Body Mass Index and Number of Steps Analysis from Activity Tracker Data

Renata Rego
25 October 2017

Abstract. The general goal of this project is to analyze steps and body weight data from users of activity trackers, and to provide insights on how these data could be used to help users on reaching their fitness goals. The database used in this project was kindly provided by Medisana® and consists on user data stored on their Vitadock Online Platform.

1 Introduction

Obesity and sedentary lifestyle are known to belong together. The question we want to investigate in this project is how daily steps, measured with a popular and accessible equipment, are related to the increase or decrease of the body mass index (bmi), and whether it is possible to provide users insights of how increasing their daily steps could help with decreasing their bmi (i.e, loosing weight). The users population considered in this project doesn't have any restriction of sex, age, height, etc. Further statistics about the available data, such as age range, amount of female and male users, etc. will be presented in the Exploratory Data Analysis.

2 Data Acquisition

The Vitadock database schema is available at the Vitadock API repository . The data was kindly provided by Medisana® for this project, and due to privacy reasons it cannot be publicly shared. The tables and fields of interest for this project are:

- user settings: user_id, sex, height, birthday
- targetscale: user_id, measurement_date, body_weight, bmi
- targetscale settings: user_id, measure_unit
- tracker stats: user_id, measurement_date, steps, calories, distance

The data of interest were initially read from a MySQL database and saved as json files (one file per table listed above). We observed that some users have several entries in the user_settings table, i.e., there are multiple entries of user_settings with the same value for the field user_id, which was expected to be unique. The explanation for this is

that the database keeps a history of all changes of each particular user settings. In this project, we took the most recently updated setting for each user with multiple entries. This is reasonable because users usually fill in few a required information when registering their account, and as they get interested in the device they update their accounts, adding new data, or updating values inserted automatically by default.

Once we read the data from the relational database into json files, we were ready to start the next steps of the project: data exploration and data preparation.

3 Data Exploration and Data Preparation

The main data exploration tasks performed in this project were:

- Getting an overview of the data. For example, number of entries available, number of null values, possible values for the attributes (domain);
- Analyzing the statistical profile of data: ex. max, min, mean values, histograms, boxplots, etc.
- Analysing the correlations between bmi and number of steps.

During the statistical analysis we could also identify some data problems, as for instance: outliers and noisy values, duplicated measurements values of steps and weights, inconsistent measurements (e.g., different weight/bmi values at the same timestamp, inconsistent distance/steps or calories/steps values, etc). Adderssing these problems to get a cleaner dataset is part of the "Data Preparation" activities.

Data Preparation activities includes the data cleaning mentioned above and a couple of other transformations, as for example:

- Date transformations: for instance, adding new fields (year, month, day) from timestamps, summarizing users measurements per month, converting birthday to age;
- Filtering weight unit measures: considering that the great majority of the weight measurements are in kilos, we selected only weights measured in kilos;
- Recomputing bmi from user weights and height: in order to fill missing values or fix possible inconsistent values;
- Merging bmi and steps datasets to get a bmi value for each steps measurement. When the corresponding bmi in not available at the same day as the steps measurement, interpolation is performed to fill missing data.

The findings of the exploratory data analysis and the data preparation process are presented in the links below:

- Users: Visualize statistical profile of user data, filter noisy data.
- Weights: Visualize statistical profile of weight data, filter noisy data, filter by metric (kilos), handle multiple measure at the same timestamp, compute bmi.

- Activities: Visualize statistical profile of activity data, filter noisy steps counts based on threshold, steps *vs* distance ratio, and steps *vs* calories ratio.
- Bmi, Steps: Merge bmi and steps data: each row of the merged dataset consists of user_id, measurement day (year, month, day), bmi, and steps. All measurements of steps are kept in the merged dataset. Missing bmi values are interpolated.
- Bmi, Steps correlation: Visualize correlation between bmi and number of steps considering all users, and inidividual users.

3.1 Body Mass Index and Steps Correlation

In this Section we discuss the correlation between body mass index (bmi) and steps, which is the central topic of this work. We first describe the analysis performed to investigate the relationship between steps and bmi, and then we present a brief summary of our findings.

3.1.1 Analysis Description: In order to investigate the relationship between number of steps and bmi value the following analysis were performed:

- Correlation Analysis: The Pearson correlation coefficient ($corr$) measures the strength and direction of a linear relationship between two variables, in our case the variables are *steps* and *bmi*. The direction of the relationship is given by the sign of $corr$: positive sign means that high values of *steps* are related to high values of *bmi*, while negative sign means that high values of *steps* are related to low values of *bmi*. The strength is given by the absolute value of $corr$. More specifically, in this project he $corr$ values are interpreted as follows :
  - $corr = -1.0$: Perfect negative linear relationship
  - $-1.0 < corr \leq -0.7$: Strong negative linear relationship
  - $-0.7 < corr \leq -0.3$: Medium negative linear relationship
  - $-0.3 < corr < 0$: Weak negative linear relationship
  - $corr = 0.0$: No linear relationship
  - $0 < corr < 0.3$: Weak positive linear relationship
  - $0.3 < corr < 0.7$: Medium positive linear relationship
  - $0.7 < corr < 1.0$: Strong positive linear relationship
  - $corr = 1.0$: Perfect positive linear relationship

  The Pearson correlation coefficient is symmetric: $corr(steps, bmi) = corr(bmi, steps)$.

