preprocess

2023-04-19

```
library(ggplot2)
library(reshape2)
library(dslabs)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## -- 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(dplyr)
library(ggcorrplot)
library(tidyverse)
library(lares)
```

```
application_train <- read_csv("./application_train.csv")
```

```
## Rows: 307511 Columns: 122
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (16): NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, N...
## dbl (106): SK_ID_CURR, TARGET, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, A...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(application_train)
```

```
## # A tibble: 6 x 122
##   SK_ID_CURR TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY
##        <dbl>  <dbl> <chr>              <chr>       <chr>        <chr>
## 1     100002      1 Cash loans         M           N            Y
## 2     100003      0 Cash loans         F           N            N
## 3     100004      0 Revolving loans    M           Y            Y
## 4     100006      0 Cash loans         F           N            Y
## 5     100007      0 Cash loans         M           N            Y
## 6     100008      0 Cash loans         M           N            Y
## # i 116 more variables: CNT_CHILDREN <dbl>, AMT_INCOME_TOTAL <dbl>,
## #   AMT_CREDIT <dbl>, AMT_ANNUITY <dbl>, AMT_GOODS_PRICE <dbl>,
## #   NAME_TYPE_SUITE <chr>, NAME_INCOME_TYPE <chr>, NAME_EDUCATION_TYPE <chr>,
## #   NAME_FAMILY_STATUS <chr>, NAME_HOUSING_TYPE <chr>,
## #   REGION_POPULATION_RELATIVE <dbl>, DAYS_BIRTH <dbl>, DAYS_EMPLOYED <dbl>,
## #   DAYS_REGISTRATION <dbl>, DAYS_ID_PUBLISH <dbl>, OWN_CAR_AGE <dbl>,
## #   FLAG_MOBIL <dbl>, FLAG_EMP_PHONE <dbl>, FLAG_WORK_PHONE <dbl>, ...
```

```
dim(application_train)
```

```
## [1] 307511    122
```

```
perc_na <- as.data.frame(round((colSums(is.na(application_train))/nrow(application_train))*100, 2))
perc_na <- perc_na|> rename("perc" = "round((colSums(is.na(application_train))/nrow(application_train))
perc_na
```

```
##                               perc
## SK_ID_CURR                    0.00
## TARGET                        0.00
## NAME_CONTRACT_TYPE            0.00
## CODE_GENDER                   0.00
## FLAG_OWN_CAR                  0.00
## FLAG_OWN_REALTY               0.00
## CNT_CHILDREN                  0.00
## AMT_INCOME_TOTAL              0.00
## AMT_CREDIT                    0.00
## AMT_ANNUITY                   0.00
## AMT_GOODS_PRICE               0.09
## NAME_TYPE_SUITE               0.42
## NAME_INCOME_TYPE              0.00
## NAME_EDUCATION_TYPE           0.00
## NAME_FAMILY_STATUS            0.00
## NAME_HOUSING_TYPE             0.00
## REGION_POPULATION_RELATIVE    0.00
## DAYS_BIRTH                    0.00
## DAYS_EMPLOYED                 0.00
## DAYS_REGISTRATION             0.00
## DAYS_ID_PUBLISH               0.00
## OWN_CAR_AGE                  65.99
## FLAG_MOBIL                    0.00
## FLAG_EMP_PHONE                0.00
## FLAG_WORK_PHONE               0.00
## FLAG_CONT_MOBILE              0.00
## FLAG_PHONE                    0.00
## FLAG_EMAIL                    0.00
## OCCUPATION_TYPE              31.35
## CNT_FAM_MEMBERS               0.00
## REGION_RATING_CLIENT          0.00
## REGION_RATING_CLIENT_W_CITY   0.00
## WEEKDAY_APPR_PROCESS_START    0.00
## HOUR_APPR_PROCESS_START       0.00
## REG_REGION_NOT_LIVE_REGION    0.00
## REG_REGION_NOT_WORK_REGION    0.00
## LIVE_REGION_NOT_WORK_REGION   0.00
## REG_CITY_NOT_LIVE_CITY        0.00
## REG_CITY_NOT_WORK_CITY        0.00
## LIVE_CITY_NOT_WORK_CITY       0.00
## ORGANIZATION_TYPE             0.00
## EXT_SOURCE_1                 56.38
## EXT_SOURCE_2                  0.21
## EXT_SOURCE_3                 19.83
## APARTMENTS_AVG               50.75
## BASEMENTAREA_AVG             58.52
## YEARS_BEGINEXPLUATATION_AVG  48.78
## YEARS_BUILD_AVG              