HW 01 (dataset "Flights" 5-10 questions)

```
library(tidyverse)
df_flights <- read_csv("flights.csv")
glimpse(df_flights)
df_airlines <- read_csv("airlines.csv")
df_airports <- read_csv("airports.csv")
df_airports <- rename(df_airports,dest = faa)
df_planes <- read_csv("planes.csv")
```

```
Rows: 336,776
Columns: 19
$ year           <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,
$ month          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
$ day            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
$ dep_time       <dbl> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558
$ sched_dep_time <dbl> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600
$ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2,
$ arr_time       <dbl> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 84
$ sched_arr_time <dbl> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 85
$ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7,
$ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6",
$ flight         <dbl> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301,
$ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N3
$ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA
$ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD
$ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149,
$ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733
$ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6,
$ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59
```

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
```

Q1: In 2013, what season did most people travelled from New York City to other destination?

```
group_ss <- mutate(df_flights,
                   seasonal = case_when(month %in% c(6,7,8) ~ "Summer",
                                        month %in% c(9,10,11) ~ "Autumn",
                                        month %in% c(12,1,2) ~ "Winter",
                                        month %in% c(3,4,5) ~ "Spring"))
group_ss %>% count(seasonal) %>% arrange(desc(n))
```

A spec_tbl_df: 4 × 2

seasonal	n
<chr>	<int>
Summer	86995
Spring	85960
Autumn	83731
Winter	80090

Q2:Which airlines were the most popular in summer 2013?

```
group_ss %>% filter(seasonal == "Summer") %>% count(carrier) %>% arrange(desc(n))
```

Q3: Which the top five most popular destinations have been visited in summer 2013?

```
group_ss %>%
    filter(seasonal == "Summer") %>% count(dest) %>% arrange(desc(n)) %>% head(5)
```

Q4: What the date was frequently delayed?

```
delay_t <- df_flights %>%
filter(!is.na(dep_delay), !is.na(arr_delay)) #ต้องไม่เท่ากับNA

delay_t %>%
  group_by(month,day) %>%
  summarise(mean = round(mean(dep_delay),2),
              sd = round(sd(dep_delay),2)) %>%
  arrange(desc(mean)) %>%
  head(15)
```

A grouped_df: 15 × 4

month	day	mean	sd
<dbl>	<dbl>	<dbl>	<dbl>
3	8	83.65	90.41
7	1	56.22	68.36
9	2	53.06	90.22
12	5	52.45	89.91
7	10	51.20	91.36
5	23	50.63	83.93
9	12	49.43	93.35
6	28	49.14	77.02
6	24	47.54	79.96
7	22	46.67	90.77
4	19	45.74	87.18
6	13	45.65	71.56
7	23	44.57	60.23
6	30	44.08	72.78
8	8	43.21	68.93

```
`summarise()` has grouped output by 'month'. You can override using the `.group
```

Q5: Airplanes flew from JFK to LAS, the Turbo Fans more efficient and economical than other types of engines? (JFK to

LAS 2248 miles)`

```
planes <- mutate(df_flights,milespermins = distance/((hour*60.0)+minute)) %>%
    left_join(df_planes, by = "tailnum") %>% #select(engine,tailnum,distance,hour
    filter(!is.na(engine),origin == "JFK", dest == "LAS") %>%
    group_by(engine) %>%
    summarise(max_mpm = round(max(milespermins,na.rm = TRUE),2),
              mean_mpm = round(mean(milespermins,na.rm = TRUE),2),
              median_mpm = round(median(milespermins,na.rm = TRUE),2),
              count = n()) %>%
              arrange(desc(count),desc(max_mpm))
planes
```

A tibble: 6 × 5

engine	max_mpm	mean_mpm	median_mpm	count
<chr>	<dbl>	<dbl>	<dbl>	<int>
Turbo-fan	5.87	3.48	3.75	2037
Turbo-jet	4.68	2.64	2.35	1332
Reciprocating	5.76	3.90	3.75	17
Turbo-shaft	3.77	2.62	2.02	8
4 Cycle	5.42	2.83	2.18	5
Turbo-prop	5.42	4.35	5.42	3

HW 02 (3-4 df, RPostgreSQL)

```
#install.packages("RPostgreSQL") Bug
#library(RPostgreSQL) Bug
df_friends <- data.frame(
    id_friends <- 1:4,
    name = c("Brown","Sally","Cony","Choco"),
    fav_color = c("Yellow","Orange","Pink","Purple"),
    age = c(25,20,22,18)
)
df_foods <- data.frame(
    id_menu <- 1:4,
    menu = c("Salmon","Burger","Cake","Brownie"),
    price = c(159,89,99,59)
)
```

```
df_tracks <- data.frame(
    tracks_id = c(1,2,5,8),
    tracks_name = c("Magic Man","Pink Venom","Love Story","Gone")
)
```

```
install.packages("RPostgreSQL")
library(RPostgreSQL)

con <- dbConnect(
  PostgreSQL(),
  host = "arjuna.db.elephantsql.com",
  dbname = "hbixmfrb",
  port = 5432,
  user = "hbixmfrb",
  password = "JzVYFb76HjqY8Gv4XAL7n13DiAVSS0KZ"
)
dbListTables(con)
df_friends <- data.frame(
    id_friends <- 1:4,
    name = c("Brown","Sally","Cony","Choco"),
    fav_color = c("Yellow","Orange","Pink","Purple"),
    age = c(25,20,22,18)
)
df_foods <- data.frame(
    id_menu <- 1:4,
    menu = c("Salmon","Burger","Cake","Brownie"),
    price = c(159,89,99,59)
)
df_tracks <- data.frame(
    tracks_id = c(1,2,5,8),
    tracks_name = c("Magic Man","Pink Venom","Love Story","Gone")
)

dbWriteTable(con,"df_friends", df_friends)
dbWriteTable(con,"df_foods", df_foods)
dbWriteTable(con,"df_tracks", df_tracks)


df1 <- dbGetQuery(con, "SELECT * FROM df_friends")
df2 <- dbGetQuery(con, "SELECT * FROM df_foods")
df3 <- dbGetQuery(con, "SELECT * FROM df_tracks")

dbDisconnect(con)
```