# HW1

Liancheng Lu

2024-01-28

GitHub Link: https://github.com/lianchenglu/BIOSTAT620_HW1

## PATT I: DATA COLLECTION AND DATA PROCESSING

### 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.**

Our purpose is to explore the relationship between screen time and the physical and mental health of adolescents. We will analyze the relationship between screen time and physical health indicators, explore the correlation between screen time and mental health, and examine the association between screen time and social behavior and cognitive development.

Hypothesis: Reducing screen time in children and adolescents will significantly improve their physical and mental health, in the form of improved sleep quality, reduced risk factors for cardiovascular disease, reduced manifestations of depression and externalizing behaviors, and improved social coping and concentration skills.

We can use the time we wake up each day as an indicator to assess the regularity and quality of sleep, and use screen time and screen pickup times to reflect the degree of personal dependence on mobile phones, which may be related to anxiety and sleep problems. In addition, the amount of time spent on social software may affect people's mental state and sleep quality. For example, overuse may make it difficult to relax before bed and interfere with sleep.

Reference: Lissak G. (2018). Adverse physiological and psychological effects of screen time on children and adolescents: Literature review and case study. Environmental research, 164, 149–157. https://doi.org/10.1016/j.envres.2018.01.015

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

In out study, informed consent ensured that participants were clear about how information such as their screen usage data, wake-up times and phone unlock times would be collected and used. It informs participants

that the data will be used to study the link between their behavior and health, and ensures that the process is voluntary, while also protecting their privacy and data security.

**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.**

| **Start Date** | **Data Freeze Date:** | **Data Points Collected** | **Source of Data** |
|---|---|---|---|
| 2024-01-16 | 2024-02-15 | Daily Screen Time, Daily Social Media Time, Daily First Pickup Time (Wake-up Time), Daily Pickups | Students enrolled in BIOSTAT620 |
| **Number of Participants:** 3 | | | |

Table 1: Data Collection Plan

**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).**

```
convert_to_minutes <- function(time) {
  if (!grepl("h", time)) {
    return(as.numeric(sub("m", "", time)))
  }
  parts <- strsplit(time, "h|m")[[1]]
  as.numeric(parts[1]) * 60 + as.numeric(parts[2])
}
df$Total.ST.min <- sapply(df$Total.ST, convert_to_minutes)
df$Social.ST.min <- sapply(df$Social.ST, convert_to_minutes)
df$prop_ST <- df$Social.ST.min / df$Total.ST.min
df$duration_per_use <- df$Total.ST.min / df$Pickups
head(df[,9:10])
```

```
## # A tibble: 6 x 2
##   prop_ST duration_per_use
##     <dbl>            <dbl>
## 1   0.432             1.53
## 2   0.495             1.77
## 3   0.449             2.23
## 4   0.584             2.68
## 5   0.607             3.33
## 6   0.625             6.21
```

**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.**

Daily total screen time: There is a significant peak in screen time on January 21st, which suggests a higher usage of the device during the weekend. This is followed by a sharp decline on January 22nd, indicating a return to a more routine or restricted use during the weekdays.

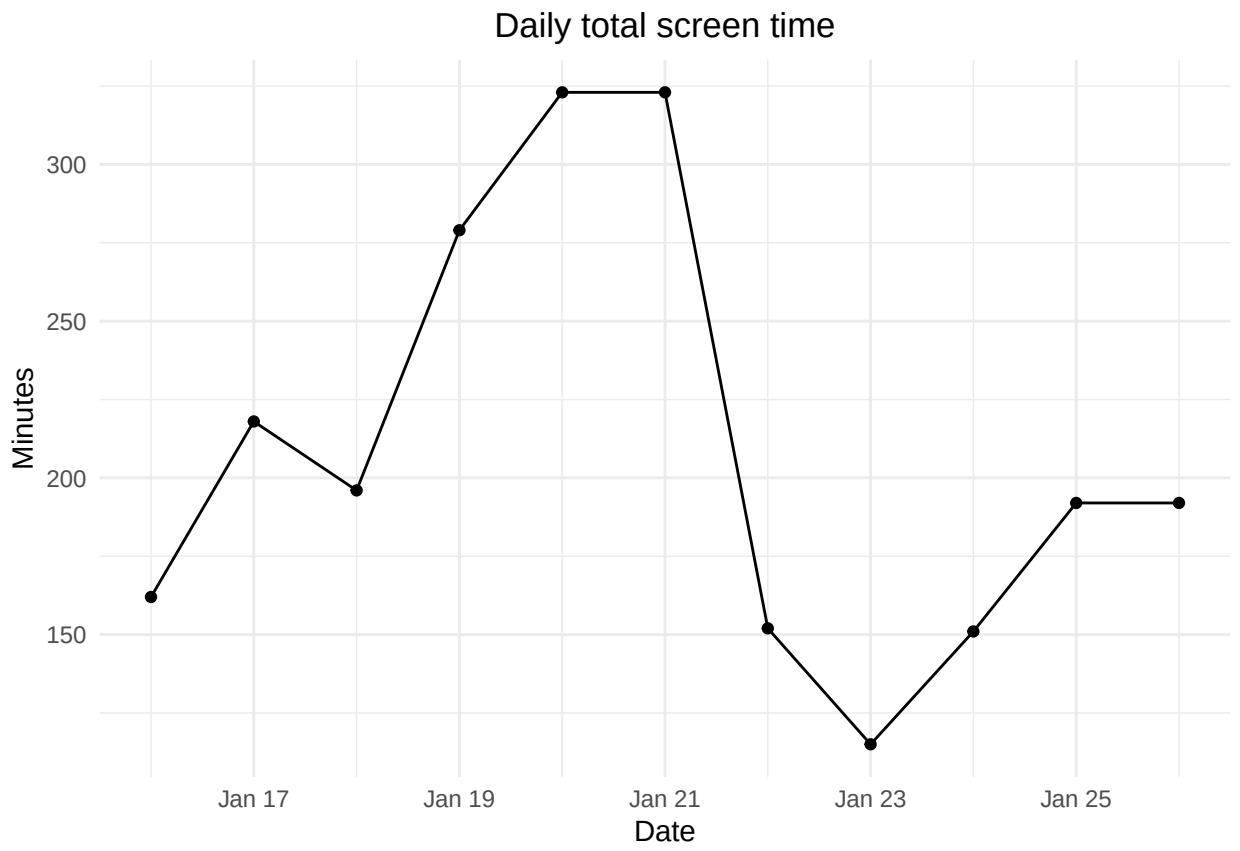
Daily Total Social Screen Time: Similar to the total screen time, social media usage peaks during the weekend, particularly on January 21st. This might reflect more leisure time available for social media engagement during non-working days.

Daily Number of Pickups: The number of times the device is picked up dips lowest on January 22nd, right after the weekend, and then shows an upward trend throughout the week. The highest number of pickups happens towards the end of the observed period, indicating an increase in the frequency of interaction with the device.

