# STA490 EDA - Xiaodong(Jenna) Fu

```
# setwd("/Users/JennaFu/Desktop/STA490/Fall Project")
data = read.csv("sta490_cognitive_flexibility_data.csv",header = TRUE)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

## Wide to Long

```
data1 <- data %>%
  select(ID, color_blind, yrs_english, video_games, device, headphones, order_of_levels,
         level1, sleep1, start_time1, OffTime1, OnTime1,
         Total_runs_Stroop_Off1, Total_runs_Stroop_On1, OnTime_minus_OffTime1) %>%
  mutate(order = 1);

data2 <- data %>%
  select(ID, color_blind, yrs_english, video_games, device, headphones,
         order_of_levels,
         level2, sleep2, start_time2, OffTime2, OnTime2,
         Total_runs_Stroop_Off2, Total_runs_Stroop_On2, OnTime_minus_OffTime2) %>%
  mutate(order = 2);

data3 <- data %>%
  select(ID, color_blind, yrs_english, video_games, device, headphones,
         order_of_levels,level3, sleep3, start_time3, OffTime3, OnTime3,
         Total_runs_Stroop_Off3, Total_runs_Stroop_On3, OnTime_minus_OffTime3) %>%
  mutate(order = 3);

names(data1)[8:15] <- c("distraction_level", "sleep", "start_time",
                        "OffTime", "OnTime", "Total_runs_Stroop_Off",
                        "Total_runs_Stroop_On", "OnTime_minus_OffTime")
names(data2)[8:15] <- c("distraction_level", "sleep", "start_time",
                        "OffTime", "OnTime", "Total_runs_Stroop_Off",
                        "Total_runs_Stroop_On", "OnTime_minus_OffTime")
names(data3)[8:15] <- c("distraction_level", "sleep", "start_time",
                        "OffTime", "OnTime", "Total_runs_Stroop_Off",
                        "Total_runs_Stroop_On", "OnTime_minus_OffTime")

data.long <- rbind(data1, data2, data3)

# Sort by ID
```

```
data.long <- data.long %>% arrange(ID, order)
```

## Data Cleaning

Remove the Total Number of Stroop On/Off that are less than 5, as these are not aligned with experiment design.

```
data.long = data.long[data.long$Total_runs_Stroop_On >= 5 && data.long$Total_runs_Stroop_Off >= 5]
```

Segment the start time into Morning(5AM-11PM),Afternoon(11PM-5PM),Evening(5PM-11AM),Midnight(11AM-5AM).

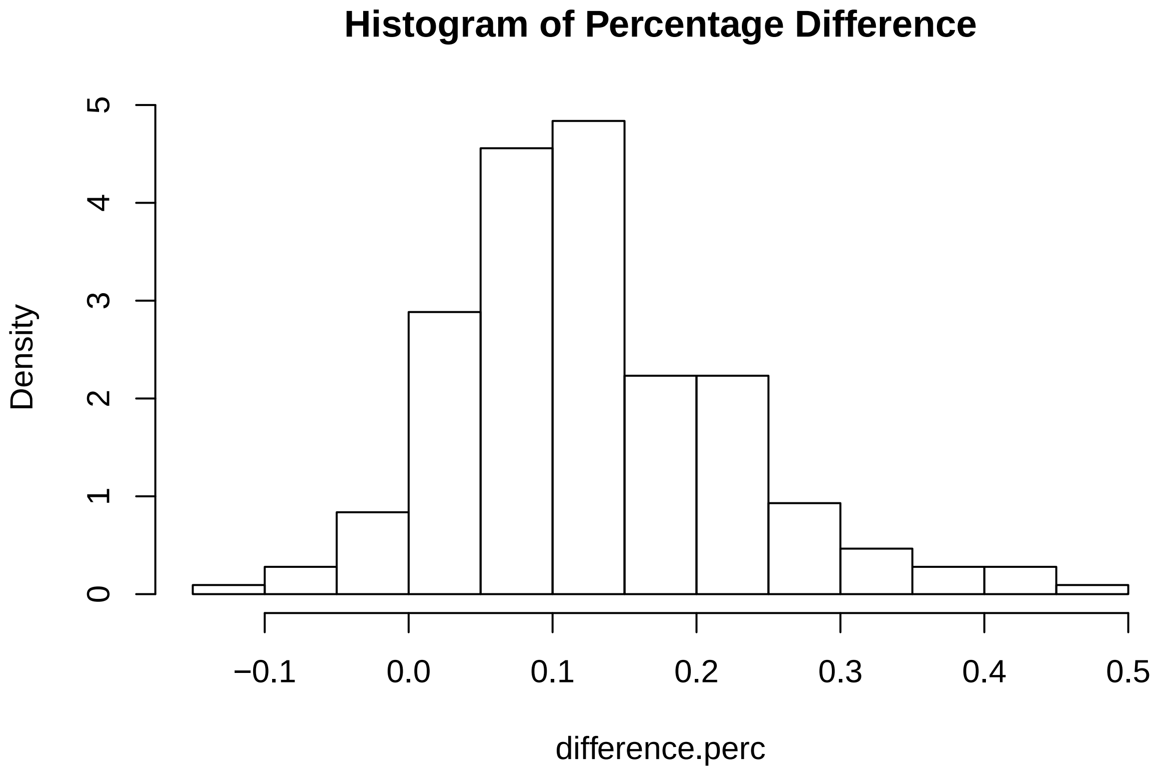
```
time <- as.POSIXct(strptime(data.long$start_time,"%H:%M:%S"),"UTC")