66.50
## COMMONAREA_AVG               69.87
## ELEVATORS_AVG                53.30
## ENTRANCES_AVG                50.35
## FLOORSMAX_AVG                49.76
## FLOORSMIN_AVG                67.85
## LANDAREA_AVG                 59.38
## LIVINGAPARTMENTS_AVG         68.35
## LIVINGAREA_AVG               50.19
## NONLIVINGAPARTMENTS_AVG      69.43
## NONLIVINGAREA_AVG            55.18
## APARTMENTS_MODE              50.75
## BASEMENTAREA_MODE            58.52
## YEARS_BEGINEXPLUATATION_MODE 48.78
## YEARS_BUILD_MODE             66.50
## COMMONAREA_MODE              69.87
## ELEVATORS_MODE               53.30
## ENTRANCES_MODE               50.35
## FLOORSMAX_MODE               49.76
## FLOORSMIN_MODE               67.85
## LANDAREA_MODE                59.38
## LIVINGAPARTMENTS_MODE        68.35
## LIVINGAREA_MODE              50.19
## NONLIVINGAPARTMENTS_MODE     69.43
## NONLIVINGAREA_MODE           55.18
## APARTMENTS_MEDI              50.75
## BASEMENTAREA_MEDI            58.52
## YEARS_BEGINEXPLUATATION_MEDI 48.78
## YEARS_BUILD_MEDI             66.50
## COMMONAREA_MEDI              69.87
## ELEVATORS_MEDI               53.30
## ENTRANCES_MEDI               50.35
## FLOORSMAX_MEDI               49.76
## FLOORSMIN_MEDI               67.85
## LANDAREA_MEDI                59.38
## LIVINGAPARTMENTS_MEDI        68.35
## LIVINGAREA_MEDI              50.19
## NONLIVINGAPARTMENTS_MEDI     69.43
## NONLIVINGAREA_MEDI           55.18
## FONDKAPREMONT_MODE           68.39
## HOUSETYPE_MODE               50.18
## TOTALAREA_MODE               48.27
## WALLSMATERIAL_MODE           50.84
## EMERGENCYSTATE_MODE          47.40
## OBS_30_CNT_SOCIAL_CIRCLE      0.33
## DEF_30_CNT_SOCIAL_CIRCLE      0.33
## OBS_60_CNT_SOCIAL_CIRCLE      0.33
## DEF_60_CNT_SOCIAL_CIRCLE      0.33
## DAYS_LAST_PHONE_CHANGE        0.00
## FLAG_DOCUMENT_2               0.00
## FLAG_DOCUMENT_3               0.00
## FLAG_DOCUMENT_4               0.00
## FLAG_DOCUMENT_5               0.00
## FLAG_DOCUMENT_6               0.00
## FLAG_DOCUMENT_7               0.00
## FLAG_DOCUMENT_8               0.00
## FLAG_DOCUMENT_9               0.00
## FLAG_DOCUMENT_10              0.00
## FLAG_DOCUMENT_11              0.00
## FLAG_DOCUMENT_12              0.00
## FLAG_DOCUMENT_13              0.00
## FLAG_DOCUMENT_14              0.00
## FLAG_DOCUMENT_15              0.00
## FLAG_DOCUMENT_16              0.00
## FLAG_DOCUMENT_17              0.00
## FLAG_DOCUMENT_18              0.00
## FLAG_DOCUMENT_19              0.00
## FLAG_DOCUMENT_20              0.00
## FLAG_DOCUMENT_21              0.00
## AMT_REQ_CREDIT_BUREAU_HOUR   13.50
## AMT_REQ_CREDIT_BUREAU_DAY    13.50
## AMT_REQ_CREDIT_BUREAU_WEEK   13.50
## AMT_REQ_CREDIT_BUREAU_MON    13.50
## AMT_REQ_CREDIT_BUREAU_QRT    13.50
## AMT_REQ_CREDIT_BUREAU_YEAR   13.50
```

```
perc_na$names <- rownames(perc_na)
perc_na
```

```
##                               perc                        names
## SK_ID_CURR                    0.00                   SK_ID_CURR
## TARGET                        0.00                       TARGET
## NAME_CONTRACT_TYPE            0.00           NAME_CONTRACT_TYPE
## CODE_GENDER                   0.00                  CODE_GENDER
## FLAG_OWN_CAR                  0.00                 FLAG_OWN_CAR
## FLAG_OWN_REALTY               0.00              FLAG_OWN_REALTY
## CNT_CHILDREN                  0.00                 CNT_CHILDREN
## AMT_INCOME_TOTAL              0.00             AMT_INCOME_TOTAL
## AMT_CREDIT                    0.00                   AMT_CREDIT
## AMT_ANNUITY                   0.00                  AMT_ANNUITY
## AMT_GOODS_PRICE               0.09              AMT_GOODS_PRICE
## NAME_TYPE_SUITE               0.42              NAME_TYPE_SUITE
## NAME_INCOME_TYPE              0.00             NAME_INCOME_TYPE
## NAME_EDUCATION_TYPE           0.00          