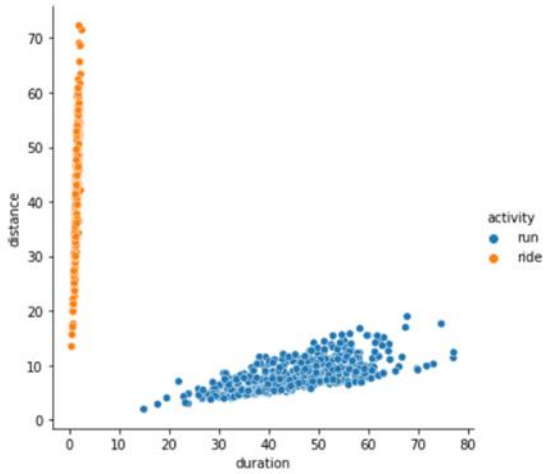
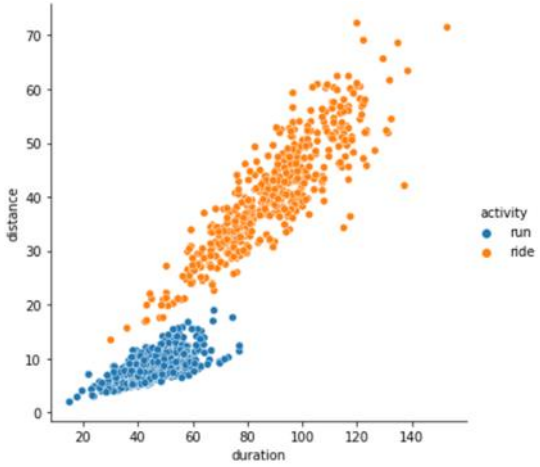
Part 1: Data Exploration

Data Cleaning

First I explored the data and I found a few irregularities and missing values in the data. First of all, I noticed that the bike rides lasted only one to two minutes (graph 1). This is unrealistically short and it is impossible to travel 70 km in that time. The users must have imputed their time for runs in minutes and their time for rides in hours. I propose to change the time for bike rides to minutes. I did this by writing a small program that multiplied the time by 60 if the duration was under 10. This results in much more realistic image, as seen in graph 2.



Graph 1. Original Data



Graph 2. Duration and Distance with Modified Times

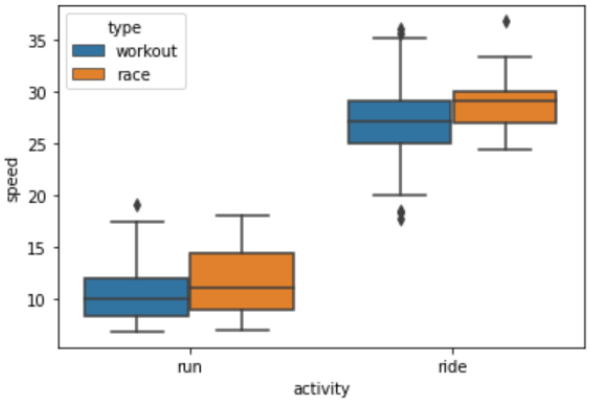
Thirdly, not all the dates follow the same format. Some dates are written with the month first: 01/03/2022 and some are written with the day first (fe. 03/01/2022). However, the dataset only contains data for January and February of 2022, so it is safe to assume that both of this dates refer to the 3th of January and not the 1st of March. I decided to write some code to make all dates fit the same template, namely month, day, year. The code first determines whether the first 2 digits are the month (namely smaller then 2 and the string is not '10' or '11') and switches the date and month when necessary. The date '01/02/2022' however might refer to the first of February or the second of January and might thus stay an error in the dataset. A last problem I found was that about 37% of the locations were missing values. Since users seemed to either always fill in their location or never it is hard to guess the locations. The missing values will thus remain missing.

The Average Distance

The average distance in this document is 23.82 km. However the distance for a ride and a run are very different. Therefor speaking about the average distancec travelled does not make a lot of sense, but saying

this duration does tell us much when we look at graph 2. Rides are generally longer. It thus makes more sense to say that the average ride is 40.77km and in the average run a distance of 8.04km is travelled.

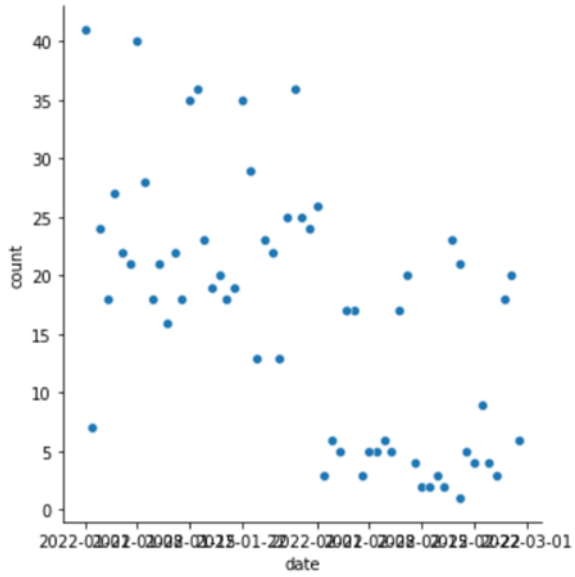
Races versus work-outs



People participated in races for both runs and rides. During those races participants seem to have put there best foot forward, as they on average run faster and cycle faster during the races then they do during their work-outs. Speed is here calculated by dividing distance by the time they took. This time is devided by 60 so the resulting speed is km/h.

