NYC Flight Data – Data Analytics with Python Project Report

By Shikhar Raj

Agenda

- About the Project
- Dataset description
- Business questions identified
- Hidden insights of the dataset

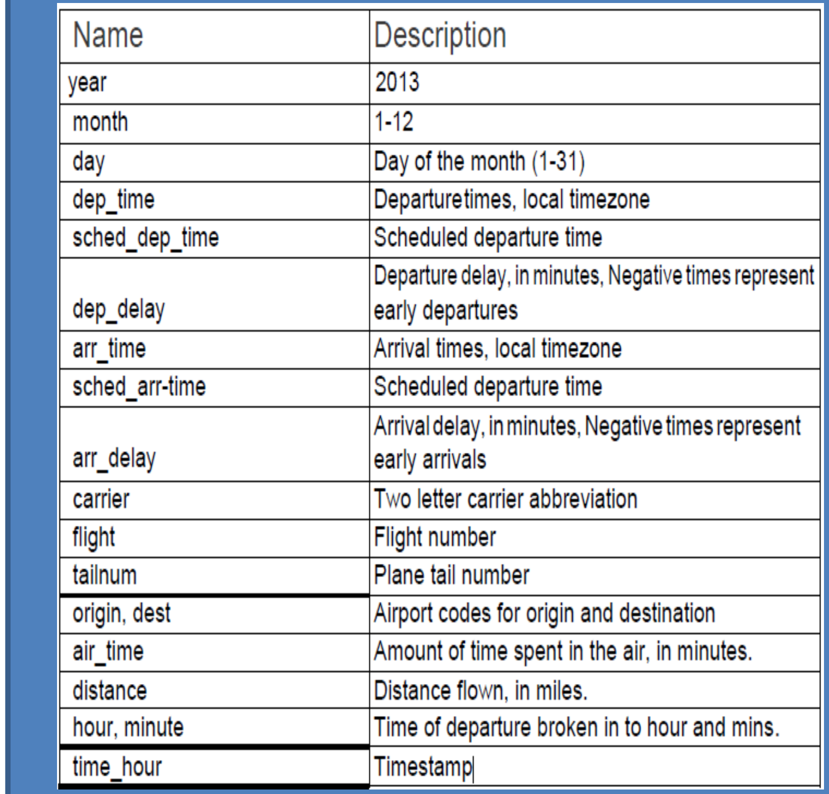
NYC Flight Data – Project Description

The Dataset

This dataset contains information about all flights that departed from NYC (e.g. EWR, JFK and LGA) in 2013: 336,776 flights in total

The objective is perform exploratory data analysis (EDA) to find hidden insides of the dataset

Dataset Attributes



Name	Description
year	2013
month	1-12
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	Arrival delay, in minutes, Negative times represent 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

NYC Flight Data – Business Questions to be addressed

Airport Performance

1. What is the concentration of Carriers at the origin airports?
2. Which are the best airports for on-time arrival of flights?
3. Which is the Best Airport for on-time departure?
4. Which month are have the least airline traffic?
5. What is the hourly traffic of flights?

Carrier Performance

1. Which are the 5 best airlines in terms of arrival delays?
2. Which airlines has fastest and slowest speed?

Flight Delays

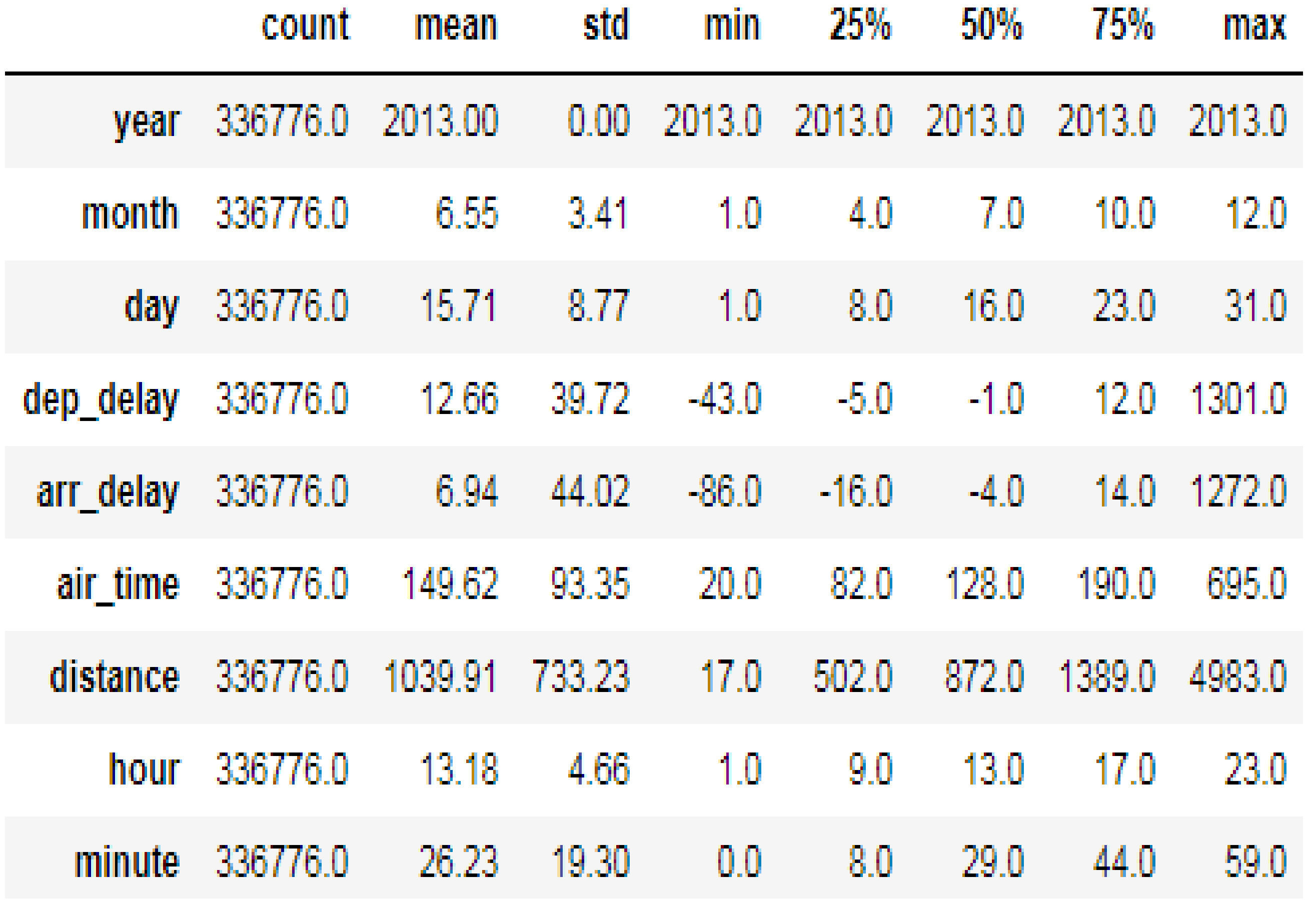
1. What is the pattern of delay of short, medium and long distance flights?
2. What is the average departure delay of each airline at Origin?
3. Which flights are delayed by more than 30 minutes?
4. On average how do delays of flights vary over the day?

Sector Performance

1. Which sectors are the most busy and which are least busy?
2. Is there a relation between flight delays and sectors congestion?

