SLEEP ONSET AND WAKE UP:

Accelerometer data to explain and detect sleep patterns among subjects

- Our analysis aims to predict sleep onset and wake-up among child subjects. This kaggle data comes from wrist-worn accelerometers, the new-age data collection method that can help researchers understand sleep patterns.

- Our work will improve the ability of researchers to analyze accelerometer data for sleep monitoring and enable them to conduct large-scale studies of sleep.

Data description

Train events: Sleep logs for series in the training set recording onset and wake events.

- Event: Code representing different events during the night (e.g., falling asleep (Onset), waking up).
- Step: A step or sequence number associated with the event.
- Timestamp: The timestamp indicating when the event occurred.
- Id_map: Identifier for a specific individual or subject.

Train Series: Each series is a continuous recording of accelerometer data for a single subject spanning many days.

- Step: A step or sequence number associated with the accelerometer data.
- Timestamp: The timestamp indicating when the accelerometer data was recorded.
- Anglez: The angle variation (in degrees) captured by the accelerometer.
- Enmo: A measure of motion intensity or energy in the recorded data.
- Id_map: Identifier for a specific individual or subject.

There are total 277 subjects

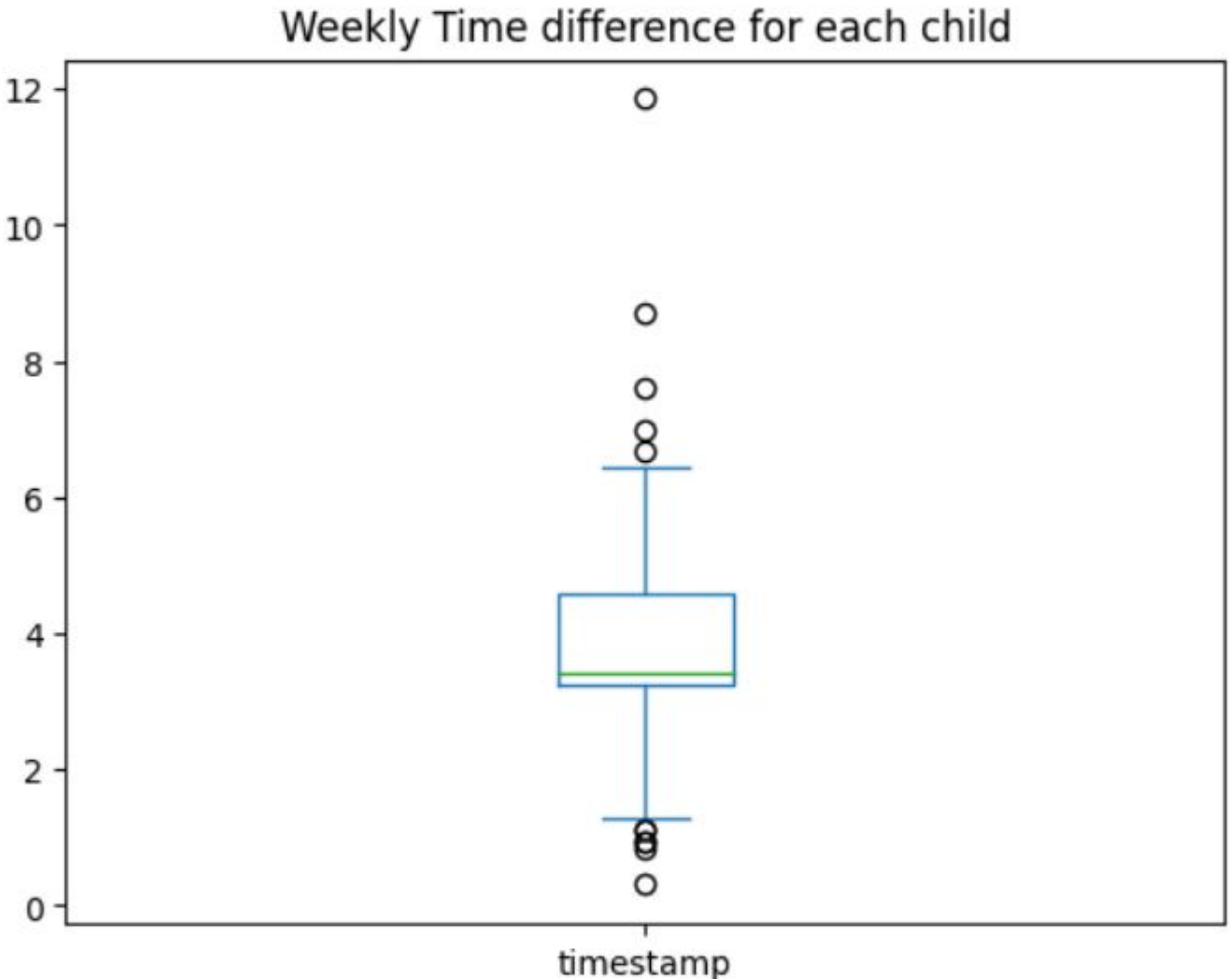
Predictors for sleep state

Here are important features needed to detect sleep state of a person:

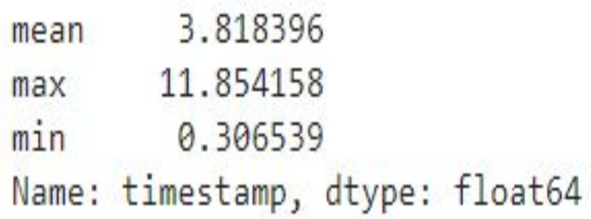
1. Movement Patterns (e.g., Accelerometer Data): Accelerometer data reflecting the subject’s movements during sleep can be crucial for identifying different sleep states.
2. Heart Rate Variability (HRV): Heart rate variability measures can provide insights into the autonomic nervous system and may be indicative of different sleep stages.
3. Sleep Duration: The total duration of sleep or specific sleep stages (e.g., deep sleep, REM sleep) is a fundamental predictor.
4. Age: The subject’s age can influence sleep patterns, and models often take this into account when predicting sleep states.
5. Environmental Factors: Room temperature, light exposure, and other environmental conditions can impact sleep. Including such features in the model might enhance predictive accuracy.

Our dataset contains 1, and 3 only. As per data description the CMI has more variables, but didn’t post for this competition.

What if every subject was a baby and they were observed from birth?

