Week 2: Exploratory Data Analysis

Basic exploratory data analysis of data; discussion of preprocessing techniques needed
● Have your dataset uploaded to your workspace
Yes
● Do some basic exploratory data analysis (EDA) on the data: type of EDA will depend on data type, such as continuous, ordinal, image, or audio data
Timeseries

● List insights into data preprocessing techniques needed.
We have large parquet files and for that reason we used "dask" for data preprocessing. We also took only a fraction of the data (0.01) and did our EDA on it.

We also looked at the number of total target values presented in the whole dataset for each household in the UK, AUS, GE.

We discussed batch processing for the synthetic data in order to handle the large amount of data (several GB)

Examples:
1. Should all data be included in the model or is inclusion / exclusion criteria needed?
- **We will focus on the UK data for now, since the synthetic data is trained on UK data.**
- **We should exclude outlier households with usage profiles that are very different from most of the households.**
- **We will look for a time period during which we have the most number of households with normal-looking usage profiles and limit our training on these households during this time period.**
- **We should only use data of 30 minute granularity.**

2. What types of transformations / preprocessing needs to be done before they can be put in the model?
Same as answers for #1.

3. How much data will go in the train / valid / test sets? Is this enough that deep learning won't overfit (it overfits way more easily than traditional ML)
80-20 split, preserving temporal order