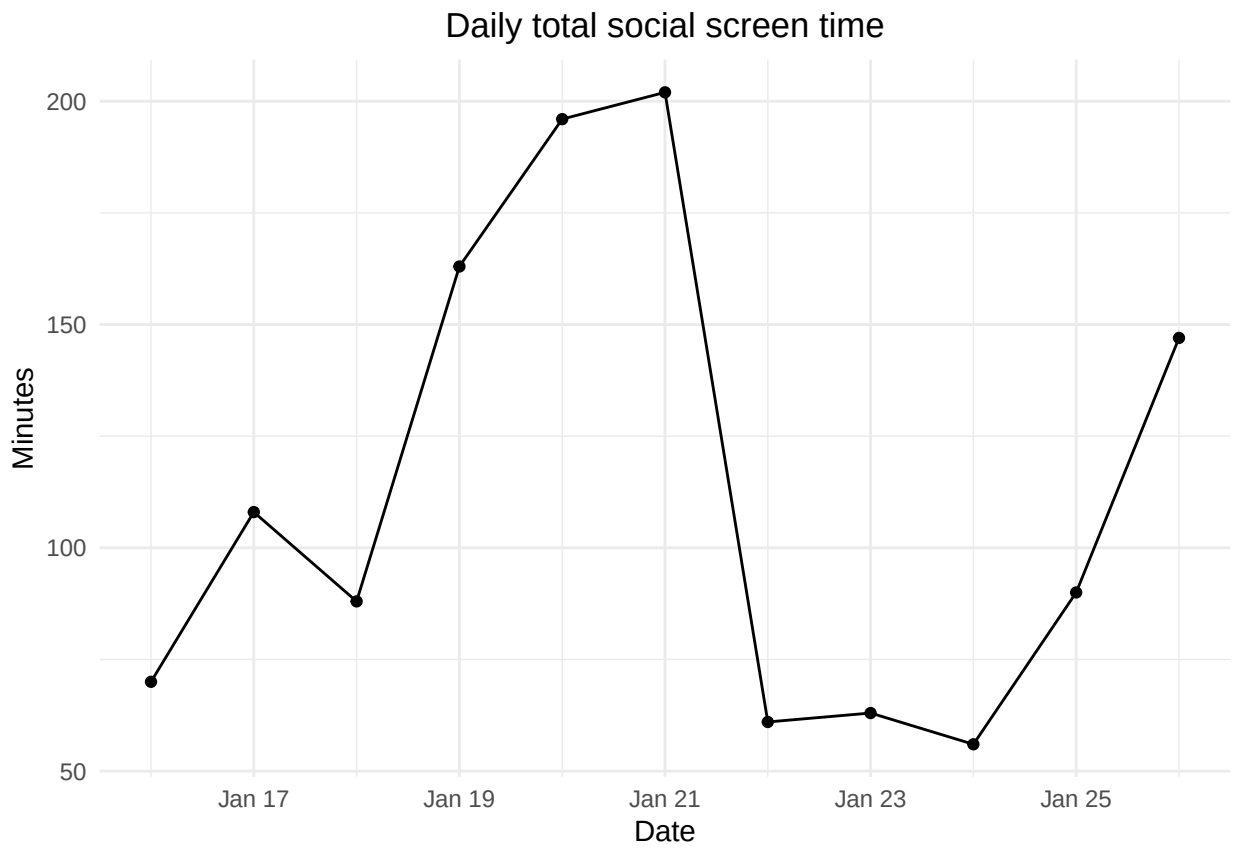
Daily Proportion of Social Screen Time: There's a noticeable fluctuation in the proportion of social media time to total screen time, with the highest proportion occurring on January 25th. This suggests a day with a particularly high engagement in social media compared to other uses of the device.

Daily Duration Per Use: On January 21st, there is a spike in the duration per use, which is the highest of all the observed days. This