2017학년도 제 2학기

빅데이터를 이용한 통계그래픽스

개인 프로젝트 2차

보고서

담당 교수님 성함 : 이은경

이름 : 정유진

학번 : 1602018

학과 : 통계학과

제출일 : 2017년 11월 24일

목차

서론

Project-data.cvs는 변수가 31개고 자료의 개수가 5819079개인 자료로 2015년의 비행에 대한 정보를 나타내주고 있다. 이 보고서는 2015년 비행의 특성을 찾아내기 위함이 목적이다. 이 특성을 찾기 위해 R-studio프로그램을 이용해 자료를 시각화 하는 작업을 행했다.

2015년 자료는 실제 비행이 이루어진 것과 비행이 아예 취소되어 가지 못하는 경우가 있어 크게 2개, '취소된 비행기' 그리고 '취소되지 않은 비행기'로 나누어 보았다. 첫째로 '취소된 비행기'는 시기별, 항공사별 2가지로 나누어 특성을 살펴보고 각각 취소된 이유를 2가지에 적용해 비교해서 분석해보았다. 두 번째로 '취소되지 않은 비행기'는 지연에 대해 알아보기도 하고 시기, 항공사, 비행기, 출발공항 별 비행횟수에 대해서 알아보았다.

1. 자료 살펴보기

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
-
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R version 3.4.2
```

자료 읽어들이기

```
flights0 <-read.csv("c:/temp/project-data.csv")
flights1 <-as.tibble(flights0)
```

자료는 2015 년에 움직인 비행이므로 취소된 비행과 취소 하지 않은 비행으로 나누어 살펴보도록 하겠다. 그러기 위해 일단 취소된 것의 개수를 알아보겠다.

#취소된 비행 개수 찾기

```
 flights1 %>%
 count(CANCELLED)
```

```
## # A tibble: 2 x 2
##   CANCELLED       n
##       <int>   <int>
## 1         0 5729195
## 2         1   89884
```

여기서 0 는 취소되지 않은 것 것이고 1 은 취소된 것이다. 취소가 되지 않은 비행의 개수는 5729195 개이고 취소된 비행의 개수는 89884 개이다. 우리가 주목해야 될 것은 취소되지

않고 운행한 비행이므로 먼저 취소된 것에 대해 살펴보고 나서 취소되지 않은 자료를 살펴보도록 하겠다.

2. 취소된 2015 년 비행의 특징 살펴보기

취소된 비행을 head 를 이용하여 살펴보도록 하겠다.

```
cancelled <-filter(flights1,CANCELLED=="1")
head(cancelled)
```

```
## # A tibble: 6 x 31
##    YEAR MONTH   DAY DAY_OF_WEEK AIRLINE FLIGHT_NUMBER TAIL_NUMBER
##   <int> <int> <int>       <int>  <fctr>         <int>      <fctr>
## 1  2015     1     1           4      AS           136      N431AS
## 2  2015     1     1           4      AA          2459      N3BDAA
## 3  2015     1     1           4      OO          5254      N746SK
## 4  2015     1     1           4      MQ          2859      N660MQ
## 5  2015     1     1           4      OO          5460      N583SW
## 6  2015     1     1           4      MQ          2926      N932MQ
## # ... with 24 more variables: ORIGIN_AIRPORT <fctr>,
## #   DESTINATION_AIRPORT <fctr>, SCHEDULED_DEPARTURE <int>,
## #   DEPARTURE_TIME <int>, DEPARTURE_DELAY <int>, TAXI_OUT <int>,
## #   WHEELS_OFF <int>, SCHEDULED_TIME <int>, ELAPSED_TIME <int>,
## #   AIR_TIME <int>, DISTANCE <int>, WHEELS_ON <int>, TAXI_IN <int>,
## #   SCHEDULED_ARRIVAL <int>, ARRIVAL_TIME <int>, ARRIVAL_DELAY <int>,
## #   DIVERTED <int>, CANCELLED <int>, CANCELLATION_REASON <fctr>,
## #   AIR_SYSTEM_DELAY <int>, SECURITY_DELAY <int>, AIRLINE_DELAY <int>,
## #   LATE_AIRCRAFT_DELAY <int>, WEATHER_DELAY <int>
```

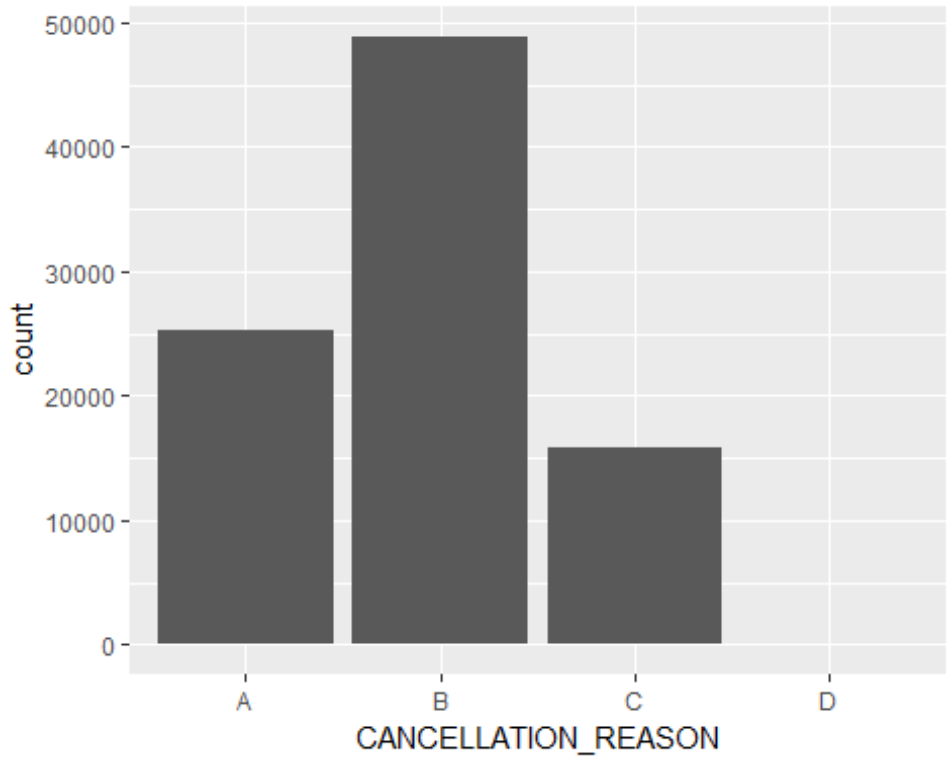
그리고 취소된 이유는 다음과 같다.

```
cancelled %>%
  group_by(CANCELLATION_REASON) %>%
  summarise(count=n())
```

```
## # A tibble: 4 x 2
##   CANCELLATION_REASON count
##                 <fctr> <int>
## 1                   A 25262
## 2                   B 48851
## 3                   C 15749
## 4                   D    22
```

여기서 A 는 'Airlines/Carrier'을 나타내고 B 는 'Weather' C 는 'National Air System' D 는 'Security'을 보여준다. 그래프로 보면 시각적으로 갯수의 차이가 뚜렷하다.

