LOAD DATA:

Data is loaded using the .sql script provided.

Step 1:

How many different firms are represented in at least 30 days in the main dataset?

SELECT count(*), symbol FROM (SELECT DISTINCT(datestart),symbol FROM Tweets) AS t GROUP BY symbol HAVING count(*)>=30;

Explanation: This query will give the list of different firms which were represented in the data set in at least 30 days. The sub query will select distinct start dates and symbols of firms from Tweets table and then it will return count and symbols.

Output:

The file Step_1.csv contains the result.

SELECT count(*) FROM (SELECT count(*) , symbol FROM (SELECT DISTINCT (datestart), symbol FROM Tweets) AS t GROUP BY symbol HAVING count(*)>30) AS c;

Explanation: This query will give only the number of firms which were represented in at least 30 days.

Output:

count(*)
96

Step 2:

Create and populate a table which shows average Twitter levels immediately following each earnings release, and a comparative baseline average for the firm and specific time period.

a) Average tweet levels (tweets per minute) in the 40 minutes following earnings release.

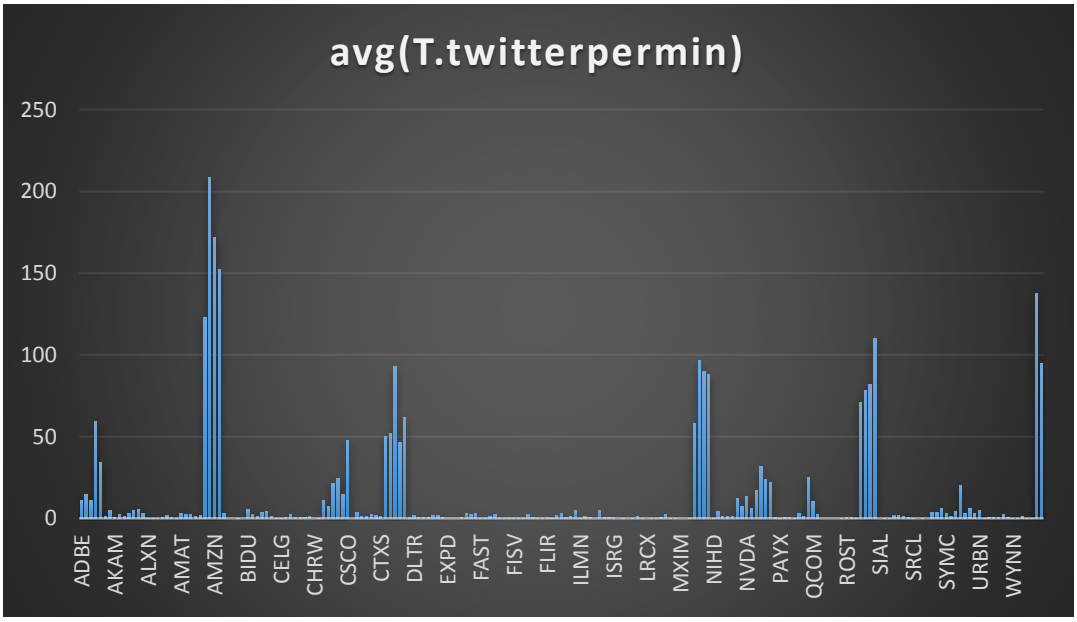
SELECT T.smblid, avg(T.twitterpermin) FROM Tweets T, EarnRelMatched E1 where T.datestart = E1.earnrelease_date and T.smblid = E1.ticker and T.timestart between E1.earnrelease_time and date_add(E1.earnrelease_time, interval 40 minute) group by T.smblid, T.datestart;

Explanation: This query will return the list of firms and their average twitter levels within 40 minutes of their earnings release. The query is written by matching tweets start date and earnings release date and tweets symbol and symbol id in earning release matched. Starting time of tweets should be between an interval of earnings release time and 40 minutes added to it.

Output:

The file Step_2a.csv contains result.

Graphical representation:



From the graph, we can observe that firm Amazon (AMZN) has highest average tweets per minute within 40 minutes of their earnings release date.

b) Average tweet levels (tweets per minute) in the 2 hours following earnings release.

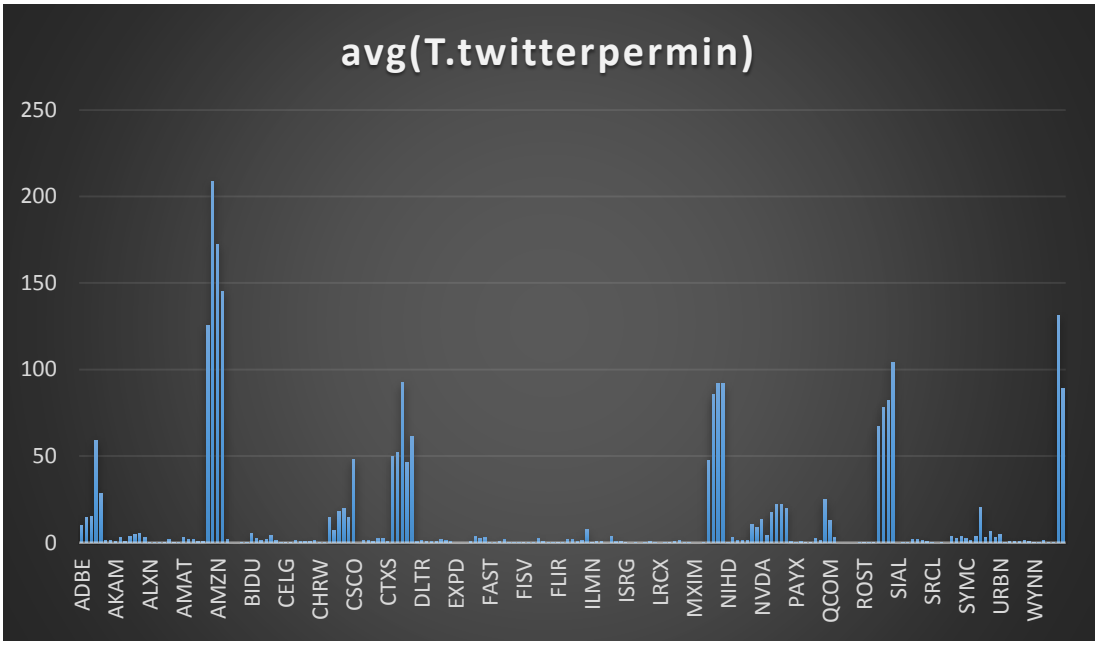
SELECT T.smblid, avg(T.twitterpermin) FROM Tweets T, EarnRelMatched E1 where T.datestart = E1.earnrelease_date and T.smblid = E1.ticker and T.timestart between E1.earnrelease_time and date_add(E1.earnrelease_time, interval 2 hour) group by T.smblid, T.datestart;

Explanation: This query will return the list of firms and their average twitter levels within 2 hours of their earnings release. The query is written by matching tweets start date and earnings release date and tweets symbol and symbol id in earning release matched. Starting time of tweets should be between an interval of earnings release time and 2 hours added to it.

Output:

The file Step_2b contains the result.

Graphical representation:



From the graph, we can observe that firm Amazon (AMZN) has highest number of average tweets per minute within 2 hours of its earnings release.

c) Average tweet levels (tweets per minute) in the 24 hours following earnings release.

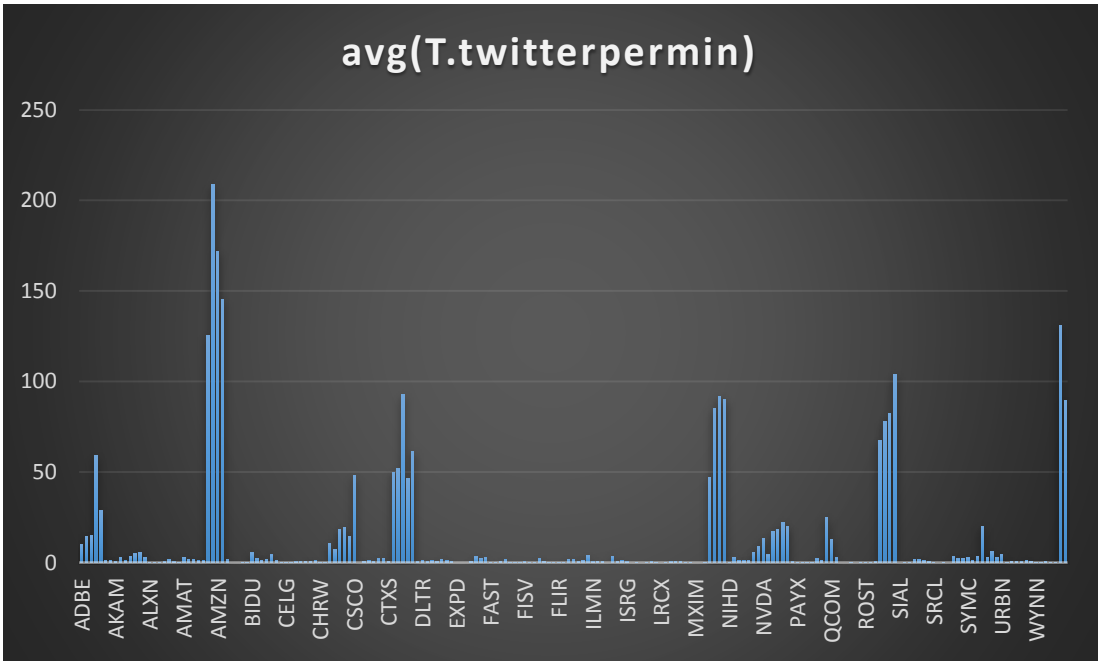
SELECT T.smblid, avg(T.twitterpermin) FROM Tweets T, EarnRelMatched E1 where T.datestart = E1.earnrelease_date and T.smblid = E1.ticker and T.timestart between E1.earnrelease_time and date_add(E1.earnrelease_time, interval 24 hour) group by T.smblid, T.datestart;

Explanation: This query will return the list of firms and their average twitter levels within 24 hours of their earnings release. The query is written by matching tweets start date and earnings release date and tweets symbol and symbol id in earning release matched table. Starting time of tweets should be between an interval of earnings release time and 24 hours added to it.

