## Business understanding

- A real-life chemical detection platform is exposed to a mixture of carbon monoxide (CO) and humid synthetic air in its gas chamber. The platform is composed of 14 temperature-modulated metal oxide semiconductor (MOX) gas sensors.

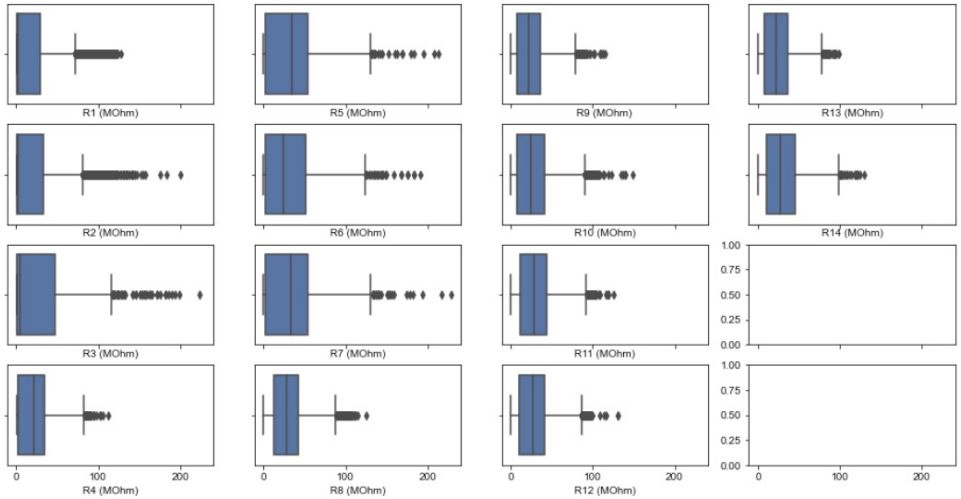
- We count with 13 datasets with 20 variables.

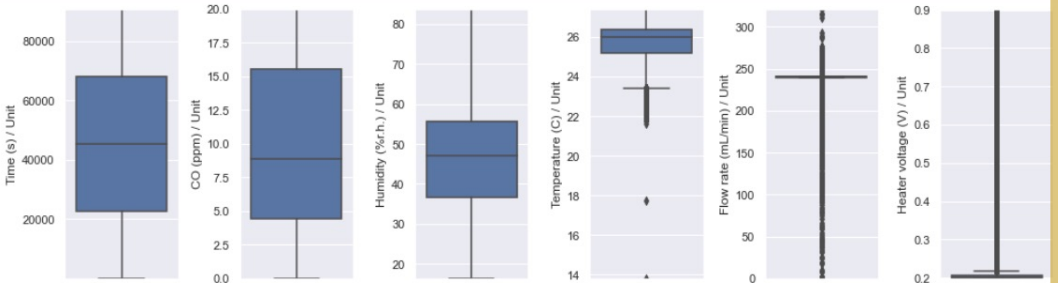
- In this work, an exploratory data analysis of the dataset is going to be made as well the evaluation of several machine learning algorithms that will be able to predict the presence of CO (ppm) in the gas chamber.

Box plot

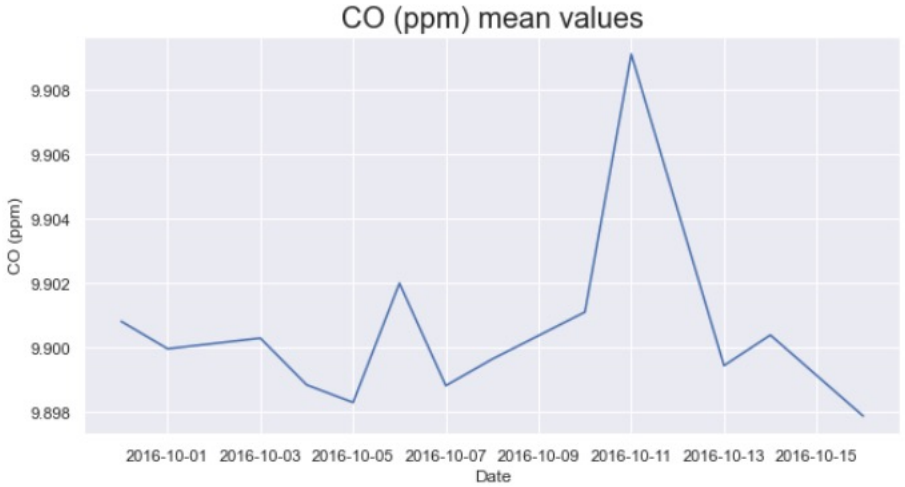
# Exploratory Data Analysis (EDA)

