

Analysis of Cybersecurity Risks and Teenage Digital Behavior Patterns

Eric McCloy
School of Technology and Computing
City University of Seattle
Seattle WA, USA
0009-0003-7177-744X

Samuel Nimako-Mensah
School of Technology and Computing
City University of Seattle
Seattle WA, USA
0009-0000-7119-0807

Albert Samigullin
School of Technology and Computing
City University of Seattle
Seattle WA, USA
0009-0004-5465-0807

Abstract—As teenagers increasingly engage with digital technology, cybersecurity vulnerabilities present significant risks to their online safety and privacy. Adolescents who lack awareness of secure online practices are particularly vulnerable to malicious actors seeking to exploit them. This paper is an empirical study investigating the relationship between real-world online behavior of teenagers, cybersecurity risks, and device interactions. The primary data set used for this analysis is teenage online behavior and cybersecurity risks. First, we consider demographic information: age, education, time spent online. We correlate this with online behaviors: use of a VPN, type of equipment (computer, mobile), use of public internet, engagement with risky websites. Finally, we analyze the data set using a combination of demographic and behavioral patterns to search for high-risk, negative outcomes. For our research, we analyze teenage online behavior patterns to identify key risk factors, develop predictive models for cybersecurity vulnerabilities, and produce actionable visualizations that illustrate the relationship between digital literacy and online safety. Our findings utilize data and business analytics to provide evidence-based recommendations for parents, educators, and policymakers to enhance teenage cybersecurity awareness and protective strategies.

Keywords—*online behavior, teenagers, adolescent, privacy, cybersecurity risk, data analysis*

I. INTRODUCTION

The internet has permeated nearly every aspect of human life, including the daily experiences of children and adolescents. While it offers numerous benefits—such as enhanced communication, access to information, entertainment, and educational opportunities—it also introduces significant cybersecurity risks, particularly for younger users. Due to their limited experience and cognitive development, teenagers may not immediately recognize security threats and may fall victim before realizing the danger [7]. Cybersecurity awareness practices aim to equip users with the knowledge and skills to identify and mitigate such risks. However, while these practices are generally considered effective, research suggests that their implementation may

not be sufficient. Ondrušková and Pospíšil [5] argue that a single training session does not meaningfully improve cybersecurity awareness in children. Similarly, Mwagwabi and Jiow [4] found that even when teenagers suspected their computers might be compromised, this suspicion did not influence their password choices—indicating that passive guidelines alone are inadequate unless enforced through stronger authentication mechanisms. Moreover, Mwagwabi and Jiow [4] highlight a notable lack of theory-based studies on teenage cybersecurity behavior, suggesting a need for deeper exploration in this area.

This paper seeks to address this gap by analyzing the Teenage Online Behavior and Cybersecurity Risks dataset. By examining variables such as age, time spent online, device usage, and specific online behaviors (e.g., VPN use, risky website engagement, public network access), the study identifies key risk factors associated with cybersecurity vulnerabilities in teenagers. The results are used to develop correlations and actionable visualizations that illustrate the relationship between digital behavior and online safety. These insights aim to provide evidence-based recommendations for parents, educators, and policymakers to enhance cybersecurity awareness and protection strategies for adolescents.

II. LITERATURE REVIEW

An examination of studies related to adolescent online behaviors and exposures to negative outcomes revealed a complex interplay of factors, including some counterintuitive findings. For example, teens who had a higher “digital literacy” (technical understanding and skillset, including being able to recognize online threats) were still exposed to a greater number of negative online experiences because they tended to spend significantly more time online [12]. Another study revealed that strict parental control of social media usage resulted in a small but significant increase in the likelihood of teens sharing private information online [3]. Considering these observations, it is important to take a careful and nuanced look at the literature in order to challenge potentially erroneous preconceptions.

Adorjan and Ricciardelli [1] explored the issue of “online addiction” as a sensitizing concept. By this, they meant that

instead of trying to confirm previous studies findings about the extent of addiction or trying to define it precisely themselves, they used a focus group with a discussion-based methodology to understand teens' self-perception of online addiction. During their study, they avoided using the term "online addiction" until it was brought up by a focus group participant. In all 35 independent focus groups they ran during the study, participants brought up the term "online addiction" to describe themselves without prompting or previous mention of the topic. The study found that the primary focus of online activity tended to be school-based networks of online friends implying that perhaps the addiction isn't to technology itself but to the social interaction that the technology enables. Many of the focus group participants also noted their parents' behavior modeled online addiction for them.

The question is, does online addiction have negative consequences? There were several direct, broad categories of online risks identified in these studies, as well as negative well-being outcomes associated with extensive internet usage and the exposure to these risks. The major types of direct risks to teens can be categorized as Content, Contact, and Criminal Risks.

1. Content Risks are related to potentially harmful content, such as sexual images, violent images, or hate speech [8].
2. Contact Risks are unwanted interactions with people online. This includes bullying, being drawn into unwanted conversations and arguments, attempted real-life contact by strangers, recruitment by hate groups, and sexual solicitation [1] [8]
3. Criminal Risk includes offers to sell alcohol or drugs, gambling, and blackmail [8][9].

In addition to direct risks, the studies identified a number of negative well-being or mental health outcomes. These included: low self-esteem, depression, anxiety, sleep deprivation, loneliness, and feelings of inadequacy [1][12]. Specifically, the "Fear of Missing Out" (FOMO) was cited as a major reason for constant addictive behavior on social media sites and was positively correlated with "boredom, loneliness, depression, and feelings of inadequacy and anxiety, as well as diminished well-being, overall mood, and life satisfaction" [1] (p. 51).

Poor academic performance may also be correlated with addictive online behaviors. However, this relationship is complex. Use of online tools can be correlated with positive performance, but high internet use at school and home has been linked with low academic performance, specifically in declining mathematics grades. Online addiction is also associated with a decreased motivation to study, poor cognitive behavior control, and deteriorating relationships with teachers and schoolmates. All these issues correlated with degraded academic performance [6].

Parental intervention is a particularly complex aspect of teen online addiction. Álvarez-García *et al.* [2] found a small

but statistically significant negative relationship. With increased parental restriction, there was an increase in adolescent risky behaviors online. Kang *et al.* [3], found that the type of parental intervention mattered. They categorized parental intervention as "restrictive" when a tight set of rules were expected to be followed and "active" when the parents' helped teens develop an understanding of the issues and involved the teens in creating boundaries. The study showed that teens with restrictive parents were less likely to use tight privacy controls on Douyin (a Tik Tok type of app popular in China) than teens with active parents. The researchers concluded that teens who were actively included in the process were better able to make considered judgements regarding protecting themselves online.

Finally, the issue of digital literacy was examined in the literature. Vissenberg *et al.* [12] define digital literacy as "the skills, knowledge and attitudes that make learners able to use digital media in a critical, responsible and creative manner" (p. 77). However, as they examined the issue more closely, they came to the conclusion that, even though digitally literate teenagers are better able to avoid risks, they still encountered direct risks more often, because they spend more time online. When the researchers questioned teens who had risky online experiences, they also discovered that teens with higher levels of digital literacy also demonstrated higher levels of "Online Resilience", that is, they were better able to process and cope with the negative experience without a long-term effect on their well-being.

III. BACKGROUND

The data used in our analysis was collected from network activity logs and e-safety monitoring systems across various educational institutions and households in Texas and California over a 7-year period (2017-2024). The dataset comprises 67,921 observations and 30 columns including 19 numerical and 11 categorical variables.

The quality of the dataset is demonstrated by its use in the literature. Most notably, Xu, *et al.* used the dataset to train an AI to create a framework for teen learning related to cybersecurity [12].

To enhance our understanding of the data, we organized the columns into the following groups: Time and User Demographics, Device and Network Information, Security Threats and Incidents, User Behavior and Activities, Authentication and Access, Safety and monitoring Controls and Risk Assessment. (Appendix A).

