K-means Clustering of Fitbit Heart Rate Data

Daniela del Río

Tuesday, December 14, 2021

1 Introduction

We are currently amid the covid-19 pandemic which has taken the lives of a total of 5.3 million people based on data published by John Hopkins (updated December 13, 2021) [1]. Nevertheless, the World Health Organization reports a total of 17.9 million deaths just in 2019 due to cardiovascular diseases [2]. That is, cardiovascular diseases cause at least 3 times more deaths in a single year than the total of covid deaths. If the world stopped because of the covid-19 pandemic, we should be taking at least a moment to address the public health crisis caused by cardiovascular diseases. Another relevant factor besides mortality, is the reversibility of cardiovascular diseases (if not caused by genetic factors). The reversibility nature of these diseases indicates that they are preventable because they are related with our lifestyle, for example, amount of exercise and diet.

Currently there are several companies which sell wearable technology for recording biosignals, like heart frequency. For example, Fitbit is a company owned by Google [3] which creates wearable devices for tracking an individual's variables such as: amount of steps, heart rate, amount of sleep time and oxygen saturation. This company allows their users to download from their website their data collected by their device. This opens the possibility of analyzing this data using machine learning techniques. This allows the extraction of valuable information from these recordings. One of such machine learning techniques is K-means, which will be further discussed below.

K-means is a clustering technique based on the euclidean distance between its elements. The objective of this clustering technique is minimizing the cost function J, and that is achieved by minimizing the distance between the cluster centers $\bar{\mu}_k$ and the datapoints $\bar{x}_n$. The entries r_{nk} are values of 1 if the n-th datapoint is in the k-th cluster and if not, 0. The cost function is defined by:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\bar{x}_n - \bar{\mu}_k\|^2 \tag{1}$$

The way we minimize the cost function is by iterating between assigning r_{nk} values and recalculating the cluster centers $\bar{\mu}_k$. We start by choosing random cluster centers, then we assign r_{nk} values and calculate the cost function. This is iterated and in each step, the cost function must decrease. The convergence criterion used is that the change of J should be small [4, pg. 348-353].

The author speculates that implementing the Fitbit as a device that helps monitor your health could potentially reduce the incidence of cardiovascular diseases. Thus, the motivation for this work in analyzing the Fitbit data with K-means clustering.

2 Methodology

Opening the Fitbit website from a web browser, the data was downloaded following this sequence: Settings, then Data Export and Export Your Account Archive. The .zip file was uncompressed, then the folder of the account holder was opened, and finally opening the Physical Activity folder. All the files are available as .json files, for example: *heart_rate-2021-08-30.json*.

The notebook was developed in Jupyter and written in Python. It is available in the following Github repository: https://github.com/DanidelRio/Heart. In this link you can also find the recordings of 55 days, nevertheless, the following analysis uses the first 20 recordings unless otherwise specified.

3 Results and Discussion

3.1 Time series

First of all was plotting the heart rate time series. This can be seen for the first recording in figure 1. Notice how the heart frequency oscillates between 50 and 70 until since the beginning of the recording until around 10 am and then it increases, this is probably because of a sleeping or resting period.

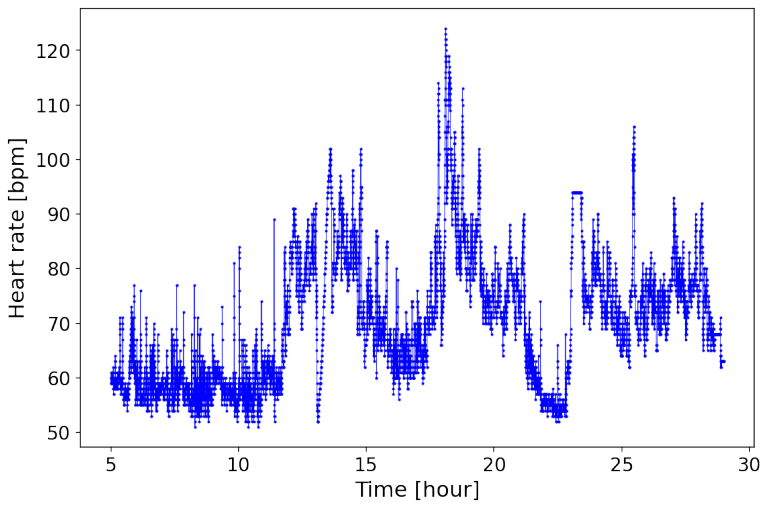

Figure 1: Time series extracted from one day's recording from the Fitbit.