x=as.POSIXct(strptime(c("5:00:00","10:59:00",
                        "11:00:00","16:59:00",
                        "17:00:00","22:59:00"),"%H:%M:%S"),"UTC")
data.long$start_time = case_when(
between(time,x[1],x[2]) ~"Morning",
between(time,x[3],x[4]) ~"Afternoon",
between(time,x[5],x[6]) ~"Evening",
TRUE ~"Midnight")
```

## Variables Used

```
# I converted the difference between OnTime and OffTime into percentage
difference.perc = data.long$OnTime_minus_OffTime/data.long$OffTime
# These are the explanatory variables I will be using
start_time = data.long$start_time
level = data.long$distraction_level
sleep = data.long$sleep
video_games = data.long$video_games
year_of_eng = data.long$yrs_english
```

## Check for Normality

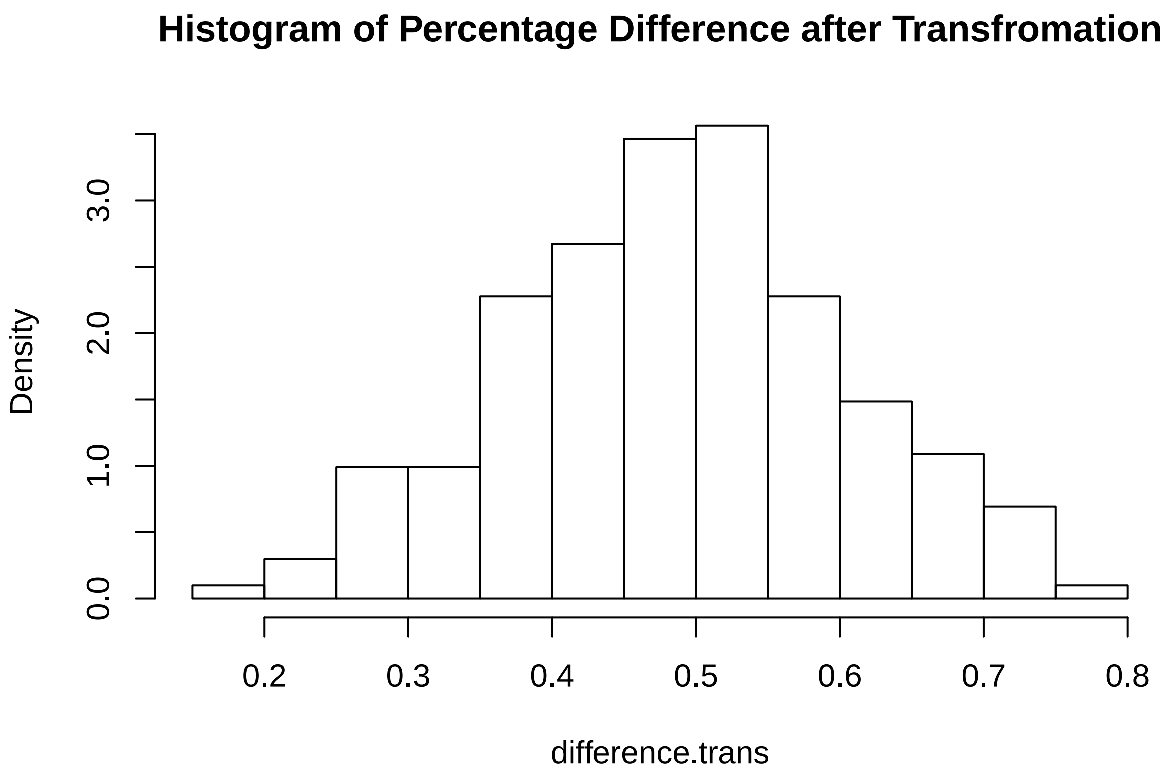
```
# I want to check for normality of the data
# From the histogram, the data looks right-skewed.
hist1 = hist(difference.perc, probability = TRUE,
             main = "Histogram of Percentage Difference")
```



