--1. Which destination in the flights database is the furthest distance away?

--Answer: "Honolulu Intl";"HNL";4983

--Used the following query:

select b.name airport\_name, a.dest airport\_code, a.distance from

(SELECT dest, max(distance) as distance from flights group by dest) as a,

airports b

where a.dest = b.faa

order by 3 desc

fetch first 1 row only

--2. What are the different numbers of engines in the planes table? For each number of engines, which aircraft have the most number of seats?

--Different number of engines in the planes are: 1, 2, 3, 4

--Used the following query:

select distinct engines from planes

--Second question's part 2 "For each number of engines, which aircraft have the most number of seats"

--I assumed that the question is asking for which manufacturer has the Maximum number of seats for each number of engines. Obtained the following results:

**Manufacturer Seats Engines**

"DEHAVILLAND" 16 1

"BOEING" 400 2

"AIRBUS" 379 3

"BOEING" 450 4

--Used the following query to get the previous result.

select distinct b.manufacturer,

A.seats as seats,

A.engines as engines

from

(select engines, max(seats) seats

from planes group by engines) A,

(select engines,

tailnum, manufacturer, seats, model from planes) b

where a.engines = b.engines and a.seats = b.seats

order by engines;

--3. What weather conditions are associated with New York City departure delays?

--The most probable weather condition that could cause departure delays is related to Wind.

-- Depending on the wind conditions, the flights departure may be delayed.

-- This might be because of increase in the inflight duration of the arrival flights (due to wind), which might have caused departure flights delay

-- I used the following query to find the correlations between the weather and departure delays. To make it simple, I used the average delay time on all days, and found the coefficient of correlation between the average delay and the weather conditions. Then the same calculations were made for all the flights which were not delayed. The coefficient of correlations for the wind factors are negatively correlated between the on-time flights and delayed flights.

-- I used the following query to obtain this:

select 'delayed' rec\_type, corr(avg\_dep\_delay,avg\_temp) temp\_corr

,corr(avg\_dep\_delay,dewp) dewp\_corr

,corr(avg\_dep\_delay,humid) humid\_corr

,corr(avg\_dep\_delay,wind\_dir) wind\_dir\_corr

,corr(avg\_dep\_delay,wind\_speed) wind\_speed\_corr

,corr(avg\_dep\_delay,wind\_gust) wind\_gust\_corr

,corr(avg\_dep\_delay,precip) precip\_corr

,corr(avg\_dep\_delay,pressure) pressure\_corr

,corr(avg\_dep\_delay,visib) visib\_corr

from

(select avg(dep\_delay) avg\_dep\_delay,sum(dep\_delay),year,month

,day

from flights where dep\_Delay >= 0

group by day,month,year)

as a,

(select

avg(temp) avg\_temp,

avg(dewp) dewp,

avg(humid) humid,

avg(wind\_dir) wind\_dir,

avg(wind\_speed) wind\_speed,

avg(wind\_gust) wind\_gust,

avg(precip) precip,

avg(pressure) pressure,

avg(visib) visib,

year,

month,

day

from

weather b

group by year,

month,

day) b where a.year = b.year and

a.month = b.month and

a.day = b.day

union all

select 'on-time' record\_type, corr(avg\_dep\_delay,avg\_temp) temp\_corr

,corr(avg\_dep\_delay,dewp) dewp\_corr

,corr(avg\_dep\_delay,humid) humid\_corr

,corr(avg\_dep\_delay,wind\_dir) wind\_dir\_corr

,corr(avg\_dep\_delay,wind\_speed) wind\_speed\_corr

,corr(avg\_dep\_delay,wind\_gust) wind\_gust\_corr

,corr(avg\_dep\_delay,precip) precip\_corr

,corr(avg\_dep\_delay,pressure) pressure\_corr

,corr(avg\_dep\_delay,visib) visib\_corr

from

(select avg(dep\_delay) avg\_dep\_delay,sum(dep\_delay),year,month

,day

from flights where dep\_Delay <= 0

group by day,month,year)

as a,

(select

avg(temp) avg\_temp,

avg(dewp) dewp,

avg(humid) humid,

avg(wind\_dir) wind\_dir,

avg(wind\_speed) wind\_speed,

avg(wind\_gust) wind\_gust,

avg(precip) precip,

avg(pressure) pressure,

avg(visib) visib,

year,

month,

day

from

weather b

group by year,

month,

day) b where a.year = b.year and

a.month = b.month and

a.day = b.day

--Obtained the following results. I re-arranged the results for better presentation

|  |  |  |
| --- | --- | --- |
|  | Delayed | On-Time |
| Temp\_Corr | 0.2206 | 0.0361 |
| Dewp\_Corr | 0.3304 | 0.1323 |
| Humid\_Corr | 0.4283 | 0.2998 |
| Wind\_Dir\_Corr | -0.215 | 0.0269 |
| Wind\_Speed\_Corr | -0.046 | 0.0682 |
| Wind\_Gust\_Corr | -0.046 | 0.0682 |
| Precip\_Corr | 0.3158 | 0.2349 |
| Pressure\_Corr | -0.259 | -0.223 |
| Visib\_Corr | -0.365 | -0.227 |

We can infer that amongst all the weather conditions, the Wind direction varies significantly on the flight delays (since for on-time flights the wind-direction is positively correlated and for delayed flights it is negatively correlated). But note that the coefficient of correlations are not strong (linear relationship) between flight dep-delays and weather conditions (including the Wind conditions). However when compared to the same coeff of correlations with the on-time flights, only the wind has different signs of correlation. So I concluded that the WIND Direction might have an impact on the departure delay.

