Compare DBF and server data for Rio de Janeiro Municipality

Marcelo Ferreira da Costa Gomes

3 de fevereiro de 2017

Loading data

We'll use dengue data from Rio de Janeiro Municipality from 2012 to 2016, taken from the DENGEON database.

```
library(foreign)
flist <- list.files('./data/', pattern='*.dbf', full.names = T)
df.dbf <- read.dbf(flist[1], as.is=T)
geocod <- 330455
for (fname in flist[2:(length(flist)-1)]){
  df.dbf <- rbind(df.dbf, read.dbf(fname, as.is=T))
}
df.dbf <- df.dbf[df.dbf$ID_MUNICIP == geocod, ]
df.dbf2016 <- read.dbf(flist[length(flist)], as.is=T)
df.dbf2016 <- df.dbf2016[df.dbf2016$ID_MUNICIP == geocod, ]
all(names(df.dbf2016) %in% names(df.dbf))
```

```
## [1] FALSE
```

The dataset from 2016 have different set of columns than those from 2012-2015. Filter by list of columns of interest, append datasets and drop possible duplicates:

```
filter.cols <- c('NU_NOTIFIC', 'ID_MUNICIP', 'ID_UNIDADE', 'DT_NOTIFIC')
df.dbf.clean <- df.dbf[, c(filter.cols, 'DT_DIGITA')]
df.dbf2016.clean <- df.dbf2016[, c(filter.cols, 'DT_DIGITA')]
all(names(df.dbf2016.clean) %in% names(df.dbf.clean))
```

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

```
df.dbf.clean <- rbind(df.dbf.clean, df.dbf2016.clean)

nrow(df.dbf.clean[duplicated(df.dbf.clean[, filter.cols]), ])
```

```
## [1] 0
```

We can see above that there are no duplicates in the DBFs.

Read data downloaded from server:

```
df.server <- readRDS('data/dengue.munRJ.2012.2016.v2.rds')
names(df.server)
```

```
## [1] "municipio_geocodigo" "dt_notific"          "dt_digita"
## [4] "nu_notific"
```

```
filter.cols <- c('municipio_geocodigo', 'nu_notific', 'dt_notific')

nrow(df.server[duplicated(df.server[, filter.cols]), ])
```

```
## [1] 0
```

This particular dataset have already been cleaned of duplicates regarding those columns, as well as against rows with empty digitization date.

Compare datasets:

```
nrow(df.dbf.clean) - nrow(df.server)
```

```
## [1] -1616
```

```
summary(df.dbf.clean$DT_NOTIFIC)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2012-01-01" "2012-04-18" "2012-06-01" "2013-03-04" "2013-04-19"
##         Max.
## "2016-12-31"
```

```
summary(df.server$dt_notific)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2012-01-01" "2012-04-19" "2012-06-04" "2013-03-16" "2013-04-24"
##         Max.
## "2016-12-31"
```

```
summary(df.dbf.clean$DT_DIGITA)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2012-01-03" "2012-05-10" "2012-07-16" "2013-03-29" "2013-05-15"
##         Max.         NA's
## "2017-02-01"        "212"
```

```
summary(df.server$dt_digita)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2012-01-03" "2012-05-11" "2012-07-19" "2013-04-11" "2013-05-21"
##         Max.
## "2017-02-06"
```

As seen from the tests above, there are a few discrepancies between the datasets, such as the presence of NA in DT_DIGITA on the dbf and the last input in the server is more recent then that on the dbf. The data from the server has already been filtered regarding NAs, we must do the same in the dbf and discard data entered after 2017-02-01 in order to have a proper comparison. Unfortunalely, even after those procedures we still have more entries in the server than in the dbf's. In fact, the difference is even bigger now.

```
df.dbf.clean <- df.dbf.clean[!is.na(df.dbf.clean$DT_DIGITA), ]
df.server <- df.server[df.server$dt_digita <= '2017-02-01', ]

nrow(df.dbf.clean) - nrow(df.server)
```

```
## [1] -1804
```

```
summary(df.dbf.clean$DT_NOTIFIC)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2012-01-01" "2012-04-18" "2012-06-01" "2013-03-03" "2013-04-19"
##         Max.
## "2016-12-31"
```

```
summary(df.server$dt_notific)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2012-01-01" "2012-04-19" "2012-06-04" "2013-03-16" "2013-04-24"
```

```
##         Max.
## "2016-12-31"
```

```
summary(df.dbf.clean$DT_DIGITA)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2012-01-03" "2012-05-10" "2012-07-16" "2013-03-29" "2013-05-15"
##         Max.
## "2017-02-01"
```

```
summary(df.server$dt_digita)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2012-01-03" "2012-05-11" "2012-07-19" "2013-04-11" "2013-05-21"
##         Max.
## "2017-02-01"
```