STA490_Winter_EDA

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
```

Data Cleaning

```
data = read.csv("new_data.csv")
data$Exercise2[287] = 0
```

Goals of EDA

I want to investigate whether or not exercises have a positive effect on the mental health of subjects/students. In this analysis, I want to assess the relationship between the overall health state and minutes of exercise of the subject.I have created a mental health indicator using question 15-19 of the questionnaire, which assess the subject's concentration, energy level, feeling and sleep quality. On a scale of 1-7, the higher the indicator the subjects are at, the better mental state they are at.

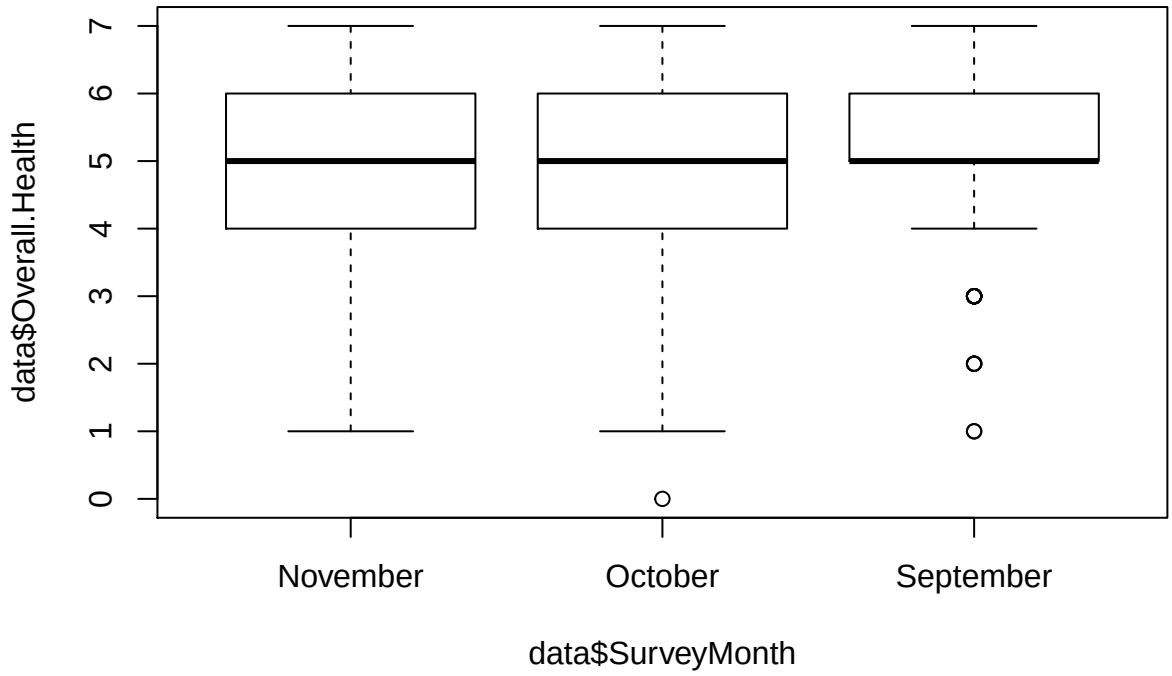
Other than the minutes of exercise, I also want to investigate the effects of activeness, stressors and hours of sleep on the subjects' mental health.

Visualizing distributions

1. The mental health indicator of the subjects seems to be following a left-skewed distribution, and most of the samples fall under the range of 4-6.
2. I want to see if there is a strong relationshio between the mental health indicator and the month the survey is taken at. From the boxplot, the means of each month are very similar. However, looking at the range of the boxplot, the mental health indicators in September are relatively more left-skewed than the other two months.
3. The minutes of exercise of the subjects seems to be right-skewed distributed, with the mean at approximately 200 minutes. There also exists five outliers beyond 600 minutes. After further investigation, it is found that all of these outliers come from 2 subjects (190114,190206).

