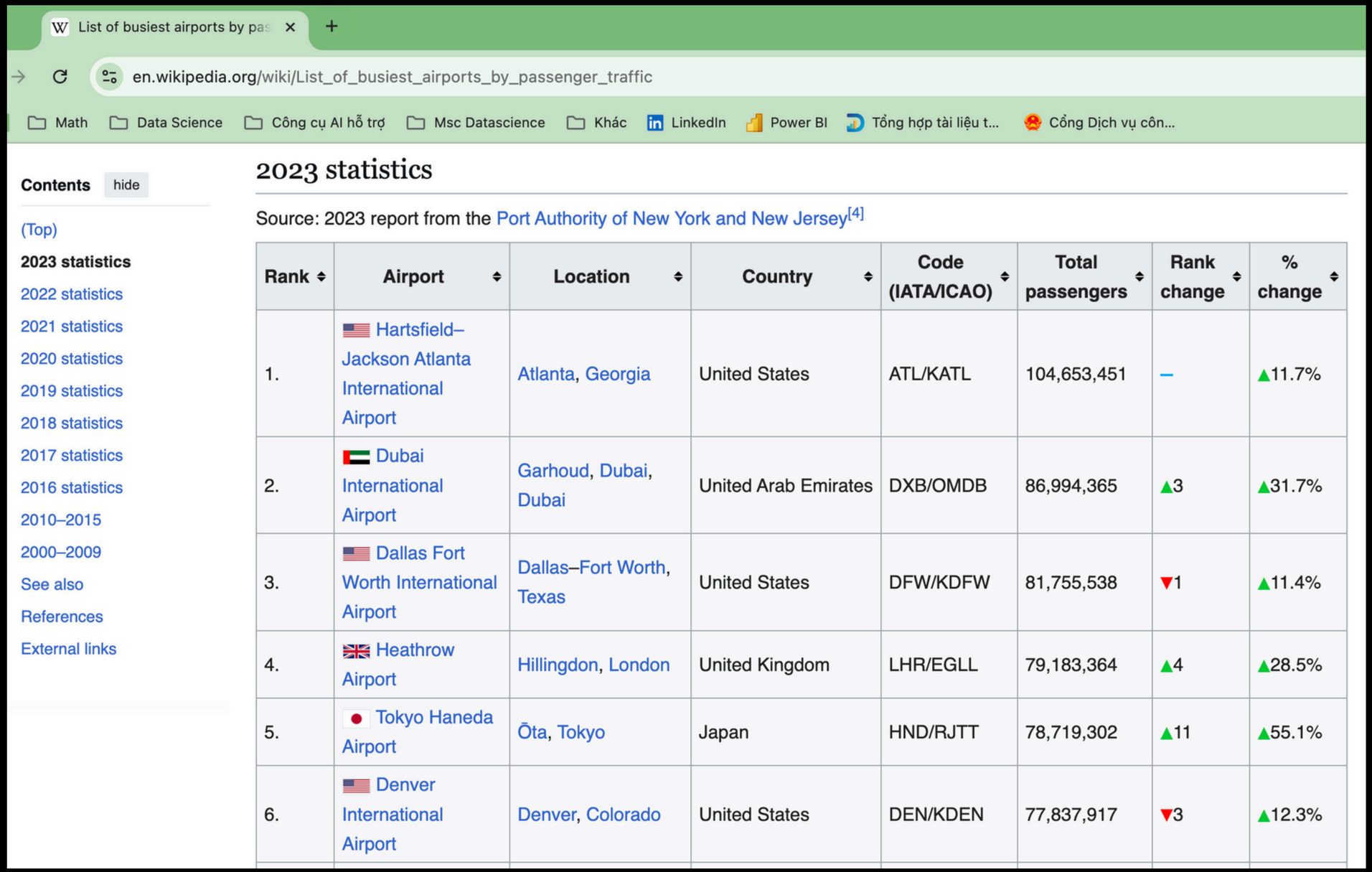
GIẢNG VIÊN HƯỚNG DẪN: PGS. TS. NGUYỄN THANH BÌNH

PHÂN TÍCH DỮ LIỆU SÂN BAY

Thành viên nhóm:
23C01036 - Nguyễn Lê Thành Phước
23C01041 - Lê Thị Mai Thảo
23C01042 - Vũ Thi Thi


Introduction

Nguồn dữ liệu



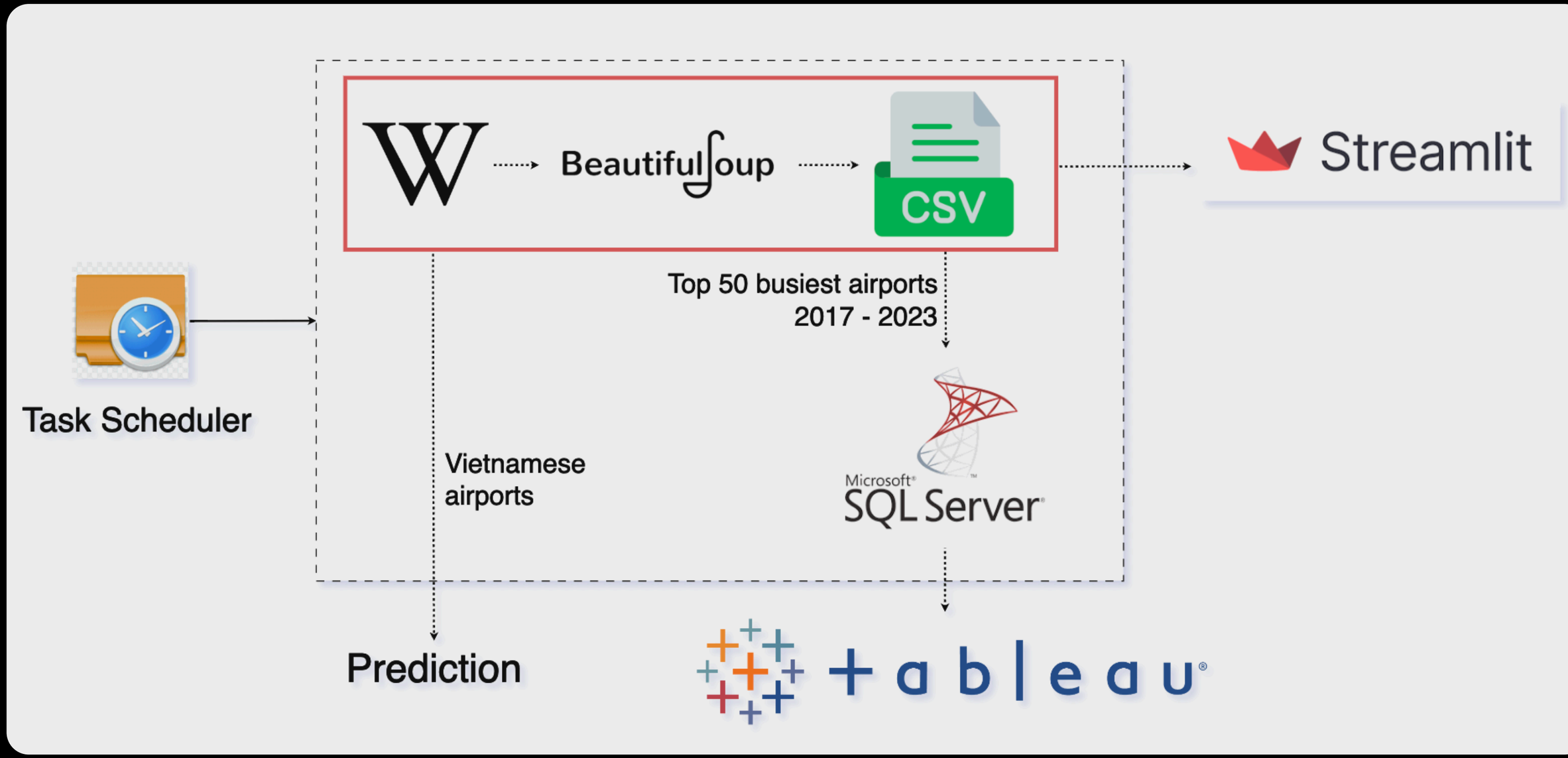
Contents hide

2023 statistics

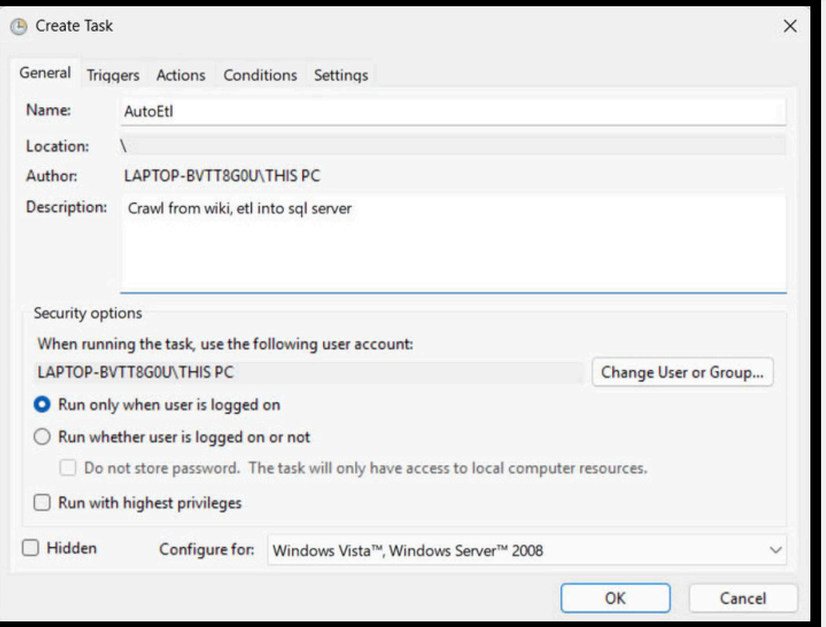
Source: 2023 report from the Port Authority of New York and New Jersey[4]