-- Note that most of the departure flight delays have occurred a day before the beginning of day lights savings (3/8/2013), in the summer months, during Christmas. Since we have not excluded such special holidays/events in our analysis, there is a potential chance that wind might not be a significant factor in the flight delays.

--4. Are older planes more likely to be delayed?

-- Not necessarily. There is NO strong proof that older flights are more likely to be delayed. Here is how I arrived at this conclusion.

--Used the following query to obtain the average filght delay grouped by the year of manufacture:

SELECT YEAR, AVG(DEP\_DELAY) AVG\_DEP\_DELAY, AVG(ARR\_DELAY) AVG\_ARR\_DELAY, COUNT(\*) AS FLIGHT\_FREQ FROM

(SELECT YEAR,TAILNUM FROM PUBLIC.PLANES) A, (SELECT DEP\_DELAY, ARR\_DELAY, TAILNUM FROM PUBLIC.FLIGHTS) B

WHERE A.TAILNUM = B.TAILNUM

GROUP BY YEAR

The above SQL statement’s output is collected to a data.frame in R, using the following commands in R:

library("RPostgreSQL")

library("sqldf")

## loads the PostgreSQL driver

drv <- dbDriver("PostgreSQL")

## Open a connection

#con <- dbConnect(drv, dbname="flights")

con <- dbConnect(drv, dbname="flights",host="localhost",port=5432,user="postgres",password="XXXX")

## Submits a statement

rs <- dbSendQuery(con, "

SELECT YEAR, AVG(DEP\_DELAY) AVG\_DEP\_DELAY, AVG(ARR\_DELAY) AVG\_ARR\_DELAY, COUNT(\*) AS FLIGHT\_FREQ FROM

(SELECT YEAR,TAILNUM FROM PUBLIC.PLANES) A, (SELECT DEP\_DELAY, ARR\_DELAY, TAILNUM FROM PUBLIC.FLIGHTS) B

WHERE A.TAILNUM = B.TAILNUM

GROUP BY YEAR

"

)

## fetch all elements from the result set

year\_of\_made\_delay <- fetch(rs,n=-1)

## Submit and execute the query

dbGetQuery(con, "select \* from R\_packages")

## Closes the connection

dbDisconnect(con)

## Frees all the resources on the driver

dbUnloadDriver(drv)

Plotted scatter plots between the YEAR and AVG\_DEP\_DELAY, YEAR and AVG\_ARR\_DELAY, and YEAR and FLIGHT\_FREQ.

Where YEAR is the year of manufacture

AVG\_DEP\_DELAY is the average departure delay

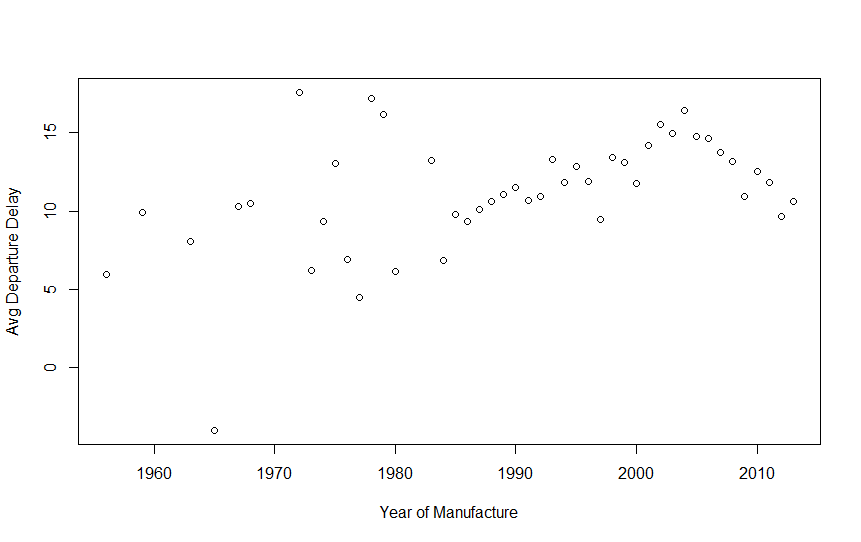
AVG\_ARR\_DELAY is the average arrival delay

FLIGHT\_FREQ is the number of times flights are operated grouped by their year of manufacture.

