Patient Survival Prediction

Aleksandra Bundovska, Miranda Bay, Bahram Khanlarov

3/26/2022

Introduction

The goal of this project is to use the tools and functions we have learned in class..... This dataset contains 91713 observations and 85 variables. For the purpose of this project we will not use all of the variables given. There are various factors given, which are involved when a patient is hospitalized. On the basis of these factors, we will try to predict whether the patient will survive or not. The predictors of in-hospital mortality for admitted patients remain poorly characterized. We aimed to develop and validate a prediction model for all-cause in-hospital mortality among admitted patients

Basic understanding of the data

```
d.patient_survival <- read.csv("Patient_survival_prediction.csv", header = TRUE,
                   stringsAsFactors = TRUE)
```

Calling the head function we get the first observations of the data frame, which allows us to get an initial understanding of the data and its structure.

The str() function allows us to get a compact display of the internal structure of the data frame. There are 91713 observations of 85 variables. For this project we will only use .... variables ‘

```
str(d.patient_survival)
```

```
## 'data.frame':    91713 obs. of  85 variables:
##  $ encounter_id                : int  66154 114252 119783 79267 92056 33181 82208 120995 80471 4287
##  $ patient_id                  : int  25312 59342 50777 46918 34377 74489 49526 50129 10577 90749 .
##  $ hospital_id                 : int  118 81 118 118 33 83 83 33 118 118 ...
##  $ age                         : int  68 77 25 81 19 67 59 70 45 50 ...
##  $ bmi                         : num  22.7 27.4 31.9 22.6 NA ...
##  $ elective_surgery            : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ ethnicity                   : Factor w/ 7 levels "","African American",..: 4 4 4 4 4 4 4 4 4 1 .
##  $ gender                      : Factor w/ 3 levels "","F","M": 3 2 2 2 3 3 2 3 3 3 ...
##  $ height                      : num  180 160 173 165 188 ...
##  $ icu_admit_source            : Factor w/ 6 levels "","Accident & Emergency",..: 3 3 2 4 2 2 2 2 5
##  $ icu_id                      : int  92 90 93 92 91 95 95 91 114 114 ...
##  $ icu_stay_type               : Factor w/ 3 levels "admit","readmit",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ icu_type                    : Factor w/ 8 levels "Cardiac ICU",..: 4 5 5 4 5 5 5 5 2 2 ...
##  $ pre_icu_los_days            : num  0.541667 0.927778 0.000694 0.000694 0.073611 ...
##  $ weight                      : num  73.9 70.2 95.3 61.7 NA ...
##  $ apache_2_diagnosis          : int  113 108 122 203 119 301 108 113 116 112 ...
##  $ apache_3j_diagnosis         : num  502 203 703 1206 601 ...
##  $ apache_post_operative       : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ arf_apache                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ gcs_eyes_apache             : int  3 1 3 4 NA 4 4 4 4 4 ...
##  $ gcs_motor_apache            : int  6 3 6 6 NA 6 6 6 6 6 ...
```

```
##  $ gcs_unable_apache           : int  0 0 0 0 NA 0 0 0 0 0 ...
##  $ gcs_verbal_apache           : int  4 1 5 5 NA 5 5 5 5 5 ...
##  $ heart_rate_apache           : int  118 120 102 114 60 113 133 120 82 94 ...
##  $ intubated_apache            : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ map_apache                  : int  40 46 68 60 103 130 138 60 66 58 ...
##  $ resprate_apache             : num  36 33 37 4 16 35 53 28 14 46 ...
##  $ temp_apache                 : num  39.3 35.1 36.7 34.8 36.7 36.6 35 36.6 36.9 36.3 ...
##  $ ventilated_apache           : int  0 1 0 1 0 0 1 1 1 0 ...
##  $ d1_diasbp_max               : int  68 95 88 48 99 100 76 84 65 83 ...
##  $ d1_diasbp_min               : int  37 31 48 42 57 61 68 46 59 48 ...
##  $ d1_diasbp_noninvasive_max   : int  68 95 88 48 99 100 76 84 65 83 ...
##  $ d1_diasbp_noninvasive_min   : int  37 31 48 42 57 61 68 46 59 48 ...
##  $ d1_heartrate_max            : int  119 118 96 116 89 113 112 118 82 96 ...
##  $ d1_heartrate_min            : int  72 72 68 92 60 83 70 86 82 57 ...
##  $ d1_mbp_max                  : int  89 120 102 84 104 127 117 114 93 101 ...
##  $ d1_mbp_min                  : int  46 38 68 84 90 80 97 60 71 59 ...
##  $ d1_mbp_noninvasive_max      : int  89 120 102 84 104 127 117 114 93 101 ...
##  $ d1_mbp_noninvasive_min      : int  46 38 68 84 90 80 97 60 71 59 ...
##  $ d1_resprate_max             : int  34 32 21 23 18 32 38 28 24 44 ...
##  $ d1_resprate_min             : int  10 12 8 7 16 10 16 12 19 14 ...
##  $ d1_spo2_max                 : int  100 100 98 100 100 97 100 100 97 100 ...
##  $ d1_spo2_min                 : int  74 70 91 95 96 91 87 92 97 96 ...
##  $ d1_sysbp_max                : int  131 159 148 158 147 173 151 147 104 135 ...
##  $ d1_sysbp_min                : int  73 67 105 84 120 107 133 71 98 78 ...
##  $ d1_sysbp_noninvasive_max    : int  131 159 148 158 147 173 151 147 104 135 ...
##  $ d1_sysbp_noninvasive_min    : num  73 67 105 84 120 107 133 71 98 78 ...
##  $ d1_temp_max                 : num  39.9 36.3 37 38 37.2 36.8 37.2 38.5 36.9 37.1 ...
##  $ d1_temp_min                 : num  37.2 35.1 36.7 34.8 36.7 36.6 35 36.6 36.9 36.4 ...
##  $ h1_diasbp_max               : int  68 61 88 62 99 89 107 74 65 83 ...
##  $ h1_diasbp_min               : int  63 48 58 44 68 89 79 55 59 61 ...
##  $ h1_diasbp_noninvasive_max   : int  68 61 88 NA 99 89 NA 74 65 83 ...
##  $ h1_diasbp_noninvasive_min   : int  63 48 58 NA 68 89 NA 55 59 61 ...
##  $ h1_heartrate_max            : int  119 114 96 100 89 83 79 118 82 96 ...
##  $ h1_heartrate_min            : int  108 100 78 96 76 83 72 114 82 60 ...
##  $ h1_mbp_max                  : int  86 85 91 92 104 111 117 88 93 101 ...
##  $ h1_mbp_min                  : int  85 57 83 71 92 111 117 60 71 77 ...
##  $ h1_mbp_noninvasive_max      : int  86 85 91 NA 104 111 117 88 93 101 ...
##  $ h1_mbp_noninvasive_min      : int  85 57 83 NA 92 111 117 60 71 77 ...
##  $ h1_resprate_max             : int  26 31 20 12 NA 12 18 28 24 29 ...
##  $ h1_resprate_min             : int  18 28 16 11 NA 12 18 26 19 17 ...
##  $ h1_spo2_max                 : int  100 95 98 100 100 97 100 96 97 100 ...
##  $ h1_spo2_min                 : int  74 70 91 99 100 97 100 92 97 96 ...
##  $ h1_sysbp_max                : int  131 95 148 136 130 143 191 119 104 135 ...
##  $ h1_sysbp_min                : int  115 71 124 106 120 143 163 106 98 103 ...
##  $ h1_sysbp_noninvasive_max    : int  131 95 148 NA 130 143 NA 119 104 135 ...
##  $ h1_sysbp_noninvasive_min    : int  115 71 124 NA 120 143 NA 106 98 103 ...
##  $ d1_glucose_max              : int  168 145 NA 185 NA 156 197 129 365 134 ...
##  $ d1_glucose_min              : int  109 128 NA 88 NA 125 129 129 288 134 ...
##  $ d1_potassium_max            : num  4 4.2 NA 5 NA 3.9 5 5.8 5.2 4.1 ...
##  $ d1_potassium_min            : num  3.4 3.8 NA 3.5 NA 3.7 4.2 2.4 5.2 3.3 ...
##  $ apache_4a_hospital_death_prob: num  0.1 0.47 0 0.04 NA 0.05 0.1 0.11 NA 0.02 ...
##  $ apache_4a_icu_death_prob    : num  0.05 0.29 0 0.03 NA 0.02 0.05 0.06 NA 0.01 ...
##  $ aids                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cirrhosis                   : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ diabetes_mellitus           : int  1 1 0 0 0 1 1 0 0 0 ...
##  $ hepatic_failure             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ immunosuppression           : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ leukemia                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lymphoma                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ solid_tumor_with_metastasis : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ apache_3j_bodysystem        : Factor w/ 12 levels "","Cardiovascular",..: 11 10 7 2 12 9 10 11 2
##  $ apache_2_bodysystem         : Factor w/ 11 levels "","Cardiovascular",..: 2 8 5 2 9 6 8 2 2 2 ..
##  $ X                           : logi  NA NA NA NA NA NA ...
##  $ hospital_death              : int  0 0 0 0 0 0 0 0 1 0 ...
```

