INTRODUCTION

Understanding the network of factors that contribute to academic success is an essential challenge in education and psychological research. A student’s Grade Point Average (GPA) is a connected factor to their overall well-being, including lifestyle behaviors like physical activity and sleep. Historically, research in this domain has relied on subjective self-reporting, a method often cluttered with inconsistencies and biases. The usage of personal sensor technology, primarily through smartphones and wearables, provides an important opportunity to gather continuous and objective data on daily behaviors. These data streams allow for more exploration and analysis on the complex relationship between lifestyle patterns and academic performance.

This study takes advantage of that opportunity, performing an exploratory analysis on sensor data collected from a cohort of college students over approximately two months, from April 1st to June 1st, 2013. The research investigates the correlations between these passively collected sensor streams, particularly activity inferences and their temporal distribution, and key outcome variables such as GPA, online engagement metrics, and self-reported sleep data. Our data collection frames an essential research question: **can physical activity, when analyzed alongside other variables, lead to higher student academic performance?**

Early analysis reveals that stationary activity was overwhelmingly dominant, accounting for more than 75% of all activity readings for most participants. Also, a general decreasing trend in daily activity was observed across the cohort as the academic term progressed. This finding is consistent with the data collected from the original StudentLife study conducted at Dartmouth (“StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones”), which reported that,” student activity levels tend to decline as academic workloads intensify over a term.” The data’s completeness also varied significantly among users, highlighting the inherent "messiness" of real-world sensor data that must be addressed as a technical challenge. Studies predicting workplace performance from similar noisy and incomplete data sources also identify this as a real challenge, demonstrating the need for exploratory analysis techniques capable of handling the data (“Jointly Predicting Job Performance, Personality, Cognitive Ability, Affect, and Well-Being”).

By applying various exploratory analytical techniques, this paper details these findings. It aims to contribute to the understanding of how quantitative data can uncover a qualitative narrative about the connections between the daily lives of college students and their academic success.

METHODS

This analysis is based on the publicly available StudentLife dataset (“StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones”). The methodology for this study can be broken down into several stages: participant and data selection, data cleaning and imputation and predictive modeling.

Participants and Data Collection

The dataset is collected from 48 undergraduate and graduate students at Dartmouth College, who used a university-provided Android smartphone for the duration of a 10-week (2 and a half months) academic term in the Spring of 2013. The primary means of data collection was the "StudentLife" application, a custom-built Android app that ran passively in the background to collect and upload data from the phone sensors. For this study, the key data stream was activity inferences, which collected and provided output markers for the user's physical state every 2-3 seconds. The primary markers for this analysis are: stationary, walking, running, and unknown.

The primary variables selected for analysis were the daily activity counts for each of the four activity inference types:

- activity_0 (stationary)
- activity_1 (walking)
- activity_2 (running)
- activity_3 (unknown)

A total activity reading variable was also used to represent the sum of each activity.

Data Cleaning and Imputation

Initial exploratory analysis showed significant data sparsity and missingness for certain days and users. To construct a complete dataset suitable for time-series modeling, an imputation strategy was developed. A "complete day" was defined by a threshold of about 6,000 readings, a value derived from the 25th percentile of daily reading counts for the entire data set. Using this lower-bound value was more appropriate than using the mean, as it better preserves the natural downward trend present in the time-series data. Days falling below this threshold were removed and subsequently replaced with this value to provide consistent time series data for the ARIMA model.

Exploratory Data Analysis and Visualization

To better understand trends in data completeness, several exploratory visualizations were created. These included a heatmap of daily data completeness to identify user-specific data gaps and a bar chart of cross-subject completeness to show the number of users providing sufficient data on each day of the study.

Predictive Models

To model and predict daily activity levels, both regression and time-series forecasting approaches were utilized. First, Simple Linear Regression and Polynomial Regression (degree 2) models were applied to capture the overall activity trend across the course of the study. A comparison of the Root Mean Squared Error (RMSE) and Symmetric Mean Absolute Percentage Error (sMAPE) showed the polynomial model provided a slightly better fit (RMSE: 128,918.49, sMAPE: 62.71%) than the linear model (RMSE: 139,820.58, sMAPE: 63.87%).

However, the high sMAPE values for both models indicate that predicting daily activity with simple regression models is still a significant challenge. To better account for the time-dependent structure of the data, an ARIMA (Autoregressive Integrated Moving Average) model was implemented to predict the time series of each of the four activity types. The goal was to improve predictive accuracy of activity patterns over time.

RESULTS

The results of this exploratory analysis are described in three parts: an overview of the activity patterns and data quality, the performance of predictive models for activity trends, and a correlation table connecting behavioral data to academic performance.

Activity Patterns and Data Quality

Total daily activity inferences ranged from 80,000 to 150,000 readings, with peaks in mid-April before showing a downward trend toward the end of the school year. Among the activity types, stationary time (activity_0) was the most dominant, accounting for 85% to 98% of the day for nearly all participants. The remaining activities, in decreasing order of prevalence, were walking, running, and unknown.