NAME_EDUCATION_TYPE
## NAME_FAMILY_STATUS            0.00           NAME_FAMILY_STATUS
## NAME_HOUSING_TYPE             0.00            NAME_HOUSING_TYPE
## REGION_POPULATION_RELATIVE    0.00   REGION_POPULATION_RELATIVE
## DAYS_BIRTH                    0.00                   DAYS_BIRTH
## DAYS_EMPLOYED                 0.00                DAYS_EMPLOYED
## DAYS_REGISTRATION             0.00            DAYS_REGISTRATION
## DAYS_ID_PUBLISH               0.00              DAYS_ID_PUBLISH
## OWN_CAR_AGE                  65.99                  OWN_CAR_AGE
## FLAG_MOBIL                    0.00                   FLAG_MOBIL
## FLAG_EMP_PHONE                0.00               FLAG_EMP_PHONE
## FLAG_WORK_PHONE               0.00              FLAG_WORK_PHONE
## FLAG_CONT_MOBILE              0.00             FLAG_CONT_MOBILE
## FLAG_PHONE                    0.00                   FLAG_PHONE
## FLAG_EMAIL                    0.00                   FLAG_EMAIL
## OCCUPATION_TYPE              31.35              OCCUPATION_TYPE
## CNT_FAM_MEMBERS               0.00              CNT_FAM_MEMBERS
## REGION_RATING_CLIENT          0.00         REGION_RATING_CLIENT
## REGION_RATING_CLIENT_W_CITY   0.00  REGION_RATING_CLIENT_W_CITY
## WEEKDAY_APPR_PROCESS_START    0.00   WEEKDAY_APPR_PROCESS_START
## HOUR_APPR_PROCESS_START       0.00      HOUR_APPR_PROCESS_START
## REG_REGION_NOT_LIVE_REGION    0.00   REG_REGION_NOT_LIVE_REGION
## REG_REGION_NOT_WORK_REGION    0.00   REG_REGION_NOT_WORK_REGION
## LIVE_REGION_NOT_WORK_REGION   0.00  LIVE_REGION_NOT_WORK_REGION
## REG_CITY_NOT_LIVE_CITY        0.00       REG_CITY_NOT_LIVE_CITY
## REG_CITY_NOT_WORK_CITY        0.00       REG_CITY_NOT_WORK_CITY
## LIVE_CITY_NOT_WORK_CITY       0.00      LIVE_CITY_NOT_WORK_CITY
## ORGANIZATION_TYPE             0.00            ORGANIZATION_TYPE
## EXT_SOURCE_1                 56.38                 EXT_SOURCE_1
## EXT_SOURCE_2                  0.21                 EXT_SOURCE_2
## EXT_SOURCE_3                 19.83                 EXT_SOURCE_3
## APARTMENTS_AVG               50.75               APARTMENTS_AVG
## BASEMENTAREA_AVG             58.52             BASEMENTAREA_AVG
## YEARS_BEGINEXPLUATATION_AVG  48.78  YEARS_BEGINEXPLUATATION_AVG
## YEARS_BUILD_AVG              66.50              YEARS_BUILD_AVG
## COMMONAREA_AVG               69.87               COMMONAREA_AVG
## ELEVATORS_AVG                53.30                ELEVATORS_AVG
## ENTRANCES_AVG                50.35                ENTRANCES_AVG
## FLOORSMAX_AVG                49.76                FLOORSMAX_AVG
## FLOORSMIN_AVG                67.85                FLOORSMIN_AVG
## LANDAREA_AVG                 59.38                 LANDAREA_AVG
## LIVINGAPARTMENTS_AVG         68.35         LIVINGAPARTMENTS_AVG
## LIVINGAREA_AVG               50.19               LIVINGAREA_AVG
## NONLIVINGAPARTMENTS_AVG      69.43      NONLIVINGAPARTMENTS_AVG
## NONLIVINGAREA_AVG            55.18            NONLIVINGAREA_AVG
## APARTMENTS_MODE              50.75              APARTMENTS_MODE
## BASEMENTAREA_MODE            58.52            BASEMENTAREA_MODE
## YEARS_BEGINEXPLUATATION_MODE 48.78 YEARS_BEGINEXPLUATATION_MODE
## YEARS_BUILD_MODE             66.50             YEARS_BUILD_MODE
## COMMONAREA_MODE              69.87              COMMONAREA_MODE
## ELEVATORS_MODE               53.30               ELEVATORS_MODE
## ENTRANCES_MODE               50.35               ENTRANCES_MODE
## FLOORSMAX_MODE               49.76               FLOORSMAX_MODE
## FLOORSMIN_MODE               67.85               FLOORSMIN_MODE
## LANDAREA_MODE                59.38                LANDAREA_MODE
## LIVINGAPARTMENTS_MODE        68.35        LIVINGAPARTMENTS_MODE
## LIVINGAREA_MODE              50.19              LIVINGAREA_MODE
## NONLIVINGAPARTMENTS_MODE     69.43     NONLIVINGAPARTMENTS_MODE
## NONLIVINGAREA_MODE           