

# Assignment 1:

## Analysis of Bank Filings Data – Data Wrangling and Exploratory data analysis using Tableau Case Study

**Under guidance of :**

Prof. Sri Krishnamurthy

Ashwin Dinoriya

**Team 1**

Amulya Aankul

Nand Govind Modani

Saksham Agrawal

Part 1:

Analyzing Banks with asset values greater  
than \$10 Billion

# Process workflow:

Step 1 : Web Scraping using Python 3

Step 2 : Cleaning the data

Step 3 : Creating stacked and unstacked CSV file using Pandas

Step 4 : Pre-processing the data for Analysis

Step 5 : Creating Dashboards in Tableau

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import re

# hitting the ffiec website with a "get" request
r = requests.get("https://www.ffiec.gov/nicpubweb/nicweb/HCSGreaterThan100.aspx")
# saving response
html_doc = r.text
# parsing the page using BeautifulSoup module
soup = BeautifulSoup(html_doc, 'html.parser')
# reading option tag to know all the quarters
option_tags = soup.findAll('option')
# list to save data of all the files
records = []
# list to header of all the columns
headers = []
flag = 0
# for loop to traverse data for all quarters
for tags in option_tags:
    date = tags.contents
    # preparing the AJAX call to load data "quarter" wise
    date_str = str(date[0])
```

## Step 2 : Cleaning the data

- Extracting the RSSD ID from the bank name
- Made RSSD ID unique column so that no new row gets created if Bank Changes their name in any quarter
- Sanitizing the bank name
- Extracting column headers
- Creating a new column for quarters
- Removing the Rank column

## Step 3 : Creating stacked and unstacked CSV file using Pandas

- Creating a 2 dimensional structure for the CSV file.
- Pivoting the data frame to create unstacked version of the file



quest1.py

*File: Python code for Step 1-3*

# Output CSV : Stacked Version

	A	B	C	D	E
1	RSSD ID	Institution Name	Location	TotalAssets	Date
2	1039502	JPMORGAN CHASE & CO.	NEW YORK, NY	\$2,466,096,000	20160630
3	1073757	BANK OF AMERICA CORPORATION	CHARLOTTE, NC	\$2,189,811,000	20160630
4	1120754	WELLS FARGO & COMPANY	SAN FRANCISCO, CA	\$1,889,235,000	20160630
5	1951350	CITIGROUP INC.	NEW YORK, NY	\$1,818,771,000	20160630
6	2380443	GOLDMAN SACHS GROUP, INC., THE	NEW YORK, NY	\$896,870,000	20160630
7	2162966	MORGAN STANLEY	NEW YORK, NY	\$828,873,000	20160630
8	1119794	U.S. BANCORP	MINNEAPOLIS, MN	\$438,463,000	20160630
9	3587146	BANK OF NEW YORK MELLON CORPORATION, T	NEW YORK, NY	\$372,351,000	20160630
10	1069778	PNC FINANCIAL SERVICES GROUP, INC., THE	PITTSBURGH, PA	\$361,528,406	20160630
11	2277860	CAPITAL ONE FINANCIAL CORPORATION	MCLEAN, VA	\$339,247,718	20160630
12	3232316	HSBC NORTH AMERICA HOLDINGS INC.	NEW YORK, NY	\$295,534,689	20160630
13	3606542	TD GROUP US HOLDINGS LLC	WILMINGTON, DE	\$276,317,370	20160630
14	1607170	TEACHERS INSURANCE & ANNUITY ASSOCIATIC	NEW YORK, NY	\$276,045,408	20160630
15	1111435	STATE STREET CORPORATION	BOSTON, MA	\$255,396,733	20160630
16	1074156	BB&T; CORPORATION	WINSTON SALEM, NC	\$221,858,615	20160630
17	1131787	SUNTRUST BANKS, INC.	ATLANTA, GA	\$199,276,480	20160630
18	1026632	CHARLES SCHWAB CORPORATION, THE	SAN FRANCISCO, CA	\$198,052,000	20160630
19	1275216	AMERICAN EXPRESS COMPANY	NEW YORK, NY	\$159,632,000	20160630
20	1562859	ALLY FINANCIAL INC.	DETROIT, MI	\$157,931,000	20160630
21	3226762	RBC USA HOLDCO CORPORATION	NEW YORK, NY	\$151,710,605	20160630
22	1132449	CITIZENS FINANCIAL GROUP, INC.	PROVIDENCE, RI	\$145,568,297	20160630
23	1447376	UNITED SERVICES AUTOMOBILE ASSOCIATION	SAN ANTONIO, TX	\$144,819,412	20160630
24	3840207	STATE FARM MUTUAL AUTOMOBILE INSURANC	BLOOMINGTON, IL	\$143,801,553	20160630
25	1070345	FIFTH THIRD BANCORP	CINCINNATI, OH	\$143,625,325	20160630
26	1245415	BMO FINANCIAL CORP.	WILMINGTON, DE	\$132,007,952	20160630
27	3981856	SANTANDER HOLDINGS USA, INC.	BOSTON, MA	\$126,502,203	20160630
28	3242838	REGIONS FINANCIAL CORPORATION	BIRMINGHAM, AL	\$126,378,482	20160630
