



# PYTHON JA DATA-ANALYTIikka

Lopputehtävä

2025

TEKIJÄ/T Wefky Hamed

# SISÄLTÖ

1. SUUNNITTELU .....	3
1.1 Valittu aineisto .....	3
1.2 Mitä halutaan selvittää.....	3
1.3 Datan käsittely ja esikäsittely .....	3
1.4 Analyysi ja visualisoinnit.....	4
1.5 Miksi projekti toteutetaan.....	4
1.6 Miksi toimenpiteet ovat tarpeellisia .....	4
2. TOTEUTUS.....	5
2.1 Johdanto ja tavoite .....	5
2.2 Datan lataus ja esikäsittely.....	6
2.2.1 JSON-aineiston lukeminen ja muuntaminen taulukkomuotoon .....	9
2.2.2 Datan siivous ja muokkaaminen analyysia varten .....	10
2.2.3 Taulukon yksinkertaistaminen ja siistiminen analyysiä varten.....	10
2.3 Datan analyysi.....	11
2.3.1 Lentoasemien ryhmittely vertailua varten .....	11
2.3.2 Tunnusluvut: Suuret ja pienet lentoasemat vuonna 2024 .....	12
2.3.3 Tunnusluvut: Yksittäiset lentoasemat vuonna 2024 .....	12
2.3.4 Ristiintaulukointi: Lentoasemaryhmät ja vuodenajat .....	13
2.3.5 Aikasarjakuvaaja: Pandemia-ajan vaikutukset Helsinki-Vantaan lentoasemalla.....	15
2.3.6 Kausivaihtelun vertailu pylväsdiagrammilla (Oulu ja Kuopio 2023) .....	16
2.3.7 Korrelaatioanalyysi: Helsinki-Vantaa ja Rovaniemi .....	17
2.3.8 Taulukko: Rovaniemen matkustajamäärät joulukuussa (2019–2022).....	17
2.3.9 Piirakkakaavio: Rovaniemen matkustajamäärien jakauma joulukuussa (2019–2022) .....	18
2.4 Tilastolliset testit: Suuret vs. pienet lentoasemat .....	18
2.4.1 Ryhmien muodostaminen (Large/Small) .....	19
2.4.2 Levene-testi (varianssien vertailu) .....	19
2.4.3 T-testi ja/tai Mann–Whitney U -testi (keskiarvojen/medioiden vertailu).....	20
3. YHTEENVETO JA TULKINTA.....	20

## 1. SUUNNITTELU

### 1.1 Valittu aineisto

Kotimaan lentoasemien kuukausittaiset matkustajamäärät (2019–2024) haetaan StatFin-PX-Webistä JSON-stat2-muodossa. Aineistosta poimitaan kunkin kuukauden kokonaismatkustajamäärät lentoasemittain (esim. Helsinki-Vantaa, Oulu, Kuopio). Havaintoja kertyy noin 15–20 lentoasemalta joka kuukausi, mikä riittää korrelaatioiden ja tilastollisten testien tekemiseen.

### 1.2 Mitä halutaan selvittää

- Selvitetään, miten kuukausittaiset matkustajamäärät kehittyvät vuosina 2019–2024 (pandemian vaikutus ja toipuminen).
- Tarkastellaan, ovatko eri lentoasemien kausivaihtelut samankaltaisia (esim. Helsinki-Vantaa vs. Oulu). Lasken Pearson-korrelaation Helsinki-Vantaan ja Oulun kuukausimäärien välillä vuosilta 2019–2024.
- Suoritan tilastollisen testin (Levene + t-test tai Mann–Whitney U, jos normaalijakautumisen oletus ei toteudu) vertaamaan “suurten lentoasemien” (Helsinki, Oulu) ja “pienten maakuntakenttien” (Kuopio, Rovaniemi) kuukausimääriä vuonna 2023.

### 1.3 Datan käsittely ja esikäsittely

- Lataus: Teen JSON-stat2-pyyntöä, jossa rajaan “Lennon tyyppi” = “Saapuneet/lähenteet yhteensä” ja “Saa” = “Yhteensä”. Näin saan pelkät kuukausikohtaiset kokonaismatkustajamäärät.
- Sarakenimet ja tyypit: Uudelleen nimeän sarakkeet seuraavasti:
  - Vuosi → Year
  - Kuukausi → MonthCode
  - Ilmoittava lentoasema → Airport
  - Value → PassengersMuun muassa muutan MonthCode (esim. “2023M05”) datetime-muotoon ja asetan sen indeksiksi.
- Ryhmittely: Lisään sarakkeen AirportGroup, jossa arvona on "Large" (Helsinki-Vantaa, Oulu) tai "Small" (Kuopio, Rovaniemi ja muut maakuntakentät), jotta voin vertailla lentoasemaryhmiä tilastollisesti.
- Puuttuvien arvojen tarkistus: Tarkistan ja poistan (dropna) kuukaudet, joilta matkustajamäärä puuttuu.

## 1.4 Analyysi ja visualisoinnit

- Aikasarjakuvaaja: Piirrän line plotin Helsinki-Vantaan kuukausimäärille vuosilta 2019–2024 trendin ja kausivaihtelun havainnollistamiseksi.
- Kausivertailu: Vertailen Oulun ja Kuopion kesä- vs. talvikuukausien matkustajamääriä vuoden 2023 osalta pylväsdiagrammilla.
- Korrelaatio: Lasken korrelaation `df["Helsinki-Vantaa"].corr(df["Oulu"])` ajanjaksolle 2019–2024.
- Tilastollinen testi:
  - i. Teen Levene-testin varianssien homogeenisuuden tarkistamiseksi (`stats.levene(large_2023, small_2023)`).
  - ii. Suoritan t-testin (tai Mann–Whitney U -testin, jos normaalijakautumisoletama ei toteudu) vertaamaan “suuret” vs. “pienet” lentoasemat vuoden 2023 kuukausimääriin. Tulostan testisuureet ja p-arvot.

## 1.5 Miksi projekti toteutetaan

Projektin tarkoituksena on saada ymmärrys siitä, miten Suomen lentoasemien matkustajamäärät kehittyivät pandemia-ajan jälkeen ja onko eri lentoasemilla samankaltaisia kausivaihteluita. Lisäksi haluan selvittää, kuinka tiiviisti suurten lentoasemien (Helsinki-Vantaa ja Oulu) matkustajamäärät seuraavat toisiaan (korrelaatio) sekä onko suurten ja pienten lentoasemien kuukausimääriin eroja tilastollisesti merkitsevästi (t-test tai Mann–Whitney U). Näin hankittu tieto tukee päätöksentekoa esimerkiksi lentoasemien resurssisuunnittelussa ja antaa mallin aikasarjaennusteisiin.

## 1.6 Miksi toimenpiteet ovat tarpeellisia

- JSON-stat2-muoto mahdollistaa dimensioiden suoran lataamisen ilman monimutkaista “skiprows”-puhdistusta.
- Sarakenimien uudelleennimeäminen ja datetime-indeksin käyttö helpottavat Pythonissa suodatusta ja ryhmittelyä.
- Ryhmittely “Large” vs. “Small” lentoasemiin on tarpeen, jotta voin vertailla kahta ryhmää tilastollisesti.
- Aikasarjakuvaajat antavat ymmärryksen pandemian vaikutuksista ja toipumisesta.
- Korrelaatio havainnollistaa, kuinka synkronisesti suuret lentoasemat liikkuvat.
- Levene + t-test (tai Mann–Whitney U) paljastavat, onko keskimääräisissä kuukausimäärissä merkitseviä eroja ryhmien välillä ( $\alpha = 0,05$ ).
- Ennustemalli (jos toteutetaan) näyttää, kuinka hyvin menneet havainnot ennustavat tulevia arvoja.

*” Suunnitelma hyväksyttiin maanantaina 2. kesäkuuta 2025 Petteri Muuruvirran toimesta.”*

## 2. TOTEUTUS

### 2.1 Johdanto ja tavoite

Tässä projektissa tarkastellaan Suomen lentoasemien kuukausittaisia matkustajamääriä vuosilta 2019–2024. Aineisto on ladattu Tilastokeskuksen PX-Web-palvelusta *Tilastokeskuksen maksuttomat tilastotietokannat*, ja se kattaa useiden eri lentoasemien matkustajamäärät kuukausitasolla.

Analyysin tavoitteena on selvittää, miten matkustajamäärät ovat muuttuneet erityisesti koronapandemian aikana ja sen jälkeen. Lisäksi pyritään tunnistamaan, esiintyykö eroja suurten ja pienten lentoasemien välillä sekä mahdollisia kausivaihteita. Saatuja tuloksia voidaan hyödyntää esimerkiksi lentoasemien resurssien suunnittelussa ja päätöksenteossa.

