

Correlations

Eduardo Yuki Yada

Imports

```
library(tidyverse)
library(yaml)
library(kableExtra)
library(ggcorrplot)
```

Loading data

```
load('../dataset/processed_data.RData')
load('../dataset/processed_dictionary.RData')

columns_list <- yaml.load_file("../auxiliar/columns_list.yaml")

outcome_column <- params$outcome_column
threshold <- params$threshold
```

Functions

```
niceFormatting = function(df, caption="", digits = 2, font_size = NULL){
  df %>%
    kbl(booktabs = T, longtable = T, caption = caption, digits = digits, format = "latex") %>%
    kable_styling(font_size = font_size,
                  latex_options = c("striped", "HOLD_position", "repeat_header"))
}
```

Correlation

```
na_eligible_columns <- df %>%
  summarise(across(everything(), ~ mean(is.na(.)))) %>%
  select_if(function(.) last(.) < 0.8) %>%
  names

unique_eligible_columns <- df %>%
  summarise(across(everything(), ~ length(unique(.)))) %>%
  select_if(function(.) last(.) > 1) %>%
  names

pre_columns = df_names %>%
  filter(momento.aquisicao == 'Admissão t0') %>%
  .$variable.name

weird_columns <- c('dieta_parentral', 'dieta_enteral')

eligible_columns <- intersect(na_eligible_columns,
                             unique_eligible_columns) %>%
```

```

intersect(pre_columns)

eligible_columns <- setdiff(eligible_columns, weird_columns)

corr <- df %>%
  select(all_of(intersect(columns_list$numerical_columns,
                           eligible_columns))) %>%

  drop_na %>%
  cor %>%
  as.matrix

## Warning in cor(.): o desvio padrão é zero

corr_table <- corr %>%
  as.data.frame %>%
  tibble::rownames_to_column(var = 'row') %>%
  tidyr::pivot_longer(-row, names_to = 'column', values_to = 'correlation') %>%
  filter(row < column)

rename_column <- function(df, column_name){
  variable.name <- 'variable.name'
  df <- df %>%
    left_join(df_names %>% select(variable.name, abbrev.field.label),
              by = setNames(variable.name, column_name)) %>%
    select(-all_of(column_name)) %>%
    rename(!sym(column_name) := abbrev.field.label) %>%
    relocate(!sym(column_name))
}

corr_table %>%
  filter(correlation > 0.9) %>%
  rename_column('row') %>%
  rename_column('column') %>%
  select(row, column, correlation) %>%
  niceFormatting(caption = "Pearson Correlation", font_size = 9)

```

Table 1: Pearson Correlation

| row | column | correlation |
|---|---|-------------|
| Idade no momento do primeiro procedimento | Idade no Procedimento 1 | 1.00 |
| Núm. de hospitalizações pré-procedimento | Número da Admissão T0 | 0.98 |
| Ano da admissão T0 | Ano do procedimento 1 | 1.00 |
| Antibióticos | Quantidade de antimicrobianos | 1.00 |
| Quantidade de procedimentos invasivos | Suporte cardiocirculatório | 0.97 |
| ECG | Quantidade de exames por métodos gráficos | 1.00 |
| Exames laboratoriais | Radiografias | 0.90 |
| Quantidade de exames de análises clínicas | Exames laboratoriais | 1.00 |
| Quantidade de exames de análises clínicas | Radiografias | 0.90 |
| Quantidade de exames de análises clínicas | Quantidade de exames diagnóstico por imagem | 0.93 |
| Quantidade de exames diagnóstico por imagem | Exames laboratoriais | 0.93 |
| Quantidade de exames diagnóstico por imagem | Radiografias | 0.98 |
| Quantidade de classes medicamentosas de ação cardiovascular | Quantidade de classes medicamentosas utilizadas | 0.91 |

Hypothesis Tests

```

df_wilcox <- tibble()

for (variable in intersect(columns_list$numerical_columns,
                           eligible_columns)){
  if (mean(is.na(df[[variable]])) > 0.95) next

```

```

x <- filter(df, !!sym(outcome_column) == 0)[[variable]]
y <- filter(df, !!sym(outcome_column) == 1)[[variable]]

test = tryCatch(wilcox.test(x, y, alternative = "two.sided", exact = FALSE),
  error=function(cond) {
    message("Can't calculate Wilcox test for variable ", variable)
    message(cond)
    return(list(statistic = NaN, p.value = NaN))
  })

df_wilcox = bind_rows(df_wilcox,
  list("Variable" = variable,
    "Statistic" = test$statistic,
    "p-value" = test$p.value))
}

significant_num_cols <- df_wilcox %>%
  filter(`p-value` <= threshold) %>%
  select(Variable) %>%
  pull

df_wilcox <- df_wilcox %>%
  arrange(`p-value`) %>%
  mutate(`Statistic` = round(`Statistic`, 3)) %>%
  rename_column('Variable')

df_wilcox %>%
  mutate(`p-value` = case_when(`p-value` == 1 ~ sprintf('> 0%s999', getOption("OutDec")),
    `p-value` < 0.001 ~ sprintf('< 0%s001', getOption("OutDec")),
    TRUE ~ as.character(round(`p-value`, 3)))) %>%
  niceFormatting(caption = "Mann-Whitney Test")