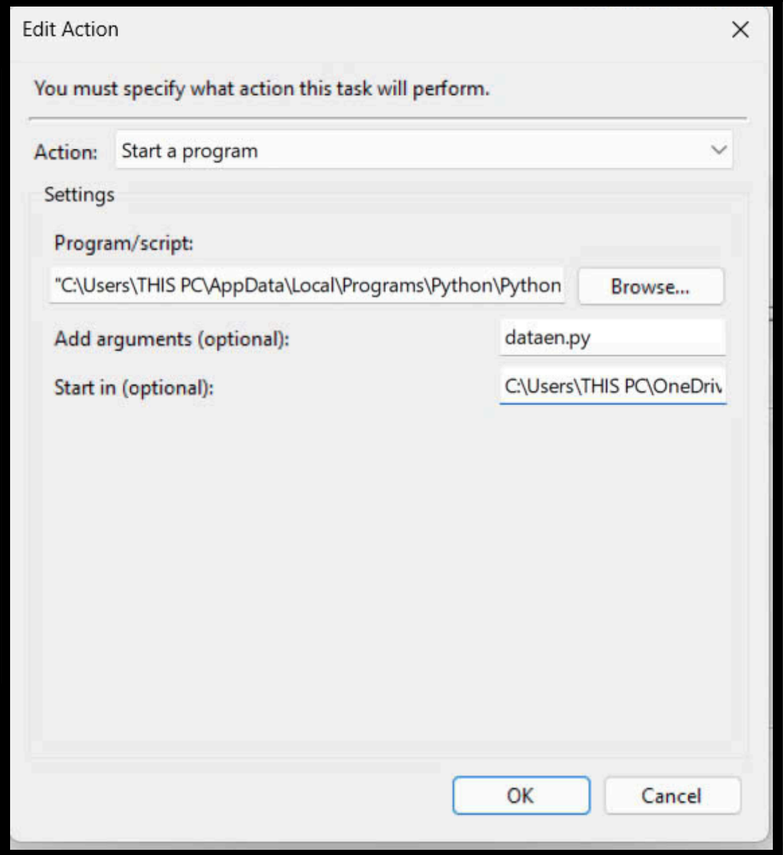
Rank	Airport	Location	Country	Code (IATA/ICAO)	Total passengers	Rank change	% change
1.	Hartsfield–Jackson Atlanta International Airport	Atlanta, Georgia	United States	ATL/KATL	104,653,451	–	▲11.7%
2.	Dubai International Airport	Garhoud, Dubai, Dubai	United Arab Emirates	DXB/OMDB	86,994,365	▲3	▲31.7%
3.	Dallas Fort Worth International Airport	Dallas–Fort Worth, Texas	United States	DFW/KDFW	81,755,538	▼1	▲11.4%
4.	Heathrow Airport	Hillingdon, London	United Kingdom	LHR/EGLL	79,183,364	▲4	▲28.5%
5.	Tokyo Haneda Airport	Ōta, Tokyo	Japan	HND/RJTT	78,719,302	▲11	▲55.1%
6.	Denver International Airport	Denver, Colorado	United States	DEN/KDEN	77,837,917	▼3	▲12.3%

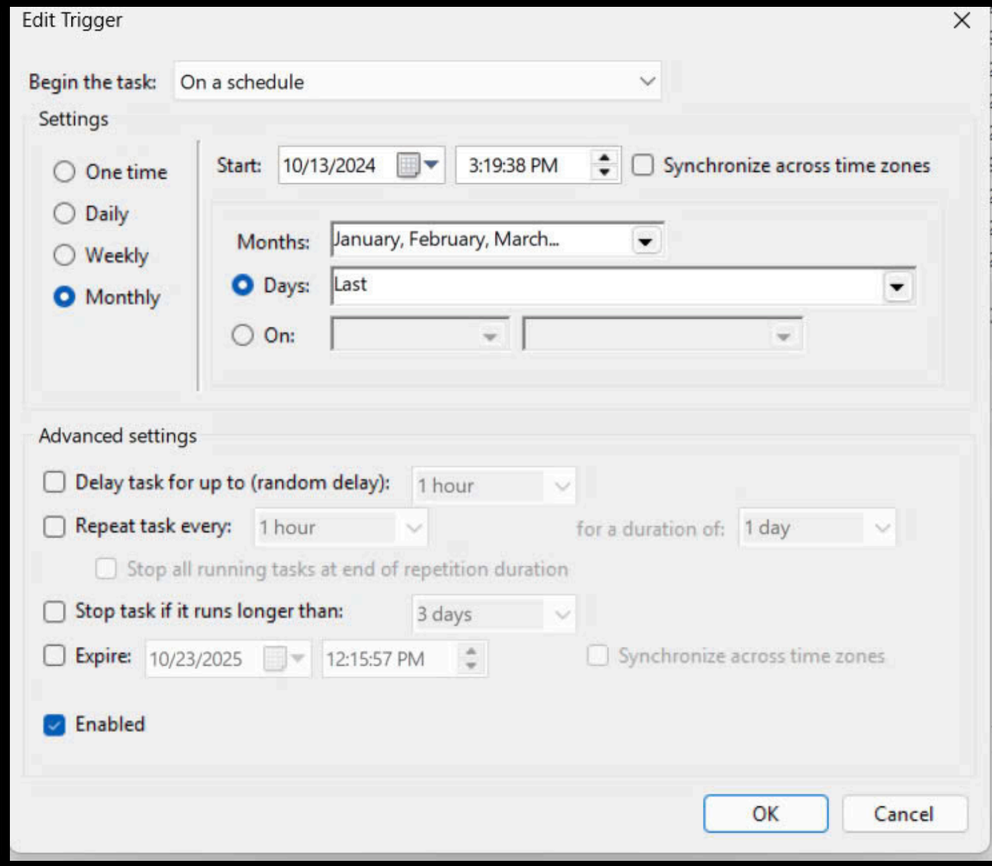
Qui trình


Beautiful Soup
CSV
Streamlit
Task Scheduler
Top 50 busiest airports
2017 - 2023
Vietnamese
airports
Microsoft
SQL Server
Prediction
tableau