```
# The Shapiro Wilk Test shows that the data is not normally distributed
shapiro.test(difference.perc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  difference.perc
## W = 0.96187, p-value = 1.589e-05
```

```
# Hence I want to transform the data using cube root transformation,
# as it can be applied on zero or negative values
difference.trans = (difference.perc)^(1/3)
# The data looks more normally distributed,
hist2 = hist(difference.trans, probability = TRUE,
             main = "Histogram of Percentage Difference after Transfromation")
```



```
# that conclusion is supported by the Shapiro Wilk Test
shapiro.test(difference.trans)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  difference.trans
## W = 0.99457, p-value = 0.6795
```

## Goals of the analysis

I wanted to investigate the effects of auditory distraction on the cognitive flexibility. In this experiment design, auditory distraction occurs in three levels: music with lyrics, music with no lyrices and control (quiet). This is represented by the variable "level" in our R code. Whereas the cognitive flexibility is represented by the percentage difference between OnTime and OffTime, which is an isolated measure of cognitive flexibility.

Beside auditory distraction, I also wanted to investigate the effect of other variables, such as the time of experiment, the hours of sleep on the day before and video game playing status of the research participants.

## Features of Data Observed

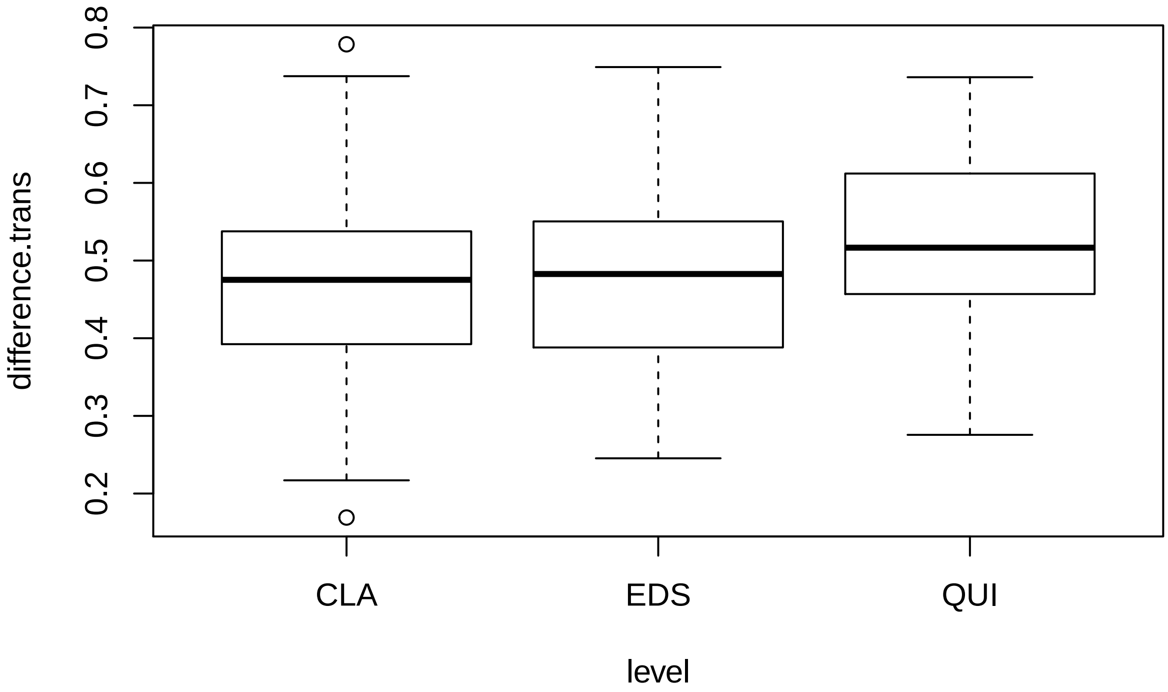
In model 1, we investigated percentage difference as a function of level of auditory distraction.From this summary statistics, there seems to be a significant difference between the control group and the music with no lyrics group. However, there is no significant difference between the two groups with music.

