Homework#1 Report Biostat620

Zihao Han

2024-02-04

GITHUB LINK:https://github.com/ZihaoHanGitHub/Biostat620Hw1

PART1: DATA COLLECTION AND DATA PROCESSING

Problem 1:

(a)

Describe the purpose of the data collection, in which you state a scientic hypothesis of interest to justify your effort of data collection. Cite at least one reference to support your proposed hypothesis to be investigated. This hypothesis may be the one of a few possible hypotheses that you like to investigate in your rst group project with your teammates.

This data collection we collect the total screen time, total social app used time, number of pickups, and the first pick up time, also the first pickup time could be considered as the sleep wake up time by adjusting the time zone of your mobile phone to avoid picking up your mobile phone in the early morning and interfering with the recording of sleep and wake time.

From these step, I'm interested in the scientific hypothesis that is the correlation between the total screen time and sleep wake up time (the first pickup times). In 2020, Pauliina Hiltunen et al. do the similar statistical inference among preschool children in Finland, they concluded that the used hours of smart phone and pad have a significant effect on the later bedtime, later wake-up time, less consistent sleep[2]. More than that, Alba do the similar research in 2018 among the adolescents (aged from 17 to 18), the higher tablet use found to be associated with reduced sleep efficiency and increased wake time after falling asleep [1].

Therefore, depending on the purpose and description of data collection, and the previous scientific literature, the hypothesis, the correlation between the total screen time and sleep wake up time (the first pickup times), would be a reasonable and meaningful hypothesis for this project.

(b)

Explain the role of Informed Consent Form in connection to the planned study and data collection

With the rapid development of science and technology, data plays a vital role in many enterprises and research fields. In most cases, the personal information of participants can be clearly restored through data. This is unethical in data theory if the Informed Consent Form is not signed. In addition, the Informed Consent Form can give a clearer understanding of the purpose of the research.

For example, this Informed Consent Form tells the participants what data will be collected, which researchers will have permission to use the data, and the signature must not The option to share with others protects the privacy of participants, ensures the rationality, security, and data ethics of this project, and reasonably

protects the legitimate rights and interests of participants. For researchers, the Informed Consent Form can regulate researchers' obligations and responsibilities.

Overall, the Informed Consent Form is essential for a project that relies on data. It can protect the legitimate rights and interests of participants, standardize the obligations and responsibilities of researchers, and ensure the orderly and legal progress of the project.

(c)

Describe the data collection plan, including when the data is collected, which types of variables in the data are collected, where the data is collected from, and how many data are collected before the data freeze. You may use tables to summarize your answers if necessary

In this project, we would collect the total screen time, total social app used time, number of pickup times, and the first pickup times these four variables from the participants' phone, their data type is text, text, numeric, and date corresponding, moreover, it would be recommended to collect some useful basic demographic variables, for example, whether it is the weekday.

In the example dataset provided by the instructor, we have to convert the screen time (text) and social app used time(text) to the number of minutes (numeric), it would be helpful when we do the modeling and analysis, the intuitive table is given by Table .

As mentioned by the instructor of this course, the participants are recommended collect the data since the the phone stored (generally, the phone will retain data within 30 days) till the end of this project. However, in this assignments, we are aimed to collect the data from the stored till the Jan/26/2024.

Variable Name	Type	Counts	Freeze Date (hw1)	Freeze Date
Total.ST	Text	34	Jan/26/2024	end of project
Total.ST.min	Numeric	34	Jan/26/2024	end of project
Social.ST	Text	34	Jan/26/2024	end of project
Social.ST.min	Numeric	34	Jan/26/2024	end of project
Pickups	Numeric	34	Jan/26/2024	end of project
Pickup.1st	Date	34	Jan/26/2024	end of project
Demongraphic	Text/Boolean		Jan/26/2024	end of project

Table 1: Example of the Data Type and Freezing date

(d)

Create and add two new variables into your dataset; they are, "daily proportion of social screen time" (defined as the ratio of daily total social screen time over daily total screen time) and "daily duration per use" (defined as the ratio of daily total screen time over daily total of pickups).

```
##       Date Total.ST Total.ST.min Social.ST Social.ST.min Pickups Pickup.1st
## 1 12/24/23    4h51m          291     2h11m           131      58      11:32
## 2 12/25/23    4h33m          273      1h2m            62      74      12:51
## 3 12/26/23    6h14m          374     1h26m            86      50      12:28
## 4 12/27/23   10h34m          634     2h27m           147      97      12:01
## 5 12/28/23    9h54m          594      2h1m           121      53      11:27
## 6 12/29/23   10h22m          622     2h16m           136      34      12:48
##   CourseToday dayOfWeek Vacation
## 1           0    Sunday        1
```

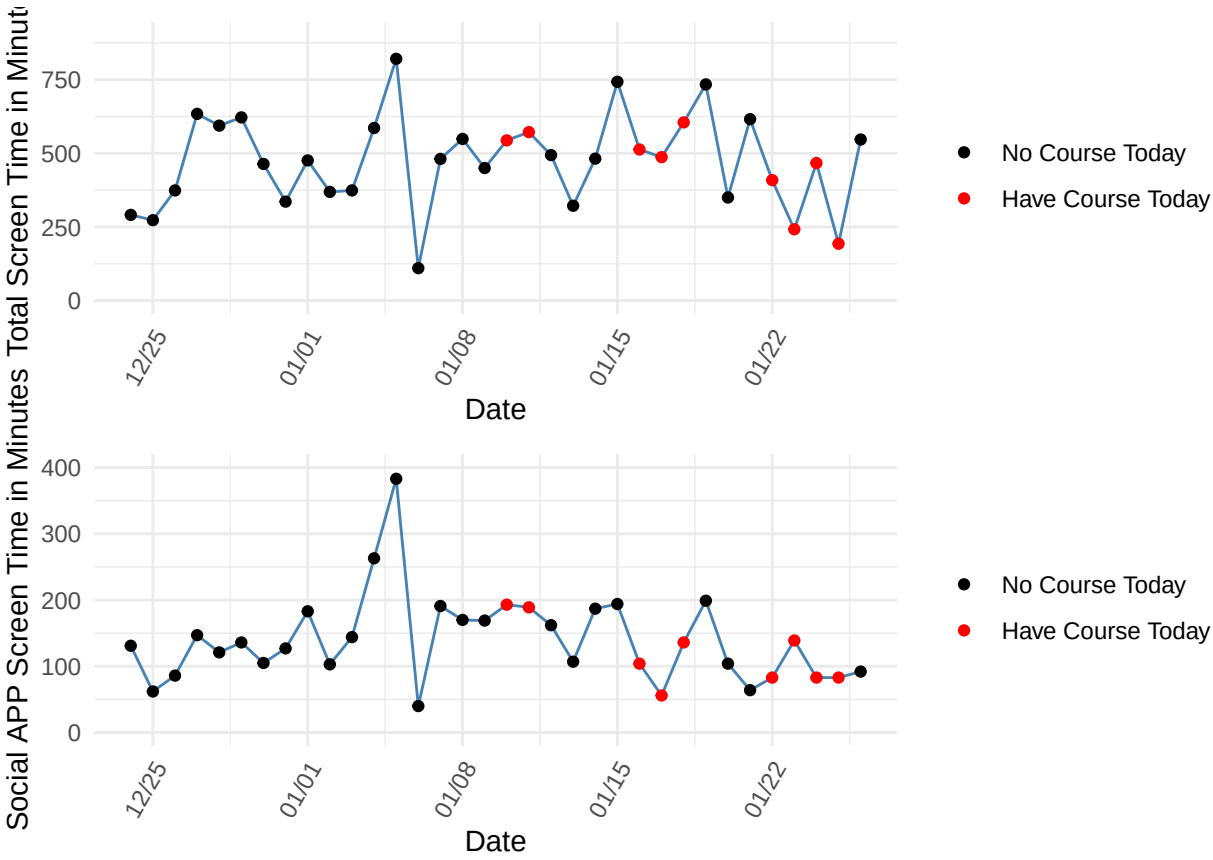
```
## 2           0    Monday        1
## 3           0   Tuesday        1
## 4           0 Wednesday        1
## 5           0  Thursday        1
## 6           0    Friday        1
```

