WdML no. 1: Cleaning Data & Exploratory Data Analysis

Hoang Thien Ly

4/14/2022

In this script, we will include steps of **importing dataset**, **cleaning** and **EDA** for OPTN waiting list history from the **STAR (Standard Transplant Analysis and Research)** file for adult liver transplant candidates. Our dataset would be later divided into training, validation and test set respecting to the proportion of 50-20-30%.

Loading needed libraries:

```r
library(haven) # to load data in SAS form
library(dplyr) # for data processing
library(readr)
```

Import dataset:

To get access to STAR_SAS datasets of organ transplants, we need to write a form to United Network for Organ Sharing. The received file contains info of CODE Dictionary - formats, SAS Dataset, STAR File Documentation and National Star File Documentation. To be specific, we can find in folder SAS Dataset all data related to different organ transplantation performed in the United States of America: **Intestine**, **Kidney**, **Liver**, **Thoracic** and **VCA**. But we will use only data of liver.

To have more info about meaning of features in dataset, we need to look up in file STAR File Documentation.xlsx in folder STAR_SAS. Now, we will load the liver dataset. This file includes one repord per liver waiting list registration and/or transplant. Transplant evens identified by TRR_ID_CODE. Then we will define our ML problem is: binary classification problem. Whether patient receives transplantation (TRR_ID_CODE not NULL) or being removed due to too sick/ death while waiting on waiting list.

```r
setwd("C:/Users/DELL/OneDrive - Politechnika Warszawska/STAR_SAS/STAR_SAS/SAS Dataset 202109/Liver")
df <- read_sas("liver_data.sas7bdat")
```

```r
dim(df)
```

This dataset contains: 329468 observations and 426 features, some features are: AGE (recipient age), CREAT_TX (recipient serum Creatinine at time of Tx), Death_Date (date of death for patient that died on waiting list), education (recipient highest education), ethnicity (recipient ethnicity),… (the excel file of info: finding in STAR File Documentation.xlsx).

Cleaning dataset:

Firstly, we only choose features that seem to be significant for our model (by reading to their meaning, especially, all lab test indices are chosen). Infomation of medical record of deceased donors would be removed.

```
liver_data <- df    %>% select("ALBUMIN_TX","ASCITES_TX","BW4","BW6","C1","C2",
                               "CREAT_TX","DQ1","DQ2","DR51","DR51_2","DR52",
                               "DR52_2","DR53","DR53_2","ENCEPH_TX",
                               "FINAL_ALBUMIN","FINAL_ASCITES","FINAL_BILIRUBIN",
                               "FINAL_CTP_SCORE","FINAL_DIALYSIS_PRIOR_WEEK",
                               "FINAL_ENCEPH","FINAL_INR","FINAL_MELD_OR_PELD",
                               "FINAL_MELD_PELD_LAB_SCORE","FINAL_SERUM_CREAT",
                               "FINAL_SERUM_SODIUM", "INIT_ALBUMIN","INIT_ASCITES",
                               "INIT_BILIRUBIN","INIT_CTP_SCORE",
                               "INIT_MELD_PELD_LAB_SCORE",
                               "INIT_SERUM_CREAT",
                               "INIT_SERUM_SODIUM",
                               "INR_TX",
                               "NUM_PREV_TX",
                               "REM_CD", "TBILI_TX", "TRR_ID_CODE")
```

For the patients get transplanted, they would be assigned a code to the column **TRR_ID_CODE**.

```
liver_data$TRR_ID_CODE[liver_data$TRR_ID_CODE == ""]<- NA
```

Now, for the patients currently in our dataframe, one group for patients transplated livers, remaining we don't have info. But the info is in feature: **REM_CD** - reason for removal from the waiting list. We can read the code SAS Analysis Format of this variable: *REMCD*, then using the table to decode: *LIVER_FORMATS_FLATFILE.DAT*.

