Inference of Daily Mobile Device Usage of Graduate-Level Students

Authors: Vikram Bala, Neyan Deng, Ethan Werner (Team VENTURE)

GitHub Link: https://github.com/CAROLINeeeeDeng/VENTURE

Abstract: The harmful effects of unrestrained screen time are well documented, with increased screen time being associated with greater instances of depression, sleep disturbances, and other health issues (Nakshine et al., 2022). The goal of this study is to identify potential causes of increased screen time by examining the behaviors of university students. By identifying potential causal factors, new methods of limiting screen time can target specific behaviors to aid in preventing future health complications. To achieve this goal, federated learning and distributed computing techniques were applied to each project group member's screen time data to create a functioning global model using each individual's summary statistics. Summary statistics were calculated for the predictors (daily number of pickups and classes, daily proportion of social screen time (DPSST)) on the outcome (daily total screen time). The federated learning model not only showed that it is a reliable way of creating a model without using raw data but also showed trends suggesting that higher daily pickups and days with fewer classes were correlated with increased screen time. Focusing intervention techniques on discouraging pickups and increasing efforts on days when students have fewer classes could be fruitful in decreasing total screen time.

Key Phrases: Federated Learning, Summary Statistics, Regression

Introduction:

The objective of this project is to gain insight into students' mobile device usage (daily total in minutes) by using federated learning techniques and distributed computing. This project is critical, as excessive screen time has been linked to negative well-being and development, obesity, depression, sleep disturbances, cardiovascular disorders, and more (Nakshine et al., 2022; Zhu et al., 2023). The motivation is to better understand causal variables of screen time and identify potential intervention strategies to ultimately reduce mobile device usage and improve user health outcomes. By using federated learning techniques and distributed computing, we can achieve this goal by using only summary statistics, illustrating the potential of these methods when data access is limited (Wen et al., 2022).

We decided to analyze the effect of daily number of pickups, daily number of classes, and daily proportion of social screen time (DPSST) on daily total screen time. Daily number of pickups has been shown to affect total screen time usage (Shaw et al., 2020). Also, social media usage has been linked to total screen time usage (Odgers et al., 2020). We hypothesize that an increased daily number of pickups and DPSST leads to increased daily screen time because of users' tendency to interact more with their mobile devices following these actions. We also hypothesize that an increased number of classes leads to decreased screen time, as users have a larger percentage of their day occupied, leading to decreased device usage.

Our findings suggest that an increase in the number of classes correlates with a reduction in daily total screen time. Conversely, a higher number of pickups is associated with an increase in screen time. DPSST does not appear to have a significant impact on screen time.

The significance of this study is gaining insights into potential causal variables for daily total screen time. With this information, we can now develop intervention strategies to focus on limiting pickups and filling time on days with fewer classes, as this would lead to reduced screen time. This, would then lead to improved well-being of users and reduced health concerns, such as obesity, depression, and cardiovascular issues. The innovation of this study is that we implemented federated learning and distributed computing methods that showed corroboration with oracle results, illustrating the viability of these methods when only limited data (summary statistics) is available. This has applications in numerous health fields, as data is often limited or incomplete due to data privacy concerns and confirmation of these innovative methods with oracle results shows its potential to address these challenges without losing relevant information and findings. Appendix 1 illustrates the procedure (design) our team implemented in this project to achieve the defined objectives and test our hypotheses.

Data Description:

Data was collected by team members from the period of 01/03/2024 to 02/13/2024 (42 days). Raw data for each individual was collected separately then merged to create a large dataset, where an "ID" column was used to distinguish between individuals. The data collected included daily total screen time, daily social screen time, daily number of pickups, first pickup time, and number of classes per day. From this raw data, more fields were calculated, including the daily proportion of social screen time, defined as the ratio of daily social screen time over daily total screen time, and daily duration per use, defined as the ratio of daily total screen time over daily number of pickups. Covariate data was collected for all group members, including number of teammates worked with before, number of teammates talked to about academic matters, number of teammates talked to about non-academic matters, number of pets, sex, age, credit hours, country of previous degree, job status, number of siblings, number of social apps installed, number of mobile devices, and self-reported procrastination scores based on an online assessment. We decided to study the effect of daily number of pickups, number of classes per day, and daily proportion of social screen time (predictors) on the daily total screen time in minutes (response). For the selected variables, the summary statistics for both the aggregated and individual data were calculated and are shown in Appendix 2.