Each group offers a unique perspective for understanding and examining the relationship between teenage behavior and cybersecurity risks. In the User Demographics category, we analyze the timestamp of each incident, age group of the teenager and the number of hours they were online. The age groups are categorized as follows: children under 13, teens between 13 and 16 and teens aged 17-19. Table I shows the age distribution in the dataset indicating a predominant representation of middle teens (13-16) that comprises 70.2%

(47,695 records) of all observations followed by 19.9% (13,491 records) for older teens and 9.9% (6,735 records) for pre-teens.

TABLE I. Age Distribution of Participants

Age Group	Count(n)	Percent (%)
Under 13	6,735	9.9
13-16	47,695	70.2
17-19	13,491	19.9
Total	67,921	100.0

In the device and network information category, we examine several key variables including device type (smartphones, laptops, desktops, and tablets). These device types will be used to create a derived feature called Device Security Index, which assigns values based on the relative security of each device type.

The dataset includes several risk indicators related to cybersecurity threats, such as malware incidents, phishing attempts and visits to risky websites. Positive online security practices such as the use of strong passwords used, whether a teenager used a VPN and public network usage is also captured along with the individual's level of e-safety awareness.

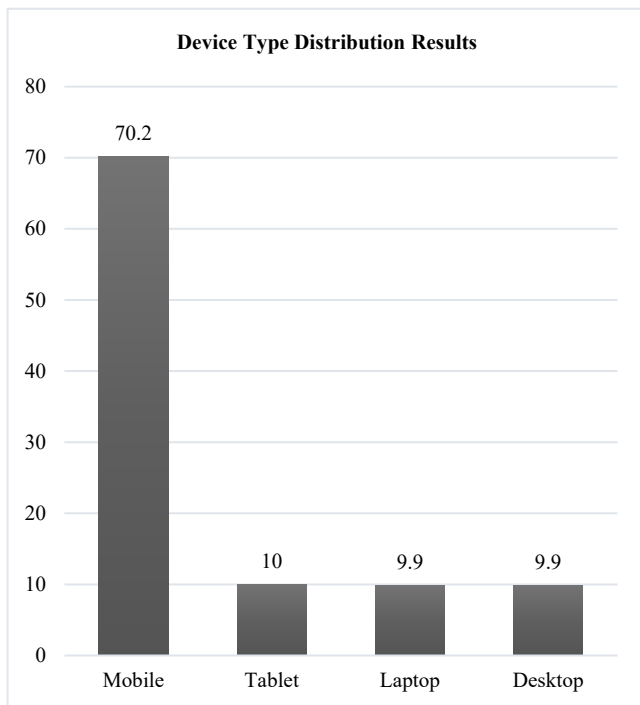


Fig. 1. Distribution of Device Types

As shown in Figure 1, mobile phones are the most dominant platform accounting for 70.17% (47,662 records) of usage. This suggests a mobile first mindset in how teenagers interact online. Traditional devices (Laptops, Tablets and Desktops) account for 29.8% (20,259 records) of usage suggesting these devices are used in more targeted or specific contexts.

IV. METHODOLOGY

A. Data Acquisition

The data was downloaded from Kaggle – the data science competition platform at <https://www.kaggle.com/datasets/datasetengineer/teenage-online-behavior-and-cybersecurity-risks> as a single 8.8Mb CSV file. The repository was first uploaded on October 9th, 2024, and has been downloaded 338 times with 1987 views as of May 23, 2025.

Our initial task was to examine the structure of the dataset using the pandas commands `df.shape` and `df.info` to understand its dimensions and data types. The full dataset comprises 67,921 observations with 30 columns.

B. Data Cleaning Strategy

In our analysis of the data, we identified several issues with consistency and completeness. As such, we crafted a detailed strategy to clean the data and determine which elements should be discarded, replaced or otherwise modified.

1) Replacing Missing Values

The `Education_Content_Usage` column contained 47595 null values, significantly affecting the completeness of the dataset. To address this, we removed the column entirely. For other columns, we standardized any missing values by replacing them with `np.nan`.

2) Clean Dates

Visually inspecting the `TimeStamp` column, we found the date format to mainly be consistent. However, to ensure consistency we applied logic to parse the data in the correct format.

3) Handling other data inconsistencies

Here is the approach we took to clean the data:

- Replace missing values
- Parse dates as noted above
- Convert to numeric and handle non-numeric numbers
- Create derived columns

TABLE II. Derived Features for Data Analysis

Feature Name	Type	Formula
Usage_Pattern_Category	category	Based on Hours_Online: Light (≤ 2), Moderate (≤ 5), Heavy (> 5)
Time_Period_Risk	category	Hour-based risk: Low (7–15), Medium (15–22), High (22–7)
Device_Security_Index	float	Mapped from Device_Type with fixed scores
Connection_Safety_Score	float	Composite score: VPN (0.4), Public_Network (0.3 inverse), Network_Type (0.3)
Total_Threat_Exposure	int	Sum of Malware_Detection, Phishing_Attempts, Data_Breach_Notifications
Threat_Severity_Index	float	Weighted threat score: Malware (1), Phishing (2), Breach (3)
Defensive_Posture_Score	float	Firewall_Logs (0.6) + VPN_Usage (0.4) scaled to 10
Risky_Browsing_Ratio	float	Risky_Website_Visits / Website_Visits
Digital_Consumption_Index	float	Weighted activity: Website (0.4), Ads (0.3), Cloud (0.3)
Password_Security_Score	int	Mapped from Password_Strength: weak=2, moderate=6, strong=10
Parental_Oversight_Level	category	Based on Parental_Control_Alerts: Low (0), Medium (≤ 3), High (> 3)
Supervision_Effectiveness	float	Prevention / (Threats + Prevention)
Safety_Behavior_Score	float	Avg. normalized VPN, Connection_Safety, Defensive_Posture
Awareness_Behavior_Gap	float	Numeric_Awareness - Safety_Behavior_Score
Overall_Security_Posture	float	Weighted average of multiple security scores
Security_Posture_Category	category	Categorized from Overall_Security_Posture (Vulnerable, Adequate, Secure)
Behavior_vs_Protection_Gap	float	Normalized risk - avg. protection scores
Protection_Match_Category	category	Categorized from Behavior_vs_Protection_Gap (Overprotected, Balanced, Underprotected)

C. Summary of cleaned data

Appendix B contains the full summary of cleaned data. Our initial analysis shows that 91.68% of the 67,921 records were 100% complete; Average completeness for the dataset was 99.84%. Observations in the dataset were made from Jan 1, 2017, to Jan 1, 2024, a period of 6 years.

The most predominant age group was the 13–16-year-olds who represent 70.2% followed by older teens (17–19) and children under 13. We observed an average daily usage of 2 hours when teens went online. In terms of devices, Mobile phones represent the largest device platform with a total of 70.2%. Traditional devices (Laptops, Tablets and Desktops) account for 29.8%. We see a clear trend in the data where 70.2% of 13–16-year-olds use mobile devices indicating the core teenage demographic is driving the mobile-first behavior.

V. DATA ANALYSIS

This section describes data analysis of Teenage online behavior and cybersecurity risks data analysis dataset from Kaggle by Sik, (2024). The data were cleaned as described in the previous section. To explore the relationships among key behavioral and security-related features, a subset of derived features was selected for correlation analysis presented in Table II. This subset included both numerical metrics (e.g., Digital_Consumption_Index, Threat_Severity_Index) and encoded categorical constructs (e.g., Usage_Pattern_Category, Security_Posture_Category). Prior to the analysis, selected categorical features were ordinally encoded to allow for correlation computation.

A Spearman correlation matrix was computed to assess monotonic relationships between the features. According to Hauke & Kossowski [11], this method is well-suited for ordinal and non-linear associations, which are expected in behavioral data. The correlation matrix was visualized using a heatmap to highlight strong associations (Figure 2).

