CTU Journal of Innovation and Sustainable Development
ISSN 2615-9422 | e-ISSN 2815-5602

DOI:10.22144/ctujoisd.2023.034

Factorizing social advanced aspects in modern life on human health

Phan Yen Ngan Nguyen[1*], Dung Hai Dinh[2], and Ngoc Hong Tran[1]

[1]*Computer Science Program, Vietnamese-German University, Viet Nam*

[2]*Business Information System Program, Vietnamese-German University, Viet Nam*

**Corresponding author (yenngan2401@gmail.com)*

Article info.

Received 30 Jul 2023
Revised 11 Sep 2023
Accepted 23 Sep 2023

Keywords

Data analysis, healthcare, life style, R language

ABSTRACT

The emergence of technology has brought about a dramatic shift in different aspects of life, especially the aspect of societal advancement. As people adapt to such transformation, it is undoubted that their lifestyles adapt as well. However, whether such changes oppose the well-being of an individual and whether there is a relationship between daily habits and health condition is an interesting topic that this research is going to focus on. Meanwhile, the benefits of implementing data analysis have been proven regarding understanding statistical problems. On such account, this work is going to use the powerful method of data analysis into investigating the relationship between lifestyle and health conditions, with the practice of big data sets deploying R programming language. The procedure from gathering to interpreting data is going to be introduced, and the result is then placed in comparison with related existing studies. By doing so, it has shown that the result has reflected accurately the topics and thus, given evidence for the potential of analyzing the connection between habitual practice and health status.

1. INTRODUCTION

1.1. Habit and Health in Modern Society

Routine is described as a course of actions adopted by a person in a repeated manner over a long period, such as daily or weekly (Arklinghaus et al., 2019). It has been suggested that the formation of a habit takes grounds from the close association between an action within the circumstances (de Vries et al., 2014). Different people, coming from various backgrounds, could adopt a distinct set of actions depending on their occupations, schedules, family status and other aspects. Meanwhile, health has always proven its importance and reflection of human life quality throughout history. Hence, it is obvious to view the strong influence of lifestyle on health status. Statistically, the percentage of such influence could be approximately 60 percent, as reported by WHO, and thus, becoming one of the most correlated factors contributing to individual health (Ziglio et al., 2004). Not only does an unhealthy lifestyle have an impact on physical health but also mental well-being. There are numerous investigations with extensive attempts to comprehend such relationships and provide statistical proof. Regardless of the difference in methods and data, the association between the two factors is unquestionably determined (Hautekiet et al., 2022).

1.2. Relationship Between Habit and Health

Habitual routines are commonly regarded to be one of the most correlated factors to well-being. As health is an indispensable element of a human being, the study of such an association would reveal how people can strive for better life quality by having control over what they normally do every day. With the rise of data analysis in recent years, there are

more robust analytical tools introduced (Stevens, 2023) and thus, creating more opportunities in interpreting health and lifestyle (The Healthcare Insight, 2020). One direction that is widely adopted by researchers is to use an existing database, typically a national survey, in combination with other techniques to study a similar subject (Deborah et al., 2013). To serve the purpose of the topic, the methods of correlation coefficient and regression are employed to a great extent. These two methods pair exceedingly well with the R programming language for data calculation and visualization. Along with that, the theoretical perspectives behind them are not only relatively appropriate to comprehend, but also acknowledge the use of statistics in solving social topics. By that means, the preference of the research method for this paper is chosen.

1.3. Research progress on modern life style and health

Data analytics is rather an immense definition to perceive as it includes various techniques applied to make conclusions from a set of raw data. It has the ability to disclose the trends and behaviors of data in such a manner that is understandable to humans. One outstanding use of data analytics in the business field is the "Six Sigma" process, developed in the 1980s, which aims to aid business people in quality control, eliminating unwanted defects and thus, enhancing the process (Hayes, 2022). One noticeable case study for applying Six Sigma to business analysis is Microsoft, which then develop its own platform known as Windows CE OS (Hayes, 2022). The product has become one of Microsoft's most successful innovations, confirming the prominence of data analytics in the world of finance and business. Henceforth, data analytics has immense advantages to today's society, emerging to become one of the most impactful technological successes in recent years.

