Time Spent Analysis

Richard Ge
SID: 916466262
STA160 Section A02
rwge@ucdavis.edu

Jonathan Jiang
SID: 916619792
STA160 Section A01
jsjjiang@ucdavis.edu

Kunteng Miao
SID: 916878494
STA160 Section A01
ktmiao@ucdavis.edu

Daniel Momeni
SID: 918389787
STA160 Section A02
dmomeni@ucdavis.edu

June 6, 2022

Abstract

One of the greatest anxieties of the past decade, especially prevalent among those using smartphones or other devices, is how much time is wasted throughout their day as a result of unnecessary distractions. The purpose of this project is to determine how one subject's time is spent, what underlying patterns exist, and how different hypotheses regarding human behavior may come into play. The main portion of this project will consist of a spectral analysis (i.e. Fourier transform) of the encoded activity data, after splitting time blocks into 5-minute intervals. Further analysis will involve hierarchical clustering and a generalized linear model (GLM) on extracted features, which can shine some light on numerous hypotheses such as ego depletion and the predictive potential of activities performed during the first hour of any given day. Significant predictors in this analysis will be crucial in optimizing the subject's productivity, in addition to confirming or rejecting the applicability of the aforementioned hypotheses.

1 About the Data

1.1 General Description

For ease of entry, the dataset was originally entered as time blocks; a few example days are shown here. A C program (for processing speed) was used to format the data as a .csv and strip away comments, while more cleaning and anonymizing was performed in Python. The time blocks were then shortened to their most relevant concept, the Activity (i.e. details and secondary activities were removed), and the dataset was extended to have multiple 5-minute intervals of each time blocks' activities. The original log contained data from November 11, 2020 to April 30, 2022, for a total of 536 days of data and 14,221 observed time blocks. Extending this dataset to contain standardized-duration time intervals, instead of time blocks that varied in duration, resulted in 96,304 rows, each containing a 5-minute interval and the activity performed in that interval. Sleeping and bedtime routines were omitted, but may be imputed if necessary, as the subject took roughly 45 minutes to an hour to prepare for bed (therefore, the rest of the time would consist of the subject sleeping).

The variables in the original long dataset are as follows:

- `Date`: The day on which the activity was performed.
- `Time`: The start time when the activity was performed.
- `Activity`: What was done at that date and time.

All the possible activities are defined below:

- `break`: anything except the bottom activities; usually slacking off, being unproductive, etc.
- `breakfast`: first meal of the day
- `planning`: preparing things (usually plans) in a way that does require thinking; often further in the future than "prep", but not necessarily
- `lunch`: meal taken around noon or after
- `career`: anything related to career stuff (e.g. applying, interviewing, working, etc.)
- `exercise`: exercising for health; doesn't count exercise to get somewhere (e.g. biking or walking to campus)
- `dinner`: last big meal of the day
- `appt`: scheduled appointments that are necessary; e.g. doctor's appointments or advising appointments; or health-related issues that are unavoidable
- `help`: performing chores
- `prep`: preparing things that are in the near future, which do not require thinking
- `bike`: biking somewhere
- `walk`: walking somewhere (rarely recreational)
- `travel`: getting somewhere, not including any of the other methods
- `bus`: taking the bus somewhere
- `108`/`135`/`141a`, `CLA`/`FYS`/`UDCE`, etc.: classes that the subject took

1.2 Feature Extraction

The differing lengths of each day necessitated the creation of a new dataset, where each row was a single day, and the columns were the extracted metrics/statistics that were of interest to this study's questions. This dataset would be similarly useful in fulfilling both the primary and secondary goals of this analysis, as the tidy data works better in machine learning and linear models.

The variables in this new dataset are defined below:

- `date`: The day that the row pertains to.
- `getup`: The first non-break time of the day (i.e. when the subject gets out of bed).
- `dinner`: Amount of 5-minute time periods spent on dinner (e.g. a value of 7 means the subject ate dinner for 35 minutes total).
- `weekend`: Whether the day is a weekend (1 for weekends, 0 for weekdays).
- `afterdinner_break`: Amount of 5-minute time periods spent on breaks after 7 P.M. (e.g. a value of 14 means the subject spent 70 minutes - of the time after dinner - being unproductive).
- `help`: Amount of 5-minute time periods spent doing chores.
- `wakeup_breaks`: Amount of 5-minute time periods spent non-productively, before getting out of bed.
- `last_productive`: The last time at which the subject was productive for the day (e.g. a value of "22:30:00" means that the subject was last productive at 10:30 PM, and didn't do anything productive after that, only taking breaks).
- `exercise`: Amount of 5-minute time periods spent exercising.
- `walk`: Amount of 5-minute time periods spent walking somewhere.
- `breakfast`: Amount of 5-minute time periods spent eating breakfast.
- `total_class_time`: Amount of 5-minute time periods spent on any class, including lectures, homework, and exams.
- `lunch`: Amount of 5-minute time periods spent eating lunch.
- `bedtime`: The last time of the day (i.e. the time at which the subject started getting ready for bed).
- `prev_bedtime`: The previous day's bedtime.
- `bike`: Amount of 5-minute time periods spent biking somewhere.
- `career`: Amount of 5-minute time periods the subject spent on career-related activities (e.g. applying, interviewing, working, etc.).

- `afterlunch_break`: Amount of 5-minute time periods spent non-productively, between 1 P.M. and 7 P.M.
- `afterlunch_nonbreak`: Amount of 5-minute time periods spent productively or semi-productively, between 1 P.M. and 7 P.M.
- `bedtime_breaks`: Amount of 5-minute time periods spent non-productively, before the subject's bedtime routine but after the last_productive time.
- `prep`: Amount of 5-minute time periods spent preparing for a short-term goal or task.
- `beforelunch_break`: Amount of 5-minute time periods spent non-productively, before 1 P.M.
- `appt`: Amount of 5-minute time periods spent on scheduled appointments (e.g. doctor's appointments, advising appointments) or health-related issues that are unavoidable.
- `beforelunch_nonbreak`: Amount of 5-minute time periods spent productively or semi-productively, before 1 P.M.
- `wakeup`: The earliest time that the subject recorded (i.e. when the subject woke up; recorded to be before the subject sleeps in).
- `break`: Amount of 5-minute time periods spent non-productively, for the entire day.
- `total_productive`: Amount of 5-minute time periods not spent on break, for the entire day. Includes both productive and semi-productive activities.
- `bus`: Amount of 5-minute time periods spent taking the bus somewhere.
- `afterdinner_nonbreak`: Amount of 5-minute time periods spent productively or semi-productively, after 7 P.M.
- `planning`: Amount of 5-minute time periods spent preparing for a long-term goal or task, or a task that requires thinking.
- `actualday`: The day of the week (0 = Monday, . . . , 6 = Sunday).

This shortened dataset will be used for the Generalized Linear Model (GLM) and Machine Learning (ML) methods of interest, while the original (longer) dataset will be formatted and encoded for spectral analysis, through the use of a Fourier transformation.

1.3 Missingness and Encoding

One issue with the original data is that approximately 1/3rd of the 24-hour day is unobserved, due to the subject being asleep. This sleeping period also varies in length, which complicates the numerical encoding necessary for spectral analysis/Fourier transform. For the encoding, we attempted to remove any values in our data that fell between the sleep to wake-up time, since it affected our implementation of the Fourier method. However, the results were still inconclusive since the daily activities have different times that affect the Fourier method; this will be discussed in Section 3.3 of the report.

2 E.D.A.

2.1 Activity Counts

The simplest EDA, and the one we first performed, was to look at the total counts of activities over the entire time period. The table of the top 15 most common activities is shown below.

Figure 1: Activities, Sorted by Commonality

Activity	Num. of Occurrences
break	34571
help	10119
career	9185
dinner	4727
lunch	3382
planning	3350
exercise	2669
131b	2390
bike	1977
171	1857
breakfast	1811
prep	1748
137	1603
141a	1408
141b	1322

This simple analysis shows that "break", meals, and chores account for the majority of the subject's time. The lower numbers for different classes also suggest that all classes will need to be aggregated somehow, because regular activities like eating and chores will always overtake class activities that only occur for up to one quarter.

2.2 Plotting Activity Counts

In order to get a narrower look at how the data looked, we then plotted all of the data points by time of day, to see what activities were most common at each time.

