620HW1

Zeyue Zhang

2024-02-06

```
library(readxl)
library(ggplot2)
ST_data = read_excel("/Users/zeyuezhang/Desktop/620/ScreenTime_SPH.xlsx")
```

Problem 1: Explore the your own screen activity data that you collect until the end of Friday (inclusive), January 26, 2024. This type of 'break' time set up by scientists in practice is often referred to as data freeze date during data collection. If you were unable to collect such data due to the previous setting of your mobile device or other logistic reasons, please let the instructor or GSI know immediately, some backup data would be provided to you.

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.

I collected data to investigate whether there is a potential relationship between screen time and risk of depression. Excessive use of electronic devices keeps people indoors, reduces normal social activities and opportunities for contact with nature, and may even lead to lack of sleep. These factors may increase the risk of depression. So I hypothesized that more screen time would lead to a higher risk of depression.

Reference: Madhav, K. C., Shardulendra Prasad Sherchand, and Samendra Sherchan. "Association between screen time and depression among US adults." Preventive medicine reports 8 (2017): 67-71.

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

Informed consent is a legal document. It informs the subjects and the people who provide the data about the details of the study (team members, purposes, risks, etc.), ensures that people are informed about their participation in the experiment and research, and protects people's right to know. It can also help build trust between people involved in research and research teams and ensure that data collection is open, voluntary, transparent and trustworthy.

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.

1.when the data is collected: 12/31/2023 to 1/26/2024 2.which types of variables in the data are collected: Date, Total Screen Time Use Per Day, Social Media Screen Time Use Per Day, Total Pickups Per Day, Time of First Pickup Everyday 3.where the data is collected from: The setting app from mobile phone 4.how many data are collected before the data freeze: 27 days of everything mentioned in part 2

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).

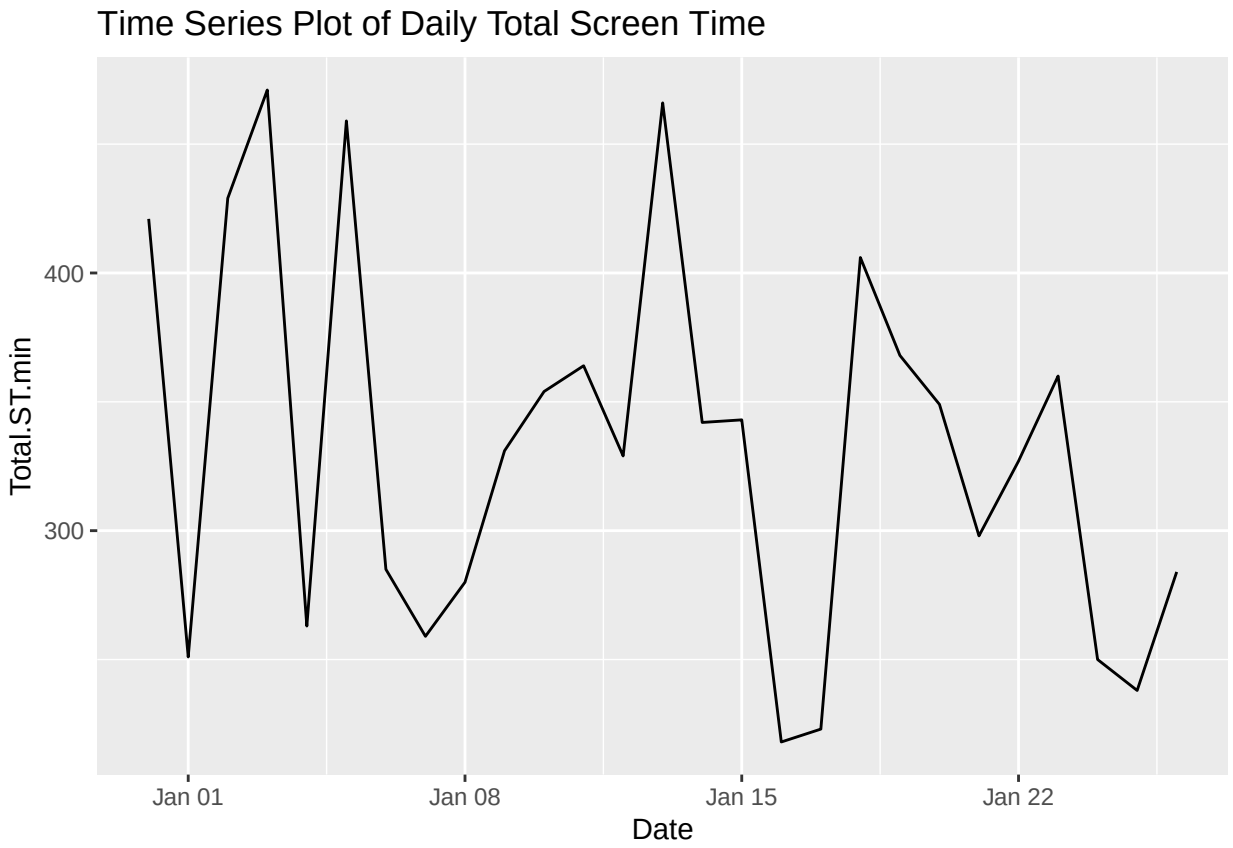
```
# Create daily_proportion_of_social_screen_time
ST_data$Proportion.social.ST <- ST_data$Social.ST.min / ST_data$Total.ST.min

# Create daily_duration_per_use
ST_data$Duration <- ST_data$Total.ST.min / ST_data$Pickups

ST_data
```

```
## # A tibble: 27 x 9
##    Date                `Total ST` Total.ST.min `Social ST` Social.ST.min Pickups
##    <dttm>              <chr>             <dbl> <chr>               <dbl>   <dbl>
##  1 2023-12-31 00:00:00 7h01m               421 2h12m                 122     220
##  2 2024-01-01 00:00:00 4h11m               251 1h36m                  96     215
##  3 2024-01-02 00:00:00 7h09m               429 1h39m                  99     137
##  4 2024-01-03 00:00:00 7h51m               471 58m                    58     132
##  5 2024-01-04 00:00:00 4h23m               263 1h56m                 116     277
##  6 2024-01-05 00:00:00 7h39m               459 1h25m                  85     174
##  7 2024-01-06 00:00:00 4h45m               285 1h51m                 111     169
##  8 2024-01-07 00:00:00 4h19m               259 2h47m                 167     174
##  9 2024-01-08 00:00:00 4h40m               280 2h09m                 129     174
## 10 2024-01-09 00:00:00 5h31m               331 1h22m                  82     183
## # i 17 more rows
## # i 3 more variables: Pickup.1st <dttm>, Proportion.social.ST <dbl>,
## #   Duration <dbl>
```

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.

