Analytics Data

In [1]:
```
import os
print(os.environ['PATH'])
```

```
C:\ProgramData\Anaconda3;C:\ProgramData\Anaconda3\Library\mingw-w64\bin;C:\Prog
ramData\Anaconda3\Library\usr\bin;C:\ProgramData\Anaconda3\Library\bin;C:\Progr
amData\Anaconda3\Scripts;C:\ProgramData\Anaconda3\bin;C:\ProgramData\Anaconda3
\condabin;C:\ProgramData\Anaconda3;C:\ProgramData\Anaconda3\Library\mingw-w64\b
in;C:\ProgramData\Anaconda3\Library\usr\bin;C:\ProgramData\Anaconda3\Library\bi
n;C:\ProgramData\Anaconda3\Scripts;C:\Program Files\AdoptOpenJDK\jdk-8.0.232.09
-hotspot\bin;C:\Program Files (x86)\Intel\iCLS Client;C:\Program Files\Intel\iC
LS Client;C:\Windows\system32;C:\Windows;C:\Windows\System32\Wbem;C:\Windows\Sy
stem32\WindowsPowerShell\v1.0;C:\Program Files\Intel\Intel(R) Management Engine
Components\DAL;C:\Program Files (x86)\Intel\Intel(R) Management Engine Componen
ts\DAL;C:\Program Files\Intel\Intel(R) Management Engine Components\IPT;C:\Prog
ram Files (x86)\Intel\Intel(R) Management Engine Components\IPT;C:\Program File
s (x86)\Common Files\Lenovo;C:\Program Files\Condusiv Technologies\ExpressCach
e;C:\Program Files (x86)\Common Files\lenovo\easyplussdk\bin;C:\SWTOOLS\ReadyAp
ps;C:\bigdata\hadoop-3.1.2\bin;C:\bigdata\hadoop-3.1.2\sbin;C:\bigdata\hadoop-
3.1.2\bin;C:\alaa\bigdata\spark-2.4.4-bin-hadoop2.7\bin;C:\Program Files\Intel
\WiFi\bin;C:\Program Files\Common Files\Intel\WirelessCommon;C:\Program Files\G
it\cmd;C:\Program Files\Intel\WiFi\bin;C:\Program Files\Common Files\Intel\Wire
lessCommon
```

(1) Read data from CSV File

In [250]:
```
import pandas as pd

flights = pd.read_csv('data/flights.csv',header=0)
flights.head(5)
```

Out[250]:

	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum	TailNum	Elapse
0	1	3	4	1512	1802	WN	706	N491WN	
1	1	3	4	919	1132	WN	643	N756SA	
2	1	3	4	1801	1900	WN	962	N302SW	
3	1	3	4	1631	1749	WN	1006	N628SW	
4	1	3	4	1331	1528	WN	2284	N409WN	

In [251]:
```
planes = pd.read_csv('data/plane.csv',header=0)
planes.head(5)
```

Out[251]:

	tailnum	type	manufacturer	issue_date	model	status	aircraft_type	engine_type	year
0	N10156	Corporation	EMBRAER	02/13/2004	EMB-145XR	Valid	Fixed Wing Multi-Engine	Turbo-Fan	2004
1	N102UW	Corporation	AIRBUS INDUSTRIE	05/26/1999	A320-214	Valid	Fixed Wing Multi-Engine	Turbo-Fan	1998
2	N10323	Corporation	BOEING	07/01/1997	737-3TO	Valid	Fixed Wing Multi-Engine	Turbo-Jet	1986
3	N103US	Corporation	AIRBUS INDUSTRIE	06/18/1999	A320-214	Valid	Fixed Wing Multi-Engine	Turbo-Fan	1999
4	N104UA	Corporation	BOEING	01/26/1998	747-422	Valid	Fixed Wing Multi-Engine	Turbo-Fan	1998

(2) Pick up sepecific columns

In [220]:
```
flight2=flights[['Month','DayofMonth','DayOfWeek','DepTime','ArrTime','UniqueCarr
flight2.head(5)
```

Out[220]:

	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum	TailNum	Elapse
0	1	3	4	1512	1802	WN	706	N491WN	
1	1	3	4	919	1132	WN	643	N756SA	
2	1	3	4	1801	1900	WN	962	N302SW	
3	1	3	4	1631	1749	WN	1006	N628SW	
4	1	3	4	1331	1528	WN	2284	N409WN	

(3) Combine columns in DataFrame

In [235]:
```
flight2['flightDate']='2019-'+flight2['Month'].astype(str)+'-'+flight2['DayofMon
flight2.head(5)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: SettingWith
CopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stab
le/indexing.html#indexing-view-versus-copy (http://pandas.pydata.org/pandas-doc
s/stable/indexing.html#indexing-view-versus-copy)
  """Entry point for launching an IPython kernel.
```

Out[235]:

	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum	TailNum	Elapse
0	1.0	3.0	4.0	1512.0	1802.0	WN	706.0	N491WN	
1	1.0	3.0	4.0	919.0	1132.0	WN	643.0	N756SA	
2	1.0	3.0	4.0	1801.0	1900.0	WN	962.0	N302SW	
3	1.0	3.0	4.0	1631.0	1749.0	WN	1006.0	N628SW	
4	1.0	3.0	4.0	1331.0	1528.0	WN	2284.0	N409WN	