suggests that on this day, each interaction with the device was longer, consistent with the pattern of increased usage on weekends.

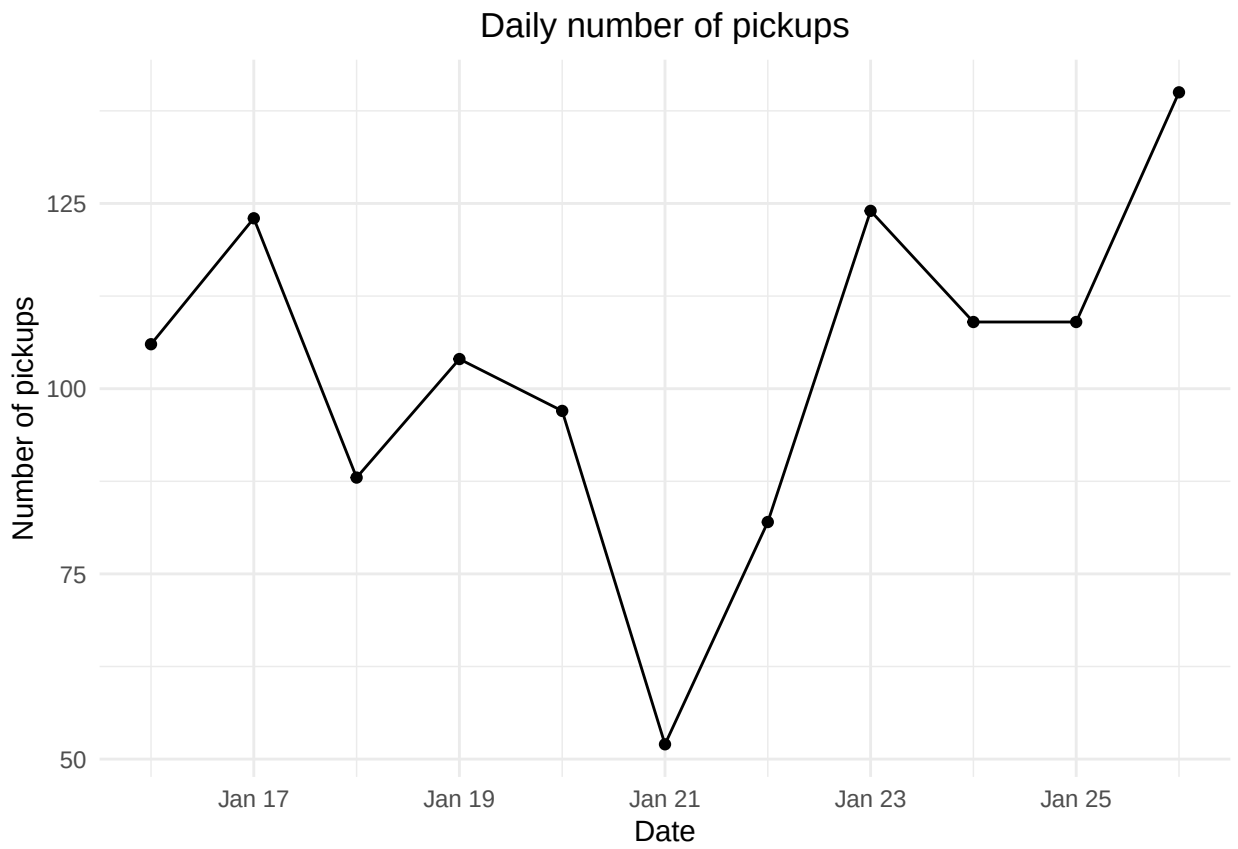
```
# Daily total screen time
ggplot(data = df, aes(x = Date, y = Total.ST.min)) +
  geom_line() +
  geom_point() +
  labs(title = "Daily total screen time",
       x = "Date",
       y = "Minutes") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



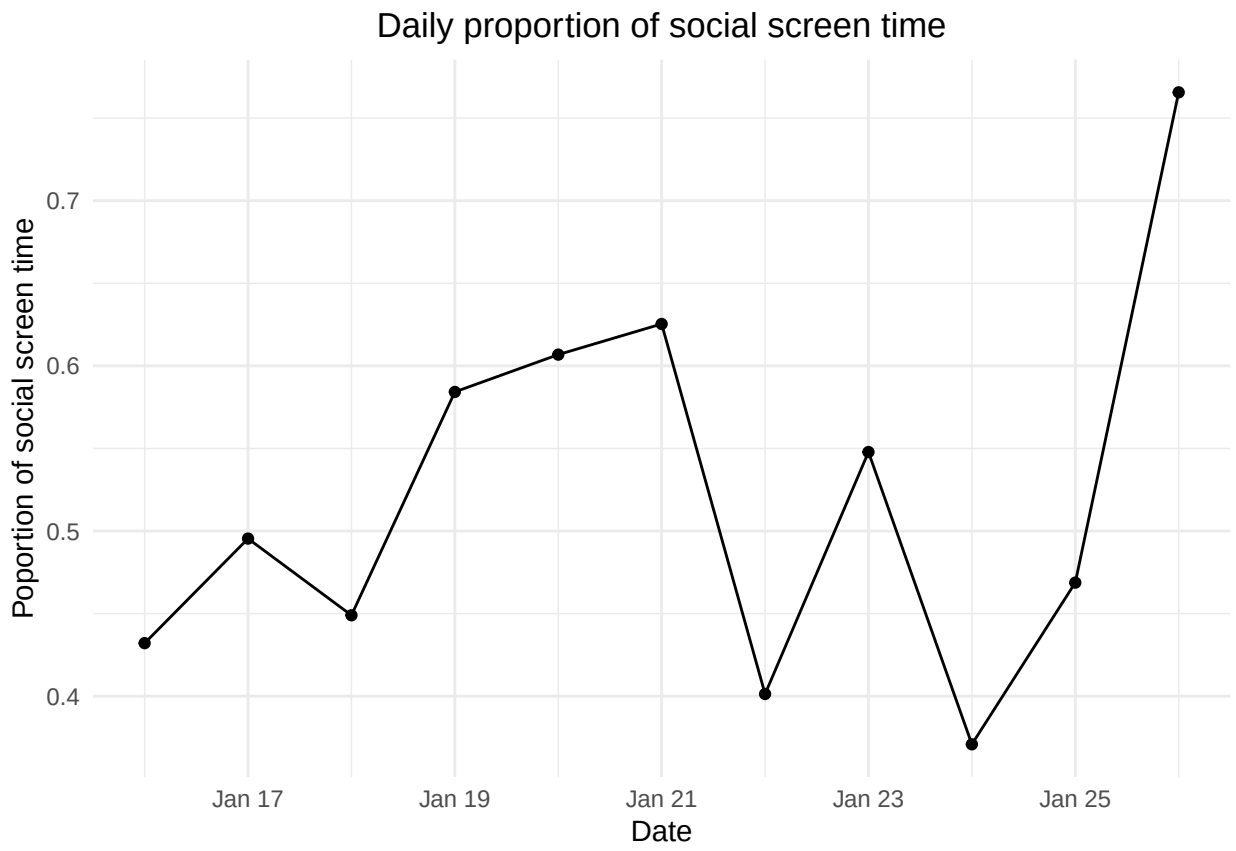

