# The Analysis Of Sleep Data



Rensselaer Polytechnic Institute, Tetherless World Constellation, Troy, NY, United States,





### **Abstract**

In today's world there are number of applications used by user which collect and produce data from day to day. This randomly organized data brings the opportunity to analyzing it and use it for the for providing complimentary suggestions and insights. However, mainly the data produced is in the raw format without any structure and field which cannot be directly utilized. Therefore, with an aim to develop useful insights, it is necessary to clean and process the data in an useful manner.

This poster mainly focusses on data processing, cleaning and providing some useful insights on the Sleep Data collected from Sleep Cycle iOS App. The data require a lot of preliminary data cleaning and processing it to other form for better utilization in terms of model building and analysis.

### Motivation

This dataset is collected by Sleep Cycle iOS App and consist of a number of fields requiring transformation and cleaning. The data is collected day to day as logged by the user, the data cleaning step can be used by other such App produced dataset. The main aim is to produce some useful insights which can be used to interpret how sleep quality may be affect by different factors. Thus providing user with information which they can use to further make some effective decision about their sleeping habits.

## **Data Description**

Dataset consists of 887 rows and 8 variables. Start and End is the date and time the activity was locked in. Sleep quality is the quality percentage inserted by the user, time in bed gives the number of hours dedicated for sleeping. Wake up is a emoticon logged by the user. Sleep notes are a list of notes which consists of Drank Tea, Drank coffee, Worked out and Stressful day. Heart rate is the heart rate logged by the user and Activity steps is the number of steps covered in that day.



### **Data Cleaning**

# Some steps taken for data processing

move the percentage from the sleep quality c(data\$Sleep.quality) n <= υ(uata>sleep.quality)
data\$Sleep.quality = as.numeric(sub("%","",x))
write.csv(data, "sleepdata-preprocess.csv")

### Converting time in bed to numeric

# converting the time in bed to numeric value
for (i in 1: length(data5Time.in.bed)){
 x = data5Time.in.bed[i]
 data5Time.in.bed[i]
 data5me.time[i]=as.numeric(sub(":",".",x))

### Transforming data and time

dataSdate\_start <- format(as.POSIXct(strptime(dataSStart,"%Y-%m-%d %H:%M",tz=""))
.format = "%Y-%m-%d")



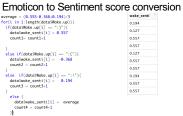






## **Data Cleaning**

hours_start	date_start	hours_end =	date_end
22:42:49	12/30/14	7:13:31	12/31/14
22:31:01	12/31/14	6:03:01	1/1/15
22:12:10	1/1/15	4:56:35	1/2/15
0:34:57	1/3/15	7:47:23	1/3/15
0:23:06	1/4/15	7:37:09	1/4/15
21:34:44	1/4/15	4:53:34	1/5/15



## for(i in 1:length(pre\$Sleep.Notes)){



## Creating Class for Sleep Quality

# creating class for sleep quality in order to apply classificate
extra\_variable <- as.integer(as.character(dataSleep.quality))
for (in in!empth(extra\_vairable )) {
 if(extra\_vairable [i] >= 0 & extra\_vairable [i] <=45){
 dataSsa\_class[i] = "low\_quality"
} else if(extra\_vairable [i] >=65 & extra\_vairable [i] <=65){
 dataSsa\_class[i] = "low\_dum\_quality"
} else { dataSsa\_class[i] = "minim\_quality"
}

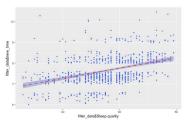


Figure1: Plot between sleep quality and the time in bed

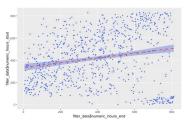


Figure2: Plot between hour start and Hour end of sleep

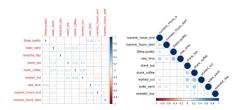


Figure3: Heat map giving idea about the Correlation between numeric factors

## Insights

data<-data[which(data\$Sleep.quality >25),]



Figure 4: Boxplot Before and after filtering Sleep quality

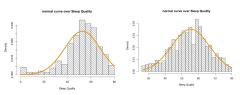


Figure 5: Frequency distribution before and after filtering Heart rate

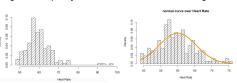
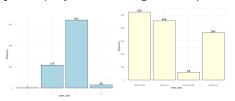
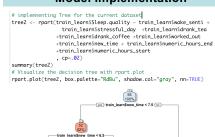


Figure 6: Frequency distribution of Wake\_senti and Sleep notes



# Model Implementation



## Conclusion

The response variable for this dataset is sleep quality. I tried tree models and classification model over the dataset. Since, the size of dataset was small with lots of missing values, the prediction rates turns out not significant for general purpose. Though, data analysis shows that that there is a relationship between Sleep quality and the time a person goes into bed and wakes up. Also there is a relationship between the wake up emoticon which is converted into sentiment score and the sleeping

Excel – For primary Data analysis

R - A program to process data and perform statistical analysis

Package (P) or Library (R): software package to be loaded to perform extra tasks Df. dataframe - Data manipulation structure in R & python pandas

Levels in R dataframe - for factor data, the possible number of choices are levels

For transforming emoticons to sentimental score: #http://kt.ijs.si/data/Emoji\_sentiment\_ranking/ Ggpl plots in R: http://www.sihda.com/english/wiki/doplot2-barrolots-aniick-start-aniide-r-sentimene-

Read excel files in R: https://www.datacamp.com/community/tutorials/r-tutorial-read-excel-into-r R deal with missing data: https://www.statmethods.net/input/missingdata.html R visualization: https://www.analyticsvidhya.com/blog/2015/07/quide-data-visualization-r/