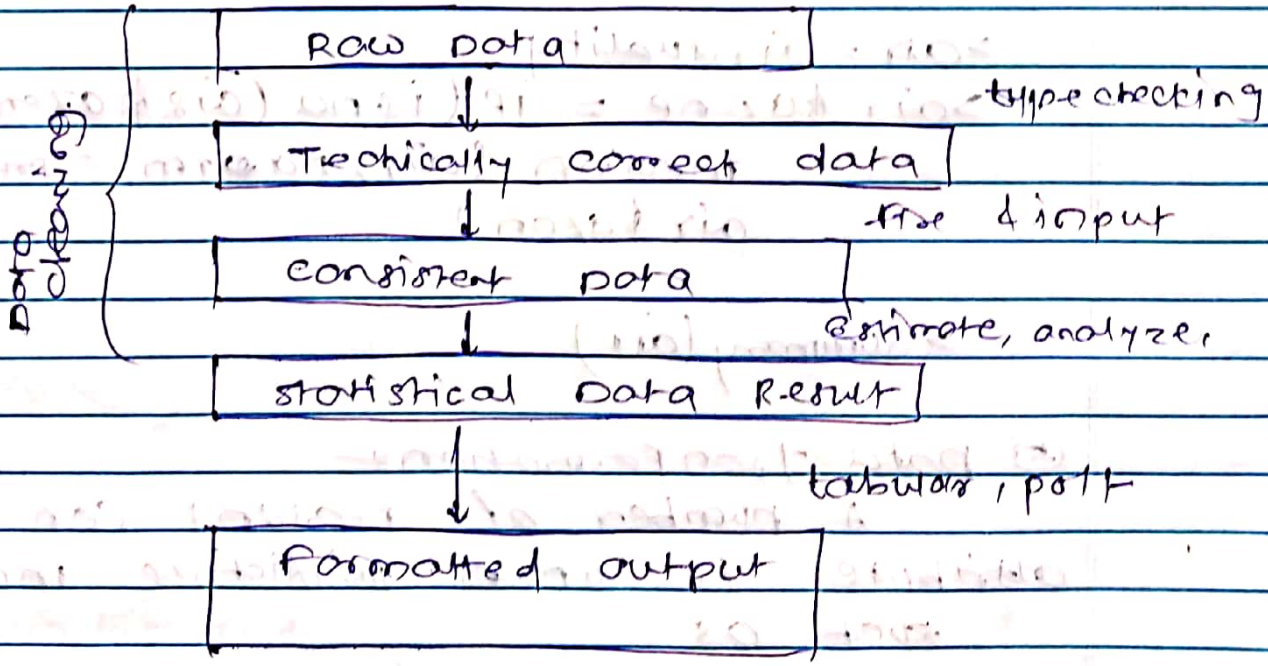
Page No.
Date

Assignment No. 6

→ Problem statement :-
Perform following operations using Python on Air quality & Heart disease dataset.

ⓐ Data cleaning
ⓑ Data integration
ⓒ Data transformation
ⓓ Error correcting
ⓔ Data model building.

→ Theory :-
Diagram of statistical analysis value chain



① The dataset
* Airquality

② Errors & Corrections
> mean (airquality $Ozone)
[1] NA
> mean (airquality $Ozon, na.rm = True)
[1] 42.12931

③ check summary

> summary (airquality)

④ Data cleaning (removing NAs).

> air = airquality
> air $Ozone = ifelse (is.na (air$Ozone),
median (air, na.rm, = True),
air $Ozon)

> summary (air)

⑤ Data transformation:
A number of reasons can be attribute to when predictive model such as
- Inadquate Data
- Inadquate model validation
- over-fitting

Page No.
Date

⑥ missing Values treatment.
   (a) Input Value with median or mode
   (b) Input value with KNN

Eg.
```
>head
> air $solar.r.anger = air $solar.r - R > 100
> head(air)
```

* Conclusion :-
Hence, I have studied & implemented the data preprocessing techniques.