## 2.2 Datan lataus ja esikäsittely

Valitse taulukko
Valitse muuttujat
Näytä taulukko

12ib – Kotimaan lentoasemien matkustajamäärät ja rahtitonnit kuukausittain, 2019M01-2025M04

### Valitse muuttujat

▼ Tietoja taulukosta
Listanäkymä

**Tiedot** Pakollinen\*

☒ Valitse kaikki

☐ Poista valinnat

---

Valittu 4 Yhteensä 4

Matkustajamäärä
Rahti ja posti yhteensä, tonnia
Matkustajamäärä, kumulatiivinen vuoden alusta
Rahti ja posti, kumulatiivinen vuoden alusta

**Kuukausi** Pakollinen\*

☒ Valitse kaikki

☐ Poista valinnat

---

☐ Sanan alusta

Hae

---

Valittu 76 Yhteensä 76

2019M06
2019M05
2019M04
2019M03
2019M02
2019M01

**Ilmoittava lentoasema**

☒ Valitse kaikki

☐ Poista valinnat

---

☐ Sanan alusta

Hae

---

Valittu 21 Yhteensä 21

<input checked="" type="checkbox"/> Valinnainen muuttuja
Pori
Rovaniemi
Savonlinna
Tampere-Pirkkala
Turku
Vaasa

**Lennon tyyppi**

☒ Valitse kaikki

☐ Poista valinnat

---

Valittu 3 Yhteensä 3

<input checked="" type="checkbox"/> Valinnainen muuttuja
Yhteensä
Reittilento
Tilauslento

**Saapuneet/lähteneet**

☒ Valitse kaikki

☐ Poista valinnat

---

Valittu 3 Yhteensä 3

<input checked="" type="checkbox"/> Valinnainen muuttuja
Saapuneet/lähteneet yhteensä
Saapuneet
Lähteneet

**Toinen lentoasema**

☒ Valitse kaikki

☐ Poista valinnat

---


Valittu 4 Yhteensä 4


<input checked="" type="checkbox"/> Valinnainen muuttuja
Yhteensä
Helsinki-Vantaa
Muut kotimaan lentoasemat
Kansainvälinen


Näytä taulukko

Muuttujien valinta Tilastokeskuksen PX-Web-palvelussa: Kuvassa näkyy Tilastokeskuksen PX-Web-verkkopalvelun näkymä, jossa valitaan analyysiin tarvittavat muuttujat. Tällä sivulla käyttäjä voi rajata, mitä tietoja halutaan mukaan ladattavaan aineistoon. Tässä projektissa valittiin muun muassa matkustajamäärä, halutut kuukaudet, lentoasemat sekä lennon tyyppi ja muut olennaiset kentät. Näin varmistetaan, että ladattava aineisto sisältää juuri ne tiedot, joita analyysissä tarvitaan.

Valintojen jälkeen data ladataan koneelle esimerkiksi JSON-tiedostona, joka on jatkokäsittelyn lähtökohta Pythonilla.

  
Valitse taulukko

  
Valitse muuttujat

  
Näytä taulukko

## Haun tulos

▼ Tietoja taulukosta

▼ Vaihda esitysmuotoa

▼ Muuta ja laske

^ Lataa taulukko

☐ Lataa PC-Axis-tiedosto (px)

☐ Excel-työkirja (xml)

☐ Excel-työkirja (xml) (koodi ja teksti)

☐ Sarkaineroitettu (otsikollinen)

☐ Sarkaineroitettu (otsikoton)

☐ Plikkueroitettu (otsikollinen)

☐ Plikkueroitettu (otsikoton)

☐ Väilyöntieroitettu (otsikollinen)

☐ Väilyöntieroitettu (otsikoton)

☐ Lataa puolipiste-eroitettu csv-tiedosto (otsikollinen)

☐ Puolipiste-eroitettu (otsikoton)

☐ HTML-tiedosto (htm)

☐ Relaatiotiedosto (txt)

☐ Lataa Excel-tiedosto (xlsx)

☐ Excel (xlsx) (koodi ja teksti sarakekellia)

☐ JSON-stat-tiedosto (json)

☒ JSON-stat2-tiedosto (json)


☐ Html5-tilausta (html)

☐ Json file (json)


Tallenna


▼ Tallenna poiminta


▼ Taulukon asetukset





**Tämä taulukko on työstetty**  
 Huomautus: taulukkoa on työstetty (ainoastaan näytöllä), koska taulukon enimmäiskoko (1000 riviä ja 30 saraketta) on ylittetty.


 Käännä manuaalisesti


 Käännä myötäpäivään

 Kokoruututila

 Käännä vastapäivään

 Taulukkonäkymä 1

 Pylväskuvio

 Viivakuvi

**Kotimaan lentoasemien matkustajamäärät ja rahtitonnit muuttujina Kuukausi, Ilmoittava lentoasema, Lennon tyyppi, Saa**

	Yhteensä				Helsinki-Var		
	Matkustajamäärä	Rahti ja postin yhteensä, tonnia	Matkustajamäärä, kumulatiivinen vuoden alusta	Rahti ja postin, kumulatiivinen vuoden alusta, tonnia	Matkustajamäärä	Rahti ja postin yhteensä, tonnia	Matkustajamäärä
2018M01							
Yhteensä							
Yhteensä							
Saapuneet/ lähteneet yhteensä	1 808 108	16 265	1 808 108	16 265	276 832	55	
Saapuneet	884 340	7 550	884 340	7 550	128 975	41	
Lähteneet	923 828	8 715	923 828	8 715	146 857	13	
Reittilento							
Saapuneet/ lähteneet yhteensä	1 857 115	14 743	1 857 115	14 743	276 648	55	
Saapuneet	831 756	6 733	831 756	6 733	128 718	41	
Lähteneet	925 358	8 010	925 358	8 010	146 930	13	
Tilauslento							
Saapuneet/ lähteneet yhteensä	112 054	1 522	112 054	1 522	283	0	
Saapuneet	52 584	817	52 584	817	256	0	

Taulukkomuotoinen esikatselu ja tiedoston lataus: Tässä ruutukaappauksessa nähdään, miltä valitut tiedot näyttävät Tilastokeskuksen PX-Web-palvelun taulukkomuodossa. Taulukossa on rivejä esimerkiksi eri lentoasemien, kuukausien ja tietotyyppien mukaan – kuten “Saapuneet/lähteneet yhteensä”, “Saapuneet” ja “Lähteneet”.

Vasemmasta reunasta voi valita ladattavan tiedoston tiedostomuodon, esimerkiksi JSON-stat2 tai Excel. Tässä projektissa data ladattiin JSON-muodossa, joka mahdollistaa automaattisen käsittelyn Pythonilla.



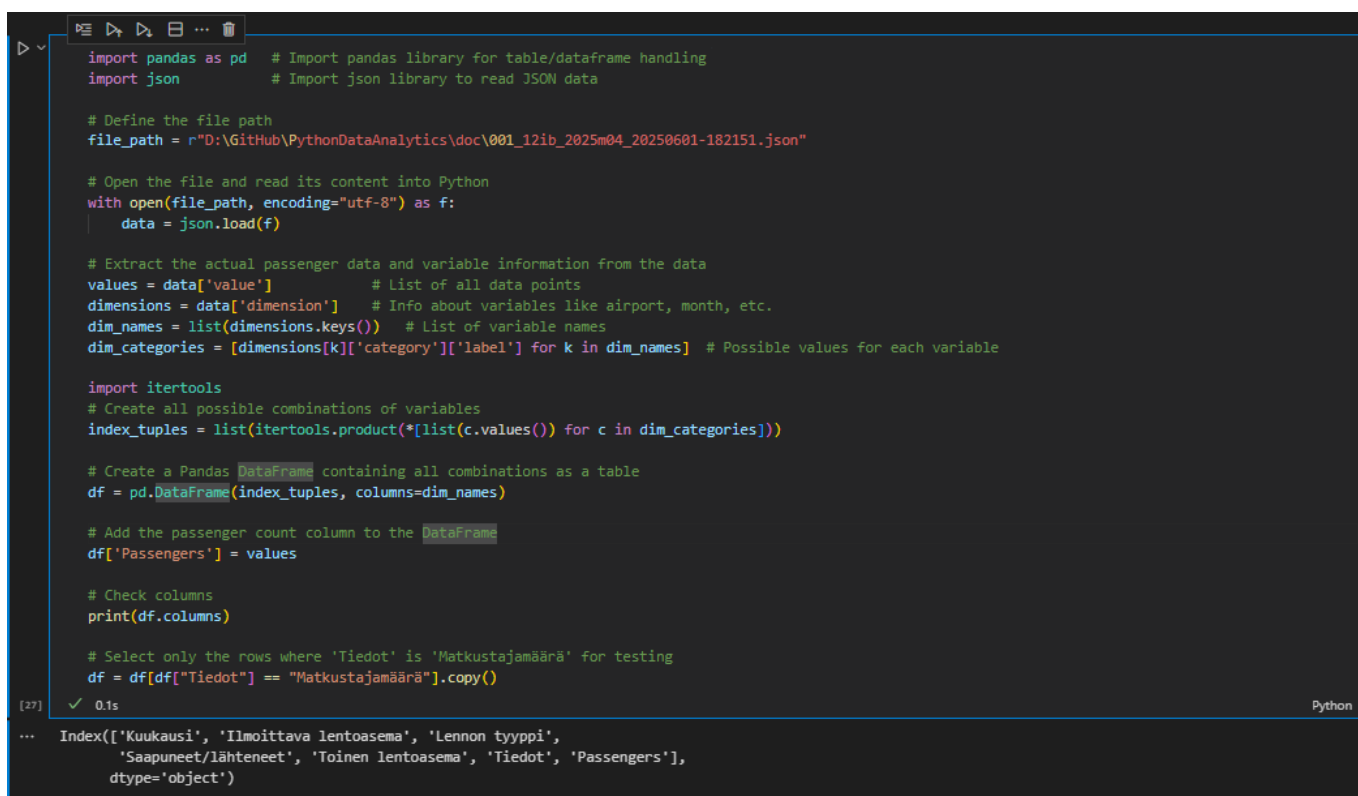


lentoasemia keskenään. Näin esikäsitelty data on valmis tarkempaan tutkimukseen ja visualisointien tekemiseen.