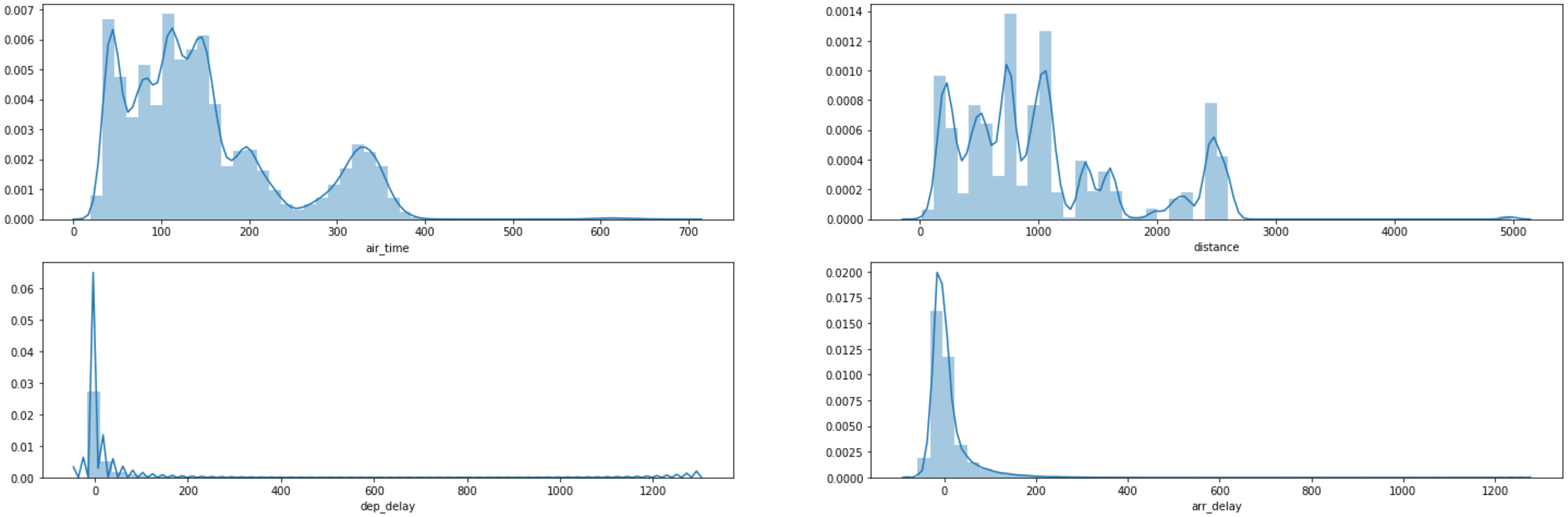
NYC Flight Data – Description of the dataset?

1. The dataset had 336776 rows. There were considerable NaN values
2. Departure and Arrival Delay seems to have some outliers as the Max value to too large
3. Distance seem to take some discrete values

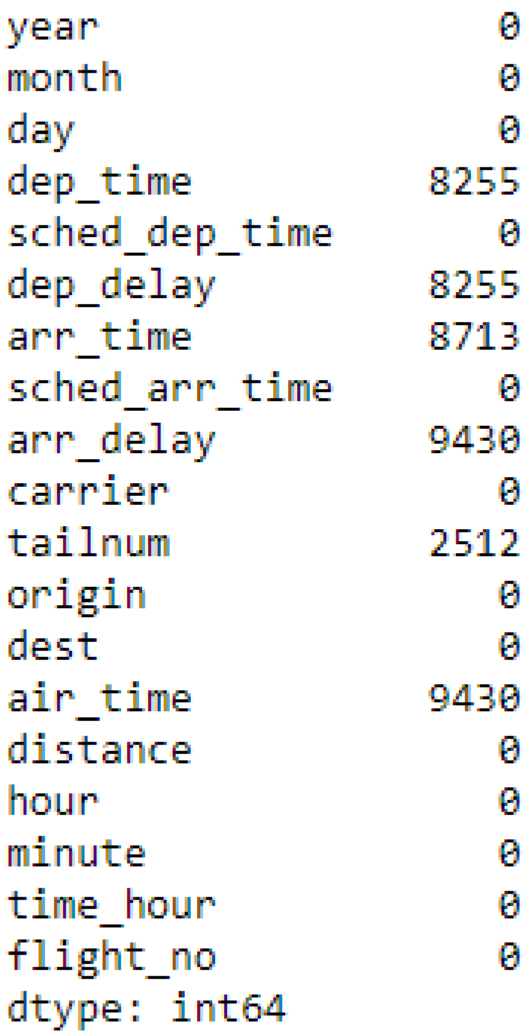


	count	mean	std	min	25%	50%	75%	max
year	336776.0	2013.00	0.00	2013.0	2013.0	2013.0	2013.0	2013.0
month	336776.0	6.55	3.41	1.0	4.0	7.0	10.0	12.0
day	336776.0	15.71	8.77	1.0	8.0	16.0	23.0	31.0
dep_delay	336776.0	12.66	39.72	-43.0	-5.0	-1.0	12.0	1301.0
arr_delay	336776.0	6.94	44.02	-86.0	-16.0	-4.0	14.0	1272.0
air_time	336776.0	149.62	93.35	20.0	82.0	128.0	190.0	695.0
distance	336776.0	1039.91	733.23	17.0	502.0	872.0	1389.0	4983.0
hour	336776.0	13.18	4.66	1.0	9.0	13.0	17.0	23.0
minute	336776.0	26.23	19.30	0.0	8.0	29.0	44.0	59.0

Air Quality – How good is the quality of data collected?



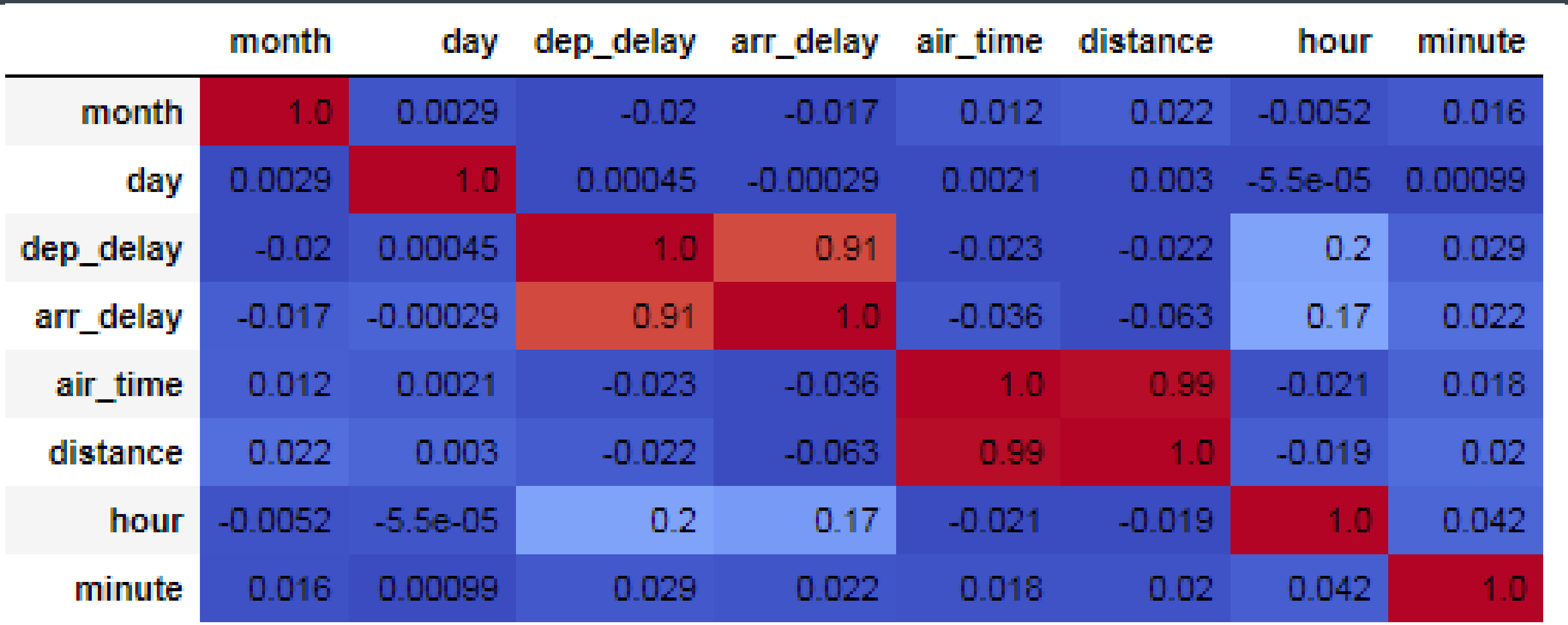
Missing Values



```
year                 0
month                0
day                  0
dep_time          8255
sched_dep_time       0
dep_delay         8255
arr_time          8713
sched_arr_time       0
arr_delay         9430
carrier              0
tailnum           2512
origin               0
dest                 0
air_time          9430
distance             0
hour                 0
minute               0
time_hour            0
flight_no            0
dtype: int64
```

