EDA

Team

2024-12-08

Loading our Data

```
library(readr)
train <- read_csv("~/Desktop/15.072_AAE/Project/tabular_data/train.csv")
```

```
## Rows: 3960 Columns: 82
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (12): id, Basic_Demos-Enroll_Season, CGAS-Season, Physical-Season, Fitne...
## dbl (70): Basic_Demos-Age, Basic_Demos-Sex, CGAS-CGAS_Score, Physical-BMI, P...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
test <- read_csv("~/Desktop/15.072_AAE/Project/tabular_data/test.csv")
```

```
## Rows: 20 Columns: 59
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (11): id, Basic_Demos-Enroll_Season, CGAS-Season, Physical-Season, Fitne...
## dbl (48): Basic_Demos-Age, Basic_Demos-Sex, CGAS-CGAS_Score, Physical-BMI, P...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Here we aim to show some exploratory data analysis to further understand the tabular data that we have. There is a great deal of missing data in these data sets, which makes the problem we aim to solve both interesting and difficult. Recall that the subjects/observations in the model are indeed humans. Human nature is unpredictable and often the information that we have on human subjects is sporadic. This data displays this uncertainty regarding human behavior.

```
# Number of NA values in the training data:
total_NAs_train <- sum(is.na(train))
cat("Number of NAs in train: ", total_NAs_train, "\n")
```

```
## Number of NAs in train:  131717
```

```
# Number of NA values in the testing data:
total_NAs_test <- sum(is.na(test))
cat("Number of NAs in test: ", total_NAs_test, "\n")
```

```
## Number of NAs in test:  590
```

```
total_NAs <- total_NAs_test + total_NAs_train
cat("Total Number of NAs: ", total_NAs, "\n")
```

```
## Total Number of NAs:  132307
```

There are a total of 132307 NA values in the data.

We can now look at the number of NA values per column in this data.

```
NAs_per_column_train <- colSums(is.na(train))

for (col_name in names(NAs_per_column_train)) {
  cat("Number of NAs in column", col_name, ":", NAs_per_column_train[col_name], "\n")
}
```

```
## Number of NAs in column id : 0
## Number of NAs in column Basic_Demos-Enroll_Season : 0
## Number of NAs in column Basic_Demos-Age : 0
## Number of NAs in column Basic_Demos-Sex : 0
## Number of NAs in column CGAS-Season : 1405
## Number of NAs in column CGAS-CGAS_Score : 1539
## Number of NAs in column Physical-Season : 650
## Number of NAs in column Physical-BMI : 938
## Number of NAs in column Physical-Height : 933
## Number of NAs in column Physical-Weight : 884
## Number of NAs in column Physical-Waist_Circumference : 3062
## Number of NAs in column Physical-Diastolic_BP : 1006
## Number of NAs in column Physical-HeartRate : 993
## Number of NAs in column Physical-Systolic_BP : 1006
## Number of NAs in column Fitness_Endurance-Season : 2652
## Number of NAs in column Fitness_Endurance-Max_Stage : 3217
## Number of NAs in column Fitness_Endurance-Time_Mins : 3220
## Number of NAs in column Fitness_Endurance-Time_Sec : 3220
## Number of NAs in column FGC-Season : 614
## Number of NAs in column FGC-FGC_CU : 1638
## Number of NAs in column FGC-FGC_CU_Zone : 1678
## Number of NAs in column FGC-FGC_GSND : 2886
## Number of NAs in column FGC-FGC_GSND_Zone : 2898
## Number of NAs in column FGC-FGC_GSD : 2886
## Number of NAs in column FGC-FGC_GSD_Zone : 2897
## Number of NAs in column FGC-FGC_PU : 1650
## Number of NAs in column FGC-FGC_PU_Zone : 1689
## Number of NAs in column FGC-FGC_SRL : 1655
## Number of NAs in column FGC-FGC_SRL_Zone : 1693
## Number of NAs in column FGC-FGC_SRR : 1653
## Number of NAs in column FGC-FGC_SRR_Zone : 1691
## Number of NAs in column FGC-FGC_TL : 1636
## Number of NAs in column FGC-FGC_TL_Zone : 1675
## Number of NAs in column BIA-Season : 1815
## Number of NAs in column BIA-BIA_Activity_Level_num : 1969
## Number of NAs in column BIA-BIA_BMC : 1969
## Number of NAs in column BIA-BIA_BMI : 1969
## Number of NAs in column BIA-BIA_BMR : 1969
## Number of NAs in column BIA-BIA_DEE : 1969
## Number of NAs in column BIA-BIA_ECW : 1969
## Number of NAs in column BIA-BIA_FFM : 1969
## Number of NAs in column BIA-BIA_FFMI : 1969
## Number of NAs in column BIA-BIA_FMI : 1969
## Number of NAs in column BIA-BIA_Fat : 1969
## Number of NAs in column BIA-BIA_Frame_num : 1969
## Number of NAs in column BIA-BIA_ICW : 1969
## Number of NAs in column BIA-BIA_LDM : 1969
## Number of NAs in column BIA-BIA_LST : 1969
## Number of NAs in column BIA-BIA_SMM : 1969
## Number of NAs in column BIA-BIA_TBW : 1969
## Number of NAs in column PAQ_A-Season : 3485
## Number of NAs in column PAQ_A-PAQ_A_Total : 3485
## Number of NAs in column PAQ_C-Season : 2239
## Number of NAs in column PAQ_C-PAQ_C_Total : 2239
## Number of NAs in column PCIAT-Season : 1224
## Number of NAs in column PCIAT-PCIAT_01 : 1227
## Number of NAs in column PCIAT-PCIAT_02 : 1226
## Number of NAs in column PCIAT-PCIAT_03 : 1229
## Number of NAs in column PCIAT-PCIAT_04 : 1229
## Number of NAs in column PCIAT-PCIAT_05 : 1231
## Number of NAs in column PCIAT-PCIAT_06 : 1228
## Number of NAs in column PCIAT-PCIAT_07 : 1231
## Number of NAs in column PCIAT-PCIAT_08 : 1230
## Number of NAs in column PCIAT-PCIAT_09 : 1230
## Number of NAs in column PCIAT-PCIAT_10 : 1227
## Number of NAs in column PCIAT-PCIAT_11 : 1226
## Number of NAs in column PCIAT-PCIAT_12 : 1229
## Number of NAs in column PCIAT-PCIAT_13 : 1231
## Number of NAs in column PCIAT-PCIAT_14 : 1228
## Number of NAs in column PCIAT-PCIAT_15 : 1230
## Number of NAs in column PCIAT-PCIAT_16 : 1232
## Number of NAs in column PCIAT-PCIAT_17 : 1235
## Number of NAs in column PCIAT-PCIAT_18 : 1232
## Number of NAs in column PCIAT-PCIAT_19 : 1230
## Number of NAs in column PCIAT-PCIAT_20 : 1227
## Number of NAs in column PCIAT-PCIAT_Total : 1224
## Number of NAs in column SDS-Season : 1342
## Number of NAs in column SDS-SDS_Total_Raw : 1351
## Number of NAs in column SDS-SDS_Total_T : 1354
## Number of NAs in column PreInt_EduHx-Season : 420
## Number of NAs in column PreInt_EduHx-computerinternet_hoursday : 659
## Number of NAs in column sii : 1224
```

