620_hw1

project

2024-02-02

PATT I: DATA COLLECTION AND DATA PROCESSING

a. Describe the purpose of the data collection, in which you state a scientific 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 first group project with your teammates.

solution: The purpose of our data collection is to investigate the hypothesis that excessive mobile screen usage is associated with an increased risk of obesity among individuals[1]. This hypothesis is grounded in the premise that higher screen time can lead to sedentary behavior, which in turn may contribute to obesity due to reduced physical activity and possible changes in eating habits during screen engagement.

Reference: 1.Domoff, Sarah E., et al. "Excessive use of mobile devices and children's physical health." Human Behavior and Emerging Technologies 1.2 (2019): 169-175.

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

solution: The Informed Consent Form plays a critical role in the planned study by ensuring that all participants are fully aware of the study's purpose, procedures, potential risks. It serves as a means to obtain participants' voluntary agreement to partake in the research, safeguarding their rights.

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.

solution: When Data is Collected: The data collection period spans 11 days, from January 16, 2023, to January 26, 2023. Data entries are recorded daily.

Types of Variables Collected: 1. Total Screen Time (Total.ST): Recorded in both HH-MM format and MM format to capture the total amount of time spent on the mobile device each day. Social App Screen Time (Social.ST): Recorded in both HH-MM format and MM format, this variable specifically tracks the time spent on social media applications daily. Total Number of Phone Pickups (Pickups): Captures how many times the participant picks up the phone throughout the day. Time of First Pickup (Pickup.1st): Indicates the first time the participant picks up the device after waking up, marking the start of the day's device usage.

Where Data is Collected From: Data is collected directly from my mobile device using apple's tracking of screen activity.

Amount of Data Collected Before Data Freeze: Daily entries across the 11-day period are collected.

```
##
## Attaching package: 'lubridate'

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

##
## 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
```

```
## # A tibble: 13 x 11
##    Date                Total.ST Total.ST.min Total.ST.min.true Total.ST.match
##    <dttm>              <chr>           <dbl>             <dbl> <lgl>
##  1 2023-01-14 00:00:00 11h1m             194               661 FALSE
##  2 2023-01-15 00:00:00 9h10m             195               550 FALSE
##  3 2023-01-16 00:00:00 8h41m             196               521 FALSE
##  4 2023-01-17 00:00:00 8h39m             252               519 FALSE
##  5 2023-01-18 00:00:00 8h19m             242               499 FALSE
##  6 2023-01-19 00:00:00 12h24m            216               744 FALSE
##  7 2023-01-20 00:00:00 11h24m            379               684 FALSE
##  8 2023-01-21 00:00:00 5h41m             194               341 FALSE
##  9 2023-01-22 00:00:00 5h41m             207               341 FALSE
## 10 2023-01-23 00:00:00 8h58m             176               538 FALSE
## 11 2023-01-24 00:00:00 8h43m             186               523 FALSE
## 12 2023-01-25 00:00:00 6h5m              310               365 FALSE
## 13 2023-01-26 00:00:00 9h17m             218               557 FALSE
## # i 6 more variables: Social.ST <chr>, Social.ST.min <dbl>,
## #   Social.ST.min.true <dbl>, Social.ST.match <lgl>, Pickups <dbl>,
## #   Pickup.1st <dttm>
```

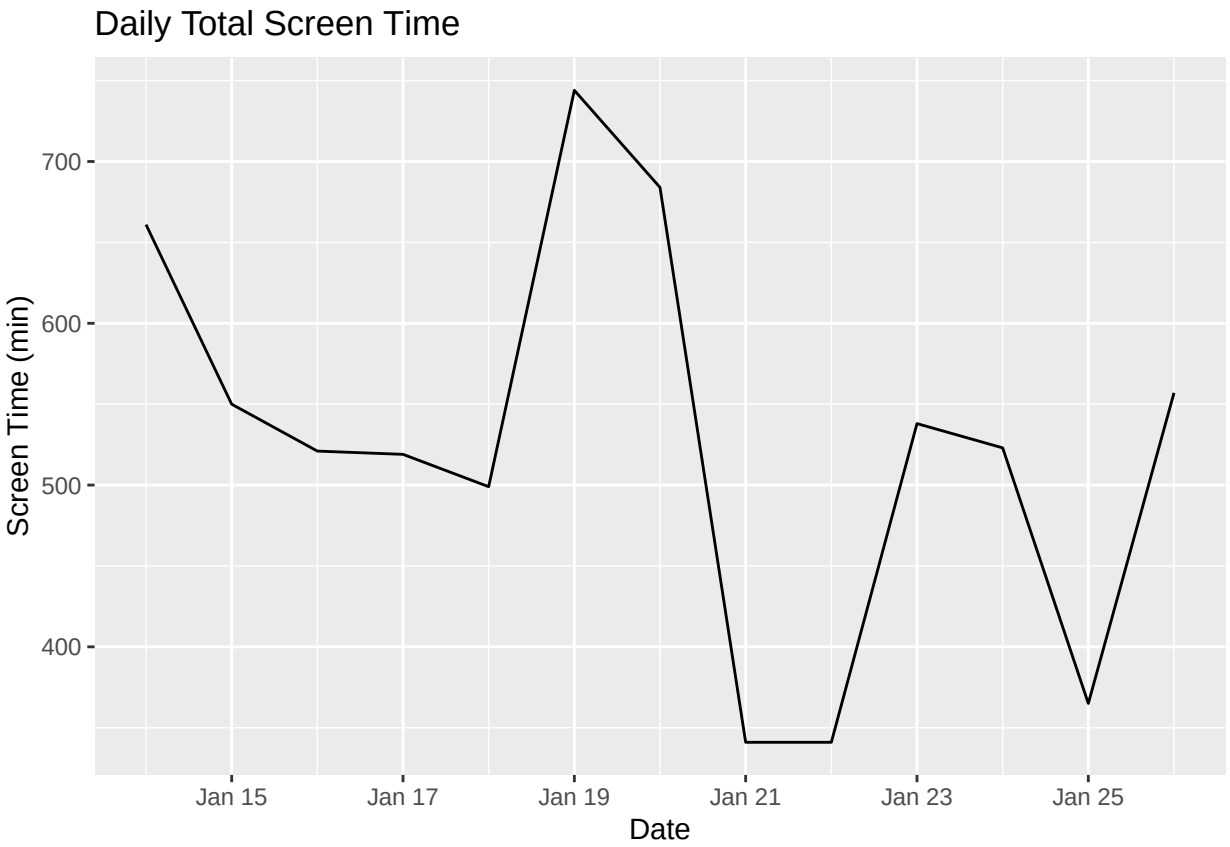
d. Create and add two new variables into your dataset; they are, "daily proportion of social screen time" (defined as the ratio of daily total social screen time over daily total screen time) and "daily duration per use" (defined as the ratio of daily total screen time over daily total of pickups)