55.18           NONLIVINGAREA_MODE
## APARTMENTS_MEDI              50.75              APARTMENTS_MEDI
## BASEMENTAREA_MEDI            58.52            BASEMENTAREA_MEDI
## YEARS_BEGINEXPLUATATION_MEDI 48.78 YEARS_BEGINEXPLUATATION_MEDI
## YEARS_BUILD_MEDI             66.50             YEARS_BUILD_MEDI
## COMMONAREA_MEDI              69.87              COMMONAREA_MEDI
## ELEVATORS_MEDI               53.30               ELEVATORS_MEDI
## ENTRANCES_MEDI               50.35               ENTRANCES_MEDI
## FLOORSMAX_MEDI               49.76               FLOORSMAX_MEDI
## FLOORSMIN_MEDI               67.85               FLOORSMIN_MEDI
## LANDAREA_MEDI                59.38                LANDAREA_MEDI
## LIVINGAPARTMENTS_MEDI        68.35        LIVINGAPARTMENTS_MEDI
## LIVINGAREA_MEDI              50.19              LIVINGAREA_MEDI
## NONLIVINGAPARTMENTS_MEDI     69.43     NONLIVINGAPARTMENTS_MEDI
## NONLIVINGAREA_MEDI           55.18           NONLIVINGAREA_MEDI
## FONDKAPREMONT_MODE           68.39           FONDKAPREMONT_MODE
## HOUSETYPE_MODE               50.18               HOUSETYPE_MODE
## TOTALAREA_MODE               48.27               TOTALAREA_MODE
## WALLSMATERIAL_MODE           50.84           WALLSMATERIAL_MODE
## EMERGENCYSTATE_MODE          47.40          EMERGENCYSTATE_MODE
## OBS_30_CNT_SOCIAL_CIRCLE      0.33     OBS_30_CNT_SOCIAL_CIRCLE
## DEF_30_CNT_SOCIAL_CIRCLE      0.33     DEF_30_CNT_SOCIAL_CIRCLE
## OBS_60_CNT_SOCIAL_CIRCLE      0.33     OBS_60_CNT_SOCIAL_CIRCLE
## DEF_60_CNT_SOCIAL_CIRCLE      0.33     DEF_60_CNT_SOCIAL_CIRCLE
## DAYS_LAST_PHONE_CHANGE        0.00       DAYS_LAST_PHONE_CHANGE
## FLAG_DOCUMENT_2               0.00              FLAG_DOCUMENT_2
## FLAG_DOCUMENT_3               0.00              FLAG_DOCUMENT_3
## FLAG_DOCUMENT_4               0.00              FLAG_DOCUMENT_4
## FLAG_DOCUMENT_5               0.00              FLAG_DOCUMENT_5
## FLAG_DOCUMENT_6               0.00              FLAG_DOCUMENT_6
## FLAG_DOCUMENT_7               0.00              FLAG_DOCUMENT_7
## FLAG_DOCUMENT_8               0.00              FLAG_DOCUMENT_8
## FLAG_DOCUMENT_9               0.00              FLAG_DOCUMENT_9
## FLAG_DOCUMENT_10              0.00             FLAG_DOCUMENT_10
## FLAG_DOCUMENT_11              0.00             FLAG_DOCUMENT_11
## FLAG_DOCUMENT_12              0.00             FLAG_DOCUMENT_12
## FLAG_DOCUMENT_13              0.00             FLAG_DOCUMENT_13
## FLAG_DOCUMENT_14              0.00             FLAG_DOCUMENT_14
## FLAG_DOCUMENT_15              0.00             FLAG_DOCUMENT_15
## FLAG_DOCUMENT_16              0.00             FLAG_DOCUMENT_16
## FLAG_DOCUMENT_17              0.00             FLAG_DOCUMENT_17
## FLAG_DOCUMENT_18              0.00             FLAG_DOCUMENT_18
## FLAG_DOCUMENT_19              0.00             FLAG_DOCUMENT_19
## FLAG_DOCUMENT_20              0.00             FLAG_DOCUMENT_20
## FLAG_DOCUMENT_21              0.00             FLAG_DOCUMENT_21
## AMT_REQ_CREDIT_BUREAU_HOUR   13.50   AMT_REQ_CREDIT_BUREAU_HOUR
## AMT_REQ_CREDIT_BUREAU_DAY    13.50    AMT_REQ_CREDIT_BUREAU_DAY
## AMT_REQ_CREDIT_BUREAU_WEEK   13.50   AMT_REQ_CREDIT_BUREAU_WEEK
## AMT_REQ_CREDIT_BUREAU_MON    13.50    AMT_REQ_CREDIT_BUREAU_MON
## AMT_REQ_CREDIT_BUREAU_QRT    13.50    AMT_REQ_CREDIT_BUREAU_QRT
## AMT_REQ_CREDIT_BUREAU_YEAR   13.50   AMT_REQ_CREDIT_BUREAU_YEAR
```

```r
perc_na_mt_40 <- perc_na |> filter(perc > 40)
row.names(perc_na_mt_40) <- NULL
perc_na_mt_40
```

```