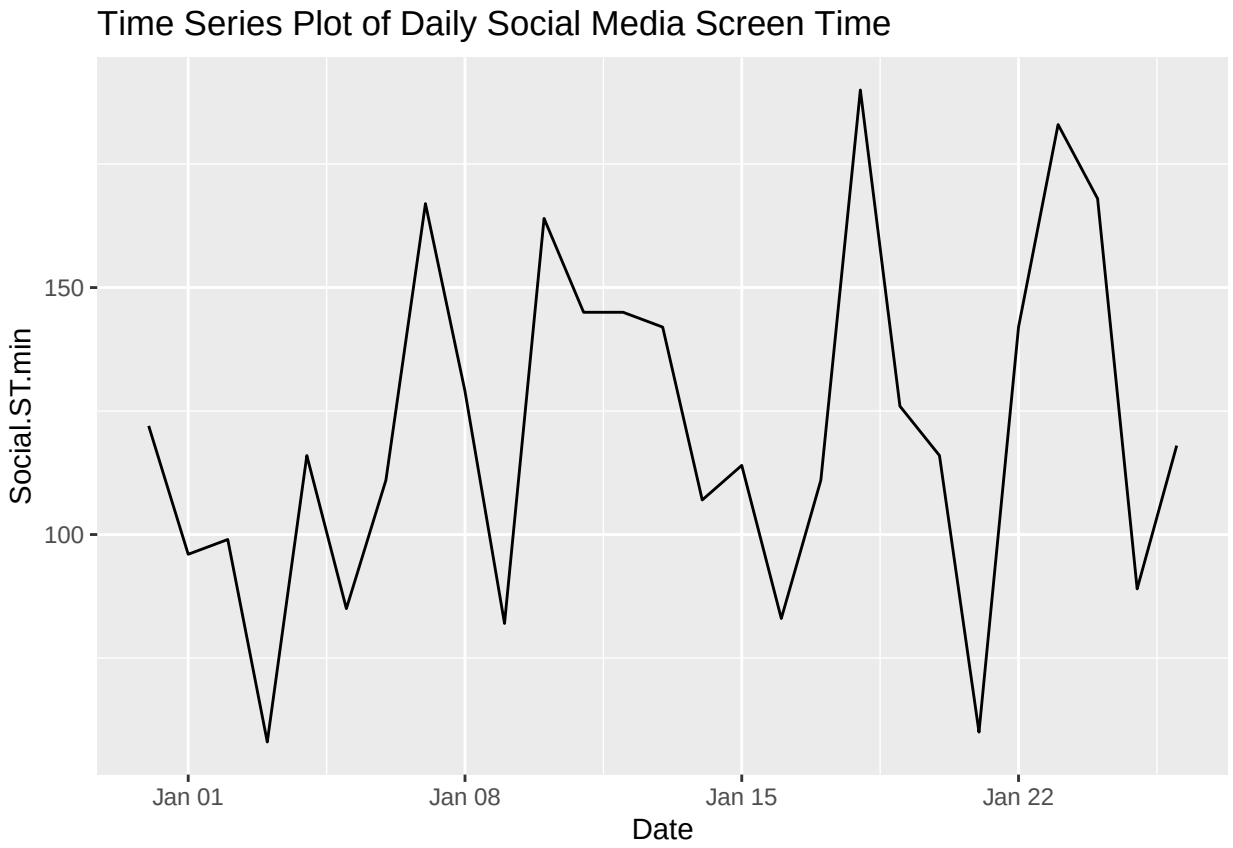
a. Make a time series plot of each of the five variables in your data. Describe temporal patterns from these time series plots.

```
library(ggplot2)
ggplot(ST_data, aes(x = Date, y = Total.ST.min)) +
  geom_line() +
  ggtitle("Time Series Plot of Daily Total Screen Time") +
  xlab("Date") +
  ylab("Total.ST.min")
```



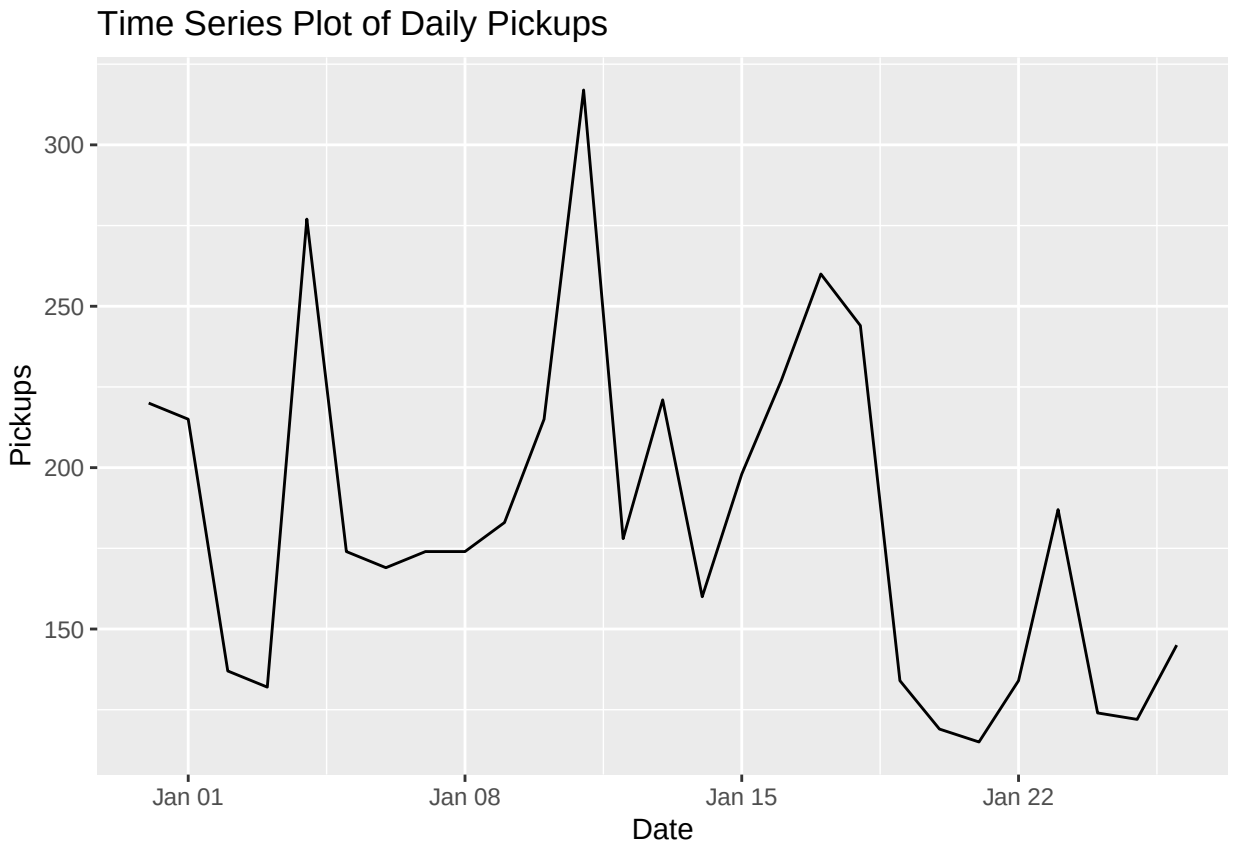
The time series graph shows changes in daily total screen time, fluctuating between about 300 and 450 minutes throughout January, with no clear long-term trend, but a cyclical pattern with peaks and troughs.

```
ggplot(ST_data, aes(x = Date, y = Social.ST.min)) +
  geom_line() +
  ggtitle("Time Series Plot of Daily Social Media Screen Time") +
  xlab("Date") +
  ylab("Social.ST.min")
```