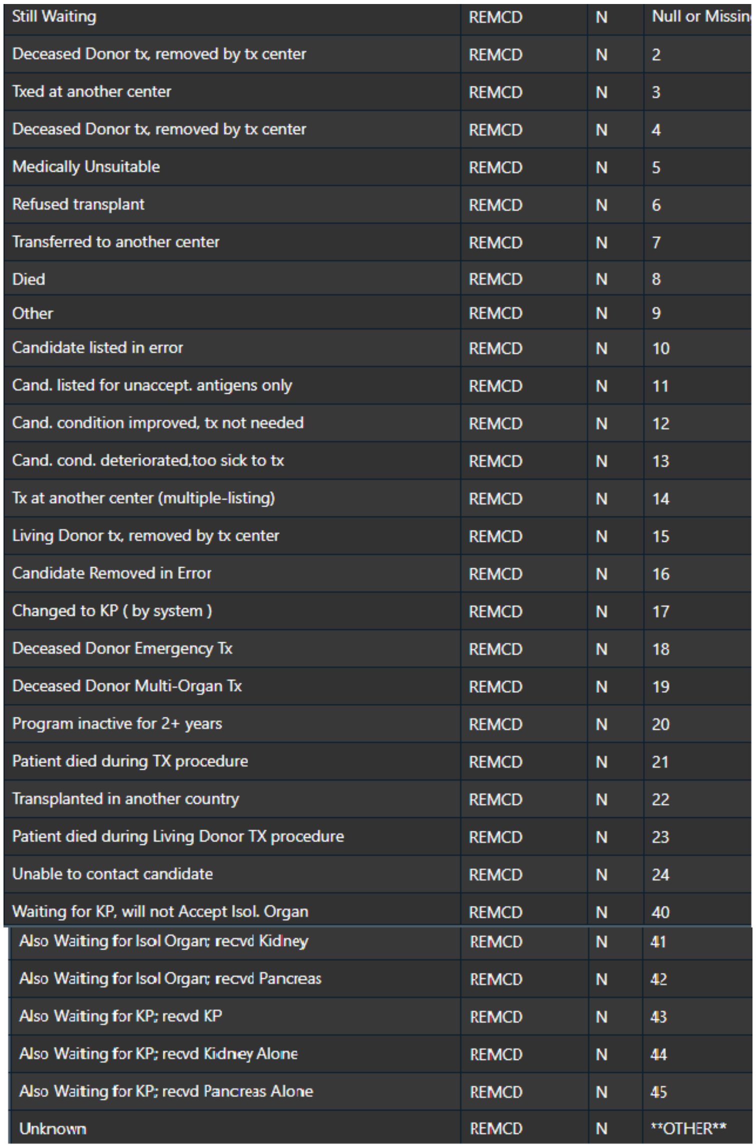
So for our ML problem of classifying and understanding factors influencing the outcome of transplantation (as well as, determine, what influence to cause the death of patients on waiting list). So, we will only extract info of sucessful transplant cases and those patients removed from waiting list due to reason 8, 13, 21, 23.

```
liver_data2 <- liver_data %>% filter( !is.na(TRR_ID_CODE) | REM_CD %in% c(8,13,21,23))
liver_data2 <- liver_data2 %>% mutate(Transplanted_perform = ifelse(!is.na(TRR_ID_CODE),1,0))
liver_data2 <- liver_data2 %>% select(-c(REM_CD,TRR_ID_CODE))

UNOS_liver <- liver_data2 %>% select(-FINAL_DIALYSIS_PRIOR_WEEK,
                                     -FINAL_CTP_SCORE,
                                     -FINAL_MELD_OR_PELD,
                                     -INIT_MELD_OR_PELD,
                                     -INIT_CTP_SCORE)

dim(UNOS_liver)
```

Eventually, we have the dataset *UNOS_liver* of 257315 observations and 38 features.

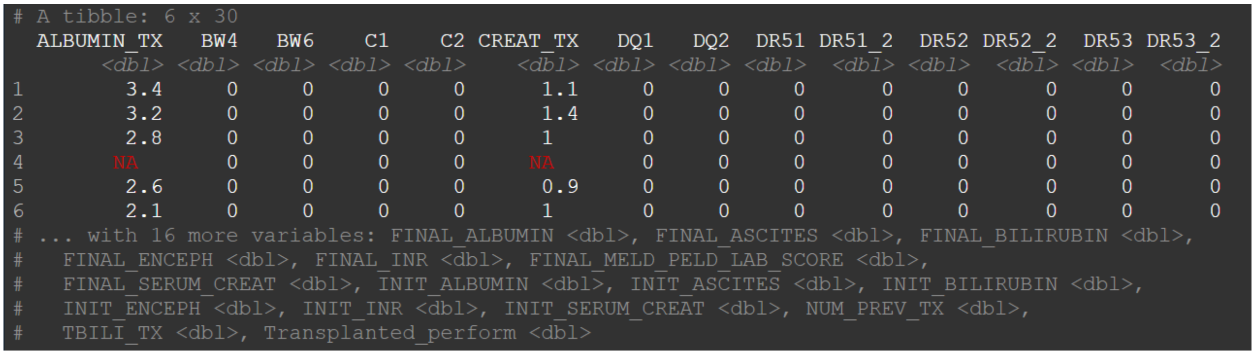
Step of Exploratory Data Analysis:

```
head(UNOS_liver)
```



Still Waiting	REMCD	N	Null or Missin
Deceased Donor tx, removed by tx center	REMCD	N	2
Txed at another center	REMCD	N	3
Deceased Donor tx, removed by tx center	REMCD	N	4
Medically Unsuitable	REMCD	N	5
Refused transplant	REMCD	N	6
Transferred to another center	REMCD	N	7
Died	REMCD	N	8
Other	REMCD	N	9
Candidate listed in error	REMCD	N	10
Cand. listed for unaccept. antigens only	REMCD	N	11
Cand. condition improved, tx not needed	REMCD	N	12
Cand. cond. deteriorated,too sick to tx	REMCD	N	13
Tx at another center (multiple-listing)	REMCD	N	14
Living Donor tx, removed by tx center	REMCD	N	15
Candidate Removed in Error	REMCD	N	16
Changed to KP (by system)	REMCD	N	17
Deceased Donor Emergency Tx	REMCD	N	18
Deceased Donor Multi-Organ Tx	REMCD	N	19
Program inactive for 2+ years	REMCD	N	20
Patient died during TX procedure	REMCD	N	21
Transplanted in another country	REMCD	N	22
Patient died during Living Donor TX procedure	REMCD	N	23
Unable to contact candidate	REMCD	N	24
Waiting for KP, will not Accept Isol. Organ	REMCD	N	40
Also Waiting for Isol Organ; recvd Kidney	REMCD	N	41
Also Waiting for Isol Organ; recvd Pancreas	REMCD	N	42
Also Waiting for KP; recvd KP	REMCD	N	43
Also Waiting for KP; recvd Kidney Alone	REMCD	N	44
Also Waiting for KP; recvd Pancreas Alone	REMCD	N	45
Unknown	REMCD	N	**OTHER**