```
mod1 = glm(difference.trans ~ level,family = gaussian())
summary(mod1)
```

```
##
## Call:
## glm(formula = difference.trans ~ level, family = gaussian())
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.295872  -0.074356   0.005917   0.080952   0.313477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.464996   0.014354   32.39  < 2e-16 ***
## levelEDS    0.008411   0.020008    0.42  0.67464
## levelQUI    0.060082   0.020300    2.96  0.00345 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.0135988)
##
##     Null deviance: 2.8467  on 201  degrees of freedom
## Residual deviance: 2.7062  on 199  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: -289.92
##
## Number of Fisher Scoring iterations: 2
```

Looking at the boxplot, it seems like the mean for the quiet group is slightly higher than that of the two music groups. It also seems to have a narrower range comparing to the other two. There also are some outliers in the CLA(music with no lyrics) group.

```
boxplot(difference.trans ~ level)
```



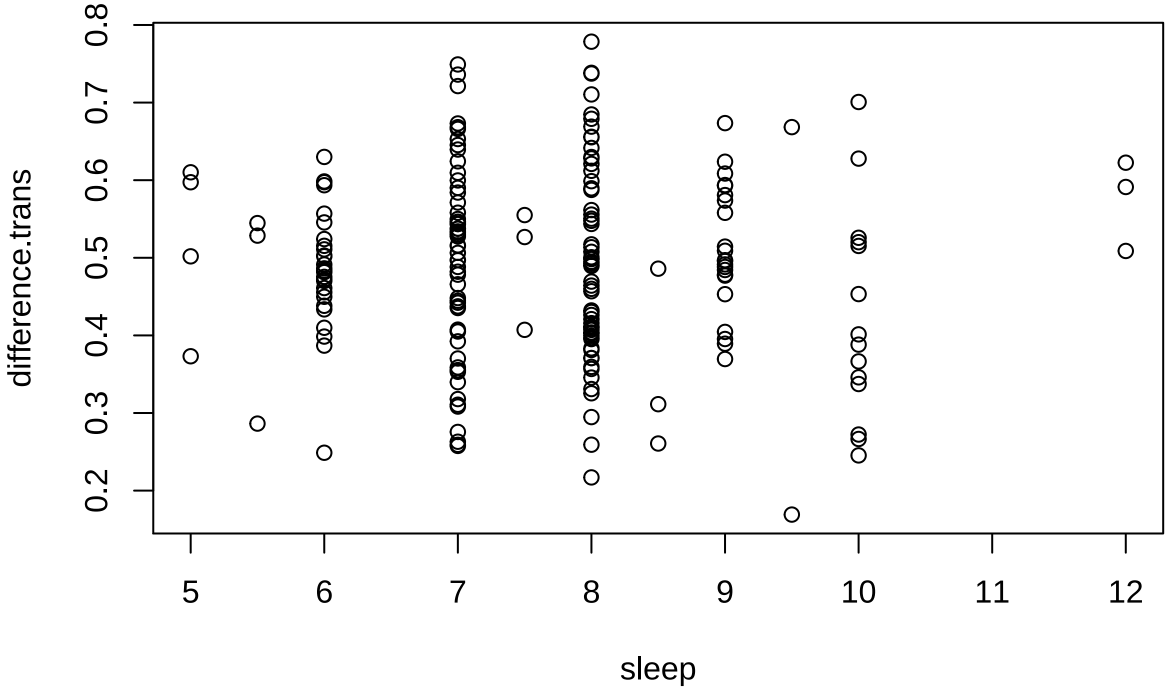
In model 2, we investigated percentage difference as a function of the other three variables.Looking at the other variables, there does not seem to be a significant factor that leads to the change in percentage difference.

