APPLIED DATA SCIENCE PROJECT
VINAYSHEEL KAPGATE

INTRODUCTION:

Dow jones stock market is one of the largest stock market in the world, with over millions of stock traded daily in large volumes, and with over more than 100 fortune companies listed in its capacity, all the other markets in other countries have some correlation with it. In my project I will try to predict which of the stocks will be successful in the long run based on the data collected on previous weeks data, a prediction of a stock being successful by more than 50% would be a much profitable deal. Stock market investing is also considered a huge gamble because it is difficult to know when the stock would fall or rise. In my applied data science project I will try to predict whether a given stock will be profitable or not in the next week given the data for present week. A stock market to be defined in layman terms is the aggregation of buyers and sellers, which represent the ownership claim of the business in terms of equity market. The dataset that I collected was from the UCI repository. The DOW jones is an aggregate of publicly 30 traded stocks of value of each stock , which measures the how wealthy a owner of the companies are, it also measures how the strength of the us economy is comparison with other countries.

THE DATA SECTION:

The data section consists of the stock name, date ,opening price , closing price, high, close,open,volume, percent_change_price , percent_change_volume_over_last_wk.The dataset description can be found in the fields given below , the data has 750 observations and 16 variables.

```
> str(e)
'data.frame':         750 obs. of  16 variables:
 $ quarter                     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ stock                       : Factor w/ 30 levels "AA","AXP","BA",..: 1 1 1 1 1 1 1 1 1 1
 $ date                        : Factor w/ 25 levels "1/14/2011","1/21/2011",..: 4 1 2 3 8 5 6 7 12 9
 $ open                        : num  65 81 75 66 74 112 115 88 83 79
 $ high                        : num  77 76 72 75 106 107 111 103 82 78
 $ low                         : num  61 59 58 62 76 86 106 65 75 56
 $ close                       : num  75 65 63 73 108 111 110 81 80 71
 $ volume                      : num  2.40e+08 2.43e+08 1.38e+08 1.51e
 $ percent_change_price        : num  3.79 -4.43 -2.47 1.64 5.93
 $ percent_change_volume_over_last_wk: num  NA 1.38 -43.02 9.36 1.99
 $ previous_weeks_volume       : num  NA 2.40e+08 2.43e+08 1.38e+08
 $ next_weeks_open             : num  80 74 65 73 112 115 87 82 78 67
 $ next_weeks_close            : num  68 65 76 111 114 113 84 83 74 75
 $ percent_change_next_weeks_price   : Factor w/ 2 levels "0","1": 1 1 2 2 2
 $ days_to_next_dividend       : num  26 19 12 5 97 90 83 76 69 62
 $ percent_return_next_dividend      : num  0.183 0.188 0.19 0.186 0.175

>summary(e)
```

```
> summary(e)
    quarter          stock            date          open            high            low             close
 Min.   :1.00   AA      : 25   1/14/2011: 30   Min.   :  1.0   Min.   :  1.0   Min.   :  1.0   Min.   :  1.0
 1st Qu.:1.00   AXP     : 25   1/21/2011: 30   1st Qu.:183.2   1st Qu.:177.0   1st Qu.:178.2   1st Qu.:177.2
 Median :2.00   BA      : 25   1/28/2011: 30   Median :359.5   Median :352.5   Median :354.5   Median :351.5
 Mean   :1.52   BAC     : 25   1/7/2011 : 30   Mean   :361.4   Mean   :354.7   Mean   :355.9   Mean   :354.1
 3rd Qu.:2.00   CAT     : 25   2/11/2011: 30   3rd Qu.:539.8   3rd Qu.:532.8   3rd Qu.:532.8   3rd Qu.:530.8
 Max.   :2.00   CSCO    : 25   2/18/2011: 30   Max.   :722.0   Max.   :713.0   Max.   :711.0   Max.   :711.0
                (Other):600   (Other)  :570
     volume          percent_change_price percent_change_volume_over_last_wk previous_weeks_volume next_weeks_open
 Min.   :9.719e+06   Min.   :-15.42290    Min.   :-61.4332                   Min.   :9.719e+06     Min.   :  1.0
 1st Qu.:3.087e+07   1st Qu.: -1.28805    1st Qu.:-19.8043                   1st Qu.:3.068e+07     1st Qu.:182.2
 Median :5.306e+07   Median :  0.00000    Median :  0.5126                   Median :5.295e+07     Median :358.5
 Mean   :1.175e+08   Mean   :  0.05026    Mean   :  5.5936                   Mean   :1.174e+08     Mean   :360.2
 3rd Qu.:1.327e+08   3rd Qu.:  1.65089    3rd Qu.: 21.8006                   3rd Qu.:1.333e+08     3rd Qu.:538.8
 Max.   :1.453e+09   Max.   :  9.88223    Max.   :327.4089                   Max.   :1.453e+09     Max.   :720.0
                                          NA's   :30                         NA's   :30
 next_weeks_close percent_change_next_weeks_price days_to_next_dividend percent_return_next_dividend
 Min.   :  1.0    0:363                           Min.   :  0.00         Min.   :0.06557
 1st Qu.:180.2    1:387                           1st Qu.: 24.00         1st Qu.:0.53455
 Median :354.5                                    Median : 47.00         Median :0.68107
 Mean   :357.0                                    Mean   : 52.53         Mean   :0.69183
 3rd Qu.:533.8                                    3rd Qu.: 69.00         3rd Qu.:0.85429
 Max.   :715.0                                    Max.   :336.00         Max.   :1.56421
```