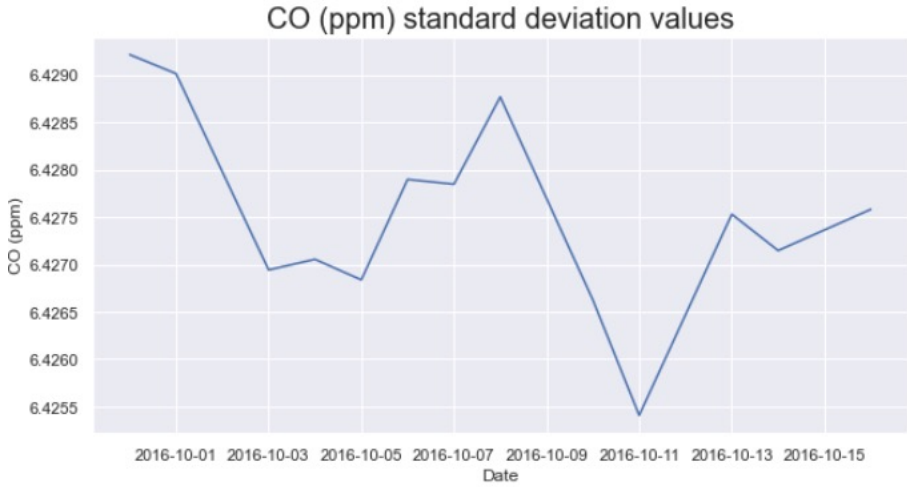
1. The resistances from R8 to 14 look very alike with an apparently normal distribution.
2. R1, R2 and R3 has a positive skewness, and the resistances from R4-R7 have a normal distribution.
3. The time, CO, and humidity have approximately a normal distribution.
4. The temperature has negative skewness.
5. The mean of the flow rate variable is equal to the constant released specify in the experimental protocol.
6. The heater voltage variable maintains the limits recommended by the manufacturer.



