Web Exercise 03: Tableau Visualization Exercise

Due Date: September 29, 2016. 3:30pm on Blackboard.

Grade: 10 points.

1. You can use the Computer Lab's account to open **Tableau 10** or download the Tableau 10 student version into your own computer using the following instruction:

Each student should go to the following landing page to download Tableau and enter the key noted below. This key will activate enough licenses for your entire class for the duration of the course

- Landing Page: http://www.tableau.com/tft/activation
- Desktop Key: **TDO8-DC26-D950-00B9-3B1A**
- Instructions: Click on the link above and select Get Started. On the form, enter your university email address for "Business email"; and under "Organization", please input the name of your school.

For any students interested in using Tableau after this key has expired, you can request a one-year Desktop key through our Tableau for Students program (www.tableau.com/students). Each year if you are enrolled as a full-time student, you can request a new one-year key.

2. When you open Tableau 10.0, you'll see the main interface.

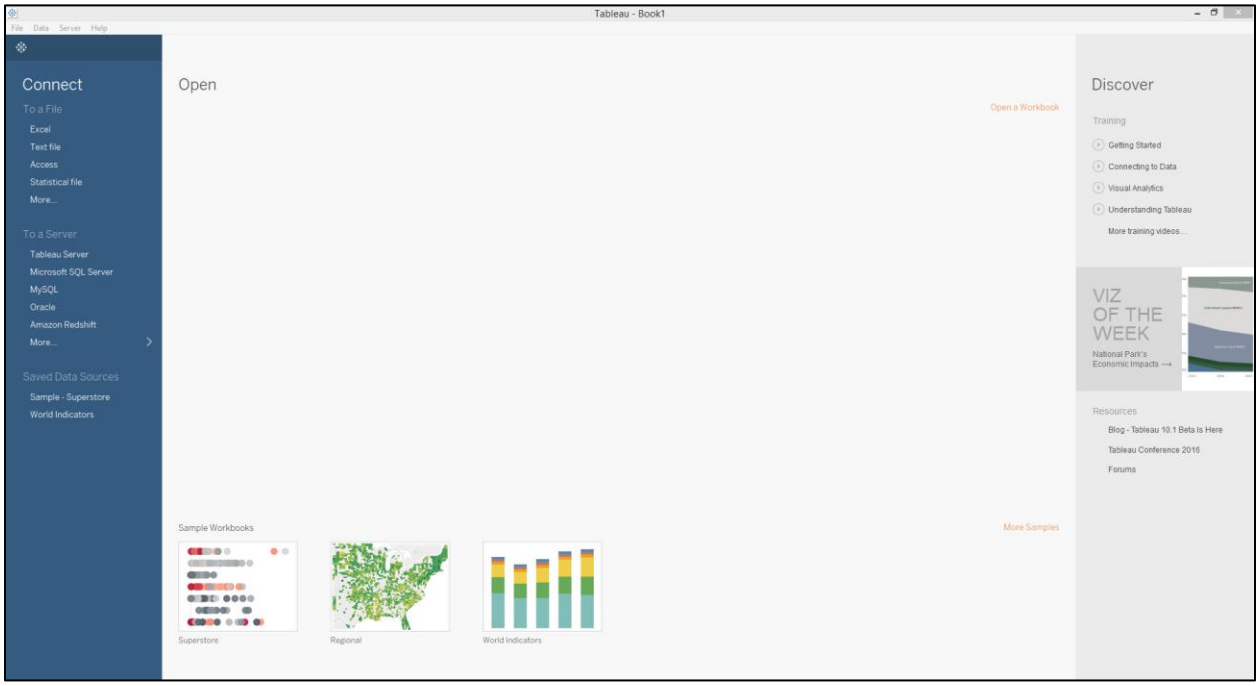


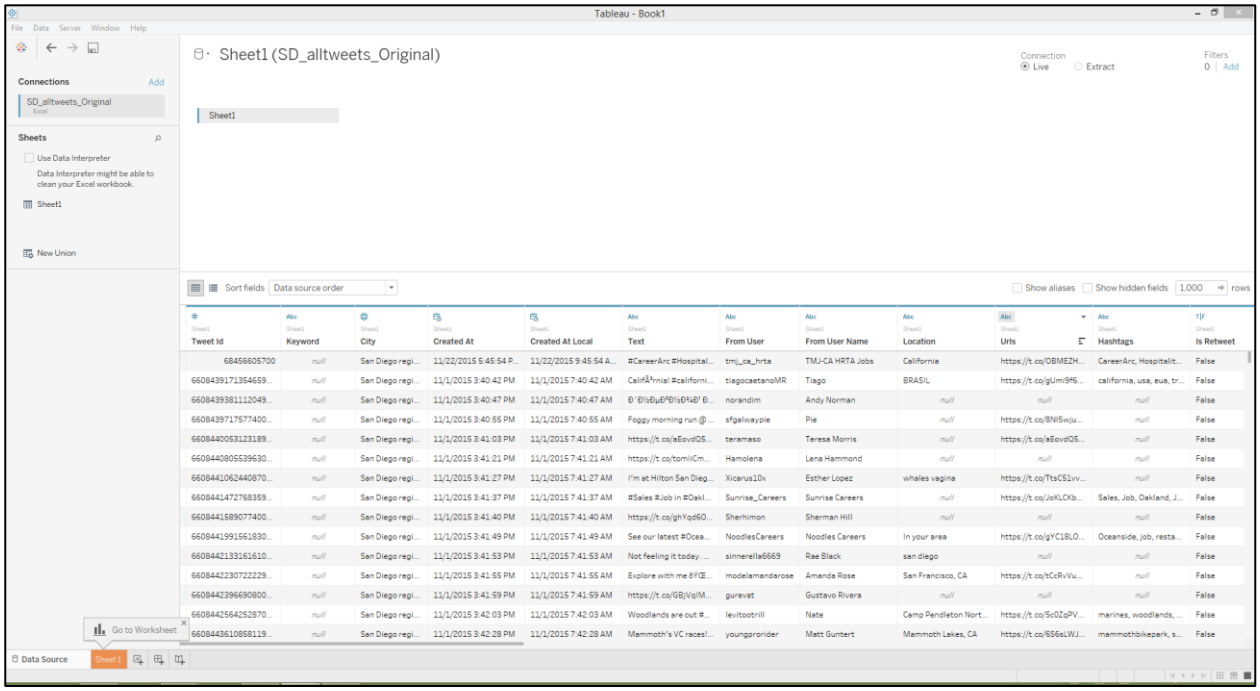
Tableau is a very powerful data visualization tool which can be used to connect many different files together and even linked to different data servers. Before we work on the actual data, let's watch a few introduction video from Tableau. In the [Upper Right Corner], click on "Getting Started" and open the Web page in your Browser. Watch all three videos in this page (Getting Started, The Tableau Interface, and Distributing and Publishing). (Total: 33 mins).