Figures were generated to display detailed patterns for the selected variables. For individual data (grouped by "ID"), boxplots are shown in Appendices 3-6, pairwise scatterplots are shown in Appendix 7, time series plots are shown in Appendices 8-11, and Autocorrelation Function (ACF) plots are shown in Appendices 12-14. The boxplots show high screen time and pickups by neyan and eawerner, and high DPSST by vikbala. The pairwise scatterplots show strong correlation between pickups and total screen time for all individuals, while DPSST is more random. The time series plots illustrate high variability and unpredictability across days (for all individuals) for total screen time, daily pickups, and DPSST, whereas number of classes follows a more repeated pattern. The ACF graphs for daily total screen time of each one of us show that there is no significant autocorrelation at various lags, implying that screen time is random and unpredictable from day to day. The lack of a clear pattern suggests that external factors may influence daily screen usage, indicating that past screen time is not a reliable predictor of future use.

Data Preprocessing:

Data was collected from all project members in a Google Sheet consisting of individual data and combined raw data. For each member and the whole project group, summary statistics were calculated for the response (total screen time) and the three covariates (pickups, number of classes, and DPSST). Each member's data was observed to see when the earliest common starting date for data collection, which was determined to be January 3, 2024. The freeze date for data collection was February 13, 2024. Due to an issue with one member's phone keeping track of screen time, for three days, screen time was not properly collected by the device. To make up for this, data from three days past February 13 were added so each member had the same number of days of screen time recorded. Also, one member's device did not record any first pickup time values, and thus, this variable was not selected for further analysis. Ultimately, each member's summary statistics were used to create a federated learning model.

Federated Learning:

We implemented a federated learning approach to collaboratively train a linear regression model while maintaining the privacy of each participant's data. Each member had access to our own data sets of equal size. In isolation, each member computed the cross-product of the matrix of predictors (XTX) and the product of the matrix of predictors with the response variable (XTY) from their respective data. These matrices represent the essential components required to derive the global regression coefficients without revealing any individual data points. Upon calculating these intermediate values, we shared our XTX and XTY results with a designated aggregator within our team. The aggregator combined these results to form the global XTX and XTY matrices. The aggregation process was as follows:

$$X^TX_{global} = X^TX_{neyan} + X^TX_{vikbala} + X^TX_{eawerner}$$

$$X^TY_{global} = X^TY_{neyan} + X^TY_{vikbala} + X^TY_{eawerner}$$

Finally, the global coefficients β_{global} were calculated based on the global matrices through matrix inversion and multiplication, yielding a single set of coefficients that encapsulate the insights from all the datasets collectively.

Based on the global regression coefficients (Table 1) and summary statistics, each member computed their Residual Sum of Squares (RSS) and Total Sum of Squares (TSS) with their data. These individual RSS and TSS calculations were then shared among us to aggregate and compute the global RSS and TSS, enabling us to calculate the global standard error for the regression coefficients and the global R^2, which is approximately equal to 0.9710.

	Estimate	**Standard Error**
Intercept	292.7527	50.9104
Pickups	0.9713	0.2890
Num.Classes	-47.8042	12.1588
DPSST	-26.5248	71.2123

Table 1. Results From Federated Learning Procedure.