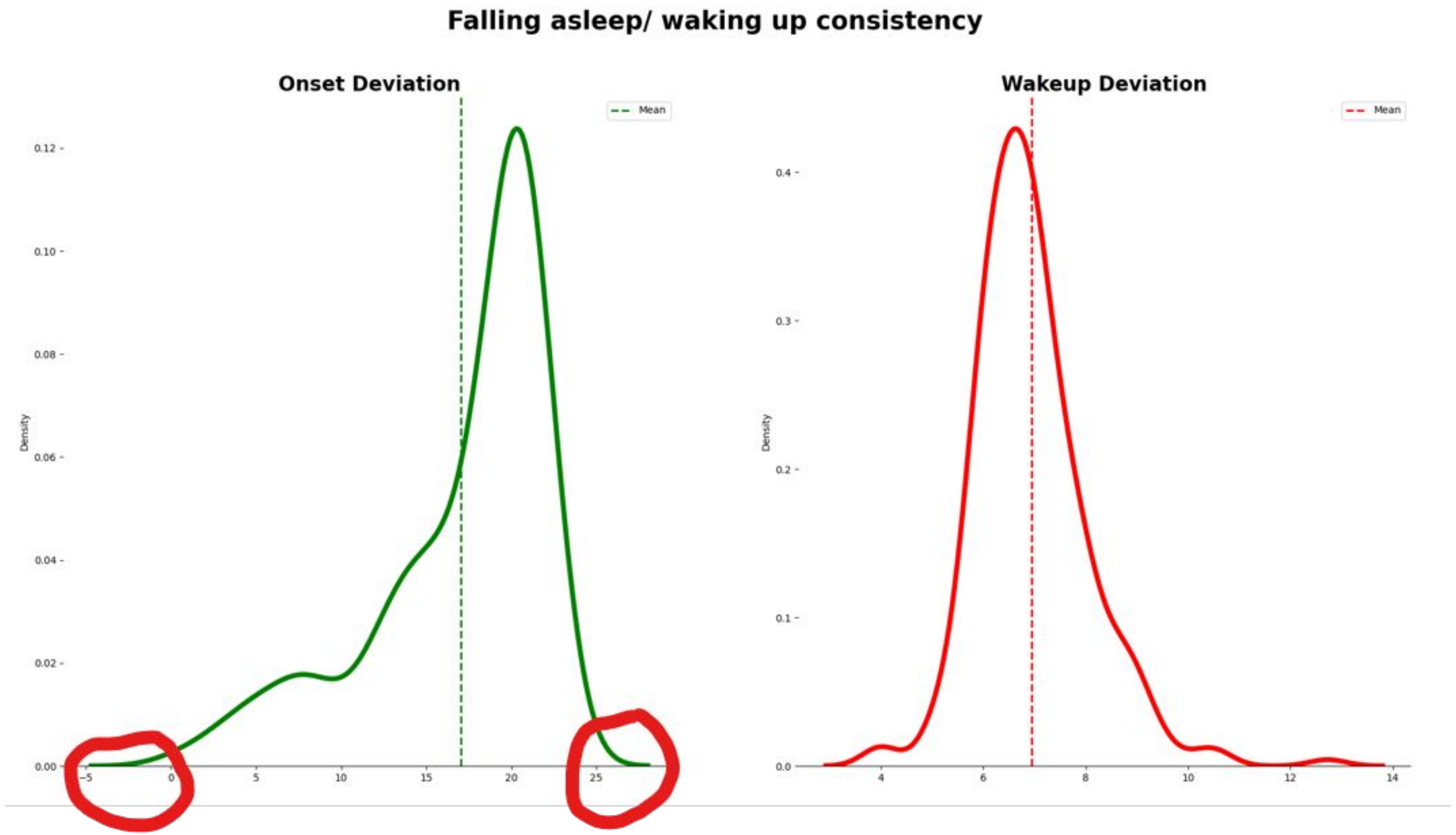


As per this graph, we would have concluded that all participants don't have enough data to check for changes in sleep patterns, say over a period of 3 months from birth. But we do not know anything about subject's age, and why these timestamps vary. We can at least note the substantial variability in observation period of each subject.



```
mean     3.818396
max     11.854158
min      0.306539
Name: timestamp, dtype: float64
```

Are there any outliers for days?