There are other great introduction videos available online. Feel free to watch them when you have time.

After watching the Getting Started video, we will start to work on the real world data. The dataset you're going to use is **SD_alltweets_Original** in excel format. It is a large set of geo-tagged twitter data downloaded from the **GeoViewer** web application (http://vision.sdsu.edu/ec2/geoviewer/sanDiego#) for one month in November, 2015. You can download the dataset from the Class **Google Drive Shared folder**: **/Class-Big-Data-Science-Shared/Web-Exercises.**

Copy the SD_alltweets_Original.xlsx from the Google Drive to your local drive (or the Z: drive in your lab machine).

3. Click on **[File] – [Open]** and then select the "SD_alltweets_Original" excel file. Tableau will open the file and you'll see the data. Click on "**Update Now**" if needed. It may take 2 mins to open the file.



Click on the orange panel **[Sheet 1] (at the bottom of your window).** You'll see all the variables and features from the original dataset on the left which are now called **Dimensions** and **Measures**.

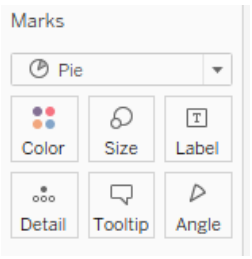
In Tableau, **Dimensions** are the "Fields" contain ***discrete categorical information*** such as ID, Location, Language, etc. **Measures** are the fields contain ***quantitative, numerical information***, such as "Friends Counts, Retweet Counts, Latitude, and Longitude.

In order to wisely choose which visualization method you want to use, Tableau has a Panel Tool function called **[Show Me] (right side of the window).** By moving your mouse over each of the graphs, you'll see what kind of variables you need to have and how many dimensions and measures you need to use to create for specific visualization graphs.

You can hold the [Ctrl] and click any combination of the dimensions and measures, **[Show Me]** will recommend the appropriate visualization method for you.

Before you start the visualization, the best way is to do the backward selection and determine which variables you want use for the visualization. The variables that this exercise will be use are**: CREATED_AT_LOCAL, FROM_USER, GEO (**split into **Longitude** and **Latitude), URLS, LANGUAGE, HASHTAGS, and SOURCE**.

Sheet 1:

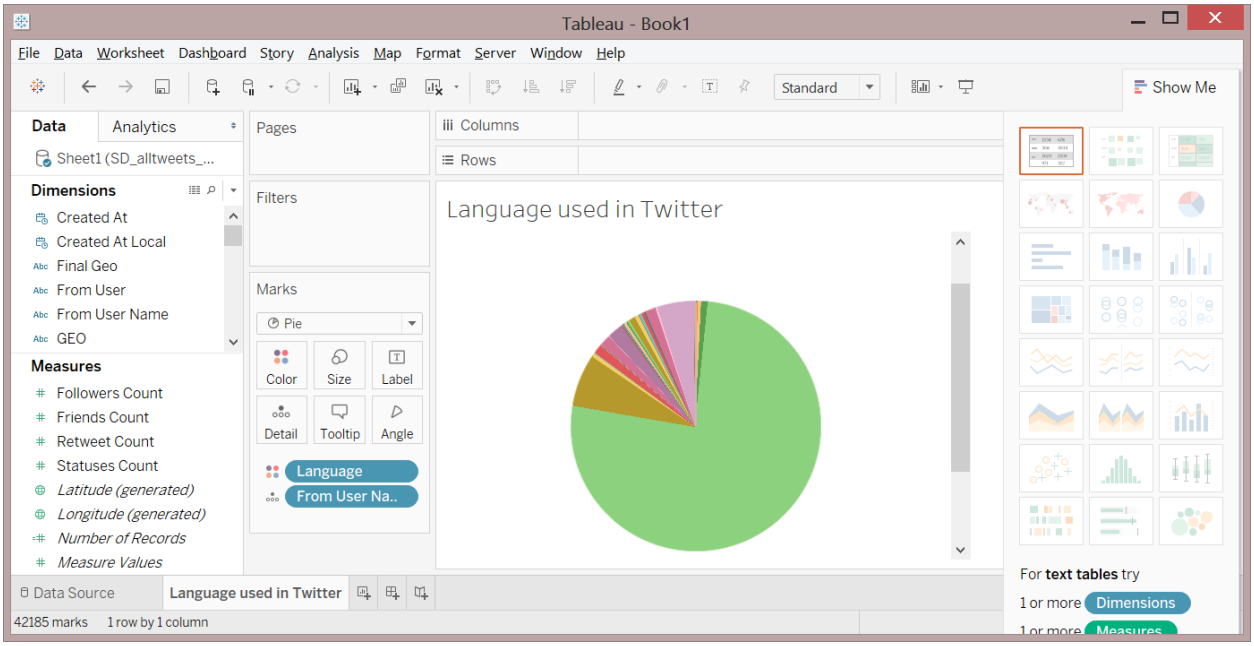


To begin with the first sheet, we want to know the language that the twitter users are using. Pie chart would be the most straightforward way to show.

1. Go the **[Marks]** Panel and click on the drop-down menu (the Automatic Bar), then choose **[Pie].**

2. Drag the variable "**Language"** from the **Dimensions** on the left panel to **[Color] cell**. (It will take 30 seconds to load the whole data). A Pie Chart will appear. Hold **[Ctrl], [Shift]** and **[b]** to enlarge the chart. Then drag the variable "**From User Name"** on top of the **[Detail] cell.** The software may pop up a window saying that "the field being added may contain too many members". Choose "Add all members". Since we want to know the actual distribution of the variable. Therefore, the detail pie chart will display the language proportional based on different user names. On the right panel, each color represent one language. Take a look at this Pie Chart. Which language is the most popular language in Twitter users in San Diego? Which one is the second most popular language in San Diego? You can check the Twitter language code in this website: https://dev.twitter.com/web/overview/languages

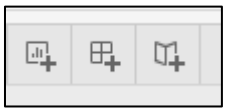
3. (Double click on the <sheet1> text (at the bottom of window) and rename this sheet 1 as "**Language used in Twitter"**. You should see the title of this sheet (on the top of pie chart) have changed to "Language used in Twitter". By moving te mouse over the chart, you can see each user's name and the language that each user use.



