So, we have started a new competition. This one is actually very complicated. There are several things to consider:

1. Since this is a time series challenge, how to narrow done the range of the input is very important. Think of it this way, we want to predict April and May, and data from December may as well serves as outliers since they bear different patterns. Do we create different models for different time zones (e.g. a model that predicts May and another model that predicts April)? This should come as no surprise that we need to explore it first. This would be the most important thing throughout this competition. Also, this is very similar to the Marcarri one, with small dataset.

2. CV scheme. Currently I have very little idea as to how to do cross validation on this one, I am assuming what we could do it to use 2016 as an example to compare the behavior of 2017. However, it has been mentioned that the behavior of 2016 is different from 2017.

3. With a different coding scheme, how to efficiently code? I do not want to script. I want to plan it out first.

So, my plans are going to be first read through some EDAs and other simple ones. Reproduce it, get an idea and start planning.

What I’ve learnt today is that I am still not very confident with this kind of datasets. When watching the datafiles, I just have no idea what to do. I need more training on Kaggle. This is a long battle.