Our dataset includes data for a class of materials, known as metal organic frameworks (MOFs) that were obtained from QMOF database. The original authors generated the database by running DFT simulations for ~ 10 k – 20 k MOFs (rows). DFT simulations were done at 4 different accuracy *levels* – PBE, HLE17, HSE06\*, HSE06, and each property (column) in the database thus is reported at all the 4 DFT *levels*. Among those calculated properties, our **target property** is the **band gap**. We conducted our exploratory data analysis (EDA) to understand trends in the dataset, explore distributions, discover correlations, uncover biases, and get any insights that will help us in building our machine learning (ML) model. We treated our dataset in two different ways during EDA. Since each property is calculated using 4 DFT simulations, and also not all MOFs are calculated at all *levels of theories* (i. e. ~ 20 k at PBE *level*, ~ 10 k at other *levels*), we first investigated a version of the dataset where each MOF at each *level of theory* was taken as an independent data point. This gives us 52754 rows × 23 columns (excluding multiple MOF identifiers) dataset. The second approach was to filter the dataset such that we end up with a subset of MOFs, where properties (columns) are available at all four *levels of theories* – 10719 rows × 94 columns. It should be mentioned foremost that regardless of the way we treat our dataset the trends and conclusions across both versions of the dataset remain the same. We first examined the distribution of band gaps across 4 DFT *levels* using combined violin + box plots, we see that the shape of the distribution and median is different across the *levels*. Most accurate levels (i. e. HSE06\* and HSE06) tend to predict higher median band gaps and have unimodal shape in their distributions. To understand the differences in distributions at the other levels, the dataset was further separated based on closed-shell/open-shell character of the MOF materials. It turns due to the shortcomings of those other DFT *levels*, the band gaps are computed differently between the two subsets. These observations help us decide the most accurate DFT *levels* (i. e. HSE06) which we can use to train our ML model. Next, we generated pair plots to understand how band gap (target property) is distributed among other numerical properties (columns) in our dataset. We unraveled that except for PBE *level*, DFT was done at other *levels* only for MOFs with smaller system size (150 or less). We pondered that if the system size could introduce a bias into our ML model, but our EDA showed that smaller MOFs seem to span the full range as larger MOFs for structural properties, and most importantly, the band gap distribution among small and large MOFs were similar. Also, from domain knowledge of how these systems are treated in DFT simulations, which is not as isolated systems, but as periodic systems, can also tell us system size differences among the MOFs are not going to be a problem. Further analyses helped us to discover formulaic relationships among properties, most importantly those between target (band gap) and other independent properties. We could come to the conclusion since band gap is computed form these other independent properties, we can also use those properties as targets, in case direct band gap prediction from our ML model fails. Finally, we carried out some correlation analyses between target and other properties, and among all properties themselves. This could also help us in figuring out which columns we can eliminate in training our ML model. Overall, EDA served as a good way to know the landscape of our data, gather insights about our data, and when necessary, transform/reformat our dataset.