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
Número da Admissão T0	11160003	< 0.001
Quantidade de classes medicamentosas utilizadas	4581393	< 0.001
Quantidade de medicamentos de ação cardiovascular	6411970	< 0.001
Quantidade de exames diagnóstico por imagem	7426412	< 0.001
Quantidade de classes medicamentosas de ação cardiovascular	3896762	< 0.001
Núm. de hospitalizações pré-procedimento	11672486	< 0.001
Antiarritmicos	7296807	< 0.001
Quantidade de exames por métodos gráficos	7542502	< 0.001
DVA	7167456	< 0.001
ECG	7618336	< 0.001
Equipe Multiprofissional	7818573	< 0.001
Insuficiência cardíaca	7188501	< 0.001
Antagonista da Aldosterona	7273832	< 0.001
UTI durante a admissão T0	12176855	< 0.001
Diuretico	6937008	< 0.001
Exames laboratoriais	7701962	< 0.001
Quantidade de exames de análises clínicas	7702891	< 0.001
Radiografias	7751583	< 0.001
Ultrassom	8841632	< 0.001
Ecocardiograma	8253132	< 0.001
Holter	8870338	< 0.001
Quantidade de procedimentos invasivos	8684567	< 0.001
Biopsias	9550337	< 0.001

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

Variable	Statistic	p-value
Número de comorbidades	12156588	< 0.001
Transplante cardíaco	9632925	< 0.001
Ressonancia magnetica	9112218	< 0.001
Cateterismo	8964854	< 0.001
Psicofármacos	7484464	< 0.001
Anticoagulantes orais	8161569	< 0.001
Culturas	8904264	< 0.001
Quantidade de exames histopatológicos	9503326	< 0.001
Tomografia	9061021	< 0.001
Vasodilator	7691487	< 0.001
Cateter venoso central	9394909	< 0.001
Cintilografia	9319038	< 0.001
Estatinas	7660021	< 0.001
Quantidade de antimicrobianos	7545157	< 0.001
Antibióticos	7562438	< 0.001
Digoxina	8240059	< 0.001
Bloqueador do canal de calcio	8418223	< 0.001
Antiviral	8542832	< 0.001
Eletrofisiologia	9369088	< 0.001
IECA/BRA	7725944	< 0.001
Exames endoscópicos	9534421	< 0.001
Bomba de infusão contínua	8316602	< 0.001
Antifúngicos	8420932	< 0.001
Diárias no serviço de Emergência na admissão T0	5219540	< 0.001
Betabloqueador	8227091	< 0.001
Instalação de CEC	9589283	< 0.001
Outros procedimentos cirúrgicos	9427237	< 0.001
Intervenção coronária percutânea	9625715	< 0.001
Suporte cardiocirculatório	9673705	< 0.001
Antiplaquetario EV	8577460	< 0.001
Idade no momento do primeiro procedimento	14988108	< 0.001
Idade no Procedimento 1	14988108	< 0.001
Espirometria / Ergoespirometria	9669024	< 0.001
Transfusão de hemoderivados	9628707	< 0.001
Angio RM	9689412	< 0.001
Citologias	9660226	< 0.001
Angio TC	9582023	< 0.001
Cardioversão/ Desfibrilação	8485050	< 0.001
Insulina	8452934	< 0.001
Angioplastia	9709114	< 0.001
Arteriografia	9720577	0.002
Intervenção cardiovascular em laboratório de hemodinâmica	9686578	0.002
Anticonvulsivante	8529225	0.003
Diálise durante a admissão T0	14048229	0.004
Ano da admissão T0	14470565	0.028
Tilt Test	9713421	0.028
Interconsulta médica	9561432	0.029
Flebografia	9668818	0.031
Ano do procedimento 1	14509474	0.033
Teste de esforço	9689764	0.043
PET-CT	9704773	0.05
Ventilação não invasiva	9766558	0.057
Antiretroviral	8647148	0.063

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

Variable	Statistic	p-value
Marca-passo temporário	8494219	0.07
Aortografia	9722464	0.111
Cirurgia Toracica	9721714	0.141
Polissonografia	9728991	0.224
Número de procedimentos na admissão T0	14053236	0.282
Traqueostomia	9733096	0.395
Antihipertensivo	8623535	0.447
Cirurgia Cardiovascular	9772230	0.5
Trombolitico	8654989	0.584
Hipoglicemiante	8632775	0.584
Stent	9741204	0.703
Drenagem de tórax e punção pericárdica ou pleural	9734408	0.718
Angiografia	9742356	0.859
Cavografia	9738336	0.927
Antiplaquetario VO	8658767	NaN
Hormonio tireoidiano	8658767	NaN
Broncodiltador	8658767	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	23.15	< 0.001
Escolaridade	51.50	< 0.001
Doença cardíaca	97.97	< 0.001
Doença cardíaca	46.32	< 0.001
Classe funcional de IC	64.73	< 0.001
Infarto do miocárdio prévio / Doença arterial coronariana	30.86	< 0.001
Insuficiência cardíaca	189.25	< 0.001
Fibrilação / flutter atrial	14.24	< 0.001
Parada cardíaca prévia/ Taquicardia ventricular instável	18.71	< 0.001
Transplante cardíaco prévio	20.20	< 0.001
Valvopatias/ Prótese valvares	19.95	< 0.001
Tipo de Procedimento 1	129.22	< 0.001
Tipo de Reoperação 1	153.74	< 0.001
Tipo de Procedimento 1	153.74	< 0.001
Tipo de Dispositivo ao final do procedimento 1	290.68	< 0.001
Tipo de Dispositivo ao final do procedimento 1	121.11	< 0.001
Admissão em até 180 dias antes da T0	244.84	< 0.001
Desfecho principal da admissão T0	37.65	< 0.001
Estado de residência	61.25	< 0.001
Insuficiência renal crônica	10.95	0.001
Diabetes melittus	8.44	0.005
Hemodiálise	7.44	0.014
Raça	14.72	0.035
Endocardite prévia	3.92	0.051
Doença pulmonar obstrutiva crônica	3.05	0.087
Acidente Vascular Cerebral/ Acidente isquêmico transitório prévios	1.33	0.275
Neoplasia em tratamento ou tratada recentemente	1.01	0.305
Óbito intraoperatório 1	1.01	0.6
Hipertensão arterial	0.23	0.64

```
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] 25
```

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

```
## [1] 77
```