Obtained the following scatter plots:

Plot-1 is obtained by using the following R Command:

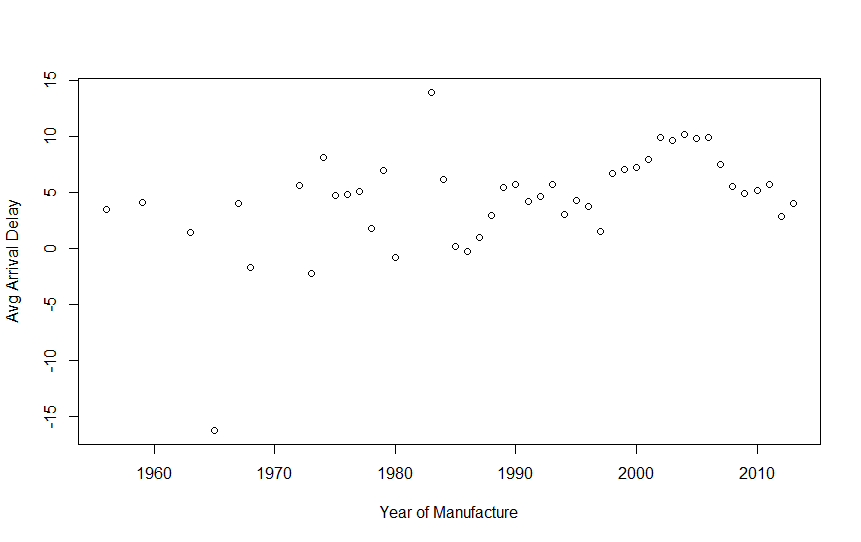
plot(year\_of\_made\_delay$year,year\_of\_made\_delay$avg\_dep\_delay,xlab="Year of Manufacture",ylab="Avg Departure Delay")



Plot – 1: Scatter plot between “Average Departure delay” and “Flight’s year of manufacture”. The plot shows that there is NO strong relationship between Average Departure Delay and the year of manufacture of the flight.

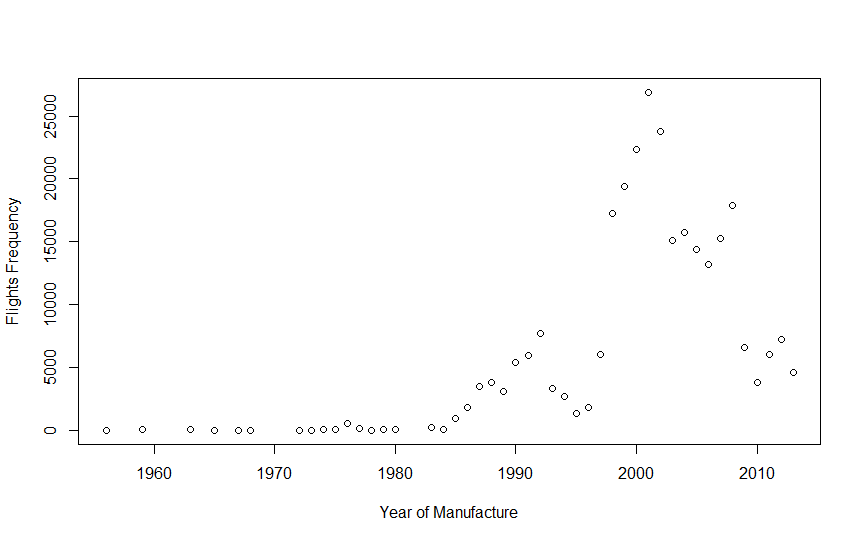
To obtain Plot-2, I used the following command:

plot(year\_of\_made\_delay$year,year\_of\_made\_delay$avg\_arr\_delay,xlab="Year of Manufacture",ylab="Avg Arrival Delay")



Plot – 2: Scatter plot between “Average Arrival delay” and “Flight’s year of manufacture”. The plot 2 also shows that there is NO strong relationship between Average Arrival Delay and Year of manufacture.

We have to also check, if there is any relationship between the year of manufacture and the flights frequency. This is to confirm if we have any relationship between the flight’s year of manufacture and the number of times it is flown. This plot is shown in Plot-3 below. The plot shows that the flights frequency increased 1985 (manufacture year) till 2000 (manufacture year). The plot however shows that there is no strong relationship between the frequency of flights and the flight year of manufacture.



Plot – 3: Scatter plot between “Flights Frequency” and “Year of Manufacture”

I used the R to plot the scatter plot. None of the scatter plots have shown any strong relationship between the variables (YEAR and AVG\_DEP\_DELAY, YEAR and AVG\_ARR\_DELAY, and YEAR and FLIGHT\_FREQ). Also I used the following query to calculate the coefficient of correlation.

