HW 1

Anran Yao

2024-02-06

github link: https://github.com/anranyao/project_620/tree/master

Problem 1:

(a) Studies have showed effects of screen time usage, such as associating with BMI[1], social media screen time usage associated with mental health [2]. We collected the data of screen time usage in order to study the possible behavioral patterns of individuals, and to give suggestions for reducing the screen time usage. Some of the possible patterns maybe the association between procrastination and daily screen time[3], social screen time and mental health[2].

[1] Duch, H., Fisher, E. M., Ensari, I., & Harrington, A. (2013). Screen time use in children under 3 years old: a systematic review of correlates. International journal of behavioral nutrition and physical activity, 10, 1-10.

[2] Barthorpe, A., Winstone, L., Mars, B., & Moran, P. (2020). Is social media screen time really associated with poor adolescent mental health? A time use diary study. Journal of affective disorders, 274, 864-870.

[3] Hammoudi, S. F., Mreydem, H. W., Abou Ali, B. T., Saleh, N. O., Chung, S., Hallit, S., & Salameh, P. (2021). Smartphone screen time among university students in Lebanon and its association with insomnia, bedtime procrastination, and body mass index during the COVID-19 pandemic: a cross-sectional study. Psychiatry investigation, 18(9), 871.

(b) The Informed Consent Form is used to recruit participants to a study. In our case, we recruited the students from biostats 620 for the study. An ICF usually contains study plans and how data are collected and used in the study. Given this information, the participants can understand their roles and be more adherence to the study, resulting in a better data quality.

(c) Screen activity is recorded in real-time by the mobile device. We collected daily entries of total screen time: (Total.ST: total screen time in HH-MM format, and Total.ST.min: total screen time in MM format), social app screen time (Social.ST: social app screen time in HH-MM format, and Social.ST.min: social app screen time in MM format), total number of times the user picked up the phone (Pickups), and the time of the first pick-up (Pickup.1st). For the data of first pickup, we use the record after user's wake-up to mark the beginning and end of the user's day. We are using the screen time data over 34 days (from 12/24/2023 to 01/26/2024) stored in the mobile device.

(d) Added new variables into the dataset

```
dat<-read.csv("ScreenTime_AnranYao.csv") %>% slice_head(n = 34)
head(dat)
```

```
##         Date Total.ST Total.ST.min Social.ST Social.ST.min Pickups Pickup.1st
## 1 12/24/2023    8h39m          519     2h25m           145     257       9:00
```

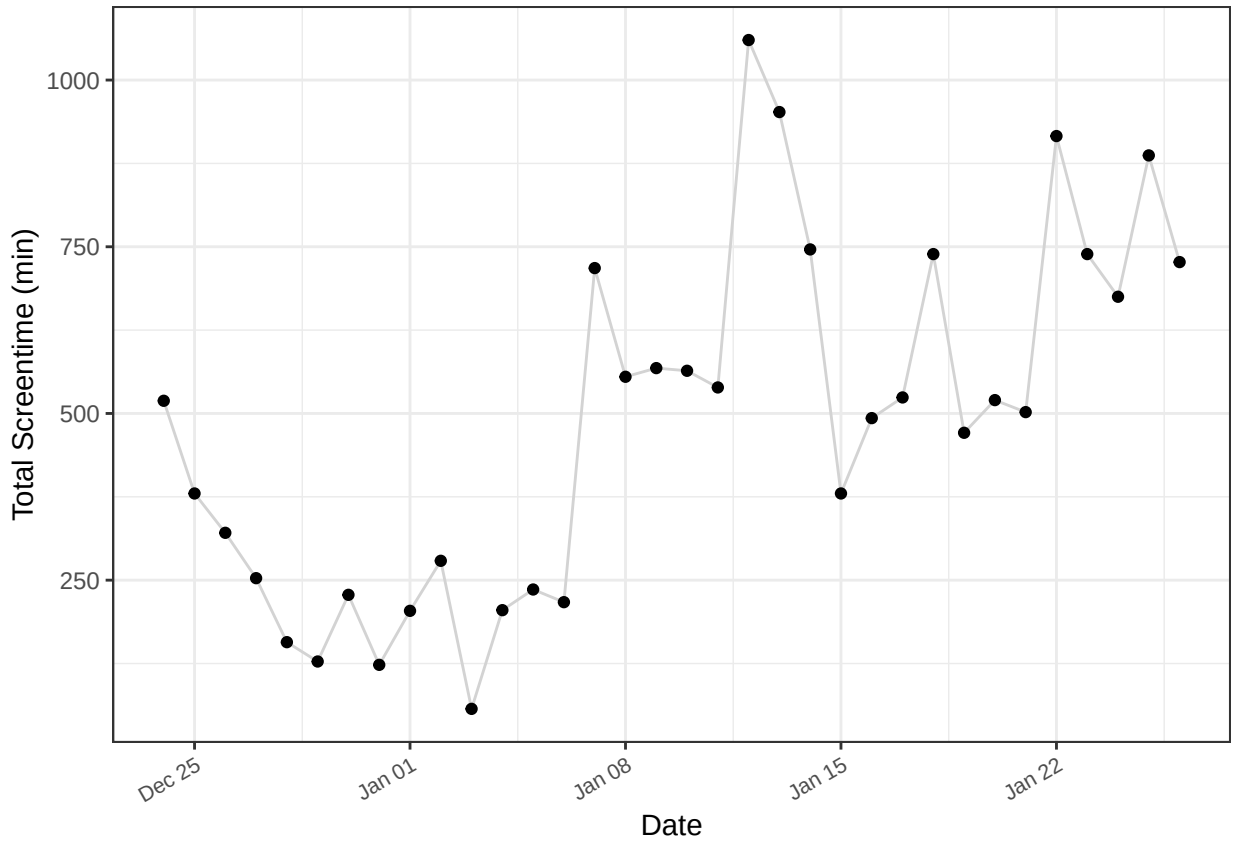
```
## 2 12/25/2023    6h20m          380     0h47m           47     240       9:45
## 3 12/26/2023    5h21m          321     0h31m           31     150       8:00
## 4 12/27/2023    4h13m          253     0h32m           32      55       9:00
## 5 12/28/2023    2h37m          157     1h37m           97      67       9:28
## 6 12/29/2023     2h8m          128     0h44m           44      67      12:40
##   Daily.Social.Prop Daily.Duration.Use
## 1        0.27938343           2.019455
## 2        0.12368421           1.583333
## 3        0.09657321           2.140000
## 4        0.12648221           4.600000
## 5        0.61783439           2.343284
## 6        0.34375000           1.910448
```

