

11/11/2018

New York

**STOCK EXCHANGE**



Pankti Parikh  
306613767

**A. Dataset URL :**

Main dataset URL

<https://www.kaggle.com/dgawlik/nyse>

Related Website Link

<https://www.kaggle.com/dgawlik/nyse#securities.csv>

**B. Data Insights:**

- In present situation Where Stock market is going up and down, it is necessary to invest consciously in the market whatever it is, the findings and trends of the New-York Stock markets and shares which enables the investors in taking decision regarding investments.
- A stock exchange is the platform where financial instruments like stocks and derivatives are traded. Market participants must be registered with the stock exchange and NYSE to conduct trades. This includes companies issuing shares, brokers conducting the trades, as well as traders and investors. All of this is regulated by the New-York Stock Exchange, which makes the rules of conduct.
- The New York Stock Exchange is still by far the most important equity market in the world. With a market cap of about \$21 trillion, the NYSE about three times larger than the NASDAQ, and the two US exchanges together have a larger market cap than the next ten exchanges combined.
- The New York Stock Exchange provides a means for buyers and sellers to trade shares of stock in companies registered for public trading.

- NYSE stocks can be traded via its electronic hybrid market (except for a small group of very high-priced stocks). Customers can now send orders for immediate electronic execution, or route orders to the floor for trade in the auction market.
- NYSE works with US regulators like the SEC and CFTC to coordinate risk management measures in the electronic trading environment through the implementation of mechanisms like circuit breakers and liquidity replenishment points.

C. **Data description:**

- New-York Stock exchange dataset from Kaggle includes four different CSV files as following:
  1. **Fundamentals.csv:** most of popular fundamental indicators are derive from annual SEC 10 K filling (2012-2016), Prices are fetched from Yahoo Finance. Whereas fundamentals are fetched from Nasdaq Financials. There are 78 columns in this CSV file describes different properties of company's share and earnings.
  2. **Prices-Split-Adjusted.csv:** There have been added adjustments for splits which was same as price.
  3. **Prices.csv:** raw, as-is daily prices. Most of data spans from 2010 to the end 2016, for Companies new on stock market date range are shorter. There have been approx. 140 stock Splits in that time, this set doesn't account for that.
  4. **Securities.csv:** Each company with division on sectors in general description. Includes

#### D. **Data Cleaning:**

Here are several key benefits that come out of the data cleaning process:

- It removes major errors and inconsistencies that are inevitable when multiple sources of data are getting pulled into one dataset.
- Using tools to cleanup data will make everyone more efficient since they'll be able to quickly get what they need from the data.
- Fewer errors means happier customers and fewer frustrated employees.
- The ability to map the different functions and what your data is intended to do and where it is coming from your data.

#### **Categories:**

##### **1. Missing values:**

- The goal of such cleaning operations is to prevent problems caused by missing data that can arise when creating a Visuals. This module supports multiple type of operations for "cleaning" missing values, including:
  - Replacing missing values with a placeholder, mean, or other value.
  - Completely removing rows and columns that have missing values.
  - Inferring values based on statistical methods
- As, you can see in following figure. That, I used 2<sup>nd</sup> type of cleaning by completely removing rows and columns that have missing values .

## Before

Power Query Editor Screenshot (Before):

**Queries [2]**

	Total Revenue	Treasury Stock	For Year	Earnings Per Share	Estimated Shares Outstanding
59	64406000000	-19218000000	2014	5.27	1428652751
60	58327000000	-30098000000	2015	1.69	1299408284
61	958511000	0	2012	0.61	217140983.6
62	974053000	0	2013	1.4	148064285.7
63	984363000	0	2014	2.06	150120873.8
64	981310000	0	2015	1.52	163625000
65	9047657000	-3098241000	2013	6.38	76631191.22
66	10381653000	-3316511000	2014	6.52	72225000
67	10325494000	-3601162000	2015	2.08	68055288.46
68	7531780000	-4470551000	null	null	null
69	3179600000	0	2013	2.08	129134615.4
70	4626500000	0	2014	1.98	15323232.3
71	5392400000	0	2015	2.07	172367149.8
72	5594800000	0	null	null	null
73	1373947000	-624462000	2012	1.15	177381738.1
74	1577922000	0	2013	1.65	177870909.1
75	1963874000	0	2014	1.87	178581818.2
76	2197448000	0	2015	1.8	178558888.9
77	2519154000	0	2012	3.49	89265329.51
78	2394270000	0	2013	4.93	83807505.07
79	2445548000	0	2014	1.69	78885207.1
80	3651335000	0	2015	3.01	11126451.8
81					

PREVIEW DOWNLOADED ON FRIDAY  
12:36 AM 11/11/2018

## After

Power Query Editor Screenshot (After):

**Queries [2]**

	Total Revenue	Treasury Stock	For Year	Earnings Per Share	Estimated Shares Outstanding
1	24855030000	-367000000	2013	0.1	3310000
2	26743000000	0	2014	-11.25	16300222.2
3	42650000000	0	2014	4.02	71691542
4	40990000000	0	2015	11.39	66812993
5	62050030000	-27095000	2012	5.29	73283553
6	6493814000	-107800000	2013	5.36	7308917
7	9843861000	-113044000	2014	6.75	73159259
8	9737018000	-119709000	2015	6.45	73395038
9	1.7091E+11	0	2013	40.03	925231076
10	1.82795E+11	0	2014	6.49	60878274
11	2.33715E+11	0	2015	9.28	57536637
12	2.15639E+11	0	2016	8.35	54714976
13	187900000000	-320000000	2013	2.58	160000000
14	199600000000	-9720000000	2014	1.11	15981981
15	228590000000	-8839000000	2015	3.15	16330158
16	87959167000	-1516856000	2013	1.88	23060921
17	1.19569E+11	-2313380000	2014	1.22	22472868
18	1.35962E+11	-4150997000	2015	-0.62	22284677
19	1.46854E+11	-4396008000	2016	6.73	221217369
20	190500000000	-5591000000	2012	3.76	15859042
21	196570000000	-6844000000	2013	1.64	15707317
22	202470000000	-8678000000	2014	1.5	15226666
23					

PREVIEW DOWNLOADED AT 12:37 AM 11/11/2018

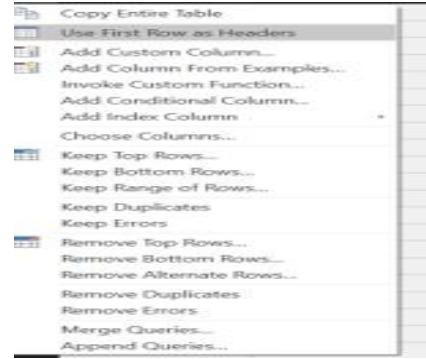
## 2. Erroneous values :

- removing typographical errors or validating and correcting values against a known list of entities. The validation may be strict or fuzzy (such as correcting records that partially match existing, known records). Some data cleansing solutions will clean data by cross checking with a validated data set.
- You can see In the following screenshot error is removed by selecting remove error option from the table properties option.

The screenshot shows the Power Query Editor interface. On the left, the 'Queries [2]' pane lists two queries: 'fundamentals' and 'securities'. The main area displays a table with columns: Column1, Ticker Symbol, Period Ending, Accounts Payable, Accounts Receivable, and Add'l Income. The 'fundamentals' query contains 23 rows of data. On the right, the 'QUERY SETTINGS' pane is open, showing the 'PROPERTIES' section with 'Name' set to 'fundamentals' and the 'APPLIED STEPS' section. The 'APPLIED STEPS' section lists several steps: Source, Changed Type, Promoted Headers, Changed Type1, Removed Duplicates, Filtered Rows1, Removed Columns1, and **Removed Errors**. The 'Removed Errors' step is circled in red.

## 3. Inconsistencies:

- Sometimes data contains some inconsistent values like here, when I inserted data from the csv file it shows that column headers are the elements considered as the first row of the tables. So, I chose the table properties option and selected option called choose first row as the header.



