Practical Data Science II
2024-06-05

Kristianstad University
www.hkr.se

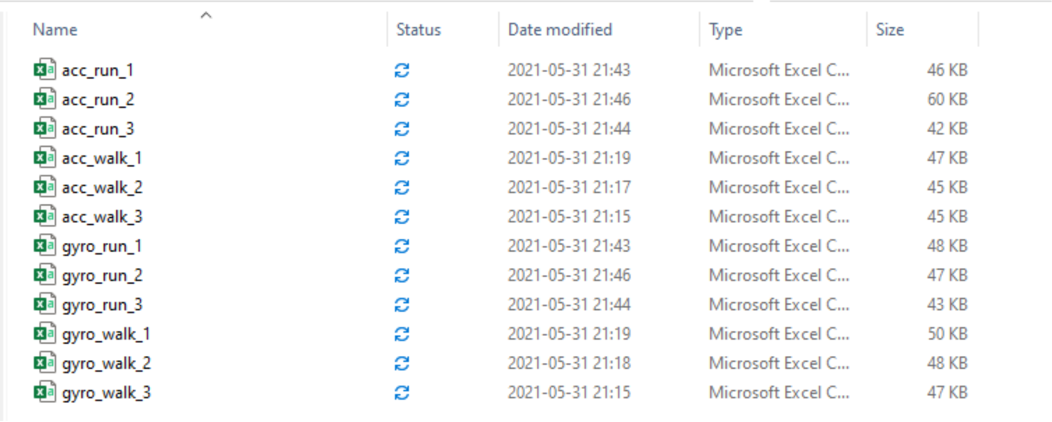
Fredrik Frisk
Computer Science Department

Content and purpose

The purpose of this practical is to give some experience of pre-processing data using Pandas. The datafiles we will analyse are movement data recorded from a mobile phone.

Movement data

You have access to 12 comma separated files (csv-files). In each file there is four datapoints on each row. First row is a timestamp, second to fourth row is the sensor data given, in the order of x, y and z-direction.



Name	Status	Date modified	Type	Size
acc_run_1	🔄	2021-05-31 21:43	Microsoft Excel C...	46 KB
acc_run_2	🔄	2021-05-31 21:46	Microsoft Excel C...	60 KB
acc_run_3	🔄	2021-05-31 21:44	Microsoft Excel C...	42 KB
acc_walk_1	🔄	2021-05-31 21:19	Microsoft Excel C...	47 KB
acc_walk_2	🔄	2021-05-31 21:17	Microsoft Excel C...	45 KB
acc_walk_3	🔄	2021-05-31 21:15	Microsoft Excel C...	45 KB
gyro_run_1	🔄	2021-05-31 21:43	Microsoft Excel C...	48 KB
gyro_run_2	🔄	2021-05-31 21:46	Microsoft Excel C...	47 KB
gyro_run_3	🔄	2021-05-31 21:44	Microsoft Excel C...	43 KB
gyro_walk_1	🔄	2021-05-31 21:19	Microsoft Excel C...	50 KB
gyro_walk_2	🔄	2021-05-31 21:18	Microsoft Excel C...	48 KB
gyro_walk_3	🔄	2021-05-31 21:15	Microsoft Excel C...	47 KB

The files contain sensor values for two types of motion, running and walking. In total there are three recordings, with six channels of data for each recording. Both acceleration (using an accelerometer) and angular velocity (using a rate gyro) have been recorded.

During this lab we will analyse the sensor data. We will look at the correlation and autocorrelation of some of the files in the dataset, as well as some statistics.

1. Import an accelerometer file and a corresponding rategyro file. Choose one of the files including **running**. Create *one* dataframe from these two files. You should have a dataframe with 6 columns, 3 acceleration and 3 from

the rategyro. Name the columns ax, ay, az, angx; angy, angz. Do the same with one file including walking.

