HW1Bios620

Mimi Li

2024-02-04

Data

```
#install.packages("lubridate")
#install.packages(("GGally"))
#install.packages(("circular"))
library(readxl)
library(lubridate)
```

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

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

```
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
```

```
library(ggplot2)
library("GGally")
```

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

```
library(circular)
```

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

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

```
data = read_excel(path = "/Users/mimi/BIOSTAT620hw1/ST_Mimi.xlsx")
```

```r
# Change datatype of Date
data <- data %>%
  mutate(Date = as.Date(Date))

# Function to extract the time component from 'Pickup.1st'
extract_time <- function(x) {
  time_part <- strsplit(as.character(x), split = " ")[[1]][2]
  return(time_part)
}

# Correct dates of pickup 1st
data <- data %>%
  mutate(Pickup.1st = as.POSIXct(paste(as.character(Date), sapply(Pickup.1st, extract_time))))


# Convert time strings to minutes
convert_min <- function(time) {
  split_hr_min <- strsplit(time, "h|m")[[1]]
  hours <- as.numeric(split_hr_min[1])
  minutes <- as.numeric(split_hr_min[2])
  return(hours * 60 + minutes)
}

# Convert Total.ST and Social.ST
data <- data %>%
  mutate(Total.ST.min = sapply(Total.ST, convert_min))

data <- data %>%
  mutate(Social.ST.min = sapply(Social.ST, convert_min))
```

Problem 1

(a)

The purpose of this data collection is to explore the potential association between screen time and poor mental health among college students. Existing research indicates that increased screen time may inversely affect physical activity levels, illustrating that higher screen time is associated with reduced exercise time among adolescents (Penglee et al. 2019). Additionally, excessive screen time has been linked to reduced sleep quality, which is a factor critical for cognitive function and overall well-being (Xu et al. 2019). Given the importance of both physical activity and sleep to mental health (Scott et al. 2021) (Mikkelsen et al. 2017), there is a compelling need to investigate how screen time contributes to mental health challenges. By collecting and analyzing screen activity data, we aim to identify patterns that may suggest a relationship between screen usage and indicators of poor mental health among graduate students.

(b)

The Informed Consent Form serves as a formal communication tool between the researchers and participants, establishing trust and promoting transparency about the research process. It ensures participants are fully aware of the study's purpose, procedures, risks, benefits, and other information before agreeing to take part, which respects the participant's right of autonomy to make an informed decision about their involvement. For the data collection, the Informed Consent Form would detail how the data will be collected, how data privacy and safety will be maintained, and how the data will be used. For researchers, it provides evidence that participants have consented to the study, which legally protects the researchers and the institution

conducting the study.

(c)

Table 1: Data Collection Plan

Data collection Time	2/4/24
Variables	Total screen time (Total.ST), social screen time (Social.ST), total pickups (Pickups), and first pickup time (Pickup.1st)
Data source	The recorded screen activity in real-time by the personal mobile device
Total counts	14 days (1/14/24 - 1/27/24)
Mobile device settings	The time zone was adjusted to GMT-8 to collect the first pickup time.

(d)

```
# Create the two new variables
data <- data %>%
  mutate(Daily.prop.social = Social.ST.min / Total.ST.min)

data <- data %>%
  mutate(Daily.duration = Total.ST.min / Pickups)
```

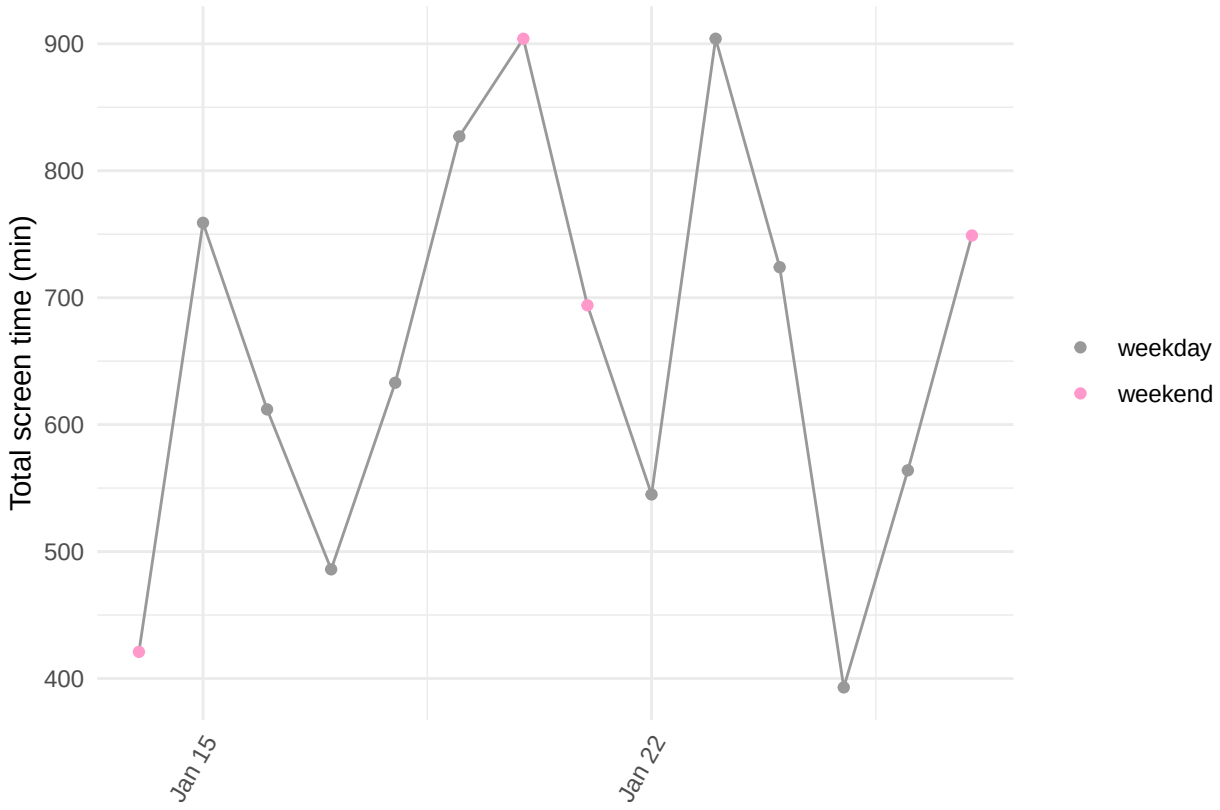
Problem 2

(a)

Based on the plots, the patterns of total screen time, the social screen time, and daily duration are similar. There is no obvious difference between the weekdays and weekends in all five variables.

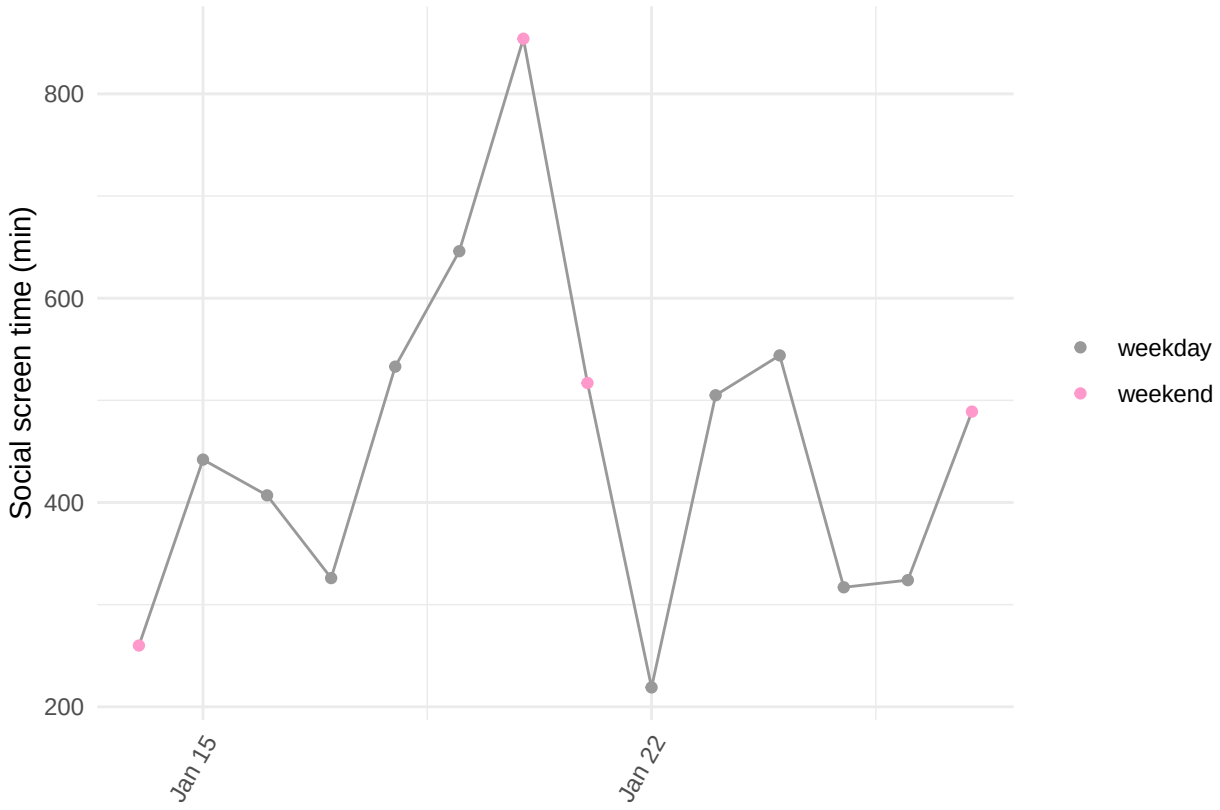
Time series plot of total screen time

```
# Check if the date is a weekend
data$if_weekend <- ifelse(weekdays(data$Date) %in% c("Saturday", "Sunday"), "weekend", "weekday")

# Create the plot
total_st_plot <- ggplot(data, aes(x = Date, y = Total.ST.min, group = 1)) +
  geom_line(color = "#999999") +
  geom_point(aes(color = if_weekend)) +
  labs(x = "", y = "Total screen time (min)") +
  scale_color_manual(values = c("weekday" = "#999999", "weekend" = "#FF99CC")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        legend.title = element_blank())

# Print the plot
total_st_plot
```