- air time, distance: There seems to be lot of randomness in the distribution of data. The distance data takes discrete values
- Departure Delay: it takes a sudden peak value between 0 - 20 minutes and then tapers down. There are apparent outliers in the dataset
- Arrival Delay: It follows somewhat normal distribution with high kurtosis and +ve skewness

Air Quality – How good is the quality of data collected?



	month	day	dep_delay	arr_delay	air_time	distance	hour	minute
month	1.0	0.0029	-0.02	-0.017	0.012	0.022	-0.0052	0.016
day	0.0029	1.0	0.00045	-0.00029	0.0021	0.003	-5.5e-05	0.00099
dep_delay	-0.02	0.00045	1.0	0.91	-0.023	-0.022	0.2	0.029
arr_delay	-0.017	-0.00029	0.91	1.0	-0.036	-0.063	0.17	0.022
air_time	0.012	0.0021	-0.023	-0.036	1.0	0.99	-0.021	0.018
distance	0.022	0.003	-0.022	-0.063	0.99	1.0	-0.019	0.02
hour	-0.0052	-5.5e-05	0.2	0.17	-0.021	-0.019	1.0	0.042
minute	0.016	0.00099	0.029	0.022	0.018	0.02	0.042	1.0

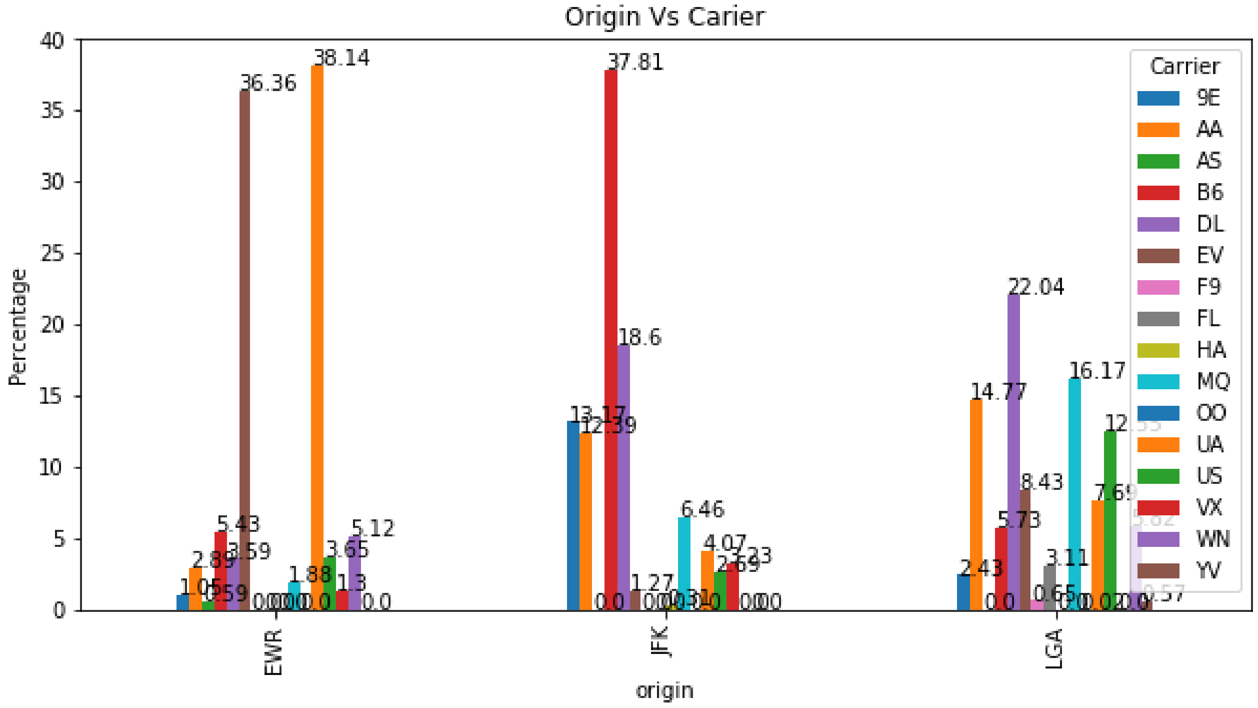
- Arrival Delay and Departure Delay are +vely correlated
- Distance and air time have high degree of correlation.
- Arrival delay and distance need to be dropped for machine learning algorithm implementation

Airport Performance: Concentration of Carriers at the origin airports?

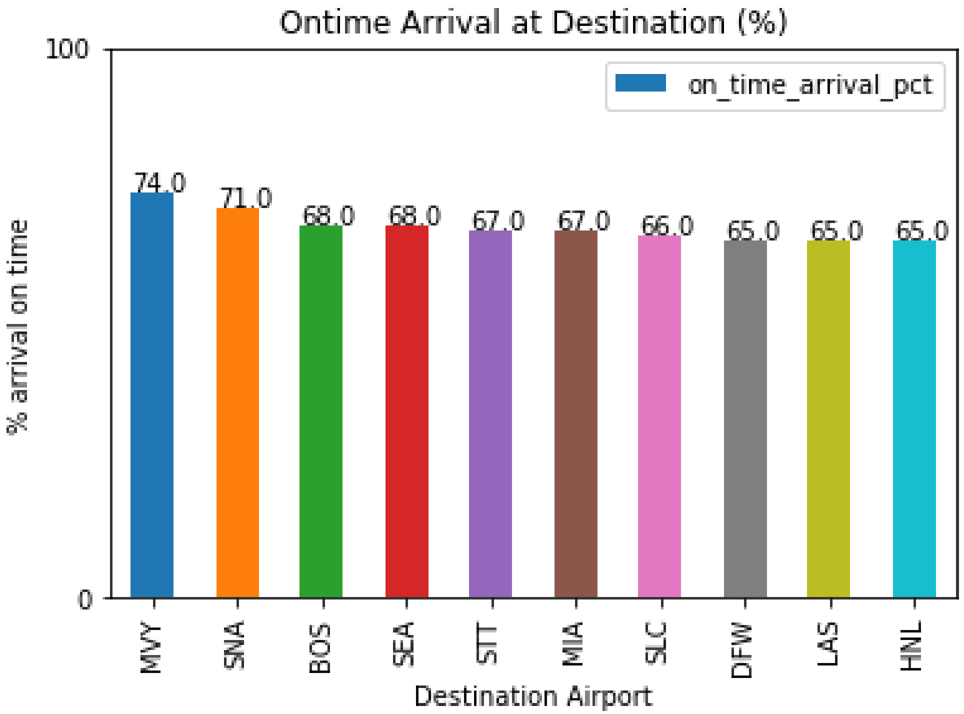
- LGA seems to have much more equal distribution of carrier than EWR and JFK. At least 7 carrier have 5% share in total flights