- Cross correlation Analysis
  In the relationship between two time series ($steps(t)$ and $bmi(t)$), the series $bmi(t)$ may be related to past lags of the series $steps(t)$. The cross correlation function (CCF) is helpful to identify lags (h) of steps that might be useful predictors of *bmi*, i.e. lags that maximize the correlation between $steps_{t+h}$ and bmi_t, where $steps_{t+h}$ is the *(t+h)*-th value of *steps* and bmi_t is the t-th value of *bmi* in the time series.

In this project we want to investigate if we can predict *bmi* from previous values of *steps*, therefore we are particularly interested in negative values of h, in other words, we aim to investigate the correlation between *steps* at a time before t and *bmi* at time t. For instance, consider $h = -2$, the CCF value would give the correlation between $steps_{t-2}$ and bmi_t.
The correlation values are between -1 and 1 and have the same interpretation as discussed above in the topic about *"Correlation Analysis"* .

- Linear Regression Analysis:
  A simple linear regression model, defined as $bmi = a + b \cdot steps$ was fitted to the data. The model coefficients were estimating using ordinary least squares.

  A 95% confidence interval was computed for the model coefficients, which is interpreted as: if the population from which this sample was drawn was sampled 100 times, approximately those confidence intervals would contain the "true" coefficient in 95 of theses samples.

  A hypothesis test was performed with the following conventional hypothesis:

  - Null hypothesis: There is no relationship bewteen bmi and steps ($b = 0$)
  - Alternative hypothesis: There is a relationship bewteen bmi and steps ($b \neq 0$)

  The Null Hypothesis is rejected if the 95% confidence interval does not include zero. Otherwise the test fails to reject the null hypothesis. In order to test the Hypothesis the p-value is computed, which represents the probability that the coefficient is actually zero. The p-values are interpreted as follows:

  - $p\text{-}value <= 0.05$ means that the 95% confidence interval does not include zero, and the test rejects the null Hypothesis (indicating a relationship between steps and bmi)
  - $p\text{-}value > 0.05$ means that the 95% confidence interval includes zero, and the test fails to reject the null Hypothesis (indicating that there is no relationship between steps and bmi)

3.1.2 Analysis Results: A summary of the above mentioned analysis is presented the following.

- Steps and bmi values of all users: The goal is to investigate if the number of steps is in general correlated to the bmi of the users. In other words, is it true that users who walks more steps have lower bmi? The investigation considered:

  - *Correlation Analysis*: The Pearson correlation coefficient computed was $corr = -0.1068$, indicating a weak but yet negative correlation between the number of steps and the bmi value.

  - *Linear Regression Analysis*: The hypothesis test rejected the null hypothesis with $p\text{-}value = 0.0$, suggesting a relationship between steps and bmi.

- Steps and bmi values of individual users: The goal is to investigate if the number of steps of an individual user is correlated to their own bmi. In other words, is it true that increasing or decreasing the number of steps affects the bmi of a given user? And how is it affected?
  - *Pearson Correlation Analysis*: The Pearson correlation coefficient was computed for each user, who had as least 60 bmi/steps registers. The median and mean correlation values are very close to zero (-0.0026 and -0.0174, respectively). Only 3 users (0.08%) showed a strong and 206 (5.63%) a medium negative correlation between steps and bmi. Yet less users presented strong or medium positive correlation between steps and bmi, 2 (0.05%) and 144 (3.94%), respectively. The strongest negative correlation observed was -0.739, and the strongest negative positive correlation was 0.799. Excluding outliers, the strongest negative and positive correlations observed were -0.4160 and 0.378, respectively.
  - *Cross-correlation analysis*: Since the data being investigated are series, we also computed the cross-correlation function (CCF) between steps and bmi. As for the pearson correlation analysis, here only users who had as least 60 bmi/steps registers were considered. The maximum lag defined was $h = 10$, i.e., considering the set of steps/bmi measurements $((s_0, b_0), (s_1, b_1), (s_2, b_2), ..., (s_n, b_n))$, the CCF considers the correlations between s_i and b_{i+10}, b_{i+9}, ..., b_i, for $i = 0$ to $i = n-10$. For each user we searched the smallest correlation value from the CCF and the lag leading this value. The median and mean correlation values found were very close to zero (-0.067 and -0.078, respectively). Only 5 users (0.14%) showed a strong and 299 (8.18%) a medium negative correlation between steps and bmi. Even less users presented strong or medium positive correlation between steps and bmi, 1 (0.03%) and 43 (1.18%), respectively. The strongest negative correlation observed was -0.788, and the strongest negative positive correlation was 0.736. Excluding outliers, the strongest negative and positive correlations observed were -0.445 and 0.288, respectively.
  - *Linear regression analysis*: Hypothesis test were performed for the same users for which we computed the pearson correlations and CCF, i.e., users with more than 60 steps/bmi registers. For 25% percent of the users the null hypothesis was rejected and the bmi/steps correlation is negative, suggesting a negative relationship between steps and bmi for these users. The tests failed to reject the null hypothesis for a total of 39% of the users suggesting a relationship between steps/bmi (including positive and negative relationships) for these users.