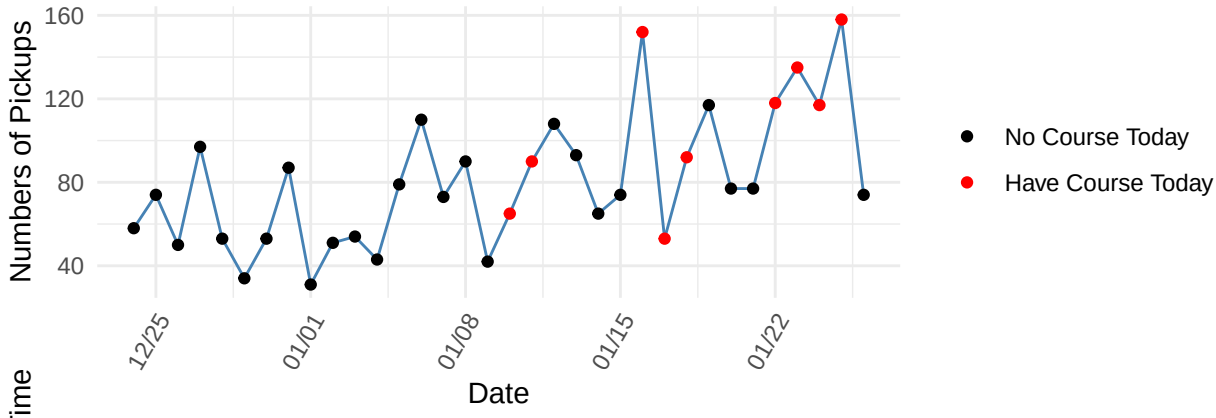
```
##       Date proportionOfSocial  duration
## 1 12/24/23          0.4501718  5.017241
## 2 12/25/23          0.2271062  3.689189
## 3 12/26/23          0.2299465  7.480000
## 4 12/27/23          0.2318612  6.536082
## 5 12/28/23          0.2037037 11.207547
## 6 12/29/23          0.2186495 18.294118
```

