BAX 421 Final Project
Business Questions with Queries

Business questions:

1.To better select the flight routes, which are the most **popular flight routes** from New York?

2.To optimize flights and seats availability, when is the peak **season/hour** ?

3.To increase customer satisfaction, what's the **delay** and **cancellation rate**?

4.To improve flight schedule accuracy, what's the **relationship between weather and delay/cancellation** ?

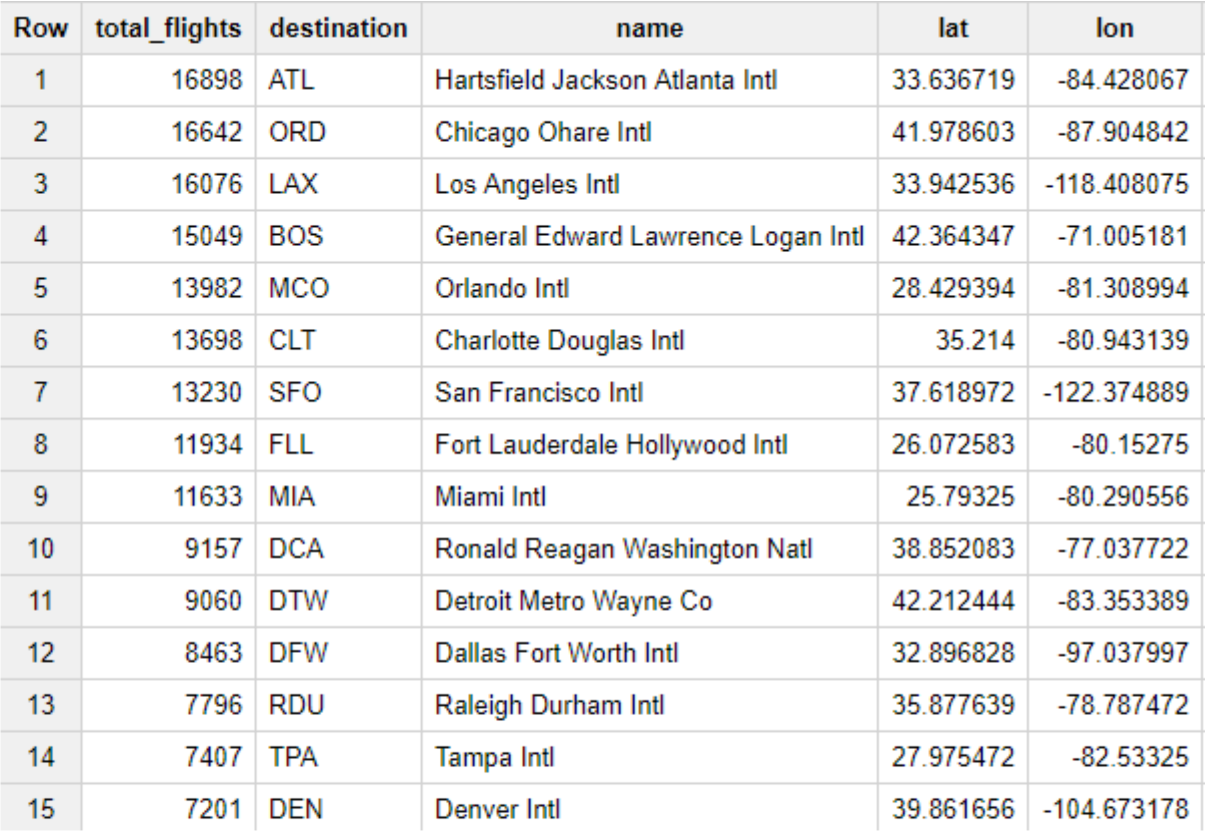
5.To reduce operations cost, what's the **airtime** distribution ?

Question 1:
Which are the most **popular flight routes** from New York?

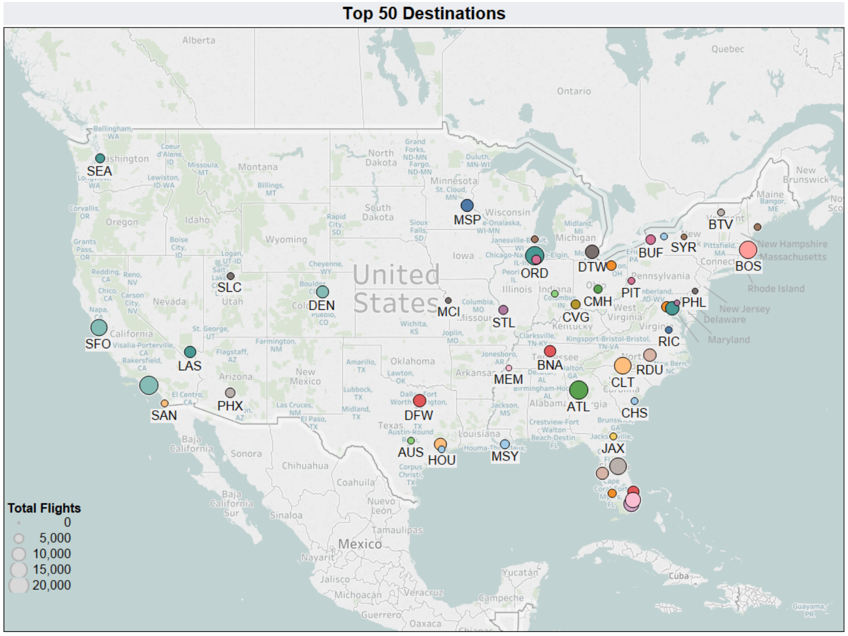
1.1 What are the overall popular destinations ?

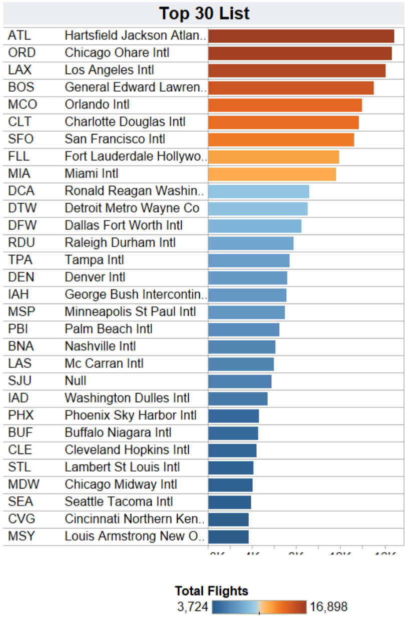
```
SELECT
        COUNT(F.dest) AS total_flights,
        F.dest AS destination,
        A.name,
        A.lat,
        A.lon
FROM `bax421final.data.flights` AS F
LEFT JOIN `bax421final.data.airports`AS A
ON F.dest = A.faa
WHERE F.dep_time IS NOT NULL
GROUP BY 2,3,4,5
```

https://drive.google.com/open?id=1e4KlogmsvN_HUyu2qPgTBTh20I-690Ax



Row	total_flights	destination	name	lat	lon
1	16898	ATL	Hartsfield Jackson Atlanta Intl	33.636719	-84.428067
2	16642	ORD	Chicago Ohare Intl	41.978603	-87.904842
3	16076	LAX	Los Angeles Intl	33.942536	-118.408075
4	15049	BOS	General Edward Lawrence Logan Intl	42.364347	-71.005181
5	13982	MCO	Orlando Intl	28.429394	-81.308994
6	13698	CLT	Charlotte Douglas Intl	35.214	-80.943139
7	13230	SFO	San Francisco Intl	37.618972	-122.374889
8	11934	FLL	Fort Lauderdale Hollywood Intl	26.072583	-80.15275
9	11633	MIA	Miami Intl	25.79325	-80.290556
10	9157	DCA	Ronald Reagan Washington Natl	38.852083	-77.037722
11	9060	DTW	Detroit Metro Wayne Co	42.212444	-83.353389
12	8463	DFW	Dallas Fort Worth Intl	32.896828	-97.037997
13	7796	RDU	Raleigh Durham Intl	35.877639	-78.787472
14	7407	TPA	Tampa Intl	27.975472	-82.53325
15	7201	DEN	Denver Intl	39.861656	-104.673178