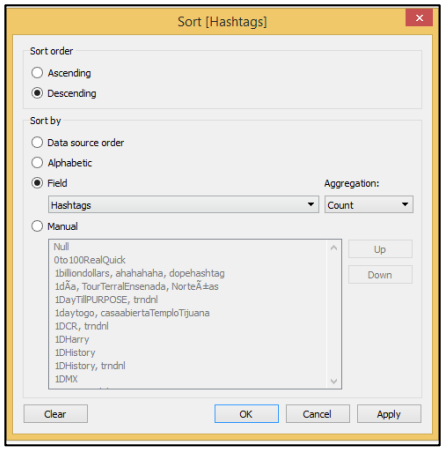
Sheet 2:

The second sheet, we want to check what the popular hashtags in our collected twitter data are.

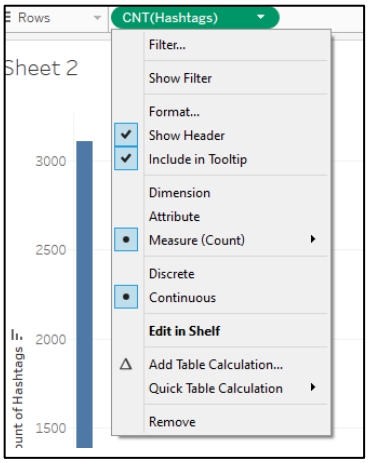
1. First, click on the add sheet button  at the bottom of the screen. (Just like in Excel). The first one is create a new sheet . The second one is create dashboard and the last one is create storyline. We'll talk about dashboard and storyline function later.

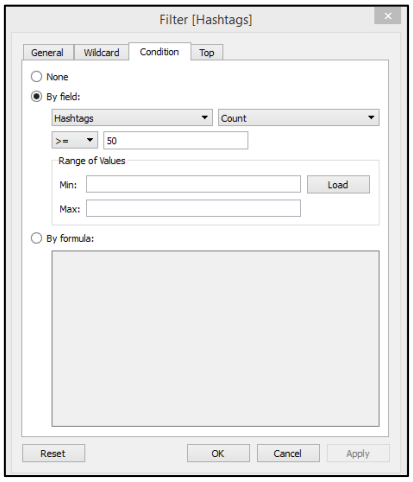


2. After creating a new "Sheet 2", drag the variable Hashtags to the [Columns] (at the top of window). Click on the bar of the "Hashtags" variable to open the drop-down menu, and select "**sort**" to open the Short window. Choose [**Descending]**, Sort by **[Field]** and choose "**Hashtags**", Aggregation by "**Count**" then click **OK**. Make sure the variable Hashtags is showed as Dimension with the check-box items for [Show Header] and [Include in Tootip].



3. Drag the variable **Hashtags** again to [Rows]. Different from the Column option, you should use "count" - CNT (Hashtags). To do so, click on the Hashtag variable. In the drop-down menu, check [Measure] – [Count]. And check [Continuous].



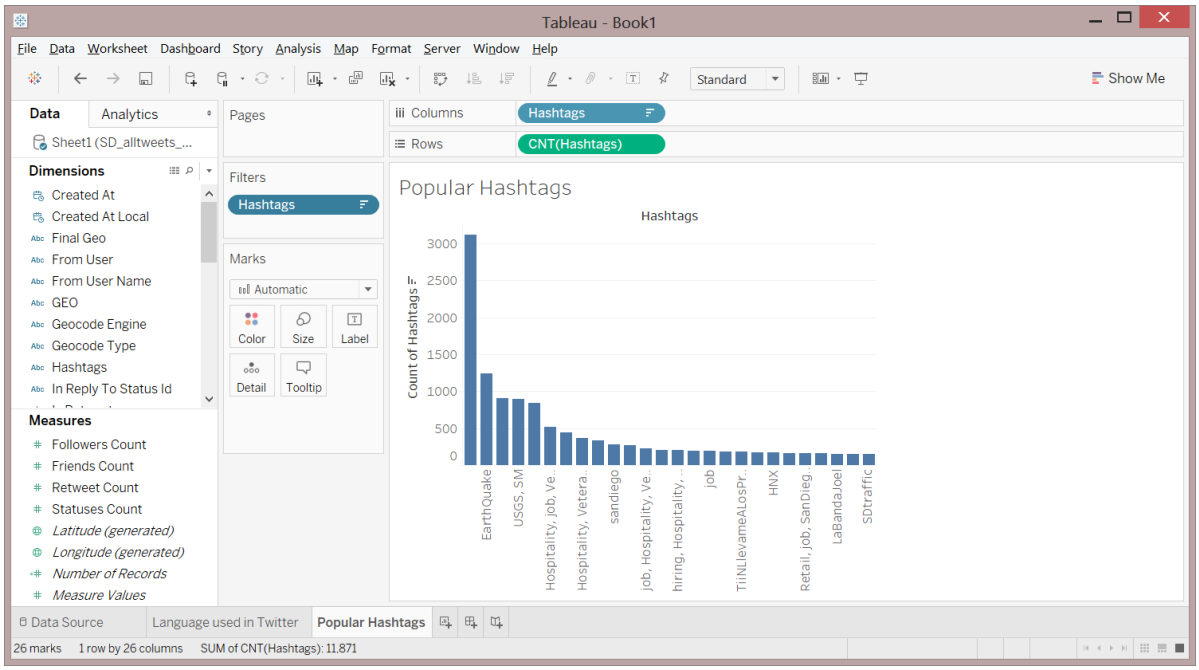


Since there are too many hashtags and we only need to know the most popular ones. Drag the variable Hashtags from "**Dimensions**" to the **[Filter] BOX** (above the Marks Panel). (If needed, click on the variable and select **[Edit Filter].)**

In the Filter box, select **[Condition]** in the tab selection – **[By field] –> [Hashtags] –> [Count] –> [>=] –> [150]. Then press OK.**

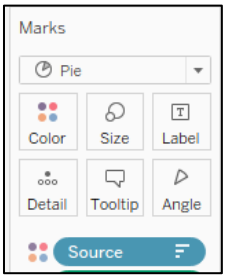
Now you can see the popular hashtags used over by 150 tweets.

