Anthony Sainez
asainez@ucmerced.edu
2022846322 / 100341001

Data Mining Project I
Data Exploration

Note

A lot of the assignment questions were very closely related, so instead of submitting an itemized report, I am submitting an essay-like paper that should answer all the questions when taken holistically.

The most critical dataset provided is device uplink information. When we want to examine individual users or uplink data over a range of users, we can relate the device uplink database to the user information database with the foreign key of `owner_id`. Here's the breakdown for the types of data in the uplink database.

Column	Dtype	What kind?
uplink_id	int64	Numerical
owner_id	int64	Numerical
client_time	datetime64[ns]	Numerical
tag_id	int64	Categorical (20 categories)
step	int64	Numerical
battery_low	int64	Numerical
is_charge	int64	Categorical (binary)
tag_battery_low	int64	Categorical (binary)

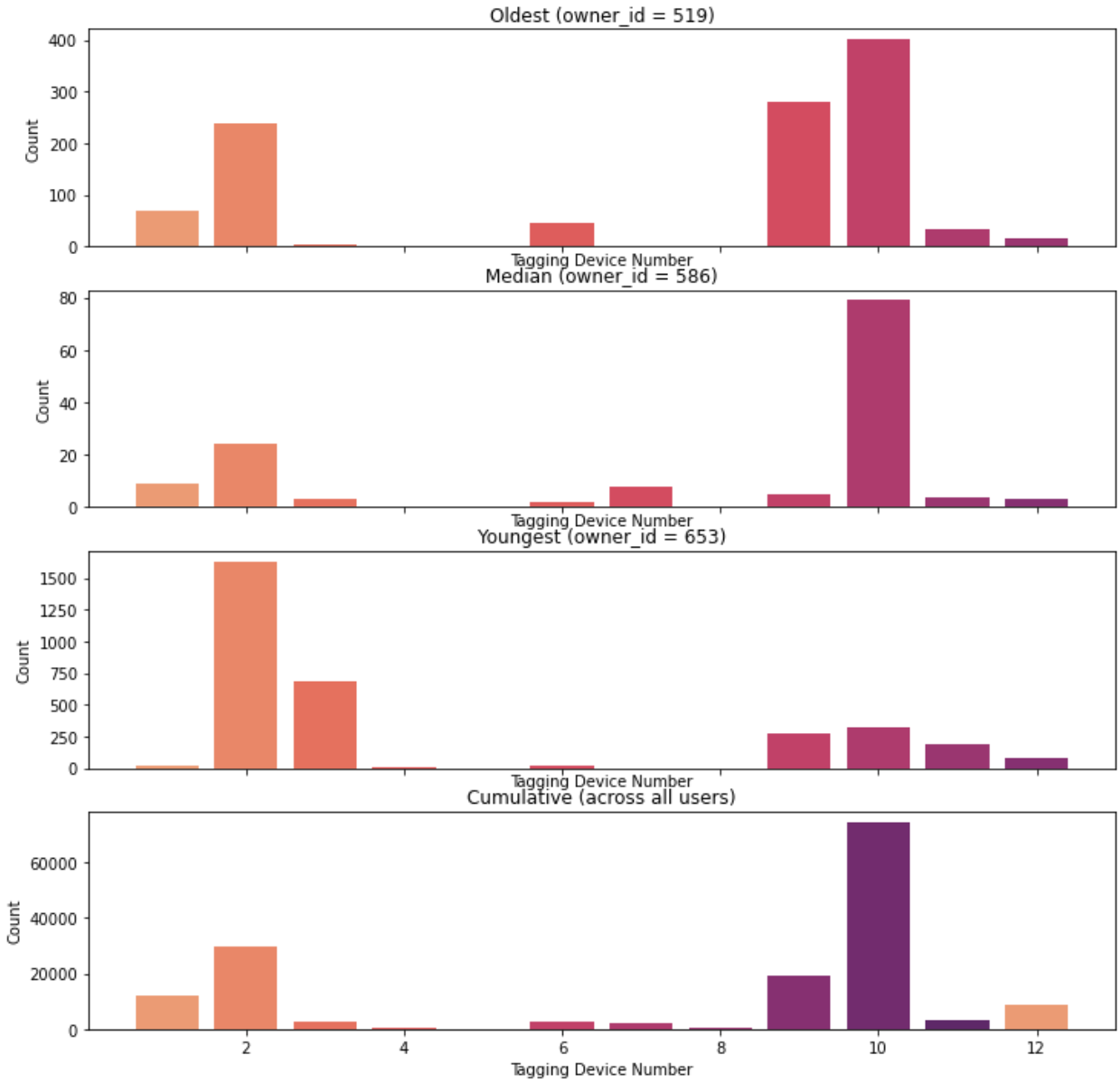
During sampling I decided to look at the youngest, oldest, median aged users, and finally all users cumulatively for a more holistic perspective. I chose age because the range spans twenty years and seemed like a good starting point; there is otherwise not very much personal information provided per user. The next variable to sample over and compare data would be sex, e.g., do male users have different behavioral patterns than female users?

