

Juha Peltomäki

DATAN ANALYSOINTI PYTHON-KIELELLÄ

Datan analysointi Python-kielellä

Juha Peltomäki

Esipuhe

Data-analytiikka on noussut 2000-luvulla tärkeäksi tietotekniikan sovellusalueeksi. Samalla laskentatehon kasvu mahdollistaa yhä suurempien datamäärien käsittelyn myös kotitietokoneilla. Data-analytiikasta onkin tullut tärkeää niin liiketoiminnassa, tutkimuksessa kuin tekoälyjärjestelmien kehittämisessä.

Kirjan esimerkit on tehty Python-kielellä. Sen suosio perustuu selkeään syntaksiin ja laajoihin kirjastoihin. Erityisen paljon Pythonia käytetään data-analytiikassa ja tekoälysovellusten toteuttamisessa.

Tämän kirjan esimerkeissä käytetään useampaa Python-kirjastoa. Niistä NumPy sisältää työkalut numeeriseen laskentaan ja taulukoiden käsittelyyn. Vastaavasti Pandas tarjoaa työkalut datan lukemiseen, luontiin, analysointiin ja muokkaamiseen. Matplotlib- ja Seaborn-kirjastoilla sekä datan että analyysin tulokset voi visualisoida kuvaajiksi.

Kirjan alussa käydään läpi NumPy- ja Pandas-kirjastojen tärkeimmät ominaisuudet. Perusteiden jälkeen jatketaan datan esikäsittelyyn, ryhmittelyyn ja yhdistämiseen. Kirja sisältää runsaasti pieniä mutta käytännönläheisiä esimerkkejä, joita voi soveltaa omiin projekteihin. Lopussa toteutetaan pieni data-analyysiprojekti, jossa edetään datan lataamisesta analysointiin ja visualisointiin saakka.

Kirja kokoaa lukijalle keskeisimmät data-analytiikan taidot yhteen, tiiviiseen pakettiin. Se sopii hyvin kaikille data-analytiikasta kiinnostuneille. Kirjaa voi käyttää kurssikirjana, mutta se soveltuu hyvin myös itseopiskeluun. Aiempi kokemus Python-ohjelmoinnista helpottaa asioiden oppimista. Toivon, että kirja tarjoaa sinulle oivaltavia lukuhetkiä data-analytiikan opiskeluun.

Tämä kirja on saanut tukea Suomen tietokirjailijat ry:ltä.

Juha Peltomäki lokakuussa 2026.

Tämän teoksen osittainenkin kopiointi on tekijänoikeuslain mukaisesti kielletty ilman oikeudenhaltijan lupaa.

Nimikkeen automaattinen analysointi tietojen, erityisesti mallien, trendien ja korrelaatioiden, saamiseksi tekijänoikeuslain 13b § ("tekstin- ja tiedonlouhinta") mukaisesti on kielletty.

© 2026 Juha Peltomäki

Kustantaja: BoD · Books on Demand, Mannerheimintie 12 B, 00100 Helsinki, bod@bod.fi

Kirjapaino: Libri Plureos GmbH, Friedensallee 273, 22763 Hampuri, Saksa

ISBN: 978-952-89-4775-2

Sisällysluettelo

Python-kieli Data-analytiikassa 6
Python-ohjelmointikieli 7
Python-tulkki 7
NumPy-kirjasto 8
NumPyn historia 8
Johdanto NumPy-kirjastoon 8
NumPy-taulukko 9
NumPy-taulukon luonti 9
dtype-tietotyyppi 10
Omatekoiset tietotyypit 11
Vektorointi 13
Overflow-virheet 14
Matriisi 17
NumPyn taulukko-operaatiot 18
Taulukon uudelleenmuokkaaminen 18
Transpoosi 19
Viipalointi 19
Matemaattisia funktioita taulukkojen käsittelyyn 21
Summa ja axis-parametrit 22
NumPy axis 23
Minimi ja maksimi 24
Keskiarvo 25
Kumulatiivinen summa 25
Tulo ja kumulatiivinen tulo 26
Taulukon luonti eri tavoilla 27
linspace 27
zeros 28

