Name :- Manvi Pandya
Class:- TE10 (L-10)
Roll No. 33235

Assignment-6

AIM: Perform different data cleaning operations using R/Python.

PROBLEM STATEMENT :
Perform the following operations using R/Python on the Air quality and Heart Diseases data sets.
a. Data cleaning.
b. Data integration.
c. Data transformation.
d. Error correcting.
e. Data model building.

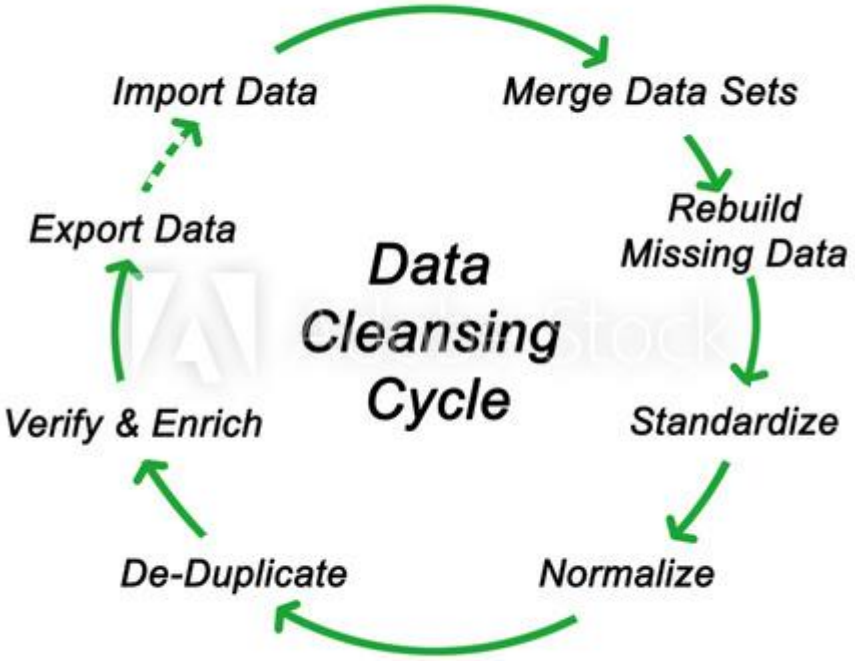
OBJECTIVE:
To learn data processing methods.
To learn building data model.

THEORY:

Data Cleaning:

Data cleansing or **Data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.[1] Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.



R has profound functions to perform data cleaning.
Eg. First import data into R and save the data frame.

Data <- read.csv("abc.csv",na.strings="")

Now analyze the dataset and remove unwanted content.

Trimming white spaces:

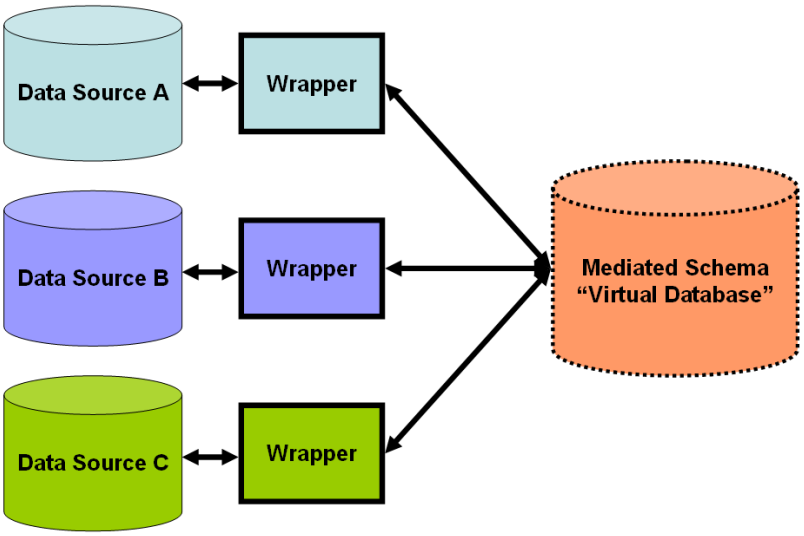
```
data$dist <- str_trim(data$dist)
```

Remove missing values:

```
na.omit(data)
```

Data Integration:

Data integration is the process of combining data from different sources into a single, unified view. Integration begins with the ingestion process, and includes steps such as cleansing, ETL mapping, and transformation.



Data Transformation:

A number of reasons can be attributed to when a predictive model crumples

a. Inadequate data pre-processing
b. Inadequate model validation
c. Unjustified extrapolation

1. Centering and Scaling:
   To center a predictor variable, the average predictor value is subtracted from all values
2. Resolving Skewness:
   Skewness is measure of shape. A common approach to check for skewness is to plot the predictor variable
3. Resolving Outliers:
   The outlier package provides a number of useful functions like outlier and score function

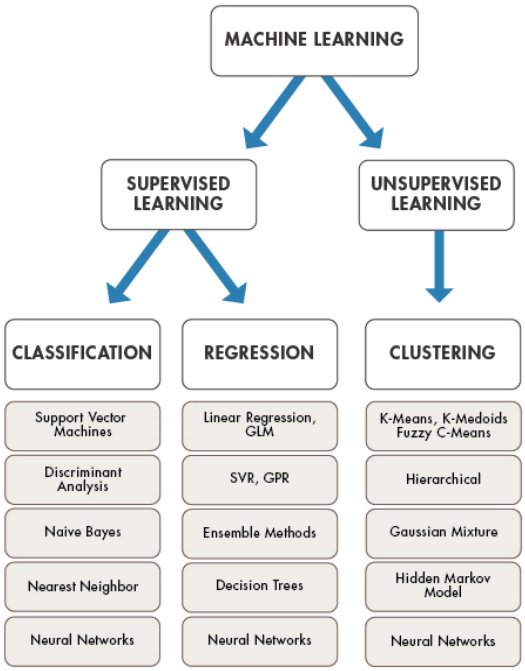
Error correction:

These functions are used to compute sequencing error correction in a library

estimate.errors.mean(lib)
compute.sequence.neighbours(tags, taglength=10, output="character")

Data Model Building:

We can build model in form of linear regression and much more



Eg:

```
linearMod <- lm(dataset$NOx.GT., data=dataset)
print (linearMod)
summary(linearMod)
```

CONCLUSION:

In this assignment we learnt basic data cleaning operations on various datasets.