```
NAs_per_column_train_df <- as.data.frame(NAs_per_column_train)

NAs_per_column_train_df
```

```
##                                             NAs_per_column_train
## id                                                             0
## Basic_Demos-Enroll_Season                                      0
## Basic_Demos-Age                                                0
## Basic_Demos-Sex                                                0
## CGAS-Season                                                 1405
## CGAS-CGAS_Score                                             1539
## Physical-Season                                              650
## Physical-BMI                                                 938
## Physical-Height                                              933
## Physical-Weight                                              884
## Physical-Waist_Circumference                                3062
## Physical-Diastolic_BP                                       1006
## Physical-HeartRate                                           993
## Physical-Systolic_BP                                        1006
## Fitness_Endurance-Season                                    2652
## Fitness_Endurance-Max_Stage                                 3217
## Fitness_Endurance-Time_Mins                                 3220
## Fitness_Endurance-Time_Sec                                  3220
## FGC-Season                                                   614
## FGC-FGC_CU                                                  1638
## FGC-FGC_CU_Zone                                             1678
## FGC-FGC_GSND                                                2886
## FGC-FGC_GSND_Zone                                           2898
## FGC-FGC_GSD                                                 2886
## FGC-FGC_GSD_Zone                                            2897
## FGC-FGC_PU                                                  1650
## FGC-FGC_PU_Zone                                             1689
## FGC-FGC_SRL                                                 1655
## FGC-FGC_SRL_Zone                                            1693
## FGC-FGC_SRR                                                 1653
## FGC-FGC_SRR_Zone                                            1691
## FGC-FGC_TL                                                  1636
## FGC-FGC_TL_Zone                                             1675
## BIA-Season                                                  1815
## BIA-BIA_Activity_Level_num                                  1969
## BIA-BIA_BMC                                                 1969
## BIA-BIA_BMI                                                 1969
## BIA-BIA_BMR                                                 1969
## BIA-BIA_DEE                                                 1969
## BIA-BIA_ECW                                                 1969
## BIA-BIA_FFM                                                 1969
## BIA-BIA_FFMI                                                1969
## BIA-BIA_FMI                                                 1969
## BIA-BIA_Fat                                                 1969
## BIA-BIA_Frame_num                                           1969
## BIA-BIA_ICW                                                 1969
## BIA-BIA_LDM                                                 1969
## BIA-BIA_LST                                                 1969
## BIA-BIA_SMM                                                 1969
## BIA-BIA_TBW                                                 1969
## PAQ_A-Season                                                3485
## PAQ_A-PAQ_A_Total                                           3485
## PAQ_C-Season                                                2239
## PAQ_C-PAQ_C_Total                                           2239
## PCIAT-Season                                                1224
## PCIAT-PCIAT_01                                              1227
## PCIAT-PCIAT_02                                              1226
## PCIAT-PCIAT_03                                              1229
## PCIAT-PCIAT_04                                              1229
## PCIAT-PCIAT_05                                              1231
## PCIAT-PCIAT_06                                              1228
## PCIAT-PCIAT_07                                              1231
## PCIAT-PCIAT_08                                              1230
## PCIAT-PCIAT_09                                              1230
## PCIAT-PCIAT_10                                              1227
## PCIAT-PCIAT_11                                              1226
## PCIAT-PCIAT_12                                              1229
## PCIAT-PCIAT_13                                              1231
## PCIAT-PCIAT_14                                              1228
## PCIAT-PCIAT_15                                              1230
## PCIAT-PCIAT_16                                              1232
## PCIAT-PCIAT_17                                              1235
## PCIAT-PCIAT_18                                              1232
## PCIAT-PCIAT_19                                              1230
## PCIAT-PCIAT_20                                              1227
## PCIAT-PCIAT_Total                                           1224
## SDS-Season                                                  1342
## SDS-SDS_Total_Raw                                           1351
## SDS-SDS_Total_T                                             1354
## PreInt_EduHx-Season                                          420
## PreInt_EduHx-computerinternet_hoursday                       659
## sii                                                         1224
```

```r
NAs_per_column_test <- colSums(is.na(test))

for (col_name in names(NAs_per_column_test)) {
  cat("Number of NAs in column", col_name, ":", NAs_per_column_test[col_name], "\n")
}
```

