# Project I: BIOTAT 620W2022

Team Hot Pot: Xueting Tao, Isabel Gomez

2/14/2022

## Abstract: Write a high-level summary of project I.

## Introduction:

This section should include objective, hypothesis, and motivation of your project, followed by a brief discussion why your data are relevant to your proposed study. In addition, it presents an outline of your data analysis as well as some major findings.

### Data Description

This section covers background of data collection, list of variables (see a separate file), and characterization of a team. It is important to create the so-called Table 1 that lists summary descriptive statistics for all collected variables to describe the team. You may also consider using figures (e.g. boxplot, histogram, scatterplot, time series plot, ACF plot and circular plot of 24 hours), if necessary, to display distributions of some variables for which you like to show more detailed information.

Two sets of self-reported data were collected for this study. The first, self-reported screen usage data between January 3rd, 2022 – February 13, 2022 (20 days) from two students in the Biostatistics 620 class is used in this study. There are a total of eight variables reported: ID or identification of participant, total daily screen time reported in minutes,daily social media time reported in minutes, number of pick-ups in a day, and first pick up time reported as HH:MM. A second set of baseline data used to characterize participants and teams, is also collected. There are a total of 13 variables collected: number of team members participant has ever wok wih in any previous group project reported as a continuous variable with maximum value 2, number of team members participant regular talks about academic matters with reported as a continuous variable with maximum value 2, number of team members participant ever talks about topics other than academic matters with reported as a continuous variable with maximum value 2, if the participant lives with pets at home that they look after reported as a binary variables with 0 = no, the sex of the participant reported as a binary variable with 0 = female, the age of the participant reported as a continuous variable, the course credit hours in the winter semester reported as a continuous variable, the country where participated previously received their degree from reported as a binary with 0 = Non- U.S, current employment status with 0 = not employed over 10 hours/week, number of siblings reported as a continuous variable, number of social apps installed on regularly used mobile device reported as a continuous variables, number of personal mobile devices owned reported as a continuous variable, and finally the procrastination score reported as a score from 0 to 100 from the psychology today procrastination test abridge.

Table 1 shows the drastic difference in screen usage and number of pickups between the two participants in this study. As a team, the average amount of total screen time between January 3rd - February 13, 2022 is 421 minutes or approximately 7 hours. The average social screen time is 105 minutes or 1 hour and 45 minutes and the average number of daily pickups was 96.

Figure 1 displays the pairwise scatter plots of total screen time, social screen time and number of pickups for the group. There is moderately weak significant negative correlation between total screen time and number of pickups (Pearson correlation = -0.292). Social screen time has a weak positive correlation with both total

Table 1: Screen Data Summary Descriptive Statistic per participant

| ID | min | max | mean | median | sd |
|---|---|---|---|---|---|
| **Total Screen Time (in minutes)** | | | | | |
| isgomez | 212 | 644 | 381 | 342 | 127 |
| xuetao | 127 | 788 | 462 | 484 | 179 |
| hotpot | 127 | 788 | 421 | 388 | 159 |
| **Total Social Time (in minutes)** | | | | | |
| isgomez | 44 | 317 | 129 | 113 | 65 |
| xuetao | 12 | 280 | 82 | 68 | 63 |
| hotpot | 12 | 317 | 105 | 87 | 68 |
| **Number of daily pickups** | | | | | |
| isgomez | 60 | 331 | 161 | 157 | 58 |
| xuetao | 14 | 60 | 30 | 28 | 12 |
| hotpot | 14 | 331 | 96 | 60 | 78 |

screen time (pearson correlation = 0.035) and number of pickups (pearson correlation = 0.181), both of these are non-significant.