29	1037003	M&T; BANK CORPORATION	BUFFALO, NY	\$123,820,584	20160630
30	1199611	NORTHERN TRUST CORPORATION	CHICAGO, IL	\$121,509,559	20160630

# Output CSV : Unstacked Version

	A	B	C	D	E	F	G	H	I	J	K
1	RSSD ID	Institution Name	Location	20121231	20130331	20130630	20130930	20131231	20140331	20140630	20140930
2	1020180	BREMER FINANCIAL CORPORATION	SAINT PAUL, MN								
3	1020902	FIRST NATIONAL OF NEBRASKA, INC.	OMAHA, NE	\$16,409,360	\$15,890,934	\$15,466,494	\$15,977,890	\$16,271,487	\$16,573,510	\$16,754,901	\$17,082,411
4	1025309	BANK OF HAWAII CORPORATION	HONOLULU, HI	\$13,789,923	\$13,563,474	\$13,787,068	\$13,916,694	\$14,127,598	\$14,333,411	\$14,884,625	\$14,577,411
5	1025608	BANCWEST CORPORATION	HONOLULU, HI	\$79,869,488	\$78,851,727	\$80,069,917	\$81,729,444	\$83,527,474	\$84,945,155	\$86,617,152	\$86,894,411
6	1025608	FIRST HAWAIIAN, INC.	HONOLULU, HI								
7	1026632	CHARLES SCHWAB CORPORATION, THE	SAN FRANCISCO, CA	\$133,637,000	\$133,324,000	\$135,907,000	\$140,211,000	\$143,642,000	\$144,066,000	\$143,401,000	\$147,445,000
8	1027004	ZIONS BANCORPORATION	SALT LAKE CITY, UT	\$55,511,918	\$54,110,564	\$54,904,540	\$55,188,312	\$56,031,127	\$56,080,844	\$55,111,275	\$55,458,000
9	1027518	CITY NATIONAL CORPORATION	LOS ANGELES, CA	\$28,618,492	\$27,433,754	\$27,379,501	\$29,059,404	\$29,717,951	\$29,851,542	\$30,819,092	\$32,015,000
10	1031449	SVB FINANCIAL GROUP	SANTA CLARA, CA	\$22,766,602	\$22,802,242	\$22,165,054	\$23,757,075	\$26,417,306	\$29,724,778	\$33,322,533	\$36,044,000
11	1032473	DEUTSCHE BANK TRUST CORPORATION	NEW YORK, NY	\$74,148,000	\$75,523,000	\$71,992,000	\$66,067,000	\$66,926,000	\$72,603,000	\$69,406,000	\$60,715,000
12	1036967	CIT GROUP INC.	LIVINGSTON, NJ	\$44,012,251	\$44,563,888	\$44,631,022	\$46,223,981	\$47,138,960	\$48,578,081	\$44,152,666	\$46,480,000
13	1037003	M&T; BANK CORPORATION	BUFFALO, NY	\$82,985,468	\$82,794,833	\$83,229,005	\$84,427,485	\$85,162,391	\$88,530,360	\$90,835,002	\$97,230,000
14	1039502	JPMORGAN CHASE & CO.	NEW YORK, NY	\$2,359,141,000	\$2,389,349,000	\$2,439,494,000	\$2,463,309,000	\$2,415,689,000	\$2,476,650,000	\$2,519,995,000	\$2,526,655,000
15	1048773	VALLEY NATIONAL BANCORP	WAYNE, NJ	\$16,012,646	\$16,028,703	\$15,977,202	\$15,976,943	\$16,156,541	\$16,344,464	\$16,335,967	\$16,726,000
16	1049341	COMMERCE BANCSHARES, INC.	KANSAS CITY, MO	\$22,176,815	\$22,240,695	\$21,921,322	\$22,462,282	\$23,081,892	\$22,820,527	\$23,016,162	\$22,710,000
17	1049828	UMB FINANCIAL CORPORATION	KANSAS CITY, MO	\$14,927,196	\$15,705,470	\$15,253,217	\$16,184,233	\$16,911,852	\$15,945,830	\$15,562,690	\$16,284,000
18	1060627	FIRSTBANK HOLDING COMPANY	LAKEWOOD, CO	\$12,874,974	\$13,078,898	\$13,022,264	\$13,169,177	\$13,384,848	\$14,009,813	\$13,941,733	\$14,052,000
19	1068025	KEYCORP	CLEVELAND, OH	\$89,425,613	\$89,441,210	\$90,858,779	\$91,016,413	\$92,991,716	\$90,928,218	\$91,934,784	\$89,884,000
20	1068191	HUNTINGTON BANCSHARES INCORPORATED	COLUMBUS, OH	\$56,153,185	\$56,054,966	\$56,113,687	\$56,648,251	\$59,476,344	\$61,145,753	\$63,797,113	\$64,330,000
21	1069778	PNC FINANCIAL SERVICES GROUP, INC.	PITTSBURGH, PA	\$305,285,879	\$300,945,933	\$304,547,667	\$308,912,603	\$320,596,232	\$323,586,973	\$327,251,474	\$334,602,000
22	1070345	FIFTH THIRD BANCORP	CINCINNATI, OH	\$121,500,604	\$120,789,641	\$122,642,935	\$124,939,419	\$129,685,180	\$129,654,487	\$132,562,382	\$134,187,000
23	1070804	FIRSTMERIT CORPORATION	AKRON, OH	\$14,916,022	\$15,273,159	\$23,534,261	\$24,140,030	\$23,912,451	\$24,500,602	\$24,566,414	\$24,610,000
24	1073757	BANK OF AMERICA CORPORATION	CHARLOTTE, NC	\$2,212,004,452	\$2,176,625,000	\$2,125,686,000	\$2,128,706,000	\$2,104,995,000	\$2,152,533,000	\$2,172,001,000	\$2,126,138,000
25	1074156	BB&T; CORPORATION	WINSTON SALEM, NC	\$183,872,371	\$180,836,757	\$182,735,153	\$181,050,008	\$183,009,992	\$184,651,158	\$188,012,330	\$187,021,000