```
ggplot(cancelled,aes(CANCELLATION_REASON))+geom_bar()
```



시각적으로 B 의 이유가 가장 많고 48851 개로 가장 많고 그 뒤를 A 가 25262 개로 많다는 것을 알 수 있다. 또한 이것을 통해 보안상의 이유로 취소된 비행기는 다른 이유와 비교해보았을 때 거의 없다는 것을 알 수 있다. 취소비행을 분석해 보면서 취소비행이유가 얼마나 영향을 끼치는지도 앞으로 알아보도록 하겠다.

##1) **월별 취소된 비행 분석**

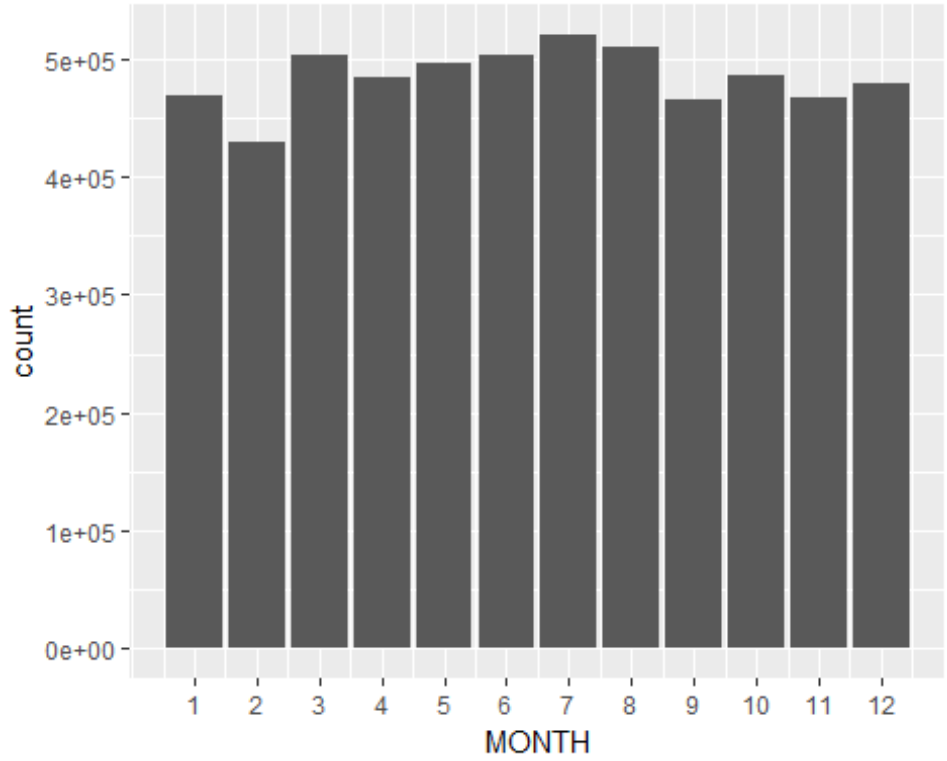
시기에 따른 취소비행 횟수 연, 일, 요일에 따른 취소비행횟수는 별로 중요하지 않을 듯 하여 달에 따른 취소비행 횟수를 그래프를 통해 알아보도록 하겠다. 먼저 달에 따른 모든 비행횟수(취소된 비행 + 취소되지 않은 비행)를 보도록 하겠다.

```
flights1 %>%
  group_by(MONTH)%>%
  summarise(count=n())
```

```
## # A tibble: 12 x 2
##    MONTH  count
##    <int>  <int>
##  1     1 469968
##  2     2 429191
```

```
##  3     3 504312
##  4     4 485151
##  5     5 496993
##  6     6 503897
##  7     7 520718
##  8     8 510536
##  9     9 464946
## 10    10 486165
## 11    11 467972
## 12    12 479230
```

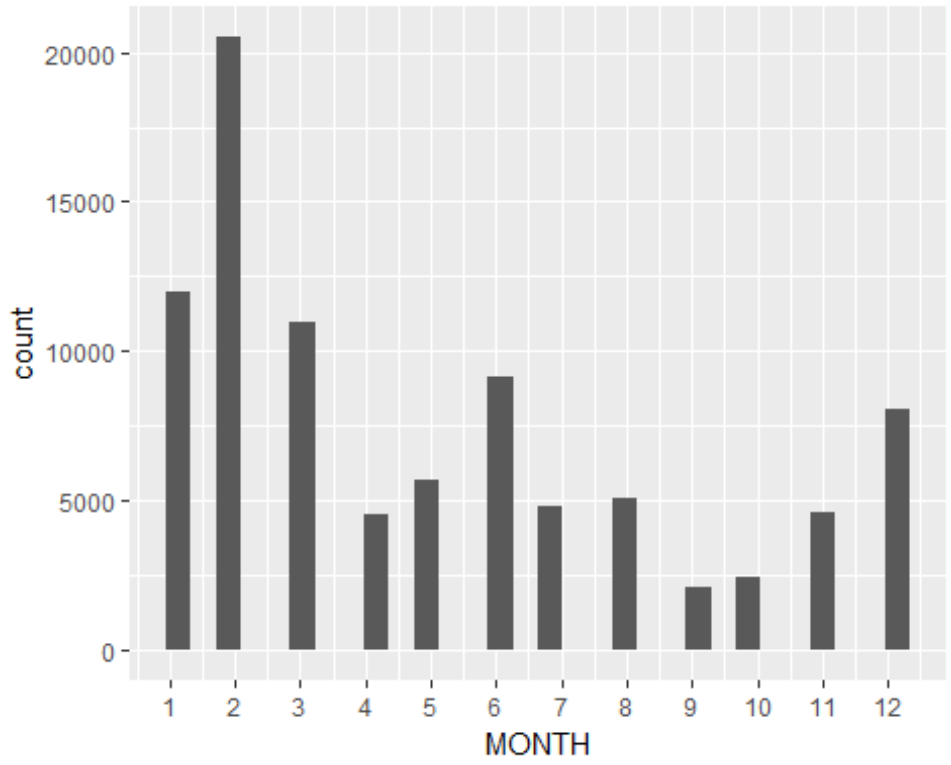
```
ggplot(flights1,aes(MONTH))+geom_bar()+scale_x_continuous(breaks=c(1:12))
```



시각적으로는 월 별로 차이가 별로 나지 않아 보인다. 그리고 수치적으로 봐도 2 월의 최저 개수인 429191 개부터 7 월 520718 개까지 엄청난 차이는 보이지 않는다. 이제 취소된 항공편을 먼저 그래프로 살펴보도록 하겠다.

```
ggplot(cancelled, aes(MONTH)) +geom_histogram()+scale_x_continuous(breaks=c(1:
12))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



시각적으로 1, 2, 3 월에 가장 취소비행횟수가 많고 9 월에 가장 적다는 것을 확인할 수 있다. 정확히 수치로 확인해보도록 하겠다.

```
cancelled %>%
group_by(MONTH)%>%
summarise(count=n())
```

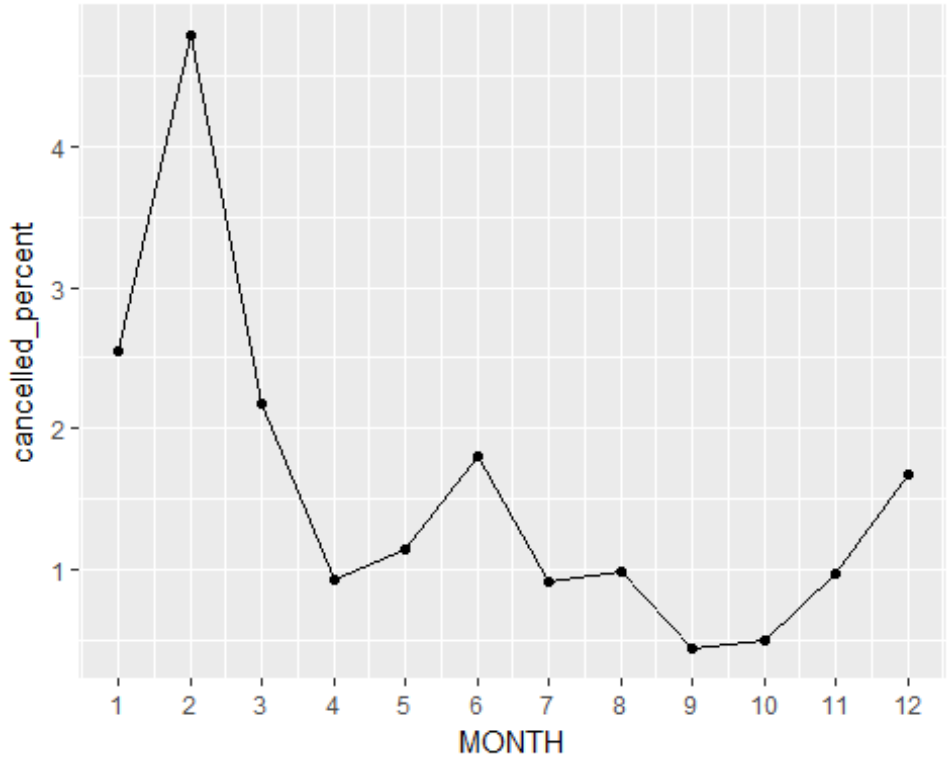
```
## # A tibble: 12 x 2
##    MONTH count
##    <int> <int>
##  1     1 11982
##  2     2 20517
##  3     3 11002
##  4     4  4520
##  5     5  5694
##  6     6  9120
##  7     7  4806
##  8     8  5052
##  9     9  2075
## 10    10  2454
## 11    11  4599
## 12    12  8063
```

특히 1 월, 2 월, 3 월에서 10000 번이 넘어가면서 다른 달보다 월등히 많이 취소된 것을 확인할 수 있다. 그런데 이것만으로는 판단하기 정확하지 않아 전체 항공과의 비율을 나타내서 보도록 하겠다.

```
all_percentage <-(flights1 %>%
   group_by(MONTH)%>%
   summarise(count2=n()))
cancel_percentage <- (cancelled %>%
 group_by(MONTH)%>%
 summarise(count1=n()))
percentage <-(cancel_percentage %>%
 left_join(all_percentage) %>% mutate (cancelled_percent=count1/count2*100))

## Joining, by = "MONTH"

  ggplot(percentage,aes(MONTH,cancelled_percent))+geom_line()+geom_point()+sc
ale_x_continuous(breaks=c(1:12))
```