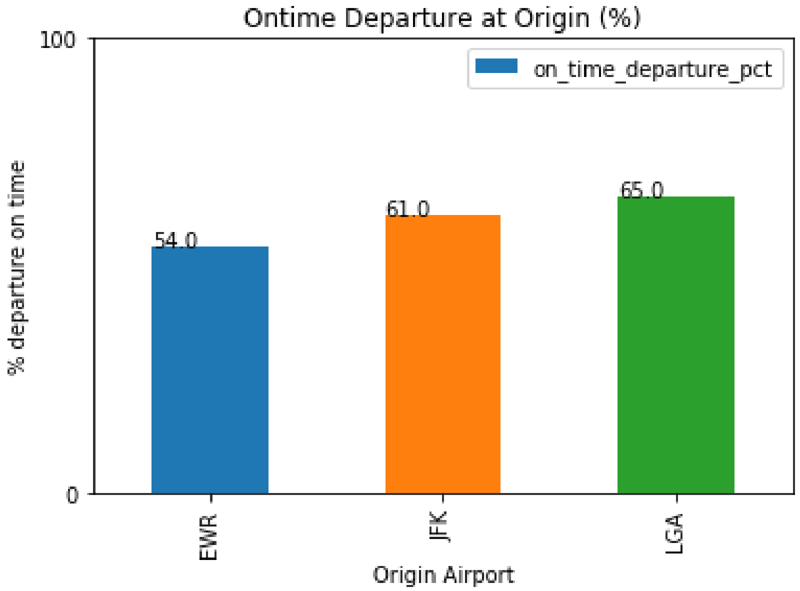
- JFK seem to have high number of B6 carrier(38%) and 4 other airlines have more than 5% share

- EWR seem to have high concentration of 2 carrier - YV(36%) and UA(38%). It has over reliance on these2 carriers



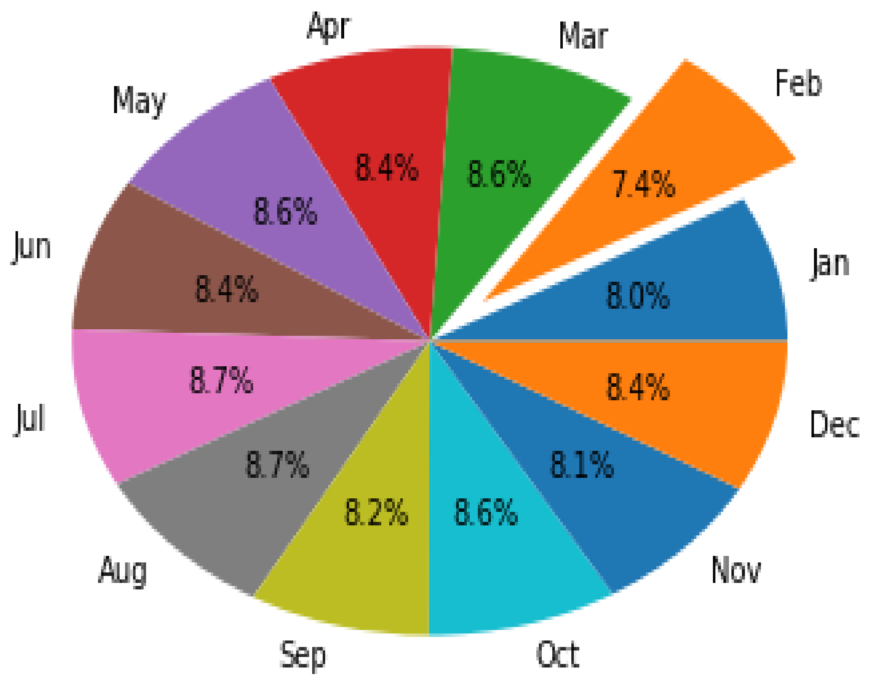
Airport Performance: Best airports for on-time arrival & departure of flights?





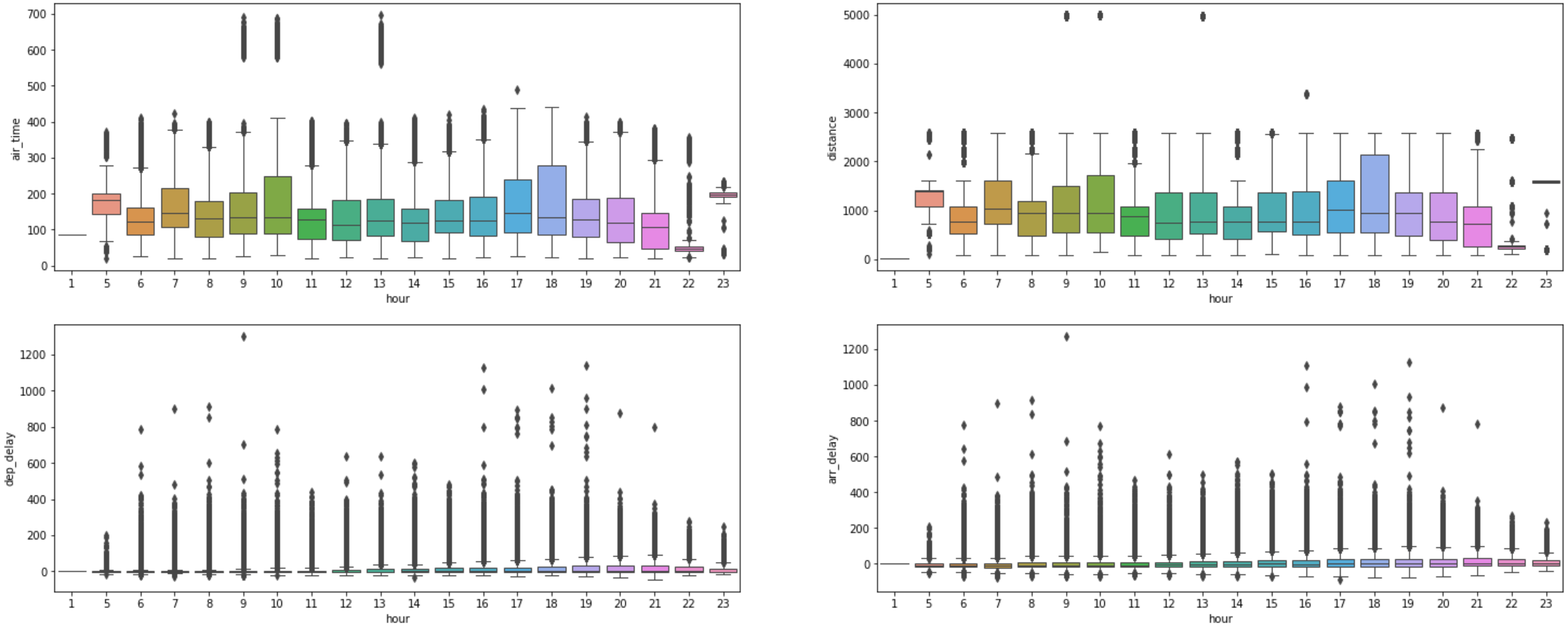
- MVY and SNA leads from other airports in terms of on time arrival with a 69% on-time flights
- However, there is not significant gap between the next 5-6 airports(SEA, MIA, STT, BOS, DFW, SLC, HNL, LAS) which range between 64% to 67%
- LGA is the best on time departure airport with a score of 65%

Airport Performance: Which month are have the least airline traffic?