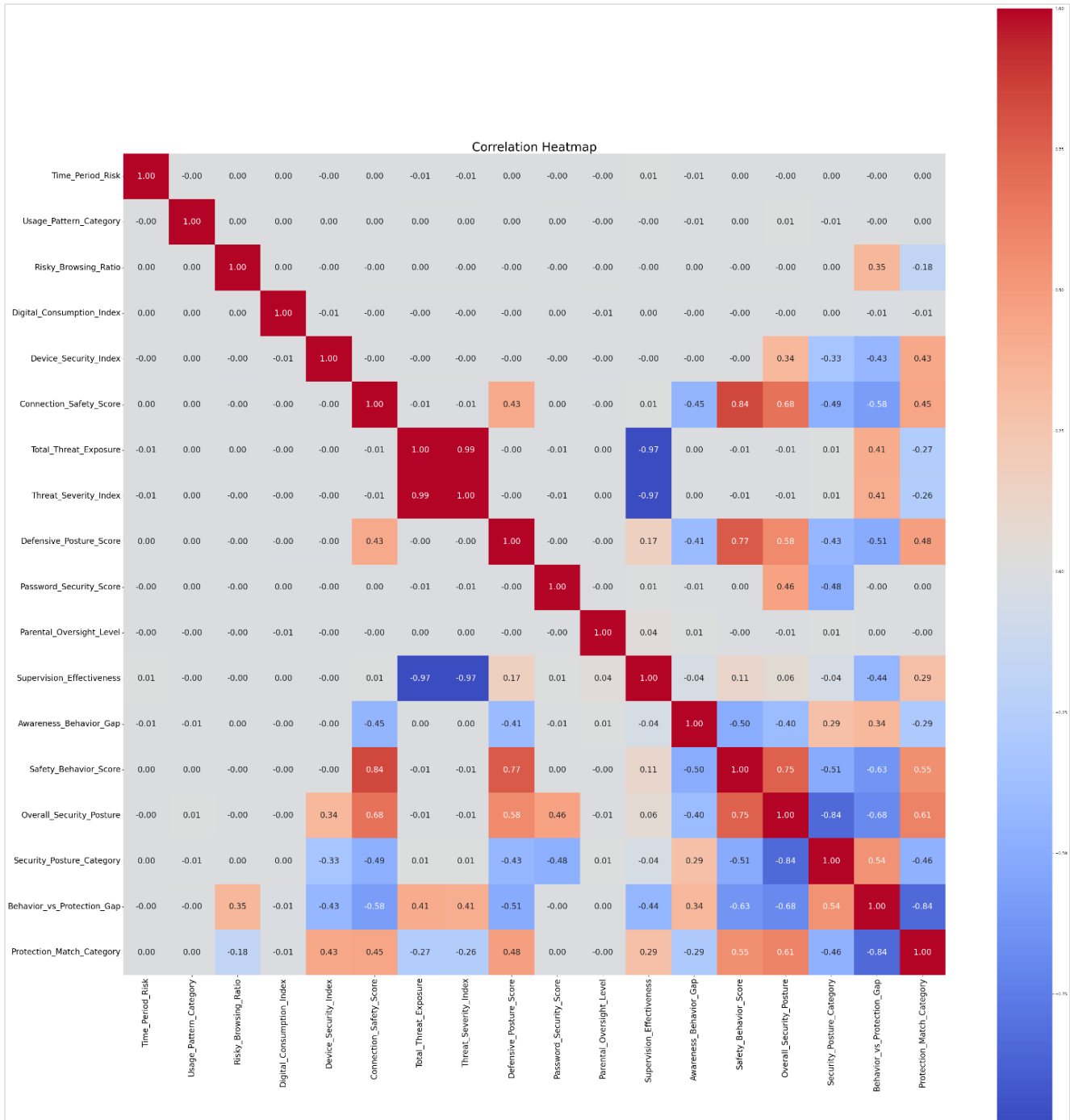


Fig. 2. Heatmap based on a Spearman correlation matrix to identify correlations between derived feature

Two sets of feature pairs were extracted from a Spearman correlation matrix:

- Positively correlated pairs, Spearman coefficient ≥ 0.5
- Negatively correlated pairs, Spearman coefficient ≤ -0.5

To ensure statistical rigor, each of these pairs was tested for significance using hypothesis testing. All extracted pairs demonstrated from moderate to strong correlation as shown in Table III. This process helped identify key patterns in the dataset, such as how certain online behaviors may co-vary with cybersecurity posture, awareness, or parental oversight.

TABLE III. Results of Testing for Significance

Feature X	Feature Y	Method	Corr. Coef	P-value
<i>Positively correlated pairs</i>				
Connection_Safety_Score	Safety_Behavior_Score	Spearman	0.84	0
Connection_Safety_Score	Overall_Security_Posture	Spearman	0.68	0
Total_Threat_Exposure	Threat_Severity_Index	Spearman	0.99	0
Defensive_Posture_Score	Safety_Behavior_Score	Spearman	0.77	0
Defensive_Posture_Score	Overall_Security_Posture	Spearman	0.58	0
Safety_Behavior_Score	Overall_Security_Posture	Spearman	0.75	0
Safety_Behavior_Score	Protection_Match_Category	Spearman	0.55	0
Overall_Security_Posture	Protection_Match_Category	Spearman	0.61	0
Security_Posture_Category	Behavior_vs_Protection_Gap	Spearman	0.54	0
<i>Negatively correlated pairs</i>				
Connection_Safety_Score	Behavior_vs_Protection_Gap	Spearman	-0.58	0
Total_Threat_Exposure	Supervision_Effectiveness	Spearman	-0.97	0
Threat_Severity_Index	Supervision_Effectiveness	Spearman	-0.97	0
Defensive_Posture_Score	Behavior_vs_Protection_Gap	Spearman	-0.51	0
Safety_Behavior_Score	Security_Posture_Category	Spearman	-0.51	0
Safety_Behavior_Score	Behavior_vs_Protection_Gap	Spearman	-0.63	0
Overall_Security_Posture	Security_Posture_Category	Spearman	-0.84	0
Overall_Security_Posture	Behavior_vs_Protection_Gap	Spearman	-0.68	0
Behavior_vs_Protection_Gap	Protection_Match_Category	Spearman	-0.84	0

To uncover distinct patterns in teenage online behavior and security posture, a K-Means clustering analysis was conducted using eleven behavioral and security-related features. These included indices related to digital activity (e.g., Digital_Consumption_Index, Device_Security_Index), security practices (e.g., Password_Security_Score, Defensive_Posture_Score), exposure indicators (Total_Threat_Exposure, Threat_Severity_Index), and demographic information (Age_Group_Encoded). All features were standardized prior to clustering to ensure equal contribution to distance computations. The clustering of behavioral profile is illustrated in Figure 3.

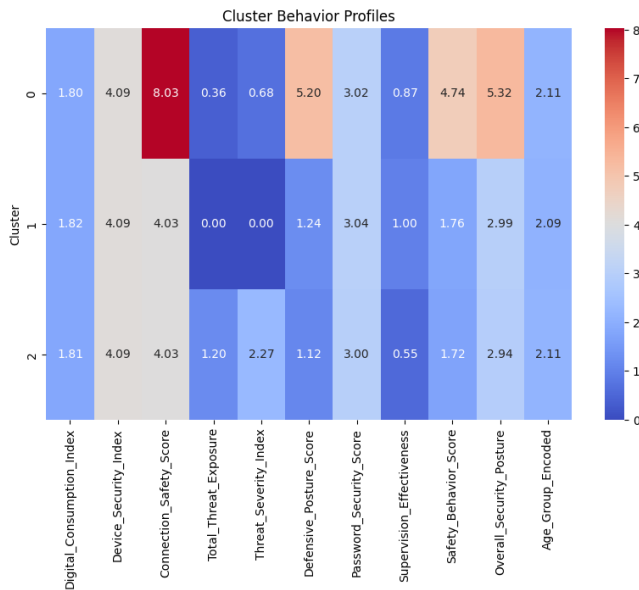


Fig. 3. Cluster Behavior Profiles

The heatmap in Figure 3 illustrates average scores of behavioral and security features across three user clusters derived via K-Means clustering. Cluster 0 represents risk-aware users with strong security behaviors. This cluster is characterized by a high Connection_Safety_Score (8.03) and Password_Security_Score (5.20), indicating strong adherence to secure online behaviors. The pattern is further reinforced by a moderate to high Safety_Behavior_Score (4.74) and an elevated Overall_Security_Posture (5.32). Additionally, very low levels of Total_Threat_Exposure (0.36) and Threat_Severity_Index (0.68) suggest that these defensive practices are effective in minimizing online risks. This cluster likely represents cyber-aware, well-supervised, and safety-conscious users.