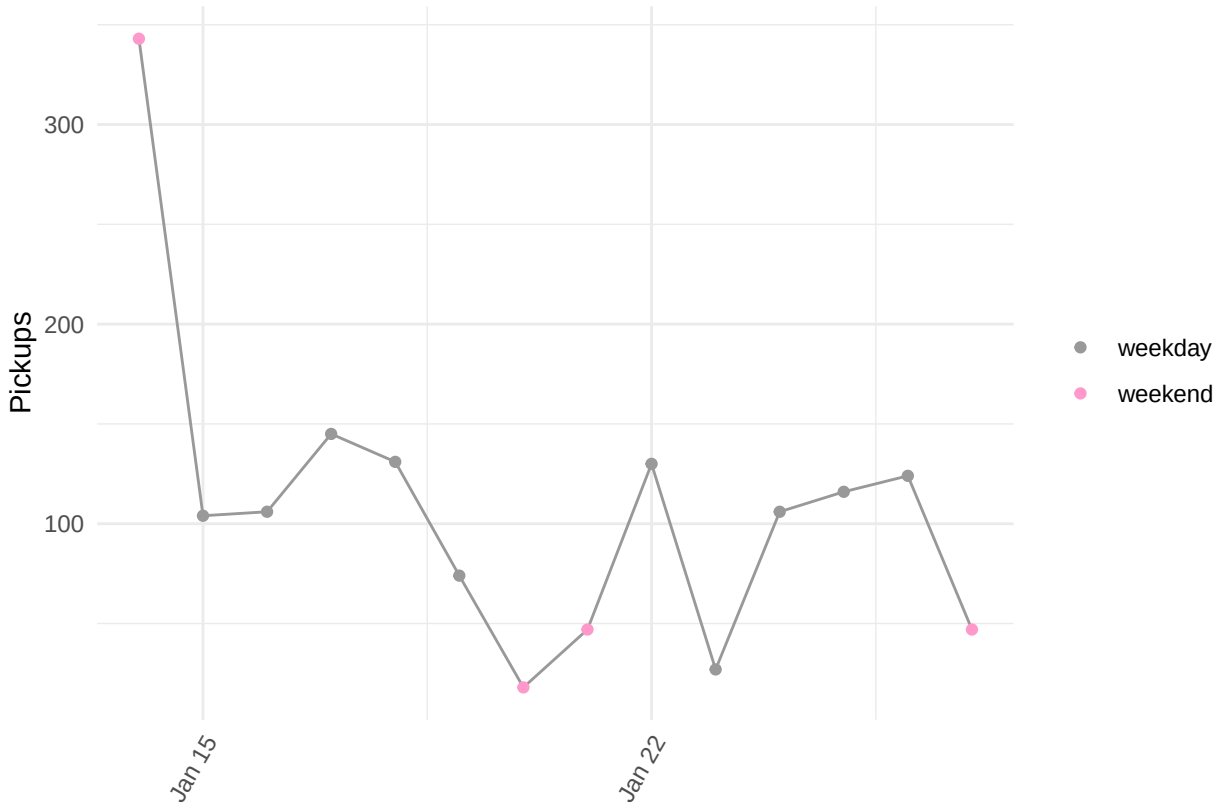
Time series plot of social screen time

```
# Create the plot
total_scl_plot <- ggplot(data, aes(x = Date, y = Social.ST.min, group = 1)) +
  geom_line(color = "#999999") +
  geom_point(aes(color = if_weekend)) +
  labs(x = "", y = "Social screen time (min)") +
  scale_color_manual(values = c("weekday" = "#999999", "weekend" = "#FF99CC")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        legend.title = element_blank())

# Print the plot
total_scl_plot
```



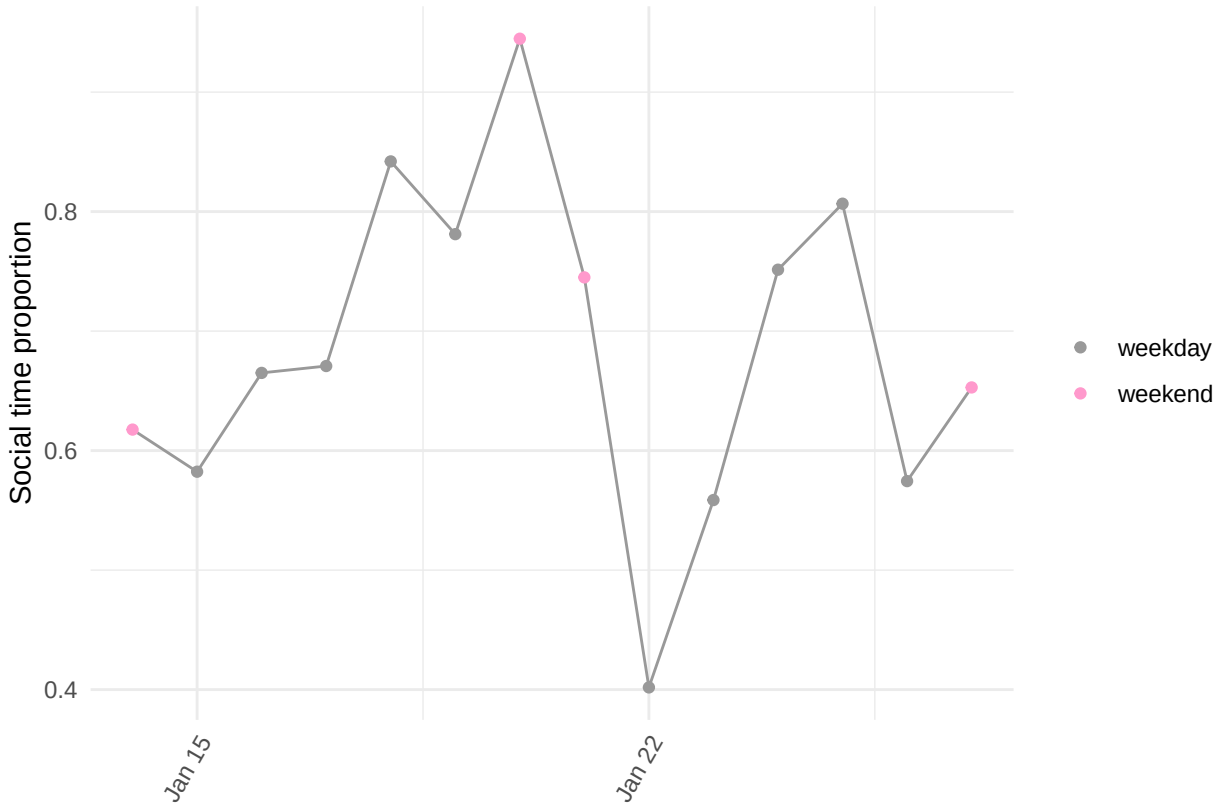
Time series plot of total pickups

```
# Create the plot
total_pu_plot <- ggplot(data, aes(x = Date, y = Pickups, group = 1)) +
  geom_line(color = "#999999") +
  geom_point(aes(color = if_weekend)) +
  labs(x = "", y = "Pickups") +
  scale_color_manual(values = c("weekday" = "#999999", "weekend" = "#FF99CC")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        legend.title = element_blank())

# Print the plot
total_pu_plot
```



