suunnittelu

June 1, 2025

SUUNNITTELU

1. Valittu aineisto Kotimaan lentoasemien kuukausittaiset matkustajamäärät (2019–2024) StatFin-PX-Webistä JSON-stat2-muodossa. Aineisto sisältää kustakin kuukaudesta tiedot lentoasemittain (esim. Helsinki-Vantaa, Oulu, Kuopio) ja kokonaismatkustajamäärät. Yhteensä havaintoja on noin 15–20 lentoasemalta kuukaudessa, mikä riittää korrelaatioiden ja tilastollisten testien suorittamiseen.

2. Mitä halutaan selvittää

   Kuinka matkustajamäärät ovat kehittyneet 2019–2024 (pandemian vaikutus ja toipuminen).

   Ovatko eri lentoasemien kuukausimääriin liittyvät kausivaihtelut samankaltaisia (esim. Helsinki-Vantaa vs. Oulu).

   Pearson-korrelaatio Helsinki-Vantaan ja Oulun kuukausimäärien välillä vuosina 2019–2024.

   Tilastollinen testi: vertailla “suurten lentoasemien” (Helsinki, Oulu) ja “pienten maakuntakenttien” (Kuopio, Rovaniemi) kuukausimääriä vuonna 2023 t-testillä (tai Mann–Whitney U -testillä, jos normaalijakautumisen oletus ei toteudu).

   (Valinnainen) Ennustemalli: Holt–Winters-ennuste Helsinki-Vantaan matkustajamäärille 2024 ja vertailu todellisiin 2024 arvoihin.

3. Datan käsittely ja esikäsittely

   Lataus: JSON-stat2-haulla rajataan “Lennon tyyppi” = “Saapuneet/lähenteet yhteensä” ja “Saa” = “Yhteensä” siten, että DataFrame-riveillä on vain kuukausikohtainen kokonaismatkustajamäärä.

   Sarakenimet ja tyypit: Nimetään sarakkeet (Vuosi → Year, Kuukausi → MonthCode, Ilmoittava lentoasema → Airport, Value → Passengers) ja muutetaan “2023M05” → datetime (toimii indeksinä).

   Ryhmittely: Lisätään sarake AirportGroup arvoilla “Large” (Helsinki-Vantaa, Oulu) ja “Small” (Kuopio, Rovaniemi ja muut maakuntakentät) tilastollisia vertailuja varten.

   Puuttuvien arvojen tarkistus: Poistetaan (dropna) mahdolliset kuukaudet, joilta kokonaismatkustajamäärä puuttuu.

4. Analyysi ja visualisoinnit

   Aikasarjakuvaaja: Piirretään line plot Helsinki-Vantaan kuukausimääriä 2019–2024 varten trendin ja kausivaihtelun havainnollistamiseksi.

Kausivertailu: Piirretään esimerkiksi Oulun ja Kuopion “kesä vs. talvi” -kuukausien vertailu pylväsdiagrammina (2023).

Korrelaation laskenta: Lasketaan df[“Helsinki-Vantaa”].corr(df[“Oulu”]) koko aineistolla (2019–2024).

Tilastollinen testi:

```
Levene-test: stats.levene(large_passengers, small_passengers) ( =0,05) varianssien tarkis

t-test (tai Mann-Whitney U, jos normaalijakautuminen epäonnistuu) vuoden 2023 kuukausimää

Tulostetaan t/U-arvot, p-arvot ja 95 % luottamusvälit ryhmien keskiarvoille.
```

(Valinnainen) Ennustemalli: Sovitetaan Holt-Winters Helsinki-Vantaan dataan (2019–2023) ja ennustetaan 2024. Piirretään hajontakaavio “Todellinen 2024 vs. Ennuste 2024”.

5. Miksi toimenpiteet ovat tarpeellisia

   JSON-stat2-muoto antaa kaikki dimensiot suoraan, eikä tarvitse manuaalista “skiprows”-käsittelyä.

   Sarakenimien uudelleennimeäminen ja päivämäärä-indeksi helpottaa suodatusta ja ryhmittelyä Pythonissa.

   Ryhmittely Large vs. Small mahdollistaa tilastollisen vertailun kahden lentoasemaluokan välillä.

   Aikasarjakuvaajat näyttävät trendit ja kausivaihtelun, jotta voin ymmärtää pandemiavaiheen vaikutukset ja toipumisen.

   Korrelaatio kertoo, miten synkronisesti suurten lentoasemien kuukausi-arvot liikkuvat.

   Levene/test + t-test (tai Mann–Whitney) havaitsevat, onko keskimääräisissä kuukausimääriä eroa ryhmien välillä tilastollisesti merkitsevästi (=0,05).

   Ennustemalli (jos toteutetaan) havainnollistaa aikaisempaan dataan perustuvan mallin ennustetarkkuutta.