Stock Market Analysis

By Swapna Samir Shukla

- **MySQL**

Inside the database "BDHS_PROJECT" we have the two tables(STOCK_COMPANIES and STOCK_PRICES) which will be used for analysis .

```
use BDHS_PROJECT;
```

SHOW TABLES;

```
+-------------------------+
| Tables_in_BDHS_PROJECT |
+-------------------------+
| STOCK_COMPANIES         |
| STOCK_PRICES            |
+-------------------------+
2 rows in set (0.00 sec)
```

Let us have a quick view at the datasets

select * from STOCK_COMPANIES limit 5;

```
+--------+--------------------------+------------------------+----------------------+--------------------------+
| Symbol | Company_name             | Sector                 | Sub_industry         | Headquarter              |
+--------+--------------------------+------------------------+----------------------+--------------------------+
 |A      | Agilent Technologies Inc | Health Care            | Health Care Equipment | Santa Clara; California
       | | American Airlines Group  | Industrials            | Airlines             | Fort Worth; Texas
       | | Advance Auto Parts       | Consumer Discretionary | Automotive Retail    | Roanoke; Virginia
   |PL   | Apple Inc.               | Information Technology | Computer Hardware    | Cupertino; California
 |ABBV   | AbbVie                   | Health Care            | Pharmaceuticals      | North Chicago; Illinois
+--------+--------------------------+------------------------+----------------------+--------------------------+
5 rows in set (0.00 sec)
```

select * from STOCK_PRICES limit 5;

```
+--------------+--------+------------+------------+------------+------------+---------+
| Trading_date | Symbol | Open       | Close      | Low        | High       | Volume  |
+--------------+--------+------------+------------+------------+------------+---------+
| 2016-01-05   | WLTW   |     123.43 | 125.839996 | 122.309998 |     126.25 | 2163600 |
| 2016-01-06   | WLTW   | 125.239998 | 119.980003 | 119.940002 | 125.540001 | 2386400 |
| 2016-01-07   | WLTW   | 116.379997 | 114.949997 |     114.93 | 119.739998 | 2489500 |
| 2016-01-08   | WLTW   | 115.480003 | 116.620003 |      113.5 | 117.440002 | 2006300 |
| 2016-01-11   | WLTW   | 117.010002 | 114.970001 | 114.089996 | 117.330002 | 1408600 |
+--------------+--------+------------+------------+------------+------------+---------+
```

- **Sqoop**

We create a data pipeline using sqoop to pull the data from the MySQL server into Hive.

sqoop import --connect jdbc:mysql://ip-10-0-1-10.ec2.internal/BDHS_PROJECT --username labuser -password simplilearn --table Stock_companies –hive-import -hive-database stock_db --m 1

sqoop import --connect jdbc:mysql://ip-10-0-1-10.ec2.internal/BDHS_PROJECT --username labuser -password simplilearn --table Stock_prices –hive-import -hive-database stock_db --m 1

- **Hive**

Now we create a new hive table (stock_data) by joining the above two hive tables (stock_companies and stock_prices).

create table stock_data as select trading_year, trading_month, sc.symbol,
company_name, trim(split(headquarter,"\;")[1]) state, sector, sub_industry,
open, close, low, high, volume from stock_companies sc,
(select symbol, year(trading_date) trading_year, month(trading_date) trading_month,
round(avg(open),2) open, round(avg(close),2) close, round(avg(low),2) low, round(avg(high),2) high,
round(avg(volume),2) volume from stock_prices group by symbol,
month(trading_date),year(trading_date)) sp where sc.symbol=sp.symbol;

select * from stock_data limit 5;

```
2010    1       A       Agilent Technologies Inc        California      Health Care     Health Care Equipment   21.72   21.61
.86   4208442.11
2011    1       A       Agilent Technologies Inc        California      Health Care     Health Care Equipment   30.29   30.29
.65   4496845.0
2012    1       A       Agilent Technologies Inc        California      Health Care     Health Care Equipment   28.54   28.78
.08   5069975.0
2013    1       A       Agilent Technologies Inc        California      Health Care     Health Care Equipment   31.2    31.26
.45   4567819.05
2014    1       A       Agilent Technologies Inc        California      Health Care     Health Care Equipment   42.01   42.04
.36   3494200.0
```

Analysis

1) Find the top five companies that are good for investment

create table stock_table1 as select company_name, min(trading_year) min_year,
max(trading_year) max_year, min(trading_month) min_month, max(trading_month)
max_month from stock_data group by company_name;

i)Next we find the growth percent for each company over the years

select stock_start.company_name, ((close-open)/open)*100 growth_percent
from(select t1.company_name, open from stock_data sd, stock_table1 t1
where sd.trading_year = t1.min_year and sd.trading_month = t1.min_month and
sd.company_name = t1.company_name) stock_start, (select t1.company_name,

```
close from stock_data sd, stock_table1 t1
where sd.trading_year = t1.max_year and sd.trading_month = t1.max_month
and sd.company_name = t1.company_name) stock_end where
stock_start.company_name = stock_end.company_name sort by
growth_percent desc limit 5;
```

Top 5 Companies by growth

```
Netflix Inc.    1536.0158311345647
Regeneron       1382.2714681440443
Ulta Salon Cosmetics & Fragran  1174.9378418697165
United Rentals; Inc.    1064.340239912759
Alaska Air Group Inc    878.5555555555554
```