SELECT

CORR(YEAR,AVG\_DEP\_DELAY) AS COEFF\_CORR\_YEAR\_DEP\_DELAY,

CORR(YEAR,AVG\_ARR\_DELAY) AS COEFF\_CORR\_YEAR\_ARR\_DELAY,

CORR(YEAR,FLIGHTS\_FREQ) AS COEFF\_CORR\_YEAR\_FLIGHTS\_FREQ

FROM

(

SELECT YEAR, AVG(DEP\_DELAY) AVG\_DEP\_DELAY, AVG(ARR\_DELAY) AVG\_ARR\_DELAY, COUNT(\*) AS FLIGHTS\_FREQ FROM

(SELECT YEAR,TAILNUM FROM PUBLIC.PLANES) A, (SELECT DEP\_DELAY, ARR\_DELAY, TAILNUM FROM PUBLIC.FLIGHTS) B

WHERE A.TAILNUM = B.TAILNUM

GROUP BY YEAR

) AS A

--Obtained the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| Coeff of Corr | Avg Dep Delay | Avg Arr Delay | Frequency |
| Year of manufacture | 0.461453 | 0.433125 | 0.651205 |

Since the scatter plot has shown no strong pattern and also the coefficient of correlations are also not very strong I concluded that the delay is flights is not strongly related to the year of manufacture.

--5. Ask (and if possible answer) a question that also requires joining information from two or more tables in the flights database, and/or assumes that additional information can be collected in advance of answering your question.

It would have been better, if we have Day of the week mentioned in the flight table.

Such information will enable us to find the relative frequencies of all the flights for all the days of week. My intention is to check, if there is any difference of the number of flights flown on any week day. If there is such difference, are there any flight carriers, who operate more flights on specific days of the week, as any underlying business strategy. Here are the steps I followed:

I created another table named "flights\_modified" similar to the structure of flights table, and added two new columns to the table flights\_modified. The new columns will contain the date of the flight and the day of the week of the flight.

create table flights\_modified (like flights);

alter table flights\_modified add column date\_obj date;

alter table flights\_modified add column day\_of\_week int;

insert into flights\_modified (select \* from flights);

update flights\_modified set date\_obj = to\_date(to\_char(year,'9999')||'-'||ltrim(to\_char(month,'99'))||'-'||ltrim(to\_char(day,'99')),'YYYY-MM-DD');

update flights\_modified set day\_of\_week = EXTRACT(dow from to\_date(to\_char(year,'9999')||'-'||ltrim(to\_char(month,'99'))||'-'||ltrim(to\_char(day,'99')),'YYYY-MM-DD') );

Let us get the number of flights that flew on various week days.

select Day\_of\_week, Num\_of\_flights, ((cast(Num\_of\_flights as real))/all\_flights) \* 100

relative\_flight\_freq

from

(select count(\*) Num\_of\_flights, case when day\_of\_week = 0 then 'Sunday'

when day\_of\_week = 1 then 'Monday'

when day\_of\_week = 2 then 'Tuesday'

when day\_of\_week = 3 then 'Wednesday'

when day\_of\_week = 4 then 'Thursday'

when day\_of\_week = 5 then 'Friday'

when day\_of\_week = 6 then 'Saturday'

end as Day\_of\_week

from flights\_modified group by day\_of\_week) as a,

(select count(\*) as all\_flights from flights) b

order by 2 desc

**day\_of\_week num\_of\_flights relative\_flight\_freq**

Monday 50690 15.05155

Tuesday 50422 14.97197

Friday 50308 14.93812

Thursday 50219 14.91169

Wednesday 50060 14.86448

Sunday 46357 13.76494

Saturday 38720 11.49726

Gives the following result (when plotted using barchart function of R. Use the following code to produce the barchart):

con <- dbConnect(drv, dbname="flights",host="localhost",port=5432,user="postgres",password="XXXXXXX")

## Submits a statement

rs <- dbSendQuery(con, "

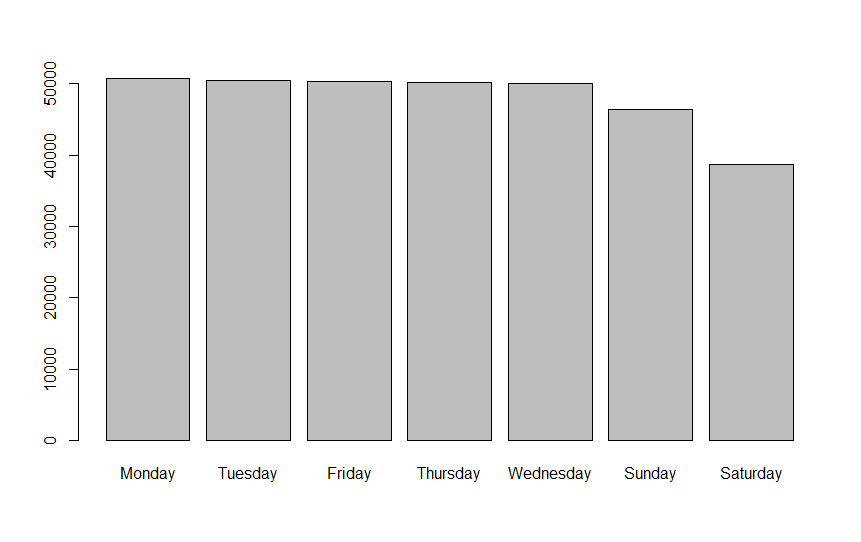
select count(\*) Num\_of\_flights, day

from flights\_modified group by day order by 2 desc

"

)

barplot(flight\_freq\_by\_week\_day$num\_of\_flights,names.arg=flight\_freq\_by\_week\_day$day\_of\_week)



**Plot – A: Relative flights frequency on all the days of the week.**

The least number of flights are flown on Saturday, and the most on Monday (approximately same for Monday, Tuesday, Friday, Thursday, Wednesday)

Let us find any hidden strategy of any carriers to capture the demand by operating more flights on Saturday/Sunday. Our intention is to find if there is any airline, who operates more flights on Saturday or Sunday.

