Problematic Internet Use: Mental Health Detection Based on Physical Activity

Child Mind Institute

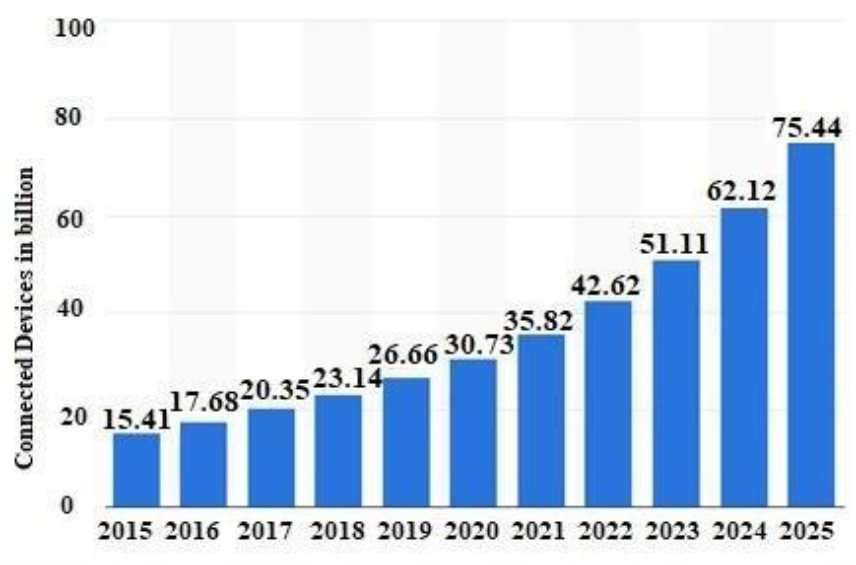
Capstone Proposal

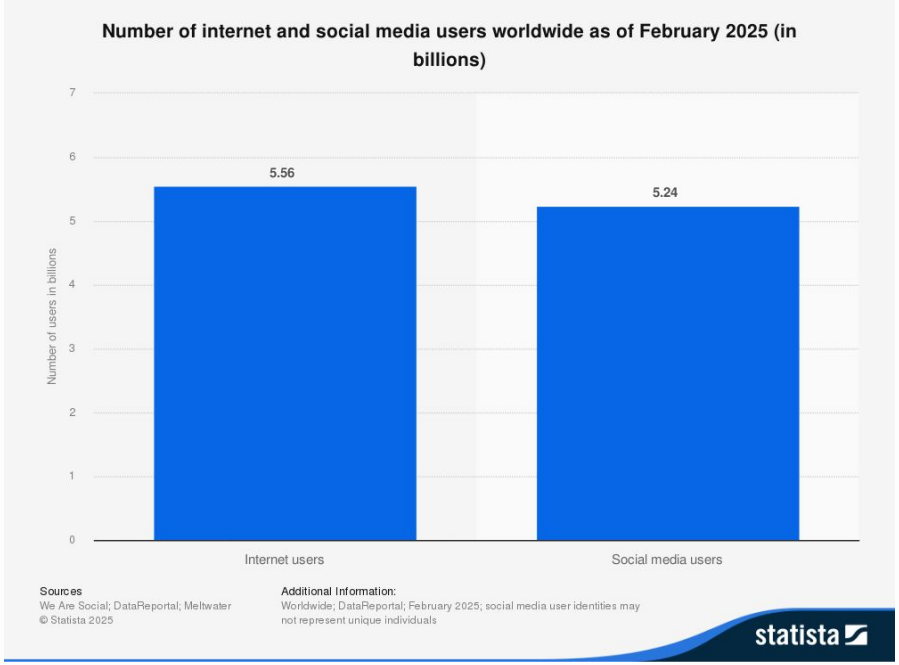
By: Daniel Epstein

Email: daniel.tepstein@gmail.com

THE Community TECH

What is Problematic Internet Use





https://www.aacap.org/AACAP/Families_and_Youth/Facts_for_Families/FFF-Guide/Children-And-Watching-TV-054.aspx?hl=en-CA#:~:text=Children%20and%20adolescents'
0spend%20a,day%20watching%20or%20using%20screens.

What is Problematic Internet Use

93%

Children had access to the Internet through a computer

97%

Children had home internet access

95%

Teens reported having access to a smartphone

https://www.commonsensemedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens-2021

What is Problematic Internet Use, Really?

Problematic Internet Use (PIU) is defined as:

"*A compulsive need to spend a great deal of time on the Internet, to the point where relationships, work, and health are allowed to suffer*."

Dr. Kimberly Young
Pioneer of Internet Addiction Research
https://netaddiction.com/

THE Community TECH

https://academic.oup.com/book/1235/chapter/140127088

Why it is hard to diagnose

- The effects of PIU can be hard to diagnose
  - Often they aren't noticed until it is too late
  - “Symptoms “ appear in the form of depression, lethargy and other hard to identify symptoms and even when those symptoms do show up they are often mistaken for other issues and treated with those issues in mind.

- Additionally, the testing criteria to identify Piu or the way its manifested are not standardized
- **Diagnosis is difficult:** Traditional assessments are complex, which means PIU often goes undiagnosed.
- **Too late:** We often only see the effects of PIU—behavioral and emotional issues—after the problem has already progressed.

The Answer: “Physical Tests”

- Physical activity and fitness measures
  - Easily captured through wearable devices and routine health assessments
  - Require minimal clinical oversight
  - Widely applicable across diverse populations.

THE Community TECH

Enter the HBN and CMI

The Healthy Brain Network (HBN)

- Publicly and privately funded research study
- Transforming the way scientists study mental health
- How families access treatment and support

Child Mind Institute

Goal

- Discover **biological markers** that support more objective, data-driven approaches to diagnosing and treating **psychiatric and cognitive conditions**.

Data

- HBN created a Kaggle competition to develop a predictive model capable of analyzing children's physical activity data to detect early indicators of problematic internet and technology use.

The Data Set

The data set provided by HBN included demographic information (age and sex) as well as a number of physical measures and the results of 5 physical assessments and questionnaires.

1. CGAS
2. The FitnessGram
3. Bio-electric Impedance Analysis
4. A Physical Assessment Questionnaire for children and teens
5. Sleep disturbance scale
6. A parent-child internet addiction test
7. A questionnaire around total hours of daily internet use

The Data Set

- The **Children's Global Assessment Scale** (CGAS) is a measure developed by Schaffer and colleagues at the Department of Psychiatry, Columbia University, to provide a global measure of the level of functioning in children and adolescents.
- **FitnessGram** is a comprehensive fitness-assessment battery for youth. It includes a variety of health-related physical fitness tests designed to assess aerobic capacity, muscular strength, muscular endurance, flexibility, and body composition.

The Data Set

- **Bioelectrical impedance analysis (BIA)** is a method for estimating body composition, in particular body fat and muscle mass.
- The **Physical Activity Questionnaire for Older Children (PAQ-C)** and the **Physical Activity Questionnaire for Adolescents (PAQ-A)** provide a general measure of physical activity for youth from grades 4-12 (approximately ages 8-20).
- The **Sleep Disturbance Scale for Children (SDSC)** evaluates various sleep disorders, including difficulties in falling asleep and staying asleep, sleep-disordered breathing, arousal disorders, sleep-wake transition disorders, excessive sleepiness, and excessive sweating during sleep (hyperhidrosis).

