Import library

In [23]:
```
# import library
import opendatasets as od
import os
import pandas as pd
```

Import data from Kaggle

In [10]:
```
# identify the dataset's URL on Kaggle
url_data = 'https://www.kaggle.com/datasets/iamsouravbanerjee/airline-dataset'
```

In [11]:
```
# download dataset form kaggle by identify URL of dataset
# used Kaggle AIP username and key
od.download(url_data)
```

```
Please provide your Kaggle credentials to download this dataset. Learn more: http://bit.ly/kaggle-creds
Your Kaggle username: sasitpiasai
Your Kaggle Key: ········
Downloading airline-dataset.zip to .\airline-dataset
100%|██████████████████████████████████████████████████████████████████████████████| 12.5M/12.5M [00:01<00:00, 6.66MB/s]
```

In [12]:
```
# identify directory for save data
data_dir = './airline-dataset'
data_dir
```

Out[12]:
```
'./airline-dataset'
```

In [13]:
```
# download file in directory
# output is file name
os.listdir(data_dir)
```

Out[13]:
```
['Airline Dataset Updated - v2.csv',
 'Airline Dataset Updated.csv',
 'Airline Dataset.csv']
```

Create data frame

In [14]:
```
# create data frame
file_path = data_dir + '\Airline Dataset Updated.csv' # file_path = data_dir + 'file name'
df = pd.read_csv(file_path)
df
```

Out[14]:

	Passenger ID	First Name	Last Name	Gender	Age	Nationality	Airport Name	Airport Country Code	Country Name	Airport Continent	Continents	Departure Date	Arrival Airport	Pilot Name	Flight Status
0	ABVWIg	Edithe	Leggis	Female	62	Japan	Coldfoot Airport	US	United States	NAM	North America	6/28/2022	CXF	Edithe Leggis	On Time
1	jkXXAX	Elwood	Catt	Male	62	Nicaragua	Kugluktuk Airport	CA	Canada	NAM	North America	12/26/2022	YCO	Elwood Catt	On Time
2	CdUz2g	Darby	Felgate	Male	67	Russia	Grenoble-Isère Airport	FR	France	EU	Europe	1/18/2022	GNB	Darby Felgate	On Time
3	BRS38V	Dominica	Pyle	Female	71	China	Ottawa / Gatineau Airport	CA	Canada	NAM	North America	9/16/2022	YND	Dominica Pyle	Delayed
4	9kvTLo	Bay	Pencost	Male	21	China	Gillespie Field	US	United States	NAM	North America	2/25/2022	SEE	Bay Pencost	On Time
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98614	hnGQ62	Gareth	Mugford	Male	85	China	Hasvik Airport	NO	Norway	EU	Europe	12/11/2022	HAA	Gareth Mugford	Cancelled
98615	2omEzh	Kasey	Benedict	Female	19	Russia	Ampampamena Airport	MG	Madagascar	AF	Africa	10/30/2022	IVA	Kasey Benedict	Cancelled
98616	VUPiVG	Darrin	Lucken	Male	65	Indonesia	Albacete-Los Llanos Airport	ES	Spain	EU	Europe	9/10/2022	ABC	Darrin Lucken	On Time
98617	E47NtS	Gayle	Lievesley	Female	34	China	Gagnoa Airport	CI	Côte d'Ivoire	AF	Africa	10/26/2022	GGN	Gayle Lievesley	Cancelled
98618	8JYEcz	Wilhelmine	Touret	Female	10	Poland	Yoshkar-Ola Airport	RU	Russian Federation	EU	Europe	4/16/2022	JOK	Wilhelmine Touret	Delayed

98619 rows × 15 columns

Overview of Data

In [32]:
```
# show first 5 rows of data
df.head()
```

Out[32]:

	Passenger ID	First Name	Last Name	Gender	Age	Nationality	Airport Name	Airport Country Code	Country Name	Airport Continent	Continents	Departure Date	Arrival Airport	Pilot Name	Flight Status	Day_of_week	Month
0	ABVWIg	Edithe	Leggis	Female	62	Japan	Coldfoot Airport	US	United States	NAM	North America	2022-06-28	CXF	Edithe Leggis	On Time	1	6
1	jkXXAX	Elwood	Catt	Male	62	Nicaragua	Kugluktuk Airport	CA	Canada	NAM	North America	2022-12-26	YCO	Elwood Catt	On Time	0	12
2	CdUz2g	Darby	Felgate	Male	67	Russia	Grenoble-Isère Airport	FR	France	EU	Europe	2022-01-18	GNB	Darby Felgate	On Time	1	1
3	BRS38V	Dominica	Pyle	Female	71	China	Ottawa / Gatineau Airport	CA	Canada	NAM	North America	2022-09-16	YND	Dominica Pyle	Delayed	4	9
4	9kvTLo	Bay	Pencost	Male	21	China	Gillespie Field	US	United States	NAM	North America	2022-02-25	SEE	Bay Pencost	On Time	4	2