Output:

The file Step_2c contains the result.

Graphical representation:



From the graph, firm Amazon (AMZN) has highest tweets per minute within 24 hours of its earnings release time.

d) Average tweet levels (tweets per minute) in the 1 week following earnings release.

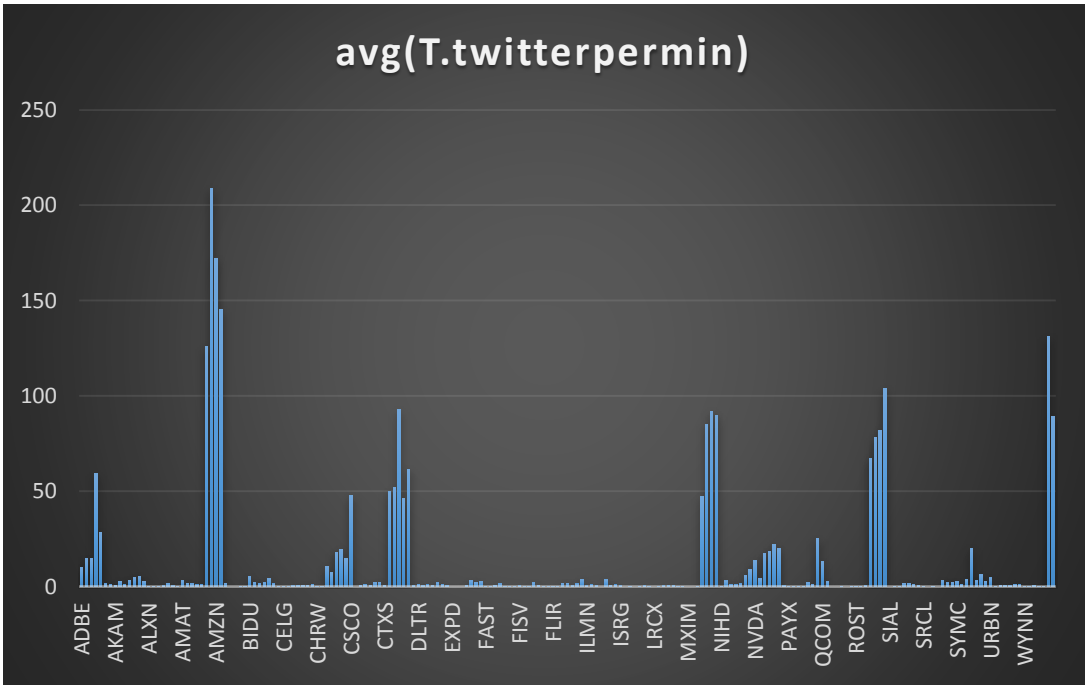
SELECT T.smblid, avg(T.twitterpermin) FROM Tweets T, EarnRelMatched E1 where T.datestart = E1.earnrelease_date and T.smblid = E1.ticker and T.timestart between E1.earnrelease_time and date_add(E1.earnrelease_time, interval 1 week) group by T.smblid, T.datestart;

Explanation: This query will return the list of firms and their average twitter levels within 1 week of their earnings release. The query is written by matching tweets start date and earnings release date and tweets symbol and symbol id in earning release matched table. Starting time of tweets should be between an interval of earnings release time and 1 week added to it.

Output:

The file Step_2d.csv contains the result.

Graphical representation:



From the graph, firm Amazon (AMZN) has highest average tweets within 1 week of its earnings release time.

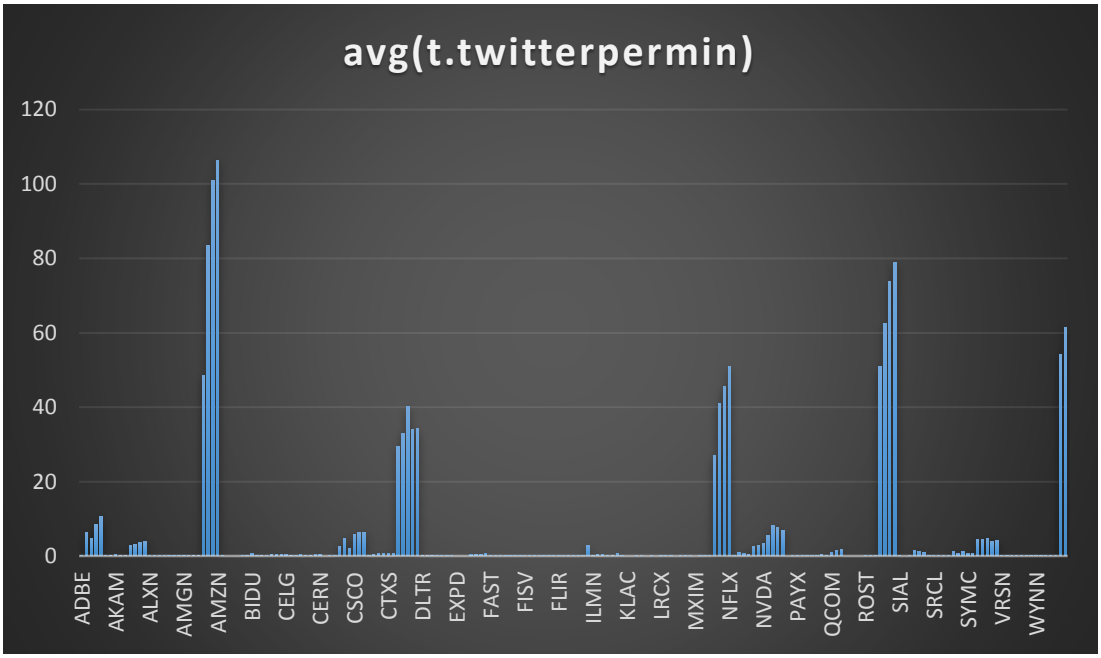
e & i) Calculate Baseline Average

select t.symbol, avg(t.twitterpermin) from tweets t, earnrelmatched e where e.ticker=t.smblid and t.timestart between e.earnrelease_time and date_add(e.earnrelease_time, interval 30 minute) and dayofweek(t.datestart)=dayofweek(e.earnrelease_date) and t.datestart between 2012/05/10 and e.earnrelease_date-7 group by t.smblid, e.earnrelease_date;

Explanation: earnrelease dates are compared with before weeks and twitter per minute are averaged.