Pythonin tehokkaat kirjastot, kuten Pandas ja Matplotlib, mahdollistavat datan puhdistamisen, muuntamisen ja analysoinnin nopeasti ja joustavasti. Näiden työkalujen ansiosta monimutkainen raakadata voidaan muuttaa helposti analysoitavaan muotoon ja esittää tulokset selkeästi kaavioina ja taulukoina.

Tässä raportissa tulen tekemään erilaisia data-analyyysejä, vertailemaan matkustajamäärien kehitystä, visualisoimaan tuloksia kaavioiden avulla sekä suorittamaan tilastollisia testejä esimerkiksi suurten ja pienten lentoasemien välillä. Näin saadaan kokonaisvaltainen käsitys siitä, miten matkustajamäärät ovat muuttuneet viime vuosina ja millaisia eroja eri lentoasemien välillä esiintyy.

## 2.2.1 JSON-aineiston lukeminen ja muuntaminen taulukkomuotoon



```
import pandas as pd # Import pandas library for table/dataframe handling
import json         # Import json library to read JSON data

# Define the file path
file_path = r"D:\GitHub\PythonDataAnalytics\doc\001_12ib_2025m04_20250601-182151.json"

# Open the file and read its content into Python
with open(file_path, encoding="utf-8") as f:
    data = json.load(f)

# Extract the actual passenger data and variable information from the data
values = data['value'] # List of all data points
dimensions = data['dimension'] # Info about variables like airport, month, etc.
dim_names = list(dimensions.keys()) # List of variable names
dim_categories = [dimensions[k]['category']['label'] for k in dim_names] # Possible values for each variable

import itertools
# Create all possible combinations of variables
index_tuples = list(itertools.product(*[list(c.values()) for c in dim_categories]))

# Create a Pandas DataFrame containing all combinations as a table
df = pd.DataFrame(index_tuples, columns=dim_names)

# Add the passenger count column to the DataFrame
df['Passengers'] = values

# Check columns
print(df.columns)

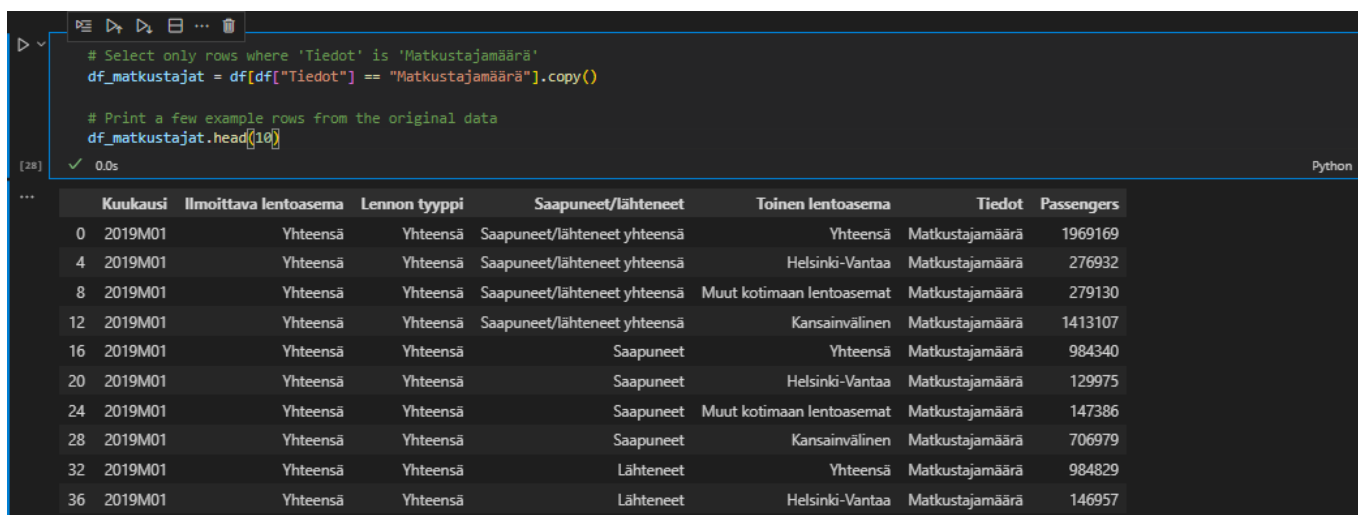
# Select only the rows where 'Tiedot' is 'Matkustajamäärä' for testing
df = df[df["Tiedot"] == "Matkustajamäärä"].copy()
```

[27] ✓ 0.1s Python

```
... Index(['Kuukausi', 'Ilmoittava lentoasema', 'Lennon tyyppi',
        'Saapuneet/lähteneet', 'Toinen lentoasema', 'Tiedot', 'Passengers'],
        dtype='object')
```

Tässä vaiheessa data haetaan Tilastokeskuksen verkkosivulta ja muutetaan sellaiseen muotoon, että tietokone ymmärtää sen taulukkona. Koodissa avataan ensin tiedosto ja tuodaan tiedot Python-ohjelmaan. Sitten tiedot järjestellään riveiksi ja sarakkeiksi, jolloin jokaiselle lentoasemalle, kuukaudelle ja tietotyypille (esim. matkustajamäärä) tulee oma rivinsä. Näin raakadata saadaan analyysia varten valmiiksi.

## 2.2.2 Datan siivous ja muokkaaminen analyysia varten



```

# Select only rows where 'Tiedot' is 'Matkustajamäärä'
df_matkustajat = df[df["Tiedot"] == "Matkustajamäärä"].copy()

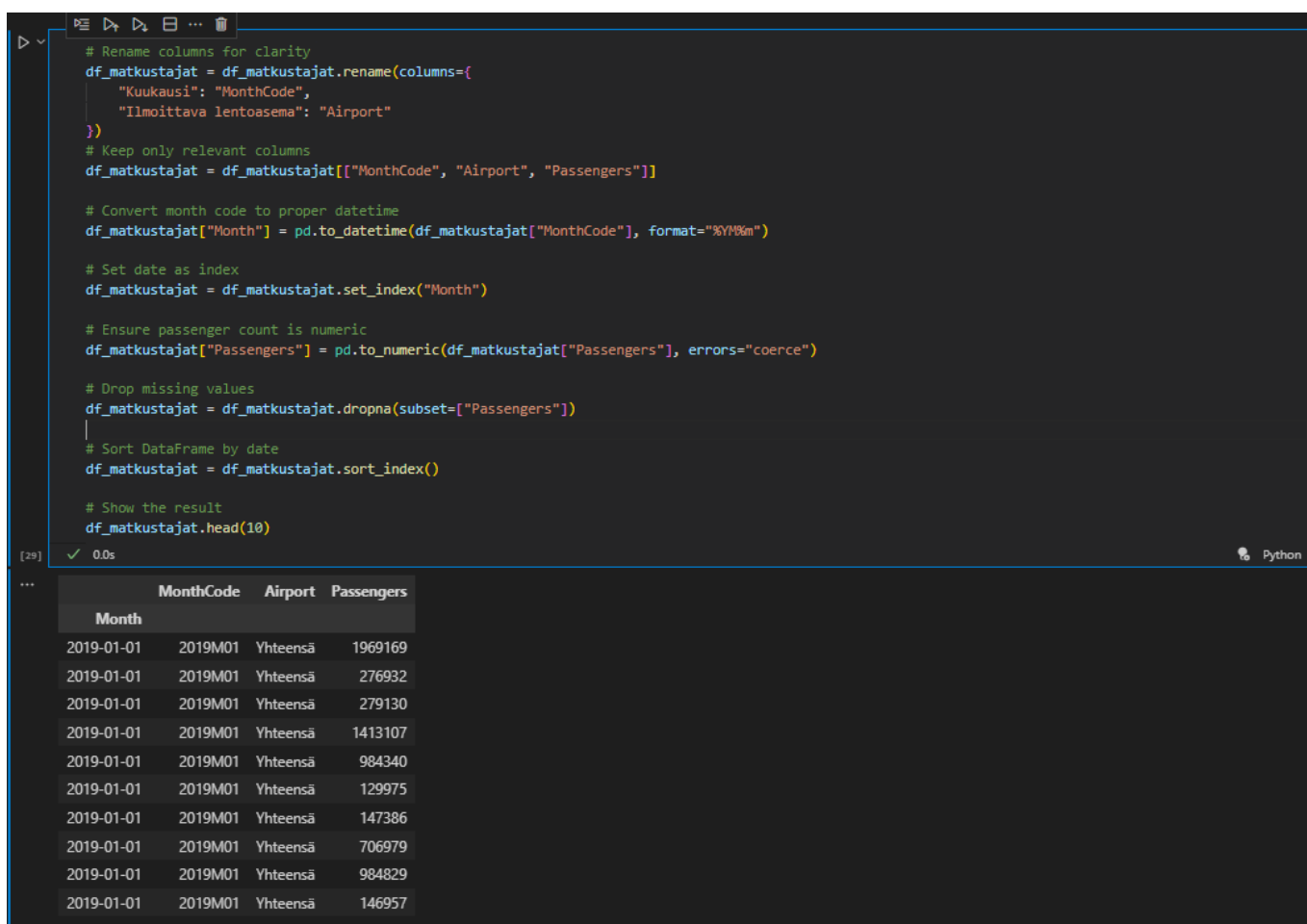
# Print a few example rows from the original data
df_matkustajat.head(10)