```r
mod2 = glm(difference.trans ~ sleep + start_time + video_games,family = gaussian())
summary(mod2)
```

```
##
## Call:
## glm(formula = difference.trans ~ sleep + start_time + video_games,
##     family = gaussian())
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.302246  -0.083864   0.008302   0.080410   0.268309
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.512305   0.053261   9.619   <2e-16 ***
## sleep              -0.003192   0.006677  -0.478    0.633
## start_timeEvening  -0.012540   0.018686  -0.671    0.503
## start_timeMidnight  0.021466   0.031588   0.680    0.498
## start_timeMorning   0.030276   0.032263   0.938    0.349
## video_gamesYes      0.001931   0.017252   0.112    0.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.01431659)
##
##     Null deviance: 2.8467  on 201  degrees of freedom
## Residual deviance: 2.8061  on 196  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: -276.6
##
## Number of Fisher Scoring iterations: 2
```

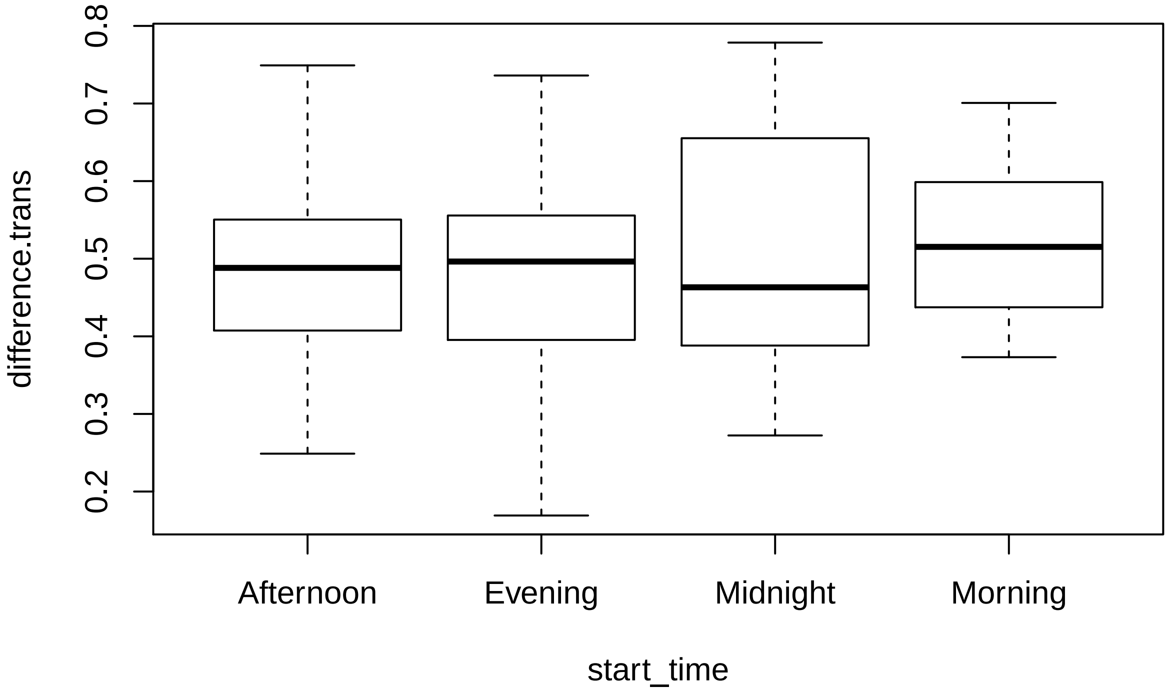
Looking at the plot, there appears to be no pattern between the percentage difference and the hours of sleep of the participants

```r
plot(sleep,difference.trans)
```



Looking at the boxplot, the means of the four categories are very similiar. However, the ranges look narrower for category 'Morning' comparing to the other three. The category 'Midnight' seems especially positively skewed comparing to the other three, which look normally distributied.

```r
boxplot(difference.trans ~ start_time)
```


difference.trans
Afternoon
Evening
Midnight
Morning
start_time