Problem 2

(a) Time series plots

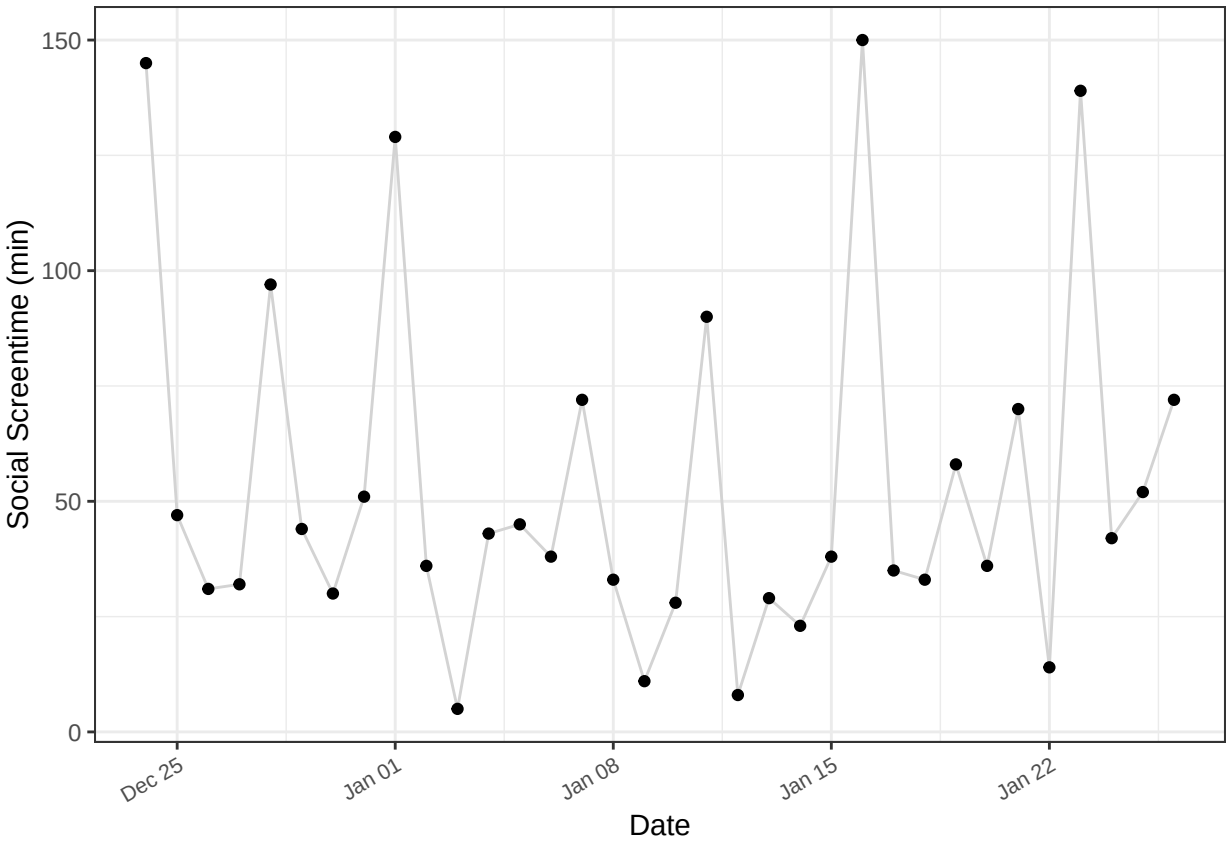
```
dat <- dat %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%Y"),
         Day = weekdays(Date, abbreviate = TRUE),
         weekend = factor(Day %in% c("Sun", "Sat")))

dat %>%
  ggplot(aes(x=Date,y=Total.ST.min)) +
  geom_line(color="lightgrey") +
  geom_point() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, size = rel(0.9)))+
  labs(x="Date",y="Total Screentime (min)")
```



