Sport activities among Trento's university students

Chemotti mattia.chemotti@studenti.unitn.it,
Dall'Acqua maja.dallacqua@studenti.unitn.it,
Vandelli davide.vandelli@studenti.unitn.it,

Abstract

Sport is one of the most common leisure activities among people, especially in youth. To describe and understand the relationship between sports and living as an adult student helps us understand more about human and social behavior, and may serve as a solid indication for policies on health and wellness. Often, research on the effect of practicing sports shows that it can positively affect many areas of life, ranging from individual bodily health to social integration and overall wellness. The current project aims to identify the main characteristics of sport activities across different dimensions. In order to provide a comprehensive analysis of students' sport behavior in the university of Trento, the available data is submitted, modeled and visualized through a thorough process of selection for trustworthiness.
The code implemented for the project is available at: project's GitHub repository.

Keywords
Sport activities, Temporal analysis, Attitudes.

Introduction

According to the World Health Organization (2022), people of age 18-64 should do at least 150-300 minutes of moderate-intensity physical activity grouped on two or more days per week and at the same time limit the sedentary activities. This behavior is usually difficult to apply in university students' everyday life due to the overlapping of many activities, such as attending classes, studying time, social life time and extra activities. Therefore, studying students' behavior could be of use to understand how sport activities can be encouraged in the university students' community.
Since sport activities may vary considerably, adopting different approaches could be useful to understand deeper the complexity of this subject, hence the approach used in this report.
Our analysis considers three different dimensions to analyze the data, as shown in Table A.

Table A. Diagram for different dimensions accounted.

main dimension	*research question(s)*
Socio-demographic dimension	How are sport activities distributed in the students' community? Which are the socio-demographic factors that influence the type of activity practiced by students?
Psychological dimension	Is there a propensity from students with certain traits and attitudes towards sport?
Temporal dimension	How are sport activities distributed across the week? How are sport activities distributed across the day?
Sport and Steps	Is it possible to validate users' sport activities through the step detector sensor? Are there any differences between students who practice sports and students who's step detector is often activated? Are there any socio-demographic factors that influence the intensity of sport and step activity?

Processing raw data

Data collection

The datasets used for the analysis are from the WeNet project (first established in 2019) and refer to an initial sample of 249 students from the University of Trento, whose activities were collected during the month of November and December in 2021 through the iLog application on users' smartphones.
For the current report three different datasets were used:

- Time diary: contains all the self-reported events that each user was involved in every 30 minutes. This dataset refers to the 249 subjects who participated in the project.
- Demographic survey: contains social, psychological and demographic information about the users. This dataset refers to the 249 subjects who participated in the project.
- Step detector: contains information from the step detector sensor from each user's smartphone. This dataset refers to the 124 subjects who had the step detector sensor switched on during the entire period of the project.

Several pre-processing steps were applied on all datasets before conducting the analyses.

Data cleaning

In the first two weeks of the WeNet survey, from the 13th to the 30th of November 2021 regard an interval where participants received a prompt for their time diary every half hour, and later it had gotten less frequent, meaning that this interval has more frequent eventful data. Only the users who participated consistently during this interval were considered.

Given the structure of the data and WHO's definition of physical activities, only those who reported at least consecutively twice[1] "Sport" in their time diary were considered, and the same criterion is applied for those who expressed "Walking". Even though we possess walking sensor data, at times the users who expressed "Walking" did not submit their sensor data, or vice versa those who did submit sensor data had not expressed in their time diary what they were doing - and this discrepancy calls for a separate analysis until a better solution allows to proceed.
From this selection it was possible to identify a subset of 128 users who did at least one physical activity during the period of study.

Because of the step detector data's specificity[2], the latter was first aggregated to obtain the total number of steps every 30 minutes[3] per each user and day, in order to compare this dataset with the time diary data.

Data preparation

Considering the different types of analyses to be performed, several ways of selecting data were produced during the data preparation process. To our favor, all the users who participated in sports at least once also were consistent with their time diary submittal every day and never dropped out.

1. Operations for socio-demographic analyses

In order to estimate how long were sport sessions, the consecutives answers in this new dataset were aggregated by time adjacency for each user, leading to 1028 observations. Each row then represents an event in which one user expresses doing sport for a time that lasts from anywhere between one hour to no upper limit, depending on how many times the user originally reported it.

[1] This choice allows us to make sure the duration of sport activity is more than half-hour, reducing the possibility that it was a mistake from the user' multiple answer selection in the time diary survey.
[2] The way a step detector creates data is: each time the sensor is activated, one event is produced and associated with a timestamp. In order to know how many steps a user took, we needed to count / aggregate them in a given time frame.
[3] The 30 minutes time frame is to match the same interval present in the time diary data, as anything more accurate could be useless and anything more broad could be lossful.

From this data two more operations are performed to analyze "Walking". **First**, only the events in which the walking session fell in the range from one hour to no upper limit are considered as not simply a commute, but rather a physical activity fitting WHO's approach. **Second**, the times in which a user expressed they were walking "indoors" (such as 'grocery', 'university', …) meant they were probably just moving from some room to another, and "walking" is likely to be not a consistent session. These choices do not shield the remaining observations from the possibility of bias, especially from the user's side, but they may be the most appropriate solution for avoiding semantically unfitting observation[4].
Ultimately, if the filtering is done as we did, one should obtain our dataset of 864 events, related to 128 users who practiced physical activity at least one time over the two weeks.

Considering the initial categories in which the variables were encoded, to simplify the analyses some of these variables were reclassified. The location of the events became labeled as "**indoor**" or "**outdoor**"; the social setting of the event became either "**alone**" or "**with company**". Only a few of the possible answers for the type of sports were used by users, and all fall in these 5 categories: "**jogging and running**", "**walking, trekking and hiking**", "**outdoor activities**" (ball games, cycling, skiing), "**gymnastics and fitness**", "**indoor activities**" (water sports and unspecified others). In addition to these answers that belong to users who declared they were doing sports, we remind you that we added the times in which they *did not* select sport but (probably) mistakenly selected just "walking" from another option, when they had been walking for one hour or more. After this features' reclassification, we incorporated demographic information on the user involved in event data.

2. Operations for psychological traits analyses

To prepare our dataset for ***clustering psychological traits***, we first addressed the issue of missing values, and employing the SimpleImputer from sklearn's impute module, we replaced missing entries, ensuring a complete dataset for more accurate analysis. Following this, we utilized both StandardScaler and RobustScaler for data normalization. The use of these scaling techniques is crucial as K-means clustering is sensitive to the scale of the data, and our aim was to mitigate the impact of outliers and ensure that each personality dimension contributed equally to the analysis.

3. Operations for sport and steps comparison

When preparing data for ***sport and steps*** comparison we found one practice to be useful: comparing side-by-side steps and sport events. This was not always possible, due to the fact that only 124 users submitted their sensor data and there are 55 different users from the sample of those who did sport in

[4] To give a more concrete example of our reasoning, if a user says they were walking at 11:30, in their university department, and then the same answer at 12:00, it's likely they didn't mean "walking" the same way multiple users expressed from 10:00 until 13:00 they were walking in the mountains of Trento, IT, as the latter may be recognized as hiking, a spread physical activity between students.

the time diary. For this dimension, we preferred starting over with the users sample: all users we had data over were being classified regarding sport activity.

Table B. Shaping features to compare.

