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

Renata Rego
23 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 coaching insights for users to reach 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 simple, wide accessible equipment, are related to the increase or decrease of the body mass index (vmi), and wether 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. 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 Methods

2.1 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.

2.2 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 outliers and noisy values, and perform the necessary operations to clean the data.

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

- converting birthday to age;
- filtering weight unit measures: considering that the great majority of entries are in kilos, we selected only weights measured in kilos;
- recomputing bmi, in order to fill missing values or make sure it is coherent with the user's weight and height.

  The findings of the exploratory data analysis 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 correlation: Visualize correlation between bmi and number of steps considering all users, and inidividual users.

2.2.1 BMI Steps Correlation

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

- Correlation Analysis: The Pearson correlation coefficient ($corr$) is computed to measure the strength and direction of a linear relationship between two variables, which the following interpretation for the values of $corr$:
  - -1.0: Perfect negative linear relationship
  - -0.7: Strong negative linear relationship
  - -0.3: Weak negative linear relationship
  - 0.0: No linear relationship
  - 0.3: Weak positive linear relationship
  - 0.7: Strong positive linear relationship
  - 1.0: Perfect positive linear relationship
- Cross 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)

  A summary of the above mentioned analysis is presented the following.

(i) 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 that there is a relationship between steps and bmi.

(ii) 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?

- Correlation Analysis: The Pearson correlation coefficient was computed for each user, who had as least 60 bmi/steps registers. A summary of the results is presentes below:
  - The median, and mean value of the correlations are respectively -0.0026 and -0.0174
  - Only 3 users (0.08%) have a strong and 206 (5.63%) users have 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%) users respectively.
  - 25% of the users has correlation smaller or equals to -0.119
  - The highest negative correlation is: -0.739
  - The highest negative correlation excluding outliers is -0.4160
  - 25% of the users has correlation greater or equals to 0.080
  - The highest positive correlation is 0.799
  - The highest positive correlation excluding outliers is 0.378
- Cross-correlation analysis: Since the data being investigated are time series, we also computed the cross-correlation between steps and vmi. The maximum lag defined was 10. For each user we stored the smallest correlation and the lag leading this value. A summary of the correlation values found is given below:
  - The median, and mean value of the correlations are respectively -0.067 and 0.078
  - 5 users (0.14%) have a strong and 299 (8.18%) users have 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%) users respectively.
  - 25% of the users has correlation smaller or equals to -0.170
  - The highest negative correlation is: -0.788
  - The highest negative correlation excluding outliers is -0.445
  - 25% of the users has correlation greater or equals to 0.013
  - The highest positive correlation is 0.736
  - The highest positive correlation excluding outliers is 0.288
- Linear regression analysis: Data from two users were selected for the linear regression analysis. The user presented with the most negative

correlation coefficient, and the second presented the most negative cross correlation coefficient (with a lag of 6). In both cases the hypothesis test rejected the null hypothesis with $p\text{-}value = 5.3160e{-}118$ for the first user, and $p\text{-}value = 1.683e{-}30$ for the second user. The results suggest a relationship bewteen bmi and steps 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.

2.3 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); $> 10000 (veryactive)$

2.4 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 am 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

3 Conclusion