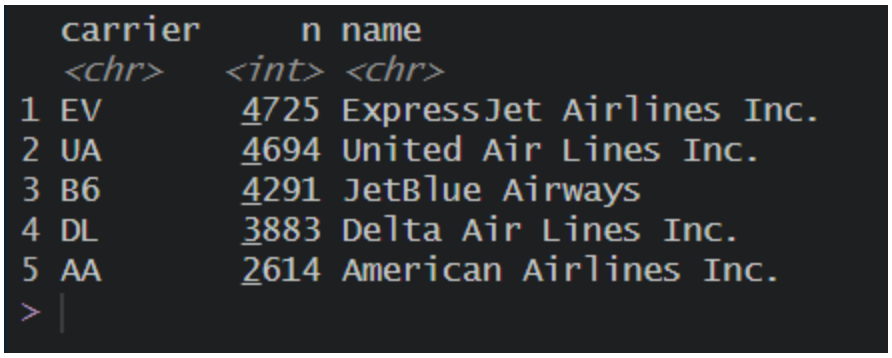
Hw-data-transform-Pham

Hw1 - analyze flights from nycflights13 library.

```
## homework data-transform

library(nycflights13)
## nycflights13 This package contains information about all flig
## to destinations in the United States, Puerto Rico, and the Ar
## To help understand what causes delays, it also includes a nur

library(tidyverse)

## Ask questions about this datasets
data("flights")
data("airlines")
data("airports")
data("planes")
data("weather")
```

```
## Q1. most flight carrier in Sep 2013
flights %>%
  filter(month == 9, year == 2013) %>%
  count(carrier) %>%
  arrange(desc(n)) %>%
  head(5) %>%
  left_join(airlines)
```