Before

Untitled - Power Query Editor

**Queries [1]**

Column1	Column2	Column3	Column4	Column5	Column6
null	Ticker Symbol	Period Ending	Accounts Payable	Accounts Receivable	Add'l Income/expense items
1	AAL	12/31/2012	3068000000	-222000000	1951000000
2	AAL	12/31/2013	4973000000	-160000000	-2723000000
3	AAL	12/31/2014	4668000000	-160000000	-1500000000
4	AAL	12/31/2015	5102000000	352000000	-708000000
5	AAP	12/29/2012	2409453000	-89482000	6000000
6	AAP	12/28/2013	2609239000	-32428000	26980000
7	AAP	1/3/2015	3616038000	-48209000	30920000
8	AAP	1/2/2016	3757085000	-21476000	-74840000
9	AAPL	9/28/2013	36223000000	-1949000000	11560000000
10	AAPL	9/27/2014	48649000000	-6452000000	9800000000
11	AAPL	9/26/2015	60671000000	-3124000000	12850000000
12	ABBY	12/31/2012	5734000000	223000000	-8000000
13	ABBY	12/31/2013	6448000000	681000000	-54000000
14	ABBY	12/31/2014	6954000000	-172000000	-65100000
15	ABBY	12/31/2015	8463000000	-1076000000	-206000000
16	ABC	9/30/2013	14870635000	-2312518000	-44000
17	ABC	9/30/2014	17250160000	-938286000	-285940000
18	ABC	9/30/2015	21578227000	-1478793000	-44220000
19	ABC	9/30/2016	24670159000	-912724000	50480000
20	ABT	12/31/2012	10889000000	36000000	-126000000
21	ABT	12/31/2013	5948000000	-113000000	
22	ABT	12/31/2014			
23					

79 COLUMNS, 999+ ROWS

PREVIEW DOWNLOADED AT 5:54 PM  
11/5/2018

After

Untitled - Power Query Editor

**Queries [1]**

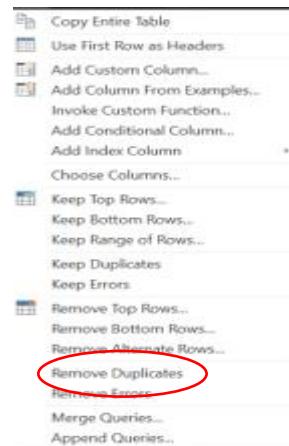
Column1	Ticker Symbol	Period Ending	Accounts Payable	Accounts Receivable	Add'l Income
1		12/31/2012	3068000000	-222000000	1951000000
2	AAL	12/31/2013	4973000000	-160000000	-2723000000
3	AAL	12/31/2014	4668000000	-160000000	-1500000000
4	AAL	12/31/2015	5102000000	352000000	-708000000
5	AAP	12/29/2012	2409453000	-89482000	6000000
6	AAP	12/28/2013	2609239000	-32428000	26980000
7	AAP	1/3/2015	3616038000	-48209000	30920000
8	AAPL	9/28/2013	36223000000	-1949000000	11560000000
9	AAPL	9/27/2014	48649000000	-6452000000	9800000000
10	AAPL	9/26/2015	60671000000	-3124000000	12850000000
11	ABBY	12/31/2012	5734000000	223000000	-8000000
12	ABBY	12/31/2013	6448000000	681000000	-54000000
13	ABBY	12/31/2014	6954000000	-172000000	-65100000
14	ABBY	12/31/2015	8463000000	-1076000000	-206000000
15	ABC	9/30/2013	14870635000	-2312518000	-44000
16	ABC	9/30/2014	17250160000	-938286000	-285940000
17	ABC	9/30/2015	21578227000	-1478793000	-44220000
18	ABC	9/30/2016	24670159000	-912724000	50480000
19	ABT	12/31/2012	10889000000	36000000	-126000000
20	ABT	12/31/2013	5948000000	-113000000	
21	ABT	12/31/2014			
22					
23					

79 COLUMNS, 999+ ROWS

PREVIEW DOWNLOADED AT 5:54 PM  
11/5/2018

#### 4. Duplicate records:

- Duplicate rows are a common problem when you import data. It is a good idea to filter for unique values first to confirm that the results are what you want before you remove duplicate values. By choosing following step you can remove duplicate data from the table.



Screenshot of the Power Query Editor interface showing a table of financial data and the 'QUERY SETTINGS' pane.

**Table Data Preview:**

	Column1	Ticker Symbol	Period Ending	1.2 Accounts Payable	1.2 Accounts Receivable	1.2 Add'l Income
1	0	AAL	12/31/2012	3068000000	-222000000	
2	1	AAL	12/31/2013	4975000000	-93000000	
3	2	AAL	12/31/2014	4668000000	-160000000	
4	3	AAL	12/31/2015	5102000000	352000000	
5	4	AAP	12/29/2012	2409453000	-89482000	
6	5	AAP	12/28/2013	2609239000	-32428000	
7	6	AAP	1/3/2015	3616038000	-48209000	
8	7	AAP	1/2/2016	3757085000	-21476000	
9	8	AAPL	9/28/2013	36223000000	-1949000000	
10	9	AAPL	9/27/2014	48649000000	-6452000000	
11	10	AAPL	9/26/2015	60671000000	-3124000000	
12	11	AAPL	9/24/2016	59321000000	10440000000	
13	12	ABBV	12/31/2012	5734000000	223000000	
14	13	ABBV	12/31/2013	6448000000	681000000	
15	14	ABBV	12/31/2014	6954000000	-172000000	
16	15	ABBV	12/31/2015	8463000000	-1076000000	
17	16	ABC	9/30/2013	14870635000	-2312518000	
18	17	ABC	9/30/2014	17250160000	-938286000	
19	18	ABC	9/30/2015	2157827000	-1478793000	
20	19	ABC	9/30/2016	24670159000	-912724000	
21	20	ABT	12/31/2012	10889000000	36000000	
22	21	ABT	12/31/2013	5948000000	-113000000	

**QUERY SETTINGS - APPLIED STEPS:**

- Source
- Changed Type
- Promoted Headers
- Changed Type1
- Removed Duplicates** (highlighted with a red circle)
- Filtered Rows1
- Removed Columns1
- Removed Errors
- Removed Columns
- Filtered Rows

## 5. Out of date :

- Means Data having old format or Data Type. No one wants to use the data which is out of the date or old format. But luckily in my dataset not a single column have this kind of problem.

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table with 23 rows and 76 columns. The columns include 'Companies', 'Period Ending', '1.2 Accounts Payable', '1.2 Accounts Receivable', and '1.2 After Tax ROE'. The 'APPLIED STEPS' pane on the right lists 22 steps, such as 'Source', 'Changed Type', 'Promoted Headers', and 'Removed Duplicates'. The status bar at the bottom indicates 'PREVIEW DOWNLOADED AT 1:58 AM' and '2:25 AM 11/12/2018'.

	Companies	Period Ending	1.2 Accounts Payable	1.2 Accounts Receivable	1.2 After Tax ROE
1	0 AAL	12/31/2012	3068000000	-222000000	
2	1 AAL	12/31/2013	4975000000	-93000000	
3	2 AAL	12/31/2014	4668000000	-160000000	
4	3 AAL	12/31/2015	5102000000	352000000	
5	4 AAP	12/29/2012	2409453000	-89482000	
6	5 AAP	12/28/2013	2609239000	-32428000	
7	6 AAP	1/3/2015	3616038000	-48209000	
8	7 AAP	1/2/2016	3757085000	-21476000	
9	8 AAPL	9/28/2013	36223000000	-1949000000	
10	9 AAPL	9/27/2014	48649000000	-6452000000	
11	10 AAPL	9/26/2015	60671000000	-3124000000	
12	11 AAPL	9/24/2016	59321000000	1044000000	
13	13 ABBV	12/31/2013	6448000000	681000000	
14	14 ABBV	12/31/2014	6954000000	-172000000	
15	15 ABBV	12/31/2015	8463000000	-1076000000	
16	16 ABC	9/30/2013	14870635000	-2312518000	
17	17 ABC	9/30/2014	17250160000	-938286000	
18	18 ABC	9/30/2015	21578227000	-1478793000	
19	19 ABC	9/30/2016	24670159000	-912724000	
20	20 ABT	12/31/2012	10889000000	36000000	
21	21 ABT	12/31/2013	5948000000	-113000000	
22	22 ABT	12/31/2014	5350000000	-195000000	
23					

## 6. Leading or trailing issues :

- Having something before or after the original value of the data called leading or trailing problem in my dataset one column called **CIK** have certain numeric values and that values have zeroes before the actual value. And this data set have this column as a text by default. But by changing datatype from text to whole numbers of that column I removed this problem.

## Before

The screenshot shows the Power Query Editor interface with two queries loaded. The left pane lists 'fundamentals' and 'securities'. The right pane displays a table with columns: ICS Sector, GICS Sub Industry, Address of Headquarters, Date first added, and CIK. The 'Date first added' column is highlighted with a red circle. The 'APPLIED STEPS' pane on the right shows steps like 'Promoted Headers', 'Filtered Rows1', 'Changed Type1', and 'Removed Duplicates'.