Let us get the relative frequencies of flights from all the 16 carriers. Plot 4 shows this.

**R Code is given below:**

rs <- dbSendQuery(con, "

select carrier, count(\*) num\_of\_flights from flights

group by carrier

order by 2 desc

"

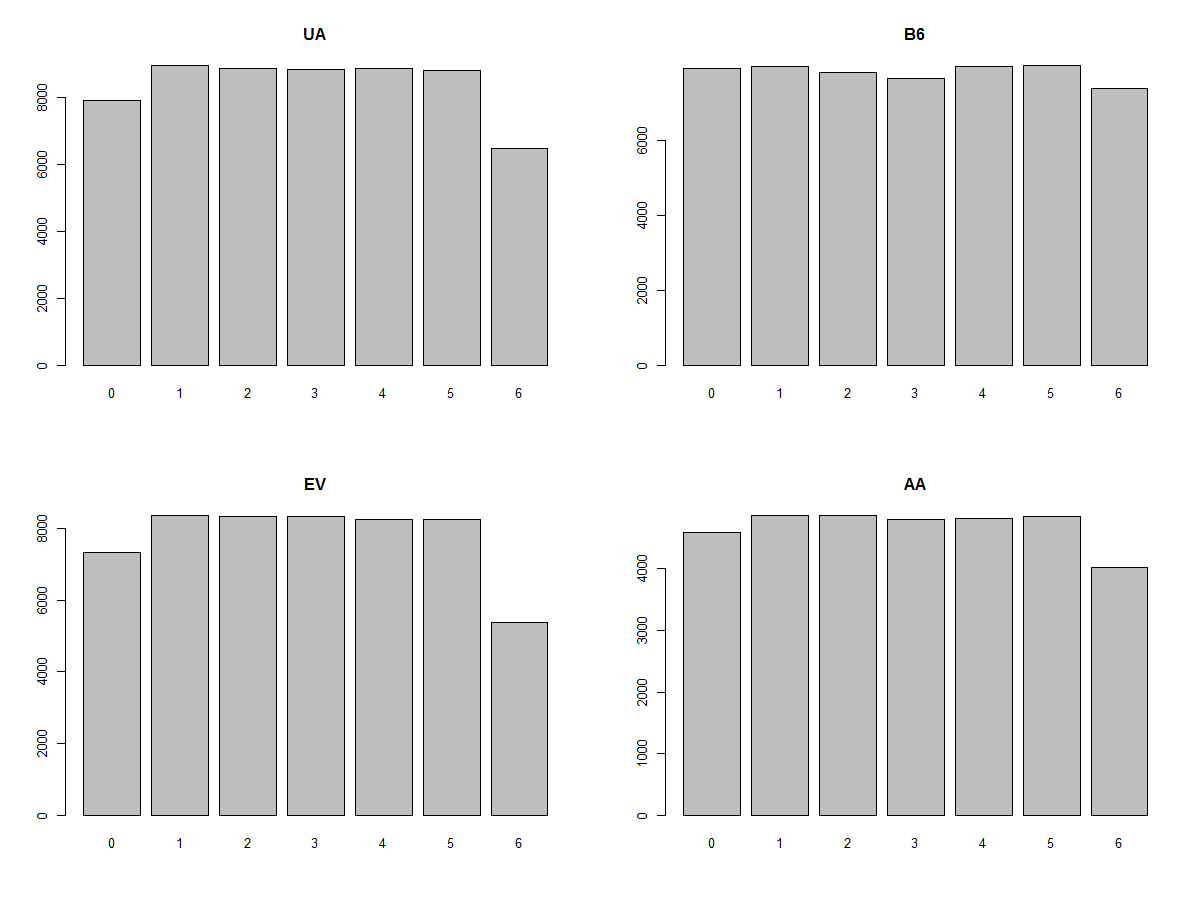
)

