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_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))
}

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	2206162	< 0.001
Quantidade de exames diagnóstico por imagem	3509135	< 0.001
Número da Admissão T0	5594499	< 0.001
Radiografias	3634841	< 0.001
Quantidade de classes medicamentosas de ação cardiovascular	1892069	< 0.001
Quantidade de medicamentos de ação cardiovascular	3207001	< 0.001
Quantidade de exames por métodos gráficos	3696397	< 0.001
ECG	3719972	< 0.001
Equipe Multiprofissional	3823620	< 0.001
UTI durante a admissão T0	5852811	< 0.001
Antiarritmicos	3614202	< 0.001
Ecocardiograma	3917659	< 0.001
Ultrassom	4302707	< 0.001
DVA	3570018	< 0.001
Exames laboratoriais	3809629	< 0.001
Quantidade de exames de análises clínicas	3809814	< 0.001
Diuretico	3431767	< 0.001
Antagonista da Aldosterona	3621286	< 0.001
Núm. de hospitalizações pré-procedimento	5888541	< 0.001
Quantidade de procedimentos invasivos	4239531	< 0.001
Biopsias	4701222	< 0.001
Transplante cardíaco	4744190	< 0.001
Insuficiência cardíaca	3717640	< 0.001

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

Variable	Statistic	p-value
Culturas	4325301	< 0.001
Ressonancia magnetica	4475738	< 0.001
Anticoagulantes orais	3998260	< 0.001
Vasodilator	3714341	< 0.001
Cateterismo	4412942	< 0.001
Psicofármacos	3658792	< 0.001
Número de comorbidades	5958287	< 0.001
Antifúngicos	4097729	< 0.001
Suporte cardiocirculatório	4737426	< 0.001
Quantidade de exames histopatológicos	4678059	< 0.001
Antiviral	4195578	< 0.001
Cateter venoso central	4620862	< 0.001
Holter	4460779	< 0.001
Exames endoscópicos	4675980	< 0.001
Digoxina	4036938	< 0.001
Cintilografia	4596561	< 0.001
Betabloqueador	3963899	< 0.001
Tomografia	4481479	< 0.001
Diárias no serviço de Emergência na admissão T0	2480492	< 0.001
Quantidade de antimicrobianos	3754158	< 0.001
Antibióticos	3757923	< 0.001
Bloqueador do canal de calcio	4148968	< 0.001
Estatinas	3829690	< 0.001
Eletrofisiologia	4638403	< 0.001
IECA/BRA	3834341	< 0.001
Instalação de CEC	4713800	< 0.001
Bomba de infusão contínua	4082548	< 0.001
Outros procedimentos cirúrgicos	4614026	< 0.001
Insulina	4107341	< 0.001
Anticonvulsivante	4148288	< 0.001
Transfusão de hemoderivados	4737641	< 0.001
Citologias	4755503	0.001
Diálise durante a admissão T0	6810550	0.001
Angio TC	4710937	0.002
Angio RM	4775333	0.003
Idade no momento do primeiro procedimento	7218507	0.006
Idade no Procedimento 1	7218507	0.006
Espirometria / Ergoespirometria	4770913	0.008
Intervenção coronária percutânea	4760306	0.012
Antiplaquetario EV	4234141	0.013
Teste de esforço	4758793	0.014
Tilt Test	4780798	0.014
Arteriografia	4791836	0.02
Cirurgia Toracica	4781540	0.021
Cardioversão/ Desfibrilação	4158206	0.033
PET-CT	4775117	0.035
Angioplastia	4789058	0.048
Interconsulta médica	4708945	0.107
Ventilação não invasiva	4817016	0.121
Ano da admissão T0	7019535	0.135
Ano do procedimento 1	7041442	0.153
Antihipertensivo	4219363	0.163
Marca-passo temporário	4157330	0.174

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

Variable	Statistic	p-value
Flebografia	4770967	0.183
Intervenção cardiovascular em laboratório de hemodinâmica	4786112	0.191
Polissonografia	4795678	0.334
Trombolitico	4268704	0.411
Drenagem de tórax e punção pericárdica ou pleural	4810278	0.476
Antiretroviral	4261729	0.494
Cirurgia Cardiovascular	4817381	0.644
Angiografia	4798761	0.68
Aortografia	4798763	0.68
Número de procedimentos na admissão T0	6839454	0.746
Stent	4802405	0.804
Hipoglicemiante	4259805	0.883
Traqueostomia	4801161	0.886
Cavografia	4800697	0.932
Antiplaquetario VO	4264722	NaN
Hormonio tireoidiano	4264722	NaN
Broncodiltador	4264722	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
Escolaridade	27.50	< 0.001
Infarto do miocárdio prévio / Doença arterial coronariana	16.57	< 0.001
Insuficiência cardíaca	63.52	< 0.001
Tipo de Procedimento 1	79.09	< 0.001
Tipo de Reoperação 1	91.86	< 0.001
Tipo de Procedimento 1	91.86	< 0.001
Tipo de Dispositivo ao final do procedimento 1	88.64	< 0.001
Tipo de Dispositivo ao final do procedimento 1	38.37	< 0.001
Admissão em até 180 dias antes da T0	98.84	< 0.001
Desfecho principal da admissão T0	15.71	< 0.001
Classe funcional de IC	24.30	0.001
Doença cardíaca	28.98	0.002
Doença cardíaca	15.37	0.002
Fibrilação / flutter atrial	9.59	0.002
Transplante cardíaco prévio	15.38	0.006
Valvopatias/ Prótese valvares	6.16	0.013
Parada cardíaca prévia/ Taquicardia ventricular instável	6.63	0.014
Diabetes melittus	6.33	0.017
Hemodiálise	6.48	0.024
Endocardite prévia	2.41	0.119
Estado de residência	36.85	0.154
Insuficiência renal crônica	1.58	0.222
Neoplasia em tratamento ou tratada recentemente	1.02	0.41
Doença pulmonar obstrutiva crônica	0.60	0.465
Acidente Vascular Cerebral/ Acidente isquêmico transitório prévios	0.44	0.563
Hipertensão arterial	0.35	0.585
Raça	2.06	0.874
Sexo	0.03	0.894
Óbito intraoperatório 1	0.42	> 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] 19
```

```
## [1] 144
```

```
## [1] 70
```