Correlations - death_3year

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
print(threshold)
```

```
## [1] 0.1
```

```
df[columns_list$outcome_columns] <- lapply(df[columns_list$outcome_columns], as.character)
df[columns_list$outcome_columns] <- lapply(df[columns_list$outcome_columns], as.integer)
```

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

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
Quantidade de exames de análises clínicas	Exames laboratoriais	1.00
Quantidade de exames de análises clínicas	Quantidade de exames diagnóstico por imagem	0.93
Biopsias	Quantidade de exames histopatológicos	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	2034009	< 0.001
Número da Admissão T0	5082513	< 0.001
Antagonista da Aldosterona	3345280	< 0.001
Quantidade de classes medicamentosas de ação cardiovascular	1722010	< 0.001
Insuficiência cardíaca	3347435	< 0.001
Diuretico	3135912	< 0.001
Quantidade de medicamentos de ação cardiovascular	3073062	< 0.001
Núm. de hospitalizações pré-procedimento	5287875	< 0.001
DVA	3524957	< 0.001
Antiarritmicos	3650603	< 0.001
Exames laboratoriais	3649933	< 0.001
Quantidade de exames de análises clínicas	3651377	< 0.001
Ultrassom	4334058	< 0.001
Quantidade de exames diagnóstico por imagem	3708821	< 0.001
Quantidade de exames por métodos gráficos	3712509	< 0.001
ECG	3718904	< 0.001
Número de comorbidades	5278593	< 0.001
Equipe Multiprofissional	3842870	< 0.001
UTI durante a admissão T0	5798095	< 0.001
Radiografias	3849785	< 0.001

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

Variable	Statistic	p-value
Culturas	4305842	< 0.001
Anticoagulantes orais	4074720	< 0.001
Digoxina	4090119	< 0.001
Ecocardiograma	4160407	< 0.001
Psicofármacos	3698149	< 0.001
Vasodilator	3791313	< 0.001
Cintilografia	4663236	< 0.001
Tomografia	4522663	< 0.001
Quantidade de antimicrobianos	3770435	< 0.001
Antibióticos	3774954	< 0.001
Ressonancia magnetica	4621764	< 0.001
Quantidade de procedimentos invasivos	4457829	< 0.001
Estatinas	3864771	< 0.001
Diálise durante a admissão T0	6719935	< 0.001
Citologias	4846574	< 0.001
Holter	4586955	< 0.001
Insulina	4168429	< 0.001
Bomba de infusão contínua	4246381	< 0.001
IECA/BRA	3901168	< 0.001
Cateterismo	4656021	< 0.001
Quantidade de exames histopatológicos	4855588	< 0.001
Idade no momento do primeiro procedimento	6207045	< 0.001
Idade no Procedimento 1	6207045	< 0.001
Cateter venoso central	4812604	< 0.001
Outros procedimentos cirúrgicos	4757490	< 0.001
Antiplaquetario EV	4389998	< 0.001
Transfusão de hemoderivados	4869600	< 0.001
Diárias no serviço de Emergência na admissão T0	2736406	0.001
Exames endoscópicos	4870636	0.002
Antifúngicos	4368761	0.003
Intervenção coronária percutânea	4879623	0.004
Angio TC	4850116	0.007
Flebografia	4866693	0.009
Eletrofisiologia	4839483	0.01
PET-CT	4899142	0.021
Tilt Test	4906994	0.021
Ano do procedimento 1	6488652	0.029
Ano da admissão T0	6474177	0.033
Angioplastia	4914383	0.05
Angiografia	4911494	0.051
Teste de esforço	4959526	0.067
Anticonvulsivante	4380933	0.083
Suporte cardiocirculatório	4913080	0.103
Ventilação não invasiva	4913110	0.104
Antiviral	4418402	0.148
Cardioversão/ Desfibrilação	4370591	0.15
Aortografia	4917195	0.198
Intervenção cardiovascular em laboratório de hemodinâmica	4912619	0.237
Cirurgia Toracica	4917994	0.256
Polissonografia	4920497	0.294
Interconsulta médica	4868646	0.31
Cirurgia Cardiovascular	4897670	0.368
Arteriografia	4923384	0.374

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

Variable	Statistic	p-value
Espirometria / Ergoespirometria	4917474	0.417
Bloqueador do canal de calcio	4415554	0.43
Marca-passo temporário	4371734	0.44
Trombolitico	4435585	0.464
Antiretroviral	4435585	0.464
Hipoglicemiante	4408327	0.476
Instalação de CEC	4916266	0.573
Biopsias	4931748	0.7
Betabloqueador	4414660	0.724
Cavografia	4931977	0.745
Transplante cardíaco	4929025	0.787
Traqueostomia	4925802	0.822
Antihipertensivo	4439322	0.835
Angio RM	4927345	0.976
Drenagem de tórax e punção pericárdica ou pleural	4927245	0.988
Número de procedimentos na admissão T0	6780614	0.993
Antiplaquetario VO	4432505	NaN
Hormonio tireoidiano	4432505	NaN
Broncodiltador	4432505	NaN
Stent	4927078	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	23.76	< 0.001
Escolaridade	86.94	< 0.001
Doença cardíaca	77.56	< 0.001
Doença cardíaca	39.60	< 0.001
Classe funcional de IC	126.49	< 0.001
Infarto do miocárdio prévio / Doença arterial coronariana	32.80	< 0.001
Insuficiência cardíaca	195.04	< 0.001
Fibrilação / flutter atrial	20.88	< 0.001
Valvopatias/ Prótese valvares	67.01	< 0.001
Diabetes melittus	37.77	< 0.001
Insuficiência renal crônica	71.19	< 0.001
Tipo de Procedimento 1	34.13	< 0.001
Tipo de Reoperação 1	36.98	< 0.001
Tipo de Procedimento 1	36.98	< 0.001
Tipo de Dispositivo ao final do procedimento 1	195.17	< 0.001
Tipo de Dispositivo ao final do procedimento 1	134.87	< 0.001
Admissão em até 180 dias antes da T0	129.79	< 0.001
Hemodiálise	23.19	0.001
Acidente Vascular Cerebral/ Acidente isquêmico transitório prévios	11.10	0.001
Doença pulmonar obstrutiva crônica	8.15	0.009
Hipertensão arterial	5.70	0.019
Raça	15.42	0.027
Parada cardíaca prévia/ Taquicardia ventricular instável	4.13	0.04
Neoplasia em tratamento ou tratada recentemente	3.24	0.114
Estado de residência	26.05	0.464
Endocardite prévia	0.07	0.845
Transplante cardíaco prévio	0.14	> 0.999

```
dir.create(file.path("./auxiliar/significant_columns/"), showWarnings = FALSE)

saveRDS(significant_cat_cols,
        file = sprintf("./auxiliar/significant_columns/categorical_%s.rds", outcome_column))

saveRDS(significant_num_cols,
        file = sprintf("./auxiliar/significant_columns/numerical_%s.rds", outcome_column))
```

```
## [1] 78
```

```
## [1] 23
```

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

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