We also concatenated the first 3 consecutive days and plotted each day with a different color. The result is figure 2. Note how there are marked increases and decreases of heart rate throughout the day. The time series we focused on for the rest of the analysis is shown in figure 3 where the first 20 recordings are concatenated.

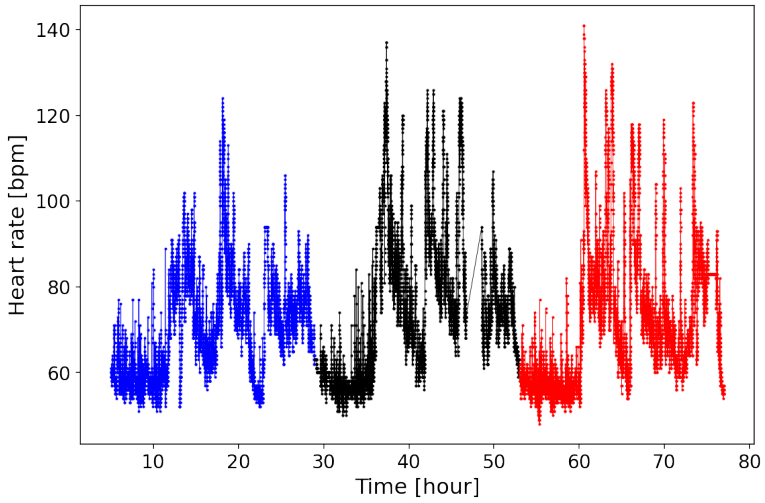

Figure 2: Concatenated time series from 3 days.

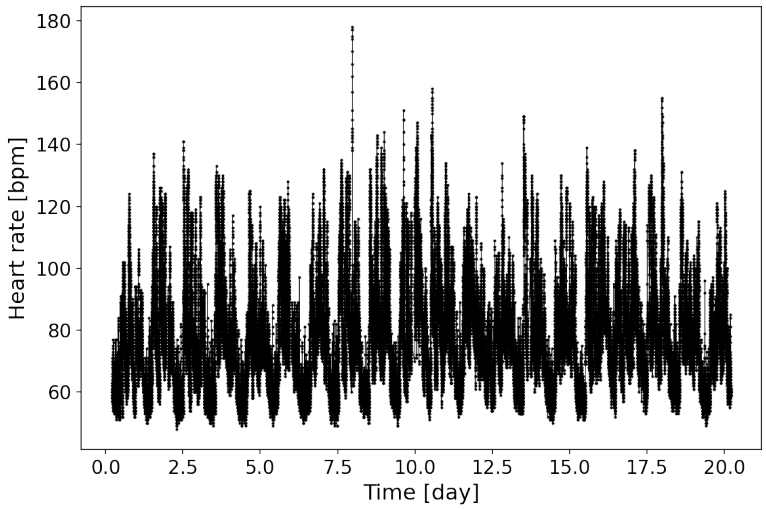

Figure 3: Concatenated time series from 20 days.

3.2 Mean and standard deviation for 5 s and 1 s timebins

The recording was divided in either 5 s or 1 s bins in order to obtain the mean heart rate and standard deviation of this timebin. These two quantities were plotted in figures 4 for the first 3 recordings and 5 for the first 20 recordings. Notice how there is a wider dispersion of points using the 1 s timebin compared to the 5 s timebin. This is expected because we are averaging in a smaller timebin in the 1 s data. Also notice how there are almost no elements with the 5s timebin with a 0 standard deviation.

A relevant remark is that in figures 4 and 5 there is information loss because there is no representation of the density of points in a single location in space. To address this concern, a heatmap was created, see figure 6, which is basically a histogram of figure 5 using the 1s and 5 s timebins. There is also a normalized histogram amplification in figure 7. Notice how most of the points are clustered in a heart rate between 50 and 60 and the standard deviation is between 0 and 2.5. This corresponds to a resting heart frequency, which can be thought of as the period of time when the subject was asleep.