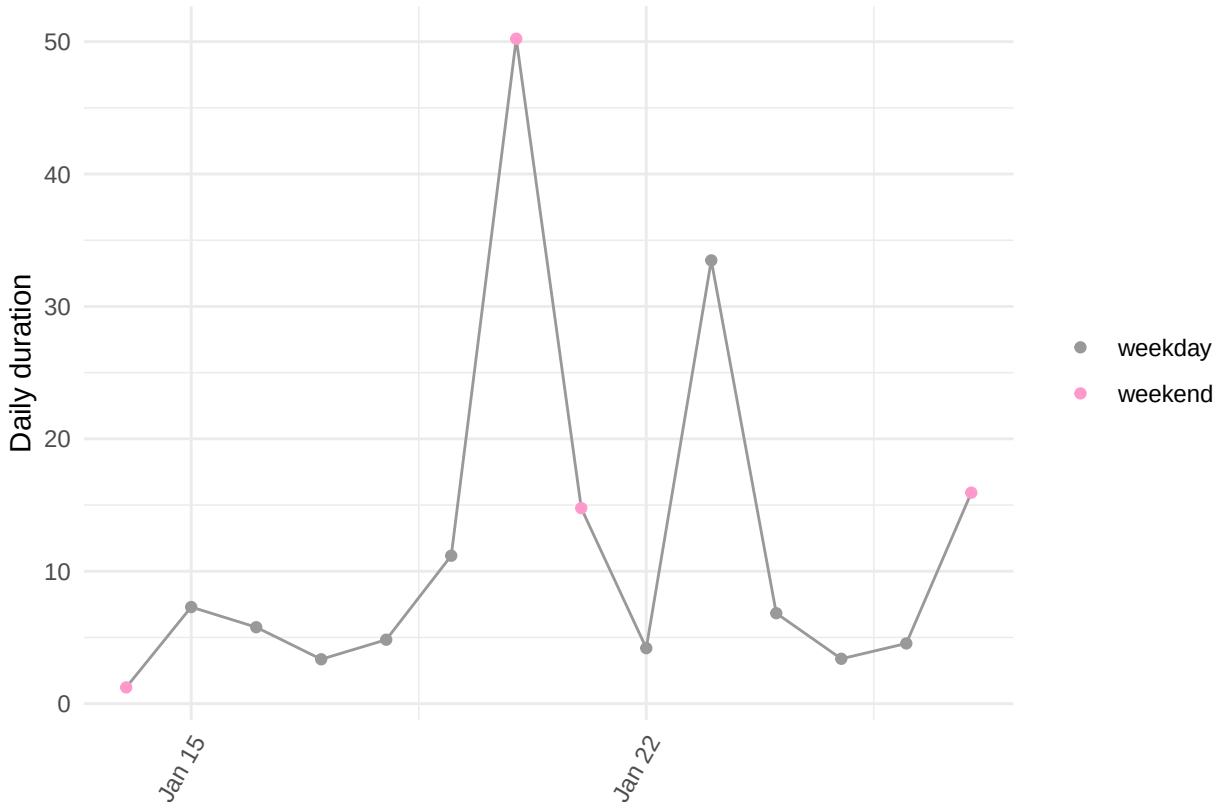
Time series plot of social screen time proportion

```r
# Create the plot
total_scl_ppt_plot <- ggplot(data, aes(x = Date, y = Daily.prop.social, group = 1)) +
  geom_line(color = "#999999") +
  geom_point(aes(color = if_weekend)) +
  labs(x = "", y = "Social time proportion") +
  scale_color_manual(values = c("weekday" = "#999999", "weekend" = "#FF99CC")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        legend.title = element_blank())

# Print the plot
total_scl_ppt_plot
```



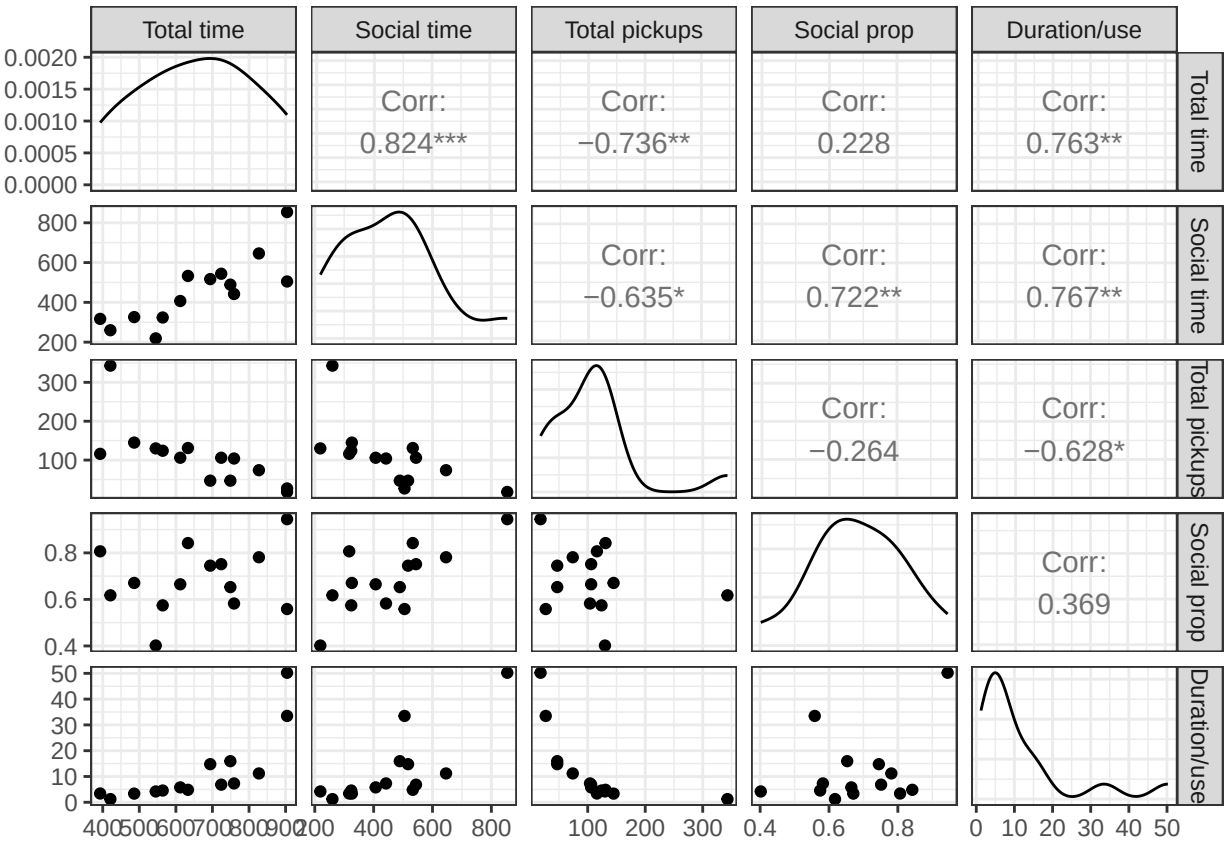
Time series plot of daily duration

```r
# Create the plot
total_duration_plot <- ggplot(data, aes(x = Date, y = Daily.duration, group = 1)) +
  geom_line(color = "#999999") +
  geom_point(aes(color = if_weekend)) +
  labs(x = "", y = "Daily duration") +
  scale_color_manual(values = c("weekday" = "#999999", "weekend" = "#FF99CC")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        legend.title = element_blank())

# Print the plot
total_duration_plot
```



(b) Pairwise

```
ggpairs(data, columns = c("Total.ST.min", "Social.ST.min", "Pickups", "Daily.prop.social",
                          "Daily.duration"),
        columnLabels = c("Total time", "Social time", "Total pickups", "Social prop",
                         "Duration/use")) + theme_bw()
```



The positive correlation between social screen time and the total screen time is the highest, with the Pearson correlation equal to 0.824. The positive correlations between the duration and total time, the proportion of social screen time and social screen time, as well as the duration and social screen time, are also significant. At the same time, there are significantly negative correlations between total pickups and total time, total pickups and social screen time, the proportion of social screen time and total pickups, as well as the duration and total pickups.