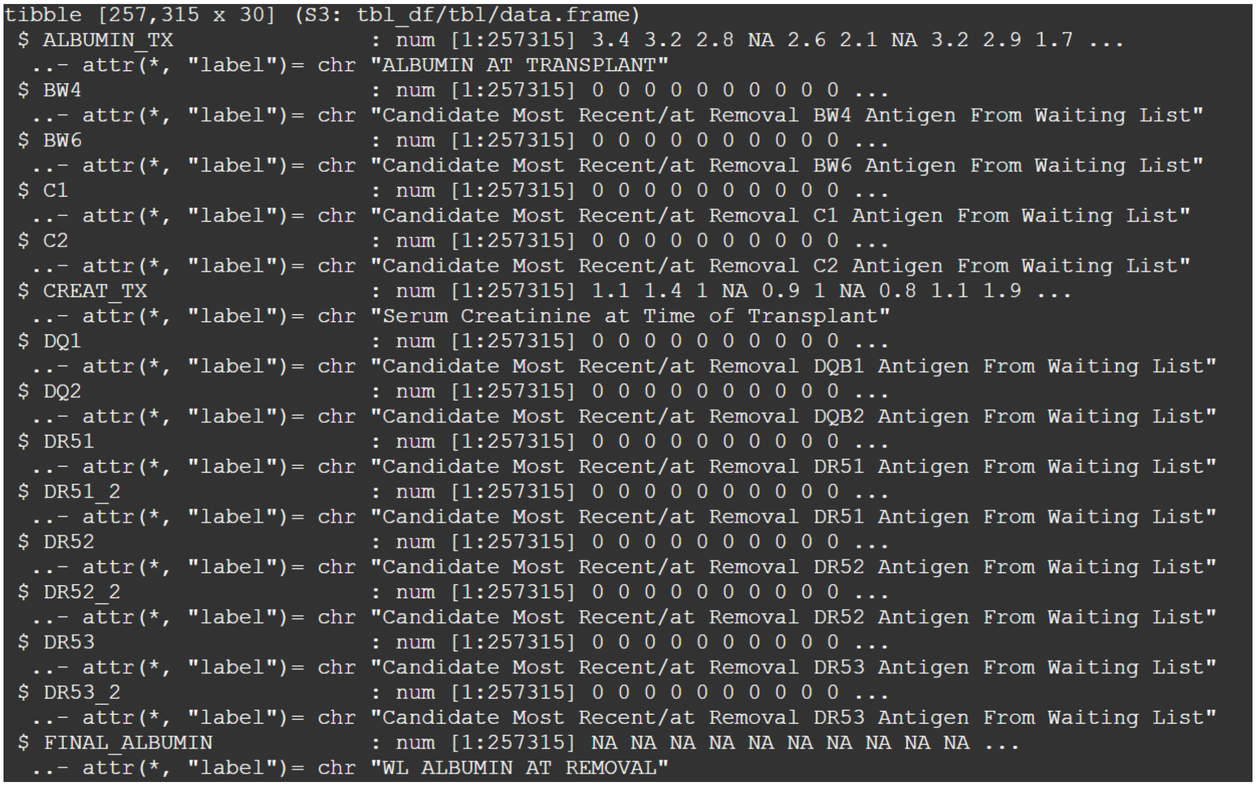
Figure 1: Reason for removal from waiting list



```
# A tibble: 6 x 30
  ALBUMIN_TX   BW4   BW6    C1    C2 CREAT_TX   DQ1   DQ2  DR51 DR51_2  DR52 DR52_2  DR53 DR53_2
       <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl>
1        3.4     0     0     0     0      1.1     0     0     0      0     0      0     0      0
2        3.2     0     0     0     0      1.4     0     0     0      0     0      0     0      0
3        2.8     0     0     0     0      1       0     0     0      0     0      0     0      0
4       NA       0     0     0     0     NA       0     0     0      0     0      0     0      0
5        2.6     0     0     0     0      0.9     0     0     0      0     0      0     0      0
6        2.1     0     0     0     0      1       0     0     0      0     0      0     0      0
# ... with 16 more variables: FINAL_ALBUMIN <dbl>, FINAL_ASCITES <dbl>, FINAL_BILIRUBIN <dbl>,
#   FINAL_ENCEPH <dbl>, FINAL_INR <dbl>, FINAL_MELD_PELD_LAB_SCORE <dbl>,
#   FINAL_SERUM_CREAT <dbl>, INIT_ALBUMIN <dbl>, INIT_ASCITES <dbl>, INIT_BILIRUBIN <dbl>,
#   INIT_ENCEPH <dbl>, INIT_INR <dbl>, INIT_SERUM_CREAT <dbl>, NUM_PREV_TX <dbl>,
#   TBILI_TX <dbl>, Transplanted_perform <dbl>
```

