전체 데이터셋 정제

```
df_concat <- rbind(df_2016,df_2017,df_2018)
df_concat <- df_concat %>% filter(!is.na(시도))
df_concat <- df_concat %>% arrange(시도,일시)
```

```
df_concat_temp <- df_concat  %>%  group_by(시도,시도코드,일시) %>% summarise(`평균기온(°C)` = mean(`평균기온(°C)`
,na.rm=TRUE),
                                                                          `최저기온(°C)` = mean(`최저기온(°C)`,na.rm
=TRUE),
                                                                          `최고기온(°C)` = mean(`최고기온(°C)`,na.rm
=TRUE),
                                                                          `평균 풍속(m/s)` = mean(`평균 풍속(m/s)`,n
a.rm=TRUE),
                                                                          `평균 현지기압(hPa)` = mean(`평균 현지기압(
hPa)`,na.rm=TRUE),
                                                                          `일 최심신적설(cm)` = mean(`일 최심신적설(c
m)`,na.rm=TRUE),
                                                                          `일강수량(mm)` = mean(`일강수량(mm)`,na.rm
=TRUE),
                                                                          `강수 계속시간(hr)` = mean(`강수 계속시간(h
r)`,na.rm=TRUE),
                                                                          )
df_concat_temp
```

```
## # A tibble: 18,632 x 11
## # Groups:   시도, 시도코드 [17]
##    시도  시도코드 일시       `평균기온(°C)` `최저기온(°C)` `최고기온(°C)`
##    <chr>    <dbl> <date>              <dbl>          <dbl>          <dbl>
##  1 강원        42 2016-01-01          0.531         -4.65           4.93
##  2 강원        42 2016-01-02          5.18           1.48           9.05
##  3 강원        42 2016-01-03          4.92           0.869         10.3
##  4 강원        42 2016-01-04          3.95          -0.515          7.97
##  5 강원        42 2016-01-05         -1.07          -4.92           3.65
##  6 강원        42 2016-01-06         -2.12          -7.04           3.42
##  7 강원        42 2016-01-07         -2.53          -6.83           2.48
##  8 강원        42 2016-01-08         -3.69          -7.97           1.59
##  9 강원        42 2016-01-09         -1.95          -7.33           2.82
## 10 강원        42 2016-01-10          0.215         -4.23           4.87
## # ... with 18,622 more rows, and 5 more variables: `평균 풍속(m/s)` <dbl>, `평균
## #   현지기압(hPa)` <dbl>, `일 최심신적설(cm)` <dbl>, `일강수량(mm)` <dbl>, `강수
## #   계속시간(hr)` <dbl>
```

```
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 20 , 1 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 50 , 2 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 70 , 3 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 90 , 4 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 110 ,5 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 140 ,6 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 160 ,7 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 180 ,8 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 200 ,9 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 230 ,10 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 250 ,11 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 270 ,12 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 290 ,13 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 320 ,14 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 340 ,15 , df_concat $`최다풍향(16방위)` )
df_concat $`최다풍향(16방위)` <- ifelse(df_concat $`최다풍향(16방위)` == 360 ,16 , df_concat $`최다풍향(16방위)` )
```

```
#---------------------------------------
# 일별 16방위 빈도수 계산
#---------------------------------------


windlist <- df_concat$`최다풍향(16방위)`
daylist <- df_concat$일시
volumelist <- df_concat$`평균 풍속(m/s)`
```

```
beforeday <- substr(daylist[1],1,10)

beforewind <- windlist[1]
beforevolume <- volumelist[1]

windfreq <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
