Roll No: 33231
Name : Prajakta Keer

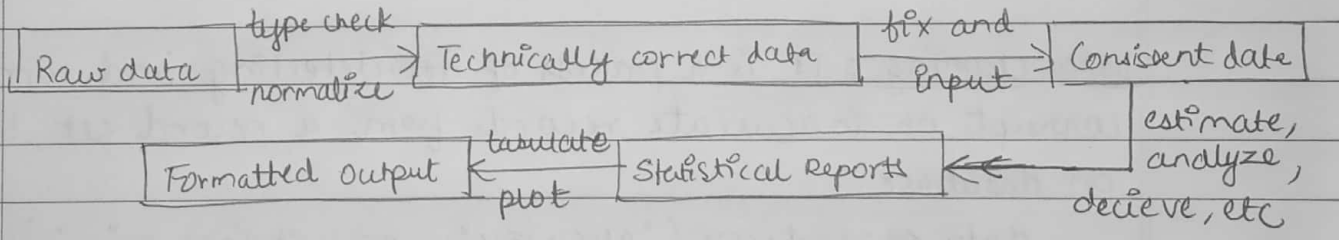
ASSIGNMENT 6

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

Problem statement : Perform the following using R/Python on air quality and heart disease datasets :
- Data cleaning
- Data integration
- Data transformation
- Error correcting
- Data modelling

Objective :
- To learn data processing methods
- To learn building data model

Theory : statistical analysis in 5 steps. A statistical analysis is viewed as result of a number of data processing steps where each step increases the "value" of data (raw data)



Variable types and indexing techniques :

vector have variables of one-type

```
c(1, 2, "three")
```

shorter arguments are recycled

```
(1:3) * 2 , (1:4) * (1:2)
```

warning ! (Why?)

```
(1:4) * (1:3)
```

Each element of a vector can be given a name. This can be done by passing named arguments to the c() function or later with named functions.

```
x <- c("red", "green", "blue")
capColor = c(huey = "red", duey = "blue", louie = "green")
capColor = ["louie"]
```

33231

classmate
Date
Page

names (capcolor) [capcolor = "blue"]

x ← s (4, 7, 6, 5, 2, 8)

I ← x < 6 J ← x > 7 x [I|J]

x [c (TRUE, FALSE)]

x [c(-1, -2)]

Data Transformation : A number of reasons can be attributed to when a predictive model crumples such as • Inadequate data pre-processing • Inadequate model validation • Unjustified extrapolation • overfitting

→ Common data pre-processing steps :

1) Creating and scaling 2) Resolving skewness

3) Resolving antilizers 4) mirrioring value treatment

Data cleaning : It is a process of ~~data~~ detecting and correcting corrupt or inaccurate records from a record set, table or database.

data ← read.csv ("abc.csv", na strings = " ")

Now analyze dataset and remove unwanted content.

Trimming white spaces - data $dist ← str_trim (data $dist)

Remove missing values - na omit (data)

Data Integration : It is a 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, etc.

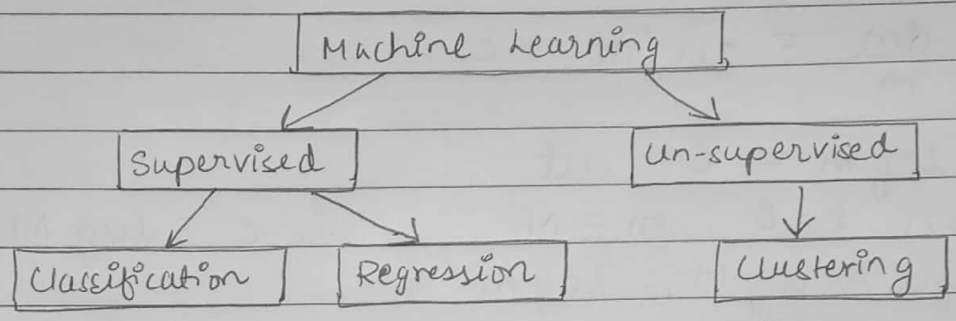
Error correction : These functions are used to compute sequencing error correction in a library estimate errors mean (lib)

compute sequence neighbours (tags, taglength = 10, output = "character")

33231

classmate
Date
Page

Data Model Building : We can build model in form of linear regression and much more.



Eg: linearmod ← lm (dataset $NOX.C++ , data = dataset)
print (linearmod)
summary (linearmod)

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