Cài đặt Task Scheduler







API Tableau

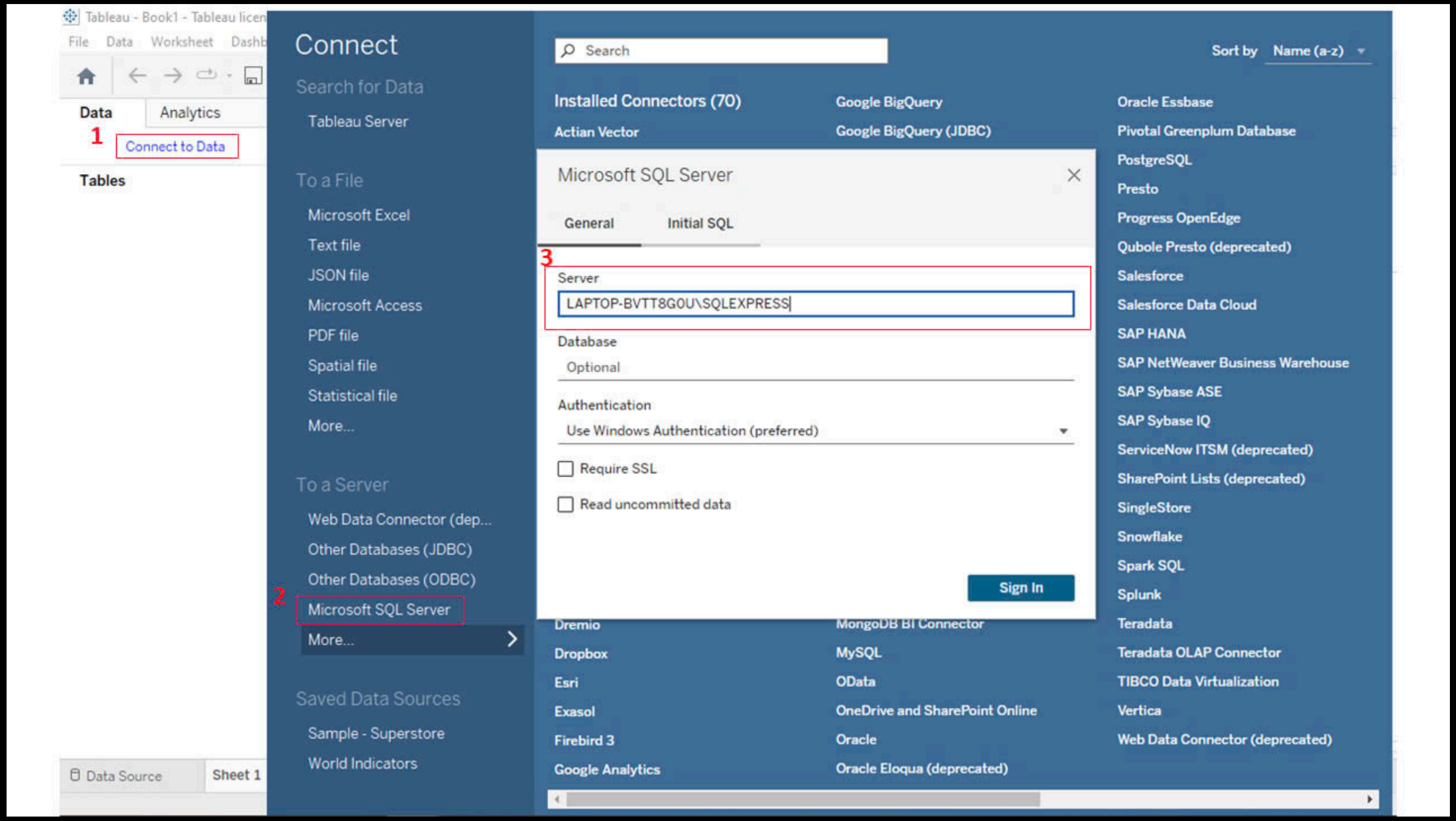
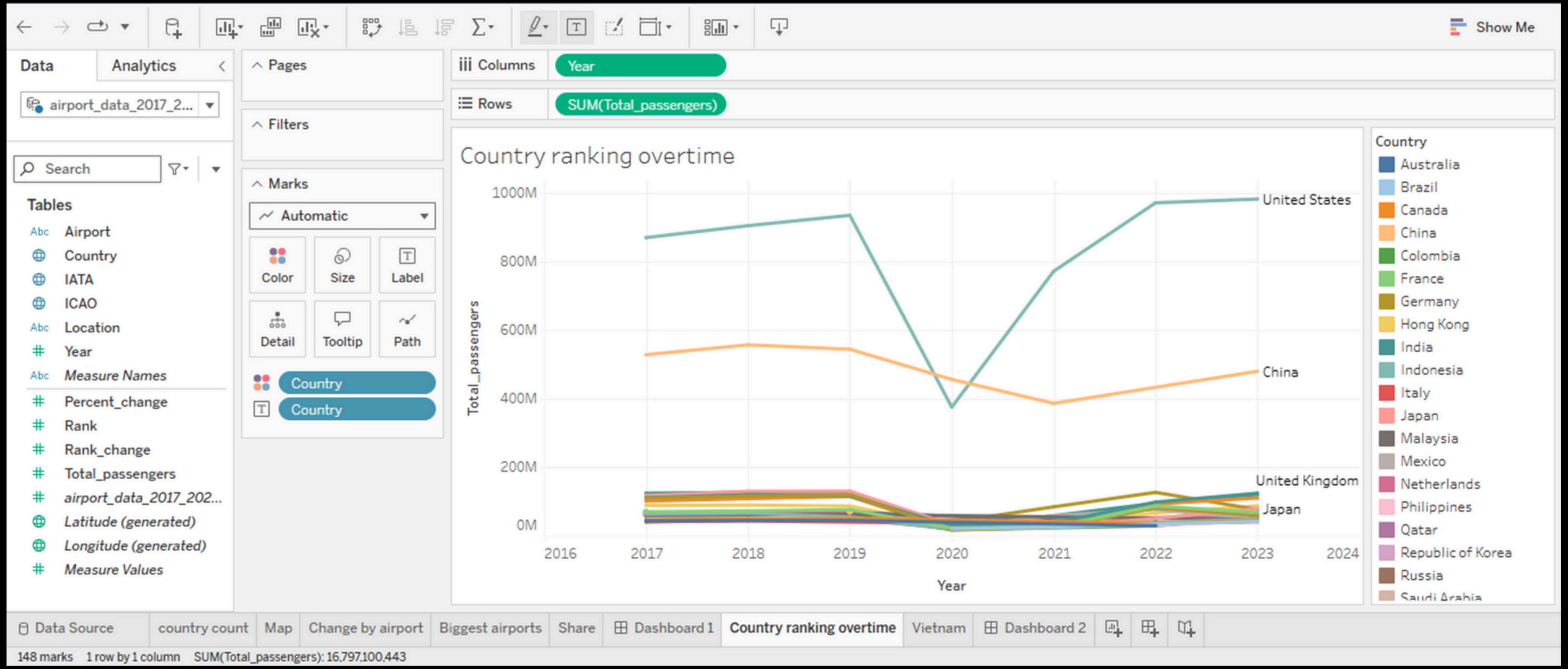
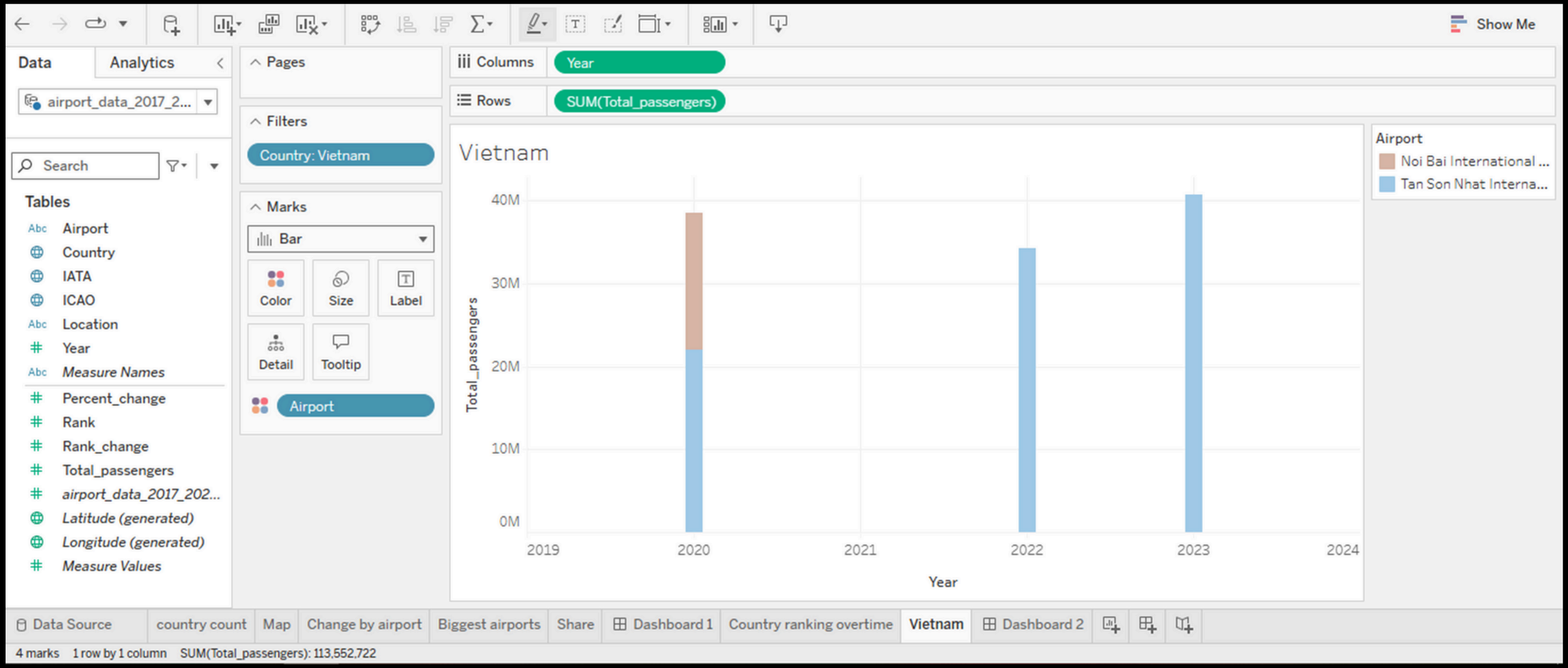


Tableau - Book1 - Tableau licen
File Data Worksheet Dashb
Data
Analytics
1
Connect to Data
Tables
Data Source
Sheet 1
Connect
Search for Data
Tableau Server
To a File
Microsoft Excel
Text file
JSON file
Microsoft Access
PDF file
Spatial file
Statistical file
More...
To a Server
Web Data Connector (dep...
Other Databases (JDBC)
Other Databases (ODBC)
2
Microsoft SQL Server
More...
Saved Data Sources
Sample - Superstore
World Indicators
Search
Sort by Name (a-z)
Installed Connectors (70)
Actian Vector
Google BigQuery
Google BigQuery (JDBC)
Microsoft SQL Server
General
Initial SQL
3
Server
LAPTOP-BVTT8G0U\SQLEXPRESS
Database
Optional
Authentication
Use Windows Authentication (preferred)
Require SSL
Read uncommitted data
Sign In
Dremio
Dropbox
Esri
Exasol
Firebird 3
Google Analytics
MongoDB BI Connector
MySQL
OData
OneDrive and SharePoint Online
Oracle
Oracle Eloqua (deprecated)
Oracle Essbase
Pivotal Greenplum Database
PostgreSQL
Presto
Progress OpenEdge
Qubole Presto (deprecated)
Salesforce
Salesforce Data Cloud
SAP HANA
SAP NetWeaver Business Warehouse
SAP Sybase ASE
SAP Sybase IQ
ServiceNow ITSM (deprecated)
SharePoint Lists (deprecated)
SingleStore
Snowflake
Spark SQL
Splunk
Teradata
Teradata OLAP Connector
TIBCO Data Virtualization
Vertica
Web Data Connector (deprecated)


Show Me
Dashboard
Layout
Default
Phone
Device Preview
Size
Desktop Browser (100...
Sheets
Share
Country ranking ove...
Vietnam
Objects
Extension
Image
Blank
Workflow
Web Page
Navigation
Download
Tiled
Floating
Show dashboard title
Country ranking overtime
1000M
500M
0M
Total_passengers
United States
China
United Kingdom
Japan
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year
Vietnam
40M
20M
0M
Total_pas...
2019
2020
2021
2022
2023
2024
Year
Airport
Noi Bai International ...
Tan Son Nhat Interna...
Data Source
country count
Map
Change by airport
Biggest airports
Share
Dashboard 1
Country ranking overtime
Vietnam
Dashboard 2