Identiteettimatriisi 28
arange 29
arange vai linspace 30
Matriisien laskuoperaatiot 30
Alkioiden summa 30
Alkioiden tulo 31
Alkioiden jakolasku 31
Broadcasting 31
Lineaarialgebra 32
Matriisitulo 32
Käänteismatriisi 34
vstack ja hstack 35
Taulukon litistäminen yksiulotteiseksi 36
Merkkijonot 36
Kuvioiden piirto 37
Matplotlib-esimerkkejä 38
Pandas-kirjasto 41
Pandasin tärkeimmät tietorakenteet 41
Pandas Series 42
Series-rakenteen luonti 42
Series-operaatiot 44
Viipalointi (slicing) 45
Lajittelu 46
Suodattaminen 47
Series-käsittely NumPy-funktiolla 48
Series ja DataFrame 50
Indeksityypit 52
Series NumPy-taulukoksi 54

Series ja silmukat 54
DataFrame 55
DataFramen luonti 56
Sarakkeiden valinta 59
.loc-valinta 59
Sarakkeiden valinta ilman .loc-indeksointia 60
.loc ja [] sarakkeiden valinnassa 61
DataFramen suodattaminen 62
DataFrame-taulukoiden yhdistäminen 65
concat 66
join 67
merge 69
DataFrame ja tilastolliset operaatiot 70
describe 73
Oman funktion kutsuminen 74
Esimerkki funktioiden käyttämisestä 76
DataFramen Iterointi 78
Järjestäminen 82
Indeksointi 83
Indeksisarakkeen asettaminen indeksiksi 84
Viipalointi 85
Korrelaation laskeminen 87
Aikasarja 88
Tasoitusmenetelmät 89
Liukuva keskiarvo 89
ewm 91
expanding 91
pct_change 92

DataFramen datan esikäsittely 95
NaN -arvot 95
Puuttuvien arvojen käsittely 97
NaN-arvojen korvaaminen 97
Arvojen korvaaminen 98
Arvojen täydentäminen 99
Duplikaattien poistaminen 100
Datan ryhmittely 100
groupby 101
agg - funktioiden yhdistäminen 106
Ryhmiteltyjen tulosten iterointi 107
Tiedon lukeminen DataFrameen ulkoisesta datalähteestä 110
HTML-taulukkodatan lukeminen 111
Visualisointi Matplotlib- ja Seaborn-kirjastoilla 114
Visualisoinnin perustekniikat 114
Aikasarjan visualisointi 118
Pylväsdiagrammi 121
Useita kuvioita yhteen kuvaan 125
Histogrammi 128
Ympyräkuvio 130
Miniprojekti: Applen pörssikurssin data-analyysi 134
Datan lukeminen 134
Päivämäärän käsittely 134
Tilastot 135
Aikaväli 135
Datan tutkiminen (EDA) 136
Puuttuvat arvot 137
Hintojen kehitys 138

Aikasarjan tiheys ja uudelleennäytteenotto 138
Kuukausi- ja kvartaalitason vertailu 140
Tuotot ja muutokset 141
Liukuvat keskiarvot 142
Kvantiilit 143
Poikkeavat arvot 144
Korrelaatio 145
Miniprojektin yhteenveto 146

Python-kieli Data-analytiikassa

Python on korkeatasoinen, rakenteinen, oliopohjainen avoimen lähdekoodin (*Open Source*) ohjelmointikieli, jota voi käyttää monipuolisesti erilaisiin ohjelmointitehtäviin, koska siihen on kehitetty erittäin runsaasti erilaisia laajennuskirjastoja. Sen alkuperäinen kehittäjä on hollantilainen **Guido Van Rossum** vuonna 1991. Ensimmäinen **Python 3** -versio eli **Python 3.0** julkaistiin joulukuussa 2008. Se ei ole täysin yhteensopiva **Python 2** - version kanssa, mikä oli Python-kielen vuonna 2000 julkaista edellinen pääversio.