ICS Sector	GICS Sub Industry	Address of Headquarters	Date first added	CIK
1 Care	Health Care Equipment	North Chicago, Illinois	3/31/1964	0000001800
2 Care	Pharmaceuticals	North Chicago, Illinois	12/31/2012	155115152
3 Information Technology	IT Consulting & Other Services	Dublin, Ireland	7/6/2011	1467373
4 Information Technology	Home Entertainment Software	Santa Monica, California	8/31/2015	1000718877
5 Trials	Electrical Components & Equipment	Atlanta, Georgia	5/3/2016	000144215
6 Information Technology	Application Software	San Jose, California	5/5/1997	0000796343
7 Consumer Discretionary	Automotive Retail	Roanoke, Virginia	7/9/2015	0001158449
8 Care	Managed Health Care	Hartford, Connecticut	6/30/1976	000122304
9 Trials	Asset Management & Custody Banks	Beverly, Massachusetts	7/1/2014	0001004434
10 Trials	Industrial Gases	Allentown, Pennsylvania	4/30/1985	0000002969
11 Information Technology	Internet Software & Services	Cambridge, Massachusetts	7/19/2007	0001086222
12 Trials	Airlines	Seattle, Washington	5/13/2016	0000766421
13 Trials	Specialty Chemicals	Baton Rouge, Louisiana	7/1/2016	0009159113
14 Care	Biotechnology	Cheshire, Connecticut	5/25/2012	0000899866
15 Trials	Building Products	Dublin, Ireland	12/31/2013	0001579241
16 Information Technology	Data Processing & Outsourced Services	Plano, Texas	12/23/2013	0001101215
17 Trials	Electric Utilities	Madison, Wisconsin	7/1/2016	0000352541
18 Information Technology	Internet Software & Services	Mountain View, California	4/3/2014	0001652044
19 Consumer Discretionary	Internet & Direct Marketing Retail	Seattle, Washington	11/18/2005	0001018724
20 Trials	MultUtilities	St. Louis, Missouri	9/19/1991	0001002910
21 Trials	Airlines	Fort Worth, Texas	3/23/2015	6201
22 Trials	Consumer Finance	New York, New York	6/30/1976	4962

## After

The screenshot shows the Power Query Editor interface with the same two queries loaded. The table structure is identical to the 'Before' state. The 'APPLIED STEPS' pane on the right shows steps like 'Promoted Headers', 'Filtered Rows1', 'Changed Type1', 'Removed Duplicates', 'Removed Errors', and 'Changed Type'.

ICS Sector	GICS Sub Industry	Address of Headquarters	Date first added	CIK
1 Care	Health Care Equipment	North Chicago, Illinois	3/31/1964	1800
2 Care	Pharmaceuticals	North Chicago, Illinois	12/31/2012	155115152
3 Information Technology	IT Consulting & Other Services	Dublin, Ireland	7/6/2011	1467373
4 Information Technology	Home Entertainment Software	Santa Monica, California	8/31/2015	1000718877
5 Trials	Electrical Components & Equipment	Atlanta, Georgia	5/3/2016	1344225
6 Information Technology	Application Software	San Jose, California	5/5/1997	796343
7 Consumer Discretionary	Automotive Retail	Roanoke, Virginia	7/9/2015	1158449
8 Care	Managed Health Care	Hartford, Connecticut	6/30/1976	1122304
9 Trials	Asset Management & Custody Banks	Beverly, Massachusetts	4/30/1985	2969
10 Information Technology	Internet Software & Services	Cambridge, Massachusetts	7/12/2005	1086222
11 Trials	Airlines	Seattle, Washington	5/13/2016	766421
12 Trials	Specialty Chemicals	Baton Rouge, Louisiana	7/1/2016	9159113
13 Care	Biotechnology	Cheshire, Connecticut	5/25/2012	899866
14 Trials	Building Products	Dublin, Ireland	12/23/2013	1579241
15 Information Technology	Data Processing & Outsourced Services	Plano, Texas	12/23/2013	1101215
16 Trials	Electric Utilities	Madison, Wisconsin	7/5/2016	352541
17 Information Technology	Internet Software & Services	Mountain View, California	4/3/2014	1652044
18 Consumer Discretionary	Internet & Direct Marketing Retail	Seattle, Washington	11/18/2005	1018724
19 Trials	MultUtilities	St. Louis, Missouri	9/19/1991	1002910
20 Trials	Airlines	Fort Worth, Texas	3/23/2015	6201
21 Trials	Consumer Finance	New York, New York	6/30/1976	4962

## 7. Date format:

- Dates are always in mm-dd-yyyy or date or date/time format by sometimes raw data contain other data types as a by default like in this dataset one column named **Date first added** have text date type by default so by correcting that I changed that data type to mm/dd/yyyy format.you can see before and after snapshot as a following.

Before

The screenshot shows the Power Query Editor interface with two queries: 'fundamentals' and 'securities'. The 'securities' query is selected. In the 'Date first added' column, the data type is currently set to 'Text'. The 'APPLIED STEPS' pane on the right shows the history of changes made to the query, including 'Filtered Rows1' and 'Changed Type'.

ICS Sector	GICS Sub Industry	Address of Headquarters	Date first added	CIK
1 Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	0000001800
2 Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152
3 nation Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	0001467373
4 nation Technology	Home Entertainment Software	Santa Monica, California	2015-08-31	0000718877
5 s	Electrical Components & Equipment	Atlanta, Georgia	2016-05-03	0001144215
6 nation Technology	Application Software	San Jose, California	1997-05-05	0000796343
7 Discretionary	Automotive Retail	Roanoke, Virginia	2015-07-09	0001158449
8 re	Managed Health Care	Hartford, Connecticut	1976-06-30	0001122304
9 s	Asset Management & Custody Banks	Beverly, Massachusetts	2014-07-01	0001004434
10 s	Industrial Gases	Allentown, Pennsylvania	1985-04-30	0000002969
11 on Technology	Internet Software & Services	Cambridge, Massachusetts	2007-07-12	0001086222
12 s	Airlines	Seattle, Washington	2016-05-13	0000766421
13 s	Specialty Chemicals	Baton Rouge, Louisiana	2016-07-01	0000915913
14 re	Biotechnology	Cheshire, Connecticut	2012-05-25	0000899866
15 s	Building Products	Dublin, Ireland	2013-12-02	0001579241
16 on Technology	Data Processing & Outsourced Services	Plano, Texas	2013-12-23	0001101215
17 s	Electric Utilities	Madison, Wisconsin	2016-07-01	0000352541
18 on Technology	Internet Software & Services	Mountain View, California	2014-04-03	0001652044
19 Discretionary	Internet & Direct Marketing Retail	Seattle, Washington	2005-11-18	0001018724
20 s	Multilevel Utilities	St. Louis, Missouri	1991-09-19	0001002910
21 s	Airlines	Fort Worth, Texas	2015-03-23	0000006201
22 s	Consumer Finance	New York, New York	1976-06-30	0000004962

After

The screenshot shows the Power Query Editor interface with the same 'securities' query selected. The 'Date first added' column has been converted to a date type. The 'APPLIED STEPS' pane shows the step 'Changed Type1'.

ICS Sector	GICS Sub Industry	Address of Headquarters	Date first added	CIK
1 Care	Health Care Equipment	North Chicago, Illinois	3/31/1964	0000001800
2 Care	Pharmaceuticals	North Chicago, Illinois	12/31/2012	0001551152
3 nation Technology	IT Consulting & Other Services	Dublin, Ireland	7/6/2011	0001467373
4 nation Technology	Home Entertainment Software	Santa Monica, California	8/31/2015	0000718877
5 trials	Electrical Components & Equipment	Atlanta, Georgia	5/3/2016	0001144215
6 nation Technology	Application Software	San Jose, California	5/5/1997	0000796343
7 inner Discretionary	Automotive Retail	Roanoke, Virginia	7/9/2015	0001158449
8 Care	Managed Health Care	Hartford, Connecticut	6/30/1976	0001122304
9 trials	Asset Management & Custody Banks	Beverly, Massachusetts	7/1/2014	0001004434
10 trials	Industrial Gases	Allentown, Pennsylvania	7/22/2007	0000002969
11 nation Technology	Internet Software & Services	Seattle, Washington	5/23/2016	0000766421
12 s	Airlines	Baton Rouge, Louisiana	7/1/2016	0000915913
13 trials	Specialty Chemicals	Cheshire, Connecticut	5/25/2012	0000899866
14 Care	Biotechnology	Dublin, Ireland	12/3/2013	0001579241
15 trials	Building Products	Plano, Texas	12/23/2013	0001101215
16 nation Technology	Data Processing & Outsourced Services	Madison, Wisconsin	7/1/2016	0000352541
17 s	Electric Utilities	Mountain View, California	4/3/2014	0001652044
18 nation Technology	Internet Software & Services	Seattle, Washington	11/18/2005	0001018724
19 inner Discretionary	Internet & Direct Marketing Retail	St. Louis, Missouri	9/19/1991	0001002910
20 trials	Multilevel Utilities	Fort Worth, Texas	3/23/2015	0000006201
21 trials	Airlines	New York, New York	6/30/1976	0000004962