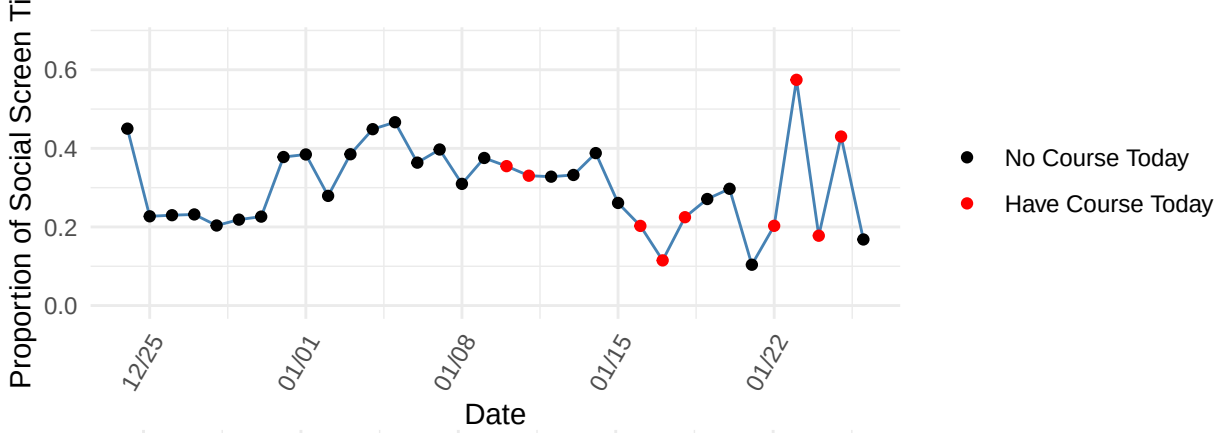
Problem 2

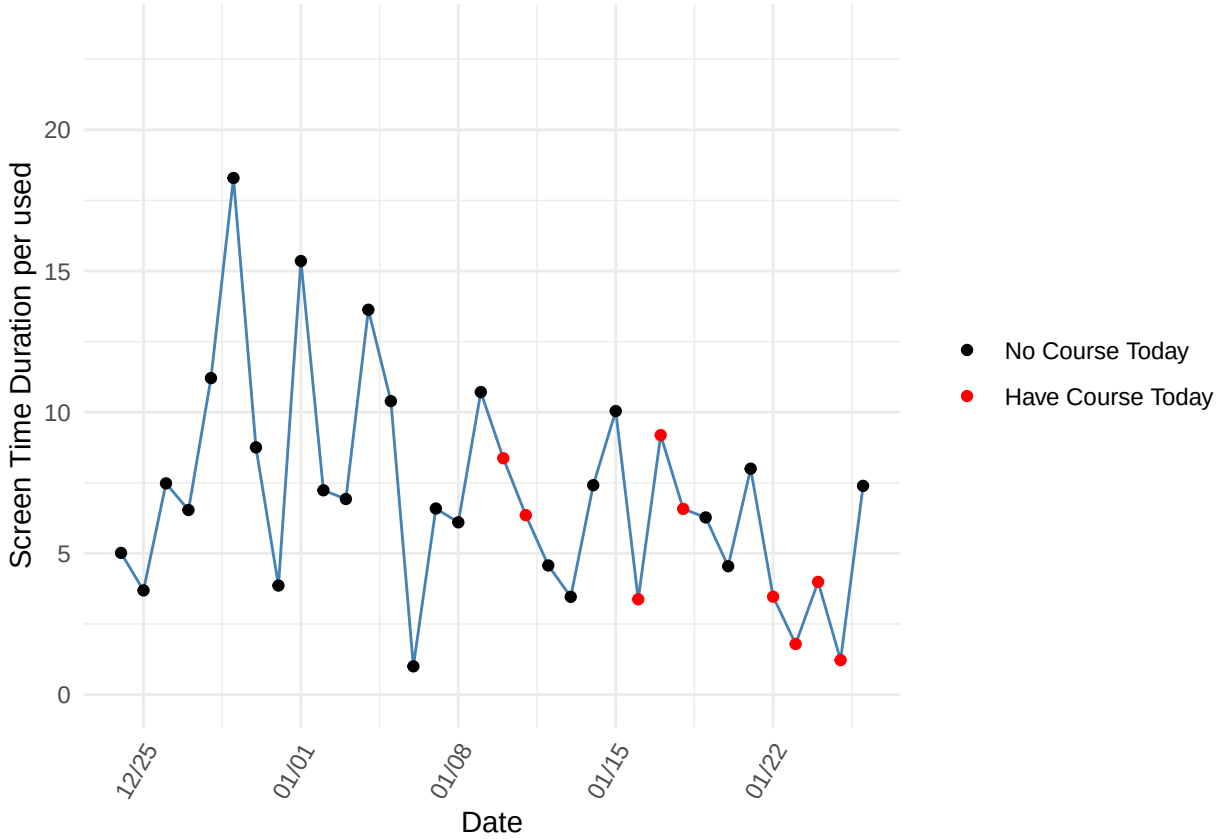
(a)

Make a time series plot of each of the five variables in your data. Describe temporal patterns from these time series plots.









According to these five time series plot, I defined red point as the day I have course, and black points

indicate that today I do not have course, usually the vocation or weekend, another additional information is my winter vocation is end at Jan 10, therefore, we could explore these temporal patterns.

In total Screen Time plot (Figure 1), from Jan 10, the semester begin, the total screen time is started to be stable gradually, especially, the total screen time between whether I have course is significant different.

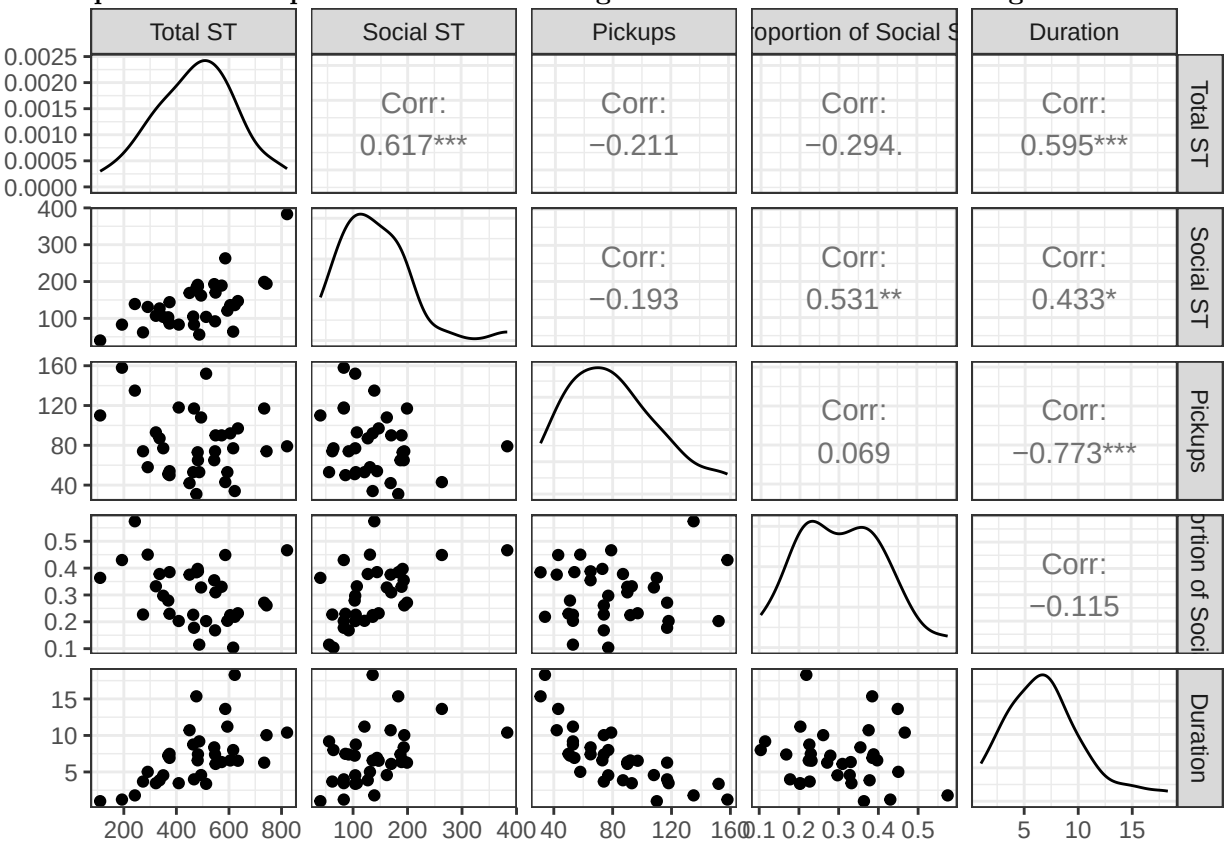
In the Social Screen Time plot (Figure 2), except the highest point (NOTE: that day I forget locked down my screen when I have a long-time phone call with my friend via Wechat, it causing my social time so long), the social screen time do not have intuitive trend shown in figure, but a step-conclusion is start from Jan 10, semester begin, the social screen time is significant different between course day and relax day, but it need more data to verify and support this step-conclusion.

In the number of pickups plot (Figure 3), it has a significant increasing trend by time, especially in the day I have course, combining with the plot of duration time per use (Figure 5), the the duration per pickups of course day is really shorter than the day I do not have course, more than that, it also has a significant decreasing trend than the day before the semester begin (Jan 10).