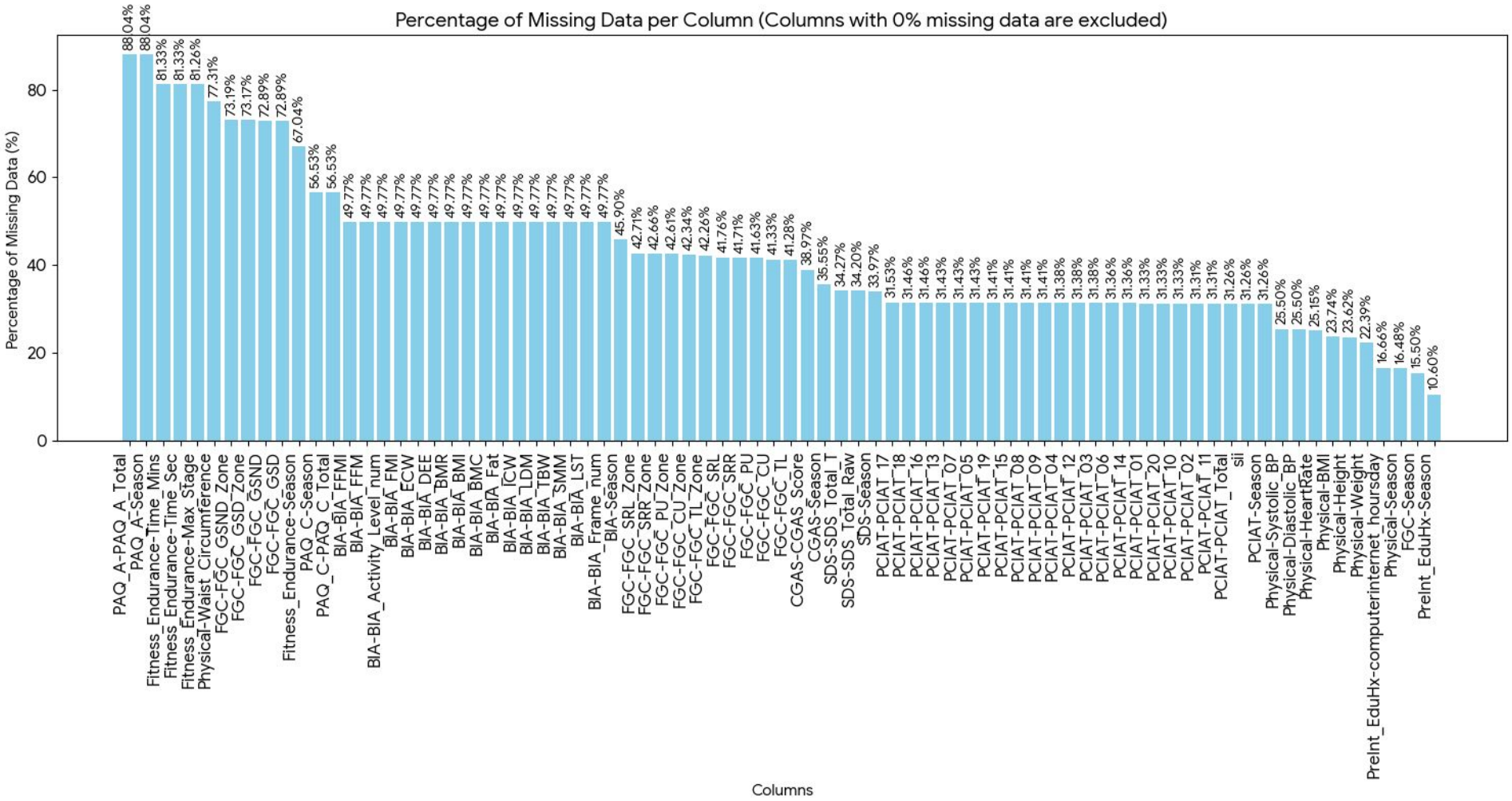
The Data Set

- The **Parent Child Internet Addiction** test is a series of questions about the children's internet habits and moods as they relate to internet use. Each question is scored 0 – 5, after which all twenty questions are added up, and the total is scored on a scale from 0 – 100. **Scores from 0 – 30 having no impairment, 31-49 having a mild impairment, 50 – 79 having moderate impairment, and 80+ having severe Impairment.**
- The **Questionnaire around total hours of internet use** measures internet use in hours. 0 meaning no hours, 1 meaning 1 hour, 2 meaning 2 hours, and 3 meaning 3 or more hours.

Objective and PS

Objective: To develop predictive models that leverage multimodal data from the Healthy Brain Network (HBN) to improve the identification of mental health and learning disorders in youth caused by PIU.

Problem Statement: How can data science be used to analyze data from physical activity and predict problematic internet use and mental health issues?