The determined matter of contention within the study is to reveal the interdependence between daily habits and health status. The paper is primarily centered on the frequency of a routine with minimum mention of the reasons for any of the choices or different categories of a habit. Not included within the research scope is the prediction of data, comparison with data in the past, or making efforts to justify the motives behind it. Also, data with insignificant results and containing little values are going to be ruled out of the paper.

The paper is structured as follows. In Section 2, the data preprocessing method presents the process adopted, aiming to collect sufficient data for the research. Henceforth, such resources are then cleaned and reorganized. The process of how the data insight is analyzed and retrieved is described in Section 3. Eventually, Section 4 concludes the work and discusses future works.

2. DATA PREPROCESSING METHOD

2.1. Data Scraping Process

The prioritized activity at the beginning of the research was collecting an adequate amount of data to construct a good dataset. Predetermined requirements that the survey must comply with and follow strictly are listed, which ensure that the collected dataset serves the research (Figure 1).

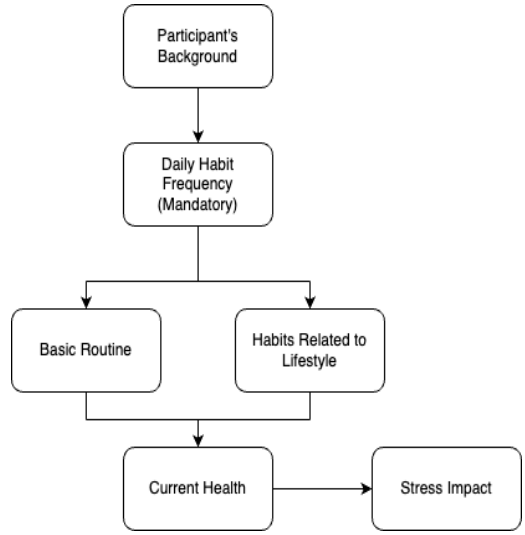

Figure 1. Data collection process

In this process, first, a set of questions to have a better grasp of the participants' background, including age, occupation, and gender, is deemed as important. The ages are divided into 18-25 (represents students and young workers), 26-40 (represents adults), 40-60 (represents adults in more stable conditions), and above 60 (retirement age). Then, questions about the frequency of each basic routine, which are mandatory to the participants, include the act of eating habit, sleep cycle, and physical activities. The answers are in the form of "Never - Always" or frequency during a week or month, depending on each question. To extend participants are asked about habits that have appeared in recent modern society, namely: electronic devices and addictive substances abuse

(smoking/drinking alcohol). There are also optional parts where the reasons, preferences, and extended aspects of several attributes are introduced to the participants. These give a better understanding of one's decision and motivation. After that, participants are presented with questions about their awareness of current health, both physical and how they feel internally, and whether they are going through any type of sickness. They are then followed by the question to investigate the level of consciousness one has upon his own condition. Last but not least, the questionnaire ends with an aspect of mental health - stress factor - which is believed to have a tremendous role in determining the well-being of a human body.

Overall, the survey successfully attained 114 replies from the participants, of which:

– 74.8 percent are from 18 to 24 years old and 20.9 percent are in the 26-40 age group. No data is collected for the age group above 60.

– 34.8 percent are females and 64.3 percent are males.

– 46.1 percent are doing office work, 41.7 percent are students. The remaining are unemployed (4.4 percent), freelancer (2.6 percent), and others (5.2 percent).

The dataset achieves 30 attributes (columns) and 114 objects (rows). However, not every attribute has the same amount of objects. For instance, the smoking frequency only receives 44 responses. The content of the survey is broken down into three segments:

– General questions: through which the overview background is gathered. This includes one's age group, gender, and occupation.

– Daily habits: the questions conduct an inquiry into various aspects in terms of daily activities that people adopt.

– Health status: While the second part of the questionnaire focuses on the influences, the third part shifts its concern to the consequences - the current health status of the participants. In this section, they are asked about any of the discomforts that occurred related to how the body functions. It is the prolonged physical pains or any complications related to the internal systems. The survey then explores the level of awareness that participants have towards their own problems and whether they recognize such issues would bring discomfort to their daily activities.