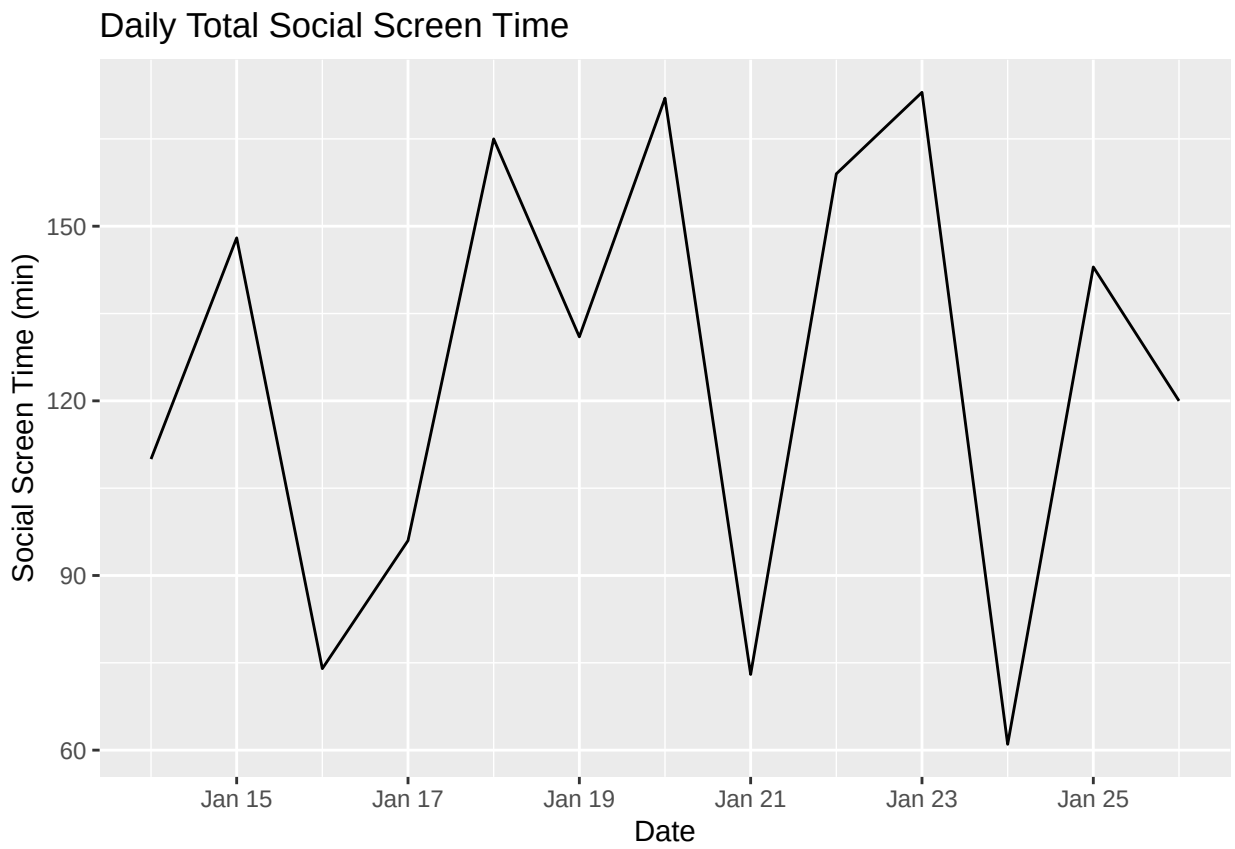
```
##  [1] "Date"                  "Total.ST"              "Total.ST.min"
##  [4] "Total.ST.min.true"     "Total.ST.match"        "Social.ST"
##  [7] "Social.ST.min"         "Social.ST.min.true"    "Social.ST.match"
## [10] "Pickups"               "Pickup.1st"            "Daily_Prop_Social_ST"
## [13] "Daily_Duration_Per_Use"
```

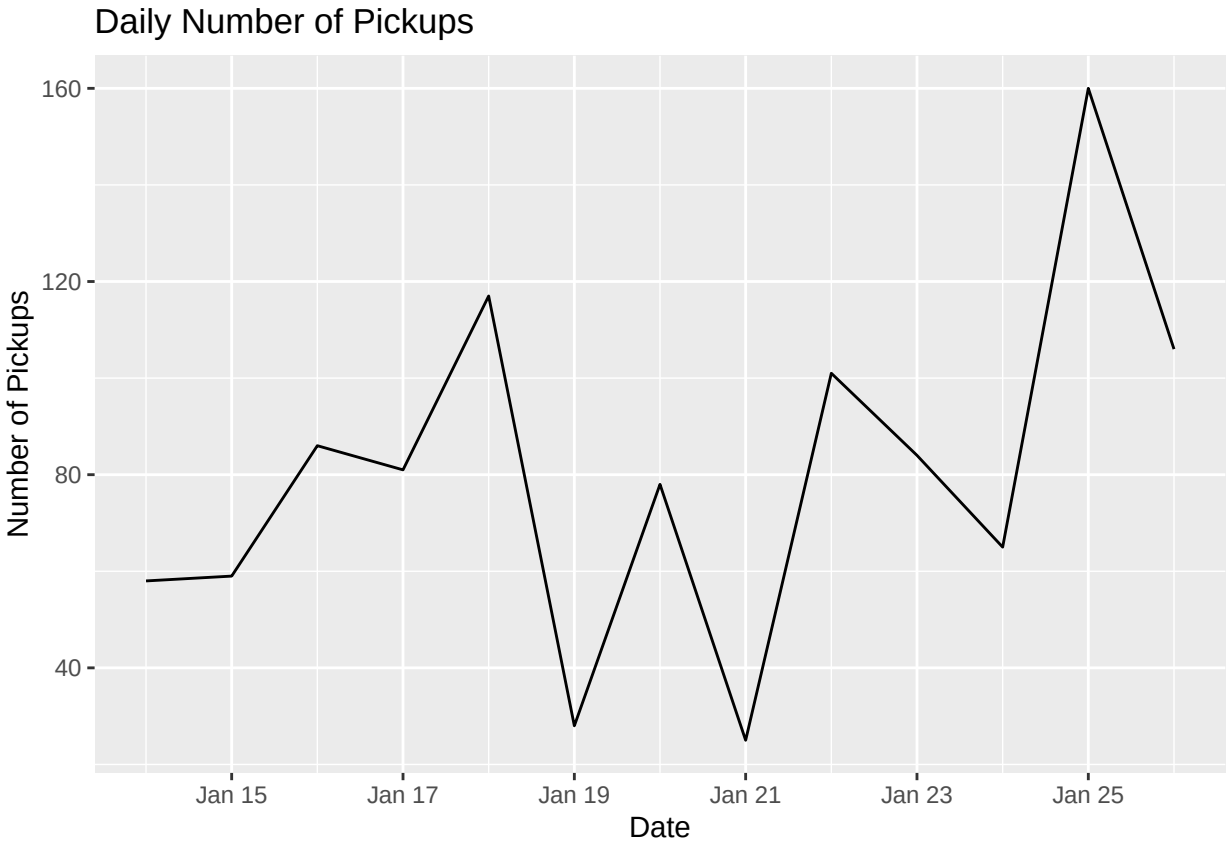
Problem 2: Data visualization is one of the early steps taken to see the data at hand. Consider the variables measured in the screen activity data, including daily total screen time, daily total social screen time, and daily number of pickups as well as two new variables derived from the raw data, daily proportion of social screen time and daily duration per use.

