클린업 2주차 오답노트 2회차

서희나

2023-01-31

```
pacman::p_load(tidyverse,magrittr,data.table,lubridate)
```

```
train <- fread("train.csv")
stores <- fread("stores.csv")
oil <- fread("oil.csv")
holidays<-fread("holidays_events.csv")
```

문제 1.

데이터 설명서를 참고하여 각각이 어떤 데이터인지 자유롭게 파악해주세요.

```
head(train)
```

```
##        id       date store_nbr       family    sales onpromotion
## 1: 971190 2014-07-01         1   AUTOMOTIVE    2.000           0
## 2: 971191 2014-07-01         1    BABY CARE    0.000           0
## 3: 971192 2014-07-01         1       BEAUTY    6.000           0
## 4: 971193 2014-07-01         1    BEVERAGES 1868.000           0
## 5: 971194 2014-07-01         1        BOOKS    0.000           0
## 6: 971195 2014-07-01         1 BREAD/BAKERY  336.592           0
```

```
glimpse(train)
```

```
## Rows: 2,029,698
## Columns: 6
## $ id          <int> 971190, 971191, 971192, 971193, 971194, 971195, 971196, 97…
## $ date        <IDate> 2014-07-01, 2014-07-01, 2014-07-01, 2014-07-01, 2014-07-…
## $ store_nbr   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ family      <chr> "AUTOMOTIVE", "BABY CARE", "BEAUTY", "BEVERAGES", "BOOKS",…
## $ sales       <dbl> 2.000, 0.000, 6.000, 1868.000, 0.000, 336.592, 16.000, 720…
## $ onpromotion <int> 0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 2, 2, 0, 0, 0, 0, 0, 1, 0…
```

```
summary(train)
```

```
##        id               date              store_nbr       family         
##  Min.   : 971190   Min.   :2014-07-01   Min.   : 1.0   Length:2029698    
##  1st Qu.:1478614   1st Qu.:2015-04-12   1st Qu.:14.0   Class :character  
##  Median :1986039   Median :2016-01-23   Median :27.5   Mode  :character  
##  Mean   :1986039   Mean   :2016-01-22   Mean   :27.5                     
##  3rd Qu.:2493463   3rd Qu.:2016-11-03   3rd Qu.:41.0                     
##  Max.   :3000887   Max.   :2017-08-15   Max.   :54.0                     
##      sales          onpromotion     
##  Min.   :     0   Min.   :  0.000  
##  1st Qu.:     1   1st Qu.:  0.000  
##  Median :    18   Median :  0.000  
##  Mean   :   415   Mean   :  3.831  
##  3rd Qu.:   238   3rd Qu.:  1.000  
##  Max.   :124717   Max.   :741.000  
```

```
head(stores)
```

```
##    store_nbr          city                          state type cluster
## 1:         1         Quito                      Pichincha    D      13
## 2:         2         Quito                      Pichincha    D      13
## 3:         3         Quito                      Pichincha    D       8
## 4:         4         Quito                      Pichincha    D       9
## 5:         5 Santo Domingo Santo Domingo de los Tsachilas    D       4
## 6:         6         Quito                      Pichincha    D      13
```

```
glimpse(stores)
```

```
## Rows: 54
## Columns: 5
## $ store_nbr <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1…
## $ city      <chr> "Quito", "Quito", "Quito", "Quito", "Santo Domingo", "Quito"…
## $ state     <chr> "Pichincha", "Pichincha", "Pichincha", "Pichincha", "Santo D…
## $ type      <chr> "D", "D", "D", "D", "D", "D", "D", "D", "B", "C", "B", "C", …
## $ cluster   <int> 13, 13, 8, 9, 4, 13, 8, 8, 6, 15, 6, 15, 15, 7, 15, 3, 12, 1…
```

```
summary(stores)
```

```
##    store_nbr         city              state               type          
##  Min.   : 1.00   Length:54          Length:54          Length:54         
##  1st Qu.:14.25   Class :character   Class :character   Class :character  
##  Median :27.50   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :27.50                                                           
##  3rd Qu.:40.75                                                           
##  Max.   :54.00                                                           
##     cluster      
##  Min.   : 1.000  
##  1st Qu.: 4.000  
##  Median : 8.500  
##  Mean   : 8.481  
##  3rd Qu.:13.000  
##  Max.   :17.000  
```

```
head(oil)
```

```
##          date dcoilwtico
## 1: 2014-07-01     106.06
## 2: 2014-07-02     105.18
## 3: 2014-07-03     104.76
## 4: 2014-07-04         NA
## 5: 2014-07-07     104.19
## 6: 2014-07-08     104.06
```

```
glimpse(oil)
```

```
## Rows: 828
## Columns: 2
## $ date       <IDate> 2014-07-01, 2014-07-02, 2014-07-03, 2014-07-04, 2014-07-0…
## $ dcoilwtico <dbl> 106.06, 105.18, 104.76, NA, 104.19, 104.06, 102.93, 103.61,…
```

```
summary(oil)
```

```
##       date              dcoilwtico    
##  Min.   :2014-07-01   Min.   : 26.19  
##  1st Qu.:2015-04-15   1st Qu.: 44.75  
##  Median :2016-01-30   Median : 48.49  
##  Mean   :2016-01-30   Mean   : 52.99  
##  3rd Qu.:2016-11-15   3rd Qu.: 53.51  
##  Max.   :2017-08-31   Max.   :106.06  
##                       NA's   :29
```

```
head(holidays)
```

```
##          date    type   locale   locale_name
## 1: 2014-07-01   Event National       Ecuador
## 2: 2014-07-03 Holiday    Local     El Carmen
## 3: 2014-07-03 Holiday    Local Santo Domingo
## 4: 2014-07-04   Event National       Ecuador
## 5: 2014-07-05   Event National       Ecuador
## 6: 2014-07-08   Event National       Ecuador
##                                    description transferred
## 1: Mundial de futbol Brasil: Octavos de Final       FALSE
## 2:                 Cantonizacion de El Carmen       FALSE
## 3:                 Fundacion de Santo Domingo       FALSE
## 4: Mundial de futbol Brasil: Cuartos de Final       FALSE
## 5: Mundial de futbol Brasil: Cuartos de Final       FALSE
## 6:      Mundial de futbol Brasil: Semifinales       FALSE
```

```
glimpse(holidays)
```

```
## Rows: 233
## Columns: 6
## $ date        <IDate> 2014-07-01, 2014-07-03, 2014-07-03, 2014-07-04, 2014-07-…
## $ type        <chr> "Event", "Holiday", "Holiday", "Event", "Event", "Event", …
## $ locale      <chr> "National", "Local", "Local", "National", "National", "Nat…
## $ locale_name <chr> "Ecuador", "El Carmen", "Santo Domingo", "Ecuador", "Ecuad…
## $ description <chr> "Mundial de futbol Brasil: Octavos de Final", "Cantonizaci…
## $ transferred <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA…
```

```
summary(holidays)
```

```
##       date                 type              locale           locale_name       
##  Min.   :2014-07-01   Length:233         Length:233         Length:233        
##  1st Qu.:2015-06-23   Class :character   Class :character   Class :character  
##  Median :2016-04-29   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :2016-03-30                                                           
##  3rd Qu.:2016-12-22                                                           
##  Max.   :2017-12-26                                                           
##  description        transferred    
##  Length:233         Mode :logical  
##  Class :character   FALSE:223      
##  Mode  :character   TRUE :10       
##                                    
##                                    
##                                    
```