The percent change in the price which is the target variable and the factor variable is converted into factor variable is converted into 1 for positive change and 0 for the negative change,the opening price and closing price being in dollars which is obvious,the code for the percent change in the next week price is given below,

e$ percent_change_next_weeks_price [e$percent_change_next_weeks_price<0]=0

e$ percent_change_next_weeks_price [e$percent_change_next_weeks_price>0]=1

The table for stock name for various companies is given in the table, the companies include tech giants like apple, cisco,ibm,etc.

> table(e$stock)

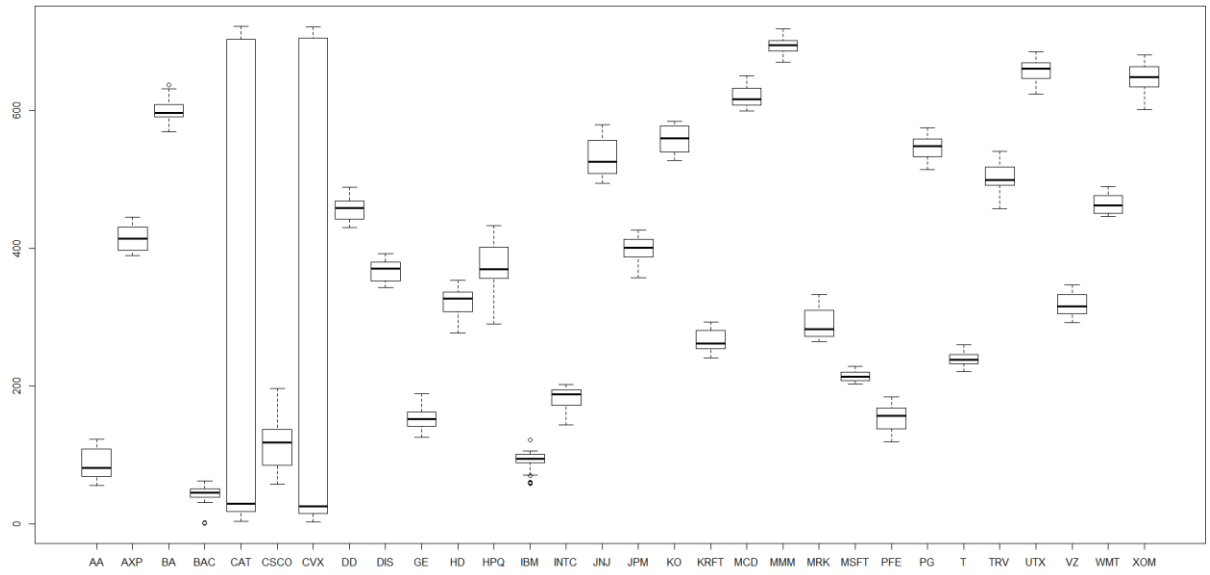
AA AXP BA BAC CAT CSCO CVX DD DIS GE HD HPQ IBM INTC JNJ JPM KO KRFT MCD MMM MRK MSFT PFE PG