Graph 3. Speed per activity

Do more people run after new year?



It is often said that people start sporting when the new year starts as a new year's resolution but they do not manage to keep doing that. There is an indication that this is also the case for this dataset as the numbers of activities registered per day drops dramatically. This is seen in graph On the 1st of January 41 activities were registered. The last day, 28th of February, only 6 activities were done.

Graph 4. Activities registered per day

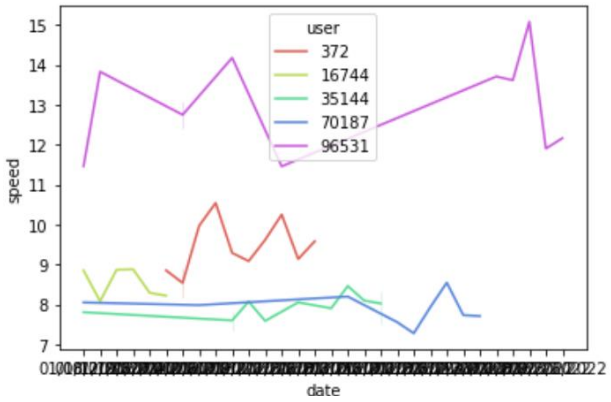
Are people who sport regularly different from the average?

	Top 5 users	Average
Age	19,67	25,21
Weight	72,27 *kg*	65,62 kg
% women	20%	57%

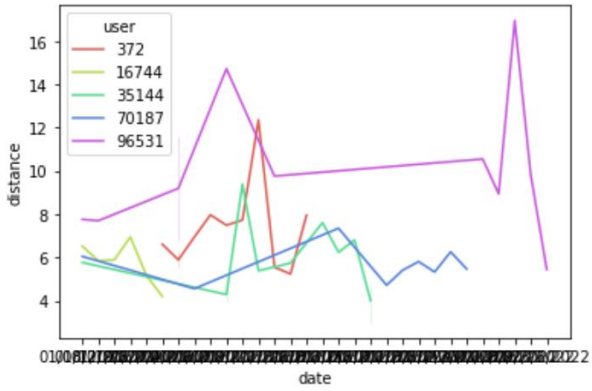
The 5 users who registered most activities make up 8,9% of the activities while only being 5% of the users. These 5 users are younger then than the average user, namely 19,67 years compared to 25,21 years. They are also heavier with averaging a weight of 72, 27kg instead of 65,62kg. However, this higher average weight is not surprising since there are more men in the top 5 users. In the 5 top users there is 1 woman and 4 men. In total there are 100 people in the dataset and 57% of them are women and 43% men. So taking gender into account the average man in this dataset weighs 71.73 kg and the average woman weighs 60.64 kg. The 4 men in the top 5 have an average weight of 72.56 kg, which is pretty close to the overall male average. The woman in the top 5 weighs 71.1 kg, which is considerably more than the average woman. To answer the question, the top 5 users are different than the average user in regards to age and gender.

Do People get Better?

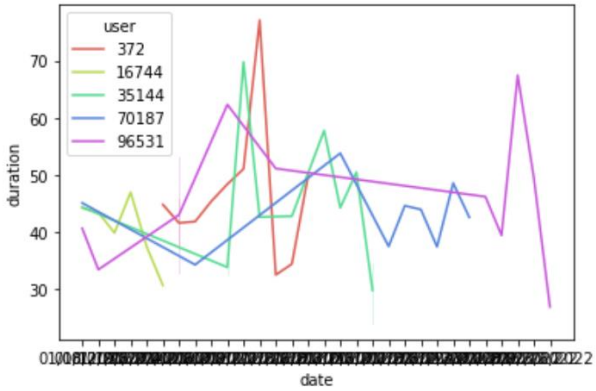
I was curious whether people improved over time. Improvement can be seen in getting faster over time, running longer distances or doing longer runs. I chose to check the patterns for the 5 most active users as they did all did a decent amount of runs. In graph 5 we see that the tempo varies widely between users, more than it does between activities from 1 user. Over a period of 2 months we do not see remarkable change in the speed, distance (graph 6) or duration (graph 7) of the runs of the top 5 users. This might be do the fact that more activities are needed to see change or change might simply not have been the objective of these runners.



Graph 5. Speed per workout per user



Graph 6. Distance per workout per user



Graph 7. Duration per workout per user