```
## Number of NAs in column id : 0
## Number of NAs in column Basic_Demos-Enroll_Season : 0
## Number of NAs in column Basic_Demos-Age : 0
## Number of NAs in column Basic_Demos-Sex : 0
## Number of NAs in column CGAS-Season : 10
## Number of NAs in column CGAS-CGAS_Score : 12
## Number of NAs in column Physical-Season : 6
## Number of NAs in column Physical-BMI : 7
## Number of NAs in column Physical-Height : 7
## Number of NAs in column Physical-Weight : 7
## Number of NAs in column Physical-Waist_Circumference : 15
## Number of NAs in column Physical-Diastolic_BP : 9
## Number of NAs in column Physical-HeartRate : 8
## Number of NAs in column Physical-Systolic_BP : 9
## Number of NAs in column Fitness_Endurance-Season : 16
## Number of NAs in column Fitness_Endurance-Max_Stage : 17
## Number of NAs in column Fitness_Endurance-Time_Mins : 17
## Number of NAs in column Fitness_Endurance-Time_Sec : 17
## Number of NAs in column FGC-Season : 3
## Number of NAs in column FGC-FGC_CU : 7
## Number of NAs in column FGC-FGC_CU_Zone : 7
## Number of NAs in column FGC-FGC_GSND : 15
## Number of NAs in column FGC-FGC_GSND_Zone : 15
## Number of NAs in column FGC-FGC_GSD : 15
## Number of NAs in column FGC-FGC_GSD_Zone : 15
## Number of NAs in column FGC-FGC_PU : 7
## Number of NAs in column FGC-FGC_PU_Zone : 7
## Number of NAs in column FGC-FGC_SRL : 7
## Number of NAs in column FGC-FGC_SRL_Zone : 7
## Number of NAs in column FGC-FGC_SRR : 7
## Number of NAs in column FGC-FGC_SRR_Zone : 7
## Number of NAs in column FGC-FGC_TL : 7
## Number of NAs in column FGC-FGC_TL_Zone : 7
## Number of NAs in column BIA-Season : 12
## Number of NAs in column BIA-BIA_Activity_Level_num : 12
## Number of NAs in column BIA-BIA_BMC : 12
## Number of NAs in column BIA-BIA_BMI : 12
## Number of NAs in column BIA-BIA_BMR : 12
## Number of NAs in column BIA-BIA_DEE : 12
## Number of NAs in column BIA-BIA_ECW : 12
## Number of NAs in column BIA-BIA_FFM : 12
## Number of NAs in column BIA-BIA_FFMI : 12
## Number of NAs in column BIA-BIA_FMI : 12
## Number of NAs in column BIA-BIA_Fat : 12
## Number of NAs in column BIA-BIA_Frame_num : 12
## Number of NAs in column BIA-BIA_ICW : 12
## Number of NAs in column BIA-BIA_LDM : 12
## Number of NAs in column BIA-BIA_LST : 12
## Number of NAs in column BIA-BIA_SMM : 12
## Number of NAs in column BIA-BIA_TBW : 12
## Number of NAs in column PAQ_A-Season : 19
## Number of NAs in column PAQ_A-PAQ_A_Total : 19
## Number of NAs in column PAQ_C-Season : 11
## Number of NAs in column PAQ_C-PAQ_C_Total : 11
## Number of NAs in column SDS-Season : 10
## Number of NAs in column SDS-SDS_Total_Raw : 10
## Number of NAs in column SDS-SDS_Total_T : 10
## Number of NAs in column PreInt_EduHx-Season : 2
## Number of NAs in column PreInt_EduHx-computerinternet_hoursday : 4
```

```r
NAs_per_column_test_df <- as.data.frame(NAs_per_column_test)