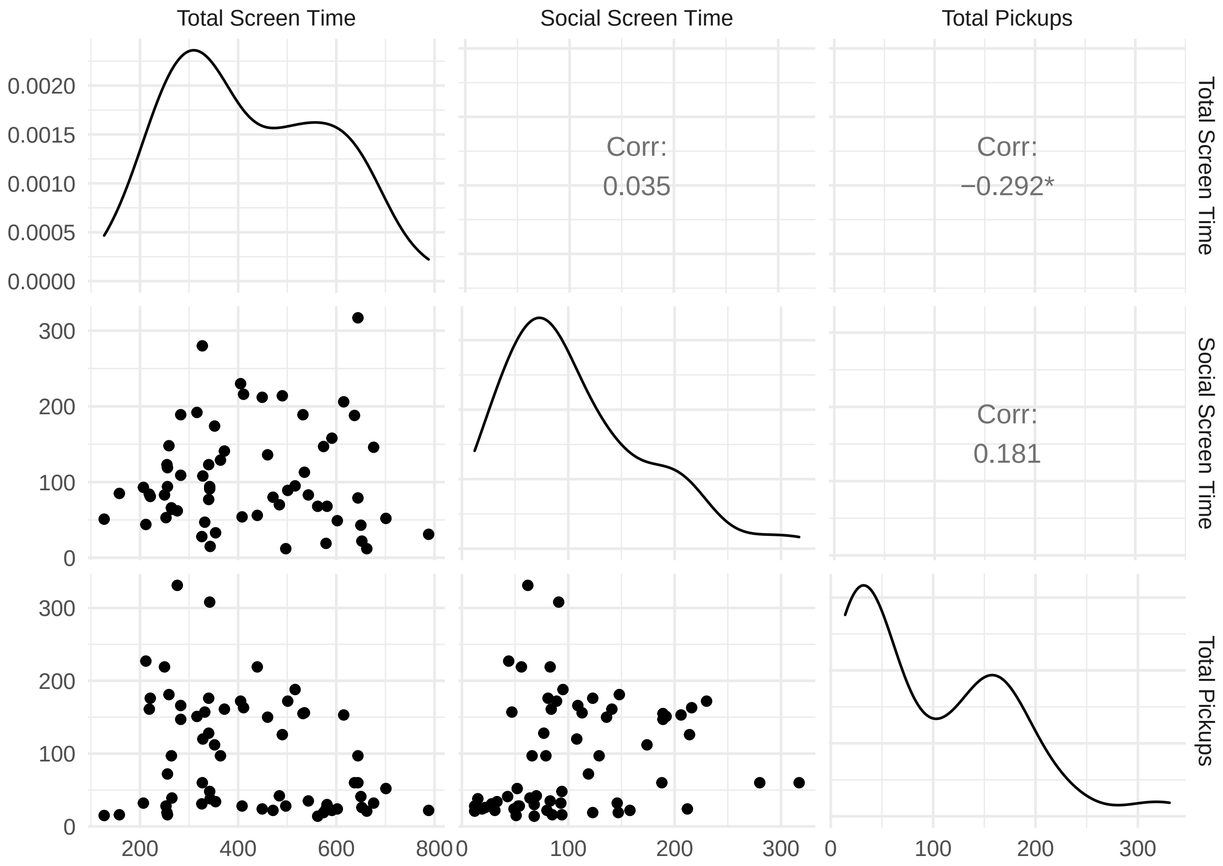