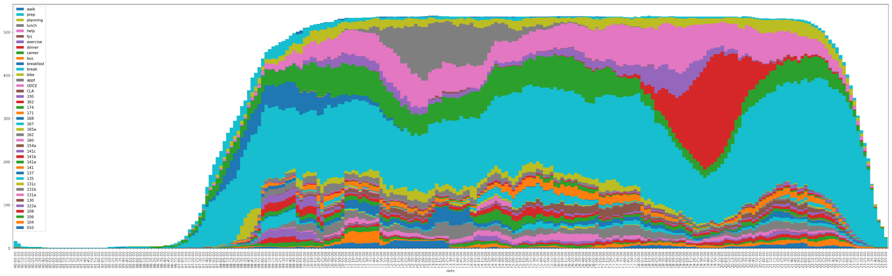

Figure 2: Counted Activities, by Time

It is clear that, usually, the subject is most likely to be taking a break at any given time instead of other activities. The only times where this is not the case is during lunch (the gray area near the top left) and dinner (the red area near the top right). This implies that the subject eats those meals at around the same time every day, but not breakfast (the thin, dark blue area on the left), of which those times are spread out more evenly. Furthermore, a third of the days recorded start at around 7:45 A.M. or earlier, while two-thirds of the days recorded start at around 8:45 A.M. or earlier.

2.3 Plotting Activity Proportions

The varying lengths of the days seen previously necessitate the use of proportions in this EDA, rather than counts only. Figure 3 (below) shows that breaks (teal) are taken slightly less during meal times (dark blue, gray, and red), and significantly more in the evening (and, to some extent, in the morning). This seems to indicate that ego depletion (or some similar phenomenon) occurs in the subject at around 8:30 to 9:00 P.M., becoming vastly more prominent after 10:15 to 11:00 P.M. - the subject performs almost no work after this time, despite often staying up this late.

Figure 4 shows that, throughout the day, only slightly more time is spent on breaks (green) than on classes (red); however, as previously noted, the number of breaks rapidly overtakes the classwork done in the later evening (i.e. around 9:45 P.M.).

The low variation in the data after summing up all values, as well as the varying times at which activities (even including regularly-occurring activities like meals) start and end, both hint that the Fourier Method may not detect lower frequencies due to their corresponding activities being out of sync across different days (e.g. eating dinner at 7:40 P.M. one day and 7:15 P.M. the next day may break up any daily "dinner frequency"). Instead, extracting features

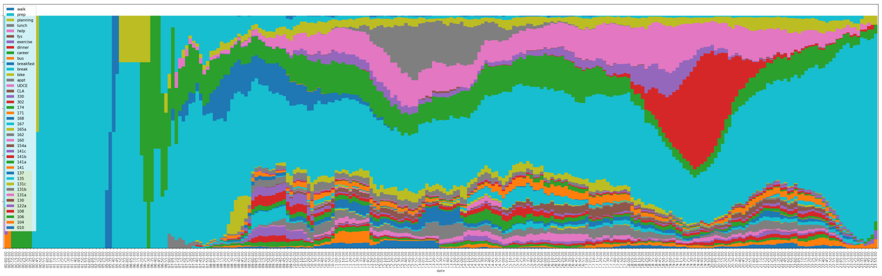

Figure 3: Proportions of Activities by Time, With Separated Classes

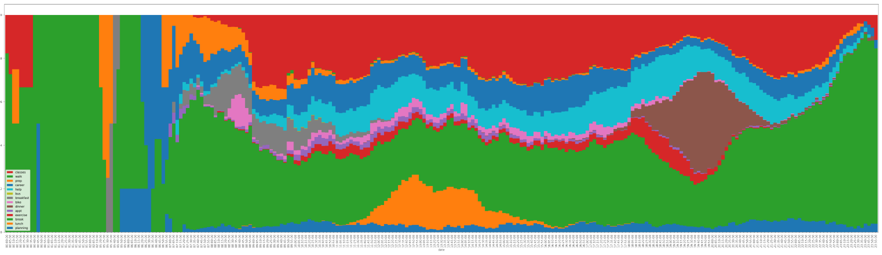

Figure 4: Proportions of Activities by Time, With Agglomerated Classes

for each individual day should associate those varying times with whatever other metrics change as a result (e.g. finding out that dinner at 7:40 P.M. is associated with more productivity before dinner, or less productivity after dinner).