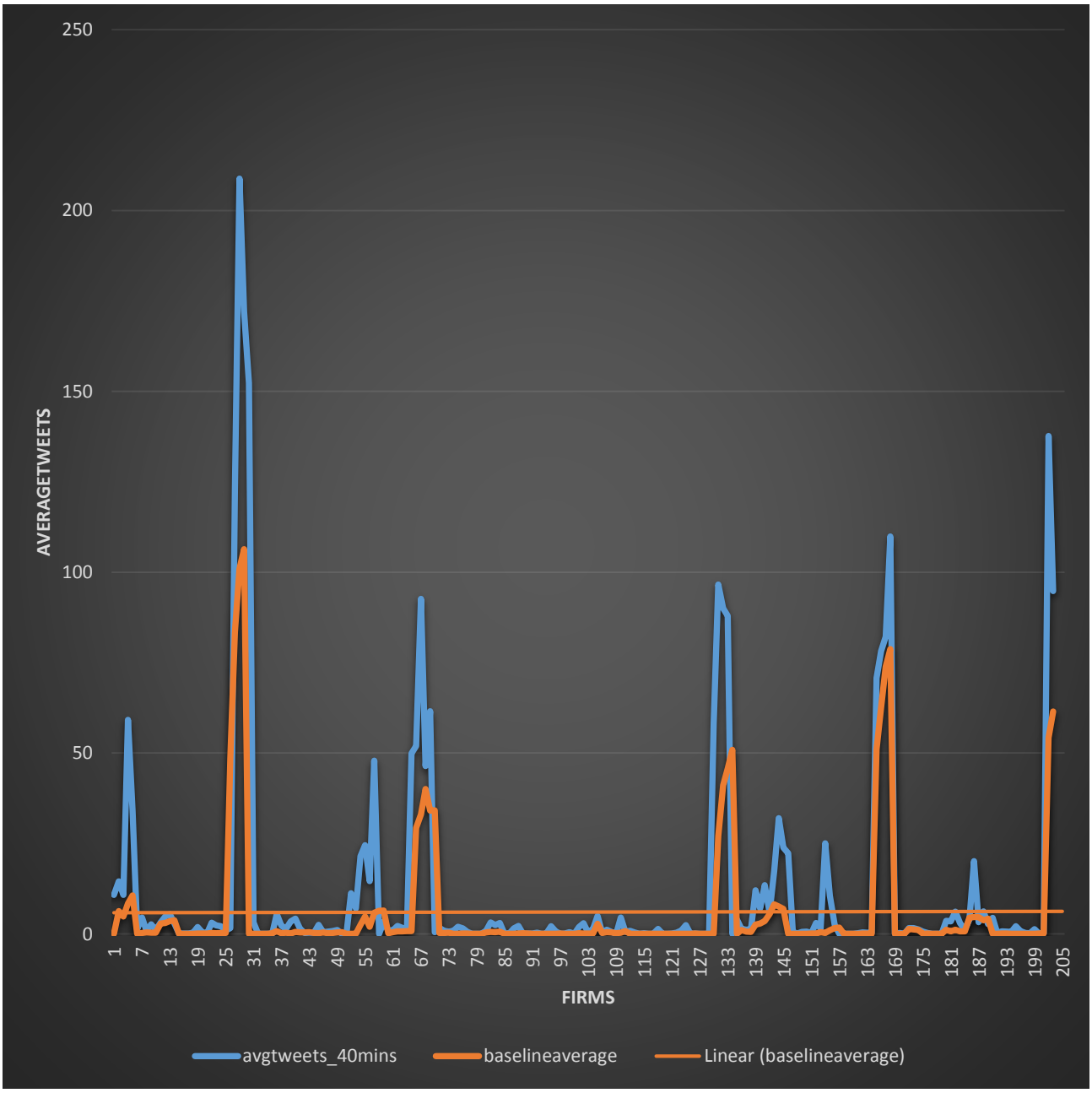
Output:

The file Step_2e.csv contains the result.



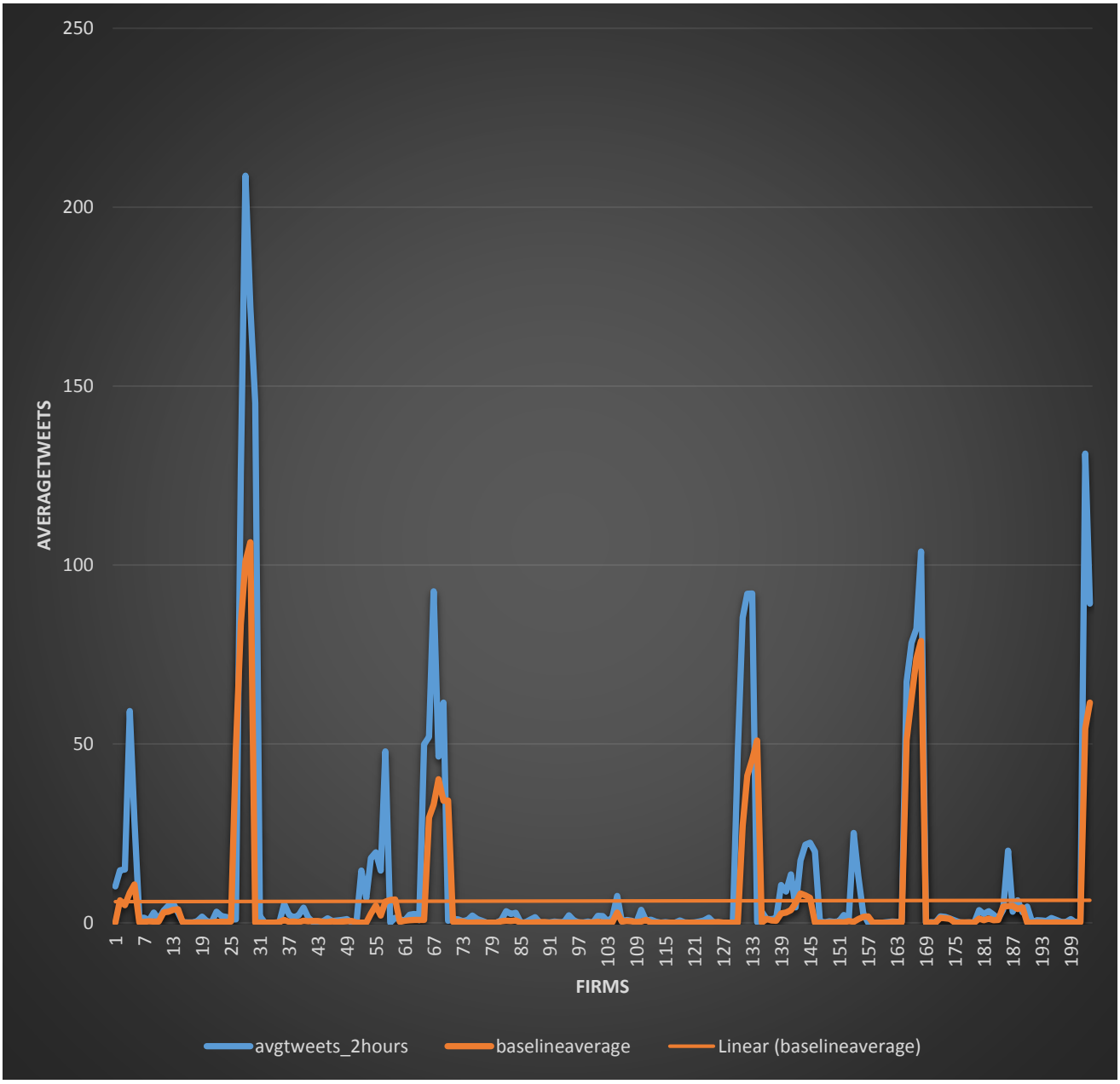
f) Compare the average Tweet levels in (a) through (d) with the baseline average described in (e).

Comparing average tweets per minute within 40 minutes and baseline average



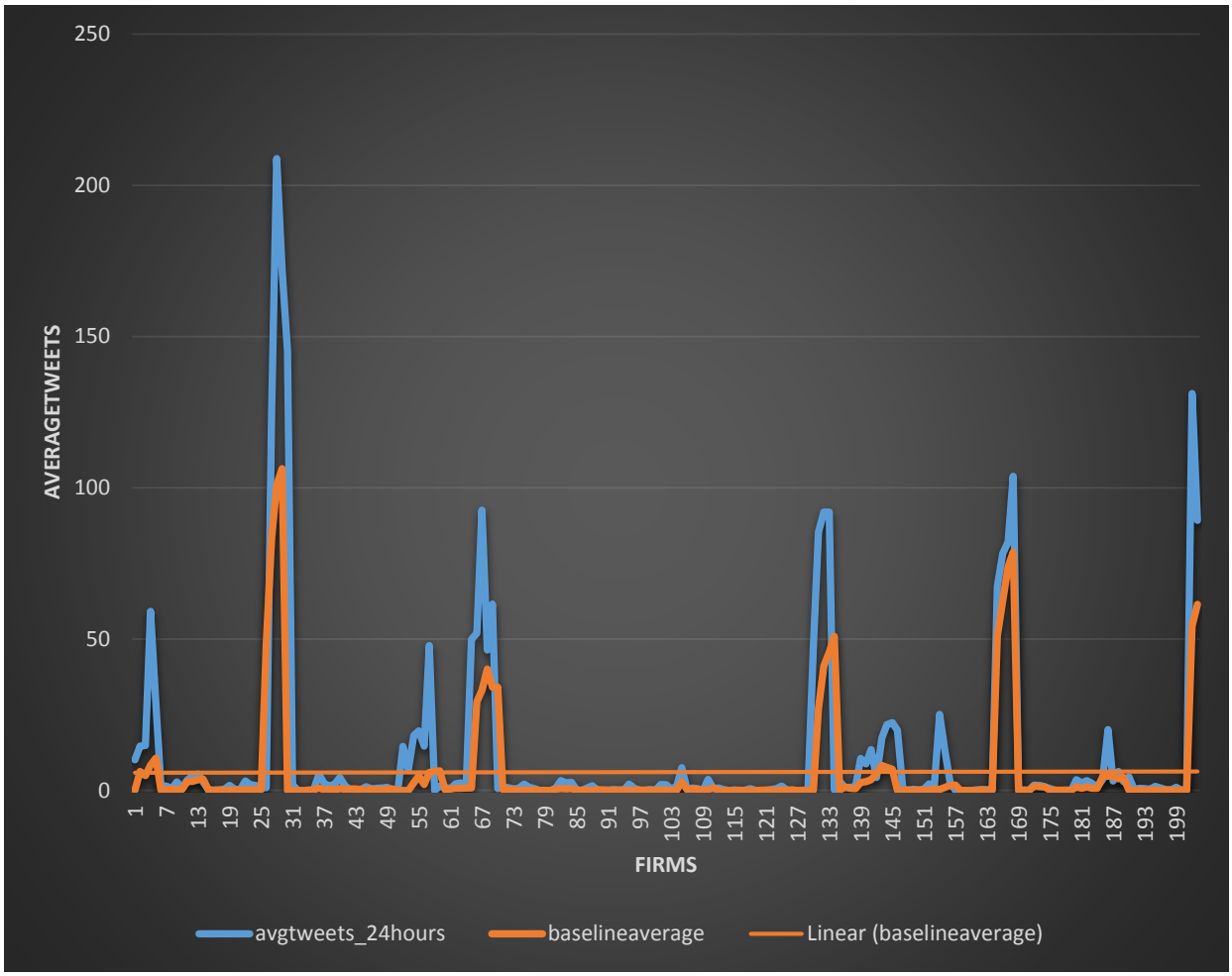
Blue line indicates average tweets per minute within 40 minutes and orange line indicates baseline average. For all firms, average tweets per minute within 40 minutes is greater than baseline average.