Examining the raw data as-is, I saw that there were 324,823 entries in the sensor data. However, 166,034 entries have a `tag_id` of `0`. If you consider this invalid and remove such entries (as well as `tag_id` of `NaN`, `18`, `19`, and `20`, which are also considered invalid), then there were only 155,773 entries left, which represents a 52% decrease in size. Although this is an alarming amount of data to delete, this removed data either had no patterns of information in them or were deemed OK for removal based on domain knowledge. If we run `uplink['tag_id'].value_counts()`, we can observe that we don't have information for tags: `13` (fitness equipment), `14` (desk), `15` (living room), `16` (vacuum cleaner), and `17` (washer). This is technically missing but can't be handled without just collecting more data.

As you can see from the below figure, it looks like `tag_id` of `4`, `5`, and `8` are completely irrelevant in all cases, and so these columns can safely be eliminated. The same cannot be said for `3`, because it seems to be important in the case of the youngest user. In the case of `10`, we can see it has a high value in nearly all cases. I do not, however, think this value is an outlier or overrepresented. This makes sense from a domain knowledge perspective, since the

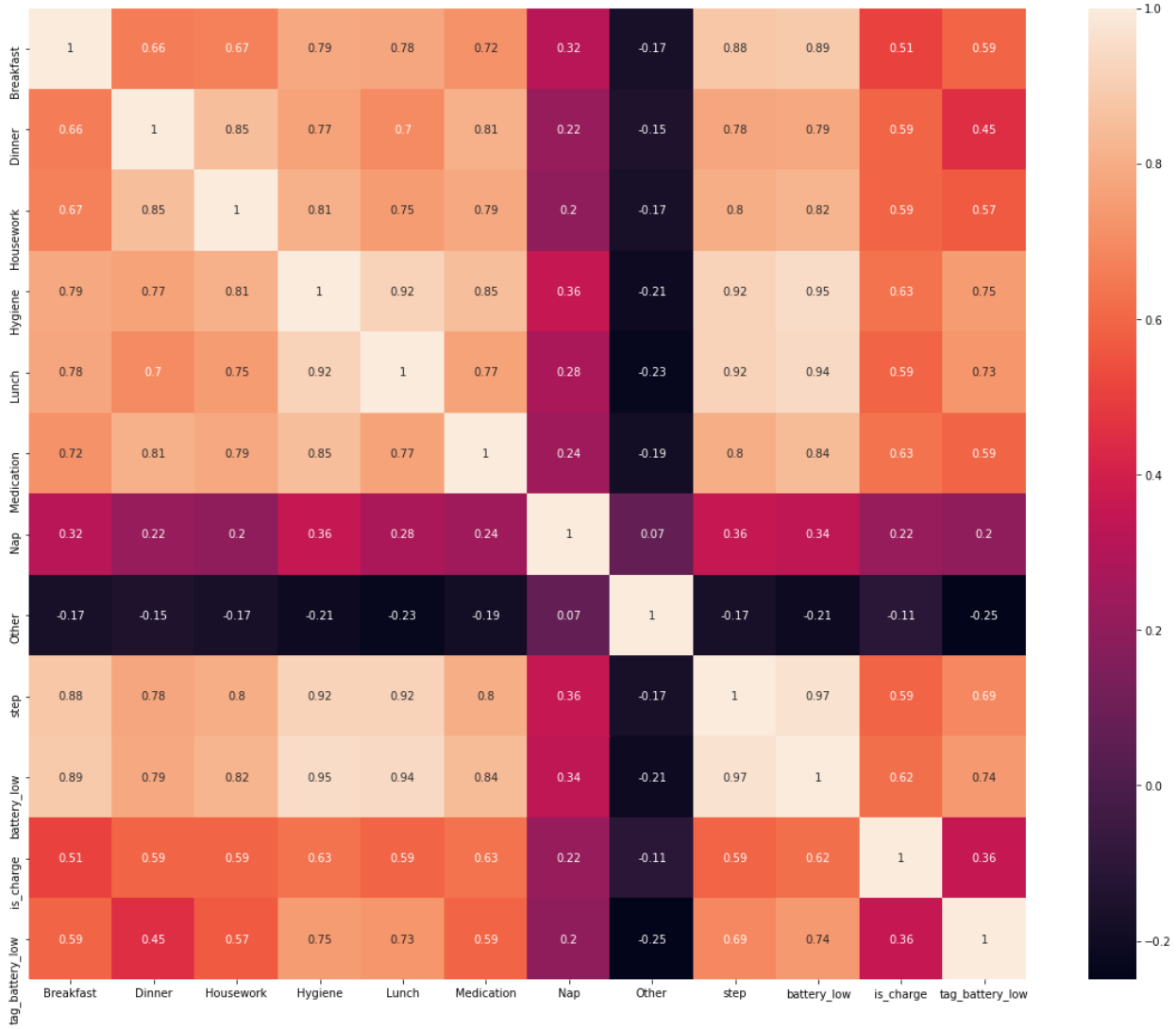
microwave might be vital for making all meals for an eldery person assuming they have no other cooking alternatives.