Confirmation Analysis:

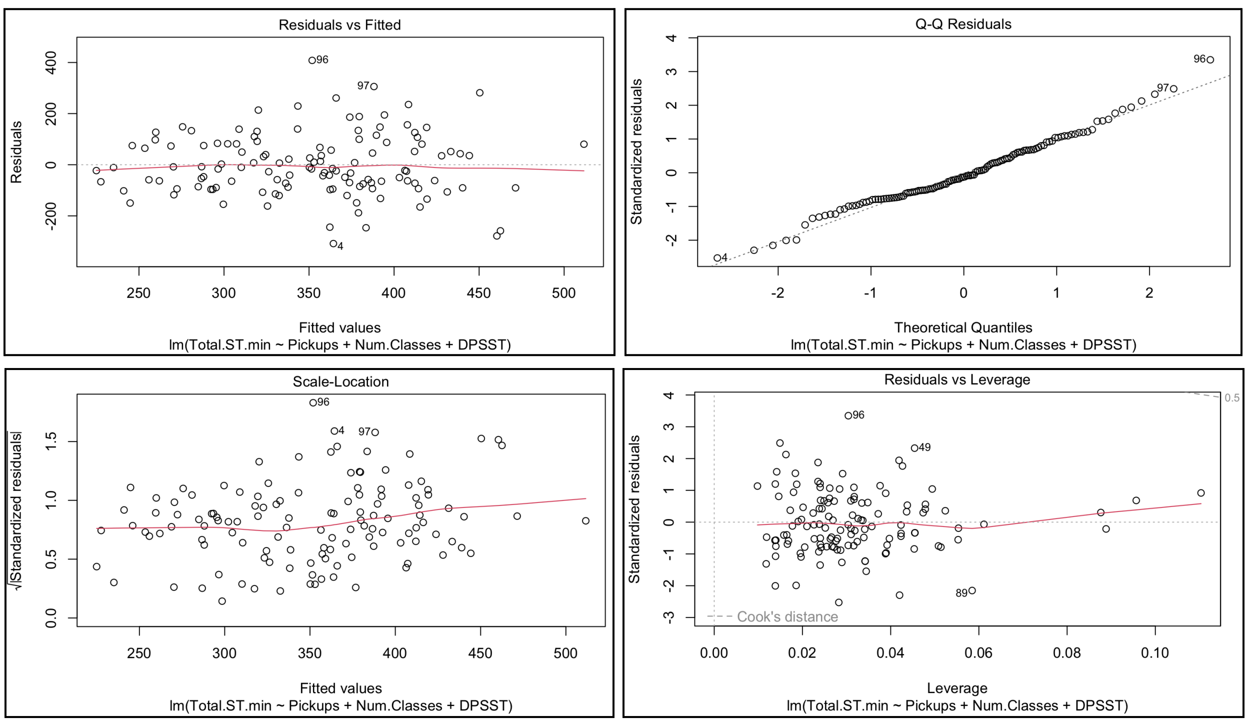
Upon completion of the federated learning phase, we combined our datasets into a single dataset for further analysis. Utilizing the lm() function in R, we directly fitted the merged data into a multiple linear regression model (Figure 1). Intriguingly, the coefficients, standard errors, and R^2 derived from this direct regression were identical to those obtained from our federated learning approach. This concordance validates the efficacy of our federated learning method, demonstrating that it can yield results comparable to traditional centralized approaches while maintaining the confidentiality of individual data contributions (Figures 2-5).

```
Call:
lm(formula = Total.ST.min ~ Pickups + Num.Classes + DPSST, data = combined_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-308.32  -85.61  -15.62   80.55  408.15 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 292.7527    50.9104   5.750 6.73e-08 ***
Pickups       0.9713     0.2890   3.361 0.001037 ** 
Num.Classes -47.8042    12.1588  -3.932 0.000141 ***
DPSST       -26.5248    71.2123  -0.372 0.710186    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 123.8 on 122 degrees of freedom
Multiple R-squared:  0.1871,	Adjusted R-squared:  0.1671 
F-statistic: 9.361 on 3 and 122 DF,  p-value: 1.288e-05
```



Figures 1-5. Federated Learning Model Implementation and Validation.