Show Me
Data
Analytics
airport_data_2017_2...
Search
Tables
Airport
Country
IATA
ICAO
Location
Year
Measure Names
Percent_change
Rank
Rank_change
Total_passengers
airport_data_2017_202...
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Path
Country
Country
Columns
Year
Rows
SUM(Total_passengers)
Country ranking overtime
1000M
800M
600M
400M
200M
0M
Total_passengers
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year
United States
China
United Kingdom
Japan
Country
Australia
Brazil
Canada
China
Colombia
France
Germany
Hong Kong
India
Indonesia
Italy
Japan
Malaysia
Mexico
Netherlands
Philippines
Qatar
Republic of Korea
Russia
Saudi Arabia
Data Source
country count
Map
Change by airport
Biggest airports
Share
Dashboard 1
Country ranking overtime
Vietnam
Dashboard 2
148 marks
1 row by 1 column
SUM(Total_passengers): 16,797,100,443


Show Me
Data
Analytics
airport_data_2017_2...
Search
Tables
Abc Airport
Country
IATA
ICAO
Abc Location
Year
Abc Measure Names
Percent_change
Rank
Rank_change
Total_passengers
airport_data_2017_202...
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Country: Vietnam
Marks
Bar
Color
Size
Label
Detail
Tooltip
Airport
Columns
Year
Rows
SUM(Total_passengers)
Vietnam
40M
30M
20M
10M
0M
Total_passengers
2019
2020
2021
2022
2023
2024
Year
Airport
Noi Bai International ...
Tan Son Nhat Interna...
Data Source
country count
Map
Change by airport
Biggest airports
Share
Dashboard 1
Country ranking overtime
Vietnam
Dashboard 2
4 marks
1 row by 1 column
SUM(Total_passengers): 113,552,722

Show Me

Dashboard | Layout

Default
Phone

Device Preview

Size

Custom size (1100 x 5...