##     perc                        names
## 1  65.99                  OWN_CAR_AGE
## 2  56.38                 EXT_SOURCE_1
## 3  50.75               APARTMENTS_AVG
## 4  58.52             BASEMENTAREA_AVG
## 5  48.78  YEARS_BEGINEXPLUATATION_AVG
## 6  66.50              YEARS_BUILD_AVG
## 7  69.87               COMMONAREA_AVG
## 8  53.30                ELEVATORS_AVG
## 9  50.35                ENTRANCES_AVG
## 10 49.76                FLOORSMAX_AVG
## 11 67.85                FLOORSMIN_AVG
## 12 59.38                 LANDAREA_AVG
## 13 68.35         LIVINGAPARTMENTS_AVG
## 14 50.19               LIVINGAREA_AVG
## 15 69.43      NONLIVINGAPARTMENTS_AVG
## 16 55.18            NONLIVINGAREA_AVG
## 17 50.75              APARTMENTS_MODE
## 18 58.52            BASEMENTAREA_MODE
## 19 48.78 YEARS_BEGINEXPLUATATION_MODE
## 20 66.50             YEARS_BUILD_MODE
## 21 69.87              COMMONAREA_MODE
## 22 53.30               ELEVATORS_MODE
## 23 50.35               ENTRANCES_MODE
## 24 49.76               FLOORSMAX_MODE
## 25 67.85               FLOORSMIN_MODE
## 26 59.38                LANDAREA_MODE
## 27 68.35        LIVINGAPARTMENTS_MODE
## 28 50.19              LIVINGAREA_MODE
## 29 69.43     NONLIVINGAPARTMENTS_MODE
## 30 55.18           NONLIVINGAREA_MODE
## 31 50.75              APARTMENTS_MEDI
## 32 58.52            BASEMENTAREA_MEDI
## 33 48.78 YEARS_BEGINEXPLUATATION_MEDI
## 34 66.50             YEARS_BUILD_MEDI
## 35 69.87              COMMONAREA_MEDI
## 36 53.30               ELEVATORS_MEDI
## 37 50.35               ENTRANCES_MEDI
## 38 49.76               FLOORSMAX_MEDI
## 39 67.85               FLOORSMIN_MEDI
## 40 59.38                LANDAREA_MEDI
## 41 68.35        LIVINGAPARTMENTS_MEDI
## 42 50.19              LIVINGAREA_MEDI
## 43 69.43     NONLIVINGAPARTMENTS_MEDI
## 44 55.18           NONLIVINGAREA_MEDI
## 45 68.39           FONDKAPREMONT_MODE
## 46 50.18               HOUSETYPE_MODE
## 47 48.27               TOTALAREA_MODE
## 48 50.84           WALLSMATERIAL_MODE
## 49 47.40          EMERGENCYSTATE_MODE
```

```r
perc_na_mt_40 <- perc_na_mt_40 |> pivot_wider(
    names_from = names,
    values_from = perc,
  )

perc_na_mt_40 <- as.data.frame(perc_na_mt_40)
perc_na_mt_40
```

```
##   OWN_CAR_AGE EXT_SOURCE_1 APARTMENTS_AVG BASEMENTAREA_AVG
## 1       65.99        56.38          50.75            58.52
##   YEARS_BEGINEXPLUATATION_AVG YEARS_BUILD_AVG COMMONAREA_AVG ELEVATORS_AVG
## 1                       48.78            66.5          69.87          53.3
##   ENTRANCES_AVG FLOORSMAX_AVG FLOORSMIN_AVG LANDAREA_AVG LIVINGAPARTMENTS_AVG
## 1         50.35         49.76         67.85        59.38                68.35
##   LIVINGAREA_AVG NONLIVINGAPARTMENTS_AVG NONLIVINGAREA_AVG APARTMENTS_MODE
## 1          50.19                   69.43             55.18           50.75
##   BASEMENTAREA_MODE YEARS_BEGINEXPLUATATION_MODE YEARS_BUILD_MODE
## 1             58.52                        48.78             66.5
##   COMMONAREA_MODE ELEVATORS_MODE ENTRANCES_MODE FLOORSMAX_MODE FLOORSMIN_MODE
## 1           69.87           53.3          50.35          49.76          67.85
##   LANDAREA_MODE LIVINGAPARTMENTS_MODE LIVINGAREA_MODE NONLIVINGAPARTMENTS_MODE
## 1         59.38                 68.35           50.19                    69.43
##   NONLIVINGAREA_MODE APARTMENTS_MEDI BASEMENTAREA_MEDI
## 1              55.18           50.75             58.52
##   YEARS_BEGINEXPLUATATION_MEDI YEARS_BUILD_MEDI COMMONAREA_MEDI ELEVATORS_MEDI
## 1                        48.78             66.5           69.87           53.3
##   ENTRANCES_MEDI FLOORSMAX_MEDI FLOORSMIN_MEDI LANDAREA_MEDI
## 1          50.35          49.76          67.85         59.38
##   LIVINGAPARTMENTS_MEDI LIVINGAREA_MEDI NONLIVINGAPARTMENTS_MEDI
## 1                 68.35           50.19                    69.43
##   NONLIVINGAREA_MEDI FONDKAPREMONT_MODE HOUSETYPE_MODE TOTALAREA_MODE
## 1              55.18              68.39          50.18          48.27
##   WALLSMATERIAL_MODE EMERGENCYSTATE_MODE
## 1              50.84                47.4
```

```
cols_to_remove <- colnames(perc_na_mt_40)
application_train_drop_na <- select(application_train, setdiff(colnames(application_train), cols_to_remo
head(application_train_drop_na)
```

