# Untitled

Xidan Kou

1/20/2022

```
library(nycflights13)
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ dplyr   1.0.2
## ✓ tibble  3.0.3     ✓ stringr 1.4.0
## ✓ tidyr   1.1.2     ✓ forcats 0.5.0
## ✓ purrr   0.3.4
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## Loading required package: effects
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(dplyr)
library(effects)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(grid)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(rlang)
```

```
##
## Attaching package: 'rlang'
```

```
## The following objects are masked from 'package:purrr':
##
##     %@%, as_function, flatten, flatten_chr, flatten_dbl, flatten_int,
##     flatten_lgl, flatten_raw, invoke, list_along, modify, prepend,
##     splice
```

```r
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-2
```

```r
library(ggplot2)
```

```r
flights = data.frame(flights)
```

```r
flights_new = flights
flights = na.omit(flights)
flights_new = flights
# recognized as delay when delay is greater than 15
flights = subset(flights,dep_delay>=15)
flights$delayed = ifelse(flights$dep_delay >=15, "Delayed","Not Delayed" )
```

```r
# Using outside data
weather = data.frame(weather)
total = merge(flights,weather,by =c("time_hour", "origin"))

# total = na.omit(total)
summary(total)
```

```
##    time_hour                       origin             year.x    
##  Min.   :2013-01-01 06:00:00   Length:72051       Min.   :2013  
##  1st Qu.:2013-04-11 06:00:00   Class :character   1st Qu.:2013  
##  Median :2013-06-27 17:00:00   Mode  :character   Median :2013  
##  Mean   :2013-06-30 07:27:18                      Mean   :2013  
##  3rd Qu.:2013-09-13 13:00:00                      3rd Qu.:2013  
##  Max.   :2013-12-30 18:00:00                      Max.   :2013  
##                                                                 
##     month.x          day.x          dep_time    sched_dep_time
##  Min.   : 1.00   Min.   : 1.00   Min.   :   1   Min.   : 500  
##  1st Qu.: 4.00   1st Qu.: 9.00   1st Qu.:1337   1st Qu.:1300  
##  Median : 6.00   Median :16.00   Median :1716   Median :1620  
##  Mean   : 6.45   Mean   :15.78   Mean   :1619   Mean   :1548  
##  3rd Qu.: 9.00   3rd Qu.:23.00   3rd Qu.:1956   3rd Qu.:1900  
##  Max.   :12.00   Max.   :31.00   Max.   :2400   Max.   :2359  
##                                                               
##    dep_delay          arr_time    sched_arr_time   arr_delay    
##  Min.   :  15.00   Min.   :   1   Min.   :   1   Min.   : -62.00  
##  1st Qu.:  25.00   1st Qu.:1320   1st Qu.:1428   1st Qu.:  19.00  
##  Median :  44.00   Median :1822   Median :1820   Median :  43.00  
##  Mean   :  64.66   Mean   :1617   Mean   :1707   Mean   :  60.89  
##  3rd Qu.:  82.00   3rd Qu.:2117   3rd Qu.:2100   3rd Qu.:  84.00  
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.00  
##                                                                   
##    carrier              flight       tailnum             dest          
##  Length:72051       Min.   :   1   Length:72051       Length:72051      
##  Class :character   1st Qu.: 587   Class :character   Class :character  
##  Mode  :character   Median :1611   Mode  :character   Mode  :character  
##                     Mean   :2154                                        
##                     3rd Qu.:3836                                        
##                     Max.   :8500                                        
##                                                                         
##     air_time        distance        hour.x         minute     
##  Min.   : 20.0   Min.   :  80   Min.   : 5.0   Min.   : 0.00  
##  1st Qu.: 80.0   1st Qu.: 488   1st Qu.:13.0   1st Qu.:10.00  
##  Median :124.0   Median : 764   Median :16.0   Median :29.00  
##  Mean   :145.1   Mean   :1005   Mean   :15.2   Mean   :27.59  
##  3rd Qu.:181.0   3rd Qu.:1325   3rd Qu.:19.0   3rd Qu.:45.00  
##  Max.   :666.0   Max.   :4983   Max.   :23.0   Max.   :59.00  
##                                                               
##    delayed              year.y        month.y          day.y      
##  Length:72051       Min.   :2013   Min.   : 1.00   Min.   : 1.00  
##  Class :character   1st Qu.:2013   1st Qu.: 4.00   1st Qu.: 9.00  
##  Mode  :character   Median :2013   Median : 6.00   Median :16.00  
##                     Mean   :2013   Mean   : 6.45   Mean   :15.78  
##                     3rd Qu.:2013   3rd Qu.: 9.00   3rd Qu.:23.00  
##                     Max.   :2013   Max.   :12.00   Max.   :31.00  
##                                                                   
##      hour.y          temp             dewp           humid       
##  Min.   : 5.0   Min.   : 10.94   Min.   :-9.04   Min.   : 12.74  
##  1st Qu.:13.0   1st Qu.: 42.80   1st Qu.:28.04   1st Qu.: 45.90  
##  Median :16.0   Median : 60.08   Median :46.94   Median : 62.40  
##  Mean   :15.2   Mean   : 58.70   Mean   :44.52   Mean   : 62.78  
##  3rd Qu.:19.0   3rd Qu.: 75.02   3rd Qu.:62.06   3rd Qu.: 81.14  
##  Max.   :23.0   Max.   :100.04   Max.   :78.08   Max.   :100.00  
##                 NA's   :10       NA's   :10      NA's   :10      
##     wind_dir       wind_speed       wind_gust         precip        
##  Min.   :  0.0   Min.   : 0.000   Min.   :16.11   Min.   :0.000000  
##  1st Qu.:130.0   1st Qu.: 8.055   1st Qu.:20.71   1st Qu.:0.000000  
##  Median :210.0   Median :11.508   Median :24.17   Median :0.000000  
##  Mean   :200.2   Mean   :11.657   Mean   :25.59   Mean   :0.008623  
##  3rd Qu.:280.0   3rd Qu.:14.960   3rd Qu.:28.77   3rd Qu.:0.000000  
##  Max.   :360.0   Max.   :42.579   Max.   :66.75   Max.   :1.210000  
##  NA's   :1779    NA's   :9        NA's   :53635                     
##     pressure         visib       
##  Min.   : 985    Min.   : 0.000  
##  1st Qu.:1011    1st Qu.:10.000  
##  Median :1016    Median :10.000  
##  Mean   :1016    Mean   : 8.986  
##  3rd Qu.:1021    3rd Qu.:10.000  
##  Max.   :1042    Max.   :10.000  
##  NA's   :11673
```

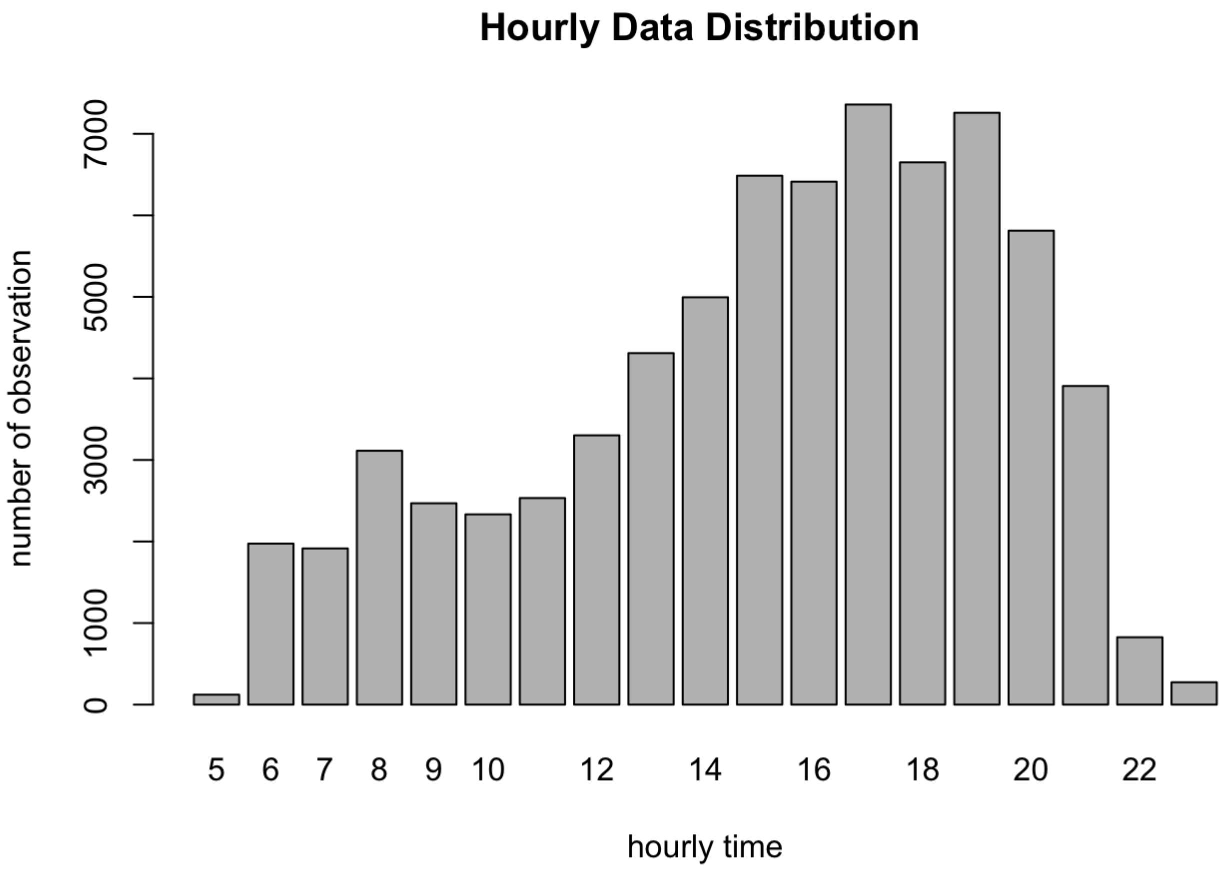
```
# wind guest contains lots of NA's, So I deleted this columnbus
total = total[,-30]
```

```
# Creating subset only containing delay and weather conditions
total_sub = total[,c(-1:-7,-9:-17,-19,-21:-24)]
summary(total_sub)
```

```
##    dep_delay          hour.x       delayed              temp       
##  Min.   :  15.00   Min.   : 5.0   Length:72051       Min.   : 10.94  
##  1st Qu.:  25.00   1st Qu.:13.0   Class :character   1st Qu.: 42.80  
##  Median :  44.00   Median :16.0   Mode  :character   Median : 60.08  
##  Mean   :  64.66   Mean   :15.2                      Mean   : 58.70  
##  3rd Qu.:  82.00   3rd Qu.:19.0                      3rd Qu.: 75.02  
##  Max.   :1301.00   Max.   :23.0                      Max.   :100.04  
##                                                      NA's   :10      
##       dewp           humid          wind_dir       wind_speed    
##  Min.   :-9.04   Min.   : 12.74   Min.   :  0.0   Min.   : 0.000  
##  1st Qu.:28.04   1st Qu.: 45.90   1st Qu.:130.0   1st Qu.: 8.055  
##  Median :46.94   Median : 62.40   Median :210.0   Median :11.508  
##  Mean   :44.52   Mean   : 62.78   Mean   :200.2   Mean   :11.657  
##  3rd Qu.:62.06   3rd Qu.: 81.14   3rd Qu.:280.0   3rd Qu.:14.960  
##  Max.   :78.08   Max.   :100.00   Max.   :360.0   Max.   :42.579  
##  NA's   :10      NA's   :10       NA's   :1779    NA's   :9       
##      precip            pressure         visib       
##  Min.   :0.000000   Min.   : 985    Min.   : 0.000  
##  1st Qu.:0.000000   1st Qu.:1011    1st Qu.:10.000  
##  Median :0.000000   Median :1016    Median :10.000  
##  Mean   :0.008623   Mean   :1016    Mean   : 8.986  
##  3rd Qu.:0.000000   3rd Qu.:1021    3rd Qu.:10.000  
##  Max.   :1.210000   Max.   :1042    Max.   :10.000  
##                     NA's   :11673
```

```
# Creating new / converting variables
total_sub$visib_class = cut(total_sub$visib, breaks = c(0,9,10), labels = c("low","high"))
```

```
total_sub$hour.x = as.factor(total_sub$hour.x)
plot(total_sub$hour.x, xlab = "hourly time", ylab = "number of observation" , main = "Hourly Dat
a Distribution ")
```



```
total_sub$hour.x = as.numeric(total_sub$hour.x)
```

```
library(gbm)
```

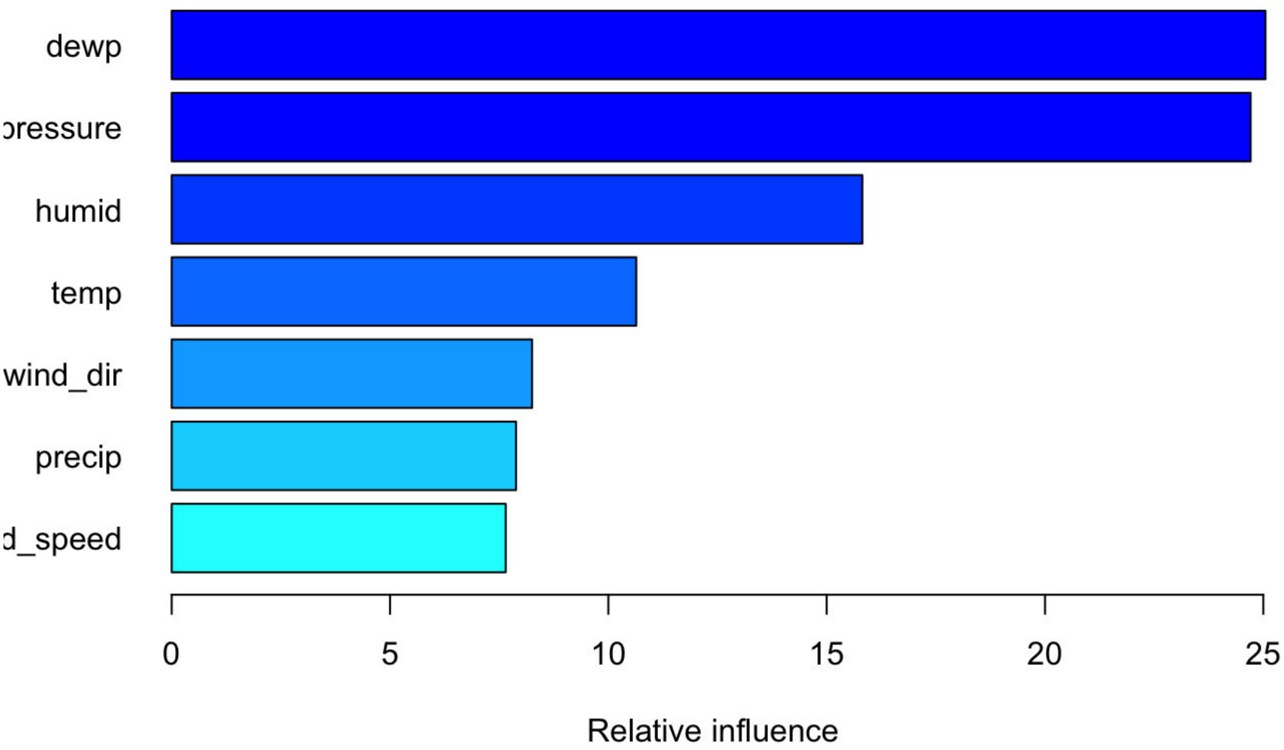
```
## Loaded gbm 2.1.8
```

```
set.seed(4911)
names(total_sub)
```

```
##  [1] "dep_delay"   "hour.x"      "delayed"     "temp"        "dewp"
##  [6] "humid"       "wind_dir"    "wind_speed"  "precip"      "pressure"
## [11] "visib"       "visib_class"
```

```
boost.fit = gbm(dep_delay~ temp + dewp + humid + wind_dir + wind_speed + precip + pressure ,data
=total_sub,distribution="gaussian",n.trees=500,interaction.depth=2)
```

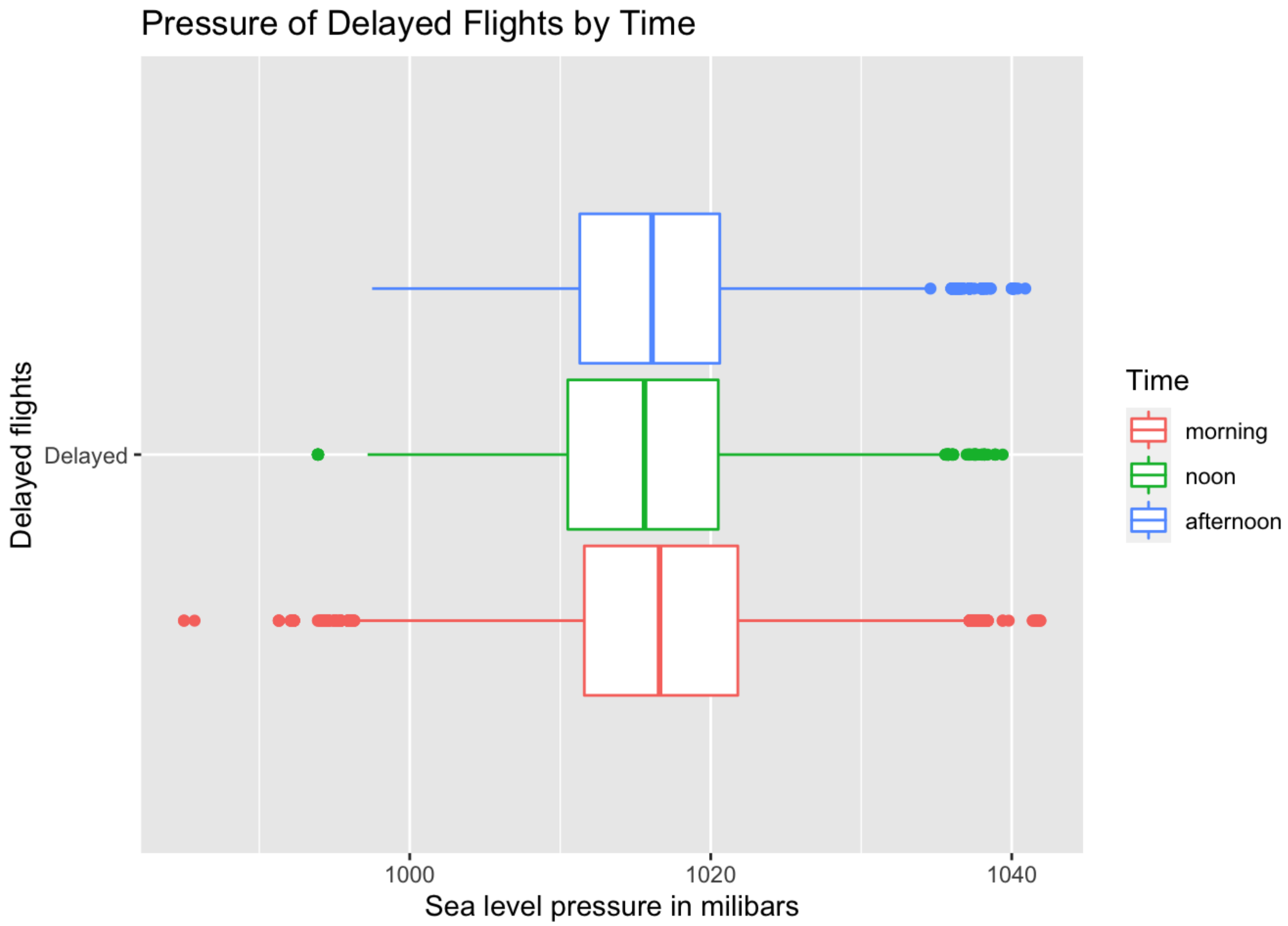
```
summary(boost.fit,las=1)
```



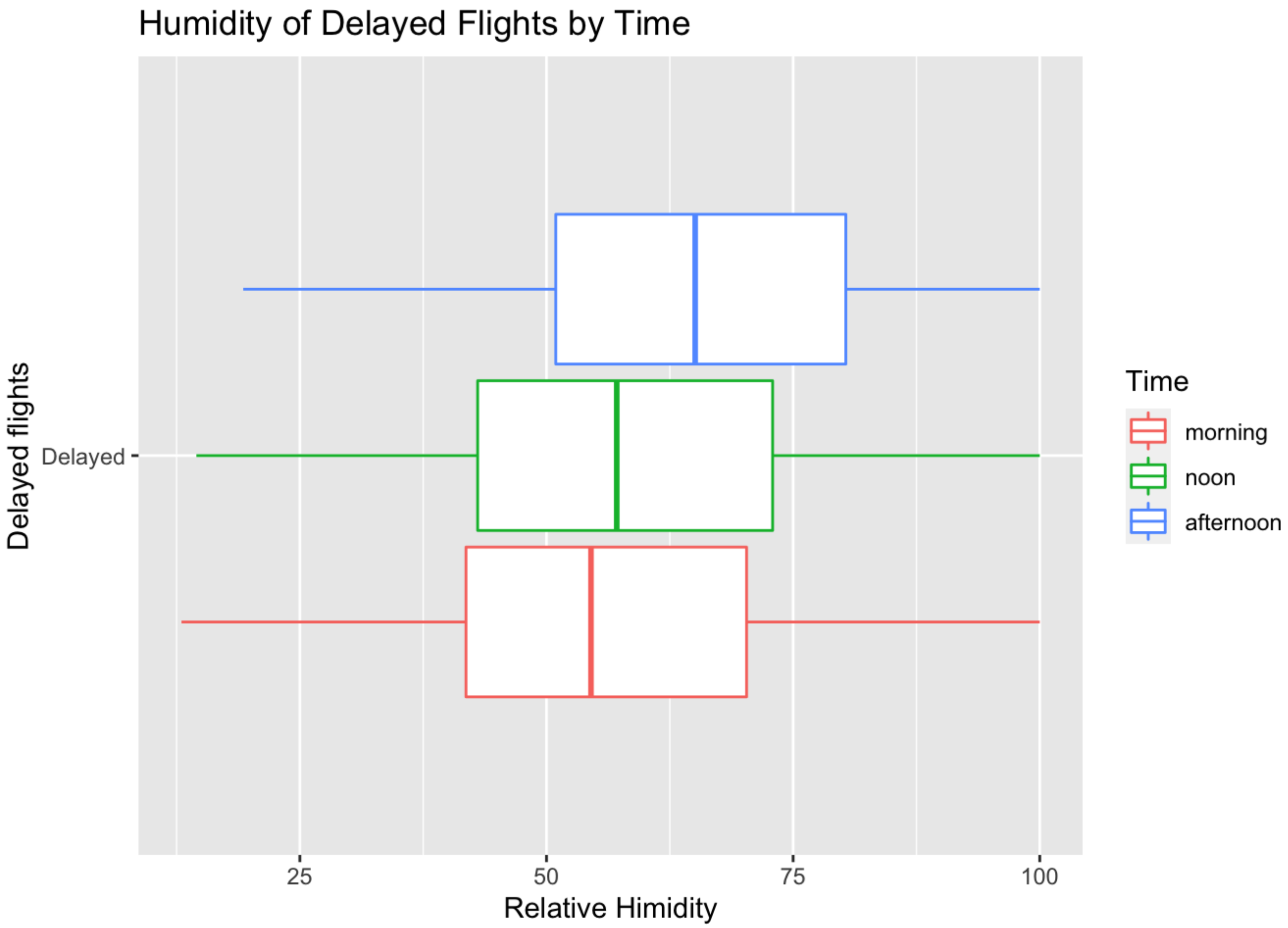
```
##                   var   rel.inf
## dewp             dewp 25.048016
## pressure     pressure 24.708103
## humid           humid 15.816552
## temp             temp 10.638823
## wind_dir     wind_dir  8.253186
## precip         precip  7.885960
## wind_speed wind_speed  7.649360
```

```
# Creating new / converting variables
total_sub$hour_class = cut(total_sub$hour.x, breaks = c(1,11,15,21,24), labels = c("morning","no