25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25 25

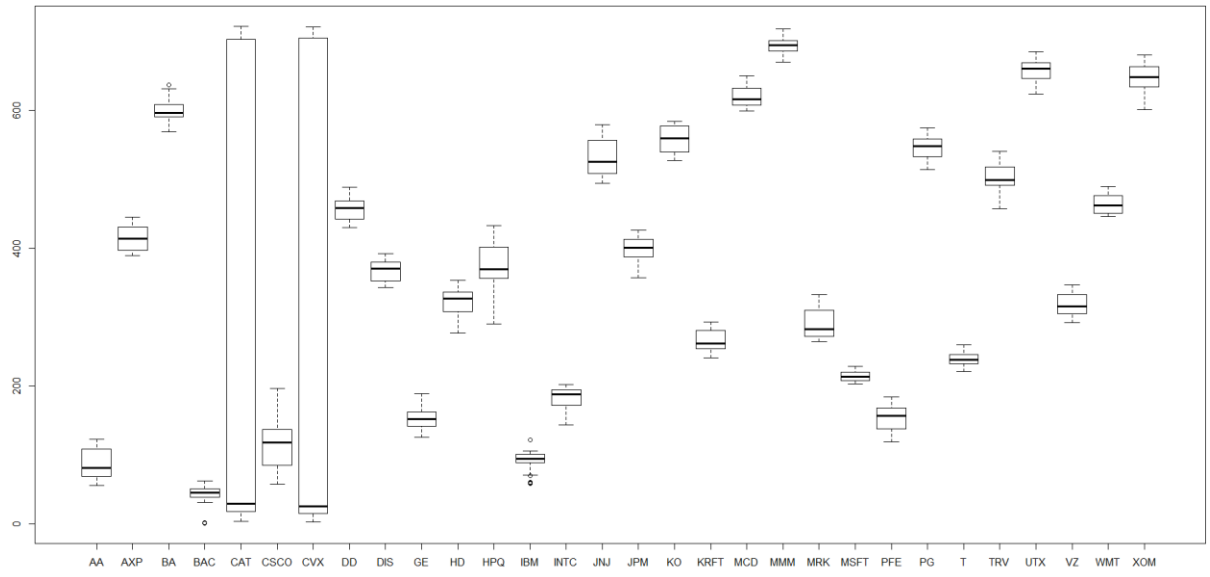
T TRV UTX VZ WMT XOM

25 25 25 25 25 25

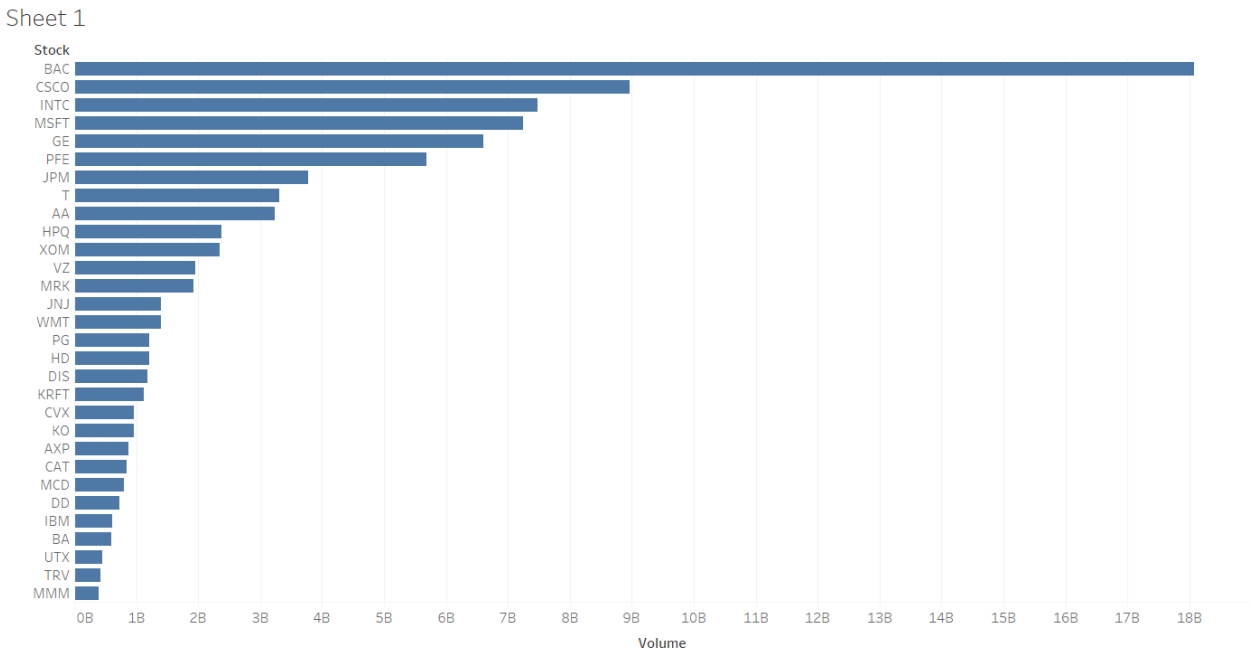
The boxplot for the stock name and the opening price is given below , we can see that there is a huge variation in the some stocks like cat,cvx while the variation is low in companies in companies like IBM,Microsoft.The x axis shows the stockname , while the y axis is the opening price is dollars.



The boxplot for the stock name and the closing price is shown in the plot given , we can see there is no change as the variation is same companies ,

