# DATA 607 Assignment 2

*Samuel I Kigamba*

*September 7, 2019*

**DATA 607: Assignment 2 SQL & R**

**install the RMySQL package**

```
#install.packages("RMySQL")
```

**load library**

```
library(RMySQL)
```

```
## Warning: package 'RMySQL' was built under R version 3.5.3
```

```
## Loading required package: DBI
```

```
## Warning: package 'DBI' was built under R version 3.5.3
```

**Execute block and hide actual(password) - note actual password not displayed**

**Create connection to database in MySQL**

```
mydb = dbConnect(MySQL(), user='root', password=pswd, dbname='movies', host='localhost')
```

**Show list of tables in database**

```
dbListTables(mydb)
```

```
## [1] "movies"
```

**show fields/columns in the database**

```
dbListFields(mydb, 'movies')
```

```
## [1] "nameid"          "Respondent"      "Game_of_Thrones" "The_Well"
## [5] "It_Chapter_Two"  "Ad_Astra"        "Jacobs_Ladder"   "Captain_Marvel"
```

**Display the contents of the table**

```
rs = dbSendQuery(mydb, "select * from movies ORDER BY Respondent")
ratings = fetch(rs)
ratings
```

```
##   nameid Respondent Game_of_Thrones The_Well It_Chapter_Two Ad_Astra
## 1      4    Jacinta               3        5             NA        3
## 2      2    Maureen               5        5              2        4
## 3      5      Naomi               5        5              3        4
## 4      3       Paul               5        4              5        4
## 5      1     Samuel               4        2              4        5
##   Jacobs_Ladder Captain_Marvel
## 1            NA              4
## 2             3              1
## 3             2              5
## 4             5             NA
## 5             4              3
```

## To view and impute missing values

## install MICE

```
#install.packages("mice")
library(mice)
```

```
## Warning: package 'mice' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

## install missForest

```
#install.packages("missForest")
library(missForest)
```

```
## Warning: package 'missForest' was built under R version 3.5.3
```

```
## Loading required package: randomForest
```

```
## Warning: package 'randomForest' was built under R version 3.5.3

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

## Loading required package: foreach

## Warning: package 'foreach' was built under R version 3.5.3

## Loading required package: itertools

## Warning: package 'itertools' was built under R version 3.5.3

## Loading required package: iterators

## Warning: package 'iterators' was built under R version 3.5.3
```

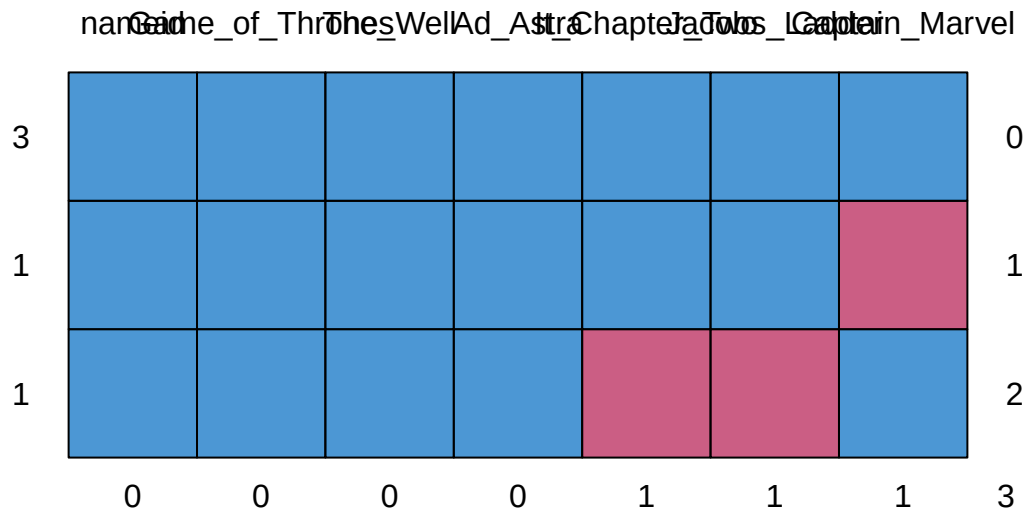
## remove categorical variables

```
ratings_mice <- subset(ratings, select = -c(Respondent))
summary(ratings_mice)
```

```
##      nameid  Game_of_Thrones    The_Well   It_Chapter_Two    Ad_Astra
##  Min.   :1   Min.   :3.0     Min.   :2.0   Min.   :2.00   Min.   :3
##  1st Qu.:2   1st Qu.:4.0     1st Qu.:4.0   1st Qu.:2.75   1st Qu.:4
##  Median :3   Median :5.0     Median :5.0   Median :3.50   Median :4
##  Mean   :3   Mean   :4.4     Mean   :4.2   Mean   :3.50   Mean   :4
##  3rd Qu.:4   3rd Qu.:5.0     3rd Qu.:5.0   3rd Qu.:4.25   3rd Qu.:4
##  Max.   :5   Max.   :5.0     Max.   :5.0   Max.   :5.00   Max.   :5
##                                            NA's   :1
##  Jacobs_Ladder  Captain_Marvel
##  Min.   :2.00   Min.   :1.00
##  1st Qu.:2.75   1st Qu.:2.50
##  Median :3.50   Median :3.50
##  Mean   :3.50   Mean   :3.25
##  3rd Qu.:4.25   3rd Qu.:4.25
##  Max.   :5.00   Max.   :5.00
##  NA's   :1      NA's   :1
```

## use md.pattern() to return a tabular form of missing value present in each variable in a data set.