4. Rename the sheet as "Popular Hashtags". And then create another sheet in your work book.

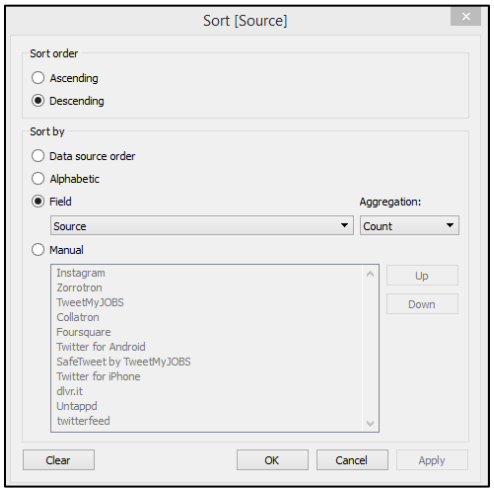


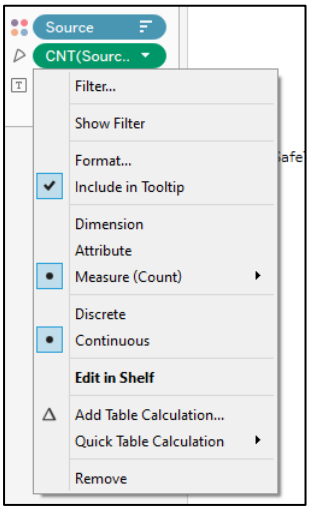
Sheet 3:

The third sheet we want to know what the major sources of twitter are. To visualize the proportion of each source, again, pie chart would be the most straightforward method.



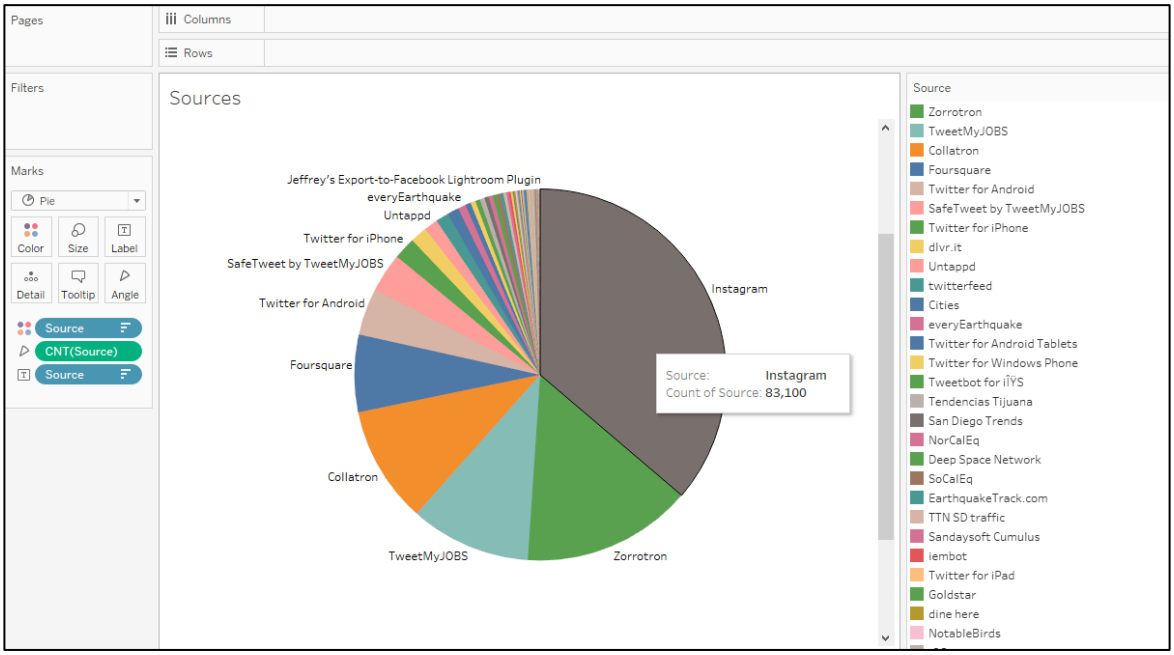
1. Create the "Sheet 3" in Tableau by click on the add sheet button.
2. Go the **[Marks]** and click on the bar, choose **[Pie].**
3. Drag the variable Source from the **Dimensions** list on the left panel to **[Color]**. A Pie Chart will appear. Hold **[Ctrl], [Shift]** and **[b]** to enlarge the chart.
4. Click on the variable and then click **[Sort].** Check **[descending]** and Sort by **Field – Source**. Aggregate by **Count**. Press [OK].





4. Then drag the variable Source on top of the **[Angle] cell.** The software may pop up a window saying that "the field being added may contain too many members". Choose "Add all members" since we want to know the actual distribution of the variable. Click on the variable and choose **[Measure] – [Count].**

5. The last step is to mark the pie chart. To do so, drag the variable **Source** again and put it on top of **[Label].** Repeat the procedure as in step 4, sort the variable with descending and sort by **Source** and aggregate by **Count**. Therefore, when you move the mouse on top the chart, you'll see the name of the source and the count of the tweets from that source. Rename the sheet as "**Sources**".

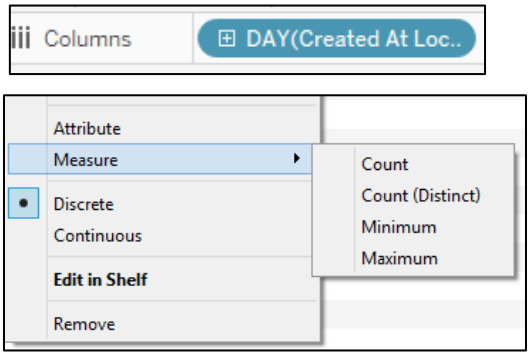


Sheet 4:

For the fourth sheet, we want to visualize the number of tweets for each day of the month and maybe further look into the temporal pattern in each day.

1. First, create a new sheet (Sheet 4).

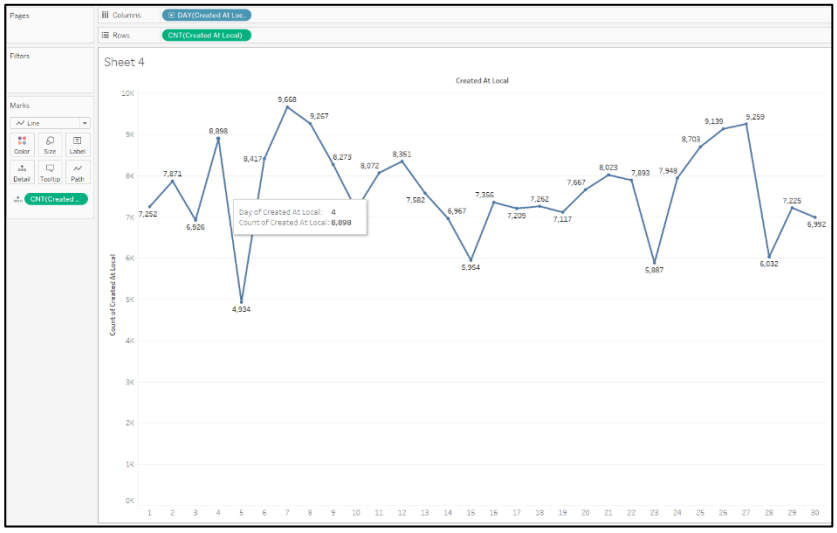
2. Drag the variable Created At Local from the left panel to the **Columns.** The default set of the variable is YEAR. Click on the variable and check on **DAY**.
3. Drag the variable Create At Local from the left panel again and put it to the **Rows**. Click on the variable and redo the process as in step 2, choose DAY. And then check **[Meansure] – [Count].**