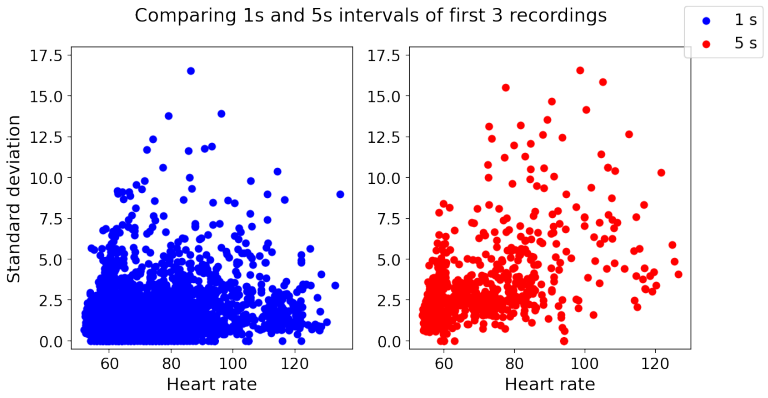

Figure 4: When using only the first 3 recordings, we see a wider dispersion of points using the 1s timebin compared to the 5s timebin.

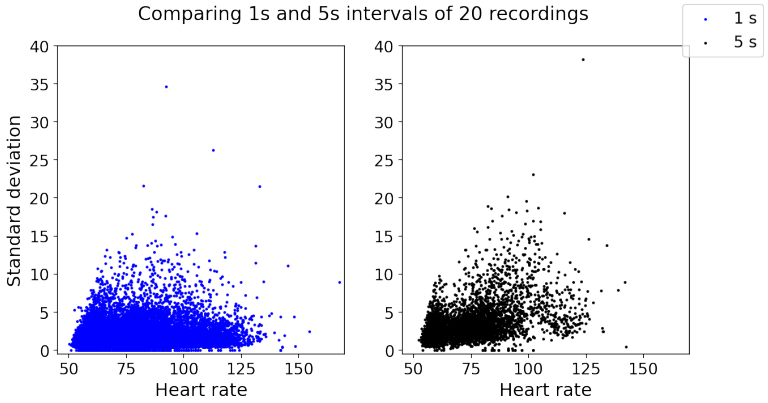

Figure 5: When comparing 20 recordings, we see the same trend as in the first 3 recordings; a wider cloud of points using the 1s timebin compared to the 5 s timebin.

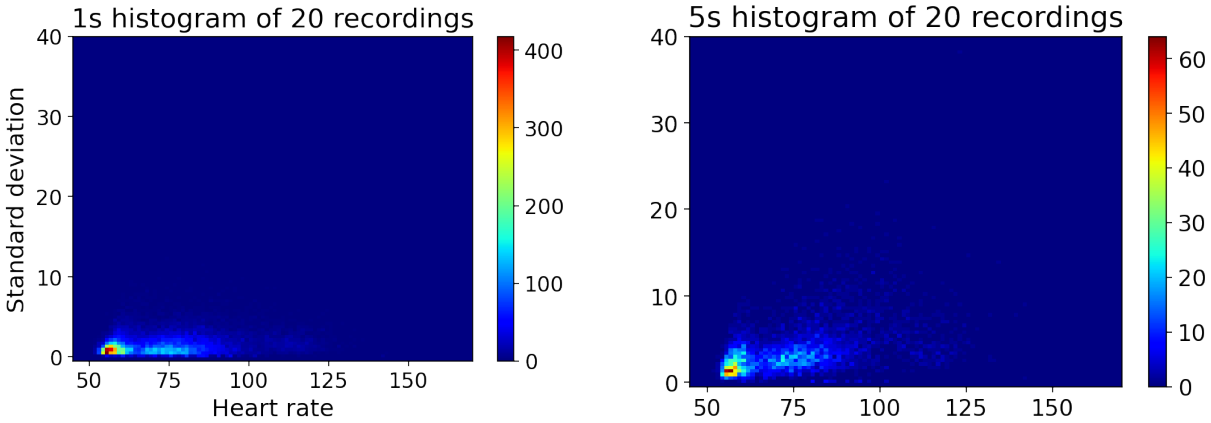

Figure 6: Histogram of the 20 first recordings using 1 s timebins left and 5 s timebins right. The colorbar indicates the number of datapoints. Notice there is a factor of 5 that should be considered in the 5 s histogram to make it comparable to the scale in the 1 s histogram.