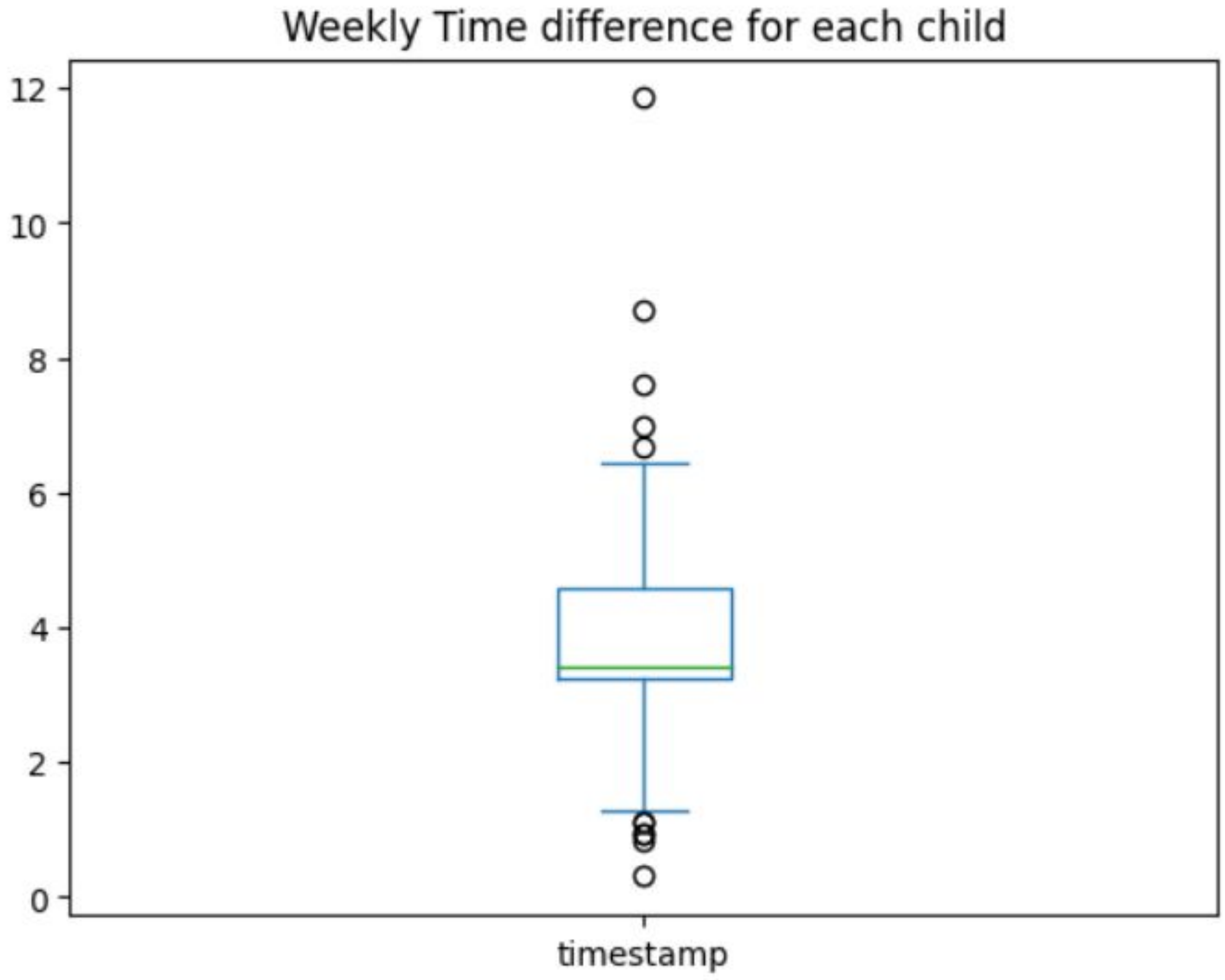
Falling asleep/ waking up consistency

Onset Deviation

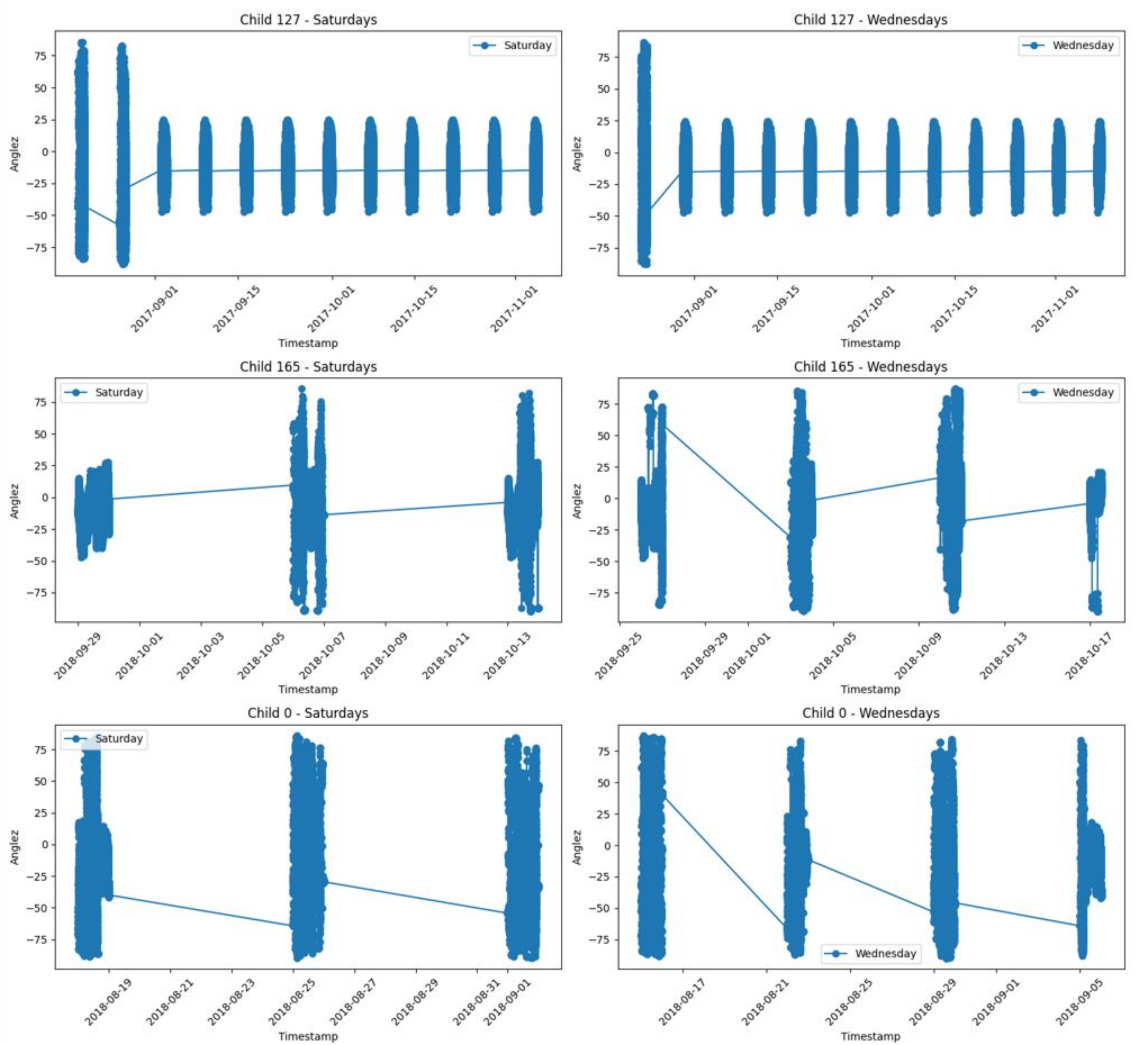
Wakeup Deviation

