DATE __/__/____
PAGE NO _____

33304

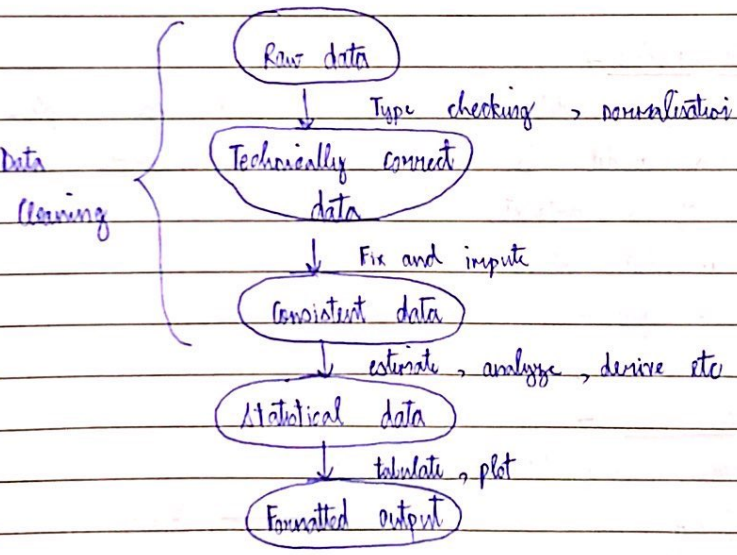
myCompanion

Assignment - 06

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

* **Problem statement →** :- Perform the following operations using R/Python on the air quality and heart disease datasets.
  1. Data cleaning
  2. Data ~~transformation~~ integration
  3. Data ~~integration~~ transformation
  4. Error correcting
  5. Data model building

* **Theory :-**



. **Data cleaning →**
Data cleaning is the process of preparing data for analysis by removing or ~~identifying~~ modifying data that is incorrect, incomplete, irrelevant, duplicate or improperly formatted.
Eg – Removing NAs.

```
airquality = read.csv("Airquality.csv")
airquality$Ozone = ifelse(is.na(airquality$Ozone),
    median(airquality$Ozone, na.rm = True), airquality]
```

33304

- **Data Transformation** →
It is the process of converting data from one format or structure.
Eg - airquality $ danger = airquality $ Temp > 100

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

// dataset 1 and dataset 2 is already available

dataset = Merge (dataset 1, dataset 2, by = "FN")

- **Build data models** →
Using simple models to better understand complex dataset. The goal of a model is to provide a simple low dimensional summary of a dataset.
Models : 1. list column
2. gapminder

- **Error correction** →
Eg - mean (airquality $ Temp)
[1] NA ← Error
mean (airquality $ Temp, na.rm = TRUE)
↑
Remove the error

33304

* Conclusion :- Thus, in this assignment we have performed data cleaning operation on air quality dataset using R