Flights to Seattle

a) How many flights were there from NYC airports to Seattle in 2013?

```
SELECT COUNT (dest)
FROM rodriglr."flights.csv"
WHERE (dest='SEA' AND year=2013)
```

Answer: 3885

b) How many airlines fly from NYC to Seattle?

```
SELECT COUNT (distinct carrier) AS "Number of unique airlines"
FROM rodriglr."flights.csv"
WHERE dest='SEA'
```

Answer: 5

c) How many unique air planes fly from NYC to Seattle?

```
SELECT COUNT (DISTINCT tailnum) AS "Number of unique airplanes"
FROM rodriglr."flights.csv"
WHERE dest='SEA'
```

Answer: 933

d) What is the average arrival delay for flights from NYC to Seattle?

```
SELECT AVG (arr_delay) AS "Average arrival delay"
FROM rodriglr."flights.csv"
WHERE dest='SEA'
```

Answer: -1.0990990990990991

e) What proportion of flights to Seattle come from each NYC airport?

```
SELECT origin, (COUNT(dest)* 1.0 /
    (SELECT COUNT (origin)
    FROM rodriglr."flights.csv"
    WHERE dest = 'SEA')) AS "Proportion"
FROM rodriglr."flights.csv"
WHERE dest='SEA'
GROUP BY origin
```

origin	Proportion
JFK	0.53410553410553410553
EWR	0.46589446589446589447

Flight Delays

a) Which date has the largest average departure delay?

```
SELECT year, month, day, AVG(dep_delay) as Average_dep_delay
FROM rodriglr."flights.csv"
GROUP BY year, month, day
ORDER BY Average_dep_delay DESC
LIMIT 1
```

year	month	day	average_dep_delay
2013	3	8	83.6478696741854637

Which date has the largest average arrival delay?

```
SELECT year, month, day, AVG(arr_delay) as Average_arr_delay
FROM rodriglr."flights.csv"
GROUP BY year, month, day
ORDER BY Average_arr_delay DESC
LIMIT 1
```

year	month	day	average_arr_delay
2013	3	8	85.8621553884711779

b) What was the worst day to fly out of NYC in 2013 if you dislike delayed flights?

*I would consider the worst day to fly out of NYC as the one which has the highest count of flights were delayed, which comes out to be **23-Dec-2013**, which makes sense because people are flying out for their Christmas holidays and hence rush at airport is high*

```
SELECT year, month, day, COUNT(flight) AS number_of_delayed_flights
FROM rodriglr."flights.csv"
WHERE dep_delay > 0
GROUP BY year,month,day
ORDER BY number_of_delayed_flights DESC
LIMIT 1
```

year	month	day	number_of_delayed_flights
2013	12	23	673

c) Is Autumn (September, October, November) worse than Summer (June, July, August) for flight delays for flights from NYC?

Summer was worse than autumn in terms for average flight departure delays from NYC

Autumn Average Departure Delays: 6.0946001233501496

```
SELECT AVG ("Avg. Monthly Dep Delay") AS "Average Autumn Dep Delay"
FROM
    (SELECT month, AVG (dep_delay) AS "Avg. Monthly Dep Delay"
    FROM rodriglr."flights.csv"
    WHERE month IN (9,10,11)
    GROUP BY month) as temp
```

Summer Average Departure Delays: 18.2727723593803359

```
SELECT AVG ("Avg. Monthly Dep Delay") AS "Average Summer Dep Delay"
FROM
    (SELECT month, AVG (dep_delay) AS "Avg. Monthly Dep Delay"
    FROM rodriglr."flights.csv"
    WHERE month IN (6,7,8)
    GROUP BY month) as temp
```

d) On average, how do departure delays vary over the course of a day? You can compute the average delay by hour of day, such that your result will have 24 records (be careful -- there are records with hour 0 and hour 24. Consider lumping these together or justify any other solution you come up with.)

```
SELECT
    (CASE
    WHEN hour = 24 THEN 0
    ELSE hour
    END) AS grouped_hour,
AVG(dep_delay) as Mean_Hourly_Delays
FROM rodriglr."flights.csv"
GROUP BY grouped_hour
```