26	1075612	FIRST CITIZENS BANCSHARES, INC.	RALEIGH, NC	\$21,283,651	\$21,351,012	\$21,308,822	\$21,511,352	\$21,199,091	\$22,154,997	\$22,062,840	\$21,942,000
27	1076217	UNITED BANCSHARES, INC.	CHARLESTON, WV						\$11,886,320	\$12,051,710	\$12,085,000
28	1078530	BBVA COMPASS BANCSHARES, INC.	HOUSTON, TX			\$60,674,076	\$70,087,078	\$71,065,476	\$74,057,337	\$75,747,100	\$70,187,000



# Step 4 : Pre-processing the data for Analysis

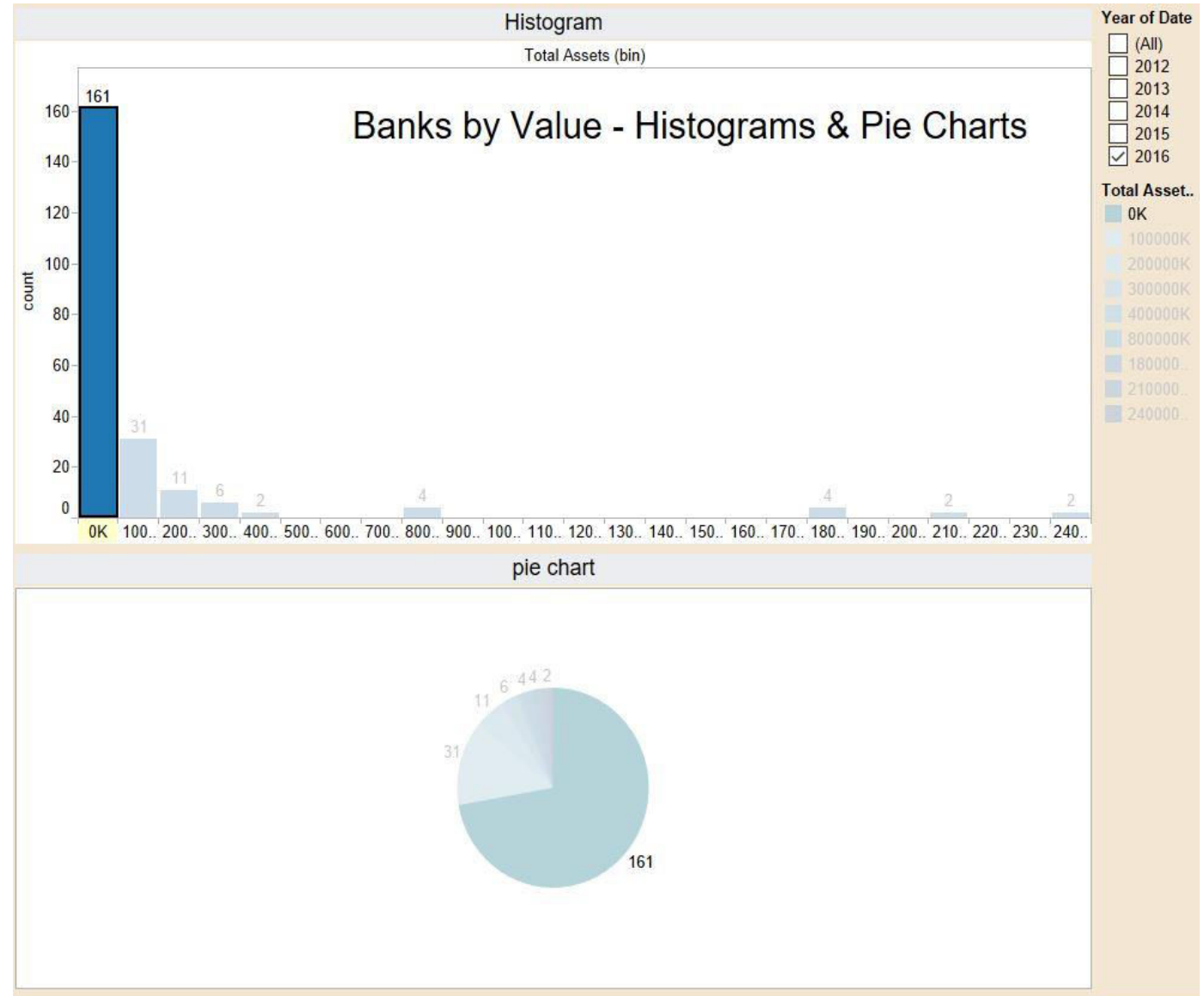
- Splitting the Location into city and state for analysis purpose on Tableau

Rssd Id # nic_ma...	Institution Name Abc nic_main_data....	Location Abc nic_main_data....	City =🌐	State =🌐	Total Assets # nic_main_d...	Date 📅 nic_ma...	Total Assets (bin) .ill.	count =#
1039502	JPMORGAN CHAS...	NEW YORK, NY	NEW YORK	NY	2,466,096,000.00	6/30/2016	2,400,000,000	1
1073757	BANK OF AMERIC...	CHARLOTTE, NC	CHARLOTTE	NC	2,189,811,000.00	6/30/2016	2,100,000,000	1
1120754	WELLS FARGO & ...	SAN FRANCISCO, ...	SAN FRANCISCO	CA	1,889,235,000.00	6/30/2016	1,800,000,000	1
1951350	CITIGROUP INC.	NEW YORK, NY	NEW YORK	NY	1,818,771,000.00	6/30/2016	1,800,000,000	1
2380443	GOLDMAN SACH...	NEW YORK, NY	NEW YORK	NY	896,870,000.00	6/30/2016	800,000,000	1
2162966	MORGAN STANLEY	NEW YORK, NY	NEW YORK	NY	828,873,000.00	6/30/2016	800,000,000	1
1119794	U.S. BANCORP	MINNEAPOLIS, MN	MINNEAPOLIS	MN	438,463,000.00	6/30/2016	400,000,000	1
3587146	BANK OF NEW YO...	NEW YORK, NY	NEW YORK	NY	372,351,000.00	6/30/2016	300,000,000	1
1069778	PNC FINANCIAL S...	PITTSBURGH, PA	PITTSBURGH	PA	361,528,406.00	6/30/2016	300,000,000	1
2277860	CAPITAL ONE FIN...	MCLEAN, VA	MCLEAN	VA	339,247,718.00	6/30/2016	300,000,000	1
3232316	HSBC NORTH AM...	NEW YORK, NY	NEW YORK	NY	295,534,689.00	6/30/2016	200,000,000	1
3606542	TD GROUP US HO...	WILMINGTON, DE	WILMINGTON	DE	276,317,370.00	6/30/2016	200,000,000	1
.....	.....	.....	.....	....	.....	.....	.....	.

# Step 5 : Creating Dashboards in Tableau

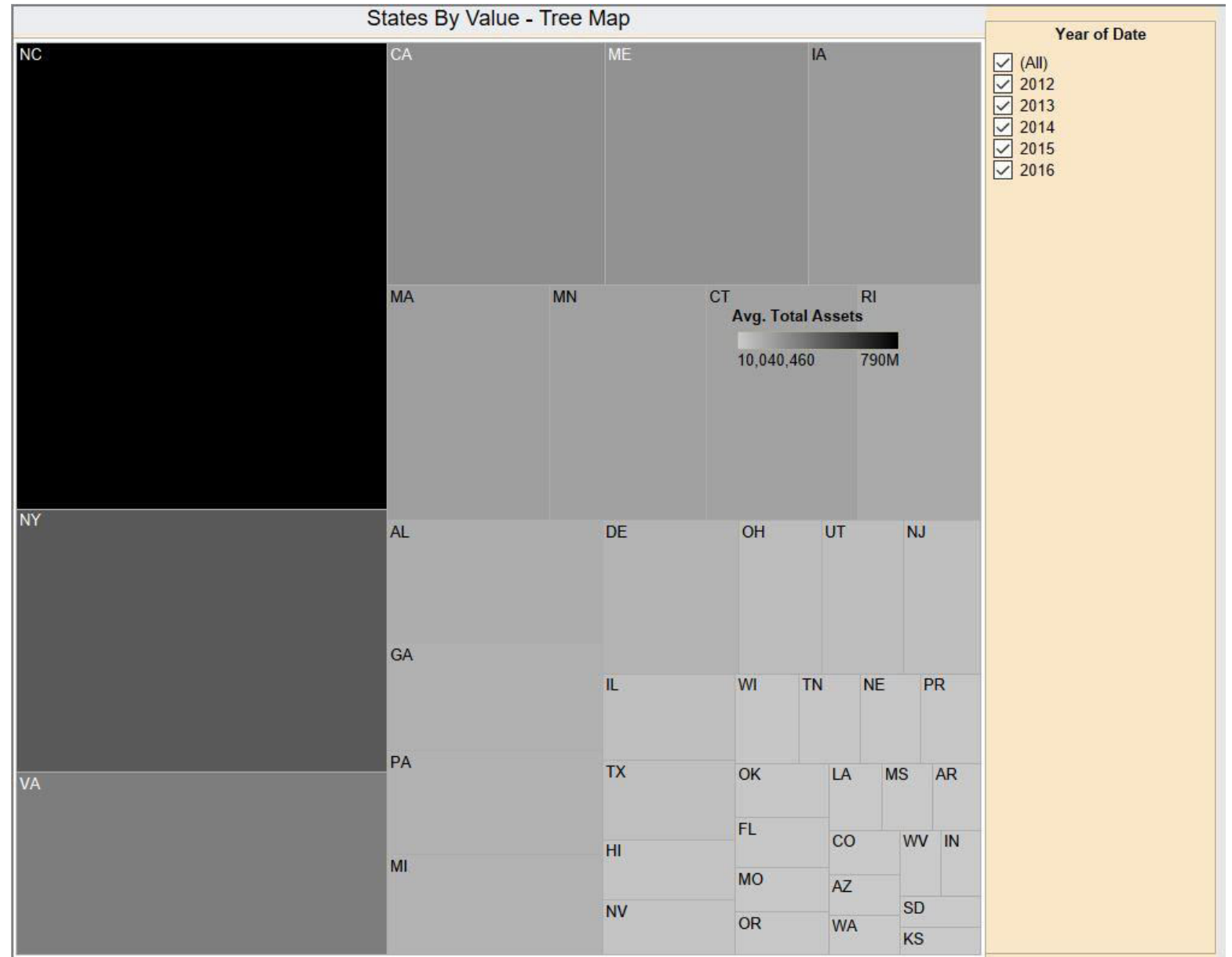
## Banks by Value:

- Histograms shows the count of companies in the range on the x axis.
- Pie chart shows the count of companies in the range. Gives us insights about the percentage of the total value.



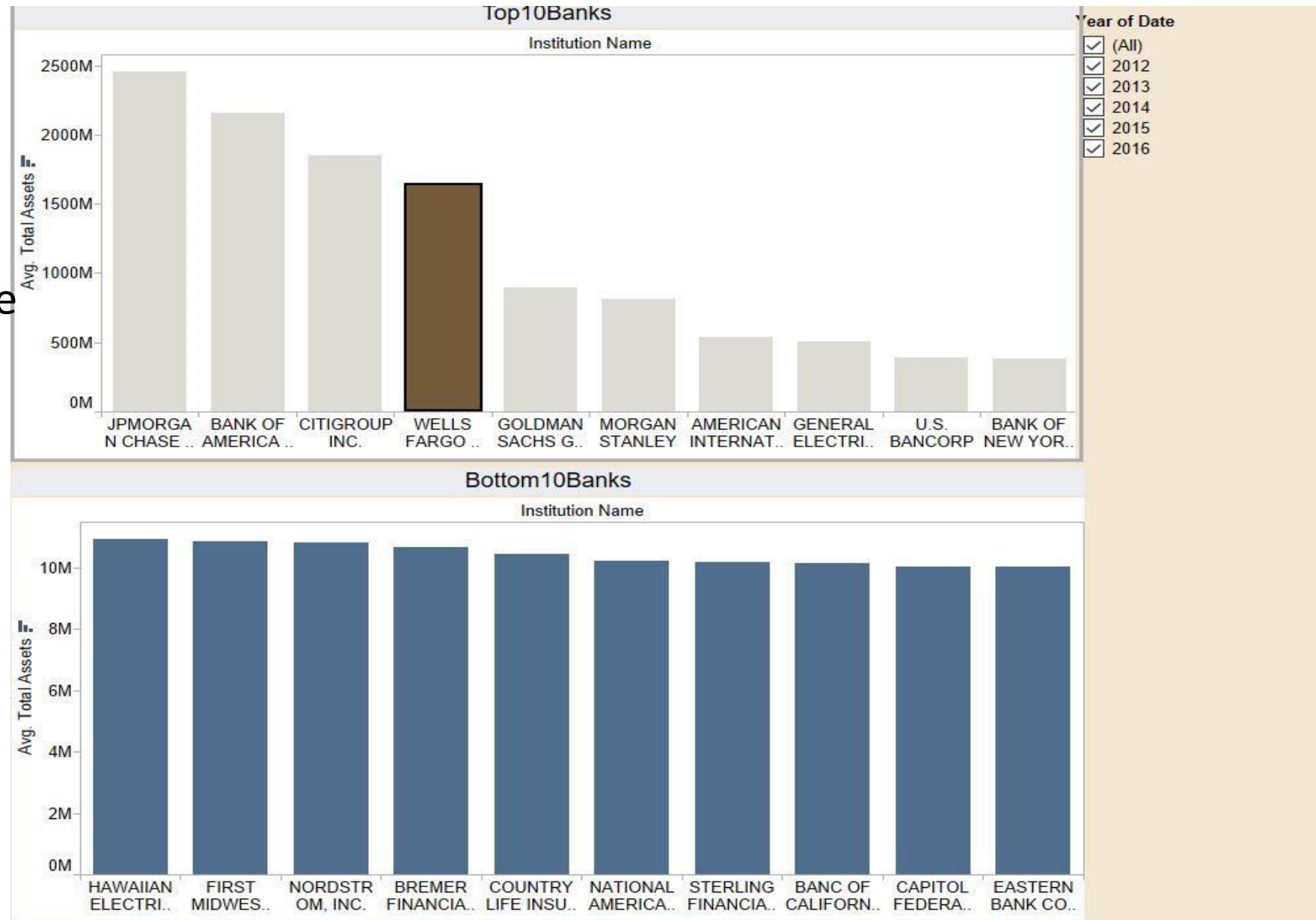
# States by Value:

- Tree map here shows the assets of companies grouped by the state of where they are located.
- The values are filtered according to the Year. If multiple years are selected it takes the average of the values



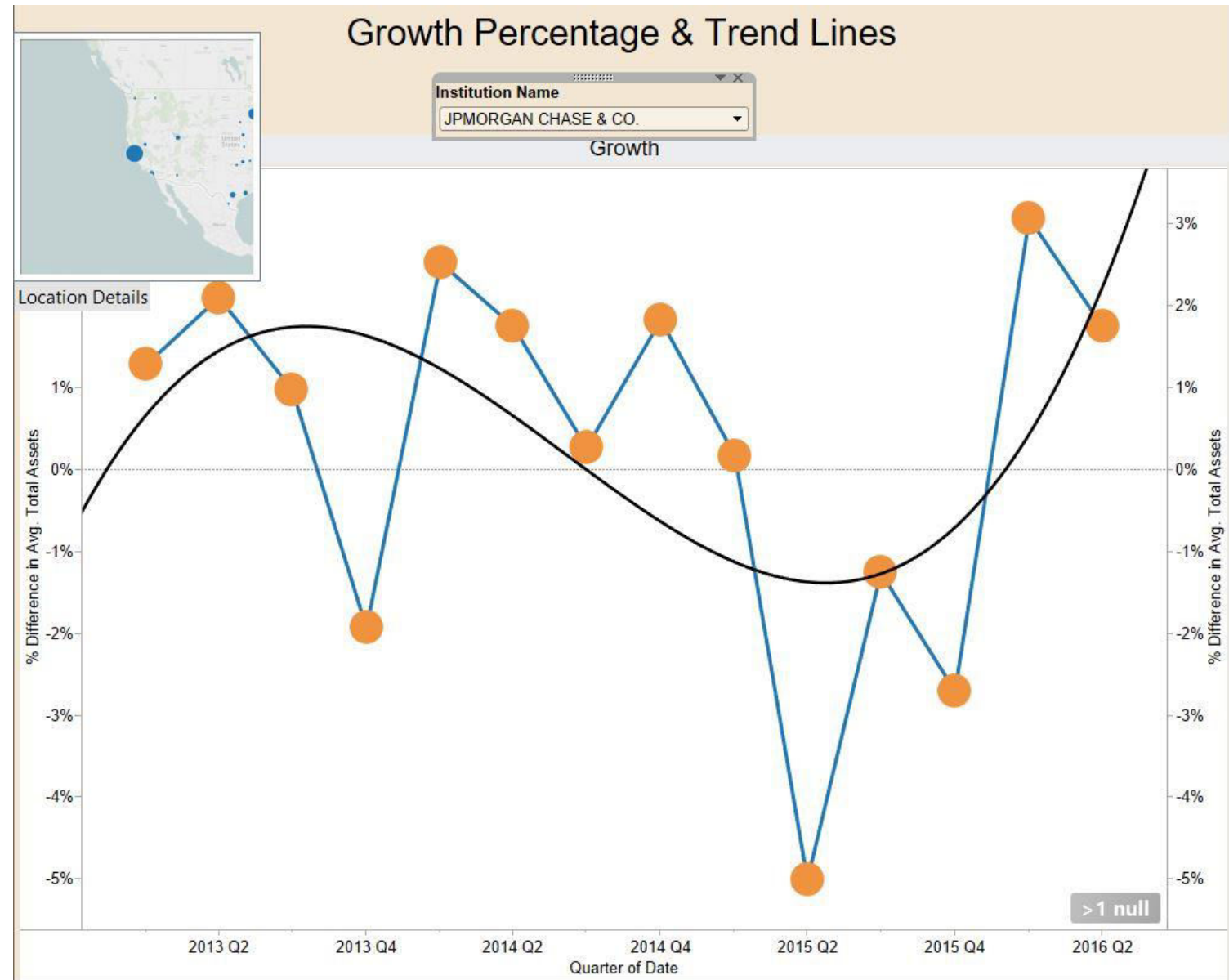
# Top and Bottom 10 banks:

- The bar chart shows the top 10 and bottom 10 banks by value of their assets
- The values are filtered according to the Year. If multiple years are selected it takes the average of the values



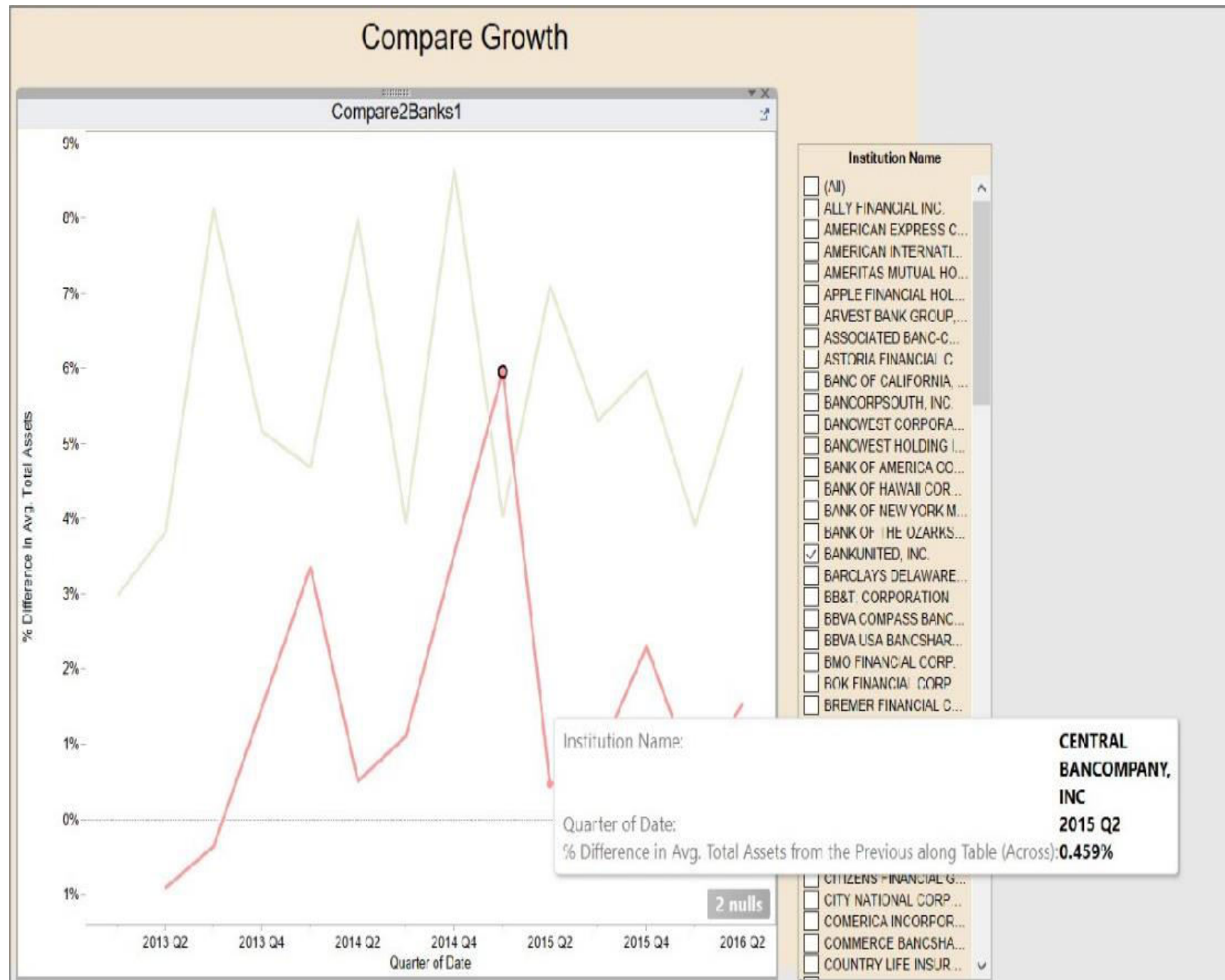
# Growth Percentage:

- Shows the percentage growth of assets & trend lines with respect to successive quarters
- The figure shows the trend line for JP Morgan, which shows their assets are expected to increase in the next quarter.



## Growth comparison:

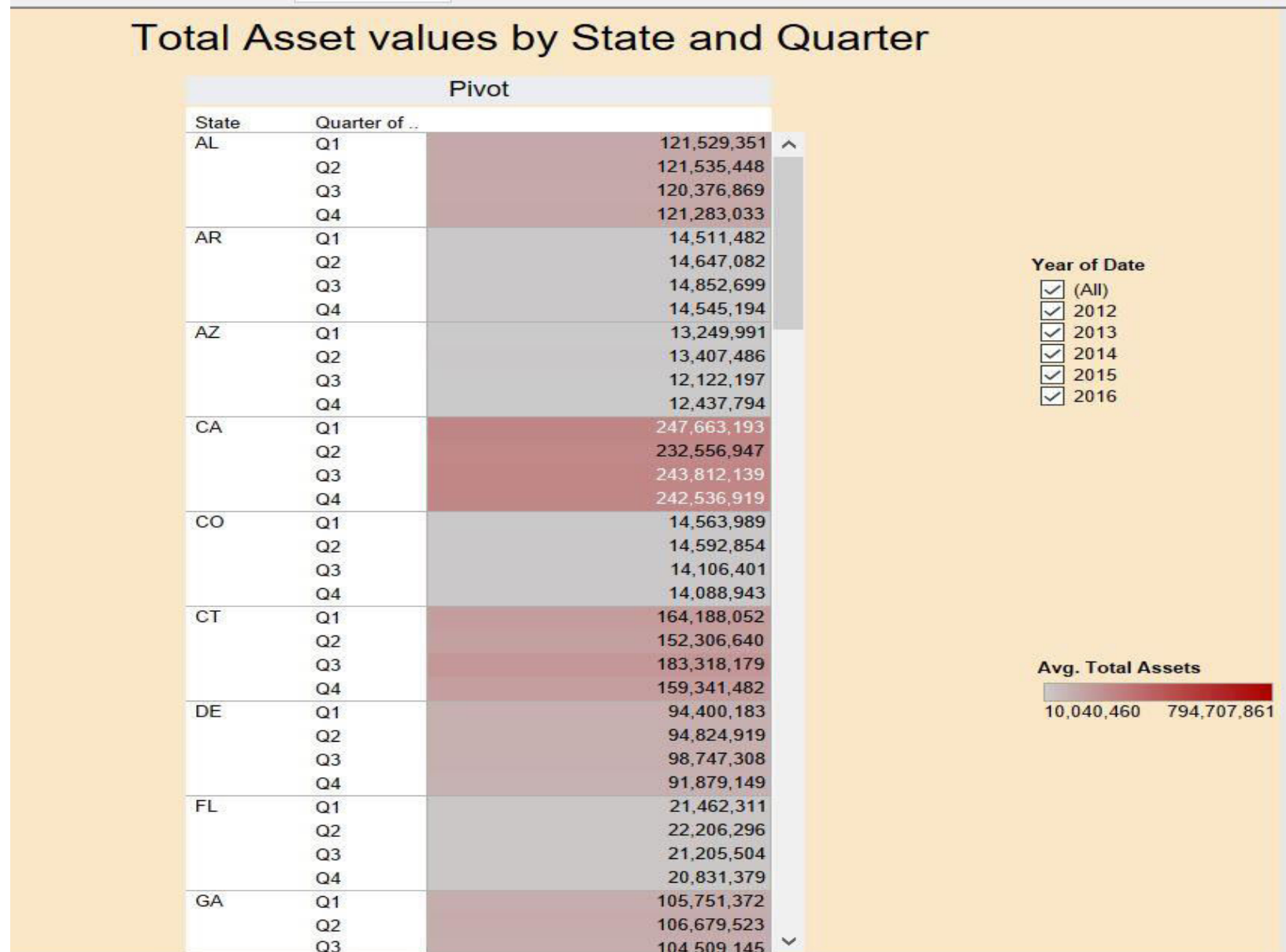
- The graph compares the asset growth of companies selected for every quarter.
- The figure shows that % asset growth of Central bankcompany Inc. is generally lower than Bankunited Inc





# Total Asset value:

- The figure shows total asset values grouped by the states for every quarter.



Part 2:

Data Scraping for BHCPR Average Reports



# Process workflow:

Step 1 : Scrape WebPage to get all the reqd. PDF files (Peer 1)

Step 2 : Scraping the PDF file

Step 3 : Cleaning the data

Step 4 : Creating and saving the data into CSV file

Step 5 : Repeating Step 2-4 for all Peer 1 files

*Note : Used Python2.7*

# Step 1 : Scrape WebPage to get all the PDFs

## Modules used:

- Beautiful Soup

To parse HTML document

- Requests

To create a request and  
a response to the website

- Slate

To scrape PDFs

- Urllib

Fetch file from HTTPResponse