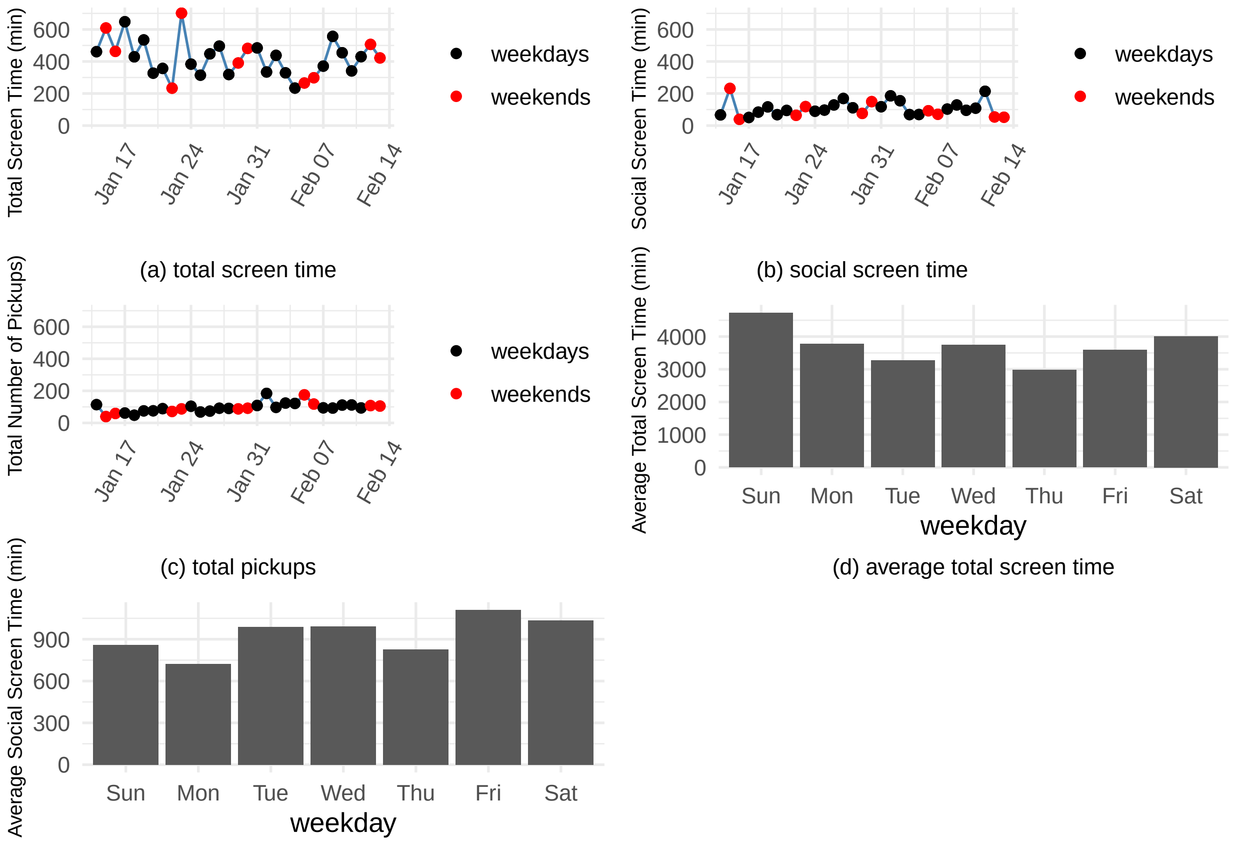


