Martin Kostelník, Marianne Jakonen, Ahmed Ansari

1 Introduction

The dataset[1] contains data collected from 54 sensors deployed in the Intel Berkeley Research lab. The collection period ranges from 28.2.2004 to 5.4.2004 and contains a total of roughly 2.3 million data samples. Each sensor collects data every 31 seconds and the data collected are: temperature, humidity, light and voltage. Each sensor also has a unique ID and stores the data with a timestamp and measurement ID.

2 Modelling goal

As per the assignment, the end goal is to use PCA to detect seasonality in the data. This will be done on both weekly and monthly basis. We will also study the correlation between the individual variables and create a dynamic model to predict future values for the most correlated variable. When we have this model, we can calculate how much ahead we can predict, it's window frame and the sampling frequency.

3 Data summary

As described in section 1, we have roughly 2.3 million data samples from 54 sensors placed at different location. The scheme of the data is shown in table 1.

Name	Data type	Unit
Date	String	-
Time	String	-
Epoch	Int	-
Mote-id	Int	-
Temperature	Float	°C
Humidity	Float	%
Light	Float	Lux
Voltage	Float	V

Table 1: Table showing column names, data types and units in the dataset.

When loading the dataset, we transform the `Date` and `Time` columns. We combine these into one column containing both using the correct datatype from pandas library.

3.1 Value ranges

Now let's take a look at the value ranges of each column. We can find theoretical value ranges for some of the columns in the dataset documentation. There are no limits to the temperature column. We also find the actual value ranges present in the dataset for each column. These can be seen in 2. We can see that for some columns, there are non-sensical values that need to be removed and this also shows that there might be many outliers in the data.

[1]Dataset publicly available at: `http://db.csail.mit.edu/labdata/labdata.html`

Name	Min_T	Max_T	Min_R	Max_R
Date	28.2.2004	5.4.2004	28.2.2004	5.4.2004
Time	00:00	23:59	00:58:15	11:02:32
Epoch	0	-	0	65535
Mote-id	1	54	1	65407
Temperature	-273.15	-	-38.4	385.568
Humidity	0	100	-8983.13	137.512
Light	0	-	0	1847.36
Voltage	2	3	0.009101	18.56

Table 2: Table showing theoretical and actual value ranges in the dataset.

3.2 Distribution

There are four columns that contain numerical values. These are temperature, humidity, light and voltage. For now, we are only going to be interested in those columns. Figure 1 shows kernel density estimation across the value ranges of each of these columns. Note that data were cleaned of outliers behind the theoretical ranges. Figure 2 uses boxplots to visualize the distribution. We can more clearly see where the mean sits and it seems to be around the usual values for all variables.