```

	Kuukausi	Ilmoittava lentoasema	Lennon tyyppi	Saapuneet/lähteneet	Toinen lentoasema	Tiedot	Passengers
0	2019M01	Yhteensä	Yhteensä	Saapuneet/lähteneet yhteensä	Yhteensä	Matkustajamäärä	1969169
4	2019M01	Yhteensä	Yhteensä	Saapuneet/lähteneet yhteensä	Helsinki-Vantaa	Matkustajamäärä	276932
8	2019M01	Yhteensä	Yhteensä	Saapuneet/lähteneet yhteensä	Muut kotimaan lentoasemat	Matkustajamäärä	279130
12	2019M01	Yhteensä	Yhteensä	Saapuneet/lähteneet yhteensä	Kansainvälinen	Matkustajamäärä	1413107
16	2019M01	Yhteensä	Yhteensä	Saapuneet	Yhteensä	Matkustajamäärä	984340
20	2019M01	Yhteensä	Yhteensä	Saapuneet	Helsinki-Vantaa	Matkustajamäärä	129975
24	2019M01	Yhteensä	Yhteensä	Saapuneet	Muut kotimaan lentoasemat	Matkustajamäärä	147386
28	2019M01	Yhteensä	Yhteensä	Saapuneet	Kansainvälinen	Matkustajamäärä	706979
32	2019M01	Yhteensä	Yhteensä	Lähteneet	Yhteensä	Matkustajamäärä	984829
36	2019M01	Yhteensä	Yhteensä	Lähteneet	Helsinki-Vantaa	Matkustajamäärä	146957

Alkuperäinen aineisto sisältää paljon muutakin tietoa kuin pelkät matkustajamäärät, kuten rahtiluvut ja vuoden alusta kertyneet summat. Tässä vaiheessa data siivotaan niin, että jäljelle jää vain kiinnostava tieto eli kuukausittaiset matkustajamäärät lentoasemittain. Näin analyysista tulee selkeämpi ja lopputulokset ovat helpommin ymmärrettävissä.

## 2.2.3 Taulukon yksinkertaistaminen ja siistiminen analyysiä varten



```

# Rename columns for clarity
df_matkustajat = df_matkustajat.rename(columns={
    "Kuukausi": "MonthCode",
    "Ilmoittava lentoasema": "Airport"
})

# Keep only relevant columns
df_matkustajat = df_matkustajat[["MonthCode", "Airport", "Passengers"]]

# Convert month code to proper datetime
df_matkustajat["Month"] = pd.to_datetime(df_matkustajat["MonthCode"], format="%Y%M")

# Set date as index
df_matkustajat = df_matkustajat.set_index("Month")

# Ensure passenger count is numeric
df_matkustajat["Passengers"] = pd.to_numeric(df_matkustajat["Passengers"], errors="coerce")

# Drop missing values
df_matkustajat = df_matkustajat.dropna(subset=["Passengers"])

# Sort DataFrame by date
df_matkustajat = df_matkustajat.sort_index()

# Show the result
df_matkustajat.head(10)

```

	MonthCode	Airport	Passengers
Month			
2019-01-01	2019M01	Yhteensä	1969169
2019-01-01	2019M01	Yhteensä	276932
2019-01-01	2019M01	Yhteensä	279130
2019-01-01	2019M01	Yhteensä	1413107
2019-01-01	2019M01	Yhteensä	984340
2019-01-01	2019M01	Yhteensä	129975
2019-01-01	2019M01	Yhteensä	147386
2019-01-01	2019M01	Yhteensä	706979
2019-01-01	2019M01	Yhteensä	984829
2019-01-01	2019M01	Yhteensä	146957

Tässä vaiheessa taulukosta poistetaan kaikki turhat tiedot ja jätetään mukaan vain kolme tärkeintä asiaa: kuukausi, lentoasema ja matkustajamäärä. Kuukausitieto muunnetaan koneen ymmärtämään aikamuotoon, ja puutteelliset

*tai virheelliset rivit siivotaan pois. Näin varmistetaan, että data on selkeä ja valmis tarkempaan tarkasteluun ja vertailuun eri lentoasemien ja aikajaksojen välillä.*

## 2.3 Datan analyysi

Kun aineisto on ensin saatu valmiiksi ja siivottu, voin siirtyä itse analyysiin. Tässä osiossa vertaillaan lentoasemien matkustajamääriä eri kuukausina ja vuosina sekä tutkitaan, miten esimerkiksi koronapandemia on vaikuttanut lentomatkustukseen Suomessa. Lisäksi analyysissä selvitetään, poikkeavatko suuret lentoasemat pienistä, ja esiintyykö matkustajamäärissä kausivaihteluita. Eri vaiheissa käytetään sekä kuvaajia että tilastollisia testejä, jotta tulokset ovat helposti ymmärrettäviä ja perusteltuja. Tavoitteena on löytää selkeitä vastauksia siihen, miten matkustajamäärät ovat muuttuneet ja mitkä tekijät niihin mahdollisesti vaikuttavat.

### 2.3.1 Lentoasemien ryhmittely vertailua varten

Jotta analyysissä voidaan vertailla isojen ja pienten lentoasemien eroja, ryhmitellään kaikki kentät kahteen kategoriaan. Isoihin lentoasemiin luetaan tässä Helsinki-Vantaa ja Oulu, pieniin taas Kuopio, Rovaniemi ja muut maakuntakentät. Näin voidaan myöhemmin helposti vertailla, miten matkustajamäärät käyttäytyvät eri kokoluokan kentillä.

```
# Define the list of large airports
large_airports = ["Helsinki-Vantaa", "Oulu"]

# Add a new column indicating whether the airport is "Large" or "Small"
df_matkustajat["AirportGroup"] = df_matkustajat["Airport"].apply(
    lambda x: "Large" if x in large_airports else "Small"
)

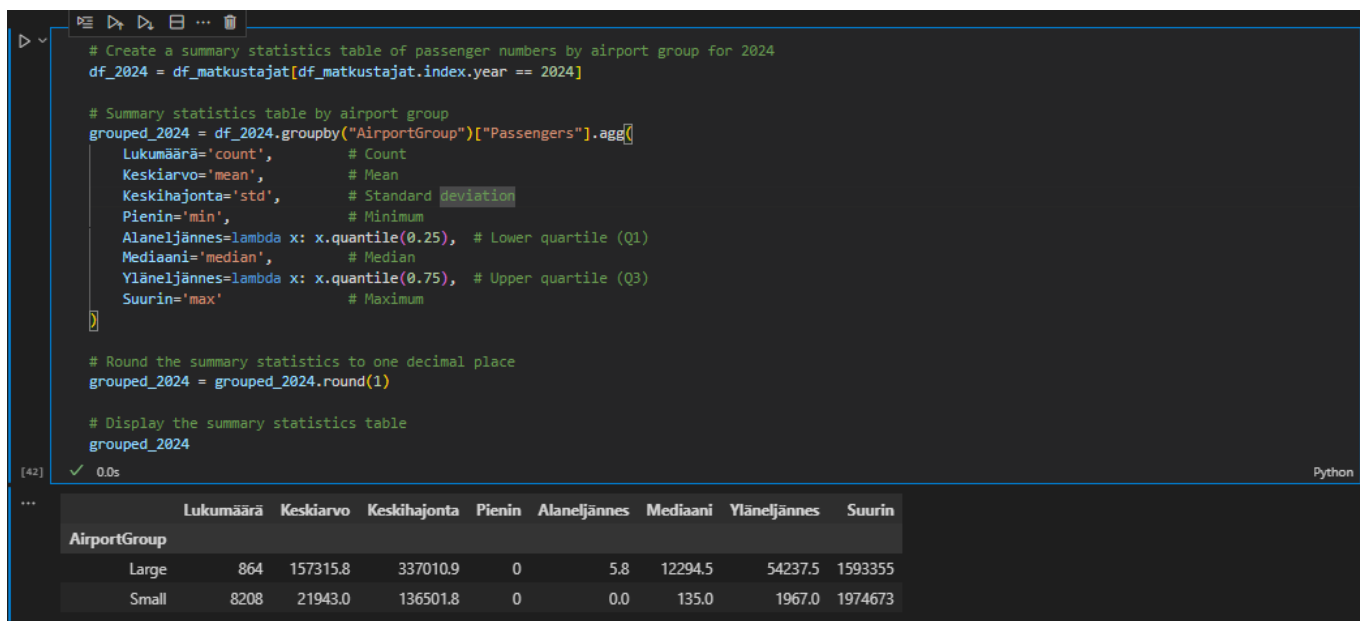
# Show examples of the grouping
df_matkustajat[["Airport", "AirportGroup"]].drop_duplicates().head(10)
```