Sheets

- country count
- Map
- Change by airport

Objects

- Horizontal Container
- Vertical Container
- Text
- Extension
- Image
- Blank

Tiled | Floating

Show dashboard title

Count of data_engineerairports1.csv

The World's 50 Busiest Airports

Most Airtports in top 50

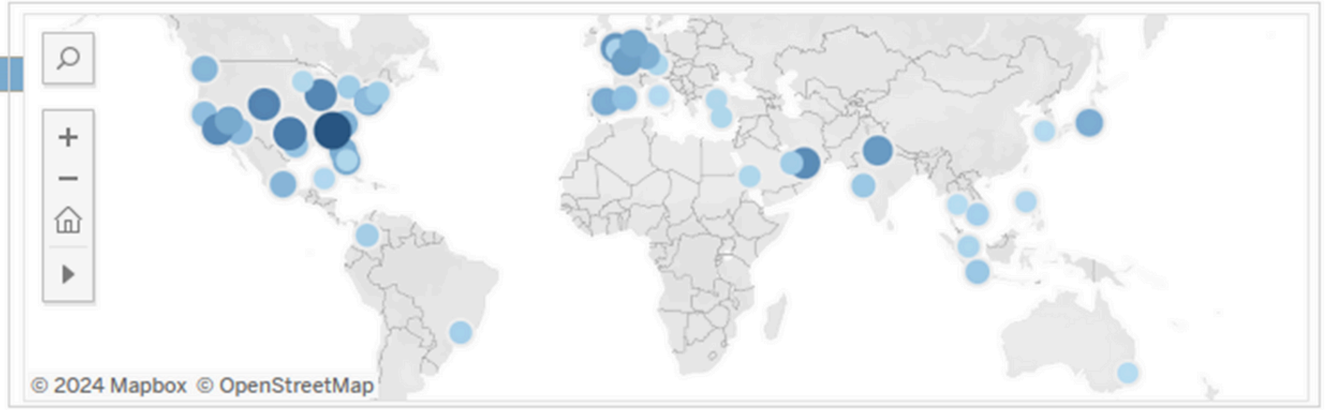
USA

Busiest Airtport

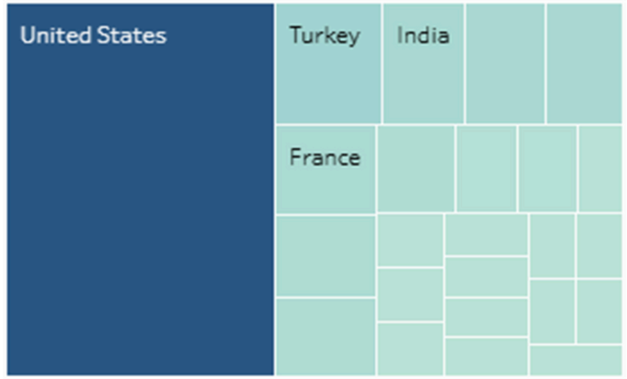
ATL- Hartsfield- Jackson Atlanta

Fast Growing Passenger Numbers

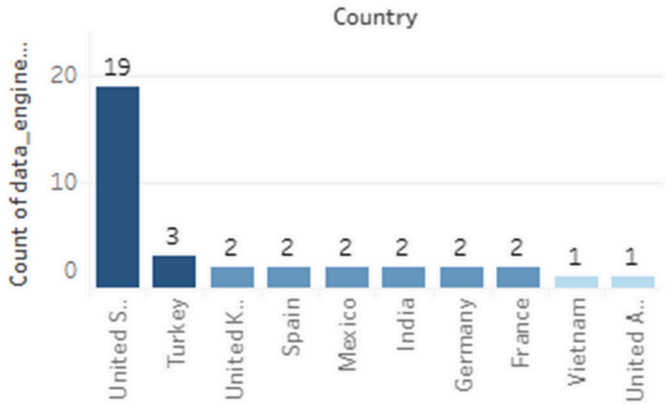
SIN- Singapore Changi



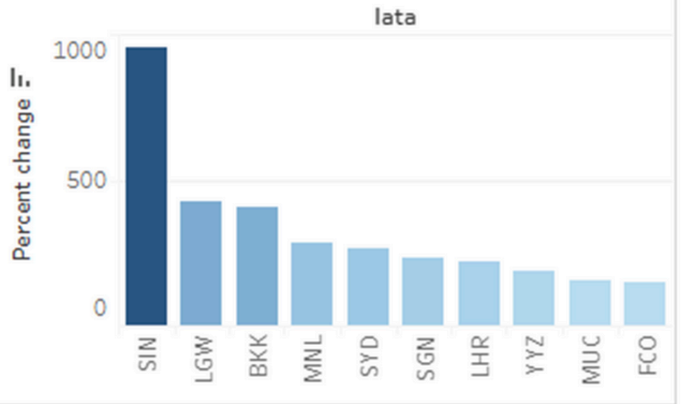
Country's share of passengers in top 50



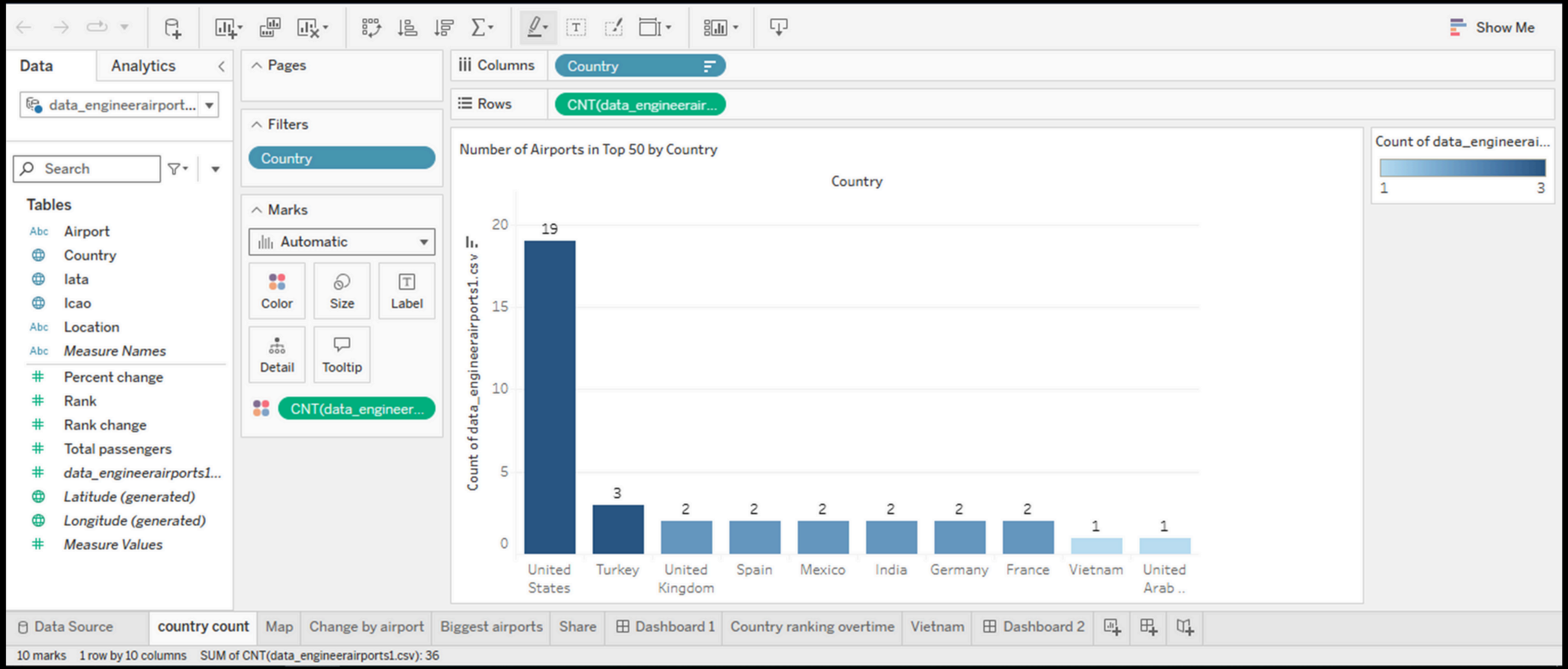
Number of Airports in Top 50 by Country

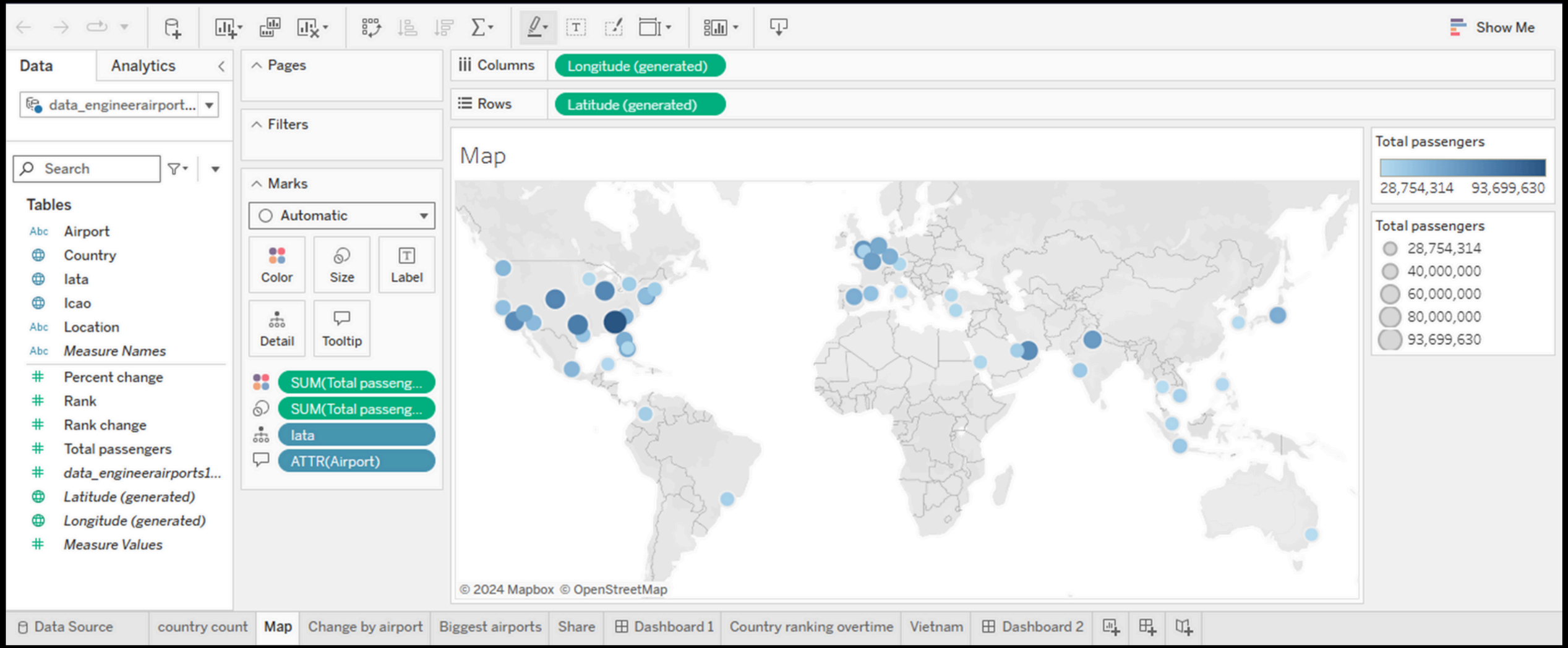


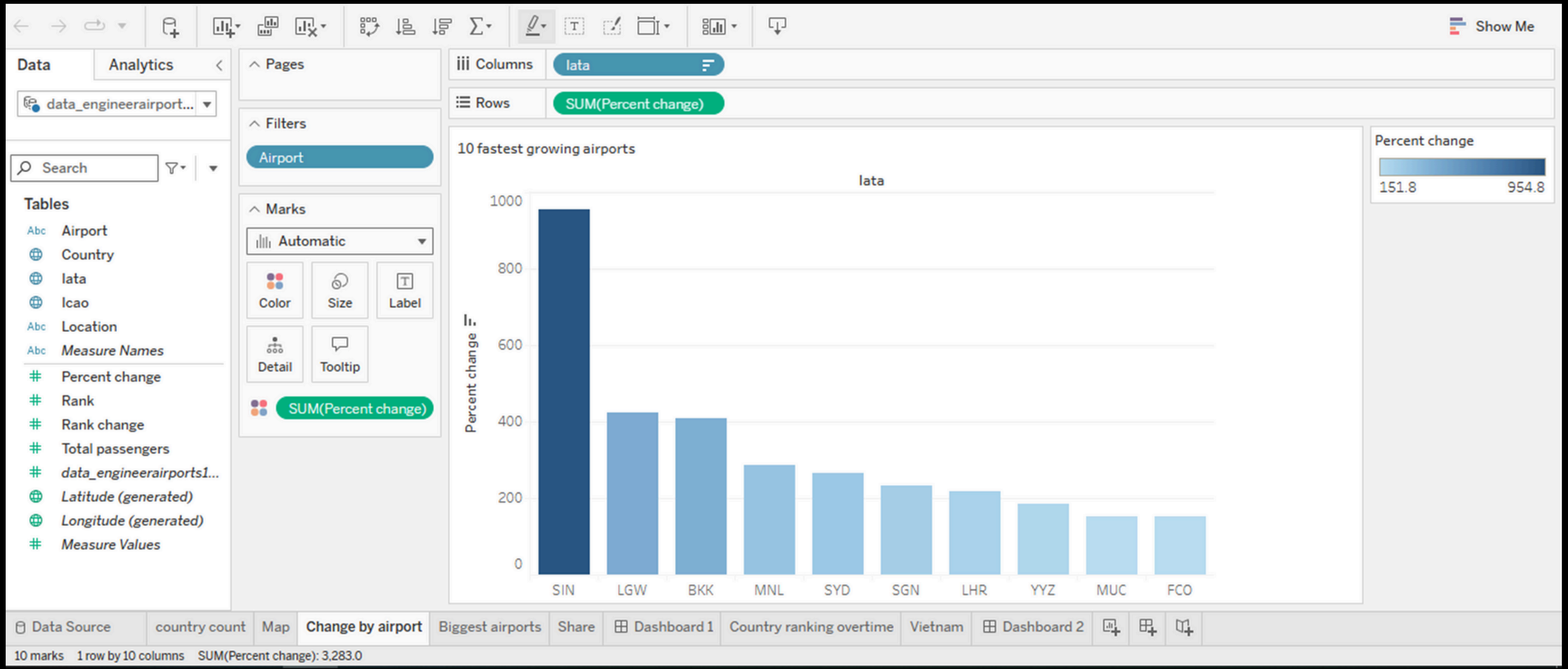
10 fastest growing airports

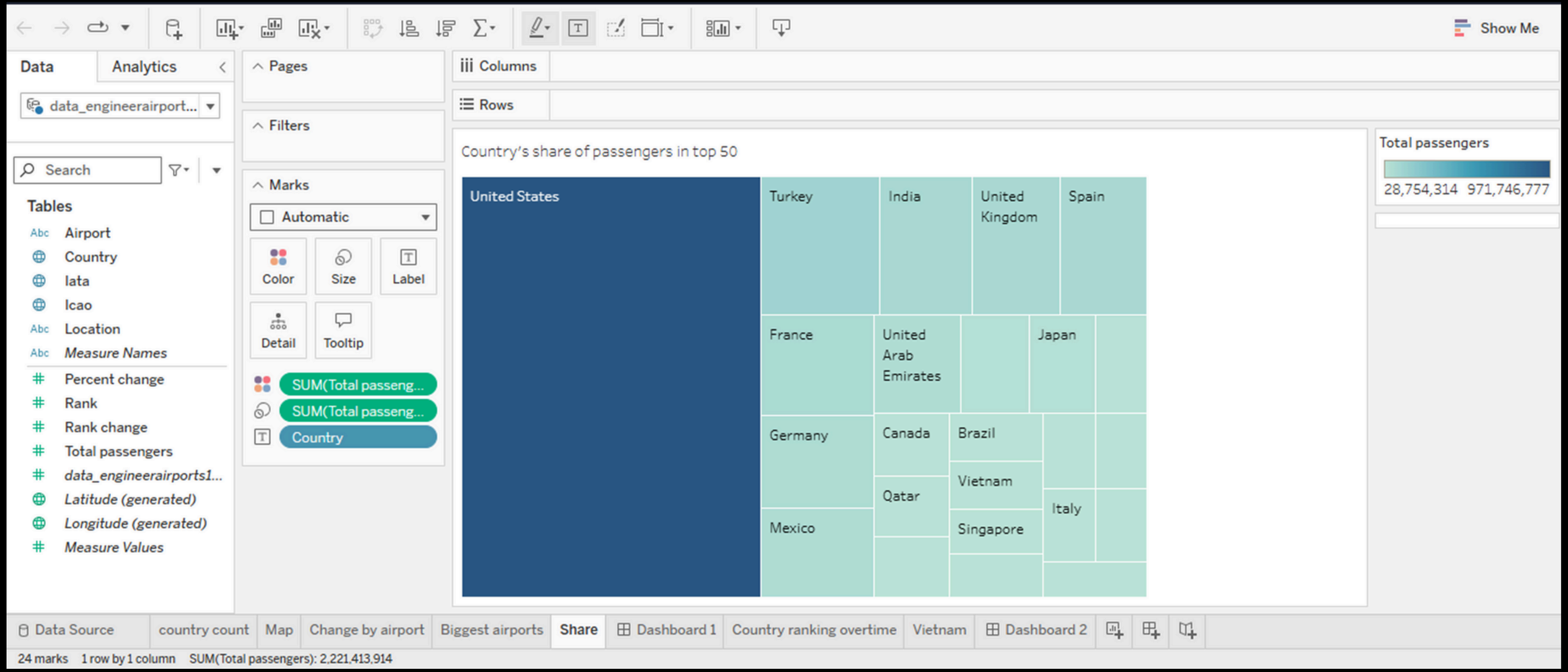


Data Source | country count | Map | Change by airport | Biggest airports | Share | Dashboard 1 | Country ranking overtime | Vietnam | Dashboard 2