```
# Response Variable: Overall Health
data1 <- data %>%
  group_by(Overall.Health,SurveyMonth) %>%
```

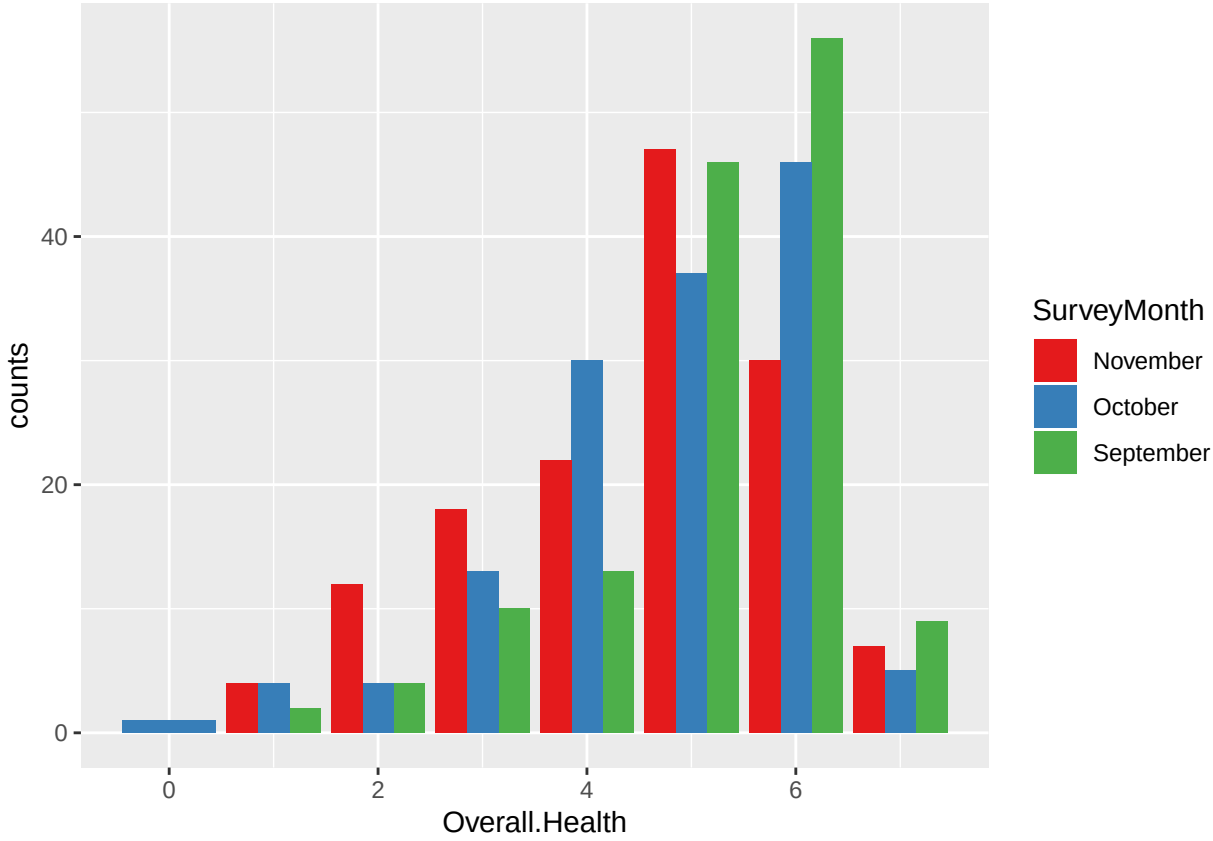
```
  summarise(counts = n())

boxplot(data$Overall.Health~data$SurveyMonth)
```




```
ggplot(data1, aes(x=Overall.Health, y= counts,fill = SurveyMonth)) +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1")
```

Does an exercise habit changes your mental health? (Logistic model)
- Generalized linear mixed model with Gamma or Beta distribution
If exercise 2 = 150 or exercise 3 = 75;
- Define treatment and control groups; Using the WHO report
- Use Study ID and Survey Month (?) as random effects
- Confounders: Hours of sleep and number of stressors
- Treatment effect should take in account of CI

Results: More exercise = Better mental health



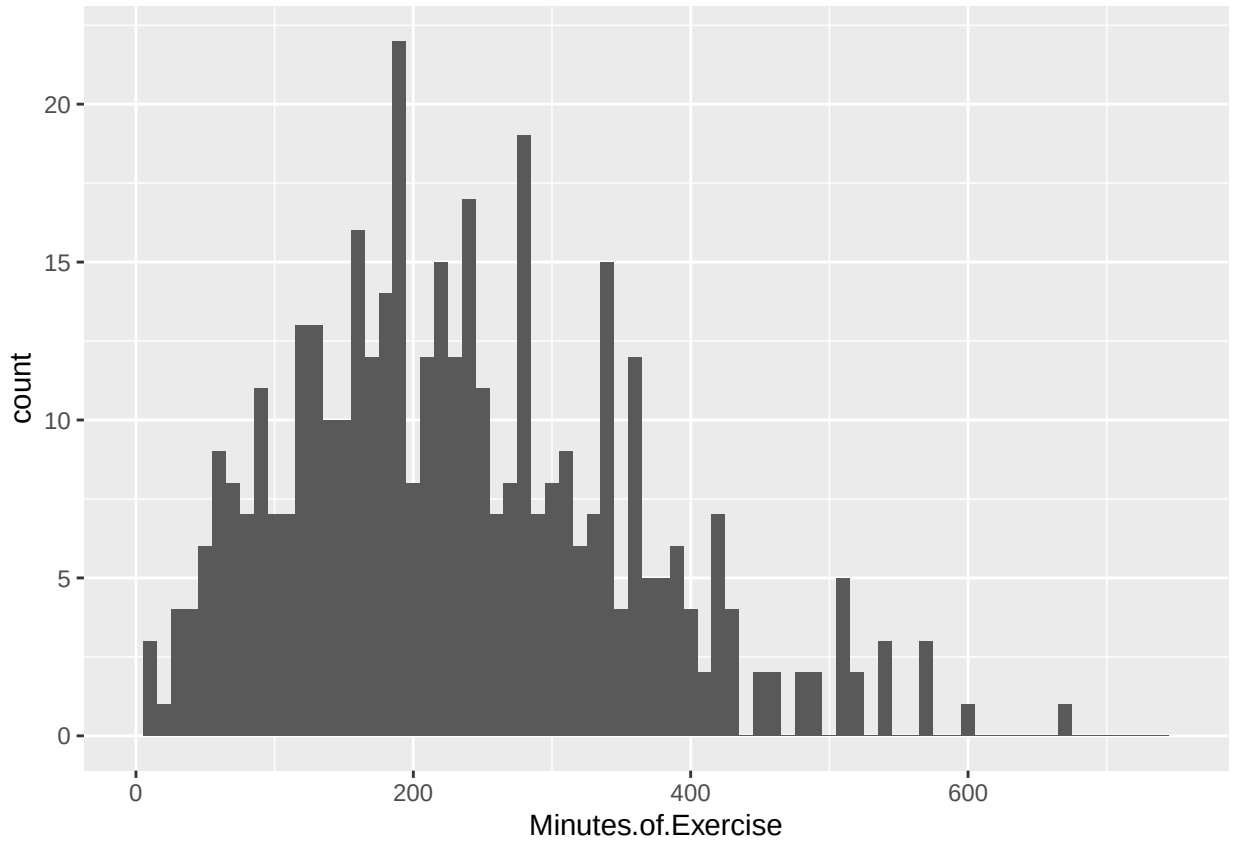