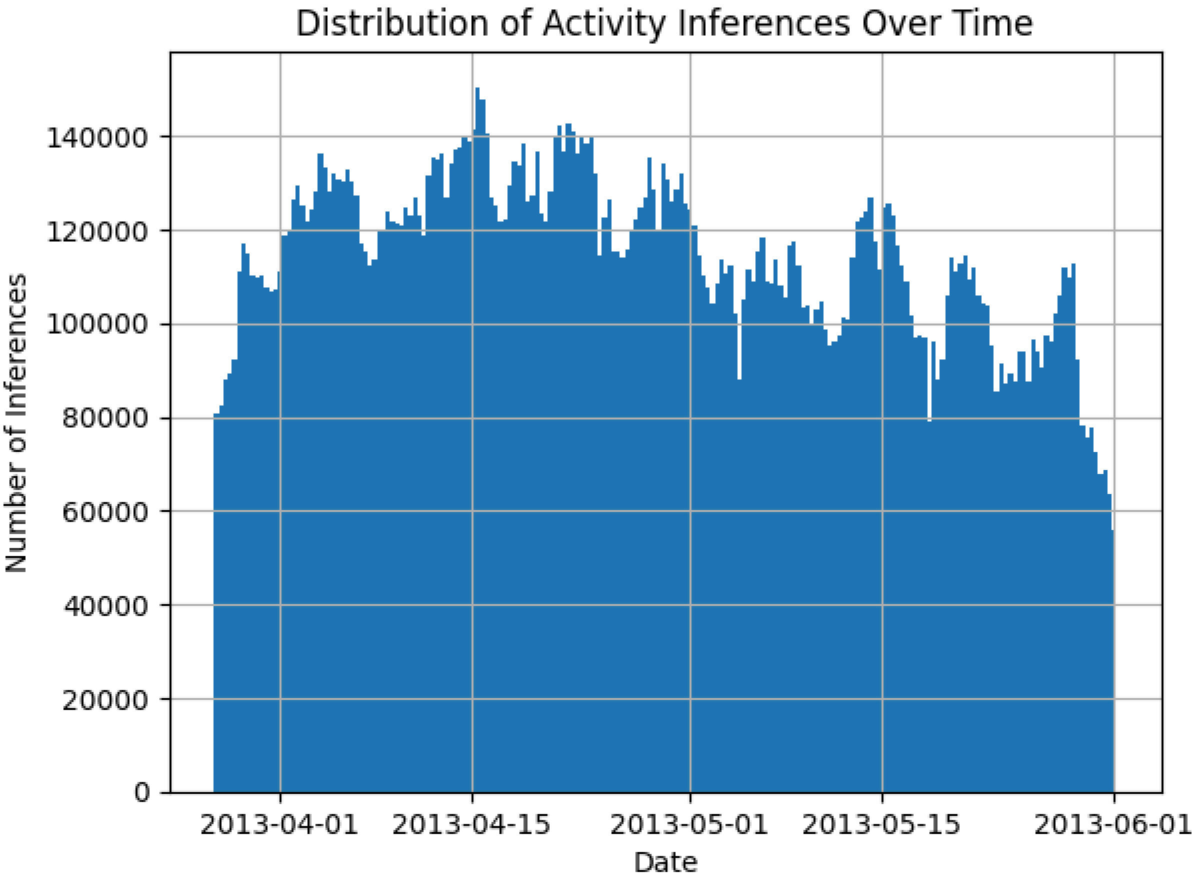


Figure 1: Distribution of Daily Activity Inferences Over Time. This plot shows the total volume of activity data collected across all users for each day of the study, showing an overall decrease in data collection toward the end of the term.

While data collection was mostly consistent, analysis revealed instances of missing data for certain days and users. For most of the study, between 40 and 45 participants contributed to the study, but this number declined significantly near the end of May.