We select the colunms we need for further data analysis and create new data set called n.patient_survival.

```
n.patient_survival <- d.patient_survival%>% select(hospital_id,age,bmi,elective_surgery,ethnicity, gende
```

```
head(n.patient_survival)
```

```
##   hospital_id age   bmi elective_surgery ethnicity gender height
## 1         118  68 22.73               0 Caucasian      M  180.3
## 2          81  77 27.42               0 Caucasian      F  160.0
## 3         118  25 31.95               0 Caucasian      F  172.7
## 4         118  81 22.64               1 Caucasian      F  165.1
## 5          33  19    NA               0 Caucasian      M  188.0
## 6          83  67 27.56               0 Caucasian      M  190.5
##   pre_icu_los_days weight apache_2_diagnosis cirrhosis diabetes_mellitus
## 1      0.541666667   73.9                113         0                 1
## 2      0.927777778   70.2                108         0                 1
## 3      0.000694444   95.3                122         0                 0
## 4      0.000694444   61.7                203         0                 0
## 5      0.073611111     NA                119         0                 0
## 6      0.000694444  100.0                301         0                 1
##   hepatic_failure immunosuppression leukemia lymphoma
## 1               0                 0        0        0
## 2               0                 0        0        0
## 3               0                 0        0        0
## 4               0                 0        0        0
## 5               0                 0        0        0
## 6               0                 0        0        0
##   solid_tumor_with_metastasis apache_3j_bodysystem apache_2_bodysystem
## 1                           0              Sepsis      Cardiovascular
## 2                           0         Respiratory         Respiratory
## 3                           0           Metabolic           Metabolic
## 4                           0       Cardiovascular      Cardiovascular
## 5                           0              Trauma              Trauma
## 6                           0        Neurological          Neurologic
##   hospital_death
## 1              0
## 2              0
## 3              0
## 4              0
## 5              0
## 6              0
```

```
summary(n.patient_survival)
```

```
##   hospital_id         age             bmi        elective_surgery
```

```
##  Min.   :  2.0   Min.   :16.00   Min.   :14.85   Min.   :0.0000
##  1st Qu.: 47.0   1st Qu.:52.00   1st Qu.:23.64   1st Qu.:0.0000
##  Median :109.0   Median :65.00   Median :27.66   Median :0.0000
##  Mean   :105.7   Mean   :62.31   Mean   :29.19   Mean   :0.1837
##  3rd Qu.:161.0   3rd Qu.:75.00   3rd Qu.:32.93   3rd Qu.:0.0000
##  Max.   :204.0   Max.   :89.00   Max.   :67.81   Max.   :1.0000
##                  NA's   :4228    NA's   :3429
##              ethnicity     gender        height      pre_icu_los_days
##                  : 1395    :   25   Min.   :137.2   Min.   :-24.94722
##  African American: 9547   F:42219   1st Qu.:162.5   1st Qu.:  0.03542
##  Asian           : 1129   M:49469   Median :170.1   Median :  0.13889
##  Caucasian       :70684             Mean   :169.6   Mean   :  0.83577
##  Hispanic        : 3796             3rd Qu.:177.8   3rd Qu.:  0.40903
##  Native American :  788             Max.   :195.6   Max.   :159.09097
##  Other/Unknown   : 4374             NA's   :1334
##      weight       apache_2_diagnosis   cirrhosis      diabetes_mellitus
##  Min.   : 38.60   Min.   :101.0      Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 66.80   1st Qu.:113.0      1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 80.30   Median :122.0      Median :0.0000   Median :0.0000
##  Mean   : 84.03   Mean   :185.4      Mean   :0.0157   Mean   :0.2252
##  3rd Qu.: 97.10   3rd Qu.:301.0      3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :186.00   Max.   :308.0      Max.   :1.0000   Max.   :1.0000
##  NA's   :2720     NA's   :1662       NA's   :715      NA's   :715
##  hepatic_failure immunosuppression    leukemia         lymphoma
##  Min.   :0.000   Min.   :0.0000    Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000    1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000    Median :0.0000   Median :0.0000
##  Mean   :0.013   Mean   :0.0262    Mean   :0.0071   Mean   :0.0041
##  3rd Qu.:0.000   3rd Qu.:0.0000    3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.000   Max.   :1.0000    Max.   :1.0000   Max.   :1.0000
##  NA's   :715     NA's   :715       NA's   :715      NA's   :715
##  solid_tumor_with_metastasis       apache_3j_bodysystem
##  Min.   :0.0000              Cardiovascular  :29999
##  1st Qu.:0.0000              Neurological    :11896
##  Median :0.0000              Sepsis          :11740
##  Mean   :0.0206              Respiratory     :11609
##  3rd Qu.:0.0000              Gastrointestinal: 9026
##  Max.   :1.0000              Metabolic       : 7650
##  NA's   :715                 (Other)         : 9793
##       apache_2_bodysystem hospital_death
##  Cardiovascular  :38816    Min.   :0.0000
##  Neurologic      :11896    1st Qu.:0.0000
##  Respiratory     :11609    Median :0.0000
##  Gastrointestinal: 9026    Mean   :0.0863
##  Metabolic       : 7650    3rd Qu.:0.0000
##  Trauma          : 3842    Max.   :1.0000
##  (Other)         : 8874
```