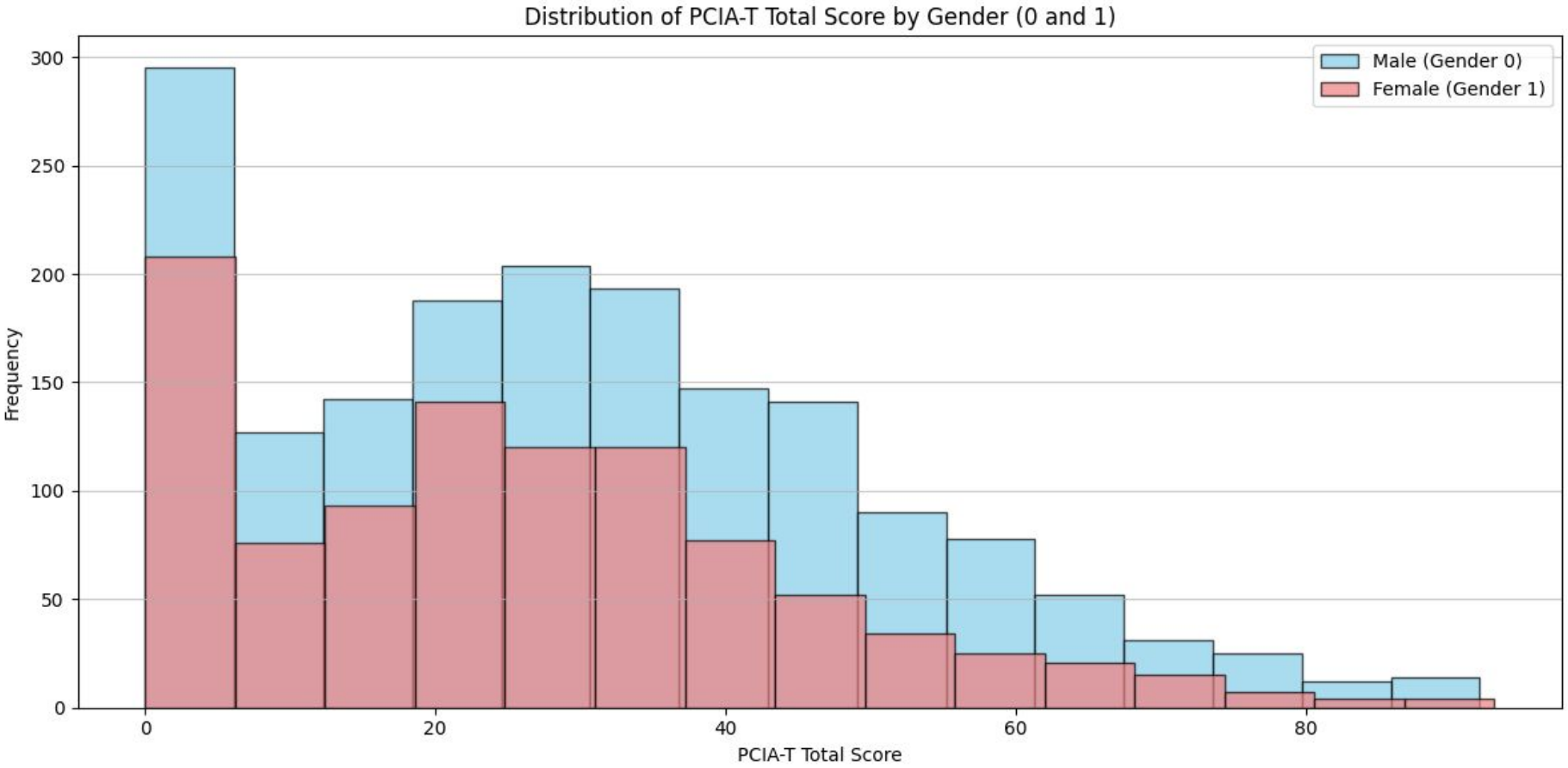

Percentage of Missing Data per Column (Columns with 0% missing data are excluded)
Percentage of Missing Data (%)
0
20
40
60
80
88.04%
88.04%
81.33%
81.33%
81.26%
77.31%
73.19%
73.17%
72.89%
72.89%
67.04%
56.53%
56.53%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
49.77%
45.90%
42.71%
42.66%
42.61%
42.34%
42.26%
41.76%
41.71%
41.63%
41.33%
41.28%
38.97%
35.55%
34.27%
34.20%
33.97%
31.53%
31.46%
31.46%
31.43%
31.43%
31.43%
31.41%
31.41%
31.41%
31.41%
31.41%
31.38%
31.38%
31.38%
31.36%
31.36%
31.36%
31.33%
31.33%
31.33%
31.31%
31.31%
31.26%
31.26%
31.26%
25.50%
25.50%
25.15%
23.74%
23.62%
22.39%
16.66%
16.48%
15.50%
10.60%
PAQ_A-PAQ_A_Total
PAQ_A-Season
Fitness_Endurance-Time_Mins
Fitness_Endurance-Time_Sec
Fitness_Endurance-Max_Stage
Physical-Waist_Circumference
FGC-FGC_GSND_Zone
FGC-FGC_GSD_Zone
FGC-FGC_GSND
FGC-FGC_GSD
Fitness_Endurance-Season
PAQ_C-Season
PAQ_C-PAQ_C_Total
BIA-BIA_FFMI
BIA-BIA_FFM
BIA-BIA_Activity_Level_num
BIA-BIA_FMI
BIA-BIA_ECW
BIA-BIA_DEE
BIA-BIA_BMR
BIA-BIA_BMI
BIA-BIA_BMC
BIA-BIA_Fat
BIA-BIA_ICW
BIA-BIA_LDM
BIA-BIA_TBW
BIA-BIA_SMM
BIA-BIA_LST
BIA-BIA_Frame_num
BIA-Season
FGC-FGC_SRL_Zone
FGC-FGC_SRR_Zone
FGC-FGC_PU_Zone
FGC-FGC_CU_Zone
FGC-FGC_TL_Zone
FGC-FGC_SRL
FGC-FGC_SRR
FGC-FGC_PU
FGC-FGC_CU
FGC-FGC_TL
CGAS-CGAS_Score
CGAS-Season
SDS-SDS_Total_T
SDS-SDS_Total_Raw
SDS-Season
PCIAT-PCIAT_17
PCIAT-PCIAT_18
PCIAT-PCIAT_16
PCIAT-PCIAT_13
PCIAT-PCIAT_07
PCIAT-PCIAT_05
PCIAT-PCIAT_19
PCIAT-PCIAT_15
PCIAT-PCIAT_08
PCIAT-PCIAT_09
PCIAT-PCIAT_04
PCIAT-PCIAT_12
PCIAT-PCIAT_03
PCIAT-PCIAT_06
PCIAT-PCIAT_14
PCIAT-PCIAT_01
PCIAT-PCIAT_20
PCIAT-PCIAT_10
PCIAT-PCIAT_02
PCIAT-PCIAT_11
PCIAT-PCIAT_Total
sii
PCIAT-Season
Physical-Systolic_BP
Physical-Diastolic_BP
Physical-HeartRate
Physical-BMI
Physical-Height
Physical-Weight
PreInt_EduHx-computerinternet_hoursday
Physical-Season
FGC-Season
PreInt_EduHx-Season
Columns