Month	Airport	AirportGroup
2019-01-01	Yhteensä	Small
2019-01-01	Helsinki-Vantaa	Large
2019-01-01	Enontekiö	Small
2019-01-01	Ivalo	Small
2019-01-01	Joensuu	Small
2019-01-01	Jyväskylä	Small
2019-01-01	Kajaani	Small
2019-01-01	Kemi-Tornio	Small
2019-01-01	Kittilä	Small
2019-01-01	Kokkola-Pietarsaari	Small

*Tässä taulukossa näkyy esimerkkejä siitä, miten lentoasemat on jaettu kahteen ryhmään. Kaikki Helsinki-Vantaan ja Oulun tiedot kuuluvat "Large"-ryhmään, kun taas muut kentät, kuten Kuopio ja Rovaniemi, kuuluvat "Small"-ryhmään. Tämä jako helpottaa myöhempää vertailua ja tilastollisia testejä.*

### 2.3.2 Tunnusluvut: Suuret ja pienet lentoasemat vuonna 2024

Tässä taulukossa esitetään suurten ("Large") ja pienten ("Small") lentoasemien kuukausittaisten matkustajamäärien tunnusluvut vuodelta 2024. Tunnusluvuista näkee mm. keskiarvon, mediaanin, vaihteluvälin sekä havaintojen määrän kummassakin ryhmässä.



```
# Create a summary statistics table of passenger numbers by airport group for 2024
df_2024 = df_matkustajat[df_matkustajat.index.year == 2024]

# Summary statistics table by airport group
grouped_2024 = df_2024.groupby("AirportGroup")["Passengers"].agg(
    Lukumäärä='count',          # Count
    Keskiarvo='mean',          # Mean
    Keskihajonta='std',        # Standard Deviation
    Pienin='min',              # Minimum
    Alaneljännes=lambda x: x.quantile(0.25), # Lower quartile (Q1)
    Mediaani='median',         # Median
    Yläneljännes=lambda x: x.quantile(0.75), # Upper quartile (Q3)
    Suurin='max'               # Maximum
)

# Round the summary statistics to one decimal place
grouped_2024 = grouped_2024.round(1)

# Display the summary statistics table
grouped_2024
```

AirportGroup	Lukumäärä	Keskiarvo	Keskihajonta	Pienin	Alaneljännes	Mediaani	Yläneljännes	Suurin
Large	864	157315.8	337010.9	0	5.8	12294.5	54237.5	1593355
Small	8208	21943.0	136501.8	0	0.0	135.0	1967.0	1974673

### 2.3.3 Tunnusluvut: Yksittäiset lentoasemat vuonna 2024

Tässä taulukossa esitetään yksittäisten lentoasemien kuukausittaisten matkustajamäärien tunnusluvut vuodelta 2024. Taulukko mahdollistaa nopean vertailun eri kenttien välillä.

```

# Calculate summary statistics for each airport in 2024
airports_summary_2024 = df_2024.groupby("Airport")["Passengers"].agg(
    Lukumäärä='count',          # Count of records
    Keskiarvo='mean',          # Mean passenger count
    Keskihajonta='std',        # Standard deviation
    Pienin='min',              # Minimum value
    Alaneljännes=lambda x: x.quantile(0.25), # Lower quartile (Q1)
    Mediaani='median',         # Median value
    Yläneljännes=lambda x: x.quantile(0.75), # Upper quartile (Q3)
    Suurin='max'               # Maximum value
)
airports_summary_2024 = airports_summary_2024.round(1) # Round results to one decimal
airports_summary_2024

```

Airport	Lukumäärä	Keskiarvo	Keskihajonta	Pienin	Alaneljännes	Mediaani	Yläneljännes	Suurin
Enontekiö	432	709.8	3291.9	0	0.0	0.0	0.0	27894
Helsinki-Vantaa	432	303840.4	429138.8	0	12.0	52655.0	624394.5	1593355
Ivalo	432	4623.7	8689.7	0	0.0	63.0	4676.5	69673
Joensuu	432	744.2	1073.6	0	0.0	10.5	1542.0	4869
Jyväskylä	432	478.1	723.0	0	0.0	13.5	880.0	3843
Kajaani	432	740.2	1106.1	0	0.0	2.0	1579.0	3995
Kemi-Tornio	432	611.4	921.5	0	0.0	0.0	1189.5	5108
Kittilä	432	7662.7	14882.5	0	4.0	900.0	8007.5	120888
Kokkola-Pietarsaari	432	569.5	842.8	0	0.0	34.0	1077.5	5086
Kuopio	432	2458.6	3443.8	0	0.0	831.5	4592.2	14481
Kuusamo	432	2352.8	4560.5	0	0.0	132.0	2561.0	37153
Lappeenranta	432	500.7	876.3	0	0.0	2.0	911.5	3539
Maarianhamina	432	736.7	887.0	0	3.8	463.5	1144.2	4924
Oulu	432	10791.3	14749.2	0	3.0	2194.0	21173.5	57187
Pori	432	232.4	382.4	0	0.0	7.0	404.5	3102
Rovaniemi	432	17548.6	31128.6	0	2.0	1012.0	24846.0	270139
Savonlinna	432	131.5	190.1	0	0.0	0.0	223.0	1101
Tampere-Pirkkala	432	3002.8	4381.4	0	0.0	499.0	5685.8	18920
Turku	432	4734.8	7018.2	0	0.0	472.0	9282.0	27233
Vaasa	432	3304.2	3783.8	0	0.0	2255.0	4772.2	17199
Yhteensä	432	365774.1	477518.6	11	42666.5	102628.5	695159.0	1974673

### 2.3.4 Ristiintaulukointi: Lentoasemaryhmät ja vuodenaajat

Tässä osiossa tutkitaan, miten matkustajamäärät jakautuvat suurten (“Large”) ja pienten (“Small”) lentoasemien välillä eri vuodenaikoina. Ristiintaulukoinnin avulla voidaan havainnollistaa, esiintyykö matkustajamäärissä selviä kausivaihteluita eri kenttäryhmien välillä.

```

def get_season(month):
    # Map month number to Finnish season name
    if month in [12, 1, 2]:
        return "Talvi" # Winter
    elif month in [3, 4, 5]:
        return "Kevät" # Spring
    elif month in [6, 7, 8]:
        return "Kesä" # Summer
    else:
        return "Syksy" # Autumn

# Add a new column 'Season' to the DataFrame based on the month
df_matkustajat["Season"] = df_matkustajat.index.month.map(get_season)

import pandas as pd

# Crosstab: AirportGroup vs. Season, sum of passengers
crosstab_season = pd.crosstab(
    df_matkustajat["AirportGroup"],
    df_matkustajat["Season"],
    values=df_matkustajat["Passengers"],
    aggfunc="sum"
)

print(crosstab_season)
# Print the number of unique airports in each group
print(df_matkustajat.groupby("AirportGroup")["Airport"].nunique())

```

[31] ✓ 0.0s Python

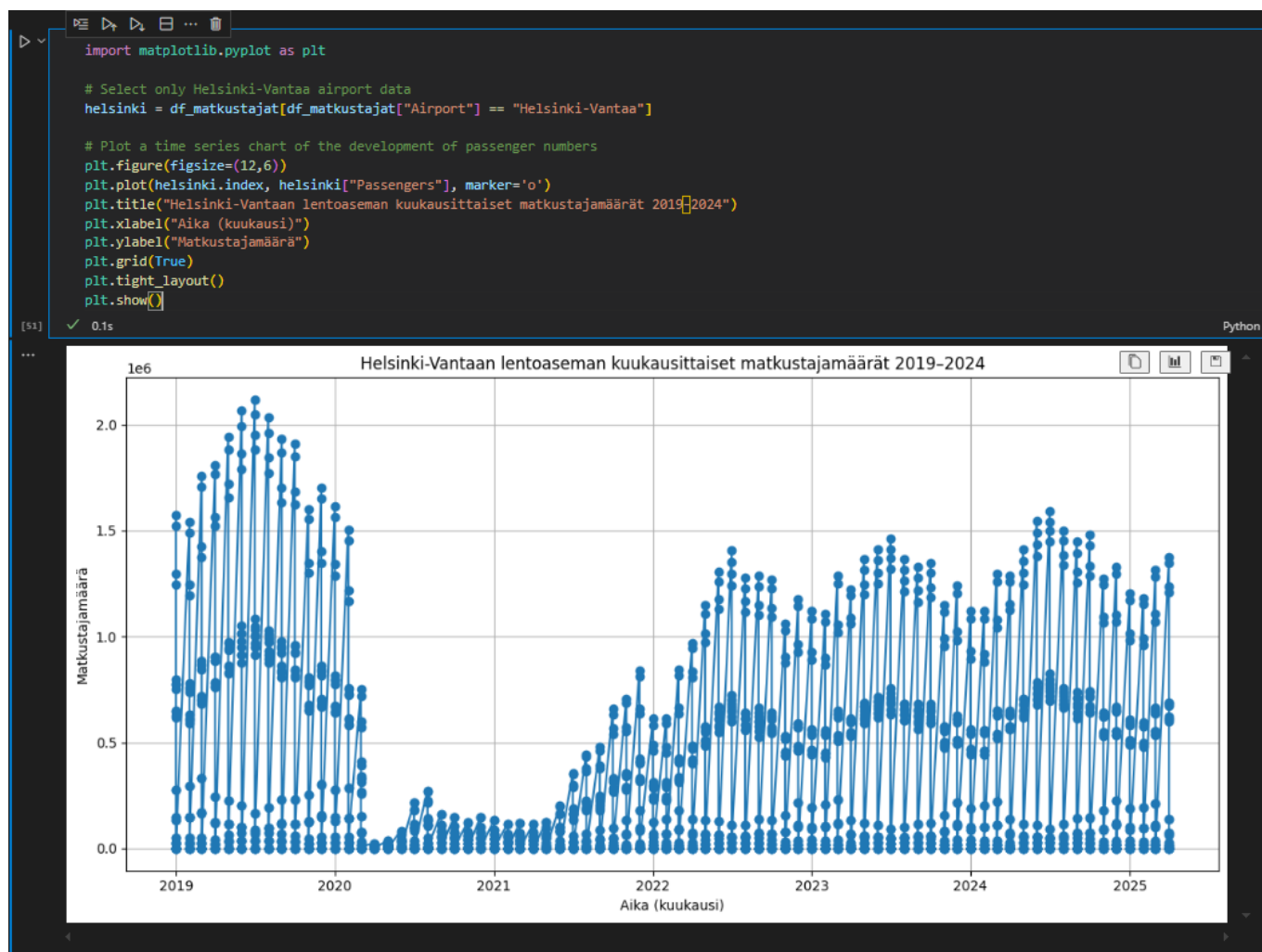
Season	Kesä	Kevät	Syksy	Talvi
AirportGroup				
Large	171588952	169045416	162863584	175649552
Small	200421816	218716184	202585024	275065216
AirportGroup				
Large	2			
Small	19			