```
# Daily total social screen time
ggplot(data = df, aes(x = Date, y = Social.ST.min)) +
  geom_line() +
  geom_point() +
  labs(title = "Daily total social screen time",
       x = "Date",
       y = "Minutes") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



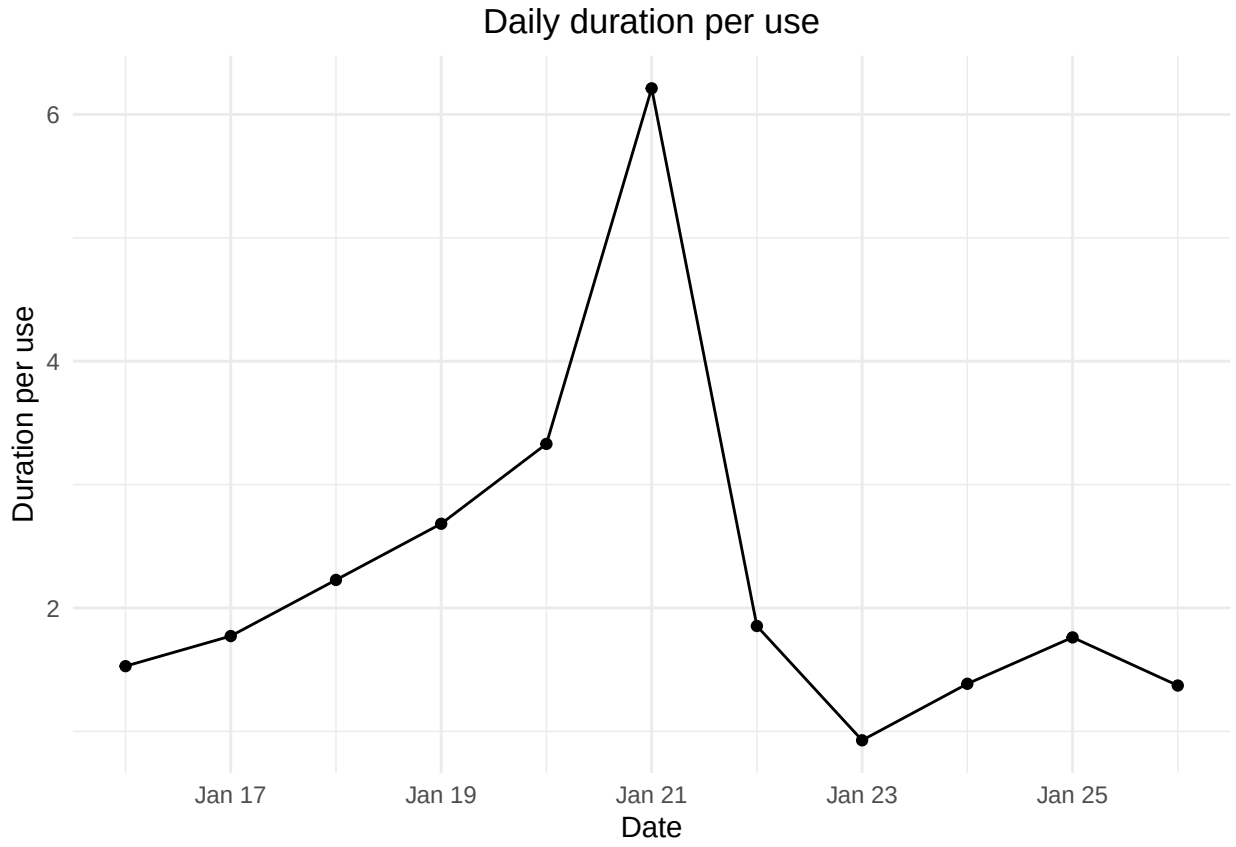

```
# Daily number of pickups
ggplot(data = df, aes(x = Date, y = Pickups)) +
  geom_line() +
  geom_point() +
  labs(title = "Daily number of pickups",
       x = "Date",
       y = "Number of pickups") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Daily proportion of social screen time
ggplot(data = df, aes(x = Date, y = prop_ST)) +
  geom_line() +
  geom_point() +
  labs(title = "Daily proportion of social screen time",
       x = "Date",
       y = "Poportion of social screen time") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Daily duration per use
ggplot(data = df, aes(x = Date, y = duration_per_use)) +
  geom_line() +
  geom_point() +
  labs(title = "Daily duration per use",
       x = "Date",
       y = "Duration per use") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




**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?**

There is a strongest positive correlation (Corr: 0.941***) between total screen time (Total.ST.min) and social screen time (Social.ST.min), indicating that as total screen time increases, social screen time also increases proportionally.

This is the highest positive correlation observed among the pairs, suggesting that social media use is a significant component of overall screen time.

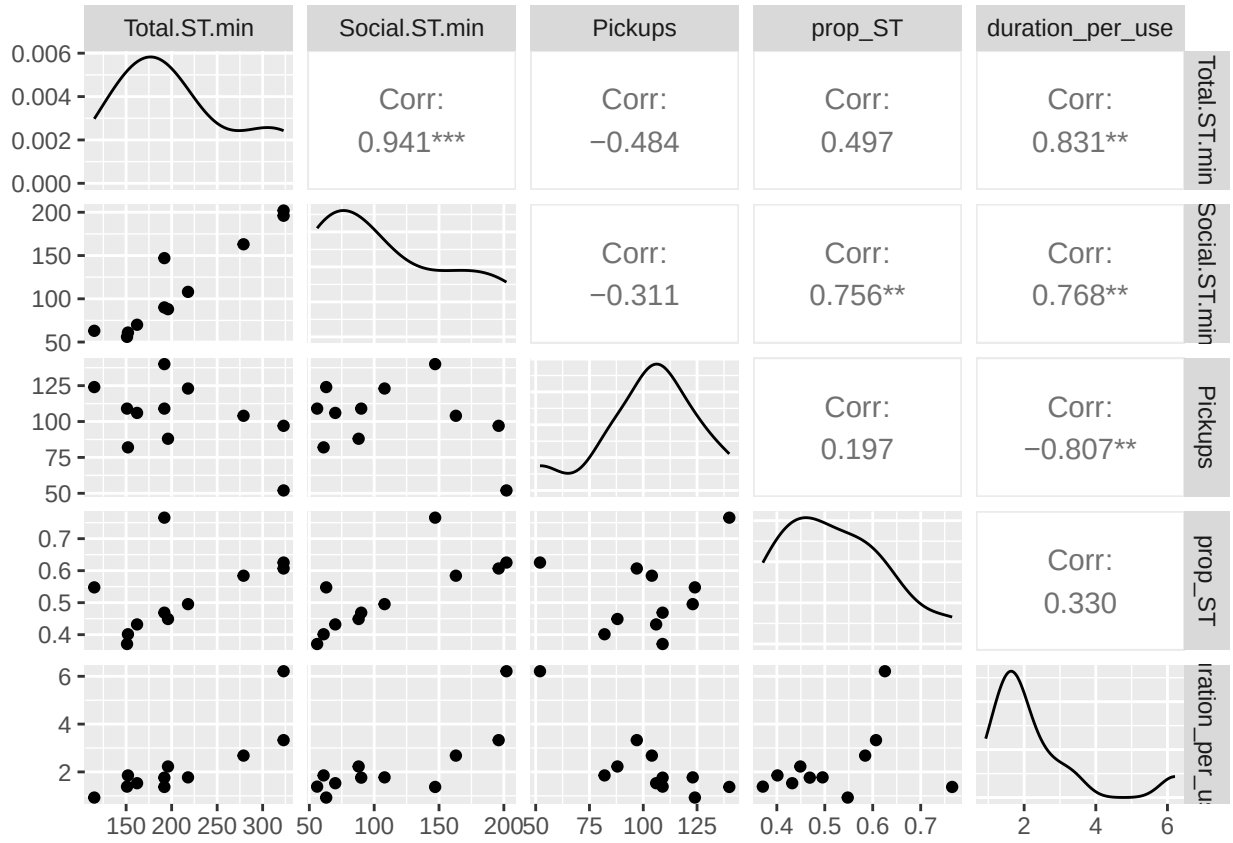
In addition, total screen time has a strong positive correlation with Duration Per Use (duration_per_use) (Corr: 0.831**), suggesting that longer periods of screen engagement are associated with increased total screen time. The relationship between total screen time and the Proportion of Social Screen Time (prop_ST) is moderately positive (Corr: 0.497), indicating that days with more screen time also tend to have a higher proportion of that time devoted to social media. Conversely, the correlation between total screen time and the Number of Pickups (Pickups) is moderately negative (Corr: -0.484), which might imply that on days with more overall screen usage, users tend to pick up their devices less frequently, possibly due to longer durations of continuous use.

Following the same trend, the Duration Per Use also has a moderate positive correlation with Social Screen Time (Corr: 0.768**), implying that longer usage sessions are often linked to social media activity. Furthermore, there is a moderate positive correlation between Social Screen Time and the Proportion of Social Screen Time (prop_ST) (Corr: 0.756**), reinforcing the idea that social media occupies a substantial portion of the time spent on screens.

On the contrary, the Number of Pickups shows a strong negative correlation with Duration Per Use (Corr: -0.807**), suggesting that more frequent interactions with the device are associated with shorter usage

periods. This might reflect a pattern of checking the device often but for brief durations, such as glancing at notifications rather than engaging in longer activities.