Comparing average tweets per minute within 2 hours and baseline average



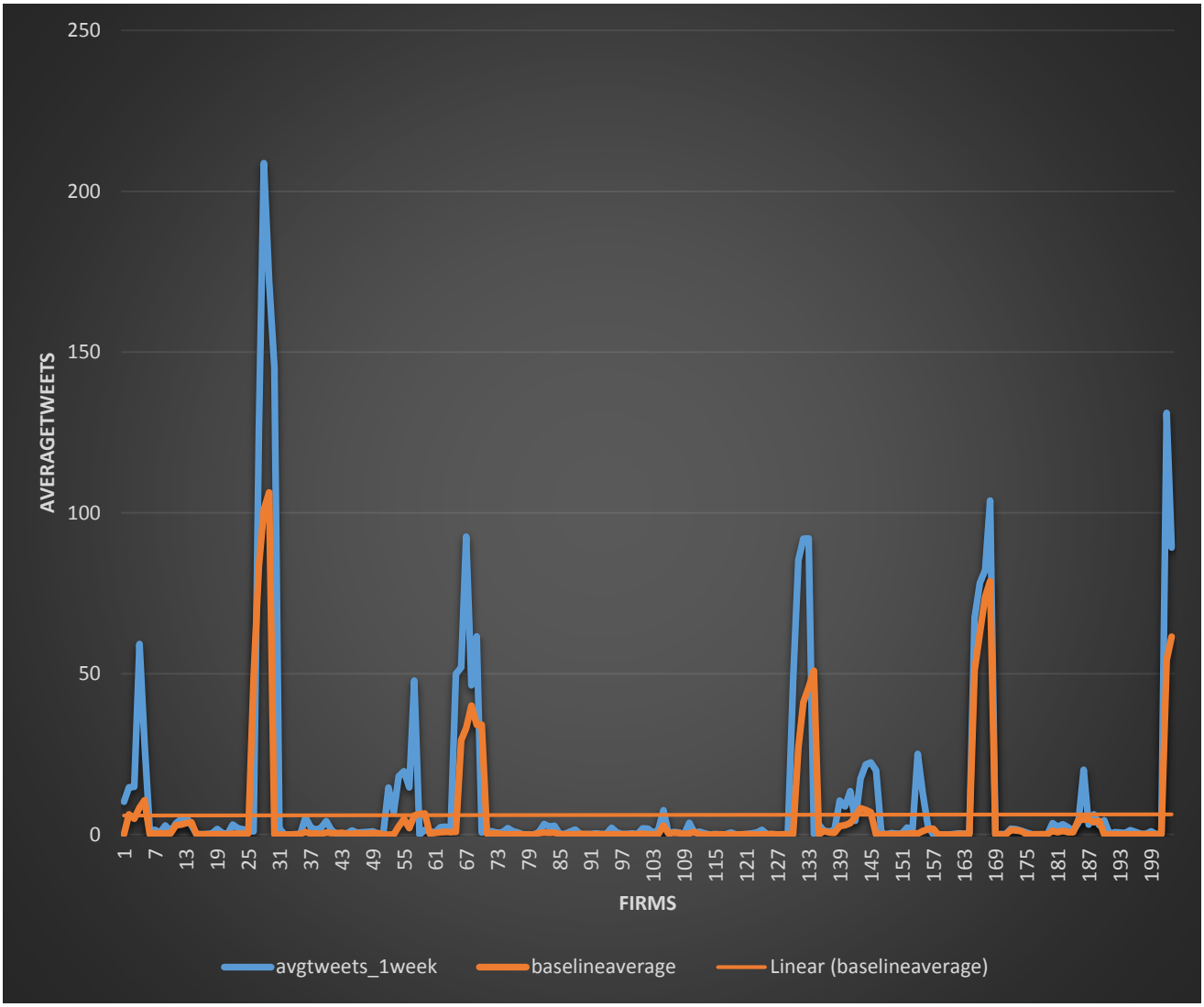
Blue line indicates average tweets per minute within 2 hours and orange line indicates baseline average. For all firms, average tweets per minute within 2 hours is greater than baseline average.

Comparing average tweets per minute within 24 hours and baseline average



Blue line indicates average tweets per minute within 24 hours and orange line indicates baseline average. For all firms, average tweets per minute within 24 hours is greater than baseline average.

Comparing average tweets per minute within 1 week and baseline average



Blue line indicates average tweets per minute within 1 week and orange line indicates baseline average. For all firms, average tweets per minute within 1 week is greater than baseline average.

Step 3:

Create and populate a table which shows average Trading volume immediately following each earnings release, and a comparative baseline average for the firm and specific time period.

a) Average trading volume levels for the firm in the 40 minutes following earnings release.

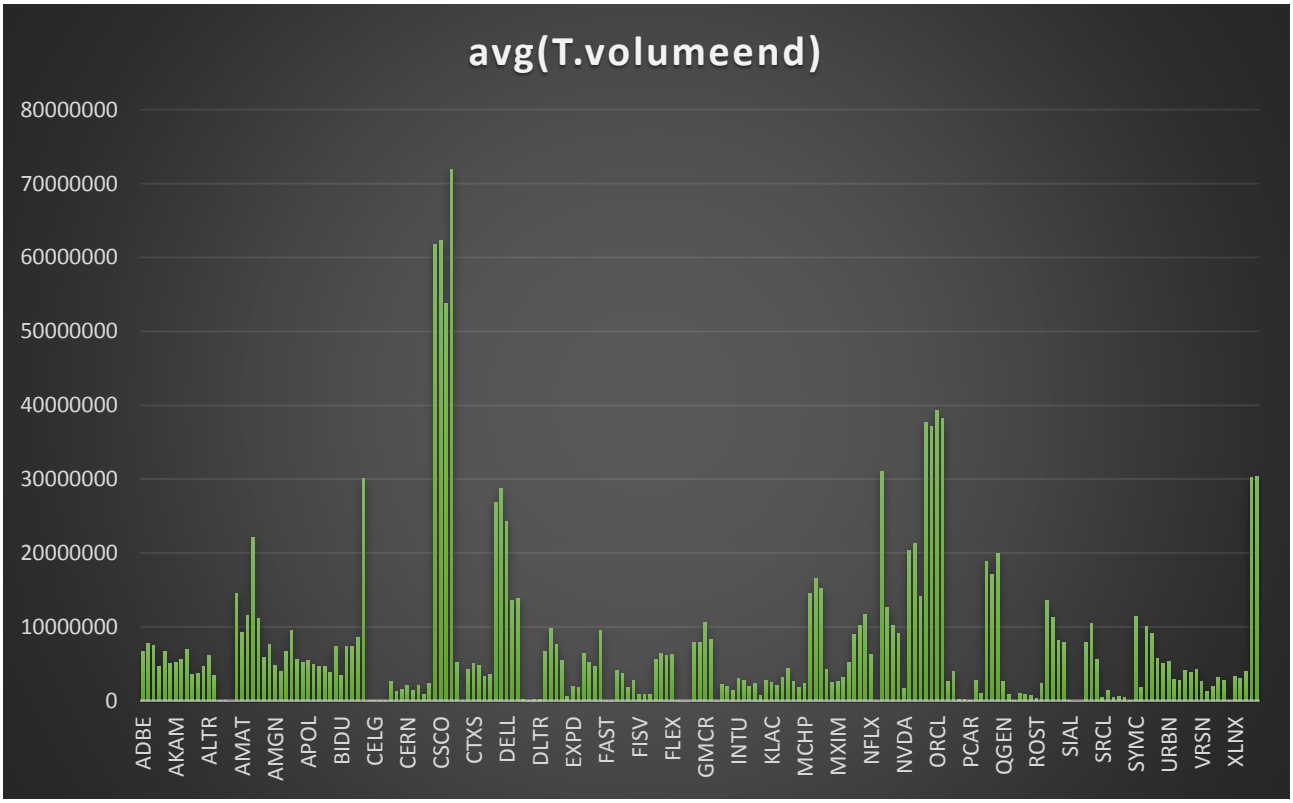
SELECT T.smblid, avg(T.volumeend) FROM Tweets T, EarnRelMatched E1 where T.datestart = E1.earnrelease_date and T.smblid = E1.ticker and T.timestart between E1.earnrelease_time and date_add(E1.earnrelease_time, interval 40 minutes) group by T.smblid, T.datestart;

