Data Summary

Matheus Hoffmann Fernandes Santos

April 2018

Collect

Create data frames from tables, which are comorbidity, demography, non-hematological and SNP(single nucleotide polymorphism)

```
setwd("C:\\Users\\Hoffmann\\Desktop\\dados_dissert")

df_como <- read.csv("comorbidade.csv", sep = ";", stringsAsFactors = FALSE)
df_demo <- read.csv("dadosdemo.csv", sep = ";",stringsAsFactors = FALSE)
df_nhem <- read.csv("nhemato.csv", sep = ";", stringsAsFactors = FALSE)
df_snp <- read.csv("snpcorrect.csv", sep = ";",stringsAsFactors = FALSE)

cat("Table Summary: Observations and Variables\n")
```

```
## Table Summary: Observations and Variables
```

```
cat(sprintf("Comorbidity ------------ %i | %i\n",nrow(df_como),ncol(df_como)))
```

```
## Comorbidity ------------ 205 | 10
```

```
cat(sprintf("Demographic ------------- %i | %i\n",nrow(df_demo),ncol(df_demo)))
```

```
## Demographic ------------- 494 | 20
```

```
cat(sprintf("Non-hemotological ------ %i | %i\n",nrow(df_nhem),ncol(df_nhem)))
```

```
## Non-hemotological ------ 278 | 219
```

```
cat(sprintf("SNP --------------------- %i | %i\n",nrow(df_snp),ncol(df_snp)))
```

```
## SNP --------------------- 271 | 15
```

Missing Values

Showing percentage of missing values, by table.

Comorbidity data

```
#IMC has one value 999
count <- sqldf("SELECT COUNT(*) AS total FROM df_como")

queryBuilderMissing <- function(column, total, df){
 if(column == "IMC"){
      result <- sprintf("SELECT (1 - CAST(COUNT(%s) AS FLOAT)/%i) AS missing FROM %s WHERE %s < 60 OR %s
                    column, total, df,column, column, column)
  } else {
      result <- sprintf("SELECT (1 - CAST(COUNT(%s) AS FLOAT)/%i) AS missing FROM %s WHERE %s <> '' OR %
                    column, total, df, column, column)
```

```
  }

  #print(result)
  return(result)
}
getMissingByHeader <- function(){
  results <- c()
  percent <- "%"
  for(index in colnames(df_como)){
    missing_in_column <- sqldf(queryBuilderMissing(index, count$total, "df_como"))
    print(sprintf("%s missing values ratio >> %.2f%s", index, missing_in_column*100, percent))
    #results <- c(results, missing_in_column)
  }
  #return(results)
}

getMissingByHeader()
```

```
## [1] "prontuario missing values ratio >> 0.00%"
## [1] "protocolo missing values ratio >> 0.00%"
## [1] "idade missing values ratio >> 0.00%"
## [1] "IMC missing values ratio >> 10.24%"
## [1] "cor missing values ratio >> 0.00%"
## [1] "menopausa missing values ratio >> 1.95%"
## [1] "comorbidade missing values ratio >> 0.49%"
## [1] "hipertensao missing values ratio >> 0.49%"
## [1] "diabetes missing values ratio >> 1.46%"
## [1] "dislipidemia missing values ratio >> 0.98%"
```