Our new dataset has 91713 obs. of 20 variables

```
str(n.patient_survival)
```

```
apply(n.patient_survival, MARGIN = 2, FUN = anyNA)
```

```
##                hospital_id                        age
```

```
##                         FALSE                          TRUE
##                           bmi             elective_surgery
##                          TRUE                         FALSE
##                     ethnicity                        gender
##                         FALSE                         FALSE
##                        height             pre_icu_los_days
##                          TRUE                         FALSE
##                        weight           apache_2_diagnosis
##                          TRUE                          TRUE
##                     cirrhosis            diabetes_mellitus
##                          TRUE                          TRUE
##               hepatic_failure            immunosuppression
##                          TRUE                          TRUE
##                      leukemia                      lymphoma
##                          TRUE                          TRUE
## solid_tumor_with_metastasis         apache_3j_bodysystem
##                          TRUE                         FALSE
##           apache_2_bodysystem               hospital_death
##                         FALSE                         FALSE
```

We can already spot out columns with NAs.(all TRUE values)

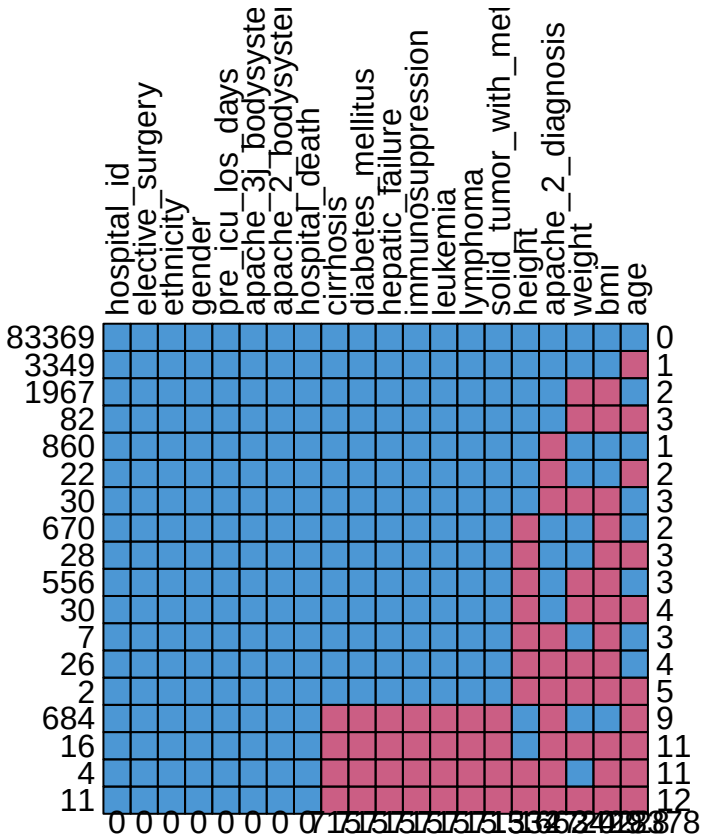
```
library("mice")
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
missing_pattern <- md.pattern(n.patient_survival, rotate.names = TRUE)
```