Explanation: This query will return the list of firms and their average trading volume within 40 minutes of their earnings release. The query is written by matching tweets start date and earnings release date and tweets symbol and symbol id in earning release matched table. Starting time of tweets should be between an interval of earnings release time and 40 minutes added to it.

Output:

The Step_3a.csv contains the result.

Graphical representation:



From the graph, firm CSCO has highest average trading volume within 40 minutes of earnings release time.

b) Average trading volume levels for the firm in the 2 hours following earnings release.

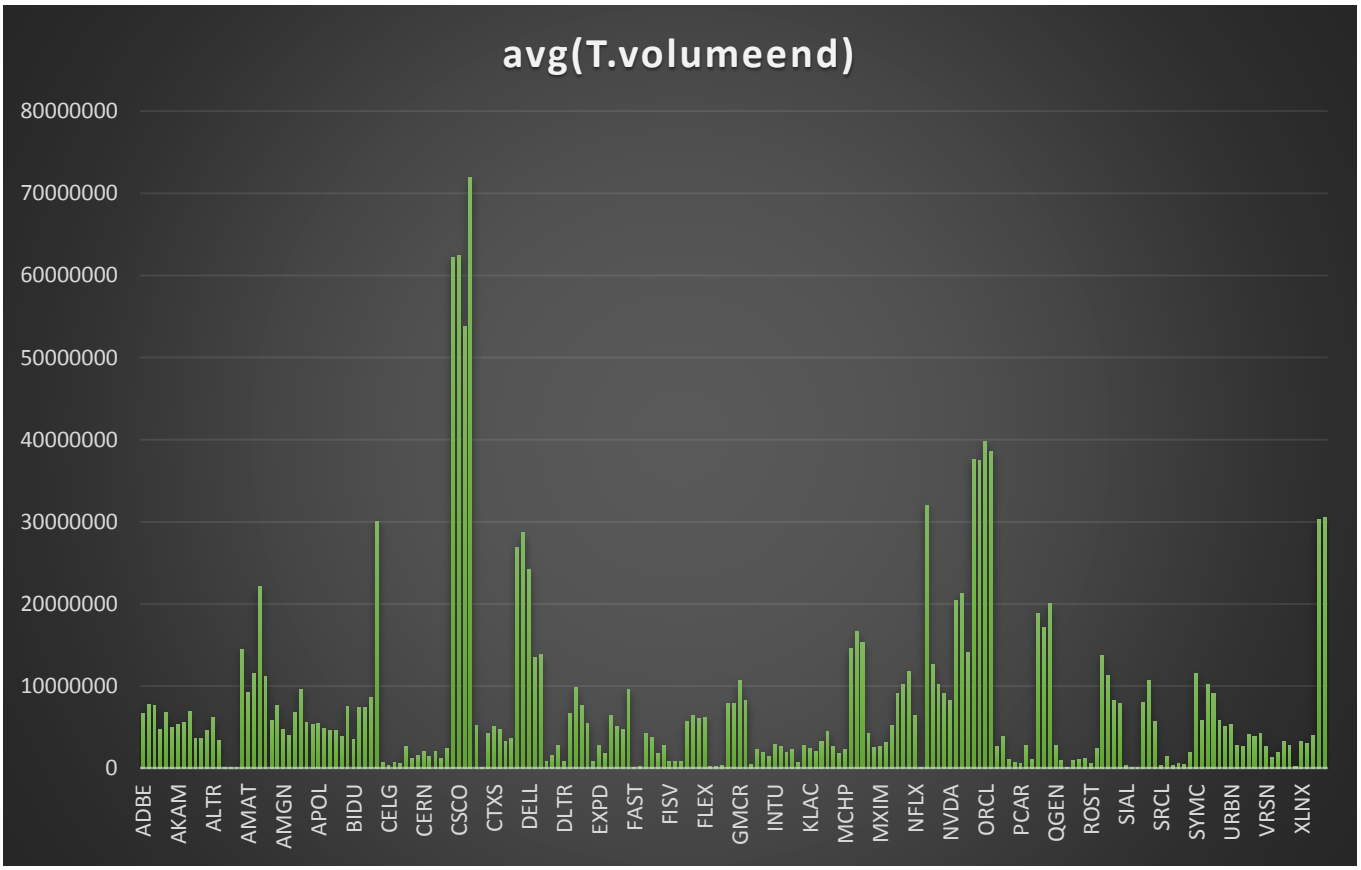
SELECT T.smblid, avg(T.volumeend) FROM Tweets T, EarnRelMatched E1 where T.datestart = E1.earnrelease_date and T.smblid = E1.ticker and T.timestart between E1.earnrelease_time and date_add(E1.earnrelease_time, interval 2 hours) group by T.smblid, T.datestart;

Explanation: This query will return the list of firms and their average trading volume within 2 hours of their earnings release. The query is written by matching tweets start date and earnings release date and tweets symbol and symbol id in earning release matched table. Starting time of tweets should be between an interval of earnings release time and 2 hours added to it.

Output:

The file Step_3b.csv contains the result.

Graphical representation:



From the graph, firm CSCO has highest average trading volume levels within 2 hours of its earnings release time.

c) Average trading volume levels for the firm in the 24 hours following earnings release.

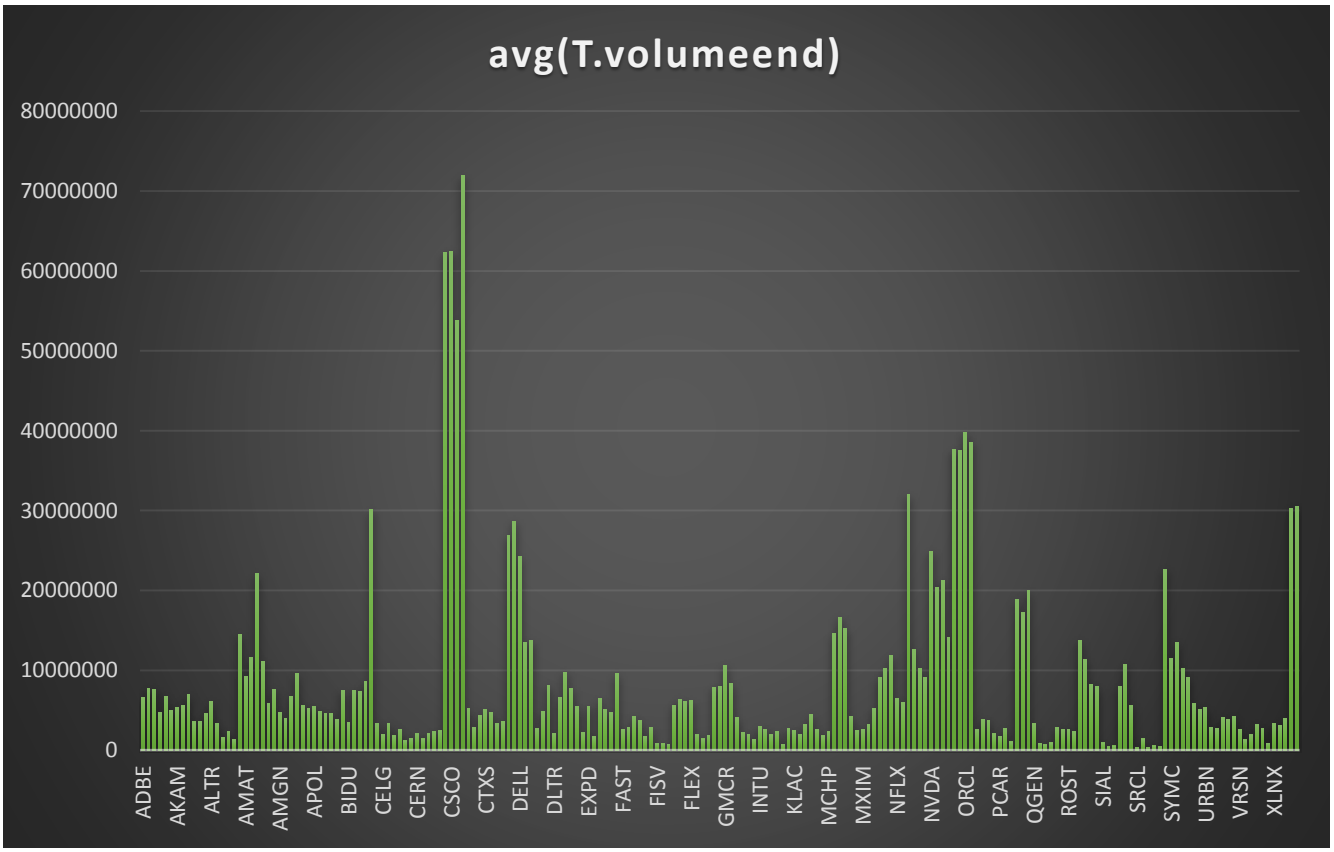
SELECT T.smblid, avg(T.volumeend) FROM Tweets T, EarnRelMatched E1 where T.datestart = E1.earnrelease_date and T.smblid = E1.ticker and T.timestart between E1.earnrelease_time and date_add(E1.earnrelease_time, interval 24 hour) group by T.smblid, T.datestart;