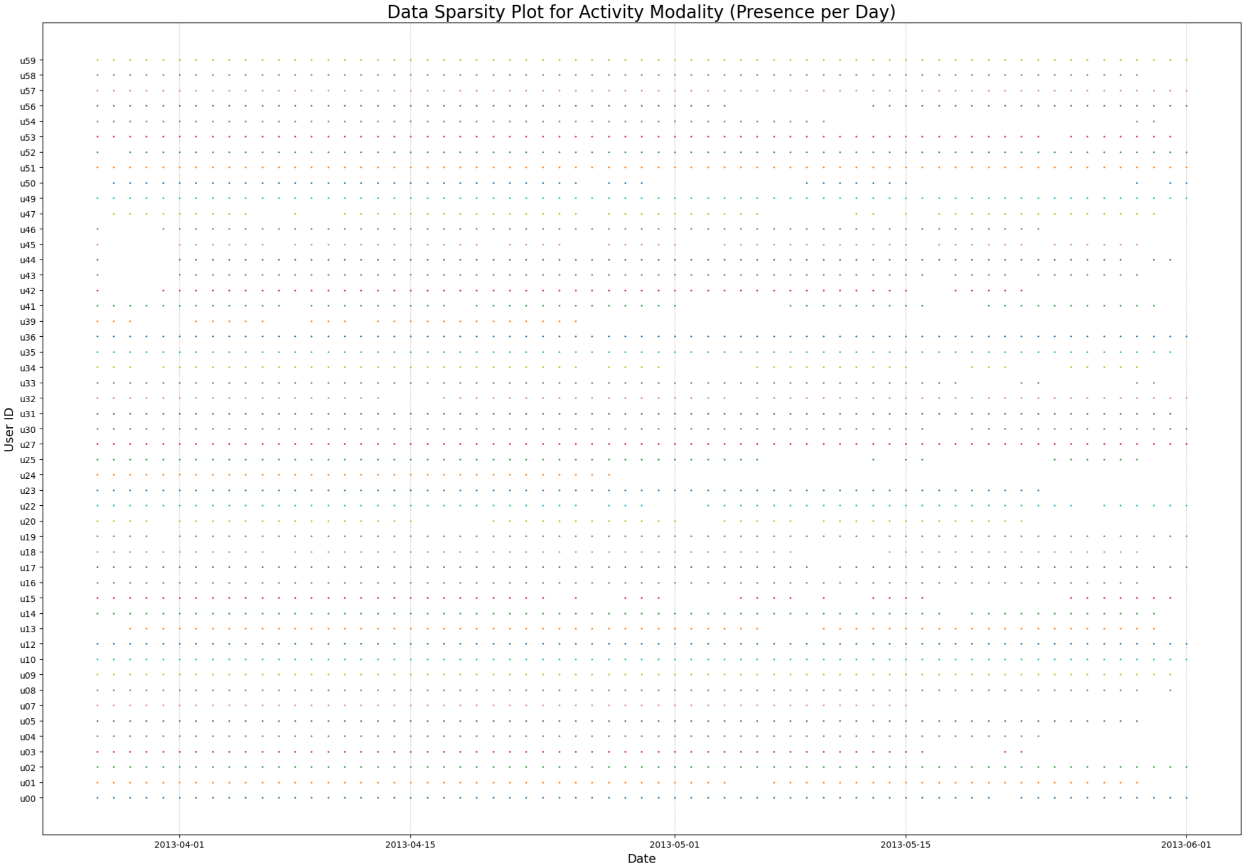


Figure 2: Data Sparsity Plot for Activity Modality. Each dot represents a day where a user recorded a specific type of activity.

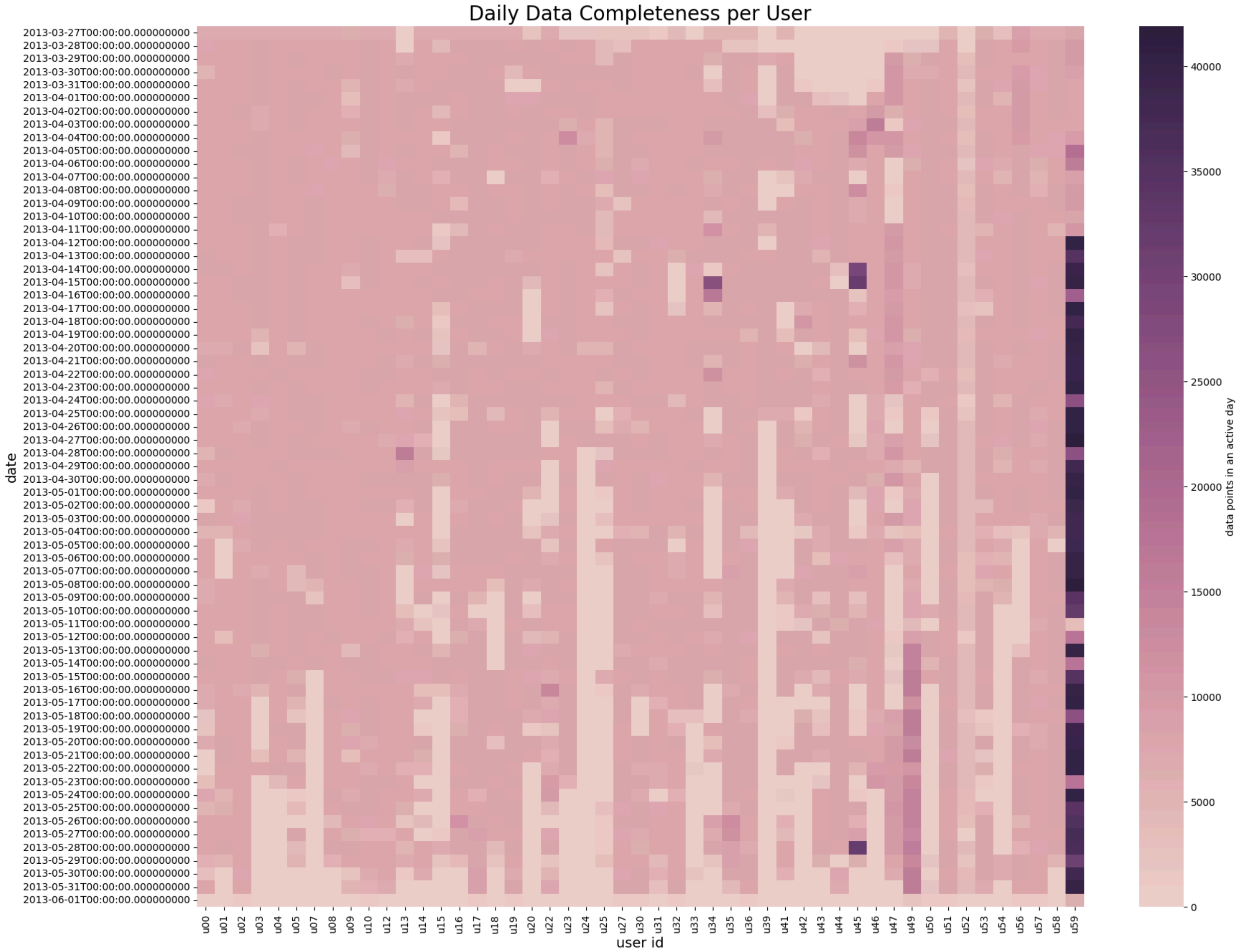


Figure 3: Heatmap of Daily Data Completeness per User. This visual shows user-specific (y-axis) and day-specific (x-axis) data missingness, providing the evidence needed for imputation.