## 8. Other Data Type:

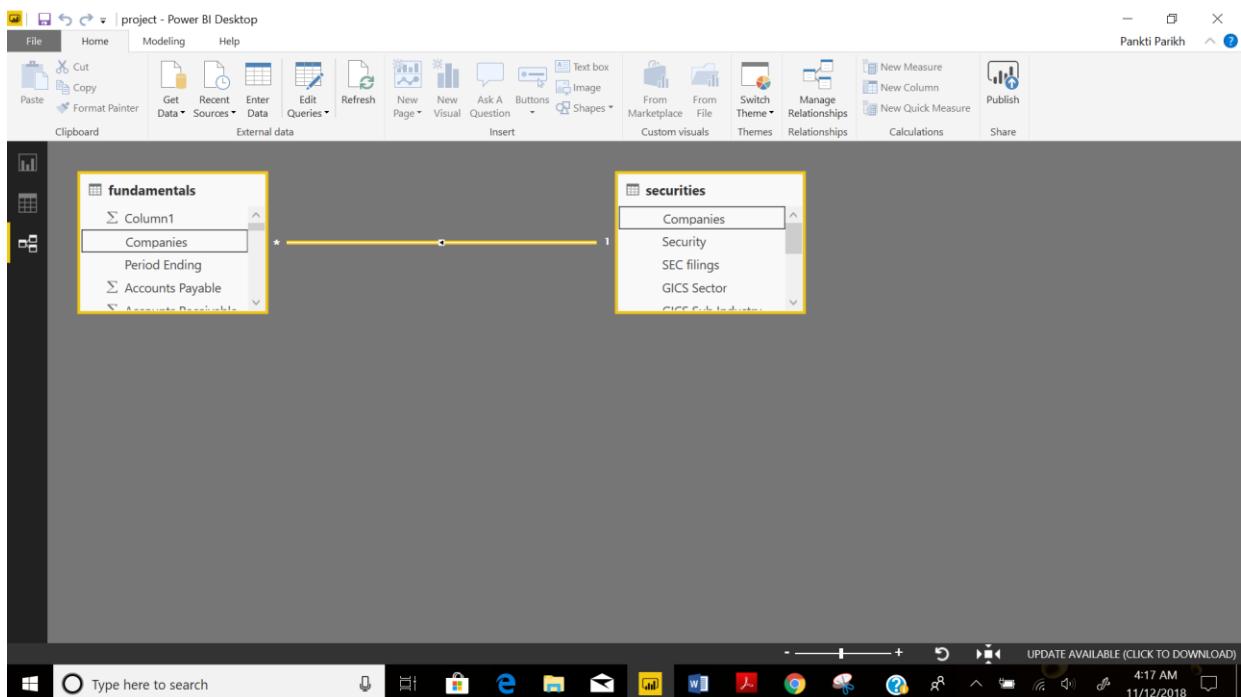
- This represents the uncommon system characters or line breaks. Any uncertain entry in the columns or rows which is not suitable for the selected datatype. This can be overcome by replacing the error values or removing the complete row or column.

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table with 23 rows of data. The columns are labeled: 'Companies', 'Period Ending', 'Accounts Payable', 'Accounts Receivable', and 'After Tax ROE'. The data includes entries for companies like AAL, AAPL, ABC, and ABT, with dates ranging from 2012 to 2014. The 'APPLIED STEPS' pane on the right lists several data cleaning and transformation steps, such as 'Promoted Headers', 'Changed Type1', 'Removed Duplicates', and 'Renamed Columns1'. The status bar at the bottom indicates 'PREVIEW DOWNLOADED AT 11:52 PM 11/12/2018'.

	Companies	Period Ending	Accounts Payable	Accounts Receivable	After Tax ROE
1	AAL	12/31/2012	3068000000	-22000000	
2	AAL	12/31/2013	4975000000	-9300000	
3	AAL	12/31/2014	4668000000	-16000000	
4	AAL	12/31/2015	5102000000	35200000	
5	AAP	12/29/2012	2409453000	-8948200	
6	AAP	12/28/2013	2609239000	-3242800	
7	AAP	1/3/2015	3616038000	-48209000	
8	AAP	1/2/2016	3757085000	-21476000	
9	AAPL	9/28/2013	36223000000	-1949000000	
10	AAPL	9/27/2014	48649000000	-6452000000	
11	AAPL	9/26/2015	60671000000	-3124000000	
12	AAPL	9/24/2016	59321000000	1044000000	
13	ABBV	12/31/2013	6448000000	68100000	
14	ABBV	12/31/2014	6954000000	-17200000	
15	ABBV	12/31/2015	8463000000	-107600000	
16	ABC	9/30/2013	14870635000	-2312518000	
17	ABC	9/30/2014	17250160000	-938286000	
18	ABC	9/30/2015	21578227000	-478793000	
19	ABC	9/30/2016	24670159000	-912724000	
20	ABT	12/31/2012	10889000000	36000000	
21	ABT	12/31/2013	59480000000	-113000000	
22	ABT	12/31/2014	53500000000	-195000000	
23					

### E. Relationships between Tables:

- If we query two or more tables at the same time, when the data is loaded, Power BI Desktop will attempt to find and create relationships for us. Cardinality, Cross filter direction, and Active properties are automatically set.
- Power BI Desktop looks at column names in the tables we are querying to determine if there are any potential relationships. If there are, those relationships are created automatically. If Power BI Desktop cannot determine with a high-level of confidence there is a match, it will not automatically create the relationship. we can still use the Manage Relationships dialog to create or edit relationships.



- Here, you can see in the snapshot that two tables are connected through the common column called Companies.

- In this Relationship Companies column of Fundamentals became **Primary key** while Companies column of securities called **Foreign key**.

#### F. **Measure:**

Here, I created one measure called **Net\_Income**.

Following , one formula I used in visual to calculate Net Income :

$$\text{EBIT} = \text{Net\_income} + \text{Interest expense} + \text{Tax expense}$$

```
Net_Income = sum(fundamentals[Earnings Before Interest and Tax]) - sum(fundamentals[Interest Expense]) + sum(fundamentals[Income Tax]))
```

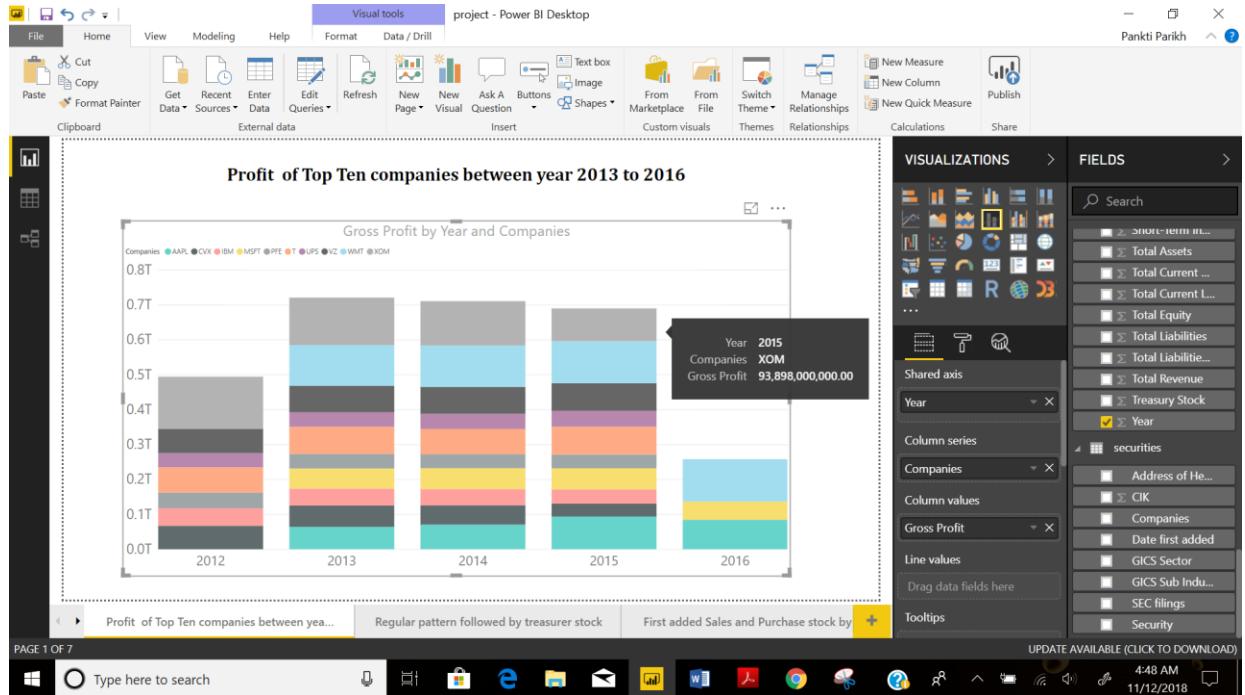
#### G. **Visuals and Dashboard :**

##### 1. Stake Column Chart:

- Stake charts are best for comparing data that is grouped by discrete categories. Stake bar charts are best when you don't have too many groups (less than 10 is usually good). Each bar is separated by blank space which indicates that there is no inherent order to your groups.

**Question 1 : How much profit made by top ten companies between year 2012 to 2016?**

- Here, we can see Profit of Top Ten companies Between year 2012 to 2016. I took **Gross Profit** in Y-axis, and **Years** in X-axis. Also took Companies in Column series having filter for top ten values.
- Top Companies made less profit during year 2016 and made more profit in 2013.