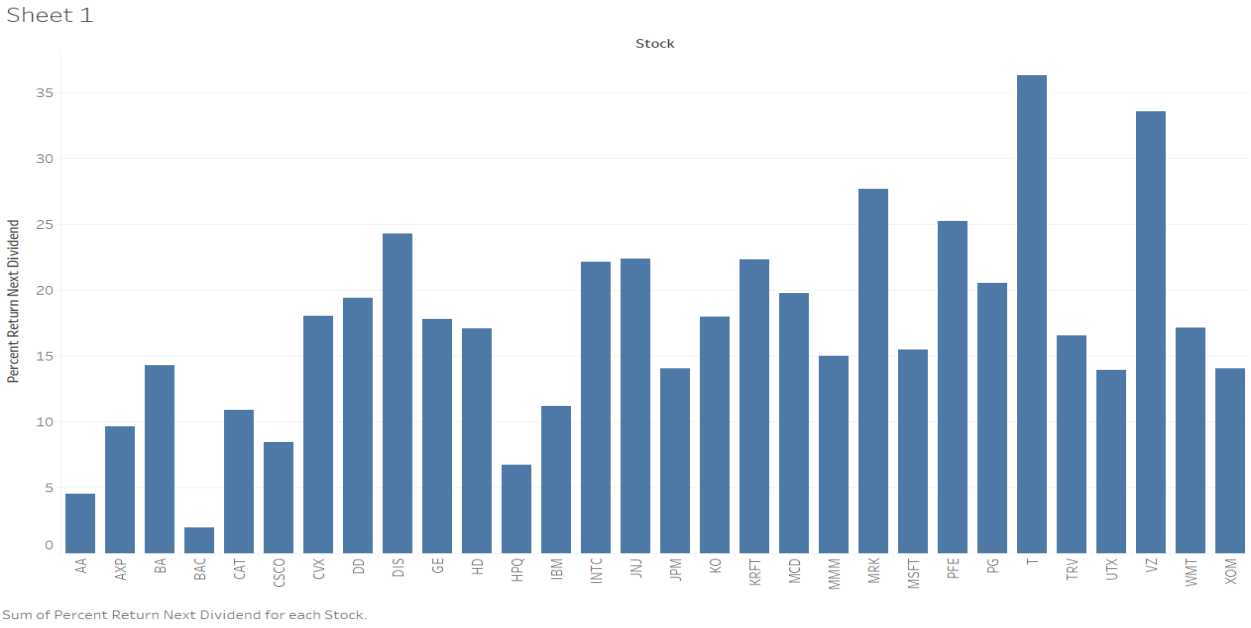


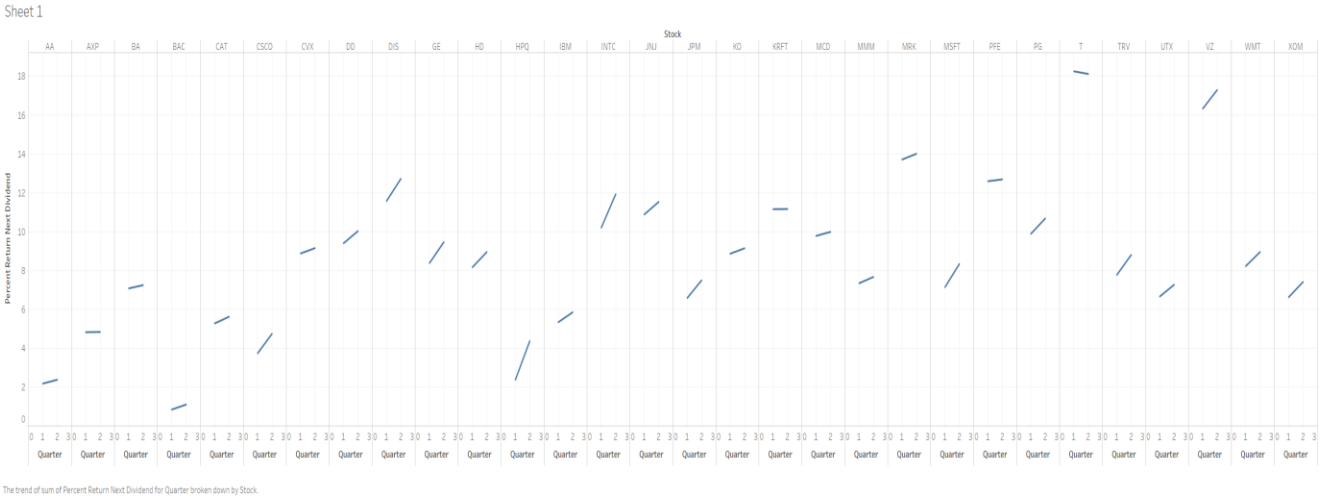
The boxplot in the above shows the stock name and the low price variation, from the plot the variation is closed as the previous one.



The plot above the stock traded of various companies and the volume of the traded stocks ,from the plot we can see that the bank of America , cisco , intel has the highest number of stocks traded,while on the other hand stocks such as trv and mmm has the lowest number of traded stocks.

