Data Science_Foundation Projects_ Analyze NYC-Flight

Lakshya Kumar

Domain: Airlines
Project 01: Analyze NYC-Flight data

Introduction

In this project we are going to analyze the data provided in the NYC_Flight Database, and infer different conclusions on the basis of charts and supportive data. The analysis is done using numpy, pandas and matplotlib in python

Dataset description

Name	Description
year	2013
month	12-Jan
day	Day of the month (1-31)
dep_time	Departuretimes, local timezone
sched_dep_time	Scheduled departure time
dep_delay	Departure delay, in minutes, Negative times represent early departures
arr_time	Arrival times, local timezone
sched_arr-time	Scheduled departure time
arr_delay	Arrivaldelay,inminutes,Negativetimesrepresent early arrivals
carrier	Two letter carrier abbreviation
flight	Flight number
tailnum	Plane tail number
origin, dest	Airport codes for origin and destination
air_time	Amount of time spent in the air, in minutes
distance	Distance flown, in miles
hour, minute	Time of departure broken in to hour and mins.
time_hour	Timestamp

1. Departure delays : finding out the mean and SD of the flights on different airports

Inference : the average delay of the three airports are:

JFK : 15.10

EWR: 12.11

LGA: 10.03

And the standard deviation is shown in the figure

Hence the least delay time is of the LGA airport

In [35]:
```
from statistics import mean
#df.groupby('origin')['dep_delay'].apply(list)

df1=df.groupby(['origin']).mean()
print(df1)
df1.std()
```

```
            year     month        day     dep_time  sched_dep_time  dep_delay  \
origin
EWR     2013.0  6.492564  15.698192  1336.704497     1322.465114  15.107954
JFK     2013.0  6.496931  15.734748  1398.569670     1401.925736  12.112159
LGA     2013.0  6.667941  15.699853  1310.169029     1308.094648  10.346876

            arr_time  sched_arr_time  arr_delay       flight    air_time  \
origin
EWR     1491.875882     1527.981082   9.107055  2373.513833  153.300025
JFK     1520.070385     1564.975997   5.551481  1365.751004  178.349050
LGA     1494.423727     1515.673568   5.783488  2152.773681  117.825806

            distance       hour     minute
origin
EWR     1056.742790  12.952257  27.239393
JFK     1266.249077  13.744237  27.501990
LGA      779.835671  12.843821  23.712541
```

Out[35]:

year	0.000000
month	0.100017
day	0.020643
dep_time	45.361702
sched_dep_time	50.538391
dep_delay	2.406896
arr_time	15.594724
sched_arr_time	25.660697
arr_delay	1.989222
flight	529.735284
air_time	30.410900
distance	243.983756
hour	0.491552
minute	2.116111

dtype: float64

Arrival delays : finding out the mean and SD of the flights on different airports

inference : the average delay of the three airports are:

JFK : 9.10

EWR: 5.55

LGA: 5.78

And the standard deviation is shown in the figure

Hence the least delay time is of the EWR airport

In [35]:
```
from statistics import mean
#df.groupby('origin')['dep_delay'].apply(list)

df1=df.groupby(['origin']).mean()
print(df1)
df1.std()
```

```
          year     month        day     dep_time  sched_dep_time  dep_delay  \
origin
EWR     2013.0  6.492564  15.698192  1336.704497     1322.465114  15.107954
JFK     2013.0  6.496931  15.734748  1398.569670     1401.925736  12.112159
LGA     2013.0  6.667941  15.699853  1310.169029     1308.094648  10.346876

           arr_time  sched_arr_time  arr_delay       flight    air_time  \
origin
EWR     1491.875882     1527.981082   9.107055  2373.513833  153.300025
JFK     1520.070385     1564.975997   5.551481  1365.751004  178.349050
LGA     1494.423727     1515.673568   5.783488  2152.773681  117.825806

           distance       hour     minute
origin
EWR     1056.742790  12.952257  27.239393
JFK     1266.249077  13.744237  27.501990
LGA      779.835671  12.843821  23.712541
```

Out[35]:

year	0.000000
month	0.100017
day	0.020643
dep_time	45.361702
sched_dep_time	50.538391
dep_delay	2.406896
arr_time	15.594724
sched_arr_time	25.660697
arr_delay	1.989222
flight	529.735284
air_time	30.410900
distance	243.983756
hour	0.491552
minute	2.116111

dtype: float64

Best airport in terms of departure

Best Airport in terms of departure according to mean delay time from the analysis is LGA with an average of 10.03