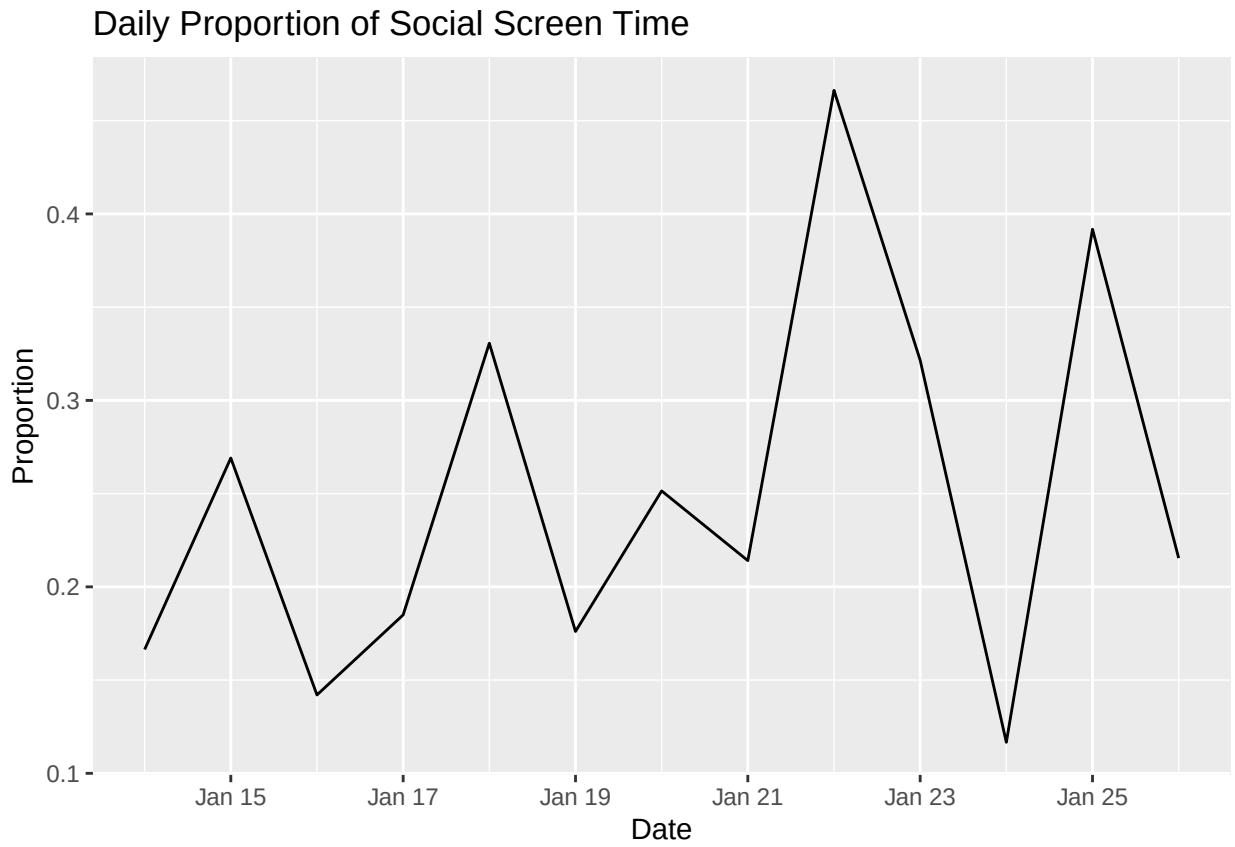
a. Make a time series plot of each of the five variables in your data. Describe temporal patterns from these time series plots.