Python on data-analytiikan käytetyin ohjelmointikieli, koska se tarjoaa helposti opittavan ohjelmointikielen, laajan ekosysteemin sekä tehokkaat kirjastot data-analytiikkaan. Se tarjoaa valmiit ja useimmiten ilmaiset kirjastot ja työkalut data-analytiikkaan. Näistä tärkeimpiä ovat:

- **Pandas**, joka soveltuu datan hakuun, muokkaamiseen ja analyysiin.
- **NumPy**, joka soveltuu numeeriseen laskentaan ja matemaattisten funktioiden toteuttamiseen.
- **Matplotlib** ja **Seaborn**, jotka ovat yleisimmät kirjastot datan visualisointiin.
- **Scikit-learn** taas mahdollistaa koneoppimismallien kehittämisen esikäsitellyn datan pohjalta.

Python toimii useissa eri ympäristöissä ja periaatteessa kehittäminen vaatii vain tekstieditorin ja Python-tulkin asentamisen. Mutta nopeaan kehittämiseen on tarjolla mm. J**upyter Notebook**, joka yhdistää koodin ja dokumentaation tarjoten pohjan interaktiiviseen data-analyysiin. Sovelluskehittimistä suosituimpia ovat **Visual Studio Code** joustavaan kehitykseen sekä **PyCharm** ammattimaiseen projektinhallintaan.

Nykyisin Python on teollisuusstandardi tieteellisessä laskennassa, data-analytiikassa, koneoppimisessa sekä tekoälysovelluksissa. Myös tutkijat käyttäjät Python-kieltä runsaasti, joskin tutkimuksessa myös **R-kieli** on hyvin suosittu vaihtoehto. Python-kieltä voi opiskella useilla verkkokursseja ja sekä oppimateriaalia että kirjallisuutta löytyy runsaasti. Myös yhteisön tuki on laajaa, koska kieli pohjautuu avoimeen lähdekoodiin ja runsaaseen määrään ilmaisia kirjastoja.

Vaikka Python on kielenä helppokäyttöinen, niin se skaalautuu suuriin datamääriin. Tämän mahdollistaa se, että Python käyttää tehokkaita kirjastoja, jotka on kehitetty **C**- ja **C++**-kielillä, kuten esimerkiksi NumPy- ja Pandas-kirjastot.

Python-kieli on suosittu data-analytiikassa, koska se:

- on helppo oppia, koska syntaksi on helppolukuista
- on tehokas datan analysointiin soveltuen myös koneoppimiseen
- sisältää runsaasti käyttövalmiita kirjastoja (mm. Pandas, NumPy, Matplotlib)
- on laajalti tuettu eri käyttöjärjestelmissä (mm. Windows, MacOS, Linux)
- on avoimen lähdekoodin ohjelmointiympäristö ollen täten vapaasti käytettävä

Datan analysointi Python- kielellä on käytännönläheinen johdatus data- analytiikkaan Python- ohjelmointikielellä. Kirjassa tutustutaan Python-ohjelmointikielen käyttöön data-analytiikassa ja opitaan käyttämään suosittuja NumPy- , Pandas- , Matplotlib- ja Seaborn- kirjastoja. Teos etenee perusteista käytännön sovelluksiin ja käsittelee muun muassa datan lukemista, muokkaamista, yhdistämistä, ryhmittelyä sekä visualisointia havainnollisten kuvaajien avulla. Kirjan avulla lukija oppii käyttämään Python- kieltä omissa data-analytiikan projekteissaan. Kirja soveltuu opiskelijoille, opettajille, ohjelmistokehittäjille sekä kaikille data-analytiikasta kiinnostuneille. Kirja sopii oppimateriaaliksi eritasoisille data- analytiikan kursseille ja sisältö toimii erinomaisesti myös verkko- ja itseopiskelussa.

BoD
9 789528 947752