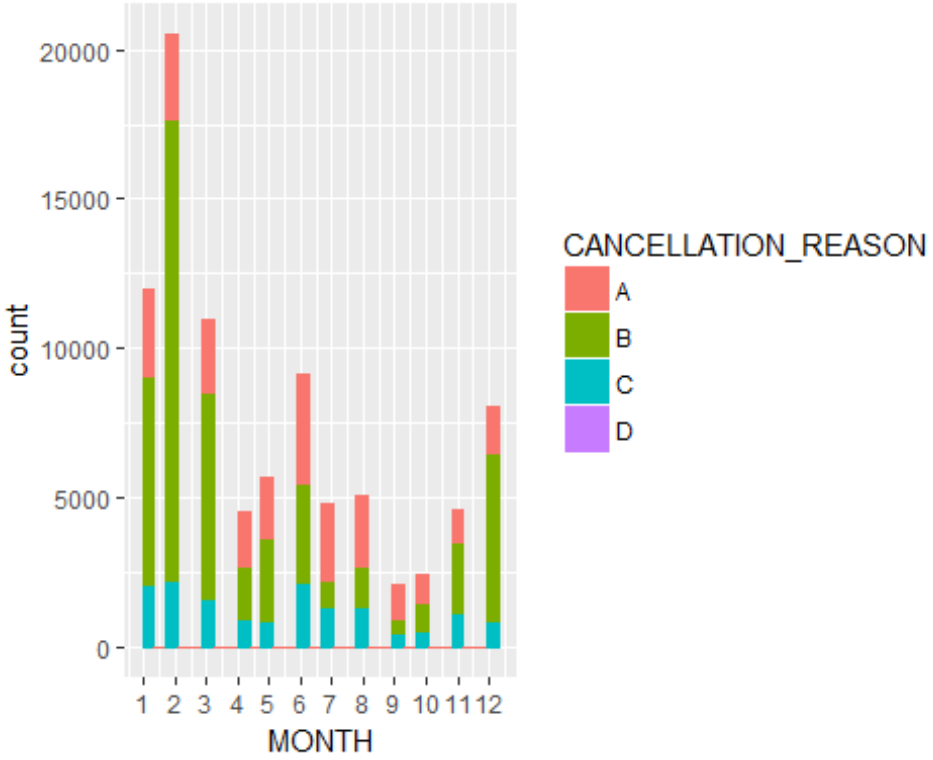


이것으로 볼 때 각 달마다 전체 항공횟수가 많이 차이가 나지 않았기 때문에 전체비율과 비교해서 보더라도 1,2,3 월 달이 취소된 비율이 많고 2 월이 특히나 많다는 것을 확신할 수 있다.

각각의 취소 원인이 월 별로 비율을 얼마나 차지하는지 그림을 통해 알아보도록 하겠다.

```
ggplot(cancelled, aes(MONTH,color=CANCELLATION_REASON,fill=CANCELLATION_REASO
N)) +geom_histogram()+scale_x_continuous(breaks=c(1:12))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

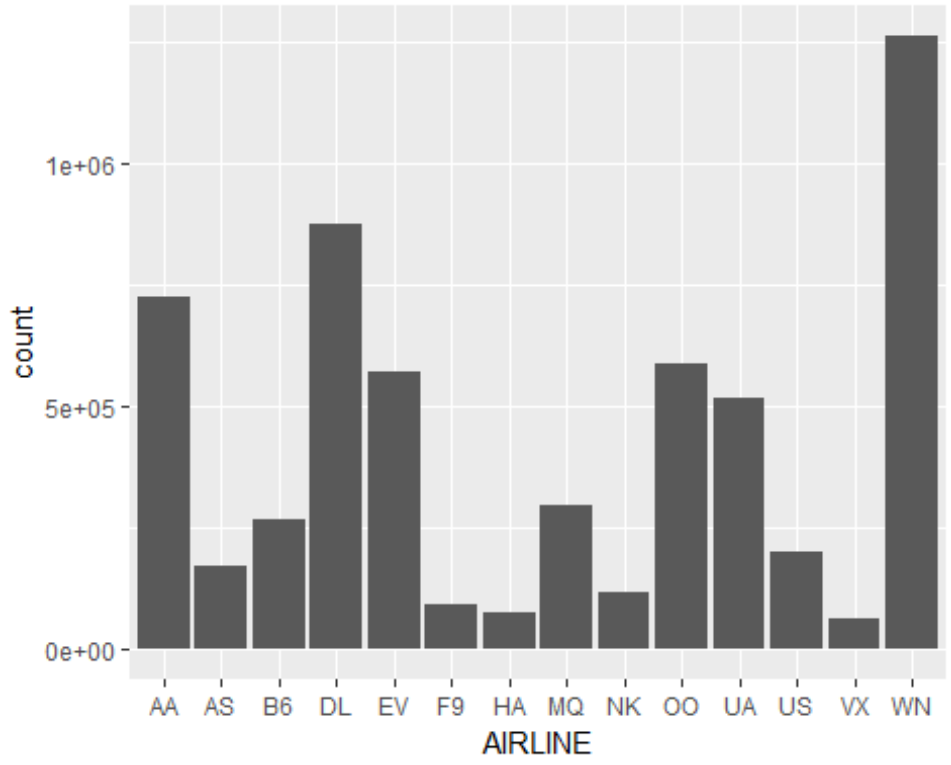


또한 1 월, 2 월, 3 월달에 날씨상의 이유로 많은 비행기가 취소됐다는 것을 알 수 있다. 특히 2 월에 비행취소가 많은 이유가 특히나 날씨 때문에 일어난 것을 알 수 있었다.

##2) 항공사에 따른 취소 비행

전체 공항에 따른 비행의 횟수를 먼저 알아보도록 하겠다.

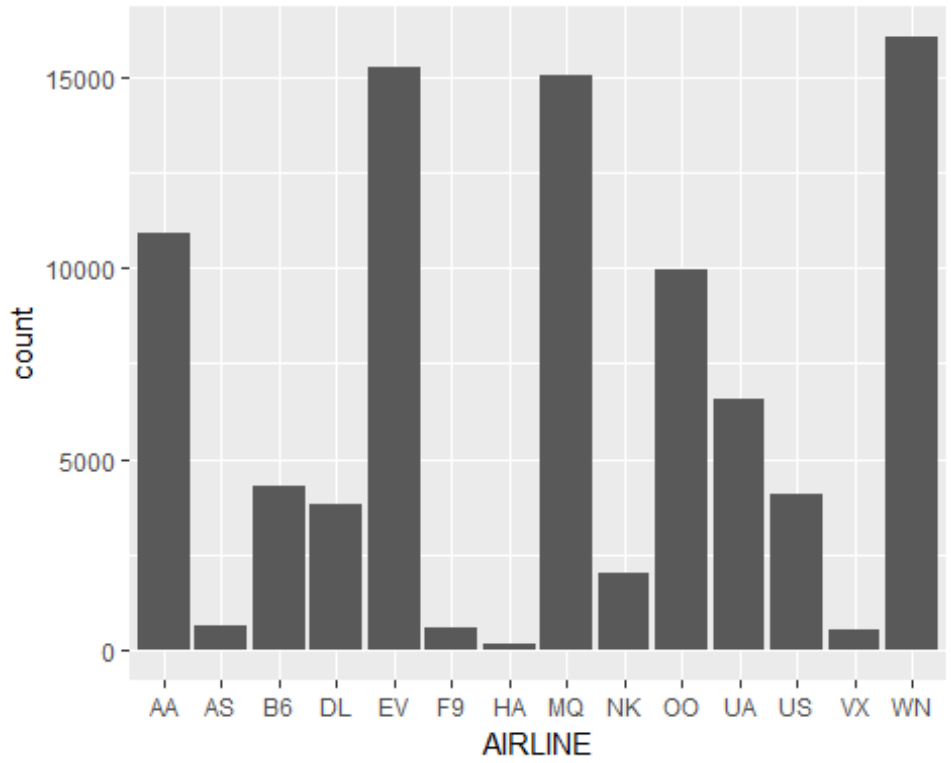
```
all_percentage1 <- flights1 %>% group_by(AIRLINE) %>% summarise(count1=n())
ggplot(flights1,aes(AIRLINE))+geom_bar()
```



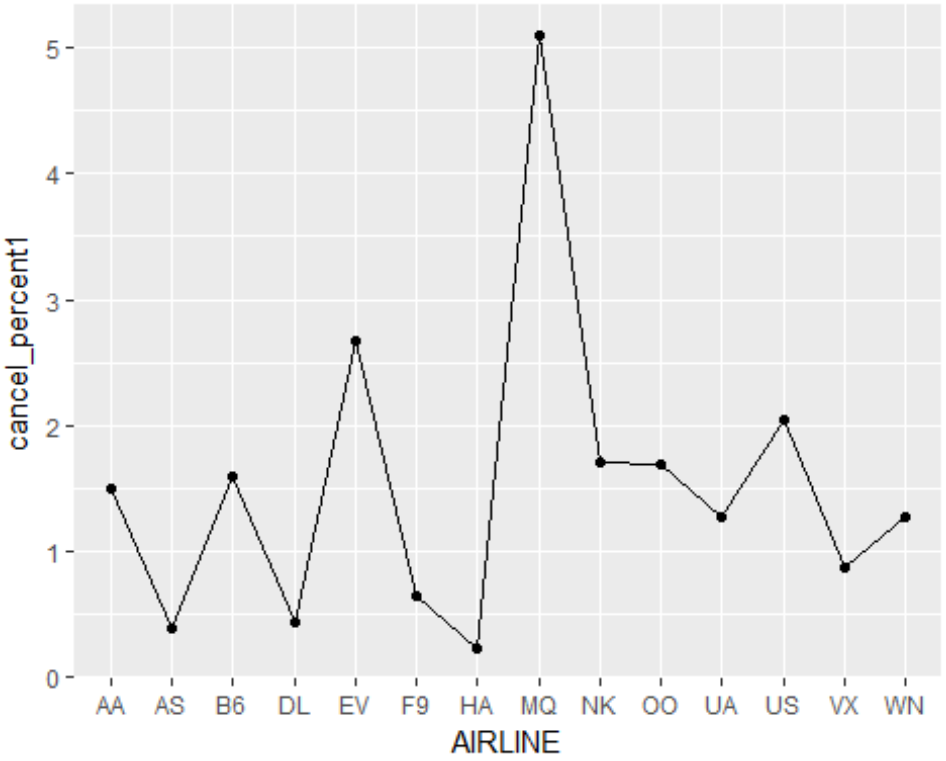
WN 의 공항에서 항공횟수가 가장 많고 그 뒤를 DL 회사가 따른다. 이제 공항에 따른 취소비행을 알아보도록 하겠다

#공항에 따른 취소비행.