```
data %>%
    group_by(SurveyMonth, Overall.Health) %>%
    tally() %>%
  arrange(n)
```

```
## # A tibble: 22 x 3
## # Groups:   SurveyMonth [3]
##    SurveyMonth Overall.Health     n
##    <fct>                <int> <int>
##  1 October                  0     1
##  2 September                1     2
##  3 November                 1     4
##  4 October                  1     4
##  5 October                  2     4
##  6 September                2     4
##  7 October                  7     5
##  8 November                 7     7
##  9 September                7     9
## 10 September                3    10
## # ... with 12 more rows
```

```
# Predictor Variabl: Minutes of Exercise
ggplot(data = data) + geom_histogram(mapping = aes(x = Minutes.of.Exercise),binwidth = 10) + xlim(0,750)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```




```
data[data$Minutes.of.Exercise >600,]
```

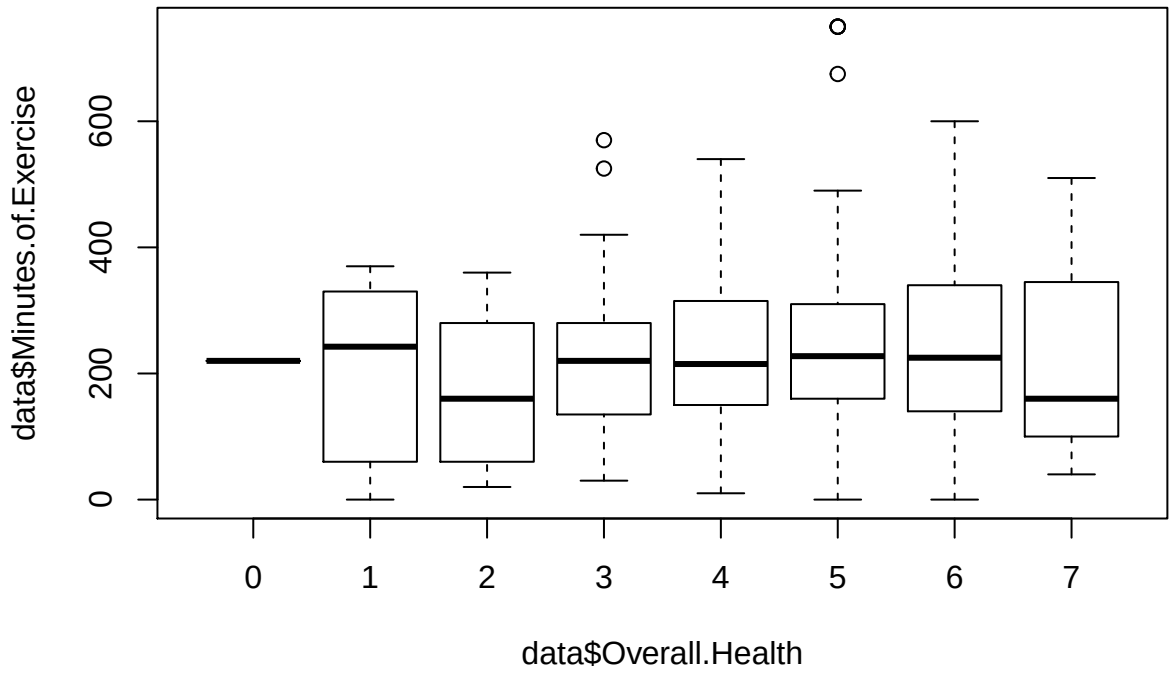
```
##       X study.ID SurveyMonth Uoft_year Enrolment     cGPA Gender Age
## 104 104   190206   September         3 full-time >2.5-3.5 female  19
## 105 105   190206    November         3 full-time >2.5-3.5 female  19
## 379 379   190114     October         4 full-time >3.5-3.9 female  22
## 380 380   190114   September         4 full-time >3.5-3.9 female  21
## 381 381   190114    November         4         0 >3.5-3.9 female  21
##           BMI AvgSittingOrLying ModerateOrHighExercise Exercise1 Exercise2
## 104     <18.5            7-10.9                      2       150       150
## 105     <18.5             4-6.9                8.0-10.0       150       150
## 379 18.5-24.9             4-6.9                  5.0-7.0       150       150
## 380 18.5-24.9            7-10.9                  5.0-7.0       150       150
## 381 18.5-24.9               >16                      >10       150       150
##     Exercise3 Exercise4 Exercise5 NumStressors Overall.Health Health1
## 104       150       150       150           >4              5       7
## 105       150       150       150            3              5       4
## 379       150       150       150           >4              5       6
## 380       150       150        75           >4              5       6
## 381       150       150       150           >4              5       6
##     Health2 Health3 Health4 Health5 AvgHoursOfSleep
## 104       5       3       5       6               8
## 105       5       3       5       6               7
## 379       5       5       5       5               6
## 380       5       6       5       5               7
## 381       5       6       5       5               5
##     NightsConsistentBedtime SignupReason Option Minutes.of.Exercise
```

```
## 104                 6.0-7.0            B      A                750
## 105                       5         <NA>   <NA>                750
## 379                       4         <NA>   <NA>                750
## 380                       5            A      A                675
## 381                       3         <NA>   <NA>                750
```