```
df2 <- df[,c(1,3,5,6,9,10)]
ggpairs(df2[,-1])
```




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

Total Screen Time: The curve will show the probability of screen time exceeding different thresholds. The curve drops sharply as the threshold increases, especially when magnitude c close to 200, it means that high levels of screen time are less common.

Social Screen Time: This curve will similarly show the probability of social screen time exceeding different levels. A smoother decline could indicate more regular high usage periods for social media as compared to overall screen time.

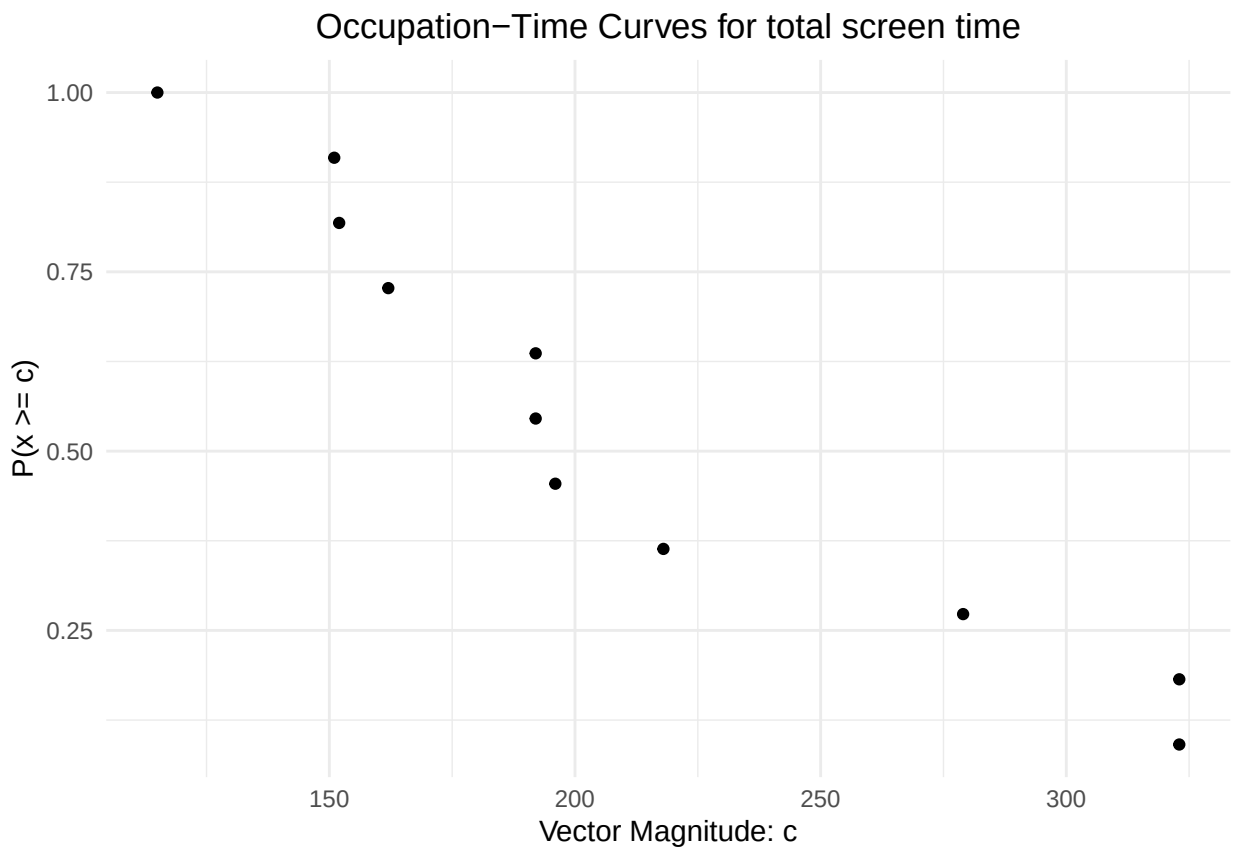
Number of Pickups: The curve show how often users exceed a certain number of pickups in a day. A flatter curve could suggest that the behavior of picking up the phone is more consistent across different days, regardless of the threshold.

Proportion of Social Screen Time: This curve illustrate the probability of the proportion of social screen time being above various points. A slowly drop-off could imply that social media often dominate screen use.

Duration Per Use: This would show the likelihood of the duration per use of the device exceeding certain lengths. The curve decreases fast, this could suggest that longer usage sessions are uncommon.

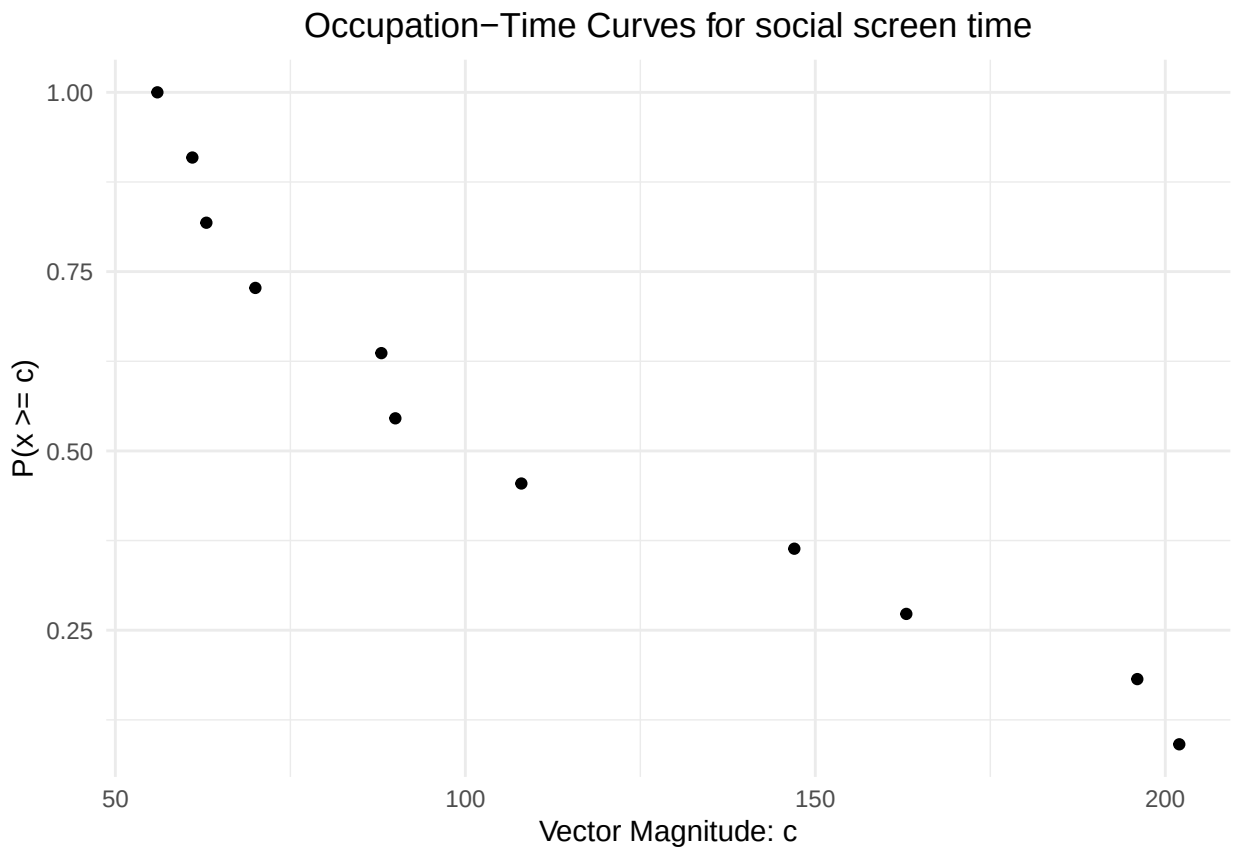
```r
# Calculate the P(x >= c) for Total.ST.min
df2 <- df2[order(df2$Total.ST.min, decreasing = TRUE),]  # Sorting by Total.ST.min
df2$Total.ST.prob <- seq_along(df2$Total.ST.min) / nrow(df2)

# Plot
ggplot(df2, aes(x = Total.ST.min, y = Total.ST.prob)) +
  geom_point() +
  labs(x = "Vector Magnitude: c",
       y = "P(x >= c)",
       title = "Occupation-Time Curves for total screen time") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