문제 2.

Holidays 데이터를 먼저 전처리 하겠습니다. Type 이 transferred 인 경우는 제외시켜주고 type 의 column 명을 holiday 로 변경해주세요.

```
holidays %<>%
  filter(!type=='transferred') %>%
  rename('holiday'='type')
```

```
holidays
```

```
##            date    holiday   locale   locale_name
##   1: 2014-07-01      Event National       Ecuador
##   2: 2014-07-03    Holiday    Local     El Carmen
##   3: 2014-07-03    Holiday    Local Santo Domingo
##   4: 2014-07-04      Event National       Ecuador
##   5: 2014-07-05      Event National       Ecuador
##  ---
## 229: 2017-12-22 Additional National       Ecuador
## 230: 2017-12-23 Additional National       Ecuador
## 231: 2017-12-24 Additional National       Ecuador
## 232: 2017-12-25    Holiday National       Ecuador
## 233: 2017-12-26 Additional National       Ecuador
##                                     description transferred
##   1: Mundial de futbol Brasil: Octavos de Final       FALSE
##   2:                 Cantonizacion de El Carmen       FALSE
##   3:                 Fundacion de Santo Domingo       FALSE
##   4: Mundial de futbol Brasil: Cuartos de Final       FALSE
##   5: Mundial de futbol Brasil: Cuartos de Final       FALSE
##  ---
## 229:                                  Navidad-3       FALSE
## 230:                                  Navidad-2       FALSE
## 231:                                  Navidad-1       FALSE
## 232:                                    Navidad       FALSE
## 233:                                  Navidad+1       FALSE
```

문제 3.

데이터 프레임 합치기(join 계열의 함수 활용)

```
data<-left_join(train,stores,by='store_nbr')
data<-plyr::join_all(list(data,oil,holidays),by='date',type='left',match='first')
```

```
nrow(train)==nrow(data)
```

```
## [1] TRUE
```

```
data %>% is.na %>% colSums
```

```
##        date          id   store_nbr      family       sales onpromotion
##           0           0           0           0           0           0
##        city       state        type     cluster  dcoilwtico     holiday
##           0           0           0           0      627264     1700028
##      locale locale_name description transferred
##     1700028     1700028     1700028     1700028
```

```
b<-oil %>%
  filter(is.na(dcoilwtico)) %>%
  select(date) %>% unique

a<-data %>%
  filter(is.na(dcoilwtico)) %>%
  select(date) %>% unique

anti_join(a,b,by='date') %>%
  mutate(days=wday(date,label=TRUE)) %>%
  distinct(days)
```

```
##    days
## 1:   토
## 2:   일
```

```
b %>% mutate(days=wday(date,label=TRUE)) %>% distinct(days)
```

```
##    days
## 1:   금
## 2:   월
## 3:   목
## 4:   화
```

```
a %<>% pull
b %<>% pull

a[which((a %in% b)==F)] %>%
  as.data.frame %>%
  mutate_at(vars(.),function(x){wday(x,label=T)}) %>%
  apply(2,unique)
```

```
##      .
## [1,] "토"
## [2,] "일"
```

3. holidays 의 NA 는 0 으로, NA 가 아닌 값은 1 로 바꿔주세요.

```
data %<>%
  mutate(holiday = if_else(is.na(holiday), 0, 1))
```

문제 4.

train 데이터의 총 기간인 2014-07-01 ~ 2017-08-15 중 data 에서 누락된 날짜가 있는지 확인해주세요

```
setdiff(
  seq(ymd('2014-07-01'), ymd('2017-08-15'), by ='1 day'),
  data$date %>% unique %>% ymd)
```

```
## [1] 16429 16794 17160
```

문제 5.

마지막 15 일을 test 기간으로 설정하여 data 를 train set 과 test set 으로 분리하겠습니다.(lubridate 패기지 함수 활용) 이때, 각 dataframe 을 test_set, train_set 으로 저장해주세요.

문제 6.

파생변수를 생성하겠습니다. date column 을 활용하여 요일을 나타내는 wday, 년도를 나타내는 year, 달을 의미하는 month 변수를 데이터프레임에 새롭게 생성해주세요.

```
data%<>%
  mutate(wday=wday(date,label=TRUE),
         year=year(date),
         month=month(date))
```

```
last15<-data %>%
  select(date) %>%
  unique %>%
  arrange(desc(date)) %>%
  head(15) %>%
  pull()

train<-data %>% filter(date<ymd('2017-08-01'))
test<-data %>% filter(date>=ymd('2017-08-01'))
```

```
train<-data %>% filter(!date %in% last15)
test<-data %>% filter(date %in% last15)
```

문제 7.