In the plot the total values of missing within the effected columns are not displayed properly. The next line shows us exactly how much values are missing in each column.

```
apply(n.patient_survival, MARGIN = 2, FUN = function(x) {sum(is.na(x))})
```

```
##                  hospital_id                          age
##                            0                         4228
##                          bmi              elective_surgery
##                         3429                            0
##                    ethnicity                       gender
##                            0                            0
##                       height              pre_icu_los_days
##                         1334                            0
##                       weight           apache_2_diagnosis
##                         2720                         1662
##                    cirrhosis            diabetes_mellitus
##                          715                          715
##              hepatic_failure            immunosuppression
##                          715                          715
##                     leukemia                     lymphoma
##                          715                          715
## solid_tumor_with_metastasis          apache_3j_bodysystem
##                          715                            0
##          apache_2_bodysystem               hospital_death
##                            0                            0
```

Since NAs could have an impact on analysis, it is decided that rows contaning NAs will be droped. The script will dropout any row that has missing data on it remaining with only the untouched rows and save them into another object called n.patient_survival_dropna. This way we can keep both the original dataset

and also the modified dataset in the working environment.Later on, we separate the survivor and nonsurvivor data from the modified dataset.

```
n.patient_survival_dropna <- n.patient_survival[rowSums(is.na(n.patient_survival)) <=0,]
```

Since categorical variables enter into statistical models differently than continuous variables, storing data as factors insures that the modeling functions will treat such data correctly.

```
n.patient_survival_dropna$age <- as.factor(n.patient_survival_dropna$age)
n.patient_survival_dropna$gender <- as.factor(n.patient_survival_dropna$gender)
n.patient_survival_dropna$ethnicity <- as.factor(n.patient_survival_dropna$ethnicity)
```

We do realize that 6904 patients died during hospitalization.

```
unique(n.patient_survival_dropna$hospital_death)
```

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

```
sum(n.patient_survival_dropna$hospital_death==1)
```

```
## [1] 6904
```

```
sum(n.patient_survival_dropna$hospital_death==0)
```

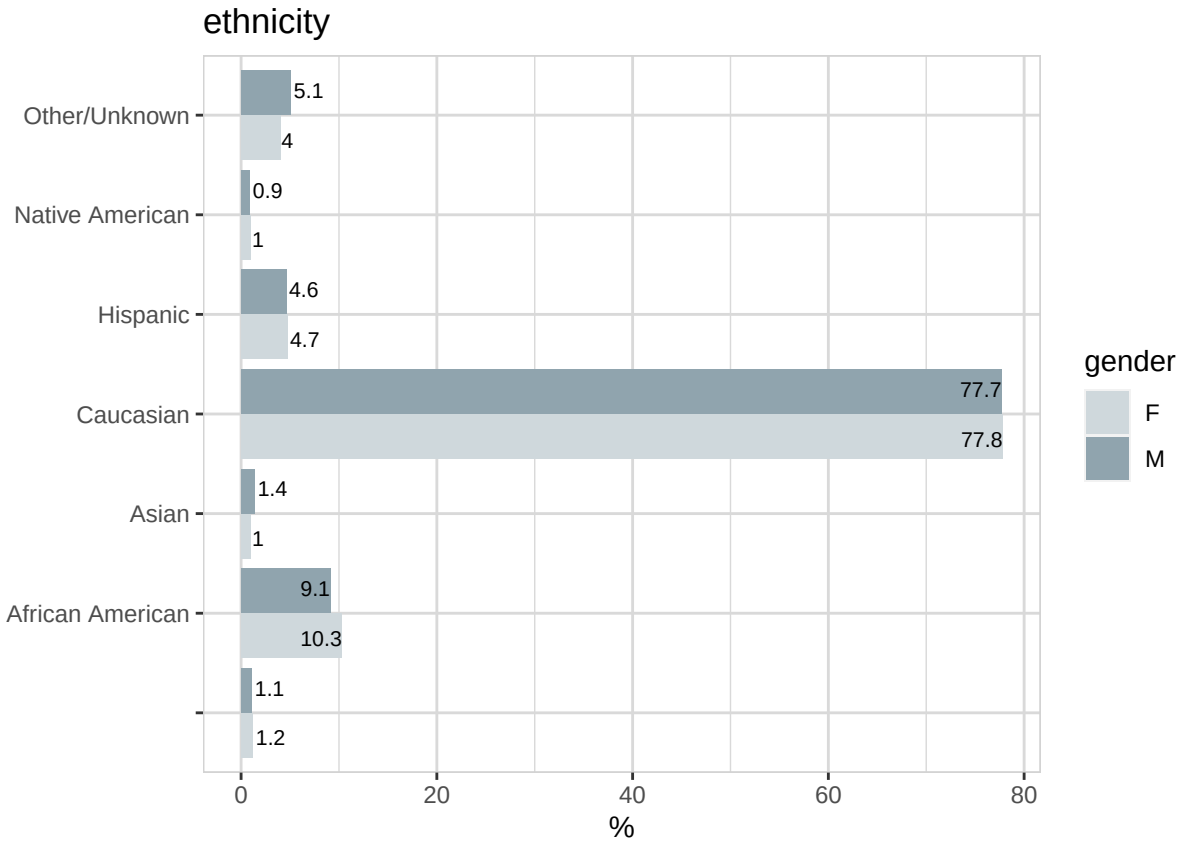
```
## [1] 76465
```

We separate the patients the ones died in hospital and survived data from dataset.