2) Show the best-growing industry by each state, having at least two or more industries mapped.

i) First we calculate the growth of companies belonging to every state and capital.

```
create table stock_table2 as select
state, sub_industry, stock_start.company_name, ((stock_end.close-
stock_start.open)/stock_start.open)*100 growth_percent
from  (select  t1.company_name,open
         from stock_data sd, stock_table1 t1
         where sd.trading_year=t1.min and
                  sd.trading_month=t1.min_month and
                  sd.company_name=t1.company_name)stock_start,
       (select t1.company_name, close
         from stock_data sd, stock_table1 t1
         where sd.trading_year=t1.max and
                  sd.trading_month=t1.max_month and
                  sd.company_name=t1.company_name)stock_end,
        (select company_name, state, sub_industry
         from stock_data
         group by company_name,state,sub_industry)sd
where (stock_end.close-stock_start.open)>0 and
        stock_start.company_name=stock_end.company_name and
        sd.company_name=stock_start.company_name;
```

select * from stock_table2 limit 5;

```
Minnesota       Industrial Conglomerates        3M Company      112.60817307692305
Georgia Life & Health Insurance AFLAC Inc       38.92215568862274
Pennsylvania    Electrical Components & Equipm  AMETEK Inc      192.15801886792448
Texas   Integrated Telecommunications   AT&T Inc        55.250282273240494
Illinois        Pharmaceuticals AbbVie  73.49295774647888
```

ii) Then we group by state and capital and filter industries listed atleast twice

create table stock_table3 as select state,sub_industry,
avg(growth_percent)ind_growth from stock_table2 group by state, sub_industry
having count(sub_industry>=2);

Select * from stock_table3 limit 10;

```
California      Application Software    120.31201428397627
California      Health Care Equipment   147.29321666303753
California      Internet Software & Services    336.56041800779366
California      REITs   131.08399426163498
California      Semiconductor Equipment 147.9653267908011
California      Semiconductors  282.4593566657631
Massachusetts   Health Care Equipment   169.22809835926253
New Jersey      Health Care Equipment   120.88893239332754
New York        Apparel; Accessories & Luxury   58.380432276158544
New York        Banks   86.82510314646063
Time taken: 0.054 seconds, Fetched: 10 row(s)
```

iii) Finally we find the industry which has maximum growth by each state

select t3.state, sub_industry, ind_growth from stock_table3 t3,
(select state,max(ind_growth) max_growth from stock_table3
group by state) max_ind where max_ind.state = t3.state and
t3.ind_growth = max_ind.max_growth;

best growing industry by each state

```
California      Internet Software & Services    336.56041800779366
Massachusetts   Health Care Equipment   169.22809835926253
New Jersey      Health Care Equipment   120.88893239332754
New York        Diversified Financial Services  244.27936670090116
Ohio    Banks   171.58635235361513
Texas   Oil & Gas Refining & Marketing  247.75563965526214
```

3) For each sector find the following.

- **Worst year**
- **Best year**

i) AS a first step we find the growth for each sector and year

```
create table stock_table4 as select open.sector, open.trading_year,
(close-open) growth from (select sector,trading_year,avg(open) open from
stock_data where trading_month = 1 group by sector,trading_year) open,
(select sector,trading_year,avg(close) close from stock_data
where trading_month=12 group by sector,trading_year) close
where open.sector = close.sector and open.trading_year = close.trading_year;
```

```
select * from stock_table4 limit 5;
```

```
Consumer Discretionary  2010    14.404047387211953
Consumer Discretionary  2011    4.860126582278504
Consumer Discretionary  2012    7.730506329113922
Consumer Discretionary  2013    24.392572590011653
Consumer Discretionary  2014    9.603809523809531
```

ii) In the year 2011, 3 sectors attained their new low

```
select x.sector,x.trading_year,x.growth from stock_table4 x,
(select sector,min(growth) growth from stock_table4
group by sector) y where x.sector=y.sector and
x.growth=y.growth;
```

Year in which each of the sectors were worst hit

```
Consumer Discretionary  2011    4.860126582278504
Consumer Staples        2016    3.183055555555569
Energy  2015    -10.09944444444443
Financials      2011    -6.860689655172415
Health Care     2016    2.0805084745762485
Industrials     2015    -2.640000000000029
Information Technology  2011    -2.9025396825396896
Materials       2011    -3.967083333333335
Real Estate     2013    -4.463448275862078
Telecommunications Services     2015    -2.29399999999999
Utilities       2015    -6.473928571428566
```

iii) In the year 2016 and 2014 most of the sectors enjoyed a very high growth.

```
select a.sector,a.trading_year,a.growth from stock_table4 a,

(select b.sector,max(growth) growth from stock_table4
group by sector) b where a.sector=b.sector and
a.growth=b.growth;
```

Best year for each sector

Consumer Discretionary	2013	24.392572590011653
Consumer Staples	2014	9.322857142857139
Energy	2016	18.642222222222223
Financials	2016	16.326612903225815
Health Care	2014	24.325254237288092
Industrials	2016	19.804925404944584
Information Technology	2013	15.538281249999997
Materials	2016	19.036400000000015
Real Estate	2014	18.892413793103472
Telecommunications Services	2014	5.059999999999995
Utilities	2014	10.268571428571427