```
ggplot(cancelled,aes(AIRLINE))+geom_bar()
```




```
  cancel_percentage1 <- (cancelled %>% group_by(AIRLINE)) %>% summarise(count
2=n())
percentage1 <- all_percentage1 %>% left_join(cancel_percentage1,by="AIRLINE")
 %>%
mutate(cancel_percent1=count2/count1*100) %>% arrange(desc(cancel_percent1))
ggplot(percentage1,aes(AIRLINE,cancel_percent1))+geom_point()+geom_line(group
="AIRLINE")
```



첫 번째 그래프를 통해 보았을 때 놀라운 결과를 확인할 수 있다. WN 회사와 MQ 회사는 약 4 배정도 전체 항공횟수가 차이가 나지만, MQ 의 취소 하는 개수가 15025 개고 WN 의 개수는 16043 개로 약 1000 개 밖에 차이가 나지 않는다는 것을 확인할 수 있다.

두 번째 그래프를 통해서 MQ 의 항공사가 5.0995819%로 월등하게 가장 높은 취소비율을 나타내고 있다는 것을 알 수 있다. MQ 항공사는 13 개의 항공사 중 7 위의 항공횟수를 자랑했는데 가장 큰 비율로 취소 중이다. 그에 반해 항공횟수가 가장 많은 WN 항공사는 생각보다 많은 취소 비율을 보이고 있지 않았다.

이를 모두 통합해서 볼 때 MQ 항공사의 취소비율은 실로 놀라운 것이라 볼 수 있다. 취소비율 2 위를 차지한 EV 항공사도 취소횟수가 15231 개로 엄청난 횟수를 자랑한다는 것을 알 수 있다.

그렇다면 각 항공사가 왜 취소하는지를 먼저 그래프를 통해 알아보도록 하겠다.

```
ggplot(cancelled,aes(AIRLINE,color=CANCELLATION_REASON,fill=CANCELLATION_REAS
ON))+geom_bar()
```

WN 항공사는 A, B 의 이유로 취소된 횟수가 많았다.

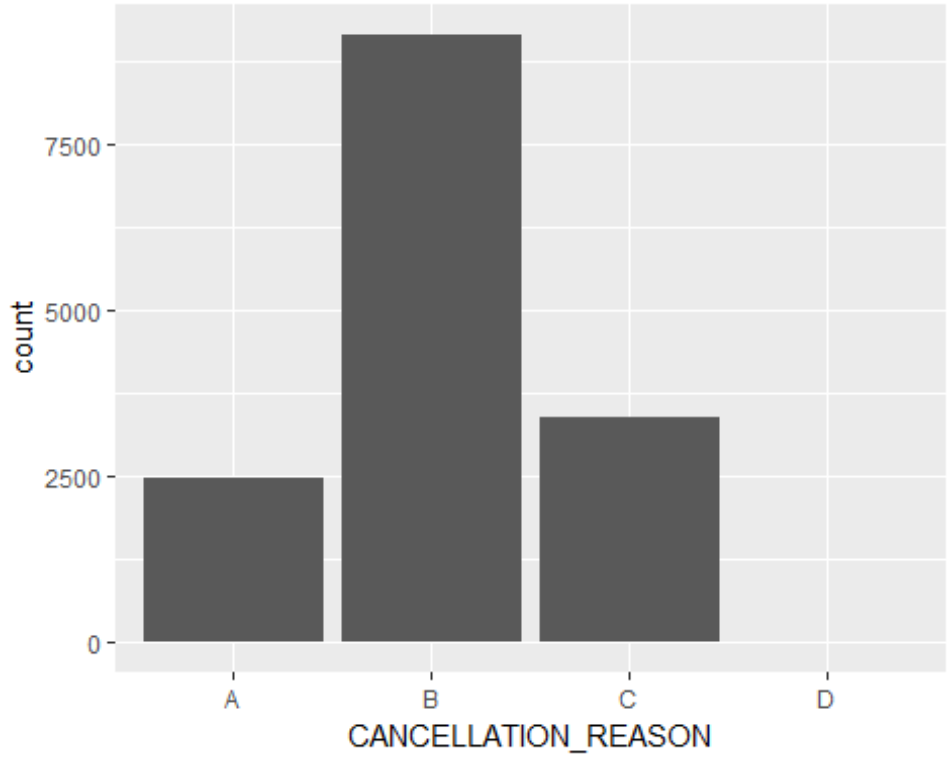
- **MQ 와 EV 항공사의 취소 이유**

우린 문제의 MQ 항공사와 EV 항공사를 더 알아보도록 하겠다.

```
MQ <- filter(cancelled,AIRLINE=="MQ")
MQ %>% group_by(CANCELLATION_REASON) %>% summarise(count=n(),percent=count/(2
475+9164+3385+1)*100)
```

```
## # A tibble: 4 x 3
##   CANCELLATION_REASON count      percent
##                 <fctr> <int>        <dbl>
## 1                    A  2475 16.472545757
## 2                    B  9164 60.991680532
## 3                    C  3385 22.529118136
## 4                    D     1  0.006655574
```

```
  ggplot(MQ,aes(CANCELLATION_REASON))+geom_bar()
```

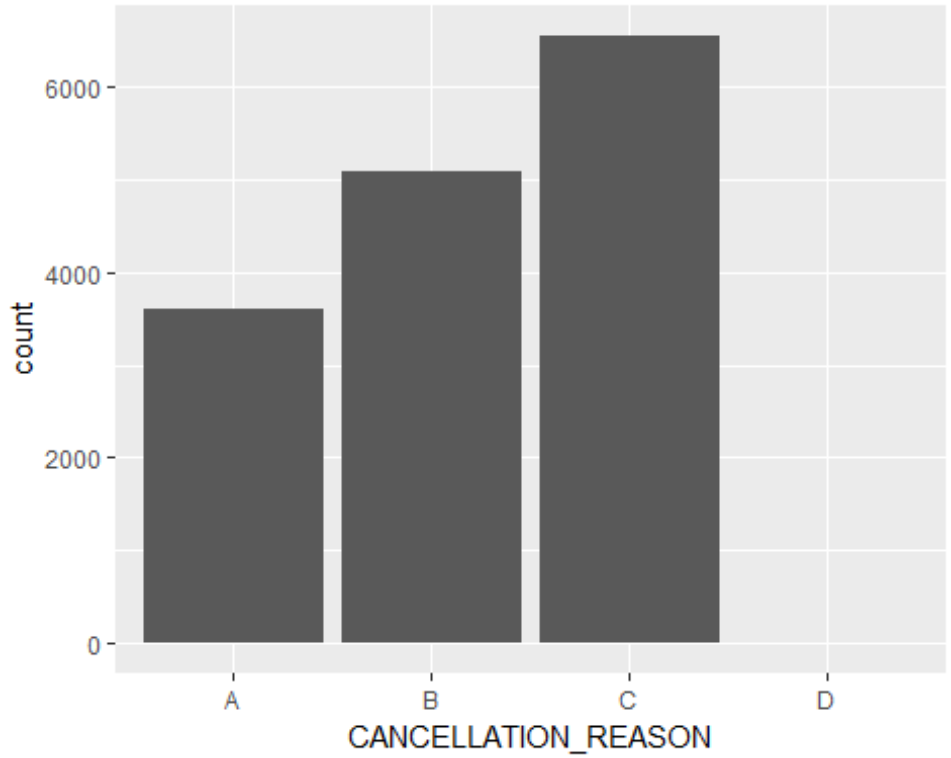


MQ 는 B>C>A 의 순으로 점점 줄어드는 추세를 보여주고 있다. 원래 모든 항공사에서는 B>A>C 순이므로 이 항공사에서 C 의 이유가 생각보다 많이 발생했다는 것을 알 수 있ㄷ.