## 2. Line Chart :

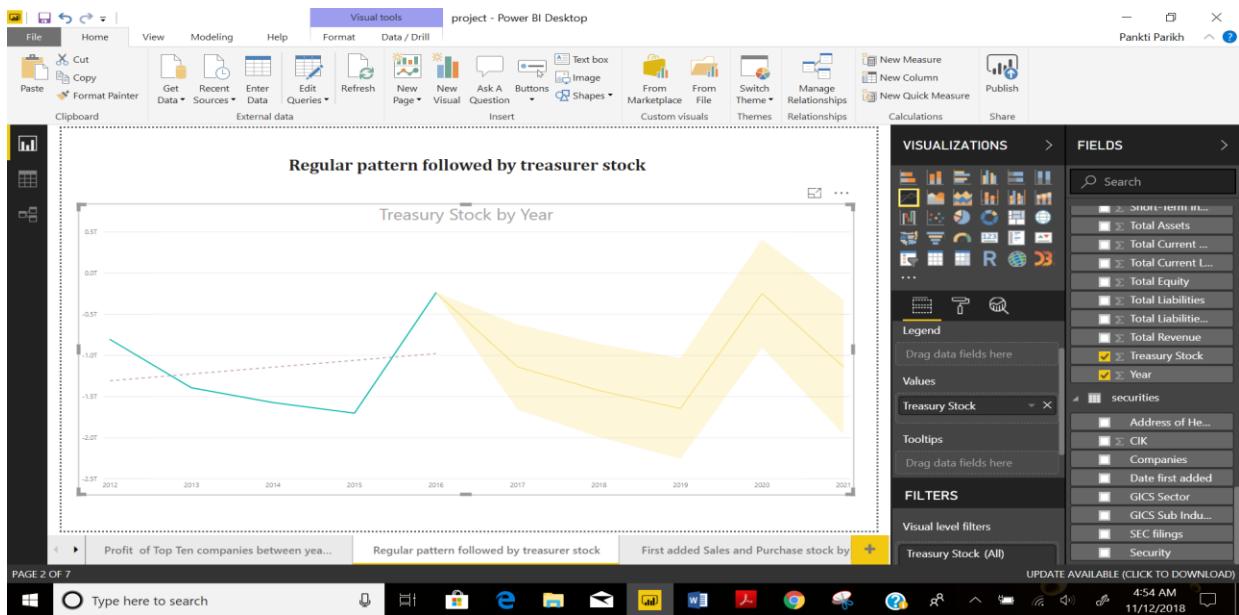
- Line charts are used to show resulting data relative to a continuous variable - most commonly time or money. They are great for projections of performance beyond your data. If you plot your sales vs. month on a line chart over the past two years, it is easy for the reader to identify any trends that may be useful as you plan for the upcoming year.
- A dual axis chart allows you to plot data using two y-axes and a shared x-axis. It's used with three data sets, one of which is based on a continuous set of data and another which is better suited to being grouped by category. This should be used to visualize a correlation or the lack thereof between these three data sets.

### Question 2: Is there any regular pattern followed by treasurer stock?

- Y-axis represent treasurer Stock. X-axis represents year given in the dataset (2012-2016) and also have future trend line for years 2016 to 2021 having upper bound and lower bound

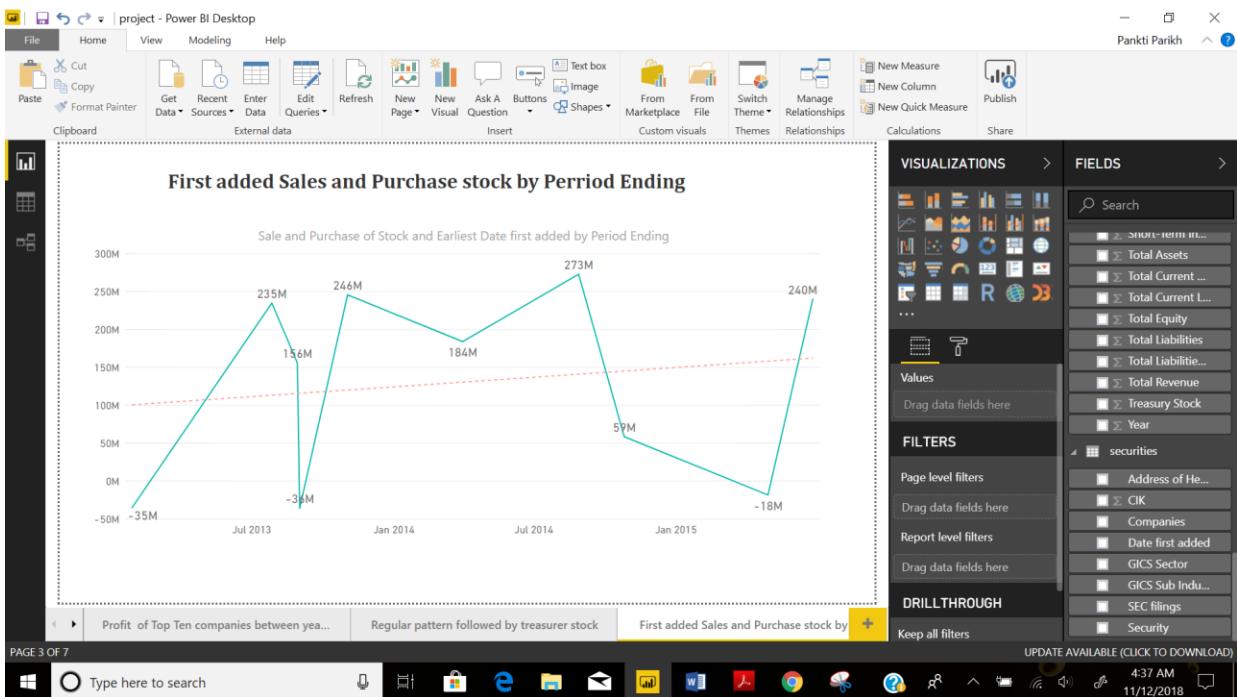
Values. I created forecast trend line using tooltip and also have trend line for year 2012 to 2016 so that we can see specific pattern followed in this year And predict the future trend line using different points.

- In future, Stock price of the treasurer stock will decrease during year 2016 to 2019 and then in between year 2019 and 2020 prices will be increase almost nearest to the stock price of year 2016.



### 3. Line Chart having Trend Line:

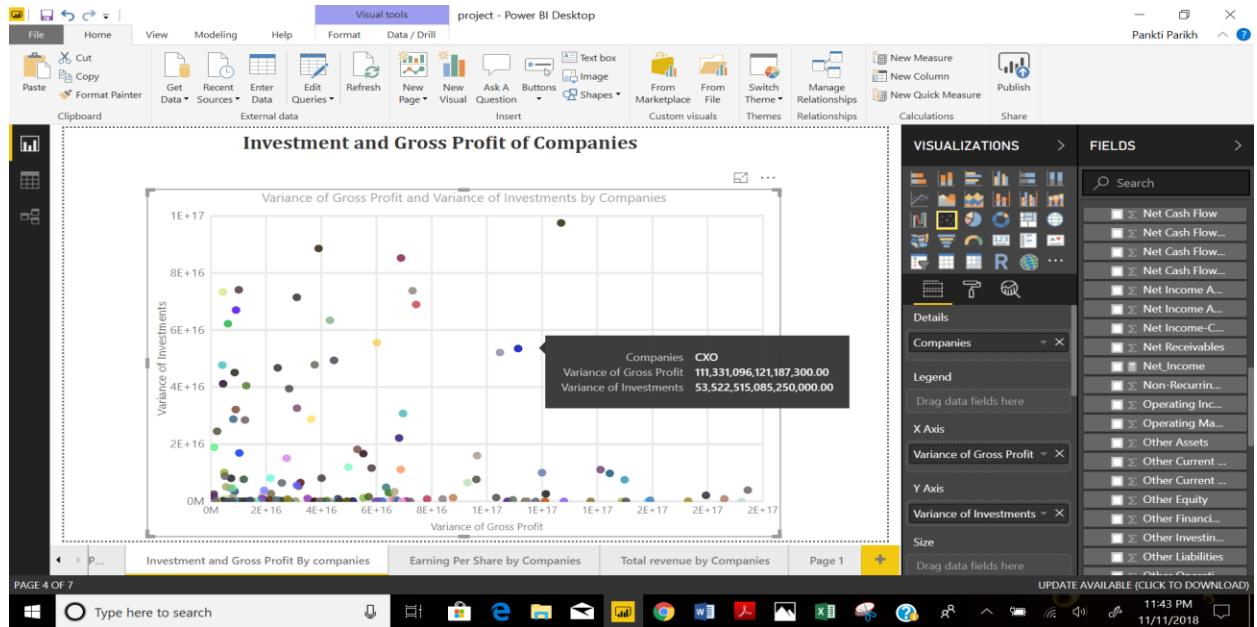
- Following line chart shows the sales and product stock added earliest before the ending period. You can see the sales and product stock prices fluctuation during first end period Jan to July and second end period July to Jan of next year.
- Here, you can see more fluctuation in the stock price during July 2013 to Jan 2014. And lowest fluctuation between year Jan 2014 to July 2014.



