Data Cleanliness Rubric

In the process of the data cleanliness, we did several examinations in order to check the cleanliness of the data. We examined the dataset by the following steps:

1. Remove useless columns
   - In order to find the valid data for our project, we screened the dataset and found out all the useful columns and removed all the unrelated and useless columns such as the external URLs etc.
2. Correct column names
   - In order to make the column names more readable and easier to understand, we want to correct the column names for columns that have a vague column name
3. Description of the dataset
   - In order to have a better understanding of the dataset, we displayed all the descriptive statistics for the dataset
   - Find out the correct dimension for the dataset
   - Check the unique value for each column in the dataset and the range for all the numerical values in each column
   - Check the certain pattern in each string type variable columns
4. Check duplicate values
   - Find out the duplicate values including duplicate rows and duplicate columns
   - Check the object type like json strings in variable
5. Check missing values
   - Check missing values by rows first and then by column
   - For different type of data, we will leave the missing value blank or fill the missing value using the average value of the column or the value in previous row
6. Check bad values
   - Check values that is not in range, for example, for all the sound features, make sure that all the values are in the correct range
   - Check values that has wrong type and strange pattern, for example the format of the id etc.