```
EV <- filter(cancelled,AIRLINE=="EV")
EV %>% group_by(CANCELLATION_REASON) %>% summarise(count=n(),percent=count/(2
475+9164+3385+1)*100)
```

```
## # A tibble: 4 x 3
##   CANCELLATION_REASON count      percent
##                 <fctr> <int>        <dbl>
## 1                    A  3604 23.986688852
## 2                    B  5082 33.823627288
## 3                    C  6544 43.554076539
## 4                    D     1  0.006655574
```

```
ggplot(EV,aes(CANCELLATION_REASON))+geom_bar()
```



EV 는 C 의 이유가 가장 많이 도출되었다. 이는 매우 신기한 결과이다. 원래 이유는 B>A>C 순서인데 EV 는 확실히 C 인 NATIONAL AIR SYSTEM 의 문제가 크다는 것을 알게 되었다.

3. 취소되지 않은 2015 년 비행의 특징 살펴보기

지금까지 취소된 비행기의 특성을 살펴봤으므로 취소되지 않은 2015 년 비행의 특성을 다양한 방법으로 나눠서 분석해 보도록 하겠다.

```
not_cancelled <- flights1 %>% filter(CANCELLED=="0")

head(not_cancelled)
```

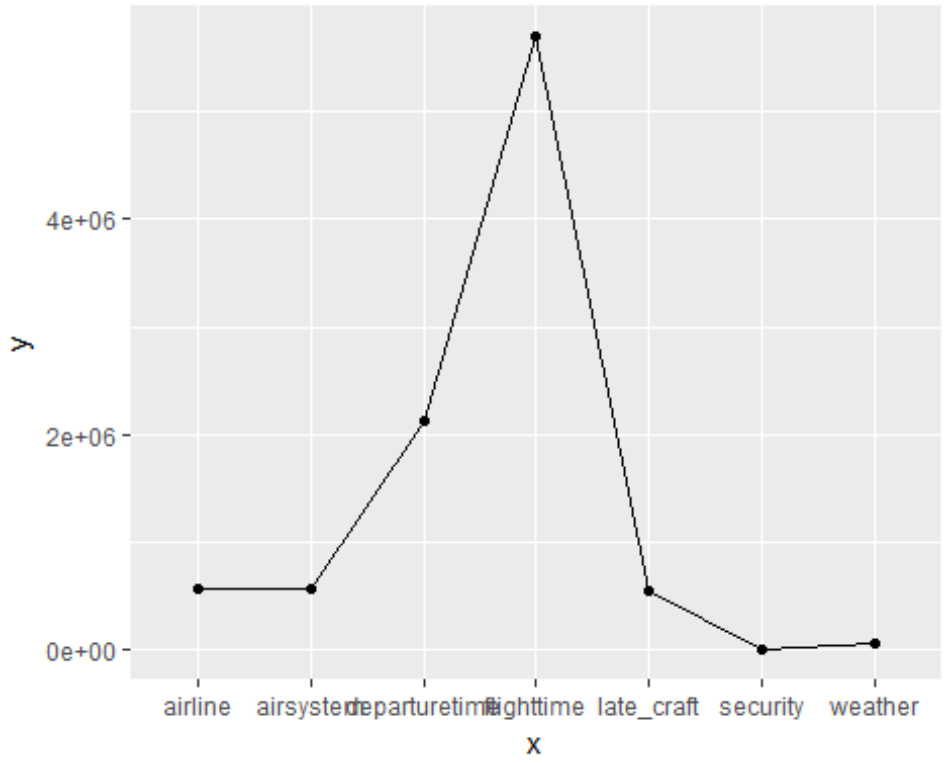
```
# A tibble: 6 x 31
   YEAR MONTH   DAY DAY_OF_WEEK AIRLINE FLIGHT_NUMBER TAIL_NUMBER
ORIGIN_AIRPORT DESTINATION_AIRPORT SCHEDULED_DEPARTURE
  <int> <int> <int>       <int>  <fctr>         <int>      <fctr>
<fctr>              <fctr>               <int>
1  2015     1     1           4      AS            98      N407AS
ANC                 SEA                   5
2  2015     1     1           4      AA          2336      N3KUAA
LAX                 PBI                  10
3  2015     1     1           4      US           840      N171US
SFO                 CLT                  20
4  2015     1     1           4      AA           258      N3HYAA
LAX                 MIA                  20
5  2015     1     1           4      AS           135      N527AS
SEA                 ANC                  25
6  2015     1     1           4      DL           806      N3730B
SFO                 MSP                  25
# ... with 21 more variables: DEPARTURE_TIME <int>, DEPARTURE_DELAY <int>,
TAXI_OUT <int>, WHEELS_OFF <int>,
#   SCHEDULED_TIME <int>, ELAPSED_TIME <int>, AIR_TIME <int>, DISTANCE <int>,
WHEELS_ON <int>, TAXI_IN <int>,
#   SCHEDULED_ARRIVAL <int>, ARRIVAL_TIME <int>, ARRIVAL_DELAY <int>,
DIVERTED <int>, CANCELLED <int>,
#   CANCELLATION_REASON <fctr>, AIR_SYSTEM_DELAY <int>, SECURITY_DELAY <int>,
AIRLINE_DELAY <int>, LATE_AIRCRAFT_DELAY <int>,
#   WEATHER_DELAY <int>
```

#1) 지연 살펴보기

취소되지 않은 지연 중에 무슨 지연이 가장 횟수가 많은지 살펴보도록 하겠다. 지연에는 출발지연, 도착지연, air system 으로 인한 지연, security 로 인한 지연, airline 으로 인한 지연, late aircraft 로 인한 지연, 날씨로 인한 지연으로 8 개가 존재한다. 무슨 지연이 가장 많이 발생되었는지 살펴보도록 하겠다.

그런데 도착지연은 출발지연의 영향을 받아서 지연된 것이 있을 수 있으므로 예정비행시간보다 실제비행시간이 긴 지연으로 바꾸도록 하겠다.