```
from bs4 import BeautifulSoup
import re
import requests
import slate
import csv
import urllib

# hitting the ffilec website with a "get" request
r = requests.get("https://www.ffi.ec.gov/nicpubweb/content/BKCPRPPT/BKCPR_Peer.htm")
# saving response
html_doc = r.text
# parsing the page using BeautifulSoup module
soup = BeautifulSoup(html_doc, 'html.parser')
# reading div tag of our interest
mydivs = soup.find_all("div", {"class": 'contentfull'})
# list to save link of all peer 2 files
list_of_links = []
# "for loop" to read all links
for tables in mydivs[0].findAll('table'):
    # reading all rows
    rows = tables.findAll('tr')
    # reading all cells with links
    links = rows[1].findAll('td')
    # for loop to save all links in a list
    for link in links:
        # "if condition" to check whether a link is empty
        if link.a != None:
            # saving links into a list
            list_of_links.append("https://www.ffi.ec.gov/nicpubweb/content/BKCPRPPT/" + link.a.get('href'))

# traversing the above list i.e. all PDF files to scrape into a CSV
for linked_file in list_of_links:
    # fetching file from the list
    response = urllib.urlopen(linked_file)
    # opening file
    pdf_file = open("document.pdf", "w")
```

## Step 2 : Scraping the PDF file

- Used “slate 0.5.2” package for scraping all PDFs.
- “slate” is a wrapper package for PDFMiner
- Fetched all PDF file links using “urllib” library

## Step 3 : Cleaning the Data

- Identified the rows & columns by recognizing the pattern of formatting in the PDF file