Name: Airport, dtype: int64

Taulukosta nähdään, kuinka suuri osa matkustajamäärästä kohdistuu suuriin ja pieniin lentoasemiin eri vuodenaikoina. Esimerkiksi talvikaudella Lapin kenttien matkustajamäärät kasvavat, kun taas kesällä matkustus voi painottua enemmän Etelä-Suomen ja isojen kenttien kautta. Tällainen ristiintaulukointi auttaa hahmottamaan kausivaihteluita ja resurssien suunnittelutarpeita.

Pienten lentoasemien ("Small") yhteenlaskettu matkustajamäärä voi olla suurempi kuin suurten kenttien ("Large"), koska Small-ryhmään kuuluu suuri määrä maakuntakenttiä eri puolilta Suomea. Yksittäinen suuri lentoasema, kuten Helsinki-Vantaa, on vilkkaampi kuin yksittäinen pieni kenttä, mutta kun kaikki pienet maakuntakentät lasketaan yhteen, niiden yhteismatkustajamäärä saattaa ylittää suurten kenttien kokonaismäärän.

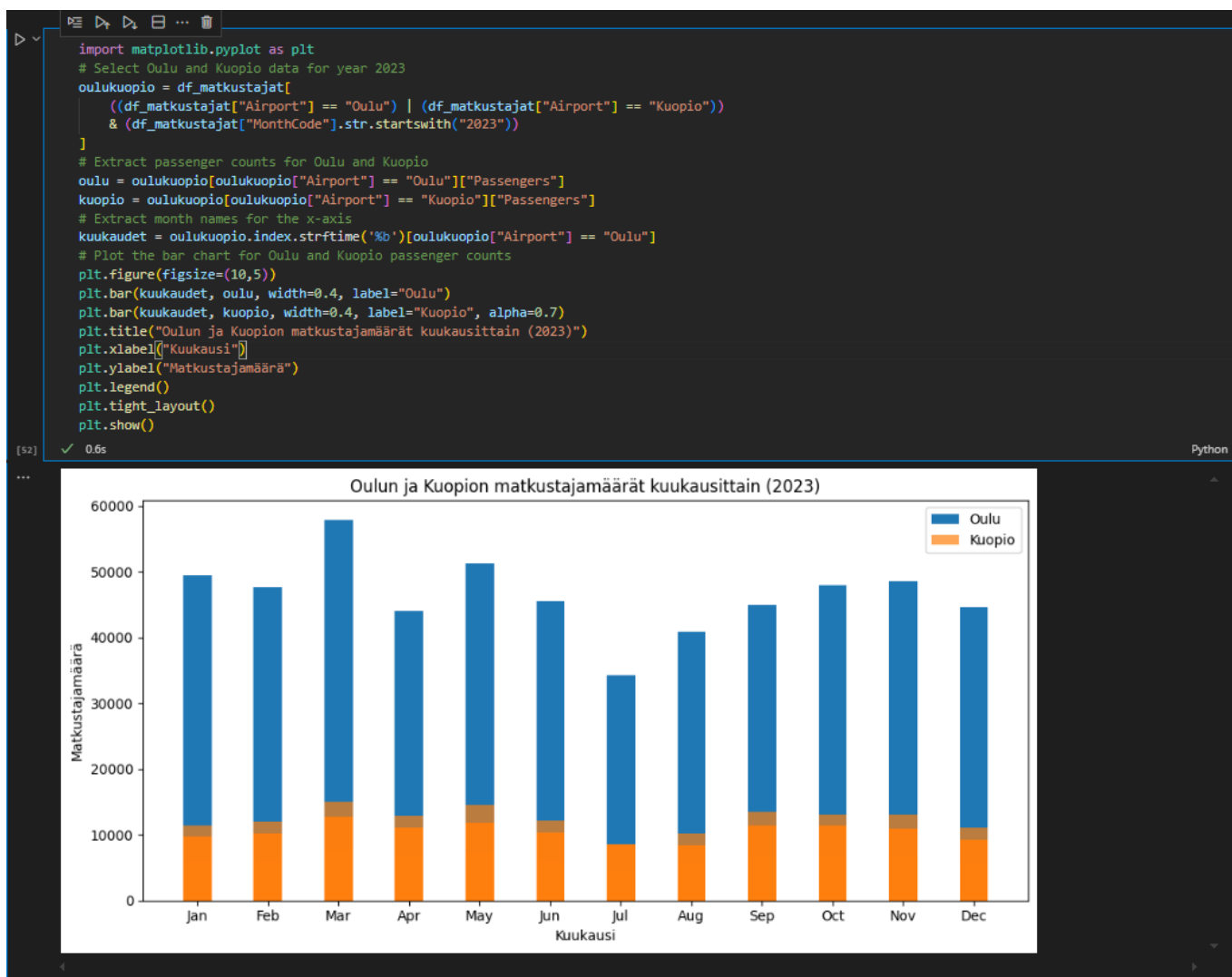
### 2.3.5 Aikasarjakuvaaja: Pandemia-ajan vaikutukset Helsinki-Vantaan lentoasemalla



Huom: Kuvaajan pystyakseli (matkustajamäärä) alkaa nolasta, jotta muutokset näkyvät selvästi ja vertailu on mahdollisimman havainnollista.

Kuvaajasta nähdään, että matkustajamäärät laskivat erittäin voimakkaasti vuoden 2020 alussa pandemian vaikutuksesta. Seuraavina vuosina määrät ovat vähitellen kasvaneet, mutta pandemian aikaiset rajoitukset ja niiden purku näkyvät vielä selvästi useiden vuosien ajan. Kuvasta on helppo nähdä sekä pandemia-ajan romahdus että hiljainen palautuminen.