NAs_per_column_test_df
```

```
##                                          NAs_per_column_test
## id                                                         0
## Basic_Demos-Enroll_Season                                  0
## Basic_Demos-Age                                            0
## Basic_Demos-Sex                                            0
## CGAS-Season                                               10
## CGAS-CGAS_Score                                           12
## Physical-Season                                            6
## Physical-BMI                                               7
## Physical-Height                                            7
## Physical-Weight                                            7
## Physical-Waist_Circumference                              15
## Physical-Diastolic_BP                                      9
## Physical-HeartRate                                         8
## Physical-Systolic_BP                                       9
## Fitness_Endurance-Season                                  16
## Fitness_Endurance-Max_Stage                               17
## Fitness_Endurance-Time_Mins                               17
## Fitness_Endurance-Time_Sec                                17
## FGC-Season                                                 3
## FGC-FGC_CU                                                 7
## FGC-FGC_CU_Zone                                            7
## FGC-FGC_GSND                                              15
## FGC-FGC_GSND_Zone                                         15
## FGC-FGC_GSD                                               15
## FGC-FGC_GSD_Zone                                          15
## FGC-FGC_PU                                                 7
## FGC-FGC_PU_Zone                                            7
## FGC-FGC_SRL                                                7
## FGC-FGC_SRL_Zone                                           7
## FGC-FGC_SRR                                                7
## FGC-FGC_SRR_Zone                                           7
## FGC-FGC_TL                                                 7
## FGC-FGC_TL_Zone                                            7
## BIA-Season                                                12
## BIA-BIA_Activity_Level_num                                12
## BIA-BIA_BMC                                               