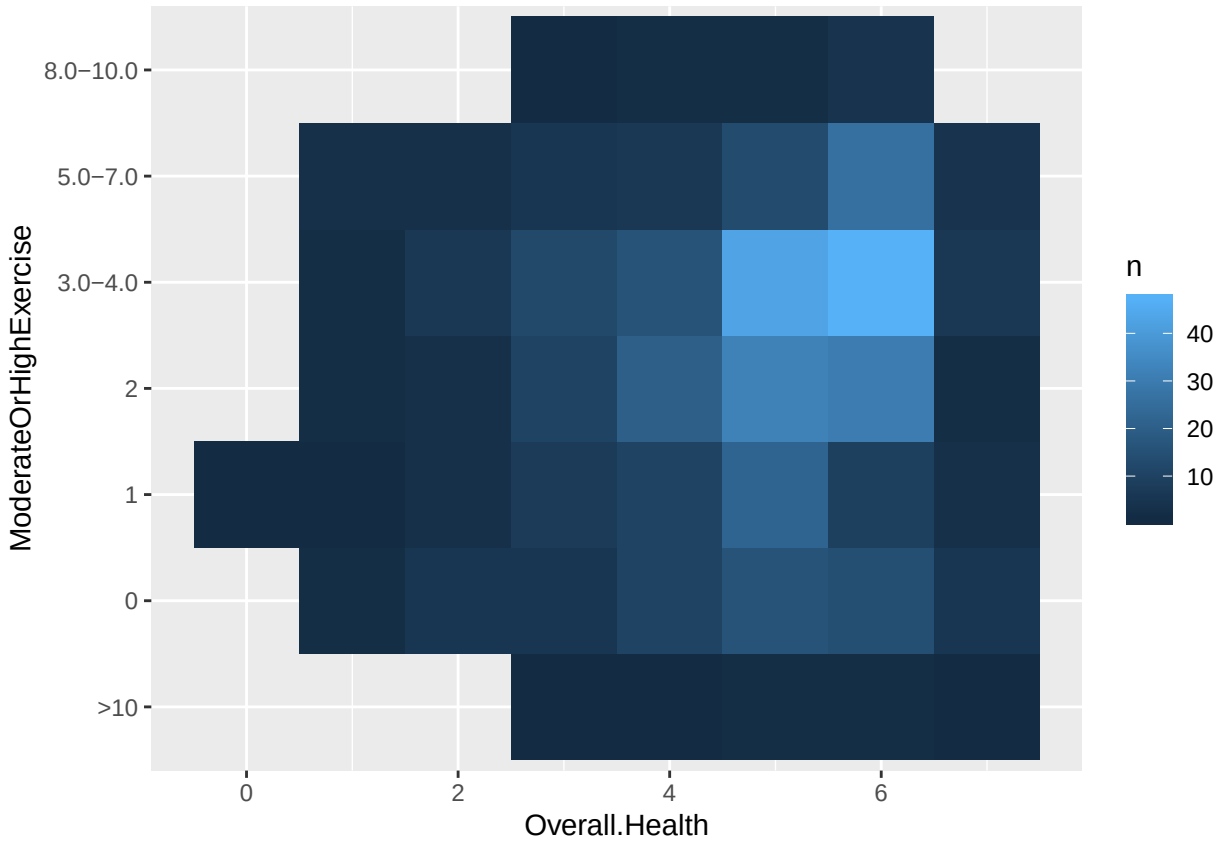
Relationship between Overall Health and predictor variables

1. Looking at boxplot showing the mental health and minutes of exercise, there does not seem to be an obvious pattern, that indicates more exercises will lead to better mental health.
but there's quite a bit of variation between the groups
2. From the heatmaps, there does not seem to be significant relationship between the overall mental health and number of stressors in the week. However, the hours of sleep and activeness of the subject seems to have a positive relationship the overall mental health.
3.