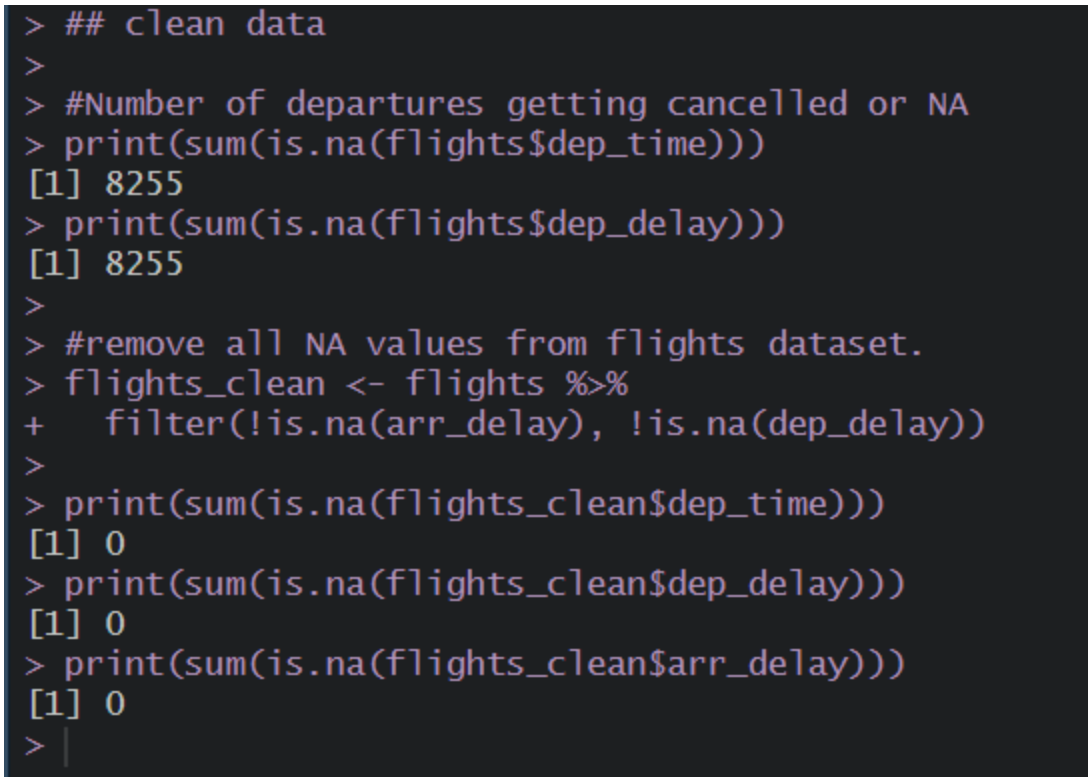
	carrier	n	name
	<chr>	<int>	<chr>
1	EV	4725	ExpressJet Airlines Inc.
2	UA	4694	United Air Lines Inc.
3	B6	4291	JetBlue Airways
4	DL	3883	Delta Air Lines Inc.
5	AA	2614	American Airlines Inc.

```
> 
```

```
## Q2. find avg departure delay and avg arrival delay for each a
flights %>%
  filter(arr_delay > 0 & arr_delay != "NA" & dep_delay > 0 & dep
  group_by(carrier) %>%
  summarise(avg_arr_delay = mean(arr_delay, na.rm = TRUE),
            avg_dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
  arrange(desc(avg_arr_delay)) %>%
  left_join(airlines,by= "carrier")
```

```
   carrier avg_arr_delay avg_dep_delay name
   <chr>           <dbl>         <dbl> <chr>
 1 OO               74.6          64.8 SkyWest Airlines Inc.
 2 HA               72.4          75.6 Hawaiian Airlines Inc.
 3 F9               63.6          57.2 Frontier Airlines Inc.
 4 YV               62.9          60.7 Mesa Airlines Inc.
 5 9E               60.6          63.4 Endeavor Air Inc.
 6 EV               58.5          58.7 ExpressJet Airlines Inc.
 7 VX               57.4          55.9 Virgin America
 8 MQ               54.6          50.3 Envoy Air
 9 AA               53.1          52.6 American Airlines Inc.
10 DL               53.1          52.0 Delta Air Lines Inc.
11 FL               51.9          47.2 AirTran Airways Corporation
12 B6               51.6          49.4 JetBlue Airways
13 WN               47.9          49.3 Southwest Airlines Co.
14 AS               45.4          50.7 Alaska Airlines Inc.
15 US               45.0          40.1 US Airways Inc.
16 UA               44.4          44.0 United Air Lines Inc.
>
```

```
## clean data

#Number of departures getting cancelled or NA
print(sum(is.na(flights$dep_time)))
print(sum(is.na(flights$dep_delay)))

#remove all NA values from flights dataset.
flights_clean <- flights %>%
  filter(!is.na(arr_delay), !is.na(dep_delay))

print(sum(is.na(flights_clean$dep_time)))
print(sum(is.na(flights_clean$dep_delay)))
print(sum(is.na(flights_clean$arr_delay)))
```



```
> ## clean data
>
> #Number of departures getting cancelled or NA
> print(sum(is.na(flights$dep_time)))
[1] 8255
> print(sum(is.na(flights$dep_delay)))
[1] 8255
>
> #remove all NA values from flights dataset.
> flights_clean <- flights %>%
+   filter(!is.na(arr_delay), !is.na(dep_delay))
>
> print(sum(is.na(flights_clean$dep_time)))
[1] 0
> print(sum(is.na(flights_clean$dep_delay)))
[1] 0
> print(sum(is.na(flights_clean$arr_delay)))
[1] 0
>
```