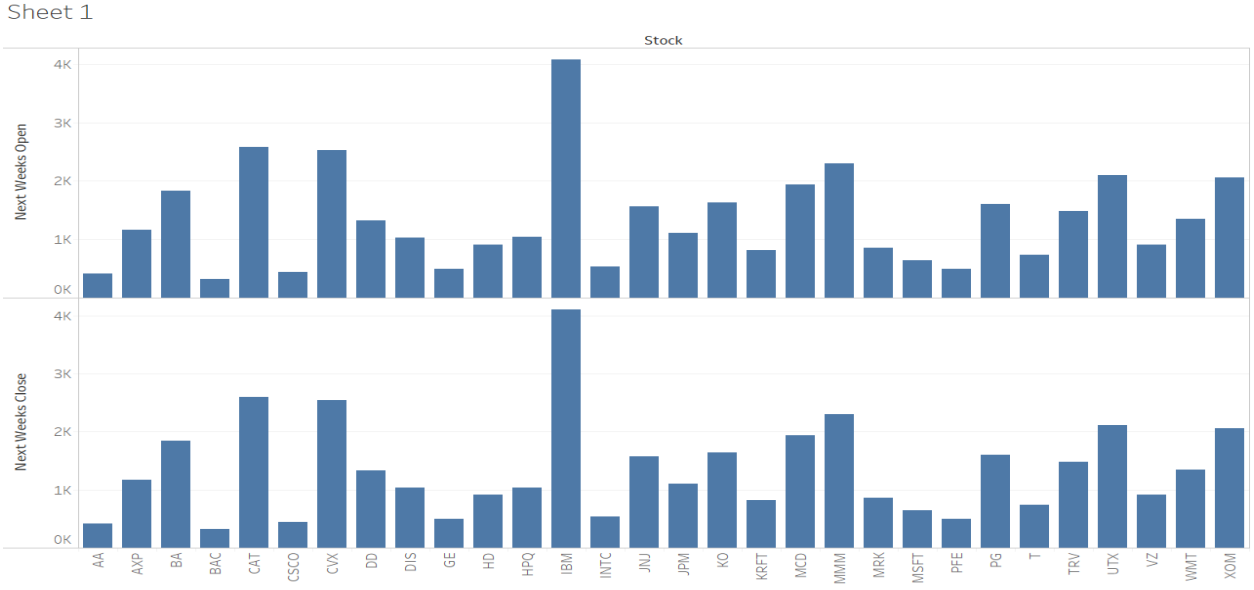
The number of percentage per dividend is the dividend growth is the rate of growth a particular stock undergoes over a period of time, the dividend is nothing but the stock earning that a particular company pays to its shareholders in terms of rate of stock growth, in histogram it can be clearly seen that t mobile, Verizon have the highest dividend return while the stock such as bank of America have the highest dividend return.



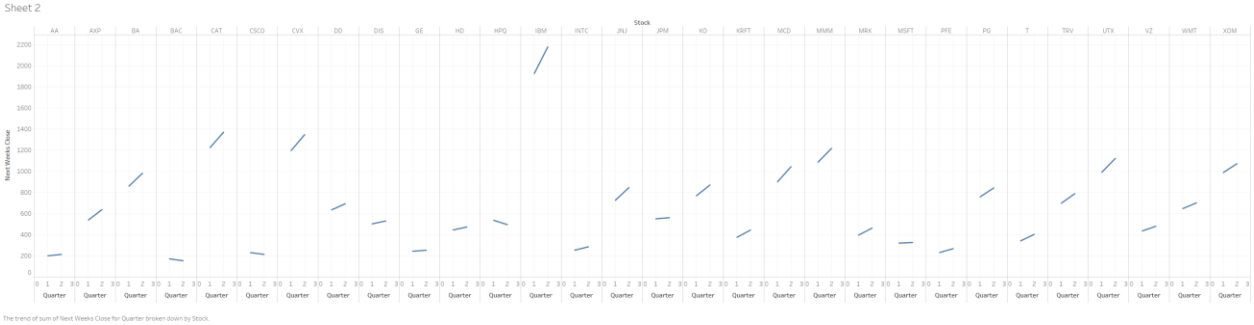


The percentage return next dividend is the value in the change in the value ,ie. change in the value of the dividend , the change is positive and highest for stock value of HPq.T mobile has the decrease.

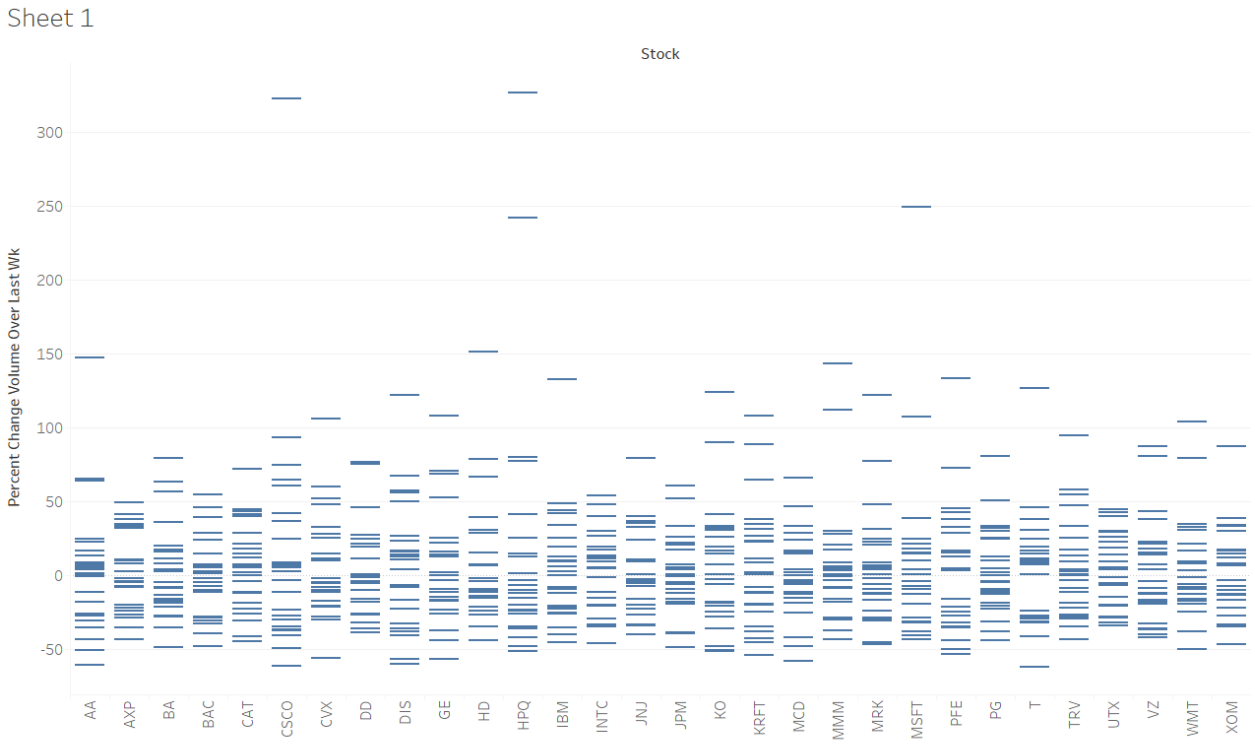
The next open and next close are the values that are predicted in the graph indicated given below, the graph show that IBM has the highest price where as the value of AA and BAC have the lowest price.