Show Me
Data
Analytics
data_engineerairport...
Search
Tables
Abc Airport
Country
Iata
Icao
Abc Location
Abc Measure Names
Percent change
Rank
Rank change
Total passengers
data_engineerairports1...
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Country
Marks
Automatic
Color
Size
Label
Detail
Tooltip
CNT(data_engineer...
Columns
Country
Rows
CNT(data_engineerair...
Number of Airports in Top 50 by Country
Country
Count of data_engineerairports1.csv
20
15
10
5
0
19
3
2
2
2
2
2
2
1
1
United States
Turkey
United Kingdom
Spain
Mexico
India
Germany
France
Vietnam
United Arab ..
Count of data_engineerai...
1
3
Data Source
country count
Map
Change by airport
Biggest airports
Share
Dashboard 1
Country ranking overtime
Vietnam
Dashboard 2
10 marks
1 row by 10 columns
SUM of CNT(data_engineerairports1.csv): 36


Show Me
Data
Analytics
data_engineerairport...
Search
Tables
Airport
Country
Iata
Icao
Location
Measure Names
Percent change
Rank
Rank change
Total passengers
data_engineerairports1...
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Total passeng...
SUM(Total passeng...
Iata
ATTR(Airport)
Columns
Longitude (generated)
Rows
Latitude (generated)
Map
© 2024 Mapbox © OpenStreetMap
Total passengers
28,754,314
93,699,630
Total passengers
28,754,314
40,000,000
60,000,000
80,000,000
93,699,630
Data Source
country count
Map
Change by airport
Biggest airports
Share
Dashboard 1
Country ranking overtime
Vietnam
Dashboard 2


Show Me
Data
Analytics
data_engineerairport...
Search
Tables
Airport
Country
Iata
Icao
Location
Measure Names
Percent change
Rank
Rank change
Total passengers
data_engineerairports1...
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Airport
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Percent change)
Columns
Iata
Rows
SUM(Percent change)
10 fastest growing airports
Iata
Percent change
1000
800
600
400
200
0
SIN
LGW
BKK
MNL
SYD
SGN
LHR
YYZ
MUC
FCO
Percent change
151.8
954.8
Data Source
country count
Map
Change by airport
Biggest airports
Share
Dashboard 1
Country ranking overtime
Vietnam
Dashboard 2
10 marks
1 row by 10 columns
SUM(Percent change): 3,283.0


Show Me
Data
Analytics
data_engineerairport...
Search
Tables
Abc Airport
Country
Iata
Icao
Abc Location
Abc Measure Names
Percent change
Rank
Rank change
Total passengers
data_engineerairports1...
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Total passeng...
SUM(Total passeng...
Country
Columns
Rows
Country's share of passengers in top 50
United States
Turkey
India
United Kingdom
Spain
France
United Arab Emirates
Japan
Germany
Canada
Brazil
Vietnam
Qatar
Italy
Mexico
Singapore
Total passengers
28,754,314
971,746,777
Data Source
country count
Map
Change by airport
Biggest airports
Share
Dashboard 1
Country ranking overtime
Vietnam
Dashboard 2
24 marks
1 row by 1 column
SUM(Total passengers): 2,221,413,914

Analysis of domestic airlines

Passenger traffic [edit]

2020s [edit]

No.	Airport name
1	Tan Son Nhat International Airport
2	Noi Bai International Airport
3	Da Nang International Airport
4	Cam Ranh International Airport
5	Phu Quoc International Airport
6	Cat Bi International Airport

2020	2021	2022	2023
22,062,893[1]	10,287,611[2]	34,278,320[2]	40,738,295[3]
16,473,214[1]	Unknown	24,430,000[4]	29,800,000[4]
Unknown	Unknown	8,900,000[5]	Unknown
3,305,057	972,817[6]	3,860,541[6]	5,700,000[7]
Unknown	Unknown	5,500,000[8]	Unknown
Unknown	Unknown	2,979,000[9]	2,670,000

2010s [edit]

No.	Airport name	Province	City served	IATA	ICAO	2012[17]
1	Tan Son Nhat International Airport	Ho Chi Minh City	Ho Chi Minh City	SGN	VVTS	17,538,353
2	Noi Bai International Airport	Hanoi	Hanoi	HAN	VVNB	11,341,039
3	Da Nang International Airport	Da Nang	Da Nang	DAD	VVDN	3,090,877
4	Cam Ranh International Airport	Khánh Hòa	Nha Trang	CXR	VVCR	1,095,776
5	Phu Quoc International Airport	Kiên Giang	Phú Quốc	PQC	VVPQ	493,434
6	Cat Bi International Airport	Hai Phong	Hai Phong	HPH	VVCI	683,574
7	Vinh Airport	Nghệ An	Vinh	VII	VVVH	635,277
8	Phu Bai International Airport	Thừa Thiên Huế	Hue	HUI	VVPB	673,044

2014[17]	2015[18]	2016[17]	2017	2018[19]	2019[20][21]
22,153,349	26,546,475	32,486,537	35,996,014	38,414,737	41,243,240
14,190,675	17,213,715	20,596,632	23,824,400	25,908,048	29,304,631
4,989,687	6,722,587	8,783,429	10,801,927	13,229,663	15,543,598
2,062,494	2,722,833	4,858,362	6,500,000[22]	8,250,000	9,747,172
1,002,750	1,467,043	2,278,814	3,000,000	3,200,000	3,700,205
927,001	1,256,719	1,800,000	2,089,000[23]	2,373,700[24]	2,639,000[25]
1,222,698	1,300,000	1,563,387	1,800,000[26]	1,880,000	1,950,000
1,159,499	1,300,000	1,550,000	1,750,000[27]	1,831,000[28]	1,931,939
675,607	862,164	1,262,513	1,530,000[29]	1,690,000	2,000,000[30]
695,147	830,000	1,220,000	Unknown	909,907	1,003,419

```python
import requests
from bs4 import BeautifulSoup
import pandas as pd
import re

# URL của trang web để lấy dữ liệu
url = "https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_Vietnam"

# Gửi yêu cầu GET tới trang web
response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

# Tìm tất cả các bảng có class 'wikitable'
tables = soup.find_all('table', class_='wikitable')

# Hàm để lấy dữ liệu từ một bảng với tiêu đề đã chỉ định
def scrape_airport_data(required_headers):
    data = []  # Danh sách để lưu trữ dữ liệu
    for table in tables:
        # Lấy tiêu đề cột
        headers = [th.text.strip() for th in table.find_all('th')]

        # Kiểm tra xem tất cả các tiêu đề yêu cầu có trong tiêu đề cột hay không
        if all(header in headers for header in required_headers):
            # Lấy dữ liệu từ bảng
            rows = table.find_all('tr')[1:]  # Bỏ qua hàng tiêu đề
            for row in rows:
                cols = [td.text.strip() for td in row.find_all('td')]
                # Chỉ thêm dữ liệu nếu số lượng cột khớp với tiêu đề
                if len(cols) == len(required_headers):
                    # Xử lý từng cột: loại bỏ số trong ngoặc và chuyển đổi giá trị null/unknown thành
                    processed_cols = []
                    for col in cols:
                        # Loại bỏ số trong ngoặc
                        col = re.sub(r'\[\d+\]', '', col).strip()
                        # Chuyển đổi 'null' hoặc 'unknown' thành '0'
                        if col.lower() in ['null', 'unknown']:
                            col = '0'
                        processed_cols.append(col)  # Thêm cột đã xử lý vào danh sách
                    data.append(processed_cols)
    return data

# Định nghĩa tiêu đề yêu cầu cho cả hai tập dữ liệu
required_headers_2020s = ['No.', 'Airport name', 'Province', 'City served', 'IATA', 'ICAO', '2020', '2021', '2022', '2023']
required_headers_2012s = ['No.', 'Airport name', 'Province', 'City served', 'IATA', 'ICAO', '2012[17]', '2013[17]', '2014[17]',
'2015[18]', '2016[17]', '2017', '2018[19]', '2019[20][21]']
```

```
Statistic for busiest airports in <a href="/wiki/Vietnam" title="Vietnam">Vietn
</p>
<meta property="mw:PageProp/toc" />
<div class="mw-heading mw-heading2"><h2 id="Passenger_traffic">Passenger traffi
<div class="mw-heading mw-heading3"><h3 id="2020s">2020s</h3><span class="mw-ed
<table class="wikitable sortable">

</table>
<div style="overflow-x:auto; white-space:nowrap;">
<table class="wikitable" style="margin:0;">
<tbody><tr>
<th>No.
</th>
<th>Airport name
</th>
<th>Province
</th>
<th>City served
</th>
<th>IATA
</th>
<th>ICAO
</th>
<th data-sort-type="number">2020
</th>
<th data-sort-type="number">2021
</th>
<th data-sort-type="number">2022
</th>
<th data-sort-type="number">2023
</th></tr>
<tr>
<td>1
</td>
<td><a href="/wiki/Tan Son Nhat International Airport" title="Tan Son Nhat Inte
</td>
<td><a href="/wiki/Ho Chi Minh City" title="Ho Chi Minh City">Ho Chi Minh City<
</td>
<td><a href="/wiki/Ho Chi Minh City" title="Ho Chi Minh City">Ho Chi Minh City<
</td>
<td>SGN
</td>
<td>VVTS
</td>
<td>22,062,893<sup id="cite_ref-world2020_1-0" class="reference"><a href="#cite
</td>
```

```
1 merged_df.head()
```