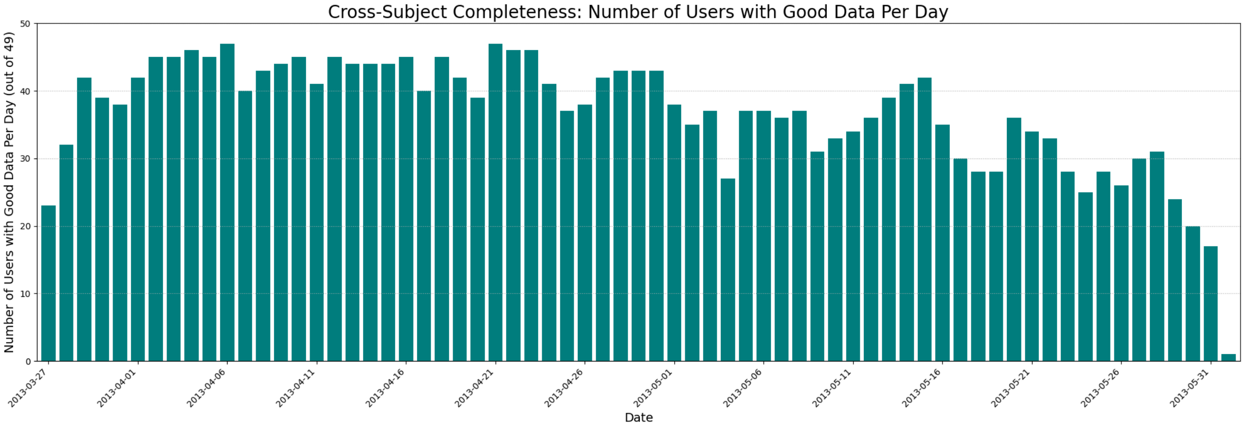


Figure 4: Cross-Subject Completeness. This chart illustrates the daily changes in the number of participants providing a sufficient amount of data, highlighting specific dates with data loss. A good day is defined by the volume of data a participant provides. Specifically, a day is considered “good” if it contains at least 6,000 activity readings (established as the 25th percentile of all daily reading counts).

Modeling Activity Trends

A general decrease in daily activity levels was observed over the study period. To model this trend, regression and time-series models were applied. The Polynomial Regression (degree 2) model provided a better fit for the semester-long activity patterns than a Simple Linear Regression model, indicated by a lower RMSE (128,918 vs. 139,820) and sMAPE (62.7% vs. 63.9%). An ARIMA model was also implemented to forecast future activity based on past data.

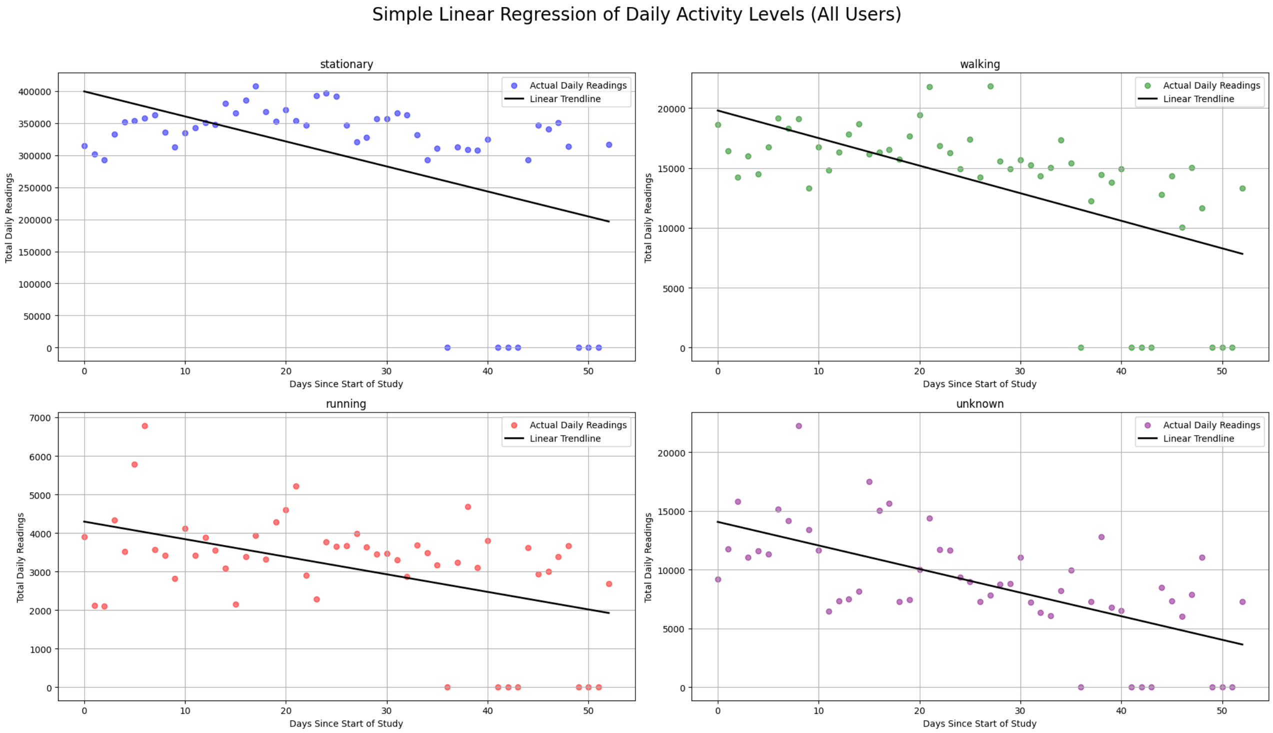


Figure 5: Simple Linear Regression of Daily Activity Levels. This model fits a straight line to the activity data, showing a general decreasing trend for all activity types over the course of the semester.