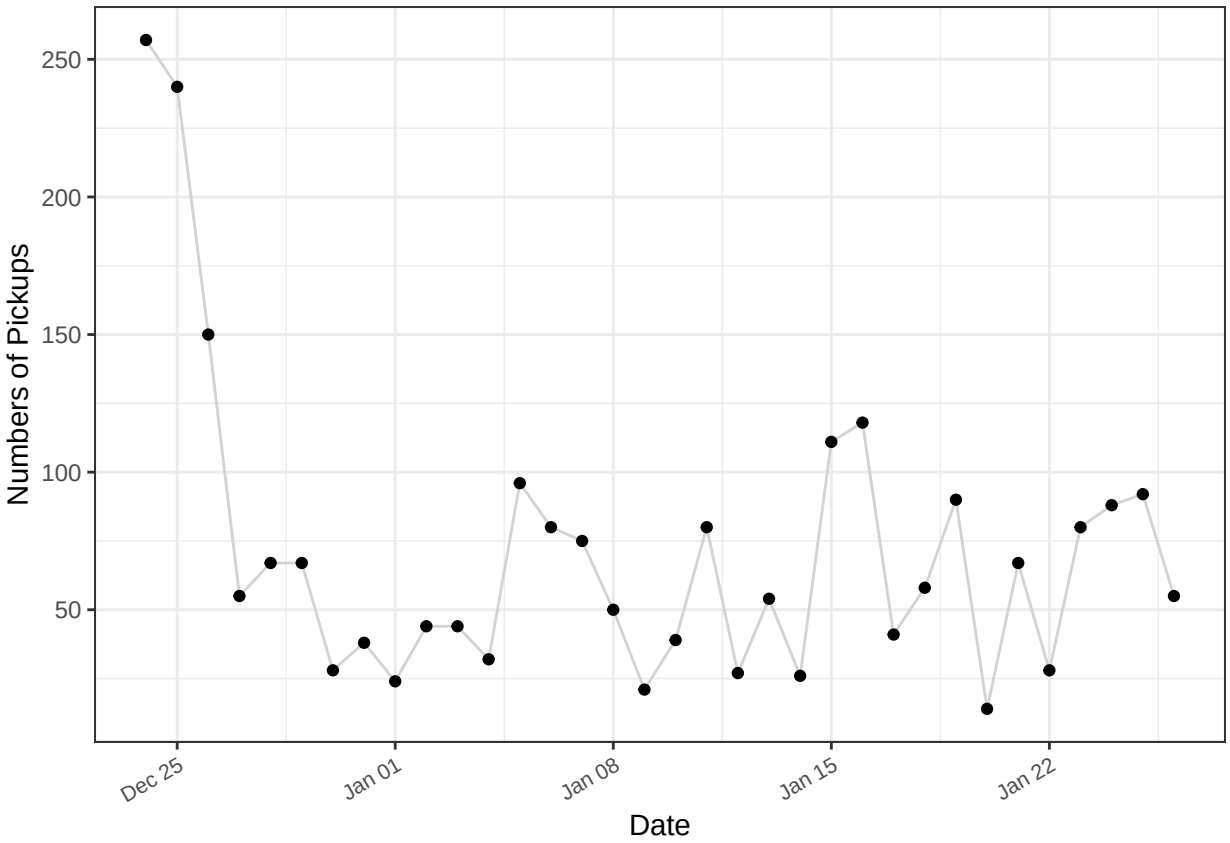
- Total Screen Time variates between about 50 to 1050. The highest value is on Jan 12, and lowest on Jan 03. The time series have a decreasing trend at the end of December and increase after Jan 05.

```
dat %>%
  ggplot(aes(x=Date,y=Social.ST.min)) +
  geom_line(color="lightgrey") +
  geom_point() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, size = rel(0.9)))+
  labs(x="Date",y="Social Screentime (min)")
```



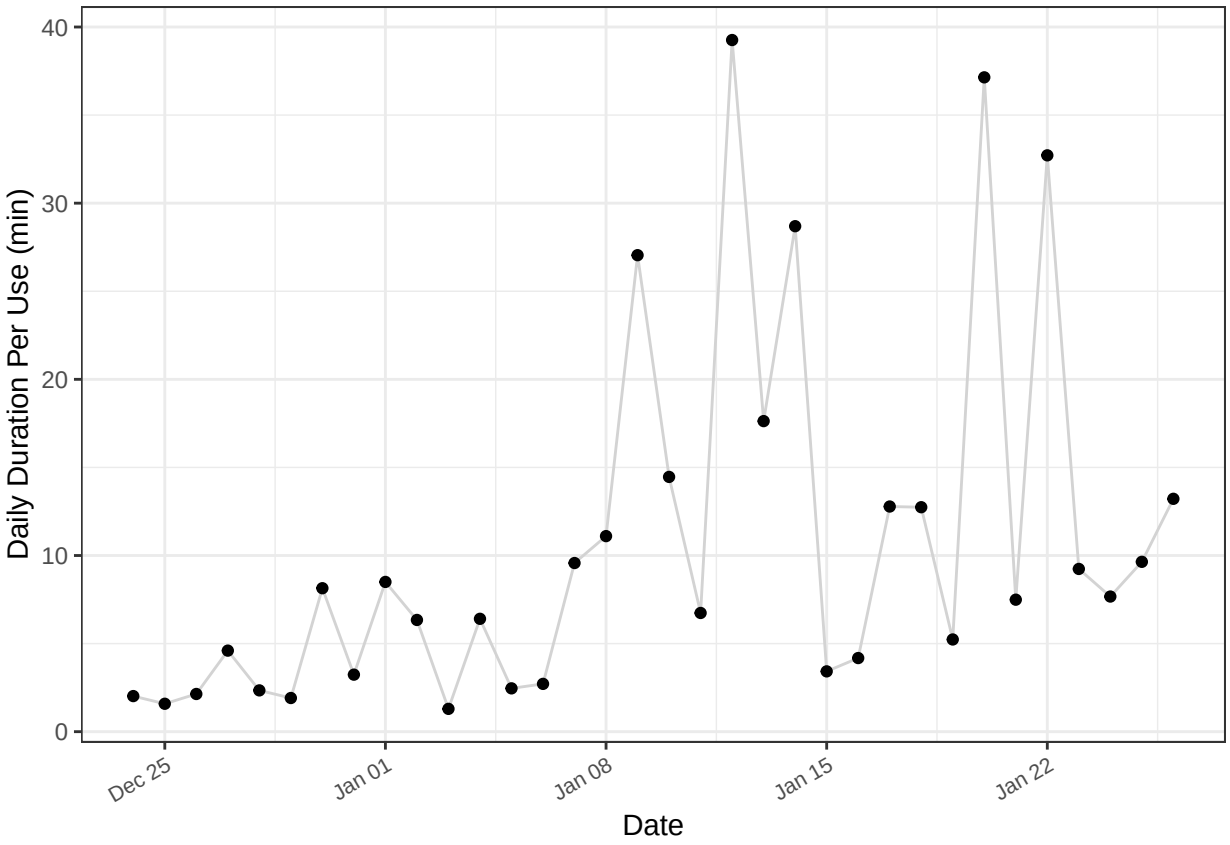
- Social Screen Time variates between about 0 to 150. The highest value is on Jan 16, and lowest on Jan 03. The time series fluctuates and doesn't have a particular pattern.

```
dat %>%
  ggplot(aes(x=Date,y=Pickups)) +
  geom_line(color="lightgrey") +
  geom_point() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, size = rel(0.9)))+
  labs(x="Date",y="Numbers of Pickups")
```



