Arrests 2010 Census Tracts analysis

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(Matrix)
```

```
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
rm(list = ls())
gc()
```

```
##           used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells 1975873 105.6    2927291 156.4  2927291 156.4
## Vcells 3424208  26.2    8388608  64.0  6537861  49.9
```

Preprocessing

```
# read the data
my.df = read.csv("../../data/core_datasets/arrests/arrests_2010_cta.csv", stringsAsFactors = T)
```

Remove date and NTA variables, convert to factor location, month and KY_CD.

```
my.df$ARREST_DATE = NULL
my.df$nta2020 = NULL
my.df$GeoID = NULL # redundant with geoid

my.df$geoid = factor(my.df$geoid)
```

```
my.df$MONTH = factor(my.df$MONTH)

# add NA category
my.df$KY_CD = ifelse(is.na(my.df$KY_CD), "MISSING", my.df$KY_CD)
my.df$KY_CD = factor(my.df$KY_CD)
```

```
str(my.df)
```

```
## 'data.frame':    419420 obs. of  14 variables:
##  $ KY_CD     : Factor w/ 70 levels "101","102","103",..: 70 4 70 15 70 70 70 4 70 4 ...
##  $ LAW_CAT_CD: Factor w/ 5 levels "","F","I","M",..: 1 2 4 2 4 2 4 2 4 2 ...
##  $ AGE_GROUP : Factor w/ 5 levels "<18","18-24",..: 3 2 3 1 4 4 3 2 2 3 ...
##  $ PERP_SEX  : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ PERP_RACE : Factor w/ 7 levels "AMERICAN INDIAN/ALASKAN NATIVE",..: 6 3 3 3 3 7 6 7 7 3 ...
##  $ geoid     : Factor w/ 2308 levels "36005000100",..: 236 636 1273 369 43 2187 983 1422 1592 1880 .
##  $ MONTH     : Factor w/ 12 levels "1","2","3","4",..: 1 11 3 12 12 12 11 11 10 9 ...
##  $ Pop1      : num  751 4544 183 2394 5337 ...
##  $ Male.P    : num  26.9 41.8 44.8 41.6 50.1 46.1 46.6 47.7 49.6 48.4 ...
##  $ MdAge     : num  20.4 31.4 30.8 32.2 30.4 33.8 36.8 34.2 35.1 37.3 ...
##  $ Hsp1P     : num  19.2 14 2.7 25 64.4 26.4 14.1 42.2 21.1 30.1 ...
##  $ WNHP      : num  61.5 1.2 39.9 19.8 3.2 20.5 3.8 8.8 53.3 17.7 ...
##  $ BNHP      : num  9.9 82.6 3.3 42.4 29 44.7 71.8 44.6 2.8 12.9 ...
##  $ ANHP      : num  7.1 0.2 49.7 10.3 1.8 4 7.6 1.5 19.1 33.1 ...
```

Check for NA

```
apply(my.df, 2, function(col) (sum(is.na(col))))
```

```
##      KY_CD LAW_CAT_CD  AGE_GROUP   PERP_SEX  PERP_RACE      geoid      MONTH
##          0          0          0          0          0          0          0
##       Pop1     Male.P      MdAge      Hsp1P       WNHP       BNHP       ANHP
##          0       3860       3796       4790       4410       4969       9686
```

A possibility is to get rid of all NAs rows, the portion of deleted rows would be relatively small (of course we're introducing some bias here).

```
nrow(na.omit(my.df)) / nrow(my.df)
```

```
## [1] 0.9752515
```

Other possibilities would be to impute values for numerical variables (using median, mean or more sofisticated methods). For simplicity we just delete missing values rows.

```
my.df = na.omit(my.df)
```

Description

Ideally 2010 data are our training set and 2011 data are the test set.

Variables Description

Explorative analysis

A reasonable response variable would be the count of crimes

Models

Model selection and issues

Here are present both spatial and temporal variables.