Sport activity (time diary data)	**Step activity** (sensor data)
Inactive (no physical activity)	**Not very active** (I° tertile)
Not very active (below mean of physical activity)	**Active** (II° tertile)
Active (above mean physical activity)	**Very active** (III° tertile)

Levels of activity had to be classified in an ordinal way, in order to match the continuous nature of steps count. The user is then placed in one of three ordinal categories, that put on the same "scale" two different variables, as shown in Table B.
Physical activities is a categorical variable, and can only become quantitative if the events of that variable occurring are counted. Differently, Step Count is not easily interpretable in this setting, so ordering it in three categories matches more Sport Activity's nature.

4. Operations for temporal analyses

In this report we use the terms "temporal analyses" and similar to indicate our efforts for visualizing patterns rather than time-series analysis. Due to the many possibilities given by incorporating time into our analyses, we chose to define time the same way many humans perceive it. Although we cannot define a specific circadian time unit that fits all users' specific characteristics, a day is exactly 24 hours. Furthermore, we know that around 05:00 nearly 100% of users' activity is at an all time low; they are not using their phone, and their step sensors are not emitting data. That time counts as the turning point for the users' day: if no one is active, it means that 05:00 is the best threshold for discerning if an event was still belonging to the previous day but past midnight, or the early morning of the day "after". Although it might seem trivial, this threshold divided the sample user activity in a clean manner, and it fits more the meaning of the datum on a semantic level for interpreting events: if a user is awake, practicing sport Friday at 00:30 (and the data shows that some do), using only the timestamp makes it seem like the user was practicing sport the day after their Thursday ended, when in reality the user still perceives their day to be not ended, and that they were practicing sport on a late Thursday night. To account for this dynamic, when trying to group activities by day following human time perception, every event occurring before 05:00 is labeled as part of the previous calendar day.

Data analysis

Socio-demographic analysis

The current analysis comprises two distinct sections, each serving a specific purpose. In this section, our focus is on understanding students' sport sessions and predicting the likelihood of engaging in a particular type of sport activity over another. Our approach incorporates various factors:

- **Event-Related Features:** examining specifics of the event, including location and participants.
- **Socio-Demographic Factors:** considering user characteristics such as sex, department and related information.

We employed regression models to facilitate these analyses. The objective of the second section is to predict the frequencies of users' step and sport activities based on their socio-demographic features.

Structural Overview of the data analysis:

1. **Model Descriptions:** the initial part briefly outlines the implemented models.
2. **Results:** the subsequent part unveils the primary outcomes.

Model choices and explanation

Regarding the sport session analysis, it was considered important to apply some recodification to the dataset before proceeding any further. In order to use machine learning (implementing methods from Sci-Kit)[5] we decided to transform the variables into dummies through a one-hot encoding computation, avoiding the risks of creating new information through a numerical encoding that the regression model would have then used during training. Having then a dataset with binomial variables, we decided to implement a binomial regression model.

Regarding the sport-step activity the decision on the model to implement was more debatable. Having an ordinal target variable the most suitable approach would have been an ordinal probit regression (Daykin *et al*., 2002), nevertheless we opted for a logistic regression model in order to still be able to validate the results through a cross-validation approach[6]. Therefore the choice was between:

- adopting an ordinal probit model which considered an additional feature of the data (the ordered categories) not validated.

[5] For a full description of the methodology behind these models, please refer to the official documentation of python's Scikit-learn library.

[6] Indeed the Scikit-learn library (for the cross-validation procedure) and the statsmodel library (for the ordinal probit regression) are incompatible and cannot be implemented together.

- using a logistic regression losing some information on the data but being able to validate the model's performance through cross-validation.

Considering advantages and disadvantages of both implementations, in the end it was decided to consider the latter approach. Thus a **logistic regression model** was implemented for predicting both the type of sport activity and the level of sport-step activity.

The main steps of this implementation are then the following:

Dealing with unbalanced classes: from the univariate analysis of the dataset, it appeared that some of the variables had highly unbalanced classes. Thus it was decided to eliminate the predictors with a high disproportion (the threshold used was 10-90) and to robustify the analysis through a balancing technique for the remaining predictors. A mixed method was implemented, combining the SMOTE technique (for undersampling) and the Tome Links technique (for oversampling) to obtain a bigger dataset (with balanced classes) for both predictors and the target variable (Swana *et al*., 2022).

Model computation: the logistic regression model is initialized. Since all variables can assume only values 0-1 a binomial model is preferred. A 5-folds cross-validation through a GridSearch approach is first applied to the model to identify the best combination of parameters to use. A set of values for each of these parameters[7] was chosen to be tested to find indeed the combination with the best performance. Then this optimal model is selected.

Model Evaluation: a 5-fold cross-validation approach was implemented to provide accurate results. The model trained is then evaluated through the following classification evaluation metrics:

- *accuracy*: measures how accurate the model is in predicting the outcome.
  It is obtained through the formula $\frac{correct\ predictions}{overall\ predictions}$.
- *precision*: measures how accurate the model is in predicting the positive class.
  Considering the confusion matrix, this metric is calculated through the formula $\frac{True\ positives}{True\ positives + False\ Positives}$.
- *recall*: measures how accurate the model is into identifying the positive instances from the actual positives observation in the dataset. It is computed through the formula $\frac{True\ positives}{True\ positives + False\ Negatives}$.
- *F1-score*: combines precision and recall scores to compute the harmonic mean of these two measures through the formula $\frac{2*Precision*Recall}{Precision + Recall}$.
- *ROC_AUC*: measures how accurate the model is in distinguishing between classes.

Feature selection: through bivariate analysis a subset of predictors is selected from the initial set, considering their association with the target variable. Cramer's V test is used to measure the strength

[7] The parameters considered were: cost, penalty, class weight, solver, number of iterations.

of the relation between each predictor and the target variable and the significance is measured through the p-value. The initial threshold is settled to 0.9 and the p-value is considered as strongly significant. These values are recursively lowered until at least one predictor satisfying the current requirements is found. Then this subset of predictors is selected for the following steps of the analysis.

Second computation: the model is re-fitted considering the new set predictors.
As for the previous case, a GridSearch method is previously applied to identify the model with the best combination of parameters. Then, the model is fitted to the subset of predictors and the target variable, after this computation the new model is evaluated through the same evaluation metrics of the first model.

Model selection: by comparing the accuracy of the first model with the accuracy of the second model, the algorithm selects the one with the highest accuracy.

Results

Table C and Table D show respectively the hyper-parameters used and the results provided for each of the models implemented, corresponding to the categories of the **sport activity** variable.
It's important to notice that in Table C the columns *predictors* and *Cramer's V* report the parameters of the models in which the feature selection produced better results, an empty box then implies that the initial model had the best performances. In such cases the predictors used were the available balanced categories, as :

- location variable (indoor, outdoor)
- company variable (alone, in company)
- degree variable (BSc, MSc)
- sex (male, female)
- department (Business/economics, Engineering, Law, Natural sciences, Social sciences)
- neighborhood (Trento center, Trento south)

In almost all the models the implementation using all predictors performed better than the one using the predictors obtained through the feature selection. The two exceptions are the "Walking" model and the "Gymnastics and fitness" model, whose implementations were applied using respectively the location variable and the outdoor category of the location variable as predictors.