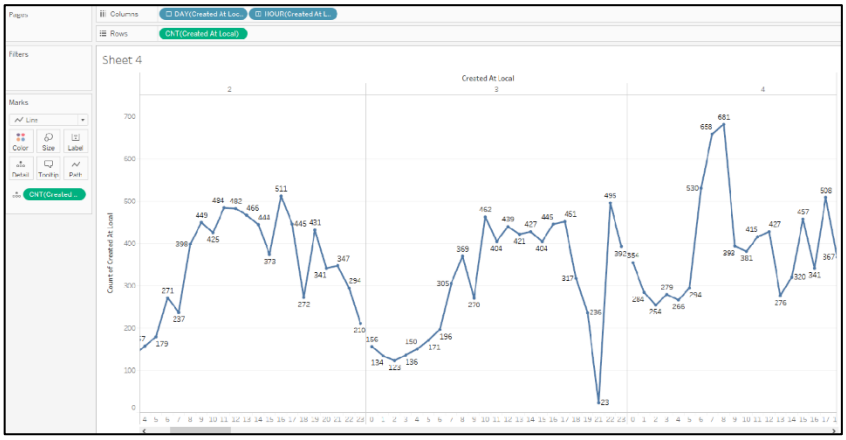
4. To further detial the graph, you can mark each day with its actual number. To do so, drag the variable Creat At Local from the left panel again and this time put it on the **[Lables].** Then click on the variable, check **DAY** and check **[Measure] – [Count].** The graph is acutomatically formed as Line Graphy. You can click on the bar below **[Marks]** to change it into "Bar" if you want. Here we'll use the "Line" graph.

5. Now that you'll see a line graph and when you move your mouse on it you can see the detail (day of create at local and count of tweets). The X-Axis is the days of the Month (Nov 2015) and the Y-Axis is the number of total tweets recorded on that day.

6. If you want to look into the detail hourly patterns in each day's tweets count, you can easily do so by click on the little **[+]** button on the left corner on the **X-Axis**. Therefore, the monthly chart will be divided into day chart. And the X-Axis will become each hour of the day (In the Columns you will see an added variable "**HOUR**".



7. Since we only need a general monthly pattern for now. Go ahead and "Remove" the HOUR variable from the Columns (by selecting remove in the dropdown list). Then rename the sheet as "Nov 2015 Tweets Daily Count".



Sheet 5

1. Create a new "Sheet 5" first.
2. The next visualization method we want to do is to display the geo-tagged tweets on the map. The latitudes and longitudes coordinates of tweets are stored in the "**GEO**" field (This is the old Twitter API version, the new version will be saved in the "coordinates" field.) [Right click] at the "**GEO**" in the Dimensions list on the left panel. Then select "**Describe**". You will see the general information about the "GEO" field and the examples (Domain) by click on "Load". In the example, you can see that the first column should be longitude (including values over 90), and the second column should be latitude. [Close] the Describe window.

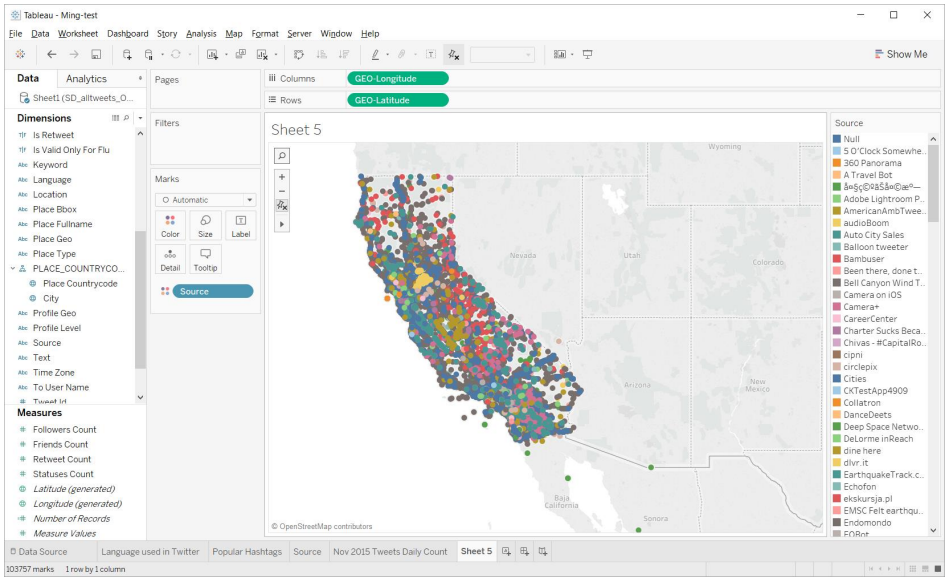
3. Now we need to split the GEO field into two fields (Lat/Long). To do so, right click on the **GEO** field, then select "Transform" → Split. You will see the two new fields (Dimensions): **GEO-Split 1** and **GEO-Split 2**.

4. Rename the GEO-Split 1 to **Geo-Longitude**, and GEO-Split 2 to "**GEO-Latitude**". (Right click and select "rename").
5. Click on the "Abc" icon at the beginning of GEO-Latitude, → select "Number (decimal)" , then select "Geographic Role → Latitude"
6. Click on the "Abc" icon at the beginning of GEO-Longitude, → select "Number (decimal)" then select "Geographic Role → Longitude".

7. As mentioned in the very beginning of the tutorial, Tableau has a function called **[Show Me].** It can recommend the best visualization method with the variables you select. Hold **[Ctrl]** and click on GEO-Latitude and GEO-Longitude, then **[Show Me**] will provide you the best method which in this case **is [Symbol Map]**. Click on it and you'll see the dots on the map.

8. To further look into the data, you can used source as filter to see the users' behavior on tweets. Drag variable Source from lest panel and put it on top of **[Color].** A window will pop up saying that "the field being added may contain too much members."

