We’ve met for an extended period of time once to initially discuss what dataset we were choosing and other basics, but more often keep in touch and communicate with each other through Discord in case we have something to discuss that doesn’t require a dedicated meeting. The dataset we plan to use has been acquired from the Kaggle competition site and we’ve already begun digging through it. The data itself has come pre-divided, with a clearly defined training csv file and a test csv file. While this is convenient in many ways, it does remove much of the inherent randomness that comes from pulling train and test from the same data pool. We have other means of addressing this situation and is something we plan to look into going forward to determine the best methods for producing a consistent and well-fitted result. Part of this determination was that we will approach this project with Python and a handful of its machine learning libraries like sklearn instead of R, in part due to greater confidence and experience with the former over the latter.

Looking deeper into the dataset, it contains about 79 predictors, excluding the Id column and treating SalePrice as the response. This is before breaking down qualitative variables into their necessary dummy variables, of which many have 4 or more categories that need to be broken down. This means we could be looking at over 200 variables for this regression problem, which for simplicities sake we may look to trim down to a more manageable number that still retains a high level of accuracy.