2.2. Essential Information Refining

Initially, the questionnaire contains various questions to retrieve as diverse a dataset as possible. As the topic only aims to inspect the connection that occurs between habits and health, optional questions that dive into the reasons or how an individual prefers one action over another are relatively not related. To settle upon which variable must be disposed of, any column that is not used during the data visualization, or does not participate in forming a cause-result relationship, is going to be cleaned from the dataset.

– Firstly, it is the variable "Where are you currently living in?" that is being eradicated. The reason behind such action is because this question was inserted at a much later stage of the survey.

– The variable of the question "How long have you been experiencing these health issues?" should be taken out as it has no valuable contribution to the further analysis.

– Another variable to be removed is the data related to how people cope with the discomfort brought by physical pain. This is also extra data that aids to understand the participants' experiences better and has no value to the actual research.

– Similar to the above variable, the data asking how people react to stress has no significance worth to be visualized in the graph.

In the survey, there are multiple questions that do not require the participants to specify any answer. Henceforth, in the retrieved collection of answers, there might include N/A values, which are considered to be difficult in the later stage of data visualization. First and foremost, the checking of missing values for numerical data is inquired for. The first variable to be inspected is the Smoking variable.

The result has shown that, fortunately, no N/A value occurs in the dataset. Considering the nonnumerical values, the approach is much easier to practice. When importing data into R studio, there is an option available to automatically fill in the missing data with the character "NA".

3. ANALYTICAL METHODOLOGY

3.1. Data Visualization

This section concentrates on the explanation of the data visualization attempts with implementing R. The research has earlier deliberated on exploring the relationship of stimulation and outcome between habits and health and thus, the correlation coefficient method is deemed to be the most reasonable in this situation. It is the ability of coefficient correlation to study the strength of the linear relationship that might occur (Nickolas, 2021). The theoretical perspectives are seen first, which then follows the establishment of such mathematical theories into R.

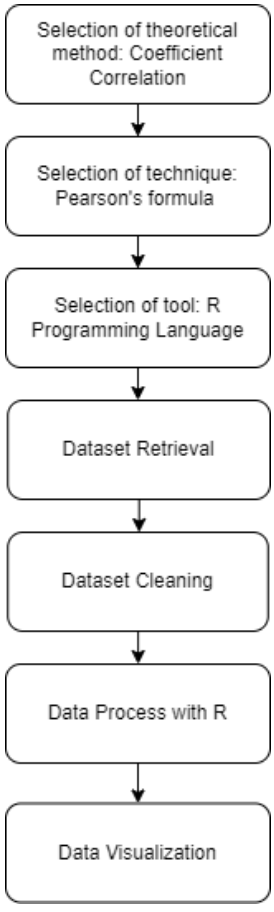

Figure 2. Analytical Method Flowchart

The correlation method is chosen for its ability to present the linear relationship between two variables, which studies the strength bonded between them. The method illustrates such a relationship with the value from -1 to 1, which can be understood that:

– The inverse association is going to have negative values (-1 to 0). One variable escalates would cause the reduction of the other, and vice versa. If -1 is achieved, it is the perfect negative correlation.

– The number 0 to 1 illustrates a direct linkage, which could either rise or decline together. The perfect positive correlation, in this case, would happen when value 1 appears.

– Otherwise, no connection between two variables is equivalent to value 0 in calculation.

Amongst the variety of techniques, Pearson's formula is selected for its ability to work with raw datasets and its suitability when implementing R.

Compared to Pearson, there are formulas by Spearman and Kendall, which worked with ranked variables to discover the monotonic relationship (Zinda, 2021). Considering the scope and ability of the retrieved dataset, linear relationship would be deemed as more suitable to investigate. With R, multiple functions are available to support the plotting of the heatmap with a basic R function called heatmap included already in the installation of R. Another strategy is by employing the package "plotly". Regardless, compared to the above solutions, the function "ggplot" supports the best with more options to define the matrix and modify the heatmap. Since the dataset is in the format of the data frame, it must be transformed into the matrix. The function "sapply()" is put into use for this case. Then, it is calculated with function "cor()" and Pearson's formula is denoted in the command. The data proceedsto be reshaped with the utilization of the function "melt()". In the command, the part "na.rm=TRUE" is showed so that the reforming would demolish any Not Available (NA) values. At this stage, the dataset is qualified for "ggplot()" to do its work.

