# IMT 575

# Flights in SQL

# Prem Shah

1. **Flights to Seattle:**
   1. How many flights were there from NYC airports to Seattle in 2013?

SELECT count(\*)

FROM rodriglr."table\_flights.csv"

WHERE dest = 'SEA'

ANS: 3885

* 1. How many airlines fly from NYC to Seattle?

SELECT count(DISTINCT carrier)

FROM rodriglr."table\_flights.csv"

WHERE dest = 'SEA'

ANS: 5

* 1. How many unique air planes fly from NYC to Seattle?

SELECT count(DISTINCT tailnum) AS 'No. of unique airlines'

FROM rodriglr."table\_flights.csv"

WHERE dest = 'SEA'

ANS: 933

* 1. 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'

ANS: -1.0990990

* 1. What proportion of flights to Seattle come from each NYC airport?

SELECT (count(dest)\*1.0/

(SELECT COUNT(\*)

FROM rodriglr."flights.csv"

WHERE dest = 'SEA')) as Percentage\_of\_flights, origin

FROM rodriglr."flights.csv"

WHERE dest = 'SEA'

GROUP BY origin

ANS:



1. **Flights Delays**
   1. Which date has the largest average departure delay? Which date has the largest average arrival delay?

SELECT year, month, day, AVG (arr\_delay) AS "average\_arrival\_delay"

FROM rodriglr."flights.csv"

GROUP BY year,month,day

ORDER BY Average\_arrival\_delay DESC

LIMIT 1

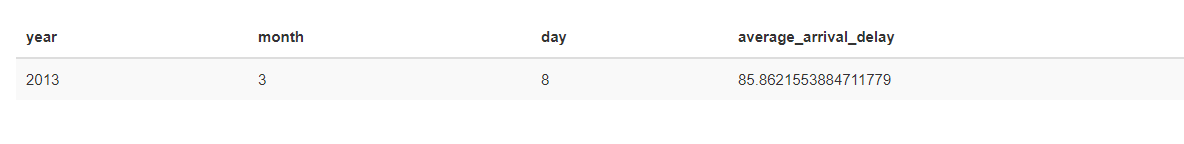
SELECT year, month, day, AVG (dep\_delay) AS "average\_departure\_delay"

FROM rodriglr."flights.csv"

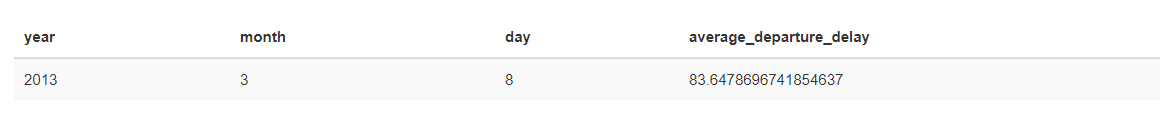
GROUP BY year,month,day

ORDER BY Average\_departure\_delay DESC

LIMIT 1

Average arrival delay:

Average departure delay:



* 1. What was the worst day to fly out of NYC in 2013 if you dislike delayed flights?  (This one is less straightforward in SQL than you may expect.)

I tried to execute this query in two ways. One is maximum delay in a particular year, month and day only considering delay > 0.

The other way was to select delay\_per\_flight

SELECT year, month, day ,( AVG (dep\_delay)\* 1.0 / (COUNT(flight))) as delay\_per\_flight

FROM rodriglr."flights.csv"

WHERE dep\_delay >0

GROUP BY year,month,day

ORDER BY delay\_per\_flight DESC

LIMIT 1



SELECT year, month, day ,AVG (dep\_delay) as delay\_per\_flight

FROM rodriglr."flights.csv"

WHERE dep\_delay >0

GROUP BY year,month,day

ORDER BY ratio DESC

LIMIT 1



Both of the results point to 12th September.

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

From the below results, you can see that Summer flight delays are almost three times than Autumn flight delays.