- Feb is the most lean month in terms of air traffic.
- Air traffic is evenly distributed in the remaining months with 8-9 percent in each month
- Hence, if any planned maintenance is to be done, February should be a preferred choice

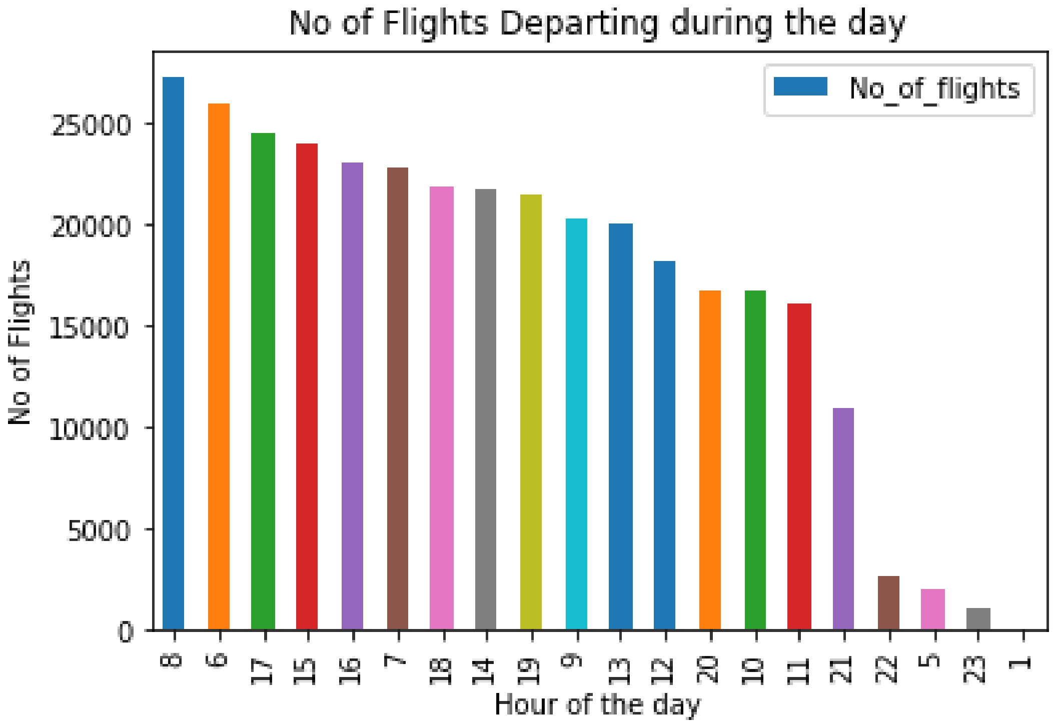
Airport Performance: What is the hourly traffic of flights?



- longer flight tend to be between 7 am to 10 am and again between 5 pm to 6 pm

Carrier Performance: What is the hourly traffic of flights?

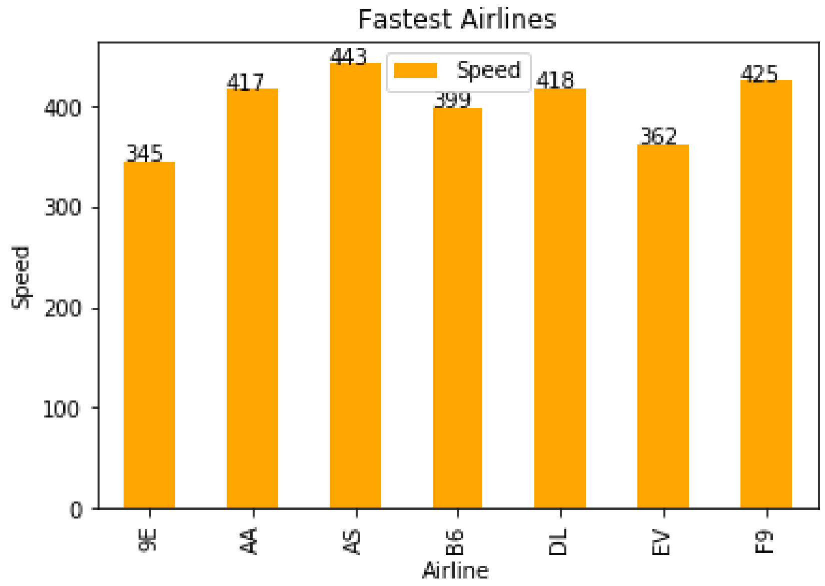
- The peak time traffic is in the time zone 8:00-9:00, 6:00-7:00, 17:00-18:00, 15:00-16:00 and 16:00-17:00 region
- There are no flights scheduled between 2:00-4:00 am
- Other lean period of the day are 23:00-24:00, 5:00-6:00 and 22:00-23:00. Any high priority maintenance can be planned in lean hours

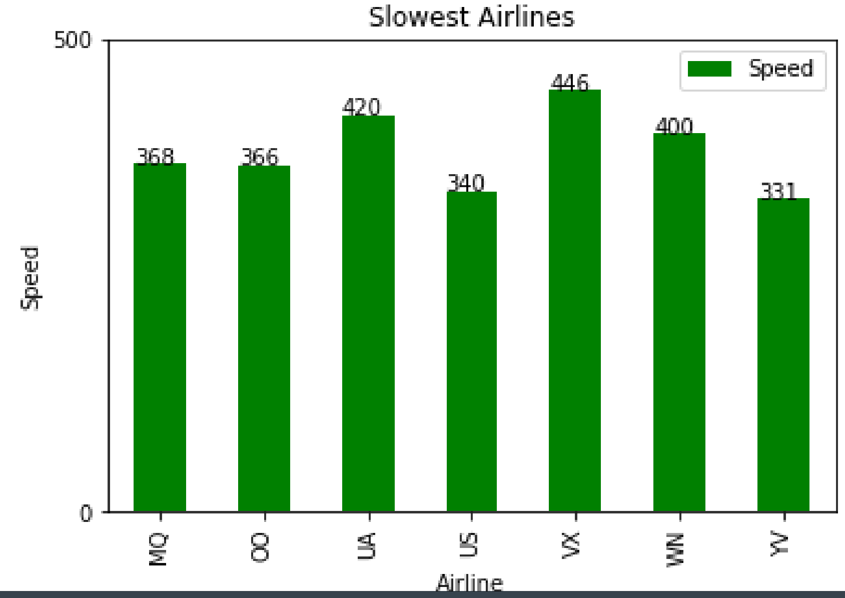


Carrier Performance : Which airlines has fastest and slowest speed?