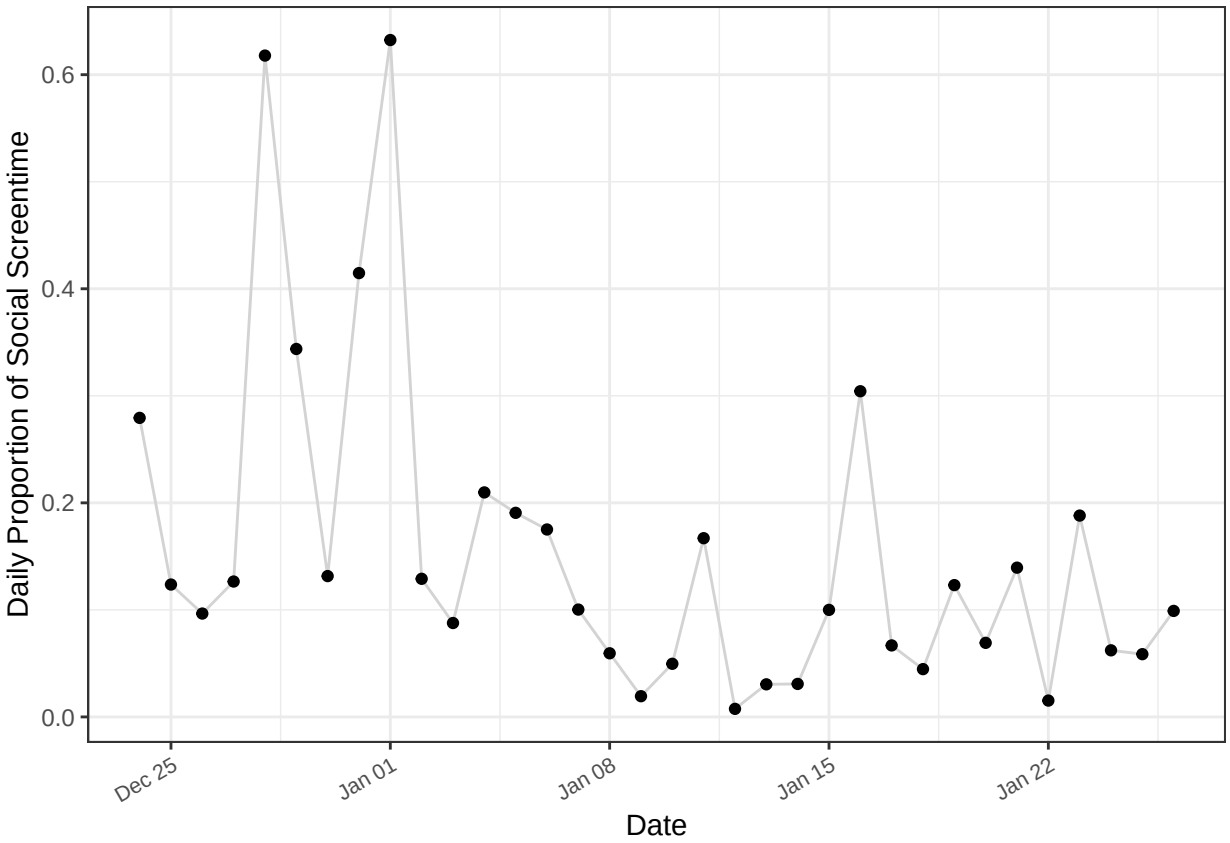
- Numbers of Pickups variates between about 50 to more than 250. The highest value is on Dec 24, and lowest on Jan 20. The time series fluctuates and doesn't have a particular pattern, except three days from Dec 24 to Dec 26.

```
dat %>%
  ggplot(aes(x=Date,y=Daily.Duration.Use)) +
  geom_line(color="lightgrey") +
  geom_point() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, size = rel(0.9)))+
  labs(x="Date",y="Daily Duration Per Use (min)")
```



- Daily Duration Per Use variates between about 2min to 40min. The highest value is on Jan 12, and lowest on Jan 3. The time series is getting more fluctuating.