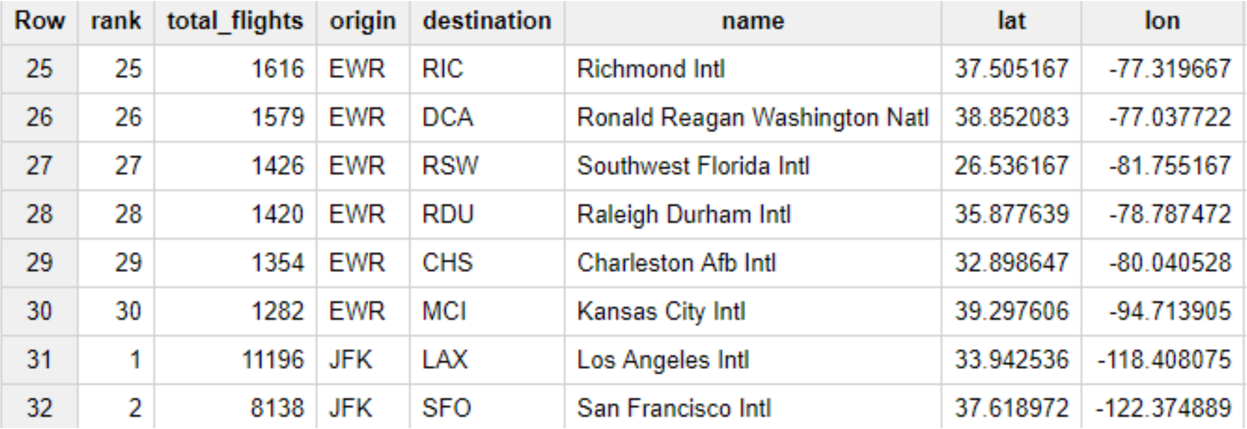




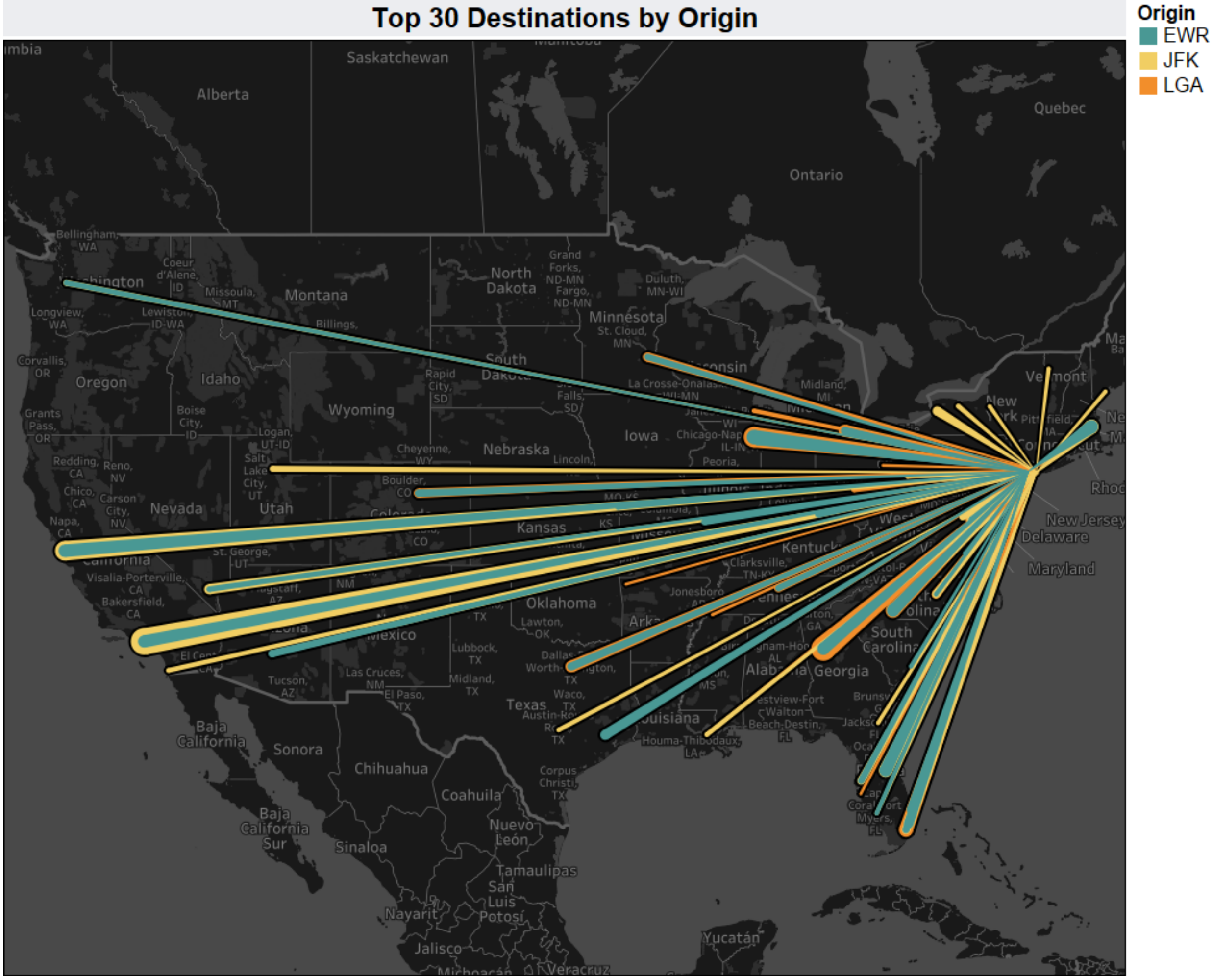
1.2 Ranking of TOP 30 destination by origin airport (EWR, JFK, LGA)

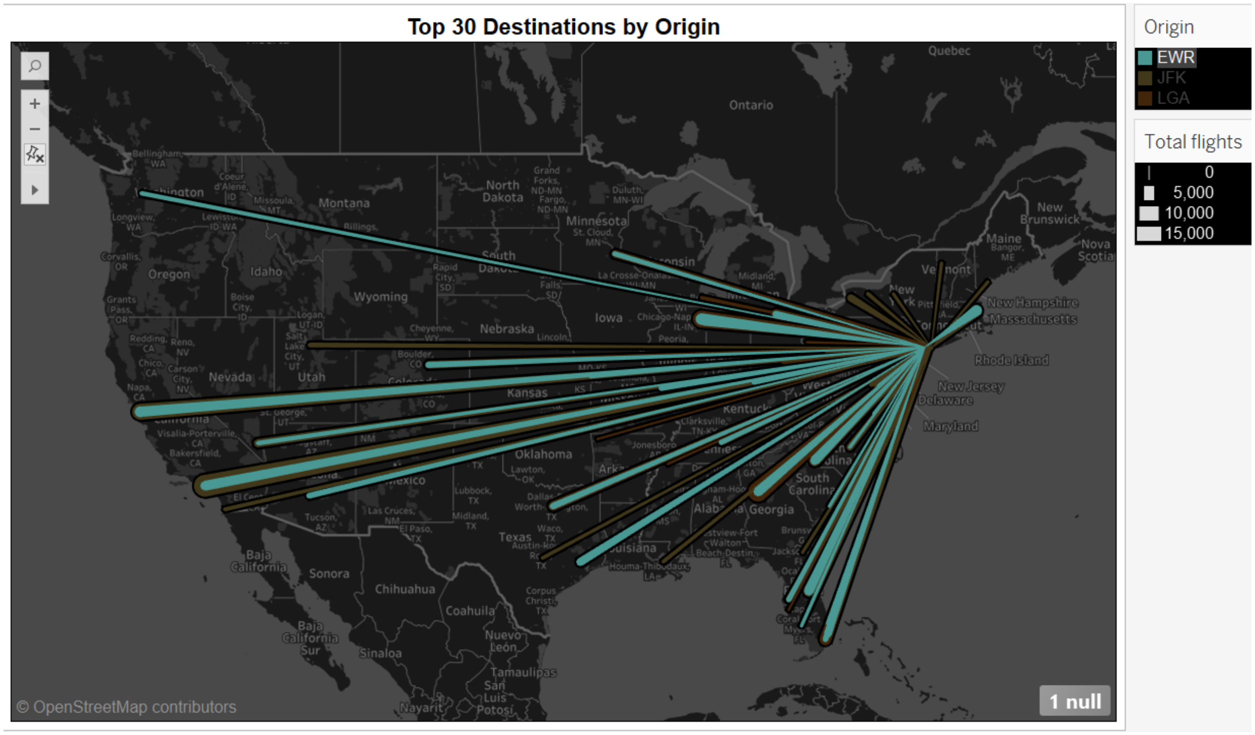
```
SELECT *
FROM
(SELECT RANK() OVER (PARTITION BY F.origin ORDER BY SUM(F.flight_number) DESC) AS
rank,
               SUM(F.flight_number) AS total_flights,
                F.origin,
                F.destination,
                A.name,
                A.lat,
                A.lon
  FROM (SELECT EXTRACT(DATE FROM time_hour) AS date,
                             origin,
                             dest AS destination,
                            COUNT(dest) AS flight_number
               FROM `bax421final.data.flights`
               WHERE dep_time IS NOT NULL
               GROUP BY origin, dest, date
               ORDER BY date, origin) AS F
LEFT JOIN `bax421final.data.airports`AS A
ON F.destination = A.faa
GROUP BY 3,4,5,6,7)
WHERE rank BETWEEN 1 AND 30
```

https://drive.google.com/open?id=1ZNRyeSyikMSPhnfI1AbjJJy3ikyTMh2B

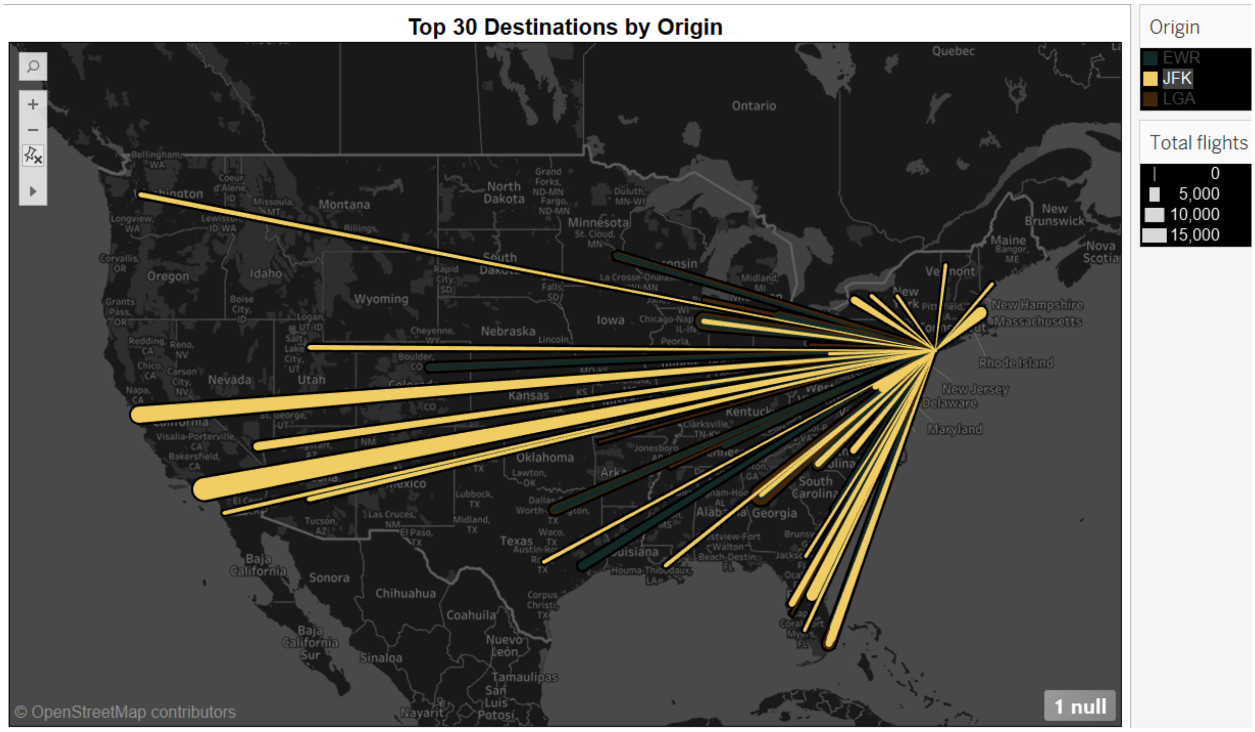


Row	rank	total_flights	origin	destination	name	lat	lon
25	25	1616	EWR	RIC	Richmond Intl	37.505167	-77.319667
26	26	1579	EWR	DCA	Ronald Reagan Washington Natl	38.852083	-77.037722
27	27	1426	EWR	RSW	Southwest Florida Intl	26.536167	-81.755167
28	28	1420	EWR	RDU	Raleigh Durham Intl	35.877639	-78.787472
29	29	1354	EWR	CHS	Charleston Afb Intl	32.898647	-80.040528
30	30	1282	EWR	MCI	Kansas City Intl	39.297606	-94.713905
31	1	11196	JFK	LAX	Los Angeles Intl	33.942536	-118.408075
32	2	8138	JFK	SFO	San Francisco Intl	37.618972	-122.374889