> Comment: Lots of NAs

```r
str(UNOS_liver)
```

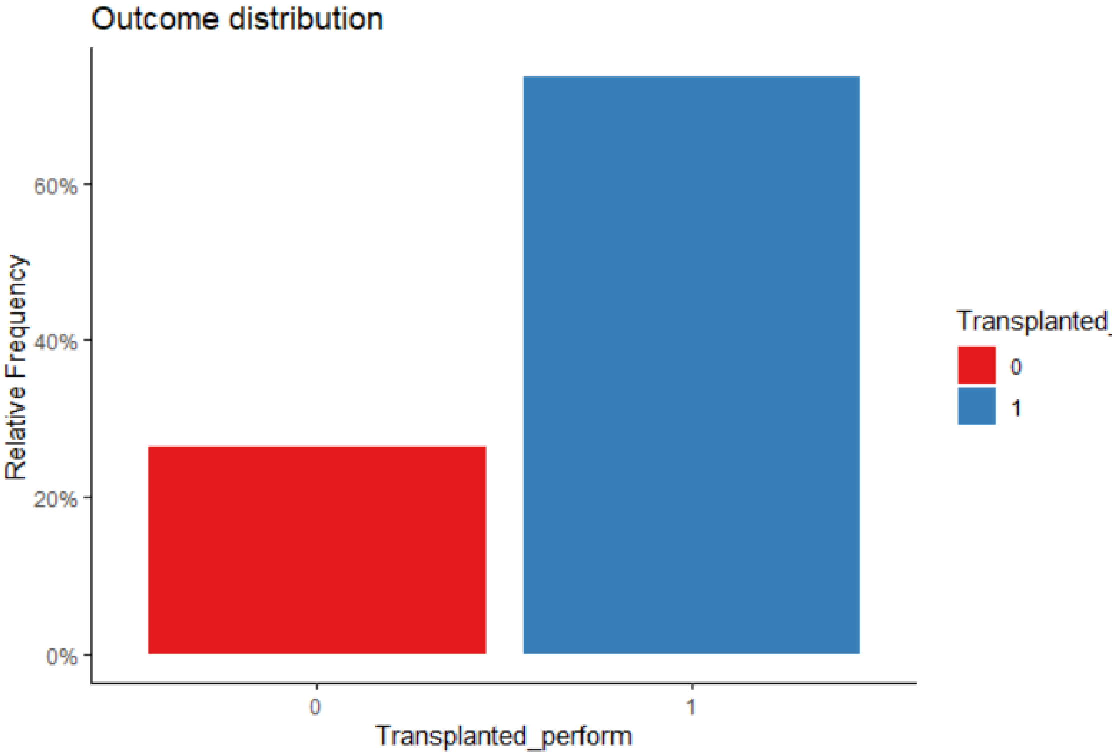


```
tibble [257,315 x 30] (S3: tbl_df/tbl/data.frame)
 $ ALBUMIN_TX               : num [1:257315] 3.4 3.2 2.8 NA 2.6 2.1 NA 3.2 2.9 1.7 ...
  ..- attr(*, "label")= chr "ALBUMIN AT TRANSPLANT"
 $ BW4                      : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal BW4 Antigen From Waiting List"
 $ BW6                      : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal BW6 Antigen From Waiting List"
 $ C1                       : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal C1 Antigen From Waiting List"
 $ C2                       : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal C2 Antigen From Waiting List"
 $ CREAT_TX                 : num [1:257315] 1.1 1.4 1 NA 0.9 1 NA 0.8 1.1 1.9 ...
  ..- attr(*, "label")= chr "Serum Creatinine at Time of Transplant"
 $ DQ1                      : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal DQB1 Antigen From Waiting List"
 $ DQ2                      : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal DQB2 Antigen From Waiting List"
 $ DR51                     : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal DR51 Antigen From Waiting List"
 $ DR51_2                   : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal DR51 Antigen From Waiting List"
 $ DR52                     : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal DR52 Antigen From Waiting List"
 $ DR52_2                   : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal DR52 Antigen From Waiting List"
 $ DR53                     : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal DR53 Antigen From Waiting List"
 $ DR53_2                   : num [1:257315] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "label")= chr "Candidate Most Recent/at Removal DR53 Antigen From Waiting List"
 $ FINAL_ALBUMIN            : num [1:257315] NA NA NA NA NA NA NA NA NA NA ...
  ..- attr(*, "label")= chr "WL ALBUMIN AT REMOVAL"
```

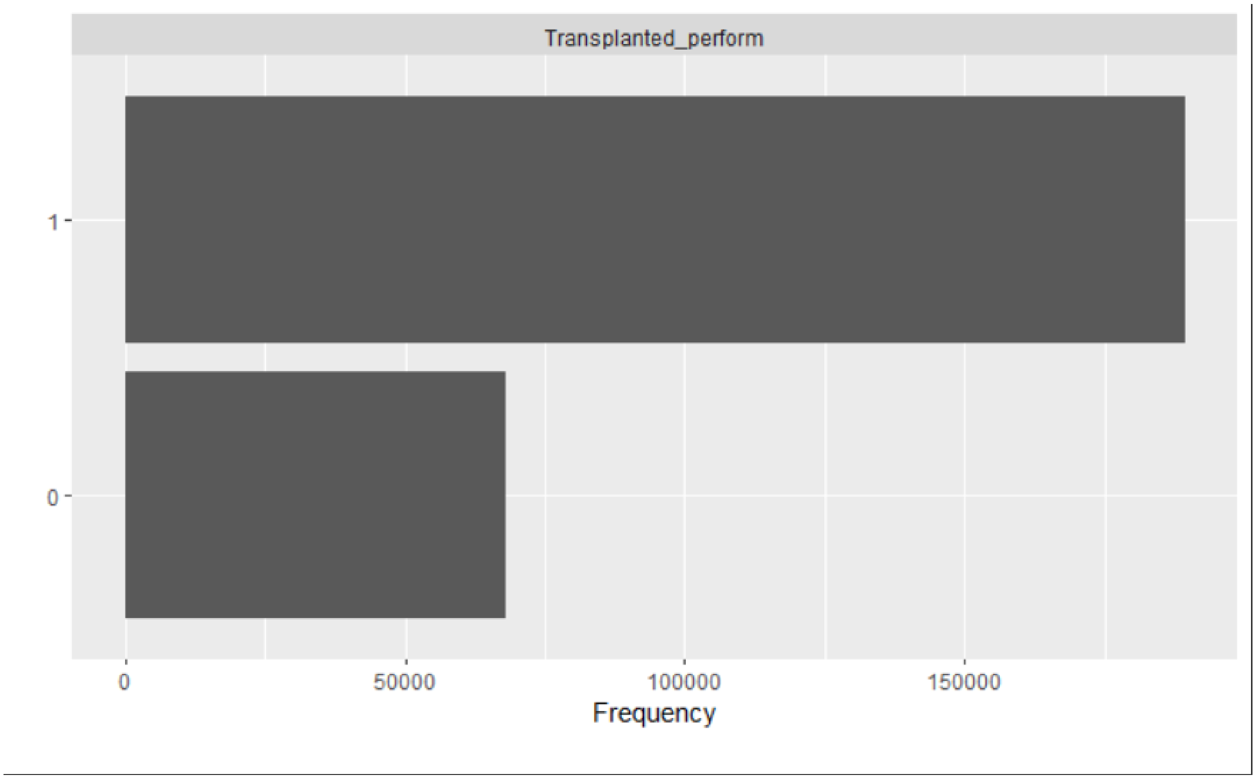
Figure 2: str of dataset.