In the case of the youngest person in the data set, they used the microwave far less than the median and eldest. Unlike the others, the oldest in the group used both “kitchen” and “microwave” a high amount. Because of this, I would need more information as to whether or not “kitchen” and “microwave” can be considered duplicate data. There seems to be different patterns per user.



The next step is to put these device tagging events in more human-readable terms via recontextualization of raw data. For this specific data set, that means I had to think about *when* devices are tagged and *what sort* of activity that tagging event suggests. I was able to come up with the following action categories on my first pass: breakfast, dinner, housework, hydration, hygiene, indoors recreation, lunch, medication, nap, other, outdoors recreation, and sleep. It yielded the following correlation matrix (heatmap) over all users.

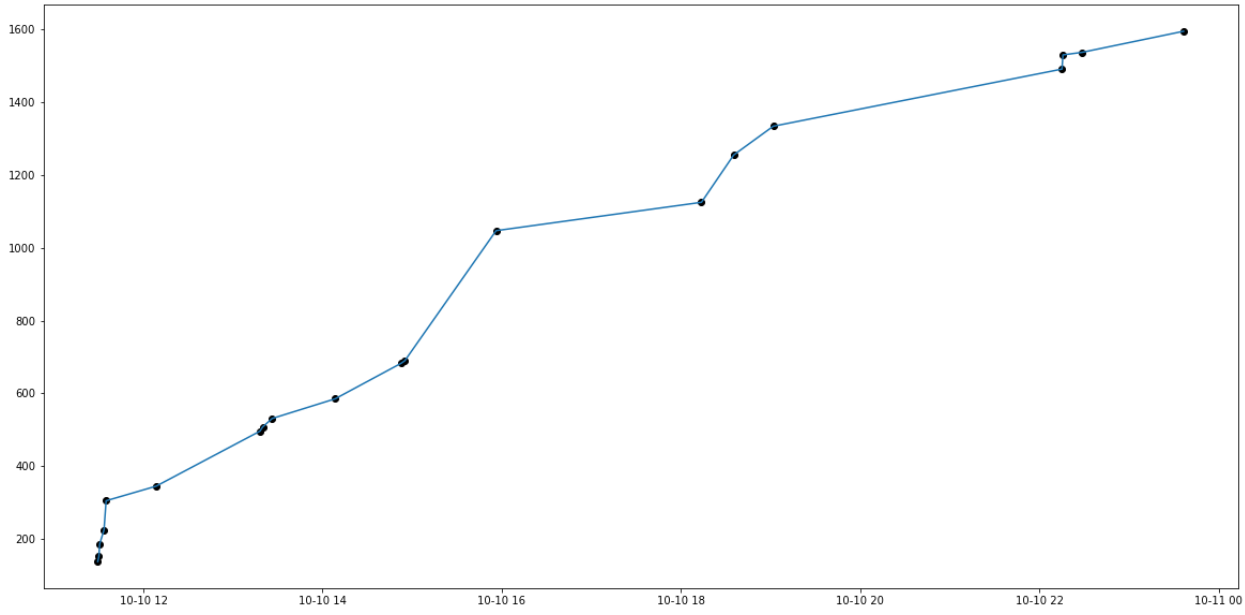
I thought my initial heatmap was incredibly complicated and not exactly enlightening, so I decided to simplify things by removing a few categories: I grouped the three categories (hydration, outdoors recreation, and indoors recreation) together into a new “other” category and produced the following new heatmap.