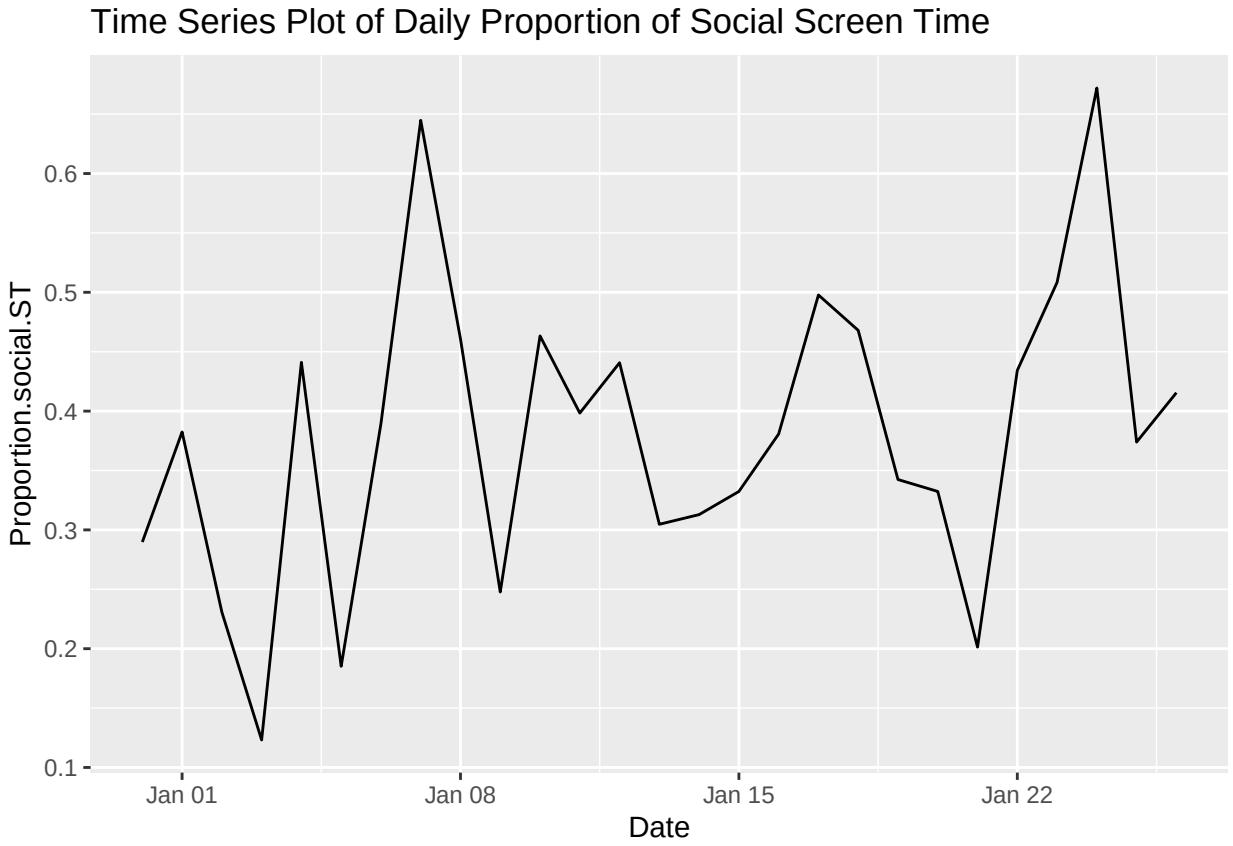
The time series graph shows changes in daily social media screen time, fluctuating between about 100 and 150 minutes throughout January, with no clear long-term trend, but a cyclical pattern with peaks and troughs.

```
ggplot(ST_data, aes(x = Date, y = Pickups)) +
  geom_line() +
  ggtitle("Time Series Plot of Daily Pickups") +
  xlab("Date") +
  ylab("Pickups")
```