```

Table 2: Mann-Whitney Test

| Variable | Statistic | p-value |
|---|-----------|---------|
| Quantidade de classes medicamentosas utilizadas | 1841228 | < 0.001 |
| Antagonista da Aldosterona | 2889486 | < 0.001 |
| Número da Admissão T0 | 4429622 | < 0.001 |
| Insuficiência cardíaca | 2877501 | < 0.001 |
| Quantidade de classes medicamentosas de ação cardiovascular | 1552991 | < 0.001 |
| Diuretico | 2739336 | < 0.001 |
| Quantidade de medicamentos de ação cardiovascular | 2690400 | < 0.001 |
| Núm. de hospitalizações pré-procedimento | 4579096 | < 0.001 |
| Exames laboratoriais | 3077022 | < 0.001 |
| Quantidade de exames de análises clínicas | 3078022 | < 0.001 |
| Número de comorbidades | 4367503 | < 0.001 |
| DVA | 3034247 | < 0.001 |
| Quantidade de exames diagnóstico por imagem | 3123052 | < 0.001 |
| Quantidade de exames por métodos gráficos | 3152882 | < 0.001 |
| ECG | 3154118 | < 0.001 |
| Equipe Multiprofissional | 3249324 | < 0.001 |
| Ultrassom | 3683860 | < 0.001 |
| Antiarrítmicos | 3178272 | < 0.001 |
| Radiografias | 3258900 | < 0.001 |
| Culturas | 3635630 | < 0.001 |
| UTI durante a admissão T0 | 4952422 | < 0.001 |
| Anticoagulantes orais | 3492210 | < 0.001 |
| Ecocardiograma | 3546944 | < 0.001 |

Table 2: Mann-Whitney Test (*continued*)

| Variable | Statistic | p-value |
|---|-----------|---------|
| Cintilografia | 3952710 | < 0.001 |
| Vasodilator | 3244828 | < 0.001 |
| Psicofármacos | 3185034 | < 0.001 |
| Tomografia | 3821948 | < 0.001 |
| Digoxina | 3563510 | < 0.001 |
| Estatinas | 3276719 | < 0.001 |
| Ressonancia magnetica | 3936010 | < 0.001 |
| Quantidade de antimicrobianos | 3241760 | < 0.001 |
| Antibióticos | 3247172 | < 0.001 |
| Holter | 3913411 | < 0.001 |
| Insulina | 3578401 | < 0.001 |
| Quantidade de procedimentos invasivos | 3834204 | < 0.001 |
| Citologias | 4134334 | < 0.001 |
| Bomba de infusão contínua | 3664356 | < 0.001 |
| IECA/BRA | 3404105 | < 0.001 |
| Cateterismo | 3969258 | < 0.001 |
| Idade no momento do primeiro procedimento | 5194318 | < 0.001 |
| Idade no Procedimento 1 | 5194318 | < 0.001 |
| Diálise durante a admissão T0 | 5715057 | < 0.001 |
| Antiplaquetario EV | 3762062 | < 0.001 |
| Cateter venoso central | 4114162 | < 0.001 |
| Ano do procedimento 1 | 5331004 | < 0.001 |
| Ano da admissão T0 | 5316110 | < 0.001 |
| Quantidade de exames histopatológicos | 4147622 | < 0.001 |
| Intervenção coronária percutânea | 4157518 | 0.002 |
| Diárias no serviço de Emergência na admissão T0 | 2413188 | 0.01 |
| Flebografia | 4149451 | 0.011 |
| Outros procedimentos cirúrgicos | 4103762 | 0.013 |
| Angioplastia | 4191188 | 0.024 |
| Teste de esforço | 4241520 | 0.026 |
| Transfusão de hemoderivados | 4166822 | 0.027 |
| Exames endoscópicos | 4163320 | 0.028 |
| Tilt Test | 4188900 | 0.045 |
| Angio TC | 4154193 | 0.062 |
| Antifúngicos | 3766746 | 0.079 |
| Interconsulta médica | 4122072 | 0.124 |
| Aortografia | 4193848 | 0.129 |
| Angiografia | 4193858 | 0.129 |
| Antiviral | 3793187 | 0.131 |
| Suporte cardiocirculatório | 4191540 | 0.135 |
| Ventilação não invasiva | 4191588 | 0.136 |
| PET-CT | 4188010 | 0.153 |
| Eletrofisiologia | 4161329 | 0.171 |
| Polissonografia | 4197817 | 0.239 |
| Intervenção cardiovascular em laboratório de hemodinâmica | 4192663 | 0.276 |
| Cirurgia Toracica | 4196178 | 0.286 |
| Arteriografia | 4201104 | 0.335 |
| Traqueostomia | 4210070 | 0.372 |
| Trombolitico | 3811554 | 0.44 |
| Biopsias | 4213857 | 0.454 |
| Antiretroviral | 3810912 | 0.485 |
| Anticonvulsivante | 3788005 | 0.494 |
| Espirometria / Ergoespirometria | 4197586 | 0.494 |