SELECT AVG (dep\_delay) AS "Avg. Autumn Monthly Dep Delay"

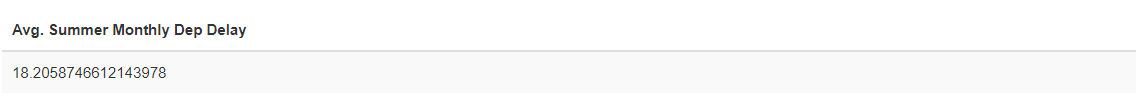
FROM rodriglr."flights.csv"

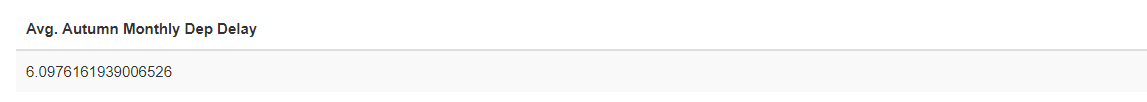
WHERE month IN (9,10,11)

SELECT AVG (dep\_delay) AS "Avg. Monthly Dep Delay"

FROM rodriglr."flights.csv"

WHERE month IN (6,7,8)



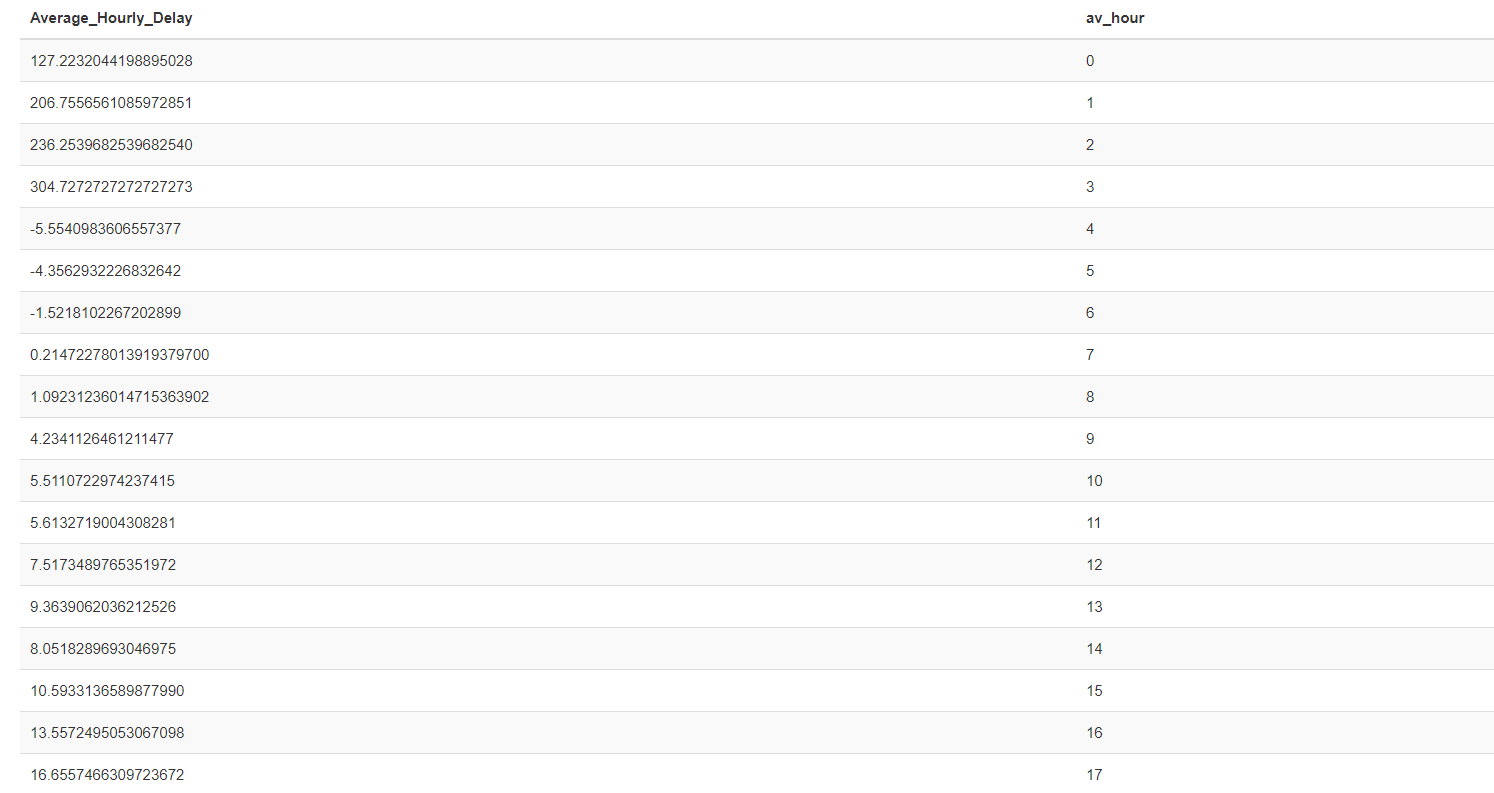


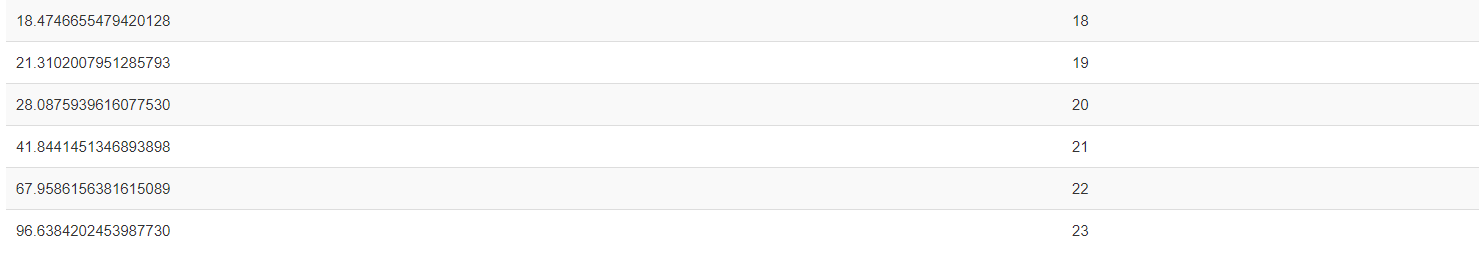
* 1. On average, how do departure delays vary over the course of a day?  You can compute the average delay by hour of day, such that your result will have 24 records (be careful -- there are records with hour 0 and hour 24. Consider lumping these together, or justify any other solution you come up with.)  No need to plot the results.

 SELECT AVG (dep\_delay) AS "Average\_Hourly\_Delay", (CASE WHEN hour = 24 THEN 0 ELSE hour END) as av\_hour

FROM rodriglr."flights.csv"

GROUP BY av\_hour





From the above results, you can see that the delay is high in early mornings and late nights.