12
## BIA-BIA_BMI                                               12
## BIA-BIA_BMR                                               12
## BIA-BIA_DEE                                               12
## BIA-BIA_ECW                                               12
## BIA-BIA_FFM                                               12
## BIA-BIA_FFMI                                              12
## BIA-BIA_FMI                                               12
## BIA-BIA_Fat                                               12
## BIA-BIA_Frame_num                                         12
## BIA-BIA_ICW                                               12
## BIA-BIA_LDM                                               12
## BIA-BIA_LST                                               12
## BIA-BIA_SMM                                               12
## BIA-BIA_TBW                                               12
## PAQ_A-Season                                              19
## PAQ_A-PAQ_A_Total                                         19
## PAQ_C-Season                                              11
## PAQ_C-PAQ_C_Total                                         11
## SDS-Season                                                10
## SDS-SDS_Total_Raw                                         10
## SDS-SDS_Total_T                                           10
## PreInt_EduHx-Season                                        2
## PreInt_EduHx-computerinternet_hoursday                     4
```

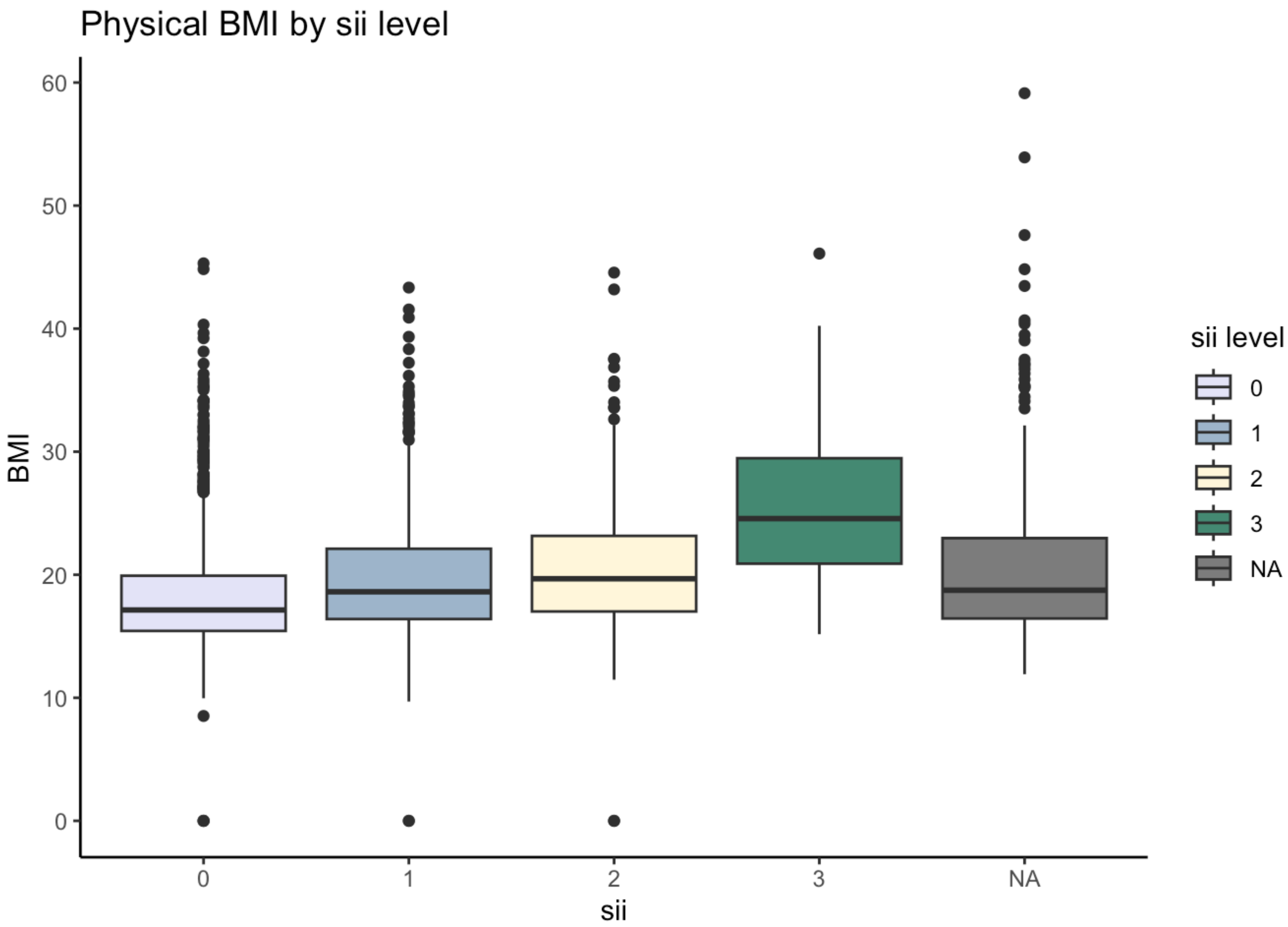
According to the output above, we can see that there is a substantial amount of NA values in nearly every column included in both the training and testing set of data relative to the total number of observations included in either data set. This high amount of missing data again is a major challenge, which is also a main point of this project. Data imputation could be important in order to appropriately predict the target variable, `ssi`. Furthermore, it is important to note that the target variable and the variables that are used to calculate the target variable, which begin with `PCIAT` all have a very high amount of NA values as well.

Let's try to glean some information about the target variable, `ssi`.

Box Plot series

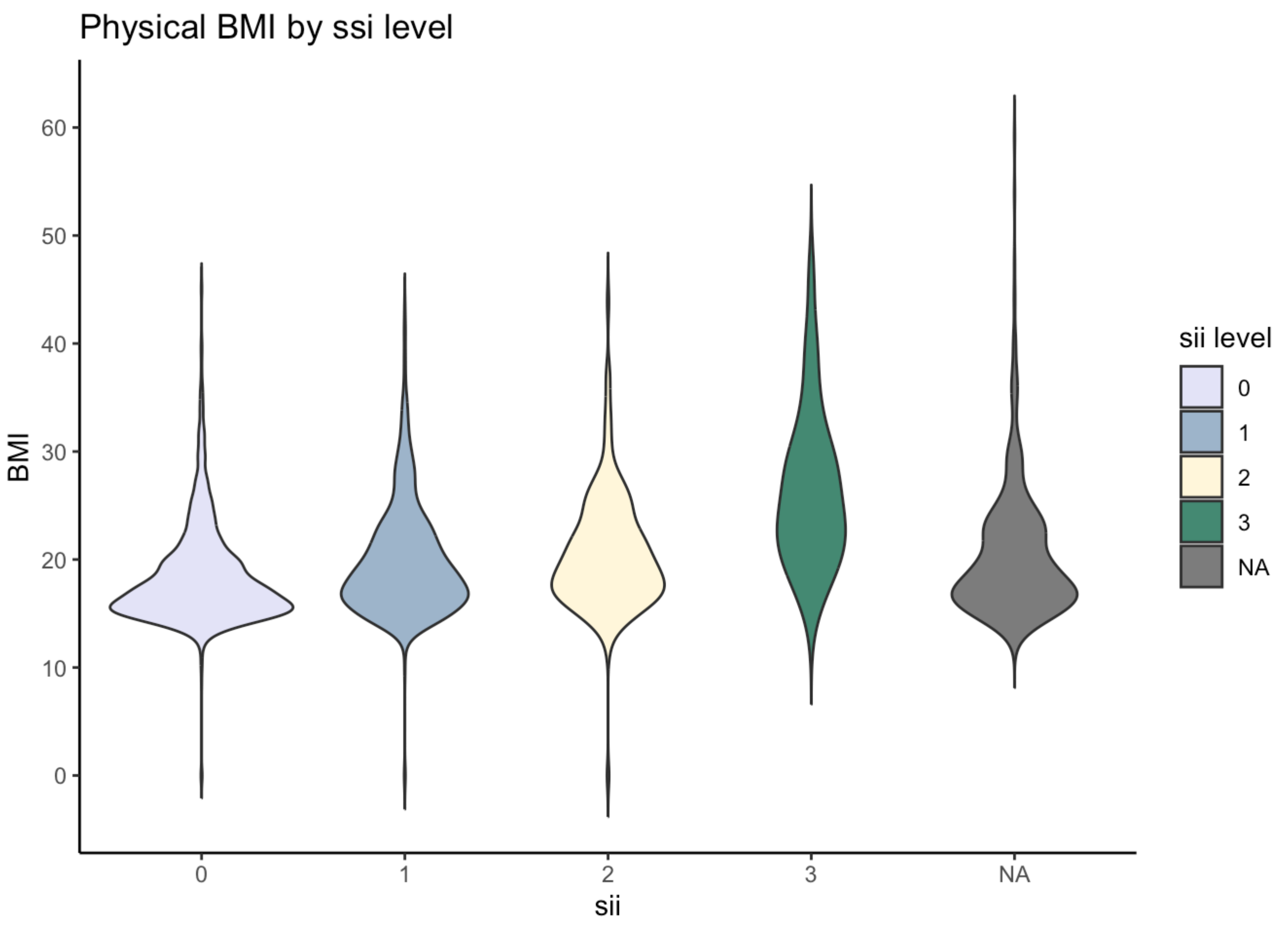
```r
train$sii <- as.factor(train$sii)