Demographic data

```
count <- sqldf("SELECT COUNT(*) AS total FROM df_demo")
#One register filho = 99
queryBuilderMissing <- function(column, total, df){
 if(column == "filho"){
      result <- sprintf("SELECT (1 - CAST(COUNT(%s) AS FLOAT)/%i) AS missing FROM %s WHERE %s < 20 OR %s
                    column, total, df,column, column, column)
  } else {
      result <- sprintf("SELECT (1 - CAST(COUNT(%s) AS FLOAT)/%i) AS missing FROM %s WHERE %s <> '' OR %
                    column, total, df, column, column)
  }
  #print(result)
  return(result)
}
getMissingByHeader <- function(){
  results <- c()
  percent <- "%"
  for(index in colnames(df_demo)){
    missing_in_column <- sqldf(queryBuilderMissing(index, count$total, "df_demo"))
    print(sprintf("%s missing values ratio >> %.2f%s", index, missing_in_column*100, percent))
    #results <- c(results, missing_in_column)
  }
```

```
  #return(results)
}

getMissingByHeader()
```

```
## [1] "prontuario missing values ratio >> 0.00%"
## [1] "idade missing values ratio >> 0.00%"
## [1] "peso missing values ratio >> 35.83%"
## [1] "fumo missing values ratio >> 1.62%"
## [1] "alcool missing values ratio >> 4.05%"
## [1] "altura missing values ratio >> 38.46%"
## [1] "estcivil missing values ratio >> 0.40%"
## [1] "nivelescol missing values ratio >> 1.42%"
## [1] "ocupa missing values ratio >> 4.45%"
## [1] "vincu missing values ratio >> 28.95%"
## [1] "cor missing values ratio >> 1.82%"
## [1] "menarca missing values ratio >> 1.01%"
## [1] "idadegestaçao missing values ratio >> 12.35%"
## [1] "filho missing values ratio >> 10.32%"
## [1] "menopausa missing values ratio >> 2.02%"
## [1] "usocontracephormonal missing values ratio >> 0.81%"
## [1] "usorephormonal missing values ratio >> 2.02%"
## [1] "antecendcancer missing values ratio >> 2.02%"
## [1] "cardiov missing values ratio >> 0.40%"
## [1] "comorbidade missing values ratio >> 4.45%"
```

Single Nucleotide Polymorphism data

Old column names renamed to allow sqldf use, columnnames containing some special chars are mapped to '.' Changes: '>' = bt, ''='','-'=" Obs: Numeric (first letter)headers are mapped to X{{Name}}

```
count <- sqldf("SELECT COUNT(*) AS total FROM df_snp")

queryBuilderMissing <- function(column, total, df){
  if(column == "N" || column == "Pront"){
          result <- sprintf("SELECT (1 - CAST(COUNT(%s) AS FLOAT)/%i) AS missing FROM %s WHERE %s<>'' O
  } else {
      result <- sprintf("SELECT (1 - CAST(COUNT(%s) AS FLOAT)/%i) AS missing FROM %s WHERE %s='0' OR %s=
  }

  #print(result)
  return(result)
}
getMissingByHeader <- function(){
  #colnames old c("N", "Pront", "15631 G>T", "18053 A>G", "25505C>T", "6986 A>G", "C1236T", "G2677T/A",
  #results <- c()
  percent <- "%"
  for(index in colnames(df_snp)){
    missing_in_column <- sqldf(queryBuilderMissing(index, count$total, "df_snp"))
    print(sprintf("%s missing values ratio >> %.2f%s", index, missing_in_column*100, percent))
    #results <- c(results, missing_in_column)
  }
  #return(results)
```

```
}

getMissingByHeader()
```

```
## [1] "N missing values ratio >> 0.00%"
## [1] "Pront missing values ratio >> 0.74%"
## [1] "X15631GbtT missing values ratio >> 16.61%"
## [1] "X18053AbtG missing values ratio >> 10.33%"
## [1] "X25505CbtT missing values ratio >> 9.23%"
## [1] "X6986AbtG missing values ratio >> 4.43%"
## [1] "C1236T missing values ratio >> 11.07%"
## [1] "G2677TorA missing values ratio >> 13.28%"
## [1] "C3435T missing values ratio >> 5.17%"
## [1] "AbtGILE105VAL missing values ratio >> 3.69%"
## [1] "X02Ala114Val missing values ratio >> 12.18%"
## [1] "X11GbtA missing values ratio >> 11.07%"
## [1] "Ile655ValAbtG missing values ratio >> 48.34%"
## [1] "AbtG missing values ratio >> 6.64%"
## [1] "CbtG missing values ratio >> 13.65%"
```