In the proportion of social app screen time plot (Figure 4), overall, it is stable. Except for 5 days among the 34 data points, it is higher than 0.4 and 4 days are lower than 0.2. The other data points are concentrated between 0.2 and 0.4. For the time being, no obvious demographic relationship has been found. information, this staged conclusion requires more data to support.

(b)

Make pairwise scatterplots of five variables. Describe correlation patterns from these pairwise scatterplots. Which pair of variables among the five variables has the highest correlation?



From this pairwise correlation plot, we could find that there are five pair variables show a significant correlation, they are Social Screen Time vs. Total Screen Time ($corr = 0.617 * **$), Duration vs. Total

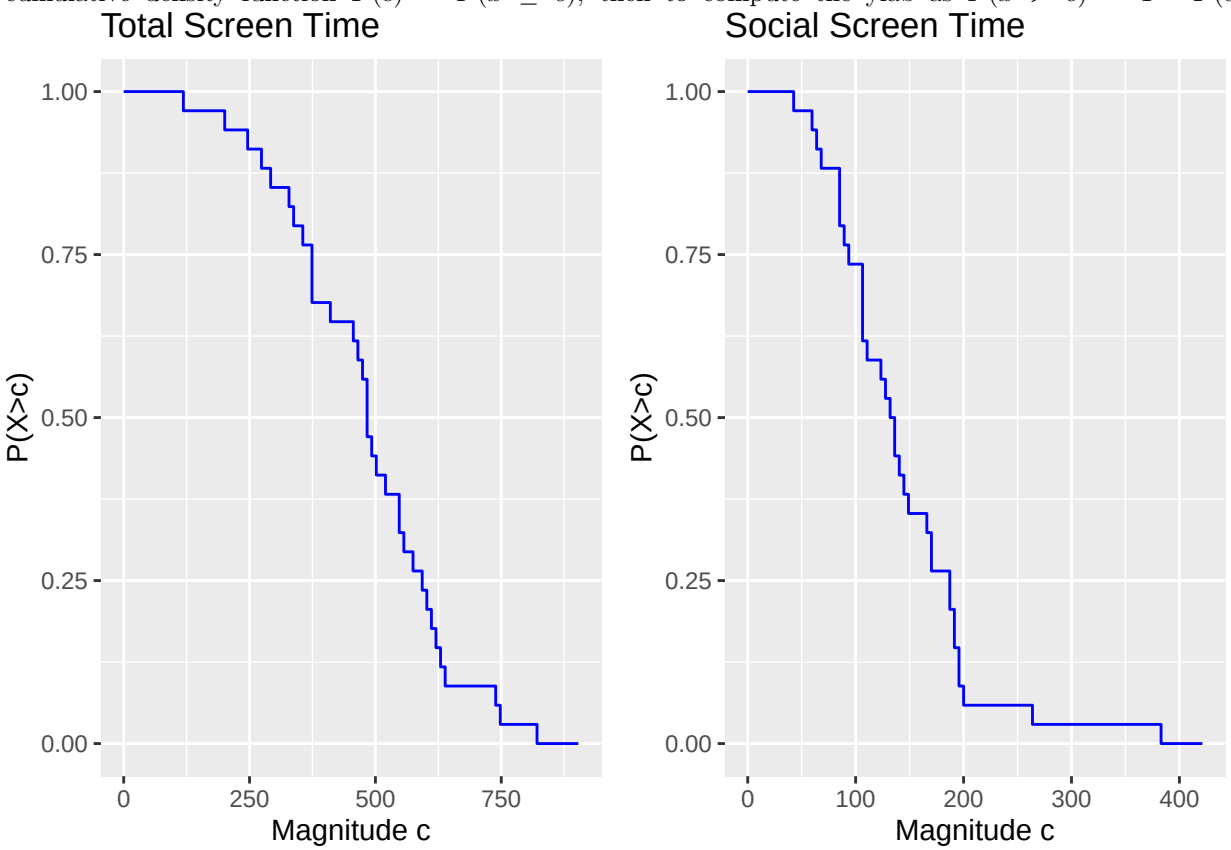
Screen Time ($corr = 0.595 * **$), Proportion of Social Screen Time vs. Social Screen Time ($0.531 * *$), Duration vs. Social Screen Time ($corr = 0.433*$), Duration vs. Pickups ($corr = -0.773 * **$).

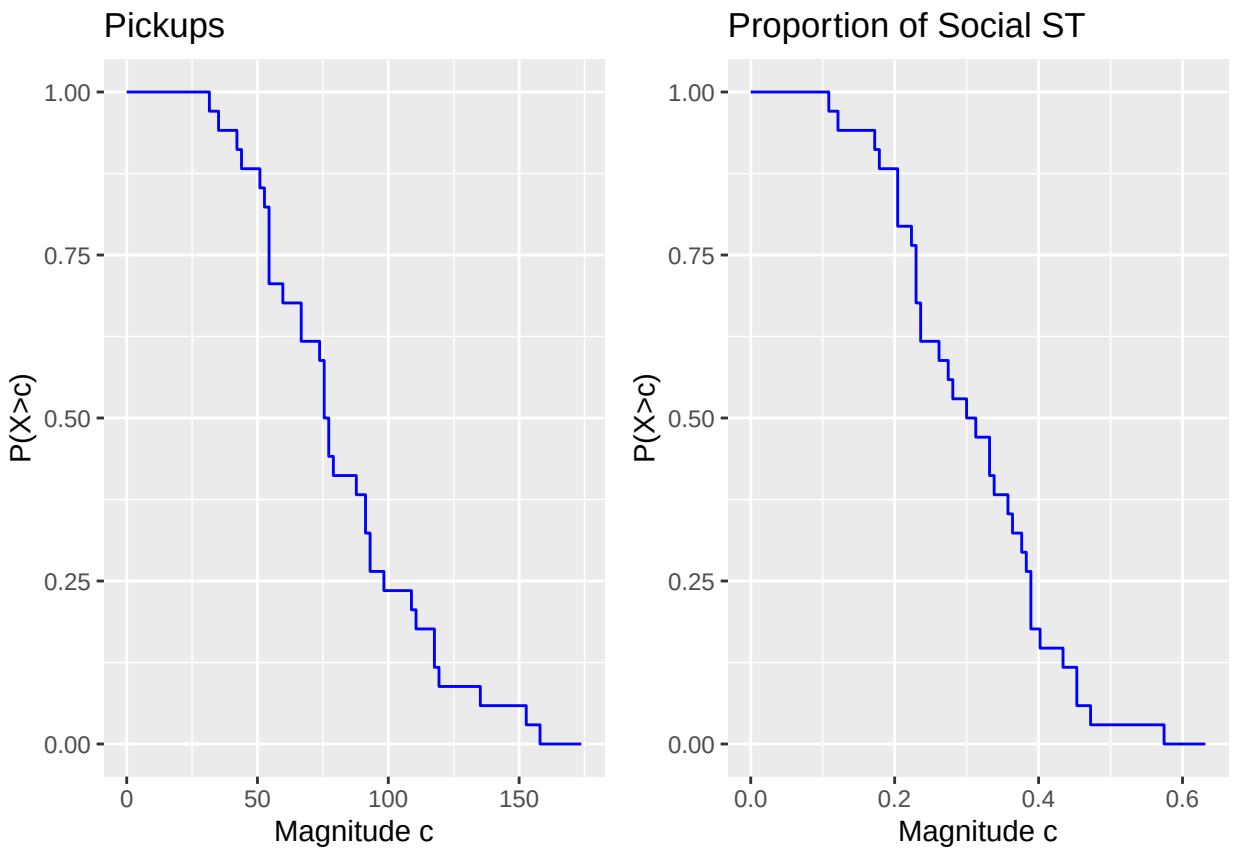
To explain this, when I have more screen time leads to have a higher social screen time during the screen time significantly (Total Screen Time contain the social screen time), however we could find there is a nonsignificant and negative correlation between proportion of social ST vs. Total ST ($corr = -0.294$), meaning that, When my total screen time increases, although the social screen time increases, according to the correlation between proportion social ST and total ST, it is known that the social screen time does not increase proportionally, but increases compared to the original proportion. The value is smaller than the value corresponding to this ratio. More generally, as my overall screen time increases, I may be more inclined to use other apps (other than social apps).