Sum of Next Weeks Open and sum of Next Weeks Close for each Stock.

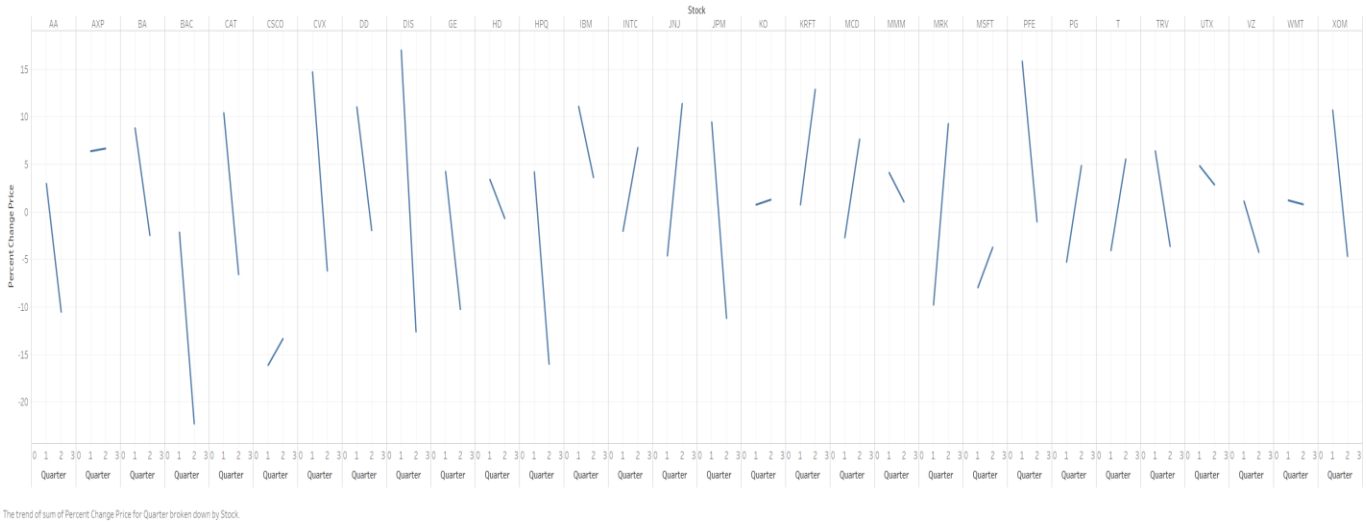


Percentage change in the volume over last week is the change in the percentage volume of each stock,the percentage change from what we can see is mostly negative,i.e below the 0,with stock like T and cisco having the value less than 50%, and stock values Cisco,HP having a positive change more than 300%.



Percent Change Volume Over Last Wk for each Stock.

The percentage change Price is the price change over the 3 quarters , from the graph the value of Disney changes / decreases from 1,2,3 quarters and the stock price of merk and co increases from the 3 quarters.



The trend of sum of Percent Change Price for Quarter broken down by Stock.

<u>METHODS</u> :

The dataset consisted of 750 observations and 16 variation,since the task is of prediction and seeing the factors of all the variables, the date , the next weeks opening price ,the next weeks closing price since they would not be known in real life, the training set consisted of first 360 observations from january till march while test set consisted of remaining observations from 361 -750 from april to june,the final variable were:

- Opening price
- Closing price
- Weekly high price
- Weekly low price
- Volume
- Percentage change in price
- Percentage change in volume
- Previous week volume
- Days to next dividend
- Percent return on next dividend
- Final variable(percentage change in next weeks price)

In the dataset i have replaced the missing values in the columns with respective median of the column with the help of following r code since removing the columns with na's will make the analysis/prediction as the extremes would not change and standard deviation would not affect much.