In order to increase the quality of the research, more attention should be paid to how smaller groups of variables have a connection with each other. Pearson's formula, described above, is effective in the computation of the pair-wise plot.

The calculation of the p-value depicts the level regarding the significance, indicated for each correlation coefficient in the plot.

3.2. Analytical Outcome

From the previously obtained dataset, the process of analyzing is currently in readiness to emerge. In this section, the following data analysis would be expected:

– Feeling of Tiredness connection with all-cause variables, depicted by a heatmap.

– Internal Dysfunctional connection with all-cause variables, depicted by a heatmap.

– Physical Pain associated with other variables, depicted by a pairwise plot.

Feeling of Tiredness: The participants were questioned about how frequently they experience fatigue, which could have a relation with their lifestyle. The heatmap has introduced an insight into the correlation matrix of how the participants' habits have potential linkage to their habits. There were 114 responses for this variable, which is showed through the "Yes"/ "No" question. The data is then compared with the frequency of each activity that occurs in the daily life of a participant. To ensure such a comparison can be made, each attribute is required to have the same amount of responses, which is 114. The frequency is varied between "None" to "Always". The most positively correlated is illustrated with the color red, while blue represents the negative correlation and white being no relation existed. For a better understanding of the visualization, the values are segregated as follows:

– $[-0.3, 0.3]$

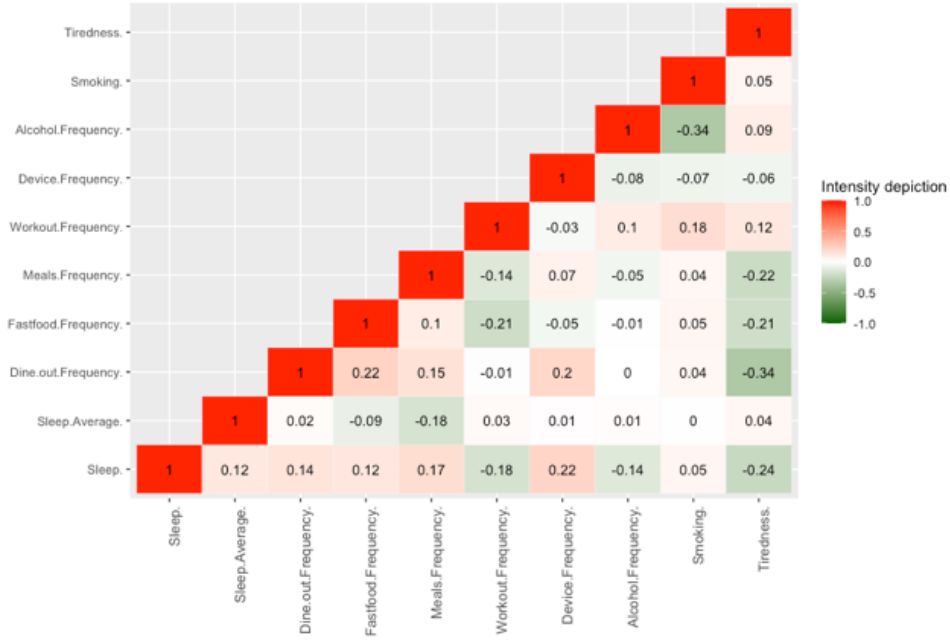

Figure 3. Heatmap for the relation between the feeling of tiredness and other elements

Values have the range from 0 to -0.3/+0.3 indicating the level of correlation as none too weakly correlated.

– $[-0.65, -0.3) \cup (0.3, 0.65]$

Values from above -0.3/+0.3 to -0.65/+0.65 consider a moderate degree of correlation.