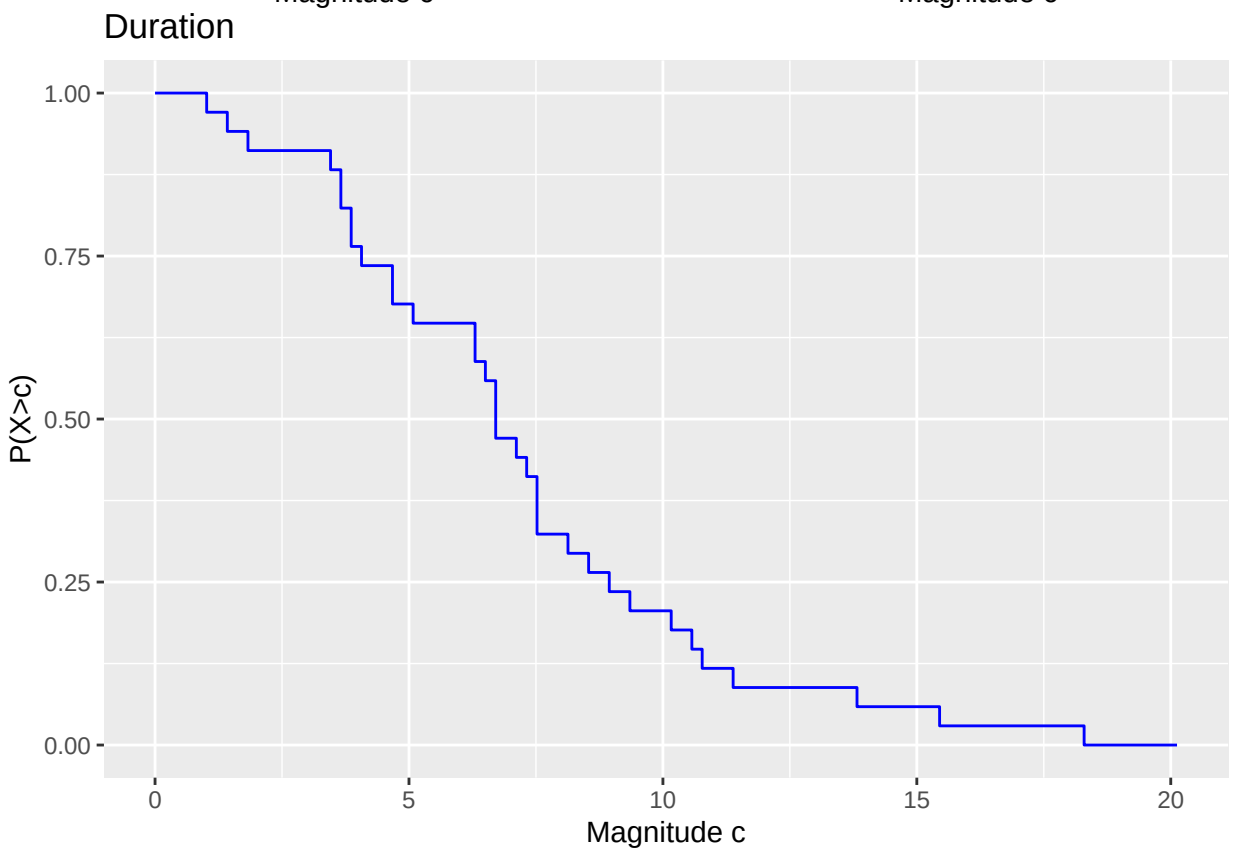
For Duration vs. Total Screen Time ($corr = 0.595 * **$) Duration vs. Pickups ($corr = -0.773 * **$), it shows that there is a significant positive relationship between total ST and Duration per use, and a significant negative relationship between Duration per use and pickups. Since the correlation between pickups vs. total ST ($corr = -0.211$) is not significant, it could be assumed to be independent, therefore, when pickups times increase it would leads to a shorter duration adjusting for total ST; when the total ST increase, it would leads to a longer duration, adjusting for pickups.

(c)

Make an occupation time curve for each of the five time series. Explain the pattern of individual curves. Since in the occupation time curve, the y-lab is shown by the $P(x > c)$, it could be rewrite as the $P(x > c) = 1 - P(x \leq c)$, and we could using the R to compute the cumulative density function $F(c) = P(x \leq c)$, then to compute the ylab as $P(x > c) = 1 - F(c)$.