### 2.3.6 Kausivaihtelun vertailu pylväsdiagrammilla (Oulu ja Kuopio 2023)

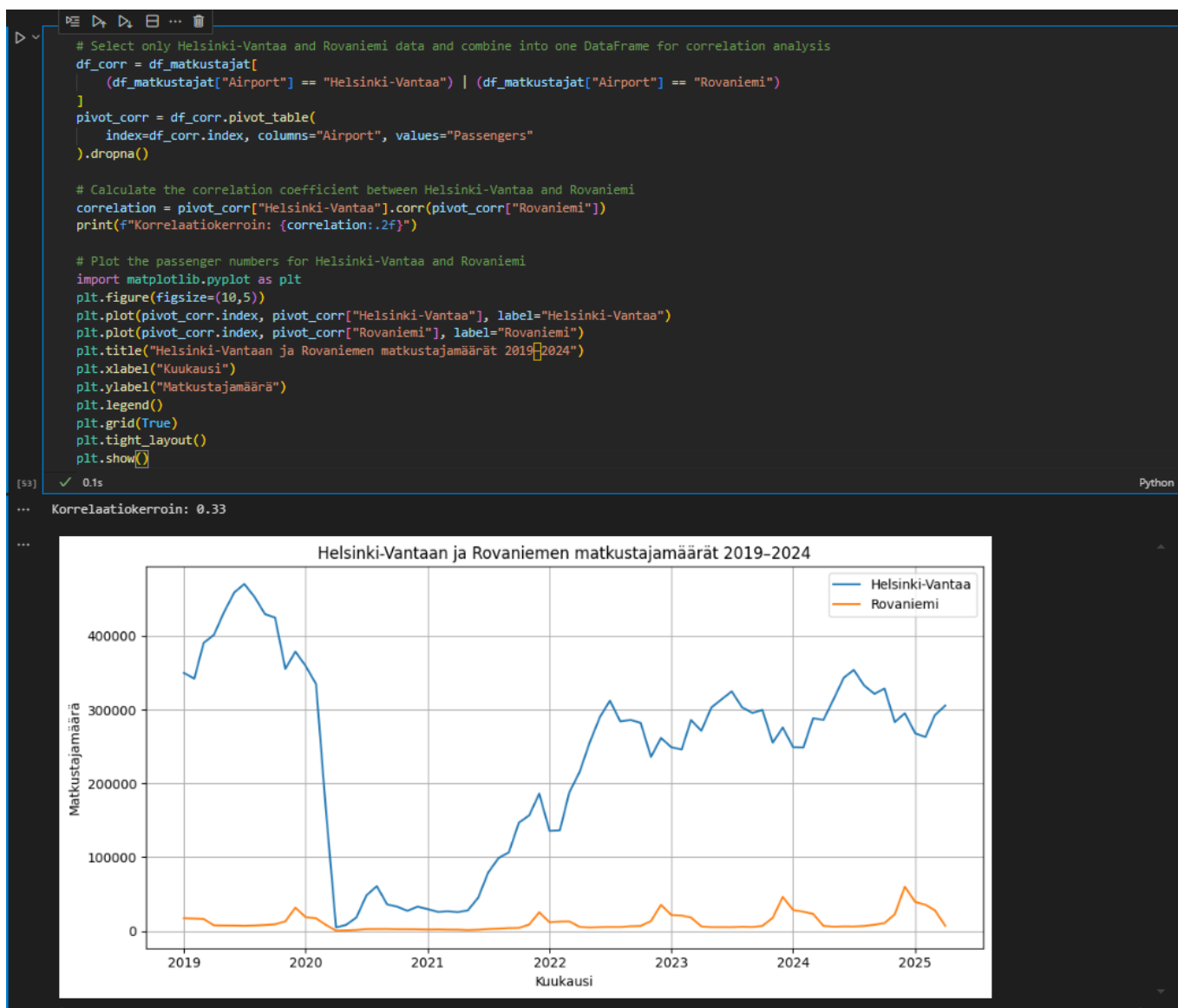


Pylväsdiagrammista nähdään, miten Oulun ja Kuopion lentoasemien matkustajamäärät vaihtelevat vuoden aikana. Kuvaajasta voi havaita esimerkiksi, onko kesäkuukausina enemmän matkustajia kuin talvella ja ovatko molempien kenttien vaihtelut saman suuntaisia. Tämä auttaa tunnistamaan kausivaihte-luita ja vertaamaan lentoasemia keskenään.

Lisäksi voidaan huomata, että Suomessa saatetaan matkustaa enemmän omalla autolla kesäaikaan esi-merkiksi lomien ja mökkireissujen vuoksi, kun taas talvella pitkät välimatkat ja vaikeat keliolosuhteet voi-vat lisätä lentomatkustuksen suosiota erityisesti Pohjois-Suomessa.

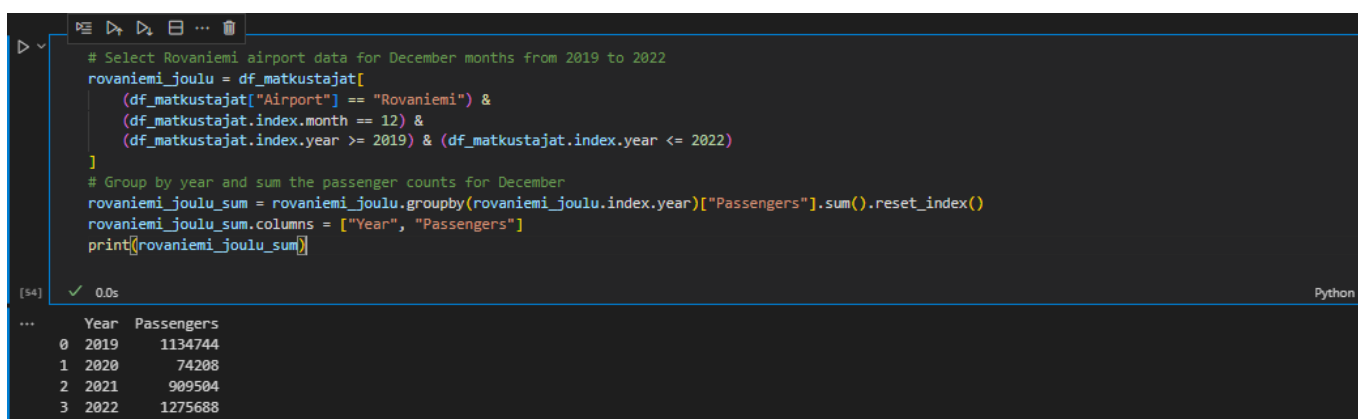


### 2.3.7 Korrelaatioanalyysi: Helsinki-Vantaa ja Rovaniemi



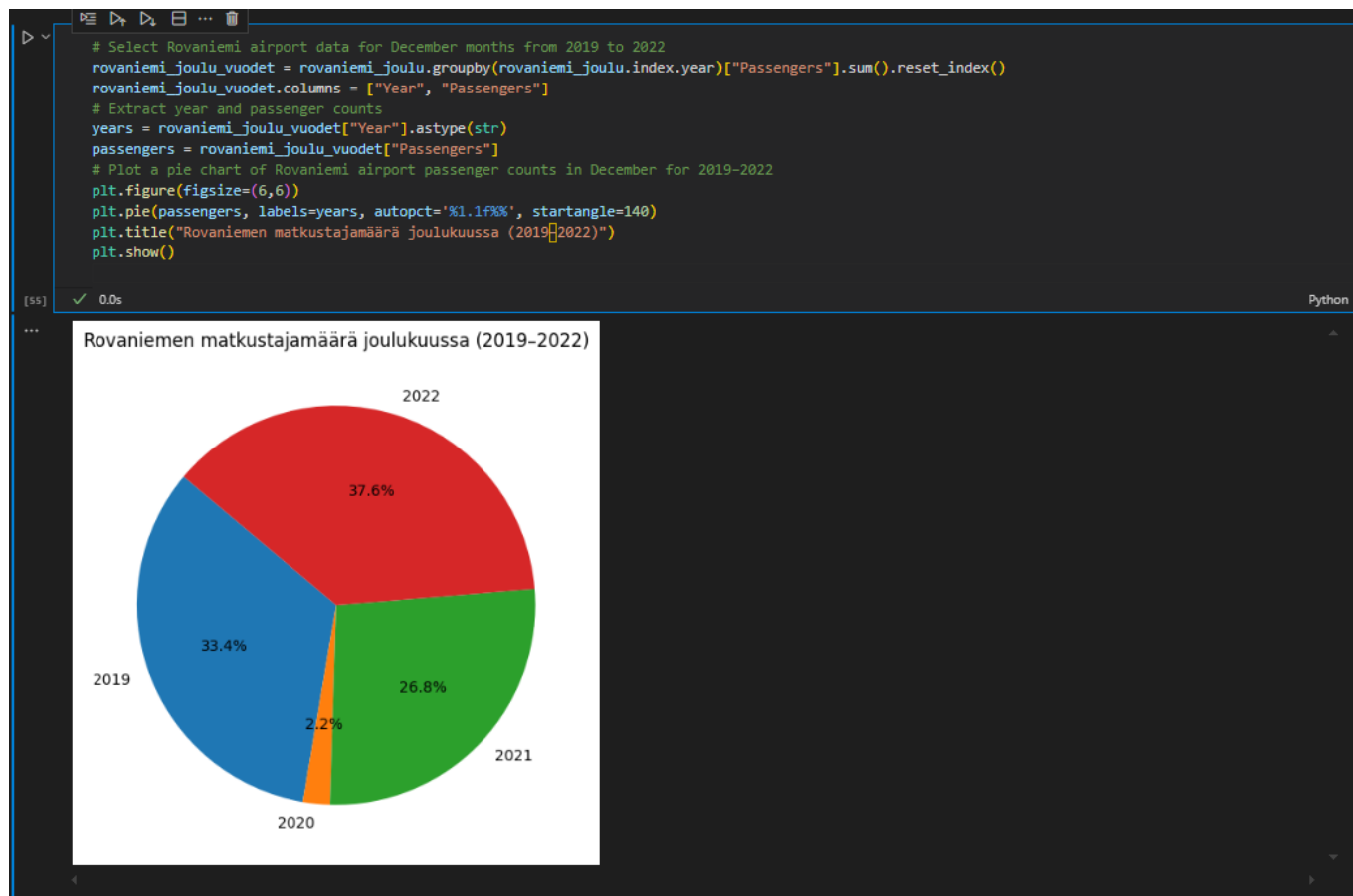
Tässä kuvassa nähdään, miten Rovaniemen ja Helsinki-Vantaan lentoasemien matkustajamäärät kehittyvät eri kuukausina ja vuosina. Korrelaatiokerroin kertoo, kuinka samankaltaisia niiden muutokset ovat: Jos luku on korkea, trendit seuraavat toisiaan, mutta Rovaniemi voi erottua selvästi kausihuipuillaan, erityisesti talven matkailusesongin aikana.