1. AS airline is the fastest with 443 MPH of average speed. The next 6 fastest airlines are F9, DL, AA and B6 whose average speed is in the range of 399 to 425 MPH

2. YV is the slowest airline with an average speed of 331 MPH.



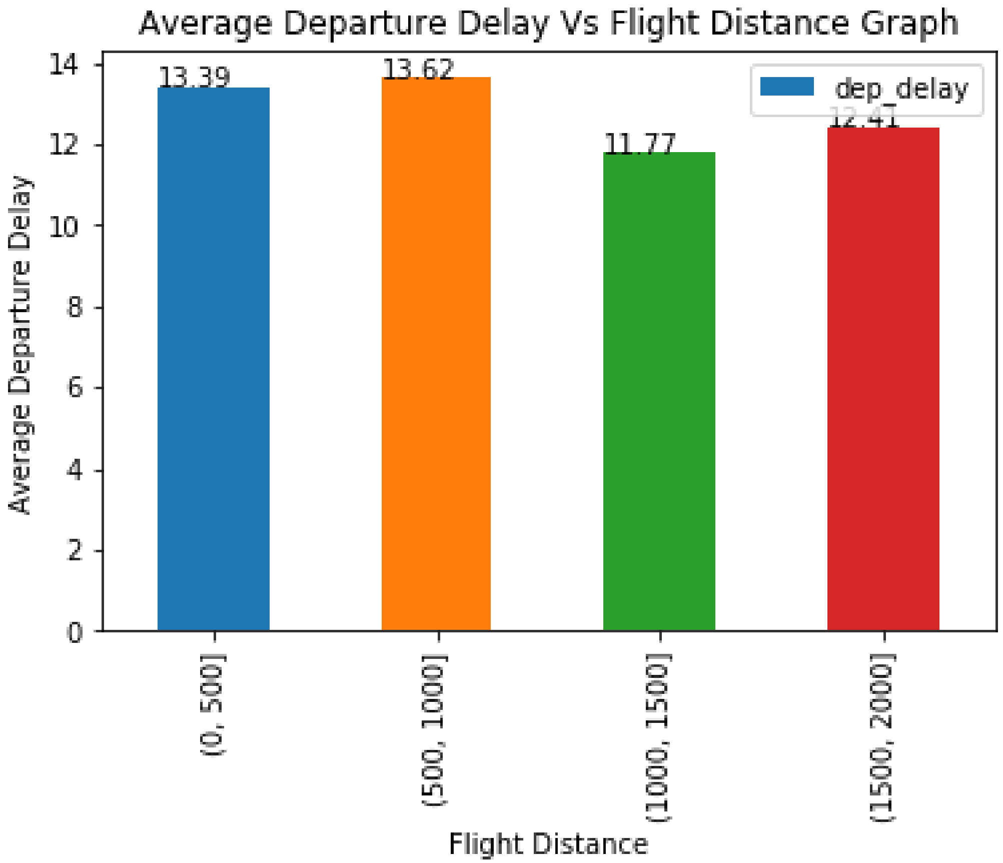


Flight Delays: Pattern of delay of short, medium and long distance flights?

- The departure delay of medium distance flights (1000 - 1500 miles) is the lowest (11.77 minutes)

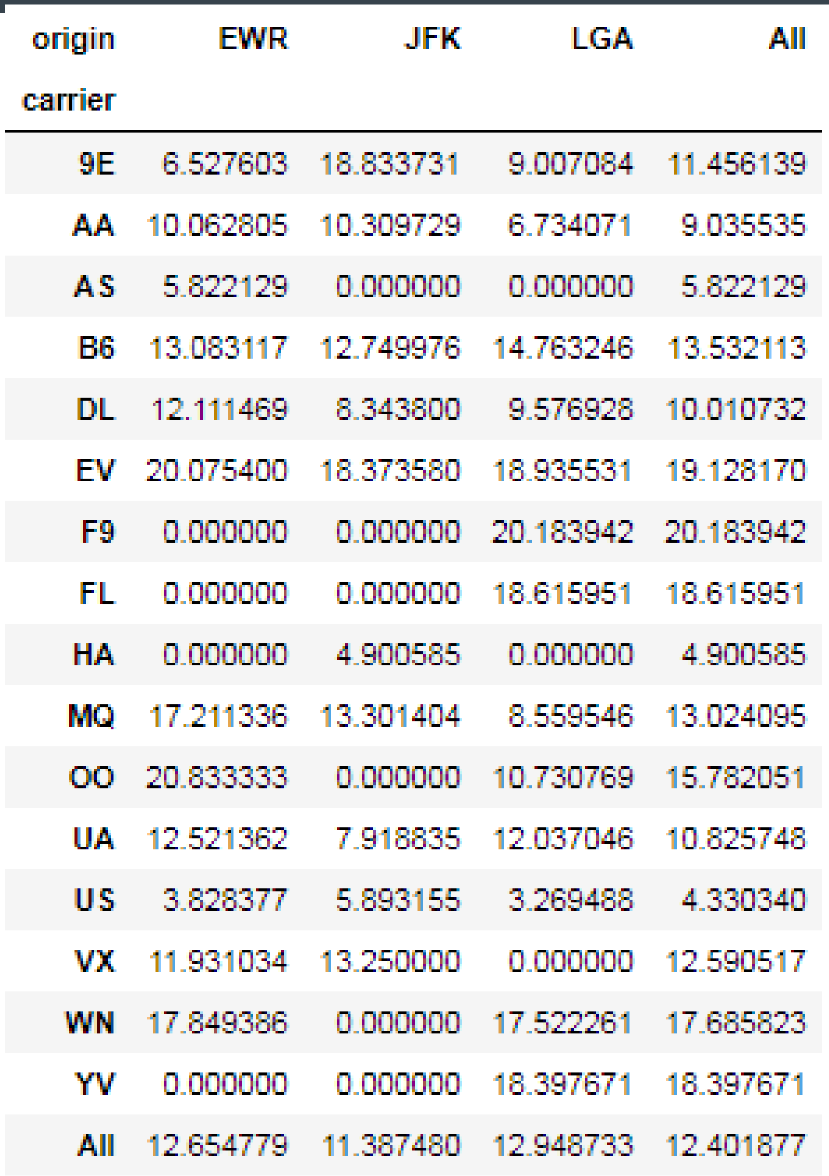
- The departure delay of short distance flights (500 - 1000 miles) is the highest (13.62 minutes)

- The delays are higher in short distance flights (< 1000 miles) than is long distance flights (>1000 miles)



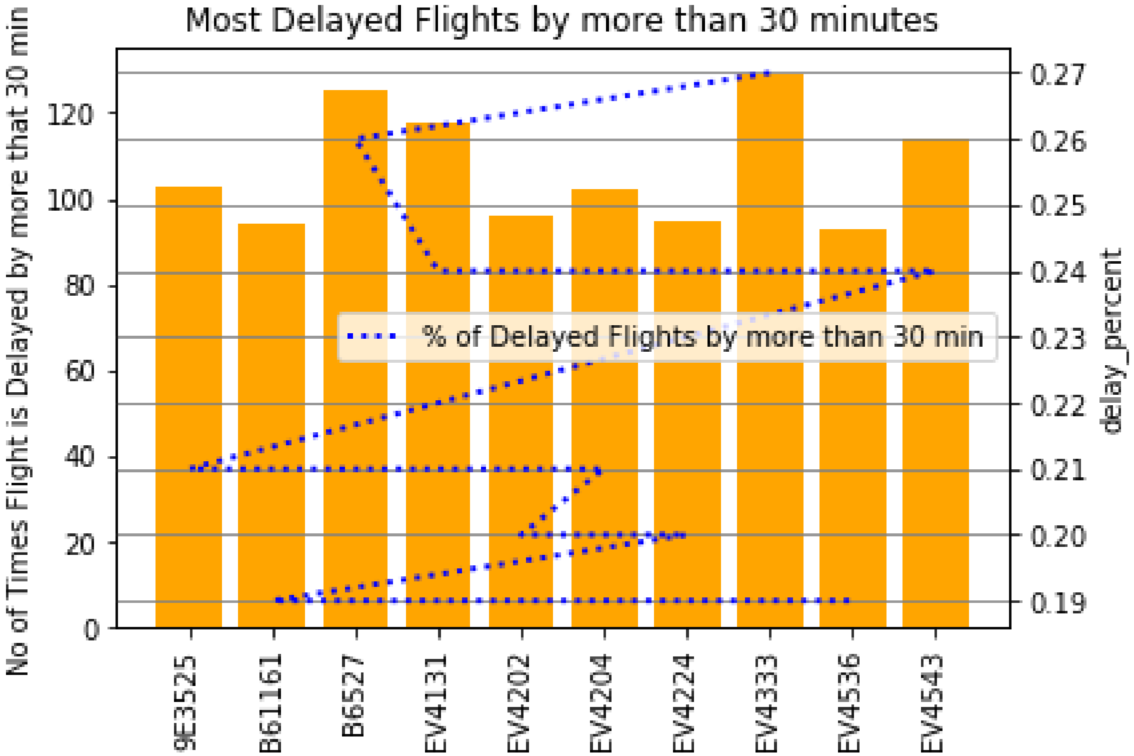
Flight Delays: What is the average departure delay of each airline at Origin?