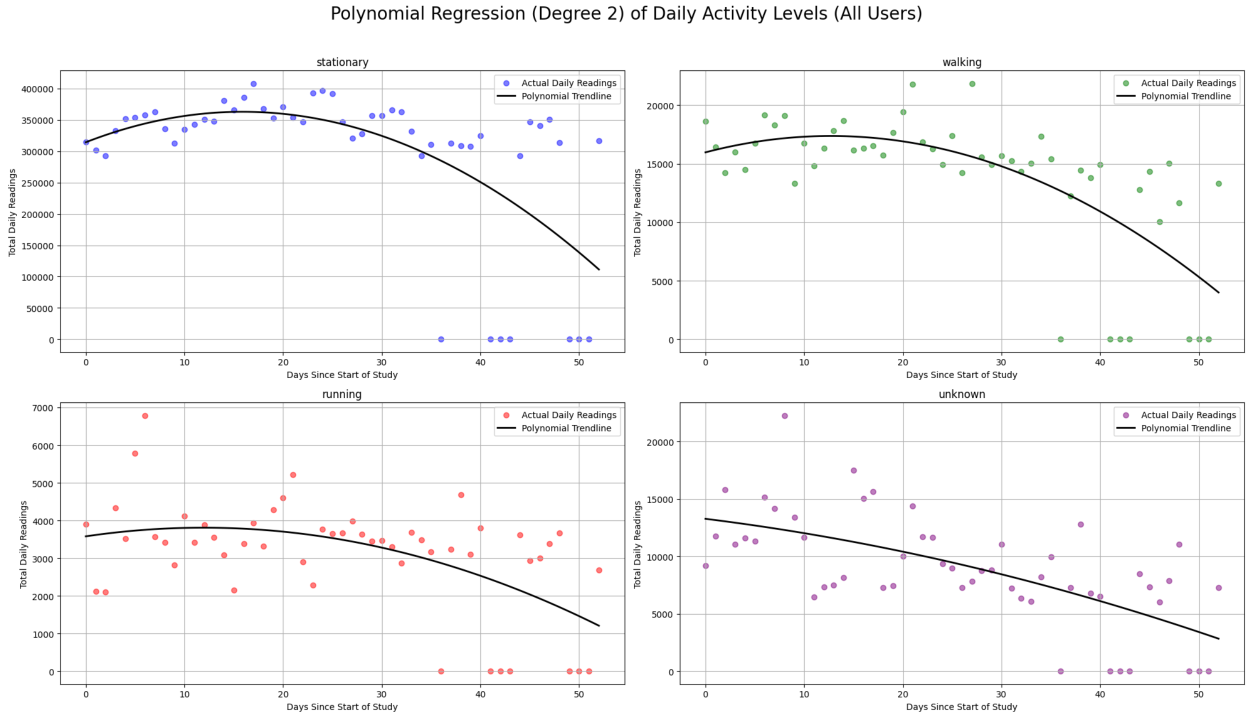


Figure 6: Polynomial Regression (Degree 2) of Daily Activity Levels. The curved trendline captures a better pattern of activity, providing a better fit than the linear model, particularly for stationary and walking activity.