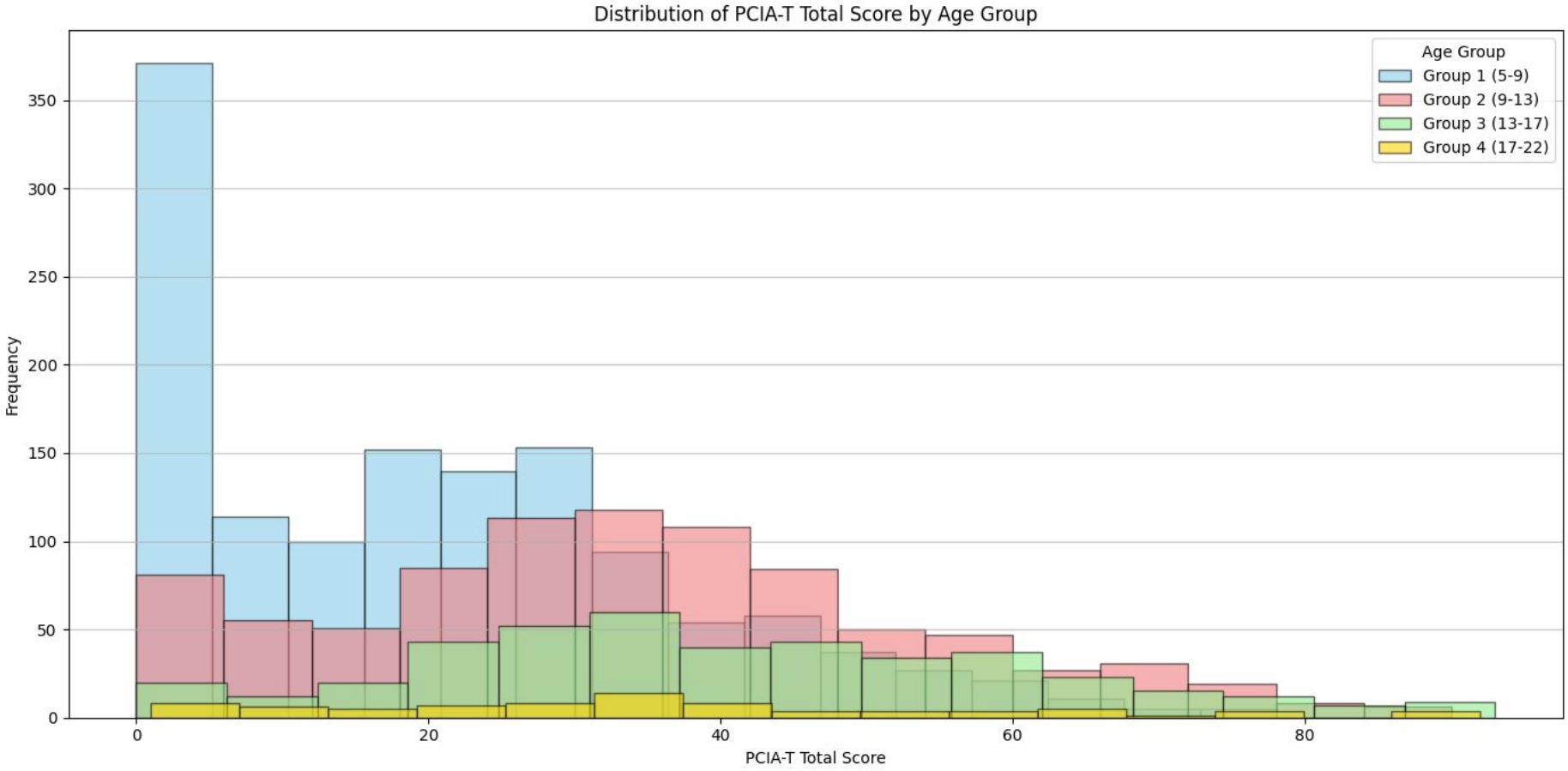
Methodology

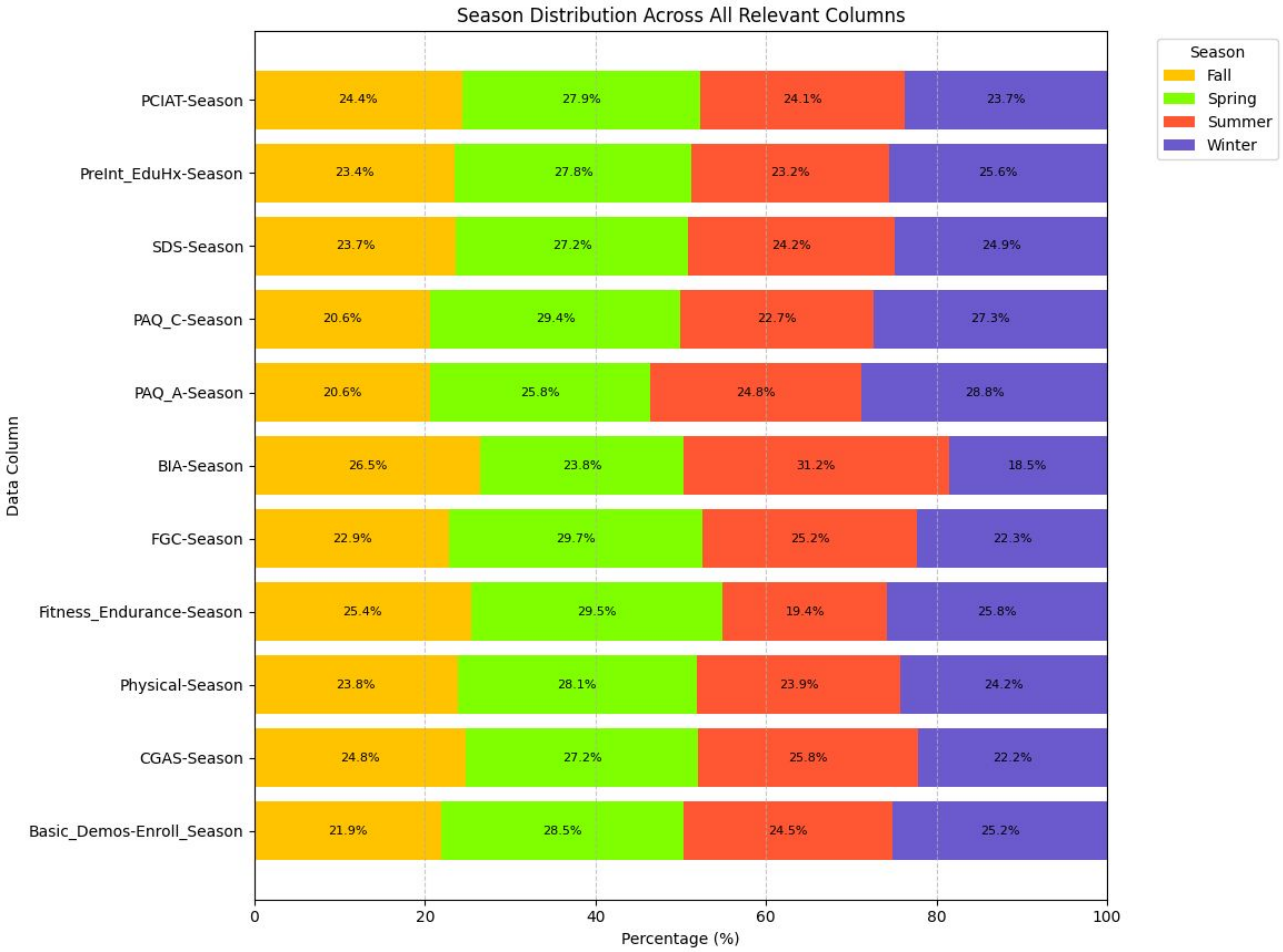
Title: Predicting High Psychological Capital and Impulsive Action Test (PCIAT) Scores

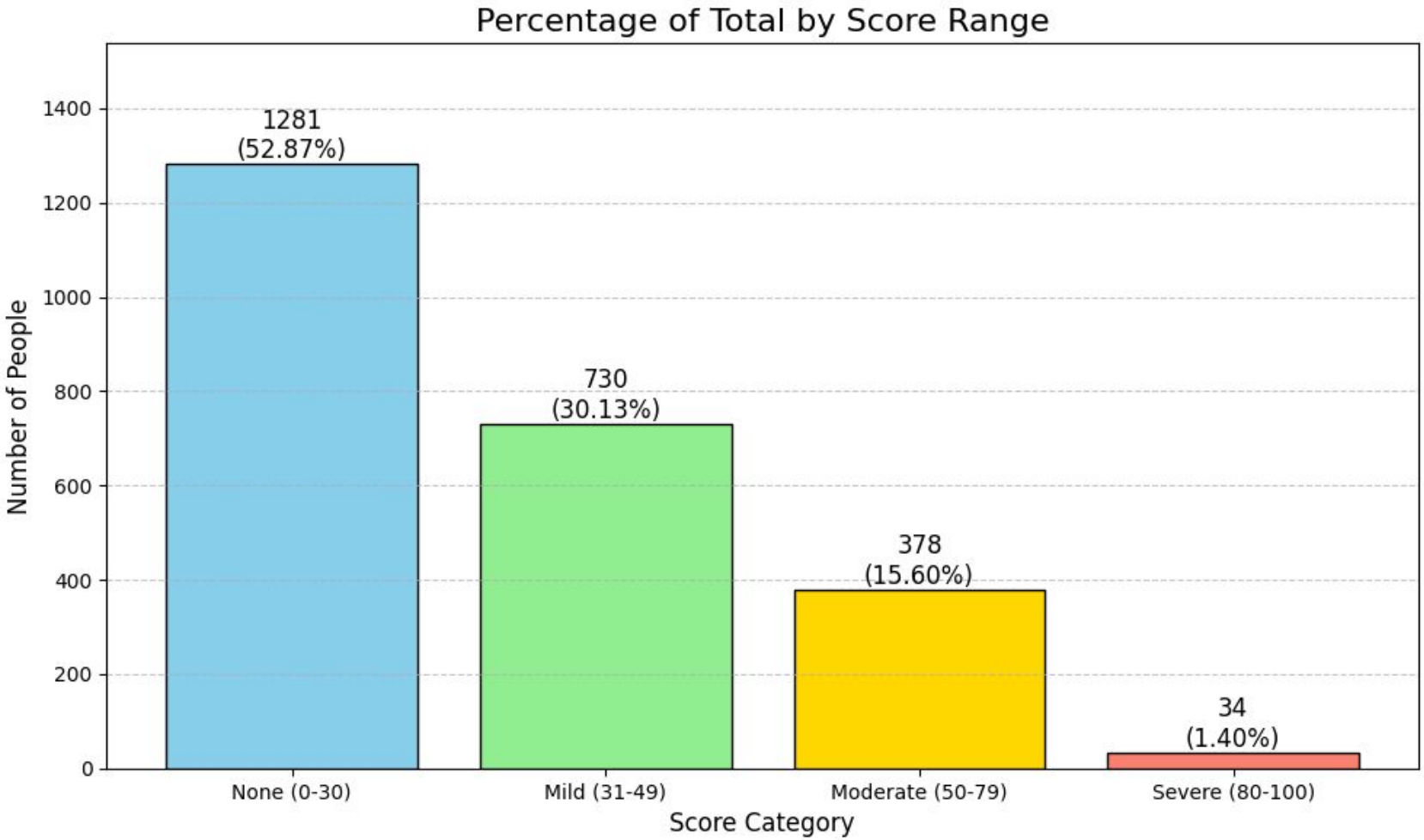
Purpose: Our study aimed to identify the key behavioral and physiological factors that predict a high score on the PCIAT.