### 4. Scatter Chart :

- A scatter plot chart will show the relationship between two different variables. Scatter plots are useful for quickly understanding if there is a relationship between two variables. If the data forms a band extending from lower left to upper right, there is most likely a positive correlation between the two variables. If the band runs from upper left to lower right, a negative correlation is probable. If it is hard to see a pattern, there is probably no correlation.

- This scatter plot portrayed Correlation between Variance of Gross Profit and variance of Investment made by Companies.



## 5. Pie Chart :

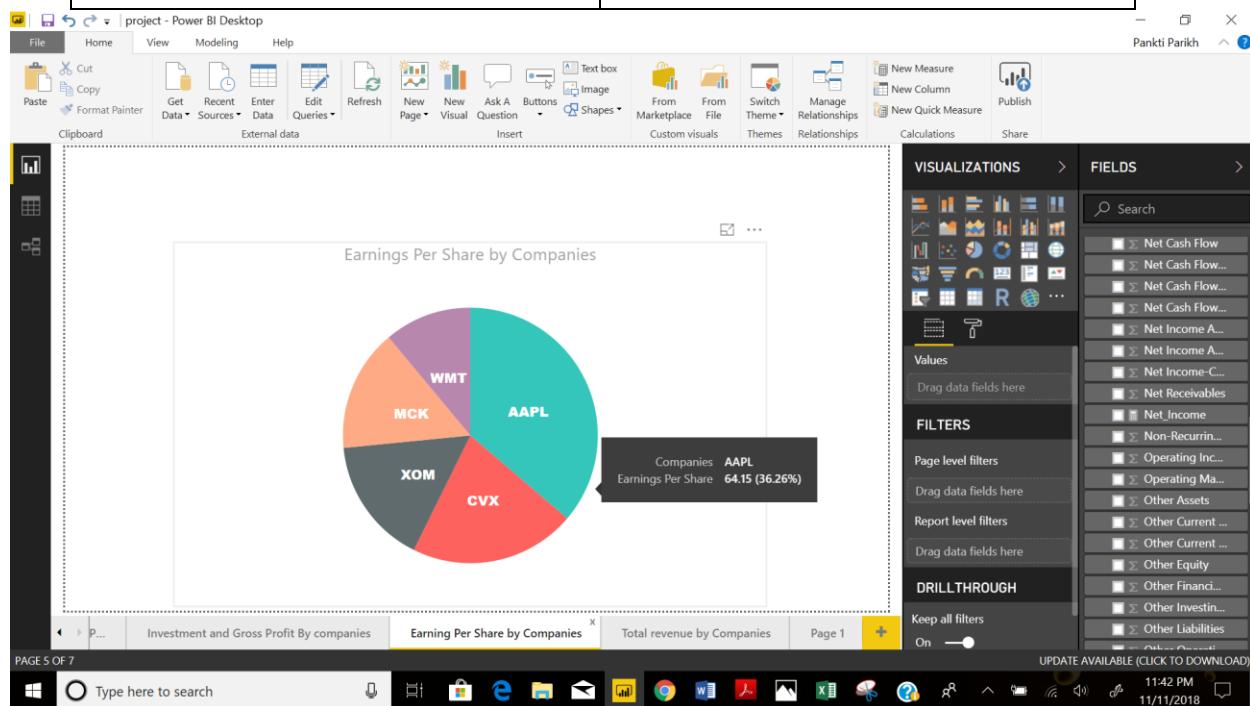
- Pie charts are easy to read and fun to look at making them a great choice if you want to understand the parts of a whole. It's a good practice to order the pieces of your pie according to size and always ensure the total of all the pieces add up to 100%.

**Question 3 : Display the top 5 company's Earning per share.**

- As you see Earning per share of top 5 companies. Highest Earning made per share by AAPL which is 36.26% having 64.15 shares .

Company Name	Earning per share (in percentage)
AAPL	36.26%
CVX	21.06%

XOM	16.09%
MCK	15.27%
WMT	11.07%

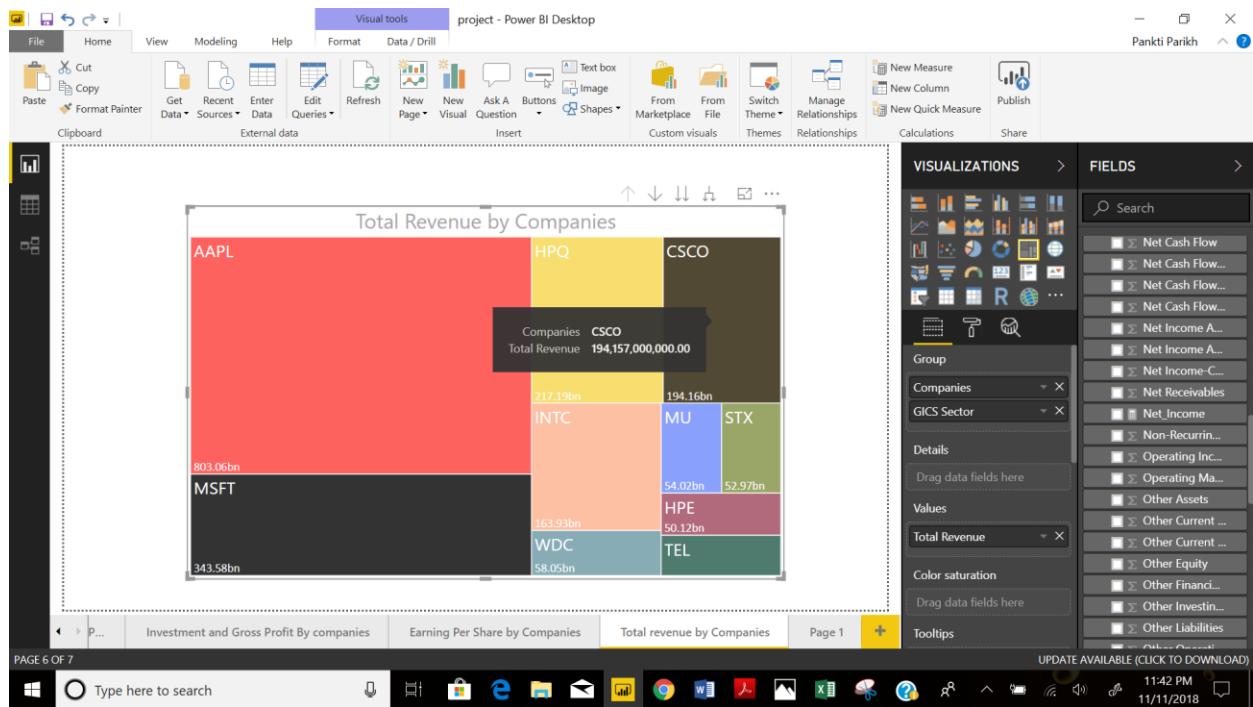


## 6. Tree Map :

- Treemapping is a data visualization technique that is used to display hierarchical data using nested rectangles; the treemap chart is created based on this technique of data visualization.

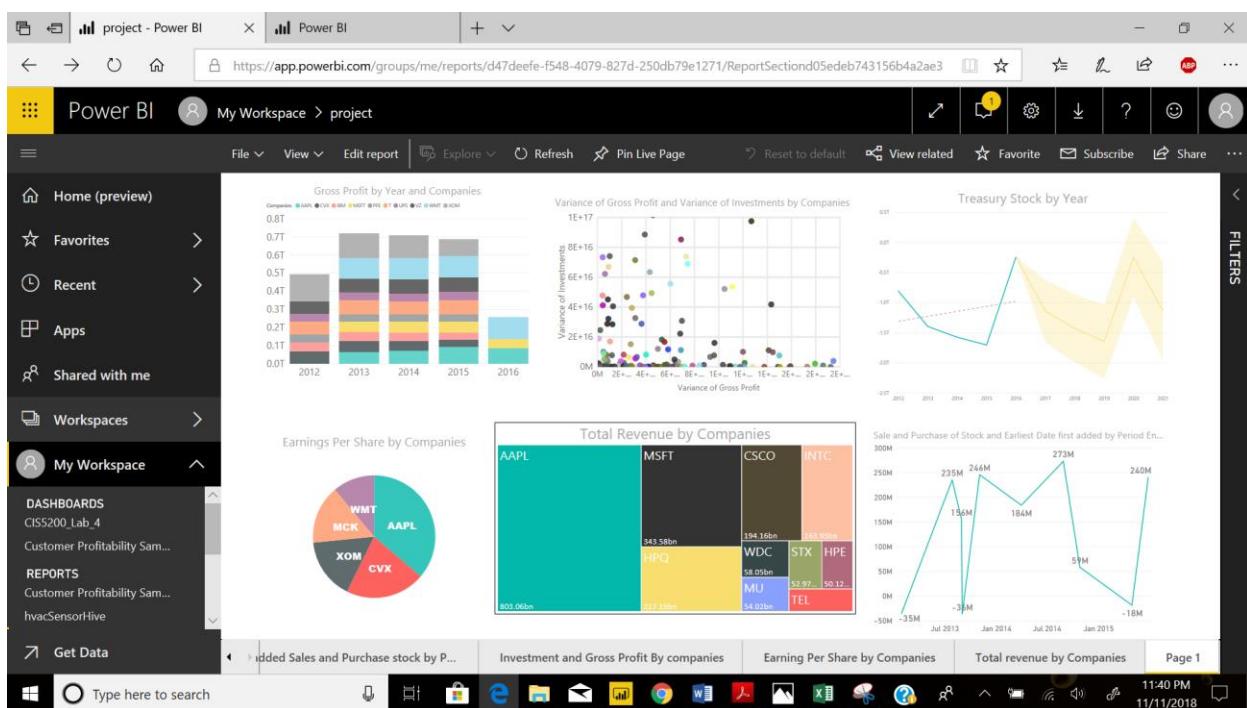
### Question 4 : Total revenue generated by Top 10 IT companies.

