Tugas Mandiri

Tugas mandiri ini digunakan pada kegiatan Kursus Data Science yang merupakan pembekalan bagi mahasiswa Universitas Gunadarma untuk Skema Associate Data Scientist

Pertemuan 2 - Semester 8

Tugas Mandiri Data Preparation

In [1]:
```
# 1. Import Library Pandas
import pandas as pd
```

In [2]:
```
# 2. Memuat data berformat CSV
df = pd.read_csv('delay_maskapai.csv')
```

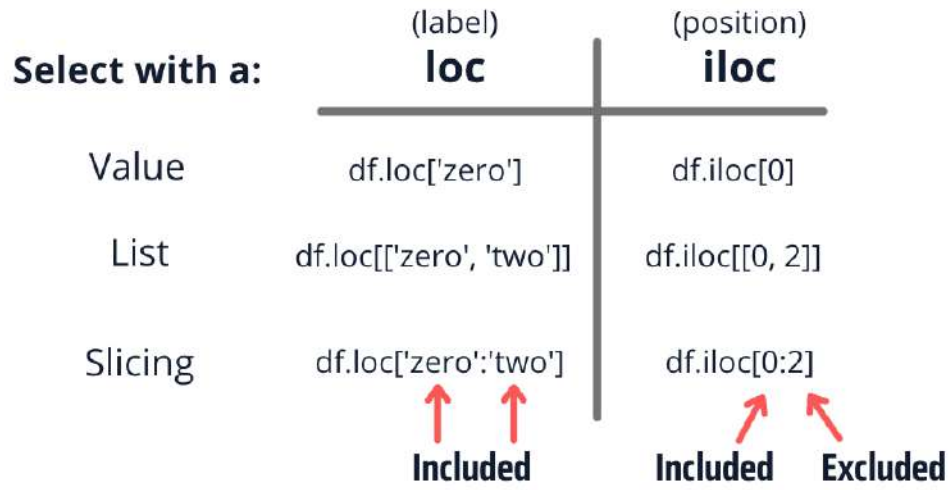
In [3]:
```
# 3. Menampilkan 5 baris teratas dari dataset
df.head(5)
```

Out[3]:

	Flight	Time	Length	Airline	AirportFrom	AirportTo	DayOfWeek	Class
0	320.0	870.0	180.0	CO	IAH	PHX	5	0
1	3045.0	365.0	66.0	OO	DSM	MKE	2	0
2	1560.0	740.0	257.0	DL	SEA	CVG	7	0
3	1156.0	595.0	235.0	WN	SNA	MDW	7	1
4	1873.0	530.0	133.0	CO	IAH	MCO	6	1

Data cleansing - membuang kolom

Dari dataset di atas dapat diketahui bahwa kolom Flight merupakan kolom untuk ID masing-masing penerbangan, kolom ini dapat dibuang karena tidak memiliki makna rill. Untuk membuang kolom dapat menggunakan fungsi iloc maupun drop, berikut adalah perbedaan penggunaan loc dan iloc



In [4]:
```
# 4. Memuat dataset
df_drop1 = pd.read_csv('delay_maskapai.csv')
```

In [5]:
```
# 5. Menampilkan 5 baris pertama dari dataset
df_drop1.head(5)
```

Out[5]:

	Flight	Time	Length	Airline	AirportFrom	AirportTo	DayOfWeek	Class
0	320.0	870.0	180.0	CO	IAH	PHX	5	0
1	3045.0	365.0	66.0	OO	DSM	MKE	2	0
2	1560.0	740.0	257.0	DL	SEA	CVG	7	0
3	1156.0	595.0	235.0	WN	SNA	MDW	7	1
4	1873.0	530.0	133.0	CO	IAH	MCO	6	1

In [6]:
```
# 6. Pada variabel df_drop1 pilihlah hanya kolom Airline, AirportFrom, AirportTo
df_drop1 = df_drop1[['Airline', 'AirportFrom', 'AirportTo']]
```

In [7]:
```
df_drop1.head()
```

Out[7]:

	Airline	AirportFrom	AirportTo
0	CO	IAH	PHX
1	OO	DSM	MKE
2	DL	SEA	CVG
3	WN	SNA	MDW
4	CO	IAH	MCO