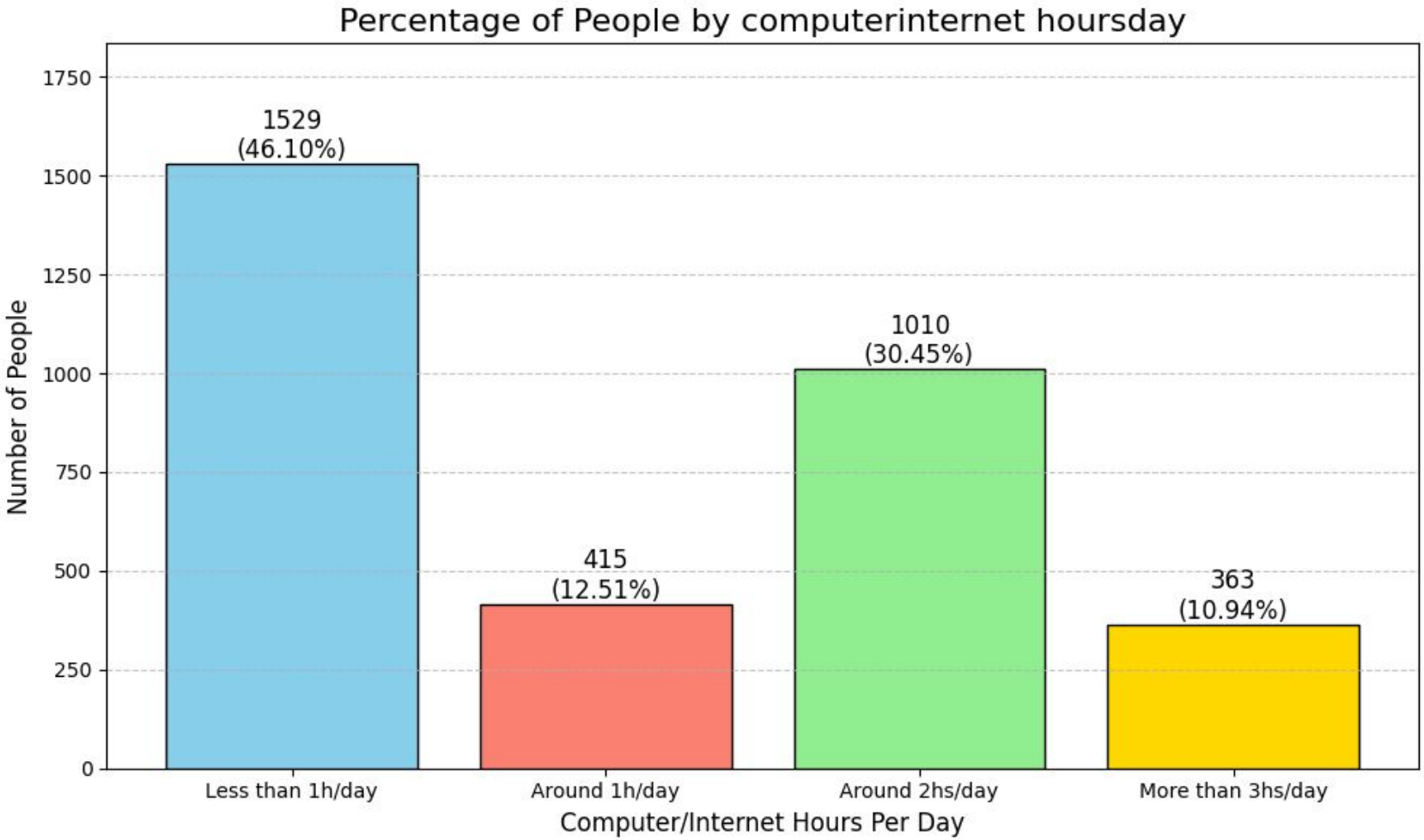
Dependent Variable: We created a binary dependent variable called is_pciat_high. This was a 1 if the raw PCIAT score was over 80 and a 0 otherwise.


Distribution of PCIA-T Total Score by Gender (0 and 1)
Male (Gender 0)
Female (Gender 1)
Frequency
PCIA-T Total Score
0
50
100
150
200
250
300
0
20
40
60
80

