Migraine Dataset

In December 2016 the company "BM" launched the app "MigraineTracker". This app was developed to collect data from people suffering from migraine. Users are able to create a record in this app after they had a sick headache. So far this app has 4485 users and they created in total 22645 entries in the year 2017.

As we are not able to analyze all the 22645 records manually, we are glad about the opportunity to use R to get some information about the data produced in the last year.

!) Load the file "records.csv" into a variable called "records" and show the head of it.

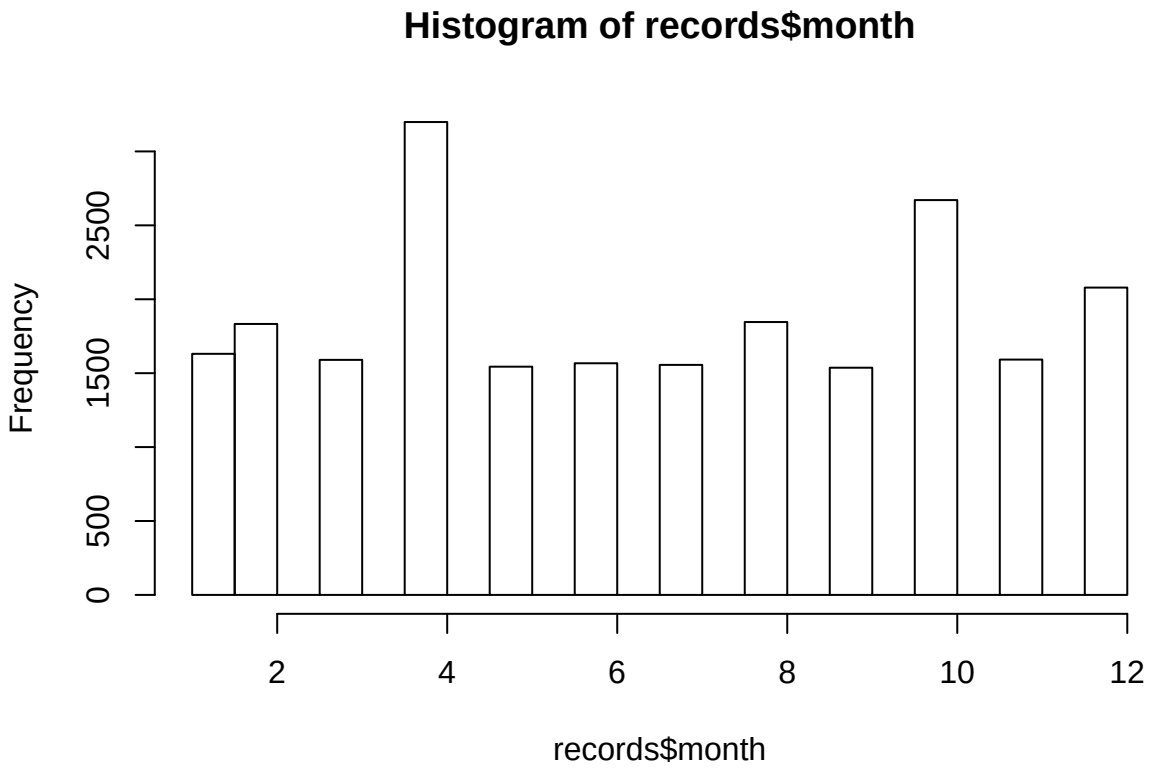
```
##   X user.ID month day beta.blocker pain.level duration visual.disorder
## 1 1     286     1   4          yes          4        3              no
## 2 2    1331     8  19           no          3        2              no
## 3 3    2283     2  15          yes          5        1              no
## 4 4    3365     9  29           no          2        3              no
## 5 5     974     8  13          yes          3        2              no
## 6 6     730     3  19           no          5        5              no
```

Each record consists of the id of the user, that is just an integer. The starting day is defined by month and day, as we only analyze the records from the year 2017. The user has to insert if she or he used beta-blocker. Beta-blockers are pharmaceutical drugs which can be used for the medication of migraine. The pain level is a value between 1 and 5, 1 means light pain and 5 means strong pain. The duration is measured in days and the user has to insert if she or he had visual disorders.

Exercise 1

Lets have a look at the data.

a) Show a histogram of the frequencies of all records for the different months. What are the two month with the most records?



b) Find the max duration that was recorded.

```
## [1] 8
```

If you use the function *nrow()* with the a data frame as parameter, you will get the total number of rows. You can use this function also for subsets.

c) Show the number of records in the dataset defined above.

```
## [1] 22645
```

Exercise 2

a) Now calculate the percentage of records in which the patients had visual disorders. You can first create a data frame named "subset.vd"" that contains the subset of records with visual disorder. Hint: use a logical vector to specify the rows for the new data frame "subset.vd", but keep all columns. Then you can calculate the percentage with the help of *nrow()* and mathematical operators.

b) Also, try to do it in only one line.