```
## # A tibble: 6 x 73
##   SK_ID_CURR TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY
##        <dbl>  <dbl> <chr>              <chr>       <chr>        <chr>
## 1     100002      1 Cash loans         M           N            Y
## 2     100003      0 Cash loans         F           N            N
## 3     100004      0 Revolving loans    M           Y            Y
## 4     100006      0 Cash loans         F           N            Y
## 5     100007      0 Cash loans         M           N            Y
## 6     100008      0 Cash loans         M           N            Y
## # i 67 more variables: CNT_CHILDREN <dbl>, AMT_INCOME_TOTAL <dbl>,
## #   AMT_CREDIT <dbl>, AMT_ANNUITY <dbl>, AMT_GOODS_PRICE <dbl>,
## #   NAME_TYPE_SUITE <chr>, NAME_INCOME_TYPE <chr>, NAME_EDUCATION_TYPE <chr>,
## #   NAME_FAMILY_STATUS <chr>, NAME_HOUSING_TYPE <chr>,
## #   REGION_POPULATION_RELATIVE <dbl>, DAYS_BIRTH <dbl>, DAYS_EMPLOYED <dbl>,
## #   DAYS_REGISTRATION <dbl>, DAYS_ID_PUBLISH <dbl>, FLAG_MOBIL <dbl>,
## #   FLAG_EMP_PHONE <dbl>, FLAG_WORK_PHONE <dbl>, FLAG_CONT_MOBILE <dbl>, ...
```

```
application_train_zero <-as.data.frame(round((colSums(application_train_drop_na==0, na.rm =T)/nrow(appli
application_train_zero <- application_train_zero |> rename("perc_zero" = "round((colSums(application_tra

application_train_zero_90 <- application_train_zero |> filter(perc_zero > 90)
list(rownames(application_train_zero_90))
```

```
## [[1]]
##  [1] "TARGET"                      "FLAG_EMAIL"
##  [3] "REG_REGION_NOT_LIVE_REGION"  "REG_REGION_NOT_WORK_REGION"
##  [5] "LIVE_REGION_NOT_WORK_REGION" "REG_CITY_NOT_LIVE_CITY"
##  [7] "DEF_60_CNT_SOCIAL_CIRCLE"    "FLAG_DOCUMENT_2"
##  [9] "FLAG_DOCUMENT_4"             "FLAG_DOCUMENT_5"
## [11] "FLAG_DOCUMENT_6"             "FLAG_DOCUMENT_7"
## [13] "FLAG_DOCUMENT_8"             "FLAG_DOCUMENT_9"
## [15] "FLAG_DOCUMENT_10"            "FLAG_DOCUMENT_11"
## [17] "FLAG_DOCUMENT_12"            "FLAG_DOCUMENT_13"
## [19] "FLAG_DOCUMENT_14"            "FLAG_DOCUMENT_15"
## [21] "FLAG_DOCUMENT_16"            "FLAG_DOCUMENT_17"
## [23] "FLAG_DOCUMENT_18"            "FLAG_DOCUMENT_19"
## [25] "FLAG_DOCUMENT_20"            "FLAG_DOCUMENT_21"
```

```r
application_train_df <- application_train_drop_na |> select(!c("FLAG_EMAIL","FLAG_DOCUMENT_2","FLAG_DOCU
application_train_df
```

```