- LGA has highest mean departure delay of 12.948 minutes across all flights while JFK has least mean departure delay
- Carrier F9 has the highest average departure delay of 20.18 minutes
- US airlines has least average departure delay of 4.33 minutes
- Best airlines at respective airports
  - EWR: US airlines – 3.82 minutes
  - JFK: HA airlines – 4.9 minutes
  - LGA: AA airlines – 6.73 minutes



origin	EWR	JFK	LGA	All
carrier				
9E	6.527603	18.833731	9.007084	11.456139
AA	10.062805	10.309729	6.734071	9.035535
AS	5.822129	0.000000	0.000000	5.822129
B6	13.083117	12.749976	14.763246	13.532113
DL	12.111469	8.343800	9.576928	10.010732
EV	20.075400	18.373580	18.935531	19.128170
F9	0.000000	0.000000	20.183942	20.183942
FL	0.000000	0.000000	18.615951	18.615951
HA	0.000000	4.900585	0.000000	4.900585
MQ	17.211336	13.301404	8.559546	13.024095
OO	20.833333	0.000000	10.730769	15.782051
UA	12.521362	7.918835	12.037046	10.825748
US	3.828377	5.893155	3.269488	4.330340
VX	11.931034	13.250000	0.000000	12.590517
WN	17.849386	0.000000	17.522261	17.685823
YV	0.000000	0.000000	18.397671	18.397671
All	12.654779	11.387480	12.948733	12.401877

Flight Delays: Which flights are delayed by more than 30 minutes?

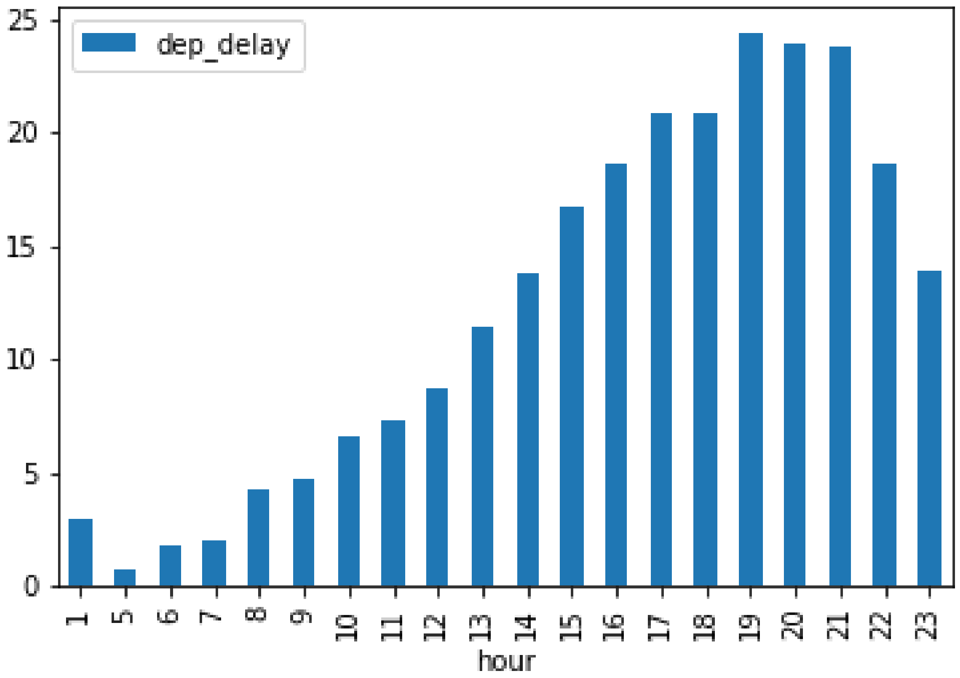


- EV4333 has the maximum no of flights delayed – 130 flights or 27%

- Flight B6527 comes next which is delayed approx 127 times or 26%

- Below is the list of other 8 airlines that had maximum frequency of delays:
  - EV4131 - 122 times or 26%
  - EV4543 – 117 times or 26%
  - 9E3525 – 103 times or 25.3%
  - EV4204 – 103 time or 25.3%
  - EV4202 – 98 times or 25%
  - EV4224 – 97 times or 24.8%
  - EV4536 – 96 times or 24.5%
  - B61161 – 97 times or 24.8%

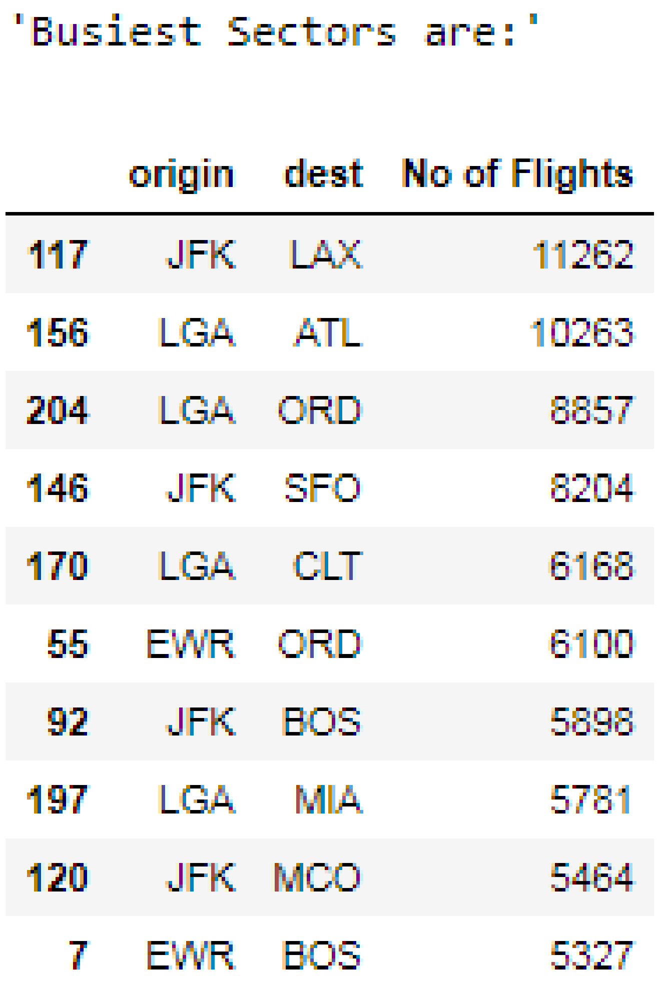
Flight Delays: How do delays of flights vary over the day?