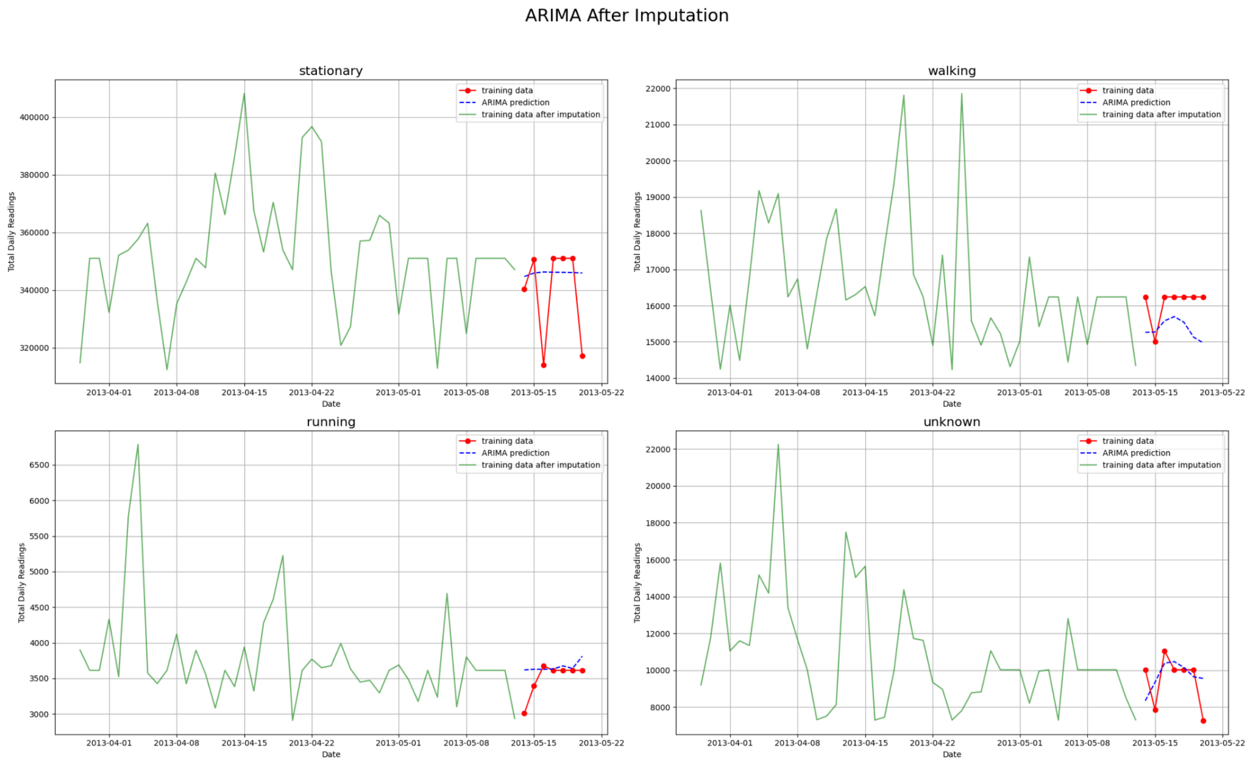
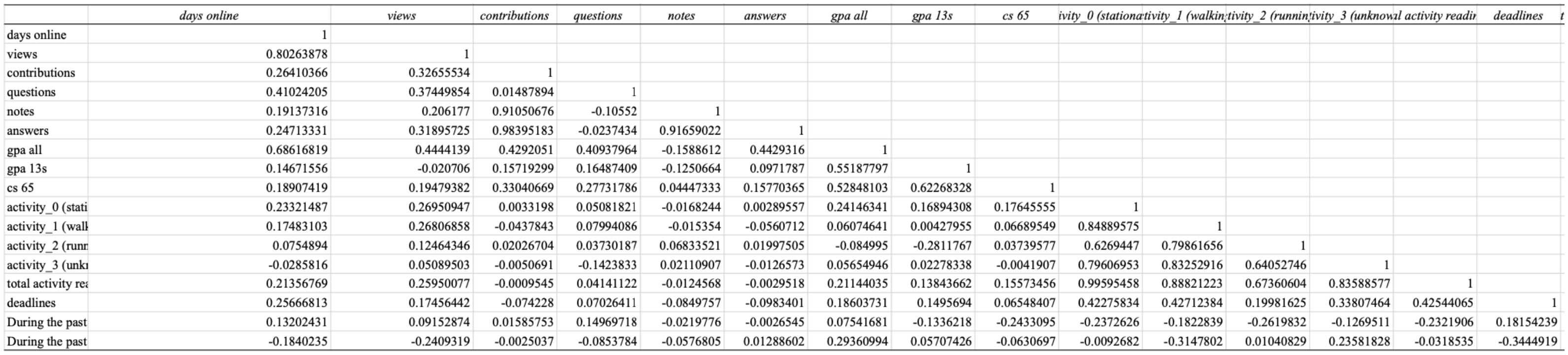


Figure 7: ARIMA Model Forecast After Imputation. This plot shows the model's prediction (blue dashed line) against the actual training data (red dots), showing its ability to forecast a trend for each activity type.

Correlation Tables

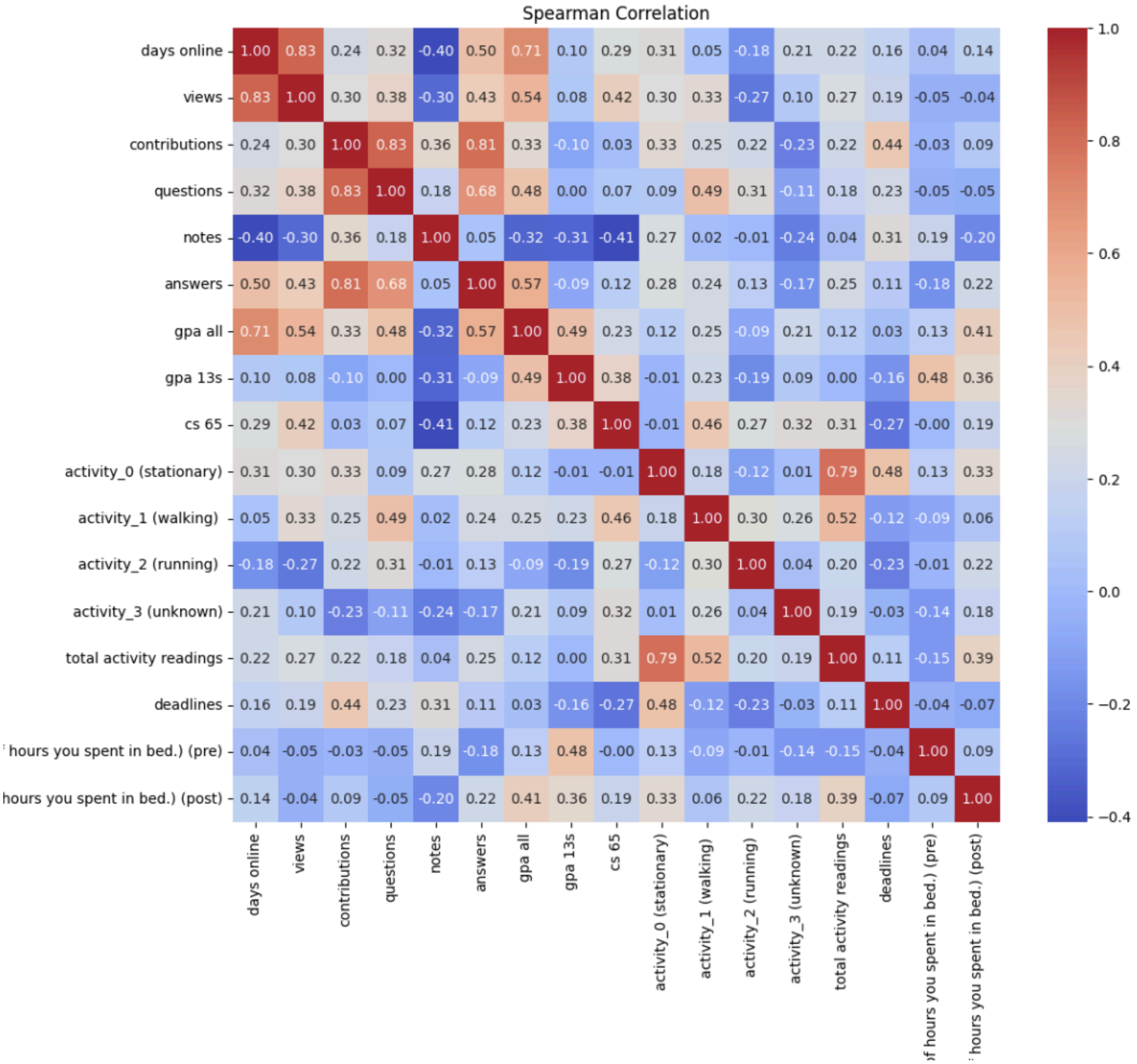
Correlation tables were built to investigate relationships between physical activity, online academic engagement (via the Piazza app), and overall GPA. The primary finding was that relationships between physical activity and GPA were significantly weaker than the stronger correlations between online engagement and GPA.