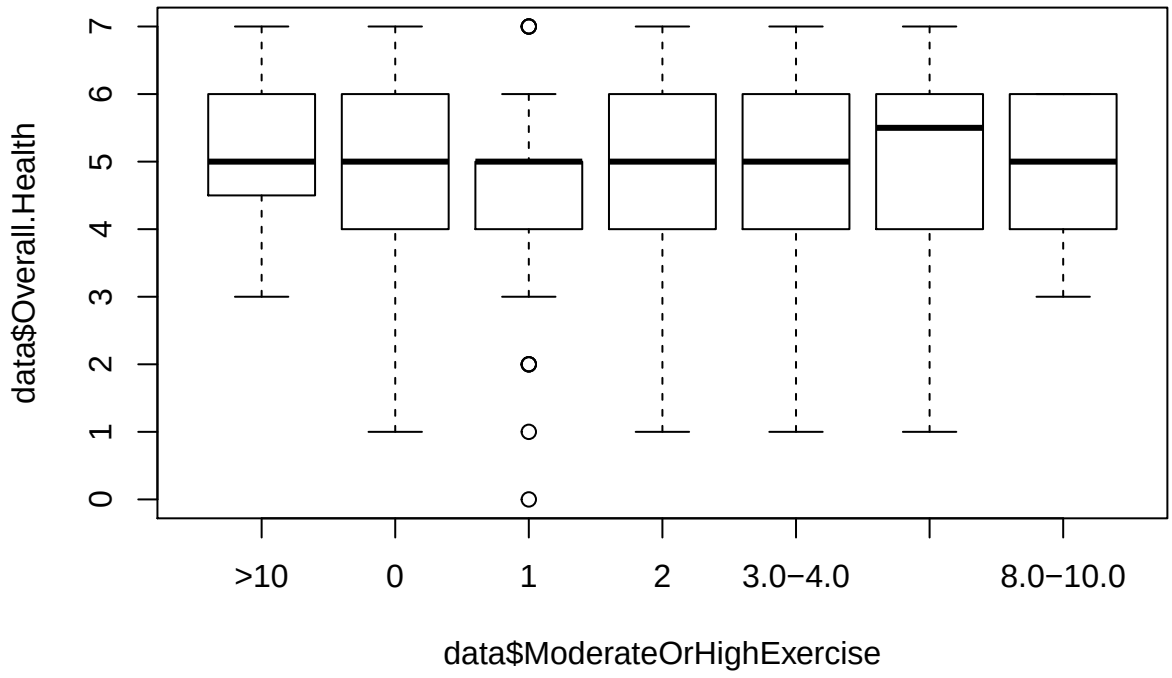
```
# Overall Health and Minutes of Exercise (Continous and Categorical)
boxplot(data$ Minutes.of.Exercise ~ data$Overall.Health,ylim = c(0,750))
```




```
# Overall Health and Activeness (Two Categorical)

data %>%
  count(Overall.Health,ModerateOrHighExercise) %>%
  ggplot(mapping = aes(x = Overall.Health, y = ModerateOrHighExercise)) +
  geom_tile(mapping = aes(fill = n))
```

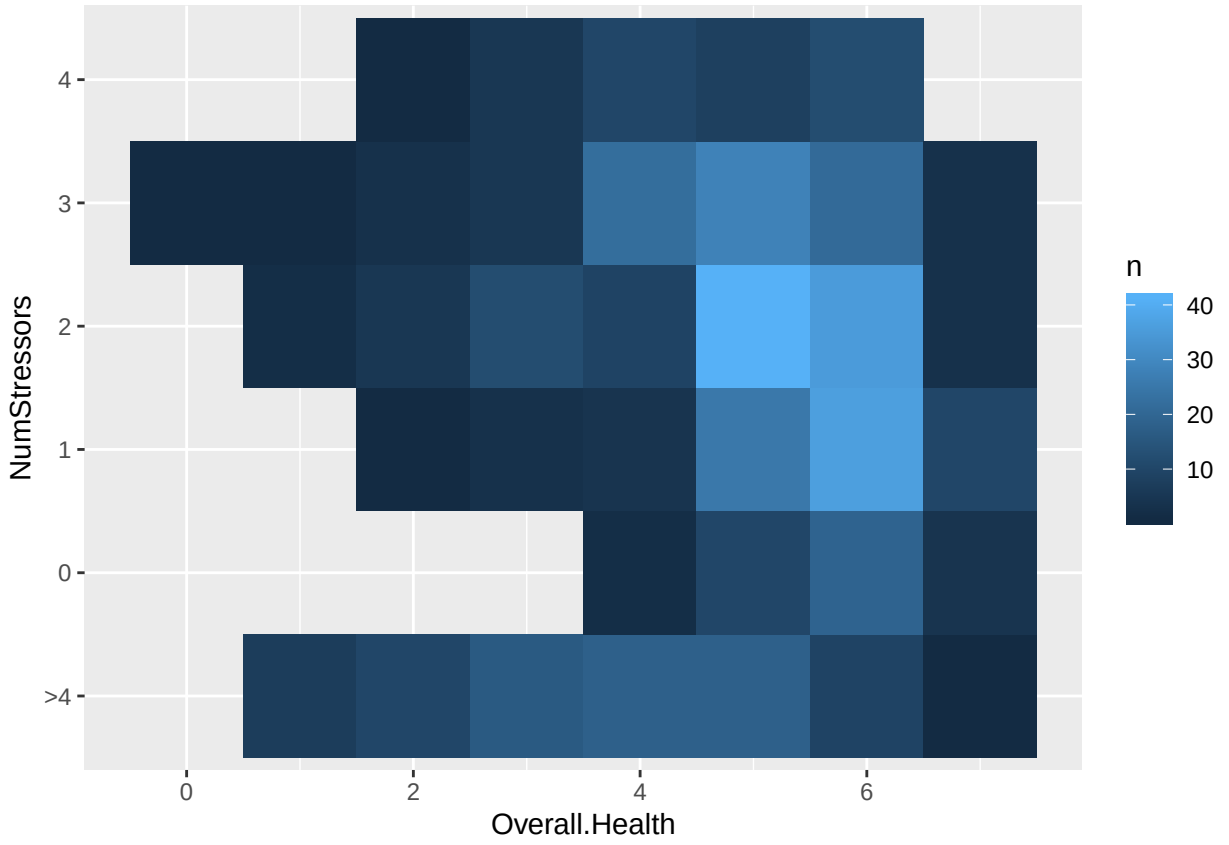



```
boxplot(data$Overall.Health~ data$ModerateOrHighExercise)
```

