GEMINI - Data Exploration

Jennifer Yueh

May 25, 2017

```
library(data.table)
library(dplyr)
library(lubridate)
library(ggplot2)
library(Hmisc)
library(psych)
library(GGally)
library(knitr)
```

```
calcLOS <- function(startDate, startTime, endDate, endTime){
  t1 <- paste(startDate, startTime)
  t2 <- paste(endDate, startTime)
  tt.interval <- t1 %--% t2
  time.duration <- as.duration(tt.interval)
  return (as.numeric(time.duration, "hours"))
}


getTestData <- function(fileName){
  filePath <- paste0("Data/Clinical/",fileName)
  target <- read.csv(filePath, header=T, stringsAsFactors=F)
  target <- target %>% select(-c(Site, Test.Item))

  # how many different types of 'Test.ID' in 'albumin' table
  testDF <- data.frame(table(target$Test.ID))
  #colnames(testDF) <- c("Test.ID", "Count")
  head(testDF)


  id_occurCount <- target %>% group_by(EncID.new) %>%
    summarise(ID_Count=n(), Mean=round(mean(Result.Value),1)) %>%
    arrange(desc(ID_Count))

  target_joined <- left_join(target, id_occurCount, by='EncID.new')

  result <- target_joined %>% select(EncID.new, Mean, ID_Count, Admit.Date, Admit.Time, Discharge.Date,
    rename(target.code = Mean) %>%
    select(EncID.new, Admit.Date, Admit.Time, Discharge.Date, Discharge.Time, target.code, ID_Count) %>%
    distinct()

  return (result)
}


doTest <- function(){
  thenames <- list.files("./Data/Clinical")
  thenames <- thenames[2:16]   # BE VERY very CAREFUL HERE
```

```
  codes <- c("ALP", "ALT", "AST",
             "CA", "CR", "GLUF", "GLUR",
             "HGB", "VLACT", "MCV", "PLT",
             "k", "sodium", "TNI", "IWBCR"
  )

  for (i in 1:length(thenames)){
    thefile <- thenames[i]
    codeName <- codes[i]
    testCount <- paste0(codeName, "_Count")

    tmp <- getTestData(thefile)

    colnames(tmp)[which(colnames(tmp) == "target.code")] <- codeName
    colnames(tmp)[which(colnames(tmp) == "ID_Count")] <- testCount

    combined <- full_join(combined, tmp, by = c("EncID.new", "Admit.Date", "Admit.Time", "Discharge.Dat
  }

  return (combined)
}
```

load in 'albumin' test data

```
ALB <- read.csv("Data/Clinical/lab.albumin.csv", header=T, stringsAsFactors=F)
ALB <- ALB %>% select(-c(Site, Test.Item)) # 10750    11 (10750 observations)
length(unique(ALB$EncID.new))
```

```
## [1] 10166
```

how many different types of 'Test.ID' in the table

```
testDF <- data.frame(table(ALB$Test.ID))
colnames(testDF) <- c("Test.ID", "Count")
testDF
```

```
##   Test.ID Count
## 1     ALB 10720
## 2   ALBPE    29
## 3    CALB     1
```

number of the 'albumin' tests having been done

```
id_occurCount <- ALB %>% group_by(EncID.new) %>%
  summarise(ALB_Count=n(), Mean=round(mean(Result.Value),1)) %>%
  arrange(desc(ALB_Count))
head(id_occurCount, 10) #show 10 patients only
```

```
## # A tibble: 10 × 3
##    EncID.new ALB_Count  Mean
##        <int>     <int> <dbl>
```

```
## 1   11384748         4  34.8
## 2   11916073         4  20.4
## 3   11986124         4  42.0
## 4   11104078         3  39.0
## 5   11155168         3  38.2
## 6   11171499         3  35.7
## 7   11182572         3  31.7
## 8   11201344         3  34.9
## 9   11206232         3  25.0
## 10  11279141         3  34.0
```

```
ALB_joined <- left_join(ALB, id_occurCount, by='EncID.new')

albumin <- ALB_joined %>% select(EncID.new, Mean, ALB_Count, Admit.Date, Admit.Time, Discharge.Date, Dis
          rename(ALB = Mean) %>%
          select(EncID.new, Admit.Date, Admit.Time, Discharge.Date, Discharge.Time, ALB, ALB_Count) %>%
                   distinct()

combined <- albumin # 10166     7
```