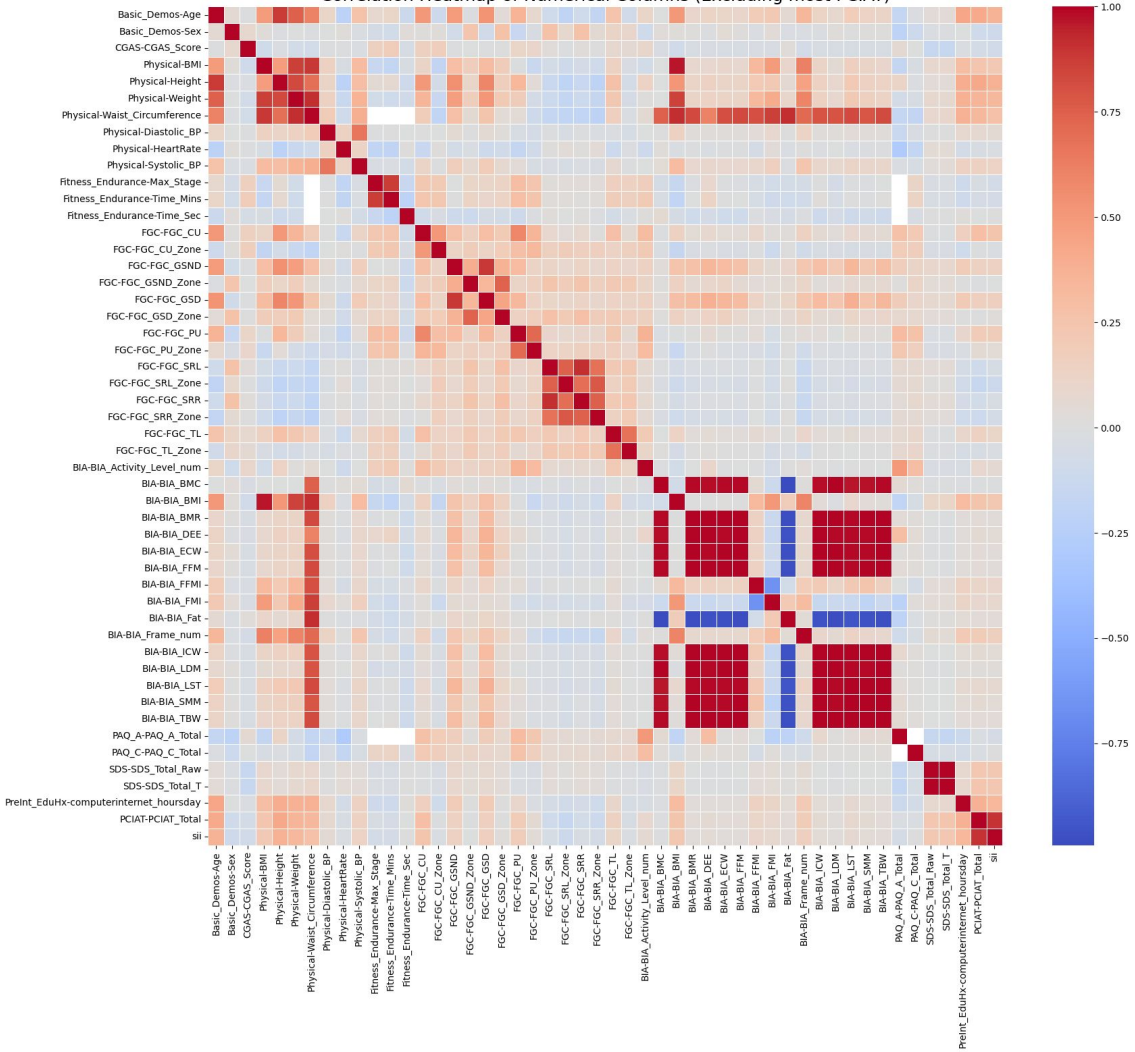

Distribution of PCIA-T Total Score by Age Group
Age Group
Group 1 (5-9)
Group 2 (9-13)
Group 3 (13-17)
Group 4 (17-22)
Frequency
0
50
100
150
200
250
300
350
0
20
40
60
80
PCIA-T Total Score

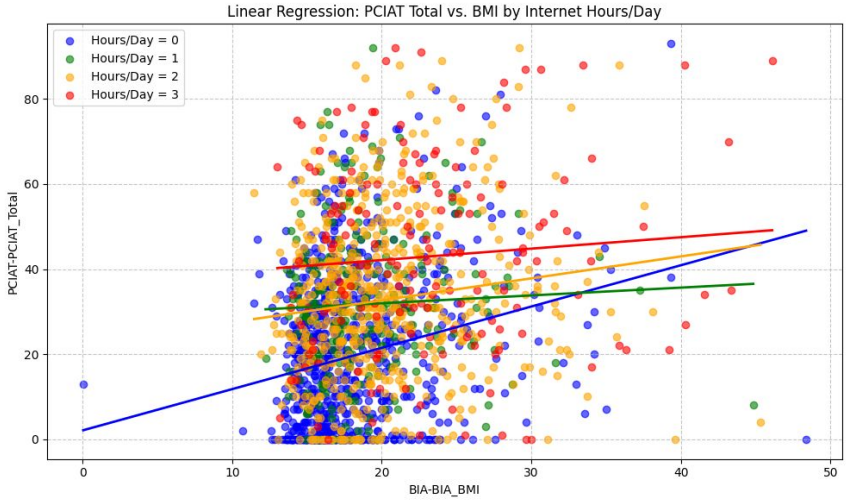

Season Distribution Across All Relevant Columns
Season
Fall
Spring
Summer
Winter
PCIAT-Season
24.4%
27.9%
24.1%
23.7%
PreInt_EduHx-Season
23.4%
27.8%
23.2%
25.6%
SDS-Season
23.7%
27.2%
24.2%
24.9%
PAQ_C-Season
20.6%
29.4%
22.7%
27.3%
PAQ_A-Season
20.6%
25.8%
24.8%
28.8%
BIA-Season
26.5%
23.8%
31.2%
18.5%
FGC-Season
22.9%
29.7%
25.2%
22.3%
Fitness_Endurance-Season
25.4%
29.5%
19.4%
25.8%
Physical-Season
23.8%
28.1%
23.9%
24.2%
CGAS-Season
24.8%
27.2%
25.8%
22.2%
Basic_Demos-Enroll_Season
21.9%
28.5%
24.5%
25.2%
Data Column
0
20
40
60
80
100
Percentage (%)

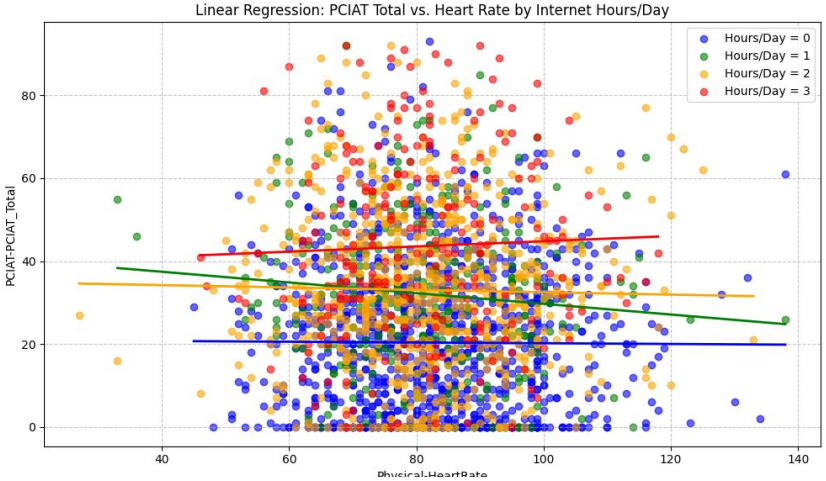

Percentage of Total by Score Range
Number of People
Score Category
1281
(52.87%)
730
(30.13%)
378
(15.60%)
34
(1.40%)
None (0-30)
Mild (31-49)
Moderate (50-79)
Severe (80-100)
0
200
400
600
800
1000
1200
1400

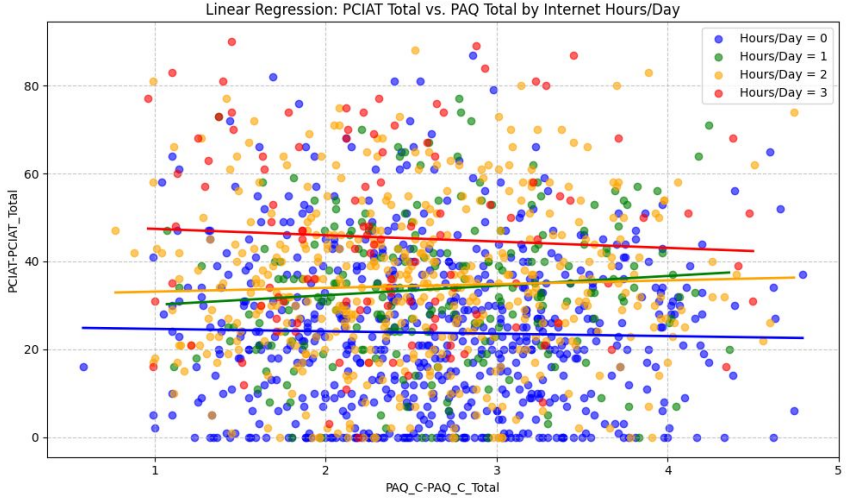

Percentage of People by computerinternet hoursday
1750
1500
1250
1000
750
500
250
0
Number of People
1529
(46.10%)
415
(12.51%)
1010
(30.45%)
363
(10.94%)
Less than 1h/day
Around 1h/day
Around 2hs/day
More than 3hs/day
Computer/Internet Hours Per Day