– $[-1, -0.65) \cup (0.65, 1]$

Values from above -0.65/+0.65 to 1 portray from strong to highest correlation.

From observation, the results show that both addictive substances are directly associated with the feeling of exhaustion. Even though the data only denotes slight relation, it can be denied that the extreme usage of cigarettes could cause weariness. The harmful impression of cigarettes has been widely promoted for an extensive amount of time. It is recorded that an average number of over 8 million deaths are caused every year by tobacco, in which the direct use of such is responsible for about 7 million (World Health Organization, 2022). The active ingredient in cigarettes, nicotine, has a mental-relief influence on human brains. Upon that, it is believed to temporarily ease anxiety, increase concentration, and stimulate the production of dopamine. However, these effects are immediate and soon the withdrawal would lure the smokers into smoking more in order to find such influence again (Mental Health Foundation, 2021). Similar to smoking, alcohol abuse is another major role contribution to the downfall of health status. There have been numerous studies dedicated to showing

how such dependencies are becoming a destructive factor to human health.

In the meantime, attributes regarding nutrients are shown to have an inverse relationship with the participants' health. Specifically, the lower the meal frequency is, the higher chance that the attribute of tiredness increases. An easily neglected contribution to a healthy body is the nutrient values, which are made up through eating habits. Either over or under-nutrients would cause the lack of energy, or even choosing the wrong food groups. Sleep is another factor that plays a role of utmost importance to well-being, not only short but in the long term. When the body is put to sleep, the process of energy conservation and restoration comes about. According to the US Department of Health and Human Services, the benefits of sleep are recognized as maintaining healthy body weight, increasing concentration, improving mental states, and reducing health issues by letting the body have time for recovery (U.S. Department of Health and Human Services, 2021). There is a surprising relationship between the average amount of sleep and the nutrient habits of an individual. Henceforth, it is revealed that the lesser the sleep average is, the higher both the data for meals and fastfood frequency is. This has scientifically explained that sleep deprivation has resulted in a change of appetite and, thus, weight management (Goldman, 2022). Multiple researches discovered the reason was the lack of sleep would provoke hormones impacting the hunger and fullness cues, which developed into stronger cravings for calorie-dense and fattened food choices. Therefore, people with insufficient sleep have the tendency to fall into obesity, in both children and adults with 89 percent and 55 percent of risk, respectively. In the long run, the deprivation of sleep is associated with cardiovascular issues, promoting higher cholesterol, hypertension, and body mass index (BMI) (Narang et al., 2012). The consequences do not stop there. Higher chances of obesity and metabolic disorder are discovered in people lacking in sleep (Lauderale et al., 2009). This is primarily related to the body being more sensitive to insulin and an increase in appetite, which motivates higher levels of food intake (Cedernaes et al., 2015). Sleep quality is highly involved in the determination of body compositions, such as BMI, fat percentage, insulin sensitivity, and resistance (Jennings et al., 2007).

Condition of Internal Dysfunction:. The data is correspondingly computed with the attributes of activity frequencies. With each activity, it is then examined to which extent a participant might experience issues related to internal functions. The most insightful information achievable from the heatmap is that most daily habits negatively, impact health, more specifically, the internal health condition. The strongest correlated with the attribute internal dysfunction is shockingly, the device frequency attribute, with +0.2/1.