Table C. Parameters of the logistic regression models on the type of sport activity.

target	hyper-parameters				
	predictors	Cramer's V[8]	C	penalty	solver

[8] '***' for a p-value lower than or equal to 0.01, '**' for a p-value between 0.01 and 0.05, '*' for a p-value between 0.05 and 0.1, a p-value bigger than 0.1 is not significant.

Walking	location	0.6***	1.0	l1	saga
Walking, Trekking, and hiking		-	1.0	l2	saga
Gymnastics and fitness	location (outdoor)	0.8***	0.1	l1	liblinear
Jogging and running		-	1.0	l1	liblinear
Indoor activities		-	0.1	l1	saga
Outdoor activities		-	10.0	l1	liblinear

By comparing the models' hyper-parameters[9] one can see that most of them vary across the models, which means that there are consistent differences across the categories of the sport activity. Indeed, none of the models utilizes the same combination of parameters of another model.

From Table D we can see that overall the best results are the ones in the "**Gymnastics and fitness**" and the "**Outdoor activities**" models. On the other hand, the "**Indoor activities**" model has the worst results. Overall, the models seem to predict quite accurately the probability of doing a sport activity instead of another although with some differences in the performances[10]. The "**Walking**" and the "**Walking, trekking and hiking**" models, which differ for the pre-processing implementation, have different results, with the latter performing overall better than the former.

To conclude it's important to notice that the difference in the number of predictors doesn't influence directly the models' performances: the "**Gymnastics and fitness**" model has the best results using only one predictor, while the "**Walking**" models, which uses two predictors, has similar results to the models adopting all the available predictors. At the same time the "**Outdoor activities**" model, which also performs quite well, needs all the predictors.

Table D. Results of the logistic regression models on the type of sport activity.

target	evaluation metrics				
	accuracy	precision	recall	F1-score	ROC AUC
Walking	0.83	0.77	0.94	0.85	0.83
Walking, Trekking, and hiking	0.86	0.84	0.89	0.86	0.92
Gymnastics and fitness	0.89	0.83	0.99	0.90	0.89
Jogging and running	0.81	0.76	0.90	0.82	0.89

[9] the number of iterations and the class weight are available at source code.

[10] To compare more accurately the models consult the not-rounded results displayed at the source code.

Indoor activities	0.76	0.72	0.83	0.77	0.80
Outdoor activities	0.89	0.86	0.93	0.89	0.95

In Table E and Table F are shown respectively the parameters and results of the logistic regression model for the step activity and sport activity variables. By comparing step and sport activity results we can derive that overall the model predicting step activity is slightly better than the other one since the evaluation metrics in all categories tend to be higher than the respective ones in the step activity. However it appears that none of the implementations for both models predicts very well users' probability of doing a certain level of step activity and sport activity since the evaluation metrics do not assume very high values.

Table E. Parameters of the logistic regression models on the level of step and sport activity.

target		hyper-parameters				
		predictors	Cramer's V[11]	C	penalty	solver
Step activity	Not very active		-	1.0	l2	saga
	Active		-	10.0	l1	liblinear
	Very active		-	10.0	l1	liblinear
	Inactive	department (Natural Sciences)	0.1**	0.1	l2	liblinear
Sport activity	Active		-	10.0	l2	saga
	Very active		-	10.0	l1	liblinear

By comparing the hyper-parameters one can notice a preference for a high cost and a liblinear solver. As for the sport activity analysis, in Table E are displayed the predictors only for the models in which the feature selection provided better results, while in all the other cases the predictors used are the ones with balanced classes and related to the users' demographics details[12]. The only model in which the feature selection was useful is the one predicting the **sport inactivity**, with the department of Natural Sciences as predictor. Further analysis would be necessary to find an exhaustive explanation to this pattern, but it is important to notice how this model is also the one with the worst

[11] '***' for a p-value lower than or equal to 0.01, '**' for a p-value between 0.01 and 0.05, '*' for a p-value between 0.05 and 0.1, a p-value bigger than 0.1 is not significant.
[12] Therefore one should look at the categories of the degree, sex, department and neighborhood variables from the list at the beginning of the results section.

performances. All models reported can provide some insights on the differences between the levels of step and sport sport activities. Therefore, further analysis and models could be implemented to improve the current results.

Table F. Results of the logistic regression models on the level of step and sport activity.

	target	hyper-parameters				
		accuracy	precision	recall	F1-score	ROC AUC
Step activity	Not very active	0.71	0.67	0.80	0.73	0.76
	Active	0.73	0.72	0.70	0.70	0.82
	Very active	0.69	0.67	0.72	0.69	0.79
Sport activity	Inactive	0.54	0.52	0.93	0.67	0.54
	Active	0.61	0.61	0.61	0.60	0.65
	Very active	0.69	0.66	0.70	0.67	0.77

Step detector analysis

The current analysis is mainly a preprocessing and descriptive section in which we prepare the sensor data to be analyzed in other sections by focusing on temporal patterns and potential correlations with socio-demographic and psychological traits. Utilizing the step count data extracted every 30 minutes, we employed several statistics to discern daily and weekly patterns from physical activity. Our analysis revealed significant variations in step counts across different times of the day and days of the week (discussed in other sections), reflecting students' lifestyles and schedules. However the dataset highlights some challenges due to the nature of the sensor and the choices of the students.

Model choices and explanation

The analysis of step activity is reliant on data from the step detector sensor. However, it is important to note that the sharing of data pertaining to it was at the discretion of individual students during the data collection phase. Therefore, out of the initial 249 users, only 126 were included in this analysis. After the preprocessing and cleaning phase, we proceeded to label each user as “Not very active”, “Active”, “Very Active” dividing the entire distribution in three tertiles based on the student’s average daily step count. By merging this dataset with the sport sessions of our students and analyzing the results, it emerged that in ~ 64.95% of the sport sessions the step detector sensor was turned off or the smartphone was simply not with his owner, meaning that only ~ 35.05% of the observations were

suitable for describing the relationship between sport sessions and step activity. Moreover, we investigated which type of sport activities were influenced the most from this peculiarity, and as can be seen from the following table, “Gymnastics and Fitness” and “Walking” are the most impacted.

Table F. Distribution of sport activity sessions in which the number of steps detected was 0.

Activity	sport sessions with 0 steps	
	N	%
Gymnastics and fitness	311	30.52
Jogging and running	72	7.07
Other indoor activities	89	8.73
Outdoor Activities	34	3.34
Walking	471	46.23
Walking, Trekking and hiking	42	4.11

Personality analysis

The Big Five personality traits, also known as the Five-Factor Model or OCEAN, is a widely accepted framework in psychology for understanding human personality. Each trait represents a continuum (in our dataset a value from 0 to 100), and individuals can fall anywhere along this spectrum, shaping their behavior, preferences, and interaction styles in various settings. In this section, we use clustering methods to explore the peculiarities of different personality clusters present in our students' sample. Therefore, we compare every user of the 4 personality clusters with their step activity type (“Not Very Active”, “Active”, “Very Active”), in order to identify possible correlations between personality traits and students' step activity.

Model choices and explanation

We use **K-means** clustering algorithm to categorize students based on their similarity in terms of the combination of Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. To identify the optimal number of clusters for the K-Means algorithm, we utilize the Elbow method and the Silhouette score. The Elbow method is an empirical approach that looks for a sudden drop in the average distance of each point to its centroid, while the Silhouette method computes the average silhouette of observations for different values of K and maximizes the average silhouette over a range of K. Both methods suggest that 4 should be the optimal number of clusters. For correlation, we use The Pearson correlation coefficient (PCC) to measure the potentially linear relationship between personality and step activity type, ranging from a negative association (-1) to a positive one (1).