ggplot(train, aes(x=sii, y=`Physical-BMI`, fill=sii)) +
  scale_fill_manual(values = c("lavender","slategray3","cornsilk", "aquamarine4", "dimgrey")) +
  labs(title="Physical BMI by sii level",
       x="sii",
       y = "BMI")+ #labels such as titles, and axis labels.
  geom_boxplot(width=0.8)+ #sets the width of the boxplots used
  #coord_cartesian(ylim =  c(0, 2800))+: this is optional, and allows one to set the range on the y-axis
  theme_classic() +
  guides(fill=guide_legend(title = "sii level"))+
  scale_y_continuous(breaks = extended_breaks(n = 10)) #determines intervals on the y-axis
```

```
## Warning: Removed 938 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



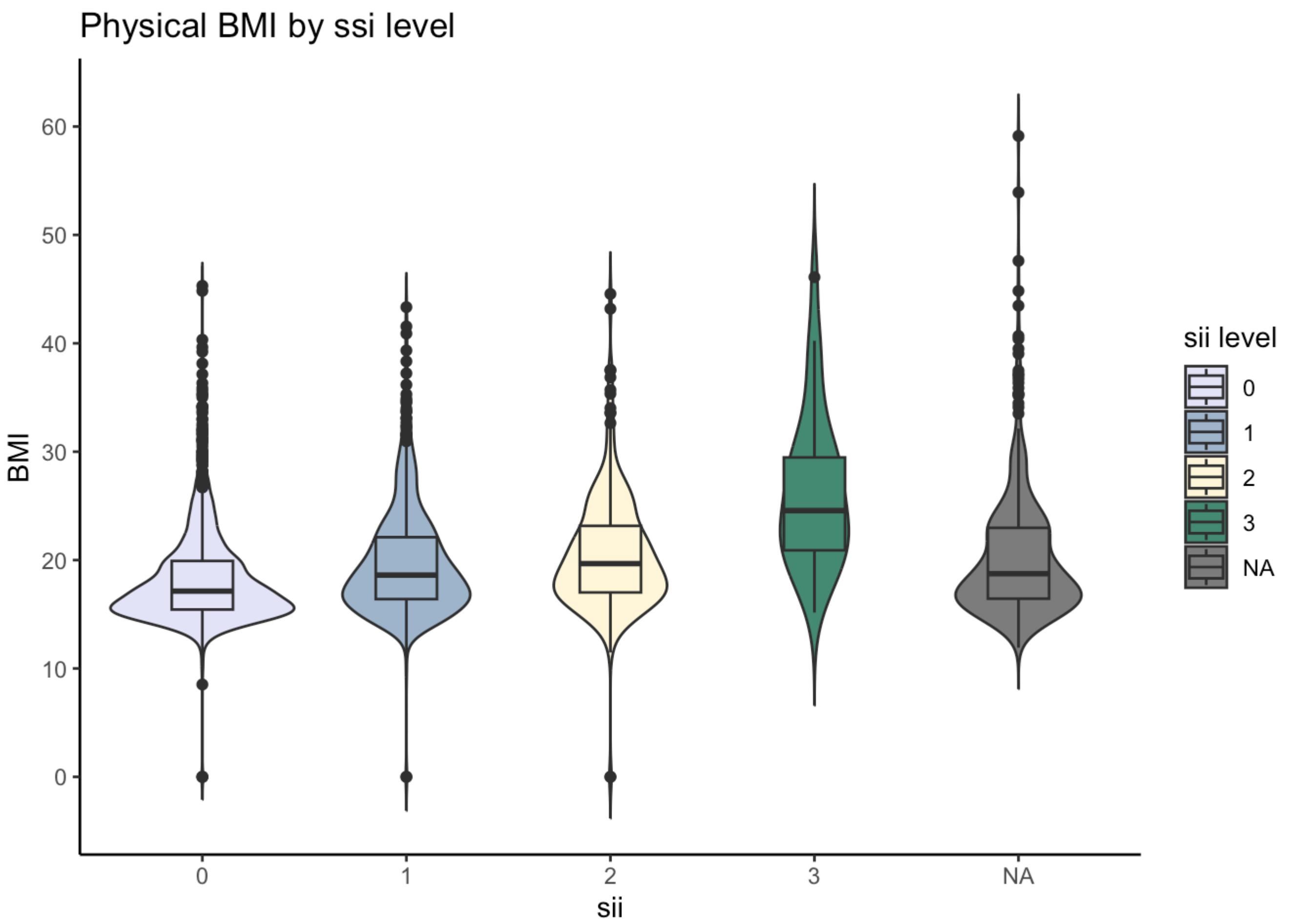

```r
#Violin Plot
ggplot(train, aes(x=sii, y=`Physical-BMI`, fill=sii)) +
  geom_violin(trim=FALSE)+
  scale_fill_manual(values = c("lavender","slategray3","cornsilk", "aquamarine4", "dimgrey")) + #set color values
for your legend
  labs(title="Physical BMI by ssi level",
       x="sii",
       y = "BMI")+ #labels for the plot
  theme_classic() +
  guides(fill=guide_legend(title = "sii level"))+ #legend title
  scale_y_continuous(breaks = extended_breaks(n = 10)) #determines intervals on the y-axis
```

```
## Warning: Removed 938 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```




```
# Violin and Boxplot Overlay
ggplot(train, aes(x=sii, y=`Physical-BMI`, fill=sii)) +
  geom_violin(trim=FALSE)+
  scale_fill_manual(values = c("lavender","slategray3","cornsilk", "aquamarine4", "dimgrey")) +
  labs(title="Physical BMI by ssi level",
       x="sii",
       y = "BMI")+ #labels such as titles, and axis labels.
  geom_boxplot(width=0.3)+ #sets the width of the boxplots used
  theme_classic() +
  guides(fill=guide_legend(title = "sii level"))+
  scale_y_continuous(breaks = extended_breaks(n = 10)) #determines intervals on the y-axis
```

```
## Warning: Removed 938 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
## Removed 938 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
favstats(`Physical-BMI` ~ sii, data = train)
```

```
##   sii      min       Q1   median       Q3      max     mean       sd    n
## 1   0  0.00000 15.43632 17.13579 19.92165 45.30603 18.32955 4.450628 1467
## 2   1  0.00000 16.39930 18.61484 22.11088 43.33770 19.78944 5.045417  676
## 3   2  0.00000 17.01367 19.67327 23.15733 44.55410 20.50330 5.203170  351
## 4   3 15.16685 20.90048 24.56353 29.47007 46.10291 26.26632 7.084858   33
##   missing
## 1     127
## 2      54
## 3      27
## 4       1
```

According to the box plot overlayed with a violin plot shown, we can learn from some important observations. Namely, the wider width of the lavender plot for an `sii` level of 0 shown that a lot of values in the data fall within the range of about a BMI of 15 and 19 for an `sii` level of 0, which suggests that many observations in the `sii` level 0 group have lower BMIs. On the other hand, when observing the aquamarine colored plot for an `sii` level of 3, we can see that the width of the violin is much narrower than the other `sii` values included in the plot. However, from BMI values of about 20 to 29, it appears most of the data falls within this range for observations in the `sii` level of 3. We can also note that according to the box plots there does not appear to be meaningful differences in the BMI between the `ssi` levels; however, we can notice a slight increase in the BMI distributions as the `sii` level increases. Namely, it appears that an `sii` level of 0 includes mostly observations with lower BMIs, while an `sii` level of 3 includes more observations with higher BMIs. Perhaps these observations may provide indications about the habits of individual subjects.

Recall that `sii` refers to “Severe Impairment Index” and higher values mean that an observation has more of a problem pertaining to problematic internet usage. It could be the case that the physical attributes of the observations can allow for inferences to be made regarding how problematic their internet usage might be. This plot tends to make some sense under a managerial lens as children who exercise less as a result of a high amount of interaction with the internet may develop higher BMIs. However, at this point, this is simply an inference, and should not be taken as a ground truth. Moving forward into modeling processes, we should however remember to consider the `Physical-BMI` as a potential variable of importance for predicting `sii`.

We can create similar plots for all of the variables that pertain to physical features in the data:

```
# Find all columns pertaining to physical features:

physical_features <- c("Physical-Height", "Physical-Weight",
                       "Physical-Waist_Circumference", "Physical-Diastolic_BP",
                       "Physical-HeartRate", "Physical-Systolic_BP")
```

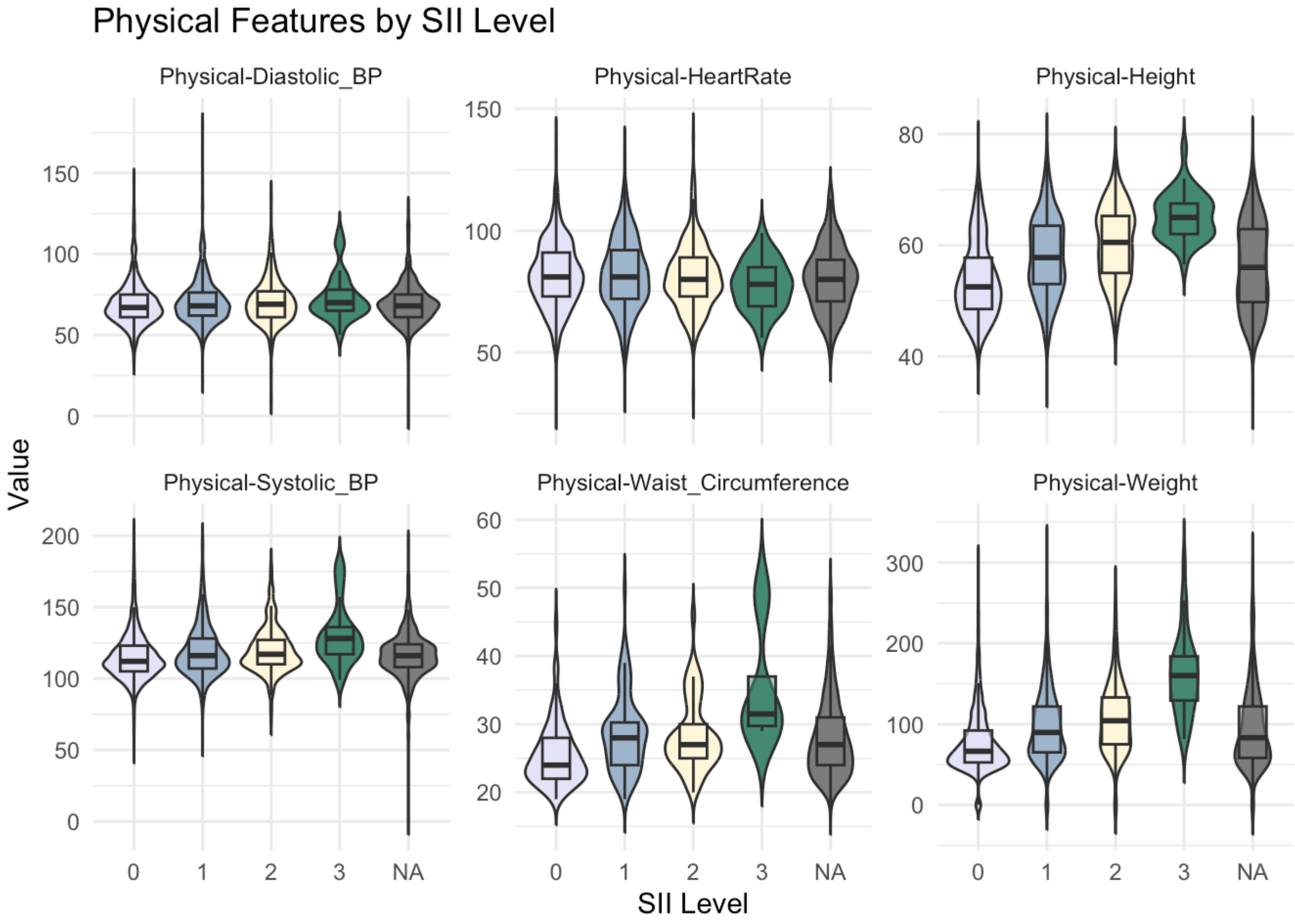
```
# Data preprocessing for plotting
df_physical_train <- train %>%
  dplyr::select(all_of(physical_features), sii) %>%
  pivot_longer(cols = all_of(physical_features),
               names_to = "Feature",
               values_to = "Value")  # Convert to long format for use with ggplot
```

```
# Function to generate violin and box plots for the remaining physical features.
generate_plots <- function(data, feature) {
  ggplot(data %>% filter(Feature == feature), aes(x = factor(sii), y = Value, fill = factor(sii))) +
    geom_violin(trim = FALSE, alpha = 0.6) +  # Violin plot for density
    geom_boxplot(width = 0.1, outlier.size = 0.5, alpha = 0.8) +  # Box plot inside violin
    labs(title = paste("Distribution of", feature, "by sii Level"),
         x = "sii", y = feature, fill = "sii") +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5, size = 14))
}
```