## # A tibble: 307,511 x 54
##    SK_ID_CURR TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY
##         <dbl>  <dbl> <chr>              <chr>       <chr>        <chr>
##  1     100002      1 Cash loans         M           N            Y
##  2     100003      0 Cash loans         F           N            N
##  3     100004      0 Revolving loans    M           Y            Y
##  4     100006      0 Cash loans         F           N            Y
##  5     100007      0 Cash loans         M           N            Y
##  6     100008      0 Cash loans         M           N            Y
##  7     100009      0 Cash loans         F           Y            Y
##  8     100010      0 Cash loans         M           Y            Y
##  9     100011      0 Cash loans         F           N            Y
## 10     100012      0 Revolving loans    M           N            Y
## # i 307,501 more rows
## # i 48 more variables: CNT_CHILDREN <dbl>, AMT_INCOME_TOTAL <dbl>,
## #   AMT_CREDIT <dbl>, AMT_ANNUITY <dbl>, AMT_GOODS_PRICE <dbl>,
## #   NAME_TYPE_SUITE <chr>, NAME_INCOME_TYPE <chr>, NAME_EDUCATION_TYPE <chr>,
## #   NAME_FAMILY_STATUS <chr>, NAME_HOUSING_TYPE <chr>,
## #   REGION_POPULATION_RELATIVE <dbl>, DAYS_BIRTH <dbl>, DAYS_EMPLOYED <dbl>,
## #   DAYS_REGISTRATION <dbl>, DAYS_ID_PUBLISH <dbl>, FLAG_MOBIL <dbl>, ...
```

```r
nums <- application_train_df |> select(where(is.numeric))
```

```r
dim(nums)
```

```
## [1] 307511     42
```

```r
corr_cross(nums, # name of dataset
           plot = FALSE,
  max_pvalue = 0.05, # display only significant correlations (at 5% level)
  top = 10 # display top 10 couples of variables (by correlation coefficient)
)
```

```
## Returning only the top 10. You may override with the 'top' argument

## # A tibble: 10 x 8
## # Rowwise:
##    key                         mix          corr pvalue group1 cat1  group2 cat2
##    <chr>                       <chr>       <dbl>  <dbl> <chr>  <chr> <chr>  <chr>
##  1 DAYS_EMPLOYED               FLAG_EMP_~ -1.00       0 DAYS_~ DAYS~ FLAG_~ FLAG~
##  2 OBS_30_CNT_SOCIAL_CIRCLE    OBS_60_CN~  0.998      0 OBS_3~ OBS_~ OBS_6~ OBS_~
##  3 AMT_CREDIT                  AMT_GOODS~  0.987      0 AMT_C~ AMT_~ AMT_G~ AMT_~
##  4 CNT_CHILDREN                CNT_FAM_M~  0.879      0 CNT_C~ CNT_~ CNT_F~ CNT_~
##  5 REG_REGION_NOT_WORK_REGION LIVE_REGI~  0.861      0 REG_R~ REG_~ LIVE_~ LIVE~
##  6 DEF_30_CNT_SOCIAL_CIRCLE   DEF_60_CN~  0.861      0 DEF_3~ DEF_~ DEF_6~ DEF_~
##  7 REG_CITY_NOT_WORK_CITY     LIVE_CITY~  0.826      0 REG_C~ REG_~ LIVE_~ LIVE~
##  8 AMT_ANNUITY                AMT_GOODS~  0.775      0 AMT_A~ AMT_~ AMT_G~ AMT_~
##  9 AMT_CREDIT                 AMT_ANNUI~  0.770      0 AMT_C~ AMT_~ AMT_A~ AMT_~
## 10 DAYS_BIRTH                 FLAG_EMP_~  0.620      0 DAYS_~ DAYS~ FLAG_~ FLAG~
```

FLAG_EMP_PHONE, OBS_30_CNT_SOCIAL_CIRCLE, AMT_GOODS_PRICE are removed due to high multicollinearity.