Cluster 1 includes low-engagement users with minimal threat exposure. This cluster demonstrates very low Total_Threat_Exposure (0.00) and Threat_Severity_Index (0.00), indicating minimal engagement with online risks. Users exhibit low to moderate scores in key security behaviors, such as Password Security (1.24) and Safety Behavior (1.76).

Additionally, Supervision_Effectiveness (1.00) and Defensive_Posture_Score (1.24) are relatively low, suggesting limited external guidance or self-initiated precautions. These patterns imply that users in this cluster may have restricted or minimal interaction with online environments—possibly due to parental controls, limited internet access, or inherently low digital engagement.

Cluster 2 represents moderately engaged users with some risk and limited supervision. Cluster 2 includes users with higher exposure to online threats compared to Cluster 1, as indicated by a Total_Threat_Exposure score of 1.20 and a Threat_Severity_Index of 2.27. While their scores for Connection Safety (4.03), Password Security (3.00), and Overall Security Posture (2.94) are moderate, the notably low Supervision_Effectiveness (0.55) and Defensive_Posture_Score (1.12) raise concerns about their vulnerability. These children appear to have some awareness of cybersecurity risks but may lack consistent supervision or comprehensive protective practices, placing them at an intermediate level of risk.

Figure 4 demonstrates the demographic distribution of age groups across the three behavioral clusters, highlighting how users from different age ranges are represented within each cluster.

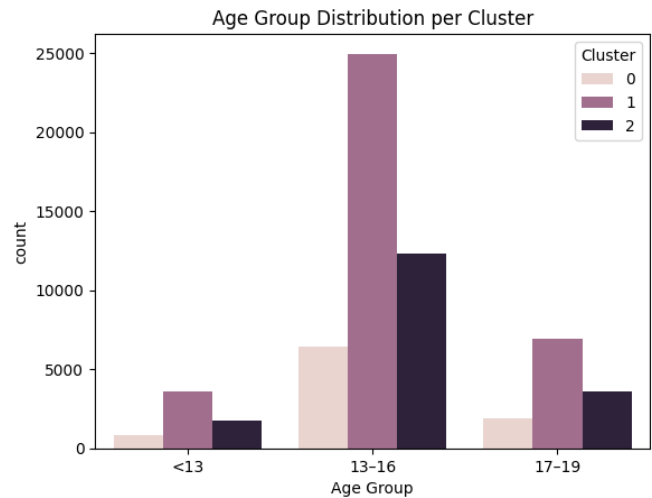


Fig. 4. Age Group Distribution per cluster

The largest segment of users across all clusters belongs to the 13–16 age group, with a significant concentration in Cluster 1, which denotes low-engagement or minimally exposed users. Clusters 0 and 2 have fewer users overall and display a more balanced distribution of age groups. This indicates that while age can affect behavioral clustering, it does not solely determine cluster formation, as similar age groups are present across different behavioral profiles.

To support interpretation, Principal Component Analysis (PCA) was applied to reduce the behavioral feature space to two principal components, enabling 2D visualization of cluster

separability. The resulting scatter plot (Figure 5) demonstrates visible differentiation among the three clusters, with some overlap between Cluster 0 and Cluster 2, while Cluster 1 forms a clearly distinct linear group.

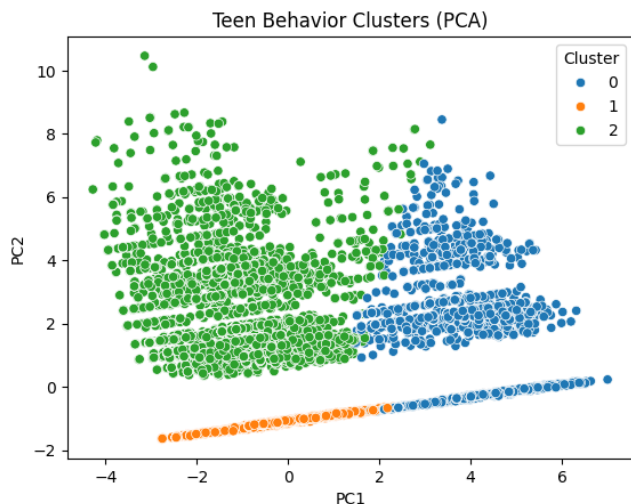


Fig. 5. PCA Behavior Clustering

The source code of the data analysis is in Appendix E.

VI. FUTURE RESEARCH

There are a number of gaps that stand out as opportunities for further research. The Kang *et al.* study [3] lays out an interesting relationship between restrictive and active parenting styles. However, the categories are broad, and the outcomes were only measured with reference to a single social media platform. This study was primarily an empirical look at a longitudinal data set. Further theoretical research would be beneficial. A study including additional apps and more nuanced categories regarding parenting restrictions could uncover more about teens and risky online behavior. A longitudinal study relating digital literacy to online resilience would also be interesting. A larger sample size and time period would show whether the observed resilience is more than a short-term coping mechanism.

VII. CONCLUSION

This paper is an analysis of teenage online behavior and its associated cybersecurity risks, leveraging the Teenage online behavior and cybersecurity risks dataset [9] to identify key patterns and vulnerabilities. The data shows that digital technology is an integral part of life for teenagers and that there are significant challenges in ensuring their online safety.

After careful data cleaning, a number of calculated fields were created. Examples include features such as: Time spent online, age, parental oversight, and threat scoring. With this data, we performed analyses such as Spearman correlations and K-Means clustering. This allowed us to define three distinct behavioral profiles among teenagers.

The profiles are: risk-aware users with strong security behaviors (Cluster 0), low-engagement users with minimal threat exposure (Cluster 1), and moderately engaged users with some risk and limited supervision (Cluster 2).

These profiles gave us insight into how different levels of digital activity, security practices, and parental oversight influence a teenager's overall cybersecurity standing.

Our findings support several conclusions:

- Digital literacy is important for threat recognition but is not sufficient to guarantee online safety. The digitally literate spend more time online and are exposed to more risk.
- The literature shows that strict parental measures are less effective than parents helping teens to understand and create boundaries to keep themselves safe.
- The different profiles provide a foundation for personalized training and intervention, instead of the traditional one-size-fits-all approach.

In conclusion, this analysis provides a framework for creating effective, personalized strategies for raising awareness and keeping adolescents safe online. It is hoped that these insights can lead to informed, resilient teenagers capable of making decisions and setting boundaries to keep themselves secure online.

REFERENCES

- [1] M. Adorjan and R. Ricciardelli, "Smartphone and social media addiction: Exploring the perceptions and experiences of Canadian teenagers," *Can. Rev. Sociol.*, vol. 58, no. 1, pp. 45–64, 2021. [Online]. Available: <https://doi.org/10.1111/cars.12319>
- [2] D. Álvarez-García, T. García, and Z. Suárez-García, "The relationship between Parental Control and high-risk internet behaviours in adolescence," *Soc. Sci. (Basel)*, vol. 7, no. 6, p. 87, 2018. [Online]. Available: <https://doi.org/10.3390/socsci7060087>
- [3] H. Kang, W. Shin, and J. Huang, "Teens' privacy management on video-sharing social media: the roles of perceived privacy risk and parental mediation," *Internet Res.*, vol. 32, no. 1, pp. 312–334, 2022. [Online]. Available: <https://doi.org/10.1108/intr-01-2021-0005>
- [4] F. Mwangi and J. Hee Jiow, "Compliance with security guidelines in teenagers: The conflicting role of peer influence and personal norms," *Aust. J. Inf. Syst.*, vol. 25, 2021. [Online]. Available: <https://doi.org/10.3127/ajis.v25i0.2953>
- [5] D. Ondrušková and R. Pospíšil, "The good practices for implementation of cyber security education for school children," *Contemp. Educ. Technol.*, vol. 15, no. 3, p. ep435, 2023. [Online]. Available: <https://doi.org/10.30935/cedtech/13253>
- [6] B. Ortega-Ruipérez, A. Castellanos Sánchez, and B. Marcano, "Risks in adolescent adjustment by Internet exposure: Evidence from PISA," *Front. Psychol.*, vol. 12, p. 763759, 2021. [Online]. Available: <https://doi.org/10.3389/fpsyg.2021.763759>
- [7] F. Quayyum, D. S. Cruzes, and L. Jaccheri, "Cybersecurity awareness for children: A systematic literature review," *Int. J. Child Comput. Interact.*, vol. 30, no. 100343, p. 100343, 2021. [Online]. Available: <https://doi.org/10.1016/j.ijcci.2021.100343>
- [8] E. Savoia, N. W. Harriman, M. Su, T. Cote, and N. Shortland, "Adolescents' exposure to online risks: Gender disparities and vulnerabilities related to online behaviors," *Int. J. Environ. Res. Public Health*, vol. 18, no. 11, p. 5786, 2021. [Online]. Available: <https://doi.org/10.3390/ijerph18115786>