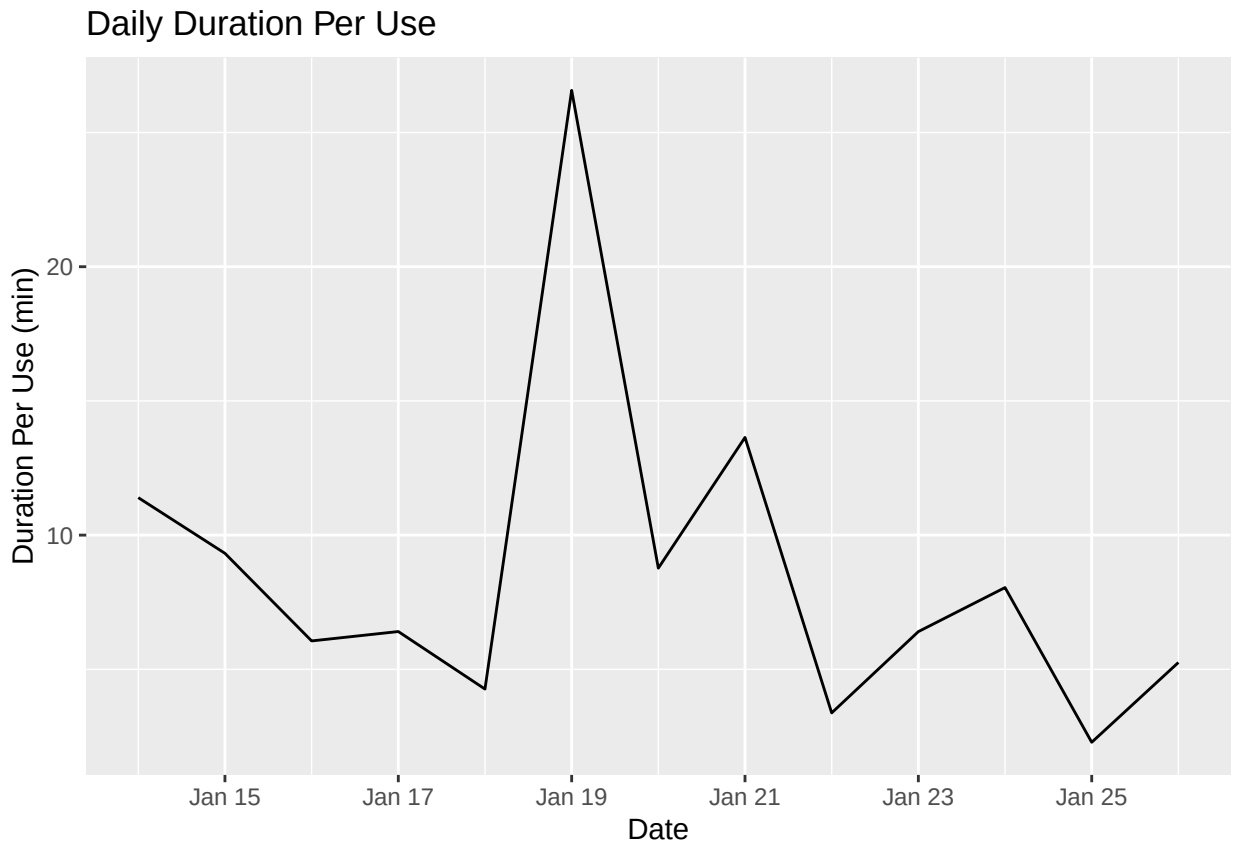
```
## [1] "C"
```


Daily Total Screen Time
Screen Time (min)
700
600
500
400
Jan 15
Jan 17
Jan 19
Jan 21
Jan 23
Jan 25
Date


Daily Total Social Screen Time
Social Screen Time (min)
150
120
90
60
Jan 15
Jan 17
Jan 19
Jan 21
Jan 23
Jan 25
Date


Daily Number of Pickups
Number of Pickups
160
120
80
40
Jan 15
Jan 17
Jan 19
Jan 21
Jan 23
Jan 25
Date


Daily Proportion of Social Screen Time
Proportion
0.4
0.3
0.2
0.1
Jan 15
Jan 17
Jan 19
Jan 21
Jan 23
Jan 25
Date



1. Daily Total Screen Time: There appears to be a sharp increase in screen time peaking around the middle of the time period, followed by a sharp decrease.

2.Daily Total Social Screen Time: There is a noticeable peak that suggests a day with significantly higher social media use compared to other days.

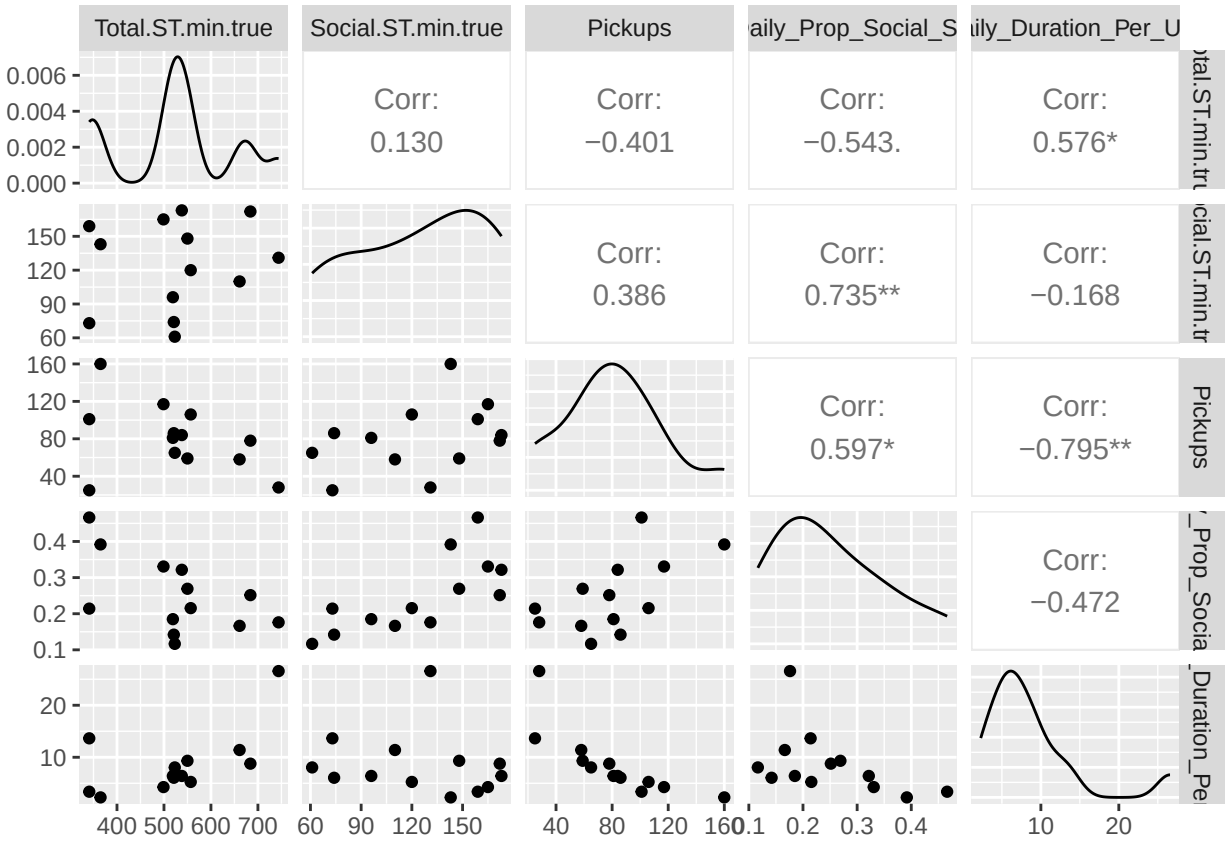
3.Daily Number of Pickups: There is a general trend where the number of pickups increases towards the end of the period.