volumefreq <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)

windfreq[beforewind]<-windfreq[beforewind] + 1
volumefreq[beforewind]<-beforevolume

maxlist <- c();
end <- TRUE

for(i in (2:length(df_concat$`최다풍향(16방위)`))){
  nextday <- substr(daylist[i],1,10)
  nextwind <- windlist[i]
  nextvolume <- volumelist[i]


    if(beforeday == nextday){
      if(!is.na(nextwind)){
      windfreq[nextwind] <- windfreq[nextwind] + 1
      volumefreq[nextwind] <- volumefreq[nextwind] + nextvolume
      }
    }else{
      if(!is.na(nextwind)){
        windfreq[nextwind] <- windfreq[nextwind] + 1
        volumefreq[nextwind] <- volumefreq[nextwind] + nextvolume
      }
      if(i == length(df_concat$`최다풍향(16방위)`)){
        end <- FALSE
      }

      if(length(which(windfreq==max(windfreq)))>1){
        maxwind <- -1
        for(i in which(windfreq==max(windfreq))){
          if(maxwind<volumefreq[i]){
            maxwind <- volumefreq[i]
          }
        }
        maxlist <- append(maxlist,which(volumefreq==maxwind)[1])
      }else{
        maxlist <- append(maxlist,which(windfreq==max(windfreq))[1])
      }

      volumefreq <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
      windfreq <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
    }



  beforevolume <- nextvolume
  beforeday <- nextday
  beforewind <- nextwind
}


if(end){
  if(length(which(windfreq==max(windfreq)))>1){
    maxwind <- -1
    for(i in which(windfreq==max(windfreq))){
      if(maxwind<volumefreq[i]){
        maxwind <- volumefreq[i]
      }
    }
    maxlist <- append(maxlist,which(volumefreq==maxwind)[1])
  }else{
    maxlist <- append(maxlist,which(windfreq==max(windfreq))[1])
  }
}
```

```
df_concat_temp$`최다풍향(16방위)`<-maxlist
df_concat<-df_concat_temp

df_concat$일시 <- as.character(df_concat$일시)

df_concat
```

```
## # A tibble: 18,632 x 12
## # Groups:   시도, 시도코드 [17]
##    시도  시도코드 일시  `평균기온(°C)` `최저기온(°C)` `최고기온(°C)`
##    <chr>    <dbl> <chr>          <dbl>          <dbl>          <dbl>
##  1 강원        42 2016~          0.531         -4.65           4.93
##  2 강원        42 2016~          5.18           1.48           9.05
##  3 강원        42 2016~          4.92           0.869         10.3
##  4 강원        42 2016~          3.95          -0.515          7.97
##  5 강원        42 2016~         -1.07          -4.92           3.65
##  6 강원        42 2016~         -2.12          -7.04           3.42
##  7 강원        42 2016~         -2.53          -6.83           2.48
##  8 강원        42 2016~         -3.69          -7.97           1.59
##  9 강원        42 2016~         -1.95          -7.33           2.82
## 10 강원        42 2016~          0.215         -4.23           4.87
## # ... with 18,622 more rows, and 6 more variables: `평균 풍속(m/s)` <dbl>, `평균
## #   현지기압(hPa)` <dbl>, `일 최심신적설(cm)` <dbl>, `일강수량(mm)` <dbl>, `강수
## #   계속시간(hr)` <dbl>, `최다풍향(16방위)` <int>
```

```
#----------------------------------------
# 최다풍향 및 기타 널값 채우기
#----------------------------------------



temp <-df_concat %>%  filter(시도코드==26 & substr(일시,1,7)=='2016-12'&(!is.na(`평균 현지기압(hPa)`)))
temp <- mean(temp$`평균 현지기압(hPa)`)
df_concat$`평균 현지기압(hPa)` <-ifelse(is.na(df_concat$`평균 현지기압(hPa)`),temp,df_concat$`평균 현지기압(hPa)`)



temp <-df_concat %>%  filter(시도코드==11 & substr(일시,1,7)=='2017-10'&(!is.na(`평균 풍속(m/s)`)))
temp <- mean(temp$`평균 풍속(m/s)`)
df_concat$`평균 풍속(m/s)` <-ifelse(((df_concat$시도코드==11)&(df_concat$일시=='2017-10-14')),temp,df_concat$`평균 풍속(m/s)`)
temp <-df_concat %>%  filter(시도코드==11 & substr(일시,1,7)=='2017-12'&(!is.na(`평균 풍속(m/s)`)))
temp <- mean(temp$`평균 풍속(m/s)`)
df_concat$`평균 풍속(m/s)` <-ifelse(((df_concat$시도코드==11)&(df_concat$일시=='2017-12-05')),temp,df_concat$`평균 풍속(m/s)`)
df_concat$`평균 풍속(m/s)` <-ifelse(((df_concat$시도코드==11)&(df_concat$일시=='2017-12-06')),temp,df_concat$`평균 풍속(m/s)`)



