빅데이터를이용한통계그래픽스 HW #4

1678226 유찬미 (Chanmi Yoo)

2019년 11월 8일

###### \* nycflights13 패키지에 있는 flights 자료를 이용

# Load Packages  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.4.4

## -- Attaching packages ----------------------- tidyverse 1.2.1 --

## √ ggplot2 3.0.0 √ purrr 0.2.5  
## √ tibble 1.4.2 √ dplyr 0.7.6  
## √ tidyr 0.8.1 √ stringr 1.3.1  
## √ readr 1.2.1 √ forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.4.4

## Warning: package 'tibble' was built under R version 3.4.4

## Warning: package 'tidyr' was built under R version 3.4.4

## Warning: package 'readr' was built under R version 3.4.4

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'dplyr' was built under R version 3.4.4

## Warning: package 'stringr' was built under R version 3.4.4

## Warning: package 'forcats' was built under R version 3.4.4

## -- Conflicts -------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(nycflights13)  
  
# Load Data  
data("flights")

###### 1. flights자료에서 출발이 2시간 이상 늦어진 비행 자료 중 year, month, day, dep\_delay, arr\_delay, carrier, origin, dest 만을 뽑아서 Late2에 저장하시오.

Late2 <- flights %>%   
 filter(dep\_delay >= 2\*60) %>%   
 select(year, month, day, dep\_delay, arr\_delay, carrier, origin, dest)

## Warning: package 'bindrcpp' was built under R version 3.4.4

###### 2. flights 자료에서 뉴욕에서 시카고로의 비행만을 뽑아서 Chicago에 저장하시오.

###### - Chicago의 공항코드: ORD, MDW

###### - 자료 중 year, month, day, dep\_delay, arr\_delay, carrier, origin, dest만을 저장

Chicago <- flights %>%   
 filter(dest %in% c('ORD', "MDW")) %>%   
 select(year, month, day, dep\_delay, arr\_delay, carrier, origin, dest)

###### 3. flights 자료에서 2013년 1년간 출발이 2시간 이상 늦어진 비행이 가장 많은 항공사를 알아보려고 한다. Late2 자료를 이용하여 각 항공사별로 2시간이상 늦어진 비행 건수를 구하고 이를 건수가 많은 항공사부터 나열하시오.

Late2 %>%   
 group\_by(carrier) %>%   
 summarise(count = n()) %>%   
 arrange(desc(count))

## # A tibble: 16 x 2  
## carrier count  
## <chr> <int>  
## 1 EV 2489  
## 2 B6 1652  
## 3 UA 1385  
## 4 DL 1104  
## 5 9E 793  
## 6 AA 728  
## 7 MQ 620  
## 8 WN 456  
## 9 US 244  
## 10 VX 183  
## 11 FL 153  
## 12 F9 34  
## 13 YV 23  
## 14 AS 17  
## 15 HA 5  
## 16 OO 2

###### 4. 30분 이상 늦게 출발했으나 늦어진 출발시간 중 10분 이상을 비행으로 단축한 비행을 추출하여 비행거리가 큰 순서대로 나열하시오.

###### - month, day, distance, dest, dep\_delay,arr\_delay 만 표시

flights %>%   
 filter(dep\_delay >= 30) %>%   
 select(month, day, distance, dest, dep\_delay, arr\_delay) %>%   
 mutate(delay = dep\_delay - arr\_delay) %>%   
 filter(delay >= 10) %>%   
 arrange(desc(distance))

## # A tibble: 20,618 x 7  
## month day distance dest dep\_delay arr\_delay delay  
## <int> <int> <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1 6 4983 HNL 79 28 51  
## 2 1 7 4983 HNL 102 50 52  
## 3 1 9 4983 HNL 1301 1272 29  
## 4 1 18 4983 HNL 123 65 58  
## 5 1 23 4983 HNL 101 82 19  
## 6 2 9 4983 HNL 186 154 32  
## 7 2 23 4983 HNL 206 126 80  
## 8 2 24 4983 HNL 36 -33 69  
## 9 2 26 4983 HNL 60 -27 87  
## 10 3 8 4983 HNL 48 35 13  
## # ... with 20,608 more rows

###### 5. 60분 이상의 출발지연 건수가 가장 많은 달을 찾고 해당 달에 어떤 항공사(carrier)가 60분 이상의 지연이 가장 많은지를 알아보고자 한다. 이를 위한 코드를 작성하고 해당 달, 항공사를 찾으시오.

flights %>%   
 filter(dep\_delay >= 60) %>%   
 group\_by(month) %>%   
 summarise(count = n())

## # A tibble: 12 x 2  
## month count  
## <int> <int>  
## 1 1 1852  
## 2 2 1688  
## 3 3 2391  
## 4 4 2572  
## 5 5 2357  
## 6 6 3555  
## 7 7 3877  
## 8 8 2338  
## 9 9 1345  
## 10 10 1366  
## 11 11 1121  
## 12 12 2597

flights %>%   
 filter(dep\_delay >= 60 & month == 7) %>%   
 group\_by(carrier) %>%   
 summarise(count = n()) %>%   
 arrange(desc(count))

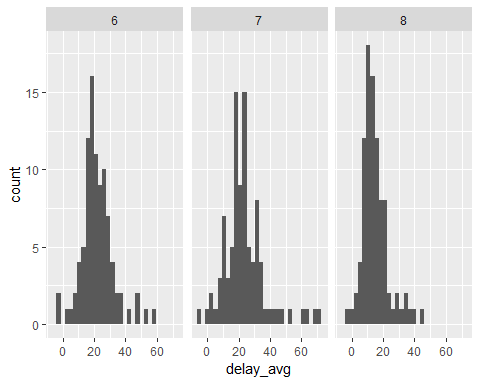
## # A tibble: 15 x 2  
## carrier count  
## <chr> <int>  
## 1 EV 787  
## 2 B6 783  
## 3 UA 578  
## 4 DL 507  
## 5 MQ 303  
## 6 9E 263  
## 7 AA 236  
## 8 WN 139  
## 9 US 127  
## 10 VX 74  
## 11 FL 56  
## 12 YV 11  
## 13 F9 9  
## 14 AS 3  
## 15 HA 1

[결과] 60분 이상 출발지연 건수가 가장 많은 달은 7월(3877건)이다. 7월에 60분 이상 출발지연이 가장 많았던 항공사는 EV(787회)이다.

###### 6. 여름 (6,7,8월)에 도착지별로 출발이 지연되는 정도를 파악하고자 한다. 비행의 도착지 별로 월평균 출발지연시간을 구하고 이를 적절한 그림으로 나타낸 후 그림을 해석하시오.

montly\_delay <- flights %>%   
 filter(month %in% c(6:8)) %>%   
 group\_by(dest, month) %>%   
 summarise(delay\_avg = mean(dep\_delay, na.rm = T))  
  
ggplot(montly\_delay, aes(delay\_avg)) +  
 geom\_histogram(bins = 30) +  
 facet\_wrap(~month)

## Warning: Removed 1 rows containing non-finite values (stat\_bin).



[설명] 6, 7월에는 약 20분, 8월은 약 10분의 출발지연이 발생한 도착지가 많았다.