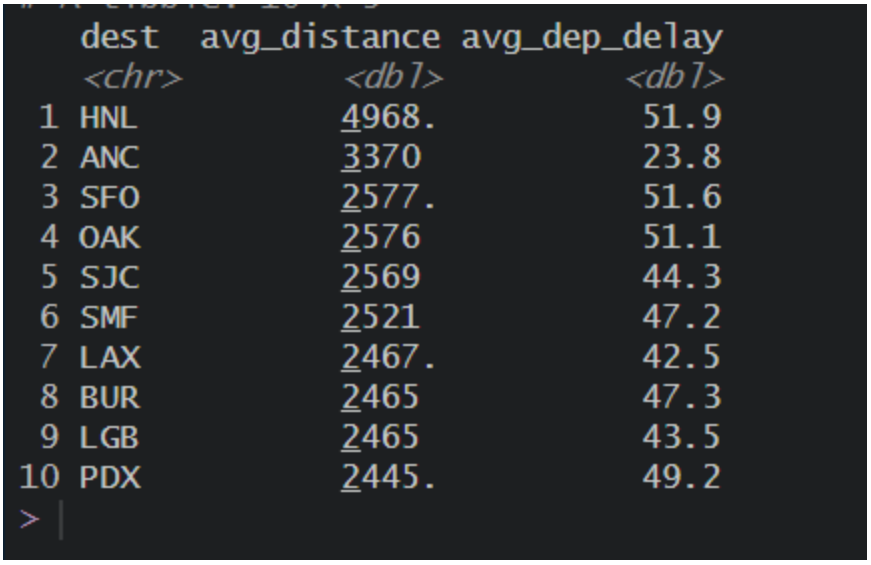
```
## Q3. avg departure delays and avg arrival delay by month with
flights_per_m <- flights_clean %>%
  group_by(month) %>%
  count(month)
colnames(flights_per_m)[2] ="flights per month"

flights_clean %>%
  filter(arr_delay>0 & dep_delay>0) %>%
  group_by(month) %>%
  summarise(avg_arr_delay = mean(arr_delay),
            avg_dep_delay = mean(dep_delay)) %>%
  left_join(flights_per_m,by= "month")
```

	month	avg_arr_delay	avg_dep_delay	`flights per month`
	<int>	<dbl>	<dbl>	<int>
1	1	47.6	45.5	26398
2	2	45.2	45.4	23611
3	3	51.1	52.2	27902
4	4	57.5	54.1	27564
5	5	51.6	53.5	28128
6	6	65.6	62.9	27075
7	7	65.1	62.4	28293
8	8	50.6	50.0	28756
9	9	51.2	54.1	27010
10	10	41.2	43.3	28618
11	11	38.7	38.3	26971
12	12	49.3	45.6	27020

```
## Q4. avg departure delays and avg arrival delay by weekdays
flights_clean %>%
  mutate(weekdays = weekdays(time_hour)) %>%
  filter(arr_delay>0 & dep_delay>0) %>%
  group_by(weekdays) %>%
  summarise(avg_arr_delay = mean(arr_delay),
            avg_dep_delay = mean(dep_delay)) %>%
  arrange(desc(avg_arr_delay))
```

	weekdays	avg_arr_delay	avg_dep_delay
	<chr>	<dbl>	<dbl>
1	Monday	57.1	55.5
2	Thursday	56.4	55.0
3	Wednesday	53.0	51.0
4	Friday	52.3	52.1
5	Sunday	50.6	49.8
6	Tuesday	49.8	48.3
7	Saturday	42.6	44.3

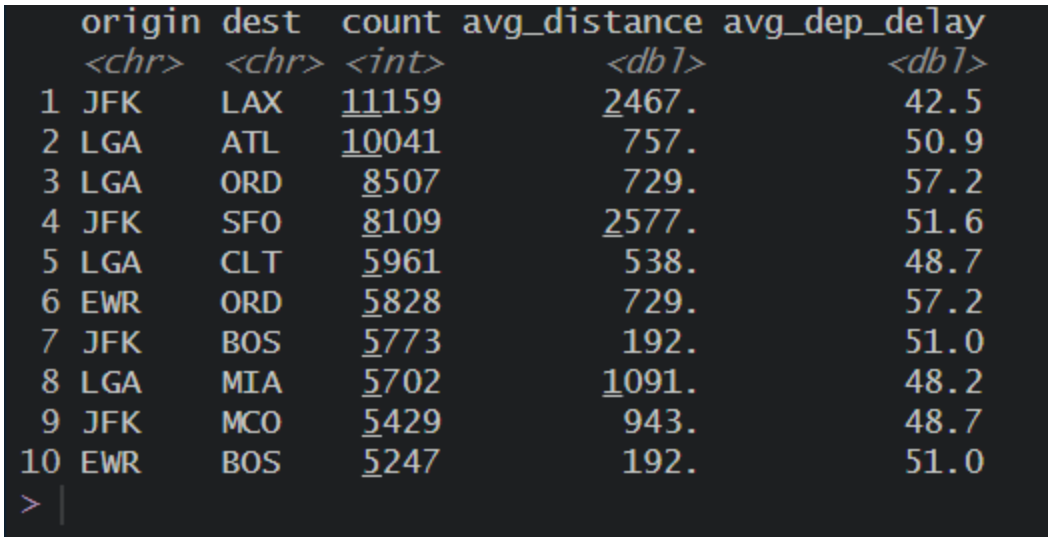
```
## Q5. avg departure delays by different destination
avg_delays_dest <- flights_clean %>%
  filter(arr_delay>0 & dep_delay>0) %>%
  group_by(dest) %>%
  summarise(avg_distance = mean(distance, na.rm = TRUE),avg_dep_
  arrange(desc(avg_distance))

avg_delays_dest %>%
  head(10)
```