```
ggplot(df_physical_train, aes(x = as.factor(sii), y = Value, fill = as.factor(sii))) +
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.4, outlier.shape = NA, alpha = 0.5) +
  scale_fill_manual(values = c("lavender","slategray3","cornsilk", "aquamarine4", "dimgrey")) +
  facet_wrap(~Feature, scales = "free_y") +
  labs(title = "Physical Features by SII Level", x = "SII Level", y = "Value") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: Removed 7884 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

```
## Warning: Removed 7884 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

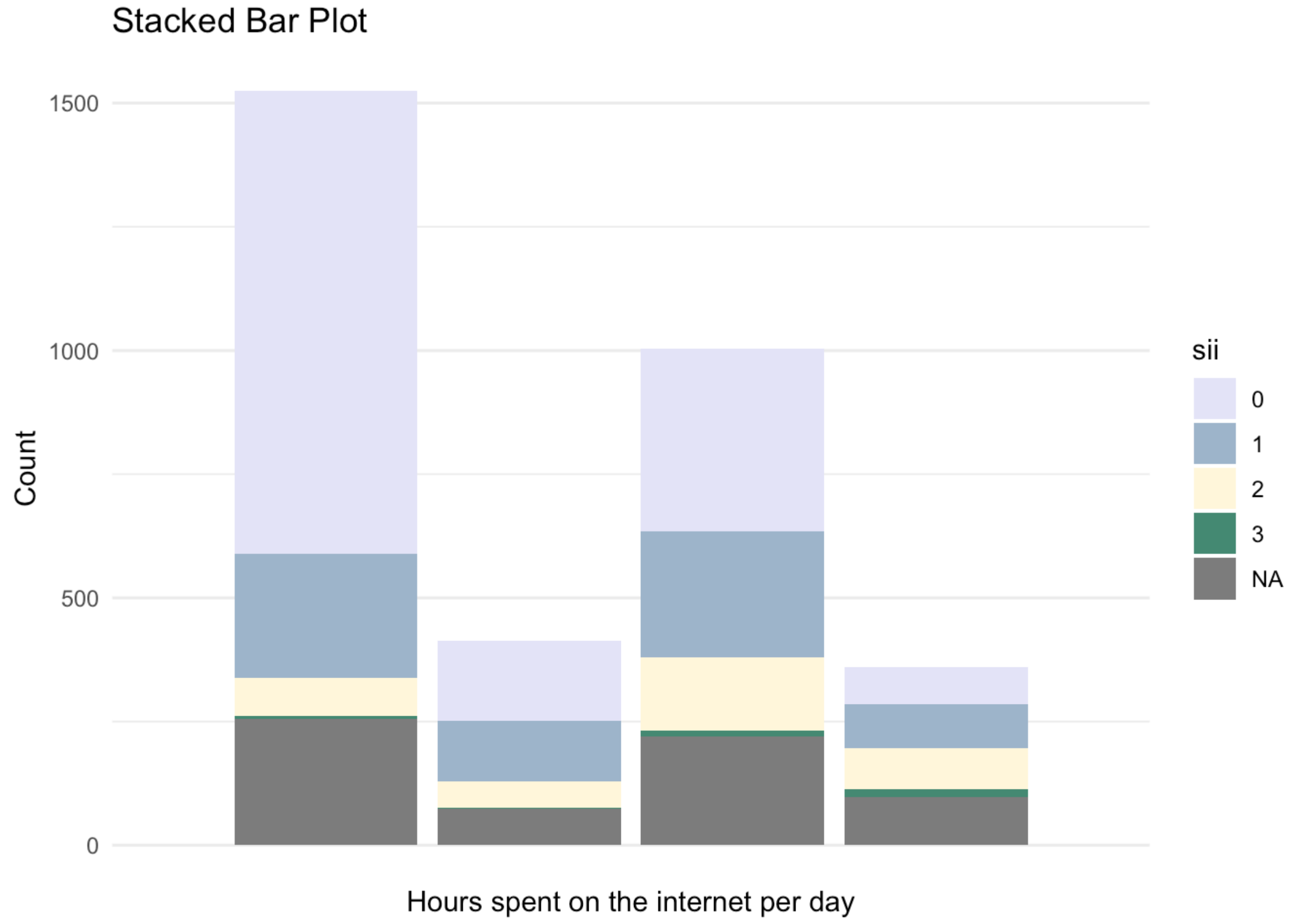


The boxplots shown here can be interpreted in a very similar fashion as the box plot shown previously that relates `Physical-BMI` to `sii`. Overall, of the physical features included in the data set, it appears that the diastolic blood pressures, heart rates, and systolic blood pressures are fairly similar across all `sii` groups. However, we do see some differences in the distributions for weight, waist circumference, and BMI, which may be important features to consider when we move into modeling.

We can now investigate the variable, `PreInt_EduHx-computerinternet_hoursday`, which is closely related to the target variable `sii`. `PreInt_EduHx-computerinternet_hoursday` refers to the number of hours that an observation spends using the compute or being engaged with the internet. This is a categorical variable where 0=Less than 1h/day, 1=Around 1h/day, 2=Around 2hs/day, and 3=More than 3hs/day.

```
x_true_labels <- c(
  "0" = "Less than 1 hour",
  "1" = "Around 1 hour",
  "2" = "Around 2 hours",
  "3" = "More than 3 hours"
)


plot <- ggplot(train, aes(x = `PreInt_EduHx-computerinternet_hoursday`,  fill = sii)) +
  geom_bar(position = "stack") +
  labs(
    title = "Stacked Bar Plot",
    x = "Hours spent on the internet per day",
    y = "Count",
    fill = "sii",
  ) + scale_fill_manual(values = c("lavender","slategray3","cornsilk", "aquamarine4", "dimgrey")) +
  scale_x_discrete(labels = x_true_labels)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+ # Rotate x-axis
  theme_minimal()

plot + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## Warning: Removed 659 rows containing non-finite outside the scale range
## (`stat_count()`).
```



According to the stacked bar plot, it is again noticeable that there are many NA values for `sii` across the various different levels of hours that a subject spends on the internet per day. However, we do notice that there is an abundance of observations that fall into a `sii` level of 0 that also use the internet for less than 1 hour per day. This observation is interesting and may be useful moving forward. It is reasonable to infer that a child that spends less time on the internet per day has less of a chance to develop bad habits pertaining to over usage and reliance on the internet. Again, thinking about the observations from our EDA under a managerial lens is important in moving to the next steps regarding what features may be important predictors of the `sii`, which is our target variable.

Other important features to consider moving forward:

There is a wide array of features in the data set that pertain to various features regarding a child's physical health and fitness. These features include information about a child's test performance on a series of fitness tests as well information on a child's bio electrical impedance analysis, which refers to body composition data.

Moving forward into the modeling process, we aim to use as many of the important features as possible in an effort to predict `sii`. It should be noted that from a managerial perspective it makes good sense to use all of the data included as these metrics largely give insight into a child's physical health and fitness, which may allow us to make inferences on the child's time spent using the internet. It follows that children who are more fit and healthy may be exercising more and spending less time interacting with the internet. However, the nature of this problem is inherently difficult because typically, information regarding problematic internet usage is self-reported. In this case, we aim to leverage physical fitness data as well, which is clearly measured and less susceptible to bias than self-reported data.