Table 2: Mann-Whitney Test (continued)

| Variable | Statistic | p-value |
|---|-----------|---------|
| Angio RM | 4209778 | 0.576 |
| Bloqueador do canal de calcio | 3815782 | 0.702 |
| Antihipertensivo | 3797402 | 0.73 |
| Betabloqueador | 3791912 | 0.734 |
| Cirurgia Cardiovascular | 4196908 | 0.793 |
| Cardioversão/ Desfibrilação | 3771921 | 0.803 |
| Stent | 4205394 | 0.818 |
| Hipoglicemiante | 3801395 | 0.833 |
| Instalação de CEC | 4201190 | 0.835 |
| Drenagem de tórax e punção pericárdica ou pleural | 4203844 | 0.911 |
| Transplante cardíaco | 4205779 | 0.916 |
| Número de procedimentos na admissão T0 | 5761821 | 0.93 |
| Marca-passo temporário | 3768478 | 0.974 |
| Cavografia | 4205315 | 0.986 |
| Antiplaquetario VO | 3808023 | NaN |
| Hormonio tireoidiano | 3808023 | NaN |
| Broncodilator | 3808023 | NaN |

```

df_chisq <- tibble()

for (variable in intersect(columns_list$categorical_columns,
                           eligible_columns)){
  if (length(unique(df[[variable]])) > 1){
    test <- tryCatch(chisq.test(df[[outcome_column]],
                               df[[variable]] %>% replace_na('NA'), # counting NA as cat
                               simulate.p.value = TRUE),
                     error = function (cond) {
                       message("Can't calculate Chi Squared test for variable ", variable)
                       message(cond)
                       return(list(statistic = NaN, p.value = NaN))
                     })

    df_chisq <- bind_rows(df_chisq,
                         list("Variable" = variable,
                              "Statistic" = test$statistic,
                              "p-value" = test$p.value))
  }
}

significant_cat_cols <- df_chisq %>%
  filter(`p-value` <= threshold) %>%
  select(Variable) %>%
  pull

df_chisq <- df_chisq %>%
  arrange(`p-value`) %>%
  mutate(`Statistic` = round(`Statistic`, 3)) %>%
  rename_column('Variable')

df_chisq %>%
  mutate(`p-value` = case_when(`p-value` == 1 ~ sprintf('> 0%s999', getOption("OutDec")),
                               `p-value` < 0.001 ~ sprintf('< 0%s001', getOption("OutDec")),
                               TRUE ~ as.character(round(`p-value`, 3)))) %>%
  niceFormatting(caption = "Chi-squared test")

```

Table 3: Chi-squared test

| Variable | Statistic | p-value |
|--|-----------|---------|
| Sexo | 18.37 | < 0.001 |
| Escolaridade | 60.00 | < 0.001 |
| Doença cardíaca | 59.56 | < 0.001 |
| Doença cardíaca | 33.76 | < 0.001 |
| Classe funcional de IC | 113.27 | < 0.001 |
| Infarto do miocárdio prévio / Doença arterial coronariana | 29.79 | < 0.001 |
| Insuficiência cardíaca | 155.56 | < 0.001 |
| Fibrilação / flutter atrial | 28.42 | < 0.001 |
| Valvopatias/ Prótese valvares | 67.86 | < 0.001 |
| Diabetes mellitus | 35.73 | < 0.001 |
| Insuficiência renal crônica | 73.53 | < 0.001 |
| Tipo de Procedimento 1 | 36.50 | < 0.001 |
| Tipo de Reoperação 1 | 39.68 | < 0.001 |
| Tipo de Procedimento 1 | 39.68 | < 0.001 |
| Tipo de Dispositivo ao final do procedimento 1 | 164.29 | < 0.001 |
| Tipo de Dispositivo ao final do procedimento 1 | 118.28 | < 0.001 |
| Admissão em até 180 dias antes da T0 | 89.22 | < 0.001 |
| Hemodiálise | 24.98 | 0.001 |
| Acidente Vascular Cerebral/ Acidente isquêmico transitório prévios | 13.38 | 0.001 |
| Desfecho principal da admissão T0 | 12.97 | 0.001 |
| Hipertensão arterial | 13.35 | 0.002 |
| Doença pulmonar obstrutiva crônica | 9.66 | 0.003 |
| Parada cardíaca prévia/ Taquicardia ventricular instável | 5.56 | 0.018 |
| Neoplasia em tratamento ou tratada recentemente | 2.50 | 0.11 |
| Raça | 8.48 | 0.21 |
| Estado de residência | 29.73 | 0.326 |
| Transplante cardíaco prévio | 0.64 | 0.678 |
| Endocardite prévia | 0.04 | 0.854 |
| Óbito intraoperatório 1 | 0.35 | > 0.999 |

```
saveRDS(significant_cat_cols,
        file = sprintf("../EDA/auxiliar/significant_columns/categorical_%s.rds", outcome_column))

saveRDS(significant_num_cols,
        file = sprintf("../EDA/auxiliar/significant_columns/numerical_%s.rds", outcome_column))
```

```
## [1] 78
## [1] 23
## [1] 144
## [1] 58
```