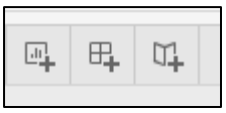
9. Click [Add all members]. Zoom in the map to California. You'll see the geo-tagged tweets with their sources on the map. Rename this sheet as "Sources on Map".



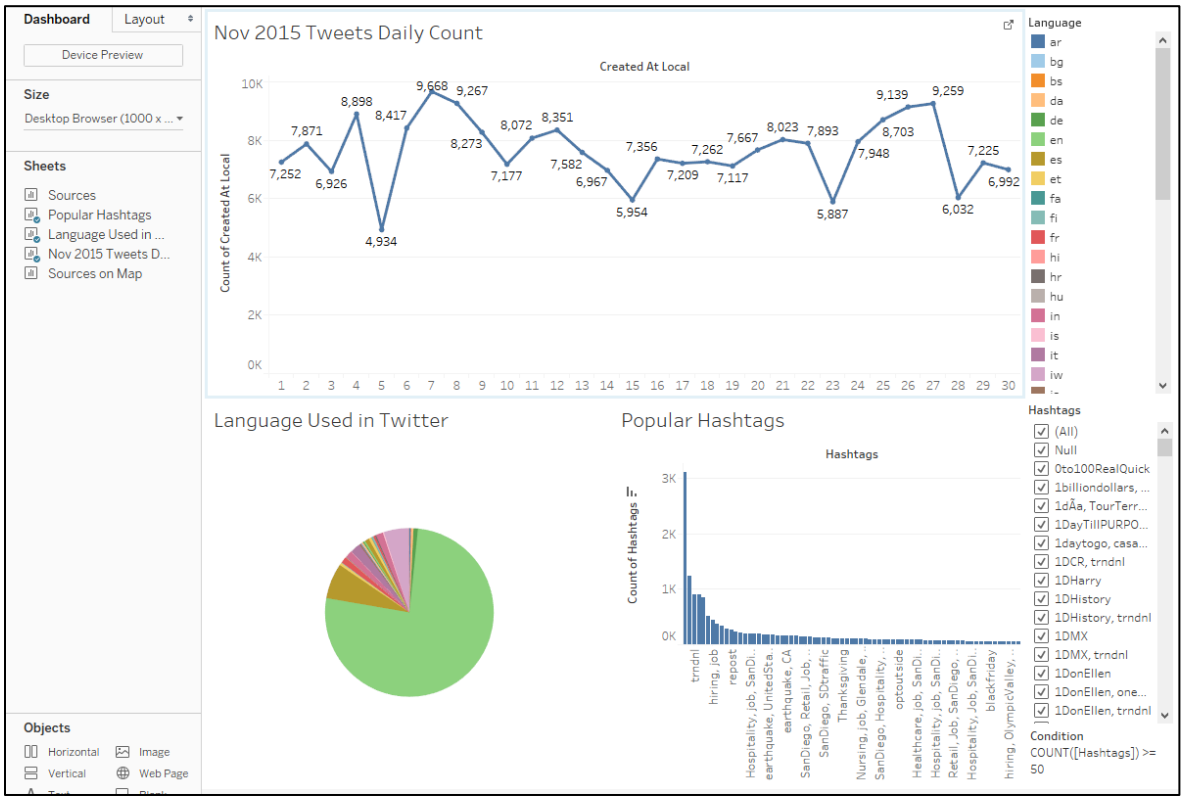
Dashboard 1:

Tableau allow you to combine various “sheets” together to create a dashboard. With the 5 sheets you just created, you can then form them into a dashboard for comprehensive visualization.

1. Click on the bottom **[Create New Dashboard] icon.** (the second one)

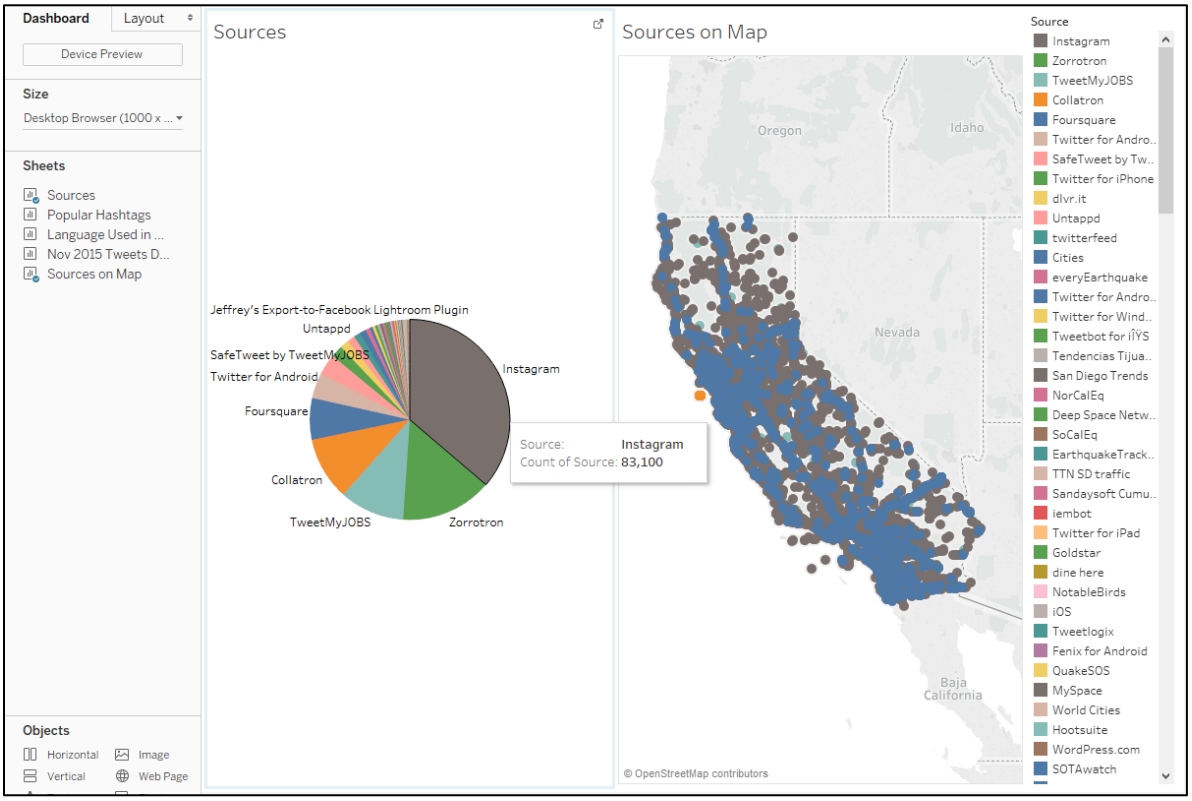


2. Choose Desktop Browser in the Size. You’ll see all the sheets you just created on the left panel. Drag Nov 2015 Tweets Daily Count, Language used in Twitter and Popular Hashtags to the dashboard. You can adjust the size of each sheet by clicking on the triangle button on the right upper corner of each sheet. **[Fit] – [Entire View]**. Rename the Dashboard as “Basic Facts About the Data”.