1. **Velocity**

Which flight departing NYC in 2013 flew the fastest

1

From the below results, you can see flight 1499 has the highest speed

SELECT year, month, day, flight, tailnum, (distance\* 60.0/air\_time) as mph\_speed

FROM rodriglr."flights.csv"

ORDER BY mph\_speed DESC

LIMIT 1



1. **Routine Flights:**

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

Here, I see the total distinct combination number of days and month the flight has flown. If it is 365, the flight has flown on every single day of the year 2013.

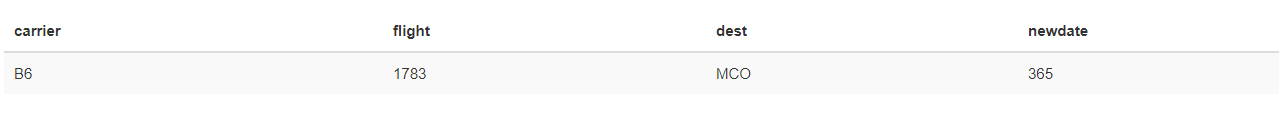
SELECT carrier, flight,dest , COUNT(DISTINCT (CONCAT (day, '-', month, '-', year))) AS newdate

FROM rodriglr."table\_flights.csv"

GROUP BY carrier,flight,dest

ORDER BY newdate DESC

LIMIT 1



1. **Open Ended Research Question**

For flights from New York to Seattle, which airlines have the best performance in terms of delays?