on","afternoon","night"))
```

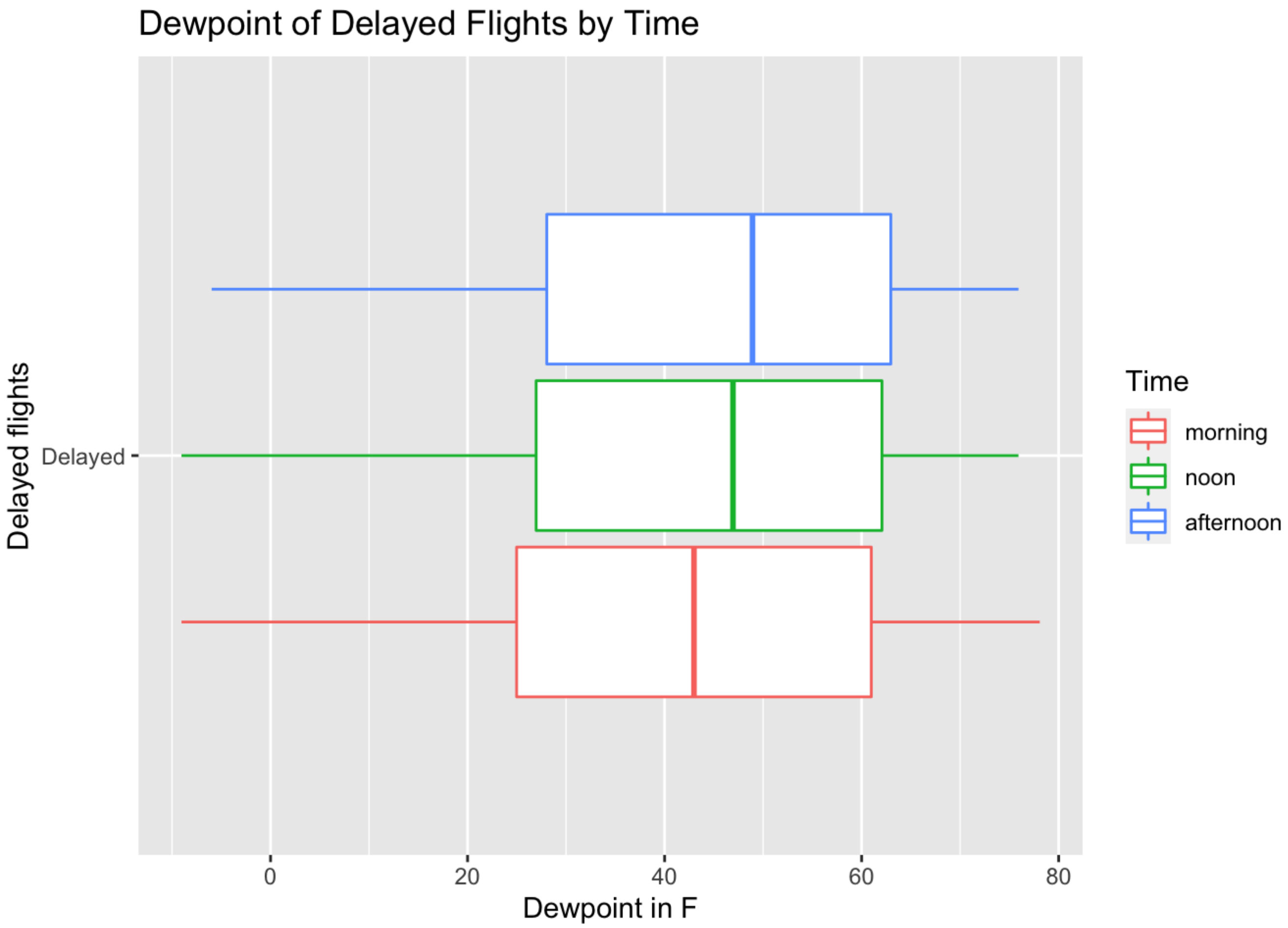
```
library(ggplot2)
total_sub = na.omit(total_sub)
ggplot(total_sub, aes(x = pressure, y = delayed )) + geom_boxplot(aes(color = hour_class)) + lab
s(x = "Sea level pressure in milibars", y = "Delayed flights" ,title = "Pressure of Delayed Flig
hts by Time")+ guides(color = guide_legend("Time"))
```



```
ggplot(total_sub, aes(x = humid, y = delayed)) + geom_boxplot(aes(color = hour_class)) + labs(x = "Relative Himidity", y = "Delayed flights", title = "Humidity of Delayed Flights by Time")+ guides(color = guide_legend("Time"))
```



```
ggplot(total_sub, aes(x = dewp, y = delayed)) + geom_boxplot(aes(color = hour_class)) + labs(x =
"Dewpoint in F", y = "Delayed flights" ,title = "Dewpoint of Delayed Flights by Time")+ guides(c
olor = guide_legend("Time"))
```



```
flightc <- flights_new %>%
  mutate(delay = ifelse(dep_delay >= 15 | is.na(dep_delay) == TRUE, 1, 0))
head(flightc)
```

```
##   year month day dep_time sched_dep_time dep_delay arr_time sched_arr_time
## 1 2013     1   1      517            515         2      830            819
## 2 2013     1   1      533            529         4      850            830
## 3 2013     1   1      542            540         2      923            850
## 4 2013     1   1      544            545        -1     1004           1022
## 5 2013     1   1      554            600        -6      812            837
## 6 2013     1   1      554            558        -4      740            728
##   arr_delay carrier flight tailnum origin dest air_time distance hour minute
## 1        11      UA   1545  N14228    EWR  IAH      227     1400    5     15
## 2        20      UA   1714  N24211    LGA  IAH      227     1416    5     29
## 3        33      AA   1141  N619AA    JFK  MIA      160     1089    5     40
## 4       -18      B6    725  N804JB    JFK  BQN      183     1576    5     45
## 5       -25      DL    461  N668DN    LGA  ATL      116      762    6      0
## 6        12      UA   1696  N39463    EWR  ORD      150      719    5     58
##             time_hour delay
## 1 2013-01-01 05:00:00     0
## 2 2013-01-01 05:00:00     0
## 3 2013-01-01 05:00:00     0
## 4 2013-01-01 05:00:00     0
## 5 2013-01-01 06:00:00     0
## 6 2013-01-01 05:00:00     0
```