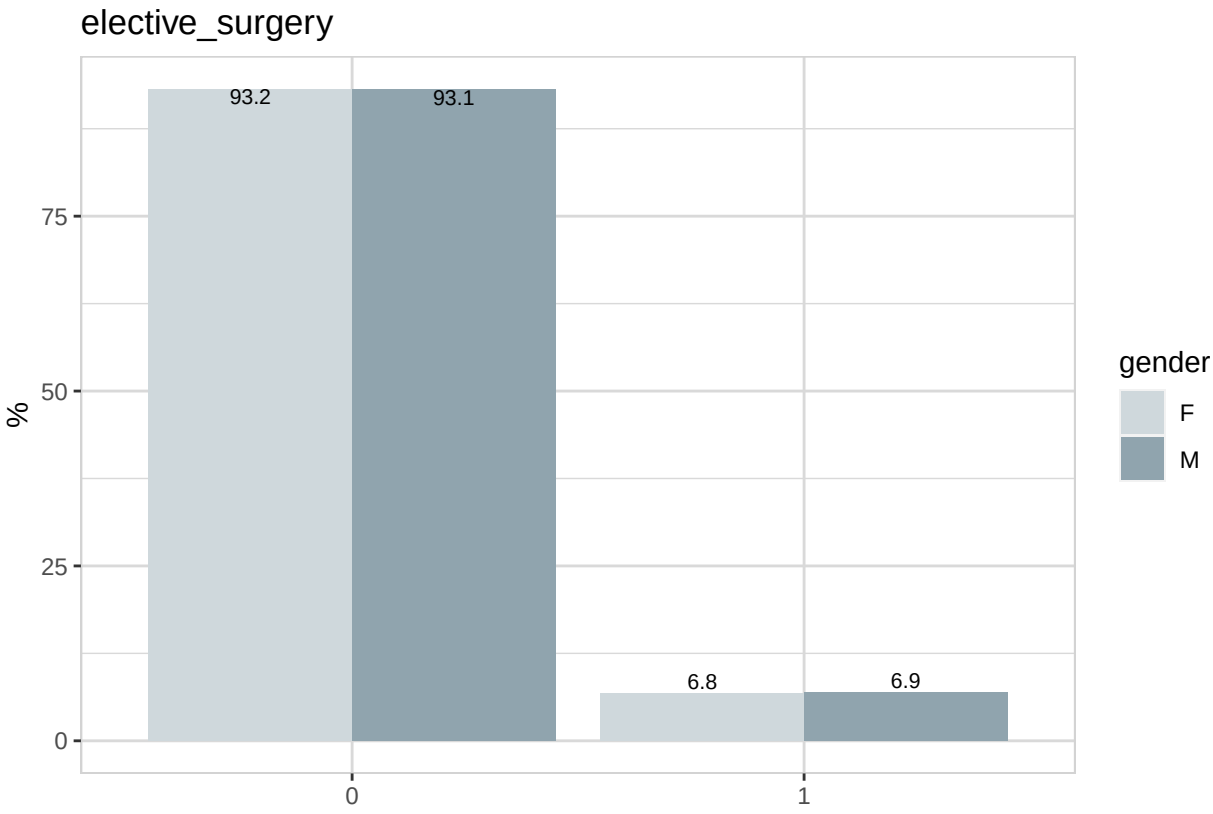
```
n.patient_survival_dropna_death <- n.patient_survival_dropna[n.patient_survival_dropna$hospital_death==
n.patient_survival_dropna_non_death <- n.patient_survival_dropna[n.patient_survival_dropna$hospital_dea
```

Vizualisation

```
library(explore)
library(dplyr)
n.patient_survival_dropna_death %>% explore(ethnicity, target = gender, split = TRUE)
```

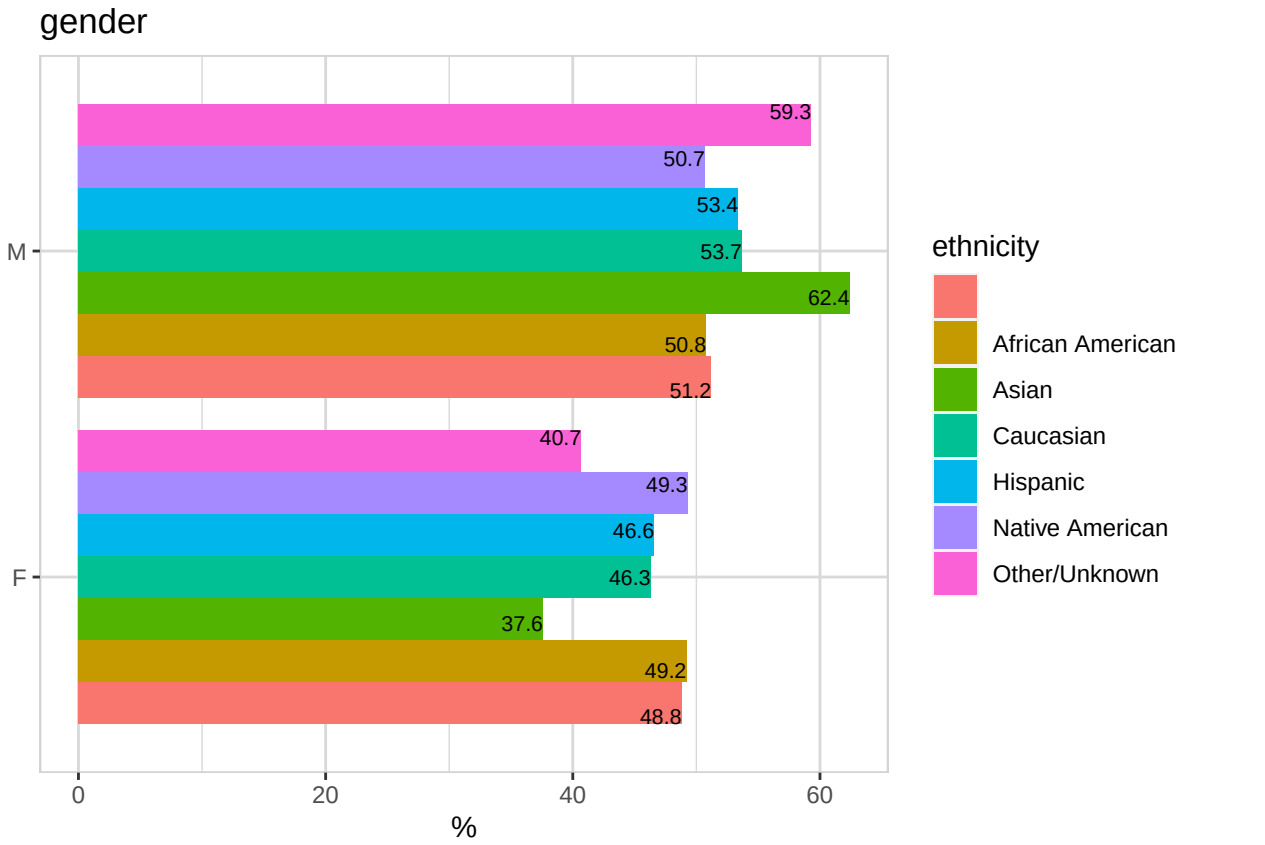



```
n.patient_survival_dropna_death %>% explore(elective_surgery, target = gender, split = TRUE)
```