- [9] L. Sik, "Teenage online behavior and cybersecurity risks." Kaggle, 2024. [Online]. Available: <https://doi.org/10.34740/KAGGLE/DSV/9587284>
- [10] K. Stewart, G. Brodowsky, and D. Sciglimpaglia, "Parental supervision and control of adolescents' problematic internet use: understanding and predicting adoption of parental control software," *Young Consum. Insight Ideas Responsible Mark.*, vol. 23, no. 2, pp. 213–232, 2022. [Online]. Available: <https://doi.org/10.1108/yc-04-2021-1307>
- [11] Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2), 87-93.
- [12] J. Vissenberg, L. d'Haenens, and S. Livingstone, "Digital literacy and online resilience as facilitators of young people's well-being?: A systematic review," *Eur. Psychol.*, vol. 27, no. 2, pp. 76–85, 2022. [Online]. Available: <https://doi.org/10.1027/1016-9040/a000478>
- [13] S. Xu, S. Peng, and J. An, "TransDenseInceptionNet: A deep learning framework for teenage cybersecurity awareness using real-world E-safety data," *Informatica (Ljubl.)*, vol. 49, no. 18, 2025.

APPENDIX A – DATASET CATEGORIES AND FIELDS

CATEGORY	VARIABLES
TIME AND USER DEMOGRAPHICS	Timestamp
	Age_Group
	Hours_Online
DEVICE AND NETWORK INFORMATION	Device_Type
	Network_Type
	Geolocation
	Public_Network_Usage
	VPN_Usage
SECURITY THREATS AND INCIDENTS	Malware_Detection
	Phishing_Attempts
	Data_Breach_Notifications
	Malware_Exposure_Risk
	Firewall_Logs
	Unencrypted_Traffic
USER BEHAVIOR AND ACTIVITIES	Social_Media_Usage
	Website_Visits
	Peer_Interactions
	Risky_Website_Visits
	Cloud_Service_Usage
	Education_Content_Usage
	Ad_Clicks
	Online_Purchase_Risk
AUTHENTICATION AND ACCESS	Login_Attempts
	Insecure_Login_Attempts
	Password_Strength
SAFETY AND MONITORING CONTROLS	Cyberbullying_Reports
	Parental_Control_Alerts
	Download_Risk
RISK ASSESSMENT	E_Safety_Awareness_Score
	Cybersecurity_Behavior_Category

APPENDIX B – SUMMARY OF DATASET

Dataset Shape: 67921 rows, 30 columns

=====

TEEN CYBERSECURITY DATASET SUMMARY

=====

Total records: 67921

=====

AGE GROUP DISTRIBUTION

=====

Age Group	Count	Percentage
13-16	47695	70.2%
17-19	13491	19.9%
<13	6735	9.9%

=====

DEVICE USAGE

=====

Device Type	Count	Percentage
Mobile	47662	70.2%
Tablet	6800	10.0%
Laptop	6759	10.0%
Desktop	6700	9.9%

Device Usage by Age Group (percentages):

Device_Type	Desktop	Laptop	Mobile	Tablet
Age_Group				
13-16	9.9	10.0	70.2	10.0
17-19	9.7	9.7	70.5	10.1
<13	10.1	10.4	69.3	10.1

=====

ONLINE HOURS

=====

Basic Statistics:

Average (mean): 1.99 hours/day
 Median: 1.38 hours/day
 Minimum: 0.00 hours/day
 Maximum: 22.66 hours/day
 Standard deviation: 2.00 hours

Hours Online Distribution:

Hours Range	Count	Percentage
0-2	43193	63.6%
2-4	15664	23.1%
4-6	5716	8.4%
6-8	2095	3.1%
8-10	763	1.1%
10-12	299	0.4%
12+	191	0.3%

Average Hours Online by Age Group:

13-16: 1.99 hours/day
 17-19: 2.00 hours/day
 <13: 1.99 hours/day

Average Hours Online by Device Type:

Laptop: 2.00 hours/day
 Tablet: 2.00 hours/day
 Mobile: 1.99 hours/day
 Desktop: 1.97 hours/day

High Usage (>8 hours/day): 1.8% of users

Percentage of High Usage (>8 hours/day) by Age Group:

13-16: 1.9%
 17-19: 1.6%
 <13: 1.8%

APPENDIX C – CODE TO SUMMARIZE DATA

```

IMPORT PANDAS AS PD
IMPORT NUMPY AS NP

DEF GENERATE_DATASET_SUMMARY(CSV_FILE_PATH):
    """GENERATE SUMMARY STATISTICS FOR THE TEEN CYBERSECURITY DATASET."""
    # LOAD DATASET
    DF = PD.READ_CSV(CSV_FILE_PATH)

    # BASIC DATASET INFORMATION
    PRINT(F"DATASET SHAPE: {DF.SHAPE[0]} ROWS, {DF.SHAPE[1]} COLUMNS\n")

    PRINT("=" * 50)
    PRINT("TEEN CYBERSECURITY DATASET SUMMARY")
    PRINT("=" * 50)
    PRINT(F"TOTAL RECORDS: {DF.SHAPE[0]}")
    PRINT("\n")

    # 1. AGE GROUP DISTRIBUTION
    PRINT("=" * 50)
    PRINT("AGE GROUP DISTRIBUTION")
    PRINT("=" * 50)

    AGE_DIST = DF['AGE_GROUP'].VALUE_COUNTS().SORT_INDEX()
    AGE_PCT = DF['AGE_GROUP'].VALUE_COUNTS(NORMALIZE=TRUE).SORT_INDEX() * 100

    # CREATE A TABLE FOR AGE DISTRIBUTION
    PRINT(F{'AGE_GROUP':<15} {'COUNT':<10} {'PERCENTAGE':<10}")
    PRINT("-" * 35)
    FOR AGE, COUNT IN AGE_DIST.ITEMS():
        PRINT(F"AGE:<15} {COUNT:<10} {AGE_PCT[AGE]:.1F}%")

    PRINT("\n")

    # 2. DEVICE USAGE
    PRINT("=" * 50)
    PRINT("DEVICE USAGE")
    PRINT("=" * 50)