4.Daily Proportion of Social Screen Time: The proportion of social screen time relative to total screen time shows a few significant peaks.

5.Daily Duration Per Use: The plot indicates that there are days with longer average usage per pickup and other days with shorter durations.

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?

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

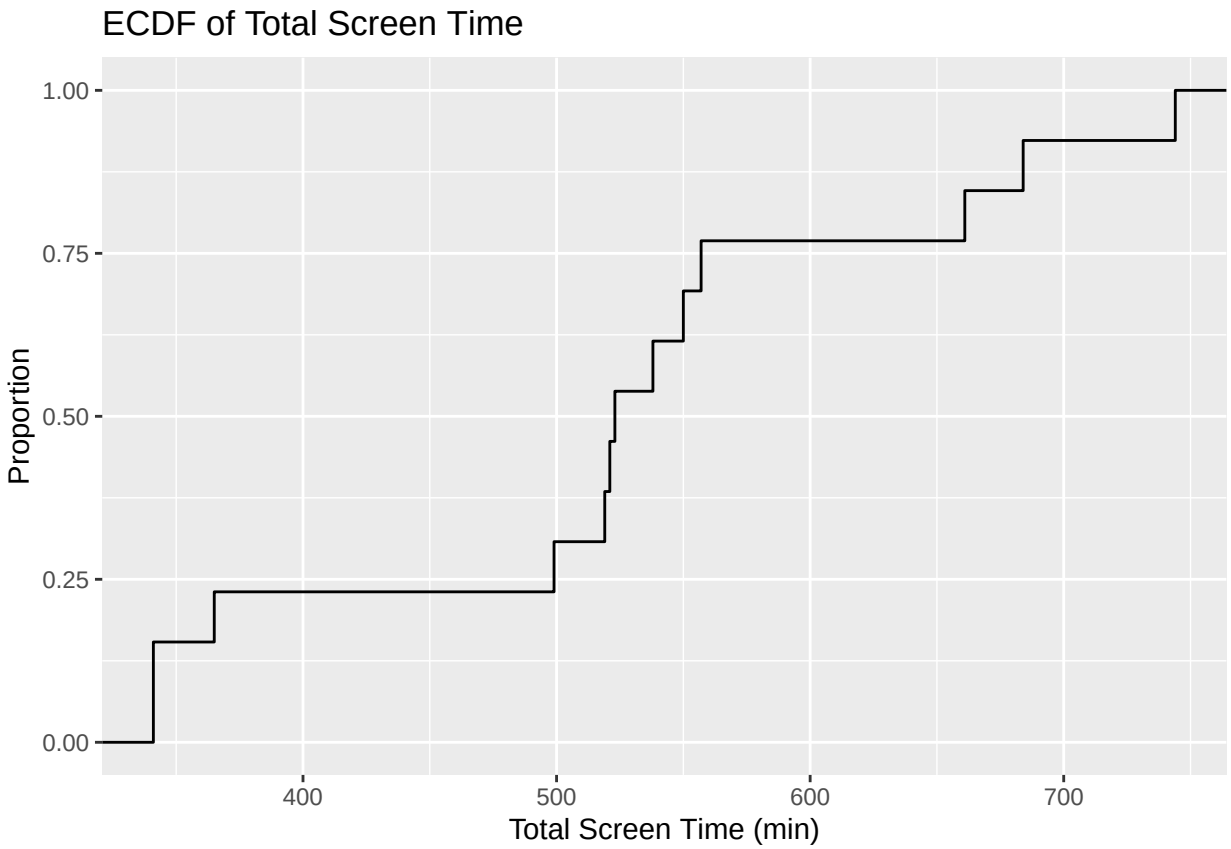


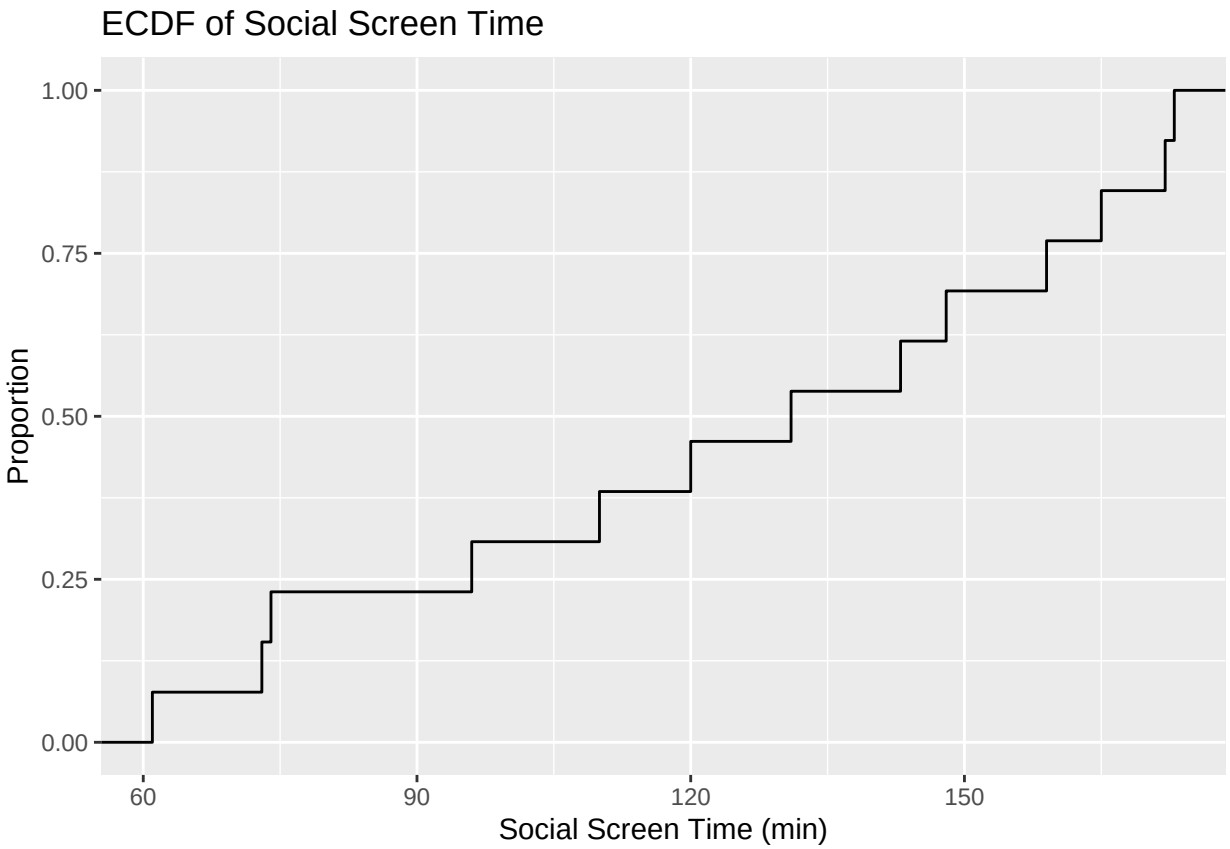
The plot shows a negative correlation between Total Screen Time and Pickups, suggesting that on days with more pickups, the total screen time might be lower. The plot shows a negative correlation between Total Screen Time and Daily Proportion of Social Screen Time, suggesting that on days with more pickups, the total screen time might be lower. There is a moderate positive correlation between Total Screen Time and Daily Duration Per Use, which suggests that longer total screen time is associated with longer average durations per use.

