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
```

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_parenteral', '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(.): the standard deviation is 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))
}

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

significant_num_cols <- df_wilcox %>%
  filter(`p-value` <= 0.3) %>%
  select(Variable) %>%
  pull

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	847009.0	< 0.001
Número de comorbidades	1749118.5	< 0.001
Culturas	1546569.5	< 0.001
Antagonista da Aldosterona	1356431.0	< 0.001
Diuretico	1251877.5	< 0.001
Exames laboratoriais	1327331.5	< 0.001
Quantidade de exames de análises clínicas	1327370.5	< 0.001
Equipe Multiprofissional	1381304.5	< 0.001
Ultrassom	1631507.0	< 0.001
Quantidade de classes medicamentosas de ação cardiovascular	743220.5	< 0.001
Quantidade de medicamentos de ação cardiovascular	1227111.0	< 0.001
Quantidade de exames diagnóstico por imagem	1362436.0	< 0.001
DVA	1381676.5	< 0.001
ECG	1389308.0	< 0.001
Quantidade de exames por métodos gráficos	1389836.0	< 0.001
Insuficiência cardíaca	1399314.5	< 0.001
Radiografias	1427085.5	< 0.001
Número da Admissão T0	2112425.5	< 0.001
Vasodilator	1394132.0	< 0.001
Antiarritmicos	1466966.5	< 0.001
Insulina	1574614.0	< 0.001
Tomografia	1680975.5	< 0.001
Ecocardiograma	1528524.0	< 0.001
Núm. de hospitalizações pré-procedimento	2138992.5	< 0.001

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

Variable	Statistic	p-value
UTI durante a admissão T0	2199105.0	< 0.001
Anticoagulantes orais	1615578.5	< 0.001
Psicofármacos	1434905.0	< 0.001
Citologias	1877557.0	< 0.001
Estatinas	1476555.0	< 0.001
Ano do procedimento 1	2127174.5	< 0.001
Ano da admissão T0	2118315.5	< 0.001
Cintilografia	1812433.5	< 0.001
Idade no momento do primeiro procedimento	2149989.0	< 0.001
Idade no Procedimento 1	2149989.0	< 0.001
Ressonancia magnetica	1789074.0	< 0.001
Antiplaquetario EV	1731514.5	< 0.001
Interconsulta médica	1752823.0	< 0.001
Holter	1771284.5	< 0.001
Diálise durante a admissão T0	2556060.0	< 0.001
Quantidade de antimicrobianos	1500475.0	< 0.001
Antibióticos	1502393.0	< 0.001
Quantidade de exames histopatológicos	1879837.0	< 0.001
Quantidade de procedimentos invasivos	1744527.5	< 0.001
Cateter venoso central	1873069.5	< 0.001
Transfusão de hemoderivados	1892782.5	0.001
Cateterismo	1818569.5	0.001
Diárias no serviço de Emergência na admissão T0	1120606.5	0.001
Aortografia	1915018.5	0.001
Intervenção coronária percutânea	1901847.0	0.006
Bomba de infusão contínua	1718524.0	0.017
IECA/BRA	1635714.5	0.02
Suporte cardiocirculatório	1916948.5	0.021
Ventilação não invasiva	1916975.5	0.022
Digoxina	1717475.0	0.023
Outros procedimentos cirúrgicos	1872032.0	0.033
Antifúngicos	1739577.0	0.049
Arteriografia	1925603.5	0.05
Angiografia	1921657.0	0.061
Teste de esforço	1950261.0	0.084
Tilt Test	1922402.0	0.114
Betabloqueador	1723267.0	0.138
Flebografia	1912151.5	0.201
Cavografia	1919424.0	0.235
Exames endoscópicos	1916911.0	0.271
Anticonvulsivante	1750518.0	0.3
Hipoglicemiante	1749668.0	0.317
Angioplastia	1927094.0	0.344
Polissonografia	1927094.5	0.344
Antihipertensivo	1751653.0	0.352
Angio RM	1936106.0	0.376
Eletrofisiologia	1913367.0	0.415
Antiviral	1766100.0	0.449
Transplante cardíaco	1927689.0	0.467
Drenagem de tórax e punção pericárdica ou pleural	1926109.0	0.502
Traqueostomia	1933275.0	0.557
Intervenção cardiovascular em laboratório de hemodinâmica	1927019.5	0.603
Trombolitico	1772770.0	0.609

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

Variable	Statistic	p-value
Antiretroviral	1772480.0	0.644
Biopsias	1927627.0	0.668
Instalação de CEC	1936040.5	0.692
PET-CT	1927918.0	0.699
Número de procedimentos na admissão T0	2594919.0	0.751
Cirurgia Toracica	1929317.0	0.76
Angio TC	1925567.5	0.767
Espirometria / Ergoespirometria	1932905.0	0.801
Marca-passo temporário	1750616.0	0.839
Stent	1931189.0	0.88
Cirurgia Cardiovascular	1928142.5	0.891
Cardioversão/ Desfibrilação	1753377.0	0.944
Bloqueador do canal de calcio	1771716.0	0.969
Antiplaquetario VO	1771175.0	NaN
Hormonio tireoidiano	1771175.0	NaN
Broncodiltador	1771175.0	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))
  }
}

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

significant_cat_cols <- df_chisq %>%
  filter(`p-value` <= 0.3) %>%
  select(Variable) %>%
  pull

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
Escolaridade	32.17	< 0.001
Doença cardíaca	26.90	< 0.001
Classe funcional de IC	91.24	< 0.001
Hipertensão arterial	32.64	< 0.001
Infarto do miocárdio prévio / Doença arterial coronariana	39.18	< 0.001
Insuficiência cardíaca	65.90	< 0.001
Fibrilação / flutter atrial	25.12	< 0.001
Valvopatias/ Prótese valvares	41.08	< 0.001
Diabetes melittus	58.60	< 0.001
Insuficiência renal crônica	74.51	< 0.001
Hemodiálise	46.66	< 0.001
Acidente Vascular Cerebral/ Acidente isquêmico transitório prévios	24.56	< 0.001
Tipo de Procedimento 1	21.96	< 0.001
Tipo de Dispositivo ao final do procedimento 1	74.16	< 0.001
Tipo de Dispositivo ao final do procedimento 1	69.37	< 0.001
Admissão em até 180 dias antes da T0	52.16	< 0.001
Doença cardíaca	32.08	< 0.001
Tipo de Procedimento 1	23.34	< 0.001
Tipo de Reoperação 1	23.34	0.003
Sexo	8.90	0.004
Neoplasia em tratamento ou tratada recentemente	9.31	0.012
Doença pulmonar obstrutiva crônica	6.92	0.017
Desfecho principal da admissão T0	5.51	0.022
Parada cardíaca prévia/ Taquicardia ventricular instável	3.85	0.06
Estado de residência	27.44	0.398
Raça	4.20	0.549
Endocardite prévia	0.26	0.781
Transplante cardíaco prévio	0.27	> 0.999
Óbito intraoperatório 1	0.15	> 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))
```