Assignment 6

Vrishti Jain
661983527
6000 Level

1. Abstract and Introduction (2%) Describe your motivation, initial hypothesis/ idea that you wanted to investigate, and if applicable any prior work, interest in the topic (like an intro for a paper, with references).

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 a useful manner.
Why is sleep important:

- It's important to have a healthy lifestyle.
- Sleep is when your body recovers from the day.
- It can affect your health both physical and mental.

The thing which got me interested in this dataset is that I don't get enough sleep and I used to use fit band which used to collect data related to sleep, exercise etc. With that in mind, I want to explore if I can do data analysis on similar dataset and maybe in future implement something similar with personal dataset that I can collect using similar apps.
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.

---

2. Data Description (3%) 1 NOTE: 6000-level students must develop at least two different types of models, not just change the number of variables for a given model type. Describe how you determined which datasets you used in this project, the criteria, source, data and information- types in detail, associated documentation and any other supporting materials. Min. 1/2 page text (+graphics if applicable).

The dataset is present on the Kaggle (https://www.kaggle.com/danagerous/sleep-data) which is personal sleep data collected by Sleep Cycle iOS app by Dana Diotte. The data was acquired between 2014-2018. It consists of 8 attributes.
Start – start date and time
End – end date and time
Sleep quality – sleep quality percentage entered by the user
Time in bed – total time in bed in the format 5:90
Wake up – It's an emoticon that user has entered
Sleep Notes – It's a list which consists of notes for the day like drank tea, drank coffee etc
Heart rate – It's the measured heart rate for that duration
Activity (steps) – Steps covered during the day.
The dataset was really unstructured and requires a lot of cleaning and transformation. The sleep quality is the target variable for the dataset. The dataset is small in size and contains 887 rows.

sleepdata

Start;End;Sleep quality;Time	in	bed;Wake up;Sleep	Notes;Heart rate;Activity	(steps)
2014-12-29 22:57:49;2014-12-3	0	07:30:13;100%;8:32;	:);;59;0	
2014-12-30 21:17:50;2014-12-3	0	21:33:54;3%;0:16;;|	;Stressful day;72;0	
2014-12-30 22:42:49;2014-12-3	1	07:13:31;98%;8:30;;	|;;57;0	
2014-12-31 22:31:01;2015-01-0	1	06:03:01;65%;7:32;;	;;0	

The initial downloaded version of data looks as shown in the picture. It is required to do cleaning and transformation.
Sleep notes looks like: drank tea: stressful day, therefore perform one hot encoding for this field and divide them into numerous columns.
Since the sleep quality is a continuous attribute, the initial step to have it as a target variable is to apply linear model and then try some tree-based models.

---

3. Analysis (5%) Explore the statistical aspects of your datasets. Perform any transformations, interpolations, smoothing, cleaning, etc. required on the data, to begin to explore your hypothesis/ questions. Analyze the distributions; provide summaries of the relevant statistics and plots of any fits you made. Discuss and specify or estimate possible sources of error, uncertainty or bias in the data you used (or did not use). Min. 2 pages text + graphics.

