DS & BDA Lab

Group B - Experiment 5

TITLE : Perform the following operations using R/Python on the Air quality & Heart Diseases data sets.

OBJECTIVE :

1. To understand & apply the Analytical concept of Big data using R/Python.
2. To study detailed concept R/Python.

SOFTWARE REQUIREMENTS:

1. Ubuntu 14.04 / 14.10
2. GNU C Compiler
3. Hadoop
4. Java
5. Python platform.

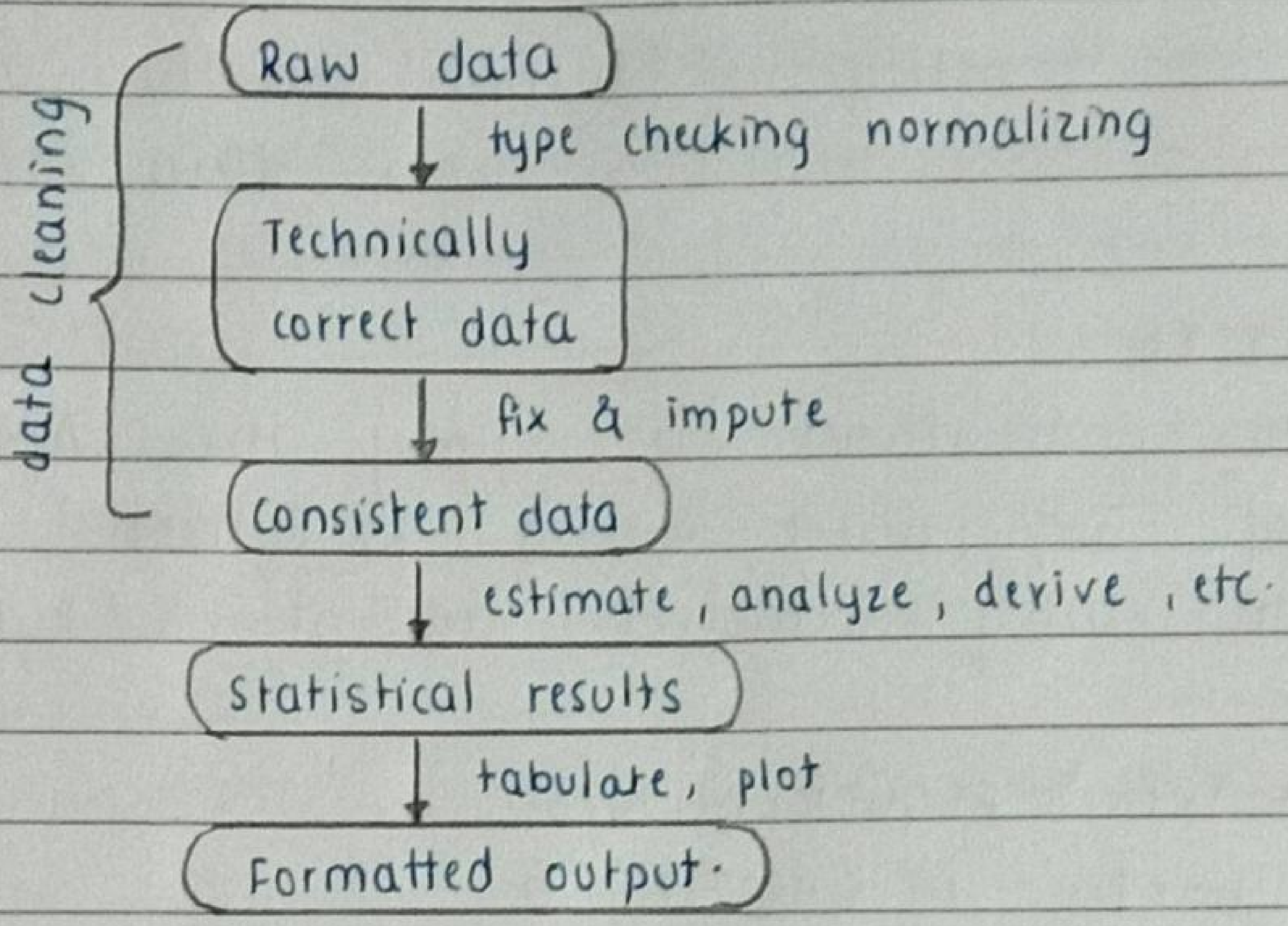
PROBLEM STATEMENT : Perform the following operations using R/Python on the Air quality & Heart Diseases data sets :-

1) Data cleaning
2) Data integration
3) Data transformation
4) Error correcting
5) Data model building.

THEORY :

Data cleaning, or data preparation is an essential part of statistical analysis.

* Statistical analysis in five steps :-



The statistical Value Chain (SVC) handles data in a cradle-to-grave prespective, from the extraction of raw data to its use for decision support. The better the data is handled at each step of the Statistical Value Chain, the better the resulting decision support - & therefore the better the final decisions.

* Variable types & indexing techniques :-
By indexing, we mean all methods & tricks that allow you to select & manipulate data using logical, integer or named indicies.

* Special values like most programming languages, R has special values like NA, NULL, ± Inf & NaN.

* Data Transformations :-

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

- Inadequate data pre-processing.
- Inadequate model validation.
- Unjustified extrapolation.
- Over- fitting

- Predictor / Independent / Attributes / Descriptors : are different terms that are used as input for the prediction equation
- Response / Dependent / Target / Class / Outcome : are terms that refer to the outcome event that is to predicted.

1. Centering & scaling:

Variable centering is perhaps the most intuitive approach used in predictive modeling. To centre a predictor variable, the average predictor value is subtracted from all the values, as a result of centering, the predictor has zero mean.

To scale the data, each predictor value is divided by its standard deviation (sd).

2. Resolving Skewness:

Skewness is measure of shape. A common approach to check for skewness is to plot the predictor value. Negative skewness indicates that mean of data value is less than median, & data distribution is left-skewed. Positive skewness would indicate that mean of data values is larger than median, & data distribution is right-skewed.

3. Resolving Outliers :
The function outliers () gets the extreme most observation from the mean.
Outlier Treatment :-
A. Imputation : Imputation with mean / median / mode.
B. Capping : For missing values that lie outside, all the values will be considered as outliers & the numbers of outliers in the dataset gives that capping number.

4. Missing value treatment :
- Impute Missing values with median or mode.
- Impute missing values based on K-nearest neighbours.

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

CONCLUSION : Thus we have learnt how to perform the different data cleaning & data modeling operations using python.

DS & BDA Lab
Group A - Experiment 2
TITLE: Design a distributed application using MapReduce.

OBJECTIVE:

1. To explore different Big data processing techniques with use cases.
2. To study detailed concept of Map-Reduced.

SOFTWARE REQUIREMENTS:

1. Ubuntu 14.04 / 14.10
2. GNU C compiler
3. Hadoop
4. Java

PROBLEM STATEMENT: Design a distributed application using MapReduce (Using Java) which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log file from the Internet & process it using a pseudo distribution mode on Hadoop platform.

THEORY:

* Introduction:
MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner. MapReduce is a processing technique & a program model for distributed

computing based on java.
The MapReduce algorithm contains two important tasks: namely Map & Reduce. Map takes a set of data & converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).
Secondly, reduce task, which takes the output from a map as an input & combines into those data tuples into a smaller set of tuples.