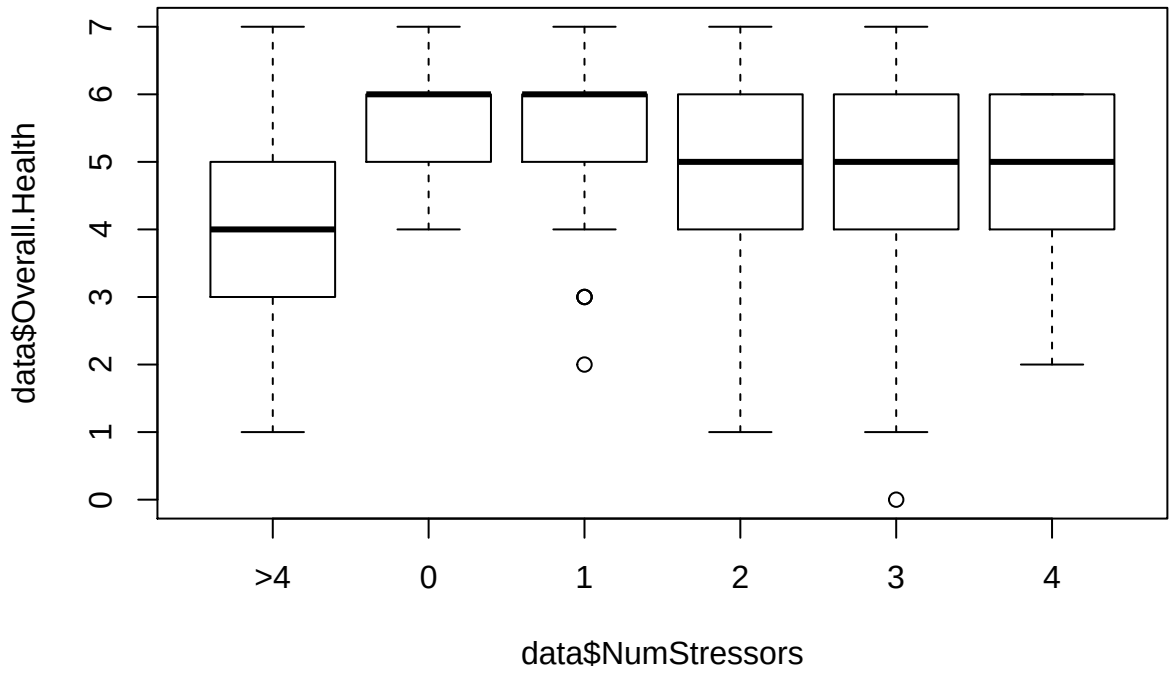



```
# Overall Health and Stressor (Two Categorical)

data %>%
  count(Overall.Health,NumStressors) %>%
```