Results

With k known, the clustering algorithm reveals our 4 distinct clusters of students grouped by personality similarity.

Fig. 1. Histogram with mean Trait Values for each cluster.

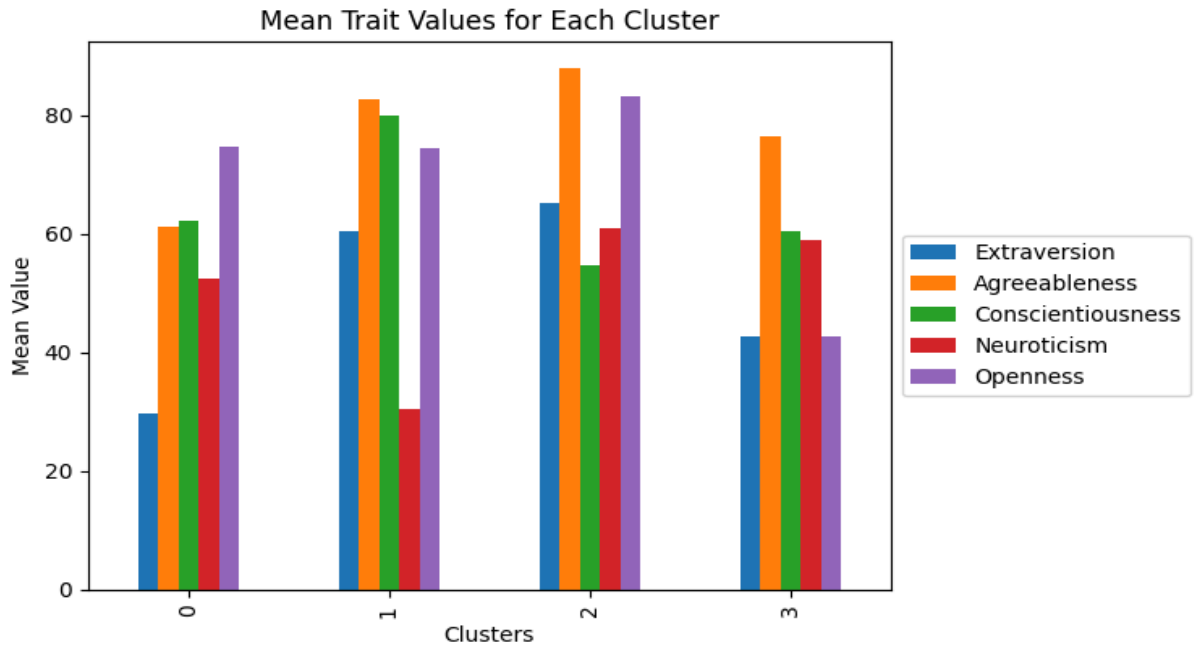


Table G. Personality traits scores for each cluster.

Cluster	Personality traits				
	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
0	59.666668	81.333336	75.250000	30.000000	74.666664
1	34.975960	54.567307	60.336540	50.360577	76.32211
2	57.115383	89.423080	57.692307	63.257576	83.75000
3	38.325470	74.305557	63.657406	59.837963	47.33796

As we can see in Fig. 1, and also evince from Table G, there are relevant differences between each cluster. However, only when we compare step activity and personality clusters by merging the two datasets and performing the correlation, it results in a Pearson Coefficient of 0.03, which is a weak positive relationship between personality and step activity. As can be seen in Table H we finally explored each single personality trait in correlation with the step activity, and the results were similar to the previous, meaning no significant correlation between the two dimensions, therefore indicating that a linear relationship is not evident nor significant in the data.

Table H. Correlation test between personality traits and step activity of the students.

	Personality traits				
	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Pearson's Coefficient	-0.023	-0.029	0.037	-0.089	-0.021

Temporal analysis

The relationship between time and practicing sports is unclear in certain sections, but regular patterns are present when dissecting the data by some attributes. In this section, visualizing and modeling are intertwined.

Model choices and explanation

Visualizing sports and walks is done by either 1) comparing the frequency of these activities against all the others that could be experienced by participants, using **histograms** or 2) by estimating the focused activities' time distribution's kernel density, using **violin plots**. The latter choice is due to the fact that violin plots - that use the benefits of summary statistics of a boxplot on top of a kernel density plot - contribute to highlighting the continuity of the distribution rather than discrete bars of human behavior. The data being visualized in this section won't have more than a total of 2.660 physical activity events out of 110.592 time diary events. Around 55% of physical activities are labeled as "Sport", and the remainder is "Walking".

It is important to mention that in histograms physical activities are compared **against** all the other events experienced by participants, as our model assumes that students possess a finite amount of time and a set of possible activities, each requiring some time available. Participants, during the observation period, likely had to choose between a set of many activities (*doing nothing* included). Furthermore, it's unlikely that participants were forced to practice sport, hence they must have chosen *one activity* rather than anything else at a *specific time*. In this practical framework, following habits or falling into a routine of physical practice is still considered a consequence of a users' total agency regarding their physical activities. Only one activity is not considered in this set of possible choices: it is sleeping, because after it starts its participants cannot make choices until they wake up. Sleeping also takes the most time compared to fulfilling other primary needs, so in order to get an adjusted comparison (time dedicated to sports Vs time dedicated to other choices) we had to exclude sleeping from the total set of activities against which to compare sports and walks. For this reason, the total time spent on any activity of choice will be referred to as "**awake time**"[13], which is approximately

[13] "Awake time" indicates that the total time for chosen activities in one day for one user is not 24 hours, but 24 - 7 and a half hours= **16 and a half hours**. The amount of sleep hours is set according to exploratory actigraphy research (Cellini *et al.,* 2020) in which a young adult student in more than 80 Italian universities sleeps for 7 and

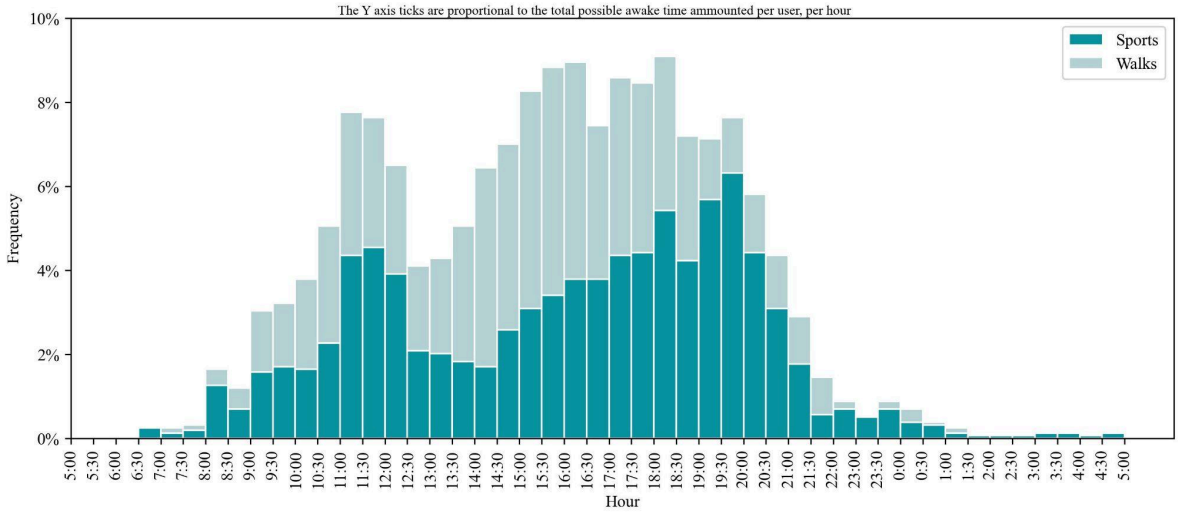
68% of the whole day. We did not adjust it by hour (even though a human is more likely to sleep at 23:00 than 13:00) because of the extreme diversity and lack of data regarding sleep schedules. As a consequence, we must consider that each hour has the same probability of being ≈ 68% "awake time", which is when we assume users made active choices and ≈ 32% of no-choice sleep events. In other words, each bar on the hourly histogram (where frequencies are computed on a sum of physical activities out of all the remaining time available) represents only 68% of the total time they have, because we excluded sleep; therefore, it represents how much time was spent on physical activities out of the total average "awake time", to give the reader a sense of quantity on how much time was spent out of each user's day on a certain activity.