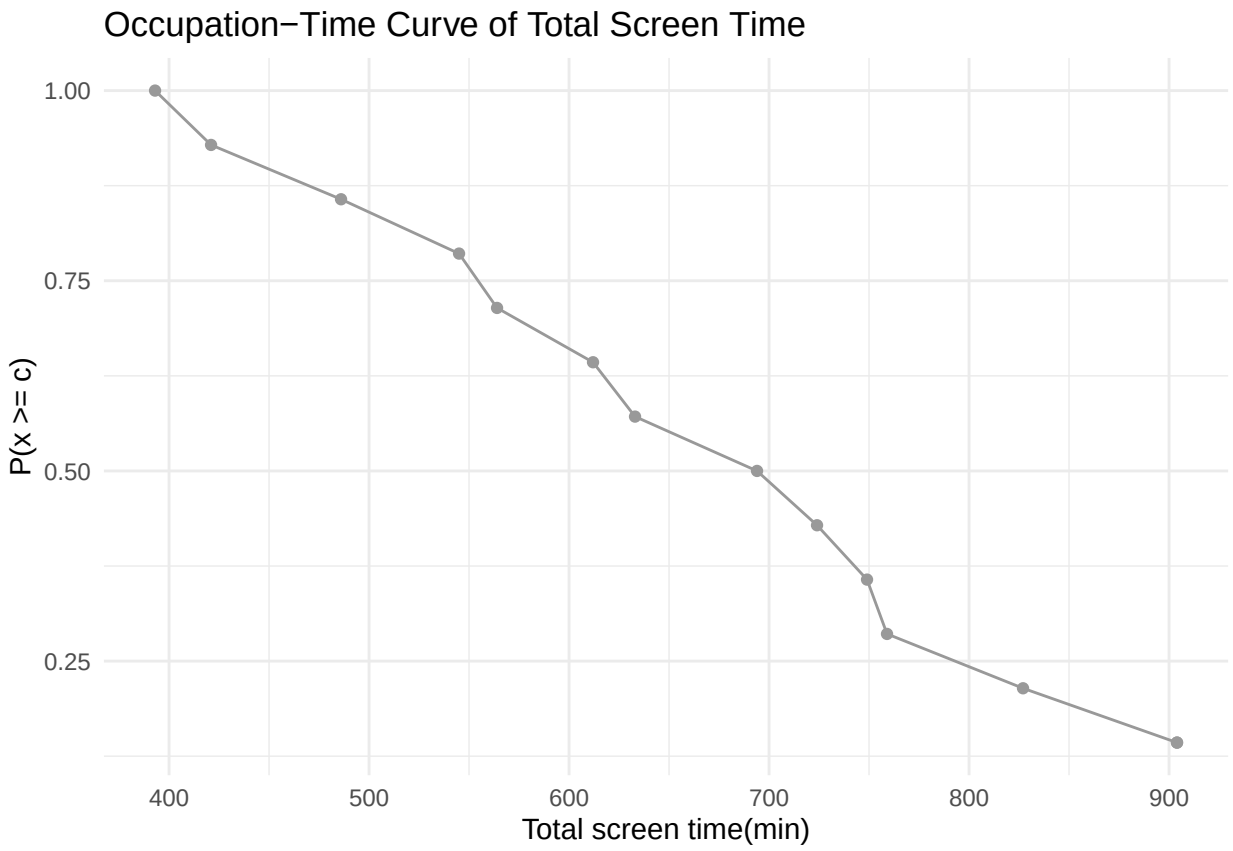
(c)

Occupation time curve of total screen time

There is no sharp decline in probability with increased screen time. This could suggest a more varied usage pattern in total screen time.

```
# Sort the data from smallest to largest
sorted_st <- sort(data$Total.ST.min, decreasing = TRUE)

# Generate a table of the frequencies of each value
value_counts <- table(sorted_st)

# Cumulative sum of the frequencies
cumulative_counts <- rev(cumsum(rev(value_counts)))

occupation_df <- data.frame(
  Vector_Magnitude = rep(as.numeric(names(value_counts)), value_counts),
  Probability = rep(cumulative_counts / length(sorted_st), value_counts)
)
```

```r
# Plot the curve
ggplot(occupation_df, aes(x = Vector_Magnitude, y = Probability)) +
  geom_point(color = "#999999") +
  geom_line(color = "#999999") +
  labs(title = "Occupation-Time Curve of Total Screen Time",
       x = "Total screen time(min)", y = "P(x >= c)") +
  theme_minimal()
```

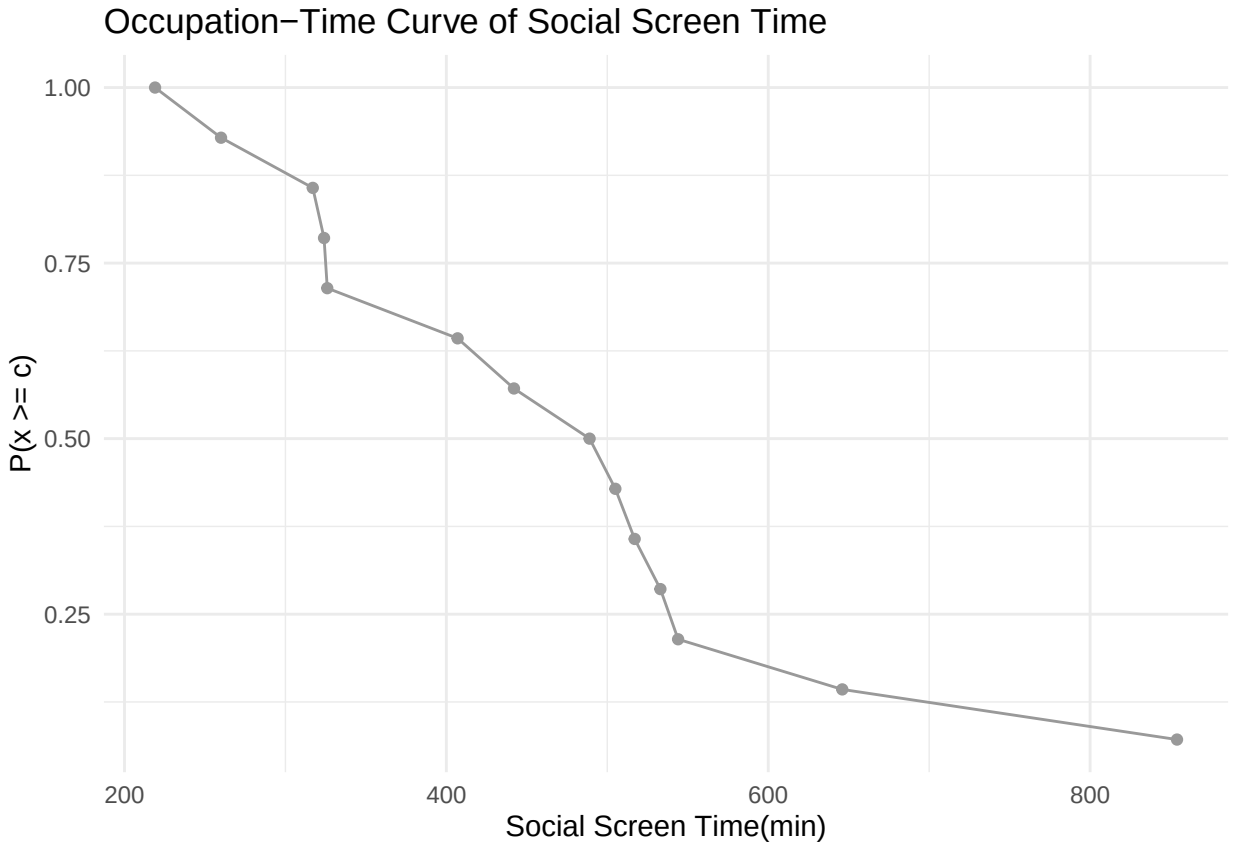


Occupation time curve of social screen time

The probability drops quickly as the social screen time increases when below 150. This suggests that most data points are clustered at lower values of social screen time, indicating that days with a very long social screen time (above 150) are less common.

```r
# Sort the data from smallest to largest
sorted_ss <- sort(data$Social.ST.min, decreasing = TRUE)

# Generate a table of the frequencies of each value
value_counts_ss <- table(sorted_ss)

# Cumulative sum of the frequencies
cumulative_counts <- rev(cumsum(rev(value_counts_ss)))

df_ss <- data.frame(
  Vector_Magnitude = rep(as.numeric(names(value_counts_ss)), value_counts_ss),
  Probability = rep(cumulative_counts / length(sorted_ss), value_counts_ss)
)
```

```r
# Plot the curve
ggplot(df_ss, aes(x = Vector_Magnitude, y = Probability)) +
  geom_point(color = "#999999") +
  geom_line(color = "#999999") +
  labs(title = "Occupation-Time Curve of Social Screen Time",
       x = "Social Screen Time(min)", y = "P(x >= c)") +
  theme_minimal()
```