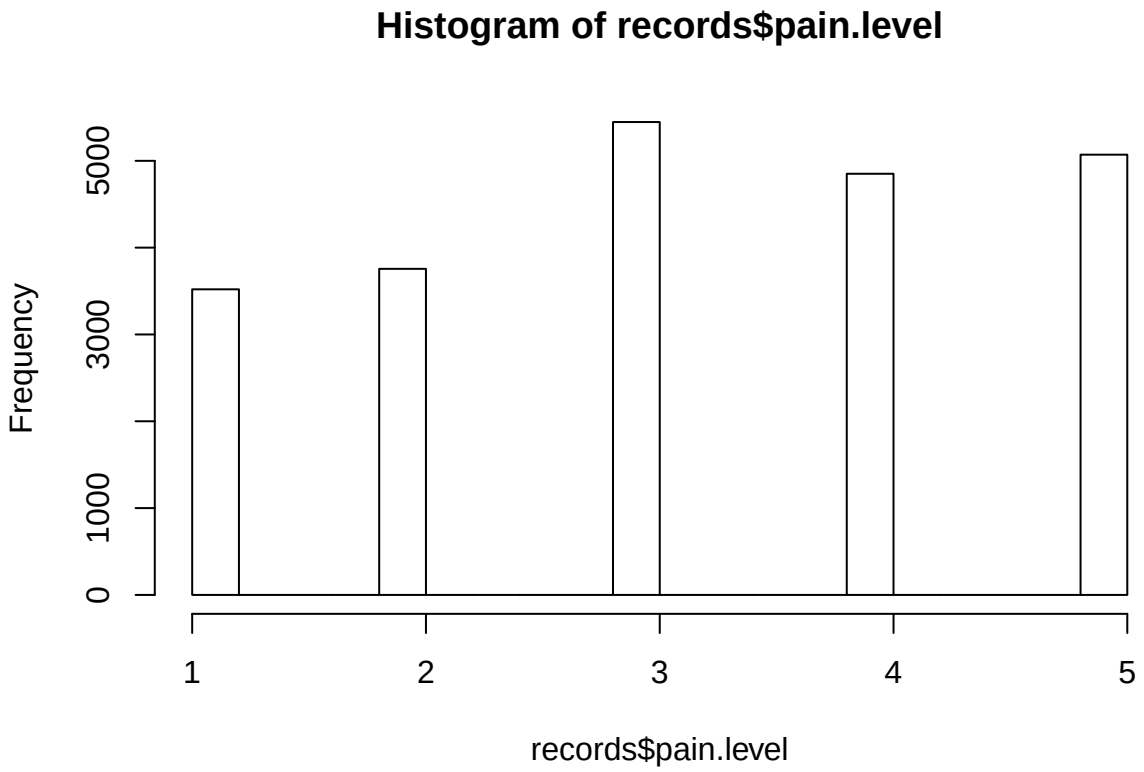
```
## [1] 23.98764
```

```
## [1] 23.98764
```

c) Calculate how often the users took beta-blockers.

```
## [1] 14116
```

d) Plot a histogram for the pain level. Can you see something conspicuous?



Exercies 3

Wouldn't it be interesting to know something about the users?

!) Load the file "users.csv" into variable "users" and show the head of it.

```
##   X user.ID   age    sex parent.suffers
## 1 1       1 20-30   male            yes
## 2 2       2 40-50 female             no
## 3 3       3 20-30 female             no
## 4 4       4 40-50   male            yes
## 5 5       5 30-40 female             no
## 6 6       6 60-70 female             no
```

As you see, in this table we get information about the age and sex of each user. Furthermore, for each user we know if one of her or his parents also suffers from migraine. The user IDs used in this table match with the IDs in the records table. However, let's start with some basic steps to get to know our new list.

a) Get the number of female and male users. (combine *logical vectors* and *nrow()*)

b) Show a simple overview for the ranges of age? (use *summary()*)

c) In this table all columns, but the ID, are categorical. Show a table that includes sex and parent.suffers and shows the number of records for the different combinations of the values for those two columns. (use *table()*)

d) The same as in c), do it also for the pair "sex <-> age"?

```
## [1] 1309
```

```
## [1] 3176
```

```
## 20-30 30-40 40-50 50-60 60-70 70-80 80-90
```

```
##   448  1061   888   897   507   359   325
```

```
##
##           no  yes
##   female 1332 1844
##   male    536  773
```

```
##
##          20-30 30-40 40-50 50-60 60-70 70-80 80-90
##   female   324   744   634   638   354   252   230
##   male     124   317   254   259   153   107    95
```

How many users are in the dataset? Does this number fit to the number mentioned in the initial description at the very beginning of this document?

e) Show the number of rows in users. We already know that there are much more records than users. In average, how many records are created by one user?

f) Calculate the average number of records per user.

```
## [1] 4485
```

```
## [1] 5.049052
```

