data_preprocessing_final

데이터 파일 읽기

```
df1 <- read.csv("seoul_pharmacy.csv")
View(df1)
class(df1) # 읽어온 데이터 파일은 data.frame 형태임

## [1] "data.frame"
```

행정구 컬럼 만들기(행정구 추출)

- 5 명 모두 유사한 방식 활용

```
addr_li<- strsplit(df1$주소,split = " ")
head(addr_li,n = 3)

## [[1]]
## [1] "서울특별시" "강남구"     "강남대로"   "292"        "3 층"
## [6] "(도곡동"    "뱅뱅빌딩)"
##
## [[2]]
## [1] "서울특별시" "동대문구"   "전농로"     "60-1"       "1 층"
## [6] "(답십리동)"
##
## [[3]]
## [1] "서울특별시"    "강남구"        "봉은사로 114 길" "42"
## [5] "(삼성동)"

# 문자열 슬라이싱 결과 list 형태로 저장됨
# -> 각각의 벡터에서 두 번째 내부 원소가 구이다.
```

- 리스트의 경우 구성 벡터의 내부 원소에 접근하려면 [[ ]]을 이용하여야 함
- 각 벡터의 두 번째 요소에 접근하는 함수를 만들고, 각 행마다(각 벡터마다) 적용

```
# 두 번째 요소에 접근하는 함수 정의
search <- function(x){ # 두 번째 요소 찾기
  x[2]
}
```

```
# 각 행(각 벡터)마다 적용
df1$행정구 <- sapply(addr_li,search)
View(df1)
```

필요힌 정보만을 가져오기

- 가져와야 할 컬럼의 수 > 없앨 컬럼의 수
- 기존의 data frame 에서 약국 ID, 우편번호 1/2, 병원경도/위도, 작업일시 컬럼 삭제

```
# 불필요한 컬럼 삭제
df2 <- df1[,-c(1,21:25)]
View(df2)

# 보기 좋은 형태로 컬럼 순서 변경하기
df3 <- df2[,c(2,1,3,20,12,4,13,5,14,6,15,7,16,8,17,9,18,10,19,11)]
View(df3)

# 컬럼명 변경하기
new_colname <- c("약국명","주소","대표전화","행정구","시작(월)","마감(월)","시작(화)","마감(화)","시작(수)","마감(수)","시작(목)","마감(목)","시작(금)","마감(금)","시작(토)","마감(토)","시작(일)","마감(일)","시작(공휴일)","마감(공휴일)")
colnames(df3) <- new_colname
View(df3)
```

새로운 정보 저장을 위한 컬럼 생성

1. 일요일 운영 여부

```
# 권지수, 이수미 ------------------------------------------------------------
df3$일운영 <- (!is.na(df3$`시작(일)`)) & (!is.na(df3$`마감(일)`))
View(df3)
```

2. 공휴일 운영 여부

```
# 권지수, 이수미 ------------------------------------------------------------
df3$공휴일운영 <- (!is.na(df3$`시작(공휴일)`)) & (!is.na(df3$`마감(공휴일)`))
View(df3)
```

3. 야간 운영 여부

- 마감 시간에서 새벽 시간대를 표시하는 형식이 통일되어있지 않음 -> 새벽 3 시의 경우 2700(2400 + 300)으로 표시하는 형식 채택 -> 데이터 상의 모든 약국들이 530 <= 운영시간 <= 2929(2400 + 529)이도록 데이터 가공

a) 24 시간 운영하는 약국

- 찾아내는 방법 -> (마감) – (시작) = 0 or 2400
- 해당 약국들의 경우, 진료시작시간을 530 으로, 진료마감시간을 2929 으로 통일시켜주기

```
# 함수 선언
open_24 <- function(df,day){
  open_filter = paste0('시작(',day,')') # 해당 요일의 시작 시간 가져오기
  close_filter = paste0('마감(',day,')') # 해당 요일의 마감 시간 가져오기
  opentime <- df[,open_filter]
  closetime <- df[,close_filter]
  df[,open_filter] <- ifelse(((closetime - opentime == 0)|(closetime - openti
me == 2400)),530,opentime)
  df[,close_filter] <- ifelse(((closetime - opentime == 0)|(closetime - opent
ime == 2400)),2929,closetime)
  return(df)
}

# 함수 호출
df3 <- open_24(df3,"월")
df3 <- open_24(df3,"화")
df3 <- open_24(df3,"수")
df3 <- open_24(df3,"목")
df3 <- open_24(df3,"금")
df3 <- open_24(df3,"토")
df3 <- open_24(df3,"일")
df3 <- open_24(df3,"공휴일")


View(df3)
```

b) 새벽 운영을 하는 약국 처리해주기

```
# 함수 선언
midnight <- function(df,day){
  close_filter = paste0('마감(',day,')')
  closetime <- df[,close_filter]
  df[,close_filter] <- ifelse(((closetime >= 0)&(closetime < 530)),closetime
+ 2400,closetime)
  return(df)
}

# 함수 호출
df3 <- midnight(df3,"월")
df3 <- midnight(df3,"화")
df3 <- midnight(df3,"수")
df3 <- midnight(df3,"목")
df3 <- midnight(df3,"금")
df3 <- midnight(df3,"토")
df3 <- midnight(df3,"일")
df3 <- midnight(df3,"공휴일")

View(df3)
```

c) 야간운영여부 저장하기

- 강효은/차수빈/최은빈

```
# 함수 선언
open_midnight <- function(df,day){
  close_filter = paste0('마감(',day,')')
  closetime <- df[,close_filter]
  midnight_filter = paste0('야간운영(',day,')')
  # 강효은/차수빈/최은빈----------------------------
  df[,midnight_filter] <- ifelse(((closetime >= 2030)&(closetime <= 2929)),TR
UE,FALSE)
  return(df)
}
```

```
# 함수 호출
df3 <- open_midnight(df3,"월")
df3 <- open_midnight(df3,"화")
df3 <- open_midnight(df3,"수")
df3 <- open_midnight(df3,"목")
df3 <- open_midnight(df3,"금")
df3 <- open_midnight(df3,"토")
df3 <- open_midnight(df3,"일")
df3 <- open_midnight(df3,"공휴일")

View(df3)
```

- 일요일과 공휴일의 경우 운영을 하지 않는 약국들도 있음

  -> 운영시간이 NA 이기 때문에 비교 연산의 결과 NA 가 나온다.

  -> 해당 값들은 별도로 FALSE 처리를 해주어야 한다.

```
df3[which(is.na(df3$`시작(일)`)),"야간운영(일)"] <- FALSE
df3[which(is.na(df3$`시작(공휴일)`)),"야간운영(공휴일)"] <- FALSE

View(df3)
```

csv 파일로 내보내기

```
write.csv(df3,"C:/waterbean/Ewha/comp/project/seoul_pharmacy_final.csv",
          row.names = FALSE)
```