Data Transformation R

Tonnsaai Anantrechai

```
library(nycflights13)
library(tidyverse)
library(dplyr)
## Q1. Top Frequently delayed 5 airlines of 2013 (unit in flights)

Q1 <- flights %>%
  select(carrier,arr_delay) %>%
  filter(arr_delay > 0) %>%
  count(carrier) %>%
  arrange(desc(n)) %>%
  group_by(carrier) %>%
  left_join(airlines) %>%
  head (5)
Q1
```

```
## # A tibble: 5 x 3
## # Groups:   carrier [5]
##   carrier     n name
##   <chr>   <int> <chr>
## 1 EV      24484 ExpressJet Airlines Inc.
## 2 B6      23609 JetBlue Airways
## 3 UA      22222 United Air Lines Inc.
## 4 DL      16413 Delta Air Lines Inc.
## 5 MQ      11693 Envoy Air
```

```
## Q2. Top 5 largest capacity of aircraft in 2013
Q2 <- planes %>%
  select(manufacturer,model,seats) %>%
  arrange(desc(seats)) %>%
  head(5)
Q2
```

```
## # A tibble: 5 x 3
##   manufacturer model   seats
##   <chr>        <chr>   <int>
## 1 BOEING       747-451   450
## 2 BOEING       777-222   400
## 3 BOEING       777-222   400
## 4 BOEING       777-200   400
## 5 BOEING       777-224   400
```

```r
## Q3. Most frequently used 5 planes in origin from JFK airport in 2013
Q3 <- flights %>%
  select(flight,tailnum,origin) %>%
  filter(origin == "JFK") %>%
  count(tailnum) %>%
  arrange(desc(n)) %>%
  group_by(tailnum) %>%
  rename(Dep_flights=n) %>%
  left_join(planes)

ANS_Q3 <- Q3 %>%
  select(manufacturer,model,tailnum,Dep_flights) %>%
  filter(tailnum != "") %>%
  head(5)

ANS_Q3
```

```
## # A tibble: 5 x 4
## # Groups:   tailnum [5]
##   manufacturer model           tailnum Dep_flights
##   <chr>        <chr>           <chr>         <int>
## 1 BOEING       767-223         N328AA          393
## 2 EMBRAER      ERJ 190-100 IGW N258JB          391
## 3 BOEING       767-223         N338AA          388
## 4 BOEING       767-223         N327AA          387
## 5 BOEING       767-223         N335AA          385
```

```r
## Q4.Average departure delay in each month
Q4 <- flights %>%
  select(carrier,month,dep_delay) %>%
  filter(dep_delay !=0 & dep_delay > 0) %>%
  group_by(month) %>%
  summarise(avg_delay_month = mean(dep_delay),
            max_delay_month = max(dep_delay))
Q4
```

```
## # A tibble: 12 x 3
##    month avg_delay_month max_delay_month
##    <int>           <dbl>           <dbl>
##  1     1            35.3            1301
##  2     2            35.3             853
##  3     3            39.6             911
##  4     4            44.2             960
##  5     5            39.2             878
##  6     6            49.8            1137
##  7     7            48.8            1005
##  8     8            37.3             520
##  9     9            35.7            1014
## 10    10            31.6             702
## 11    11            28.7             798
## 12    12            37.2             896
```