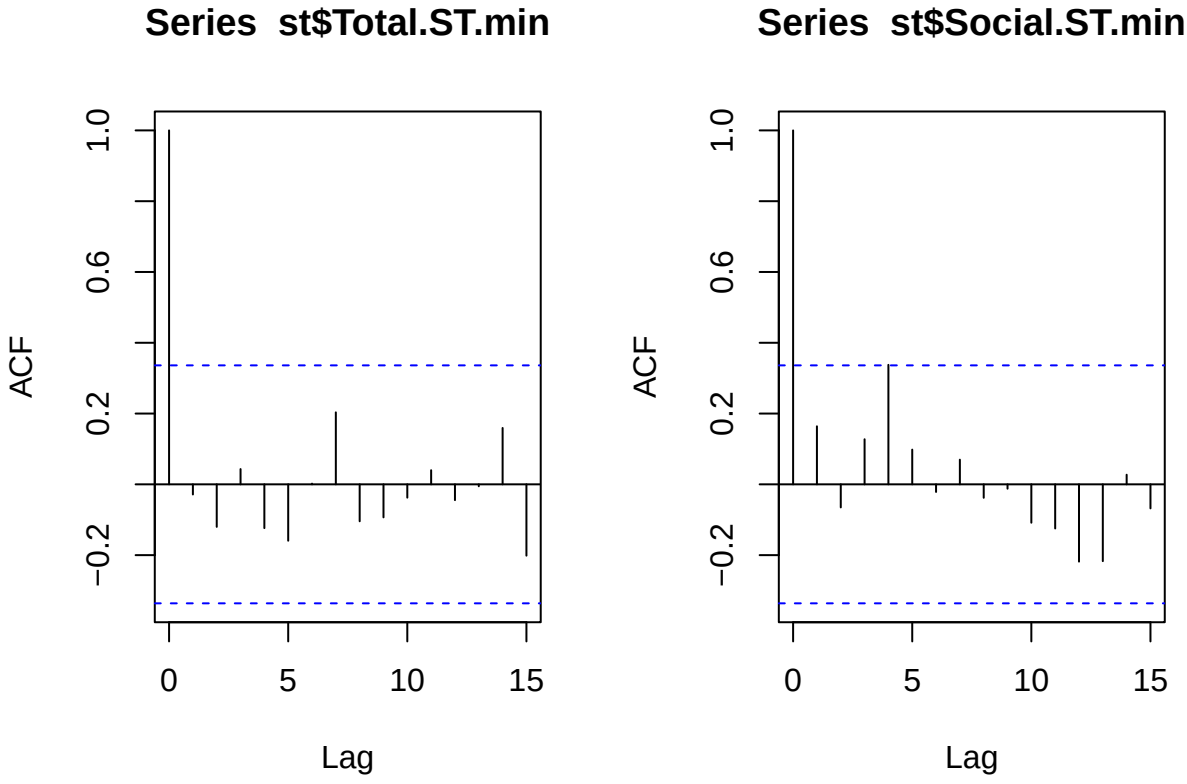


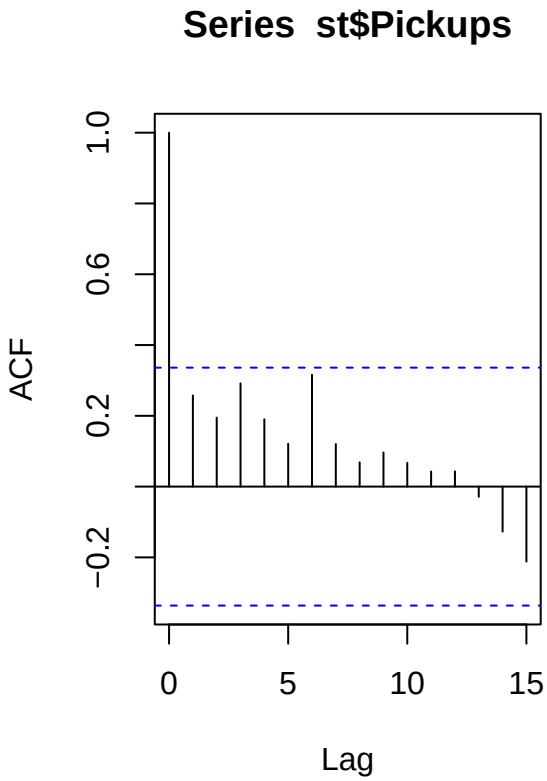
From the Occupation-Time plot of these five variable generated by the cumulative density function (CDF),

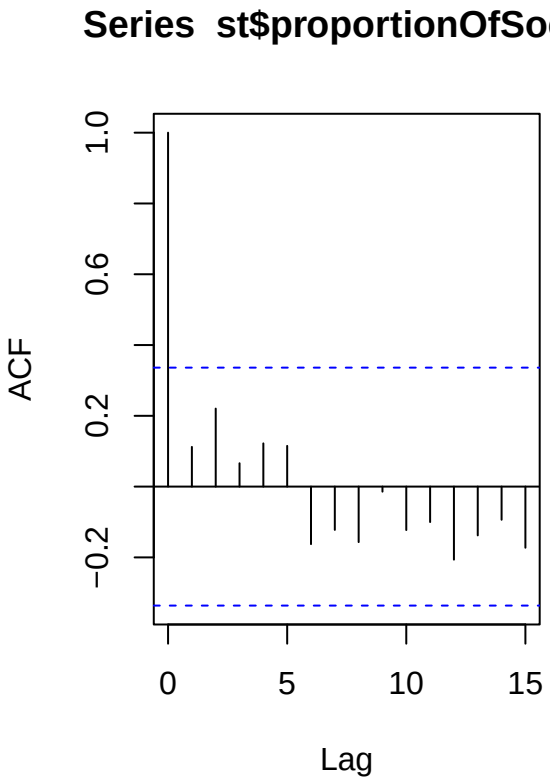
we could see this curve decrese slowly at the edge of the range, and faster at the middle of these five variables. One has a extremly difference is the figure 2, at the magnitude $c = 200$, it has already near to the 0, it is the social screen time occupation time curve, so it mean that majority of the social ST is below the 200, there are few days social ST beyond the magnitude 200.