This question is interesting because it helps people who do not like flight delays and who want the best flight in terms of the least arrival and departure delay.

As you can see from the above plots, AS (Alaska Airlines) has the least average departure delay and the least average arrival delay as well while United Airlines (UA) has the highest average departure delay for flights to Seattle. B6 has the highest average arrival delay for flights from NYC to Seattle. Hence, people who want departures & arrivals on time should choose Alaska Airlines. This inference have one problem which we did not consider is that since Alaska Airlines is based out of Seattle, hence might have more flights in this route which might have resulted in the lower departure delay. But as we can see from the below table, that is not the case. Hence we can safely say that Alaska Airlines has a good track record of arriving and departing on time.

SELECT ROUND(AVG(dep\_delay),2) as avg\_dep\_delay,

ROUND(AVG(arr\_delay),2) as avg\_arr\_delay,

carrier

FROM rodriglr."flights.csv" f

WHERE dest='SEA'

group by carrier

ORDER by avg\_arr\_delay, avg\_dep\_delay DESC

SELECT COUNT(tailnum),

carrier

FROM rodriglr."flights.csv" f

WHERE dest='SEA'

group by carrier

ORDER by COUNT(tailnum) ASC

SELECT ROUND(((AVG(dep\_delay) + AVG(arr\_delay))/2/COUNT(tailnum)),7) as avg\_total\_delay\_per\_flight,

ROUND(AVG(dep\_delay),2) as average\_dep\_delay,

Round(AVG(arr\_delay),2) as average\_arr\_delay,

COUNT(tailnum) as total\_flights,

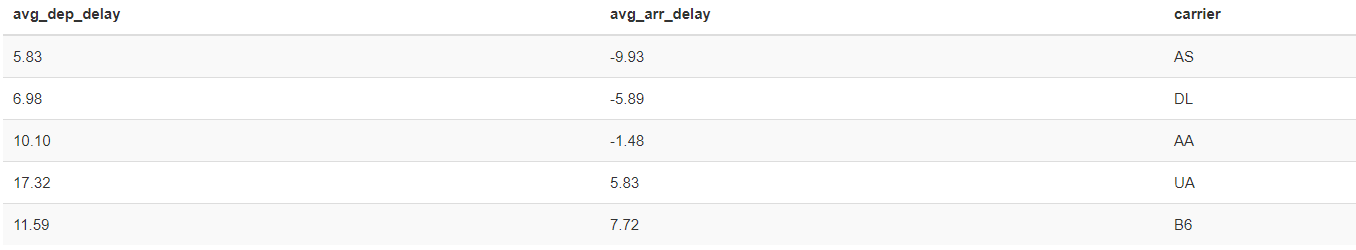
carrier

FROM rodriglr."flights.csv" f

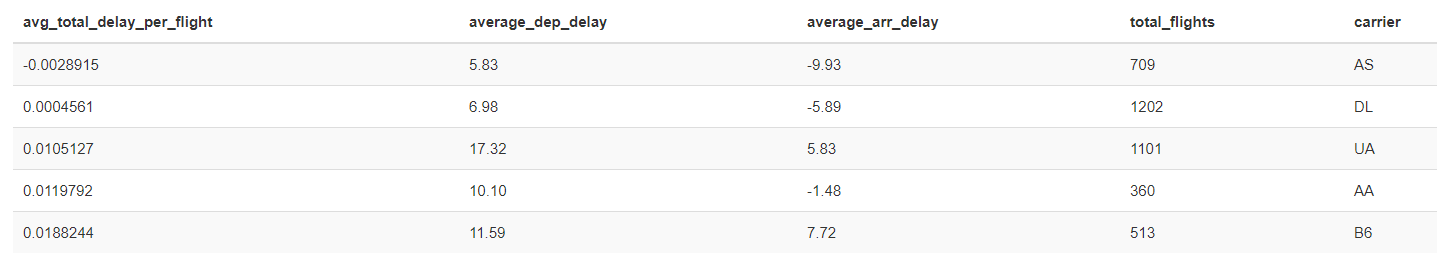
WHERE dest='SEA'