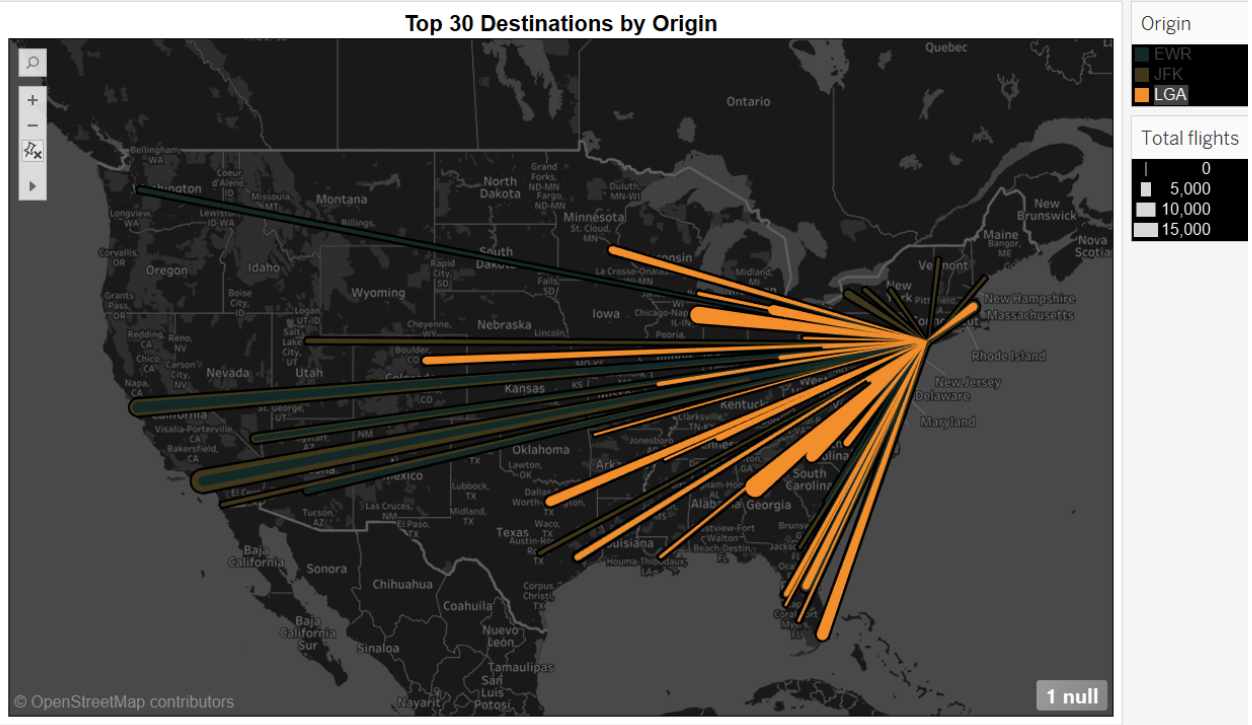

Top 30 Destinations by Origin
Origin
EWR
JFK
LGA
Alberta
Saskatchewan
Ontario
Quebec
Montana
North Dakota
South Dakota
Minnesota
Wisconsin
Oregon
Idaho
Wyoming
Nebraska
Iowa
Nevada
Utah
Colorado
Kansas
California
Oklahoma
Texas
Arkansas
Kentucky
Alabama
Georgia
South Carolina
Louisiana
Vermont
New York
New Jersey
Delaware
Maryland
Baja California
Sonora
Chihuahua
Coahuila
Nuevo León
Sinaloa
Tamaulipas
San Luis Potosí
Nayarit
Jalisco
Yucatán
Veracruz



Top 30 Destinations by Origin



Top 30 Destinations by Origin



Question 2:

When is the peak/off **season/hour**? Top month/quarter with highest/lowest traffic (#flight, *seat or count by seat)

Traffic by month/quarter/hour

2.1: Number of flights by month:

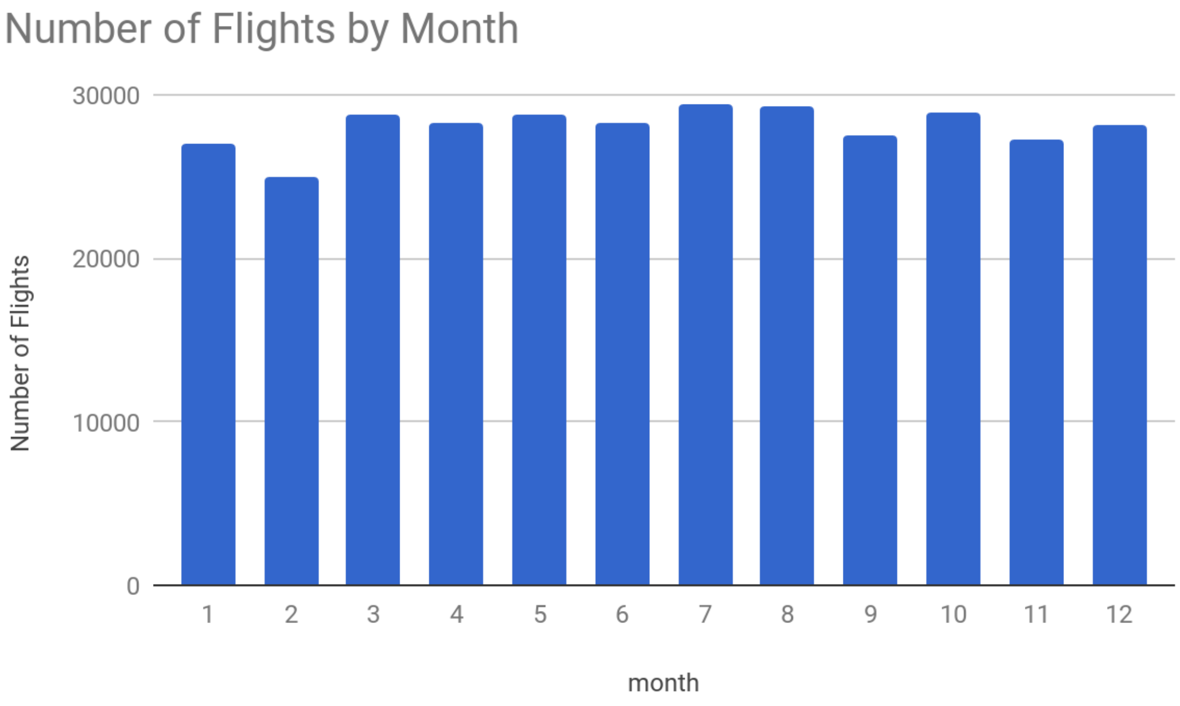
```
SELECT month, count(sched_dep_time) as num_of_fl
FROM `bax421final.data.flights`
group by month
order by num_of_fl desc
LIMIT 100
```

2.2: Number of plane seats by month

```
SELECT fl.month , sum(pl.seats) as max_total_pass
FROM `bax421final.data.flights` as fl
join `bax421final.data.plans` as pl on (fl.tailnum = pl.tailnum)
group by month
order by max_total_pass desc
LIMIT 100
```

Flights:

month	num_of_fl
7	29425
8	29327
10	28889
3	28834
5	28796
4	28330
6	28243
12	28135
9	27574
11	27268
1	27004
2	24951



--Slight differences, February is lower partially because there are only 28 days in the month.

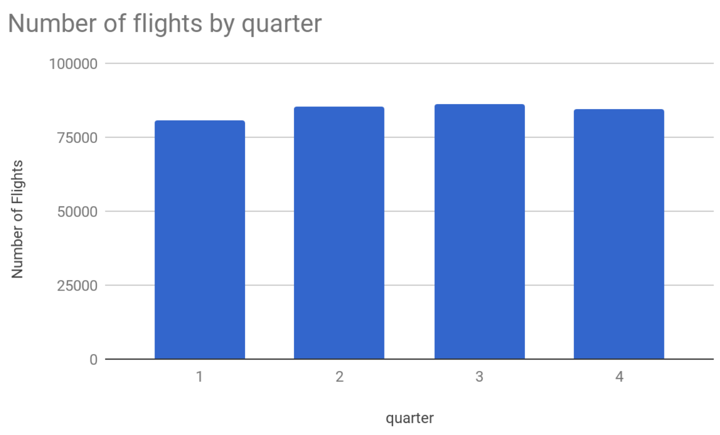
Seats:

month	max_total_pass
8	3407554
7	3380565
10	3378417
3	3299678
12	3298589
5	3284725
4	3275186
6	3253530
11	3216635
9	3179846
1	3075040
2	2801552

2.3 Number of flights by quarter:

```
with tbl_with_quart as(
SELECT CASE
            when month in (1,2,3) then 1
            when month in (4,5,6) then 2
            when month in (7,8,9) then 3
            when month in (10,11,12) then 4
               ELSE 0
       END as quarter, *
from `bax421final.data.flights`
)
SELECT quarter, count(sched_dep_time) as num_of_fl
FROM tbl_with_quart
group by quarter
order by num_of_fl desc
LIMIT 100
```

quarter	num_of_fl
3	86326
2	85369
4	84292
1	80789



2.4 Number of seats by quarter:

```
with tbl_with_quart as(
SELECT CASE
        when month in (1,2,3) then 1
        when month in (4,5,6) then 2
        when month in (7,8,9) then 3
        when month in (10,11,12) then 4
          ELSE 0
    END as quarter, *
from `bax421final.data.flights`
)
SELECT quarter , sum(pl.seats) as max_total_pass
FROM tbl_with_quart
join `bax421final.data.plans` as pl on (tbl_with_quart.tailnum = pl.tailnum)
group by quarter
order by max_total_pass desc
LIMIT 100
```