- Created a list of lists containing the data of PDF file
- Removed all the white spaces
- Removed all redundant lines
- Aligned the headers & data in respective positions

## Step 4 : Creating and saving the data into CSV

- Created File name as reqd. by extracting it from the online PDF file itself i.e. from  
[https://www.ffiec.gov/nicpubweb/content/BHCPRRPT/REPORTS/BHCPR\\_PEER/June2016/PeerGroup\\_1\\_June2016.pdf](https://www.ffiec.gov/nicpubweb/content/BHCPRRPT/REPORTS/BHCPR_PEER/June2016/PeerGroup_1_June2016.pdf)  
to “Peer1\_2016\_June.csv”
- Using “csv” module, made a CSV file for each quarter

# Output CSV

Less: Provision for Loan and Lease Losses	0.21	0.32	0.33	0.49	1.1
Plus: Realized G/L on HTM Sec	0	0	0	0	0
Plus: Realized G/L on AFS Sec	0.03	0.03	0.04	0.05	0.06
Plus: Other Tax Equiv Adjustments	0	0	0	0	0
Equals: Pretax Net Oper Inc (TE)	1.44	1.23	1.28	1.12	0.77
Less: Applicable Income Taxes (TE)	0.48	0.42	0.42	0.39	0.29
Less: Minority Interest	0	0	0.01	0.01	0
Equals: Net Operating Income	0.95	0.81	0.88	0.72	0.51
Plus: Net Extraordinary Items	0	0	0	0	0
Equals: Net Income	0.95	0.81	0.87	0.72	0.51
Memo: Net Income (Last Four Qtrs)	0.94	0.75	0.86	0.72	0.52
MARGIN ANALYSIS:					
Avg Earning Assets / Avg Assets	90.18	90.01	89.87	90.3	89.93
Avg Int-Bearing Funds / Avg Assets	68.17	69.01	68.65	70.86	72.85
Int Income (TE) / Avg Earning Assets	3.83	4.11	4.07	4.32	4.54
Int Expense / Avg Earning Assets	0.62	0.75	0.74	0.88	1.07
Net Int Inc (TE) / Avg Earn Assets	3.15	3.31	3.28	3.37	3.39
YIELD OR COST:					