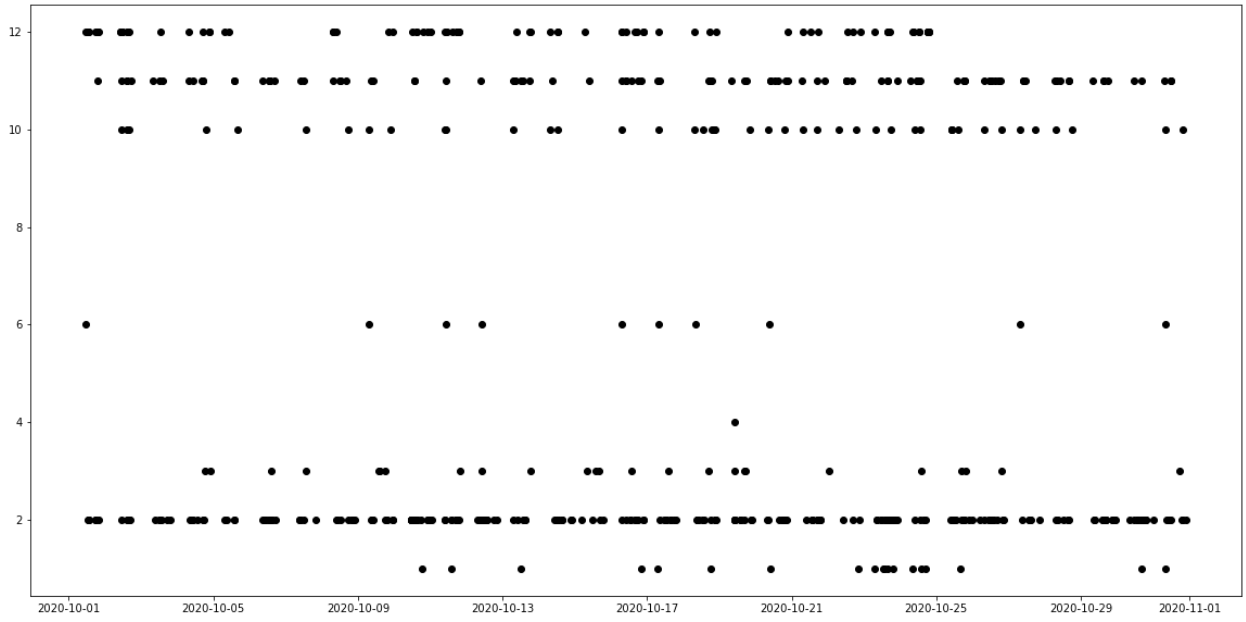
When it comes to concerns for dimensionality, this process of narrowing down which columns (or in this case devices) are relevant is important. In future parts of the data mining process, it could make the model training and testing process simpler (less variables means less complex) and therefore faster. The process of reducing dimensions in this case involves identifying which tags appear in significant quantities or which tags contain duplicate information (appear the same amount and same time as another tag). Using that information you can choose to keep less tags than the original raw data and build a more concise and lean database.

Side note, I do think there is something worthwhile in trying to normalize some of the values in the dataset since they seem to be underrepresented. I haven't gotten the chance to do this because of time constraints, however it could reveal new patterns in the data exploration process.