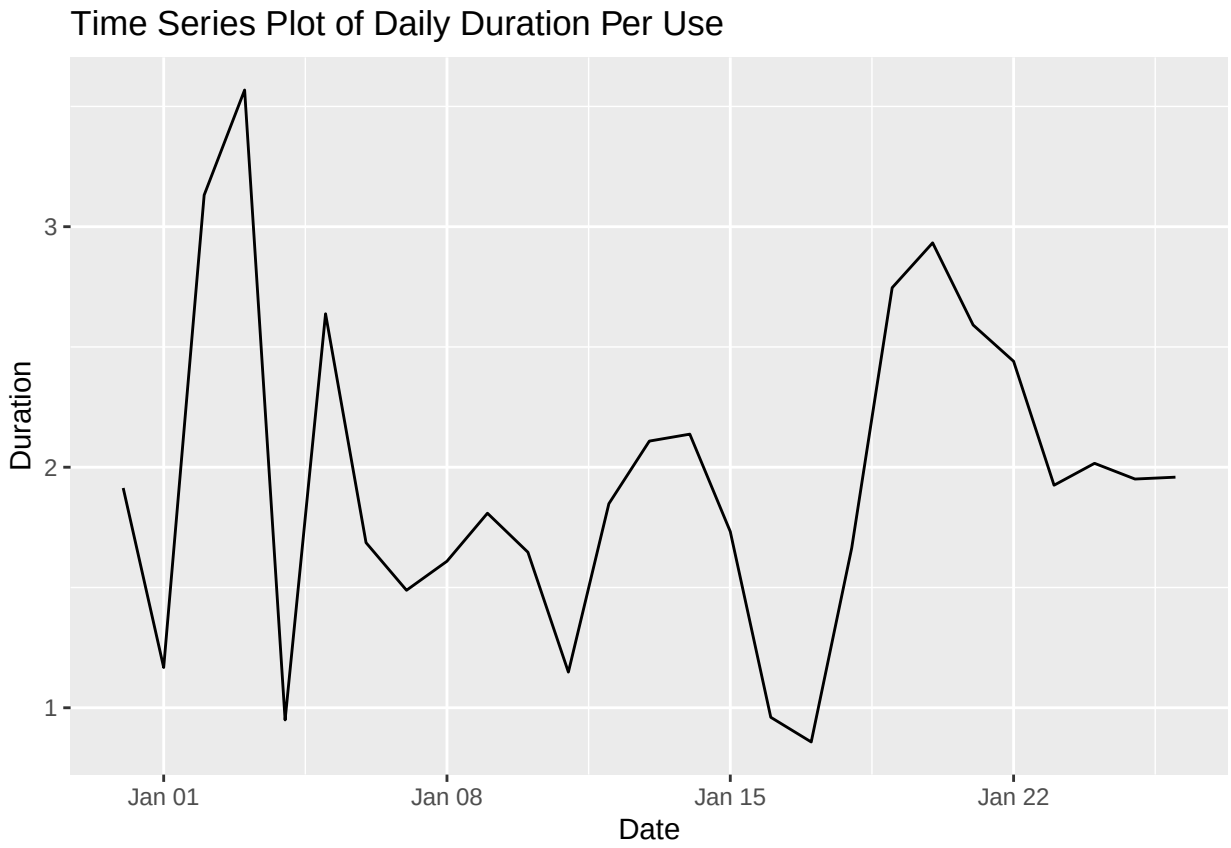
The time series graph shows changes in daily pickups, fluctuating between about 50 and 350 minutes throughout January, with a decreasing trend, in the end of the January the pickups go down obviously.

```
ggplot(ST_data, aes(x = Date, y = Proportion.social.ST)) +
  geom_line() +
  ggtitle("Time Series Plot of Daily Proportion of Social Screen Time") +
  xlab("Date") +
  ylab("Proportion.social.ST")
```



The time series graph shows changes in daily proportion of social screen time, fluctuating between about 0.1 and 0.7 throughout January, with no clear long-term trend, but a cyclical pattern with peaks and troughs.

```
ggplot(ST_data, aes(x = Date, y = Duration)) +
  geom_line() +
  ggtitle("Time Series Plot of Daily Duration Per Use") +
  xlab("Date") +
  ylab("Duration")
```



The time series graph shows changes in daily duration per use, fluctuating between about 0.5 and 4 throughout January, with no clear long-term trend, but a cyclical pattern with peaks and troughs.

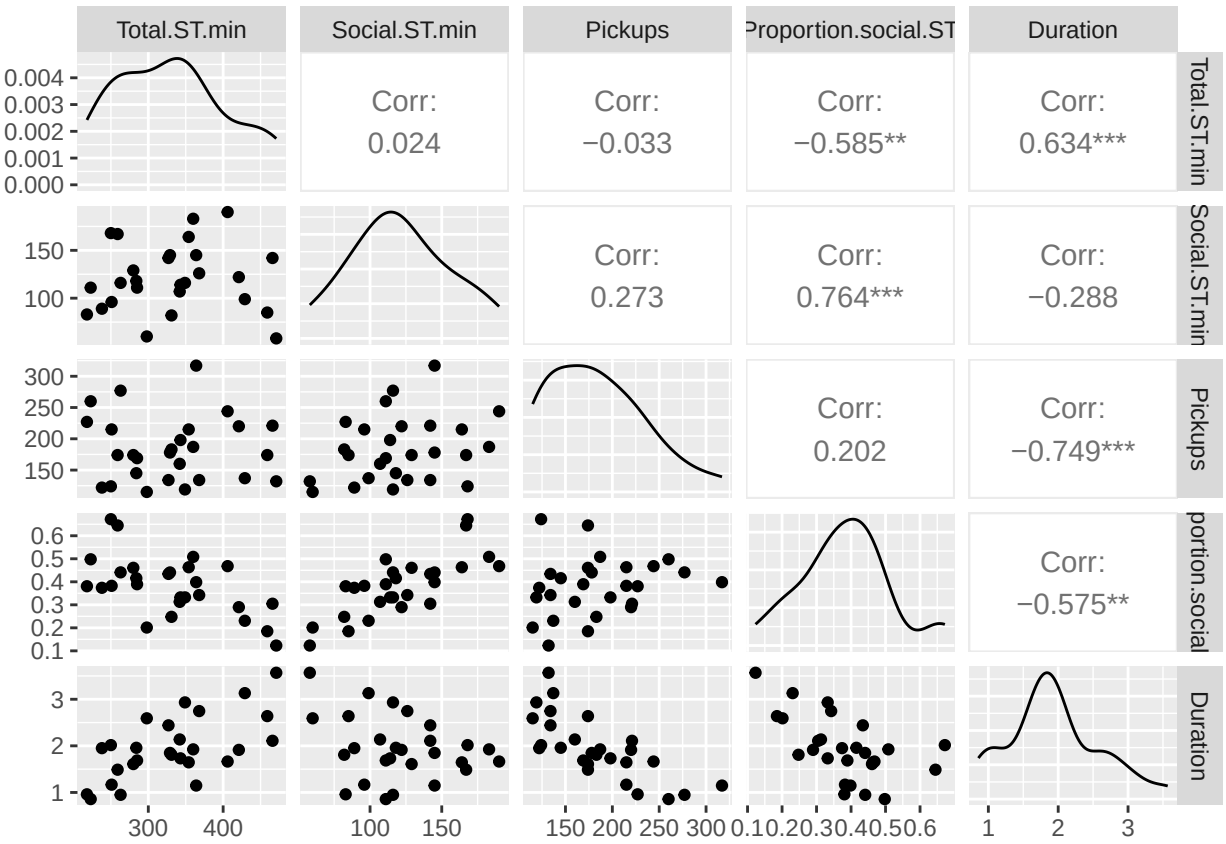
b. Make pairwise scatterplots of five variables. Describe correlation patterns from these pairwise scatterplots. Which pair of variables among the five variables has the highest correlation?