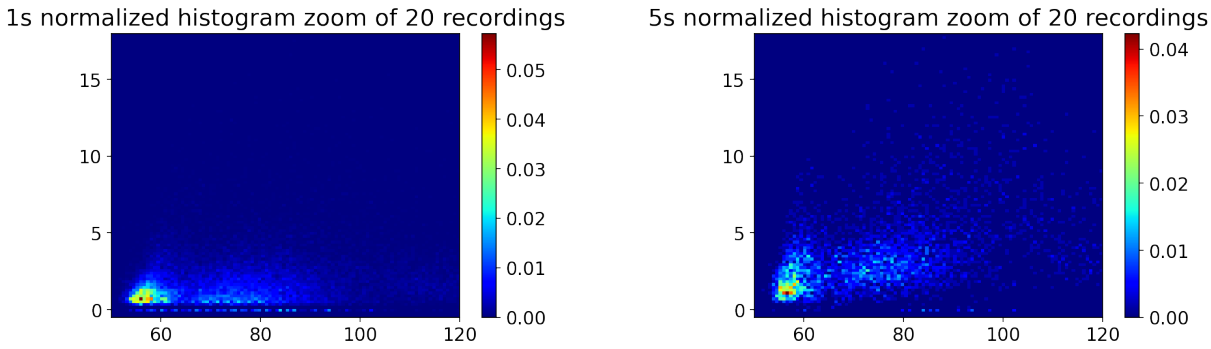

Figure 7: Normalized amplified histogram of the 20 first recordings using 1 s timebins left and 5 s timebins right.

3.3 K means clustering

The K-means clustering algorithm was described in the Introduction. In homework 4, we used this algorithm to minimize the distance of a vector in 1D, specifically voltage. In this case, we have data points $\bar{x}_n$ in 2 dimensions: heart rate and standard deviation. So the functions used in homework 4 were modified for considering these two variables. The K-means clustering algorithm was written in python and was used for obtaining the figures associated with 3 clusters (figure 8) and the cost function (figure 9). For using more than 3 clusters, the `KMeans` function from the sklearn package was used.

Note in figure 8 how the K-means algorithm basically marks vertical lines to separate the clusters, that is, the relevant parameter for clustering is the mean, not the standard deviation. And this can be observed both using the 1 s or the 5 s timebins. The next question one could ask is how many iterations did it take for the algorithm to reach a converging value. This depends on what our converging criteria is, but we can see how the cost function decreases with each iteration until practically having a constant value after 5 iterations in figure 9.

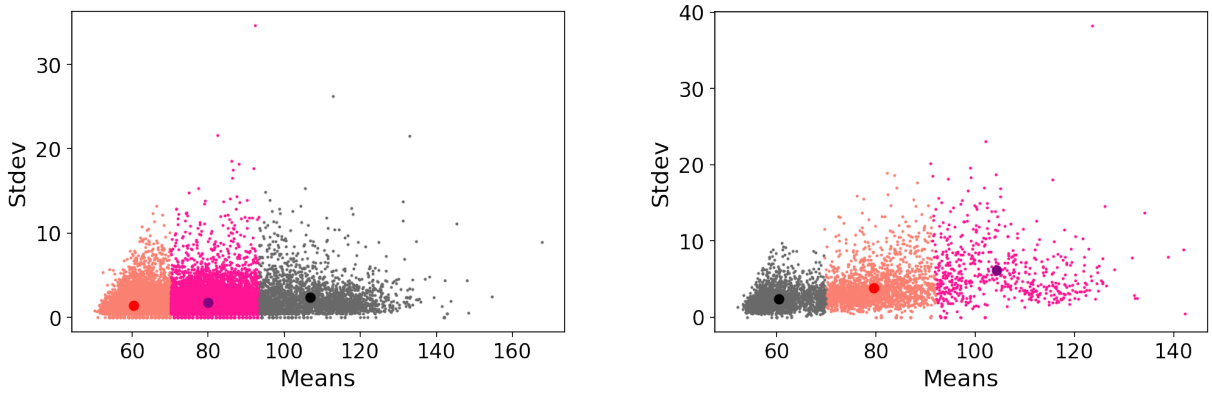

Figure 8: K means clustering of data using 1 s timebins left and 5 s timebins right using 3 clusters.