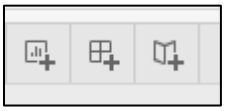
Dashboard 2:

Repeat the same procedure as in the first dashboard, drag the sheet Sources and Sources on Map to the interface and then adjust the size. Rename the dashboard as “Twitter_Source”.



Storyline

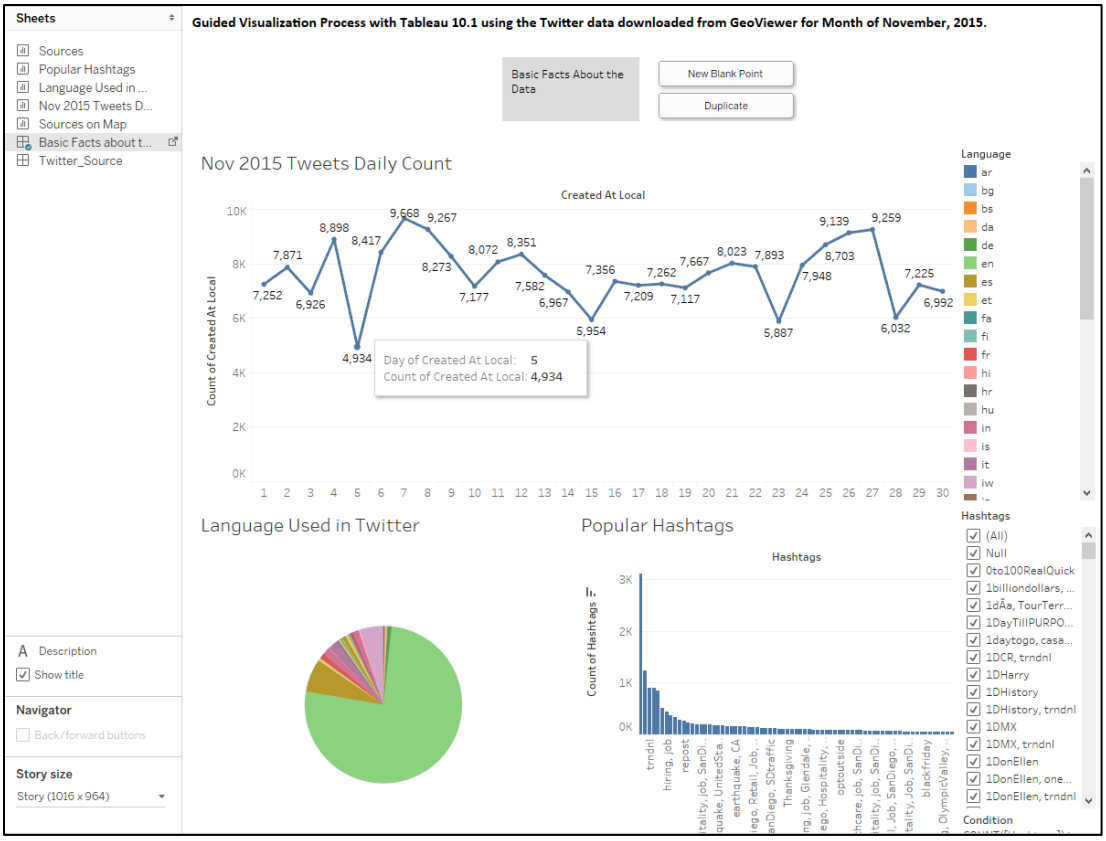
Storyline is another powerful function of Tableau which allow you to combine the dashboards into one for presentation. You can create a logic workflow of the project.