According to the plots for our fitted model, the Residuals vs Fitted plot does not display any obvious patterns, indicating good fit without systematic errors. The residuals are reasonably well-distributed along the reference line, suggesting normality. Residuals are spread evenly across the range of fitted values, implying consistent variance. Lastly, the "Residuals vs Leverage" plot shows there are a few points with higher leverage, while none appear to be excessively influential by checking the Cook's distance. Overall, these diagnostics suggest all the model assumptions are reasonably met.

Conclusion and Discussion:

By harnessing the power of federated learning, we ensured that each participant's data remained in their control, upholding the principle of data privacy while still enabling us to benefit from the collective power of our combined data. The model's validity can be seen from its comparison to a model of the raw data points from each member's dataset. The predictors and the response variables have the same values regardless of the method (federated learning or raw data) used to produce them. Homogeneity in the regression parameters shows the federated learning approach was successful in producing a valid regression model. While the R^2 value was not very high (0.1671), the p-value suggests (1.3e-5) the model itself was highly significant.

The regression model also produced trends that were hypothesized: greater number of pickups meant more screen time and greater daily number of classes meant less screen time. Surprisingly, higher DPSST resulted in reduced screen time, but the predictor did not indicate a significant difference in the model. Despite the apparent contradictory nature of this predictor, its effect on the model itself was minimal. In computing these analyses, it was interesting to see the federated approach's homogeneity with the global model of raw data and how well it fit the assumptions of a linear regression model. Working in a team, it was interesting to see how each member of the team thought about the data and their own perceptions of their behaviors. Despite the differences in screen time for each day, analyses went smoothly for both the individual local models and the global federated model. There was little, to no, confusion about how variables were measured.

Overall, the federated learning model was very accurate to a raw, global dataset, but did have some limitations. One such limitation was the small number of local models (members of the group), with only 3. Future studies could potentially benefit from having a greater number of local models to coordinate data with the global federated model, and this would also increase statistical power. Additionally, though significant, the global model had a very low R^2 value. This suggests that there are some other predictor variables for total screen time (either measured or unmeasured) that could offer some explanation for the response. Some potential variables to explore in future iterations of the model for total screen time data could be time of first device pickup or daily number of hours asleep. Additionally, including some relevant indicator variables (weekday, sex, etc.) may raise the R^2 value slightly and thus could be explored in future studies.

This project showed that the federated learning procedure could be used to reliably create a model for linear regression. With the reliance on technology that people have, other analyses on screen time data could look at training different federated learning models using the same local data but studying different relationships between them. Considering the previously stated relationship between screen time and daily number of pickups, it would be interesting to examine a relationship of a multiple linear regression model that treats both screen time and pickups as independent variables of a model to see if another variable can be explanatory for both outcomes.

Acknowledgments: Our team would like to thank our professor, Dr. Peter Song, and our GSI, Yijun Li, for their instruction and guidance during this project.

References:

Nakshine, V. S., Thute, P., Khatib, M. N., & Sarkar, B. (2022). Increased screen time as a cause of declining physical, psychological health, and sleep patterns: A literary review. *Cureus*. https://doi.org/10.7759/cureus.30051

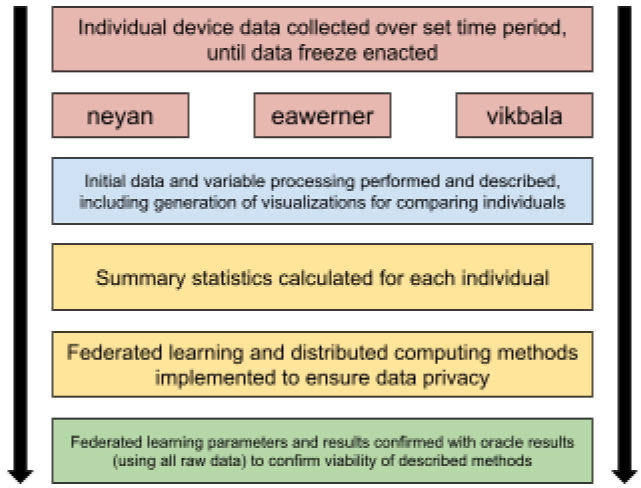
Odgers, C. L., Schueller, S. M., & Ito, M. (2020). Screen Time, social media use, and adolescent development. *Annual Review of Developmental Psychology*, 2(1), 485–502. https://doi.org/10.1146/annurev-devpsych-121318-084815