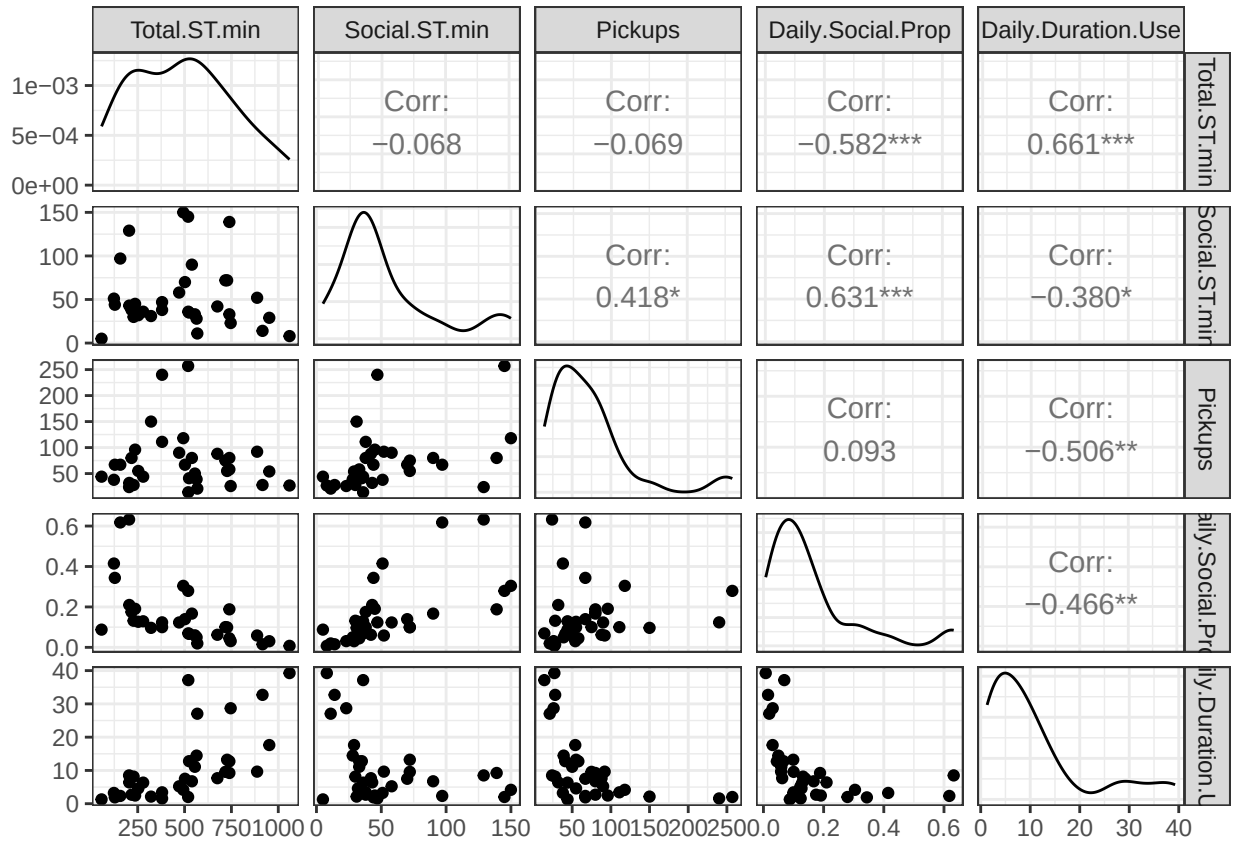
```
dat %>%
  ggplot(aes(x=Date,y=Daily.Social.Prop)) +
  geom_line(color="lightgrey") +
  geom_point() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, size = rel(0.9)))+
  labs(x="Date",y="Daily Proportion of Social Screentime")
```



- Daily Proportion of Social Screentime variates between about 0% to 65%. The highest value is on Jan 1, and lowest on Jan 12. The time series is getting less fluctuating.

(b) scatterplots

```
dat %>%
  ggpairs(columns = c(3,5,6,8,9), progress = FALSE)+ theme_bw()
```



- We observed highest positive correlation between Daily Duration per Use and Total Screen Time(0.66), high negative correlation between Daily proportion of social screentime and total screen time(-0.58), high correlation between Social screentime and Daily proportion of social screentime(0.63), high negative correlation between daily pickups and daily duration per use. The scatter plots showed some linear trend.