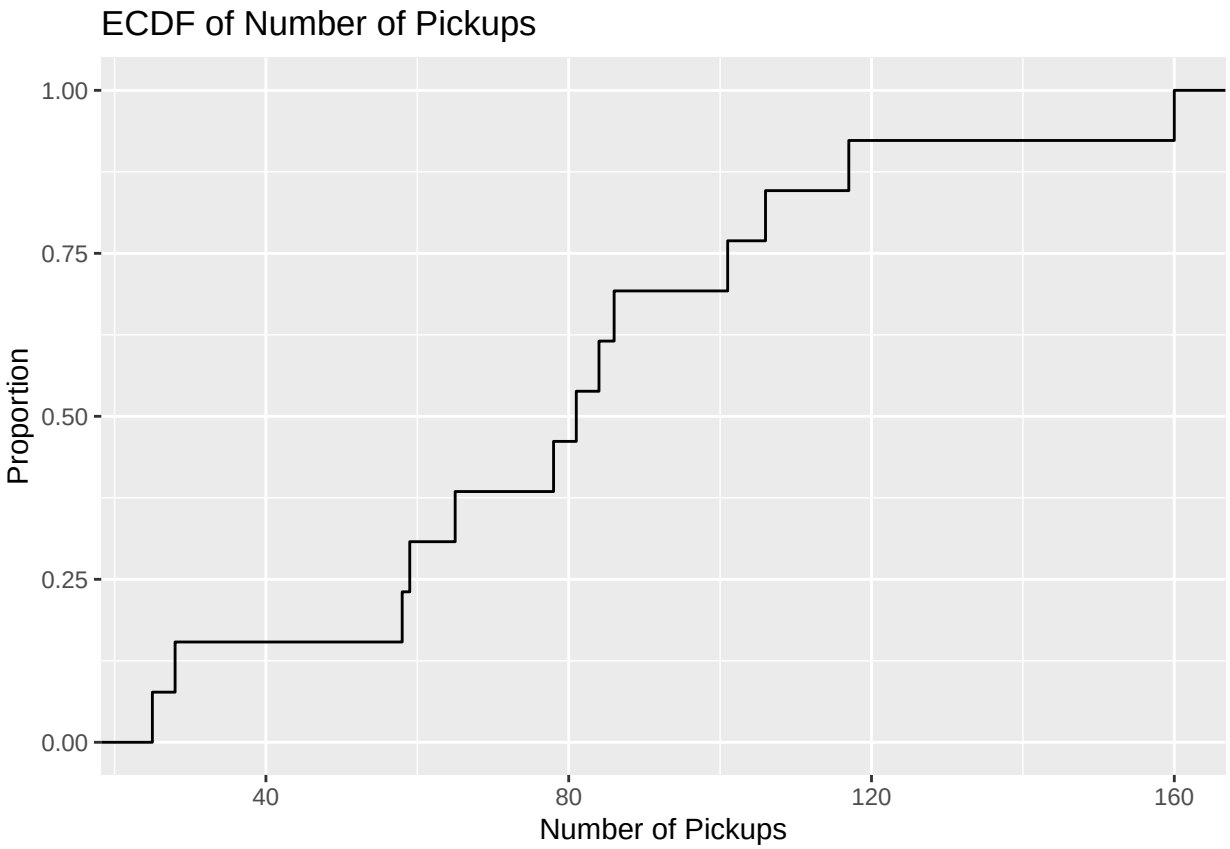
```
## [1] "Total.ST.min.true" "Total.ST"
```