Shaw, H., Ellis, D. A., Geyer, K., Davidson, B. I., Ziegler, F. V., & Smith, A. (2020). Quantifying smartphone "use": Choice of measurement impacts relationships between "usage" and health. *Technology, Mind, and Behavior*, 1(2). https://doi.org/10.1037/tmb0000022

Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J., & Zhang, W. (2022). A survey on Federated Learning: Challenges and Applications. *International Journal of Machine Learning and Cybernetics*, 14(2), 513–535. https://doi.org/10.1007/s13042-022-01647-y

Zhu, X., Griffiths, H., Xiao, Z., Ribeaud, D., Eisner, M., Yang, Y., & Murray, A. L. (2023). Trajectories of screen time across adolescence and their associations with adulthood mental health and behavioral outcomes. *Journal of Youth and Adolescence*, 52(7), 1433–1447. https://doi.org/10.1007/s10964-023-01782-x

Appendix:

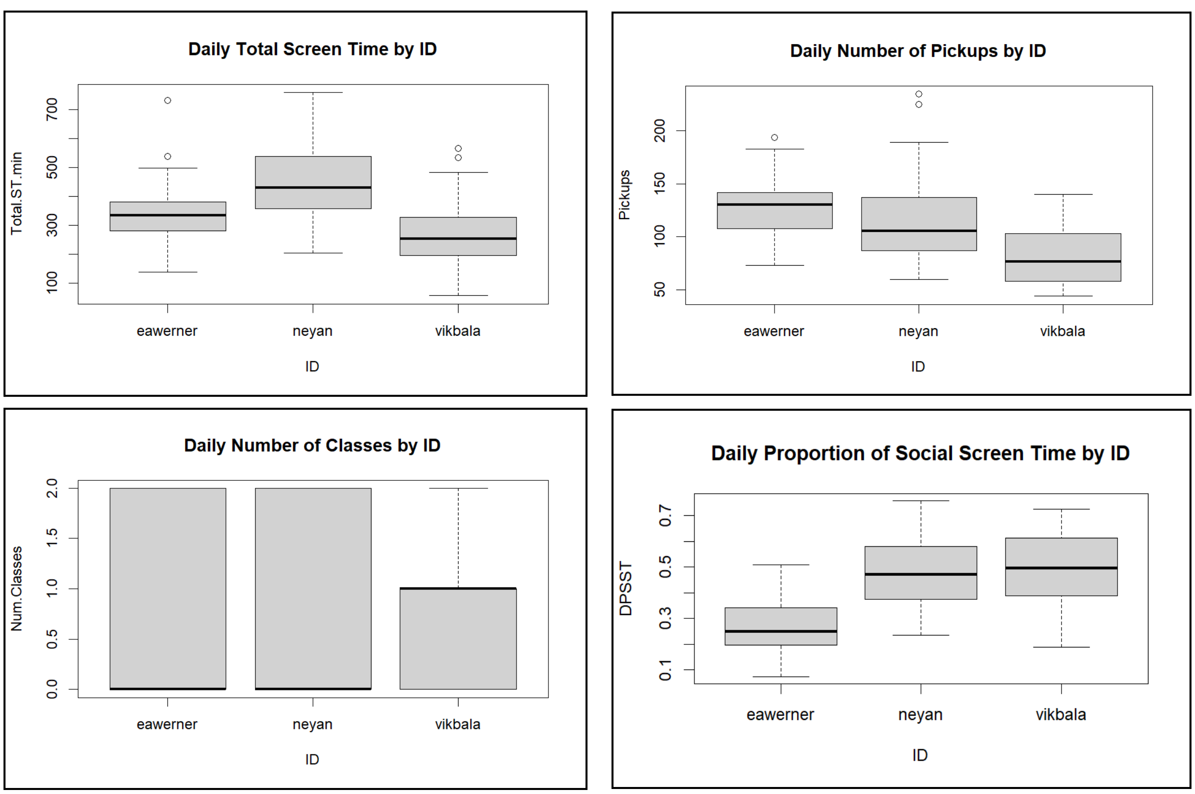


Appendix 1. Workflow of Procedure Implemented for Project Analysis.

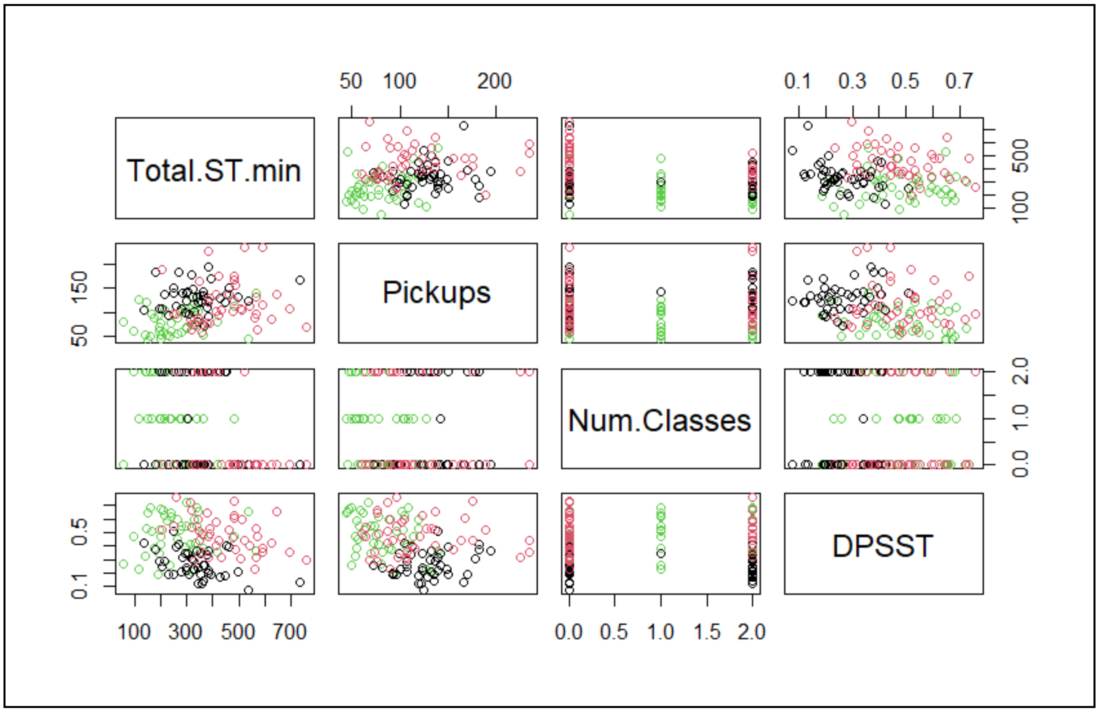
	Pickups	**Num.Classes**	**DPSST**	**Total.ST.min**	**ID**
Mean	109.6	0.7778	0.4111	351.1	all
Q1	78	0	0.2851	260.5	all
Median	106.5	0	0.4	340	all
Q3	133	2	0.5317	424	all
Min	44	0	0.0743	56	all
Max	235	2	0.7595	760	all
Variance	1590.929	0.8302	0.0262	18046.54	all
Sample Size	126	126	126	126	all
Mean	81.119	0.7857	0.4867	264.381	vikbala
Q1	58.25	0	0.3878	196.25	vikbala
Median	76.5	1	0.4944	252.5	vikbala
Q3	102.75	1	0.6084	326.5	vikbala
Min	44	0	0.1885	56	vikbala
Max	140	2	0.7248	565	vikbala
Variance	689.1806	0.6115	0.021	12042.144	vikbala
Sample Size	42	42	42	42	vikbala
Mean	129.5476	0.881	0.2667	338.881	eawerner
Q1	108.75	0	0.1956	281.25	eawerner