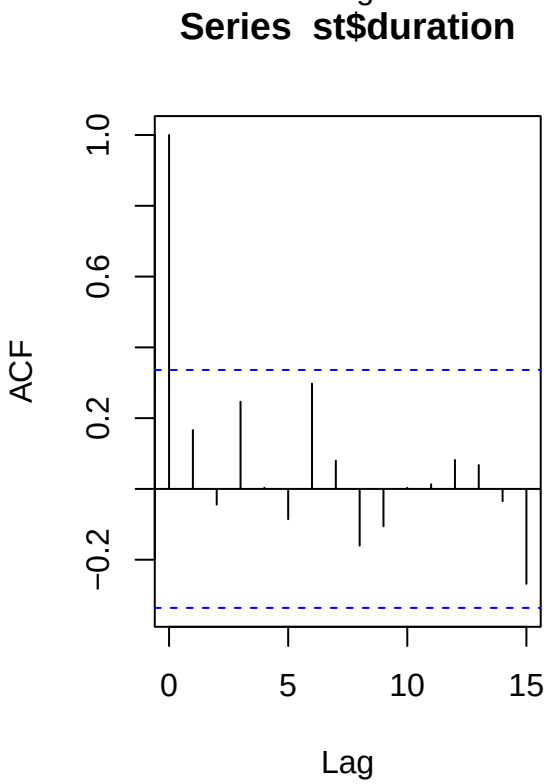
(d)

Use the R function acf to display the serial dependence for each of the five time series. Are there any significant autocorrelations? Explain your results. Note that in this R function, you may set plot=FALSE to yield values of the autocorrelations.










```
##
## Autocorrelations of series 'st$Total.ST.min', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
```

```
##  1.000 -0.029 -0.120  0.043 -0.124 -0.159  0.002  0.203 -0.104 -0.093 -0.038
##     11     12     13     14     15
##  0.040 -0.045 -0.005  0.159 -0.202

##
## Autocorrelations of series 'st$Social.ST.min', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.164 -0.065  0.127  0.338  0.098 -0.022  0.070 -0.038 -0.012 -0.108
##     11     12     13     14     15
## -0.125 -0.218 -0.217  0.027 -0.068

##
## Autocorrelations of series 'st$Pickups', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.258  0.195  0.292  0.190  0.121  0.316  0.120  0.069  0.096  0.067
##     11     12     13     14     15
##  0.043  0.043 -0.029 -0.127 -0.212

##
## Autocorrelations of series 'st$proportionOfSocial', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.112  0.220  0.066  0.122  0.115 -0.163 -0.122 -0.157 -0.014 -0.123
##     11     12     13     14     15
## -0.100 -0.206 -0.138 -0.094 -0.173

##
## Autocorrelations of series 'st$duration', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.166 -0.044  0.246  0.003 -0.085  0.297  0.080 -0.160 -0.106  0.003
##     11     12     13     14     15
##  0.013  0.082  0.068 -0.034 -0.268
```

From these five auto-correlation plot and the value computed by auto-correlation function (ACF), we could conclude that, these five variable (Total Screen Time, Social Screen Time, pickups, proportion of social Screen time, and duration) all shows nonsignificant autocorrelation over lags, since the vertical bars in the plot (except first 1, highest) all containing in the 95 % confident interval which is showed in plot as two horizontal lines. Conclude that these five variables are recorded independents on the records of passed days.

Problem 3

(a)

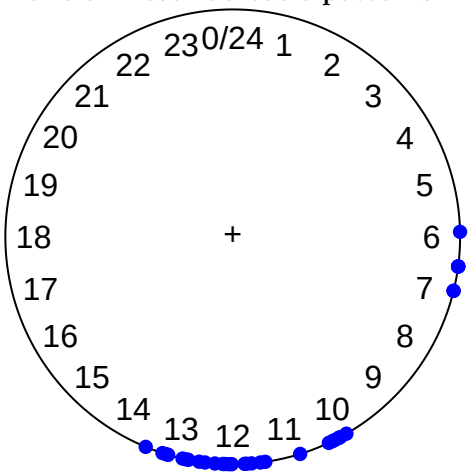
Transform (or covert) the time of first pickup to an angle ranged from 0 to 360 degree, treating midnight as 0 degree. For example, 6AM is 90 degree and noon is 180 degree.

```
##         Date Pickup.1st.angle
## 1 2023-12-24           173.00
## 2 2023-12-25           192.75
```

```
## 3 2023-12-26           187.00
## 4 2023-12-27           180.25
## 5 2023-12-28           171.75
## 6 2023-12-29           192.00
```

(b)

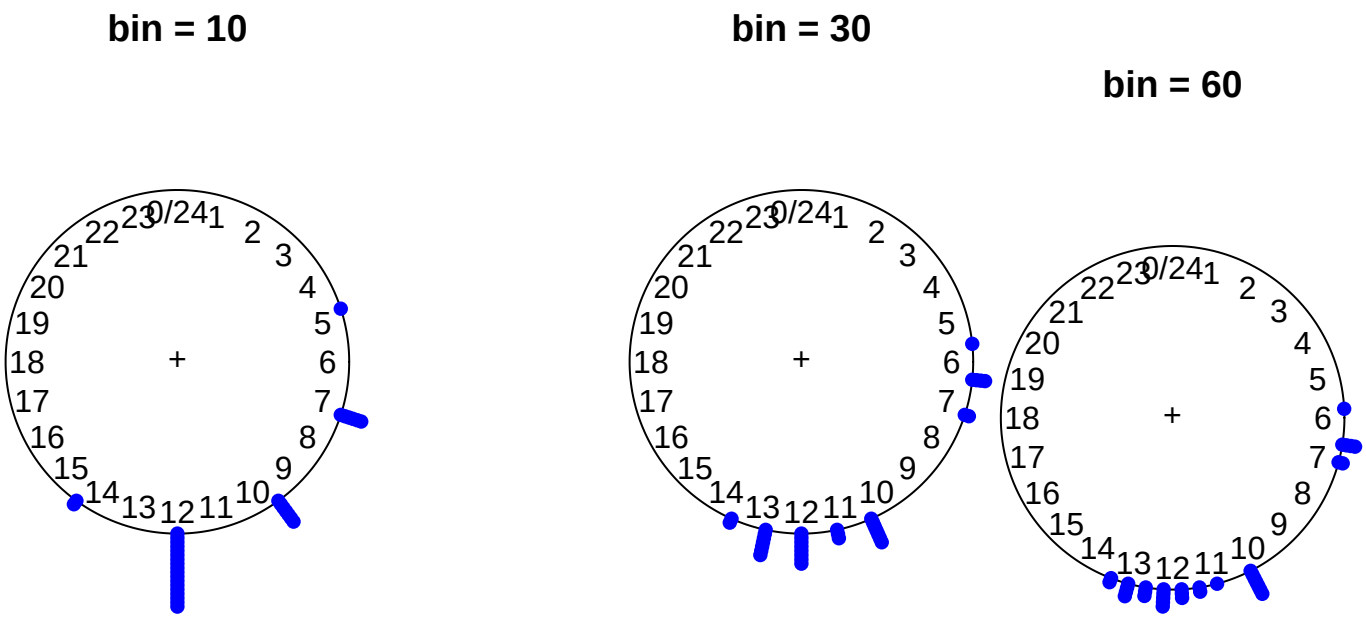
Make a scatterplot of the first pickup data on a 24-hour clock circle. Describe basic patterns from this scatterplot in terms of personal habit of first pickup.