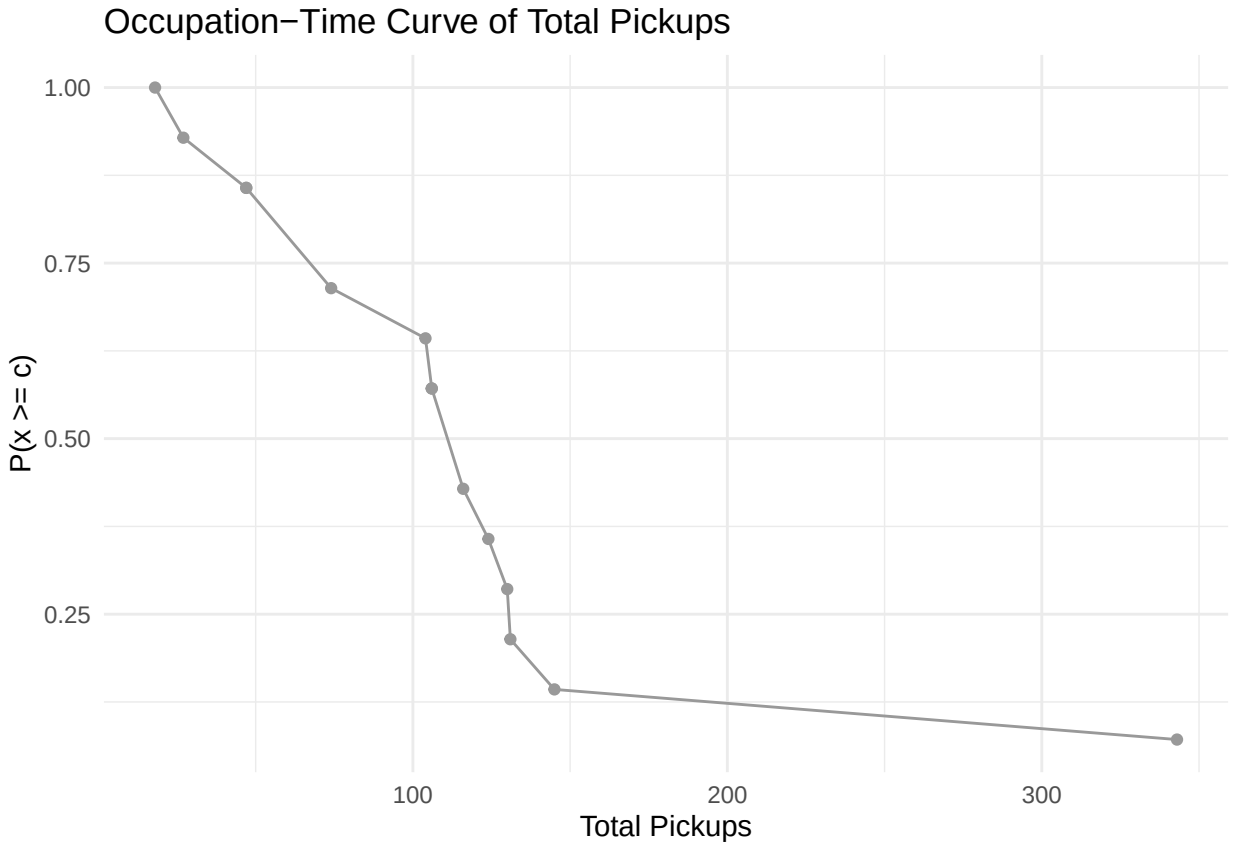


Occupation time curve of total pickups

The probability drops relatively quickly as the number of total pickups increases but below 600. This suggests that most data points are clustered at lower values (below 600) of total pickups, indicating that days with a very high number of pickups are less common.

```r
# Sort the data from smallest to largest
sorted_pu <- sort(data$Pickups, decreasing = TRUE)

# Generate a table of the frequencies of each value
value_counts_pu <- table(sorted_pu)

# Cumulative sum of the frequencies
cumulative_counts <- rev(cumsum(rev(value_counts_pu)))

df_pu <- data.frame(
  Vector_Magnitude = rep(as.numeric(names(value_counts_pu)), value_counts_pu),
  Probability = rep(cumulative_counts / length(sorted_pu), value_counts_pu)
)
```

```r
# Plot the curve
ggplot(df_pu, aes(x = Vector_Magnitude, y = Probability)) +
  geom_point(color = "#999999") +
  geom_line(color = "#999999") +
  labs(title = "Occupation-Time Curve of Total Pickups",
       x = "Total Pickups", y = "P(x >= c)") +
  theme_minimal()
```

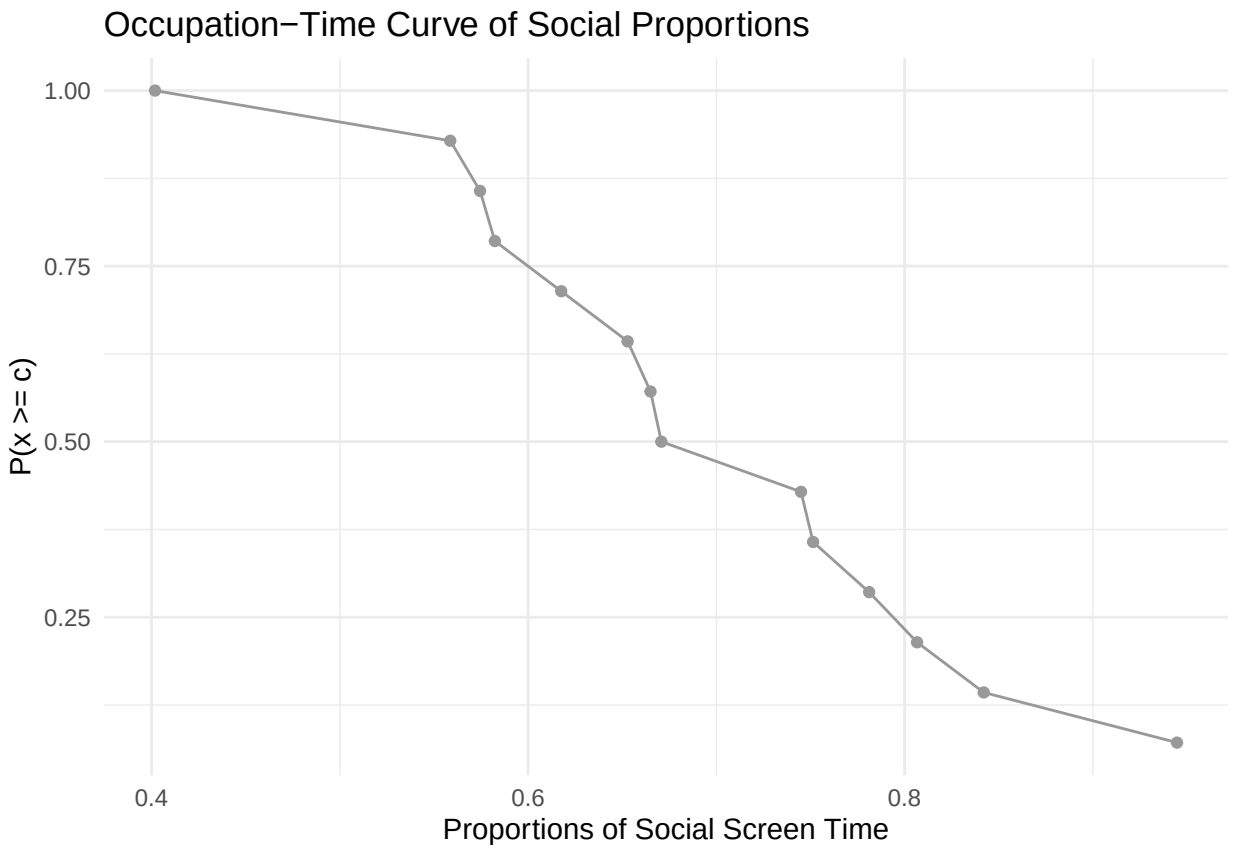


Occupation time curve of proportions of social screen time

There is no sharp decline in probability with increased proportion of screen time. This could suggest a more varied usage pattern in social screen time proportion.

```r
# Sort the data from smallest to largest
sorted_ss_prop <- sort(data$Daily.prop.social, decreasing = TRUE)

# Generate a table of the frequencies of each value
value_counts_ss_prop <- table(sorted_ss_prop)

# Cumulative sum of the frequencies
cumulative_counts <- rev(cumsum(rev(value_counts_ss_prop)))

df_ss_prop <- data.frame(
  Vector_Magnitude = rep(as.numeric(names(value_counts_ss_prop)), value_counts_ss_prop),
  Probability = rep(cumulative_counts / length(sorted_ss_prop), value_counts_ss_prop)
)
```

```r
# Plot the curve
ggplot(df_ss_prop, aes(x = Vector_Magnitude, y = Probability)) +
  geom_point(color = "#999999") +
  geom_line(color = "#999999") +
  labs(title = "Occupation-Time Curve of Social Proportions",
       x = "Proportions of Social Screen Time", y = "P(x >= c)") +
  theme_minimal()
```