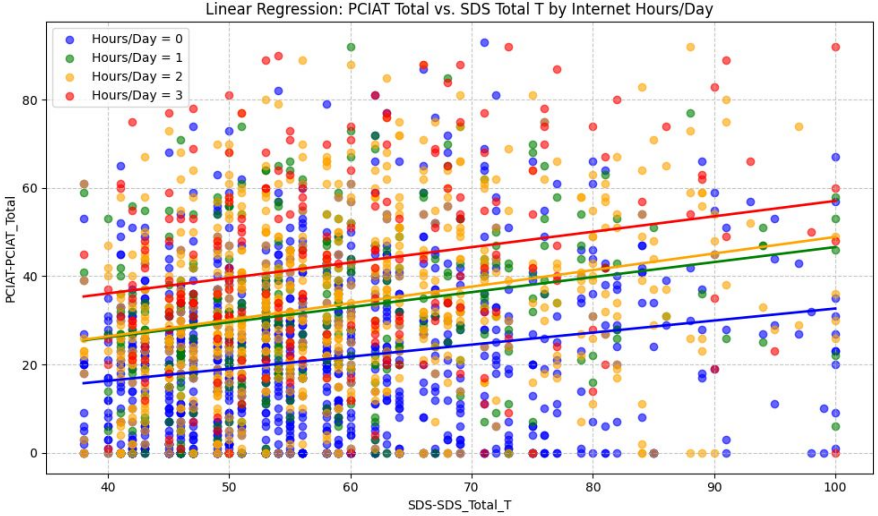
Correlation Heatmap of Numerical Columns (Excluding most PCIAT)

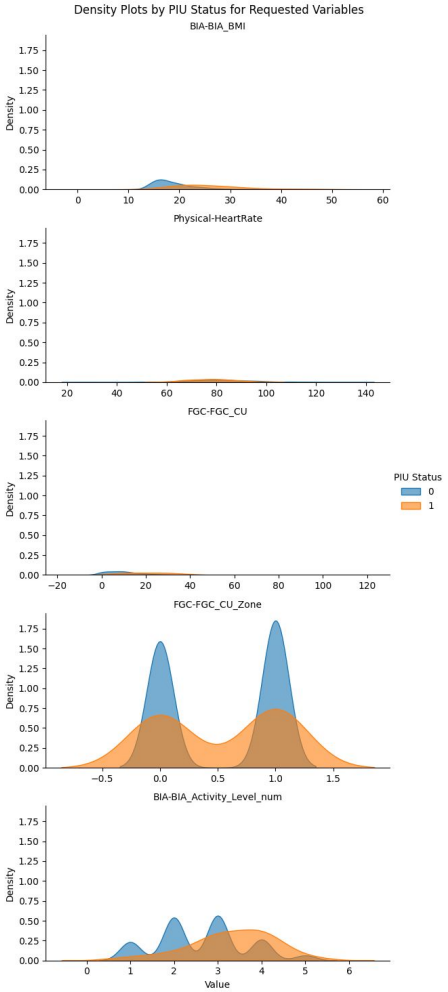


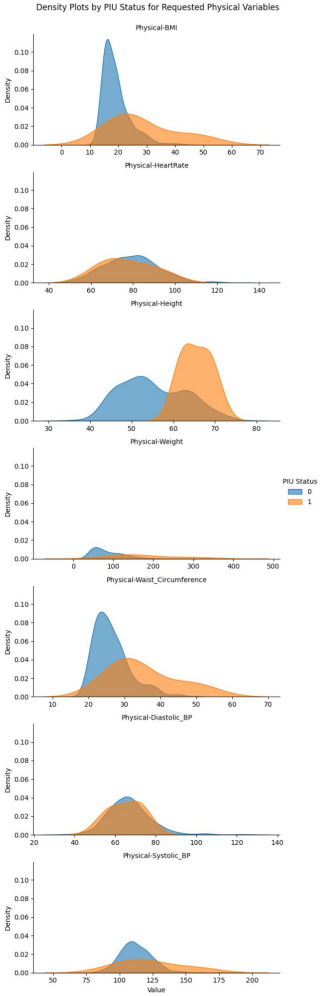

Linear Regression: PCIAT Total vs. BMI by Internet Hours/Day
Hours/Day = 0
Hours/Day = 1
Hours/Day = 2
Hours/Day = 3
PCIAT-PCIAT_Total
BIA-BIA_BMI

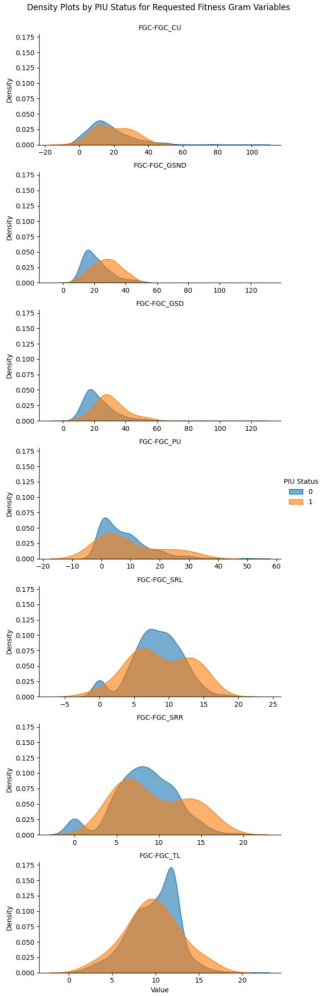

Linear Regression: PCIAT Total vs. Heart Rate by Internet Hours/Day
Hours/Day = 0
Hours/Day = 1
Hours/Day = 2
Hours/Day = 3
PCIAT-PCIAT_Total

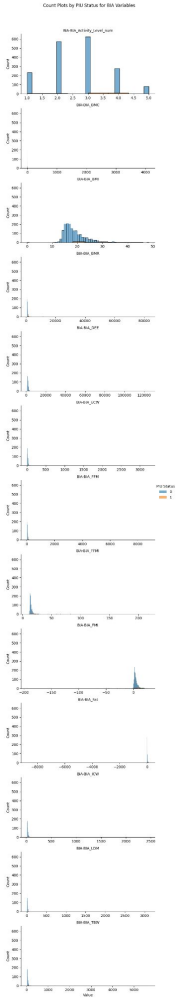

Linear Regression: PCIAT Total vs. PAQ Total by Internet Hours/Day
Hours/Day = 0
Hours/Day = 1
Hours/Day = 2
Hours/Day = 3
PCIAT-PCIAT_Total
PAQ_C-PAQ_C_Total


Linear Regression: PCIAT Total vs. SDS Total T by Internet Hours/Day
Hours/Day = 0
Hours/Day = 1
Hours/Day = 2
Hours/Day = 3
PCIAT-PCIAT_Total
SDS-SDS_Total_T


Density Plots by PIU Status for Requested Variables
BIA-BIA_BMI
Physical-HeartRate
FGC-FGC_CU
FGC-FGC_CU_Zone
BIA-BIA_Activity_Level_num
Density
Value
PIU Status
0
1


Density Plots by PIU Status for Requested Physical Variables
Physical-BMI
Physical-HeartRate
Physical-Height
Physical-Weight
Physical-Waist_Circumference
Physical-Diastolic_BP
Physical-Systolic_BP
Density
Value
PIU Status
0
1