There are many outliers where subject are in sleep state where hours are in -ve as well as +24 hrs.

What’s the range of time difference between each child?



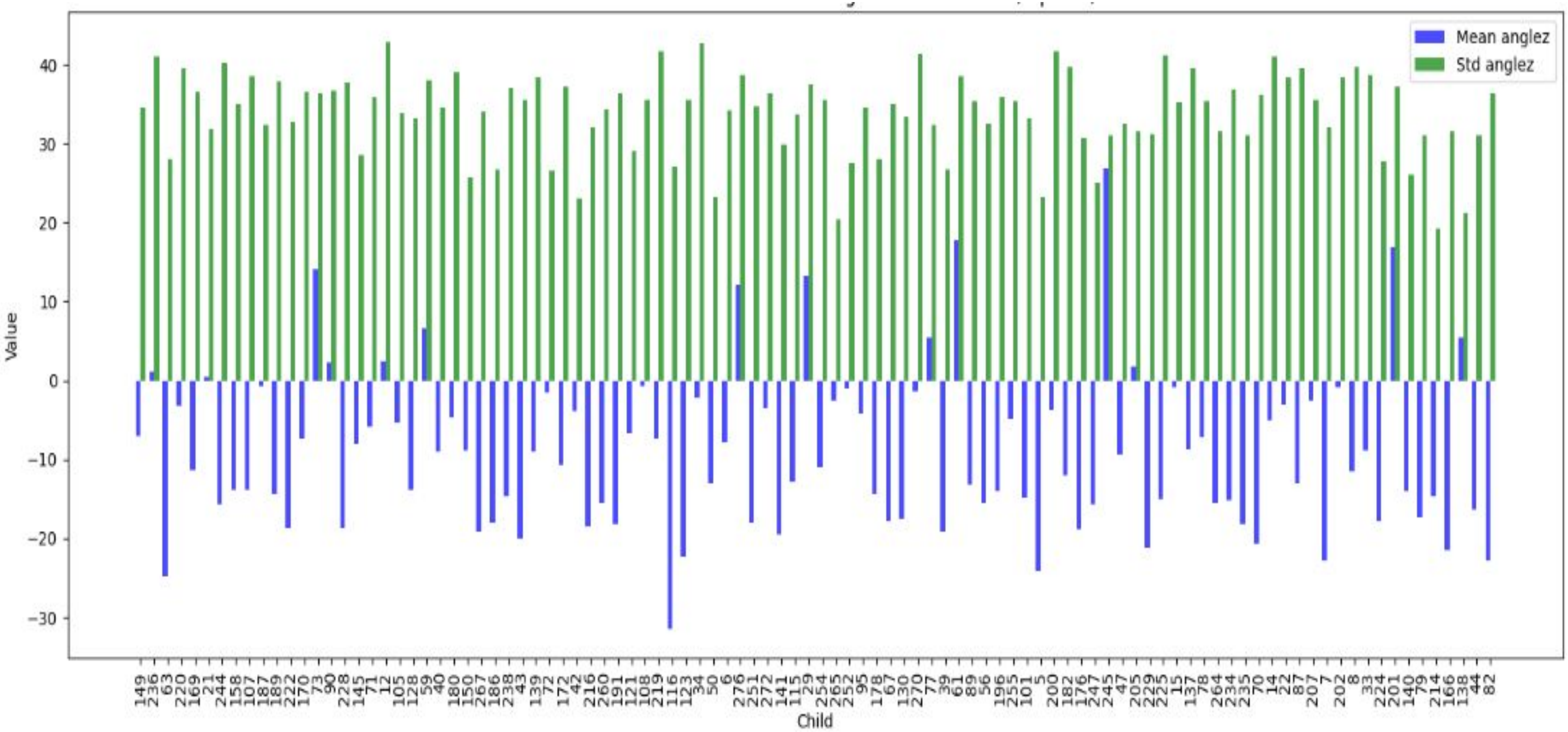
It can seen that most of the child have 3 weeks of dataset which will be used to split the dataset into