판매량이 0 인 날이 많은 store 들을 확인해보겠습니다.

```
train%>%
  group_by(date,store_nbr) %>%
  summarise(total_sales=sum(sales)) %>%
  filter(total_sales==0) %>%
  group_by(store_nbr) %>%
  summarise(zero_days=n_distinct(date)/n_distinct(train$date)) %>%
  arrange(-zero_days)->zero_days
```

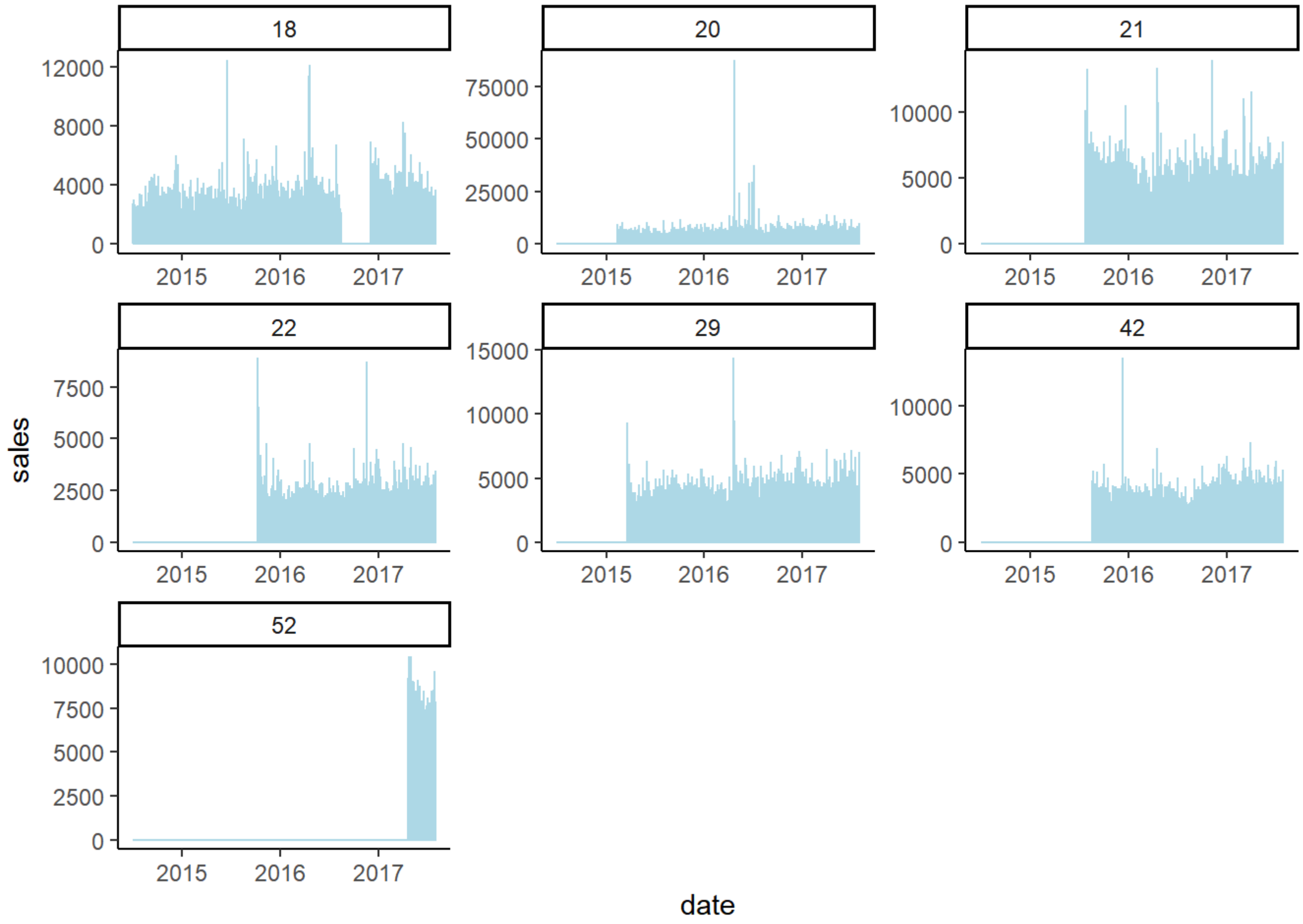
```
## `summarise()` has grouped output by 'date'. You can override using the
## `.groups` argument.
```

```
zero_days %>%
  filter(zero_days>=0.1) %>%
  select(store_nbr) %>% pull()->percent10
```

##문제 8.

판매량이 0 인 날이 10 퍼센트 이상인 52,22,42,21,29,20 의 판매량을 아래와 같이 시각화해서 알아본 후 train_set 에서 52,22,42,21,29,20 에 해당하는 row 는 제거해주세요

```
train %>%
  filter(store_nbr %in% percent10) %>%
  ggplot(aes(x=date,y=sales))+
  geom_line(color='lightblue')+
  facet_wrap(vars(store_nbr),nrow=3,scale='free')+
  theme_classic()
```



```
train %<>%
  filter(!store_nbr %in% percent10)
```

문제 9.

family 도 분포를 살펴본 후 제거해주겠습니다.

```
train %>%
  group_by(date,family) %>%
  summarise(sales_sum=sum(sales)) %>%
  filter(sales_sum==0) %>%
  group_by(family) %>%
  summarise(zero_family=n_distinct(date)/n_distinct(train$date)) %>%
  arrange(zero_family)->zero_family
```

```
## `summarise()` has grouped output by 'date'. You can override using the
## `.groups` argument.
```

```
zero_family %>%
  filter(!zero_family<0.1) %>%
  select(family) %>% pull->family_list

train %<>%
  filter(!family %in% family_list)
```

##문제 10.