a half hours. The distribution of physical events remains unchanged from this adjustment, but the height of the bars indicates **on average** how much time was spent on a certain activity by participants for 30 minutes. For example, from hours 15:30 to 16:00 of all the 5 weekends present in the data, 15% of the students were walking or doing sports. 15% of 128 students x 5 weekends per hour adjusted for "awake time" (16 and a half hours divided by 24) is equal to 66, which is exactly the number of events that occurred then.

Results

When it comes to people choosing some time out of their schedule to routinely practice sport, the simplest model may assume that every day is not different from any other, and there are some hours that are just more fit for physical activity than others, hence the distribution does not differ, i.e. in a Sunday or in a Tuesday. For simplicity purposes, this is the first approach we chose, as shown in Fig. 2 (Stacked Histogram (all days in one)).

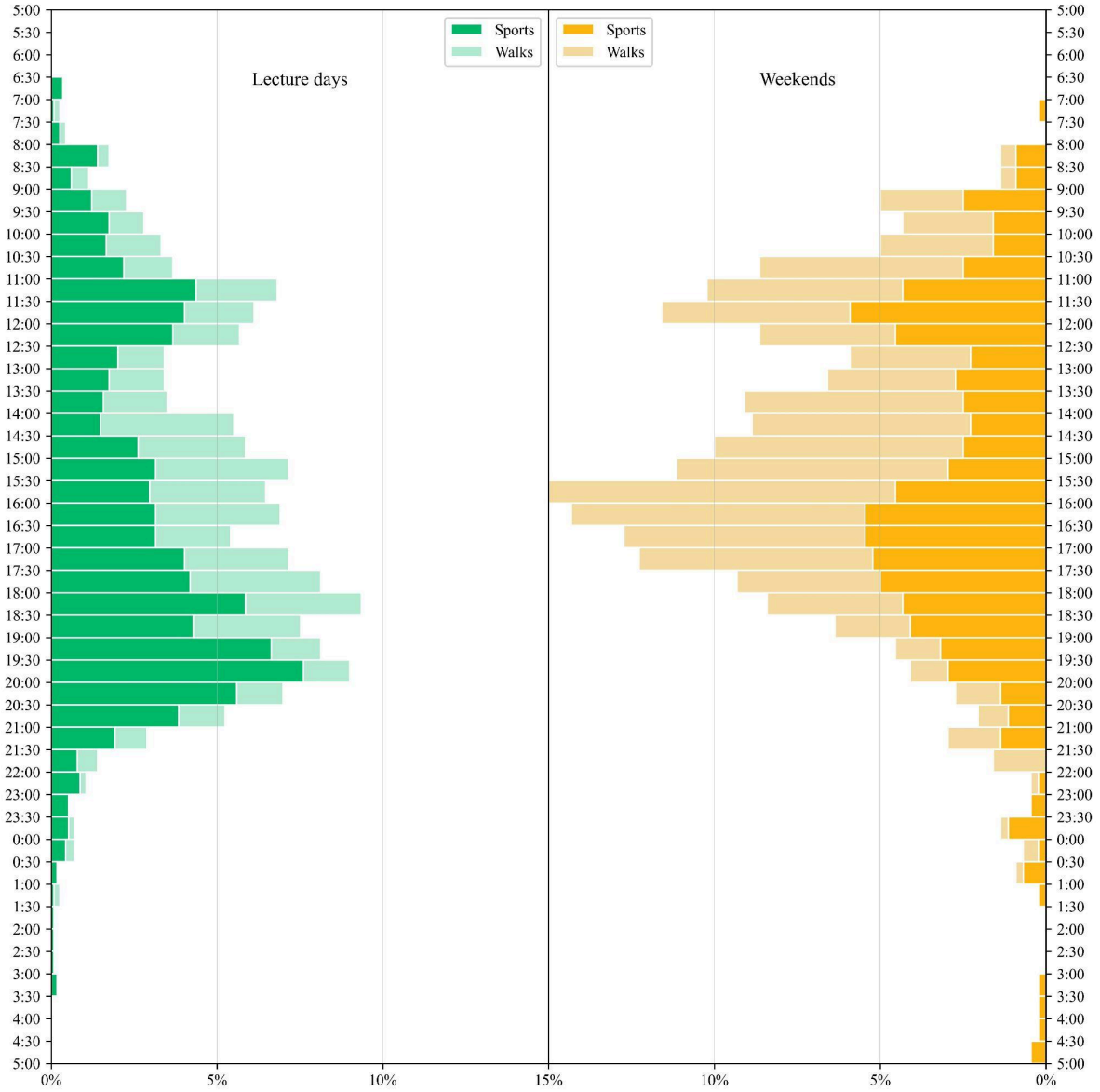
Fig. 2. Stacked histogram (all days in one). 128 users, 2660 sport and walk events (30 minutes).



The afternoon hours (14:30 - 20:00) are the most eventful and peak around 18:00. The morning has only spikes. Walking and practicing other sports seem to mirror each other in their hourly distribution, but there are some differences. *Walking* spikes after lunch time and begins to **decrease**, while *Sports* **steadily increases** until dinner time (20:30 and after). Some clear outliers represent some users "practicing sports" at 02:30 or 03:30.

The "every day is similar to any other" assumption made in the previous visualization may be too rough and may hinder visualization of smaller patterns, so more action was taken. There are many ways to dissect the data using different time units, but each time it is done we consider that it may be a step further in overanalyzing insignificant data spikes that occurred in random days. The hourly distribution between Monday-Friday vs Saturday-Sunday is shown below in Fig. 3 (Stacked histograms, (all days in two)). We did not expect data from Lecture days to be particularly more interpretable than the previous histogram, because students' routines differ greatly from one another depending on their academic year (in Trento younger students are expected to attend earlier lectures), their attitude towards following lectures, and other factors.

