**BDA Assignment 2**

Team Members

|  |  |
| --- | --- |
| Name | Registration No. |
|  |  |
|  |  |

**Selected Dataset**: Dow Jones Index Data Set

Abstract: This dataset contains weekly data for the Dow Jones Industrial Index. It has been

used in computational investing research.

Number of Instances: 750

Number of Attributes: 16

Attribute Information: Quarter, Stock, Date, Open, High, Low, Close, Volume,

Percent\_Change\_Price, Percent\_Change\_Volume\_Over\_Last\_Wk, Previous\_Weeks\_Volume,

Next\_Weeks\_Open, Next\_Weeks\_Close, Percent\_Change\_Next\_Weeks\_Price,

Days\_To\_Next\_Dividend, Percent\_Return\_Next\_Dividend.

Explain what fields are present in csv file used to import data to Hive:

Quarter: the quarter when the stock was measured.

Stock: the name used to identify the stock.

Date: the date when the stock was measured.

Close: the measured when the stock exchange closed.

**Explain what problem is solved using R and Hive:**

Data set is loaded using read.table function then a subset is taken of the table which contain the symbol ‘IBM’ after taking the subset we count the number of stocks in retrieved table by using nrow function. The mean, standard deviation of the closing price is taken using mean, sd function respectively.

To see the relative difference between change in price of stock and attribute percentage return next dividend the script first calculates difference between opening and closing price of the stock on every day then we take the minimum value of these difference taken and then add them to the difference to get a relative positive value. We then plot a graph containing these relative differences and attribute percentage return next dividend to check for linear regression. Resultant graph is given in the visualization given the graph now the script tries to fit linear model using lm function then we summarize the result using summary function.

In hive the command tries to calculate average closing price of stock named with symbol IBM. Then partition is done using ticket symbol of stocks after partitioning each partition contains all the rows of that stock. In Bucketing three buckets are created according to low, medium and high closing value of stock.

**Hive commands:**

**Load data into Hive warehouse**

CREATE DATABASE IF NOT EXISTS STOCKS COMMENT 'Dow Jones Details' WITH DBPROPERTIES ('creator' = 'MIT');

SHOW DATABASES;

DESCRIBE DATABASE EXTENDED STOCKS;

CREATE EXTERNAL TABLE IF NOT EXISTS EXT\_STOCK4(quarter INT, symbol STRING, dateandtime STRING, close FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','LOCATION '/user/hive/warehouse/stock';

LOAD DATA LOCAL INPATH '/home/cloudera/Downloads/djindex1.csv' OVERWRITE INTO TABLE EXT\_STOCK4;

SELECT \* from EXT\_STOCK4;

**Solved particular problem**

SELECT avg(close) from EXT\_STOCK HAVING symbol='IBM';

**Partitioning**

CREATE TABLE IF NOT EXISTS STATIC\_PART\_STOCK(quarter INT,symbol STRING, dateandtime STRING, close FLOAT) PARTITIONED BY (symbol STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

INSERT OVERWRITE TABLE STATIC\_PART\_STOCK PARTITION (symbol='IBM') SELECT symbol, reported from EXT\_STOCK4 where symbol='IBM';

**Bucketing**

CREATE TABLE IF NOT EXISTS DJI\_BUCKET(quarter INT,symbol STRING, dateandtime STRING, close FLOAT)CLUSTERED BY (close) into 3 buckets;

FROM EXT\_STOCK INSERT OVERWRITE TABLE DJI\_BUCKET SELECT quarter, symbol, dateandtime, close;

SELECT DISTINCT close FROM DJI\_BUCKET SAMPLE(BUCKET 1 OUT OF 3 ON close);

**R script:**

dji <- read.table("E:/dow\_jones\_index/dow\_jones\_index.data", header = TRUE)

ibm\_stock = subset(df, stock=='IBM')

nrow(ibm\_Stock)

mean(ibm\_stock$close)

sd(ibm\_stock$close)

ibm\_stock$diff <- ibm\_stock$open - ibm\_stock$close

min <- ibm\_stock$diff[which.min(ibm\_stock$diff)]

ibm\_stock$rdiff <- ibm\_stock$diff - min

plot(ibm\_stock$percent\_return\_next\_dividend, ibm\_stock$rdiff, main = "Diffrence in price vs Percentage return next divident", ylab = "Difference in price", xlab = "Percentage return next divident")

gives = lm(ibm\_stock$percent\_return\_next\_dividend~ibm\_stock3$rdiff, data = ibm\_\_stock)