범주형 변수를 범주형으로 바꿔주세요.

```
train %>%
  select_if(summarise_all(.,n_distinct)<=75) %>%
  lapply(unique)
```

```
## $store_nbr
##  [1]  1 10 11 12 13 14 15 16 17 19  2 23 24 25 26 27 28  3 30 31 32 33 34 35 36
## [26] 37 38 39  4 40 41 43 44 45 46 47 48 49  5 50 51 53 54  6  7  8  9
## 
## $family
##  [1] "AUTOMOTIVE"          "BEAUTY"              "BEVERAGES"          
##  [4] "BREAD/BAKERY"        "CLEANING"            "DAIRY"              
##  [7] "DELI"                "EGGS"                "FROZEN FOODS"       
## [10] "GROCERY I"           "GROCERY II"          "HARDWARE"           
## [13] "HOME AND KITCHEN I"  "HOME AND KITCHEN II" "HOME APPLIANCES"    
## [16] "LAWN AND GARDEN"     "LINGERIE"            "LIQUOR,WINE,BEER"   
## [19] "MEATS"               "PERSONAL CARE"       "POULTRY"            
## [22] "PREPARED FOODS"      "PRODUCE"             "SEAFOOD"            
## 
## $city
##  [1] "Quito"         "Cayambe"       "Latacunga"     "Riobamba"     
##  [5] "Ibarra"        "Santo Domingo" "Guaranda"      "Ambato"       
##  [9] "Guayaquil"     "Salinas"       "Daule"         "Babahoyo"     
## [13] "Quevedo"       "Playas"        "Libertad"      "Cuenca"       
## [17] "Loja"          "Machala"       "Esmeraldas"    "Manta"        
## [21] "El Carmen"    
## 
## $state
##  [1] "Pichincha"                      "Cotopaxi"                      
##  [3] "Chimborazo"                     "Imbabura"                      
##  [5] "Santo Domingo de los Tsachilas" "Bolivar"                       
##  [7] "Tungurahua"                     "Guayas"                        
##  [9] "Santa Elena"                    "Los Rios"                      
## [11] "Azuay"                          "Loja"                          
## [13] "El Oro"                         "Esmeraldas"                    
## [15] "Manabi"                        
## 
## $type
## [1] "D" "C" "B" "E" "A"
## 
## $cluster
##  [1] 13 15  6  7  3 12  9  1 10  8  2  4  5 11 14 17
## 
## $holiday
## [1] 1 0
## 
## $locale
## [1] "National" NA         "Local"    "Regional"
## 
## $locale_name
##  [1] "Ecuador"                        NA                              
##  [3] "El Carmen"                      "Cayambe"                       
##  [5] "Guayaquil"                      "Esmeraldas"                    
##  [7] "Riobamba"                       "Ambato"                        
##  [9] "Ibarra"                         "Quevedo"                       
## [11] "Santo Domingo de los Tsachilas" "Santa Elena"                   
## [13] "Guaranda"                       "Latacunga"                     
## [15] "Quito"                          "Loja"                          
## [17] "Salinas"                        "Manta"                         
## [19] "Cotopaxi"                       "Cuenca"
## [21] "Libertad"                       "Puyo"
## [23] "Machala"                        "Imbabura"
## 
## $transferred
## [1] FALSE    NA  TRUE
## 
## $wday
## [1] 화 수 목 금 토 일 월
## Levels: 일 < 월 < 화 < 수 < 목 < 금 < 토
## 
## $year
## [1] 2014 2015 2016 2017
## 
## $month
##  [1]  7  8  9 10 11 12  1  2  3  4  5  6
```

```
train %<>%
  mutate_if(summarise_all(.,n_distinct)<=75,
            as.factor)
```

```
colnames(train)[which(summarise_all(train,n_distinct)<=75)]->category
```

Part2. NA imputation

문제 1.

먼저, date, dcoilwtico 로만 이루어진 데이터프레임을 생성해줍니다. 이때, dataframe 에서 중복되는 row 가 존재하지 않는 dataframe 을 만들어주세요.

```
train %>%
  select(date,dcoilwtico) %>%
  filter(!duplicated(date))->oil_price
```

문제 2.

NA 가 있는 행은 그 전날의 oil price 로 대체해주세요

```
na_index<-which(oil_price$dcoilwtico %>% is.na)

for (i in na_index){
  oil_price$dcoilwtico[i]<-oil_price$dcoilwtico[i-1]
}
```

문제 3.

train_set 에 oil_price 라는 이름의 변수로 결합시켜주고 이전의 dcoilwtico 는 제거합니다.