Now one could ask, what is the optimal number of clusters? This is a tricky question because we want enough clusters to group our data, but if we use too many, it will be harder to analyze even though the cost function will be lower. So we need to find the correct

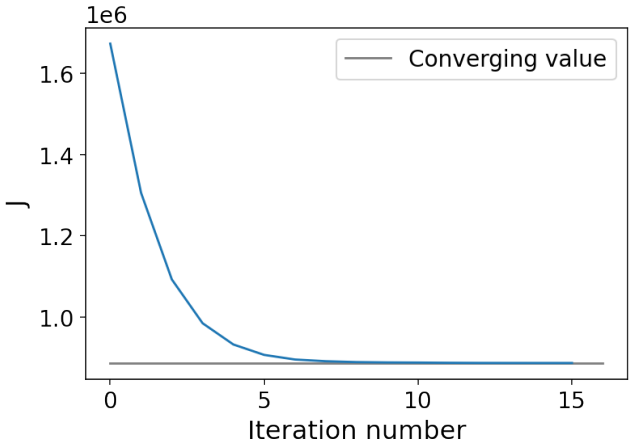

Figure 9: Number of iterations for algorithm to converge using 1 s timebins and 3 clusters.

balance. This balance is obtained when the change of the cost function (that is, the first derivative) changes by a small fraction of its value. In figure 10, one can visually appreciate this change occurs around 4 or 5 clusters, both using 1 s or 5 s timebins. Nevertheless, this can be confirmed in figure 11 where the derivative of the cost function is plotted. There we can see how the elbow of the change of the cost function is small with 4 or 5 clusters. So we clustered the data in these 4 (figure 12) or 5 (figure 13) clusters. Note in these figures how the relevant clustering parameter is the mean heart frequency instead of the standard deviation. This is because the clusters have very well defined vertical separations between them.

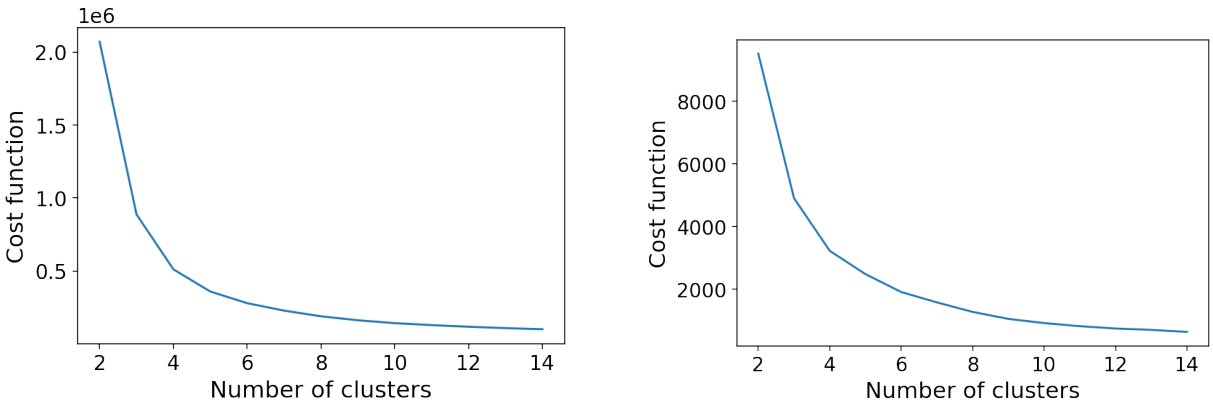

Figure 10: Left cost function for 1 s timebin data and right for the 5 s timebin data.

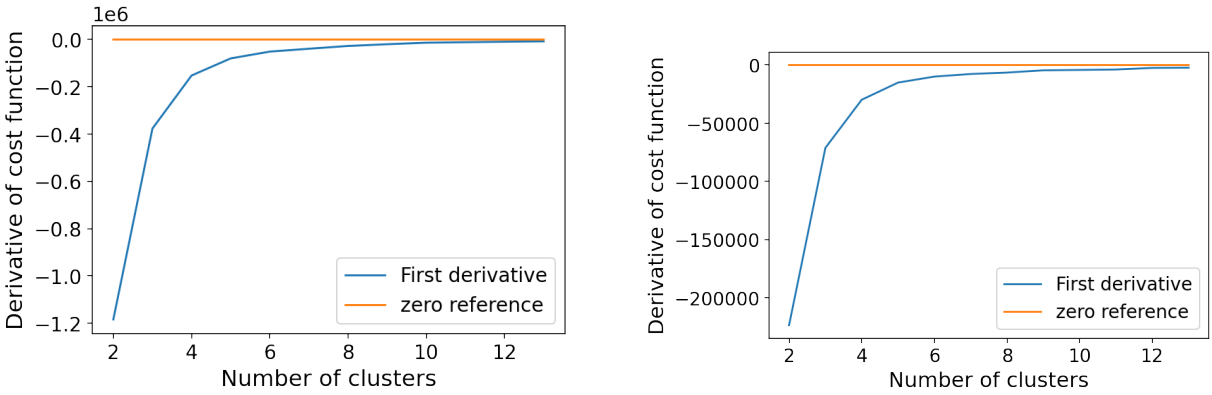

Figure 11: Convergence criteria, left for 1 s and right for 5 s. We plot the first derivative of the cost function. Note how in 4 or 5 clusters, the distance to 0 is small, that is the cost function's change due to number of clusters is close to zero.