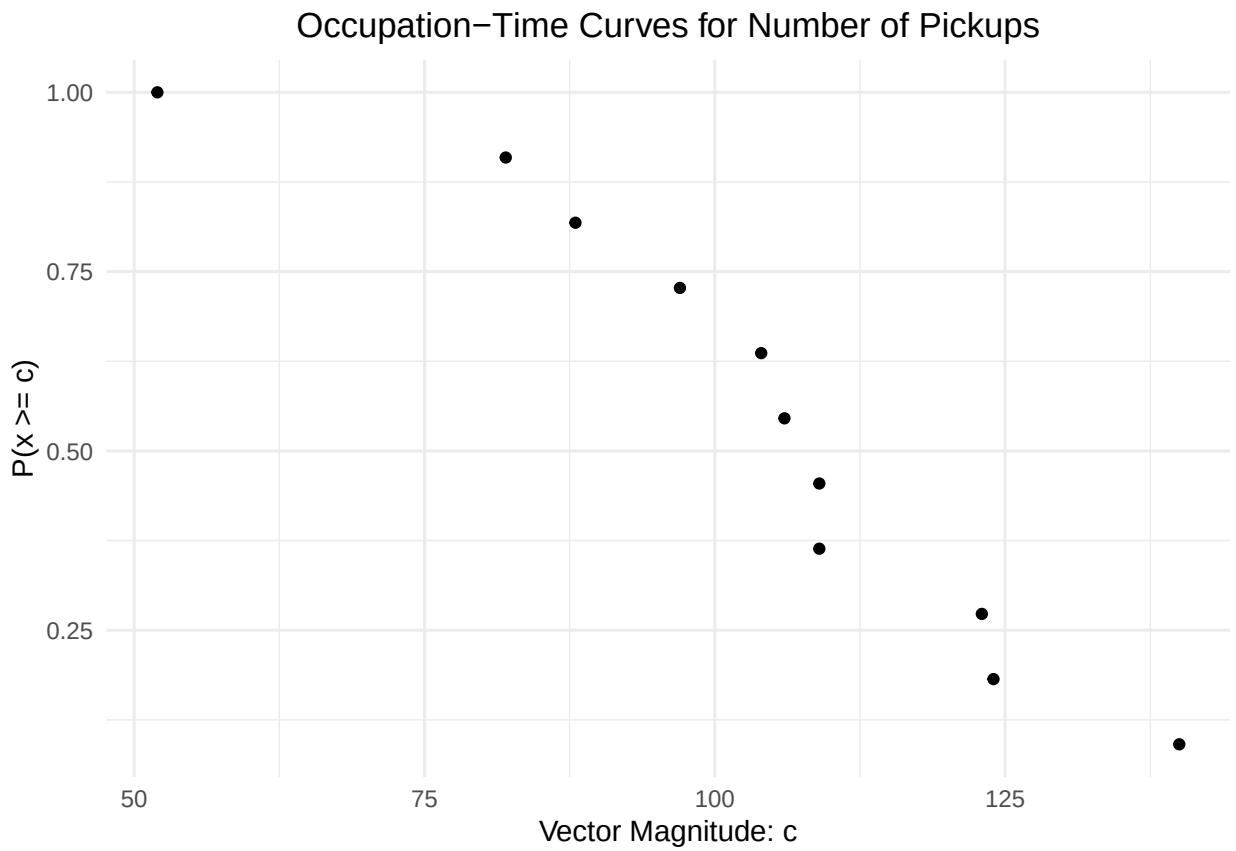

```r
# Calculate the P(x >= c) for Social.ST.min
df2 <- df2[order(df2$Social.ST.min, decreasing = TRUE),]  # Sorting by Social.ST.min
df2$Social.ST.prob <- seq_along(df2$Social.ST.min) / nrow(df)

# Plot
ggplot(df2, aes(x = Social.ST.min, y = Social.ST.prob)) +
  geom_point() +
  labs(x = "Vector Magnitude: c",
       y = "P(x >= c)",
       title = "Occupation-Time Curves for social screen time") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



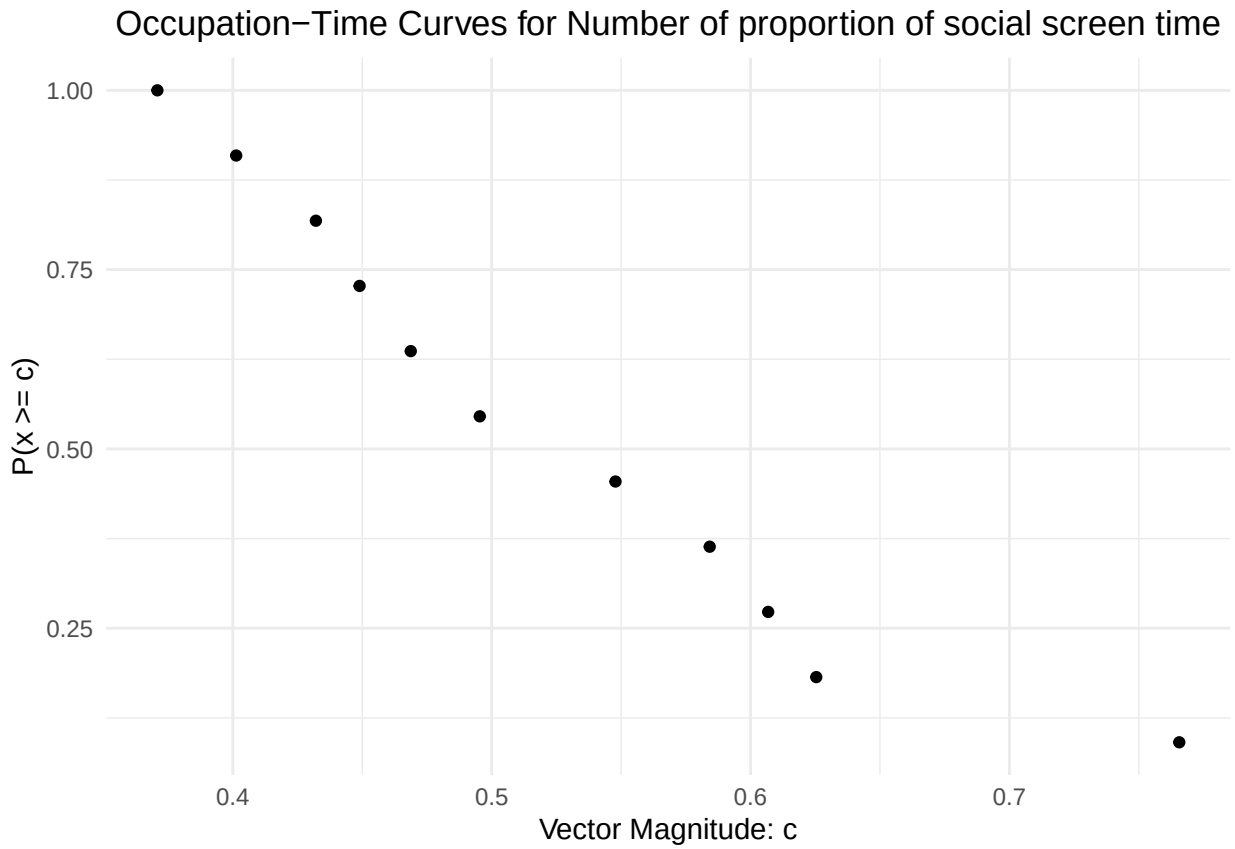

```
# Calculate the P(x >= c) for Pickups
df2 <- df2[order(df2$Pickups, decreasing = TRUE),]  # Sorting by Pickups
df2$Pickups.prob <- seq_along(df2$Pickups) / nrow(df)

# Plot
ggplot(df2, aes(x = Pickups, y = Pickups.prob)) +
  geom_point() +
  labs(x = "Vector Magnitude: c",
       y = "P(x >= c)",
       title = "Occupation-Time Curves for Number of Pickups") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