Non-hematological data

Check if percentage by reaction make more sense.

```
count <- sqldf("SELECT COUNT(*) AS total FROM df_nhem")

queryBuilderMissing <- function(column, total, df){
  if(column == "Prontuario" || column == "Protocolo"){
          result <- sprintf("SELECT (1 - CAST(COUNT(%s) AS FLOAT)/%i) AS missing FROM %s WHERE %s<>'' O
  } else {
      result <- sprintf("SELECT (1 - CAST(COUNT(%s) AS FLOAT)/%i) AS missing FROM %s WHERE %s='0' OR %s=
  }

  #print(result)
  return(result)
}
getMissingByHeader <- function(){
  percent <- "%"
  for(index in colnames(df_nhem)){
    missing_in_column <- sqldf(queryBuilderMissing(index, count$total, "df_nhem"))
    print(sprintf("%s missing values ratio >> %.2f%s", index, missing_in_column*100, percent))
    #results <- c(results, missing_in_column)
  }
  #return(results)
}

getMissingByHeader()
```

```
## [1] "Prontuario missing values ratio >> 0.00%"
## [1] "Protocolo missing values ratio >> 0.00%"
## [1] "Fad0 missing values ratio >> 6.83%"
## [1] "Fad1 missing values ratio >> 6.83%"
## [1] "Fad2 missing values ratio >> 11.15%"
## [1] "Fad3 missing values ratio >> 17.99%"
```

```
## [1] "Fad4 missing values ratio >> 23.38%"
## [1] "Fad5 missing values ratio >> 36.33%"
## [1] "Fad6 missing values ratio >> 39.57%"
## [1] "Fraq0 missing values ratio >> 7.55%"
## [1] "Fraq1 missing values ratio >> 6.83%"
## [1] "Fraq2 missing values ratio >> 11.51%"
## [1] "Fraq3 missing values ratio >> 17.99%"
## [1] "Fraq4 missing values ratio >> 23.38%"
## [1] "Fraq5 missing values ratio >> 36.33%"
## [1] "Fraq6 missing values ratio >> 39.57%"
## [1] "Disp0 missing values ratio >> 6.83%"
## [1] "Disp1 missing values ratio >> 7.19%"
## [1] "Disp2 missing values ratio >> 11.15%"
## [1] "Disp3 missing values ratio >> 17.27%"
## [1] "Disp4 missing values ratio >> 24.10%"
## [1] "Disp5 missing values ratio >> 37.05%"
## [1] "Disp6 missing values ratio >> 39.57%"
## [1] "Anor0 missing values ratio >> 6.83%"
## [1] "Anor1 missing values ratio >> 6.83%"
## [1] "Anor2 missing values ratio >> 11.51%"
## [1] "Anor3 missing values ratio >> 17.63%"
## [1] "Anor4 missing values ratio >> 23.38%"
## [1] "Anor5 missing values ratio >> 35.97%"
## [1] "Anor6 missing values ratio >> 38.85%"
## [1] "Naus0 missing values ratio >> 7.19%"
## [1] "Naus1 missing values ratio >> 7.19%"
## [1] "Naus2 missing values ratio >> 10.79%"
## [1] "Naus3 missing values ratio >> 17.99%"
## [1] "Naus4 missing values ratio >> 23.74%"
## [1] "Naus5 missing values ratio >> 36.33%"
## [1] "Naus6 missing values ratio >> 39.57%"
## [1] "Azia0 missing values ratio >> 6.83%"
## [1] "Azia1 missing values ratio >> 7.19%"
## [1] "Azia2 missing values ratio >> 11.87%"
## [1] "Azia3 missing values ratio >> 18.35%"
## [1] "Azia4 missing values ratio >> 24.10%"
## [1] "Azia5 missing values ratio >> 35.97%"
## [1] "Azia6 