3 Fourier Method

3.1 Fourier Method Goal

The Fourier transformation is especially useful when it comes to time series decomposition. It decomposes functions depending on space or time into functions depending on frequency, essentially simplifying a periodic time series into its constituent sine and cosine waves. Our continuous signal will be represented by a infinite series of sinusoids. This will follow the function:

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k cos 2kt\pi + b_k sin 2kt\pi) \tag{1}$$

The coefficients of a and b subscript k from the fourier method are derived via the following integral:

$$\int_{-\infty}^{\infty} x(t)^{-2iFt\pi} dt \tag{2}$$

For our time series, we have the time variable logged by date (year), hours, and minutes of time spent on each activity. If, for this current section, we assume seasonality, then we can attempt to describe our data as potential waves and frequencies; for example, a daily or biweekly frequency. There are many factors we want to look at in our dataset to first see if there are any trends in seasonality. Then, we want to further simplify what we have to see if there are seasonal trends in terms of workload by minutes as a factor of years and/or months.

3.2 Initial Analysis

What we see is that time spent on most activities has a range of very low time spent on those activities. However, for activities with more time spent on them, there does seem to be an initial trend: A sudden increase in time followed by a general increase in time which is then followed by one or two sudden drops until that activity completely tails off like the other variables. This could be for one of many reasons; for example, the data petering away as collection stopped (if that variable was a university class then it would make sense to have a large amount of time spent only for it to sharply decline). For daily activities there is, of course, a lot of time spent on those activities but we expect that to remain fairly consistent in terms of time spent. Later on in our time series analysis we will specifically look at factors that have high feature importance to have a closer look at seasonality trends and model selection.

3.3 Fast Fourier Plot Analysis and Conclusion

Now we will apply the Fourier transformation to gain further insight on our data. First we must understand the differences between a discrete Fourier transformation and fast Fourier transformation. The discrete version is where the variables are measured as discrete variables and the time series is not a continuous function. The input is a sequence of numbers from the original variable and the output is signal strength for each frequency. The fast Fourier method gives the same output as the discrete method; however, the algorithm is much more computationally efficient. The first plot here is the discrete Fourier transformation, which is fairly hard to interpret, so we applied the faster method. We then had to smooth the graph out for better interpretability, and the result can be seen below. The x-axis unit is defined as 5-minute time periods. It is clear that there is a huge peak in the beginning of the graph, but this can be ignored as it is not relevant to our analysis. What we want to be looking at is the peak at the end of the graph at time 100,000. What this demonstrates to us is that there is a yearly trend in our data. This 100,000 is due to the fact that our data is in 5 minute intervals; after converting the x-axis to represent the time intervals of our data, the maximum frequency present is that of (just over 100,000 intervals) / (12 intervals/hr.) / (24 hrs. a day) = just over 347.2 days, which seems to indicate one year. This means that there is a very strong yearly trend present in the data, as the peak is fairly large. However, we believe this method to be inconclusive for lower frequencies (e.g. monthly, quarterly, or daily) since many factors such as time spent sleeping and other slightly-offset daily activities that cause those frequencies to not be.

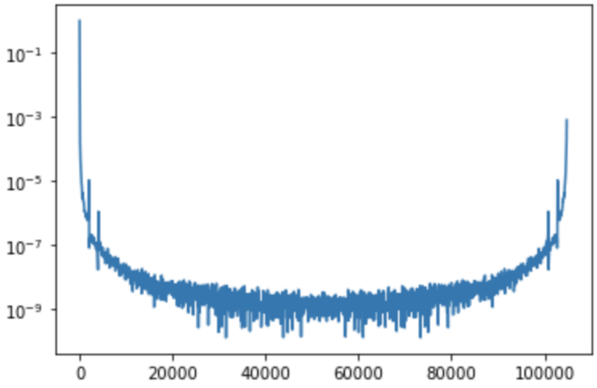

Figure 5: Fast Fourier Transformation

4 Time Series Modeling

4.1 Time Series Goal

Time series analysis helps us to understand the underlying causes of trends or systemic patterns over time. The objective is to consider which factors play a large role in the relationship between time spent on certain classes and time spent on career. Once those factors have been decided we will then do time series analysis to create a final model. We will focus on stationarity as that is an assumption in statistical time series modeling. This will be done by attempting to do different transformations on our data to see which one can create the most stationary outcome for which we could use as our baseline for a forecasting model. We will check for stationarity by looking at moving averages and conducting hypothesis testing. As we saw in the Fourier transformation section there was a yearly seasonal trend so we would like to construct an ARIMA model on our data that could best fit a prediction model; the AR stands for "auto regressive", and the MA stands for "moving average." The I stands for "integrated", which is what allows us to combine the two to create our final model.