   The initial summary of dataset.

```
> summary(final_data)
                Start                         End      Sleep.quality  Time.in.bed  Wake.up
 2014-12-29 22:57:49:  1   2014-12-30 07:30:13:  1   81%    : 42   8:01   : 18     :641
 2014-12-30 21:17:50:  1   2014-12-30 21:33:54:  1   83%    : 39   7:32   : 16   :(:  1
 2014-12-30 22:42:49:  1   2014-12-31 07:13:31:  1   79%    : 37   8:00   : 15   :):216
 2014-12-31 22:31:01:  1   2015-01-01 06:03:01:  1   77%    : 36   8:02   : 15   :|: 29
 2015-01-01 22:12:10:  1   2015-01-02 04:56:35:  1   74%    : 33   7:35   : 14
 2015-01-03 00:34:57:  1   2015-01-03 07:47:23:  1   75%    : 31   7:29   : 13
 (Other)            :881   (Other)            :881   (Other):669   (Other):796
                          Sleep.Notes    Heart.rate    Activity..steps.
                                :235   Min.   :49.0   Min.   :    0
 Drank coffee:Drank tea:Worked out:164   1st Qu.:57.0   1st Qu.:    0
 Drank coffee:Drank tea           :123   Median :60.0   Median :  255
 Drank coffee:Worked out          : 91   Mean   :60.6   Mean   : 2776
 Drank coffee                     : 75   3rd Qu.:64.0   3rd Qu.: 5317
 Drank tea:Worked out             : 62   Max.   :98.0   Max.   :21870
 (Other)                          :137   NA's   :725
```

```
> str(final_data)
'data.frame':   887 obs. of  8 variables:
 $ Start           : Factor w/ 887 levels "2014-12-29 22:57:49",..: 1 2 3 4 5 6 7 8 9
 $ End             : Factor w/ 887 levels "2014-12-30 07:30:13",..: 1 2 3 4 5 6 7 8 9
 $ Sleep.quality   : Factor w/ 79 levels "0%","10%","100%",..: 3 12 78 42 50 62 56 56
 $ Time.in.bed     : Factor w/ 229 levels "0:00","0:15",..: 173 3 171 113 66 94 96 100
 ...
 $ Wake.up         : Factor w/ 4 levels "",":(",":)",":|": 3 4 4 1 3 3 1 3 3 4 ...
 $ Sleep.Notes     : Factor w/ 20 levels "","Ate late:Drank coffee",..: 1 19 1 1 8 8 1
 ...
 $ Heart.rate      : int  59 72 57 NA 68 60 NA 57 56 64 ...
 $ Activity..steps.: int  0 0 0 0 0 0 0 0 0 0 ...
```

Using the excel initially, I divided the sleep notes into 4 columns. Since there is no particular order in these notes. And it depends entirely on the user the dataset is inserted, one hot encoding is performed on the dataset

Drank coffee	Drank tea		
Drank coffee	Drank tea		
Drank tea			
Ate late	Drank coffee		
Drank coffee	Drank tea	Worked out	
Drank tea	Worked out		
Drank coffee	Drank tea	Stressful day	

stressful_day	drank_tea	drank_coffee	worked_out
0	0	0	0
1	0	0	0
0	0	0	0
0	0	0	0
0	1	1	0
0	1	1	0
0	1	0	0
0	0	1	0
0	1	1	1
0	1	0	1
1	1	1	0
0	1	1	0
0	1	1	0

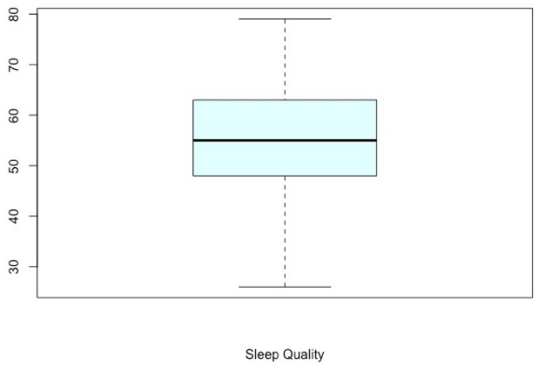
Now the date and time are put into separated column:

Start	End
2014-12-29 22:57:49	2014-12-30 07:30:13
2014-12-30 21:17:50	2014-12-30 21:33:54
2014-12-30 22:42:49	2014-12-31 07:13:31
2014-12-31 22:31:01	2015-01-01 06:03:01

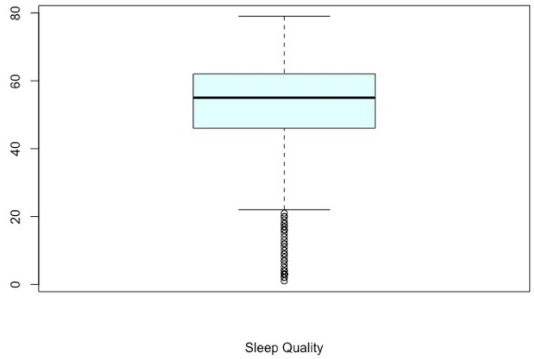
hours_start	date_start	hours_end	date_end
22:57:49	12/29/14	7:30:13	12/30/14
21:17:50	12/30/14	21:33:54	12/30/14
22:42:49	12/30/14	7:13:31	12/31/14
22:31:01	12/31/14	6:03:01	1/1/15
22:12:10	1/1/15	4:56:35	1/2/15
0:34:57	1/3/15	7:47:23	1/3/15
0:23:06	1/4/15	7:37:09	1/4/15

Converted the wake up variable which is a emoticon to sentiment score. Also where there was no emoticon assigned, that was replaced with the average sentiment score.
(reference: #http://kt.ijs.si/data/Emoji_sentiment_ranking/)
Cleaning:

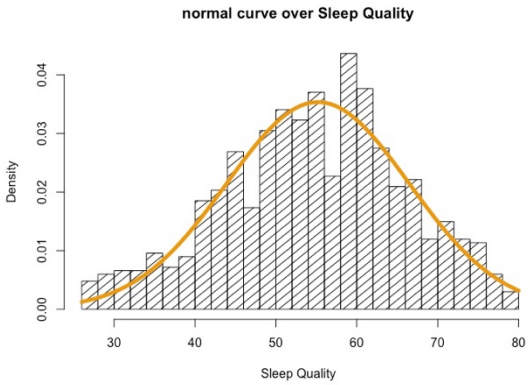
- Removed the percentage from the sleep quality and converted it into percentage
- Converted the time in bed to numeric and changing its form from 7:80 to 7.80.
- One hot encoding of the wake-up notes
- Converted the time to numeric value

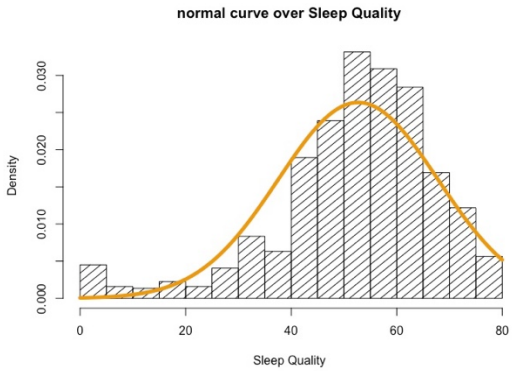


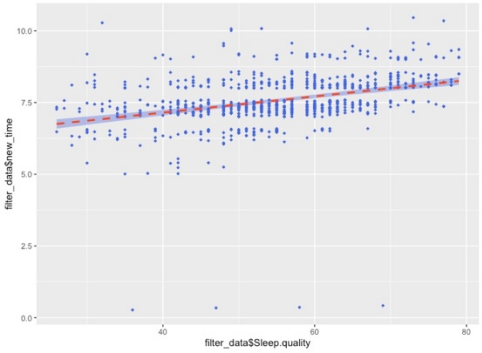
After filtering Sleep quality

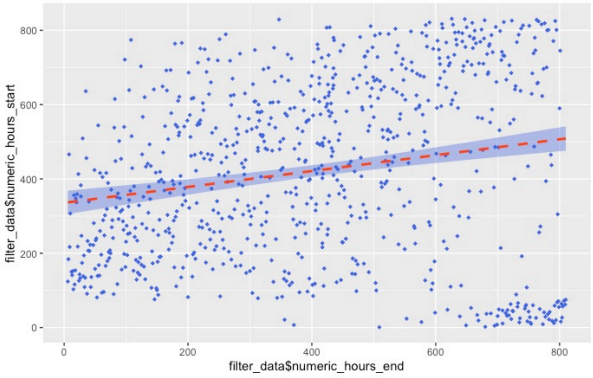


Before filtering Sleep quality

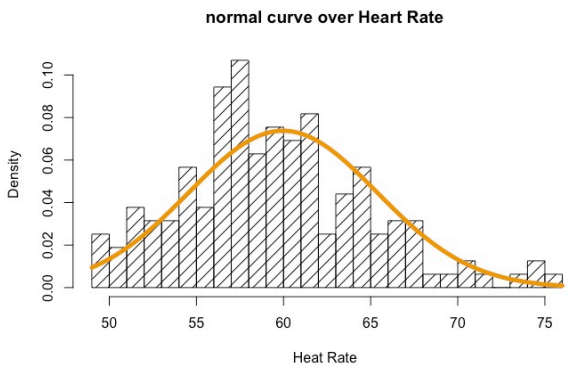


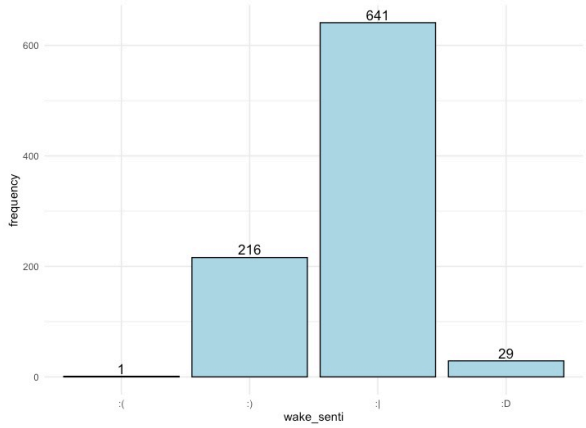


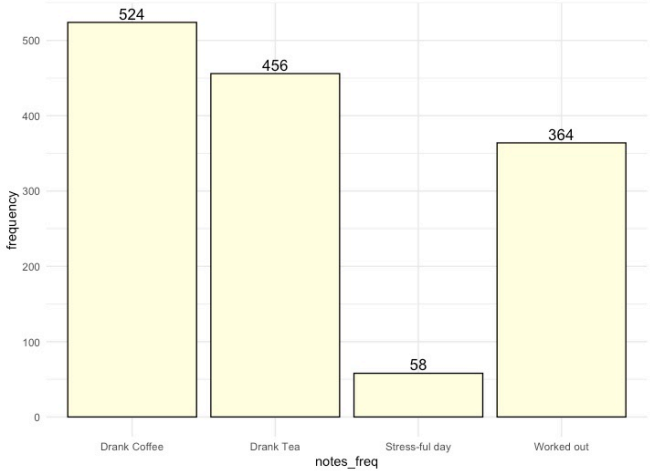




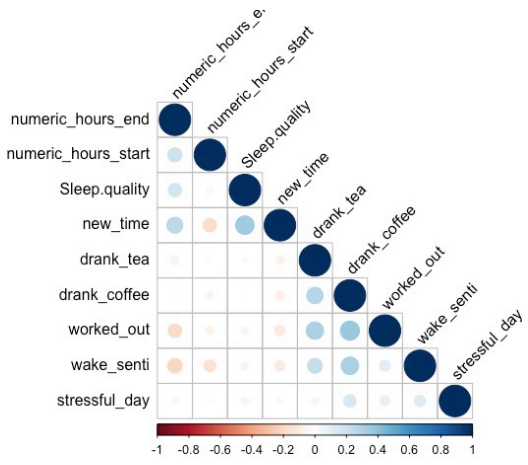
Plot between the sleep quality and time shows positive relationship
Also, the plot between the hour end and hour start, shows a positive relationship though it is a little bit scattered.

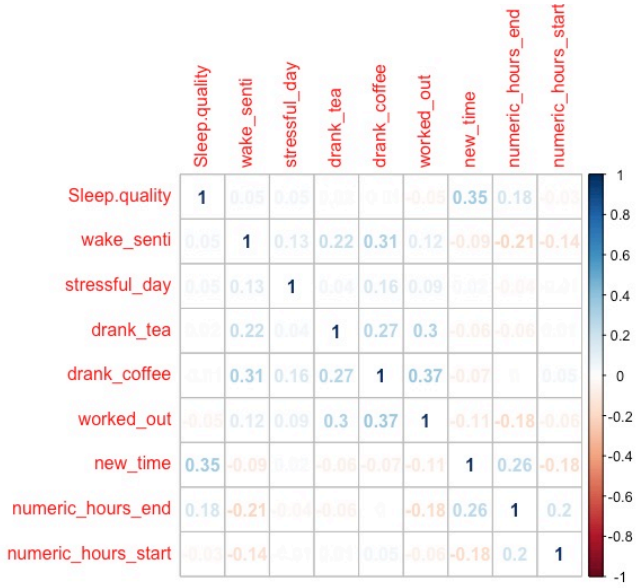






The frequency distribution of the wake senti and the sleep notes





We can see that there is a positive correlation between the new time and the sleep quality and numeric hour end. Also, the slight positive correlation can be seen between the sleep notes.

4. Model Development and Application of model(s) (12%) What types of models you used to describe the data (regression, classification, clustering, etc.), patterns/ trends you found, visual approaches that helped you choose models, and or variables (type/ number) in the model, other parameter choices or settings for the models (e.g. distance metrics, kernels, etc.). Apply the models to assess model performance (i.e. predict). Discuss the confidence in your results including any statistic measures. Discuss how you validated your models and performed any optimization (give details). Min. 6 pages text + graphics.