	dest	avg_distance	avg_dep_delay
	<chr>	<dbl>	<dbl>
1	HNL	4968.	51.9
2	ANC	3370	23.8
3	SFO	2577.	51.6
4	OAK	2576	51.1
5	SJC	2569	44.3
6	SMF	2521	47.2
7	LAX	2467.	42.5
8	BUR	2465	47.3
9	LGB	2465	43.5
10	PDX	2445.	49.2

```
> 
```

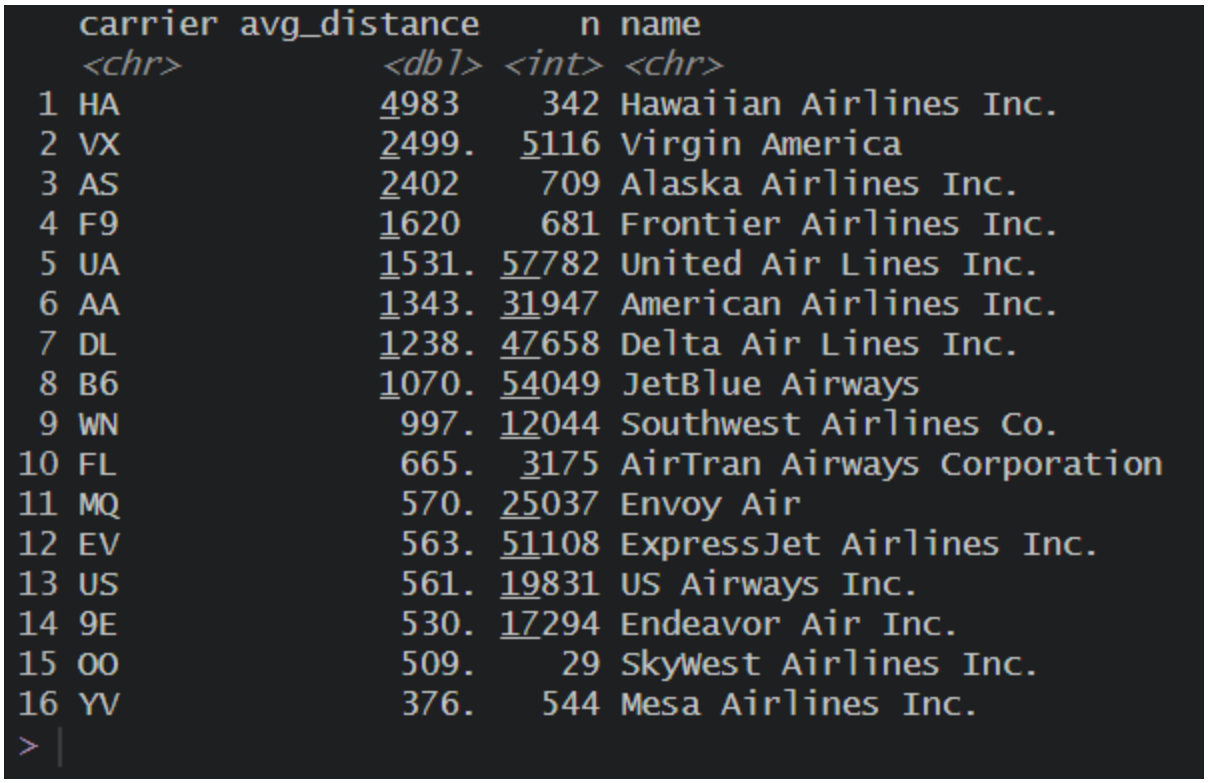
```
##Q6. The route with the most frequent flights with avg departu
route01 <- flights_clean %>%
  group_by(origin,dest) %>%
  summarise(count=n()) %>%
  arrange(desc(count)) %>%
  left_join(avg_delays_dest, by= "dest")

route01 %>%
  head(10)
```