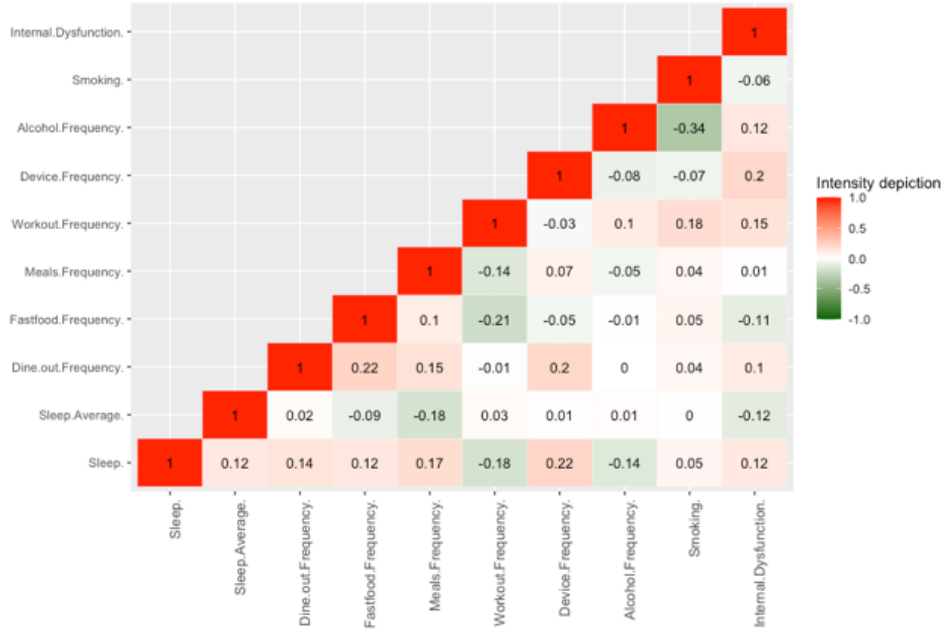

Figure 4. Heatmap for the relation between internal health conditions and other elements

It is quite reasonable to explain such a trend as the 21st century has been famously known as the "digital age" (Thomas, 2019). It has been studied that the energy emitted from these devices brings about multiple health concerns (Naeem, 2014). The electromagnetic radiation coming from the screen is believed to be consumed by human tissues and thus, has biological effects on the human cells. Not to mention that there have been numerous arguments brought about the connection of extensive electronic device exposure to brain issues, for instance, the development of brain cancer. From the figure, another easily seen aspect is that alcohol has a positive correlation with health issues. This is not a new discovery, but a topic previously deliberated on. Multiple papers and articles have identified the devastating outcome that alcohol might have on the body. Several body organs are found to undergo significant suffering from the influence of excessive alcohol, including the heart, brain, stomach, and specifically, the liver. In the long term, addiction to alcohol would make way for dangerous health conditions and, in the worst scenario, shorten life expectancy. It can also be perceived from the figure that sleep is another constituent of the well-being of the human body. An explanation for this point is that there are some internal reactions that only function when the body is asleep and thus, if the sleep average is reduced, those functions are interrupted tremendously. A prolonged period of those disruptions could foster a disordered behavior on the hormones and biological reactions to the body. The only attribute that appears to have no connection, only +0.01, is the frequency of meals attribute. The heatmap fails to understand the number of meals one individual consumes a day could have a potential effect on the body. Therefore, the related information on this topic will not be delved into in this research.

Physical Pains Feeling and Related Factors: A pair-wise plot noted from the graph:

- Red color stands for the 18-25 age group
- Green color stands for the 26-40 age group
- Blue color stands for 40-60 age group

The aim of this visualization is to disclose the data acquired from analyzing the physical pains and their related variables. The Figure illustrates the correlation values between the attributes as well as the density distribution of each variable. There are 86 samples from the age group of 18-25, while 23 and 5 answers were received from the 26-40 and 4.