Comparison of top 3 childrens.

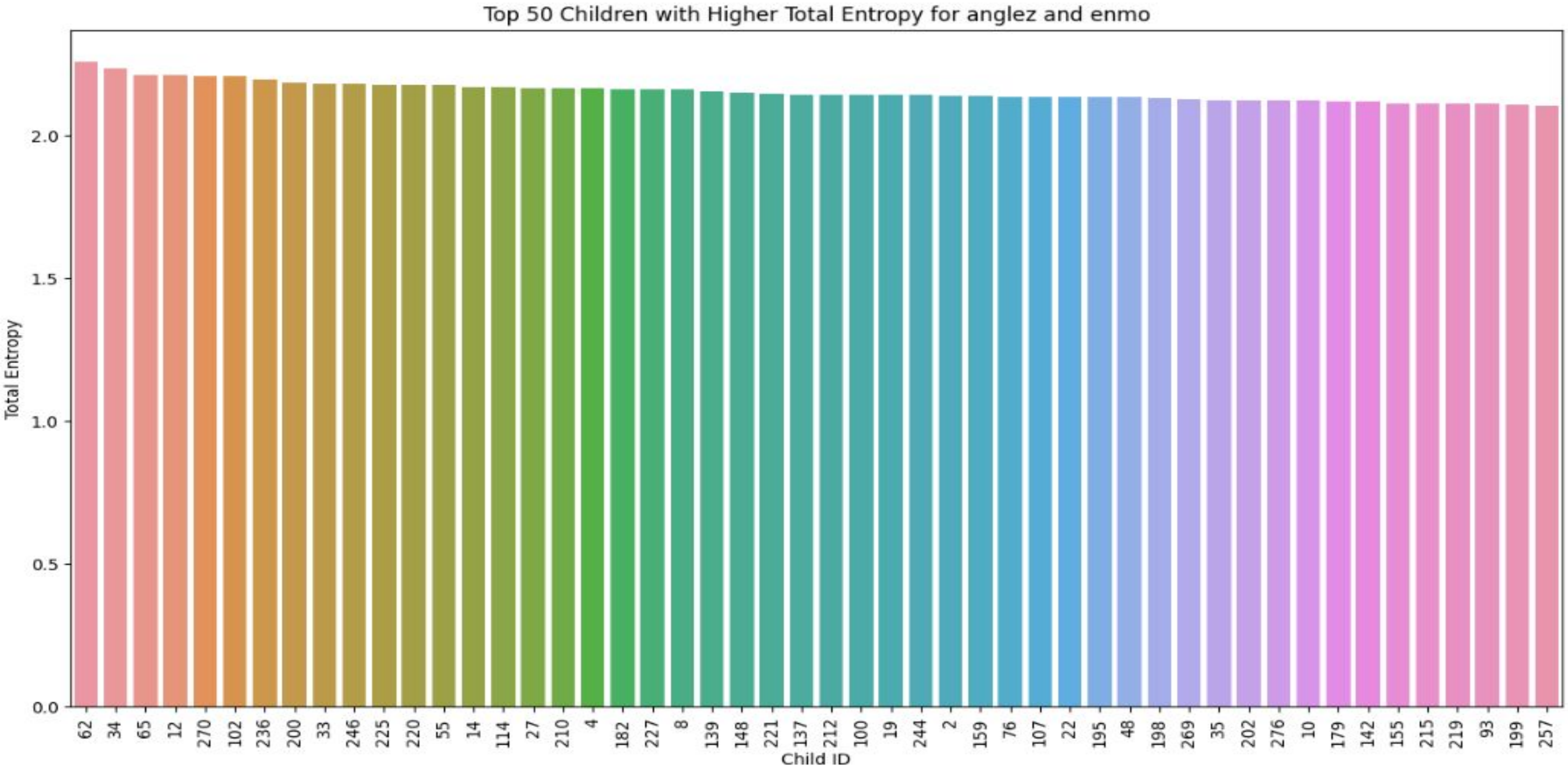
- There are lots of missing values for specific days for other childrens
- Tried to impute the dataset, but the data is already huge, so we’ve removed random samples from the child no. 127
- Child 0 doesn’t come in top 70 count