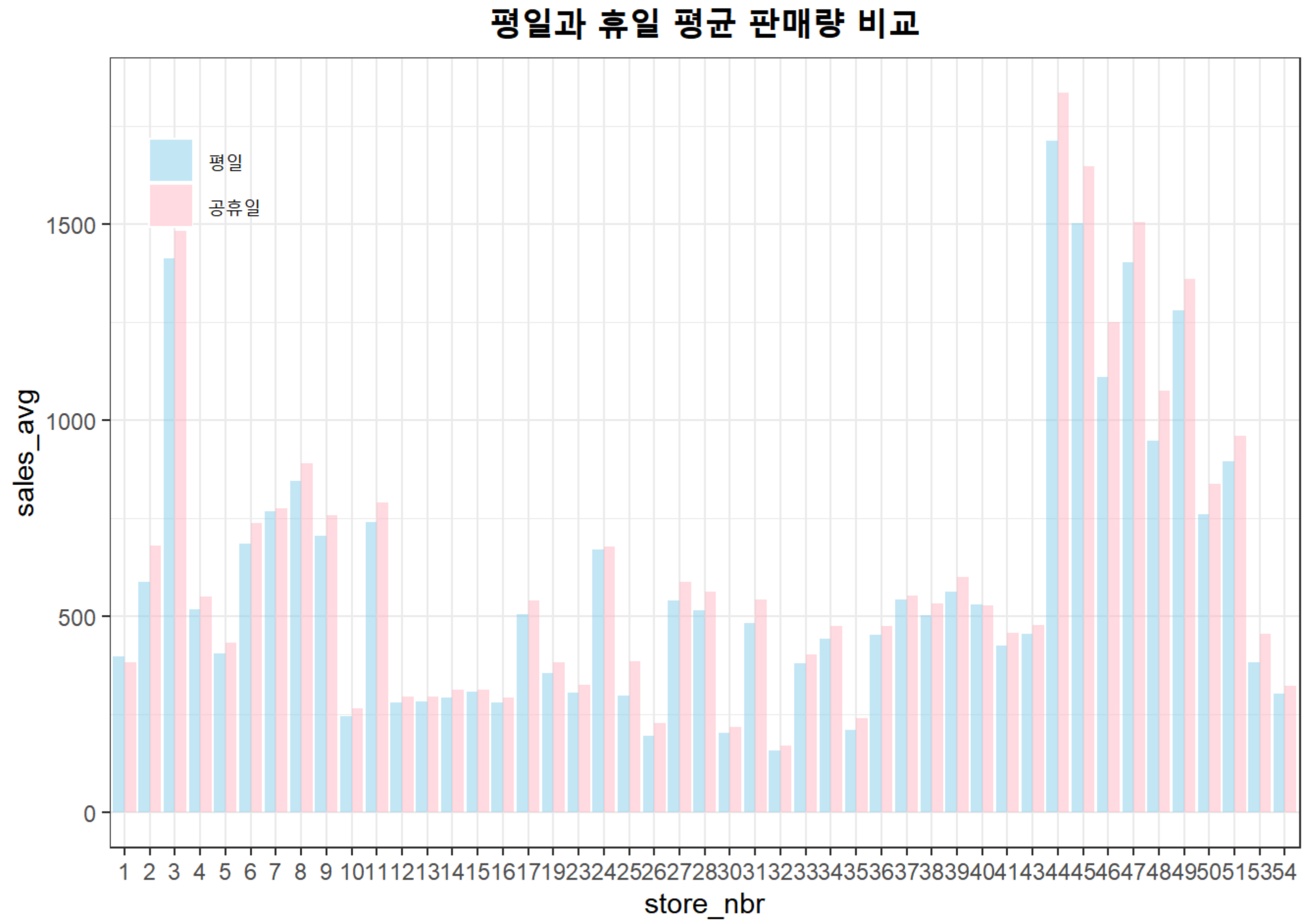
```
left_join(train %>% select(-dcoilwtico),
          oil_price,
          by='date')->train
```

Part3. EDA

문제 1. holiday & store 에 따른 sales 의 차이를 확인하는 아래의 플랏을 그린 후 알 수 있는 점을간략히 적어주세요.

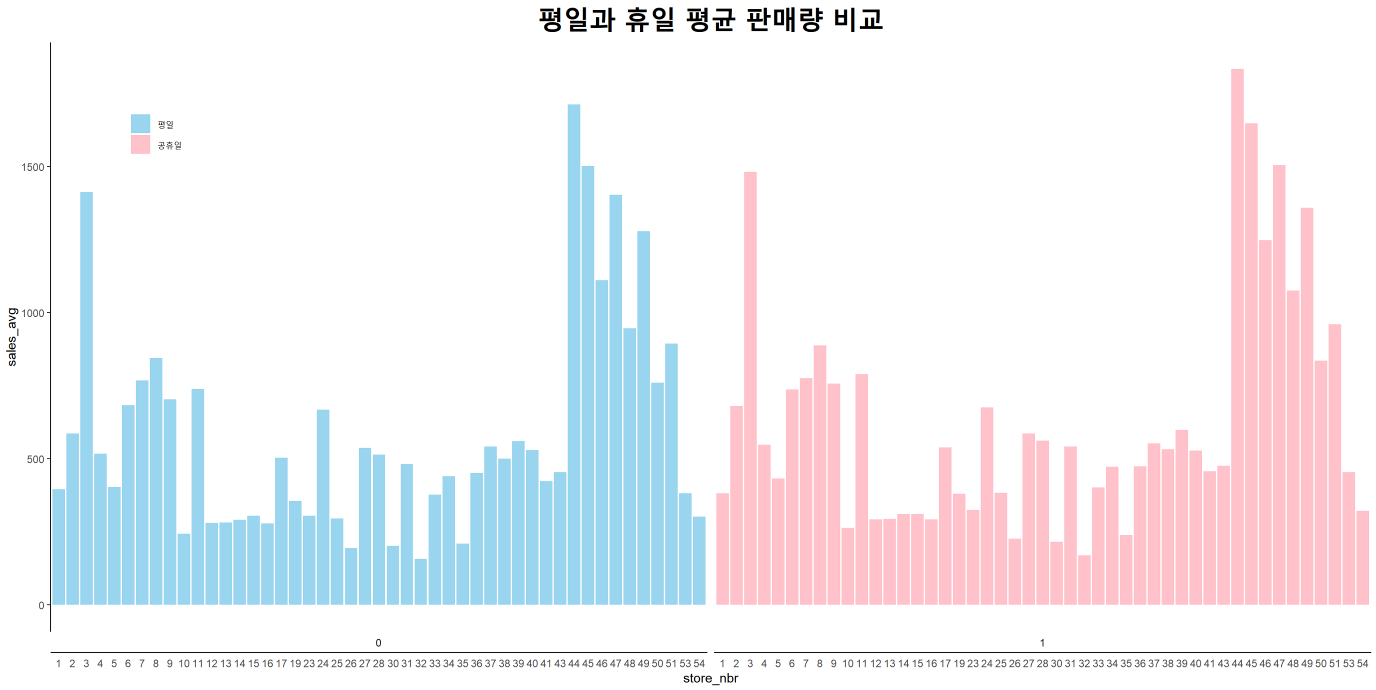
```
color<-c('skyblue','lightpink')
label<-c("평일","공휴일")

train %>%
  group_by(store_nbr,holiday) %>%
  summarise(sales_avg=mean(sales)) %>%
  ggplot(aes(store_nbr,sales_avg,fill=holiday))+
  geom_bar(stat='identity',position='dodge',alpha=0.5)+
  theme_bw()+
  scale_fill_manual(values=color,labels=label)+
  ggtitle('평일과 휴일 평균 판매량 비교')+
  theme(plot.title=element_text(size=15,hjust=0.5,face='bold'))+
  theme(legend.position=c(0.08,0.85),
        legend.background = element_rect(fill='transparent'),
        legend.title=element_blank())
```

```
## `summarise()` has grouped output by 'store_nbr'. You can override using the
## `.groups` argument.
```



```
train %>%
  group_by(store_nbr,holiday) %>%
  summarise(sales_avg=mean(sales)) %>%
  ggplot(aes(x = store_nbr,  y =sales_avg, fill = holiday))  +
  geom_col(position = "dodge",alpha=0.85) +
  facet_grid(~holiday, scales = "free_x", space = "free_x", switch = "x") +
  scale_fill_manual(values=color,labels=label)+
  theme_classic()+
  theme(#axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        #axis.title.x=element_blank(),
        #axis.title.y=element_blank(),
        strip.background = element_blank(),
        legend.position=c(0.08,0.85),
        legend.background = element_rect(fill='transparent'),
        legend.title=element_blank())+
  ggtitle('평일과 휴일 평균 판매량 비교')+
  theme(plot.title=element_text(size=25,hjust=0.5,face='bold'))
```

```
## `summarise()` has grouped output by 'store_nbr'. You can override using the
## `.groups` argument.
```