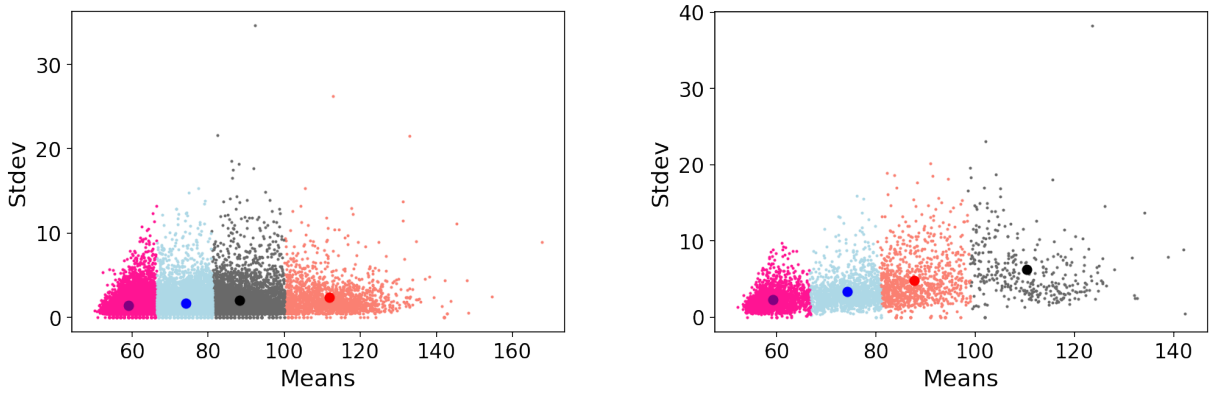

Figure 12: K means clustering of data using 1 s timebins left and 5 s timebins right using 4 clusters.

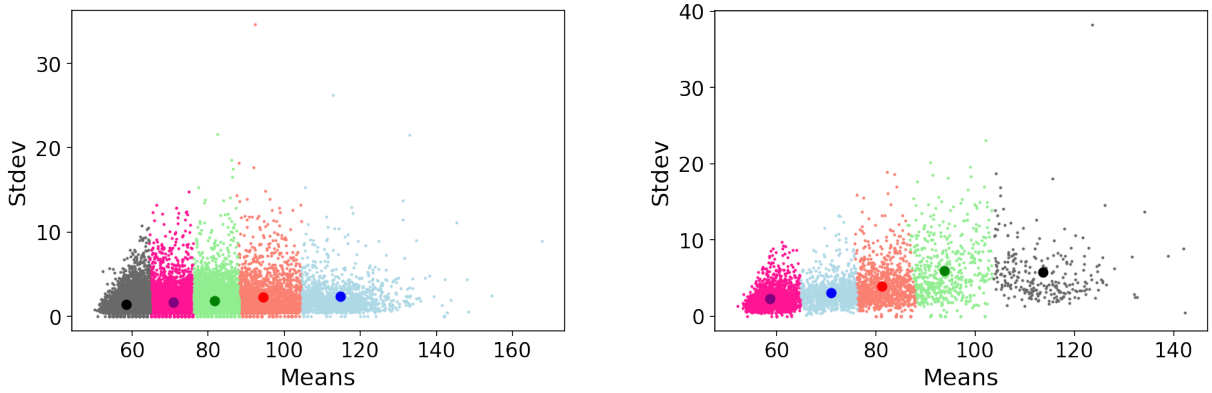

Figure 13: K means clustering of data using 1 s timebins left and 5 s timebins right using 5 clusters.

3.3.1 Clusters based on time-of-day labels

The amount of physical activity we perform during the day changes depending on the time. As the dataset had no associated labels, we created some based on the time of day. The

underlying idea being that physical activity is related with the mean heart frequency. The data was clustered based on this labels, thus obtaining figure 14 and 15. Even though the best value for k found in the previous section was 4 or 5 clusters, in this section there were 6 labels because we wanted to find if there was a distinctive pattern.

