our week 2

Jialuo Li

9/19/2021

Git hub link:

https://github.com/JialuoLi/New-repo

it's at the bottom of the file list

(1)

```
filter(flights,is.na(dep_time))
```

```
## # A tibble: 8,255 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1       NA           1630        NA       NA
##  2  2013     1     1       NA           1935        NA       NA
##  3  2013     1     1       NA           1500        NA       NA
##  4  2013     1     1       NA            600        NA       NA
##  5  2013     1     2       NA           1540        NA       NA
##  6  2013     1     2       NA           1620        NA       NA
##  7  2013     1     2       NA           1355        NA       NA
##  8  2013     1     2       NA           1420        NA       NA
##  9  2013     1     2       NA           1321        NA       NA
## 10  2013     1     2       NA           1545        NA       NA
## # ... with 8,245 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

8255 flights have missing dep_time. seems that dep_delay, arr_time and arr_delay also have the same missing value. They may mean the delay of departure, arrive of time and delay of arrival. Bacause they are canceled thus no such values.

(2)

```
transmute(flights,deptime=60*dep_time %/% 100+dep_time %% 100, scedtime=60*sched_dep_time %/% 100+sched
```

```
## # A tibble: 336,776 x 2
##    deptime scedtime
##      <dbl>    <dbl>
##  1     317      315
##  2     333      329
##  3     342      340
##  4     344      345
```

```
##  5     354      360
##  6     354      358
##  7     355      360
##  8     357      360
##  9     357      360
## 10     358      360
## # ... with 336,766 more rows
```

since the thousand and hundred digit are hours, ten and unit are minute, we just need to multiply hours by 60 and add it by minute.

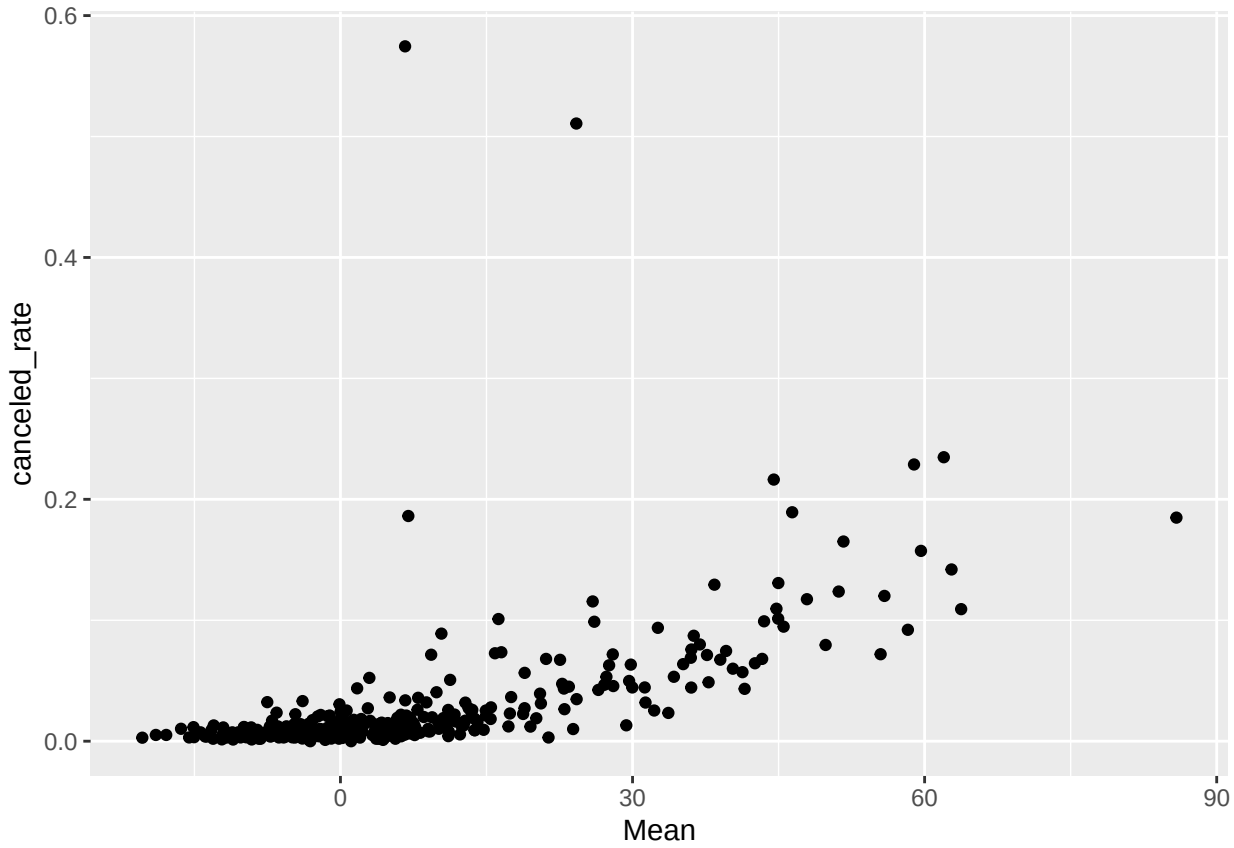
(3)

```
Delay<-flights %>% group_by(year,month,day) %>% summarise(Mean = mean(arr_delay,na.rm = TRUE),cancel=su
```

```
head(Delay)
```

```
## # A tibble: 6 x 6
## # Groups:   year, month [1]
##    year month   day  Mean cancel canceled_rate
##   <int> <int> <int> <dbl>  <int>         <dbl>
## 1  2013     1     1 12.7      11       0.0131
## 2  2013     1     2 12.7      15       0.0159
## 3  2013     1     3  5.73     14       0.0153
## 4  2013     1     4 -1.93      7       0.00765
## 5  2013     1     5 -1.53      3       0.00417
## 6  2013     1     6  4.24      3       0.00361
```

```
ggplot(data = Delay, mapping = aes(x=Mean, y=canceled_rate)) +
  geom_point()
```



Seems that the canceled rate increases when the mean of delay time increases