A more extensive presentation of the relationships between steps and bmi, including charts and bmi predictions from steps can be found here.

4 Clustering

The users where grouped in a total of 36 groups according to their activity lebel, bmi and age.

- age: < 30; 30 to 59, $>= 60$ years old
- bmi: < 18 (underweight); 18 to 25 (normal); 26 to 30 (overweight); > 30 (obese)
- activity level: < 5000 (low active); 5000 to 10000 (active); > 10000 (very active)

We also experimented using *K-means* to cluster users according to their age, bmi, and steps. However we concluded that the above described groups give a more intuitive visualisation of a user's profile. Currently we use standard values to define ranges of age, bmi, and activity level. Another option would be to define the ranges based on the histograms constructed in the exploratory data analysis .

4.1 User results report

The last phase of the project was to provide a report for a given user showing the characteristics of their group, and how they compare to other users in the same group, through histograms and box plot charts. The bmi/steps evolution curves for a from their group (the one with lowest steps/bmi correlation) is also presented as an insight of how increasing the average number of daily steps in a month, could lead to a weight lost (bmi decrease).

Take a look at two examples os such reports:

- uset data not from the dataset
- user data from the dataset

5 Conclusion and Next Steps

Despite it is a common sense that sedentary lifestyle and overweight are highly related, it is not so obvious that we can infer the bmi of a user uniquely from their daily steps. First, the activity level of a person is not the variable determining their weight, but a many other variables are involved, as their dietary habits and their personal metabolism. Second, the number of daily steps is not enough to measure the activity level of a person, since there are many sports modalities that can not be measured through steps.

Considering the values of the cross-correlation functions between steps and bmi series, our analysis show that around 8% of the users have strong to medium negative correlations. Another interesting results is that for 25% of the users the steps/bmi correlation is negative and the hypothesis test could reject the null hypothesis, suggesting a negative relationship between steps and bmi.

It is worth to notice that during our "Data Preparation" activities, we interpolated the missing weight/bmi values, which could have lead to users with practically constant measurements. Therefore, as next steps it is worthy to investigate the bmi/steps relationship only for users without missing values.

As further development of this work, we would like to analyse the bmi auto-correlation to have an insight on how previous values affects future values of bmi. Next, we would apply lagged regression to predict next bmi values, from previous bmi and number of steps. By lagged regression we mean including (past) lags of bmi and steps to the regression model. This approach could be improved by including the residual error in the regression model, such that the predicted error can be subtracted from the model prediction to enhance performance. This can be accomplished by including a number of lagged error values in the linear regression model. Note that we have then an autoregression of the residual error time series, which is called a Moving Average (MA) model.

A more sophisticated approach are Artificial Neural Networks, in particular the Long-Short-Term Memory (LSTM) Recurrent Networks. A great advantage of this approach is that it can easily model problems with multiple inputs.

While we don't expect, and we don't intend to give precise predictions, since the bmi depends on a lot of other variables, we could present the user how their bmi is likely to develop in a near future, giving them the possibility to identify early enough any tendency of a bmi increase, and hopefully take the necessary actions to get/stay fit. With a similar purpose we could also create a model to predict the number of steps, warning the user for any tendency of a decrease on their steps counter, and to motivate them to keep active.

Predicting the bmi is not the only thing, a couple of other insights could be provided to the user though their personal data analysis, For examples, what are their most active days of the week? Are they more active on weekend/holidays than on week days? When was their most active period? Etc.

As for the user clustering, it is interesting to see that the age, bmi and steps distributions differ from one group to another, and to the overall distributions of all samples. As an example one could see that the bmi from all users looks normally distributed, but this is not the case for a specific group of overweight people.

The user reports gives an interesting view if distributions inside the user particular group as well as how the user compares to other people in the same group.

To summarize, this works gives a good overview of the bmi/steps data, and leaves a couple of open possibilities for future developments towards helping users of activity trackers to reach their weight/bmi goals.