library(tidyverse)

library(stringr)

park = read\_csv("parking\_2017\_noNA.csv")

ticket.price = read\_csv("Violation\_Costs.csv",col\_names = FALSE)

names(ticket.price) <- c("Violation.code","ticket.price")

street.name = read\_csv("location-0224-street.csv")

# violation location

# location.na = which(is.na(park$`Violation Location`))

# length(location.na)

# park = park[-location.na,]

# violation location with duplicated county

#park %>%

# group\_by(`Violation Location`) %>%

# summarise(county = n\_distinct(`Violation County`))

dup <- c(103,114,102,113,100,94,1,10,12,17,26,23,48,50,63,70,67,68,69,81,75,76,77,71,72,60,61,62,52,40,41,42,44,45,46,47,20,9,4,5,90,83,84,66,88,108,104,105,106,110,111,112,120,122,123,121,118,128,147,115,107,109,101,87,91,96,97,93,99,6,7,8,2,3,16,14,11,21,22,18,19,27,28,24,25,30,32,34,43,49,37,73,74,78,79,80

)

park$`Violation Location` <- as.numeric(park$`Violation Location`)

vio = park %>%

dplyr::filter(`Violation Location` %in% dup) %>%

select(`Summons Number`,`Violation Location`,`Violation County`) %>%

count(`Violation Location`,`Violation County`) %>%

group\_by(`Violation Location`) %>%

dplyr::filter(n == max(n)) %>%

select(`Violation Location`,`Violation County`)

vio = vio[!duplicated(vio$`Violation Location`),]

vio <- rename(vio,unique.county = `Violation County`)

vio <- rename(vio,violation.location = `Violation Location`)

park <- rename(park,violation.location = `Violation Location`)

park = park %>% left\_join(vio,by='violation.location')

park = park %>%

mutate(unique.county = ifelse(is.na(unique.county),violation.location,unique.county))

############

fact = park

### dim.location ###

## Street name

fact <- rename(fact,street.name = `Street Name`)

street.name <- rename(street.name,street.name.clean = `Street Name` )

fact <- rename(fact,summons.id = `Summons Number`)

street.name <- rename(street.name,summons.id = `Summons Number`)

fact = fact %>% left\_join(street.name, by="summons.id")

fact = fact %>% select(-street.name) %>%

mutate(street.name = street.name.clean) %>%

select(-street.name.clean)

### dim.car ###

# color

fact$`Vehicle Color` <- str\_replace(fact$`Vehicle Color`,"^W[A-Z]+","WHITE")

fact$`Vehicle Color` <- str\_replace(fact$`Vehicle Color`,"[A-Z]+K$","BLACK")

fact$`Vehicle Color` <- str\_replace(fact$`Vehicle Color`,"^BLA[A-Z]\*$","BLACK")

fact$`Vehicle Color` <- str\_replace(fact$`Vehicle Color`,"^G[A-Z]\*Y$","GREY")

fact$`Vehicle Color` <- str\_replace(fact$`Vehicle Color`,"^BL$","BLACK")

fact$`Vehicle Color` <- str\_replace(fact$`Vehicle Color`,"^Y[A-Z]\*","YELLOW")

fact$`Vehicle Color` <- str\_replace(fact$`Vehicle Color`,"^R[A-Z]\*","RED")

fact$`Vehicle Color` <- str\_replace(fact$`Vehicle Color`,"^C[A-Z]\*","GREY")

fact$`Vehicle Color`[fact$`Vehicle Color` %in% c('SILVE','SILV','SIL','SL','SLVR')]<-'SILVER'

fact$`Vehicle Color`[fact$`Vehicle Color` %in% c('GLD','GD')]<-'GOLD'

fact$`Vehicle Color`[fact$`Vehicle Color` %in% c('BURG','MAROO','MR','BURGU')]<-'RED'

fact$`Vehicle Color`[fact$`Vehicle Color` %in% c('GRN','GN','G','GR','GREE')]<-'GREEN'

fact$`Vehicle Color`[fact$`Vehicle Color` %in% c('B','BLU')]<-'BLUE'

fact$`Vehicle Color`[fact$`Vehicle Color` %in% c('BR','BRN','BRO')]<-'BROWN'

fact$`Vehicle Color`[fact$`Vehicle Color` %in% c('TN')]<-'TAN'

fact$`Vehicle Color`[fact$`Vehicle Color` %in% c('ORANG')]<-'ORANGE'

fact$`Vehicle Color`[fact$`Vehicle Color` %in% c('W')]<-'WHITE'

fact$`Vehicle Color`[! fact$`Vehicle Color` %in% c('BLUE','RED','YELLOW','BLACK','GREEN','GREY','PURPL','BEIGE','ORANGE','TAN',