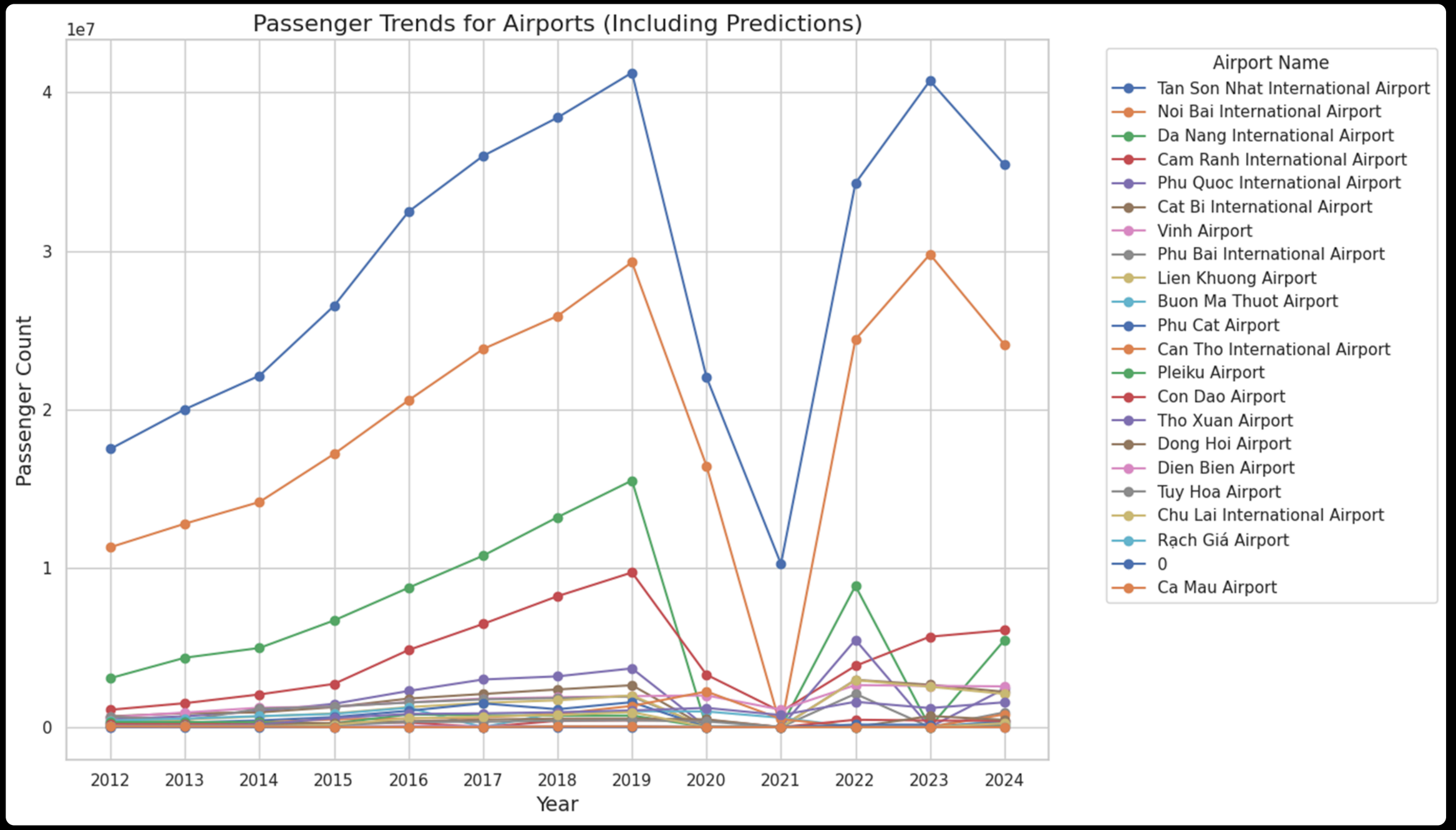
	Airport name	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
0	Tan Son Nhat International Airport	17538353	20035152	22153349	26546475	32486537	35996014	38414737	41243240	22062893	10287611	34278320	40738295
1	Noi Bai International Airport	11341039	12825784	14190675	17213715	20596632	23824400	25908048	29304631	16473214	0	24430000	29800000
2	Da Nang International Airport	3090877	4376775	4989687	6722587	8783429	10801927	13229663	15543598	0	0	8900000	0
3	Cam Ranh International Airport	1095776	1509212	2062494	2722833	4858362	6500000	8250000	9747172	3305057	972817	3860541	5700000
4	Phu Quoc International Airport	493434	685036	1002750	1467043	2278814	3000000	3200000	3700205	0	0	5500000	0

```
# DataFrame đã được định hình lại
pivot_df.head()
```

Airport name	Year	NaN	Buon Ma Thuot Airport	Ca Mau Airport	Cam Ranh International Airport	Can Tho International Airport	Cat Bi International Airport	Chu Lai International Airport	Con Dao Airport	Da Nang International Airport	...	Noi Bai International Airport	Phu Bai International Airport	Phu Cat Airport	Phu Quoc International Airport
0	2012	0	410724	37995	1095776	200751	683574	53753	191039	3090877	...	11341039	673044	236254	493434
1	2013	0	535084	34400	1509212	241307	872762	50974	175574	4376775	...	12825784	427582	290832	685036
2	2014	0	695147	30698	2062494	305015	927001	40198	188549	4989687	...	14190675	1159499	420520	1002750
3	2015	0	830000	0	2722833	481447	1256719	154549	231679	6722587	...	17213715	1300000	630935	1467043
4	2016	0	1220000	0	4858362	550090	1800000	553285	293932	8783429	...	20596632	1550000	1030000	2278814

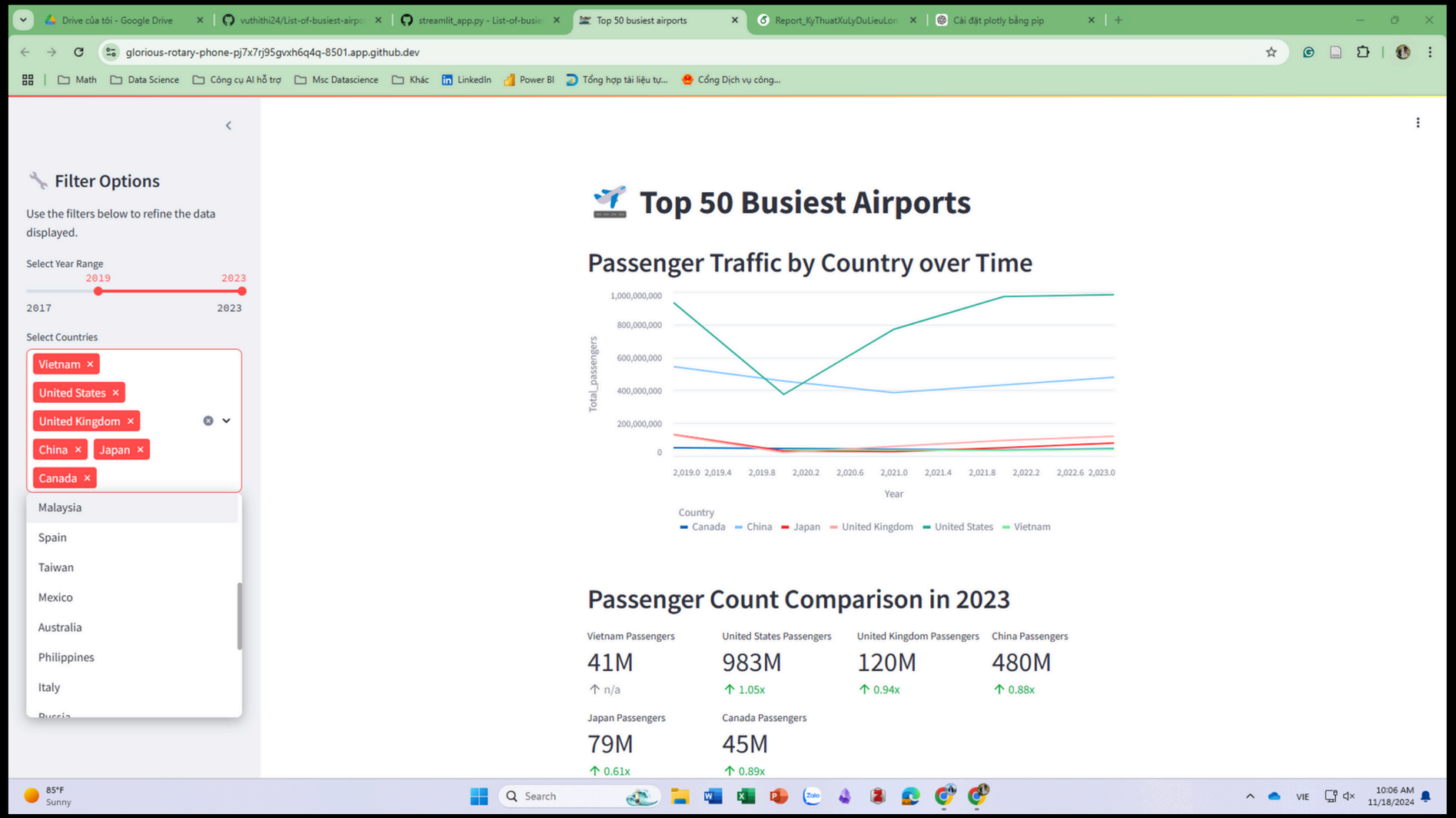
```python
import pandas as pd
from sklearn.linear_model import LinearRegression
import numpy as np


# Chuẩn bị một DataFrame để lưu trữ các dự đoán
predictions = []

# Vòng lặp qua từng sân bay để xây dựng mô hình và dự đoán
for airport in pivot_df.columns[1:]:
    # Chuẩn bị dữ liệu cho sân bay này
    X = pivot_df['Year'].values.reshape(-1, 1)  # Năm là đặc trưng
    y = pivot_df[airport].values  # Số hành khách là mục tiêu

    # Xóa các giá trị NaN nếu có
    if np.any(np.isnan(y)):
        valid_indices = ~np.isnan(y)
        X = X[valid_indices]
        y = y[valid_indices]

    # Huấn luyện mô hình nếu có đủ dữ liệu
    if len(X) >= 2:  # Đảm bảo có đủ dữ liệu để xây dựng mô hình
        model = LinearRegression()
        model.fit(X, y)

        # Dự đoán cho năm tiếp theo
        next_year = np.array([[pivot_df['Year'].max() + 1]])
        predicted_passengers = model.predict(next_year)[0]

        # Lưu kết quả dưới dạng số nguyên
        predictions.append({'Airport name': airport,
                            'Predicted Passengers':
int(round(predicted_passengers))})
    else:
        predictions.append({'Airport name': airport, 'Predicted Passengers':
None})  # Không đủ dữ liệu

# Chuyển đổi các dự đoán thành DataFrame
predictions_df = pd.DataFrame(predictions)

# Hiển thị các dự đoán
print(predictions_df)
```