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.2.3
```

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

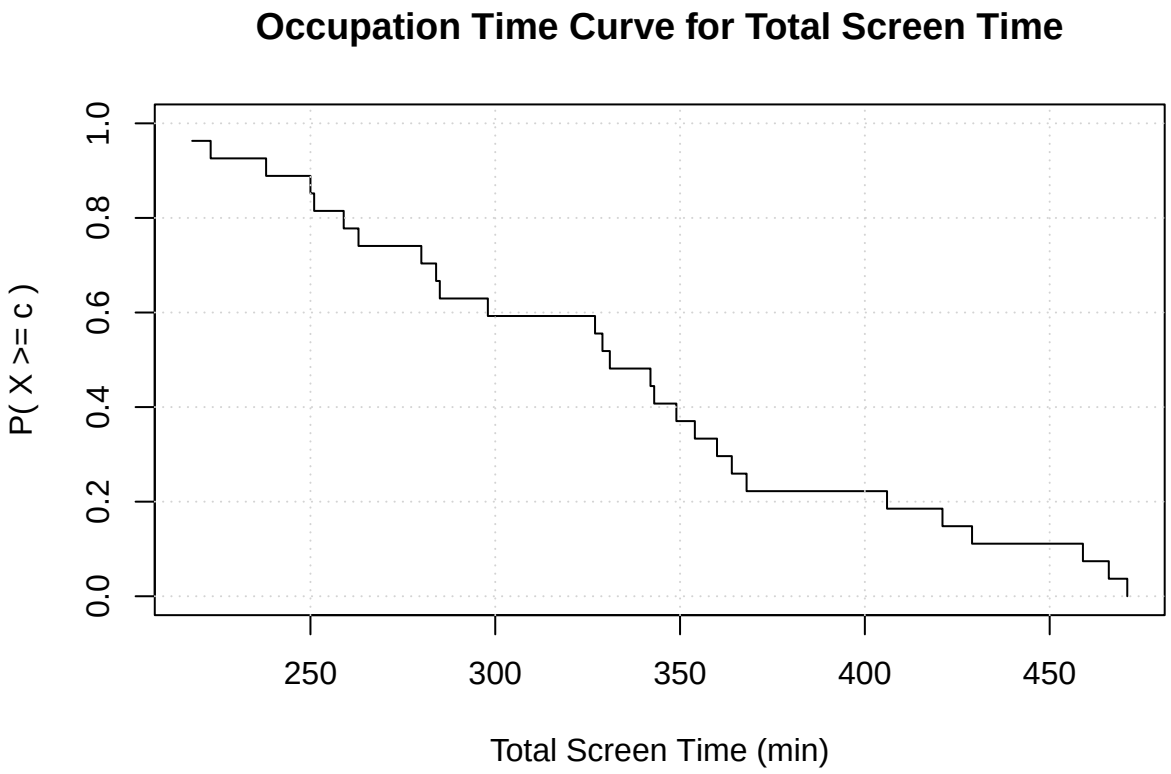
```
selected_vars <- ST_data[, c("Total.ST.min", "Social.ST.min", "Pickups", "Proportion.social.ST", "Durati
ggpairs(selected_vars)
```



The plot and correlations indicates the linear relationships between five variables. The strongest positive correlation is between Social.ST.min and Proportion.social.ST (0.764), suggesting a strong relationship that is statistically significant. When one goes up, the other does too. On the other hand, the strongest negative correlation exists between Pickups and Duration (-0.749), indicating a strong inverse relationship. When one goes up, the other goes down.

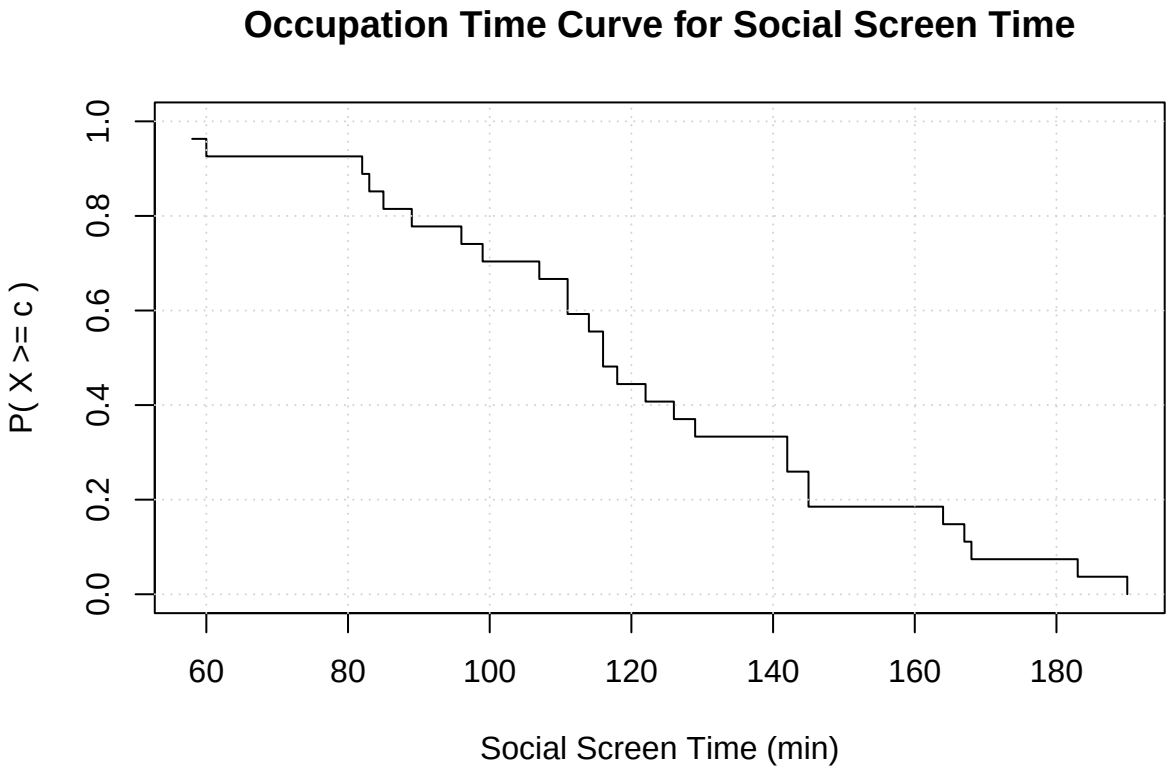
c. Make an occupation time curve for each of the five time series. Explain the pattern of individual curves.

