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	3505151	< 0.001
Número da Admissão T0	8743766	< 0.001
Quantidade de medicamentos de ação cardiovascular	4886966	< 0.001
Quantidade de exames diagnóstico por imagem	5656440	< 0.001
Antiarritmicos	5551729	< 0.001
Quantidade de classes medicamentosas de ação cardiovascular	3007520	< 0.001
DVA	5489048	< 0.001
Quantidade de exames por métodos gráficos	5825363	< 0.001
Radiografias	5860553	< 0.001
UTI durante a admissão T0	9262145	< 0.001
ECG	5873147	< 0.001
Antagonista da Aldosterona	5609445	< 0.001
Equipe Multiprofissional	6032727	< 0.001
Insuficiência cardíaca	5554513	< 0.001
Exames laboratoriais	5910127	< 0.001
Quantidade de exames de análises clínicas	5910723	< 0.001
Diuretico	5370033	< 0.001
Ultrassom	6780199	< 0.001
Núm. de hospitalizações pré-procedimento	9197237	< 0.001
Transplante cardíaco	7431750	< 0.001
Ecocardiograma	6340792	< 0.001
Biopsias	7366318	< 0.001
Quantidade de procedimentos invasivos	6659612	< 0.001

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

Variable	Statistic	p-value
Culturas	6770465	< 0.001
Número de comorbidades	9225813	< 0.001
Quantidade de exames histopatológicos	7315755	< 0.001
Cateterismo	6898638	< 0.001
Anticoagulantes orais	6278302	< 0.001
Ressonancia magnetica	7051356	< 0.001
Holter	6951056	< 0.001
Psicofármacos	5765074	< 0.001
Vasodilator	5860701	< 0.001
Antiviral	6588344	< 0.001
Tomografia	6973425	< 0.001
Cateter venoso central	7248483	< 0.001
Cintilografia	7182434	< 0.001
Quantidade de antimicrobianos	5776176	< 0.001
Antibióticos	5782588	< 0.001
Estatinas	5928379	< 0.001
Digoxina	6364111	< 0.001
Bloqueador do canal de calcio	6505132	< 0.001
Exames endoscópicos	7353950	< 0.001
Antifúngicos	6485399	< 0.001
Betabloqueador	6272007	< 0.001
IECA/BRA	5949670	< 0.001
Eletrofisiologia	7276612	< 0.001
Diárias no serviço de Emergência na admissão T0	4061492	< 0.001
Antiplaquetario EV	6618990	< 0.001
Outros procedimentos cirúrgicos	7268545	< 0.001
Bomba de infusão contínua	6460618	< 0.001
Instalação de CEC	7416797	< 0.001
Citologias	7452002	< 0.001
Suporte cardiocirculatório	7482158	< 0.001
Insulina	6476057	< 0.001
Idade no momento do primeiro procedimento	11551514	< 0.001
Idade no Procedimento 1	11551514	< 0.001
Intervenção coronária percutânea	7450866	< 0.001
Transfusão de hemoderivados	7448073	< 0.001
Anticonvulsivante	6566585	< 0.001
Diálise durante a admissão T0	10771004	< 0.001
Angio RM	7498943	< 0.001
Espirometria / Ergoespirometria	7493516	0.003
Cardioversão/ Desfibrilação	6561263	0.004
Angio TC	7434508	0.005
Interconsulta médica	7337473	0.006
Tilt Test	7510325	0.013
Antiretroviral	6688604	0.014
Teste de esforço	7485620	0.019
Flebografia	7469550	0.02
PET-CT	7500403	0.021
Intervenção cardiovascular em laboratório de hemodinâmica	7503339	0.026
Ventilação não invasiva	7561748	0.043
Angioplastia	7522048	0.065
Marca-passo temporário	6563291	0.1
Arteriografia	7529019	0.133
Número de procedimentos na admissão T0	10774821	0.183

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

Variable	Statistic	p-value
Ano da admissão T0	10984271	0.229
Ano do procedimento 1	11016782	0.25
Cirurgia Toracica	7524811	0.264
Drenagem de tórax e punção pericárdica ou pleural	7552528	0.292
Polissonografia	7528669	0.297
Aortografia	7527061	0.303
Trombolitico	6696043	0.324
Cirurgia Cardiovascular	7567473	0.468
Antihipertensivo	6675546	0.516
Hipoglicemiante	6678666	0.576
Cavografia	7527137	0.601
Traqueostomia	7540049	0.71
Angiografia	7533699	0.719
Stent	7537884	0.747
Antiplaquetario VO	6702032	NaN
Hormonio tireoidiano	6702032	NaN
Broncodiltador	6702032	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	40.01	< 0.001
Doença cardíaca	64.20	< 0.001
Doença cardíaca	34.21	< 0.001
Classe funcional de IC	47.38	< 0.001
Infarto do miocárdio prévio / Doença arterial coronariana	28.39	< 0.001
Insuficiência cardíaca	136.36	< 0.001
Fibrilação / flutter atrial	15.94	< 0.001
Parada cardíaca prévia/ Taquicardia ventricular instável	19.41	< 0.001
Valvopatias/ Prótese valvares	15.33	< 0.001
Tipo de Procedimento 1	121.85	< 0.001
Tipo de Reoperação 1	139.58	< 0.001
Tipo de Procedimento 1	139.58	< 0.001
Tipo de Dispositivo ao final do procedimento 1	218.03	< 0.001
Tipo de Dispositivo ao final do procedimento 1	96.18	< 0.001
Admissão em até 180 dias antes da T0	168.52	< 0.001
Desfecho principal da admissão T0	26.86	< 0.001
Transplante cardíaco prévio	20.97	< 0.001
Diabetes melittus	11.26	< 0.001
Sexo	10.30	0.002
Hemodiálise	8.45	0.013
Estado de residência	52.50	0.014
Insuficiência renal crônica	5.24	0.023
Endocardite prévia	4.47	0.043
Doença pulmonar obstrutiva crônica	3.21	0.084
Neoplasia em tratamento ou tratada recentemente	1.93	0.194
Acidente Vascular Cerebral/ Acidente isquêmico transitório prévios	0.87	0.391
Hipertensão arterial	0.33	0.576
Raça	4.38	0.61
Óbito intraoperatório 1	0.72	0.631

```
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] 24
```

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

```
## [1] 74
```