```
departuretime <- not_cancelled %>% filter(DEPARTURE_DELAY>0)
flighttime <- not_cancelled %>% filter(DEPARTURE_TIME>SCHEDULED_TIME)
airsystem <- not_cancelled %>% filter(AIR_SYSTEM_DELAY>0)
security <- not_cancelled %>% filter(SECURITY_DELAY>0)
airline <- not_cancelled %>% filter(AIRLINE_DELAY>0)
late_aircraft <- not_cancelled %>% filter(LATE_AIRCRAFT_DELAY>0)
weather <- not_cancelled %>% filter(WEATHER_DELAY>0)
delay <- data.frame(x=c("departuretime","flighttime","airsystem","security","
airline","late_craft","weather"),y=c(nrow(departuretime),nrow(flighttime),nro
w(airsystem),nrow(security),nrow(airline),nrow(late_aircraft),nrow(weather)))
ggplot(delay,aes(x,y))+geom_point()+geom_line(group="x")
```



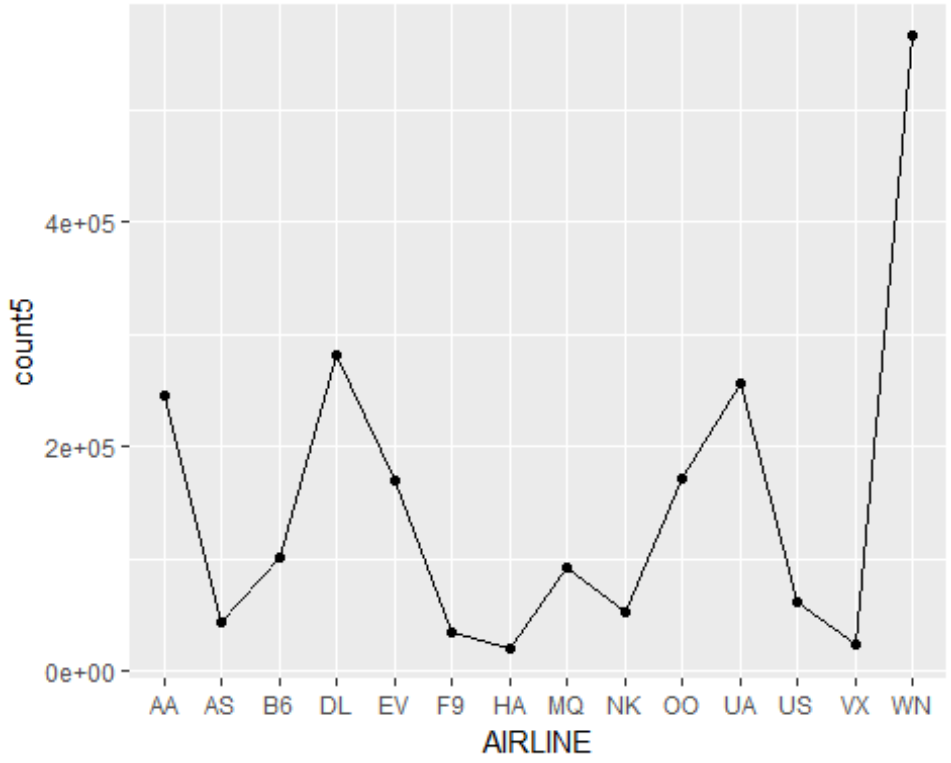
이것으로 운행시간에서 운행에서 지연이 가장 많다는 것을 알게 되었다. 그 뒤로는 출발시간의 지연이 뒤를 따랐다. 그래서 이 두 가지를 좀 더 살펴보도록 하겠다.

먼저 출발시간이 가장 늦는 공항이 어딘지 알아보도록 하였다.

```
departuredelay <- departuretime %>% group_by(AIRLINE) %>% summarise(count5=n
()) %>% arrange(desc(count5))
departuredelay
```

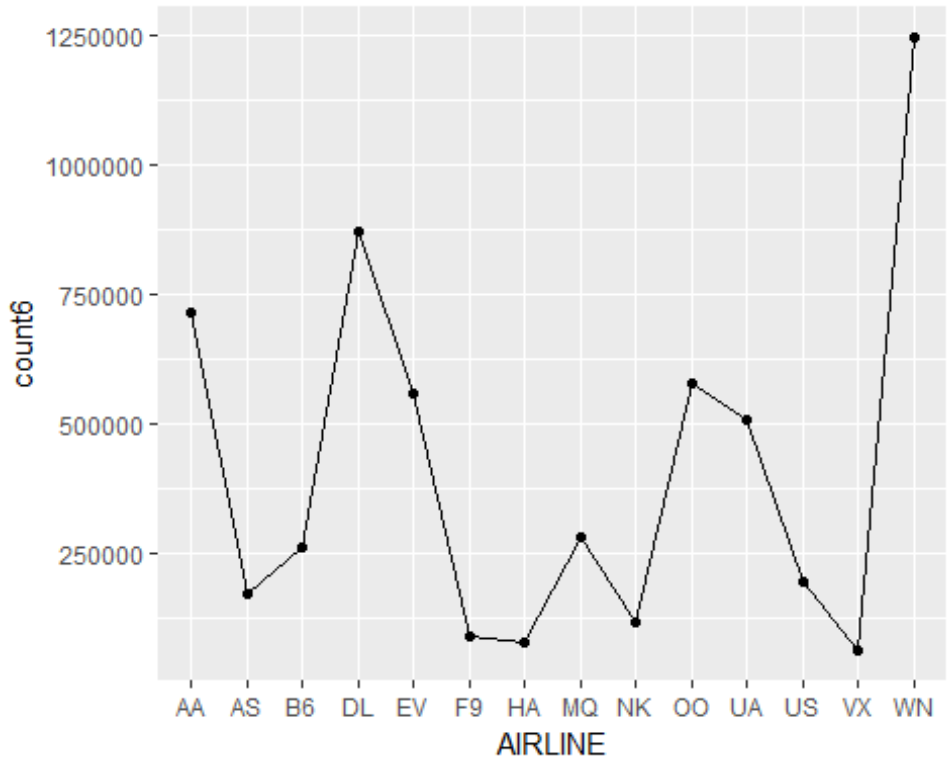
```
## # A tibble: 14 x 2
##    AIRLINE count5
##     <fctr>  <int>
##  1      WN 566583
##  2      DL 282385
##  3      UA 256241
##  4      AA 245550
##  5      OO 171181
##  6      EV 169503
##  7      B6 102012
##  8      MQ  93232
##  9      US  62452
## 10      NK  52033
## 11      AS  43541
## 12      F9  34859
## 13      VX  23366
## 14      HA  20140
```

```
ggplot(departuredelay,aes(AIRLINE,count5))+geom_line(group="AIRLINE")+geom_po
int()
```



WN 항공사가 가장 delay 가 많다는 것을 알 수 있었다. 그 뒤로는 DL 항공사가 따르는데 무려 약 400000 번의 차이로 WN 이 압승을 거뒀다. 그런데 사실 다시 취소되지 않고 운행한 전체적인 횟수를 따져볼 때와 다르다. 다음은 전체적인 그래프이다.

```
notcancelleddelay <-not_cancelled %>% group_by(AIRLINE) %>% summarise(count6=
n()) %>% arrange(desc(count6))
ggplot(notcancelleddelay,aes(AIRLINE,count6))+geom_point()+geom_line(group="M
ONTH")
```

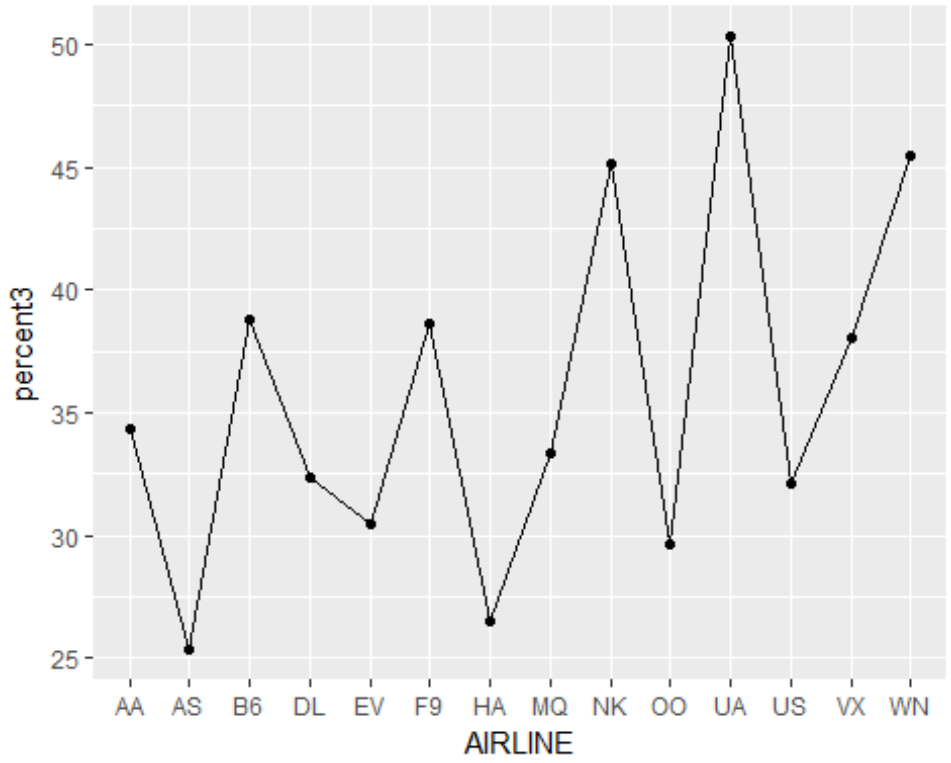


둘이 모양 매우 비슷하다는 것을 알 수 있다. 특출한 변화는 보이지 않아 더 자세히 보기 위해 수치로 나타내보겠다.