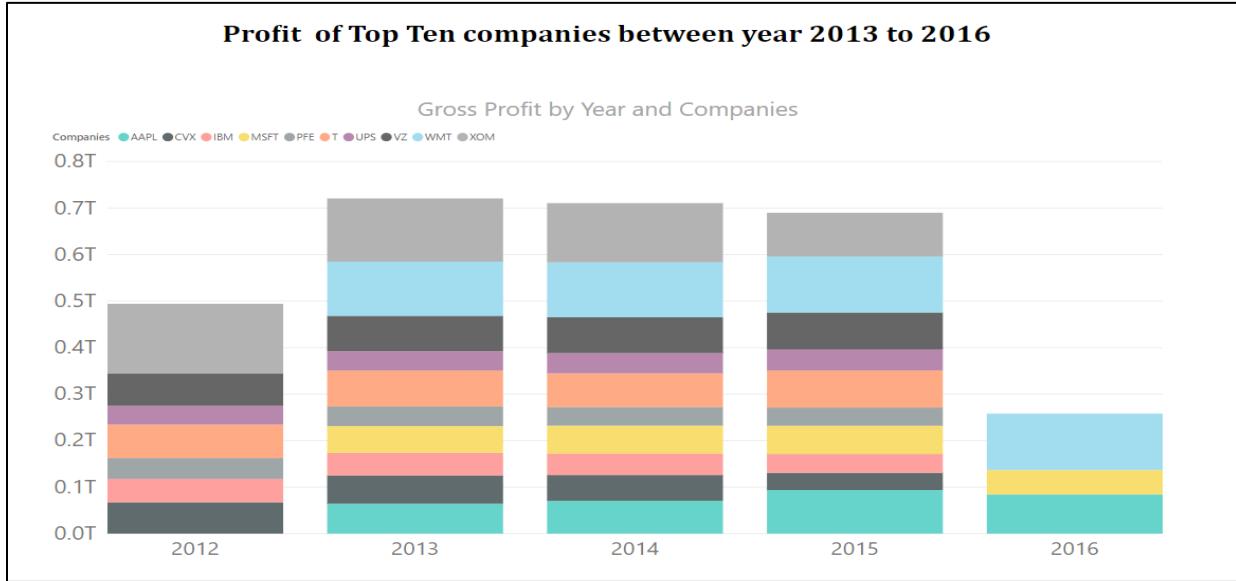
- There are top 10 IT companies filtered in following Tree map in which AAPL company made highest revenue because in previous visual you can see they have more Earning per share so they made more revenue 806.03 Billoions.



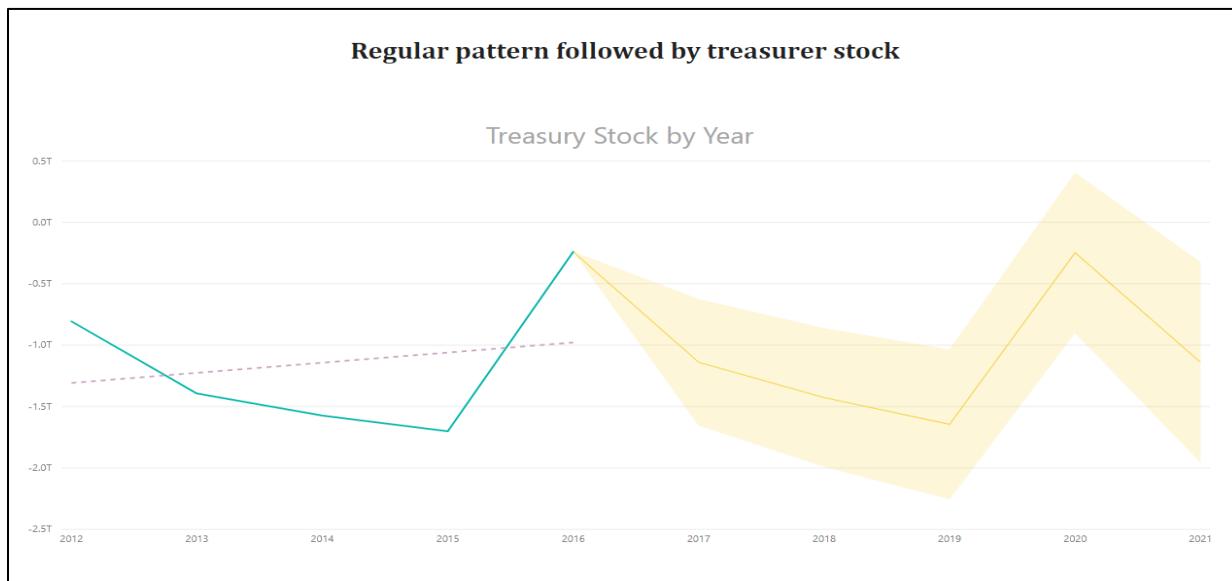
## H. Dash Board and Story Telling:

- A Power BI dashboard is a single page, often called a canvas, that uses visualizations to tell a story. Because it is limited to one page, a well-designed dashboard contains only the most-important elements of that story.