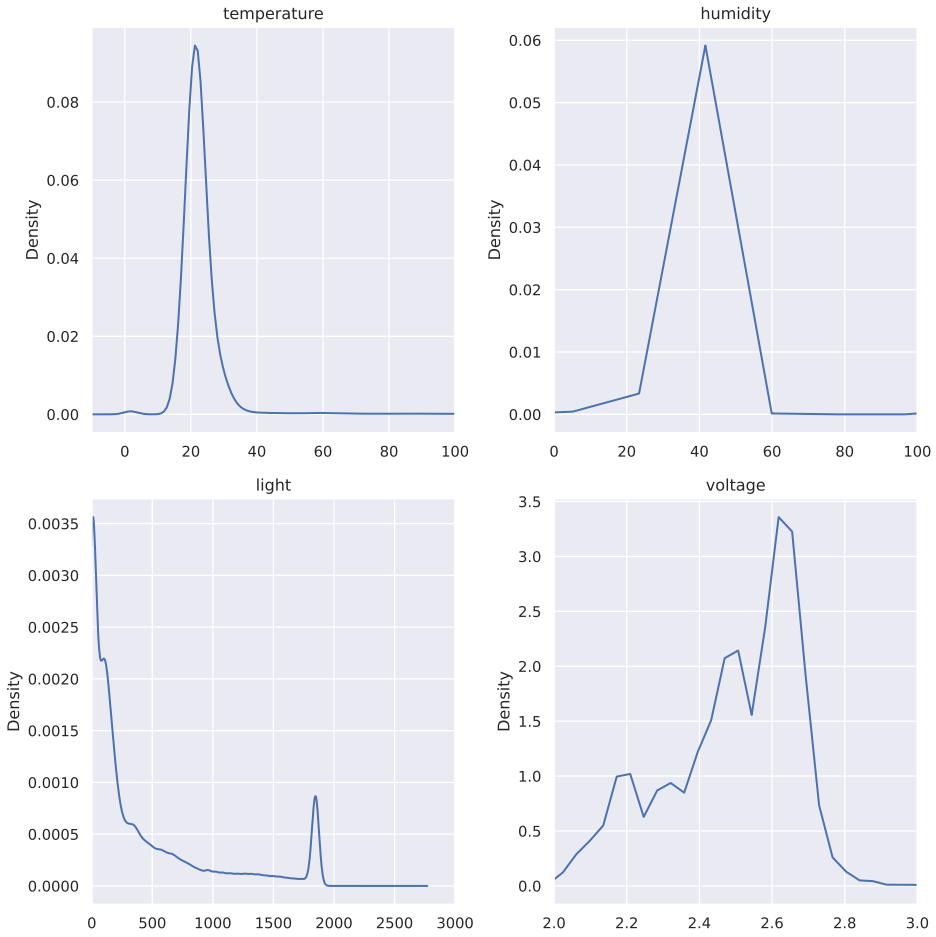

Figure 1: Kernel density estimation for the numerical columns of the dataset.

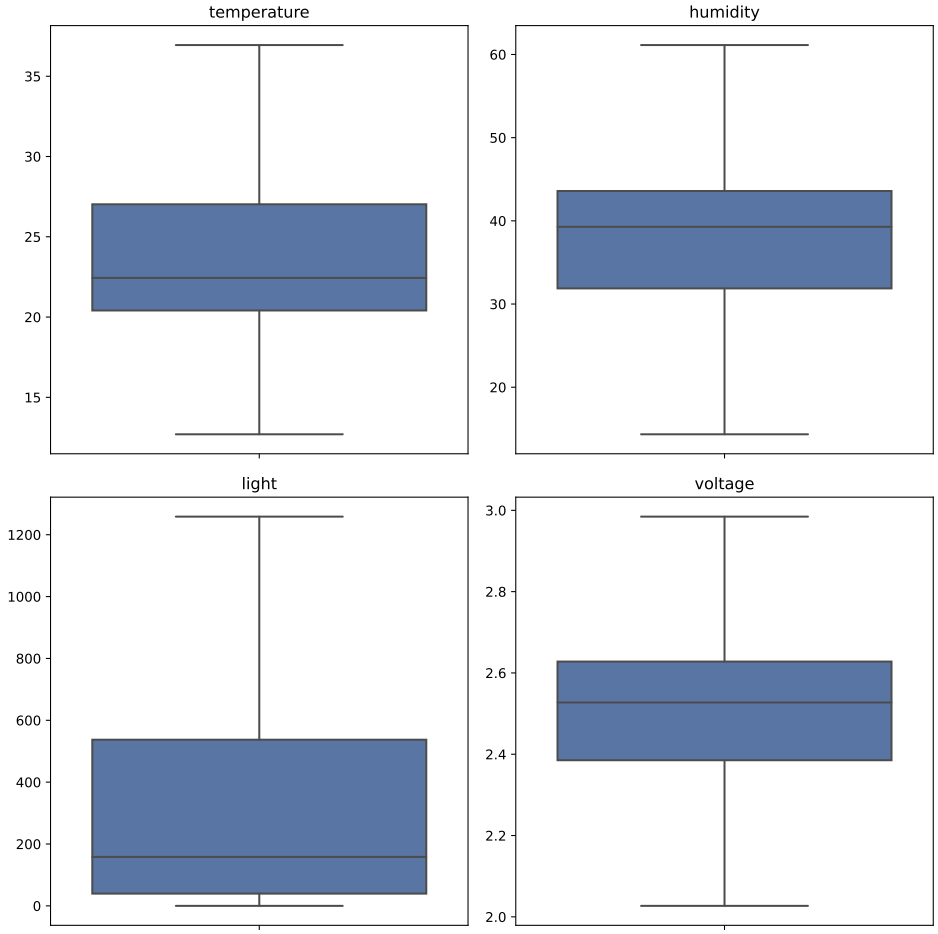

Figure 2: Data distribution visualized using boxplots.