Median	130.5	0	0.2479	333.5	eawerner
Q3	141.5	2	0.3423	380.75	eawerner
Min	73	0	0.0743	137	eawerner
Max	194	2	0.508	732	eawerner
Variance	762.2538	0.9855	0.0095	11034.0587	eawerner
Sample Size	42	42	42	42	eawerner
Mean	118.1429	0.6667	0.4798	450.119	neyan
Q1	87	0	0.3769	357.75	neyan
Median	105.5	0	0.4704	428.5	neyan
Q3	137	2	0.5727	544.25	neyan
Min	60	0	0.2339	204	neyan
Max	235	2	0.7595	760	neyan
Variance	2085.6376	0.9106	0.0173	15140.8391	neyan
Sample Size	42	42	42	42	neyan

Appendix 2. Summary Statistics for Selected Variables.

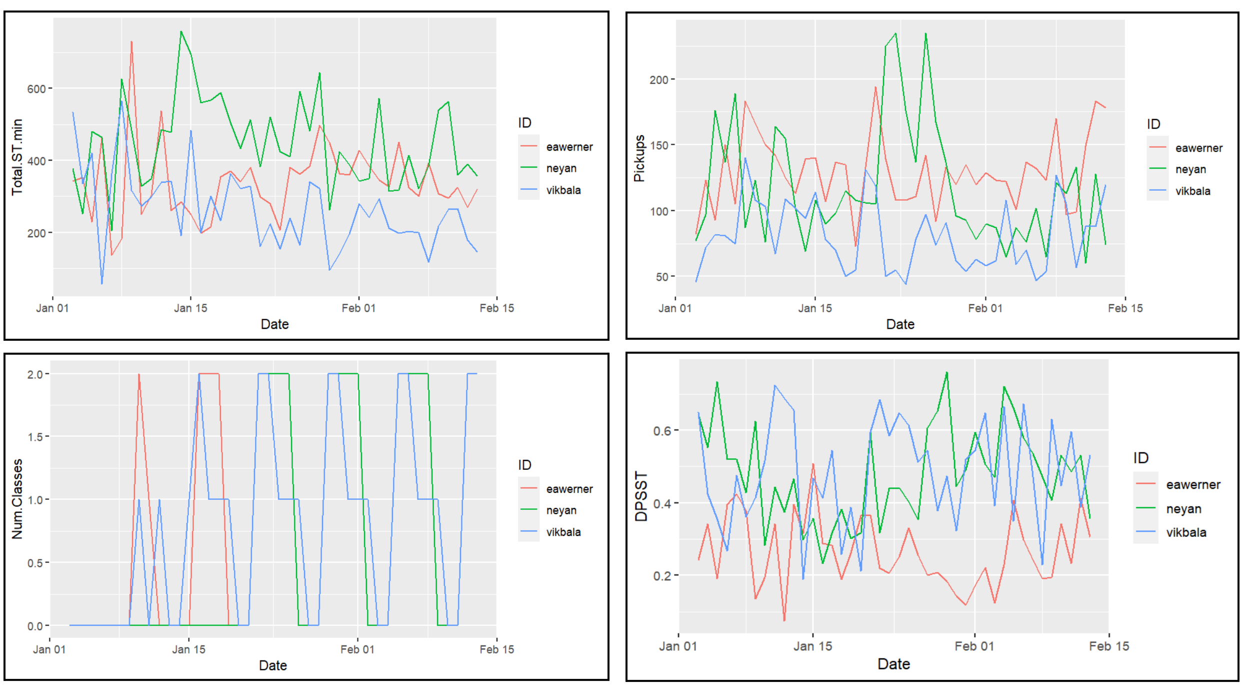


Appendices 3-6. Boxplots of Selected Variables by “ID”.