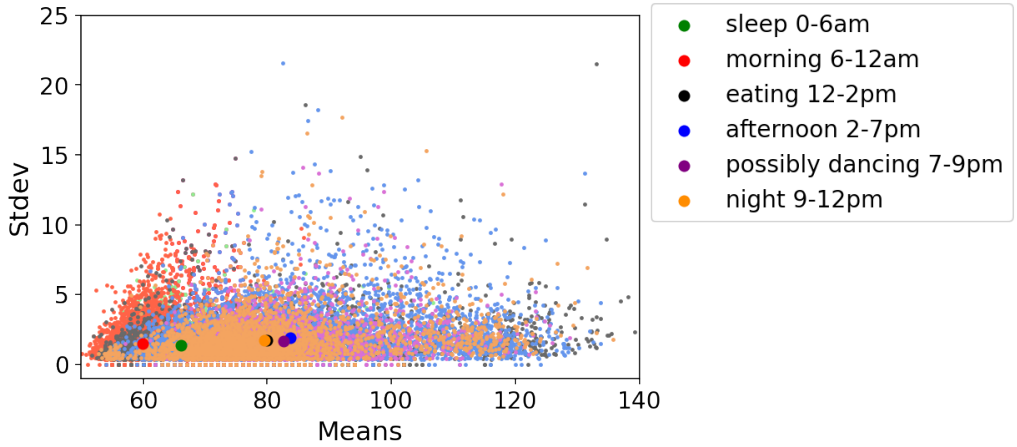

Figure 14: Clustered data based on labels which indicate the time of day using the 1 s timebin.

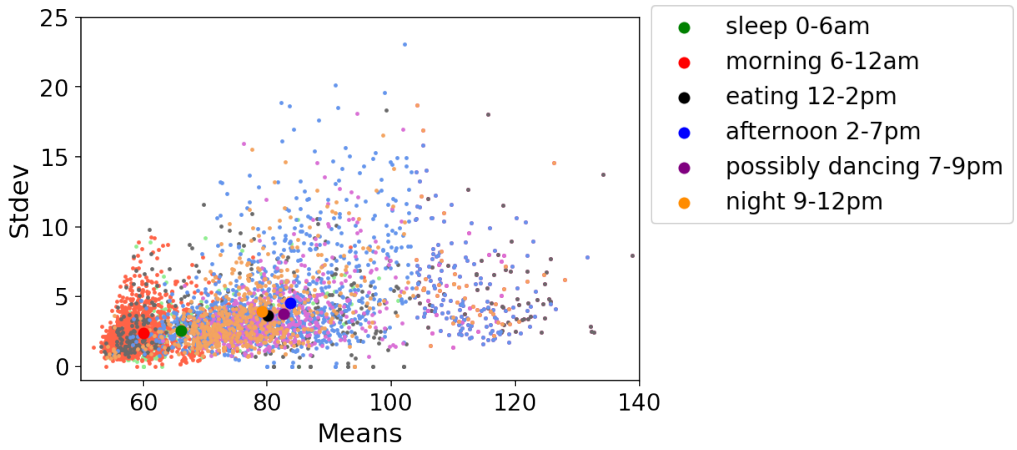

Figure 15: Clustered data based on labels which indicate the time of day using the 5 s timebin.

At least from the previous figures, we cannot say there is a clear pattern associated with the labels. It seems the labels are not separating adequately the data because there are no clearly marked clusters. The only possibly distinctive cluster is the morning cluster. But this is not the best visualizing technique because of overlap, problem we ran into before and thus created the heatmaps.

4 Future work

In the json files downloaded from Fitbit, there is a variable called *confidence*. This is an integer value associated to each data sample ranging from 0 to 3. Nevertheless, in none

of the documents downloaded or in the website was it possible to find an explanation to this variable. So the Fitbit support was contacted and after some email exchange, they replied with an answer. As seen in figure 16, "Confidence scores indicate confidence in the accuracy of the captured data with "3" being the highest and "0" being the lowest." With this confidence score, the heart rate time series could be weighted, giving more relevance or weight to entries with the highest confidence scores and low weight to entries with zeros.