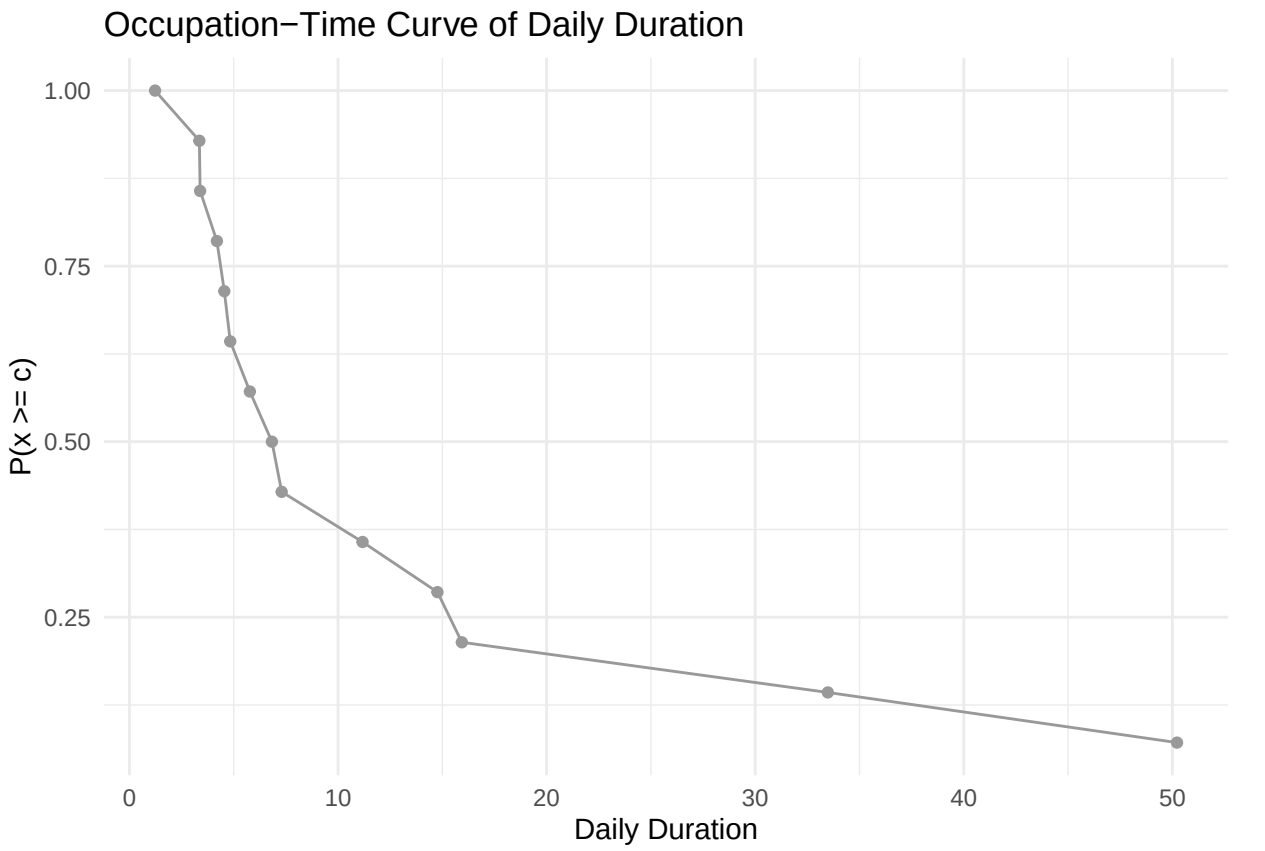


Occupation time curve of daily duration

The probability drops quickly as the duration increases below 20. This suggests that most data points are clustered at shorter duration (below 20), indicating that days with a very high duration are less common.

```r
# Sort the data from smallest to largest
sorted_dura <- sort(data$Daily.duration, decreasing = TRUE)

# Generate a table of the frequencies of each value
value_counts_dura <- table(sorted_dura)

# Cumulative sum of the frequencies
cumulative_counts <- rev(cumsum(rev(value_counts_dura)))

df_dura <- data.frame(
  Vector_Magnitude = rep(as.numeric(names(value_counts_dura)), value_counts_dura),
  Probability = rep(cumulative_counts / length(sorted_dura), value_counts_dura)
)
```

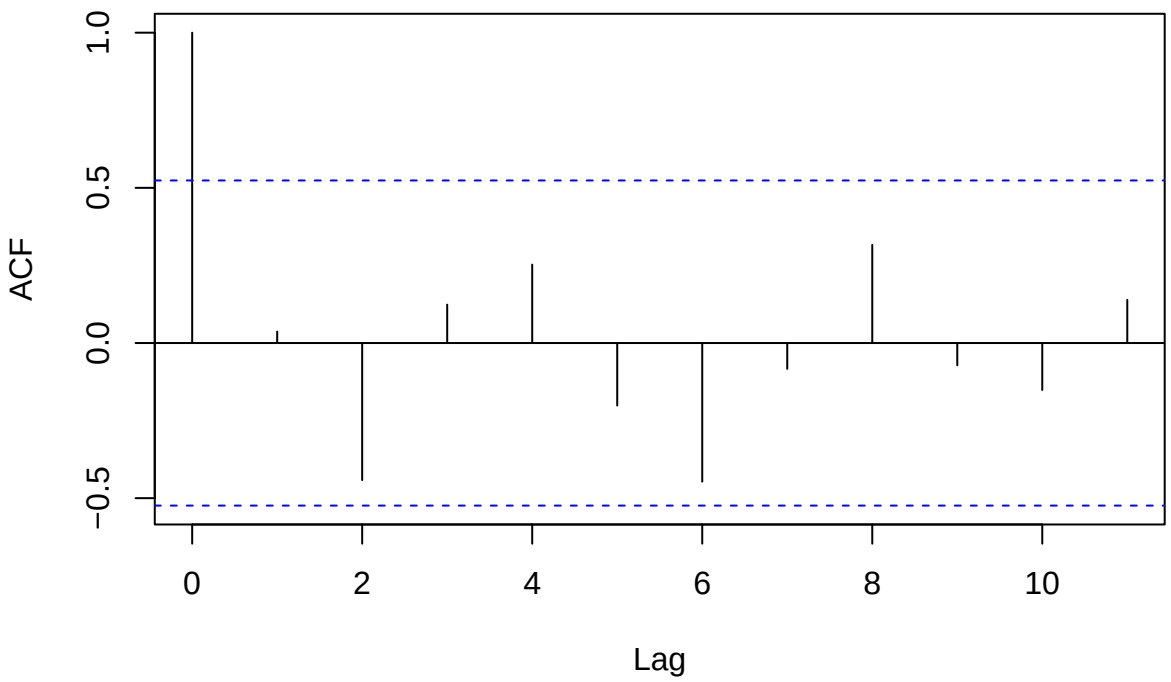
```r
# Plot the curve
ggplot(df_dura, aes(x = Vector_Magnitude, y = Probability)) +
  geom_point(color = "#999999") +
  geom_line(color = "#999999") +
  labs(title = "Occupation-Time Curve of Daily Duration",
       x = "Daily Duration", y = "P(x >= c)") +
  theme_minimal()
```



(d)

All variables demonstrated no meaningful autocorrelation as all observed vertical bars remained within the boundaries of the 95% confidence intervals. This suggests that daily records for these variables are statistically independent from preceding days' data.

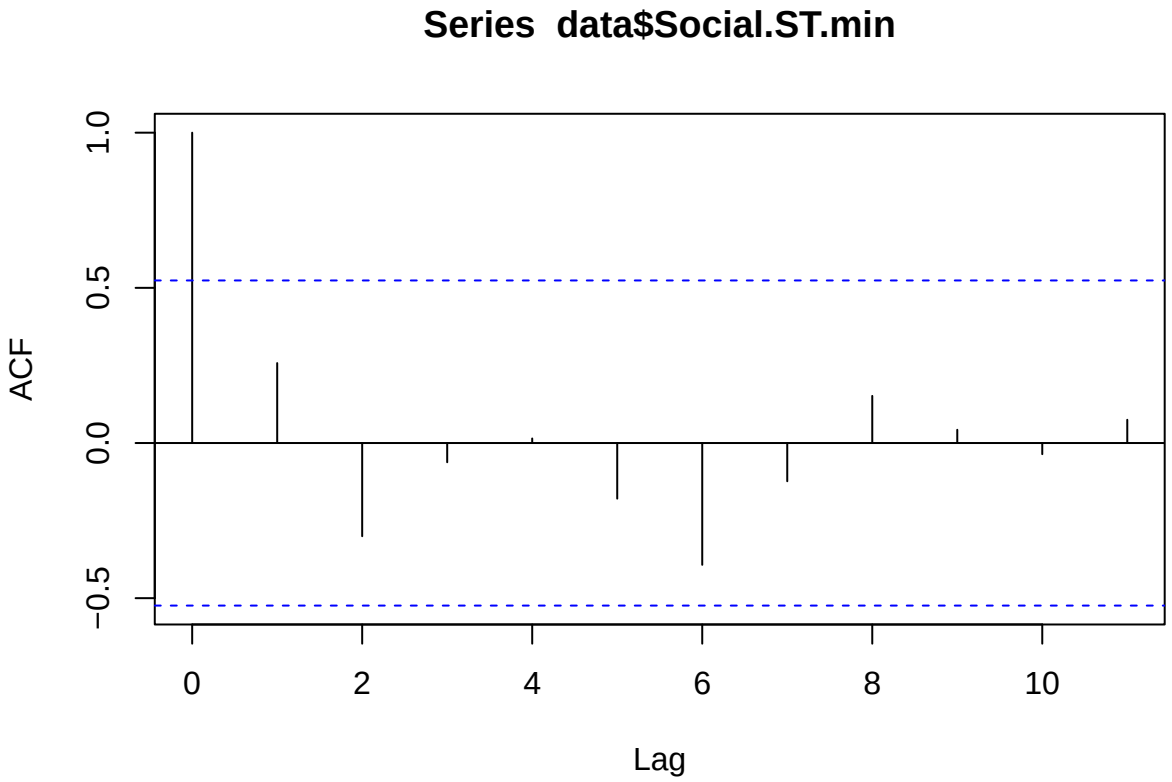
```r
# Apply the acf function for total screen time
acf(data$Total.ST.min)
```