display randomly selected observations in the dataset here

```
testSet <- doTest()
testSet$LOS <- calcLOS(testSet$Admit.Date, testSet$Admit.Time, testSet$Discharge.Date, testSet$Discharg
thesample <- sample_n(testSet,5)

df <- thesample %>% select(1,6,8,10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38)
kable(df)
```

	EncID.new	ALB	ALP	ALT	AST	CA	CR	GLUF	GLUR	HGB	VLACT	MCV	PLT	k	s
13296	11544196	NA	NA	NA	NA	NA	59	NA	14.4	117	NA	88.4	189	4.8	
2237	11305873	43	99	27	20	2.4	76	NA	19.7	130	2.2	92.2	266	3.5	
12122	11325143	NA	NA	NA	NA	NA	129	NA	9.2	96	NA	91.6	109	4.4	
6292	11655014	32	99	17	39	2.2	83	NA	6.4	127	NA	89.6	416	4.3	
2923	11365770	32	126	37	55	NA	55	NA	7.0	123	NA	84.6	361	NA	

```
df_count <- thesample %>% select(1,7,9,11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 38)
kable(df_count)
```

	EncID.new	ALB_Count	ALP_Count	ALT_Count	AST_Count	CA_Count	CR_Count	GLUF_Coun
13296	11544196	NA	NA	NA	NA	NA	1	N
2237	11305873	1	1	1	1	1	1	N
12122	11325143	NA	NA	NA	NA	NA	1	N
6292	11655014	1	1	1	1	1	1	N
2923	11365770	1	1	1	1	NA	1	N

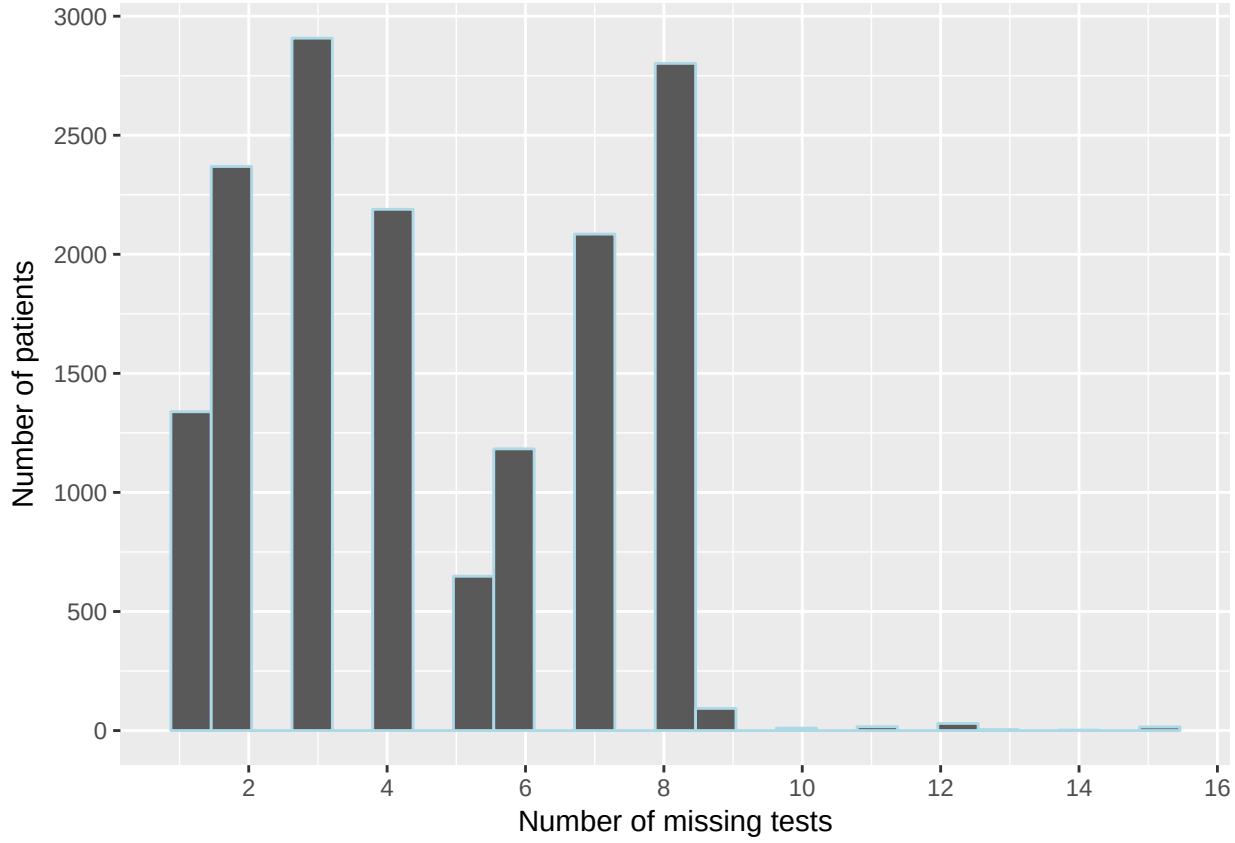
number of missing tests per patient

```
smallSet <- testSet %>% select(EncID.new, ALB, ALP, ALT, AST, CA, CR, GLUF,
                               GLUR, HGB, VLACT, MCV, PLT, k, sodium, TNI, IWBCR, LOS)
```

```
smallSet$NA_count <- rowSums(is.na(smallSet))
kable(head(smallSet))
```