All features are continuous (results of lab tests).

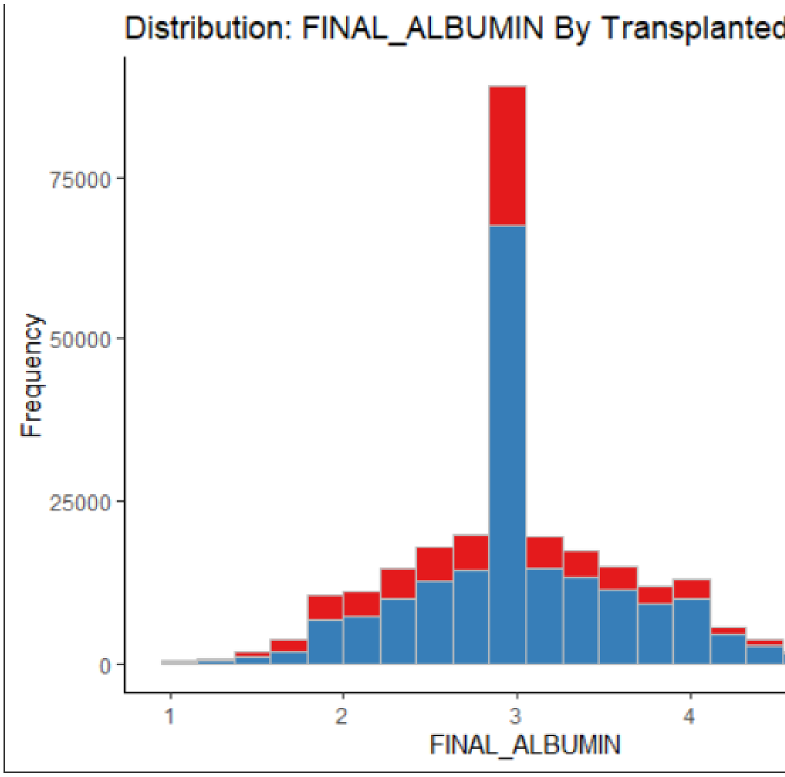
```r
## Checking numbers of NAs:
x <- colSums(is.na(UNOS_liver))/nrow(UNOS_liver) # Some columns include over 40% of NAs.
                                                # except Transplanted_perform, all of
                                                # remained features having NAs.
x[x>0.3]

## We remove columns having higher 30% of NAs:
UNOS_liver <- UNOS_liver %>% select(-c("ASCITES_TX",
                                       "ENCEPH_TX",
                                       "FINAL_SERUM_SODIUM",
                                       "INIT_MELD_PELD_LAB_SCORE",
                                       "INIT_SERUM_SODIUM",
                                       "INR_TX"))
```



Outcome distribution:



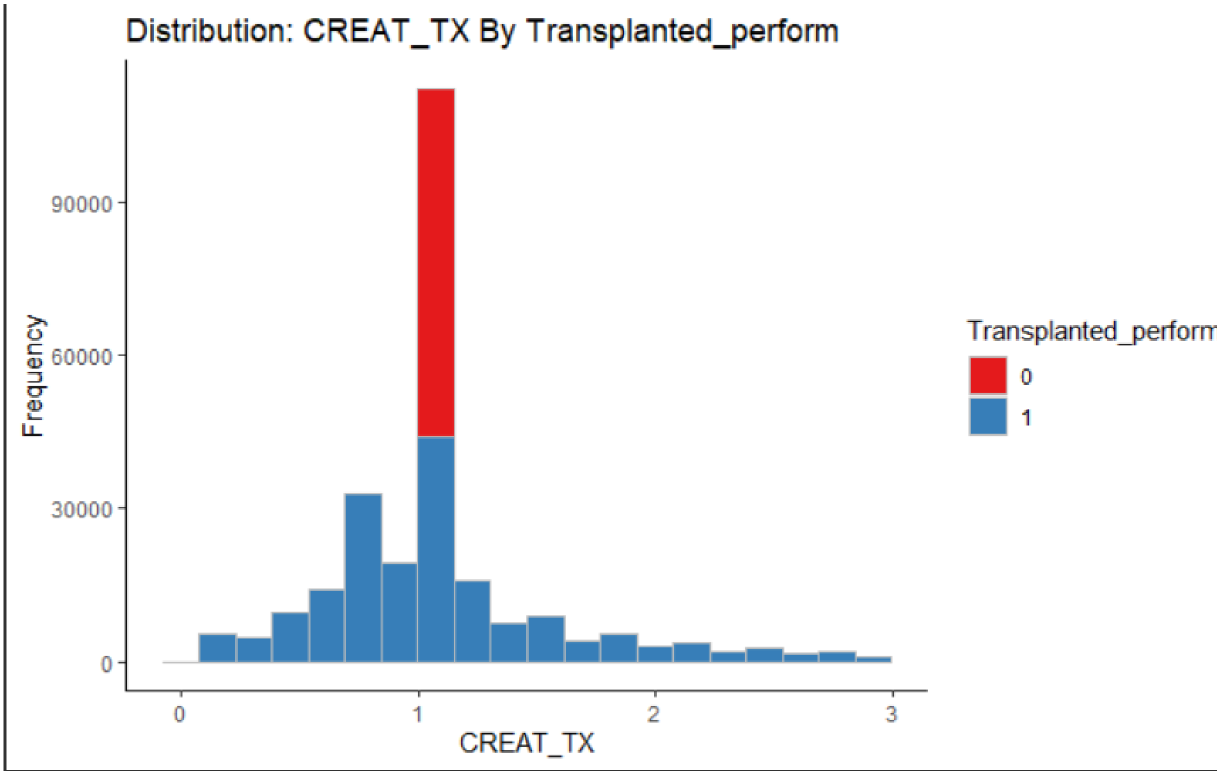
Comment: Our dataset is quite balanced for binary classification problem.



Final Albumin when performed transplantation:

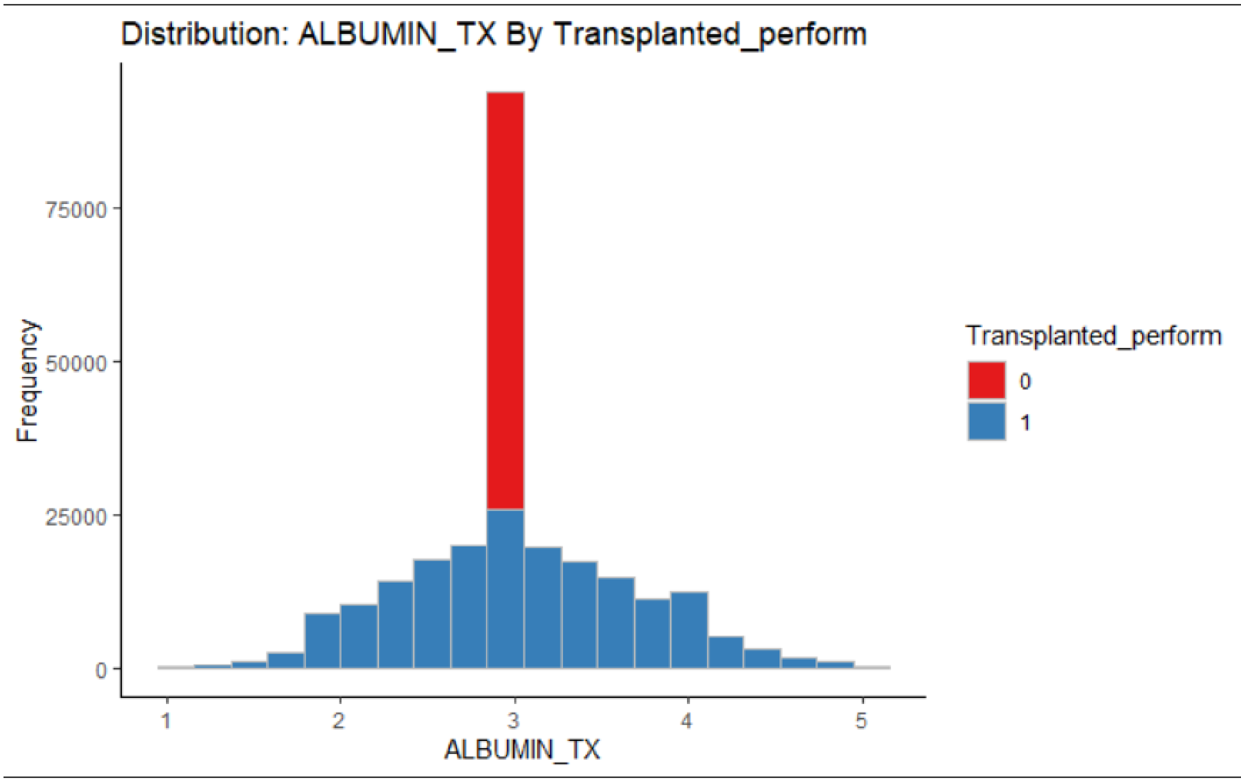
Comment: Most patients died at Final_Albumin 3 and also most transplant cases are performed sucessfully at Final Albumin 3.