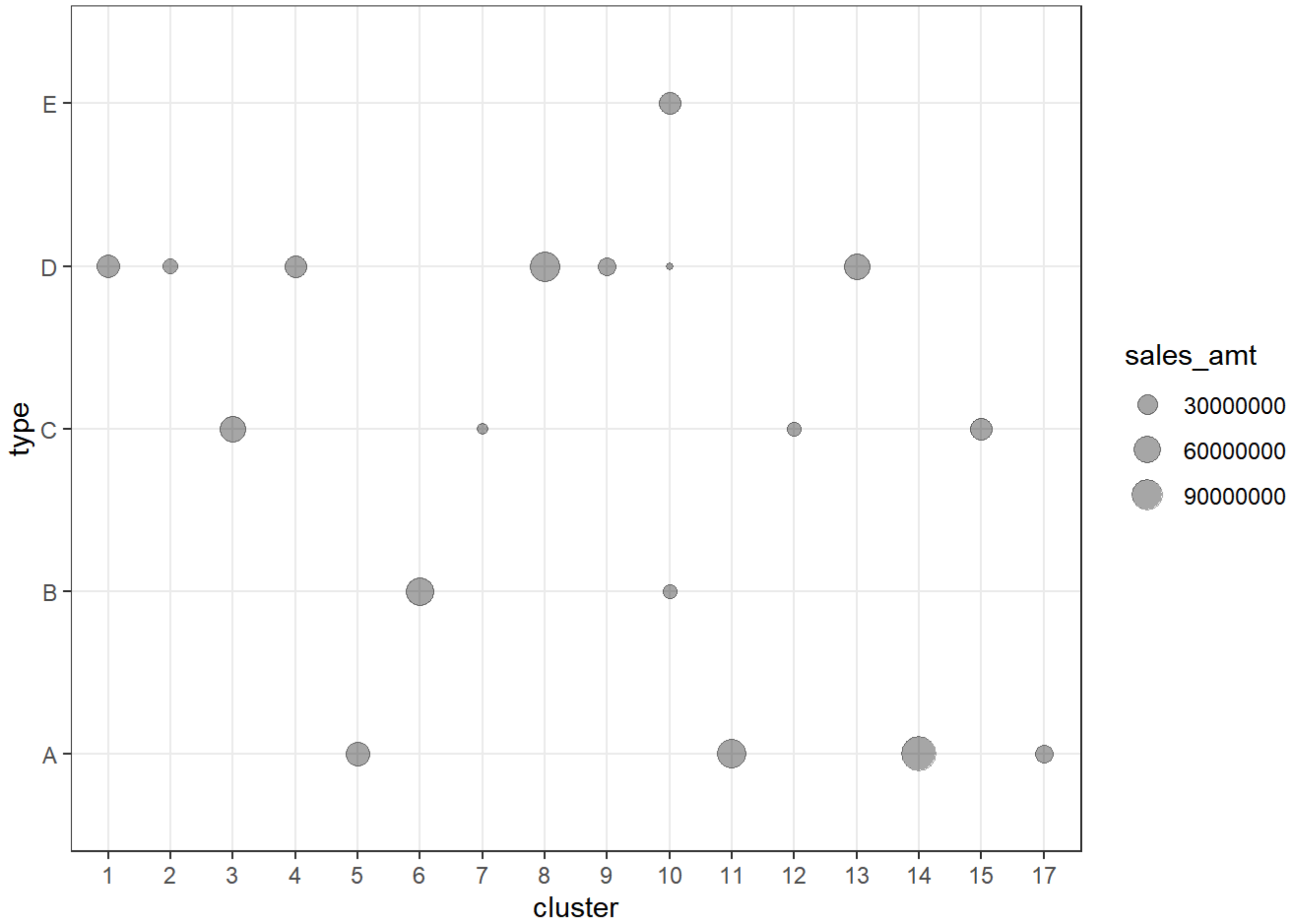


문제 2.

store 의 type 이 판매량과 연관이 있는지 다음의 플랏을 통해 확인해보세요.

```
options(scipen=999)

train %>%
  group_by(cluster,type) %>%
  summarise(sales_amt=sum(sales)) %>%
  ggplot(aes(x=cluster,y=type,size=sales_amt))+
  geom_point(alpha=0.35)+
  theme_bw()
```

```
## `summarise()` has grouped output by 'cluster'. You can override using the
## `.groups` argument.
```