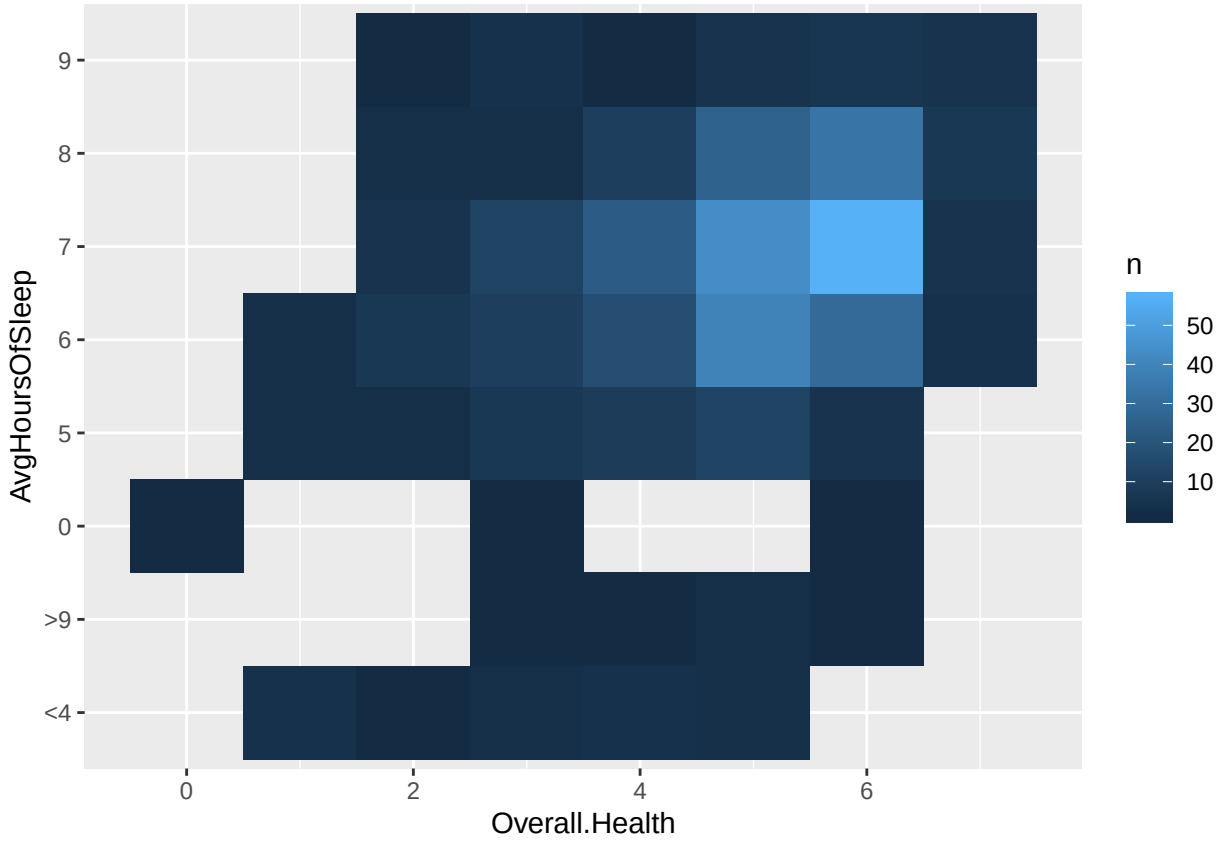
```
ggplot(mapping = aes(x = Overall.Health, y = NumStressors)) +
geom_tile(mapping = aes(fill = n))
```



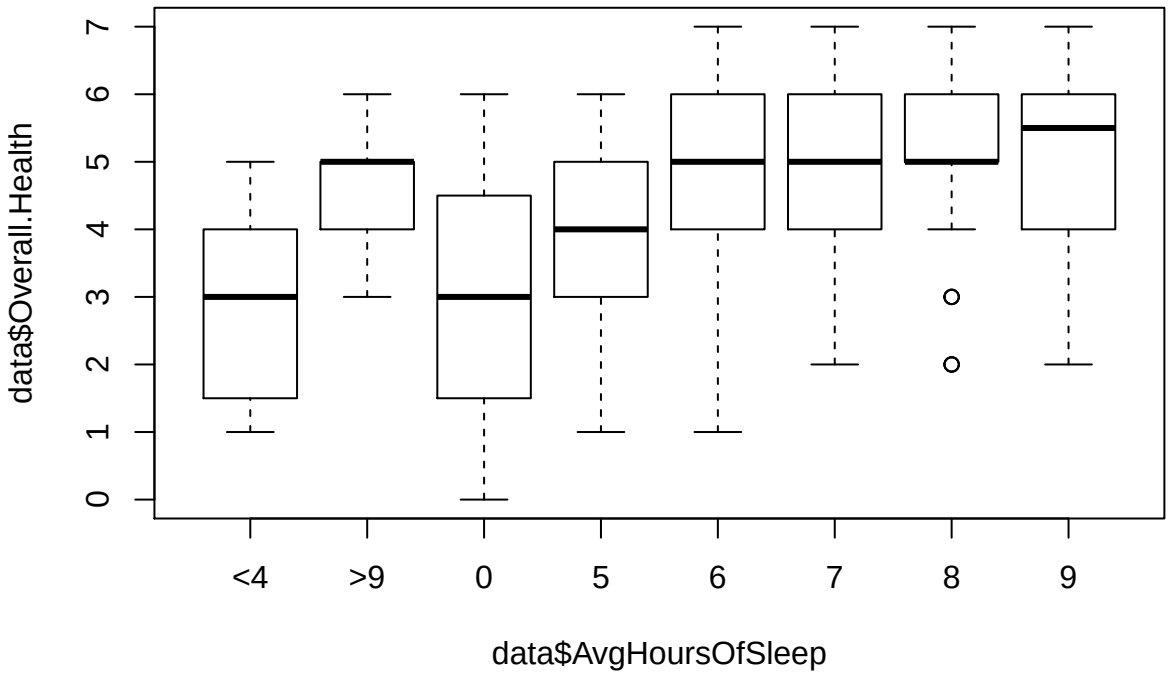

```
boxplot(data$Overall.Health~data$NumStressors)
```



```r
# Overall Health and Sleep Hours (Two Categorical)

data %>%
  count(Overall.Health,AvgHoursOfSleep) %>%
  ggplot(mapping = aes(x = Overall.Health, y = AvgHoursOfSleep)) +
  geom_tile(mapping = aes(fill = n))
```