```

```

DEVICE_DIST = DF['DEVICE_TYPE'].VALUE_COUNTS()
DEVICE_PCT = DF['DEVICE_TYPE'].VALUE_COUNTS(NORMALIZE=TRUE) * 100

# CREATE A TABLE FOR DEVICE DISTRIBUTION
PRINT(F"{'DEVICE TYPE':<20} {'COUNT':<10} {'PERCENTAGE':<10}")
PRINT("-" * 40)
FOR DEVICE, COUNT IN DEVICE_DIST.ITEMS():
    PRINT(F"{'DEVICE':<20} {'COUNT':<10} {'DEVICE_PCT[DEVICE]:.1F}%")

# DEVICE USAGE BY AGE GROUP
PRINT("\nDEVICE USAGE BY AGE GROUP (PERCENTAGES):")
DEVICE_BY_AGE = PD.CROSSTAB(DF['AGE_GROUP'], DF['DEVICE_TYPE'], NORMALIZE='INDEX')
* 100
PRINT(DEVICE_BY_AGE.ROUND(1))

PRINT("\n")

# 3. ONLINE HOURS
PRINT("=" * 50)
PRINT("ONLINE HOURS")
PRINT("=" * 50)

HOURS_STATS = DF['HOURS_ONLINE'].DESCRIBE()

PRINT("BASIC STATISTICS:")
PRINT(F" AVERAGE (MEAN): {HOURS_STATS['MEAN']:.2F} HOURS/DAY")
PRINT(F" MEDIAN: {HOURS_STATS['50%']:.2F} HOURS/DAY")
PRINT(F" MINIMUM: {HOURS_STATS['MIN']:.2F} HOURS/DAY")
PRINT(F" MAXIMUM: {HOURS_STATS['MAX']:.2F} HOURS/DAY")
PRINT(F" STANDARD DEVIATION: {HOURS_STATS['STD']:.2F} HOURS")

# HOURS DISTRIBUTION
PRINT("\nHOURS ONLINE DISTRIBUTION:")
HOURS_BINS = [0, 2, 4, 6, 8, 10, 12, DF['HOURS_ONLINE'].MAX() + 0.1]
HOURS_LABELS = ['0-2', '2-4', '4-6', '6-8', '8-10', '10-12', '12+']
DF['HOURS_BINNED'] = PD.CUT(DF['HOURS_ONLINE'], BINS=HOURS_BINS, LABELS=HOURS_LABELS)

HOURS_DIST = DF['HOURS_BINNED'].VALUE_COUNTS().SORT_INDEX()
HOURS_PCT = DF['HOURS_BINNED'].VALUE_COUNTS(NORMALIZE=TRUE).SORT_INDEX() * 100

```

```

PRINT(F"{'HOURS_RANGE':<15} {'COUNT':<10} {'PERCENTAGE':<10}")
PRINT("-" * 35)
FOR HOURS_RANGE, COUNT IN HOURS_DIST.ITEMS():
    PRINT(F"{'HOURS_RANGE':<15} {'COUNT':<10} {'HOURS_PCT[HOURS_RANGE]:.1F}%")

# HOURS BY AGE GROUP
PRINT("\NAVERAGE HOURS ONLINE BY AGE GROUP:")
HOURS_BY_AGE = DF.GROUPBY('AGE_GROUP')['HOURS_ONLINE'].MEAN().SORT_INDEX()
FOR AGE, HOURS IN HOURS_BY_AGE.ITEMS():
    PRINT(F" {AGE}: {HOURS:.2F} HOURS/DAY")

# HOURS BY DEVICE TYPE
PRINT("\NAVERAGE HOURS ONLINE BY DEVICE TYPE:")
HOURS_BY_DEVICE = DF.GROUPBY('DEVICE_TYPE')['HOURS_ONLINE'].MEAN()
.SORT_VALUES(ASCENDING=FALSE)
FOR DEVICE, HOURS IN HOURS_BY_DEVICE.ITEMS():
    PRINT(F" {DEVICE}: {HOURS:.2F} HOURS/DAY")

# HIGH USAGE STATISTICS
HIGH_USAGE = DF[DF['HOURS_ONLINE'] > 8]
HIGH_USAGE_PCT = (LEN(HIGH_USAGE) / LEN(DF)) * 100
PRINT(F"\NHIGH USAGE (>8 HOURS/DAY): {HIGH_USAGE_PCT:.1F}% OF USERS")

# HIGH USAGE BY AGE
HIGH_BY_AGE = HIGH_USAGE.GROUPBY('AGE_GROUP').SIZE() / DF.GROUPBY('AGE_GROUP').SIZE()
* 100
PRINT("\NPERCENTAGE OF HIGH USAGE (>8 HOURS/DAY) BY AGE GROUP:")
FOR AGE, PCT IN HIGH_BY_AGE.ITEMS():
    PRINT(F" {AGE}: {PCT:.1F}%")

# IDENTIFY NUMERICAL AND CATEGORICAL COLUMNS
NUMERICAL_COLS = DF.SELECT_DTYPES(INCLUDE=['INT64', 'FLOAT64']).COLUMNS.TOLIST()
CATEGORICAL_COLS = DF.SELECT_DTYPES(INCLUDE=['OBJECT']).COLUMNS.TOLIST()

# SUMMARY STATISTICS FOR NUMERICAL COLUMNS
PRINT("NUMERICAL COLUMN STATISTICS:")
PRINT(DF[NUMERICAL_COLS].DESCRIBE().ROUND(2))
PRINT()

```

```

# KEY RISK METRICS
PRINT("\nKEY RISK METRICS:")
PRINT(F" MALWARE INCIDENTS: {(DF['MALWARE_DETECTION'] >= 1).MEAN():.1%} OF USERS")
PRINT(F" PHISHING ATTEMPTS: {(DF['PHISHING_ATTEMPTS'] >= 1).MEAN():.1%} OF USERS")
PRINT(F" CYBERBULLYING REPORTS: {(DF['CYBERBULLYING_REPORTS'] >= 1).MEAN():.1%}
OF USERS")
PRINT(F" AVERAGE HOURS ONLINE: {DF['HOURS_ONLINE'].MEAN():.2F} HOURS/DAY")
PRINT(F" VPN USAGE: {(DF['VPN_USAGE'] > 0).MEAN():.1%} OF USERS")
PRINT(F" PUBLIC NETWORK USAGE: {(DF['PUBLIC_NETWORK_USAGE'] > 0).MEAN():.1%}
OF USERS")

# AGE GROUP CORRELATIONS
PRINT("\nAGE GROUP CORRELATIONS:")
AGE_GROUPS = DF['AGE_GROUP'].UNIQUE()
FOR COL IN ['MALWARE_DETECTION', 'PHISHING_ATTEMPTS', 'CYBERBULLYING_REPORTS',
            'PASSWORD_STRENGTH', 'VPN_USAGE', 'E_SAFETY_AWARENESS_SCORE']:
    IF COL IN NUMERICAL_COLS:
        PRINT(F"\n {COL} BY AGE GROUP:")
        FOR AGE IN AGE_GROUPS:
            PRINT(F" {AGE}: {DF[DF['AGE_GROUP'] == AGE][COL].MEAN():.2F}")
    ELIF COL IN CATEGORICAL_COLS:
        PRINT(F"\n {COL} BY AGE GROUP:")
        FOR AGE IN AGE_GROUPS:
            DIST = DF[DF['AGE_GROUP'] == AGE][COL].VALUE_COUNTS(NORMALIZE=TRUE)
            PRINT(F" {AGE}:")
            FOR VAL, PCT IN DIST.ITEMS():
                PRINT(F" {VAL}: {PCT:.1%}")

RETURN "SUMMARY STATISTICS GENERATED SUCCESSFULLY."

SUMMARY = GENERATE_DATASET_SUMMARY("TEEN.CSV")
PRINT(SUMMARY)

```

APPENDIX D – CODE TO CLEAN DATA

```

DEF REPLACE_MISSING_VALUES (DATA) :
    """
    REPLACE COMMON MISSING VALUE INDICATORS WITH NAN IN THE DATAFRAME.
    PARAMETERS:
    DATA (DATAFRAME): THE DATAFRAME TO CLEAN.
    RETURNS:
    DATAFRAME: THE CLEANED DATAFRAME WITH NAN VALUES.
    """

    DATA.DESCRIBE ()
    MISSING_VALUES = ['N/A', 'NA', '', 'N/A', 'NONE', 'NONE', 'UNKNOWN']

    DATA = DATA.DROP(COLUMNS=['EDUCATION_CONTENT_USAGE'])
    DF_CLEAN = DATA.REPLACE(MISSING_VALUES, NP.NAN)
    RETURN DF_CLEAN

DEF CLEAN_DATES (DATE_STR) :
    """
    CLEAN DATES WITH ERROR HANDLING FOR VARIOUS FORMATS
    """

    IF PD.ISNA (DATE_STR) :
        RETURN NP.NAN

    # TRY COMMON DATE FORMATS
    FORMATS = [
        '%Y-%M-%D %H:%M:%S%Z',      # 2023-06-22 00:00:00+00:00
        '%Y-%M-%D',                # 2021-09-16
        '%M/%D/%Y',                # MM/DD/YYYY
        '%D/%M/%Y'                 # DD/MM/YYYY
    ]

    FOR FMT IN FORMATS:
        TRY:
            RETURN PD.TO_DATETIME (DATE_STR, FORMAT=FMT)
        EXCEPT (VALUEERROR, TYPEERROR) :
            CONTINUE