	origin	dest	count	avg_distance	avg_dep_delay
	<chr>	<chr>	<int>	<dbl>	<dbl>
1	JFK	LAX	11159	2467.	42.5
2	LGA	ATL	10041	757.	50.9
3	LGA	ORD	8507	729.	57.2
4	JFK	SFO	8109	2577.	51.6
5	LGA	CLT	5961	538.	48.7
6	EWR	ORD	5828	729.	57.2
7	JFK	BOS	5773	192.	51.0
8	LGA	MIA	5702	1091.	48.2
9	JFK	MCO	5429	943.	48.7
10	EWR	BOS	5247	192.	51.0

```
>
```

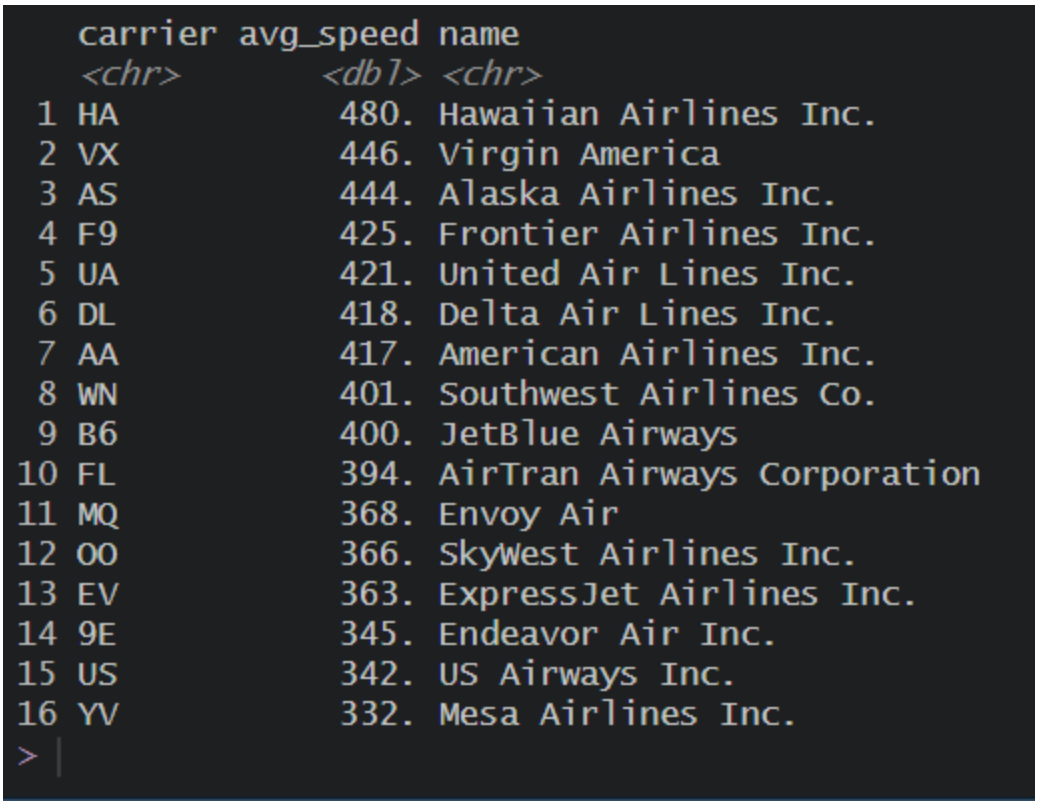
```
##Q7. Number of flights and Average flight distance for each ai
n_of_flights <- flights_clean %>%
  count(carrier) %>%
  arrange(desc(n))

flights_clean %>%
  group_by(carrier) %>%
  summarise( avg_distance = mean(distance)) %>%
  arrange(desc(avg_distance)) %>%
  left_join(n_of_flights, by= "carrier") %>%
  left_join(airlines, by= "carrier")
```