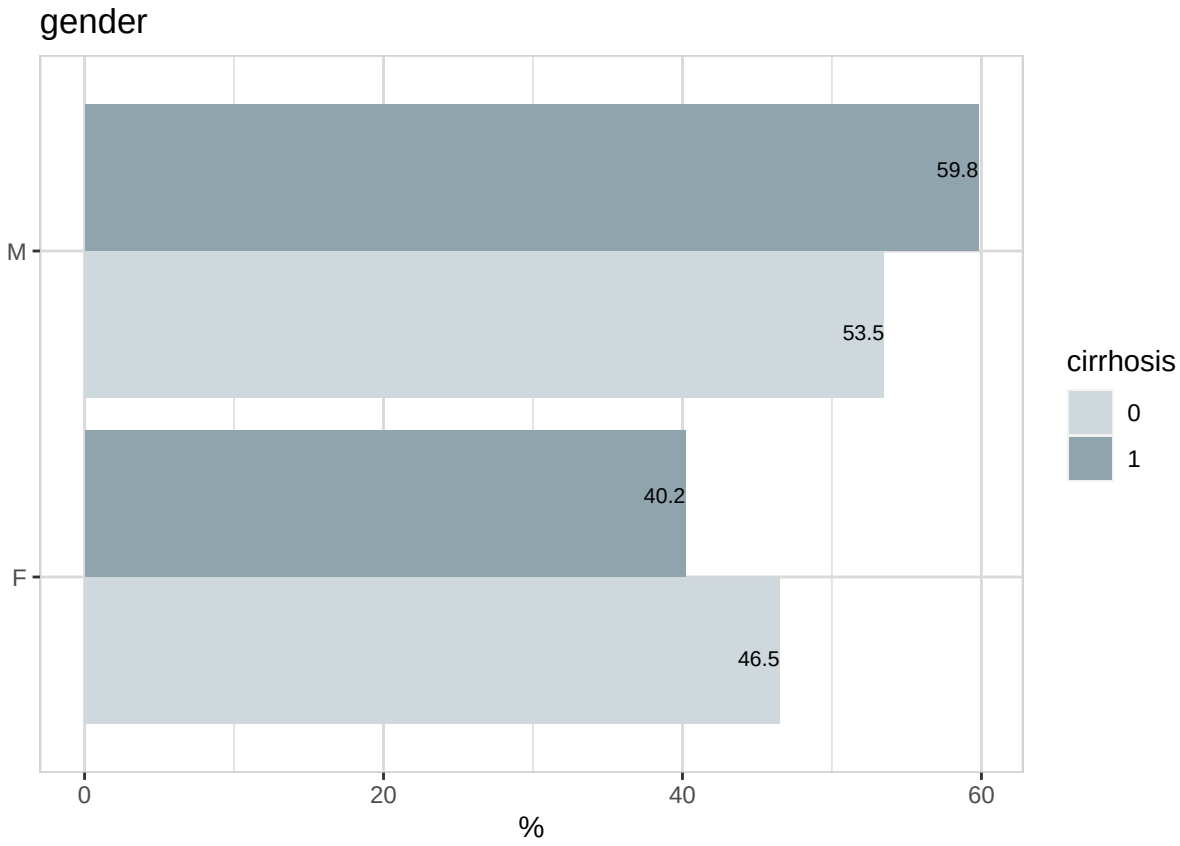


Among patients died during hospitalization Asians takes majority with 62.4% and females with Native American origin 49.3%. 59.8% male patients patient has a history of heavy alcohol use with portal hypertension and varices, other causes of cirrhosis while this is only 40.2% for female patients. In general among patients died Males are 53.7% more than females 46.3%

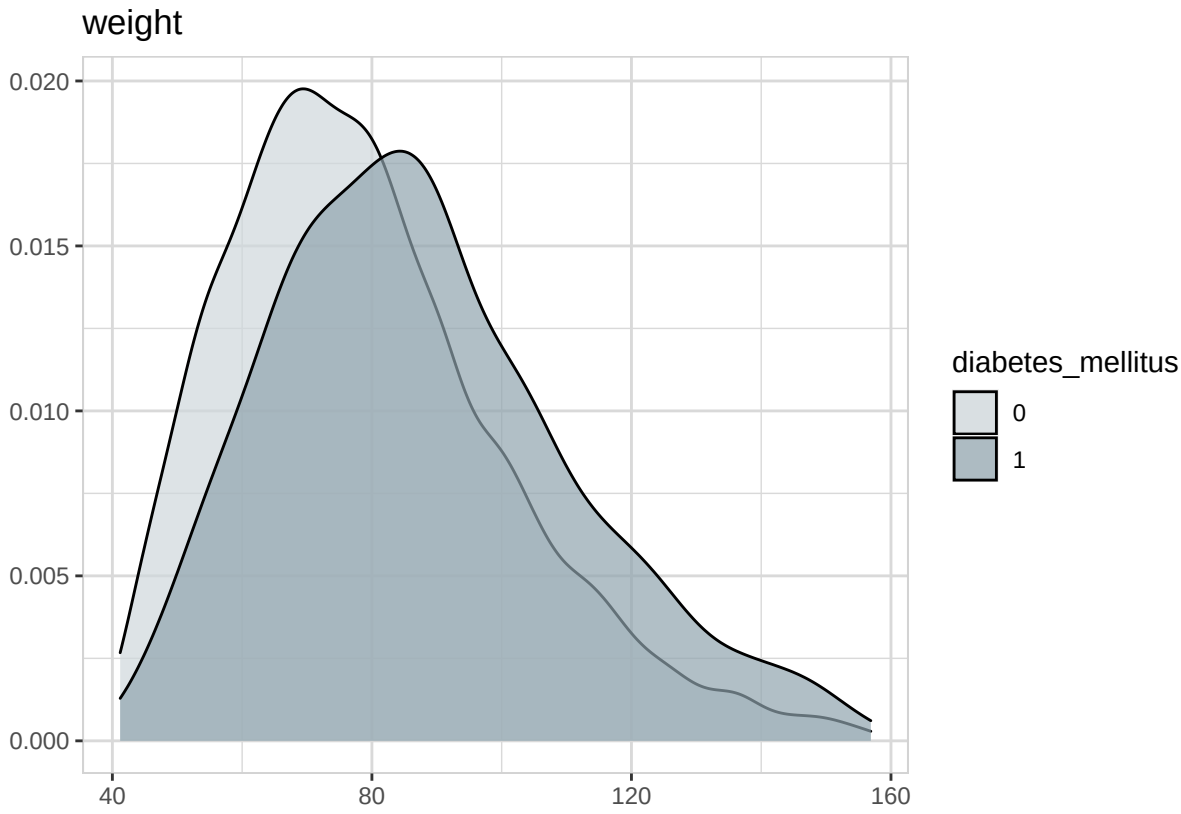
```
n.patient_survival_dropna_death %>% explore(gender, target = ethnicity)
```




```
n.patient_survival_dropna_death %>% explore(gender, target=cirrhosis)
```

