Lab assignment 2

Group B

→ Title : Perform the following operations using python on the air quality and heart disease attacks.

→ Objective :

a) To understand and apply the analytical concept of big data using python.

b) To study detailed concept in R.

→ Software requirements :

a) Ubuntu 14.04 / 14.10

b) GNC C compiler

c) Hadoop

d) Java

e) Python platform

→ Problem statement : Perform the following operations using python on the air quality and heart disease attacks.

a) Data cleaning

b) Data integration

c) Data transformation

d) Error correcting

e) Data model building

→ Theory :

· Data cleaning or data preparation is an essential part of statistical analysis. Infact, in practice it is often more time consuming than the statistical analysis itself.

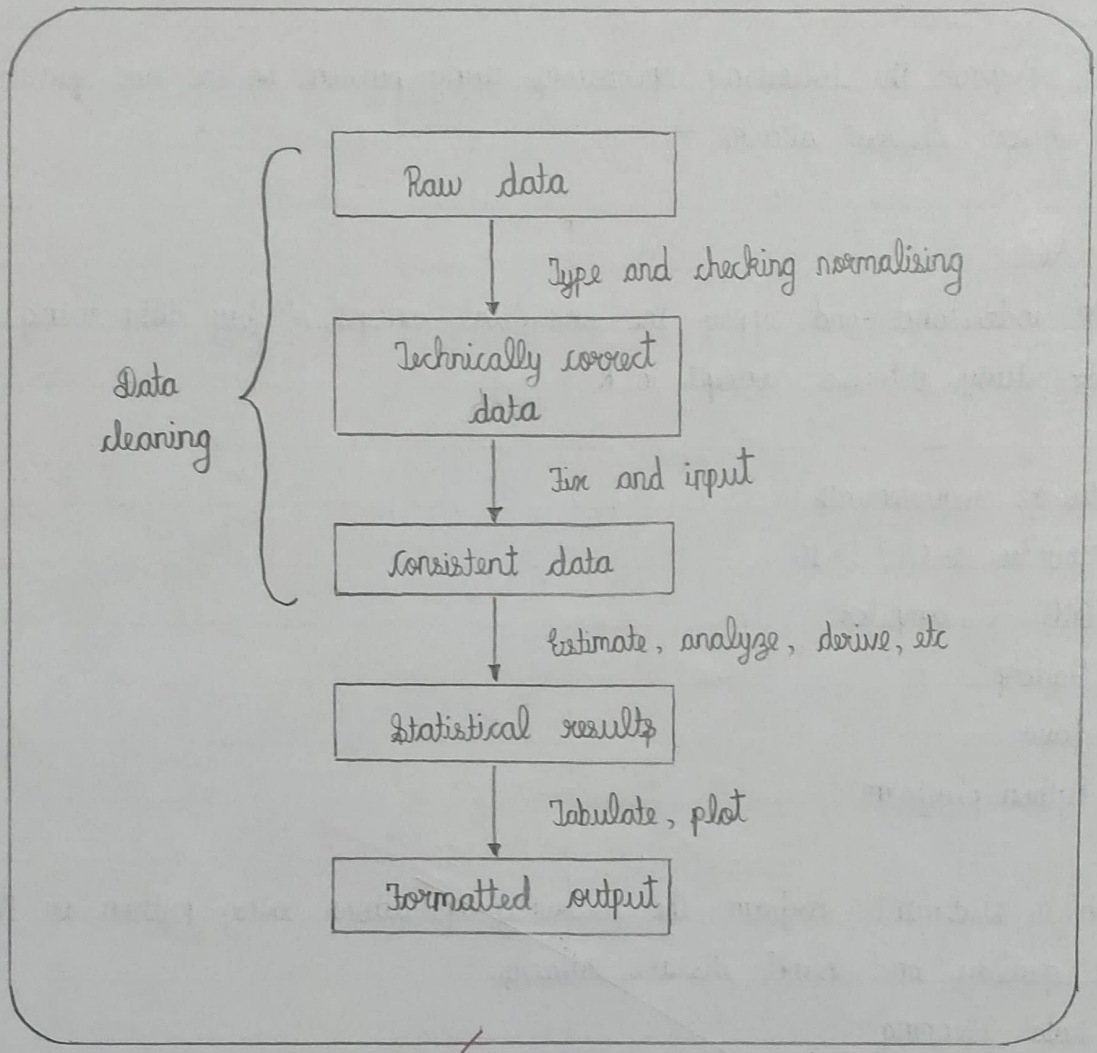

Fig. Statistical analysis value chart

- The statistical value chain handles data in a cradle to grave perspective from the extraction of raw data to its use for decision support.
- The better the data is handled at each step of statistical value chain, the better the resulting decision support and therefore the better the final decisions.

Variable types and indexing techniques -

- By indexing, we mean all methods and tricks that allow you to select and manipulate data using logical integer or normal indices.

Data transformations -

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

a) inadequate data pre-processing
b) inadequate model validation
c) unjustified extrapolation
d) over-fitting

a) Centering and skewness -

- Variable scaling is perhaps the most intuitive approach used in predictive modeling.
- To center a predictor variable, the average predictor value is subtracted from all values.

b) Resolving skewness -

- Skewness is a measure of shape.
- A common approach to check for skewness is to plot the predictor variable.
- As a rule, negative skewness indicates that the mean of the data values is

less than median and data distribution is left skewed.

c) Resolving outliers -

- The function outlier() gets the extreme most observation from mean.
- If you set the argument opposite = TRUE, it fetches from after side.

1. Imputation : Imputation with mean / median / mode.
2. Capping : For missing values that lie outside the 1.5 * IQR limits, we would cap it by replacing those observations outside the lower limit with value of 5th and those that lie above the upper limit, with the value for 95% percentile.

d) Missing value treatment -

1. Impute missing values with median or mode.
2. Impute missing values based on K-nearest neighbour.

There are many other types of transformations like treating collinearity, dummy variables, encoding, covariance, treatment, etc.

→ Conclusion : Thus, we have learnt how to perform the different data cleaning and data modelling operations using python.