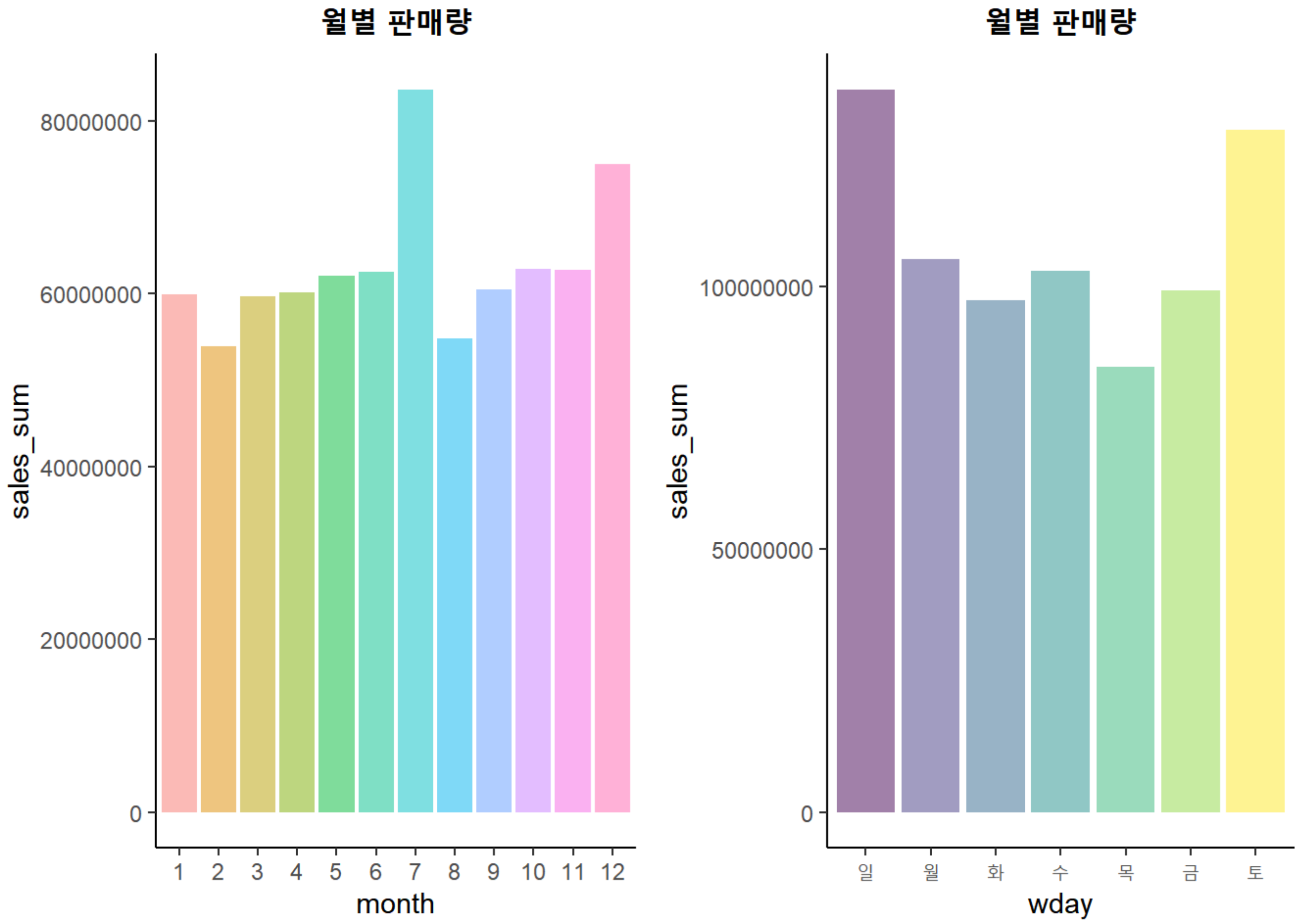


```
library(gridExtra)
```

```
##
## 다음의 패키지를 부착합니다: 'gridExtra'
```

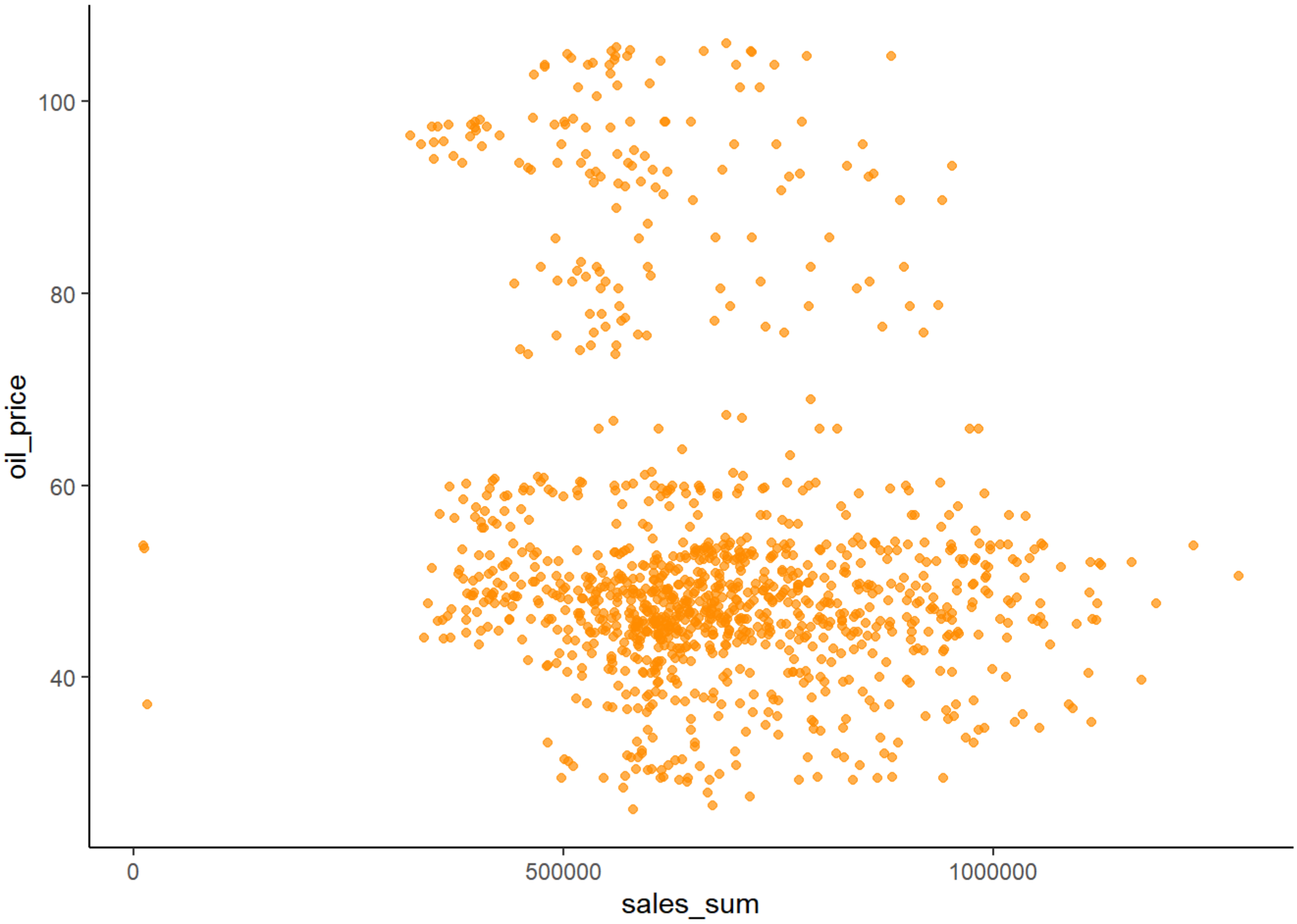
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
month<-train %>%
  group_by(month) %>%
  summarise(sales_sum=sum(sales)) %>%
  ggplot(aes(x=month,y=sales_sum,fill=month))+
  geom_bar(stat='identity',alpha=0.5)+
  theme_classic()+
  labs(title='월별 판매량')+
  theme(legend.position='none',
        plot.title=element_text(hjust=0.5,face='bold'))