Physical Activity & GPA: The activity correlations themselves were weak and inconsistent. Stationary activity showed a weak positive correlation (Pearson r = +0.24, Spearman r_s = +0.12), which is consistent with a weak positive linear relationship. Walking showed a divergence where the Spearman correlation (r_s = +0.25) was stronger than the Pearson correlation (r = +0.06), suggesting that its relationship with GPA is non-linear. The similar weak negative values for running (Pearson r = -0.08, Spearman r_s = -0.09) suggests no evidence of a non-linear relationship. These mixed results show that the link between physical activity and GPA (whether linear or non-linear) depends on the specific activity being measured.



	days online	views	contributions	questions	notes	answers	gpa all	gpa 13s	cs 65	ivity_0 (stationa	tivity_1 (walkin	tivity_2 (runnin	tivity_3 (unknow	al activity readin	deadlines	t
days online	1															
views	0.80263878	1														
contributions	0.26410366	0.32655534	1													
questions	0.41024205	0.37449854	0.01487894	1												
notes	0.19137316	0.206177	0.91050676	-0.10552	1											
answers	0.24713331	0.31895725	0.98395183	-0.0237434	0.91659022	1										
gpa all	0.68616819	0.4444139	0.4292051	0.40937964	-0.1588612	0.4429316	1									
gpa 13s	0.14671556	-0.020706	0.15719299	0.16487409	-0.1250664	0.0971787	0.55187797	1								
cs 65	0.18907419	0.19479382	0.33040669	0.27731786	0.04447333	0.15770365	0.52848103	0.62268328	1							
activity_0 (stati	0.23321487	0.26950947	0.0033198	0.05081821	-0.0168244	0.00289557	0.24146341	0.16894308	0.17645555	1						
activity_1 (walk	0.17483103	0.26806858	-0.0437843	0.07994086	-0.015354	-0.0560712	0.06074641	0.00427955	0.06689549	0.84889575	1					
activity_2 (runn	0.0754894	0.12464346	0.02026704	0.03730187	0.06833521	0.01997505	-0.084995	-0.2811767	0.03739577	0.6269447	0.79861656	1				
activity_3 (unkı	-0.0285816	0.05089503	-0.0050691	-0.1423833	0.02110907	-0.0126573	0.05654946	0.02278338	-0.0041907	0.79606953	0.83252916	0.64052746	1			
total activity rea	0.21356769	0.25950077	-0.0009545	0.04141122	-0.0124568	-0.0029518	0.21144035	0.13843662	0.15573456	0.99595458	0.88821223	0.67360604	0.83588577	1		
deadlines	0.25666813	0.17456442	-0.074228	0.07026411	-0.0849757	-0.0983401	0.18603731	0.1495694	0.06548407	0.42275834	0.42712384	0.19981625	0.33807464	0.42544065	1	
During the past	0.13202431	0.09152874	0.01585753	0.14969718	-0.0219776	-0.0026545	0.07541681	-0.1336218	-0.2433095	-0.2372626	-0.1822839	-0.2619832	-0.1269511	-0.2321906	0.18154239	
During the past	-0.1840235	-0.2409319	-0.0025037	-0.0853784	-0.0576805	0.01288602	0.29360994	0.05707426	-0.0630697	-0.0092682	-0.3147802	0.01040829	0.23581828	-0.0318535	-0.3444919	

Table 1: Pearson Correlation Matrix. This table shows the linear correlation coefficients (r) between all study variables.