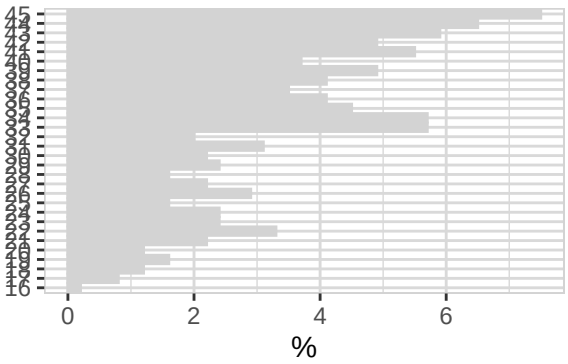


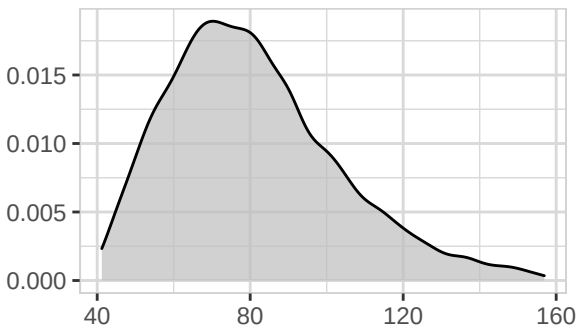

```
n.patient_survival_dropna_death %>% explore(weight, target=diabetes_mellitus,split=TRUE)
```

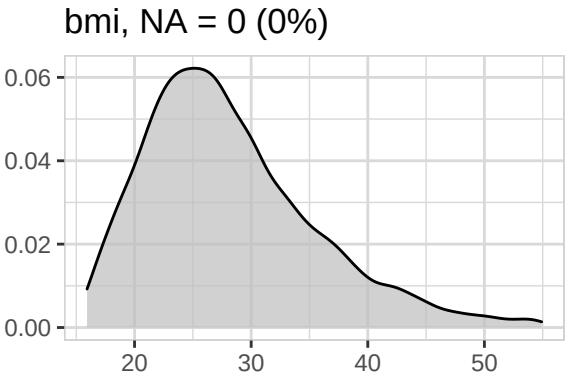


```
n.patient_survival_dropna_death %>%
  select(age,weight,bmi) %>%
  explore_all()
```

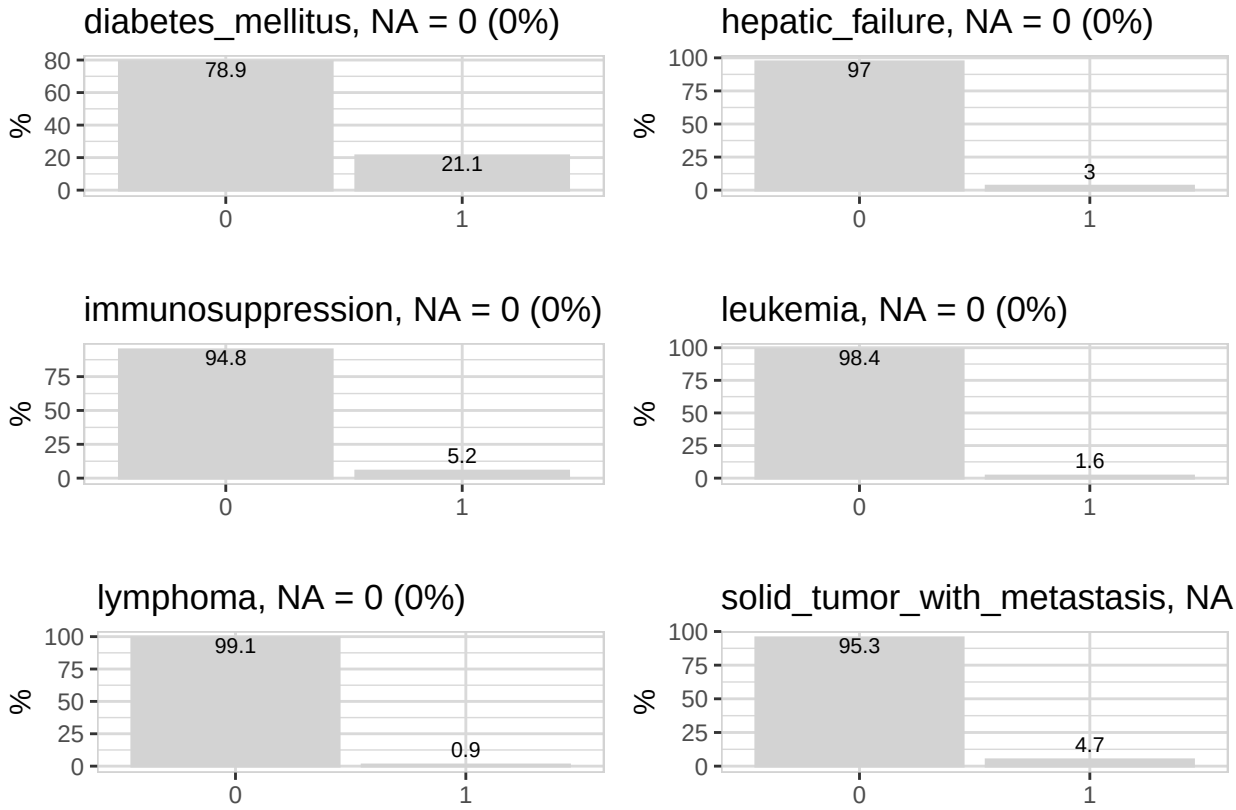
```
## Warning in explore_bar(data_tmp, !!sym(var_name)): number of bars limited to 30
## by parameter max_cat
```