Mean and Standard Deviation of anglez for each Child



- Most of the children have –ve means and higher standard deviation
- This tells us that data is highly variable

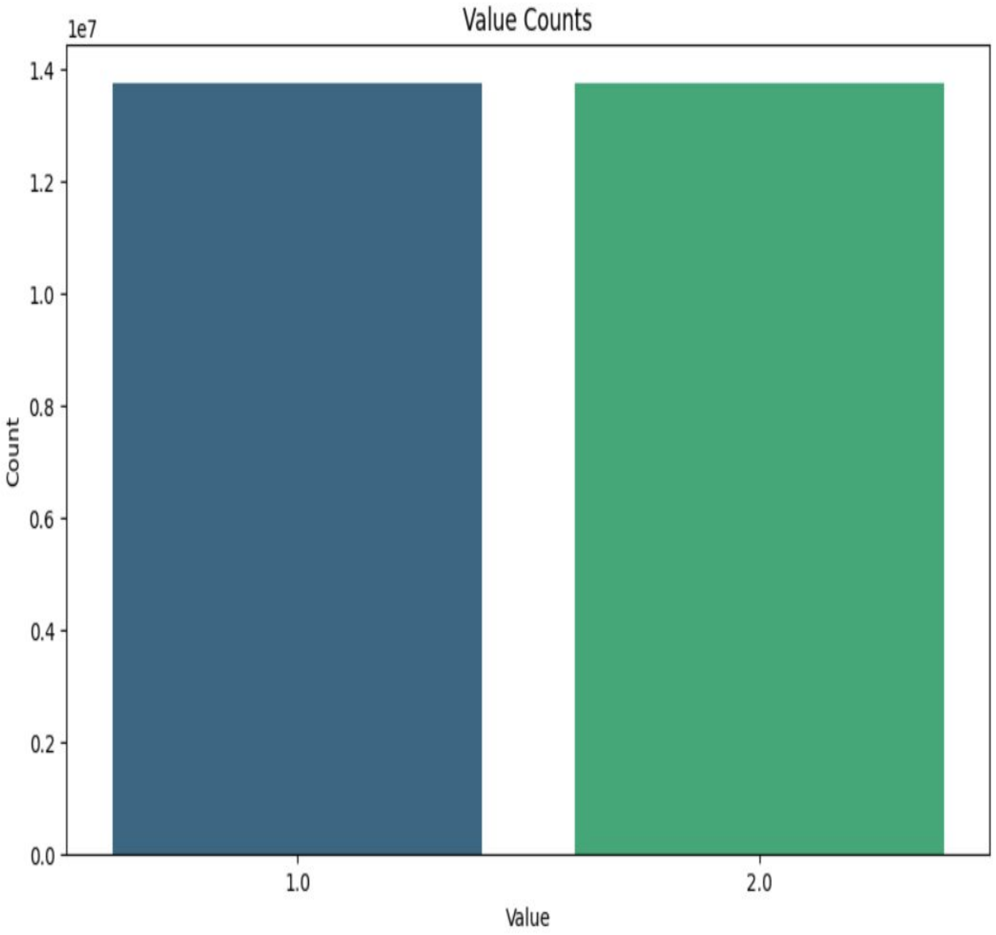
Which subject’s data contributes the most towards to study?



