Assignment 4

Title: Perform the operations using python on the facebook metrics data sets.

AIM: Perform following operations using python on facebook metrics data sets.

a) Create data subsets
b) merge data
c) sort data
d) Transposing data
e) shape & reshape data.

Objective - 1) To understand & apply the analytical concept of big data using python.
2) To study detailed concept python.

Software requirements: 1) Ubuntu 14.04/14.10
2) GNU C compiler
3) Hadoop
4) Java
5) Python platform

Theory -

Python is an easy to learn, powerful programming language. It has efficient high level data structure & a simple but effective approach to object oriented programming Python elegant syntax & dynamic typing, together with its interpreted nature, make it an ideal language for scripting & rapid application development in many areas on most platform.

The python interpreter & the extensive standard library are truly available in source or binary form for all major platforms from python website

- features of python.

Python is a dynamic, high level free open source & interpreted programming language.

1) easy to use
2) object oriented language
3) ~~ob~~ Free & open source
4) GUI Programming support
5) High level language
6) Extensible feature
7) Python is portable.

* Data Reshaping -

It is about changing the way data is organised into rows & columns most of time data processing in python is done by taking input data from rows & columns of data frame but there are situations when we need data frame in a format that is different from format in which we recived it. Python has many functions to split, merge & change rows to columns & vice versa in data frames.

1) Joining columns & Rows in data frame.

- We can join multiple vectors using cbind() function we can merge using rbind() function.

2) Subsets

Subsetting vectors, matrices & data frames Return subsets of vectors, matrices or data frames which meet conditions.

Arguments :-

x : object to be subsetted

subset : logical expression indicating elements or rows to keep

PCCOE

missing values are taken as false.
select : expression, indicating columns
to select from a data frame.
drop: passed on to indexing operator.

3) melting & Casting
The functions used to do this are
called melt() & cast()
Steps: melt the data, cast molten
data, transpose using t()
function.

4) Sorting Data:
to sort a dataframe use order()
function by default sorting is
ASCENDING prepand sorting variable
by a minus sign to indicate
DESCENDING order.

Conclusion - Thus we have learnt
how to perform different reshape
operations using python.

Assignment 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 following operations using R| Python on air quality & heart diseases data sets.

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

```
2 rows x 13 columns]
```

Theory:

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

* Statistical analysis in 5 steps:

Raw data

type checking normalizing

data cleaning

technically correct data

fix & impute

Consistent data

estimate, analyze, derive, etc

statistical results

tabulate, plot

formatted output

The statistical values chain (SVC) handles data in a cradle-to-grave prespective from extraction of raw data to its use for decision support the better data is handled at each step of statistical value chain better resulting decision support. The better data is handled at each step of statistical value chain better resulting decision support - f -

```
[435742 rows x 13 columns]
```

therefore better final decisions.

variables types & indexing techniques
By indexing , we mean all methods & tricks that allow you to select & manipulate data using logical , integer or named indices.

* special values like most programming languages , R has special values like NA , NULL , & NAN.

* Data Transformations :-
A no. of reasons can be attributed to when a predictive model crumples such as
- inadequate data preprocessing
- inadequate model validation
- unjustified extrapolation.
- Response / Dependant / target / class / Outcome are terms that refer to the outcome event that is to predicted.

i) 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 values , as a result of centering the predictor has zero mean. To scale 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 predictor values is subtracted from all values as a result of centering the predictor has zero mean. To scale data each predictor value is divided by its standard deviation (sd) Positive skewness would indicate that mean of data is larger than median & data distributed is right skewed.

3) Resolving outliers.

The function outliers() gets the extreme most observation from mean.

Outlier treatment.

A) Imputation : Imputation with mean/ median /mode.

B) Capping : for missing values that lie outside, all values will be considered as outliers.

```
[435742 rows x 13 columns]
```

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 transformation 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.