4.2 Factor Importance

In order to decide which factor was associated with the highest feature importance in our potential model, we have to do some basic analysis. We did a simple regression model that took into account time spent in classes and the time spent on work to see which classes had the most impact on work and vice versa. What we found was that STA 135, STA 106, and STA 104 had the largest negative impact on time spent on work and that no classes had a very large positive impact on time spent with work. The next step was to plot each class as a time series function to check for stationarity and none of which did at first glance. This was important to do because if any resembled a stationary function then

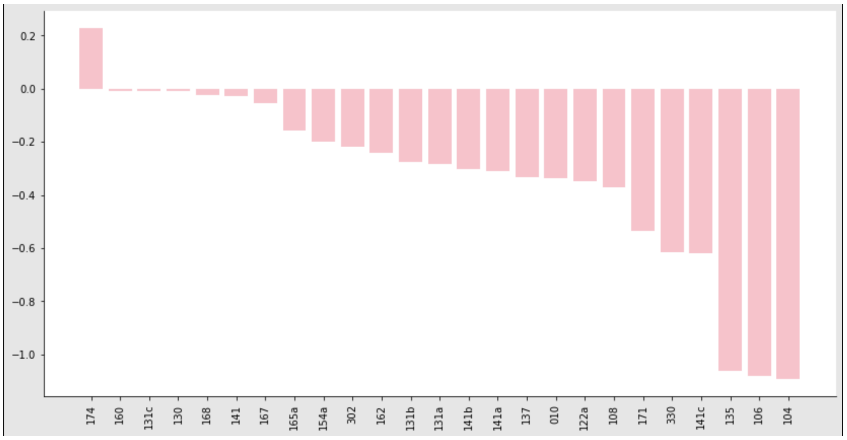

Figure 6: Factor Importance

the analysis for a final model would be much easier to conduct. Since none of them were stationary, we chose STA 135 to do analysis on as we figured the impacts were relatively similar, so our models would be conclusive regardless.

4.3 Model Description and Parameters

The most useful parameters in deciding our final model will be dependence, stationarity, differencing, exponential/log smoothing, and rolling averages. We will follow the standard ARIMA model, which is defined as follows:

AR:

$$Y_t = \alpha + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} + \epsilon \tag{3}$$

MA:

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + ... + \phi_q \epsilon_{t-q} \tag{4}$$

The assumptions of an functioning ARIMA model are that it must be stationary and univariate.

When looking at the AR portion of our model Y of t depends only on its own lags and the betas are the coefficients of the respective lags. Alpha would be the intercept of the model. When looking at the MA portion of our model Y of t depends on the lagged forecast errors where the error terms are the errors of the autoregressive models of the respected lags. An ARIMA model is where the time series is differenced to make it stationary so the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} \epsilon_t + \alpha_1 \epsilon_{t-1} + ... + \alpha_q \epsilon_{t-q} \tag{5}$$

In simple terms the equation is calculating our predicted variable Yt by adding a constant plus a linear combination of our lags plus the linear combination of our lag forecast errors.

4.4 Finding A Stationary Model

The first step was to actually test our time series model for stationarity. This was done by calculating the rolling mean and rolling standard deviation. The model itself was then plotted over our original time series to have a side by side overview. From that graph, it becomes clear that the model is not stationary, as the rolling mean and rolling standard deviation are not constant in relation to our initial time series plot. A hypothesis test was also conducted by using the Dickey-Fuller test with a focus on AIC, as a lower AIC is associated with a better model. The test statistic is -3.14 and our p-value is .02, which means we can reject the null hypothesis and say it is stationary. This result is fairly unexpected because our rolling mean and rolling standard deviation were not constant, so we will try to create more models to find the best model with respects to stationarity. We perform a logarithmic transformation which did not end up meeting our criteria; instead we created a new model consisting of the log transformation minus the rolling average. We then plotted the rolling average and rolling standard deviation, which seemed to be much more constant than our original model and our log transformation model. When we conducted the Dickey-Fuller test, we had a p-value of 7.9e-07, with which we can reject the null hypothesis at any reasonable value of α; this strongly indicates a stationary model. Two additional models were created (specifically, a decaying logarithmic difference shifting model and a logarithmic-minus-exponential-decay model), but these both failed the Dickey-Fuller test as well. As a result, our final model was the logarithmic transformation minus the rolling average.