In [8]:
```
# 7. Memuat Dataset ke dalam variabel df_drop2
df_drop2 = pd.read_csv('delay_maskapai.csv')
```

In [9]:
```
# 8. Pada variabel df_drop2 buanglah kolom Flight, Airline, AirportFrom, Airport
df_drop2.drop(['Flight', 'Airline', 'AirportFrom', 'AirportTo', 'DayOfWeek', 'Cl
```

In [10]:
```
df_drop2.head()
```

Out[10]:

	Time	Length
0	870.0	180.0
1	365.0	66.0
2	740.0	257.0
3	595.0	235.0
4	530.0	133.0

Data cleansing - Mengubah indeks di dataframe

Karena tidak ada kolom yang bernilai unik pada dataframe, maka kita akan mencoba mengubah kolom Airline menjadi indeks dari dataset

In [11]:
```
# 9. Import Library Numpy
import numpy as np
```

In [12]:
```
# 10. Memuat data
df_index = pd.read_csv('delay_maskapai.csv')
```

In [13]:
```
# 11. Menjadikan kolom Airline sebagai index
df_index = df_index.set_index('Airline')
```

In [14]:
```
# 12. Membuang kolom Flight dengan fungsi drop
df_index.drop(['Flight'], axis=1, inplace=True)
```

In [15]:
```
df_index.head()
```

Out[15]:

Airline	Time	Length	AirportFrom	AirportTo	DayOfWeek	Class
CO	870.0	180.0	IAH	PHX	5	0
OO	365.0	66.0	DSM	MKE	2	0
DL	740.0	257.0	SEA	CVG	7	0
WN	595.0	235.0	SNA	MDW	7	1
CO	530.0	133.0	IAH	MCO	6	1

Data cleansing - Merapikan *Fields* dalam data

Pada dataset ini, kolom "Time" merupakan waktu keberangkatan dari sebuah maskapai. Kondisi nilai pada kolom tersebut adalah waktu yang dikonversi ke menit, untuk melihat jumlah delay terbanyak dari semua waktu keberangkatan, kita perlu mengonversi menit tersebut ke dalam jam dan menit

In [16]:
```
# 13. Mengubah format kolom Time ke dalam Jam dan Menit
Time = pd.to_datetime(df['Time'], unit='m').dt.strftime('%H:%M')
```

In [17]:
```
# 14. Membuat kolom baru dengan nama DepartTime dengan isi datanya berasal dari
df['DepartTime'] = Time
```

In [18]:
```
df.head()
```

Out[18]:

	Flight	Time	Length	Airline	AirportFrom	AirportTo	DayOfWeek	Class	DepartTime
0	320.0	870.0	180.0	CO	IAH	PHX	5	0	14:30
1	3045.0	365.0	66.0	OO	DSM	MKE	2	0	06:05
2	1560.0	740.0	257.0	DL	SEA	CVG	7	0	12:20
3	1156.0	595.0	235.0	WN	SNA	MDW	7	1	09:55
4	1873.0	530.0	133.0	CO	IAH	MCO	6	1	08:50

In [19]:
```
# 15. Menghitung jumlah penerbangan dalam 24 jam
Time_Flight_Count = df.groupby(['DepartTime']).size().reset_index(name='FlightCo
Time_Flight_Count.head()
```

Out[19]:

	DepartTime	FlightCount
0	00:15	2
1	00:30	1
2	00:35	3
3	00:40	2
4	00:49	1

In [20]:
```
# 16. Mengurutkan 20 data dari jumlah penerbangan terbanyak
Time_Flight_Count.sort_values('FlightCount', ascending=False).head(20)
```

Out[20]:

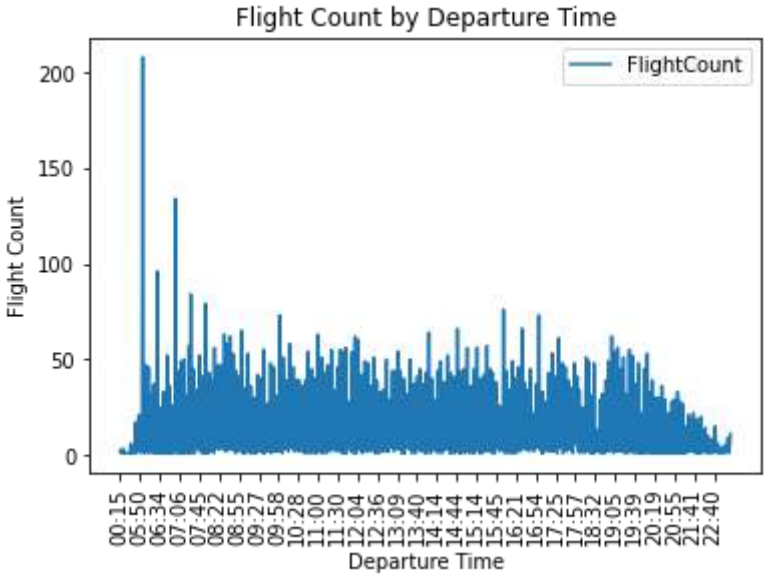
	DepartTime	FlightCount
34	06:00	208
84	07:00	134
56	06:30	96
107	07:30	84
130	08:00	79
581	16:00	76
242	10:00	73
635	17:00	73
609	16:30	66
511	14:45	66
184	09:00	65
468	14:00	64
300	11:00	63
158	08:30	63
167	08:40	62
745	19:00	62
356	12:00	62
665	17:30	61
361	12:05	60
163	08:35	58

Dari kode di atas dapat diketahui bahwa waktu 06.00 merupakan waktu dengan jumlah terbanyak. Berikut kita akan lakukan visualisasi dengan line chart

In [21]:
```
# 17. Mengimport library pyplot
import matplotlib.pyplot as plt
```

In [22]:
```
# 18. Mengatur ukuran gambar
plt.figure(figsize=(10, 6))

# 19. Melakukan plot data
Time_Flight_Count.plot(kind='line', x='DepartTime', y='FlightCount')

# 20. Mengatur label sumbu x
ticks = range(0, len(Time_Flight_Count), 30)  # Menentukan posisi label setiap 3
labels = Time_Flight_Count.iloc[ticks]['DepartTime']  # Mengambil label berdasar
plt.xticks(ticks=ticks, labels=labels, rotation=90) # Mengatur posisi rotasi lab
plt.xticks(ticks=ticks, labels=labels)  # Mengatur posisi dan label pada sumbu x

# 21. Mengatur label sumbu x, y dan judul grafik
plt.xlabel('Departure Time')
plt.ylabel('Flight Count')
plt.title('Flight Count by Departure Time')

# 22. Menampilkan grafik
plt.show()
```

```
<Figure size 720x432 with 0 Axes>
```



Data cleansing - Mengganti Nama Kolom

In [23]:
```
# 23. Memuat Dataset
df_rename_col = pd.read_csv('delay_maskapai.csv')
```

In [24]:
```
# 24. Ubah Nama Kolom Flight -> ID Flight, Airline -> Maskapai, Class -> Kelas
df_rename_col = df_rename_col.rename(columns=
                                {"Flight":"ID Flight", "Airline":"Maskapai", "Class"
```

In [25]:
```
# 25. Menampilkan 10 data teratas
df_rename_col.head(10)
```

Out[25]:

	ID Flight	Time	Length	Maskapai	AirportFrom	AirportTo	DayOfWeek	Kelas
0	320.0	870.0	180.0	CO	IAH	PHX	5	0
1	3045.0	365.0	66.0	OO	DSM	MKE	2	0
2	1560.0	740.0	257.0	DL	SEA	CVG	7	0
3	1156.0	595.0	235.0	WN	SNA	MDW	7	1
4	1873.0	530.0	133.0	CO	IAH	MCO	6	1
5	3641.0	1045.0	160.0	MQ	MIA	CMH	1	0
6	138.0	1160.0	185.0	WN	BDL	MCO	7	0
7	586.0	700.0	142.0	AS	PDX	SNA	2	1
8	5220.0	895.0	102.0	EV	AEX	ATL	2	0
9	5122.0	516.0	137.0	EV	DTW	MCI	7	0