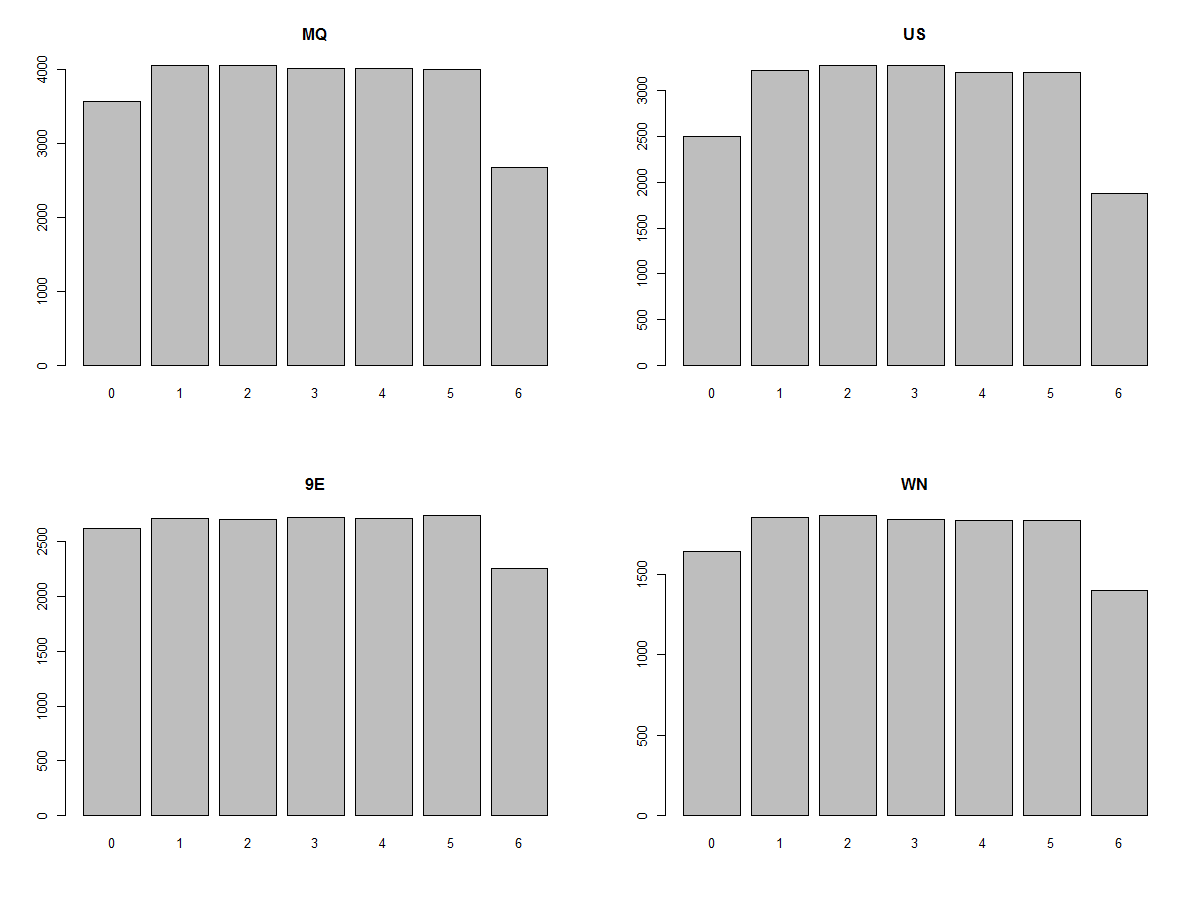
con <- dbConnect(drv, dbname="flights",host="localhost",port=5432,user="postgres",password="XXXXXX")