```

```

DEF SAFE_NUMERIC(VALUE):
    """CONVERT TO NUMERIC, HANDLING NON-NUMERIC CHARACTERS"""
    IF PD.ISNA(VALUE):
        RETURN NP.NAN

    IF ISINSTANCE(VALUE, (INT, FLOAT)):
        RETURN VALUE

    TRY:
        # REMOVE ANY NON-NUMERIC CHARACTERS EXCEPT DECIMAL POINTS
        CLEANED = RE.SUB(R'^0-9.', '', STR(VALUE))
        RETURN FLOAT(CLEANED) IF CLEANED ELSE NP.NAN
    EXCEPT (VALUEERROR, TYPEERROR):
        RETURN NP.NAN

DEF ADD_QUALITY_INDICATORS(DF):
    # CALCULATE COMPLETENESS SCORE FOR EACH ROW
    DF['COMPLETENESS_SCORE'] = DF.NOTNA().SUM(AXIS=1) / LEN(DF.COLUMNS)

    # FLAG ROWS WITH HIGH-RISK INDICATORS
    RISK_COLUMNS = [
        'MALWARE_DETECTION', 'PHISHING_ATTEMPTS', 'CYBERBULLYING_REPORTS',
        'RISKY_WEBSITE_VISITS', 'INSECURE_LOGIN_ATTEMPTS'
    ]
    VALID_RISK = [COL FOR COL IN RISK_COLUMNS IF COL IN DF.COLUMNS]

    IF VALID_RISK:
        DF['HAS_HIGH_RISK_INDICATORS'] = (DF[VALID_RISK] > 0).ANY(AXIS=1)

    RETURN DF

```

APPENDIX E – CODE FOR DATA ANALYSIS

```

IMPORT PANDAS AS PD
IMPORT NUMPY AS NP
IMPORT SEABORN AS SNS
IMPORT MATPLOTLIB.PYPLOT AS PLT
FROM SKLEARN.PREPROCESSING IMPORT ORDINALENCODER
FROM SCIPY.STATS IMPORT PEARSONR, SPEARMANR
FROM SKLEARN.CLUSTER IMPORT KMEANS
IMPORT MATPLOTLIB.PYPLOT AS PLT
FROM SKLEARN.PREPROCESSING IMPORT STANDARDSCALER
FROM SKLEARN.DECOMPOSITION IMPORT PCA

IN [ ]:
DEF GET_CLEANED_DATA():
    %RUN DATA_CLEANING.IPYNB
    FILE_PATH = '../ARCHIVE/TEEN_E_SAFETY_DATASET.CSV'
    RAW_DATA = PD.READ_CSV(FILE_PATH)
    CLEANED, SUMMARY = CLEAN_TEEN_SAFETY_DATA(RAW_DATA)
    RETURN CLEANED, SUMMARY

IN [ ]:
DEF ENCODE_SELECTED_CATEGORICAL_FEATURES(DF, CATEGORICAL_FEATURES):
    """
    ENCODE STRING-BASED AND ORDERED CATEGORICAL FEATURES USING ORDINAL ENCODING.
    """
    FEATURES_TO_ENCODE = [
        COL FOR COL IN CATEGORICAL_FEATURES IF COL IN DF.COLUMNS
    ]
    # CONVERT ANY CATEGORICAL TYPES TO STRING (OBJECT) SO ORDINALENCODER CAN
    # PROCESS THEM
    DF[FEATURES_TO_ENCODE] = DF[FEATURES_TO_ENCODE].ASTYPE(STR)

    ENCODER = ORDINALENCODER()
    DF[FEATURES_TO_ENCODE] = ENCODER.FIT_TRANSFORM(DF[FEATURES_TO_ENCODE])

    RETURN DF

```

```

IN [ ]:
DEF PLOT_HEATMAP(CORR_MATRIX):
    PRINT("=== DESCRIPTIVE STATISTICS ===")

    PLT.FIGURE(FIGSIZE=(40, 40))
    SNS.HEATMAP(
        CORR_MATRIX,
        ANNOT=TRUE,
        FMT=".2F",
        CMAP="COOLWARM",
        SQUARE=TRUE,
        ANNOT_KWS={"SIZE": 20} # FONT SIZE FOR ANNOTATIONS INSIDE THE HEATMAP
    )
    # INCREASE FONT SIZE FOR AXIS TICK LABELS
    PLT.XTICKS(FONTSIZE=20)
    PLT.YTICKS(FONTSIZE=20)

    # INCREASE FONT SIZE FOR THE TITLE
    PLT.TITLE("CORRELATION HEATMAP", FONTSIZE=30)

    PLT.TIGHT_LAYOUT()
    PLT.SHOW()

IN [ ]:
DEF GET_HIGH_CORRELATION_PAIRS(CORR_MATRIX, THRESHOLD):
    """
    RETURNS FEATURE PAIRS WITH ABSOLUTE CORRELATION GREATER THAN THE GIVEN THRESHOLD.
    ONLY CONSIDERS THE UPPER TRIANGLE OF THE CORRELATION MATRIX
    (EXCLUDING THE DIAGONAL).
    """
    HIGH_CORR_PAIRS = []
    COLS = CORR_MATRIX.COLUMNS

    FOR I IN RANGE(LEN(COLS)):
        FOR J IN RANGE(I + 1, LEN(COLS)):
            CORR_VALUE = CORR_MATRIX.ILOC[I, J]
            ACCEPTED = CORR_VALUE > THRESHOLD IF THRESHOLD > 0 ELSE
            CORR_VALUE < THRESHOLD
            IF ACCEPTED:
                HIGH_CORR_PAIRS.APPEND((COLS[I], COLS[J], ROUND(CORR_VALUE, 2)))

```

```

RETURN PD.DATAFRAME(HIGH_CORR_PAIRS, COLUMNS=['FEATURE_X',
'FEATURE_Y', 'CORRELATION'])

IN [ ]:
DEF GET_STATISTICS(DF, FEATURE_X, FEATURE_Y, CATEGORICAL_FEATURES):
    X_IS_CAT = FEATURE_X IN CATEGORICAL_FEATURES
    Y_IS_CAT = FEATURE_Y IN CATEGORICAL_FEATURES

    # DROP MISSING VALUES FOR BOTH
    SUB_DF = DF[[FEATURE_X, FEATURE_Y]].DROPNAN()

    # IF BOTH CATEGORICAL → INVALID FOR PEARSON/SPEARMAN
    IF X_IS_CAT AND Y_IS_CAT:
        RETURN {
            "METHOD": "INVALID (CATEGORICAL-CATEGORICAL)",
            "CORRELATION": NONE,
            "P_VALUE": NONE
        }

    CORR, P = SPEARMANR(SUB_DF[FEATURE_X], SUB_DF[FEATURE_Y])
    RETURN {
        "METHOD": "SPEARMAN",
        "CORRELATION": CORR,
        "P_VALUE": FLOAT(F"{P:.2E}")
    }

IN [ ]:
DEF GET_STATISTICALLY_SIGNIFICANT_CORRELATIONS(DF, CORRELATION_PAIR,
CATEGORICAL_FEATURES):
    RESULTS = []

    FOR _, ROW IN CORRELATION_PAIR.ITERROWS():
        COL1 = ROW["FEATURE_X"]
        COL2 = ROW["FEATURE_Y"]

        RESULT = GET_STATISTICS(DF, COL1, COL2, CATEGORICAL_FEATURES)
        # DF IS YOUR MAIN DATASET