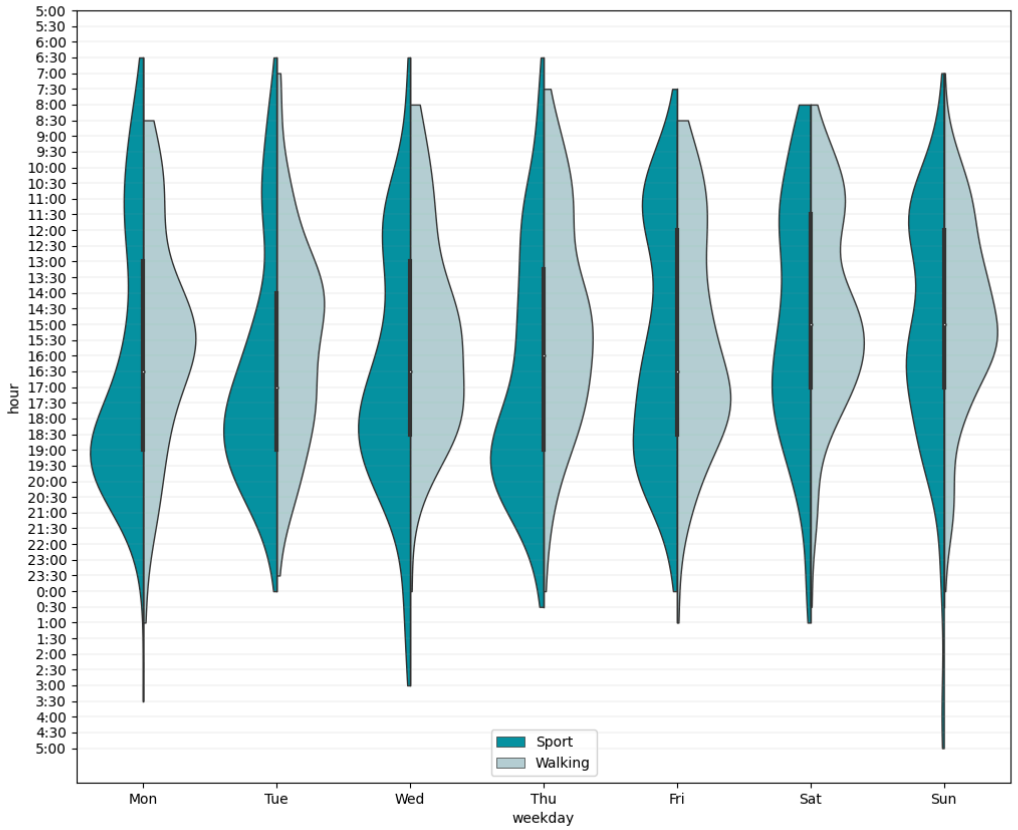
Fig. 3. Stacked histograms, (all days in two). Lecture days: 1.738 events; Weekends: 922.



As it can be noted in the latter figure, weekends are overall dedicated to walking. Indeed, the frequency of **walking** is 39.93% during Lecture days (N=1.738), and becomes 57.81% when we look at Weekends (N=922)[14].

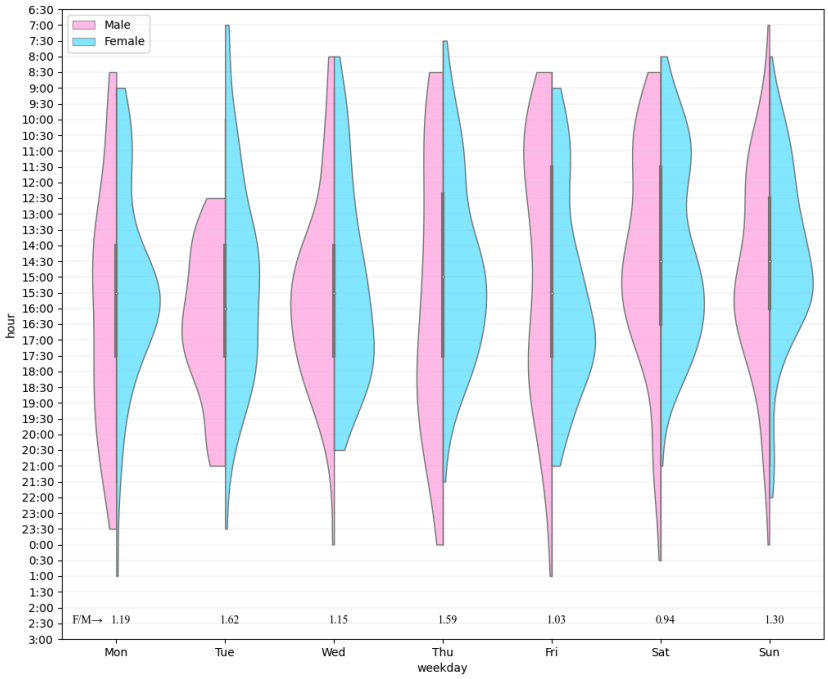
[14] Confidence intervals at level 99% do not overlap (lecture days' walking CI = {37%; 43%} and weekends CI = {54%; 62%}) and are quite distant, so the difference is statistically significant.

Fig. 4. Violin plots (all days in one week) for sports and walks. Average N of events per weekday = 380. *Range {*271*; 536}*



Mornings are generally more spread out, while after dinner there is a steep change: the main dynamics shown in the simplest visualization model do not differ much for most of the week when breaking data down by weekday. Violin plots enhance hourly differences, but one fact that does not appear clearly must be mentioned: Sunday and Saturday are the days in which there are **the most** sport and walk events: 536 and 532, against an average 318 from other weekdays.

Fig 5. Walking and sex. Violin plots of ***walking events*** *(all days in one week), by sex. General F/M ratio:* 1.25