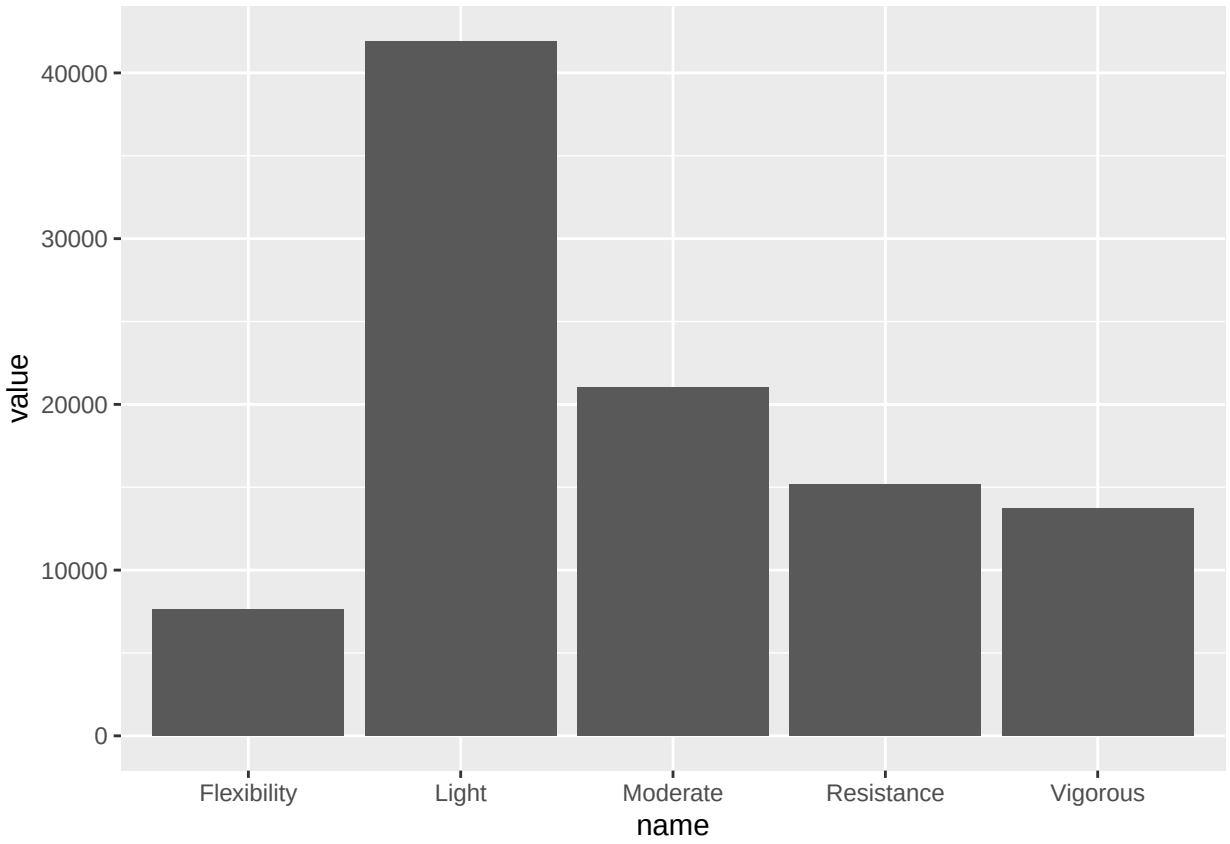

```r
boxplot(data$Overall.Health~ data$AvgHoursOfSleep)
```



Potential Alteration/Improvements

1. Given that there seems to be a relatively strong relationship between mental health status and the month of survey, it would be good to remove the strong relationship, to help me understand the relationship between mental health status and minutes of exercise.
2. The 420 samples are collected from 140 subjects throughout September-November,meaning multiple samples are collected from the same subject. To prevent pseudoreplication, Study ID could be used as the random effect in the model.

```
typeOfAerobic <- c("Light","Moderate","Vigorous","Resistance","Flexibility")
totalMinutes <- c(sum(data$Exercise1,na.rm = TRUE),
                  sum(data$Exercise2,na.rm = TRUE),
                  sum(data$Exercise3,na.rm = TRUE),
                  sum(data$Exercise4,na.rm = TRUE),
                  sum(data$Exercise5,na.rm = TRUE))

data3 <- data.frame(name =typeOfAerobic ,value=totalMinutes)

# Barplot
ggplot(data3, aes(x=name, y=value)) + geom_bar(stat = "identity")
```



3. In our predictor variable 'minutes of exercise', light aerobic exercises are taking a significant proportion of the exercises. Hence, it would be worth exploring the relationship of minutes of exercises without light aerobic and the other variables, as a lot of light aerobic activities are conducted passively (e.g. commuting to school, walking between classes).