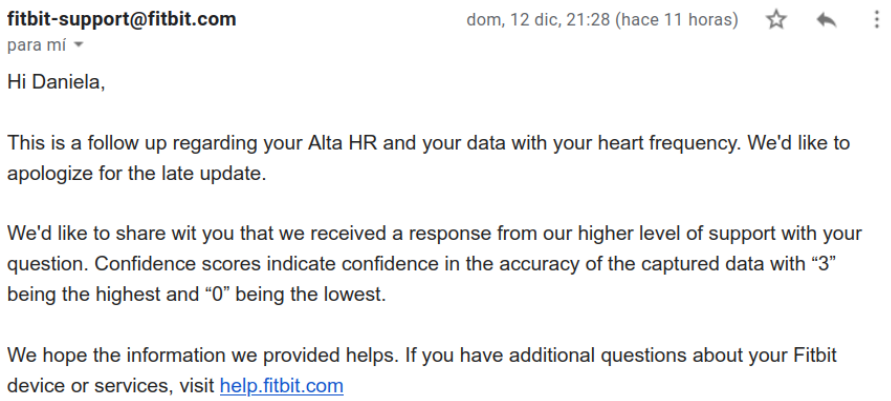

Figure 16: Fitbit's reply to question about confidence scores.

Another aspect we missed is there are 2 files in the Physical Activity folder downloaded from the Fitbit website called: *steps-2021-08-29.json* and *distance-2021-08-29.json*; which contain the number of steps per minute and the distance value per minute per day. This would have been useful because now we could define some intervals of physical activity intensity based on the values of these files. Even though there is no reference to the units of the distance file.

Another idea was to cluster the data using other clustering techniques, such as gaussian mixtures or principal components analysis. Citing Kevin Murphy, "Since K-means is not a proper EM algorithm, it is not maximizing likelihood. Instead, it can be interpreted as a greedy algorithm for approximately minimizing a loss function related to data compression."[4]. As we are using a time series, it is time-dependent, so we could use a Kalman filter or a hidden Markov model [5] to predict future states based on the past history. A potential application of this would be predicting a heart stroke. We did not see it in class, but researchers have even proposed an agent based model for clustering [6]. Using another clustering technique would allow the comparison between it and K-means.

We are only using the first 20 recordings, but this could be extended to using the 55 of them. A detail that was ran into, but not looked further was that there seems to be no mean heart rate in recording 25.

An idea for improving the labels is to label the heart rate data based on the physical activity performed at that time of day. For example:

1. Sleeping.
2. Non vigorous awake activity, like taking a class, or sitting at a desk.
3. Walking.
4. Riding a bike or dancing.

But for keeping this registrar of physical activity, one should be truly committed to do this rigorously.

For a better visualization of the labeled data, we could create a histogram for each label. This could help see the distribution of the datapoints.

5 Conclusions

When comparing 1 s to 5 s timebins, there was a wider dispersion in the 1 s timebins. This was expected because the we are averaging over a wider time period in the 5 s timebin.

We found the optimal number of clusters was 4 or 5 when using timebins of 1 s or 5 s. Nevertheless, other groups have used K-means as well and found the optimal number of clusters is 2 [7].

The most relevant parameter for k-means clustering between mean heart frequency and its standard deviation is heart frequency. This was visualized by the vertical lines which divided one cluster from another.

Finally, it is of utter importance to have an adequate data labeling which must be related to a relevant variable, such as physical activity in this case.

6 References

[1] E. Dong, H. Du, and L. Gardner. *An Interactive Web-Based Dashboard to Track COVID-19 in Real Time*. https://doi.org/10.1016/S1473-3099(20)30120-1. Accessed December 13, 2021. 2020.

[2] World Health Organization. *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed December 13, 2021. 2017.

[3] A. Williams. *Google Now Owns Fitbit: What It Means For Your Fitness Data Privacy*. https://www.forbes.com/sites/andrewwilliams/2021/01/14/google-now-owns-fitbit-what-it-means-for-your-fitness-data-privacy/?sh=17c9ef4039e1. Accessed December 13, 2021. 2021.

[4] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[5] S. Yun et al. "Forecasting of heart rate variability using wrist-worn heart rate monitor based on hidden Markov model". In: *2018 International Conference on Electronics, Information, and Communication (ICEIC)*. IEEE. 2018, pp. 1–2.

[6] M. Bursa and L. Lhotska. "Modified ant colony clustering method in long-term electrocardiogram processing". In: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2007, pp. 3249–3252.

[7] W. Materko. "Stratification Fitness Aerobic Based on Heart Rate Variability during Rest by Principal Component Analysis and K-means Clustering." In: *Journal of Exercise Physiology Online* 21.1 (2018).