- The departure delay tend to be longer as the day progresses. They peak during 17:00 - 22:00 hours to more than 17 minutes

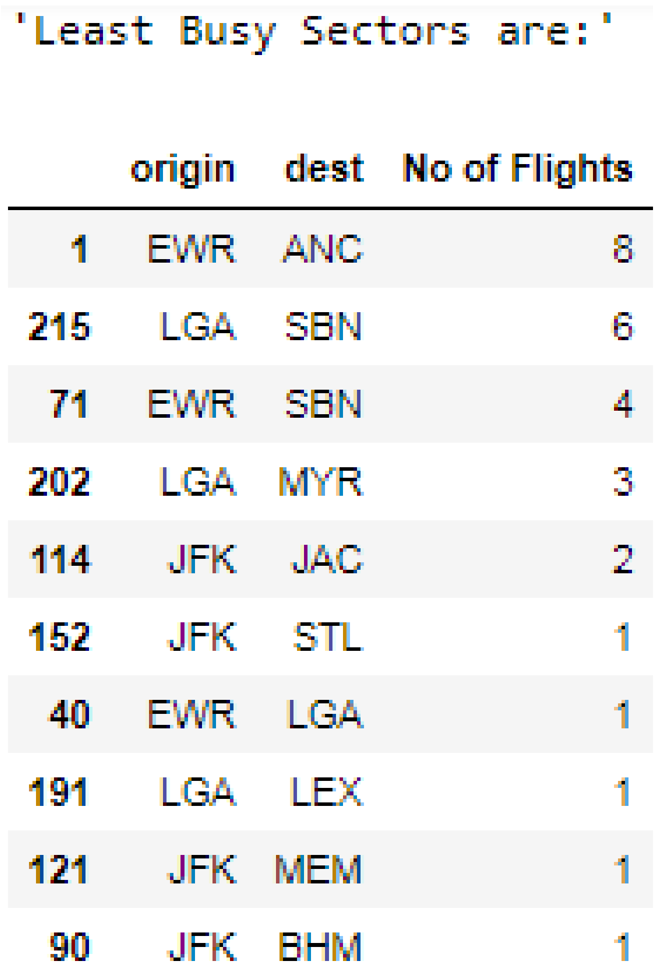
- Morning flights (6 am to 9 am) have east amount of departure delays (2-5 minutes)

Sector Performance: Which sectors are the most busy and which are least busy?



'Busiest Sectors are:'

	origin	dest	No of Flights
117	JFK	LAX	11262
156	LGA	ATL	10263
204	LGA	ORD	8857
146	JFK	SFO	8204
170	LGA	CLT	6168
55	EWR	ORD	6100
92	JFK	BOS	5898
197	LGA	MIA	5781
120	JFK	MCO	5464
7	EWR	BOS	5327



'Least Busy Sectors are:'

	origin	dest	No of Flights
1	EWR	ANC	8
215	LGA	SBN	6
71	EWR	SBN	4
202	LGA	MYR	3
114	JFK	JAC	2
152	JFK	STL	1
40	EWR	LGA	1
191	LGA	LEX	1
121	JFK	MEM	1
90	JFK	BHM	1

- JFK - LAX and LGA - ATL are the two busiest sectors with more than 10,000 flight operating in each sector

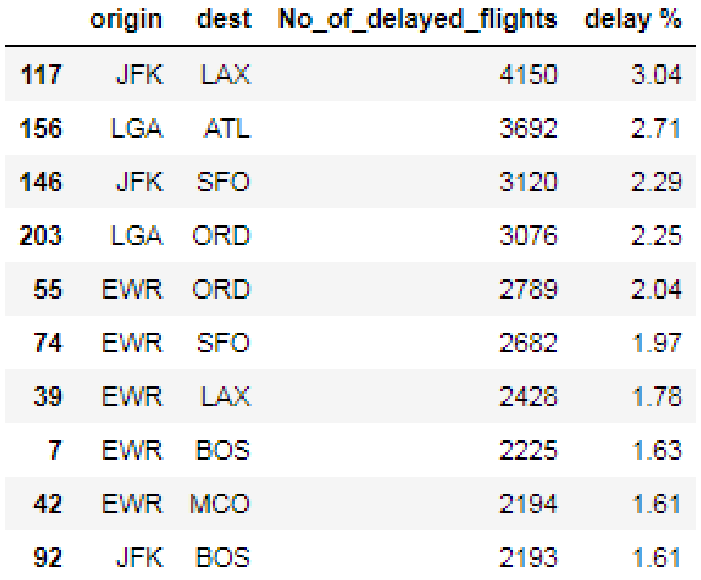
- LGA - ORD, JFK - SFO, LGA - CLT and EWR - ORD are next busiest sectors with flights operating between 8857 to 6100 respectively

- JFK - STL, JFK - LGA, LGA - LEX, JFK - MEM and JFK - BHM are the least popular sectors with just 1 fight operating in the entire year

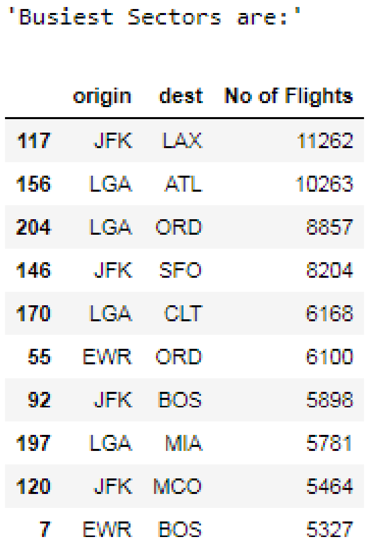
- Other least operating sectors are JFK - JAC, LGA - MYR, EWR - SBN, LGA - SBN and EWR - ANC with less that 10 flights

Is there a relation between flight delays and sectors congestion?

Top 10 sectors with highest flight delays



	origin	dest	No_of_delayed_flights	delay %
117	JFK	LAX	4150	3.04
156	LGA	ATL	3692	2.71
146	JFK	SFO	3120	2.29
203	LGA	ORD	3076	2.25
55	EWR	ORD	2789	2.04
74	EWR	SFO	2682	1.97
39	EWR	LAX	2428	1.78
7	EWR	BOS	2225	1.63
42	EWR	MCO	2194	1.61
92	JFK	BOS	2193	1.61



```
'Busiest Sectors are:'
```

	origin	dest	No of Flights
117	JFK	LAX	11262
156	LGA	ATL	10263
204	LGA	ORD	8857
146	JFK	SFO	8204
170	LGA	CLT	6168
55	EWR	ORD	6100
92	JFK	BOS	5898
197	LGA	MIA	5781
120	JFK	MCO	5464
7	EWR	BOS	5327

- JFK LAX sector has maximum number of delayed flights (4105) which account for 3.04% of all the flights that are delayed
- LGA ATL, JFK SFO, LGA ORD and EWR ORD, each of these sectors have very high number of delayed flights and each of them account for more than 2% of overall delayed flights

Relation between Flight Delays and Sectors

- **We observe complete correlation between no of flights operating in the sectors and highest flight delays in those sectors.**
- **It is clear that reason for flight delays is primarily the congestion of the sectors – higher the flights, higher the % of flights getting delayed in those sectors**

Thank You