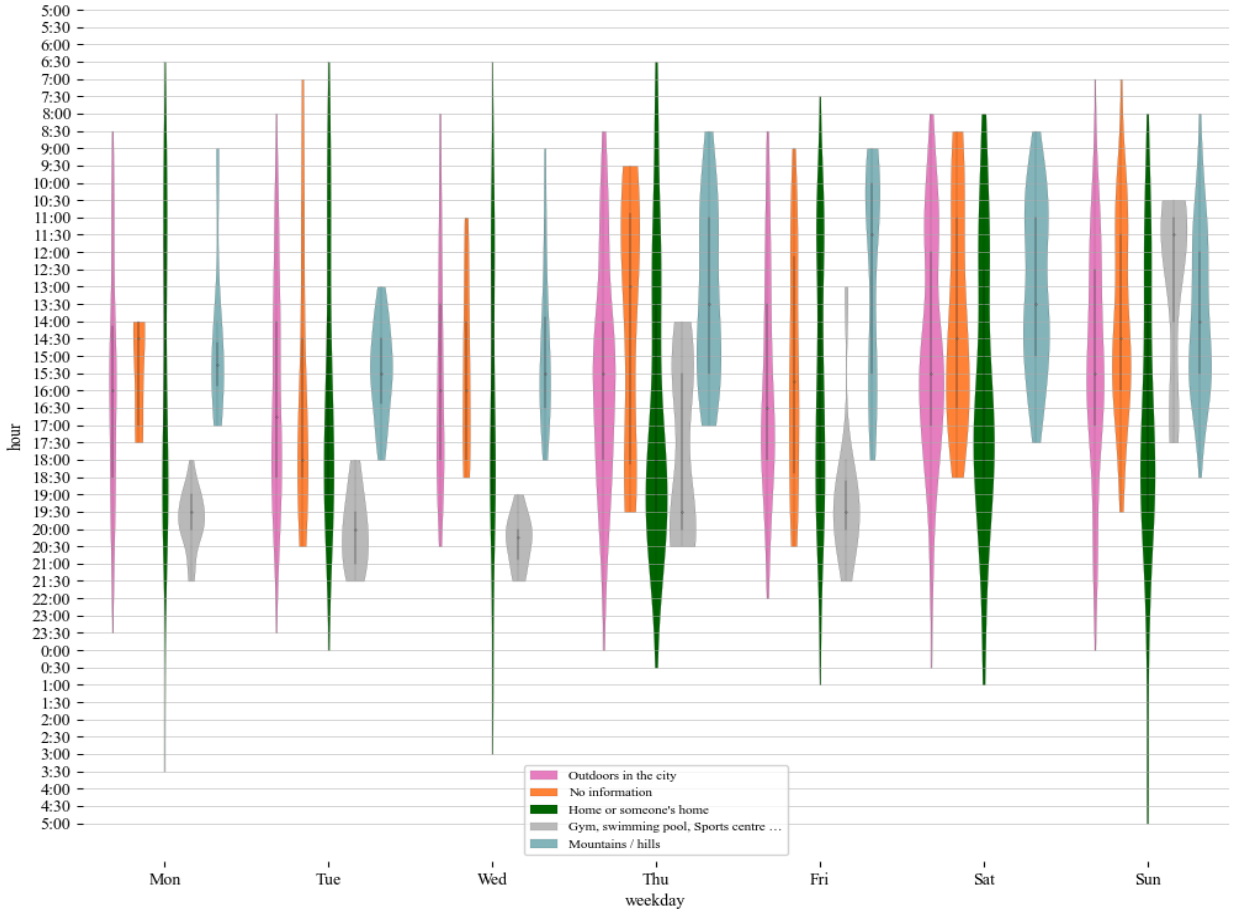


On the bottom of each violin plot the Female / Male ratio can be seen for their respective day. As shown when selecting only walking events, the violin plots make it possible to find tangible differences between users. Female users do not have very different sports hours than men (which is the reason we chose the graph to **not** display both

sports and walks), but do have shorter intervals in which they go on a walk, and are much more concentrated on specific times. This might be due to the fact that female users may experience less security in less crowded (Caiazza, 2005) and darker hours of November nights. Another difference stands solely in the F/M ratio: for 6 out of 7 days of the week, most of the events belong to female users. While the original sample (249 users) had F/M = 1.33, it must not be assumed that a similar disproportion would also turn out to be present in users who actively walk.

From the previous sections we have found several associations between the location of sports and walks. The location can be easily assumed to be a changing variable, based on the hour of the day, other than the activity itself, making it possible for certain windows of time to be more associated with certain locations as shown in Fig. 6 (Violin plots on activity and location).

Fig. 6. Violin plots on activity and location, sorted by day of the week. No information: 7.47% of 2660 events.



One location is expectedly regular over time: **gyms, swimming pools or sports centers**. Even though these commercial facilities are open usually from morning until night, participants find themselves to be practicing fitness indoors later in the day, with an exception for Sunday. Another interesting pattern occurs with outdoors activities: likely because of only 9h 21' hours of average daylight[15] in November, participants never practice physical activities in **mountains or hills** after 17:30 and never

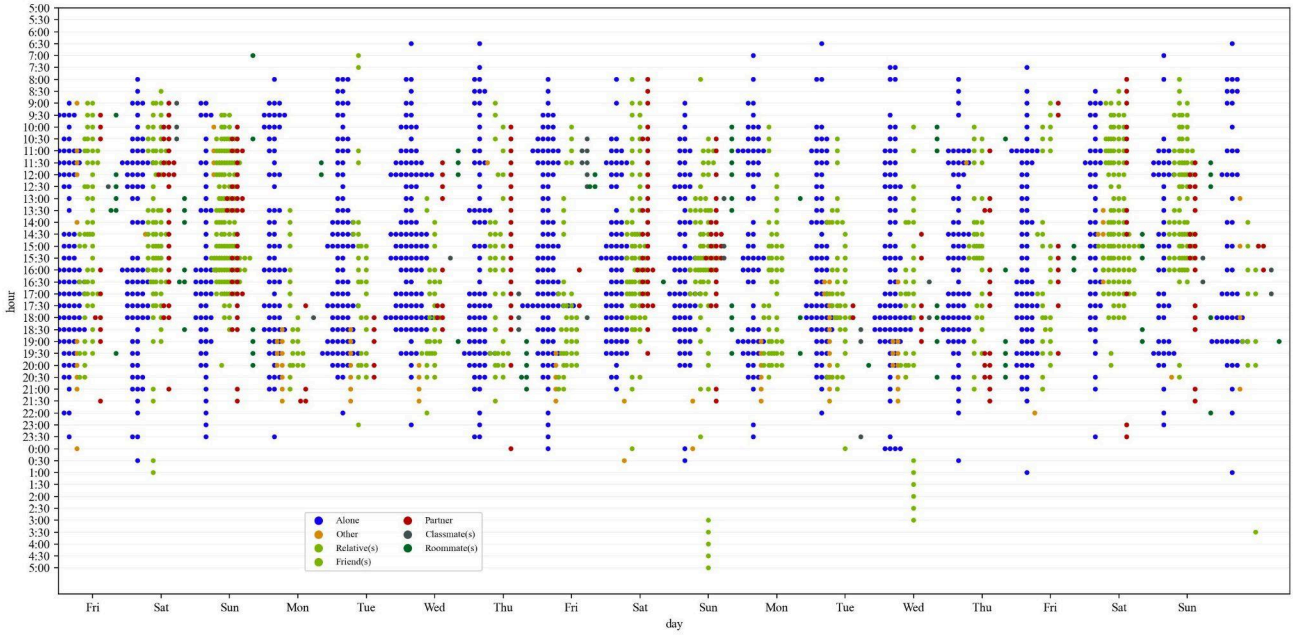
[15] Estimate using neighboring city Bozen (ClimateTemps, 2017)

before 8:00. **Urban outdoor** sports and walks and "**commercial**" **outdoor** sports make it possible for participants to run, jog, cycle, ski and trek for more dissimilar hours, likely thanks to street lighting, ski slopes being active overnight and public transportation. On the other hand, the participants who practice physical activities indoors in a **private location** (such as their home), time of occurrence varies greatly, possibly due to lack of societal and environmental influence (there is no closing time of the gym, or lack of sunlight, and more, although we note this could be detrimental to roommates and family members). We also note an uncanny similarity between "No Information" and "Outdoors in the city", possibly indicating the two being an equivalent answer for participants.

Observing with a more general outlook, it can be concluded that as time ranges from the head of the week to its tail (the weekend), many physical activities take place progressively earlier due to more availability of time or energy, and specific locations contribute to discern at what time the physical activity may take place.