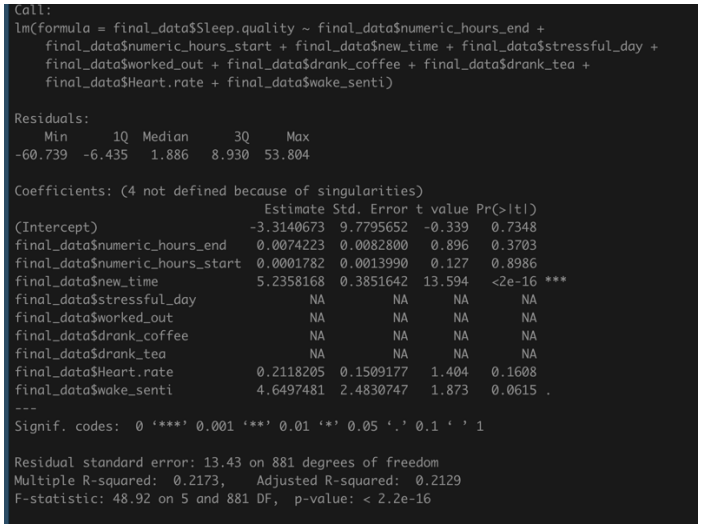
ANSWER

First I tried linear model for the given dataset, but as we can see that the R square is 0.2156 which is not so great when I did not included the heart rate as there were a lot of rows with NA and removing them can results into getting 167 rows which can affect the overall prediction.

```
Call:
lm(formula = final_data$Sleep.quality ~ final_data$numeric_hours_end +
    final_data$numeric_hours_start + final_data$new_time + final_data$stressful_day +
    final_data$worked_out + final_data$drank_coffee + final_data$drank_tea +
    final_data$wake_senti)

Residuals:
    Min      1Q  Median      3Q     Max
-60.112  -6.343   1.879   8.955  53.684

Coefficients: (4 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 9.0804585  4.2042250   2.160   0.0311 *
final_data$numeric_hours_end   0.0089438  0.0082133   1.089   0.2765
final_data$numeric_hours_start 0.0001314  0.0013994   0.094   0.9252
final_data$new_time            5.2222224  0.3852540  13.555   <2e-16 ***
final_data$stressful_day              NA         NA      NA       NA
final_data$worked_out                 NA         NA      NA       NA
final_data$drank_coffee               NA         NA      NA       NA
final_data$drank_tea                  NA         NA      NA       NA
final_data$wake_senti          4.6964586  2.4842165   1.891   0.0590 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.43 on 882 degrees of freedom
Multiple R-squared:  0.2156,    Adjusted R-squared:  0.212
F-statistic: 60.59 on 4 and 882 DF,  p-value: < 2.2e-16
```

```
mean_heart =mean(final_data$Heart.rate , na.rm = TRUE)
mean_heart

for ( i in 1:length(final_data$Heart.rate)){
  if ( is.na(final_data$Heart.rate[i])){
    print(i)
    final_data$Heart.rate[i] = mean_heart   }}
```

Since there were a lot of rows with heart rate missing, assigning heart rate with average value for those cases. And then again trying the linear model. It turns out that the r square improved with .2173 but the change is not significant.