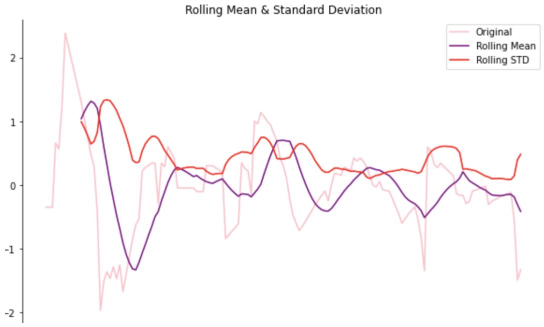

Figure 7: Log Minus Moving Average Model

4.5 Final ARIMA Model

Now that our stationary model is determined, we can form our ARIMA model. We first plotted the ACF and PACF graphs to get our respective p and q values for the model (the p pertains to the AR and the q pertains to the MA). The PACF gives us our value for p, and the ACF graphs gives us our value for q. We get these values from looking at the value on the x axis, when the graph reaches zero on the y axis. The values on the x axis give us our p and q values. From the graph we see we have a value of 2 for p and 7 for q. For differencing, we will use the value 1 for integration (i) as the overall trend in time spent studying appears to decrease linearly over time. As a result, it will be important to include integration because there is a trend in the data, which needs to be accounted for to create a stationary model. Therefore, our final model is ARIMA(p = 2,i = 1,q = 7). When we plot the ARIMA model, the RSS is quite small, which means that this model fits our data well. We can conclude from our Fourier Transformation and ARIMA model that this data contains a yearly trend that can be stationary when accounting for AR, MA, and integration factors accordingly. This specific model took into account STA 135, and the time spent in that class, as it appeared to be an important factor in predicting overall workload.

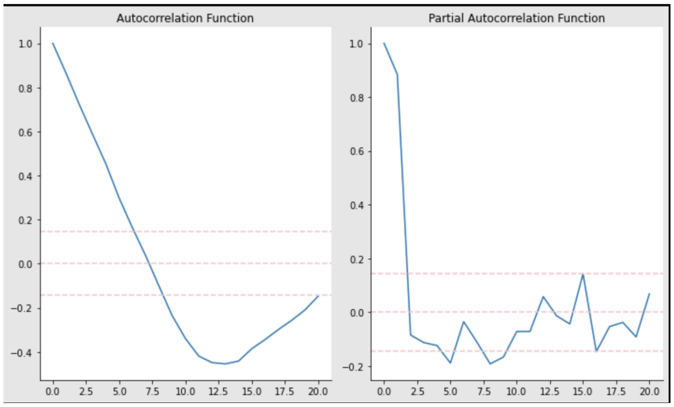

Figure 8: ACF and PACF Graph

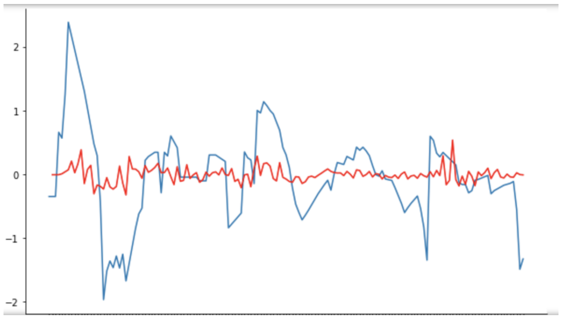

Figure 9: RSS

5 GLM

5.1 Model Selection

In order to find out how time spent is related to the subject's productivity, we should start with some model selection. First, we have to find some variables from a set of candidates. Once any set of candidate variables has been chosen, the statistical analysis allows us to select the best of these models, by the Principle of Parsimony (given candidate models of similar predictive or explanatory power, the simplest model is most likely to be the best choice). Model selection techniques are used for several reasons; these include shortened training times, improved interpretability by researchers, and better compatibility with a learning model class.