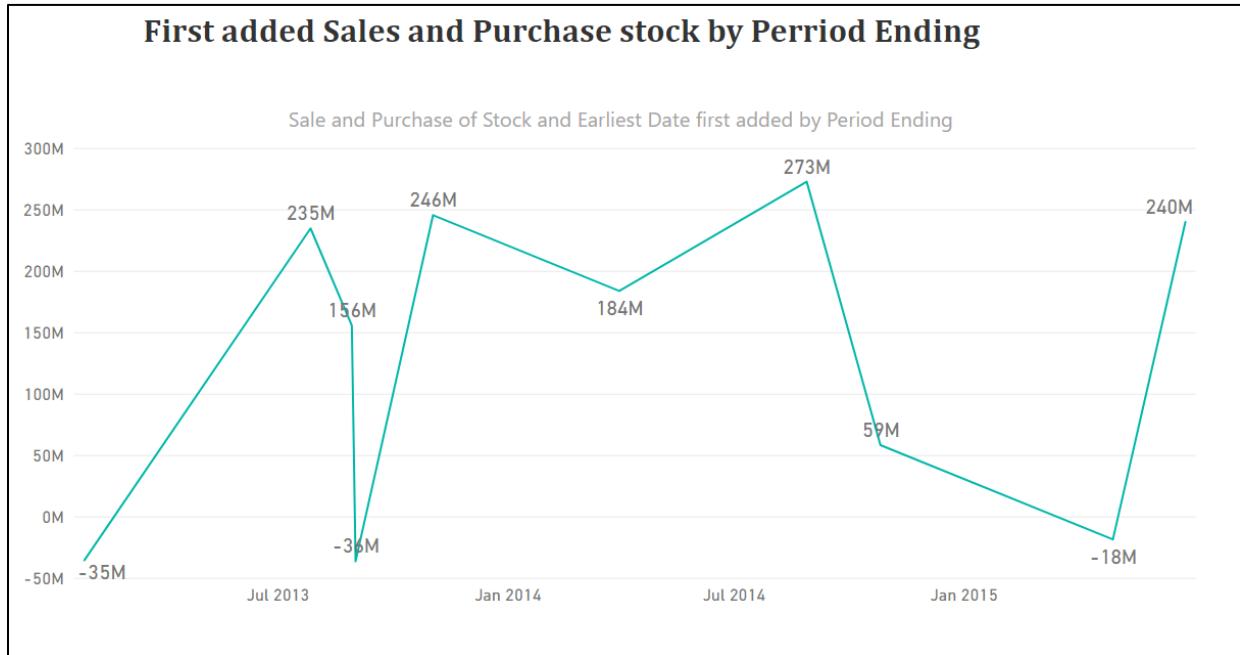


**Story Telling:****Chart 1:**

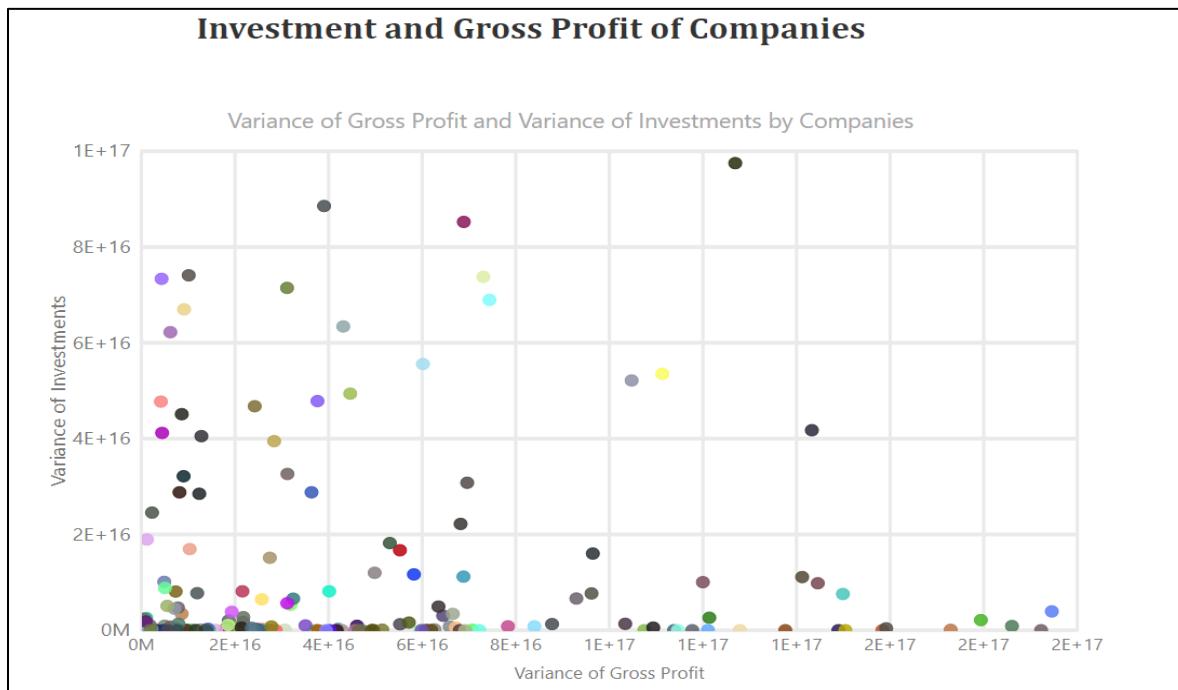
- From above visual we can conclude that only 3 companies out of top 10 companies are able to stay on top [profit wise] in year 2016 compare to last three years.
- WMT and XOM Companies are on top continuously by making more profit during 2012 to 2016.

**Chart 2:**

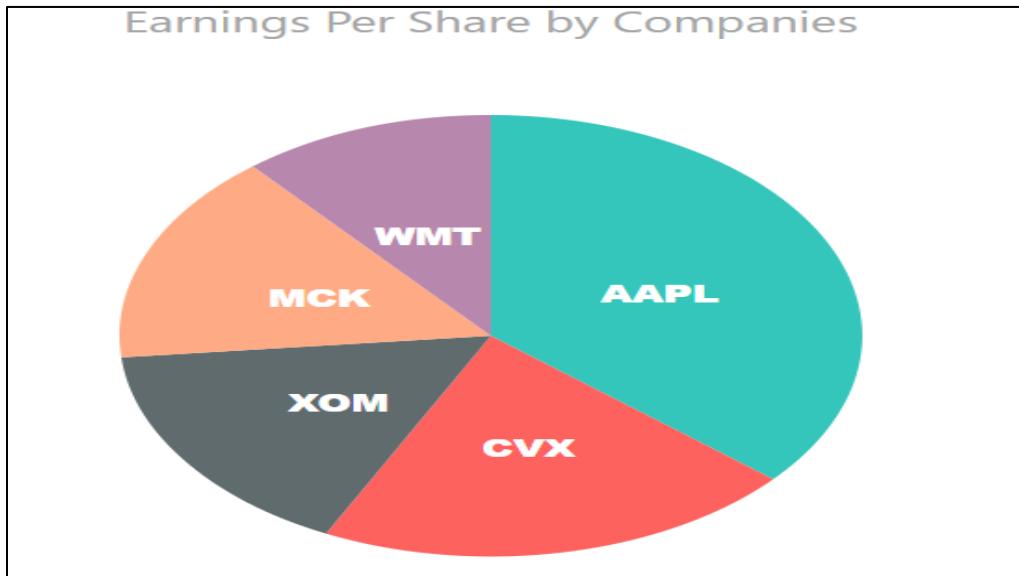
- In future, Stock price of the treasurer stock will decrease during year 2016 to 2019 and then in between year 2019 and 2020 prices will be increase almost nearest to the stock price of year 2016.

**Chart 3:**

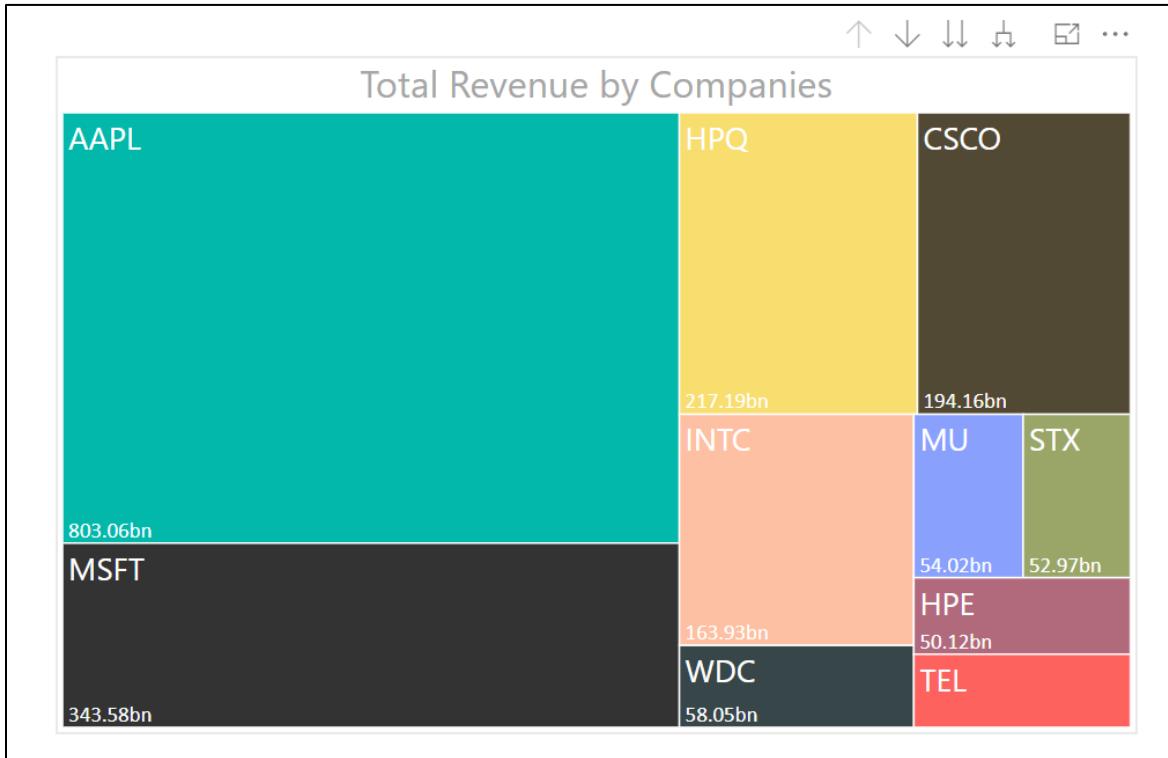
- Prices of stock decreases before January of each year and stock prices are on their lowest level during January. After January Stock prices are suddenly increases.
- Here, you can see more fluctuation in the stock price during July 2013 to Jan 2014. And lowest fluctuation between year Jan 2014 to July 2014.

**Chart 4:**

- Investors make more profit by investing less in companies. While those who invest more their chances become less to get more profit.
- We can also say that people used to make less investment and get less profit.

**Chart 5:**

- AAPL company has almost double earning per share compare to WMT,MCK,XOM , and CWX .
- As you see Earning per share of top 5 companies. Highest Earning made per share by AAPL which is 36.26% having 64.15 shares .

**Chart 6:**

- Revenue generated by company AAPL is almost 40% in top 10 IT companies.
- There are top 10 IT companies filtered in following Tree map in which AAPL company made highest revenue because in previous visual you can see they have more Earning per share so they made more revenue 806.03 Billions.

## I. Purpose Map:

- Special purpose maps are designed or created for the special purpose. They are not standard map, specialized map displaying particular information.
- There are 3 main elements of purpose Map:
  - Explanatory:** Viewer is assisted to interpreting stage beyond perceiving stage in this stage user is scanning document and reading the data.
  - Exhibitory:** The viewer has to work to interpret the meaning, relying on their own capacity.
  - Exploratory:** Viewer is assisted to the extent of perceiving and then left with finding their own insights.