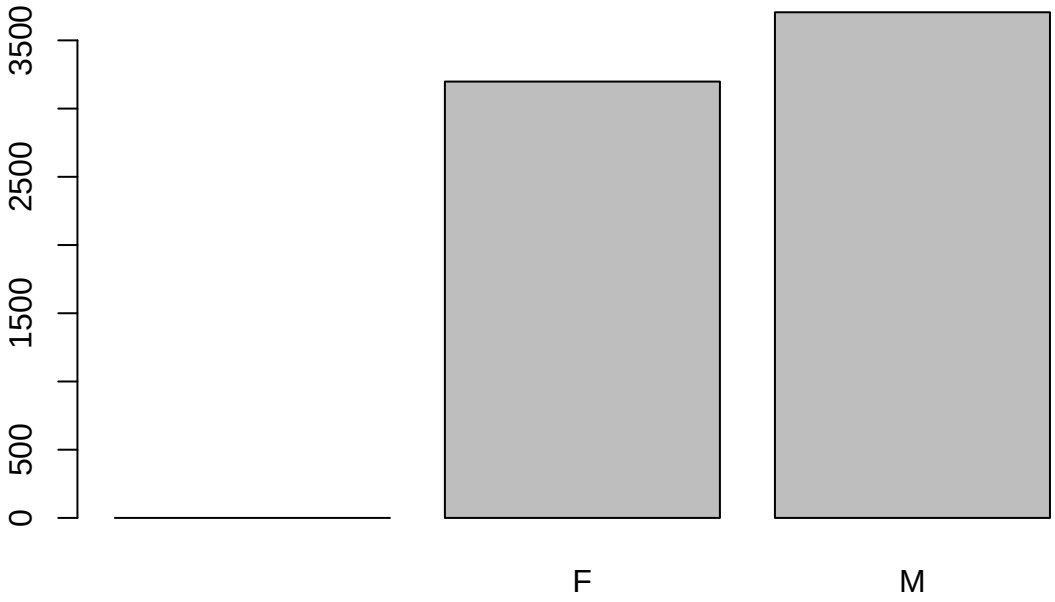




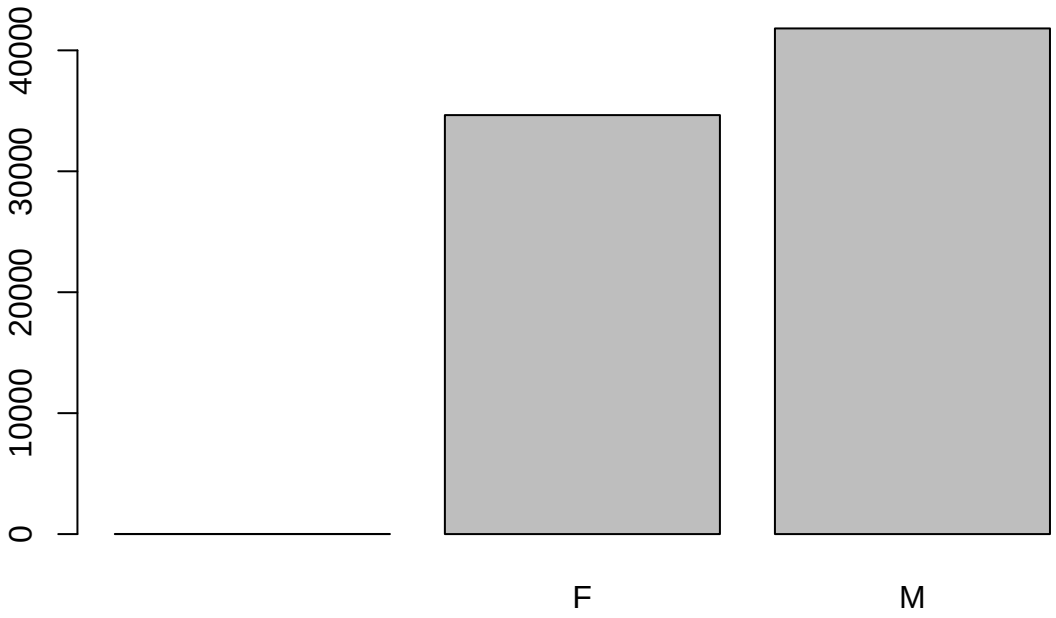

```
n.patient_survival_dropna_death %>%
  select(diabetes_mellitus, hepatic_failure, immunosuppression,leukemia,lymphoma,solid_tumor_with_metas
  explore_all()
```




```
barplot(table(n.patient_survival_dropna_death$gender))
```




```
barplot(table(n.patient_survival_dropna_non_death$gender))
```


0
10000
20000
30000
40000
F
M