```
flights_weather <- left_join(x = flightc, y = weather, by = c("origin","time_hour", "year", "month", "day", "hour"))
head(flights_weather)
```

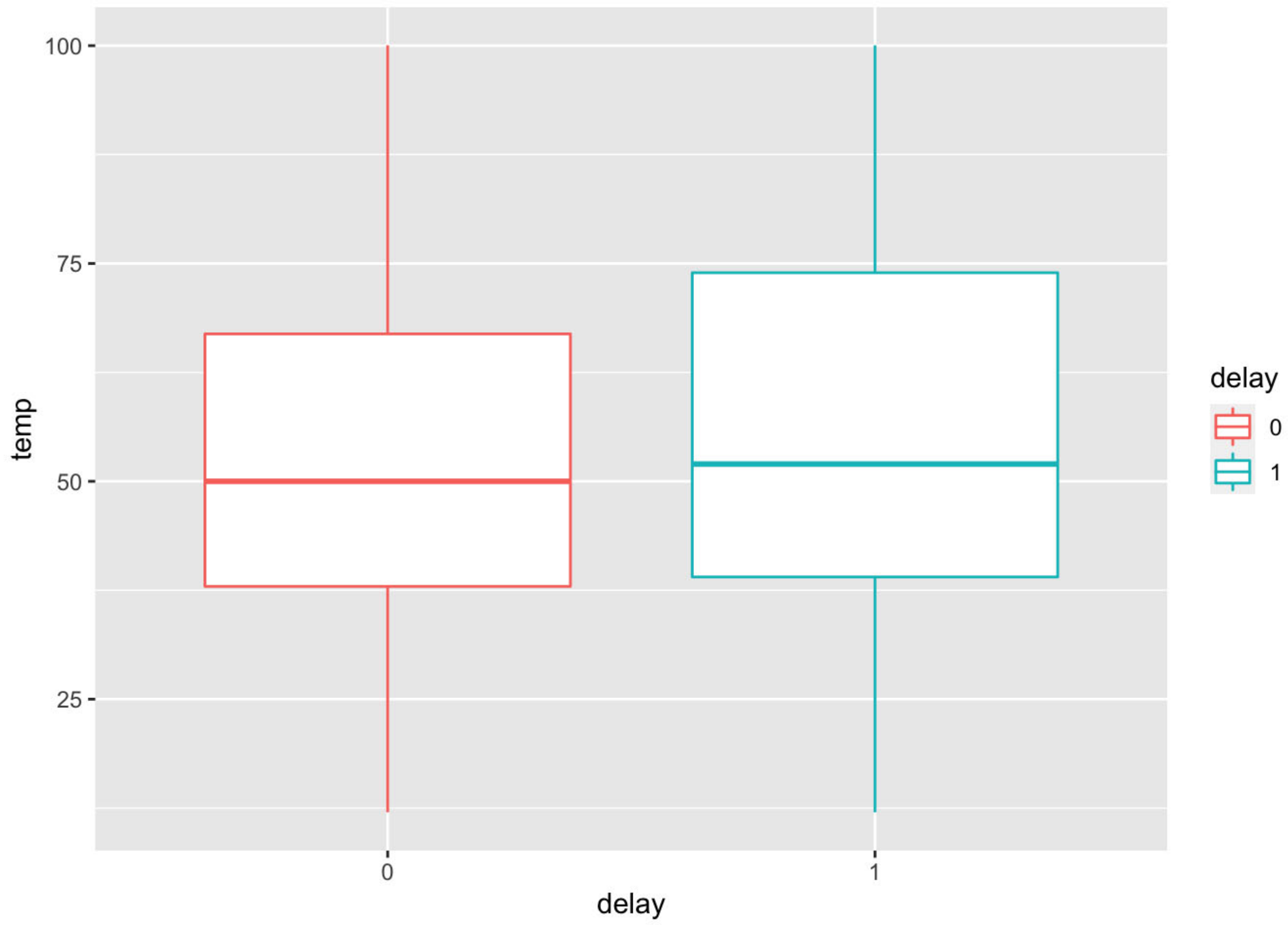
```
##   year month day dep_time sched_dep_time dep_delay arr_time sched_arr_time
## 1 2013     1   1      517            515         2      830            819
## 2 2013     1   1      533            529         4      850            830
## 3 2013     1   1      542            540         2      923            850
## 4 2013     1   1      544            545        -1     1004           1022
## 5 2013     1   1      554            600        -6      812            837
## 6 2013     1   1      554            558        -4      740            728
##   arr_delay carrier flight tailnum origin dest air_time distance hour minute
## 1        11      UA   1545  N14228    EWR  IAH      227     1400    5     15
## 2        20      UA   1714  N24211    LGA  IAH      227     1416    5     29
## 3        33      AA   1141  N619AA    JFK  MIA      160     1089    5     40
## 4       -18      B6    725  N804JB    JFK  BQN      183     1576    5     45
## 5       -25      DL    461  N668DN    LGA  ATL      116      762    6      0
## 6        12      UA   1696  N39463    EWR  ORD      150      719    5     58
##             time_hour delay  temp  dewp humid wind_dir wind_speed wind_gust
## 1 2013-01-01 05:00:00     0 39.02 28.04 64.43      260   12.65858        NA
## 2 2013-01-01 05:00:00     0 39.92 24.98 54.81      250   14.96014  21.86482
## 3 2013-01-01 05:00:00     0 39.02 26.96 61.63      260   14.96014        NA
## 4 2013-01-01 05:00:00     0 39.02 26.96 61.63      260   14.96014        NA
## 5 2013-01-01 06:00:00     0 39.92 24.98 54.81      260   16.11092  23.01560
## 6 2013-01-01 05:00:00     0 39.02 28.04 64.43      260   12.65858        NA
##   precip pressure visib
## 1      0   1011.9    10
## 2      0   1011.4    10
## 3      0   1012.1    10
## 4      0   1012.1    10
## 5      0   1011.7    10
## 6      0   1011.9    10
```

```
summary(flights_weather)
```

```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 500
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 905