Creatinine TX:

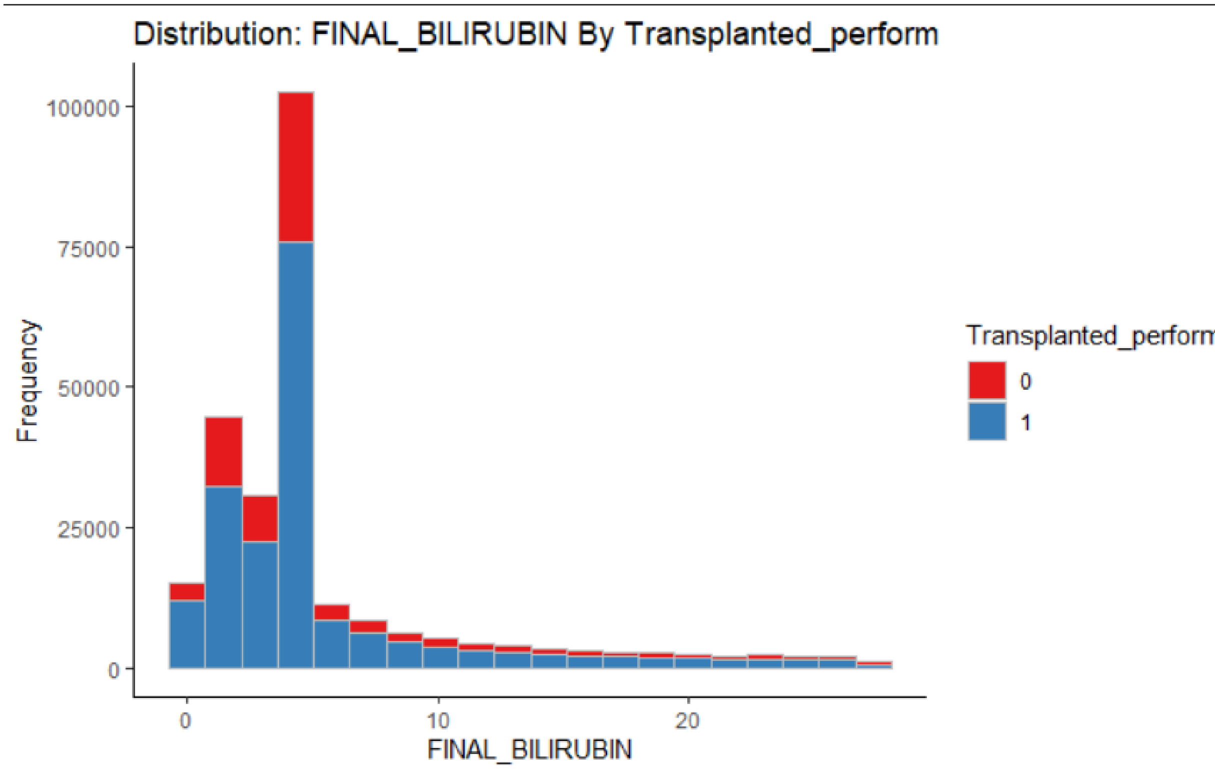


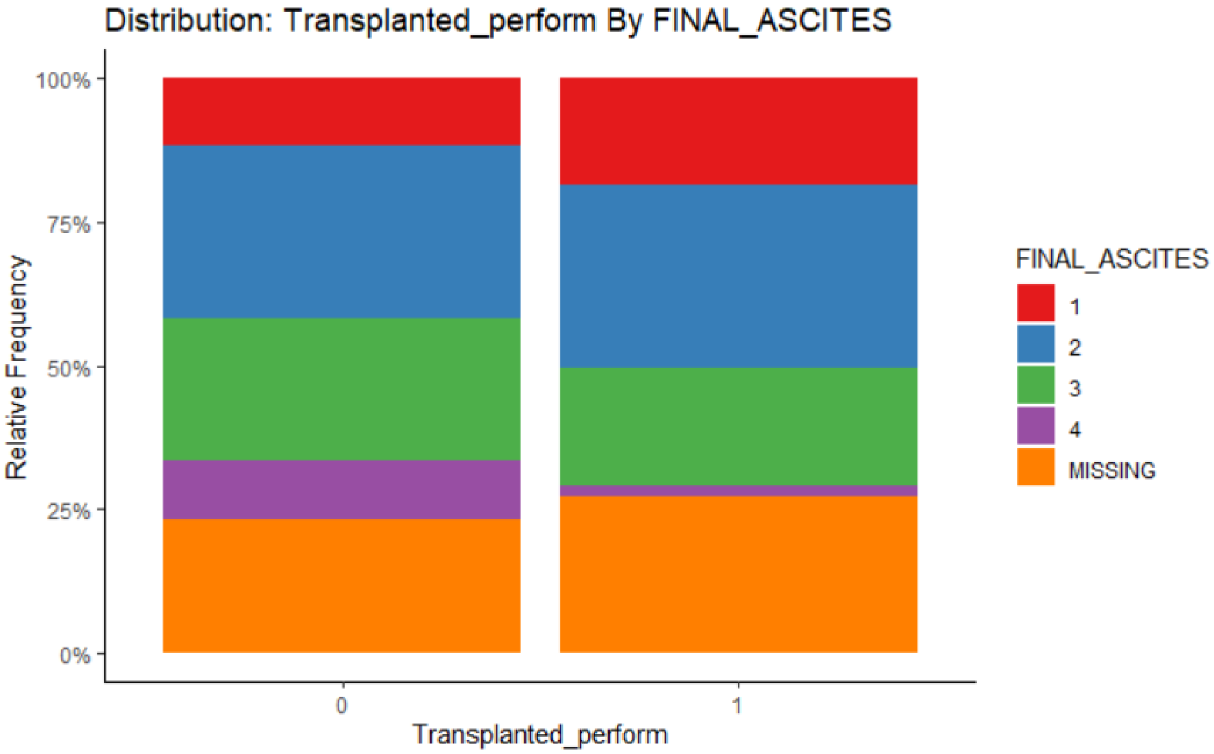
Comment: dataset is right-skewed. All critically deceased recipient passed away while transplanting at Creat_tx around 1.1