```
md.pattern(ratings_mice)
```




```
##   nameid Game_of_Thrones The_Well Ad_Astra It_Chapter_Two Jacobs_Ladder
## 3      1               1        1        1              1             1
## 1      1               1        1        1              1             1
## 1      1               1        1        1              0             0
##        0               0        0        0              1             1
##   Captain_Marvel
## 3              1 0
## 1              0 1
## 1              1 2
##                1 3
```

## We can also create a visual which represents missing values.

```
#install.packages("VIM")
library(VIM)
```

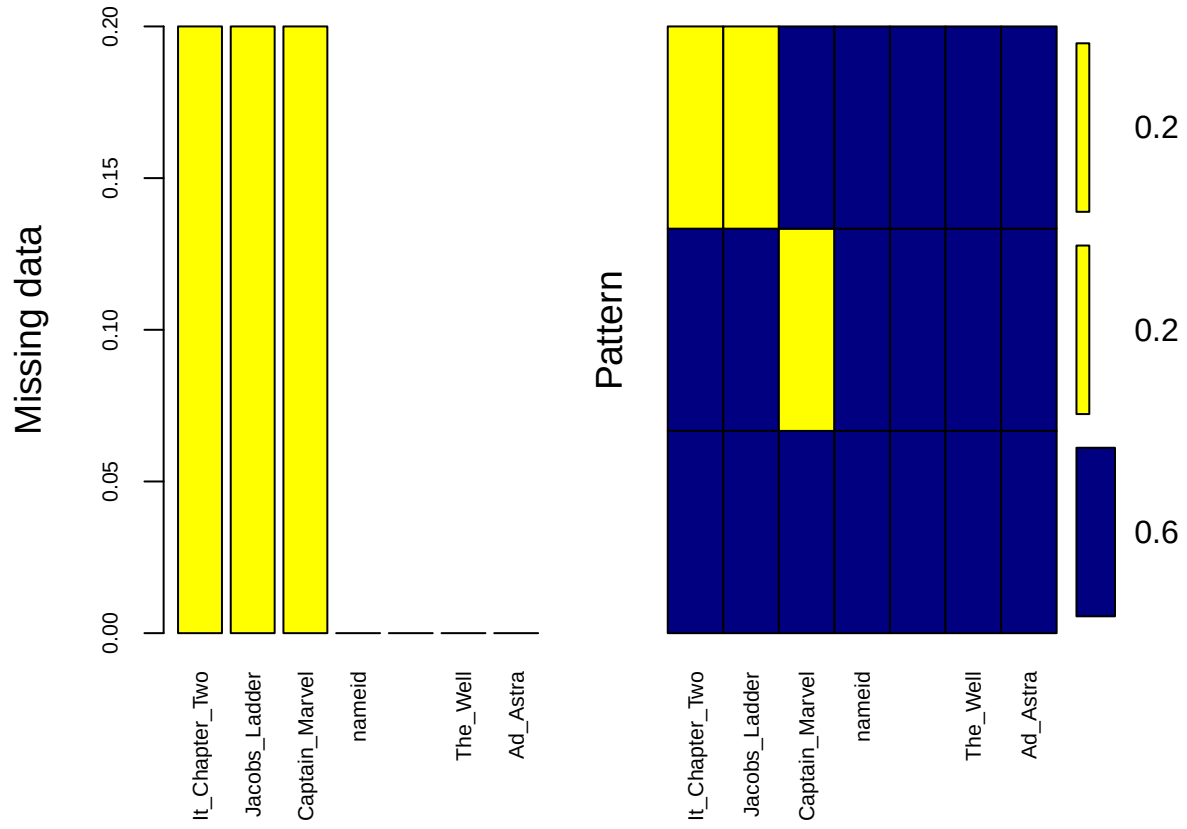
```
## Warning: package 'VIM' was built under R version 3.5.3
```

```
## Loading required package: colorspace
```

```
## Warning: package 'colorspace' was built under R version 3.5.3
```

```
## Loading required package: grid

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 3.5.3

## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```r
mice_plot <- aggr(ratings_mice, col=c('navyblue','yellow'),
                  numbers=TRUE, sortVars=TRUE,
                  labels=names(ratings_mice), cex.axis=.7,
                  gap=3, ylab=c("Missing data","Pattern"))
```



```
##
##  Variables sorted by number of missings:
##         Variable Count
##   It_Chapter_Two   0.2
##    Jacobs_Ladder   0.2
##   Captain_Marvel   0.2
##           nameid   0.0
##  Game_of_Thrones   0.0
##         The_Well   0.0
##         Ad_Astra   0.0
```

## Imputation steps to follow