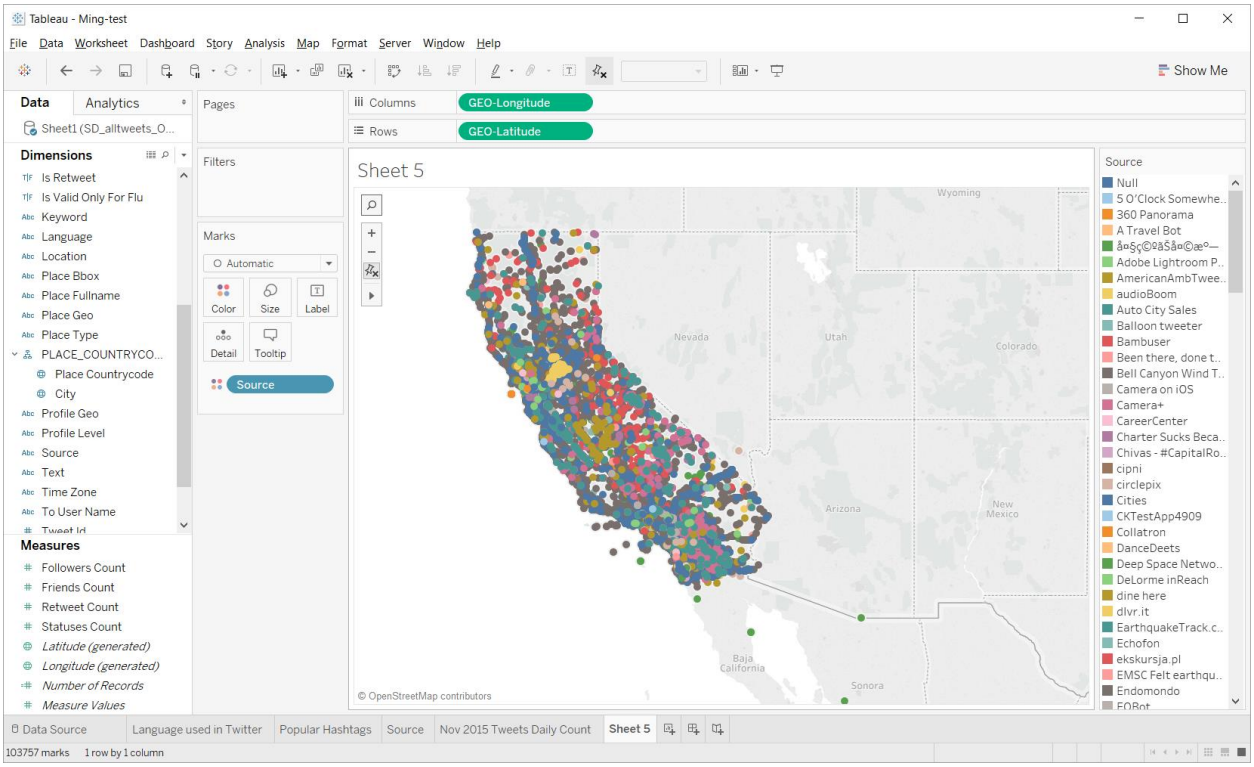


1. To do so, click on **[Create Storyline]** (the last one) at the bottom of the interface.

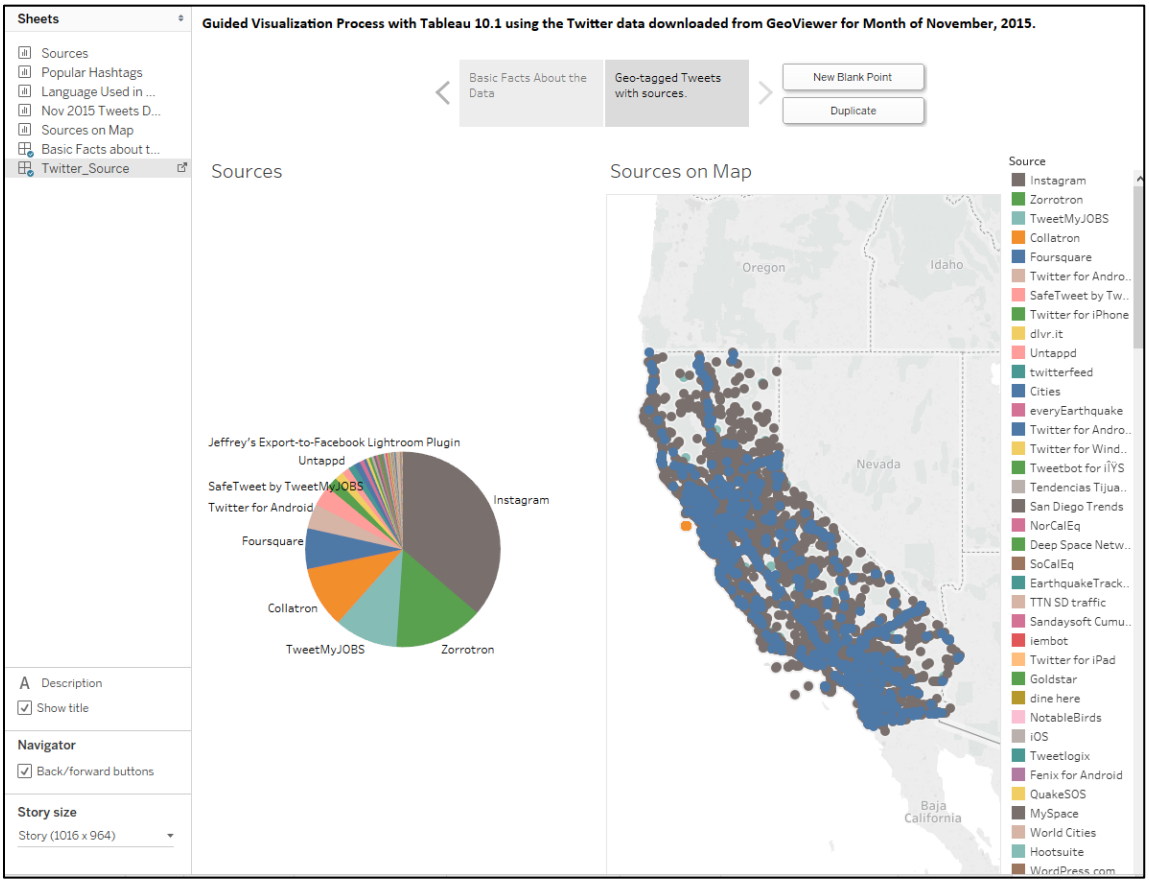
2. Double click on the [Story Title] and Type in “**Guided Visualization Process with Tableau 10.1 using the Twitter data downloaded from GeoViewer for Month of November, 2015.”**

3. On the gray box “Add a caption”, type in “Basic Facts about the Data”. And then drag the dashboard 1 “Basic Facts about the Data” to the main screen. This is your first page of the storyline.





4. Then click on the box [New Blank Point]. A new grey box "Add a caption" appears. Type in "Geo-tagged Tweets with Sources". Then drag the dashboard "Twitter_Source to the screen.



5. Rename the storyline as "Nov 2015 Tweets" and right click on the sheet, you can choose **[Export Sheet] – Save to your computer**. This is a complete project that can be present with several visualization methods.

There are more online courses for Tableau learning, below are some websites. You can register as a student to take the online course.

- http://www.tableau.com/academic/students
- http://www.tableau.com/learn/training
- http://www.tableau.com/learn/live-training
- https://www.youtube.com/channel/UCK6g6lLDdplHTEQW6dEyphw

After finishing this Web Exercise, Please use your own words to answer the following questions (next page): (DO NOT COPY any web resources or Wikipedia texts. We will check your answers with Blackboard tools to verify that your responses are uniquely yours.) By submitting your answers (paper) to Blackboard, you agree: (1) that you are submitting your paper to be used and stored as part of the SafeAssign™ services in accordance with the Blackboard Privacy Policy; (2) that your institution may use your paper in accordance with your institution's policies; and (3) that your use of SafeAssign will be without recourse against Blackboard Inc. and its affiliates.

SafeAssign accepts files in .doc, .docx, .docm, .ppt, .pptx, .odt, .txt, .rtf, .pdf, and .html file formats only. Files of any other format will not be checked through SafeAssign.

LAB-3 Additional Assignment:

1. Please use Tableau to compare the **mean, median, and standard variation** of "Followers Count", "Friends Count", and "Retweet Count" in graphs.
2. Please use Tableau to analyze the **Lung_Cancer_Death** data in the Web Exercise 02. Create a Dashboard to show your visualization results and include the Screenshot in the report. Write a short paragraph for each sheet to explain how you create each sheet in the dashboard.
3. Compare the functions between R and Tableau. What's their advantages and disadvantages for each?

Please submit your LAB-3 Answers (in a MS Word or a PDF file format only) to the Blackboard System BEFORE the DUE DATE/TIME.