	carrier	avg_distance	n	name
	<chr>	<dbl>	<int>	<chr>
1	HA	4983	342	Hawaiian Airlines Inc.
2	VX	2499.	5116	Virgin America
3	AS	2402	709	Alaska Airlines Inc.
4	F9	1620	681	Frontier Airlines Inc.
5	UA	1531.	57782	United Air Lines Inc.
6	AA	1343.	31947	American Airlines Inc.
7	DL	1238.	47658	Delta Air Lines Inc.
8	B6	1070.	54049	JetBlue Airways
9	WN	997.	12044	Southwest Airlines Co.
10	FL	665.	3175	AirTran Airways Corporation
11	MQ	570.	25037	Envoy Air
12	EV	563.	51108	ExpressJet Airlines Inc.
13	US	561.	19831	US Airways Inc.
14	9E	530.	17294	Endeavor Air Inc.
15	OO	509.	29	SkyWest Airlines Inc.
16	YV	376.	544	Mesa Airlines Inc.

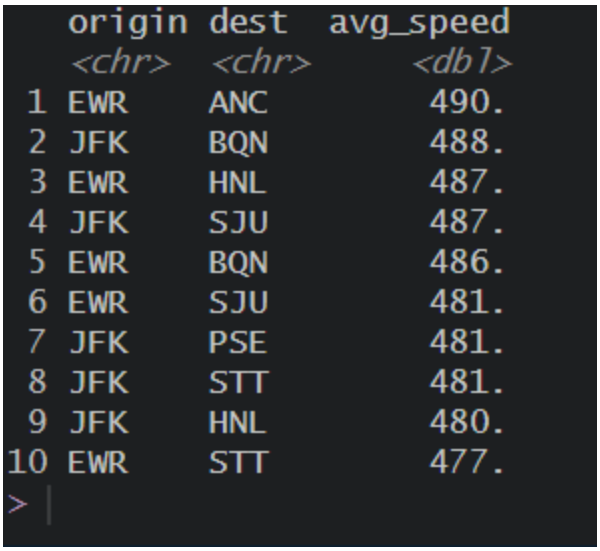
```
> 
```

```
##Q8. average speed for each airlines
#speed in the unit of miles per hour
flights_clean %>%
  mutate(speed = distance /(air_time / 60) ) %>%
  group_by(carrier) %>%
  summarise(avg_speed = mean(speed)) %>%
  arrange(desc(avg_speed)) %>%
  left_join(airlines, by= "carrier")
```



```
   carrier avg_speed name
   <chr>       <dbl> <chr>
 1 HA            480. Hawaiian Airlines Inc.
 2 VX            446. Virgin America
 3 AS            444. Alaska Airlines Inc.
 4 F9            425. Frontier Airlines Inc.
 5 UA            421. United Air Lines Inc.
 6 DL            418. Delta Air Lines Inc.
 7 AA            417. American Airlines Inc.
 8 WN            401. Southwest Airlines Co.
 9 B6            400. JetBlue Airways
10 FL            394. AirTran Airways Corporation
11 MQ            368. Envoy Air
12 OO            366. SkyWest Airlines Inc.
13 EV            363. ExpressJet Airlines Inc.
14 9E            345. Endeavor Air Inc.
15 US            342. US Airways Inc.
16 YV            332. Mesa Airlines Inc.
>
```

```
##Q9. average speed for different route
flights_clean %>%
  mutate(speed = distance /(air_time / 60) ) %>%
  group_by(origin,dest) %>%
  summarise(avg_speed = mean(speed)) %>%
  arrange(desc(avg_speed)) %>%
  head(10)
```