Total Loans and Leases (TE)	4.74	5.09	5.06	5.34	5.44
Interest-Bearing Bank Balances	0.31	0.32	0.31	0.34	0.35
Fed Funds Sold & Reverse Repos	0.43	0.36	0.4	0.38	0.45
Trading Assets	1.29	1.4	1.25	1.43	1.37
Total Earning Assets	3.77	4.06	4.03	4.26	4.46

# Step 5 : Repeating Step 2-4 for all Peer 1 files

- Traverse through the list of links of PDF files fetched in Step 1
- For every PDF file, scrape, clean & save a new CSV



quest2.py

*File: Python code for Part 2*

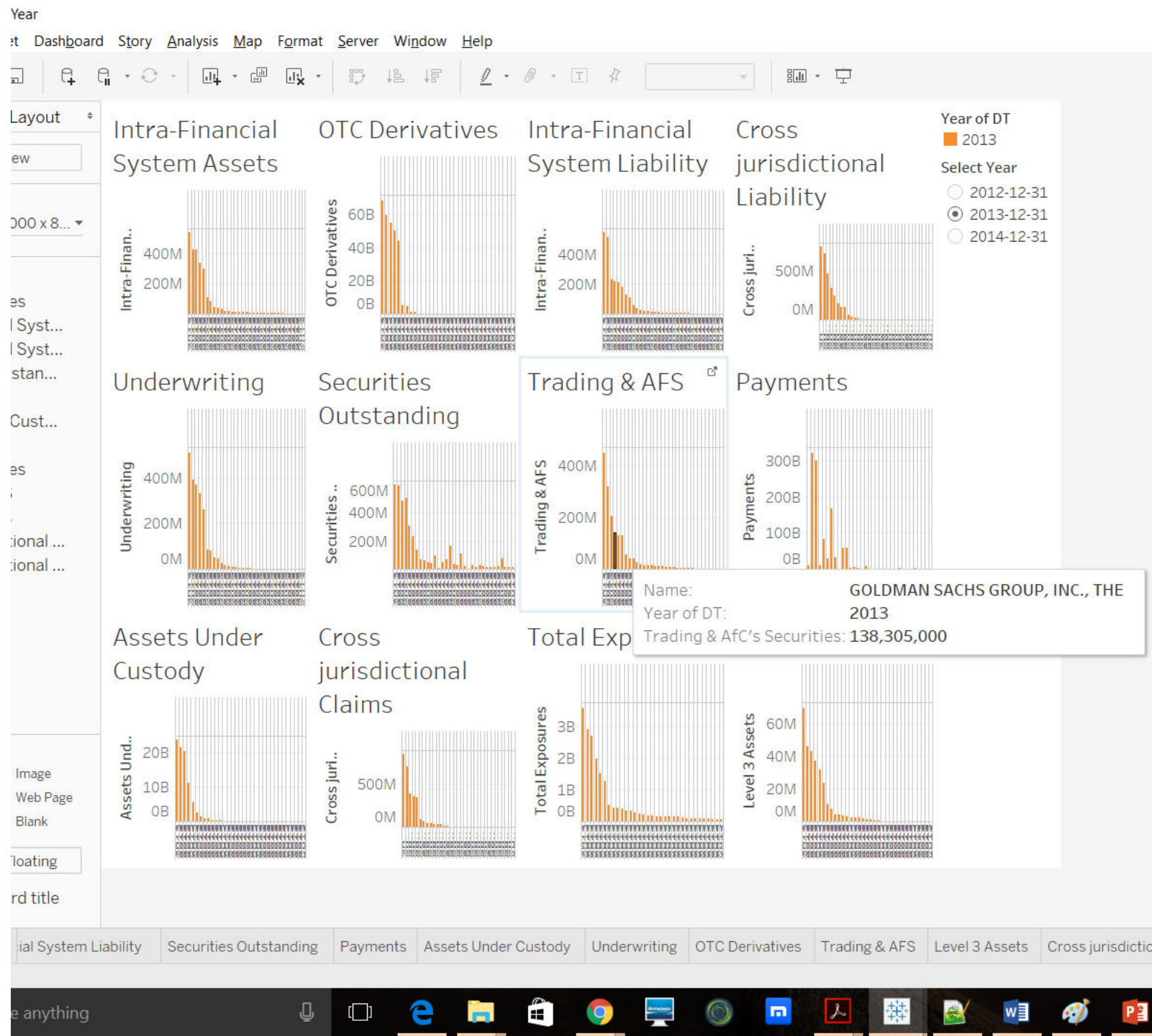
Part 3:

Analysis of Banking Organization Systemic Risk  
Reports for the years 2012, 2013 and 2014



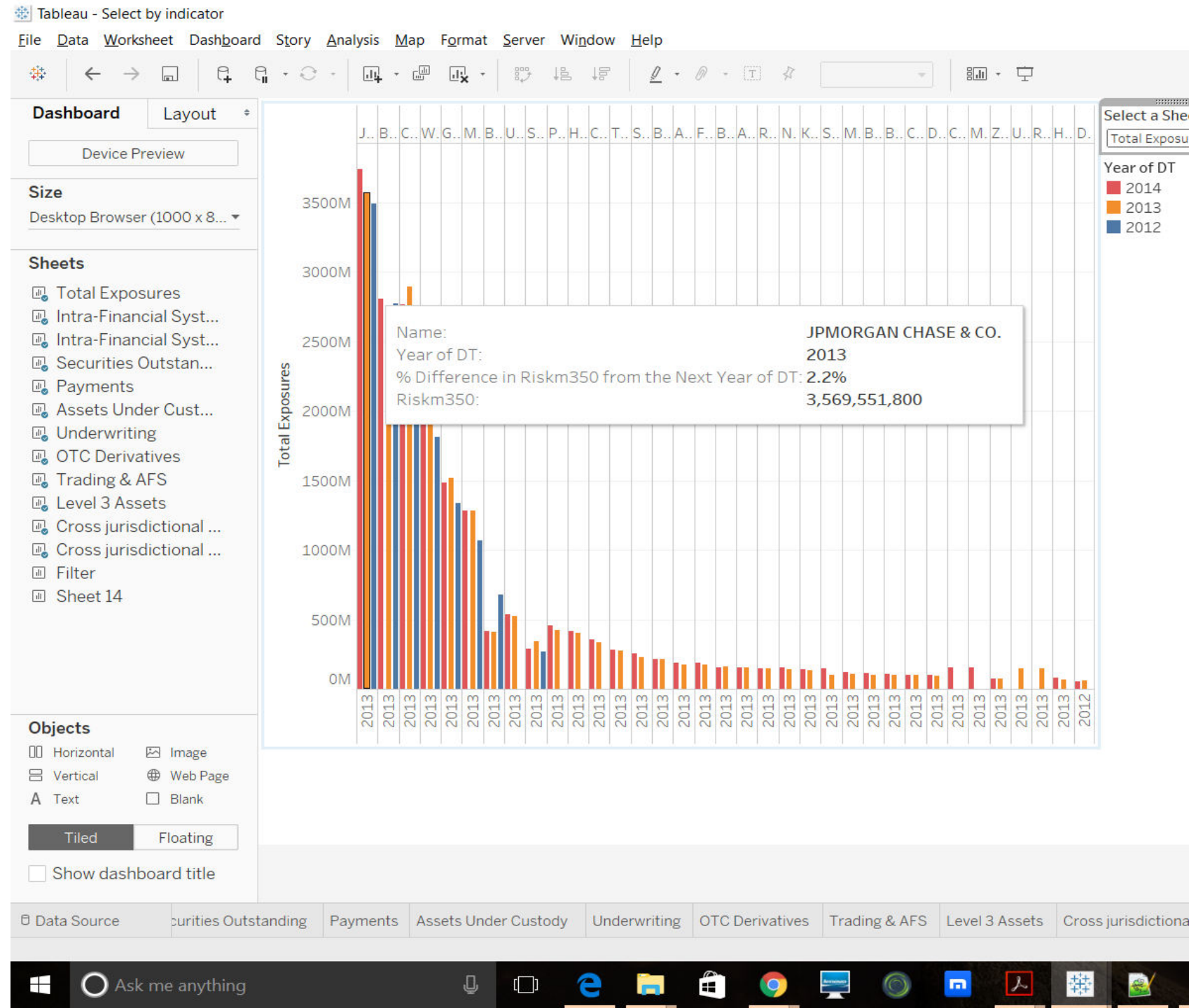
# Graphs Dashboard

- 12 bar charts for Measures for each Bank.
- Filtering based on Year.



# Time Series Dashboard

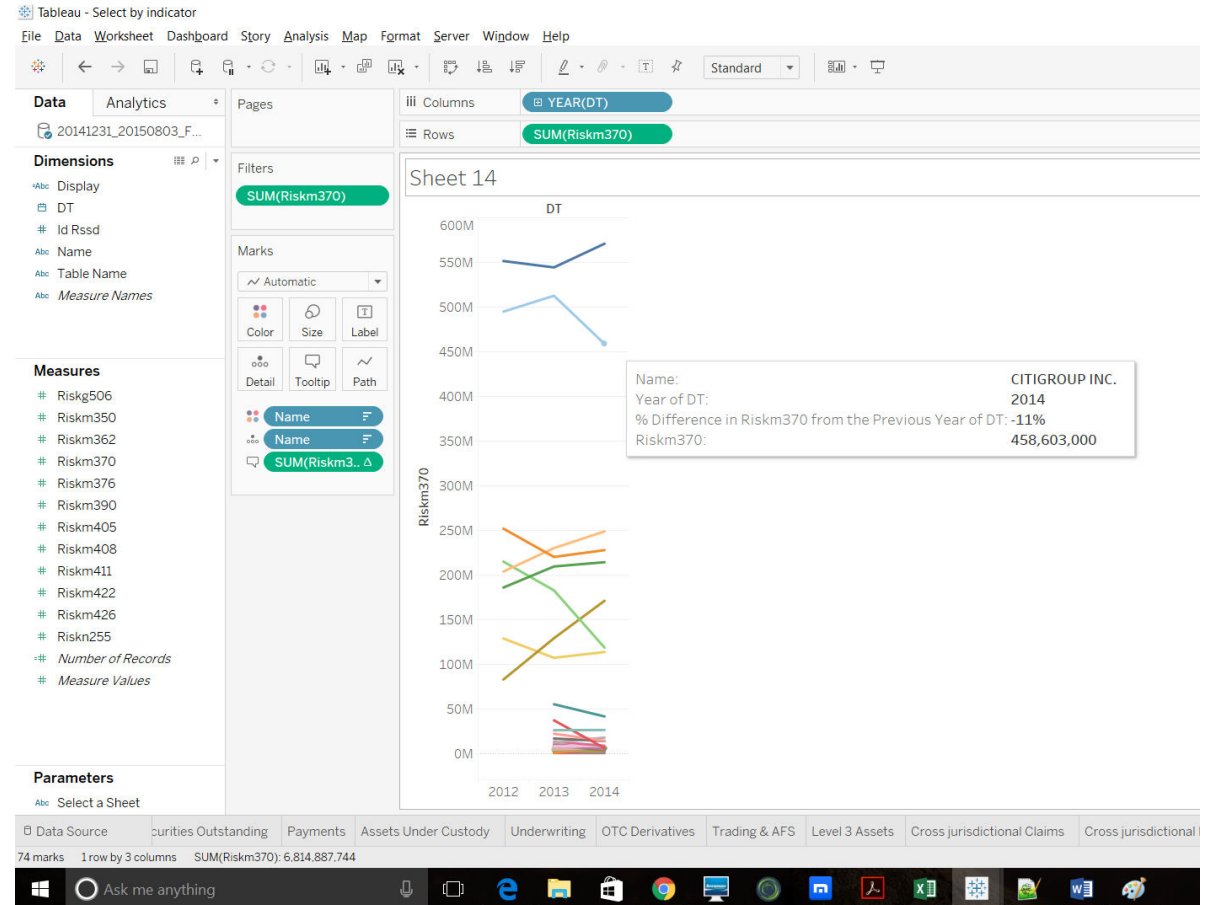
- Shows Year to Year comparison in form of Percentage Change values.
- Sheet Selector as Dashboard filter, to select Measure value over the 3 years.



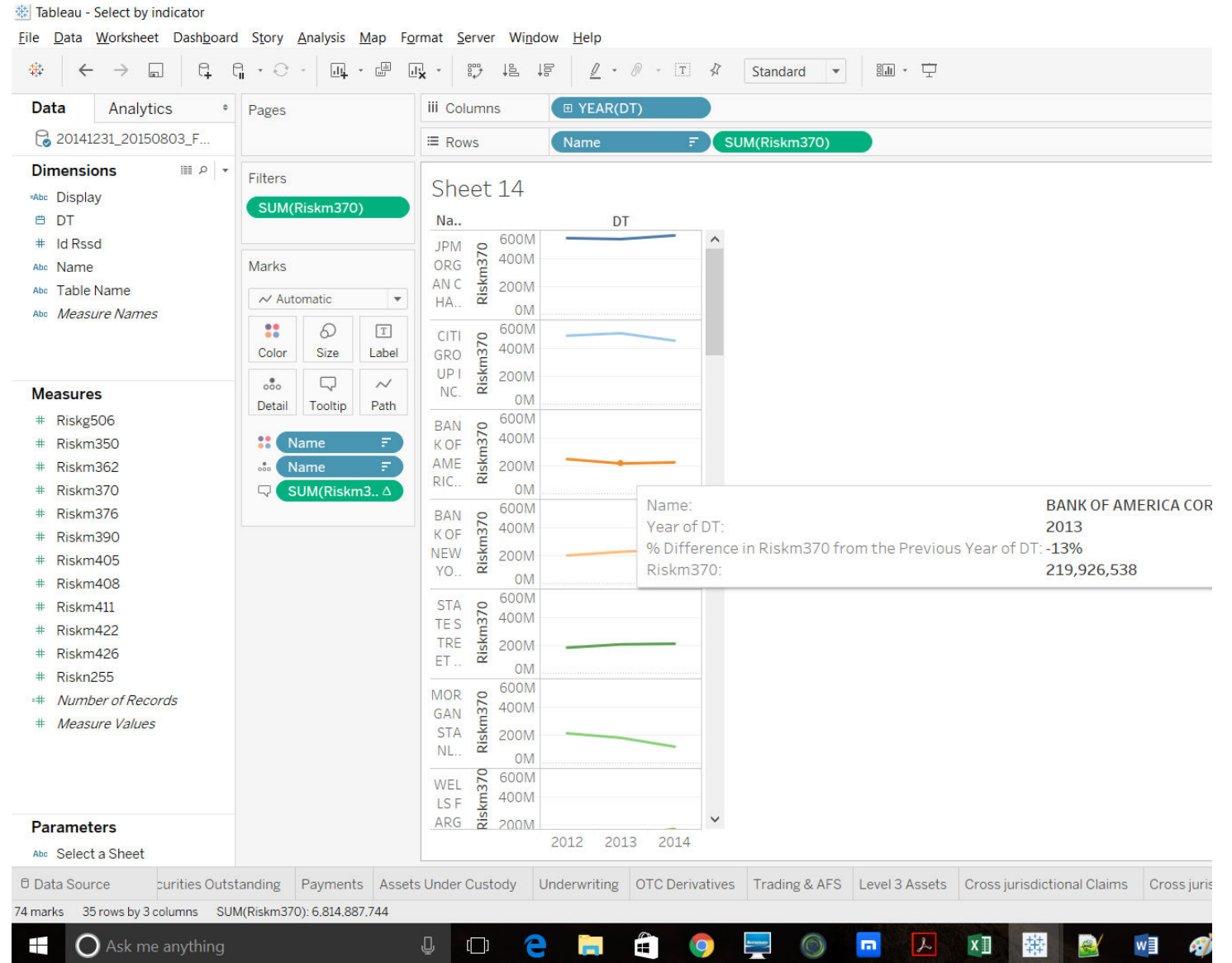
# Alternative Visualization techniques

## Line Graph

- Progressive visual of a measure for each bank over the years.
- Tooltip Indicates percentage change.
- Indicates the rise & fall of a measure for each bank.

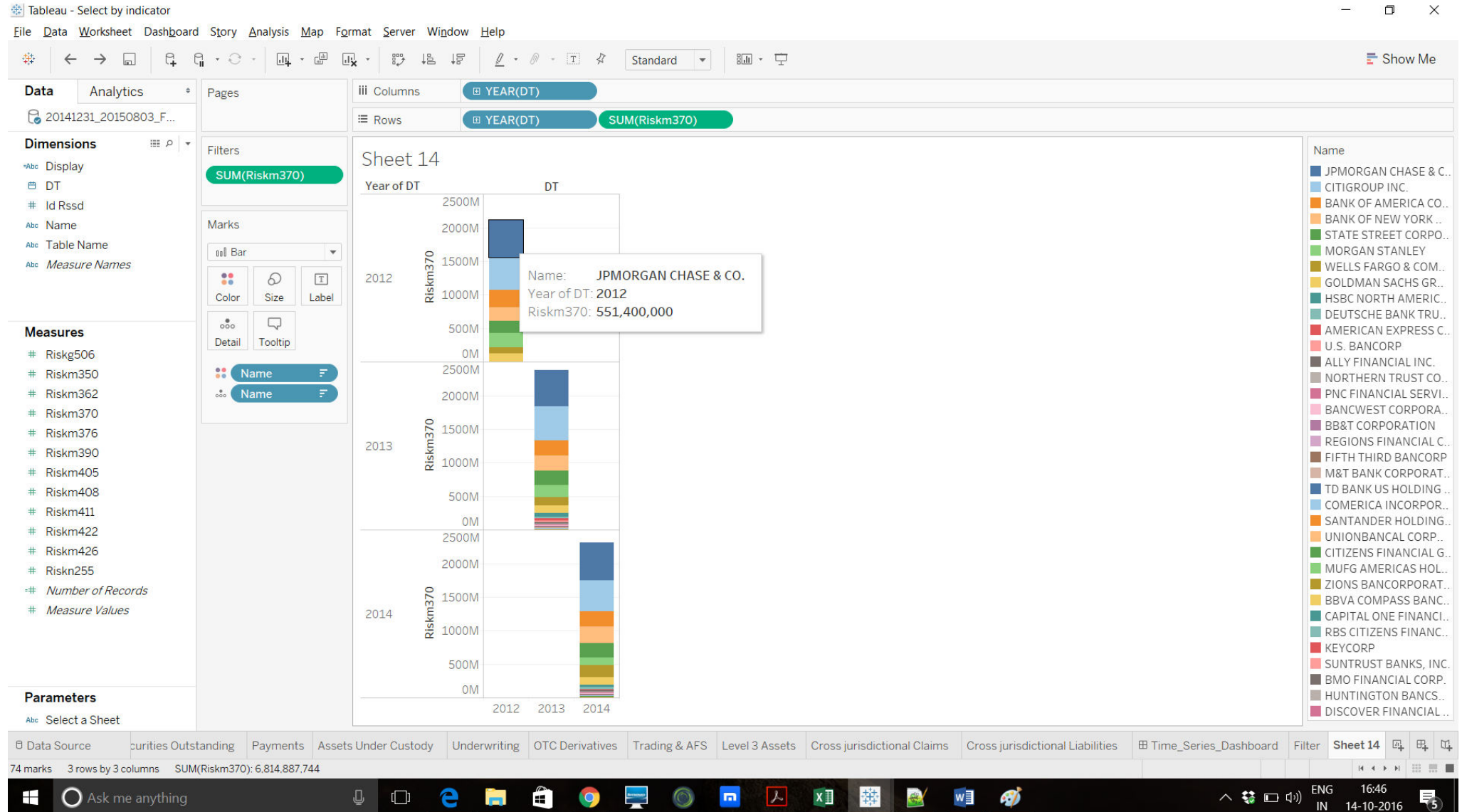


# Contd..

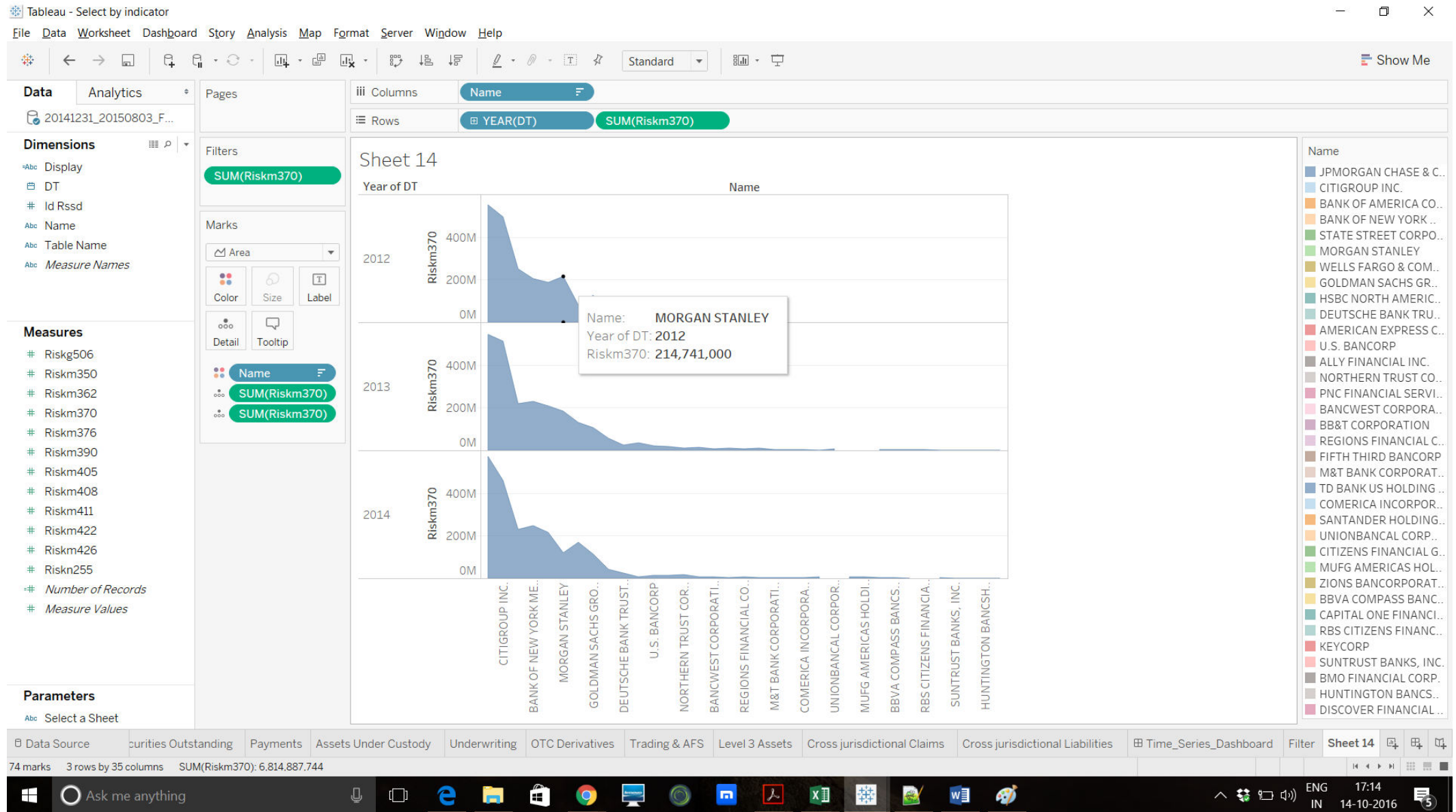




# Contd..

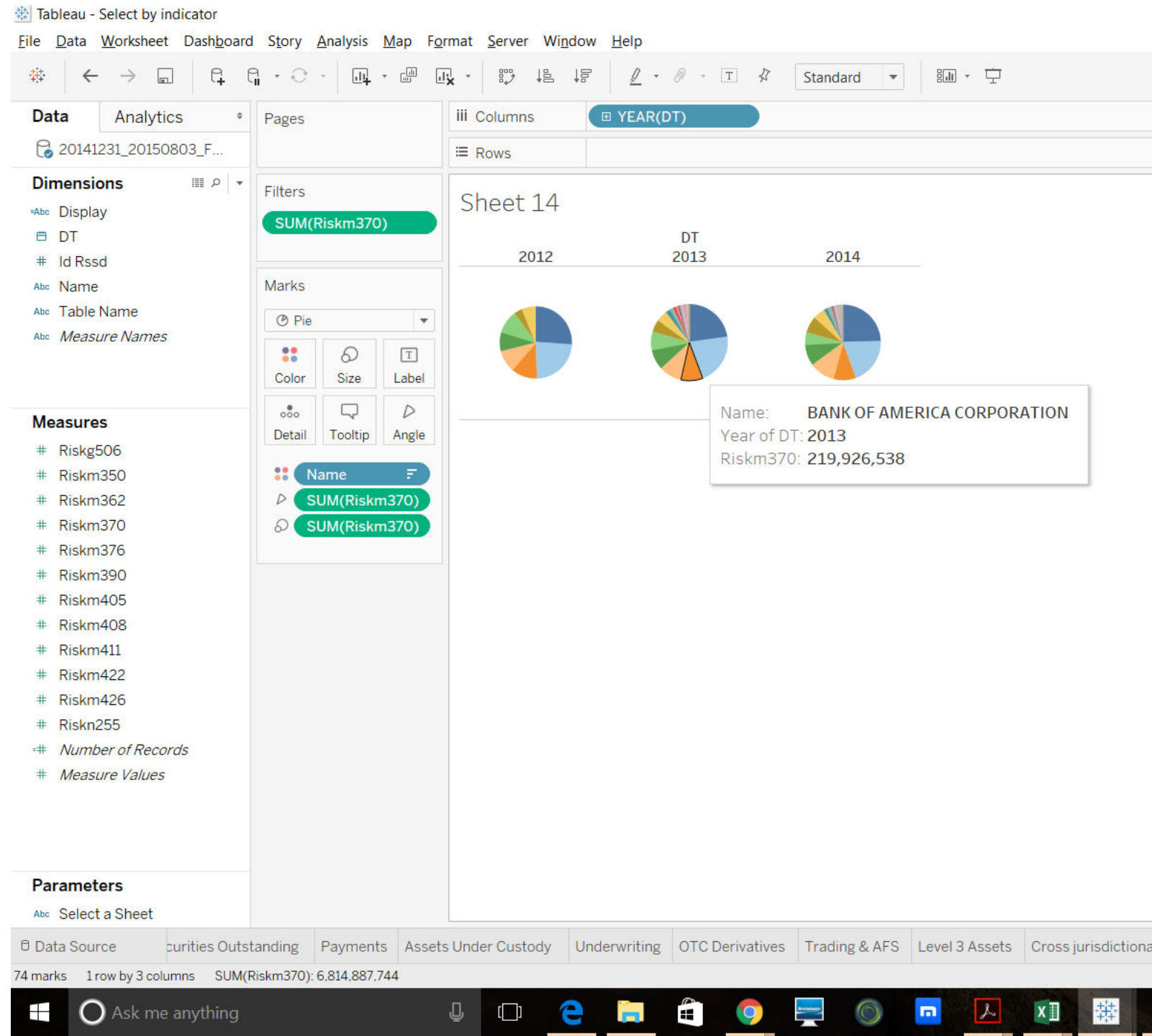


# Contd..



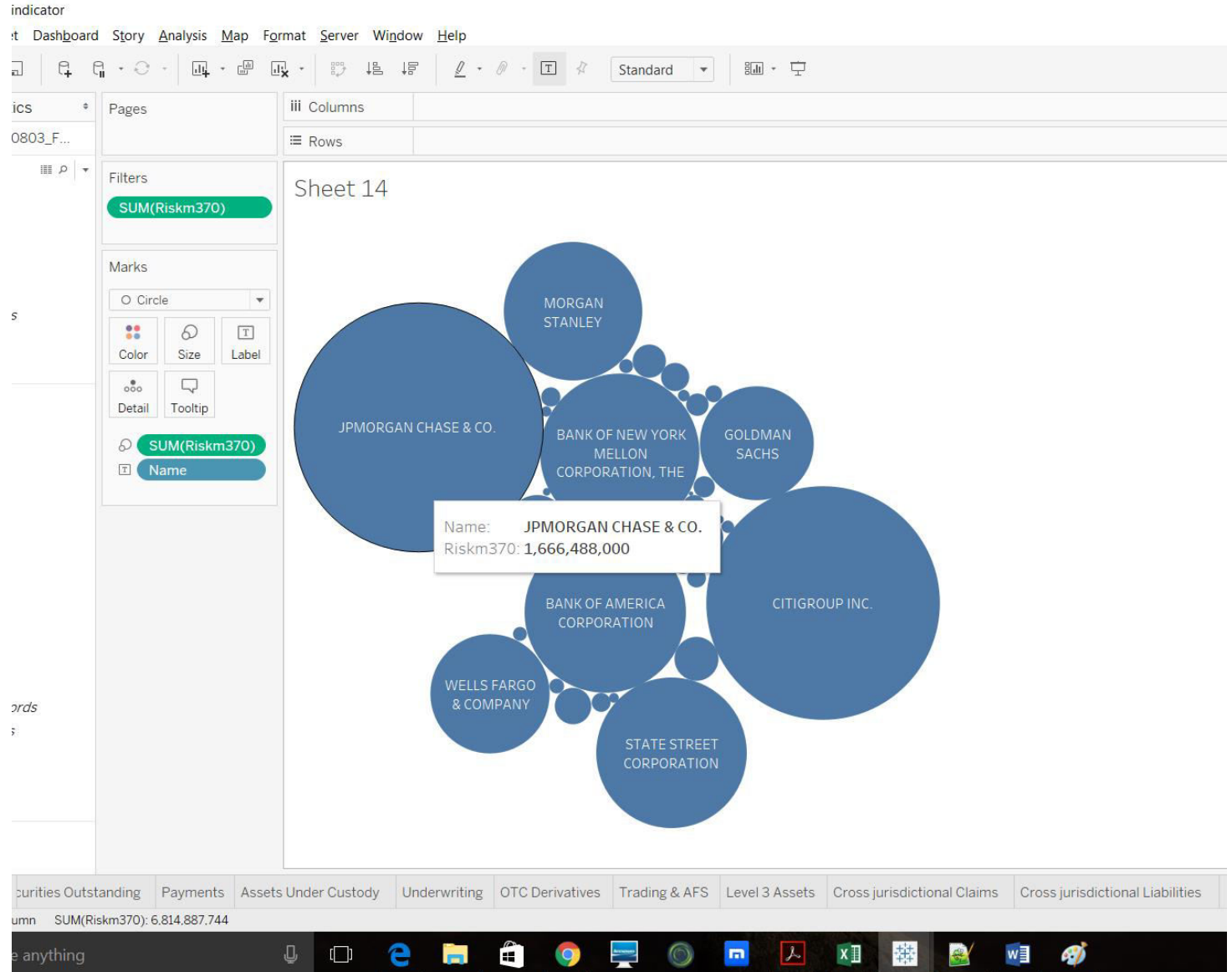
# Pie Chart

- Year Wise Distribution of individual measure for all the banks.