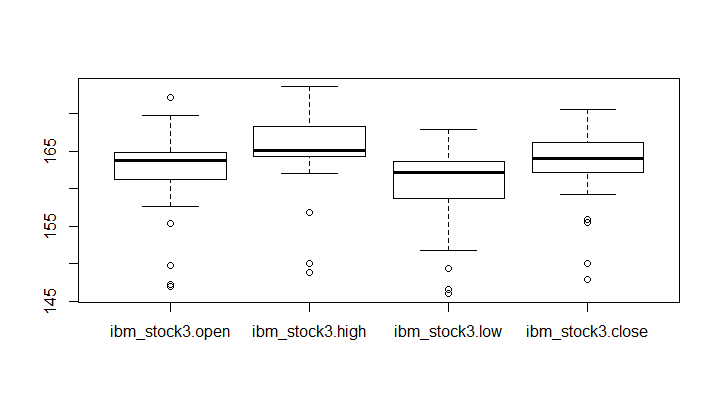
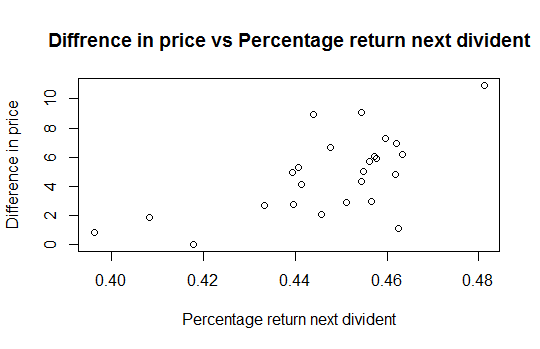
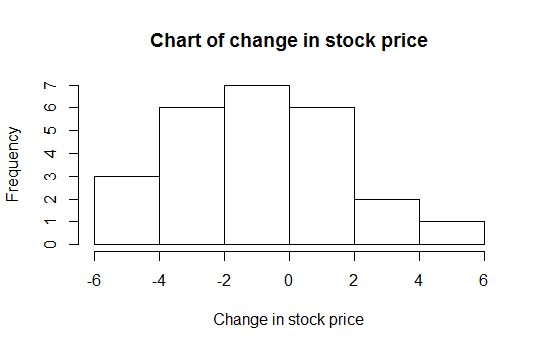
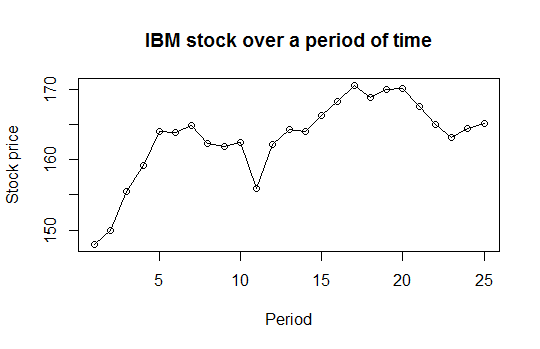
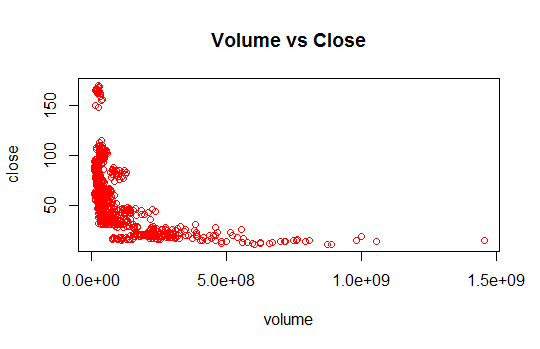
summary(gives)

plot(df$volume,df$close,col = "red", xlab = "volume", ylab = "close", main = "Volume vs Close")

hist(ibm\_stock3$diff, xlab = "Change in stock price", main = "Chart of change in stock price")

data <- data.frame(ibm\_stock3$open,ibm\_stock3$high,ibm\_stock3$low,ibm\_stock3$close)

boxplot(data)

**Screenshot of visualization in R:** xxxxx