Next, I made some other nice visualizations using the median user's device uplink data. For example, what does a randomly sampled user's step count look like in a day? This graph helps confirm my hypothesis that these users might be pretty sedentary. 1,600 steps in a whole day is admittedly not very many, but it makes sense given the domain knowledge: these are eldery people confined mostly to their home and also are participating in some sort of data collection regarding life-logging, so they must already spend most of their time inside their home.

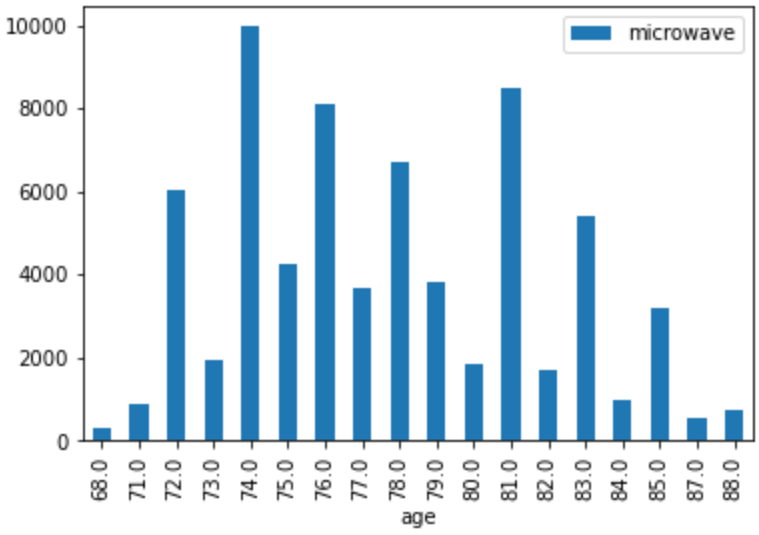


The next visualization I tried for this sample user was to understand their device usage over about a month's time. I chose a specific range which seemed to have a lot of device usage, compared to others which were relatively more sparse and had less data to work with. This visualization helps understand whether or not there is a routine regarding specifically which devices users interact with and at what intervals.



From the above graphs, we can see that their medicine usage is on a relatively consistent but sparse pattern. Checking this user's data, we can see it notes "차상위, 혈압, 당뇨, 관절," which mentions diabetes and blood pressure. It could be possible that this user is taking medication to treat these symptoms on a periodic basis. If you wanted to scale this graph to a single day you could take a look at how many times the user goes to the toilet in a day, which can be specifically useful for medical diagnosis as sometimes infrequent or too frequent usage of the bathroom can be considered a medical symptom indicative of something else. Additionally, we can make the inference that this user is mostly sedentary.

I attempted to create a bar chart showcasing the differences in microwave usage based on age. The distribution almost looks normal, but I suspect there might be differences in the distribution of age itself, e.g. there is only one 71 year-old, but there are two 72 year-olds, which explains why this graph is not as helpful as I thought it might be. I think it requires normalization.



But otherwise, this graph might suggest that there is not as much correlation between age and microwave usage as I hypothesized. I would need a wider range of data to be certain, since the sample size is admittedly pretty small ($n = 52$). This investigation strays a bit from the intentions of the original study (or whatever generated this data we're working with), since there is obviously more of a focus on the device uplink data than the user information.

For developing a service based on this information, I specifically thought of the "Meals on Wheels" program in America, which is a service to provide cooked meals to the eldery and combat social isolation. For one, it may be the case that as users approach greater age, kitchens become more inaccessible and microwaved meals become the only option. I theorize that increased consumption of microwaved meals positively correlates with depression levels, however I would need the data on depression levels to examine this hypothesis more closely.

Additionally, non-usage is another observation we can make. In our sample user, they didn't use the remote controller a single time in a month. While this might be this specific person's habit, perhaps this is an area in which the services provided by the remote controller can be improved? The "improvement" can be measured possibly in the percent increase in usage of this particular device.

Generally, being able to make observations about the habits and patterns in a user's daily life in their home is generally a good starting point for many services. The heatmap is especially good for identifying specifically which routines might need the most attention. The parts of their lives that see heavy daily usage are especially important for any person's quality of life, but especially so when you reach into your older years and everything becomes harder.