day<-train %>%
  group_by(wday) %>%
  summarise(sales_sum=sum(sales)) %>%
  ggplot(aes(x=wday,y=sales_sum,fill=wday))+
  geom_bar(stat='identity',alpha=0.5)+
  theme_classic()+
  labs(title='월별 판매량')+
  theme(legend.position='none',
        plot.title=element_text(hjust=0.5,face='bold'))

grid.arrange(month,day,ncol=2)
```



```
train %>%
  group_by(date,dcoilwtico) %>%
  summarise(sales_sum=sum(sales)) %>%
  ggplot(aes(sales_sum,dcoilwtico,color=dcoilwtico))+
  geom_point(alpha=0.7,color='darkorange')+
  theme_classic()+
  labs(y='oil_price')
```

```
## `summarise()` has grouped output by 'date'. You can override using the
## `.groups` argument.
```



```
#  coord_cartesian(xlim = c(0, 3000000))
```

```
train %>%
  group_by(date,dcoilwtico) %>%
  summarise(sales_sum=sum(sales))->eda
```

```
## `summarise()` has grouped output by 'date'. You can override using the
## `.groups` argument.
```

```
cor(eda$dcoilwtico,eda$sales_sum,method='pearson')
```

```
## [1] -0.2049756
```

Part4. Time series CV

```
#library(catboost)
library(Metrics)
```

문제 1.

모델링에 필요한 column 만 남깁니다.(date, id, oil_price, state, type, cluster 제거해주세요)

```
train<-train%>%
  select(-date,-id,-dcoilwtico,-state,-type,-cluster)
```

문제 2.

아래와 같은 데이터프레임을 만들고 catboost 모델과 오늘 사용할 두 파라미터에 대해 간단히 설명해주세요.(데이터프레임명:result)

```
result<-expand.grid(learning_rate=c(0.10,0.01),
                    iterations=c(50,100),
                    rmse=NA)

result
```

```
##   learning_rate iterations rmse
## 1          0.10         50   NA
## 2          0.01         50   NA
## 3          0.10        100   NA
## 4          0.01        100   NA
```

문제 3.

Time series cv 를 위해 index list 를 생성해주세요. 아래의 그림과 같이 train set 과 validation set 이 구성될 수 있도록 만들어주세요

```
index<-c()
for (i in seq(5,1)){
  cv<-nrow(train)-i*26730
  index<-c(index,cv)
}

index
```

```
## [1] 1134222 1160952 1187682 1214412 1241142
```

3 번에서 생성한 index 를 활용하여 시계열 교차검증을 통해 구한 rmse 를 result data 에 저장한 뒤 rmse 가 가장 낮은 행을 출력해보세요.

```
set.seed(1003)

rmse<-c()
x<-train%>% select(sales)
y<-train %>% select(-sales)

for (i in 1:nrow(result)){

  lr=result$learning_rate[i]
  iter=result$iterations[i]

  for(j in index){

    x_train<-x[1:(j-1),]%>% as.matrix %>% as.integer
    y_train<-y[1:(j-1),]

    x_valid<-x[j:nrow(train),]%>% as.matrix %>% as.integer
    y_valid<-y[j:nrow(train),]

    train_pool<-catboost.load_pool(data=x_train, label = y_train)
    val_pool<-catboost.load_pool(data=x_val, label=y_val)

    params <- list(iterations=lr,
                   learning_rate=lr,
                   loss_function='RMSE',
                   random_seed = 1003,
                   logging_level='Silent')

    model<-catboost.train(learn_pool=train_pool,params=params)
    predict<-catboost.predict(model,val_pool)
    loss<-c(loss,rmse(predict,y_val))}

  result$rmse[i]<-mean(loss)
}
```

```
## Error in catboost.load_pool(data = x_train, label = y_train): 함수 "catboost.load_pool"를 찾
을 수 없습니다
```

```
result[which(result$rmse==min(result$rmse)),]
```

```
## [1] learning_rate iterations    rmse
## <0 행> <또는 row.names의 길이가 0입니다>
```

```
##   learning_rate iterations     rmse
## 3           0.1        100 56.41204
```

#Part5. Modelling & Prediction

Test_set 에 train_set 에 했던 전처리 과정을 똑같이 진행해주세요.(파이프 연산자 활용하여 한번에 처리)

```
test %<>%
  filter(!store_nbr %in% percent10) %>%
  filter(!family %in% family_list) %>%
  mutate_at(vars(all_of(category)),as.factor) %>%
  select(-date,-id,-dcoilwtico,-state,-type,-cluster)
```