grouped_hour	mean_hourly_delays
0	127.2232044196895026
1	206.7556561085972851
2	236.2539682539682540
3	304.7272727272727273
4	-5.5540983606557377
5	-4.3562932226832642
6	-1.5218102267202899
7	0.21472278013919379700
8	1.09231236014715363902
9	4.2341126461211477
10	5.5110722974237415
11	5.6132719004308281
12	7.5173489765351972
13	9.3639062036212526
14	8.0518289693046975
15	10.5933136589677990
16	13.5572495053067098
17	16.6557466309723672
18	18.4746655479420128
19	21.3102007951285793
20	28.0875939616077530
21	41.8441451346893898
22	67.9586156381615089
23	96.6384202453987730

It looks like flights from midnight till 3 am are the worst when it comes to average departure delays. This could be possibly due to shift changes, less crew, etc. during early morning hours at the airports.

Velocity

Which flight departing NYC in 2013 flew the fastest?

```
SELECT year, month, day, carrier, tailnum, flight, origin, dest,
    (distance * 1.0 / air_time) AS Miles_per_minute
FROM rodriglr."table_flights.csv"
ORDER BY Miles_per_minute DESC
LIMIT 1
```

year	month	day	carrier	tailnum	flight	origin	dest	miles_per_minute
2013	5	25	DL	N666DN	1499	LGA	ATL	11.7230769230769231

*Flight **DL 1499 (tail number N666DN),** from LGA to ATL which flew on date **25-May-2013** flew with max speed of **11.72 miles/minutes***

Routine Flights

Which flights (i.e. carrier + flight + dest) happen every day?

```
SELECT newname, COUNT (DISTINCT newdate)
FROM
    (SELECT CONCAT (day, '-', month, '-', year) AS newdate,
            CONCAT(carrier, ' ', flight, ' ', dest) AS newname
    FROM rodriglr."table_flights.csv"
    GROUP BY newname,newdate) AS temp
GROUP BY newname
ORDER BY COUNT DESC
LIMIT 1
```

Flight B6 – 1783 flying to MCO from New York operates 365 days a year.

newname	count
B6 1783 MCO	365

Open-ended: Research Question

Which two airlines flying to Seattle were most reliable in terms of having the minimum average departure delay in 2013?

Explanation: *This is very important for people to know which airline options they should be looking for when they are booking flights to Seattle from New York, since it has the best reputation in terms of least average departure delay and flying higher number of flights.*

```
SELECT carrier, dest, ROUND(AVG (dep_delay),2) AS Mean_Dep_Delay,
        COUNT(*) AS Number_of_Operated_flights
FROM rodriglr."table_flights.csv"
WHERE dest='SEA'
GROUP BY carrier, dest
ORDER BY Mean_Dep_Delay ASC
```

carrier	dest	mean_dep_delay	number_of_operated_flights
AS	SEA	5.83	709
DL	SEA	6.98	1202
AA	SEA	10.10	360
B6	SEA	11.59	513
UA	SEA	17.32	1101

From the results above, the best airlines to fly to Seattle are Alaska and Delta as they have nearly equal departure delays which are lowest among all the carriers flying to Seattle, as well as they operate greater number of flights while still holding the reputation of least delays. So, if the person must be assured that on any given day he can fly to Seattle from NYC and wants to fly with minimum departure delays, then the two options that he should be first looking out for are Alaska & Delta.

Exogenous effects

Is there any link between visibility and delay? What about temperature?