'BROWN','GOLD','SILVER')]<-'OTHER'

fact$unique.county[fact$unique.county %in% c('BX')]<-'BRONX'

fact$unique.county[fact$unique.county %in% c('K')]<-'BROOKLYN'

fact$unique.county[fact$unique.county %in% c('Q')]<-'QUEENS'

fact$unique.county[fact$unique.county %in% c('R')]<-'STATEN ISLAND'

fact$unique.county[fact$unique.county %in% c('NY')]<-'MANHATTAN'

fact$unique.county[! fact$unique.county %in% c('BRONX','BROOKLYN','QUEENS','STATEN ISLAND','MANHATTAN')]<-'OTHER'

# ticket price

fact <- rename(fact,Violation.code = `Violation Code`)

fact <- fact %>% left\_join(ticket.price,by="Violation.code")

# create dimension tables

dim.issuer = fact %>% dplyr::select(`Issuer Code`,`Issuer Precinct`,`Issuing Agency`,`Issuer Command`) %>% unique()

dim.car = fact %>% dplyr::select(`Vehicle Body Type`,`Plate Type`,`Registration State`,`Vehicle Color`,`Vehicle Year`,`Vehicle Make`) %>% unique()

dim.location = fact %>% dplyr::select(violation.location,unique.county,street.name,`Intersecting Street`,`Violation Precinct`) %>% unique()

# add natural key in dim.location

dim.location$key <- seq(1:nrow(dim.location))

join.key = c("violation.location","unique.county","street.name","Intersecting Street","Violation Precinct")

fact = fact %>% left\_join(dim.location,by=join.key)

# 24-format time

time\_to\_24 <- function(time){

time <- trimws(time)

hour <- as.integer(substr(time, 1,2))

hour <- hour%%24

minute <- as.integer(substr(time, 3,4))

minute <- minute%%60

minute <- ifelse(is.na(minute),0,minute)

am\_pm <- substr(time,5,5)

hour <- ifelse(am\_pm == 'P', hour+12, hour)

hour <- ifelse(hour >= 24, hour - 24, hour)

hour <- ifelse(hour >= 24, 0, hour)

hour <- str\_pad(hour,2, pad='0')

minute <- str\_pad(minute,2, pad='0')

out <- paste0(hour,':',minute)

out <- ifelse(is.na(hour) | is.na(minute), '00:00', out)

return(out)

}

# year to 2017

fact$`Issue Date` <- str\_replace(fact$`Issue Date`,'/.{4,4}$', '/2017')

fact$`Violation Time`<- time\_to\_24(fact$`Violation Time`)

# remove blank issue data

fact <- fact %>%

filter(!is.na(`Issue Date`))

# select columns from fact table

fact = fact %>% select(summons.id,`Violation Time`,`Issue Date`,`Issuer Code`,`Vehicle Body Type`,`Plate Type`,`Registration State`,`Vehicle Color`,

`Vehicle Year`,`Vehicle Make`,Violation.code,`Plate ID`,`House Number`,ticket.price,key)

names(dim.location) <- c("Violation location","unique county","street name","Intersecting Street","Violation Precinct"

,"key")

#write\_csv(park,"parking\_2017\_noNA.csv")

write\_csv(dim.issuer,"dim.issuer\_0224.csv",na='')

write\_csv(dim.car,"dim.car\_0224.csv",na='')

write\_csv(dim.location,"dim.location\_0224.csv",na='')

write\_csv(fact,"fact.table\_0224.csv",na='')

#SCD2

new\_row<-data.frame(dim.location[37,])

names(new\_row)<-names(dim.location)

new\_row$`Violation Precinct`<-102

dim.location<-rbind(dim.location,new\_row)

new\_row<-data.frame(dim.location[808,])

names(new\_row)<-names(dim.location)

new\_row$`Violation Precinct`<-30

dim.location<-rbind(dim.location,new\_row)

new\_row<-data.frame(dim.location[884,])

names(new\_row)<-names(dim.location)

new\_row$`Violation Precinct`<-45

dim.location<-rbind(dim.location,new\_row)

new\_row<-data.frame(dim.location[920,])

names(new\_row)<-names(dim.location)

new\_row$`Violation Precinct`<-25

dim.location<-rbind(dim.location,new\_row)

new\_row<-data.frame(dim.location[1055,])

names(new\_row)<-names(dim.location)

new\_row$`Violation Precinct`<-60

dim.location<-rbind(dim.location,new\_row)

##### Check #####

nrow(unique(dim.location))

nrow(unique(dim.car))

nrow(unique(dim.issuer))

apply(dim.car,2,n\_distinct)

apply(dim.location,2,n\_distinct)

apply(dim.issuer,2,n\_distinct)

str\_view(c("RE","REED","YRAS"),"^R[A-Z]\*")

str\_replace(black,"^G[A-Z]\*Y$","GRAY")

dim.car$`Vehicle Color`[str\_detect(dim.car$`Vehicle Color`,"[A-Z]+K$")]

unique(dim.car$`Vehicle Color`)

park\_check = read\_csv("parking\_2017\_noNA.csv")

#find column name

colnames(fact)[str\_detect(str\_to\_lower(colnames(fact)),"unique.county")]

#####

fact$Violation.code