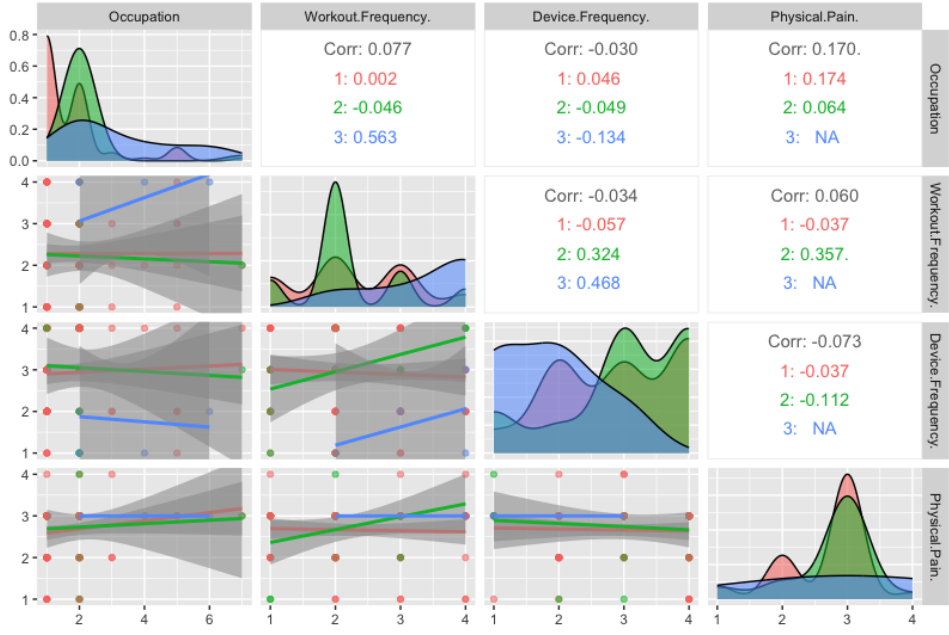

Figure 5. Pair-wise plot for the relation between physical pains and other variables

60 age groups, respectively. Those objects are then assessed with the response to the occurrence of physical pain, which establishes the diagram.

The age group 18-25, which is made up of mostly students, fluctuates in the data on workout frequency. However viewing the distribution graph, there is no significant point, as the data appears to

be quite constant. Meanwhile, for the Device Frequency graph, the data is tilted immensely to the right. This has exposed the habits of indulging in electronic devices in young people. Last but not least, it is displayed that most of the participants aged 18-25 experienced somewhat physical discomfort, with a peak point skewed to the right of the visualization.

Mentioning the 26-40 age group, it is noticeable from the graph that the most common occupation for this age group is office-related jobs. In contrast to the graph of workout frequency, the result for device frequency has a colossal slope to the far right of the distribution. It corresponds to the extensive usage of devices in this age group. The number is even greater compared to the 18-25 age group.

Considering the last age group of 40-60, it is surprising that the graph for Workout Frequency in this group has a tendency to spread to the right, which shows that people from 40 to 60 spend more time in physical activities. The differences are even emphasized by how older people spend less time on electronic devices. The graph is immensely distributed to the left and contains a sharp slope to the right, showing the scarcity of devices that occur in their daily routines. The most astounding point from the density plot is how the age group 40-60 has a much lower distribution in the Physical Pains attribute. Unfortunately, R cannot compute the correlation value for the group age 40-60 with the Physical Pain variable, as shown in "NA" in Figure 6. The reason behind this is that among the 114 results, there are only 5 results from this age group, which all accidentally responded to the same answer of "Sometimes". Hence, the correlation coefficient could not calculate correctly the data with the lack of data variation. From the comparison above between the ages, it is without denying that occupation, workout, and Device Usage Frequency have a distinctive interconnection with Physical Pains. Participants whose jobs demanding of working inside offices have a higher chance of undergoing aches and cramps in the body. The same applies to students who must spend much time sitting inside classrooms.

4. CONCLUSION

The research has attempted to provide insight into the relationship between lifestyle and health conditions in modern society. As the main tool for analyzing data, the use of R programming language has provided support for the preparation, calculation, and visualization of the retrieved dataset, which was collected through the help of Google Forms. The obtained result has successfully investigated the topic, creating an opportunity for comparison with the studies that have also been conducted on the subject. There are great opportunities for the research to be further developed, given that it has already established a solid foundation on the topic. One possibility includes retrieving more results from a wider range of participants, such as, from different regions or diverse backgrounds, which would increase the reliability of the research. Another option is to engage more data regarding the biological clock of each individual, given that different bodies have significantly dissimilar ways of functioning and adapting to the environment. However, it is rather a varied result and thus, there is flawed data that could be deemed as conflicting, when being put side by side with some other studies. The precision did outweigh, and such inaccuracies could be potentially eliminated if more responses are collected and observed, given more time allowed for the research.