Figure 1: Pairwise scatterplot of total screen time, social screen time and total pickup for team hotpot.



(a) total screen time

(b) social screen time

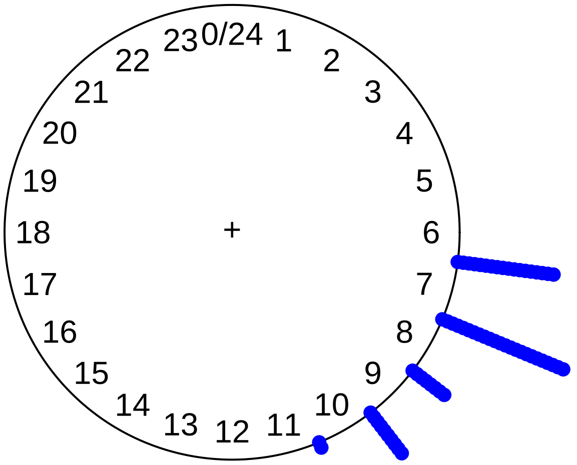
(c) total pickups

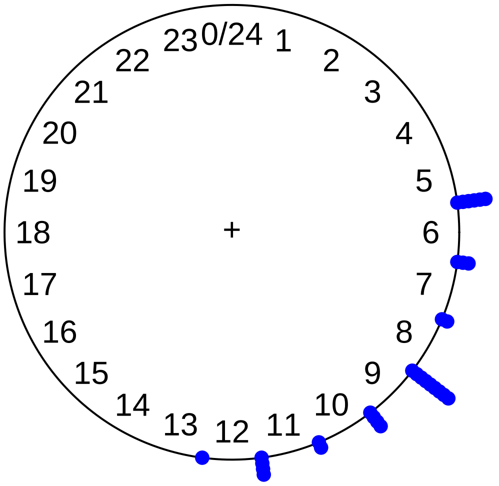
(d) average total screen time

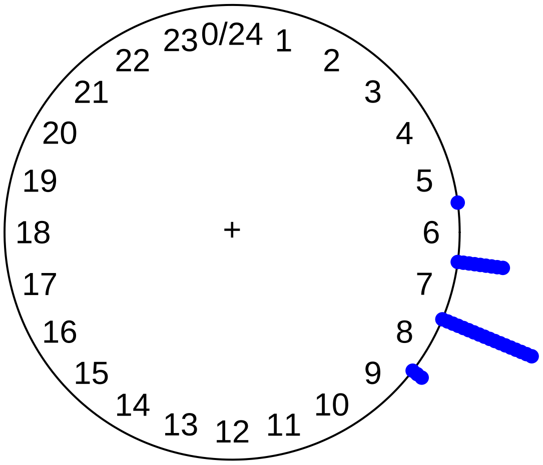
(e) average social screen time

Figure 2: Time Series Plots (a) - (e) team average total screen time, social screen time, team average total number of pickups, average total screen time by weekday, average social screen time by weekday

Figure 2 shows the first pick-up time for each member, along with the average of all the participants together. From this figure we can see that the most popular pickup time is around 7:30am. Whereas if we look at the individual pickup time, isgomez has a more variable pickup time with the earliest being at 8:30am and xuetao has a more consolidated pickup time around 7:30am.


0/24
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
+


0/24
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
+



## Data Preprocessing:

This section discusses issues related to data merging from individual data sheets into a group data sheet, data cleaning and validation, data harmonization in data merging, variable transformation and so on.

#Federated Learning: Based on your study objective and hypothesis given in the Introduction section as well as preliminary data analysis in the Data Description section, in this case you choose an outcome variable y (e.g. daily social screen time), and a predictor x (e.g. number of pickups), to run a linear regression y ~ x via the federated learning method. Here, each device user represents a data source where raw data cannot be shared, but summary statistics are allowed to be shared. (a) First, establish a federated learning procedure for the calculation of regression parameters (intercept, slope and variance) and their standard errors. Second, (b) describe a distributed computing platform to implement your federated learning machinery developed in part (a). (c) Third, report your analysis results.

## Meta Learning:

You may now extend the above federated learning based on the simple linear regression with one covariate to the case of multiple regression with many covariates (e.g. the number of pickups and duration per use etc.), that is, y ~ x1 + ... + xp. The meta learning allows you to establish a distributed computing platform to carry out the federated learning without sharing individual data. Demonstrate your meta learning platform by one example with multiple predictors.

## Confirmation analysis:

Note that your team has a combined data sheet. Thus, you may repeat the above analyses of the simple regression and multiple regression using the combined raw data to obtain the oracle results. Compare and confirm numerically if the results from the federated learning method are the same as the oracle results. Conclusion & Discussion: Summarize your main contributions and findings in this project. What was your experience of data analysis, especially in the aspect of team collaboration? What have you found to be most interesting and surprising? What are the limitations of your study (e.g. the inclusion of confounding factors in the analysis)? What is the future work?