temp <-df_concat %>%  filter(시도코드==11 & substr(일시,1,7)=='2017-10'&(!is.na(`최고기온(°C)`)))
temp <- mean(temp$`최고기온(°C)`)
df_concat$`최고기온(°C)` <-ifelse(((df_concat$시도코드==11)&(df_concat$일시=='2017-10-12')),temp,df_concat$`최고기온(°C)`)



temp <-df_concat %>%  filter(시도코드==27 & substr(일시,1,7)=='2017-07'&(!is.na(`평균기온(°C)`)))
temp <- mean(temp$`평균기온(°C)`)
df_concat$`평균기온(°C)` <-ifelse(((df_concat$시도코드==27)&(df_concat$일시=='2017-07-29')),temp,df_concat$`평균기온(°C)`)
```

```
df_concat <- df_concat %>% arrange(일시)
df_concat$일시 <- as.character(df_concat$일시)
df_concat$일시 <- as.Date(df_concat$일시,tryFormats = c("%Y-%m-%d", "%Y/%m/%d"))
#----------------------------------------
# 16방위 변환
#----------------------------------------

df_concat <- df_concat %>% arrange(시도,일시)
df_concat
```

```
## # A tibble: 18,632 x 12
## # Groups:   시도, 시도코드 [17]
##    시도  시도코드 일시       `평균기온(°C)` `최저기온(°C)` `최고기온(°C)`
##    <chr>    <dbl> <date>              <dbl>          <dbl>          <dbl>
##  1 강원        42 2016-01-01          0.531         -4.65           4.93
##  2 강원        42 2016-01-02          5.18           1.48           9.05
##  3 강원        42 2016-01-03          4.92           0.869         10.3
##  4 강원        42 2016-01-04          3.95          -0.515          7.97
##  5 강원        42 2016-01-05         -1.07          -4.92           3.65
##  6 강원        42 2016-01-06         -2.12          -7.04           3.42
##  7 강원        42 2016-01-07         -2.53          -6.83           2.48
##  8 강원        42 2016-01-08         -3.69          -7.97           1.59
##  9 강원        42 2016-01-09         -1.95          -7.33           2.82
## 10 강원        42 2016-01-10          0.215         -4.23           4.87
## # ... with 18,622 more rows, and 6 more variables: `평균 풍속(m/s)` <dbl>, `평균
## #   현지기압(hPa)` <dbl>, `일 최심신적설(cm)` <dbl>, `일강수량(mm)` <dbl>, `강수
## #   계속시간(hr)` <dbl>, `최다풍향(16방위)` <int>
```

```
write.csv(df_concat,"../../refinedata/weather/weather_data.csv")
```

```
dust_total <- dust_total %>% dplyr::select(-X1)
weather_concat  <- weather_concat  %>% dplyr::select(-X1)




#----------------------------------------
#미세먼지와 날씨 결합
#----------------------------------------

weather_concat$일시 <- as.character(weather_concat$일시)
weather_concat$일시 <- as.Date(weather_concat$일시,tryFormats = c("%Y-%m-%d", "%Y/%m/%d"))

dust_total$일시 <- as.character(dust_total$일시)
dust_total$일시 <- as.Date(dust_total$일시,tryFormats = c("%Y-%m-%d", "%Y/%m/%d"))

climate_total<- inner_join(weather_concat,dust_total,by=c("일시","시도"))

climate_total
```

```
## # A tibble: 18,632 x 18
##    시도  시도코드 일시       `평균기온(°C)` `최저기온(°C)` `최고기온(°C)`
##    <chr>    <dbl> <date>              <dbl>          <dbl>          <dbl>
##  1 강원        42 2016-01-01          0.531         -4.65           4.93
##  2 강원        42 2016-01-02          5.18           1.48           9.05
##  3 강원        42 2016-01-03          4.92           0.869         10.3
##  4 강원        42 2016-01-04          3.95          -0.515          7.97
##  5 강원        42 2016-01-05         -1.07          -4.92           3.65
##  6 강원        42 2016-01-06         -2.12          -7.04           3.42
##  7 강원        42 2016-01-07         -2.53          -6.83           2.48
##  8 강원        42 2016-01-08         -3.69          -7.97           1.59
##  9 강원        42 2016-01-09         -1.95          -7.33           2.82
## 10 강원        42 2016-01-10          0.215         -4.23           4.87
## # ... with 18,622 more rows, and 12 more variables: `평균 풍속(m/s)` <dbl>,
## #   `평균 현지기압(hPa)` <dbl>, `일 최심신적설(cm)` <dbl>,
## #   `일강수량(mm)` <dbl>, `강수 계속시간(hr)` <dbl>, `최다풍향(16방위)` <dbl>,
## #   SO2 <dbl>, CO <dbl>, O3 <dbl>, NO2 <dbl>, PM10 <dbl>, PM25 <dbl>
```