```
percentage_departure_delay <- notcancelleddelay %>% left_join(departuredelay,
by="AIRLINE") %>% mutate(percent3=count5/count6*100)
percentage_departure_delay %>% arrange(desc(percent3))
```

```
## # A tibble: 14 x 4
##    AIRLINE  count6 count5 percent3
##     <fctr>   <int>  <int>    <dbl>
##  1      UA  509150 256241 50.32721
##  2      WN 1245812 566583 45.47901
##  3      NK  115375  52033 45.09902
##  4      B6  262772 102012 38.82149
##  5      F9   90248  34859 38.62579
##  6      VX   61369  23366 38.07460
##  7      AA  715065 245550 34.33954
##  8      MQ  279607  93232 33.34394
##  9      DL  872057 282385 32.38148
## 10      US  194648  62452 32.08458
## 11      EV  556746 169503 30.44530
## 12      OO  578393 171181 29.59597
## 13      HA   76101  20140 26.46483
## 14      AS  171852  43541 25.33634
```

```
  ggplot(percentage_departure_delay,aes(AIRLINE,percent3))+geom_line(group="A
IRLINE")+geom_point()
```



이것을 볼 때 가장 지연이 많은 항공사는 UA 이고 가장 비율이 적은 항공사는 AS 라는 것을 확인할 수 있다.

##2. 시기별 비행횟수

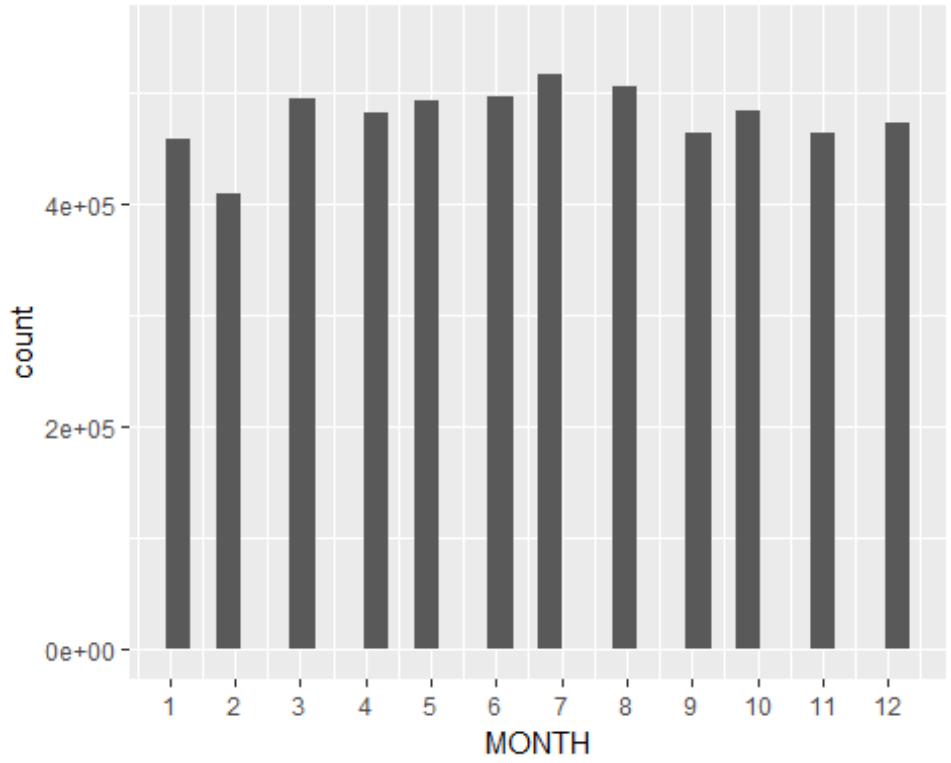
시기에 따라 언제가 가장 비행횟수가 많은지를 알아보도록 하겠다. 2015 년의 비행이므로 연은 계산하지 않고 또한 날도 많은 정보를 포함할 것 같지 않아 생략하도록 하겠다.

1) month

달에 따라 비행횟수를 히스토그램을 이용해 비교해보도록 하겠다.

```
ggplot(not_cancelled, aes(MONTH)) +geom_histogram()+scale_x_continuous(breaks
=c(1:12))+scale_y_continuous(limits=c(0,550000))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



그림에서 볼 수 있다시피 1 월, 2 월의 항공횟수가 다른 달들보다 떨어지는 것을 알 수 있다. 특히 7 월에서 가장 많은 횟수를 보이는 것으로 관측되었다. 2 월 달에 많은 비행을 하지 못한 것을 빼면 월 별로는 그렇게 크게 차이가 나지는 않는다. 그리고 우리가 취소된 비행을 살펴보던 중 2 월에 특히 날씨 때문에 취소가 많이 된 것을 보아 2 월에 비행의 횟수가 낮은 것이 그러한 이유가 영향을 끼쳤을 것이라고 볼 수 있다. 월 별로 항공횟수를 정확한 수치로 나타내보았다.

```
not_cancelled <- filter(flights1,CANCELLED=="0")
cancelled %>%
 group_by(MONTH) %>%
 summarise(count=n())
```

```
## # A tibble: 12 x 2
##    MONTH count
##    <int> <int>
##  1     1 11982
##  2     2 20517
##  3     3 11002
##  4     4  4520
##  5     5  5694
##  6     6  9120
##  7     7  4806
##  8     8  5052
##  9     9  2075
## 10    10  2454
## 11    11  4599
## 12    12  8063
```

그림에서 확인했던 것처럼 7 월달이 515912 번으로 가장 많이 운행했고 2 월이 208674 번으로 가장 조금 운행했다는 것을 확인할 수 있다.

2) DAY_OF_WEEK

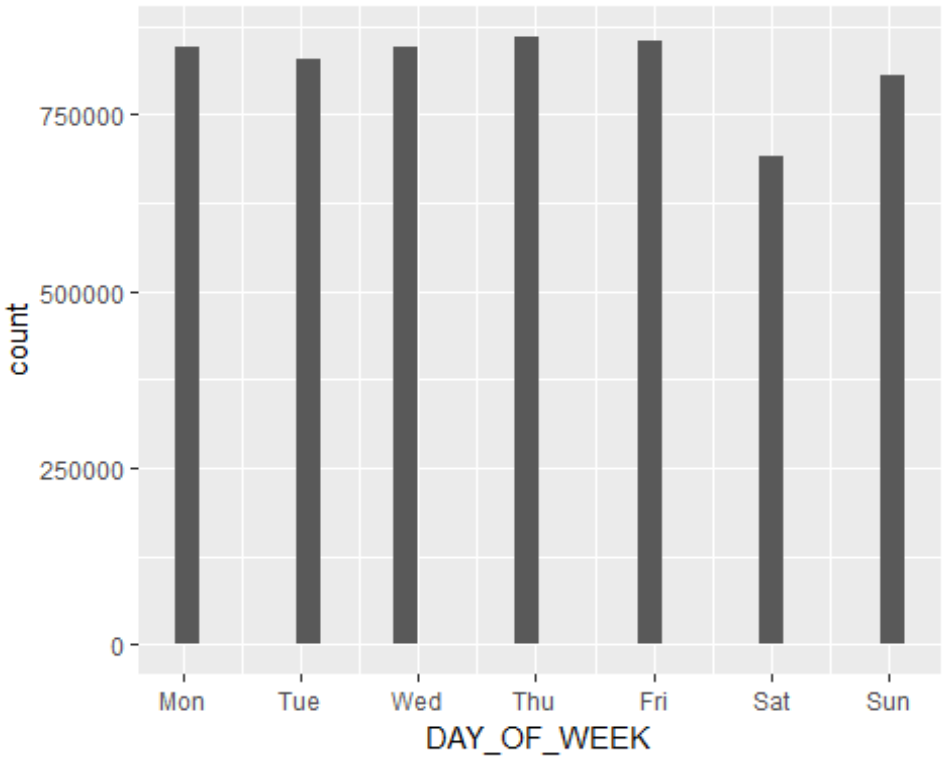
요일 별로 운행횟수를 알아보도록 하겠다.

```
not_cancelled %>%
  group_by(DAY_OF_WEEK) %>%
  summarise(count=n())
```

```
## # A tibble: 7 x 2
##   DAY_OF_WEEK  count
##         <int>  <int>
## 1           1 844470
## 2           2 829528
## 3           3 845168
## 4           4 860230
## 5           5 853404
## 6           6 691796
## 7           7 804599
```

```
ggplot(not_cancelled,aes(DAY_OF_WEEK))+geom_histogram()+scale_x_continuous(br
eaks=c(1:7),labels=c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

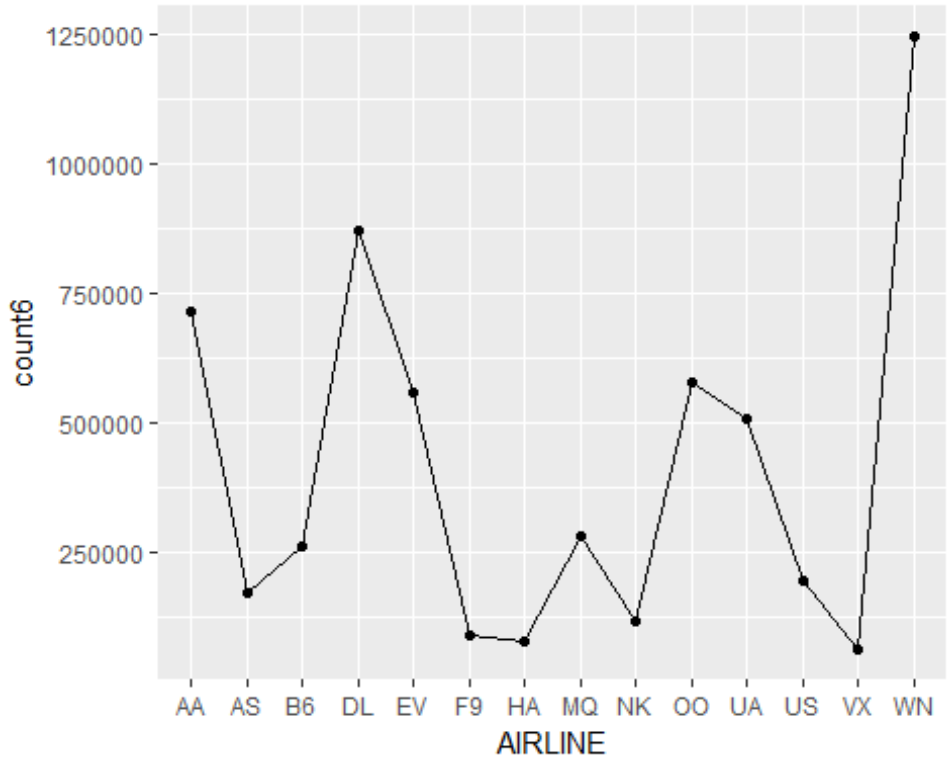


토요일이 가장 적은 횟수를 보이고 있었다. 이는 다들 휴일 날 쉬기 위함으로 추측할 수 있다.

#3) 각종 변수를 이용한 비행 분석

앞서 이미 보았지만 먼저 항공사부터 알아보겠다. 항공사로 나누어 실제 비행횟수에 대한 자료를 분석해보도록 하겠다.