```
application_train_df <- application_train_df|> select(-c(SK_ID_CURR, FLAG_EMP_PHONE, OBS_30_CNT_SOCIAL_C
head(application_train_df)
```

```
## # A tibble: 6 x 50
##   TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY
##    <dbl> <chr>              <chr>       <chr>        <chr>
## 1      1 Cash loans         M           N            Y
## 2      0 Cash loans         F           N            N
## 3      0 Revolving loans    M           Y            Y
## 4      0 Cash loans         F           N            Y
## 5      0 Cash loans         M           N            Y
## 6      0 Cash loans         M           N            Y
## # i 45 more variables: CNT_CHILDREN <dbl>, AMT_INCOME_TOTAL <dbl>,
## #   AMT_CREDIT <dbl>, AMT_ANNUITY <dbl>, NAME_TYPE_SUITE <chr>,
## #   NAME_INCOME_TYPE <chr>, NAME_EDUCATION_TYPE <chr>,
## #   NAME_FAMILY_STATUS <chr>, NAME_HOUSING_TYPE <chr>,
## #   REGION_POPULATION_RELATIVE <dbl>, DAYS_BIRTH <dbl>, DAYS_EMPLOYED <dbl>,
## #   DAYS_REGISTRATION <dbl>, DAYS_ID_PUBLISH <dbl>, FLAG_MOBIL <dbl>,
## #   FLAG_WORK_PHONE <dbl>, FLAG_CONT_MOBILE <dbl>, FLAG_PHONE <dbl>, ...
```

```
colnames(application_train_df)
```

```
##  [1] "TARGET"                      "NAME_CONTRACT_TYPE"
##  [3] "CODE_GENDER"                 "FLAG_OWN_CAR"
##  [5] "FLAG_OWN_REALTY"             "CNT_CHILDREN"
##  [7] "AMT_INCOME_TOTAL"            "AMT_CREDIT"
##  [9] "AMT_ANNUITY"                 "NAME_TYPE_SUITE"
## [11] "NAME_INCOME_TYPE"            "NAME_EDUCATION_TYPE"
## [13] "NAME_FAMILY_STATUS"          "NAME_HOUSING_TYPE"
## [15] "REGION_POPULATION_RELATIVE"  "DAYS_BIRTH"
## [17] "DAYS_EMPLOYED"               "DAYS_REGISTRATION"
## [19] "DAYS_ID_PUBLISH"             "FLAG_MOBIL"
## [21] "FLAG_WORK_PHONE"             "FLAG_CONT_MOBILE"
## [23] "FLAG_PHONE"                  "OCCUPATION_TYPE"
## [25] "CNT_FAM_MEMBERS"             "REGION_RATING_CLIENT"
## [27] "REGION_RATING_CLIENT_W_CITY" "WEEKDAY_APPR_PROCESS_START"
## [29] "HOUR_APPR_PROCESS_START"     "REG_REGION_NOT_LIVE_REGION"
## [31] "REG_REGION_NOT_WORK_REGION"  "LIVE_REGION_NOT_WORK_REGION"
## [33] "REG_CITY_NOT_LIVE_CITY"      "REG_CITY_NOT_WORK_CITY"
## [35] "LIVE_CITY_NOT_WORK_CITY"     "ORGANIZATION_TYPE"
## [37] "EXT_SOURCE_2"                "EXT_SOURCE_3"
## [39] "DEF_30_CNT_SOCIAL_CIRCLE"    "OBS_60_CNT_SOCIAL_CIRCLE"
## [41] "DEF_60_CNT_SOCIAL_CIRCLE"    "DAYS_LAST_PHONE_CHANGE"
## [43] "FLAG_DOCUMENT_3"             "FLAG_DOCUMENT_21"
## [45] "AMT_REQ_CREDIT_BUREAU_HOUR"  "AMT_REQ_CREDIT_BUREAU_DAY"
## [47] "AMT_REQ_CREDIT_BUREAU_WEEK"  "AMT_REQ_CREDIT_BUREAU_MON"
## [49] "AMT_REQ_CREDIT_BUREAU_QRT"   "AMT_REQ_CREDIT_BUREAU_YEAR"
```

```
write.csv(application_train_df, "./application_train_preprocessed.csv", row.names = F)
```