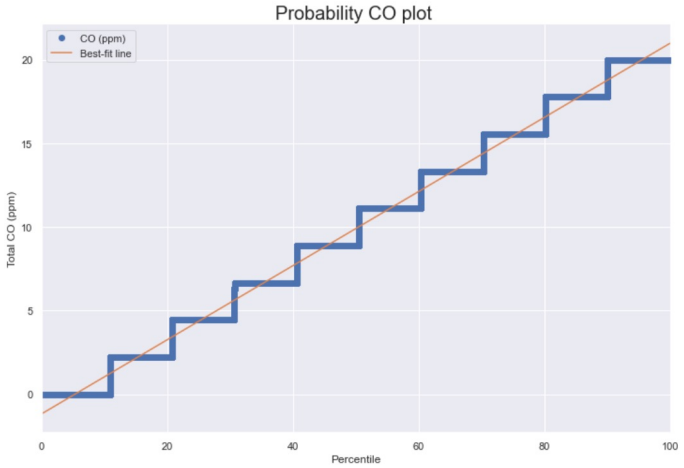
# Mean and standard deviation plots

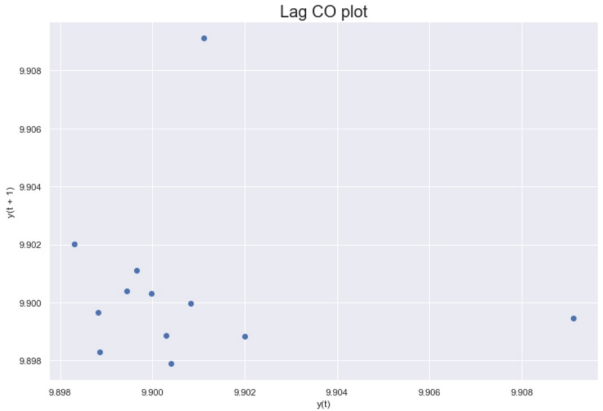


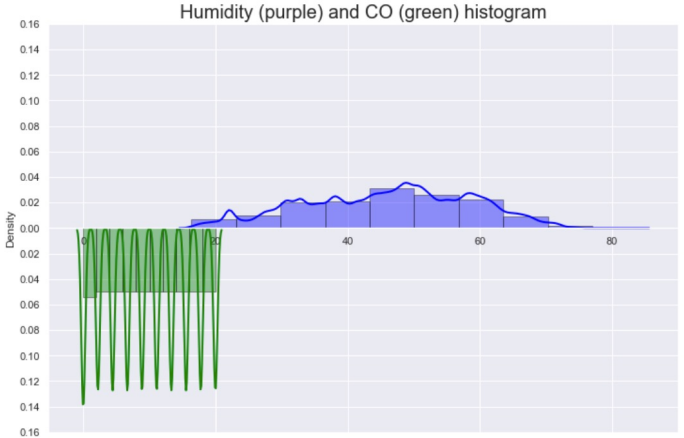




- The mean and standard deviation are very similar over time.
- On November the 10th the mean reaches its highest point, so the standard deviation is low, which mean that most of the values are closer to the mean value.

# Distribution plots

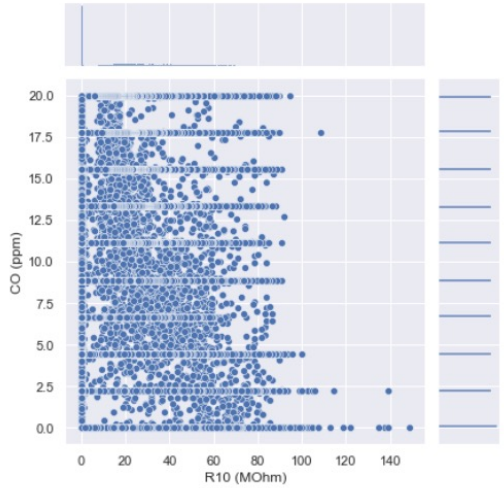



- The lag plot demonstrates that the variable CO hasn't any seasonality. The data is not randomness, almost all the data is concentrated almost in the same place.
- The probability plot and paired histogram show that the CO values have a normal distribution.
- The humidity data has a normal distribution. It demonstrates that the relative humidity randomly chosen were in fact chosen in a distributed way.