(c) Occupation time curve

```
dat_melted <- melt(dat, id.vars = "Date",
                   measure.vars = c("Total.ST.min","Social.ST.min", "Pickups",
                                    "Daily.Social.Prop", "Daily.Duration.Use"))
calculate_cdf <- function(ts, threshold) {
  mean(ts >= threshold)
}
cdf_data <- data.frame()
for (variable in c("Total.ST.min","Social.ST.min", "Pickups",
                                    "Daily.Social.Prop", "Daily.Duration.Use")) {
  thresholds <- seq(min(dat[[variable]]), max(dat[[variable]]), length.out = 100)

  # Calculate CDF for each threshold
  cdf_values <- sapply(thresholds, function(c) calculate_cdf(dat[[variable]], c))

  # Results
  temp_df <- data.frame(
    Date = rep(dat$Date[1], length(thresholds)),
    threshold = thresholds,
    cdf = cdf_values,
```

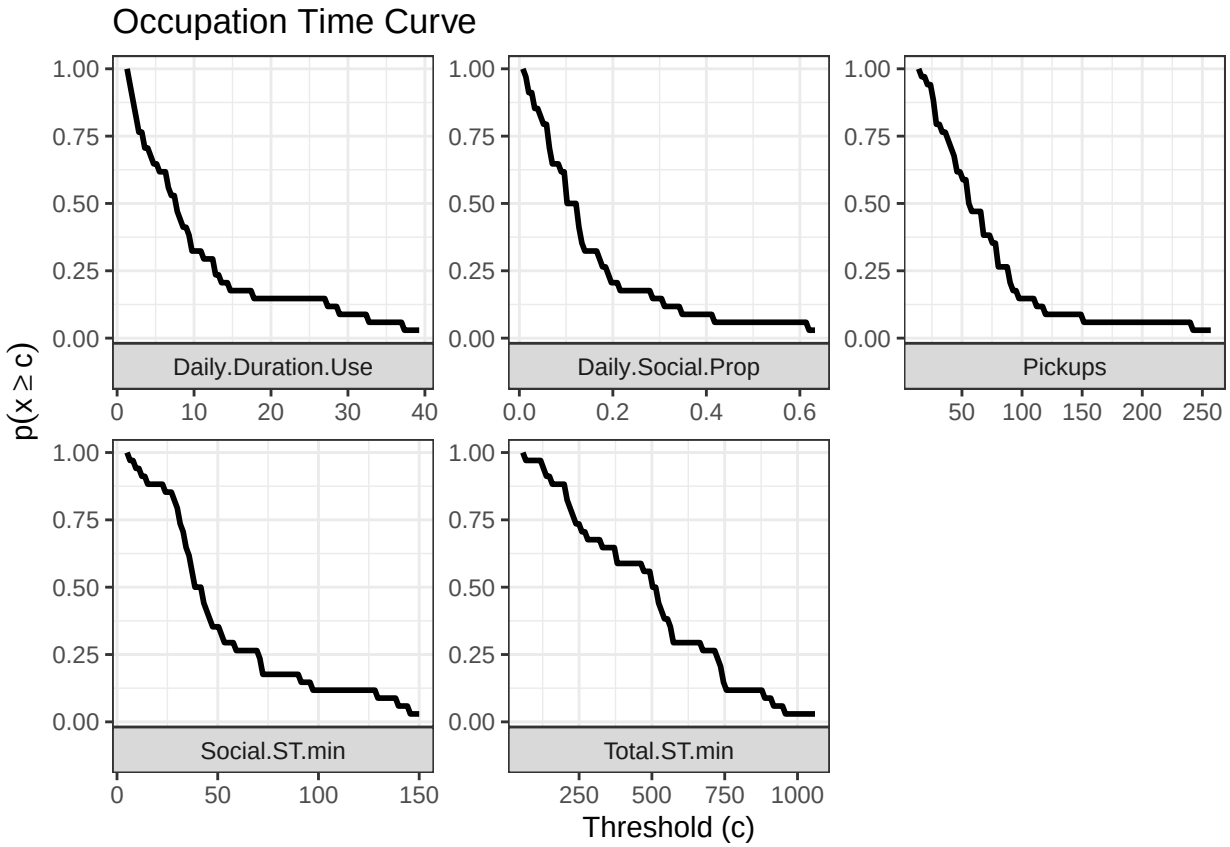
```
    variable = rep(variable, length(thresholds))
  )

  cdf_data <- rbind(cdf_data, temp_df)
}

# Plot separate occupation time curves for each variable
ggplot(cdf_data, aes(x = threshold, y = cdf)) +
  geom_line(size = 1) +
  labs(title = "Occupation Time Curve",
       x = "Threshold (c)",
       y = expression(p(x >= c))) +
  facet_wrap(~variable, scales = "free", switch = "x", ncol = 3) +
  theme_bw()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

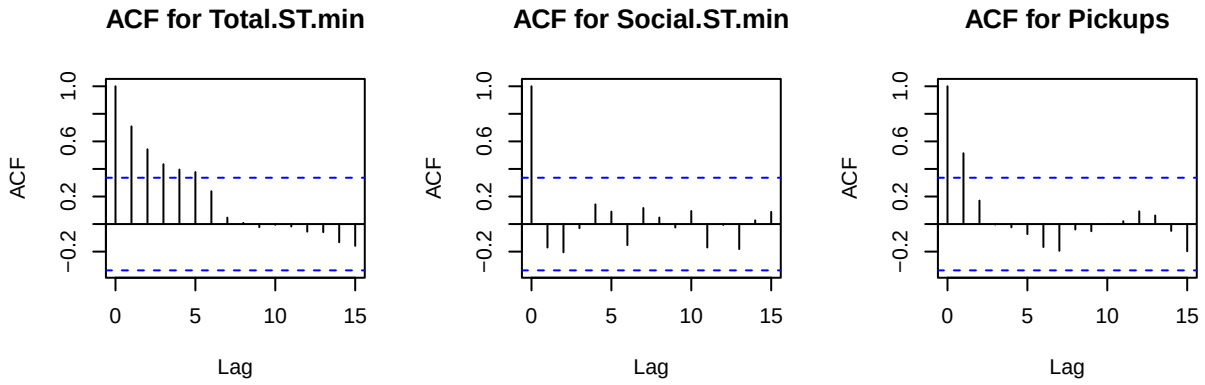
```
## Warning: The ‘switch‘ argument of ‘facet_wrap()‘ is deprecated as of ggplot2 2.2.0.
## i Please use the ‘strip.position‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