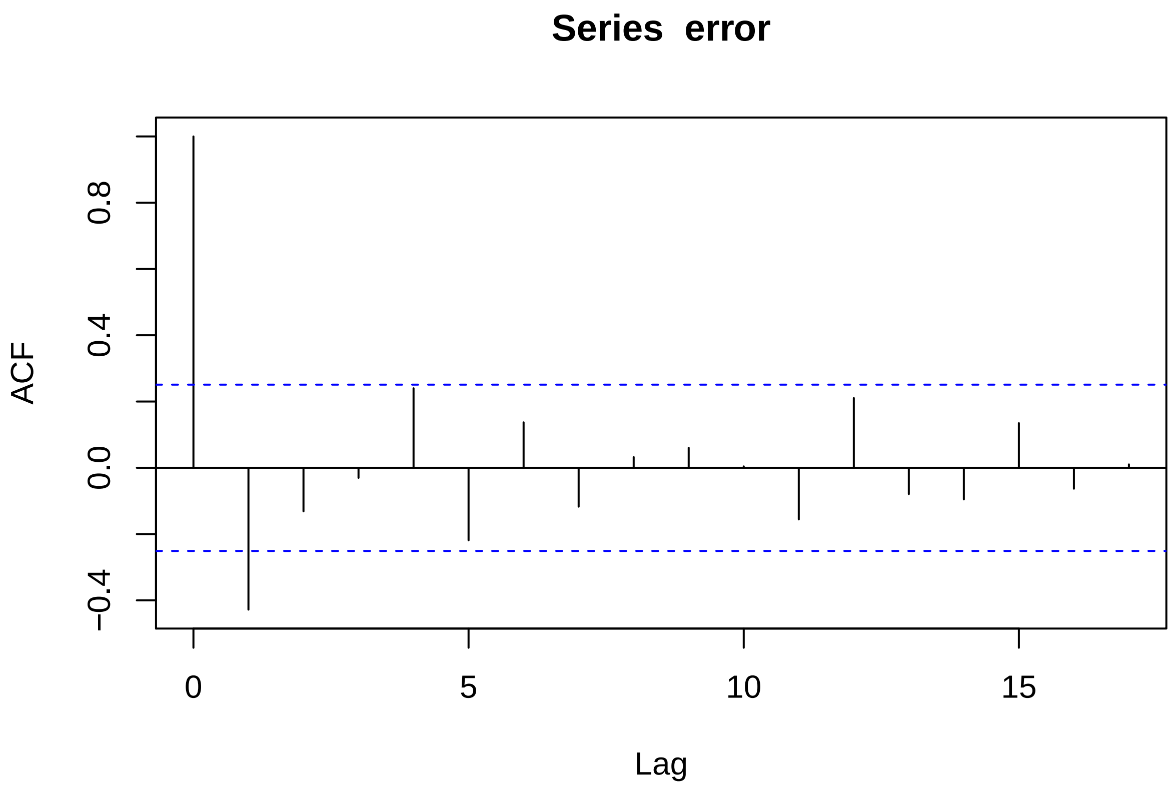
## Acknowledgement:

You may write some additional notes related to help given by people outside of the authorship and roles that individual authors played in the project.

## Appendix:

You can always create an appendix to include more detailed supplementary information of your project if necessary. Format Required by the Project Each project should be prepared with one-inch margins, in 12-point size letters and no more than 25 lines per page, double-spaced throughout. The first page should include a title, authorship, key phrases, and a one-paragraph abstract. The abstract should not exceed 200 words. Each project should not exceed 8 pages, including title, authors, abstract, key phrases, figures and tables as well as references. You are allowed to submit an appendix that contains some technical details, R code, and additional analysis results. The appendix should not exceed 5 pages.

```r
ff <- function(x) dvonmises(x, mu=circular(2*pi-2.04), kappa=13.44)
curve.circular(ff, join=TRUE, xlim=c(-2.3, 1),
  main="Density of a VonMises Distribution \n mu=2.337, kappa=3.941")
#plot(density.circular(density_V, bw = 3))
```


Series error
ACF
−0.4
0.0
0.4
0.8
0
5
10
15
Lag