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 de comorbidades	2755344	< 0.001
Quantidade de classes medicamentosas utilizadas	1364262	< 0.001
Diuretico	1934446	< 0.001
Quantidade de classes medicamentosas de ação cardiovascular	1131901	< 0.001
Antagonista da Aldosterona	2118491	< 0.001
Ultrassom	2534481	< 0.001
Exames laboratoriais	2097018	< 0.001
Quantidade de exames de análises clínicas	2097552	< 0.001
Insuficiência cardíaca	2108613	< 0.001
Equipe Multiprofissional	2184145	< 0.001
Quantidade de exames diagnóstico por imagem	2121773	< 0.001
Quantidade de medicamentos de ação cardiovascular	1934855	< 0.001
DVA	2130494	< 0.001
ECG	2155713	< 0.001
Culturas	2474234	< 0.001
Quantidade de exames por métodos gráficos	2158922	< 0.001
Número da Admissão T0	3242783	< 0.001
Radiografias	2210965	< 0.001
Antiarritmicos	2291141	< 0.001
Núm. de hospitalizações pré-procedimento	3312021	< 0.001
Tomografia	2624403	< 0.001
Ecocardiograma	2440787	< 0.001
UTI durante a admissão T0	3453775	< 0.001

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

Variable	Statistic	p-value
Anticoagulantes orais	2502242	< 0.001
Vasodilator	2278551	< 0.001
Psicofármacos	2256240	< 0.001
Cintilografia	2797889	< 0.001
Citologias	2901540	< 0.001
Insulina	2521544	< 0.001
Quantidade de antimicrobianos	2307375	< 0.001
Antibióticos	2307544	< 0.001
Ano do procedimento 1	3437087	< 0.001
Ano da admissão T0	3422504	< 0.001
Ressonancia magnetica	2788387	< 0.001
Holter	2754427	< 0.001
Idade no momento do primeiro procedimento	3459589	< 0.001
Idade no Procedimento 1	3459589	< 0.001
Estatinas	2386544	< 0.001
Digoxina	2572999	< 0.001
Diálise durante a admissão T0	3962826	< 0.001
Quantidade de procedimentos invasivos	2710831	< 0.001
Quantidade de exames histopatológicos	2908814	< 0.001
Bomba de infusão contínua	2627344	< 0.001
Cateter venoso central	2896928	< 0.001
Antiplaquetario EV	2690069	< 0.001
IECA/BRA	2481959	< 0.001
Cateterismo	2838386	0.001
Interconsulta médica	2852569	0.007
Outros procedimentos cirúrgicos	2892936	0.018
Ventilação não invasiva	2956942	0.022
Antifúngicos	2679401	0.025
Aortografia	2960522	0.026
Angiografia	2960532	0.026
Teste de esforço	3004365	0.03
Exames endoscópicos	2941704	0.04
Intervenção coronária percutânea	2948251	0.046
Diárias no serviço de Emergência na admissão T0	1785293	0.06
Tilt Test	2961695	0.062
Transfusão de hemoderivados	2948056	0.07
Flebografia	2941730	0.075
Cirurgia Toracica	2962141	0.08
PET-CT	2957048	0.084
Antiviral	2710531	0.106
Angio TC	2939462	0.128
Suporte cardiocirculatório	2963538	0.154
Arteriografia	2969604	0.167
Antihipertensivo	2696232	0.286
Cavografia	2964478	0.415
Hipoglicemiante	2703471	0.442
Drenagem de tórax e punção pericárdica ou pleural	2967473	0.449
Traqueostomia	2977860	0.46
Espirometria / Ergoespirometria	2967741	0.47
Trombolitico	2726464	0.521
Betabloqueador	2699567	0.542
Antiretroviral	2726012	0.561
Anticonvulsivante	2709651	0.562

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

Variable	Statistic	p-value
Angioplastia	2971930	0.637
Polissonografia	2971930	0.637
Instalação de CEC	2980063	0.716
Eletrofisiologia	2965618	0.745
Transplante cardíaco	2972859	0.791
Cardioversão/ Desfibrilação	2696984	0.797
Número de procedimentos na admissão T0	4003303	0.804
Biopsias	2976471	0.832
Stent	2974605	0.849
Angio RM	2975640	0.858
Cirurgia Cardiovascular	2971188	0.903
Intervenção cardiovascular em laboratório de hemodinâmica	2975529	0.904
Marca-passo temporário	2699620	0.932
Bloqueador do canal de calcio	2723442	0.975
Antiplaquetario VO	2723978	NaN
Hormonio tireoidiano	2723978	NaN
Broncodiltador	2723978	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	16.27	< 0.001
Escolaridade	32.62	< 0.001
Doença cardíaca	48.60	< 0.001
Doença cardíaca	36.31	< 0.001
Classe funcional de IC	115.13	< 0.001
Hipertensão arterial	34.98	< 0.001
Infarto do miocárdio prévio / Doença arterial coronariana	38.80	< 0.001
Insuficiência cardíaca	112.89	< 0.001
Fibrilação / flutter atrial	30.29	< 0.001
Valvopatias/ Prótese valvares	80.68	< 0.001
Diabetes melittus	57.20	< 0.001
Insuficiência renal crônica	90.72	< 0.001
Hemodiálise	40.80	< 0.001
Acidente Vascular Cerebral/ Acidente isquêmico transitório prévios	14.07	< 0.001
Tipo de Procedimento 1	27.40	< 0.001
Tipo de Reoperação 1	30.70	< 0.001
Tipo de Procedimento 1	30.70	< 0.001
Tipo de Dispositivo ao final do procedimento 1	108.05	< 0.001
Tipo de Dispositivo ao final do procedimento 1	89.89	< 0.001
Admissão em até 180 dias antes da T0	47.02	< 0.001
Doença pulmonar obstrutiva crônica	14.63	< 0.001
Desfecho principal da admissão T0	8.74	0.006
Parada cardíaca prévia/ Taquicardia ventricular instável	6.88	0.011
Neoplasia em tratamento ou tratada recentemente	5.16	0.035
Raça	4.95	0.455
Estado de residência	20.73	0.634
Endocardite prévia	0.49	0.636
Transplante cardíaco prévio	0.43	> 0.999
Óbito intraoperatório 1	0.23	> 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] 24
```

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

```
## [1] 62
```