EncID.new	ALB	ALP	ALT	AST	CA	CR	GLUF	GLUR	HGB	VLACT	MCV	PLT	k	sodium
11100037	44	51	101	97	NA	58.0	NA	4.6	147	NA	90.9	290	4.0	133
11100066	42	79	57	56	2.2	84.0	NA	6.8	142	1.4	92.5	163	3.9	135
11100114	42	NA	NA	NA	2.4	65.0	NA	5.3	142	NA	94.2	178	3.6	135
11100155	37	58	19	11	NA	79.0	NA	7.9	92	NA	83.4	316	4.1	138
11100241	35	104	13	28	NA	472.5	NA	6.4	129	1.9	103.0	228	3.8	134
11100368	33	123	12	24	NA	120.0	NA	7.1	111	NA	83.4	280	4.4	127

```
ggplot(data=smallSet, aes(x=NA_count)) + geom_histogram(bins=25, color="lightblue") +
  scale_x_continuous(breaks = seq(0,16,2)) +
    scale_y_continuous(breaks = seq(0,3000,500)) +
    xlab("Number of missing tests") +
    ylab("Number of patients")
```



percentage of patients have missed (have not taken) a particular test

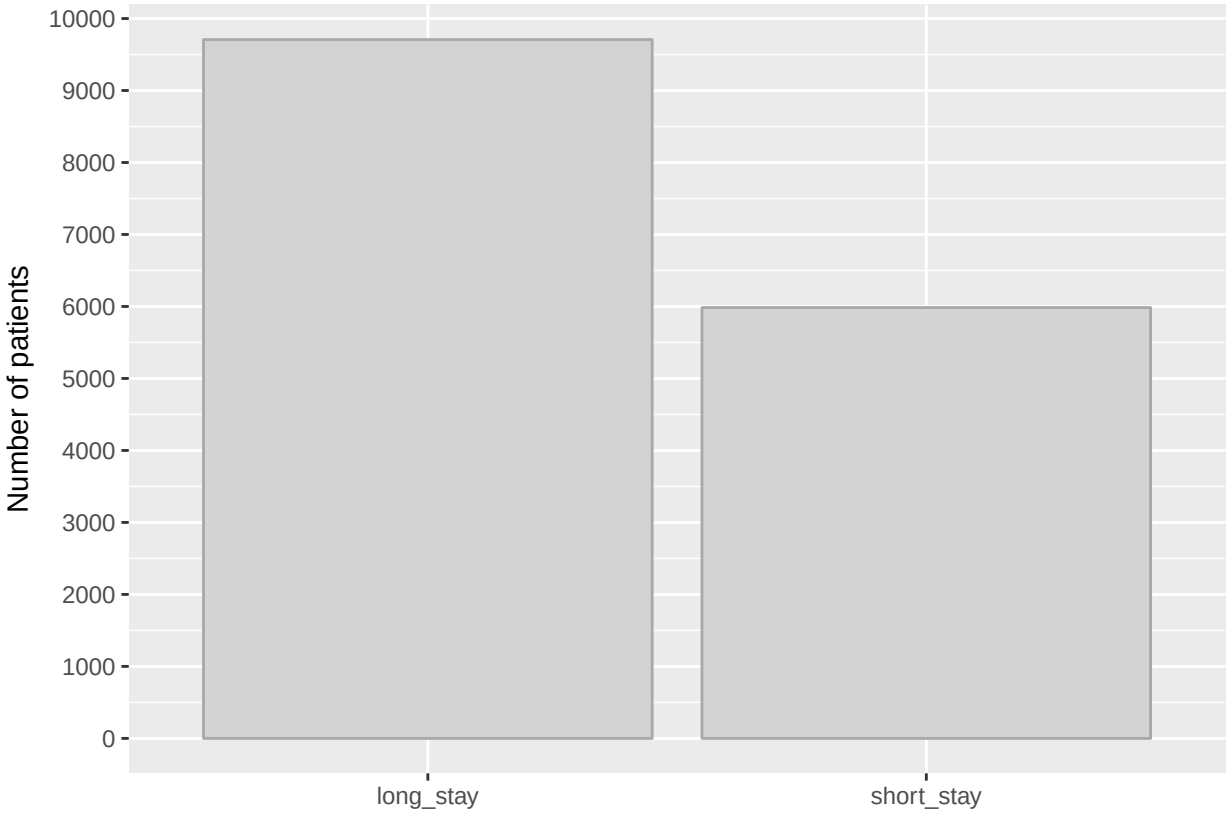
```
smallSetTestOnly <- smallSet %>% select(-c(EncID.new, LOS, NA_count))
NA_count_PerTest <- smallSetTestOnly %>%
          summarise_all(funs(Total = sum(is.na(.))))

md <- melt(NA_count_PerTest)
md$percent <- round((md$value / nrow(smallSetTestOnly)) * 100, 1)
names(md) <- c("missing_Test.ID", "missing_Count", "Percent")
md$"missing_Test.ID" <- gsub("_Total", "", md$"missing_Test.ID")
md %>% arrange(desc(Percent))
```

```
##    missing_Test.ID missing_Count Percent
## 1             GLUF         15684    99.9
## 2            VLACT         11410    72.7
## 3              TNI         10145    64.6
## 4               CA          8221    52.4
## 5              ALT          6873    43.8
## 6              AST          6868    43.8
## 7              ALP          6864    43.7
## 8              ALB          5527    35.2
## 9                k           485     3.1
## 10             PLT           127     0.8
## 11             MCV            93     0.6
## 12              CR            72     0.5
## 13            GLUR            85     0.5
## 14             HGB            86     0.5
## 15           IWBCR            86     0.5
## 16          sodium            58     0.4
```

long_stay vs short_stay

```
smallSet$typeStay <- if_else(smallSet$LOS <= 72, "short_stay", "long_stay")
smallSetTally <- smallSet %>% count(typeStay) %>% rename(Count = n)
ggplot(data=smallSetTally, aes(x=typeStay, y=Count)) +
  geom_bar(stat="identity", fill='lightgrey', colour='darkgrey') +
  scale_y_continuous(breaks = seq(0,10000,1000)) +
  xlab("") + ylab("Number of patients")
```


Number of patients
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
long_stay
short_stay