5.2 Method

(a) R^2 and R^2_{adj} Larger values indicate larger portion of variability explained:

$$R^2 = 1 - \frac{SSE}{SSTO} R^2_{adj} = 1 - \frac{MSE}{MSTO} \tag{6}$$

(b) AIC: the preferred model is the one with the smallest AIC value

$$AIC = -2log(\hat{L}) + 2k \tag{7}$$

(c) BIC: Models with lower BIC are always preferred

$$BIC = -2log(\hat{L}) + klog(n) \tag{8}$$

(d) Mallows's C_P: The general consensus is that smaller Mallows's C_P values are better as they indicate smaller amounts of unexplained error.

$$C_P = \frac{SSE_p}{MSE(X_1, ..., X_{p-1})} - (n - 2p) \tag{9}$$

5.3 Result

R^2 works poorly since it keeps increasing as more predictors are included, and even insignificant predictors can reduce SSE. Therefore R^2_{adj} will work better than R^2. The best subset of data by using R^2_{adj} consists of afterdinner_nonbreak, afterlunch_nonbreak and beforelunch_nonbreak. After applying Mallows's C_P and AIC, the model now consists of all variables but breakfast, lunch, bus, walk and any of the variables involving breaks; the variables mentioned here have no association with productivity. The model indicates that, for the subject, more time spent on breakfast, lunch, bus, walk and any of the variables involving breaks is not associated with productivity. Therefore, all other unmentioned activities are associated with increased productivity. After using bidirectional stepwise selection, the dropped variables are planning, all variables involving breaks, weekend, bus, and prep; this means they are not associated with the subject's productivity.

6 Unsupervised Learning

6.1 General Purpose

We attempted two Unsupervised Learning techniques: Hierarchical Clustering with heat mapping and Agglomerative Hierarchical Clustering to find hidden patterns. We will ignore pairing variables with high correlation such as afterlunch break with other breaks such as beforelunch break, and actualday with weekend because those are known, not hidden patterns. Furthermore, preliminary clustering on these highly-correlated variables were inconclusive, possibly due to how similarly these variables were created (e.g. what day of the week vs. whether a day is a weekend). We will look for correlations between possible variables with a correlation heat map and conduct unsupervised learning.

6.2 Exploratory Data Analysis

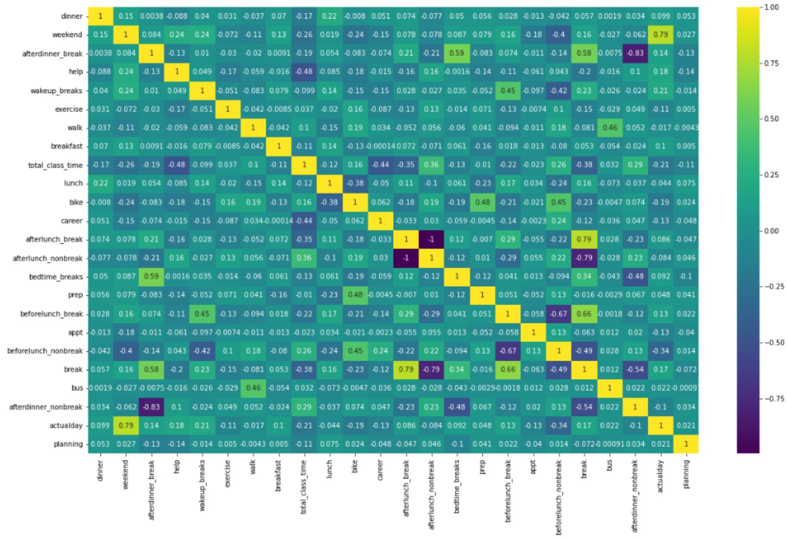

Figure 10: Correlation Heat map

We want to look at potential missed patterns between the predictor variables. Above, we constructed a correlation heat map to obtain potential patterns before using cluster analysis.