R code for filling the values of NA:

```
>e$percent_change_volume_over_last_wk[is.na(e$percent_change_volume_over_last_wk)]=
median(e$percent_change_volume_over_last_wk, na.rm=TRUE)
>e$previous_weeks_volume[is.na(e$previous_weeks_volume)]=median(e$previous_weeks_volu
me,na.rm=TRUE)
```

Knn:

Since we have been given the task of classification and clustering the algorithms that we will use are the Knn and the decision trees,Knn is a non parametric method used for classification and clustering, it is a method in which input consists of k closest training examples in the feature space, the output here is a class membership, meaning the majority vote is classified by output of its class membership,the output is the property value of the object,the value is the average value of the k neighbours.It is one of the simplest classification algorithms,it can give highly competitive results,this algorithm computes distance between each test example and the training examples to determine its nearest neighbor list.Choosing the proper value of k is very important because if the value of k is too large the classifier may misclassify the test instance and may include the data points that are far away from its neighbourhood, on the other hand a lower value of the k the classify may be susceptible to overfitting because of noise in the training dataset.

R Code for Knn:

```
> k1=knn(train[,1:10],test[,1:10],cl=train$final.change,k=1)
> k5=knn(train[,1:10],test[,1:10],cl=train$final.change,k=5)
> k10=knn(train[,1:10],test[,1:10],cl=train$final.change,k=10)
```

```
> kpred=data.frame(test,k1,k5,k10)
> k=data.frame(test,k1,k5,k10)
> with(k,table(final.change,k1))
          k1
final.change   0   1
         0        99 107
         1        90  94          (Accuracy= 49.48%)
> with(k,table(final.change,k5))
          k5
final.change   0   1
         0       66 140
         1       73 111          (Accuracy=45.34%)
> with(k,table(final.change,k10))
          k10
final.change   0   1
         0     67 139
         1     69 115          (Accuracy=46.66%)
```

The knn with different values was calculated and observed ,with different values i.e 1,5,10 and various accuracy rate was calculated the accuracy rate was highest for k=1, k=10 followed by k=5.

Decision tree:

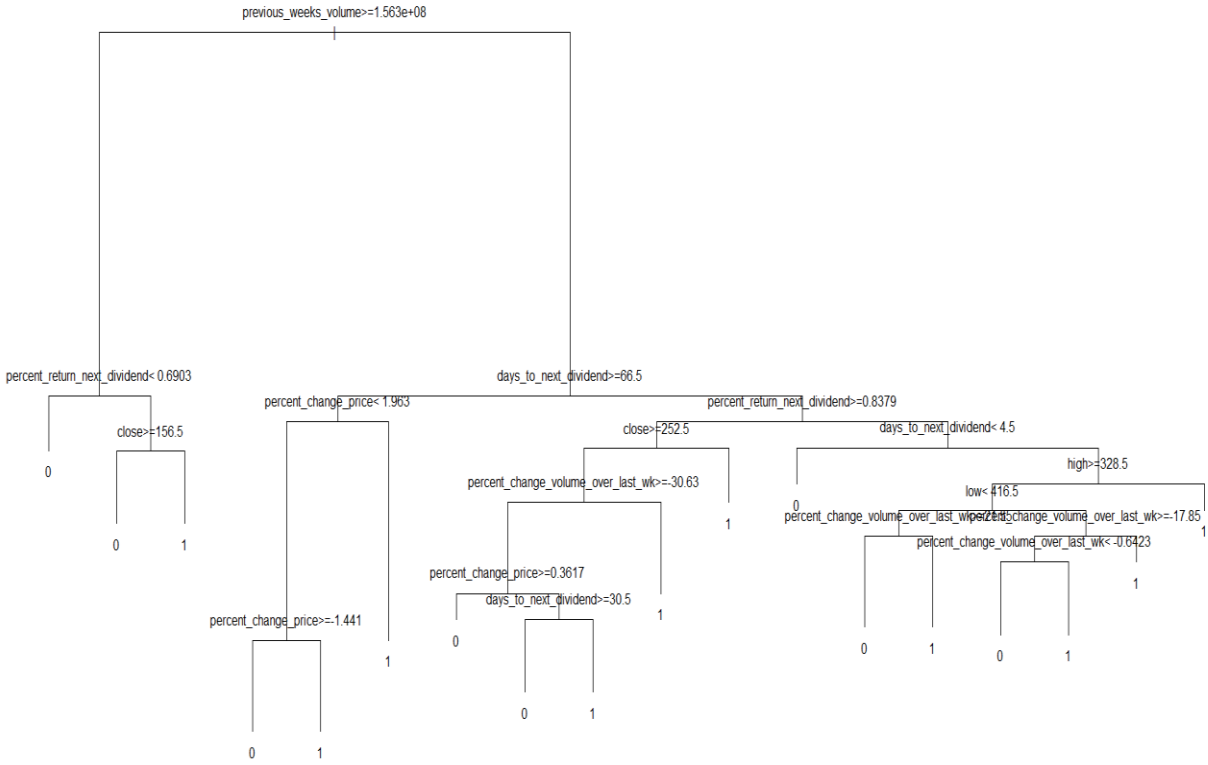
Decision tree is an algorithm in which direct algorithm is applied without preprocessing or tuning of the learning algorithm ,they are non-parametric non supervised used for classification and regression .It is an algorithm in which each internal node denotes a test on a attribute, each branch represents the outcome of a test. Decision tree is an algorithm in which the it involves deciding which features to choose and what are the conditions used for splitting .In decision tree whole training set is considered as a root , features are preferred to be categorical, the popular attribute selection measures include information gain and the gini index, the step in calculating information gain includes calculation of entropy of an target , one of the problem in the decision trees is the overfitting , information gain in the decision tree with categorical variable gives biased response and the calculations also sometimes become too complex.