missing values ratio >> 39.57%"
## [1] "Diarr0 missing values ratio >> 6.83%"
## [1] "Diarr1 missing values ratio >> 7.55%"
## [1] "Diarr2 missing values ratio >> 11.51%"
## [1] "Diarr3 missing values ratio >> 18.71%"
## [1] "Diarr4 missing values ratio >> 23.74%"
## [1] "Diarr5 missing values ratio >> 35.97%"
## [1] "Diarr6 missing values ratio >> 39.93%"
## [1] "Vomit0 missing values ratio >> 6.83%"
## [1] "Vomit1 missing values ratio >> 7.19%"
## [1] "Vomit2 missing values ratio >> 11.87%"
## [1] "Vomit3 missing values ratio >> 18.35%"
## [1] "Vomit4 missing values ratio >> 23.74%"
## [1] "Vomit5 missing values ratio >> 35.97%"
## [1] "Vomit6 missing values ratio >> 39.57%"
## [1] "Const0 missing values ratio >> 6.47%"
## [1] "Const1 missing values ratio >> 6.83%"
```

```
## [1] "Const2 missing values ratio >> 11.51%"
## [1] "Const3 missing values ratio >> 18.35%"
## [1] "Const4 missing values ratio >> 23.74%"
## [1] "Const5 missing values ratio >> 35.97%"
## [1] "Const6 missing values ratio >> 39.57%"
## [1] "Dorabd0 missing values ratio >> 6.83%"
## [1] "Dorabd1 missing values ratio >> 7.55%"
## [1] "Dorabd2 missing values ratio >> 12.23%"
## [1] "Dorabd3 missing values ratio >> 18.35%"
## [1] "Dorabd4 missing values ratio >> 23.74%"
## [1] "Dorabd5 missing values ratio >> 36.33%"
## [1] "Dorabd6 missing values ratio >> 39.57%"
## [1] "Alop0 missing values ratio >> 7.19%"
## [1] "Alop1 missing values ratio >> 6.83%"
## [1] "Alop2 missing values ratio >> 12.23%"
## [1] "Alop3 missing values ratio >> 18.35%"
## [1] "Alop4 missing values ratio >> 23.74%"
## [1] "Alop5 missing values ratio >> 35.97%"
## [1] "Alop6 missing values ratio >> 39.57%"
## [1] "Hiperpig0 missing values ratio >> 6.83%"
## [1] "Hiperpig1 missing values ratio >> 7.91%"
## [1] "Hiperpig2 missing values ratio >> 12.59%"
## [1] "Hiperpig3 missing values ratio >> 18.71%"
## [1] "Hiperpig4 missing values ratio >> 24.10%"
## [1] "Hiperpig5 missing values ratio >> 36.69%"
## [1] "Hiperpig6 missing values ratio >> 39.93%"
## [1] "Altunhas0 missing values ratio >> 6.83%"
## [1] "Altunhas1 missing values ratio >> 7.91%"
## [1] "Altunhas2 missing values ratio >> 11.87%"
## [1] "Altunhas3 missing values ratio >> 19.06%"
## [1] "Altunhas4 missing values ratio >> 24.46%"
## [1] "Altunhas5 missing values ratio >> 37.05%"
## [1] "Altunhas6 missing values ratio >> 40.29%"
## [1] "Eritmulti0 missing values ratio >> 7.19%"
## [1] "Eritmulti1 missing values ratio >> 7.55%"
## [1] "Eritmulti2 missing values ratio >> 12.23%"
## [1] "Eritmulti3 missing values ratio >> 18.71%"
## [1] "Eritmulti4 missing values ratio >> 24.46%"
## [1] "Eritmulti5 missing values ratio >> 37.05%"
## [1] "Eritmulti6 missing values ratio >> 40.29%"
## [1] "Prur0 missing values ratio >> 6.83%"
## [1] "Prur1 missing values ratio >> 7.19%"
## [1] "Prur2 missing values ratio >> 11.87%"