From the basic scatter plot in a circle of clock 24 hours, we could find a simple distribution of first pick up times (could be considered as the wake up time), main distributed in 6:00 AM - 7:00 AM, and 10:00 AM to 1:30 PM, since we cannot see the histogram of this circle plot, according to the dataset recorded the data of during winter vocation and winter semester, the wake up time maybe could split by whether that day is the vocation.

(c)

Make a histogram plot on the circle in that you may choose a suitable bin size to create stacking. For example, you may set a bin size at 2.5 degree, which corresponds an interval of 10 minutes. Adjust the bin size to create different forms of histogram, and explain the reason that you choose a particular value to report your final histogram plot.



Finally, I choose the bin size of 60, corresponding to the time interval 240 mins, in this setting, Such a histogram is more intuitive and reduces information loss. For example, in images where the bin size is set to 10 and 30, many data points are all recorded in one column of the histogram, resulting in serious information loss; while setting the bin size to In the case of 90, the image becomes less intuitive. For example, at the time point of 10 o'clock, when the bin size is set to 60, it can be clearly seen that there are many data points that get up around 10 AM, and the bin size is This can be found less intuitively in 90 cases. So, in the end my bin size was set to 60.

PART II: DATA ANALYSIS

Problem 4

(a)

Explain why the factor S_t is needed in the Poisson distribution above.

Since

$$Y_t \sim \text{Poisson}\left(S_t\lambda\right), t = 1, \ldots, T$$

and Y_t defined as the daily number of pickups at day t, S_t defined as the total screen time, λ defined as the hourly rate of pickups. We know that the mean of Poisson distribution is the $E[Y|S_t\lambda] = S_t\lambda$, all so in definition, $Y_t = S_t * \lambda_t$, λ_t defined as the hourly rate of pickups at day t, the total screen time is important here to contribute the distribution fitness to the Y_t.

(b)

Use the R function glm to estimate the rate parameter λ in which ln(St) is included in the model as an offset.

Since we estimate the parameter λ and it is defined as the **hourly** rate pickups, therefore, we need to convert the S_t from minutes to hourly.

```
##
```

```
## Call:
## glm(formula = Pickups ~ offset(log(Total.ST.min/60)), family = poisson,
##     data = st)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -8.153  -3.089  -1.084   3.369  15.686
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.32345    0.01909   121.7   <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: 1040.3  on 33  degrees of freedom
## Residual deviance: 1040.3  on 33  degrees of freedom
## AIC: 1251.5
##
## Number of Fisher Scoring iterations: 5
```

Since we used the poisson regression model, and the link function is log-link function, the model could be described as:

$$\ln(\text{ Pickups }) = 2.32345 + \ln(\text{ Total.ST.min }/60)$$

the estimate of $\hat{\lambda}$ would be$\hat{\lambda} = exp(\frac{ln(Pickups)}{ln(Total.ST.min/60)})$

Therefore, the estimate of λ is given by:

```
## The estimate is 10.21087
```

(c)

Define two dummy variables: $X_t = 1$ for day t being a weekday and 0 for day t being a weekend day; and $Z_t = 1$ for day t being January, 10 (the first day of the winter semester) or after, and 0 for day t before January, 10 (the winter holiday day).

Repeat part (b) for a model $ln(\lambda) = \beta_0 + \beta_1 * X_t + \beta_2 * Z_t$, under which the rate parameter λ differs between weekdays and weekends as well as between the winter semester and the winter holiday. This model is called log-linear model. Cleary, this rate parameter depends on day t. Use the R function glm to estimate the regression coefficients and answer the following questions.

After create these two dummy variables, the summary of model generated by glm is given by:

```
##
## Call:
## glm(formula = Pickups ~ Xt + Zt + offset(log(Total.ST.min/60)),
##     family = poisson, data = st)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -6.2546  -3.0878  -0.6344   1.9851  14.7818
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.28175    0.04472  51.023  < 2e-16 ***
## Xt          -0.21690    0.04394  -4.936 7.96e-07 ***
## Zt           0.37041    0.03908   9.478  < 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: 1040.32  on 33  degrees of freedom
## Residual deviance:  925.43  on 31  degrees of freedom
## AIC: 1140.6
##
## Number of Fisher Scoring iterations: 5
```

(c1)

Is there data evidence for significantly different behavior of daily pickups between weekdays and weekends? Justify your answer using the significance level $\alpha = 0.05$.

Since the P-value of regression coefficients of X_t is $Pr(> |z|) = 7.96e-07 < 0.05$, there is a significant correlation between the weekdays and weekends on number of daily pickups.

(c2)

Is there data evidence for a significant change on the behavior of daily pickups after the winter semester began? Justify your answer using the significance level $\alpha = 0.05$.

Since the P-value of regression coefficients of Z_t is $Pr(> |z|) < 2e-16 < 0.05$, there is a significant correlation between the winter vocation and semester begin on number of daily pickups.

Problem 5

(a)

Use the R function mle.vonmises from the R package circular to obtain the estimates of the two model parameters μ and λ from your data of first pickups.

Using the MLE vonmises to estimate the two parameter of μ and λ needs to input a numerical value to compute, thus, the angle degree we calculated before would be suitable for this estimator.

```
##
## Call:
## mle.vonmises(x = st$Pickup.1st.angle)
##
## mu: 2.786  ( 0.6064 )
##
## kappa: 0.404  ( 0.25 )
```

Therefore, the estimate of $\hat{\mu} = 2.786$ with standard error equal to 0.6064, the estimate of $\hat{\lambda} = 0.404$, with $se = 0.25$

(b)

Based on the estimated parameters from part (a), use the R function pvonmises from the Rpackage circular to calculate the probability that your first pickup is 8:30AM or later.

Since we use the function of pvonmises, it give us the probability of $P(X < a)$, and we need the probability of later than 8:30 AM, thus, we need one more step of calculating, $P(X >= a) = 1 - P(X < a)$

```
## The probability of first pickup time later than 8:30 AM with 1e-4 tolerence
## , and the parameters are estimated by (a), is 0.6256397
```

References

[1] Alba Cabré-Riera, Maties Torrent, David Donaire-Gonzalez, Martine Vrijheid, Elisabeth Cardis, and Mònica Guxens. Telecommunication devices use, screen time and sleep in adolescents. *Environmental research*, 171:341–347, 2019.

[2] Pauliina Hiltunen, Marja H Leppänen, Carola Ray, Suvi Määttä, Henna Vepsäläinen, Leena Koivusilta, Nina Sajaniemi, Maijaliisa Erkkola, and Eva Roos. Relationship between screen time and sleep among finnish preschool children: results from the dagis study. *Sleep Medicine*, 77:75–81, 2021.