Rcode:

```
>rtree=rpart(final.change~open+high+low+close+volume+percent_change_price+percent_change_volume_over_last_wk+previous_weeks_volume+days_to_next_dividend+percent_return_next_dividend,data=train,method='class')
> pred1 = predict(rtree, test, type="class")
> plot(rtree)
> text(rtree)
> confusionmatrix = table(test$final.change, pred1)
> accuracy=sum(diag(confusionmatrix))/390
> accuracy
[1] 0.5102564=    51.0%
```

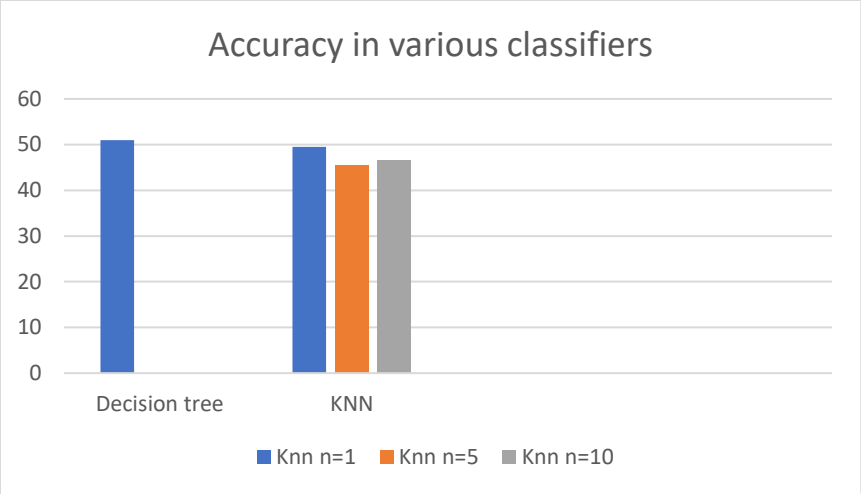
> error=(390-sum(diag(confusionmatrix)))/390
> error
[1] 0.4897436= 48.9%

Classification tree:



RESULTS:
From the results we can see that the accuracy in the decision is more compared to then knn ,and the , accuracy in the algorithms is as follows:
(1)KNN with k=1, Accuracy=49.48%
k=5, Accuracy= 45.34%
k=10, Accuracy= 46.66%

(2)Decision tree , Accuracy=51%
From the results we can see that the accuracy in the decision tree is more than compared to the knn , the accuracy is decision tree which is more than 50% is considered to be a positive in the stock market since an algorithm with more than 50% accuracy could be profitable in stock market.



DISCUSSION:

The algorithms that were used in the analysis while most of them had accuracy around 50, while the accuracy in knn was less than 50%,the accuracy in the decision tree generated on the test set was not that too impressive with cross a mere 50%, but still in the unpredictable world of stock market even a classifier with accuracy rate more than 50% can be considered worth a good deal.While decision trees had accuracy more than 50%, other algorithms such as random forest , Xgboost , and other could be used to check the accuracy and error rate and report the same.Stock market prediction involves years and years of experience , even the most experienced traders sometimes loose money , infact 90% of people exist in the market who loose money and are in losses and there exist a mere 10% of people who are making, however to advancement of data and technology, new machine algorithms are being created and implemented which is making the profit percentage to grow more and increase, with the age of data scientists, and the data science being called as the sexiest job of the 21st century by the Harvard business magazine , data scientists are being in wall street in large numbers and thus the algorithmic trading and other normal trading is being replaced by machine learning.

CONCLUSION:

The overall work of algorithm for modelling the stock data on training and test set did the job as per the requirement, the overall accuracy was actually achieved being around 50% , though it would have been great if it was gearing near 60s to 70s.Though various combinations of closing high , opening low, weekly high , dividend percentages and other stuff could have been used which could have improved the performance accuracy of algorithm,also the use of the algorithms on the data from other markets and other years could have made the algorithm more reliable and robust, the index pattern could have used and determined whether a stock a particular index is doing well , or particular index is doing well or bad.(Index here refers to various sectors of the market example for auto it is the auto sector, for IT is the IT sector,bank etc., if the auto sector is doing well i.e all the other stocks or the majority of stocks that come under that sector would perform well, same happens when the particular sector performs bad i.e stocks falling under that sector would also perform pretty low.In conclusion, with chances of

ups and downs pretty high, the classifier performed quite decently and would help the particular portfolio or hedge fund manager to gain some profits.