## [1] "Prur3 missing values ratio >> 18.35%"
## [1] "Prur4 missing values ratio >> 23.74%"
## [1] "Prur5 missing values ratio >> 36.33%"
## [1] "Prur6 missing values ratio >> 39.93%"
## [1] "Reamaospes0 missing values ratio >> 6.83%"
## [1] "Reamaospes1 missing values ratio >> 7.19%"
## [1] "Reamaospes2 missing values ratio >> 11.87%"
## [1] "Reamaospes3 missing values ratio >> 18.71%"
## [1] "Reamaospes4 missing values ratio >> 23.74%"
## [1] "Reamaospes5 missing values ratio >> 36.33%"
## [1] "Reamaospes6 missing values ratio >> 39.93%"
```

```
## [1] "Neuroperif0 missing values ratio >> 7.19%"
## [1] "Neuroperif1 missing values ratio >> 7.19%"
## [1] "Neuroperif2 missing values ratio >> 11.87%"
## [1] "Neuroperif3 missing values ratio >> 18.71%"
## [1] "Neuroperif4 missing values ratio >> 23.74%"
## [1] "Neuroperif5 missing values ratio >> 36.33%"
## [1] "Neuroperif6 missing values ratio >> 39.93%"
## [1] "Perdmemor0 missing values ratio >> 6.12%"
## [1] "Perdmemor1 missing values ratio >> 6.83%"
## [1] "Perdmemor2 missing values ratio >> 11.87%"
## [1] "Perdmemor3 missing values ratio >> 19.06%"
## [1] "Perdmemor4 missing values ratio >> 24.46%"
## [1] "Perdmemor5 missing values ratio >> 36.69%"
## [1] "Perdmemor6 missing values ratio >> 39.93%"
## [1] "Confusao0 missing values ratio >> 6.12%"
## [1] "Confusao1 missing values ratio >> 6.47%"
## [1] "Confusao2 missing values ratio >> 12.95%"
## [1] "Confusao3 missing values ratio >> 18.71%"
## [1] "Confusao4 missing values ratio >> 24.46%"
## [1] "Confusao5 missing values ratio >> 36.69%"
## [1] "Confusao6 missing values ratio >> 40.29%"
## [1] "Perdaudicao0 missing values ratio >> 6.83%"
## [1] "Perdaudicao1 missing values ratio >> 7.19%"
## [1] "Perdaudicao2 missing values ratio >> 12.59%"
## [1] "Perdaudicao3 missing values ratio >> 18.71%"
## [1] "Perdaudicao4 missing values ratio >> 23.74%"
## [1] "Perdaudicao5 missing values ratio >> 36.33%"
## [1] "Perdaudicao6 missing values ratio >> 39.93%"
## [1] "Zumb0 missing values ratio >> 6.12%"
## [1] "Zumb1 missing values ratio >> 7.19%"
## [1] "Zumb2 missing values ratio >> 12.59%"
## [1] "Zumb3 missing values ratio >> 18.35%"
## [1] "Zumb4 missing values ratio >> 24.10%"
## [1] "Zumb5 missing values ratio >> 36.33%"
## [1] "Zumb6 missing values ratio >> 39.93%"
## [1] "Disturbmenst0 missing values ratio >> 69.42%"
## [1] "Disturbmenst1 missing values ratio >> 69.78%"
## [1] "Disturbmenst2 missing values ratio >> 71.94%"
## [1] "Disturbmenst3 missing values ratio >> 74.82%"
## [1] "Disturbmenst4 missing values ratio >> 74.10%"
## [1] "Disturbmenst5 missing values ratio >> 79.86%"
## [1] "Disturbmenst6 missing values ratio >> 81.65%"
## [1] "Cistite0 missing values ratio >> 7.19%"
## [1] "Cistite1 missing values ratio >> 7.91%"
## [1] "Cistite2 missing values ratio >> 12.59%"
## [1] "Cistite3 missing values ratio >> 19.78%"
## [1] "Cistite4 missing values ratio >> 24.10%"
## [1] "Cistite5 missing values ratio >> 36.33%"
## [1] "Cistite6 missing values ratio >> 39.93%"
## [1] "Febre0 missing values ratio >> 7.91%"
## [1] "Febre1 missing values ratio >> 9.35%"
## [1] "Febre2 missing values ratio >> 14.39%"
## [1] "Febre3 missing values ratio >> 20.14%"
## [1] "Febre4 missing values ratio >> 23.74%"
```