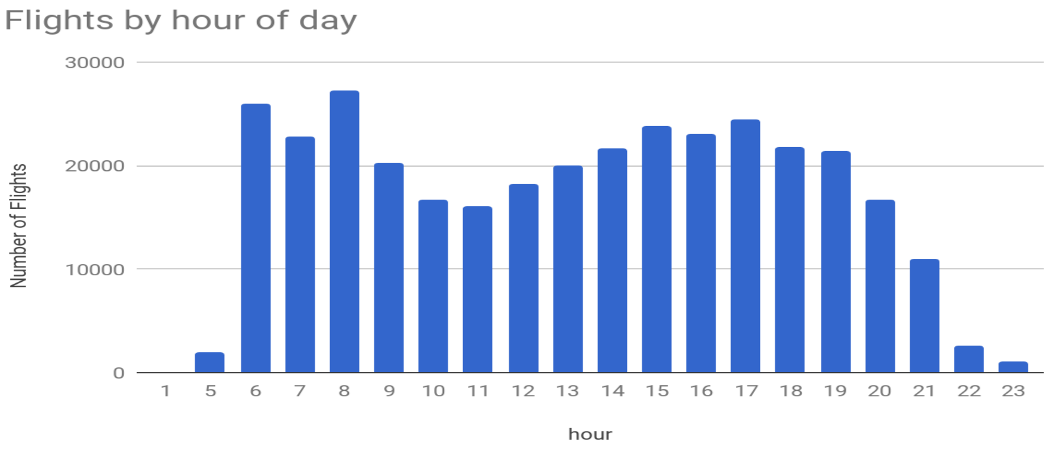
quarter	max_total_pass
3	9967965
4	9893641
2	9813441
1	9176270

2.5 Peak hours

```
SELECT hour, count(sched_dep_time) as num_of_fl
FROM `bax421final.data.flights`
group by hour
order by num_of_fl desc
LIMIT 100
```

hour	num_of_fl
8	27242
6	25951
17	24426
15	23888
16	23002
7	22821
18	21783
14	21706
19	21441
9	20312
13	19956
12	18181
20	16739
10	16708
11	16033
21	10933
22	2639
5	1953
23	1061
1	1

--NOTE: 4 hours have no data whatsoever: 0,2,3,4 . Hour “1” only has 1 flight.

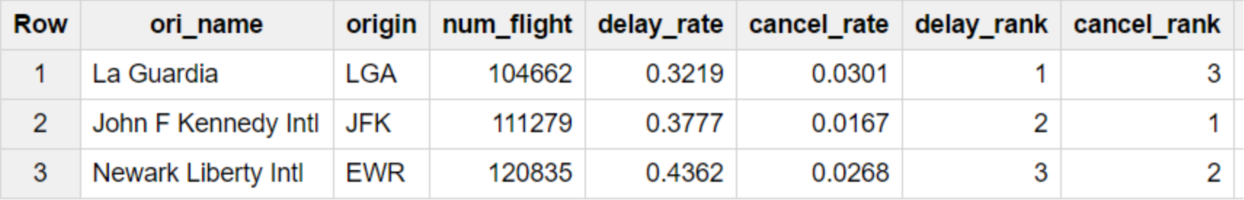


Question 3:
In order to attract more business travelers (who will pay higher price than leisure travelers), our company should reduce **delay and** minimize **cancellation**. (if do not have actual departure time, assume it is cancelled) **--Yiqing**
Average delay time/rate, cancellation rate. (by carrier, by city, by month)

https://public.tableau.com/profile/yiqing.yang5803#!/vizhome/Question3Viz/Dashboard1?publish=yes

3.1 by departure airport

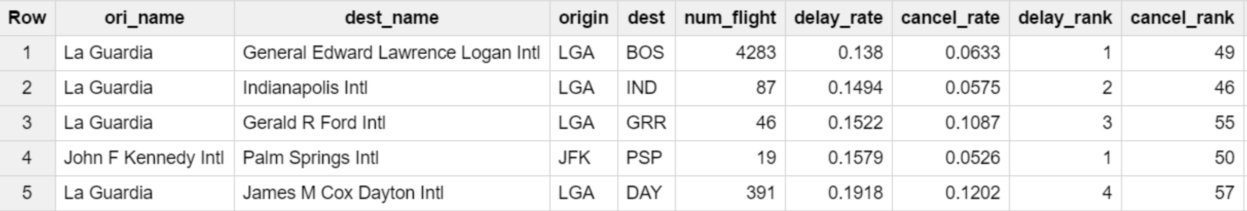
```
select ori.ori_name, t1.*,
rank() over (order by delay_rate) as delay_rank,
rank() over (order by cancel_rate) as cancel_rank
from
(select origin,
count(1) as num_flight,
round(sum(if(dep_delay>0,1,0))/count(1),4) as delay_rate,
round(sum(if(dep_delay is null,1,0))/count(1),4) as cancel_rate
from `bax421final.data.flights`
group by origin
) as t1
left join
(select faa, name as ori_name from `bax421final.data.airports`) as ori
on t1.origin = ori.faa
order by delay_rate
```



Row	ori_name	origin	num_flight	delay_rate	cancel_rate	delay_rank	cancel_rank
1	La Guardia	LGA	104662	0.3219	0.0301	1	3
2	John F Kennedy Intl	JFK	111279	0.3777	0.0167	2	1
3	Newark Liberty Intl	EWR	120835	0.4362	0.0268	3	2

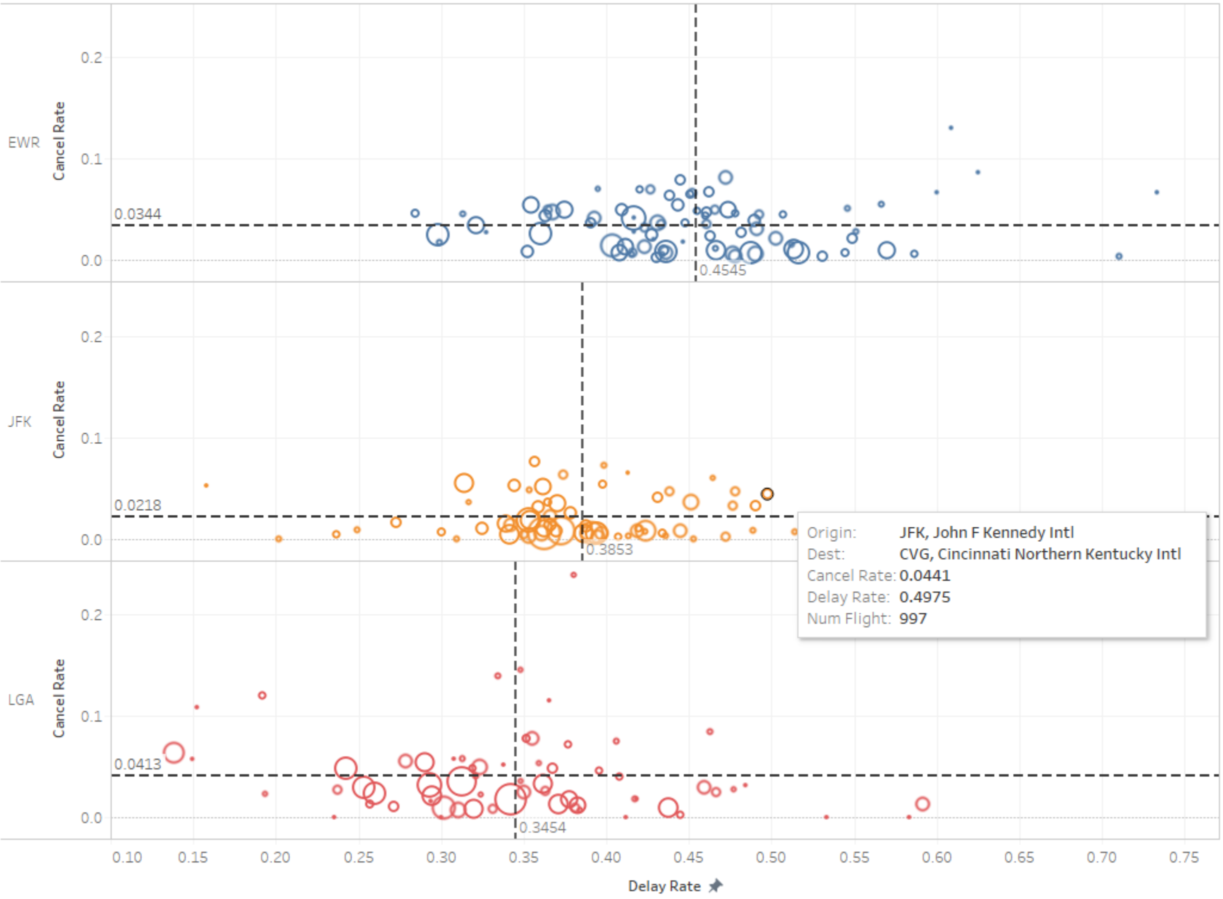
3.2 by departure and destination airport
rank in departure airport, only consider lines have more than 10 flights

```
select ori.ori_name,dest.dest_name, t1.*,
rank() over (partition by origin order by delay_rate) as delay_rank,
rank() over (partition by origin order by cancel_rate) as cancel_rank
from
(select origin,dest,
count(1) as num_flight,
round(sum(if(dep_delay>0,1,0))/count(1),4) as delay_rate,
round(sum(if(dep_delay is null,1,0))/count(1),4) as cancel_rate
from `bax421final.data.flights`
where dest is not null
group by 1,2
) as t1
left join
(select faa, name as ori_name from `bax421final.data.airports`) as ori
on t1.origin = ori.faa
left join
(select faa, name as dest_name from `bax421final.data.airports`) as dest
on t1.dest = dest.faa
where num_flight >= 10  --only consider the lines have more than 10 flights
order by delay_rate
```