REFERENCES

Arlinghaus, K.R., Johnston, C.A., (2015). The imporatnce of creating habits and routine. *American Journal of Lifestyle Medicine*, *13*(2), 142–144. https://doi.org/10.1177/1559827618818044

Cedernaes, J., Schioth, H. B., Benedict , C. (2015). Determinants of shortened, disrupted, and mistimed sleep and associated metabolic health consequences in healthy humans. *Diabetes*, *64*(4), 1073–1080. https://doi.org/10.2337/db14-1475

Deborah, A. C., Sonja C. K., Stefanie S. (2013) Healthy habits: The connection between diet, exercise, and locus of control. *Journal of Economic Behavior and Organization*, *98*, 1-28. https://doi.org/https://doi.org/10.1016/j.jebo.2013.10.011

de Vries, H., Eggers, S. M., Lechner, L. et al. (2014). Predicting fruit consumption: the role of habits, previous behavior and mediation effects. *BMC Public Health*, *730*(14). https://doi.org/10.1186/1471-2458-14-730

Goldman S. (2022). *How Does Sleep Affect Your Weight?*. https://comprehensivesleepcare.com/2022/01/04/weight-loss-and-sleep/

Hautekiet, P., Saenen, N. D., Martens, D. S., Debay, M., Van der Heyden, J., Nawrot, T. S., De Clercq, E. M. (2022). A healthy lifestyle is positively associated

with mental health and well-being and core markers in ageing. *BMC Med*, *20*, 328. https://doi.org/10.1186/s12916-022-02524-9

Hayes, A. (2022). *Six Sigma: Methodology and Belt Rankings*. https://www.investopedia.com/terms/s/six-sigma.asptoc-what-is-six-sigma

Jennings, J. R., Muldoon, M. F., Hall, M., Buysse, D. J., Manuck, S. B. (2007). self-reported sleep quality is associated with the metabolic syndrome. *Sleep*, *30*(2), 219– 223. https://doi.org/10.1093/sleep/30.2.219

Lauderdale, D. S., Knutson, K. L., Rathouz, P. J., Yan, L. L., Hulley, S. B., Liu, K. (2009). crosssectional and longitudinal associations between objectively measured sleep duration and body mass index: The CARDIA sleep study. *Am J Epidemiol*, *170*(7), 805–813. https://doi.org/10.1093/aje/kwp230

Mental Health Foundation. (2021). *Smoking and mental health*. https://www.mentalhealth.org.uk/explore-mental-health/a-z-topics/smokingand-mental-health

Naeem, Z. (2014) Health risks associated with mobile phones use. *Int J Health Sci (Qassim)*, *8*(4), V–VI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4350886/

Narang, I., Manlhiot, C., Davies-Shaw, J., Gibson, D., Chahal, N., Stearne, K., Fisher, A., Dobbin, S., McCrindle, B. W. (2012). Sleep disturbance and cardiovascular risk in adolescents. *CMAJ*, *184*(17), 913–920. https://doi.org/10.1503/cmaj.111589

Nickolas, S. (2021) *Correlation coefficients: Positive, negative, and zero*. https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

Stevens, E. (2023). *The 7 most useful data analysis methods and techniques*. https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/

The Healthcare Insight. (2020). *Importance of data analysis in healthcare.* https://thehealthcareinsights.com/importanceof-data-analysis-in-healthcare/

Thomas, D. (2019) *What is the digital age?*. https://www.ventivtech.com/blog/whatis-the-digital-age

U.S. Department of Health and Human Services. (2021) *Get enough sleep*. https://health.gov/myhealthfinder/healthy-living/mental-health-andrelationships/get-enough-sleep

World Health Organization. (2022) *Tobacco*. https://www.who.int/news-room/factsheets/detail/tobacco

Ziglio, E., Currie, C., Rasmussen, V. B. (2004) The WHO cross-national study of health behavior in school-aged children from 35 countries: Findings from 2001–2002. *J School Health*, *74*, 204–206.

Zinda, Z. (2021) *Data science stats review: Pearson's, Kendall's, and Spearman's Correlation for feature selection*. https://www.phdata.io/blog/data-science-stats-review/#:~:text=Spearman's%20Rank%20Correlation,preferred%20method%20of%20the%20two.