```
total_screen_time <- sort(ST_data$Total.ST.min)
ccdf_values <- 1 - ecdf(ST_data$Total.ST.min)(total_screen_time)
plot(total_screen_time, ccdf_values, type = "s",
     main = "Occupation Time Curve for Total Screen Time",
     xlab = "Total Screen Time (min)", ylab = "P( X >= c )",
     xlim = range(total_screen_time), ylim = c(0, 1))
grid()
```



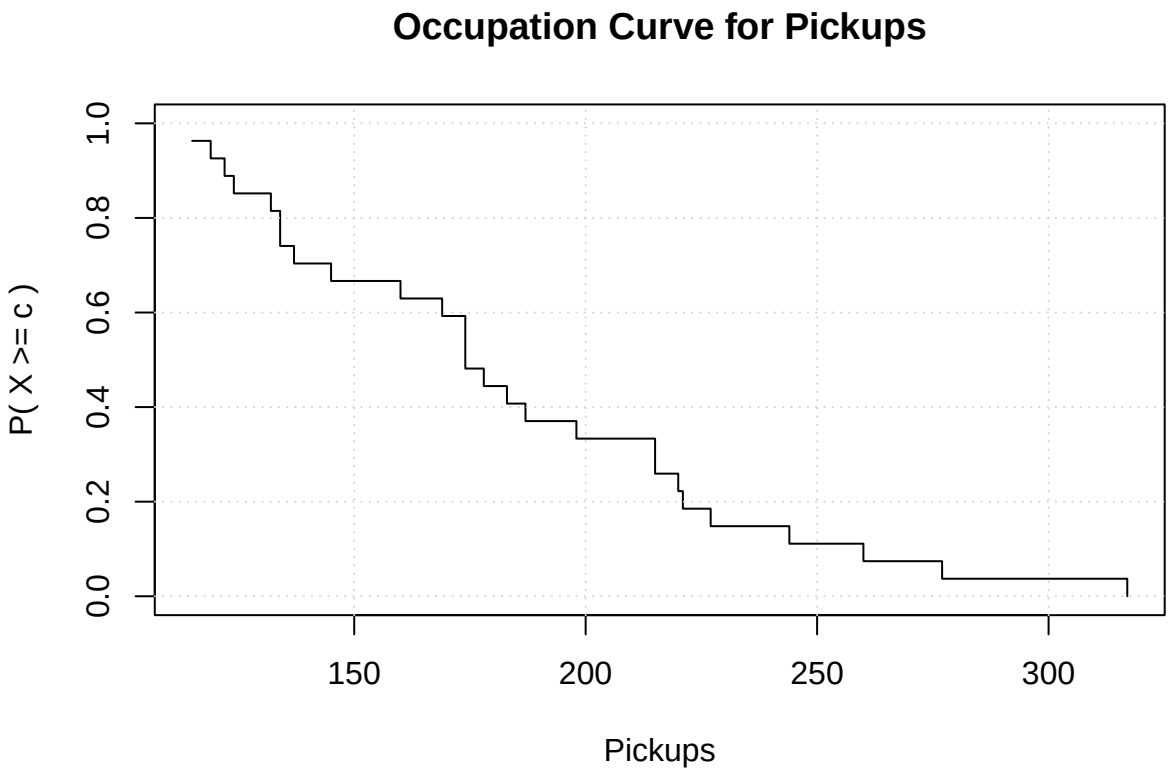
The overall downward trend of the curve indicates that the number of days with lower screen use time is greater. Conversely, days with longer screen time are fewer.

```
social_screen_time <- sort(ST_data$Social.ST.min)
ccdf_values <- 1 - ecdf(ST_data$Social.ST.min)(social_screen_time)
plot(social_screen_time, ccdf_values, type = "s",
     main = "Occupation Time Curve for Social Screen Time",
     xlab = "Social Screen Time (min)", ylab = "P( X >= c )",
     xlim = range(social_screen_time), ylim = c(0, 1))
grid()
```



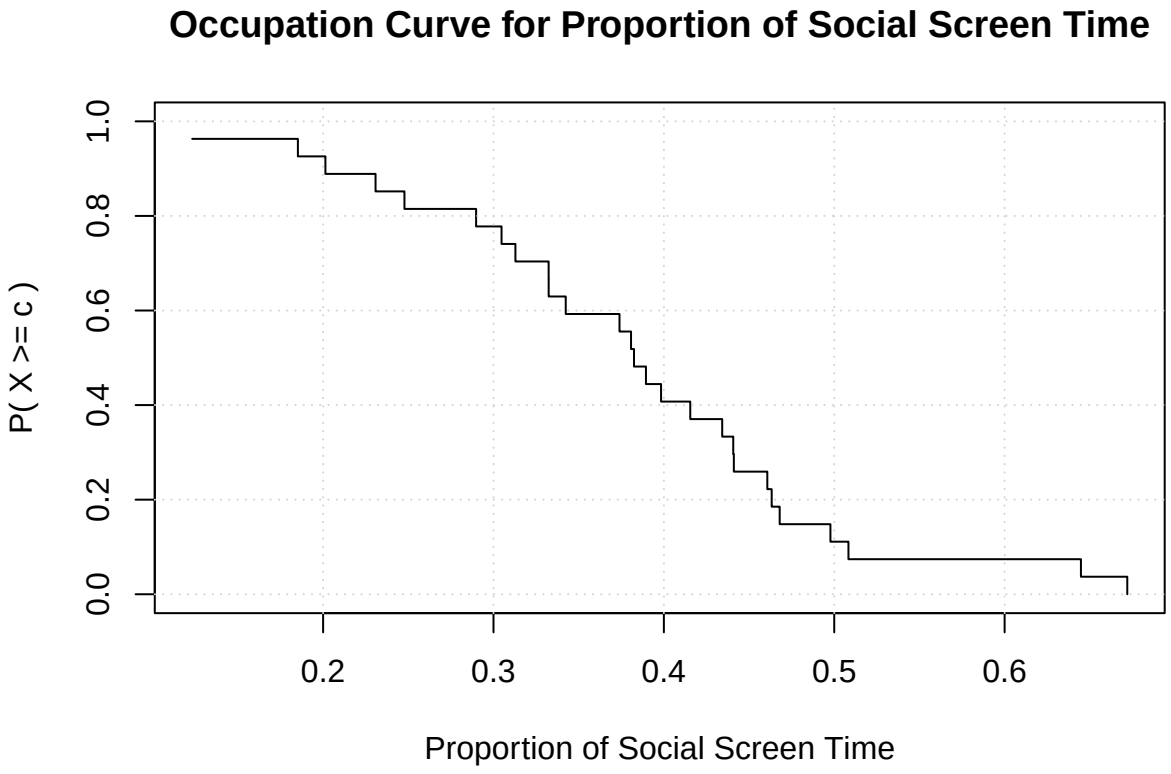
The overall downward trend of the curve indicates that the number of days with lower social media screen time is greater. Conversely, days with longer social media screen time were lower.