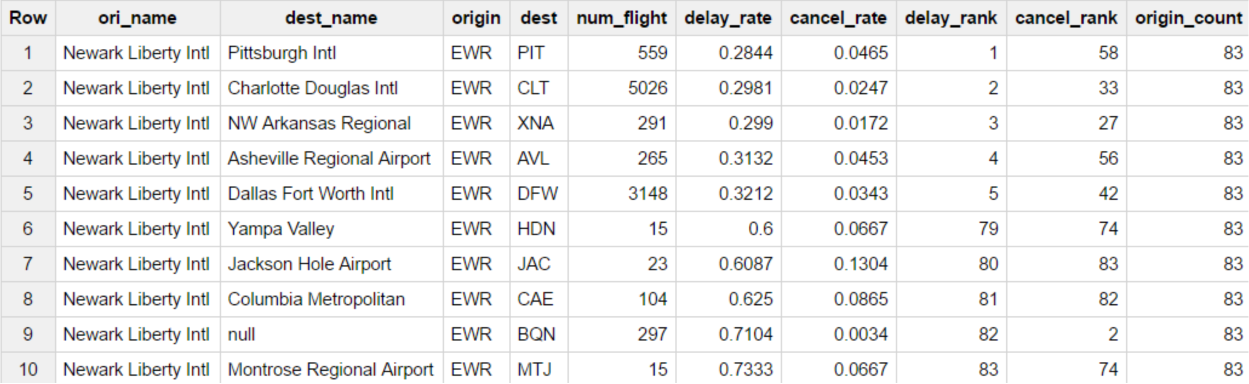
Row	ori_name	dest_name	origin	dest	num_flight	delay_rate	cancel_rate	delay_rank	cancel_rank
1	La Guardia	General Edward Lawrence Logan Intl	LGA	BOS	4283	0.138	0.0633	1	49
2	La Guardia	Indianapolis Intl	LGA	IND	87	0.1494	0.0575	2	46
3	La Guardia	Gerald R Ford Intl	LGA	GRR	46	0.1522	0.1087	3	55
4	John F Kennedy Intl	Palm Springs Intl	JFK	PSP	19	0.1579	0.0526	1	50
5	La Guardia	James M Cox Dayton Intl	LGA	DAY	391	0.1918	0.1202	4	57

Delay and Cencel Rate by Airlines

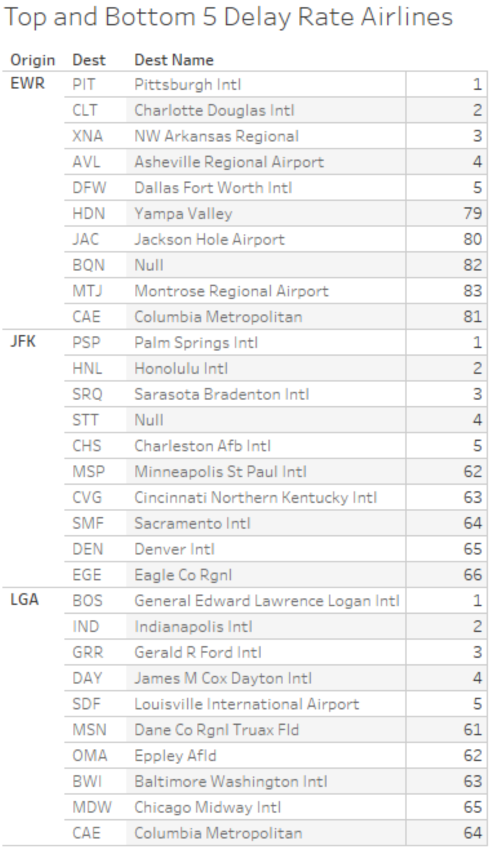


Top and bottom 5 cancel rate and delay rate from each origin airport

```
select t2.*
from
(select ori.ori_name,dest.dest_name, t1.*,
rank() over (partition by origin order by delay_rate) as delay_rank,
rank() over (partition by origin order by cancel_rate) as cancel_rank,
count(*) over (partition by origin) as origin_count
from
(select origin,dest,
count(1) as num_flight,
round(sum(if(dep_delay>0,1,0))/count(1),4) as delay_rate,
round(sum(if(dep_delay is null,1,0))/count(1),4) as cancel_rate
from `bax421final.data.flights`
where dest is not null
group by 1,2
) as t1
left join
(select faa, name as ori_name from `bax421final.data.airports`) as ori
on t1.origin = ori.faa
left join
(select faa, name as dest_name from `bax421final.data.airports`) as dest
on t1.dest = dest.faa
where num_flight >= 10  --only consider the lines have more than 10 flights
) t2
where delay_rank <=5 or delay_rank > origin_count-5
order by origin,delay_rate
```



Row	ori_name	dest_name	origin	dest	num_flight	delay_rate	cancel_rate	delay_rank	cancel_rank	origin_count
1	Newark Liberty Intl	Pittsburgh Intl	EWR	PIT	559	0.2844	0.0465	1	58	83
2	Newark Liberty Intl	Charlotte Douglas Intl	EWR	CLT	5026	0.2981	0.0247	2	33	83
3	Newark Liberty Intl	NW Arkansas Regional	EWR	XNA	291	0.299	0.0172	3	27	83
4	Newark Liberty Intl	Asheville Regional Airport	EWR	AVL	265	0.3132	0.0453	4	56	83
5	Newark Liberty Intl	Dallas Fort Worth Intl	EWR	DFW	3148	0.3212	0.0343	5	42	83
6	Newark Liberty Intl	Yampa Valley	EWR	HDN	15	0.6	0.0667	79	74	83
7	Newark Liberty Intl	Jackson Hole Airport	EWR	JAC	23	0.6087	0.1304	80	83	83
8	Newark Liberty Intl	Columbia Metropolitan	EWR	CAE	104	0.625	0.0865	81	82	83
9	Newark Liberty Intl	null	EWR	BQN	297	0.7104	0.0034	82	2	83
10	Newark Liberty Intl	Montrose Regional Airport	EWR	MTJ	15	0.7333	0.0667	83	74	83

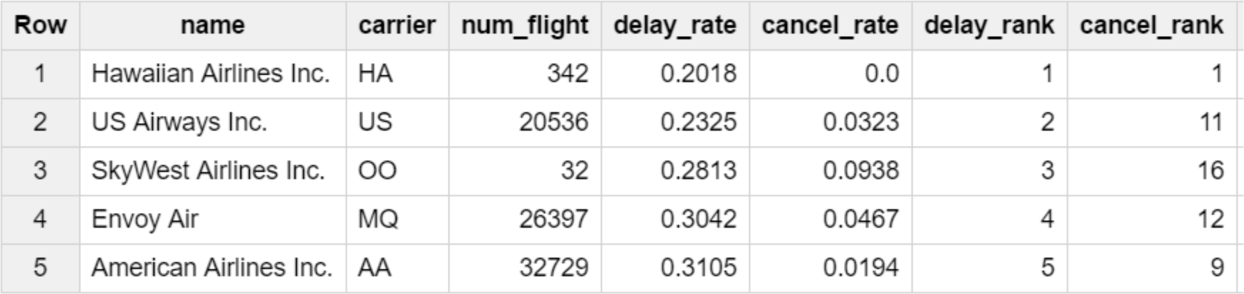


Top and Bottom 5 Delay Rate Airlines

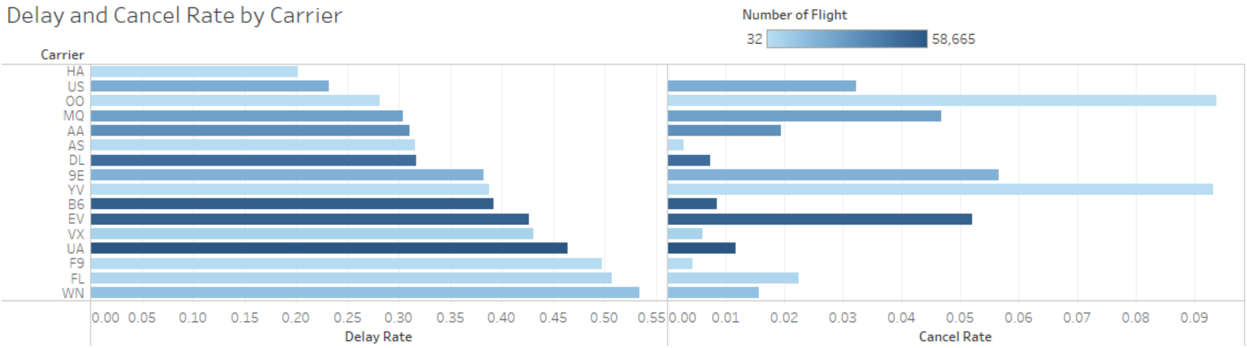
Origin	Dest	Dest Name	
EWR	PIT	Pittsburgh Intl	1
	CLT	Charlotte Douglas Intl	2
	XNA	NW Arkansas Regional	3
	AVL	Asheville Regional Airport	4
	DFW	Dallas Fort Worth Intl	5
	HDN	Yampa Valley	79
	JAC	Jackson Hole Airport	80
	BQN	Null	82
	MTJ	Montrose Regional Airport	83
	CAE	Columbia Metropolitan	81
JFK	PSP	Palm Springs Intl	1
	HNL	Honolulu Intl	2
	SRQ	Sarasota Bradenton Intl	3
	STT	Null	4
	CHS	Charleston Afb Intl	5
	MSP	Minneapolis St Paul Intl	62
	CVG	Cincinnati Northern Kentucky Intl	63
	SMF	Sacramento Intl	64
	DEN	Denver Intl	65
	EGE	Eagle Co Rgnl	66
LGA	BOS	General Edward Lawrence Logan Intl	1
	IND	Indianapolis Intl	2
	GRR	Gerald R Ford Intl	3
	DAY	James M Cox Dayton Intl	4
	SDF	Louisville International Airport	5
	MSN	Dane Co Rgnl Truax Fld	61
	OMA	Eppley Afld	62
	BWI	Baltimore Washington Intl	63
	MDW	Chicago Midway Intl	65
	CAE	Columbia Metropolitan	64

3.3 by carrier

```
select t2.name, t1.*,
rank() over (order by delay_rate) as delay_rank,
rank() over (order by cancel_rate) as cancel_rank
from
(select carrier,
count(1) as num_flight,
round(sum(if(dep_delay>0,1,0))/count(1),4) as delay_rate,
round(sum(if(dep_delay is null,1,0))/count(1),4) as cancel_rate
from `bax421final.data.flights`
group by 1
) as t1
left join
`bax421final.data.airlines`  as t2
on t1.carrier = t2.carrier
order by delay_rate
```