	days online	views	contributions	questions	notes	answers	gpa all	gpa 13s	cs 65	activity_0 (stationary)	activity_1 (walking)	activity_2 (running)	activity_3 (unknown)	total activity readings	deadlines	f hours you spent in bed.) (pre)	hours you spent in bed.) (post)
days online	1.00	0.83	0.24	0.32	-0.40	0.50	0.71	0.10	0.29	0.31	0.05	-0.18	0.21	0.22	0.16	0.04	0.14
views	0.83	1.00	0.30	0.38	-0.30	0.43	0.54	0.08	0.42	0.30	0.33	-0.27	0.10	0.27	0.19	-0.05	-0.04
contributions	0.24	0.30	1.00	0.83	0.36	0.81	0.33	-0.10	0.03	0.33	0.25	0.22	-0.23	0.22	0.44	-0.03	0.09
questions	0.32	0.38	0.83	1.00	0.18	0.68	0.48	0.00	0.07	0.09	0.49	0.31	-0.11	0.18	0.23	-0.05	-0.05
notes	-0.40	-0.30	0.36	0.18	1.00	0.05	-0.32	-0.31	-0.41	0.27	0.02	-0.01	-0.24	0.04	0.31	0.19	-0.20
answers	0.50	0.43	0.81	0.68	0.05	1.00	0.57	-0.09	0.12	0.28	0.24	0.13	-0.17	0.25	0.11	-0.18	0.22
gpa all	0.71	0.54	0.33	0.48	-0.32	0.57	1.00	0.49	0.23	0.12	0.25	-0.09	0.21	0.12	0.03	0.13	0.41
gpa 13s	0.10	0.08	-0.10	0.00	-0.31	-0.09	0.49	1.00	0.38	-0.01	0.23	-0.19	0.09	0.00	-0.16	0.48	0.36
cs 65	0.29	0.42	0.03	0.07	-0.41	0.12	0.23	0.38	1.00	-0.01	0.46	0.27	0.32	0.31	-0.27	-0.00	0.19
activity_0 (stationary)	0.31	0.30	0.33	0.09	0.27	0.28	0.12	-0.01	-0.01	1.00	0.18	-0.12	0.01	0.79	0.48	0.13	0.33
activity_1 (walking)	0.05	0.33	0.25	0.49	0.02	0.24	0.25	0.23	0.46	0.18	1.00	0.30	0.26	0.52	-0.12	-0.09	0.06
activity_2 (running)	-0.18	-0.27	0.22	0.31	-0.01	0.13	-0.09	-0.19	0.27	-0.12	0.30	1.00	0.04	0.20	-0.23	-0.01	0.22
activity_3 (unknown)	0.21	0.10	-0.23	-0.11	-0.24	-0.17	0.21	0.09	0.32	0.01	0.26	0.04	1.00	0.19	-0.03	-0.14	0.18
total activity readings	0.22	0.27	0.22	0.18	0.04	0.25	0.12	0.00	0.31	0.79	0.52	0.20	0.19	1.00	0.11	-0.15	0.39
deadlines	0.16	0.19	0.44	0.23	0.31	0.11	0.03	-0.16	-0.27	0.48	-0.12	-0.23	-0.03	0.11	1.00	-0.04	-0.07
f hours you spent in bed.) (pre)	0.04	-0.05	-0.03	-0.05	0.19	-0.18	0.13	0.48	-0.00	0.13	-0.09	-0.01	-0.14	-0.15	-0.04	1.00	0.09
hours you spent in bed.) (post)	0.14	-0.04	0.09	-0.05	-0.20	0.22	0.41	0.36	0.19	0.33	0.06	0.22	0.18	0.39	-0.07	0.09	1.00

Table 2: Spearman Correlation Matrix. This table shows the correlation coefficients (r_s), calculated based on the rank of the data, between all study variables.

DISCUSSION

A large body of research suggests that physical activity is an important factor in academic success. The physiological and mental health benefits of exercise are well-established. Therefore, an assumption can be made that a more physically active student might be more focused, less stressed, and more cognitively aware, leading to a higher GPA. However, the evidence from this exploratory analysis provides a more ambiguous picture, aligning with assumptions where the direct link between physical activity and GPA is surprisingly weak.

This study found that correlations between specific activities and GPA were inconsistent. For example, stationary activity showed a weak positive correlation with GPA (Pearson r = +0.24, Spearman r_s = +0.12), while walking showed a divergence where the Spearman correlation (r_s = +0.25) was stronger than the Pearson correlation (r = +0.06). Running showed a very weak negative correlation (Pearson r = -0.08, Spearman r_s = -0.09). This suggests that for this cohort of students, increased sedentary activity may be more time dedicated to studying. Also, while walking is generally positive for academic performance, the benefit isn't a straight line.

Looking holistically at the data, this analysis suggests that the type of activity matters in relation to GPA. It's possible that a low positive signal for low-intensity physical activity like walking and a very low negative signal for higher-intensity activities like running suggests that, for this cohort of students, low-impact activity is more beneficial than more intense forms of exercise, which showed no positive link at all. Tying everything together, these thoughts allow for drawing a conclusion that the relationship between physical activity and GPA is more complex than we think and dependent on the type of the activity.

In contrast, the relationship between online academic engagement and GPA provided much stronger and more consistent analysis. Variables such as answering questions on Piazza were positive predictors of GPA, indicating that direct academic engagement is more strongly correlated with performance than the physical activity metrics captured in this dataset.

Limitations

It is important to acknowledge the limitations of this exploratory study. The findings are based on a small sample size (n=48) from a single ivy-league college, which may not be applicable in other academic settings. Also, this analysis only implies correlation and not causation. There are many potential variables, such as major, class difficulty, and previous academic success that were not applied to this analysis.

The full StudentLife dataset contains other valuable data streams, such as mood and stress surveys, which could be integrated for a more holistic behavioral picture. I also believe that repeating this study with a larger and more diverse student population would be necessary to validate assumptions.

Conclusion

Passive sensor data from smartphones is a powerful tool for measuring human behavior. While this study did not find a direct link between physical activity and GPA, the key takeaway here isn’t that exercise is not important to student performance, but that it is just one piece of a complex picture of student behavior. This reinforces the need to promote a more holistic view of student success that considers a wide range of behavioral health, mental health and contextual variables. Future studies may aim to discover the link between these multiple variables to measure the connection between physical activity and GPA.