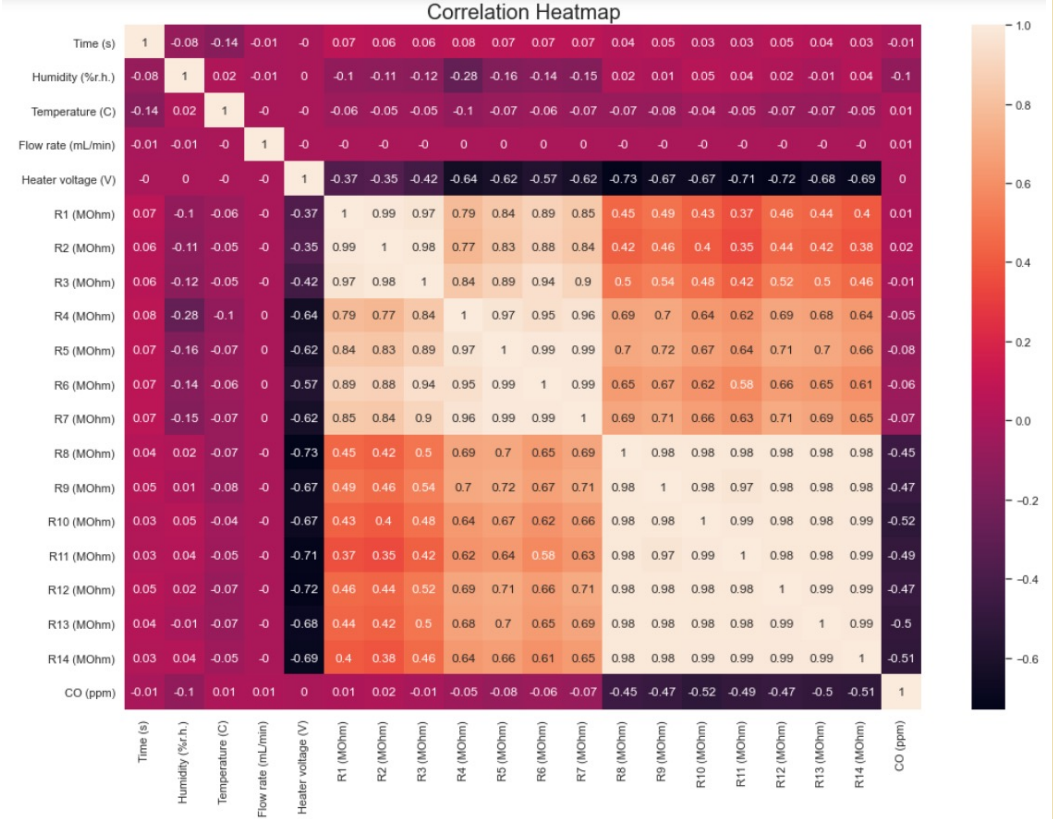
# Correlation plots




1. There are a high correlation among the gas sensors from the same distributor.

2. There is a high negative correlation between the CO and the gas sensors from R8 to R14 (one distributor). The CO variable has a small correlation with the other seven gas sensors.