2. Plot the data walking data and the running data.
   - Is there any difference between the data from the walking and the data from the running?
   - What is seen in the beginning and end of the recorded data?
3. Do the same for the remaining files. You should end up with six dataframes as a total.

We will discuss your result here.

Some statistics

4. Use the `.describe` function and investigate the data.
   - Compare the mean value and the median value for all features/attributes (that is the columns) for both walking and running. If you find a (“significant”) difference note which column and try to explain the difference.
   - Compare the mean value and standard deviation for running and walking. Is there any difference? Can you explain why or why not?
5. Use the `.boxplot`() method to investigate if there are any outliers in the data.
   - Compare the different features/attributes (columns)
   - Compare running and walking
6. Use the zscore method to investigate if there are any outliers in the data.
   - Compare the different features/attributes (columns)
   - Compare running and walking

We will discuss your result here.

Cleaning the data

7. Create a new dataframes from the old ones. Remove the beginning and the end so you have a cleaner dataframe. Plot these dataframes as well.

8. Use the `.describe` function and compare the statistics from the “raw” data and the cleaned one. Do the same analysis as above, that is

    - Compare the mean value and the median value for all features/attributes (that is the columns) for both walking and running. If you find a (“significant”) difference note which column and try to explain the difference.

    - Compare the mean value and standard deviation for running and walking. Is there any difference? Can you explain why or why not?

9. Use the `.boxplot()` method to investigate if there are any outliers in the data.

    - Compare the different features/attributes (columns)

    - Compare running and walking

10. Use the zscore method to investigate if there are any outliers in the data.

    - Compare the different features/attributes (columns)

    - Compare running and walking

We will discuss your result here.

Correlation

11. Create a correlation matrix for the cleaned and the raw files. Start with one walking file and one running file, that is in total 4 files, cleaned and “raw” as well.

    - Is there any difference between the cleaned and “raw” file?

    - Is there a difference between walking and running?

    - For which features do you find the highest correlation?

- Can you explain why?

Calculate the correlation matrix for the remaining files.

- Is there any difference between the different “raw” walking files (3 files)?
- Is there any difference between the different cleaned walking files (3 files)?
- Do the same for the running files

We will discuss your result here.

Autocorrelation

From here on we will only work with the cleaned data.

12. Plot the autocorrelation for all the six features (columns). Look at one walk file and one run file.
    - Is there a difference how far you can see any sign of correlation if you compare the six features? How far can you see the correlation. How many samples (lag)?
13. Is there any difference regarding the autocorrelation between the walking and running file?
    - Is there any difference regarding the autocorrelation between the acceleration and rategyro?
    - Extract the periodicity from the plot. How many samples (lag) is it between the repetition? That is the period of the data in the plot.
14. Given that the sample time is 20 ms. How long time does a period correspond to?
    - Investigate the running data. Is the period the same for all features, including acceleration and rotation (rategyro)?
    - Do the same for walking data.
15. What does the periodicity say? What can it be used to?

We will discuss your result here.

Rolling mean value

There is a method called `.rolling` that can be used for dataframes. To avoid lag we can center the mean value. This can be done as:

`rolling(window = #samples).mean()`

16. Experiment with different #samples. Can you make the curve look as good as for the autocorrelation? Compare the best features, that is the features that gives the best result. But use the same for the rolling mean value as for the autocorrelation.

    - Which window size is the best? And why is it the best?

17. Explain the difference between the mean value curve and the autocorrelation plot.

We will discuss your result here.

Extra - Record data by yourself

There exist several apps, both android apps and apps for iPhone that records data from the internal sensors and store the data, as well. For this part of the practical you need an app that can record the data and store the result in an csv-file.

You will find information how to use the “PhyPhox” app in the document named “Practical ML II - preparation - Data Recording App.pdf”. This is the one I have used but there exist others.

18. Record some different movements and analyse the recorded data. Use the methods shown above to try understand what kind of movement it is.