Exercise 4

Now we got some basic information about the users. But we would like to get some statistics about the information stored in our records table connected to the user information. Create a new table stored in the variable "all.info" that contains all information. Use the *merge()* function. Which column do both tables have in common, you will need this column for this step.

a) Create "all.info" and show the head of it.

```
##   user.ID   X.x month day beta.blocker pain.level duration visual.disorder
## 1       1 22176     8  25          yes          5        6              no
## 2       1 20464    10   3           no          4        5              no
## 3       1  5762     4  26          yes          2        2             yes
## 4       1 20993     7   5          yes          3        1              no
## 5       1 13587    11  14          yes          3        3              no
## 6       1 17140     4   8          yes          1        4              no
##   X.y   age  sex parent.suffers
## 1   1 20-30 male            yes
## 2   1 20-30 male            yes
## 3   1 20-30 male            yes
## 4   1 20-30 male            yes
## 5   1 20-30 male            yes
## 6   1 20-30 male            yes
```

Exercise 5

Let's use the *tapply()* function to get some information.

a) The mean of the duration for the records differing by the usage of meta-blockers.

b) The mean of the duration for the records differing by the information about the parents. Is there a difference between a) and b)?

c) Do the same as in a) and b) but for the pain level instead of the duration. What is your observation?

```
##       no      yes
## 3.897409 2.787475
```

```
##       no      yes
## 3.199367 3.209947
```

```
##       no      yes
## 3.207293 3.172074
```

```
##       no      yes
## 3.058259 3.276765
```

Exercise 6

Now we would like to get some information about the visual disorder.

a) Use *table()* to see if there is a difference between male and female users with respect to the visual disorders.

b) Maybe table is not the best function you can use here. Use in addition *prop.table()*.

With *prop.table()* it is easy to see that the ratios between male and female differ. But we are interested in the percentage of records from male users with visual disorders. And the same for female users.

c) Calculate those percentages. Hint: It often helps to do it step by step. Create new variables. E.g. "only.male" that is a new data frame containing only rows from male users and all columns. Do the same for "only.female". In the following lines you can use "only.male" and "only.female" for the calculations.

```
##
##             no   yes
##   female 11195  4792
##   male    6018   640
```

```
##
##                   no        yes
##   female 0.49436962 0.21161404
##   male   0.26575403 0.02826231
```

```
## [1] 29.97435
```

```
## [1] 9.612496
```

Advanced Exercise 7 (Calculate the score)

So far there are two main values describing the "strength" of a sick headache, the pain level and the duration. We would like to calculate a main score that is a combined value that we call simply **score**. This score is calculated as follows. For both the pain level and the duration we want to compute the percentage based on the maximum pain level or duration, respectively. Afterwards, the score is just the mean of those two percentages. To keep control of what has to be done, let's do it step by step.

a) Create two variables "pl.max" and "d.max" which are assigned to the maximum pain level and maximum duration respectively.

b) Add two columns (pl.perc and d.perc) to all.info which contain the percentage of the pain level and duration. Hint: "pl.perc" and "d.perc" are new vectors calculated with the vectors "pain.level" and "duration" from "all.info" and the variables "pl.max" and "d.max" respectively.

c) Now use the two new columns to create a new column **score** that contains the mean of pl.perc and d.perc. Again, **score** is a new vector that is added to "all.info" as a new column.

d) Now get some information about the score. Calculate the mean score for the different categories of age, beta.blocker, sex or parent.suffers. Do you see a difference?

```
##     20-30     30-40     40-50     50-60     60-70     70-80     80-90
## 0.5185949 0.5208333 0.5215572 0.5184792 0.5207401 0.5090951 0.5152907
```

```
##        no       yes
## 0.5643173 0.4914246
```

```
##    female      male
## 0.5240508 0.5064603
```

```
##        no       yes
## 0.5057863 0.5282982
```

Additional Exercise 8

The user support reported problems with a certain user account. It is about the user with ID **2**.

a) Display all information about the records from user 2. Records are not sorted, but isn't there something strange?

```
##    user.ID   X.x month day beta.blocker pain.level duration
## 7        2  2192     5   5          yes          4        1
## 8        2  9645     5   6          yes          1        4
## 9        2 12020     9  29           no          4        2
## 10       2   973    12   8          yes          2        2
## 11       2 16437     8  26           no          2        7
##    visual.disorder X.y   age    sex parent.suffers pl.perc d.perc  score
## 7              yes   2 40-50 female             no     0.8  0.125 0.4625
## 8              yes   2 40-50 female             no     0.2  0.500 0.3500
## 9               no   2 40-50 female             no     0.8  0.250 0.5250
## 10              no   2 40-50 female             no     0.4  0.250 0.3250
## 11              no   2 40-50 female             no     0.4  0.875 0.6375
```

Final Remark

This dataset was inspired by existing apps and some very basic information about the disease migraine. However, this is an artificial dataset, all values were sampled randomly.