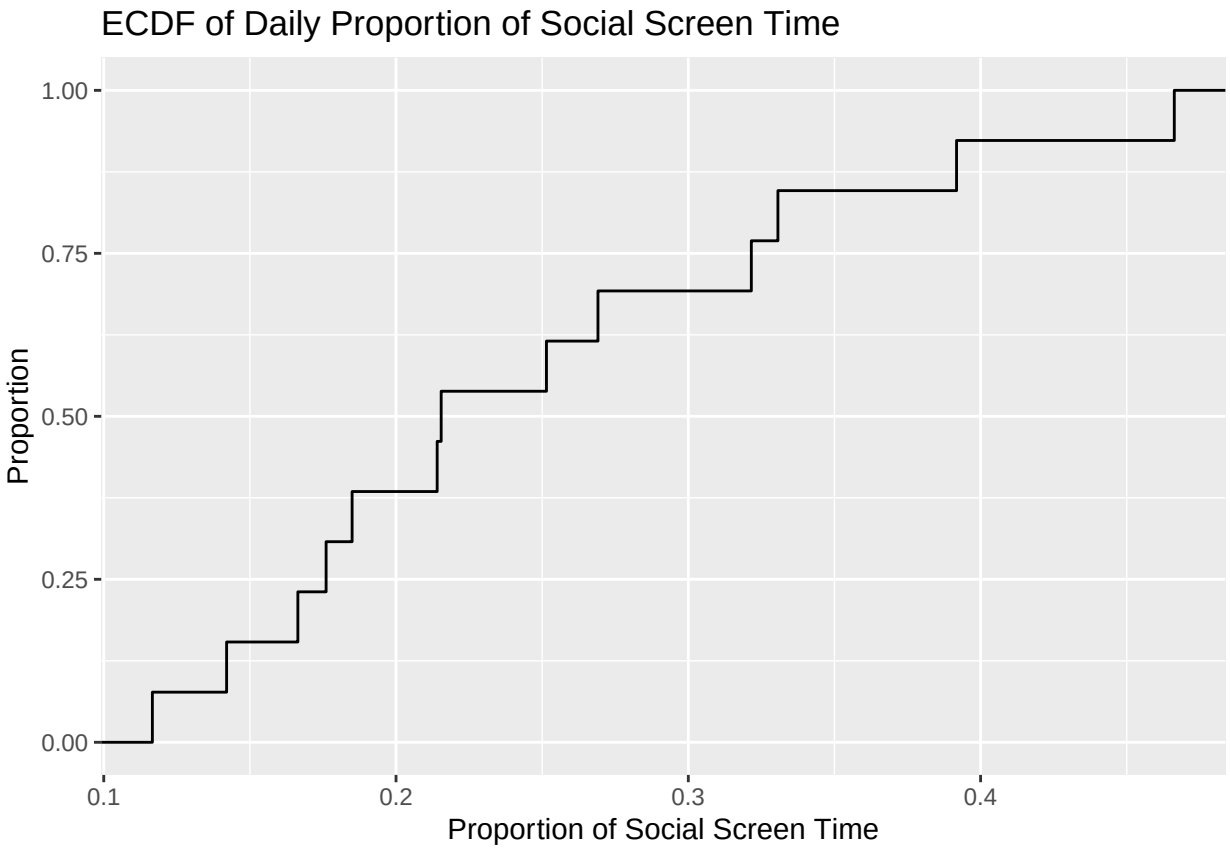
Total.ST.min.true and Total.ST have the highest correlation, among the five variables has the highest correlation.

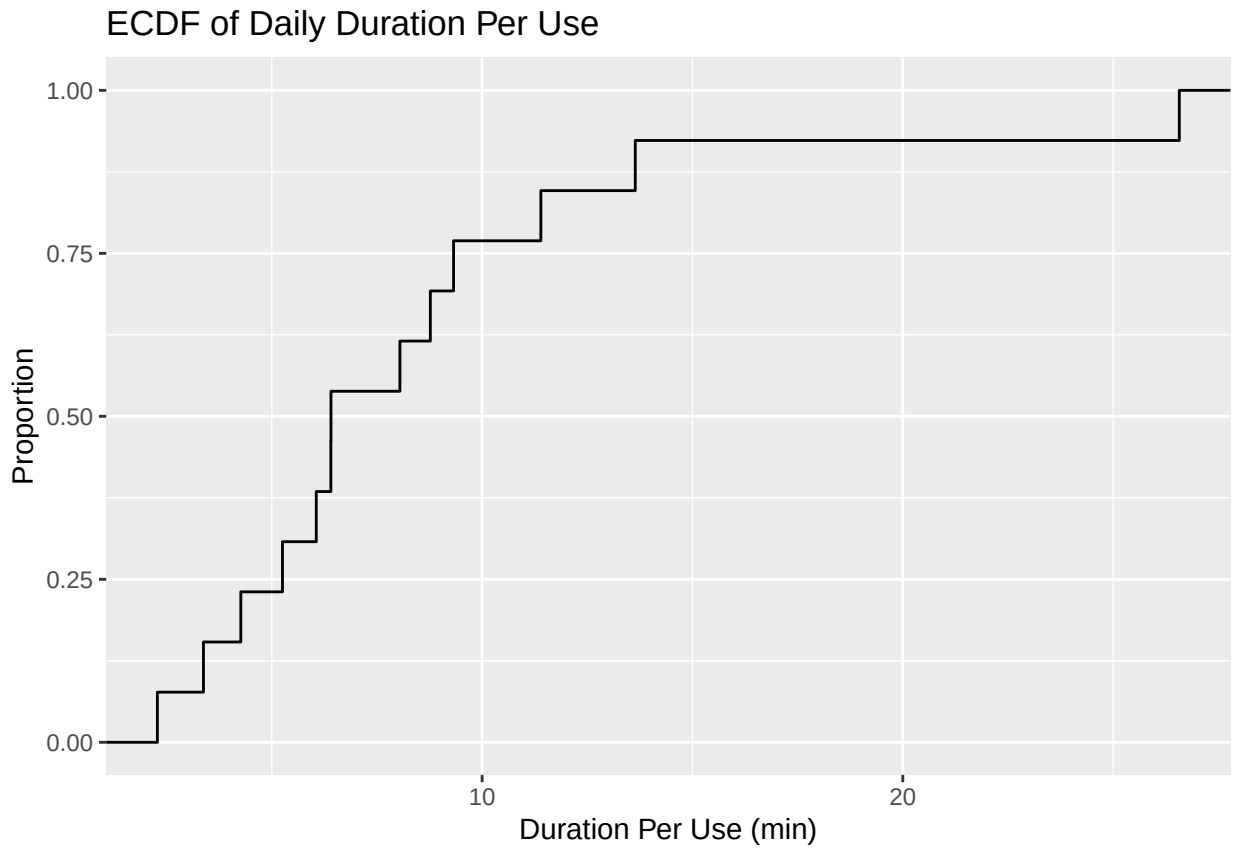
c. Make an occupation time curve for each of the five time series. Explain the pattern of individual curves.


ECDF of Total Screen Time
1.00
0.75
0.50
0.25
0.00
Proportion
400
500
600
700
Total Screen Time (min)


ECDF of Social Screen Time
1.00
0.75
0.50
0.25
0.00
Proportion
60
90
120
150
Social Screen Time (min)


ECDF of Number of Pickups
1.00
0.75
0.50
0.25
0.00
Proportion
40
80
120
160
Number of Pickups


ECDF of Daily Proportion of Social Screen Time
Proportion
1.00
0.75
0.50
0.25
0.00
0.1
0.2
0.3
0.4
Proportion of Social Screen Time



1. ECDF of Total Screen Time: This curve shows a relatively steady rise, indicating a more uniform distribution of total screen time values among the days observed.