	Airport name	Predicted Passengers
0	Tan Son Nhat International Airport	35449899
1	Noi Bai International Airport	24113336
2	Cam Ranh International Airport	6113270
3	Da Nang International Airport	5480717
4	Vinh Airport	2568648
5	Phu Quoc International Airport	2414015
6	Cat Bi International Airport	2225046
7	Lien Khuong Airport	2095871
8	Tho Xuan Airport	1565522
9	Phu Bai International Airport	929769
10	Can Tho International Airport	802589
11	Dong Hoi Airport	458692
12	Con Dao Airport	399062
13	Buon Ma Thuot Airport	393873
14	Phu Cat Airport	338533
15	Chu Lai International Airport	249989
16	Tuy Hoa Airport	178673
17	Pleiku Airport	122683


Passenger Trends for Airports (Including Predictions)
1e7
Passenger Count
Year
0
1
2
3
4
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Airport Name
Tan Son Nhat International Airport
Noi Bai International Airport
Da Nang International Airport
Cam Ranh International Airport
Phu Quoc International Airport
Cat Bi International Airport
Vinh Airport
Phu Bai International Airport
Lien Khuong Airport
Buon Ma Thuot Airport
Phu Cat Airport
Can Tho International Airport
Pleiku Airport
Con Dao Airport
Tho Xuan Airport
Dong Hoi Airport
Dien Bien Airport
Tuy Hoa Airport
Chu Lai International Airport
Rạch Giá Airport
0
Ca Mau Airport

Web - app

Link Github: https://github.com/vuthithi24/List-of-busiest-airports-by-passenger-traffic.git



Web – app

Filter Options

Use the filters below to refine the data displayed.

Select Year Range

2019 2023

2017 2023

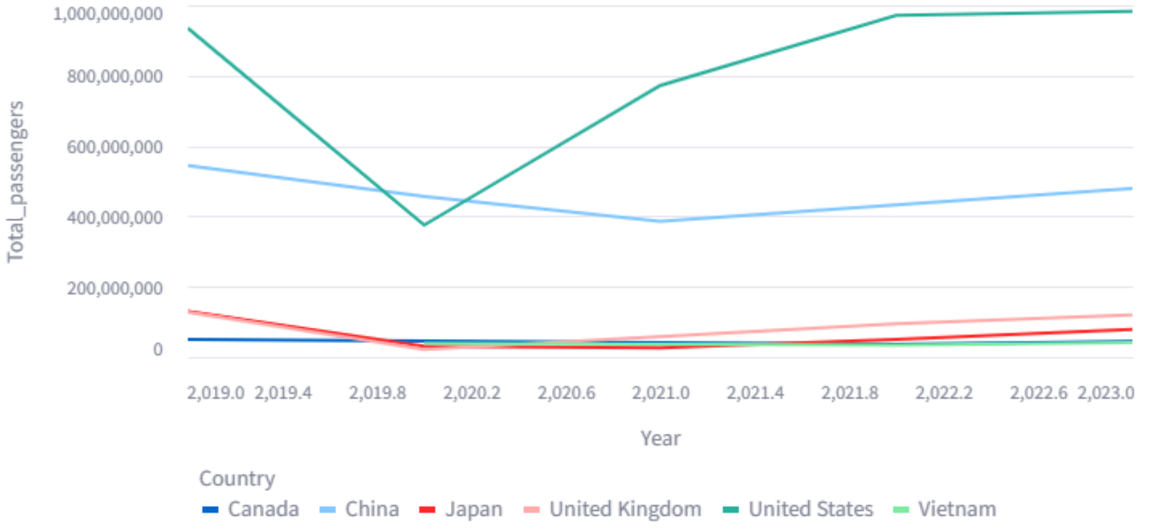
Select Countries

Vietnam × United States × United Kingdom × China × Japan × Canada ×

Malaysia
Spain
Taiwan
Mexico
Australia
Philippines
Italy
Russia

Top 50 Busiest Airports

Passenger Traffic by Country over Time



Passenger Count Comparison in 2023

Vietnam Passengers
41M
↑ n/a

United States Passengers
983M
↑ 1.05x

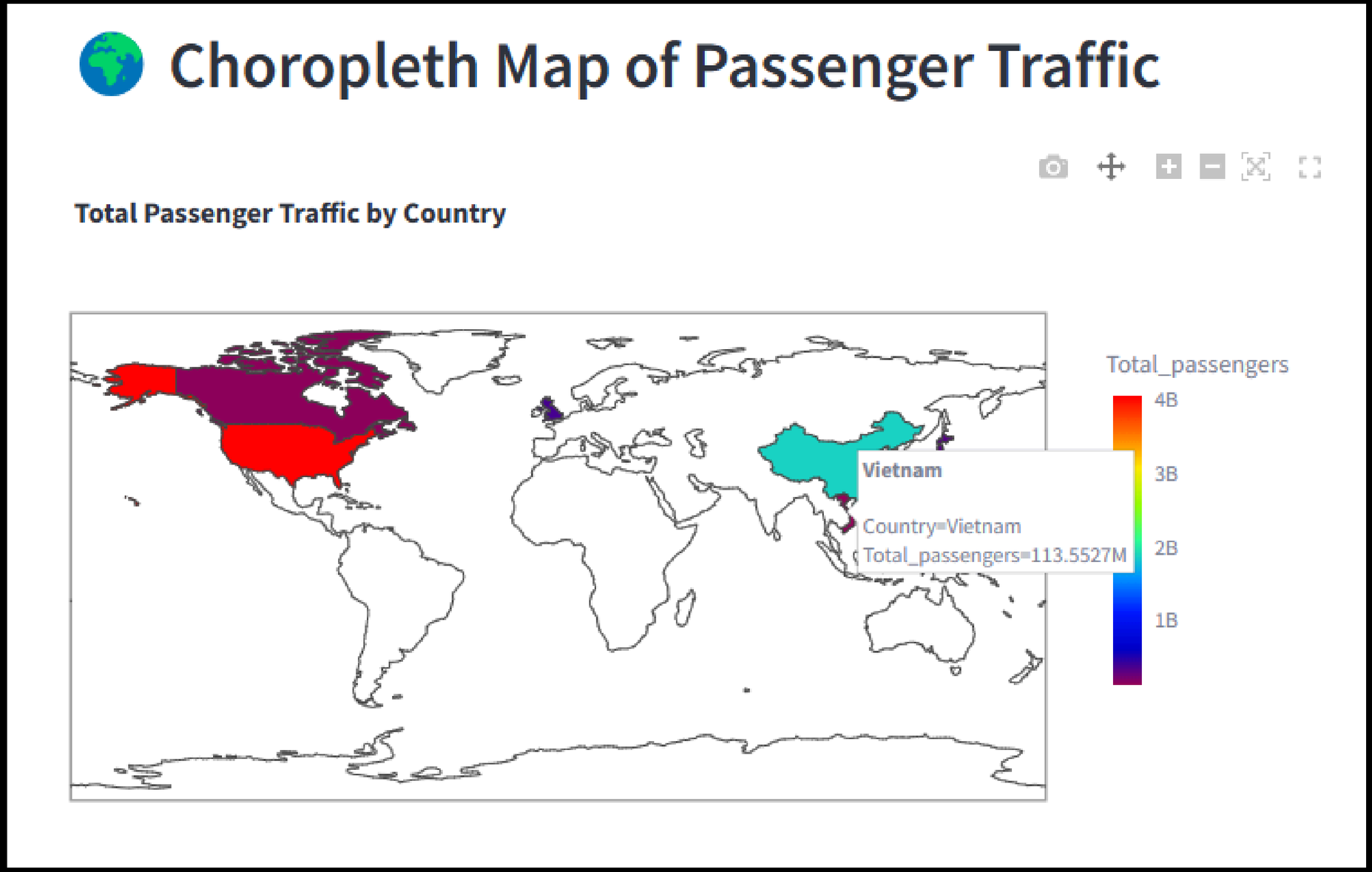
United Kingdom Passengers
120M
↑ 0.94x

China Passengers
480M
↑ 0.88x

Japan Passengers
79M
↑ 0.61x

Canada Passengers
45M
↑ 0.89x

Web - app


Choropleth Map of Passenger Traffic
Total Passenger Traffic by Country
Vietnam
Country=Vietnam
Total_passengers=113.5527M
Total_passengers
4B
3B
2B
1B

THANK YOU!