```
## [1] "Febre5 missing values ratio >> 36.33%"
## [1] "Febre6 missing values ratio >> 40.65%"
## [1] "Artral0 missing values ratio >> 24.46%"
## [1] "Artral1 missing values ratio >> 23.74%"
## [1] "Artral2 missing values ratio >> 26.98%"
## [1] "Artral3 missing values ratio >> 33.81%"
## [1] "Artral4 missing values ratio >> 36.33%"
## [1] "Artral5 missing values ratio >> 47.12%"
## [1] "Artral6 missing values ratio >> 51.08%"
## [1] "Mialg0 missing values ratio >> 24.82%"
## [1] "Mialg1 missing values ratio >> 23.38%"
## [1] "Mialg2 missing values ratio >> 26.26%"
## [1] "Mialg3 missing values ratio >> 33.09%"
## [1] "Mialg4 missing values ratio >> 35.97%"
## [1] "Mialg5 missing values ratio >> 47.48%"
## [1] "Mialg6 missing values ratio >> 51.08%"
## [1] "Muco0 missing values ratio >> 24.82%"
## [1] "Muco1 missing values ratio >> 22.66%"
## [1] "Muco2 missing values ratio >> 26.98%"
## [1] "Muco3 missing values ratio >> 33.09%"
## [1] "Muco4 missing values ratio >> 35.61%"
## [1] "Muco5 missing values ratio >> 47.12%"
## [1] "Muco6 missing values ratio >> 50.72%"
## [1] "Arrepio0 missing values ratio >> 25.90%"
## [1] "Arrepio1 missing values ratio >> 26.26%"
## [1] "Arrepio2 missing values ratio >> 28.06%"
## [1] "Arrepio3 missing values ratio >> 34.17%"
## [1] "Arrepio4 missing values ratio >> 37.41%"
## [1] "Arrepio5 missing values ratio >> 48.56%"
## [1] "Arrepio6 missing values ratio >> 52.52%"
## [1] "Cefaleia0 missing values ratio >> 26.62%"
## [1] "Cefaleia1 missing values ratio >> 26.62%"
## [1] "Cefaleia2 missing values ratio >> 28.06%"
## [1] "Cefaleia3 missing values ratio >> 33.81%"
## [1] "Cefaleia4 missing values ratio >> 37.77%"
## [1] "Cefaleia5 missing values ratio >> 48.92%"
## [1] "Cefaleia6 missing values ratio >> 52.52%"
## [1] "Tontura0 missing values ratio >> 26.98%"
## [1] "Tontura1 missing values ratio >> 26.62%"
## [1] "Tontura2 missing values ratio >> 28.42%"
## [1] "Tontura3 missing values ratio >> 34.17%"
## [1] "Tontura4 missing values ratio >> 37.77%"
## [1] "Tontura5 missing values ratio >> 48.56%"
## [1] "Tontura6 missing values ratio >> 52.52%"
## [1] "Tosse0 missing values ratio >> 26.62%"
## [1] "Tosse1 missing values ratio >> 26.26%"
## [1] "Tosse2 missing values ratio >> 27.70%"
## [1] "Tosse3 missing values ratio >> 33.81%"
## [1] "Tosse4 missing values ratio >> 37.41%"
## [1] "Tosse5 missing values ratio >> 48.20%"
## [1] "Tosse6 missing values ratio >> 51.80%"
```