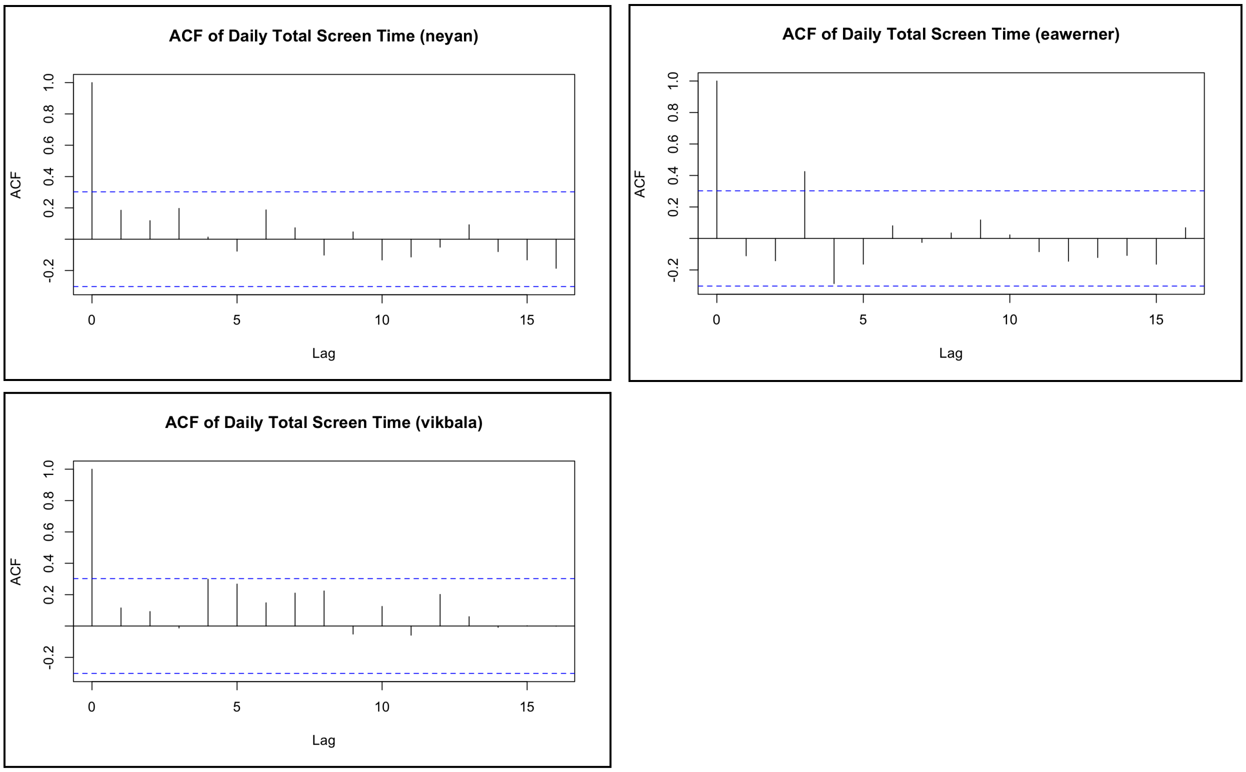


Appendix 7. Pairwise Scatterplots of Selected Variables by "ID".

Green = vikbala; Black = eawerner; Red = neyan.



Appendices 8-11. Time Series Plots of Selected Variables by "ID".



Appendices 12-14. ACF Plots of Daily Total Screen Time by "ID".