In [33]:
```
# Check information of data as number of rows and columns, all column name, data type and Non-Null Count
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 98619 entries, 0 to 98618
Data columns (total 17 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Passenger ID          98619 non-null  object
 1   First Name            98619 non-null  object
 2   Last Name             98619 non-null  object
 3   Gender                98619 non-null  object
 4   Age                   98619 non-null  int64
 5   Nationality           98619 non-null  object
 6   Airport Name          98619 non-null  object
 7   Airport Country Code  98619 non-null  object
 8   Country Name          98619 non-null  object
 9   Airport Continent     98619 non-null  object
 10  Continents            98619 non-null  object
 11  Departure Date        98619 non-null  datetime64[ns]
 12  Arrival Airport       98619 non-null  object
 13  Pilot Name            98619 non-null  object
 14  Flight Status         98619 non-null  object
 15  Day_of_week           98619 non-null  int32
 16  Month                 98619 non-null  int32
dtypes: datetime64[ns](1), int32(2), int64(1), object(13)
memory usage: 12.0+ MB
```

In [18]:
```
# Check missing value in dataset
df.isnull().sum()
```

Out[18]:
```
Passenger ID            0
First Name              0
Last Name               0
Gender                  0
Age                     0
Nationality             0
Airport Name            0
Airport Country Code    0
Country Name            0
Airport Continent       0
Continents              0
Departure Date          0
Arrival Airport         0
Pilot Name              0
Flight Status           0
dtype: int64
```

In [19]:
```
# Show statistics about data (show only numeric data)
df.describe()
```

Out[19]:

	Age
count	98619.000000
mean	45.504021
std	25.929849
min	1.000000
25%	23.000000
50%	46.000000
75%	68.000000
max	90.000000

Data Cleaning

In [20]:
```
# Cleaning data
# separate day and month from Departure Date
df['Departure Date'] = pd.to_datetime(df['Departure Date']) # convert column 'Departure Date' to datetime
df['Day_of_week'] = df['Departure Date'].dt.dayofweek # add column 'Day_of_Week' by .dt.dayofweek
df['Month'] = df['Departure Date'].dt.month
```

In [21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 98619 entries, 0 to 98618
Data columns (total 17 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Passenger ID          98619 non-null  object
 1   First Name            98619 non-null  object
 2   Last Name             98619 non-null  object
 3   Gender                98619 non-null  object
 4   Age                   98619 non-null  int64
 5   Nationality           98619 non-null  object
 6   Airport Name          98619 non-null  object
 7   Airport Country Code  98619 non-null  object
 8   Country Name          98619 non-null  object
 9   Airport Continent     98619 non-null  object
 10  Continents            98619 non-null  object
 11  Departure Date        98619 non-null  datetime64[ns]
 12  Arrival Airport       98619 non-null  object
 13  Pilot Name            98619 non-null  object
 14  Flight Status         98619 non-null  object
 15  Day_of_week           98619 non-null  int32
 16  Month                 98619 non-null  int32
dtypes: datetime64[ns](1), int32(2), int64(1), object(13)
memory usage: 12.0+ MB
```

In [22]:
```
# Check data after clean
df.head()
```

Out[22]:

	Passenger ID	First Name	Last Name	Gender	Age	Nationality	Airport Name	Airport Country Code	Country Name	Airport Continent	Continents	Departure Date	Arrival Airport	Pilot Name	Flight Status	Day_of_week	Month
0	ABVWIg	Edithe	Leggis	Female	62	Japan	Coldfoot Airport	US	United States	NAM	North America	2022-06-28	CXF	Edithe Leggis	On Time	1	6
1	jkXXAX	Elwood	Catt	Male	62	Nicaragua	Kugluktuk Airport	CA	Canada	NAM	North America	2022-12-26	YCO	Elwood Catt	On Time	0	12
2	CdUz2g	Darby	Felgate	Male	67	Russia	Grenoble-Isère Airport	FR	France	EU	Europe	2022-01-18	GNB	Darby Felgate	On Time	1	1
3	BRS38V	Dominica	Pyle	Female	71	China	Ottawa / Gatineau Airport	CA	Canada	NAM	North America	2022-09-16	YND	Dominica Pyle	Delayed	4	9
4	9kvTLo	Bay	Pencost	Male	21	China	Gillespie Field	US	United States	NAM	North America	2022-02-25	SEE	Bay Pencost	On Time	4	2

In [28]:
```
# save data frame to csv file
df.to_csv('df_airline.csv', index=False)
```

Connect to MySQL

In [29]:
```
# Import the neccessary modules
from sqlalchemy import create_engine as ce
```

In [26]:
```
# Detail for connect MySQL
host = 'localhost' #localhost
user = 'root' # user name
password = 'rootuser' # Your password
database = 'datasciencedb' # database name
```

In [27]:
```
# Create connection
engine = ce(f'mysql+pymysql://{user}:{password}@{host}/{database}')
```

In [30]:
```
# Import DataFrame DB
df.to_sql(name='airline', con=engine, if_exists='replace', index=False)host = 'localhost'
user = 'root'
password = 'rootuser'
database = 'datasciencedb'
```

Out[30]: 98619

In [31]:
```
# close connections
engine.dispose()
```