##  Median :2013   Median : 7.000   Median :16.00   Median :1400   Median :1355
##  Mean   :2013   Mean   : 6.565   Mean   :15.74   Mean   :1349   Mean   :1340
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
##
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1122   1st Qu.: -17.000
##  Median :  -2.00   Median :1535   Median :1554   Median :  -5.000
##  Mean   :  12.56   Mean   :1502   Mean   :1533   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1944   3rd Qu.:  14.000
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##
##    carrier              flight       tailnum             origin
##  Length:327346      Min.   :   1   Length:327346      Length:327346
##  Class :character   1st Qu.: 544   Class :character   Class :character
##  Mode  :character   Median :1467   Mode  :character   Mode  :character
##                     Mean   :1943
##                     3rd Qu.:3412
##                     Max.   :8500
##
##      dest              air_time        distance         hour
##  Length:327346      Min.   : 20.0   Min.   :  80   Min.   : 5.00
##  Class :character   1st Qu.: 82.0   1st Qu.: 509   1st Qu.: 9.00
##  Mode  :character   Median :129.0   Median : 888   Median :13.00
##                     Mean   :150.7   Mean   :1048   Mean   :13.14
##                     3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4983   Max.   :23.00
##
##      minute        time_hour                       delay
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00   Min.   :0.0000
##  1st Qu.: 8.00   1st Qu.:2013-04-05 06:00:00   1st Qu.:0.0000
##  Median :29.00   Median :2013-07-04 09:00:00   Median :0.0000
##  Mean   :26.23   Mean   :2013-07-03 17:56:45   Mean   :0.2212
##  3rd Qu.:44.00   3rd Qu.:2013-10-01 18:00:00   3rd Qu.:0.0000
##  Max.   :59.00   Max.   :2013-12-31 23:00:00   Max.   :1.0000
##
##       temp             dewp           humid           wind_dir
##  Min.   : 10.94   Min.   :-9.94   Min.   : 12.74   Min.   :  0.0
##  1st Qu.: 42.08   1st Qu.:26.06   1st Qu.: 43.74   1st Qu.:130.0
##  Median : 57.20   Median :42.80   Median : 57.22   Median :220.0
##  Mean   : 57.01   Mean   :41.50   Mean   : 59.21   Mean   :201.9
##  3rd Qu.: 71.96   3rd Qu.:57.92   3rd Qu.: 74.67   3rd Qu.:290.0
##  Max.   :100.04   Max.   :78.08   Max.   :100.00   Max.   :360.0
##  NA's   :1544     NA's   :1544    NA's   :1544     NA's   :9574
##    wind_speed       wind_gust          precip          pressure
##  Min.   : 0.000   Min.   :16.11    Min.   :0.0000   Min.   : 983.8
##  1st Qu.: 6.905   1st Qu.:20.71    1st Qu.:0.0000   1st Qu.:1012.9
##  Median :10.357   Median :24.17    Median :0.0000   Median :1017.6
##  Mean   :11.060   Mean   :25.15    Mean   :0.0042   Mean   :1017.9
##  3rd Qu.:14.960   3rd Qu.:27.62    3rd Qu.:0.0000   3rd Qu.:1022.9
##  Max.   :42.579   Max.   :66.75    Max.   :1.2100   Max.   :1042.1
##  NA's   :1605     NA's   :249912   NA's   :1527     NA's   :36142
##      visib
##  Min.   : 0.00
##  1st Qu.:10.00
##  Median :10.00
##  Mean   : 9.29
##  3rd Qu.:10.00
##  Max.   :10.00
##  NA's   :1527
```

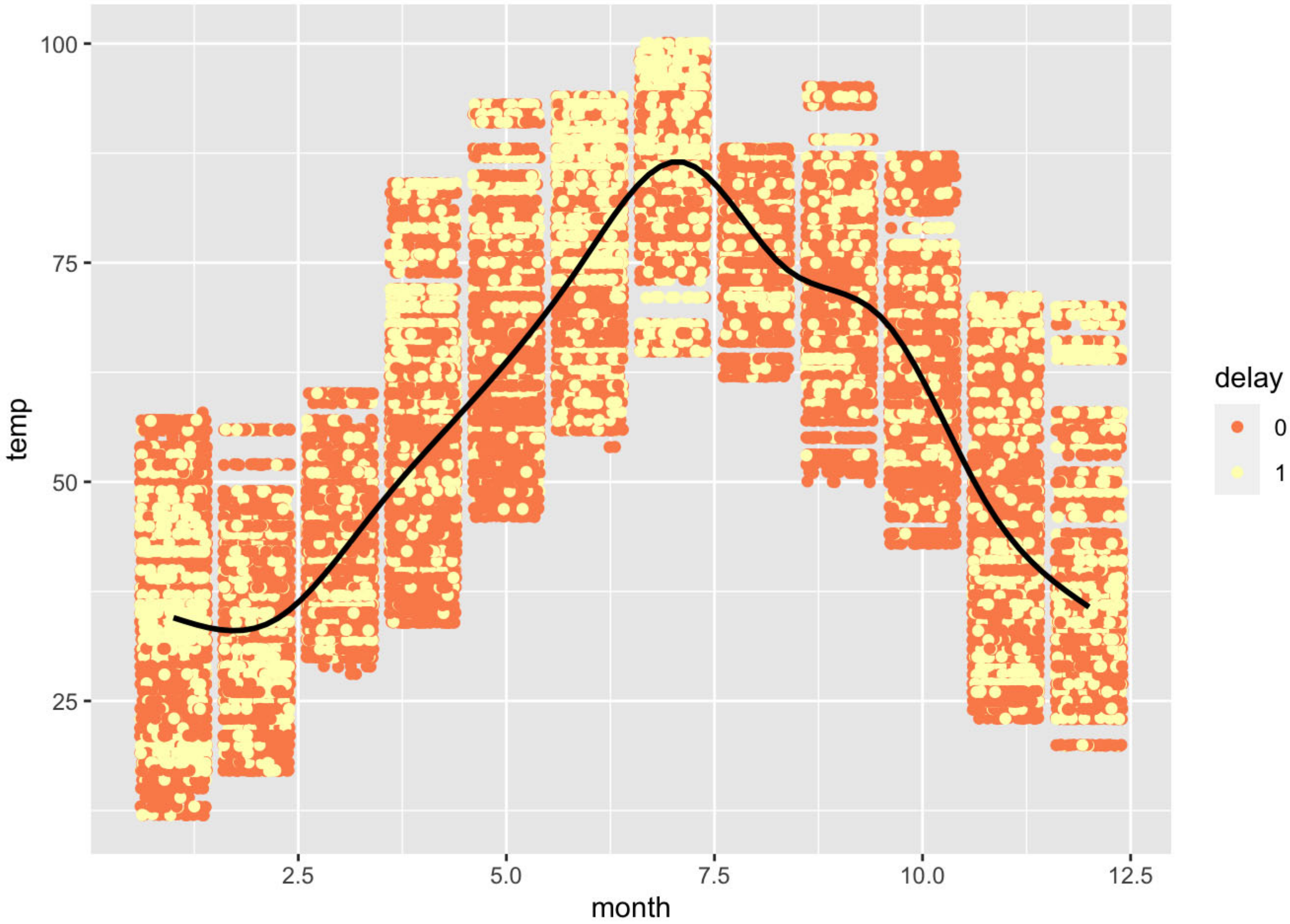
```
flights_weather_rm <- na.omit(flights_weather)
```

```
flights_weather_rm %>% mutate(delay = factor(delay)) %>%
  ggplot(aes(x = delay, y=temp, col = delay)) +
  geom_boxplot()
```

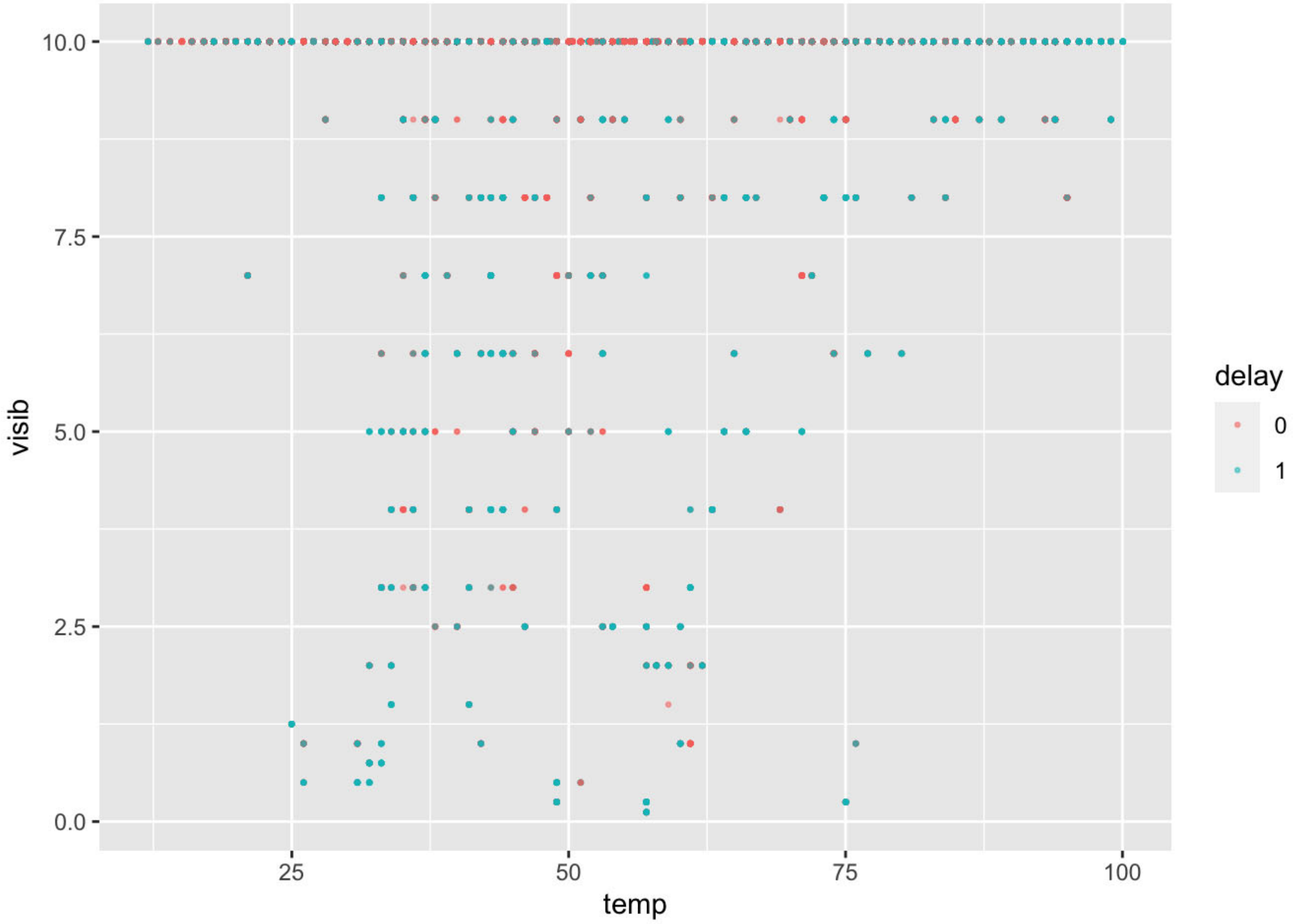


```
plot1 = flights_weather_rm %>% mutate(delay = factor(delay)) %>%
  ggplot(aes(y = temp, x=month, col = delay)) +
  geom_jitter() +
  geom_smooth(se=FALSE,fullrange=TRUE,color="black")
plot1 + scale_color_brewer(palette = "Spectral")
```

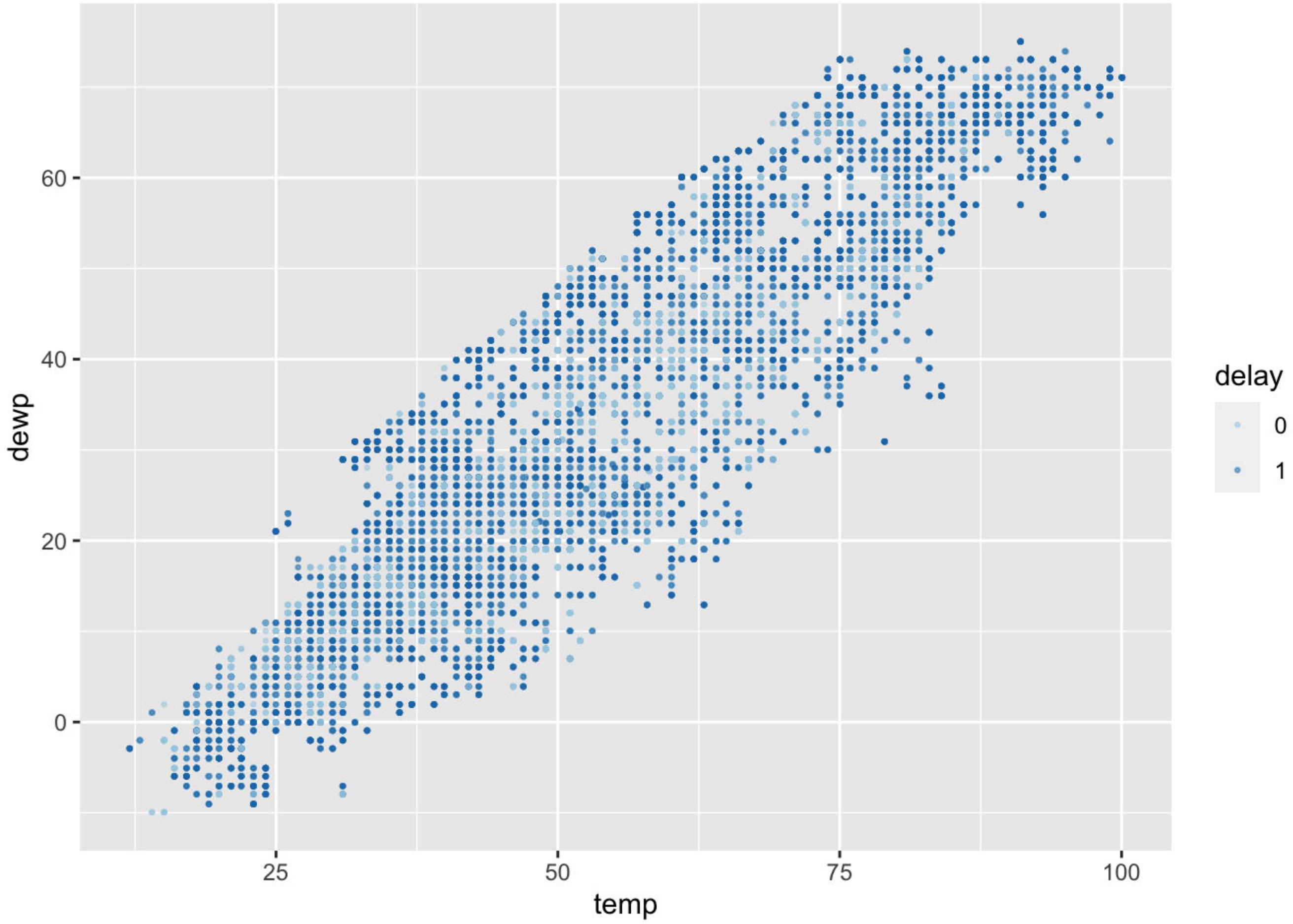
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



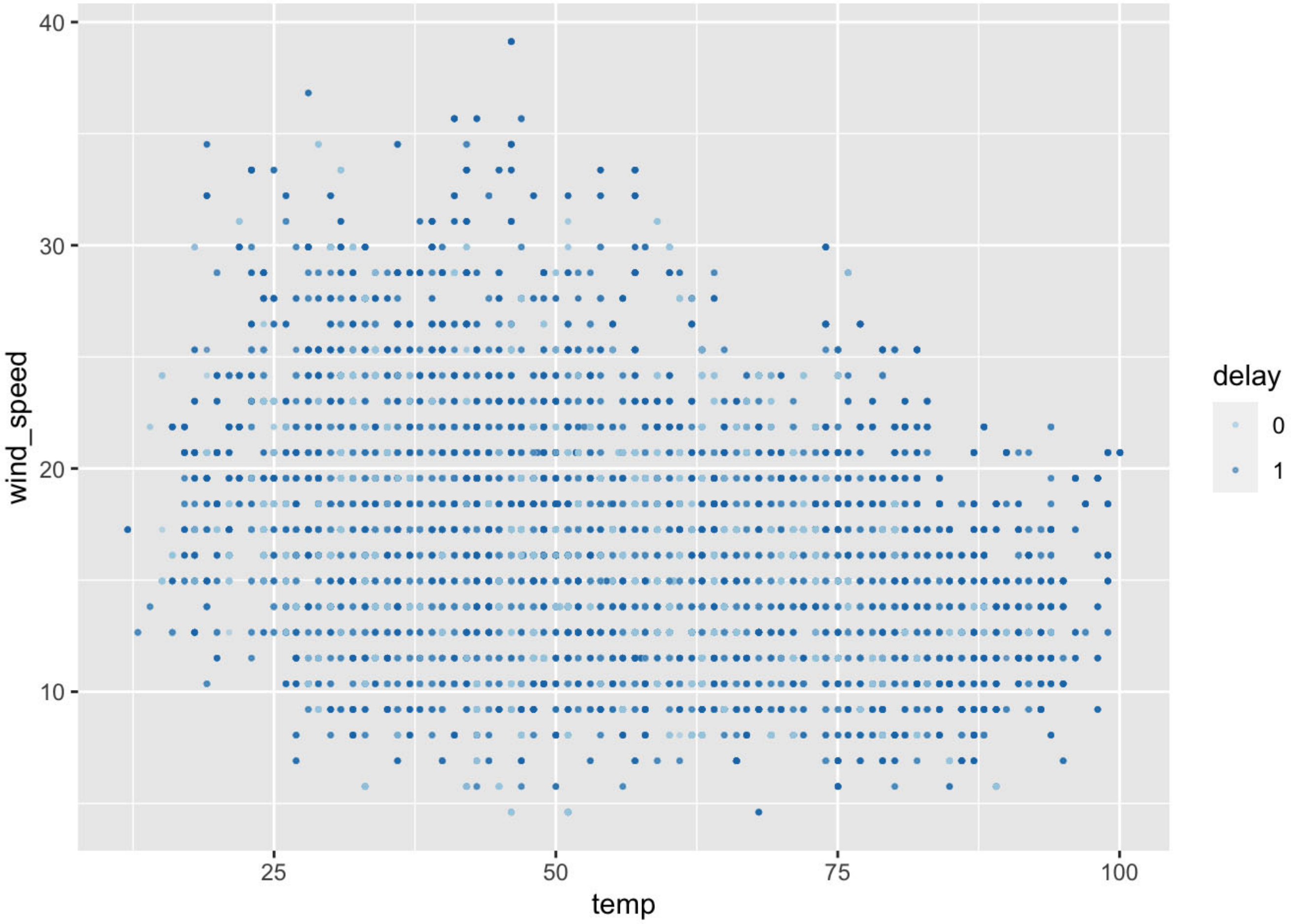
```
flights_weather_rm %>% mutate(delay = factor(delay)) %>%
  ggplot(aes(x = temp, y=visib, col = delay)) +
  geom_point(alpha=0.5, size = 0.5)
```



```
plot2=flights_weather_rm %>% mutate(delay = factor(delay)) %>%
  ggplot(aes(x = temp, y=dewp, col = delay)) +
  geom_point(alpha=0.5, size = 0.5)
plot2 + scale_color_brewer(palette = "Paired")
```

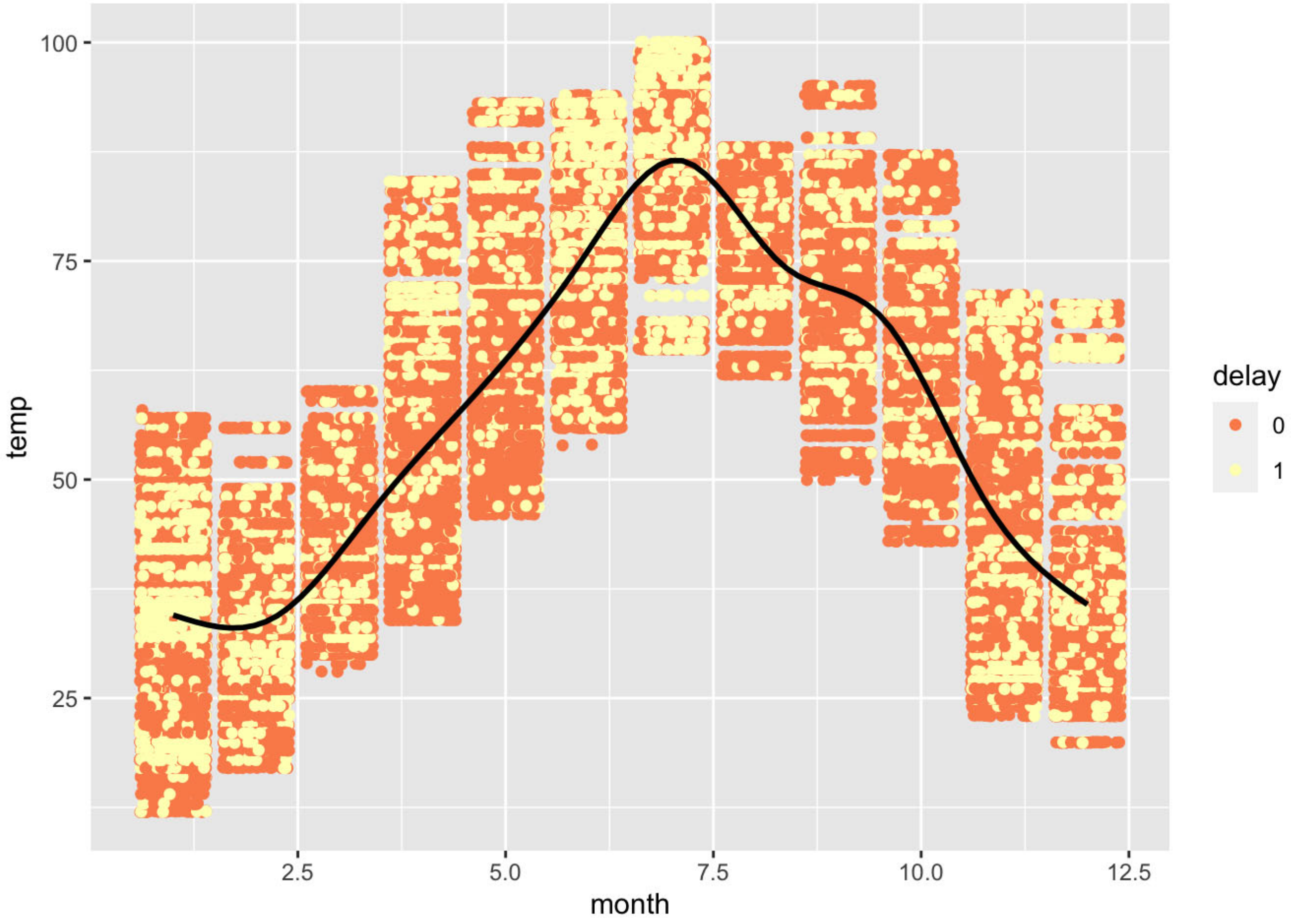


```
plot3=flights_weather_rm %>% mutate(delay = factor(delay)) %>%
  ggplot(aes(x = temp, y=wind_speed, col = delay)) +
  geom_point(alpha=0.5, size = 0.5)
plot3 + scale_color_brewer(palette = "Paired")
```



```
plot1 = flights_weather_rm %>% mutate(delay = factor(delay)) %>%
  ggplot(aes(y = temp, x=month, col = delay)) +
  geom_jitter() +
  geom_smooth(se=FALSE,fullrange=TRUE,color="black")
plot1 + scale_color_brewer(palette = "Spectral")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```


100
75
50
25
temp
2.5
5.0
7.5
10.0
12.5
month
delay
0
1