```

```

RESULTS.APPEND({
    "FEATURE_X": COL1,
    "FEATURE_Y": COL2,
    "METHOD": RESULT["METHOD"],
    "CORRELATION": ROUND(RESULT["CORRELATION"], 2) IF RESULT["CORRELATION"]
        IS NOT NONE ELSE NONE,
    "P-VALUE": RESULT["P_VALUE"]
})

# CONVERT TO DATAFRAME TO VIEW OR ANALYZE
VALIDATED_CORR_DF = PD.DATAFRAME (RESULTS)

# OPTIONAL: SHOW ONLY STATISTICALLY SIGNIFICANT ONES
SIGNIFICANT = VALIDATED_CORR_DF[
    (VALIDATED_CORR_DF["CORRELATION"].ABS() > 0.3) &
    (VALIDATED_CORR_DF["P-VALUE"] < 0.05)
]

RETURN SIGNIFICANT

IN [ ]:
# THIS TAKES A WHILE TO RUN, SEPARATE IT FROM THE REST OF THE CODE
CLEANED, SUMMARY = GET_CLEANED_DATA()

IN [ ]:
CATEGORICAL_FEATURES = [
    'USAGE_PATTERN_CATEGORY',
    'TIME_PERIOD_RISK',
    'PARENTAL_OVERSIGHT_LEVEL',
    'SECURITY_POSTURE_CATEGORY',
    'PROTECTION_MATCH_CATEGORY'
]

DERIVED_FEATURE_SUBSET = [
    "TIME_PERIOD_RISK",
    "USAGE_PATTERN_CATEGORY",
    "RISKY_BROWSING_RATIO",
    "DIGITAL_CONSUMPTION_INDEX",
    "DEVICE_SECURITY_INDEX",
    "CONNECTION_SAFETY_SCORE",

```

```

    "TOTAL_THREAT_EXPOSURE",
    "THREAT_SEVERITY_INDEX",
    "DEFENSIVE_POSTURE_SCORE",
    "PASSWORD_SECURITY_SCORE",
    "PARENTAL_OVERSIGHT_LEVEL",
    "SUPERVISION_EFFECTIVENESS",
    "AWARENESS_BEHAVIOR_GAP",
    "SAFETY_BEHAVIOR_SCORE",
    "OVERALL_SECURITY_POSTURE",
    "SECURITY_POSTURE_CATEGORY",
    "BEHAVIOR_VS_PROTECTION_GAP",
    "PROTECTION_MATCH_CATEGORY"
]
ENCODE_SELECTED_CATEGORICAL_FEATURES (CLEANED, CATEGORICAL_FEATURES)

# # ALL DERIVED CATEGORICAL FEATURES ARE NOW ENCODED
# FOR COL IN CATEGORICAL_FEATURES:
#     IF COL IN CLEANED.COLUMNS:
#         PRINT(F"UNIQUE VALUES IN '{COL}':")
#         PRINT(CLEANED[COL].UNIQUE())
#         PRINT()

# GET CORRELATION AND P-VALUE FOR POSITIVE CORRELATIONS FROM HEATMAP
CORR_MATRIX = CLEANED[DERIVED_FEATURE_SUBSET].CORR(METHOD='SPEARMAN')
PLOT_HEATMAP (CORR_MATRIX)
POSITIVE_CORRELATIONS = GET_HIGH_CORRELATION_PAIRS (CORR_MATRIX, THRESHOLD=0.5)
POSITIVE_SIGNIFICANT = GET_STATISTICALLY_SIGNIFICANT_CORRELATIONS (CLEANED,
POSITIVE_CORRELATIONS, CATEGORICAL_FEATURES)
FOR _, ROW IN POSITIVE_SIGNIFICANT.ITERROWS():
    COL1 = ROW["FEATURE_X"]
    COL2 = ROW["FEATURE_Y"]
    SNS.PAIRPLOT (CLEANED[[COL1, COL2]])

PRINT (POSITIVE_SIGNIFICANT)

# GET CORRELATION AND P-VALUE FOR NEGATIVE CORRELATIONS FROM HEATMAP
NEGATIVE_CORRELATIONS = GET_HIGH_CORRELATION_PAIRS (CORR_MATRIX, THRESHOLD=-0.5)
NEGATIVE_SIGNIFICANT = GET_STATISTICALLY_SIGNIFICANT_CORRELATIONS (CLEANED,
NEGATIVE_CORRELATIONS, CATEGORICAL_FEATURES)

```

```

FOR _, ROW IN NEGATIVE_SIGNIFICANT.ITERROWS():
    COL1 = ROW["FEATURE_X"]
    COL2 = ROW["FEATURE_Y"]
    SNS.PAIRPLOT(CLEANED[[COL1, COL2]])

PRINT(NEGATIVE_SIGNIFICANT)

IN [ ]:
BEHAVIORAL_FEATURES = [
    "DIGITAL_CONSUMPTION_INDEX",
    "DEVICE_SECURITY_INDEX",
    "CONNECTION_SAFETY_SCORE",
    "TOTAL_THREAT_EXPOSURE",
    "THREAT_SEVERITY_INDEX",
    "DEFENSIVE_POSTURE_SCORE",
    "PASSWORD_SECURITY_SCORE",
    "SUPERVISION_EFFECTIVENESS",
    "SAFETY_BEHAVIOR_SCORE",
    "OVERALL_SECURITY_POSTURE",
    "AGE_GROUP_ENCODED"
]

DF_CLUSTER = CLEANED[BEHAVIORAL_FEATURES].DROPNA() # DROP MISSING ROWS

# KMEANS CLUSTERING
SCALER = STANDARDSCALER()
SCALED_DATA = SCALER.FIT_TRANSFORM(DF_CLUSTER)
KMEANS = KMEANS(N_CLUSTERS=3, RANDOM_STATE=42)
CLUSTER_LABELS = KMEANS.FIT_PREDICT(SCALED_DATA)
DF_CLUSTER['CLUSTER'] = CLUSTER_LABELS

# AVERAGE SCORES PER CLUSTER
CLUSTER_SUMMARY = DF_CLUSTER.GROUPBY("CLUSTER").MEAN()

# HEATMAP TO VISUALIZE PATTERNS
PLT.FIGURE(FIGSIZE=(10, 6))
SNS.HEATMAP(CLUSTER_SUMMARY, CMAP="COOLWARM", ANNOT=TRUE, FMT=".2F")
PLT.TITLE("CLUSTER BEHAVIOR PROFILES")
PLT.SHOW()

```

```

# SHOW THE DISTRIBUTION OF AGE GROUPS PER CLUSTER WITH CUSTOM LABELS
SNS.COUNTPLOT(DATA=DF_CLUSTER, X='AGE_GROUP_ENCODED', HUE='CLUSTER')
PLT.TITLE("AGE GROUP DISTRIBUTION PER CLUSTER")

# REPLACE NUMERIC CODES WITH AGE GROUP LABELS
PLT.XTICKS(TICKS=[0, 1, 2], LABELS=["<13", "13-16", "17-19"])
PLT.XLABEL("AGE GROUP")
PLT.SHOW()

# PCA
PCA = PCA(N_COMPONENTS=2)
COMPONENTS = PCA.FIT_TRANSFORM(SCALED_DATA)

DF_CLUSTER['PC1'] = COMPONENTS[:, 0]
DF_CLUSTER['PC2'] = COMPONENTS[:, 1]

SNS.SCATTERPLOT(DATA=DF_CLUSTER, X="PC1", Y="PC2", HUE="CLUSTER", PALETTE="TAB10")
PLT.TITLE("TEEN BEHAVIOR CLUSTERS (PCA)")
PLT.SHOW()

# CLUSTER 0 (RISK-AWARE, CAUTIOUS USERS)
# CLUSTER 1 (LOW ENGAGEMENT / UNEXPOSED)
# CLUSTER 2 (MODERATE USERS WITH SOME THREATS)

```