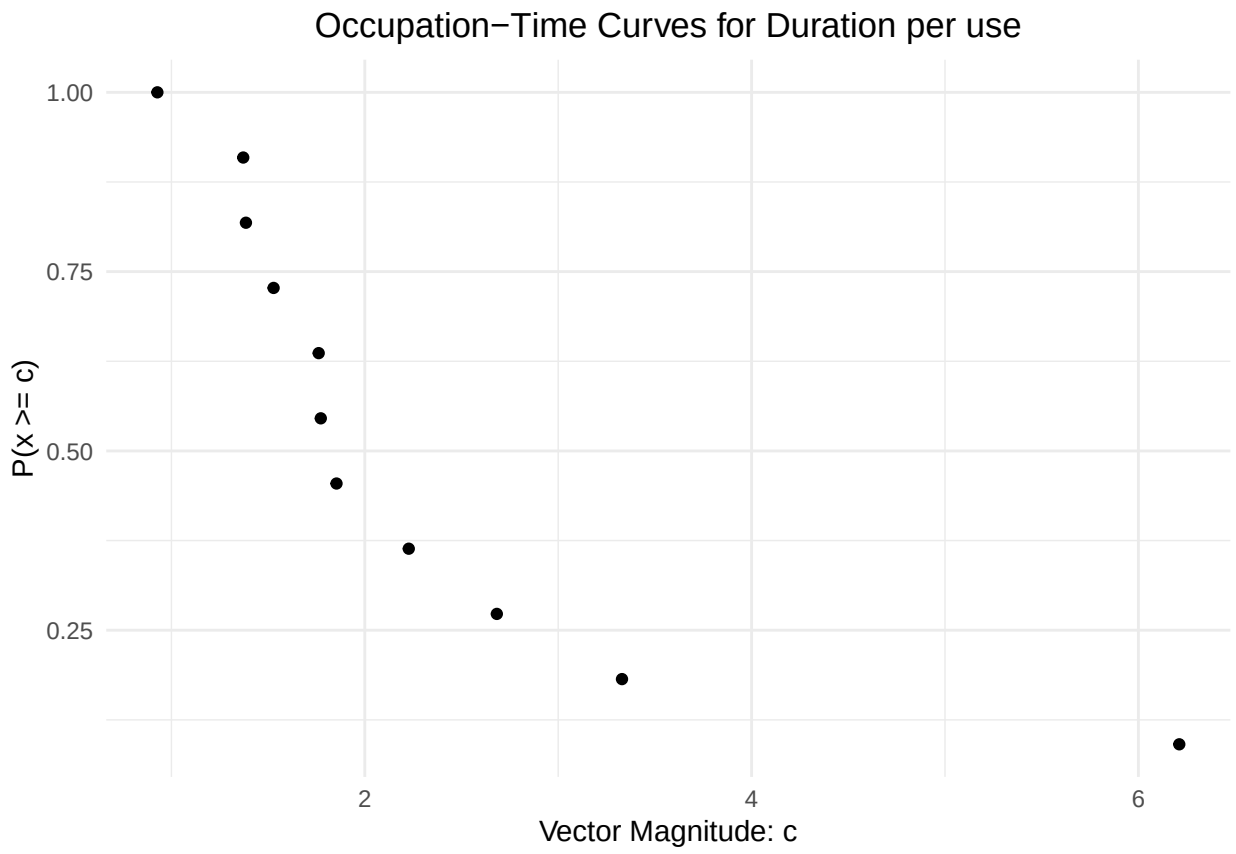

```r
# Calculate the P(x >= c) for prop_ST
df2 <- df2[order(df2$prop_ST, decreasing = TRUE),]  # Sorting by prop_ST
df2$prop_ST.prob <- seq_along(df2$prop_ST) / nrow(df)

# Plot
ggplot(df2, aes(x = prop_ST, y = prop_ST.prob)) +
  geom_point() +
  labs(x = "Vector Magnitude: c",
       y = "P(x >= c)",
       title = "Occupation-Time Curves for Number of proportion of social screen time") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```r
# Calculate the P(x >= c) for prop_ST
df2 <- df2[order(df2$duration_per_use, decreasing = TRUE),]  # Sorting by duration_per_use
df2$duration_per_use.prob <- seq_along(df2$duration_per_use) / nrow(df)

# Plot
ggplot(df2, aes(x = duration_per_use, y = duration_per_use.prob)) +
  geom_point() +
  labs(x = "Vector Magnitude: c",
       y = "P(x >= c)",
       title = "Occupation-Time Curves for Duration per use") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




**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.**

An autocorrelation bar for lag 1 (or any other lag) is only very close to the blue line, but does not exceed it, suggesting that the autocorrelation at this lag is not statistically significant at the 95% confidence level. This means that we do not have enough evidence to say that there is a true autocorrelation at this lag. No bar plots exceed the blue line, indicating that there is no significant autocorrelation in any of the lagging data examined. This means that the past values of the sequence do not provide reliable information for predicting future values.

```
par(mfrow=c(3,2))
for(i in 2:6) {
  acf_result <- acf(df2[[i]], main = paste("ACF for series", names(df2)[i]), plot = T)
  print(paste("Autocorrelations for series:", names(df2)[i]))
  print(acf_result$acf)
  acf_result
}
```

```
## [1] "Autocorrelations for series: Total.ST.min"
## , , 1
##
##               [,1]
##  [1,]  1.000000000
##  [2,]  0.531744084
##  [3,]  0.206876086
```

```
##  [4,] -0.004454385
##  [5,] -0.025860306
##  [6,] -0.067419336
##  [7,] -0.116948657
##  [8,] -0.242272153
##  [9,] -0.307362782
## [10,] -0.257131820
## [11,] -0.217170732
```

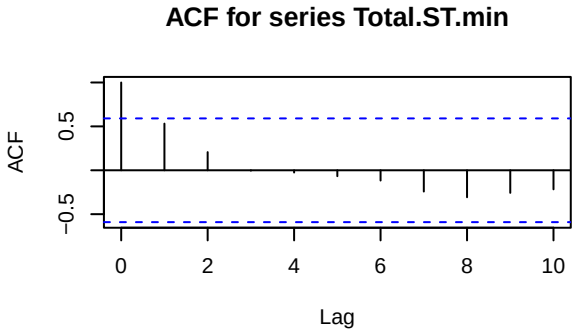
```
## [1] "Autocorrelations for series: Social.ST.min"
## , , 1
##
##              [,1]
##  [1,]  1.00000000
##  [2,]  0.39294360
##  [3,]  0.13437361
##  [4,] -0.07725573
##  [5,] -0.03896096
##  [6,] -0.15349119
##  [7,] -0.22428038
##  [8,] -0.18750952
##  [9,] -0.15909891
## [10,] -0.03800610
## [11,] -0.14871442
```

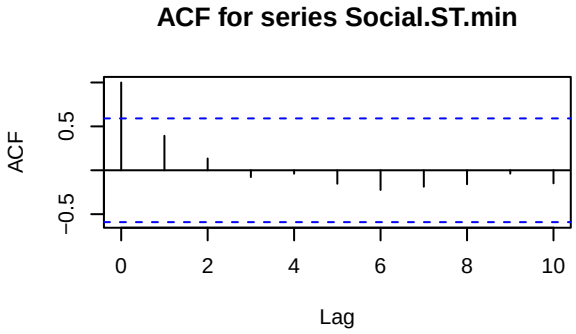
```
## [1] "Autocorrelations for series: Pickups"
## , , 1
##
##              [,1]
##  [1,]  1.00000000
##  [2,]  0.23810459
##  [3,] -0.01340800
##  [4,]  0.20797685
##  [5,]  0.29518127
##  [6,] -0.26843323
##  [7,] -0.23454026
##  [8,] -0.08339784
##  [9,] -0.09074365
## [10,] -0.35980466
## [11,] -0.19093509
```

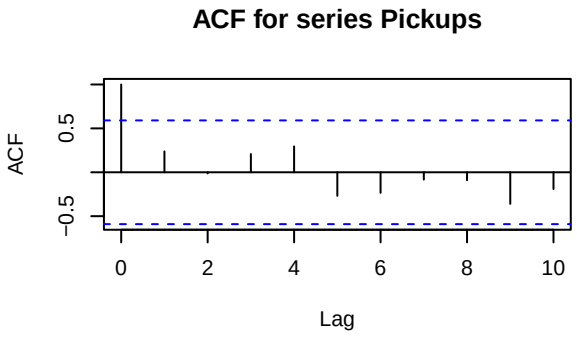
```
## [1] "Autocorrelations for series: prop_ST"
## , , 1
##
##               [,1]
##  [1,]  1.000000000
##  [2,]  0.079654144
##  [3,] -0.103214996
##  [4,] -0.116085192
##  [5,] -0.007330646
##  [6,] -0.234323468
##  [7,] -0.319208700
##  [8,] -0.065642097
##  [9,]  0.047500864
## [10,]  0.199492098
```

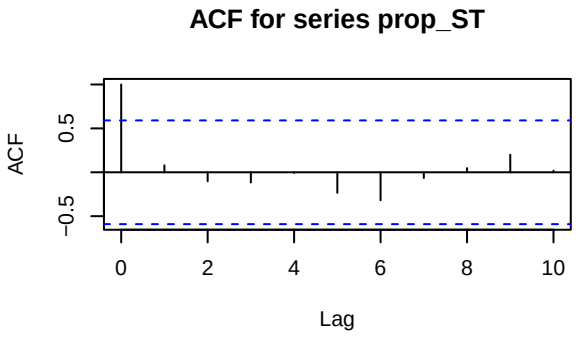
```
## [11,]  0.019157993
```

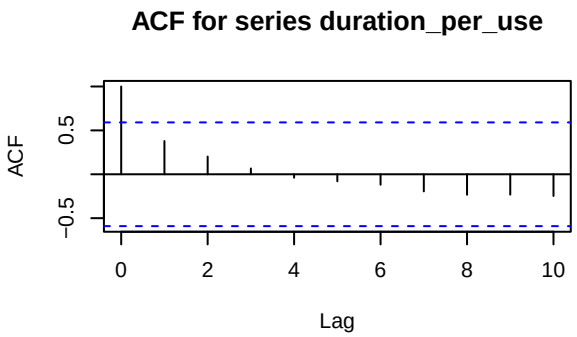
```
## [1] "Autocorrelations for series: duration_per_use"
## , , 1
##
##              [,1]
##  [1,]  1.00000000
##  [2,]  0.37890832
##  [3,]  0.20209022
##  [4,]  0.06640660
##  [5,] -0.03906728
##  [6,] -0.08032067
##  [7,] -0.11943644
##  [8,] -0.19509996
##  [9,] -0.23354449
## [10,] -0.23237144
## [11,] -0.24756486
```



## 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.**