	origin	dest	avg_speed
	<chr>	<chr>	<dbl>
1	EWR	ANC	490.
2	JFK	BQN	488.
3	EWR	HNL	487.
4	JFK	SJU	487.
5	EWR	BQN	486.
6	EWR	SJU	481.
7	JFK	PSE	481.
8	JFK	STT	481.
9	JFK	HNL	480.
10	EWR	STT	477.

```
> 
```

```
##Q10. find the oldest passenger aircraft models that were still

flights_clean %>%
  select(tailnum) %>%
  left_join(planes, by = "tailnum") %>%
  arrange(year) %>%
  distinct(tailnum) %>%
  pull()

flights_clean %>%
  filter(tailnum %in% c('N381AA')) %>%
  select(tailnum,flight,time_hour)

planes %>%
  filter(tailnum %in% c('N381AA'))
## The oldest passenger aircraft models that were still use in 2
## This aircraft has an operational lifespan of 57 years (2013-1
```

```
> ##Q10. find the oldest passenger aircraft models that were still use in 2013
>
> flights_clean %>%
+   select(tailnum) %>%
+   left_join(planes, by = "tailnum") %>%
+   arrange(year) %>%
+   distinct(tailnum) %>%
+   pull()
   [1] "N381AA" "N201AA" "N567AA" "N575AA" "N378AA" "N14629" "N615AA" "N425AA"
   [9] "N383AA" "N364AA" "N840MQ" "N621AA" "N508AA" "N675MC" "N711MQ" "N545AA"
  [17] "N762NC" "N737MQ" "N767NC" "N376AA" "N774NC" "N519AA" "N779NC" "N777NC"
  [25] "N600TR" "N202AA" "N782NC" "N525AA" "N350AA" "N519MQ" "N604DL" "N603DL"
  [33] "N602DL" "N750AT" "N757AT" "N319AA" "N613DL" "N609DL" "N320AA" "N551AA"
  [41] "N610DL" "N612DL" "N658SW" "N520AA" "N693SW" "N503US" "N614DL" "N397AA"
  [49] "N347AA" "N657SW" "N650SW" "N662SW" "N660SW" "N604SW" "N663SW" "N507US"
```

```
> flights_clean %>%
+   filter(tailnum %in% c('N381AA')) %>%
+   select(tailnum,flight,time_hour)
# A tibble: 22 × 3
   tailnum flight time_hour
   <chr>    <int> <dttm>
 1 N381AA      59 2013-01-30 07:00:00
 2 N381AA      85 2013-10-07 15:00:00
 3 N381AA    2351 2013-10-08 17:00:00
 4 N381AA      59 2013-11-07 07:00:00
 5 N381AA      85 2013-11-12 15:00:00
 6 N381AA     179 2013-12-17 10:00:00
 7 N381AA      59 2013-12-18 08:00:00
 8 N381AA      85 2013-02-01 15:00:00
 9 N381AA     179 2013-02-03 10:00:00
10 N381AA      59 2013-02-07 07:00:00
# i 12 more rows
# i Use `print(n = ...)` to see more rows
>
> planes %>%
+   filter(tailnum %in% c('N381AA'))
# A tibble: 1 × 9
  tailnum  year type                manufacturer model engines seats speed engine
  <chr>   <int> <chr>               <chr>        <chr>   <int> <int> <int> <chr>
1 N381AA   1956 Fixed wing multi e… DOUGLAS      DC-7…       4   102   232 Recip…
> ## The oldest passenger aircraft models that were still use in 2013 is N381AA
> ## This aircraft has an operational lifespan of 57 years (2013-1956).
>
```

```
##Q11. aggregate function related to weather conditions
weather %>%
  select(origin) %>%
  distinct(origin) %>%
  pull()

airports %>%
```

```
  filter(faa %in% c("JFK", "EWR","LGA")) %>%
  left_join(weather, by = c("faa" = "origin")) %>%
  group_by(faa) %>%
  summarise(avg_temp=mean(temp, na.rm = T),avg_humid=mean(humid,
            avg_wind_speed=mean(wind_speed, na.rm = T),
            avg_pressure=mean(pressure, na.rm = T))
```

	faa	avg_temp	avg_humid	avg_wind_speed	avg_pressure
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	EWR	55.5	63.1	9.46	1018.
2	JFK	54.5	65.2	11.5	1018.
3	LGA	55.8	59.3	10.6	1018.

