**Report**

Data cleaning is the process of fixing or removing incorrect, duplicate or incomplete data within a data set when combining multiple sources as stated in the data integration mentioned in class notes, data cleaning is the procedure that filters out irrelevant data.

Data cleaning is beneficial in the output of a result, Most importantly helps in identifying data that needs to be cleaned and helps in removing duplicate observations to avoid unambiguous result, it helps to fix structural mistakes, it helps to remove outliers to avoid inconsistencies in the data set. it helps in implementation process.

Data cleaning is an extremely important step and a required process because all data is prone to duplications, mislabeling, missing value and so on. It helps to avoid Garbage in and garbage out. This irony means if you throw in a bad dataset for analysis, it could result in an inaccurate result.

The steps of data cleaning involve a. identify data b. fix structural mistakes c. set data cleansing techniques d. implement process and e. filter outliers and fix missing data. It helps to prepare the dataset typically the records are filtered out or data is imputed to get a clean dataset with minimal noise or outliers, the data cleaning must be done to create a clean and uniform dataset to build a robust predictive model. the cleaning of data set helps to remove unnecessary outliers to maintain the quality of data