3. When an instance has a high CO value, the R10 is low (in most cases).



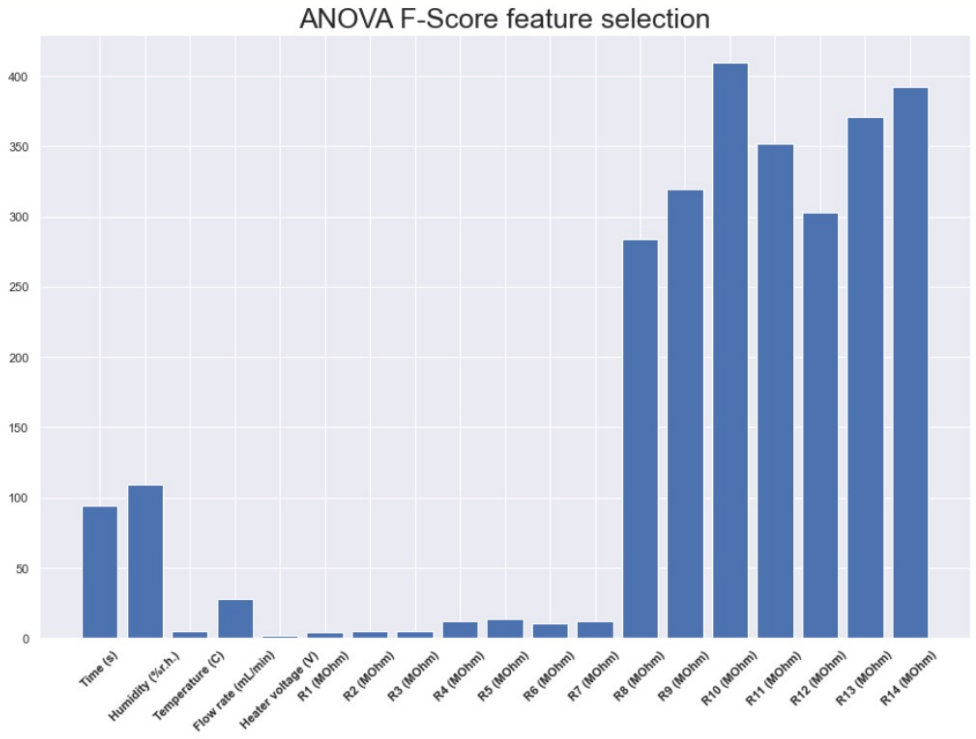
# MODELLING

Before implementing the models, it is necessary to take into consideration some modifications of our data to be able to make a better analysis of our dataset:

1. Normalize the data (in this case the MinMaxScaler method is applied).

2. Feature selection that helps in selecting the most appropriate variables of the dataset (in this case the ANOVA f-value method is applied).

The attributes with higher scores are from the R8 to the R14. Other possible relevant attributes are the time, humidity, and flow rate.



# Models

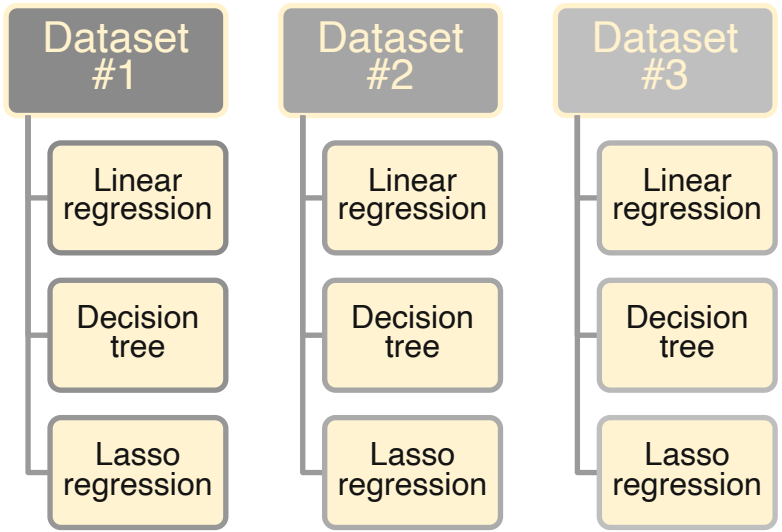
Taking into account the suggestions made by the feature selection algorithm I would propose three variations of the dataset to analyze with every model:

- Dataset #1 = The original dataset with all the features.
- Dataset #2 = 7 features of the original dataset (R8, R9, R10, R11, R12, R13, R14).
- Dataset #3 = 10 features of the original dataset (Time, humidity, flow rate, R8, R9, R10, R11, R12, R13, R14).

Three models are going to be used in this analysis: linear regression, decision tree and lasso regression.

We need to split the dataset into train and test set, but instead of making just one split of the whole dataset, k-folds cross-validation is applied to make more splits.

Three metrics are going to be evaluated: R-squared, Mean Squared Error (MSE), Mean absolute error (MAE). The bigger the value of R-squared and the lower the value of MSE and MAE it is better.



# Results

The best results are obtained with the decision tree algorithm. The linear and lasso regression algorithms performed poorly.

The results obtained with the decision tree using dataset #3 obtain the best scores in the three different metrics. The results are great, with a high R-square of .938 and a low MAE and MSE of 2.60 and 0.341, respectively.

A noticeable dimensionality reduction can be done by analyzing just 10 features of dataset #3, which represents 47.5 % less data than the original.

It is possible to invest less time running models for having fewer features and therefore, you can apply more complex and slower machine learning algorithms such as support vector machine and multilayer perceptron.