Data cleaning

- After merging datasets it seems that there is no need to save name and source_name columns(which are the same), but I tried to extract some features from them so I keep one of them.
- The first column is an index column so I drop it.
- There are 2808 duplicate rows, so I keep 1 of each.
- There aren't nan values in the data. It seems strange, but as I am going to do feature extraction later I will keep in mind this point.
- The flag column has 2 unique values Det(Detected, successful) and UL(Upper Limit, not successful). It depends on the task whether we need UL rows or not so at this point I keep them by assigning Det: 1 and UL: 0.
- It seems that the Catalog column just contains the source of data. I can either drop or do one-hot-encoding with this column, but I will go with first for now as there are 68 unique values and doing one-hot-encoding(codes are commented) will dramatically increase columns in our data, but the source information isn't worth it(this is just an intuitive opinion).
- Every source_names begins with '4FGL_J', so I drop that part of the name and work with the second half. As it seems the other half contains right ascension(hh:mm:ss) and declination(-90:+90) information so lets extract it. 360 degree=24 hours, so I changed hh:mm:ss to degree info.
- End time is always bigger than the start time, but as they are big numbers I decided to replace end time with duration(depending on the task I would prefer to drop start time too, but for now I keep it as it contains information that is not wise to lose). Also I have noticed that duration is almost always 0(start_time = end_time), but I will investigate it in the EDA part.
- It turned out that nufnu_err is the half size of the (nufnu_lower, nufnu_upper) interval (the sum of errors is not exactly 0 because of floating point error, but small enough to be treated as 0). Also nufnu is not in the center of the interval which means we can't drop both nufnu_lower and nufnu_upper so I'll drop just nufnu_upper(this is arguable as we will see in the EDA part).
- (Optional) As I've read nufnu=frequency*flux_density so we can get flux density as we have nufnu and frequency. I haven't included this in the main process, but you can find commented code for this.

Conclusion and open questions: At this point we have a modified version of our initial data. I tried to extract much more information from the initial data and not to lose any. I think the only data that I've lost was a Catalog column which could be replaced with its one-hot encoding or some kind of embedding representing it, but as the column just indicates the source of data and is very imbalanced I preferred just dropping it. Maybe at some point of the project you need it, but in my opinion doing all the complicated modifications with this column at the beginning of the project may not be a good decision.
There are also some arguable columns and samples that were kept:
Do we need the samples which flag is UL? As they are not successful observations and should be removed from our data => the flag column will be deleted too.
Do we need time info? As the columns start_time and duration are in very poor quality almost everywhere 0(duration).
Do we need nufnu_lower if we have nufnu and nufnu_err? The number showed that yes, but as I guess nufnu and nufnu_lower will have high correlation which is bad for some models(linear regression).
I hope the EDA will answer some of these questions. Also I had some issues with outlier detection so I will investigate it in the EDA part.