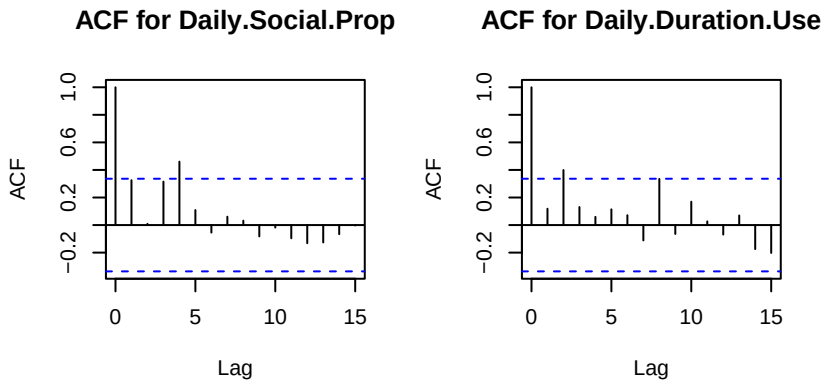


- The OTC for daily duration per use, daily social proportion and pickups showed that the participant is a less active person, while social screen time and total screen time shows that the person is an active person.

(d) ACF

```
variables_to_plot <- c("Total.ST.min", "Social.ST.min", "Pickups", "Daily.Social.Prop", "Daily.Duration
par(mfrow = c(2, 3))
acf(dat[["Total.ST.min"]], main = paste("ACF for", "Total.ST.min"), na.action = na.pass)
acf(dat[["Social.ST.min"]], main = paste("ACF for", "Social.ST.min"), na.action = na.pass)
acf(dat[["Pickups"]], main = paste("ACF for", "Pickups"), na.action = na.pass)
acf(dat[["Daily.Social.Prop"]], main = paste("ACF for", "Daily.Social.Prop"), na.action = na.pass)
acf(dat[["Daily.Duration.Use"]], main = paste("ACF for", "Daily.Duration.Use"), na.action = na.pass)
```






```
acf(dat[["Total.ST.min"]],plot=FALSE)
```

```
##
## Autocorrelations of series 'dat[["Total.ST.min"]]', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.709  0.543  0.434  0.395  0.377  0.238  0.047  0.009 -0.023 -0.005
##     11     12     13     14     15
## -0.018 -0.055 -0.059 -0.132 -0.158
```

```
acf(dat[["Social.ST.min"]], plot=FALSE)
```

```
##
## Autocorrelations of series 'dat[["Social.ST.min"]]', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000 -0.170 -0.205 -0.030  0.143  0.090 -0.153  0.116  0.047 -0.025  0.095
##     11     12     13     14     15
## -0.170 -0.006 -0.182  0.027  0.088
```

```
acf(dat[["Pickups"]], plot=FALSE)
```

```
##
## Autocorrelations of series 'dat[["Pickups"]]', by lag
```

```
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.514  0.170 -0.003 -0.024 -0.073 -0.167 -0.194 -0.040 -0.052  0.000
##     11     12     13     14     15
##  0.021  0.093  0.062 -0.050 -0.197
```

```
acf(dat[["Daily.Social.Prop"]], plot=FALSE)
```

```
##
## Autocorrelations of series 'dat[["Daily.Social.Prop"]]', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.325  0.009  0.315  0.460  0.109 -0.055  0.061  0.032 -0.082 -0.019
##     11     12     13     14     15
## -0.096 -0.132 -0.126 -0.066 -0.003
```

```
acf(dat[["Daily.Duration.Use"]], plot=FALSE)
```

```
##
## Autocorrelations of series 'dat[["Daily.Duration.Use"]]', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.119  0.400  0.131  0.060  0.114  0.071 -0.111  0.335 -0.064  0.170
##     11     12     13     14     15
##  0.027 -0.069  0.070 -0.174 -0.202
```

- According to the autocorrelation values and cutoffs(2/sqrt(34)=0.343), we observed lag-5 autocorrelation ACF(5) of the increments for total screen time (estimate = 0.377). For Pickups the estimated lag-1 ACF of increments is 0.514, showing that the correlation between two adjacent days is 0.514.

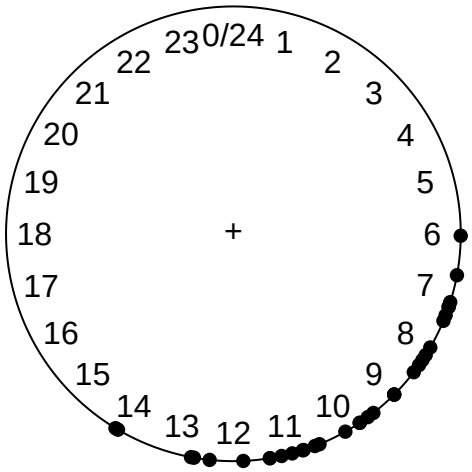
Problem 3

(a) Transform (or covert) the time of first pickup to an angle ranged from 0 to 360 degree

```
dat <- dat %>% mutate(Pickup.1st = strptime(Pickup.1st, format = "%H:%M"),
                      Pickup.1st.degree = (hour(Pickup.1st)*60 +
                                             minute(Pickup.1st))/(24*60)*360)
pickup.cir <- circular(dat$Pickup.1st.degree, units = "degrees",
                       template = "clock24")
```

