HPBOARD

Data is the fuel of important strategic decisions for any Organisation. In this respect, its important to be sure that such decisions are inspired by reliable info. So, Monitoring data quality is the major concern in a business.

Suggestions:

.t3: if the data is empty we need to consider as zero , so that in future data processing will become easy

T3: in mvps session details , multiple sessions are split by semicolon(:) ,it should be using comma (,), so that we can get each session as an object to process

Data we are capturing now is,

Ex :

Participant X:if he/she had 3 mvpa session(each 10 min ) in an hour total mvpa is 45 Minutes:

Participant Y: if he/she had only one mvpa session(30) in an hour total mvpa is 35 Minutes:

Both the participants had the same time of actual activity (30 min) but the data we are capturing is 45 minutes and 35 minutes

So the additional 5 minutes of the activity session made a tremendous difference in the data. (Leads to less Data Quality)

Create the master table to get the 360 degree view of the participants

Which includes (profile creation date, activity, mvpa, step counts, latest update date)

Machine learning use cases:

Predict the upcoming participant counts : How many participants will create profile in future,

So that we can think what we need to enhance

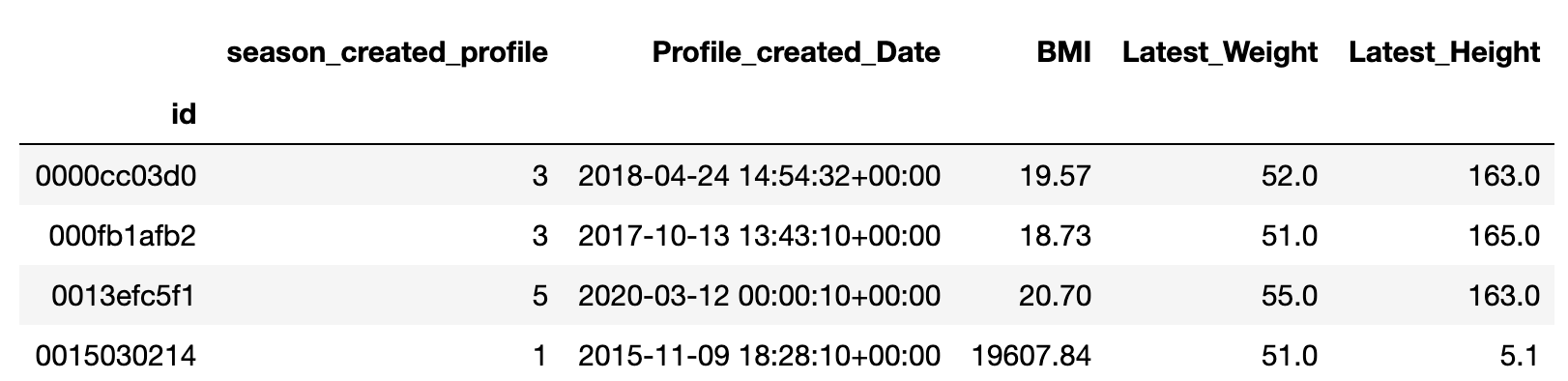
Predict the most mvpa hour in the day then run the campaign on that particular time related to new promotions

Data quality

1. Number of empty values. ...
2. Data transformation error rates. ..

Dataset1

Below Image height of the participant [5.1] Data Error ,which misleads the business values.



Dataset 3:

The output table has Null values in the empty places (Incompleteness).While processing the data for analysis it will be more user dependable

Two different strategies to think about the data:

1. User Perspective

How can we use the data to enhance the participant health and experience

1. Business Perspective

How to make the business values using the data we are capturing

To enhance the data quality

**Define data quality thresholds and rules**

Need to do root cause analysis to measure the data quality.

Error in the height

Duplicates in the data

Among the various techniques of **quality** management, **data quality metrics** must be of a top-notch and clearly defined. These **metrics** encompass different aspect of **quality**, that can be summed up with the acronym “ACCIT” standing for Accuracy, Consistency, Completeness, Integrity, and Timeliness

Dataset 1:

Profile create date is missing

<https://blog.syncsort.com/2019/08/data-quality/data-quality-dimensions-measure/>

|  |  |  |
| --- | --- | --- |
| Dimensions(Data Quality) | Ratings | Description |
| Data error rate | 5 to 1 | Lower the error rate higher the ratings |
| Validity (Data format) |  | Accessible formats |
| Uniqueness(Duplication rate) | 5 to 1 | Lower the duplication rate higher the ratings |
| Processing time | Acceptable/Not acceptable | Time taken for analysis and processing ( Lesser the time is accesstable) |
| Timeliness(Ready to access) |  |  |