Documentation

Pragyan:

The data that we received was in the raw form with a lot of missing values, structural errors and defunct columns which had to be removed before the other teams could work on the data.

We used Pandas for data manipulation. We used the **“ pandas.read\_csv** “ command for reading the csv files for our task and commands like **.head(), .shape , .describe** to get information about the kind of data we are dealing with; **.dtypes** to determine the datatypes of the columns and check if they are going to be compatible for the various future tasks for the other teams. In case we encountered incompatible datatypes we used **.astype()** to convert them into the desired datatype.

On encountering columns which are defunct **.drop()** was used to delete the columns.

We used the **.isnull()** command for checking the availability of null / missing values. In case we encountered null values , we used the **SimpleImputer** from the **impute**  module of the **sci-kit learn** library for filling in the null values with various mathematical methods like the mean , median or mode of the observations.

Finally we used the **to\_csv** command to convert the cleaned data into a csv file for the other teams to work on.