The last figure, figure 3 show the trend in the data over the measuring period. To create this figure, the dataset was averaged over 8 hour periods. We can see a spike in temperature and change in the humidity at around the same time. Light amount seems to be changing based on the daily activity in the office. And lastly, as expected, the voltage in the batteries has a downward trend until the batteries are probably replaced.

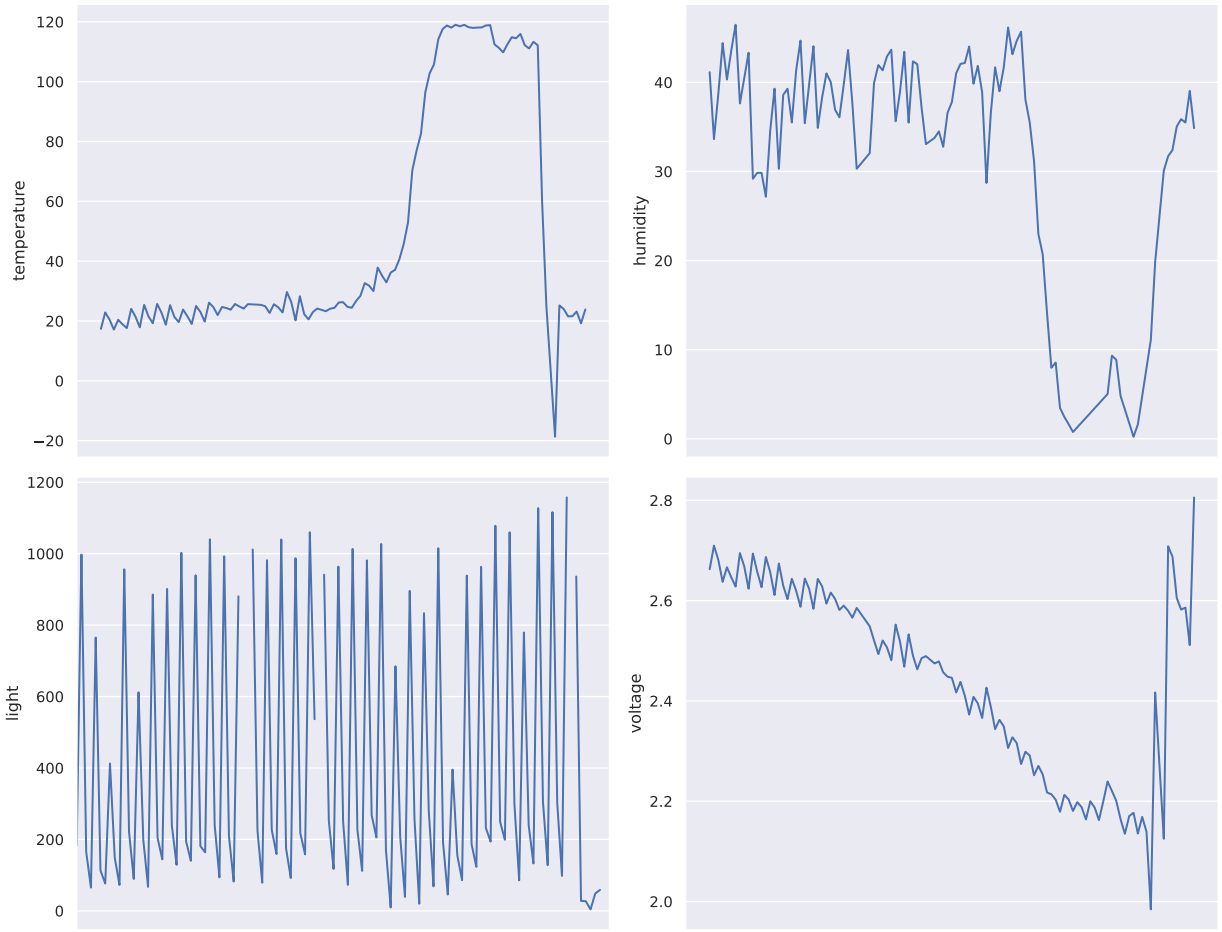

Figure 3: Time series of the data averaged over 8 hour time periods. Note that we only have light data for a shorter period of time than the other variables.