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

```
##
## Autocorrelations of series 'data$Total.ST.min', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.037 -0.442  0.124  0.253 -0.202 -0.447 -0.083  0.316 -0.072 -0.151
##     11
##  0.139
```

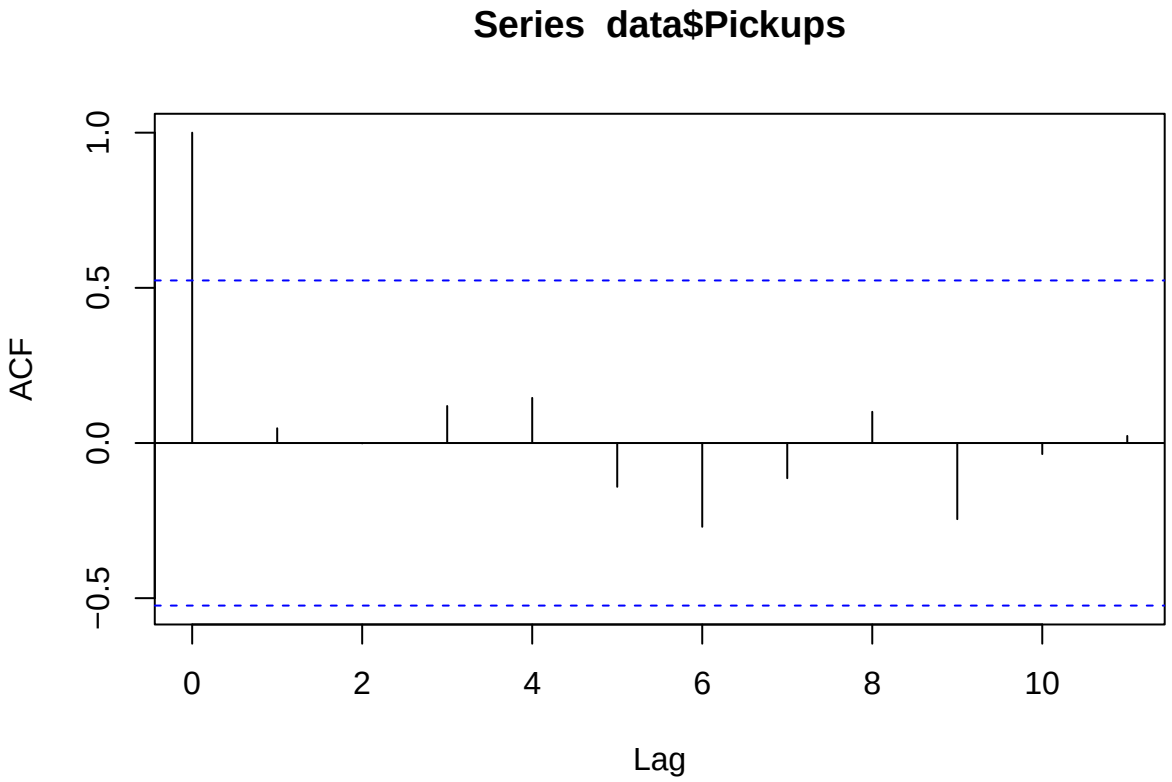
```r
# Apply the acf function for social screen time
acf(data$Social.ST.min)
```




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

```
##
## Autocorrelations of series 'data$Social.ST.min', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.258 -0.300 -0.062  0.015 -0.179 -0.393 -0.123  0.152  0.043 -0.036
##     11
##  0.075
```

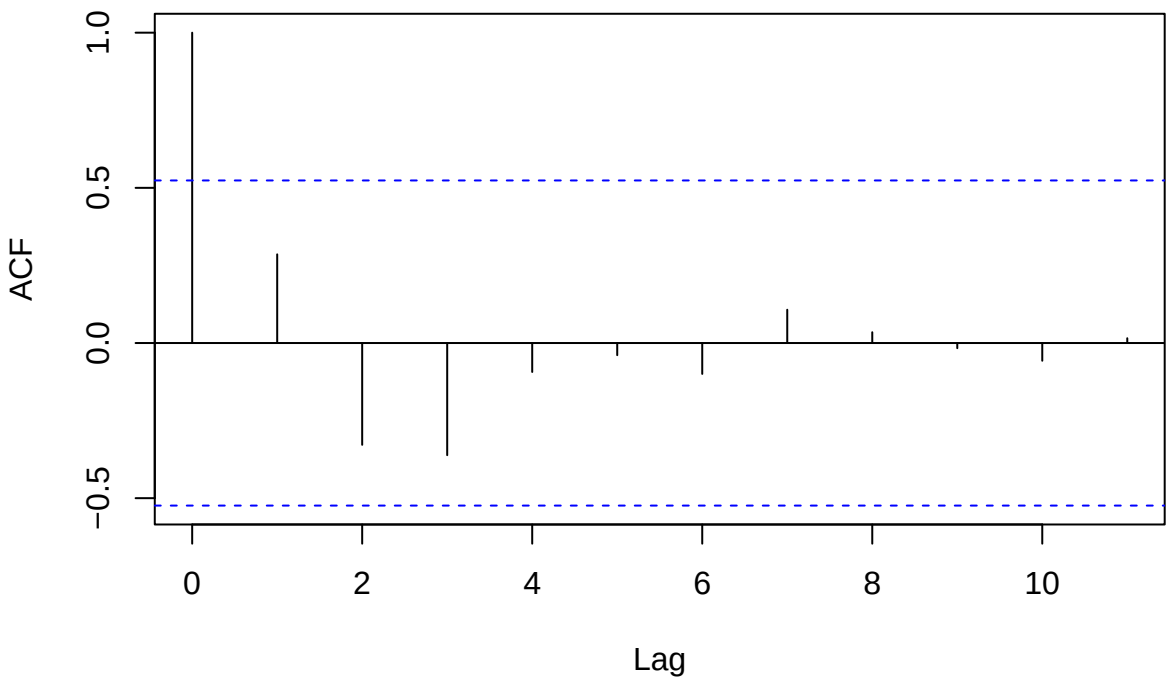
```
# Apply the acf function for total pickups
acf(data$Pickups)
```




```
acf(data$Pickups, plot = FALSE)
```

```
##
## Autocorrelations of series 'data$Pickups', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.048 -0.001  0.119  0.145 -0.141 -0.270 -0.113  0.101 -0.245 -0.036
##     11
##  0.023
```