- Chart shows data for all the banks & its measure distribution for each year.



# Packed Bubbles

- Total Bank Distribution of a measure value computed for all years.
- Combined Distribution for all banks for all the years for an individual measure.





# Text Tables/Highlight Tables

- Tabular data for bank assets over 3 years.

The screenshot shows the Tableau Desktop interface. The top menu bar includes 'Board', 'Story', 'Analysis', 'Map', 'Format', 'Server', 'Window', and 'Help'. The toolbar contains various icons for data manipulation and visualization. The left sidebar shows the 'Pages' shelf, 'Filters' shelf with 'SUM(Risk370)', and 'Marks' shelf with 'Automatic', 'Color', 'Size', 'Text', 'Detail', and 'Tooltip'. The main view displays 'Sheet 14' with a table of bank assets over 3 years. The table has columns for 'Name', '2012', '2013', and '2014'. The data is sorted by '2012' in descending order. The bottom status bar shows 'SUM(Risk370): 6.814.887.744'.

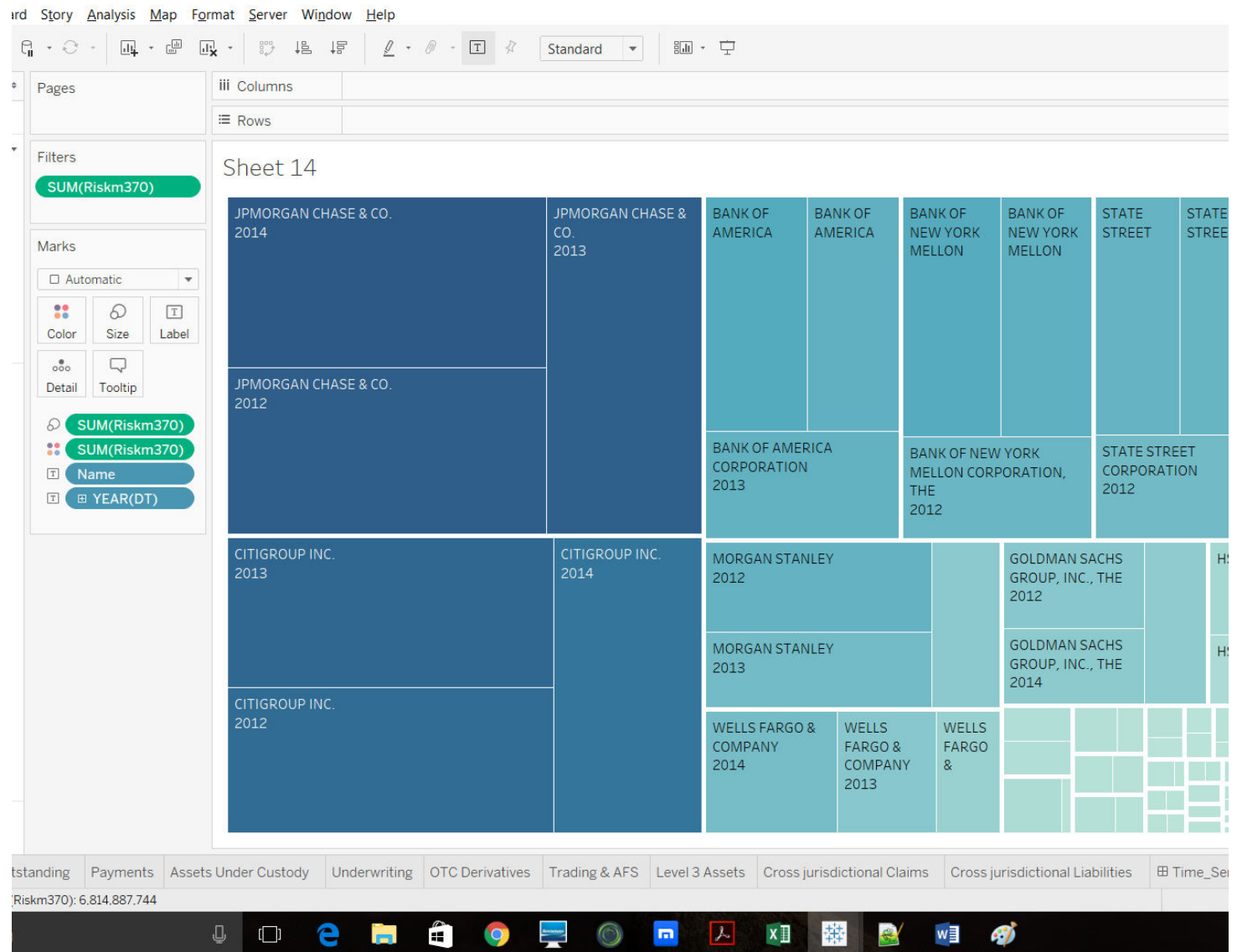
Name	2012	2013	2014
ALLY FINANCIAL INC.		16,534,000	13,796,000
AMERICAN EXPRESS..		36,921,000	5,794,000
BANCWEST CORPOR..		7,834,827	8,252,590
BANK OF AMERICA C..	251,635,813	219,926,538	227,579,653
BANK OF NEW YORK..	203,596,000	229,805,000	248,552,000
BB&T CORPORATION		10,316,636	4,558,381
BBVA COMPASS BA..		2,381,339	2,656,660
BMO FINANCIAL CO..		957,768	881,241
CAPITAL ONE FINAN..		2,466,871	1,682,001
CITIGROUP INC.	494,696,000	512,523,000	458,603,000
CITIZENS FINANCIAL..			5,785,512
COMERICA INCORPO..		4,241,765	4,756,909