6.3 Cluster Analysis Techniques

Overall, we will be using clustering techniques to find out how time/daily activities are related to productivity. Using the correlation heat map and the information from the GLM, we can identify potential patterns by observing correlations greater than, or equal to, 0.45. However, we will also consider correlations that have potential patterns but don't have high values; one such example is the pairing between break and actualday, which can be used to determine if breaks taken could be related to the day of the week. Using hierarchical cluster analysis, we construct a dendrogram with Agglomerative clustering on the variables Break + Afterlunch_break (0.79), Bus + Walk(0.46), Bike + Beforelunch_nonbreak(0.45), Prep + Bike (0.48), and Break + Actualday (0.17). Normally, linkage is a primary requisite of hierarchical cluster analysis as it measures the distance between the two clusters invarious ways. Due to the data set, the linkage will stay the same between each cluster analysis.

6.4 HCA results

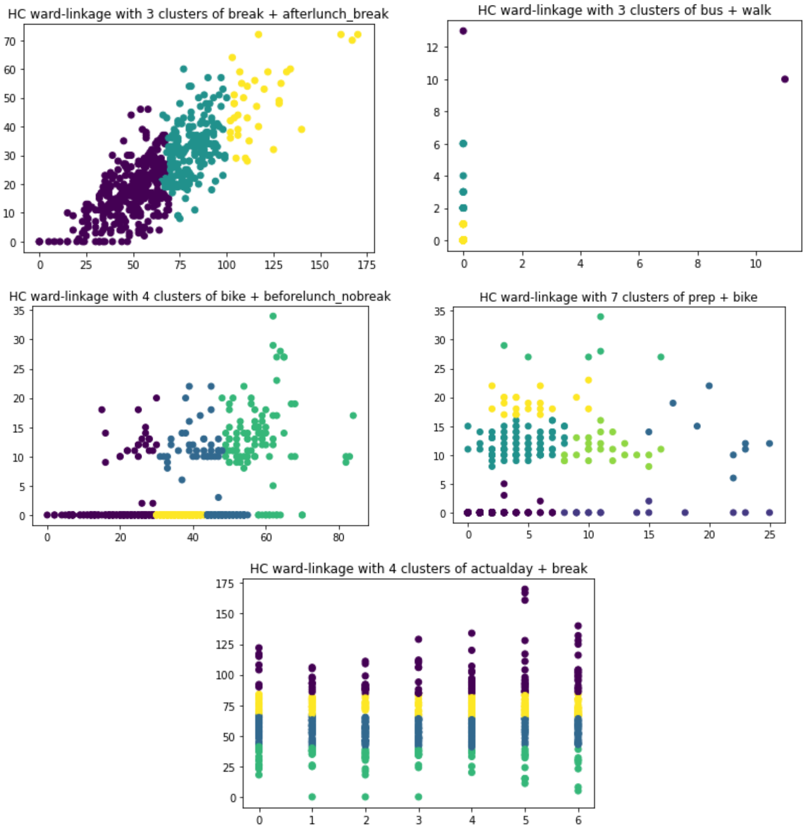

Figure 11: HCA plots

With Agglomerative Hierarchical Clustering, we observe that there are no new patterns learned from Bus + Walk(0.46), Prep + Bike (0.48), or Break + Actualday (0.17). However, Break + Afterlunch_break shows a steady positive linear trend, which tells us that a majority of breaks each day are taken in the afternoon. This is understandable as breaks are not commonly taken in the morning, and instead more likely to be taken after performing "non-break" tasks. Bike + Beforelunch_nonbreak show a small positive trend. Therefore, as time spent on one increases, so does the other; the majority of the subject's Beforelunch_nobreak time is spent on biking, and not much productivity (of any other kind) occurs in that time. It may be harder to see the correlation as there are many 0s in the bike variable; however, the number of 0s from bike slightly decrease as Beforelunch_nobreak increases. This is understandable as there are not many activities that a university student (in this case, the subject) would regularly accomplish before lunchtime, other than getting to campus. Reexamining the Break + Actualday plot, the hypothesis that weekends result in more breaks may be incorrect. Each day seems to have the same amount of breaks, with only a very weak trend and some outliers. It makes sense that these outliers would occur on weekends, as that would be when students celebrate holidays or take day-long breaks from being productive at all.

References

- EDA Bar Chart: https://towardsdatascience.com/stacked-bar-charts-with-pythons-matplotlib-f4020e4eb4a7
- Fourier Transform: https://towardsdatascience.com/fourier-transform-for-time-series-292eb887b101
- "First Hour" Theory: https://www.youtube.com/watch?v=jBwM-mCLQQo
- Ego Depletion: https://www.sciencedirect.com/topics/psychology/ego-depletion