(b) Scatterplot

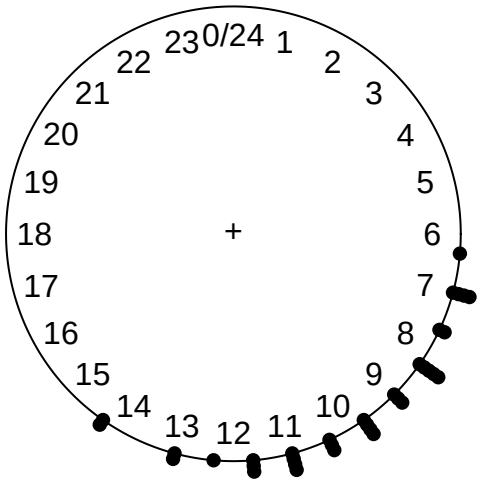
```
plot(pickup.cir)
```



- From the scatter plot we can see that most of the 1st pickup times happened between 7am and 11:30am, reflecting the wake up time of the participant. Some 1st pickup times happened at 6am and 12am to 2pm. The person may not have a pattern of waking up or differ due to the condition.

(c) histogram

```
plot(pickup.cir,stack = TRUE, bins = 36)
```



- We chose 36 as the bin size, indicating the 30 hour duration. We can see that the first pickups most frequently happened during 8:00-8:30AM.

Problem 4

(a) The inclusion of the factor S_t in the Poisson distribution accounts for the fact that the daily number of pickups Y_t is influenced by the total screen time S_t on that day. This modeling choice reflects the idea that the number of pickups is related to the amount of time people spend interacting with screens. S_t is like an offset variable, to scale the data and change the count variable to a rate.

```
attach(dat)
model1 <- glm(Pickups ~ offset(log(Total.ST.min)),family = "poisson")
summary(model1)
```

```
##
## Call:
## glm(formula = Pickups ~ offset(log(Total.ST.min)), family = "poisson")
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.91796    0.02026  -94.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1910.9  on 33  degrees of freedom
## Residual deviance: 1910.9  on 33  degrees of freedom
## AIC: 2113.3
##
## Number of Fisher Scoring iterations: 5
```

- The maximum likelihood estimate $\hat{\lambda} = exp(-1.91796) = 0.15$.
- On average, the participant picked 0.15 times per hour during the 34 days.

(c)

```
Z_t<-dat$Date>="2024-01-10"

model2<-glm(Pickups ~ offset(log(Total.ST.min))+ weekend + Z_t,family = "poisson")
summary(model2)
```

```
##
## Call:
## glm(formula = Pickups ~ offset(log(Total.ST.min)) + weekend +
##     Z_t, family = "poisson")
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.26109    0.03057 -41.254  < 2e-16 ***
## weekendTRUE -0.19514    0.04637  -4.208 2.58e-05 ***
## Z_tTRUE     -1.06663    0.04111 -25.944  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1910.9  on 33  degrees of freedom
## Residual deviance: 1238.3  on 31  degrees of freedom
## AIC: 1444.7
##
## Number of Fisher Scoring iterations: 5
```

- (c.1) P-value of the weekend variable is <0.001, we have evidence to reject the null hypothesis that there is no difference in behavior between weekdays and weekends. There is -0.19 decrease in the number of pickups for one-unit time on the weekends, comparing to weekdays.
- (c.2) P-value of the weekend variable is <0.001, we have evidence to conclude that there is difference in behavior between winter semester and holiday, where the average number of pickups decrease 1.07 for one-unit time on winter semester, comparing to holiday.

Problem 5

(a)

```r
c(min(pickup.cir),max(pickup.cir))
```

```
## [1]  90.50 211.25
```

```r
pickup.rad <- circular((pickup.cir)*pi/180-pi, units = "radians")
c(min(pickup.rad),max(pickup.rad))
```

```
## [1] -1.5620697  0.5454154
```

```r
estimate_rad <- mle.vonmises(pickup.rad)
estimate_rad_mu <- estimate_rad$mu
estimate_rad_lambda <- estimate_rad$kappa
estimate_rad
```

```
##
## Call:
## mle.vonmises(x = pickup.rad)
##
## mu: -0.5965  ( 0.09309 )
##
## kappa: 3.941  ( 0.8598 )
```

```r
(estimate_rad_mu+pi)*180/pi
```

```
## Circular Data:
## Type = angles
## Units = radians
## Template = none
## Modulo = asis
## Zero = 0
## Rotation = counter
## [1] 145.8253
```

- We transform y to the range of $[-\pi, \pi)$, which means transforming degrees to radians.
- $\hat{\mu} = -0.5965(\text{rad})$, equals to $(-0.5965 + \pi) * 180/\pi = 145.8253(\text{degree})$.
- $\hat{\lambda} = 3.941$

(b)

```r
degree <- (8*60+30)/(24*60)*360
radians <- (-degree * (pi/180))-pi
1- pvonmises(circular(radians),mu=estimate_rad_mu, kappa = estimate_rad_lambda)
```

```
## [1] 0.005009683
```

- The probability of first pickup time at 8:30 AM or later is 0.005.