2. ECDF of Social Screen Time: The ECDF for social screen time has more pronounced steps, indicating that there are more distinct groupings of social screen time values.

3. ECDF of Number of Pickups: The curve for the number of pickups shows a more gradual increase, which suggests a wider variation in the number of times phones are picked up each day.

4. ECDF of Daily Proportion of Social Screen Time: This curve starts steep and then becomes more gradual, indicating that a significant proportion of days have a lower proportion of social screen time.

5. ECDF of Daily Duration Per Use: The curve for daily duration per use has a sharper rise initially and then levels off, which suggests that most of the days have shorter durations per use.

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 'ST00_hy$Total.ST.min.true', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.161 -0.425 -0.035  0.177  0.017 -0.066 -0.098 -0.130  0.011 -0.024
##     11
## -0.112
```

```
##
## Autocorrelations of series 'ST00_hy$Social.ST.min.true', by lag
##
```

```
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000 -0.406 -0.131  0.202 -0.205  0.284 -0.414  0.089  0.168 -0.159  0.087
##     11
## -0.021
```

```
##
## Autocorrelations of series 'ST00_hy$Pickups', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000 -0.100  0.138 -0.102 -0.191  0.018 -0.217  0.151  0.021  0.044 -0.078
##     11
## -0.148
```

```
##
## Autocorrelations of series 'ST00_hy$Daily_Prop_Social_ST', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000 -0.264 -0.076  0.228 -0.051  0.103 -0.393  0.173 -0.112 -0.170  0.138
##     11
## -0.098
```

```
##
## Autocorrelations of series 'ST00_hy$Daily_Duration_Per_Use', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000 -0.102  0.126 -0.285 -0.089  0.065 -0.190 -0.033  0.029  0.036  0.005
##     11
## -0.043
```

I don't think there any significant autocorrelations. The value of autocorrelation at lag 1 is small.

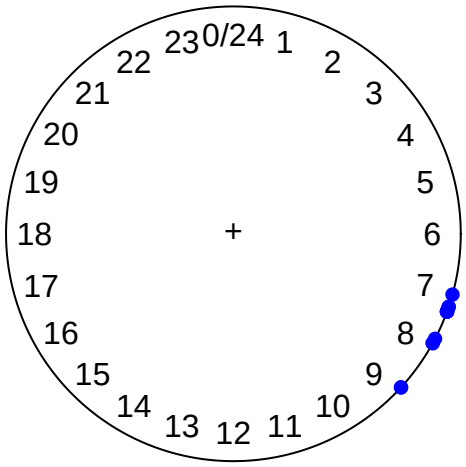
Problem 3 a. Transform (or covert) the time of first pickup to an angle ranged from 0 to 360 degree, treating midnight as 0 degree. For example, 6AM is 90 degree and noon is 180 degree.

```
## [1] 132.50 105.50 108.75 110.00 108.75 118.75
```

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.

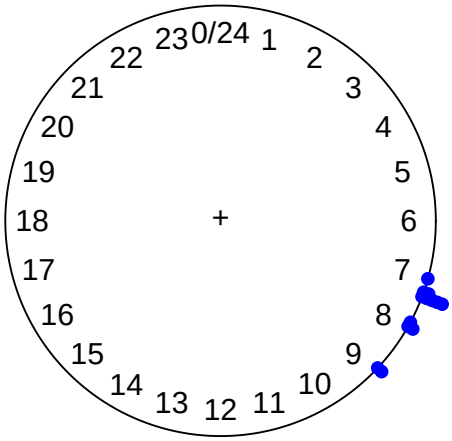
```
##
## Attaching package: 'circular'
```

```
## The following objects are masked from 'package:stats':
##
##     sd, var
```



This plot shows that the time of first pickup basically distributed around 7-9 clock. And the distribution of first pickup is very concentrated.

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.



I set a bin size at 1.25 degree, which corresponds an interval of 5 minutes. The reason why I choose 5 minutes as time interval cause particularly, my wake up time likely to concentrate at some time intervals. It's necessary to set a shorter time interval to see the differences between each days.

Problem 4: a. Explain why the factor S_t is needed in the Poisson distribution above.

The factor S_t is needed in the Poisson distribution because the Poisson model assumes that events occur independently and at a constant average rate. On days with more screen time , there are more opportunities for pickups to occur, hence the rate of pickups would be multiplied by the amount of screen time to adjust the expected number of pickups to the actual screen time for that day.

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

```
##
## Call:
## glm(formula = Pickups ~ offset(log(Total.ST.min.true/60)), family = poisson,
##     data = ST00_hy)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.21800    0.03089    71.8   <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: 339.45  on 12  degrees of freedom
## Residual deviance: 339.45  on 12  degrees of freedom
## AIC: 421
```

```
##
## Number of Fisher Scoring iterations: 5
```

c.

```
##
## Call:
## glm(formula = Pickups ~ Xt + Zt + offset(log(Total.ST.min.true/60)),
##     family = poisson, data = ST00_hy)
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.04149    0.06415  31.824  < 2e-16 ***
## Xt           0.23656    0.07319   3.232  0.00123 **
## Zt                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 339.45  on 12  degrees of freedom
## Residual deviance: 328.57  on 11  degrees of freedom
## AIC: 412.11
##
## Number of Fisher Scoring iterations: 5
```

(c.1) 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. solution: It's statistically significant (Xt's p-value < 0.05), there is evidence to suggest a significant difference in the number of pickups between weekdays and weekends.

(c.2) 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. solution: Since I started counting my phone usage time after January 14, all Zt values are 1. All the results are NA. I'm so sorry I can't answer this question.

Problem 5: a. Use the R function mle.vonmises from the R package circular to obtain the estimates of the two model parameters mu and lambda from your data of first pickups.

```
##
## Call:
## mle.vonmises(x = ST00_hy$Pickup.1st.angular)
##
## mu: -2.979  ( 1.02 )
##
## kappa: 0.388  ( 0.4033 )
```

The estimate for mu is approximately -2.979 radians. The value in parentheses is the standard error of the estimate, which is 1.02 in this case. the estimate for kappa is 0.388, and the standard error is 0.4033.

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

```
## [1] 0.7275834
```

The probability of first pickup being at 8:30 AM or later is 0.7275834.