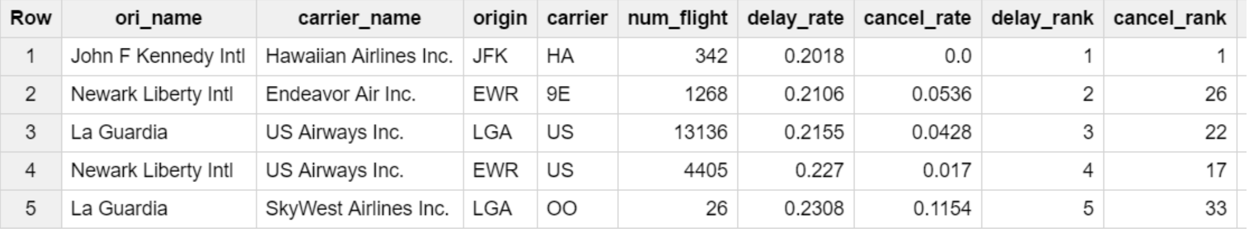


Row	name	carrier	num_flight	delay_rate	cancel_rate	delay_rank	cancel_rank
1	Hawaiian Airlines Inc.	HA	342	0.2018	0.0	1	1
2	US Airways Inc.	US	20536	0.2325	0.0323	2	11
3	SkyWest Airlines Inc.	OO	32	0.2813	0.0938	3	16
4	Envoy Air	MQ	26397	0.3042	0.0467	4	12
5	American Airlines Inc.	AA	32729	0.3105	0.0194	5	9

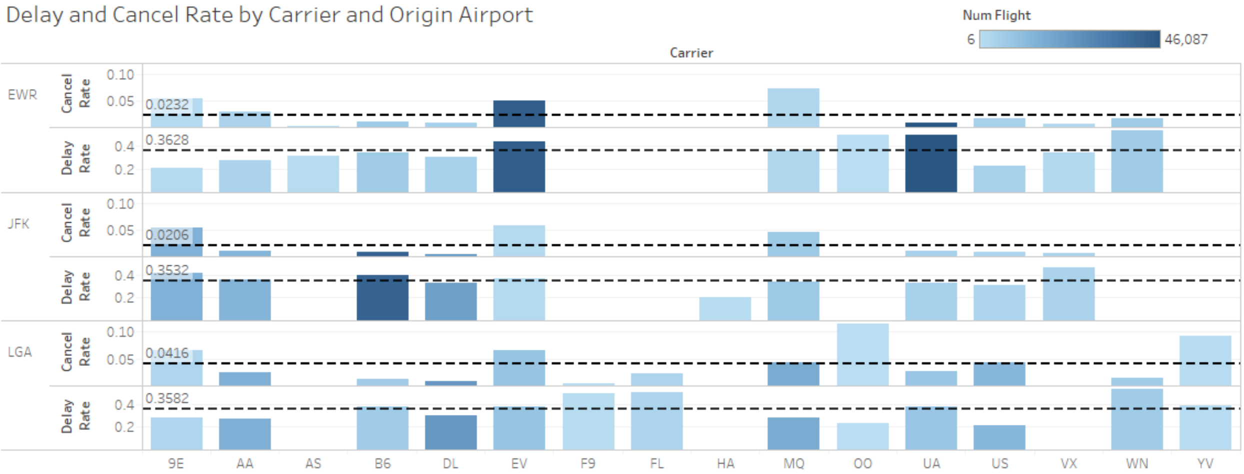


3.4 by departure airport and carrier

```
select t3.ori_name,t2.name as carrier_name,
t1.*,
rank() over (partition by origin order by delay_rate) as delay_rank,
rank() over (partition by origin order by cancel_rate) as cancel_rank
from
(select origin,carrier,
count(1) as num_flight,
round(sum(if(dep_delay>0,1,0))/count(1),4) as delay_rate,
round(sum(if(dep_delay is null,1,0))/count(1),4) as cancel_rate
from `bax421final.data.flights`
group by 1,2
) as t1
left join
`bax421final.data.airlines`  as t2
on t1.carrier = t2.carrier
left join
(select faa, name as ori_name from `bax421final.data.airports`) as t3
on t1.origin = t3.faa
order by delay_rate
```

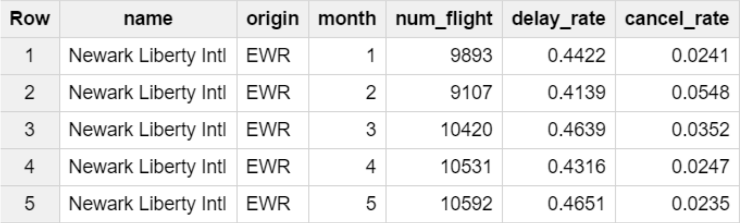


Row	ori_name	carrier_name	origin	carrier	num_flight	delay_rate	cancel_rate	delay_rank	cancel_rank
1	John F Kennedy Intl	Hawaiian Airlines Inc.	JFK	HA	342	0.2018	0.0	1	1
2	Newark Liberty Intl	Endeavor Air Inc.	EWR	9E	1268	0.2106	0.0536	2	26
3	La Guardia	US Airways Inc.	LGA	US	13136	0.2155	0.0428	3	22
4	Newark Liberty Intl	US Airways Inc.	EWR	US	4405	0.227	0.017	4	17
5	La Guardia	SkyWest Airlines Inc.	LGA	OO	26	0.2308	0.1154	5	33



3.5 by origin and month

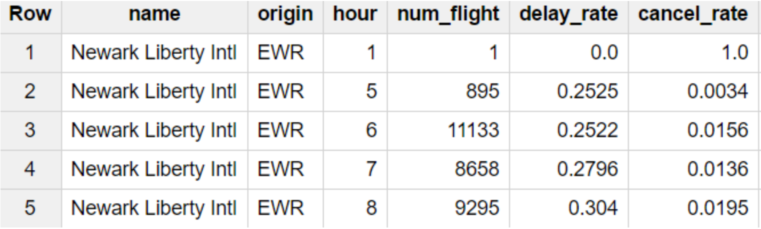
```
select t2.name, t1.* from
(select origin,month,
count(1) as num_flight,
round(sum(if(dep_delay>0,1,0))/count(1),4) as delay_rate,
round(sum(if(dep_delay is null,1,0))/count(1),4) as cancel_rate
from `bax421final.data.flights`
group by 1,2
) as t1
left join
`bax421final.data.airports`  as t2
on t1.origin = t2.faa
order by origin, month
```



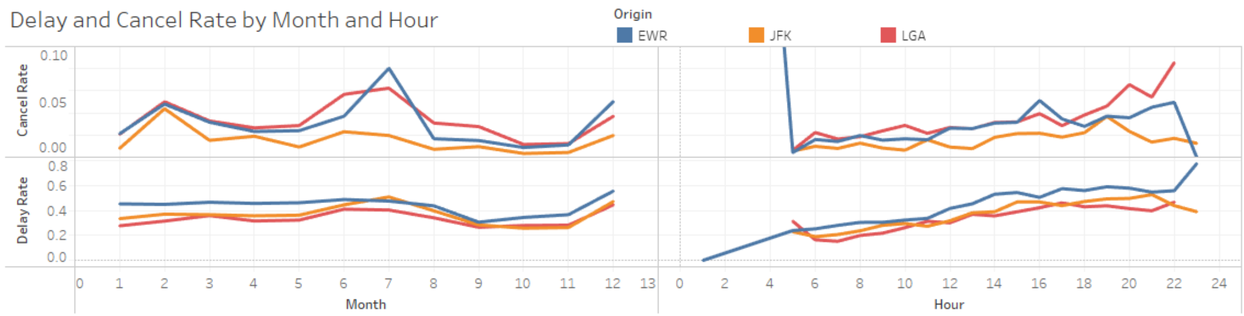
Row	name	origin	month	num_flight	delay_rate	cancel_rate
1	Newark Liberty Intl	EWR	1	9893	0.4422	0.0241
2	Newark Liberty Intl	EWR	2	9107	0.4139	0.0548
3	Newark Liberty Intl	EWR	3	10420	0.4639	0.0352
4	Newark Liberty Intl	EWR	4	10531	0.4316	0.0247
5	Newark Liberty Intl	EWR	5	10592	0.4651	0.0235

3.6 by origin and hour

```
select t2.name, t1.* from
(select origin,extract(hour from time_hour) as hour,
count(1) as num_flight,
round(sum(if(dep_delay>0,1,0))/count(1),4) as delay_rate,
round(sum(if(dep_delay is null,1,0))/count(1),4) as cancel_rate
from `bax421final.data.flights`
group by 1,2
) as t1
left join
`bax421final.data.airports`  as t2
on t1.origin = t2.faa
order by origin,hour
```



Row	name	origin	hour	num_flight	delay_rate	cancel_rate
1	Newark Liberty Intl	EWR	1	1	0.0	1.0
2	Newark Liberty Intl	EWR	5	895	0.2525	0.0034
3	Newark Liberty Intl	EWR	6	11133	0.2522	0.0156
4	Newark Liberty Intl	EWR	7	8658	0.2796	0.0136
5	Newark Liberty Intl	EWR	8	9295	0.304	0.0195



Question 4:

What's the relationship between **weather and delay/cancel,** (help to predict the delay time)

Define delay by category as (<10, 10-30, 30-1, 1-3, >3, etc) - **Yifu**

Define precipitation/wind/visibility as (no, small, medium, large)

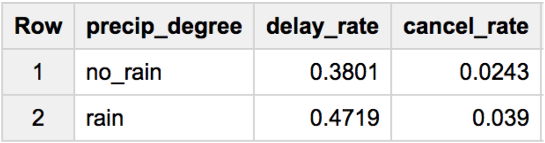
Count/percentage each type of delay by precipitation/wind/visibility type

Relationship between precip and delay:

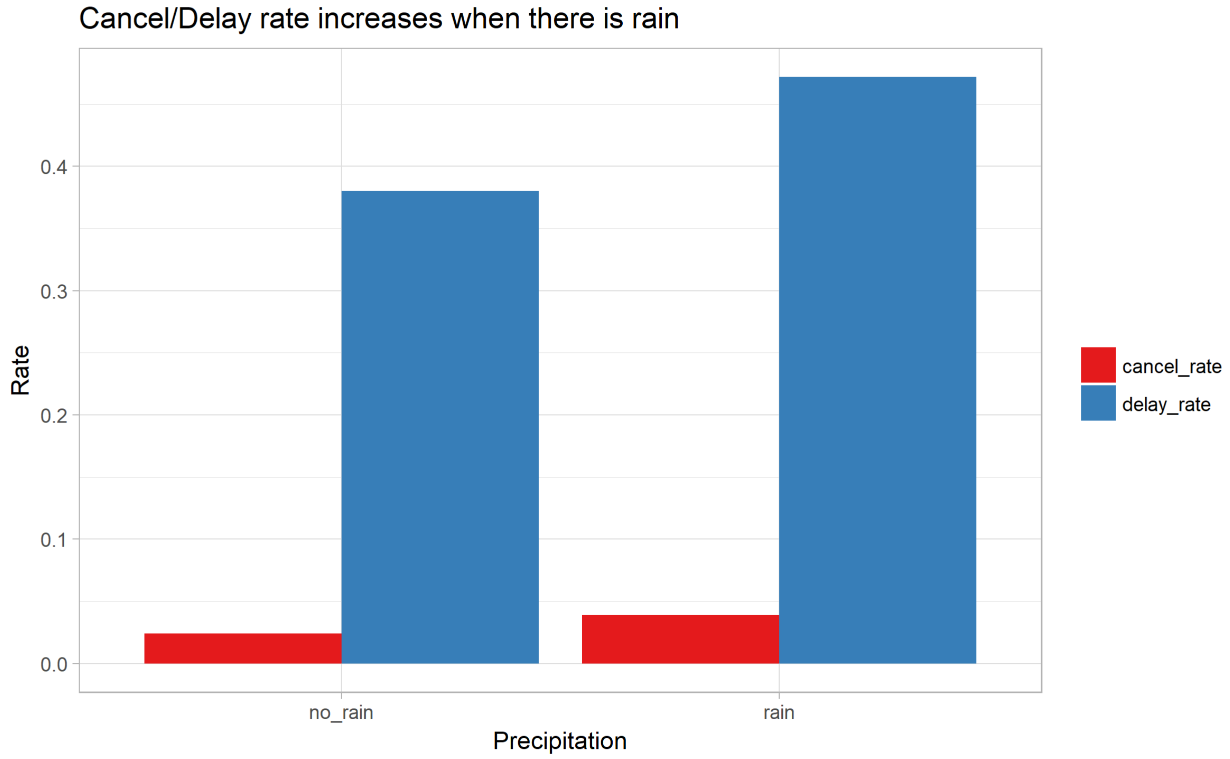
```
with tb_all as (select tb1.origin, tb1.time_hour, dep_time, sched_dep_time, dep_delay, precip,
                    case when precip < 0.1 then "no_rain"
                    else "rain" end as precip_degree
from `bax421final.data.flights` tb1
left join `bax421final.data.weathers` tb2
on tb1.time_hour = tb2.time_hour and tb1.origin = tb2.origin)

select precip_degree,
     round(sum(if(dep_delay>0,1,0))/count(*),4) as delay_rate,
     round(sum(if(dep_delay is null,1,0))/count(*),4) as cancel_rate
from tb_all
```

```
group by precip_degree;
```



Row	precip_degree	delay_rate	cancel_rate
1	no_rain	0.3801	0.0243
2	rain	0.4719	0.039

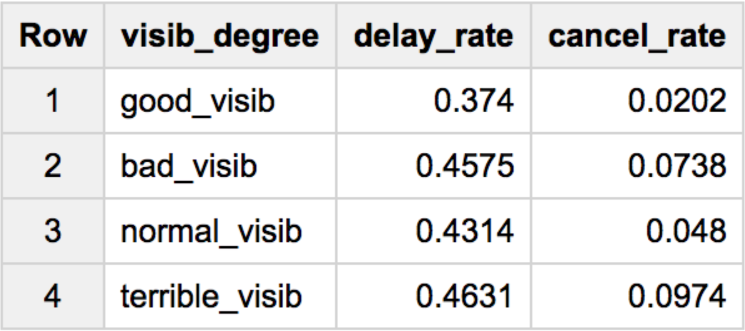


Cancel/Delay rate increases when there is rain

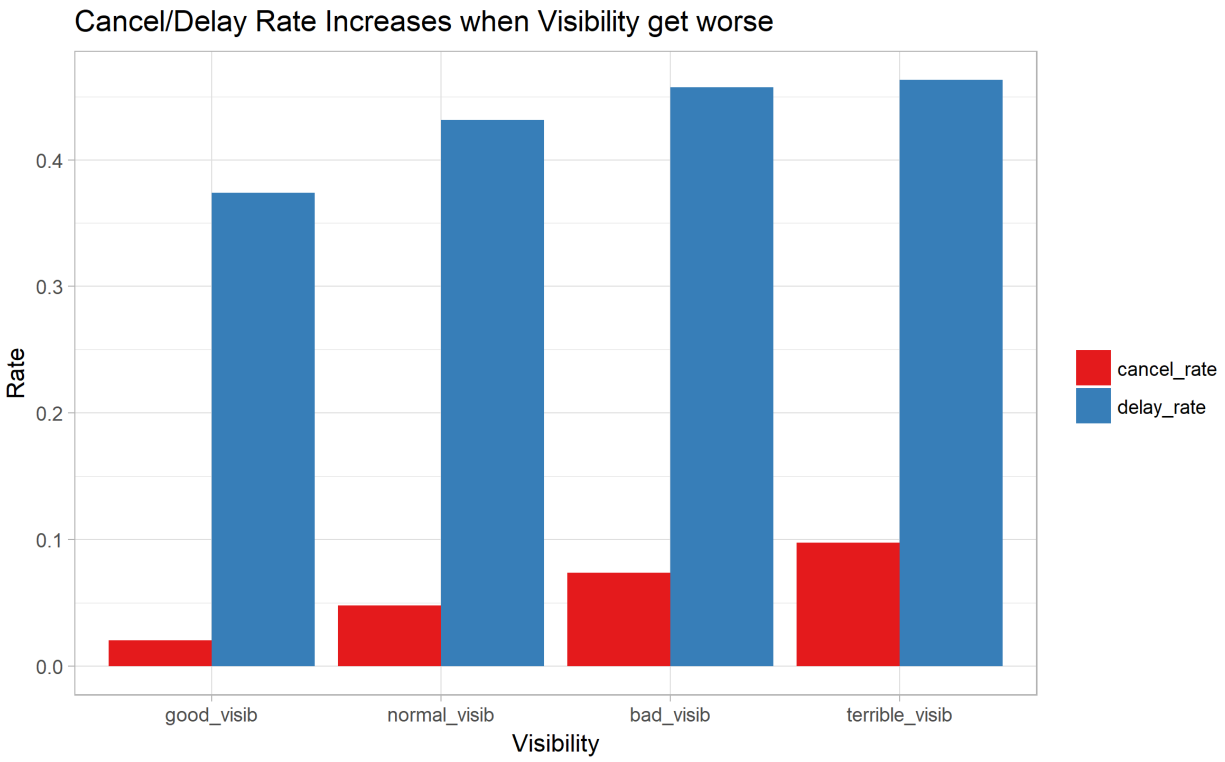
Relationship between visibility and delay:

```
with tb_all as (select tb1.origin, tb1.time_hour, dep_time, sched_dep_time, dep_delay, visib,
                  case when visib < 1 then "terrible_visib"
                      when visib >= 1 and visib < 3 then "bad_visib"
                      when visib >= 3 and visib < 8 then "normal_visib"
                  else "good_visib" end as visib_degree
from `bax421final.data.flights` tb1
left join `bax421final.data.weathers` tb2
on tb1.time_hour = tb2.time_hour and tb1.origin = tb2.origin)

select visib_degree,
      round(sum(if(dep_delay>0,1,0))/count(*),4) as delay_rate,
      round(sum(if(dep_delay is null,1,0))/count(*),4) as cancel_rate
from tb_all
group by visib_degree;
```



Row	visib_degree	delay_rate	cancel_rate
1	good_visib	0.374	0.0202
2	bad_visib	0.4575	0.0738
3	normal_visib	0.4314	0.048
4	terrible_visib	0.4631	0.0974



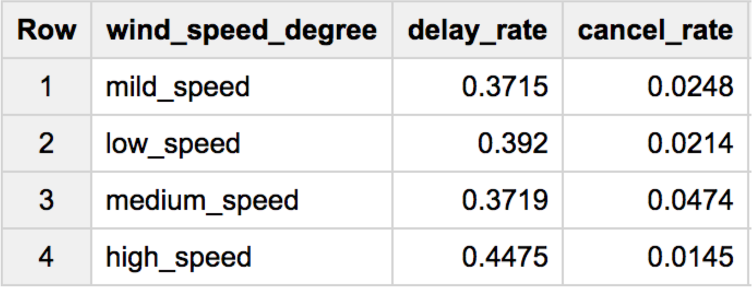
Cancel/Delay Rate Increases when Visibility get worse

Relationship between wind_speed and delay:

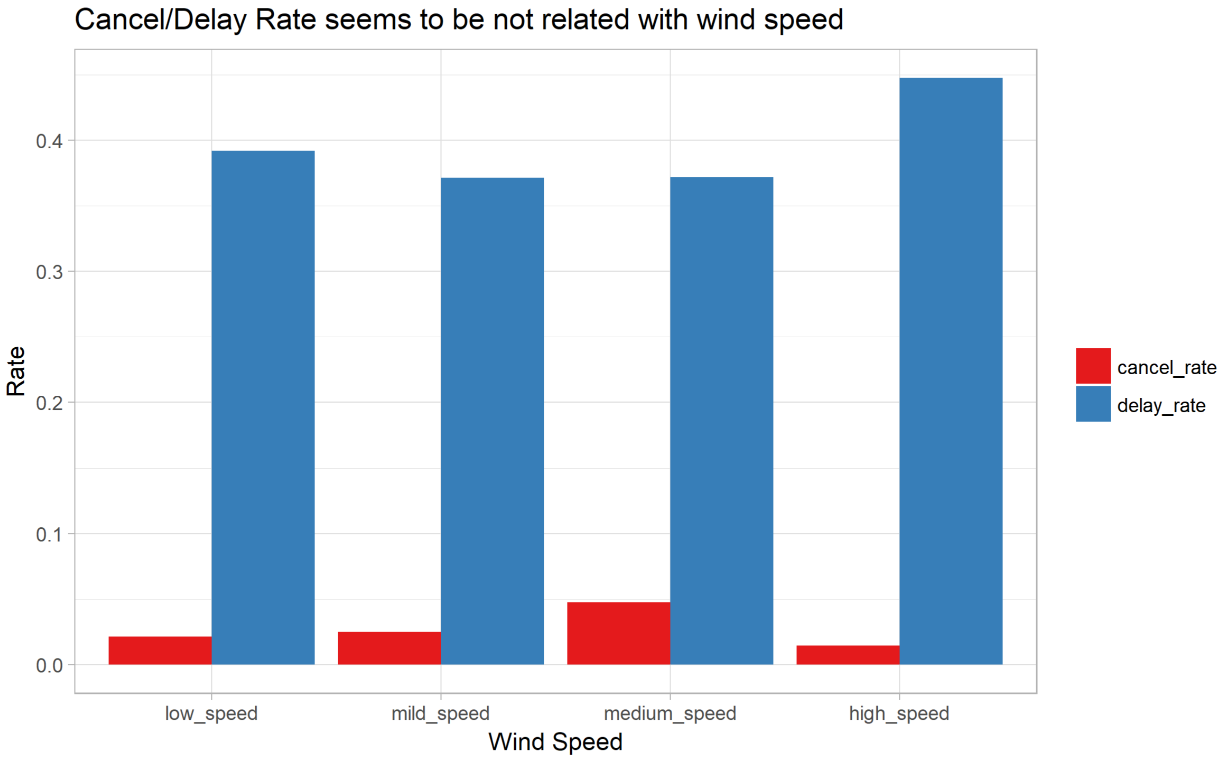
```
with tb_all as (select tb1.origin, tb1.time_hour, dep_time, sched_dep_time, dep_delay, wind_speed,
                  case when wind_speed < 10 then "low_speed"
                      when wind_speed >= 10 and wind_speed < 20 then "mild_speed"
                      when wind_speed >= 20 and wind_speed < 50 then "medium_speed"
                  else "high_speed" end as wind_speed_degree
from `bax421final.data.flights` tb1
left join `bax421final.data.weathers` tb2
on tb1.time_hour = tb2.time_hour and tb1.origin = tb2.origin)

select wind_speed_degree,
     round(sum(if(dep_delay>0,1,0))/count(*),4) as delay_rate,
     round(sum(if(dep_delay is null,1,0))/count(*),4) as cancel_rate
```

```
from tb_all
group by wind_speed_degree;
```