Explanation: This query will return the list of firms and their average trading volume within 24 hours of their earnings release. The query is written by matching tweets start date and earnings release date and tweets symbol and symbol id in earning release matched table. Starting time of tweets should be between an interval of earnings release time and 24 hours added to it.

Output:

The file Step_3c contains the result.

Graphical representation:



From the graph, firm CSCO has highest trading volume level within 24 hours of its earnings release time.

d) Average trading volume levels for the firm in the 1 week following earnings release.

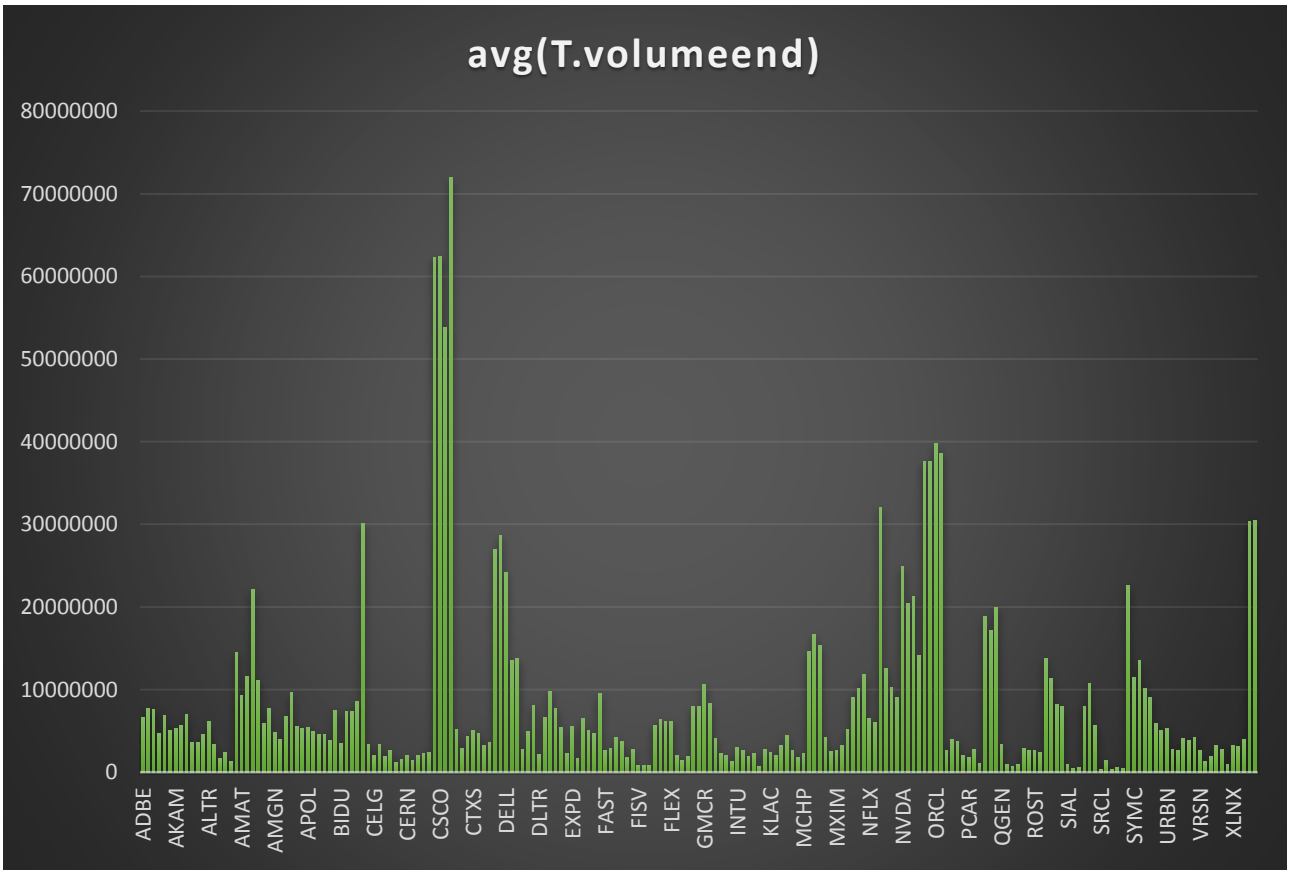
SELECT T.smblid, avg(T.volumeend) FROM Tweets T, EarnRelMatched E1 where T.datestart = E1.earnrelease_date and T.smblid = E1.ticker and T.timestart between E1.earnrelease_time and date_add(E1.earnrelease_time, interval 1 week) group by T.smblid, T.datestart;

Explanation: This query will return the list of firms and their average trading volume within 1 week of their earnings release. The query is written by matching tweets start date and earnings release date and tweets symbol and symbol id in earning release matched table. Starting time of tweets should be between an interval of earnings release time and 1 week added to it.

Output:

The file Step_3d.csv contains the result.

Graphical representation:



From the graph, firm CSCO has highest trading volume within 1 week of its earnings release time.

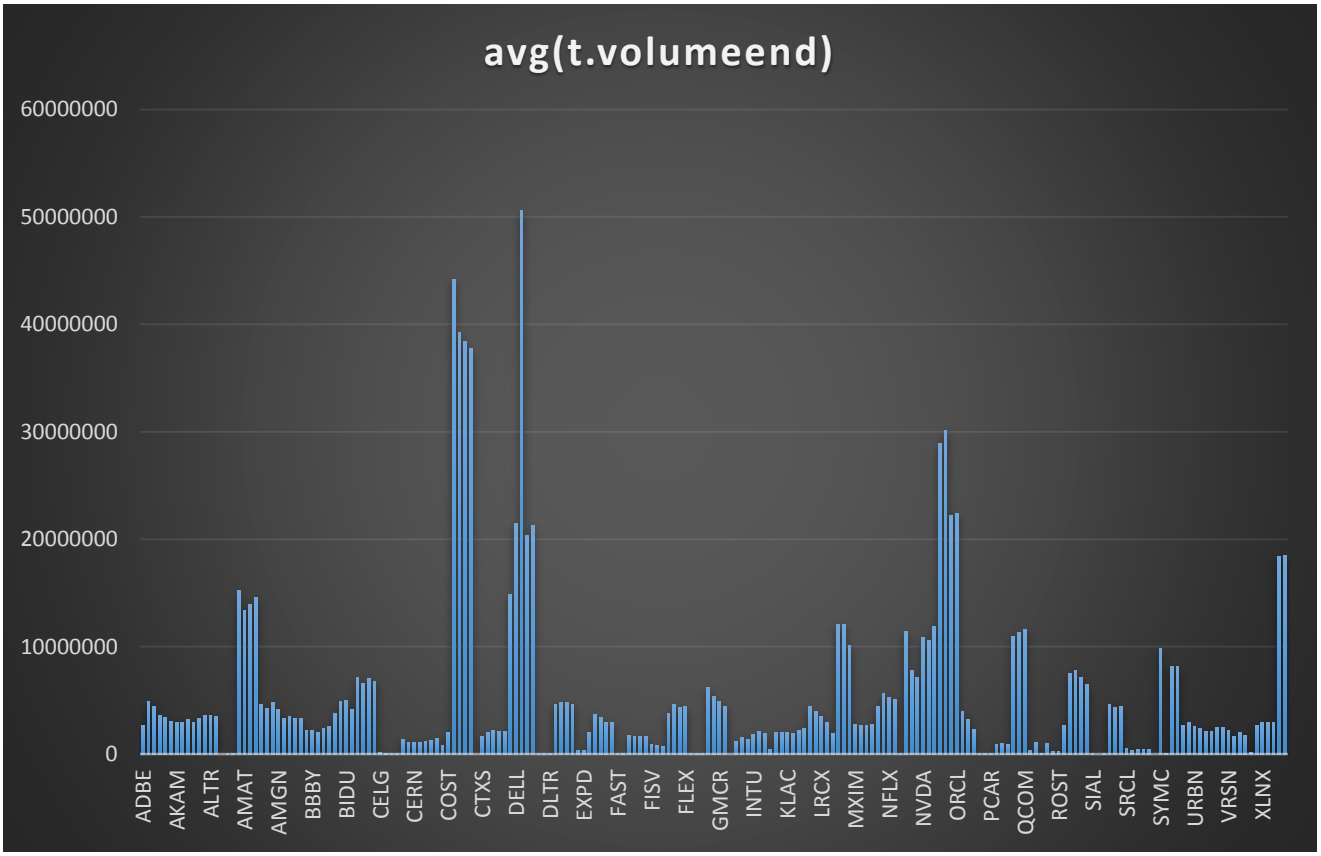
e & i) Calculate baseline average.