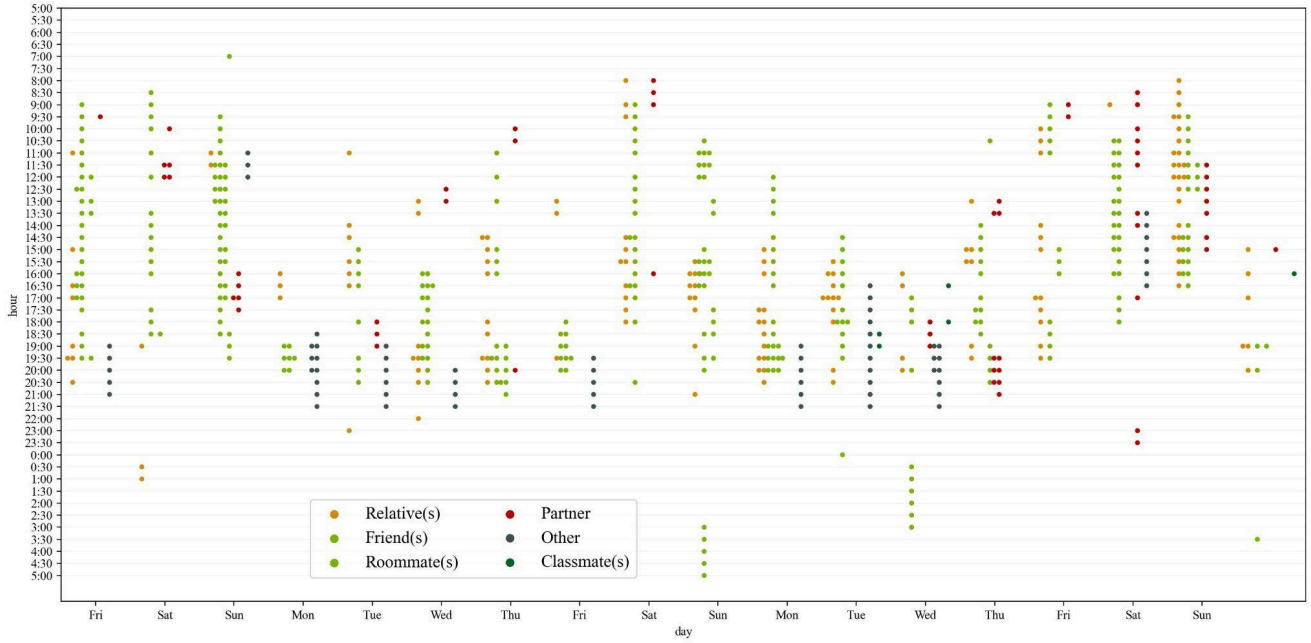
As a final visualization, the company of users may reveal the time in which physical activities take place.

*Fig 7. Swarm plot of the whole distribution of **sports** (hued on company of the user while practicing). N = 1433.*



At first glance, this indicates clearly that the dominant company for those who practice physical activities is … none. More than half of the observations belong to this modality, which in our interpretation is clearly shaped by a multitude of factors. To obtain some more insights, we chose to focus on those observations that occurred not only within the user's own choices, but also those possibly scheduled according to people's agency.

*Fig. 8. Swarm plot of **sports** minus "Alone" (hued on company of the user while practicing physical activity). N = 588.*

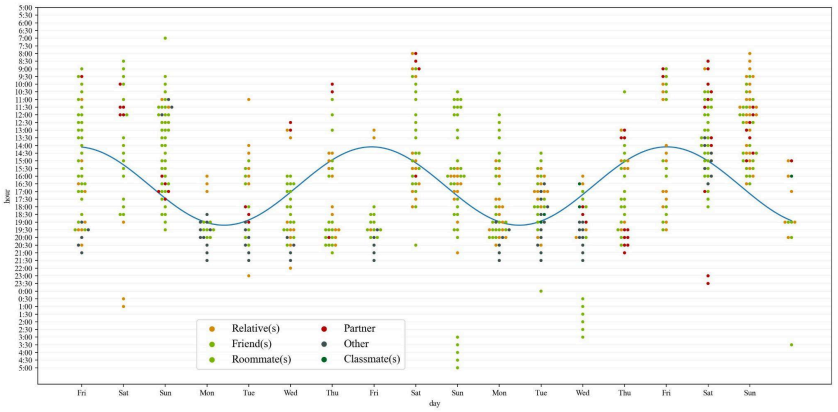


The other modalities of company in Fig. 8 indicate that the time at which the sample's physical activity fluctuates not at random, but through a "social" rhythm, that tries to locate itself in times of certain availability: later in the day during lecture days, and earlier when closer or on weekends. In itself, this distribution could also be easily emulated by a sine wave with some error. The parameters of our fit are indicated in the table below, if the readers want to replicate it. Formula used:

$$y(t) = amp \cdot sin(\omega \cdot t + \phi) + Q$$

Table I. Parameters estimated with sine curve.

Parameter	Estimate	Error (±)
amp / h	2.555	0.040
ω / d	0.889	0.0002
ϕ	4.822	0.111
Q / h	11.644	0.020



We have not researched the causes of this dynamic, but most likely it is due to a general time availability for people earlier in the day only during weekends or the days close to them; it is as cyclic as the week's schedule.

Conclusion

The current report's aim is to provide solid and generalizable insights about students' sport habits, ranging from insightful time patterns to predicting different attributes of human behavior. In order to report this level of information we used several methods that imply errors that are discussed below.
Firstly, the subset of students practicing sport does not represent the original sample, but this is due to the cold period of the experiment (indeed, in November, outdoor activities decrease) and at the time ongoing pandemic, which caused indoor structures and people (also gyms, swimming pools) to adopt more restrictive choices. Both of these major events may have negatively impacted the intensity of sport activities among students.
The sample issue is present also in the step detector dataset, although for different reasons. Here indeed it is related to the fact that the step detector sensor was not activated during the survey and therefore these data were not collected for half of the users, as smartphone devices and user compliance varied substantially.

With respect to the research questions, from the current analysis we derived the following results:

1. Depending on the type of sport activity the set of associated predictors is variable: while some categories require few predictors to produce accurate results, others need more variables to compute the model;
2. There is no simple relationship between psychological attributes of the personality and physical activity mainly because of the nature of the sample, but physically active users can be better classified by the same attributes into 4 clusters;
3. Time and physical activity have complex relationships. Participants spend less time on physical activity during normal weekdays with respect to weekends. Location and sex contribute to discern the hour and day in which physical activity occurs, but more greatly does the type of company present during sports. Gyms, mountains and hills are associated with regularity, while participants practice privately at more seemingly random times.

References

World Health Organization. Newsroom: "*Physical Activity*", 5 October 2022, Physical activity.

WeNet. Pilots: "The Pilots", 2023. Pilots - WeNet

Climate Temps, Italy, Bolzano/ Bozen, Trentino-Alto Adige.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Bibliography

- Caiazza, A. (2005). *Don't Bowl at Night: Gender, Safety, and Civic Participation,* Signs: Journal of Women in Culture and Society, 30(2), 1607–1631. DOI:10.1086/382632
- Cellini N., Menghini L., Mercurio M., Vanzetti V., Bergamo D. & Sarlo M. (2020). *Sleep quality and quantity in Italian University students: an actigraphic study*, Chronobiology International, 37:11, 1538-1551, DOI: 10.1080/07420528.2020.1773494
- Swana, Elsie Fezeka, Wesley Doorsamy, and Pitshou Bokoro. (2022). *Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset*, Sensors 22, no. 9: 3246. DOI: 10.3390/s22093246
- Anne R. Daykin & Peter G. Moffatt (2002) *Analyzing Ordered Responses: A Review of the Ordered Probit Model*, Understanding Statistics, 1:3, 157-166, DOI: 10.1207/S15328031US0103_02
- Bainbridge, T., Ludeke, S., & Smillie, L. (2022). *Evaluating the Big Five as an organizing framework for commonly used psychological trait scales*. Journal of personality and social psychology. DOI: 10.1037/PSPP0000395
- Binte Habib, Adria. (2021). *Elbow Method vs Silhouette Coefficient in Determining the Number of Clusters*. DOI:10.13140/RG.2.2.27982.79688.