In [35]:
```
from statistics import mean
#df.groupby('origin')['dep_delay'].apply(list)

df1=df.groupby(['origin']).mean()
print(df1)
df1.std()
```

```
            year     month        day     dep_time  sched_dep_time  dep_delay  \
origin
EWR     2013.0  6.492564  15.698192  1336.704497     1322.465114  15.107954
JFK     2013.0  6.496931  15.734748  1398.569670     1401.925736  12.112159
LGA     2013.0  6.667941  15.699853  1310.169029     1308.094648  10.346876

            arr_time  sched_arr_time  arr_delay       flight    air_time  \
origin
EWR     1491.875882     1527.981082   9.107055  2373.513833  153.300025
JFK     1520.070385     1564.975997   5.551481  1365.751004  178.349050
LGA     1494.423727     1515.673568   5.783488  2152.773681  117.825806

            distance       hour     minute
origin
EWR     1056.742790  12.952257  27.239393
JFK     1266.249077  13.744237  27.501990
LGA      779.835671  12.843821  23.712541
```

Aircraft speed analysis

The carrier wise air time and distance travelled are shown in figure 1, and the average speed is shown in figure 2.

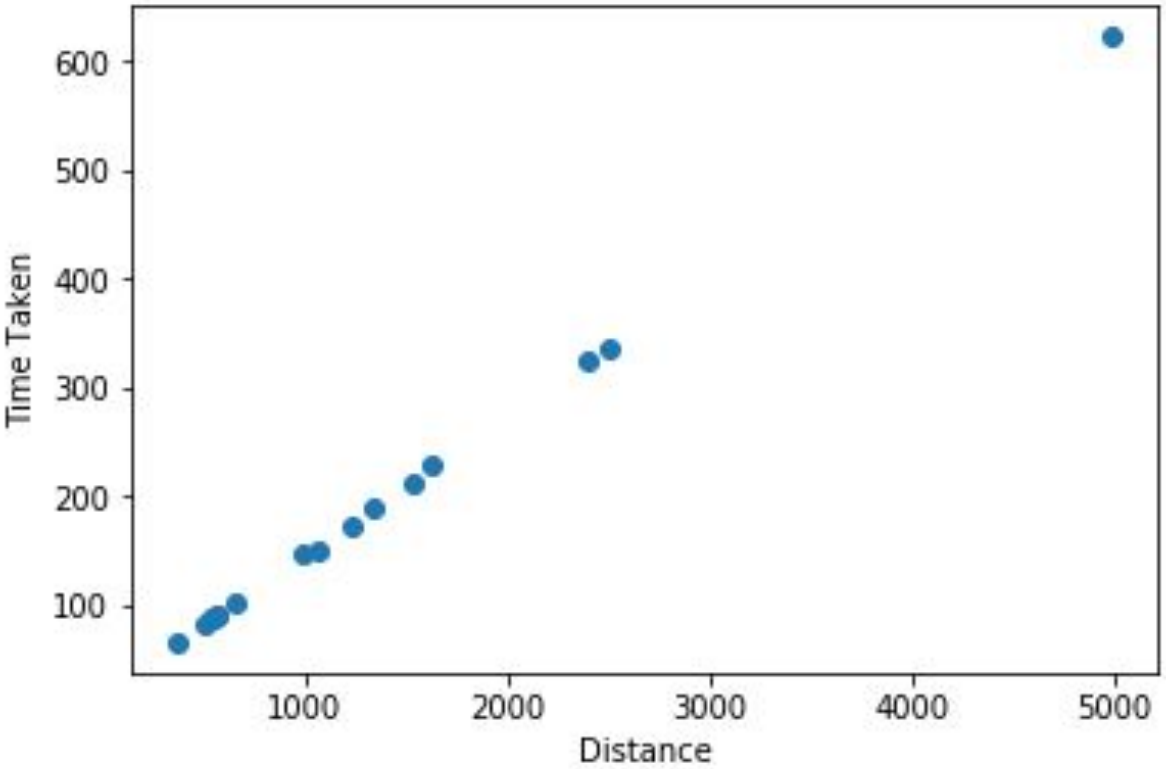
Scatter plot is shown below:

Inference: Fastest flight - HA

In [49]:
```
df2=df[['carrier','air_time','distance']]
df3=df2.groupby(['carrier']).mean()
print(df3)
avg_speed=df3.distance/df3.air_time
print(avg_speed)
```

```
            air_time     distance
carrier
9E         86.781601   530.235753
AA        188.822299  1340.235999
AS        325.617772  2402.000000
B6        151.177173  1068.621525
DL        173.688804  1236.901206
EV         90.076192   562.991730
F9        229.599119  1620.000000
FL        101.143937   664.829448
HA        623.087719  4983.000000
MQ         91.180253   569.532712
OO         83.482759   500.812500
UA        211.791354  1529.114873
US         88.573799   553.456272
VX        337.002346  2499.482177
WN        147.824809   996.269084
YV         65.740809   375.033278
carrier
```

```
carrier
9E    6.110002
AA    7.097869
AS    7.376747
B6    7.068670
DL    7.121364
EV    6.250172
F9    7.055776
FL    6.573102
HA    7.997269
MQ    6.246229
OO    5.998993
UA    7.219912
US    6.248533
VX    7.416809
WN    6.739526
YV    5.704726
dtype: float64
```


600
500
400
300
200
100
Time Taken
1000
2000
3000
4000
5000
Distance

On time arrival analysis

The origin of the flights is sorted and taken the mean of and the airport with min delay(mean) is JFK.

Out[59]:

origin	arr_delay
EWR	9.107055
JFK	5.551481
LGA	5.783488

In [59]:
```
df4= df[['origin','arr_delay']]
print(df4)
df4.groupby(['origin']).mean()
```

```
   origin  arr_delay
0     EWR       11.0
1     LGA       20.0
2     JFK       33.0
3     JFK      -18.0
4     LGA      -25.0
5     EWR       12.0
6     EWR       19.0
7     LGA      -14.0
8     JFK       -8.0
9     LGA        8.0
10    JFK       -2.0
11    JFK       -3.0
12    JFK        7.0
13    EWR      -14.0
14    LGA       31.0
15    JFK       -4.0
16    EWR       -8.0
17    LGA       -7.0
18    LGA       12.0
19    EWR       -6.0
20    LGA       -8.0
21    LGA       16.0
22    EWR      -12.0
23    JFK       -8.0
24    EWR      -17.0
```

Maximum number of flights headed to some particular destination

According to analysis, the destination where the maximum number of flights are headed are ATL

In [88]:
```
tot_flight_dest_count = df.groupby(['dest']).size()
tot_flight_dest_count_1 = tot_flight_dest_count.reset_index(name = 'tot_flight_dest_count')
print(tot_flight_dest_count_1)
```

	dest	tot_flight_dest_count
0	ABQ	254
1	ACK	265
2	ALB	439
3	ANC	8
4	ATL	17215
5	AUS	2439
6	AVL	275
7	BDL	443
8	BGR	375
9	BHM	297
10	BNA	6333
11	BOS	15508
12	BQN	896
13	BTV	2589
14	BUF	4681
15	BUR	371
16	BWI	1781
17	BZN	36
18	CAE	116
19	CAK	864
20	CHO	52
21	CHS	2884
22	CLE	4573
23	CLT	14064

Arrival and departure analysis : Finding out delay departure count, on time departure count, early departure count

```
#on time departure count data
flight_on_time = df[(df['dep_delay'] == 0)]
print ('On time departure count: ',flight_on_time['dep_delay'].count())

#early departure count
flight_early_dep = df[(df['dep_delay'] < 0)]
print ('Early departure count: ',flight_early_dep['dep_delay'].count())
flight_early_dep
```

```
Delay departure count:  128432
On time departure count:  16514
Early departure count:  183575
```

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin	dest	air_time	distance	hour	min
3	2013	1	1	544.0	545	-1.0	1004.0	1022	-18.0	B6	725	N804JB	JFK	BQN	183.0	1576	5	
4	2013	1	1	554.0	600	-6.0	812.0	837	-25.0	DL	461	N668DN	LGA	ATL	116.0	762	6	
5	2013	1	1	554.0	558	-4.0	740.0	728	12.0	UA	1696	N39463	EWR	ORD	150.0	719	5	
6	2013	1	1	555.0	600	-5.0	913.0	854	19.0	B6	507	N516JB	EWR	FLL	158.0	1065	6	

Plot of Flight vs Departure delay count