# D1

Hello, I am going to present to you our progress on our Data Mining: Energy Consumption in Oulu.

# D2

We based our work on the following problematic: How do the buildings' characteristics (their floor count, their gross area, or their location) impact their yearly energy consumption and can we cluster some different consumption profiles (and do we get similar results between the first analysis and the clustering)?

# D3

## Data set

The data is composed of 2 documents:

One which is the buildings’ metadata, that describes their characteristics: it gives some information about their floor count, their gross area or the year of construction.

The other is the buildings’ hourly energy consumption records, which describes the recorded heat and electricity consumption on a certain time window. Those records have to be retrieved from an URL.

# D4

To merge these 2 documents, we can use a common attribute, which is the buildings’ id. We simply will have to apply a join function to do so.

And here is some characteristics of different buildings, some of them or not visible here, but there are 16 of them.

# D5

Next, we have the consumption records, you can see the property id again and the consumption value and measure type, and the time where it has been recorded.

# D6

As we could expect, the dataset has to be cleaned for the following reasons.

## Data Cleaning

Some outliers were noticed, although it doesn’t impact the data that much.

However, a lot of data is missing, some is redundant, untranslated or the used format are not optimal.

# D7

## Missing Data

By exploring the dataset, I quickly noticed that some records were missing. Indeed, on the 536 buildings in metadata, only 286 actually have energy consumption records. So, we have to completely ignore the other 250 buildings.

# D8

## Data transformation

The metadata don’t need that much transformation as it is just a list of characteristics, the intended uses had to be translated. However, some information is missing for a lot of buildings, we can only manage them downstream. For example, a lot of building’s floor count is 0. And so, it is hard to predict why is that so, but if we look at its records, most of the 0 floor count buildings have invalid or mostly missing records so we can not really study them but neither can we be 100% sure that they will be uninvesting. We will see that in the next slides.

And for the consumption’s records, they also needed quite some transformation. First, the energy type (heat and electricity) had to be translated (go D5). The time was also noted as 4 different attributes (hour, day, month and year) which make the records hard to classify chronologically. So, I decide to use a datetime library which allow us to compare the times easily. The URL with which we retrieve the records is taking years as time parameter and so it had to be narrowed as an hour/day/year format.

# D9

This is what the data frame looks like after the transformation, the consumption measure is in English and the time format is just on one column.

# D10

## Feature extraction

An important objective of our project is to look for correlation between building’s characteristics and their energy consumption.

So, this is why I found interesting to split the buildings’ characteristic in 2 or 3 different types of features to search for these possible correlations:

Categorical features: Intended uses, postal code district name

Semi-continuous features: The gross area or the volume

Discrete features: Floor count, Year of construction

I think this last one is a bit tricky because some of its characteristics can be classified as categorical or semi-continuous according to the wanted precision.

# D11

### Categorical features

I started with the categorical features as so:

First, we need to create a list of characteristic categories, here we are looking at the building’s intended use. For example, we will have this list educational buildings, Kindergartens and Health centers.

Then we browse this list and retrieve the average energy consumption of a certain number of buildings per category by merging the metadata and the consumption records by using the property id. At the same time, we are checking if we can actually retrieve the consumption, it happens that the records are partly missing or invalid (for example, here we can see that heat measurement is missing for the selected educational buildings).

Finally, we group these means by category to have a wider viewpoint of energy consumption by category, here by intended use.

And we can display these results with graphs to visualize them.

# D12

So here, on this new example, we can see the proportion of the buildings we couldn’t retrieve an average consumption measure (in blue), classified by floor count. The gray bars indicate the remaining valid data. And so, this graph allows us to know if the results are consistent or not. If they are too much missing data, the results could actually be inconsistent.

# D13

And finally, the average consumption visualization. Here can see the average heat (in orange) and electricity (in blue) classified by floor count. And we can clearly see tendency to consume more energy as the floor count increases. Also notice that we went from a sample of 136 records to 60.

# D14

## Next steps

About the next steps.

Now that I studied categorical features, I have to find a way to split the discrete and semi-continuous features (again, for example: gross area, year of construction) in classes to get similar results.

# D15

I am currently working on it but it would look like something like this graph. A curve for every category, here for two intervals of gross area (one is between 50 and 300 square meters and the other on between 300 and 500). And instead of simply print the total hourly records, we reduce the time window to a monthly average. Unlike in the previous slide (D9). This would allow us to have a more precise understanding of the consumption of buildings according to their characteristics.

# Back to D14

I will also have to narrow down the gross consumption average into a monthly or yearly average to observe the changes through time and by category and a way to get easily readable results.

# D16

## Conclusion