Step by step description of homecredit_wrangling.R

P. A. Ortiz Otalvaro

3 September 2019

1. Load packages

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.6.1
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

2. Read data

For now I am working with the two main files *application_train.csv* and *application_test.csv* in order to define a starting point to improve upon later on. Data frames `homecredit_train_df` and `homecredit_test_df` contain the data from these two files respectively.

During the reading process, all blank and empty observations are replaced with NA.

```
# ********** Load data (convert any blank or empty entries into NA)*****************
homecredit_train_df <- read.csv("../LoanData_HomeCredit/application_train.csv", sep=",",
                                stringsAsFactors = FALSE, na.strings = c("","  ",NA))
homecredit_test_df <- read.csv("../LoanData_HomeCredit/application_test.csv", sep=",",
                               stringsAsFactors = FALSE, na.strings = c("","  ",NA))


# bureau <- read.csv("../LoanData_HomeCredit/bureau.csv", sep=",",
#                    stringsAsFactors = FALSE, na.strings = c("","  ",NA))
#
# bureau_balance <- read.csv("../LoanData_HomeCredit/bureau_balance.csv", sep=",",
#                    stringsAsFactors = FALSE, na.strings = c("","  ",NA))
#
# previousapp <- read.csv("../LoanData_HomeCredit/previous_application.csv", sep=",",
#                    stringsAsFactors = FALSE, na.strings = c("","  ",NA))
#
# poscashbalance <- read.csv("../LoanData_HomeCredit/POS_CASH_balance.csv", sep=",",
```

```
#                     stringsAsFactors = FALSE, na.strings = c("","  ",NA))
#
# installments <- read.csv("../LoanData_HomeCredit/installments_payments.csv", sep=",",
#                     stringsAsFactors = FALSE, na.strings = c("","  ",NA))
#
# creditcard <- read.csv("../LoanData_HomeCredit/credit_card_balance.csv", sep=",",
#                     stringsAsFactors = FALSE, na.strings = c("","  ",NA))
```

	Train	Test
Number of observations	307511	48744
Number of features	123	122
Number of duplicated observations	0	0
Number of missing values	9152465	1404419
Total features of char type	17	17
Total features of factor type	0	0
Total features of numeric type	106	105