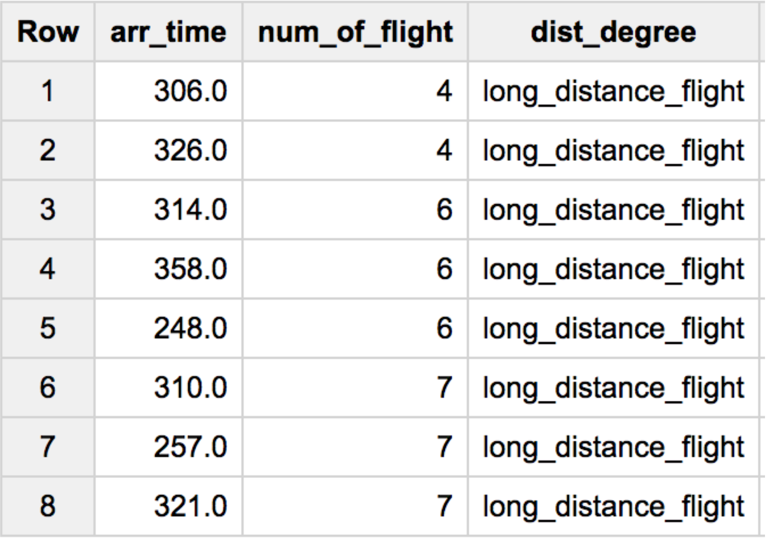
Row	wind_speed_degree	delay_rate	cancel_rate
1	mild_speed	0.3715	0.0248
2	low_speed	0.392	0.0214
3	medium_speed	0.3719	0.0474
4	high_speed	0.4475	0.0145



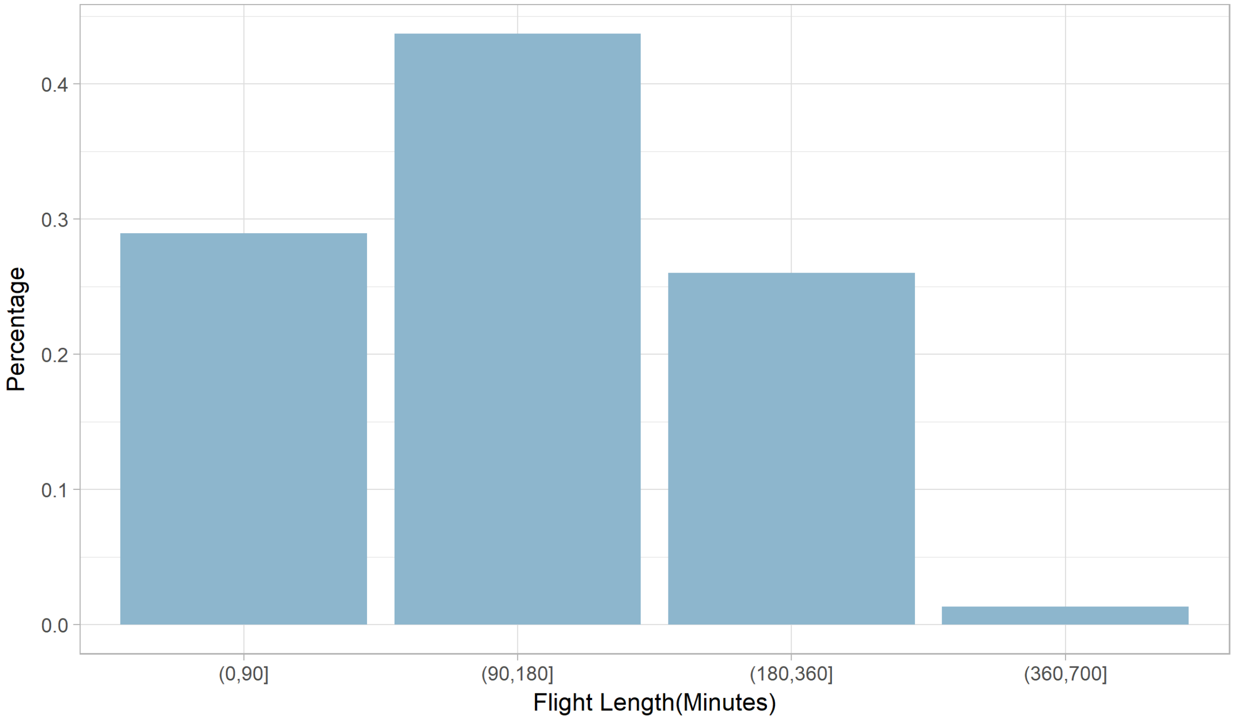
Question 5:

Airtime Distribution:

```
select arr_time, count(*) as num_of_flight,
      case when arr_time < 90 and arr_time is not null then "short_distance_flight"
          when arr_time >= 90 and arr_time < 180 and arr_time is not null then "medium_distance_flight"
          when arr_time >= 180 and arr_time < 360 and arr_time is not null then "long_distance_flight"
      else "extreme_long_distance" end as dist_degree
from `bax421final.data.flights`
group by arr_time, dist_degree
```

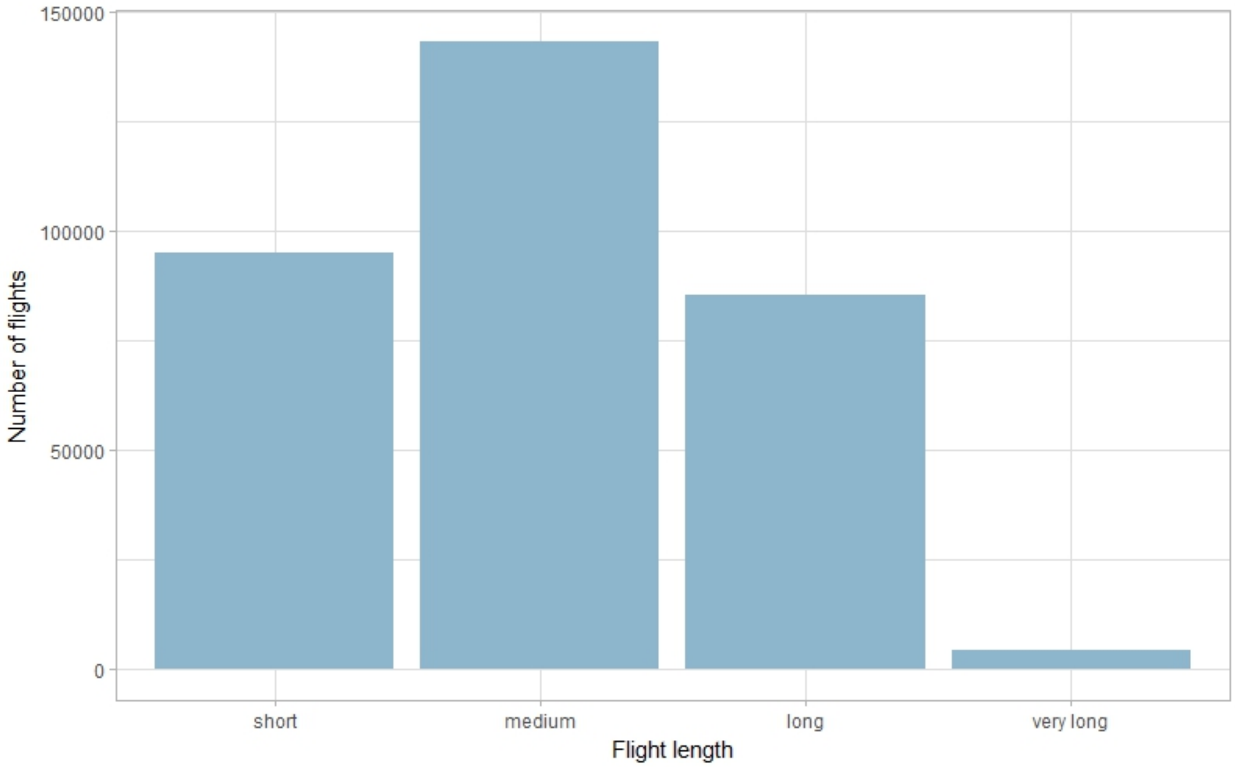


Row	arr_time	num_of_flight	dist_degree
1	306.0	4	long_distance_flight
2	326.0	4	long_distance_flight
3	314.0	6	long_distance_flight
4	358.0	6	long_distance_flight
5	248.0	6	long_distance_flight
6	310.0	7	long_distance_flight
7	257.0	7	long_distance_flight
8	321.0	7	long_distance_flight



```r
flights %>%
   filter(!is.na(air_time)) %>%
   mutate(flight_length = cut(air_time,breaks = c(0,90,180,360,700),labels =
c("short","medium","long","very long"))) %>%
   group_by(flight_length) %>%
   summarise(number_of_flights = n()) %>%
   ggplot(mapping = aes(x = flight_length, y = number_of_flights)) +
   geom_col(fill = "lightskyblue3") +
   labs(x = "Flight length", y = "Number of flights") +
   theme_light()

```


Number of flights
short
medium
long
very long
Flight length