-the-calendar>. Viitattu 10.7.2026.
- Hrabač, P. & Trkulja, V. 2020. Of cheese and bedsheets – some notes on correlation. *Croatian Medical Journal*, 61(3), 293–295. Saatavissa: <https://doi.org/10.3325/cmj.2020.61.293>. Viitattu 10.7.2026.
- Huff, D. 1954. *How to Lie with Statistics*. New York: W. W. Norton.
- Neugebauer, O. 1969. *The Exact Sciences in Antiquity*. 2. painos. New York: Dover Publications.
- Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13(2), 238–241.
- Vigen, T. 2015. *Spurious Correlations*. New York: Hachette Books.
- Wickham, H. 2014. Tidy data. *Journal of Statistical Software*, 59(10), 1–23. Saatavissa: <https://doi.org/10.18637/jss.v059.i10>. Viitattu 9.7.2026.