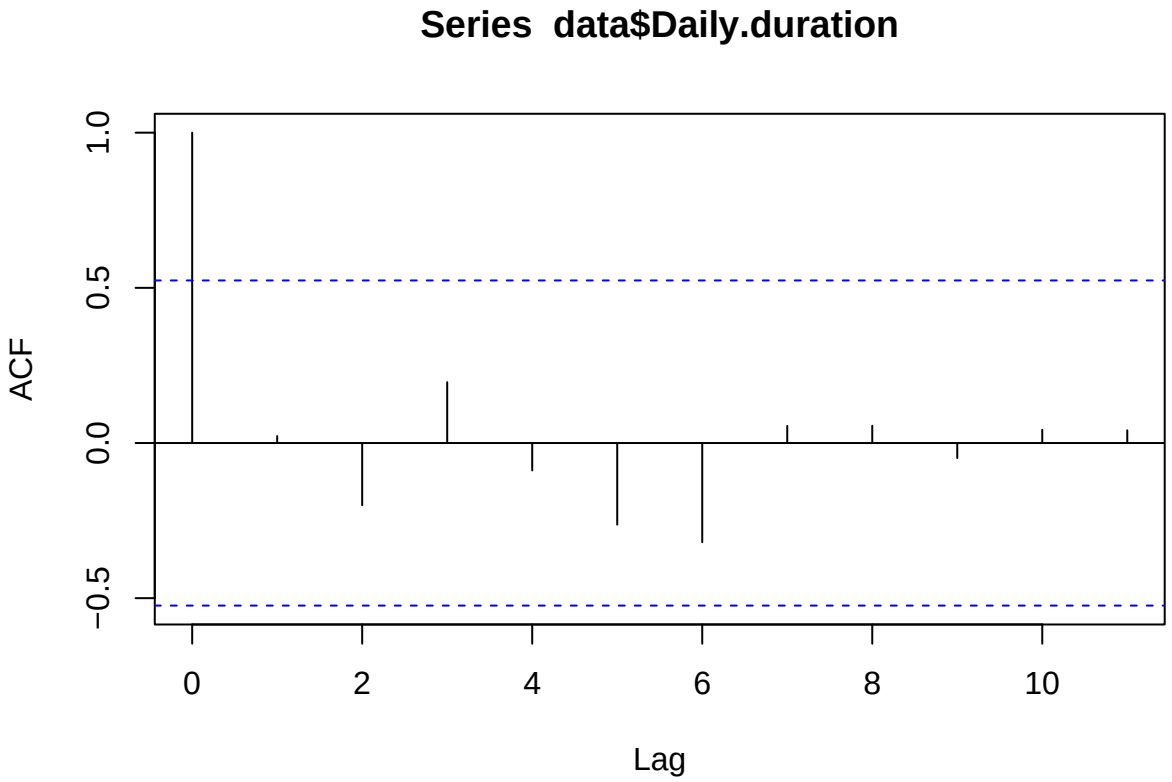
```
# Apply the acf function for total daily proportion of social screen time
acf(data$Daily.prop.social)
```




```
acf(data$Daily.prop.social, plot = FALSE)
```

```
##
## Autocorrelations of series 'data$Daily.prop.social', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.286 -0.328 -0.361 -0.093 -0.039 -0.100  0.107  0.035 -0.017 -0.057
##     11
##  0.015
```

```
# Apply the acf function for total daily duration
acf(data$Daily.duration)
```




```
acf(data$Daily.duration, plot = FALSE)
```

```
##
## Autocorrelations of series 'data$Daily.duration', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
##  1.000  0.022 -0.200  0.196 -0.088 -0.263 -0.320  0.055  0.056 -0.048  0.043
##     11
##  0.041
```

Problem 3

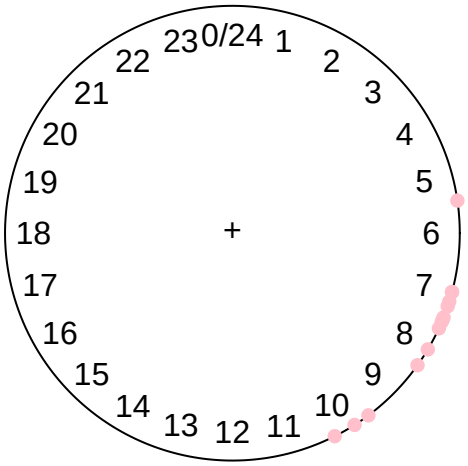
(a)

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

(b)

The initial phone usage typically occurs between 7 and 8 in the morning. There are occasional first pickups noted between 9 and 11 a.m., with an outlier around 5 a.m. This pattern suggests that the individual generally tends to start their day early, with some fluctuations in their morning routine.

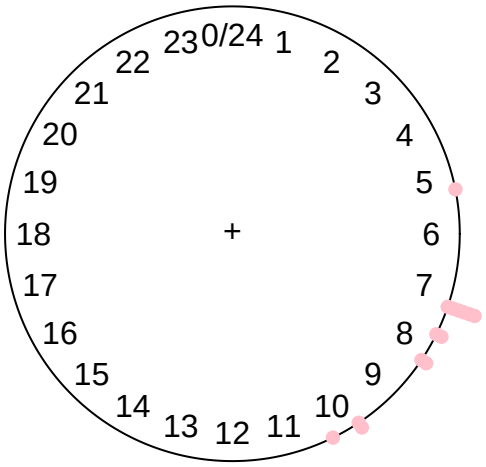
```
pickup_c = circular(data$Pickup.1st.agl, units = "degrees", template = "clock24")
plot(pickup_c, col = "pink")
```



(c)

With 48 bins across a 24-hour period, each bin represents a 30-minute interval. This level allows for a detailed analysis of the variations in the first pickup times, which can capture the early morning activities and distinguish the less frequent late morning pickups. The graph shows that the first pickup times mostly lie in 7am to 7:30 am.

```
plot(pickup_c, stack = TRUE, bins = 48, col = "pink")
```



Problem 4

(a)

Lambda is the expected hourly rate of pickups, which depends on the total screen time of that day, since it is more likely to have a higher pickup times when the total screen time is higher in a given day. Therefore, St is needed to act as a scale of lambda to address this difference.

(b)

```r
# Convert total screen time from minutes to hours
data$ST_hrs <- data$Total.ST.min / 60

# Fit the model
model1 <- glm(Pickups ~ offset(log(ST_hrs)), family = poisson, data = data)

summary(model1)
```

```
##
## Call:
## glm(formula = Pickups ~ offset(log(ST_hrs)), family = poisson,
##     data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.29091    0.02567   89.26   <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: 1146.2  on 13  degrees of freedom
## Residual deviance: 1146.2  on 13  degrees of freedom
## AIC: 1236.3
##
## Number of Fisher Scoring iterations: 5
```

```r
# Extract lambda
lambda <- exp(coef(model1)["(Intercept)"])

print(lambda)
```

```
## (Intercept)
##    9.883885
```

(c)

```r
# Create dummy variables
data$Xt <- as.numeric(!weekdays(data$Date) %in% c("Saturday", "Sunday"))
data$Zt <- as.numeric(data$Date >= as.Date("2021-01-16"))
```

```r
# Repeat part (b)
model2 <- glm(Pickups ~ offset(log(ST_hrs)) + Xt + Zt,
              family = poisson, data = data)

summary(model2)
```

```
##
## Call:
## glm(formula = Pickups ~ offset(log(ST_hrs)) + Xt + Zt, family = poisson,
##     data = data)
##
## Coefficients: (1 not defined because of singularities)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.288762   0.046881  48.821   <2e-16 ***
## Xt          0.003063   0.056023   0.055    0.956
## 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: 1146.2  on 13  degrees of freedom
## Residual deviance: 1146.2  on 12  degrees of freedom
## AIC: 1238.3
##
## Number of Fisher Scoring iterations: 6
```

c.1

Since the p-value is larger than 0.05, there's no enough evidence to reject the null hypothesis that there is no difference in the number of pickups between weekdays and weekends at the 5% significance level. Therefore, there is no enough evidence to conclude that there is a significant difference in the number of pickups between weekdays and weekends.

c.2

Since the p-value is less than 0.05, there is enough evidence to reject the null hypothesis that there is no significant change in the behavior of daily pickups after the winter semester began. Therefore, we have enough evidence to support that there is significant change in the behavior of daily pickups after the winter semester began

Problem 5

(a)

```
# Convert first pickups to radians
data = data %>%
  mutate(Pickup.1st.rad = ((hour(Pickup.1st) * 60 + minute(Pickup.1st)) / (24 * 60)) * (2 * pi))

# Convert first pickups to a circular object
pickups_c_rad <- circular(data$Pickup.1st.rad, units = "radians", template = "none")

# Estimate the parameters
vm_est <- mle.vonmises(pickups_c_rad)

print(vm_est)
```

```
##
## Call:
## mle.vonmises(x = pickups_c_rad)
##
## mu: 2.055  ( 0.08613 )
##
## kappa: 10.14  ( 3.726 )
```

(b)

```
# Convert 8:30 AM radians
t <- (8.5 / 24) * (2 * pi)

# Calculate the probability
p_t <- pvonmises(t, vm_est$mu, vm_est$kappa)
```

```
## 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.circularqradians0counter
```

```
print(p_t)
```

```
## [1] 0.7037814
```

Reference

Mikkelsen, Kathleen, Lily Stojanovska, Momir Polenakovic, Marijan Bosevski, and Vasso Apostolopoulos. 2017. "Exercise and Mental Health." *Maturitas* 106: 48–56. https://doi.org/10.1016/j.maturitas.2017.09.003.

Penglee, Nop, Richard W Christiana, Rebecca A Battista, and Eli Rosenberg. 2019. "Smartphone Use and Physical Activity Among College Students in Health Science-Related Majors in the United States and Thailand." *International Journal of Environmental Research and Public Health* 16 (8): 1315. https://doi.org/10.3390/ijerph16081315.

Scott, Alexander J, Thomas L Webb, Marrissa Martyn-St James, Georgina Rowse, and Scott Weich. 2021. "Improving Sleep Quality Leads to Better Mental Health: A Meta-Analysis of Randomised Controlled Trials." *Sleep Medicine Reviews* 60: 101556. https://doi.org/10.1016/j.smrv.2021.101556.

Xu, Fengqing, Swann Arp Adams, Steven A Cohen, Jessica E Earp, and Mary L Greaney. 2019. "Relationship Between Physical Activity, Screen Time, and Sleep Quantity and Quality in US Adolescents Aged 16–19." *International Journal of Environmental Research and Public Health* 16 (9): 1524. https://doi.org/10.3390/ijerph16091524.