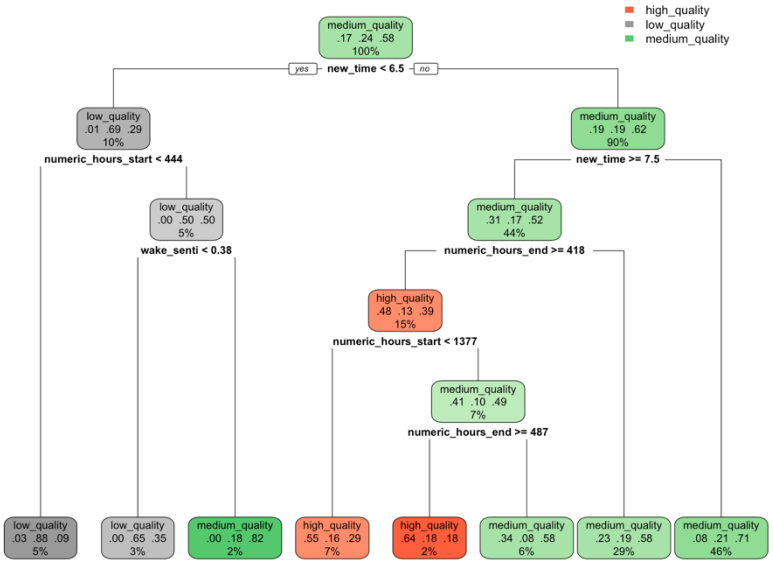


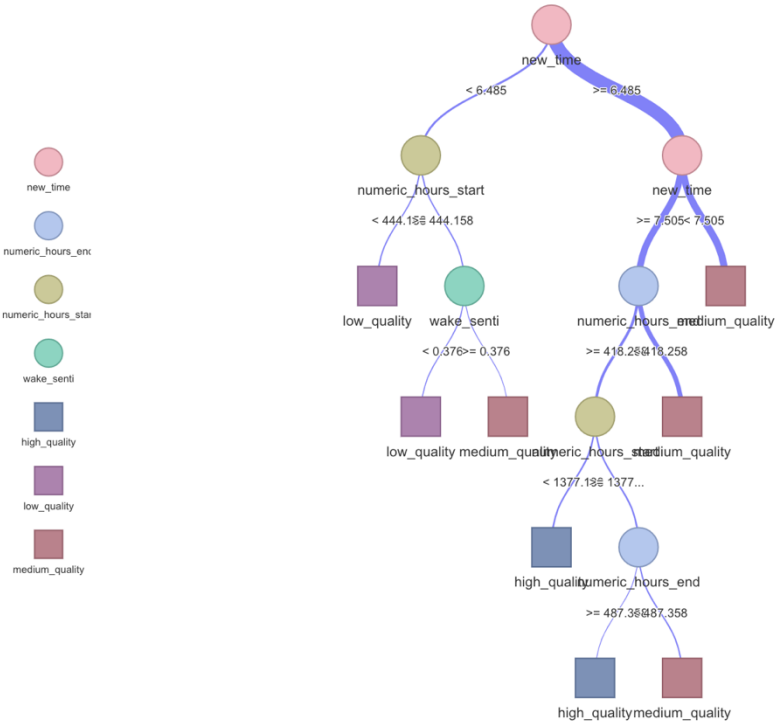
```
Call:
lm(formula = final_data$Sleep.quality ~ final_data$numeric_hours_end +
    final_data$numeric_hours_start + final_data$new_time + final_data$stressful_day +
    final_data$worked_out + final_data$drank_coffee + final_data$drank_tea +
    final_data$Heart.rate + final_data$wake_senti)

Residuals:
    Min      1Q  Median      3Q     Max
-60.739  -6.435   1.886   8.930  53.804

Coefficients: (4 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -3.3140673  9.7795652  -0.339   0.7348
final_data$numeric_hours_end  0.0074223  0.0082800   0.896   0.3703
final_data$numeric_hours_start  0.0001782  0.0013990   0.127   0.8986
final_data$new_time            5.2358168  0.3851642  13.594   <2e-16 ***
final_data$stressful_day              NA         NA      NA       NA
final_data$worked_out                 NA         NA      NA       NA
final_data$drank_coffee               NA         NA      NA       NA
final_data$drank_tea                  NA         NA      NA       NA
final_data$Heart.rate          0.2118205  0.1509177   1.404   0.1608
final_data$wake_senti          4.6497481  2.4830747   1.873   0.0615 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.43 on 881 degrees of freedom
Multiple R-squared:  0.2173,    Adjusted R-squared:  0.2129
F-statistic: 48.92 on 5 and 881 DF,  p-value: < 2.2e-16
```

With Tree based model, when we define the class for the sleep quality, we get accuracy of 59.99

```
extra_vairbale <- as.integer(as.character(subset_sleep$Sleep.quality))
 for (i in 1:length(extra_vairbale ))
 { if(extra_vairbale [i] >= 0 & extra_vairbale [i] <=45)
 { subset_sleep$sq_class[i] = "low_quality" }
   else if(extra_vairbale [i] >45 & extra_vairbale [i] <=65)
   { subset_sleep$sq_class[i] = "medium_quality" } else
      { subset_sleep$sq_class[i] = "high_quality"}}
```

It can also be seen that the new_ time, hour start, hour end and wake_senti contributed to the final decision tree. While the sleep notes cannot be seen in the decision part.