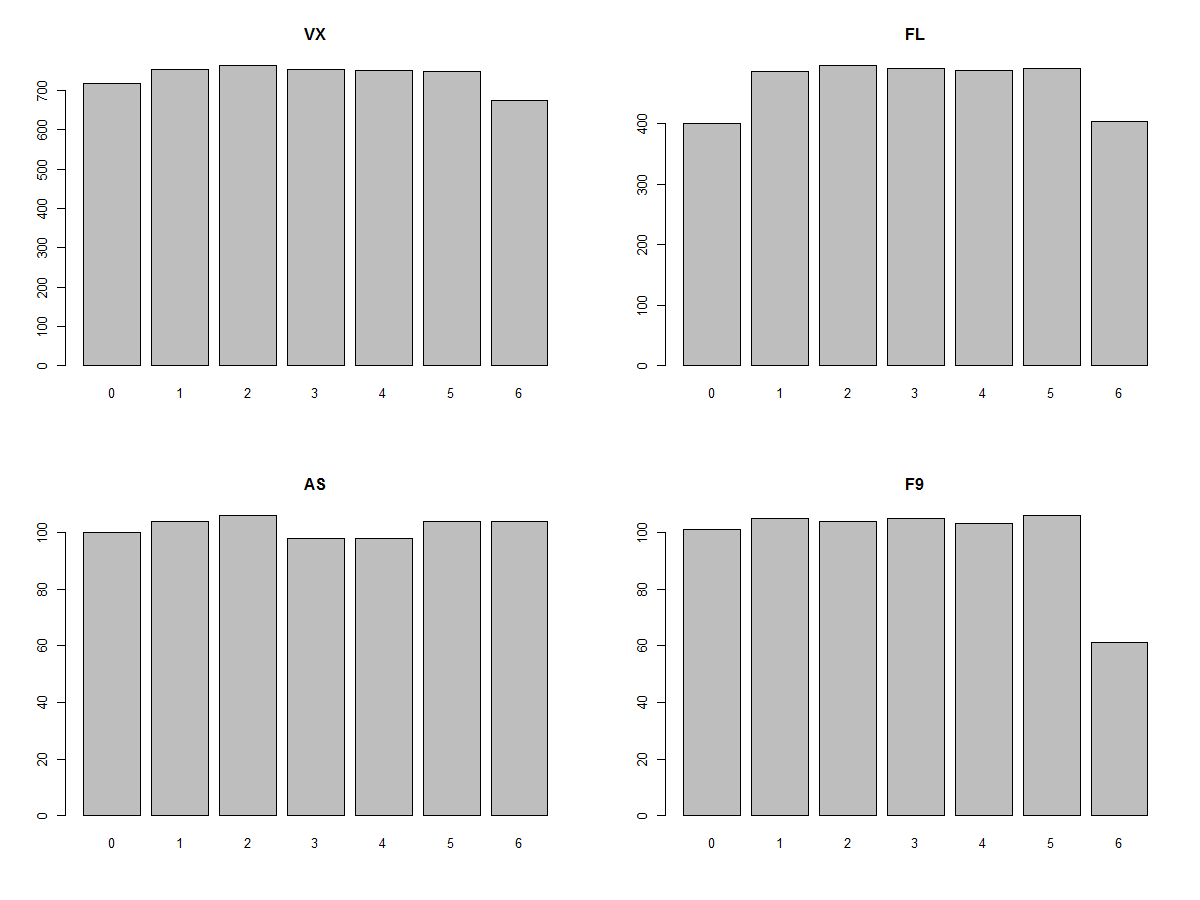
flights\_by\_carrier <- fetch(rs,n=-1)

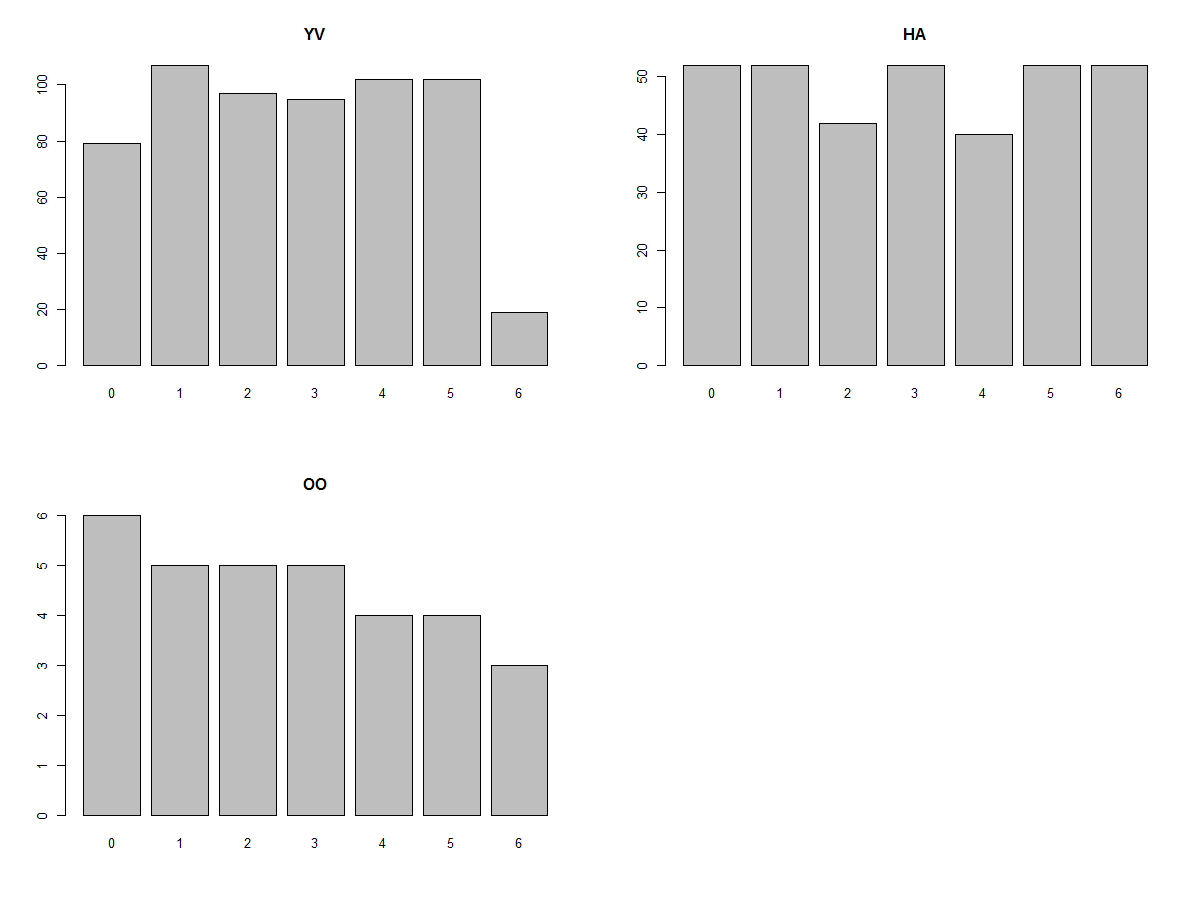
|  |
| --- |
| barplot(flights\_by\_carrier$num\_of\_flights,names.arg=flights\_by\_carrier$carrier) |
| Plot – 4: Relative frequencies of flights for all carriers, flown for the whole year |
|  |