select t.symbol, avg(t.volumeend) from tweets t, earnrelmatched e where e.ticker=t.smblid and t.timestart between e.earnrelease_time and date_add(e.earnrelease_time, interval 30 minute) and dayofweek(t.datestart)=dayofweek(e.earnrelease_date) and t.datestart between 2012/05/10 and e.earnrelease_date-7 group by t.smblid, e.earnrelease_date;

Explanation: earnrelease dates are compared with before weeks and twitter per minute are averaged.

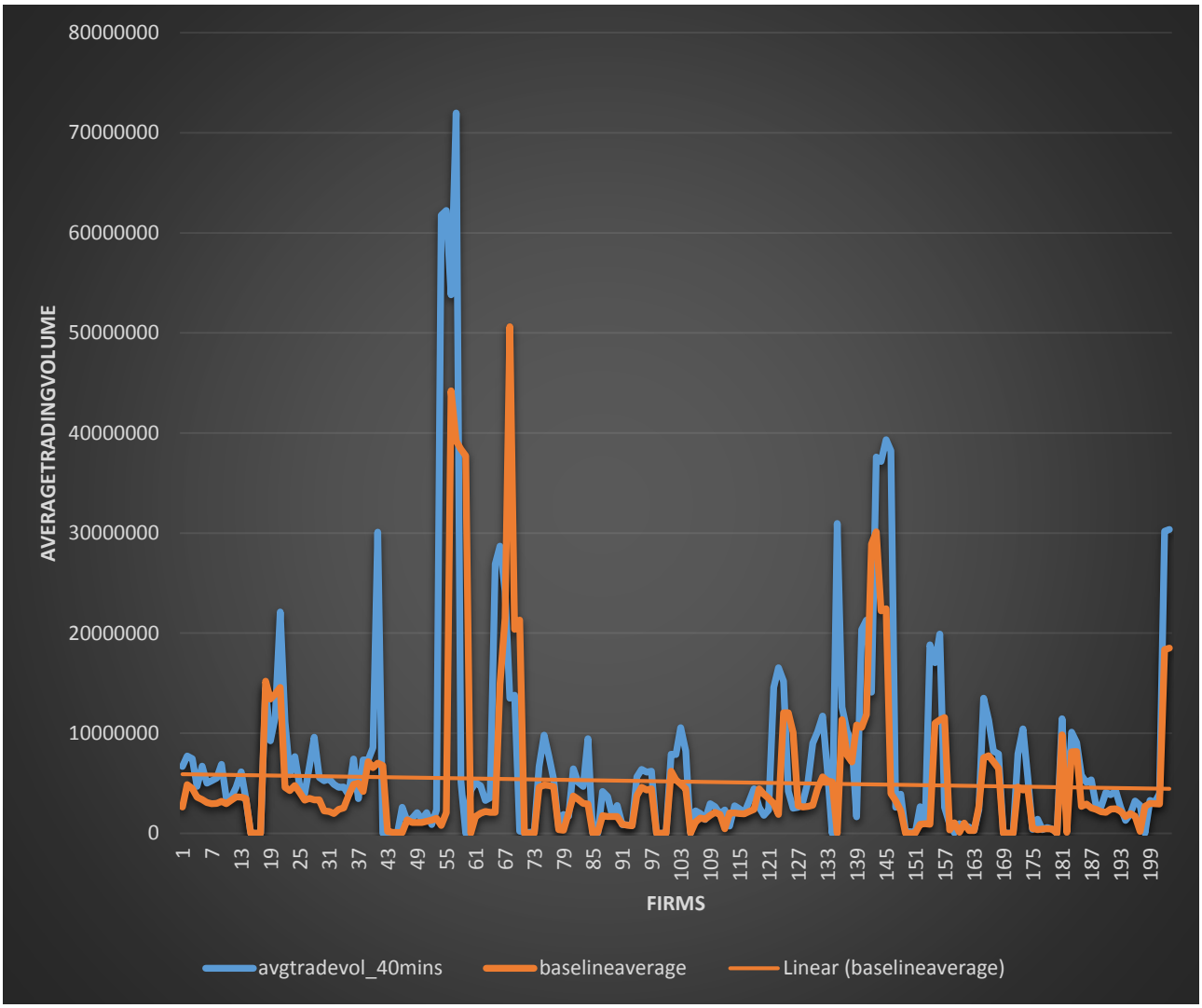
Output:

The file Step_3e.csv contains the result.



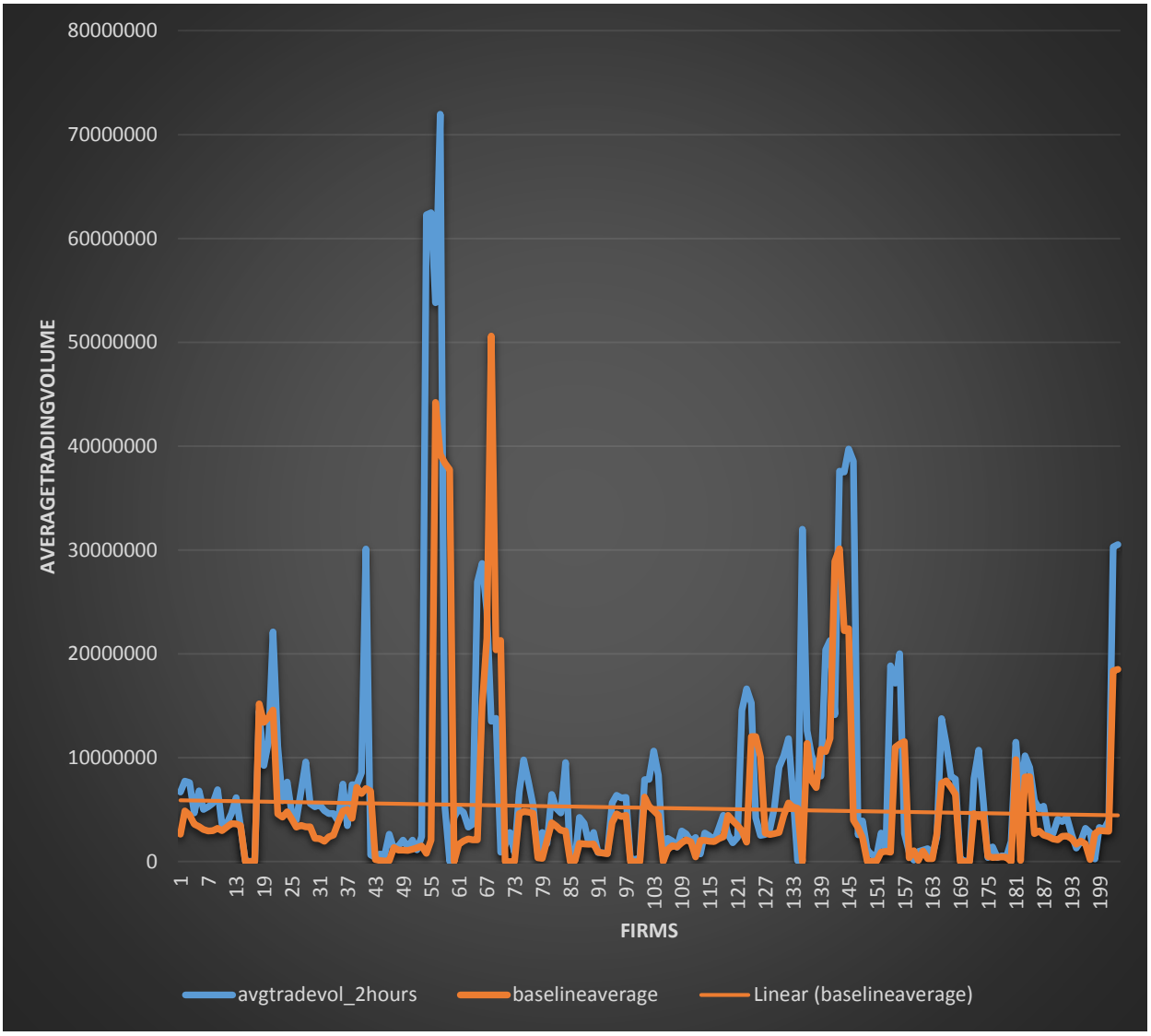
f) Compare the average trading volume levels in (a) through (d) with the baseline average described in (e).