DEUTSCHE BANK TR..		25,846,000	26,104,000
DISCOVER FINANCIA..		295,985	233,917
FIFTH THIRD BANCO..		8,942,647	3,231,856
GOLDMAN SACHS G..	128,579,000	106,942,000	113,405,000
HSBC NORTH AMERI..		54,924,874	41,279,038
HUNTINGTON BANC..		313,497	1,273,678
JPMORGAN CHASE ..	551,400,000	544,258,000	570,830,000
KEYCORP		724,429	2,139,610
M&T BANK CORPOR..		4,983,271	5,014,048
MORGAN STANLEY	214,741,000	182,491,089	118,359,866
MUFG AMERICAS H..			5,362,938
NORTHERN TRUST C..		12,215,420	17,567,798
PNC FINANCIAL SER..		13,135,000	8,342,000
RBS CITIZENS FINAN..		3,609,774	
REGIONS FINANCIAL..		6,692,841	6,831,840
SANTANDER HOLDI..		401,818	7,289,948
STATE STREET CORP..	185,648,900	209,263,087	214,038,869
SUNTRUST BANKS, I..		1,307,524	1,281,693
TD BANK US HOLDIN..		5,124,830	4,472,244
U.S. BANCORP		21,887,000	13,976,000

Outstanding Payments Assets Under Custody Underwriting OTC Derivatives Trading & AFS Level 3 Assets Cross jurisdictional Claims Cross jurisdictional L

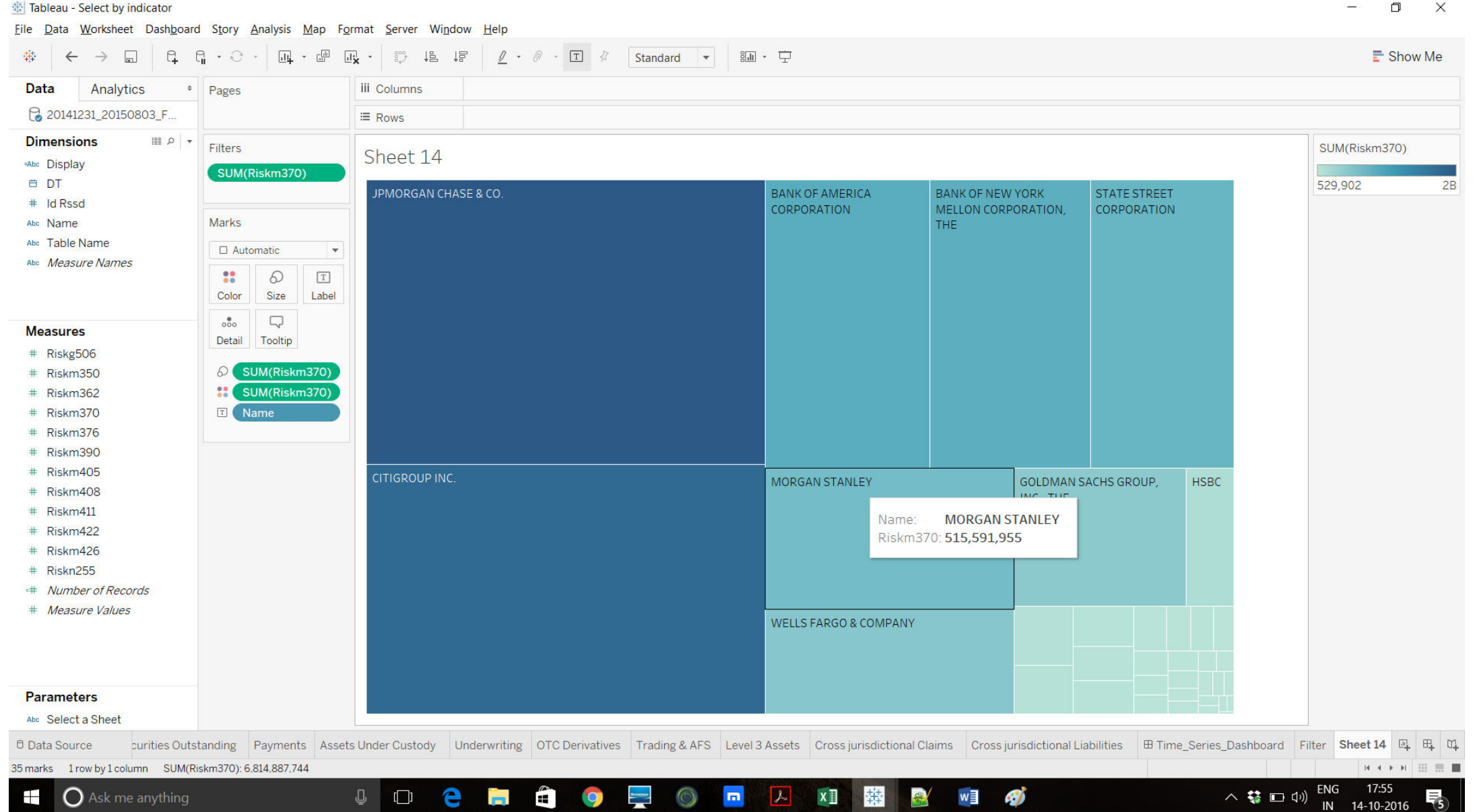
SUM(Risk370): 6.814.887.744

# Tree Maps

- Year wise measure distribution for banks.
- Coloring the rectangles by measure value

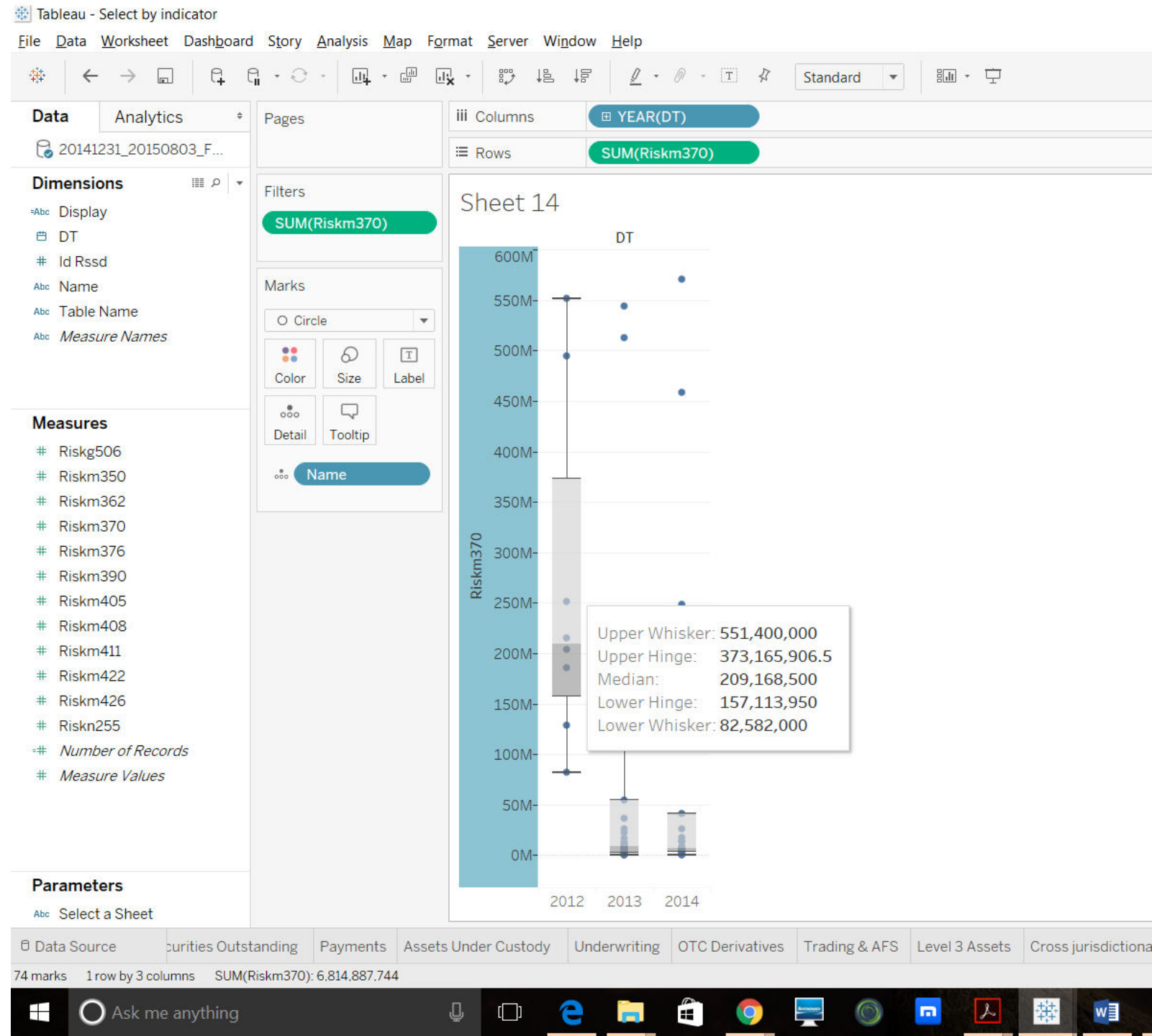


# Contd..



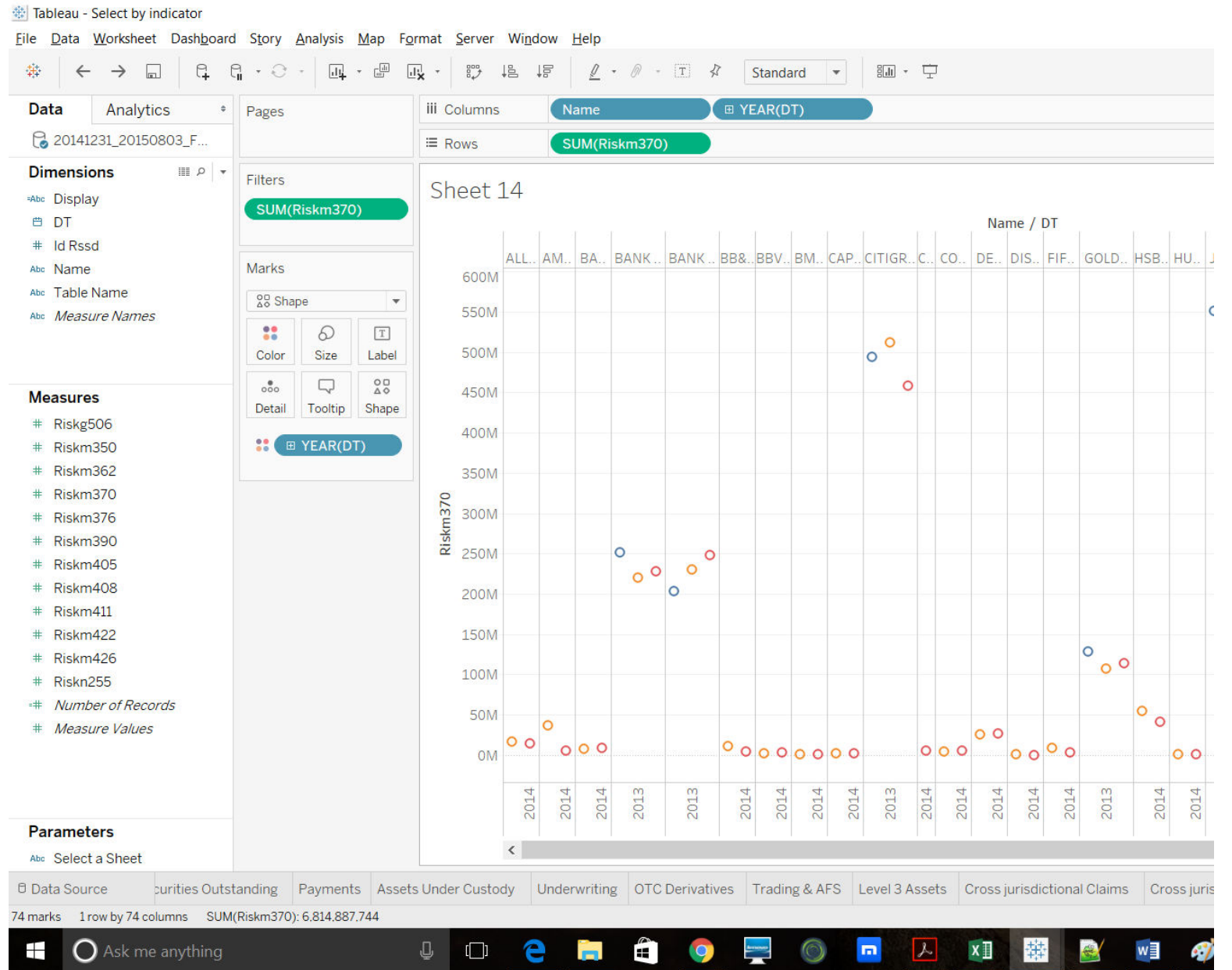
# Box Plots

- Shows Median of the measure value along with 1<sup>st</sup> & 3<sup>rd</sup> Quartiles.
- Shows the skewness of data.
- Shows details about the outliers



# Side-by-Side Circle Views

- Technique to accentuate data on scatter plots.
- Tells about relative concentration of data.



Thank you!