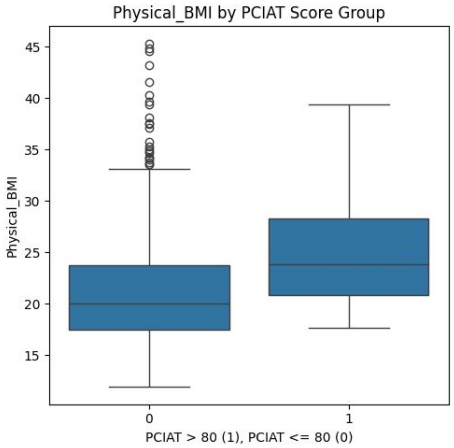

Density Plots by PIU Status for Requested Fitness Gram Variables
FGC-FGC_CU
FGC-FGC_GSND
FGC-FGC_GSD
FGC-FGC_PU
FGC-FGC_SRL
FGC-FGC_SRR
FGC-FGC_TL
Density
Value
PIU Status
0
1

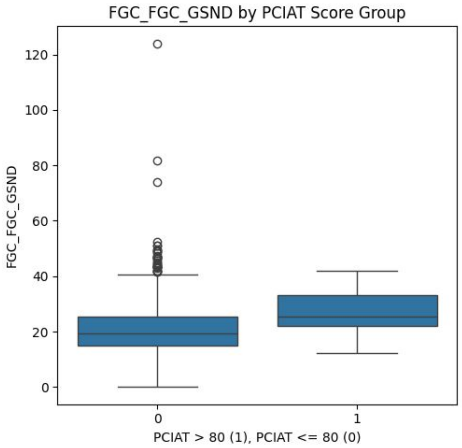

Count Plots by PIU Status for BIA Variables
Count
Value

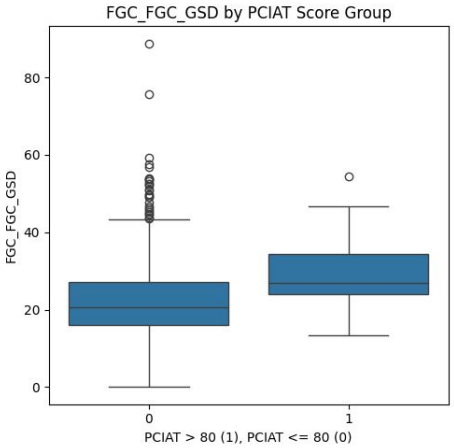
Key Findings

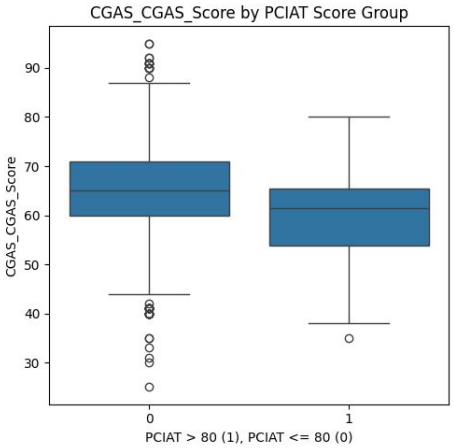
```
Optimization terminated successfully.
         Current function value: 0.115917
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:          is_pciat_high   No. Observations:                  722
Model:                          Logit   Df Residuals:                      717
Method:                           MLE   Df Model:                            4
Date:                Sun, 17 Aug 2025   Pseudo R-squ.:                 0.08479
Time:                        20:35:09   Log-Likelihood:                -83.692
converged:                       True   LL-Null:                       -91.446
Covariance Type:            nonrobust   LLR p-value:                 0.003756
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept            -4.0070      1.530     -2.619      0.009      -7.006      -1.008
Physical_BMI          0.0739      0.035      2.111      0.035       0.005       0.143
FGC_FGC_GSND         -0.0077      0.050     -0.156      0.876      -0.105       0.090
FGC_FGC_GSD           0.0456      0.045      1.008      0.313      -0.043       0.134
CGAS_CGAS_Score      -0.0360      0.019     -1.850      0.064      -0.074       0.002
=====================================================================================

==================================================
```

Key Findings









Summary

From the data provided I was able to find 4 variables that showed significant difference between those with PIU and those without.

Those variables can be used to identify PIU among children.

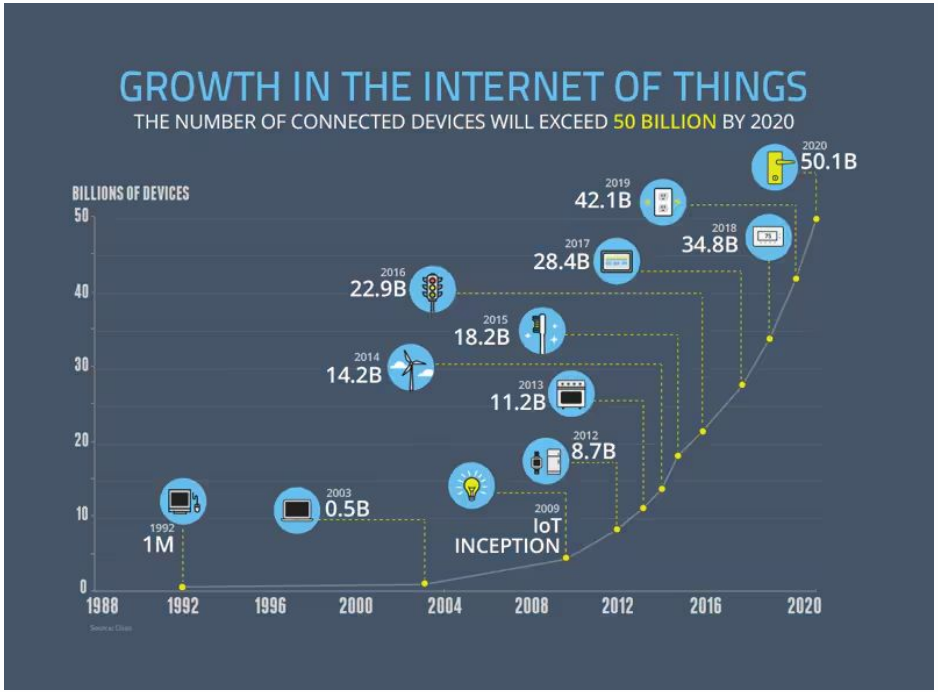
However,
The Data set was missing a lot of data and even amongst the variables I found the pseudo R squared value was only 8.5%, relatively low.

FINAL PRESENTATION

- Take your final report and convert it to presentation form.
- Make your talking points flow from one step slide to the next.
- 20-minute speaking limit. 5-10 minutes of Q&A.
- **Remember:** you are talking to the audience. Make sure your presentation is prepared in a manner that they can follow.
- **Tip:** Leverage visuals to reinforce key points about your analysis.
- **Tip:** If you do create a power BI dashboard, provide a brief demonstration and share the link for others to experiment with as well.
- **Recommended reading:** https://chemicalstatistician.wordpress.com/2017/05/16/a-comprehensive-guide-for-public-speaking-at-scientific-conferences/

Background

Over the last few years with the rapid development of the internet, access to the internet and generational gaps in the understanding of the internet, Problematic Internet Use, PUI, has become a serious issue.



https://www.weforum.org/stories/2016/04/this-is-what-the-internet-looked-like-when-it-was-first-invented/