Comparing average trading volume level within 40 minutes and baseline average



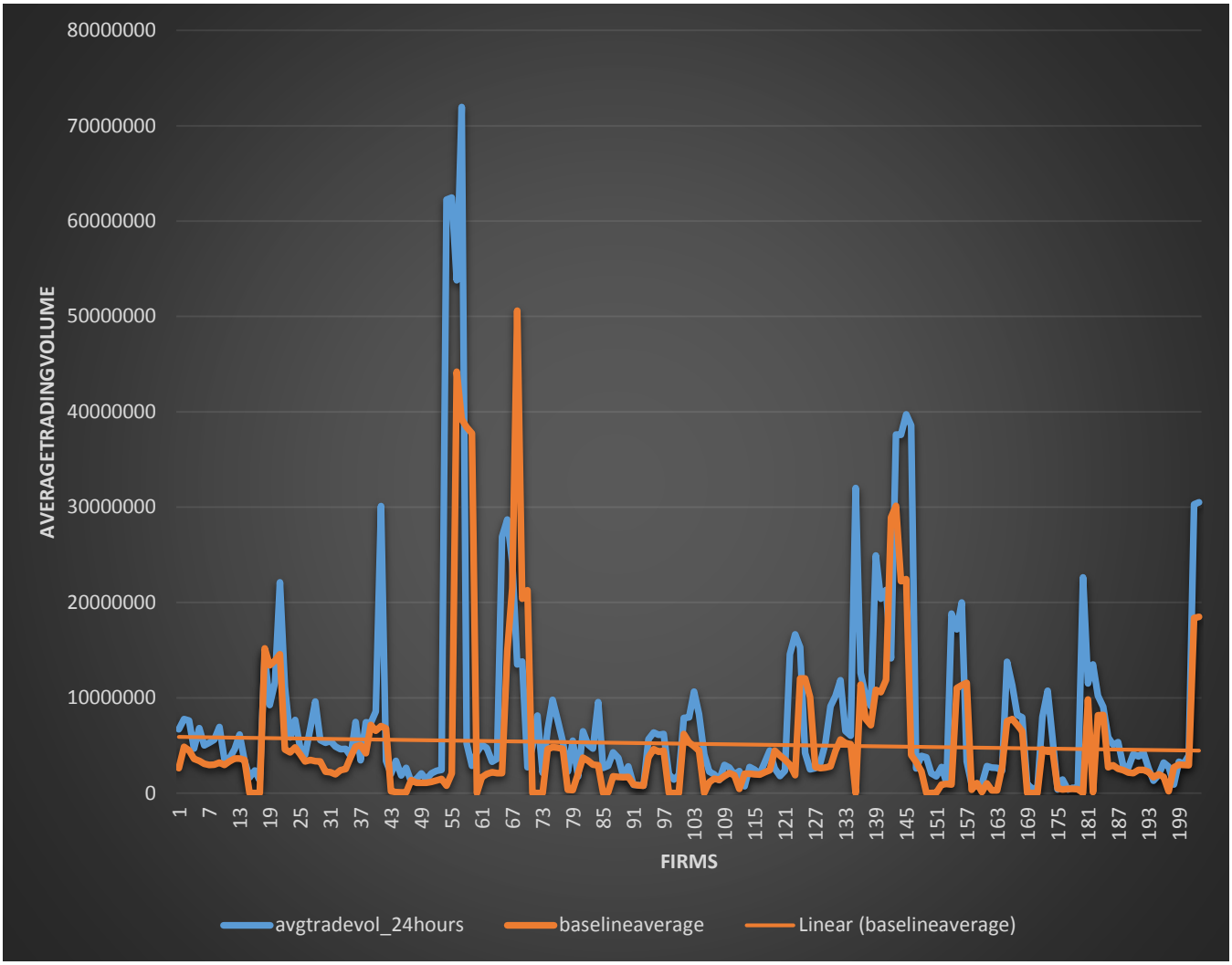
Blue line indicates average trading volume level within 40 minutes and orange line indicates baseline average. For all firms, average trading volume level within 40 minutes is greater than baseline average.

Comparing average trading volume level within 2 hours and baseline average



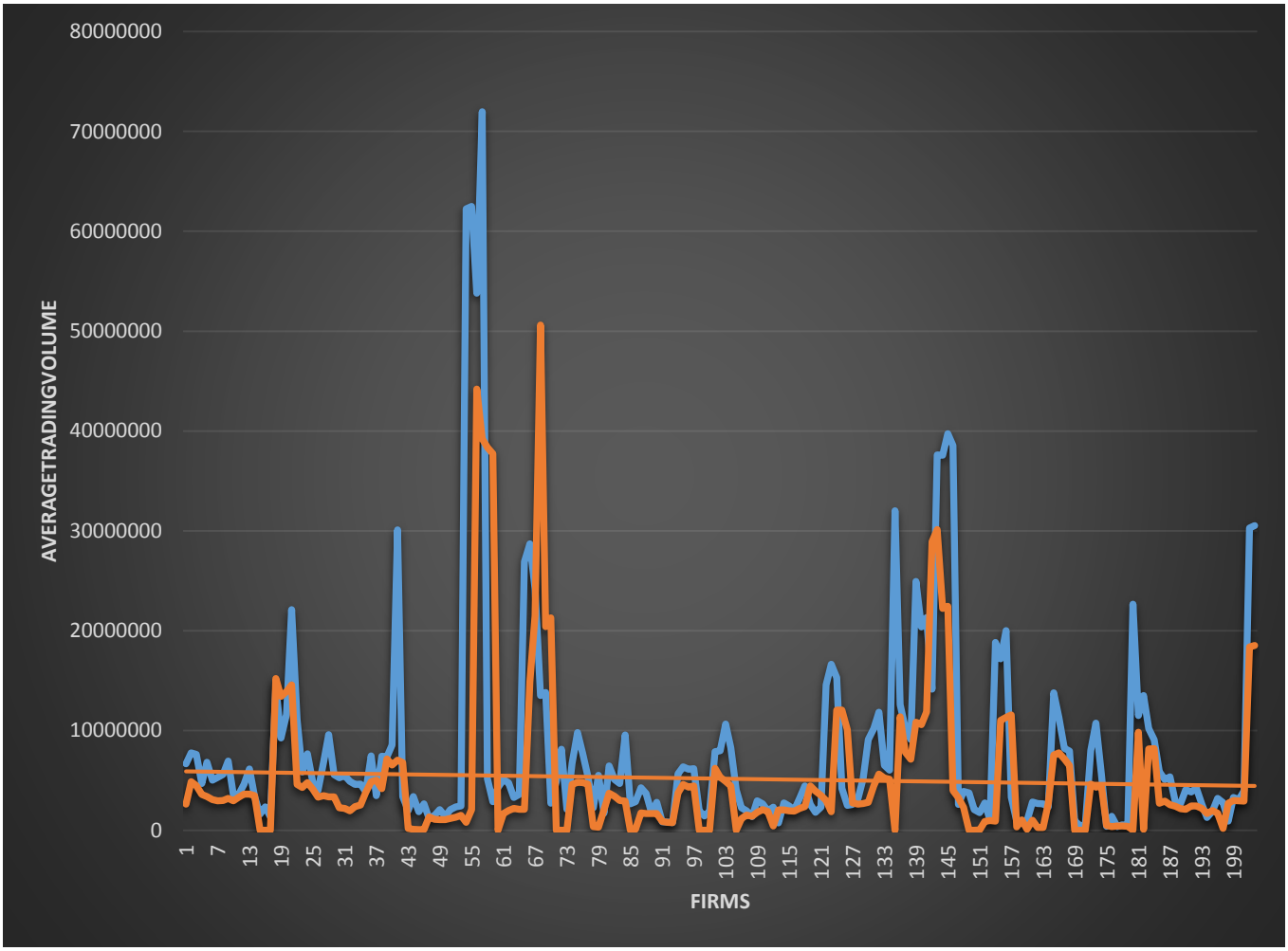
Blue line indicates average trading volume level within 2 hours and orange line indicates baseline average. For all firms, average trading volume level within 2 hours is greater than baseline average.

Comparing average trading volume level within 24 hours and baseline average



Blue line indicates average trading volume level within 24 hours and orange line indicates baseline average. For all firms, average trading volume level within 24 hours is greater than baseline average.

Comparing average trading volume level within 1 week and baseline average



Blue line indicates average trading volume level within 1 week and orange line indicates baseline average. For all firms, average trading volume level within 1 week is greater than baseline average.