Albumin TX:

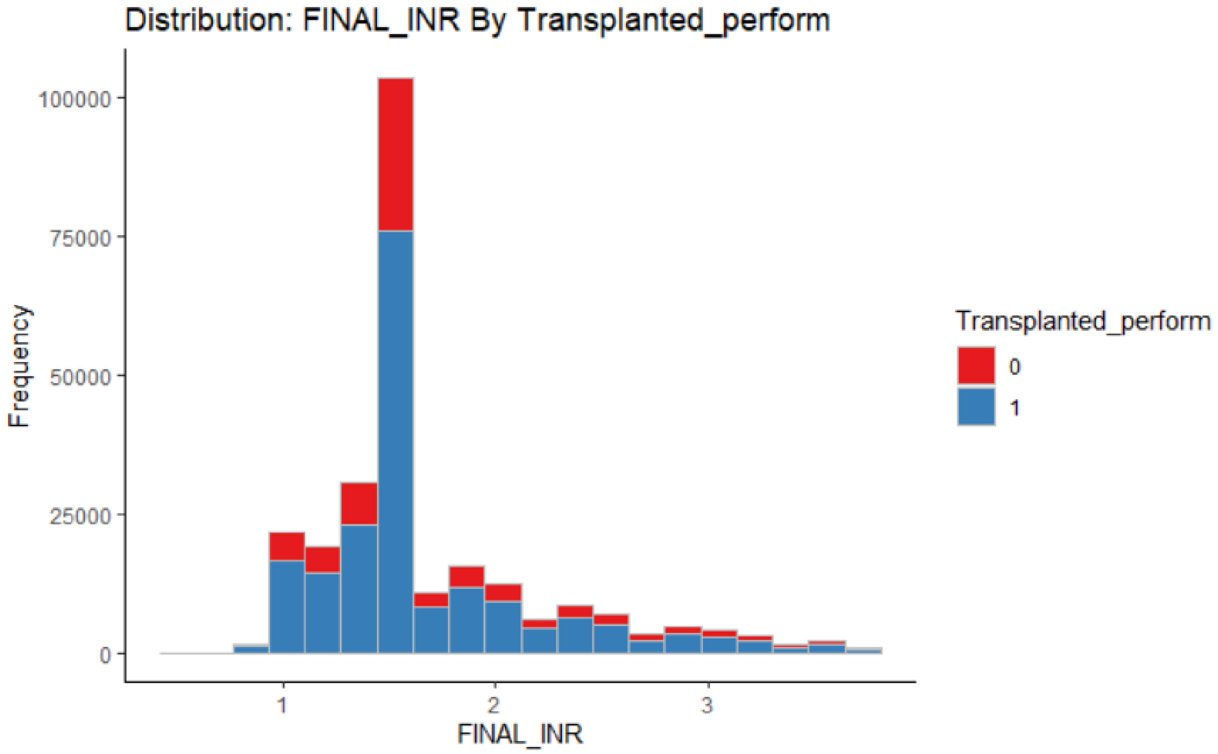


Final Bilirubin:

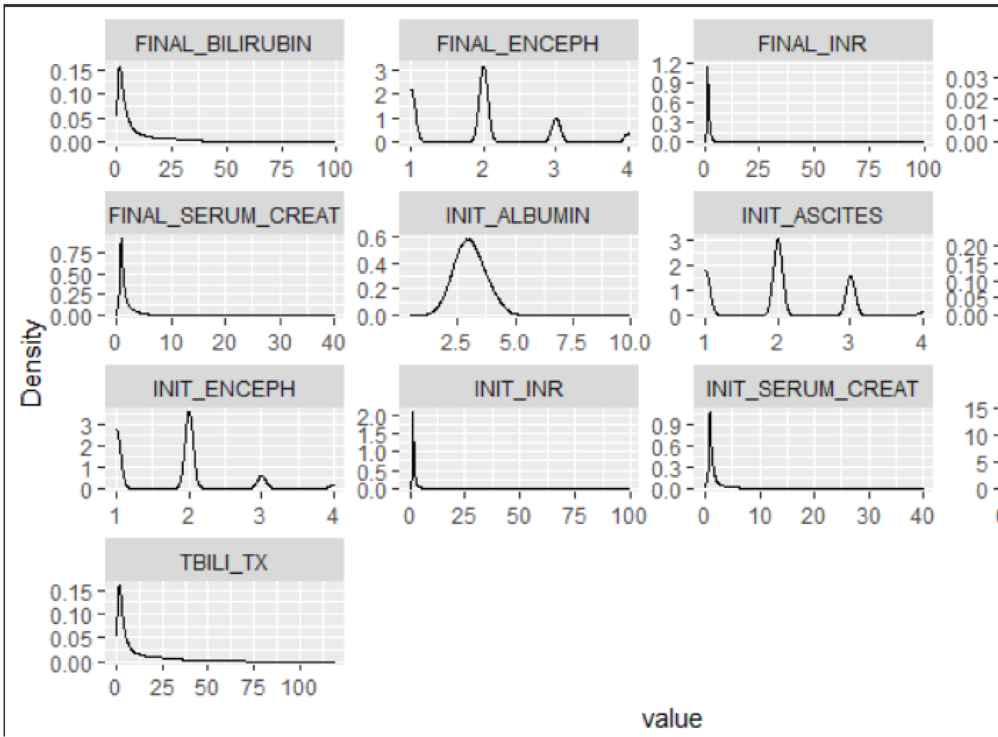




Final Ascities:



Final INR:



Density Plot for continuous vars