```
#---------------------------------------
# 데이터 전처리
#---------------------------------------

df_medical_total <- df_medical_total %>%  dplyr::select(성별코드,연령대코드,시도코드,일시)

#발생건수 카운트(성별,연령대 제외)
df_medical_total <- df_medical_total %>% group_by(시도코드,일시) %>% summarize(발생건수 = n())

df_medical_total$일시 <- as.character(df_medical_total$일시)
df_medical_total$일시 <- as.Date(df_medical_total$일시,tryFormats = c("%Y%m%d", "%Y/%m/%d"))

#기상데이터와 진료데이터 연결
analysis_total<- inner_join(df_medical_total,climate_total,by=c("일시","시도코드"))
analysis_total
```

```
## # A tibble: 18,632 x 19
## # Groups:   시도코드 [17]
##    시도코드 일시       발생건수 시도  `평균기온(°C)` `최저기온(°C)`
##       <dbl> <date>        <int> <chr>          <dbl>          <dbl>
##  1       11 2016-01-01      217 서울             1.2           -3.3
##  2       11 2016-01-02     2200 서울             5.7            1
##  3       11 2016-01-03      267 서울             6.5            5.1
##  4       11 2016-01-04     3244 서울             2             -2.5
##  5       11 2016-01-05     2163 서울            -2.7           -4.8
##  6       11 2016-01-06     2197 서울            -1.7           -4.9
##  7       11 2016-01-07     2134 서울            -3.4           -5.9
##  8       11 2016-01-08     2332 서울            -3.3           -6.9
##  9       11 2016-01-09     1824 서울            -2.1           -6.2
## 10       11 2016-01-10      207 서울             0.3           -2.7
## # ... with 18,622 more rows, and 13 more variables: `최고기온(°C)` <dbl>, `평균
## #   풍속(m/s)` <dbl>, `평균 현지기압(hPa)` <dbl>, `일 최심신적설(cm)` <dbl>,
## #   `일강수량(mm)` <dbl>, `강수 계속시간(hr)` <dbl>, `최다풍향(16방위)` <dbl>,
## #   SO2 <dbl>, CO <dbl>, O3 <dbl>, NO2 <dbl>, PM10 <dbl>, PM25 <dbl>
```

```
#---------------------------------------
#발병률 데이터 넣기
#---------------------------------------
population$name <- as.factor(population$`행정구역별(읍면동)`)


code <- c('42','41','43','44','30','47','48','45','46','11','28','27','31','29','26','49','36')
name <- c('강원도','경기도','충청북도','충청남도','대전광역시','경상북도','경상남도','전라북도','전라남도','서울특별시','
인천광역시','대구광역시','울산광역시','광주광역시','부산광역시','제주특별자치도','세종특별자치시')
df_sido <- data.frame("code"=code,"name"=name)

population <- inner_join(population,df_sido,by='name')

population <- population %>% dplyr::select(-`행정구역별(읍면동)`,-name)


pop <- melt(population, id.vars = c("연령별","code"))
```

```
## Warning in melt(population, id.vars = c("연령별", "code")): The melt generic
## in data.table has been passed a spec_tbl_df and will attempt to redirect to
## the relevant reshape2 method; please note that reshape2 is deprecated, and
## this redirection is now deprecated as well. To continue using melt methods from
## reshape2 while both libraries are attached, e.g. melt.list, you can prepend the
## namespace like reshape2::melt(population). In the next version, this warning
## will become an error.
```

```
pop <- pop %>% mutate(년도=as.factor(substr(variable,1,4)),
                      성별코드 = as.factor(substr(variable,6,8)),
                      연령대코드 = as.factor(연령별),
                      인구수 = value  )

pop <-pop %>% dplyr::select(-연령별,-variable,-value)

pop <- rename(pop,시도코드=code)


analysis_total$년도 <- substring(analysis_total$일시,1,4)


pop <- pop %>% group_by(년도, 시도코드) %>% summarise(인구수 = sum(인구수))


analysis_total$시도코드 <- as.factor(analysis_total$시도코드)

analysis_total <- inner_join(analysis_total, pop, by=c("시도코드","년도"))
```

```
## Warning: Column `년도` joining character vector and factor, coercing into
## character vector
```

```
analysis_total$발병률 <- analysis_total$발생건수/analysis_total$인구수*100

analysis_total_Fixed <- analysis_total

write.csv(analysis_total_Fixed,"../../refinedata/analysis/analysis_total_Fixed.csv")
save(analysis_total_Fixed,file="../../refinedata/analysis/analysis_total_Fixed.rda")
```