```
Pickups <- sort(ST_data$Pickups)
ccdf_values <- 1 - ecdf(ST_data$Pickups)(Pickups)
plot(Pickups, ccdf_values, type = "s",
     main = "Occupation Curve for Pickups",
     xlab = "Pickups", ylab = "P( X >= c )",
     xlim = range(Pickups), ylim = c(0, 1))
grid()
```



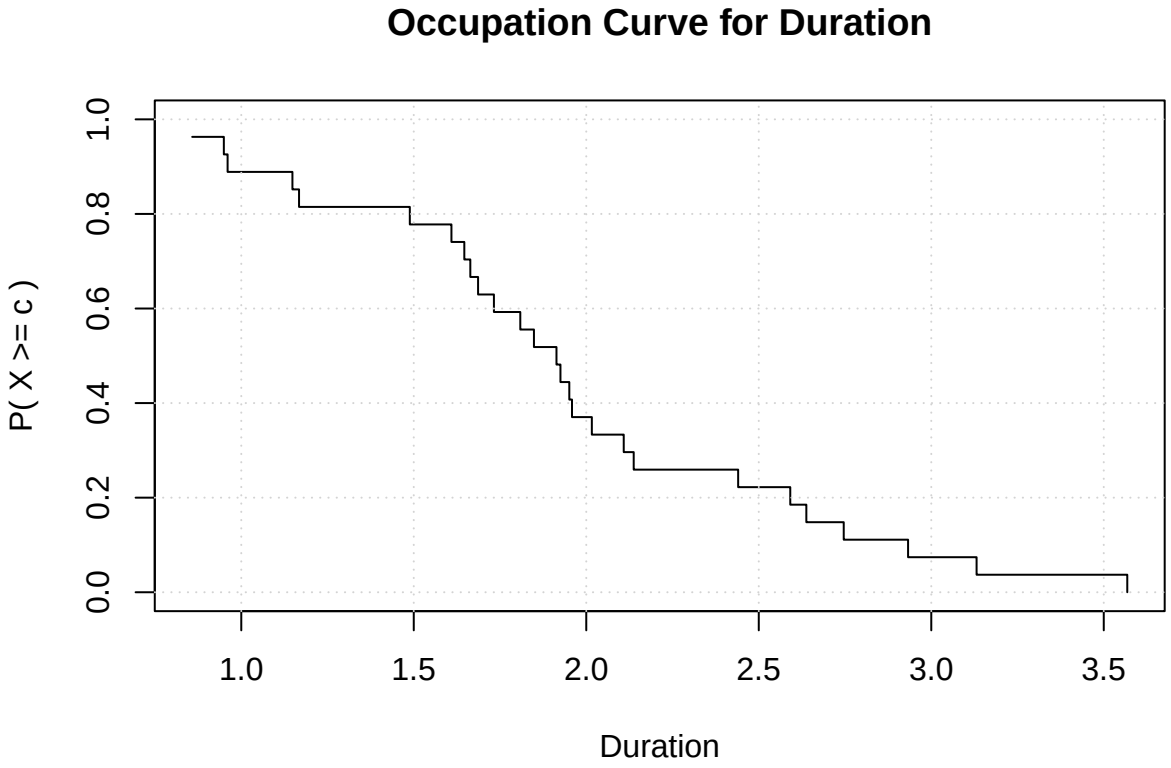
The overall downward trend of the curve indicates that there are more days when the phone is picked up less often. Conversely, days when the phone is picked up more often are fewer.

```
proportion_social_screen_time <- sort(ST_data$Proportion.social.ST)
ccdf_values <- 1 - ecdf(ST_data$Proportion.social.ST)(proportion_social_screen_time)
plot(proportion_social_screen_time, ccdf_values, type = "s",
     main = "Occupation Curve for Proportion of Social Screen Time",
     xlab = "Proportion of Social Screen Time", ylab = "P( X >= c )",
     xlim = range(proportion_social_screen_time), ylim = c(0, 1))
grid()
```



The overall downward trend of the curve indicates that there are more days with lower Proportion of Social Screen Time. On the contrary, the number of days with a larger Proportion of Social Screen Time is smaller.

```
Duration <- sort(ST_data$Duration)
ccdf_values <- 1 - ecdf(ST_data$Duration)(Duration)
plot(Duration, ccdf_values, type = "s",
     main = "Occupation Curve for Duration",
     xlab = "Duration", ylab = "P( X >= c )",
     xlim = range(Duration), ylim = c(0, 1))
grid()
```



The overall downward trend of the curve indicates that the number of days with lower Duration is greater. On the contrary, the number of days with higher Duration is smaller.

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.

```
acf(ST_data$Total.ST.min, plot = FALSE)
```

```
##
## Autocorrelations of series 'ST_data$Total.ST.min', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000 -0.008 -0.032  0.002 -0.268  0.044 -0.174 -0.073  0.264  0.153  0.237
##     11     12     13     14
## -0.092 -0.231  0.027 -0.202
```

```
acf(ST_data$Social.ST.min, plot = FALSE)
```

```
##
## Autocorrelations of series 'ST_data$Social.ST.min', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.150 -0.217 -0.204  0.030  0.186  0.095 -0.175 -0.101 -0.257 -0.064
##     11     12     13     14
##  0.177  0.177  0.059 -0.086
```

```r
acf(ST_data$Pickups, plot = FALSE)
```

```
##
## Autocorrelations of series 'ST_data$Pickups', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.295  0.045 -0.048  0.030  0.125  0.157  0.180 -0.353 -0.292 -0.175
##     11     12     13     14
## -0.087  0.038 -0.037 -0.113
```

```r
acf(ST_data$Proportion.social.ST, plot = FALSE)
```

```
##
## Autocorrelations of series 'ST_data$Proportion.social.ST', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.160 -0.076 -0.084 -0.181 -0.110  0.158 -0.009 -0.060 -0.089  0.044
##     11     12     13     14
##  0.053  0.065 -0.043 -0.109
```

```r
acf(ST_data$Duration, plot = FALSE)
```

```
##
## Autocorrelations of series 'ST_data$Duration', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.193 -0.085 -0.026 -0.236 -0.088  0.027  0.088 -0.110 -0.086  0.027
##     11     12     13     14
## -0.036 -0.073 -0.178 -0.225
```

There is no significant autocorrelation.

Problem 3: Explore the use of the R package circular to display the time of first pickup as a circular variable or angular variable. 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.