>

Hw2 - write table into PostgreSQL server.

```
# HW2  from pizza restaurant SQL
#      create 3-5 dataframes => write table into PostgreSQL serv

## connect to PostgreSQL server
library(RPostgreSQL)
library(tidyverse)

## create connection
con <- dbConnect(
  PostgreSQL(),
  host = "*****.db.elephantsql.com",
  dbname = "******",
  user = "******",
  password = "*****************************",
  port = 5432
)

##Create dataframe
```

```
menu <- tribble(
  ~menu_id, ~menu_name, ~price_menu,~type_menu,
01,"Rustica", 500, "Pizza",
02,"Bufalina", 360, "Pizza",
03,"Prosciutto_Pizza", 680, "Pizza",
04,"Roasted_Chicken", 500, "Main_course",
05,"Apple&Pineapple,", 80, "Drink",
06,"Caesar_Salad", 200, "Salad",
07,"Crab&Mango_Salad", 200, "Salad",
08,"Truffle_Bruschetta", 400, "Starter",
09,"Water", 20, "Drink",
10,"Lobster_Spaghetti",1660, "Main_course"
)

customer <- tribble(
  ~customer_id, ~first_Name,~last_Name,
~email,~country,~age,~phone,
01,"Walter","White", "Heisenberg99.1@example.com", "USA", 50, "(
02,"Hector","Salamanca", "DingDingDing@example.com", "Mexico", 1
03,"Jesse","Pinkman", "AyoMrWhite@example.com", "USA", 25, "012-
04,"Greta","Thunberg","HowDareYou@example.com","Sweden",20,"066-
05,"Nancy","Pelosi","Pelosi@example.com","USA",80,"012-896-7842"
06,"Sam","Bankman-Fried", "SamBankman@example.com", "USA", 30, "
07,"Do","Kwon","Luna888@example.com","Korea",32,"088-888-8888",
08,"Gojo","Satoru","UnlimitedVoid@example.com","Japan",28,"077-8
)

customer_order <- tribble(
  ~order_id, ~customer_id, ~date, ~menu_id, ~quantity,
01,06,"2023-01-01",08,1,
02,06,"2023-01-01",05,2,
03,06,"2023-01-01",10,1,
04,06,"2023-01-01",03,1,
05,08,"2023-01-01",01,1,
06,08,"2023-01-01",06,1,
```

```
07,08,"2023-01-01",02,1,
08,07,"2023-01-02",05,2,
09,07,"2023-01-02",03,1,
10,01,"2023-01-05",08,1,
11,01,"2023-01-05",03,1,
12,02,"2023-01-07",07,1,
13,02,"2023-01-07",03,1,
14,03,"2023-02-05",06,2,
15,03,"2023-02-10",03,1,
16,04,"2023-02-15",05,2,
17,04,"2023-02-30",03,1,
18,05,"2023-03-10",04,2,
19,05,"2023-03-15",03,1,
20,02,"2023-03-30",02,1,
21,02,"2023-04-01",03,1,
22,02,"2023-04-01",09,2,
23,04,"2023-05-01",04,1,
24,04,"2023-05-01",02,2,
25,04,"2023-05-01",09,1,
26,01,"2023-06-15",01,2,
27,01,"2023-06-15",06,1,
28,01,"2023-06-15",09,1,
29,05,"2023-07-20",08,1,
30,05,"2023-07-20",09,1,
31,07,"2023-08-05",09,1,
32,07,"2023-08-05",07,2,
33,07,"2023-08-05",02,1,
34,04,"2023-09-18",05,2,
35,04,"2023-09-18",06,1,
36,04,"2023-09-18",04,2,
37,05,"2023-10-09",05,1,
38,05,"2023-10-09",10,1,
39,05,"2023-10-09",07,1,
40,02,"2023-11-11",09,1,
41,02,"2023-11-11",08,1,
42,02,"2023-11-11",02,2,
```

```
43,06,"2023-12-22",10,2,
44,06,"2023-12-22",05,2,
45,06,"2023-12-22",02,3
)

## database List Tables
dbWriteTable(con, "menu", menu)
dbWriteTable(con, "customer", customer)
dbWriteTable(con, "customer_order", customer_order)

# check database tables ,get data
dbListTables(con)
dbGetQuery(con, "select * from menu")
dbGetQuery(con, "select * from customer")
dbGetQuery(con, "select * from customer_order")

# Disconnect from database
dbDisconnect(con)
```