```
notcancelleddelay <-not_cancelled %>% group_by(AIRLINE) %>% summarise(count6=
n()) %>% arrange(desc(count6))
ggplot(notcancelleddelay,aes(AIRLINE,count6))+geom_point()+geom_line(group="M
ONTH")
```



WN 항공사가 가장 많이 비행하고 그 뒤로 DL 이 따르고 가장 적은 항공사는 VX 라는 것을 확인하였다.

두번째로 비행을 가장 많이 한 비행기를 찾아보도록 하겠다. 비행기는 너무 많으므로 가장 비행횟수가 많은 비행기 하나와 가장 적은 비행기 하나를 알아보도록 하겠다.

```
not_cancelled %>% group_by(TAIL_NUMBER) %>% summarise(count=n()) %>% arrange
(count) %>% head()
```

```
## # A tibble: 6 x 2
##   TAIL_NUMBER count
##         <fctr> <int>
## 1      N121UA     1
## 2      N175UA     1
## 3      N180UA     1
```

```
## 4      N7LBAA     1
## 5      N7LEAA     1
## 6      N840MH     1
```

```
not_cancelled %>% group_by(TAIL_NUMBER) %>% summarise(count=n()) %>% arrange
(desc(count)) %>% head()
```

```
## # A tibble: 6 x 2
##   TAIL_NUMBER count
##         <fctr> <int>
## 1      N480HA  3766
## 2      N484HA  3723
## 3      N488HA  3721
## 4      N493HA  3584
## 5      N478HA  3577
## 6      N483HA  3528
```

첫 번째 결과에서 볼 때 N121UA, N175UA, N180UA, N7LBAA, N7LEAA, N840NW, N852NW, N860NW 비행기 모두 1 번 비행으로 가장 낮은 비행횟수를 달성했다. 그리고 두 번째 결과에서 볼 때 N480HA 가 3766 번으로 가장 많이 비행했고 N484HA 가 3723 번으로 뒤를 이었다.

마지막으로 공항 별로 나누어 확인하도록 하겠다.

```
not_cancelled %>% group_by(ORIGIN_AIRPORT) %>% summarise(count=n()) %>% arran
ge(count) %>% head()
```

```
## # A tibble: 6 x 2
##   ORIGIN_AIRPORT count
##           <fctr> <int>
## 1          11503     4
## 2          13502     6
## 3          10165     9
## 4          14222     9
## 5          13541    11
## 6          14025    13
```

```
not_cancelled %>% group_by(ORIGIN_AIRPORT) %>% summarise(count=n()) %>% arran
ge(desc(count)) %>% head()
```

```
## # A tibble: 6 x 2
##   ORIGIN_AIRPORT  count
##           <fctr>  <int>
## 1            ATL 344279
## 2            ORD 277336
## 3            DFW 233297
## 4            DEN 193932
```

```
## 5             LAX 192509
## 6             PHX 145913
```

첫 번째 결과에서 11503 이라는 공항이 4 번으로 가장 적었다. 아마 개인 여객기를 이용하는 것으로 추측할 수 있다. 그리고 두 번째 결과에서 볼 때 ATL 이라는 공항이 344279 번으로 가장 많이 출발지로 이용되었다는 것을 알 수 있다.

결론

2015년 자료는 실제 비행이 이루어진 것과 비행이 아예 취소되어 가지 못하는 경우가 있어 크게 2개, '취소된 비행기' 그리고 '취소되지 않은 비행기'로 나누어 보았다. 첫째로 '취소된 비행기'는 시기별, 항공사별 2가지로 나누어 특성을 살펴보고 각각 취소된 이유를 2가지에 적용해 비교해서 분석해보았다. 두 번째로 '취소되지 않은 비행기'는 지연에 대해 알아보기도 하고 시기, 항공사, 비행기, 출발공항 별 비행횟수에 대해서 알아보았다.

취소된 비행의 개수는 89884개이고 최소되지 않은 비행의 개수는 5729195개이다. 그 중 취소된 비행기에서 취소원인의 개수는 B, A, C 단계 순으로 적어진다. 먼저 취소된 비행을 시기별로 나누어 살펴보면 1,2,3월에 가장 취소비행횟수가 많고 취소 비율 또한 높다는 것을 알 수 있다. 특히 2월에는 월등히 높은 취소횟수를 보여주고 있는데 취소원인 B인 날씨가 영향을 많이 끼치고 있다는 것을 알 수 있었다. 두 번째로 공항 별로 취소비행을 분석해보았더니 놀라운 결과를 확인할 수 있다. MQ항공사의 취소비율이 가장 크고 EV의 항공사가 뒤를 따랐다. 가장 많은 비행을 제공하는 WN항공사는 MQ항공사보다 약 4배정도 전체 비행횟수가 많지만, MQ의 취소 비행횟수가 15025개고 WN의 취소 비행횟수는 16043개로 약 1000번 밖에 차이가 나지 않는 것을 확인할 수 있다. 또한 취소이유를 대입하여 볼 때 MQ회사는 취소이유인 C의 비율이 상대적으로 높은 것이 확인되었다. 또한 2위를 차지했던 EV항공사는 취소이유인 C의 비율이 월등히 높다는 것이 관측되었다.

그리고 취소되지 않고 정상적으로 운행한 비행에 대해서 살펴보았다. 첫째로 지연의 많은 이유 중 운행시간과 출발시간의 지연이 가장 많은 지연을 도출해냈다. 이 중 출발시간지연에서 항공사별 지연되는 비율을 조사했더니 UA항공사가 가장 많은 지연을 했다는 것을 알아냈다. 둘째로 시기별로 비행횟수에 대해 알아보았다. 뚜렷한 차이는 없지만 2월이 가장 적은 횟수로 비행을 했다는 것이 관측되었고 이는 앞서 월별 취소원인에서 봤다시피 2월달 날씨가 좋지 않아 대량으로 비행이 취소돼서 이러한 결과가 도출되었던 것으로 예상되었다. 월 별 뿐 아니라 일 별로도 비행횟수를 알아보았는데, 토요일이 가장 적은 횟수의 운행을 했다. 이는 휴일에 쉬기 위한 사람들의 일정에 영향을 받는 것으로 예상된다. 마지막으로 각종 변수를 이용해 비행횟수를 알아보았다. 비행횟수는 WN 항공사가 가장 많고 그 뒤로 DL이 따르고, 가장 적은 항공사는 VX라는 것을 확인하였다. 그리고 비행기는 번호 N480HA가 3766번으로 가장 많이 비행을 한 것으로 나왔다. 또한 ATL이라는 공항이 344279번으로 가장 많이 출발지로 이용되었다는 사실을 알 수 있었다.