high_quality
low_quality
medium_quality
medium_quality .17 .24 .58 100%
yes
new_time < 6.5
no
low_quality .01 .69 .29 10%
numeric_hours_start < 444
medium_quality .19 .19 .62 90%
new_time >= 7.5
low_quality .00 .50 .50 5%
wake_senti < 0.38
medium_quality .31 .17 .52 44%
numeric_hours_end >= 418
high_quality .48 .13 .39 15%
numeric_hours_start < 1377
medium_quality .41 .10 .49 7%
numeric_hours_end >= 487
low_quality .03 .88 .09 5%
low_quality .00 .65 .35 3%
medium_quality .00 .18 .82 2%
high_quality .55 .16 .29 7%
high_quality .64 .18 .18 2%
medium_quality .34 .08 .58 6%
medium_quality .23 .19 .58 29%
medium_quality .08 .21 .71 46%


new_time
numeric_hours_end
numeric_hours_start
wake_senti
high_quality
low_quality
medium_quality
new_time
< 6.485
>= 6.485
numeric_hours_start
new_time
< 444.158
>= 444.158
>= 7.505
< 7.505
low_quality
wake_senti
numeric_hours_end
medium_quality
< 0.376
>= 0.376
>= 418.258
low_quality
medium_quality
numeric_hours_start
medium_quality
< 1377.1...
>= 1377...
high_quality
numeric_hours_end
>= 487.358
high_quality
medium_quality

```
>  t_pred = predict(tree1,test_valid,type="class")
>
>
>  acc_table<-table(t_pred , test_valid$sq_class)
>  acc_table

t_pred           high_quality low_quality medium_quality
  high_quality             11           1              9
  low_quality               0           8              5
  medium_quality           27          47            114
>
>  accuracy_tree<- sum(diag(acc_table))/sum(acc_table)
>  accuracy_tree
[1] 0.5990991
```

With SVM model, we get a accuracy of 57.2 % with target class of sleep quality becomes class.

```
  NAs introduced by coercion
>  svm.model_sleep <- svm(sq_class ~ ., data = train_learn, type='C-classification', gamma = 0.01)
Warning message:
In svm.default(x, y, scale = scale, ..., na.action = na.action) :
  Variable(s) 'stressful_day' and 'worked_out' and 'drank_coffee' and 'drank_tea' constant. Cannot scale data.
>
>  pred_svm<- predict(svm.model_sleep,test_valid, type ="class")
>  tab_svm<-table(pred_svm , test_valid$sq_class)
>  tab_svm

pred_svm         high_quality low_quality medium_quality
  high_quality              7           2              2
  low_quality               3           2              8
  medium_quality           28          52            118
>  accuracy_svm<- sum(diag(tab_svm))/sum(tab_svm)
>  accuracy_svm
[1] 0.5720721
>   summary(svm.model_sleep)

Call:
svm(formula = sq_class ~ ., data = train_learn, type = "C-classification",
    gamma = 0.01)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  586

 ( 160 311 115 )


Number of Classes:  3

Levels:
 high_quality low_quality medium_quality
```

5. Conclusions and Discussion (3%) Describe your conclusions; interpret the results, predictions you made, the models and their characteristics, and a give summary of what changed as you went through the project (data, analysis, model choices, etc.), what you would do next, or do differently in a subsequent exploration. Min. 1 page text + graphics (optional). References – websites, papers, packages, data refs, etc. should be included at the end. Include your R scripts! (e.g. in a zip file) and also include the Github URL that contains the code. There is no specific citation format, just be consistent.

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 turn out not significant for general purpose. Trying linear model gave insignificant results and maybe rich dataset would have been better with such model.
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 notes.
The accuracy turns out to be 59.99 with tree model which cannot be considered good but average. That also results when I tried to do classification based on 3 different defined class of sleep quality. In the beginning, I thought the data will be rich in the sense that will have values and not the NA's but since the dataset was logged in by the user, it lacks structure and consistency.
The SVM gave 57.2% accuracy which is close to the accuracy of tree based decision tree model.
In the future, I would like to extend the analysis by collecting sleep cycle dataset with my own Fitbit and see the subsequent changes with each passing month.

GIT Link: https://github.com/vrishtijain/Analysis_of_sleep_data

References:
Data source: https://www.kaggle.com/danagerous/sleep-data

1: https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-decision-trees-c24dfd490abb
2: https://stackoverflow.com/questions/18171246/error-in-contrasts-when-defining-a-linear-model-in-r
3: https://stackoverflow.com/questions/40080794/calculating-prediction-accuracy-of-a-tree-using-rparts-predict-method-r-progra
4: http://kt.ijs.si/data/Emoji_sentiment_ranking/