Answer: *Let's check the visibility and temperature data for top few flights* ***delayed by more than 2 hours*** *(threshold chosen for this problem as humans become impatient after this)*

```
SELECT carrier, flight, tailnum, AVG(temp) as avg_temp,
        AVG(visib) as avg_visib, AVG(dep_delay) as avg_delay
FROM
    (SELECT *
    FROM rodriglr."table_weather.csv" as weather
    LEFT JOIN rodriglr."table_flights.csv" as flight
        ON (weather.month = flight.month AND
            weather.day = flight.day AND
            weather.hour = flight.hour AND
            weather.origin = flight.origin)) as temp
WHERE dep_delay > 120
GROUP BY carrier, flight,tailnum
ORDER BY avg_delay desc
```

carrier	flight	tailnum	avg_temp	avg_visib	avg_delay
MQ	3695	N517MQ	39.9200000000000000	10.0000000000000000	1126.0000000000000000
AA	172	N5DMAA	48.0200000000000000	4.0000000000000000	896.0000000000000000
MQ	3744	N523MQ	55.0400000000000000	10.0000000000000000	878.0000000000000000
DL	1223	N375NC	24.9800000000000000	9.0000000000000000	849.0000000000000000
AA	172	N5EMAA	24.9800000000000000	10.0000000000000000	845.0000000000000000
DL	2042	N990AT	50.0000000000000000	10.0000000000000000	798.0000000000000000
DL	575	N348NW	46.0400000000000000	10.0000000000000000	786.0000000000000000
DL	502	N943DL	64.0400000000000000	10.0000000000000000	702.0000000000000000
DL	2343	N338NB	42.0800000000000000	10.0000000000000000	592.0000000000000000
AA	1895	N3EMAA	84.9200000000000000	10.0000000000000000	580.0000000000000000
EV	4711	N12163	57.9200000000000000	5.0000000000000000	548.0000000000000000
DL	2343	N310DE	48.9200000000000000	10.0000000000000000	545.0000000000000000
US	1745	N180US	69.9800000000000000	10.0000000000000000	486.0000000000000000

*Let's check the visibility and temperature data for top few flights **which departed between on-time till one hour early** (threshold chosen for this problem as this is easily accommodated by customers most of whom come at least an hour early to airport)*

```
SELECT carrier, flight, tailnum, AVG(temp) as avg_temp,
        AVG(visib) as avg_visib, Round(AVG(dep_delay),2) as avg_delay
FROM
    (SELECT *
    FROM rodriglr."table_weather.csv" as weather
    LEFT JOIN rodriglr."table_flights.csv" as flight
        ON (weather.month = flight.month AND
            weather.day = flight.day AND
            weather.hour = flight.hour AND
            weather.origin = flight.origin)) as temp
WHERE dep_delay BETWEEN -60 AND 0
GROUP BY carrier, flight,tailnum
ORDER BY avg_delay desc
```

carrier	flight	tailnum	avg_temp	avg_visib	avg_delay
AA	2075	N528AA	41.0000000000000000	10.0000000000000000	0.00
AA	2147	N3ABAA	78.0800000000000000	10.0000000000000000	0.00
AA	1841	N581AA	71.0600000000000000	10.0000000000000000	0.00
AA	2075	N4YKAA	55.9400000000000000	10.0000000000000000	0.00
AA	1895	N5CBAA	19.0400000000000000	10.0000000000000000	0.00
AA	2083	N586AA	75.9200000000000000	10.0000000000000000	0.00
AA	1507	N571AA	84.0200000000000000	10.0000000000000000	0.00
AA	1507	N555AA	80.9600000000000000	10.0000000000000000	0.00
AA	1623	N3GVAA	37.0400000000000000	7.0000000000000000	0.00
AA	1841	N4XXAA	75.0200000000000000	9.0000000000000000	0.00
AA	1853	N4XRAA	24.9800000000000000	10.0000000000000000	0.00
AA	1895	N3CEAA	48.9200000000000000	10.0000000000000000	0.00
AA	2075	N4WSAA	44.0600000000000000	10.0000000000000000	0.00

From mere inspection of both the tables, it doesn't come out clearly if there is any relation between temperature, visibility and departure delay, since the values look pretty much scattered. However, I am sure that if we find Pearson Coefficient there might be some inference towards the fact that at least visibility might be affecting the flights during winter months.