(4) read from sklearn

In [236]:
```
from sklearn.datasets import load_boston
import pandas as pd
import numpy as np

data = load_boston()

df = pd.DataFrame(data.data, columns=data.feature_names)
df['target'] = data['target']
df.head(4)
```

Out[236]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94

(6) filter dataframe

In [237]:
```
flight2[flight2['UniqueCarrier']=='WN'].head(5)
```

Out[237]:

	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum	TailNum	Elapse
0	1.0	3.0	4.0	1512.0	1802.0	WN	706.0	N491WN	
1	1.0	3.0	4.0	919.0	1132.0	WN	643.0	N756SA	
2	1.0	3.0	4.0	1801.0	1900.0	WN	962.0	N302SW	
3	1.0	3.0	4.0	1631.0	1749.0	WN	1006.0	N628SW	
4	1.0	3.0	4.0	1331.0	1528.0	WN	2284.0	N409WN	

(7) Sort dataframe

In [238]:
```
flight2.sort_values(by='flightDate').head(4)
```

Out[238]:

	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum	TailNum	Ela
3778	1.0	1.0	2.0	547.0	829.0	MQ	3778.0	N667GB	
2475	1.0	1.0	2.0	655.0	829.0	US	1933.0	N423US	
1170	1.0	1.0	2.0	1954.0	2135.0	XE	2782.0	N13929	
2476	1.0	1.0	2.0	1444.0	1750.0	US	1509.0	N167US	

(8) Group By

In [239]:
```
planes.groupby(['manufacturer']).size()
```

Out[239]:
```
manufacturer
AIRBUS INDUSTRIE      24
BOEING               145
CANADAIR               1
DOUGLAS                2
EMBRAER              200
MCDONNELL DOUGLAS      1
dtype: int64
```

In [240]:
```
flight2.groupby(['UniqueCarrier', 'Origin','Dest']).agg({'ArrDelay': ['sum'], 'D
```

Out[240]:

			ArrDelay	DepDelay
			sum	sum
UniqueCarrier	Origin	Dest		
9E	ABE	DTW	22.0	40.0
	ALB	DTW	-16.0	28.0
	ALO	MSP	-4.0	-3.0
	ATL	BHM	-3.0	-3.0
		BNA	-26.0	-3.0

In [241]:
```
flight2.groupby(['UniqueCarrier', 'Origin','Dest'])['ArrDelay','DepDelay'].sum()
```

Out[241]:

			ArrDelay	DepDelay
UniqueCarrier	Origin	Dest		
9E	ABE	DTW	22.0	40.0
	ALB	DTW	-16.0	28.0
	ALO	MSP	-4.0	-3.0
	ATL	BHM	-3.0	-3.0
		BNA	-26.0	-3.0

(9) Reset Index

In [242]:
```
flight3=flight2.groupby(['UniqueCarrier', 'Origin','Dest'])['ArrDelay','DepDelay

flight3.head(5)
```

Out[242]:

	UniqueCarrier	Origin	Dest	ArrDelay	DepDelay
0	9E	ABE	DTW	57.0	57.0
1	9E	ALB	DTW	19.0	26.0
2	9E	ALO	MSP	-2.0	0.0
3	9E	ATL	BHM	-3.0	-3.0
4	9E	ATL	BNA	-4.0	4.0

In [243]:
```
flight4=flight2.groupby(['UniqueCarrier', 'Origin','Dest']).agg({'ArrDelay': ['s
for col in flight4.columns:
    print(col)
```

```
('ArrDelay', 'sum')
('DepDelay', 'sum')
```

In [244]:
```
flights2=flights[['FlightNum','TailNum']].drop_duplicates()
flights2['tailNum']=flights2['TailNum']
flights2.head(5)
```

Out[244]:

	FlightNum	TailNum	tailNum
0	706	N491WN	N491WN
1	643	N756SA	N756SA
2	962	N302SW	N302SW
3	1006	N628SW	N628SW
4	2284	N409WN	N409WN

(10) Merge two dataframes (left join, right join, inner join)

In [245]:
```
planes2=planes[['tailnum','type','manufacturer']]
planes3=planes[['tailnum','issue_date','model','status']]
pd.merge(planes2,planes3, left_on='tailnum', right_on='tailnum').head(5)
```

Out[245]:

	tailnum	type	manufacturer	issue_date	model	status
0	N10156	Corporation	EMBRAER	02/13/2004	EMB-145XR	Valid
1	N102UW	Corporation	AIRBUS INDUSTRIE	05/26/1999	A320-214	Valid
2	N10323	Corporation	BOEING	07/01/1997	737-3TO	Valid
3	N103US	Corporation	AIRBUS INDUSTRIE	06/18/1999	A320-214	Valid
4	N104UA	Corporation	BOEING	01/26/1998	747-422	Valid

In [246]:
```
pd.merge(planes2,planes3, left_on='tailnum', right_on='tailnum',how='left').head
```

Out[246]:

	tailnum	type	manufacturer	issue_date	model	status
0	N10156	Corporation	EMBRAER	02/13/2004	EMB-145XR	Valid
1	N102UW	Corporation	AIRBUS INDUSTRIE	05/26/1999	A320-214	Valid
2	N10323	Corporation	BOEING	07/01/1997	737-3TO	Valid
3	N103US	Corporation	AIRBUS INDUSTRIE	06/18/1999	A320-214	Valid
4	N104UA	Corporation	BOEING	01/26/1998	747-422	Valid