```
library(circular)
```

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

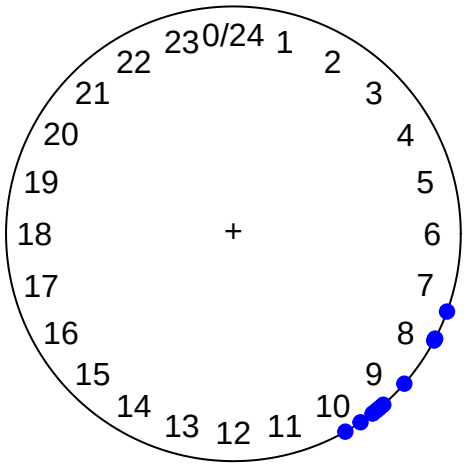
```
## The following objects are masked from 'package:stats':
##
##     sd, var
```

```
time_to_degrees <- function(time) {
  # Convert time to hours and minutes
  hours <- as.numeric(substr(time, 1, 2))
  minutes <- as.numeric(substr(time, 4, 5))
  # Calculate the total hours since midnight
  total_hours <- hours + minutes / 60
  # Calculate the angle
  angle <- total_hours / 24 * 360
  return(angle)
}
angles <- sapply(df$Pickup.1st_EST, time_to_degrees)
pickup_angles <- circular(angles, units="degrees", template="clock24")
```

**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.**

According to this pattern, the first phone pickup usually takes place in the late morning, particularly between 8 and 10 am. People clearly have a tendency to use their gadgets during these times, which could mean that they get up or begin their daily routines. It might represent how much time a person spends getting ready for school.

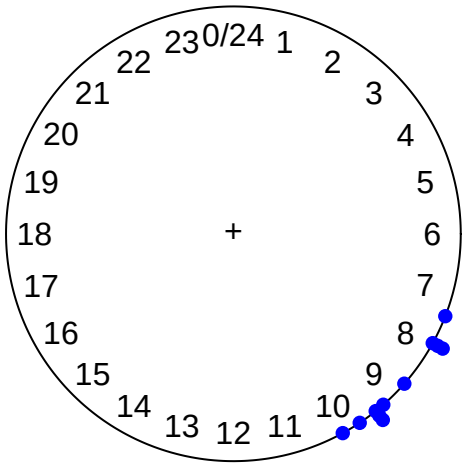
```
# Plot the circular data
plot(pickup_angles, pch=19, col='blue', main="First Pickup Times on 24-hour Clock")
```




**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 bin size as 144 because 360/2.5=144, which I want to create a histogram with a bin size of 2.5 degrees (10 minutes). A full circle is 360 degrees, which is equivalent to a full 24-hour cycle in terms of time. Since there are 1440 minutes in a day (24 hours * 60 minutes per hour), each degree on this circular representation of a 24-hour clock corresponds to 4 minutes of time (1440 minutes / 360 degrees) and 4 minutes/degree $\times$ 2.5 degrees=10 minutes. Thus, I use 144 bin size to represent 10 min.

```
plot(pickup_angles, stack=TRUE, bins=144, col=" blue " ) # 360/2.5=144
```




# PART II: DATA ANALYSIS

## P4

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

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

```
glm_model <- glm(Pickups ~ offset(log(Total.ST.min / 60)), family = poisson, data = df2)
summary(glm_model)
```

```
##
## Call:
## glm(formula = Pickups ~ offset(log(Total.ST.min/60)), family = poisson,
##     data = df2)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -9.891  -1.928   1.451   3.275   7.723
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.3859     0.0297     114   <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: 241.13  on 10  degrees of freedom
## Residual deviance: 241.13  on 10  degrees of freedom
## AIC: 314.04
##
## Number of Fisher Scoring iterations: 4
```

**c**

```
mark_weekdays <- function(date) {
  if (weekdays(date) %in% c("Saturday", "Sunday")) {
    return(0)
  } else {
    return(1)
  }
}
df2$IsWeekday <- sapply(df2$Date, mark_weekdays)
df2$IsAfterJan10 <- 1
glm_model2 <- glm(Pickups ~ IsWeekday + IsAfterJan10 + offset(log(Total.ST.min / 60)), family=poisson, d
summary(glm_model2)
```

```
##
## Call:
## glm(formula = Pickups ~ IsWeekday + IsAfterJan10 + offset(log(Total.ST.min/60)),
##     family = poisson, data = df2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1603  -1.8256  -0.4842   2.1506   6.0332
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.62749    0.08192   32.07   <2e-16 ***
## IsWeekday     0.94673    0.08790   10.77   <2e-16 ***
## IsAfterJan10       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: 241.128  on 10  degrees of freedom
## Residual deviance:  96.101  on  9  degrees of freedom
## AIC: 171.01
##
## Number of Fisher Scoring iterations: 4
```

**c1. Is there data evidence for significantly different behavior of daily pickups between weekdays and weekends? Justify your answer using the significance level $\alpha = 0{:}05$.**

There is evidence for significantly different behavior of daily pickups between weekdays and weekends. The highly significant coefficient for $IsWeekday$ confirms this (p-value $<$ 2e-16).

**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 $\alpha = 0{:}05$.**

There is no evidence for a significant change on the behavior of daily pickups after the winter semester began because all data is collected after Jan 10, which means dummy variable $Zt$ is 1 and it cannot be included in the model.

## P5

**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.**

```
estimates <- mle.vonmises(pickup_angles)
(mu <- estimates$mu)
```

```
## Circular Data:
## Type = angles
## Units = degrees
## Template = clock24
## Modulo = asis
## Zero = 1.570796
## Rotation = clock
## [1] 132.1067
```

```
(kappa <- estimates$kappa)
```

```
## [1] 19.05673
```

**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.**

```
angle_830AM <- (time_to_degrees("08:30") * 2 * pi)/360- pi
angle_830AM_circular <- circular(angle_830AM, type = 'angles', units = 'radians')
cdf <- pvonmises(angle_830AM_circular, mu, kappa)

# The probability that the first pickup is at 8:30 AM or later
probability <- 1 - cdf
probability
```

```
## [1] 0.7839057
```