```r
library(circular)
```

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

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

```r
library(lubridate)
```

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

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

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

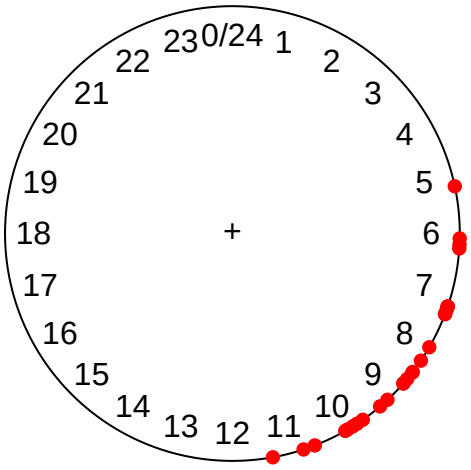
```
##
## Attaching package: 'dplyr'
```

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

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

```
ST_data = ST_data %>%
  mutate(FirstPickupTimeAngular = (hour(Pickup.1st)*60 + minute(Pickup.1st)) / (24*60) * 360)
```

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.

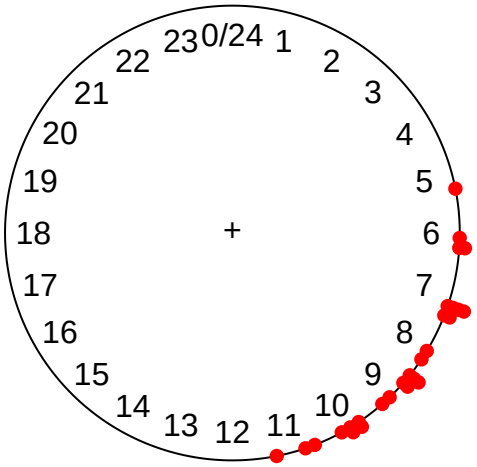
```
angularValues = ST_data$FirstPickupTimeAngular
circularData = circular(angularValues, units = "degrees", template = "clock24")
plot(circularData, col = "red")
```



The earliest time to pick up the phone for the first time is 5 a.m. and the latest time to pick up the phone for the first time is 11 a.m. In most cases, the first time the user picks up the phone is before 12 noon, indicating that the user never gets up.

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.

```r
plot(circularData, stack = TRUE, bins = 144, col = "red")
```



I set a bin size at 2.5 degree, which corresponds an interval of 10 minutes. The interval of 10 minutes is more in line with our daily habits of using mobile phones and is also very reasonable from the plot.

Problem 4: Consider the data of the daily number of pickups. Let Y_t be the daily number of pickups at day t, and let S_t be the daily total screen time at day t. Then, we assume that $Y_t \sim \text{Poisson}(S_t\lambda)$, t = 1,...,T, where T is the number of days for data collection, and λ is the expected hourly rate of pickups (or the expected number of pickups per hour). Note that here S_t needs to be recorded in unite of hour.

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

The factor S_t adjusts for daily screen time in predicting pickups. Since screen time varies daily, S_t, measured in hours, scales the hourly pickup rate (λ) to match the actual screen usage. This ensures accurate predictions of pickups based on the day's screen time, making S_t crucial for adjusting the expected number of pickups to the real world context of each day's screen activity.

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

```r
ST_data$Total.ST.hour <- ST_data$Total.ST.min / 60
model <- glm(Pickups ~ 1 + offset(log(Total.ST.hour)), data = ST_data, family = poisson())
summary(model)
```

```
##
## Call:
## glm(formula = Pickups ~ 1 + offset(log(Total.ST.hour)), family = poisson(),
##     data = ST_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.7905  -2.9855  -0.8345   1.4431  10.7120
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.50108    0.01421   246.4   <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: 648.97  on 26  degrees of freedom
## Residual deviance: 648.97  on 26  degrees of freedom
## AIC: 840.33
##
## Number of Fisher Scoring iterations: 4
```

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 $\ln(\lambda(t)) = \beta_0 + \beta_1 X_t + \beta_2 Z_t$, 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. This rate parameter depends on day t. Use the R function glm to estimate the regression coefficients and answer the following questions.

```r
ST_data$Date <- as.Date(ST_data$Date, format="%Y-%m-%d")
ST_data$X_t <- as.integer(!weekdays(ST_data$Date) %in% c("Saturday", "Sunday"))
year_of_analysis <- 2024
ST_data$Z_t <- as.integer(ST_data$Date >= as.Date(paste(year_of_analysis, "-01-10", sep="")))

model_new <- glm(Pickups ~ X_t + Z_t, data = ST_data, family = poisson())
summary(model_new)
```

```
##
## Call:
## glm(formula = Pickups ~ X_t + Z_t, family = poisson(), data = ST_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0939  -3.9913  -0.5084   2.3973   8.6202
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.13958    0.03355 153.185  < 2e-16 ***
## X_t          0.11722    0.03345   3.504 0.000458 ***
## Z_t         -0.02449    0.02943  -0.832 0.405322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 375.83  on 26  degrees of freedom
## Residual deviance: 362.97  on 24  degrees of freedom
## AIC: 558.33
##
## Number of Fisher Scoring iterations: 4
```

(c.1) Is there data evidence for significantly different behavior of daily pickups between weekdays and weekends? Justify your answer using the significance level $\alpha = 0.05$. There is data evidence for significantly different behavior of daily pickups between weekdays and weekends cause p=2e-16 (<0.05).

(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 $\alpha = 0.05$. There is no data evidence for a significant change on the behavior of daily pickups after the winter semester began cause p=0.405322 (>0.05).

Problem 5: Now analyze the first pickups data from Problem 3. The von Mises distribution is widely used to model a circular random variable Y.

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.

```
model_estimates <- mle.vonmises(ST_data$FirstPickupTimeAngular)
```

```
## Warning in as.circular(x): an object is coerced to the class 'circular' using default value for the
##   type: 'angles'
##   units: 'radians'
##   template: 'none'
##   modulo: 'asis'
##   zero: 0
##   rotation: 'counter'
## conversion.circularxradians0counter2pi
```

```
print(model_estimates)
```

```
##
## Call:
## mle.vonmises(x = ST_data$FirstPickupTimeAngular)
##
## mu: -2.856  ( 2.483 )
##
## kappa: 0.1097  ( 0.2728 )
```

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.

```
probability <- 1 - pvonmises(circular(8.5/24*2*pi), mu = model_estimates$mu, kappa = model_estimates$ka

print(probability)
```

```
## [1] 0.7077172
```

The probability that first pickup is 8:30AM or later is 70.77%.