group by carrier

ORDER by avg\_total\_delay\_per\_flight ASC







1. **Exogenous Effects**

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

SELECT ROUND(AVG(dep\_delay),2) as avg\_delay,

ROUND(AVG(visib),2) as avg\_visib,

ROUND(AVG(temp),2) as avg\_temp,

flight, tailnum

FROM rodriglr."flights.csv" f

INNER JOIN rodriglr."weather.csv" w

ON f.year = w.year AND

f.month = w.month AND

f.day = w.day AND

f.hour = w.hour

WHERE dep\_delay > 60

group by flight,tailnum

ORDER by avg\_delay ASC

SELECT ROUND(AVG(dep\_delay),2) as avg\_delay,

ROUND(AVG(visib),2) as avg\_visib,

ROUND(AVG(temp),2) as avg\_temp,

f.month, f.day

FROM rodriglr."flights.csv" f

INNER JOIN rodriglr."weather.csv" w

ON f.year = w.year AND

f.month = w.month AND

f.day = w.day AND

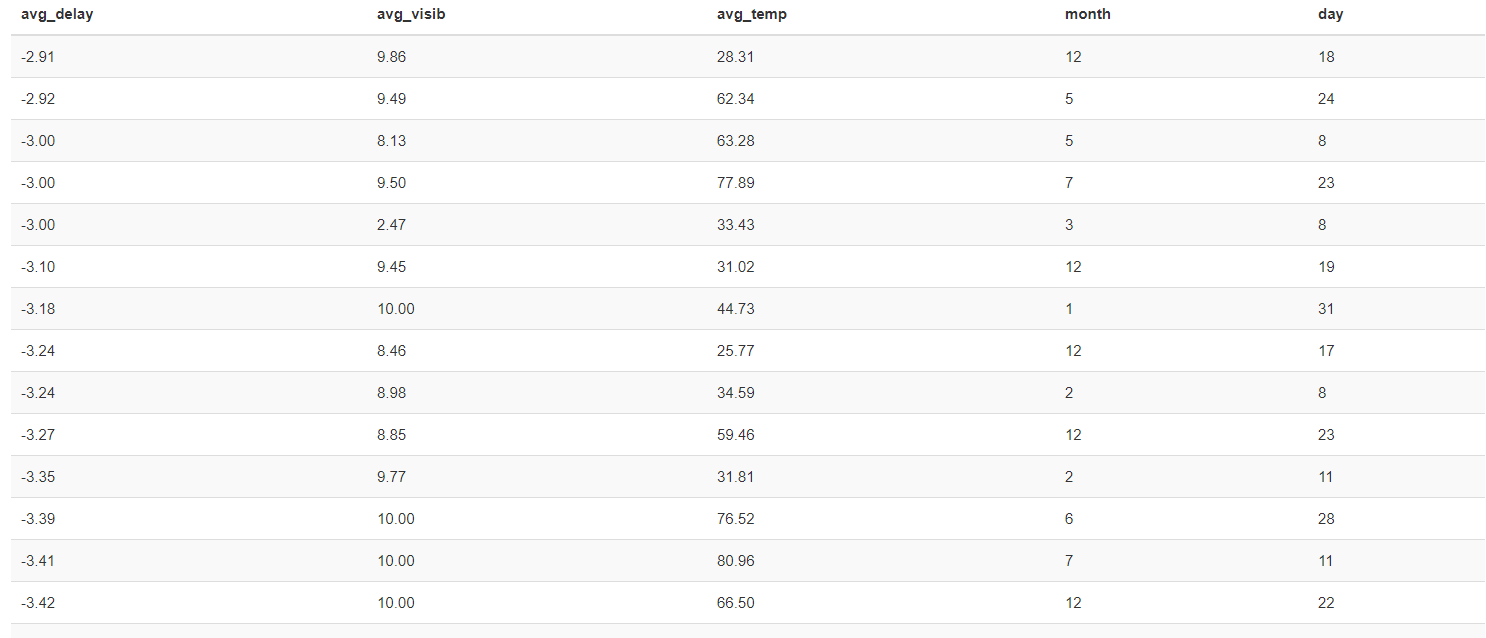
f.hour = w.hour

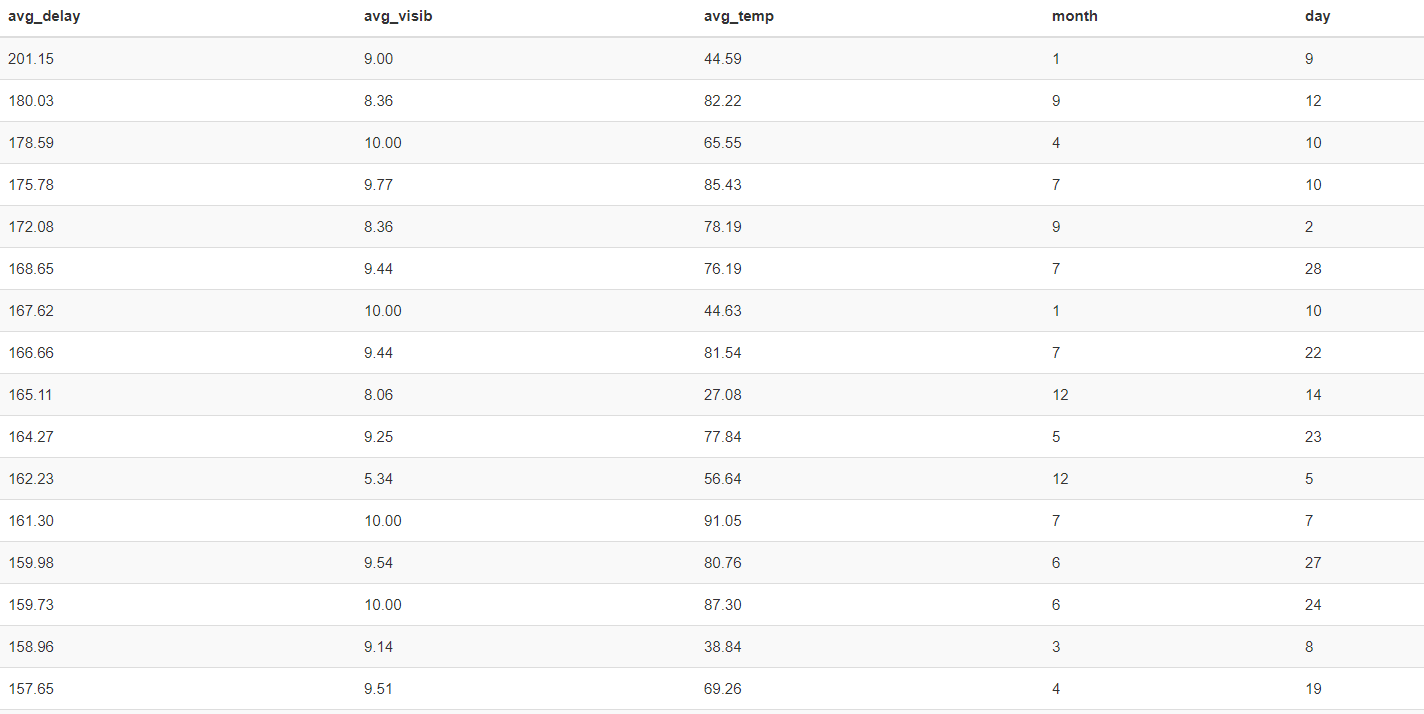
WHERE dep\_delay BETWEEN -15 and 0

group by f.month, f.day

ORDER by avg\_delay DESC

I define delayed as anything more than 60 minutes delay and on time as anything with delay between -15 and 0.





From the above results, I do not see any particular trends between visibility, temperature and delay.