### 2.3.8 Taulukko: Rovaniemen matkustajamäärät joulukuussa (2019–2022)



Tässä taulukossa näkyvät Rovaniemen lentoaseman matkustajamäärät joulukuussa vuosina 2019–2022. Jokainen rivi kertoo, kuinka monta matkustajaa kentällä oli joulukuun aikana kyseisenä vuonna. Taulukosta nähdään, että koronavuonna 2020 matkustajamäärä romahti erittäin pieneksi, kun taas muina vuosina matkustajamäärät ovat olleet huomattavasti suurempia.

### 2.3.9 Piirakkakaavio: Rovaniemen matkustajamäärien jakauma joulukuussa (2019–2022)



Tässä piirakkakaaviossa vertaillaan joulukuun matkustajamääriä neljänä peräkkäisenä vuonna. Kaavio-osta näkyy selvästi, kuinka koronavuosi 2020 oli poikkeuksellisen hiljainen – matkustajista vain noin 2 % saapui tuona vuonna. Vuosina 2019, 2021 ja 2022 matkustajamäärät ovat olleet selvästi korkeampia, ja vuonna 2022 määrät ovat jo palautuneet lähes pandemiaa edeltäneelle tasolle. Kaavio havainnollistaa hyvin, miten pandemia vaikutti erityisesti joulukuun sesonkiin Lapissa.

## 2.4 Tilastolliset testit: Suuret vs. pienet lentoasemat

Jotta voidaan varmistaa, ovatko suurten ja pienten lentoasemien matkustajamäärät oikeasti erilaisia, tarvitaan tilastollisia testejä. Näiden testien avulla voidaan päätellä, onko havaittu ero ryhmien välillä sattumaa vai aidosti merkittävä.

Ensin tarkistetaan Levene-testillä, ovatko ryhmien vaihtelut (hajonnat) samanlaisia. Sen jälkeen vertaillaan ryhmien matkustajamäärien keskiarvoja t-testillä. Jos t-testiä ei voida käyttää, koska ryhmien hajonnat eroavat paljon toisistaan tai data ei ole normaalisti jakautunutta, käytetään Mann–Whitney U -testiä. Näiden testien avulla

saadaan selville, onko suurten ja pienten lentoasemien matkustajamäärissä tilastollisesti merkitsevää eroa vuonna 2023.

#### 2.4.1 Ryhmien muodostaminen (Large/Small)

Jotta voimme vertailla isojen ja pienten lentoasemien matkustajamääriä, jaetaan lentoasemat kahteen ryhmään: "Large" ja "Small". Isoihin kuuluvat esimerkiksi Helsinki-Vantaa ja Oulu, pieniin taas Kuopio, Rovaniemi ja muut maakuntakentät. Tämä ryhmittely tehtiin jo analyysin alussa, jotta voimme helposti vertailla ryhmiä keskenään eri vaiheissa. Tässä kohdassa keskitymme vuoden 2023 tietoihin, jotta saamme vertailuun selkeät ja ajankohtaiset ryhmät tilastollisia testejä varten.

```
# Select only data for the year 2023 for analysis
df_2023 = df_matkustajat[df_matkustajat.index.year == 2023].copy()

# Separate large (Large) and small (Small) airports
large_2023 = df_2023[df_2023["AirportGroup"] == "Large"]["Passengers"]
small_2023 = df_2023[df_2023["AirportGroup"] == "Small"]["Passengers"]

# Print the first few rows of large and small airports for 2023
print("Suuret lentoasemat (esim. Helsinki-Vantaa, Oulu):")
print(large_2023.head())
print("\nPienet lentoasemat (esim. Kuopio, Rovaniemi):")
print(small_2023.head())
```

[56] ✓ 0.0s Python

```
... Suuret lentoasemat (esim. Helsinki-Vantaa, Oulu):
Month
2023-01-01    1121174
2023-01-01         0
2023-01-01    193320
2023-01-01    927854
2023-01-01    564670
Name: Passengers, dtype: int64

Pienet lentoasemat (esim. Kuopio, Rovaniemi):
Month
2023-01-01    1459126
2023-01-01    193272
2023-01-01    194431
2023-01-01    1071423
2023-01-01     726380
Name: Passengers, dtype: int64
```

*Tässä vaiheessa vuoden 2023 lentoasematiedot on jaettu kahteen ryhmään: suuriin (Large) ja pieniin (Small) lentoasemiin. Suuriin kuuluvat esimerkiksi Helsinki-Vantaa ja Oulu, pieniin taas Kuopio, Rovaniemi ja muut maakuntakentät.*

*Taulukossa näkyy muutama esimerkkirivi kummastakin ryhmästä. Jokainen rivi kertoo kyseisen lentoaseman matkustajamäärän tietyssä kuukautena. Näiden kahden ryhmän avulla voidaan seuraavaksi vertailla, onko niiden välillä tilastollisesti merkitsevää eroa matkustajamäärissä vuoden 2023 aikana.*

#### 2.4.2 Levene-testi (varianssien vertailu)

Ennen kuin vertaillaan isojen ja pienten lentoasemien matkustajamäärien keskiarvoja, täytyy tarkistaa, ovatko ryhmien vaihtelut eli hajonnat samanlaisia. Tätä varten tehdään Levene-testi. Jos hajonnat ovat samanlaisia, voimme myöhemmin käyttää t-testiä. Jos eivät, pitää käyttää vaihtoehtoisia testejä.

```

from scipy.stats import levene

# Perform Levene's test for equal variances between large and small airports in 2023
stat_levene, p_levene = levene(large_2023, small_2023)
# Print the results of Levene's test
print(f"Levene-testin testisuure: {stat_levene:.2f}")
print(f"Levene-testin p-arvo: {p_levene:.3f}")

```

[57] ✓ 0.0s Python

... Levene-testin testisuure: 529.26  
Levene-testin p-arvo: 0.000

Levene-testin tuloksista nähdään, että p-arvo on erittäin pieni (0.000). Tämä tarkoittaa, että suurten ja pienten lentoasemien matkustajamäärien vaihtelut eli hajonnat poikkeavat tilastollisesti merkitsevästi toisistaan. Näin ollen ryhmien variansseja ei voi pitää yhtä suurina. Tämän vuoksi seuraavassa vaiheessa käytetään tavanomaisen t-testin sijaan Mann–Whitney U -testiä, joka ei edellytä ryhmien yhtä suuria hajontoja.

### 2.4.3 T-testi ja/tai Mann–Whitney U -testi (keskiarvojen/medioiden vertailu)

Kun Levene-testin perusteella ryhmien hajonnat eivät olleet yhtä suuret, käytetään t-testin sijaan Mann–Whitney U -testiä. Tämän testin avulla voidaan selvittää, onko suurten ja pienten lentoasemien kuukausittaisissa matkustajamäärissä tilastollisesti merkitsevä ero vuonna 2023 – ilman oletusta siitä, että hajonnat olisivat samanlaisia tai että jakauma olisi normaalijakautunut.

```

from scipy.stats import mannwhitneyu

# Perform the Mann-Whitney U test to compare passenger counts between large and small airports in 2023
stat_mw, p_mw = mannwhitneyu(large_2023, small_2023, alternative="two-sided")
# Print the results of the Mann-Whitney U test
print(f"Mann-Whitney U -testin testisuure: {stat_mw:.2f}")
print(f"Mann-Whitney U -testin p-arvo: {p_mw:.3f}")

```

[58] ✓ 0.0s Python

... Mann-Whitney U -testin testisuure: 4747638.00  
Mann-Whitney U -testin p-arvo: 0.000

Mann–Whitney U -testin p-arvo on 0.000, mikä tarkoittaa, että suurten ja pienten lentoasemien kuukausittaisien matkustajamäärien välillä on tilastollisesti merkitsevä ero vuonna 2023. Näin pieni p-arvo tarkoittaa, että ero ryhmien välillä ei ole sattumaa – suuret lentoasemat (kuten Helsinki-Vantaa ja Oulu) ja pienet lentoasemat (kuten Kuopio, Rovaniemi) poikkeavat selvästi toisistaan matkustajamäärissä.

Tämän perusteella voidaan päätellä, että suurten ja pienten lentoasemien välillä on merkittävä ero matkustajamäärissä vuoden 2023 aikana.

## 3. YHTEENVETO JA TULKINTA

Tässä analyysissä selvitettiin Suomen lentoasemien kuukausittaisien matkustajamäärien kehitystä ja tehtiin vertailuja suurten ja pienten lentoasemien välillä. Aika-sarjakuvaajat osoittivat selvästi koronapandemian vaikutukset: matkustajamäärät laskivat rajusti vuoden 2020 alussa, mutta ovat sen jälkeen vähitellen palautuneet. Kausivaihtelut näkyivät erityisesti Lapin lentoasemilla, kuten Rovaniemellä, jossa matkustajamäärät nousevat huomattavasti talvisesongin aikana.

Korrelaatioanalyysin perusteella suuret ja pienet lentoasemat eivät liiku täysin samassa tahdissa, vaikka pandemia vaikutti kaikkiin kenttiin. Tilastolliset testit osoittivat, että suurten ja pienten lentoasemien kuukausittaisten matkustajamäärien välillä oli tilastollisesti merkitsevä ero vuonna 2023. Näitä tuloksia voidaan hyödyntää esimerkiksi resurssien ja palveluiden suunnittelussa sekä tulevaisuuden liikennemäärien ennakkoinnissa lentoasemilla.