Here are top 50 subjects sorted according to entropy where subject 62 has the highest information entropy. This will be useful for data preparation before model building

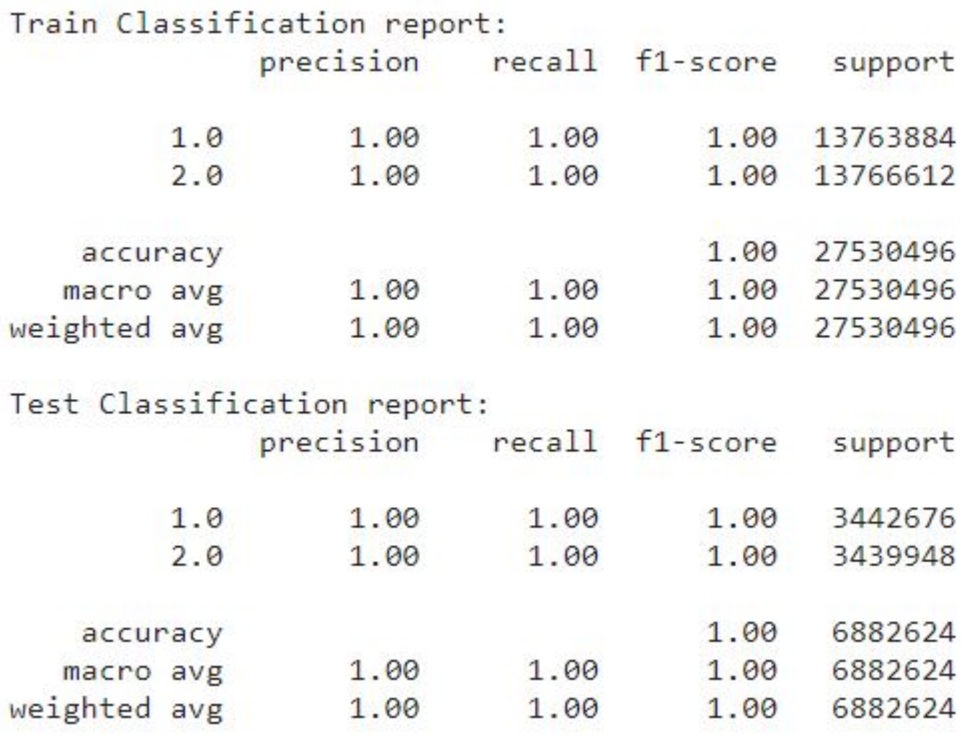
Predicting sleep onset and wakeup: ML models

- I previously noted the magnitude of second dataset. Its size hinders our ability join it to the sleep logs (data one). While we will be able to do it eventually, and I'll talk about in the next slide, I went ahead tried predicting onset and wakeup for just subject no. 127 which has most number of observations.
- Removed the outliers and standardized the dataset to fit into the model.
- Shape of the child 127 data for model training:
  - X_train: (27530496, 6), X_test: (6882624, 6)
  - Y_train: (27530496,), Y_test: (6882624,)
- Only trained on child no. 127 which has the highest number of observations because training on all the 277 childrens will take a large amount of time.



Gradient Boosting Machine

- Used gradient boosting machine in start as a base model and it performed well.
- Currently the model is able to detect the sleep state with 100 percent f1 score, which raises some concerns about the data.
- This is that it is due to the fact that the dataset it was trained on was too big and the model was trained on single child, which leads to no variation.
- **Hypothesis:** Does training the model on subset and pooled data will make a difference in the f1 score/precision of the model?



```
Train Classification report:
              precision    recall  f1-score   support

         1.0       1.00      1.00      1.00  13763884
         2.0       1.00      1.00      1.00  13766612

    accuracy                           1.00  27530496
   macro avg       1.00      1.00      1.00  27530496
weighted avg       1.00      1.00      1.00  27530496

Test Classification report:
              precision    recall  f1-score   support

         1.0       1.00      1.00      1.00   3442676
         2.0       1.00      1.00      1.00   3439948

    accuracy                           1.00   6882624
   macro avg       1.00      1.00      1.00   6882624
weighted avg       1.00      1.00      1.00   6882624
```

Next steps

1. I will investigate code in submissions made by other participants of the Kaggle competition to see what methods and packages they used
2. I will estimate three models on the entire data:
   a. Baseline Model: Apply the single-subject model specification to the entire dataset. We expect poor fit but it could serve as a good basis/test. We could also try a multi-linear regression model as the baseline
   b. Alternative Model 1: We will use {SVC, Decision Tree, Ensemble models and Meta models} as our first alternative model
   c. Alternative Model 2: We will use {some fancy model that's different from Model 1 above} as our second alternative model
3. I will compare Model 1 and 2 with the baseline models and report metric A, B, and C
4. The dataset can be considered on the basis of timeline which will be considered on the basis of most important subjects mentioned in the previous plot.

In summary, I have determined our next steps in sufficient detail after exploring the dataset, and will proceed accordingly.