Plot – 5: Relative frequencies of all flights flying on each week day for each carrier









|  |
| --- |
| In plot – 5, the horizontal axis represents the day of the week, 0 being Sunday, 1 being Monday, 2 Tuesday, 3 being Wednesday, 4 being Thursday, 5 being Friday and 6 being Saturday. |

In Plot – 5, the vertical axis represents the number of flights flown. The main heading in each plot represents the carrier information (in abbreviated form). See below for the carrier abbreviations used in Plot-5.

9E - "Endeavor Air Inc."

AA - "American Airlines Inc."

AS - "Alaska Airlines Inc."

B6 - "JetBlue Airways"

DL - "Delta Air Lines Inc."

EV - "ExpressJet Airlines Inc."

F9 - "Frontier Airlines Inc."

FL - "AirTran Airways Corporation"

HA - "Hawaiian Airlines Inc."

MQ - "Envoy Air"

OO - "SkyWest Airlines Inc."

UA - "United Air Lines Inc."

US - "US Airways Inc."

VX - "Virgin America"

WN - "Southwest Airlines Co."

YV - "Mesa Airlines Inc."

**Analysis of Plot-5:**

Almost all carriers operated less number of flights relatively on Saturday and most number of flights on Monday. But we have some exceptions as given below:

The B6 – JetBlue Carrier operated 2nd most number of flights among all the carriers. But this operator operated almost same number of flights on all days of the week, including Saturday. Perhaps, their business strategy is to capture the demand of travelers on Saturday and Sunday. The OO – “SkyWest Airlines” operated least number of flights. But it operated most of its flights on Sunday. The HA – Hawaiian Airlines operated approximately same number of flights on Monday, Sunday, Friday and Saturday. This could be because most of the people who go to Vacation to Hawaii tend to travel on Friday, Saturday, Sunday and Monday. On the similar lines the “AS – Alaska Airlines” also operated most of its flights on Friday, Saturday, Sunday and Monday perhaps to cater the needs of the people going on vacations to Alaska.

The VX – Virgin America operated around the same number of flights on all the week days (approximately). As per Wiki, the main goal of VX airlines is to cater the needs of passengers who travel between east coast and west coast, connecting the seaports. Since the seaports (or beaches) are popular destinations for vacations, the VX airlines might have some good demand due to people going on vacations during the weekends.

They have good demand during the week days also, since many people travel between coasts due to businesses or work.

*We can finally conclude that airlines which operate to popular vacation spots tend to operate more number of their flights on weekends than on weekdays, while the airlines which operate to all the cities (which are not popular vacation destinations) operate their flights more on weekdays and less number of flights on weekends (with the exception being B6 – Jet Blue airlines, which operates its flights on Saturday and Sunday also, almost equally on all the week days). Perhaps the Jet Blue airlines also serve popular vacation spots, and hence the demand for their airlines on weekends also.*

I used the following code in R to generate the graphs listed in Plot – 5.

rs <- dbSendQuery(con, "

select day\_of\_week,

Num\_of\_flights, ((cast(Num\_of\_flights as real))/all\_flights) \* 100 as freq\_by\_week, carrier

from

(

select count(\*) Num\_of\_flights, day\_of\_week, carrier

from flights\_modified group by day\_of\_week, carrier

ORDER BY day\_of\_week) as a,

(select count(\*) as all\_flights from flights ) b

ORDER BY 4,day\_of\_week

"

)

par(mfrow=c(2,2))

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="UA"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="UA"],main="UA")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="B6"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="B6"],main="B6")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="EV"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="EV"],main="EV")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="AA"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="AA"],main="AA")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="MQ"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="MQ"],main="MQ")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="US"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="US"],main="US")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="9E"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="9E"],main="9E")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="WN"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="WN"],main="WN")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="VX"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="VX"],main="VX")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="FL"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="FL"],main="FL")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="AS"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="AS"],main="AS")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="F9"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="F9"],main="F9")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="YV"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="YV"],main="YV")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="HA"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="HA"],main="HA")

barplot(flight\_rel\_freq\_by\_carrier$num\_of\_flights[flight\_rel\_freq\_by\_carrier$carrier=="OO"],names.arg=flight\_rel\_freq\_by\_carrier$day\_of\_week[flight\_rel\_freq\_by\_carrier$carrier=="OO"],main="OO")